# **Covid-19 Data Analysis Project Report**

Di Jin

**Motivation**

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. It started from **Wuhan, China** in **December 2019.** Which later declared as **Pandemic** by **WHO** due to high rate spreads throughout the world. Pandemic is spreading all over the world. Today (4/20), it has led to a total of more than 2.4 million confirmed cases and 170k+ death in the world, and it becomes more important to understand about this spread and the distribution in the world. This project is an effort to analyze the cumulative data of confirmed, death, and recovered cases over time. It also focuses on the distribution of cases in the U.S.

Four Overarching Questions:

1. Confirmed, Deaths, Recovered Cased Over Time in The World and Different Countries.
2. The distribution of cumulative confirmed, death, recovered and active cases in the world.
3. Cluster US states by confirmed, death and recovered cases.
4. The relationship between confirmed cases and the number of residents, confirmed cases and the party of governor in the U.S.

**Data Source**

The data source comes from Kaggle [Novel Corona Virus 2019 Dataset](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset) “covid\_19\_data.csv”. (<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>). This dataset has daily level information on the number of affected cases, deaths, and recovery from 2019 novel coronavirus. This is a time series data and so the number of cases on any given day is the cumulative number. The dataset is updated every day, and the range of date for this project is from 22 Jan to 14 April 2020.

This is the column description (from Kaggle), It has 8 columns and 15130 rows.

* Sno - Serial number
* ObservationDate - Date of the observation in MM/DD/YYYY
* 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 till that date
* Deaths - Cumulative number of of deaths till that date
* Recovered - Cumulative number of recovered cases till that date

**Data Cleaning**

This dataset is clean. Because it was scripted by humans and update every day. Only Province/State has the missing values, I think it is ok for this dataset. because some countries do not count the number of cases in Provinces/State level. In addition, I will only use this column for U.S. which does not have missing value.

I also change the column ‘ObservationDate’ to timestamp type because I want to analyze the change over time.

**Q1: Confirmed, Deaths, Recovered Cased Over Time in The World and Different Countries.**

I would like to analyze how confirmed, deaths and recovered cased change over time in the world, and I also want to pick China and U.S. as two examples to compare the how virus spread in China and U.S.

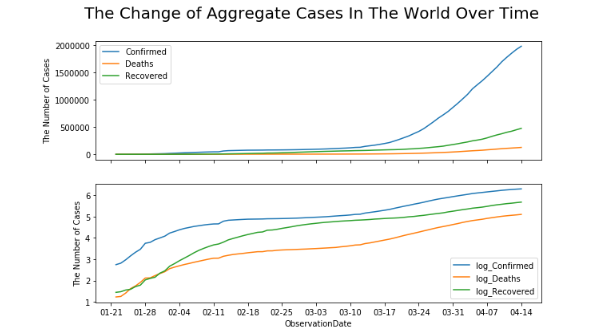
* Method

The method I used is line chart with the change of time. I need group each date and count all the cases for the same day either in the world or specific country. I plot three lines with different colors for each graph to represent confirmed, deaths, and recovered cases. I also transform the number to logarithm, and it is better to compare the rate of change and growth rate.

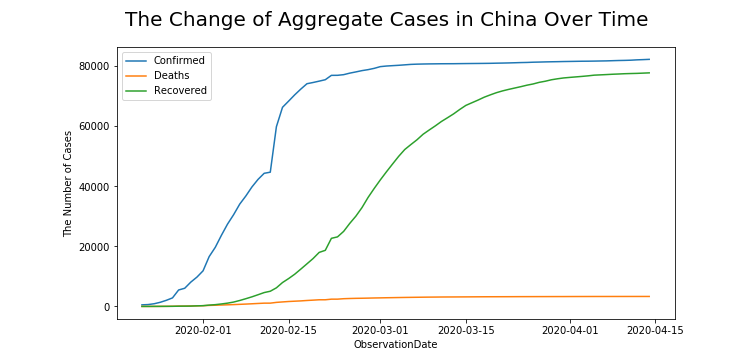
* Libraries and Techniques

Seaborn/Matplotlib

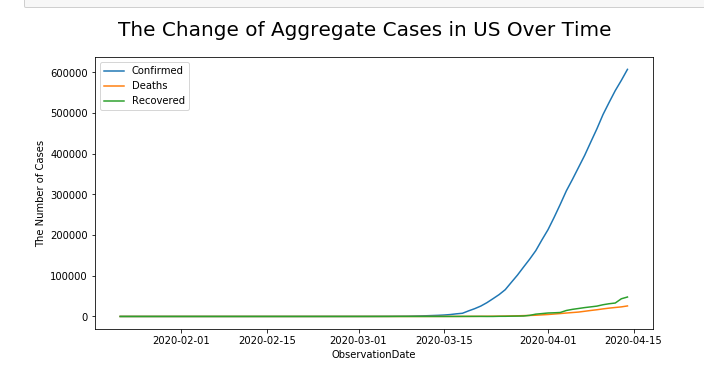
* Data Analysis and Visualization



The world graph shows that before March, the confirmed cases are less than 500000, I think most cases are in China at that time. But after March, both confirmed cases and deaths cases have rapidly increased since March.

The Logarithmic plot also shows that the situation is very serious all over the World from March, It is the result of the virus has started to grow outside of China. The slope of log\_confirmed is large from Jan to the beginning of Feb. I think it is due to the outbreak of the virus in China especially in Wuhan. 

The confirmed and deaths cases in China has a slow growth since March. The confirmed cases of around 80000 does not change a lot. It could be due to the improvement of situation. It is the good news for China. Another good news is that the line shows that the recovered cases in China increased a lot from March. The line of recovered is close to the line of confirmed cases. Actually, there are few increasing confirmed cases in China now.



The situation in the US is very concerning. The increase from 3/17 is significant, I guess it is the result of more cases getting testing for the first time. After 3/17, the rate of growth looks like to remain high. There are more than 600000 confirmed cases in the U.S.

The number of recovered cases in the U.S. is low, I think the actual recovered cases are higher than this data. I guess it could be due to the lack of the test for those are recovered. Some people can not get a test when they recovered. So, we could ignore this line.

**Q2: The distribution of cumulative confirmed, death, recovered and active cases in the world, which country has is the most infected.**

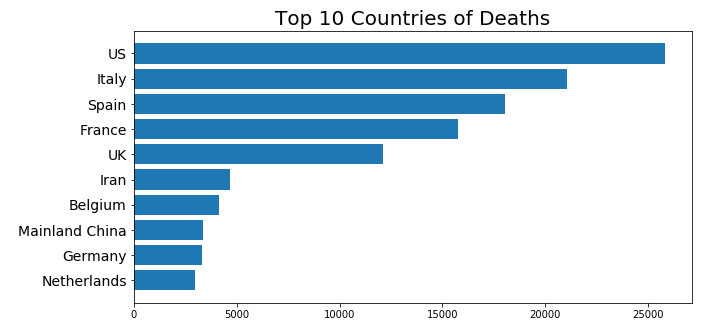
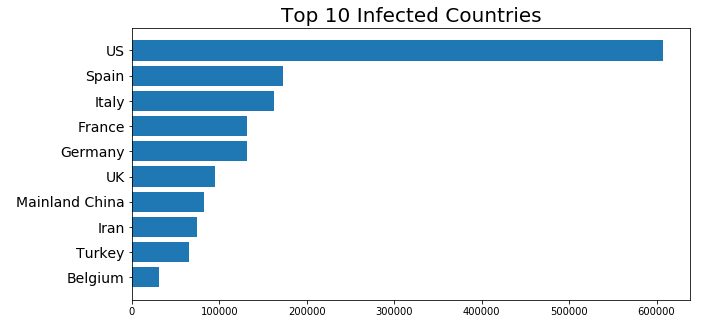
* **Method**

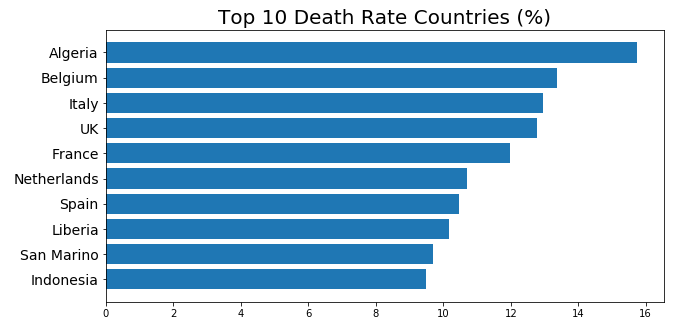
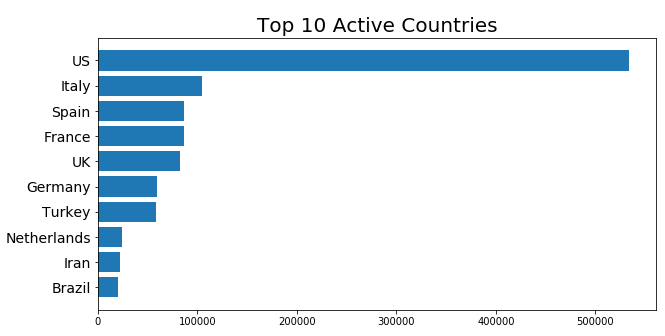
The method I used in this question is bar chart. I would like to plot the distribution of different types of cases in the world and choose the top 10 country to show. Because this dataset is cumulative data, I choose the data of latest date. I also group the country because some countries count the number by states. I also calculate the death rate and active cases for each country.

* **Libraries and Techniques**

Seaborn/Matplotlib

* **Data Analysis and Visualization**

****

****

The four visualizations above show that the top 10 countries of each of four factors. The first bar chart shows that the U.S. is the most infected country in the world, and the gap between the U.S. and other countries is also large. Spain is the second infected countries, but the number of confirmed cases is less than 200000. Another interesting point is that there are 7 European (I count Turkey as part of Europe) countries in the top 10, Which means the situation in Europe is severe now.

There are 9 countries of the top 10 infected are also in the top 10 countries of deaths. The U.S. is still the country has the most deaths.

For the death rate, I extract the countries which have more than 50 confirmed cases to calculate its death rate. Algeria became the top 1 country, and another African country Liberia also in the top 10 countries. The data shows that the medical condition is not optimistic in Africa. The death rate of the U.S. is not in the top 10 list, which is good news.

**Q3: Cluster U.S. states by confirmed, death and recovered cases**

In this question, I want to cluster U.S. states by confirmed, death, and recovered cases. It is interesting to know which states has the

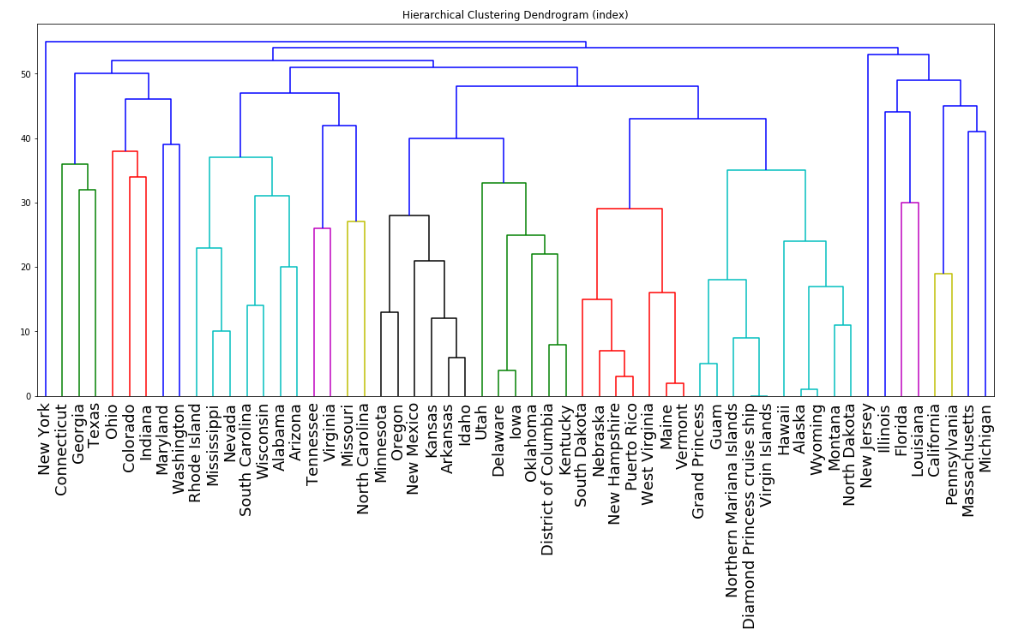
* **Method**

The method I used in this question is clustering. Because U.S. does not count recovered cases, and I would like to cluster U.S. state by the number of confirmed and death cases. I group the states and try to use two cluster technique (Agglomerative and Kmeans) to do clustering.

* **Libraries and Techniques**

sklearn, scipy

* **Data Analysis and Visualization**

****

I extract U.S. data and group by state. This dataset has more than 50 states because it counts Guam and cruises as a state. For both agglomerative and K-means models, I decide to use 5 as the number of clusters. Because I think we could cluster the severity as 5 levels.

The graph above is the hierarchical clustering dendrogram of the agglomerative method. I think the clustering works well. We all know that New York and New Jersey are the two most severe states in America, and this dendrogram also puts New York as one cluster and put New Jersey as one cluster.

The result of K-means is similar to the result of the agglomerative method. For example, CA, FL, IL, LA, MA, MI, PA are in the same cluster of the agglomerative method. These states are also in the same cluster as the K-means method.

**Q4: The relationship between confirmed cases and the number of residents, confirmed cases and the party of governor.**

There are some articles talked about why New York, Michigan has the serious condition. Some said that it is related to the number of residents, and some said that the governor of these states is democrat. So, I would like to do t- test to see whether there is significant difference between those factors.

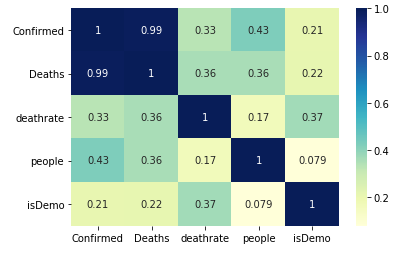
* **Method**

The method I used in this question is t-test and heatmap. I also use two html webpage to extract the number of residents and the party of governor.

* **Libraries and Techniques**

seaborn, scipy.

* **Data Analysis and Visualization**

****

After extracting the information from the two webpages, I also did some data cleaning for the number of residents and the party of governor. I create a column ‘isDemo’. The value of 1 represents the governor is democrat, otherwise republican.

The plot above is the heat map. We can see that the confirmed has a strong correlation with deaths. Because with the increase of confirmed cases, the death cases also increase. The coefficient of correlation between the number of people and confirmed cases is 0.43, which is not high, so I can say there is no relationship between the number of residents and confirmed cases. Also, the coefficient of correlation between isdemo and confirmed cases is just 0.21, there is also not too much correlation between the party of the governor and confirmed cases.

I also use t-test to test whether there are significant differences between the party of the governor and the number of confirmed cases. The P-value of t-test is around 0.13, which is large than 0.05, 0.01. So, I cannot reject the null hypothesis, they are not statistically significant.

Reference:

<https://www.kaggle.com/holfyuen/coronavirus-2019-20-visualization/data>

<https://www.kaggle.com/imdevskp/covid-19-analysis-visualization-comparisons/data>

<https://www.kaggle.com/abhinand05/covid-19-digging-a-bit-deeper#2.-Confirmed-Deaths-Over-Time>