COVID-19 Data Comparison

Lawrence Guloy
Elif Evrim Polat
Alex Goodman

Hypothesis

States and Cities with strong mitigation efforts were more effective at reducing the spread of covid than those that had weak mitigation efforts.

Questions?

Can we use trends in the data to predict case counts?

Does the data accurately show spikes around holidays and other social events?

Were COVID policies effective or not?

Did a lack of certain policies change the outcome of case numbers?

Data Collecting

- Needed to see rates of covid infections over a set period of time and observe any increases or decreases in the rates
- Needed data from different states or cities with different policies to see if the policies actually affected the outcome
- Predicting the trajectory of cases is now difficult with the vaccine. It is an unknown variable in how it will effect case numbers. We do not have the data to show the effect the vaccine will have.

Data Found Here

https://api.covidtracking.com/v1/states/fl/daily.json
https://corona.lmao.ninja/v2/historical/usacounties/florida?lastdays=all
https://data.cityofnewyork.us/resource/rc75-m7u3.json?
http://corona-api.com/countries/US
https://api.covidactnow.org/v2/state/WA.timeseries.json?

Exploring the Data Sources

Searching for API’s with the data we needed was harder than expected

Due to COVID being such a recent phenomenon, finding comprehensive data was a challenge. Many data sets were incomplete.

Covid Tracking Project was very helpful, but not a direct source/federal project

Question the validity of the data we received

Miami - no actual city data, had to use county data and infer city counts from it

Many API’s were made by other developers pulling from government sources

Insights

The daily death counts were all much lower than anticipated

There were points in the data that didn’t seem to follow the expected trends, leading us to question if the data was reported correctly

Confused as to why data ends before the end of the year in NYC data

Cleanup Process

Interested in cases per day instead of accumulated case numbers but some did not have cases per day. We had to go through accumulated case numbers and use calculations to get the daily case numbers.

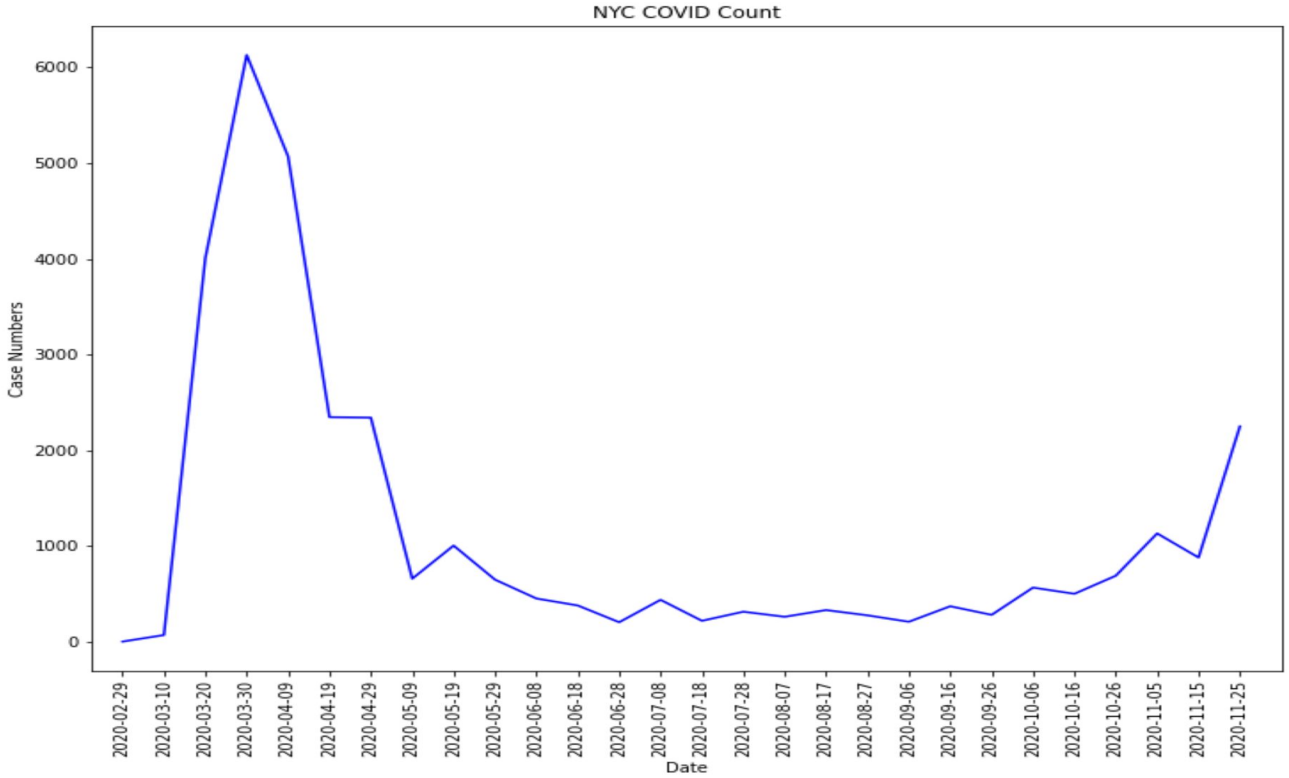
Multiple data points for the same day. Needed to clean the data to make sure only one data point per day

Organization of the json files were different for each dataset so we had to cross check the original data source and translate it so we were pulling the correct data.

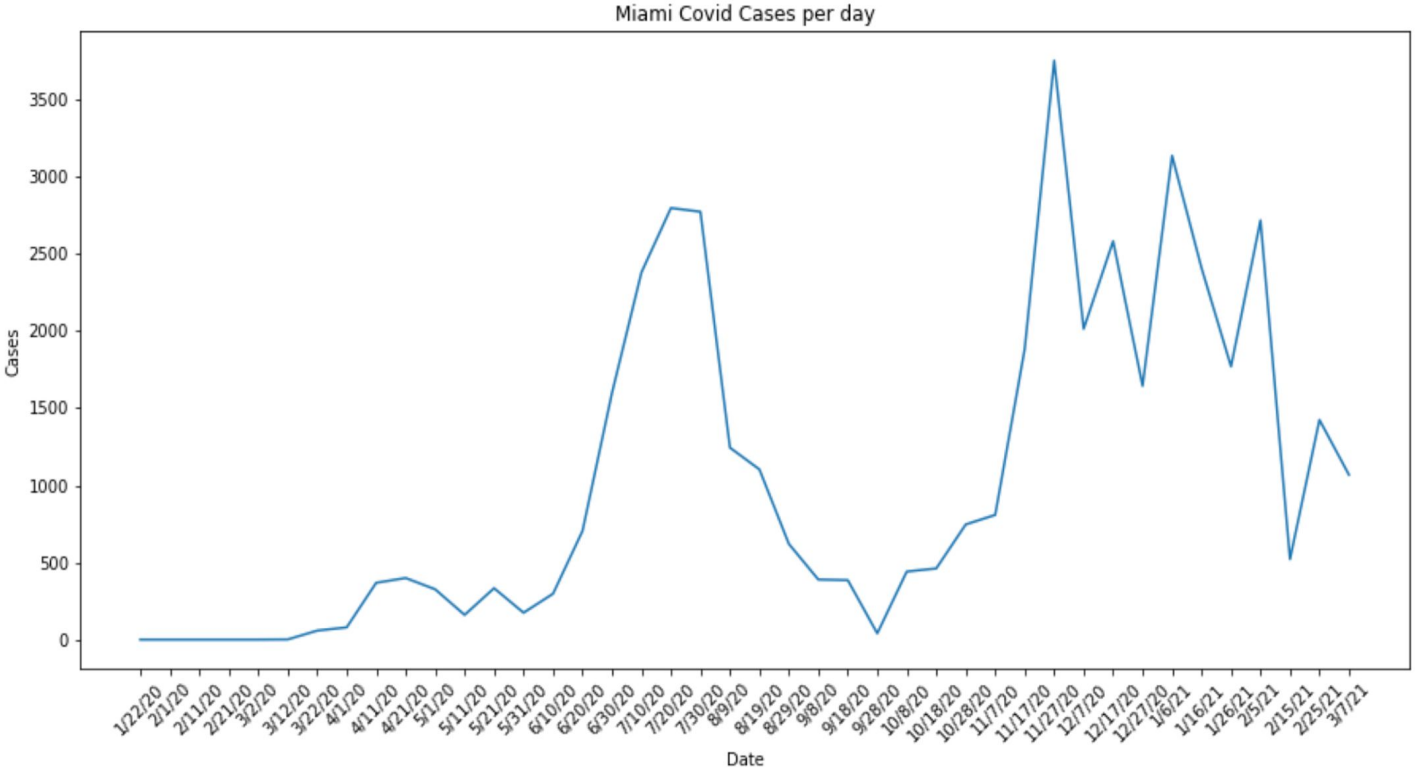
Data Analysis

Does the data accurately show spikes around holidays and other social events?

New York City Case Count

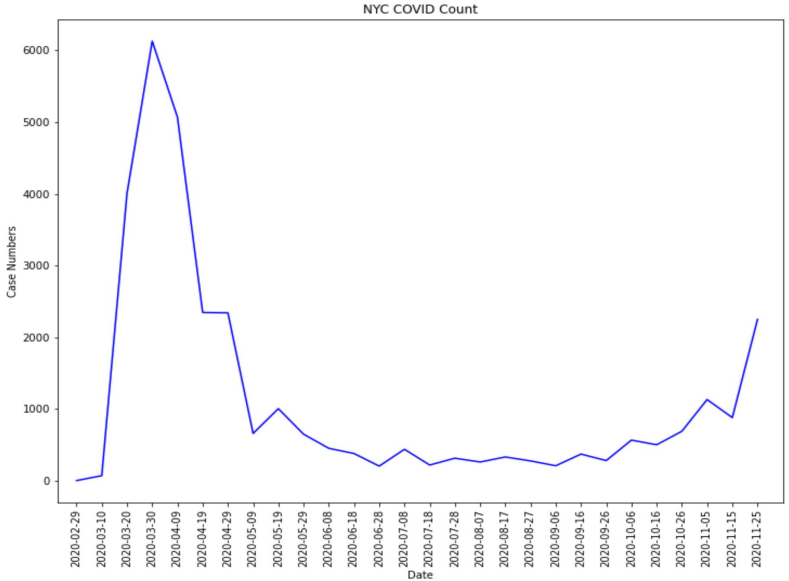


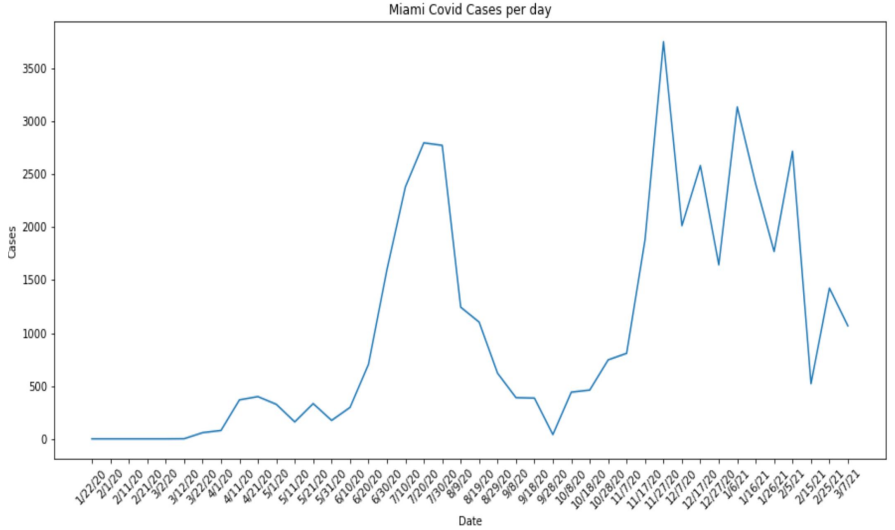
Miami Case Count



City Comparison

New York City vs Miami



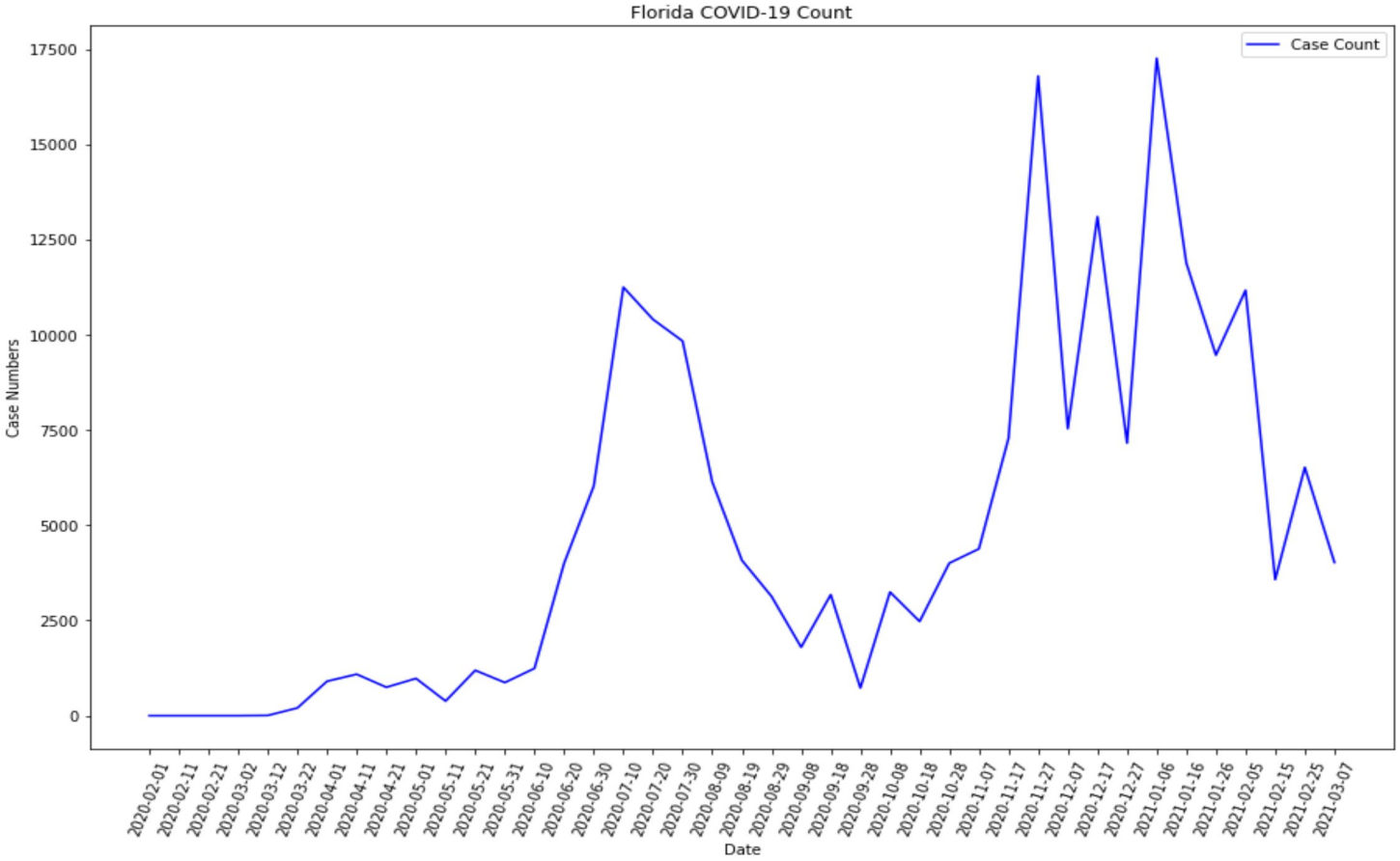


NYC Total Case Number - 771k (NY state total - 1.72m - 44.83% of state total)
Percent of US total - 2.63%
Population - 8.4M (9.17% infected)

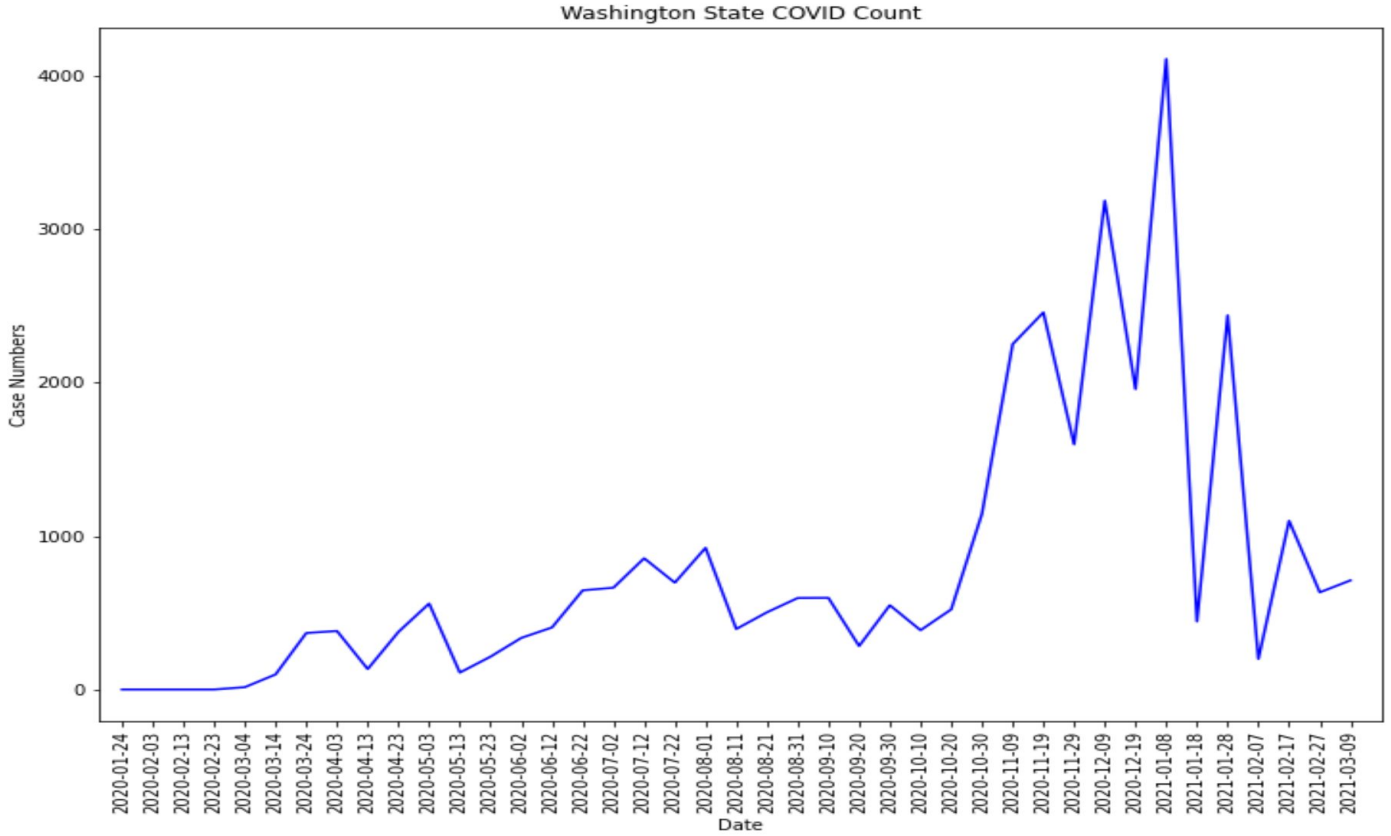
Miami Total Case Number - 424k (FL state total - 1.96m - 21.63% of state total)
Percent of US total - 1.45%
Population - 2.7M (15.7% infected)

Florida Case Count

-

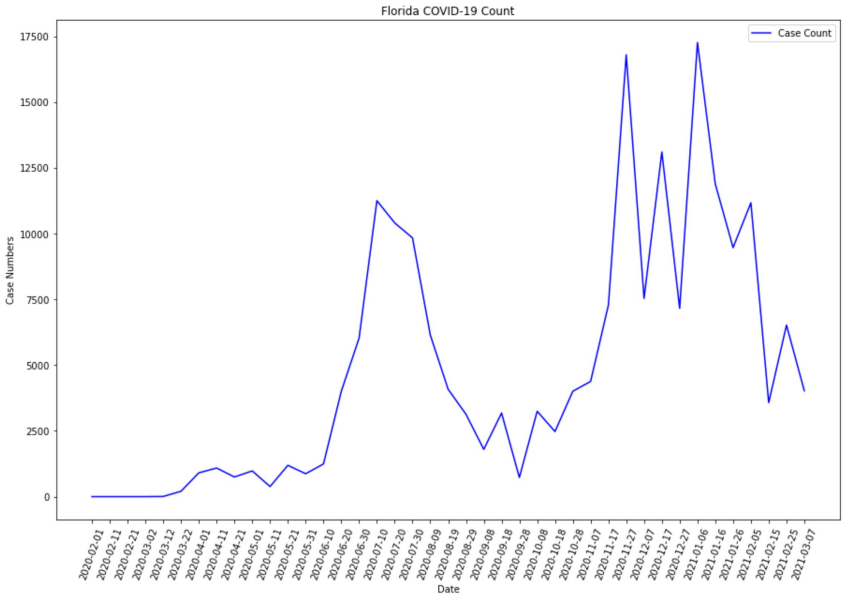


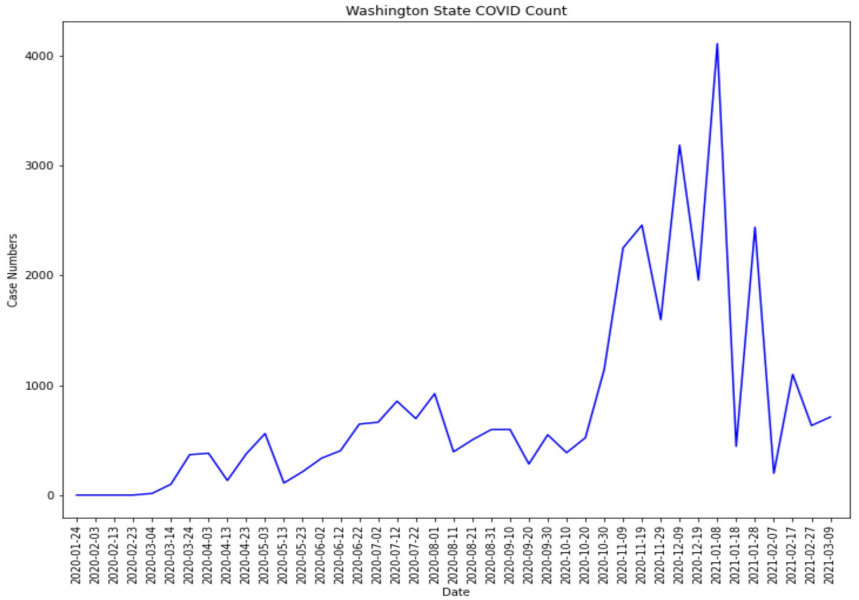
Washington State Case Count



State Comparison

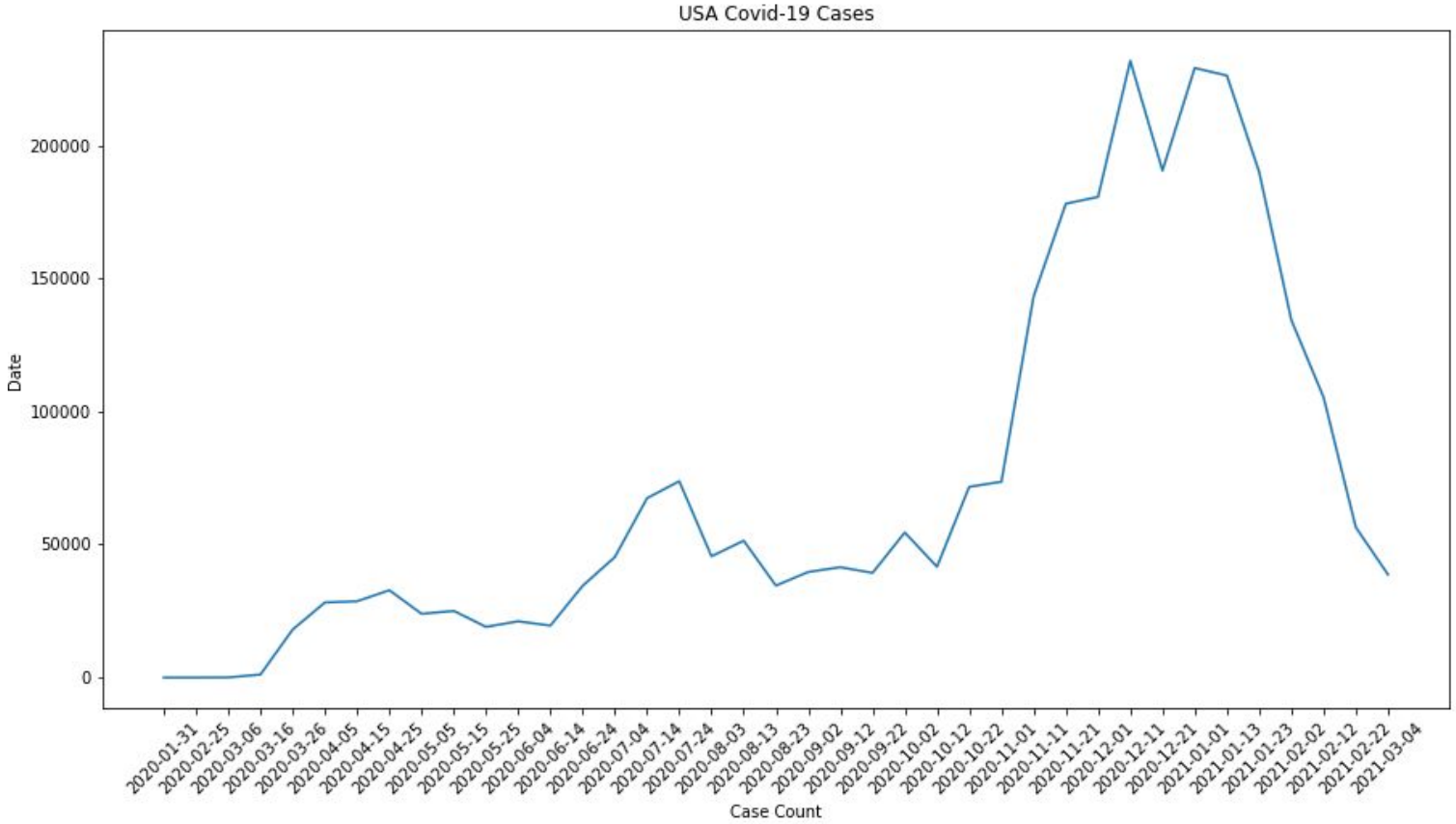
Florida vs Washington State





Florida total case numbers - 1.96M
Percent of US total - 6.69%
Population - 21.4M (9.15% infected)

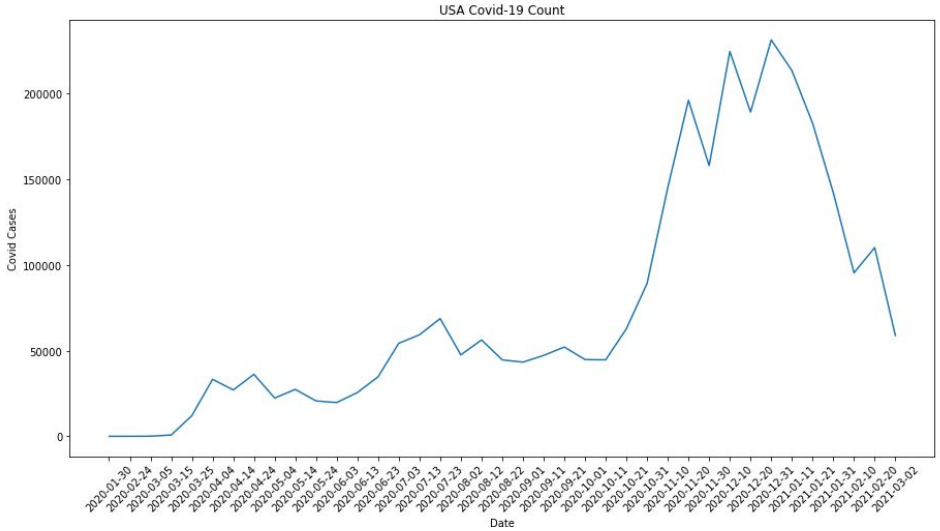
Washington Total Case Number - 351K
Percent of US total - 1.2%
Population - 7.6M (4.6% infected)

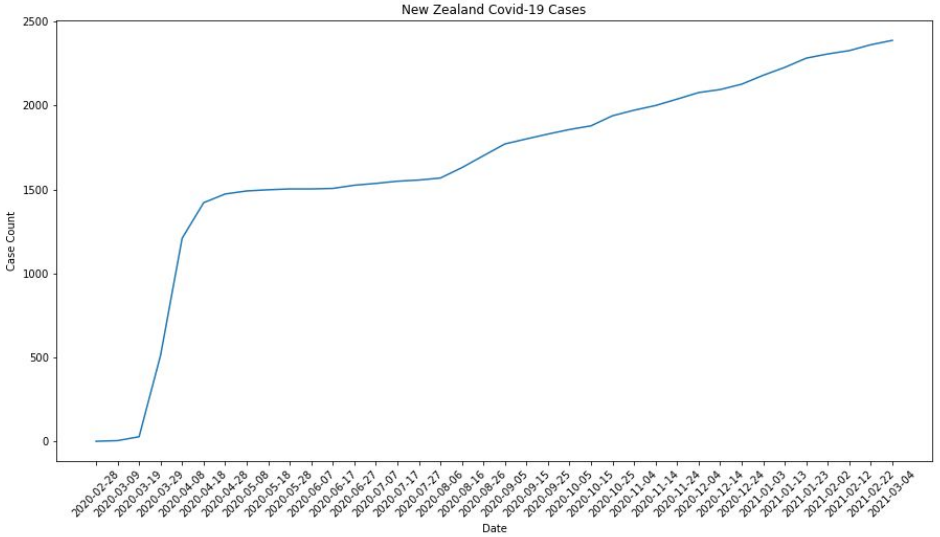

USA Covid-19 Cases
Date
Case Count
0
50000
100000
150000
200000
2020-01-31
2020-02-25
2020-03-06
2020-03-16
2020-03-26
2020-04-05
2020-04-15
2020-04-25
2020-05-05
2020-05-15
2020-05-25
2020-06-04
2020-06-14
2020-06-24
2020-07-04
2020-07-14
2020-07-24
2020-08-03
2020-08-13
2020-08-23
2020-09-02
2020-09-12
2020-09-22
2020-10-02
2020-10-12
2020-10-22
2020-11-01
2020-11-11
2020-11-21
2020-12-01
2020-12-11
2020-12-21
2021-01-01
2021-01-13
2021-01-23
2021-02-02
2021-02-12
2021-02-22
2021-03-04

Data Analysis

Were COVID policies effective or not?

Country Comparisons:





Can we use trends in the data to predict case counts?

Observations from the previous questions show us that there are multiple factors that affect the case counts

- Holidays
- Policies
- Vaccinations
- And possibly more unknowns

Limitations on the data

- Non-linear data
- Partly qualitative
- Source fidelity

So the answer?

Quantitatively, no we cannot predict future case numbers to an accurate degree due to the limitations of the data and the number of events that can affect the rate of cases

However, we can guess over a short time frame in the future. We can take a linear regression, but the question now becomes whether that prediction is useful.

Revisit the Hypothesis

States and Cities with strong mitigation efforts were more effective at reducing the spread of covid than those that had weak mitigation efforts.

At first glance of the data, it seems like it would but due to the effects of other factors such as vacations, personal habits, and communal response, it is hard to say that mitigation efforts alone had the most effect in reducing covid spread rates

Challenges

- Finding accurate data.
- Data cleaning or dropping data.
- Working around the limitations of the data.
- Changing our questions based on the data that we found.
- More questions were raised.
- Representative data.

Additional Considerations

- How has covid-19 affected the infection rates of other diseases
- Does technology or more developed countries have an effect in spread rates?
- Does wealth have to do anything with covid spread?
- Analyzing and comparing death tolls.
- More data from different states

Any Questions?