Covid19 Hackathon

By JETW -- Wanyu, Emma, Jason and Tommy

Introduction

In this project, we explored two COVID-19 datasets and one Influenza/ Pneumonia dataset to examine whether there would be correlations between diagnosis and death cases from COVID-19 and Influenza in the US.

We also looked into the dataset of community mobility during COVID-19 to investigate possible correlations between various community mobility features and confirmed cases of and/or deaths from Covid-19 in the US.

Data

We explored the following data:

- Confirmed cases of and deaths from Covid-19
  - From the *New York Times*' GitHub
    i. Locations include the United States, D.C., and 4 US territories
    ii. Dates range from January 21, 2020 through June 12, 2020
  - From JHU CSSE GitHub
    i. Locations include 58 province states
    ii. Dates range from January 22, 2020 through June 13, 2020

Data

- Death Rates and Number of Deaths from Influenza and Pneumonia in the US
  - From the Centers for Disease Control
  - Includes all 50 states and DC
  - From 2018 (the most recent year with finalized statistics)

- Community Mobility Data
  - Filtered to include only data from the United States
  - Dates range from February 15, 2020 through June 7, 2020
  - Includes statistically significant increases and decreases in the visitors and/or time spent in a location (e.g., transit station or workplace) compared to a normal day

Libraries

For this project, we mainly used numpy and pandas to read data into jupyter notebook, pre-process data and conduct the data analysis.

We used matplotlib to visualize the data. To add interactivity into some graphs, we also used plotly python package

EDA Findings from the U.S. States Data

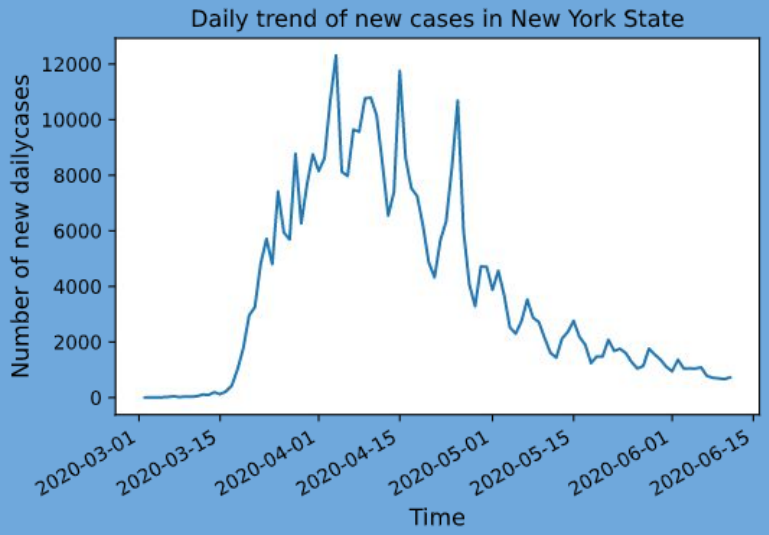
First we explored the state-level data for covid19 in the U.S and here are some initial findings. (The data ends at 2020/06/11.)

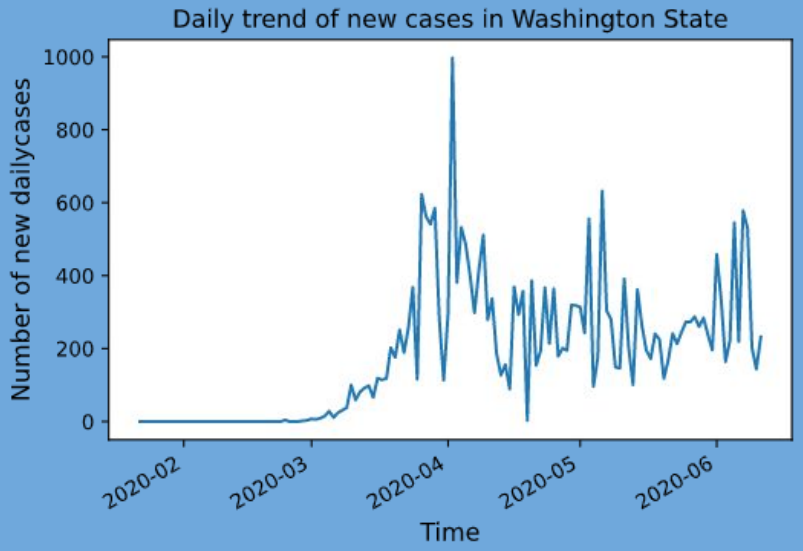
- The top 5 states with the highest number of cases are New York (385669), New Jersey (165816), California (143693), Illinois (131731) and Massachusetts (104667).
- The top 5 states with the highest number of deaths are New York (30431), New Jersey (12443), Massachusetts (7492), Illinois (6388) and Pennsylvania (6187).
- The top 5 states with the highest death rates are Connecticut (9.3%), Michigan (9.1%), Virgin Islands (8.3%), New York (7.9%) and Pennsylvania (7.6%).

We're also interested in exploring the both the overall and daily trend of the covid-19 cases and deaths. So we created two functions in the notebook to plot the overall/daily trend for any input state for cases/deaths.

For example, Here is the daily trend of new cases in New York State and Washington State.

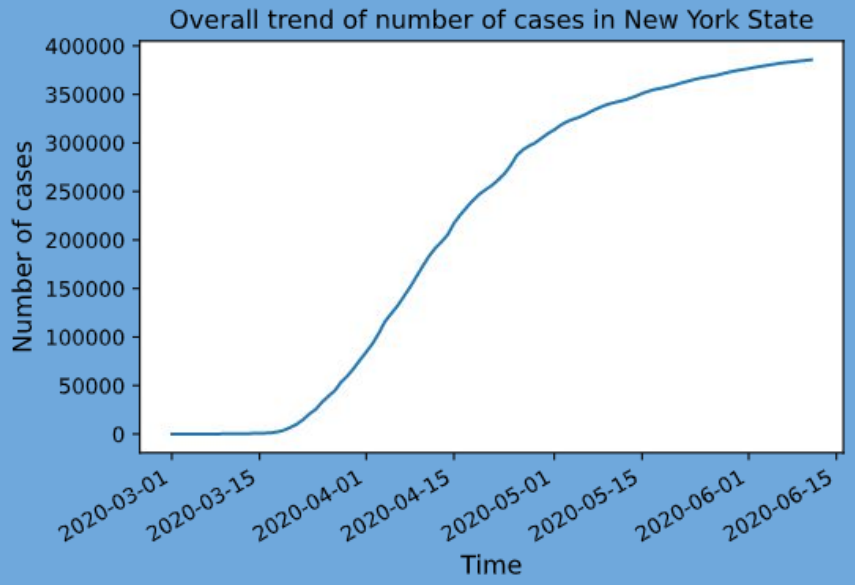
In general, the number of daily new cases in Washington is much lower than that in New York. The New York curve shows that there are a couple of surges in daily new cases before May, but the number of new cases have been dropping steadily after May. On the other hand, the Washington curve shows that the daily new cases never exceeded 1000 since the outbreak and the state is doing a fairly good job of containing the outbreak of covid-19.

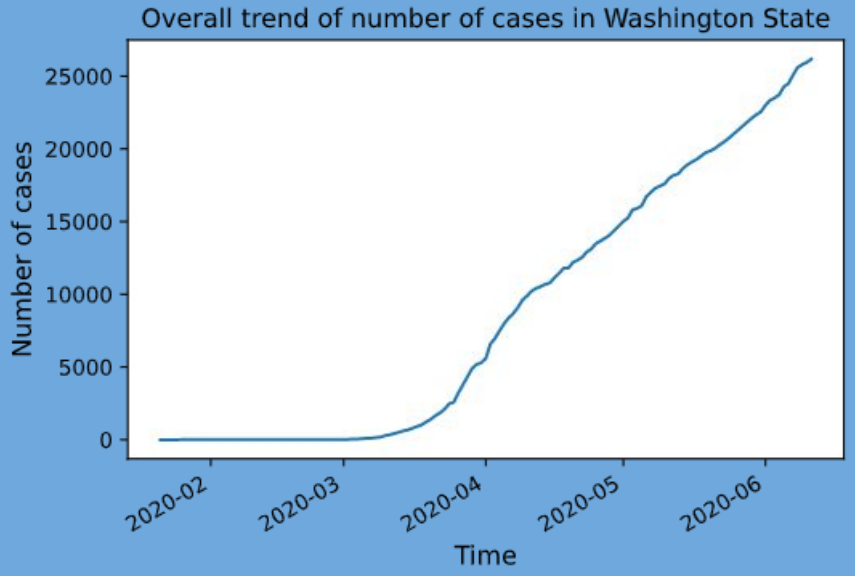




In addition, let's see the overall trend for total number of cases in New York State and Washington State.

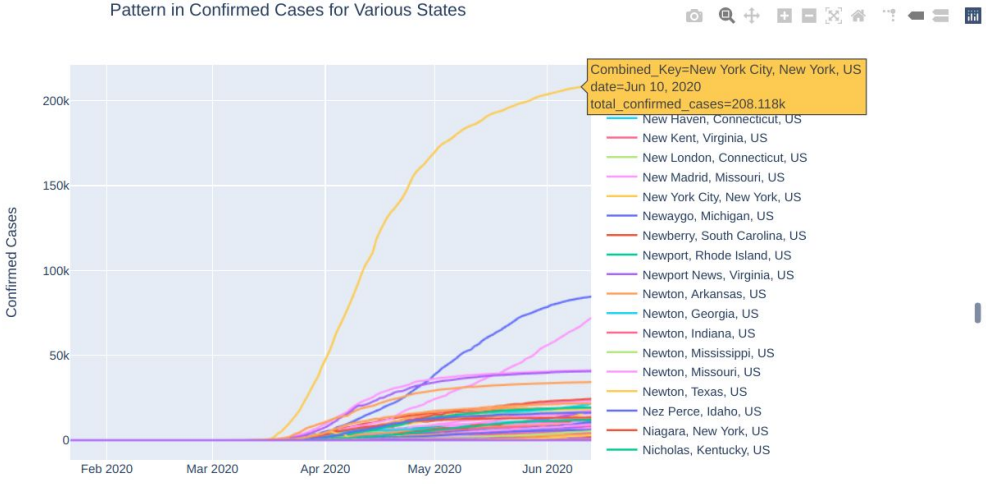
From the graphs we can see that despite the huge difference in number of cases, the New York curve starts to flatten after May whereas the Washington curve shows no signs of flattening.

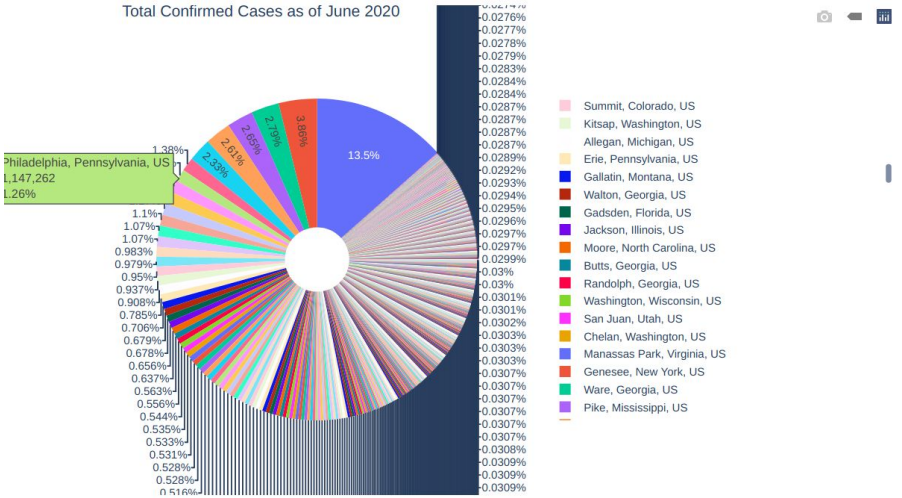




Interactive Charts

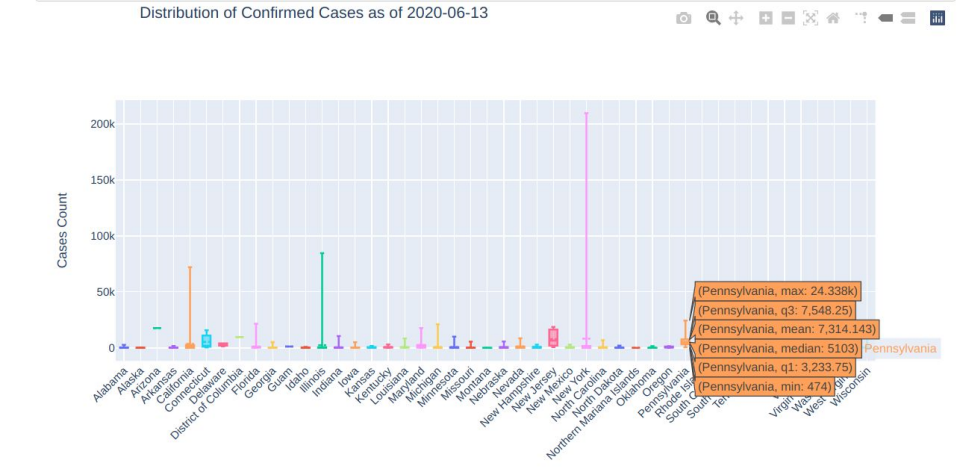
Taking a look into several states, the rate of infection in New York drastically increased than any other state. With other states reaching peak early, numbers in New York continued to rise to date with curve seemingly flattening. Proportion of confirmed cases in New York forms the bigger slice with 13.5%.

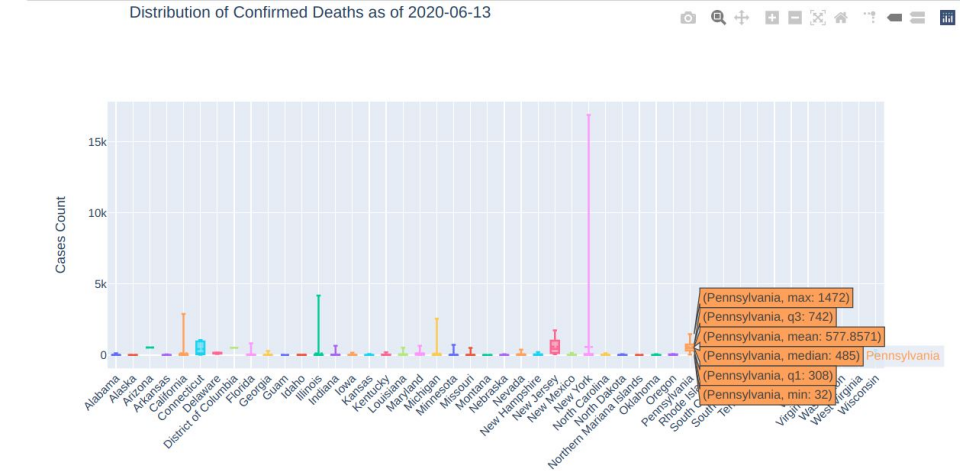




Interactive Charts

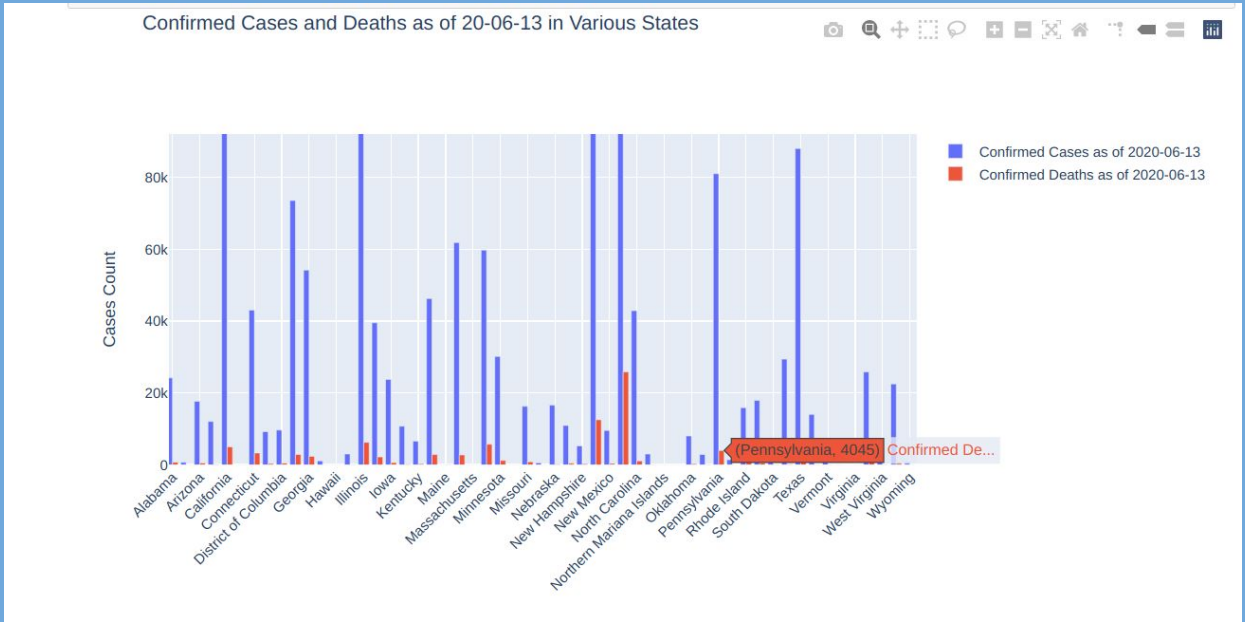
Distribution of confirmed cases and confirmed deaths for most of states shows a strong positive skew with long upper whiskers. This may be attributed to increasing infection and death rate with time.





Interactive Charts

There is a notable positive correlation between confirmed cases and confirmed deaths. A reduction in confirmed cases drastically lead to fewer deaths.



EDA Findings from COVID-19 Community Mobility dataset

When using Transit Station Percent Change to analyze mobility:

- Positive correlation happened between:
  - Transit Station Percent Change From Baseline vs. Retail and Recreation Percent Change From Baseline
  - Transit Station Percent Change From Baseline vs. Grocery and Pharmacy Percent Change From Baseline
  - Transit Station Percent Change From Baseline vs. Workplaces Percent Change From Baseline

- Negative correlation happened between:
  - Transit Station Percent Change From Baseline vs. Workplaces Percent Change From Baseline
  - Transit Station Percent Change From Baseline vs. Residential Percent Change From Baseline

When looking at various mobility factors have the most influences in which states:

- Grocery and Pharmacy Percent Change From Baseline is more likely to increase in Iowa and decrease in DC (on average).
- Parks Percent Change From Baseline is more likely to increase in Iowa and decrease in Hawaii (on average).
- Retail and Recreation Percent Change From Baseline is more likely to increase in Mississippi and decrease in DC (on average).
- Transit Stations Percent Change From Baseline is more likely to increase in Wyoming and decrease in DC (on average).

Hypotheses / Assumptions

After we cleaned and explored the datasets, we made the following assumptions and proceeded to verify them.

1. Because Covid-19 and influenza/pneumonia are both respiratory diseases, we assumed that the death rate from influenza/pneumonia in each state might be correlated with the death rate from Covid-19.
2. As most state governments have recommended social distancing and limiting travel to essential activities, we also assumed that there may be a positive correlation between community mobility and confirmed cases of and/or deaths from Covid-19.

Findings

For the assumptions, we produced and examined the correlation matrix and here are some of our findings.

- There is a high correlation between the number of influenza/pneumonia deaths and the number of tests performed.
  - Pearson Coefficient = 0.8744
- There is a high correlation between the community mobility features themselves. For example, if the rate of retail and recreation mobility increased, the rate of grocery and pharmacy mobility also increased.
  - This indicates that people may not be limiting their travel as solely for essential activities.
- Unsurprisingly, there is a high correlation between the number of confirmed cases and the number of deaths.

Future Work and Guidance

- Despite the high correlations found between various categorical variables related to COVID-19 from these three datasets, we plan to use more data analysis methods to further investigate in-depth findings and our hypotheses. For instance, we can perform logistic regression to see the significance of categorical variables based on the data we have.