COVID-19 in USA Statistical Report

1. Abstract

Since March 2020, the Corona Virus 2019 began to ravage the entire United States. Every day, many people get infected and die from diseases caused by the Corona Virus. In this report, we utilize the dataset of COVID-19 in USA to analyze the transmission trend, severity, and other factors of the Corona Virus. In addition, the linear regression model is applied to explore and predict the number of deaths by using several variables as regressors. Besides, we use four variables in the dataset and hierarchical clustering method, which is one of the techniques in unsupervised learning to divide the 50 US states into 5 clusters. The dataset utilized in the report is from Kaggle.

2. Introduction

Corona viruses are a large family of viruses which may cause illness in animals or humans. In humans, several corona viruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered corona virus causes corona virus disease COVID-19 [1].

In this paper, we will use the dataset COVID-19 in USA, which is downloaded from Kaggle [2], to analyze the transmission trend and other factors of the COVID-19 viruses. We will also analyze how severe the situation is in the whole USA and different states. The methods utilized in this report are descriptive statistics as well as different kinds of charts. In addition, the linear regression model is also used to explore the relationship among the response variable *death* and other regressors variables. Once the linear regression equation is built, we can predict the response variable *death* given other regressors variables. Besides, at the end of this report, we utilize four variables *death, positive, positiveIncrease,* and *deathIncrease* of 50 US states to calculate the Euclidean distance among different states. Then we use hierarchical clustering method, which is one of the techniques in unsupervised learning to divide the 50 US states into 5 clusters by applying the Euclidean distance matrix.

3. Methods & Results

The COVID-19 dataset in USA has 2769 cases and 25 variables in total. Each case represents a set of observation data about COVID-19 Virus epidemic in a US state in one certain day. At the time of writing this report, the data have been updated to April 23, 2020. The

variables that will be utilized in this report and their meanings are shown as below:

(1) *date* - date of observation

(2) *state* – full name of 50 US states

(3) *positive* - number of tests with positive results

(4) *negative* - number of tests with negative results

(5) *death* - number of deaths

(6) *deathIncrease* - daily increase of death

(7) positiveIncrease - daily increase of positive

(8) totalTestResults - total number of tests

(9) hospitalizedCumulative - number of people who has been hospitalized

(10) recovered - number of people who are recovered

First, we hope to analyze how the number of people with positive results, the number of death and the number of people who is recovered change over time. As is shown in Figure1, we can observe that how fast is the COVID-19 Virus epidemic spread in the US. The green line represents the cumulative number of people who are tested to be positive. And the purple line represents the cumulative number of people who are dead. The blue line represents the cumulative number of people who are recovered. It can be observed that the cumulative number of people who are positive increases very fast, and is close to exponential growth. The cumulative number of deaths and the cumulative number of recoveries have risen at a similar rate, both very slowly.

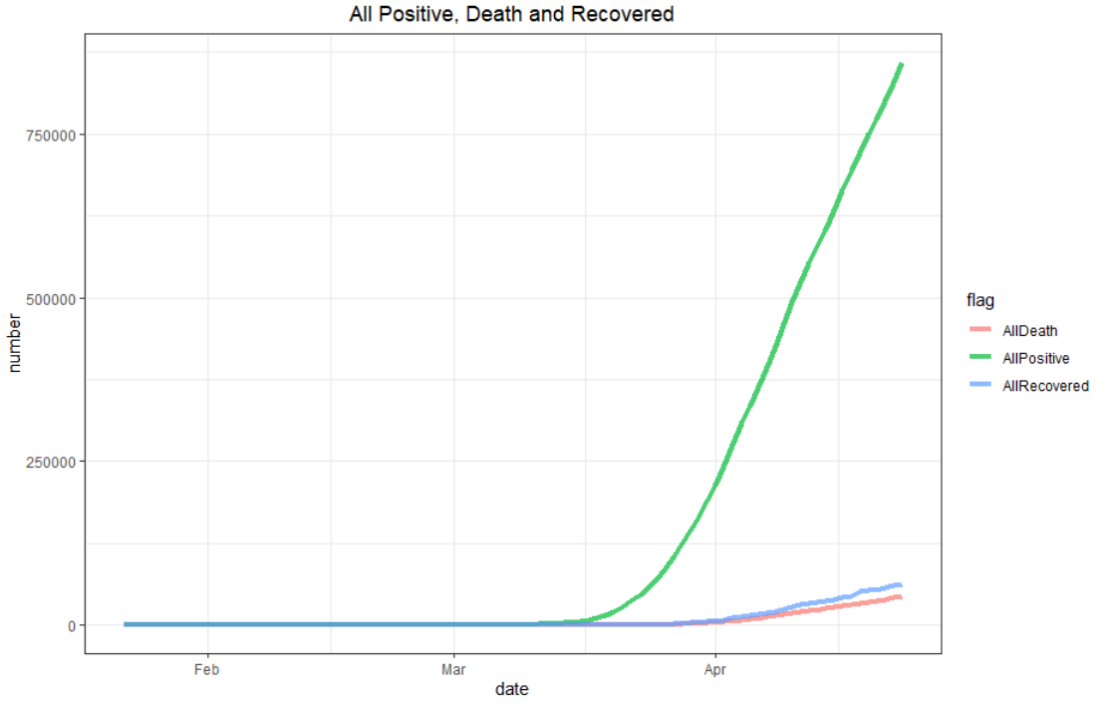

Figure1 All the positive, death and recovered change over time.

Next, we analyze the daily increase in the numbers of people who are tested to be positive and people who are dead. As is shown in Figure2, the blue line represents the daily increase in the number of people who are tested to be positive. It can be seen that the rate of increase is oscillating and going to be flattening out in the last two weeks. This pattern means that COVID-19 virus doesn't spread as much fast as before. We can give the hope that the situation is stabilizing and an inflection point in the spread of the virus is approaching. And the red line represents the daily increase in the number of people who are dead. We can observe the same pattern in the red line as in the previous blue line. Thus, this also indicates that the situation is stabilizing.

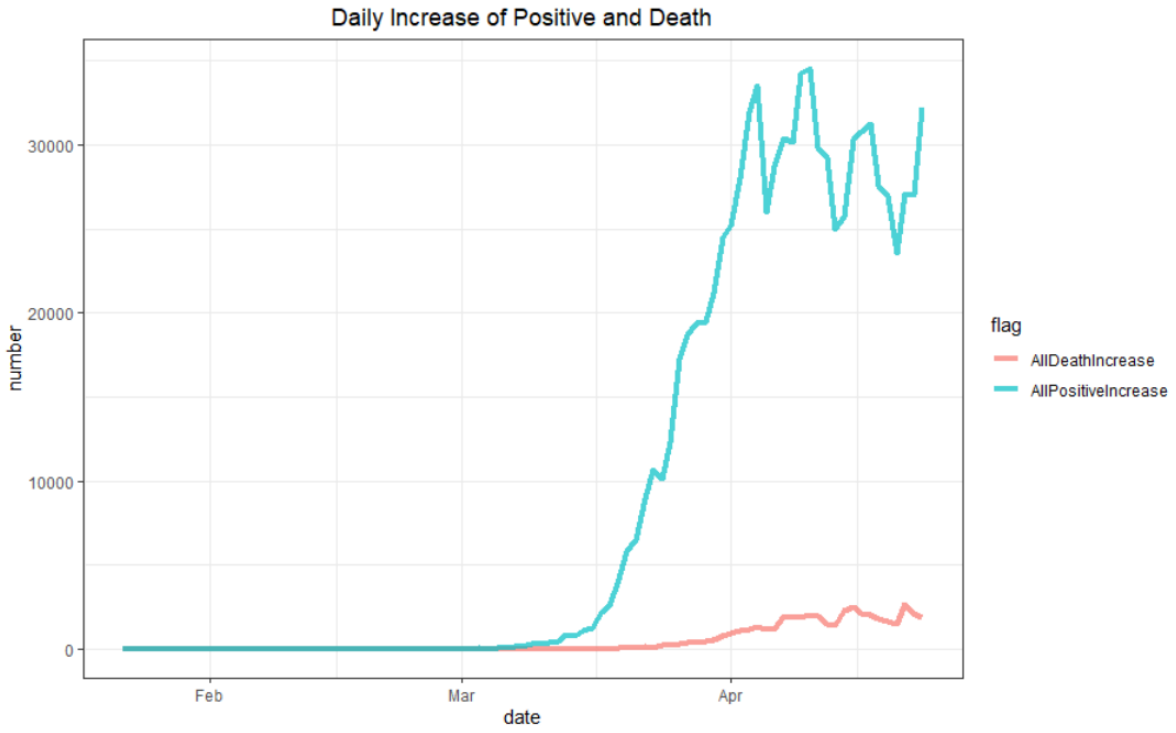

Figure2 Daily increase of positive and death over time.

Then, we hope to analyze the separate situations of COVID-19 virus in the different 50 US states. As the population and geographical location in each state are different, the transmission trend in each state is also different. Compare the cumulative number of people with positive results, the cumulative number of people who is dead in different US States. The results are shown in Figure3 and Figure4, respectively.

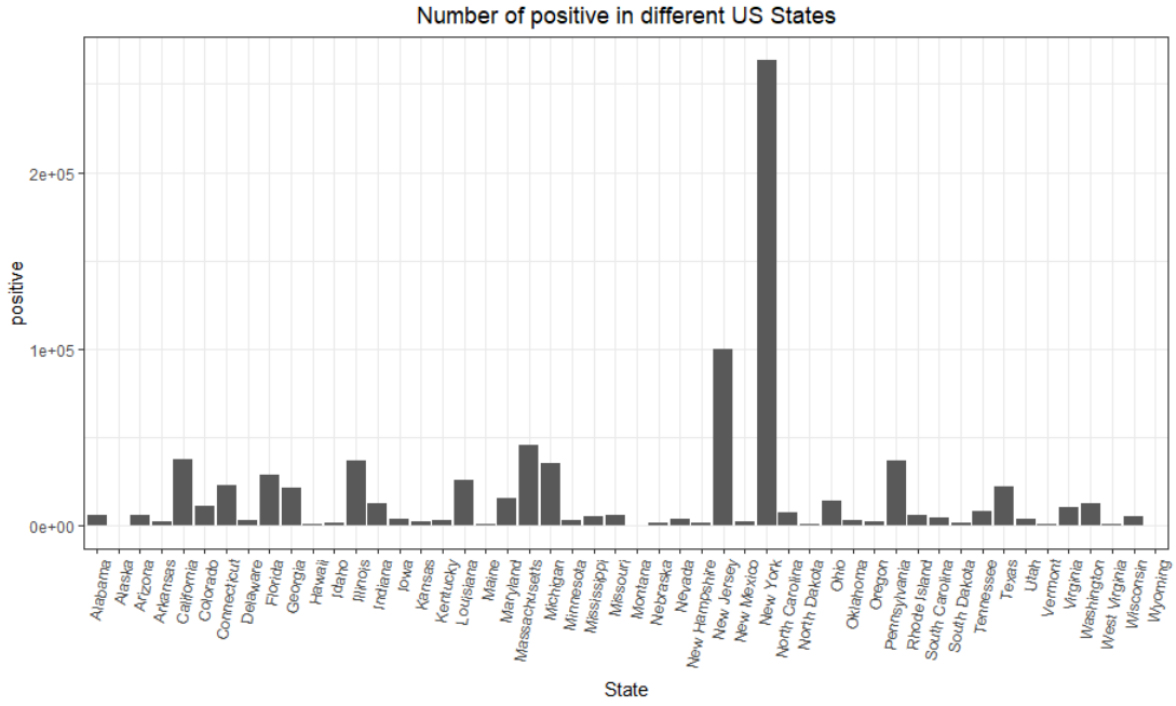

Figure3 Number of positive in 50 US states.

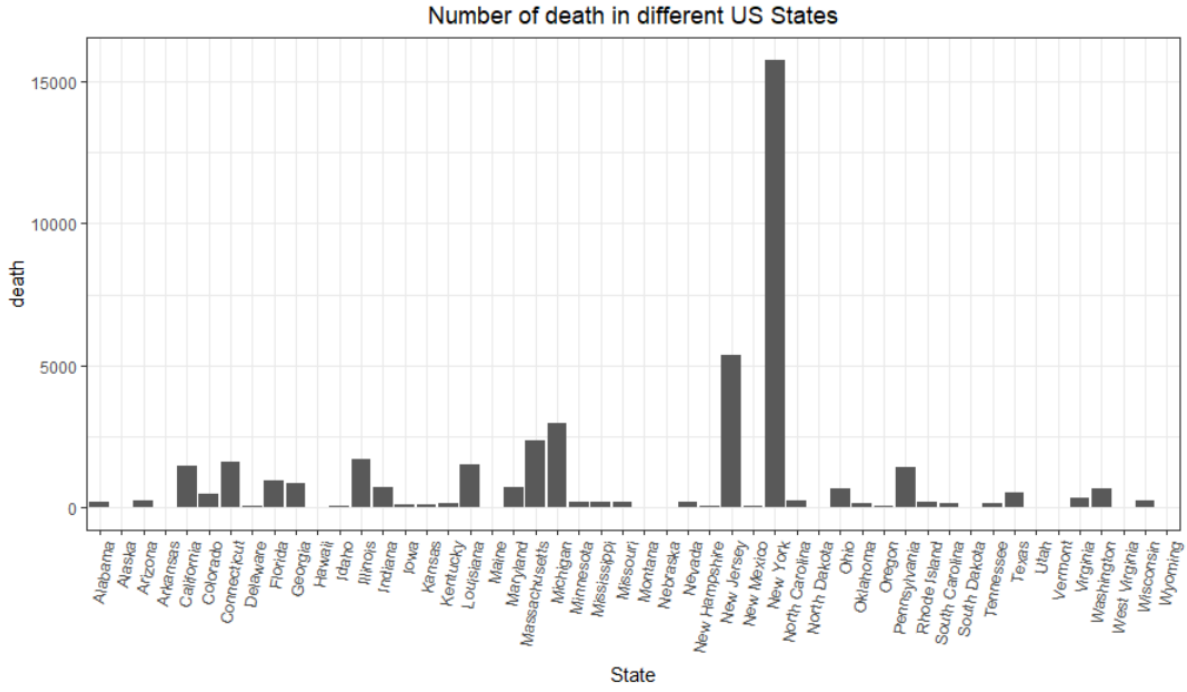

Figure4 Number of deaths in 50 US states.

From Figure3 and Figure4, it can be observed that both the cumulative number of people with positive results and the cumulative number of deaths are very high in New York and New Jersey compared to the other states. It indicates that the COVID-19 virus is very widespread in these two states. Lots of people are infected by the virus and many of them are dead from disease caused by the virus. New York and New Jersey should take more effective measures to prevent further spread of the COVID-19 virus.

Now, we use the variable *death* as response variable, the other four variables *positive*, *totalTestResults*, *positiveIncrease*, *totalTestResultsIncrease* as regressors and build a linear regression model between *death* and the other four variables.

```
Call:
lm(formula = death ~ positive + totalTestResults + positiveIncrease +
    totalTestResultsIncrease, data = modData)

Residuals:
    Min      1Q  Median      3Q     Max
-953.94  -59.11   -3.43   54.09 1586.75

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              29.1689169 11.9015725   2.451   0.0145 *
positive                  0.0719412  0.0006653 108.129  < 2e-16 ***
totalTestResults         -0.0026378  0.0002148 -12.281  < 2e-16 ***
positiveIncrease         -0.3598543  0.0084054 -42.812  < 2e-16 ***
totalTestResultsIncrease  0.0052822  0.0012515   4.221 2.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 207.3 on 734 degrees of freedom
Multiple R-squared:  0.987,     Adjusted R-squared:  0.9869
F-statistic: 1.393e+04 on 4 and 734 DF,  p-value: < 2.2e-16
```

Table1 Summary table of the linear regression model.

Once the linear regression equation is acquired, we can predict the response variable *death* when the other four variables are given. The summary table of the linear regression model is shown in Table1. It can be observed that the p-value of the four regressors are all significantly lower than 0.01, which means that the coefficients of the four regressors are significantly different from 0. And the R^2 is 0.987, which means that the linear regression equation can explain 98.7% of the variance in the response variable *death.* The linear regression equation is shown as below:

$$death = 29.17 + 0.072positive - 0.0026totalTestResults - 0.46positiveIncrease + 0.0053totalTestResultsIncrease$$

The diagnostics and residuals plots for the multiple linear regression model are shown in Figure5.

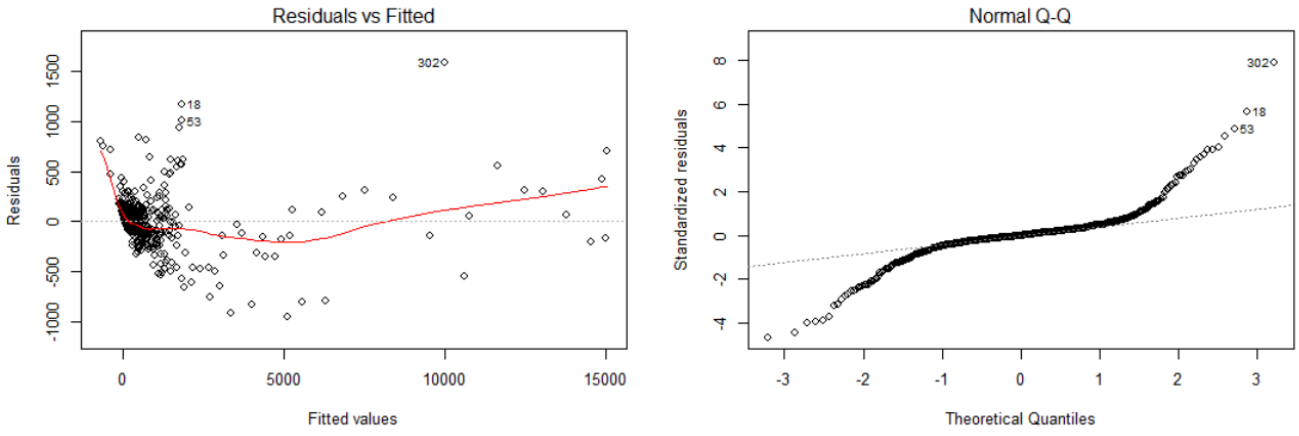

Figure5 Diagnostics and residuals plots of the multiple linear regression model.

From Figure5, we can observe from residuals vs fitted plot that the residuals are like white noise pattern. From normal Q-Q plot we can observe that the residuals are slightly different from normally distributed pattern. That may indicate that there are some nonlinear relationships between the response variable *death* and other variables. In the future researches, we can further explore these nonlinear relations and build related models.

At the end of this section, according to the different situations of COVID-19 virus in different states, we use hierarchical clustering method, which is one of the techniques in unsupervised learning to divide the 50 US states into 5 clusters. First, we obtain the data of four variables *death*, *positive, positiveIncrease*, *deathIncrease* in different states and standardize the data to a mean of 0 and a standard deviation of 1 because the variables differ widely in range. Then calculate the Euclidean distance matrix among different states. Finally, perform the average-linkage clustering method on the Euclidean distance and divide the 50 states into 5 clusters. The dendrogram of clustering is shown in Figure6.

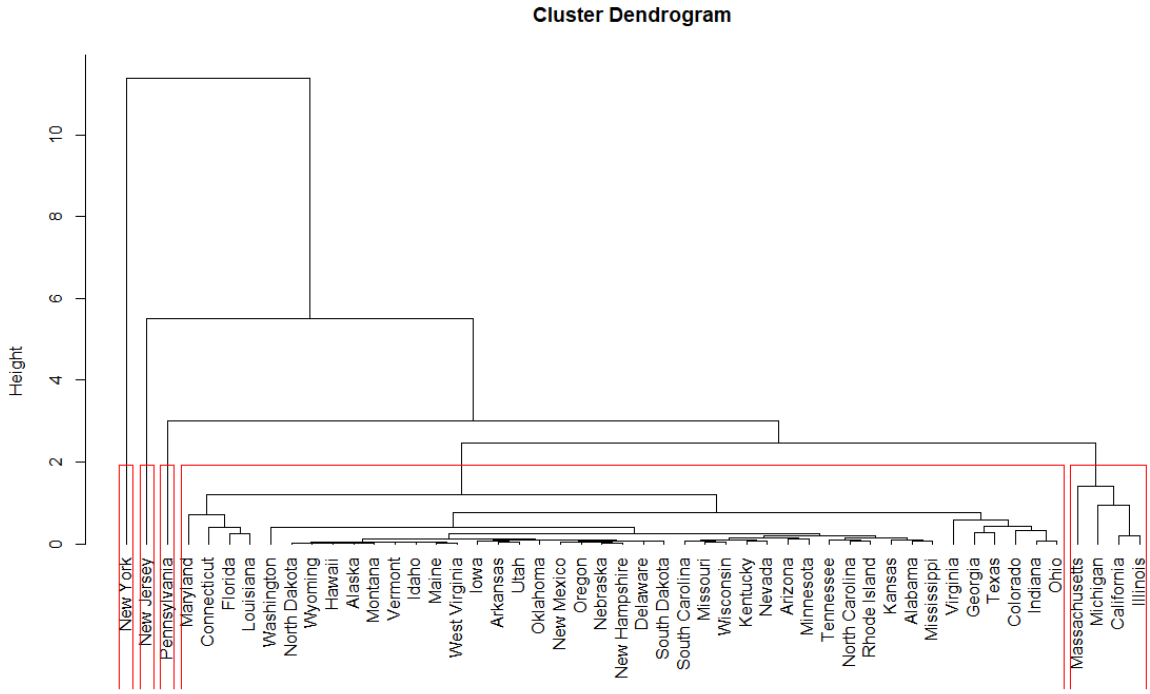

Figure6 Dendrogram of clustering

The dendrogram displays how items are combined into clusters and is read from the bottom up. Each observation starts as its own cluster. Then the two observations that are closest are combined. This continues until all observations are combined into a single cluster. The height dimension indicates the criterion value at which clusters are joined. For average-linkage clustering, this criterion is the average distance between each point in one cluster and each point in the other cluster.

According to Figure6, the red squares indicate the 5 different clusters. New York, New

Jersey, and Pennsylvania are divided into separate cluster, respectively. Massachusetts, Michigan, California, and Illinois are in the same cluster. And all the other 43 states are in the same cluster.

4. Conclusions & Discussion

In this report, descriptive statistics as well as different kinds of charts are utilized to analyze the dataset of COVID-19 in USA. We have a detailed understanding of how the COVID-19 virus is spreading in the United States. In the early stages of transmission, the virus spreads very fast and is close to exponential growth. And in the last two weeks, the rate of increase in the number of people who are tested to be positive is oscillating and going to be flattening out. We can give the hope that the situation is stabilizing and an inflection point in the spread of the virus is approaching. But the situations in the whole United States and different states are still severe. Lots of people are infected by the virus and many of them are dead from disease caused by the virus. The situations in New York and New Jersey are most terrible. Both the number of positive and death are much higher than the other states. Thus, New York and New Jersey should take more effective measures to prevent further spread of the COVID-19 virus.

Besides, we also build a linear regression model between the response variable *death* and the other four regressors. The linear regression equation is acquired and we can utilize this model to predict response variable *death* when the other four regressors is given. However, the normal Q-Q plot shows the residuals are slightly different from normally distributed pattern, which indicates that there may be some nonlinear relationships between the response variable and other regressors. And in the future research, we can further explore these nonlinear relationships and build related models.

In addition, at the end of this report, we use hierarchical clustering method, which is one of the techniques in unsupervised learning to divide the 50 US states into 5 clusters based on the different transmission trends of COVID-19 virus in different states. The features which are used in clustering are four variables *death, positive, positiveIncrease,* and *deathIncrease*. According to the dendrogram result of clustering, New York, New Jersey, and Pennsylvania are divided into separate cluster, respectively. Massachusetts, Michigan, California, and Illinois are in the same cluster. And all the other 43 states are in one cluster. It indicates that the US states in the same cluster may have similar situation and similar transmission trend of COVID-19 virus.

5. Reference

[1] The information is from World Health Organization:

https://www.who.int/news-room/q-a-detail/q-a-coronaviruses

[2] The data is downloaded from Kaggle:

https://www.kaggle.com/sudalairajkumar/covid19-in-usa#us_covid19_daily.csv

Group 14

Hongji Li, Shun Hu, Yuhan Liu, Yanfei Du