Media Engineering and Technology Faculty
German University in Cairo

Identifying High-risk Areas of COVID-19

Bachelor Thesis

Author: Sandy Sameh Fakhry Fouad
Supervisors: Assoc. Prof. Seif Eldawlatly

Submission Date: 01 August, 2021

Declaration

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowlegement has been made in the text to all other material used

Sandy Sameh Fakhry Fouad
01 August, 2021

Acknowledgments

First and foremost, I would like to thank God for His showers of blessings throughout the years and through my thesis' work. I would like to express my deep and sincere gratitude to my supervisor Assoc. Prof.Seif Eldawlatly for providing me with guidance throughout the semester. It was a great privilege to work and study under his supervision. I am expanding my heartfelt thanks to my precious mother, Iman for her faith in me and patience during the semester. I am extremely grateful for my Dad's care and support. I would like to give very special thanks to tante Amoula for filling my hear with hope. My thanks and appreciations also go to my Frisbee teammates that have lifted my spirits and motivated me throughout this project specially Naim and Waka. I would like to also thank Sharko, Yomna, Sandra Boulos, Martha, Monica Emad, and Ramy for being my rock. Last but not least, I would like to thank my sister, Sarah who helps me stay in the light.

Abstract

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has become an unprecedented public health crisis. This study aims to identify clusters of geographical locations that involve high-risk areas in the counties of New York State and in the states of the USA by using k-means clustering method. Features were selected manually. Those features were tested through 3 different attempts; as provided in the dataset , with reference to the population size of each region , and with reference to population density. They then passed by dimension reduction technique (PCA) to extract the most important information from the data table.

In addition, this study provides monitoring the disease's progression in the USA by predicting the number of total cases during different time intervals; one attempt to predict the beginning of wave 2 by using wave 1, and the other attempt is to predict the total cases in the USA later into wave 2 by using the beginning of wave 2. The predictions results were more close to the real values as long as the number of days being predicted decreased ($\sim$ 10 days). In the first time interval, SARIMA model performed best in predicting total number of confirmed cases during last 10 days in October using wave 1 and the beginning of wave 2 as a training data. In the second time interval, Holt's Linear Model had the best prediction values of the total number of confirmed cases during last 10 days in February using just wave 2 as a training data. While SVR performed poorly in all the prediction scenarios.

The GoogleCloudPlatform data set used in this study includes open, publicly sourced, licensed data relating to demographics, economy, epidemiology, geography, health, hospitalizations, mobility, government response, weather,and more. Moreover, the data merges daily time-series, +20,000 global sources, at a fine spatial resolution, using a consistent set of region keys.

Contents

Acknowledgments **V**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Aim . . . 1
1.3 Contribution . . . 1
1.4 Thesis Outline . . . 2

2 Background **3**
2.1 COVID-19 . . . 3
2.1.1 Symptoms . . . 3
2.1.2 Spread of COVID-19 . . . 3
2.2 Machine Learning . . . 4
2.2.1 Supervised Learning . . . 4
2.2.2 Unsupervised Learning . . . 5
2.2.3 Reinforcement Learning . . . 5
2.3 Dimensionality Reduction . . . 5
2.3.1 Principal Component Analysis . . . 5
2.4 Regression Models . . . 6
2.4.1 Linear Regression . . . 6
2.4.2 Support Vector Regression . . . 7
2.4.3 Exponential Smoothing . . . 7
2.4.4 ARIMA/SARIMA MODEL . . . 9
2.5 Clustering Model . . . 11
2.5.1 K-means . . . 11
2.6 Cluster Validity indices . . . 12
2.6.1 Elbow Method . . . 12
2.6.2 Silhouette Analysis . . . 12
2.6.3 Davies–Bouldin index . . . 13
2.7 Prediction Evaluation Parameters . . . 13
2.7.1 Mean Absolute Error (MAE) . . . 13
2.7.2 Mean Square Error (MSE) . . . 14
2.7.3 Root Mean Square Error (RMSE) . . . 14
2.7.4 Mean Absolute Percentage Error(MAPE) . . . 14

2.8 Literature Review . . . 15
2.8.1 Spatial Analysis Studies . . . 15
2.8.2 Predictive Models Studies . . . 17

3 Methodology **21**
3.1 Methodological Approach Overview . . . 21
3.2 Method of data collection . . . 22
3.3 Data Preparation . . . 24
3.3.1 Feature Extraction and Engineering . . . 24
3.3.2 Feature and Scaling Selection . . . 27
3.3.3 Rescaling(Min-Max Scaling) . . . 27
3.4 Clustering approach . . . 28
3.4.1 Dimensionality Reduction using PCA . . . 28
3.4.2 K-Means . . . 29
3.5 Prediction approach . . . 29
3.5.1 Data Splitting . . . 29
3.5.2 Models Applied . . . 30
3.6 Evaluation . . . 30
3.6.1 Clustering approach: K-means Evaluation . . . 30
3.6.2 Prediction approach evaluation . . . 30

4 Results **31**
4.1 Clustering NY State . . . 32
4.1.1 First Attempt: Original . . . 32
4.1.2 Second Attempt: Normalization on population . . . 39
4.1.3 Third Attempt: Normalization on Density . . . 52
4.2 Clustering USA . . . 52
4.2.1 First Attempt: Original . . . 52
4.2.2 Second Attempt: Normalization on population . . . 57
4.2.3 Third Attempt: Normalization on Density . . . 61
4.3 Predictions Using Wave 1 . . . 62
4.3.1 First Trial . . . 62
4.3.2 Second Trial . . . 66
4.3.3 Third Trial . . . 69
4.3.4 Forth Trial . . . 72
4.3.5 Discussion . . . 75
4.4 Predictions Using the start of Wave 2 . . . 75
4.4.1 First Trial . . . 75
4.4.2 Second Trial . . . 79
4.4.3 Third Trial . . . 82
4.4.4 Forth Trial . . . 85
4.4.5 Discussion . . . 88

5 Conclusion **89**
5.1 Future Works 90

Appendix **91**

A Lists **92**
List of Abbreviations 92
List of Figures 96

References **100**

Chapter 1

Introduction

1.1 Motivation

COVID-19 has drastically changed everyday lives around the globe within a short period of time leaving many health foundations paralyzed. With the rise of Artificial Intelligence and machine learning technologies, these tools can help curb the spread of the virus through their countless applications as:- identifying areas with possible high risk; with that information, governments can enforce lock-down and offer health solutions. As well as using tech methods such as Natural Language Processing to roll out chatbots offering an accurate diagnosis of COVID-19 through feeding it with the user's symptoms to decreases the overburning of healthcare givers. In addition , learning models of machine learning are used to predict positivity and mortality rates in order to prepare governments and health organizations to take the required precautions to help flattening the curve.

1.2 Aim

The aim of this study is to identify clusters of testing rates, positivity rates, and mortality rate to understand the risk level of COVID-19 during wave 1 using GoogleCloudPlatform dataset. Also the objective of this study is to predict the total confirmed cases of COVID-19 during wave 2.

1.3 Contribution

To contribute to the current human catastrophe the attempts in this study is to provide both spatial analysis by identifying regions with possible high risk of COVID-19 to understand, prepare, and to move forward with the right decisions for the upcoming situation. In addition , Temporal analysis by providing a prediction of total confirmed cases during wave 2 to help in taking precautions and correct action to contain this crises.

1.4 Thesis Outline

Chapter 2 gives an overview of some of the concepts, and that were used in this study. Next, the methodology chapter which illustrates the workflow of the execution and the implementation. Next, The results chapter mentions the attempts analyzed in this study, the results , its interpretations, and the limitations of the study. The final chapter offers the conclusions and possible future works.

Chapter 2

Background

2.1 COVID-19

The coronavirus disease (COVID-19) is a viral infection highly transmittable caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which originally appeared in Wuhan, China, and it has sequentially propagated around the world [1]. Recently at the end of 2019, the city of Wuhan, China, the epicenter of the current COVID-19 experienced an outbreak of a novel coronavirus that killed more than eighteen hundred and infected thousands of individuals within the first two months of the epidemic [2].

2.1.1 Symptoms

More recently, the epicenter has moved to other cities in Europe and then in America. The patients' most notable found symptoms (according to the collected experimental data) are dry cough, dyspnea, fever, and bilateral lung infiltrates on imaging. Initially, all the cases were associated with Wuhan's Huanan Seafood Wholesale Market, which trades in seafood and a wide variety of live animal species. Due to the many reported cases up to January 30th, 2020, the World Health Organization (WHO) declared the Chinese outbreak of COVID-19 to be a Public Health Emergency of International Concern posing a high risk to countries with vulnerable health systems around the world[3].

2.1.2 Spread of COVID-19

Close physical contact, respiratory droplets, and touching contaminated surfaces are the most common ways for the virus to spread. The most difficult element of its transmission is that a person might be infected with the virus for days without showing symptoms. Because of the causes of its development and the threat it poses, practically all governments have declared partial or complete lockdowns in the impacted regions and cities. Medical

researchers from all across the world are working on effective vaccinations and treatment for the disease. The first mass vaccination programme started in early December 2020 and the number of vaccination doses administered is updated on a daily basis. At least 13 different vaccines (across 4 platforms) have been administered.

In that reference, it is critical to develop models that are both computationally capable and realistic in order to assist policymakers, medical personnel, and the general public. Modeling the disease, identifying areas of potential high risk, and forecasting the number of likely total cases can help the medical system prepare for the incoming patients.

2.2 Machine Learning

Machine Learning is a branch of Artificial Intelligence that uses training datasets to train machines. We can discover patterns, evaluate data, make better predictions,and make the right judgments with no or minimal human interaction using machine learning.[4]. Over the last decade, machine learning has established itself as a major subject of research by addressing many of the extremely complicated and sophisticated real-world issues. Almost every real-world domain was represented, including healthcare, autonomous vehicles (AV), commercial applications, gaming, climate modeling, speech, and image processing.[5]. There are countless applications of machine learning; it has been employed in the modeling of coronary artery disease[6], cardiovascular disease prediction[7], and breast cancer prediction[8].
Machine learning may be divided into three categories: supervised learning, unsupervised learning and reinforcement learning as shown in figure 2.2

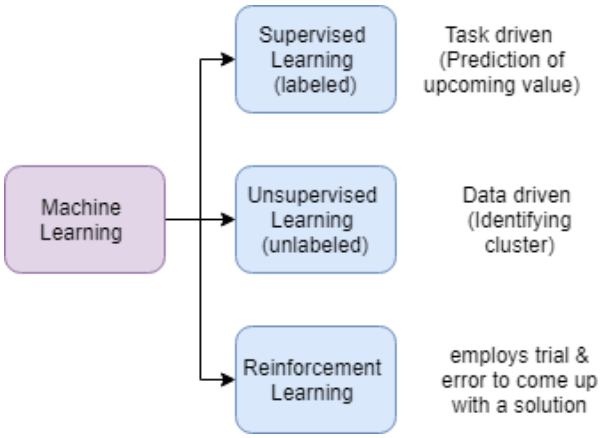


Figure 2.1: Machine Learning classifications

2.2.1 Supervised Learning

A supervised learning algorithm is built to make a prediction when it is provided with an unknown input instance. The learning algorithm uses a dataset of input instances

and their corresponding regressor to train the regression model in this learning approach. After that, the trained model makes a prediction for the unpredicted input data or valid dataset.[5]

2.2.2 Unsupervised Learning

Unsupervised Learning is a a machine learning algorithm that is used to create interruptions from unlabeled input data responses in data sets. Unsupervised learning, which is used to analyse exploratory data, includes cluster analysis. This approach has been proven to be highly beneficial for organizing data or discovering hidden patterns.[9]

2.2.3 Reinforcement Learning

Reinforcement learning is a learning algorithm that employs a reward and penalty-based system to interact with the environment.The agent is rewarded if it makes a good judgement and penalised if it makes a bad one. This technique enables the agent to develop behaviour that is optimal for achieving the intended result.[9].

2.3 Dimensionality Reduction

There are two primary reasons why dimensionality reduction is utilised. The initial goal is to aid with visualising.There are frequently too many characteristics on which the final categorization is based. The more characteristics there are, the more difficult it is to understand the training set and subsequently work on it. Dimensionality reduction techniques are useful in this situation. We may plot a summary of the data on a 2D plane by reducing or summarizing the number of characteristics to just two components.[10].One of the techniques that are is here: Principal Component Analysis (PCA).

2.3.1 Principal Component Analysis

Principal Components Analysis (PCA) is a well-known unsupervised dimensionality reduction approach that creates meaningful features/variables by combining the original variables in linear (linear PCA) or non-linear (kernel PCA) ways [11]. PCA creates a low-dimensional representation of the data that exhibits a large in which the data's variance is maximized. The importance of each feature is reflected by the magnitude of the corresponding values in the eigenvectors (higher magnitude — higher importance). shows a graph with three dimensions representing data. Since maximum variance is attained, PCA transforms the original data space into a component space with only two dimensions, making it easier to view the data and understand how various groups are grouped.

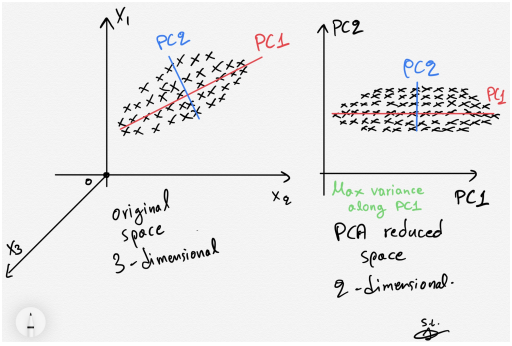


Figure 2.2: home made sketch by the author Serafeim Loukas to how the PCA may be used to minimize data dimensionality and maximise variance. Source : [11]

2.4 Regression Models

The regression method of machine learning is a forecasting methodology to examine the relationship between a dependent and an independent variable [9]. In this study , we applied Prediction using Machine Learning Models:

- Linear Regression
- Support Vector Regression (SVR)

Time Series Forecasting Models:

- Holt's Linear Model
- Holt's Winter Model
- AR Model
- MA Model
- ARIMA Model
- SARIMA Model

2.4.1 Linear Regression

In regression modeling, a target class is predicated on the independent features[8]. This approach may be used to determine the relationship between independent and dependent variables, as well as to forecast. Linear regression, a kind of regression modelling, is the most widely used statistical approach in machine learning for predictive analysis. In linear regression, each observation is based on two values: the dependent variable and the independent variable. There are two factors (x,y) that are involved in linear regression

analysis shown in equation below

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

or equivalently

$$E(y) = \beta_0 + \beta_1 x \tag{2.2}$$

ϵ is the error term of linear regression. The error term here uses to account the variability between both x and y , $\beta 0$ represents y-intercept, $\beta 1$ represents slope.
To bring the notion of linear regression into a machine learning framework, x represents the input training dataset, and y represents the class labels contained in the input dataset. The machine learning algorithm's objective is then to identify the optimal values for $\beta 0$ (intercept) and $\beta 1$ (coefficient) in order to obtain the best-fit regression line.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$g = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Here, g is called a cost function, which is the root mean square of the predicted value of y (predi) and actual y (yi), n is the total number of data points [5]

2.4.2 Support Vector Regression

For both linear and nonlinear regression types, support vector regression is a common choice for prediction and curve fitting. SVR is built on support vector machine (SVM) components, where support vectors are essentially closer points towards the produced hyperplane in an n-dimensional feature space that distinguishes the data points regarding the hyperplane.[12] The SVR model performs the fitting as shown in 2.3.The generelized equation for hyperplane may be respesented as

$$y = wX + b \tag{2.3}$$

where w is weights and b is the intercept at X = 0. The margin of tolerance is represented by epsilon ϵ. The SVR regression madel is imported from SVM class of sklearn python library.The regressor is fit on the training dataset.

2.4.3 Exponential Smoothing

Forecasting is done using data from past periods in the exponential smoothing family methods. As time passes, the effect of previous data observations diminishes exponentially.As a result, the weight attributed to various lag values decreases exponentially. ES

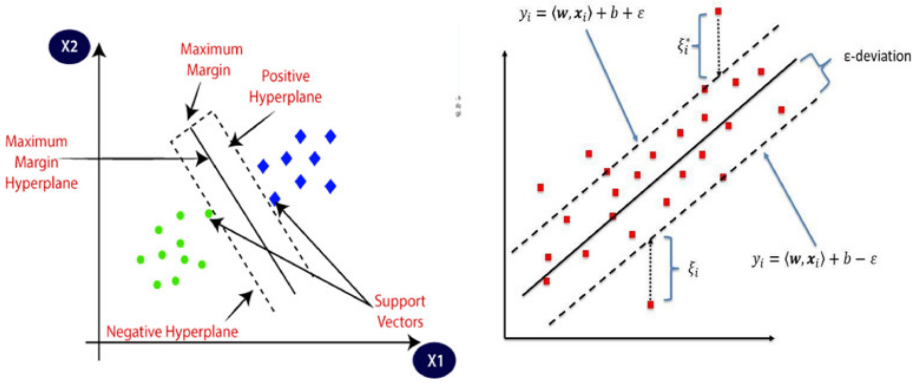


Figure 2.3: Support vector regression model for linear regression fitting where X1= X and X2 = y are the features and label in our case. [Image credit: Source]

.

is a strong time series forecasting approach for univariate data that is relatively simple to use[12].
The forecast for the current time (F_t) in ES is given by:

$$F_t = \alpha A_{t-1} + (1 - \alpha) F_{t-1} \tag{2.4}$$

Here, α smoothing cost where $0 <= \alpha <= A_{t-1}$ is the actual value of the previous period in time series, (F_t) s the forecast value of the previous forecast. There are three main types of exponential smoothing time series forecasting methods.

Single Exponential Smoothing

Single Exponential Smoothing (SES) is a time series forecasting approach for univariate data without a trend or seasonality [13].
The pace at which the effect of previous time steps' observations decays exponentially is controlled by this parameter. The value of Alpha is frequently set to a number between 0 and 1. Large values indicate that the model focuses on the most recent historical observations, whilst smaller values indicate that the model considers more of the history when generating a prediction.

Double Exponential Smoothing

Double Exponential Smoothing with an additive trend which is also refereed to as Holt's linear trend mode, is an Exponential Smoothing extension that explicitly supports trends in univariate time series[13].
In addition to the alpha parameter for controlling smoothing factor for the level, an additional smoothing factor is added to control the decay of the influence of the change in trend called beta (b). The method supports trends that change in different ways: an additive(for linear trend) and a multiplicative(for exponential trend). Holt's linear trend

method involves a forecast equation and two smoothing equations (one for the level and one for the trend):

$$\hat{y}_{t+h|t} = \boldsymbol{\ell}_t + hb_t \qquad \text{(Forescast Equation)}$$
$$\boldsymbol{\ell}_t = \alpha y_t + (1-\alpha)(\boldsymbol{\ell}_{t-1} + b_{t-1}) \qquad \text{(Level Equation)}$$
$$b_t = \beta * (\boldsymbol{\ell}_t - \boldsymbol{\ell}_{t-1}) + (1-\beta^*)b_{t-1} \qquad \text{(Trend Equation)}$$

where $\boldsymbol{\ell}_t$ is an denotes an estimate of the level of the series at time t. b_t is the estimate of trend of series at time t , α is the smoothing parameter for the level ,$0 <= \alpha <= 1$ amd β^* is the smoothing parameter for the trend,$0 <= \beta^* <= 1$

Triple Exponential Smoothing

Triple Exponential Smoothing is sometimes called Holt-Winters Exponential Smoothing. This method is an extension of Exponential Smoothing that explicitly adds support for seasonality to the univariate time series.[13]
A new parameter, gamma (g), is added to the alpha and beta smoothing parameters to manage the effect on the seasonal component. Seasonality may be represented as an additive(for linear seasonality) or multiplicative (for exponential change in seasonality). Hyperparameters:

- **Alpha :** Smoothing factor for the level.
- **Beta :**Smoothing factor for the trend.
- **Gamma :**Smoothing factor for the seasonality.
- **Trend Type :**Additive or multiplicative.
- **Dampen Type :**Additive or multiplicative.
- **Phi :**Damping coefficient.
- **Seasonality Type :**Additive or multiplicative.
- **Period :**Time steps in seasonal period.

2.4.4 ARIMA/SARIMA MODEL

Autoregressive models (AR)

In an autoregression model, we use a linear combination of the variable's previous values to forecast the variable of interest. The term autoregression means that the variable is

being regressed against itself[13]. Thus, an autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t \tag{2.5}$$

where ϵ_t is white noise.AR(p)model,an autoregressive model of order ρ.

Autoregressive models are incredibly adaptable when it comes to dealing with a variety of time series patterns. As shown in below figure 2.4 AR(1) and AR(2) model. Changing the parameters ϕ_1,..,ϕ_p results in results in different time series patterns.ϵ_t -Variance of the error term - will only change the scale of the series, not the patterns.

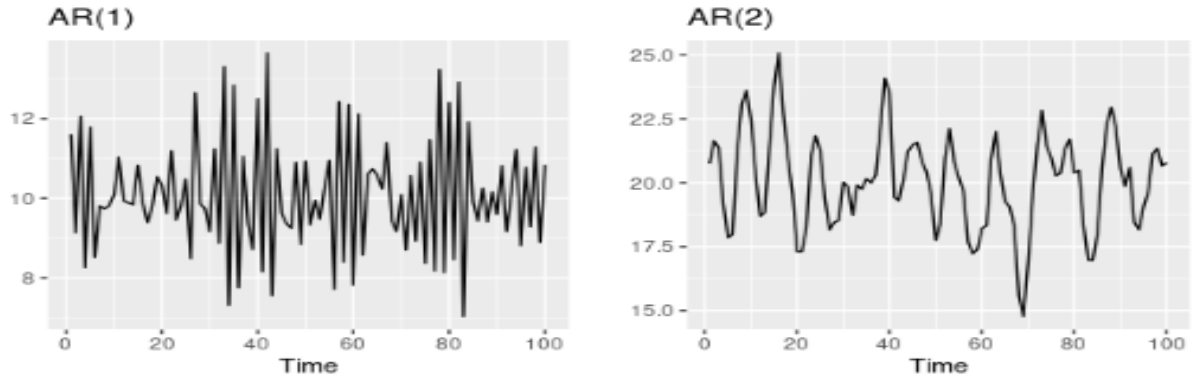


Figure 2.4: Two examples of data from autoregressive models with different parameters.Source [13]

.

Moving average models(MA)

Instead of using past values of forecast variable in a regression, a MA model uses past forecast errors in a regression-like model[13].

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + .. + \theta_q \epsilon_{t-q} \tag{2.6}$$

where ϵ_t is white noise. We refer to this as an MA(q) model, a moving average model of order q. Notice that each value of y_t can be thought of as a weighted moving average of the past few forecast errors.

ARIMA Model

Autoregressive Integrated Moving Average, or ARIMA (in this context, "integration" is the reverse of differencing), is a forecasting method for univariate time series data. A problem with ARIMA is that it does not support seasonal data. That is a time series with a repeating cycle. [13]. The full model can be written as

$$y'_t = c + \phi_1 y'_{t-1} + .. + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t \tag{2.7}$$

where y'_t the differenced series. "Predictors" on the right side include lagged values of y_t and lagged errors. This is what known is as an ARIMA(p,d,q) model where **p :** order of the autoregressive part and **d :** degree of first differencing involved and **q :** order of the moving average part.

SARIMA

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.Additional seasonal variables are added to the ARIMA models we've examined so far to create a seasonal ARIMA model [13]

$$ARIMA \quad \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}} + \underbrace{(P,D,Q)_m}_{\text{seasonal part of the model}}$$

where m= number of observations per year.The seasonal part of the model consists of terms that are similar to the non-seasonal components of the model, but involve backshifts of the seasonal period

2.5 Clustering Model

Clustering, also known as the partitioning process, is a machine learning approach in which a set of patterns is sorted into distinct clusters. The clusters that are similar are grouped in the same manner.In this study, Kmeans will be covered which is considered as one of the most used clustering algorithms [9].

2.5.1 K-means

The Kmeans algorithm is an iterative technique that attempts to split a dataset into K separate non-overlapping subgroups (clusters), each of which contains just one data point [14].It allocates data points to clusters in such a way that the sum of the squared distances between them and the cluster's centroid (arithmetic mean of all the data points in that cluster) is as little as possible.It attempts to make intra-cluster data points as comparable as feasible while maintaining clusters as distinct (far) as possible[14]. Within clusters, the less variance there is, the more homogenous (similar) the data points are. K-means work as follows :

1. determine the number of clusters K
2. Initialize the centroids by shuffling the dataset first, then picking K data points at random for the centroids without replacing them.
3. Repeat until there is no change to the centriods, illustered in fig 2.5

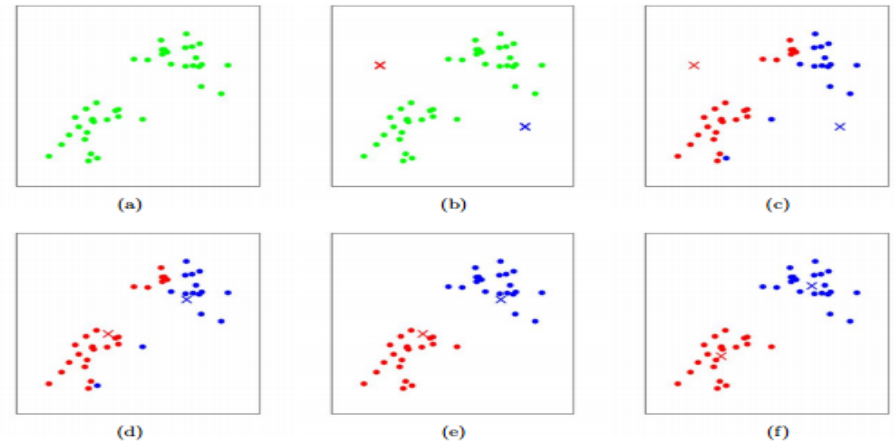


Figure 2.5: K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid Source :Source

2.6 Cluster Validity indices

Cluster Validity indices is also refered to as Evaluation Methods to evaluate how well the models are performing based on different K clusters since clusters are used in the downstream modeling [14]. The metrics used in this study :

2.6.1 Elbow Method

Based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids, the elbow method gives us an estimate of what a suitable k number of clusters might be [14]. We pick k at the spot where SSE starts to flatten out and forming an elbow as shown in 2.6 . Because the curve is monotonically declining and may not exhibit any elbow or have an evident point where the curve starts flattening out, it might be difficult to determine an appropriate number of clusters to use.

2.6.2 Silhouette Analysis

The degree of separation between clusters may be determined via silhouette analysis.Each sample :

- Calculate a^i which is the average distance from all data points in the same cluster
- Calculate b^i which is the average distance from all data points in the closest cluster

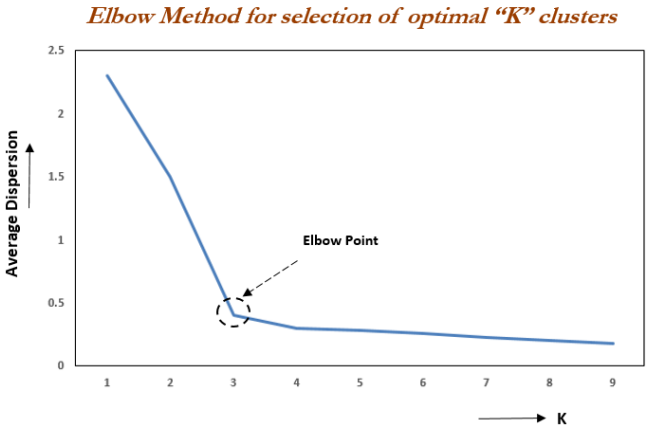


Figure 2.6: Elbow Method for selection of optimal K. Source :Source

- Compute the coefficient:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

Coefficients can take values ranging from -1,1. The bigger the coefficients and the closer to 1 the better the clusters[14].

2.6.3 Davies–Bouldin index

The Davies–Bouldin index (DBI) is an internal evaluation scheme, where the validation of how effectively the clustering has been done is accomplished using quantities and attributes inherent to the dataset, is a measure for assessing clustering algorithms[15]. The lower the DB index Value, the better the clustering. However, A disadvantage of this approach is that a good value does not always signify the best information retrieval[15].

2.7 Prediction Evaluation Parameters

To evaluate the performance of each model: mean absolute error (MAE), mean square error (MSE) , root mean square error (RMSE), and mean absolute percentage error (MAPE) are applied.

2.7.1 Mean Absolute Error (MAE)

MAE is a very good Key Performance Indicator to measure forecast accuracy. MAE is the average magnitude of the errors in the set of model predictions[16]. This is an average of the model predictions and actual data on valid data, with all individual differences

given equal weight. Its matrix value range is 0 to infinity, and lower score values indicate better learning models, which is why it's also known as negatively-oriented scores[5].MAE is calculated as :

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y_j}| \tag{2.8}$$

2.7.2 Mean Square Error (MSE)

Mean square error is a way to measure the performance of regression models[5]. MSE squares the distance between data points and the regression line. Because squaring eliminates the negative sign from the value, it lends greater weight to bigger differences. Smaller MSE demonstrates that finding a best fit is close. MSE calculated as :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{2.9}$$

2.7.3 Root Mean Square Error (RMSE)

Root mean square error can be defined as the standard deviation of the prediction errors.The distane from the best fit line and the actual data points are known as Prediction errors and also referred to as residuals. The root mean square error (RMSE) is a measurement of how concentrated the real data points are around the best fit line. It's the error rate calculated using the square root of MSE[5], as shown below :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{2.10}$$

2.7.4 Mean Absolute Percentage Error(MAPE)

The mean absolute percentage error (MAPE) measures the accuracy of a forecast system.It is determined as the average absolute percent error for each time period minus actual values divided by real values, and it is expressed as a percentage.The mean absolute percentage error (MAPE) is the most often used measure for predicting error, and it works best when the data does not include any extremes [16].MAPE is calculated as :

$$MAPE = \frac{1}{n}\sum_{j=1}^{n}|\frac{y_j - \hat{y_j}}{y_j}| \tag{2.11}$$

where **n** is the number of fitted points, y_j is the actual value, and $\hat{y_j}$ is the forecast value.

2.8 Literature Review

2.8.1 Spatial Analysis Studies

Previous study[1] aimed at the analysis of the spatial evolution of the Coronavirus pandemic through a certain type of unsupervised neural network, which is called self-organizing maps(SOM) where there is no given output target.
Depending upon the clustering abilities of self-organizing maps they were able to spatially group countries that are similar according to their coronavirus cases, analyzing countries that are behaving similarly can guide us by using similar strategies in dealing with the spread of the virus.

Publicly available datasets of coronavirus cases that have occurred from January 22, 2020, to May 13, 2020, were used in this study. The Dataset which includes data from the countries was obtained from Humanitarian Data Exchange (HDX) and also the data set of 32 states of Mexico was obtained from Mexico's Government website along with data about both diabetes and hypertension; to track the similarities between these diseases and COVID-19.

The SOM has formed clusters of countries in the world into 4 classes -according to the severity of Coronavirus cases: Very High, High, Medium, and Low. Clusters formed with the SOM method were plotted, indicating the classes for the COVID-19 recovered cases, classes for the COVID-19 Confirmed cases, and classes of death for the January 22 of 2020 to May 13 of 2020 period of time. Furthermore, in this study concerning the 32 states of Mexico, it was also noticed that there is a similarity between states with a higher number of deaths to the states with higher numbers of Hypertension cases and the same goes for the number of Diabetes cases. In this regard, constructing a model using the number of cases of both hypertension and diabetes can help us estimate the number of COVID cases.

This study concluded that the clustering abilities of self-organizing maps allowed to spatially group countries that are similar according to their coronavirus cases, in order to analyze which countries are behaving similarly and thus can benefit by using similar strategies in dealing with the spread of the virus.

Another study [17] aimed to use a spatial scan to identify clusters of testing rates, positivity rates, and proportions of tests that were positive to understand test access and case burden. In addition to evaluating contextual factors associated with these clusters as well as across all of New York City.

This study used data of a total number of COVID-19 tests and the total number of positive COVID-19 tests aggregated by zip code were provided by New York. City Department of Health as of April 12, 2020. Taking into consideration all factors provided via the IPUMS National Historical Geographic Information System such as Zip code total population, race, Hispanic ethnicity, citizenship status, health insurance status, mode

of transportation to work, educational attainment, median household income, receipt of public assistance payments, rent as a proportion of income, and poverty data.

Global Moran's I test, using simple adjacency as the neighborhood definition, was used in this study to investigate the presence of spatial autocorrelation (clustering) in each the outcome, where each covariate was created using a 4 quantile categorization.In this In the study, there were 177 zip codes for which testing data were available and the results were as follows: the mean COVID-19 testing rate across zip codes was 21.6 tests per 1000 people(MIN =8.6,MAX=42.8, SD=7.3), the mean of positivity rate was 12.1 per 1000 people (MIN=2.9, MAX=25.9, SD=4.7) and the mean proportion of COVID-19 positive tests was 0.55(MIN=0.23 ,Max=0.79,SD=0.097).

This study faced some limiting factors, first that the analysis only describes associations between COVID-19 testing patterns and contextual factors of the zip code and there was no claim made for a casual relationship between any of the variables examined. Second, this study is subject to the modifiable areal unit problem as in any spatial analysis. Third, the zip codes are relatively small spatial units. Forth, many wealthy residents in New York City have left the city, which will artificially inflate the population denominator in such areas and drive down the test rates in those areas.

This study concluded that areas with lower test rates and lower proportions of those tests being positive are likely the result of less severe illness and track with higher income, education, and white populations. Areas with higher test rates and higher proportions of positive tests point to more severe cases, which are disproportionately in areas of the black population, uninsured, and have rent$\geq$ 50 percent of income. The third pattern of lower test rate areas coupled with higher positive test proportions may indicate severe illness, but inadequate testing, which appeared among areas with non-citizens and high use of public transportation.

In another research, [18] developed by researchers at a startup company Akai Kaeru LLC. This study aimed at using a pattern mining algorithm based on the FP-growth algorithm [19] in yielding a better understanding of COVID-19 as well as assisting authorities in predicting future COVID-19 deaths rates and helping in the allocation of resources. The pattern mining algorithm is used to analyze the data set used in this research with 500 attributes for all 3,007 US counties.

The findings in this research reveal that it is usually a combination of features that exposes a country to high COVID-19 death rates than just one feature. In this study, 2 examples of patterns were extracted for predictive analysis; In Pattern 1, it was shown that poor, aging, and Rural Countries are at High COVID-19 risk and that the bar of confidence read 30 % of its full length with just counties with high poverty rates. Adding the "age greater than 65 " as a second dimension boosted the Bar of Confidence to 64 % as well as adding the "population density" as a third feature confirms that the Bar of Confidence maxed out its 100 %.On the Other hand pattern 2, reveals that counties

with a low Asian but high minority population where black children live in poverty hard hit by the COVID-19 virus, and the bar of confidence read 30 % of its full length with just counties with a low Asian population. To sharpen our risk assessment a second feature was added "low rate of Asians but a strong minority population".Fittingly the Bar of Confidence reads 71 % as well as adding the "percentage of black children living in poverty" as a third feature confirms that the Bar of Confidence maxed out its 100%. Predictive Analysis was applied to both patterns.

This study concluded that pattern analysis can be highly effective in defining the characteristics of counties at risk of elevated COVID-19 death rates and that these characteristics also allow reliable predictions of future death rates, and that the death rates in May with those in June and found that for 98% of our patterns the death rate growth was 2–3 times higher than the US average, while the remaining 2% grew at the average pace, and none slowed in growth below the US average.

A study [20] aimed to report data for the first three clusters of COVID-19 cases in Singapore; Cluster A, linked to a tour group from China, Cluster B, a group linked to exposure at a conference, and Cluster c, linked to a visit to church were identified in Singapore in February 2020. The findings of this study will be important for countries and cities to calibrate detection and response efforts during the ongoing epidemic. In this study, epidemiological and clinical data was gathered from individuals with confirmed COVID-19, via interviews and inpatient medical records, and field investigations to assess interactions and possible modes of transmission of COVID-19.

This study reported the median (IQR) incubation period, defined as the duration between estimated dates of infection and reported symptom onset, using R.One of the limiting factors this study faced is that the small sample size used to ascertain the incubation period because primary cases could not be identified with certainty. Moreover, symptom-onset dates and the movement of and exposure history of cases detected overseas were either based on media reports or were unknown. The study concluded that direct transmission could be possible by contact or indirect transmission, as the local cases in cluster C were identified through enhanced pneumonia surveillance of people who had no history of travel to China, whereas the initial cases in cluster A direct physical contact was reported. Handshaking and physical contact during team-building activities and sharing of meals were reported among participants of the business meeting (cluster B).

2.8.2 Predictive Models Studies

Previous Study [21] mainly focused on developing a decision tree algorithm on the COVID-19 global real-time data in order to utilize supervised machine learning algorithms for time-series forecasting. The aim of this study is to contribute by providing the governments and health authorities with the required information that helps in planning and decision-making and decrease the anxiety of the population as well as allowing them

to deal with the next phases of the pandemic. The data used in this study were collected from official data repositories such as Johns Hopkins University, WHO and Worldometer official website. These data features are the daily total COVID-19 confirmed positive cases, daily and total deaths, and the total and daily recoveries.

Decision tree algorithm and linear regression are the algorithms proposed for this study in predicting sequence and time-series data-related problems. The proposed method has forecasted the possible confirmed cases for the upcoming 7 days for the USA. Experimental results showed that the confirmed cases are exponentially increasing from a few hundreds of thousands to nearly two and a half million. For validation root mean square error was used. we performed a comparative study using the most up-to-date methods. A proposed model is compared with various state-of-the-art models (random forest, ARIMA, and deep learning) and the accuracy of the machine learning models on the training dataset is evaluated using root mean square error (RMSE) and mean absolute error (MAE). The proposed method also depicts the possible stoppage of the pandemic using the normal distribution. It specifically presents the statistical estimation of the slow down period of the pandemic which is extracted based on the concept of normal distribution.

Another study [22] aimed to developed a model and then employed it for forecasting future COVID-19 cases in India. The statistical prediction models are useful in forecasting as well as controlling the global epidemic threat.Auto-Regressive Integrated Moving Average (ARIMA) model for predicting the incidence of 2019-nCov disease. As compared to other prediction models, for instance support vector machine (SVM) and wavelet neural network (WNN), ARIMA model is more capable in the prediction of natural adversities. For our study, we have identified the best ARIMA model and then predicted the number the cases for the next 20 days. The main objective of the study is to find the best predictive model and apply it to forecast future incidence of COVID-19 cases in India. Confirmed, recovered and death cases of COVID-19 infection are collected for India as well as countries with highest confirmed infection (US, Spain, Italy, France, Germany, China and Iran) and countries in South-East Asia region (India, Indonesia, Thailand, Bangladesh, Sri Lanka, Maldives, Nepal, Bhutan and Timor-Leste), as per World Health Organization region classification, from the official website of Johns Hopkins University from 22 January 2020 to 13 April 2020. MAPE, MAD, MSD were used as a measure of accuracy followed by verification of the model.the MAPE, MAD and MSD values suggests that ARIMA (2, 2, 2) model is the most accurate of all for forecasting future incidences as it possesses the least value for all the measures.

Previous study [23] has employed two-layers LSTM and XGBoost models to establish daily confirmed infected cases prediction models for the time series data of America. The number of daily new confirmed COVID-19 cases in time series is collected from the World Health Organization website . The data set is available in time series format with date, month, and year to ensure that the time component is not overlooked.In order to anticipate future diseases, those models actively learn real-time data from current COVID-19

observations. This study faced some limiting factors one of them was the limiting factor : absence of a broadly available COVID-19 vaccination. This study concluded that the results of test set show that MAPE of the LSTM and XGBoost algorithms reach and, respectively that the LSTM model has the lower metrics value. In addition, the models project that the country has a tentative range between 30,000 to 70,000 new cases by October.

In another study [5] , it aims to provide an early forecast model for the spread of novel coronavirus.This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. In particular, four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and the number of recoveries in 10 days in the future. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the ES performs best among all the used models followed by LR and LASSO which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while SVM performs poorly in all the prediction scenarios given the available dataset.The modeks have been trained using g the COVID-19 stats dataset provided by Johns Hopkins.

Parbat and Chakraborty [12] have implemented the Support Vector Regression (SVR) for predicting the COVID-19 cases in India for 60 days using the time-series data reported for the period of 1st March 2020 to 30th April 2020. Their results indicate that the SVR model has an accuracy of 97 % in predicting the cumulative fatalities cases, cumulative recovered cases, cumulative confirmed cases. Their model also able to predict the daily new COVID cases with an accuracy of 87 %.

Another study [24] used Auto-Regressive (AR) models based on two-piece scale mixture normal distributions to forecast the confirmed and recovered COVID-19 cases. Their model performed well in forecasting confirmed and recovered global COVID-19 cases.

K.E. ArunKumar et al. [25] employed time-series models — Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) to forecast the epidemiological trends of the COVID-19 pandemic for top-16 countries where 70%–80% of global cumulative cases are located. Evaluation metrics Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE) were used as criteria for selecting the best model. The data used in this work is obtained from publicly available John Hopkins University's COVID-19 database.It is found that the COVID-19 forecasted value of the 60th day from the ARIMA and SARIMA models are more or less the same but to capture the seasonality or trends of the data SARIMA models outperform the ARIMA models.

Chapter 3

Methodology

3.1 Methodological Approach Overview

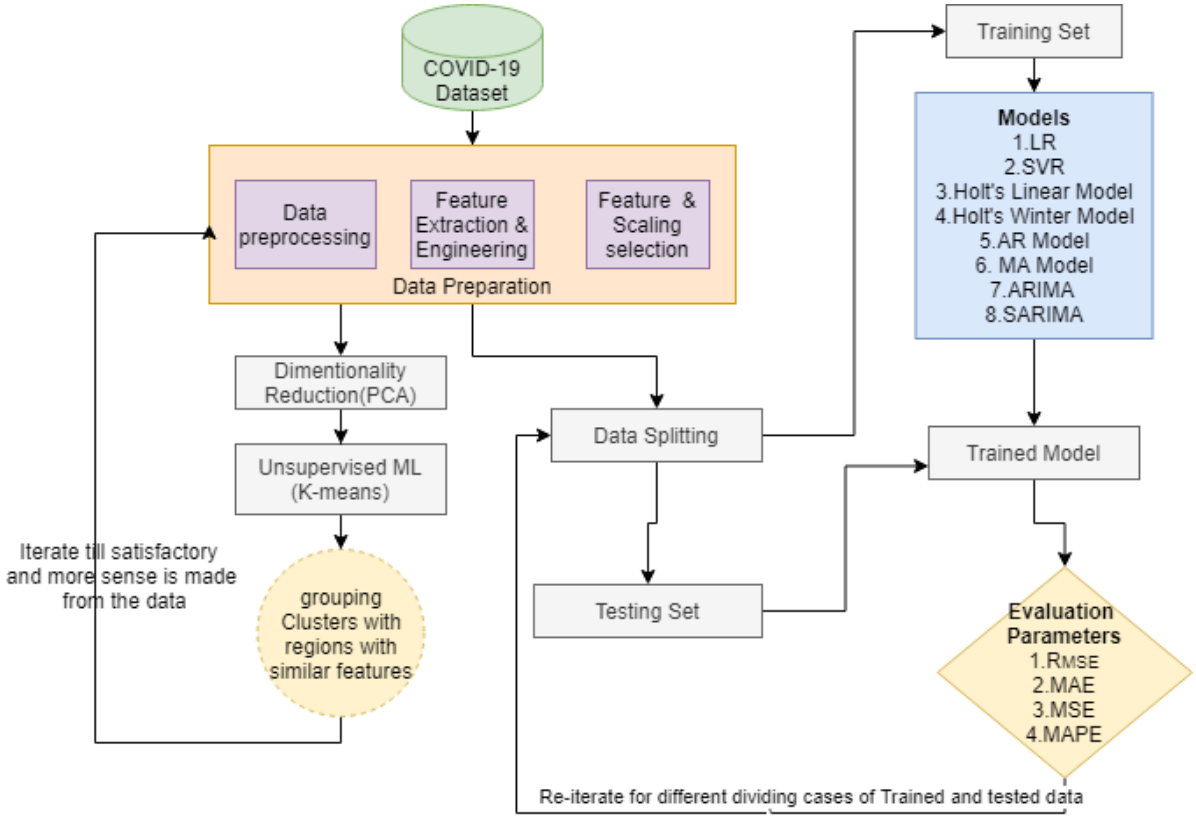


Figure 3.1: Workflow

The first step is where we start by data preprocessing; the purpose of this step is to prepare the data by retrieving it from files, cleaning it, and emphasizing important details. After that, we go through feature extraction and engineering which is about identifying

which features can aid in the creation of accurate results and finally the features go through standardization. The workflow here splits into 2 paths, one heading towards the clustering approach and the other towards the prediction approach:

1. Clustering: dimensionalty reduction set the features to produce the desired results. This path will then apply Unsupervised ML that would group my regions with similar features and then we can repeat with different Normalization to the data. K-means evaluation is then applied to identify the optimal k clusters.

2. Prediction: data splitting would split the data into training data and validation data; where the learning models will be used to train on my training data. The models have been trained on the total confirmed cases. Then the learning models have then been evaluated based on important metrics such as MSE, MAPE, RMSE, and MAE and reported in the results. Finally, we repeat with a different Data splitting percentage.

3.2 Method of data collection

This study aims to group counties in New York State, and to cluster states in the USA with potential similar risks of COVID-19 during wave 1. The second part of the study aids in understanding the COVID-19 spread in the US by predicting the total number of positive cases during different time intervals . The dataset used in the study has been obtained from the GitHub repository provided by GoogleCloudPlatform [26].

The data set includes open, publicly sourced, licensed data with several tables relating to demographics, economy, epidemiology, geography, health, hospitalizations, mobility, government response, weather, and more. Along with daily time series of +20,000 global sources.

From the several tables provided in my dataset, data samples from Epidemiology Table 3.1 ,Demographic table 3.2, Geography table 3.3, and Mobility table 3.4 were chosen as they included most valuable data concerning this COVID-19 study.

The records from all tables mentioned above were joined by date and key. These joined data could be also found in the Aggregated table in [26],but the size of the Aggregated table is so large and will take longer time to pre-process that's why the tables mentioned above were chosen and joined.

A sample of the features after pre-processing is shown in both table 3.5, and 3.6.

Epidemiology			
Name	Type	Description	Example
date	string	(yyyy-mm-dd)	2020-03-30
key	string	unique name	CN_HB
new_confirmed	int	Daily Cases	34
new_deceased	int	daily deaths	2
new_recovered	int	daily recovered	13
new_tested	int	daily tested	13
total_confirmed	int	cumulative Cases	6447
total_deceased	int	cumulative Death	6447
total_tested	int	cumulative Tests	133
total_recovered	int	cumulative Recovery	133

Table 3.1: Information related to the COVID-19 infections for each date-region pair

Demographics			
Name	Type	Description	Example
key	string	unique name	KR
population	int	Humans count	51606633
population_male	int	male count	25846211
population_female	int	female count	25760422
population_female	int	population/area in km	529.35
population_age low_up	int	pop between the age of low and up	42038247

Table 3.2: Information related to the population demographics for each region.

Mobility			
Name	Type	Description	Example
date	string	(yyyy-mm-dd)	2020-03-30
key	string	unique name	KR
mobility_grocery_and_pharmacy	double%	%change to visit shops	-15
mobility_workplaces	double%	%change to visit work	-15
mobility_retail_recreation	double%	%change visit cafes,etc	-15
mobility_residential	double%	%change to visit resident places	-15

Table 3.3: These datasets show how visits and length of stay at different places change compared to a baseline(The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020) .

Geography			
Name	Type	Description	Example
key	string	unique name	CN_HB
latitude	double	geographic coordinate	30.9756
longitude	double	geographic coordinate	30.112.2707
area	int[squared kilometers]	region area	3729

Table 3.4: Information related to the geography for each region.

key	date	coun	subregion1_	subregio	new_confirm	new_decease	new_tested	total_confirme	total_decease	total_tested
US_NY_36001	01/03/2020	US	New York	Albany Cou	0	0	0	0	0	0
US_NY_36001	02/03/2020	US	New York	Albany Cou	0	0	0	0	0	0
US_NY_36001	03/03/2020	US	New York	Albany Cou	0	0	0	0	0	0
US_NY_36001	04/03/2020	US	New York	Albany Cou	0	0	42	0	0	42
US_NY_36001	05/03/2020	US	New York	Albany Cou	0	0	-18	0	0	24
US_NY_36001	06/03/2020	US	New York	Albany Cou	0	0	-8	0	0	16
US_NY_36001	07/03/2020	US	New York	Albany Cou	0	0	8	0	0	24
US_NY_36001	08/03/2020	US	New York	Albany Cou	0	0	1	0	0	25
US_NY_36001	09/03/2020	US	New York	Albany Cou	0	0	-4	0	0	21
US_NY_36001	10/03/2020	US	New York	Albany Cou	0	0	7	0	0	28
US_NY_36001	11/03/2020	US	New York	Albany Cou	0	0	42	0	0	70

Table 3.5: Sample of my data

total_tested	population	pop_male	pop_female	pop_0_9	lat	long	area	mobility_retail_and_	mobility_grocery_an	mobility_workplace	mobility_residenti
0	307717	149017	158700	31227	43	-73.8	1381	10	13	7	-1
0	307717	149017	158700	31227	43	-73.8	1381	11	15	3	0
0	307717	149017	158700	31227	43	-73.8	1381	8	15	3	-1
42	307717	149017	158700	31227	43	-73.8	1381	7	8	3	0
24	307717	149017	158700	31227	43	-73.8	1381	5	13	3	-1
16	307717	149017	158700	31227	43	-73.8	1381	6	10	3	0
24	307717	149017	158700	31227	43	-73.8	1381	12	12	6	-1
25	307717	149017	158700	31227	43	-73.8	1381	9	13	2	-1
21	307717	149017	158700	31227	43	-73.8	1381	10	16	2	-1
28	307717	149017	158700	31227	43	-73.8	1381	0	9	2	1
70	307717	149017	158700	31227	43	-73.8	1381	3	15	0	1

Table 3.6: Continuation of my data sample

3.3 Data Preparation

Both the prediction and clustering approaches require the data to be in a particular form, this is where the data preparation techniques are used. For prediction of total cases in the US , the total confirmed cases was the only features used to predict future number of cases. However for clustering a multi phase preparation process is required. Feature extraction and engineering, feature and scaling selection and rescaling (Min-Max Scaling).

3.3.1 Feature Extraction and Engineering

To group the regions according to probable similar risk, the features will be likely related to cases, deaths, tested along with mobility, geography, and demographics. The features are computed for Wave 1 of COVID-19.The candidate features that engineered are mentioned below in 3.7

Table 3.7: My features

Table 3.8

KEY
Date
Subregion_Name1 and 2
Total_confirmed
Newly_confirmed
Total_Tested
Newly_Tested
Total_deceased
Newly_deceased
confirmedtests %
population
population_density
population_of_ male_and_female
population_age_low_Up
mobility_workplaces
mobility_residential
mobility_grocery_and_pharmacy
mobility_retail_and_recreation
area
latitude
longitude

Table 3.9

KEY
Date
Subregion_Name1 and 2
Total_confirmed_Percentage
New_confirmed_Percentage
Total_tested_Percentage
New_tested_Percentage
Total_deceased_Percentage
New_deceased_Percentage
confirmedtests %
population_density
population_ male_and_female_Percentage
population_age_low_Up_Percentage
mobility_workplaces
mobility_residential
mobility_grocery_and_pharmacy
mobility_retail_and_recreation
area
latitude
longitude

The value names presented in the previous figures each are used for a particular goal and could be explained as follows:

- **Sub region1 name:** is an American English name of the sub region, subject to change
- **Sub region2 name:** is American English name of the county (or local equivalent), subject to change
- **Total Confirmed:**Cumulative sum of cases confirmed after positive test to date
- **Total Deceased are:** Cumulative sum of deaths from a positive COVID-19 case to date
- **Total Tested:** Cumulative sum of COVID-19 tests performed to date

* Cumulative count will not always amount to the sum of daily counts, because many authorities make changes to criteria for counting cases, but not always make adjustments to the data. There is also potential missing data.

- **New Confirmed:** Count of new cases confirmed after positive test on this date.
- **New Deceased:** Count of new deaths from a positive COVID-19 case on this date.
- **New Tested:** Count of new COVID-19 tests performed on this date

  * Values can be negative, typically indicating a correction or an adjustment in the way they were measured. For example, a case might have been incorrectly flagged as confirmed one date so it will be subtracted from the following date.

- **Confirmed/tests%:** It is the percentage of Confirmed cases with reference to the number of people that did the tests.
- **Population:** Total count of humans in a region.
- **Population_Male:** Total count of males.
- **Population_female:** Total count of females.
- **Population age low to high:** Estimated population between the ages. of *low* and *high*, both inclusive.
- **Mobility_workplace:** Percentage change in visits to places of work compared to baseline.
- **Mobility_residential:** Percentage change in visits to places of residence compared to baseline.
- **Mobility_retail_and_recreation:** Percentage change in visits to restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters compared to baseline.
- **Mobility_grocery_and_pharmacy:** Percentage change in visits to places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies compared to baseline.
- **Total confirmed Percentage:**Cumulative sum of cases confirmed after positive test to date, divided by population of the region, and multiplied by 100.
- **New confirmed Percentage:**Count of new cases confirmed after positive test on this date, divided by population of the region, and multiplied by 100.
- **Total Tested Percentage:**Cumulative sum of COVID-19 tests performed to date, divided by population of the region, and multiplied by 100.
- **New Tested Percentage:**Count of new COVID-19 tests performed on this date, divided by population of the region, and multiplied by 100.
- **Total Deceased Percentage:**Cumulative sum of deaths from a positive COVID-19 case to date, divided by population of the region, and multiplied by 100.

- **New Deceased Percentage:**Count of new deaths from a positive COVID-19 case on this date, divided by population of the region, and multiplied by 100.
- **Population male percentage:** Total count of males, divided by population of the region, and multiplied by 100.
- **Population female percentage:** Total count of females, divided by population of the region, and multiplied by 100.
- **Population age low to high percentage:** Estimated population between the ages. of *low* and *high*, both inclusive, , divided by population of the region, and multiplied by 100.

3.3.2 Feature and Scaling Selection

Visualizing the data, helps in better understanding the data and the importance of the features used to help in the selection of the features. The data was visualized in an animated COVID-19 report as shown in [28] and [29]. Figures for the data visualizaton was mentioned below in figure 3.2 . After Visualizing the data, Manual selection is applied by selecting the features manually based on trial and and error with aim to perform better clustering algorithms.

The features will be covered in the study are:

1. As shown in 3.8
2. After normalizing them by population of each region.3.9
3. After normalizing features in 3.8 by population density of each region

3.3.3 Rescaling(Min-Max Scaling)

Feature scaling is a method used to normalize the range of independent variables or features of data. In this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers [30]. A Min-Max scaling is typically done via the following equation :

$$X_{sc} = \frac{X - min(X)}{max(X) - min(X)} \tag{3.1}$$

Where X is the is an original value, X_{SC} is the normalized value.
To rescale a range between an arbitrary set of values [min, max], the formula becomes:

$$X_{sc} = min + \frac{(X - min(X))(max - min)}{max(X) - min(X)} \tag{3.2}$$

where min,max are the min-max values

3.4 Clustering approach

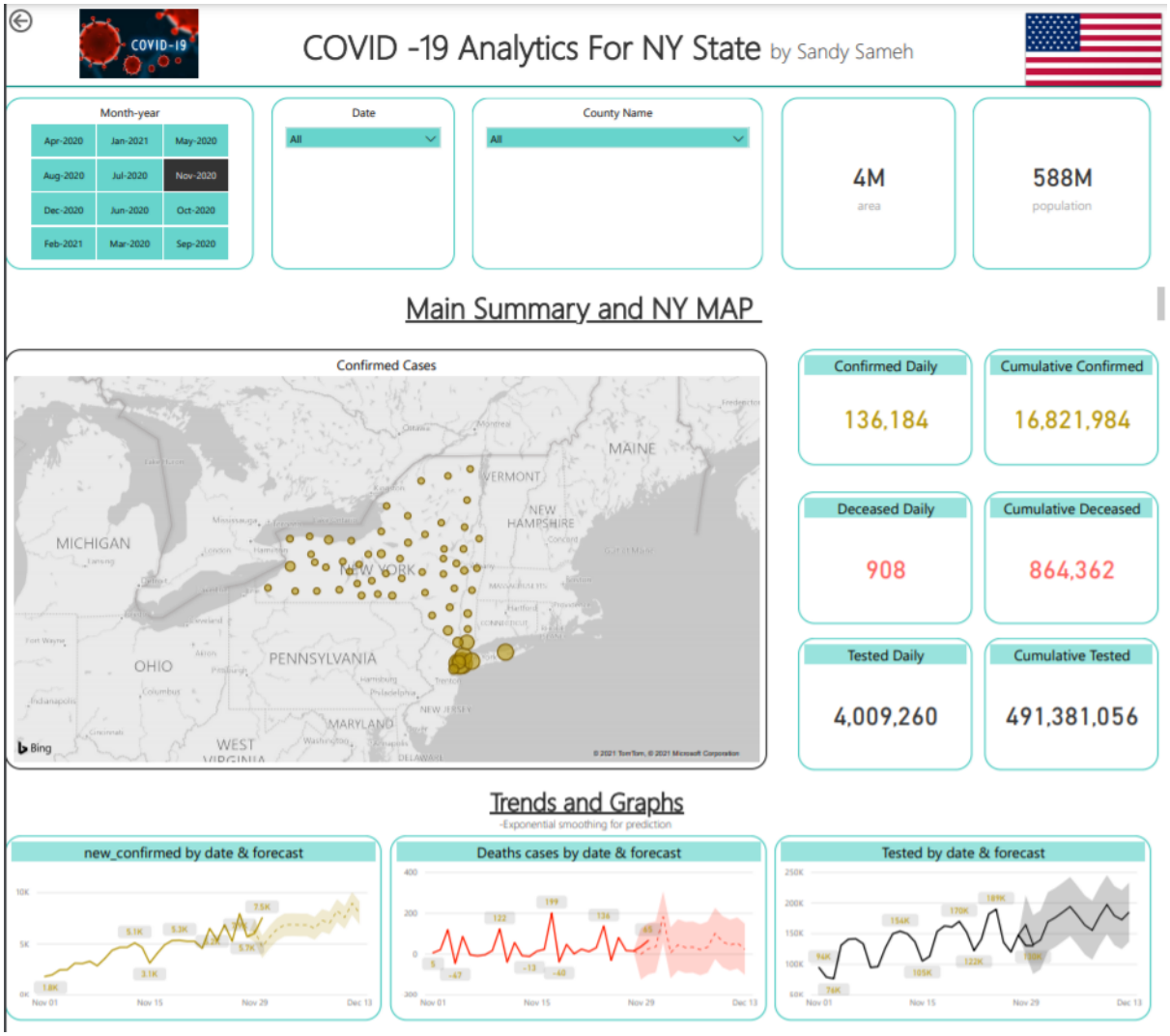


Figure 3.2: Sample of Data Visualization of NY state using PowerBI

3.4.1 Dimensionality Reduction using PCA

The features are then put through a dimensionality reduction process(PCA) in order to summarize useful data into the desired number of components. Then, Reducing the data using PCA with just two components to be able to visualize how the data is distributed. Afterwards, the number of components using Cluster Validation Indices mentioned below 3.6.1 are chosen that would help me achieve a 95% Cumulative variance explained.

3.4.2 K-Means

K-means was applied on Wave 1 of COVID-19 on 62 counties of New York state as well as on 54 states of the US. It was applied on both 3.8 and 3.9 and on 3.8 after being normalized by population density.

3.5 Prediction approach

3.5.1 Data Splitting

Each predictive model in the study has been conducted at different timestamps throughout wave 2 of COVID-19 in the United States. It was expected by intuition that as the number of days - where the total confirmed cases were being predicted - decrease the predictive accuracy should increase. The different time intervals are trying to predict the beginning of wave 2 from wave 1 (From March 1st to the end of July) The data is split and tested separately accordingly:

1. A training set (153 days;- 62.2% of my Data- of wave 1 from March till August) and A validation set (93 days; - 37.8% of my Data- of the first 3 months of wave 2).
2. A training set (214 days;- 87 % of my Data- of wave 1 from March till the beginning of wave 2) and A validation set (32 days; - 13% of my Data- to predict the confirmed cases in November).
3. A training set (231 days;- 94 % of my Data- of wave 1 from March till mid November) and A validation set (15 days; - 6% of my Data- to predict the confirmed cases in the last 15 days of November).
4. A training set (236 days;- 96 % of my Data- of wave 1 from March till end of November) and A validation set (10 days; - 4% of my Data- to predict the confirmed cases in the last 10 days of November).

Also to predict further into wave 2 the beginning of wave 2 (from the first of August to the first of January)was used to predict what will happen further into wave 2. The data is split and tested separately accordingly:

1. A training set (168 days;- 79% of my Data- of beginning of wave 2) and A validation set (45 days; - 21% of my Data- from mid January to the end of February).
2. A training set (182 days;- 86% of my Data- of beginning of wave 2) and A validation set (30 days; - 14% of my Data).
3. A training set (197 days;- 93% of my Data) and A validation set (15 days; - 7% of my Data).
4. A training set (202 days;- 4.7% of my Data) and A validation set (10 days; - 95.3% of my Data).

3.5.2 Models Applied

The learning models such as Support Vector Regression, Linear Regression are being used for Time Series Forecasting as well as applying Time Series Forecasting Models such as Holt's Linear Model, Holt-Winter's seasonal model, AR Model, MA Model, ARIMA, and SARIMA have been used in this study. These models have been trained on the days and newly confirmed cases.

3.6 Evaluation

3.6.1 Clustering approach: K-means Evaluation

The clusters have been evaluated using cluster validation indices such as Elbow Method (Intertia), Silhouettes, and Davis-Bouldin index(DBI) to gives an idea on what a good k number of clusters and reported in the results.

3.6.2 Prediction approach evaluation

The learning models have been evaluated based on important metrics such as MSE, RMSE, MAPE, and MAE and reported in the results.

Chapter 4

Results

This study attempts to achieve both spatial analysis evolution of COVID-19 around the 62 counties of New York State as well as the 54 states of the USA by using K-means to cluster counties/states based on similarity in their features and risks. Secondly, we attempt to achieve temporal analysis which is to develop models to predict future total confirmed cases of wave 2 in the USA from wave 1.

This chapter discussed both spatial and temporal results of the models applied. The first part of this chapter will discuss the spatial analysis where several trials were applied. The first trial was clustering the counties in New York state using the features in 3.8 of wave 1 (From March to the end of July) after applying PCA and standardizing my data using 'MinMaxScale' from 'sklearn.preprocessing'.

The Second trial was clustering the counties in New York state using features in 3.9 across wave 1 and applying PCA as well as MinMax Scale. The third trial in this study was clustering counties in New York State using features in 3.8 but after normalizing them by the population density of each county and applying the same techniques as mentioned in the above trials. In addition, those 3 trials were then repeated across the 54 states of the USA during waves 1.

The second part of this chapter discusses the temporal analysis of COVID-19 in the USA, regarding the total confirmed cases during wave 2. 8 Time Series models (Linear Regression, Support Vector Regression, Holt's Linear Model, Holt's Winter Model, AR Model, MA Model, ARIMA, and SARIMA) were trained during different times intervals to determine which performs better to predict the total confirmed cases and during which period. By intuition, it was expected that the more the days needed to be predicted decreases, the better the models performed. The different trials applied during different time intervals mentioned above in 3.5.1

In the part, I will divide my clustering results into 2 sections one for the counties of New York state and the other for The states of the USA.

4.1 Clustering NY State

The first section will cover 3 attempts of clustering the counties of New York state depending on the features mentioned above in both 3.3.1 during wave 1.

4.1.1 First Attempt: Original

In this attempt, the features mentioned in 3.8 were used with no normalization using population or population density. Where the period from March till the end of July was studied.

The data was scaled using $'MinMaxScale'$ from $'sklearn.preprocessing'$ ranging from 0 to 10 to standardize all the attributes of the dataset and give them equal weight so that redundant or noisy objects can be eliminated and there is valid and reliable data which enhances the accuracy of the result.

Cluster Validation Indecencies results

Three different clustering validation indices were applied; Elbow Method, DB Index, Silhouette Method. The results for the 3 Evaluation Methods suggested an optimal number of k=2 (Refer to figure 4.1) before applying any dimensionality reduction techniques (PCA).

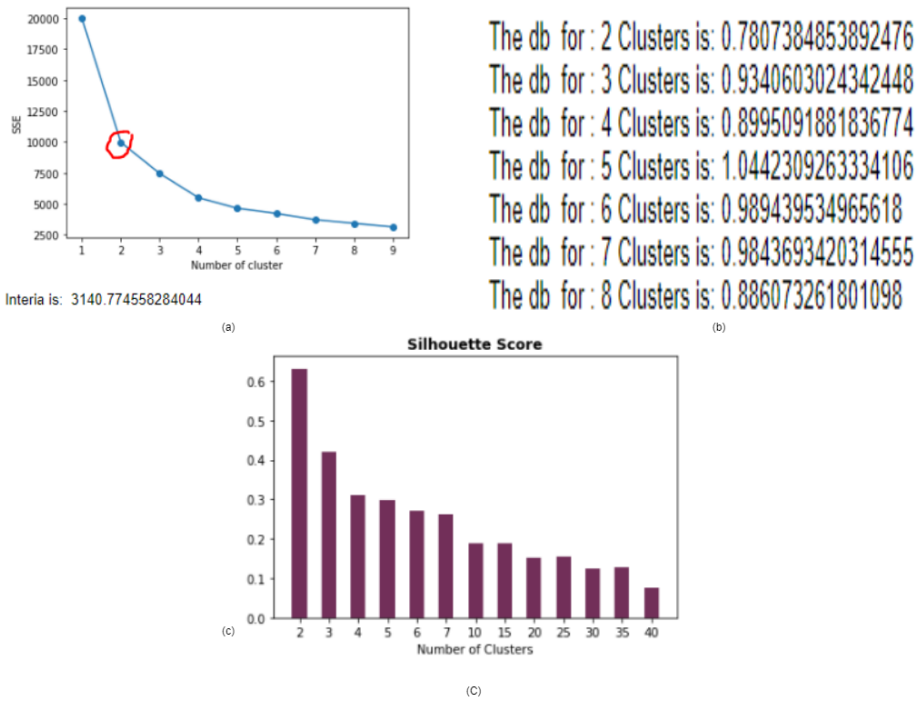


Figure 4.1: NY attempt1.(a)Elbow Method (b)DBI results (c)Silhouette. All suggesting k=2 as an optimal number of k

Dimensionality reduction could help k-means perform better as K-means is extremely sensitive to scale. This is where PCA comes in, since it whitens the data. By using PCA, global correlation will be removed giving better results. At first PCA with n components =2 was applied to visualize it on a 2D diagram as shown in fig 4.2.10 components were used in PCA to achieve 95% explained variance. After applying the PCA all the 3 evaluation methods were repeated to give better actual results (Refer to figure 4.3)

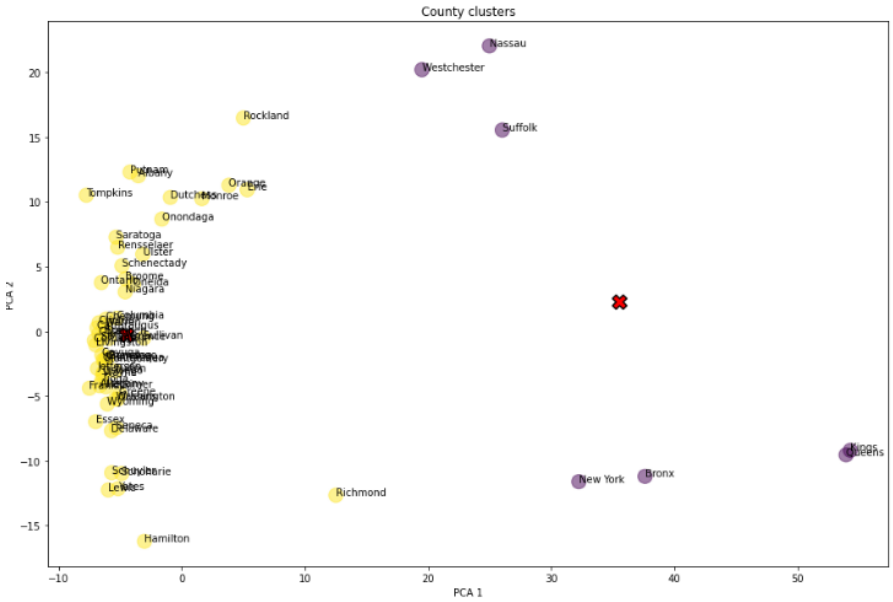


Figure 4.2: PCA of NY first attempt with n components =2 after applying k-means with k =2

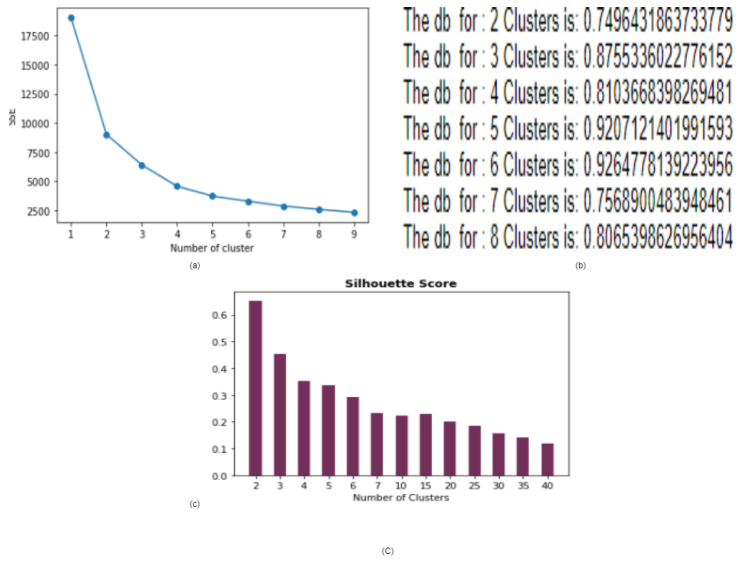


Figure 4.3: NY first attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA all suggesting k=2 as an optimal number of k

Clustering Results

Cluster 0 Counties: Bronx County, Kings County, Nassau County, New York County, Queens, Suffolk County, and Westchester County.

Feature	Min	Mean	MAX
Total_confirmed	26757.0	45314.714286	65370.00
Total_Tested	267344.00	388370.857143	545272.00
Total_deceased	1424.0	3300.714286	5814.00
.....	.....	.....	.....
.....	.....	.....	.....
Area	87.00	1365.857143	6146.00
Population	969,689	1,681,735	2,594,676

Table 4.1: Table related to NY's trial 1, containing the minimum , mean and maximum of some of the features that are related to counties of cluster 0

Cluster 1 Counties: Albany County, Allegany County, Broome County, Cattaraugus County, Cayuga County, Chautauqua County, Chemung County, Chenango County, Clinton County, Columbia County, Cortland County, Delaware County, Dutchess County, Erie County, Essex County, Franklin County, Fulton County, Genesee County, Greene County, Hamilton County, Herkimer County, Jefferson County, Lewis County, Livingston County, Madison County, Monroe County, Montgomery County, Niagara County, Oneida County, Onondaga County, Ontario County, Orange County, Orleans County, Oswego County, Otsego County, Putnam County, Rensselaer County, Richmond County, Rockland County, St. Lawrence County, Saratoga County, Schenectady County, Schoharie County, Schuyler County, Seneca County, Steuben County, Sullivan County, Tioga County, Tompkins County, Ulster County, Warren County, Washington County, Wayne County, Wyoming County, and Yates County.

Feature	Min	Mean	MAX
Total_confirmed	6.00	1347.927273	1,4062.000
Total_Tested	671.00	25249.127273	151,745.00
Total_deceased	0.0	71.872727	868.00
.....	.....	.....	.....
.....	.....	.....	.....
Area	265.00	2391.563636	7308.00
Population	4471	142,135	919,034

Table 4.2: Table related to NY's trial 1 , containing the minimum , mean and maximum of some of the features that are related to counties of cluster 0

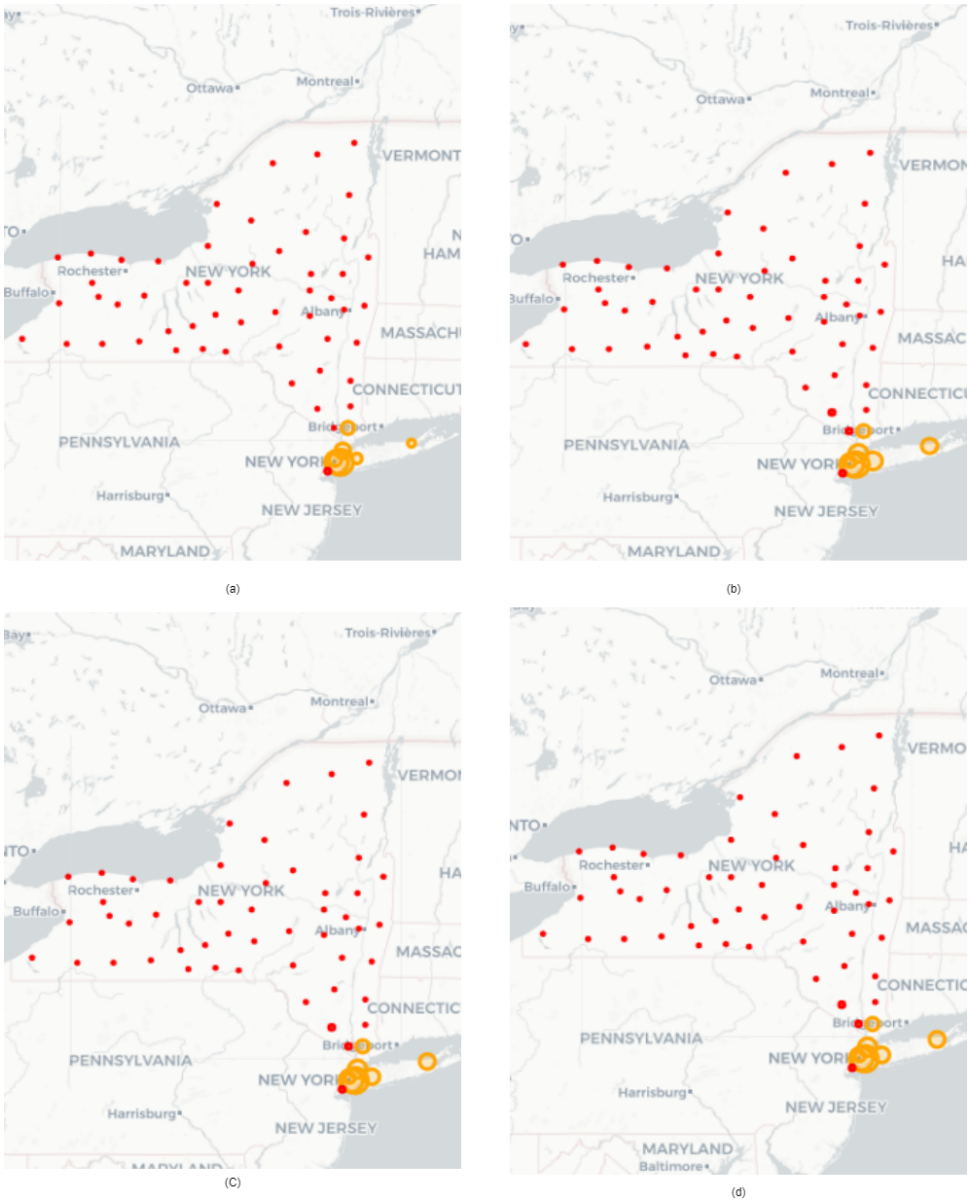


Figure 4.4: NY's first trial where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the first attempt. The orange closer are for cluster 0, and the red color is for cluster 1. The radius of the circle represents the total confirmed cases in each cluster in order to give a better visualization of each cluster

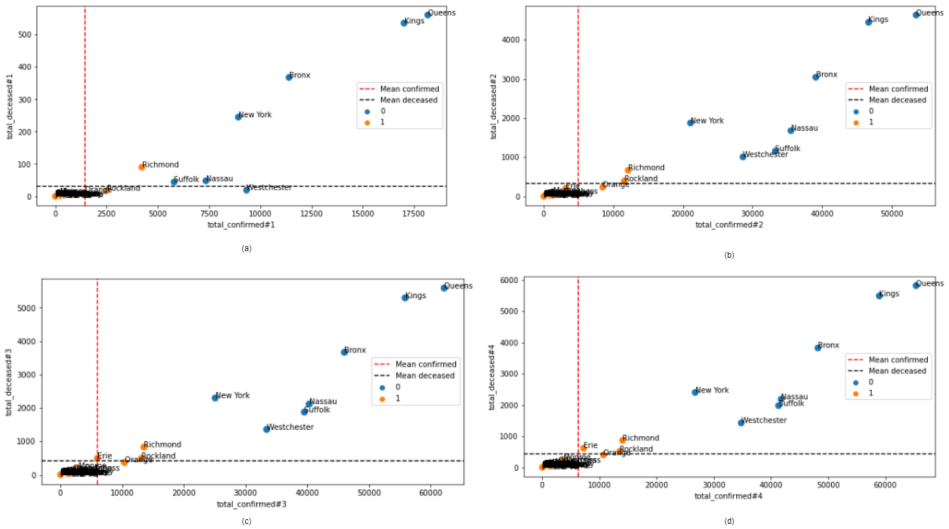


Figure 4.5: NY's trial 1 where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased

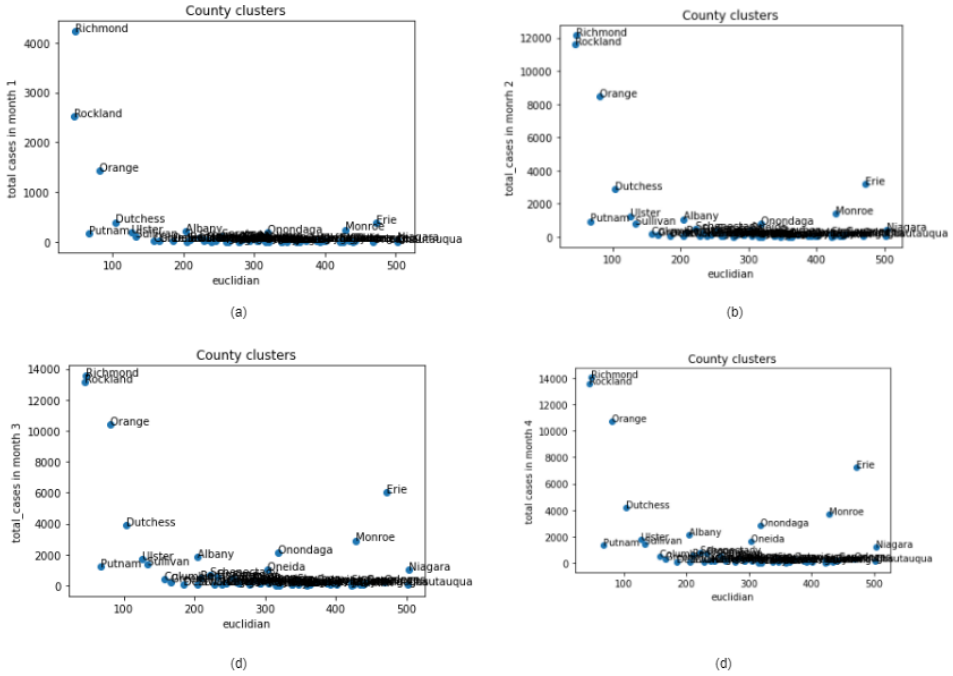


Figure 4.6: NY's first trial showing the distance between the geographic location of the center of cluster 0 and all counties of cluster 1 on the x-axis, and the total number of cases on the y-axis along the 4 months of wave 1 where (a) is this first month , and (d) is forth month.It shows that the closer the distance from center of cluster 0, the greater the number of the total cases

Interpretation

In this trial , As shown in map 4.4 along 4 months of wave 1; counties of cluster 0 had a larger number of total confirmed cases which is also shown in tables 4.1 and 4.2.

Through the figures 4.5 , It was shown how counties of cluster 0 are related with reference to both total confirmed cases and total deaths. It also shows how *Richmond county*, *Rockland land*, and *Orange county* are some how trying to approach counties of clusters 0. In order to better understand why ; in figure 4.6 the geographic distance in km square was measured from the center of cluster 0. It showed that the closer the counties were to the center the higher the total confirmed cases were; as covid is transmitted through close interactions.

To sum up this trial, **Cluster 0 :** is suggested to have a higher risk of COVID-19 as it has higher total confirmed cases as well as the counties of this cluster 0 had smaller area than counties in cluster 1, However counties of cluster 0 had a higher population number than counties in cluster 1 -(*Cluster 0 has higher population density , which means that the population is high relative to the size of county*)- which suggests that a large number of people in a smaller area are at higher risk to spread the disease. Counties of cluster 0 had higher population with ages ranging from 40-70 than counties in cluster 1 as shown in fig 4.7; these age groups have a higher risk of COVID-19 infection , and deaths according to source. In addition, Cluster 0 had a higher mobility to workplaces and retail and recreations area than Cluster 1 especially in the first 2 months of wave 1; which maybe have increases changes of spread of the disease due to interactions or other reasons. There is no major difference between both cluster in both Mobility Residential and Mobility of Grocery and Pharmacy.

Problem with this attempt, is that it doesn't fully give a true indication of the total number of confirmed cases with reference to population. To further explain, a county with population = 10 and number of total confirmed cases is equal to = 10; should be considered of a high risk as 100% of population is infected. However a county with population =100 and number of the total confirmed cases =10; is at a better state as only 10% of population is infected. In this attempt, A false indication is given as only the number of total confirmed cases is compared. That's why in the second trial, The features should be normalized against the population.

	Cluster 0	Cluster 1
Mean Population Density	6563.14857	1795
Mean Area	1365.85714	2391.563636
Mean Population	1,681,735	142,135
Mean Pop. Age 0-9	201,155	15,876
Mean pop. Age 10-19	190,935	18,077
Mean pop. Age 20-29	248,957	19,558
Mean pop. Age 30-39	245,121	16,902
Mean pop. Age 40-49	216,191	16,750
Mean pop. Age 50-59	225,250	20,720
Mean pop. Age 60-69	179,259	17,638
Mean pop. Age 70-79	105,397	10,226
Mean pop. Age 80++	69,468	6,385
Mobillity Retail & Reacreation #1	-21.57	-38.8
Mobillity Retail & Reacreation #2	-24.14	-27.98
Mobillity Retail & Reacreation #3	-18	-22.45
Mobillity Retail & Reacreation #4	-7	-3.2
Mobility Workplace #1	-25.87	-43.618
Mobility Workplace #2	-24.85	-42.95
Mobility Workplace #3	-7.51	-10.29
Mobility Workplace #4	-17.42	-30
Mobility Grocery & pharmacy #1	-9.42	-14.2
Mobility Grocery & pharmacy #2	-8.714	-4.109
Mobility Grocery & pharmacy #3	-2	6.6
Mobility Grocery & pharmacy #4	-0.57	11.4
Mobility Residential #1	11.8714	15.87
Mobility Residential #2	11.714	14.76
Mobility Residential #3	4.148	2.8909
Mobility Residential #4	5.7	6.818

Figure 4.7: NY attempt 1's mean Values of some features between both cluster 0 and 1. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk.

4.1.2 Second Attempt: Normalization on population

In this attempt, the features mentioned in 3.9 were used which are normalized using population to give a better sense of the data.The data was scaled using $'MinMaxScale'$.

Cluster Validation Indecencies Results

Then, Elbow Method, DB Index, Silhouette Method were applied before applying any Dimensionality Reduction techniques (PCA). Before applying PCA, DB Index suggested 3 cluster, Silhouette Method suggested 2 cluster, and Elbow Method is leaning more towards k=2 and k=3 as shown in fig 4.8.

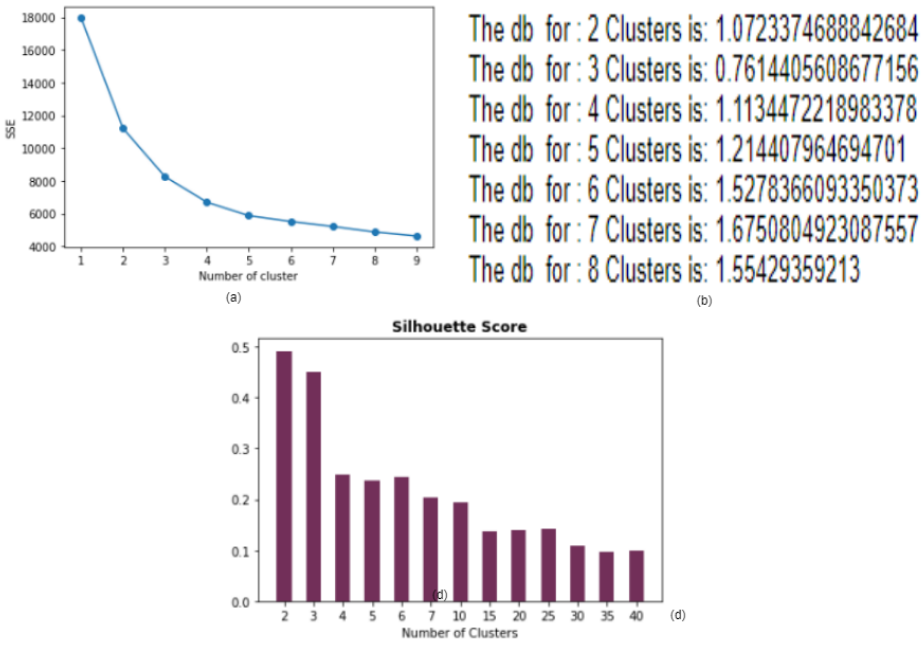


Figure 4.8: NY's second attempt where (a)Elbow Method (b)DBI results (c)Silhouettes before applying PCA

At first PCA with n components =2 was applied to visulize it on a 2D diagram as shown in fig 4.10, and 4.11. Then, 16 components were used in PCA to achieve 95% explained variance.

Afterwards all 3 Evaluation Methods were applied as shown in fig 4.3 also after applying PCA; where here DB Index suggested 3 cluster, Silhouette Method suggested 2 cluster, and Elbow Method is unclear between k=2 or k=3. As a result , both studies using k=2 and k=3 were applied.

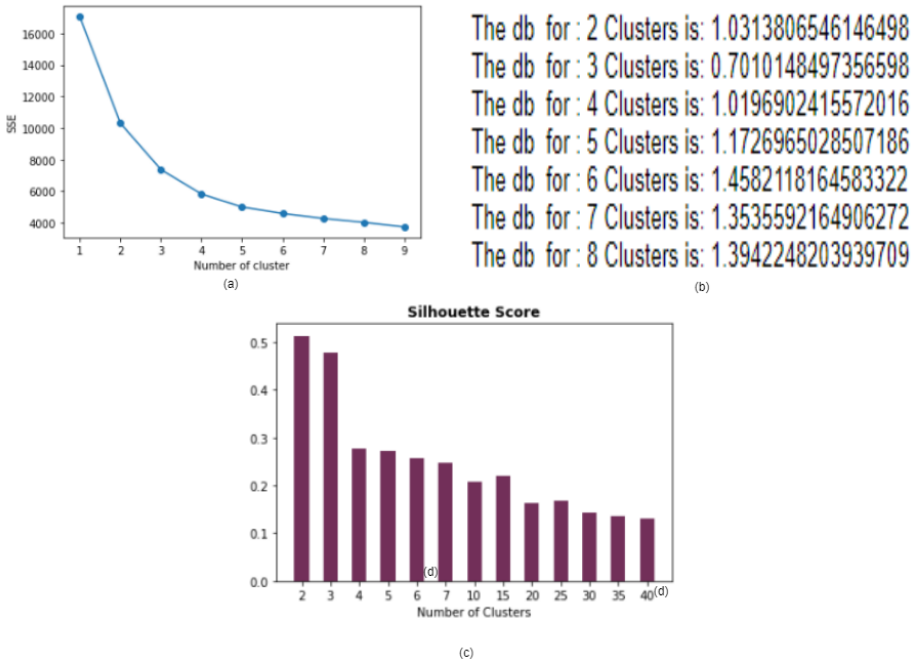


Figure 4.9: NY's second attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA

This part will be divided into two attempts : firstly, clustering using k-means with an optimal k=2, and secondly clustering using k-means with an optimal k=3.

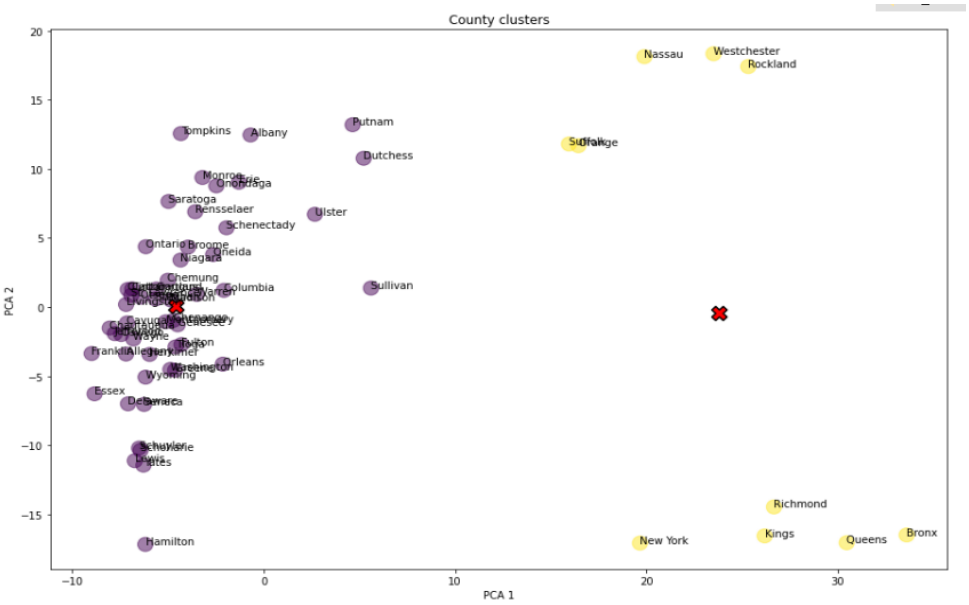


Figure 4.10: NY's second attempt applying PCA with n components =2 after applying kmeans with k =2

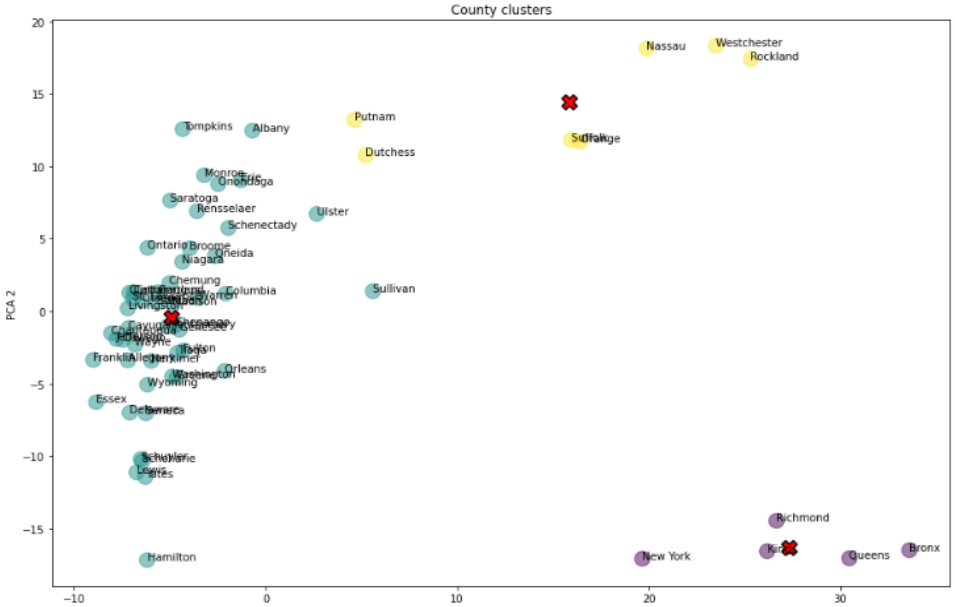


Figure 4.11: NY's second attempt applying PCA with n components =2 after applying kmeans with k =3

Clustering Results: K-means where K=2

Cluster 0 Counties: Bronx County, Kings County, Nassau County, New York County,Orange County, Queens County, Richmond County, Rockland County,Suffolk County, and Westchester County.

Feature	Min	Mean	MAX
Total Confirmed percent	1.640800	2.952170 6	4.183300
Total Tested percent	20.564700	23.538290	27.570100
Total deceased percent	0.106100	0.176570	0.265300
.....	.....	.....	.....
.....	.....	.....	.....
Area	87.0	1251.400	6146.00
Population Density	175.00	4854.10	18744.00

Table 4.3: Table related to trial 2 when k=2 , containing the minimum , mean and maximum of some of the features that are related to counties of cluster 0

Cluster 1 Counties: Albany, Allegany, Broome, Cattaraugus, Cayuga, Chautauqua, Chemung, Chenango, Clinton, Columbia, Cortland, Delaware, Dutchess, Erie, Essex, Franklin, Fulton, Genesee, Greene, Hamilton, Herkimer, Jefferson, Lewis, Livingston, Madison, Monroe, Montgomery, Niagara, Oneida, Onondaga, Ontario, Orleans, Oswego, Otsego, Putnam, Rensselaer,St. Lawrence, Saratoga, Schenectady, Schoharie, Schuyler, Seneca, Steuben, Sullivan, Tioga, Tompkins, Ulster, Warren, Washington, Wayne, Wyoming, and Yates.

Feature	Min	Mean	MAX
Total Confirmed percent	0.059400	0.392162	1.934800
Total Tested percent	9.680100	16.109152	23.095700
Total deceased percent	0.000000	0.023083	0.127500
.....	.....	.....	.....
.....	.....	.....	.....
Area	543.00	2472.750000	7308.0000
Population Density	1.00	61.326923	289.00

Table 4.4: Table related to trial 2 when k=2 , containing the minimum , mean and maximum of some of the features that are related to counties of cluster 1

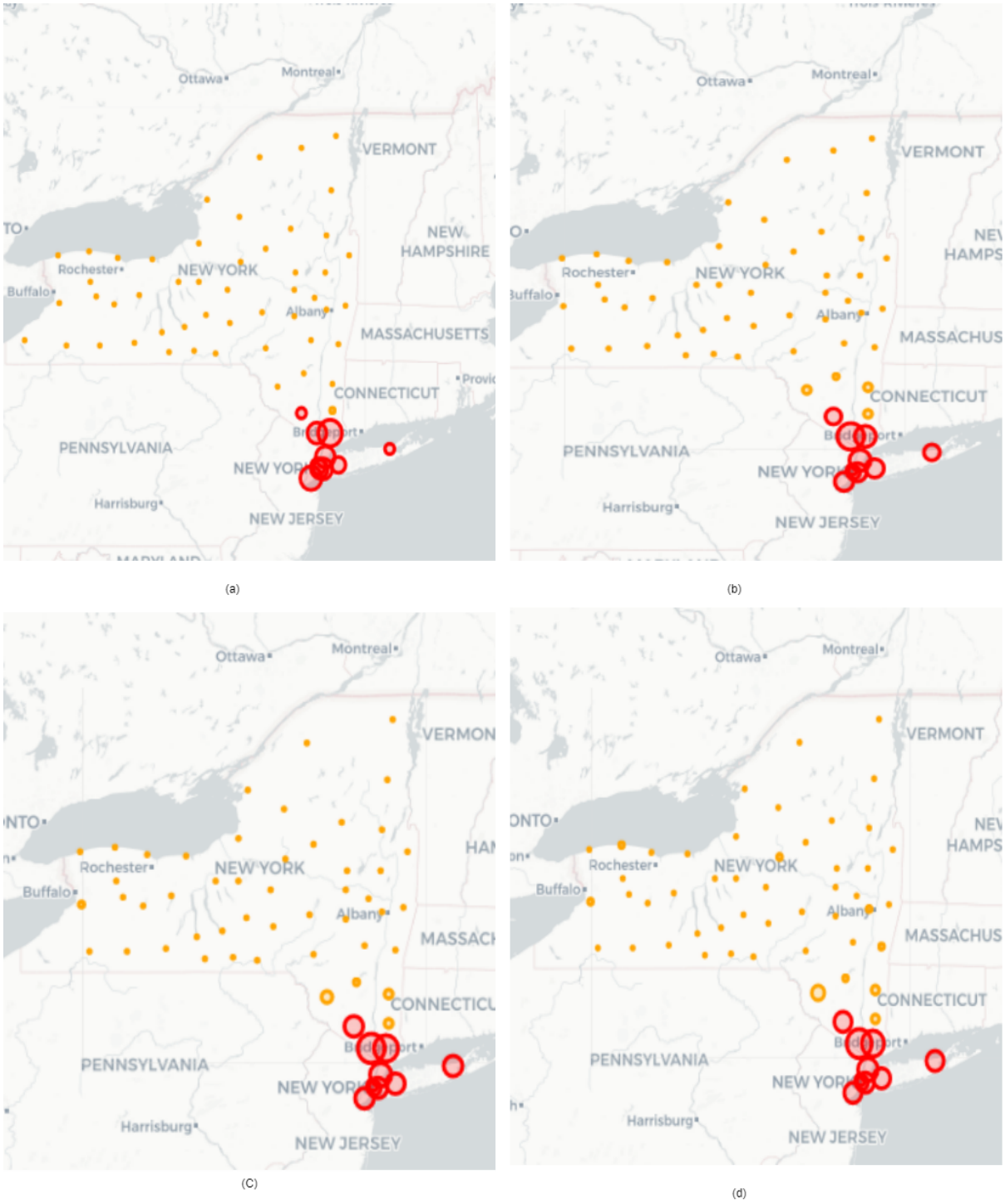


Figure 4.12: NY's second attempt with (k=2) where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the second attempt. The Orange closer are for cluster 0, and the red color is for cluster 1. The radius of the circle represents the total confirmed cases percentage in each cluster in order to give a better visualization of each cluster

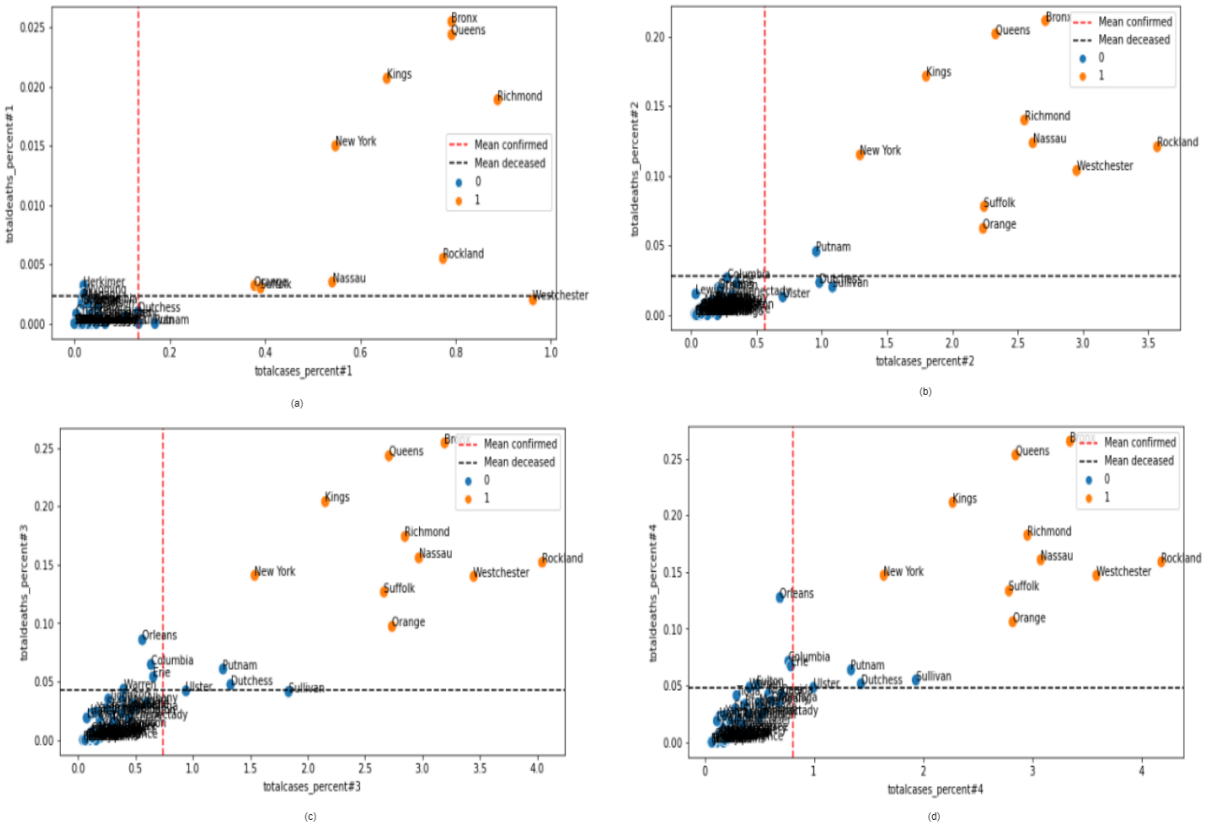


Figure 4.13: NY's second attempt with (k=2) where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased

Clustering Results: K-means where K=3

Cluster 0 Counties: Bronx, Kings, New York, Queens, and Richmond.

Feature	Min	Mean	MAX
Total Confirmed percent	1.640800	2.612580 6	3.346700
Total Tested percent	21.015000	23.065460	24.454100
Total deceased percent	0.147100	0.211900	0.265300
.....	.....	.....	.....
Area	87.00	242.40000	461.0000
Population Density	1795.00	9118.000	18,744.00

Table 4.5: NY trial 2 when k=3 , containing the minimum , mean and maximum of some of the features that are related to counties of cluster 0

Cluster 1 Counties: Albany, Allegany, Broome, Cattaraugus, Cayuga, Chautauqua, Chemung, Chenango, Clinton, Columbia, Cortland, Delaware, Erie, Essex, Franklin, Fulton, Genesee, Greene, Hamilton, Herkimer, Jefferson, Lewis, Livingston, Madison, Monroe, Montgomery, Niagara, Oneida, Onondaga, Ontario, Orleans, Oswego, Otsego, Rensselaer,St. Lawrence, Saratoga, Schenectady, Schoharie, Schuyler, Seneca, Steuben, Sullivan, Tioga, Tompkins, Ulster, Warren, Washington, Wayne, Wyoming, and Yates.

Feature	Min	Mean	MAX
Total Confirmed percent	0.059400	0.352420	1.934800
Total Tested percent	9.680100	15.944986	23.095700
Total deceased percent	0.000000	0.021704	0.127500
.....	.....	.....	.....
Area	543.00	2516.140000	7308.0000
Population Density	1.00	57.940000	289.00

Table 4.6: NY state trial 2 when k=3 , containing the minimum , mean and maximum of some of the features that are related to counties of cluster 1

Cluster 2 Counties: Dutchess, Nassau, Orange, Putnam, Suffolk, and Westchester.

Feature	Min	Mean	MAX
Total Confirmed percent	1.340300	2.747171	4.183300
Total Tested percent	18.482400	15.944986	27.570100
Total deceased percent	0.051400	0.117329	0.160800
.....	.....	.....	.....
Area	516.00	2011.142857	6146.000000
Population Density	137.00	463.285714	1157.0000

Table 4.7: NY state trial 2 when k=3 , containing the minimum , mean and maximum of some of the features that are related to counties of cluster 1

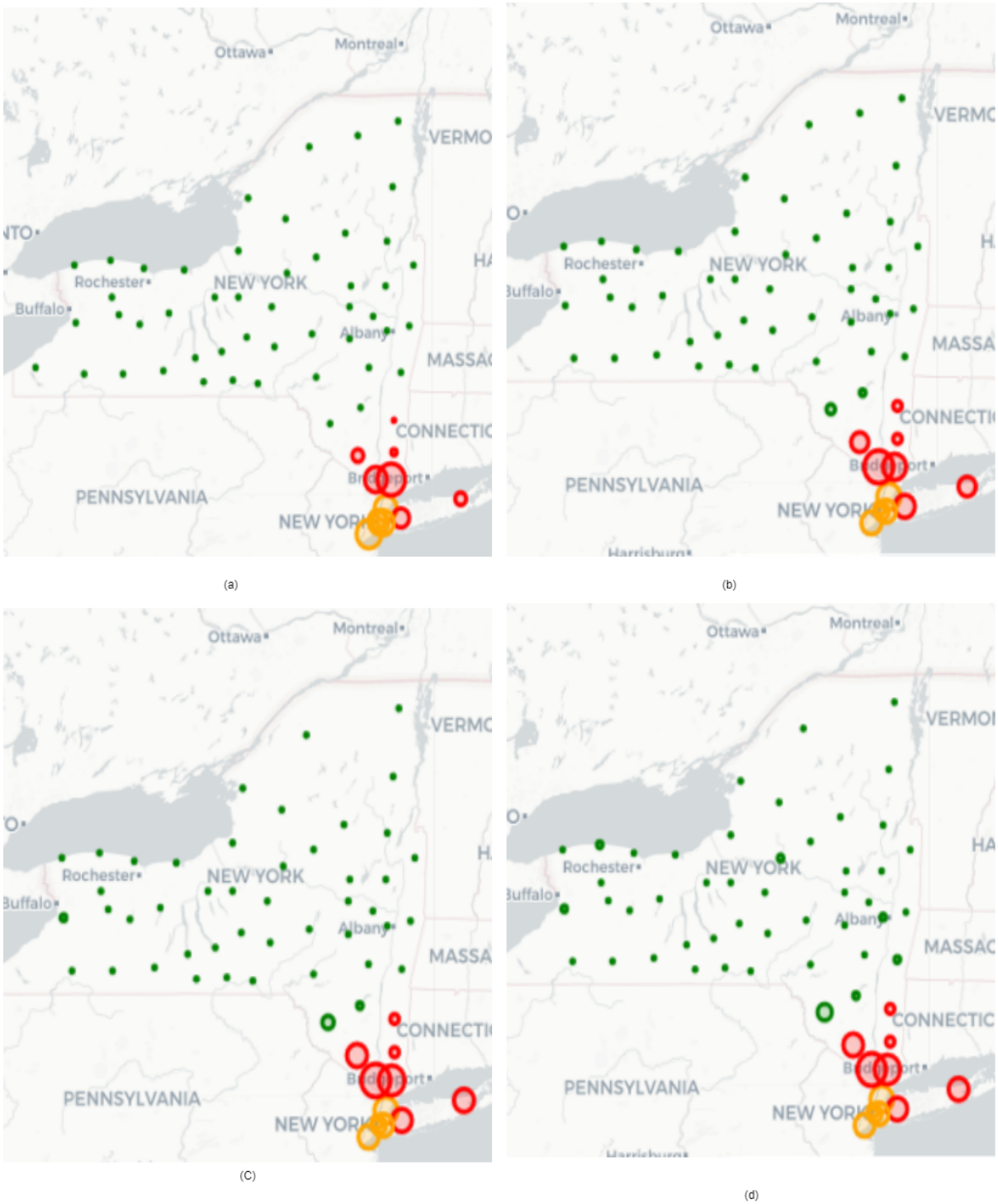


Figure 4.14: NY state second trial(k=3) where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the second attempt(k=3). The Orange closer are for cluster 0, and the green color is for cluster 1 and red is for cluster 2. The radius of the circle represents the total confirmed cases percent

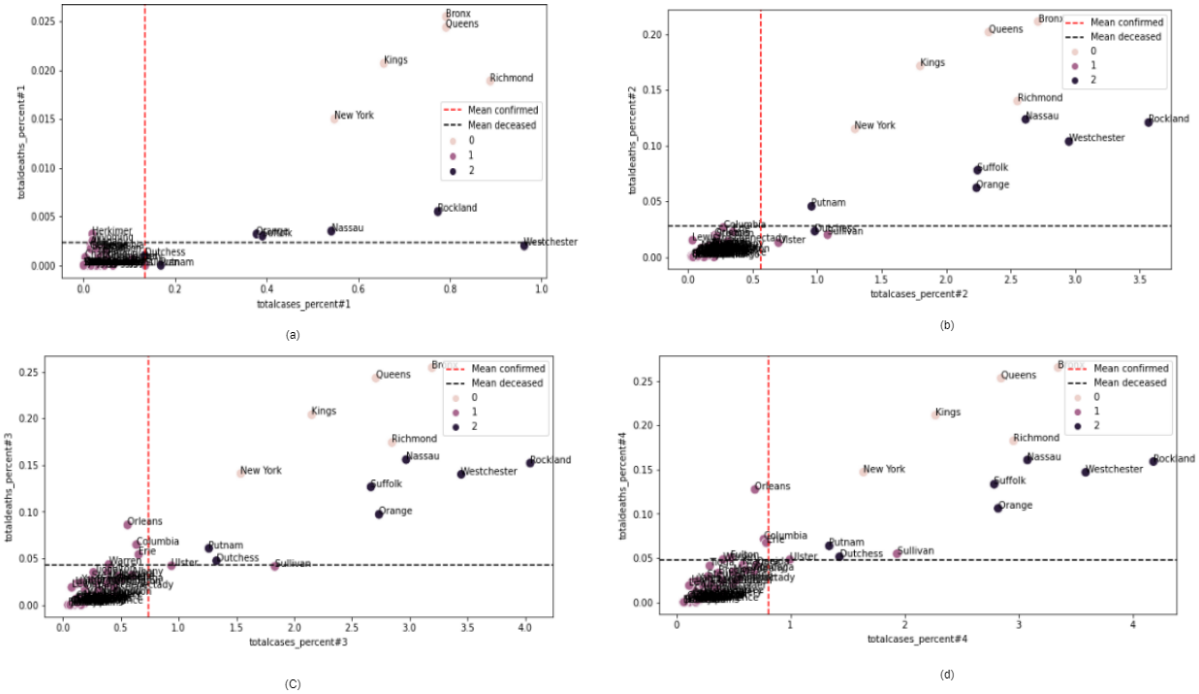


Figure 4.15: NY state trial 2 (k=3) where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased

Interpretation: where k =2

In this trial , As shown in map 4.12 along the 4 months of wave 1; counties of cluster 0 had a larger percentage of total confirmed cases, total tested, and total deceased which is also shown in tables 4.3 and 4.4.

Through the figures 4.13 , It was shown how counties of cluster 0 are related with reference to both percentage of total confirmed cases and total deaths. It also shows how *Putnam county*, and *Dutchess land* are some how trying to approach counties of clusters 0 with reference to both percentages.

To sum up this trial, **Cluster 0 :** is suggested to have a higher risk of COVID-19 as counties in cluster 0 has a higher population density than counties in cluster 1.

Having a higher population density means that the population of counties in cluster 0 is high relative to the area size of the counties - which suggests that a large number of people in a smaller area are at higher risk to spread the disease.

Counties of cluster 0 had a higher percentage of mobility to workplaces, retails and recreations through out the 4 months of wave 1 than counties of cluster 0, Mobility to pharmacies and grocery were less in counties of cluster 0 which could help us observe that those counties sought less medical supplies as shown in 4.16.

These observations along with the higher percentage of total confirmed cases, deaths, and tested could help identify that counties of cluster 0 were subject to higher risk of covid during wave 1.

The percentage of the population of different ages, and mobility residential in both clusters weren't significantly different.

	Cluster 0	Cluster 1
Mean Population Density	4854.1	61.326
Mean Area	1251.4	2472.75
Mean Pop. Age 0-9 %	12.35959	10.490338
Mean pop. Age 10-19%	12.38482	12.330308
Mean pop. Age 20-29%	14.08774	13.265848
Mean pop. Age 30-39%	13.43604	11.40525
Mean pop. Age 40-49%	12.74403	11.630644
Mean pop. Age 50-59%	13.73501	14.976456
Mean pop. Age 60-69%	10.79069	13.360498
Mean pop. Age 70-79%	6.34313	7.858796
Mean pop. Age 80++%	4.11895	4.681862
Mobillity Retail & Reacreation #1	-24.8	-39.211538
Mobillity Retail & Reacreation #2	-27.5	-27.557692
Mobillity Retail & Reacreation #3	-21.6	-22.076923
Mobillity Retail & Reacreation #4	-8.3	-2.730769
Mobility Workplace #1	-29	-44.038462
Mobility Workplace #2	-27.7	-43.346154
Mobility Workplace #3	-8.8	-10.211538
Mobility Workplace #4	-18.9	-30.461538
Mobility Grocery & pharmacy #1	-11.1	-14.153846
Mobility Grocery & pharmacy #2	-9.8	-3.634615
Mobility Grocery & pharmacy #3	-2.6	7.211538
Mobility Grocery & pharmacy #4	-0.8	12.134615
Mobility Residential #1	13.5	15.788462
Mobility Residential #2	13.2	14.653846
Mobility Residential #3	4.6	2.730769
Mobility Residential #4	6.3	6.769231

Figure 4.16: NY's second attempt with k=2 showing the mean Values of some features, between both cluster 0 and 1. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk.

Interpretation: where k =3

In this trial , As shown in map 4.14 along the 4 months of wave 1; counties of cluster 0 had a larger percentage of total total tested, and total deceased than . Counties of cluster 0 and cluster 2 had relatively equal percentage of total confirmed cases which is also shown in tables 4.5, 4.6, and 4.7

Through the figures 4.15 , It was shown how counties of cluster 0 and counties of cluster 2 are related with reference to both percentage of total confirmed cases and total deaths.

To sum up this trial with ref to 4.17,

Cluster 0 : is suggested to have a high risk of COVID-19 as counties in cluster 0 has a higher population density than all other clusters .Having a higher population density means that the population of counties in cluster 0 is high relative to the area size of the counties - which suggests that a large number of people in a smaller area are at higher risk to spread the disease.

Also Counties in cluster 0 have the highest total test as well as deceased percentage, along with highest Mobility to workplaces which could reflect on why they might have a relatively large number of total cases percentage. Counties in this clusters have the highest population of age group of 20-50 which have a high percentage of being infected, hospitialized, and deceased according to this source.

Cluster 1: is suggested to have the lowest risk of COVID-19 among the other 2 clusters as they have a very low population density, which suggests that the population isn't to high for the area of county, so less contact. Also, counties of cluster 1 recorded the least percentage of total cases, deaths and tests during wave 1.

The Mobility of workplace in this cluster remained low and less during wave 1. Even tho the percentage of population age from 60-69 was the highest in those counties, not so many deaths cases were reorded.

Cluster 2: is suggested to have also a high risk of COVID-19 but through analysis it is assumed to have taken the corrective actions to reduce this risk than counties of cluster 0. Counties of cluster 2 had relatively high percentage of total cases than cluster 0, however cluster 2 had a fewer percentage of total deaths, and tests recorded.

Counties of this cluster have a lower population density than counties of cluster 0 which suggests that population isn't relatively high for the area of the counties, this may indicate a reason behind the softer risk of COVID-19.

It was also noticed that counties in this cluster had relatively intermediate mobility to workplaces during wave 1 which is why maybe the percentage of total cases in this cluster was higher, but it was lower than those of cluster 0 indicating that maybe these clusters required a more strict workplaces closing to deal with the increase in the total confirmed cases percentage.

	Cluster 0	Cluster 1	Cluster 2
Mean Population Density	9118	57.94	463.285714
Mean Area	242.4	2516.14	2011.14286
Mean Pop. Age 0-9 %	12.08022	10.525738	11.772214
Mean pop. Age 10-19%	10.99662	12.30981	13.507229
Mean pop. Age 20-29%	15.60854	13.292368	12.5772
Mean pop. Age 30-39%	15.3144	11.41789	11.423843
Mean pop. Age 40-49%	12.76442	11.570164	12.5772
Mean pop. Age 50-59%	12.8493	14.905376	15.230071
Mean pop. Age 60-69%	10.43	13.393292	11.548314
Mean pop. Age 70-79%	6.10954	7.882936	6.7706
Mean pop. Age 80++%	3.84696	4.702422	4.3272
Mobillity Retail & Reacreation Mean	-24.203716	-24.78966	-20
Mobility Workplace Mean	-30.386563	-43.507731	-37.857143
Mobility Grocery & pharmacy Mean	2.319875	1.00575	-2.714286
Mobility Residential #1 Mean	12.19832	12.7419	12.428571

Figure 4.17: NY's second attempt with k=3. Figure shows the mean values of the features between both cluster 0, 1, 2. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk.

4.1.3 Third Attempt: Normalization on Density

4.2 Clustering USA

The second section will cover the 3 attempts of clustering the states of the USA depending on the features mentioned above in both 3.3.1 during wave 1 .

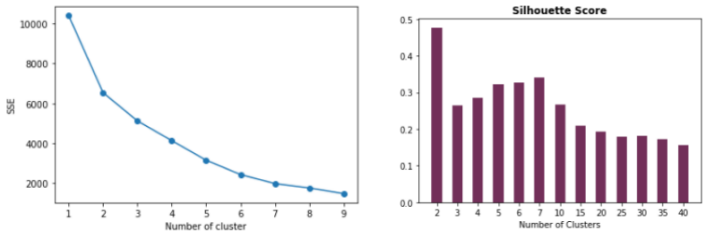
4.2.1 First Attempt: Original

In this attempt, the features mentioned in 3.8 were used with no normalization using population or population density. The data was scaled using $'MinMaxScale'$.

Cluster Validation Indecencies

Elbow Method, DB Index, Silhouette Method were applied before applying any Dimensionality Reduction techniques (PCA). Both Silhouette Method and Elbow Suggested k=2 However DB index suggested k=5. PCA was then applied with n components =2 was applied to visualize it on a 2D diagram .Then PCA with varience of 95% was applied.

After appyling PCA, Still both Silhouette and Elbow method suggested k=2 as seen in 4.18 so kmeans with 2 clusters was applied.



(a) (b)

The db for : 2 Clusters is: 1.0622746455987253
The db for : 3 Clusters is: 1.208425200472979
The db for : 4 Clusters is: 1.0045492527439048
The db for : 5 Clusters is: 0.8069837361588249
The db for : 6 Clusters is: 0.813583937582786

(c)

Figure 4.18: USA's first attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA

Clustering Results

Cluster 0 states: California, Florida, Illinois, Massachusetts, Michigan, New Jersey, New York, Pennsylvania, and Texas.

Feature	Min	Mean	MAX
Total_confirmed	70,223.000000	166,843.777778	392,930.000000
Total_Tested	912,000.00	2,049,750.000	4,061,692.000
Total_deceased	3,160.0000	8,922.111111	24,842.000000
.....	.....	.....	.....
Area	22,591.000000	222,390.666667	696,241.000000
Population	6,794,422	17,557,610	39,144,820
Population Density	39.00	141.00	397.0

Table 4.8: Table related to USA first attempt , containing the minimum , mean and maximum of some of the features that are related to states of cluster 0

Cluster 1 Counties: Alaska, Alabama, Arkansas, Arizona, Colorado, Connecticut, District of Columbia, Delaware, Georgia, Guam, Hawaii, Iowa, Idaho, Indiana, Kansas, Kentucky, Louisiana, Maryland, Maine, Minnesota, Missouri, Northern Mariana Islands, Mississippi, Montana, North Carolina, North Dakota, Nebraska, New Hampshire, New Mexico, Nevada, Ohio, Oklahoma, Oregon, Puerto Rico, South Carolina, South Dakota, Tennessee, Utah, Virginia, Virgin Islands, Vermont, Washington, Wisconsin, West Virginia, and Wyoming

Feature	Min	Mean	MAX
Total_confirmed	30.000000	23,773.111111	79,417.000000
Total_Tested	2,827.000000	347,546.333333	860,664.000000
Total_deceased	2.000	869.288889	4,320.00
.....	.....	.....	.....
Area	62,755.00	174,135.7000	1,717,856
Population	3.677432e+06	5.514400e+04	1.161342e+07
Population Density	0.00	150.733	3,987.00

Table 4.9: Table related to USA first attempt , containing the minimum , mean and maximum of some of the features that are related to states of cluster 1

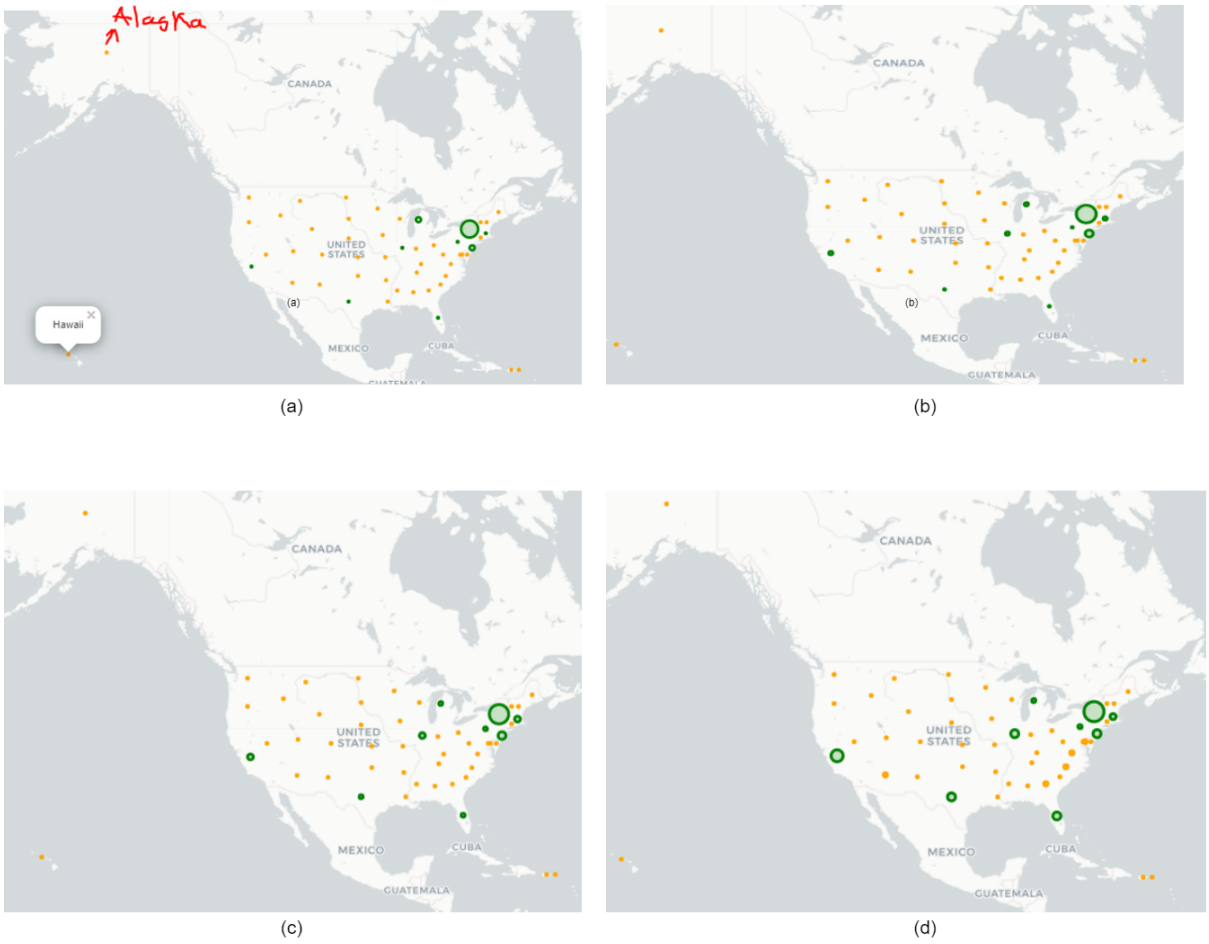


Figure 4.19: USA's First trial where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the first attempt. The orange closer are for cluster 1, and the green color is for cluster 0. The radius of the circle represents the total confirmed cases in each cluster in order to give a better visualization of each cluster

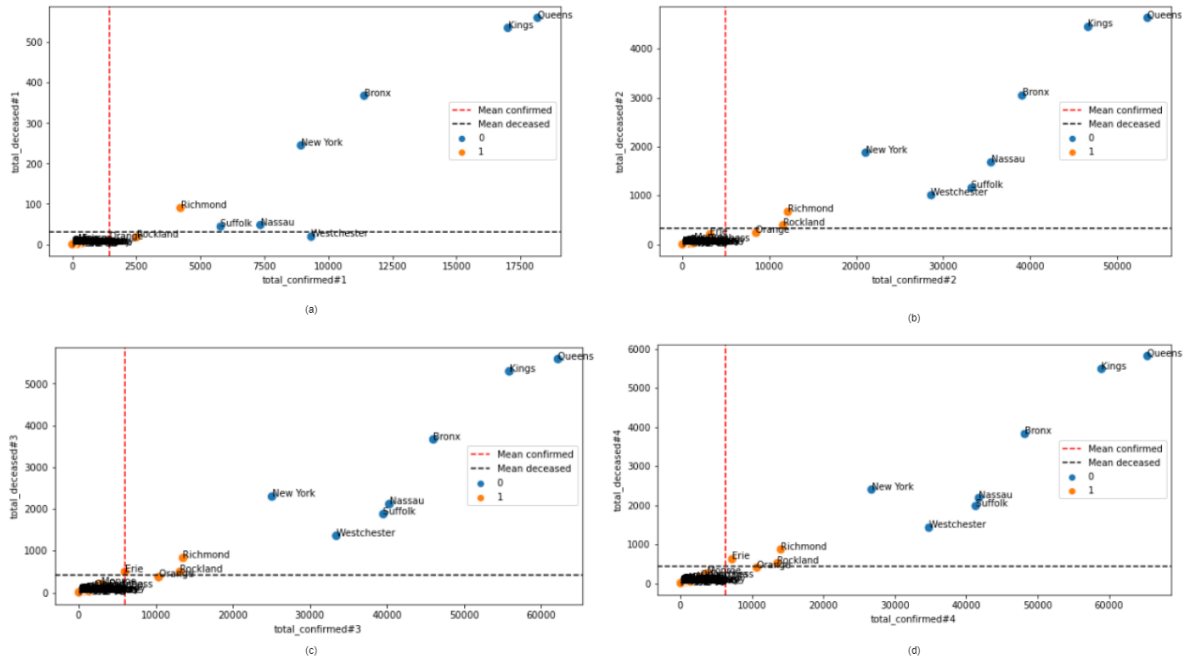


Figure 4.20: USA's trial 1 where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased

Interpretation

In this trial , As shown in map 4.19 along 4 months of wave 1; states of cluster 0 had a larger number of total confirmed cases which is also shown in tables 4.8 and 4.9. Through the figures 4.20 , It was shown how counties of cluster 0 are related with reference to both total confirmed cases and total deaths.

To sum up this trial, **Cluster 0 :** is suggested to have a higher risk of COVID-19 as it has higher number of total confirmed cases, total tested as well as total deceased. Cluster 0 states have a higher population than states of cluster 1. Despite, the population density being higher slightly in states of cluster 1. However, not all the population is well distributed, it is assumed that states in cluster 0 were all condensed in a certain area which explains the rise in total cases. It is also noticed in 4.21 that the mobility number were way less in states of cluster 0 as different states enforced lock-down during different time depending on the how severe the situation was source.Also 4.21 shows that through out the wave, the visits to pharmacy and grocery has increased indicating people seeking medications and their needs, also visits to residential places decreased through the wave indicating the effect of the lock down. This suggests that the states of cluster 0 tried to take the corrective actions by decreasing mobility to workplaces and retail and re-creations to help curb the disease.

Cluster 1 : is having a lower risk than states of cluster 0. States in cluster 0 has a lower age population which could explain the low number in total deceased, confirmed

and tested. Cluster 1 has a slightly a higher population density than cluster 0, however maybe the population isn't equally divided in states of cluster 0.

This attempt doesn't give enough insights to the reality of the cases as the numbers aren't normalized with reference to the population as mentioned above 4.1.1. That is why a second attempt was conducted taking into consideration the population of each state to normalize my features.

	Cluster 0	Cluster 1
Mean Population Density	141	150.733333
Mean Area	222,390.67	174,135.70
Mean Population	17,557,610	3,677,432
Mean Pop. Age 0-9	2,274,401	499,722.10
Mean pop. Age 10-19	2,391,350	507,177.80
Mean pop. Age 20-29	2,583,910	548,400.90
Mean pop. Age 30-39	2,522,384	536,015
Mean pop. Age 40-49	2,278,528	482,679.80
Mean pop. Age 50-59	2,314,152	510,908.60
Mean pop. Age 60-69	2,101,909	470,691.60
Mean pop. Age 70-79	1,339,069	288,083.44
Mean pop. Age 80++	735,714.20	147,758.67
Mobillity Retail & Reacreation #1	-46.444444	-36.02828
Mobillity Retail & Reacreation #2	-42.333333	-27.774641
Mobillity Retail & Reacreation #3	-31.888889	-19.465964
Mobillity Retail & Reacreation #4	-15.888889	-3.976296
Mobility Workplace #1	-53	-40.916733
Mobility Workplace #2	-52.555556	-43.641685
Mobility Workplace #3	-15.222222	-13.073203
Mobility Workplace #4	-39.111111	-31.446972
Mobility Grocery & pharmacy #1	-22.333333	-13.637543
Mobility Grocery & pharmacy #2	-15.777778	-6.068293
Mobility Grocery & pharmacy #3	-5	1.395404
Mobility Grocery & pharmacy #4	-5	3.284532
Mobility Residential #1	22.111111	16.033383
Mobility Residential #2	21.777778	16.505214
Mobility Residential #3	7.333333	5.366294
Mobility Residential #4	11.777778	8.717444

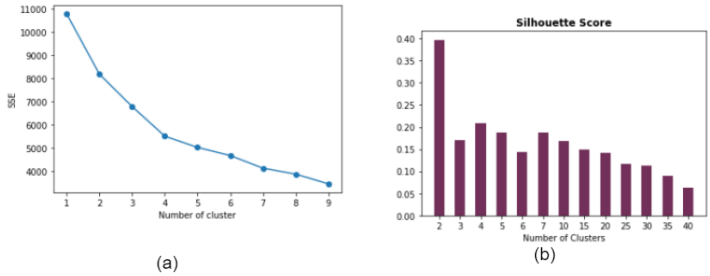
Figure 4.21: Mean Values of some features between both cluster 0 and 1. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk.

4.2.2 Second Attempt: Normalization on population

In this attempt, the features mentioned in 3.9 were used which are normalized using population to give a better sense of the data.The data was scaled using $'MinMaxScale'$.

Cluster Validation Indecencies

After Applying PCA with varience 95% , Both Silhouette method and DB index suggested an optimal k =2 as shown in 4.22. The elbow method is unclear whether it's k=2 or k=4, but since the other 2 evaluation methods suggested the same result, therfore a k=2 was applied.



(a) (b)

The db for : 2 Clusters is: 0.9676501581826124
The db for : 3 Clusters is: 1.6090586197825598
The db for : 4 Clusters is: 1.317230047986124
The db for : 5 Clusters is: 1.2483027049625384
The db for : 6 Clusters is: 1.1425516543778693

(c)

Figure 4.22: USA's second attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA

Clustering Results

Cluster 0 states: Connecticut, District of Columbia , Illinois, Louisiana, Massachusetts, Maryland, Michigan, New Jersey, New York, Pennsylvania, and Pennsylvania.

Feature	Min	Mean	MAX
Total confirmed percentage	0.671600	1.312920	1.984900
Total Tested percentage	7.123600	13.441190	19.513800
Total deceased percentage	0.051700	0.090140	0.166300
.....	.....	.....	.....
Area	177	89,305	250,493
Population Density	33.00	547.70	3,987.0

Table 4.10: Table related to USA second attempt , containing the minimum , mean and maximum of some of the features that are related to states of cluster 0

Cluster 1 Counties: Alaska, Alabama, Arkansas, Arizona, California, Colorado, Delaware, Georgia, Guam, Hawaii, Iowa, Idaho, Indiana, Kansas, Kentucky, Maine, Minnesota, Missouri, Northern Mariana Islands, Mississippi, Montana, North Carolina, North Dakota, Nebraska, New Hampshire, New Mexico, Nevada, Ohio, Oklahoma, Oregon, Puerto Rico, South Carolina, South Dakota, Tennessee, Texas, Utah, Virginia, Virgin Islands, Vermont, Virgin Islands, Vermont, Washington, Wisconsin, West Virginia, and Wyoming

Feature	Min	Mean	MAX
Total confirmed percentage	0.051300	0.512291	1.236800
Total_Tested	2.694900	9.838305	23.354300
Total_deceased	0.001300	0.015052	0.052200
.....	.....	.....	.....
Area	346.0	203,285.00	171,7856
Population Density	14.00	58.522727	351.0

Table 4.11: Table related to USA second attempt , containing the minimum , mean and maximum of some of the features that are related to states of cluster 1

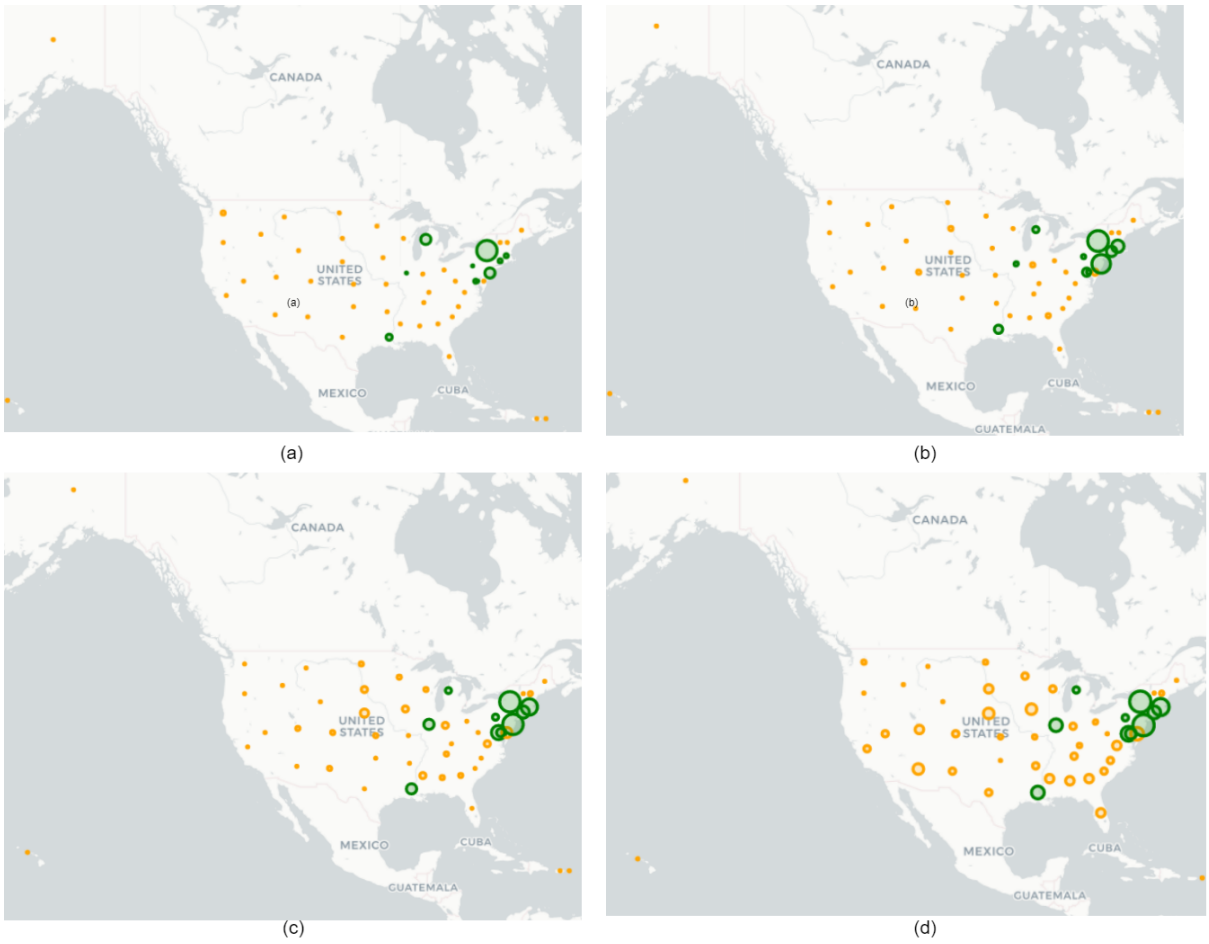


Figure 4.23: USA's second trial where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the first attempt. The orange closer are for cluster 1, and the green color is for cluster 1. The radius of the circle represents the total confirmed cases percentage in each cluster in order to give a better visualization of each cluster

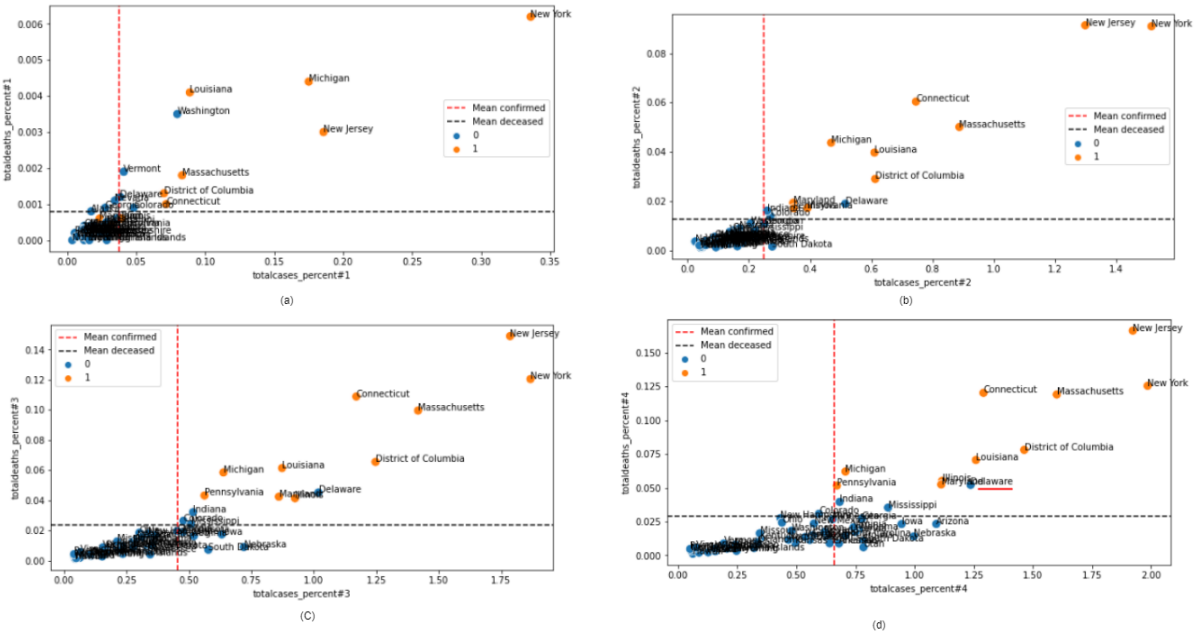


Figure 4.24: USA's second trial where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases percentage along with the to total deceased percentage

Interpretation

In this trial , As shown in map 4.23 along 4 months of wave 1; states of cluster 0 had a larger number of total confirmed cases percentage which is also shown in tables 4.10 and 4.11. Through the figures 4.24 , It was shown how counties of cluster 0 are related with reference to both total confirmed cases percentage and total deaths percentage.

The appearance of Delaware in 4.24 raised some questions, allowing a further dig into Delware's features through wave 1. As shown in fig 4.25, The figure shows that as we go further into the wave; as the distance from center of cluster 0 decreases the total confirmed cases percentage increases. Delware didn't have similar features as states in cluster 0 during first month or two but started to be similar to cluster 0 by the 3rd and 4th month. Delware's feature- population , area , total confirmed cases percentage,etc.. were similar to those of cluster 0 and was considered in emergency state over a year ago (by time of wave 1 source. In this attempt maybe, it wasn't part of the cluster due to it's performance in the first months.

To sum up this trial, **Cluster 0 :** is suggested to have a higher risk of COVID-19 as it has higher number of percentage of total confirmed cases, total tested as well as total deceased. Cluster 0 states have a very higher population density than states of cluster 1 which suggests that a large number of people in a smaller area are at higher risk to spread the disease.. It is also noticed that the mobility number were way less in states of cluster

0 as different states enforced lock-down during different time.Also the visits to pharmacy and grocery has increased. This suggests that the states of cluster 0 tried to take the corrective actions by decreasing mobility to workplaces and retail and re-creations to help curb the disease.

Both cluster 1 and 0 have very close percentage of different age groups. So age groups didn't have a high impact in this attempt.

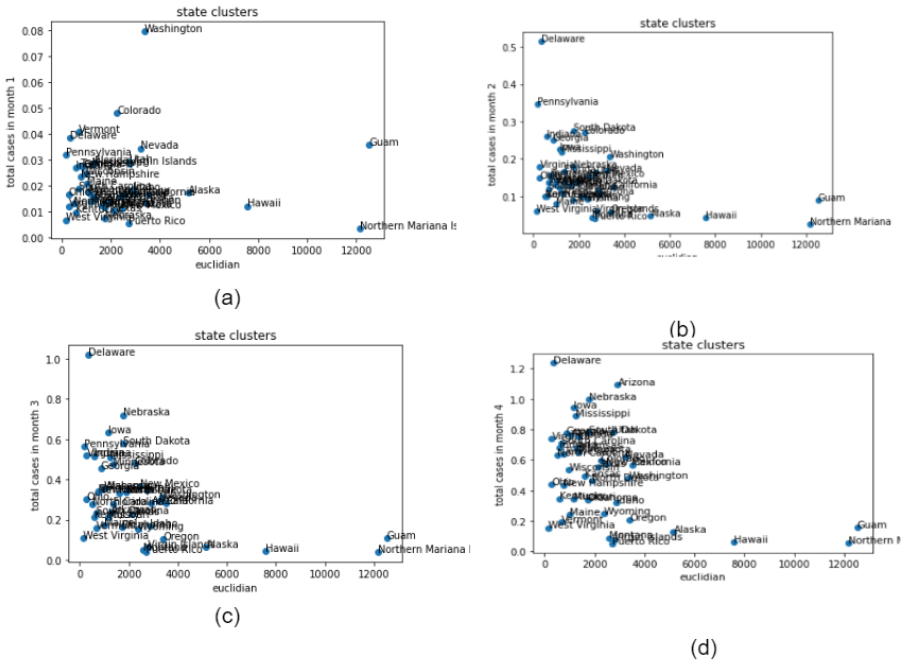


Figure 4.25: Distance between center of cluster 0 from USA's second attempt and states of cluster 1 were calculated and on the y-axis is the total confirmed cases %. (a) is the first month in wave 1 and (d) being the forth.

4.2.3 Third Attempt: Normalization on Density

Prediction of Total Confirmed Cases

In the part, I will divide my results into 2 sub sections according to time Intervals 3.5.1. The first section will cover the predictions of the total confirmed cases in the USA; where wave 1 was a part of training data to predict what will happen in wave 2 at different periods. The second section will cover the predictions of the total confirmed cases in the USA; where wave 1 wasn't part of training data, however, the beginning of wave 2 was part of training data to try and predict what happened further into wave 2 at different periods.

4.3 Predictions Using Wave 1

This sections will be divided into the 4 trials that were studied using 8 Learning models to predict the total confirmed cases of wave 2 of different period length along with their evaluation.

4.3.1 First Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 1 (From March to end of July) to predict total confirmed cases of the first 3 months $\tilde{9}3$ days in Wave 2 (From August to 1st of November). The results of this trial is in the figures below 4.27, and 4.26 and the predicted values of each model is also shown in the table below 4.12.

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2020-08-01	4582276.0	3.736949e+06	5.241082e+06	4.582661e+06	4.577062e+06	4.588692e+06	4.590847e+06	4.588758e+06	4.585708e+06
2020-08-02	4629459.0	3.765075e+06	5.365859e+06	4.645750e+06	4.635179e+06	4.653328e+06	4.658081e+06	4.649151e+06	4.644110e+06
2020-08-03	4678610.0	3.793201e+06	5.493586e+06	4.708840e+06	4.685581e+06	4.719060e+06	4.725758e+06	4.707480e+06	4.698203e+06
2020-08-04	4728239.0	3.821327e+06	5.624318e+06	4.771929e+06	4.742142e+06	4.786245e+06	4.793879e+06	4.767894e+06	4.756054e+06
2020-08-05	4781612.0	3.849452e+06	5.758111e+06	4.835019e+06	4.807506e+06	4.854350e+06	4.862444e+06	4.833333e+06	4.818907e+06
...	...	...	...	...	...	...	...	...	...
2020-10-28	8763682.0	6.212022e+06	3.572784e+07	1.013454e+07	9.808636e+06	1.208787e+07	1.220649e+07	1.205550e+07	9.649589e+06
2020-10-29	8852730.0	6.240148e+06	3.642861e+07	1.019763e+07	9.883676e+06	1.219271e+07	1.231279e+07	1.215982e+07	9.706385e+06
2020-10-30	8952086.0	6.268274e+06	3.714109e+07	1.026072e+07	9.964744e+06	1.229800e+07	1.241952e+07	1.226546e+07	9.763513e+06
2020-10-31	9032465.0	6.296400e+06	3.786541e+07	1.032381e+07	1.002206e+07	1.240373e+07	1.252671e+07	1.237343e+07	9.819277e+06
2020-11-01	9108353.0	6.324526e+06	3.860175e+07	1.038690e+07	1.007457e+07	1.250990e+07	1.263433e+07	1.248343e+07	9.874245e+06

93 rows × 9 columns

Table 4.12: First trial of prediction from wave 1 of All 8 models along with the valid confirmed cases value

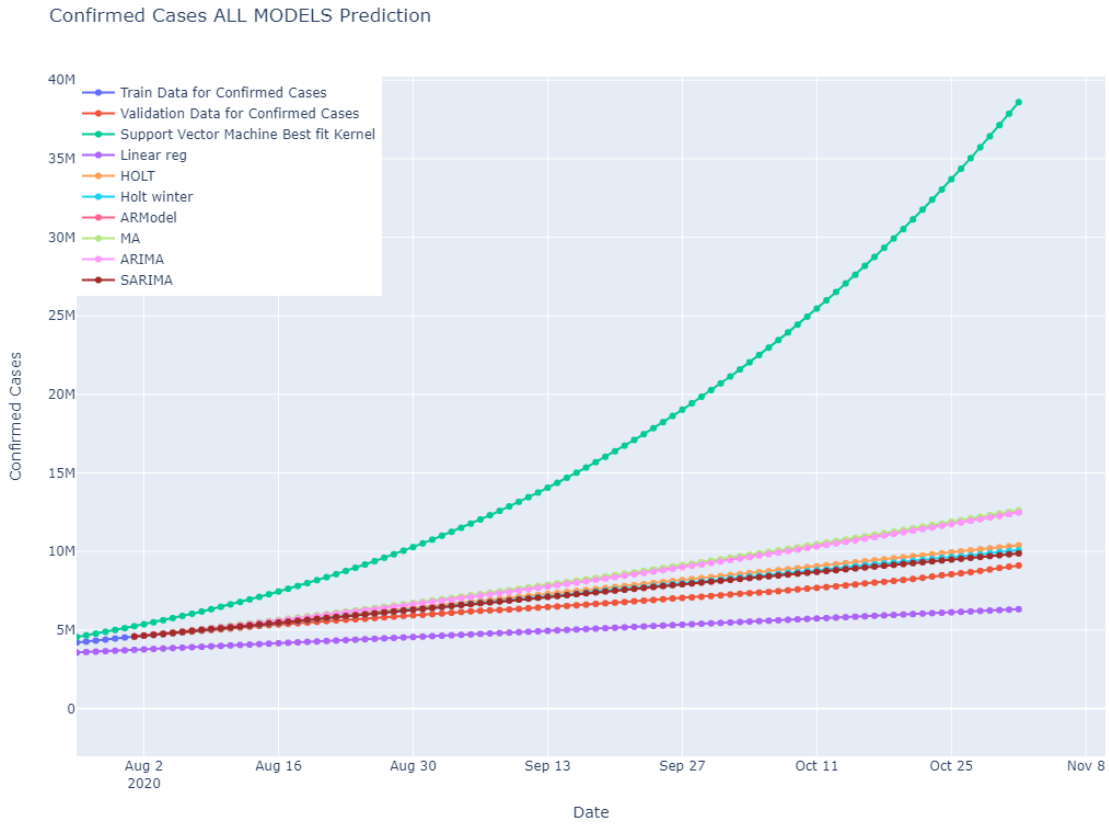


Figure 4.26: First trial of prediction from wave 1 of All 8 models joined Zoomed in

Evaluation

In The first trial , SARIMA model performed best as shown in table 4.13. SARIMA had the lowest RMSE with a value of 6.994140e+05 along with a low MAPE of 8.396200. Furthermore, in figure 4.26 SARIMA's plot was the closest to true validation total confirmed cases values. SVR performs worst in this situation

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
SARIMA	6.994140e+05	8.396200	4.891800e+11	6.029436e+05
Holt's Winter	7.723829e+05	9.057681	5.965753e+11	6.559918e+05
Holt's Linear	9.632732e+05	11.404604	9.278953e+11	8.242124e+05
LR	1.694696e+06	24.030960	2.871995e+12	1.629829e+06
ARIMA	1.888765e+06	20.705646	3567432e+12	1.531744e+06
AR	1.931419e+06	21.440914	3.730380e+12	1.580560e+06
MA	2.002229e+06	22.300358	4.008923e+12	1.642353e+06
SVR	1.359448e+07	172.778467	1.848099e+14	1.067812e+07

Table 4.13: Evaluation for trial 1

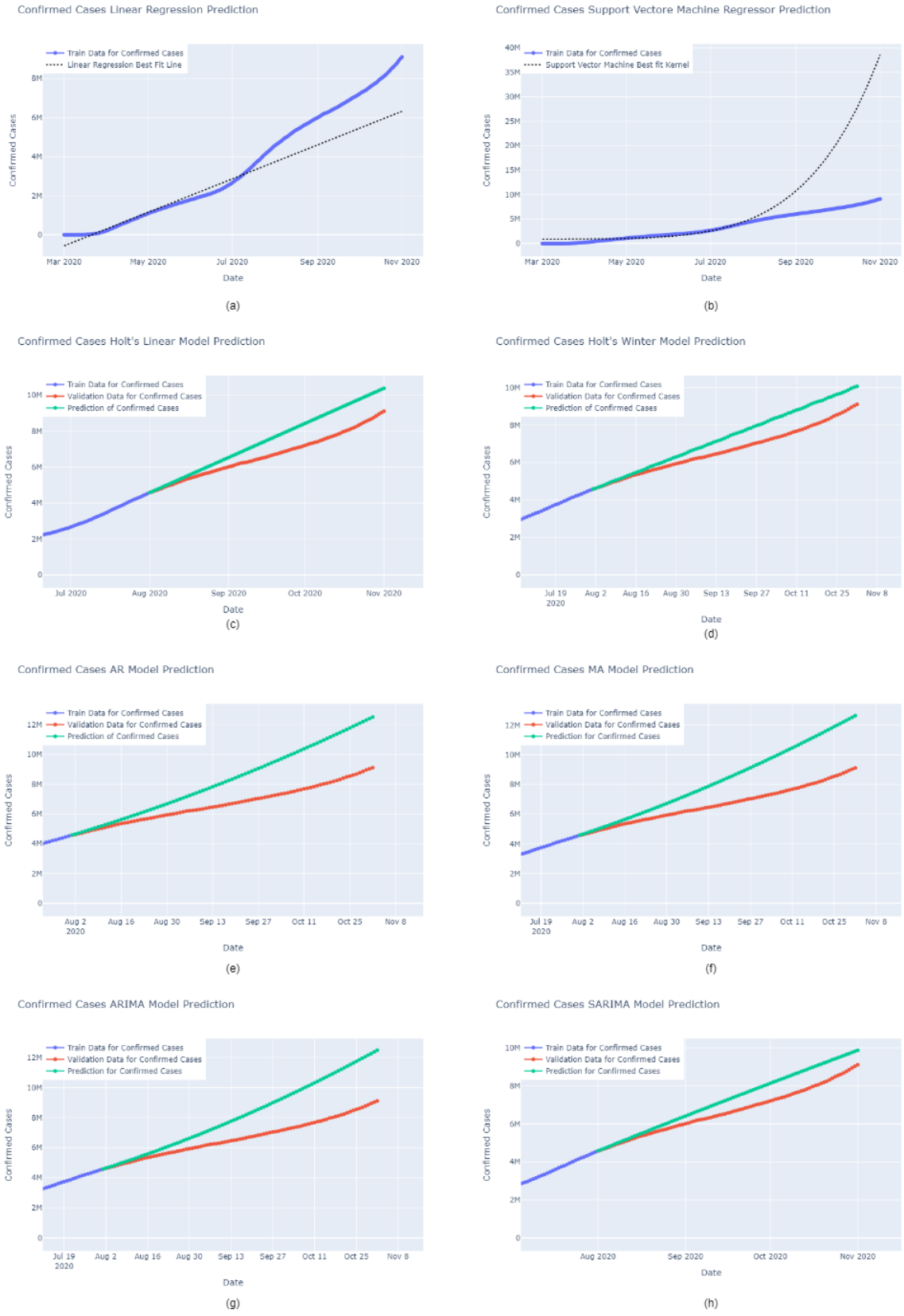


Figure 4.27: First trial of prediction from wave 1 of All 8 models

4.3.2 Second Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 1 (From March to end of July) and part of wave 2 (from August to the end of September) to predict total confirmed cases of the 1 months $\tilde{3}2$ days in the beginning of Wave 2 (from October to the first day in November). The results of this trial is in the figures below 4.29, and 4.28 and the predicted values of each model is also shown in the table below 4.14.

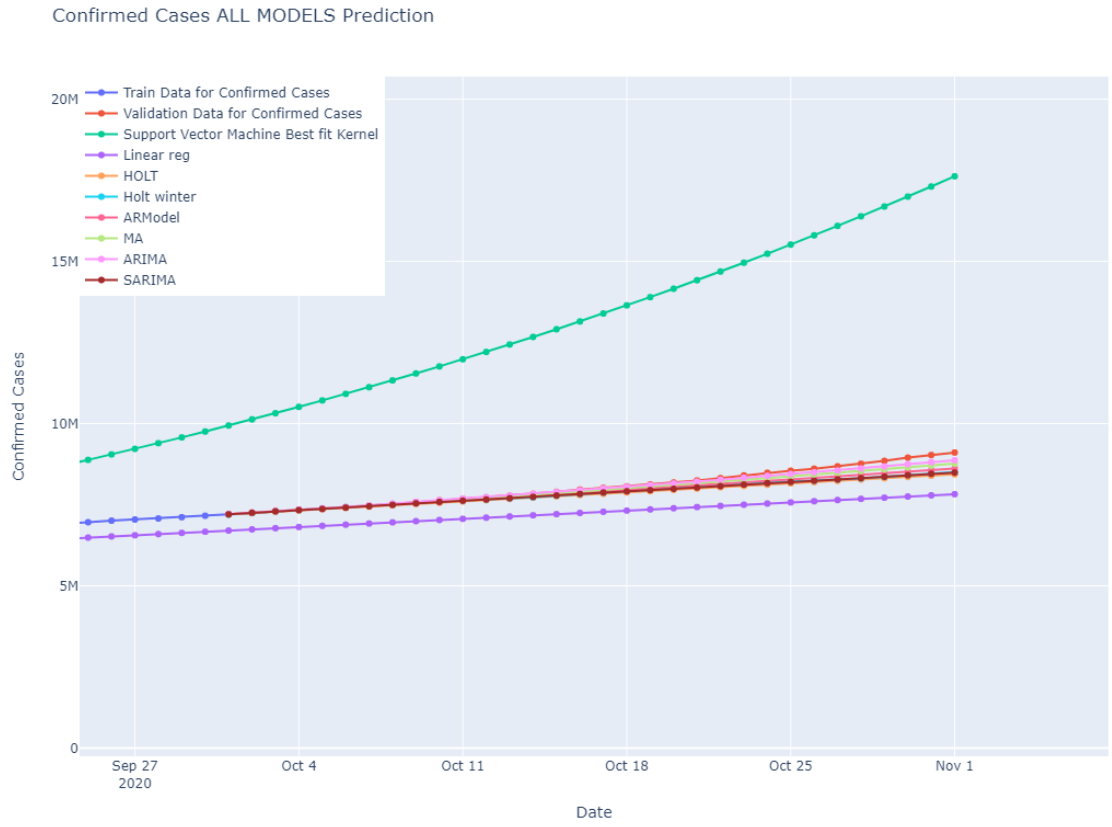


Figure 4.28: Second trial of prediction from wave 1 of All 8 models joined Zoomed in

Evaluation

In The Second trial , ARIMA model performed best as shown in table 4.15. ARIMA had the lowest RMSE with a value of 8.974896e+04 along with a low MAPE of 0.637762. Furthermore, in figure 4.28 ARIMA's plot was the closest to true validation total confirmed cases values. MA and AR models performed relatively well and achieve a low MAPE with very close values to one another along with lower RMSE than other Models. SVR performs worst in this situation.

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2020-10-01	7206769.0	6.704140e+06	9.941493e+06	7.201836e+06	7.196649e+06	7.207115e+06	7.204417e+06	7.209331e+06	7.204001e+06
2020-10-02	7256234.0	6.740318e+06	1.012964e+07	7.241613e+06	7.250816e+06	7.251117e+06	7.247768e+06	7.259564e+06	7.252232e+06
2020-10-03	7305270.0	6.776495e+06	1.032123e+07	7.281390e+06	7.297864e+06	7.292385e+06	7.291588e+06	7.307056e+06	7.297392e+06
2020-10-04	7341406.0	6.812673e+06	1.051633e+07	7.321167e+06	7.338625e+06	7.333223e+06	7.335877e+06	7.350211e+06	7.334970e+06
2020-10-05	7380326.0	6.848851e+06	1.071498e+07	7.360944e+06	7.379859e+06	7.375598e+06	7.380634e+06	7.391132e+06	7.370454e+06
2020-10-06	7419230.0	6.885028e+06	1.091724e+07	7.400721e+06	7.428549e+06	7.419589e+06	7.425860e+06	7.434011e+06	7.410225e+06
2020-10-07	7471688.0	6.921206e+06	1.112316e+07	7.440498e+06	7.466370e+06	7.464044e+06	7.471555e+06	7.481972e+06	7.452554e+06
2020-10-08	7525920.0	6.957383e+06	1.133278e+07	7.480275e+06	7.505321e+06	7.508061e+06	7.517719e+06	7.534738e+06	7.495790e+06
2020-10-09	7583748.0	6.993561e+06	1.154617e+07	7.520052e+06	7.553738e+06	7.551615e+06	7.564351e+06	7.588867e+06	7.543580e+06
2020-10-10	7636803.0	7.029738e+06	1.176337e+07	7.559829e+06	7.600380e+06	7.595251e+06	7.611453e+06	7.640360e+06	7.588408e+06
2020-10-11	7682128.0	7.065916e+06	1.198445e+07	7.599606e+06	7.632642e+06	7.639400e+06	7.659023e+06	7.687661e+06	7.625913e+06
2020-10-12	7728436.0	7.102093e+06	1.220945e+07	7.639383e+06	7.662852e+06	7.684063e+06	7.707062e+06	7.732803e+06	7.661401e+06
2020-10-13	7774745.0	7.138271e+06	1.243844e+07	7.679160e+06	7.699434e+06	7.728979e+06	7.755570e+06	7.779855e+06	7.701034e+06
2020-10-14	7833851.0	7.174448e+06	1.267146e+07	7.718937e+06	7.730183e+06	7.773947e+06	7.804546e+06	7.831852e+06	7.743143e+06
2020-10-15	7896895.0	7.210626e+06	1.290858e+07	7.758714e+06	7.778951e+06	7.818970e+06	7.853992e+06	7.888532e+06	7.786132e+06
2020-10-16	7966729.0	7.246804e+06	1.314986e+07	7.798491e+06	7.834129e+06	7.864171e+06	7.903906e+06	7.946562e+06	7.833524e+06
2020-10-17	8019237.0	7.282981e+06	1.339534e+07	7.838268e+06	7.881609e+06	7.909646e+06	7.954289e+06	8.002057e+06	7.878057e+06
2020-10-18	8065615.0	7.319159e+06	1.364509e+07	7.878045e+06	7.922295e+06	7.955392e+06	8.005141e+06	8.053502e+06	7.915515e+06
2020-10-19	8124633.0	7.355336e+06	1.389917e+07	7.917822e+06	7.963493e+06	8.001348e+06	8.056462e+06	8.102861e+06	7.951025e+06
2020-10-20	8184788.0	7.391514e+06	1.415764e+07	7.957599e+06	8.012733e+06	8.047472e+06	8.108251e+06	8.154082e+06	7.990543e+06
2020-10-21	8248149.0	7.427691e+06	1.442055e+07	7.997376e+06	8.050249e+06	8.093765e+06	8.160509e+06	8.210115e+06	8.032456e+06
2020-10-22	8320491.0	7.463869e+06	1.468798e+07	8.037153e+06	8.088986e+06	8.140255e+06	8.213236e+06	8.270714e+06	8.075220e+06
2020-10-23	8403121.0	7.500046e+06	1.495998e+07	8.076930e+06	8.137923e+06	8.186963e+06	8.266432e+06	8.332649e+06	8.122241e+06
2020-10-24	8485747.0	7.536224e+06	1.523661e+07	8.116707e+06	8.184944e+06	8.233888e+06	8.320097e+06	8.392149e+06	8.166500e+06
2020-10-25	8548111.0	7.572401e+06	1.551794e+07	8.156484e+06	8.216479e+06	8.281016e+06	8.374231e+06	8.447733e+06	8.203922e+06
2020-10-26	8611256.0	7.608579e+06	1.580403e+07	8.196261e+06	8.245814e+06	8.328338e+06	8.428833e+06	8.501305e+06	8.239462e+06
2020-10-27	8683298.0	7.644757e+06	1.609494e+07	8.236038e+06	8.282014e+06	8.375854e+06	8.483904e+06	8.556691e+06	8.278877e+06
2020-10-28	8763682.0	7.680934e+06	1.639074e+07	8.275815e+06	8.311945e+06	8.423571e+06	8.539444e+06	8.616761e+06	8.320607e+06
2020-10-29	8852730.0	7.717112e+06	1.669149e+07	8.315592e+06	8.361253e+06	8.471493e+06	8.595453e+06	8.681284e+06	8.363159e+06
2020-10-30	8952086.0	7.753289e+06	1.699727e+07	8.355368e+06	8.417443e+06	8.519620e+06	8.651931e+06	8.747129e+06	8.409826e+06
2020-10-31	9032465.0	7.789467e+06	1.730814e+07	8.395145e+06	8.465353e+06	8.567949e+06	8.708877e+06	8.810633e+06	8.453825e+06
2020-11-01	9108353.0	7.825644e+06	1.762416e+07	8.434922e+06	8.505965e+06	8.616477e+06	8.766292e+06	8.870355e+06	8.491216e+06

Table 4.14: Second trial of prediction from wave 1 of All 8 models along with the valid confirmed cases value

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
ARIMA	8.974896e+04	0.637762	8.054876e+09	5.528526e+04
MA	1.397174e+05	1.112582	1.952096e+10	9.559729e+04
AR	2.096010e+05	1.734869	4.393258e+10	1.485867e+05
Holt's WInter	2.637914e+05	2.200723	6.958592e+10	1.883663e+05
SARIMA	2.692969e+05	2.274045	7.252080e+10	1.943980e+05
Holt's Linear	3.018411e+05	2.653803	9.110805e+10	2.258064e+05
LR	8.131725e+05	9.539016	6.612494e+11	7.792933e+05
SVR	5.677271e+06	68.047396	3.223141e+13	5.408930e+06

Table 4.15: Evaluation for trial 2

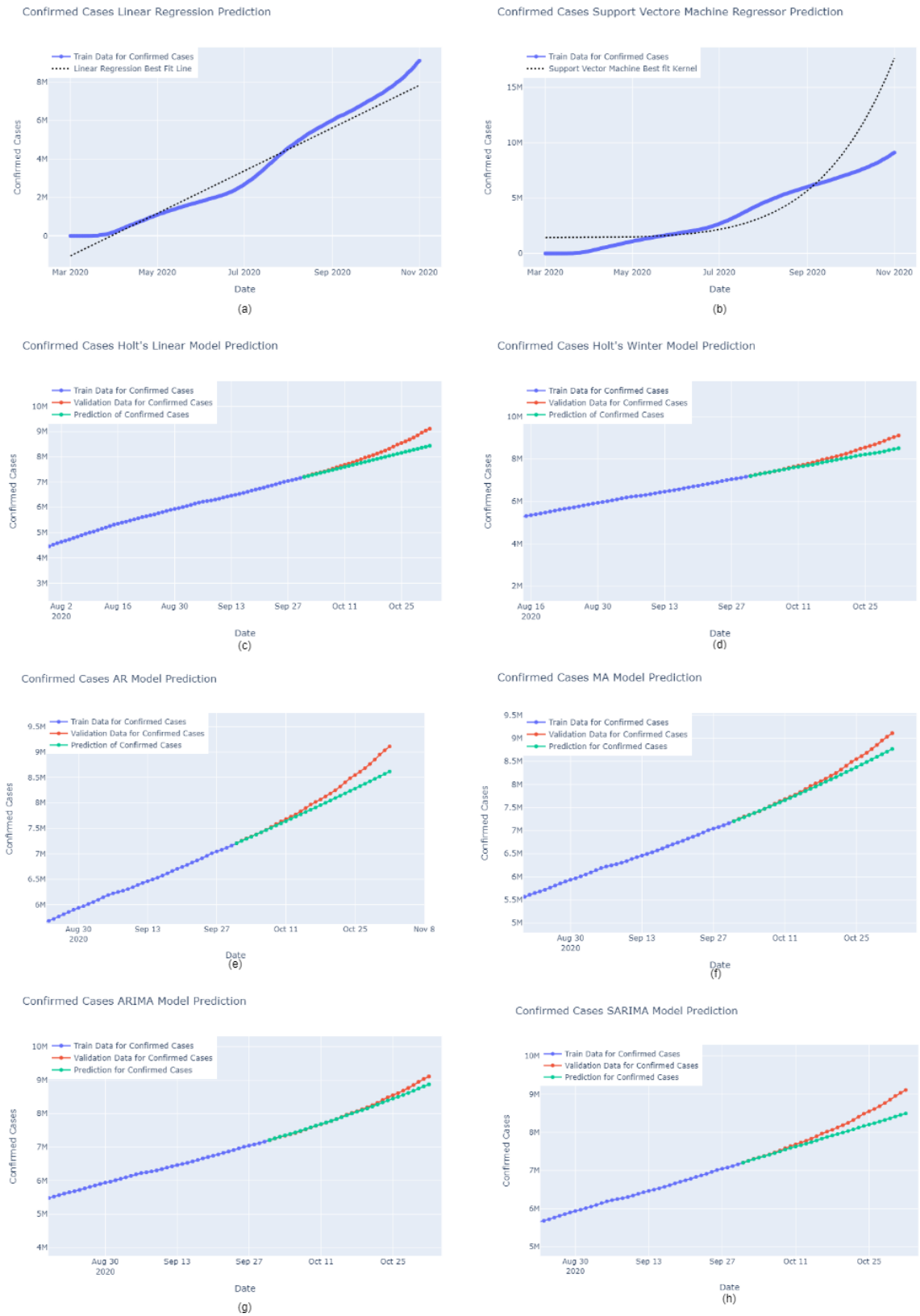


Figure 4.29: Second trial of prediction from wave 1 of All 8 models

4.3.3 Third Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 1 (From March to end of July) and part of wave 2 (from August to the mid of October) to predict total confirmed cases of the last 15 days in the beginning of Wave 2 (from mid October to the first day in November). The results of this trial is in the figures below 4.31, and 4.30 and the predicted values of each model is also shown in the table below 4.16.

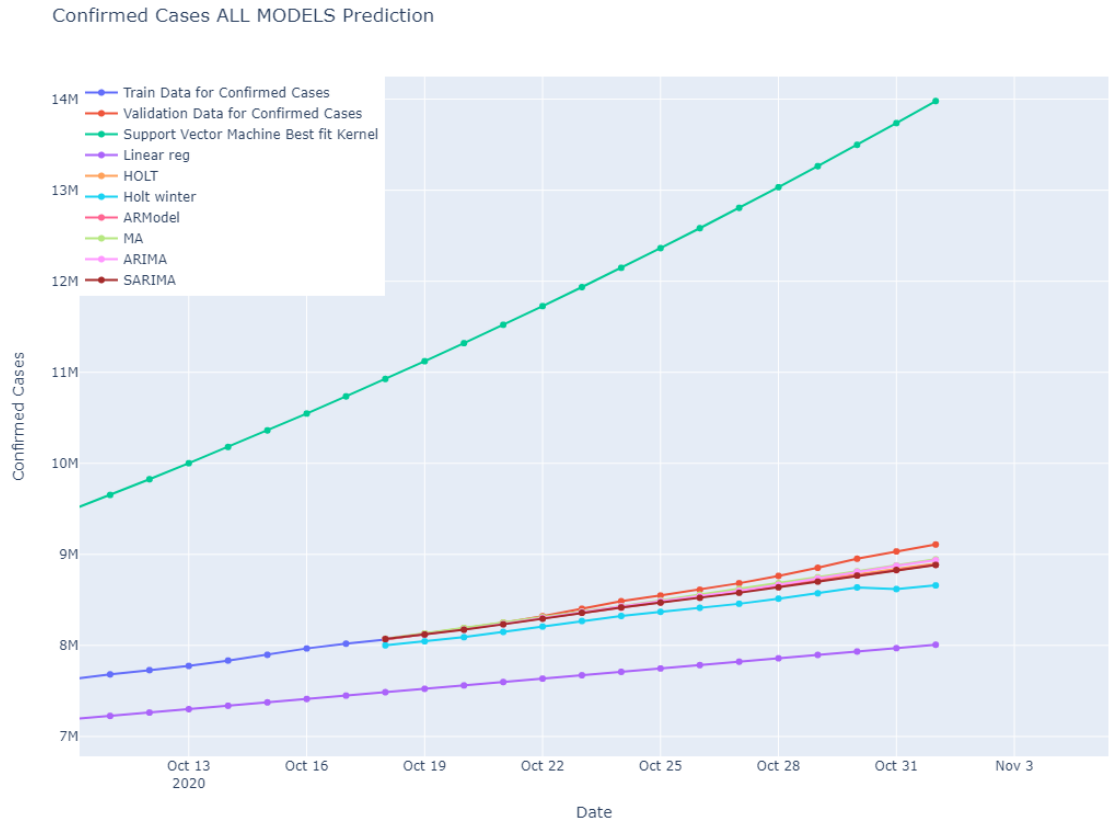


Figure 4.30: Third trial of prediction from wave 1 of All 8 models joined Zoomed in

Evaluation

In The Third trial , MA model performed best as shown in table 4.17. MA had the lowest RMSE with a value of 8.140810e+04 along with a low MAPE of 0.697435. Furthermore, in figure 4.30 MA's plot was the closest to true validation total confirmed cases values. MA and ARIMAS models performed relatively equally achieving a low MAPE with very close values to one another along with lower RMSE than other Models. SVR performs worst in this situation.

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2020-10-18	8065615.0	7.486214e+06	1.092695e+07	8.072199e+06	8.001215e+06	8.070423e+06	8.073971e+06	8.070766e+06	8.070027e+06
2020-10-19	8124633.0	7.523454e+06	1.112187e+07	8.131264e+06	8.045253e+06	8.122505e+06	8.132402e+06	8.120643e+06	8.120040e+06
2020-10-20	8184788.0	7.560693e+06	1.132017e+07	8.190329e+06	8.090477e+06	8.179104e+06	8.191421e+06	8.174473e+06	8.172418e+06
2020-10-21	8248149.0	7.597932e+06	1.152188e+07	8.249394e+06	8.148697e+06	8.240299e+06	8.251029e+06	8.235386e+06	8.231635e+06
2020-10-22	8320491.0	7.635172e+06	1.172704e+07	8.308459e+06	8.207023e+06	8.299974e+06	8.311227e+06	8.301679e+06	8.293083e+06
2020-10-23	8403121.0	7.672411e+06	1.193572e+07	8.367524e+06	8.268191e+06	8.357845e+06	8.372013e+06	8.368127e+06	8.359203e+06
2020-10-24	8485747.0	7.709650e+06	1.214794e+07	8.426589e+06	8.323482e+06	8.414146e+06	8.433388e+06	8.429945e+06	8.417983e+06
2020-10-25	8548111.0	7.746890e+06	1.236378e+07	8.485654e+06	8.368909e+06	8.470939e+06	8.495351e+06	8.486399e+06	8.471070e+06
2020-10-26	8611256.0	7.784129e+06	1.258326e+07	8.544719e+06	8.413661e+06	8.529494e+06	8.557904e+06	8.541341e+06	8.523187e+06
2020-10-27	8683298.0	7.821368e+06	1.280644e+07	8.603784e+06	8.457515e+06	8.589064e+06	8.621045e+06	8.600281e+06	8.577979e+06
2020-10-28	8763682.0	7.858608e+06	1.303338e+07	8.662850e+06	8.513800e+06	8.648819e+06	8.684775e+06	8.666214e+06	8.638572e+06
2020-10-29	8852730.0	7.895847e+06	1.326412e+07	8.721915e+06	8.572670e+06	8.708100e+06	8.749094e+06	8.737377e+06	8.701030e+06
2020-10-30	8952086.0	7.933086e+06	1.349870e+07	8.780980e+06	8.636976e+06	8.767135e+06	8.814002e+06	8.808595e+06	8.767660e+06
2020-10-31	9032465.0	7.970326e+06	1.373720e+07	8.840045e+06	8.619038e+06	8.826495e+06	8.879499e+06	8.875210e+06	8.828976e+06
2020-11-01	9108353.0	8.007565e+06	1.397964e+07	8.899110e+06	8.659215e+06	8.886450e+06	8.945584e+06	8.936592e+06	8.883941e+06

Table 4.16: Third trial of prediction from wave 1 of All 8 models along with the valid confirmed cases value

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
MA	8.140810e+04	0.697435	6.627278e+09	6.153976e+04
ARIMA	8.877659e+04	0.788573	7.881282e+09	6.945326e+04
Holt's Linear	1.023221e+05	0.859842	1.046980e+10	7.598083e+04
AR	1.123158e+05	0.969606	1.261483e+10	8.555657e+04
SARIMA	1.145916e+05	1.011203	1.313124e+10	8.910290e+04
Holt's Winter	2.340363e+05	2.336624	5.477300e+10	2.038935e+05
LR	8.285474e+05	9.429190	6.864908e+11	8.120787e+05
SVM	3.888246e+06	45.060509	1.511846e+13	3.838904e+06

Table 4.17: Evaluation for trial 3

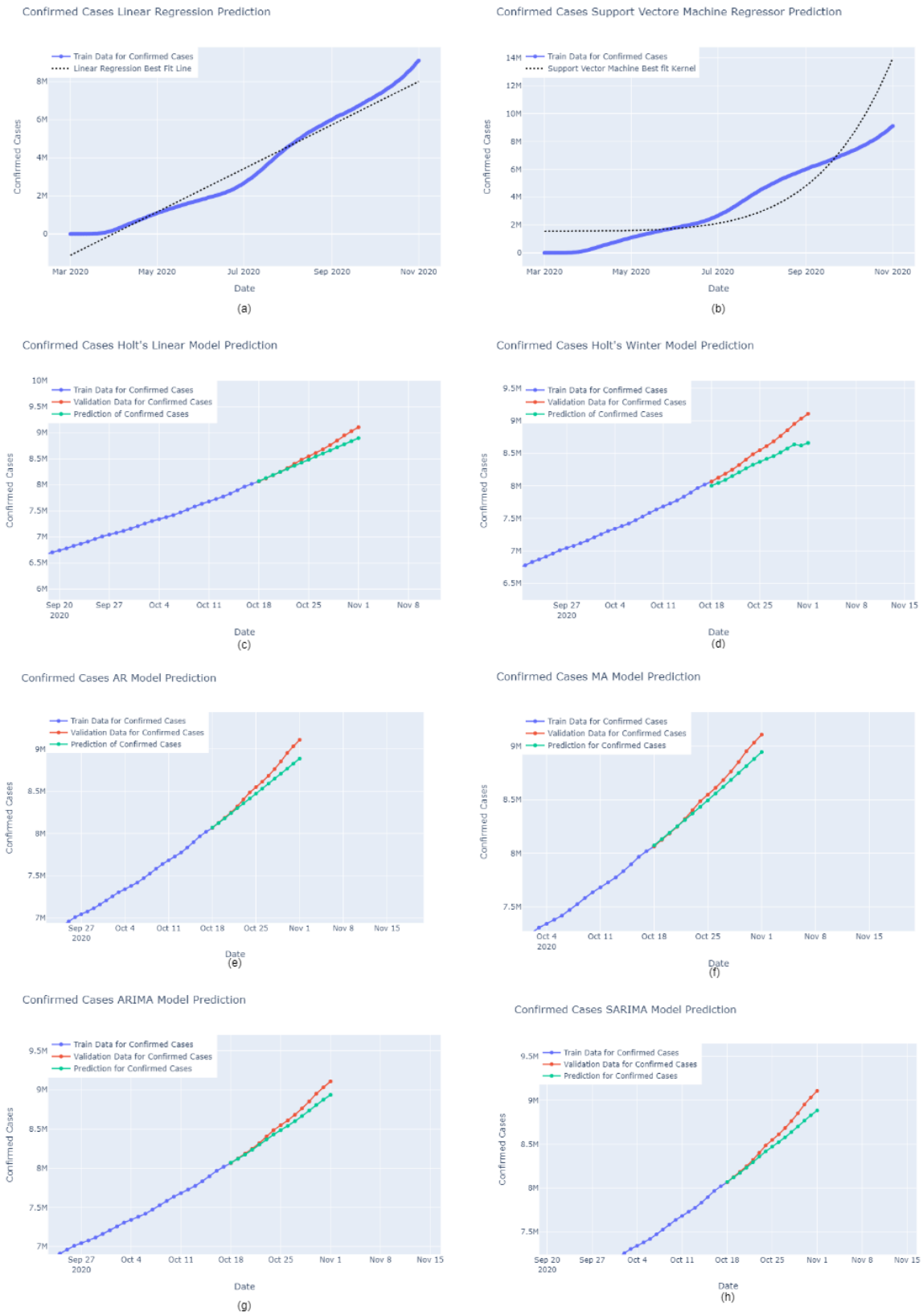


Figure 4.31: Third trial of prediction from wave 1 of All 8 models

4.3.4 Forth Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 1 (From March to end of July) and part of wave 2 (from August to the end of October) to predict total confirmed cases of the last 10 days in the beginning of Wave 2 (last 9 days in October and the 1st of November). The results of this trial is in the figures below 4.33, and 4.32 and the predicted values of each model is also shown in the table below 4.18.

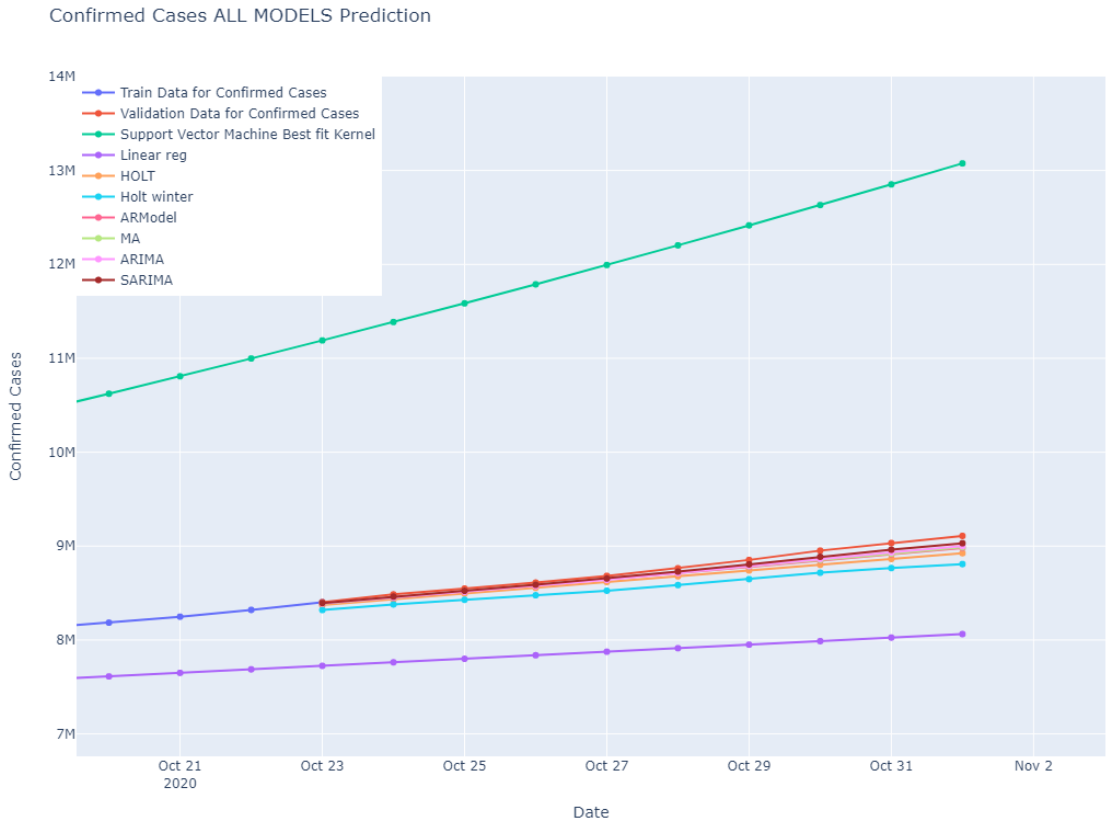


Figure 4.32: Forth trial of prediction from wave 1 of All 8 models joined Zoomed in

Evaluation

In The Forth trial , SARIMA model performed best as shown in table 4.19. SARIMA had the lowest RMSE with a value of 4.623047e+04 along with a low MAPE of 0.456509. Furthermore, in figure 4.32 SARIMA's plot was the closest to true validation total confirmed cases values. ARIMA and MA models performed relatively equally achieving a low MAPE with very close values to one another along with lower RMSE than other Models. SVR performs worst in this situation.

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2020-10-23	8403121.0	7.724984e+06	1.119049e+07	8.371506e+06	8.319755e+06	8.386249e+06	8.388409e+06	8.388805e+06	8.394231e+06
2020-10-24	8485747.0	7.762554e+06	1.138624e+07	8.432956e+06	8.379043e+06	8.450939e+06	8.453825e+06	8.453100e+06	8.462499e+06
2020-10-25	8548111.0	7.800125e+06	1.158531e+07	8.494406e+06	8.428404e+06	8.514165e+06	8.519518e+06	8.513191e+06	8.523809e+06
2020-10-26	8611256.0	7.837696e+06	1.178775e+07	8.555855e+06	8.477084e+06	8.578030e+06	8.585487e+06	8.574221e+06	8.588195e+06
2020-10-27	8683298.0	7.875266e+06	1.199360e+07	8.617305e+06	8.524857e+06	8.644937e+06	8.651734e+06	8.640452e+06	8.655129e+06
2020-10-28	8763682.0	7.912837e+06	1.220291e+07	8.678755e+06	8.585162e+06	8.712278e+06	8.718257e+06	8.713059e+06	8.728193e+06
2020-10-29	8852730.0	7.950408e+06	1.241573e+07	8.740205e+06	8.648082e+06	8.779641e+06	8.785057e+06	8.788941e+06	8.805240e+06
2020-10-30	8952086.0	7.987978e+06	1.263210e+07	8.801654e+06	8.716498e+06	8.846332e+06	8.852133e+06	8.863299e+06	8.886351e+06
2020-10-31	9032465.0	8.025549e+06	1.285207e+07	8.863104e+06	8.766141e+06	8.912740e+06	8.919487e+06	8.933179e+06	8.961975e+06
2020-11-01	9108353.0	8.063120e+06	1.307569e+07	8.924554e+06	8.808508e+06	8.979825e+06	8.987117e+06	8.999650e+06	9.030684e+06

Table 4.18: Forth trial of prediction from wave 1 of All 8 models along with the valid confirmed cases value

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
SARIMA	4.623047e+04	0.456509	2.137256e+09	4.045430e+04
ARIMA	6.471971e+04	0.646977	4.188641e+09	5.729522e+04
MA	6.914438e+04	0.652966	4.780945e+09	5.798265e+04
AR	7.432760e+04	0.716582	5.524592e+09	6.357137e+04
Holt's Linear	1.091167e+05	1.084348	1.190645e+10	9.605491e+04
Holt's Winter	1.912742e+05	2.025357	3.658582e+10	1.787315e+05
LR	8.583525e+05	9.692580	7.367689e+11	8.500332e+05
SVM	3.388952e+06	38.611571	1.148500e+13	3.368104e+06

Table 4.19: Evaluation for trial 4

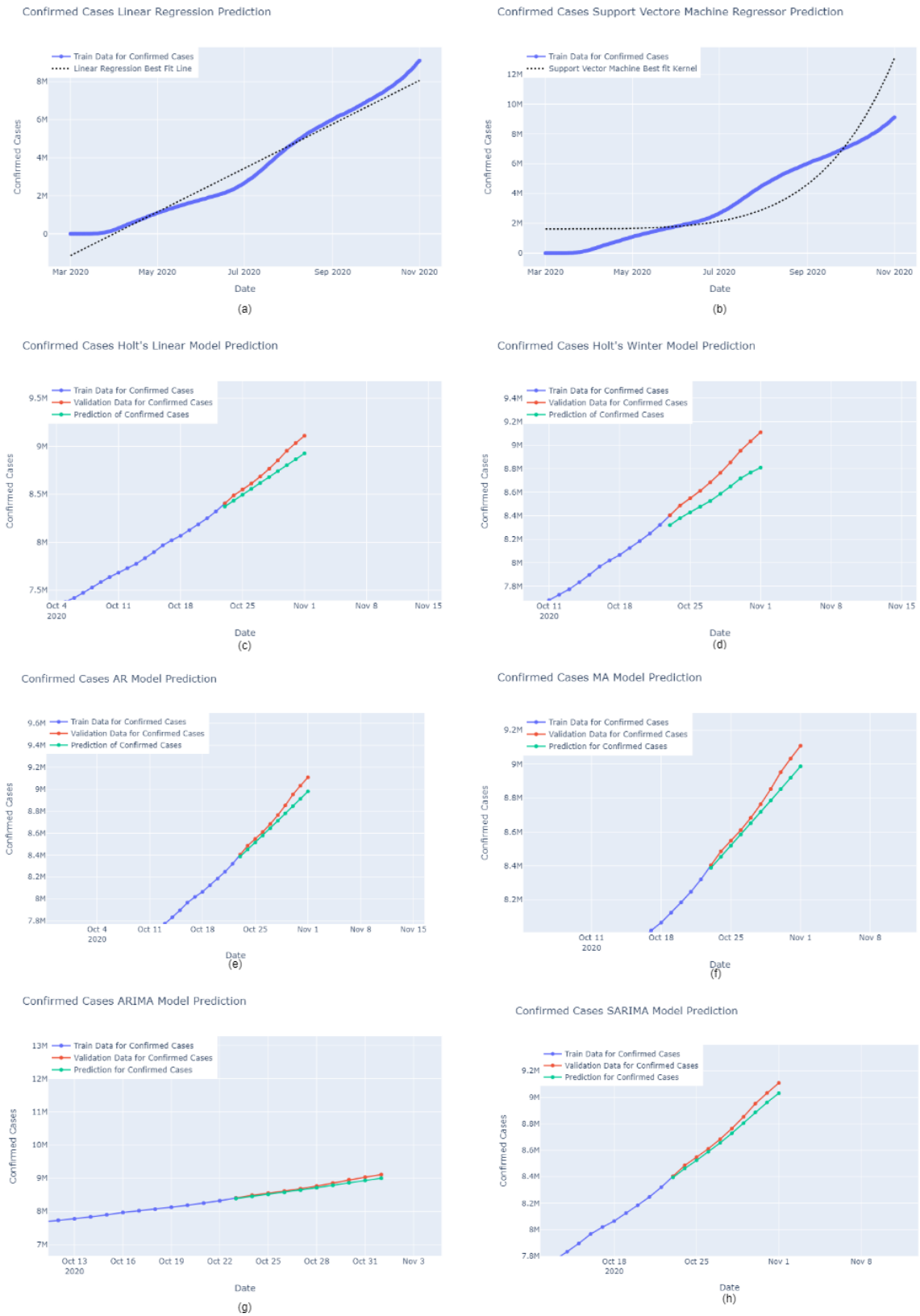


Figure 4.33: Forth trial of prediction from wave 1 of All 8 models

4.3.5 Discussion

The less days to predict the total confirmed cases, the better the performance of the learning Models and gives a better result of the reality of the growing pandemic. SARIMA Models performs best for both Trial 1 and Trial 4. Here, the SARIMA model is the model obtained by auto.arima function with an order of (0, 2, 1) and a seasonalorder of (1, 0, 2, 7).

SVR Model performed the worst among all models in all trials, it isn't providing great results , the predictions are either overshooting or really lower than what's expected as shown in the above figures 4.33 (b)

4.4 Predictions Using the start of Wave 2

This sections will be divided into the 4 trials that were studied using 8 Learning models to predict the total confirmed cases later into wave 2 of different period length along with their evaluation.

4.4.1 First Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 2 (From August to mid of January) to predict total confirmed cases of $\tilde{4}5$ days in Wave 2 (From mid January to end of February). The results of this trial is in the figures below 4.35, and 4.34 and the predicted values of each model is also shown in the table below 4.20.

Evaluation

In The Forth trial , Holt's Linear model performed best as shown in table 4.21. Holt's Linear had the lowest RMSE with a value of 2.537801e+06 along with a low MAPE of 7.390732. Furthermore, in figure 4.34 Holt's Linear's plot was the closest to true validation total confirmed cases values. ARIMA, MA, and AR models performed equally achieving a low MAPE with very close values to one another along with lower RMSE than other Models. SVR performs worst in this situation.

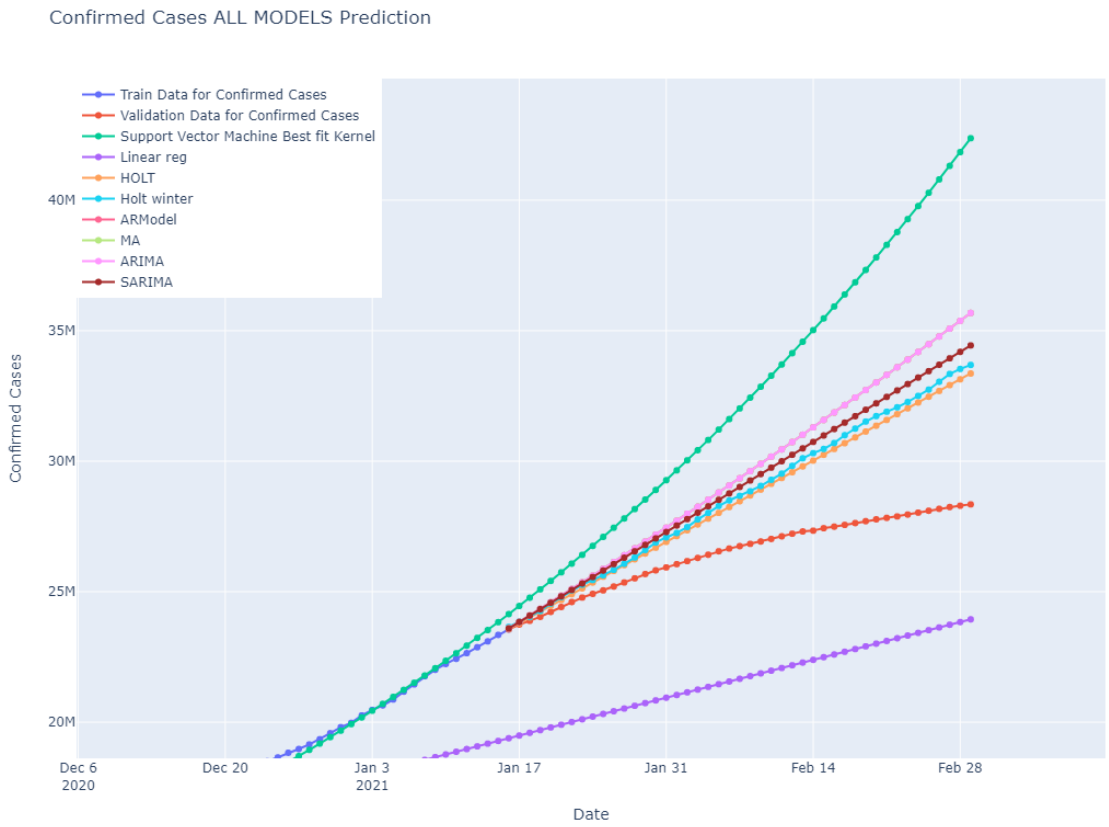


Figure 4.34: First trial of prediction using wave 2 of All 8 models joined Zoomed in

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
Holt's Linear	2.537801e+06	7.390732	6.440431e+12	2.025650e+06
Holt's Winter	2.755917e+06	8.133395	7.595077e+12	2.226525e+06
SARIMA	3.161682e+06	9.406064	9.996234e+12	2.573566e+06
AR	3.732445e+06	10.977141	1.393115e+13	3.006187e+06
MA	3.732445e+06	10.977141	1.393115e+13	3.006187e+06
ARIMA	3.732445e+06	10.977141	1.393115e+13	3.006187e+06
LR	4.784249e+06	18.101384	2.288904e+13	4.777275e+06
SVM	7.239018e+06	25.101704	5.240338e+13	6.012952e+06

Table 4.21: Evaluation for trial 1 in prediction using wave 2

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2021-01-16	23556876.0	1.938484e+07	2.414109e+07	2.357055e+07	2.365259e+07	2.359211e+07	2.359211e+07	2.359211e+07	23590908.0
2021-01-17	23742059.0	1.948838e+07	2.445305e+07	2.379305e+07	2.385548e+07	2.384099e+07	2.384099e+07	2.384099e+07	23837393.0
2021-01-18	23884299.0	1.959192e+07	2.476913e+07	2.401556e+07	2.403073e+07	2.409108e+07	2.409108e+07	2.409108e+07	24063878.0
2021-01-19	24037236.0	1.969546e+07	2.508935e+07	2.423806e+07	2.425805e+07	2.434237e+07	2.434237e+07	2.434237e+07	24330363.0
2021-01-20	24225155.0	1.979900e+07	2.541378e+07	2.446057e+07	2.454410e+07	2.459486e+07	2.459486e+07	2.459486e+07	24576848.0
2021-01-21	24413331.0	1.990254e+07	2.574244e+07	2.468307e+07	2.479235e+07	2.484855e+07	2.484855e+07	2.484855e+07	24823333.0
2021-01-22	24604325.0	2.000609e+07	2.607539e+07	2.490558e+07	2.505313e+07	2.510344e+07	2.510344e+07	2.510344e+07	25069818.0
2021-01-23	24775208.0	2.010963e+07	2.641267e+07	2.512808e+07	2.526775e+07	2.535953e+07	2.535953e+07	2.535953e+07	25316303.0
2021-01-24	24916899.0	2.021317e+07	2.675433e+07	2.535059e+07	2.544803e+07	2.561682e+07	2.561682e+07	2.561682e+07	25562788.0
2021-01-25	25050308.0	2.031671e+07	2.710040e+07	2.557309e+07	2.562983e+07	2.587531e+07	2.587531e+07	2.587531e+07	25809273.0
2021-01-26	25198841.0	2.042025e+07	2.745094e+07	2.579560e+07	2.583778e+07	2.613500e+07	2.613500e+07	2.613500e+07	26055758.0
2021-01-27	25354044.0	2.052379e+07	2.780599e+07	2.601811e+07	2.607101e+07	2.639589e+07	2.639589e+07	2.639589e+07	26302243.0
2021-01-28	25512197.0	2.062733e+07	2.816560e+07	2.624061e+07	2.630866e+07	2.665796e+07	2.665796e+07	2.665796e+07	26548728.0
2021-01-29	25676812.0	2.073087e+07	2.852981e+07	2.646312e+07	2.659765e+07	2.692128e+07	2.692128e+07	2.692128e+07	26795213.0
2021-01-30	25817939.0	2.083441e+07	2.889868e+07	2.668562e+07	2.688314e+07	2.718577e+07	2.718577e+07	2.718577e+07	27041698.0
2021-01-31	25930068.0	2.093795e+07	2.927224e+07	2.690813e+07	2.708226e+07	2.745147e+07	2.745147e+07	2.745147e+07	27288183.0
2021-02-01	26055512.0	2.104149e+07	2.965055e+07	2.713063e+07	2.725011e+07	2.771836e+07	2.771836e+07	2.771836e+07	27534668.0
2021-02-02	26172274.0	2.114503e+07	3.003365e+07	2.735314e+07	2.747708e+07	2.798646e+07	2.798646e+07	2.798646e+07	27781153.0
2021-02-03	26293150.0	2.124857e+07	3.042160e+07	2.757564e+07	2.777051e+07	2.825575e+07	2.825575e+07	2.825575e+07	28027638.0
2021-02-04	26418016.0	2.135212e+07	3.081444e+07	2.779815e+07	2.802107e+07	2.852625e+07	2.852625e+07	2.852625e+07	28274123.0
2021-02-05	26547977.0	2.145566e+07	3.121222e+07	2.802065e+07	2.828575e+07	2.879795e+07	2.879795e+07	2.879795e+07	28520608.0
2021-02-06	26654965.0	2.155920e+07	3.161499e+07	2.824316e+07	2.849628e+07	2.907085e+07	2.907085e+07	2.907085e+07	28767093.0
2021-02-07	26746377.0	2.166274e+07	3.202281e+07	2.846566e+07	2.867216e+07	2.934495e+07	2.934495e+07	2.934495e+07	29013578.0
2021-02-08	26832826.0	2.176628e+07	3.243571e+07	2.868817e+07	2.884788e+07	2.962025e+07	2.962025e+07	2.962025e+07	29260063.0
2021-02-09	26923756.0	2.186982e+07	3.285376e+07	2.891067e+07	2.905310e+07	2.989675e+07	2.989675e+07	2.989675e+07	29506548.0
2021-02-10	27020890.0	2.197336e+07	3.327699e+07	2.913318e+07	2.928678e+07	3.017445e+07	3.017445e+07	3.017445e+07	29753033.0
2021-02-11	27122583.0	2.207690e+07	3.370548e+07	2.935568e+07	2.952539e+07	3.045335e+07	3.045335e+07	3.045335e+07	29999518.0
2021-02-12	27221607.0	2.218044e+07	3.413926e+07	2.957819e+07	2.982157e+07	3.073345e+07	3.073345e+07	3.073345e+07	30246003.0
2021-02-13	27309503.0	2.228398e+07	3.457839e+07	2.980069e+07	3.011369e+07	3.101476e+07	3.101476e+07	3.101476e+07	30492488.0
2021-02-14	27337816.0	2.238752e+07	3.502292e+07	3.002320e+07	3.030904e+07	3.129726e+07	3.129726e+07	3.129726e+07	30738973.0
2021-02-15	27433718.0	2.249106e+07	3.547290e+07	3.024570e+07	3.046948e+07	3.158096e+07	3.158096e+07	3.158096e+07	30985458.0
2021-02-16	27491574.0	2.259460e+07	3.592839e+07	3.046821e+07	3.069611e+07	3.186587e+07	3.186587e+07	3.186587e+07	31231943.0
2021-02-17	27560643.0	2.269815e+07	3.638944e+07	3.069071e+07	3.099691e+07	3.215197e+07	3.215197e+07	3.215197e+07	31478428.0
2021-02-18	27628834.0	2.280169e+07	3.685611e+07	3.091322e+07	3.124980e+07	3.243928e+07	3.243928e+07	3.243928e+07	31724913.0
2021-02-19	27702074.0	2.290523e+07	3.732845e+07	3.113572e+07	3.151836e+07	3.272779e+07	3.272779e+07	3.272779e+07	31971398.0
2021-02-20	27773047.0	2.300877e+07	3.780652e+07	3.135823e+07	3.172681e+07	3.301749e+07	3.301749e+07	3.301749e+07	32217883.0
2021-02-21	27828370.0	2.311231e+07	3.829036e+07	3.158073e+07	3.189630e+07	3.330840e+07	3.330840e+07	3.330840e+07	32464368.0
2021-02-22	27883560.0	2.321585e+07	3.878004e+07	3.180324e+07	3.206593e+07	3.360051e+07	3.360051e+07	3.360051e+07	32710853.0
2021-02-23	27955338.0	2.331939e+07	3.927562e+07	3.202574e+07	3.226843e+07	3.389382e+07	3.389382e+07	3.389382e+07	32957338.0
2021-02-24	28028815.0	2.342293e+07	3.977714e+07	3.224825e+07	3.250254e+07	3.418833e+07	3.418833e+07	3.418833e+07	33203823.0
2021-02-25	28102166.0	2.352647e+07	4.028467e+07	3.247075e+07	3.274212e+07	3.448404e+07	3.448404e+07	3.448404e+07	33450308.0
2021-02-26	28174978.0	2.363001e+07	4.079826e+07	3.269326e+07	3.304548e+07	3.478095e+07	3.478095e+07	3.478095e+07	33696793.0
2021-02-27	28244591.0	2.373355e+07	4.131798e+07	3.291577e+07	3.334424e+07	3.507906e+07	3.507906e+07	3.507906e+07	33943278.0
2021-02-28	28294809.0	2.383709e+07	4.184387e+07	3.313827e+07	3.353582e+07	3.537837e+07	3.537837e+07	3.537837e+07	34189763.0
2021-03-01	28345585.0	2.394063e+07	4.237600e+07	3.336078e+07	3.368886e+07	3.567888e+07	3.567888e+07	3.567888e+07	34436248.0

Table 4.20: First trial of prediction using wave 2 of All 8 models along with the valid confirmed cases value

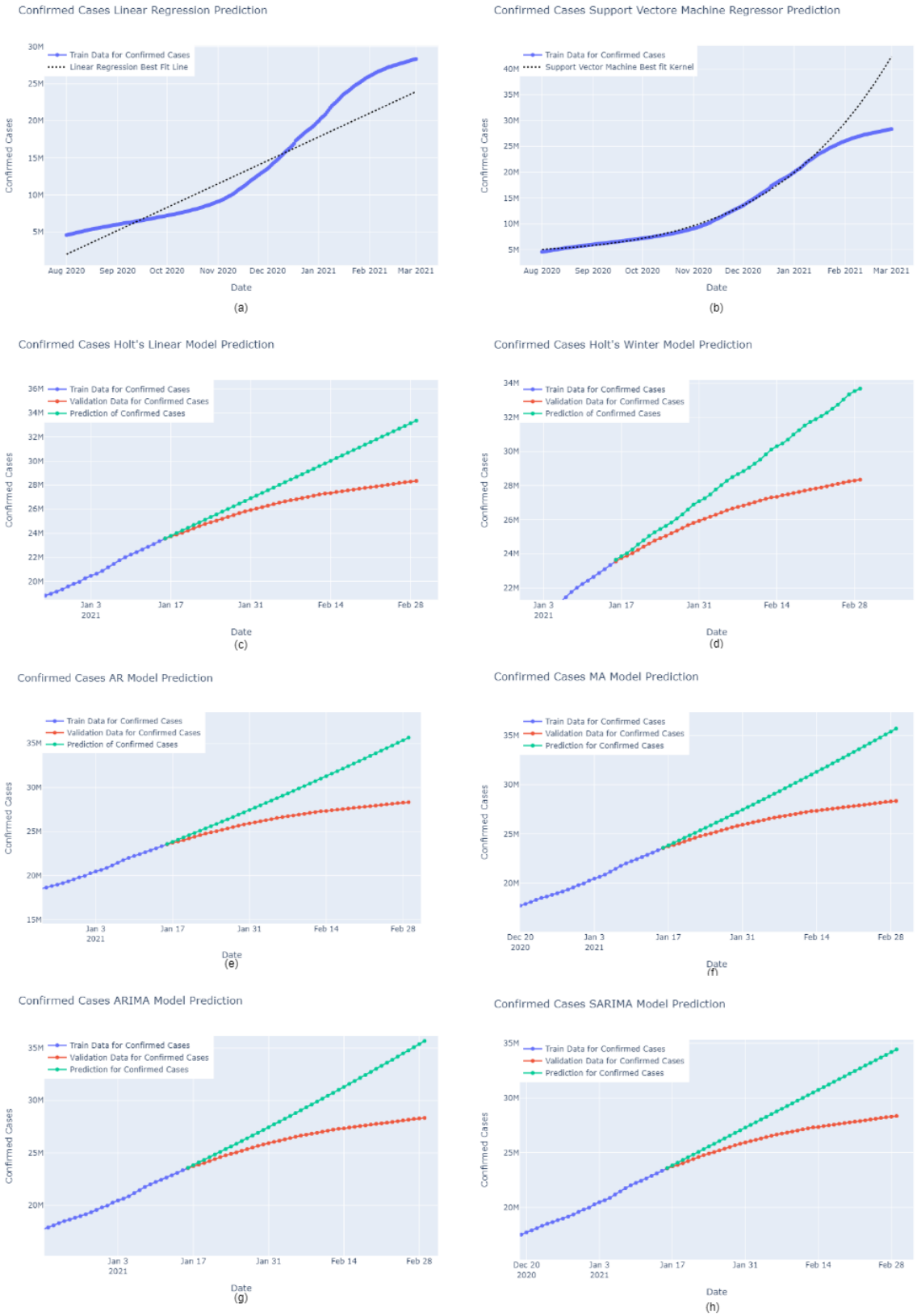


Figure 4.35: First trial of prediction from wave 1 of All 8 models

4.4.2 Second Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 2 (From August to end of January) to predict total confirmed cases of $\tilde{3}0$ days in Wave 2 (From the end of January to end of February). The results of this trial is in the figures below 4.37, and 4.36 and the predicted values of each model is also shown in the table below 4.22.

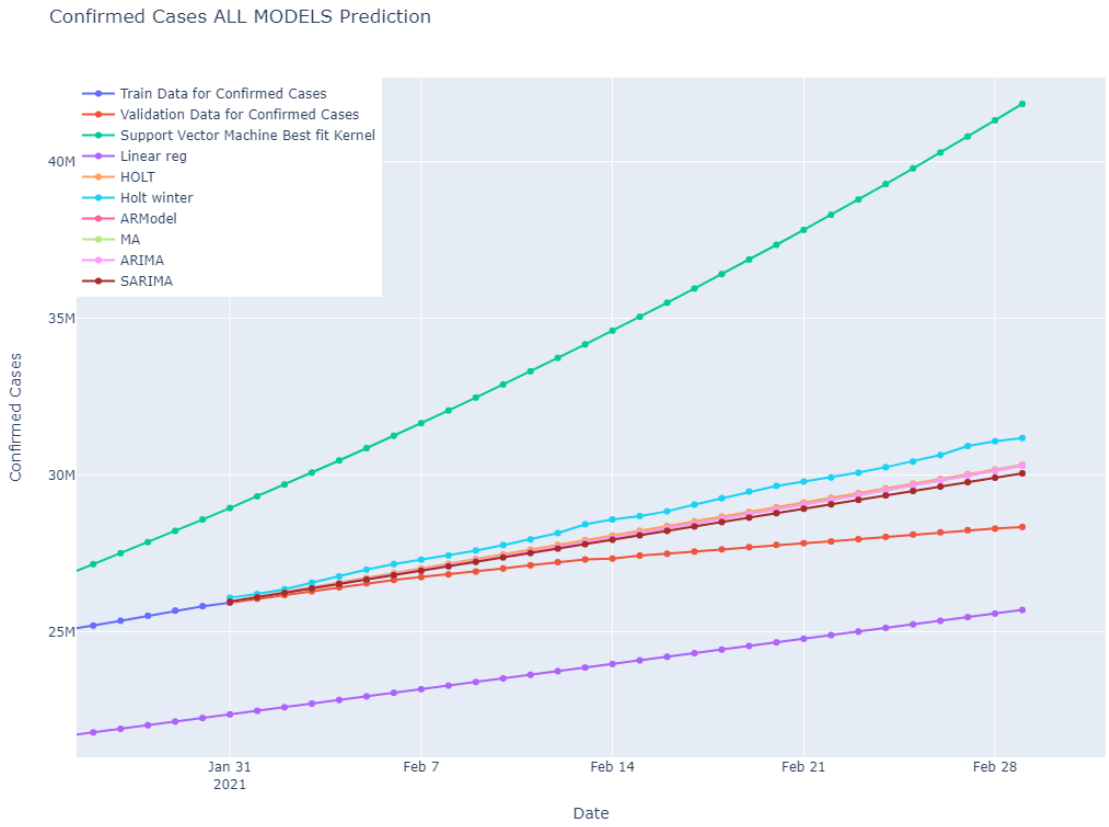


Figure 4.36: Second trial of prediction using wave 2 of All 8 models joined Zoomed in

Evaluation

In The Second trial , SARIMA model performed best as shown in table 4.23. SARIMA had the lowest RMSE with a value of 8.849428e+05 along with a low MAPE of 2.544200. Furthermore, in figure 4.36. SARIMA's plot was the closest to true validation total confirmed cases values. ARIMA, MA, and AR models performed equally achieving a low MAPE with very close values to one another along with lower RMSE than other Models. SVR performs worst in this situation.

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2021-01-31	25930068.0	2.236457e+07	2.895047e+07	2.596939e+07	2.608963e+07	2.595979e+07	2.595979e+07	2.595979e+07	25959266.0
2021-02-01	26055512.0	2.247951e+07	2.932276e+07	2.611979e+07	2.620563e+07	2.610215e+07	2.610215e+07	2.610215e+07	26100593.0
2021-02-02	26172274.0	2.259446e+07	2.969977e+07	2.627018e+07	2.635792e+07	2.624504e+07	2.624504e+07	2.624504e+07	26241920.0
2021-02-03	26293150.0	2.270940e+07	3.008155e+07	2.642058e+07	2.656972e+07	2.638845e+07	2.638845e+07	2.638845e+07	26383247.0
2021-02-04	26418016.0	2.282434e+07	3.046815e+07	2.657097e+07	2.677166e+07	2.653238e+07	2.653238e+07	2.653238e+07	26524574.0
2021-02-05	26547977.0	2.293928e+07	3.085960e+07	2.672136e+07	2.698317e+07	2.667682e+07	2.667682e+07	2.667682e+07	26665901.0
2021-02-06	26654965.0	2.305422e+07	3.125597e+07	2.687176e+07	2.719128e+07	2.682179e+07	2.682179e+07	2.682179e+07	26807228.0
2021-02-07	26746377.0	2.316917e+07	3.165730e+07	2.702215e+07	2.730233e+07	2.696728e+07	2.696728e+07	2.696728e+07	26948555.0
2021-02-08	26832826.0	2.328411e+07	3.206364e+07	2.717255e+07	2.743670e+07	2.711329e+07	2.711329e+07	2.711329e+07	27089882.0
2021-02-09	26923756.0	2.339905e+07	3.247504e+07	2.732294e+07	2.759256e+07	2.725982e+07	2.725982e+07	2.725982e+07	27231209.0
2021-02-10	27020890.0	2.351399e+07	3.289155e+07	2.747334e+07	2.776659e+07	2.740686e+07	2.740686e+07	2.740686e+07	27372536.0
2021-02-11	27122583.0	2.362893e+07	3.331322e+07	2.762373e+07	2.794839e+07	2.755443e+07	2.755443e+07	2.755443e+07	27513863.0
2021-02-12	27221607.0	2.374387e+07	3.374010e+07	2.777412e+07	2.815130e+07	2.770252e+07	2.770252e+07	2.770252e+07	27655190.0
2021-02-13	27309503.0	2.385882e+07	3.417224e+07	2.792452e+07	2.843139e+07	2.785113e+07	2.785113e+07	2.785113e+07	27796517.0
2021-02-14	27337816.0	2.397376e+07	3.460971e+07	2.807491e+07	2.858461e+07	2.800026e+07	2.800026e+07	2.800026e+07	27937844.0
2021-02-15	27433718.0	2.408870e+07	3.505254e+07	2.822531e+07	2.869469e+07	2.814991e+07	2.814991e+07	2.814991e+07	28079171.0
2021-02-16	27491574.0	2.420364e+07	3.550078e+07	2.837570e+07	2.884458e+07	2.830008e+07	2.830008e+07	2.830008e+07	28220498.0
2021-02-17	27560643.0	2.431858e+07	3.595451e+07	2.852609e+07	2.905958e+07	2.845077e+07	2.845077e+07	2.845077e+07	28361825.0
2021-02-18	27628834.0	2.443353e+07	3.641376e+07	2.867649e+07	2.926376e+07	2.860198e+07	2.860198e+07	2.860198e+07	28503152.0
2021-02-19	27702074.0	2.454847e+07	3.687858e+07	2.882688e+07	2.947837e+07	2.875371e+07	2.875371e+07	2.875371e+07	28644479.0
2021-02-20	27773047.0	2.466341e+07	3.734905e+07	2.897728e+07	2.965647e+07	2.890596e+07	2.890596e+07	2.890596e+07	28785806.0
2021-02-21	27828370.0	2.477835e+07	3.782520e+07	2.912767e+07	2.979413e+07	2.905873e+07	2.905873e+07	2.905873e+07	28927133.0
2021-02-22	27883560.0	2.489329e+07	3.830709e+07	2.927806e+07	2.992454e+07	2.921202e+07	2.921202e+07	2.921202e+07	29068460.0
2021-02-23	27955338.0	2.500824e+07	3.879479e+07	2.942846e+07	3.007843e+07	2.936583e+07	2.936583e+07	2.936583e+07	29209787.0
2021-02-24	28028815.0	2.512318e+07	3.928834e+07	2.957885e+07	3.025215e+07	2.952016e+07	2.952016e+07	2.952016e+07	29351114.0
2021-02-25	28102166.0	2.523812e+07	3.978780e+07	2.972925e+07	3.043433e+07	2.967501e+07	2.967501e+07	2.967501e+07	29492441.0
2021-02-26	28174978.0	2.535306e+07	4.029322e+07	2.987964e+07	3.063948e+07	2.983038e+07	2.983038e+07	2.983038e+07	29633768.0
2021-02-27	28244591.0	2.546800e+07	4.080467e+07	3.003003e+07	3.092855e+07	2.998627e+07	2.998627e+07	2.998627e+07	29775095.0
2021-02-28	28294809.0	2.558295e+07	4.132221e+07	3.018043e+07	3.107958e+07	3.014268e+07	3.014268e+07	3.014268e+07	29916422.0
2021-03-01	28345585.0	2.569789e+07	4.184588e+07	3.033082e+07	3.118375e+07	3.029961e+07	3.029961e+07	3.029961e+07	30057749.0

Table 4.22: Second trial of prediction using wave 2 of All 8 models along with the valid confirmed cases value

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
SARIMA	8.849428e+05	2.544200	7.831237e+11	7.073268e+05
AR	9.980357e+05	2.852833	9.960752e+11	7.933220e+05
MA	9.980357e+05	2.852833	9.960752e+11	7.933220e+05
ARIMA	9.980357e+05	2.852833	9.960752e+11	7.933220e+05
Holt's Linear	1.045021e+06	3.055810	1.092069e+12	8.489276e+05
Holt's Winter	1.568612e+06	4.763772	2.460543e+12	1.320995e+06
LR	3.284464e+06	12.012176	1.078771e+13	3.269951e+06
SVR	8.350712e+06	28.405116	6.973439e+13	7.731468e+06

Table 4.23: Evaluation for trial 2 in prediction using wave 2

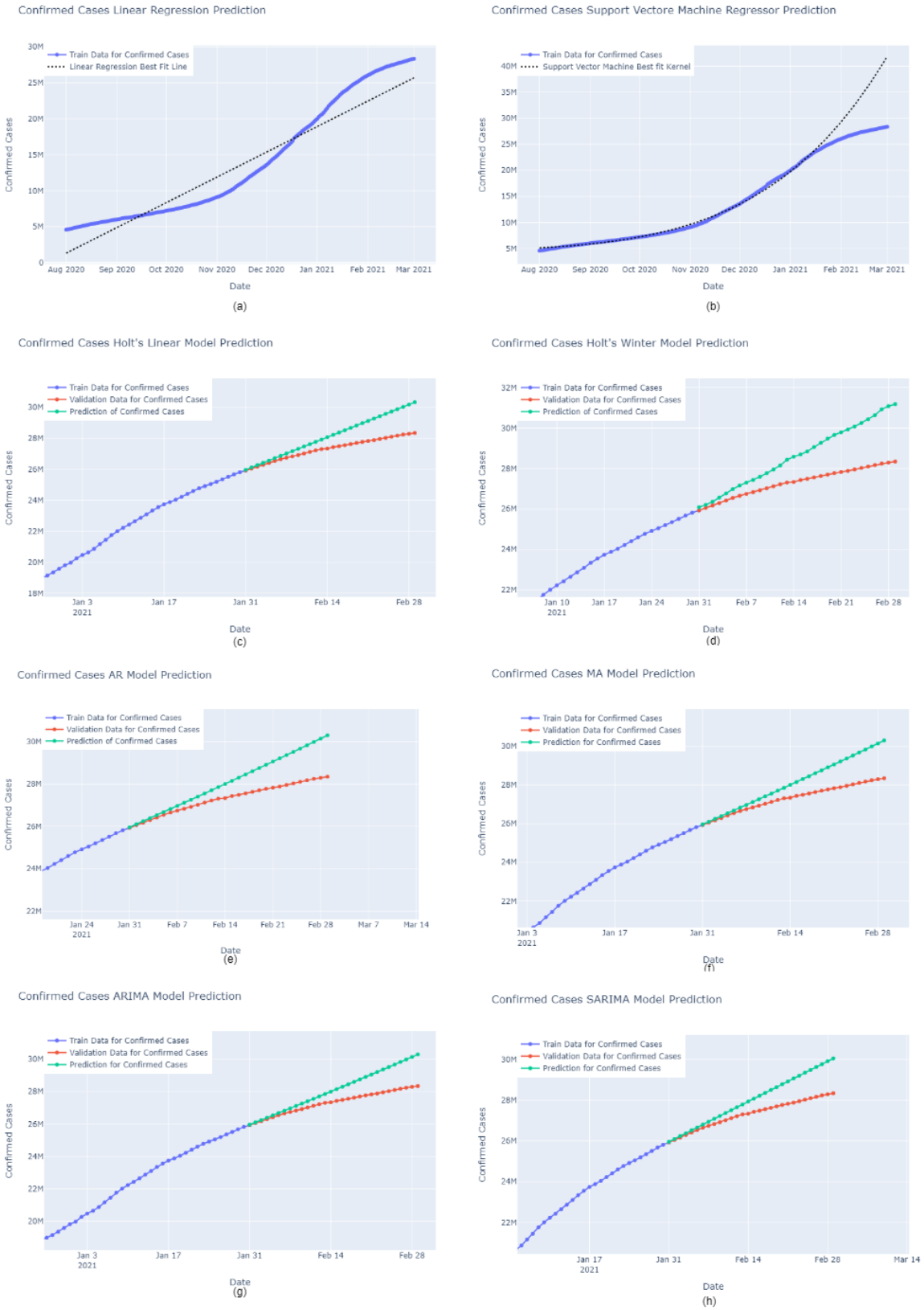


Figure 4.37: Second trial of prediction from wave 1 of All 8 models

4.4.3 Third Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 2 (From August to mid of February) to predict total confirmed cases of $\tilde{1}5$ days in Wave 2 (From the mid of February to end of February). The results of this trial is in the figures below 4.39, and 4.38 and the predicted values of each model is also shown in the table below 4.24.

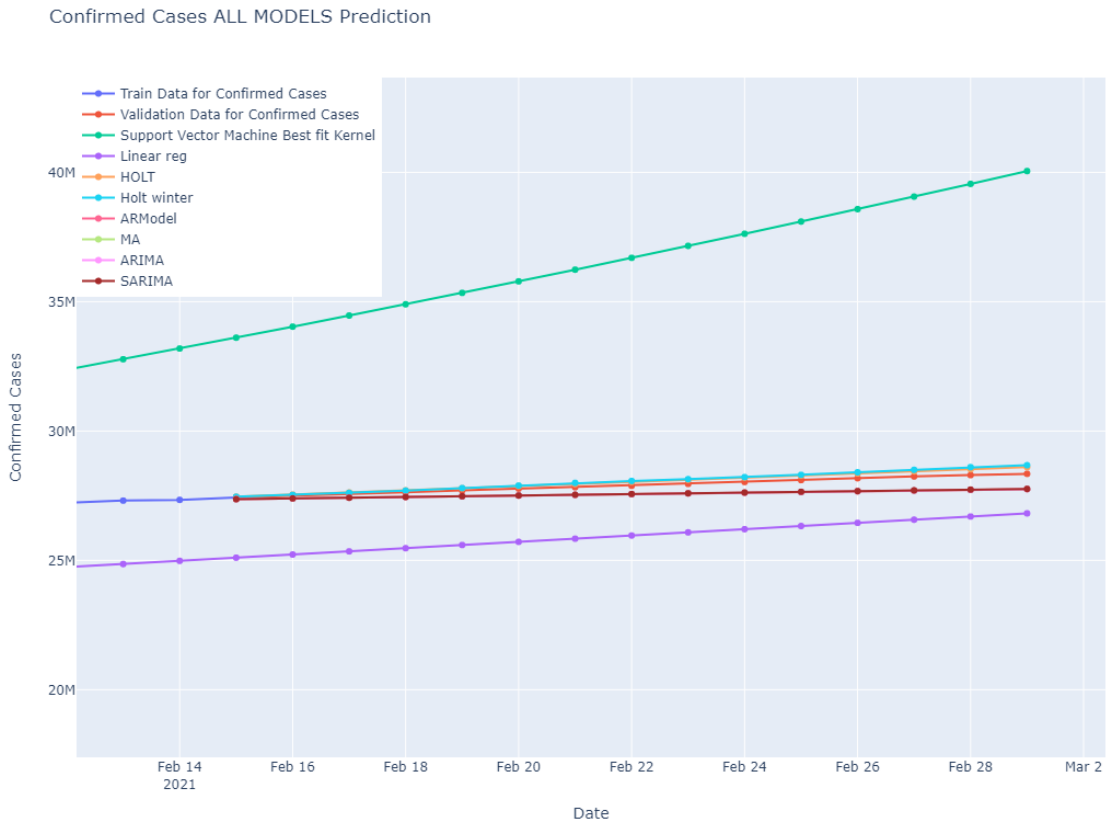


Figure 4.38: Third trial of prediction using wave 2 of All 8 models joined Zoomed in

Evaluation

In The Third trial , Holt's Linear model performed best as shown in table 4.25. Holt's Linear had the lowest RMSE with a value of 1.494727e+05 along with a low MAPE of0.478986. Furthermore, in figure 4.38. Holt's Linear plot was the closest to true validation total confirmed cases values. Holt's Winter also performed relatively close to Holt's Linear with s very close RMSE and MAPE. ARIMA, MA, and AR models performed equally achieving a low MAPE with very close values to one another along with lower RMSE than other Models. SVR performs worst in this situation.

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2021-02-15	27433718.0	2.510275e+07	3.360942e+07	2.745920e+07	2.745973e+07	2.736603e+07	2.736603e+07	2.736603e+07	27366129.0
2021-02-16	27491574.0	2.522519e+07	3.403439e+07	2.754086e+07	2.753096e+07	2.739415e+07	2.739415e+07	2.739415e+07	27394442.0
2021-02-17	27560643.0	2.534764e+07	3.446454e+07	2.762252e+07	2.760842e+07	2.742218e+07	2.742218e+07	2.742218e+07	27422755.0
2021-02-18	27628834.0	2.547008e+07	3.489994e+07	2.770418e+07	2.769613e+07	2.745011e+07	2.745011e+07	2.745011e+07	27451068.0
2021-02-19	27702074.0	2.559252e+07	3.534063e+07	2.778584e+07	2.779537e+07	2.747794e+07	2.747794e+07	2.747794e+07	27479381.0
2021-02-20	27773047.0	2.571496e+07	3.578665e+07	2.786750e+07	2.789020e+07	2.750567e+07	2.750567e+07	2.750567e+07	27507694.0
2021-02-21	27828370.0	2.583740e+07	3.623808e+07	2.794916e+07	2.799438e+07	2.753331e+07	2.753331e+07	2.753331e+07	27536007.0
2021-02-22	27883560.0	2.595984e+07	3.669494e+07	2.803082e+07	2.806897e+07	2.756085e+07	2.756085e+07	2.756085e+07	27564320.0
2021-02-23	27955338.0	2.608228e+07	3.715731e+07	2.811248e+07	2.813985e+07	2.758830e+07	2.758830e+07	2.758830e+07	27592633.0
2021-02-24	28028815.0	2.620472e+07	3.762522e+07	2.819414e+07	2.821710e+07	2.761565e+07	2.761565e+07	2.761565e+07	27620946.0
2021-02-25	28102166.0	2.632716e+07	3.809874e+07	2.827579e+07	2.830482e+07	2.764290e+07	2.764290e+07	2.764290e+07	27649259.0
2021-02-26	28174978.0	2.644960e+07	3.857792e+07	2.835745e+07	2.840433e+07	2.767006e+07	2.767006e+07	2.767006e+07	27677572.0
2021-02-27	28244591.0	2.657204e+07	3.906280e+07	2.843911e+07	2.849934e+07	2.769712e+07	2.769712e+07	2.769712e+07	27705885.0
2021-02-28	28294809.0	2.669448e+07	3.955346e+07	2.852077e+07	2.860389e+07	2.772409e+07	2.772409e+07	2.772409e+07	27734198.0
2021-03-01	28345585.0	2.681692e+07	4.004993e+07	2.860243e+07	2.867821e+07	2.775096e+07	2.775096e+07	2.775096e+07	27762511.0

Table 4.24: Third trial of prediction using wave 2 of All 8 models along with the valid confirmed cases value

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
Holt's Linear	1.494727e+05	0.478986	2.234210e+10	1.342752e+05
Holt's Winter	1.875530e+05	0.580656	3.517612e+10	1.629072e+05
SARIMA	3.712075e+05	1.184902	1.377950e+11	3.322201e+05
AR	3.766890e+05	1.200419	1.418946e+11	3.365846e+05
MA	3.766890e+05	1.200419	1.418946e+11	3.365846e+05
ARIMA	3.766890e+05	1.200419	1.418946e+11	3.365846e+05
LR	1.951694e+06	6.952095	3.809109e+12	1.936701e+06
SVR	9.011518e+06	31.737402	8.120746e+13	8.849724e+06

Table 4.25: Evaluation for trial 3 in prediction using wave 2

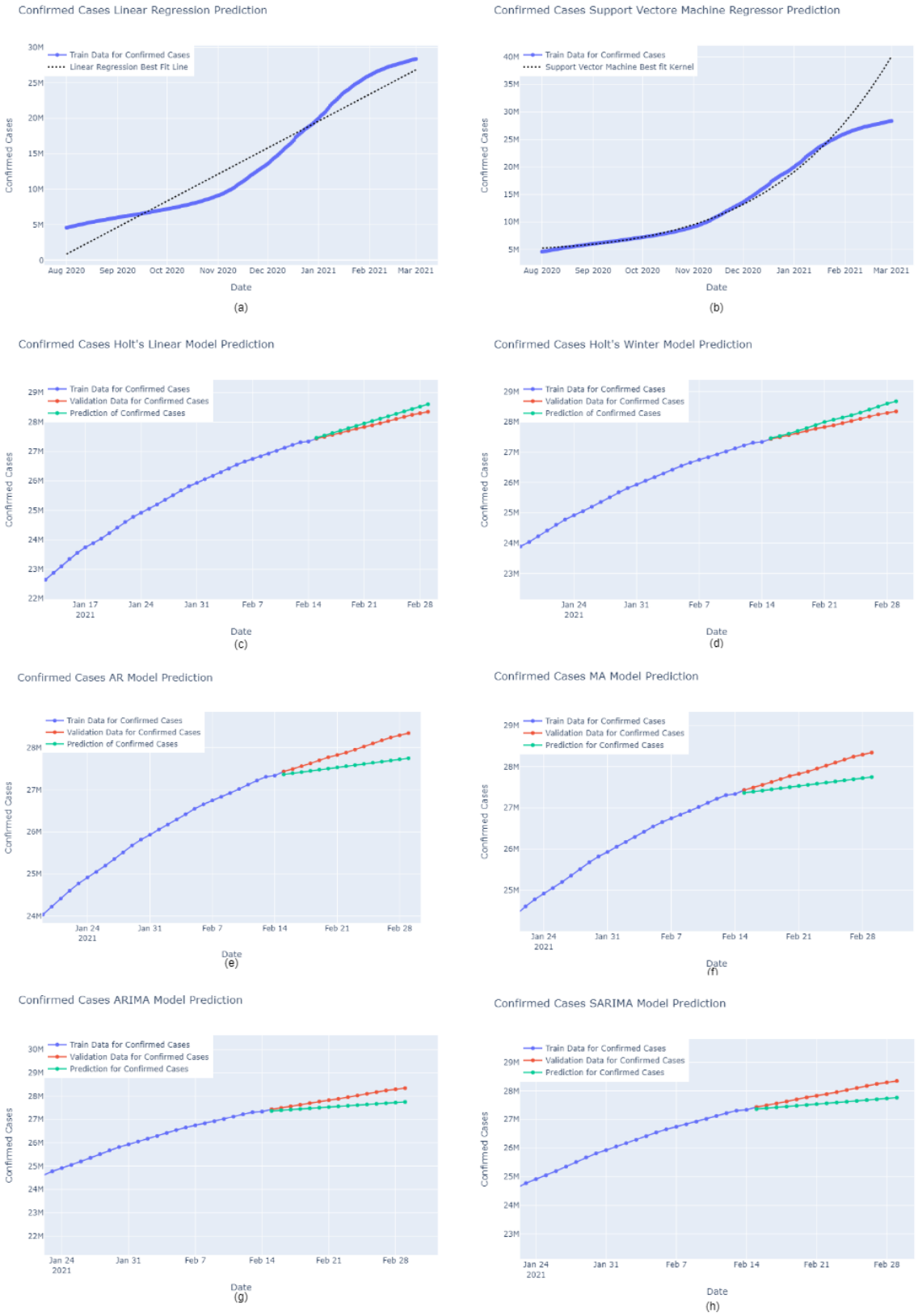


Figure 4.39: Third trial of prediction from wave 1 of All 8 models

4.4.4 Forth Trial

In this trial, the 8 models mentioned above in 3.5.2 were trained using wave 2 (From August to mid of February) to predict total confirmed cases of $\tilde{1}0$ days in Wave 2 (last 10 days in February). The results of this trial is in the figures below 4.41, and 4.40 and the predicted values of each model is also shown in the table below 4.26.

Figure 4.40: Forth trial of prediction using wave 2 of All 8 models joined Zoomed in

Evaluation

In The Forth trial , Holt's Linear model performed best as shown in table 4.27. Holt's Linear had the lowest RMSE with a value of1.037184e+04 along with a low MAPE of 1.075751e+08. Furthermore, in figure 4.40. Holt's Linear plot was the closest to true validation total confirmed cases values. SARIMA also performed relatively close to Holt's Linear with s very close RMSE and MAPE. ARIMA, MA, and AR models performed equally achieving a low MAPE with very close values to one another along with lower RMSE than other Models. SVR performs worst in this situation.

date	total_confirmed	LinearPredict	predictSVM	Holt	Holt's Winter Model	predictAR	predictMA	predictARIMA	predictSARIMA
2021-02-20	27773047.0	2.592993e+07	3.517643e+07	2.776847e+07	2.780220e+07	2.777544e+07	2.777544e+07	2.777544e+07	27775314.0
2021-02-21	27828370.0	2.605395e+07	3.561811e+07	2.783414e+07	2.785369e+07	2.784894e+07	2.784894e+07	2.784894e+07	27848554.0
2021-02-22	27883560.0	2.617796e+07	3.606511e+07	2.789981e+07	2.794803e+07	2.792257e+07	2.792257e+07	2.792257e+07	27921794.0
2021-02-23	27955338.0	2.630197e+07	3.651750e+07	2.796548e+07	2.800682e+07	2.799633e+07	2.799633e+07	2.799633e+07	27995034.0
2021-02-24	28028815.0	2.642599e+07	3.697531e+07	2.803115e+07	2.807508e+07	2.807022e+07	2.807022e+07	2.807022e+07	28068274.0
2021-02-25	28102166.0	2.655000e+07	3.743860e+07	2.809682e+07	2.814786e+07	2.814424e+07	2.814424e+07	2.814424e+07	28141514.0
2021-02-26	28174978.0	2.667401e+07	3.790744e+07	2.816249e+07	2.824695e+07	2.821838e+07	2.821838e+07	2.821838e+07	28214754.0
2021-02-27	28244591.0	2.679803e+07	3.838185e+07	2.822816e+07	2.833266e+07	2.829266e+07	2.829266e+07	2.829266e+07	28287994.0
2021-02-28	28294809.0	2.692204e+07	3.886191e+07	2.829383e+07	2.838369e+07	2.836707e+07	2.836707e+07	2.836707e+07	28361234.0
2021-03-01	28345585.0	2.704605e+07	3.934767e+07	2.835949e+07	2.847837e+07	2.844160e+07	2.844160e+07	2.844160e+07	28434474.0

Table 4.26: Forth trial of prediction using wave 2 of All 8 models along with the valid confirmed cases value

Evaluation				
Model Name	RMSE	MAPE	MSE	MAE
Holt's Linear	1.037184e+04	0.031419	1.075751e+08	8.823772e+03
SARIMA	4.725982e+04	0.148380	2.233491e+09	4.176810e+04
AR	5.074565e+04	0.158490	2.575121e+09	4.462011e+04
MA	5.074565e+04	0.158490	2.575121e+09	4.462011e+04
ARIMA	5.074565e+04	0.158490	2.575121e+09	4.462011e+04
Holt Winter	7.140892e+04	0.228878	5.099233e+09	6.440820e+04
LR	1.583835e+06	5.617074	2.508533e+12	1.575132e+06
SVR	9.236567e+06	32.667752	8.531417e+13	9.165868e+06

Table 4.27: Evaluation for trial 4 in prediction using wave 2

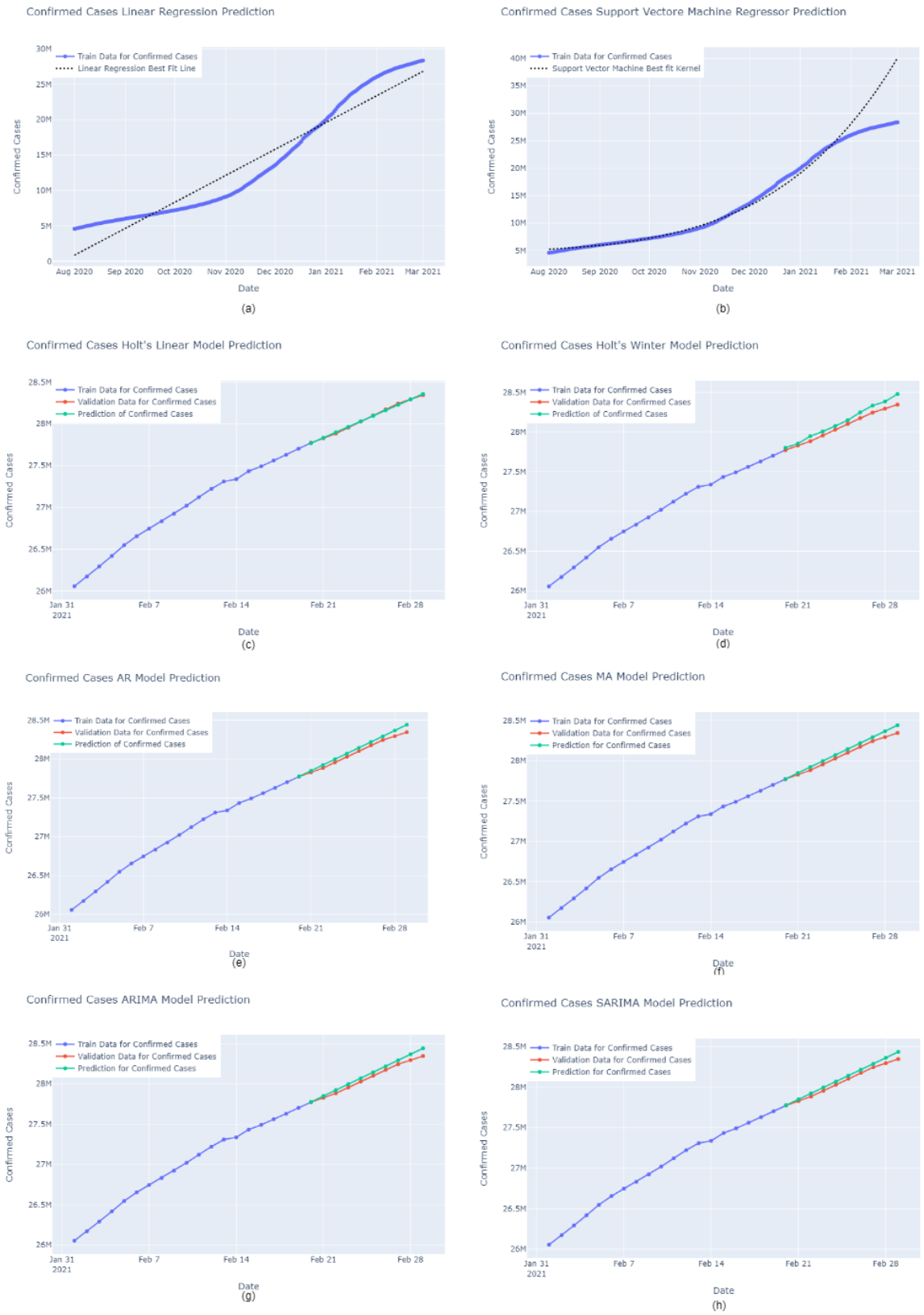


Figure 4.41: Forth trial of prediction from wave 1 of All 8 models

4.4.5 Discussion

As Noticed in section 4.3.5, The less days to predict the total confirmed cases, the better the performance of the learning Models and gives a better result of the reality of the growing pandemic. Holt's Linear Models performs best for both Trial 1 and Trial 4. In this prediction , AR, AM, and ARIMA performed equally.

SVR Model performed the worst among all models in all trials, it isn't providing great results , the predictions are either overshooting or really lower than what's expected as shown in the above figures 4.33 (b)

Chapter 5

Conclusion

In this paper, an analysis of the spatial evolution of coronavirus pandemic around the USA and New York state by K-means was presented. Based on k-means,we were able to spatially group together counties/states that are similar according to their coronavirus cases, in this way being able to analyze which counties/states are behaving similarly and thus can benefit by using similar strategies in dealing with the spread of the virus. In this part of the study, 1) features were studied as provided from GoogleCloudPlatform 2) then those features were studied with reference to population and finally 3) these features were studied with reference to the population density. Interesting conclusions have been obtained, that could be helpful in deciding the best strategies in dealing with this virus, which is the relation between population density, mobility, age groups and the number of COVID-19 cases, tests and deceased. In most cases, regions with greater population density have greater rates of transmission of COVID-19, likely due to increased contact rates in areas with greater density.

In addition, in this study a temporal analysis was provided for predicting COVID-19 total confirmed cases in The USA using machine learning algorithms to help in the risk of COVID-19 outbreak. The results of the this part of the study prove that the less the days to predict in the future the better. In predicting the total confirmed cases from wave 1, SARIMA performs best in the in predicting the last 93 days(from August to the first of November) and in the attempt of predicting the last 10 days in October than other models. Also ARIMA performed best in predicting 30 days in the month of October. MA performance was relatively good as well in predicting the last 15 days in October. SVR produces poor results in all scenarios because of the ups and downs in the dataset values. It was very difficult to put an accurate hyperplane between the given values of the dataset.

In predicting the total confirmed cases later into wave 2 by using the data from the beginning of wave 2 as a training data, Holt's Linear Model performs best in the in predicting the last 45 days(from mid January to the end of February). MA, AR and ARIMA's performance were exactly equal in this study and the results of the model preformed relatively good. It was noticed that SVR produces the poorest results in all scenarios.

Overall we conclude that both spatial and temporal analysis, may be helpful to understand the upcoming situation. The study analysis thus can also be of great guidance for the healthcare providers, authorities, and governments to take timely actions and make decisions to contain the COVID-19 crisis.

This study faced some limitations. First, information about recovered cases, Vaccinations and information about healthcare concerning COVID-19 wasn't provided for the USA in the GoogleCloudPlatform open source; these factors could have impacted my study and given a closer look to reality. I have even contacted *Integrated Public Use Microdata Series(IPUMS)*; which is the world's largest individual-level population database. However, data about New York state counties weren't available. Despite of the limitations mentioned, the GoogleCloudPlatform provided other features like the mobility, and age groups which also helped in the analysis.

5.1 Future Works

As a recommendation for future works of spatial analysis, I propose trying to cluster data of the globe to get a better sense of how different countries will combat the COVID-19 crises. In this analysis, I will also suggest studying the pandemic through a different time interval or wave. As during other waves, features like vaccination, restriction on internal travel would have appeared and would definitely impact the study.

As a recommendation for the prediction future work, I suggest predicting other variables as deaths rate and recovery rate (if details about recovery were provided). I will also propose predicting and forecasting total cases using multivariate time series, where each observation at a time t is a vector of values instead of a single value (total cases in this study). Multivariate could give us a closer look on reality as it will help us study the other factors affecting the confirmed cases, and deaths rate.

Appendix

Appendix A

Lists

List of Figures

2.1 Machine Learning classifications 4

2.2 home made sketch by the author Serafeim Loukas to how the PCA may be used to minimize data dimensionality and maximise variance. Source : [11] 6

2.3 Support vector regression model for linear regression fitting where X1= X and X2 = y are the features and label in our case. [Image credit: Source] 8

2.4 Two examples of data from autoregressive models with different parameters.Source [13] 10

2.5 K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid Source :Source 12

2.6 Elbow Method for selection of optimal K. Source :Source 13

3.1 Workflow 21

3.2 Sample of Data Visualization of NY state using PowerBI 28

4.1 NY attempt1.(a)Elbow Method (b)DBI results (c)Silhouette. All suggesting k=2 as an optimal number of k 32

4.2 PCA of NY first attempt with n components =2 after applying k-means with k =2 33

4.3 NY first attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA all suggesting k=2 as an optimal number of k 33

4.4 NY's first trial where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the first attempt. The orange closer are for cluster 0, and the red color is for cluster 1. The radius of the circle represents the total confirmed cases in each cluster in order to give a better visualization of each cluster 35

4.5 NY's trial 1 where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased 36

4.6 NY's first trial showing the distance between the geographic location of the center of cluster 0 and all counties of cluster 1 on the x-axis, and the total number of cases on the y-axis along the 4 months of wave 1 where (a) is this first month , and (d) is forth month.It shows that the closer the distance from center of cluster 0, the greater the number of the total cases 36

4.7 NY attempt 1's mean Values of some features between both cluster 0 and 1. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk. 38

4.8 NY's second attempt where (a)Elbow Method (b)DBI results (c)Silhouettes before applying PCA 39

4.9 NY's second attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA 40

4.10 NY's second attempt applying PCA with n components =2 after applying kmeans with k =2 41

4.11 NY's second attempt applying PCA with n components =2 after applying kmeans with k =3 41

4.12 NY's second attempt with (k=2) where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the second attempt. The Orange closer are for cluster 0, and the red color is for cluster 1. The radius of the circle represents the total confirmed cases percentage in each cluster in order to give a better visualization of each cluster 43

4.13 NY's second attempt with (k=2) where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased 44

4.14 NY state second trial(k=3) where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the second attempt(k=3). The Orange closer are for cluster 0, and the green color is for cluster 1 and red is for cluster 2. The radius of the circle represents the total confirmed cases percent 46

4.15 NY state trial 2 (k=3) where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased 47

4.16 NY's second attempt with k=2 showing the mean Values of some features, between both cluster 0 and 1. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk. 49

4.17 NY's second attempt with k=3. Figure shows the mean values of the features between both cluster 0, 1, 2. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk. 51

4.18 USA's first attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA 52

4.19 USA's First trial where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the first attempt. The orange closer are for cluster 1, and the green color is for cluster 0. The radius of the circle represents the total confirmed cases in each cluster in order to give a better visualization of each cluster 54

4.20 USA's trial 1 where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases along with the to total deceased 55

4.21 Mean Values of some features between both cluster 0 and 1. where the highlighted cells in red indicate a reason behind the chances of the counties in this being at a higher risk. 56

4.22 USA's second attempt where (a)Elbow Method (b)DBI results (c)Silhouettes after applying PCA 57

4.23 USA's second trial where (a) is the first month and (d) is the forth month of wave 1. These maps show the different clustering groups during the first attempt. The orange closer are for cluster 1, and the green color is for cluster 1. The radius of the circle represents the total confirmed cases percentage in each cluster in order to give a better visualization of each cluster 59

4.24 USA's second trial where (a)is the first month and (d) is the forth month of wave 1. This figure gives a better visualization of each cluster with reference to both the total confirmed cases percentage along with the to total deceased percentage 60

4.25 Distance between center of cluster 0 from USA's second attempt and states of cluster 1 were calculated and on the y-axis is the total confirmed cases %. (a) is the first month in wave 1 and (d) being the forth. 61

4.26 First trial of prediction from wave 1 of All 8 models joined Zoomed in . . 63

4.27 First trial of prediction from wave 1 of All 8 models 65

4.28 Second trial of prediction from wave 1 of All 8 models joined Zoomed in . 66

4.29 Second trial of prediction from wave 1 of All 8 models 68

4.30 Third trial of prediction from wave 1 of All 8 models joined Zoomed in . 69

4.31 Third trial of prediction from wave 1 of All 8 models 71

4.32 Forth trial of prediction from wave 1 of All 8 models joined Zoomed in . 72
4.33 Forth trial of prediction from wave 1 of All 8 models 74
4.34 First trial of prediction using wave 2 of All 8 models joined Zoomed in . 76
4.35 First trial of prediction from wave 1 of All 8 models 78
4.36 Second trial of prediction using wave 2 of All 8 models joined Zoomed in 79
4.37 Second trial of prediction from wave 1 of All 8 models 81
4.38 Third trial of prediction using wave 2 of All 8 models joined Zoomed in . 82
4.39 Third trial of prediction from wave 1 of All 8 models 84
4.40 Forth trial of prediction using wave 2 of All 8 models joined Zoomed in . 85
4.41 Forth trial of prediction from wave 1 of All 8 models 87

Bibliography

[1] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo, "Analysis of spatial spread relationships of coronavirus (covid-19) pandemic in the world using self organizing maps," *Chaos, Solitons & Fractals*, vol. 138, p. 109917, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077920303179

[2] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "Covid-19 infection: Emergence, transmission, and characteristics of human coronaviruses," *Journal of Advanced Research*, vol. 24, pp. 91–98, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2090123220300540

[3] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, and R. Agha, "World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19)," *International Journal of Surgery*, vol. 76, pp. 71–76, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1743919120301977

[4] R. S. Yadav, "Data analysis of covid-2019 epidemic using machine learning methods: a case study of india," *International Journal of Information Technology*, vol. 12, pp. 1321–1330, 2020.

[5] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B.-W. On, W. Aslam, and G. S. Choi, "Covid-19 future forecasting using supervised machine learning models," *IEEE Access*, vol. 8, pp. 101 489–101 499, 2020.

[6] P. Lapuerta, S. P. Azen, and L. Labree, "Use of neural networks in predicting the risk of coronary artery disease," *Computers and Biomedical Research*, vol. 28, no. 1, pp. 38–52, 1995.

[7] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," *American heart journal*, vol. 121, no. 1, pp. 293–298, 1991.

[8] H.-L. Hwa, W.-H. Kuo, L.-Y. Chang, M.-Y. Wang, T.-H. Tung, K.-J. Chang, and F.-J. Hsieh, "Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models," *Journal of evaluation in clinical practice*, vol. 14, no. 2, pp. 275–280, 2008.

[9] S. Kushwaha, S. Bahl, A. K. Bagha, K. S. Parmar, M. Javaid, A. Haleem, and R. P. Singh, "Significant applications of machine learning for covid-19 pandemic," *Journal of Industrial Integration and Management*, vol. 05, no. 04, pp. 453–479, 2020. [Online]. Available: https://doi.org/10.1142/S2424862220500268

[10] [Online]. Available: https://towardsdatascience.com/dimensionality-reduction-for-machine-learning80a46c2ebb7e.

[11] "Pca clearly explained—when, why, how to use it." [Online]. Available: https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e

[12] D. Parbat and M. Chakraborty, "A python based support vector regression model for prediction of covid19 cases in india," *Chaos, Solitons & Fractals*, vol. 138, p. 109942, 2020.

[13] [Online]. Available: https://otexts.com/fpp2/holt.html

[14] "K-means clustering." [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[15] [Online]. Available: https://www.geeksforgeeks.org/dunn-index-and-db-index-cluster-validity-indices-set-1/

[16] [Online]. Available: https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

[17] J. Cordes and M. C. Castro, "Spatial analysis of covid-19 clusters and contextual factors in new york city," *Spatial and Spatio-temporal Epidemiology*, vol. 34, p. 100355, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877584520300332

[18] K. Mueller and E. Papenhausen, "Using demographic pattern analysis to predict covid-19 fatalities on the us county level," *Digit. Gov.: Res. Pract.*, vol. 2, no. 1, Dec. 2020. [Online]. Available: https://doi.org/10.1145/3430196

[19] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD 29*, vol. 2, 2000.

[20] B. Y. e. a. R.Pung, C.J. Chiew, "Investigation of three clusters of covid-19 in singapore:implications for surveillance and response measures," *Chaos, Solitons & Fractals*, vol. 395, p. 10229, 2020.

[21] Z. Malki, E.-S. Atlam, A. Ewis, G. Dagnew, O. A. Ghoneim, A. A. Mohamed, M. M. Abdel-Daim, and I. Gad, "The covid-19 pandemic: prediction study based on machine learning models," *Environmental Science and Pollution Research*, 2021. [Online]. Available: https://doi.org/10.1007/s11356-021-13824-7

[22] H. Tandon, P. Ranjan, T. Chakraborty, and V. Suhag, "Coronavirus (covid-19): Arima based time-series analysis to forecast near future," 2020.

[23] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time series prediction of covid-19 transmission in america using lstm and xgboost algorithms," *Results in Physics*, vol. 27, p. 104462, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211379721005775

[24] M. Maleki, M. R. Mahmoudi, D. Wraith, and K.-H. Pho, "Time series modelling to forecast the confirmed and recovered cases of covid-19," *Travel medicine and infectious disease*, vol. 37, p. 101742, 2020.

[25] K. ArunKumar, D. V. Kalaga, C. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative covid-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (arima) and seasonal auto-regressive integrated moving average (sarima)," *Applied Soft Computing*, vol. 103, p. 107161, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494621000843

[26] "Googlecloudplatform," 2020-2021. [Online]. Available: https://github.com/GoogleCloudPlatform/covid-19-open-data#understand-the-data

[27] "Percentage of adults with diagnosed diabetes, by county, new york state," 2018. [Online]. Available: https://www.health.ny.gov/statistics/prevention/injury_prevention/information_for_action/docs/2021-01_ifa_report.pdf

[28] "Front end for covid-19 anaylsis in new york state." [Online]. Available: https://app.powerbi.com/links/HL3Juyqqjy?ctid=271e6487-2832-46aa-ab86-c6cd77140512&pbi_source=linkShare

[29] "Front end for covid-19 anaylsis in us and canada." [Online]. Available: https://app.powerbi.com/links/oisOzZPUan?ctid=271e6487-2832-46aa-ab86-c6cd77140512&pbi_source=linkShare

[30] [Online]. Available: http://rasbt.github.io/mlxtend/user_guide/preprocessing/minmax_scaling/

[31] Using machine learning to predict the growth of covid-19. [Online]. Available: https://towardsdatascience.com/using-machine-learning-to-model-the-growth-of-covid-19-2f3b0af304bb

[32] K. C. Santosh, "Ai-driven tools for coronavirus outbreak: Need of active learning and cross-population train/test models on multitudinal/multimodal data," *Journal of Medical Systems*, vol. 44, no. 7, 2020. [Online]. Available: https://doi.org/10.1007/s10916-020-01562-1

[33] A. S. R. Srinivasa Rao and J. A. Vazquez, "Identification of covid-19 can be quicker through artificial intelligence framework using a mobile phone–based survey when cities and towns are under quarantine," *Infection Control ; Hospital Epidemiology*, vol. 41, no. 7, p. 826–830, 2020.

[34] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos Santos Coelho, "Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil," *Chaos, Solitons & Fractals*, vol. 135, p. 109853, 2020.

[35] The humanitarian data exchange (hdx). [Online]. Available: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases