Novel Corona Virus 2019 – Visualization and Prediction

Bohan Zhang
Email: bzhan014@ucr.edu

Jialiang Ruan
Email: 1121153676@qq.com

Pai Zhang
Email: 1124321458@qq.com

Abstract: In this report, we used the Novel Corona Virus Dataset found from Kaggle. This dataset provides the numbers of confirmed, deaths, recovered cases in different provinces and countries in different dates. By analyzing the contents of this dataset, we primarily aim to seek answers to the following questions: 1) As COVID-19 is prevailing in the world, how can we see the circumstance of controlling this virus? 2) How worse will the condition go in the future? In order to answer these questions, we provide data visualization of this dataset in different forms and perspectives. We also applied several meaningful models to forecast and predict the confirmed case number in the future.

1 Introduction

This report illustrates visualization of the Novel Corona Virus Dataset. We analyzed the data of the world cases, China, Korea, and Hubei province for illustration. One of the files in this dataset provides daily level information on the number of 2019-nCoV affected cases across the globe. The dataset is saved in csv file. It has 9 columns listed following:

- Sno - Serial number
- ObservationDate - Date of the observation in MM/DD/YYYY format
- Province/State - Province or state of the observation (Could be empty when missing)
- Country/Region - Country of observation
- Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised and so please clean before using it)
- Confirmed - Cumulative number of confirmed cases until that date
- Deaths - Cumulative number of of deaths until that date
- Recovered - Cumulative number of recovered cases until that date

There are also another 3 files that reordered the rows and columns of the above mentioned dataset file into time series based information of various provinces' and countries' cases numbers in different dates.

The package we mainly used for graphing is pyecharts and matplotlib.

All the Jupyter notebook source codes are posted on github.[1]

2 Visualization

2.1 China cases

2.1.1 Pyecharts Analysis

From Figure 1, we can roughly tell the severity of every province due to its Concerned population. Clearly, many provinces closely around the Hubei province are greatly spread. So in a great possibility, Hubei is highly believed as the center of COVID-19.

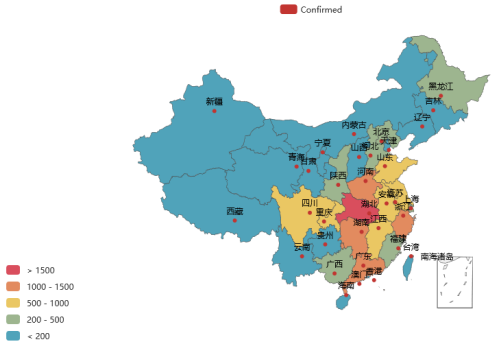

Fig. 1. China Confirmed Map

Figure 2 shows the proportion of confirms and deaths in all confirmed people, which changes along the date.

To show the trend of spread, we use line to depict the change of the existing confirmed number. We can clearly see from the Figure 3 that existing confirmed number is getting fewer and fewer,which means the situation in China is better now. But the graph also shows that there is an outbreak around 02/13. To see it clearly, we paint a line about daily

[1] https://github.com/blankzhang1/NCOV-Visualization

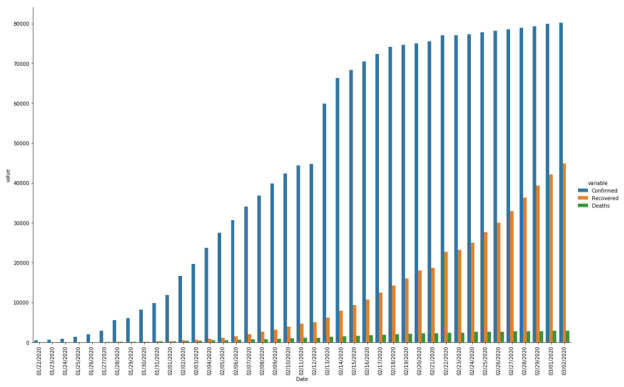

Fig. 2. China Bar chart

increase number of confirmed ones. Since in the dataset, the confirmed number is cumulative, we get the number of confirmed cases still exist by the following formula:

$$Existing = Confirmed - Recovere - Deaths \quad (1)$$

Fig. 3. Existing Confirmed chart

In figure 4, from 01/22 to 01/27, the line is smooth. On 01/28, however, the number is getting out of control, so we say that this is the first burst point. The second burst point is on 02/13, which is even worse. Coincidentally, the time interval between them is calculated to be about one incubation period. So the main reason of the sudden outbreak in middle February might be the end of one incubation period that begins in 01/28.

2.1.2 Province scatter plots

To have a closer look at the provinces in China, we use scatter plots to see the relationships among confirmed, deaths, and recovered cases. Hubei province is ignored as a special case. Figure 5 shows that the confirmed and recovered numbers of China provinces are relatively linear.

In addition, from the relationship between confirmed and death numbers, we can find out the seriousness of each province using cluster. In figure 6, we applied K-means clustering with 3 cluster centers, bigger confirmed and deaths numbers represent a more serious condition.

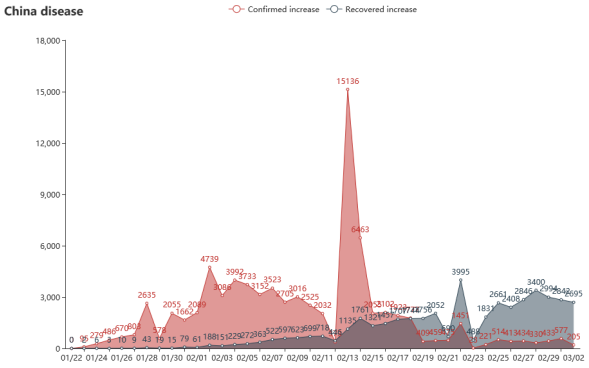

Fig. 4. Increase number

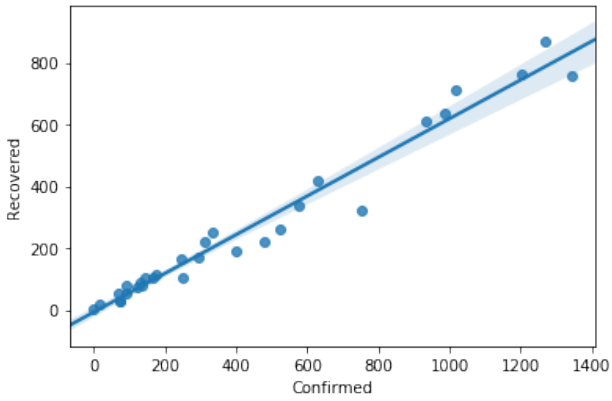

Fig. 5. Confirmed vs. Recovered in China Provinces

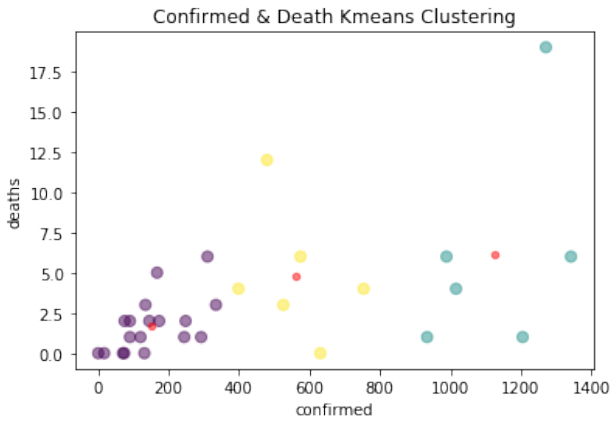

Fig. 6. K-means clustering

We can also see the 3D scatter plot of all these three features together in figure 7. Since one more feature is added into consideration, we apply four cluster centers for illustration.

2.2 World cases

From Figure 8, we can roughly tell the severity of every Country due to its Concerned population. Clearly, coronavirus has spread around the world and China is still the most

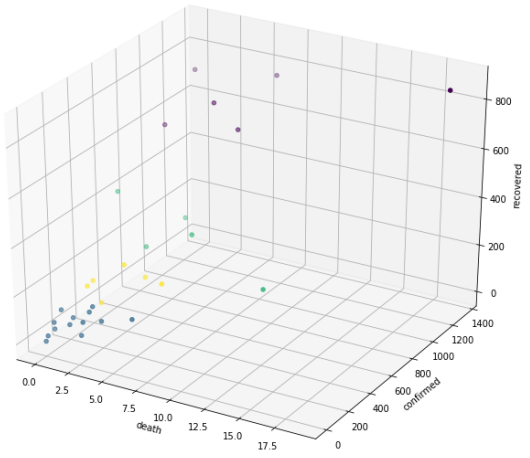

Fig. 7. 3D K-means clustering

serious Country in regard to the virus.

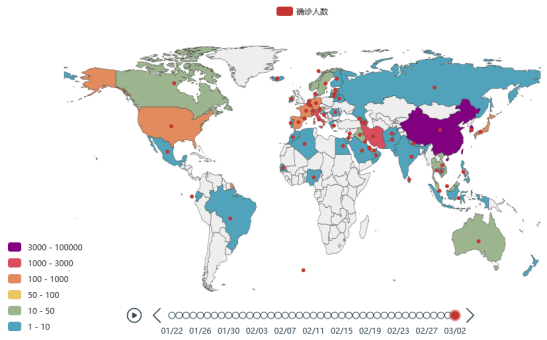

Fig. 8. World Confirmed map

As some countries are too tiny in map, so word cloud is also an image technique to visualize the data. To avoid the large scale of China, we just cancel data of China.

Figure 9 is a word cloud which represents the confirmed numbers in different countries. The bigger the font is, the more confirmed people there are in the country. From the figure, we see Korea, Iran, and Italy are large in shape, which means nCoV is now invading in and threatening those countries.

From Figure 10, by comparing in confirmed populations, we can clearly conclude how horrible and contagious the virus is.

2.3 Hubei Province Vs South Korea

As data tells us, Hubei is the burst center of nCoV-19. Now, from the Figure 11, fortunately, by Chinese government's efficient controls, virus spread seems to slow down. New Recovers have overwhelmed new Confirms eventually in late February.

As Figure 12 shows, South Korea is experiencing burst of nCoV like the situation of Hubei a month ago. If not

Fig. 9. Word cloud

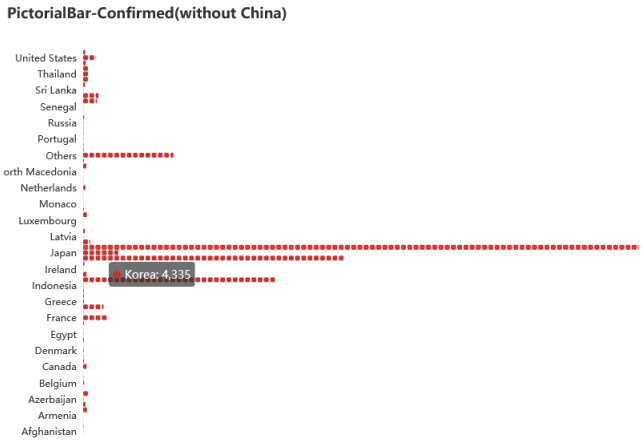

Fig. 10. World Bar chart

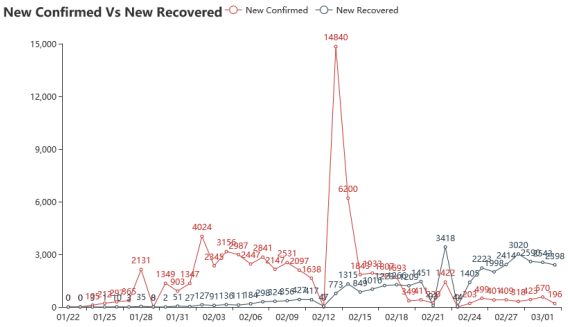

Fig. 11. Hubei Confirmed and Recovered growth chart

taking immediate measures, South Korea will ruins soon or later.

3 Prediction

3.1 China

To depict the trend of disease of China, we decide to use Polynomial Curve in figure 13, its shape is likely to go down after the peak.

$$y(x,w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \quad (2)$$

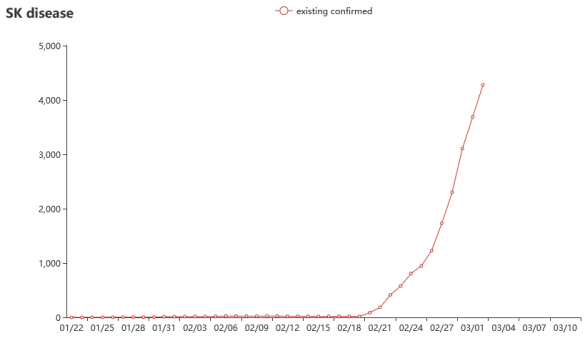

Fig. 12. South Korea Confirmed chart

It seems nCoV will stop around 03/10 in China if positively predicted.

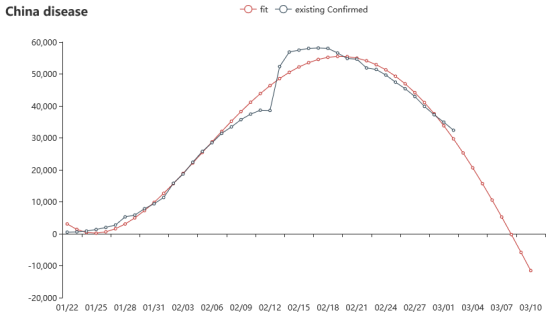

Fig. 13. China Polynomial Curve

3.2 South Korea

To fit the burst trend of NCOV in Sorth Korea, logistic curve might be a better choice. Hopefully, South Korea could control the virus soon so that the confirmed slope can decrease as shown in figure 14. The mathematical formula for this prediction is the following:

$$P(t) = \frac{KP_0e^rt}{K + (P_0e^rt - 1)} \tag{3}$$

3.3 Hubei

Since there are many aspects that can affect the condition of the virus, it is very difficult to predict in general.To give a insightful prediction, we train the dataset using Time Series Split. As shown in figure 15, we apply different regression methods such as Random Forest Regression, Bayesian Ridge Regression, and Logistic Curve. And we calculate the average performance of each prediction method in figure 16.

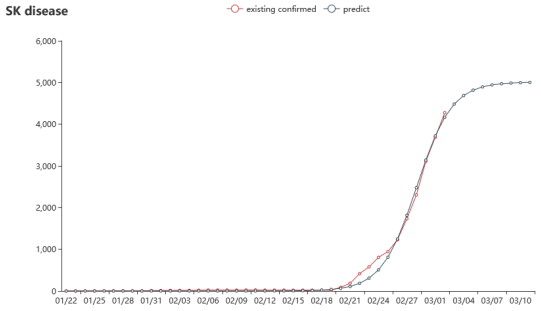

Fig. 14. South Korea Logistic Curve

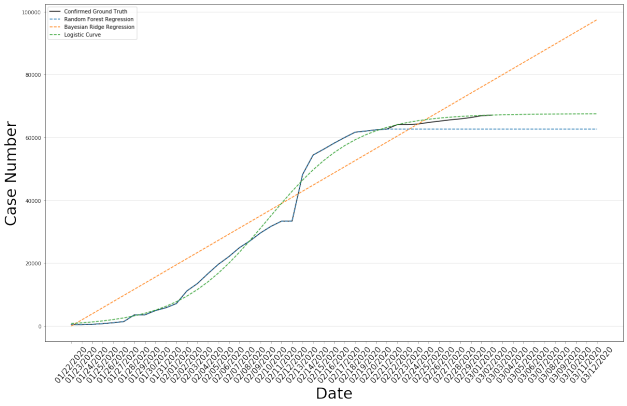

Fig. 15. Hubei Prediction

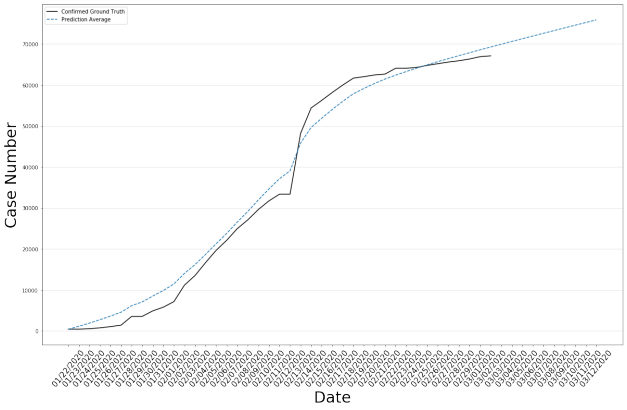

Fig. 16. Hubei Prediction Average

4 Conclusions

In the above analysis, We provide some insights to visualize the nCov-19 data of the world, China, Hubei, and South Korea. We can find that nCov is now wide-spreading all over the world. Luckily, we can see the sign that virus is gradually getting controlled. There is definitely more to explore in the dataset, but due to the limited information given to us, we cannot know what may affect these numbers. The future of the virus condition is determined by many potential factors that this dataset does not provide, and such information is essential to give a more prudent analysis and prediction on the data. While China's epidemic situation is getting better and more patients are recovered, South Korea, Iran, and Italy are

facing the burst of virus. If the local governments and people do not take serious concerns and actions, much more people will be infected.

5 Discussions

There are still some problems and challenges listed following.

1. We don't take incubation period into account, so the prediction seems too simple. Of course, our data for prediction is too limited.
2. A new dataset of patient's infomation, including symptom, traffic record, is recently updated on Kaggle, which is more helpful to make meaningful analysis and prediction.