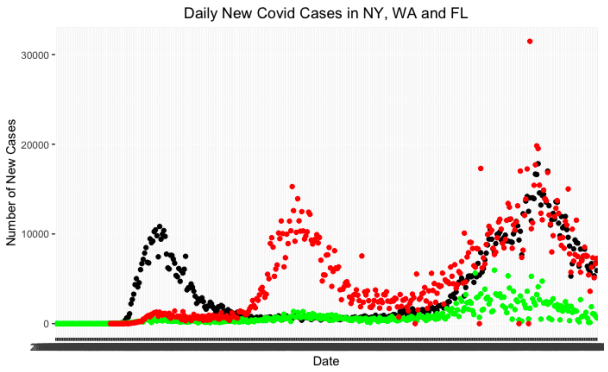
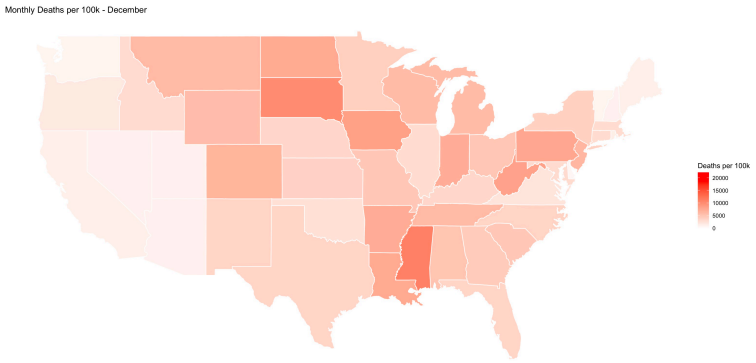
COVID-19 Case Study Summary

The goal of this case study is to discover what county-level demographic and policy interventions are associated with mortality rate in the US and to construct models to find possible factors related to county-level COVID-19 mortality rates.

From examining "covid_rate.csv," we can see case number and death number on county and state level. COVID-19 deaths start to emerge in April 2020, and it hits different parts of the United States differently in terms of number and time. Another file is "county_data.csv," which includes data with various socioeconomic characteristics, which is helpful for examing possible factor related to country-level mortality rates. To see the data clear, we drew multiple plots such as new COVID cases in NY, WA and FL by state and by day, spaghetti plots of weekly_case_per100k by state, and monthly deaths per 100k heatmap by state on US map. We can see two examples below.





The impact of COVID-19 on different regions of the United States varied over time, with the eastern coastal states being hit hardest initially, followed by a shift towards southern states and then mid to upper states towards the end of 2020.

We used LASSO for model selection. Then we utilized Cp or BIC to fine tune the model. Next, we reduced the model so that the final model includes only significant variables.

Among different age groups, COVID affects elderly the most. In different ethic groups, African Americans and Latinos have higher death rate than others do. Different Occupations are also influenced differently by COVID.

We still face some limitations during our case study. The number of people age 65 and older, African Americans, and Hispanic American are data from 2010, while the data we use for COVID-19 case and death rate is from 2020 to 2021, which might create inaccuracy. Our results would be more accurate if we have data of population of elderly, African Americans, and Hispanic American from the same time period.