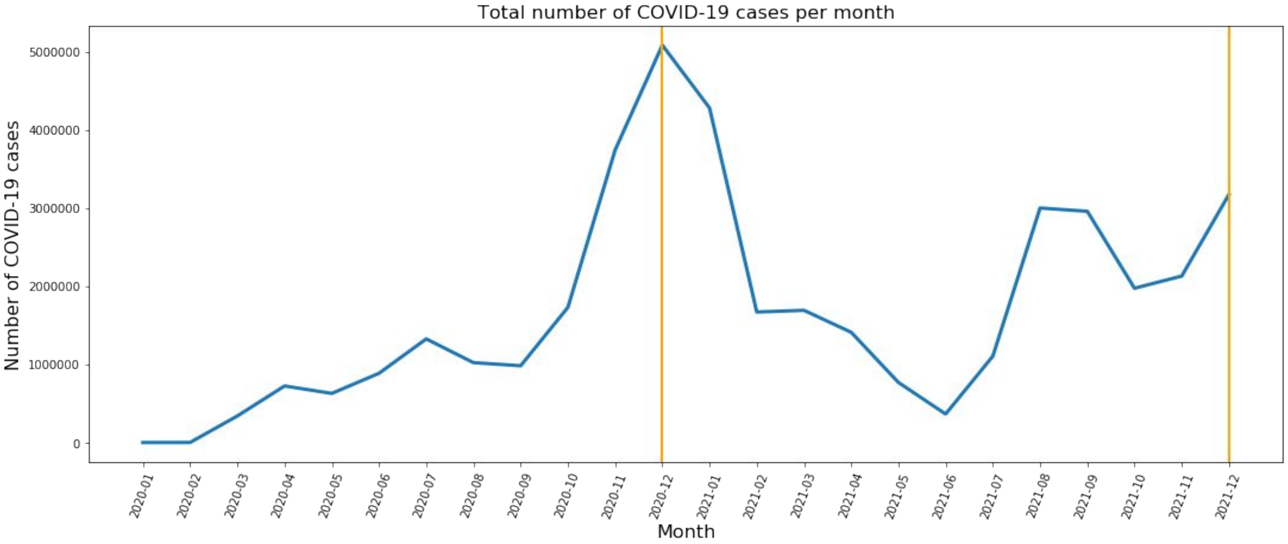
Project 5: Covid Death Rate Analysis

Dimitrios Vlahos, Douglas Hundley, Jake Ritmire, Juan Vega

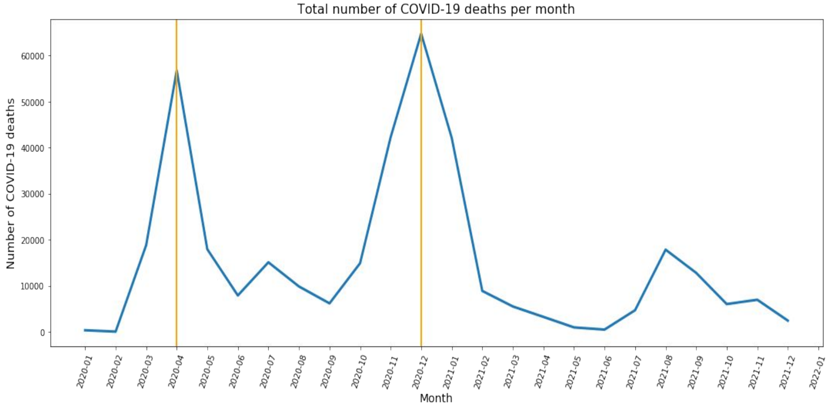
Background

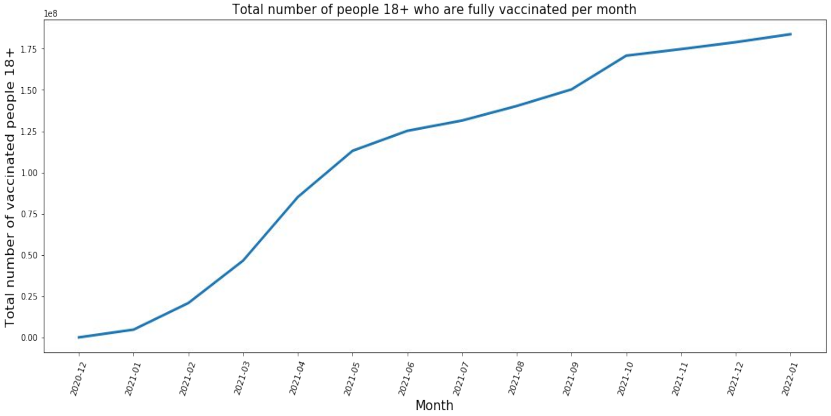
Since the first case of COVID-19 was diagnosed in January 2020, the pandemic has created multiple public health challenges across the U.S. The number of COVID-19 cases has grown very fast between 12/20 and 12/21



Background

The number of deaths from COVID-19 per capita per 100,000 population also increased significantly in this time period, reaching the highest peak in 12/20 and slowing down as vaccines started become available


Total number of COVID-19 deaths per month


Total number of people 18+ who are fully vaccinated per month

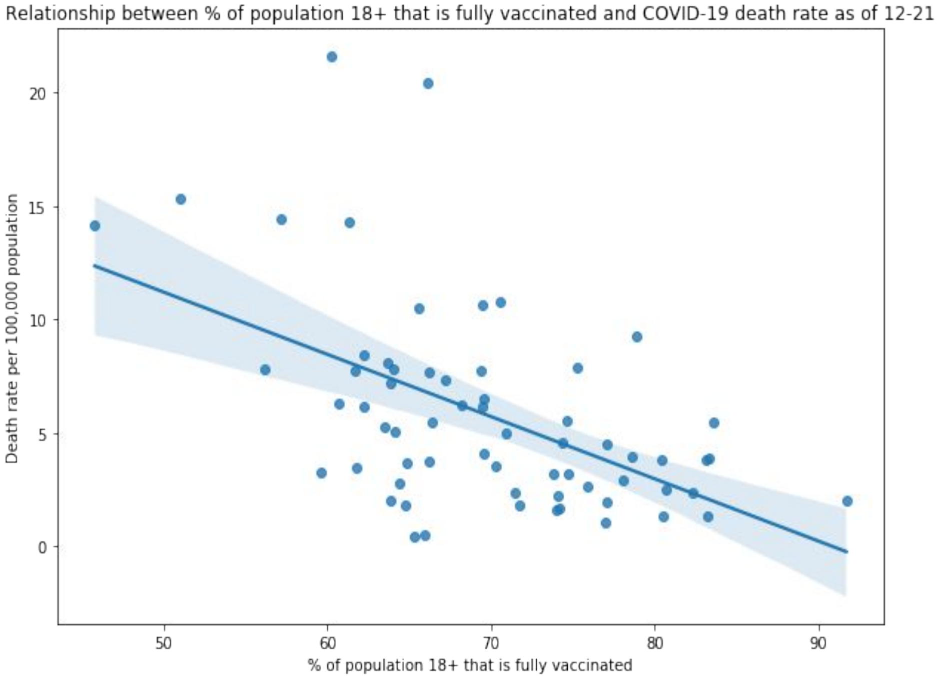
Problem Statement

- Given the importance of COVID-19 to the overall health of the U.S., the goal of this analysis is to identify counties that have the greatest need for COVID-19 relief as measured by the median number of deaths per capita per 100,000 population between 2020 and 2021
- Exploratory data analysis will be used to identify the regions where these counties are located and their characteristics
- Logistic and K-nearest neighbors models will be used to predict whether a county has a median death rate above the 75th percentile of all U.S. counties

Exploratory Data Analysis

Share of the adult population that is fully vaccinated and death rate

Using the latest data of COVID-19 case surveillance from the CDC from December 2021, there appears to be a moderate negative and linear correlation between the percentage of the adult population (18+) that is fully vaccinated and the death rate



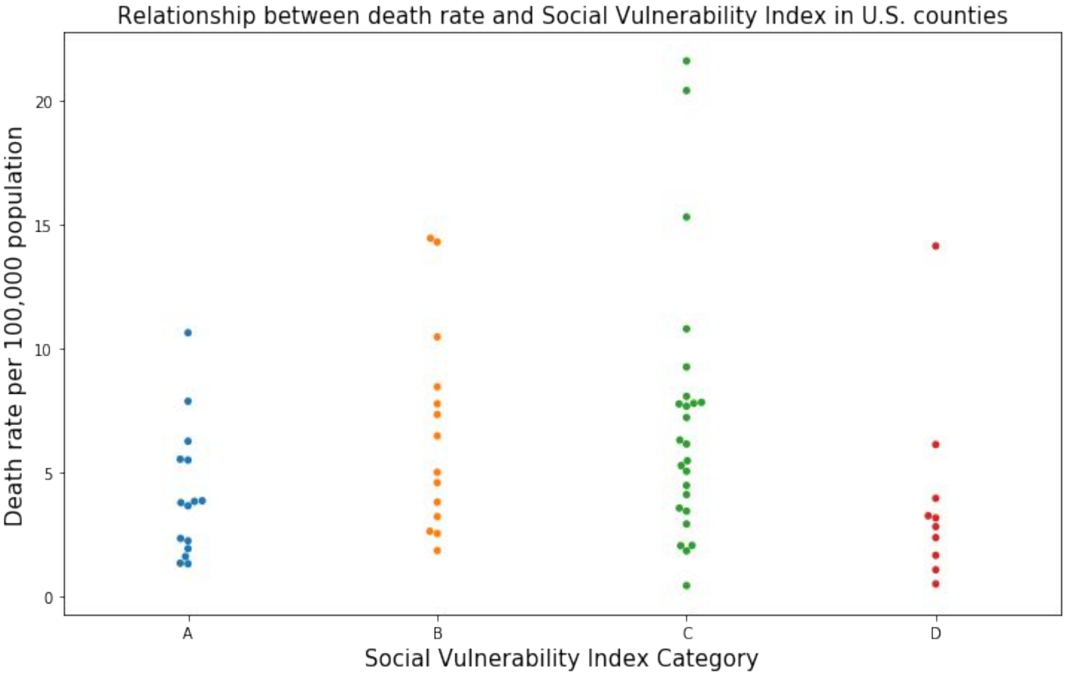
Death rate and social vulnerability

(Based on data from December, 2021)

Social Vulnerability Index (SVI) from the CDC is a measure of a community's resilience to natural disasters and disease outbreak based on socioeconomic factors such as poverty

SVI category A implies the lowest level of social vulnerability and category D the greatest level of social vulnerability

It appears that counties with greater levels of social vulnerability tends to have higher death rates, pointing to the need to support economically vulnerable communities

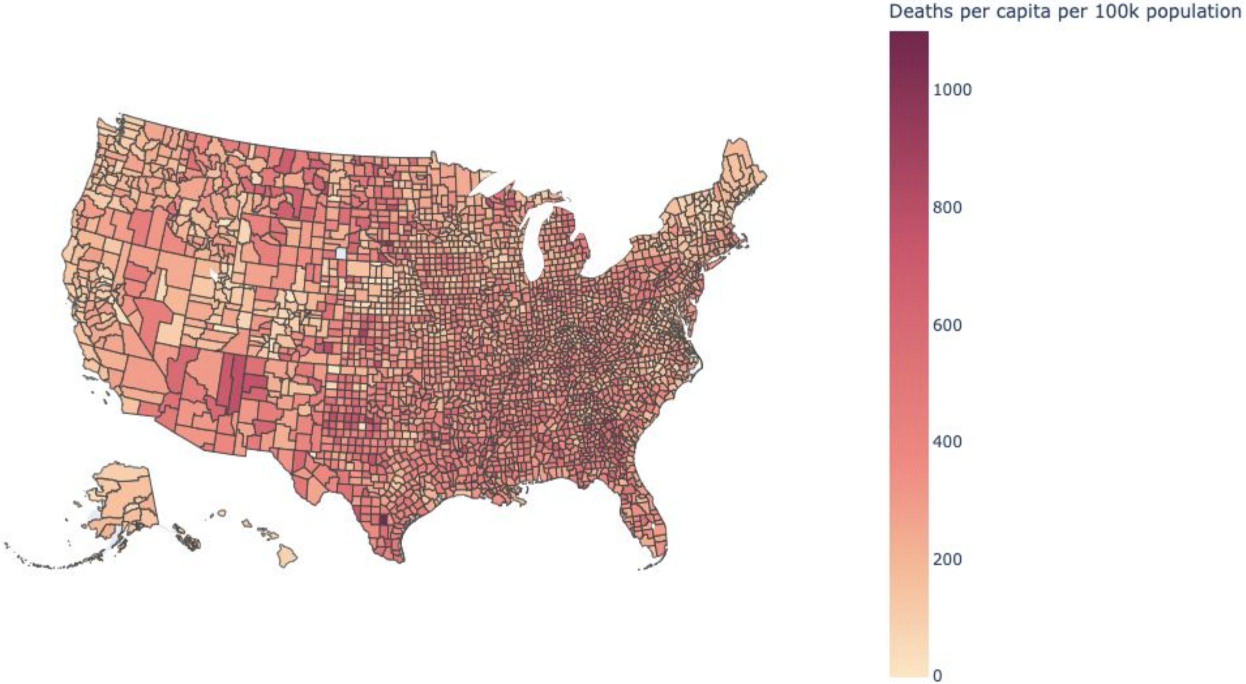


Geographic distribution of the death rate

Some counties have a disproportionately higher median death rate compared to other counties between 2020 and 2021

Counties in states like Arizona, Texas, Georgia, and Florida tend to be among the counties with the highest death rates per capita

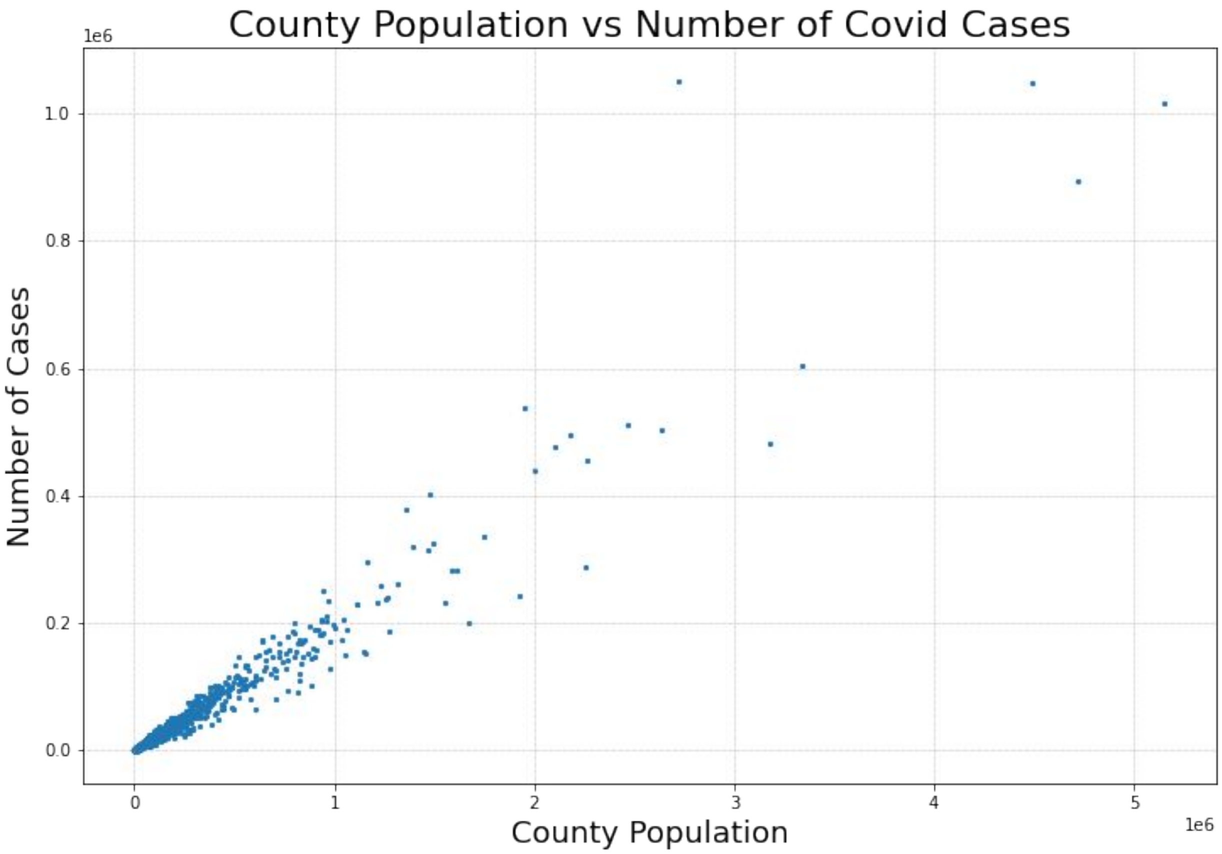
Counties with highest death rates are concentrated mostly around the southern United States



EDA

Strong linear relationship

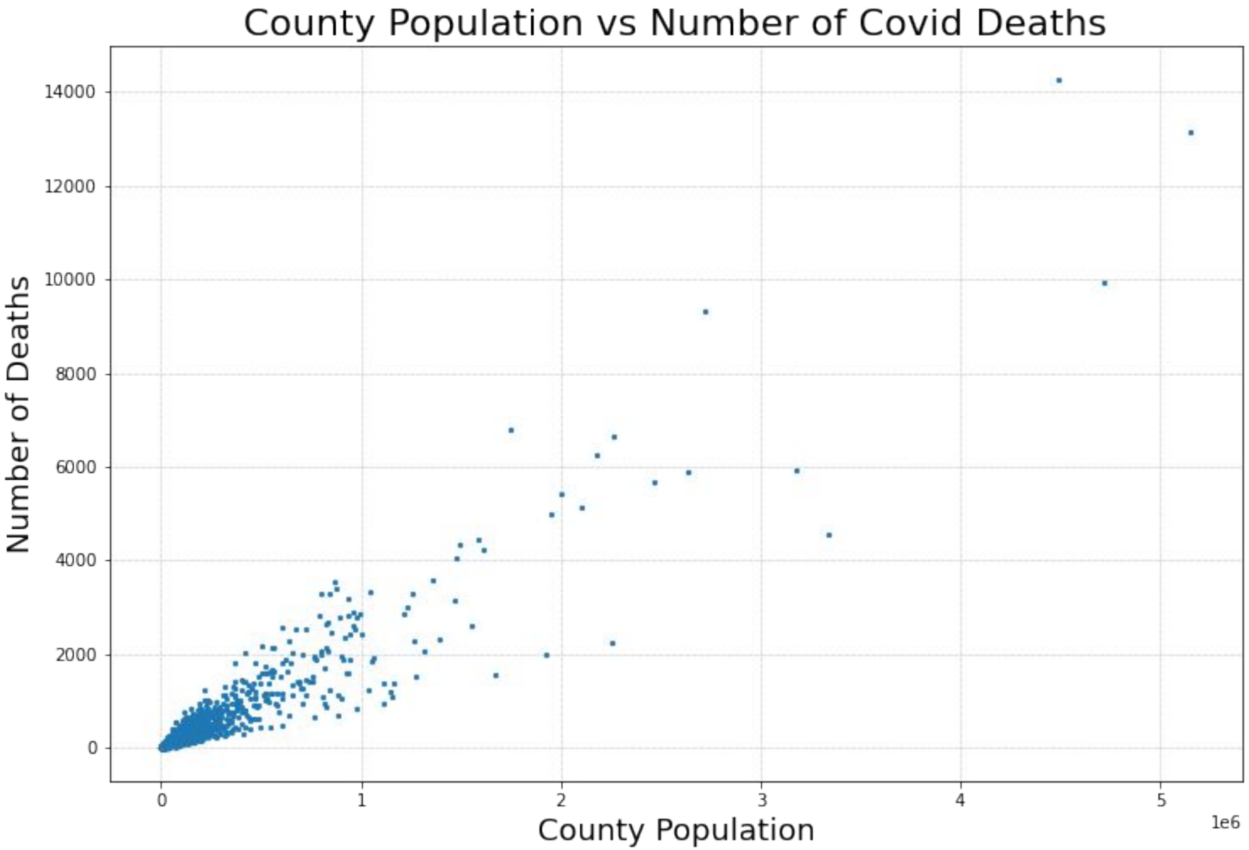
Indicates that population is a strong predictor of number of cases



EDA

Strong linear relationship

More variation than there was with number of cases



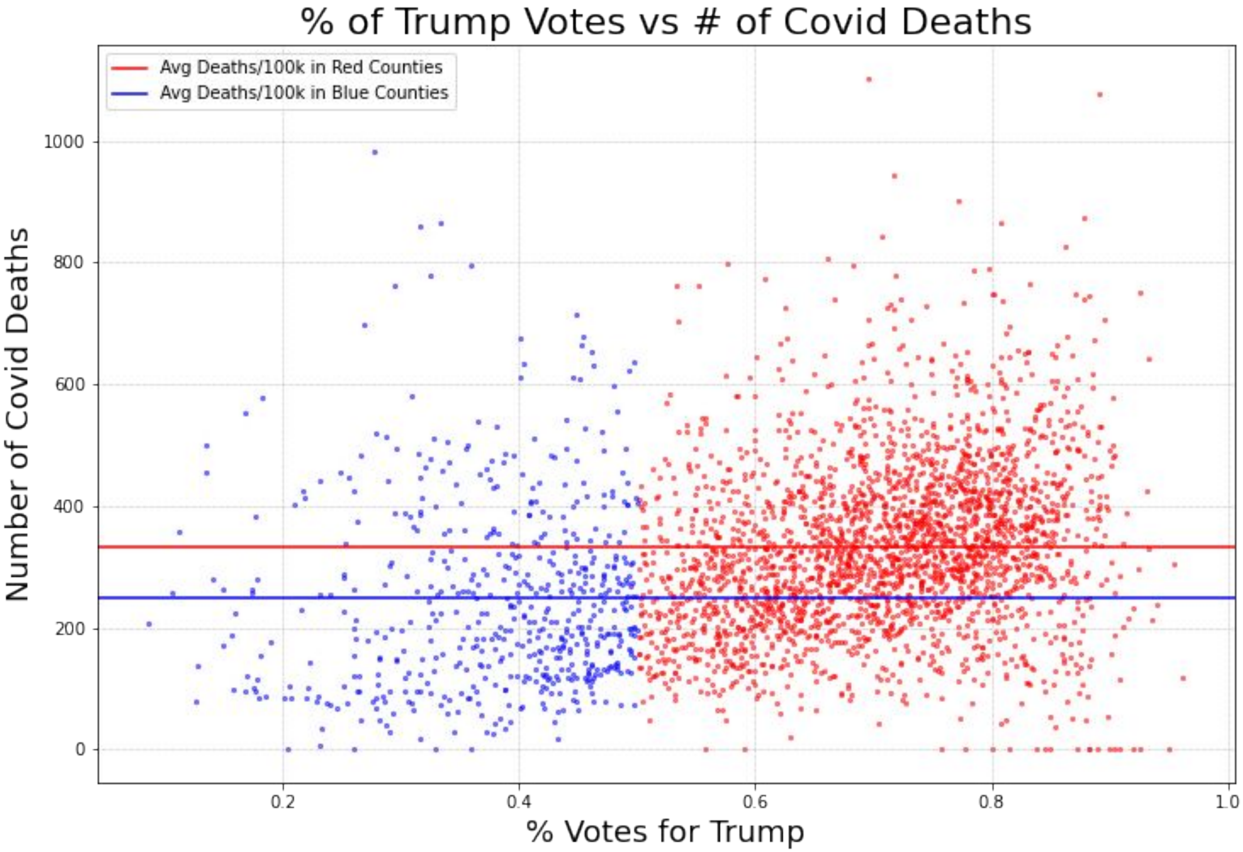
EDA

“Red” Counties >50% of voters voting for Trump

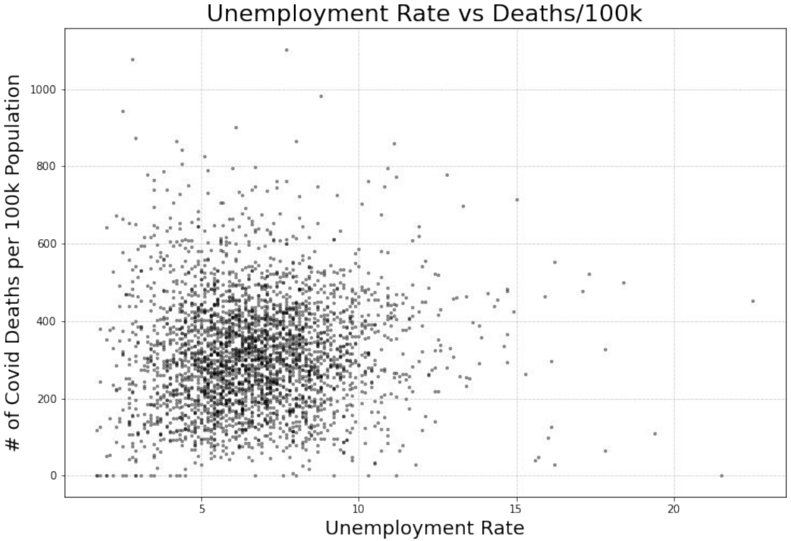
“Blue” Counties >50% of voters voting for Biden

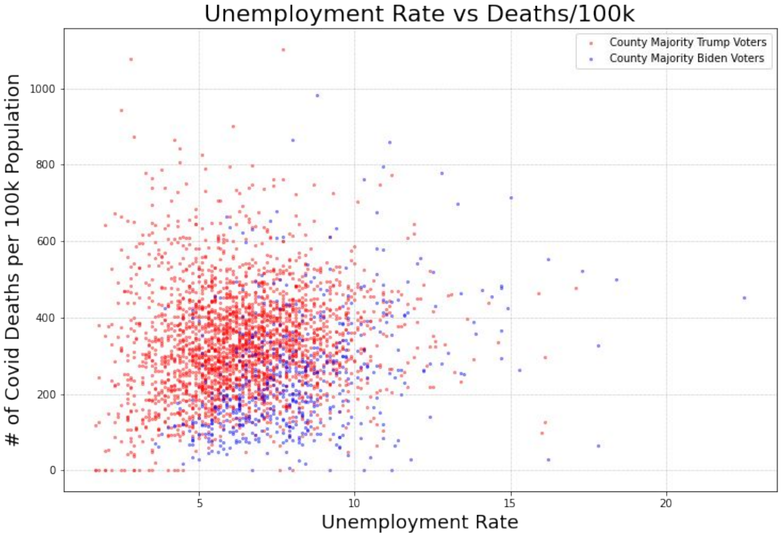
Red county deaths/100k: 333

Blue county deaths/100k: 259



EDA

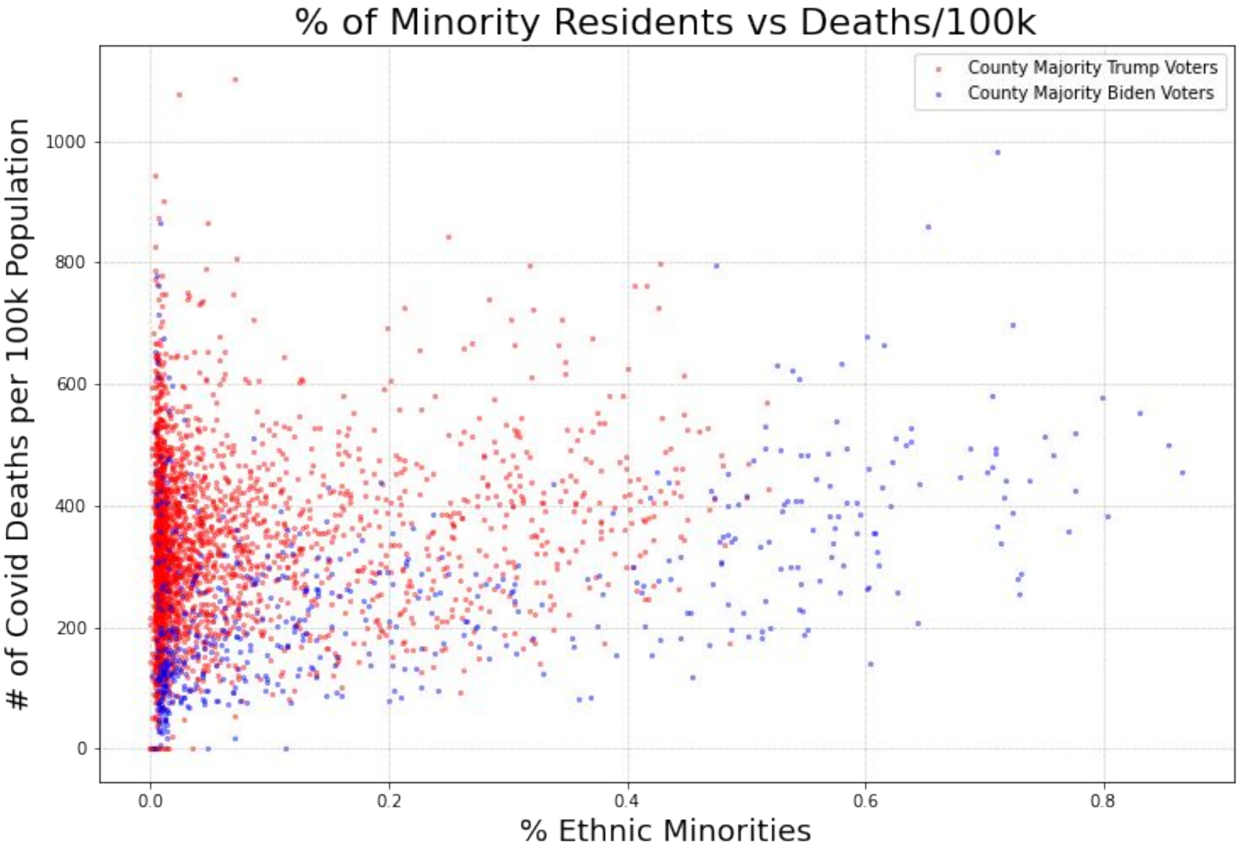

Unemployment Rate vs Deaths/100k
# of Covid Deaths per 100k Population
Unemployment Rate
0
200
400
600
800
1000
5
10
15
20


Unemployment Rate vs Deaths/100k
County Majority Trump Voters
County Majority Biden Voters
# of Covid Deaths per 100k Population
Unemployment Rate
0
200
400
600
800
1000
5
10
15
20

EDA

Counties with more than 50% of population identified as minorities are almost exclusively democrat

As % of minority residents increases, we see number of deaths/100k increase slightly



Data Sources

- Covid Data - CDC and New York Times Covid Repository
- County Demographic Data - US Census Bureau
- Unemployment Data - Bureau of Labor Statistics
- Election Data - MIT
- Hospital Capacity - Health Data.gov
- COVID cases surveillance from the Center for Disease Control

Ultimately compiled a dataset containing 3,133 observations of 54 counties.

Methodology

- Target variable made sense
- Used PCA to slim down data that we used
- Resulted in a precision score measuring how likely the model would be able to predict to the 75th percentile

Logistic Regression

- Cursory testing brought better scores than a regression model, stuck with classification
- Counties above 75th percentile was the focus, categorized as ‘high need’ for COVID relief
- Used Iterative Imputing to fill in missing data; 700+ missing hospital capacity data for counties
    - **5 principle components** made up about **70%** of the models predictive power
    - Baseline Score: **74.9%**
    - Modeling Score: **77.5%**
    - Modeling w/ PCA Score: **76% +/- 3%**

KNN

- Dataset had many nulls
- Best scoring model had .783 accuracy
- Logistic is the better way to go

Results and Conclusion

- Decided to move forward with the Logistic Regression model
- Concluded model would not be strong enough to predict target
- Recommendations:
  - use visualizations and exploratory analysis from data variables and resulting model coefficients

Coefficient	Short Desc.	Exponentiated Value
land_area_sqmi	Total land area of county in square miles	6.883
population_desity	Total pop. / Area Per sq. mile	3.787
cases_per_10k	# of cases per 10k people	1.898
biden_votes	# of votes for Biden during election	1.769
med_cases_per_100k_change	Median of new cases per 100k people in 7 days	1.533

Proposed Future Analysis/Next Steps

- Collection of data needs to be over a greater window of time
  - Analyze a cross-section data (the most recent month of data)
  - Use time-series methods to exploit the longitudinal nature of the data
- More uniform methods on what data is collected
- Compare model performance based on the data provider
  - New York Times
  - CDC