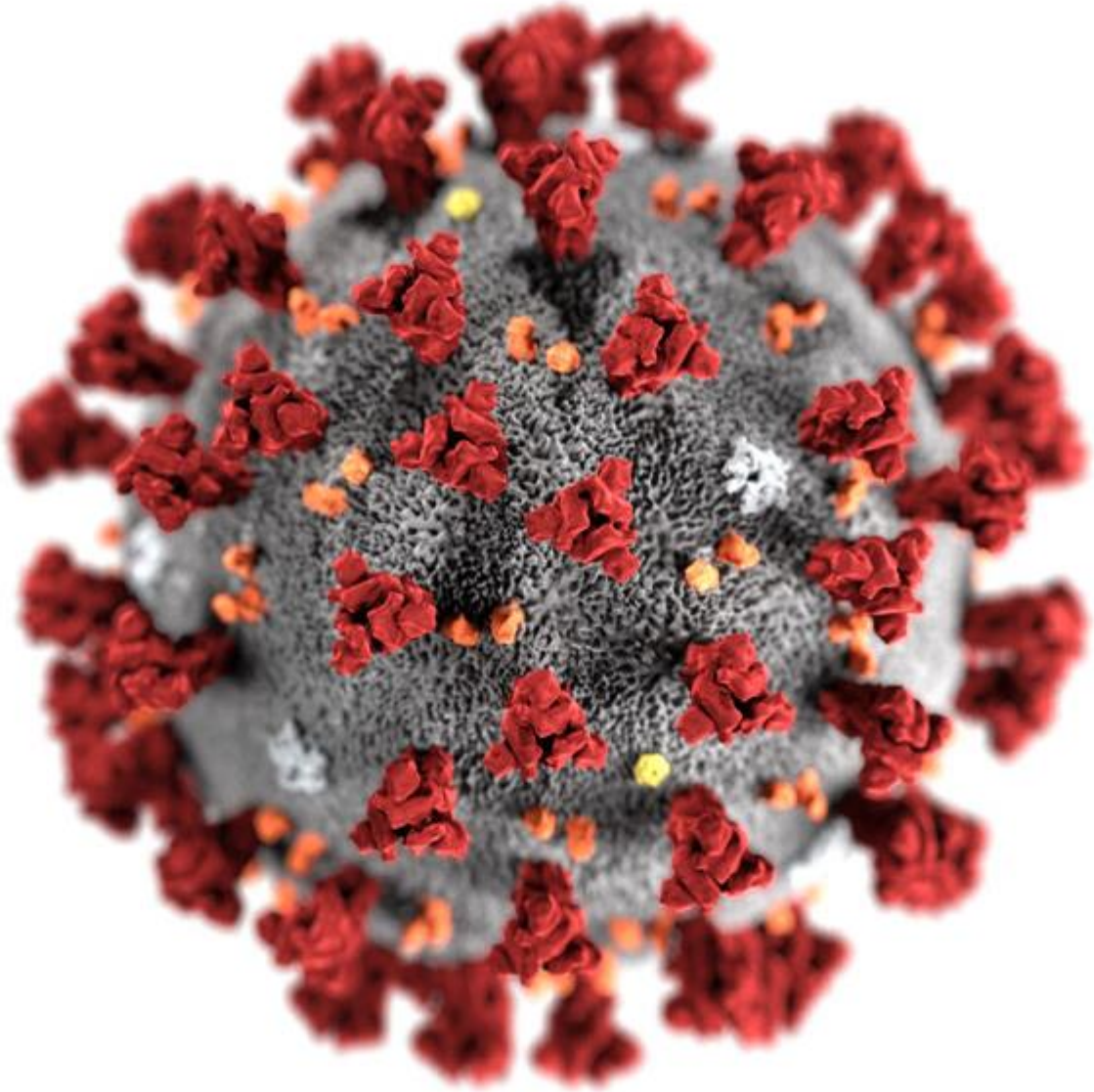
Predicting COVID-19 Using Demographic Data

Caroline Clark, Feras Atwal, James Lee

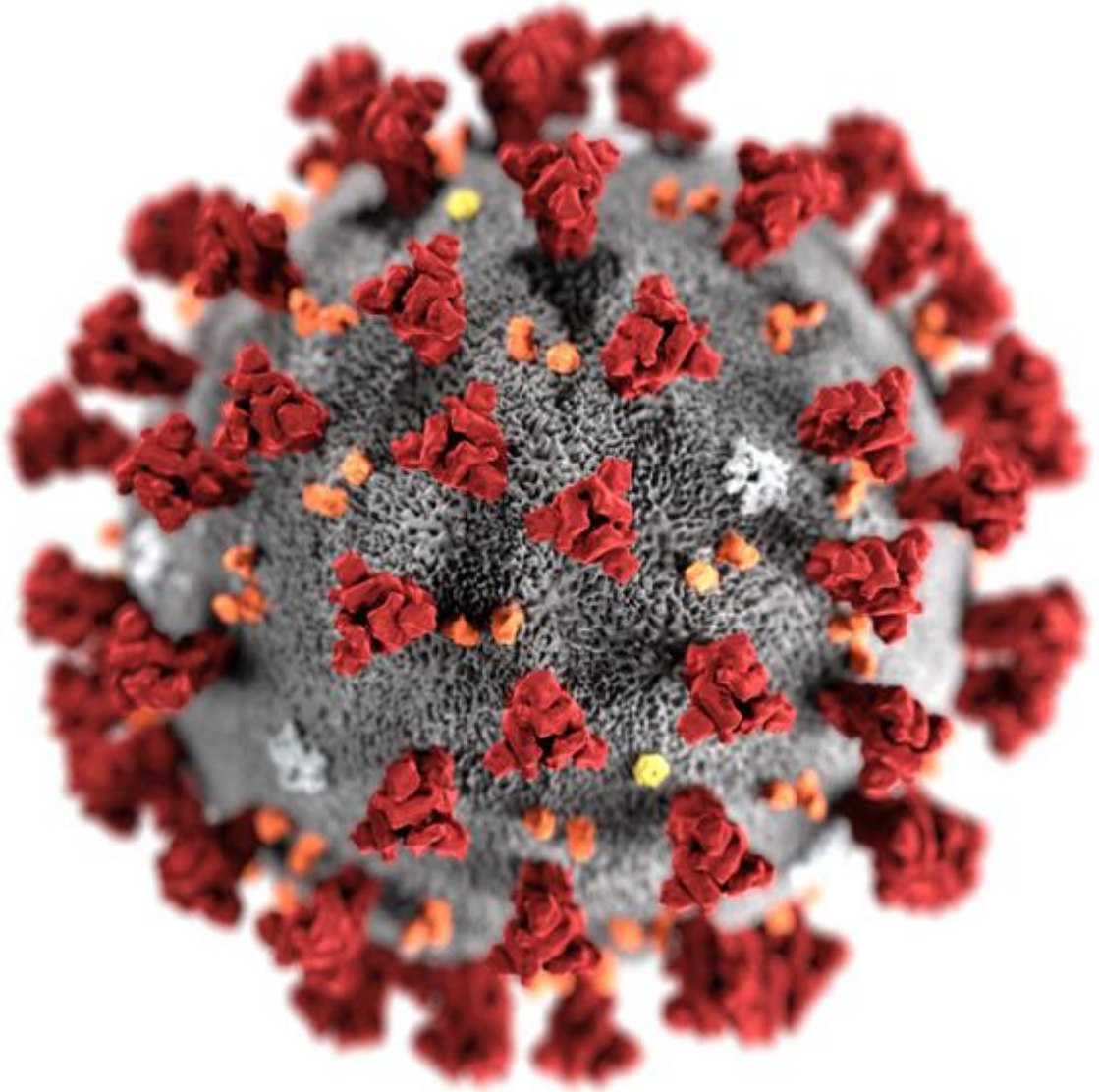
October 30th, 2020



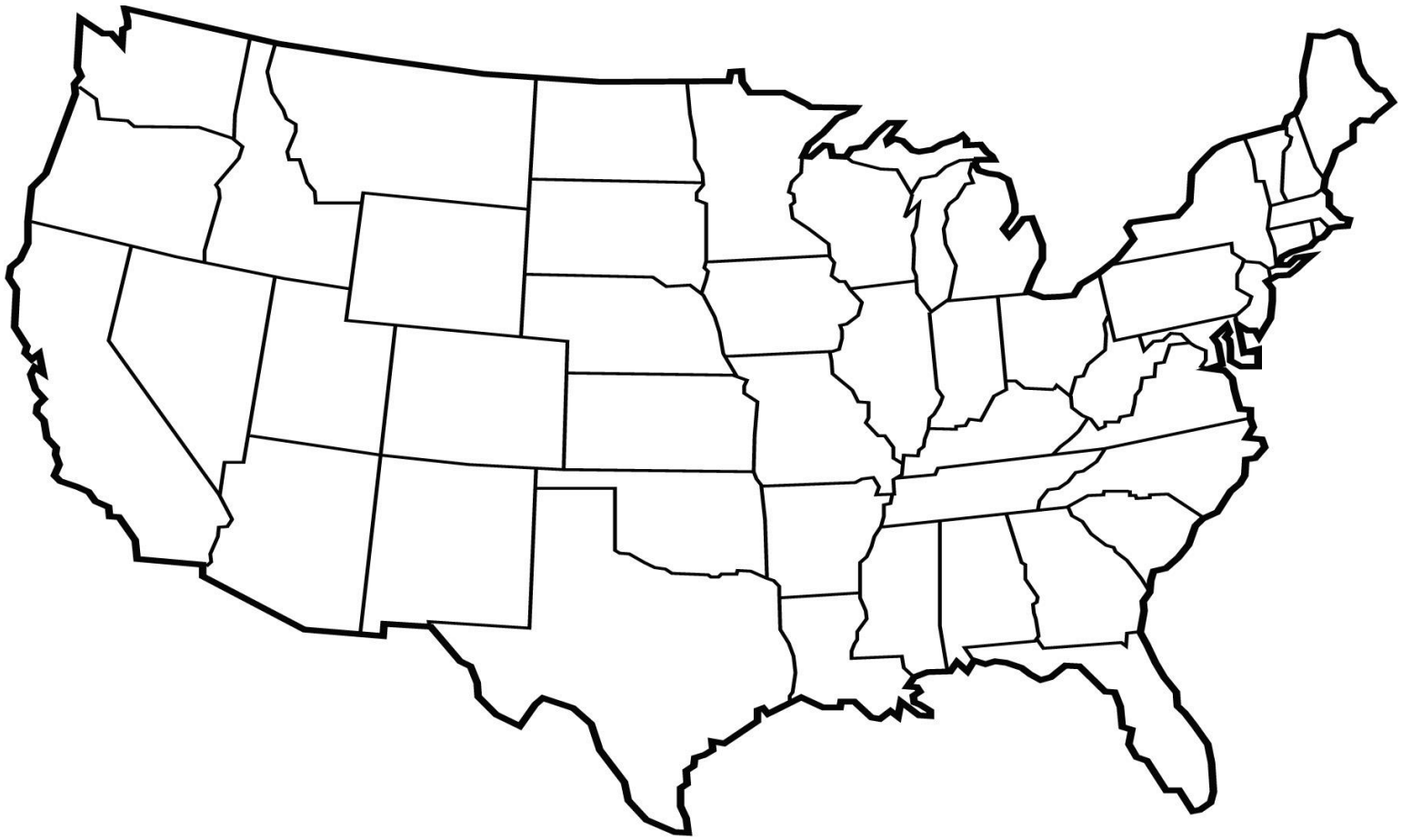
Predicting COVID-19 Using Demographic Data

Caroline Clark, Feras Atwal, James Lee

October 30th, 2020

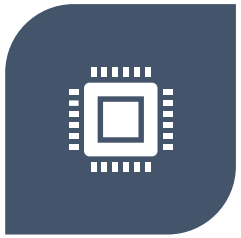


Can we predict COVID-19 severity using demographic data?



Project Pipeline

DATA COLLECTION



DATA PRE-PROCESSING

DATA VISUALIZATION

MODELING

MAKING THE DATA INTERACTIVE

Data Collection

County-level

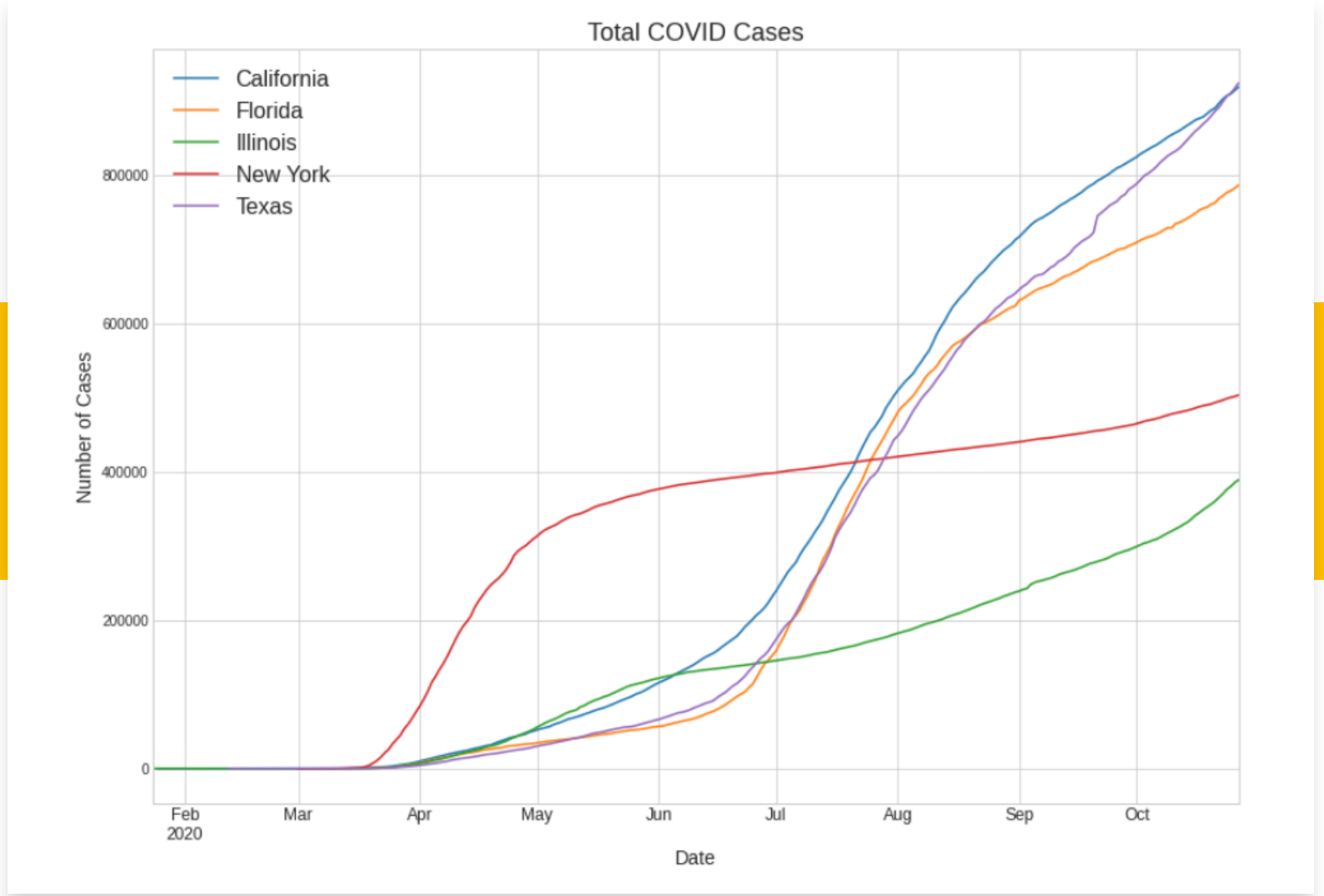
Area
- Population Density

Demographics
- Age
- Gender
- Race

Economic Indicators
- Income Per Capita
- Household Income
- Median Worker Income

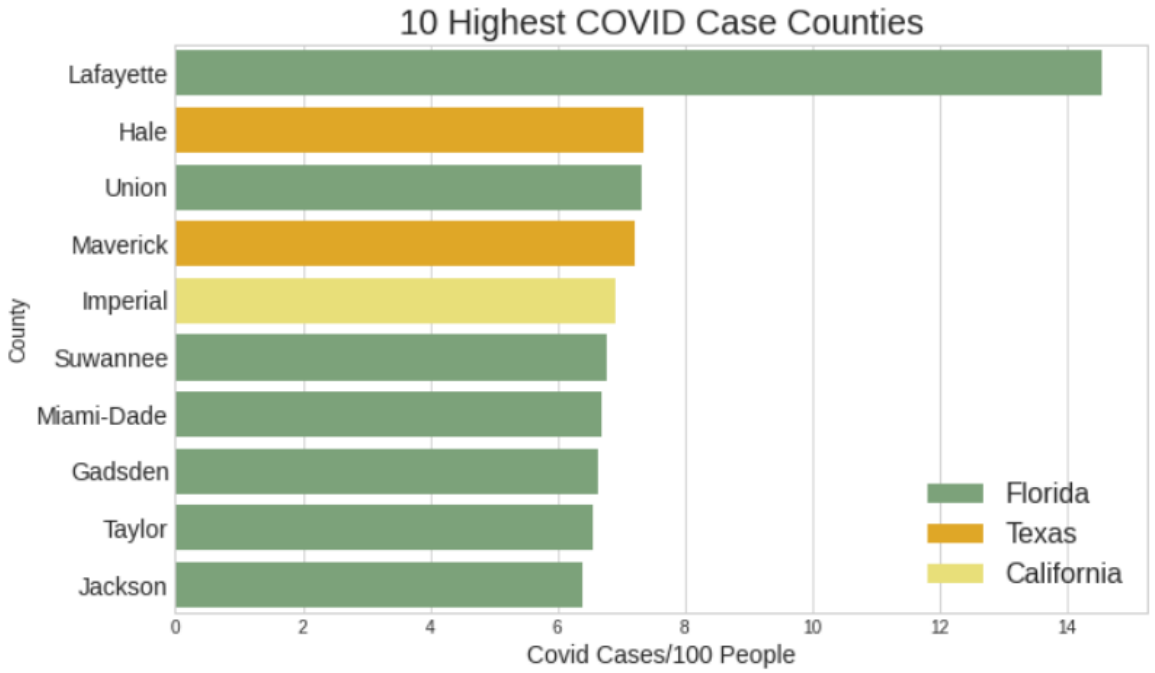
Health Indicators
- Obesity Rates

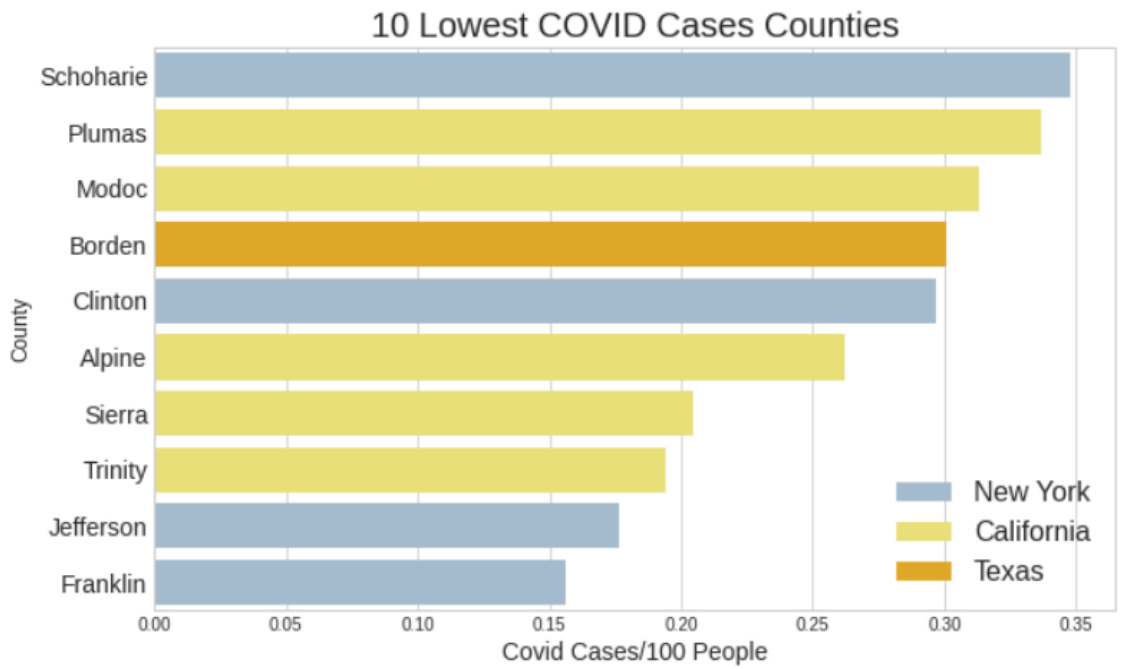
COVID-19
- Tests
- Cases
- Fatalities



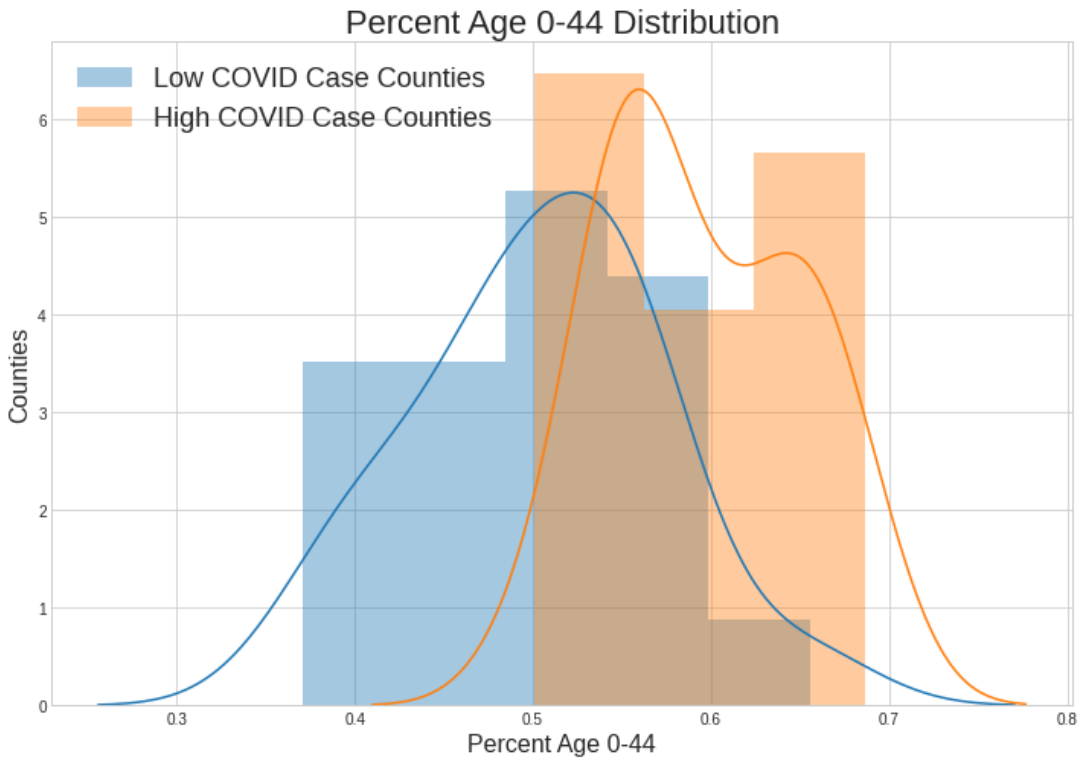
Five States with the Most COVID-19 Data

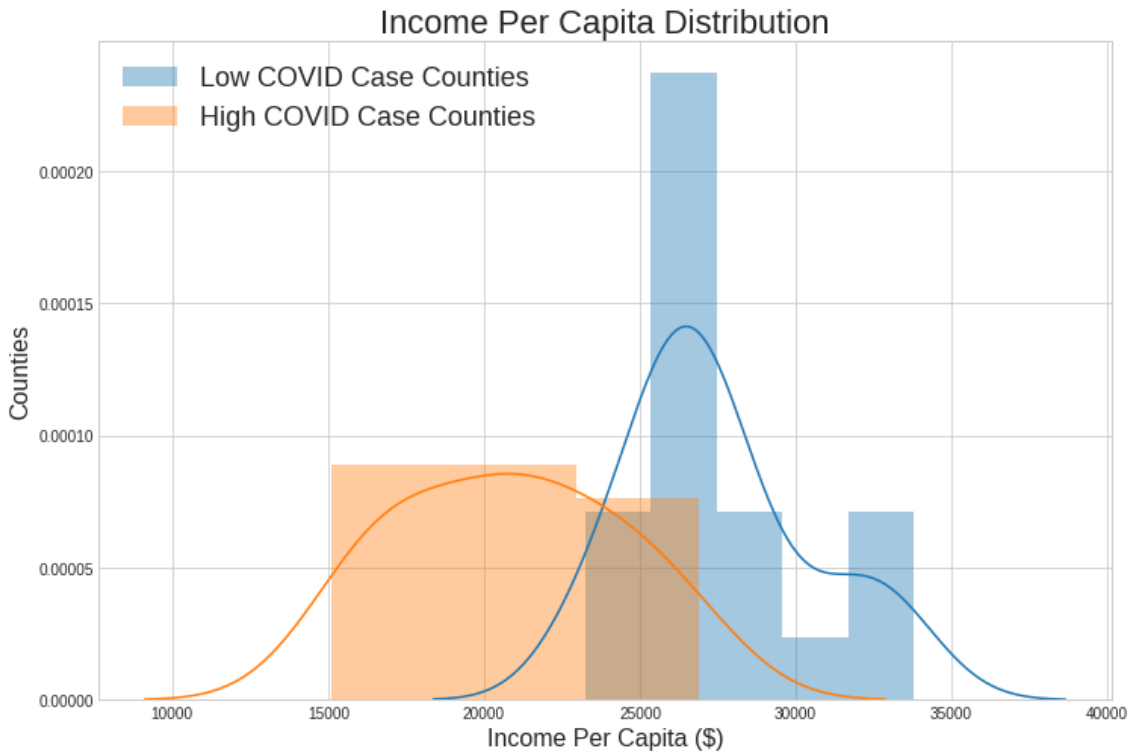
COVID-19 Statistics Vary Widely Among Counties



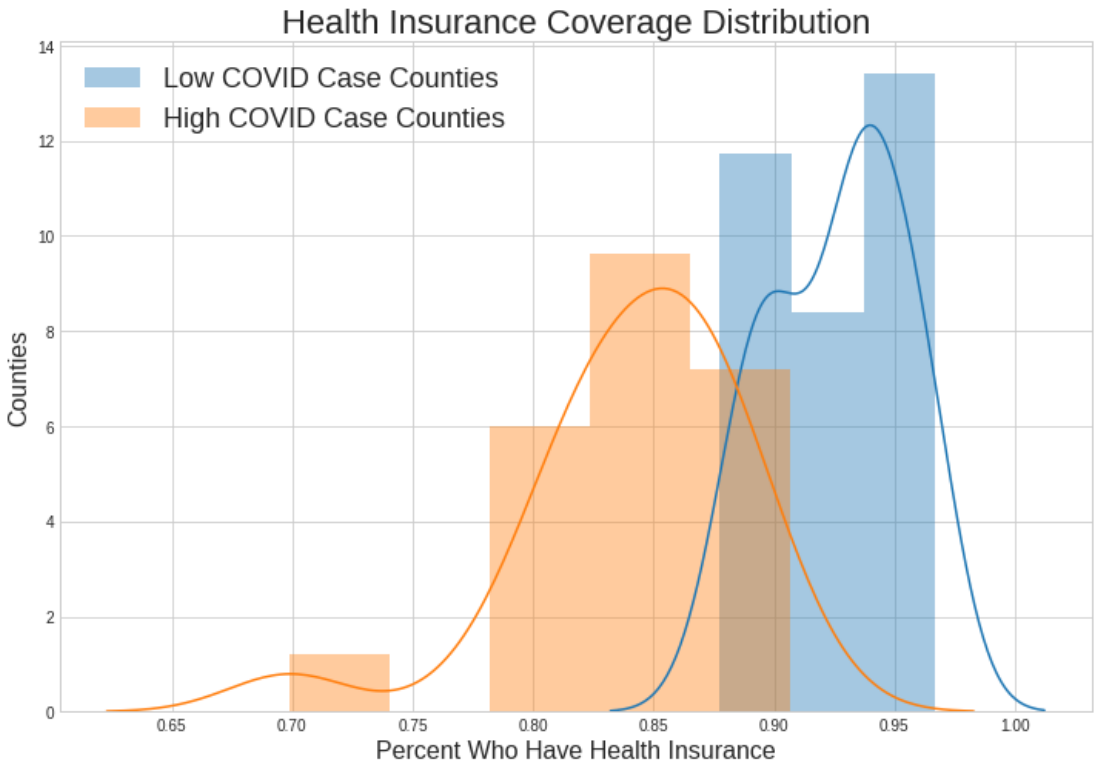


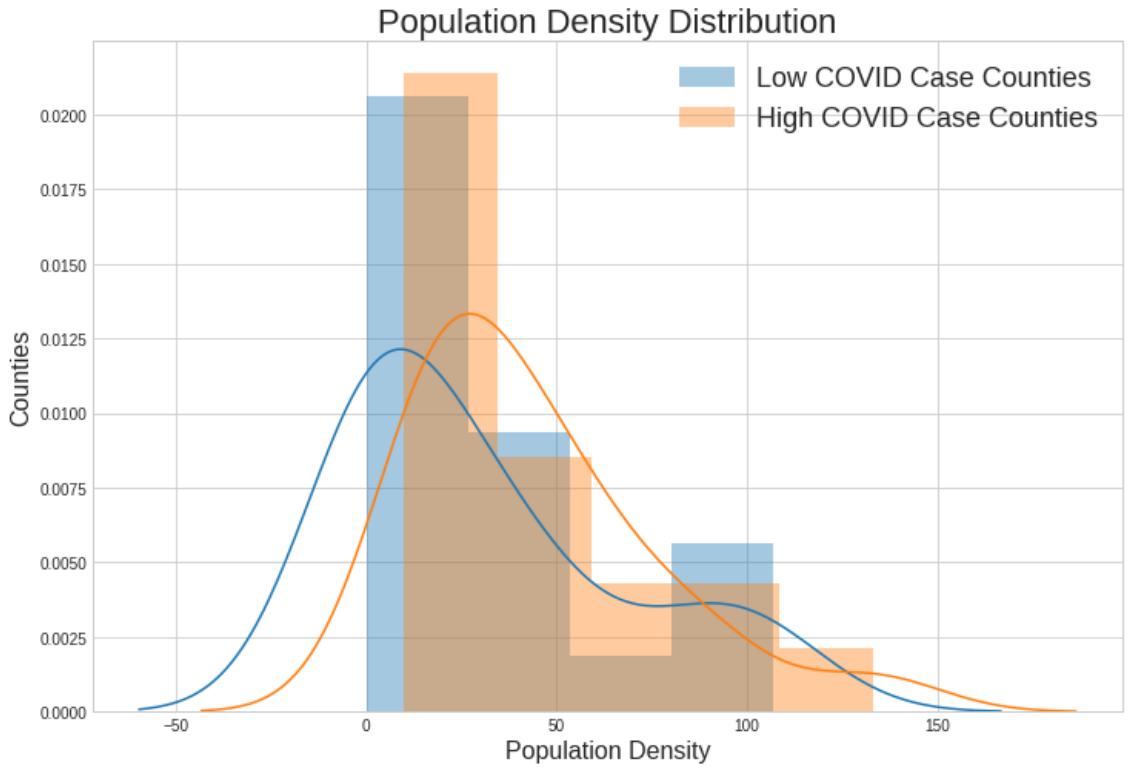
High COVID Case Counties Likely to be Younger, Have Lower Income Per Capita





Low COVID Case Counties Likely to have Insurance Coverage, Lower Pop. Density

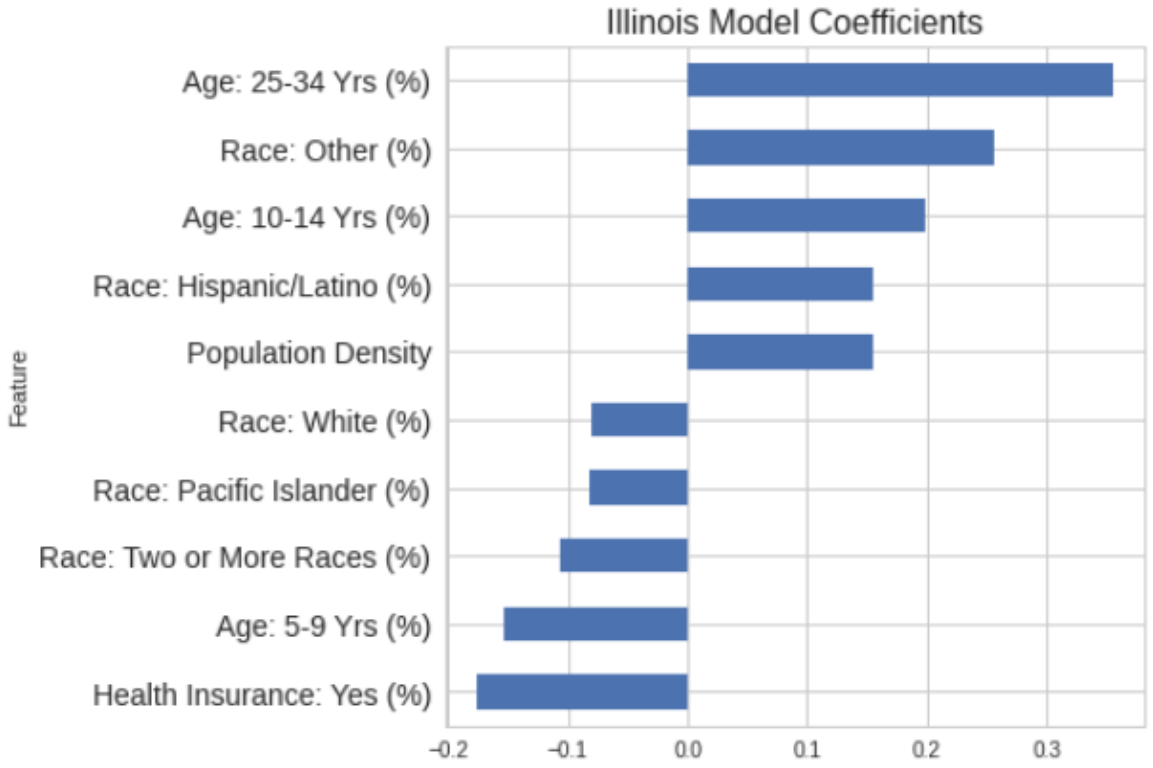


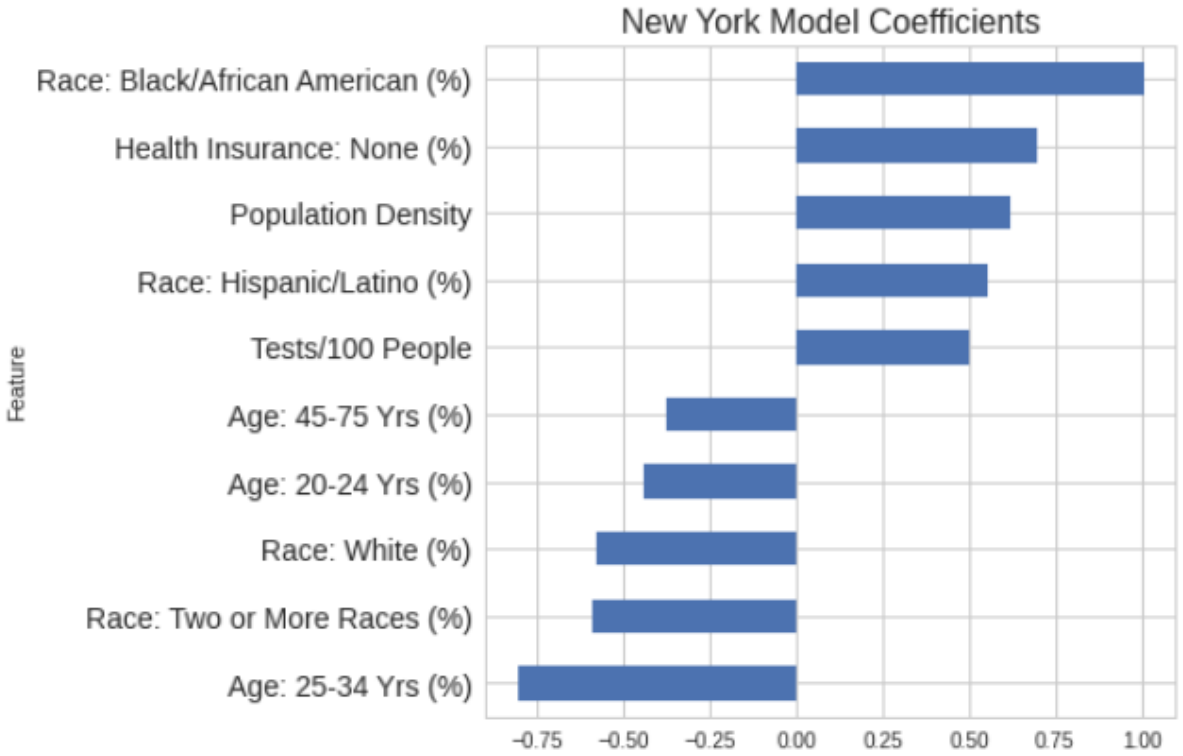


Modeling Successes and Challenges

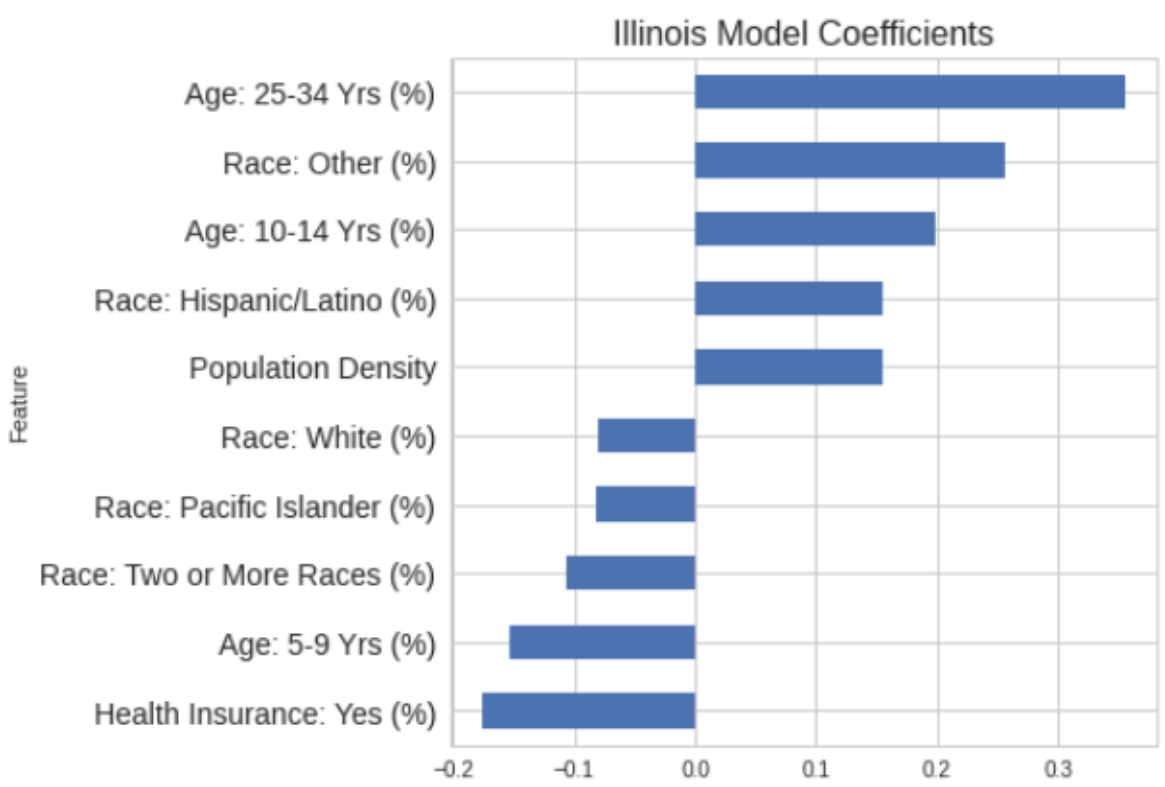
Region	Best Regression R2 Score	Best Classification Accuracy Score	Classification Baseline
All Five States	47%	63%	42%
California	75%	93%	66%
Florida	76%	71%	71%
Illinois	32%	73%	54%
New York	81%	94%	81%
Texas	49%	59%	40%

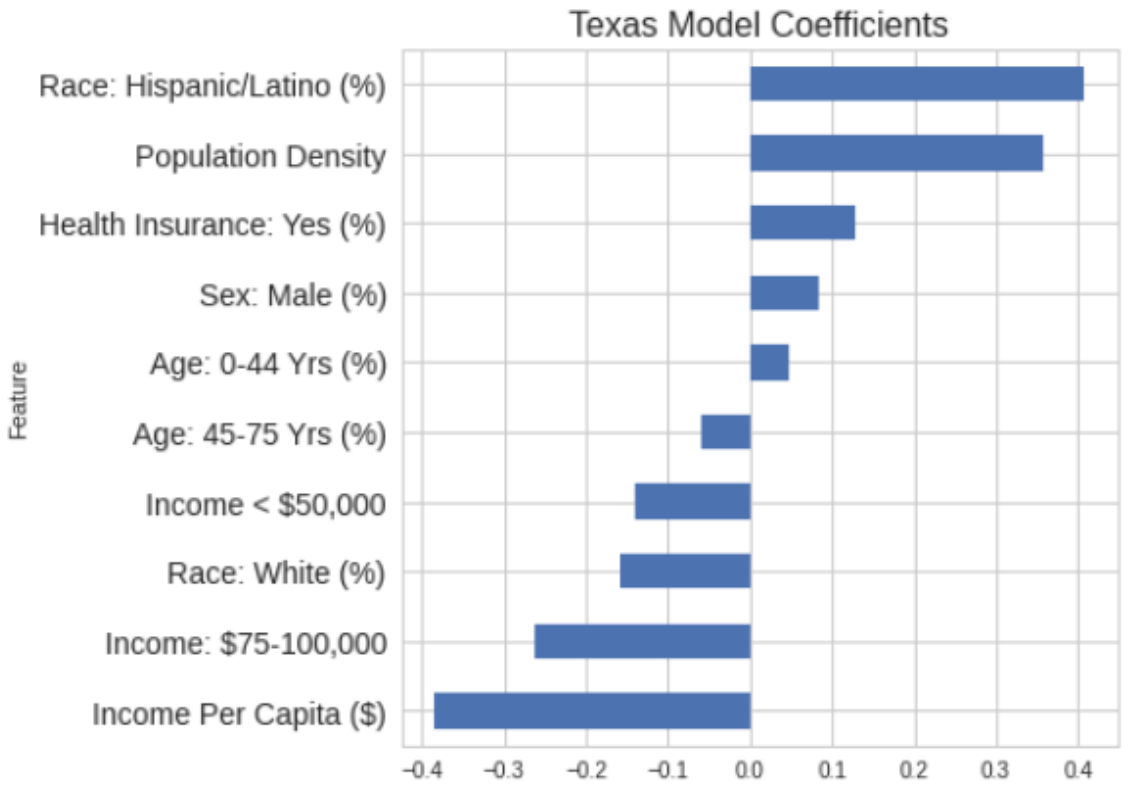
Predictors Varied in State-Level Models



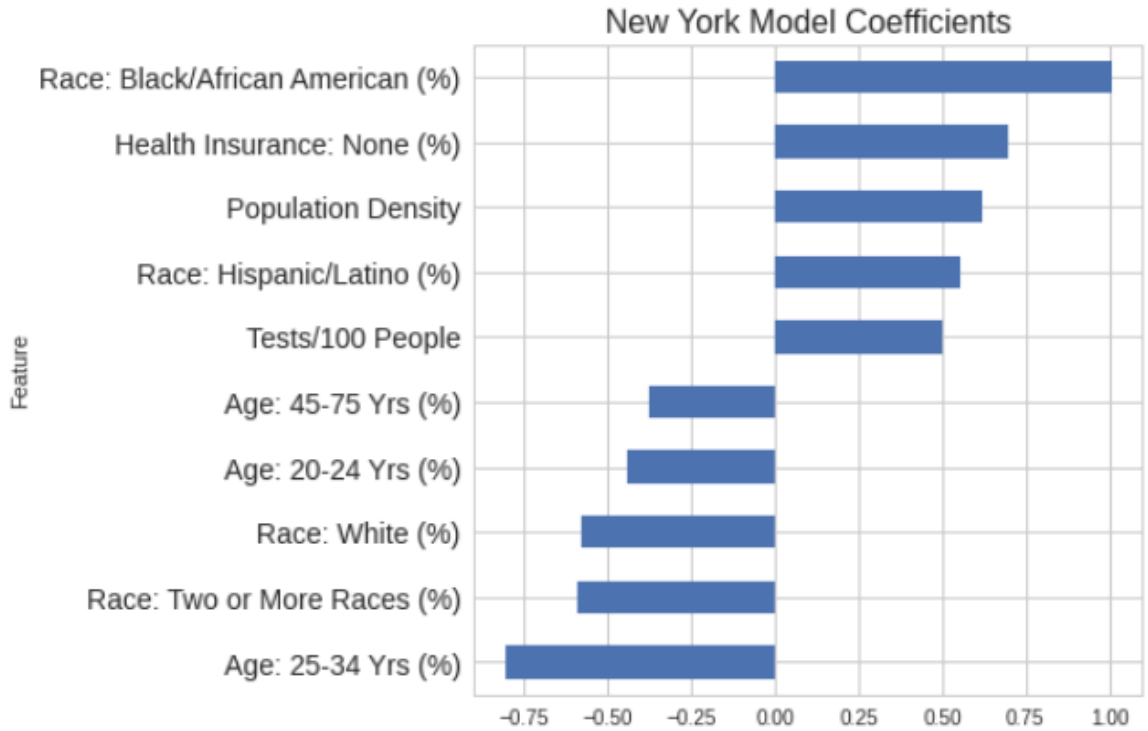


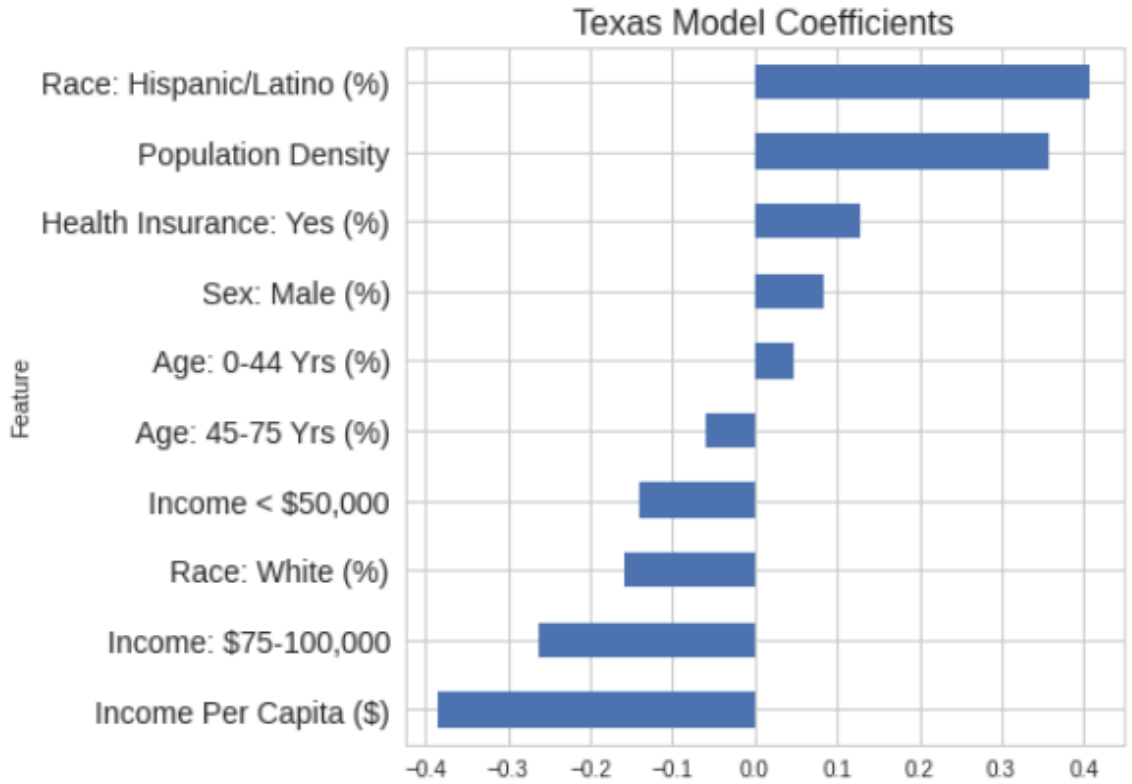
Predictors Varied in State-Level Models





Predictors Varied in State-Level Models



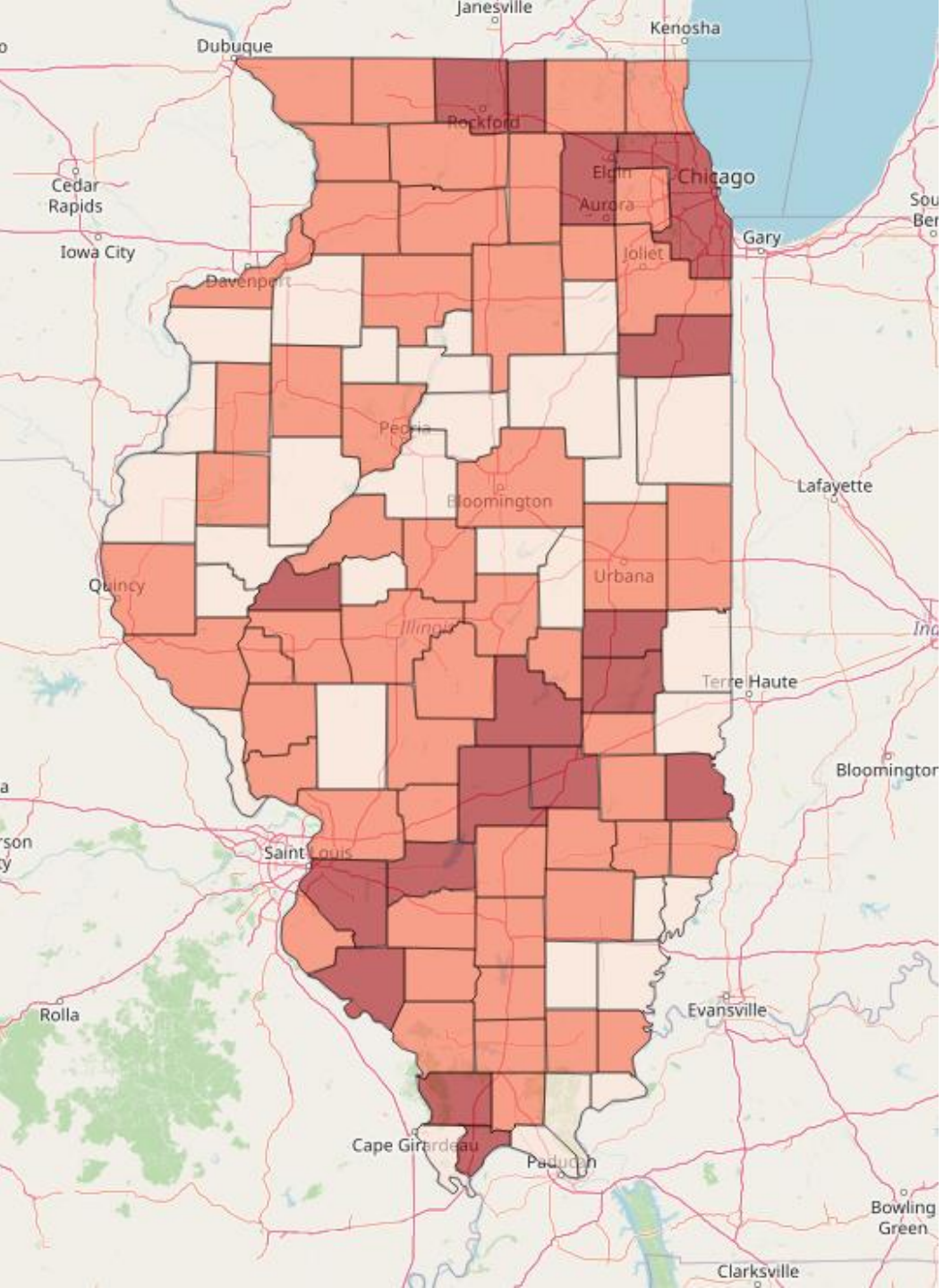


Conclusions and Key Challenges

Ongoing event

Widely varying data

More features

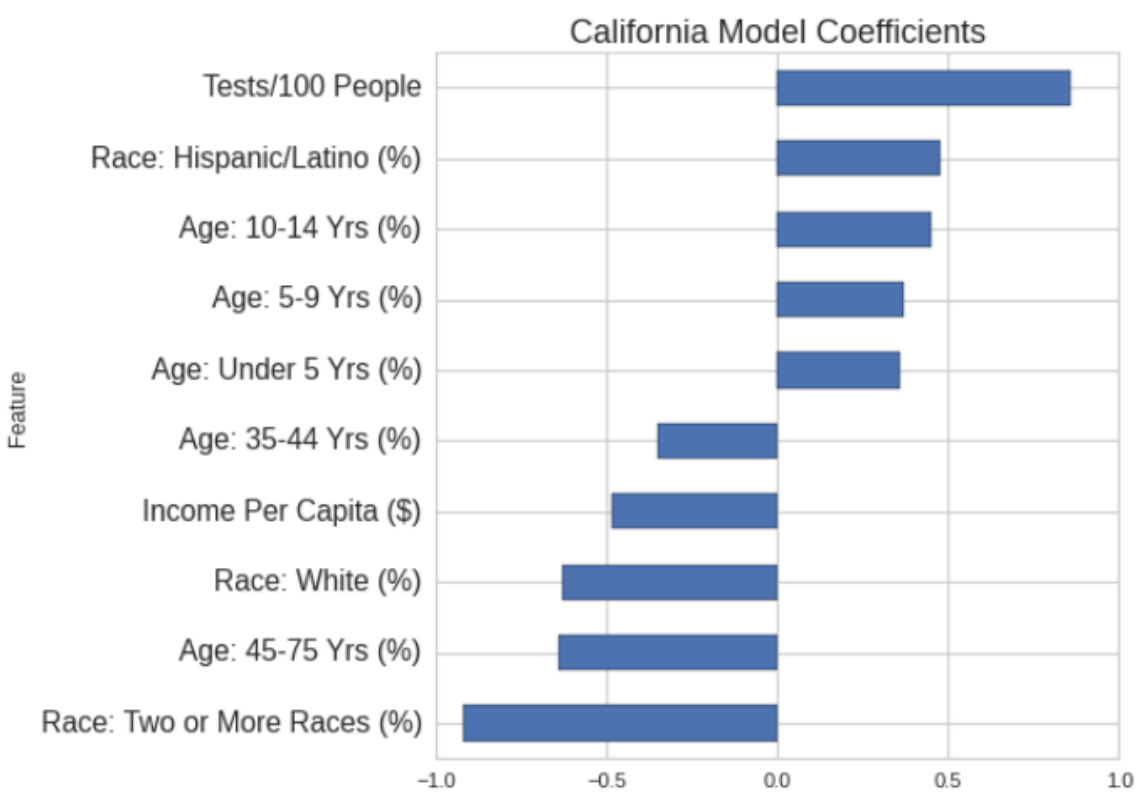


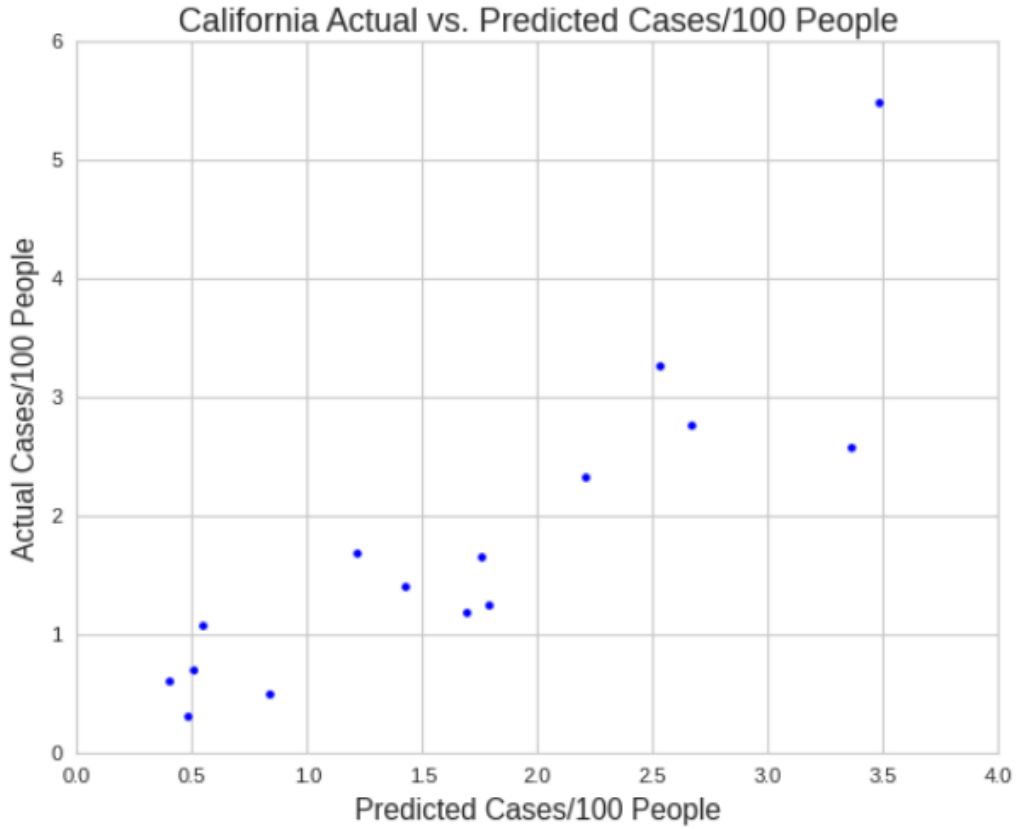
Demo: Interacting with Demographic Data and Classification Model

Thank you

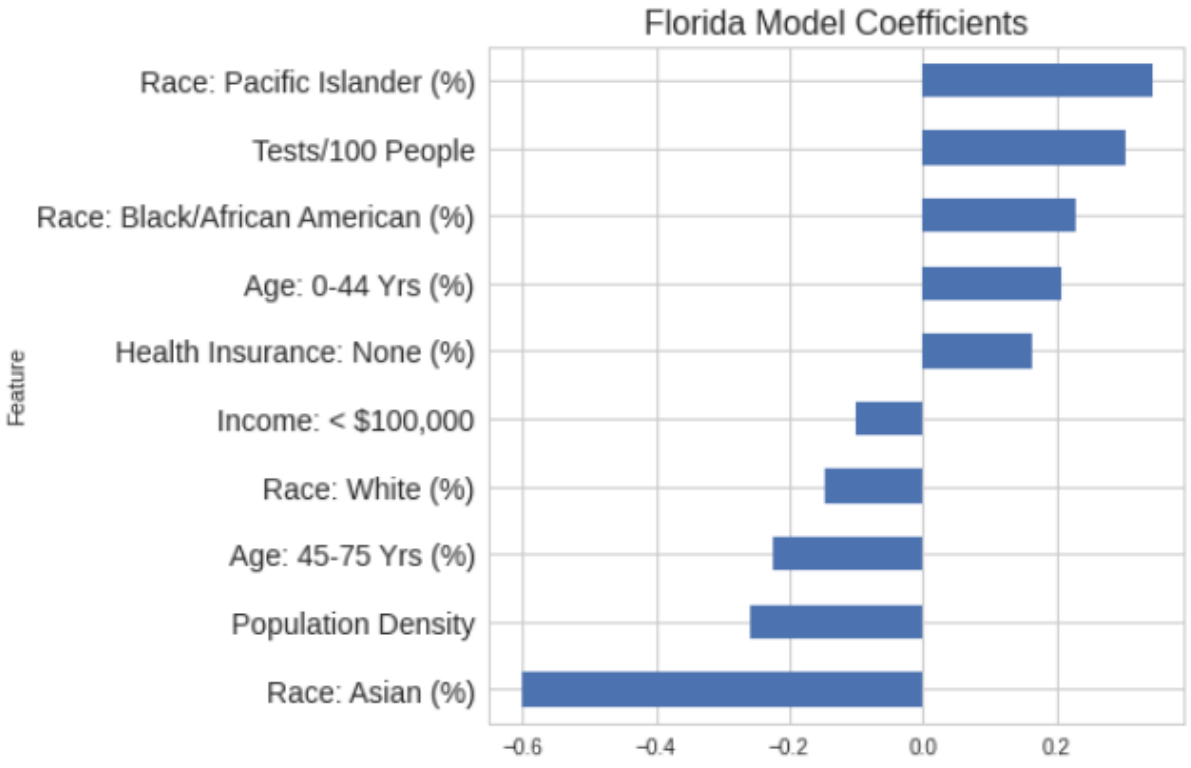
Appendix

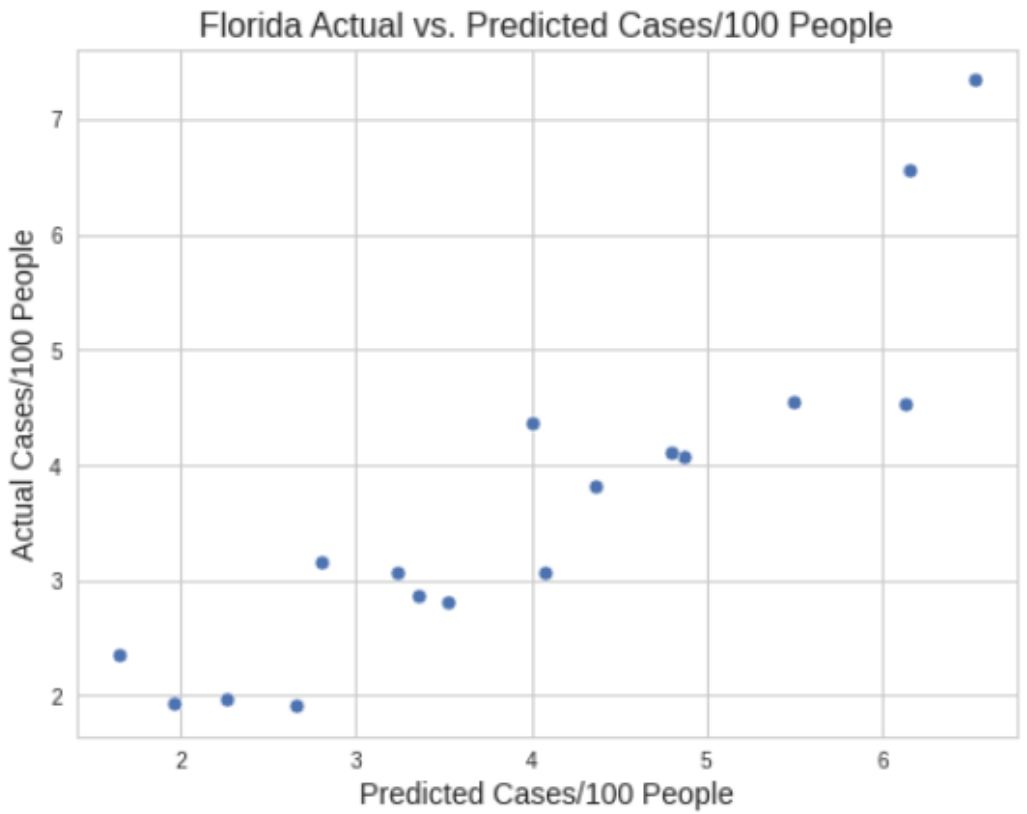
In California, Testing and Race Emerged as Strongest Predictors



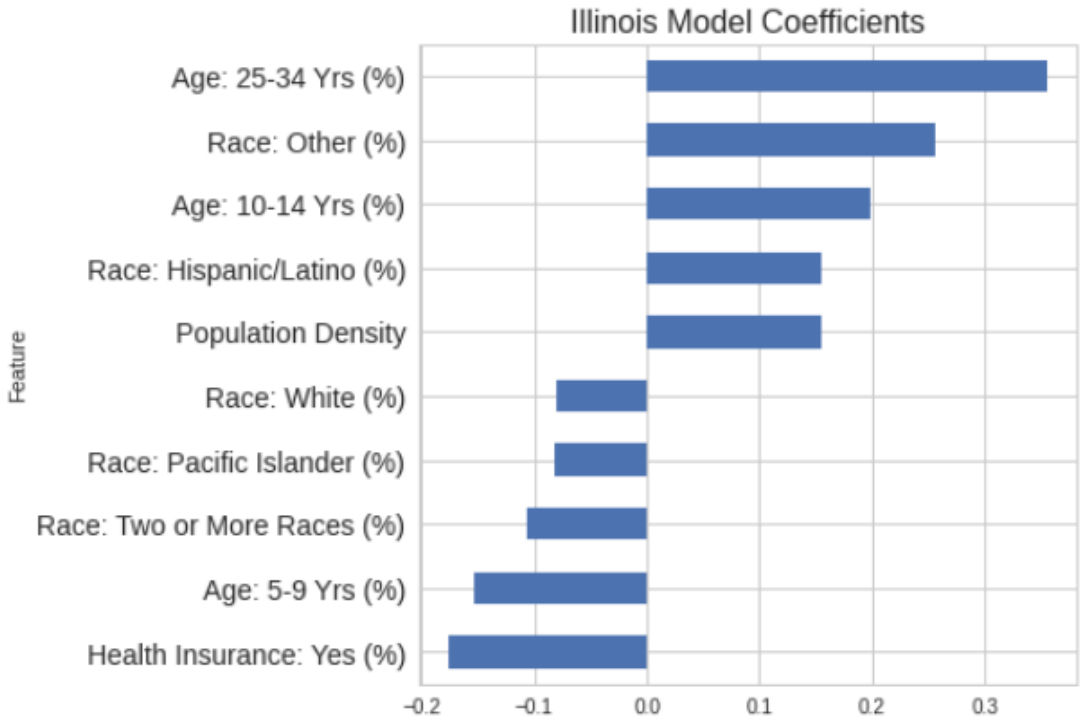


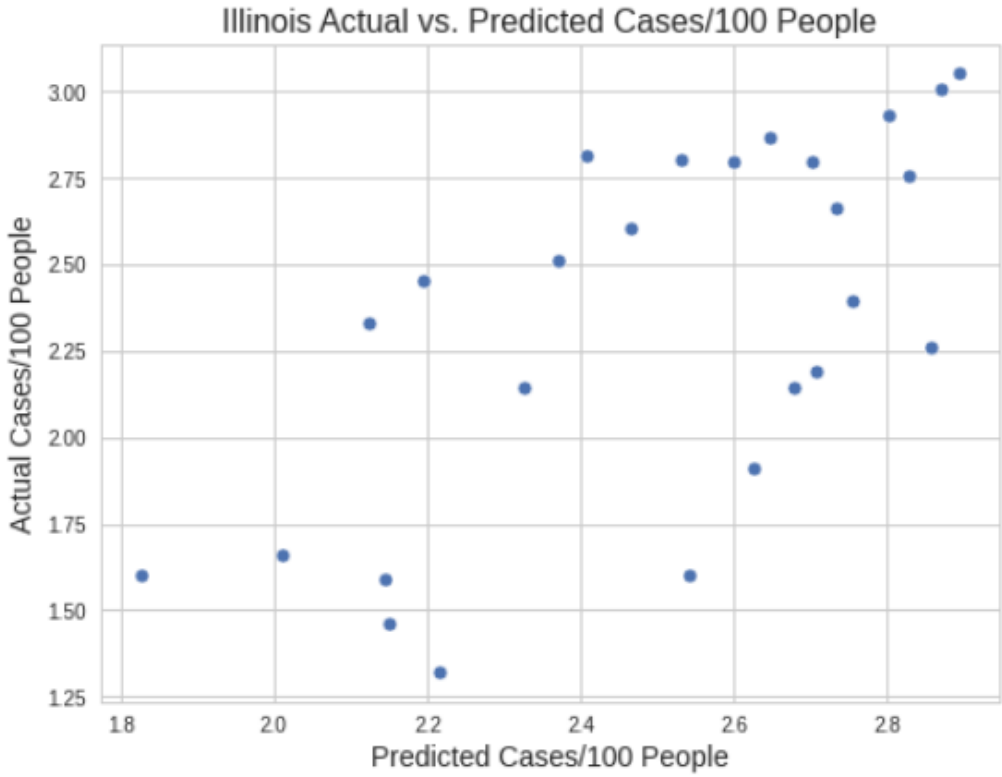
In Florida, Race Emerged as Strongest Predictors



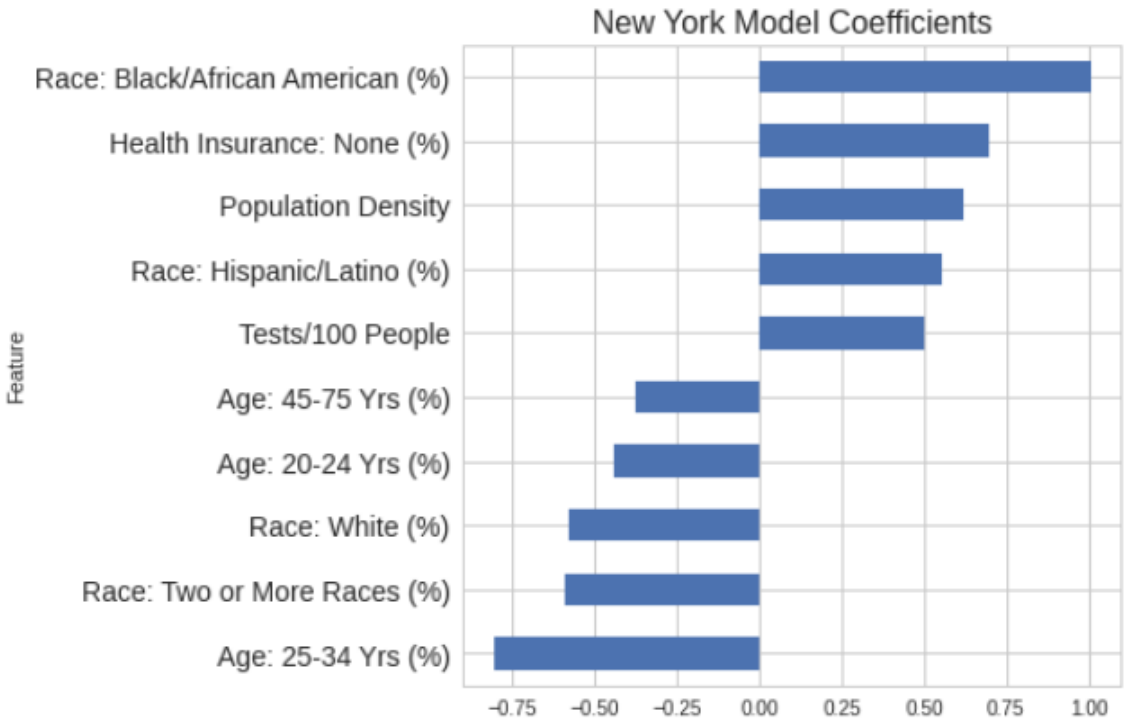


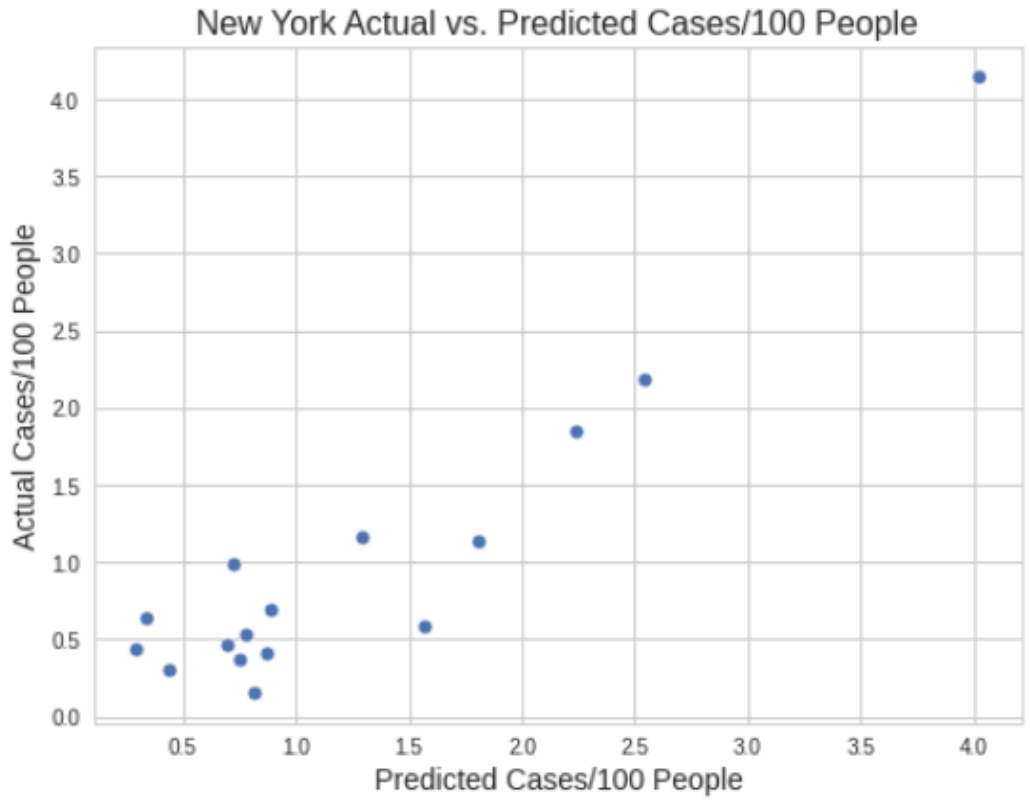
In Illinois, Age and Being Insured Emerged as Strongest Predictors



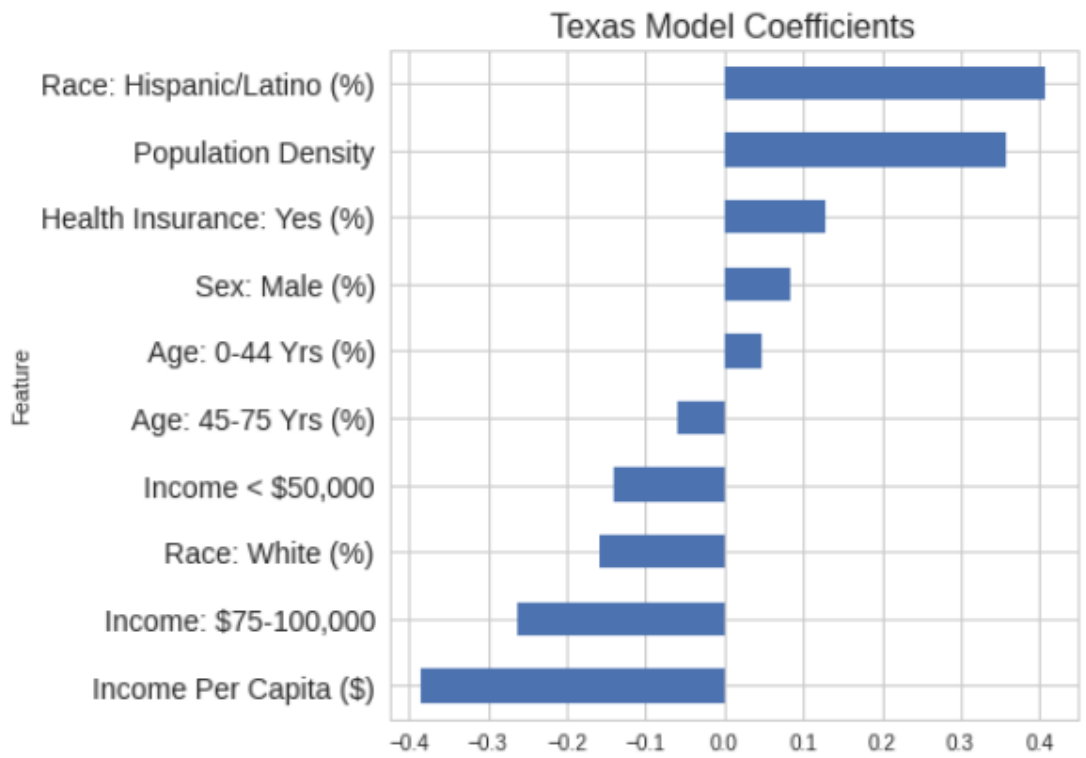


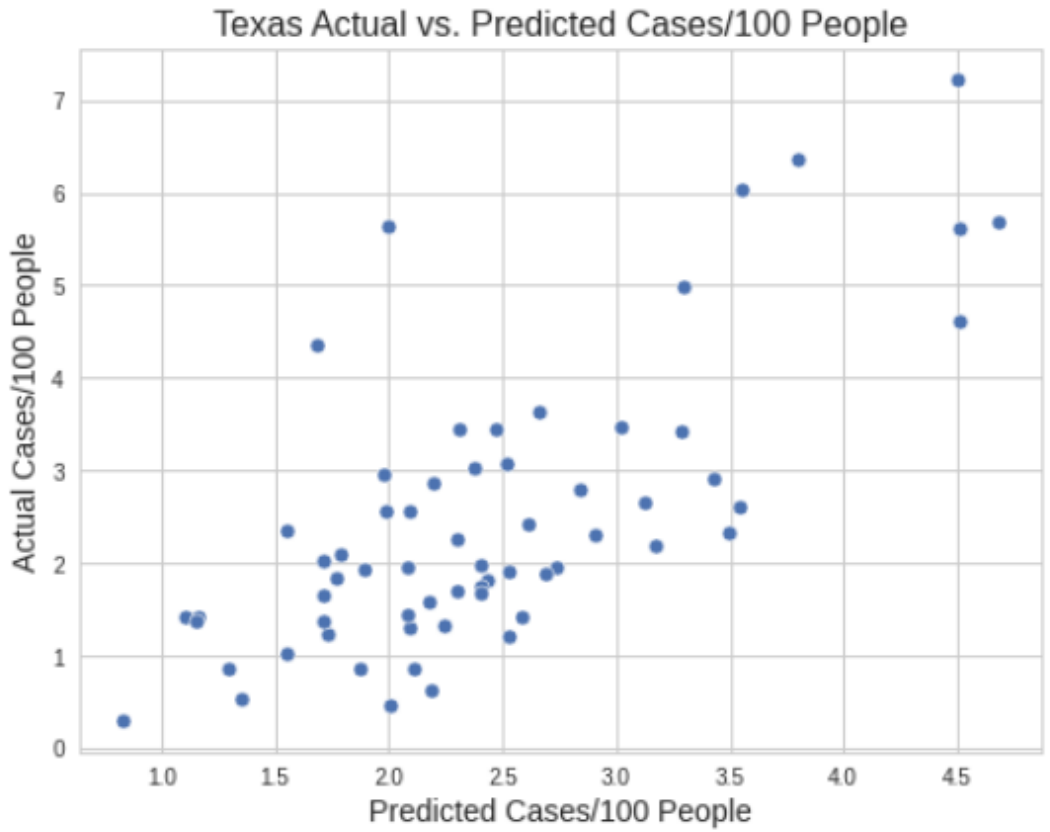
In New York, Race and Age Emerged as Strongest Predictors



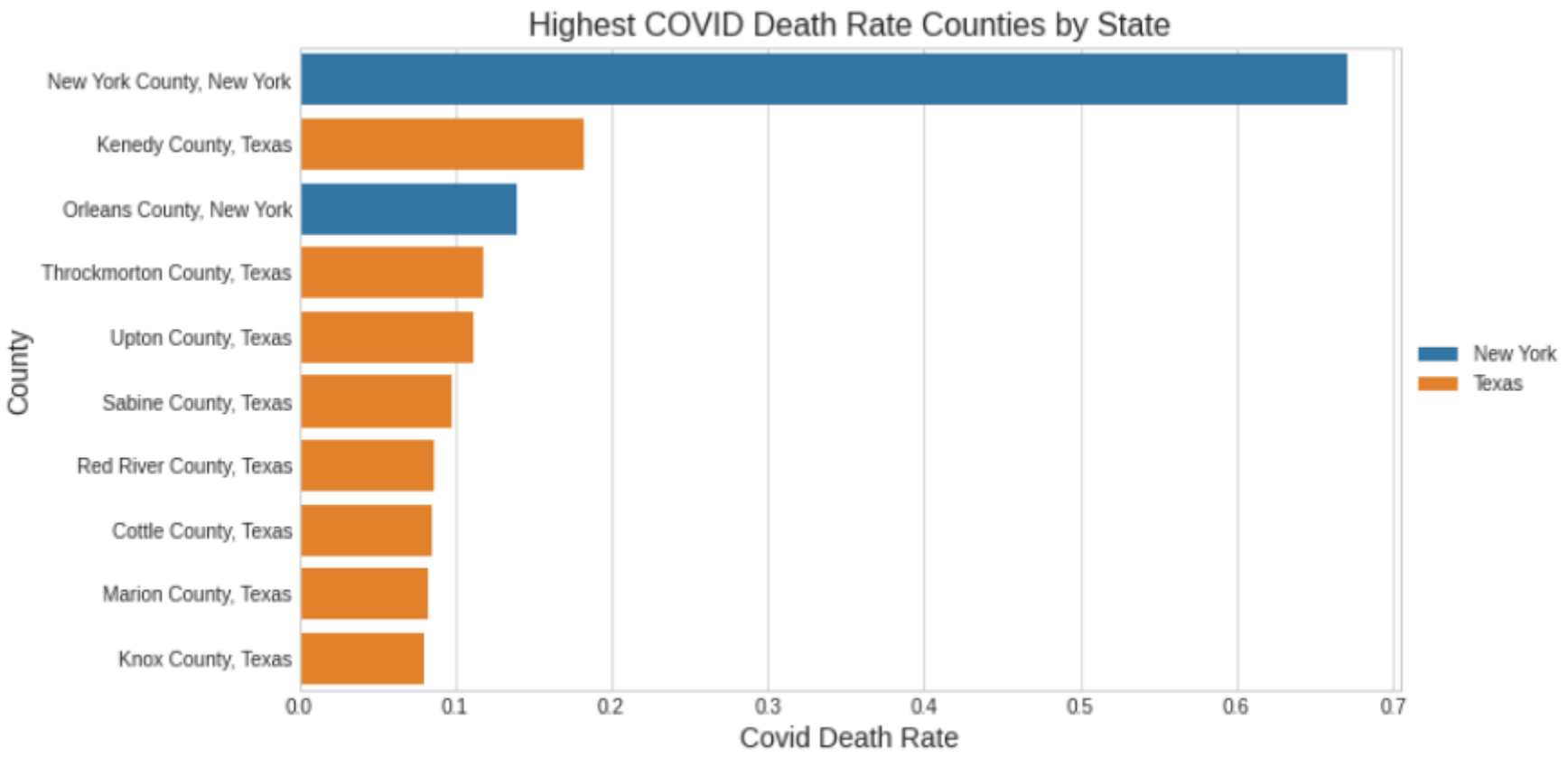


In Texas, Race and Income Emerged as Strongest Predictors

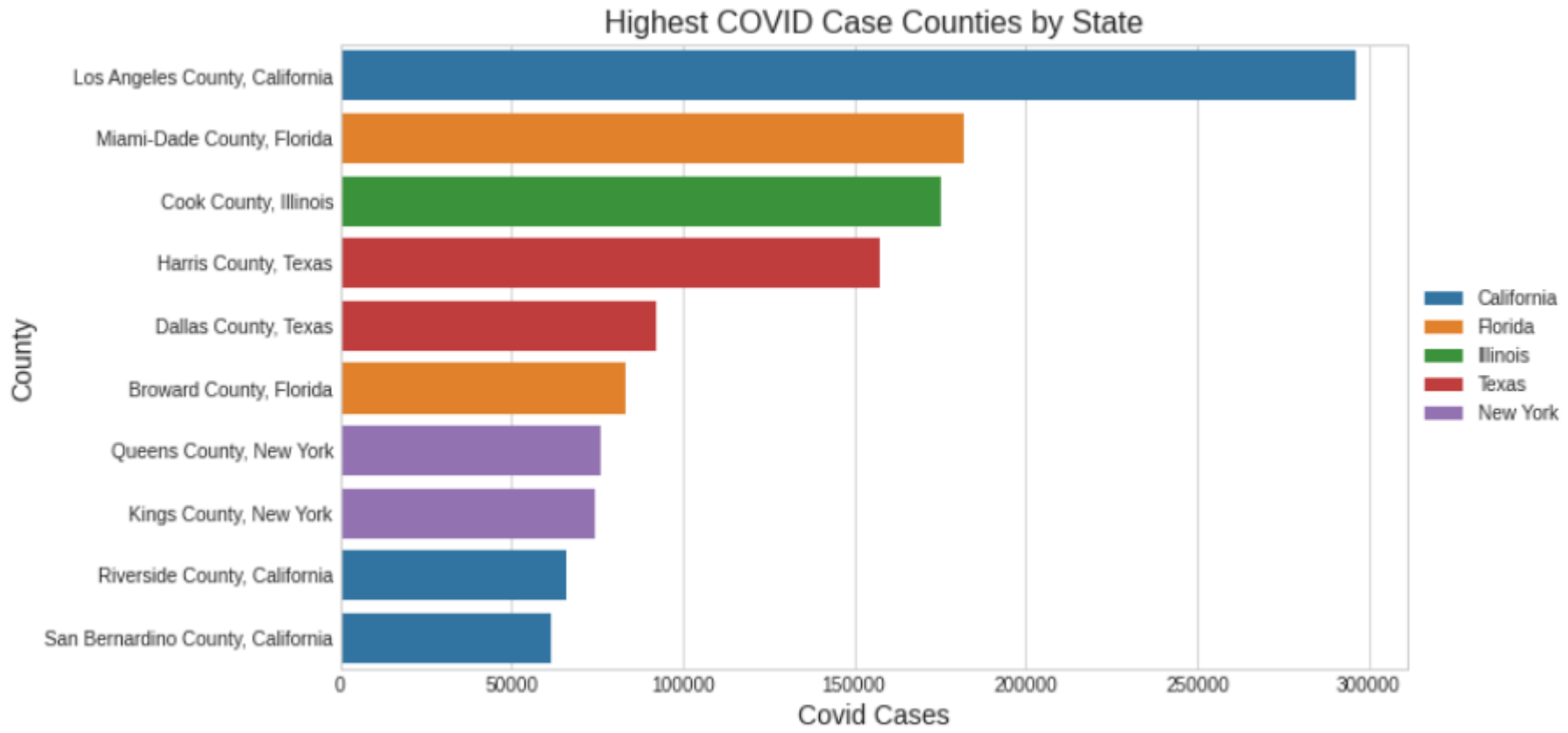




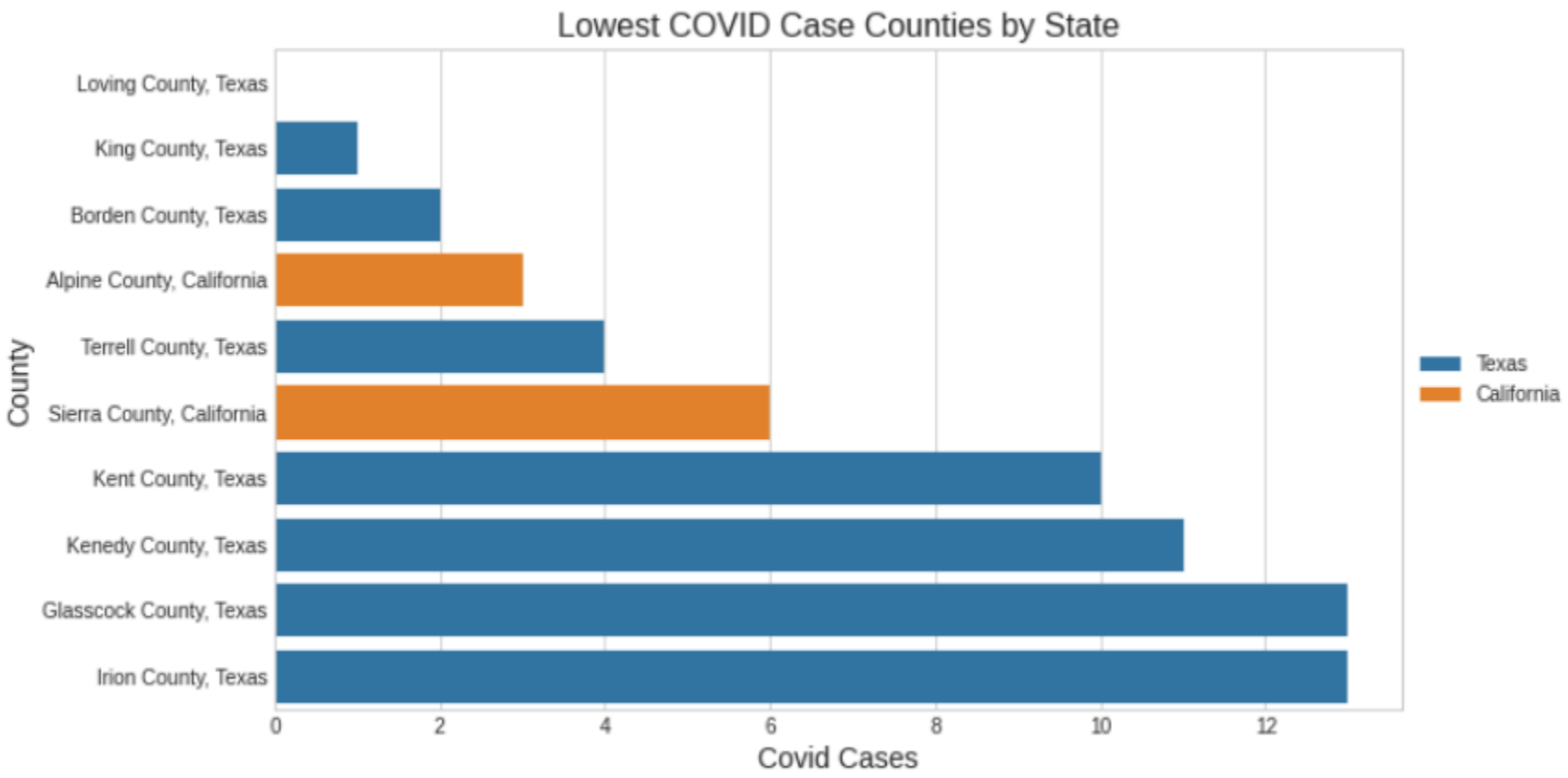
Highest Death Rate Counties by State



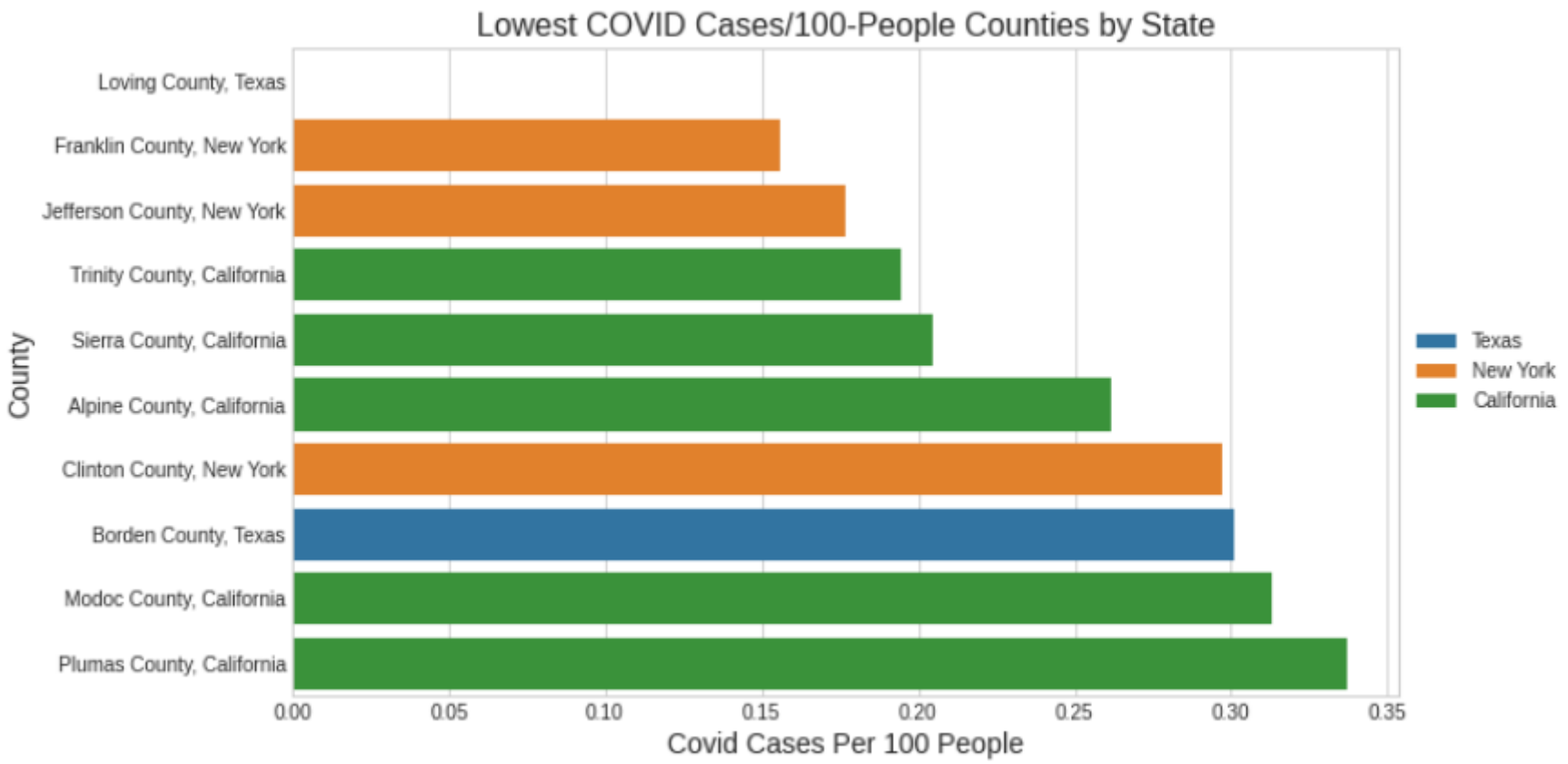
Highest Case Counties by State



Lowest Case Counties by State



Lowest Cases/100 People Counties by State



Highest Cases/100 People Counties by State