Group Name: GGWP
Group Member: Yuhui Wang (yw4479@nyu.edu) Xinchi Huang (xh1255@nyu.edu)
Project Topic: Analyzing the Pandemic Spread and the Effect of Interventions – Idea 4
Github Link: https://github.com/ywangdq/bigdataProject

Project Description:
Out project aims to investigate in the country-level and state-level COVID-19 case trends by analyzing the data of different countries. We collect data about confirmed cases, deaths and recovered cases of different countries and states in U.S. as shown below.

For the data pre-processing, because the data are collected from official websites like New York Times and China Data Lab, the data are quite clean and formal. So we mainly focus on consistency checking and data clustering. We notice that there are mismatches of city names between different datasets. It is because some of the country names are in abbreviation form, for example, 'Saint' is written as 'St.'. So we unified all country names to their full names. Besides, the raw data are simply ordered by the report date, in order to make it easy to read and draw graphs, we first group the data by country/state name then order by report date and select the top 20 countries with most confirmed/deaths/recovered cases. In order to get the daily increase data in New York, we first select all data in New York and order them by report date. Then we do subtraction between data of neighboring dates to get the daily increase cases.

To visualize the data, we do some research and decide to use python folium library to draw the graphs. Folium is a library that make it easy to visualize data in python. It provides many functions including drawing choropleth maps and dynamic heatmaps. With the data we get from previous operations, we draw the following 7 graphs:

1. New York daily increase confirmed/death cases graph
2. U.S. daily increase confirmed/death cases graph
3. U.S. confirmed heatmap
4. U.S. deaths heatmap
5. World confirmed top 20 countries choropleth map
6. World deaths top 20 countries choropleth map
7. World recovered top 20 countries choropleth map

The main difficulty we have is that drawing heatmaps requires the coordinates of states in the U.S. and the geo json of different states. So we search about the coordinates of all states and join the data of state coordinates and data of state COVID-19 cases by state name so that we can locate points in the map.

From the above research, we notice that by 18th April, U.S has become the country with most confirmed cases (732197 cases) and deaths cases (38664 cases) in the world. However, it is not the country that has most recovered cases. Germany has 143,342 confirmed cases, which is only 24% of U.S. but it has most recovered cases (85400 cases). We think there are mainly two reasons for this. First, the spread time of the pandemic in Europe is earlier than that in U.S.. The vast spread of COVID-19 in Italy forced countries in Europe started to prepare for it early. Second, although U.S.

shut down transportation between U.S. and countries suffering from the pandemic early, it did not make enough preparation to deal with those who were already infected in the nation. The government did not encourage citizens to wear masks early, leading to the fast domestic spread. Another thing we find from our investigation is that starting from April, the daily increase of confirmed cases in the U.S. were fluctuating around 30,000. We thought there might be two possible explanations. First, the spread of the virus has been gradually under control as the government were carrying out effective policies and people were aware of the seriousness of the pandemic. Another possibility is that, the U.S. did not have enough test kits and 30,000 cases per day was the maximum number of tests available.

At last, all our results and outputs described are reproducible as we have tested our codes and results multiple times. We also add comments to our codes to make sure that they are readable and easy to understand.

Datasets:

1. *World Covid-19 Daily Cases with Basemap*: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L20LOT
2. *US Covid-19 Daily Cases with Basemap*: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HIDLTK
3. *Coronavirus (Covid-19) Data in the United States*: https://github.com/nytimes/covid-19-data
4. *World COVID-19 Events Timeline:* https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OAM2JK

Tools:

Spark, Pandas, Python folium