COVID-19: Exploratory Data Analysis

Authors:
Alissa Stover
Alyssa Augsburger
John Lee
Nathan Nusaputra
Ryan Sawasaki

May 2020

Github Repository: https://github.com/CSJohnLee/COVID19_Project

Introduction

Background:

Classmates of “ Python Fundamentals for Data Science” collaborated to conduct an exploratory data analysis (EDA) on COVID-19 data. What they found was not contrary to popular beliefs and validates expectations.

Objective:

Better understand and visualize COVID-19 data.

Our Curiosities:

1) How do different states compare?
2) How does population and income affect number of cases?

Data-Sourcing & Processing

Problem:
COVID-19 data is coming in in real-time, from multiple sources, and with different data management practices. Sifting through this information is time-consuming and can be a barrier to data use.

Solution:
We compared the data quality of 3 sources and documented our findings so that other researchers could more easily choose which data to use for their own explorations. The code we used to source and process the data is documented on our Github, so that other researchers can use our code to download the most recent data themselves.

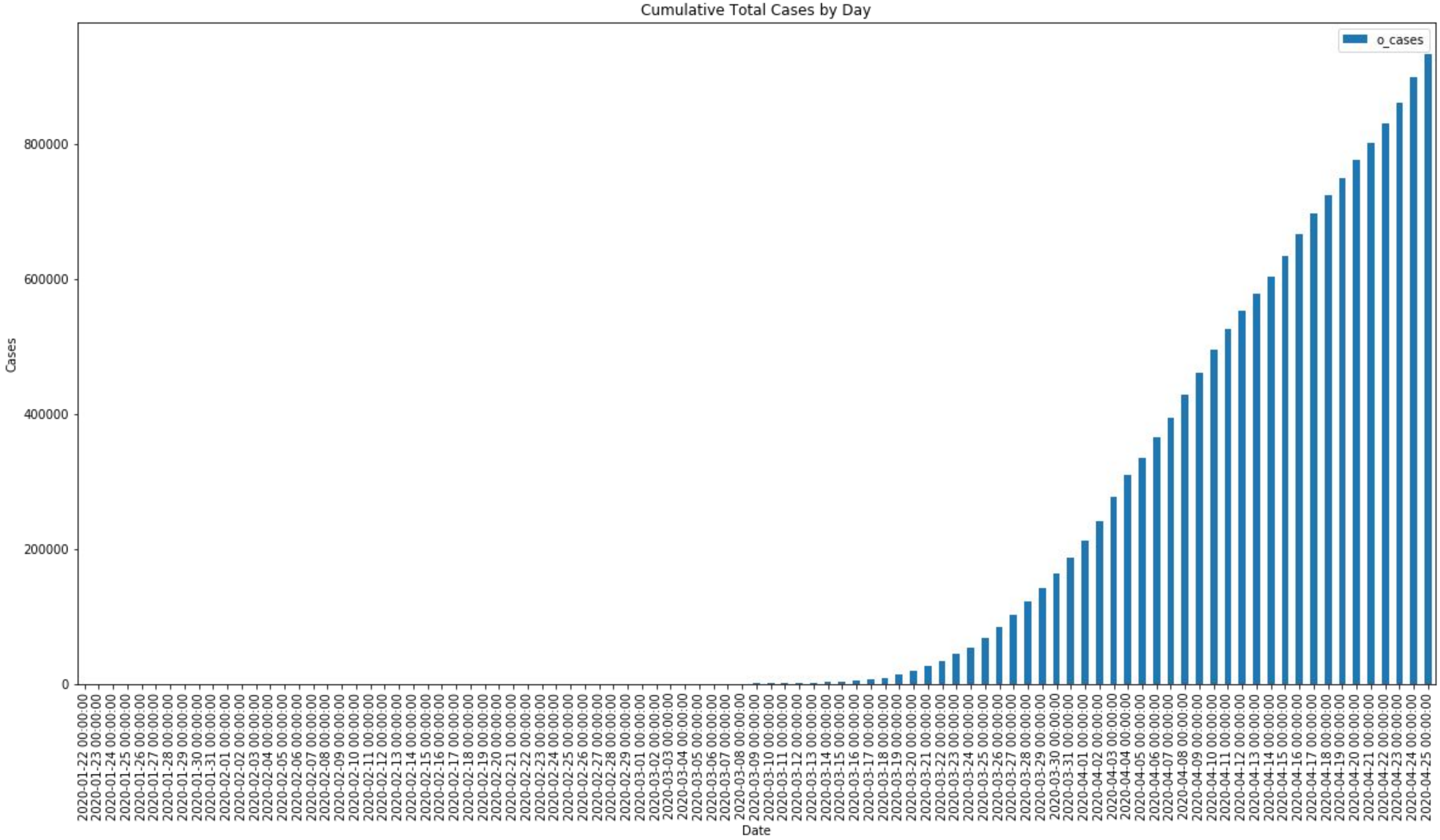
- The datasets were not large, so we were able to process them using Python in a Jupyter Notebook, by connecting to their respective website using using HTTP requests.
- To assess data quality, we focused on:
  - which information was contained in each dataset
  - what level the data are on
  - how complete the data were
  - whether the data could be joined to other sources

Data Quality Assessment

1. **New York Times COVID-19 Github repository:** https://github.com/nytimes/covid-19-data
   While this dataset was utilized for a few analyses, we recommend not using this source in the future, as the sources below seemed to have a more ideal format
   - Advantages:
     - Already in long format, at the state and county level
     - Contains information about time, place, and COVID cases/deaths
   - Disadvantages:
     - Mix of county versus city level data, so many FIPS codes are missing
     - The way data are updated makes it unclear why there are discrepancies in cases/deaths for each date for a given location, and thus which count should be used
2. **USA Facts:** https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
   We recommend using this source for county-level COVID analyses
   - Advantages:
     - County-level COVID cases, deaths, and overall population data are well documented with clean FIPS codes. These can be merged together and on other data sources.
   - Disadvantages:
     - In wide format with columns for every date, which means that it must be transformed to long format for most analyses
3. **COVID Tracking Project:** https://covidtracking.com/data
   We recommend using this source for state-level COVID analyses
   - Advantages:
     - Very detailed COVID data, including information about what treatments people are getting (e.g., how many people are in ICU)
     - Clean FIPS codes can be merged on other data.
   - Disadvantages:
     - Only at the state level

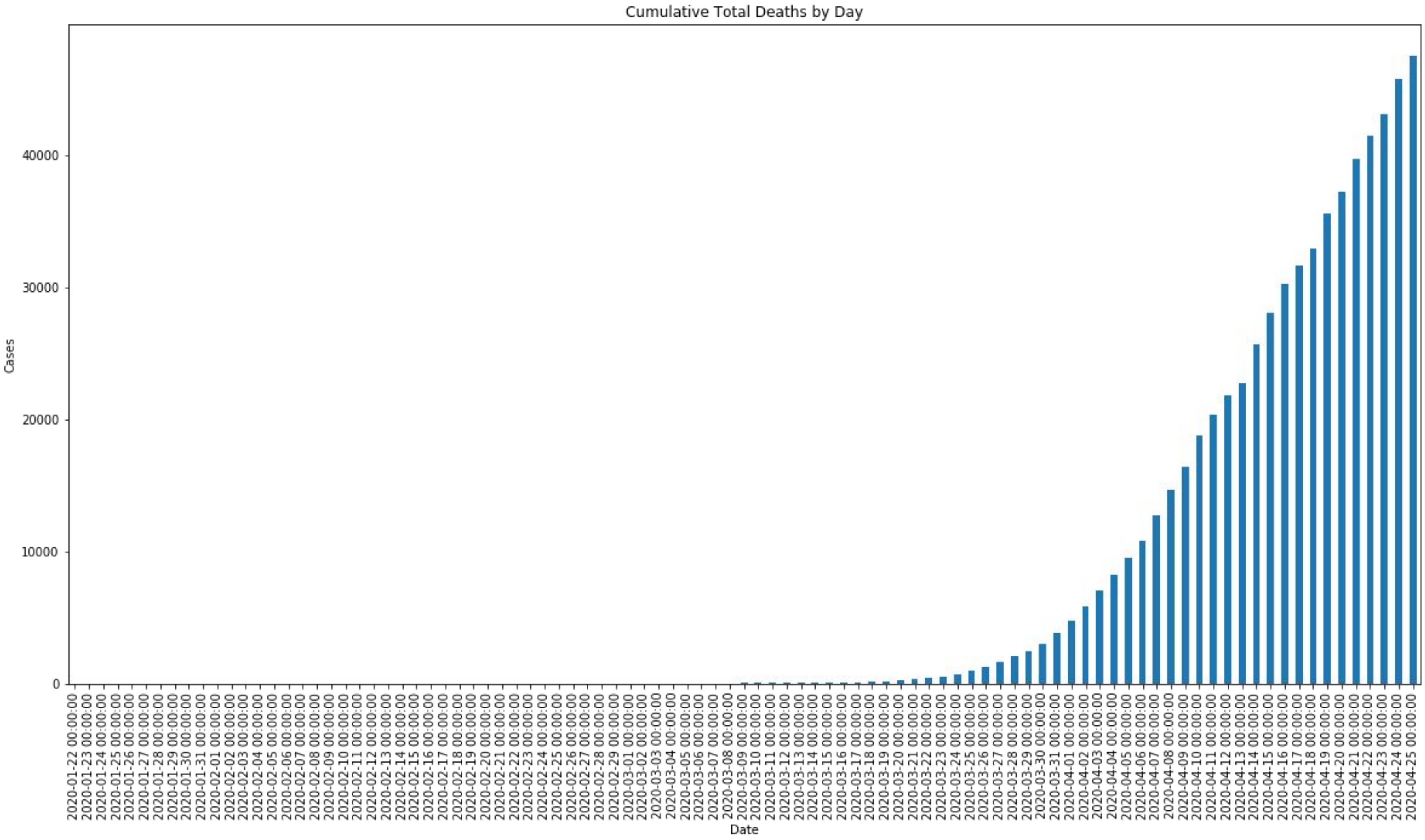
Introductory Plots

Cumulative Cases by Day



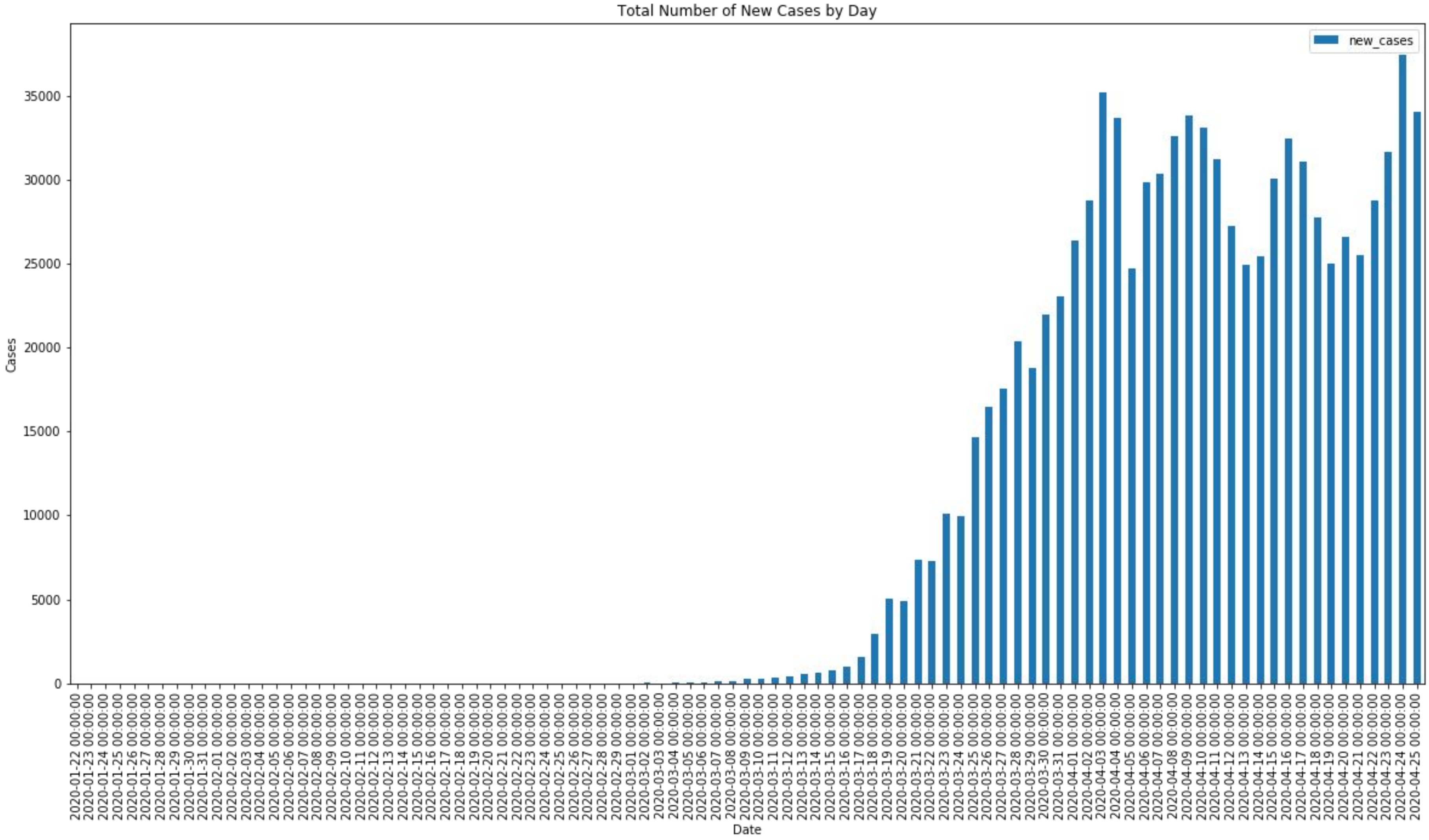
Data Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Cumulative Deaths by Day



Data Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

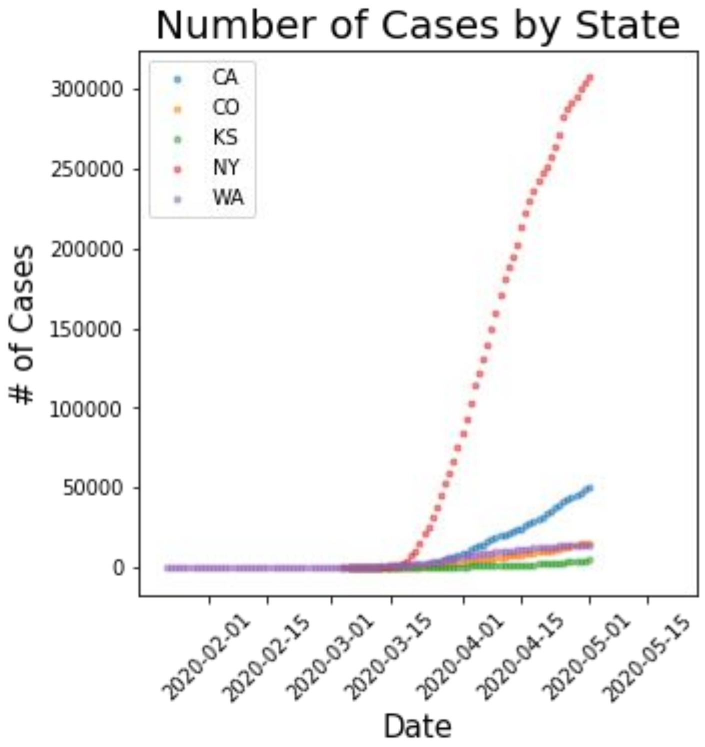
New Cases by Day

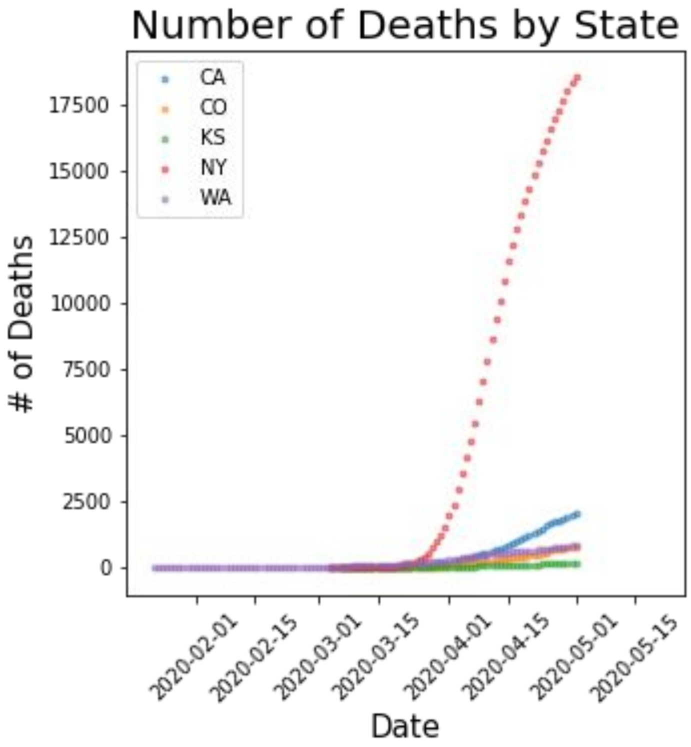


Data Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

How do different states compare?

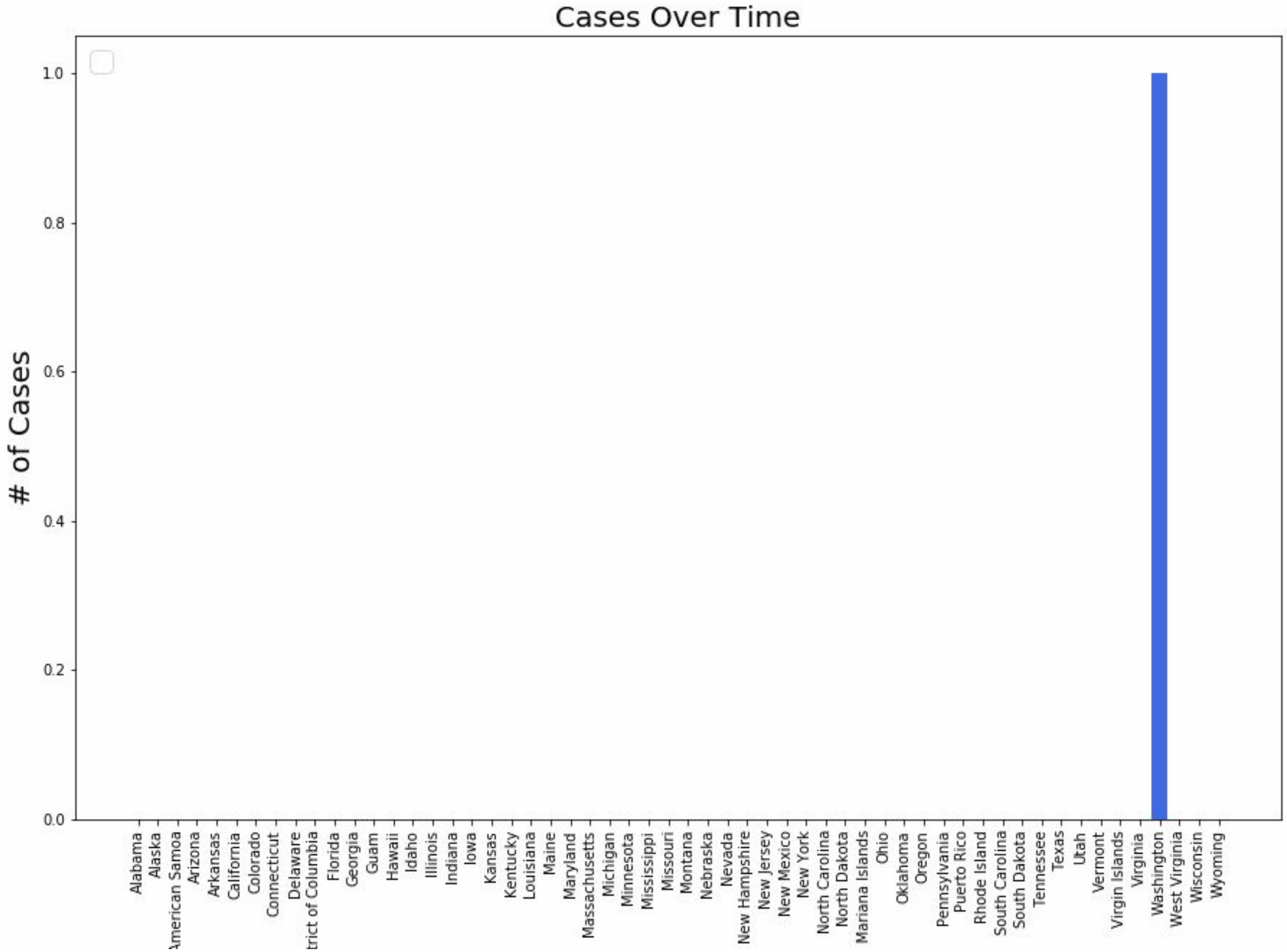
Out of our home states, New York is doing worse.


Number of Cases by State


Number of Deaths by State

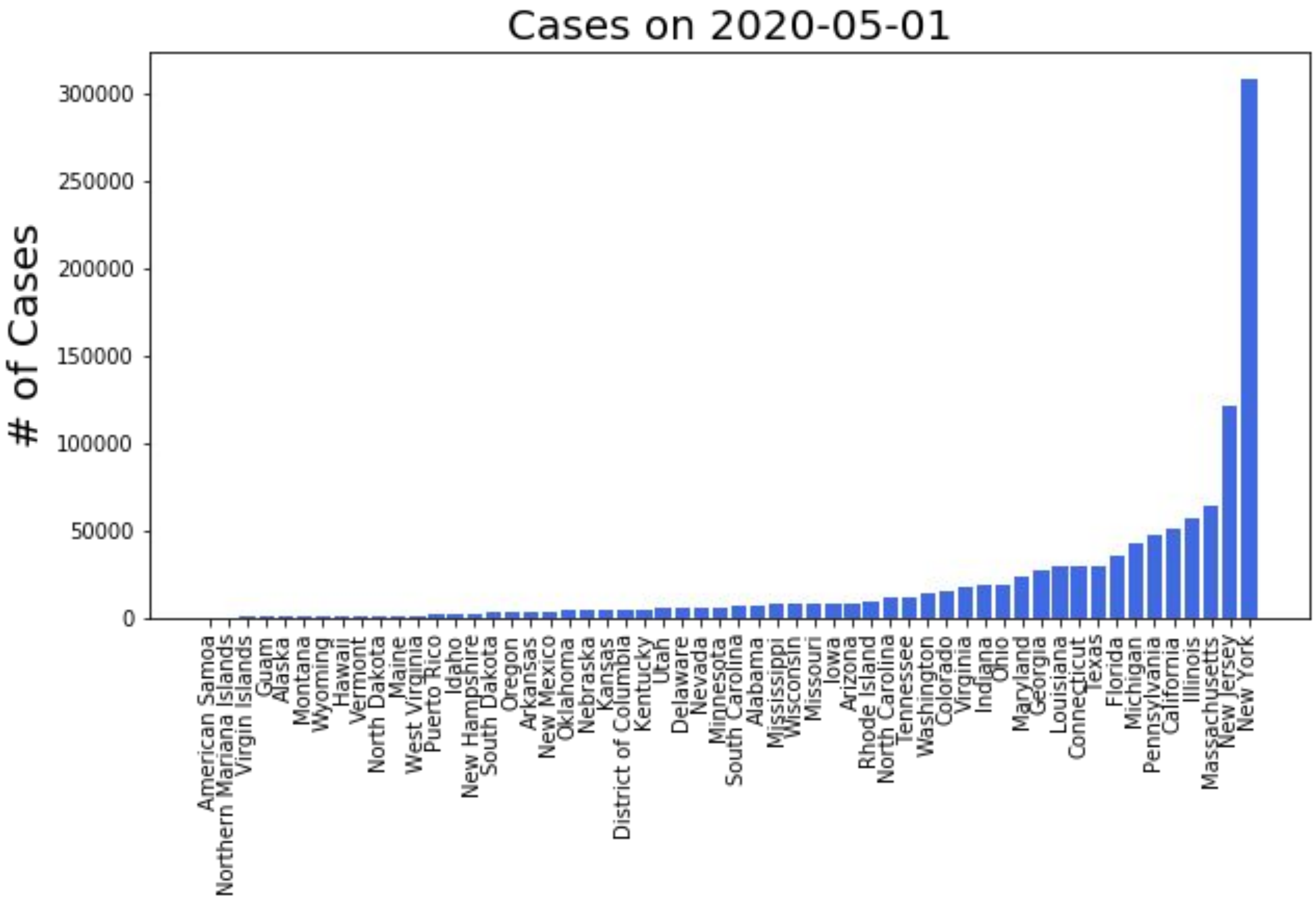
Data Source: https://covidtracking.com/data

Cases Over Time. Reported cases increase late February.



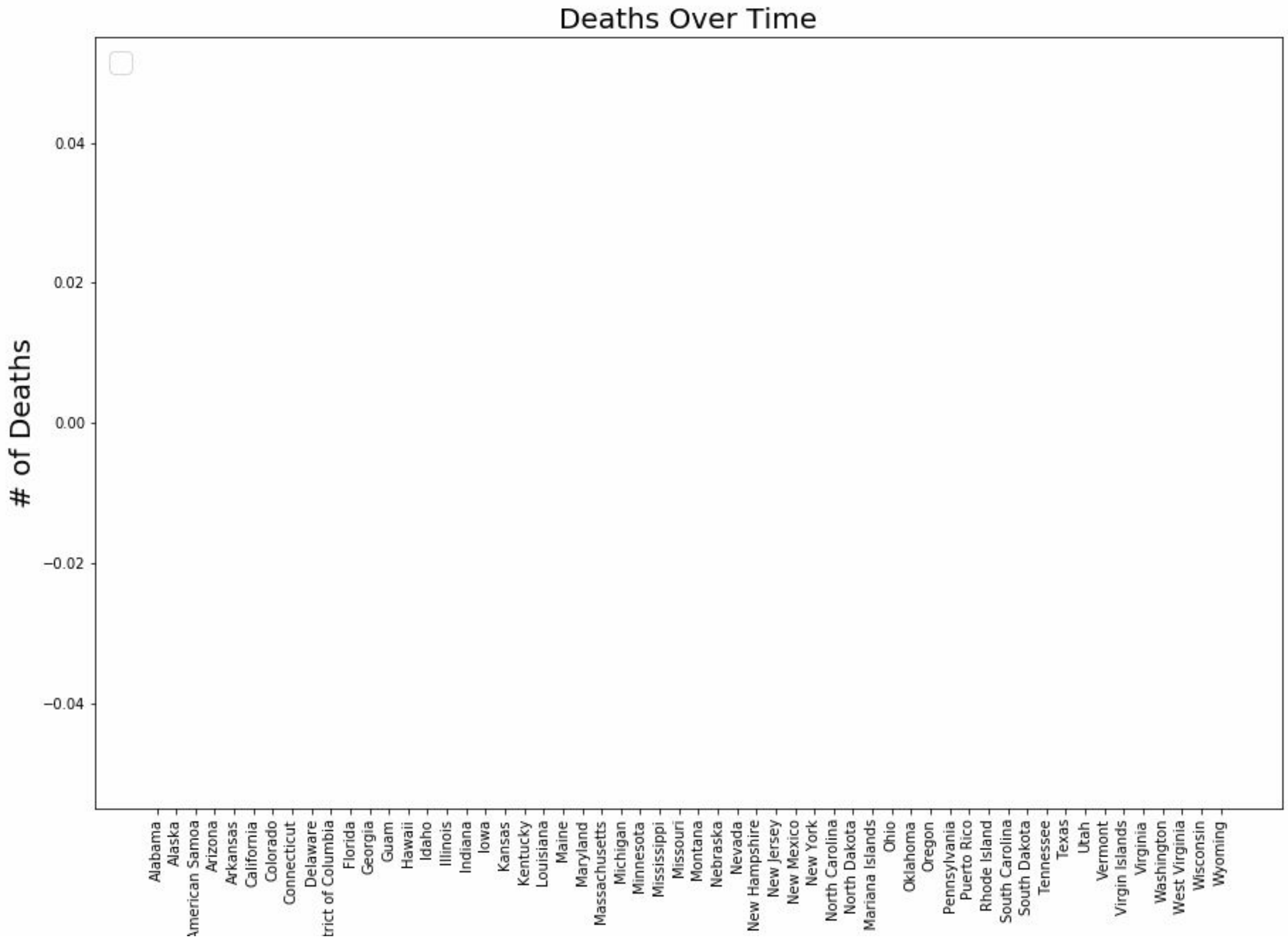
Data Source: https://covidtracking.com/data

Cases per State (as of 5/1/2020)



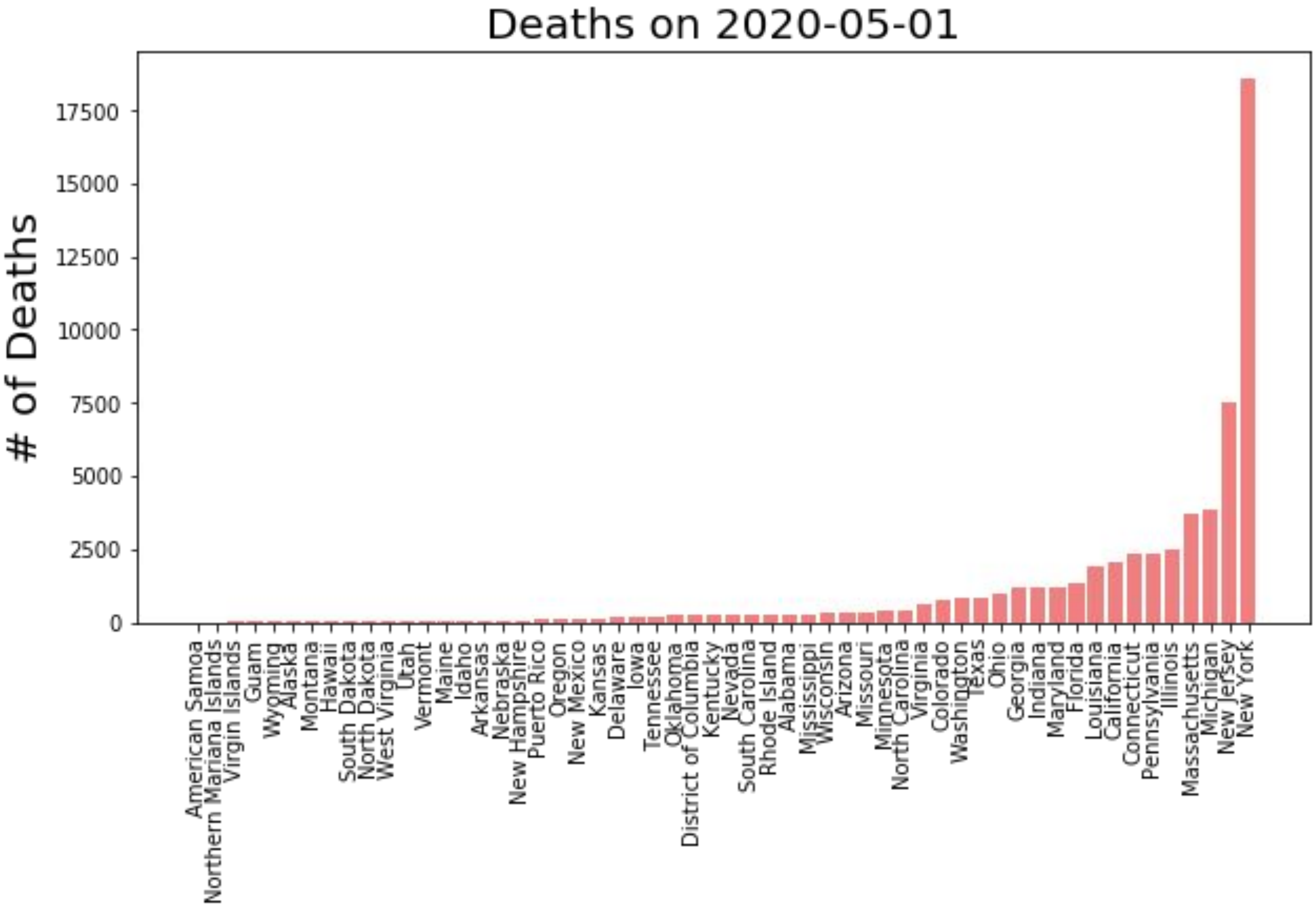
Data Source: https://covidtracking.com/data

Deaths Over Time. Reported deaths begin late February.



Data Source: https://covidtracking.com/data

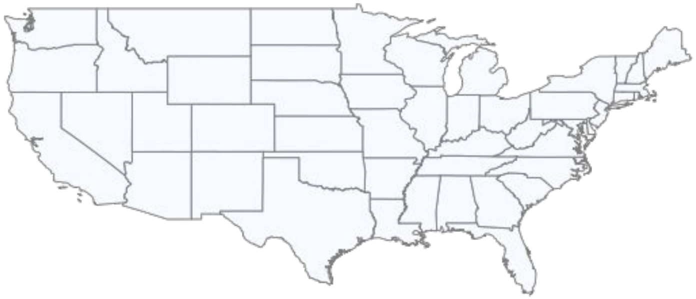
Deaths per State (as of 5/1/2020)



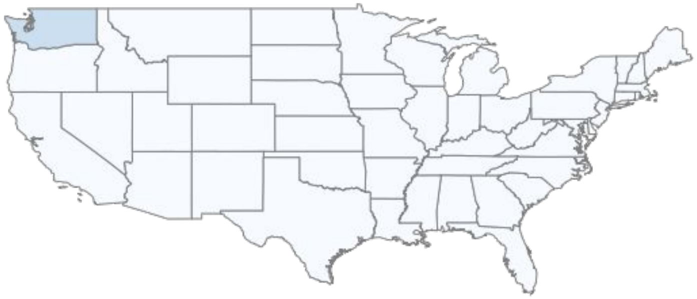
Data Source: https://covidtracking.com/data

How are cases spreading?

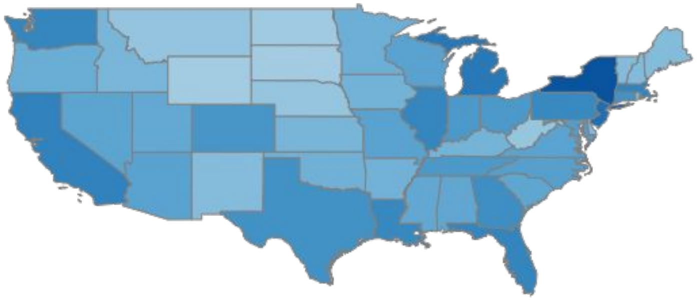
2020-01-31 COVID-19 Cases



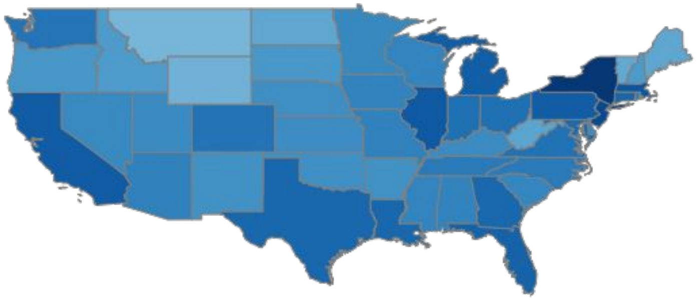
2020-02-29 COVID-19 Cases



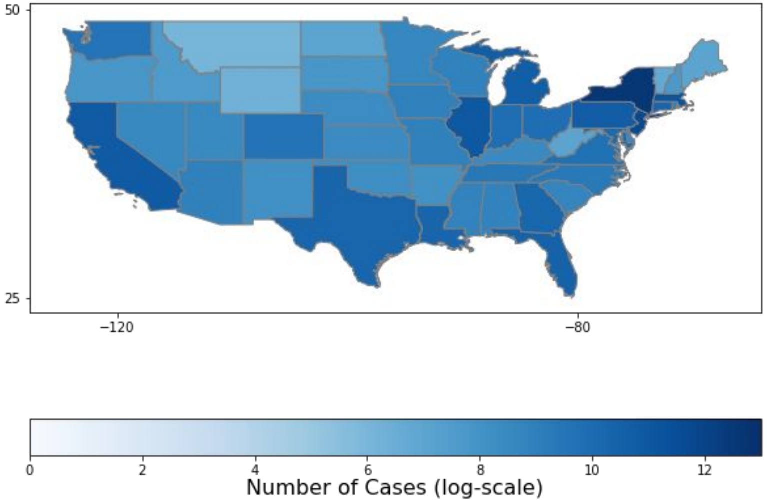
2020-03-31 COVID-19 Cases



2020-04-30 COVID-19 Cases

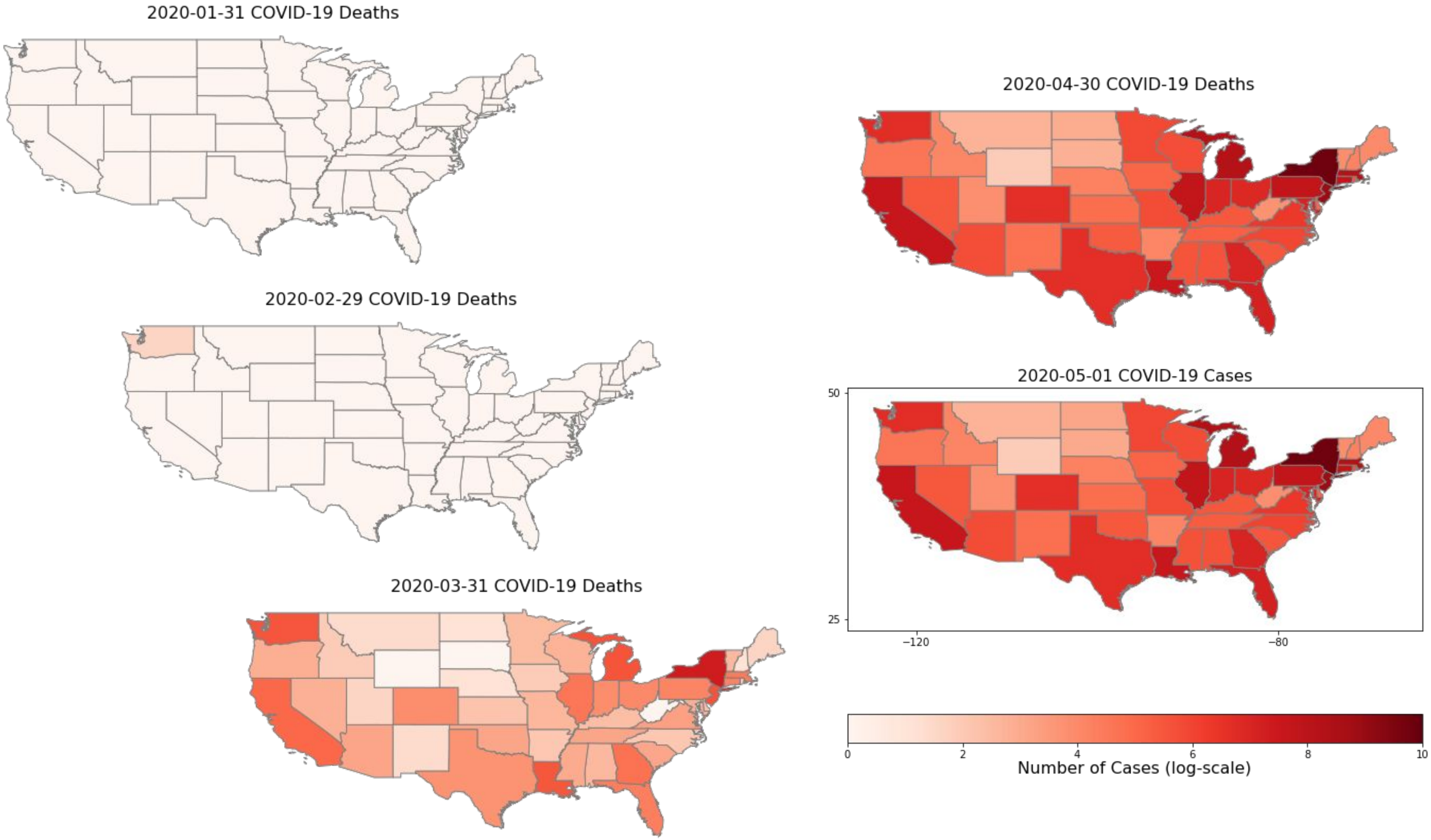


2020-05-01 COVID-19 Cases



Data Source: https://covidtracking.com/data

How are deaths spreading?



2020-01-31 COVID-19 Deaths

2020-02-29 COVID-19 Deaths

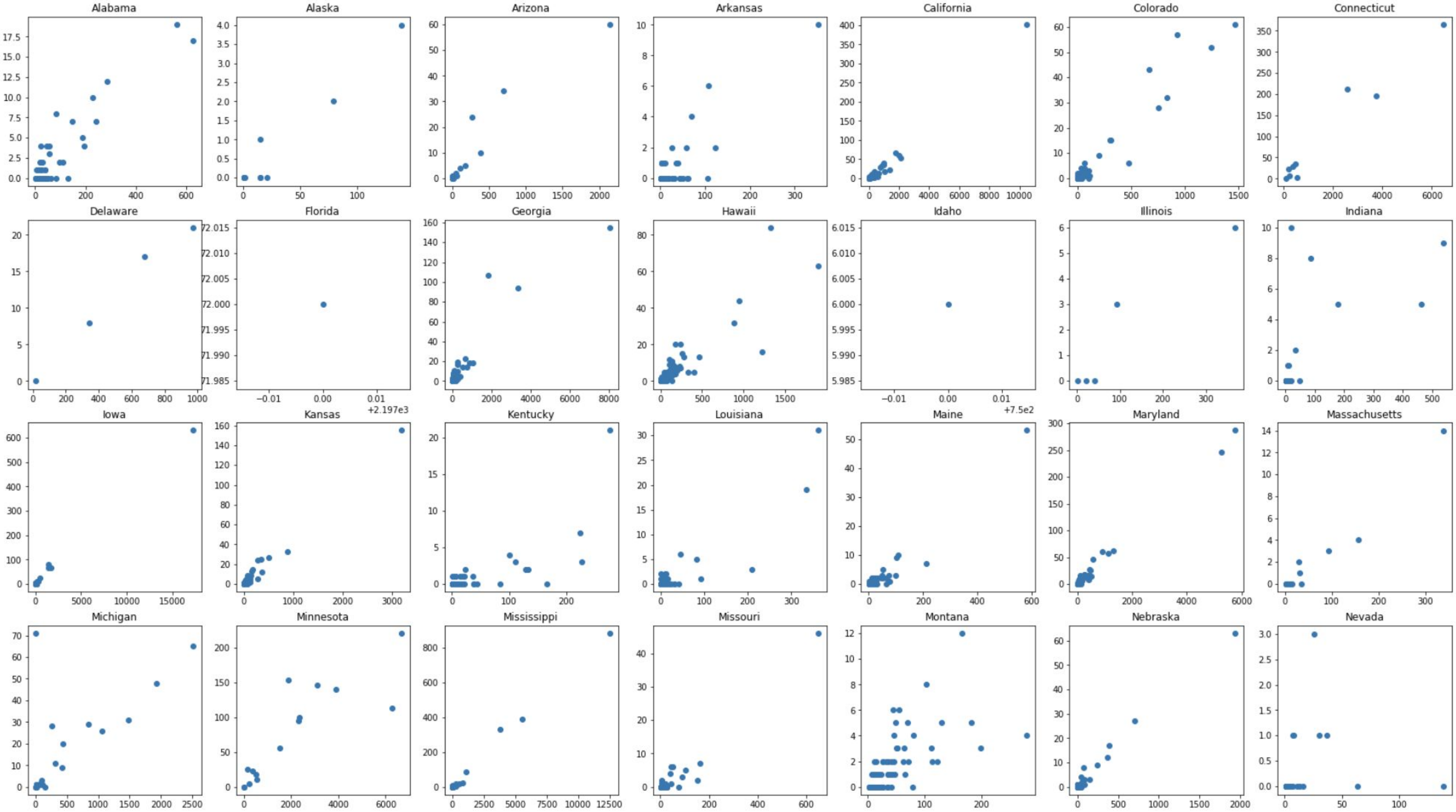
2020-03-31 COVID-19 Deaths

2020-04-30 COVID-19 Deaths

2020-05-01 COVID-19 Cases

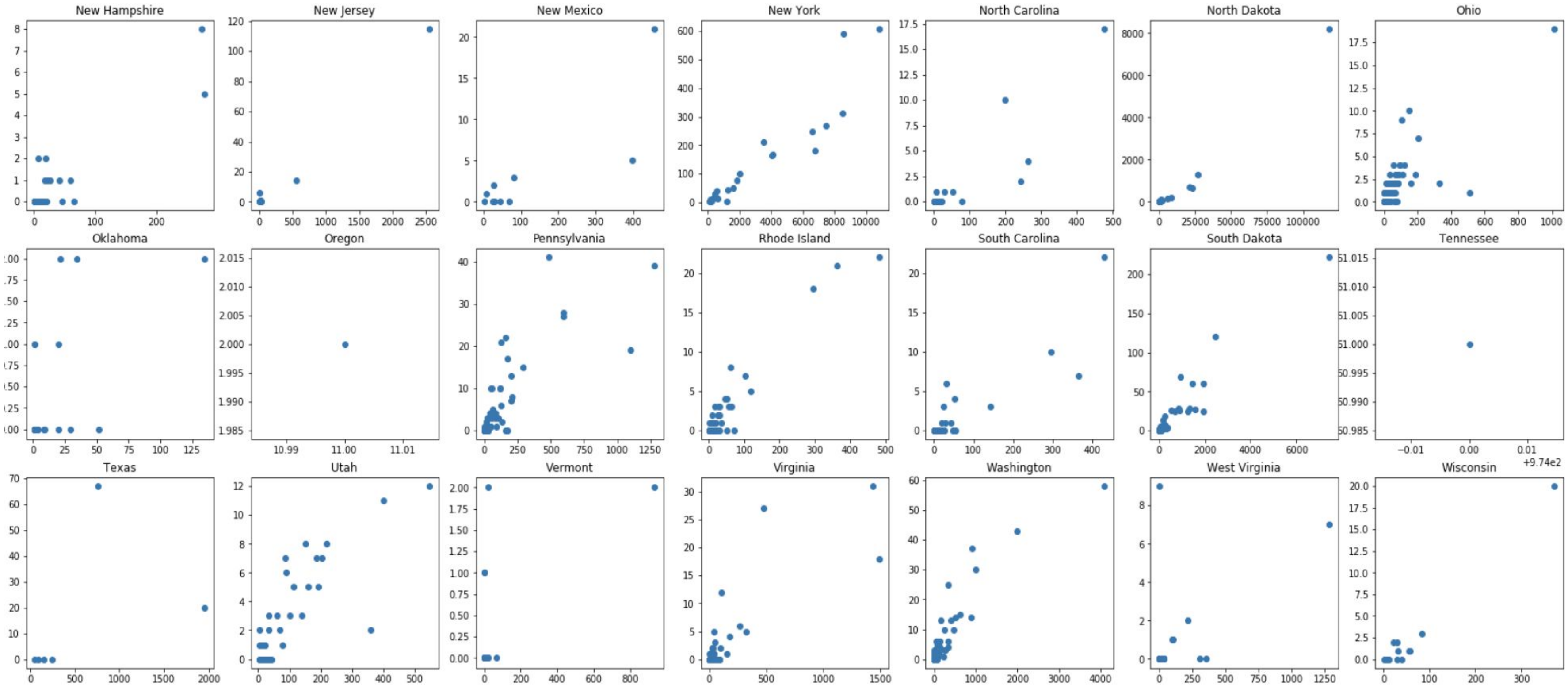
Data Source: https://covidtracking.com/data

Cases vs. Deaths Part 1



Data Source: https://github.com/nytimes/covid-19-data

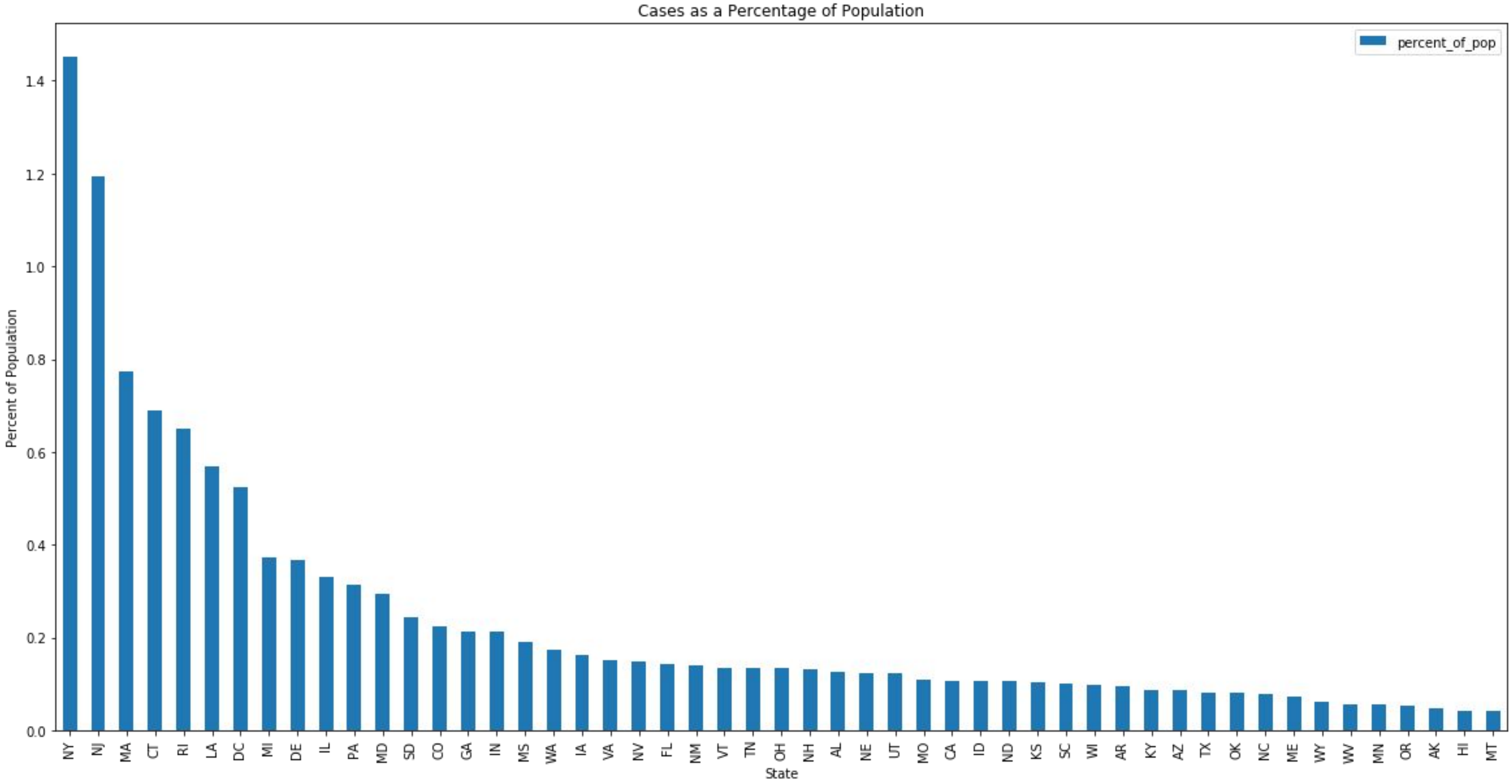
Cases vs. Deaths Part 2



Data Source: https://github.com/nytimes/covid-19-data

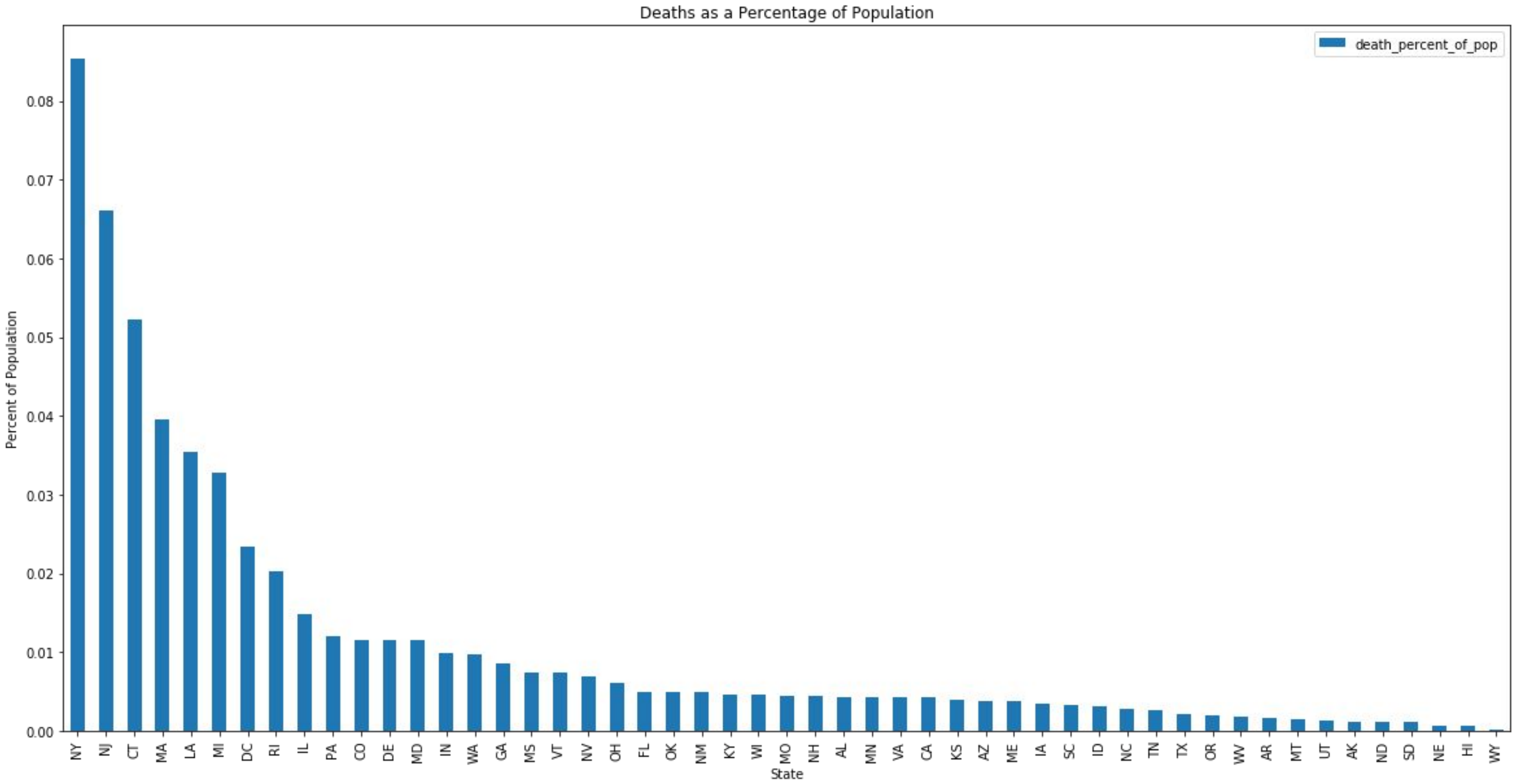
How does population affect number of cases?

Cases as a Percentage of Population



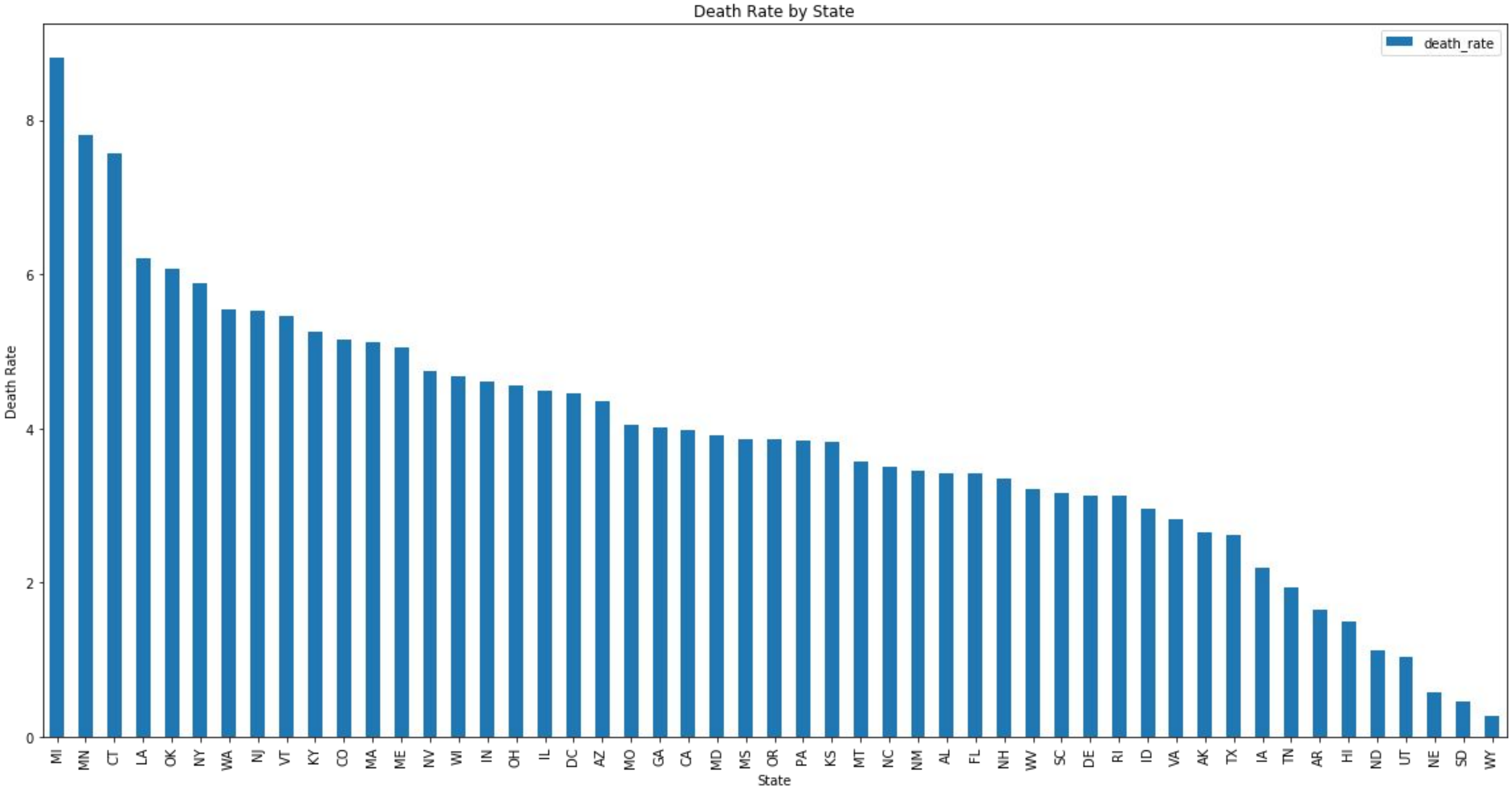
Data Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Deaths as a Percentage of Population



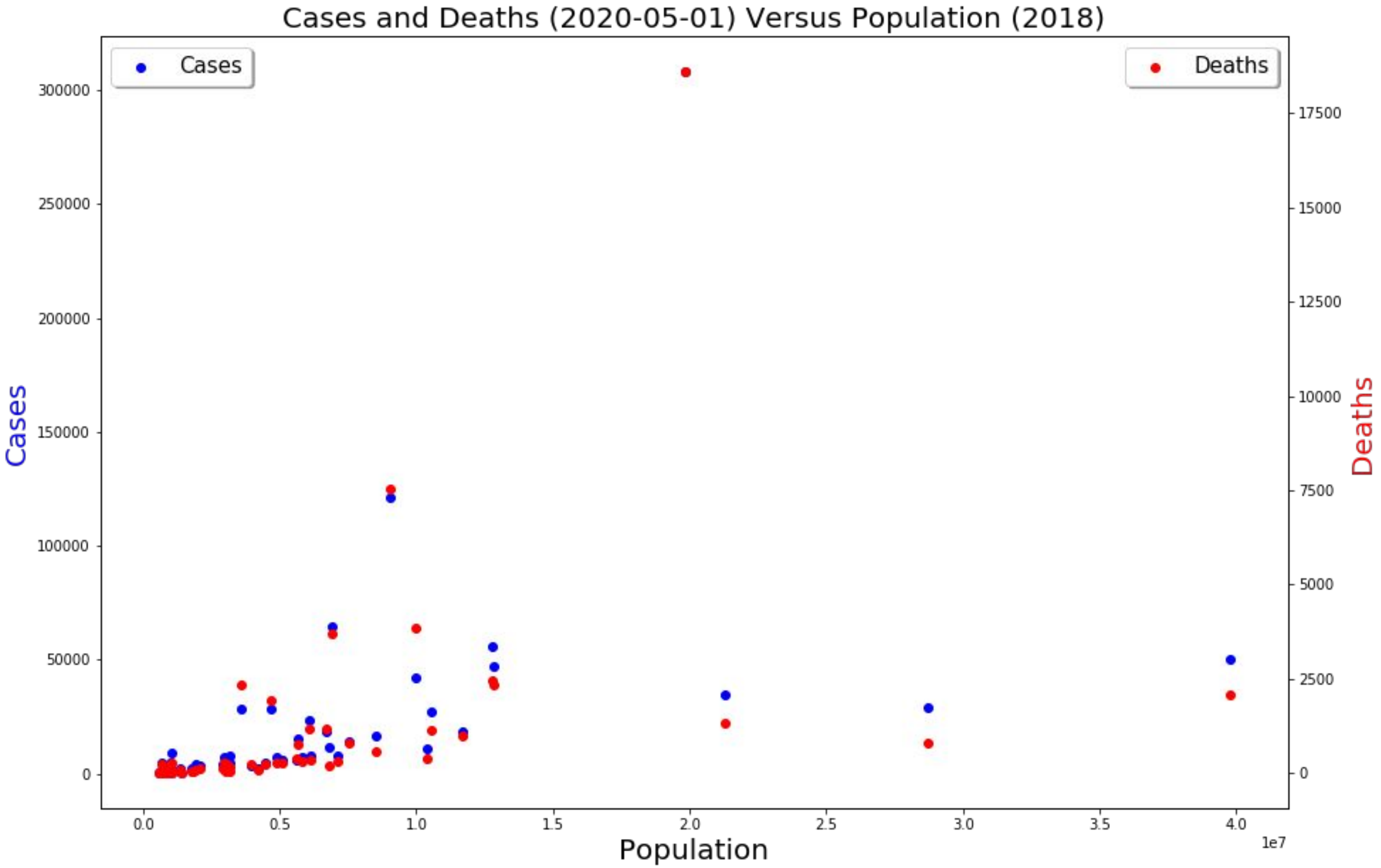
Data Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Death Rate (deaths per case) by State



Data Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

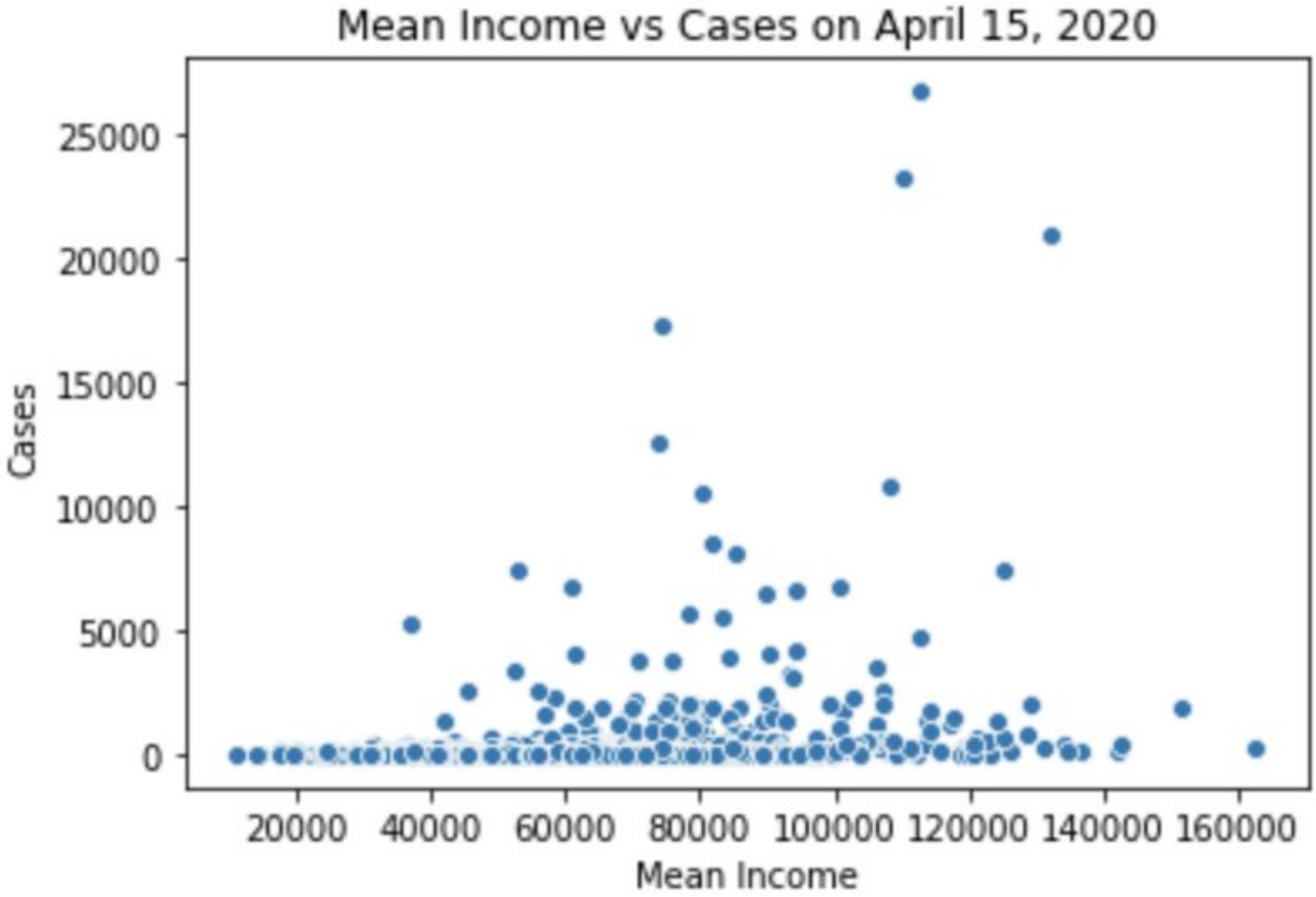
Cases & Deaths Versus State Populations



Data Source: https://covidtracking.com/data https://www.kaggle.com/lucasvictor/us-state-populations-2018#State%20Populations.csv

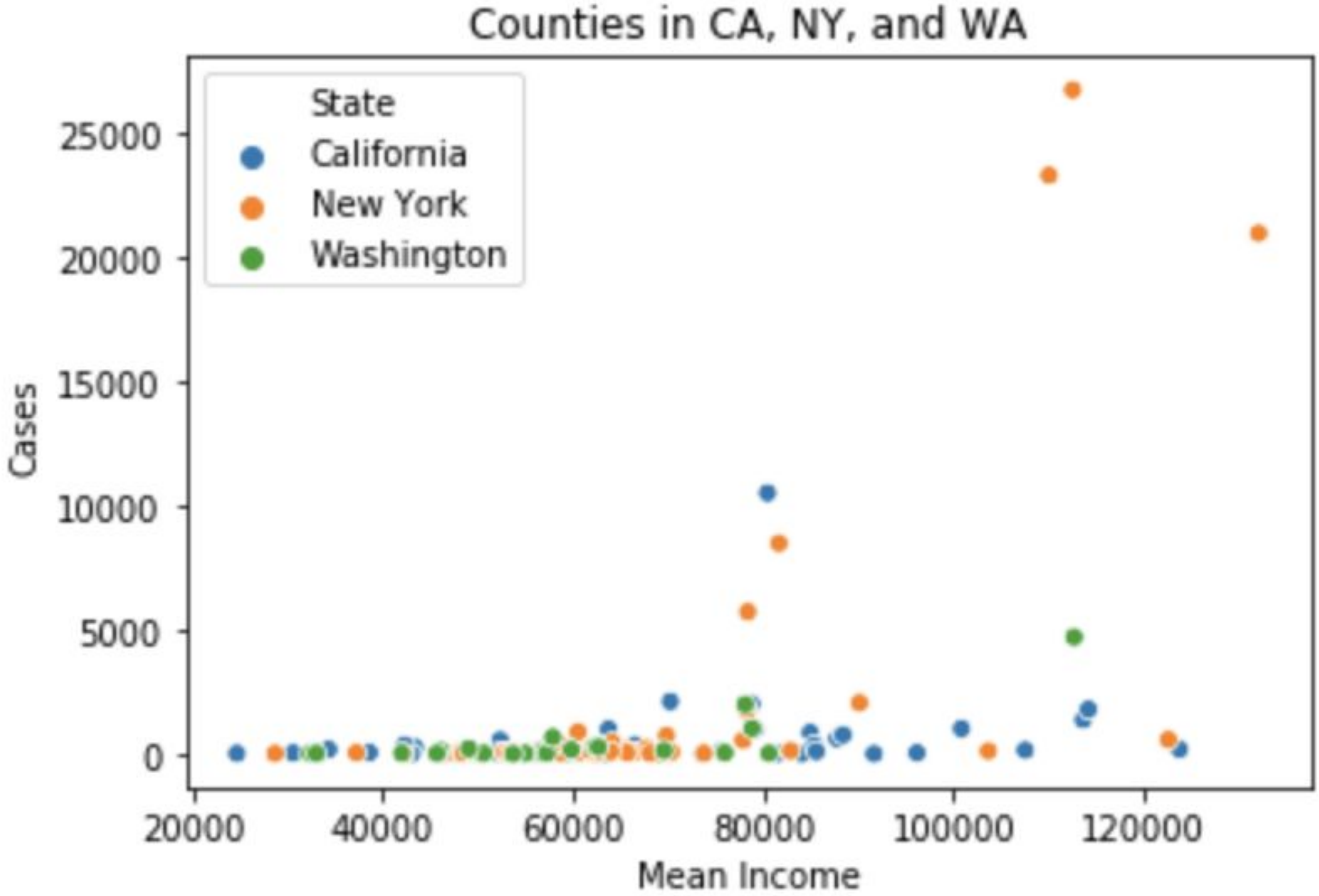
How does income affect number of cases?

Effect on Wealth



Data Source: https://github.com/nytimes/covid-19-data https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

Comparing Counties in Specific States



Data Source: https://github.com/nytimes/covid-19-data https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

Conclusion

- By 5/01/2020, cases and deaths exceed 800,000 and 40,000, respectively.
- New York has the highest number of cases and deaths.
- By looking at number of cases per capita, New York still has highest cases. However, New York is ranked 6 in number of deaths per number of cases. Michigan is number 1 in deaths per number of cases.
- Based on the map, the most populated states appear to have the highest number of cases and deaths.
- As expected, the more cases there are, the more deaths there may be.
- There’s not a strong visual relationship between mean income and number of cases.

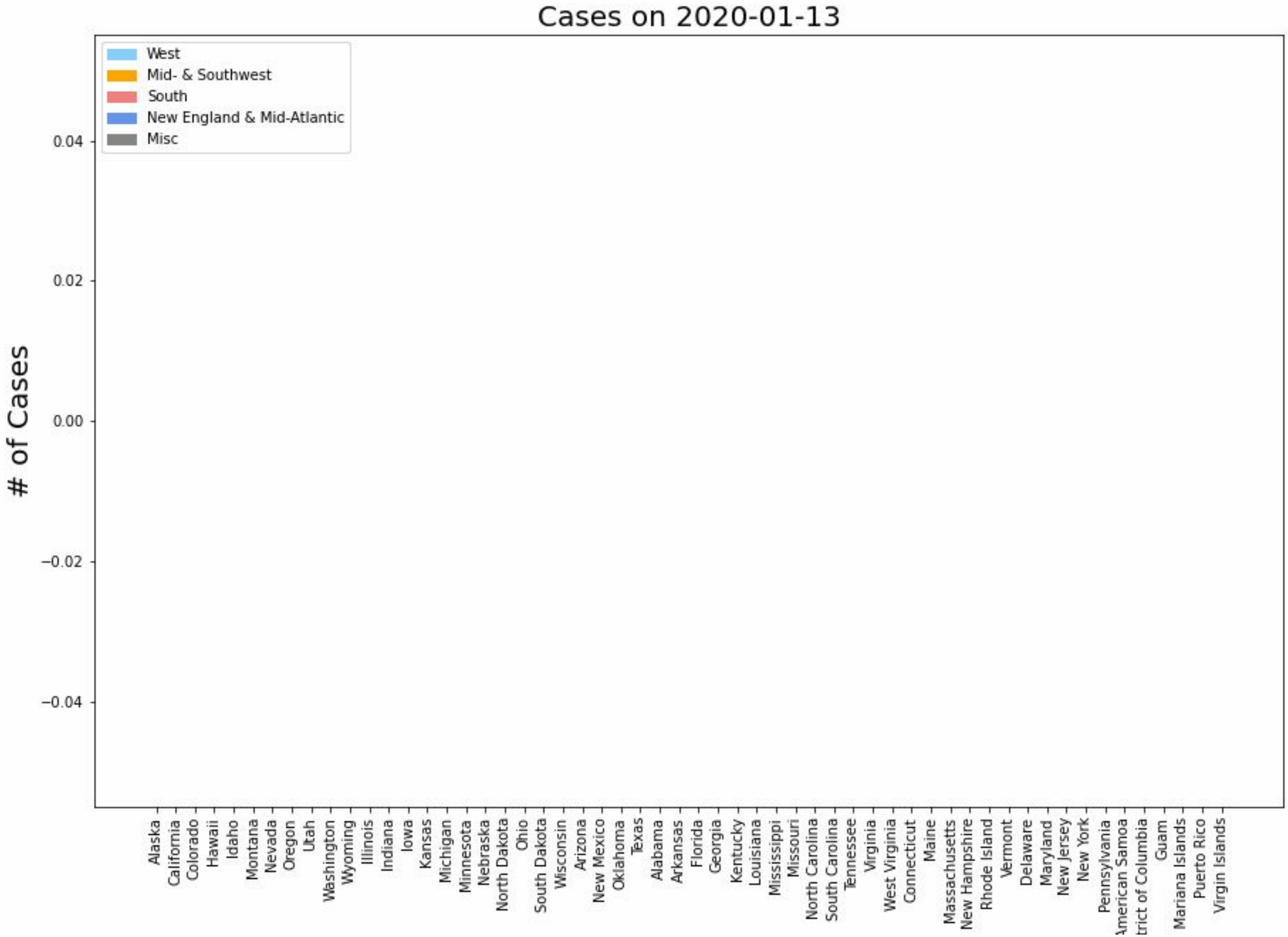
Bibliography

Github Repository: https://github.com/CSJohnLee/COVID19_Project

Datasets:

1. https://github.com/nytimes/covid-19-data
2. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
3. https://covidtracking.com/data
4. https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations
5. https://www.kaggle.com/lucasvictor/us-state-populations-2018#State%20Populations.csv

Updated December Cases Over Time

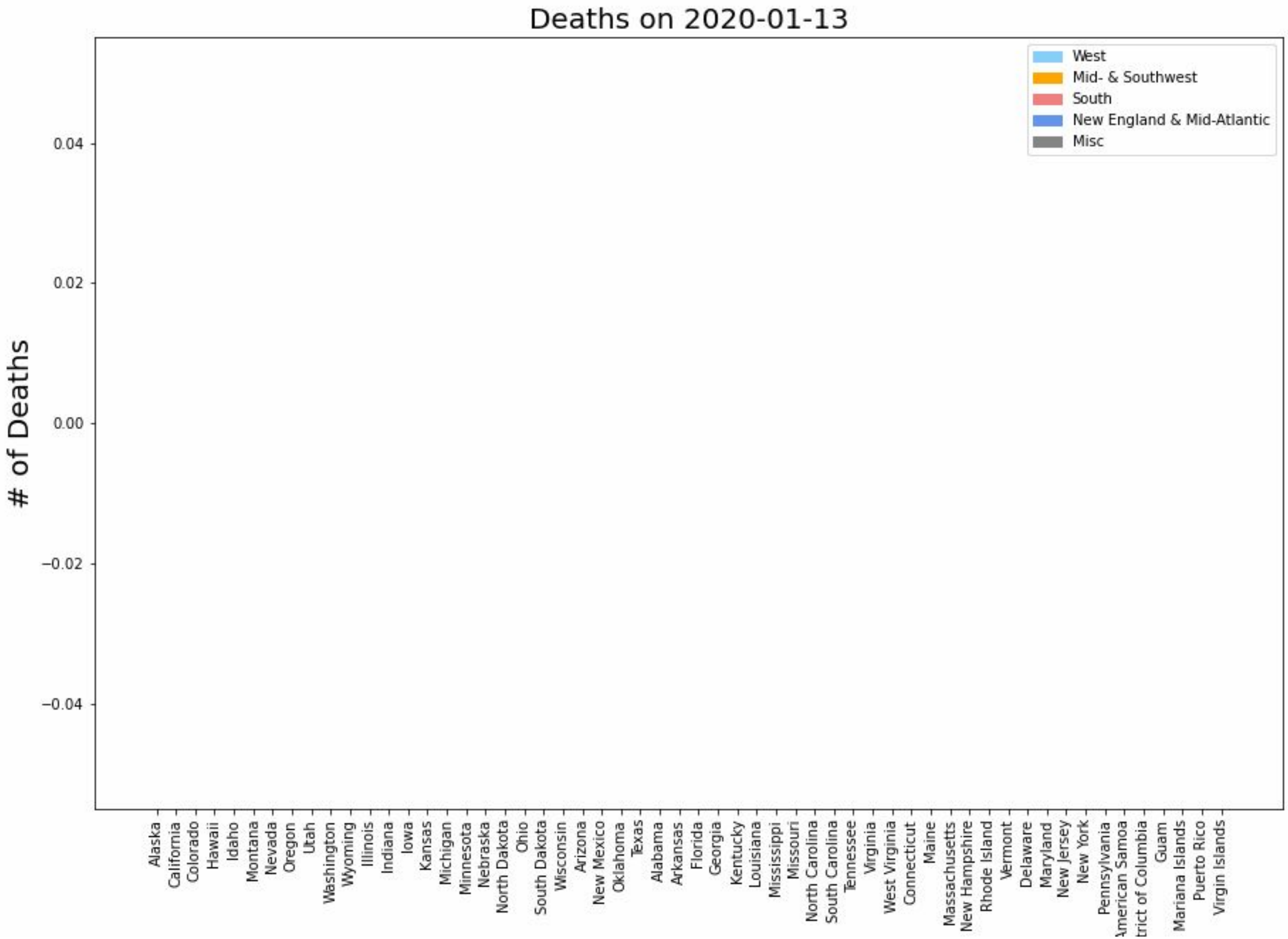


Data Source: https://covidtracking.com/data

Updated December Cases per 100K Over Time

Data Source: https://covidtracking.com/data

Updated December Deaths Over Time



Data Source: https://covidtracking.com/data

Updated December Deaths per 100K Over Time

Data Source: https://covidtracking.com/data