Summary of the Recent Novel Coronavirus (COVID-19) Pandemic from 2019 and Comparison with the Severe Acute Respiratory Syndrome(SARS) Epidemic in 2003

Authors: Wenxin Cao (1663021) Yimi Tao (1834695) Yiling Sun (1663149)

Table of Content

Research Questions and Results 2

Motivation and Background 4

Dataset 5

Methodology 6

Results 8

Result Reproduction 18

Work Plan Evaluation 19

Testing 21

Collaboration 21

Research Questions and Results

1. *What are the infection regions of the COVID-19 and SARS outbreak globally? Are there any similarities or differences? What do the similarities or differences imply?*

   The most infected regions of COVID-19 include China(Asia), Italy(Europe), and Iran(Asia). The most infected regions of SARS include China(Asia), Canada(North America), and Singapore(Asia).

   The similarities between SARS and COVID-19 include the virus homology and transmission routes. However, the situation of COVID-19 is more severe compared to SARS. The current infection region of COVID_19 is 5 times larger than SARS, and the number of confirmed cases for COVID-19 already exceeds 10 times the total number of confirmed cases for SARS, while the number is still increasing rapidly.

2. *How did SARS progress over time? How about the novel coronavirus so far?*

   The number of infected cases of SARS increased rapidly during March and May in 2003. Starting from July 2003, the total number of infected cases of SARS remained nearly constant. This epidemic was under control after four months worldwide(March -- July 2003).

   The first case of COVID-19 was reported on 01/22/2020 as recorded in the data we found. The confirmed cases started to increase from the end of January to mid-February in 2020. Surprisingly, the number started to grow rapidly from 02/13/2020, which majorly corresponded to the global spread in South Korea, Italy, and Iran. Currently the situation is still not optimistic all over the world.

3. *What is the situation of infection, death and recovery cases of the COVID-19 pandemic in China? How do the numbers vary in different provinces?*

   Hubei Province, the center of the COVID-19 outbreak, has a far greater number of infectious and death than other provinces. Based on the data, Hubei has around 70 thousand people get infected, while others number under thousands. Overall, the death rate is around 4% in the whole country. This percentage mainly comes from the data of Wuhan. For most other provinces, the death rate is under 2%. The recovery rate is relatively low in Hongkong, while there are several other northern provinces around Hubei that have low recovery rates as well.

4. *What are the differences between the infection numbers for males and females that have been reported as confirmed cases for COVID-19? What do they imply?*

Among the top 10 infectious countries, 8 countries show that the male confirmed cases of COVID-19 are much more than female confirmed cases of COVID-19.

5. *What are the differences between the numbers of infection, death, and recovery cases for different ages that have been reported for COVID-19?*

   Most infectious people are middle-age. The older people are the most vulnerable group of people. Children and young adults have lower infectious rates. There are no people under 30 being recorded dead yet.

Motivation and Background

COVID-19 is an emerging coronavirus first detected at Wuhan City, Hubei Province, China and is observed to be spreading person-to-person. Chinese health officials have reported tens of thousands of cases in China and it has now reached multiple other countries. Considering the severity of this novel virus, in this project, we will investigate the development and spread of the coronavirus in different regions in China and around the world. At the same time, we will also explore another global epidemic SARS, which happened in 2003. There are many similarities between SARS and COVID-19 from the virus homology to the origin and transmission routes. Therefore, we will compare the spread of these two outbreaks and hopefully gain more information on the current COVID-19 pandemic.

We understand the current situation of COVID-19 pandemic by knowing its infection region, number of confirmed cases, recovery and death rate. Then we can have a deeper understanding of its severity by comparing it with the historical data of SARS. By comparing the infection region of SARS and COVID-19 so far, we can observe the difference between the transmission range and extent. By comparing the progress of two diseases over time, we can see the development of the similarities and differences between the two diseases and hopefully we can make predictions on the transmission of COVID-19.

Dataset

1. https://www.kaggle.com/nattay/who-sars-cumulative-reported-cases
    a. This URL displays the development of SARS each day during March and July 2003.
2. https://www.kaggle.com/zhongtr0n/sars-who-data
    a. This URL displays the final report of SARS starting from March 13 to July 11 2003.
3. https://github.com/CSSEGISandData/COVID-19
    a. This URL displays the newly updated summary of confirmed, death, and recovered COVID-19 cases reported globally from JHU CSSE
4. http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51
    a. This URL displays the gender and age distribution of COVID-10 by Feb 11
5. http://www.naturalearthdata.com/downloads/10m-cultural-vectors/
    a. These URLs include shapefiles for cities, provinces, and states in the world.

Methodology

Research question 1:

In order to compare similarities or differences between the worldwide infection scale of the SARS and COVID-19 outbreak, we plan to:

1) plot the world map based on countries using light grey color

2) plot all the countries that have reported any confirmed cases of SARS colored by the number of confirmed cases.

3) repeat the above steps to plot another map for the confirmed cases of COVID-19.

Research question 2:

In order to show the progress of the infection of SARS and COVID-19 globally, we plan to:

1) plot a line plot with the x-axis representing the date and the y-axis representing the number of confirmed SARS cases, colored by different countries.

2) repeat the above step to plot another line plot for the COVID-19 confirmed cases. Since the number of confirmed COVID-19 cases is still increasing, we will use the data up to 03/13/2020.

3) plot the progress line of COVID-19 for each country to show how the virus spread in different regions

For the initial day of the x-axis for both viruses, we will use the official announced date of the first confirmed case for both viruses.

Research question 3:

In order to have an overview of the COVID-19 situation in China, including the death and recovery ratio for each province, we plan to:

1) plot the map of China based on provinces using light grey color

2) plot all the provinces that have reported any confirmed cases of COVID-19 colored by the number of confirmed cases.

3) plot a bar chart showing the top 10 provinces with the highest death ratio for COVID-19. The data should be sorted in descending order.

4) repeat step 3 to plot another bar chart for the recovery ratio.

Research question 4:

In order to see the difference of infection between gender for COVID-19, we plan to:

1) plot a bar chart showing the number of confirmed infected cases of COVID-19 for each gender in the top 10 countries. Male and female will be distinguished by color.

Since approximately 90% of the data for COVID-19 do not contain information about gender, thus we will filter out the portion denoted with gender for this question to receive an estimated trend.

Research question 5:

In order to see the difference of infection between age for COVID-19, we plan to:

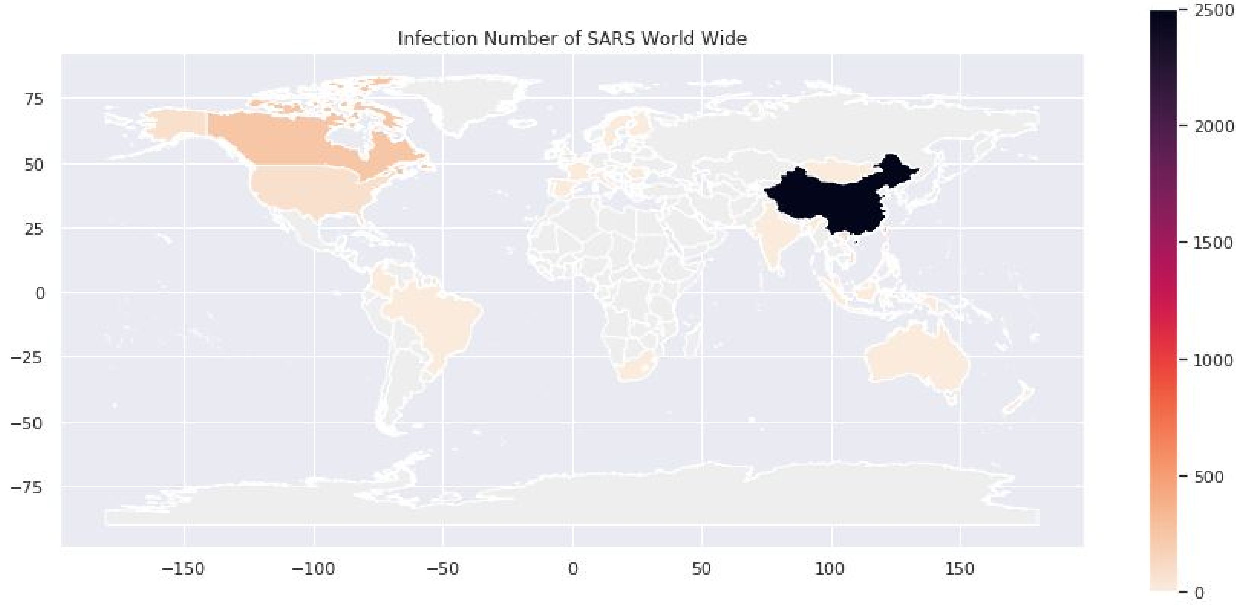
1) Plot a bar chart showing the distribution of age for the people reported infected.

2) Plot a bar plot of age to take a closer look at the statistics.

3) Compare the bar plots of the age of death and still alive infectious people.

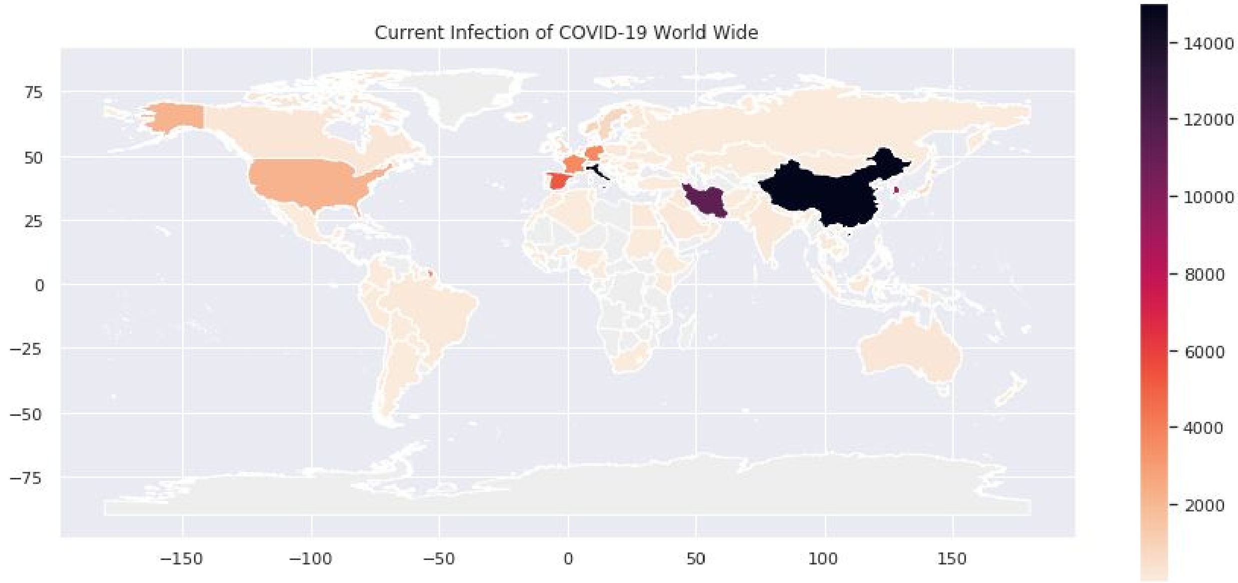
Results

Research Q1: *What are the infection regions of the COVID-19 and SARS outbreak globally? Are there any similarities or differences? What do the similarities or differences imply?*



The above graph shows the infection condition of SARS globally in 2003, with each country colored by the number of the confirmed infection cases. The maximum of the legend is set as 2500 in order to maximize the color difference to show the infection condition more clearly. Any number of confirmed cases above 2500 is shown in black, such as China, which had 5327 cases during the SARS epidemic.

According to the available data and from the above graph, 10% of the countries in the world had reported positive cases of SARS. It can be observed that the most infected region of the SARS epidemic was China, followed by Canada and Singapore. Besides these three countries, the numbers of confirmed cases for each country were all below 200, which were not considered as very serious conditions.



The above graph shows the current infection condition of COVID-19 globally before 03/13/2020, with each country colored by the number of the confirmed infection cases. The maximum of the legend is set as 15000 in order to maximize the color difference to show the infection condition more clearly. Any number of confirmed cases above 15000 is shown in black, such as China, which currently has 80945 cases, and Italy, which currently has 17660 cases during the COVID-19 pandemic so far.

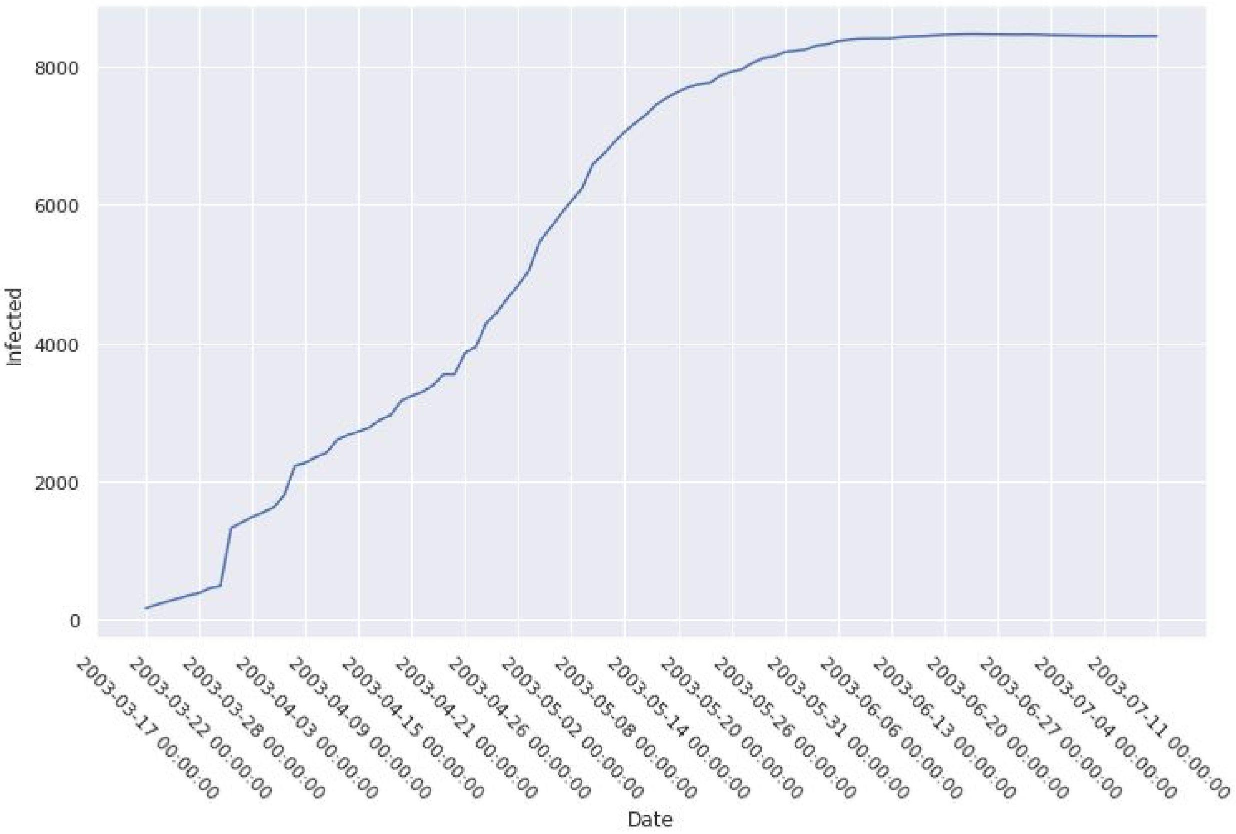
According to the available data and from the above graph, 50% of the countries in the world have reported positive cases of COVID-19. It can be observed that the most infected regions of the COVID-19 pandemic are China, Italy, and Iran so far. Besides these three countries, there are more than 15 countries with the numbers of confirmed cases exceeding 200, which shows the severeness of this pandemic.

Even though there are many similarities between SARS and COVID-19 from the virus homology to the origin and transmission routes, from the data forehead mentioned, the current situation of COVID-19 is more severe. The current infection region is 5 times larger than SARS, and the number of confirmed cases for COVID-19 already exceeds 10 times the total number of confirmed cases for SARS, while the number is still increasing rapidly. These observations indicate that it is urgent to take immediate action to prevent the person-to-person spread and come up with effective medical measures to eradicate the virus.

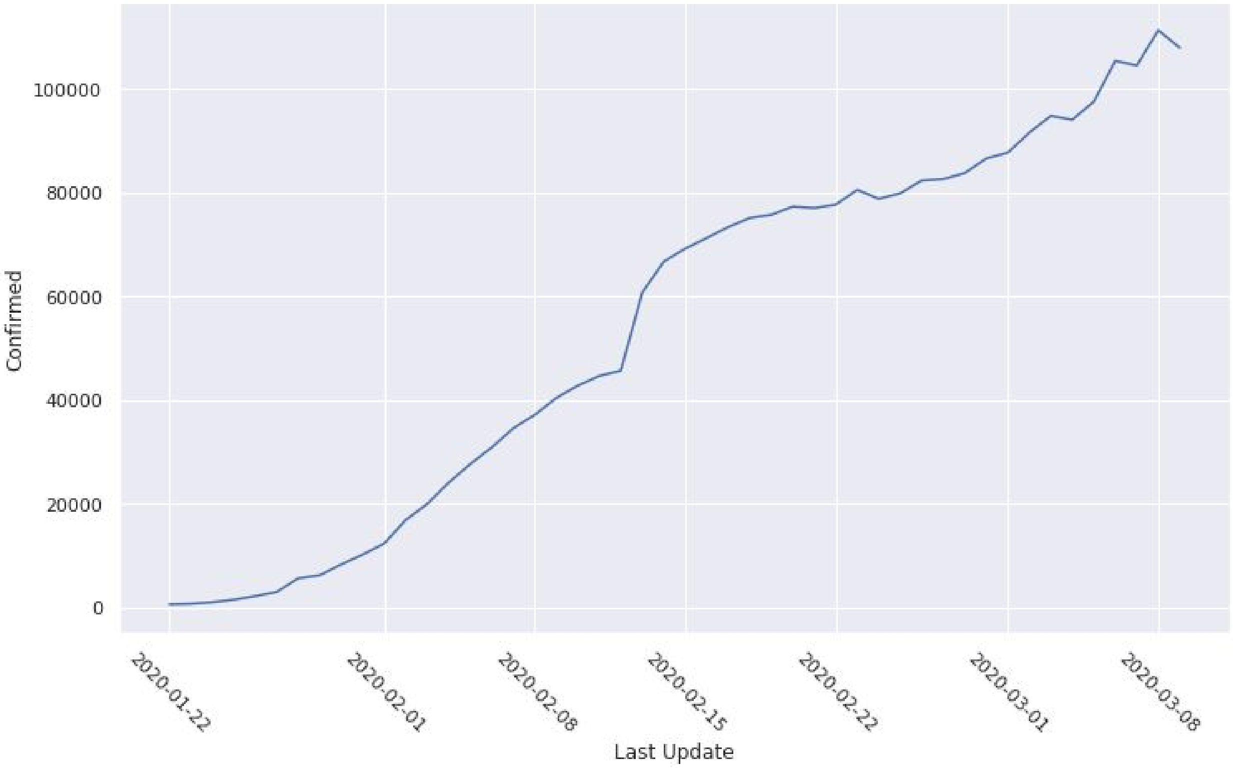
A more intuitive chart is shown below, comparing the top 10 countries with the most infection cases for COVID-19.

The chart shows that, even though China has the greatest number of confirmed cases, the ratio between the number of the infectious and total population is not the largest due to its large population. Meanwhile, for those European countries that have lower populations, especially Italy, they are also facing severe situations. By taking the population into consideration, it turns out that Italy has almost 300 infected cases among a million people. The countries contained in the chart need to take effective measures to protect their people's health.

Research Q2: *How did SARS progress over time? How about the novel coronavirus so far?*



The above graph shows the overall progress of the SARS epidemic. It can be observed that from 03/17/2003 to 05/23/2003, the number of infected cases of SARS increased rapidly. After 05/23/2003, the situation got controlled and the increase in confirmed cases was minimal. There has been a zero increase since 07/18/2003 indicating that this epidemic was overcome. It took around four-month for SARS to be eradicated.

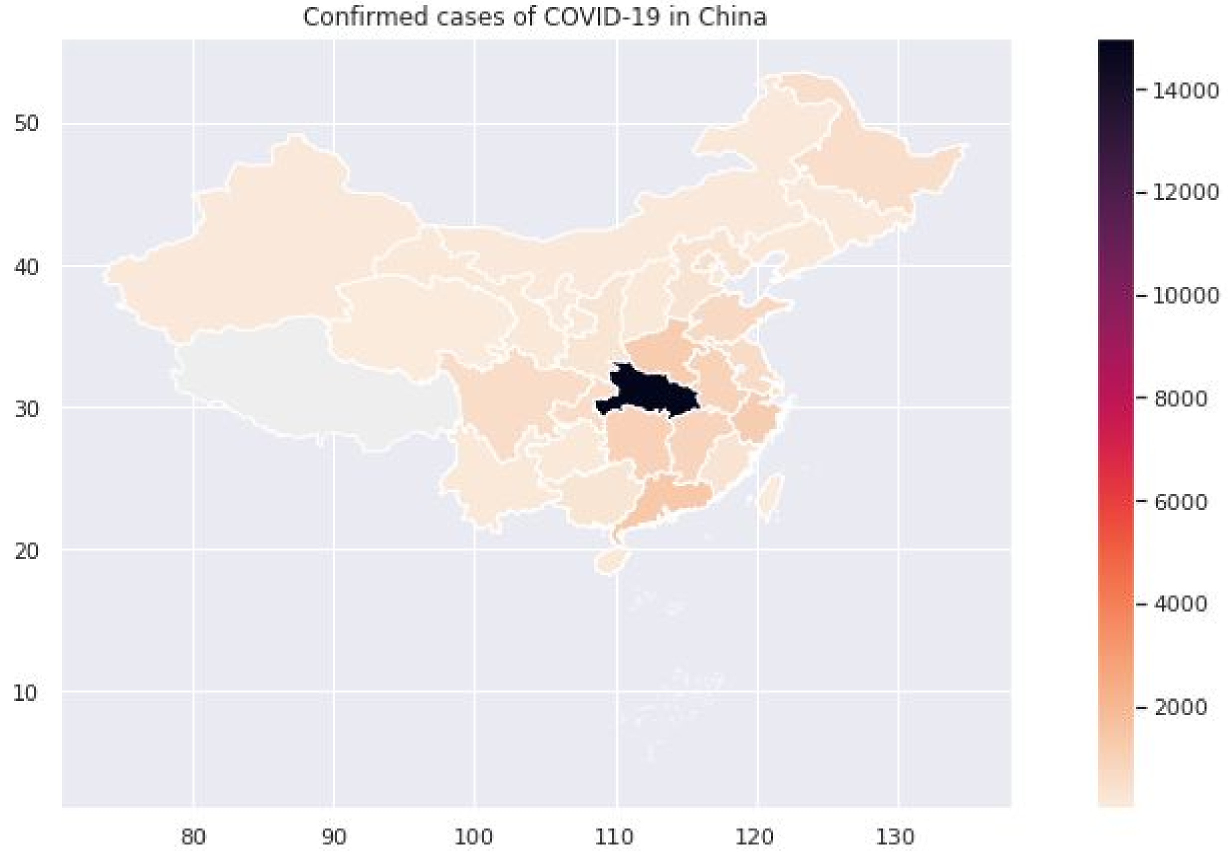


The above graph shows the current progress of COVID-19 pandemic. It can be observed that from 01/22/2020 to 02/12/2020, the number of confirmed cases of COVID-19 increased rapidly, which corresponds to the first outbreak in China. The steep change around 02/12/2020 indicates a significant increment in the number of confirmed cases which corresponds to the outbreak in Europe, especially Italy. The total number is still increasing so far and it has been two months since the first case. Compared with the progress of SARS, we can learn that it will take at least four months to eradicate COVID-19 due to the more severe condition this time. Awareness and caution are urged for everyone’s health.

From the top 10 infectious countries in the world, China is the county with the most confirmed cases of COVID-19. From 01/22/2020 to 02/15/2020, the number of confirmed cases of COVID-19 increased rapidly. However, the severe situation of COVID-19 was under control by China within one month. Starting from 02/22/2020, the rest countries displayed steep increments in the number of newly confirmed cases of COVID-19. Among those countries, Italy has the worst situation. So far, according to the data, there is no sign that the COVID-19 pandemic is controlled in those countries.

Combined with the recent news, the increase in the number of newly confirmed cases of COVID-19 may be due to the lack of attention or action, and the absence of perception of the severity of COVID-19. As known, Wuhan province in China has been in lockdown since 01/23/20, when there were 835 confirmed cases in China. And the daily increment number of confirmed cases right now(03/13/20) is minimal(around 10). Recently the Lombardy region in Italy has been in lockdown to prevent further spread while over 6000 cases have been reported already. It can be predicted that it is very likely for Italy to take more than two months to get the situation under control.

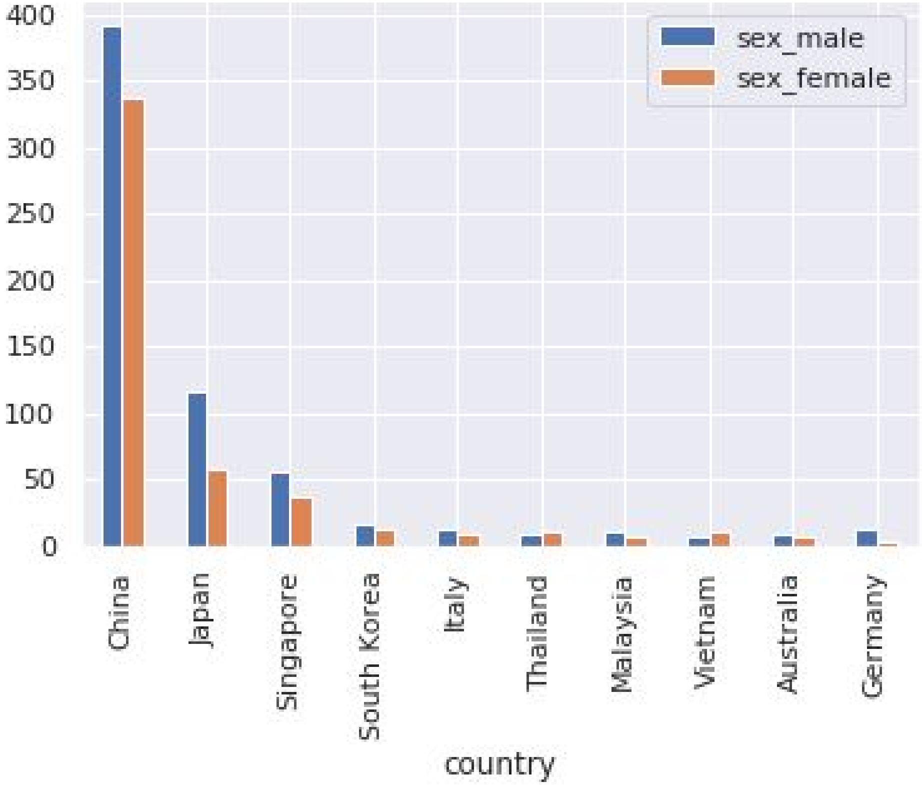
Research Q3: *What is the situation of infection, death and recovery cases of the COVID-19 pandemic in China? How do the numbers vary in different provinces?*



There is a distinct difference between the number of confirmed cases of COVID-19 in province Hubei and the rest provinces in China. Hubei is the most infected region with confirmed cases of COVID-19 exceeding 14,000. All other provinces have confirmed cases under 1500 by 03/11/2019. This was due to a government regulation that placed a lockdown on the entire Hubei province. It can be observed that this regulation effectively prevented the spread of COVID-19 to other provinces.

The above graphs provide detailed information about the death and recovery ratio in each province in China. The overall death rate is around 4% in China, while only Hubei province has a death rate over 4%. Besides, 118 people died outside Hubei in China. Only 6 provinces have death rates over 2%, while 27 provinces have death rates under 2%. As for the recovery rate, around 60% infectious people haven’t recovered from COVID-19 yet in Hong Kong. Gansu, Hubei, and Beijing have relatively low recovery rates as well.

Research Q4: *What are the differences between the infection numbers for males and females that have been reported as confirmed cases for COVID-19? What do they imply?*



As the chart shows, 8 of 10 countries have higher numbers of infections being recorded for males than females. This difference is most outstanding in Japan in that the number of female infections is half that of males. There are two possible explanations for the differences in number. One possible explanation is that male is at higher risk of being infected than females. It may be caused by the higher exposure to the virus during work or social events. Another possibility we concern is that it may imply that males tend to have stronger symptoms that drive them to go to the hospital rather than self-quarantine, which leads to more recorded cases of males.

Research Q5: *What are the differences between the numbers of infection, death, and recovery cases for different ages that have been reported for COVID-19?*

The distribution of recorded infectious people's age shows that the infectious people have ages ranging from 0.25 to 96. The smallest children get infected is only 0.25 years old, around 4 months. As the box plot, the average age of infectious people is 50 years old. The first quartile and third quartile are around 35 and around 60. As data was split into death or not, the box plot for still alive people is similar as before. For death, the average age now is around 70, with the first quartile of 60 and third quartile 80. From the data, we may conclude that middle-aged(40-60) are most prone to viruses, while older people (60-80) are the most vulnerable group. For young adults and teenagers, the number of records is relatively low and not fatal. It may be explained by the mild symptoms of COVID-19 in young people. Most people with mild symptoms would choose to home quarantine.

Result Reproduction

We have a shared google drive with all of data and code:
https://drive.google.com/open?id=1GmcPEkBEk0Y_-fE8FR3dJCwCMUWAjIFc

To reproduce the result, run the main.py file.

Work Plan Evaluation

Work Plan:

1. Perform Analysis (03/04 - 03/08)

   Use Colab to write and store our code.

- Read in the shapefile of each country and CSV file for SARS and COVID-19.
- Write a function called merge_data that takes two parameters, a shapefile indicating each country and a CSV file that contains information about infection countries. The method should join the data on the column of [‘Country’] and return a GeoDataFrame. The function should be called twice in main in order to merge the data for both SARS and COVID-19. (03/04/2020)
- Write a function called graph_infection_scale that takes a merged GeoDataFrame as the parameter. The function will first plot the world map in light grey as a background reference, and then plot all the countries that had reported confirmed cases based on the passed in data, colored by the number of the confirmed cases. The function should be called twice in main in order to have plots for both SARS and COVID-19.
- Write a function called plot_progress_line that takes a merged GeoDataFrame as the parameter. The function will plot a line plot showing the growth of the number of confirmed cases over time, colored by different countries. The function should be called twice in main in order to have line plots for both SARS and COVID-19.
- Write a function called plot_idr_china that takes a merged GeoDataFrame as the parameter. The ‘idr’ in the method name refers to infection, death, and recovery. The function will plot a bar chart based on the counts for infection, death and recovery cases in different provinces, colored by infection, death and recovery cases. The function should be called twice in main in order to have line plots for both SARS and COVID-19. (03/06/2020)
- Write a function called plot_gender that takes a merged GeoDataFrame as the parameter. The function will plot a bar chart based on the count for male patients and female patients in different provinces, colored by different genders. The function should be called twice in main in order to have bar charts for both SARS and COVID-19.
- Write a function called plot_age that takes a merged GeoDataFrame as the parameter. The function will plot a bar chart based on the count for infection, death, and recovery cases of each age group, colored by different age groups. The function should be called twice in main in order to have bar charts for both SARS and COVID-19. (03/08/2020)

2. Testing (03/09/2020)

   Final check if code works alright.

3. Project Summary and results (03/10-3/13)

Evaluation:

In the whole process, we used Collab to do the visualization and test collaboratively. We spent two more days getting familiar with data, then cleaning and filtering the data to match the name in geospatial data with the infection data. This is an unexpected but necessary step before we did the analysis. To do this, we add additional functions for cleaning.

Furthermore, we planned to implement a single method for merging data. However, the merging details vary for different data files. For instance, the column names for countries are different in SARS and COVID files. Because of the differences, for each merging, we implement a unique merge function.

Also, in our original plan, we also wanted to plot the gender distribution and age distribution for SARS epidemic in order to show the trends of infection. However, since SARS broke out in 2003, the recorded data wasn't complete with patients' gender and age, we were unable to plot the desired trend for SARS. Therefore, we focused on exploring more of the data for COVID-19 instead.

It is worth mentioning that at the time we decided the topic of our report, the scale of COVID-19 wasn't as large as the current. The most infected region was majorly China and the situation of other countries was not very severe, which was very similar to the SARS epidemic in 2003. So we decided to compare these two outbreaks since they shared many similarities. However, as COVID-19 continued to spread globally, it could be noticed that COVID-19 led to much terrible conditions, and more countries got influenced. As WHO declared COVID-19 a pandemic, the necessity of the comparison with SARS was reduced. As a result, we focused more on COVID-19 in our report at the time of submission.

Testing

Because we used collab to do the analysis, we can see the result immediately after we write code. Since the data for SARS and COVID-19 are mainly CSV files, in order to find out whether our graphs/plots match the data in CSV files, we apply different functions in Excel. For instance, for finding the number of females and males of confirmed cases of COVID-19, we use “=COUNTIF(C2:C13175, "male")” to find cases for males and “=COUNTIF(C2:C13175, "female")” for females for CSV file “COVID19_open_line_list.csv”. And then we compare the result of the CSV file to the result of our analysis and decide whether our function is correct. In addition, most CSV files contain a row that shows information about the total number(total confirmed, total deaths, and total recovered), and we use that row to compare to our result.

Since our analysis for SARS and COVID-19 focused more on displaying the data visually, we did not use asserts to test our result.

Collaboration

All members contributed equally to this project. No other assistance was received.