Data Science Project:

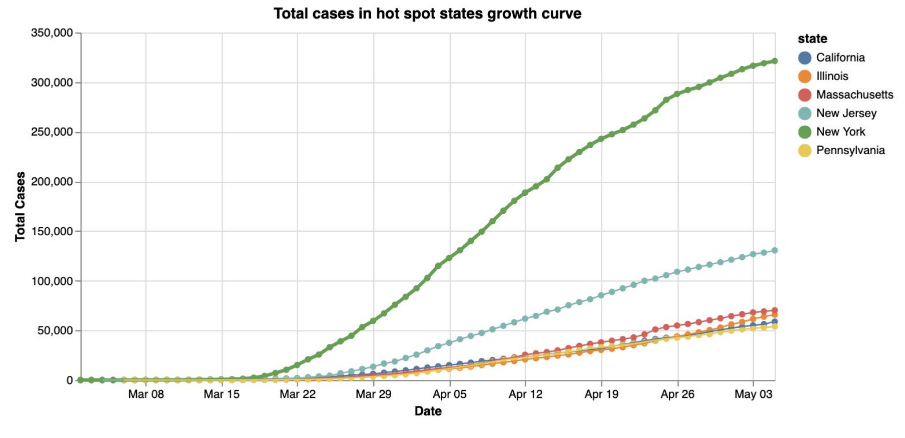
Exploring Coronavirus Data

Anna Godin & Katherine Thai

Overview - Why we chose COVID-19

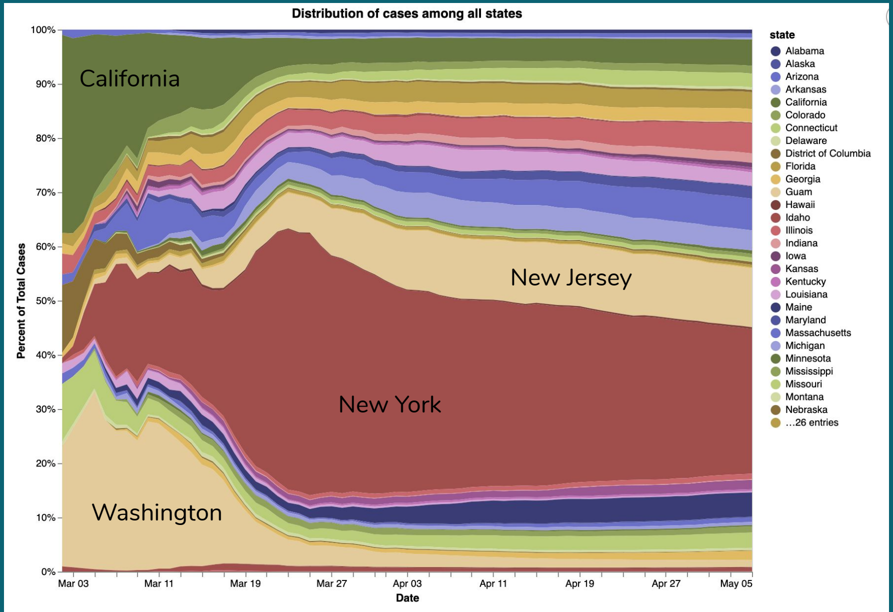
- To gain insight on the developments of the pandemic
- Current event that everyone has been researching
- Find patterns in cases based on state population & density
- Predict number of cases based on previously recorded data

Findings

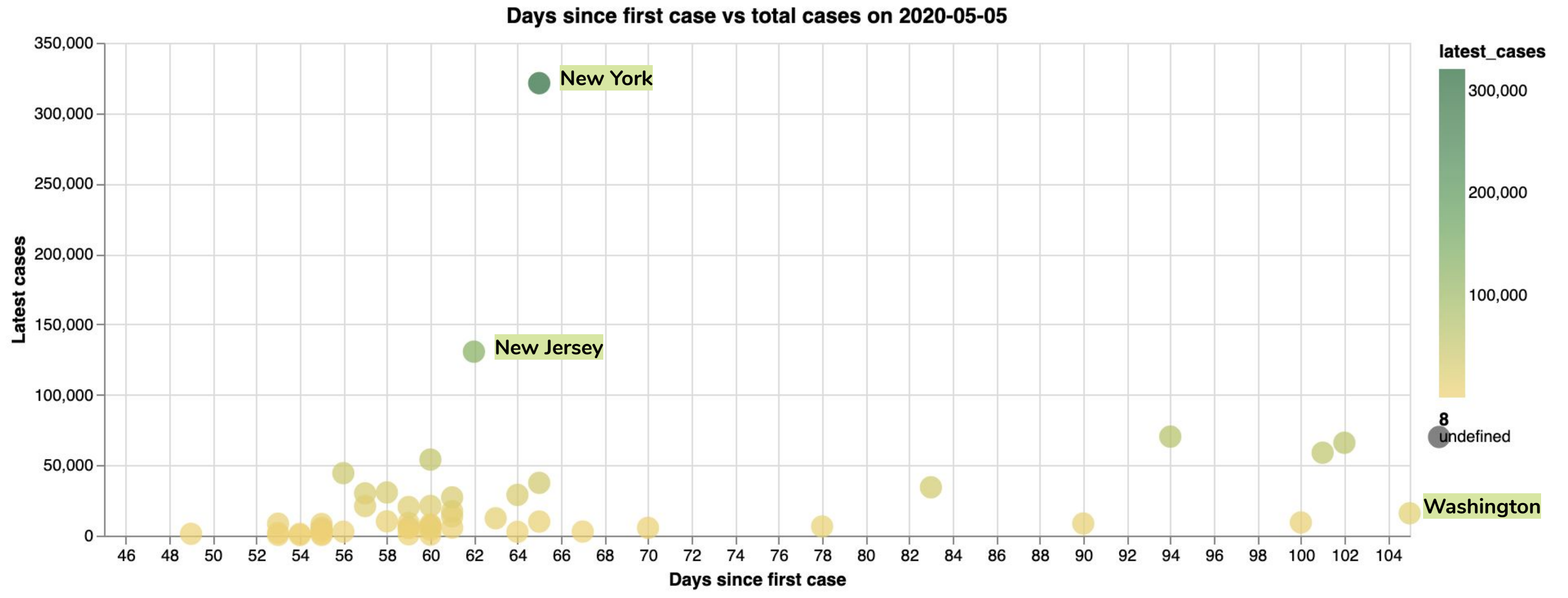


Shows how fast cases in NY & NJ have been growing, and the similar trajectories of PA, MA, CA & IL

Illustrates how fast NY/NJ cases grew to have the majority of cases in the US

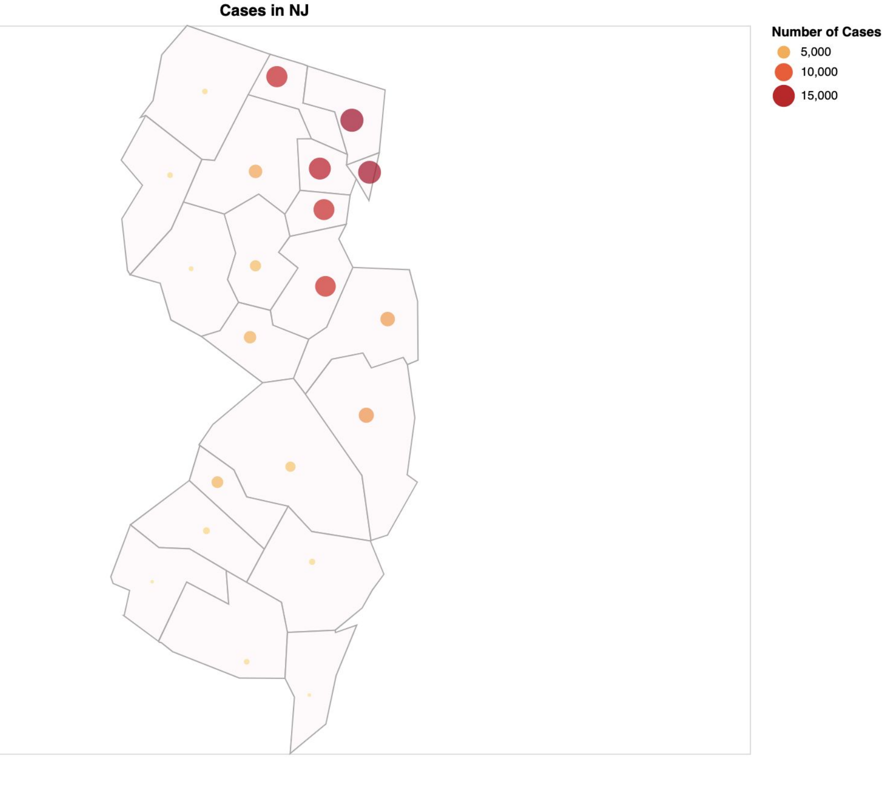


Insightful Graphs



Shows how quickly NY & NJ blew up in terms of number of cases, despite reporting their first cases much later than many states, who still have a relatively low number of cases compared to NY & NJ

Insightful Graphs

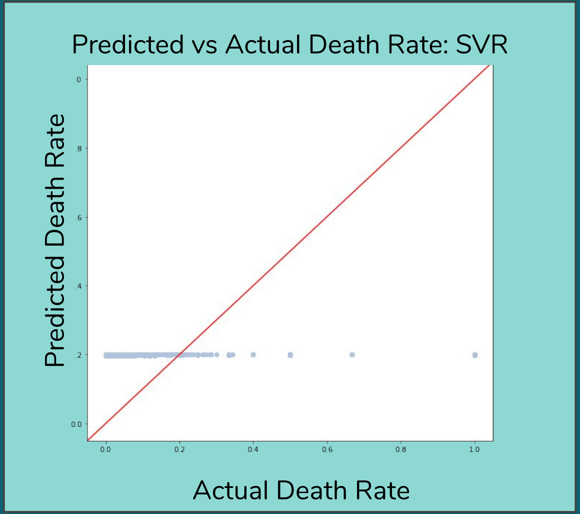


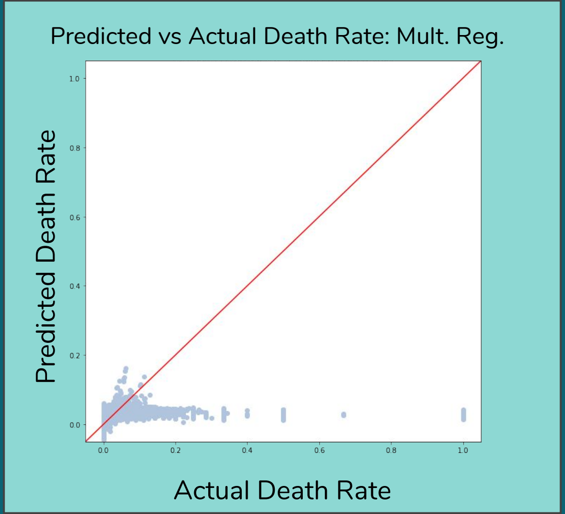
Northeastern counties have more cases than southern counties

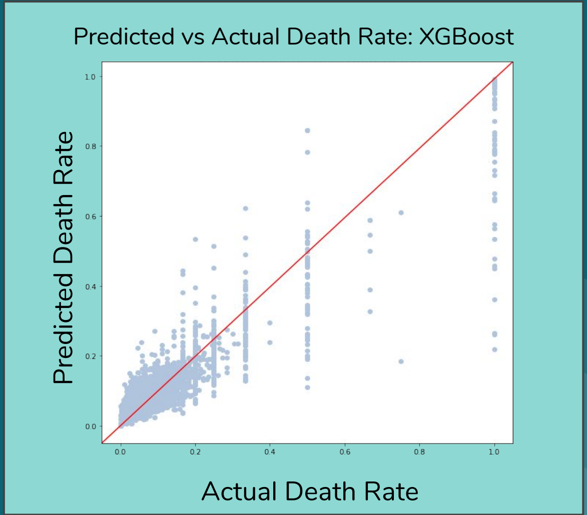
Could be due to population density & proximity to NYC

Machine Learning Summary

We tried to predict the case death rate based on number of cases and other factors we thought might be impacting death count.


Predicted vs Actual Death Rate: SVR


Predicted vs Actual Death Rate: Mult. Reg.


Predicted vs Actual Death Rate: XGBoost

Unfortunately, none of our models were particularly accurate, but XGBoost performed the best.

Conclusions

Data is so important, and ours was not fully representative of the spread of coronavirus:

- Different states had different approaches to testing.
- Some counties had a backlog of cases that they would report weeks after patients had recovered.
- Some counties had 100% death rates because the only case of reported coronavirus had died.