Final Report:
Team: Linsy Wang, Sooyoun Lee, Dana Zheng
Title: Relationship between Health Insurance and COVID-19 Rates and Deaths

Project Scope and Goal:

According to the official definition by the CDC, social determinants are conditions that people are born and live in. Differences in these conditions lead to health inequities and population health outcomes. We want to better understand how social determinant factors influence the spread and severity of diseases like COVID-19 by looking at confirmed cases and death numbers across the country.

We will primarily focus on health insurance coverage because we think it's a good reflection of a multitude of these determinant factors like income, marriage, and geography. Therefore, the main goal of our project is to determine the relationship between health insurance coverage and confirmed cases/death rates of COVID-19 in the US. Our findings will demonstrate how population health is affected by social determinants of health such as health insurance coverage.

Data collection:

The datasets that we collected for confirmed cases and death numbers of COVID-19 were obtained via Github repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (link). The two datasets for confirmed cases and death numbers were in the form of csv files. Both datasets consisted of columns detailing the county names, county area codes, state names, and corresponding values for 1/22/20 to 3/20/21.

The dataset for U.S. health insurance coverage in 2018 was obtained via The United States Census Bureau (link). The dataset included columns detailing relevant information such as county and state names, population count, uninsured/insured number of people, and uninsured/insured percentage by county.

The three datasets were collected as csv files and viewable as Excel spreadsheets.

Data processing:

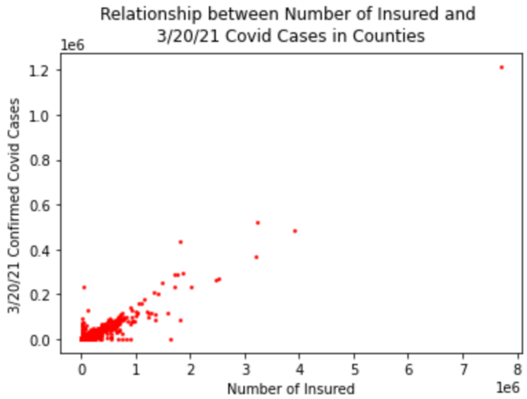
Once we collected all the relevant data, we preprocessed them in order to make the data cleaner. In the process, we dropped irrelevant columns such as 'UID', 'iso2', 'iso3', 'code3'. Other relevant columns were dropped and used based on the different datas we were attempting to plot and discover. Additionally, we made relevant changes to the data types, such as converting the population data into floats to make the plotting process more clear and efficient.

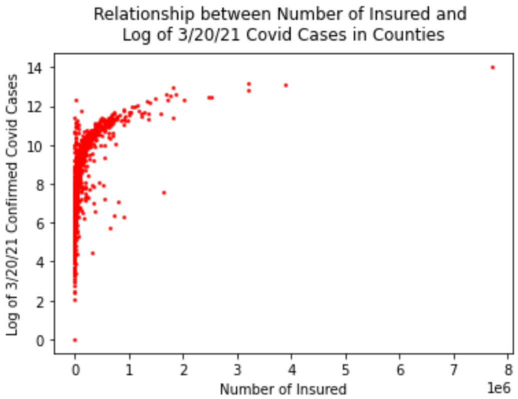
Along with this, we discovered duplicated or unassigned counties in some of the datasets so those were removed. For instance, Utah was divided into four separate sections in the dataset and was thus merged into one. We also found several counties with no data, so those counties were found and dropped manually.

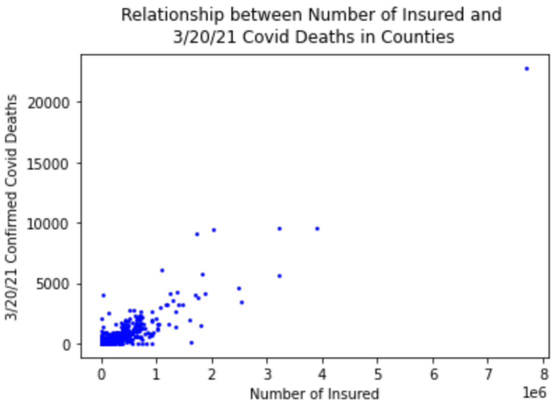
Whilst going through the datasets further, we noticed that the population, death, and covid rates were only in actual numbers rather than percentages or ratios. In order to make this more clearer when plotting, we also processed the data for the percentages to be calculated and they were renamed into new columns. For instance, in order to get the ratios of deaths and cases, death rates and number of cases were each divided by the population.
The data in the datasets are all cumulative, so when plotting the actual data we used the most recent date in the dataset.

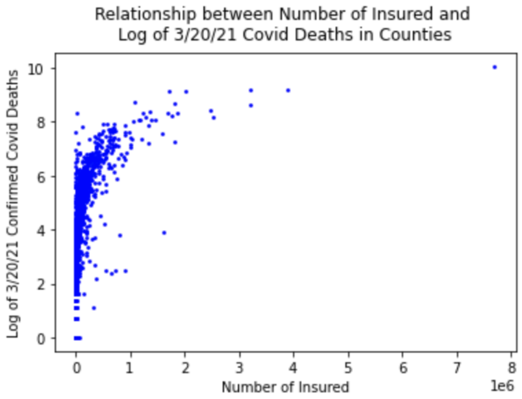
Data visualization:

The data visualizations are all scatter plots. Our first visualization is a graph showing the correlation between the number of insured and number of confirmed cases.

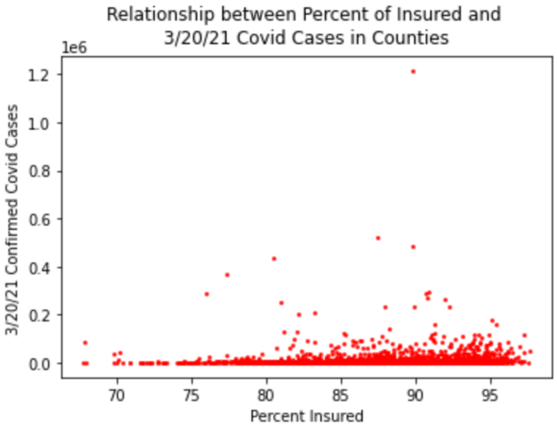

Relationship between Number of Insured and
3/20/21 Covid Cases in Counties

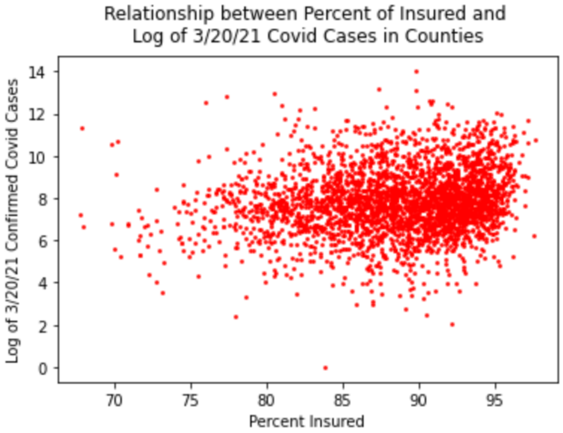

Relationship between Number of Insured and
Log of 3/20/21 Covid Cases in Counties

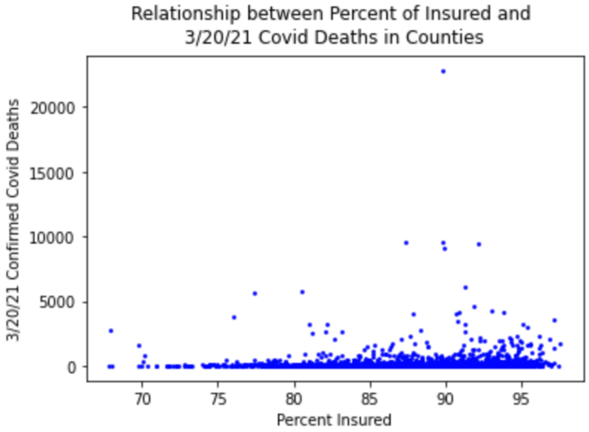

Relationship between Number of Insured and
3/20/21 Covid Deaths in Counties

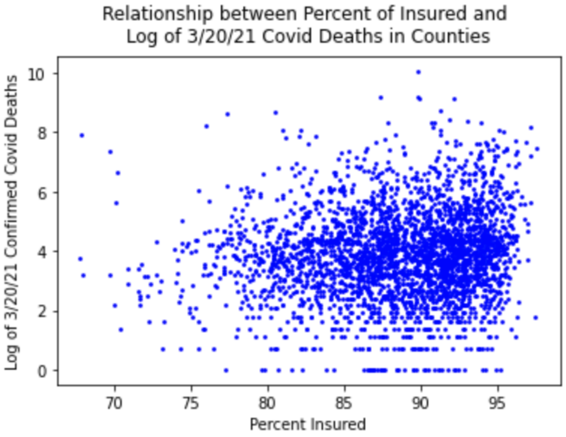

Relationship between Number of Insured and
Log of 3/20/21 Covid Deaths in Counties

Based on the plots, there seems to be a direct correlation. However, it seems to mainly be affected by population. When we plot the percent insured with the confirmed case number, the county data begins to to form one general cluster.



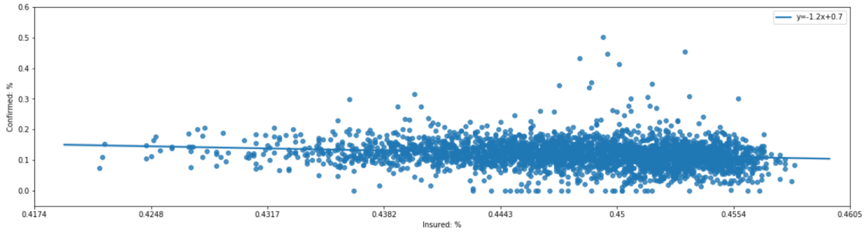




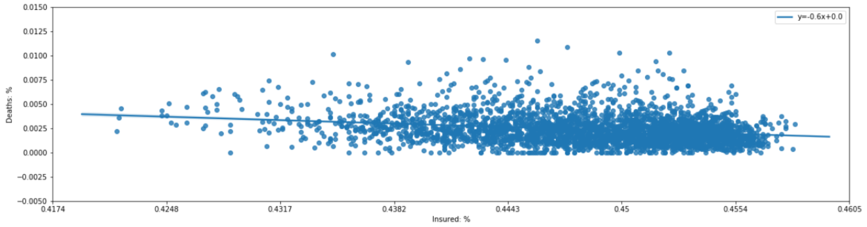


When plotting the percent insured to percent of confirmed cases or deaths, there seems to be a negative correlation.

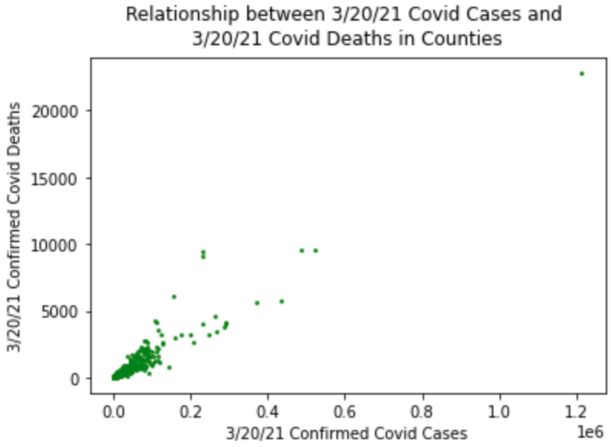
Insured % and Confirmed %

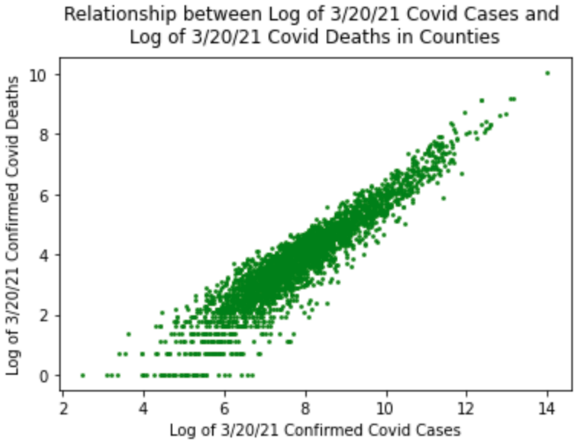


Insured % and Deaths %

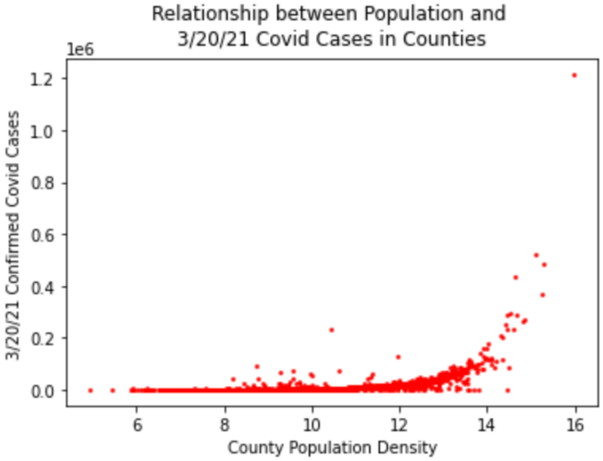


There is a positive correlation between confirmed cases and deaths, although it may also be affected by population.

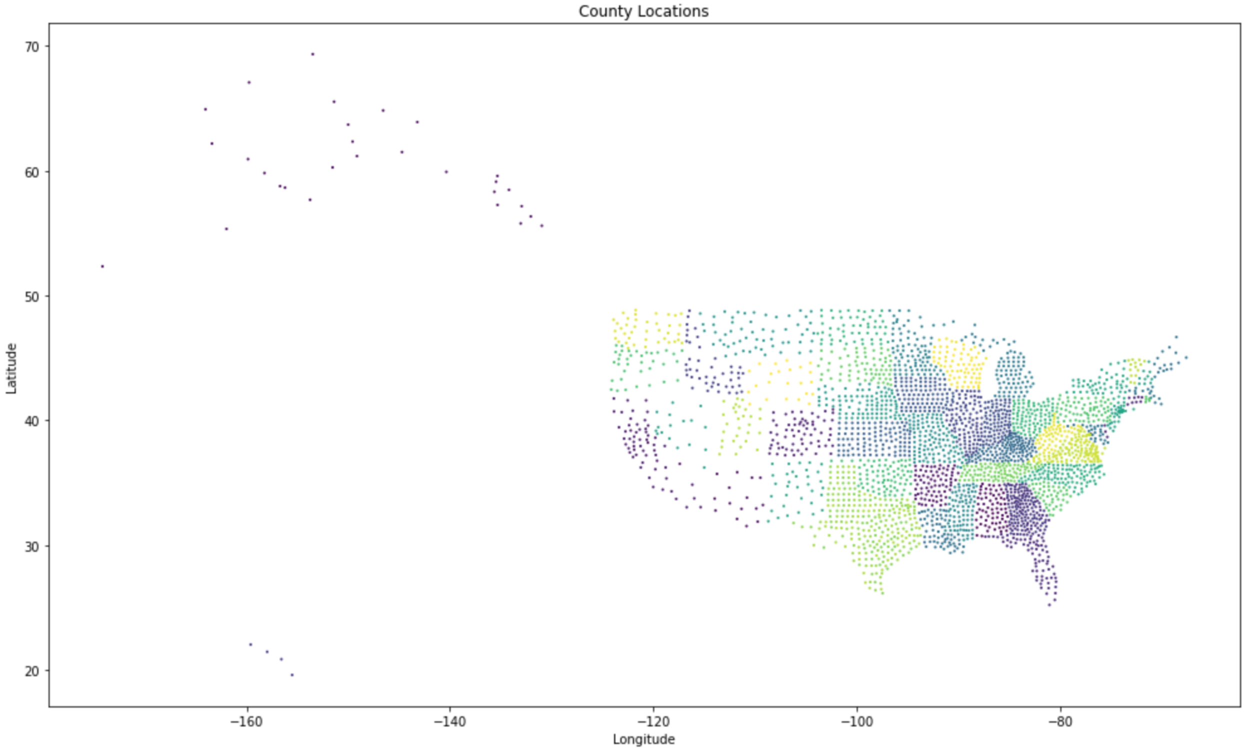




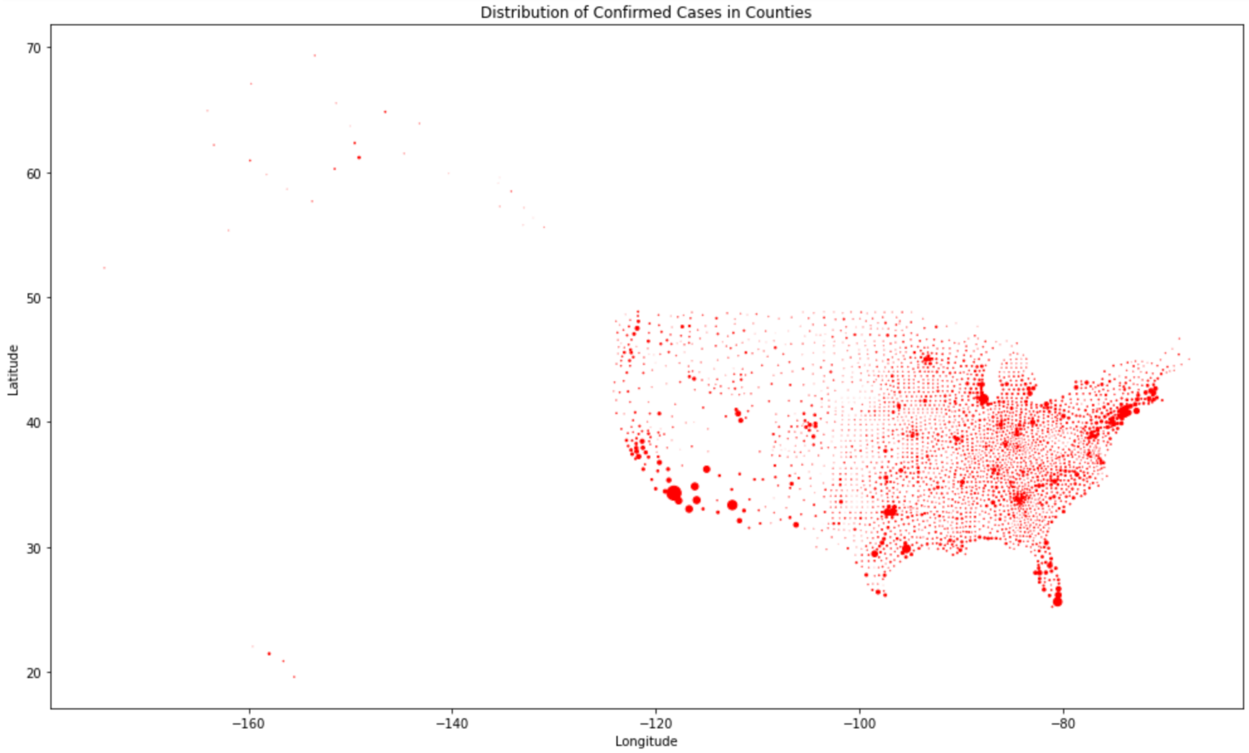
Through the relationship between county population and confirmed cases, it can be seen that covid cases grow exponentially based on the population. In addition, many counties have a similar number of confirmed cases.



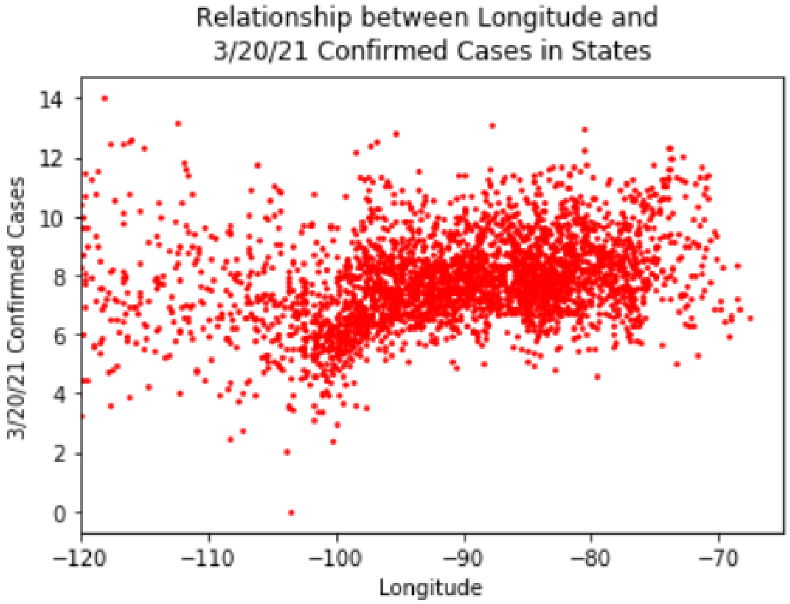
In addition to plotting covid cases, deaths, and insurance, we also plotted the counties based on their longitude and latitude.



To combine the covid and insurance data with the location of the counties, the size of the counties is based on the number of confirmed cases in the county.



We plotted the number of total confirmed cases in US counties and the longitude of each county. The scatterplot indicates there is a majority of confirmed cases between -100W and -70W longitude, which corresponds to the eastern half of the country.



Data analysis:

The regression line for percentage of confirmed cases and percentage of insured county population has a slope of -1.2, and the percentage of confirmed cases and percentage of insured population has a slope of -0.6. These findings indicate that there is a slight correlation between health insurance coverage and COVID-19 cases, where counties with increasingly insured populations demonstrate less confirmed cases and death numbers.