COVID-19 Spread and Recovery Analysis

Shravani Kasralikar

Abstract

For this project, I wanted to analyze the correlation of COVID-19 mortality rates between the US and countries that have emerged from lockdown as of 05/2020: China, South Korea, and Germany. I also analyzed the general overall COVID-19 spread and recovery globally; I observed confirmed cases, deaths and recoveries between US states and US provinces and between foreign countries. I used six different data sets (.csv) taken from an open source JHU data set published by user SRK on kaggle.com: COVID-19 global confirmed cases, global deaths, global recoveries, US confirmed cases, US deaths, and South Korea confirmed cases, deaths and recoveries.

My outcomes show that the US has the highest mortality rate preceding Germany, China, then South Korea (in that order). China was the first country to successfully flattened their growth curve by making their new daily number of confirmed, recovered, and death cases after April zero; they preceded South Korea, Germany and then the US (in that order). Locally, New York has the highest number of COVID-19 cases, and globally the US has the highest number of COVID-19 cases worldwide.

Motivation

My overarching question and problem I want to solve is: how have other countries reduced their mortality rates of COVID-19 faster than the US? To solve this, I wanted to explore the correlation between mortality rates in the US, China, Germany, and South Korea. I also wanted to observe the individual country datasets to determine when and how their COVID-19 confirmed cases growth curve was flattened.

My motivation for this project was use my findings from my COVID-19 dataset comparisons to evaluate the proficiency of global healthcare systems and global political actions taken between the US and countries who have unimplemented quarantine regulations by May 2020 (China, South Korea and Germany). What have other countries done to limit COVID-19 spread and how have their healthcare systems effectively contributed to reduced mortality rates?

This study could be useful for policy makers, healthcare professionals, and researchers.

Dataset(s)

COVID19 Data Set:

- ➔ https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
- ➔ https://github.com/CSSEGISandData/COVID-19

South Korea Data Set:

- ➔ https://www.kaggle.com/kimjihoo/coronavirusdataset

- ➔ time_series_covid_19_confirmed.csv (266 rows x 112 cols)
  - ◆ global confirmed COVID-19 cases by date (01/22/20 → 05/08/20)
  - ◆ COL: Province/State, Country/Region, Lat, Long, Data
- ➔ time_series_covid_19_recovered.csv (252 rows x 112 cols)
  - ◆ global recovered COVID-19 cases by date (01/22/20 → 05/08/20)
  - ◆ COL: Province/State, Country/Region, Lat, Long, Data

Dataset(s)

- time_series_covid_19_deaths.csv (266 rows x 112 cols)
  - global deaths COVID-19 cases by date (01/22/20 → 05/08/20)
  - COL: Province/State, Country/Region, Lat, Long, Data
- time_series_covid_19_confirmed_US.csv (3261 rows x 119 cols)
  - US confirmed COVID-19 cases by date (01/22/20 → 05/08/20)
  - COL: UID, is02, iso3, FIPS, Country, Province_State, Country_Region, Lat, Long, Combined_Key, Population
- time_series_covid_19_deaths_US.csv (3261 rows x 120 cols)
  - US confirmed COVID-19 cases by date (01/22/20 → 05/08/20)
  - COL: UID, is02, iso3, FIPS, Country, Province_State, Country_Region, Lat, Long, Combined_Key, Population
- time.csv (102 rows x 7 cols)
  - South Korea confirmed, deceased, and released COVID-19 cases by date (01/20/20 → 04/30/20)
  - COL: date, time, test, negative, confirmed, released, deceased

Data Preparation and Cleaning

To clean data, I did the following:

- Filtered all datasets and removed multiple columns so that only necessary data (name, population numbers, [confirmed, deceased, recovered] and timeline dates) were left
- To gain access to China and Germany data, I sliced the individual row whose string contained the country name and used the .groupby() and .sum() function to sum all the population data
- Transposed the China and US data frames to match the Germany series that was created during slicing so there were no dimension mishaps when plotting through matplotlib
- Split the South Korea data.csv dataset into multiple data frames that could be plotted against each other

Research Question(s)

1) What is the correlation between COVID-19 confirmed cases, deaths,and recoveries worldwide over time? In the US?

2) Does the amount of COVID-19 confirmed cases worldwide and in the US have an effect on COVID-19 deaths? COVID-19 recoveries?

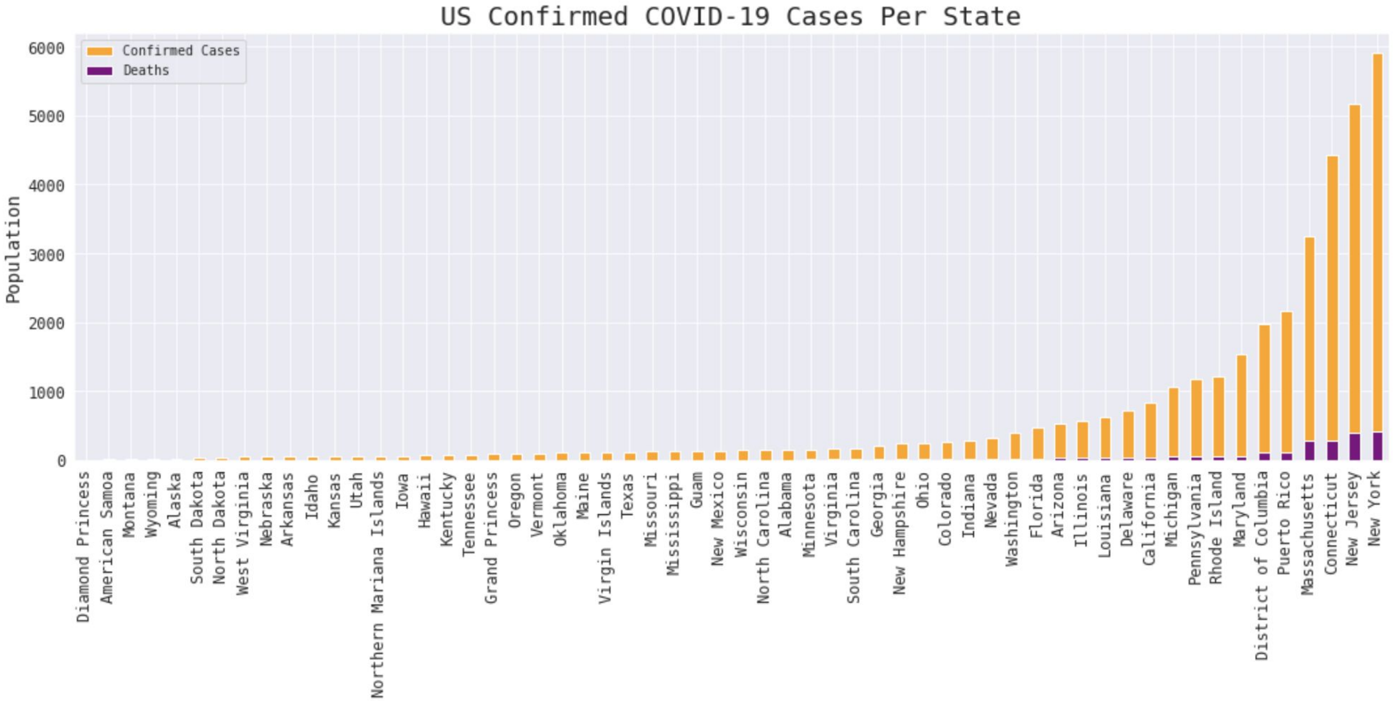
3) Do countries which have stopped quarantine measures have higher or lower mortality rates than the US over time?

Methods

I used various different quantitative data analysis methods to research my data:

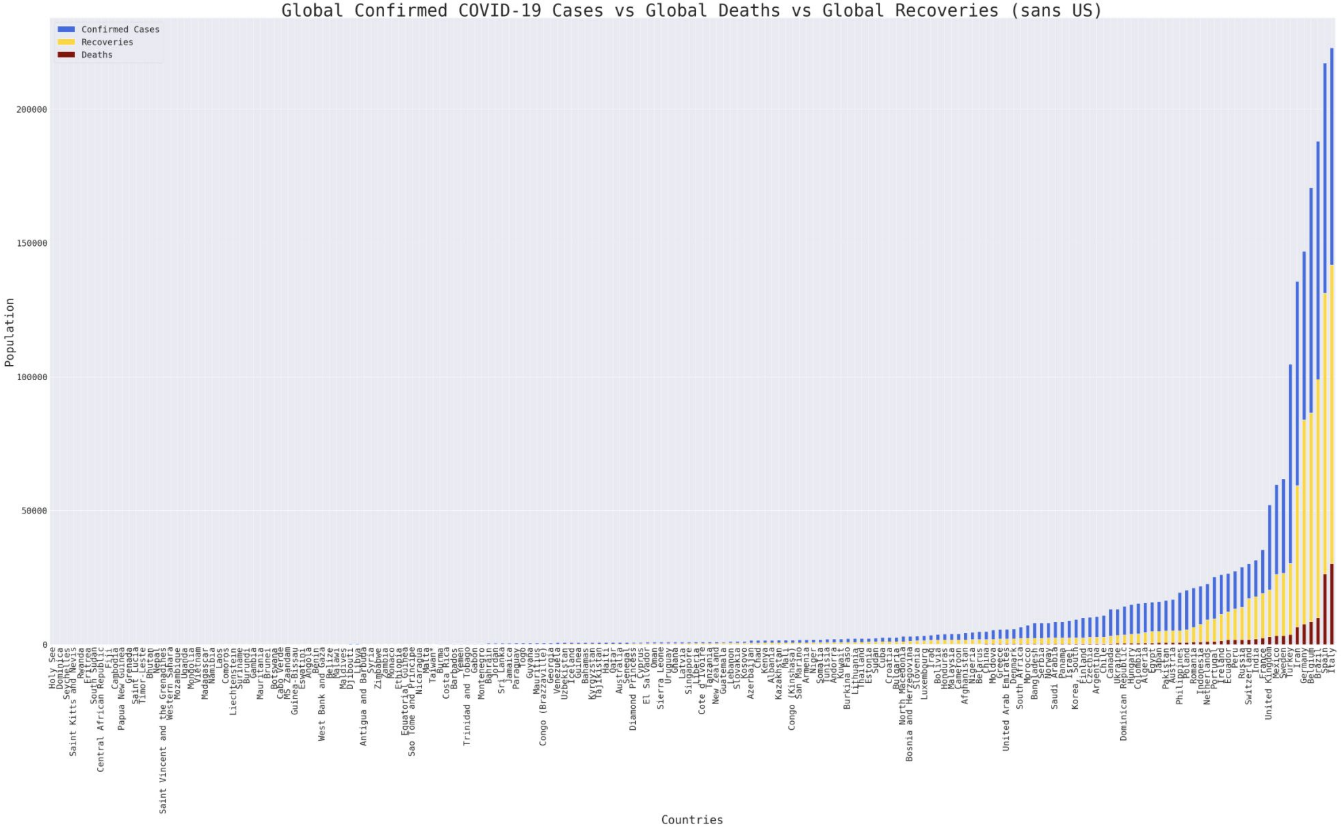
- I built a narrative around my research questions and connected it to my motivation
- I omitted useless data and cleaned up my data using pandas and numpy tools
- I conducted statistical analysis and calculated correlation coefficients between each one of my datasets
- I created multiple bar and line plots which highlighted relationships between each one of my datasets

Findings



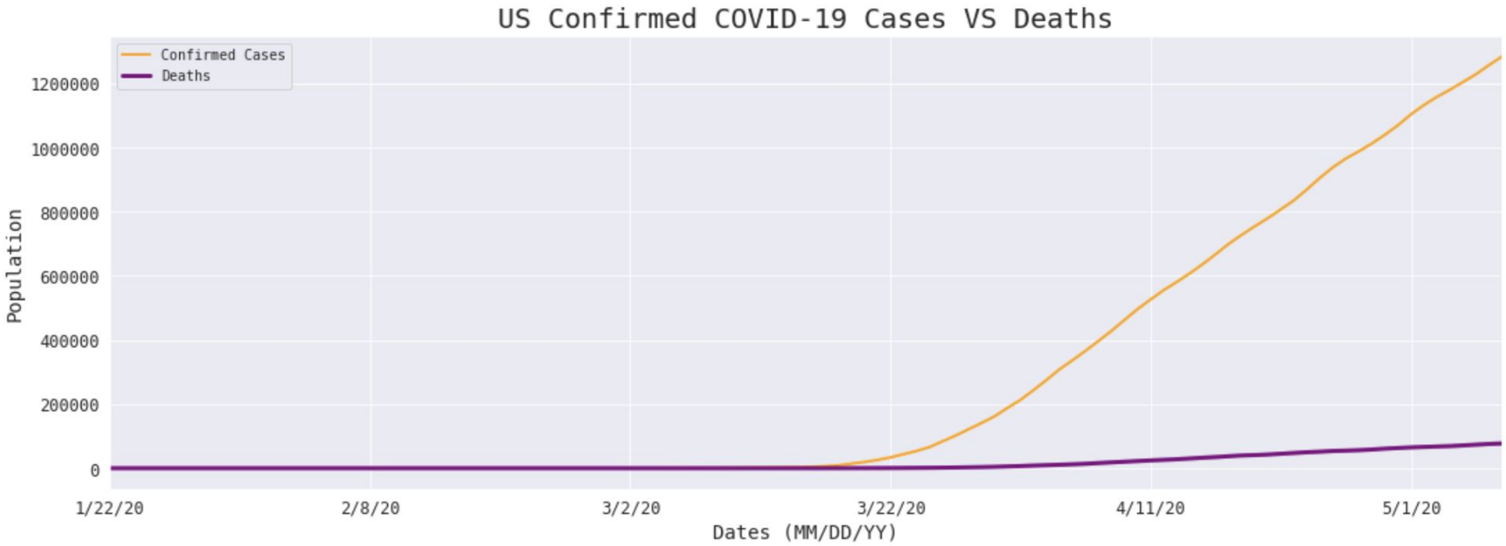
This graph helped me visualize the differences between COVID-19 cases in the US and COVID-19 deaths; this reflects on the mortality rate between each state. The graph shows that New York, New Jersey and Connecticut have the highest number of confirmed cases and deaths, however, Massachusetts has the highest death to confirmed COVID-19 cases ratio.

Findings



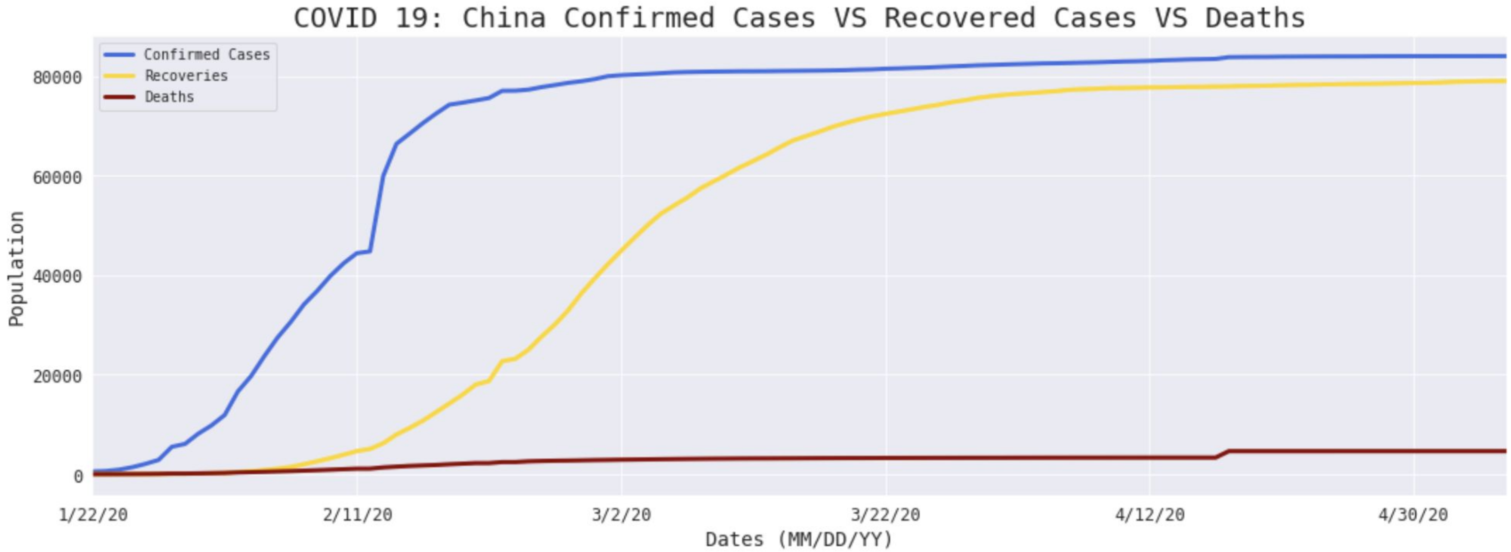
This graph helped me visualize the differences between COVID-19 cases, deaths and recoveries between each country globally. This graph does not include the US because compared to other countries, they exponentially have higher number of cases, deaths and recoveries. The graph shows that Italy, Spain and Brazil have the highest number of cases and deaths, however Spain and Brazil have almost the same amount of recoveries despite large differences in death and case count.

Findings



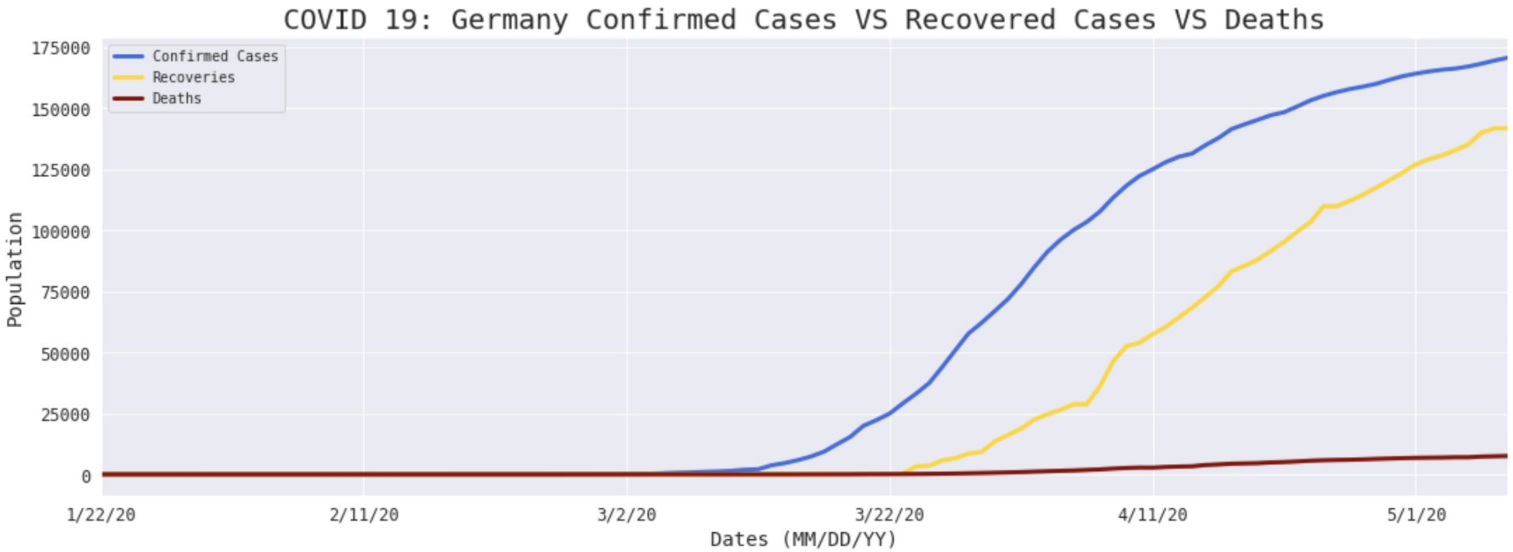
This graph shows the difference between US confirmed COVID-19 cases and deaths overtime. Around March 22, the number of confirmed cases skyrockets while the number of deaths increases at a slower rate. Although the recovery data for US cases was not provided, the data seems to reflect that the mortality rate in the US is low due to the low number of deaths over time.

Findings



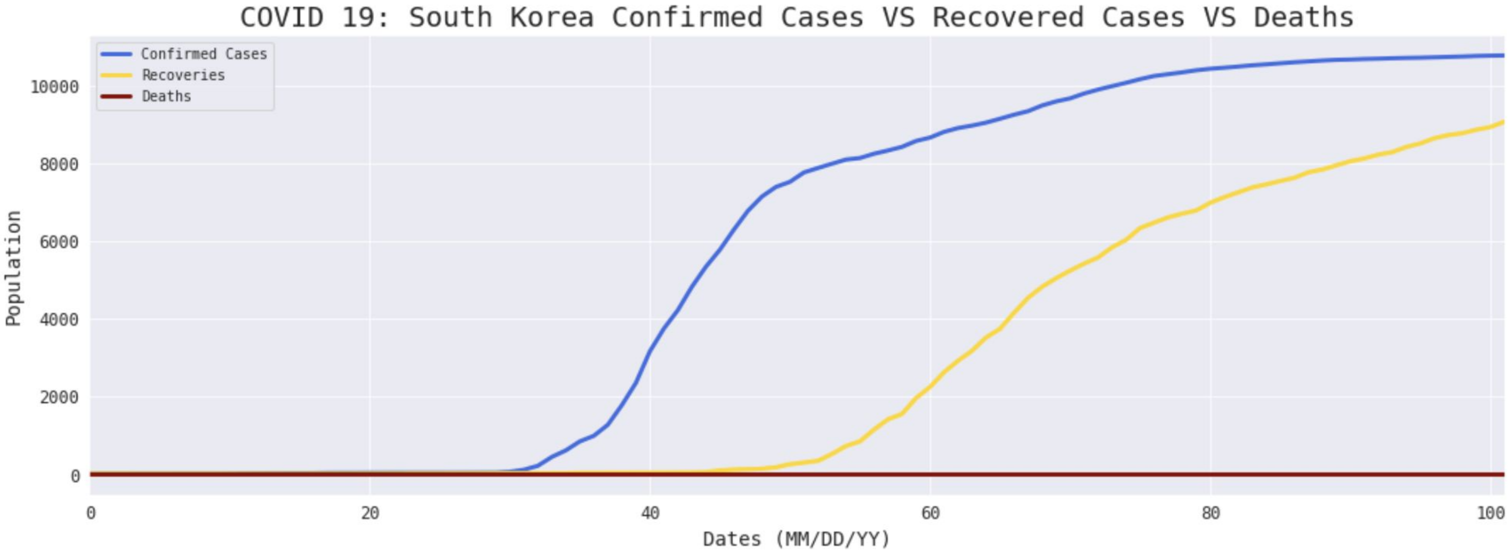
This graph shows the difference between China confirmed COVID-19 cases, recoveries and deaths overtime. The graph clearly shows the number of confirmed cases skyrocketing in January (which is when data recording started) and plateauing right around March 2 after an irregularity around February 12.The recovery rate increases around the beginning of February but flattens out around April 12. The death count rate stays constant except for a small spike around April 15th. Due to China's strict regulations, the confirmed case, recovery, and death count curves all flatten out after strong spikes in the beginning.

Findings



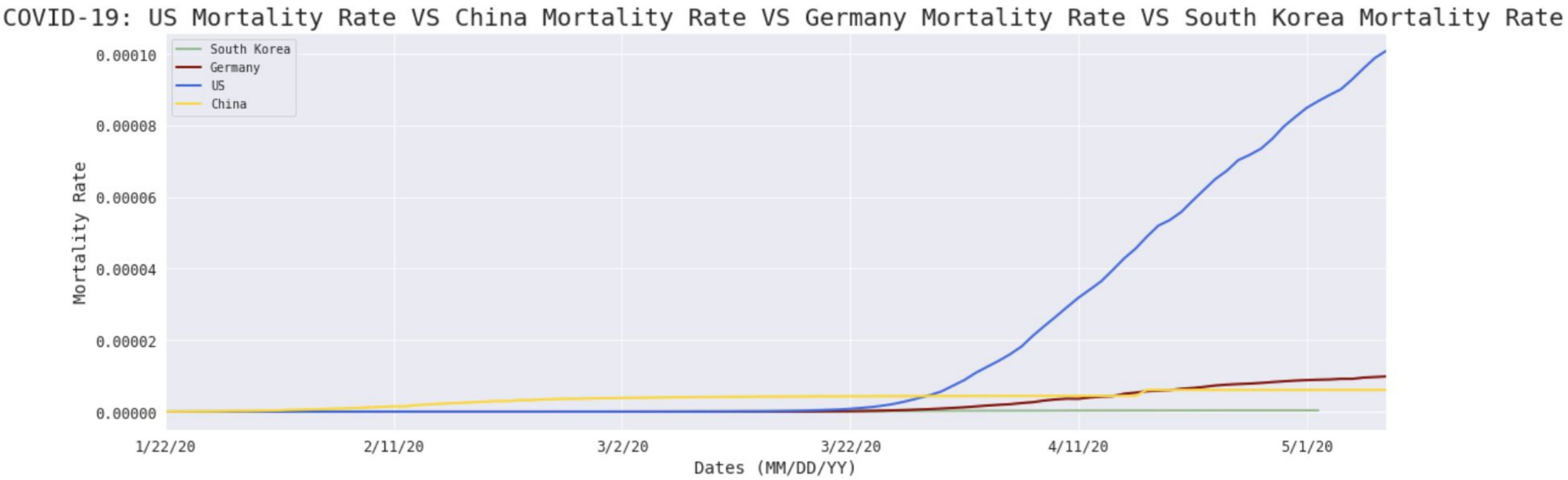
This graph shows the difference between Germany confirmed cases, recoveries and deaths over time. The number of confirmed cases spikes around mid March while the number of recoveries spikes around March 22. The number of deaths stays constant but slowly increases over time and is on an upward trend, showing that the mortality rate in Germany seems high.There are some interesting irregularities in the recovery rate, reflecting on some further research that could be done.

Findings



This graph shows the difference between South Korea confirmed COVID-19 cases, recoveries and deaths overtime. There is no new data available for South Korea after April 30, which is around when confirmed cases, recoveries, and deaths started plateauing. There is a sharp increase in COVID-19 confirmed cases around the 30 day period (February) and for recoveries around the the 50 day period (mid March). The number of deaths stays fairly low and constant, showing that the mortality rate is probably low.

Findings



This graph shows the difference between US, China, Germany, and South Korea mortality rates over time. This is the main research question I wanted to explore, and all my other data visualization have been a foundation for this visualization. This graph clearly shows the US mortality rate skyrocketing around March-end, with Germany in second place, following China then South Korea. South Korea's data-recording stops April 30th, which is why the line graph stops extending at that point. There are interesting regularities seen in China's mortality rate, as it intertwines with Germany, increasing and decreasing at certain time periods in April. This could be due to reopening issues or limitations in data provided by WHO. The US mortality rate spike makes sense as they were the last country to start quarantine, compared to these other three countries which recovered quickly and are relaxing their COVID-19 regulations as a result.

Limitations

There were a few limitations from my findings:

- South Korea COVID-19 data was incomplete and stopped recording after 4/30 despite there being new confirmed cases as of 5/23 (confirmed by: South Korea Coronavirus: 11,165 Cases and 266 Deaths)
- China COVID-19 data may have some limitations due to political implications (confirmed by: China's Coronavirus Figures Don't Add Up. 'This Never Happens With Real Data.')
- US COVID-19 data was incomplete compared to countries globally and did not have recovery information; only number of deaths and confirmed cases were provided which lead to incomplete data analysis of recovery to death rate correlation

Conclusions

According to my correlation coefficients I calculated, there is a positive correlation between confirmed cases, deaths, and recoveries worldwide over time; in fact, it is close to 1 at .831. In the US, the correlation coefficient is even higher at .981.

From these correlations coefficients and from my data visualizations, it is clear that the way COVID-19 has spread worldwide has had an effect of how COVID-19 spread through the US, but not visa versa. US seems to have the highest mortality rate and it seems independent of the rest of the globe.

As for worldwide recoveries, countries with high confirmed cases did not seem to have an effect on their individual recovery rates, however countries with high recovery rates seemed to conclusively have high death rates as well. An exception this however is Brazil, which managed to keep their recovery rate high and death toll low despite their large number of confirmed cases. I find this interesting and hope further research can be done on the healthcare systems of countries such as Brazil.

Overall, the US so far has the highest mortality rate compared to Germany, China and South Korea who managed to keep their recovery rate high and mortality rate low. South Korea's mortality rate was especially low and their recovery rate correlation coefficient was high at .985. China's death to recovery correlation coefficient was the lowest at .923 and Germany's was the highest at .999. These numbers along with the visualizations of confirmed cases and deaths flattening over time in China, Germany and South Korea all beg further research on healthcare regulations in these countries and could help policy makers visualize what bills could be passed to help the US healthcare system keep their mortality rate and confirmed case count lower.

Acknowledgements

I collected the majority of my data from kaggle at 4: www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset.

The github repository from which this data was obtained from is from John Hopkins University and is linked at: www.github.com/CSSEGISandData/COVID-19

My South Korea data was taken from a separate kaggle source at: www.kaggle.com/kimjihoo/coronavirusdataset

I would like to acknowledge my roomates for reading over my final project and helping me analytically process my data and the research questions I wanted to inquire about for this broad data set.

References

I did all the research on my own and used my knowledge of the COVID-19 pandemic and daily news I follow to guide my analyses.