Project Number: 22791

Modeling and Analysis of COVID-19 using Mathematical and Data Analytic Methods

A Major Qualifying Project submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

by

Samuel Berbeco (MA)

Alvin Lee (DS)

Larson Ost (MA)

Tom Strow (MA)

Xizhao Zhang (DS)

Approved

Professor Roger Lui
Advisor

Approved

Professor Oren Mangoubi
Advisor

Contents

List of Figures

Abstract

COVID-19 is a highly contagious infectious disease that has spread throughout the world. On January 19th, 2020, the first case in the United States was reported by a patient in Washington State. Since then, COVID-19 has killed over 850,000 people in the United States alone, and despite having readily available vaccines and being over two years into the pandemic, COVID-19 cases remain high [7]. In continuing to fight the global pandemic, it is important to study the development of COVID-19 and methods to mitigate its spread.

In order to contribute to research, we developed an epidemiological compartmental model of the pandemic using a system of differential equations from which we determined a formula for the Basic Reproduction Number. Using the model, we conducted a case study within Massachusetts to determine the effects vaccinations and other preventive measures have on mitigating the spread of COVID-19 by using the Basic Reproduction Number as an indicator.

The epidemiological model requires a set of parameters that describe the behavior of COVID-19: γ, λ, μ, α, v(t), c(t), and β. Each parameters is defined in section 2. Using online research, we were able to find resources describing the values for γ, λ, and α. The parameters v(t), c(t) and μ were provided by the Massachusetts Government Response Reporting Website [17] [18] [19] [20] [21] [22]. By conducting a statistical analysis in which we compared the model's expected cases to real reported cases, we were able to solve for β.

Once all parameters were determined, we solved for the Basic Reproduction Number and described the reasoning for changes in its behavior. This includes a comparison of the Basic Reproduction Number's time series alongside changes in travel restrictions, government mandated lockdowns, mask mandates, and the vaccination rate. The comparison sheds light on which mitigation techniques were the most effective at preventing the spread of COVID-19. Also, this information indicates whether or not COVID-19 will evolve into an endemic state, or diminish until we have a disease-free state.

Furthermore, we developed several linear regression models to assess the effectiveness of COVID-19 mitigation techniques. The multiple linear regression describes the extent that techniques such as mask mandates, social event restrictions, business closings, seasonal changes, and vaccinations have on influencing the Basic Reproduction Number that we previously solved for. Also, a logistic regression was developed to determine which restrictions are most likely to result in an endemic or disease-free state. These regressions describe the significance of various restrictions in preventing the spread of COVID-19.

We hope that this information will be helpful for future research into COVID-19 and for determining a more accurate Basic Reproduction Number. Further research will allow us to understand the behavior and spread of COVID-19, allowing for a more comprehensive solution to the pandemic.

In addition to differential equation based methods of studying COVID-19's reproduction number and modeling its spread, we studied COVID-19 through a data-driven lens as well. During the course of our study, we collected and cleaned Massachusetts COVID-19 data to

forecast future cases. Data was acquired from the Massachusetts government website. The model selected for this task was the Autoregressive Integrated Moving Average model (ARIMA). By generating a model through both manually selected and automatically selected parameters, we were able to produce forecasts of COVID cases a week in advance; however, the forecasts were not as accurate as we had hoped.

Finally, the last portion of the project was a clustering analysis as a method to determine which states or groups of states managed to keep the COVID cases low, relative to their population. To do this, all the daily vaccination and COVID case data was collected from the Center of Disease Control, and then was cleaned for analysis purposes. We then separated the data into three key time periods: Pre-Delta, Delta, and Omicron in order to separate distinct periods where we'd expect different results. By aggregating the COVID cases per month for each state, as well as the total vaccinations of the state in that month, we were able to create three scatter plots of vaccinations per population and cases per population: one for each time period. These scatter plots were then clustered after performing the min-max scaling technique, and clustered using K-Means. From the clusters generated, we are able to identify states that had high COVID positivity rate, despite having a high number of vaccinations. Another finding from this clustering analysis is the clear drop off of vaccination efficacy as a preventative measure for COVID spread as new variants emerged.

Acknowledgements

There are many people our team would like to thank for contributing to the success of this project. Firstly, we thank our families for their overwhelming support throughout this project, our academic careers, and the ongoing pandemic.

Also, the many professors who have instructed us during our time at Worcester Polytechnic Institute deserve recognition as their teachings provided us the foundational knowledge for our work on this project.

Additionally, we extend our gratitude to the countless people and institutions who contributed to COVID-19 research and data collection throughout the pandemic. Whether by collecting data or publishing research, their work supplied us with the tools and knowledge to successfully complete our project.

Last, but certainly not least, we would like to thank our advisors: Professor Lui and Professor Mangoubi. The guidance they provided to us was invaluable for both the success of this project and our preparedness for future work. We are grateful for their time and effort throughout this project.

1 Mathematical Sciences Group

The Coronavirus Disease 2019 is a deadly respiratory virus that has infected over 350 million people. In December of 2019, a cluster of patients in Wuhan, China, contracted an unknown illness and began exhibiting symptoms of pneumonia: shortness of breath and fever. In early 2020, Chinese public health officials announced the genetic sequence of the unknown respiratory virus and noted it has a similar sequence to Severe Acute Respiratory Syndrome (SARS). As a result, it was named SARS-CoV-2, and is the virus that causes the Coronavirus Disease (COVID-19) [3].

COVID-19 quickly spread across the world and in January 2020 the first case was reported in the United States in Washington State. After an alarming rate of spread, the World Health Organization declared a pandemic on March 11, 2020 [3].

Artists at the Center for Disease Control and Prevention (CDC), the national public health agency of the United States, designed the now iconic red and white image of the virus which can be seen in figure 1.1:

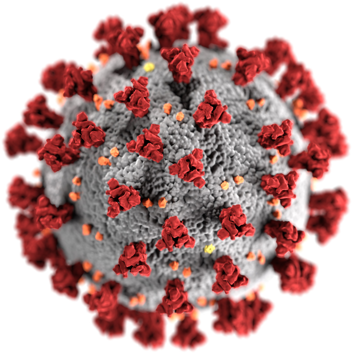


Figure 1.1: Artistic rendering of Sars-CoV-2 by the CDC[8]

The effects of COVID-19 on someone who has contracted the virus can vary widely from person to person. Symptoms may appear 2-14 days after exposure and can range from that of a common cold to life threatening respiratory issues. The CDC originally recommended that people who have been in contact with someone who has the virus or is symptomatic isolate themselves for 14 days. The recommended actions by the CDC to prevent the spread of COVID-19 has changed over time as new research has emerged, but currently the recommendations are for people to stay at least six feet apart at all times, limit large gatherings, avoid indoor gatherings, and wear a medical mask to prevent COVID-19 particulate from transferring from person to person on water vapor in the air [9].

In the United States, the coronavirus has spread in a series of massive surges due to people's behavior and different variants of the virus. A variant is a strain of the coronavirus that has undergone a mutation that distinguishes it from other versions of the coronavirus. These variants can exhibit different levels of severity, rate of spread, and resistance to treatment. The first surge of cases was in the winter months of 2020-21 when there were a large number of people traveling for the holidays, causing the largest rate of new cases thus far in the pandemic. In December and January of the same holiday season, three new variants were detected: Alpha, Beta, and Gamma. The CDC currently designates these as Variants of Concern because they all are all more contagious than the original variant and lead to more severe cases [4].

The first major surge of the coronavirus was eventually tempered by the introduction of vaccines. On december 11, 2020. The Food and Drug Administration issued an Emergency Use Authorization for the Pfizer-BioNTech COVID-19 Vaccine. Quickly after, another emergency authorization was granted for the Moderna COVID-19 vaccine. The vaccinations were released in stages to people to maximize their effectiveness. In the beginning of the pandemic, high-risk groups and essential workers such as medical staff were given the first round of vaccinations. Now, they are available to everyone over the age of five who is willing to get one [5].

In July 2021, the second major surge of cases occurred when the Delta variant emerged. The Delta variant is estimated to be twice as contagious than previous variants, and more likely to cause severe illness for people who remain unvaccinated. This surge saw less cases then that of the holiday season of 2020-2021 due to the presence of widely available vaccines [6].

This second surge concluded alongside the introduction of booster shots. Vaccines efficacy wanes over time, and Pfizer and Moderna recommended booster shots as a follow-up to vaccines in order to preserve their effectiveness. The CDC recommends that five months after becoming fully vaccinated, one should receive their booster shot [3].

The Delta variant remained the dominant variant in the United States until December of 2021 when a new variant, Omicron, surpassed it. Research into Omicron's mutations are yet to be fully understood, but early research by the CDC suggests that it is more contagious than any previously detected variant. Soon after it became dominant, the United States experienced a large surge in cases in January 2022. New cases peaked at over four times the level of the previous surge [6].

Due to the continuation of high infection rates despite the access to vaccinations, and knowledge of COVID-19 that we have gained since the beginning of the pandemic, it is important to add to our understanding of the virus to see where improvements can be made to mitigation efforts. COVID-19's affect on American Society is constantly changing, so it is important to understand with hindsight how COVID-19 changed the way we lived so in the future we can make more informed decisions regarding policy making and recommended behavior. For example, we hope to further develop our understanding of the impact vaccinations have had on COVID-19 within Massachusetts. Using data collected by the Massachusetts Government on PCR tests, we have developed a method to assess the current state of the pandemic and make recommendations for public health policy changes going forward.

The goal of this project is to create an epidemiological model that can accurately describe the spread of COVID-19 throughout Massachusetts. Namely, from the basic reproduction number which describes the virus' transmission potential, we can draw conclusions about COVID-19's long term implications.

Our MQP team was divided into two groups: The Mathematical Sciences Majors and the Data Science Majors. Therefore, our MQP report is divided into two parts. The mathematical sciences portion of this report is organized into seven sections. Section 2 introduces our epidemiological model used to describe the spread of COVID-19 among subpopulations of Massachusetts. Section 3 describes an analysis of the model where we proved the total population size remains constant over time, found the steady state solutions to the model's equations, and solved for the equation to the Basic Reproduction Number. In section 4, we calculate the transmission rate of the virus purely from Data. In section 5, we calculate the transmission rate of the virus using only our epidemiological model. Section 6 is a discussion of the resulting analysis from comparing the two transmission rates we solved for in sections 4 and 5. Section 7 describes several linear regression that provide insight into which COVID-19 mitigation techniques are most effective at preventing the spread of COVID-19 based on our previously calculated Basic Reproduction Number. In section 8, we recap what we have accomplished during this project and discus's the significance of our results.

2 Mathematical Model

The transmission diagram for this project is as follows:

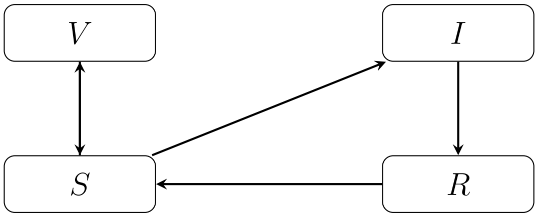


Figure 2.1: Transmission diagram for COVID-19. The compartments represent sub-populations of Massachusetts. The arrows represent which directions people can move between compartments

In Figure (2.1),

- S - People susceptible to the disease.
- I - People who are infected with the disease.
- V - People who are immune due to vaccination.
- R - People who are immune due to having recovered from the disease.

In the model of the pandemic, we use a SIVR compartmental model. This model was adapted from a SIR model created by Jana Kopfová, et. al. [27]. We built on their model by including a vaccinated population as well as accounting for waning efficacy of immunized people. This model tracks the number of people who fall into each compartment at any given time. The compartments are S, I, V, R. This model assumes homogeneous mixing of populations and a constant death rate in all compartments. Based on the above transmission diagram, the differential equations that describe it are as follows:

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta IS}{N} - v(t)S - \mu S + \lambda R + \alpha V + \mu N\\ \frac{dI}{dt} &= \frac{\beta IS}{N} - (\gamma + \mu)I\\ \frac{dV}{dt} &= v(t)S - \alpha V - \mu V\\ \frac{dR}{dt} &= \gamma I - \lambda R - \mu R\,.\end{aligned}$$

The parameters in the model are described below:

- v(t) - Vaccination rate, the number of people who become fully vaccinated per day.
- γ - Recovery rate, the time it takes for someone to recover from the virus.
- α - Waning vaccination rate, the time it takes for vaccine efficacy to wear off.
- λ - Waning natural immunity rate, the time it takes for someone's natural immunity from the virus to wear off.
- μ - The birth/death rate, the rate at which people enter and exit the model per day.
- β - The transmission rate, the probability given that when two people interact, the virus will be transmitted

3 Analysis of the Model

In order to assure the model is sound, some basic proofs were completed to demonstrate that the overall population remains constant, the steady states for the system of differential equations exist, and the basic reproduction number are derived.

Lemma 3.1. *The total population size $N = S + I + V + R$ is a constant independent of time.*

Proof. From the model equations, we have

$$\begin{aligned}\dot{N} &= \dot{S} + \dot{I} + \dot{V} + \dot{R}\\ &= \mu(N - S - I - V - R)\\ &= 0.\end{aligned}$$

Therefore, $N(t) = N(0) = N_0$. The proof of the lemma is complete. □

A steady state is a constant solution of the above model equations.

Lemma 3.2. *There are two steady state solutions to the model equations. The first steady-state is the disease-free state (DFS), where $I = 0$, and the second steady state is the endemic state, where the number of infected, I, is positive.*

Proof. To solve for the steady-states, we set the right side of the model equations to zero and find solutions to the resulting nonlinear system of algebraic equations.

$$\begin{aligned} -\frac{\beta IS}{N} - v(t)S - \mu S + \lambda R + \alpha V + \mu N = 0 \\ \frac{\beta IS}{N} - (\gamma + \mu)I = 0 \\ v(t)S - \alpha V - \mu V = 0 \\ \gamma I - \lambda R - \mu R = 0\,. \end{aligned}$$

Using the mathematical software Maple, we find the that DFS is

$$(S_1^*, I_1^*, V_1^*, R_1^*) = \left(\frac{N(\mu+\alpha)}{\alpha+\mu+v(t)}, 0, \frac{Nv(t)}{\alpha+\mu+v(t)}, 0\right)\,.$$

The endemic state is

$$(S_2^*, I_2^*, V_2^*, R_2^*) = \left(\frac{N(\gamma+\mu)}{\beta}, -\frac{N\bar{W}(\alpha+\mu)}{\beta(\alpha+\mu)(\mu+\gamma+\lambda)}, \frac{Nv(t)(\gamma+\mu)}{\beta(\alpha+\mu)}, -\frac{N\gamma\bar{W}}{\beta(\alpha+\mu)(\mu+\gamma+\lambda)}\right),$$

where

$$\bar{W} = -\alpha\beta + \alpha\gamma + \alpha\mu - \beta\mu + \gamma\mu + \gamma v(t) + \mu^2 + \mu v(t)\,.$$

Therefore, I_2^* and R_2^* are positive if and only if $\bar{W} < 0$. This is the same as

$$R_0 = \frac{\beta(\alpha+\mu)}{(\alpha+\mu+v(t))(\gamma+\mu)} > 1\,. \tag{3.1}$$

The proof of the lemma is complete. ∎

Remark 3.1. The number R_0 defined in the above proof is also called the basic reproduction number. Hence, the endemic state exists if and only if the basic reproduction number is larger than one. The proof of the lemma is complete.

In epidemiology, the basic reproduction number is the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. It is often denoted by the symbole R_0.

Lemma 3.3. *For our vaccination model, the basic reproduction number, R_0, is given by* (3.1).

Proof. We use the next generation method to find the basic reproduction number.

$$\begin{aligned}
\dot{I} &= \frac{\beta S I}{N} - (\gamma + \mu) I \\
\text{Let} \qquad F &= \frac{\beta S I}{N}, \quad V = (\gamma + \mu) I \\
\frac{\partial F}{\partial I} &= \frac{\beta S}{N}, \quad \frac{\partial V}{\partial I} = \gamma + \mu \\
F V^{-1} &= \frac{\beta S}{N} \frac{1}{\gamma + \mu} \, .
\end{aligned}$$

Evaluate the above expression FV^{-1} at the DFS, yields

$$\begin{aligned}
R_0 &= \frac{\beta}{N} \frac{N(\mu + \alpha)}{\alpha + \mu + v(t)} \frac{1}{\gamma + \mu} \\
&= \frac{\beta(\mu + \alpha)}{(\alpha + \mu + v(t))(\gamma + \mu)} \, .
\end{aligned}$$

The proof of the lemma is complete. □

4 Calculating Daily $\beta_d(t)$ from Data

This section is an explanation of how to find the transmission rate, β, for each day using only the data and some known parameter values. Subscript d in the state variables mean it is derived from the data.

4.1 Parameters

The following are parameters used in the model. They are taken from current scientific literature, adapted to our current situation.

Known Parameters:

$$\begin{aligned}
&v(t) = \text{Vector of newly vaccinated people for every day.} \\
&c(t) = \text{Vector of newly infected people for every day.} \\
&\lambda = \text{Waning natural immunity rate.} \\
&\mu = \text{Birth/death rate.} \\
&\gamma = \text{Recovery rate.} \\
&\alpha = \text{Waning vaccinated immunity rate.} \\
&N = \text{Total population} \, .
\end{aligned}$$

The vaccination rate, $v(t)$, and newly infected people, $c(t)$, are obtained from the data available online [21] [22]. The parameters λ, μ, γ and α are estimated from available research literature [17] [18] [19] [20].

The only remaining unknown parameter is $\beta(t)$. It is defined as the probability two people come in contact times the probability of infection resulting from the contact. $R_0(t)$ may be computed once $\beta(t)$ is found using (3.1).

4.2 Methodology

Suppose there are n days in the data set. Let $c(t)$ and $v(t)$ be the vectors of the number of newly infected and vaccinated individuals respectively. Four equations were created to find $S_d(t), I_d(t), V_d(t)$ and $R_d(t)$ solely from the data. The data set of newly infected persons from the Massachusetts Government Response Reporting Website begins on the first of June 2020 thus for $1 \leq t \leq n$, t denotes the tth day after June 1st 2020 [21].

$$\begin{aligned}
I_d(t) &= I_d(t-1) + c(t) - c(t - \frac{1}{\gamma}) - \mu I_d(t-1) \\
V_d(t) &= V_d(t-1) + v(t) - v(t - \frac{1}{\alpha}) - \mu V_d(t-1) \\
R_d(t) &= R_d(t-1) + c(t - \frac{1}{\gamma}) - c(t - \frac{1}{\lambda} - \frac{1}{\gamma}) - \mu R_d(t-1) \\
S_d(t) &= N - I_d(t) - V_d(t) - R_d(t)\,.
\end{aligned}$$

The infected, vaccinated and recovered population vectors are a function of four terms in order: The population the previous day, people entering the population (newly infected persons, newly vaccinated persons and newly recovered persons), people leaving the population for another compartment (recovering from infection, waning vaccinated immunity, waning natural immunity), and deaths. The parameters γ, α, and λ are rates and thus the reciprocals must be used to incorporate them. For example, α is the waning vaccinated immunity rate and therefore $\alpha = 1/$ (the number of days it takes for vaccinated immunity to wear off). Since N is constant, the susceptible population is trivially the other populations subtracted from N.

The initial condition is $I_d(1) = 326$ because that is how many people are infected on the first day of the data set from the Massachusetts Government Response Reporting Website. The initial condition of the susceptible compartment is $S_d(1) = N - I_d(1)$, and the assumption was made that the initial conditions for the vaccinated and recovered classes are $V_d(1) = R_d(1) = 0$. Using MATLAB, $I_d(t), V_d(t), R_d(t), S_d(t)$ may be found iteratively using the above formulas. Let $\mathbf{W}(t) = (S_d(t), I(t), V(t), R(t)),\ t = 1, ..., n$.

From the compartmental model, the function for movement from susceptible to the newly infected people can be represented as

$$c(t) = \frac{\beta_d(t)\, I_d(t)\, S_d(t)}{N}\,.$$

Rearranging,

$$\beta_d(t) = \frac{c(t)\,N}{I_d(t)\,S_d(t)}\,. \tag{4.1}$$

Using the previously obtained vectors for the susceptible and infected populations from MATLAB and the input function $c(t)$, a vector for $\boldsymbol{\beta_d} = (\beta_d(1), \beta_d(2), ..., \beta_d(n))$ was found from (4.1). The chart begins June 1st 2020 and ends 432 days later as shown in Figure 4.1.

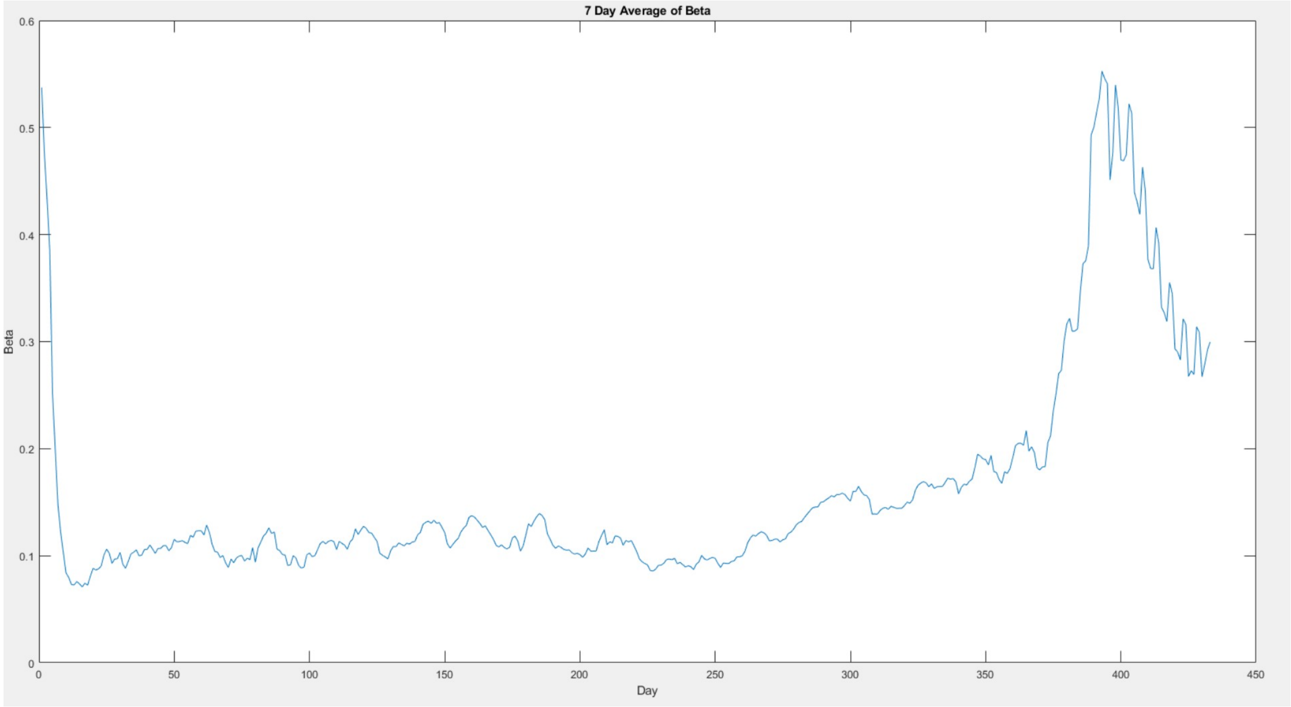


Figure 4.1: Plot of the beta vector, β_d, derived from data.

From the now acquired β_d, $R_0(t)$ may be found using the equation derived in Lemma 3.3.

$$R_0(t) = \frac{\beta_d(t)(\mu + \alpha)}{(\gamma + \mu)(\mu + v(t) + \lambda)} \tag{4.2}$$

The plot for $R_0(t)$ is found in Figure 4.2.

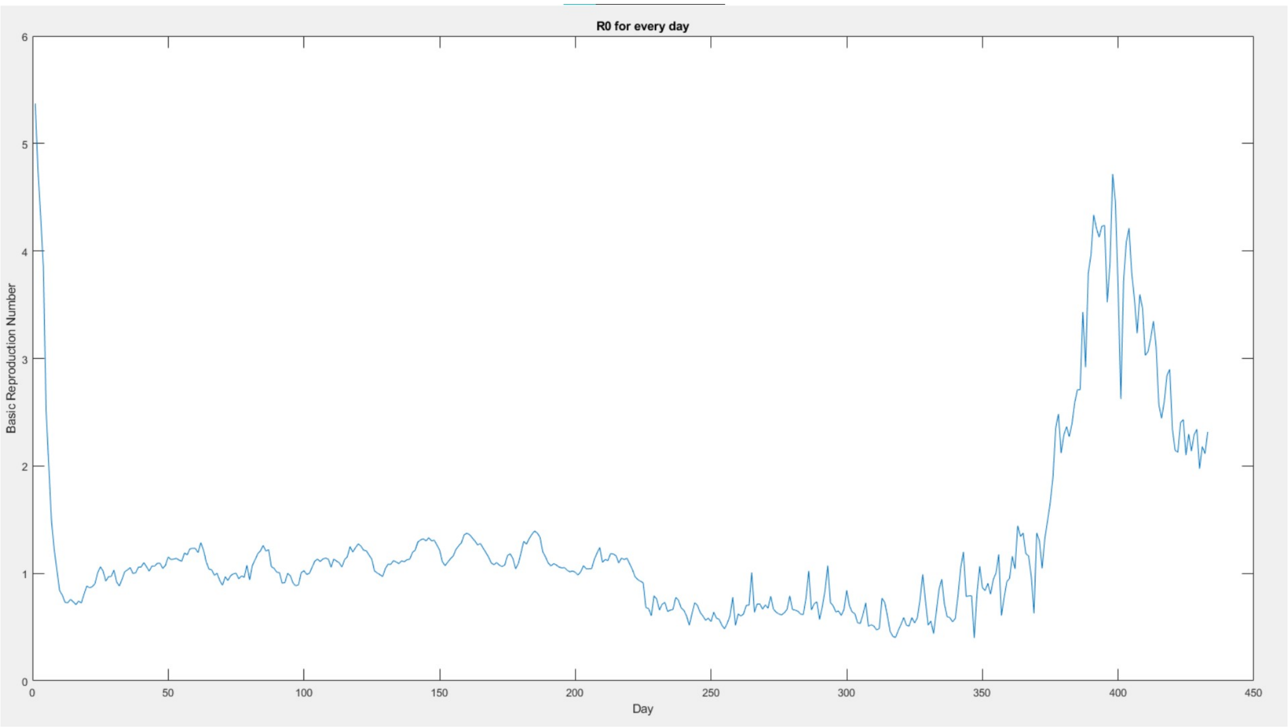


Figure 4.2: Plot of the basic reproduction number, $R_0(t)$, derived from data

5 Calculation of Daily $\beta_e(t)$ using the ODE Model

Using MATLAB, we created a program that finds a different daily β vector, $\beta_e(t)$ (subscript e means estimated), that matches $I_e(t)$ to $I_d(t)$, which are the infected populations predicted by MATLAB's ODE solver and from the data, respectively. From the data, we are given that

$$I_d(1) = 326$$

so we assume that

$$I_e(1) = 326$$

After the first day, the program will find a β value for each sequential day using the following logic:

We assume that for each day t, $\beta_e(t)$ lies in the interval $[0, 1]$, since it can be described as an aggregate probability. We partition $[0, 1]$ into 1000 equally spaced intervals

$$[b_j, b_{j+1}],\ b_j = (j-1) \times 0.001, j = 1, ..., 1001$$

such that $b_1 = 0$, $b_2 = 0.001$, and so on up to $b_{1001} = 1$. We defined $\mathbf{B} = [b_1, ..., b_{1001}]$, the vector of all possible values we are checking for β. (We did this so we don't need to use MATLAB's proprietary Optimization Toolbox.)

Using $\mathbf{W}_e(t=1)$, which we set to be the same as $\mathbf{W}_d(t=1)$, we use the MATLAB function ODE45 to solve the model equations in Section 2 to find $\mathbf{W}_e(t+1)$ using every value in $\mathbf{B}$ for every day t in our data. The value of b_j in $\mathbf{B}$ that minimizes the error function defined by (5.1) will be set to $\beta_e(t)$.

$$f(b) = \left(\log\left(\frac{I_e(t+1,b)}{I_d(t+1)}\right)\right)^2 . \tag{5.1}$$

The value of b among all components of $\mathbf{B}$ that minimizes $f(b)$ will provide the most accurate predictions of the infected population over time (I_e). We store each day's $\beta_e(t)$ in a vector $\boldsymbol{\beta}_e$. By graphing $\boldsymbol{\beta}_e$ against time, as shown in Figure 5.1, and adding lines for significant events affecting the transmission rate of COVID-19 in Massachusetts, we are able to draw conclusions about the effectiveness of specific lock down measures and/or restrictions. The analyses performed on the data are explained in greater detail in the Results section.

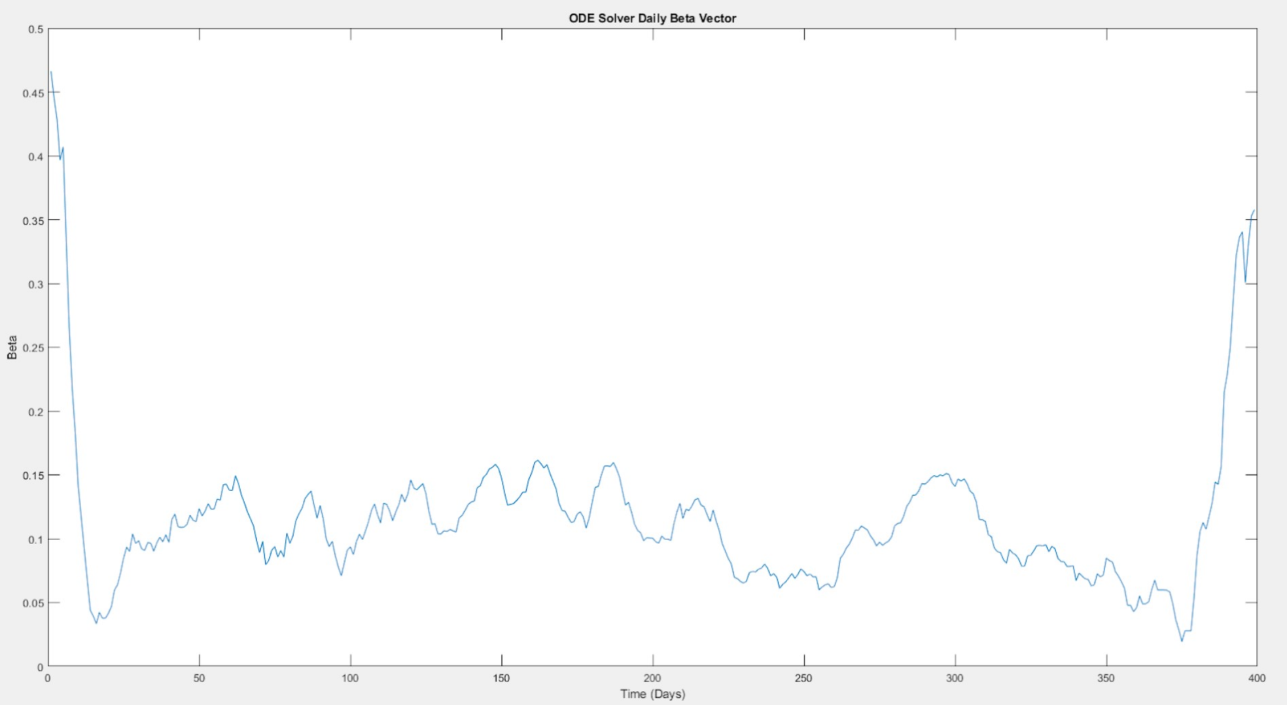


Figure 5.1: Plot of the daily $\boldsymbol{\beta}_e$ vector derived from the ODE Model

6 Results

Once the vectors for β_d and β_e were obtained, using (4.1) the vectors for $R_{0d}(t)$ and $R_{0e}(t)$ may be found from (3.1). The first vector, $R_{0d}(t)$, shown in Figure 4.2, describes the true development of the virus within Massachusetts based on the vaccination and newly infected data. The second vector, $R_{0e}(t)$, shown in Figure 5.1, describes the epidemiological compartmental model's projection of the virus predicated on the parameters input into the differential equations.

Upon comparing the two vectors, displayed in Figure 6.1, it was found that $R_{0d}(t)$ had a mean value of 0.9925 compared to $R_{0e}(t)$'s mean value of 1.0172.

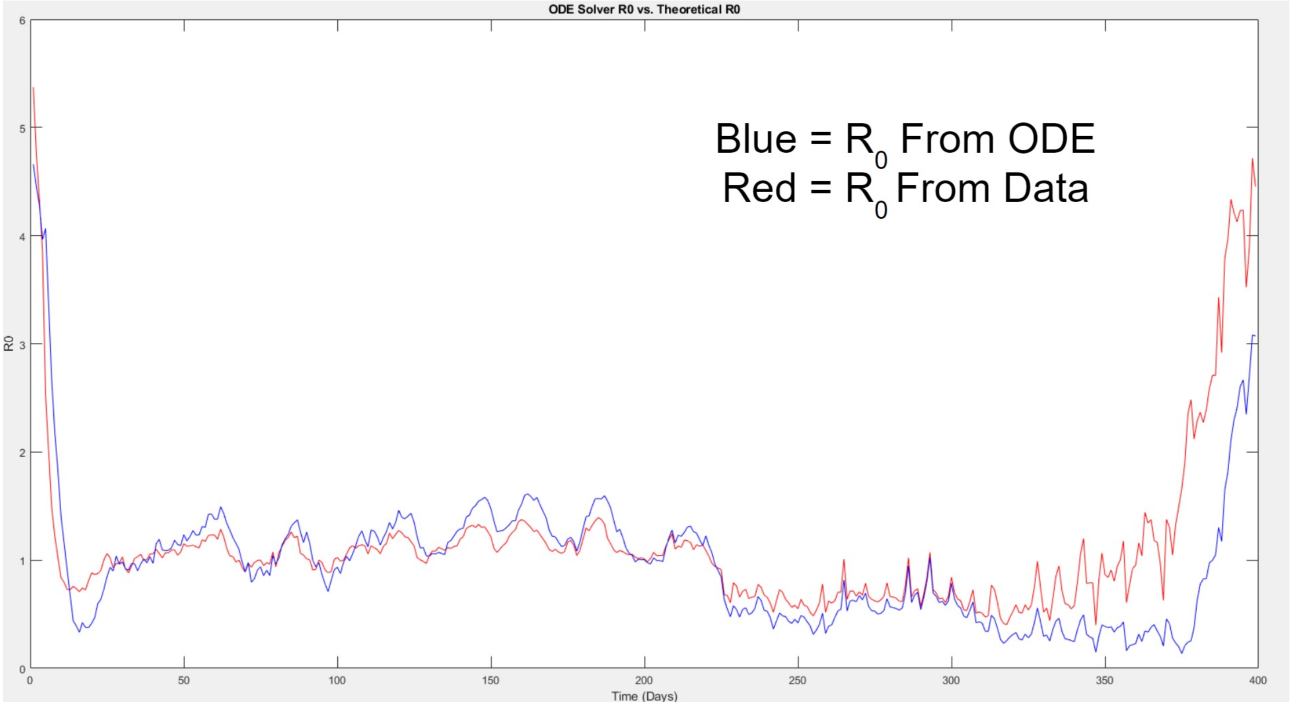


Figure 6.1: Comparison of the two basic reproduction number vectors

The root mean squared error (RMSE) between the vectors was calculated to be 0.1128 over the entire data set. The RMSE is under twelve percent of the desired vector's average value. We consider this method to produce a highly accurate model of COVID-19's dissemination throughout Massachusetts.

With an accurate model to emulate the real world data, the graph of the $R_0(t)$'s were partitioned at significant events to see if the events had any discernible effect on the spread of COVID-19. We first partitioned the graph by severity of restrictions mandated by the Massachusetts Government in Figure 6.2.

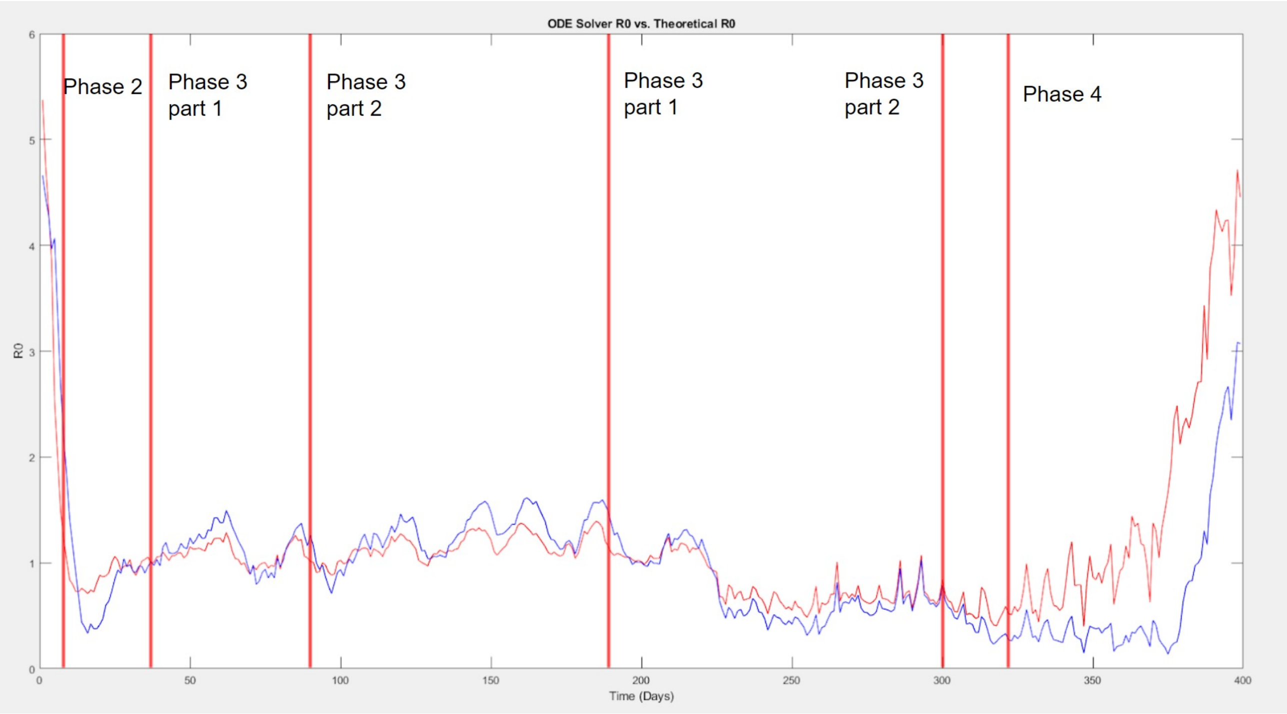


Figure 6.2: Plot of the two basic reproduction number vectors partitioned by phases of Massachusetts's COVID-19 reopening plan. Note: Phases are repeated because Massachusetts reverted to previous phases as COVID-19 cases increased.

This partition saw days being grouped based on which of the four phases in Massachusetts reopening were currently occurring. The restrictions implemented during each phase are as follows:

- Phase 1: Manufacturing facilities, construction sites, and places of worship were allowed to re-open. Hospitals were able begin to provide high priority preventative care, pediatric care and treatment for high risk patients.

- Phase 2: Retail, childcare facilities, restaurants (with outdoor table service only), hotels, and driving and flight schools were allowed to reopen. Youth and adult amateur sports were also allowed to resume.

- Phase 3 Part 1: Movie theaters, outdoor performance venues, museums, cultural and historical sites, fitness centers and health clubs and professional sports teams (without spectators) became eligible to reopen.

- Phase 3 Part 2: Indoor performance venues, indoor and outdoor exhibition and convention halls were allowed to reopen. Additionally, indoor dining is permitted.

- Phase 4: Indoor and outdoor stadiums, arenas, and ballparks were permitted to open at 12 percent capacity. Amusement parks, theme parks, and outdoor water parks also became eligible to reopen.

There was no noticeable discrepancy in the value of the $R_0(t)$s between the groups of this first partition as shown by the following root mean square error (RMSE) values for each partition, which never surpass an RMSE of 0.0344.

- Phase 2: 0.0655
- Phase 3 Part 1 (first occurrence): 0.0161
- Phase 3 Part 2 (first occurrence): 0.0216
- Phase 3 Part 1 (second occurrence): 0.03
- Phase 3 Part 2 (second occurrence): 0.0344
- Phase 4: 0.0056

The graph was then partitioned upon the presence of vaccines and the Delta variant becoming dominant in Massachusetts, shown in Figure 6.3.

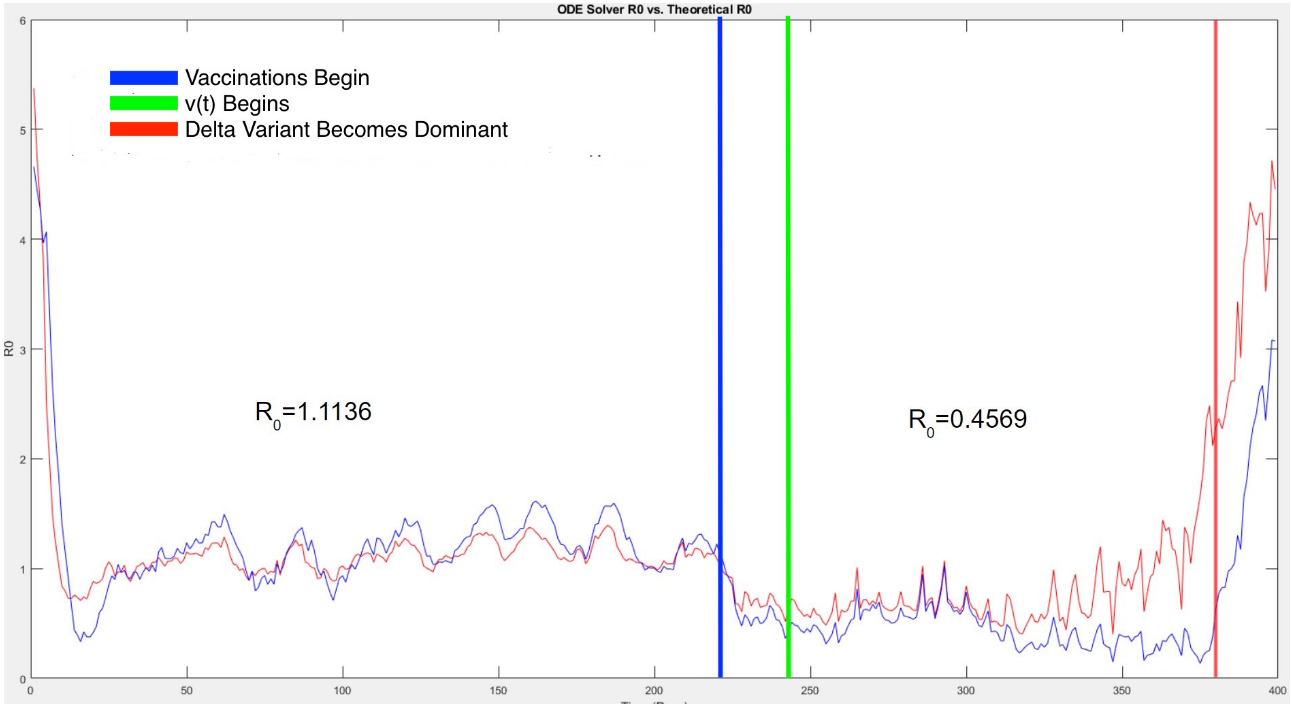


Figure 6.3: Plot of the two basic reproduction number vectors partitioned by important vaccine and COVID-19 variant dates. Note: Vaccinations start before v(t) because v(t) only counts Massachusetts residents that are fully vaccinated

This partition yielded a very recognizable difference in the value of $R_0(t)$ between the created intervals. The average value of $R_0(t)$ prior to vaccine introduction was calculated to be 1.1136 while subsequently, the value dropped by over 50% to 0.4569. Unfortunately, despite having a high percentage of the population vaccinated in Massachusetts, it is apparent that upon the Delta Variant becoming the dominant variant, $R_0(t)$ increased to exceed even its greatest value during the Alpha Variant prior to vaccine introduction.

Going forward, the model's time line will be expanded as more data comes out. Currently, based on the partitions of $R_0(t)$, it appears that governmental restrictions have had very little affect on COVID-19's Basic Reproduction Number. In contrast, $R_0(t)$s behavior subsequent to the introduction of the COVID-19 vaccine and upon the Delta Variant becoming prevalent suggest that the most impactful factors on COVID-19's development were vaccines and new virus variants.

7 Linear Regression Analysis

To further investigate the significance of COVID-19 mitigation techniques, we developed several linear regression models. The models determine the influence of each technique on the Basic Reproduction Number that we previously solved for. The tested independent variables are business closings, social gathering restrictions, the vaccination rate, mask mandates, seasonal changes, Massachusetts reopening phases, and the number of daily travelers through Logan International Airport.

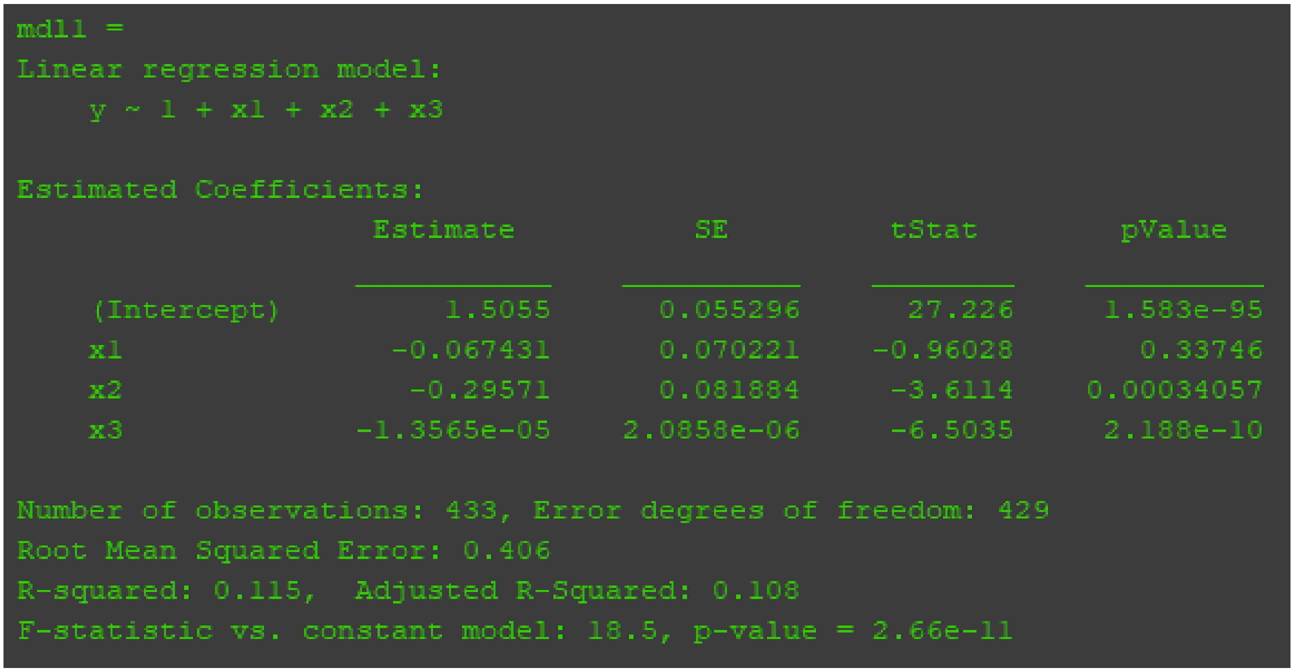
7.1 Preventive Measures Multiple Linear Regression

To analyze the influence of lockdown restrictions and other statewide preventive measures on the number of COVID-19 cases per day statewide, we used a multiple linear regression to determine which restrictions had the greatest and least effects on daily cases. The basic reproductive number, computed daily, ($R_0(t)$) was the response variable, calculated using data from the CDC [23], [24].To calculate $R_0(t)$, we used (3.1) and the same set of constant parameters used throughout the project. For the regression model, we initially considered the following:

$$y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \tag{7.1}$$

$$\begin{aligned}
X_1 &= \text{Business Closings/Staffing Limits (Binary variable)}\\
X_2 &= \text{Gathering/Social Event Restrictions (Binary variable)}\\
X_3 &= \text{Newly Vaccinated Population per Day (Quantitative variable)}
\end{aligned}$$

We tested this model over the period from Jun. 1, 2020 to Aug. 31, 2021. The intervals in which X_1 and X_2 were set to 1 were determined from the "Reopening Massachusetts" web page on ma.gov [25], which describes the phased plan for the state to reopen businesses, social events, etc. The regression showed that all predictors except for X_1, with a p-value of 0.3375, are statistically significant in the model. X_3 returned a p-value of order 10^{-10} and had a coefficient of -1.3565*10^{-5}, showing that vaccination rate had a highly significant impact on reducing R_0, considering the values of X_3 reach into the tens of thousands at times. The detailed results from this regression model are below.



```
mdl1 =
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                    Estimate         SE          tStat        pValue
                   ___________   __________   ________   __________

    (Intercept)         1.5055     0.055296     27.226    1.583e-95
    x1               -0.067431     0.070221   -0.96028      0.33746
    x2                -0.29571     0.081884    -3.6114   0.00034057
    x3             -1.3565e-05   2.0858e-06    -6.5035    2.188e-10

Number of observations: 433, Error degrees of freedom: 429
Root Mean Squared Error: 0.406
R-squared: 0.115,  Adjusted R-Squared: 0.108
F-statistic vs. constant model: 18.5, p-value = 2.66e-11
```

Figure 7.1: $y = 1.5055 - 0.0674X_1 - 0.2957X_2 - 1.3565 \times 10^{-5}X_3$

Due to the insignificance of X_1, we removed it from the model to improve the fit. This new model can be represented by the following equation:

$$y = b_0 + b_1X_1 + b_2X_2 \tag{7.2}$$

X_1 = Gathering/Social Event Restrictions (Binary variable)
X_2 = Newly Vaccinated Population per Day (Quantitative variable)

With this model, all variables were statistically significant, with p-values of order 10^{-9} or smaller. In addition, both models return an intercept approximately equal to 1.5, indicating that without preventive measures being taken, the R_0 value would have a base value of approximately 1.5: an endemic state. This simpler model returned the equation $y = 1.4950 - 0.3517X_1 -$

$1.4247 \times 10^{-5} X_2$, with all variables returning extremely small p-values. The R^2 value for this adjusted model was 0.113, which prompted further changes.

To attempt to improve the R^2 value, we added to and modified some of our predictor variables. A large number of outliers also existed in the R_0 data from the lack of weekend testing in Massachusetts, so we treated the data with a 7-day average to reduce noise and thus improve the model fit. In addition, we decided to test these predictors against the daily number of COVID-related deaths in the state, and the 7-day average of this set. The final change made to the previous models is that X_5 (Newly Vaccinated Population per Day) was replaced with the vaccination rate, defined as the newly vaccinated population divided by the current susceptible population for each day.

The final set of predictor and response variables in the model are as follows:

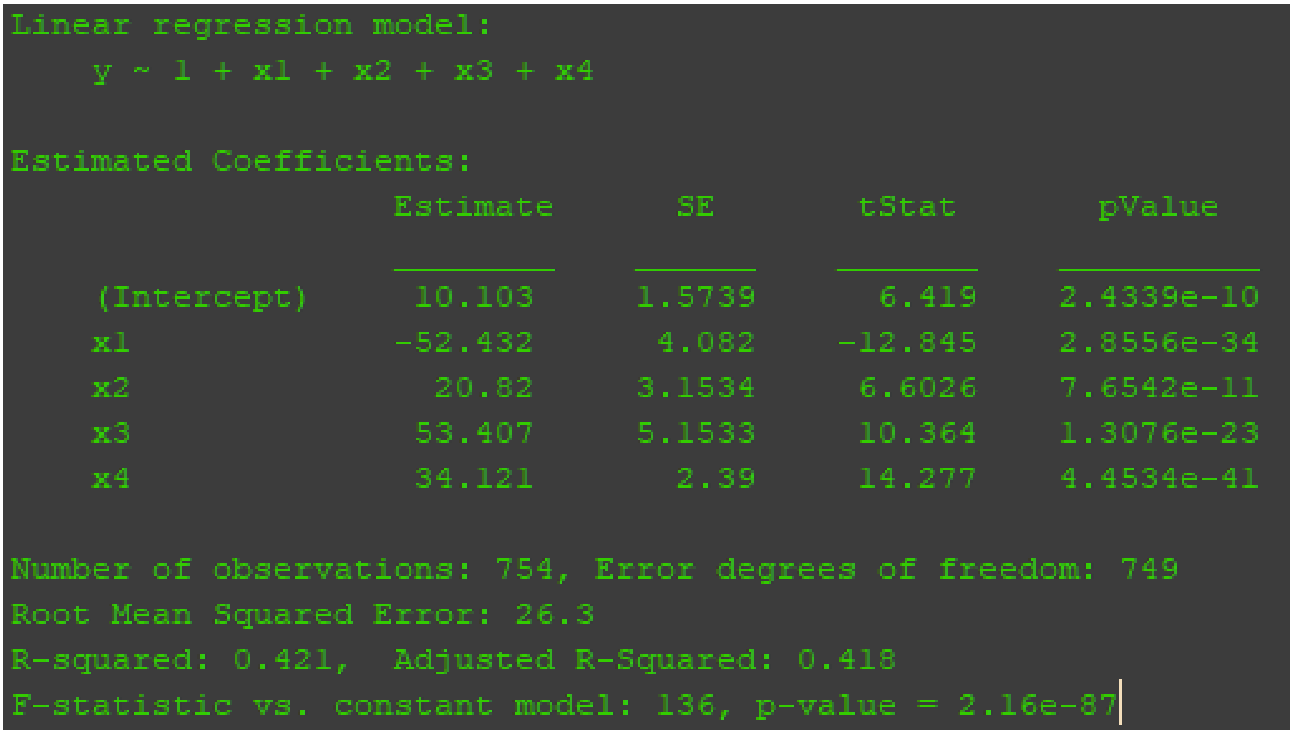
$$y_i = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 \tag{7.3}$$

$X_1 =$ Public Mask Mandate (Binary)
$X_2 =$ Gathering/Social Event Restrictions (Binary)
$X_3 =$ Business Closings/Staffing Limits (Binary)
$X_4 =$ IsWinter? (Binary)
$X_5 =$ Vaccination Rate (Proportion)
$y_1 = R_0$ (Quantitative)
$y_2 =$ 7-Day Average of y_1 (Quantitative)
$y_3 =$ Daily COVID Deaths (Quantitative)
$y_4 =$ 7-Day Average of y_3 (Quantitative)

In addition to modifying the model, we collected a much larger set of data points to operate on, spanning from February 1, 2020 until February 23, 2022. The date ranges in which X_1, X_2, and X_3 are set to 1 was obtained from Massachusetts' "State of Emergency" web page [26], where all lockdown phases and COVID-related governor's orders are archived and dated. X_4 is set to 1 between the days of December 21 and March 20, which are the winter solstice and spring equinox, respectively, and set to 0 during the rest of the year. The results of these regression models are:

$$y_2 = 1.5380 - 0.4580X_1 + 0.3326X_2 - 0.3242X_3 - 0.3459X_4 - 10.4250X_5$$
$$y_4 = 10.386 - 52.413X_1 + 19.275X_2 + 54.669X_3 + 34.042X_4 - 189.511X_5$$

The only issue with these models is the significance of the vaccination rate (X_5). The best-performing model using these variables is against y_2, since the p-value associated with X_5 is approximately 0.08. However in this model, the predictors are only able to account for 24% of the response variance. In the model against y_4, the p-value for X_5 is not within any reasonable tolerance, with a value of 0.477. As a response, we removed X_5 from the model for y_4, since it is adding excessive error. The resulting model returned only significant p-values, indicating high accuracy, yet we were still only able to account for 42% of the variance in the 7-day average of daily deaths. In addition, this model did not contain any data about vaccinations, so no conclusions about the effects of vaccines on death rate can be made using this particular model. The results of this final regression are below:



```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue  
                   ________    ______    _______    __________

    (Intercept)     10.103     1.5739      6.419    2.4339e-10
    x1             -52.432      4.082    -12.845    2.8556e-34
    x2               20.82     3.1534     6.6026    7.6542e-11
    x3              53.407     5.1533     10.364    1.3076e-23
    x4              34.121       2.39     14.277    4.4534e-41


Number of observations: 754, Error degrees of freedom: 749
Root Mean Squared Error: 26.3
R-squared: 0.421,  Adjusted R-Squared: 0.418
F-statistic vs. constant model: 136, p-value = 2.16e-87
```

Figure 7.2: $y_4 = 10.103 - 52.432X_1 + 20.82X_2 + 53.407X_3 + 34.121X_4$

In this model, all coefficients' t-statistics as well as the F-statistic have extremely small p-values, indicating high accuracy. While the R^2 value is still lower than is desirable, little other reliable daily data was available to add to the model as predictors.

7.2 Logistic regression

To further evaluate the effectiveness of various COVID-19 mitigation techniques, we created a logistic regression to determine which techniques prevented an endemic state. Although an endemic state means there is a low level of the virus present, it implies the virus is still growing and will therefore require annual vaccinations, and it has a risk of leading to a pandemic. A logistic regression models the probability of a categorical response variable given an input.

The basic reproduction number, R_0, was used as the binary response variable with a threshold of one, so when $R_0 \geq 1$, the response variable was set to class 1, otherwise it was set to class 0. This regression would indicate which mitigation techniques were significant in preventing the spread of COVID-19 and an endemic state.

For the model, we initially considered evaluating travel restrictions, mask mandates, Massachusetts reopening phases, and vaccinations; however, our preliminary findings showed that the reopening phases had no significant impact on R_0. After subset selection the model is as follows:

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 \tag{7.4}$$

$y = R_0$ (Binary variable)

$X_1 =$ Number of people traveling through Logan International Airport (Quantitative variable)

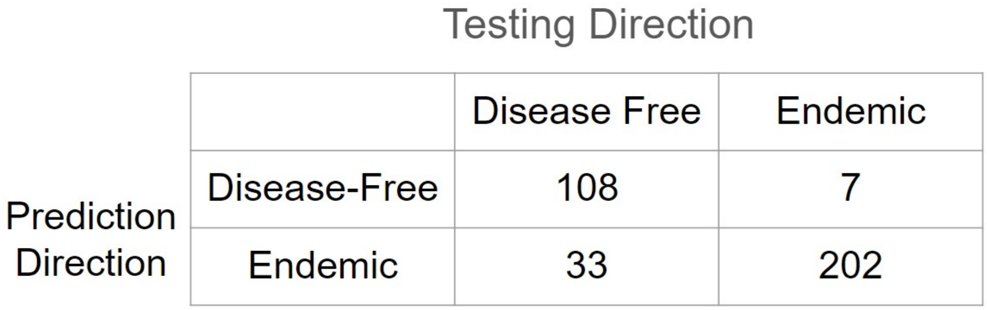
$X_2 =$ Mask mandate (Binary variable)

$X_3 =$ Vaccinations per day (Quantitative variable) .

The results of our regression, demonstrated in the equation below, showed that the number of travelers in Logan International Airport (X_1) has a statistically significant impact on the response variable. The p-value of X_1 was 0.000122. Also, the vaccination rate was highly significant as expected, with a p-value of 6.24e-15.

$$y = 1.053 + 1.774 \times 10^{-4} X_1 - 1.183 X_2 - 2.571 \times 10^{-4} X_3 \tag{7.5}$$

Using Leave One Out Cross Validation (LOOCV), the logistic regression model was tested for accuracy. LOOCV validates the model by determining the models coefficients using all but one data point (one day of data). The model then tests if it categorizes the data point that was left out correctly as either leading to an endemic state or disease-free state. This is repeated for all 350 days used to compute the model. The validation set yielded a 23% error rate when $R_0 <$ 1 and a 3% error rate when $R_0 >= 1$. The resulted in a 11% error rate overall for the validation set. The error rates refer to what percent of the time the model incorrectly categorizes a data point. The LOOCV method was used because our data set is relatively small (less than 1000 data points). This allows for us to use as much training data as possible when fitting the model.



		Testing Direction	
		Disease Free	Endemic
Prediction Direction	Disease-Free	108	7
	Endemic	33	202

Figure 7.3: Logistic regression's performance on the validation set. The correct prediction is $310/350 = 88.6\%$.

8 Math Group Conclusion

The goal of this project was to solve for the Basic Reproduction Number (R_0) of COVID-19 in order to make conclusions about the direction of the pandemic, and evaluate the methodology that has been used to mitigate its spread. We developed both an equation for R_0, as well as an epidemiological compartmental model, which were both initially missing a single parameter: β. Using available data on reported COVID-19 cases and vaccinations, we developed a theoretical β value that was derived solely from the data. We then forced our model to find a β value that would evolve its infected population to match the data.

This method yielded a β value for every single day for which we have data. Using our equation for R_0, we then solved for a solution to R_0 for every single day. We evaluated the accuracy of our R_0 found using the model against the R_0 derived solely from the data. The root mean square error between the two R_0 values is 0.0695, thus it can conclude that our model has a high degree of accuracy.

To further evaluate the state of the pandemic, we created linear regression models to evaluate which COVID-19 mitigation techniques had the greatest impact on minimizing R_0 and denying an endemic state, In addition, we used the same regression models with daily deaths as the response to determine the effects of these techniques on the death toll in Massachusetts. A multiple linear regression using R_0 as the response variable concluded that vaccinations are overwhelmingly influential in reducing R_0, while government-imposed restrictions tended to have little to no power to reduce the spread of COVID-19. The models using deaths as a response showed almost identical trends, although we had to exclude vaccination rates in one model. An additional predictor variable was added to the model to determine if winter also had an effect on R_0 and on deaths, and winter was shown to sharply increase both R_0 and average daily death count.

Also, a logistic regression was created with a binary R_0 as the response variable, indicating either an endemic ($R_0 \geq 1$) or disease-free state ($R_0 < 1$). Similar to the multiple linear regression, the conclusion demonstrated that vaccinations are the strongest mitigation technique, accounting for a majority of the response variable's variance. This model also displayed that the number of passengers travelling through Logan International Airport (used to quantify travel activity) was positively correlated with the response. All other tested mitigation techniques were deemed insignificant at preventing an endemic state.

9 Data Science Group

In each term, we pursued different project goals. In our first term, we focused on data collection, data visualization, and statistical analysis on positive COVID-19 cases and vaccination statuses in Massachusetts. We analyzed and showed how positive cases and administered vaccination changed over time, and we provided useful information for the math group to fit their model. Additionally, we researched to find the unknown parameters in the SIVR COVID model: birth and death rate μ, recovery rate γ, waning immunity for vaccinated individuals α, and recovered waning immunity λ.

During B term, we attempted to create a model that can reasonably forecast Massachusetts COVID cases ahead of time. After studying possible models, we decided on using the ARIMA time-series model as a method of forecasting future COVID cases in Massachusetts.

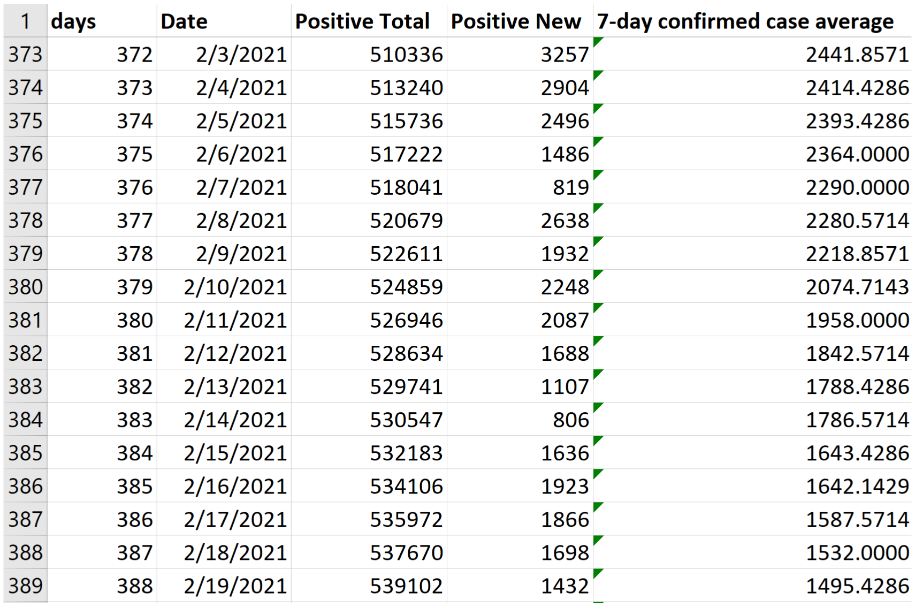
Finally, during C term, we investigated each state's COVID spread during given times. We wanted to do so in order to identify states that had low COVID infections, and more importantly find out why these states had lower COVID spreads. To do so, we clustered COVID rates over each month and analyzed each of the generated groups. After clustering was performed, we applied statistical analysis to find why there might be different clusters.

10 Data set

For A and B terms, we used the Covid-19 cases data set from Massachusetts government website. The dataset provides the following:

1. Daily COVID-19 cases by report date, as well as by testing date.
2. Pfizer and Moderna vaccination data

However, in B term we decided to start our time series data from April 1, 2020 since there were too few and unreliable cases before April due to low access to testing. To make data more smooth and generate a better forecast, we chose to use the 7-day moving average data. There are 542 data points in our data set. A snippet of the dataset we used can be seen in Figure 10.1. (https://www.mass.gov/info-details/massachusetts-covid-19-vaccination-data-and-updates)



1	days	Date	Positive Total	Positive New	7-day confirmed case average
373	372	2/3/2021	510336	3257	2441.8571
374	373	2/4/2021	513240	2904	2414.4286
375	374	2/5/2021	515736	2496	2393.4286
376	375	2/6/2021	517222	1486	2364.0000
377	376	2/7/2021	518041	819	2290.0000
378	377	2/8/2021	520679	2638	2280.5714
379	378	2/9/2021	522611	1932	2218.8571
380	379	2/10/2021	524859	2248	2074.7143
381	380	2/11/2021	526946	2087	1958.0000
382	381	2/12/2021	528634	1688	1842.5714
383	382	2/13/2021	529741	1107	1788.4286
384	383	2/14/2021	530547	806	1786.5714
385	384	2/15/2021	532183	1636	1643.4286
386	385	2/16/2021	534106	1923	1642.1429
387	386	2/17/2021	535972	1866	1587.5714
388	387	2/18/2021	537670	1698	1532.0000
389	388	2/19/2021	539102	1432	1495.4286

Figure 10.1: Data points in *CasesByDate(Test Date)*

11 Time Series Model

The first model we considered studying was the time-series model, specifically an Auto Regressive (Inegrating) Moving Average (ARIMA) model. The ARIMA model is one that makes next-step predictions based on previous data from the series. This model tends to be reasonably accurate for short-term predictions and is good for modeling complex systems without needing to know all the underlying variables. This is because the only input for ARIMA models are previous datapoints from the time series.

ARIMA models have hyperparameters p, d, and q, where p is the number of autocorrelated terms, q is the number of moving average terms, and d is the order of differencing applied to the data. Hyperparameters are parameters that are external to the model and cannot be estimated from the data itself; they are parameters that need to be set before the model learning process begins.

Let y_t, $t = 1, 2, 3, ...$ be a time series. Then an ARIMA(p,d,q) model is of the form:

$$y_t = c + \sum_{i=1}^{p} \phi_i \cdot y_{t-i} \; + \sum_{i=1}^{q} \theta_i \cdot \epsilon_{t-i}$$

Here y_t is the predicted value of the time series at time t, and y_{t-i} refers to the actual value of the time series at time $t - i$. ϕ_i refers to the auto-correlation coefficient, θ_i refers to the moving average coefficient, and ϵ_{t-i} refers to the prediction error on day $t - i$. The constant c is the expected average of the time series. This ARIMA model was the model of choice for the B-term study. The parameters that the ARIMA model will produce are θ_i and ϕ_i.

12 ARIMA(p,d,q) Model

12.1 Introduction to ARIMA model

To review, ARIMA (autoregressive integrated moving average) models have parameters p, d, and q, and they follow the equation:

$$y_t = c + \sum_{i=1}^{p} \phi_i \cdot y_{t-i} \ + \sum_{i=1}^{q} \theta_i \cdot \epsilon_{t-i}$$

The $\sum_{i=1}^{p} \phi_i \cdot y_{t-i}$ portion of the model is the autoregressive portion, where the model uses lagged data y_{t-i} and multiplies it by some coefficient ϕ_i to forecast y at time t. The $\sum_{i=1}^{q} \theta_i \cdot \epsilon_{t-i}$ portion of the model is the moving average part. This is where the model multiples ϵ_{t-i} by a coefficient θ_i.

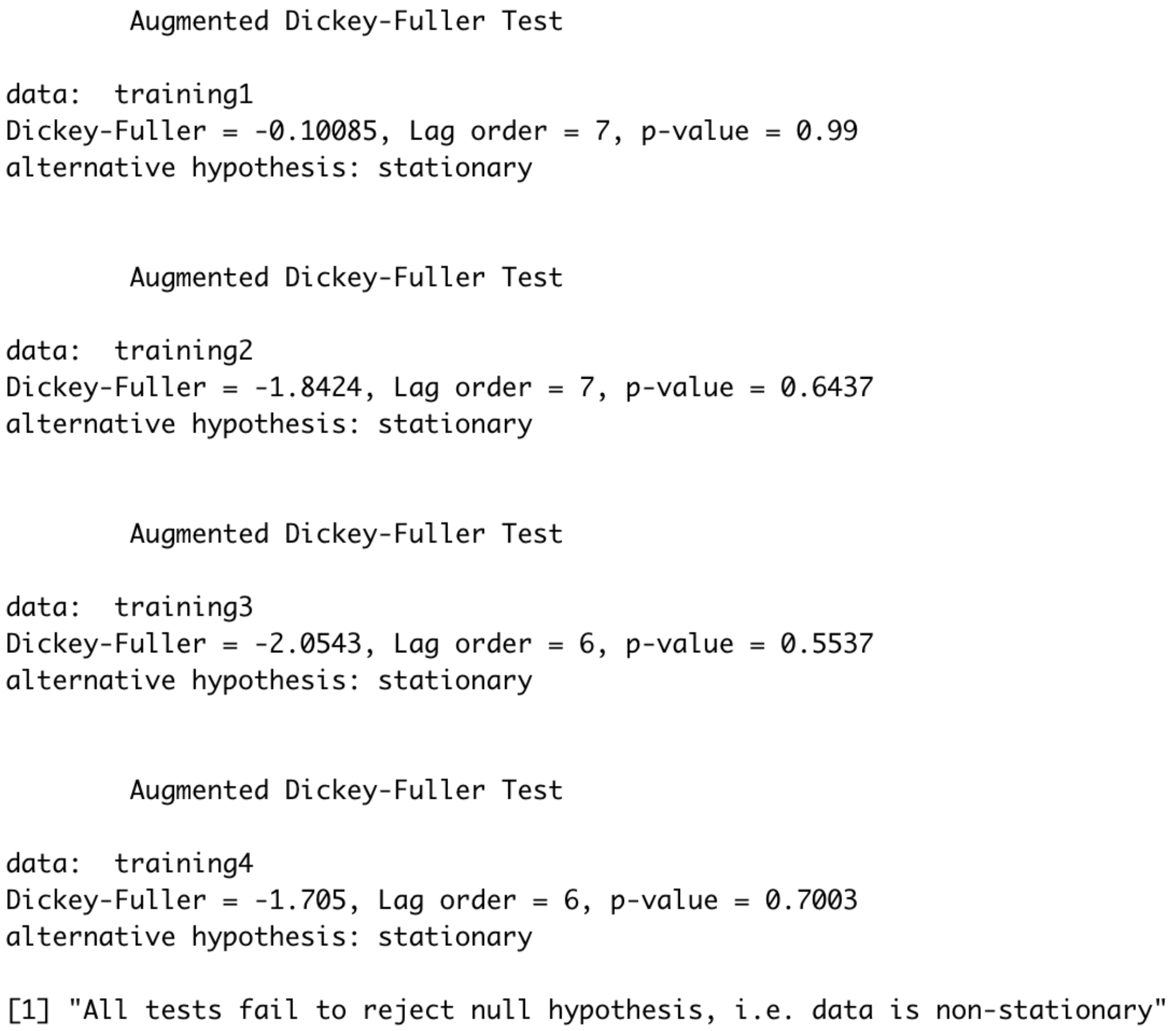
Hyperparameters p and q determine the order of the autoregressive and moving average parts of the model, respectively. Hyperparameter d determines the amount of differencing applied to the series. Each of these hyperparameters are chosen based on a variety of factors; auto and partial autocorrelation can help determine a viable p and q, while differencing order d can be applied to make the time series stationary. We shall discuss them below.

12.2 Stationarity

In order to get a reliable ARIMA model, the time series data must first be made stationary. A time series has stationarity if a shift in time doesn't cause a change in the shape of the distribution. Basic properties of the distribution like the mean, variance, and covariance are constant over time.[1]

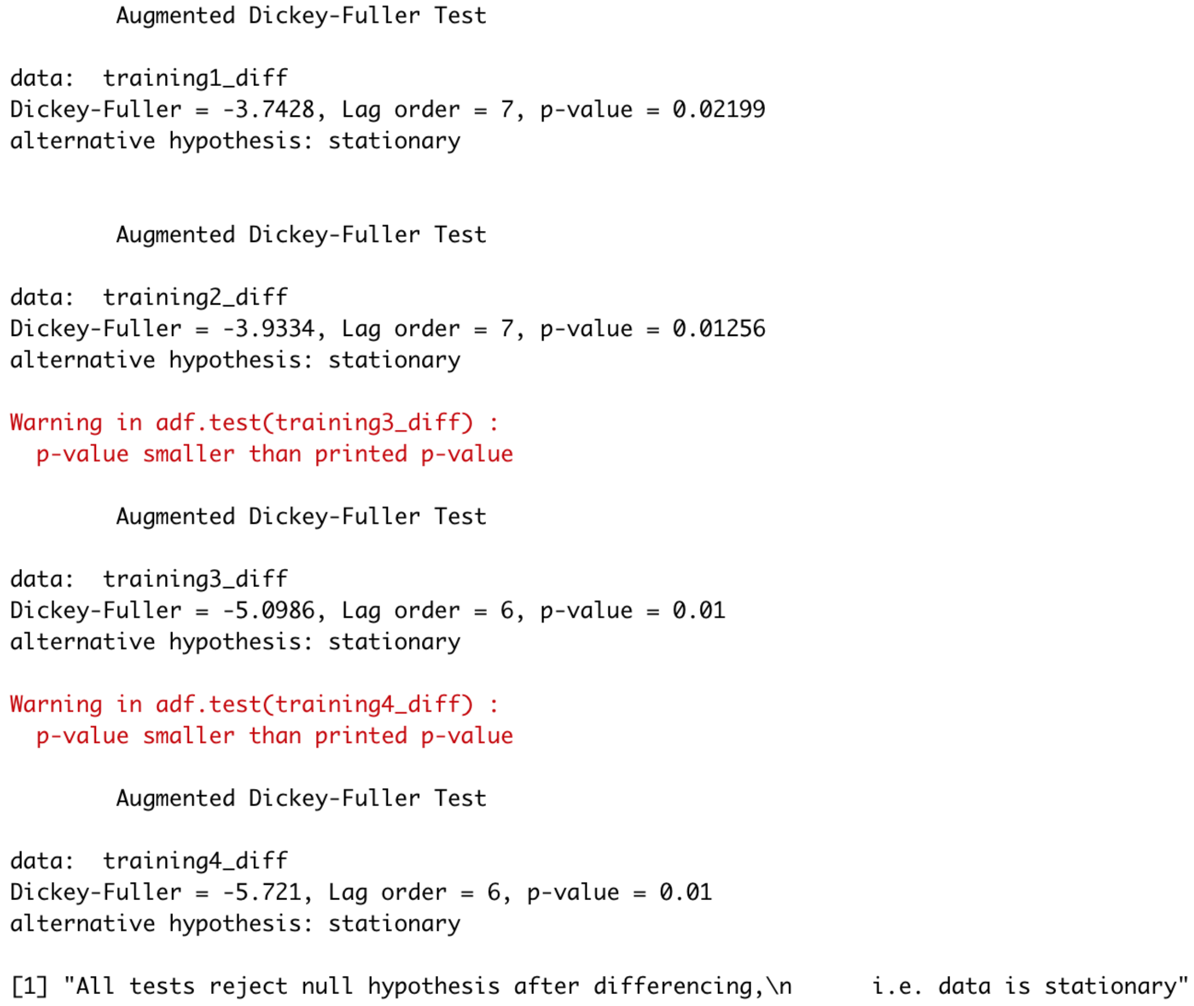
The *unit root tests* are statistical tests often employed to detect non-stationarity in time series. One such unit root test that we used is called the Augmented Dickey Fuller (ADF) test. ADF Test is a common statistical test used to test whether a given Time series is stationary or not[2]. Its null hypothesis declares that the time series has a unit root and is therefore non-stationary. The alternative hypothesis is that there is no unit root and the time series is stationary.

Upon trying the ADF test on our daily-new COVID cases series, we fail to reject the null hypothesis, as shown in Figure 12.1. Thus, we apply differencing once and re-test. Upon testing after applying difference, we are able to reject the null hypothesis and proceed with a stationary time series, as seen in Figure 12.2 .



```
        Augmented Dickey-Fuller Test

data:  training1
Dickey-Fuller = -0.10085, Lag order = 7, p-value = 0.99
alternative hypothesis: stationary


        Augmented Dickey-Fuller Test

data:  training2
Dickey-Fuller = -1.8424, Lag order = 7, p-value = 0.6437
alternative hypothesis: stationary


        Augmented Dickey-Fuller Test

data:  training3
Dickey-Fuller = -2.0543, Lag order = 6, p-value = 0.5537
alternative hypothesis: stationary


        Augmented Dickey-Fuller Test

data:  training4
Dickey-Fuller = -1.705, Lag order = 6, p-value = 0.7003
alternative hypothesis: stationary

[1] "All tests fail to reject null hypothesis, i.e. data is non-stationary"
```

Figure 12.1: Time series before differencing



```
        Augmented Dickey-Fuller Test

data:  training1_diff
Dickey-Fuller = -3.7428, Lag order = 7, p-value = 0.02199
alternative hypothesis: stationary


        Augmented Dickey-Fuller Test

data:  training2_diff
Dickey-Fuller = -3.9334, Lag order = 7, p-value = 0.01256
alternative hypothesis: stationary

Warning in adf.test(training3_diff) :
  p-value smaller than printed p-value

        Augmented Dickey-Fuller Test

data:  training3_diff
Dickey-Fuller = -5.0986, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

Warning in adf.test(training4_diff) :
  p-value smaller than printed p-value

        Augmented Dickey-Fuller Test

data:  training4_diff
Dickey-Fuller = -5.721, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

[1] "All tests reject null hypothesis after differencing,\n      i.e. data is stationary"
```

Figure 12.2: Time series after differencing

12.3 Choosing parameters p and q

There are several methods for choosing hyperparameters p and q. A common way to choose the order of the p is to generate a Partial Auto-Correlation Function (PACF) plot, while using the Auto-Correlation Function (ACF) plot to determine the order of q.

To choose the order p from the PACF plot, you can set p equal to the number of lags where the partial auto-correlation is outside the significance level. To choose the order q from the ACF plot, you can set q equal to the number of lags where the auto-correlation sharply decreases.

Another way to choose p and q is to use the Extended Auto-Correlation Function (EACF) table. From the generated EACF table, we can choose a p and q where the table marks a circle, indicating a p and q combination candidate.

Finally, a strong method to find an optimal p and q is to simulate many different ARIMA models with different p and q, and select the one that returns the best Akaike's Information Criterion (AIC). The AIC is an error metric that combines the training error with the model complexity. Lower AIC is better, as it's best to have lowest possible training error while maintaining lower model complexity. This can be done by using Auto.ARIMA(), which is a function in R that returns the best ARIMA model and p and q based on the dataset.

13 Forecasting

Our goal of forecasting is to produce accurate predictions of COVID cases as far as a week in advance. Being able to do so would be highly valuable as it could prepare medical staff and COVID prevention measures if COVID cases are predicted to rise.

13.1 Modeling Positive Cases: Training and Testing Sets)

In order to test our ARIMA model's accuracy on testing data, we decided to partition our dataset into training and testing set. We also decided to do k-fold cross-validation, and create 4 different training and test splits:

1. For the first training and test split, we trained the ARIMA model to the first 80 percent of the data, and tested the model against the remaining 20 percent of the data.

2. For the second training and test split, we trained the model to the first 70 percent and tested against the remaining 30 percent.

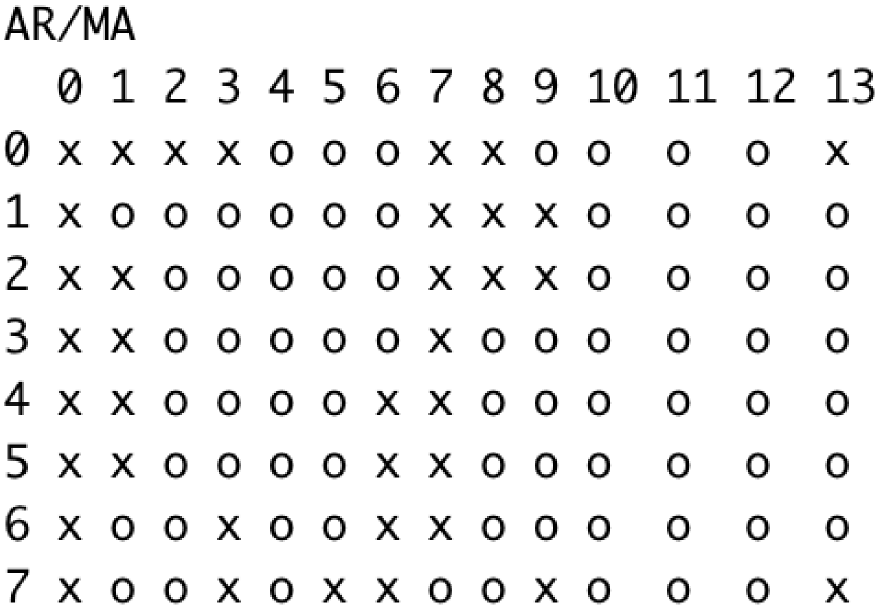
3. For the third training and test split, we trained the model to the first 60 percent and tested against the remaining 40 percent.

4. For the fourth and final training and test split, we trained the model to the first 50 percent and tested against the remaining 50 percent.

13.2 Modeling Positive Cases: Choosing Model Parameters

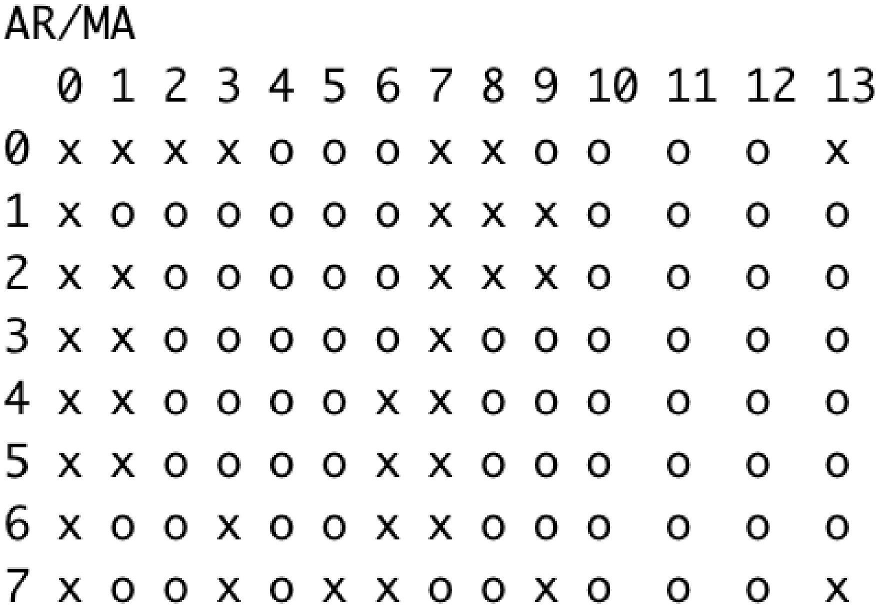
First, we tried picking p and q based on the EACF plots (Figure 13.1 to Figure 13.4) for each training set. To pick p and q, we examined all possible combinations based on the EACF table, and then put them into the ARIMA model. We then compare their AICc value: the lower the better.AICc is AIC for small data set to avoid overfitting. Finally, we pick p and q of (2, 1, 2) for all training set.

Additionally, we wanted to compare our ARIMA models with the ARIMA models chosen by the auto.arima() function. Auto.arima() selected the hyperparameters to be (9, 1, 6), (2, 1, 0), (2, 1, 0), and (3, 2, 1) for training sets 1 through 4 respectively.



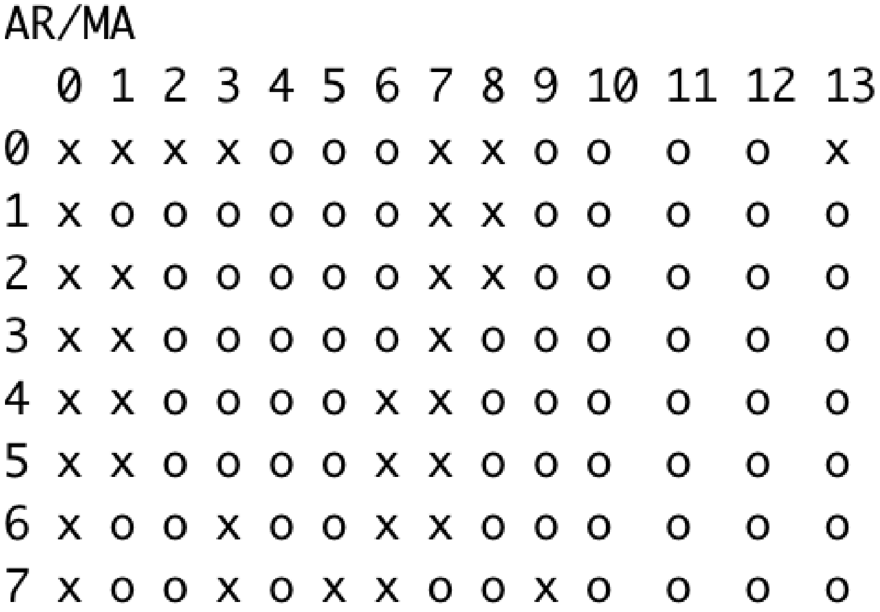
```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x o o o x x o o  o  o  x
1 x o o o o o o x x x o  o  o  o
2 x x o o o o o x x x o  o  o  o
3 x x o o o o o x o o o  o  o  o
4 x x o o o o x x o o o  o  o  o
5 x x o o o o x x o o o  o  o  o
6 x o o x o o x x o o o  o  o  o
7 x o o x o x x o o x o  o  o  x
```

Figure 13.1: EACF for training set 1



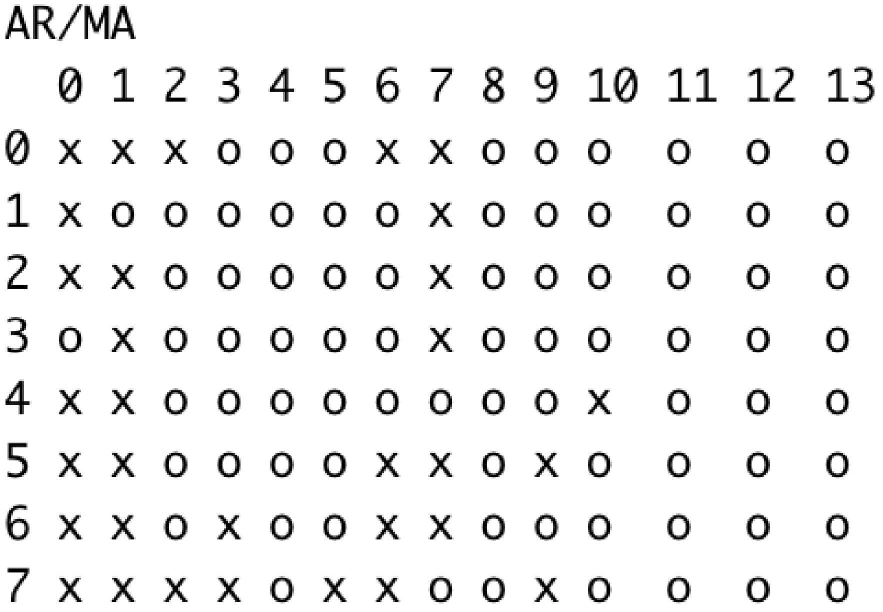
```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x o o o x x o o  o  o  x
1 x o o o o o o x x x o  o  o  o
2 x x o o o o o x x x o  o  o  o
3 x x o o o o o x o o o  o  o  o
4 x x o o o o x x o o o  o  o  o
5 x x o o o o x x o o o  o  o  o
6 x o o x o o x x o o o  o  o  o
7 x o o x o x x o o x o  o  o  x
```

Figure 13.2: EACF for training set 2



```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x o o o x x o o  o  o  x
1 x o o o o o o x x o o  o  o  o
2 x x o o o o o x x o o  o  o  o
3 x x o o o o o x o o o  o  o  o
4 x x o o o o x x o o o  o  o  o
5 x x o o o o x x o o o  o  o  o
6 x o o x o o x x o o o  o  o  o
7 x o o x o x x o o x o  o  o  o
```

Figure 13.3: EACF for training set 3



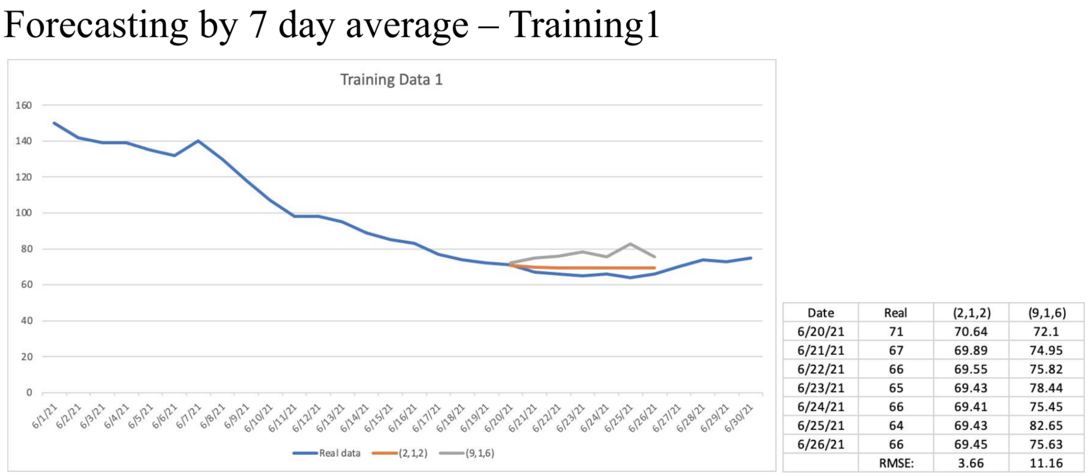
```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x o o o x x o o o  o  o  o
1 x o o o o o o x o o o  o  o  o
2 x x o o o o o x o o o  o  o  o
3 o x o o o o o x o o o  o  o  o
4 x x o o o o o o o o x  o  o  o
5 x x o o o o x x o x o  o  o  o
6 x x o x o o x x o o o  o  o  o
7 x x x x o x x o o x o  o  o  o
```

Figure 13.4: EACF for training set 4

13.3 Results of Forecasting

To evaluate the performance of our models, we used the error metric root mean squared error (RMSE). The RMSE of a model is found by squaring all residuals, or differences between the forecasted result and the actual data, and then taking the average of these squared residuals. Finally, the root of this is taken.

We compared the forecast results (Figure 13.5 to Figure 13.8) from our model(orange line) and auto.arima(gray line). Only training dataset 1 shows that our model provides better forecast result than the auto.arima() model with much lower RMSE. However, the forecast in other training sets did not provide a good result. Our model showed almost linear forecast as same as auto.arima().

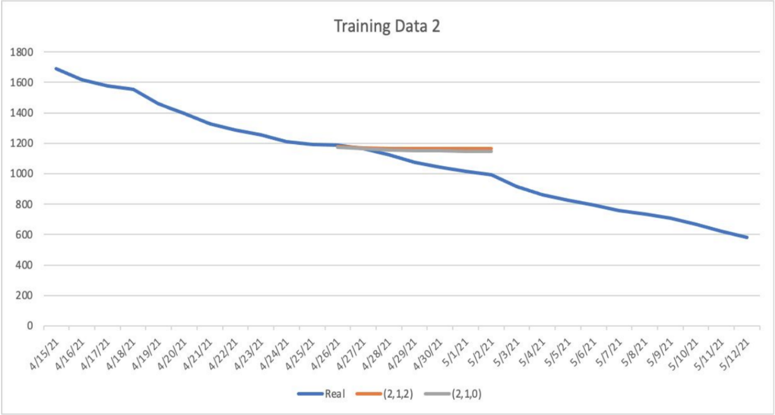


Forecasting by 7 day average – Training1

Date	Real	(2,1,2)	(9,1,6)
6/20/21	71	70.64	72.1
6/21/21	67	69.89	74.95
6/22/21	66	69.55	75.82
6/23/21	65	69.43	78.44
6/24/21	66	69.41	75.45
6/25/21	64	69.43	82.65
6/26/21	66	69.45	75.63
	RMSE:	3.66	11.16

Figure 13.5: Training dataset 1 forecast

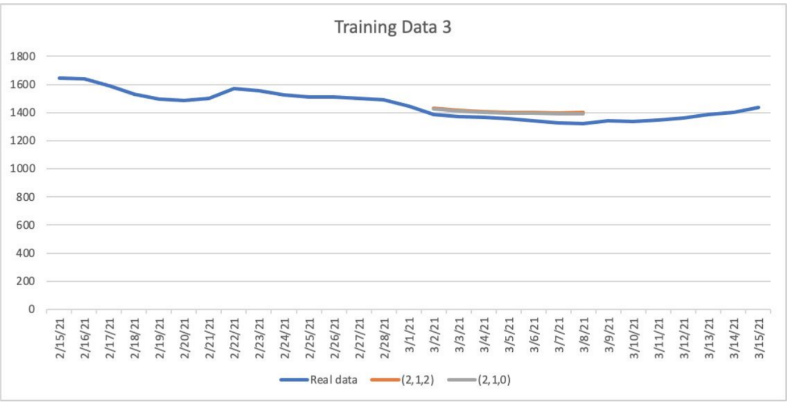
Forecasting by 7 day average – Training 2



Date	Real	(2,1,2)	(2,1,0)
4/26/21	1189	1177.01	1173.89
4/27/21	1164	1169.34	1164.13
4/28/21	1126	1165.67	1157.52
4/29/21	1074	1164.41	1153.29
4/30/21	1041	1164.33	1150.5
5/1/21	1016	1164.67	1148.69
5/2/21	992	1165.06	1147.5
	RMSE:	105.01	93.57

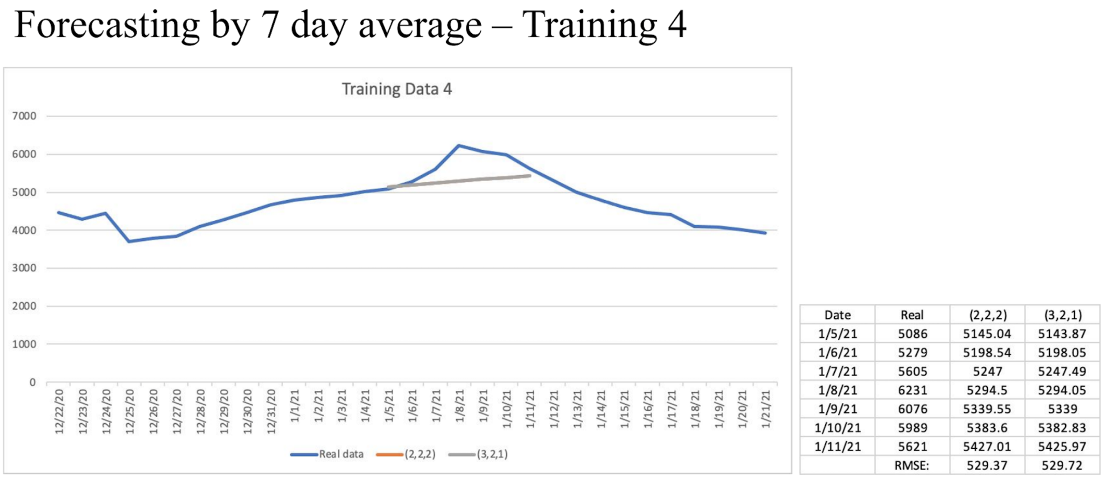
Figure 13.6: Training dataset 2 forecast

Forecasting by 7 day average – Training 3



Date	Real	(2,1,2)	(2,1,0)
3/2/21	1387	1428.8	1426.7
3/3/21	1372	1414.75	1411.86
3/4/21	1364	1405.29	1403.41
3/5/21	1355	1400.33	1397.6
3/6/21	1341	1398.53	1393.95
3/7/21	1325	1398.43	1391.56
3/8/21	1321	1399	1390
	RMSE:	56.21	51.44

Figure 13.7: Training dataset 3 forecast



Forecasting by 7 day average – Training 4

Date	Real	(2,2,2)	(3,2,1)
1/5/21	5086	5145.04	5143.87
1/6/21	5279	5198.54	5198.05
1/7/21	5605	5247	5247.49
1/8/21	6231	5294.5	5294.05
1/9/21	6076	5339.55	5339
1/10/21	5989	5383.6	5382.83
1/11/21	5621	5427.01	5425.97
	RMSE:	529.37	529.72

Figure 13.8: Training dataset 4 forecast.

14 ARIMA Modeling Conclusion and Discussions

Based on our forecast results and overall experience with ARIMA modeling, we are not convinced that this is a robust or helpful way to forecast COVID cases; especially for data further into the future than just a couple steps. These ARIMA models did not perform well outside of the training dataset, and during forecasts, the model tends to struggle to predict any peaks or sudden changes to the series. Based on the forecasts we generated, our root-mean square error was very large with respect to the scale of the data. Furthermore, ARIMA models best suited for very short-term forecasts. For our purposes of forecasting COVID cases, making only one to two week-out predictions is not so useful to our study.

We also tried average weekly data by calculating the mean of every weeks, which reduced our data set to 78 data points. Since data set is too small, it can not pass adf test after applying first difference. We have to transformed data by using ln(), which help P-value of adf test to be reduced to close to 0.05 but still larger than 0.05.

15 Introduction to Clustering

For C-Term, the data science team clustered all 50 states' COVID-19 rates with several features like vaccination rates. After creating these clusters, we want to conduct exploratory and statistical analysis on why certain clusters have higher or lower COVID rates than others. We wanted to answer the following questions:

1. Do certain policies like mask mandates, lockdowns, and others lead to significantly different COVID rates?
2. What shared state variables contribute to COVID rates? Some examples of variables we expect to lead to differing COVID rates are vaccination rates, population density, interstate travel, and more.

Since we wanted to try to keep all states' COVID rates as low as possible, the results will help us find out why or how certain states keep their COVID rates lower than others. If we can find noticeable trends within the high performing and low performing clusters, then we can produce a best-practice guideline on preventing COVID infections or even produce a statistical model that can predict a state's COVID spread based on its features.

16 Data Collection and Preprocessing

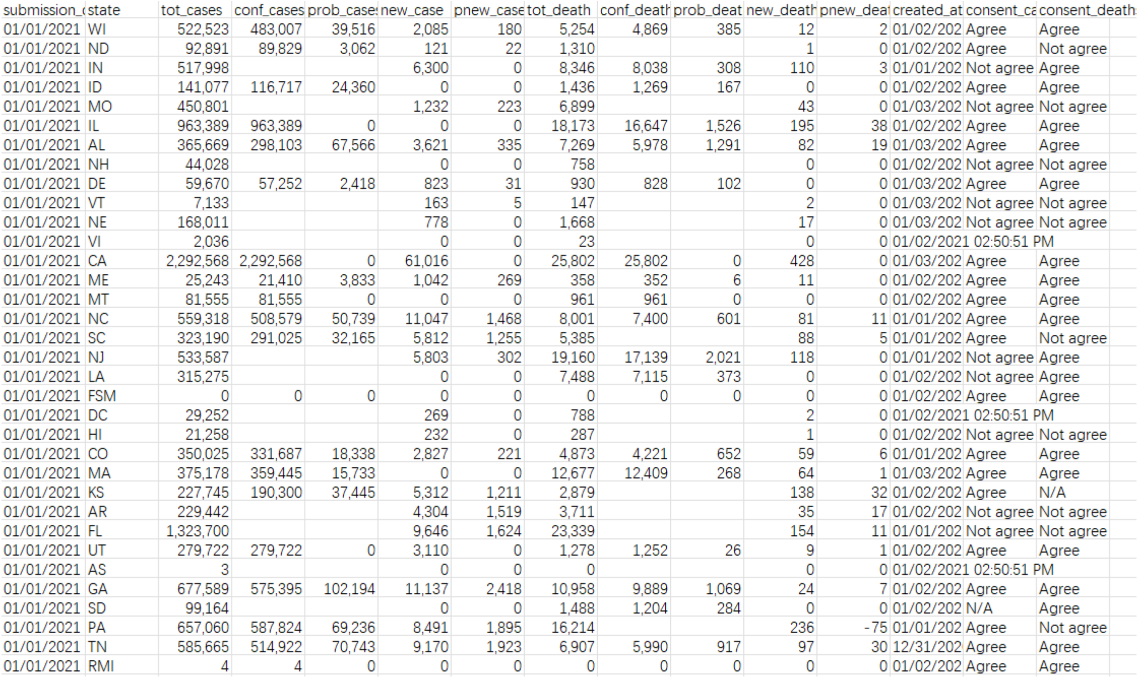
16.1 Data collection

To cluster and compare difference between each states of the United States, we collected COVID-19 data, like vaccination rate, positive cases, and population data for all states. To ensure all data are correct and reliable, we will only use government data as resources.

The data we collected for COVID-19 cases daily per state is publicly provided by the CDC. Figure 16.1 shows a part of the cases data. The shape of this data is 45841 rows and 15 columns. It contains all positive cases and death in four ways (daily cases, total cases, confidence cases and probably cases) from Jan 22 2020 to Jan 28 2022 for U.S. including 50 states and 10 non-state U.S. territories.

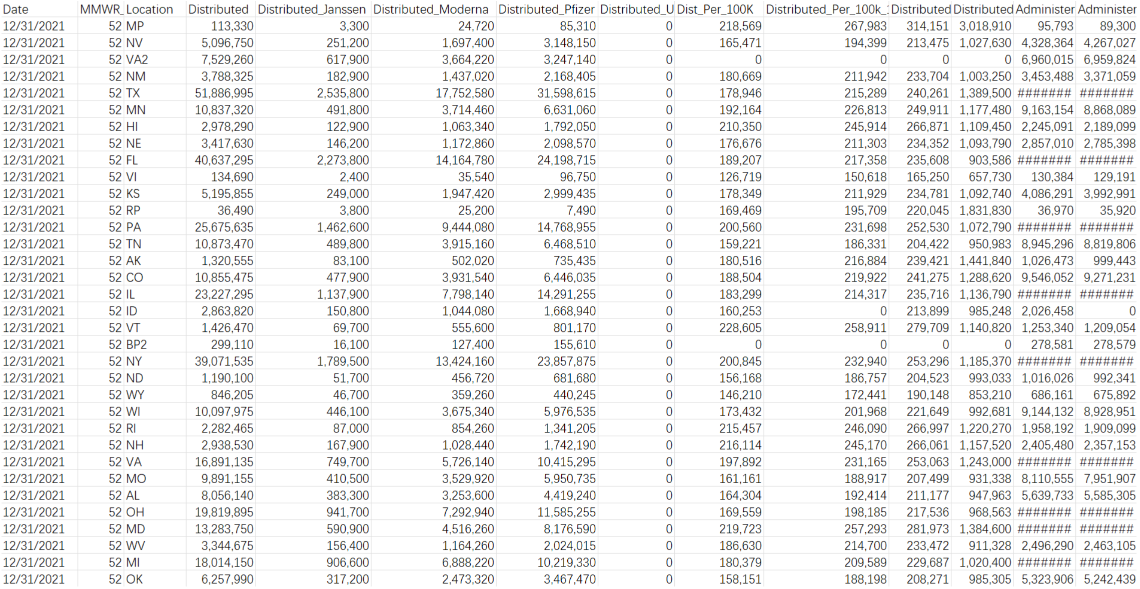
The data we collected for vaccinations daily per state is publicly provided by the CDC. As shown in Figure 16.2, the vaccination dataset is very large and detailed, containing 28312 rows and 83 columns from Dec 13 2020 to Jan 28 2020 for U.S.. In these 83 columns, there are total delivered doses for all different vaccinations, different age groups, people who only took the first dose and people who took additional doses.

The data we collected for population per state is from the 2020 US Census, and is publicly provided in Figure 16.3. The shape of this data is 54 rows and 76 columns. The relevant information included in this is just the population per state; other included information that was not relevant to us were demographics data.



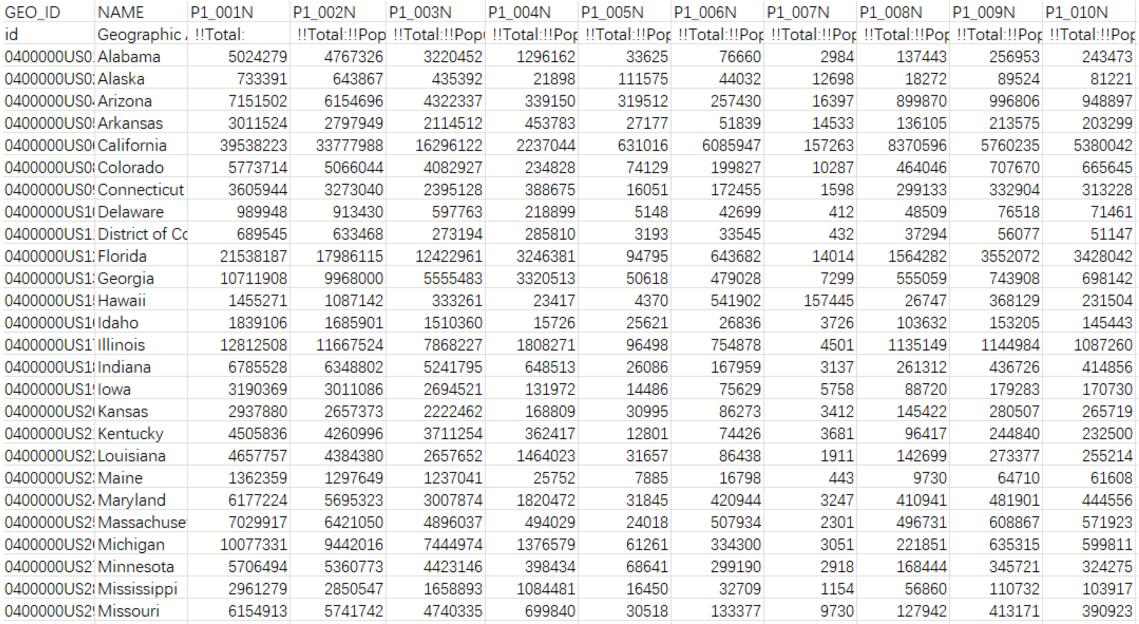
submission_	state	tot_cases	conf_cases	prob_case	new_case	pnew_case	tot_death	conf_death	prob_deat	new_death	pnew_deat	created_at	consent_ca	consent_death
01/01/2021	WI	522,523	483,007	39,516	2,085	180	5,254	4,869	385	12	2	01/02/202	Agree	Agree
01/01/2021	ND	92,891	89,829	3,062	121	22	1,310			1	0	01/02/202	Agree	Not agree
01/01/2021	IN	517,998			6,300	0	8,346	8,038	308	110	3	01/01/202	Not agree	Agree
01/01/2021	ID	141,077	116,717	24,360	0	0	1,436	1,269	167	0	0	01/02/202	Agree	Agree
01/01/2021	MO	450,801			1,232	223	6,899			43	0	01/03/202	Not agree	Not agree
01/01/2021	IL	963,389	963,389	0	0	0	18,173	16,647	1,526	195	38	01/02/202	Agree	Agree
01/01/2021	AL	365,669	298,103	67,566	3,621	335	7,269	5,978	1,291	82	19	01/03/202	Agree	Agree
01/01/2021	NH	44,028			0	0	758			0	0	01/02/202	Not agree	Not agree
01/01/2021	DE	59,670	57,252	2,418	823	31	930	828	102	0	0	01/03/202	Agree	Agree
01/01/2021	VT	7,133			163	5	147			2	0	01/03/202	Not agree	Not agree
01/01/2021	NE	168,011			778	0	1,668			17	0	01/03/202	Not agree	Not agree
01/01/2021	VI	2,036			0	0	23			0	0	01/02/2021 02:50:51 PM		
01/01/2021	CA	2,292,568	2,292,568	0	61,016	0	25,802	25,802	0	428	0	01/03/202	Agree	Agree
01/01/2021	ME	25,243	21,410	3,833	1,042	269	358	352	6	11	0	01/02/202	Agree	Agree
01/01/2021	MT	81,555	81,555	0	0	0	961	961	0	0	0	01/02/202	Agree	Agree
01/01/2021	NC	559,318	508,579	50,739	11,047	1,468	8,001	7,400	601	81	11	01/01/202	Agree	Agree
01/01/2021	SC	323,190	291,025	32,165	5,812	1,255	5,385			88	5	01/01/202	Agree	Not agree
01/01/2021	NJ	533,587			5,803	302	19,160	17,139	2,021	118	0	01/01/202	Not agree	Agree
01/01/2021	LA	315,275			0	0	7,488	7,115	373	0	0	01/02/202	Not agree	Agree
01/01/2021	FSM	0	0	0	0	0	0	0	0	0	0	01/02/202	Agree	Agree
01/01/2021	DC	29,252			269	0	788			2	0	01/02/2021 02:50:51 PM		
01/01/2021	HI	21,258			232	0	287			1	0	01/02/202	Not agree	Not agree
01/01/2021	CO	350,025	331,687	18,338	2,827	221	4,873	4,221	652	59	6	01/01/202	Agree	Agree
01/01/2021	MA	375,178	359,445	15,733	0	0	12,677	12,409	268	64	1	01/03/202	Agree	Agree
01/01/2021	KS	227,745	190,300	37,445	5,312	1,211	2,879			138	32	01/02/202	Agree	N/A
01/01/2021	AR	229,442			4,304	1,519	3,711			35	17	01/02/202	Not agree	Not agree
01/01/2021	FL	1,323,700			9,646	1,624	23,339			154	11	01/01/202	Not agree	Not agree
01/01/2021	UT	279,722	279,722	0	3,110	0	1,278	1,252	26	9	1	01/02/202	Agree	Agree
01/01/2021	AS	3			0	0	0			0	0	01/02/2021 02:50:51 PM		
01/01/2021	GA	677,589	575,395	102,194	11,137	2,418	10,958	9,889	1,069	24	7	01/02/202	Agree	Agree
01/01/2021	SD	99,164			0	0	1,488	1,204	284	0	0	01/02/202	N/A	Agree
01/01/2021	PA	657,060	587,824	69,236	8,491	1,895	16,214			236	-75	01/01/202	Agree	Not agree
01/01/2021	TN	585,665	514,922	70,743	9,170	1,923	6,907	5,990	917	97	30	12/31/202	Agree	Agree
01/01/2021	RMI	4	4	0	0	0	0	0	0	0	0	01/02/202	Agree	Agree

Figure 16.1: United States COVID-19 Cases and Deaths by State over Time.



Date	MMWR_	Location	Distributed	Distributed_Janssen	Distributed_Moderna	Distributed_Pfizer	Distributed_U	Dist_Per_100K	Distributed_Per_100k_	Distributed	Distributed	Administer	Administer
12/31/2021	52	MP	113,330	3,300	24,720	85,310	0	218,569	267,983	314,151	3,018,910	95,793	89,300
12/31/2021	52	NV	5,096,750	251,200	1,697,400	3,148,150	0	165,471	194,399	213,475	1,027,630	4,328,364	4,267,027
12/31/2021	52	VA2	7,529,260	617,900	3,664,220	3,247,140	0	0	0	0	0	6,960,015	6,959,824
12/31/2021	52	NM	3,788,325	182,900	1,437,020	2,168,405	0	180,669	211,942	233,704	1,003,250	3,453,488	3,371,059
12/31/2021	52	TX	51,886,995	2,535,800	17,752,580	31,598,615	0	178,946	215,289	240,261	1,389,500	#######	#######
12/31/2021	52	MN	10,837,320	491,800	3,714,460	6,631,060	0	192,164	226,813	249,911	1,177,480	9,163,154	8,868,089
12/31/2021	52	HI	2,978,290	122,900	1,063,340	1,792,050	0	210,350	245,914	266,871	1,109,450	2,245,091	2,189,099
12/31/2021	52	NE	3,417,630	146,200	1,172,860	2,098,570	0	176,676	211,303	234,352	1,093,790	2,857,010	2,785,398
12/31/2021	52	FL	40,637,295	2,273,800	14,164,780	24,198,715	0	189,207	217,358	235,608	903,586	#######	#######
12/31/2021	52	VI	134,690	2,400	35,540	96,750	0	126,719	150,618	165,250	657,730	130,384	129,191
12/31/2021	52	KS	5,195,855	249,000	1,947,420	2,999,435	0	178,349	211,929	234,781	1,092,740	4,086,291	3,992,991
12/31/2021	52	RP	36,490	3,800	25,200	7,490	0	169,469	195,709	220,045	1,831,830	36,970	35,920
12/31/2021	52	PA	25,675,635	1,462,600	9,444,080	14,768,955	0	200,560	231,698	252,530	1,072,790	#######	#######
12/31/2021	52	TN	10,873,470	489,800	3,915,160	6,468,510	0	159,221	186,331	204,422	950,983	8,945,296	8,819,806
12/31/2021	52	AK	1,320,555	83,100	502,020	735,435	0	180,516	216,884	239,421	1,441,840	1,026,473	999,443
12/31/2021	52	CO	10,855,475	477,900	3,931,540	6,446,035	0	188,504	219,922	241,275	1,288,620	9,546,052	9,271,231
12/31/2021	52	IL	23,227,295	1,137,900	7,798,140	14,291,255	0	183,299	214,317	235,716	1,136,790	#######	#######
12/31/2021	52	ID	2,863,820	150,800	1,044,080	1,668,940	0	160,253	0	213,899	985,248	2,026,458	0
12/31/2021	52	VT	1,426,470	69,700	555,600	801,170	0	228,605	258,911	279,709	1,140,820	1,253,340	1,209,054
12/31/2021	52	BP2	299,110	16,100	127,400	155,610	0	0	0	0	0	278,581	278,579
12/31/2021	52	NY	39,071,535	1,789,500	13,424,160	23,857,875	0	200,845	232,940	253,296	1,185,370	#######	#######
12/31/2021	52	ND	1,190,100	51,700	456,720	681,680	0	156,168	186,757	204,523	993,033	1,016,026	992,341
12/31/2021	52	WY	846,205	46,700	359,260	440,245	0	146,210	172,441	190,148	853,210	686,161	675,892
12/31/2021	52	WI	10,097,975	446,100	3,675,340	5,976,535	0	173,432	201,968	221,649	992,681	9,144,132	8,928,951
12/31/2021	52	RI	2,282,465	87,000	854,260	1,341,205	0	215,457	246,090	266,997	1,220,270	1,958,192	1,909,099
12/31/2021	52	NH	2,938,530	167,900	1,028,440	1,742,190	0	216,114	245,170	266,061	1,157,520	2,405,480	2,357,153
12/31/2021	52	VA	16,891,135	749,700	5,726,140	10,415,295	0	197,892	231,165	253,063	1,243,000	#######	#######
12/31/2021	52	MO	9,891,155	410,500	3,529,920	5,950,735	0	161,161	188,917	207,499	931,338	8,110,555	7,951,907
12/31/2021	52	AL	8,056,140	383,300	3,253,600	4,419,240	0	164,304	192,414	211,177	947,963	5,639,733	5,585,305
12/31/2021	52	OH	19,819,895	941,700	7,292,940	11,585,255	0	169,559	198,185	217,536	968,563	#######	#######
12/31/2021	52	MD	13,283,750	590,900	4,516,260	8,176,590	0	219,723	257,293	281,973	1,384,600	#######	#######
12/31/2021	52	WV	3,344,675	156,400	1,164,260	2,024,015	0	186,630	214,700	233,472	911,328	2,496,290	2,463,105
12/31/2021	52	MI	18,014,150	906,600	6,888,220	10,219,330	0	180,379	209,589	229,687	1,020,400	#######	#######
12/31/2021	52	OK	6,257,990	317,200	2,473,320	3,467,470	0	158,151	188,198	208,271	985,305	5,323,906	5,242,439

Figure 16.2: COVID-19 Vaccinations in the United States,Jurisdiction



GEO_ID	NAME	P1_001N	P1_002N	P1_003N	P1_004N	P1_005N	P1_006N	P1_007N	P1_008N	P1_009N	P1_010N
id	Geographic	!!Total:	!!Total:!!Pop	!!Total:!!Pop	!!Total:!!Pop	!!Total:!!Pop	!!Total:!!Pop	!!Total:!!Pop	!!Total:!!Pop	!!Total:!!Pop	!!Total:!!Pop
0400000US0	Alabama	5024279	4767326	3220452	1296162	33625	76660	2984	137443	256953	243473
0400000US0	Alaska	733391	643867	435392	21898	111575	44032	12698	18272	89524	81221
0400000US0	Arizona	7151502	6154696	4322337	339150	319512	257430	16397	899870	996806	948897
0400000US0	Arkansas	3011524	2797949	2114512	453783	27177	51839	14533	136105	213575	203299
0400000US0	California	39538223	33777988	16296122	2237044	631016	6085947	157263	8370596	5760235	5380042
0400000US0	Colorado	5773714	5066044	4082927	234828	74129	199827	10287	464046	707670	665645
0400000US0	Connecticut	3605944	3273040	2395128	388675	16051	172455	1598	299133	332904	313228
0400000US1	Delaware	989948	913430	597763	218899	5148	42699	412	48509	76518	71461
0400000US1	District of Co	689545	633468	273194	285810	3193	33545	432	37294	56077	51147
0400000US1	Florida	21538187	17986115	12422961	3246381	94795	643682	14014	1564282	3552072	3428042
0400000US1	Georgia	10711908	9968000	5555483	3320513	50618	479028	7299	555059	743908	698142
0400000US1	Hawaii	1455271	1087142	333261	23417	4370	541902	157445	26747	368129	231504
0400000US1	Idaho	1839106	1685901	1510360	15726	25621	26836	3726	103632	153205	145443
0400000US1	Illinois	12812508	11667524	7868227	1808271	96498	754878	4501	1135149	1144984	1087260
0400000US1	Indiana	6785528	6348802	5241795	648513	26086	167959	3137	261312	436726	414856
0400000US1	Iowa	3190369	3011086	2694521	131972	14486	75629	5758	88720	179283	170730
0400000US2	Kansas	2937880	2657373	2222462	168809	30995	86273	3412	145422	280507	265719
0400000US2	Kentucky	4505836	4260996	3711254	362417	12801	74426	3681	96417	244840	232500
0400000US2	Louisiana	4657757	4384380	2657652	1464023	31657	86438	1911	142699	273377	255214
0400000US2	Maine	1362359	1297649	1237041	25752	7885	16798	443	9730	64710	61608
0400000US2	Maryland	6177224	5695323	3007874	1820472	31845	420944	3247	410941	481901	444556
0400000US2	Massachuse	7029917	6421050	4896037	494029	24018	507934	2301	496731	608867	571923
0400000US2	Michigan	10077331	9442016	7444974	1376579	61261	334300	3051	221851	635315	599811
0400000US2	Minnesota	5706494	5360773	4423146	398434	68641	299190	2918	168444	345721	324275
0400000US2	Mississippi	2961279	2850547	1658893	1084481	16450	32709	1154	56860	110732	103917
0400000US2	Missouri	6154913	5741742	4740335	699840	30518	133377	9730	127942	413171	390923

Figure 16.3: States population.

16.2 Cleaning and Preprocessing

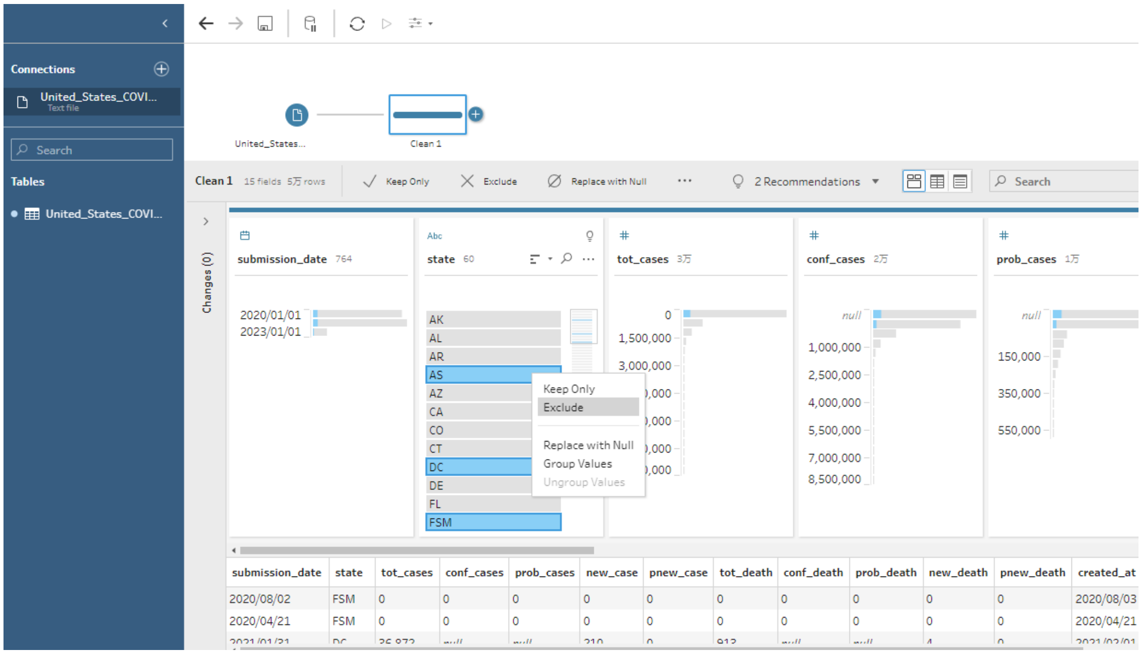
In our project, we focused strictly on the months in which the COVID-19 vaccination was widely and publicly available. As such, we decided to cut off data prior to March of 2021. This was done simply using Excel, sorting by latest to newest, and deleting entries prior to March 2021.

For all datasets, the data providers stored data for US territories outside the 50 states, such as Puerto Rico and DC. Since we are focusing on just state data, we needed to remove all entries unrelated to the states. Furthermore, the case and vaccination data was collected daily, which we believed would be too noisy to cluster effectively. As such, we needed to aggregate and total all the new cases for each state, per month.

Before removing non-territories and aggregating data, it is necessary to drop useless columns to reduce the size of the data. In the cases dataset, for quantity numbers we only need daily new cases, since we only focus on aggregated new cases but not death or total cases. Similar to the vaccination dataset, we only care the number of people fully vaccinated in each states, so we dropped all other columns. For population dataset, we only keep total states population.

These data sets contain too many rows and therefore it is time-consuming to delete all non-state territories in Excel. So we used Tableau Prep Builder to remove these areas by selecting names of all non-state territories and then choosing exclude as shown in Figure 16.4.

To aggregate data based on month, the Python library "Pandas" was extremely helpful; particularly the dataframe and the method "groupby". Method groupby allows the programmer to perform actions on the data based on grouping data by a criteria; in our case, we want



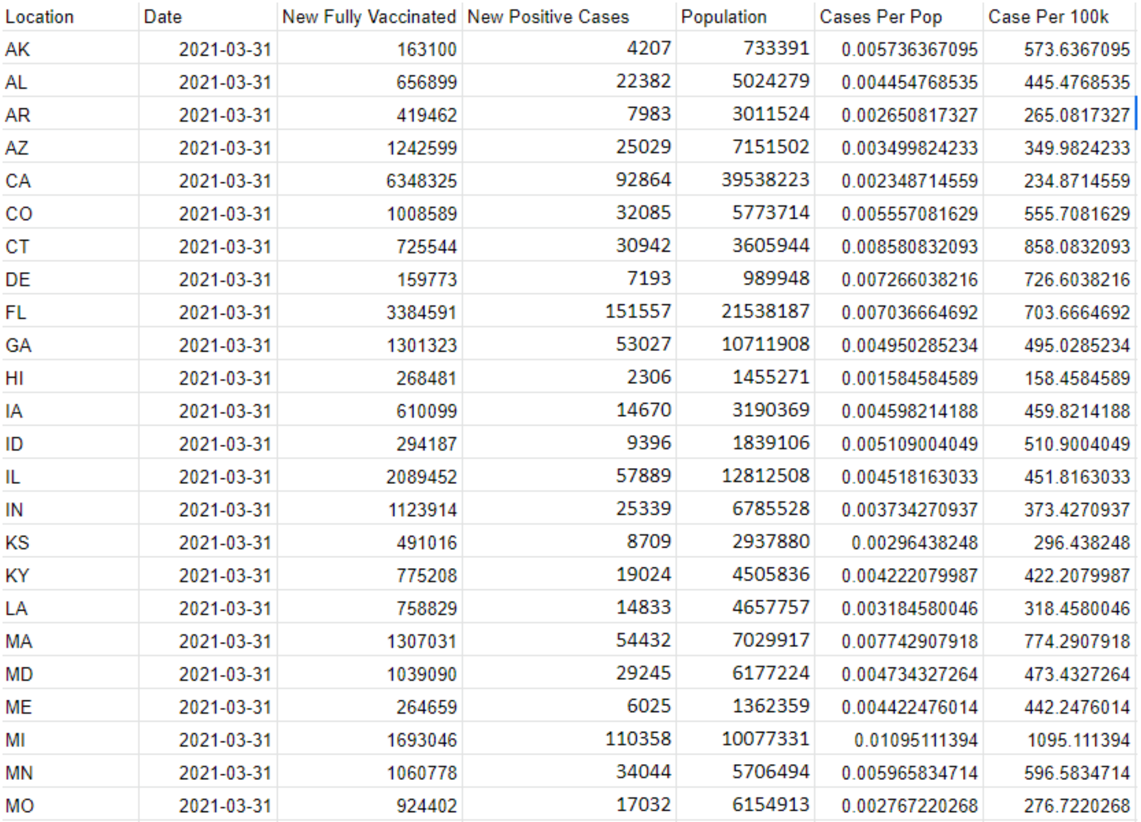
submission_date	state	tot_cases	conf_cases	prob_cases	new_case	pnew_case	tot_death	conf_death	prob_death	new_death	pnew_death	created_at
2020/08/02	FSM	0	0	0	0	0	0	0	0	0	0	2020/08/03
2020/04/21	FSM	0	0	0	0	0	0	0	0	0	0	2020/04/21
2021/01/31	DC	36,872	null	null	210	0	913	null	null	4	0	2021/02/01

Figure 16.4: A faster way to remove non-state territories is by using the standalone software Tableau, available at tableau.com

monthly sums of vaccination by day as well as monthly sums of COVID cases, grouped by state. Before being able to run this, we also had to change the datatypes of the columns such that the program would recognize the dates as a datetime datatype, and the vaccination data as integer.

The final step in our cleaning process was adding the census population data. Before doing so, we had to manually change all the state full names into the state abbreviations, and then sort by alphabetical order to make sure the rows match with the current dataset that we had. We then pasted the data into our final dataset, making sure to line up the states correctly.

As shown in Figure 16.5, we put columns we will use into a new dataset with two calculated columns—-one is cases per population and another is cases per 100k.



Location	Date	New Fully Vaccinated	New Positive Cases	Population	Cases Per Pop	Case Per 100k
AK	2021-03-31	163100	4207	733391	0.005736367095	573.6367095
AL	2021-03-31	656899	22382	5024279	0.004454768535	445.4768535
AR	2021-03-31	419462	7983	3011524	0.002650817327	265.0817327
AZ	2021-03-31	1242599	25029	7151502	0.003499824233	349.9824233
CA	2021-03-31	6348325	92864	39538223	0.002348714559	234.8714559
CO	2021-03-31	1008589	32085	5773714	0.005557081629	555.7081629
CT	2021-03-31	725544	30942	3605944	0.008580832093	858.0832093
DE	2021-03-31	159773	7193	989948	0.007266038216	726.6038216
FL	2021-03-31	3384591	151557	21538187	0.007036664692	703.6664692
GA	2021-03-31	1301323	53027	10711908	0.004950285234	495.0285234
HI	2021-03-31	268481	2306	1455271	0.001584584589	158.4584589
IA	2021-03-31	610099	14670	3190369	0.004598214188	459.8214188
ID	2021-03-31	294187	9396	1839106	0.005109004049	510.9004049
IL	2021-03-31	2089452	57889	12812508	0.004518163033	451.8163033
IN	2021-03-31	1123914	25339	6785528	0.003734270937	373.4270937
KS	2021-03-31	491016	8709	2937880	0.00296438248	296.438248
KY	2021-03-31	775208	19024	4505836	0.004222079987	422.2079987
LA	2021-03-31	758829	14833	4657757	0.003184580046	318.4580046
MA	2021-03-31	1307031	54432	7029917	0.007742907918	774.2907918
MD	2021-03-31	1039090	29245	6177224	0.004734327264	473.4327264
ME	2021-03-31	264659	6025	1362359	0.004422476014	442.2476014
MI	2021-03-31	1693046	110358	10077331	0.01095111394	1095.111394
MN	2021-03-31	1060778	34044	5706494	0.005965834714	596.5834714
MO	2021-03-31	924402	17032	6154913	0.002767220268	276.7220268

Figure 16.5: Final dataset

17 Clustering

17.1 Clustering and the K-Means Algorithm

Clustering in machine learning is the task of grouping together objects based on a similarity criterion [15]. One of the most basic examples of clustering is grouping together points on an x-y scatter plot, as seen in Figure 17.1.

One of the most simple and popular algorithms to cluster points based on similarity is the K-Means algorithm [14]. In this algorithm, K points are randomly selected to be the starting centroid of a cluster. Then, the following loop is performed until the centroid no longer changes, or the algorithm meets some termination condition:

1. Compute the sum of squared distances between the data points and the centroids.
2. Assign all data points to their nearest centroid.
3. Reassign a new centroid for each cluster by taking the average position of each point in the cluster.

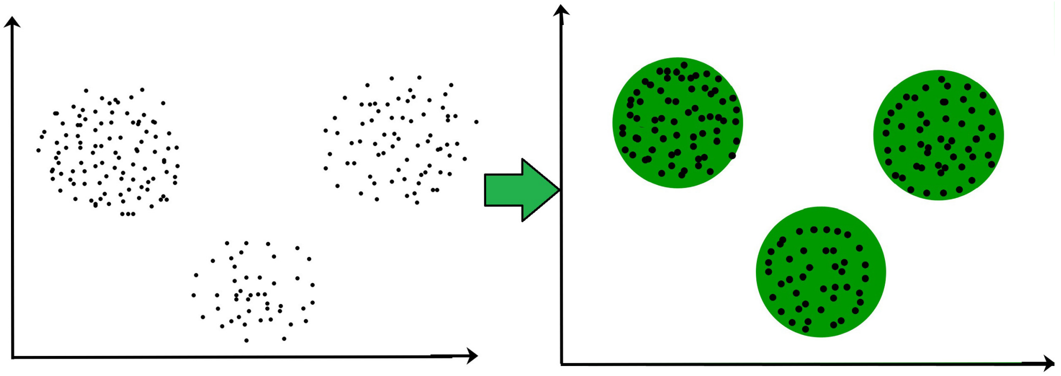


Figure 17.1: The goal of a clustering algorithm in this case would be to group together points based on their relative closeness to another point.

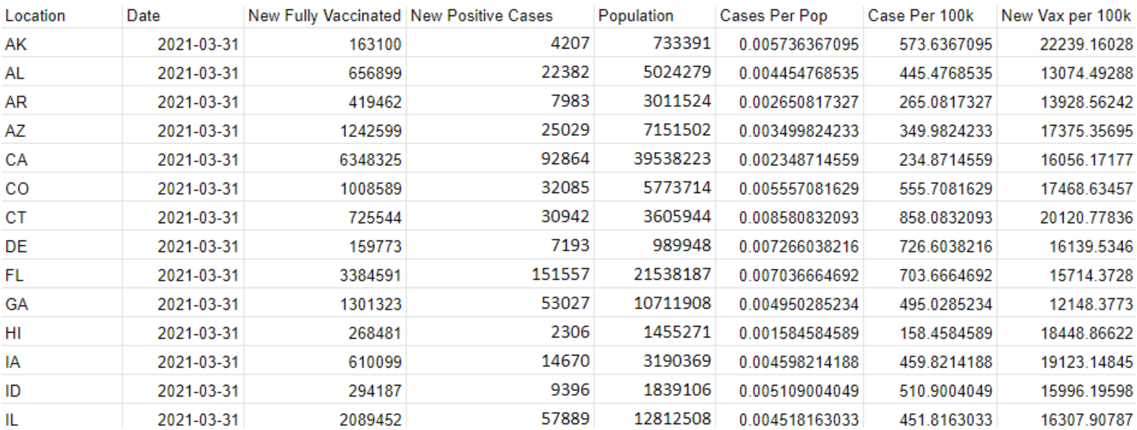
An issue that K-Means has is its reliance on good initial centroid assignment; the clusters can completely change based on the first centroids that are randomly chosen. Just because the objective function converges does not mean it has created globally optimal clusters.

Clustering is of value to us in this project as it allows us to group similar states in terms of COVID-19 prevention performance. By performing clustering, it becomes clear which states kept their COVID cases low relative to their population, and then we will figure out why these states are grouped.

17.2 Clustering Features

The final required step before clustering is choosing and preparing features to cluster. In our study, we used cases and vaccination as the two features for K-means. In order to get better results, we added another calculated column that calculating the vaccination per 100k as shown in Figure 17.2. Then we will use both vaccination per 100k and cases per 100k for K-means.

Before performing K-means clustering, we need to scale our data. In Figure 17.2, numbers in Case Per 100k are all under 1000 while numbers in New Vax per 100k are higher than 10k even higher than 20k. therefore, it is important to scale them first [16].



Location	Date	New Fully Vaccinated	New Positive Cases	Population	Cases Per Pop	Case Per 100k	New Vax per 100k
AK	2021-03-31	163100	4207	733391	0.005736367095	573.6367095	22239.16028
AL	2021-03-31	656899	22382	5024279	0.004454768535	445.4768535	13074.49288
AR	2021-03-31	419462	7983	3011524	0.002650817327	265.0817327	13928.56242
AZ	2021-03-31	1242599	25029	7151502	0.003499824233	349.9824233	17375.35695
CA	2021-03-31	6348325	92864	39538223	0.002348714559	234.8714559	16056.17177
CO	2021-03-31	1008589	32085	5773714	0.005557081629	555.7081629	17468.63457
CT	2021-03-31	725544	30942	3605944	0.008580832093	858.0832093	20120.77836
DE	2021-03-31	159773	7193	989948	0.007266038216	726.6038216	16139.5346
FL	2021-03-31	3384591	151557	21538187	0.007036664692	703.6664692	15714.3728
GA	2021-03-31	1301323	53027	10711908	0.004950285234	495.0285234	12148.3773
HI	2021-03-31	268481	2306	1455271	0.001584584589	158.4584589	18448.86622
IA	2021-03-31	610099	14670	3190369	0.004598214188	459.8214188	19123.14845
ID	2021-03-31	294187	9396	1839106	0.005109004049	510.9004049	15996.19598
IL	2021-03-31	2089452	57889	12812508	0.004518163033	451.8163033	16307.90787

Figure 17.2: Adding calculated vaccination data (New Vax per 100k) into the dataset

In python, there are four scalers in sklearn library that people usually used. For our dataset, we tried two of them. The first one is *MinMaxScaler*, it shifts all features individually into a range from 0 to 1. For our two dimensional data, cases per 100k and vaccination per 100k will be contained into a 0 to 1 range. The second one is *StandardScaler*, which scales all features to the same magnitude by ensuring that for each feature the mean is zero and the variance is 1. However, this scaler does not ensure a range for features; in other word, there is no minimum or maximum values. The results of using these two scaler are almost same. For StandardScaler, it contains negative values in both x-axis and y-axis. Since K-means only considers the distance between points, so both MinMaxScaler and StandardScaler are good for us. We will just choose MinMaxScaler for this project since it is scaled between 0 and 1 which is easier to read compared with using StandardScaler.

18 Results of Clustering

As shown in Figure 18.1, we performed a K-means with 4 clusters on the total dataset. In Figures 18.2 through 18.4, we showed how K-means worked on monthly data. However, monthly results were still hard to interpret and draw useful conclusion. Therefore, we instead tried splitting the total data into different period of variants. In Figures 18.5 through 18.7 we clustered states based on the time periods PreDelta, Delta, and Omicron. For each of the plots, the numbers in the legend simply represent the name of the cluster.

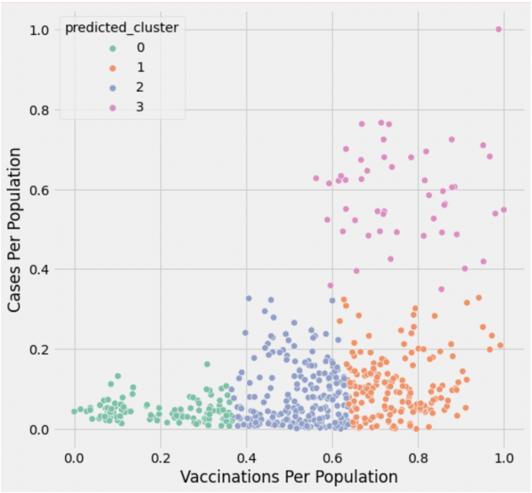


Figure 18.1: K-means with K=4 for all data

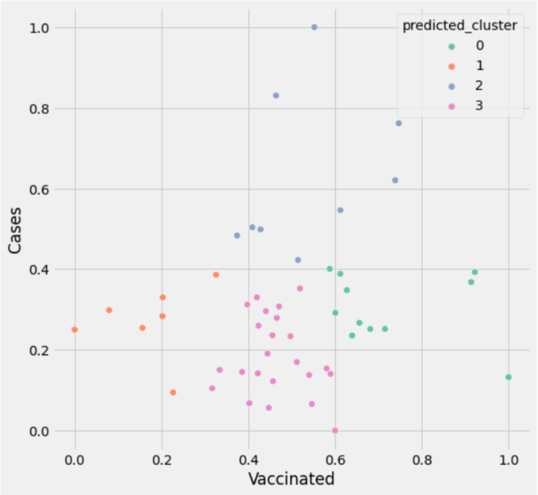


Figure 18.2: K-means with K=4 for March

We cut data into three parts: PreDelta period (2021.3 - 2021.6), Delta period (2021.7 - 2021.10)and Omicron period (2021.11 - 2022.1). This provided much more reasonable results.

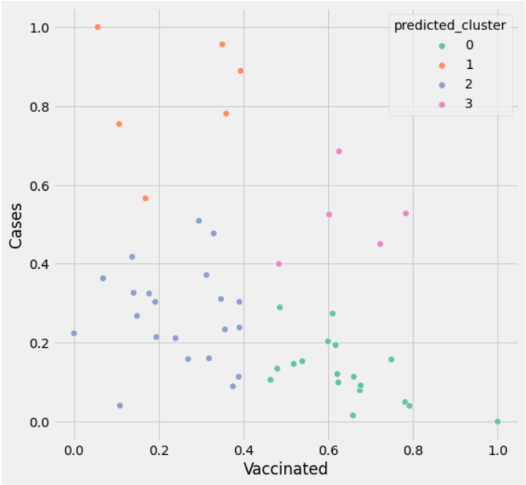


Figure 18.3: K-means with K = 4 for June

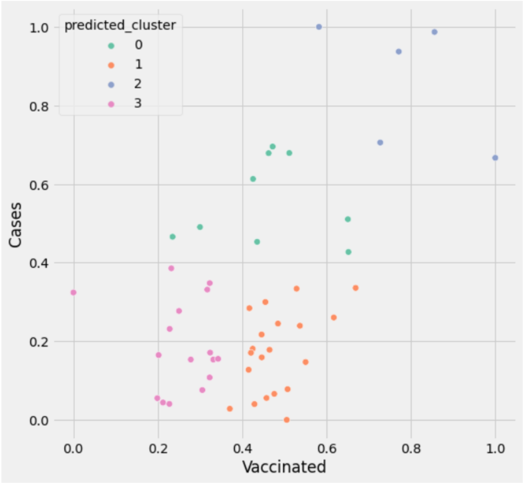


Figure 18.4: K-means with K=4 for Aug

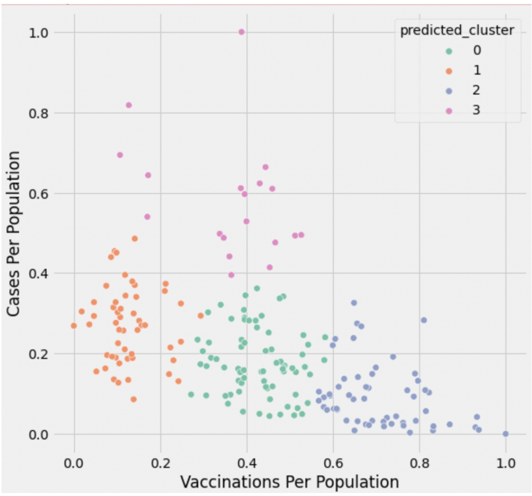


Figure 18.5: K-means with K=4 for PreDelta period.

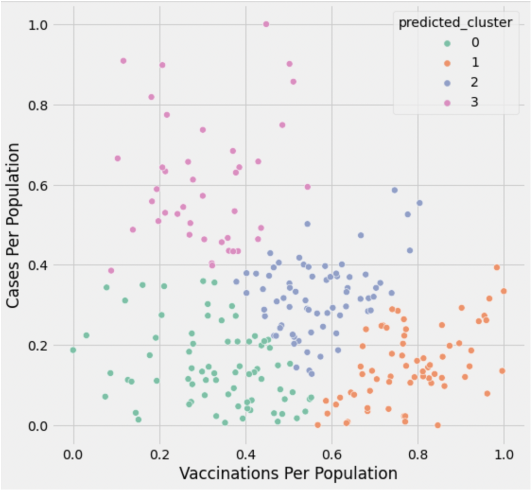


Figure 18.6: K-means with K=4 for Delta period.

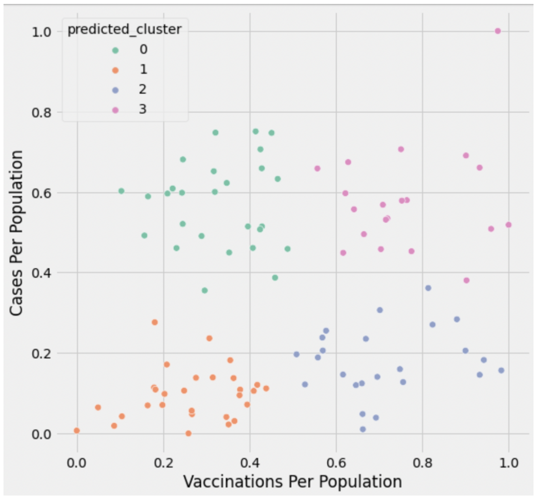


Figure 18.7: K-means with K=4 for Omicron period.

As seen in Figure 18.5, the period of PreDelta, the states that have relatively higher vaccination have much lower positive cases. In Figure 18.6, during the period of Delta, there seems to be a much lower negative correlation between vaccinations and cases per population. Many low vaccination states still had a low case rate while many had a high case rate. However, highly vaccinated states had lower-than-average case rates. For Omicron period shown in Figure 18.7, the situation changed a lot. It seems like there is no relationship between vaccination and positives cases. This suggests that vaccination may not be as useful for Omicron. Also, many people did not take booster, which reduces the protection of vaccination.

There are some interesting examples of states shifting positions in each plot based on which time period. For example, the most heavily vaccinated state during the PreDelta period was Vermont (June 2021), with a correspondingly low COVID rate for the time period. However, during the Delta variant, Vermont had the 3rd highest vaccination rate (September 2021) yet had only slightly below-average case rate. Finally, during the Omicron period, Vermont once again had the highest vaccination rate, yet had an above average case per population rate. This seems to suggest that as variants changed over time, even highly vaccinated populations became increasingly susceptible to infection.

Using Rhode Island as another example, in PreDelta period, the vaccination increased from 0.3 to 0.9, and the cases decreased from 0.65 to 0.03, which means it moved from purple area to the blue area in Figure 18.5. And during the Delta period, RI started at orange area, the right bottom in Figure 18.6, Showing that high vaccination did help. However, in the era of Omicron, even though the vaccination still increased, the cases increased hugely from 0.11 to 1, leading the RI to be the top right point in Figure 18.7. This a good example of effect of vaccination fading.

19 Clustering Conclusions

Clustering turned out to be a much more difficult pursuit than we had anticipated, with its interpretations also being more difficult than other types of analysis. Nonetheless, we were able to generate interesting results that seemed to indicate that as variants changed, the efficacy of the vaccine dramatically changed as well. For example, while there was a clear relationship between high vaccinations leading to lower case rates for the PreDelta period, as time went on even highly vaccinated states generated high COVID positivity rate.

Some future directions to take such a study would be to investigate commonalities between the states within the same cluster. These commonalities to investigate could be COVID prevention policy such as mask mandating and lockdowns.

References

[1] Stephanie. "Stationarity amp; Differencing: Definition, Examples, Types." Statistics How To, 16 Apr. 2021, www.statisticshowto.com/stationarity/

[2] Prabhakaran, Selva. "Augmented Dickey-Fuller (ADF) Test - Must Read Guide - Ml+." Machine Learning Plus, 19 Dec. 2021, www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/

[3] CDC Museum COVID-19 Timeline.(n.d.). Retrieved from https://www.cdc.gov/museum/timeline/covid19.html

[4] SARS-CoV-2 Variant Classifications and Definitions. (2021, December 1). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html

[5] Tracking SARS-CoV-2 variants. (n.d.) Retrieved from https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

[6] Kathy Katella (2021, December 20). Omicron, Delta, Alpha, and More: What To Know About the Coronavirus Variants. Retrieved from https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron

[7] COVID Data Tracker (n.d.). Retrieved from https://covid.cdc.gov/covid-data-tracker/datatracker-home

[8] CDC Image Newsroom Image Library (n.d.). Retrieved from https://www.cdc.gov/media/subtopic/images.htm

[9] Symptoms (2021, February 22). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[10] Centers for Disease Control and Prevention. (n.d.). United States covid-19 cases and deaths by State over time. Centers for Disease Control and Prevention. Retrieved February 25, 2022, from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

[11] Centers for Disease Control and Prevention. (n.d.). Covid-19 vaccinations in the United States,jurisdiction. COVID-19 Vaccinations in the United States,Jurisdiction. Retrieved February 25, 2022, from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc

[12] United States Census Bureau. Decennial Census P1 Race. Retrieved February 25, 2022, https://data.census.gov/cedsci/table?q=state/20by/20state/20population/ tid=DECENNIALPL2020.P1

[13] Olsen, NL; Markussen, B; Raket, LL (2018), "Simultaneous inference for misaligned multivariate functional data", Journal of the Royal Statistical Society, Series C, 67 (5): 1147–76

[14] Dabbura, I. (2020, August 10). K-means clustering: Algorithm, applications, evaluation methods, and drawbacks. K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Retrieved February 25, 2022, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[15] Galbraith, Sally et al. “A study of clustered data and approaches to its analysis.” The Journal of neuroscience : the official journal of the Society for Neuroscience vol. 30,32 (2010): 10601-8. doi:10.1523/JNEUROSCI.0362-10.2010

[16] Zheng, A., Casari, A. (2018). In Feature Engineering for Machine Learning: Principles and techniques for Data scientists. essay, O’Reilly.

[17] “Lasting Immunity Found after Recovery from COVID-19.” National Institutes of Health, U.S. Department of Health and Human Services, 11 Feb. 2021, https://www.nih.gov/news-events/nih-research-matters/lasting-immunity-found-after-recovery-covid-19.

[18] “National Center for Health Statistics - Massachusetts.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 17 Mar. 2021, https://www.cdc.gov/nchs/pressroom/states/massachusetts/ma.htm.

[19] “Ending Isolation and Precautions for People with Covid-19: Interim Guidance.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html.

[20] “Waning 2-Dose and 3-Dose Effectiveness of Mrna Vaccines against COVID-19–Associated Emergency Department and Urgent Care Encounters and Hospitalizations among Adults

during Periods of Delta and Omicron Variant Predominance." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 17 Feb. 2022, https://www.cdc.gov/mmwr/volumes/71/wr/mm7107e2.htm.

[21] Health, Department of Public, and Executive Office of Health and Human Services. "Covid-19 Response Reporting." Mass.gov, https://www.mass.gov/info-details/covid-19-response-reporting.

[22] Health, Department of Public. "Massachusetts Covid-19 Vaccination Data and Updates." Mass.gov, https://www.mass.gov/info-details/massachusetts-covid-19-vaccination-data-and-updates.

[23] Centers for Disease Control. 2022. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data

[24] Centers for Disease Control. 2022. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc/data

[25] State of Massachusetts. 2022. https://www.mass.gov/info-details/reopening-massachusetts

[26] State of Massachusetts. 2022. https://www.mass.gov/info-details/covid-19-state-of-emergency

[27] Kopfová, J., Nábělková, P., Rachinskii, D. et al. Dynamics of SIR model with vaccination and heterogeneous behavioral response of individuals modeled by the Preisach operator. J. Math. Biol. 83, 11 (2021). https://doi.org/10.1007/s00285-021-01629-8