Final Project Report

Keyu Chen (kc1044)

4/15/2020

1 Introduction

This project gives insights on the worldwide spread of **novel Coronavirus** (also known as **COVID-19** of which "Corona" is "Co", "Virus" is "vi", and "Disease" is "d" and "19" stands for the year "2019"; whereas, "SARS-CoV-2" being the virus that causes the disease). It is a contagious respiratory virus, outbreak of it was first identified in Wuhan, Hubei, China in December 2019, and was recognized as **pandemic** by the **World Health Organization (WHO)** on Mar 11, 2020. As of April, more than 3,624,470 cases of **COVID-19** have been reported in more than 212 countries and territories, resulting in more than 250,986 deaths and more than 1,179,863 recoveries.

Coronavirus is now the must-have loudest topic around the world. It is a new virus to human beings and it caused a lot of confusion in the world, by its rapid spread and changed the lives of bilions people. A huge amount of people has died, lose their families & friends and lose their jobs because of it. Most industries have shut down because people have to stop to work and meet each other to avoid the spread of the virus. Because the virus is brand new to everyone, no one knows the definite effective measures to fight the virus. As the result, what human beings can do is to try some methods and explore the effective ones and stick to it. The measures that countries taking changed a lot as time goes, fortunately, the spread in some countries has been controlled.

2 Data Discription

The data used in this project is from a dataset on **Kaggle** website (https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset). The dataset is extracted from **COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University** (https://github.com/CSSEGISandData/COVID-19).

Thanks so much to **Johns Hopkins University** for making the data available for educational and academic research purposes.

The main data file analysized in this project is `covid_19_data.csv`. The detailed description of the dataset is below:

- 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

3 Data Preprocessing

3.1 Data Cleaning

After importing the worldwide dataset into a data frame, I modify some column names to easy-to-use ones, delete some columns that will not used in this project and transfer `ObservationDate` to dates formats which saved as variable `Date`. In order to focus on main affected countries, I add a variable `Rank` to the dataset, which numbered the rank of confirmed cases of countries.

3.2 Data Files

I create two data frames in this part. The first one is the total number of affected cases, deaths and recovery in every country or region till now and the rank of confirmed cases of country. The second one shows the the change in confirmed cases, deaths and recovery over time at country level over time. The two data frames are saved into `.csv` files, separately.

The first five rows of the two data frames are showed below:

```
## # A tibble: 6 x 5
##   Country.Region Confirmed Deaths Recovered  Rank
##   <fct>              <dbl>  <dbl>     <dbl> <int>
## 1 US               1103461  64943    164015     1
## 2 Spain             236899  24543    132929     2
## 3 Italy             207428  28236     78249     3
## 4 UK                178685  27583       892     4
## 5 France            169053  24628     51124     5
## 6 Germany           164077   6736    126900     6
```

```
## # A tibble: 6 x 6
## # Groups:   Country.Region [1]
##   Country.Region Date       Confirmed Deaths Recovered  Rank
##   <fct>          <date>         <dbl>  <dbl>     <dbl> <int>
## 1 US             2020-01-22         1      0         0     1
## 2 US             2020-01-23         1      0         0     1
## 3 US             2020-01-24         2      0         0     1
## 4 US             2020-01-25         2      0         0     1
## 5 US             2020-01-26         5      0         0     1
## 6 US             2020-01-27         5      0         0     1
```

4 Data Analysis and Visualization

4.1 Comparison of 11 Main Affected Countries

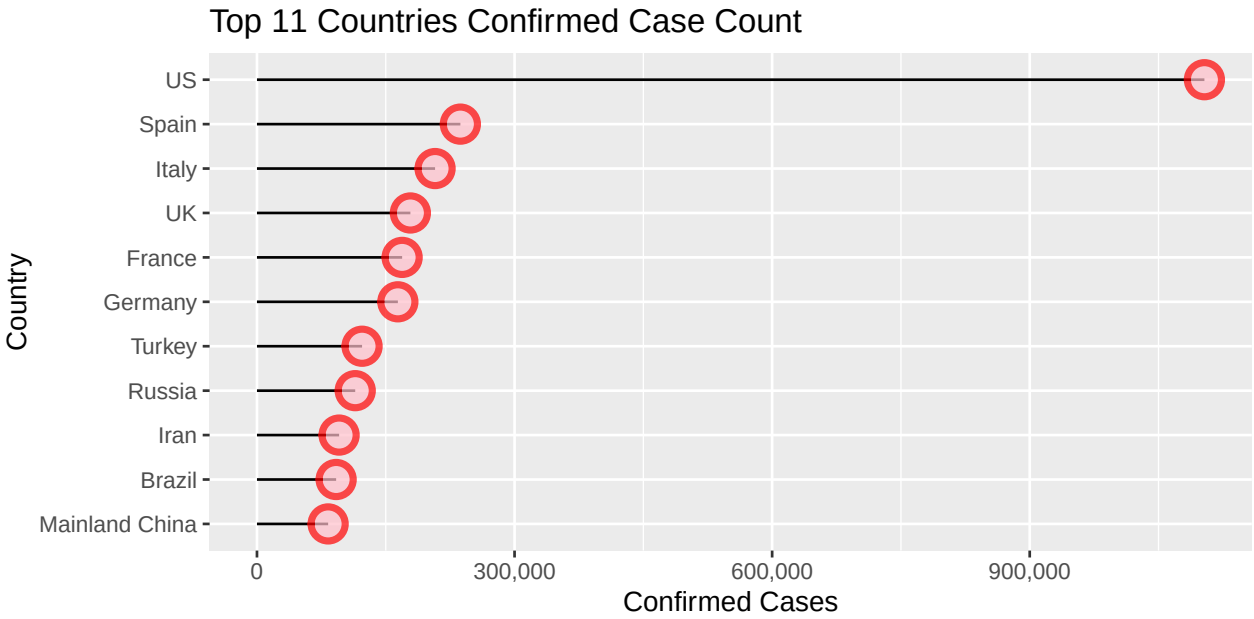
In this part, the data of 11 main affected countries are selected to compare. The reason I choose 11 instead of 10 is that my home country China is ranked 11 and I would like to include it in my analysis.

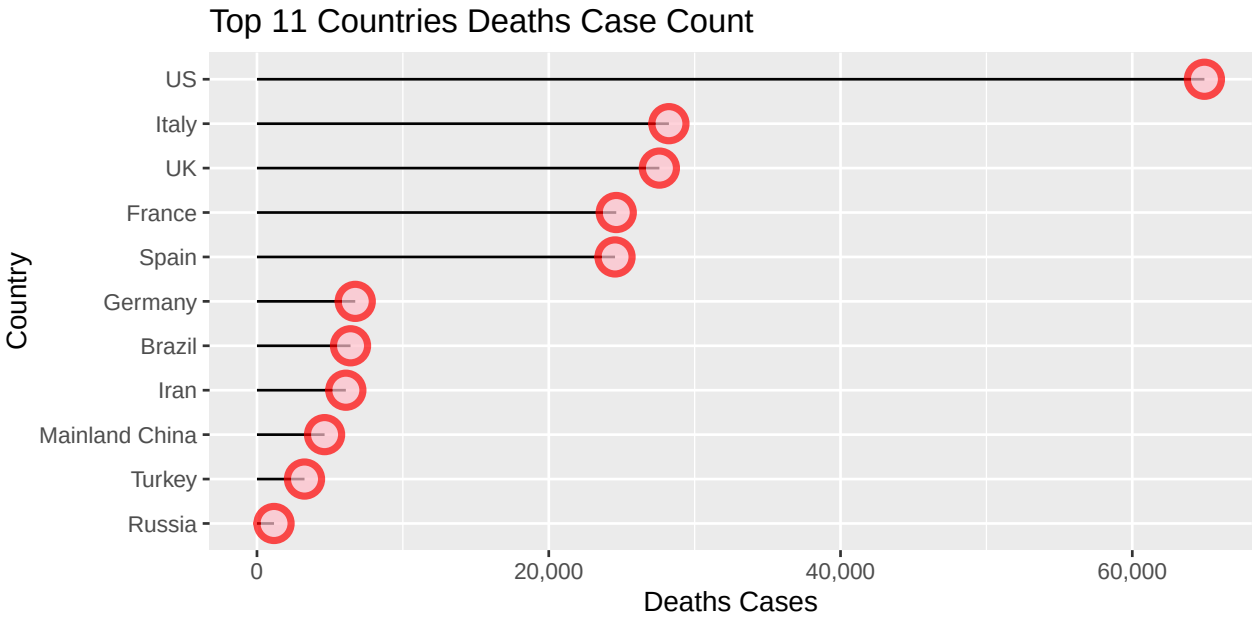
Here are the number of confirmed cases, death cases and recovered cases of the top 11 countries:

```
## # A tibble: 11 x 5
##    Country.Region Confirmed Deaths Recovered  Rank
##    <fct>              <dbl>  <dbl>     <dbl> <int>
##  1 US               1103461  64943    164015     1
##  2 Spain             236899  24543    132929     2
```

```
##  3 Italy             207428  28236     78249     3
##  4 UK                178685  27583       892     4
##  5 France            169053  24628     51124     5
##  6 Germany           164077   6736    126900     6
##  7 Turkey            122392   3258     53808     7
##  8 Russia            114431   1169     13220     8
##  9 Iran               95646   6091     76318     9
## 10 Brazil             92202   6412     38039    10
## 11 Mainland China     82875   4633     77900    11
```

Then two simple plots to display the number of confirmed cases and deaths for the top 11 countries:



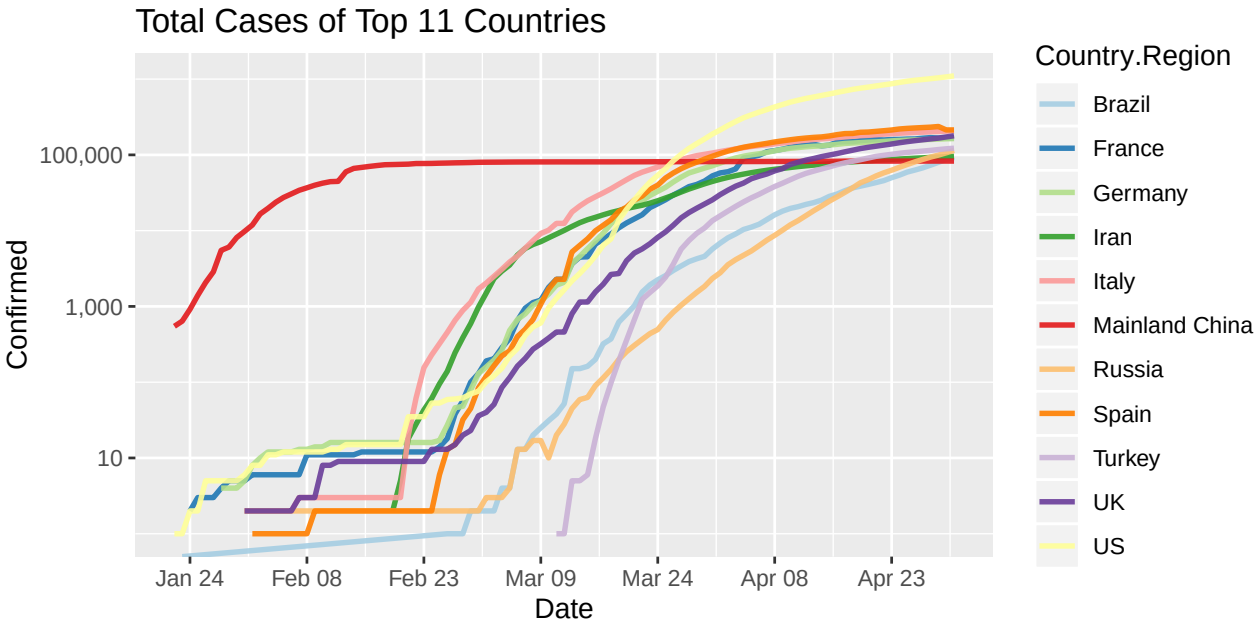


As we can see, the confirmed cases and deaths are way more than the other countries, even several times than other countries. Among the Europe countries, the death rate of Germany is the lowest.

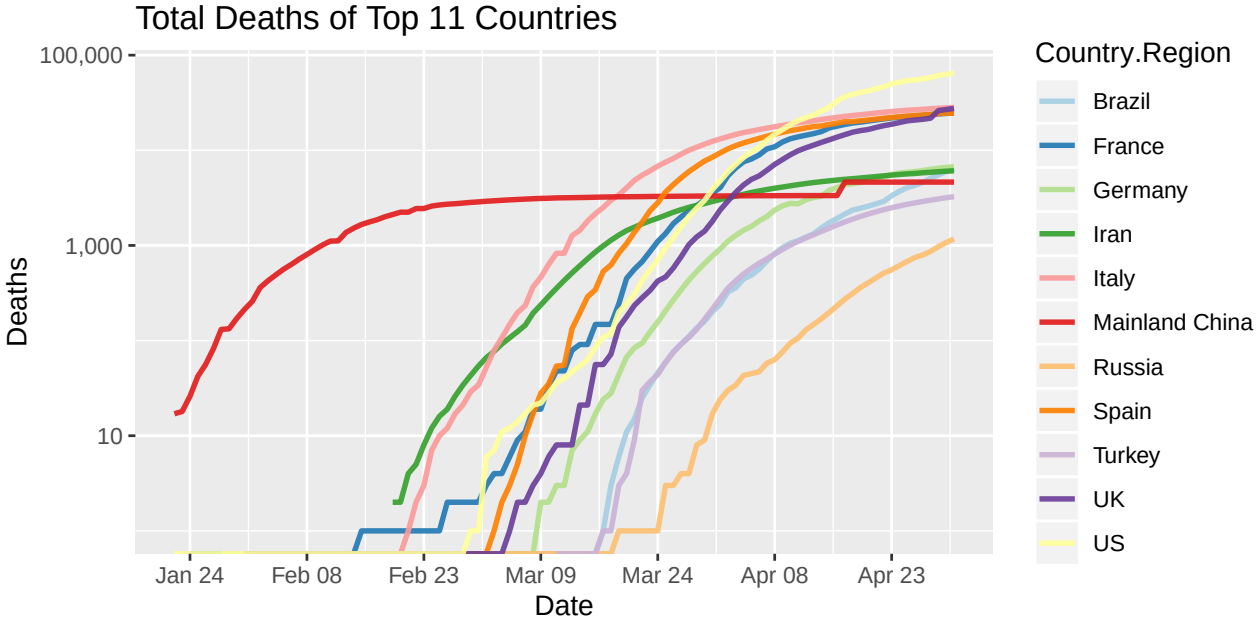
In the three plots below, I put the time series of confirmed cases, deaths and recovery of the 11 countries and

see what can we find out from the plots. Here the logarithmic scale was used, because in this way we can better see the dynamics of increases and decreases that we would not observe using a linear scale.

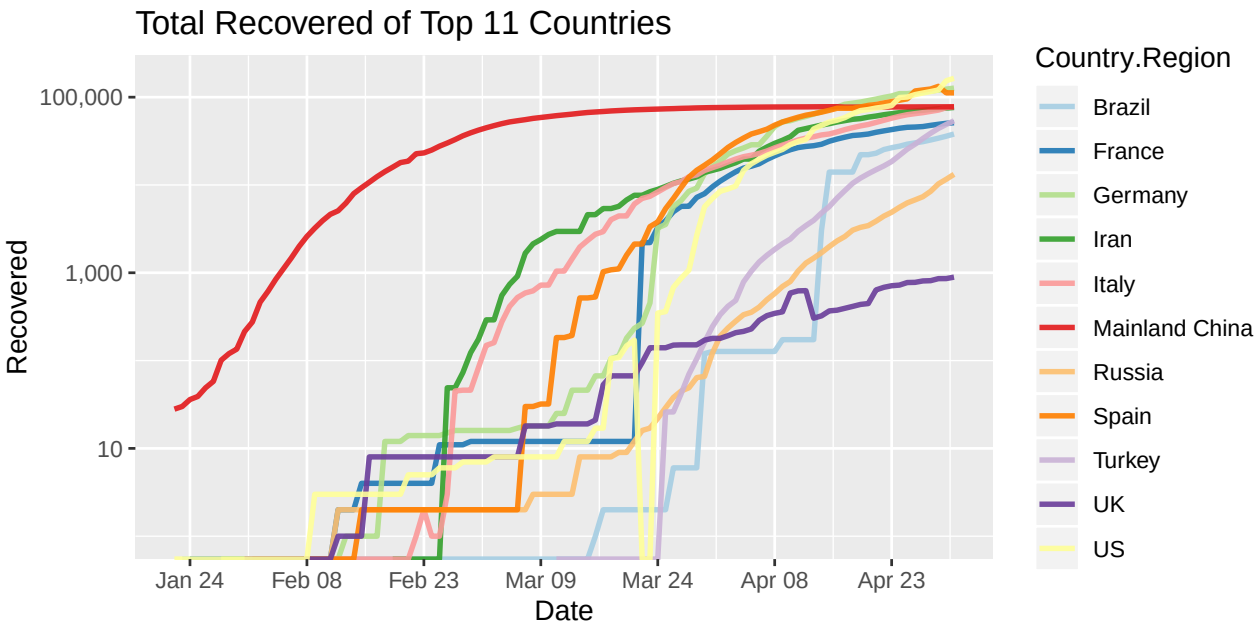
According to the plot *Total Cases of Top 11 Countries*, the United States has the most cumulative confirmed cases while Mainland China has the least cases among these 11 countries. The outbreak firstly happened in China and slowed down after the beginning of Feburary. Now the situation in China has been stable. Also, the increase speed of cases in every country except Russia has been slowed down a lot after April. Hopefully the cases can almost stop to increase in May.



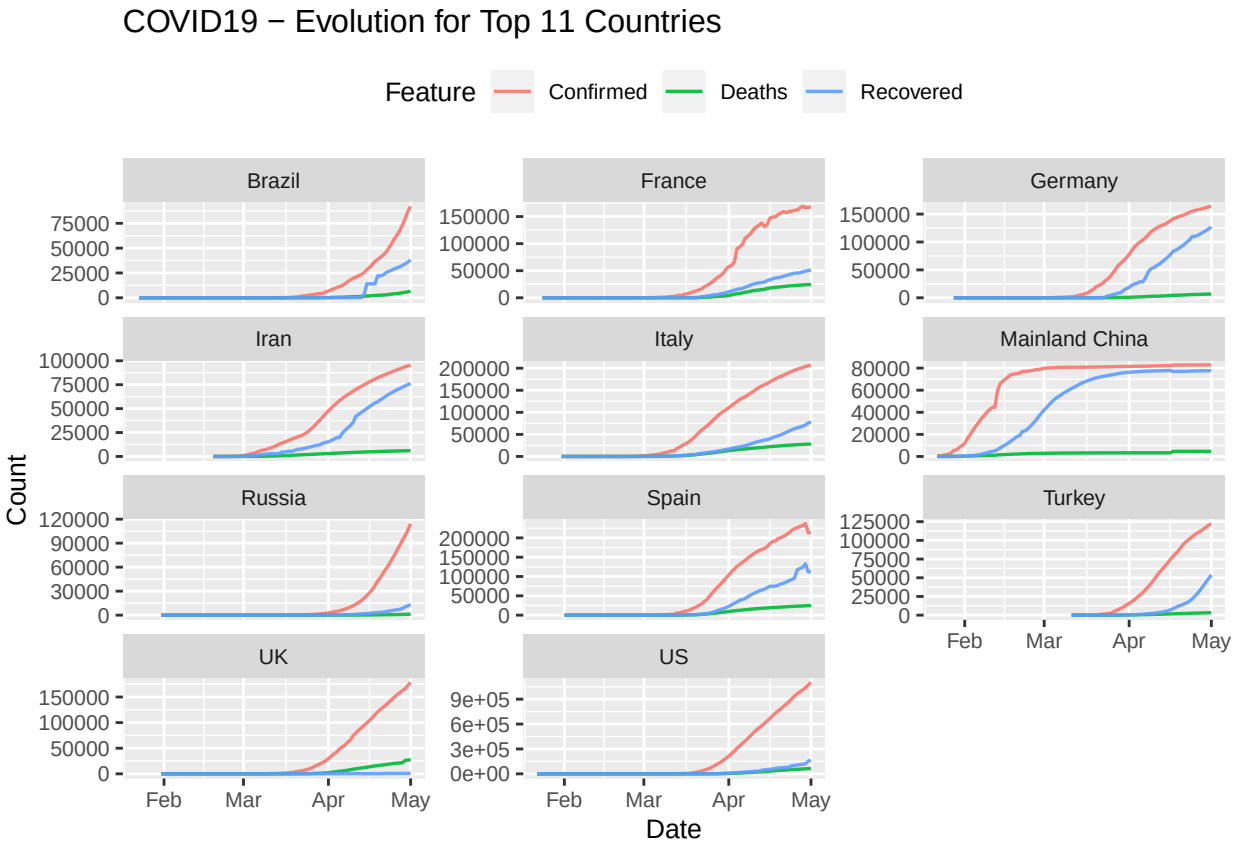
Based on plot *Total Deaths of Top 11 Countries*, currently the United States has the most death cases while Russia has the least deaths happened. The growth of deaths tend to stop in most countries among the 11, however, the increase speed of deaths in Brazil and Russia still has not obvious trend of decrease.



From plot *Total Recovered of Top 11 Countries*, the United States, Germany and Spain has the most recovered cases while the United Kingdom has the least.



In order to visualize the confirmed, recovered and death cases for each place, I made plots of the change trend of the three variables for every country. From the plots we can see that the most patients in Germany, Iran and Mainland China have been cured, while in other countries, the numbers of infected people are way more than the recovered cases.



Combining the four plots above together, we can tell that the most countries started to cure people effectively and the increase of new cases has been decreased. From the comparison, the most effective countries are Germany, Iran and Mainland China. Fortunately, among the top 11 countries, the increase of new cases

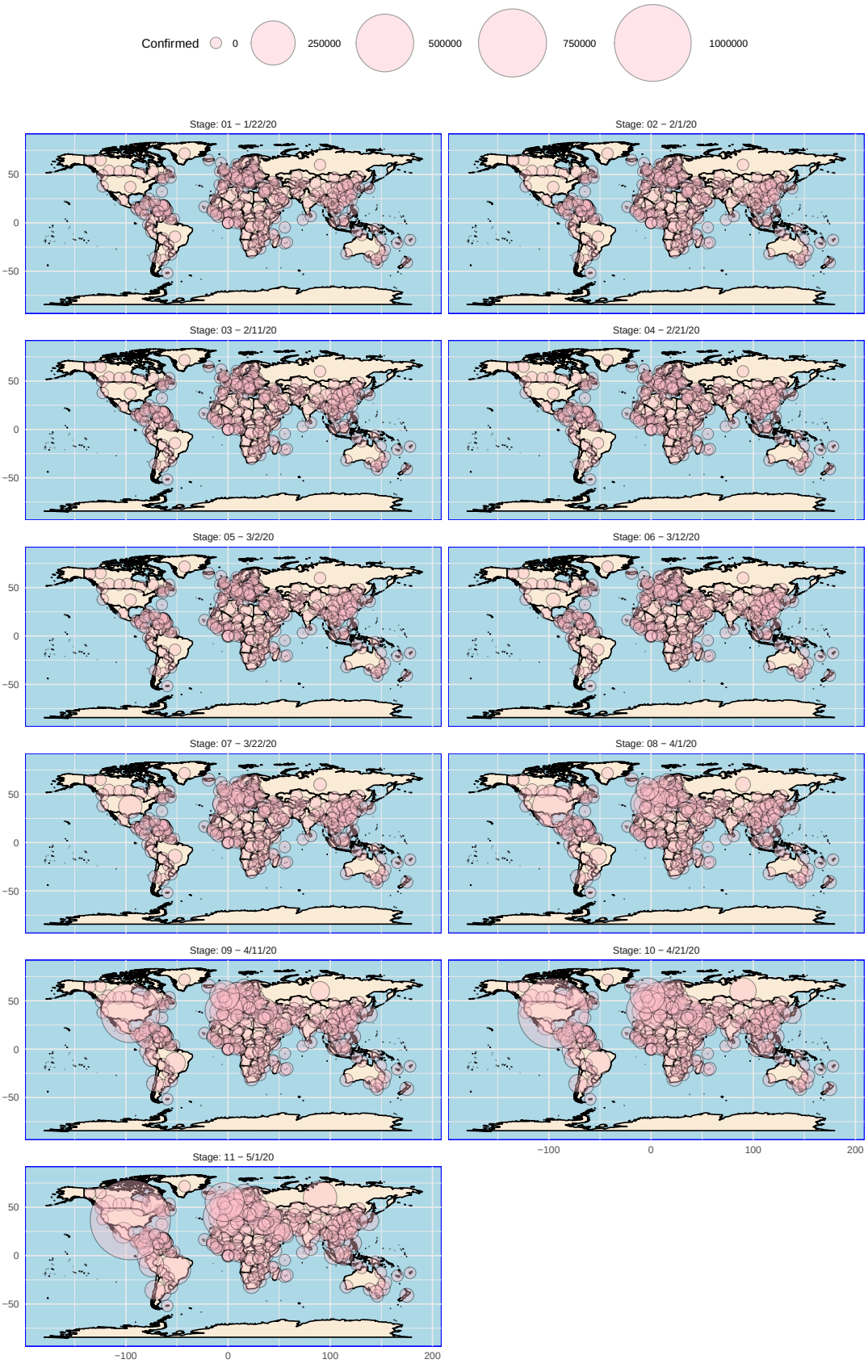
are decreasing and people cured are increasing, which will result in the successfully control of the spread. However, the situation in Russia and Brazil is still serious, the explosion of new cases is still there and the government should take more measures to control the infection growth, cure more people and stop the increase of deaths.

4.2 World Map of COVID-19 over time

For this part, I used a time series confirmed case dataset of the world `time_series_covid_19_confirmed.csv` and show the variation of the cases in different countries over time in a world map view. In order to show the change clearly in one page, I choose 10 day as the intervel. From Jan.22.2020 to May.01.2020, there are 11 maps shown below.

According to the maps, we can tell that almost every place in the world has been infected by COVID-19, and its infect speed is rapid, especially in the America and Europe (including Russia) area. The cirtles in these two areas obviously become bigger and bigger, which means the number of infected people is increasing rapidly. As for my home country China, even the whole area of Eastern Asia, the infected cases are also everywhere but the circles basically remain their sizes during this period, which means that the spread of COVID-19 has been controlled.

World Map − COVID19



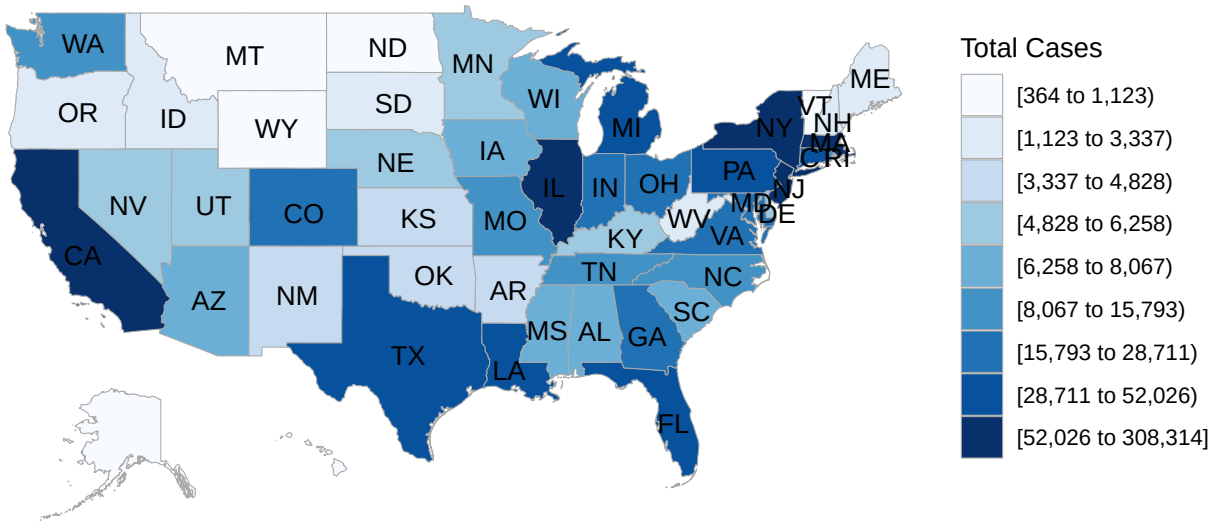
4.3 US COVID-19

Because the United States has the most confirmed cases in the world, in this part, I research more about the distribution of both infected cases and deaths in the US. Below are choropleth maps of confirmed cases and deaths in the US. We can see that the situation in New York and New Jersey is the most serious. The neighbor states of these two states are also under serious situation.

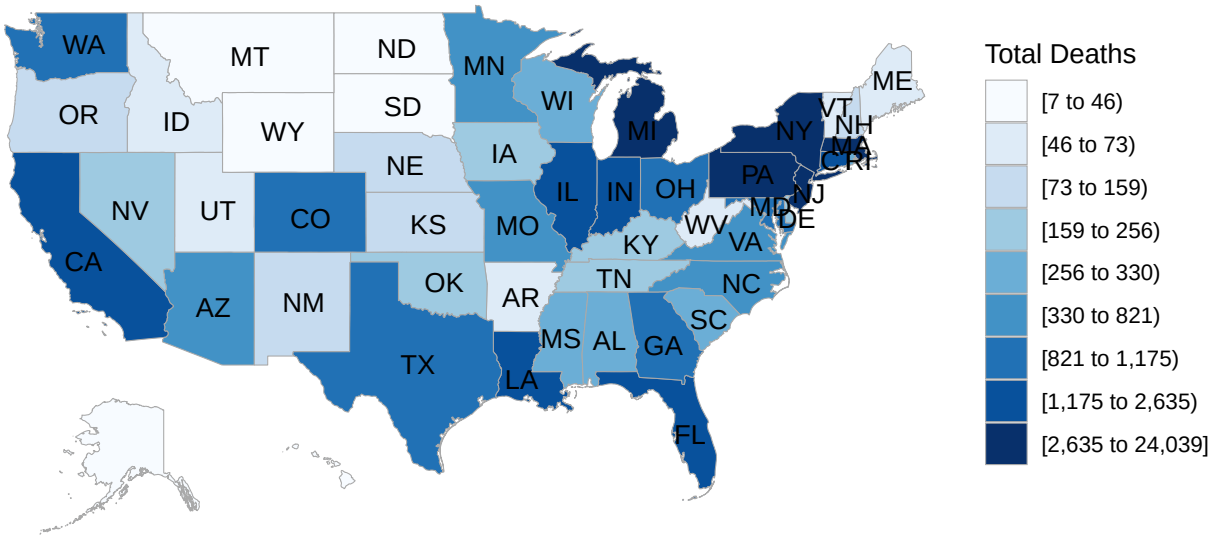
As for the western coast, California is the most serious. We can conclude that COVID-19 is easier to spread in city area because people lives closer in this kind of area and take more public transportation.

```
## # A tibble: 10 x 3
##    State         Death Confirmed
##    <fct>         <int>     <int>
##  1 New York      24039    308314
##  2 New Jersey     7538    121190
##  3 Massachusetts  3716     64311
##  4 Illinois       2457     56055
##  5 California     2126     52026
##  6 Pennsylvania   2635     49579
##  7 Michigan       3866     42356
##  8 Florida        1314     34728
##  9 Texas           840     29692
## 10 Connecticut    2339     28764
```

Confirmed cases in US



Death cases in US



5 Conclusion

Initially, rapid growth of COVID-19 was observed in China, which began to slow down around January, until growth was virtually unnoticeable in mid-Feburary. It seems that Asia has already gone through the worst period, and its rapid recovery from emipidemia raises hopes for other regions of the world. Since mid-Feburuary we have been observing a very rapid increase in the number of cases in other parts of the world, USA and Europe, which may mean that we are dealing with a similar situation as in China and the surrounding area with a delay of about a month. Around the turn of May one can notice a progressive slowdown in the growth rate of new cases in these groups. China and countries in the surrounding area are undergoing rapid stabilization, while in Europe and other parts of the world the number of victims is growing rapidly.

For the number of people cured of COVID-19 disease. Fortunately, we see continuous growth in all countries. The growth rate is lower in China, but it results directly from a smaller number of people infected. A sudden drop to zero in the US around March 20 is probably a database error. Among the top 11 countries, the increase speed of deaths in Brazil and Russia still has not obvious trend of decrease. The most countries started to cure people effectively and the increase of new cases has been decreased. From the comparison, the most effective countries are Germany, Iran and Mainland China. Fortunately, among the top 11 countries, the increase of new cases are decreasing and people cured are increasing, which will result in the successfully control of the spread. However, the situation in Russia and Brazil is still serious, the explosion of new cases is still there and the government should take more measures to control the infection growth, cure more people and stop the increase of deaths.

As the biggest infected country in the world, the United States, its confirmed cases, cure cases and deaths are still increasing quickly. Fortunately, the cure speed in going up and confirm case inscreasing speed is going down. Hopefully the patients will stop to increase in a month or so.

I left the states and come back to China on May 3, 2020. At the time I submitted the report, I am in quarantine in a hotel in China for two days and I heard that everyone on my flight has been tested negative. God bless us. I came back through JFK and ICN airport. I never saw the airports that empty. Even in Incheon airport, there was only one flight at the transfer building that whole day.Everyone at airports wore a mask and some of them even wore protective suit. We are experiencing something historically and memorable. I sincerely hope that we human beings can overcome this disaster altogether, Please don't panic, stay safe, follow the WHO, and your nation guidelines. We all can defeat this together.

6 Github Link

https://github.com/keyuchen96/stat597_final_project