Which Feasible Measures Can Other Countries Learn from Korea to Alleviate COVID-19?

Di Xin dx489@nyu,edu
Jianan Gong jg6193@nyu.edu
Zheyuan Zhang zz2498@nyu.edu

May 8 2020

Github:
https://github.com/CindyXin97/COVID-19_Reaserch_Project

Abstract

COVID-19 impacted many countries in the world, Korea is one of the countries which has only less than 1,000 infected cases till now, which is a successful case that other countries should learn. When the pandemic suddenly broke out in Korea, the government took several effective measures to avoid it. In this project we are trying to evaluate what are some measures that successfully helped Korea to alleviate COVID-19. We used time series models to predict the duration as one of the factors, and used clustering techniques to get the evaluation standard. We further conducted classification models to evaluate the measures of Korea.

1 Introduction

Four months have passed since the coronavirus started, but experts still believe that the epidemic will continue to spread for a rather long time. How to slow down the spread of coronavirus has become an issue of concern to people all over the world. South Korea was one of the worst-hit countries in the coronavirus outbreak back in February. But a series of progressive responses has made it one of the exemples in the epidemic, because of its flexible implementation of a large scale testing as well as its consistent, transparent messaging to the public in the crisis. In this project, we are trying to use South Korea as an example to find insights in what general factors contribute to alleviating the spread of the virus. We hope to analyze relevant COVID-19 data from KCDC (Korea Centers for Disease Control & Prevention) to find demographic and social insights as well as feasible measures to effectively alleviate the epidemic. In order to achieve this, we conducted clustering over provinces and studied how policy and time series data influenced the spread of the epidemic. We developed a susceptible infected recovered model to determine how severe the epidemic is, and finally we also conducted a classification model to see what factors, like city infrastructures, transportation, age distribution, search trend affect the spread most. Based on what we have analysed, we emphasize the importance of people's awareness towards the epidemic, which is clearly helpful in general.

2 Literature Review

Coronavirus disease 2019 (COVID-19), which will cause severe respiratory illness like lung failure, was reported firstly in Wuhan, China. The number of COVID-19 cases reported to the WHO has been growing since the first case reported in Dec.2019 from the WHO China Country Office[1]. Only 31 cases were reported in South Korea on Feb. 18, 2020[2] and most of them were travelers from China. But a religious group in the Daegu metropolitan area and a hospital nearby make COVID-19 spread rapidly to other cities in South Korea[3], making South Korea one of the worst-hit countries. However, South Korea has made aggressive responses, that a consistent decrease in confirmed cases could be seen since early March. Many researchers attributed this decrease to its intensive testing[4,5,6], while some other researchers believe social distancing played an important role[7] and appeal for school closure[8]. We also want to use Korea as an example and analyze how different age groups and public infrastructure in the city affect the spread of the epidemic.

3 Data

3.1 Data Source

KCDC open dataset and Korean Statistical Information System (https://www.kaggle.com/kimjihoo/coronavirusdataset, http://kosis.kr/eng/statisticsList/)

Time: Time series data of COVID-19 by region including test number, negative/positive number,
released number and deceased number along with patient info like age, sex and location.

Case: Data of COVID-19 infection cases in South Korea including location, group infection,infection case(overseas or infected place).

PatientInfo: Epidemiological data of COVID-19 patients in South Korea including age,sex,contact people number, confirm date, release date and decease date.

PatientRoute: Route data of COVID-19 patients in South Korea including where the patients visited and dated.

SearchTrend: Trend data of the keywords searched in Naver(largest portal in Korea). The keywords are cold, flu, pneumonia, coronavirus

Weather: Data of the weather including temperature, precipitation, wind speed and relative humidity.

Sickbed: Detailed Status of sickbed by Region

Region: Statistics of public infrastructure and age & educational structures grouped by city, such as the ratio of elder people, the number of hospitals and schools, etc.

KOSIS - Korean Statistical Information Service: Demographic Population Distribution in South Korea (as of 2020)

3.2 Data Preprocessing

NaN value filling: For numeric data, we filled NaN with the mean of this column.

Data source integration: Our data are from different organizations, the data structure and attribute presentation are quite different. For example, Jeju and Jeju-do mean the same location even though they seem different. We need to calibrate this presentation difference and merge all the data together properly.

Data type conversion:For date attributes, we need to convert the string to datetime.

Scaling: Since we will use the clustering method later, we need to preprocess the features by standardscaler.

3.3 Exploratory Data Analysis

We first studied the number of cases over time and the relationship between the case number and the gender. As can be seen from Figure-1, women are slightly more susceptible to infections than men, however the number of deaths is not very different between the gender.

As for the age distribution of the confirmed and released patients from Figure-2, the elders are more likely to pass away once they are infected. No patients under 30 are dead until now which indicates that the impact of COVID-19 is different among different age groups. Surprisingly there are more young patients (especially 20s) than the elders, unlike many media-consumers' impression, there are less number of elders getting infected. However, when combined with

demographic population distribution data from KOSIS in Figure-3, generally speaking, the ratio of elder patients is higher. This is in line with common sense.

Figure-4 shows the distribution of causes of infection and Figure-5 shows how the searching trend is correlated to the confirmed cases over time. Most of the infection is caused by Shincheonji church which is a huge church with more than 10,000 cultists. 1,001 cultists participated in the gathering in Daegu causing the spread of COVID-19. It's easy to understand that more people want to search for "coronavirus", when more people get infected and reported.

It can be seen from Figure-6 and Figure-7 that the current number of patients increased at the beginning and dropped down to around 1000 by the end of April while the positive rate when taking tests is constantly falling due to the rapidly increasing number of tests taken. And Figure-8 outlines the general situation about where the epidemic is most serious. The province Seoul ,Gyeongsangbuk-do, Daegu, and Gyeonggi-do are the most hit areas.

3.4 Anomaly Detection and Related Policy Analysis

In order to detect the anomaly days with infected people, we conducted cluster-based model anomaly detection and it performed the best in this situation with unlabeled data. After that, we think we are all consistent influence factors. Policy is one of the important factors that play an important role in these anomaly days so we dive deep into the policy.

By using Gaussian Mixture model and Kmeans for the cluster-based model, we conclude that 03/01, 03/02, 02/29/,03/12, 04/10 are analyzed as anomaly points (Figure 9,10) which shows the significant drop or increase. In Figure G, the distribution about policy starting date, these anomaly days are corresponding to the policy starting date. In order to have a deeper analysis, we draw the line graph(Figure 11,12,13) of two policies immigration and Education, as it occupies the highest percent. Combining the new infected number with starting dates of policy, the policy has significant impact around 5 days after it comes up. Thus, 03/01, 03/02, 02/29 are around 5-7 days later where Korea started public scale mask distribution and alerts, 03/12 is 4-7days where

Korea government started Emergency Use Authorization of Diagnostic Kit, 04/10 is 3-5 days after school starts online classes.

3.5 Susceptible Infected Recovered (SIR) Analysis

The SIR model of infection describes time dynamics of an infectious disease spreading through a homogenous closed population. The population is divided into three categories: Susceptible S, Infective I, or Recovered/Dead R. We simplify R for now and include dead people as recovered. When people die, if properly buried, they cannot infect anymore and so they are equivalent to people that have recovered and are immune to the infection.

Susceptible(S): are those that have not acquired immunity yet and are susceptible to becoming infected.

Infected(I): have been infected with the disease.

Recovered(R): are cured and not susceptible anymore to the disease.

The mathematical equation is shown as below:

$$\frac{dI}{dt} = \beta(1 - I - R - \gamma/\beta)I \quad I(0) = I_0$$

$$\frac{dR}{dt} = \gamma I \quad R(0) = R_0$$

β is the possibility of S to I and γ is the possibility of I to R.

The quantity β/γ is called R0 which indicates one infected person can infect how many susceptible individuals. If $R0>1$, then an epidemic will take place. If $R0\leq1$ then there will be no epidemic.

Importantly, a disease's R0 value only applies when everyone is completely vulnerable to the disease in a certain population. This means: 1) no one has been vaccinated 2)no one has had the disease before 3) there's no way to control the spread of the disease

We are now applying the SIR model to estimate the R0 of COVID-19 by modeling Korea's infection data with Bayesian estimation in order to figure out the critical Korean β and γ parameters.

From Figure 14, we can know that β is well estimated (2.3) by leveraging estimation of the non-dimensional parameter R0 (4.8), and γ(0.55). This means without any control, one person who is infected by COVID-19 will infect 4.8 individuals.

4 Methodology

4.1 Time series analysis

To predict when the COVID-19 in Korea will end, we need to look into 2 indicators--confirmed and recovered patients.

MLP Regression:

We used the Multi-layer Perceptron(MLP) regressor to predict when will COVID-19 get over in Korea at first. Compared with other regressors, MLP is a nonlinear function approximator which is used for regression problems. We can see that based on the results (Figure 15,16), the pandemic impact will approximately end on May 10th and the trends have shown the infected people are decreasing now.

Prophet Model:

Unlike other traditional time series models, the algorithm behind its approach is to fit the regression model. It works pretty reasonably by default, without setting any parameters explicitly. We set the predicting period with 30 in order to see the future trends in 30 days. The yhat_lower and yhat_upper shows the predicted results fluctuate in a reasonable range. Figure 17 shows the predicted infected number has an increasing trend in May and June while with a stable decreased slope. We generated the interactive graph like Figure 17, it shows that on May 20th, the infected people will reach 5506 people overall. For Figure 18, for each week, it reached the lowest point on Thursday and highest on Friday and Saturday.

ARIMA:

We are going to use the ARIMA(Auto Regressive Integrated Moving Average) model to predict these two indicators. ARIMA can explain a given time series based on its own past values, with its own lags and the lagged forecast errors. It has three parameters: p,d,q.

p: lags of the dependent variable. For example if p is 2, we will use x(t-1), x(t-2) to predict x(t)
d: These are the number of nonseasonal differences. In COVID-19 scenario, we put d=0
q: lagged forecast errors in prediction equation.

First, we need to choose p,d,q parameters of the ARIMA model by plotting Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

Then we predicted daily confirmation(Figure 21) and recovered(Figure 22) in the future 2 months and visualized the original and predicted data. Based on the predicted data, we can find the equilibrium when the total recovered equals total confirmed. The data is 2020-05-17. So the COVID-19 epidemic will basically end on that day(Figure 23).

4.2 Clustering

Clustering Algorithm is used to find provinces with similar situations. We tried Kmeans, Gaussian Mixture and used silhouette score as a guideline to pick the number of clusters. Finally, after carefully selecting and scaling the features and setting the province name as the index, we clustered using KMeans to get three categories based on total confirmed cases, total released cases, total deceased cases, epidemic duration of the province and R of the province. Duration means the time interval between the date of the first confirmed case in each province and the anticipated end date. The result is shown in Figure 24.

After the clustering, we can get three groups with quite different characteristics. The result is quite reasonable because it actually represents the epidemic situation in South Korea. It can be seen from the figure that Daegu is the worst-hit area and is a cluster itself. It is because there are more than 6800 cases and no more than 1500 cases in any other provinces. And the other two groups are basically divided by the severity of the epidemic situation in the province. Provinces labeled as 0 generally have the features that the epidemic started early but the number of confirmed cases stays low, which potentially shows that the provinces have controlled the epidemic well. And the other provinces are labeled as 1.

4.3 Classification

As cluster labels serve as an indication of how bad the situation is in each province, we used it as a target in classification to further analyse what factors most affect the severity. We have chosen 'elementary_school_count', 'kindergarten_count', 'university_count', 'nursing_home_count', 'academy_ratio', 'elderly_population_ratio', 'elderly_alone_ratio', 'vehicle', 'search trend' as our features after delicate selection from our dataset, and used CART Tree, random forest to do classifications and analysed the feature importance. The result could be seen in Figure 25, Figure 26.

As could be seen from the result, the important features are “nursing_home_count”, “academy_ratio”, “vehicle”, “search”. It is quite obvious that nursing home counts and academy ration and vehicle number all represent the scale of the province, and when the province is more prosperous and has more population, it is generally harder to control the epidemic. Despite that, we also find that search trend, which reflects the awareness of the people, matters. This indicates that when people are willing to find out what “coronavirus” is, the more likely they will be cautious, which leads to better control of the epidemic.

5 Results and Conclusion

In order to have a first glance of how COVID-19 spreads in Korea, by taking into look at the Patients, Region, Infected Reason, Social Media Trends, and Policy, we can suggest that considering the high proportion of young adults, the government should regulate their social behaviors to control the spread and infection. Especially for the policy part, as it plays an important role at this epidemic fight, we conducted the cluster-based anomaly detection model and identified 5 anomalous days. After building the relationship of these points with policy, we’ve found that mask distribution, Emergency Use Authorization of Diagnostic Kit and online courses are three critical reasons for Korea to succeed in this fight.

Through the clustering methods, we divided cities in Korea into 3 infective levels. And we conducted three time-series model related MLP, Prophet, ARIMA to predict the number of existing COVID-19 infected patients for each day, we found that this pandemic will end on May 17th, 2020 as ARIMA performs the best. By taking 'elementary_school_count', 'kindergarten_count', 'university_count', 'academy_ratio', 'elderly_population_ratio', 'vehicle', 'nursing_home_count', 'elderly_alone_ratio' and 'search trend' as features, we conducted classification model with decision tree and random forest. Finally, we identified nursing home counts, academy ration, vehicle number and search trend really influence the performance of control in Korea. In conclusion, starting emergency diagnosis for the elderly, closing the schools, limiting people's outdoor activities and raising people's awareness are feasible measures in Korea which other countries can learn from.

Reference

[1]WHO Novel Coronavirus (2019-nCoV) SITUATION REPORT - 1 21 JANUARY 2020. Available from https://www.who.int/docs/default-source/coronaviruse/situationreports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4.Accessed 28 Feb. 2020.

[2]KCDC COVID-19 situation reports in South Korea (18 Feb. 2020). Available from https://www.cdc.go.kr/board/board.es?mid=a20501000000&bid=0015&act=view&list_no=366228&tag=&nPage=3. Accessed 28 Feb. 2020.

[3]KCDC COVID-19 situation reports in South Korea (01 Mar 2020). Available from https://www.cdc.go.kr/board/board.es?mid=a20501000000&bid=0015&act=view&list_no=366410&tag=&nPage=1. Accessed 02 Mar. 2020.

[4]COVID-19 National Emergency Response Center, Epidemiology & Case Management Team, Korea Centers for Disease Control & Prevention. Contact transmission of COVID-19 in South Korea: novel investigation techniques for tracing contacts. Osong Public Health and Research Perspectives, 11(1):60–63, 2020.

[5] Dennis Normile. Coronavirus cases have dropped sharply in South Korea. What's the secret to its success? Science, 2020. https://www.sciencemag.org/news/2020/03/coronavirus-cases-have-dropped-sharply-south-korea-whats-secret-its-success. Accessed March 21, 2020.

[6] Seoul Metropolitan City Government. Official website of the Seoul Metropolitan City Government. 2020. http://www.seoul.go.kr/. Accessed January 20 – March 24, 2020.

[7] Park, Sang Woo, et al. "Potential roles of social distancing in mitigating the spread of coronavirus disease 2019 (COVID-19) in South Korea." medRxiv (2020).

[8] Kim, Soyoung, et al. "School Opening Delay Effect on Transmission Dynamics of Coronavirus Disease 2019 in Korea: Based on Mathematical Modeling and Simulation Study." *Journal of Korean medical science* 35.13 (2020).

Appendix

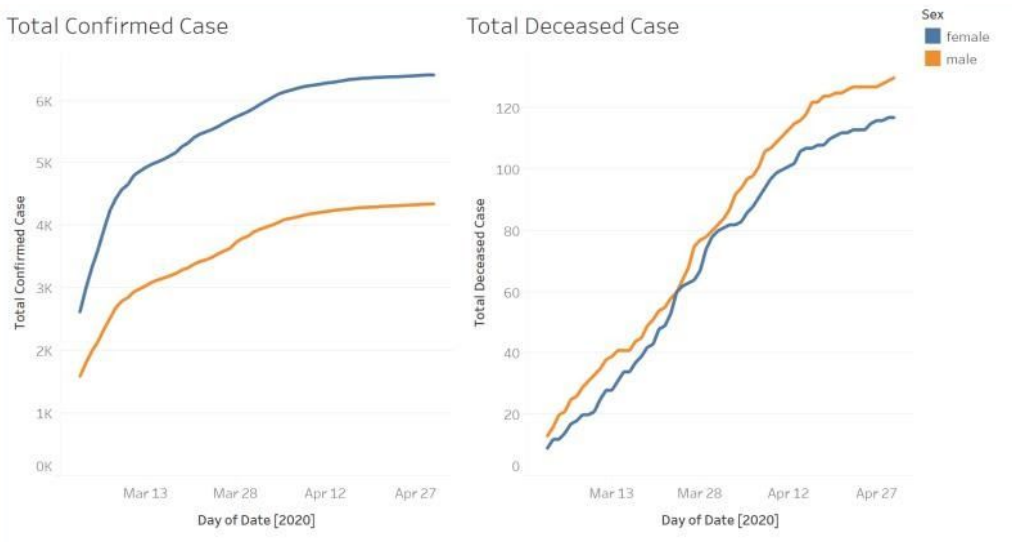

Figure 1

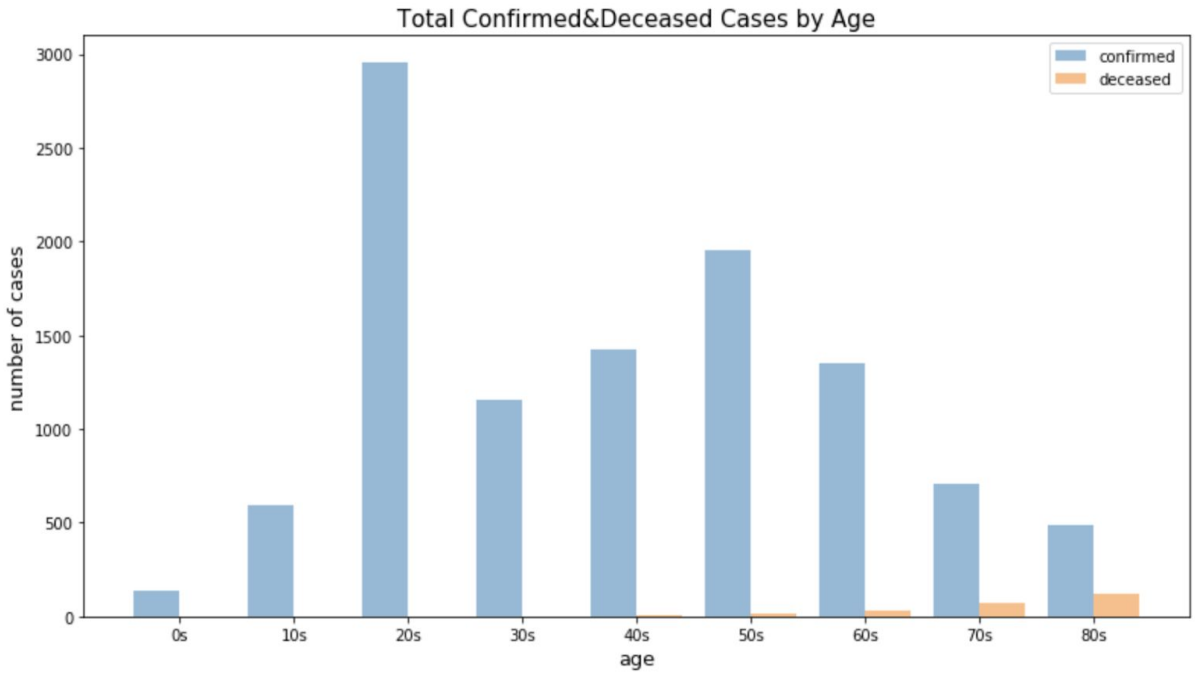

Figure 2

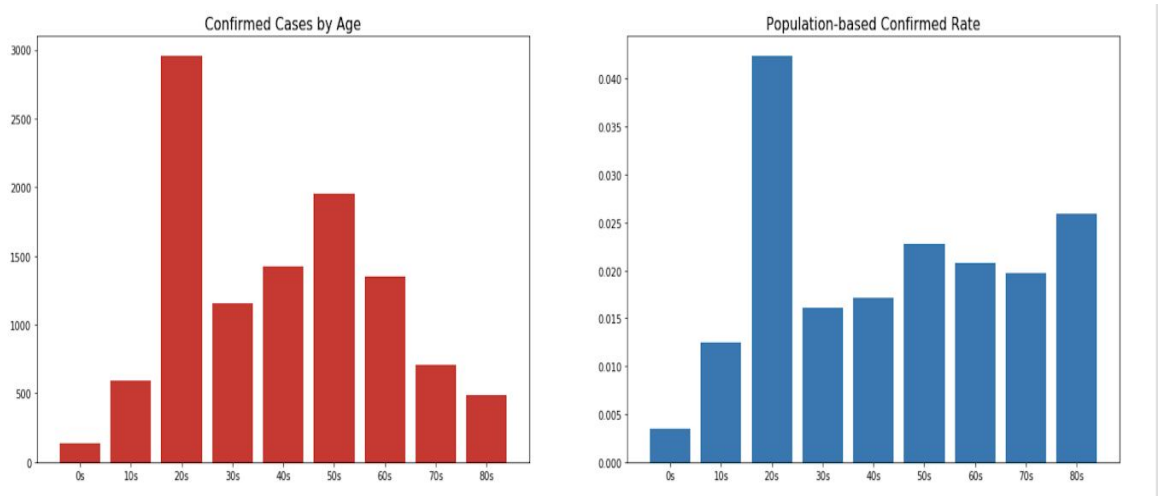

Figure 3

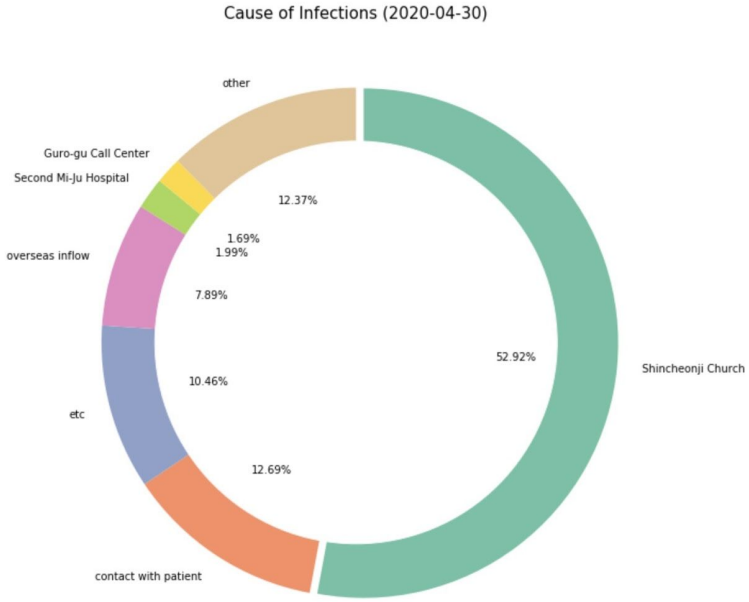

Figure 4

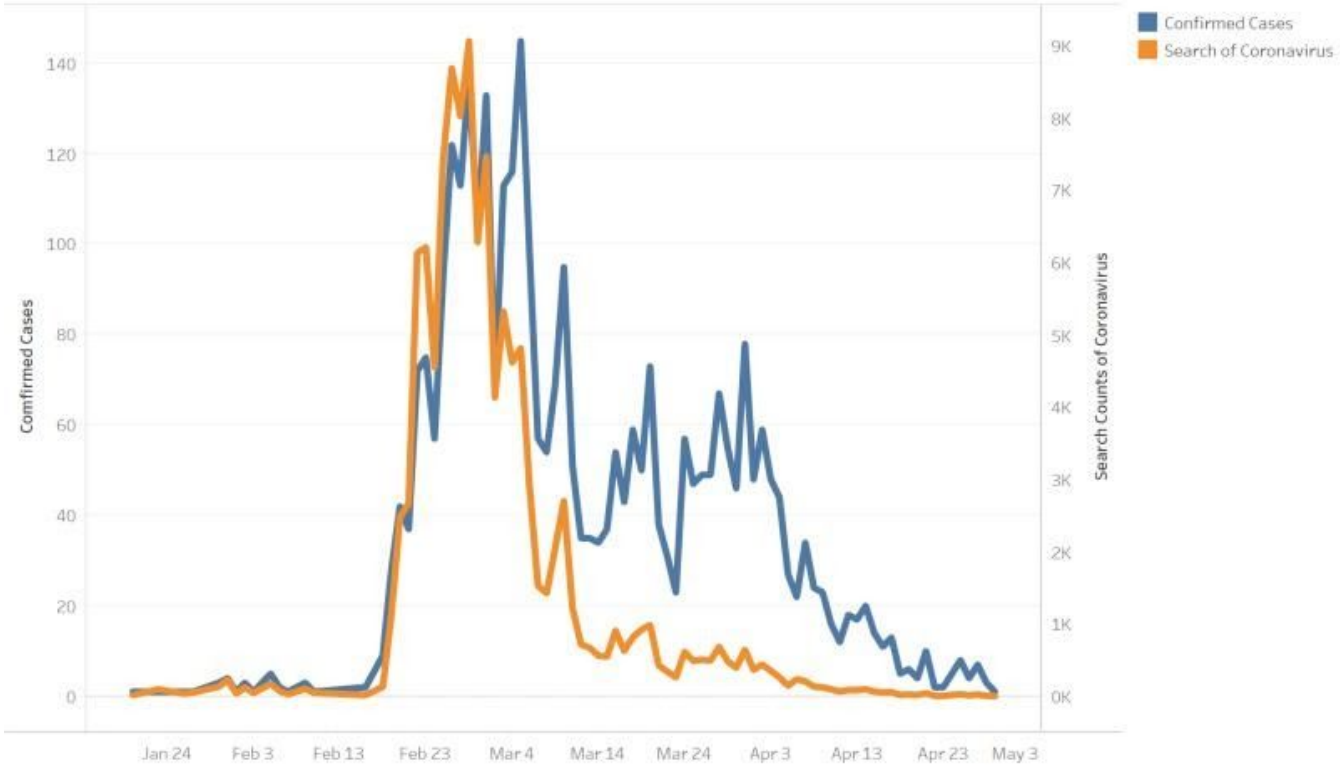

Figure 5

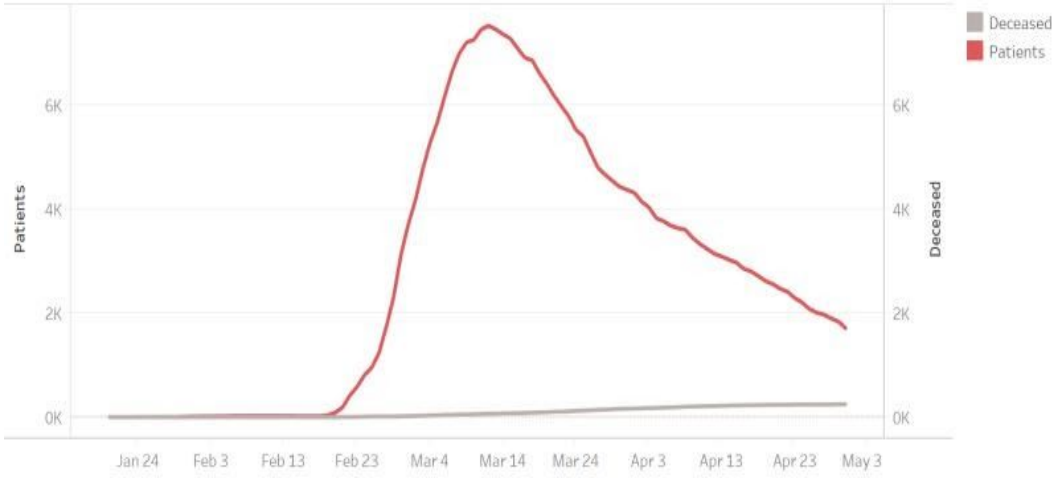

Figure 6

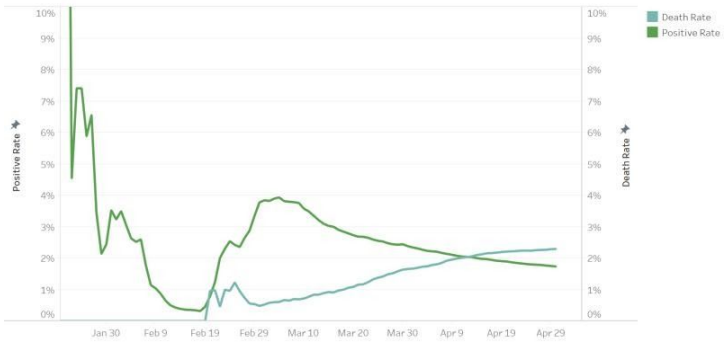

Figure 7

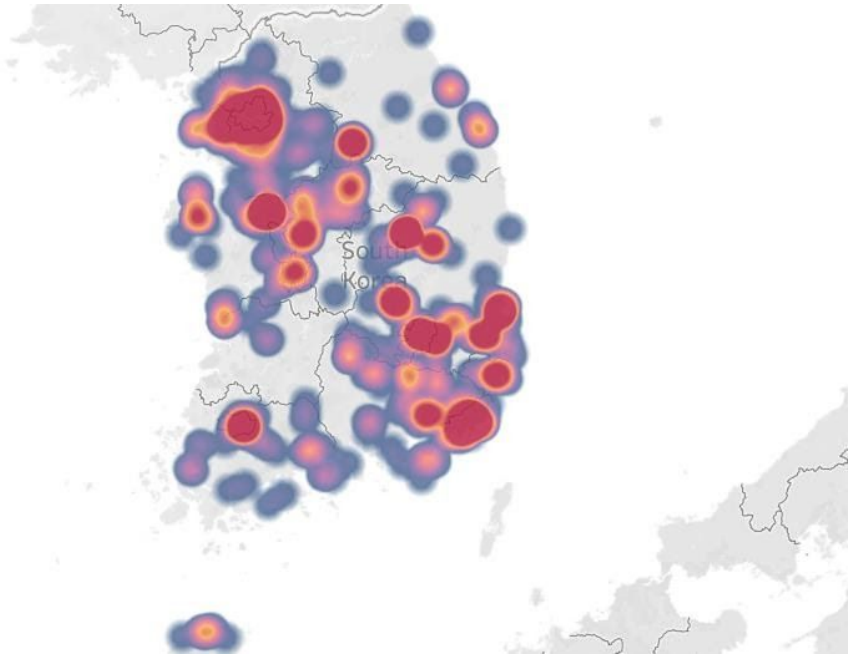

Figure 8

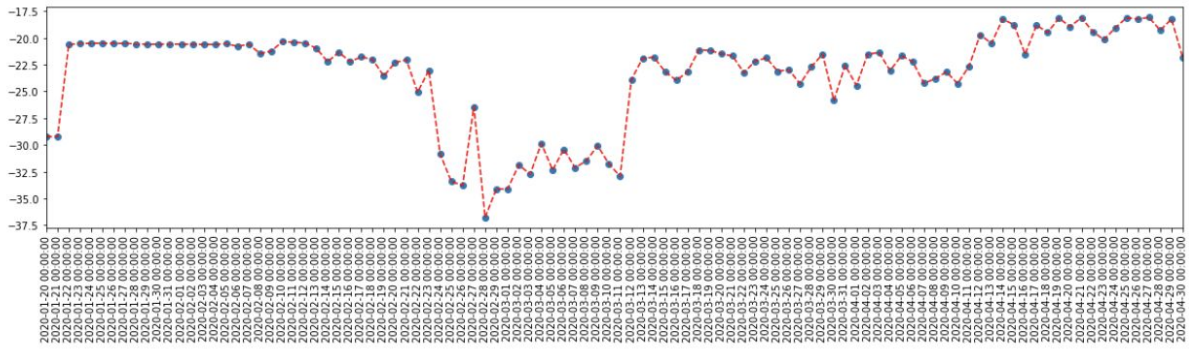

Figure 9

```
          date         score
41  2020-03-01  98833.513066
42  2020-03-02  97430.184559
40  2020-02-29  91506.354090
52  2020-03-12  88969.205871
81  2020-04-10  87619.695657
0     42
2     21
4     14
1     14
3     11
dtype: int64
```

Figure 10

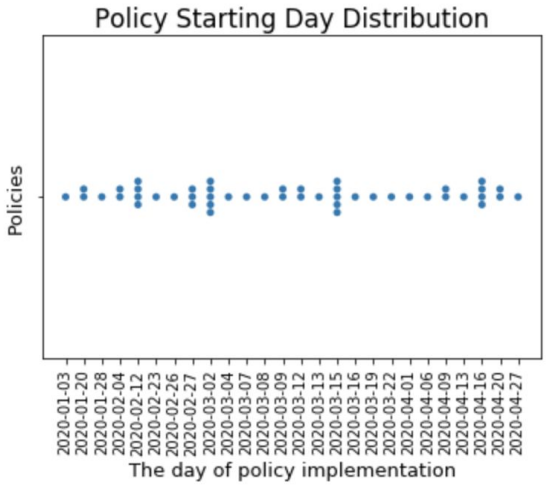

Figure 11

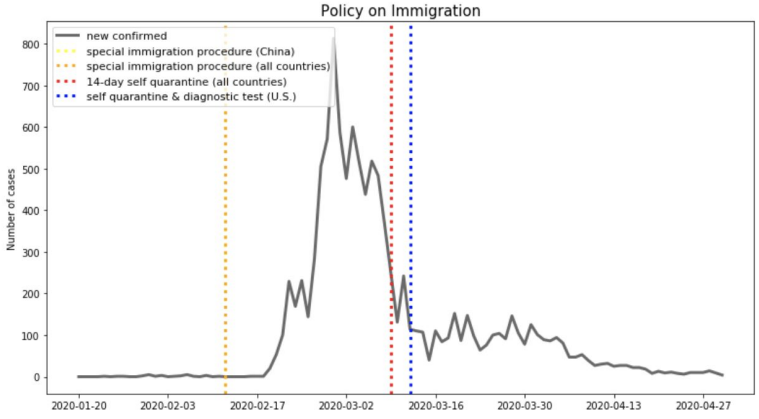

Figure 12

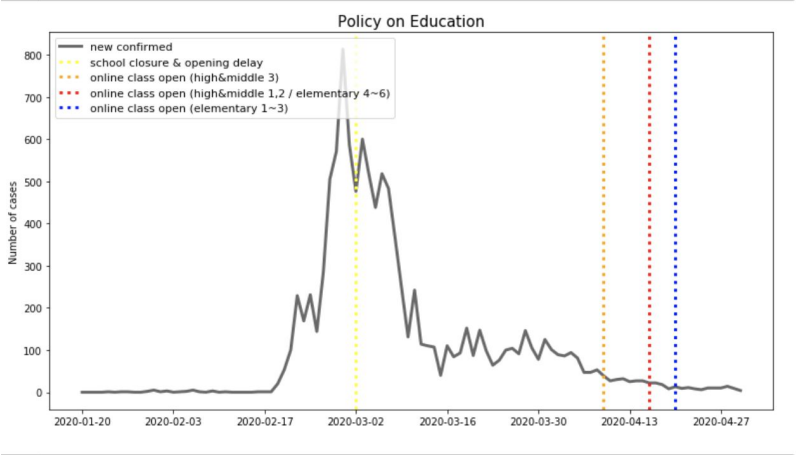

Figure 13

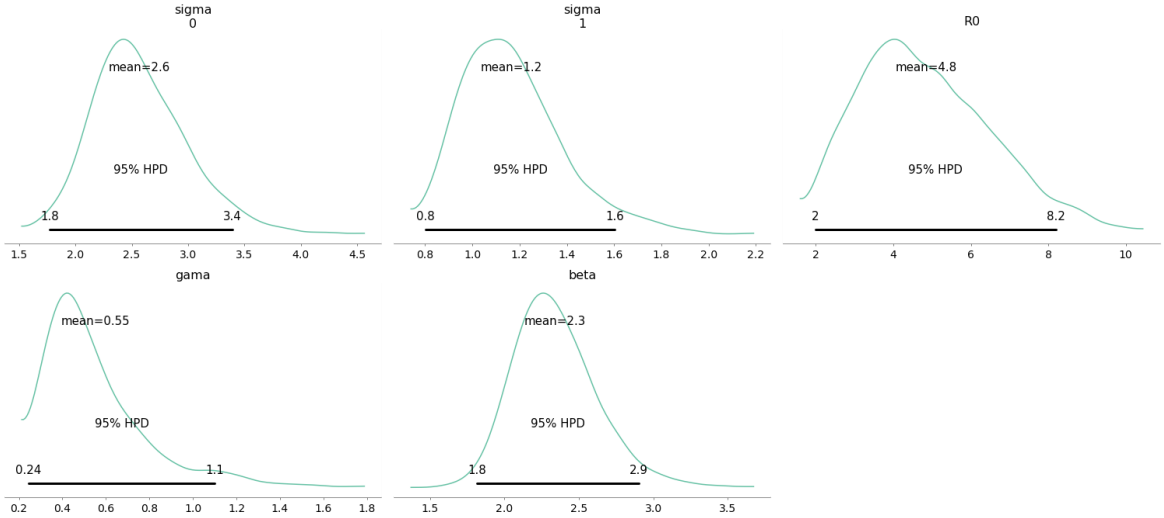

Figure 14

```
1  predicted_count
```

```
2020-01-20   -355
2020-01-21    -55
2020-01-22     10
2020-01-23     10
2020-01-24     10
              ...
2020-05-06    417
2020-05-07    296
2020-05-08    174
2020-05-09     52
2020-05-10    -70
Length: 112, dtype: int64
```

Figure 15

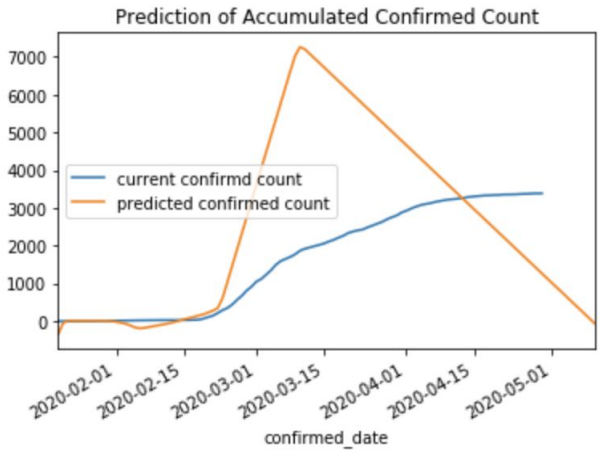

Figure 16

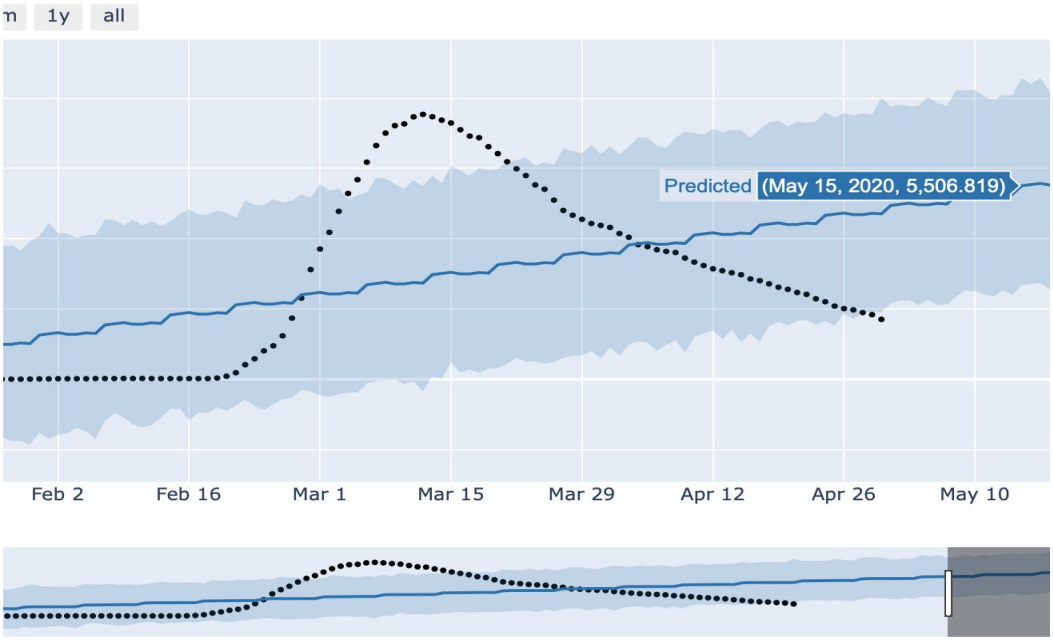

Figure 17

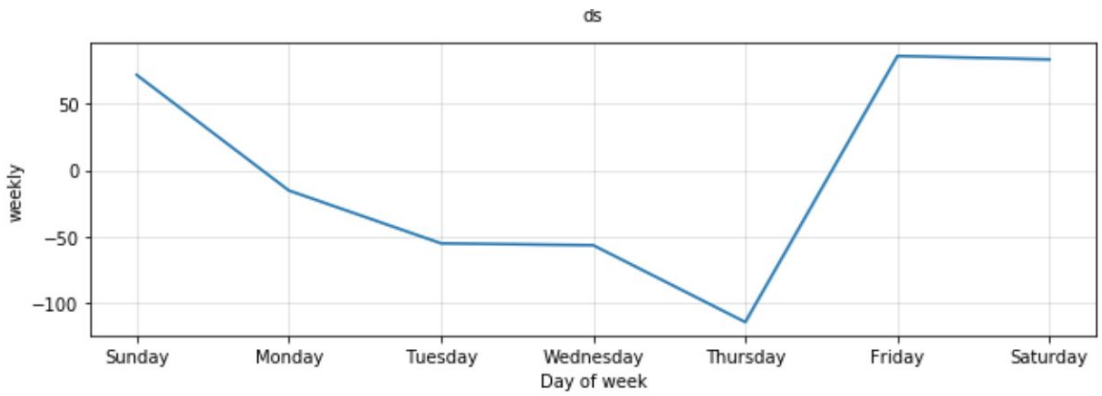

Figure 18

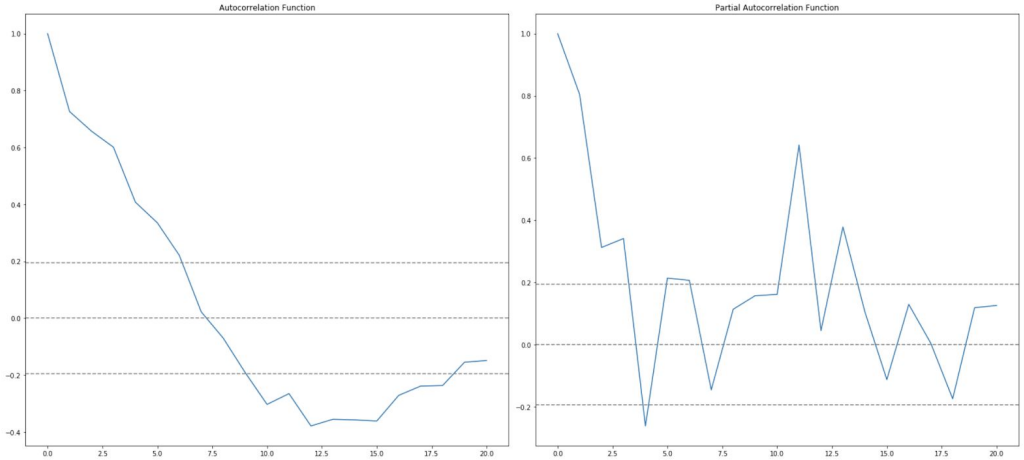

Figure 19

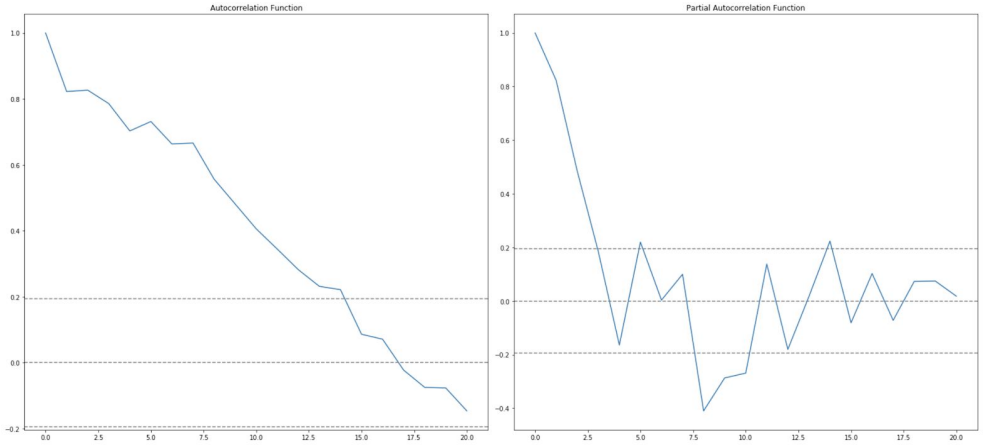

Figure 20

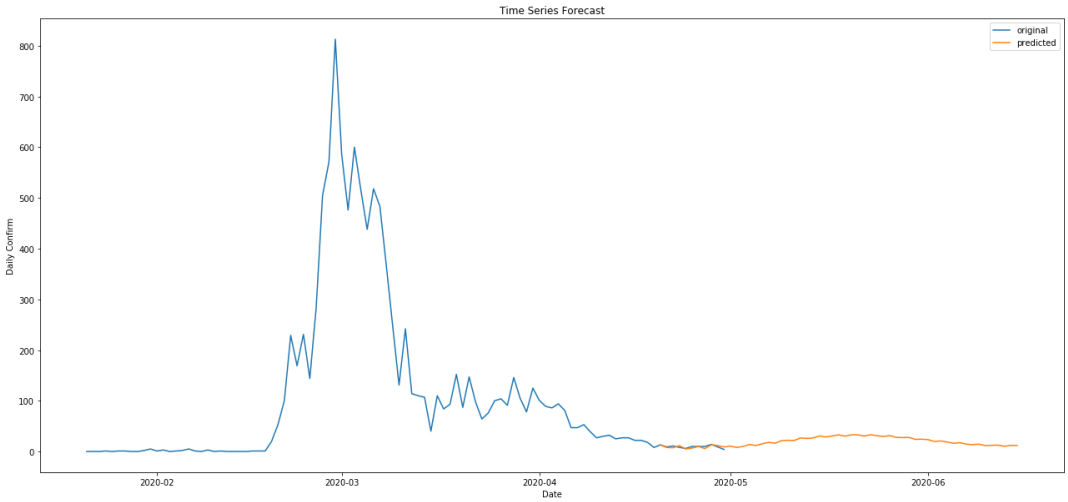

Figure 21

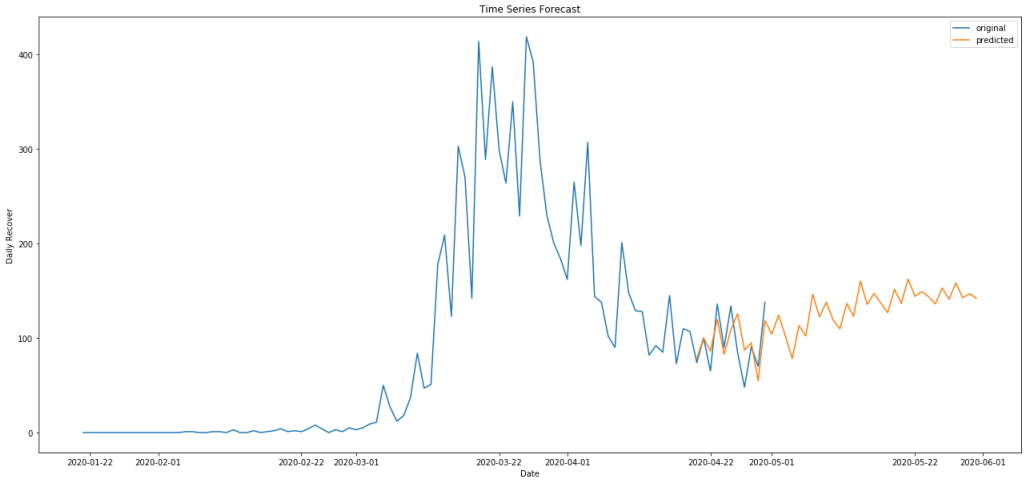

Figure 22

	confirm	recover	sum_confirm	sum_recover
2020-05-13	25.805064	106.879051	10986.378184	10685.105086
2020-05-14	26.737430	133.961570	11013.115614	10819.066655
2020-05-15	30.671126	115.901459	11043.786740	10934.968114
2020-05-16	28.625066	124.656092	11072.411806	11059.624206
2020-05-17	30.708044	110.187927	11103.119850	11169.812133

Figure 23

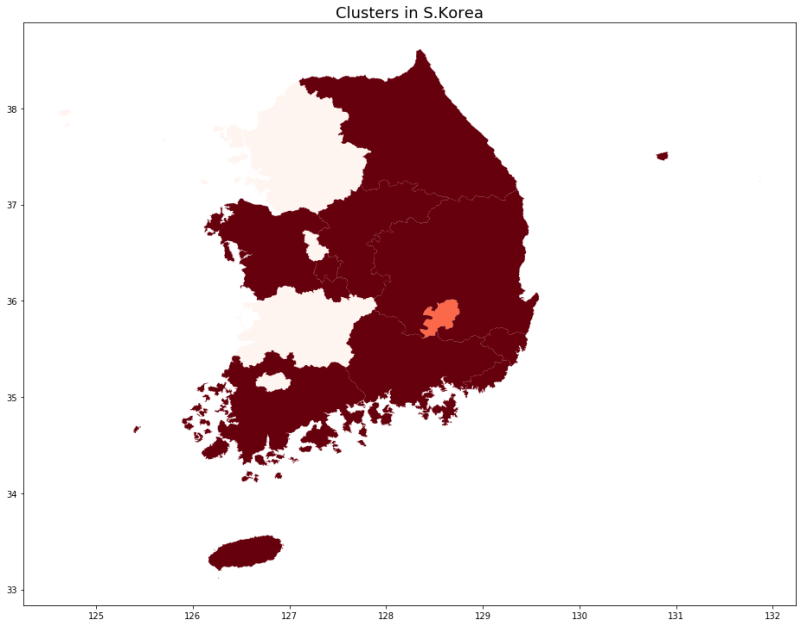

Figure 24

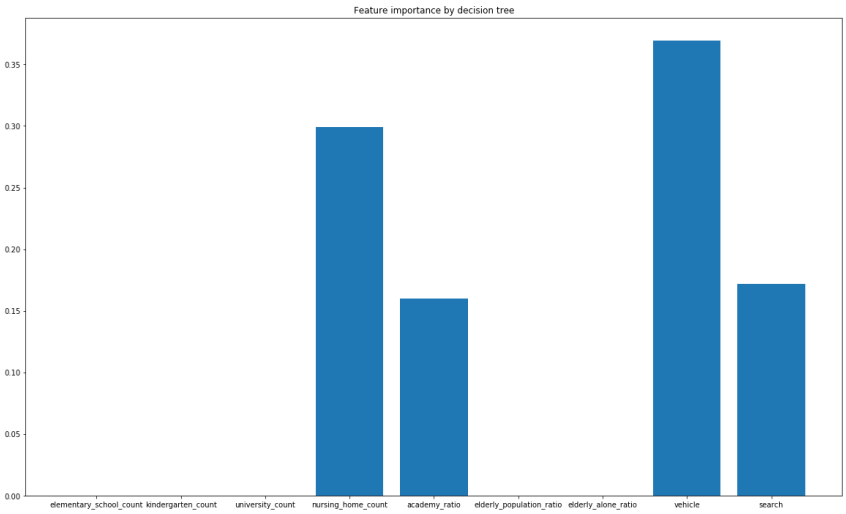

Figure 25

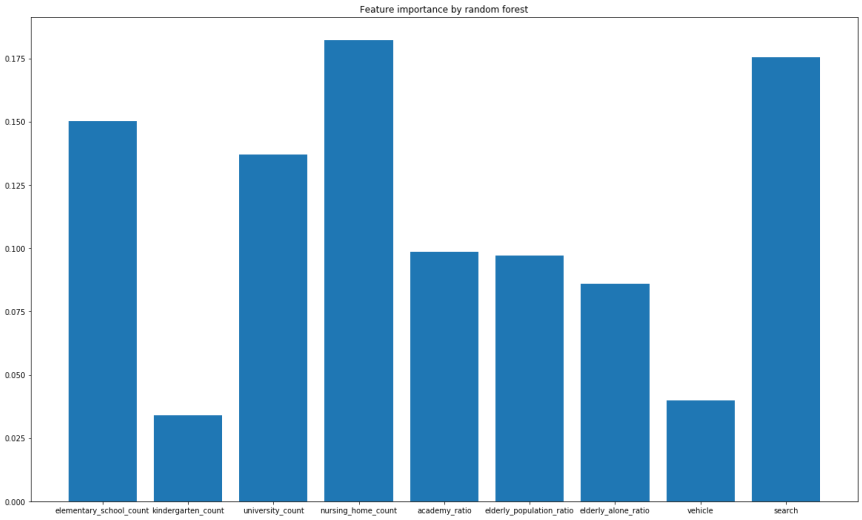
Feature importance by random forest
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
elementary_school_count
kindergarten_count
university_count
nursing_home_count
academy_ratio
elderly_population_ratio
elderly_alone_ratio
vehicle
search

Figure 26