Using AI Models to Understand the Impact of COVID-19 in the Context of Long COVID and Food Delivery Operations

by

Kushagra Kushagra

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 9, 2023

Keywords: COVID-19, Medical, Network, Machine Learning, Artificial Intelligence, Big Data

Approved by

Xiao Qin, Chair, Alumni Professor of Computer Science and Software Engineering
Ashish Gupta, Co-chair, Globe Life Professor of Business Analytics
Gerry Dozier, Charles D. McCrary Eminent Chair Professor in the Department of Computer Science and Software Engineering
Cheryl Seals, Charles W. Barkley Professor
Pankush Kalgotra, Assistant Professor of Harbert College of Business

Abstract

Long COVID is unlike the usual diseases with well-defined symptoms or any medical test parameters. According to the CDC (Centers for Disease Control and Prevention), Long COVID is broadly defined as signs, symptoms, and conditions that continue or develop after acute COVID-19 infection. The definition is so broad that any two Long COVID patients may have entirely dissimilar symptoms or medical problems and this makes it almost impossible to come up with any discernible means to predict Long COVID. This paper proposes a novel method to predict Long COVID by using network analytics and machine learning. Our methodology has no medical basis but the methodology used can help in furthering the research by medical experts. The usual method of dimensionality reduction such as PCA (Principal Component Analysis) did not help much. Community Detection algorithms using a bi-partite network between COVID-19 and Long COVID patients provide extra information to remove features from over 3,500 to fewer than 300. The dimensionality reduction obtained in such a manner coupled with improvement in the ratio of Long COVID to Non-Long COVID cases in the datasets by controlling the number of Non-Long COVID cases has a strong impact on the performance of machine learning models. We identified two demographic groups for our study adult and pediatric because of differences in their vulnerabilities and immunity levels. Although the results of network analytics are different, prediction accuracies through LSTM and neural networks are above 90% in both cases.

COVID-19 has also had an adverse global impact on various industry sectors. It has led to significant changes in societal behavior. Social distancing made public spaces hazardous, shifting consumer habits, including purchasing and spending patterns. People have been driven toward online resources and delivery services, causing disruptions and impacting industries. The study investigates and identifies new online food delivery patterns that

emerged during COVID-19. We focus on the food delivery industry in a University town, integrating 183 restaurants to understand how e-commerce and consumer behavior with respect to restaurant food delivery changed from pre-COVID to the COVID-19 times. We use AI and machine learning techniques to collect and analyze data collected over three years. Findings suggest that new emerging patterns require adaption to the variability in the types of food that consumers are ordering, ordering times, delivery locations, etc. Such insights provide for resource planning and allocation decisions.

Acknowledgments

This dissertation would not have seen the light of day without invaluable guidance, experience sharing, unflagging support, and encouragement from my advisors, colleagues in our research group, and my parents during my study at Auburn University.

First and foremost, I would like to extend my most sincere and deepest gratitude to my advisors, Dr. Ashish Gupta and Dr. Xiao Qin, for providing their unparalleled mentorship, reposing trust, supervision, and immaculate guidance. I shall ever remain enlightened by their extensive knowledge in the field of Machine Learning and Artificial Intelligence. Their passion and fervor for research have kindled enthusiasm and optimism in me for research, which will go a long way towards achieving many more milestones in my research. Their insightful advice and suggestions at all times have propelled me to conduct the research with finesse maintaining the highest standards. I would also like to convey my special thanks to Dr. Pankush Kalgotra for his constant guidance and support during my research.

I am highly obliged to my committee members, Dr. Gerry Dozier, Dr. Cheryl Seals, and Dr. Pankush Kalgotra, who reviewed my proposal and dissertation documents for providing me with their precious advice and feedback. Their valuable suggestions helped me in improving my dissertation substantially. I extend appreciation to Dr. Beverley Rilett from the core of my heart for acting as a University Reader. In addition, I am thankful to all the professors and students in the Department of Computer Science and Software Engineering who have contributed to an excellent academic atmosphere that is congenial for learning and training.

I am also indebted to Late Dr. Ashutosh Mishra, Visiting Professor Emeritus in Mechanical Engineering at Auburn University. He encouraged and guided me to join Auburn University and remained a constant source of inspiration throughout my life.

I want to thank my friends and extended family for their support during difficult moments which has given me the strength to keep fighting for my research. They motivated me to achieve my goal and to overcome all hurdles.

Most importantly, I have no words to express my gratitude towards my parents for the endless love that they have given me over the years. They have been my pillars of moral and emotional support and comfort. They have always been there for me through thick and thin.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The devastating impact of the COVID-19 pandemic extends to a global scale. COVID-19 has presented distinctive challenges for the various communities, healthcare professionals, and common individuals. The echo of devastation has reverberated across a multitude of industries, including healthcare, manufacturing, tourism, transportation, and supply chains, to name a few [90]. These challenges stem from factors such as a limited scientific understanding of the virus, its potential to cause high mortality rate, the often asymptomatic nature of carriers, its unpredictable severity, and the emergence of lingering, long-term symptoms [106]. COVID-19 has had a long-term effect on both health and economy. Research has shed light on the diverse array of symptoms that can afflict the human body as a result of COVID-19. These symptoms encompass fatigue, muscle or body aches, shortness of breath, headaches, and more. Notably, a comprehensive study revealed that all respondents reported experiencing fatigue, while 68.1% mentioned muscle or body aches, 67.0% struggled with shortness of breath and difficulty breathing, and 57.9% endured headaches [93]. More symptoms and conditions have been rampant with the ever-evolving COVID-19 with the advent of new strains, such as Omicron, Delta, Beta, alpha [30]. Unlike the strains of COVID-19, another issue of the virus is its tendency to linger in the body for longer than the expected duration. This phenomenon is termed "Long COVID" and as per the CDC, at least 4 weeks after infection with COVID-19 is the start of when Long COVID is potentially identifiable [21] [81].

In a recent article, a patient states - “I’m 29 years old and I feel like I’m 70”, as he struggles for months if not years from this disease [50]. Long-COVID is a new health disaster with the drastically changing pandemic as medical experts are studying the historical records

and diseases of patients. With the aid of the National COVID-19 Cohort Collaborative (N3C) and the National Institute of Health (NIH) [94], we have been able to work on medical records of over 1 million COVID-19 adult and pediatric patients in over 80 medical care units across the USA [49]. Due to Long COVID-19 not being investigated enough, we find a lack of papers, Artificial Intelligence and Machine Learning models, and disease networks. Among the multitude of factors that lead to a patient suffering from Long COVID-19, we will delve into the comorbidities a patient suffers and determine the possibility of such patients facing the impending adversity of Long COVID. Research regarding Long COVID is minimal and the nature of Long COVID still remains not unspecific, therefore, there is likelihood of further delay in research. As of now, Long COVID is not well defined in medical terms and there is no specific symptom(s) or sign(s) [38].

Along with complications in long-term health, COVID-19 has an economic effect leading to companies shutting down [60, 25]. Campbell discusses how ontological security leads consumers to incorporate such changes to adapt to such drastic changes. This attack on security is caused by the disruption of the consumers' beliefs, values, and norms as they encounter a sudden break from the consistency of their lives [18]. As a result, organizations are having a difficult time understanding and adapting their business to best fit the current needs and perform short-term operational planning [103]. People are required to take precautionary measures and maintain social distancing. This has made going to public or crowded places extremely dangerous [54, 95, 118]. Over 43% of businesses had temporarily closed by early April 2020 based on recent research [11]. As an example, in the Mid-Atlantic region, 54% of firms were closed down while employment went down by 47% [11]. Small or local brick-and-mortar companies are dealing with many issues as they face difficulty adapting to the drastic shift in online deliveries as they lack the resources or experience. People are losing jobs as demand from these companies declines rapidly and in-person working environments become unsafe for employees. In this dire situation, people shift to online resources and delivery to meet their demands. This has squeezed both the demand and supply leading to a downturn

in the economy and disturbing almost all industries [121]. Online delivery has witnessed a massive surge of incoming consumers since March 2020 and has strongly impacted the buying patterns of customers. Deliveries during March 2020, have seen an increase in 42% where most of the sales were from Amazon with 60% followed by Walmart with 47% [108].

Section 1.1 articulates the reasons behind our research. We discuss the nature and newness of such a disease along with the different impacts it has on people. We discuss not only the health effects of Long COVID but also the impacts of COVID-19 on a local business. Thereafter, we dive into the contributions made in the research in Section 1.2 and finally we explain the organization of the dissertation in 1.3.

1.1 Motivations

In this section, we discuss the three motivations that spur this dissertation research about COVID-19 and Long COVID. More specifically, Section 1.1.1 shows the reasons behind our research on Long COVID and why the prediction of Long COVID is crucial in medical healthcare. Finding comorbidities and their importance in this research is described in Section 1.1.2. In section 1.1.3, we look into the impact of COVID-19 on a local food delivery business.

1.1.1 Motivation 1: Long COVID and its Unspecificity

Research regarding Long COVID is very few and the nature of Long COVID still remains not very specific, there is likelihood of more delay in research. As of now, Long COVID is not well defined in medical terms and there is no specific symptom(s) or sign(s). When the problem is still in amorphous state, the solution is definitely going to be elusive. This paper is dealing with such type of issue and as a result, this is very challenging. According to the CDC [21], the symptoms of Long COVID vary widely - fatigue, shortness of breath, cognitive dysfunction (also called brain fog), post-exertional malaise, gut-related diseases,

etc [99, 113, 114]. Especially in the case of pediatric patients, it has been relatively overlooked [15, 12].

1.1.2 Motivation 2: Comorbidities and its Value

A majority of survivors of COVID-19 report manifestations that persist beyond the acute illness, so-called Post-Acute Sequelae of SARS-CoV-2 (PASC, or "long COVID"). Long COVID can affect even those who were initially mildly symptomatic or asymptomatic, may include a constellation of neurological, respiratory, cardiovascular, and gastrointestinal symptoms, and is debilitating in some affected individuals. In a study, a total of 287 phenotypic abnormalities were identified and represented as HPO (Human Phenotype Ontology) terms [29]. Hypertension and cardiovascular diseases are among the most common coexisting conditions. The most common symptoms at the initial stage of illness were fever, fatigue, dry cough, anorexia, chest tightness, myalgia, mild shortness of breath, chill, and dyspnea [125]. COVID-19 has simply not been with us long enough to decipher its long-lasting effects on infected patients. This is more prevalent in the case of patients below 18 as there are multiple children or birth-related complications relevant to Long COVID [98, 75]. Due to its newness and uncertain numbers, Long COVID has flown under the radar for many medical experts and practitioners. In fact, the percentage of Long COVID among COVID-19 patients is seen to vary from less than 5% [26, 62] to up to 60% [133, 130]. Such a range makes it nearly impossible to pinpoint the nature of Long COVID [17]. Another aspect not covered in several studies is the usage of machine learning and artificial intelligence models for Long COVID prediction. These conditions motivated us to study the long-term after-effects of such a disastrous pandemic.

Moreover, we have witnessed the escalation of delivery services during the times of the pandemic or when COVID-19 spikes occur throughout the year. One of the challenges faced by the delivery company is to identify factors affecting the performance of the key stakeholders and the demand of the customers. Remedial steps pertinent to the specific

target groups ought to be designed and implemented that may be the next order challenge in terms of engaging appropriate agency and incurring financial burden [69]. The erratic nature of demand is the greatest challenge during the pandemic. Grocery delivery helps in hedging the volatility in the food delivery demands, but the challenge is to find such correlated items to deliver and tie up with the suppliers [14]. The variability in the number of orders and the value of each order is another challenge. It puts undue pressure on the logistics and performance of the company as perceived by the customers. The company cannot have too much slack in the logistics to meet a sudden surge in demands. Food delivery companies aim to optimize efficiency to maximize value while cutting back waste. Thus, the companies have an incentive to decrease slack, as stated earlier, to meet the demands of a large group of customers in these trying times [132].

1.1.3 Motivation 3: Impact on Local Food Delivery

Due to high variability, models prior to COVID-19 for sales and customers in the field of E-commerce have necessitated recalibration and development of new models that account for this change. The large size of the industry coupled with the diversity of variables consisting of the type of food items ordered, the time of ordering, restaurants ordered from, frequency of orders, the places of orders, and to name just a few [48] enable huge data for machine learning techniques and analysis. A different demographic pattern in university towns motivates us to construct models to delve into food-delivery services, where most of the population in concern are students, faculty, and staff of the university. This company integrates 183 restaurants and provides delivery services. Various machine learning models (e.g., heatmap and K-means) have been applied to data collected over 3 years by the delivery industry [61, 79]. In fact, there exist studies that anchor into a large sector of E-commerce [82, 4], but none of these existing studies was dedicated to the food delivery industry in a university town or implemented Machine Learning techniques, especially during a pandemic.

1.2 Contributions

Throughout this dissertation study, we made the following two contributions.

- Contribution 1: Finding prevalent diseases or comorbidities related to Long COVID for Adults and Pediatrics. Please refer to Section 1.2.1 for the details.
- Contribution 2: Implementing a system for the prediction of Long COVID for Adults and Pediatrics. Please refer to Section 1.2.2 for the details.
- Contribution 3: Analyzing the impact of COVID-19 on a Local Food Delivery Service. Please refer to Section 1.2.3 for the details.

1.2.1 Contribution 1: Long COVID Comorbidities

Patients have a history of medical illness(es) which doctors and medical experts have used for their analysis and diagnosis of future diseases. Such cases indicate the possibility of relationships between diseases. In our research, our goal is to find the prevalent diseases while helping medical experts in the diagnosis of Long COVID. In a situation, where no path to the final goal is known to any community and the only source of knowledge is from the data provided by the National Institute of Health NIH in the N3C Enclave. With the multitude of patient records provided by NIH, diagnosing estimates for specific conditions has become viable using big data [131]. We were motivated to use data analytics to identify patterns of relationships/communities that may lead to future investigations by different domain experts using Gephi. With the implementation of network algorithms, we have visualized such patterns using two algorithms. Louvain is a modularity-based algorithm that is widely popular [23] and constructs well-connected communities [96, 40, 124]. Its implementation with the result of the projections formed from Bipartite graphs between Patient-Disease relationships. Our experimented data consists of confirmed cases of 1 million adults and 340 thousand pediatric COVID-19 patients. Furthermore, manage the dataset to find which patients have which diseases during set periods. This helps calculate coefficients and relevant

diseases with the help of our community detection results. With the help of certain machine learning models, we are able to create a process to predict Long COVID amount COVID-19 patients and other relevant diseases. These methods yield more information for medical practitioners and doctors to help with the ongoing research on Long COVID. The main concerns that we address in our studies for both adult and pediatric patients are - 1. What patients can have Long COVID based on their past disease history diagnosis? 2) What other diseases are likely to occur or are relevant with Long COVID?

1.2.2 Contribution 2: Prediction of Long COVID

In light of the above-ground realities and concerns, this paper aims to discover the diseases that may result from Long COVID and the patients who are likely to be victims of Long COVID. To start with, we have analyzed the data collected by NIH regarding COVID-19 patients across the US. At the next level, we are trying to establish a relationship using various techniques between the Pre-COVID and Long COVID patients. We have used network analysis and visual representations of the Pre-COVID and Long COVID diseases, correlation, etc. At the next level, we intend to apply Machine Learning techniques to develop a model for the prediction of Long COVID. This model will help medical practitioners to dive further into it in explore the medical reasoning behind the predictions. In other words, our research work will provide assistance and direction for further research.

1.2.3 Contribution 3: Analyzing Impact of COVID-19 on Food Delivery

For our food delivery study, We implement heatmaps, segmentation analysis, temporal analysis, association rule mining (ARM), and K-means clustering. We generate heatmaps using Restaurant-wise daily sales values and daily count VS company's daily sales value and count, hourly sales value VS company's daily sales value, etc. K-means clustering is implemented to figure out how the customers are grouped together on the basis of their frequency of orders. ARM is implemented with menu items in all the restaurants to figure

out related items ordered in the same basket. We further go into detail with timeline graphs and box-plots to depict the ups and downs and the overall statistical effect COVID-19 has on the delivery company. What impact COVID-19 has had on food delivery services? Can we develop a better understanding of shifts in food delivery patterns during COVID-19 that could help them effectively manage customer demand patterns and resources? - Are two questions addressed in this study? We embark on this study by investigating a relatively new food delivery company's sales and performance from January 1, 2017. The normal sales duration lasted till March 1, 2020. The period after that is considered COVID-19 impacted. The company involved in the study is relatively new and was established within the last five years and has shown a significant rise in sales over this duration.

1.3 Organization

The rest of this dissertation is organized as follows. The upcoming chapter explains related work done in this field of study. In Chapter 3, we explain the data engineering and pre-processing techniques implemented before delving into our analysis. We also explain the environment and tools used along with the workflow of the study on Long COVID. In Chapter 4, we look into the network analytics methods used on our data, including community detection algorithms. This method shows the diseases relevant to Long COVID among patients.

In Chapter 5, we discuss how the results of the network models were used for machine learning models and how that helps predict the possibilities of Long COVID. Chapter 6 presents the impact of COVID-19 on businesses and how a local food delivery company has been affected by the pandemic. Chapter 7 explores the implications of our study and how they can be helpful to researchers and practitioners. Finally, Chapter 8 concludes the dissertation study and points out the future direction of the research.

Chapter 2

Related Work

We start this chapter by delving into the related work that has been researched regarding machine learning in the realm of Long COVID and then on e-commerce. We discuss the prior studies focusing on these research topics and intend to shed light on the impacts of the pandemic. There have not been a great number of studies conducted for Long COVID, especially on network analytics with a machine learning model. This is even more true for pediatric patients. The economic Impact of COVID-19 has been delved into before but few have touched the surface of how food delivery has been affected by the pandemic.

In section 2.1 we discuss the prior research conducted on the effects of Long COVID on Adult Patients. In this section, we dive into the previous studies done on disease comorbidities and disease networks in the subsection 2.1.1 and 2.1.2. After that, in Section 2.2.1, we look into prior studies on the symptoms of Long COVID and again on disease networks in section 2.2.1 and section 2.2.2, respectively. Thereafter, we look into the impacts of COVID-19 on E-commerce in section 2.3.1 and, subsequently, on food delivery services in section 2.3.2 and section 2.3.3. Finally, we discuss the summary of the studies in section 2.4.

2.1 Long COVID-19 and its Effects on Adult Patients

According to the research paper "Long Covid—mechanisms, risk factors, and management" by Harry Crook [27], with many people having been infected and continuing to be infected with COVID-19, the long-term implications are of increasing concern. The authors of this paper have reviewed the studies that have explored the persisting symptoms of Long COVID and have addressed the possible risk factors associated with developing Long COVID and the treatment options that may be useful in alleviating its symptoms. Currently, Long

COVID-19 remains enigmatic [38, 17] and, with the question of the impact that new variants of COVID-19 will have on the incidence and severity of Long COVID still looming large, it is important that research continues to explore post-COVID-19 syndrome [37, 127]. A greater understanding of the pathogenesis, risk factors, symptoms, and methods of treating Long COVID is required to reduce the strain and demand on people with the condition and the healthcare systems that will endeavor to support them [78, 91, 92].

The article "Characterizing Long COVID in an International Cohort" [28] mentions the findings of the responses from 3762 adult participants with confirmed (diagnostic/antibody positive; 1020) or suspected (diagnostic/antibody negative or untested; 2742) COVID-19, from 56 countries, with illness lasting over 28 days and onset prior to June 2020, prevalence of 203 symptoms in 10 organ systems was estimated and 66 symptoms over seven months were traced. Davis inferred that those patients with Long COVID report prolonged, multi-system involvement and significant disability, and by seven months, many patients had not recovered (mainly from systemic and neurological/cognitive symptoms), had not returned to previous levels of work, and continued to experience significant symptom burden.

Long COVID has severely affected the health quality of thousands of people around the world [13]. Due to a lack of direction and unsure feelings looming around this new phenomenon, people are unaware of their treatment [68, 110]. This has become a deeper issue as Long COVID has not shown proper patterns and affects varying organs throughout the body [3, 107].

2.1.1 COVID-19 and Comorbidities

Our research also involves finding related diseases that can indicate a pattern regarding Long COVID. Several studies in the area of COVID-19 have helped give a direction regarding related diseases [34, 45, 57, 32]. For example, a study by Sanyaolu *et al.* [101] found that older people, especially those 65 years and older, have a higher chance of infection. According to a paper titled "Patterns and temporal trends of comorbidity among adult patients with

incident cardiovascular disease in the UK between 2000 and 2014: A population-based cohort study" by Jenny Tran *et al.* [119], the most common comorbidities in age/sex-standardized models were hypertension (28.9% [95% CI 27.7%-31.4%]), depression (23.0% [21.3%-26.0%]), arthritis (20.9% [19.5%-23.5%]), asthma (17.7% [15.8%-20.8%]), and anxiety (15.0% [13.7%-17.6%]). On average, older patients, women, and socioeconomically deprived groups had higher numbers of comorbidities, but the type of comorbidities varied by age and sex [111]. Although Cardiometabolic conditions contributed substantially to the burden, 4 out of the 10 top comorbidities were non-cardiometabolic [105].

T2DM (Type 2 Diabetes Mellitus) was also among the most common comorbidities of EH (essential hypertension) [24], CHD (coronary atherosclerosis and other heart diseases), and ACVD (acute cerebrovascular disease) and these were the most popular comorbidities of T2DM. T2DM were the first, second, and third popular comorbidities of EH, ACVD, and CHD (absolute co-occurrence risk ACoR 29.8%, 23.0%, and 25.9%, resp.). Female patients with EH, CHD, or ACVD showed consistently higher proportions of having T2DM than male patients [85, 72]. When taking patient sex or age into consideration, major comorbidities varied for the particular populations. 22 out of 27 overall major comorbidities, such as EH, DLM, and chronic renal failure (CRF), remained the major comorbidities for both male and female patients, whereas biliary tract disease and noninfectious gastroenteritis for male patients and thyroid disorders plus other two diseases for female patients could no longer be considered as major comorbidities.

Out of 5700 patients hospitalized with COVID-19 in the New York City area, the most common comorbidities were hypertension, obesity, and diabetes [100]. Among patients who were discharged or died (n=2634), 14.2% were treated in the intensive care unit, 12.2% received invasive mechanical ventilation, 3.2% were treated with kidney replacement therapy, and 21% died. In another paper on COVID-19 Comorbidities [87], the authors included a total number of 375,859 participants from 14 countries, namely, Brazil, China, India, Iran, Italy, Mexico, Oman, Saudi Arabia, South Korea, Spain, Turkey, Uganda, United

Kingdom, and United States and concluded that among SARS-CoV-2 infected patients the three most prevalent comorbidities were hypertension, obesity, and diabetes amounting to 80,093 (21.3%), 68,935 (18.3%), and 67,954 (18.1%) patients respectively.

A similar study in China "Prevalence and Risk Factors of Comorbidities among Hypertensive Patients in China", [126] concluded that patients with hypertension have three important comorbidities diabetes mellitus, hyperlipidemia, and coronary heart disease. Our research pertains to disease networks to find correlations between pre-existing diseases. Checking for similar symptoms among patients has shined light on the importance of finding comorbidities among patients. A study in 2014 by Barabási and Zhour [9, 129] used a Human Symptoms Disease Network (HSDN) to display how the weighted links between any two diseases result in the relationship between the symptoms and the disease. 94.5% of the taken records ended up providing meaningful and direct results. They have also implemented a similarity index for the Pearson Correlation Coefficient to calculate the ratio of shared disease links and disease similarity. This resulted in a high (PCC = 0.96, p = 1.4*10-5) which solidifies the foundation for using these measurements as reliable tools. The studies by Barabási and Zhou are especially relevant to our literature review as they combine comorbidities and networks. These are aspects that we tackle in our own analysis.

2.1.2 Disease Network Studies Conducted on Adult Patients

In the article titled "The Human Disease Network" by Goh [41], the authors have explained the concept of the Human Disease Network using a diseasome bipartite network on the basis of disease genes and disease phenomes. Human Disease Network is the projection of this bipartite network. From the diseasome bipartite graph, two biologically relevant network projections are generated – the human disease network and the disease gene network. In the "human disease network" (HDN) nodes represent disorders, and two disorders are connected to each other if they share at least one gene in which mutations are associated with both disorders. In the "disease gene network" (DGN) nodes represent disease genes, and two

genes are connected if they are associated with the same disorder. If each human disorder tends to have a distinct and unique genetic origin, then the HDN would be disconnected into many single nodes corresponding to specific disorders or grouped into small clusters of a few closely related disorders [42]. In contrast, the obtained HDN displays many connections between both individual disorders and disorder classes. Of 1,284 disorders, 867 have at least one link to other disorders, and 516 disorders form a giant component, suggesting that the genetic origins of most diseases, to some extent, are shared with other diseases.

Genís Calderer and Marieke L. Kuijjer in their paper titled "Community Detection in Large-Scale Bipartite Biological Networks" [16] mention the detection of communities in large biological networks. They wanted to evaluate whether the communities they discovered were enriched for specific biological processes. For each method, they ran GO enrichment analysis [65] on the selected communities. All methods resulted in communities that were significantly enriched for biological pathways. This high level of enrichment confirms that the retrieved communities likely represent true biological information. A t-test concluded that there was no difference between the significance of the results for each method. The final community structure obtained by bi-Louvain with Murata+ offers a relationship between communities of each of the bipartite sets.

While unipartite community detection has been widely applied to large-scale biological networks [23, 71, 63], community detection on bipartite networks and, in particular, on genome-wide bipartite networks, has been less studied. However, as many types of biological networks are bipartite, it is important to review community detection approaches that are specifically designed for such networks [16]. The relationships obtained through the bipartite network between symptom clusters (SCs) and disease conditions have been buttressed by underlying molecular mechanisms across different disease conditions [76]. The study found that some SCs are associated with a single body system, but some SCs have strong relationships with multiple body systems. For instance, there are multiple body systems (e.g., urinary system, circulatory system, immune system, etc.) associated with SC47 showing

diversity [122]. Such findings support our approach to applying network analysis for Long COVID cases which, as of writing, appears to encompass multiple body systems.

2.2 Long COVID-19 and its Effects on Pediatric Patients

After discussing the related work with Adult patients, we now move to research that has been covered for pediatric patients. Pediatric Long COVID research has been explored less than adults as Long COVID is more prevalent among older people. One of the studies by the CDC regarding this subject found that Long COVID has only affected 1% of the population of COVID-19 pediatric patients [102]. This number has fluctuated vastly over the years and precisely finding which pediatric patients are confirmed to have Long COVID also remains a challenge [75, 47].

An analysis of 129 studies from 31 countries comprised of 10,251 children of which 57.4% were hospitalized [55]. Children primarily faced mild forms of infection. Unfortunately, the study suggests that they are at risk of more severe outcomes. This analysis presents a comparison between the clinical symptoms, management, and outcomes of the reported pediatric patients. In fact, the severity of the disease was defined by studying each individual. This was done while considering parameters, such as admission to intensive care (ICU), usage of ventilation, multi-organ failure, along with the presence of hypoxia. Establishing the severity parameter with a quantitative value is an exceedingly challenging task because it is more of a qualitative attribute. Multiple studies define it in different ways making the evaluation and comparison among them difficult to represent a meaningful and consistent outcome.

A lot of research is conducted and summarized in the comprehensive review by [43]. The authors observed that most of the models providing good prediction accuracies used image datasets, such as X-rays or CT scans to report the results. However, the approach for analyzing chest CT images in pediatric patients suffers from two main disadvantages: cost and risk of developing cancer due to radiation exposure. In addition, an important concern

is related to confidentiality and privacy of personal medical records, especially accessing data for patients under 18 years old. Last but not least, it is crucial to use real data to avoid biased results when applying predictive models. Most of the studies have been using public repositories or synthetic data which does not guarantee the uniqueness of the records and the quality of them. Moreover, sometimes the number of records gathered is not enough to train and test predictive models, leading to poor and inaccurate outcomes.

2.2.1 Long COVID Symptoms in Pediatric Patients

As mentioned earlier, estimating children that are more likely to have a propensity for developing Long COVID symptoms is a monumental task as of writing this paper. Chahar [22] suggested machine learning techniques on pediatric cases with COVID-19 infections to predict the results of CR scans by using clinical laboratory data and RT-PCR positive results. The number of scans represented a small sample size of 200, which can result in biased conclusions when training and testing machine learning models. As mentioned above, the usage of image data for pediatric patients has some leak points. Huijing [53] reported a case study where five children from Sweden aged 9–15 had experienced COVID-19 symptoms for over 2 months after clinical diagnoses of COVID-19, in which females appeared to be represented more than males. They all reported to have fatigue, dyspnea, chest pain, and heart palpitations, and four of them suffered from headaches. They also had difficulties concentrating with muscle weakness, dizziness, and sore throats. Another study, [77] was based on 58 children and adolescents who reported to suffer Long COVID symptoms. These were reported symptoms that included fatigue in 12 (21%), shortness of breath in 7 (12%), exercise intolerance in 7 (12%), weakness in 6 (10%), and walking intolerance in 5 (9%) individuals. Older age, muscle pain on admission, and intensive care unit admission were significantly associated with Long COVID in adult patients. A very intriguing outcome by [7], they observed an increased rate of Type 1 diabetes among US children during

COVID-19. The lack of social activities appeared to cause long-term adiposity in children in Singapore, one year after the lockdown [44].

2.2.2 Disease Networks using Bipartite Graphs

Bipartite networks are important for designing systems for complicated patient or disease networks that apply in the real world [89]. Thus, our goal is to find out and identify the communities that provide meaningful outcomes. The concept of communities is very frequent in social networks, biological networks, and so on. Communities play a crucial role in understating human diseases, and they cannot be explained only on degree distribution, but it is connected to the fact of who connects to whom [115]. We require algorithms to figure out such communities because the Bell number is not an efficient strategy when dealing with millions of nodes in a network. K.K. Sum [115], explains two algorithms used to detect communities. One of them is the Ravasz algorithm, which is an agglomerative algorithm that utilizes the average cluster similarity with the Girvan-Newman algorithm. These two techniques create a hierarchical tree, called dendrogram which is further optimized by using modularity in order to decide an optimal cut for large networks. Due to this optimization, the application of the Louvain algorithm is extended to detect communities within expansive networks, including Leiden and various other versions. In the context of human diseases networks, the occurrence of overlapping communities is quite common, as nodes often belong to multiple communities simultaneously. Consequently, two algorithms, namely the clique percolation method and the link clustering algorithm, are utilized to pinpoint these overlapping communities.

Our research involves extensive data and a time-evolving network. Within this framework, associations between patients and diseases, as well as between different diseases, are generated at a massive scale across various timeframes. When the number of edges becomes exceedingly dense, it becomes impractical to comprehensively read and interpret the results. To address this issue, some studies have employed the concept of tripartite graphs to track

and uncover valuable insights pertaining to COVID-19. In fact, the implementation was based on the social network, Twitter data [74] which is different compared to our data domain. Moreover, [74], recommended an improved bipartite network projection method that detects metabolite-disease relationships based on linear neighborhood similarity. KATZ model and Bipartite Network Recommendation Algorithm (KATZBNRA) were implemented in such research to discover likely associations, such as micro-disease [70], metabolite-disease [74], CircRNA- disease [73], [33]. In our study, the methodology is unique and adapted to the environment where the data is stored.

2.3 Food Delivery

We start this section by delving into the related work that has been researched regarding machine learning in the realm of e-commerce (see Section 2.3.1). Then, we discuss the prior studies focusing on food delivery services in Section 2.3.2. Section 2.3.3 is intended to shed light on the impacts of pandemics or natural disasters on food delivery services.

2.3.1 Machine Learning with E-commerce

Machine Learning has a considerable role in the field of E-commerce and its uses are witnessed in a huge variety of cases. For example, machine learning has been adopted to automatically classify items during online shopping [31]. This study demonstrates how unstructured and unclassified data can be handled for the classifications while describing the data in multiple dimensions. The proposed processes include data collection, pre-processing of data, attribute selection, and customer grouping. Gupta and Pathak devised multiple modeling techniques, such as agent-based modeling, data-driven modeling, and auction-based modeling, where new and further frameworks are to be proposed to obtain more comprehensive results [97].

From the various approaches mentioned above, we move on to some examples where these techniques have been implemented. Habault *et al.* demonstrated that machine learning techniques have been implemented in the case of a dispatcher of commercial goods by tracking the routes and finding efficient means to offer better delivery access and maneuverability [46]. Moreover, Sytian discovered that E-commerce has had a significant boost thanks to mobile applications and the ease of access. Because of the massive influx of users, there are large amounts of data stored that have surfaced several opportunities for advanced machine learning schemes (see, for example, the aforementioned techniques) [116]. In fact, the above scenarios barely scratch the surface of the multitude of uses the pair of machine learning and E-commerce has brought to society.

2.3.2 Food Delivery Services

According to a recent study conducted in May 2020, the food delivery section of E-commerce has an enormous market with a value of $138.39 million in 2019 and is expected to grow even further at a compound annual growth rate [1]. Hirschberg *et al.* articulated that with the advent of accessible technology, people are able to easily order food and take advantage of E-commerce. Thanks to this ease of access to a large audience, one can dig into the patterns of customer behaviors such as platform preferences, delivery times, place of delivery, spike of orders during weekends, and to name just a few [19]. A statistical representation by Hirschberg *et al.* shows that an average of 77% of consumers would not or would rarely change their platform for ordering deliveries. A study by Jeng suggests that saving time is a vital factor for consumers when ordering a product; such reasons are highly important to retailers [58].

There is a positive correlation when considering behavioral intentions and attitudes to the point that e-satisfaction has a strong effect on e-loyalty, as described by the *contingency framework* designed by Anderson and Srinivasan [5]. Consumers have a certain taste for what they like and assess prices rationally at a given time when ordering through any means

of E-commerce [123]. Also, a handful of studies reported in the literature [2] [120] revealed motivations to shop also, potentially, come from values and pleasure that consumer seeks from shopping [8]. With the constant assessment of such traits, we reckon that the popularity of online services has grown at such a fast pace that even day-to-day grocery delivery has become a substantial part of food delivery to homes.

2.3.3 Impact of COVID-19 on Food Delivery Services

According to a study carried out by Jones [59], traditional E-commerce is no longer what it used to be and has changed from its core substantially. Molla and Wainright elaborates in a 2020 study that during the pandemic and post-pandemic periods, technologies will be focused more on touchless or mobile technologies to avoid the spread of the disease [82]. Even with solid and relatively new approaches to E-commerce to provide services through web applications, there are open issues that can impact the industry in a negative way. Some of these issues, of course, include pandemics. There has been a lot of panic and hoarding by people all around the world. A dramatic scenario, as explained by Krakar *et al.*, was the chaotic shopping spree of canned goods and hygiene products across the shelves of supermarkets [67].

Due to the social distancing rule, an increasing number of people have opted for delivery of food and groceries. In a recent news report, Ocado, an online food delivery company had to suspend its services due to enormously large demands, which became a strenuous task to cope with. In fact, many companies are facing difficulties with the high volume of bookings and proper timing for deliveries [56]. Also described by Mussell *et al.*, online food delivery or distribution systems are experiencing a sudden hike in orders, which has caused these systems to go out of service as the systems are unable to cope with this unexpected updraft in demand [86].

In some regions, the government has made drastic decisions to control the number of orders an individual can place to ensure there is restocking in time, thereby curbing

any vulnerabilities. Such actions orchestrated by the government are necessary for people who are more prone to the virus, especially the elderly [80]. Some food companies or restaurants have closed down as the number of customers has dwindled due to the large-scale quarantine [109]. All these issues have been actively testing the resilience of the food delivery system and even for the consumers. A growing number of consumers now have to order food and groceries online for the first time. This new adoption of a medium can improve the curve, at which more people will tend towards ordering food/groceries online. A study investigated by Hobbs in April 2020 suggests that such a shift can have a profound impact on the food economy and method of sales for food/groceries once the number of COVID-19 cases is declining [52]. Different from the above studies, our research leverages machine learning techniques to gauge the impact of the COVID-19 pandemic from the point of view of a local food delivery company in the southern states of the USA.

2.4 Summary

The main goal of this dissertation is to show and find ways of how COVID-19 has affected the livelihood of people in multiple ways. These impacts have been deemed long-term as it has affected family health and businesses. With the lack of solid groundwork figured out regarding Long COVID, it becomes ever more crucial to dive in that direction and help medical experts. Local businesses that cater to online orders have been affected by the pandemic with regard to their sale and frequency of orders. COVID-19 has affected the foundation of adulthood, childhood, and the basis of livelihood for thousands of families around the world. Moreover, COVID-19 has not shown signs of slowing down as we see new infections cycle throughout the year.

Chapter 3

Data Engineering and Pre-Processing

This part of the dissertation research is conducted on the Palantir platform through the National COVID Cohort Collaborative (N3C). The N3C currently has access to an ever-evolving size of COVID-19 patient data sets to allow scientists and experts to study COVID-19 health diagnoses, symptoms, and conditions. The data sets involved have been contributed by 82 healthcare units across the United States of America. The number of total patients involved is over 20.8 million at the time of reporting with over 8 million being COVID cases. For our research, we have taken the confirmed cases of COVID-19 for both adults and pediatrics as per the ICD-10 disease coding explained in section 3.1.3. The welcome screen for Palantir is depicted in figure 3.1.

3.1 NIH, Palantir, and Disease Encoding

The purpose of this section is to show the environment of Palantir and its role in the research analysis for Long COVID. We further explain the variables and datasets used from Palantir and their prevalence as Palantir has several datasets from NIH and medical healthcare centers around the country. These datasets and disease encoding are covered in section 3.1.1. We then explain the implementation of ICD-10 encoding in hospitals and medical care units. We elaborate on the use of RegEx for cleaning noise in the Palantir datasets and the importance of the ICD-10 encoding in section 3.1.3.

3.1.1 The Palantir Workspace

The Palantir system consists of tools and analysis applications for research purposes that range from code workbooks implementing coding environments and in-built IDE software

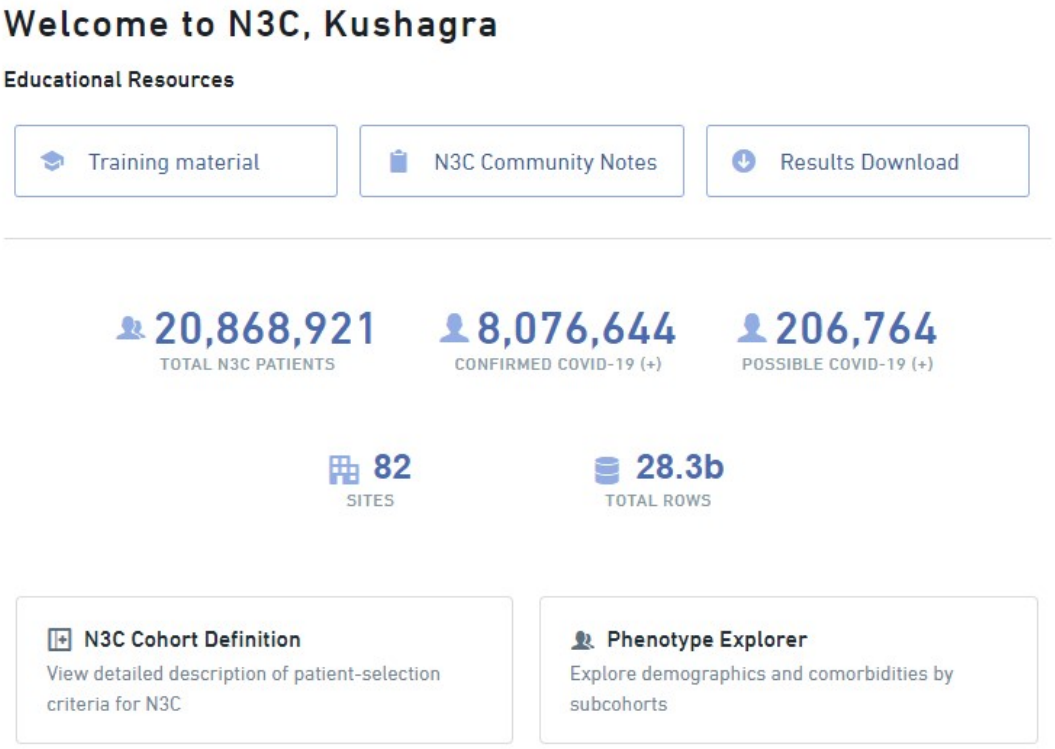


Figure 3.1: Palantir Homescreen

embedded with required libraries for machine learning approaches. My dissertation study is led in Palantir's Code Repositories and Code Workbooks that supported PySpark and Spark distributed database systems along with R. Preliminary analysis is accomplished with the help of Fusion, Reports, and Contour Tools of Palantir as illustrated in Figure 3.2. The workspace required the creation of code blocks in PySpark, R, or SQL as mentioned earlier. Our work has been primarily handled in the former two. Every code block receives an input dataset and outputs a processed dataset that is saved in the Palantir system for later use. The initial data set input for this study is Palantir's de-identified data catalogs. These catalogs consist of information regarding patients, diseases (or conditions), locations, medical tests, deaths, healthcare units, etc. Our primary dataset catalogs used are called "Condition_Occurence" and "Person". These databases provide us with the necessary variables to conduct our analysis.

- person_id: Unique De-identified ID of COVID-19 patient. This was used as our primary key to link records during analysis.
- condition_occurrence_id: De-identified unique ID of the medical case. A particular patient can have multiple medical cases. Each occurrence denotes such an instance. This instance can be a diagnosis of different diseases or the same disease multiple times.
- condition_start_date: The diagnosis date of the medical case has started.
- condition_start_date: The diagnosis date of the medical case when the disease has ended.
- visit_occurrence_id: De-identified unique ID of the particular visit of a patient
- . condition_source_value: Consists of characters used to extract ICD-10 encoding, which is explained further in section 3.1.3.
- person_id: (Same as above) Unique De-identified ID of COVID-19 patient. This was used as our primary key to link records during analysis.
- year_of_birth: birth year of patient.
- month_of_birth: birth month of the patient. Both year and month information is used to distinguish between the pediatric and adult patients
- gender_source_concept_id: Distinguishes the gender of the patient

3.1.2 How have we defined long COVID?

Based on the CDC definition of Long COVID-19, post-COVID conditions are a huge array of new, returning, or ongoing health problems that people can experience four or more weeks after they have initially been infected with the virus that causes COVID-19. We address the dataset into two splits, Pre-COVID and Post-COVID:

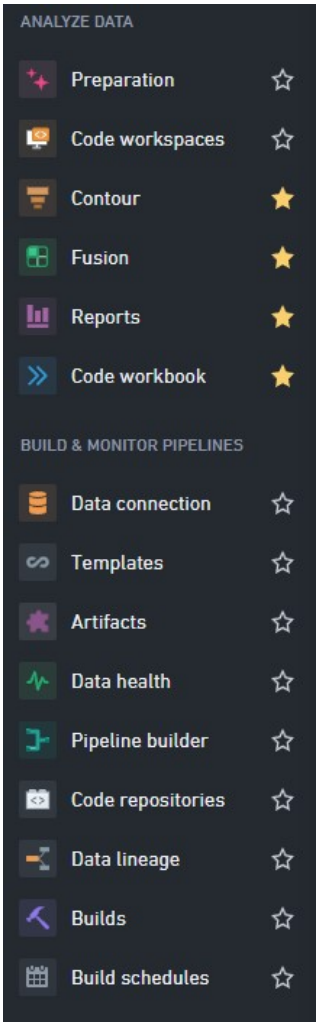
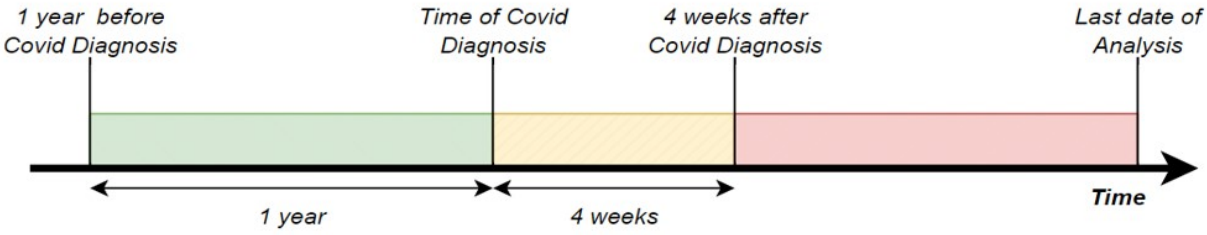


Figure 3.2: Palantir Tools

- Pre-Covid – The dataset of diseases for all the COVID-19 adult patients where the diseases have occurred within 1 year prior to the diagnosis date of COVID-19.

- Post-COVID or Long COVID – The dataset of diseases for all the COVID-19 adult patients where the diseases recorded have occurred from the end of the Normal-COVID duration (which is 4 weeks) and beyond that.

Moreover, normal COVID-19 duration is the timeline of four (4) weeks just after the diagnosis of COVID-19 among those patients. This definition is depicted the Figure 3.3.



(a) Time-windows used for a COVID-19 patient

Figure 3.3: Time-windows used for a COVID-19 patient

Adult Patients		COVID-19	Long COVID (U09)
		Total = 2,007,336	Total = 24,865
Age	Mean	51.97	55.56
	Min	18.25	18.25
	Median	51.54	56.03
Gender	Male	1,211,105	16,546
	Female	782,841	8,319

Table 3.1: Adult Patients Population Statistics

Pediatric Patients		COVID-19	Long COVID (U09)
		Total = 341,646	Total = 1,101
Age	Mean	8.57	11.24
	Min	0.14	0.11
	Median	8.48	12.55
Gender	Male	168,103	578
	Female	173,543	523

Table 3.2: Pediatric Patients Population Statistics

3.1.3 The ICD-10 Disease Encoding

The ICD-10 code, as per the CDC, was implemented for mortality coding and classification from death certificates in the United States of America [6]. ICD-10 is the 10^{th} revision of this encoding and the publication was authorized by the World Health Organization (WHO). ICD-10 has been used for several decades in healthcare units across the country.

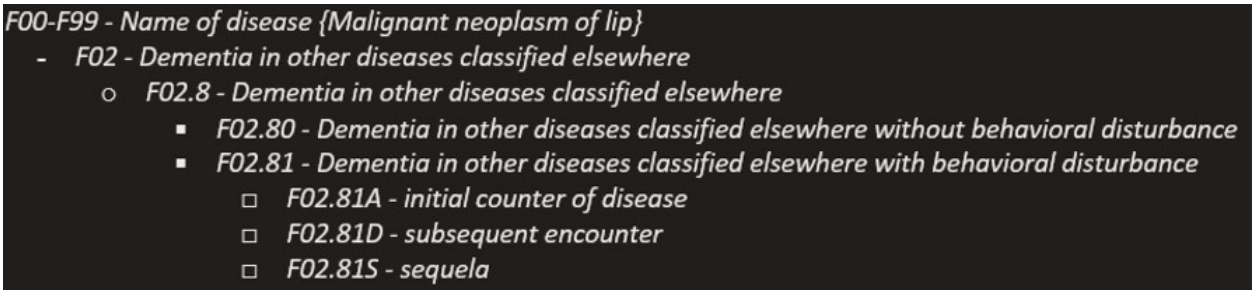
The format of ICD-10 consists of three to seven characters. The disease becomes more specific as there are more characters. This type of categorization helps pinpoint a diagnosis by medical specialists. ICD-10 encoding is categorized around the decimal point. The characters on the left of the decimal point tell us the category of the disease, condition, or injury.

The characters further to the right in this sequence tell us a more specific subcategory of that disease, condition, or injury that has been addressed before the decimal point. The final character – called an extension – signifies if this case is an initial counter, an encounter after the active phase of treatment, or a condition that has risen due to an injury.

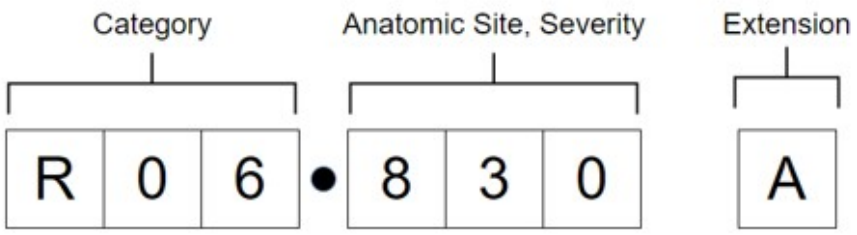
ICD-10 encoding is represented in a hierarchical manner. ICD-10 can also be said that the character more towards the right represents the specificity of the condition, disease, or injury of the diagnosis. Figure 3.5(a) shows an example of the ICD-10 functioning.

Let us take an example of R06.83 as shown in 3.5(b). The first character R represents symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified. R06 is a diagnosis of abnormalities in breathing. R06.83, specifically, is a representation of Snoring, which is a subcategory of abnormalities of breathing in the ICD-10 encoding.

Our motivation behind using ICD-10 is that we are able to acquire higher-level cases of diseases, conditions, and injuries. In the NIH datasets, diseases are recorded as Condition Occurrences, which describe the disease or medical condition at a micro-level. For instance, Unilateral femoral hernia (without obstruction or gangrene, not specified as recurrent), Transient global amnesia, Unspecified visual disturbance, etc. are ways of recording medical conditions in the NIH database. Initially, we considered the diseases mentioned



(a) ICD-10 Hierarchy



(b) ICD-10 Format

Figure 3.5: ICD-10

for our analytical purposes, but we were overwhelmed by the number of patient categories based on such diseases and any kind of analysis was thwarted by the unmanageable size of the datasets. Our approach forced us to explore the medical conditions of the patients at a higher level in order to bring patients with minor variations in their medical conditions into the same category. Here we stumbled upon ICD-10-related entries in one of the columns of the database and after a lot of juggling using RegEx and other means, we could group the patients according to the ICD-10 categories.

The ICD-10 codes are not recorded in the manner shown in Figure 3.5(a). The user inputs from the various medical healthcare units have many random characters that need to be cleaned and parsed, primarily using RegEx. An example record in the dataset found is ICD10CM:K40.90. K40 is to be extracted from this record by preprocessing. After clearing the ICD-10 codes for the diseases, we obtain 1,945 diseases. COVID-19 is denoted by U07 and Long COVID is denoted by U09.

3.2 The Workflow

This section aims to explain the workflow of the research analysis for Long COVID prediction. We explain everything according to Figure 3.6. We have divided this section into three different parts. First, we discuss the methods used for handling large amounts of medical data in subsection 3.2.1. Thereafter, we elaborate on the network analytics implementation for the workflow in subsection 3.2.2. Finally, for the workflow, we also talk about the models used for our prediction with the help of network analytics in subsection 3.2.3.

3.2.1 Phase 1: Data Engineering

In figure 3.6, we have the workflow of our analysis for Long COVID from the perspectives of Adult and Pediatric patients.

We start our process with one of the biggest and main datasets on the Palantir platform, called Condition Occurrence. This dataset includes several variables as explained in Section 3.1.1 and includes the information for every patient required. From this dataset, we use the "condition_source_value" feature and with the help of regex are able to clean the extra characters. This feature has been stored by hospitals and care units across the USA that are collaborating with Palantir and N3C (Step 1). Due to this nature, the data is supposed to show characters only relevant to the ICD-10 encoding of the diseases but human input has added unnecessary data – and please refer to Section 3.1.3 (Step 2) for the details. This process leads us to have 1,945 diseases.

After the above step, we filter all the patients based on their age. Patients above or equal to 18 are adult patients and patients below 18 are pediatric patients (Step 3). We obtain a total of adult patients to be 17.7 million and pediatric patients to be 3.1 million (Step 4). For each of these patients, we clear out the features that are not required as the Palantir system has trouble handling too many columns, especially in PySpark or R (Steps 5 and 6). This step allows us to acquire all the data for every COVID-19 patient as we have the data in the required format (Step 7).

Now, we acquire the COVID-19 patients with our extracted diseases with RegEx from Step 2 and filter all the patients that have been inflicted by U07 (code for COVID-19) as designated by CDC (Step 8). This gives us 2.01 million adult patients and 343 thousand pediatric patients in total. These patients are confirmed by hospitals and medical experts that they have COVID-19 as their condition_occurrence records indicated that they have U07. This procedure validates our subset of COVID-19 patients. We also keep all the records of these patients where they were diagnosed with any diseases 1 year prior to their COVID-19 diagnosis date (Step 9). These records will be the Pre-COVID period. This leads us to all the condition_occurrence records for COVID-19 patients, exclusively - 23.31 million adult records and 3.86 million pediatric records (Step 10). These numbers do not represent number of patients. These indicate the number of medical cases for these COVID-19 patients as a single patient has a high chance of suffering from a multitude of diseases. This medical occurrence is prevalent in the case of older people or newborns. We then filter out all the records of diseases for these patients that have been diagnosed after 4 weeks of their COVID-19 diagnosis (Step 11). Again, these records will be the Post-COVID period. Finally, we have 257.65 thousand records for adult patients and 10.2 thousand records for pediatric patients (Step 12).

Just like earlier, such records are the occurrences of diseases for the Long COVID patients and not the patients themselves. At our mid-point (Step 13), we have the final number of Long COVID patients confirmed for adults - 24.865 thousand and pediatric - 1.1 thousand. This number is the confirmed cases from hospitals and healthcare units as we again use another ICD-10 code designated by the CDC. Long COVID has been assigned code U09 and all the Long COVID patients used in our experiments have been labeled to have U09 in their medical records at least once.

3.2.2 Phase 2: Network Analytics

We then take an SQL Join for Pre-COVID and Post-COVID diseases separately while using the 1,945 diseases we extracted earlier (Step 14). We explain the process of distinguishing the Pre-COVID and Post-COVID relationships in detail in 4. This step allows the creation of bipartite networks between patients and their Pre-COVID era diseases and Post-COVID era diseases, respectively (Step 16).

Once we create the respective bipartite graphs, we form projection graphs. Bipartite connections are Patient-Disease connections. Projections are Disease-Disease connections and remove any sensitive patient information. We create separate projections for both Pre-COVID and Post-COVID durations (Step 18). These projections can then be used as input for Gephi, which is a network-analytics-driven application (Step 19.) This application was used in creating Leiden and Louvain Algorithms. For our research purposes, we have used Louvain algorithms as our community clusters were comprehensive while identifying related diseases (Steps 20a and 20b). We explain these steps further in Chapter 4.

Afterwards, we sorted out the top 100 diseases from both Pre-COVID and Post-COVID sides of the Community Clusters. This is based on the disease nodes obtained in the Louvain Algorithm and the required nodes are identified with the help of modularity and weighted degrees (Step 21). Initially, we were handling a total of 1,945 diseases on either side. A recent article by Nguyen and Holmes [88] strongly suggests that dimensionality reduction yields better and more meaningful results in the case of medical and health records. It is common to have multiple features in a dataset to not add value to our analysis and, in fact, cause noise. This procedure is addressed in our reduction process (Step 22).

3.2.3 Phase 3: Prediction and LSTM

Once we have the data ready for the top diseases, we move on to create a dataset with patients as all our rows and columns represent all our features. For example, for adult patients, this dataset will have 2.01 million rows for all the patients and 1,945 columns for

each disease. We do this for both Pre-COVID and Post-COVID sides for the same diseases. On the Pre-COVID side, we mark the disease with a 0 or 1 against each patient to represent whether they were diagnosed with that disease in the past or not (Step 23a). On the Post-COVID side, we calculate a Jaccard coefficient for each disease based on the patient's disease history (Step 23b). We explain this process in further detail in Chapter 5.

Finally, we obtain a master dataset of 2.01 million rows and 3,890 columns. We created a similar dataset for Pediatric patients that is again 3,890 columns but has 340 thousand rows. Thereafter, we use a random sampling technique to implement our Machine Learning models. We have chosen Long Short-Term Memory Networks (LSTM) and Neural Networks for our analysis. Due to the extremely large datasets and highly dense string of characters in the dataset (especially on the Post-COVID side), analysis on the Palantir system was becoming difficult (Steps 24a, 24b, and 24c). Random Sampling allows us to take a subset of the total COVID-19 patients for better results and speed. Finally, after acquiring the needed subsets for our analysis, we are able to use LSTM and neural network models for our predictions. This sequence of data engineering leads us to dive further into our Network Analytics Process for our research for Long COVID in the next chapter.

3.3 Summary

Our goal in this chapter was to explain the fundamentals and context needed to understand medical health records in the Palantir system. Since this is backed by the NIH, it is difficult to use this data anywhere outside and most of the variables are sensitive and cannot be disclosed. Creating models and utilizing the given datasets in Palantir's distributed database. Furthermore, we elaborated on the workflow of the entire research for Long COVID starting from the Palantir Datasets, the network analytics and machine learning models along with the data engineering methods needed to implement models for prediction of Long COVID and the comorbidities.

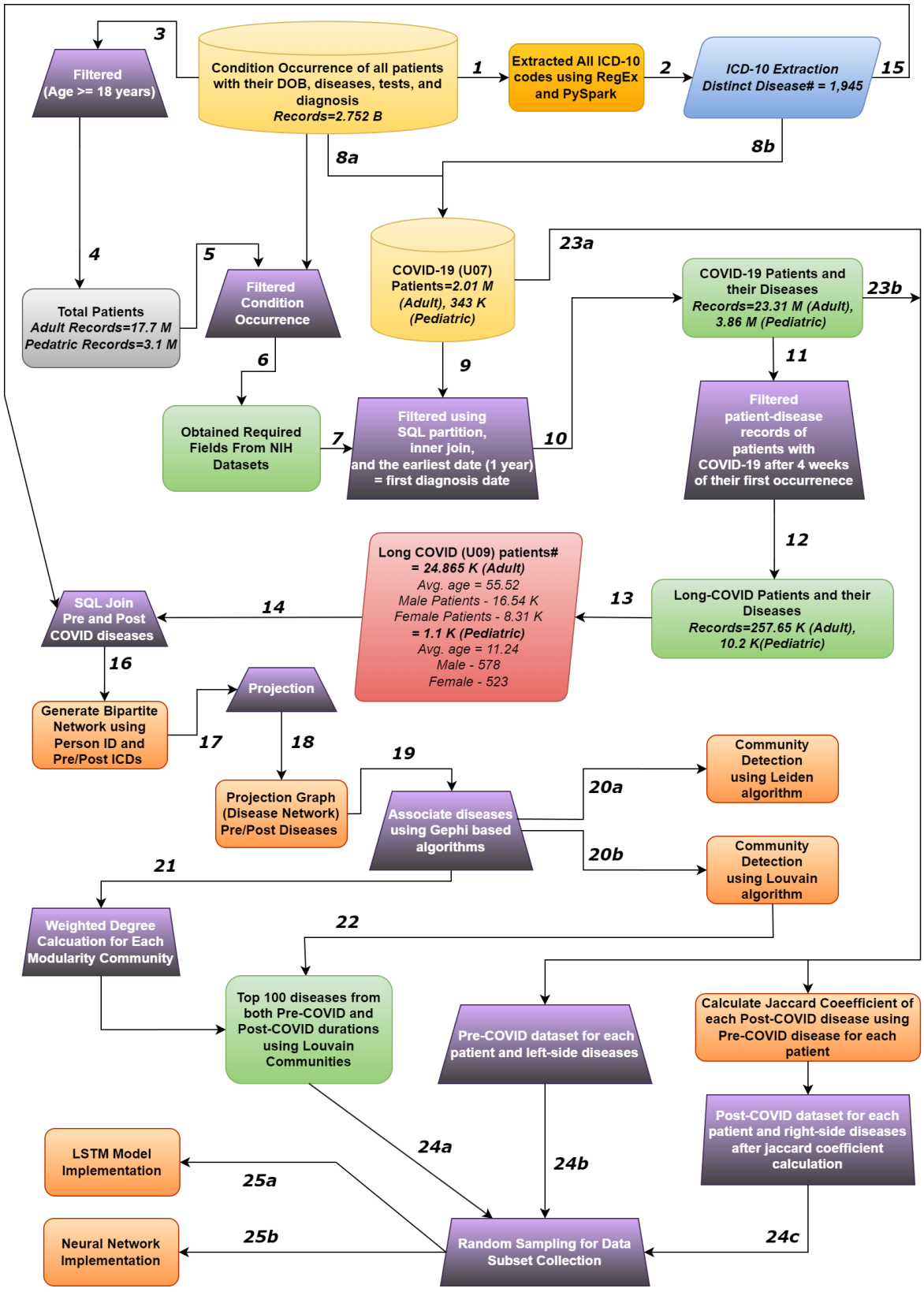


Figure 3.6: Workflow Diagram

Chapter 4

Network Analytics

This chapter deals with the relationships between Pre-COVID diseases and post-COVID diseases. One of the main aims of this paper is to find out how patients with a particular set of diseases can develop a new set of diseases in the Post-COVID phase. To achieve this objective, the relationship has been developed in two stages. In the first stage, we relate patients and diseases and in the second stage, we relate Pre-COVID diseases and Post-COVID diseases.

4.1 Bipartite Modeling

In this section, we discuss the concepts of why and how bipartite graphs have been implemented in our research. In subsection 4.1.1, we elaborate on the steps implemented while working on patient-disease connections till disease-disease projections. This also shows us the requirements when distinguishing between Pre-COVID and Post-COVID relationships of diseases and how bi-adjacency matrices can be used to find connections among the patients with their diseases. In subsection 4.1.2, we talk about the handling of Chronic diseases and how they can potentially create noise in our analysis.

4.1.1 Bipartite Disease Network

In the first stage, we represent the patient and disease relationship as an edge of a graph. Each patient can have one or more disease(s) and each disease can afflict one or more patient(s). This relationship is represented as connections in a patient-disease graph in which there is neither any connection among patients nor any connection among diseases. This provides us an opportunity to use the concept and properties of a bipartite graph. A

bipartite graph, also called a bigraph, is a set of graph vertices that can be decomposed into two disjoint sets such that no two graph vertices within the same set are adjacent. A bipartite graph can be projected to two unipartite graphs where each projection joins nodes of the same partition. The graph, so obtained, can provide one projection in which patients are connected to one another depending on their diseases and another projection in which diseases are connected to one another depending on the patients. For this paper, we focus only on the second projection to establish a disease-disease relationship.

In the second stage, we create subgraphs containing nodes with pre-COVID diseases and all other connected Post-COVID disease nodes including Long-COVID. But this subgraph is not a bipartite graph by default. In this paper, since we are exploring Pre-COVID to Post-COVID relationships, we are not concerned about the relationship between Pre-COVID diseases and Post-COVID diseases. We want to emphasize that this research is about the transformation of Pre-COVID to Post-COVID. As a result, we can easily put all the disease nodes on Pre-COVID and disease nodes on Post-COVID without any interconnection among themselves. Therefore, we remove all the edges in the network that connect Pre-COVID nodes and Post-COVID nodes excluding the edges that connected Pre-COVID to Post-COVID. This results in a network that has Pre-COVID diseases connected to Post-COVID diseases and vice-versa. Like the first stage, this enables us to apply the concept of a bipartite graph.

We have also divided the analysis considering chronic diseases and non-chronic diseases. For our analysis, we assumed that chronic diseases in Pre-COVID stage will persist in Post-COVID stage as well.

To explore the existence of disease-disease relationships from Pre-COVID to Post-COVID periods, these steps have been followed –

- On the basis of adult patients' condition occurrence data in the NIH database, a bipartite patient network has been realized. The projection graph follows this bipartite

network. The concept is depicted in Figure 4.1. Projection X corresponds to person-person network and projection Y corresponds to disease-disease network. The patient-disease network is shown in the center.

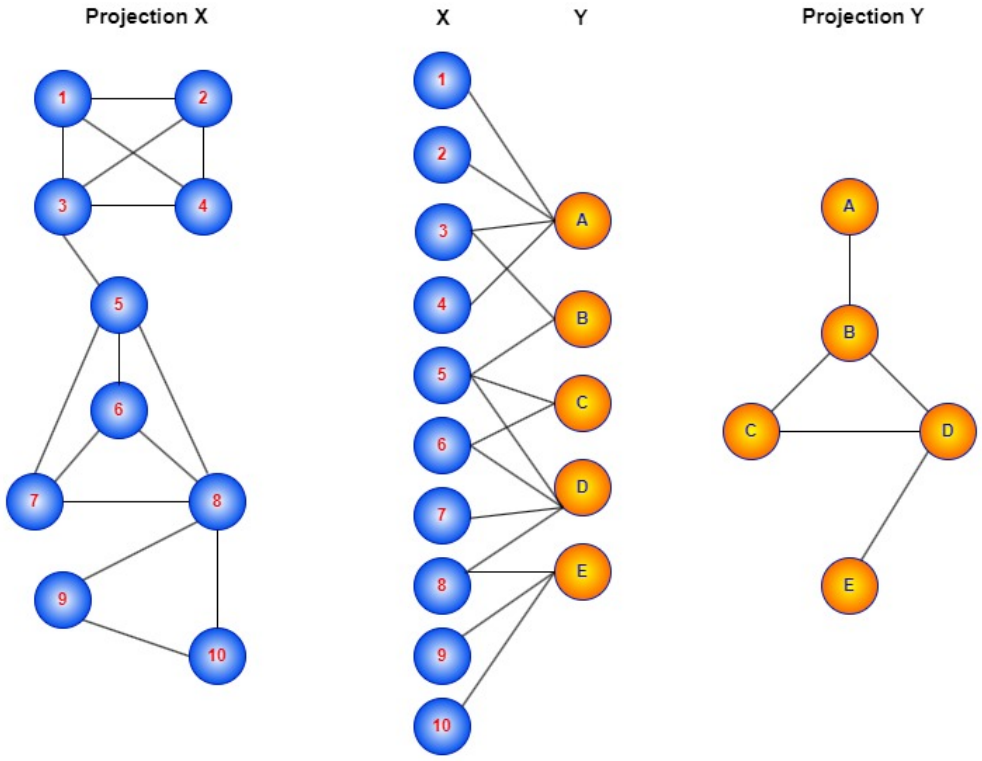


Figure 4.1: Bipartite Projection Network

- For this paper, we have focused only on the disease-disease network as explained earlier. These diseases include all diseases of the patients including Pre-COVID and Post-COVID diseases. The strategy to establish a relationship between Pre-COVID and Post-COVID is to create a bipartite network between Pre-COVID and Post-COVID diseases. Therefore, we manipulate the disease-disease network as shown in Figure 4.3.

- In order to create a Pre-COVID – Post-COVID bipartite network, we have identified the connections between the diseases that are part of Pre-COVID or Post-COVID. These are shown in green and orange colors, respectively. We have removed all these

internal connections to get a bipartite network between Pre-COVID and Post-COVID as shown in Figure 4.2(c).

Our aim is to find out how patients with a particular set of Pre-COVID diseases may develop into a new set of diseases in the Post-COVID phase. Several articles mention that patients with diseases before their COVID diagnosis have a high likelihood of developing new diseases in the Post-COVID phase [83, 112, 39]. As we can see, the patient-disease relationship can be represented in bi-adjacency matrices to identify the relationships between the patients and their corresponding diseases. Using the bi-adjacency matrices, we can create bipartite graphs that represent the connections between the patients and diseases [9].

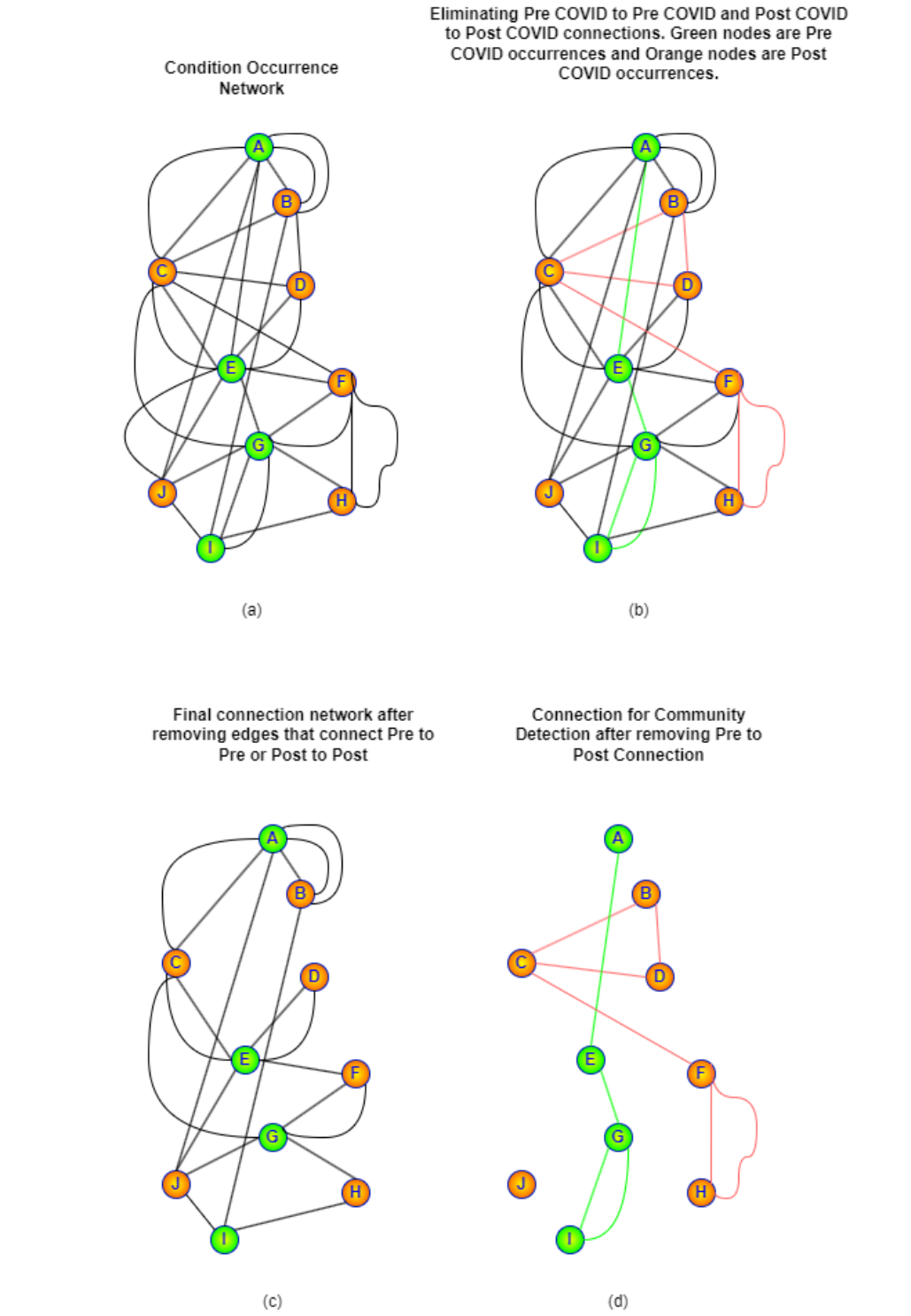


Figure 4.2: Pre to Post Bipartite Process

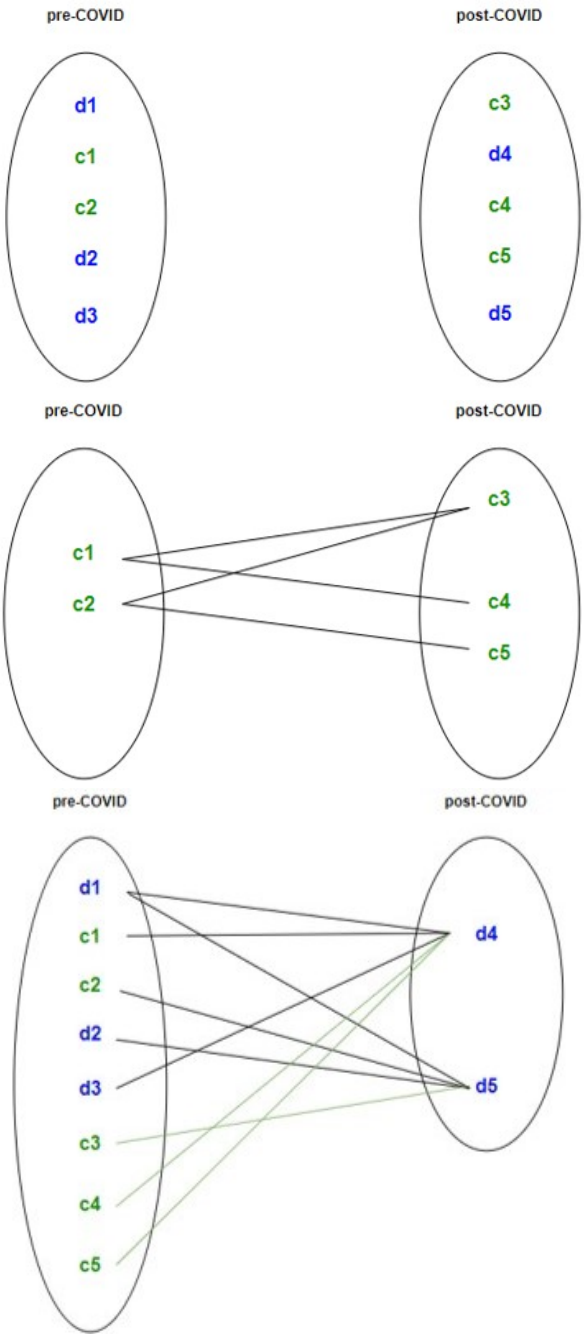


Figure 4.3: Bipartite Networking

4.1.2 Handling Chronic Diseases

For a meaningful exercise, we have divided all the diseases into two categories – chronic and non-chronic, which are represented by c and d as shown in Figure 4.3. Based on this way of categorizing, we are connecting the diseases from Pre-COVID (left) to Post-COVID (right) to identify the relationships between these diseases. If any chronic case is observed after COVID then it is very likely that that patient already had that chronic disease before their COVID diagnosis. In fact, COVID can increase the chances of certain diseases occurring in a patient but not necessarily cause such diseases [128, 35, 104]. Initially, we see chronic and non-chronic diseases to be present in both Pre-COVID and Post-COVID ends. Therefore, we have moved all chronic diseases to the Pre-COVID side of the network. By this method, we are also avoiding chronic diseases recorded on the Post-COVID side as we are considering such records to be already a part of the patients during their Pre-COVID phase.

We have considered 153 Chronic Diseases in our analysis [66]. Chronic diseases have played an important role in the management of COVID-19 cases. This network, therefore, represents all the cases such that all chronic disease records of patients will be found on the Pre-COVID side while non-chronic diseases will be recorded on both Pre-COVID and Post-COVID sides. For instance, if a patient is diabetic during Pre-COVID, that patient will remain diabetic even during the post-COVID phase. In such a situation, taking diabetes on both sides will be superfluous.

The bipartite network has an important feature in terms of the projection graph for the left side as well as the right side. Here, the left side is Pre-COVID (PC) and the right side is Post-COVID (LC). The connected graphs in the projection graph have a great significance because they show an association of a set of diseases on the Pre-COVID side or Post-COVID side and can be interpreted as a combination of diseases on the Pre-COVID side that lead to a set of diseases on the Post-COVID side.

In any network, communities are a property in which a subset of nodes of the graph is densely connected. In the case of social networking, community detection identifies a group

of individuals who interact with themselves much more than with other individuals. In this disease-disease network, such community detection technique helps us to identify how strongly diseases are related. In other words, if A and B belong to the same community then the existence of A implies a high probability of the existence of B. This technique is significant in applying unsupervised machine learning algorithms.

Louvain Algorithm is dependent on modularity, which is the quality of the sub-graphs among the communities [84]. Modularity is represented by Q in the following formula:

$$Q = \frac{1}{2m} \sum_{i,j} [A_{i,j} - (\frac{k_i k_j}{2m})] \delta(c_i, c_j)$$

1. Here, m is the number of links or connections and it represents the cases of diseases from our NIH datasets.

2. i, j are the nodes and they represent the diseases.

3. A_{ij} is the weight of the edge between nodes i and j and represents the relationship between the diseases.

4. k_i is the degree of the node (disease) i and k_j is the degree of the node (disease) j.

5. c_i is the community to which node i is assigned.

6. $\delta(c_i, c_j) = 1$, if $c_i = c_j$, otherwise 0.

4.2 Network Analytics Results

We show the results of our Network Analytics results in this section for both Adult patients and Pediatric patients. We discuss the results from our projections that lead to Community Detection from Gephi. We also show the prevalent diseases for each of these results in both Pre-COVID side and Post-COVID side and the methodology used for these results. Adult patient results are described in subsection 4.2.1 and pediatric patients results are described in subsection 4.2.2.

4.2.1 Network Analytics Results - Adults

Using the concept of network analysis and bipartite network as explained earlier, we created a patient-condition occurrence bipartite network and subsequently took a projection to create a condition occurrence (disease) sub-graph for community detection for this sub-graph, we identified the nodes that belong to Pre-COVID or Post-COVID and removed all the edges that connected any Pre-COVID condition occurrence to Post-COVID condition occurrence. This has been depicted in Figure 4.2(d) where Pre-COVID nodes and edges have been in green color and Post-COVID nodes and edges have been in orange color. Gephi was used to detect communities in Pre-COVID and Post-COVID projection graphs by using modularity with a resolution of 0.5. Louvain algorithm uses a modularity score, which helps in determining the connection between the nodes of the graph in the community. Each disease is represented by a node. The edges are the relationships between the diseases. That is, the higher the value of the modularity score, the more well-established will be the connection between the nodes within the communities. Since the data available in the NIH database on the Palantir platform is large and varied, this results in a huge network with hundreds of nodes and thousands of edges. To extract a meaningful interpretation from such a complex network graph, we used community detection tools. Community Detection tools in the present context represent disease-disease relationships - the larger the node, the greater the connections. Each such disease community has several nodes of varying sizes. For the purpose of description and better understanding, we have the community with the node of the largest size. For instance, Community 1 in Pre-COVID for Adult patients is represented by node I10, which as per ICD-10, is related to kidney/renal diseases. In table 4.2, under the column head diseases, we have mentioned the top 5 diseases for each disease community. These 5 diseases are based on the weighted degree of the nodes, nodes being the diseases. Similarly, we have listed 6 such prominent disease communities where each community represents connected diseases. The connection is based on network analytics and related methodologies applied to data in an appropriate manner. The entire pattern of

Patient Type	*Adult (Pre-COVID or Left)*	*Adult (Post-COVID or Right)*
Number of Diseases	1,555	1,569
Number of Communities	100	91
Percentage covered by the Top 6 Communities	28.5%(∼30%)	29.01%(∼30%)
Number of nodes covered by the Top 6 Communities	443	452
Avg. Weighted Degree	5,308,023,920.331	38,169,557,310.707

Table 4.1: Adult Patients Community Statistics

such a community network depicts an interesting pattern that can provide clues to medical practitioners and experts to probe further into Long COVID.

A stronger edge between the diseases would mean a higher frequency of that comorbidity. Comorbidities are diseases that occur in a patient at the same time [20]. Thus, the larger nodes have a greater weighted degree and connections. The top clusters cover approximately 30% of the Pre-COVID diseases and Post-COVID diseases among adults. We have summarized the statistics of the same in table 4.1. We also notice that all 1,945 diseases, as referred in 3.2.1, have not been covered here. In fact, we see 1,555 in the Pre-COVID side and 1,569 in the Post-COVID side. The slightly lower number of diseases is due to the fact that not all diseases occur among these COVID-19 adult patients. Another aspect of community detection is that we see the communities for COVID-19 (U07) and Long COVID (U09) in the Post-COVID side. Notably, we do not see COVID-19 or Long COVID in the Pre-COVID side as Pre-COVID was the period taken before COVID-19 diagnosis of a patient. We have explained earlier that Post-COVID is the period that is 4 weeks or more after the COVID-19 diagnosis. Therefore, U07 should not be present in these communities. But, hospitals and medical care centers have marked such patients as U07 even though it has been over 4 weeks since their initial COVID-19 diagnosis. The reason for U07 occurrence is due to the fact that Long COVID is too new for confirmed diagnosis as mentioned in 2.1. The diseases in the Pre-COVID side represent the medical history of the diseases for the patients.

On the basis of these 2 sets of disease communities, we can easily visualize the likely transition of Pre-COVID diseases into Post-COVID diseases. This visualization can be a great tool to understand the relationship between Pre-COVID and Post-COVID as well as the likelihood of Pre-COVID cases into Long COVID. In order to buttress the predictability part, we have used Jaccard similarity that can be translated into probabilities as explained in further detail in chapter 5. This information can be used by medical practitioners and experts for further analysis.

Based on the Louvain algorithm, we see that each community is thematically formed. These communities give a direction on what related diseases medical experts or practitioners can look out for during the diagnosis of COVID-19. For example, in Community 1, the diseases are tumor or infection-related. Furthermore, in Community 2 for U07 on Post-COVID side for adults, our other top diseases are immunity-related. Older patients or senior citizens are likely to suffer from such diseases [117, 10].

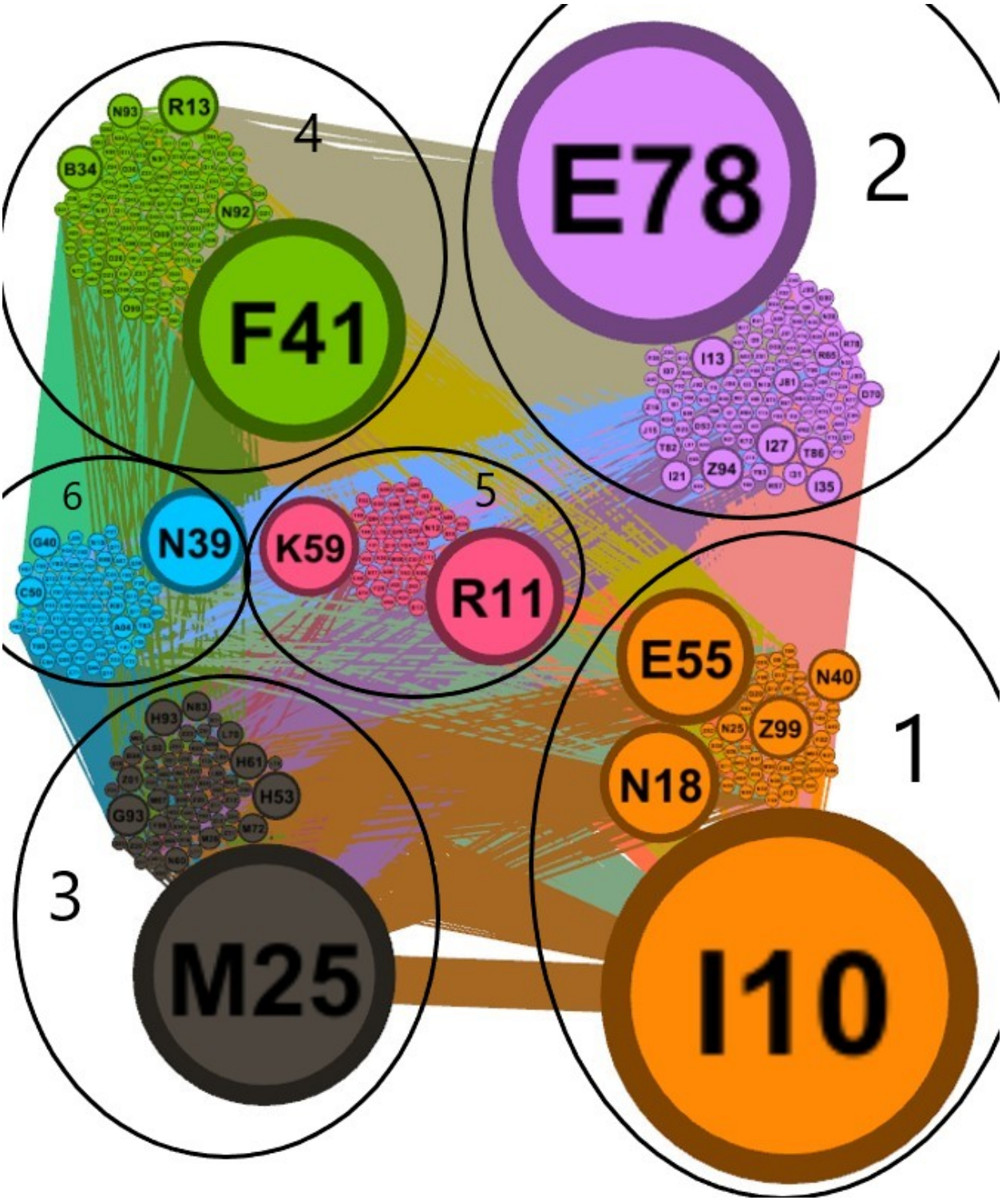


Figure 4.4: Pre-COVID Adult Community Detection

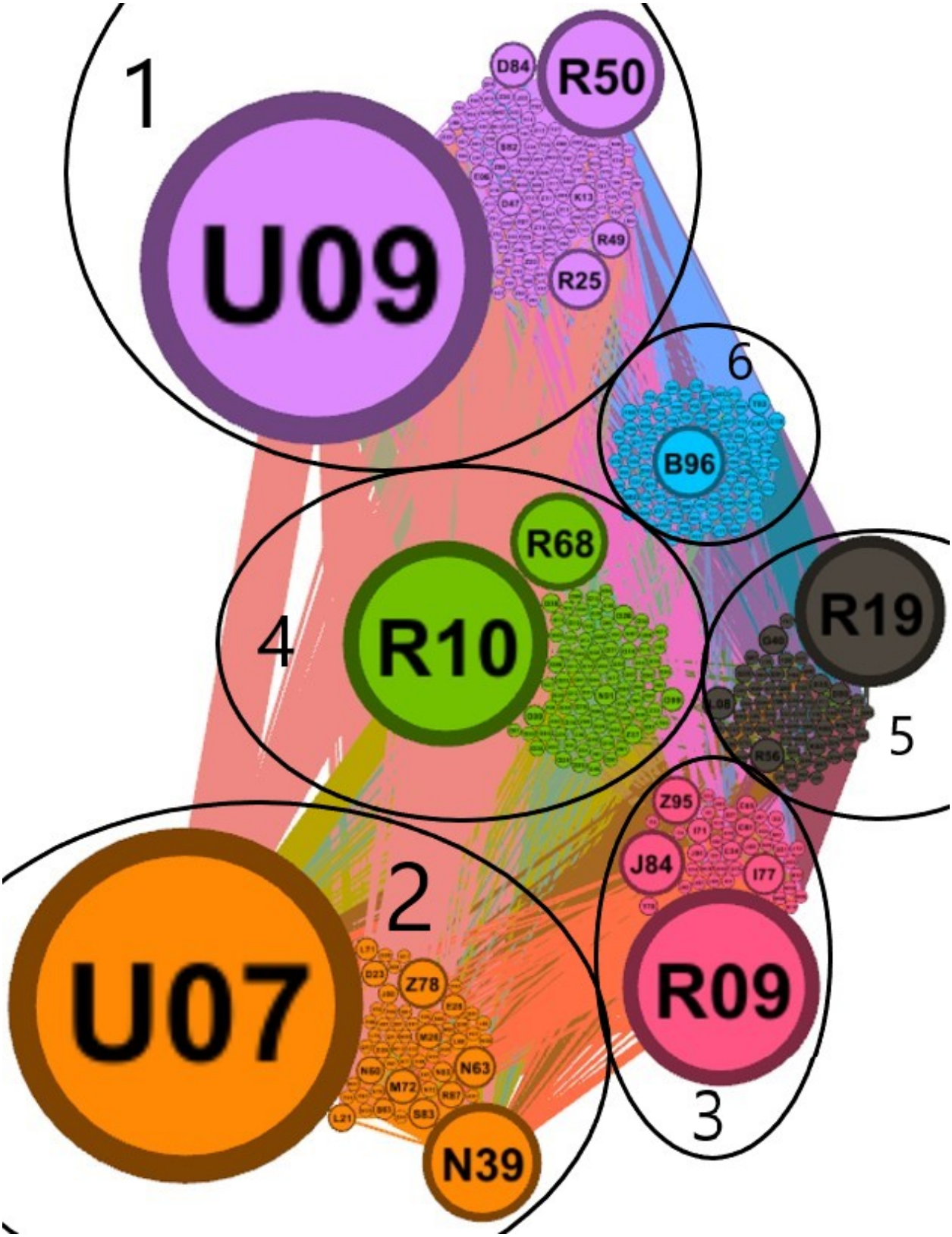


Figure 4.5: Post-COVID Adult Community Detection

Community Detected	Diseases (and ICD-10)
Kidney/Renal-related (1)	Hypertension (I10), Vitamin D Deficiency (E55), Chronic kidney disease (N18), Machine/Device Dependence (Z99), Benign prostatic hyperplasia (N40)
Cardio-related (2)	Lipoprotein Disorder (E78), Hypertensive heart disease (I13), Transplanted organ and tissue (Z94), Pulmonary heart diseases (I27), Nonrheumatic aortic valve disorders (I35)
Sensory-related (3)	Joint disorder (M25), Visual disturbances/impairment (H53), Disorders of the Brain (G93), Audio/Ear Disorders (H93), Disorders of the external ear (H61)
Pregnancy-related (4)	Anxiety disorders (F41), Aphagia and Dysphagia (R13), Viral infection (B34), Other Abnormal Uterine Disorders (N93), Childbirth/Pregnancy Complications (O99)
Digestion-related (5)	Nausea and vomiting (R11), Intestinal Disorders (K59), Tubulointerstitial nephritis (N12), Benign neoplasm and ill-defined parts of Digestive system (D13), Congenital malformations of intestine (Q43)
Infectious-related (6)	Infections of Urinary system (N39), Epileptic seizures (G40), Malignant neoplasm (C50), Enteropathogenic Escherichia coli infection (A04), Acute pyelonephritis (N10)

Table 4.2: Adult Patients Pre-COVID Prominent Disease Communities

Community Detected	Diseases (and ICD-10)
Immunity-related (1)	Long COVID (U09), Fever (R50), Abnormal involuntary movements (R25), Immunodeficiencies (D84), Hypernasality and Hyponasality (R49)
Neoplasm/Tissue-related (2)	COVID-19 (U07), Infections of Urinary system (N39), Specified Health Status (Z78), Lump in Breast (N63), Fibroblastic disorders (M72)
Cardio-related (3)	Circulatory and Respiratory system Disorders (R09), Interstitial Pulmonary diseases (J84), Cardiac and Vascular implants and grafts (Z95), Disorders of Arteries and Arterioles (I77), Aortic Aneurysm (I71)
Pregnancy-related (4)	Abdominal and pelvic pain (R10), Hypothermia (R68), Pregnancy/Birth-Complications (O99), Amenorrhea (N91), High Risk Pregnancy (O09)
Digestion-related (5)	Digestive System and Abdomen Disorders (R19), Convulsions (R56), Infections of skin and Subcutaneous tissue, Immunodeficiency with predominantly antibody defects (D80), Neoplasm of Endocrine Glands (D35), Crohn's Disease (K50)
Neoplasm-related (6)	Bacterial Diseases (B96), Malignant neoplasm of colon (C18), Malignant neoplasm of bladder (C67), Malignant neoplasm of Liver and intrahepatic bile ducts (C22), Malignant neoplasm of brain (C71)

Table 4.3: Adult Patients Post-COVID Prominent Disease Communities

4.2.2 Network Analytics Results - Pediatric

Community Detection using the Louvain methodology provided us with disease communities. As explained in the case of Adult disease communities 4.2.1, the prominent research communities are Fever, Respiratory Diseases, Abdominal Pain, Breathing Abnormalities, Throat/Chest Pain, and Cardiomyopathy. In the Post-COVID side, the most prominent

Patient Type	*Pediatric (Pre-COVID or Left)*	*Pediatric (Post-COVID or Right)*
Number of Diseases	887	1,081
Number of Communities	82	127
Percentage covered by the Top 6 Communities	39.23%(∼40%)	35.18%(∼35%)
Number of nodes covered by the Top 6 Communities	348	374
Avg. Weighted Degree	1,435,201.944	1,358,717.197

Table 4.4: Pediatric Patients Community Statistics

diseases are - Long COVID (U09), COVID-19 (U07), Fever, Cough, Abdominal and Pelvic Pain, and Nontraumatic Compartment Syndrome. In this case, we have also listed the top 5 diseases for each disease community. We see these diseases in 4.5 and 4.6. Since the number of pediatric patients is much fewer compared to adult patients, the absolute weighted degree of each node is substantially less but the relative relationship between Pre-COVID and Post-COVID disease communities presents an interesting pattern that medical experts can further probe. Since our research encompasses both adult and pediatric patients, we had the opportunity to compare the results of the two. In comparison, we find that in the case of Pediatric patients, diseases like Cough, Fever, Abdominal Pain, Neoplasms, Tumors, and infections are at the top. Moreover, based on common knowledge, these appear similar to real-life situations.

4.3 Summary

After looking into the role of bipartite graphs, we observe how disease relationships and comorbidities can be represented in projections and networks. These networks can help medical experts identify relevant diseases while diagnosing Long COVID among patients. Moreover, distinguishing between Pre-COVID and Post-COVID periods allows us to see the prominent diseases of each community, especially for Long COVID. We also learn how the newness of Long COVID has led to hospitals being unsure about labeling patients. ICD-10

Community Detected	Diseases (and ICD-10)
Respiratory-related (1)	Fever (R50), Lung Abnormalities (R91), Gastrostomy Abnormalities (Z93), Diseases of upper respiratory tract (J39), Streptococcus, Staphylococcus, and Enterococcus as the cause (B95)
Infectious-related (2)	Acute Respiratory Infection (J06), Unique Encounter Diagnosis (Z01), Birth-related (Z38), Atopic Dermatitis (L20), Acute bronchiolitis (J21)
Lower Abdominal-related (3)	Abdominal Pain (R10), Dizziness (R42), Pain in Micturition (R30), Polyuria (R35), Mittelschmerz (N94)
Pregnancy-related (4)	Breathing Abnormalities (R06), Lipoprotein Disorder (E78), Pregnancy-related (N94), Fetal Problems (O36), Postprocedural Complications (T81)
Tumor-related (5)	Throat/Chest Pain (R07), Elevated Blood Glucose Level (R73), Malignant Neoplasm (C79), Embolism and Thrombosis (I82), Sleep disorders (F51)
Cardio-related (6)	Cardiomyopathy (I42), Osteoarthritis (M19), Atrial fibrillation and flutter (I48), Chronic Obstructive Pulmonary Disease (J44), Malignant neoplasm of bronchus and lung (C34)

Table 4.5: Pediatric Patients Pre-COVID Prominent Disease Communities

for U09 has been identified this year, 2023, by the CDC, and medical care units are still working on incorporating such a disease into their systems.

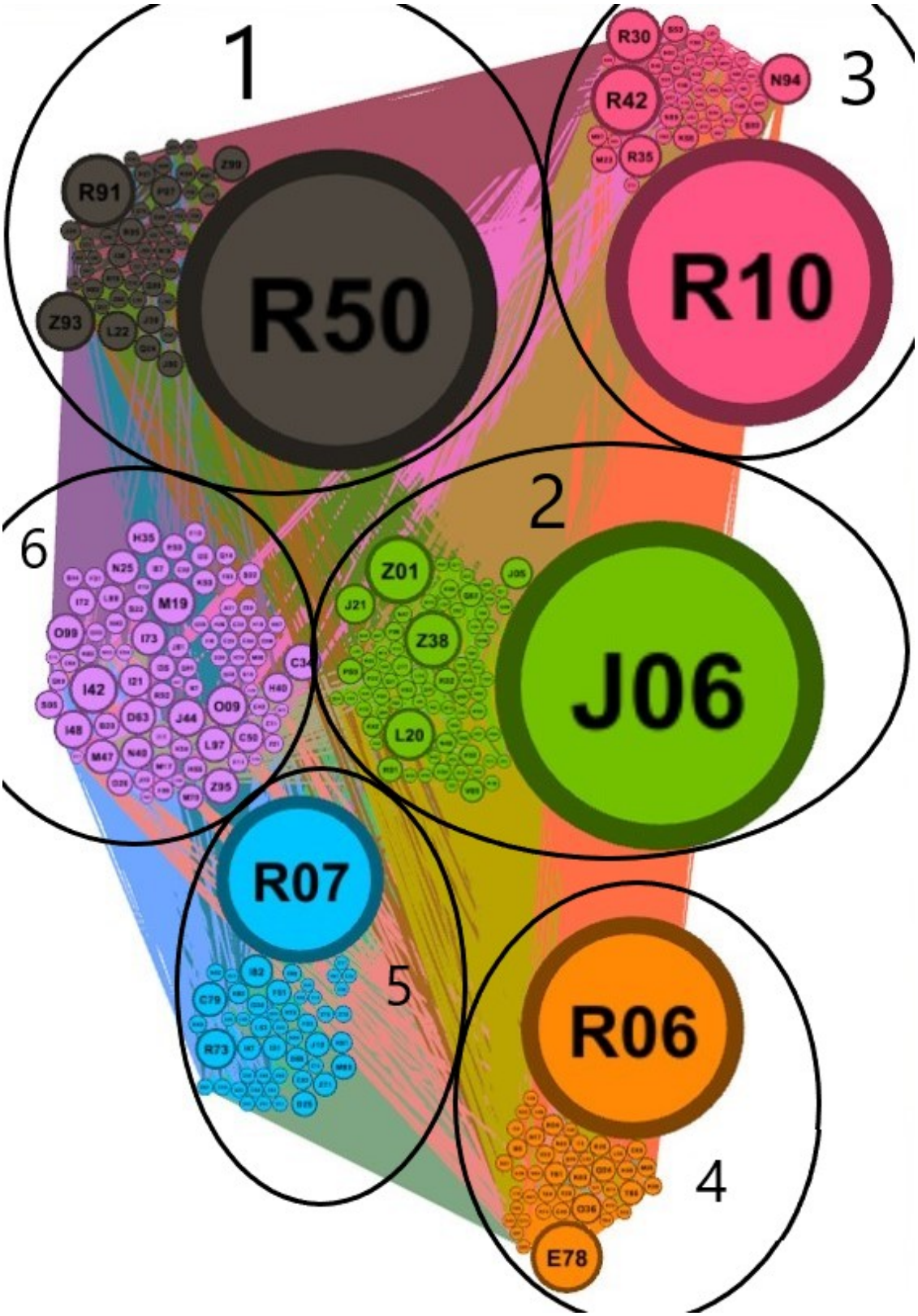


Figure 4.6: Pre-COVID Pediatric Prominent Disease Community Detection

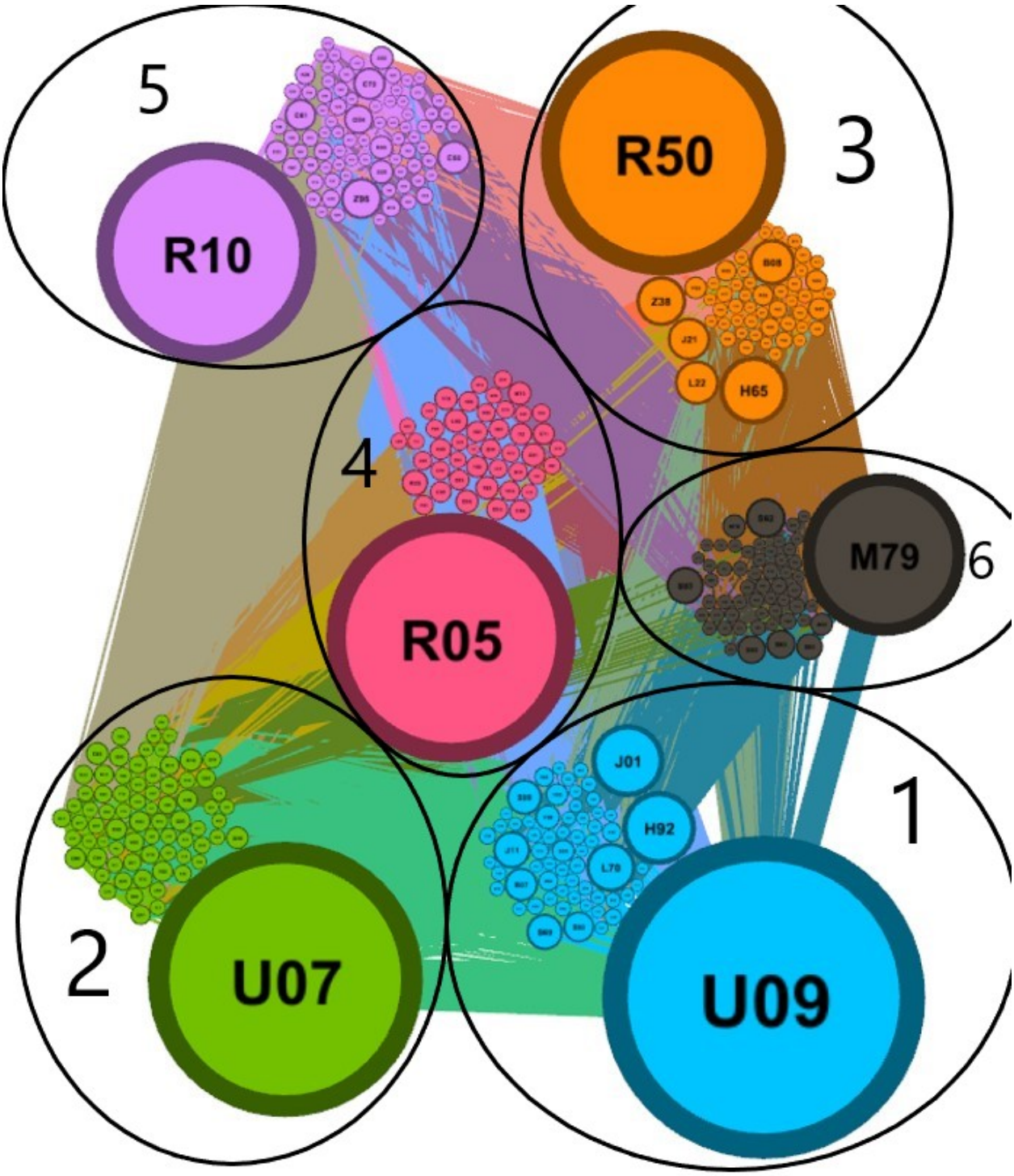


Figure 4.7: Post-COVID Pediatric Prominent Disease Community Detection

Community Detected	Diseases (and ICD-10)
ENT-related (1)	Long COVID (U09), Ear Disorders (H92), Acute sinusitis (J01), Acne (L70), Unidentified Influenza (J11)
Viral-related (2)	COVID-19 (U07), Viral Hepatitis (B18), Lymphoma (C85), Congenital malformations of great veins (Q26), Neoplasms/Tumors (D49)
Infectious-related (3)	Fever (R50), Nonsuppurative otitis media (H65), Birth-related (Z38), Skin infections/Membrane Lesions (B08), Irritant Dermatitis (L22)
Tissue/STD-related (4)	Cough (R05), HIV (B20), Asymptotic HIV (Z21), Soft tissue disorders (M70), Fetal abnormality and damage (O35)
Neoplasm/Tumor-related (5)	Abdominal and pelvic pain (R10), Malignant neoplasm (C50), Secondary malignant neoplasm of kidney and renal pelvis (C79), Malignant neoplasm of prostate (C61), Leiomyoma (D25)
Bone/Muscular-related (6)	Nontraumatic compartment syndrome (M79), Fracture Hand/Wrist (S62), Dislocation and Sprain of Joints (S93), Dislocation and sprain of joints and ligaments (S93), Ankle/foot injury (S90)

Table 4.6: Pediatric Patients Post-COVID Prominent Disease Communities

Chapter 5

LSTM and Neural Network with Network Analytics

This chapter is dedicated to the implementation of Machine Learning algorithms utilizing the information obtained through network analytics. Since the objective is to predict Long COVID and the number of such patients is abysmally low, any machine learning methodology in spite of the application of methods like SMOTE does not perform well. Like informed search, we apply the knowledge obtained through network analytics to select the features wisely and scientifically reducing the computational burden and generating excellent results.

More specifically, in this chapter, we deal with the results of our Network Analytics approach used in Long Short-Term Memory Networks and Neural Networks. Our community detection results reveal the most prevalent diseases for COVID-19 and Long COVID patients. Once we have gathered this information regarding the top diseases, we have been able to use this for our Long COVID prediction for potential COVID-19 patients. To achieve the prediction objective, first of all, we create a master dataset consisting of all patients as rows and all the diseases as columns/features. Except for a few columns, the data for most of the columns are categorical in nature – the majority of times binary. In the case of adult patients, the number of rows is about 2.01 million distinct patients and, in the case of pediatric patients, about 343,000. The Long COVID patients for adults and pediatrics are 24,864 and 1,101, respectively. As mentioned in 3.2, we have taken only those cases where U09 has been confirmed by medical health care units in the NIH records. The fact that such cases have been already established by medical experts and doctors would validate our dataset for Long COVID patients, who are labeled with ICD-10's U09 code.

5.1 Creation of Master Dataset before Modeling

The first goal in this part of the study is to distinguish the Pre-COVID side and the Post-COVID side of diseases for each patient. We have marked all the diseases in the Pre-COVID side that were diagnosed before the detection of COVID-19. Post-COVID diseases are diseases that were diagnosed 4 weeks after the original diagnosis of COVID-19. This is as explained in Section 3.1.3. For each patient, we simply mark these patients 1 or 0 depending on whether they have a particular disease during Pre-COVID or not, respectively. This process is done for all 1,945 diseases for the Pre-COVID side. For the Post-COVID side, we take an approach different from marking 1 or 0 for each disease rather it is based on Jaccard coefficients. Jaccard coefficient's role in this procedure is important. In general, Jaccard coefficient, measuring similarity between finite sample sets, is defined as the size of the intersection divided by the size of the union of the sample sets as below:

$$\frac{A \cap B}{|A| + |B| - |A \cap B|}.$$

In our case, when we form a bipartite network of Pre-COVID and Post-COVID diseases, we have connections between Pre-COVID and Post-COVID diseases represented by a number of edges between the nodes. If let us say, we want to compute the Jaccard coefficient between Pre-COVID disease x and Post-COVID disease y then the Jaccard coefficient is computed as below:

$$\frac{|\text{edges connecting x and y}|}{|\text{edges of node x}| + |\text{edges of node y}| - |\text{edges connecting x and y}|}.$$

In this manner, we calculate a matrix of Jaccard coefficients. If the number of disease columns in Pre-COVID and Post-COVID is 1,945, then this Jaccard coeffcient matrix will be size of 1,945 x 1,945.

For the Post-COVID side, the dataset generation is implemented in the following steps:

1. We will again have 1,945 columns on the Post-COVID side for each disease. We start with the first patient and multiply the row with Pre-COVID diseases with categorical

values [0, 1] for each column with the Jaccard coefficient matrix to compute column values for A01, A02, .. etc. of the Post-COVID side.

2. Although the column names are the same as those of Pre-COVID columns, the columns in Post-COVID are real numbers between 0 and 1. These numbers are based on Jaccard Coefficients as explained in the preceding bullet point. Essentially, Jaccard Coefficient results in a value that is the similarity of two given nodes or vertices. In our case, the two nodes are the given diseases from Pre-COVID and Post-COVID.

3. We repeat the same step as above for the next patient onwards and compute Post-COVID disease column values that are real numbers between 0 and 1.

4. Since we multiply a row vector of size (1, n) with a matrix of Jaccard coefficients of size (n, n), we multiply Pre-COVID disease values with each column values of the matrix and sum them up to get to compute each element of Post-COVID disease column, resulting in another row vector of size (1, n).

5. The above process is done for all adult and pediatric COVID-19 patients separately because Jaccard coefficient matrices are different.

5.2 Inclusion of Top Diseases

Recall that (see also Chapter 4) we are able to acquire information on the most prevalent diseases for Pre-COVID and Post-COVID for both Adult and Pediatric patients. Our objective is to develop a model that helps medical experts and practitioners in the prediction of Long COVID (U09). If all the columns are utilized, the machine learning models will be computationally inefficient, and despite immense computation, the results are worthless. We realized that large number of diseases on both Pre-COVID and Post-COVID sides create noise obstructing the machine learning models to learn properly. In such a situation, dimensionality reduction is found to be essential to attain good results. As we know, the

dimensionality reduction helps in the simplification of the dataset as it is not uncommon to handle a large number of features in medical health records. Such a method can be treated as latent feature extraction for the removal of redundant data [88]. Therefore, we take the top 100 diseases of each side of Pre-COVID and Post-COVID on the basis of network analysis carried out in the earlier section.

After dimensionality reduction, we use a Long Short-Term Memory (LSTM) network and a common Neural Network for our prediction. Since the LSTM model is good for temporal data and in our case, Post-COVID data temporally follows Pre-COVID data, we reshape both Pre-COVID and Post-COVID data to form an array of size (2, n) and use that as an input to the model. For the LSTM model, we require both sides of disease in Pre-COVID and Post-COVID to be of equal size. Necessitated balancing of disease columns on both sides is to ensure not only the same number of diseases on both sides but also the same disease columns in the same order. For example, for every patient we have 3,890 columns for diseases. For LSTM, we reshape this row into a (2, 1945) array. If, after dimensionality reduction, the column size is reduced to 100 then this 2-D array size is reduced to (2, 100) and the same is processed for each patient and is used as an input for the LSTM model. Another aspect to note is that the top row of this 2-D array are the diseases in Pre-COVID and the bottom row are the diseases in Post-COVID. For the neural network model, we do not carry out this reshaping phase. We use both Pre-COVID and Post-COVID disease columns as separate features.

The next issue that we confront is the imbalanced dataset. During our experiments, long COVID cases are always a lot fewer than the number of patients with COVID-19. This is easy to realize because Long-COVID is rare. Moreover, we consider the U09 cases that have been confirmed by medical healthcare units and hospitals and, likewise, U07 for COVID-19-labeled patients. To handle such an imbalanced dataset, we advocate for the implementation of random sampling to deal with large data with smaller samples – thus reducing the size of the majority to bring in some semblance of balance. For example, we have 1,101 confirmed

U09 pediatric patients. Using this small sample against an enormous sample of 340 thousand COVID-19 pediatric patients will result in very poor results due to high imbalance. Instead, we ran our analysis against 10 thousand COVID-19 pediatric patients. In order to have a robust and reliable result, we apply the random sampling of COVID-19 patients, and we repeat the experiments multiple times to take random samples of 10 thousand patients from a population of 340 thousand patients. This step ensures utilizing as much of the data as possible while making sure to obtain good, stable, and reliable prediction results from the models.

5.3 Long COVID Prediction of Adults

During our analysis, we use LSTM and Neural Networks on the adult population. We use the results from our network analytics community detection model as an input for our machine learning models. With our network analytics results, we were able to learn what relevant diseases a medical expert should look out for while diagnosing Long COVID. The use of synergy we create here with the machine learning models and the network analytics results helps us in the prediction of Long COVID.

5.3.1 Long COVID Prediction of Adults - LSTM

Our results for adult patients in the case of LSTM are experimented on 500,000 patients for most of the experiments. The accuracy measures in almost all the results are around 90% even with different parameters. We see a substantial difference in results when we raise or lower the number of COVID-19 patients. For example, in our analysis of 500,000 COVID-19 patients, our F1 score is around 88%. This score would shoot up to approximately 94% when taking 400,000 COVID-19 patients. We obtain approximately 100% accuracy if we use 250,000 COVID patients.

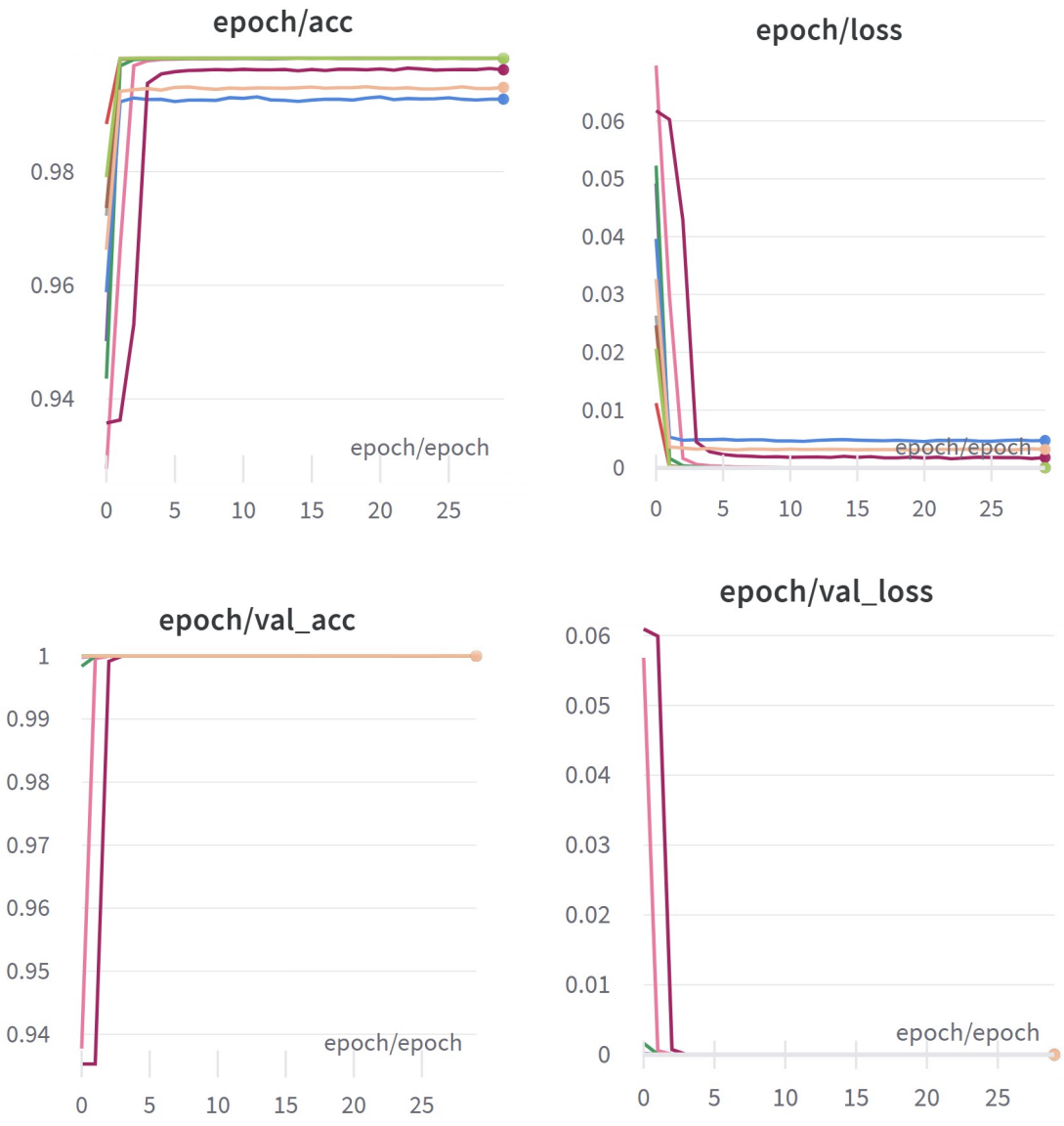


Figure 5.1: LSTM for Adults with Hyperparameter Tuning

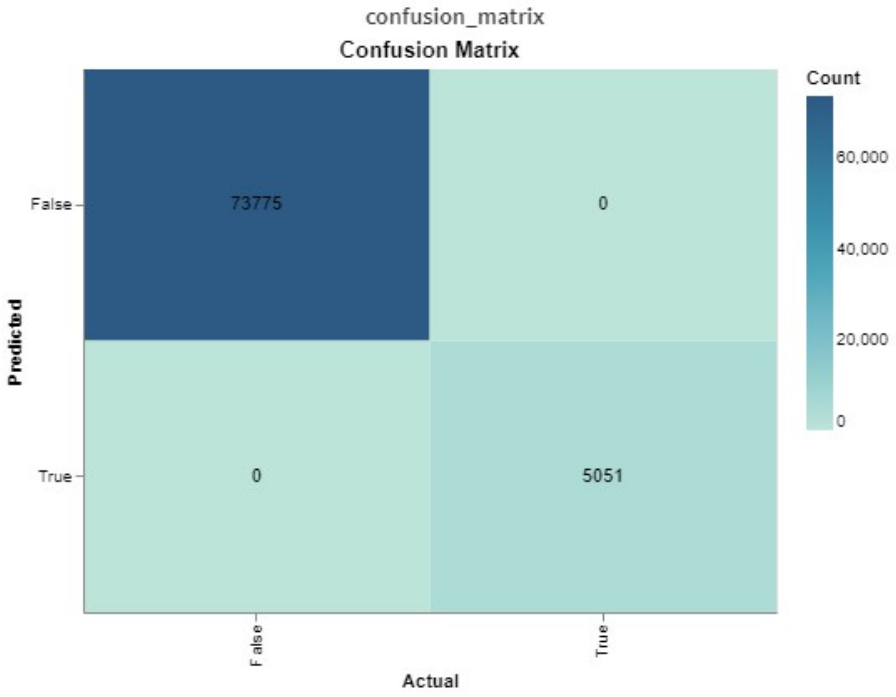


Figure 5.2: LSTM for Adults Confusion Matrix

#	**Activation Function**	**Batch Size**	**Learning Rate**	**Rows**	**Columns**	***Accuracy***	***F1 Score***	***Loss***
1	tanh	256	5e-4	500,000	100	*93.22*	*87.46*	*3.09e-5*
2	tanh	256	5e-5	500,000	100	*94.89*	*87.85*	*3.22e-5*
3	tanh	256	5e-3	500,000	100	*94.77*	*87.53*	*3.54e-4*
4	tanh	256	5e-2	500,000	100	*10.02*	*0*	*0.0629*
5	tanh	512	5e-4	500,000	100	*90.97*	*88.27*	*2.25e-5*
6	tanh	128	5e-5	500,000	100	*90.89*	*86.85*	*3.62e-5*
7	tanh	128	5e-3	500,000	100	*90.89*	*88.19*	*2.09e-4*
8	tanh	512	5e-4	500,000	100	*90.99*	*88.19*	*2.09e-4*
9	tanh	1024	5e-4	500,000	100	*89.99*	*87.55*	*0.0020*
10	tanh	1024	5e-3	500,000	100	*89.79*	*87.95*	*0.0017*
11	sigmoid	512	5e-3	500,00	100	*89.34*	*88.22*	*0.0015*
12	relu	512	5e-3	500,000	100	*84.79*	*86.19*	*0.0023*
13	tanh	256	5e-4	500,000	200	*82.22*	*87.46*	*3.09e-5*
14	tanh	256	5e-3	500,000	50	*97.65*	*94.48*	*0.0160*
15	tanh	256	5e-4	400,000	100	*94.97*	*93.96*	*0.0035*
16	tanh	1024	5e-4	600,000	100	*95.98*	*82.52*	*0.0032*

Table 5.1: Adult LSTM Results

5.3.2 Long COVID Prediction of Adults - NN

In this section, we evaluate the Long-COVID prediction results of the models for adults. With hyperparameter tuning, we use 'tanh' activation function with batch size = 256. Our learning rate is 0.0005. The number of columns and rows set are 100 and 4000,000 or 500,000, respectively. These parameters lead us to an accuracy of around 90% accompanied by F1 score in the range [82.34%, 92.57%] in our testing over 30 epochs. The results for the same are plotted in Figure 5.3. In all of the testing cases, accuracy rate is around 90%.

Raising the number of rows to 600,000 lowered some scores with the same hyperparameters but accuracy stays at around 90%. We end up with an F1 score of 82.34%. Raising the number of columns of diseases to 200 improved F1 score slightly and it did not make as much of an impact as raising the number of rows. In fact, our results yield around 90% accuracy and 87.95% F1 score. Due to the high imbalance of the data, the purpose is to nail down reasonably acceptable configurations of parameters, rows, columns, and variables that can give the best results for Long COVID prediction. We experimented with SMOTE, but our results do not enthuse us: the results deteriorated slightly when compared to our Non-SMOTE run. The results of experiments further implemented are tabulated in Table 5.2.

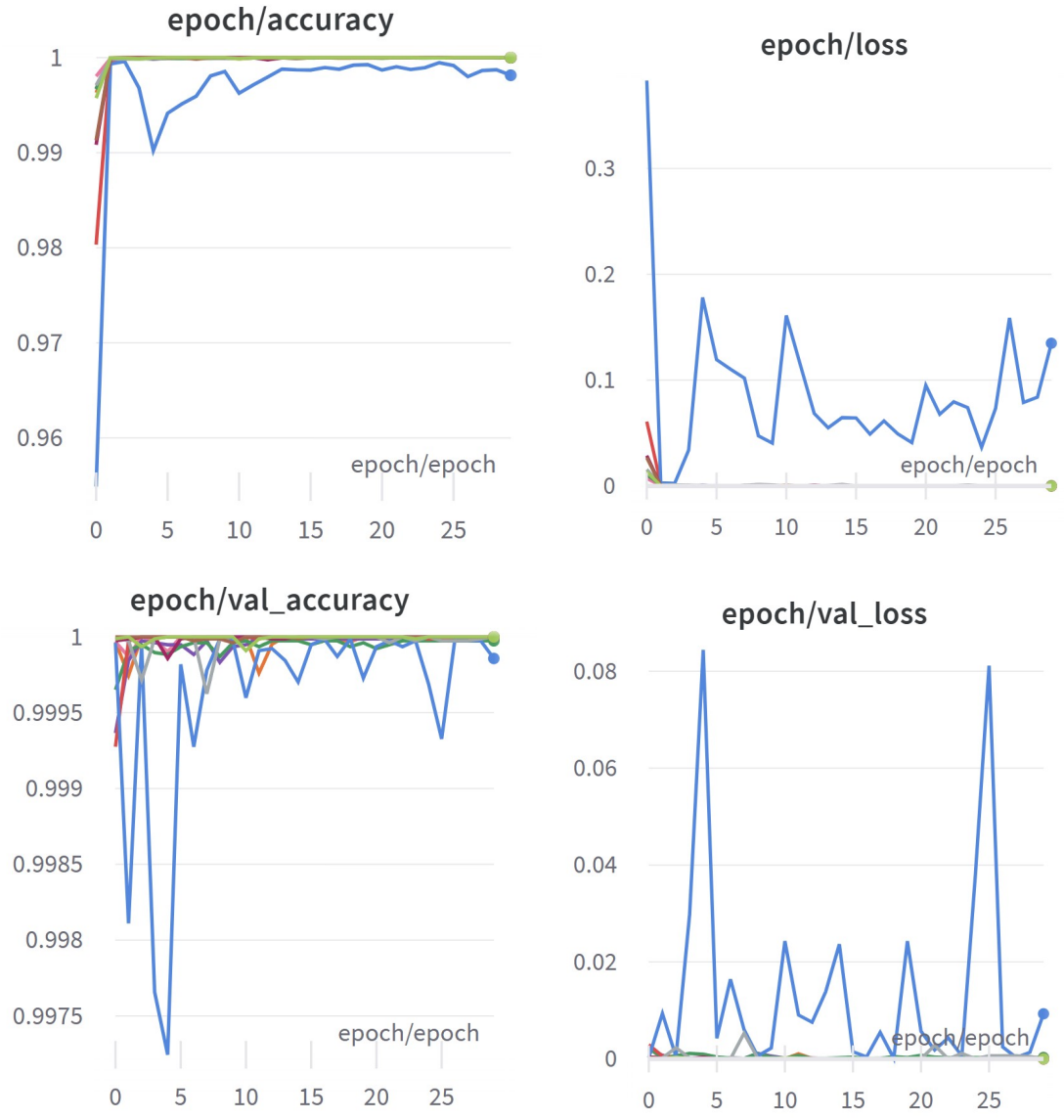


Figure 5.3: Neural Network for Adults with Hyperparameter Tuning

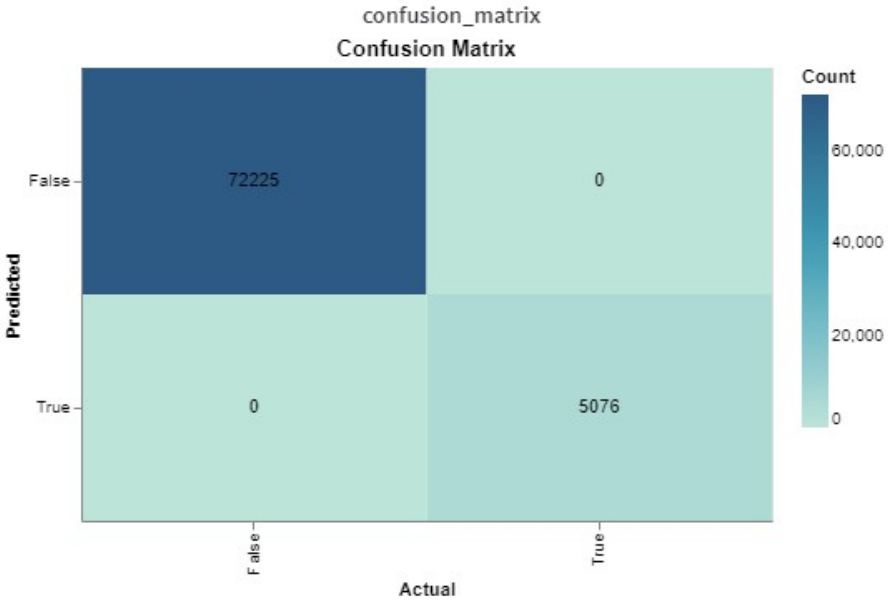


Figure 5.4: Neural Network for Adults Confusion Matrix

#	**Activation Function**	**Batch Size**	**Learning Rate**	**Rows**	**Columns**	***Accuracy***	***F1 Score***	***Loss***
1	tanh	256	5e-5	500,000	100	*89.97*	*88.78*	*0.0099*
2	tanh	256	5e-2	500,000	100	*89.89*	*88.53*	*0.0093*
3	tanh	256	5e-3	500,000	100	*89.98*	*88.65*	*1.6e-33*
4	tanh	128	5e-4	500,000	100	*92.65*	*87.30*	*4.31e-9*
5	tanh	1024	5e-4	500,000	100	*95.65*	*87.88*	*4.22e-9*
6	tanh	512	5e-5	500,000	100	*93.68*	*87.52*	*4.63e-9*
7	relu	256	5e-4	500,000	100	*95.33*	*87.55*	*4.32e-9*
8	sigmoid	256	5e-4	500,000	100	*93.22*	*87.44*	*4.30e-9*
9	tanh	256	5e-4	500,000	50	*92.68*	*87.55*	*4.33e-9*
10	tanh	256	5e-4	500,000	200	*95.88*	*87.95*	*4.30e-9*
11	tanh	256	5e-4	400,000	100	*92.88*	*92.57*	*9.93-6*
12	tanh	256	5e-4	600,000	100	*91.98*	*82.34*	*9.94e-6*
13	tanh	256	5e-4	500,000	500	*92.88*	*87.56*	*6.07e-5*
14	tanh	256	5e-4	500,000	1000	*97.88*	*87.54*	*6.07e-5*

Table 5.2: Adult Neural Network Results

5.4 Long COVID Prediction of Pediatrics

Just like in the case of adult patients, we use LSTM and Neural Networks for the pediatric population. The results for adult patients show us how well the models are trained to result in stable prediction for Long COVID. With our network analytics results, we also see the differences in diseases among the adult and pediatric populations. Moreover, we deal with an even smaller number of Long COVID patients for pediatrics. Our results vary more compared to adult patients as we see in the following subsections related to LSTM and Neural Network for pediatric patients.

5.4.1 Long COVID Prediction of Pediatrics - LSTM

Now we discuss the results of the models that have been used with the help of our findings from network analytics. With hyperparameter tuning, we use 'tanh' activation function with batch size = 256. Our learning rate is 0.0005. The numbers of columns and rows set are 100 and 10,000, respectively. These parameters give rise to an accuracy of 99.48%, precision of 92.12%, recall of 91.97%, and F1 score of 91.58% in our testing over 30 epochs. The findings for the same are displayed in Figure 5.5. The accuracy of all the tests is over 95%.

Changing the number of rows to 20,000 lowered our overall results with the same hyperparameters. We ended up with an accuracy of 99.79%, precision of 61.11%, recall of 60.90%, and F1 score of 60.85%. Raising the number of columns of diseases to 200 lowered our results but it did not have as much of an impact as raising the number of rows. In fact, our results yield an impressive 90.71% accuracy, 89.09% precision, 86.91% recall, and 87.76% F1 score. Due to the high imbalance of the data, our goal is to find reasonably acceptable configurations of parameters, rows, columns, and variables that can yield the best Long-COVID predictions. While using SMOTE, the results are slightly worsened compared to our Non-SMOTE counterpart. The results of experiments further implemented are shown in Table 5.3.

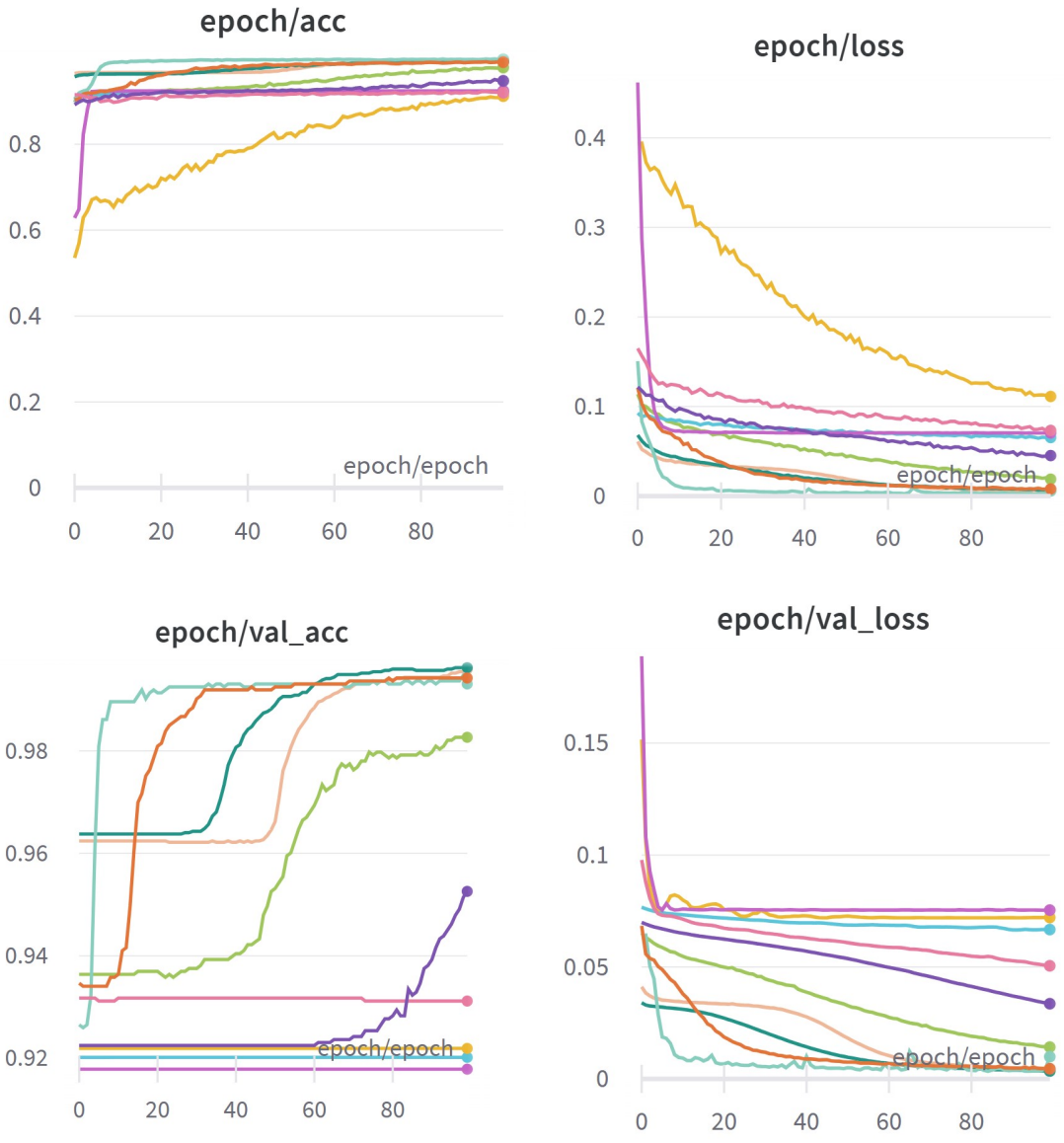


Figure 5.5: LSTM for Pediatrics with Hyperparameter Tuning

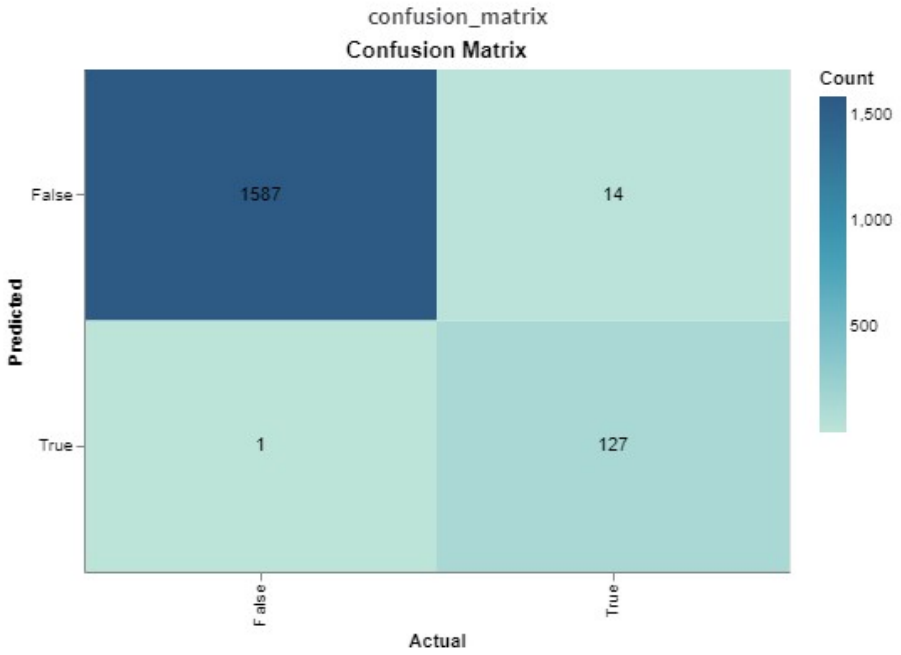


Figure 5.6: LSTM for Pediatrics Confusion Matrix

#	**Activation Function**	**Batch Size**	**Learning Rate**	**Rows**	**Columns**	**SMOTE**	***Accuracy***	***F1 Score***	***Loss***
1	tanh	256	5e-5	10,000	100	No	*90.23*	*89.41*	*0.0078*
2	tanh	256	5e-6	10,000	100	No	*9.04*	*0*	*0.0709*
3	tanh	256	5e-4	10,000	100	No	*89.94*	*88.48*	*0.0005*
4	tanh	256	5e-3	10,000	100	No	*80.71*	*78.1*	*0.0035*
5	tanh	128	5e-5	10,000	100	No	*5.55*	*0*	*0.0656*
6	tanh	512	5e-4	10,000	100	No	*4.11*	*0*	*0.0656*
7	tanh	512	5e-4	10,000	100	Yes	*85.45*	*84.84*	*0.0040*
8	tanh	1024	5e-4	10,000	100	No	*89.99*	*84.32*	*0.0057*
9	sigmoid	1024	5e-3	10,000	100	No	*4.22*	*0*	*0.0667*
10	sigmoid	1024	5e-5	10,000	100	No	*4.81*	*0*	*0.0763*
11	relu	1024	5e-5	10,000	100	No	*5.88*	*0*	*0.0602*
12	tanh	256	5e-5	20,000	100	No	*70.17*	*61.48*	*0.0041*
13	tanh	256	5e-5	20,000	100	Yes	*62.51*	*58.80*	*0.0056*
14	tanh	256	5e-5	10,000	200	No	*80.29*	*76.69*	*0.0178*
15	tanh	256	5e-5	10,000	200	Yes	*82.11*	*79.97*	*0.0147*

Table 5.3: Pediatrics LSTM Results

5.4.2 Long COVID Prediction of Pediatrics - NN

Similar to the LSTM methodology, we adopt neural networks for prediction and comparison of our results. Likewise, we implement this with 'tanh' activation function, 10 thousand records of COVID-19 patients and 1,101 Long COVID patients. Along with that, we have the top 100 columns of diseases from the network analytics results. These given parameters also help us make a comparative analysis between LSTM and Neural Networks as they are similar models. Table 5.4 summarizes the prediction results. Our accuracy has mostly been above 90% in all cases. Our best F1 scores are runs 4, 9, and 14 where all of them exceed 90%. The lowest loss value was in run 1 but with a lower F1 score of 77.77%. The learning rate slightly changes the results with differences in 2% to 3%. We see a significant change of 10% increase in F1 score results when changing batch size from 256 to 512 while keeping a learning rate of 0.005.

In cases where we increase the number of rows to above 10,000, the F1 score drops sharply. At 20,000 rows, we obtain an F1 score of around 60.5%. At 30,000 rows, we see an F1 score of 50.21%. We obtain similar outcomes when the number of columns increases from 100 to 200 where we get an F1 score of 82.46% for the latter. Increasing columns has more of an impact on accuracy than increasing the number of rows. In the case of activation functions, 'tanh' leads to the best results for Neural Networks for Pediatric patients. 'Sigmoid' and 'Relu' activation functions lead to 0% in both precision and recall, which result in a 0% for the F1 score. Even loss values are more desirable while using 'tanh'.

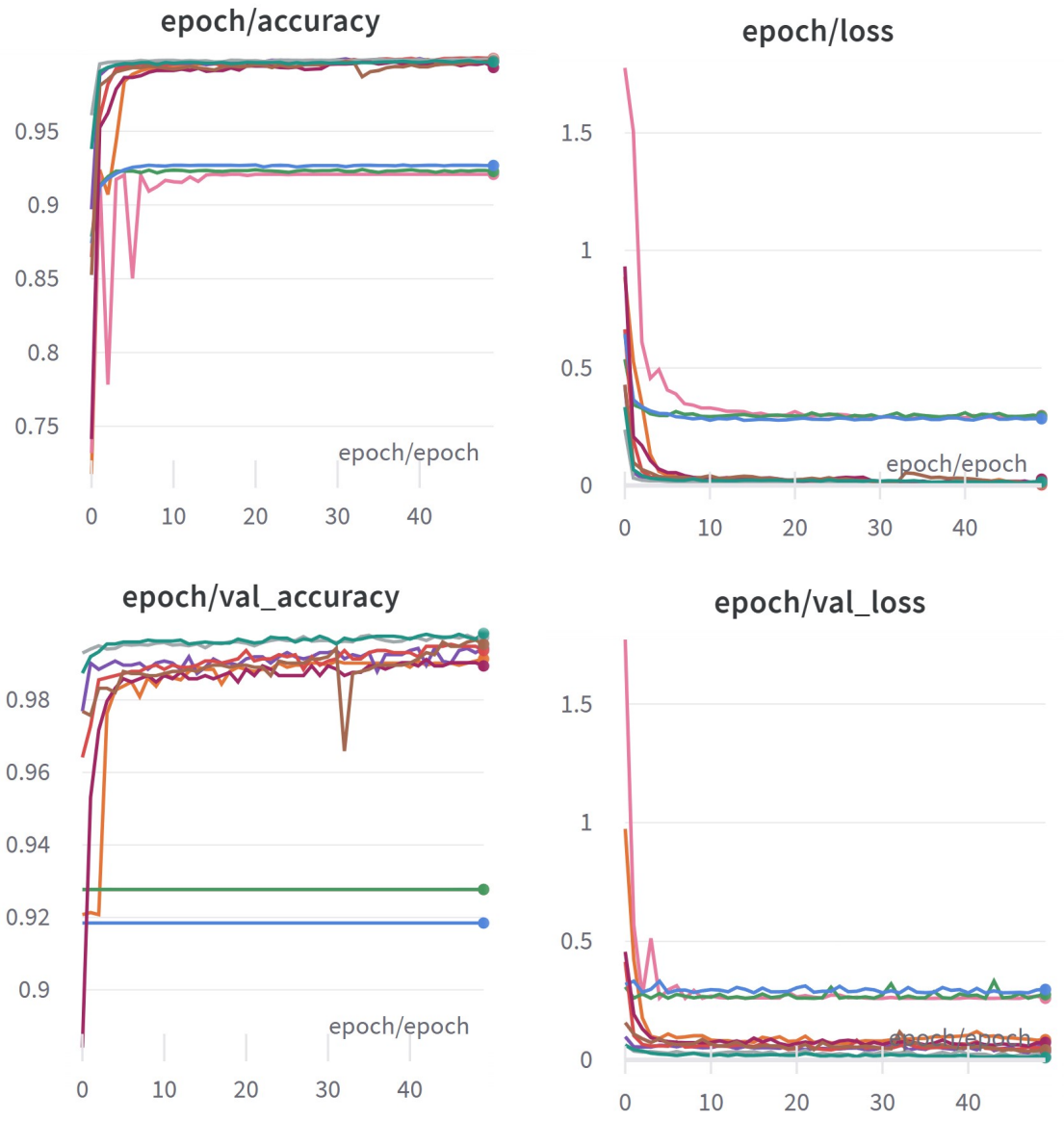


Figure 5.7: Neural Network for Pediatrics with Hyperparameter Tuning

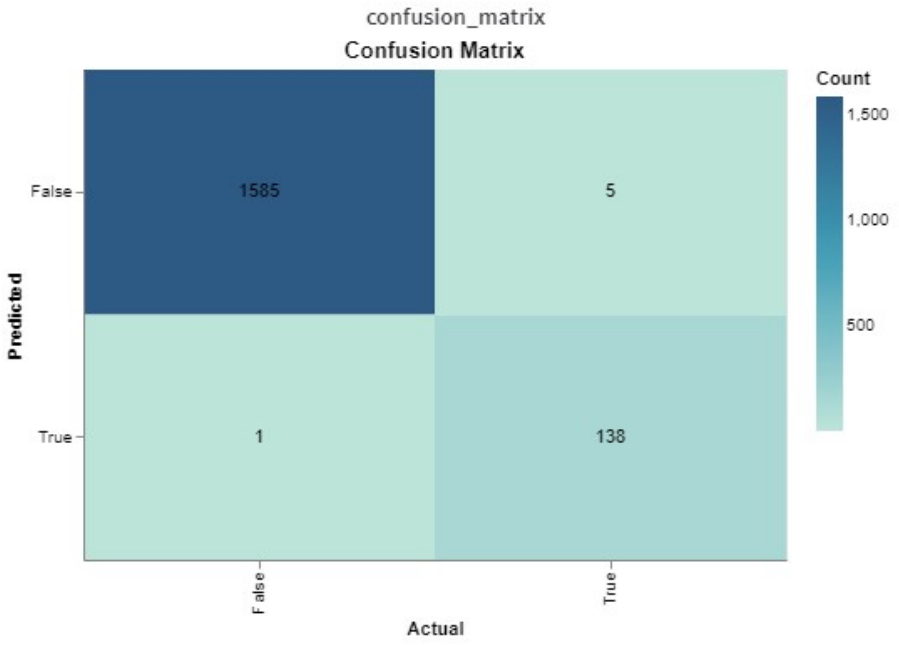


Figure 5.8: Neural Network for Pediatrics Confusion Matrix

#	**Activation Function**	**Batch Size**	**Learning Rate**	**Rows**	**Columns**	***Accuracy***	***F1 Score***	***Loss***
1	tanh	256	5e-4	10,000	100	*85.48*	*77.77*	*0.0078*
2	tanh	256	5e-5	10,000	100	*78.25*	*81.05*	*0.0256*
3	tanh	256	5e-3	10,000	100	*94.89*	*80.45*	*0.0474*
4	tanh	512	5e-3	10,000	100	*96.70*	*90.06*	*0.0216*
5	tanh	1024	5e-3	10,000	100	*82.02*	*79.85*	*0.0381*
6	tanh	1024	5e-3	20,000	100	*75.14*	*60.50*	*0.0208*
7	tanh	512	5e-3	20,000	100	*72.15*	*60.55*	*0.0165*
8	tanh	512	5e-3	10,000	200	*88.74*	*82.46*	*0.1138*
9	tanh	512	5e-3	5,000	100	*92.93*	*91.50*	*0.0784*
10	tanh	512	5e-3	30,000	100	*65.18*	*50.21*	*0.0027*
11	sigmoid	512	5e-3	10,000	100	*3.44*	*0*	*0.2829*
12	relu	512	5e-3	20,000	100	*4.67*	*0*	*0.2829*
13	sigmoid	256	5e-3	20,000	100	*2.59*	*0*	*0.0285*
14	relu	256	5e-3	10,000	100	*93.19*	*94.48*	*0.0160*
15	sigmoid	1024	5e-3	10,000	100	*12.65*	*0*	*0.2595*
16	relu	1024	5e-3	10,000	100	*11.34*	*0*	*0.2595*

Table 5.4: Pediatric Neural Network Results

5.5 Summary

The results for the adult LSTM and neural network experiments remained stable, with minimal changes when the parameters are altered, especially in the Neural Network model. SMOTE, a technique for handling imbalanced datasets, is unable to significantly optimize the results. The number of rows, particularly the ratio of COVID-19 to Long COVID patients, had a substantial impact, with more rows leading to worse results. Although the study had about 25,000 Long COVID patients, accuracy remained around 90% when tested on 500,000 COVID-19 patients. Changing the number of columns imposes a marginal impact on accuracy, but the accuracy measure never drops below 85%. The experiments are conducted over 30 epochs. In the case of the pediatric LSTM and Neural Network experiments, the prediction results are more variable compared to the adult experiments. SMOTE is not necessary again. With a ratio of 10,000 COVID-19 patients to 1,101 Long COVID patients, changing the number of rows significantly affects predictions. In fact, altering columns slightly influenced accuracy. The 'tanh' activation function outperforms the 'sigmoid' and 'ReLU' counterparts, especially in the Neural Network model. The experiments are conducted over 100 epochs.

Chapter 6

COVID-19 Impact on Local Food Delivery Business

This chapter explains the type of data and the process of how that data was acquired while shedding light on the machine learning models used for the analysis of this data. The data consists of variables and information that are in relation to a food delivery company. This means we are working on data that consists of information on restaurants, food, and menu items, customers, and their potential preferences regarding such food items that they order. We will go deeper into the implementations used to have the raw data for analysis, called pre-processing. The main steps incorporated for pre-processing involve Extraction, Cleaning, Recoding, and Merging, which are explained below -

- Extraction - Explains the process of acquiring the raw data.
- Cleaning - Clearing out the unnecessary details and/or variables that are unneeded.
- Recoding - Renaming and organizing data to incorporate it all in a CSV file or database.
- Merging - Putting all the processed data together to handle analysis.

After that, we discuss the various machine learning models ranging from graphs and visualizations to heat maps and clustering analyses showing multiple correlations in different aspects and levels of the data. As elaborated further below, the analysis used Python and its respective machine-learning packages.

Fig. 6.7 depicts the framework that handles data acquisition prior to the data pre-processing module.

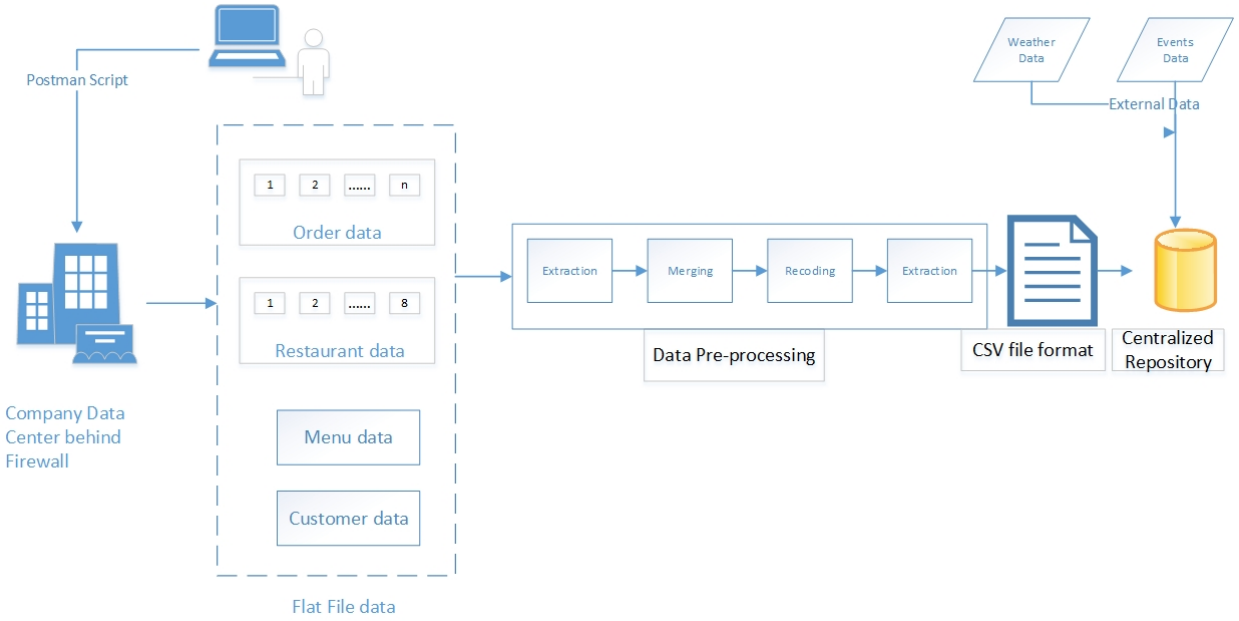


Figure 6.1: The Framework of Data Acquisition and Pre-processing

6.1 Data Acquisition

We first describe the data collection process and then articulate the pre-processing procedure. The acquired data are harvested from the company that is engaged in providing food delivery to customers. We used multiple HTTP API requests with the help of the Python and Postman applications. We leverage POST requests using an account with appropriate privileges provided by the company that allowed us to use the API endpoints to export data in JSON formats. Thereafter, we were able to use GET requests to collect data related to three different entities: customers, restaurants, and orders. The timeline for data extraction is set from January 1, 2017, to May 14, 2020. The data acquisition is carried out in two phases. In the first phase, we acquire data of approximately 60,000 orders spanning from January 1, 2017, through December 31, 2019 (3 years). In the second phase, we collect additional data in light of the advent of the COVID-19 Virus, which adversely affected most of the businesses across the US. This data spanned from January 1, 2020, to May 14, 2020. While we collected the data from January to May to make it look contiguous, we analyzed data from March 1 to May 28, 2020, for COVID-19.

The API requests offer information of 2000 orders per request. After we run each request 35 times, the data are compiled in one final CSV file for further data extraction. The variables extracted are summarized in the table below.

The variables mentioned in the table above are then added to the MS SQL tables, which are further processed using Python and its recommended libraries for analysis and machine learning. The data processing mechanism is highlighted and elaborated below to show the steps for preparing data for analysis. Below we illustrate the database diagrams, which are explained in the upcoming sections.

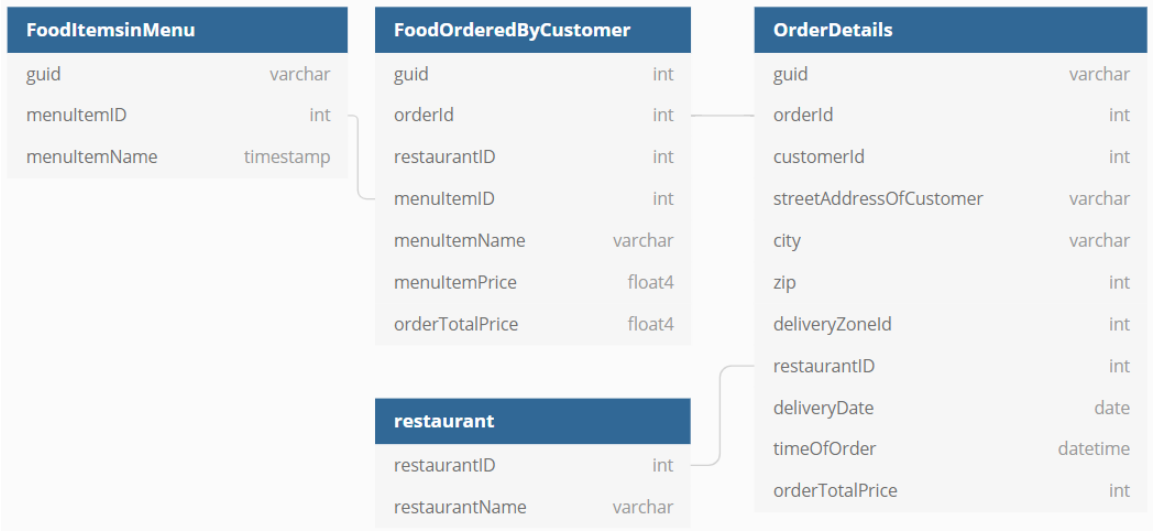


Figure 6.2: Food ordered by Customers is saved in the above format along with the important information regarding restaurants

Variable Extracted	Description
orderID	ID for each order made (therefore, we had over 60,000 orderIDs).
customerID	Each individual had his/her own ID. We are not supposed to use name, email, phone number or addresses of customers as that would be breach of privacy. This helped tracking order with their corresponding customers without any privacy issues.
AddressID	This ID is specific to an address as one customer can have multiple delivery addresses.
Address Type	This determines whether an address is a residence (RES), apartment (APT), business (BUS), hotel (HTL), hospital (HOSP).
Address	This variable was tricky to handle due to privacy concerns. We had to cut out the house/apartment number and name of residence, which only left the street name. This does not help to determine demands from a geographical basis.
City	The towns as specified by Virtual Diner to analyze.
Zip	Zip code.
Delivery Zone ID	Each ID determined distance of that address from Virtual Diner. Higher numbers meant that the address given is farther.
DeliveryType	To be delivered to address or if it is pickup.
DeliveryDate	Date of delivery.
DeliveryTime	Whether delivery is to be made ASAP or at a specified time.
ItemsInOrder	Number of items in the order.
EntityID	Each restaurant has unique ID regardless of the fact that it belongs to the same chain of restaurant. Thus, each outlet has a unique ID.
Restaurant	Name of the restaurant.
PickupDateAndTime	Date and Time when Order was picked up.
orderTotal	Total of the entire payment amount including tip, delivery fee and tax.
orderSubTotal	The payment amount excluding tip, delivery fee and tax.
orderStatusChangedTime	Time when order was delivered or picked up - What is being used to do the delivery time analysis.
CuisineID	The ID of cuisine for that particular restaurant (this is not universal).
CuisineName	Type of cuisine (for example - Chinese, Indian, American, etc).
RestaurantAddress	Address of restaurant is useful to determine from which outlet the order is being picked up from.
RestaurantCity	The towns as specified by Virtual Diner to analyze.
RestaurantZip	Zip code of that restaurant.
RestaurantGeoLat and	Latitude of that restaurant.
RestaurantGeoLong	Longitude of that Restaurant

Figure 6.3: Customers' address information is desensitized first before any analysis

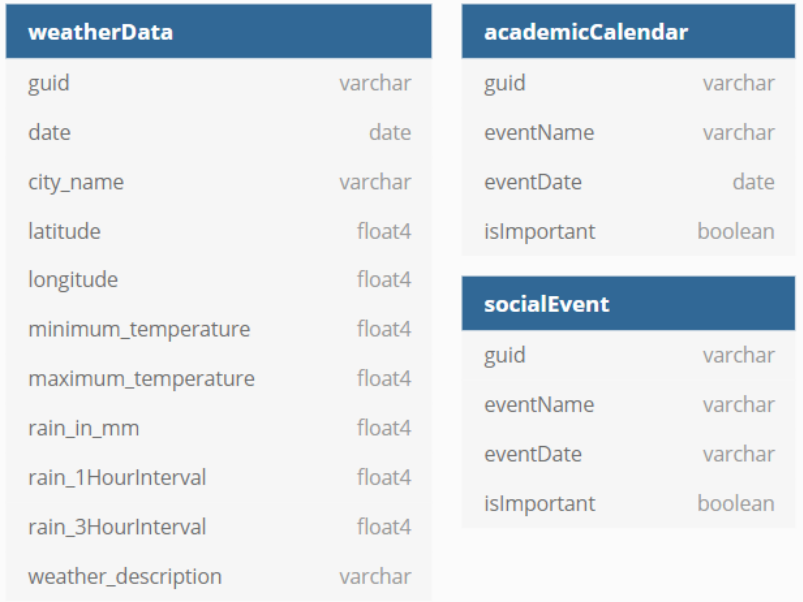


Figure 6.4: Weather, academic and social calendar information is used to analyse the external factors that have affected the sales of the company besides COVID-19

6.2 Pre-Processing

Now we elaborate on the procedure of processing raw data for analysis. This section explains the new variables and tables that are added and the reasoning behind the aforementioned modifications. We also elaborate on the steps for extraction, cleaning, recoding, and merging for all the parts below.

We start this section by explaining the acquisition of external data.

6.2.1 External Data

We extracted information about major social events in the city and academic events since January 1, 2017. The values for the academic calendar inherit the University's academic calendar released for every academic year and its semesters. This calendar is comprised of major academic events, including when the semester started and began, examination periods, and holidays. The calendar also contains the dates for each of the events and how many days they lasted.

The social events data were acquired with the help of online websites and organizations that provide us with the information in spreadsheets. This was cleaned manually as the formatting was highly inconsistent. The cleaning required all the different fields to be organized in their respective columns, which consisted of the event name, time, and date. All this was inserted into the *socialEvent* table.

Below is explained how the pre-processing has been implemented for such external data, that is, academic and social events.

- *Extraction* - The extraction process has been implemented based on the format or availability of the original data. The academic calendar data was in a PDF format that was The data and events had to be manually entered as the original files online on the University website are all in PDF format. Then, all of these events are inserted into the *academicCalendar* table.

- *Cleaning* - The academic calendar part did not require a lot of cleaning as the data was succinct. The social events spreadsheets provided by the organizations required manual cleaning and re-organization as the required fields - name of event, date and time were all out of place and not in separate columns. Moreover, the dates were given in a simple string format which proved slightly difficult to process as a date variable in the database or even in the spreadsheets itself. We also took into consideration only major sports events, college ceremonies, and inaugurations that had fixed dates during the year and that affected a large population. This had to be implemented to ascertain consistency as many smaller events caused confusion when determining their effect on sales. Smaller events never have a specific timeline or schedule for being organized. Larger events that encompass a massive amount of people can easily shake up sales to a noticeable extent that leads to a pattern when that same event occurs in the future.

- *Recoding* - The date and time fields, once formatted correctly, were added to the database. The major change implemented when recoding was the naming of academic events. Instead of using particular names that referred to the academic events (like Spring Break or Winter Break), more general terms were used, such as, "holiday" or "exam day" to find correlations between such periods throughout each year.

- *Merging* - All the data was modified in a CSV file that was imported into an MS SQL database. The tables are (as mentioned above) *academicCalendar* and *socialEvent*.

6.2.2 Ordering and Restaurant Information

The ordering information consists of important variables regarding the customers, their location, items ordered, and from which restaurant. These sets of files were the most essential during the analysis. The JSON format made it easier to extract data as the variables were assigned in a hierarchical manner. For example, under the "address" variable, there will be a list of variables that associate with *address*, such as, *street*, *zip code*, etc. This hierarchical

grouping by the JSON format also made it easier to search for any required information by using words or terms that are close in relation. The restaurant information provides material to work on which only pertains to all the restaurants and their branches. This helps to determine the location from which restaurant the order will be delivered and any other time-location factors.

Below is an explanation of how the ordering and restaurant information has been pre-processed for analysis. All the different types of raw data acquired are processed in different ways depending on the format in which they are obtained.

- *Extraction* - Two sets of files, which are of importance for the analysis, are taken from the API requests from the company. The files consist of information on orders and restaurants. As stated in Section 6.1, the API requests are executed several times to obtain all the data for the required time period for orders made by the customers. We acquire data in this fashion because the servers only allow a maximum of 2,000 orders in a single JSON file. This extraction constituted 70,155 orders.

- *Cleaning* - The data extracted, fortunately, did not require cleaning as only the required variables were taken. In other words, what data was needed was already determined.

- *Recoding* - The *customerOrderSummary* table includes all the customer IDs, restaurant IDs, delivery zone IDs, order count, and order value. This table shows the number of orders and the total value - a summation of all the orders combined by a specific customer for a certain restaurant. To protect data privacy, we substitute customer IDs for detailed customer information. The calculations are made using SQL queries by grouping together all the rows by customer IDs and then using SUM for the calculations on $orderCount$ and $orderValue$. Delivery zone IDs, restaurant IDs, and customer IDs are all extracted from the same orders.csv file, which has the compilation of all the orders needed.

- *Merging* - After obtaining all the files, we merge the files into a singular large JSON file, in which all the necessary variables are extracted using Python's *json* and *csv* libraries. Again, the hierarchical format of the JSON files allows easy extraction as the data variables have fixed formats and data types on all levels. When it comes to the restaurant data, there is no need to consolidate multiple files as there are only 165 restaurant brands in the concerned area. As such, all the restaurant data is maintained in a single file.

6.2.3 Address Information

Street information is incorporated to pinpoint popular locations from where orders are placed. Such information includes the desensitized addresses as well as street names for all the delivery addresses entered in the ordering database.

Following is the explanation and procedure that was used to pre-process the data used for the addresses and street names of customers.

- *Extraction*: The street names are taken from the address variable under each $orderID$ of every order made.

- *Cleaning*: The addresses entered by the customers are trimmed by simple string manipulation so that only the street names, city, and zip codes remain. This cleaning process is accomplished to preserve data privacy. Thereafter, we analyze the amount and number of orders being made from all the streets. One issue here is that there are thousands of errors and mistakes when the addresses are entered by the customers. The street names have to be identified individually and manually due to the high volumes of errors and inconsistency. This inconsistency problem makes it extremely cumbersome to automate the address-cleaning process. A total of $2,857$ streets are being processed.

- *Recoding*: We introduce variable $IsInvalid$, where 0 and 1 indicate invalid and valid street names, respectively. If $IsInvalid$ is set to 2, it means that the valid street has

been encountered in the data set before. Each street is also given a unique street ID once the street has been filtered out and desensitized.

- *Merging*: Finally, the cleaned and recoded data are inserted in the *customerStreet* table in the database.

6.2.4 Weather and Rain Information

The weather data for the town are obtained from the API requests passed to *api.openweathermap.org*, which returns an entire dataset in the CSV format with required variables - maximum and minimum temperatures and rain (in mm). Since the data is extracted in the CSV format, it is straightforward to import the *weatherByDate* and *weather* data into the SQL database as a table.

Following is the explanation for the procedure that was used to pre-process the weather and rain data to useful information for analysis.

- *Extraction*: The weather and rain data acquired from *openweathermap.org* encompass such information on various cities around the world. The correct geographical latitude and longitude are acquired after feeding the appropriate arguments for the API request with the help of the Postman application.

- *Cleaning*: The data obtained in the CSV format is easily imported as a separate table in the MS SQL database as *weather*.

- *Recoding*: A new weather table is created just for the variables that are needed. The variable names remain unchanged for the most part to maintain consistency.

- *Merging*: The two tables are separate from each other; only *weatherByDate* is used for the analysis purpose from the SQL Database.

6.3 Machine Learning Models

Now we shed light on the models and methodologies used for the machine-learning process in this study. We explain the versatile models coupled with their usage while describing the reasoning behind model adoptions. The data used was analyzed from the point of view of the customer and the types of orders being generated to recognize patterns while keeping in check the factors that have an effect on such a point of view. We elaborate on Heatmap visualizations and unsupervised learning, followed by the importance of customer segmentation and association rule mining with the help of clustering. These models offer theoretical underpinnings in the analysis of the data.

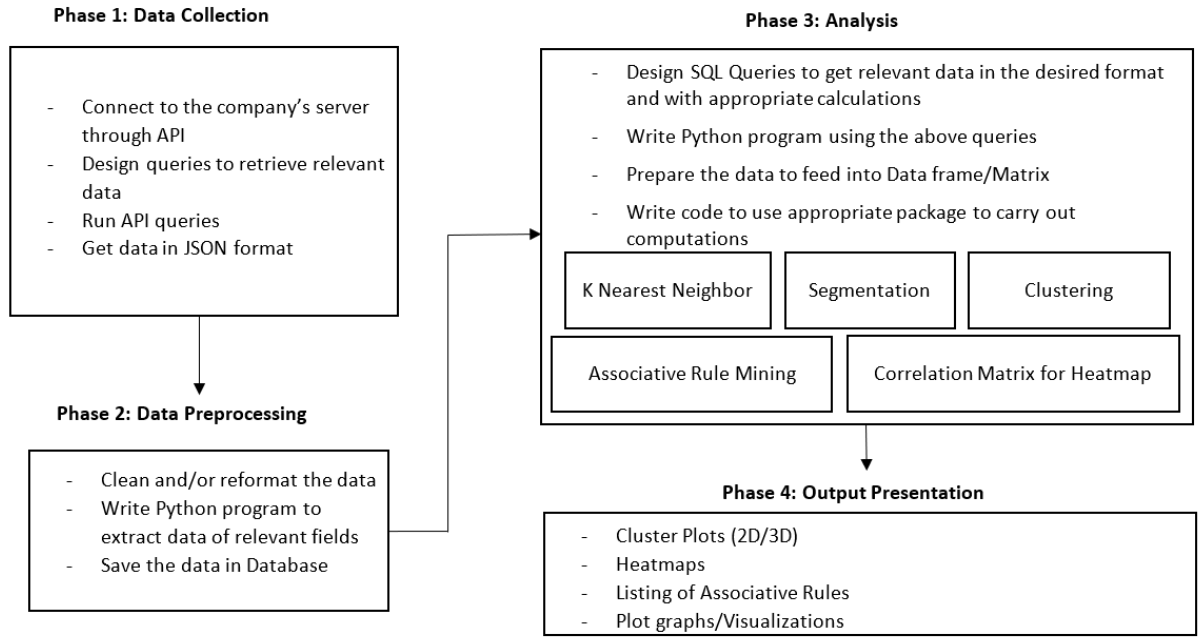


Figure 6.5: Bird's eye view of the process used for the report

6.3.1 Heatmap and Unsupervised Learning

Heatmap is a visualization technique to illustrate the relationship among various parameters. In this study, we leverage heatmaps to visualize correlation matrices related to the sales of the company and sales by restaurants while processing the same for sales by the

restaurants in the topmost cluster. This is also applicable to the sales order count of the company when calculated with the sales order count of the restaurants. Similarly, we plot the heatmaps with respect to revenue by delivery zone and menu item IDs. These heatmaps enable a company to focus more on the restaurants, delivery zone IDs, and menu item IDs that are relevant to revenue and order volume. At the same time, the heatmaps provide insights into the areas where restructuring is needed. In all the cases, the first column/row is the revenue of the company or the order count of the company. A high positive correlation signifies that an entity is important for the company to focus on for future improvements.

Customer segmentation enables a company to customize its relationships with customers. Such relationships are something every seller does with its customers but not necessarily in a scientific manner. Machine Learning addresses this customer segmentation issue using unsupervised learning. The most common algorithm used is *K-Means clustering* and; therefore, we carry out in this study customer segmentation by deploying the K-Means clustering technique.

Unsupervised learning is a type of machine learning in which there is no human-labeled data. In the case of the company's data that preserve privacy or our non-disclosure agreement (NDA) with the company allows us to look at only the order-related data that has sales value, number of orders, menu items, menu item value, delivery zone, street address of the customers. None of these parameters are labeled. Hence, some of the unsupervised learning algorithms such as clustering, anomaly detection, neural networks, etc. could be applied.

Specifically, clustering algorithms are the most appropriate in our application domain because we aim to segment the customers by putting boundaries around the clusters. Such clusters can be determined on the basis of the aforementioned parameters (see also Section 6.1). This delineation should suffice for further business processes.

6.3.2 Customer Segmentation

In K-Means clustering - a vector quantization method, n observations are partitioned into K clusters such that each observation belongs to the centroid of the nearest cluster. This clustering method minimizes within-cluster variances rather than regular Euclidean distances. Although the clustering problem is computationally NP-hard, efficient heuristics algorithms lead to local optimum within a reasonable time.

Let us look at the mathematical formulation of this method. Given a set of observations $(x_1, x_2, x_3, \ldots .x_n)$, where observation x_i is a d-dimensional real vector, k-means clustering aims to partition the n observations into k $(k \leq n)$ sets $S = \{S_1, S_2, S_3, \ldots .S_k\}$ so as to minimize the within-cluster sum of squares (WCSS). Formally, the objective is to minimize the following expression:

$$\sum_{i=1}^{k} \sum_{x \in S_i} \mid x - \mu_i \mid^2 . \tag{6.1}$$

where μ_i is the mean of points in set S_i. Vector x_i varies for different customer segmentations. We define the ith 3D customer segmentation as a vector x_i, which is expressed as:

$$x_i = [ov_i, oc_i, dz_i], \tag{6.2}$$

where ov_i represents the customer order value, oc_i is customer order count, and dz denotes delivery zone ID. Subscript i represents the i-th set of this three-dimensional vector. In the case of 2D customer segmentation based on order value and order count, the vector is simplified as

$$x_i = [ov_i, oc_i] \tag{6.3}$$

.

The computer packages implementing the K-Means algorithm take user input regarding the number of clusters to be used. The number of clusters is determined by the *Elbow*

method. Since this step is fundamental for any unsupervised algorithm, we adopt the Elbow method - one of the most popular schemes to determine the optimal value of k.

Customer segmentation is carried out in two ways. In the first method, we make use of sales revenue generated by the customers to categorize the customers into five camps, namely very low, low, medium, high, and very high. In the very-high segment, the number of customers is the lowest. In contrast, the number of customers in the very-low segment is the highest among the five categories. From the point of view of boosting revenue, the company can focus on high-value customers to ensure their retention by providing preferential services or other business strategies [64].

In the second method, menu item-wise segmentation of the customers is accomplished. The idea is to locate the most sought-after menu items and how pricing can be fixed in order to capture and retain more customers, thereby pushing up the number of orders by offering competitive prices coupled with discounts and freebies. At the same time, attention should be paid to the menu items that are not in demand. Candidate improvements on such menu items include upgrading quality, taste, or even pricing. For the menu items that have very poor demand, business decisions can be taken on whether it is worth continuing those items and these ideas can determine what all changes should be made.

As mentioned above in the first case, the customer segmentation is done on the basis of revenue generated by each customer. A few outliers are purged in order to have a meaningful plot. The K-Means clustering is employed with the help of the SciKit-Learn package. To figure out K for KMeans, we implement the ELBOW Method on the KMeans++ calculation.

Based on the graphs (as explained in the subsequent sections), we break up customers into five clusters, namely, very low, low, medium, high and very high in terms sales revenue generated by the customers. The highest value category refers to sales revenue roughly above \$4200.

The first cluster is a 3D visualization of clusters, where the three dimensions include sales order value, sales order count, and delivery zone ID of the customers. The data plotted

in Figure 6.27 clearly reveals that delivery Zone IDs 3 and 5 have the highest concentrations. The delivery Zone IDs between 15 and 25 yield the meager revenue.

Figure 6.28 plots 2D visualization where two dimensions are sales order value and sales order count of the customers. We observe from this figure that the majority of the customers have a total order count <175 and total order value $<2,500$. There are very few customers with a total order value $>4,200$. The second case is the 3D visualization of clusters, where three dimensions Sales Order Value, Sales Order Count, and Menu Item ID of the orders made by the customers have been used. The trend depicted in the figure sheds bright light on the menu items that generate high or low revenue.

6.3.3 Association Rule Mining

Now we move on to a second model spurred by the association rule mining [51]. Association rule learning is a rule-based machine learning method for discovering relations among variables in large databases. Association rule learning is a rule-based machine learning method for discovering relations among variables in large databases.

Let MI be a set of n menu items ordered by customers. We formally express set MI as

$$MI = \{mi_1, mi_2, mi_3, \ldots mi_n\} \tag{6.4}$$

where mi_i is the ith menu item ordered by the customers.

We refer to CO as a set of m customer orders placed by the customers. CO is written as

$$CO = \{co_1, co_2, co_3, \ldots co_m\} \tag{6.5}$$

where co_j is the jth customer order in set CO. Each customer order co in set CO has a unique order ID and contains a subset of the menu items in MI.

A rule is defined as an implication of the form: $X \implies Y$ where $X, Y \subseteq MI$. We explore the following three important concepts while implementing the association rule mining technique in our system.

- to determine rules from the set of all possible rules.
- to determine constraints on various measures of significance.
- to determine interests being used.

The most relevant constraints are minimum thresholds on support and confidence. If X, Y are itemsets, $X \implies Y$ is an association rule and T is a set of customer order transactions archived in a given database. *Support* is an indication of how frequently a menu item appears in a dataset. The support of X with respect to transaction set T is defined as the proportion of transactions t in the dataset, which is processed from the original raw data and contains. More formally, we express support $supp(X)$ as

$$supp(X) = \mid t \in T; X \subseteq t \mid / \mid T \mid. \tag{6.6}$$

Confidence is an indication of how often a rule has been found to be true. The confidence value of a rule, $X \implies Y$ with respect to a set of customer orders or transactions T, is the proportion of the customer orders or transactions that contain X which also contains Y. Thus, confidence is defined as:

$$conf(X \implies Y) = (supp(X \bigcup Y)/supp(X). \tag{6.7}$$

Given two itemsets X and Y, we define the lift of a rule as:

$$lift(X \implies Y) = (supp(X \bigcup Y))/(supp(X)supp(Y)). \tag{6.8}$$

If the rule will have a lift of 1, the probability of the antecedent and that of the consequent are independent of each other. Hence, no rule can exist between the two item sets. If the lift is larger than 1, two occurrences are dependent on one another, making those rules potentially useful for predicting the consequences in future data sets. The value of lift considers both the support of the rule and the overall data set. In our study, we use the above three values appropriately to come up with rules.

The aforementioned association rule mining and market basket analysis are mathematical modeling techniques [36]. Evidence confirms that if a customer buys a certain item from a group, the customer is likely to buy another item from the same group. This correlation can be envisioned as an if-then relationship; that is, if Item A from Group A then Item B is from Group B. Association rule mining has become a very powerful tool to predict the purchasing behaviors of customers. The company can easily tap this potential to suggest menus for its customers to try. Companies like Google and Amazon advocate for this technique to guide customers in a very pinpointed manner, thereby proving that the association rule mining techniques are very productive in increasing sales and customer loyalty. One of the most important aspects of association rule mining is that it does not use any personal data of the customer. Therefore, no legal issues can arise. The second important aspect of this technique is that since association rules are dynamically generated, the rules will keep changing on the current situations of the menus and the preferences of the customers. This dynamic feature ensures that fresh and relevant recommendations do not go stale.

Let us look at the topmost rule - Rule: Bacon Cheeseburger $\rightarrow$ Regular Fries. This rule states that if a customer buys a Bacon Cheeseburger, most likely she or he would like to buy Regular Fries. Other top rules are:

- Rule: Hand Cut Fries $\rightarrow$ BurgerFi Bacon Cheeseburger
- Rule: Hand Cut Fries $\rightarrow$ BurgerFi Cheeseburger
- Rule: Chips & Guacamole $\rightarrow$ Burrito

```
Rule: Bacon Cheeseburger -> Regular Fries
Rule: Hand Cut Fries -> BurgerFi Bacon Cheeseburger
Rule: Hand Cut Fries -> BurgerFi Cheeseburger
Rule: Chips & Guacamole -> Burrito
Rule: Burrito Bowl -> Chips
Rule: Burrito Bowl -> Chips & Guacamole
Rule: Burrito Bowl -> Chips & Salsa
Rule: Burrito Bowl -> Guacamole
Rule: Burrito Bowl -> Large Soda
Rule: Burrito Bowl -> Queso
Rule: Large Drink -> Cheeseburger
Rule: Little Fries -> Cheeseburger
Rule: Cheeseburger -> Regular Fries
Rule: ChickfilA Sauce -> Chick-fil-A Chicken Sandwich
Rule: Polynesian Sauce -> Chick-fil-A Chicken Sandwich
Rule: ChickfilA Sauce -> Chick-fil-A Nuggets (12ct)
Rule: ChickfilA Chicken Sandwich  -> ChickfilA Sauce
Rule: ChickfilA Chicken Sandwich  -> Waffle Potato Fries Large
Rule: ChickfilA Sauce -> Polynesian Sauce
Rule: Chips -> Large Soda
Rule: Queso -> Chips
Rule: John Coctostan -> Moe's Famous Queso
Rule: Little Cheeseburger -> Regular Fries
```

Figure 6.6: Association Rule Mining output for most popular menu item combinations

In the above example, if a customer buys Hand Cut Fries, she or he is likely to buy BurgerFi Bacon Cheeseburger or BurgerFi Cheeseburger. We demonstrate that the company can manage its logistics using association rule mining.

6.4 Experimental Results

We show in this section a practical way to determine customer preferences and to improve the churning of customers for the company while figuring out the possible factors that can increase its sales. In our experiments, we deploy the association rule mining technique in

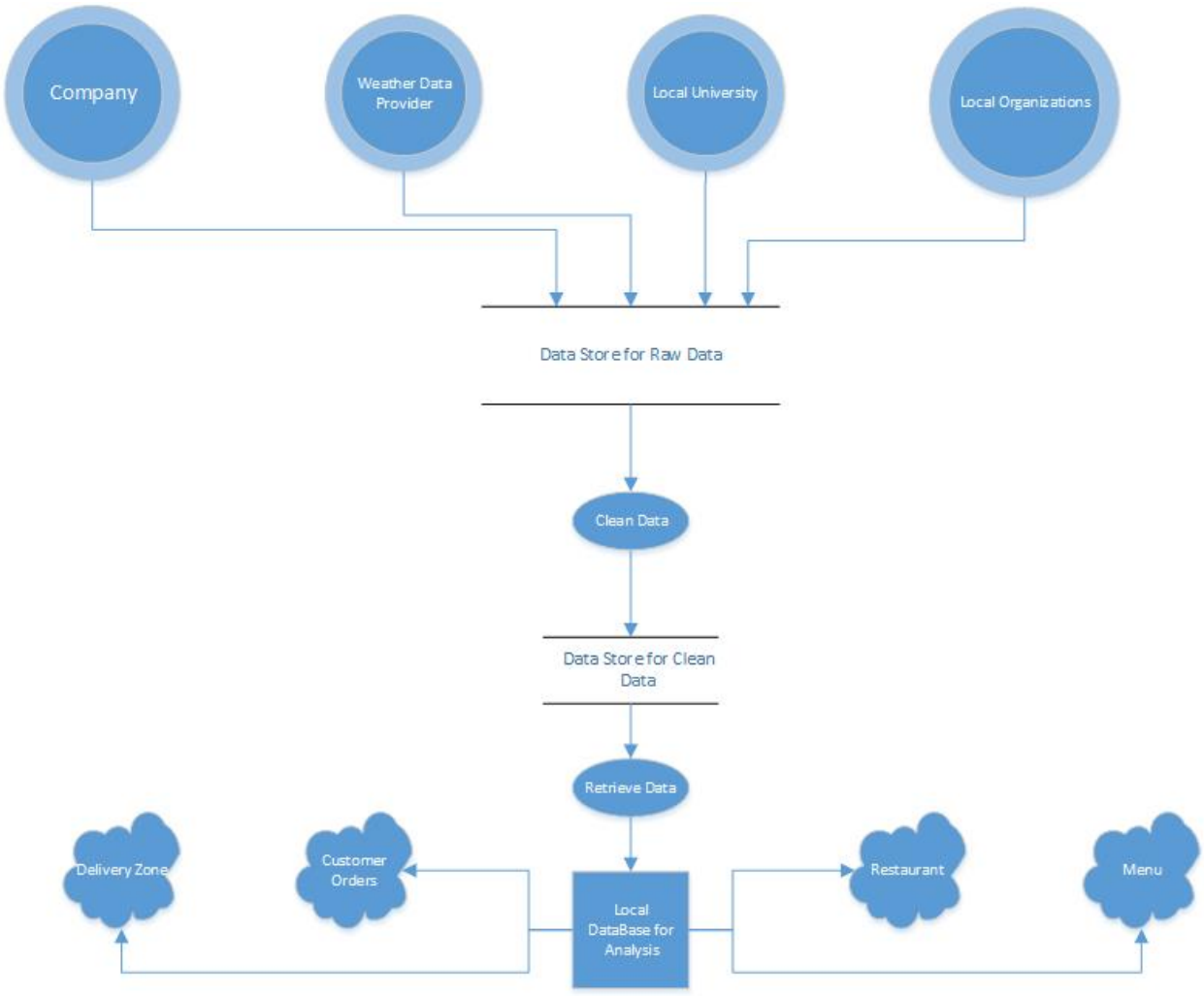


Figure 6.7: The raw data sets used and processed to useable data sets for the analysis

conjunction with the heatmaps, K-means clustering, customer segmentation, and histogram strategies.

6.4.1 Setups

Recall that this study is anchored to the analysis of sales and orders of a virtual diner delivery company. After collecting and processing their customer and sales data, we observed spatial and temporal fluctuations in sales over a period of 3 years from January 2017 through July 2020, where the duration from March 2020 onwards is regarded as the COVID-19 timeline. The overarching goal is to determine the factors that drive the demand and consequently, the sales revenue, so as to deploy the manpower and manage the logistics in a more efficient manner. This analysis includes the creation of visualizations to observe the patterns of sales and order counts by slicing in a different manner. Keeping in mind the socio-cultural milieu in which the business operates, analysis and visualization are carried out especially to delve into pertinent factors, such as the academic calendar of the local University, local weather conditions, graduation ceremonies of local universities, etc. As a part of a social event, only a few sports events by the University are added to the analysis.

As of now, there is a clear linkage between the sales and the beginning and end of the academic session of the local University. The results also show that at the time of transition, students or their parents depend upon Virtual Diner for their supply of food.

The other analyses such as clustering, customer segmentation, restaurant segmentation, cluster-wise heatmaps of restaurants, and associative rule mining are carried out that throw light on how the business is operating and which elements or entities are playing key roles in the revenue generation of the company.

6.4.2 Overall Sales Impact

For the COVID-19 Data, we are focusing on the duration from March 2020 to July 2020.

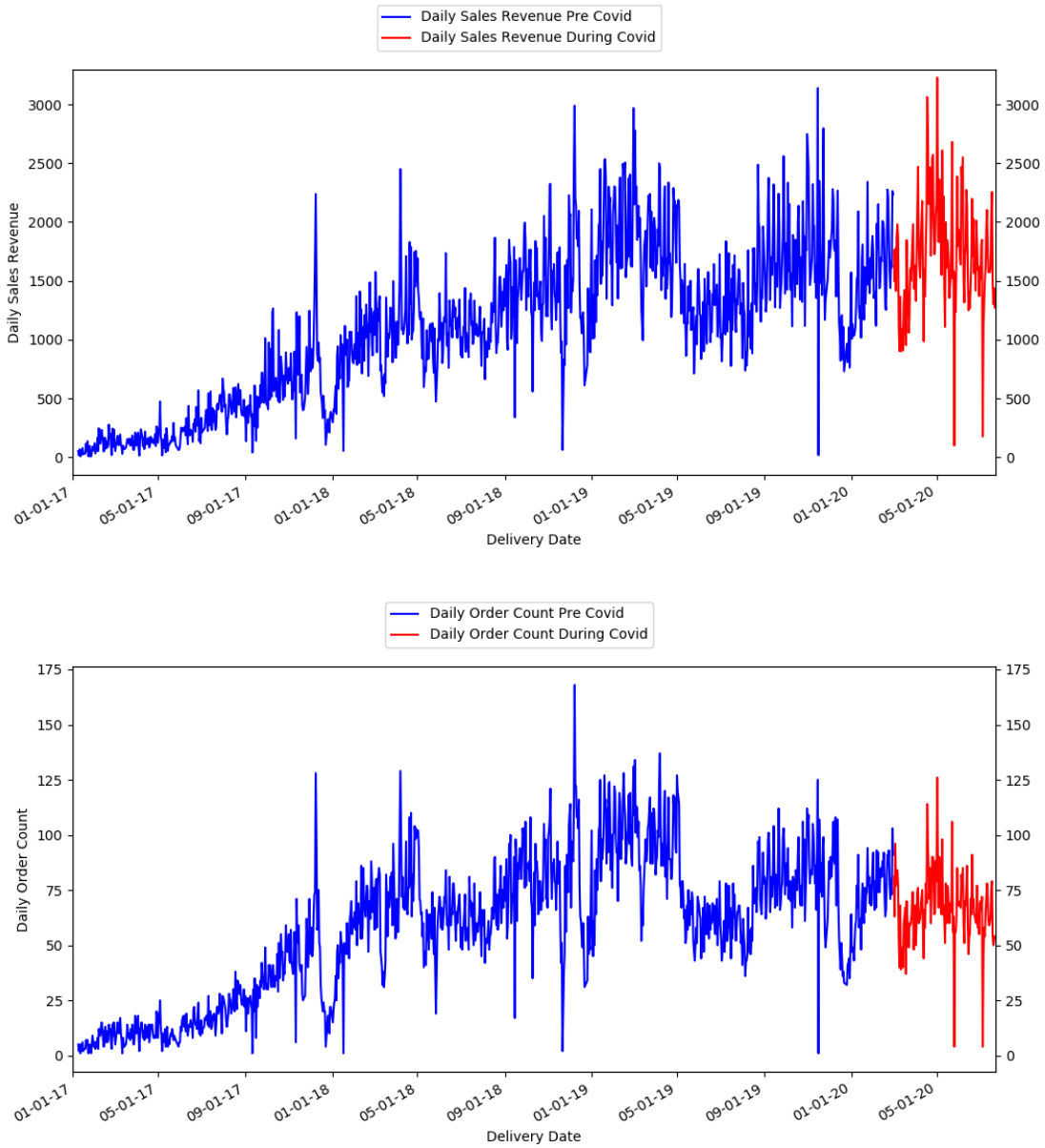


Figure 6.8: Overall sales chart in revenue and order count over the entire data set

In the above model, the blue represents the pre-COVID-19 phase and the red is when the pandemic witnessed a significant rise in the United States. We notice a trend of sales increasing with a high amount of fluctuation during the COVID-19 period. Below is a comparison of March 1, 2018, to May 14, 2018 VS March 1, 2019 to May 14, 2019 VS March

1, 2020, to May 14, 2020, to emphasize on this aspect. This duration is the initial 2-month period when the virus started to have a significant effect in the US (see Figs. 6.10 and 6.11).

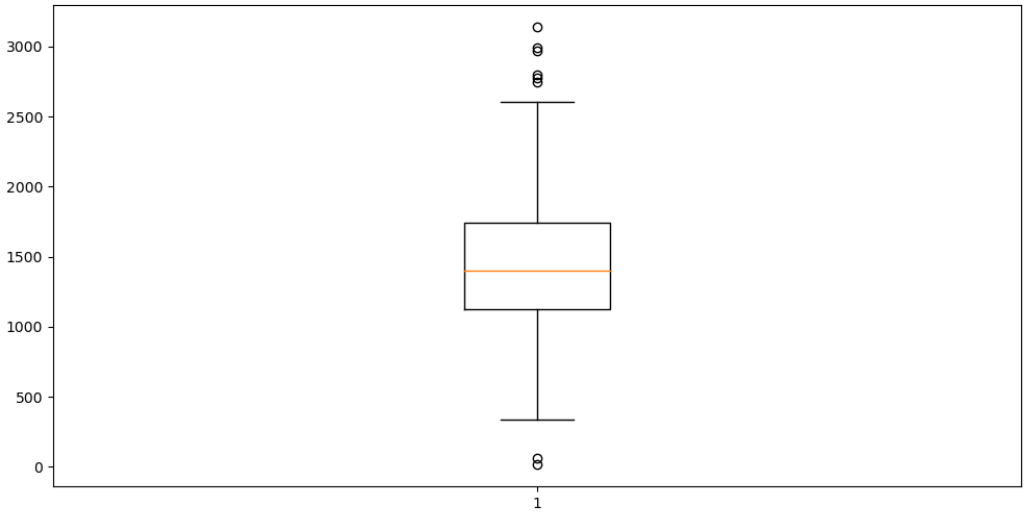


Figure 6.9: Box Plot to show sales statistics in May 2018

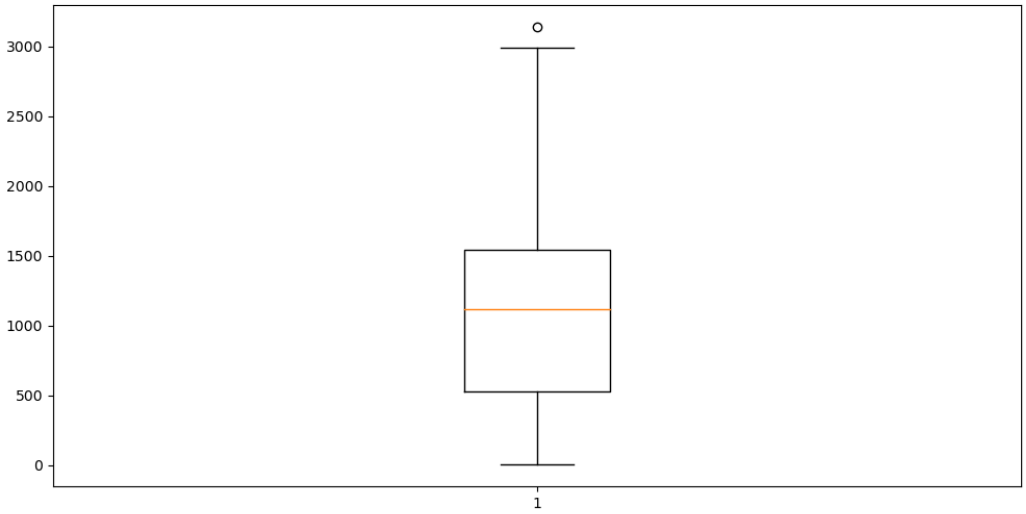


Figure 6.10: Box Plot to show sales statistics in May 2019

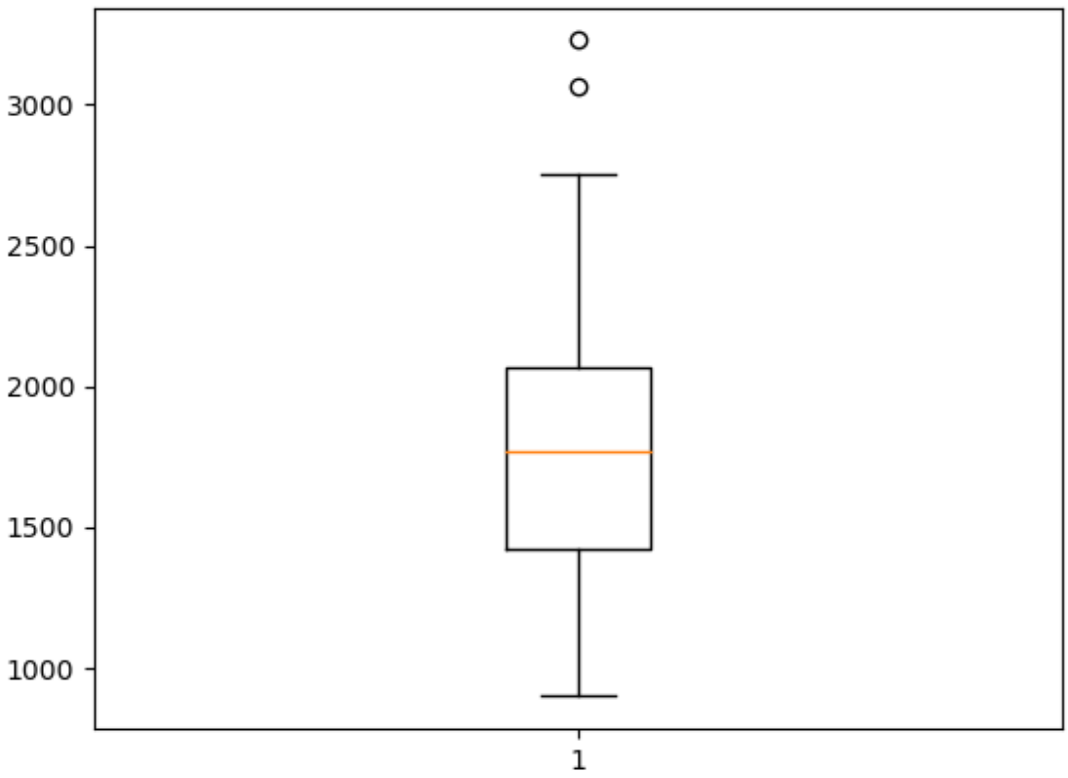


Figure 6.11: Box Plot to show sales statistics in May 2020

The average value of sales revenue during COVID-19 is $1762.41 whereas this value was $1526.64 in 2019 for the same period. The standard deviation has also been higher where the highest value in the third quartile (as per the boxplot) has reached $2753.05 revenue on a particular day. The lowest, when considering the first quartile has been marked to be $903.391 in sales revenue on a particular day during the COVID-19 period. During 2019, the same duration had a peak in the third quartile to be $2596.62 and the lowest taken from the first quartile was $709.07. In 2018, the average was $1392.86 where the highest value from the third quartile and lowest value from the first quartile were $2604.35 and $337.446, respectively. Fig. 9.43 unveils that the same trend continues till July 2020.

3-Days Window - Taking the results from Fig. 9.48, we observe a small spike on January 11, 2017. The sales were relatively low as it was the company's early years. However, the same occurrence was identified in the 2018 and 2019 periods as well. Classes began on January 22, 2018, Fig. 9.49 where we see a rise in sales on that day and an increase in sales 3 days prior to the day classes began. This is important to note so that the company can focus on the deliveries during these time windows. In fact, the time window for an increase in sales has always been that of 3 days (before/after the day of the academic event). This time window does not necessarily mean that the sales will always spike compared to the other days on the day of the academic event. Most of the students arrive at campus a few days before the classes begin. The red graph shows the spikes of academic events just before the classes begin. Mathematically speaking, this would seem as if the increase in sales and the beginning of classes have no relation as the dates do not coincide unless we take a running average of a set of days. Thus, we do discover a consistent pattern when there is a spike in sales a few days before classes begin every semester. This trend can be quite clearly observed when classes start on August 20, 2018, Fig. 9.50 where there is actually a drop in sales after the classes commenced (due to the reopening of dining halls) but a definite increase in sales with the 3-Days window as stated above.

Temporal Precedence - Holidays, a major impact on sales, affect the population of students in the University as most of the students leave for home outside the University. Figs. 6.12, 6.13, 6.14 unravel that the sales were sapped by the holidays in terms of both count of orders and revenue generated per day. As observed in these models, we do see a spike in sales during the end of classes and exam periods. The sales are consistent during the exam periods once the sales have reached a peak. Thereafter, there is little variation in the number of sales or even revenue of sales generated. This little variation is useful to know as examinations are week-long and happen three times in the entire year - late April, late July, and late December. The more notable aspect is the drastic drop in sales just after the exams are over and the spike in sales just before the beginning of these holidays. Such a drop

is consistent throughout other holidays as well (Spring break in March and Thanksgiving break in November are week-long breaks where most students leave the University). This situation of deteriorating sales is important so the company would not have to invest more in resources and labor during these times to increase efficiency. In fact, the company has seen an overall increase in sales due to these massive spikes in sales during the pre-holiday season.

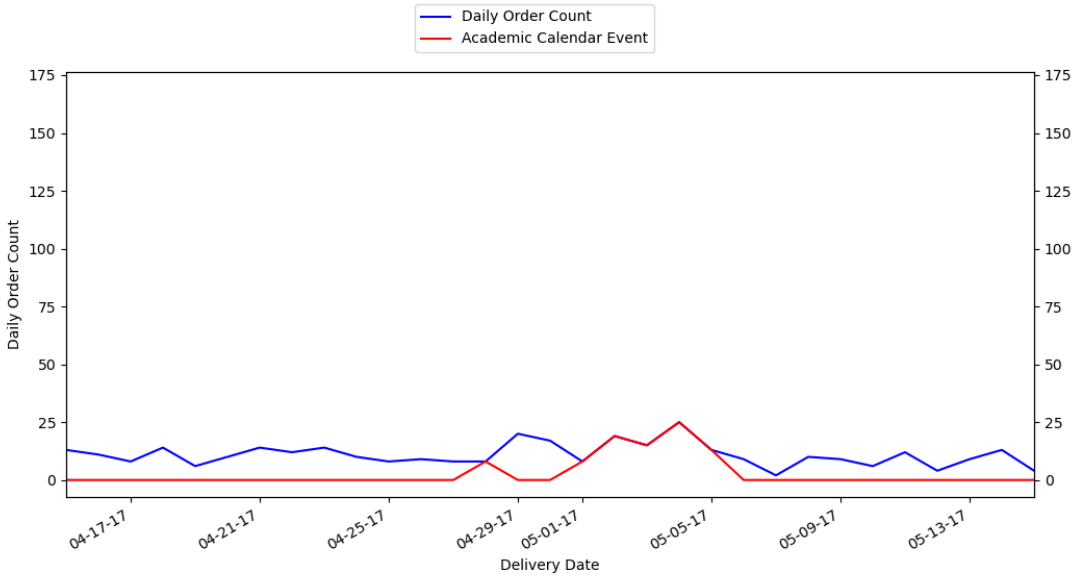


Figure 6.12: Sales count in April through May 2017

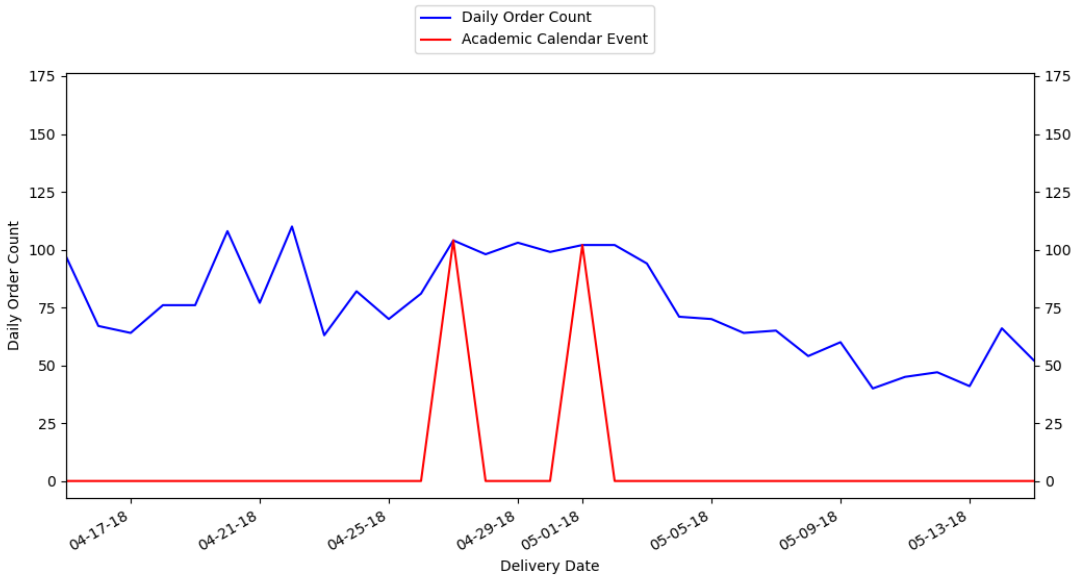


Figure 6.13: Sales count in April through May 2018

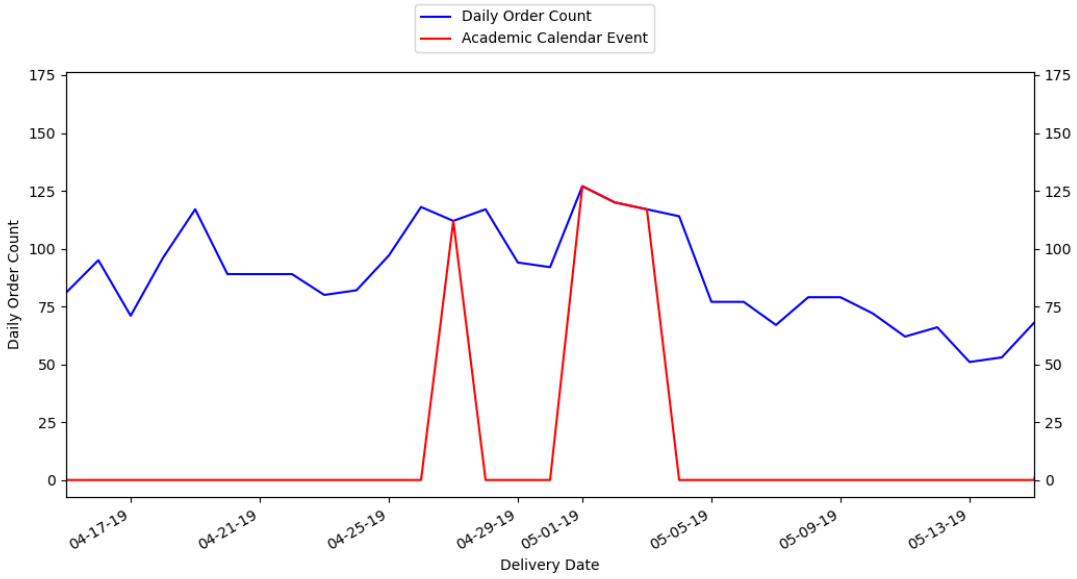


Figure 6.14: Sales count in April through May 2019

The above mentioned patterns of the occurrence of temporal precedence is evident of a strong cause-effect relationship. Moreover, we witness the same patterns happening again

in the same time for a different year. One could argue that the increase in sales could be linked to social or weather factors -

- Social events during that time are due to the fact that there are special academic events. So, such social events are more of a catalyst to the initial cause.
- Weather has not been the same during those periods of time throughout the period as found on the basis of historical weather data.

The above-itemized points strongly imply that the majority of orders are made by the students. After comparing the order count and revenue of sales generated during the summer of these three years, we pinpoint the trend taken from May through August of all 3 years. We see a drastic rise in sales every time the Fall semester starts, which is in August.

There is a drastic drop in sales once the Spring semester approaches the finishing line by May. Though, this is not completely true when considering 2017 as the sales increased even when the semester was over. Based on Facebook and Instagram social media updates by the company, we see that they have made advertisements on grocery delivery starting at $3.99. Such advertisements were posted on June 7, 2017. From that point of time, the sales increased steadily over the Summer even without the presence of students. This promotion of new sales of items shows that grocery delivery has a strong effect on households and residences and this brought the attention of local people of the University town area and nearby towns. Due to this, there has been a steady sales supply every summer. Such a situation indicates two things -

- Sales can spike based on the schedule of the students per academic year and increase the growth of the company.
- Steady sales can be maintained as local residents order from the company. The company is not entirely dependent on students for sales. A platform like Facebook is a great place to find people of that demographic.

In the models plotted in Figs. 9.42 and 9.41, the red spikes signify the periods when there was an academic event (e.g., classes begin and end, holidays and times of examinations). Immediately after such events, there is a clear rise or fall in sales revenue. The models demonstrate that these factors play a vital role in sales. There are other events during which no cause-effect relationship can be attributed. We conclude that there exists a handful of academic events that have no bearing on the company's sales.

A similar trend is true when it comes to a strong causal relationship between social events and sales. There has been an increase in sales whenever there is a major social event. Here, we are primarily observing the impact of sports events, major national events, and events organized by local universities as well. Social events can potentially determine the finer dips and spikes in a given period while the academic calendar pertains to the stronger dips and lows in the overall span of 3 years. We consider the Summers of 2018 and 2019 to compare the results by taking specific examples. The red graph above represents graduation ceremony days for the university in the local town and those nearby as an example. This model offers insight as to how social events can increase sales but not as drastically as major academic events. Most of the impact of social events is attributed to graduation ceremonies or sports events. For example, a spike and consistently high number of sales occurred two days after May 15 when major sports events began on the University campus. A drop in sales is kicked in once this three-day event is over. In 2019, the events held were different as there was a seven-day event instead of a three-day one. Such periodic drops in sales can potentially explain the long and high number of sales during that time period.

Social events may be single-day events or multi-day events. The local city has a lot of social events but only the events that generate a substantial floating population have an impact on the sales of the company. For this purpose, we look at the sales figure both in terms of sales revenue and order count and compare these metrics with the immediately preceding and succeeding sales levels to recognize either a rise or fall in sales in a clear manner (see Fig. 9.51). A rise indicates a positive correlation between the social event with

the sales of the company, whereas a fall implies a negative correlation between the social event with the sales of the company.

6.4.3 24-Hour Daily Sales Analysis

Fig. 6.15) plots a 24-hour model to analyze the what part of the day there are most sales. The data are taken for all the three-year period for the correlation matrix calculation. The dataset is of importance for a delivery company to utilize their staff members at the right times to make investments financially efficient. Another aspect covered by the model below is the probability of finding high revenue being generated for that day based on another time. We further elaborate on the model below.

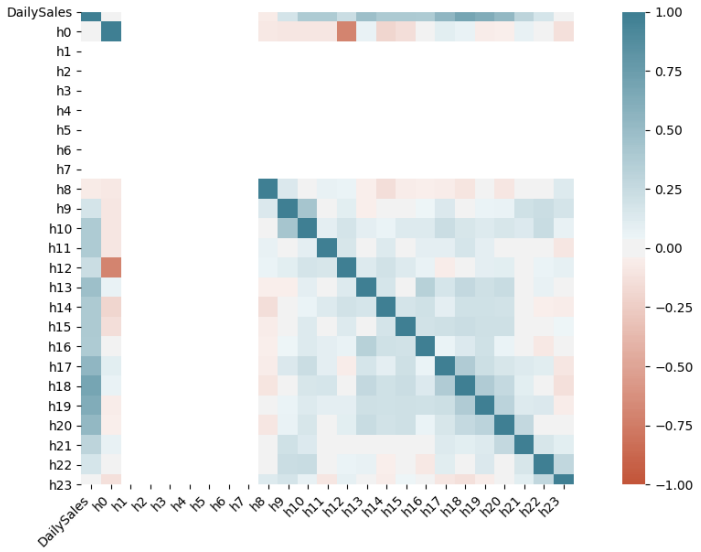


Figure 6.15: 24-hour daily sales revenue over the COVID-19 period

In the heatmap depicted in Fig. 6.15), a correlation matrix is calculated using sales during 12:00 AM to 12:59 AM as h0, 1:00 AM to 1:59 AM as h1, and so on up to 11:00 PM to 11:59 PM as h23 along with daily sales. The first row and column show the same values. If a particular day has more sales, then the time slots that are blue will have a chance of having an increase in sales. The darker the blue, the better the chance of that increase in

sales occurring for that time slot. A red block represents a chance of sales being decreased for that time slot if the overall sales for that day are increased. Fortunately, very few red blocks show up in the heatmap. This kind of information can serve as a powerful toolkit if many people are ordering for breakfast and then figuring out what the chance will be that they will order for lunch or dinner time.

As per our observations, from the above heatmap, it is clear that high sales revenue is generated with high probability from 7 PM to 7:59 PM. In fact, the correlation is very high for all hourly slots from 5 PM to 10:59 PM where the darker the color will represent a higher chance of also having greater sales. From the above results, we conclude that almost the entire revenue is generated through transactions from 11 AM to 10:59 PM. This information can be used for allocating resources accordingly in the future. These correlations show the probability of high or low sales in relation to sales during another time slot or hour. If a high number of sales from 10 am to 11 am has a high probability of resulting in a high number of sales from 6 pm to 7 pm, then the cell is marked as dark blue. The opposite is true when the cell is inclined towards red.

Above are the charts for the day-of-the-week analysis that provide information on what days represent a higher-sale pattern of the week taken from the entire dataset. Here, the numbers are represented as - 1 - Sunday; 2 - Monday; 3 - Tuesday; 4 - Wednesday; 5 - Thursday; 6 - Friday; 7 - Saturday.

In what follows, we plot a cumulative data model for all 7 days of the week.

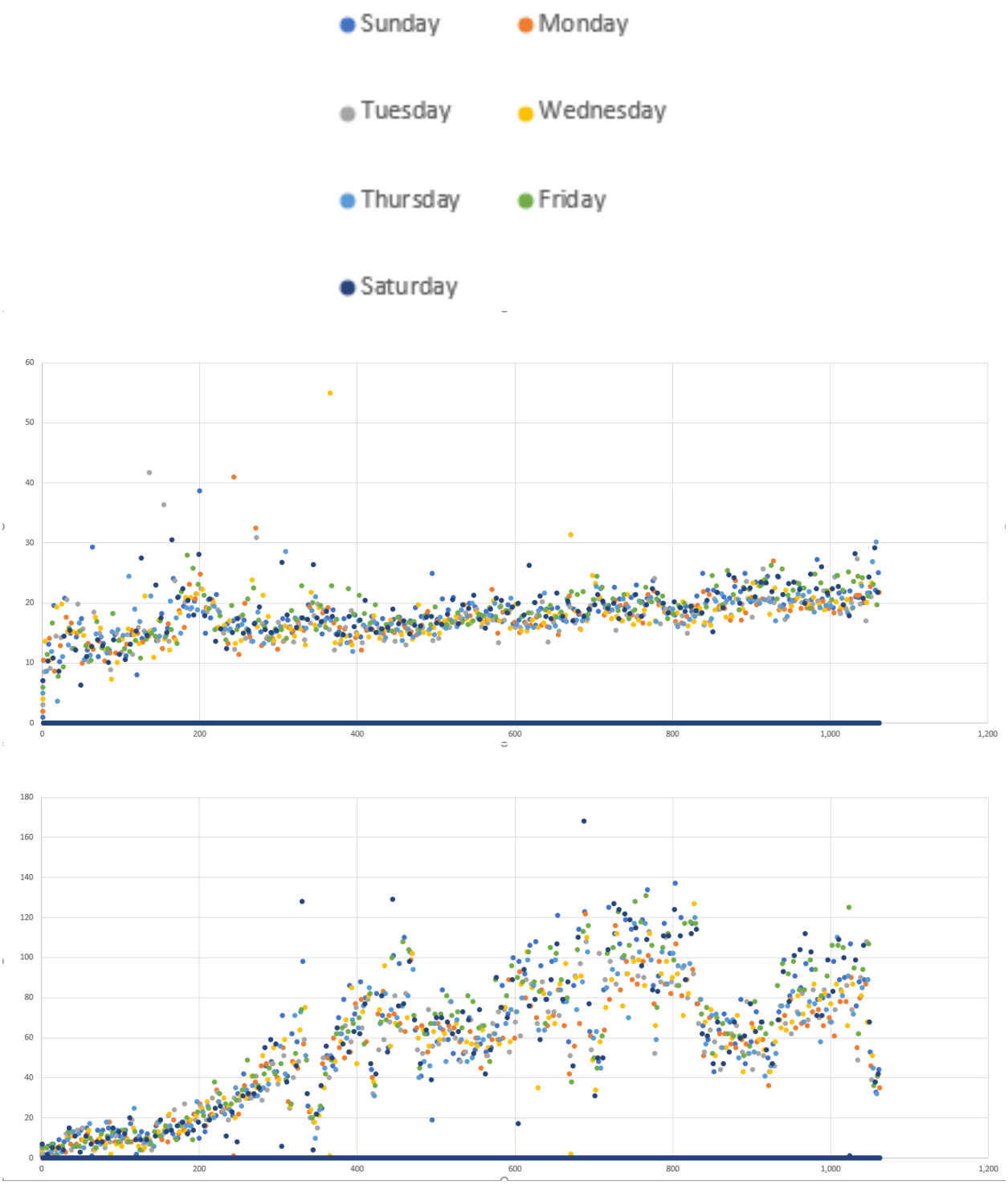


Figure 6.16: 7-Day Week analysis for each day of the week

We use the two models to capture the structure of sales per the day of the week. The first model takes the average of all the days. This is done on a day-by-day basis where each

day is added to the final cumulative and then the average is taken. These results reveal that there are some variations among the days of the week especially for particular weekends. Two parts that are notable are -

Around the 200^{th} day (July-August 2017) of the entire dataset shows a drastic increase in the average amount of sales. As mentioned in the Academic Calendar analysis. This drastic increase in the average amount of sales is contributed by the initiation of grocery delivery and the incoming students for the new semester. A significant increase in sales for the delivery company brings us to the second point.

The overall average of sales is steadily going up. There are small-term dips but after every significant spike in sales as discussed earlier, the mean (number of sales revenue till that day from the beginning/the number of days passed) increases.

The second model, which is similar to the previous visualizations from the Academic and Social analysis, represents the sale numbers as indicated by a point in the graph. The different colors show which day that order was delivered. The spike in sales is observed to be occurring during weekends, primarily as most of the points in the graphs. These sale spikes are especially notable when there are events happening around the university campus or in town. When marking the points above 80 count of orders in the y-axis, we observe that most of these are on Friday, Saturday, and Sunday. Monday is the lowest where only where it only shows over 80 sales 9 times during the past three years. More often than not, the drops are associated with Saturdays and Wednesdays, where there are instances when the order count is below 20. This observation is not a surprise and the reason is two-fold. First, weekdays in general have lower sales relative to weekends. Second, many outdoor events occur on Saturdays which can explain the sudden drastic drops on that day in particular. Such drops in sales would occur during summertime when no students are present on campus. Since the overall average has been increasing steadily, this low threshold has also witnessed a rise.

6.4.4 Delivery Zone ID and Street Analysis

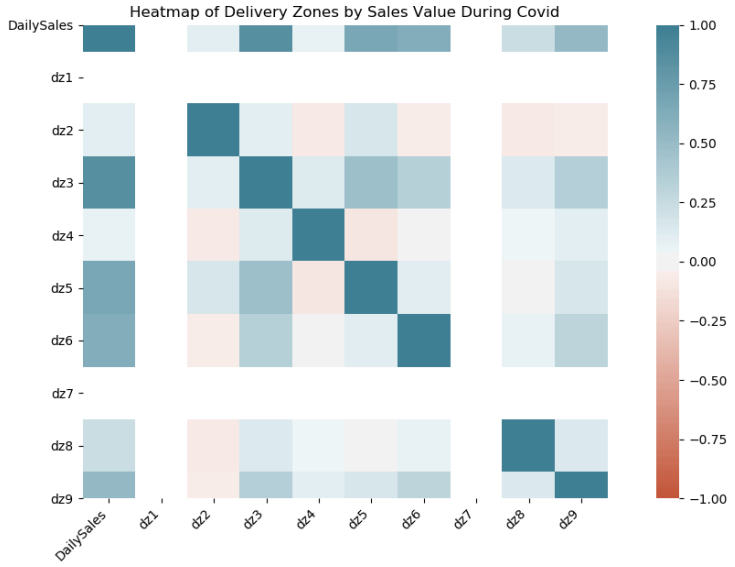


Figure 6.17: Sales count generated from each delivery zone ID

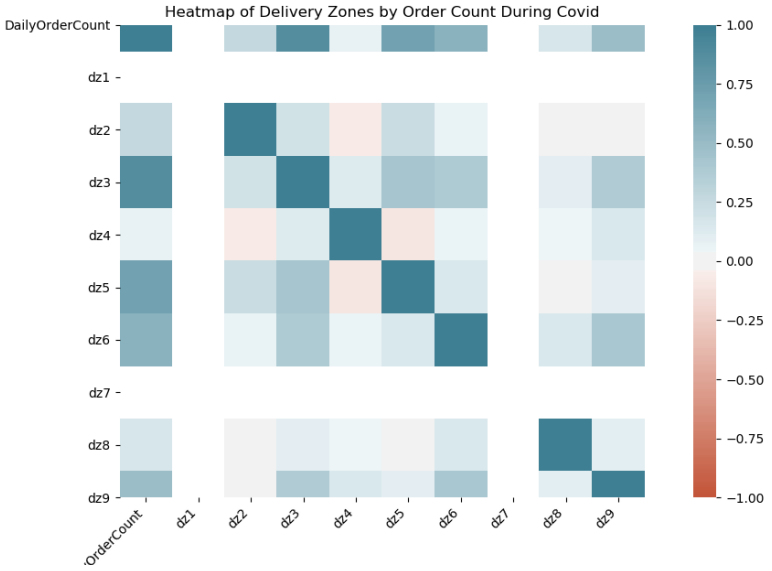


Figure 6.18: Sales revenue generated from each delivery zone ID

In the heatmap plotted in Fig. 6.18, a correlation matrix is calculated using sales on each day for delivery zones named dz1 for $deliveryZoneId = 1$, dz2 for $deliveryZoneId =$

2, and so on along with daily sales. From this heatmap, it is clear that the daily sales metric is highly correlated with sales for delivery zone IDs 3 and 5. Delivery zone IDs 2 and 6 also have a good correlation. The other delivery zones are insignificant. This information can be used for dynamic resource allocation in the future. This map can be clubbed with a heatmap for hourly sales and combining the two I conclude that delivery zone Ids 3 and 5 appear to be the most demanding from 5 PM to 11 PM. The same concept applies here similar to *hourlySales* where we see if there are a high number of deliveries in a particular ID, we can see a high number of deliveries in another *deliveryZone* ID based on probability.

Restaurants that generated zero sales are excluded from the heatmap, where the total number of restaurants is 183. In spite of discarding non-revenue generating restaurants, the number of restaurants is still large. As a result, the heatmap is not clear. A correlation matrix is calculated using sales on each day for each restaurant. A restaurant named as '1' corresponds to $RestaurantId = 1$, '2' for $RestaurantId = 2$, and so on along with daily sales. In spite of the unclear heatmap, we conclude by looking at the top edge that a good number of restaurants have a high correlation.

Figure 6.19: Street Clustering using K-Means

Fig. 6.19 depicts street clusters representing the most popular streets from where orders are being placed from. The results indicate that cluster 5 has few streets as these 17 points represent the streets from 2800 ones in the dataset. We also apply the *k-elbow* method to determine the number of clusters to be formed. We then observe an increase in a number of streets from Cluster 5 to Cluster 3 (the cluster just below). The number increases further as we move down the clusters and shows where the food delivery company can promote popularity easily by observing the sales per street. Furthermore, many of the 2800 streets, geographically, cover less than 5 to 7 houses. Compared with long streets, short streets have lower chances of sales. Many streets encompass the university area, which consists of mixed

residences of students, faculty, and staff. The university area enhances the chance of orders being made.

After creating the clusters, we again form the heatmap models using the correlation matrix we have implemented for delivery zone IDs. These heatmaps intuitively show the probability of there being a high number of orders in one street based on the orders made from another street. The same concept is applied in the previous heatmaps. Scanning the first column, we examine the overall daily sales while displaying the probability of there being high sales from a particular street. Darker blue means that a high number of orders from that street is more likely if the overall sales as number of orders being made are higher. The first row also shows the same results. From the second column(row) and onwards, we illustrate the correlation between the sales made from one street and another. The darker the blue the higher the probability of sales or orders being made from that street when there are a high number of sales. For example, if we compare street IDs 531 and 712, we see that there is a probability of about 0.5 of high sales occurring in street ID 531 if there are high sales in street ID 712. The information is maintained in a database similar to the one plotted in Fig. 6.26

6.4.5 The Weather Calendar Effect

In the above temperature models (see Section 6.2.4), we consider the overall temperature variation juxtaposed with that of sales revenue and order count. Unfortunately, the relation between temperature and sales is quite opaque due to the fact that temperatures never reach any extremes in the current town area as it is located in the South. Thus, weather has a marginal impact on sales.

Considering the magnified models Figs. 6.20 and 6.21, we observe a large variation in spikes and dips for temperature VS sales revenue(order count) that do not seem congruent with each other. Given an even smaller window between December 1, 2017, and December 15, 2017, we observe a spike in sales even though there is a drop in temperature. This trend

potentially implies that the two variables work inversely. We see, however, the opposite of our proposed hypotheses when there is a drop in sales just after January 15, 2018, accompanied by a drop in temperature. The results reveal linearity with both variables again from March 1, 2018 onwards.

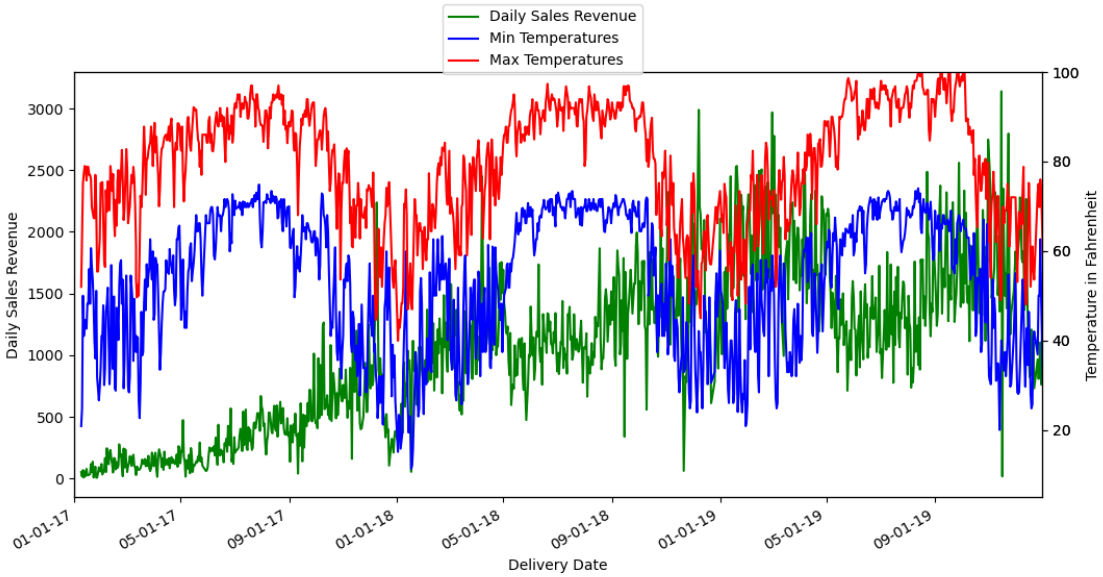


Figure 6.20: Weather analysis visualizations with sales revenue generated during the pre-COVID-19 period

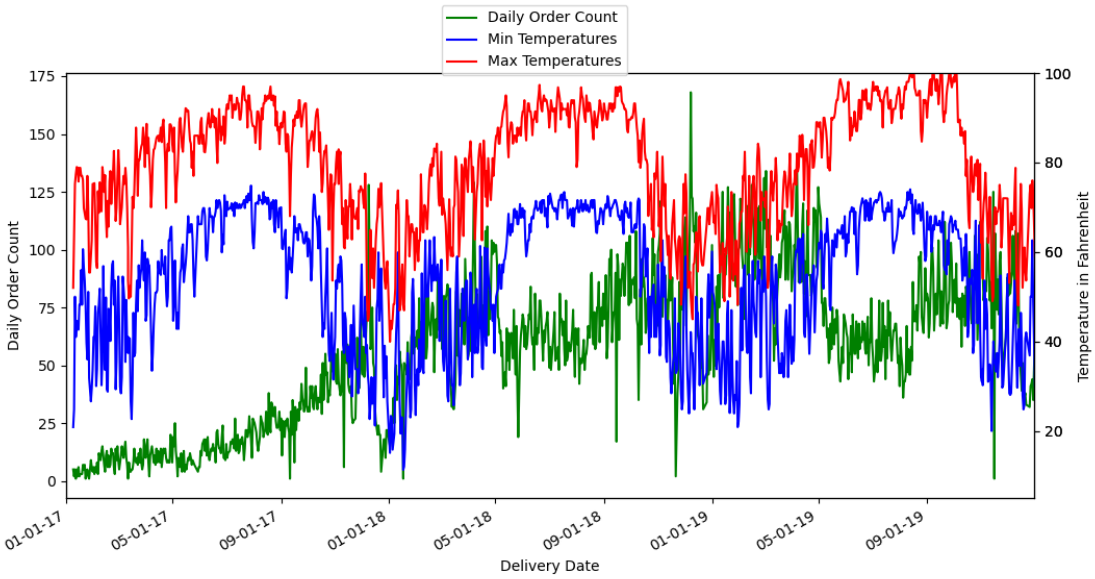


Figure 6.21: Weather analysis visualizations with sales count generated during pre-COVID-19 period

Focusing on the 15-day December window mentioned above, we must consider the situation as discussed during the academic calendar analysis where most of the students ordered food during the exam and dead week. The January window can be explained by the fact that dining courts were reopened for the semester. This scenario also explains the sudden spike a little before January 15. The students would order food while the dining courts are closed before the semester begins. Furthermore, during March, sales would remain consistent until Spring Break occurs in late March with certain dips and spikes depending on special social and academic events.

In what follows, we further elaborate on the effects of rain for the weather calendar.

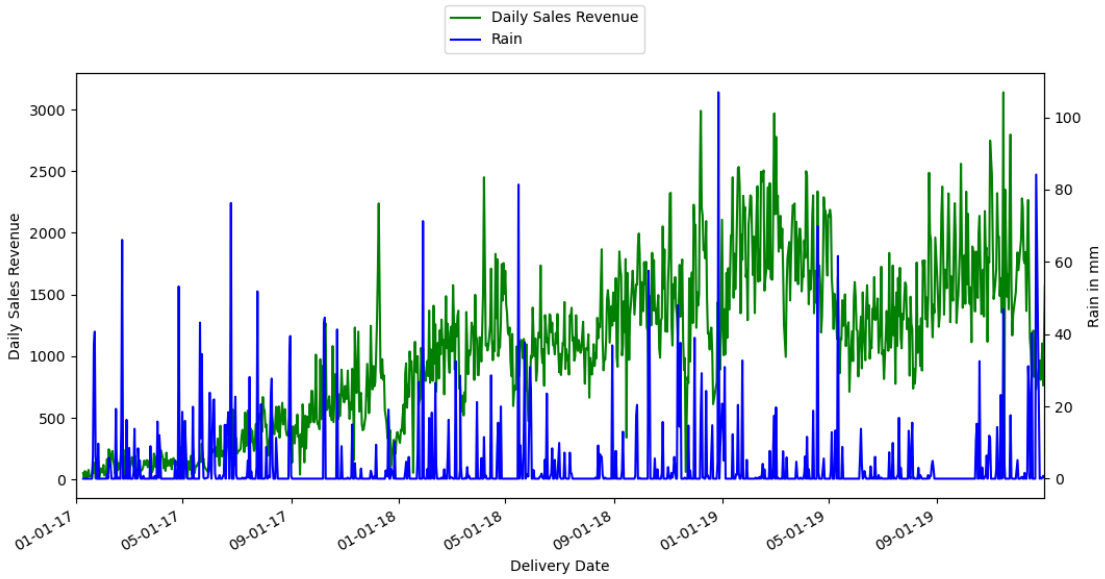


Figure 6.22: Rain analysis visualizations with sales revenue generated during pre-COVID-19 period

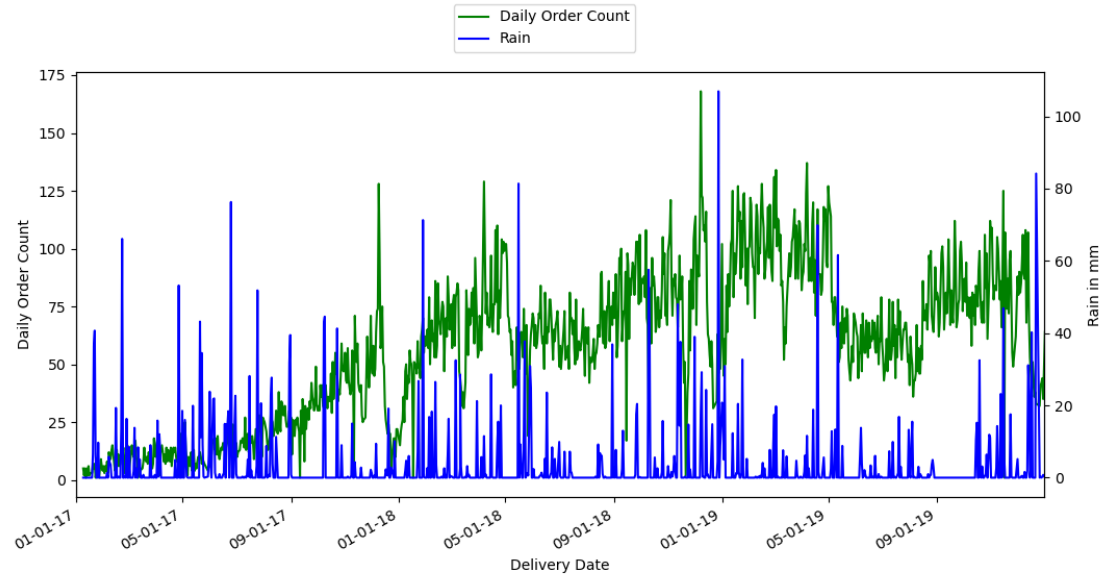


Figure 6.23: Rain analysis visualizations with sales count generated during pre-COVID-19 period

Even though higher sales tend to show up on rainy days, there is a lack of a guaranteed pattern that can predetermine the sales when taking rain into consideration independently. Nevertheless, rain still has a substantial effect in the following three ways.

- Rain can be treated as a secondary factor or catalyst for sales when we take academic and social events into consideration as primary factors; it does not necessarily mean that rain itself can impact sales in a clear-cut manner.
- The number of sales always climbs on the day it rains when compared with a window of the previous 3 days based on historical performance unless there has been a downward trend in sales.
- In case of a downward trend, rain can slow the rate of sales decreasing, but rain would not guarantee a positive turnout in sales. It is evident that rain is a catalyst in the positive direction of sales and tends to bring an increase.

6.4.6 Association Rule Mining for Market Basket Analysis

In this section, we delve into the implementation of customer segmentation and market-basket analysis using association rule mining when analyzing restaurant contributions and different menu items ordered by customers.

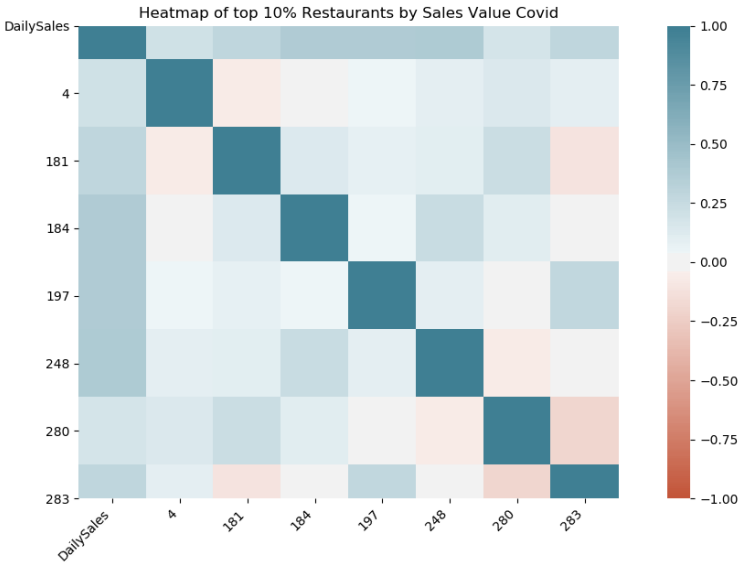


Figure 6.24: Cuisine analysis for top 10 percent of cuisines based on sales revenue generated

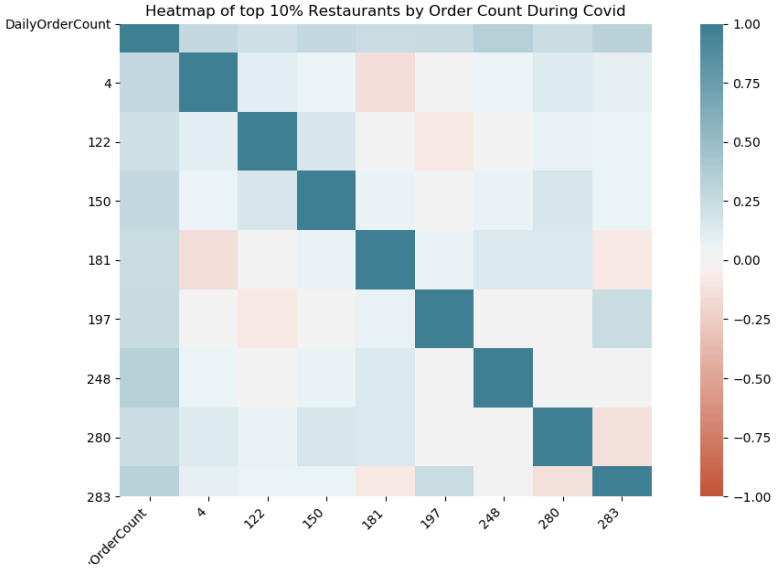


Figure 6.25: Cuisine analysis for top 10 percent of cuisines based on order count

Fig. 6.24 and 6.25 unveil the two heatmaps highlighting the restaurants that have sales revenue $\geq$ X percentage of the total revenue of the company. We are able to adjust the tier

percentages, thereby setting the model to show the top 20% or 30% of the top contributing restaurants arranged in descending order of their sales revenue. The same concept of heatmaps is applied here as elaborated for 24-hour, streetID, and deliveryID analysis (see Section 6.4.3).

customerId	cuisineId	restaurantid	restaurantName	cuisineName	cuisineWiseSalesVal
5373	162	181	Taste of Asia-Authentic Chinese Food	Chinese	2453.16
12233	162	181	Taste of Asia-Authentic Chinese Food	Chinese	2371.31
4466	51	62	Louie's Chicken- Cary Creek	Fast Food	1895.78
4466	52	62	Louie's Chicken- Cary Creek	Chicken Wings	1895.78
4466	54	62	Louie's Chicken- Cary Creek	Fish and Chips	1895.78
4466	53	62	Louie's Chicken- Cary Creek	Burger	1895.78
13577	258	257	99 Kabobs	Chinese	1718.29
12511	162	181	Taste of Asia-Authentic Chinese Food	Chinese	1695.52
2687	6	3	Chipotle	Fast Food	1604.19
2687	5	3	Chipotle	Mexican	1604.19
12837	162	181	Taste of Asia-Authentic Chinese Food	Chinese	1597.38
5373	258	257	99 Kabobs	Chinese	1594.56
4890	13	7	Fuji Sushi Bar	Japanese	1563.57
4890	14	7	Fuji Sushi Bar	Fast Food	1563.57
4890	12	7	Fuji Sushi Bar	Sushi	1563.57
4890	15	7	Fuji Sushi Bar	Chicken Wings	1563.57
8549	12	7	Fuji Sushi Bar	Sushi	1560.12
8549	14	7	Fuji Sushi Bar	Fast Food	1560.12
8549	15	7	Fuji Sushi Bar	Chicken Wings	1560.12
8549	13	7	Fuji Sushi Bar	Japanese	1560.12
3677	1	1	Chick-Fil-A- Breakfast	Fast Food	1477.56
57	6	3	Chipotle	Fast Food	1452.25
57	5	3	Chipotle	Mexican	1452.25
14317	162	181	Taste of Asia-Authentic Chinese Food	Chinese	1451.48
8021	5	3	Chipotle	Mexican	1405.6
8021	6	3	Chipotle	Fast Food	1405.6
2093	53	62	Louie's Chicken- Cary Creek	Burger	1394.42
2093	51	62	Louie's Chicken- Cary Creek	Fast Food	1394.42
2093	52	62	Louie's Chicken- Cary Creek	Chicken Wings	1394.42

Figure 6.26: IDs used for important restaurants and their respective information

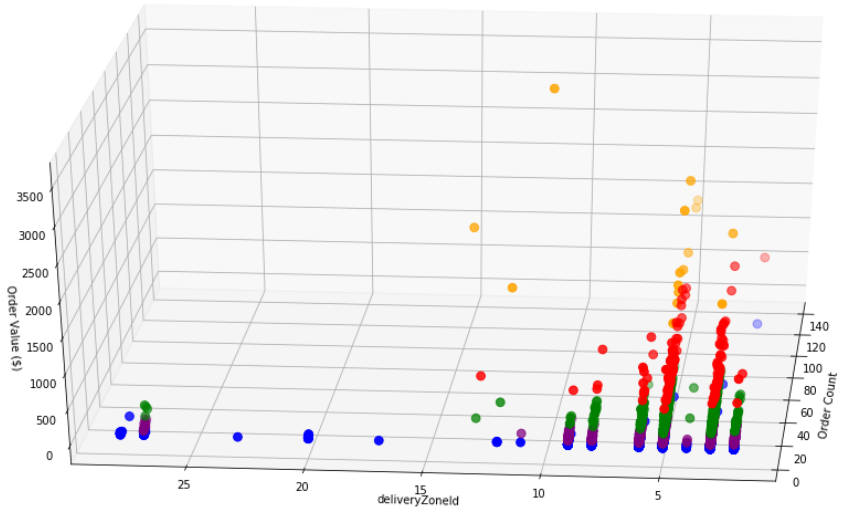


Figure 6.27: 3D Model to represent the relation between orders made, delivery, and order count using customer segmentation

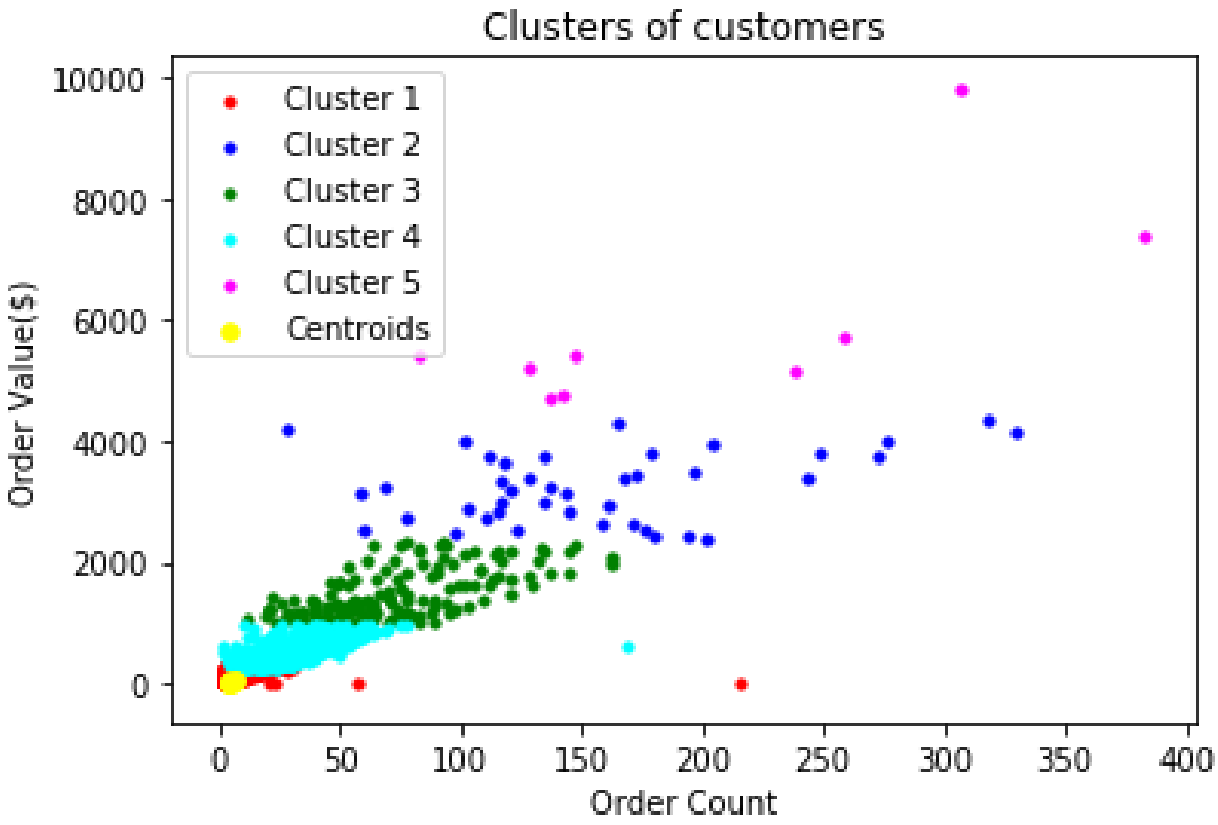
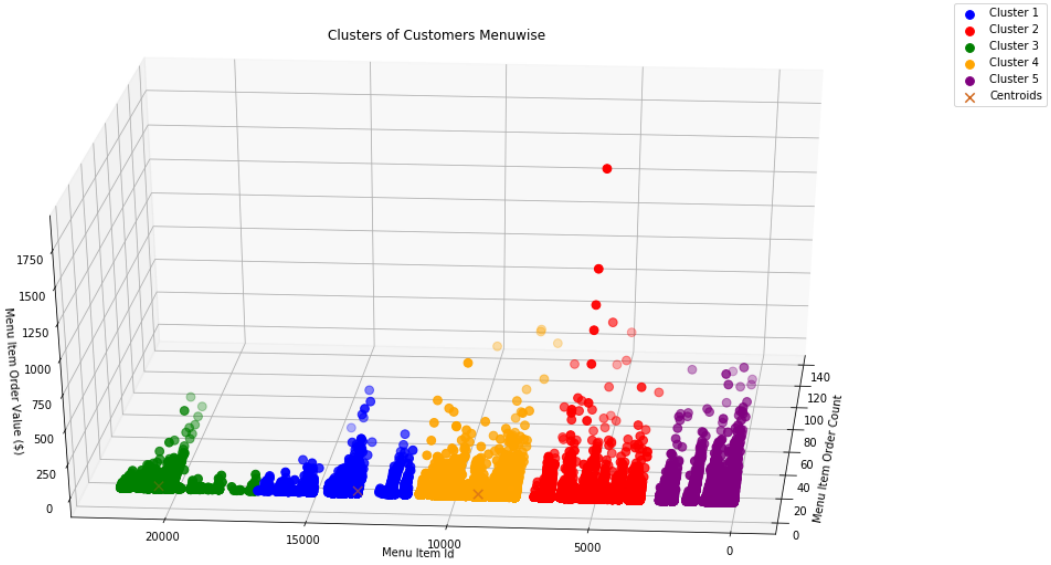
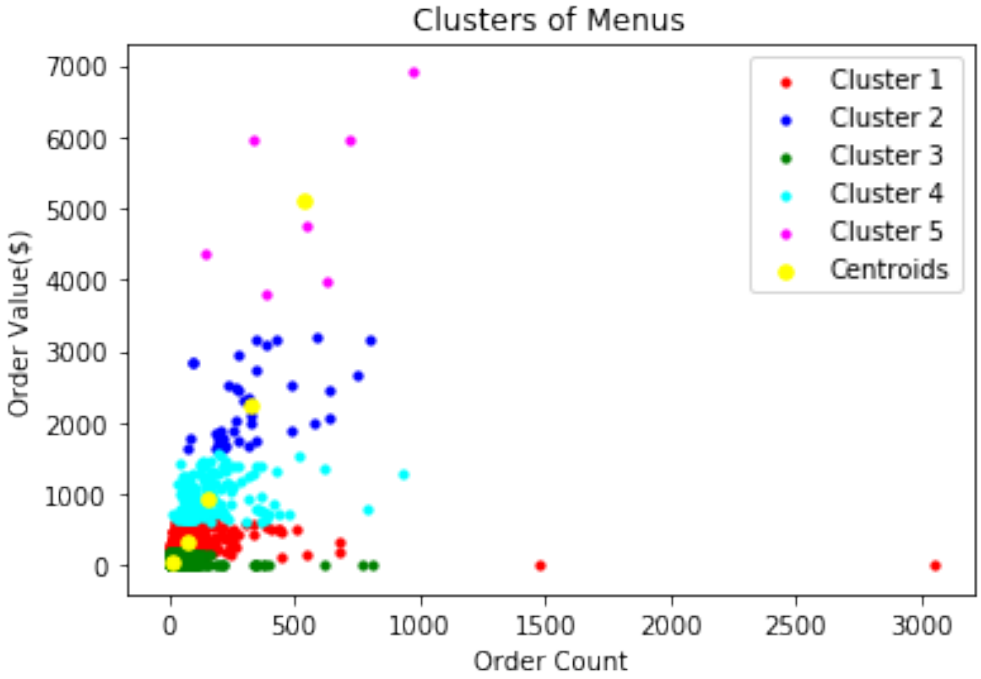
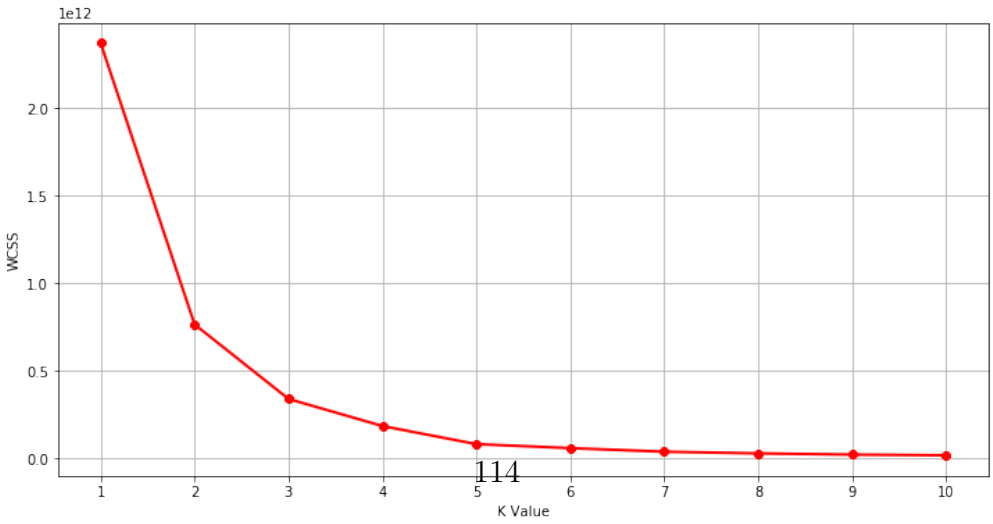


Figure 6.28: K-means clustering for customers with segmentation analysis

Restaurant segmentation enables a company to customize its relationships with the restaurants. We make use of sales revenue generated by the restaurants to put the restaurants in different sales revenue categories like very low, low, medium, high, and very high. In the highest segment, there are $8 restaurants that contribute an order value of over $5,000. Most of the restaurants generate order values less than $2,000 and order count $<$ $150. Furthermore, Clustering of restaurants carries out a similar concept as that of clustering of customers (see Section 6.4.6) on the basis of *sales order count* and *sales order value*. In the case of restaurant clustering, we use sales data for each restaurant for the entire 3-year period and; thereafter, Figs. 6.27 and 6.28 plot two heatmaps for clusters starting from $0 to $4. The number of restaurants in the highest bracket is the fewest and vice versa.


Clusters of Customers Menuwise
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Centroids
Menu Item Order Value ($)
Menu Item Id
Menu Item Order Count


Clusters of Menus
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Centroids
Order Value($)
Order Count


1e12
WCSS
K Value

The clusters are divided using k-elbow governing how many k-means clusters to divide the entire dataset of restaurants. In Fig. 9.46, the list of restaurants is categorized into multiple tiers that are represented in the above heatmaps, which in turn show the probability of food being ordered from one restaurant in a particular tier if food is being ordered from another restaurant.

6.5 Summary

In this section, we share key findings from our study on the impact of COVID-19 on businesses. While most businesses faced challenges during the pandemic, there were exceptions where revenues actually thrived. The pandemic instilled fear in customers, driving them to seek essential services while staying home due to lock-downs. As such, increased demand benefited the company, which not only delivered food but also groceries. In areas with limited competition, the company enjoyed a near-monopoly, thriving in this critical situation, highlighting how adverse conditions can benefit certain businesses. An in-depth analysis revealed that a significant number of orders concentrated between 5 to 7 PM. To meet the demand for quick and timely deliveries, effective workforce management during peak hours was crucial. Adjusting workforce deployment between peak and lean hours reduced costs and idle time, improving financial performance and employee satisfaction. Understanding these intra-day fluctuations is crucial for optimization. Identifying key contributing restaurants was essential to maintain performance, with further analysis focusing on menu items, location, and proximity to customers. Understanding both high-performing and underperforming restaurants provided insights for business improvement within the existing framework. Our study also identified menu items driving high revenue. Despite lacking customer profiles, this finding guided menu optimization and replication of successful items in other locations or restaurants.

The student community from a local university drove significant demand. The company can enhance this customer base through improved services, discounts, and loyalty programs.

Conducting surveys within this segment can uncover insights for business enhancement. Recognizing customers placing high-value orders, the company can offer special privileges and prioritize their treatment. Engaging a customer relationship manager in regular interactions can enhance satisfaction and loyalty. Market Basket Analysis (MBA) identified item combinations often ordered together. This information can simplify the ordering process and boost sales, particularly when taking phone orders. It enhances customer satisfaction and retention. In summary, our study reveals the diverse impacts of COVID-19 on businesses, showcasing a unique case of pandemic-driven success. Insights on workforce management, restaurant performance, menu items, customer segments, and optimization strategies offer guidance for data-informed decision-making in dynamic circumstances.

Chapter 7

Discussions

Developing a methodology to establish a relationship between COVID-19 and Long COVID was challenging. The patient-disease bipartite network followed by patient-patient and disease-disease projections were realized in a straightforward manner, but, it was challenging to figure out the methodology to associate COVID-19 patients with Long COVID patients. The usual methodology is to develop some correlation between the COVID-19 and Long COVID cases. The problem with such a methodology is that it cannot be easily employed in network analytics such as community detection or machine learning methods for prediction. As a result, taking clues from the bipartite network for patient-disease, we create a methodology to develop a bipartite network from the network of diseases containing both COVID-19 and Long COVID patients. This part has been dealt with in detail in Section 4.1.1. This paved the way for further progress in the research.

We conducted this research on Palantir's platform which is supported by the NIH. This platform is big data that supports PySpark and SQL. The size of the database is huge and executing standard Python-based code is difficult. As a result, we had to convert a number of Python programs into appropriate SQL queries in multiple steps. This aspect of the research was a big challenge.

There is an important fact that we want to place on record. When we started the research experiments, the database did not contain the ICD-10 code for Long COVID (U09). Therefore, we followed the definition by the CDC. Our initial number of Long COVID patients was much larger but in due course, NIH assigned the code, U09, to the patient to clearly mark them as Long COVID patients. Since it was not practically feasible to depart

from the data provided by the NIH, we modified our program to incorporate such a change and continued our experiments.

We faced the problem of the imbalanced dataset for the machine learning part of the research because, in both adult and pediatric patients, the number of Long COVID patients was fewer than 5%. For any machine learning method, such a high imbalance leads to poor prediction results. It is important to mention here that to tackle the imbalanced dataset problem, we tried to use SMOTE, but it did not improve the situation much. To overcome this issue, we extracted relevant information from network analytics and used the same for a) introducing additional features and b) dimensionality reduction. In the first step, we added almost 1,945 columns that were generated using Jaccard Coefficients. Because of a large number of feature vectors, our initial model ran into the problem of huge execution time and extremely poor results. We tried PCA (Principal Component Analysis) to reduce the dimensionality but again, this did not give us good results. At this juncture, we utilized the information obtained through community detection results. We reduced the number of columns to roughly 200. In spite of dimensionality reduction, the results were poor due to the imbalanced dataset. Since SMOTE too could not improve the situation, we reduced the imbalance of datasets by curtailing the number of records belonging to the majority group, i.e., Non-Long COVID.

After throwing light in the preceding paragraphs on some bottlenecks we encountered in the experiment, in the remaining part of this chapter we discuss the implications of our research. The theme of the research is the impact of COVID-19 on (a) adult patients leading to Long COVID, (b) pediatric patients leading to Long COVID, and (b) food delivery. For the first and second parts, we used NIH data made available to us through Palantir in the form of Big Data with the PySpark interface. For the third part, data was made available through a local food delivery company. Palantir's framework for NIH data collected from various hospitals across the entire US presents a broad spectrum of COVID cases including

Long COVID cases. Since the database is very rich and is not confined to a small geographical region or section of people, it is free from geographical or demographic biases.

7.1 Long COVID on Adult Patients

A novel method of establishing relationships has been used in this paper. Bipartite network has been used at different stages. In the first stage, it has been used to create a patient-disease network. In the next stage, we identify Pre-COVID and Post-COVID nodes. We create another bipartite network by removing edges that connect the nodes within the set of Pre-COVID diseases or Post-COVID diseases. In addition, we split the disease-disease network into two sub-graphs – one containing only the nodes and edges that exclusively connect Pre-COVID diseases and the other only the nodes and edges that exclusively connect Post-COVID diseases. These two sub-graphs have been used for community detection in terms of diseases. The method adopted in this paper can be applied in many situations involving network analytics/bipartite networks. The methodology used can be applied in any network where two identifiable groups are to be associated. This can be done by manipulating the nodes and edges as explained in this paper. The relationship obtained through different analyses can be further probed by medical experts. The results can provide them a good starting point as well as clues for delving deeper into Post-COVID.

The results of community detection reveal some important aspects:

1. The communities are cohesive and thematic because diseases have some connecting thread such as multi-organ based of the same system such as digestive system or infectious diseases

2. There are no COVID related diseases in Pre-COVID communities.

3. We see U09 and U07 in the Post-COVID communities. It seems that U09 being new some hospitals are still using U07 instead of U09 in case of Post-COVID. Correction in such records would lead to further improvement in our results.

Besides network analytics, we carried out an experiment for the prediction of Long COVID for adult patients using LSTM and Neural Networks. We fixed the teething issues as mentioned in the initial paragraphs of this chapter. Accuracy results for test datasets are quite stable and do not alter with changes in learning rate and batch size, especially, in the case of Neural Networks. This is also to be noted that we applied SMOTE but that does not help much as mentioned earlier.

Changing the number of non-Long COVID cases or Long COVID cases made the maximum difference and, undoubtedly, the model is pretty sensitive to these parameters. When we go for a large number of non-Long COVID cases, the imbalance increases, and that leads the model to train and predict very poorly. We had had approx. 25 thousand Long COVID adult patients. Even with about 500,000 COVID-19 patients, the model is able to achieve around 90% for accuracy and F1 score.

For training of the model and tuning of hyperparameters, we used 500k records.

As mentioned earlier, we used dimensionality reduction by slashing the number of columns. After reaching such a stage, the addition in a number of columns by even 100% had a slight impact on accuracy and accuracy never dipped below 95%.

For our experiments, we found that the model converged within 30 epochs and, therefore, we used 30 epochs.

7.2 Long COVID on Pediatric Patients

The results of community detection reveal some important aspects:

1. The communities are cohesive and thematic because diseases have some connecting thread such as multi-organ based of the same system such as digestive system or infectious diseases

2. There are no COVID-related diseases in Pre-COVID communities.

3. We see U09 and U07 in the Post-COVID communities. It seems that U09 being new some hospitals are still using U07 instead of U09 in case of Post-COVID. Correction in such records would lead to further improvement in our results.

Besides network analytics, as in the case of adult patients, we carried out an experiment for the prediction of Long COVID for pediatric patients using LSTM and Neural Networks. Once again, accuracy results for test datasets are quite stable and do not alter with changes in learning rate and batch size, especially, in the case of Neural Network but variation was more compared to those in the case of adult patients. We had a similar experience with SMOTE as mentioned in the case of adult patients.

In the case of pediatric patients as well, changing the number of non-Long COVID cases or Long COVID cases made the maximum difference and, undoubtedly, this model is also pretty sensitive to these parameters. When we go for a large number of non-Long COVID cases, the imbalance increases, and that leads the model to train and predict very poorly. We had 1,101 Long COVID pediatric patients. With about 10,000 COVID-19 patients, the model is able to achieve reasonably good accuracy and F1 score. For training the model and tuning of hyperparameters, we used 10,000 records. As mentioned earlier, we used dimensionality reduction by slashing the number of columns. After reaching such a stage, as in the case of adult patients, the addition of a number of columns had a slight to moderate impact on accuracy. We found 'tanh' as the best activation function, especially for the Neural Network model. Changing the activation function to 'sigmoid' or 'ReLU' did not improve the performance of the model. For our experiments, we found that the model converged within 100 epochs and, therefore, we used 100 epochs.

7.3 COVID-19 Impact on Food Delivery

We now discuss the interpretation of the results and how it can affect the people in the related fields. We first look at the viewpoint of the researchers. Researchers will attempt to comprehend the causation, patterns, and consequences of the entire process to understand

how such research can provide helpful output in their work. Thereafter, we delve into the perspectives of business practitioners, who are dealing with the pandemic and trying a large variety of measures to adapt to this global phenomenon.

Finding the factors that influence the revenue of a food delivery company during the Pre-COVID-19 and COVID-19 periods is vital for any similar business concern to get a handle on how to launch and operate a business in an efficient and more profitable manner. We investigate the factors that are essentially derived from the demand side, which is dependent on key customers and their ordering patterns. We consider these aspects by looking at the socio-economic dynamics in the local environment of the company. These factors vary from place to place and from company to company; nevertheless, the core idea behind the methodology of finding factors remains unchanged. The food delivery company provides a software-based platform to bring all the local restaurants on one platform for ordering food items from the customers. The ordered items are picked up by the delivery person of the food delivery company and delivered to the customers. A single restaurant may have insufficient sales to provide such a feature, but a food delivery company is able to afford such a facility by aggregating several restaurants as far as ordering and delivery are concerned.

In our study, we focus on orders by extracting data pertaining to the sales revenue generated by each order, ordered menu items, delivery zone of each order, and restaurants supplying the order in each order since the inception of the company. To preserve the data privacy of the customers, we ignore the other parameters in the dataset offered by the company. One of the techniques applied in this research is a temporal variation of the sales value of the company. One of the findings is that there was no decline the sales revenue during the COVID-19 period in contrast to the usual intuition that it should have gone down. One candidate reason is that during COVID-19, people remain locked down and depend more on the food delivery company. This trend maintains the sales almost at the same pre-COVID-19 level in spite of an adverse economic situation. Another technique we deploy is correlation-based heat maps. We use this technique to plot heat maps for

hourly sales revenue and the company's daily sales revenue, delivery zone-wise daily sales revenue and the company's daily sales revenue, and restaurant-wise daily sales revenue and the company's daily sales revenue. These heat maps unveil intriguing findings. It is noticed that the highest revenue-generating hourly slot is 6 p.m. to 7 p.m. The delivery zones with dense populations or student-based populations generated more revenue. About 10 to 20 percent of restaurants have major contributions to the revenue of the food delivery company.

Besides the factors that have significant results, the other factors that do not show any impact are also worth mentioning. No seasonality pattern is discovered in terms of weather or weekdays in the sales. Similarly, there is a lack of a clear relationship between sales and weather data in terms of temperature or rainfall. A factor that makes remarkable impacts on sales is the arrival and departure of the student community. This conspicuous correlation is reasonable because the local town is based on the university and student community - a major driver of the demand.

The food delivery company provides a software-based platform to bring all the local restaurants on one platform for ordering food items from customers. The ordered items are picked up by the delivery person of the food delivery company and delivered to the customers. A single restaurant may not have enough sales to warrant such a feature, but a food delivery company is able to afford such a facility by aggregating several restaurants as far as ordering and delivery are concerned. From a customer's point of view, the software platform offers a wide variety of food options to choose from and saves time by getting it delivered to the place of stay without incurring any substantial extra cost.

In our study, we extract data regarding the sales revenue generated by each order, ordered menu items, delivery zone of each order, and restaurants supplying the order in each order since the inception of the company. One of the techniques we use is the variation of the sales value of the company with time. One of the findings is that there was no decline the sales revenue during the COVID-19 period in contrast to the usual intuition that it should have gone down. We also examine how hourly sales revenue, delivery zone-wise daily sales,

and restaurant-wise daily sales revenue are related to the company's daily sales revenue. Our analysis reveals that the highest revenue-generating hourly slot is 6 PM to 7 PM, delivery zones with more population or student-based population generated more revenue, and about 10 to 20 percent of restaurants had a major contribution to the revenue of the food delivery company.

Besides the factors that have significant noticeable results, other factors that do not show any impact are also worth mentioning. For example, we could not find any clear relationship with weather either in terms of temperature or rainfall. A remarkable factor is the arrival and departure of the student community because the student community in the university town is the major driver of the demand. These results can be used by the company to reallocate its resources, carry out promotions to attract different segments of customers, enhance the order value from the existing customer base, and focus more on restaurants that are contributing more to the company's revenue.

7.4 Summary

Our study encountered challenges in establishing a methodology to link COVID-19 and Long COVID patients. To address this, we created bipartite networks and used a novel approach. They utilized a big data platform, Palantir, supported by the NIH, which required transforming Python code into SQL due to the massive database size. Initially, there was no ICD-10 code for Long COVID (U09), so the CDC's definition was used. Later, U09 codes were assigned, and the program was adapted accordingly.

Dealing with imbalanced datasets for machine learning was a challenge, especially for Long COVID patients, which comprised less than 5%. Traditional methods like SMOTE were ineffective. To address this, the study used network analytics to extract additional features and reduce dimensionality, ultimately curbing the imbalance by reducing the number of non-Long COVID records. The research aimed to understand the relationship between COVID-19, Long COVID, and food delivery. A unique method using bipartite networks was

applied, allowing the association of two distinct groups. This methodology can be used in various network analytics contexts. Community detection revealed cohesive disease groups and a need for standardizing codes. The study also delved into machine learning, using LSTM and Neural Networks for Long COVID prediction in adult and pediatric patients. Results were stable, but sensitivity to the number of non-Long COVID cases was noted. Ultimately, the model performed well with sufficient data.

Our research explored the impact of COVID-19 on adult and pediatric Long COVID patients and the food delivery industry. The analysis and observations provided insights into factors affecting food delivery sales, such as delivery zones, restaurant contributions, and student community dynamics. Despite the absence of weather-related patterns, the student community played a significant role in driving demand. These findings can guide resource allocation and promotion strategies for the food delivery company.

Chapter 8

Conclusion

We finally conclude this research and we discuss the major key points that have been brought to light during our analysis. We first discuss the Long COVID impact on adults and pediatrics. This includes the discussion of our methods and the main diseases that have been most prevalent in our findings. We then look into the impacts of COVID-19 on the local food delivery sector and see the multiple aspects, such as menu items, time and frequency of order, location, etc that have been affected.

8.1 Long COVID Impact on Adult Patients

This study helps in identifying and visualizing disease networks needed for understanding key disease clusters that could be highly associated with post-COVID. This study also identifies several condition occurrences/diseases that could reveal risk stratification for post-COVID patients and the basis for developing future prediction models or recommendation systems. For example, this study demonstrates, that post-COVID male patients with hypertension and disorders of lipoprotein metabolism have a strong likelihood of developing sleep disorders followed by breathing abnormalities, gastroesophageal reflux, and pain. Another key finding suggests that post-COVID female patients with hypertension and disorders of lipoprotein metabolism have a strong likelihood of developing sleep disorders followed by pain, soft tissue disorder, gastro-esophageal reflux, breathing abnormality, and sleep disorder. These results distinctly demonstrate that post-COVID disease/symptom's chances of male and female patients are different. The network analysis of the Louvain community detection method recognizes different types of disease clusters that are associated with key diseases/symptoms for both post-COVID and post-COVID diseases/symptoms in clusters

related either on an organ basis (e.g., kidney) or similar type of events (e.g., cardiac arrest, etc.).

Future studies could overcome the drawbacks of the approaches used in this study. While carrying out network analysis with many nodes, we notice that the cluster formation of nodes is relatively dense using the Louvain community detection method. This modularity-based method does not necessarily guarantee that the detected communities remain well-connected. Future work could compare the performance of Louvain with other community detection algorithms such as the Leiden algorithm for improved and well-connected detection of diseases among clusters.

8.2 Long COVID Impact on Pediatric Patients

Communities in disease networks exhibited cohesiveness and were thematic. Diseases within a community shared common characteristics, such as multi-organ involvement or affiliation with the same system (e.g., the digestive system or infectious diseases). Notably, COVID-related diseases were not present in Pre-COVID communities, which may indicate a clear distinction in disease patterns before and after the pandemic's advent. In Post-COVID communities, the presence of both U09 and U07 codes suggested that hospitals continue to use U07 for Post-COVID cases as U09 was new and just introduced by the CDC. Correcting this inconsistency in records could enhance the accuracy and reliability of our results.

Our experiments in Long COVID prediction for pediatric patients, utilizing LSTM and Neural Network models, resulted in stable accuracy results for test datasets. Notably, the Neural Network exhibited consistency even when varying learning rates and batch sizes, though there was slightly more variation compared to adult patients. We encountered similar experiences with Synthetic Minority Over-sampling Technique (SMOTE) as observed in adult patients. The model's sensitivity to the balance between non-Long COVID and Long COVID cases in pediatric patients was evident. Increasing the number of non-Long COVID cases significantly impacted the model's performance, leading to poor training and

prediction. We observed successful results when working with a dataset consisting of 1,101 Long COVID pediatric patients and about 10,000 COVID-19 patients. Our model training and hyperparameter tuning were conducted with a dataset of 10,000 records, allowing us to achieve meaningful results. Dimensionality reduction proved effective, and additional columns had only a slight to moderate impact on accuracy.

We identified the 'tanh' activation function as the optimal choice, particularly for the Neural Network model. Altering the activation function to 'sigmoid' or 'ReLU' did not yield performance improvements. Our experiments indicated that the model consistently converged within 100 epochs, leading us to adopt this number for future experiments. In summary, our research findings provide valuable insights into the structure of disease communities and the prediction of Long COVID in pediatric patients. These insights have the potential to inform medical decision-making and improve the quality of healthcare in the context of both disease detection and prediction. Further research and correction of inconsistencies in medical coding practices can enhance the reliability and accuracy of our models.

8.3 COVID-19 Impact on Food Delivery Business

In this section, we discuss the key takeaways and conclusions from our study on how COVID-19 has affected businesses across the globe to varying degrees. Most businesses have been adversely affected but COVID-19 had almost no impact on the company rather it boosted the revenue. The fear among the customers pushed the demand up for the company because the customers preferred or were forced to stay locked in their homes. Lockdowns and restrictions on movements spurred the activities of the company. It gained importance in the community because it supplied not only food but also groceries. Grocery is a necessity that no one can live without. There being no major company in this business in the town, it was almost a monopoly for the company and it prospered during such a critical situation.

The key takeaway is that even adverse conditions like COVID-19 can be beneficial for some businesses.

Since the company provided delivery of food items from local restaurants, and people preferred dinner from restaurants, it had a high number of orders in the afternoon around 5 to 7 PM. Since the activity of the business demands quickest and timely delivery to all the customers, there is pressure for manpower during peak hours. This finding makes it clear that the company needs to manage its workforce in such a manner that peak-hour demands are served without any delays and meeting the expectations of the customers. At the same time, this also suggests that during lean hours manpower deployment may be reduced. This will enable the company to reduce the cost and at the same time slash idle time for its workforce. This will enhance the bottom line of the company and also improve the satisfaction level of its workforce. The key takeaway is that in this kind of business, there is a lot of intraday fluctuation and this needs to be understood and taken into account for the betterment of the company.

The main contributing restaurants were identified. The company can focus more on these restaurants to ensure that if things do not improve further then at least the status quo is maintained at any cost. Factors such as menu items, location, distance from the customers, etc. may be analyzed further to understand the reason for better performance. At the same time, the restaurants that are contributing poorly to the revenue of the company need to be analyzed. The analyses of both the good-performing and bad-performing restaurants will provide clues about the parameters, and factors that will guide how to improve the business in the future. This may involve decisions about business relationships as well. The key takeaway of this aspect is that there may be solutions for improving the performance within the existing system itself.

The menu item IDs that contribute to high revenue have been identified. This is a very significant finding. This tells us what appeals to the taste buds of the customers. We did not have information about the profiles of the customers otherwise we could have a

mapping of the community with the menu items. Yet, this finding can used to fine-tune the menu items in poor-performing restaurants and further improve in the good-performing restaurants after getting feedback from the customers. Top menu items can be replicated as far as practicable in other locations/restaurants. The takeaway from this finding is that in this type of business, menu item plays an important role and must be paid adequate attention for further improvement and expansion.

The major demand is driven by the student community of the local university. The company can further enhance it by introducing better services or discounts loyalty points or other means so that revenue is further improved and the customers are retained. Understanding the major customer base is important in shaping future business. A survey may be carried out with the student community to understand how to improve the business further and add more services. For instance, services like supplying stationery materials may be a good option to include. The key takeaway of this aspect is that such a finding allows the business to focus on the customers that matter and the company must make all efforts to retain and strengthen it.

Customers who place high-value orders have been identified. They should be placed in a privileged category by the company so that they are retained in the future. The high-value customers have to be dealt with special attention with priority. The company must ensure that such customers are served with the highest professionalism and meet their expectations. Going the extra mile will help in boosting the business of the company. The key takeaway of this point is that high-value customers have to be kept on a different footing. The company may engage a customer relationship manager to have regular interaction with such customers to ensure that they have no grievances about the services being rendered.

Market Basket Analysis (MBA) is another important finding that tells us the combination of the items usually ordered together. The menu items of all the restaurants may be modified accordingly so that customers find it easy to order. Moreover, if the order is placed on the phone, the operator can suggest items based on the findings of the MBA. This may

boost sales and save time leading to an increase in revenue. The key takeaway is that this finding is that it can help in serving the customers in a better way providing them with more satisfaction and improving the chances of their retention.

8.4 Future Work

In our analysis methods, we have used network analytics with machine learning/artificial models, like LSTM and Neural Networks. Our methodology is flexible with a rich set of plug-and-play features. As a future research direction, instead of LSTM or Neural Networks, we will deploy deep neural networks and a larger set of data to obtain stable results and, potentially, enhance F1 scores besides identifying the important feature vectors. Furthermore, we will apply our analysis methods for patient-patient projections rather than disease-disease projections. This process can show the patients with a more critical condition or those who are closer to being affected by Long COVID. Another methodology that we can use other community detection or even cluster algorithms. Making use of other algorithms that can yield cohesive communities like that of Louvain algorithms can help hospitals narrow down more diseases or cases related to Long COVID. Such communities will be used for a comparative analysis to see their accuracy and F1 scores when medical diagnoses for patients are conducted.

With our food delivery research, we faced a few limitations. If we use more parameters of the customers such as age, sex, occupation, etc., the findings may be more worthwhile. More food delivery companies operating in different locations and different local conditions both in terms of demographics and climate can give us more insight into the factors that influence sales of such companies. Moreover, the data used was only a single local company where we covered the sales in two town areas and we will be using data from multiple companies or sources. Such businesses can encompass a university area and that will lead us to delve into a comparative analysis. Such areas are also heavily impacted by the population of students, faculty, and university staff. Comparing this to a large city or metropolitan area will allow

us to demonstrate a significant difference in the type of sales, frequency, and amounts. This comparison can be considered in our future research where various food delivery companies manage a multitude of locations at hand.

8.5 Final Remarks

In the first part of the dissertation studies, We identified and visualized disease networks, which are crucial for understanding key disease clusters associated with post-COVID. These findings reveal patterns in disease occurrences that can inform risk stratification for post-COVID patients and serve as a basis for future prediction models or recommendation systems. Specific insights into post-COVID patients' disease likelihood are provided. For instance, male patients with hypertension and disorders of lipoprotein metabolism are more likely to develop sleep disorders, breathing abnormalities, gastroesophageal reflux, and pain. In contrast, female patients with the same conditions are more likely to experience sleep disorders, pain, soft tissue disorders, gastroesophageal reflux, breathing abnormalities, and further sleep disorders. The Louvain community detection method in network analysis identifies different types of disease clusters related to specific diseases or symptoms. These clusters are either based on organ systems or similar types of medical symptoms.

Disease communities in pediatric patients demonstrate cohesiveness and thematic relevance, with distinct differences between Pre-COVID and Post-COVID communities. The second part of the dissertation research highlights the sensitivity of predictive models to the balance between non-Long COVID and Long COVID cases in pediatric patients, emphasizing the importance of dataset composition. The selection of the 'tanh' activation function is identified as optimal for Neural Network models, with a consistent convergence within 100 epochs. We confirmed that these insights can improve medical decision-making and healthcare quality in terms of disease detection and prediction.

Moving on to the last part of the dissertation studies – the food delivery aspect of COVID-19, we concluded that the impact on the food delivery business was mostly positive,

with increased demand due to lock-downs and restrictions. The company benefited from this situation and became a crucial community resource, delivering not only food but also groceries. In terms of peak hours, this last study emphasizes the need for workforce management to handle high demand during specific time slots. Lean-hour workforce deployment optimization is also suggested to reduce costs and idle time. We identified top-performing and underperforming restaurants and menu items, offering insights into factors contributing to success or challenges. Understanding the major customer base, primarily driven by the local university's student community, provides opportunities for improving services and retaining customers. The identification of high-value customers and their special treatment is recommended to enhance customer retention. We discovered that the market basket analysis (MBA) findings offer a valuable tool for modifying menu items, suggesting additional items to customers, and improving service, ultimately leading to increased revenue and customer satisfaction.

In conclusion, this dissertation provides valuable insights for understanding disease networks, Long COVID prediction in pediatric patients, and the positive impact of COVID-19 on a food delivery business. These findings can inform decision-making, improve healthcare practices, and enhance business strategies for various sectors. Further research and analysis are encouraged to maximize the benefits of these insights.

Chapter 9

Appendix

9.1 Adult Long COVID

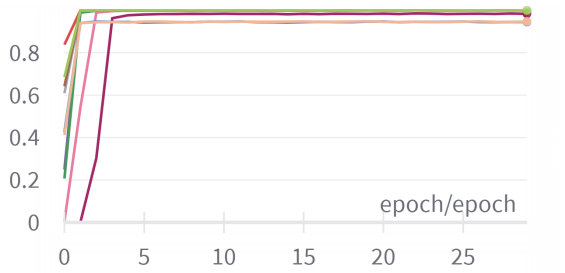


Figure 9.1: Adult LSTM F1 Score Training

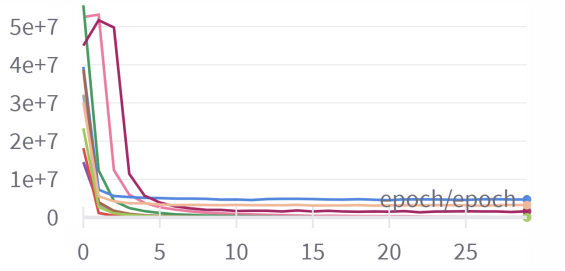


Figure 9.2: Adult LSTM MAPE Training

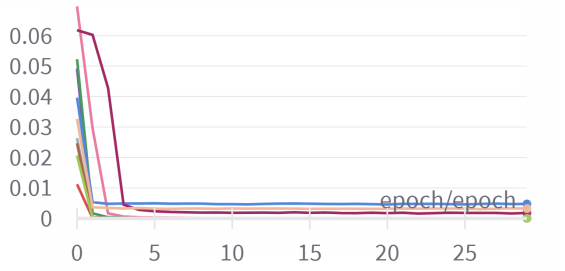


Figure 9.3: Adult LSTM MSE Training

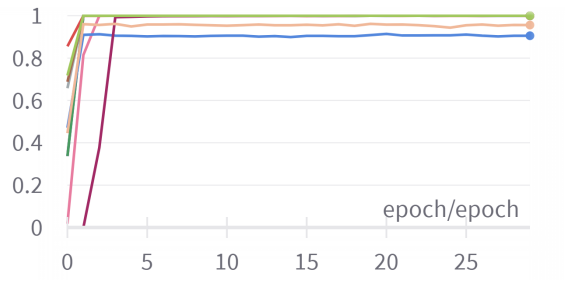


Figure 9.4: Adult LSTM Precision Training

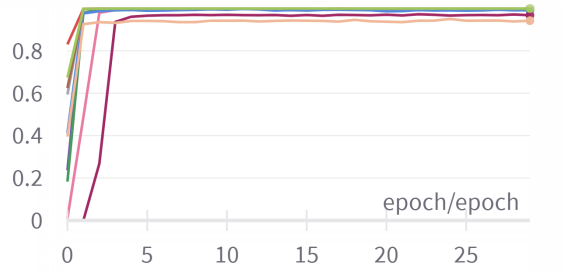


Figure 9.5: Adult LSTM Recall Training

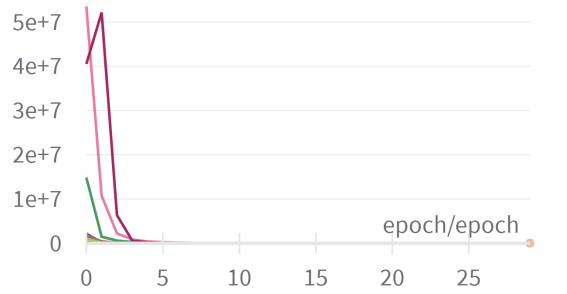


Figure 9.6: Adult LSTM MAPE Validation

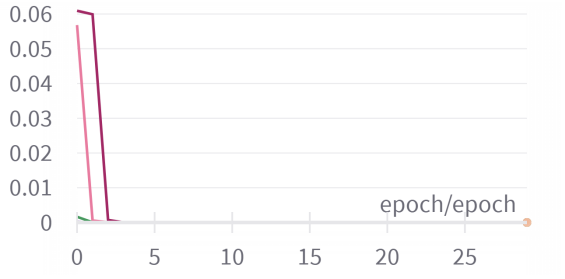


Figure 9.7: Adult LSTM MSE Validation

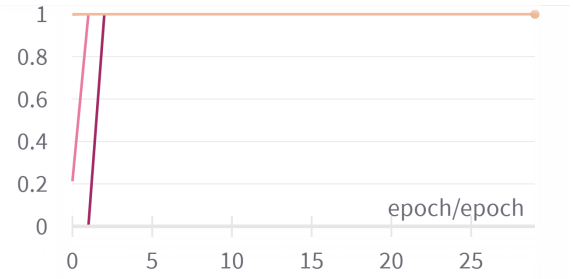


Figure 9.8: Adult LSTM Precision Validation

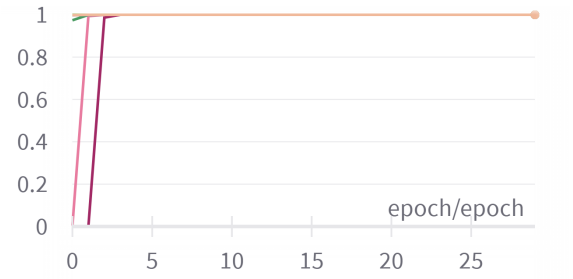


Figure 9.9: Adult LSTM Recall Validation

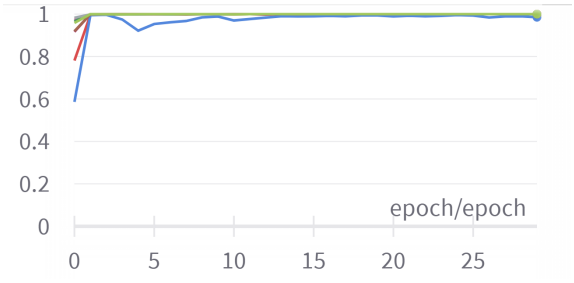


Figure 9.10: Adult Neural Network F1 Score Training

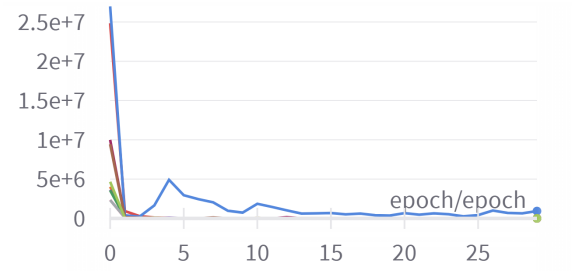


Figure 9.11: Adult Neural Network MAPE Training

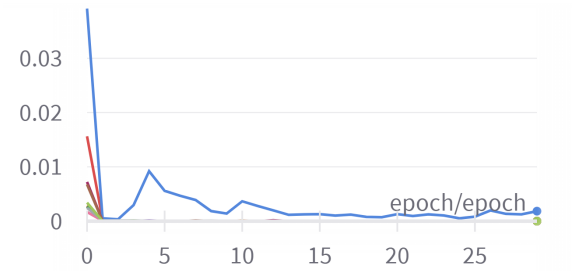


Figure 9.12: Adult Neural Network MSE Training

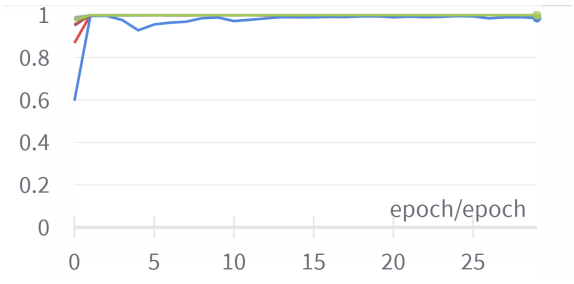


Figure 9.13: Adult Neural Network Precision Training

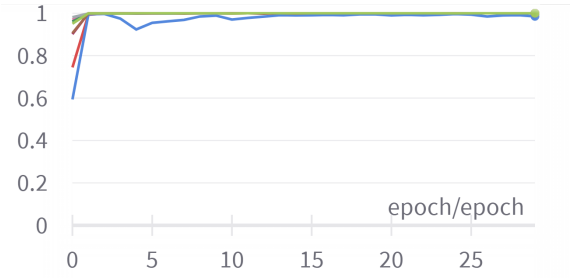


Figure 9.14: Adult Neural Network Recall Training

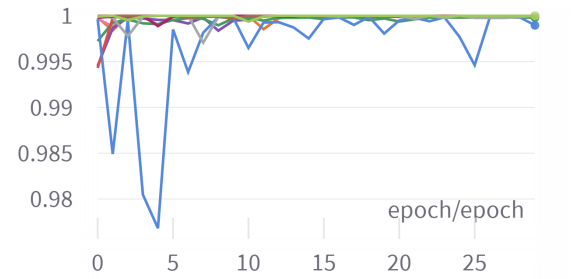


Figure 9.15: Adult Neural Network F1 Score Validation

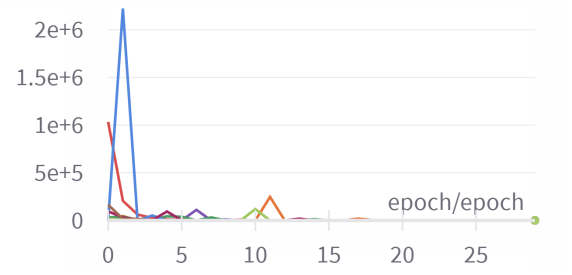


Figure 9.16: Adult Neural Network MAPE Validation

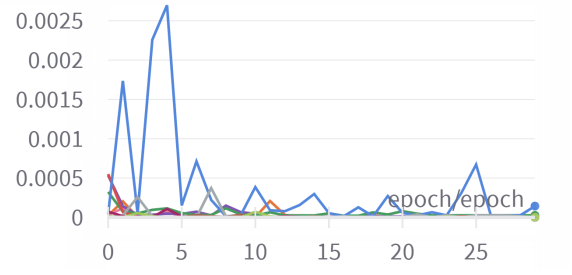


Figure 9.17: Adult Neural Network MSE Validation

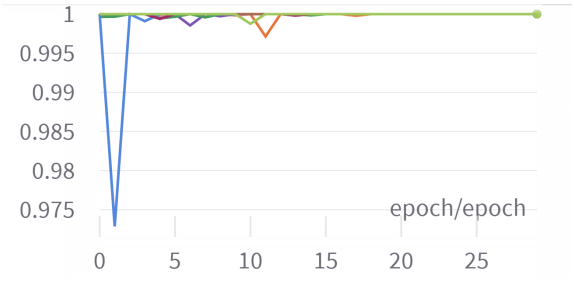


Figure 9.18: Adult Neural Network Precision Validation

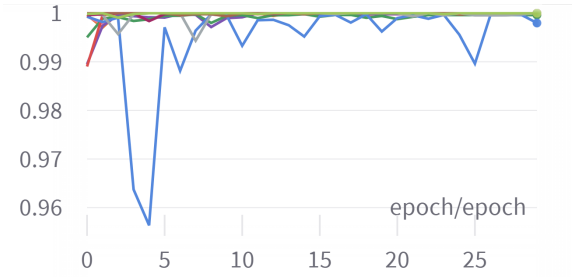


Figure 9.19: Adult Neural Network Recall Validation

9.2 Pediatric Long COVID

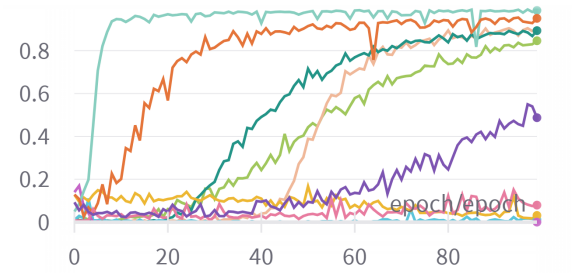


Figure 9.20: Pediatric LSTM F1 Score Training

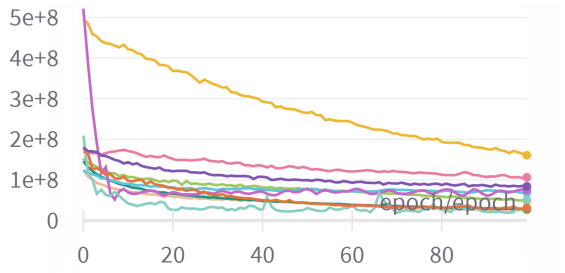


Figure 9.21: Pediatric LSTM MAPE Training

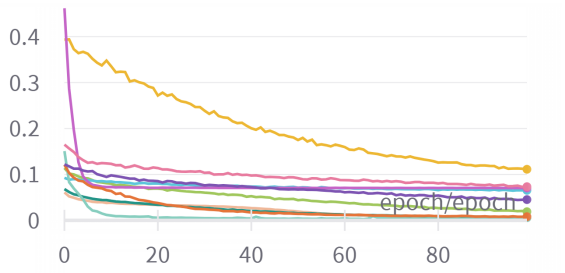


Figure 9.22: Pediatric LSTM MSE Training

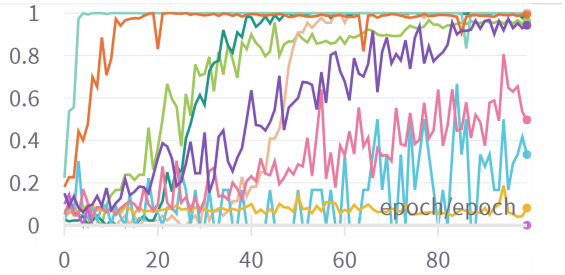


Figure 9.23: Pediatric LSTM Precision Training

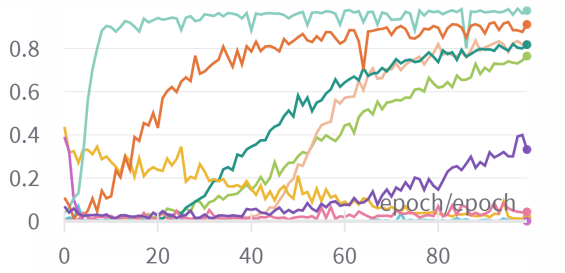


Figure 9.24: Pediatric LSTM Recall Training

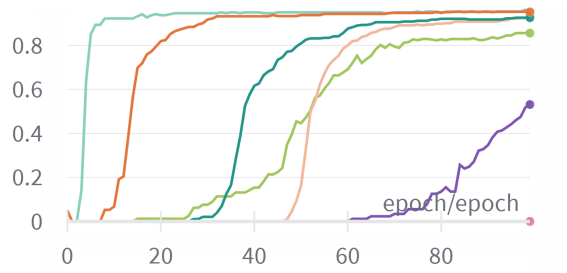


Figure 9.25: Pediatric LSTM F1 Score Validation

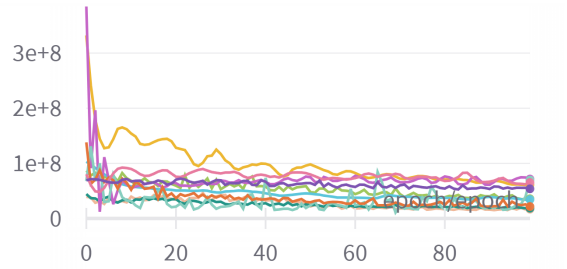


Figure 9.26: Pediatric LSTM MAPE Validation

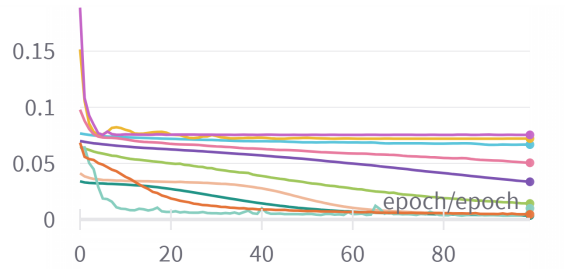


Figure 9.27: Pediatric LSTM MSE Validation

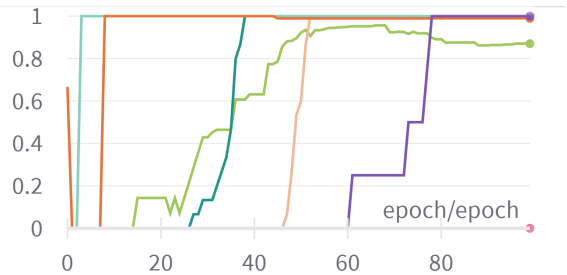


Figure 9.28: Pediatric LSTM Precision Validation

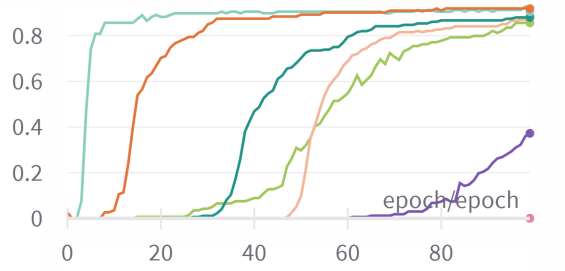


Figure 9.29: Pediatric LSTM Recall Validation

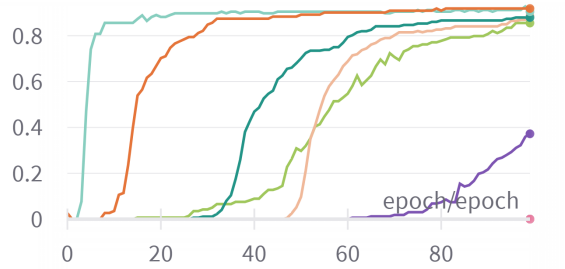


Figure 9.30: Pediatric LSTM Recall Validation

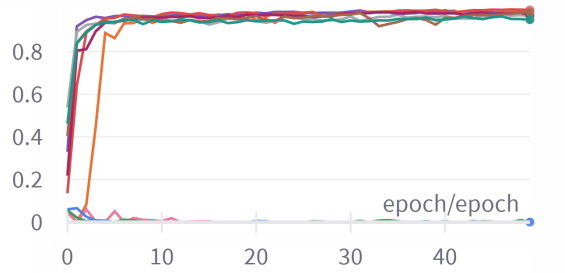


Figure 9.31: Pediatric Neural Network F1 Score Training

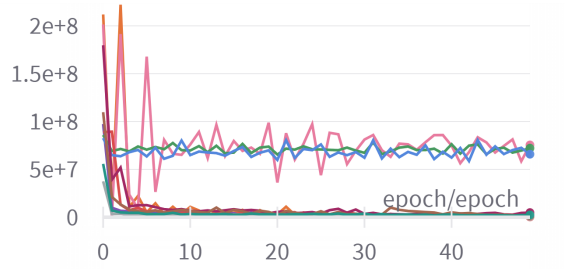


Figure 9.32: Pediatric Neural Network MAPE Training

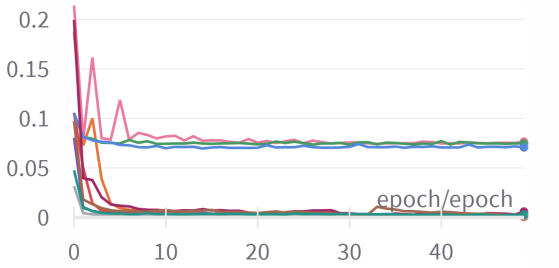


Figure 9.33: Pediatric Neural Network MSE Training

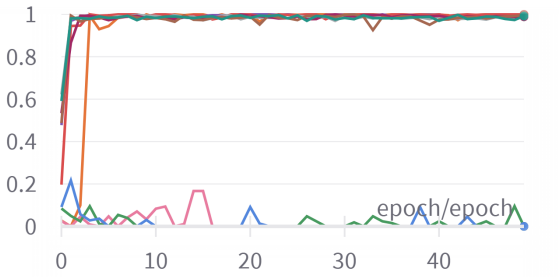


Figure 9.34: Pediatric Neural Network Precision Training

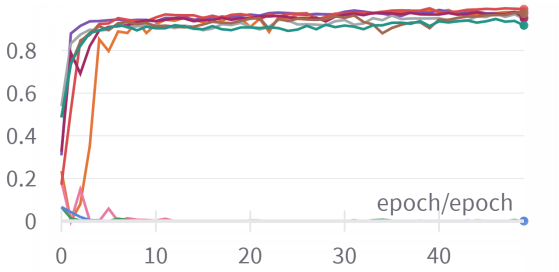


Figure 9.35: Pediatric Neural Network Recall Training

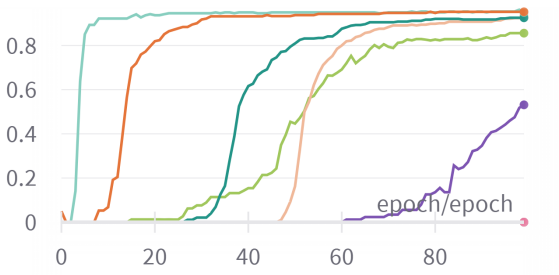


Figure 9.36: Pediatric Neural Network F1 Score Validation

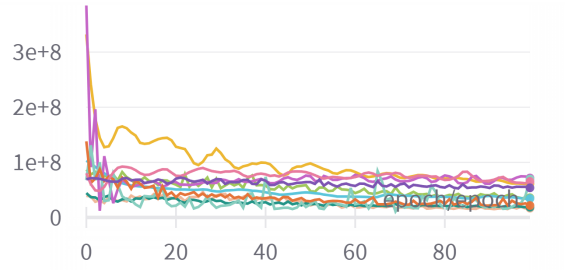


Figure 9.37: Pediatric Neural Network MAPE Validation

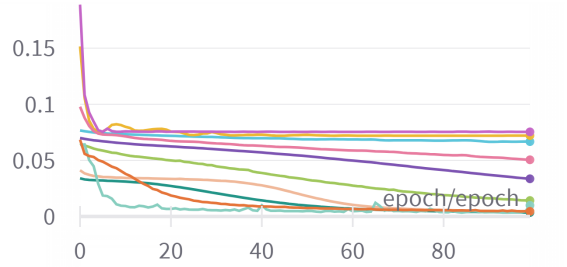


Figure 9.38: Pediatric Neural Network MSE Validation

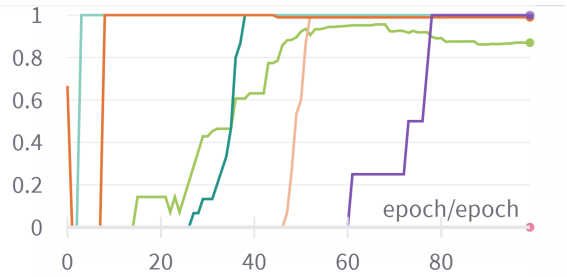


Figure 9.39: Pediatric Neural Network Precision Validation

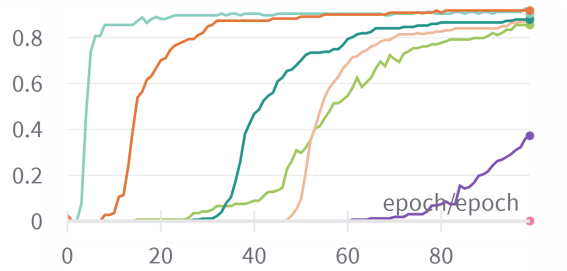


Figure 9.40: Pediatric Neural Network Recall Validation

9.3 Food Delivery

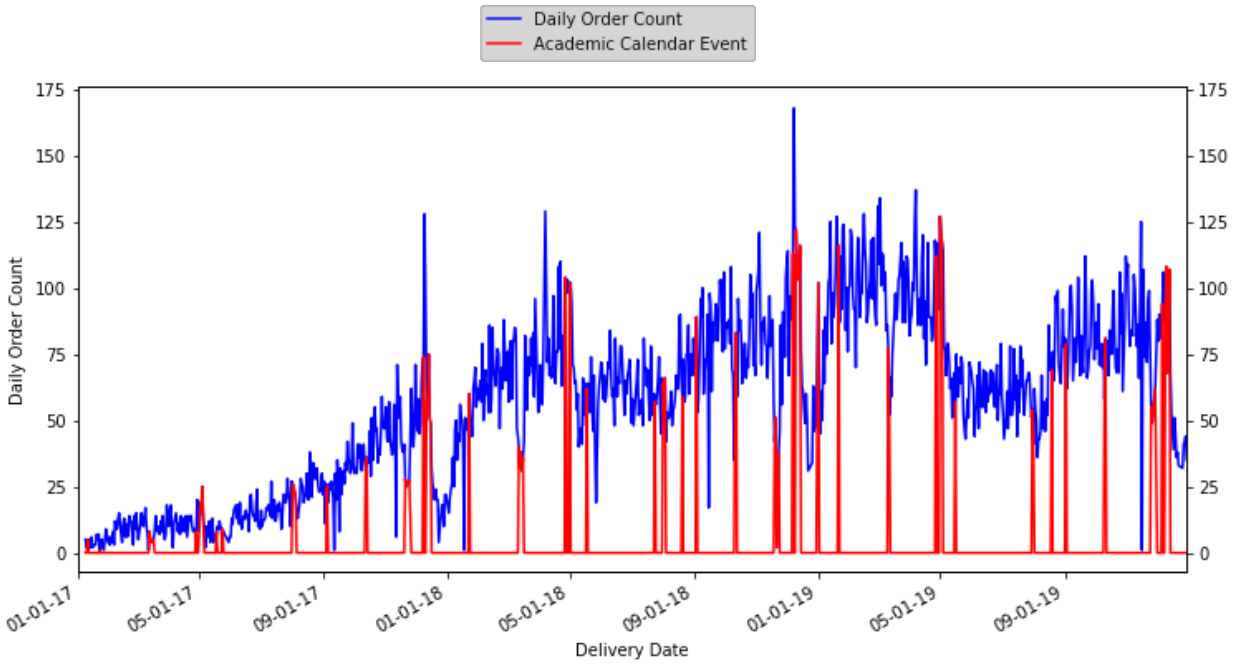


Figure 9.41: Academic Events Sales Count Generated Over Dataset

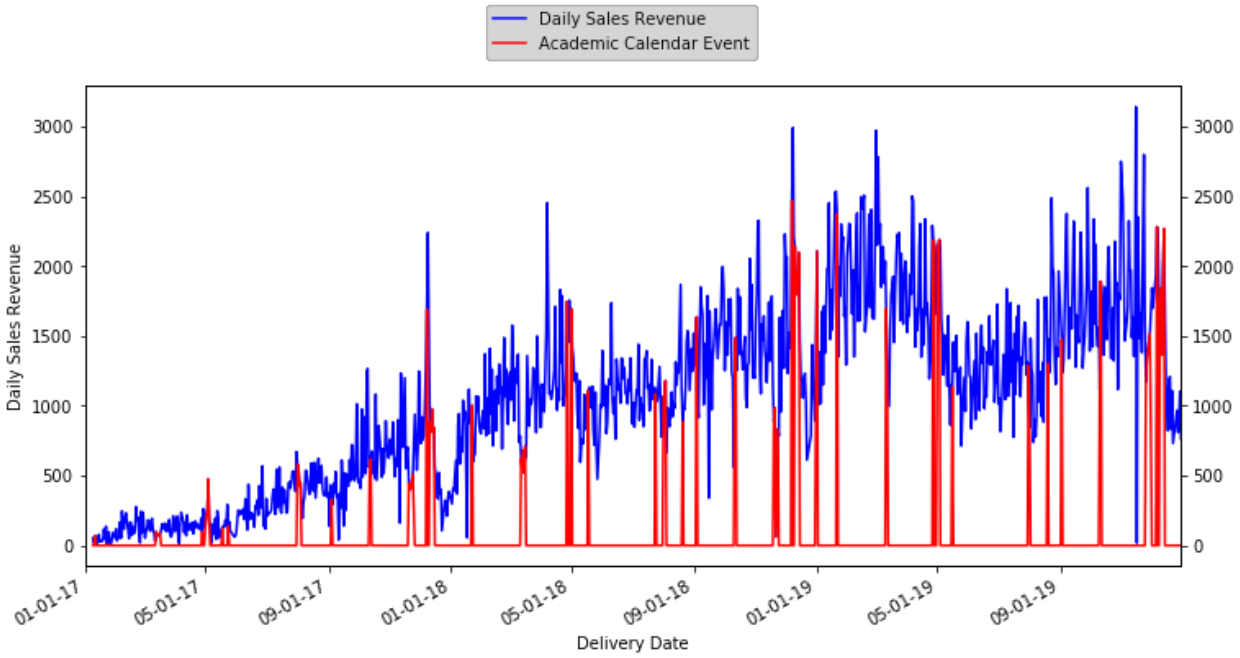
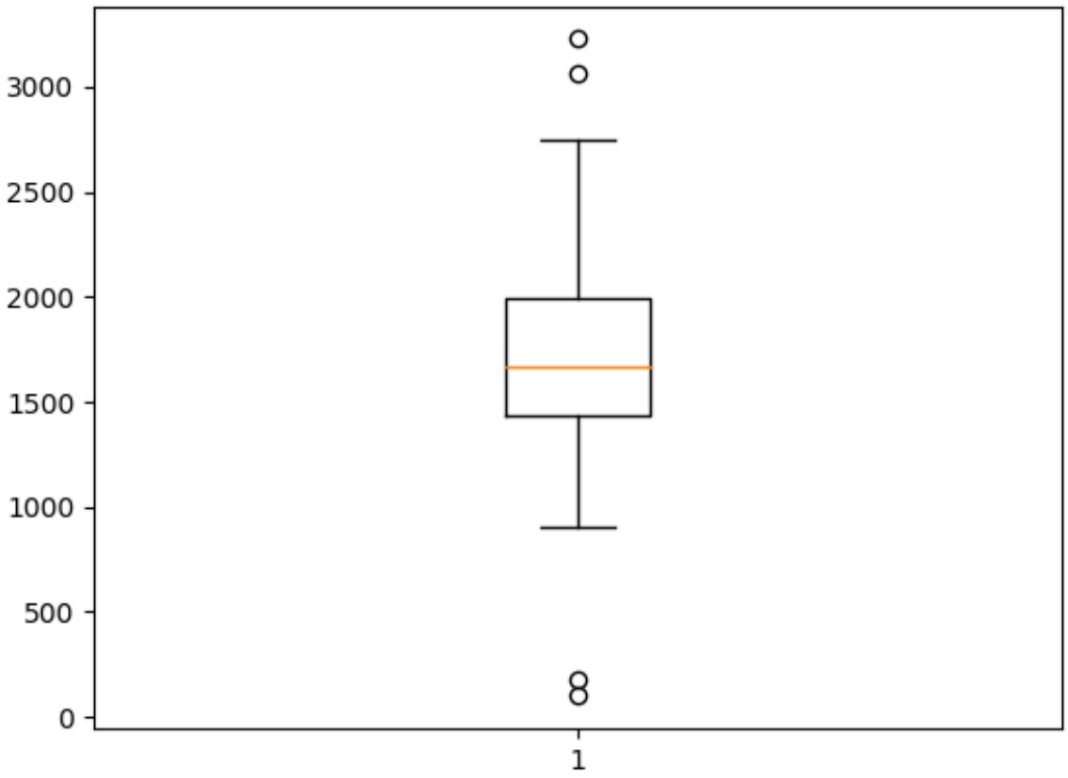


Figure 9.42: Academic Events Sales Revenue Generated Over Dataset



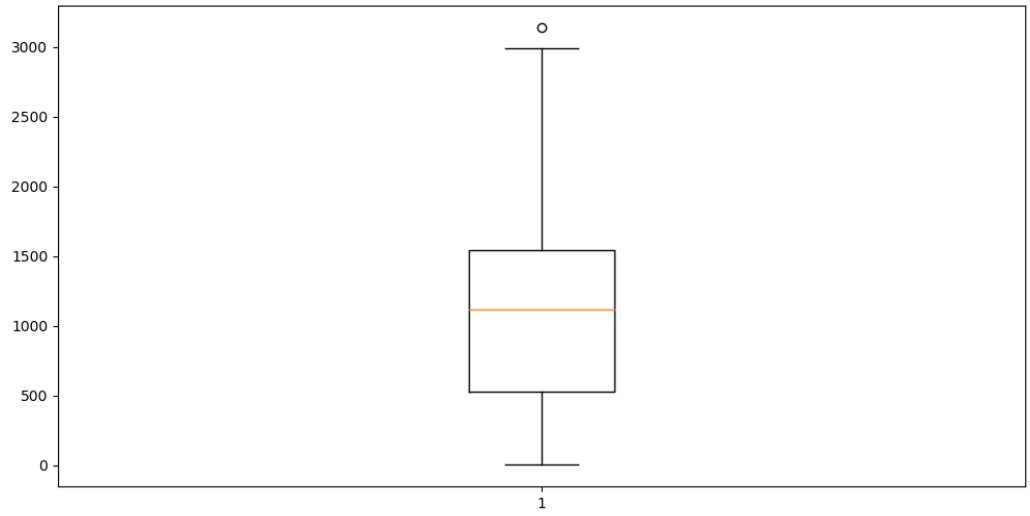


Figure 9.43: Box-Plot Sales Statistics from March 2020 to July 2020

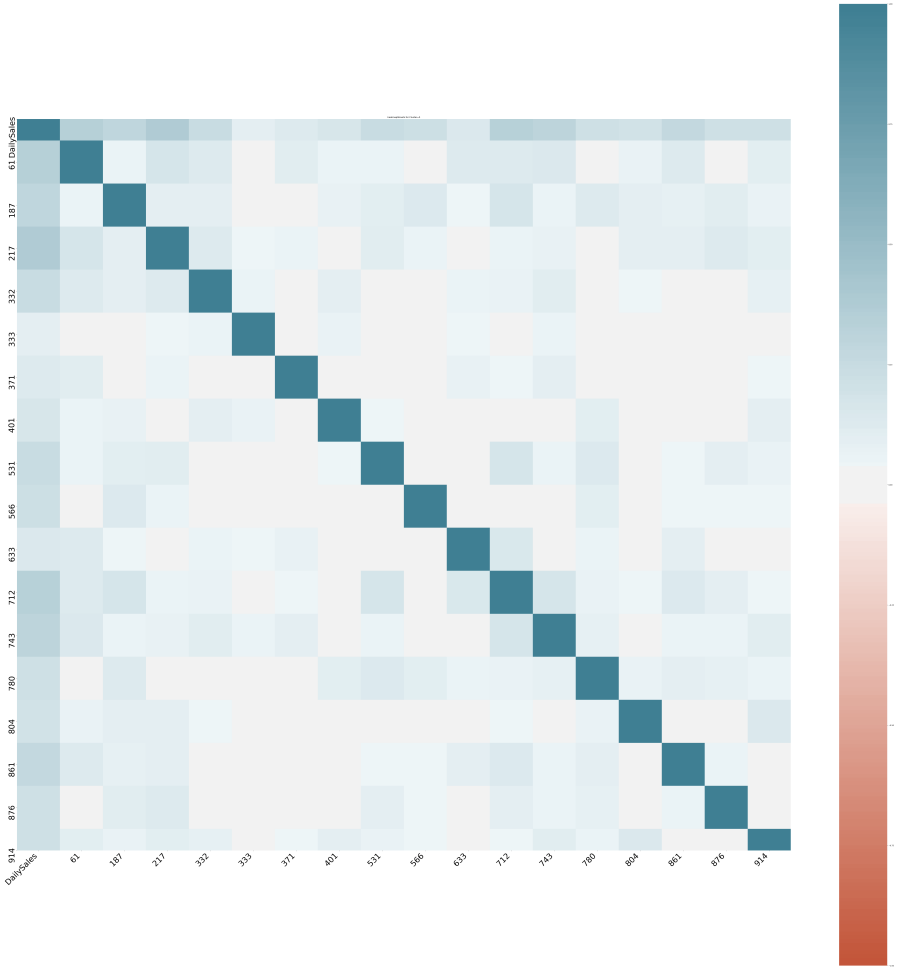


Figure 9.44: Delivery Zone ID StreetAnalysis

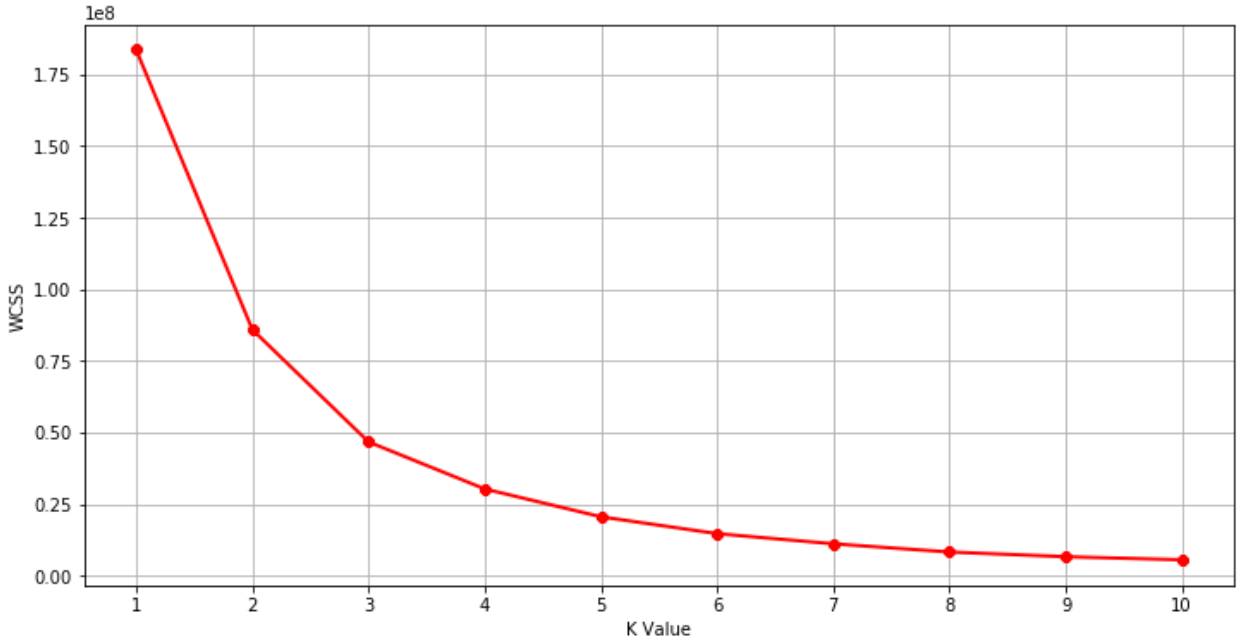


Figure 9.45: K-Elbow Method Cuisine Analysis

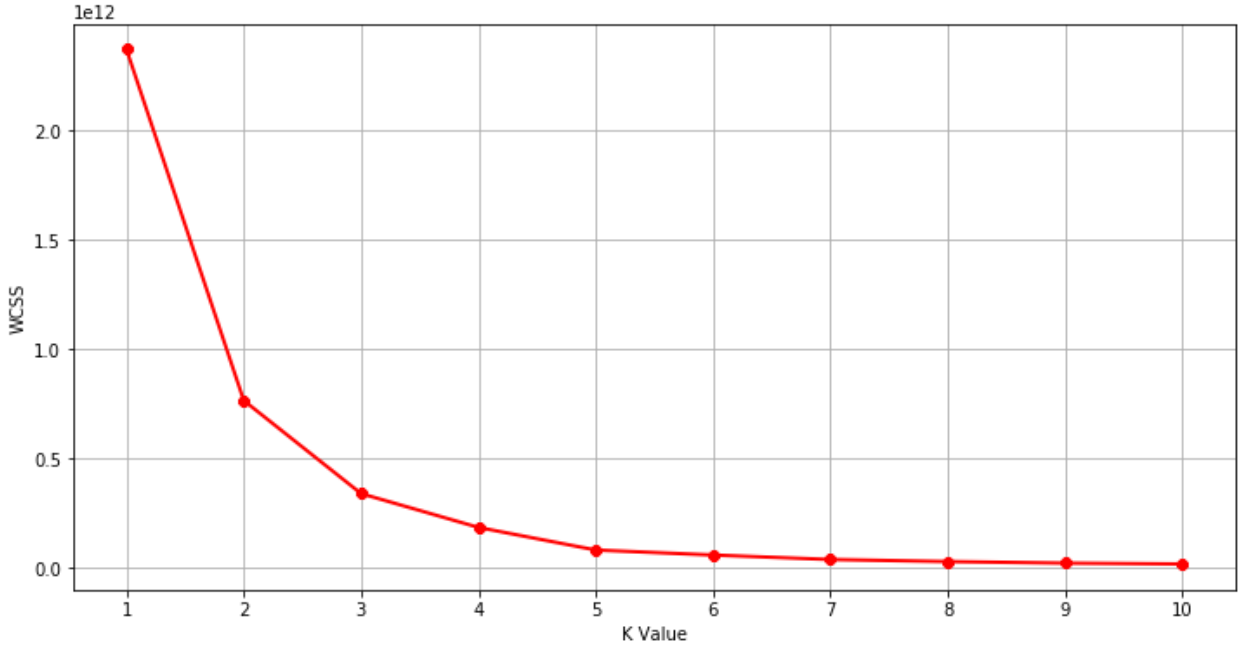


Figure 9.46: K-Elbow Method Menu Item Analysis

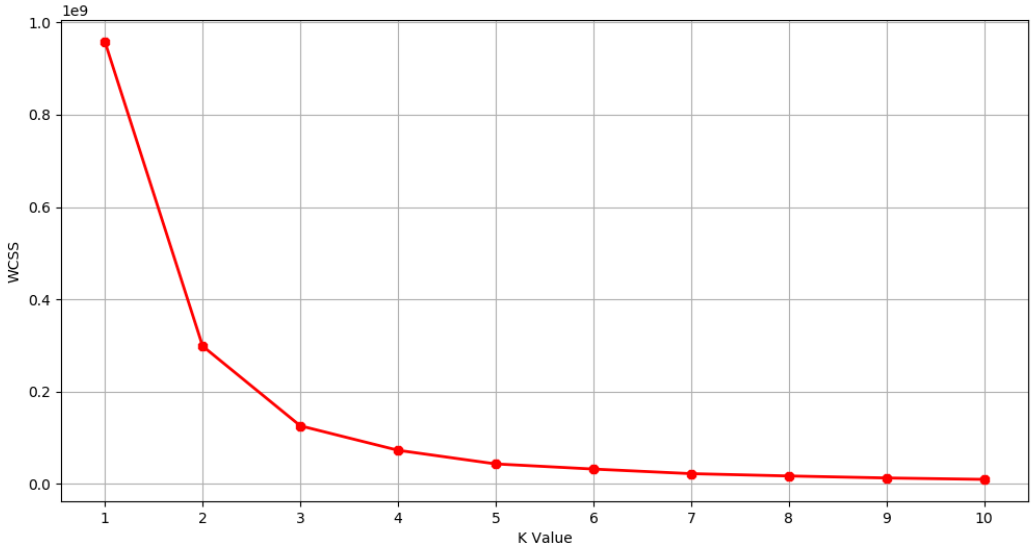


Figure 9.47: K-Elbow Method Street Clustering Analysis

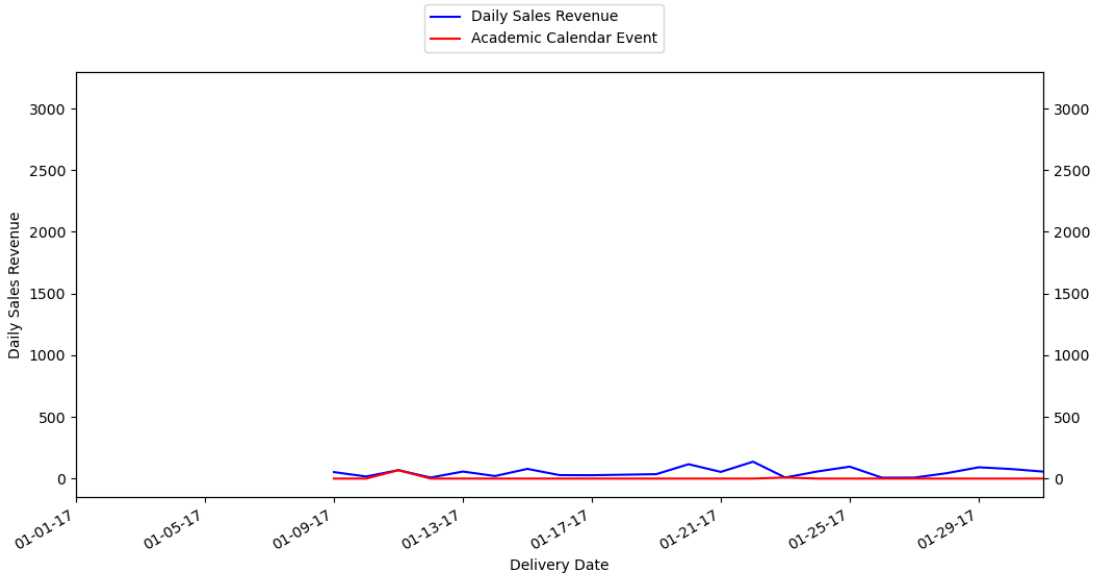


Figure 9.48: Sales Revenue January 2017

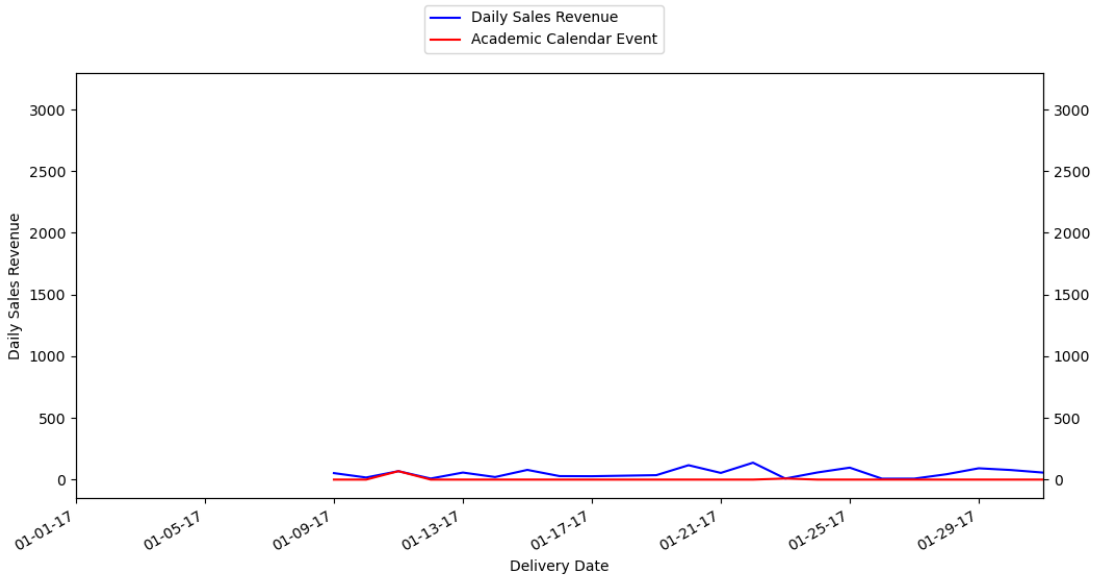


Figure 9.49: Sales Revenue January 2018

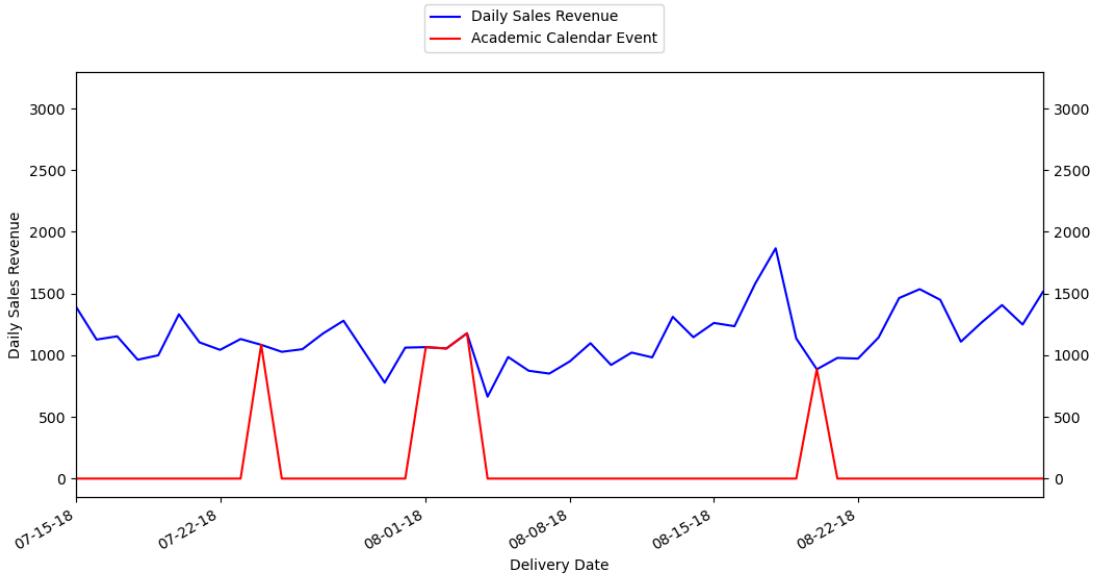


Figure 9.50: Sales Revenue July through August 2018

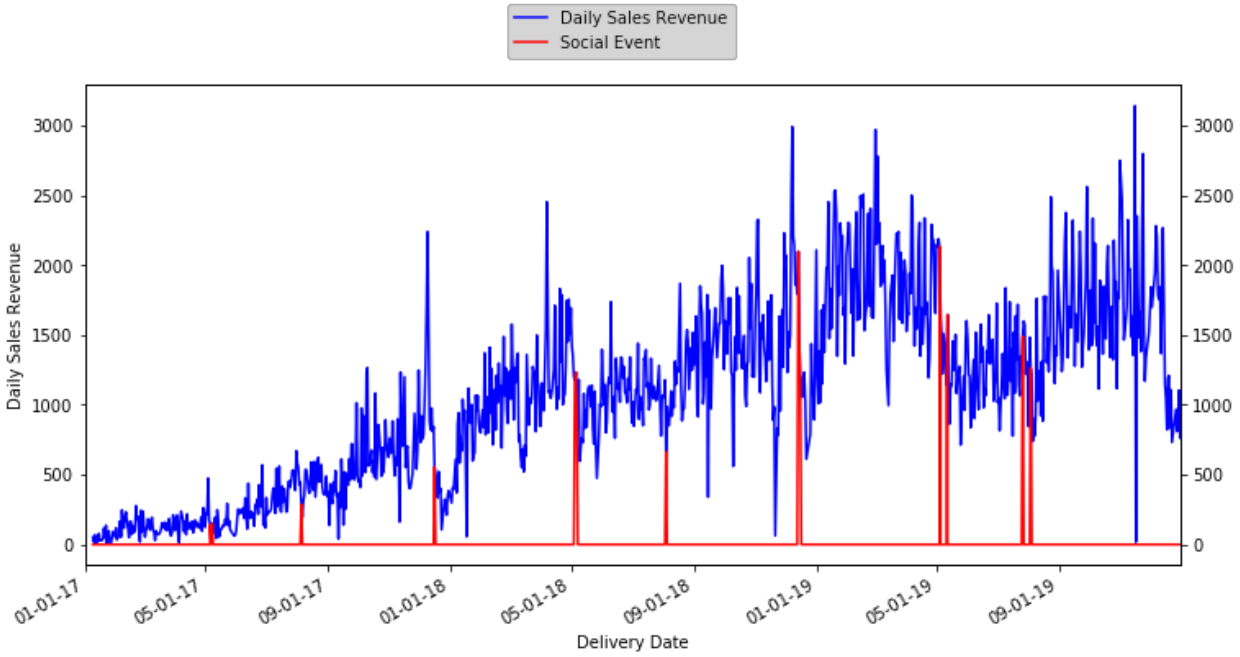


Figure 9.51: Social Events Sales Revenue Generated Over Dataset

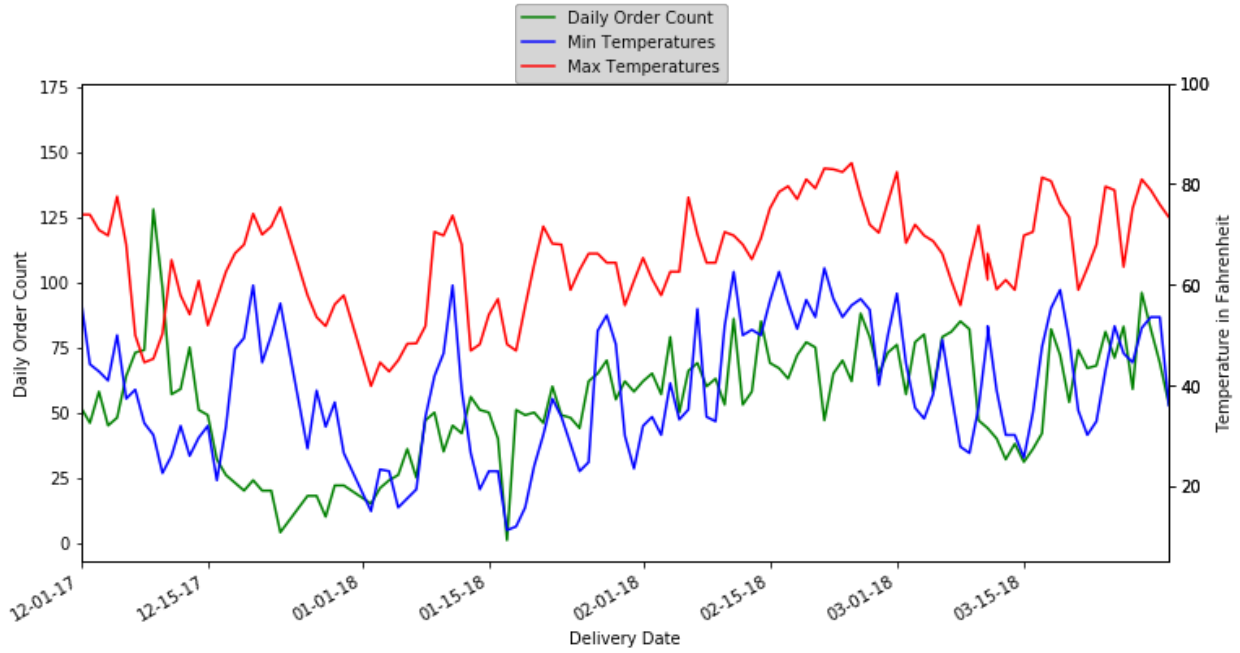


Figure 9.52: Weather Analysis Sales Count December 2017 to Early 2018

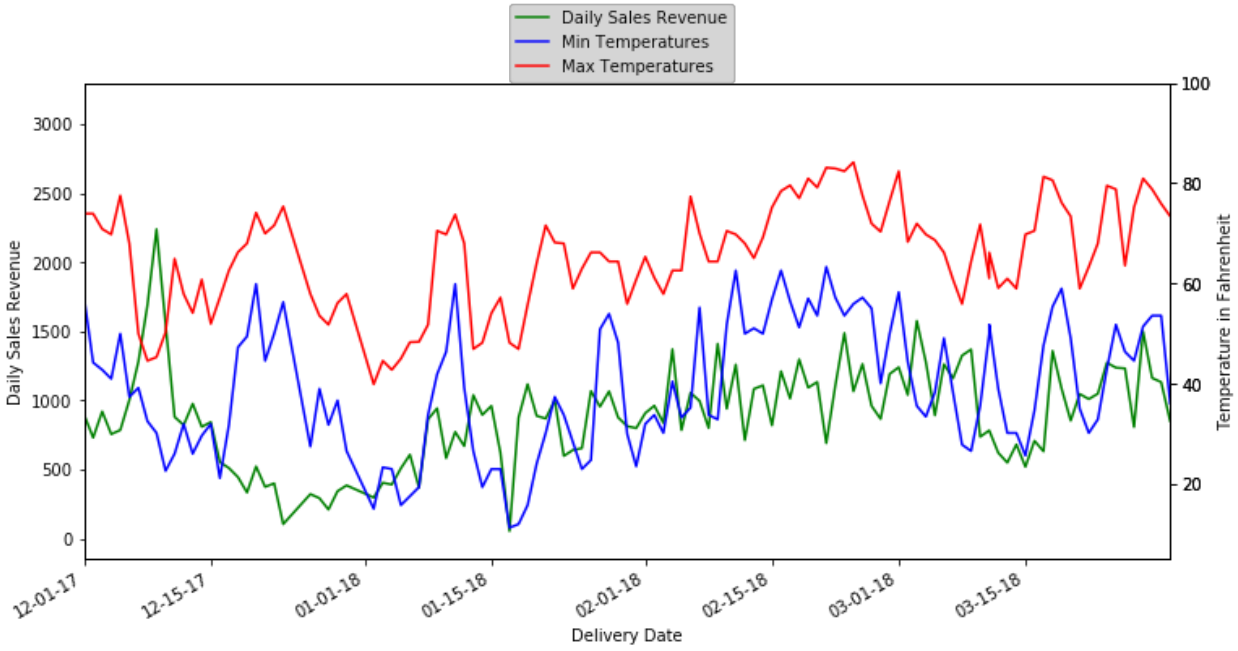


Figure 9.53: Weather Analysis Sales Revenue December 2017 to Early 2018

9.4 Algorithms/Code

GitHub Links are provided as follows:

Long COVID (Adult and Pediatric) - `https://github.com/kush4565/LongCOVID`

Food Delivery - `https://github.com/kush4565/FoodDelivery`

Algorithm 1 Algorithm 1: Customer Segmentation by Menu Item

$Input \leftarrow$ Data through Database Queries $Output \leftarrow Graphs$

Query1=Get $CustomerId$, $menuItemId$, $menuItemOrderCount$,

$menuItemOrderValue$ excluding outliers

dataframe df $\leftarrow$ Assign data obtained through Query1

Initialize $WCSS$

k in $range(1, 11)$: $kmeans \leftarrow KMeans$ ($number\ of\ clusters = k$, $init =$ "$k - means++$")

Use kmeans *package to fit the data*

Append the data to $WCSS$

Plot WCSS graph using K-Means package

$km \leftarrow KMeans$(number of clusters = 5)

$clusters \leftarrow$ KMeans Predicted data

Query2 = Get $menuItemOrderCount$, $menuItemOrderValue$ customer-wise excluding outliers df $\leftarrow$ Assign data through Query2

$X \leftarrow$ Values of x coordinates from df

kmeansmodel $\leftarrow$ Get data about cluster centroids through KMeans

$Y \leftarrow$ Predicted values of y coordinates through KMeans

SQL Query SQL3 = GET menuItemOrderCount, menuItemOrderValue menuItemwise excluding outliers df $\leftarrow$ Get data through SQL3

X $\leftarrow$ Values of x coordinates from df

kmeansmodel $\leftarrow$ Get data about cluster centroids through KMeans

Y $\leftarrow$ Predicted values of y coordinates through KMeans

Algorithm 2: Association Rule Mining for Menu Items in Orders

InputInput OutputOutput Data through Database Queries Graphs
Query1=Get $OrderId$, $menuItemName$
from $orderedItem$

$data$←Get through Query1
Initialize array $transactions$
$orderId$←0
$orderCount$←0
i in $range(length(data))$: $data[orderId][i]$!= $orderId$ $orderId > 0$
Append $newList$ to $transactions$
$orderId$ ←data[orderId][i]Create an empty $newList$ array
Append $data[menuItemName][i]$ to $newList$
Append $data[menuItemName][i]$ to $newList$

$results$ = list(apriori($transactions$, $minimum\ support$=0.002,
$minimumj\ confidence$=0.2, $minimum\ lift$=1.0,
$maximum\ length$=None))

$item$ in $results$: $pair$ ←item[0]
$items$ ←[x for x in pair]
Print "Rule: " + $items$[0] + "− > " + $items$[1])

Algorithm 3: Heatmap for Hourly Sales Revenue

InputInput OutputOutput Data through Database Queries Graphs

Query1=Get $deliveryDate$, $DailySales$,
from $orders$

$hour$ in $range(0, 24)$: Query2=Get $deliveryDate$, $DailySales$, from $orders$ during $hour$
$hour$ = 0 DataFrame $OrderSum$ ←Get data using Query1
DataFrame df ←copy of DataFrame OrderSum
col ← $Concatenate(h$,hour)
$df[col]$ ←OrderSum[DailySales] DataFrame $OrderSum$ ←Get data using Query2 for hour
col ← $Concatenate(h$,hour)
$df[col]$ ←OrderSum[DailySales]
Get correlation matrix of company's daily sales with hourly sales as -
$corr$ ←df.corr
Plot heatmap using $corr$ and seaborn package

References

[1] Global food e-commerce market: Consumer behavior analysis by countries, buying pattern analysis, demographics, trends analysis, survey findings and results, leading companies and their market strategies. *infiniumglobalresearch*.

[2] S. A. Alavi, S. Rezaei, N. Valaei, and W. K. Wan Ismail. Examining shopping mall consumer decision-making styles, satisfaction and purchase intention. *The International Review of Retail, Distribution and Consumer Research*, 26(3):272–303, 2016.

[3] D. M. Altmann, E. M. Whettlock, S. Liu, D. J. Arachchillage, and R. J. Boyton. The immunology of long covid. *Nature Reviews Immunology*, 23(10):618–634, 2023.

[4] R. Anderson and S. Karunamoorthy. E-satisfaction and e-loyalty: A contingency framework. *Psychology and Marketing*, 20:123 – 138, 02 2003.

[5] R. E. Anderson and S. S. Srinivasan. E-satisfaction and e-loyalty: A contingency framework. *Psychology & marketing*, 20(2):123–138, 2003.

[6] J. L. Annest, H. Hedegaard, L. H. Chen, M. Warner, and E. A. Smalls. Proposed framework for presenting injury data using icd-10-cm external cause of injury codes. 2014.

[7] A. A. Asadi-Pooya, H. Nemati, M. Shahisavandi, A. Akbari, A. Emami, M. Lotfi, M. Rostamihosseinkhani, Z. Barzegar, M. Kabiri, Z. Zeraatpisheh, et al. Long covid in children and adolescents. *World Journal of Pediatrics*, 17:495–499, 2021.

[8] B. J. Babin, W. R. Darden, and M. Griffin. Work and/or fun: measuring hedonic and utilitarian shopping value. *Journal of consumer research*, 20(4):644–656, 1994.

[9] A.-L. Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

[10] D. H. Barouch. Covid-19 vaccines—immunity, variants, boosters. *New England Journal of Medicine*, 387(11):1011–1020, 2022.

[11] A. W. Bartik, M. Bertrand, Z. B. Cullen, E. L. Glaeser, M. Luca, and C. T. Stanton. How are small businesses adjusting to covid-19? early evidence from a survey. Technical report, National Bureau of Economic Research, 2020.

[12] C. L. Brackel, C. R. Lap, E. P. Buddingh, M. A. van Houten, L. J. van der Sande, E. J. Langereis, M. A. Bannier, M. W. Pijnenburg, S. Hashimoto, and S. W. Terheggen-Lagro. Pediatric long-covid: An overlooked phenomenon? *Pediatric pulmonology*, 56(8):2495–2502, 2021.

[13] K. Brown, A. Yahyouche, S. Haroon, J. Camaradou, and G. Turner. Long covid and self-management. *Lancet (London, England)*, 399(10322):355, 2022.

[14] M. W. Brown, K. R. Lawrence, and M. A. Paolini. Method, system and program for specifying an electronic food menu with food preferences from a universally accessible database, Nov. 11 2003. US Patent 6,646,659.

[15] D. Buonsenso, L. D. Gennaro, C. D. Rose, R. Morello, F. D'Ilario, G. Zampino, M. Piazza, A. L. Boner, C. Iraci, S. O'Connell, et al. Long-term outcomes of pediatric infections: from traditional infectious diseases to long covid. *Future microbiology*, 17(7):551–571, 2022.

[16] G. Calderer and M. L. Kuijjer. Community detection in large-scale bipartite biological networks. *Frontiers in Genetics*, 12:520, 2021.

[17] F. Callard and E. Perego. How and why patients made long covid. *Social science & medicine*, 268:113426, 2021.

[18] M. C. Campbell, J. J. Inman, A. Kirmani, and L. L. Price. In times of trouble: A framework for understanding consumers' responses to threats, 2020.

[19] T. S. Carsten Hirschberg, Alexander Rajko and M. Wrulich. The changing market for food delivery. *Mckinsey*.

[20] CDC. Comorbidities. 2022.

[21] CDC. Clong covid or post-covid conditions. *Centers for disease control and prevention*, 2023.

[22] S. Chahar and P. K. Roy. Covid-19: A comprehensive review of learning models. *Archives of Computational Methods in Engineering*, pages 1–26, 2021.

[23] C. Chang and C. Tang. Community detection for networks with unipartite and bipartite structure. *New Journal of Physics*, 16(9):093001, 2014.

[24] H. Chen, Y. Zhang, D. Wu, C. Gong, Q. Pan, X. Dong, Y. Wu, K. Zhang, S. Wang, J. Lei, et al. Comorbidity in adult patients hospitalized with type 2 diabetes in northeast china: an analysis of hospital discharge data from 2002 to 2013. *BioMed research international*, 2016, 2016.

[25] F. Collins, S. Adam, C. Colvis, E. Desrosiers, R. Draghia-Akli, A. Fauci, M. Freire, G. Gibbons, M. Hall, E. Hughes, et al. The nih-led research response to covid-19. *Science*, 379(6631):441–444, 2023.

[26] J. Croft. Long covid is very rare among children, research finds. 2022.

[27] H. Crook, S. Raza, J. Nowell, M. Young, and P. Edison. Long covid—mechanisms, risk factors, and management. *bmj*, 374, 2021.

[28] H. E. Davis, G. S. Assaf, L. McCorkell, H. Wei, R. J. Low, Y. Re'em, S. Redfield, J. P. Austin, and A. Akrami. Characterizing long covid in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine*, 38, 2021.

[29] R. R. Deer, M. A. Rock, N. Vasilevsky, L. Carmody, H. Rando, A. J. Anzalone, M. D. Basson, T. D. Bennett, T. Bergquist, E. A. Boudreau, et al. Characterizing long covid: deep phenotype of a complex condition. *EBioMedicine*, 74, 2021.

[30] C. Del Rio, S. B. Omer, and P. N. Malani. Winter of omicron—the evolving covid-19 pandemic. *Jama*, 327(4):319–320, 2022.

[31] Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel. Goldenbullet: Automated classification of product data in e-commerce. In *Proceedings of the 5th international conference on business information systems*, volume 5. Citeseer, 2002.

[32] H. Ejaz, A. Alsrhani, A. Zafar, H. Javed, K. Junaid, A. E. Abdalla, K. O. Abosalif, Z. Ahmed, and S. Younas. Covid-19 and comorbidities: Deleterious impact on infected patients. *Journal of infection and public health*, 13(12):1833–1839, 2020.

[33] C. Fan, X. Lei, and F.-X. Wu. Prediction of circrna-disease associations using katz model based on heterogeneous networks. *International journal of biological sciences*, 14(14):1950, 2018.

[34] X. Fang, S. Li, H. Yu, P. Wang, Y. Zhang, Z. Chen, Y. Li, L. Cheng, W. Li, H. Jia, et al. Epidemiological, comorbidity factors with severity and prognosis of covid-19: a systematic review and meta-analysis. *Aging (albany NY)*, 12(13):12493, 2020.

[35] N. C. for Immunization et al. Science brief: Evidence used to update the list of underlying medical conditions associated with higher risk for severe covid-19. In *CDC COVID-19 Science Briefs [Internet]*. Centers for Disease Control and Prevention (US), 2022.

[36] N. R. P. Galit Shmueli, Peter C. Bruce. *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner, 3rd Edition*. Wiley, 2016.

[37] B. Gallo Marin, G. Aghagoli, K. Lavine, L. Yang, E. J. Siff, S. S. Chiang, T. P. Salazar-Mather, L. Dumenco, M. C. Savaria, S. N. Aung, et al. Predictors of covid-19 severity: a literature review. *Reviews in medical virology*, 31(1):1–10, 2021.

[38] M. Garg, M. Maralakunte, S. Garg, S. Dhooria, I. Sehgal, A. S. Bhalla, R. Vijayvergiya, S. Grover, V. Bhatia, P. Jagia, et al. The conundrum of 'long-covid-19: a narrative review. *International journal of general medicine*, pages 2491–2506, 2021.

[39] L. Geddes. Nine factors that could boost your risk of long covid. *VaccinesWork*, 2022.

[40] S. Ghosh, M. Halappanavar, A. Tumeo, A. Kalyanaraman, H. Lu, D. Chavarria-Miranda, A. Khan, and A. Gebremedhin. Distributed louvain algorithm for graph community detection. In *2018 IEEE international parallel and distributed processing symposium (IPDPS)*, pages 885–895. IEEE, 2018.

[41] K.-I. Goh and I.-G. Choi. Exploring the human diseasome: the human disease network. *Briefings in functional genomics*, 11(6):533–542, 2012.

[42] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[43] A. Göktuğ, A. Güngör, F. N. Öz, Z. Akelma, M. M. Güneylioğlu, R. M. Yaradılmış, I. Bodur, B. Öztürk, A. Tekeli, C. D. Karacan, et al. Evaluation of epidemiological, demographic, clinical characteristics and laboratory findings of covid-19 in the pediatric emergency department. *Journal of Tropical Pediatrics*, 67(4):fmab066, 2021.

[44] B. L. Gottesman, J. Yu, C. Tanaka, C. A. Longhurst, and J. J. Kim. Incidence of new-onset type 1 diabetes among us children during the covid-19 global pandemic. *JAMA pediatrics*, 176(4):414–415, 2022.

[45] W.-j. Guan, W.-h. Liang, Y. Zhao, H.-r. Liang, Z.-s. Chen, Y.-m. Li, X.-q. Liu, R.-c. Chen, C.-l. Tang, T. Wang, et al. Comorbidity and its impact on 1590 patients with covid-19 in china: a nationwide analysis. *European Respiratory Journal*, 55(5), 2020.

[46] N. Y. Guillaume Habault, Yuya Taniguchi. Delivery management system based on vehicles monitoring and a machine-learning mechanism. *IEEE*.

[47] M. Gupta, N. Gupta, and M. Esang. Long covid in children and adolescents. *The Primary Care Companion for CNS Disorders*, 24(2):40720, 2022.

[48] G. Habault, Y. Taniguchi, and N. Yamanaka. Delivery management system based on vehicles monitoring and a machine-learning mechanism. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pages 1–5. IEEE, 2018.

[49] M. A. Haendel, C. G. Chute, T. D. Bennett, D. A. Eichmann, J. Guinney, W. A. Kibbe, P. R. Payne, E. R. Pfaff, P. N. Robinson, J. H. Saltz, et al. The national covid cohort collaborative (n3c): rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association*, 28(3):427–443, 2021.

[50] V. Higgins, D. Sohaei, E. P. Diamandis, and I. Prassas. Covid-19: from an acute to chronic disease? potential long-term health consequences. *Critical reviews in clinical laboratory sciences*, 58(5):297–310, 2021.

[51] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1):58–64, 2000.

[52] J. E. Hobbs. Food supply chains during the covid-19 pandemic. *Wiley Online Library*.

[53] W. D. Y. J. M. W.-Z. N. X. Z. Y. G. C. W. W. M.-S. E. F. F. J. S. J. X. G. Y. Huijing Ma, Quinghao Ye. Can clinical symptoms and laboratory results predict ct abnormality? initial findings using novel machine learning techniques in children with covid-19 infections. *National Library of Medicine*, 8, 2021.

[54] A. Humphreys and C. J. Thompson. Branding disaster: Reestablishing trust through the ideological containment of systemic risk anxieties. *Journal of Consumer Research*, 41(4):877–910, 2014.

[55] O. Irfan, F. Muttalib, K. Tang, L. Jiang, Z. S. Lassi, and Z. Bhutta. Clinical characteristics, treatment and outcomes of paediatric covid-19: a systematic review and meta-analysis. *Archives of disease in childhood*, 106(5):440–448, 2021.

[56] S. Jack. Coronavirus: Online shopping website ocado suspends service. *BBC*.

[57] S. Jakhmola, O. Indari, B. Baral, D. Kashyap, N. Varshney, A. Das, S. Chatterjee, and H. C. Jha. Comorbidity assessment is essential during covid-19 treatment. *Frontiers in physiology*, 11:984, 2020.

[58] S.-P. Jeng. The influences of airline brand credibility on consumer purchase intentions. *Journal of Air Transport Management*, 55:1–8, 2016.

[59] C. Jones. Ecommerce is growing nicely while mcommerce is on a tear. *Forbes, England*.

[60] S. Kalogiannidis. Covid impact on small business. *International Journal of Social Science and Economics Invention*, 6(12):387–391, 2020.

[61] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[62] M. Kekatos. Percentage of us adults with long covid falls to 6 2023.

[63] M. Kheirkhahzadeh, A. Lancichinetti, and M. Rosvall. Efficient community detection of network flows for varying markov times and bipartite networks. *Physical Review E*, 93(3):032309, 2016.

[64] S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang. Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert systems with applications*, 31(1):101–107, 2006.

[65] D. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. Warwick Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, et al. Goatools: A python library for gene ontology analyses. *Scientific reports*, 8(1):10872, 2018.

[66] D. Koller, G. Schön, I. Schäfer, G. Glaeske, H. van den Bussche, and H. Hansen. Multimorbidity and long-term care dependency—a five-year follow-up. *BMC geriatrics*, 14(1):1–9, 2014.

[67] J. E. Krakar, K. Longtin, and S. P. Branch. *An overview of the Canadian agriculture and agri-food system*. Citeseer, 2005.

[68] T. Lancet. Understanding long covid: a modern medical challenge. *Lancet (London, England)*, 398(10302):725, 2021.

[69] S. W. Lee, H. J. Sung, and H. M. Jeon. Determinants of continuous intention on food delivery apps: extending utaut2 with information quality. *Sustainability*, 11(11):3141, 2019.

[70] X. Lei, C. Zhang, et al. Predicting metabolite-disease associations based on linear neighborhood similarity with improved bipartite network projection algorithm. *Complexity*, 2020, 2020.

[71] K. Li and Y. Pang. A unified community detection algorithm in complex network. *Neurocomputing*, 130:36–43, 2014.

[72] M. Li, P. Lei, B. Zeng, Z. Li, P. Yu, B. Fan, C. Wang, Z. Li, J. Zhou, S. Hu, et al. Coronavirus disease (covid-19): spectrum of ct findings and temporal progression of the disease. *Academic radiology*, 27(5):603–608, 2020.

[73] S. Li, M. Xie, and X. Liu. A novel approach based on bipartite network recommendation and katz model to predict potential micro-disease associations. *Frontiers in Genetics*, 10:1147, 2019.

[74] X. Liao, D. Zheng, and X. Cao. Coronavirus pandemic analysis through tripartite graph clustering in online social networks. *Big Data Mining and Analytics*, 4(4):242–251, 2021.

[75] S. Lopez-Leon, T. Wegman-Ostrosky, N. C. Ayuzo del Valle, C. Perelman, R. Sepulveda, P. A. Rebolledo, A. Cuapio, and S. Villapol. Long-covid in children and adolescents: A systematic review and meta-analyses. *Scientific reports*, 12(1):9950, 2022.

[76] K. Lu, K. Yang, E. Niyongabo, Z. Shu, J. Wang, K. Chang, Q. Zou, J. Jiang, C. Jia, B. Liu, et al. Integrated network analysis of symptom clusters across disease conditions. *Journal of Biomedical Informatics*, 107:103482, 2020.

[77] J. F. Ludvigsson. Case report and systematic review suggest that children may experience similar long-term effects to adults after clinical covid-19. *Acta Paediatrica*, 110(3):914–921, 2021.

[78] K. Macpherson, K. Cooper, J. Harbour, D. Mahal, C. Miller, and M. Nairn. Experiences of living with long covid and of accessing healthcare services: a qualitative systematic review. *BMJ open*, 12(1):e050979, 2022.

[79] C. Marcus. A practical yet meaningful approach to customer segmentation. *Journal of consumer marketing*, 1998.

[80] C. S. A. K. A. A.-J. C. I. M. A. R. A. Maria Nicola, Zaid Alsafi. The socio-economic implications of the coronavirus pandemic (covid-19): A review. *Science Direct*.

[81] R. McFee. Covid-19 laboratory testing/cdc guidelines. *Disease-a-month*, 66(9):101067, 2020.

[82] N. A. Megahed and E. M. Ghoneim. Antivirus-built environment: lessons learned from covid-19 pandemic. *Sustainable Cities and Society*, page 102350, 2020.

[83] M. E. Mikkelsen, B. Abramoff, and J. G. Elmore. Covid-19: Evaluation and management of adults with persistent symptoms following acute illness (" long covid"). *Waltham, MA: UpToDate*, 2022.

[84] A. Mishra. Demystifying louvain's algorithm and its implementation in gpu. 2019.

[85] S. Mukherjee and K. Pahan. Is covid-19 gender-sensitive? *Journal of Neuroimmune Pharmacology*, 16:38–47, 2021.

[86] A. Mussell, T. Bilyea, and D. Hedley. Agri-food supply chains and covid-19: Balancing resilience and vulnerability. *Agri-Food Economic Systems*, 2020.

[87] W. H. Ng, T. Tipih, N. A. Makoah, J.-G. Vermeulen, D. Goedhals, J. B. Sempa, F. J. Burt, A. Taylor, and S. Mahalingam. Comorbidities in sars-cov-2 patients: a systematic review and meta-analysis. *MBio*, 12(1):10–1128, 2021.

[88] L. H. Nguyen and S. Holmes. Ten quick tips for effective dimensionality reduction. *PLoS computational biology*, 15(6):e1006907, 2019.

[89] A. J. O'Malley, T. A. Bubolz, and J. S. Skinner. The diffusion of health care fraud: A bipartite network analysis. *Social Science & Medicine*, 327:115927, 2023.

[90] H. Onyeaka, C. K. Anumudu, Z. T. Al-Sharify, E. Egele-Godswill, and P. Mbaegbu. Covid-19 pandemic: A review of the global lockdown and its far-reaching effects. *Science progress*, 104(2):00368504211019854, 2021.

[91] D. V. Parums. long covid, or post-covid syndrome, and the global impact on health care. *Medical science monitor: international medical journal of experimental and clinical research*, 27:e933446–1, 2021.

[92] A. Pavli, M. Theodoridou, and H. C. Maltezou. Post-covid syndrome: Incidence, clinical spectrum, and challenges for primary healthcare professionals. *Archives of medical research*, 52(6):575–581, 2021.

[93] J. Peto, N. A. Alwan, K. M. Godfrey, R. A. Burgess, D. J. Hunter, E. Riboli, P. Romer, I. Buchan, T. Colbourn, C. Costelloe, et al. Universal weekly testing as the uk covid-19 lockdown exit strategy. *The Lancet*, 395(10234):1420–1421, 2020.

[94] E. R. Pfaff, A. T. Girvin, T. D. Bennett, A. Bhatia, I. M. Brooks, R. R. Deer, J. P. Dekermanjian, S. E. Jolley, M. G. Kahn, K. Kostka, et al. Identifying who has long covid in the usa: a machine learning approach using n3c data. *The Lancet Digital Health*, 4(7):e532–e541, 2022.

[95] M. Phipps and J. L. Ozanne. Routines disrupted: Reestablishing security through practice alignment. *Journal of Consumer Research*, 44(2):361–380, 2017.

[96] X. Que, F. Checconi, F. Petrini, and J. A. Gunnels. Scalable community detection with the louvain algorithm. In *2015 IEEE International Parallel and Distributed Processing Symposium*, pages 28–37. IEEE, 2015.

[97] C. P. Rajan Gupta. A machine learning framework for predicting purchase by online customers based on dynamic pricing. *Sciencedirect*, pages 601–602.

[98] R. Rastogi, I. H. Cerda, A. Ibrahim, J. A. Chen, C. Stevens, and C. H. Liu. Long covid and psychological distress in young adults: Potential protective effect of a prior mental health diagnosis. *Journal of Affective Disorders*, 340:639–648, 2023.

[99] A. Raveendran, R. Jayadevan, and S. Sashidharan. Long covid: an overview. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(3):869–875, 2021.

[100] S. Richardson, J. S. Hirsch, M. Narasimhan, J. M. Crawford, T. McGinn, K. W. Davidson, D. P. Barnaby, L. B. Becker, J. D. Chelico, S. L. Cohen, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area. *Jama*, 323(20):2052–2059, 2020.

[101] A. Sanyaolu, C. Okorie, A. Marinkovic, R. Patidar, K. Younis, P. Desai, Z. Hosein, I. Padda, J. Mangat, and M. Altaf. Comorbidity and its impact on patients with covid-19. *SN comprehensive clinical medicine*, 2:1069–1076, 2020.

[102] L. Searing. 1 percent of children had long covid through last year, cdc says. *The Washington Post*, page 2.

[103] P. Seetharaman. Business models shifts: Impact of covid-19. *International Journal of Information Management*, 54:102173, 2020.

[104] L. Semenzato, J. Botton, J. Drouin, F. Cuenot, R. Dray-Spira, A. Weill, and M. Zureik. Chronic diseases, health conditions and risk of covid-19-related hospitalization and in-hospital mortality during the first wave of the epidemic in france: a cohort study of 66 million people. *The Lancet Regional Health–Europe*, 8, 2021.

[105] Y. Sharifi, M. Payab, E. Mohammadi-Vajari, S. M. M. Aghili, F. Sharifi, N. Mehrdad, E. Kashani, Z. Shadman, B. Larijani, and M. Ebrahimpur. Association between cardiometabolic risk factors and covid-19 susceptibility, severity and mortality: a review. *Journal of Diabetes & Metabolic Disorders*, 20(2):1743–1765, 2021.

[106] N. R. S. Sibuyi, A. O. Fadaka, and O. B. Adewale. Understanding the epidemiology, pathophysiology, diagnosis and management of sars-cov-2. 2020.

[107] N. Smallwood, W. Harrex, M. Rees, K. Willis, and C. M. Bennett. Covid-19 infection and the broader impacts of the pandemic on healthcare workers. *Respirology*, 27(6):411–426, 2022.

[108] T. Soper. Covid-19 crisis sparks 'inflection point'for online grocery—and huge revenue for amazon. *Geek Wire*, 4(07):2020, 2020.

[109] E. Staff. The atlanta restaurants now permanently closed due to the covid-19 crisis. *Atlanta Eater*.

[110] M. Stangerup, M. B. Hansen, R. Hansen, L. P. Sode, B. Hesselbo, K. Kostadinov, B. S. Olesen, and H. Calum. Hand hygiene compliance of healthcare workers before and during the covid-19 pandemic: a long-term follow-up study. *American journal of infection control*, 49(9):1118–1122, 2021.

[111] Z. Su, A. Cheshmehzangi, B. L. Bentley, D. McDonnell, S. Šegalo, J. Ahmad, H. Chen, L. A. Terjesen, E. Lopez, S. Wagers, et al. Technology-based interventions for health challenges older women face amid covid-19: a systematic review protocol. *Systematic Reviews*, 11(1):1–8, 2022.

[112] A. Subramanian, K. Nirantharakumar, S. Hughes, P. Myles, T. Williams, K. M. Gokhale, T. Taverner, J. S. Chandan, K. Brown, N. Simms-Williams, et al. Symptoms and risk factors for long covid in non-hospitalized adults. *Nature medicine*, 28(8):1706–1714, 2022.

[113] C. H. Sudre, B. Murray, T. Varsavsky, M. S. Graham, R. S. Penfold, R. C. Bowyer, J. C. Pujol, K. Klaser, M. Antonelli, L. S. Canas, et al. Attributes and predictors of long-covid: analysis of covid cases and their symptoms collected by the covid symptoms study app (preprint). 2020.

[114] C. H. Sudre, B. Murray, T. Varsavsky, M. S. Graham, R. S. Penfold, R. C. Bowyer, J. C. Pujol, K. Klaser, M. Antonelli, L. S. Canas, et al. Attributes and predictors of long covid. *Nature medicine*, 27(4):626–631, 2021.

[115] K. K. Sum, S. Cai, E. Law, B. Cheon, G. Tan, E. Loo, Y. S. Lee, F. Yap, J. K. Y. Chan, M. Daniel, et al. Covid-19–related life experiences, outdoor play, and long-term adiposity changes among preschool-and school-aged children in singapore 1 year after lockdown. *JAMA pediatrics*, 176(3):280–289, 2022.

[116] K. Sytian. How mobile apps are reshaping the e-commerce industry. *singlegrain*.

[117] M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, and L. F. Ng. The trinity of covid-19: immunity, inflammation and intervention. *Nature Reviews Immunology*, 20(6):363–374, 2020.

[118] C. Thompson. Consumer risk perceptions in a community of reflexive doubt. *Journal of Consumer Research*, 32:235–248, 09 2005.

[119] J. Tran, R. Norton, N. Conrad, F. Rahimian, D. Canoy, M. Nazarzadeh, and K. Rahimi. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the uk between 2000 and 2014: a population-based cohort study. *PLoS medicine*, 15(3):e1002513, 2018.

[120] P. M. Tsang and S. Tse. A hedonic model for effective web marketing: an empirical examination. *Industrial Management & Data Systems*, 2005.

[121] A. Unnikrishnan and M. A. Figliozzi. A study of the impact of covid-19 on home delivery purchases and expenditures. 2020.

[122] D. L. Villeneuve, M. M. Angrish, M. C. Fortin, I. Katsiadaki, M. Leonard, L. Margiotta-Casaluci, S. Munn, J. M. O'Brien, N. L. Pollesch, L. C. Smith, et al. Adverse outcome pathway networks ii: network analytics. *Environmental toxicology and chemistry*, 37(6):1734–1748, 2018.

[123] S. R. Vincent Cheow Sern Yeo, See-Kwong Goh. Consumer experiences, attitude and behavioral intention toward online food delivery (ofd) services. *Science Direct*.

[124] C. Wang, F. Wang, and T. Onega. Network optimization approach to delineating health care service areas: Spatially constrained louvain and leiden algorithms. *Transactions in GIS*, 25(2):1065–1081, 2021.

[125] D. Wang, R. Li, J. Wang, Q. Jiang, C. Gao, J. Yang, L. Ge, and Q. Hu. Correlation analysis between disease severity and clinical and biochemical characteristics of 143 cases of covid-19 in wuhan, china: a descriptive study. *BMC infectious diseases*, 20:1–9, 2020.

[126] J. Wang, J. J. Ma, J. Liu, D. D. Zeng, C. Song, and Z. Cao. Prevalence and risk factors of comorbidities among hypertensive patients in china. *International journal of medical sciences*, 14(3):201, 2017.

[127] P. Wasilewski, B. Mruk, S. Mazur, G. Półtorak-Szymczak, K. Sklinda, and J. Walecki. Covid-19 severity scoring systems in radiological imaging–a review. *Polish journal of radiology*, 85(1):361–368, 2020.

[128] C. Watson. Diabetes risk rises after covid, massive study finds. *Nature (Lond.)*, 2022.

[129] A.-L. B. A. S. XueZhong Zhou, Jörg Menche. Human symptoms-disease network. *PubMed*, (10):1038, 2014.

[130] L. Yao, G. Wang, L. Aleya, M. Maida, J. C. Graff, D. Sun, and W. Gu. Was the rate of long covid as high as 45%–a scary report with flaw. *EClinicalMedicine*, 59, 2023.

[131] J. Yates, A. Gutiérrez-Sacristán, V. Jouhet, K. LeBlanc, C. Esteves, U. D. Network, T. N. DeSain, N. Benik, J. Stedman, N. Palmer, et al. Finding commonalities in rare diseases through the undiagnosed diseases network. *Journal of the American Medical Informatics Association*, 28(8):1694–1702, 2021.

[132] V. C. S. Yeo, S.-K. Goh, and S. Rezaei. Consumer experiences, attitude and behavioral intention toward online food delivery (ofd) services. *Journal of Retailing and Consumer Services*, 35:150–162, 2017.

[133] S. M. Yoo. About 302022.