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

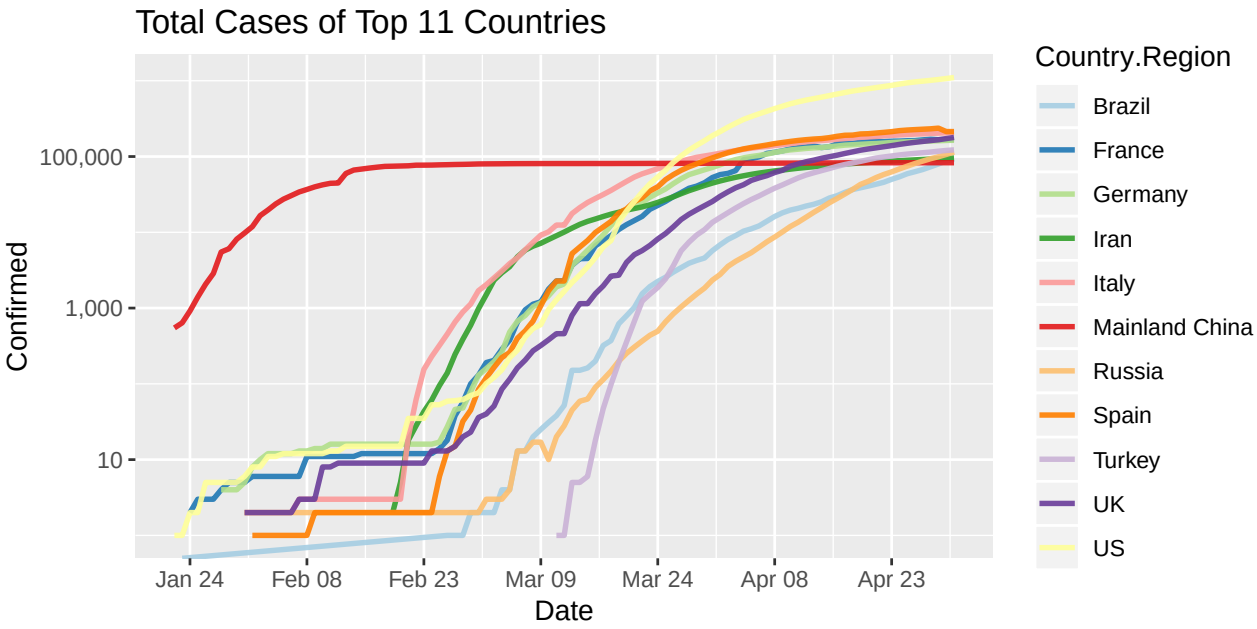
4.1 Comparison of 11 Main Affected Countries

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.

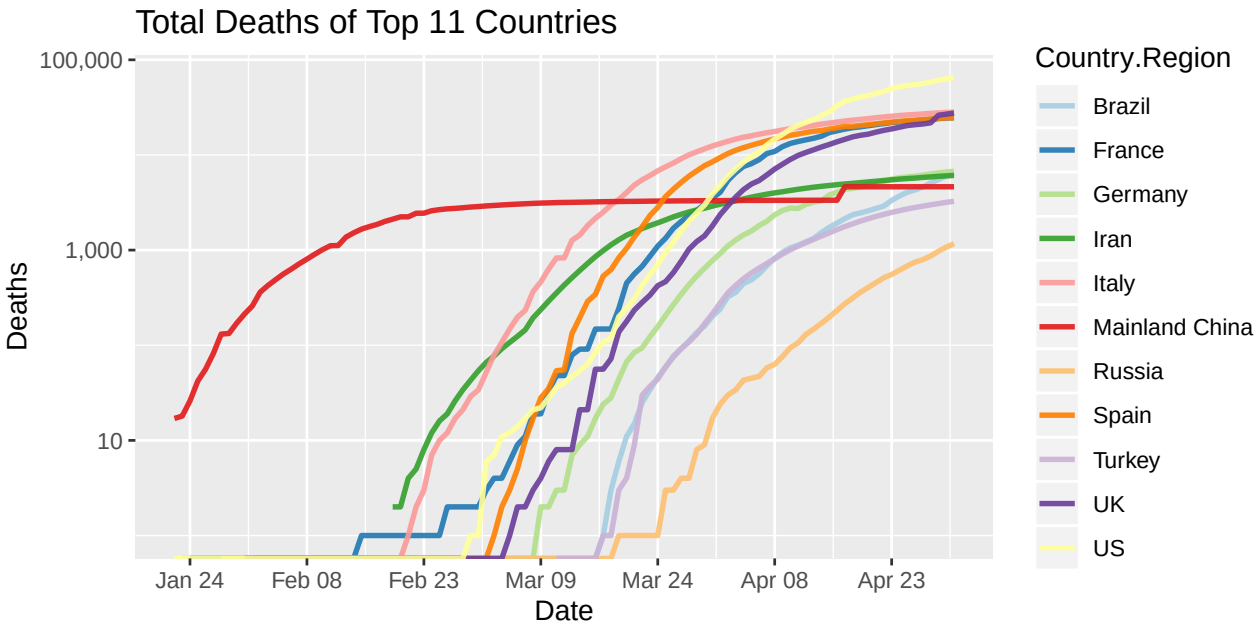
In the three plots below, I put the time series of confirmed cases, deaths and recovery of the 11 countries and see what can we find out from the plots. Here the logarithmic scale was used, because in this way we can better see the dynamics of increases and decreases that we would not observe using a 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,

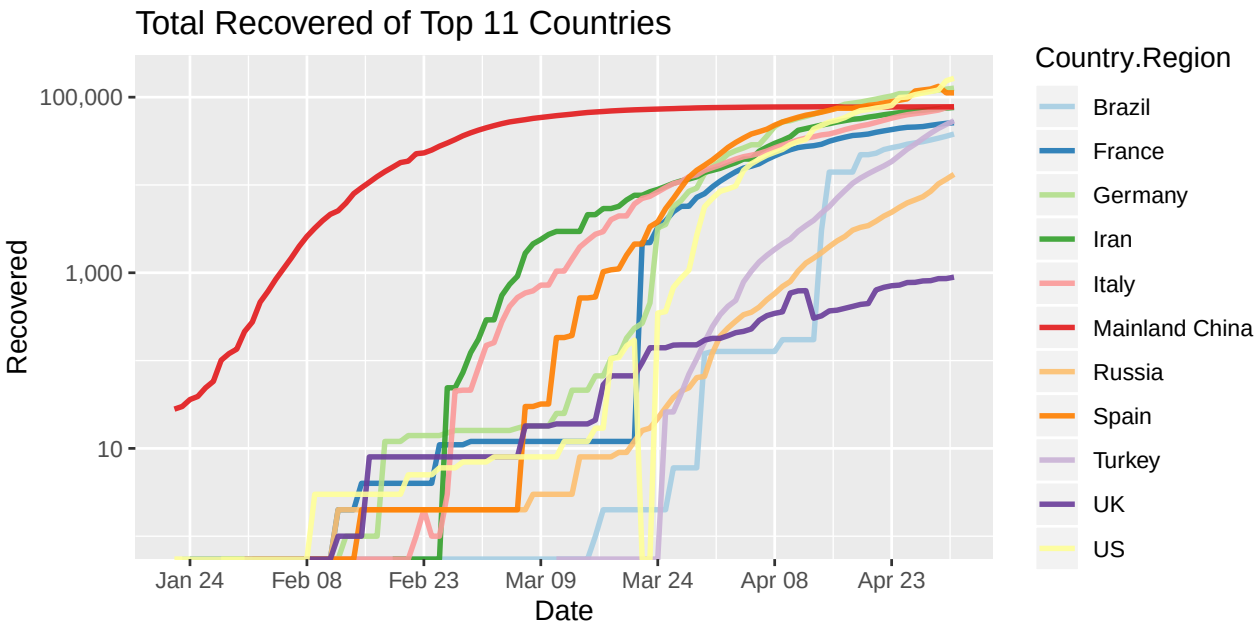
the increase speed of cases in every country except Russia has been slowed down a lot after April. Hopefully the cases can almost stop to increase in May.



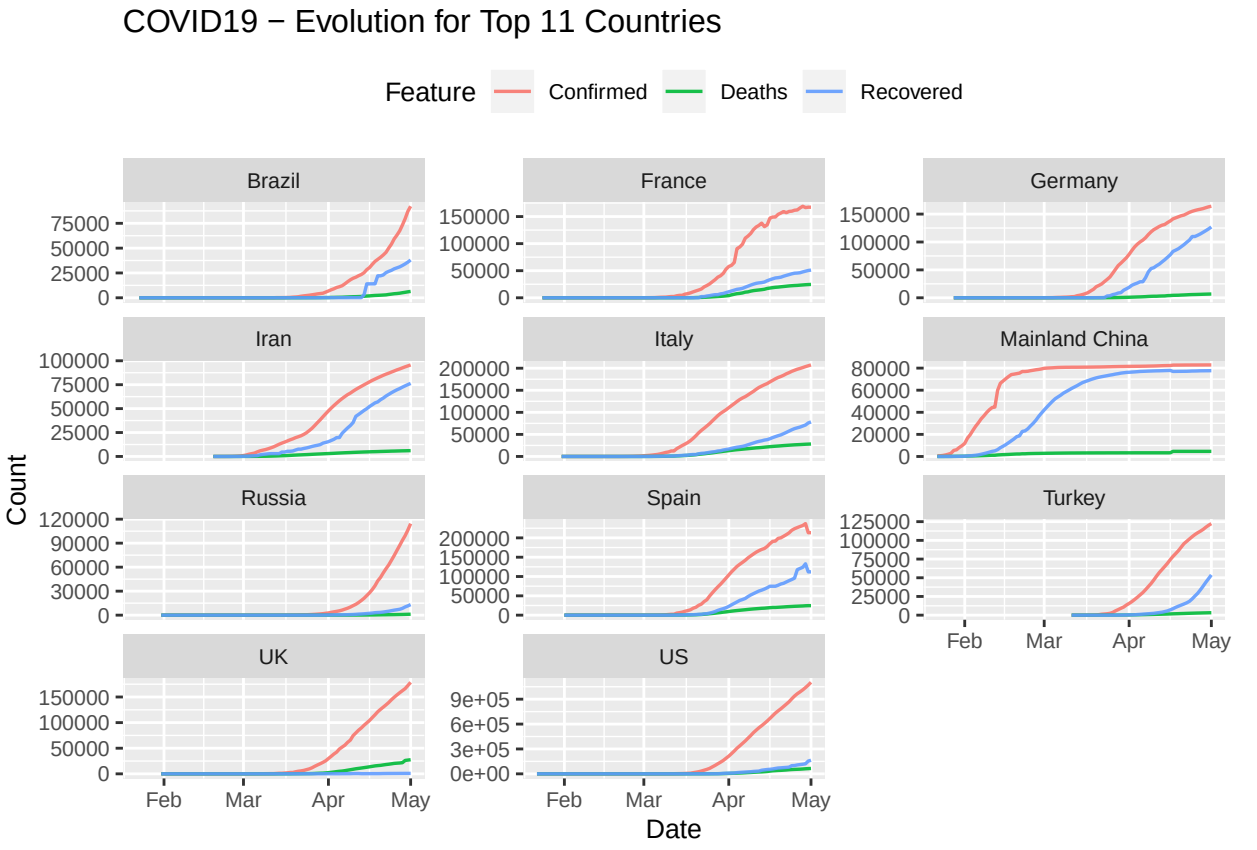
Based on plot *Total Deaths of Top 11 Countries*, currently the United States has the most death cases while Russia has the least deaths happened. The growth of deaths tend to stop in most countries among the 11, however, the increase speed of deaths in Brazil and Russia still has not obvious trend of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

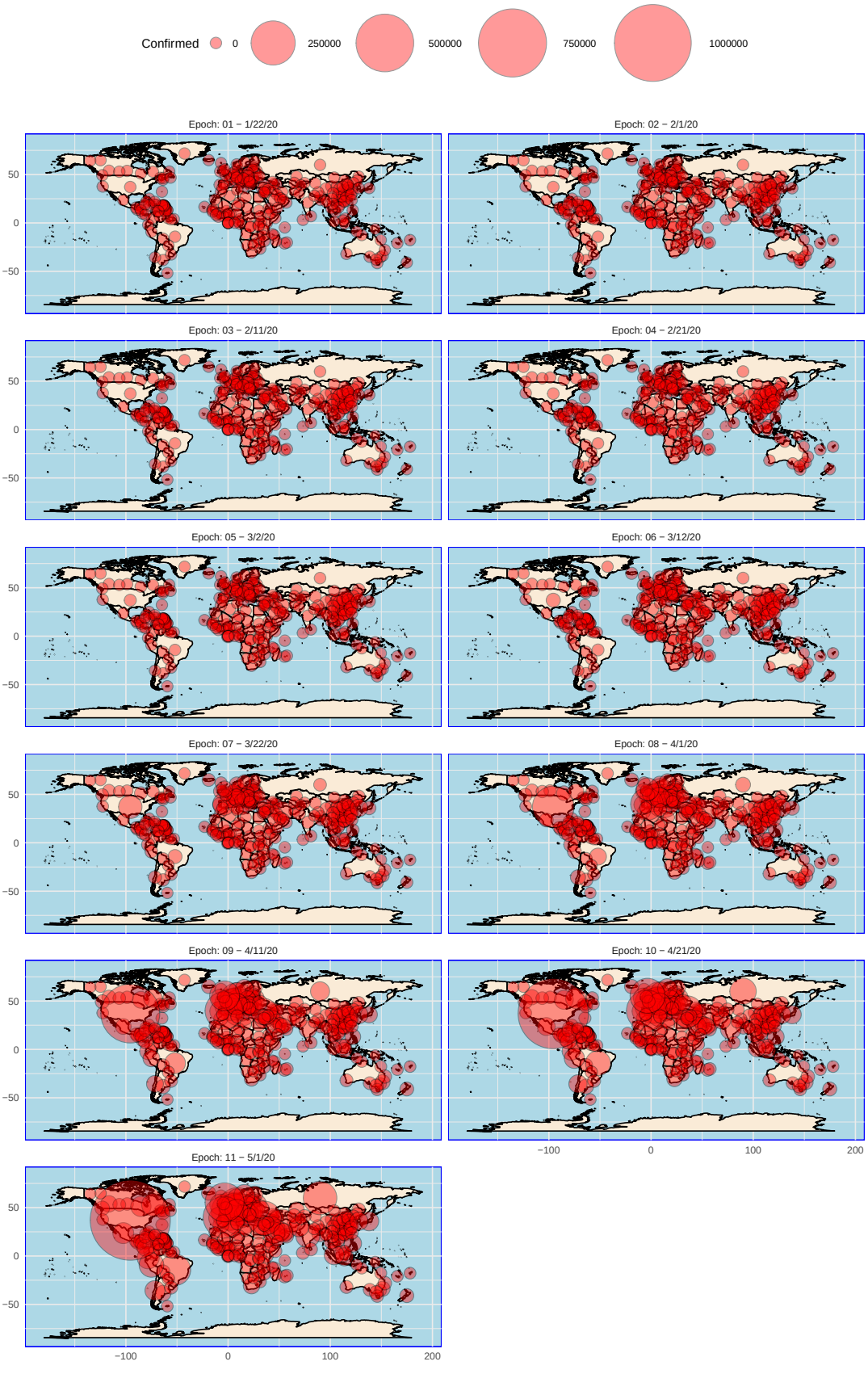
are decreasing and people cured are increasing, which will result in the successfully control of the spread. However, the situation in Russia and Brazil is still serious, the explosion of new cases is still there and the government should take more measures to control the infection growth, cure more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

5 Ana

```
## [1] 101
```

```
##  [1]   1  11  21  31  41  51  61  71  81  91 101
```

Occurrences Map − COVID19



6 Conclusion

Please don't PANIC, stay safe, follow the WHO, and your nation guidelines. We all can defeat this together. Please don't spread rumors.

7 Github Link