Final Report

Title: Relationship between Health Insurance and COVID-19 Rates and Deaths

Team:
Linsy Wang, linsyw@bu.edu
Sooyoun Lee, selena98@bu.edu
Dana Zheng, daynah@bu.edu

Project Manager:
Yifu Hu, yfhu@bu.edu

Project Scope and Goal:

According to the official definition by the CDC, social determinants of health are specific conditions people are born and live in which lead to health inequities and population health outcomes. Examples of social determinants include socioeconomic and physical factors, such as education, health care coverage, natural environment, and poverty. We want to better understand how social determinant factors influence the spread and severity of diseases such as COVID-19 by looking at confirmed cases and death numbers across the U.S.

We refined our project scope into two key questions:

1. What is the relationship between COVID-19 cases and access to health insurance? How does this relationship change across various states and counties in the United States?
2. What are other social determinants of health that affect coronavirus cases? How do these relationships compare with health insurance?

We decided to primarily focus on health insurance coverage because it is a good reflection of several determinant factors such as income, marriage, and geographical location. Therefore, the main goal of our project is to determine the relationship between health insurance coverage and confirmed cases and death rates of COVID-19 in the US. We hypothesized that there would be less confirmed cases and death numbers in highly insured populations. Our findings will demonstrate how population health is affected by social determinants of health such as health insurance coverage, and if given more time, we aim to analyze the effects of additional determinant factors.

Data Collection:

The datasets we collected for confirmed cases and death numbers of COVID-19 were obtained through a 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 county names, county area codes, state names, and corresponding case or death values from 1/22/20 to 3/20/21. The number of cases were cumulative by dates, so the values in the “3/20/21” column represent the total number of cases for each recorded county.

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

The three datasets were collected as csv files and viewable as Excel spreadsheets. We converted them into dataframes to clean and process the data.

Data Processing:

Once we collected all the relevant data into dataframes, we dropped irrelevant columns and identified which columns needed to be further manipulated to visualize our data. Additionally, we made relevant changes to the data types, such as converting the population data from comma separated integers into floats to make the plotting process more clear and efficient.

All the datasets sorted their county information alphabetically based on state name, then county names. However, we still ran into issues aligning corresponding counties between the datasets and discovered duplicated and unassigned counties in the COVID-19 datasets. For instance, the state of Utah was divided into three separate sections (Central Utah, Southeast Utah, Southwest Utah) and considered counties in the datasets. We also found several counties with no data in all of the datasets. The misaligned and counties with no data had to be manually found and removed. Along with this, we had to match the counties between the two datasets by matching their respective FIPS and ID codes in order to combine the Insurance and COVID-19 datasets.

In addition, there were also misalignments in state data. The COVID-19 datasets included US Provinces as states, while the insurance dataset did not include any provinces. Another issue was that the insurance dataset included insurance data for each state while the COVID-19 datasets did not. To solve this issue, we found the sums of the COVID-19 cases and deaths for counties in the same state, and used that data for our models. Because the Insurance dataset had state data, it duplicated Washington DC by considering it a state and county. One of the duplicates was manually removed.

The population, death, and COVID-19 rates were in actual numbers and did not include percentages or ratios. In order to make this clearer when plotting, we processed the data for the percentages to be calculated and they were added back into the dataframe in 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. Once the data was plotted, we also noticed that there are a lot of outliers on the graph. We removed the outliers by removing all values with a z-score larger than 3 in the data for insured, death, and COVID-19 cases. The data in the datasets are all cumulative, so when plotting the actual data we used the most recent date in the dataset.

Final Dataframe Schema:

US_Confirmed and US_Deaths

The datasets for Confirmed Cases and Death Rates contained the same set of columns with the only difference being that one contained data for confirmed cases and the other contained data for number of deaths.

Columns:

- FIPS: code number for the county
- Admin2: county name
- Province_State: state name
- Lat: latitude of the county
- Long: longitude of the county
- Combined_Key: county, state format
- 1/22/2020 to 3/20/21: number of confirmed cases / deaths in a county for that day

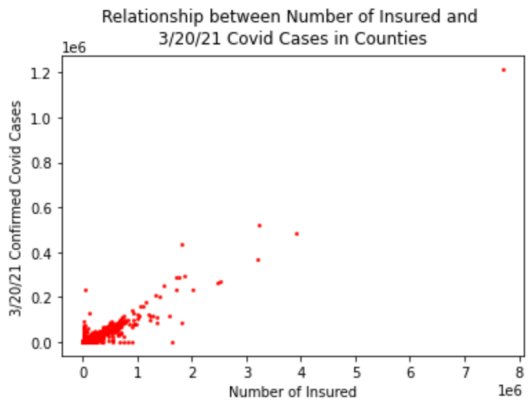
US_Insurance

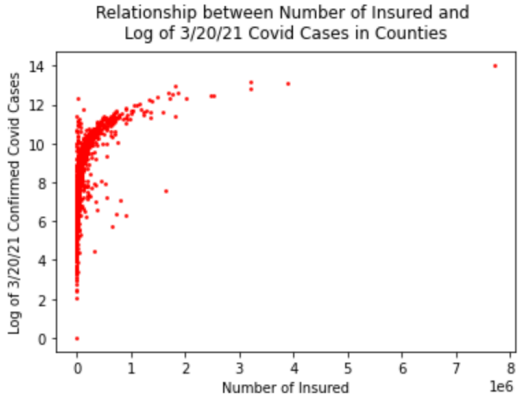
Columns:

- ID: code number for the county (matches FIPS)
- Name: county name (matches Admin2)
- Demographic Group Number: county population
- Uninsured Number: number of uninsured population
- Uninsured MOE: uninsured population margin of error
- Insured Number: number of insured population
- Insured MOE: insured population margin of error
- Insured %: percentage of insured population

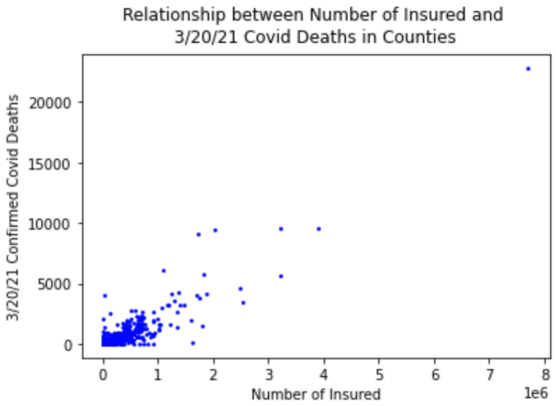
Data Visualization:

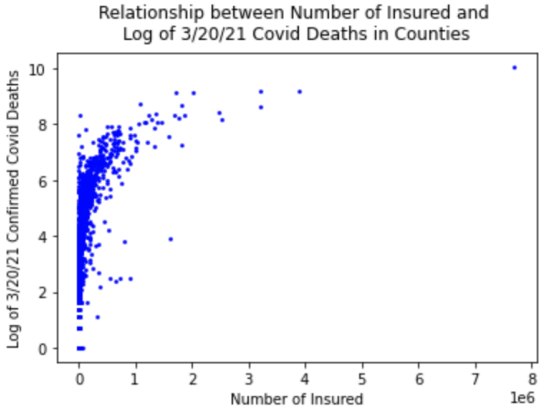
The data visualizations are all scatter plots. Our first visualizations show the correlation between the number of insured and number of confirmed cases.



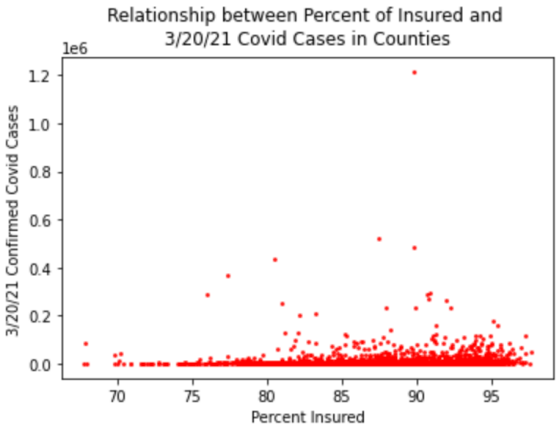


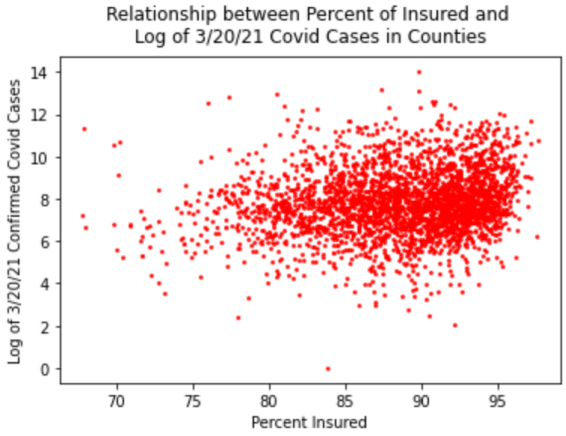
Similar plots were created to show the correlation between the number of insured and number of confirmed deaths.

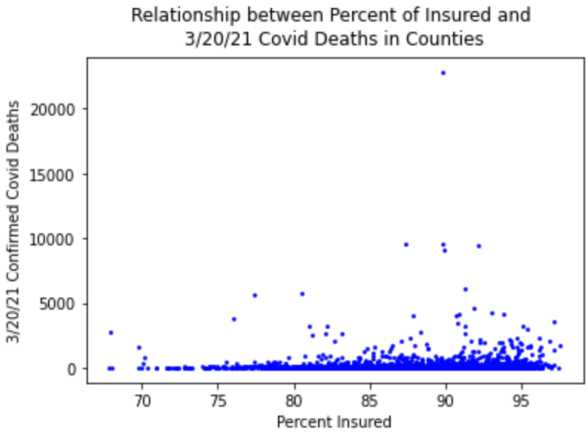


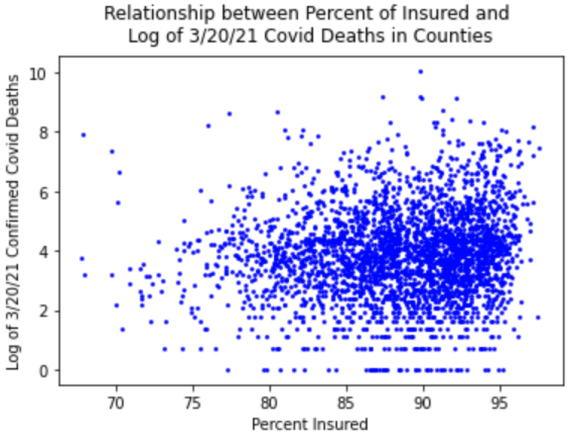


Based on the plots, there seems to be a direct correlation. However, the increase of insurance number and case number is being affected by population. When we plot the percent insured with the confirmed cases and deaths, the county data begins to form one general cluster.





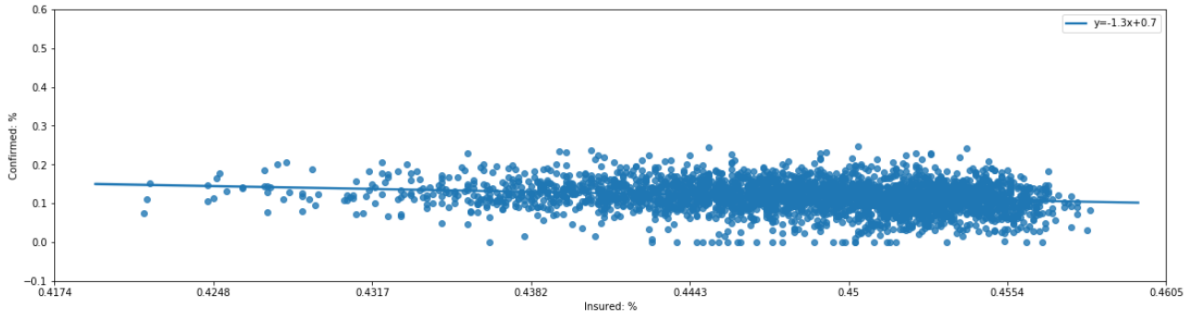




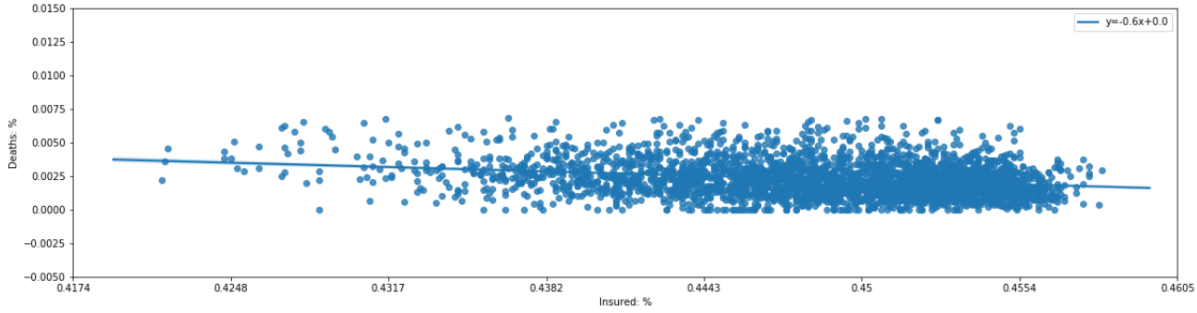
To better compare the variables in respect to county population, we plotted the percent insured and percent confirmed cases and deaths.

For March 20, 2021, the relationship between insured and confirmed rates is relatively constant with a slope of -1.3 for the linear regression line. The R^2 value was 0.0380. Similarly, the relationship between insured and death rates is also constant with a slope of -0.6. The R^2 value was 0.0625.

March 20, 2021 - Insured % and Confirmed %

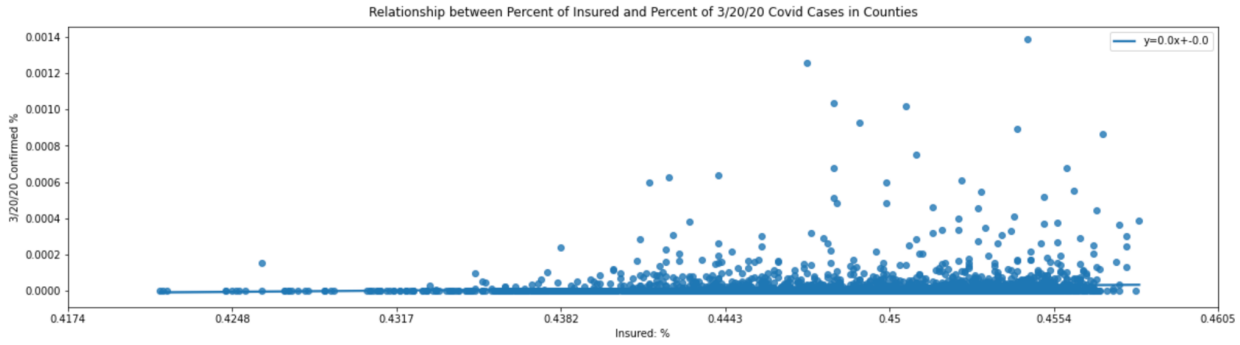


March 20, 2021 - Insured % and Deaths %

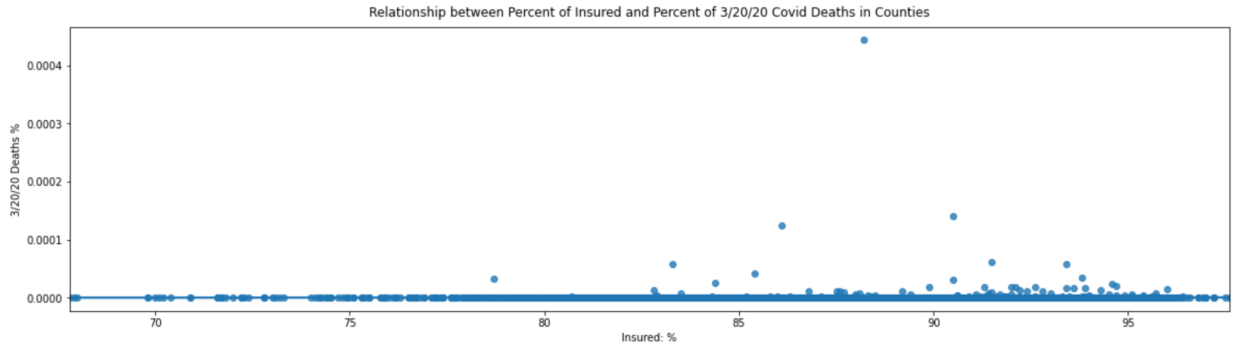


We also visualized data from exactly one year earlier. For March 20, 2020, the relationship between insured and confirmed rates is constant with a slope of 0.0 for the linear regression line. The R^2 value was 0.0076. Similarly, the relationship between insured and death rates is also constant with a slope of 0. The R^2 value was 0.00007.

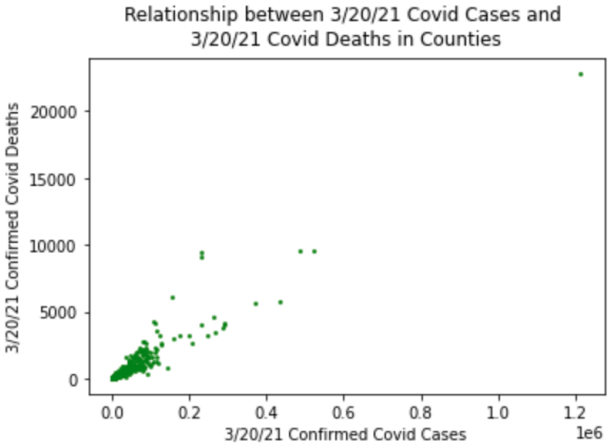
March 20, 2020 - Insured % and Confirmed %

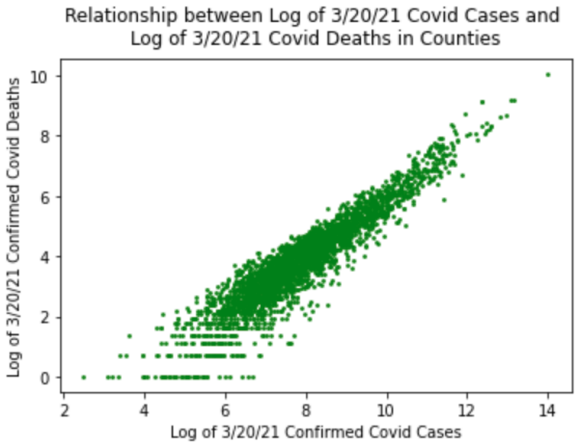


March 20, 2020 - Insured % and Deaths %

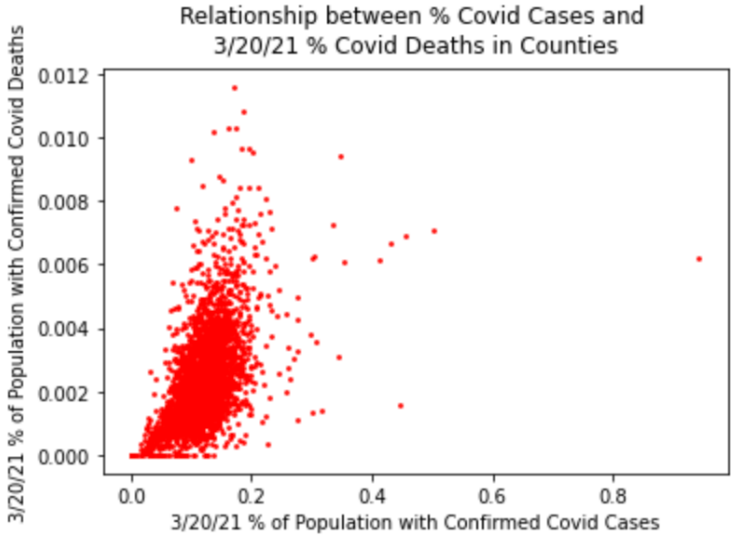


There is a positive correlation between confirmed case and death numbers, although based on previous graphs we knew it was being affected by population.

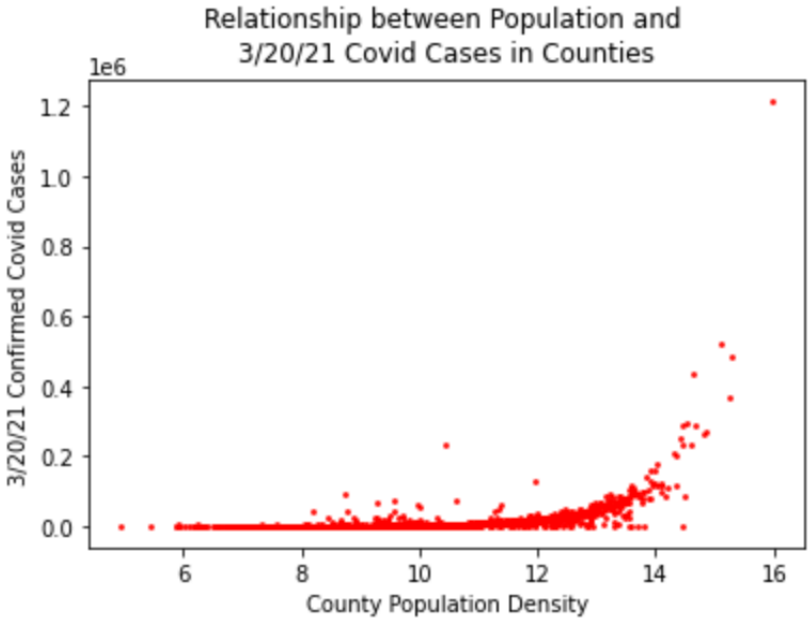




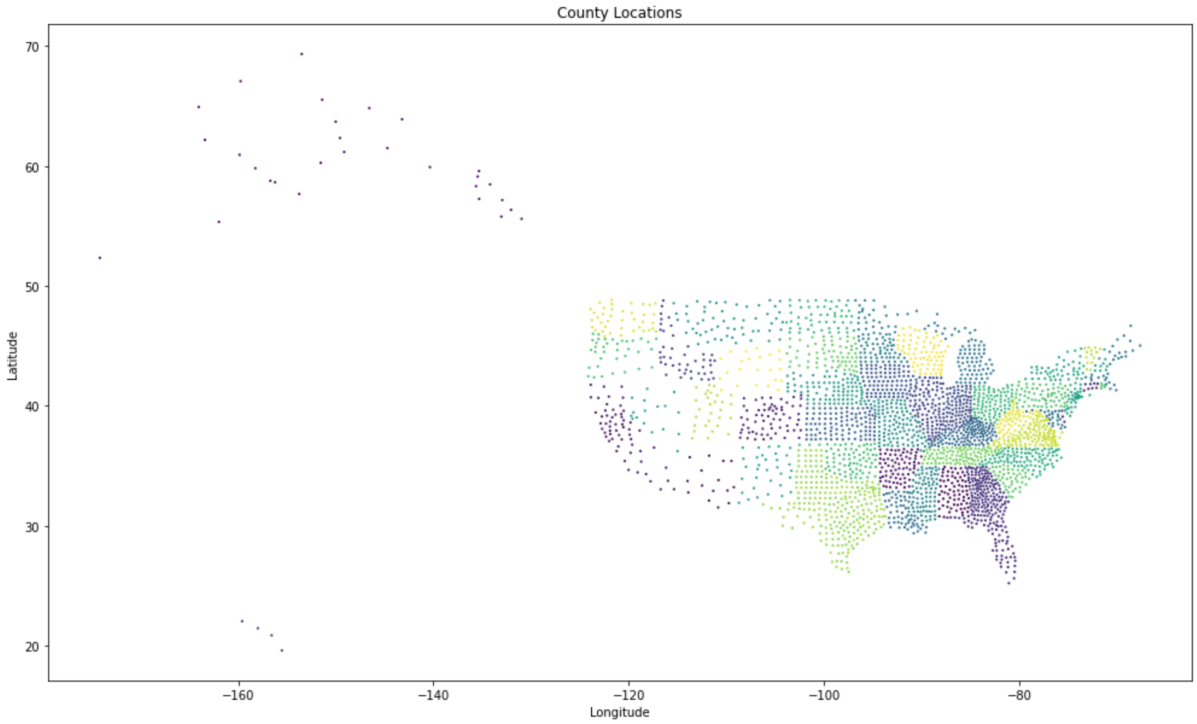
To remove the effect the population may have on the data, we plotted the case and death percentages. It confirms some correlation between COVID-19 cases and deaths.



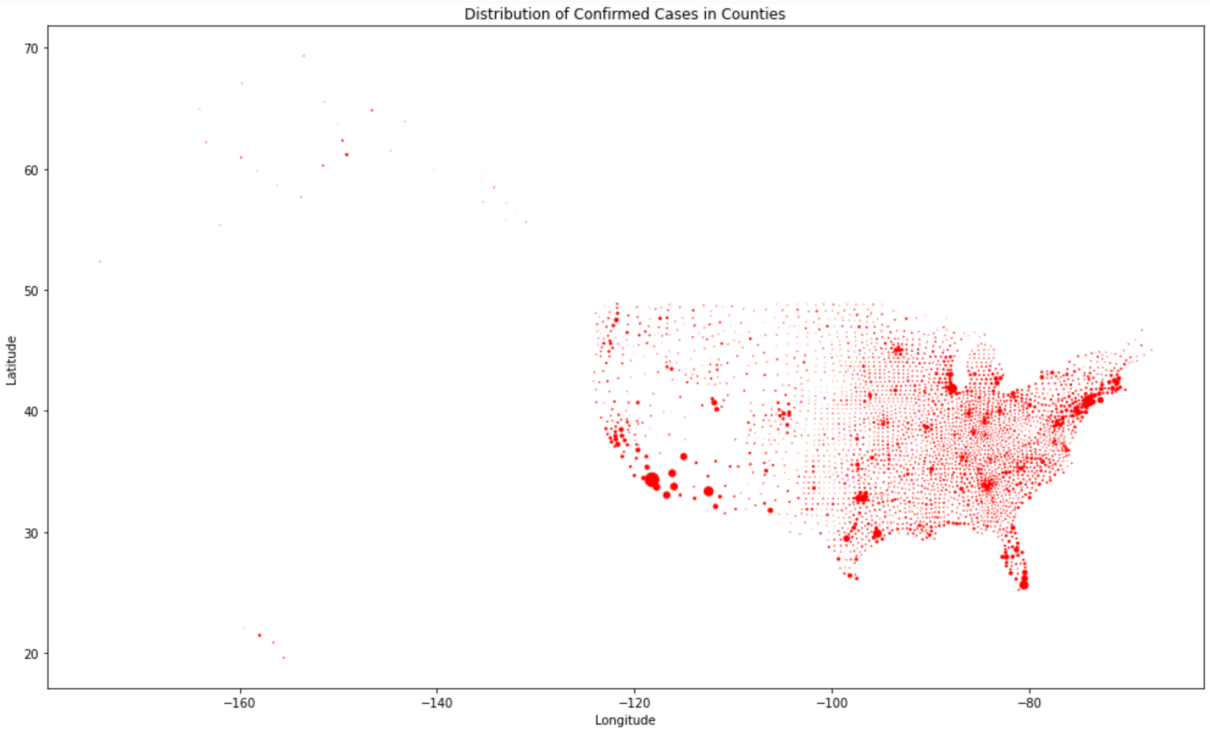
Through the relationship between county population and confirmed cases, it can be seen that COVID-19 cases grow exponentially based on the population. In addition, many counties have a similar number of confirmed cases.



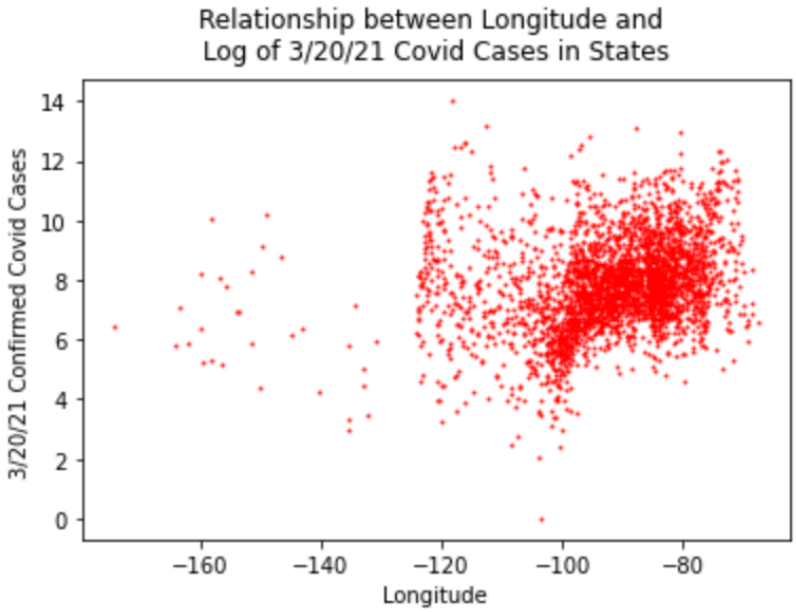
In addition to plotting COVID-19 cases, deaths, and insurance, we also plotted the counties based on their longitude and latitude.

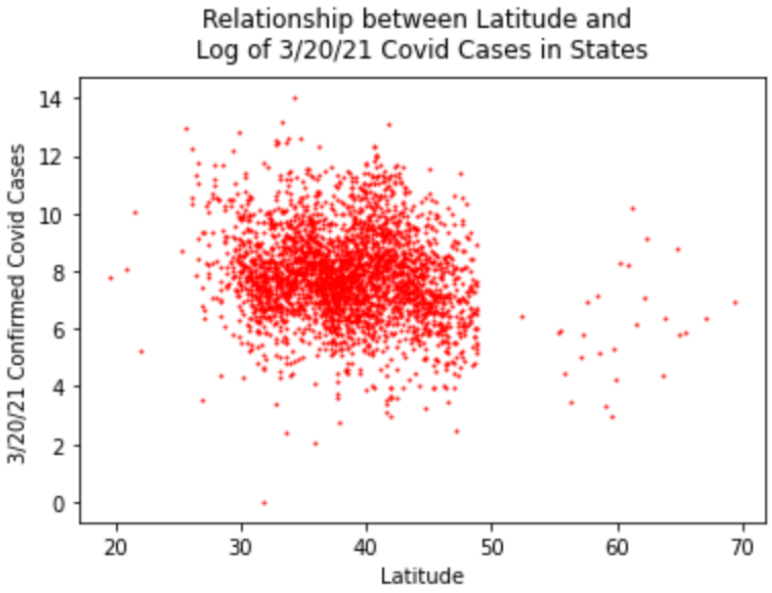


To combine the COVID-19 and insurance data with the location of the counties, we changed the size of the county points based on the number of confirmed cases in the county.



We plotted the number of total confirmed cases in US counties and the longitude of each county. These scatter plots indicate that there is a slight increase in confirmed cases in more densely packed counties which are packed in higher longitude and lower latitude parts of the country, or northern and eastern parts of the country.





Data Analysis:

No relationship between Health Insurance and Confirmed/Deaths

According to our statistical analysis, there are no significant relationships between insured population and confirmed cases/deaths. For March 20, 2021, our linear regression models for percent insured county population and percent county confirmed/deaths have near constant slopes of -1.3 and -0.6 and small R-squared values of 0.0380 and 0.0625 - respectively for confirmed cases and death numbers. We also obtained data from last March 2020. Likewise, the linear regression models had constant 0.0 slopes and small R-squared values that were both less than 0.01.

We can conclude that there is no significant correlation between the insured population (independent variable) and confirmed cases/death rates (dependent variables) based on our regression models and R-squared values < 0.1. The slopes for each model also indicate that the relationship between our variables is constant.

Our findings disprove our hypothesis that there would be less confirmed cases and death rates in an increasingly insured population. We discovered that there is actually no direct relationship between health insurance and COVID-19 cases. An explanation as to why there is no relationship between health insurance and COVID-19 cases may be that insurance is not required for an individual to be tested in the U.S. Furthermore, many people became unemployed during the pandemic and therefore lost their health insurance. The amount of insured population may have decreased during the past year even though cases have continued to rise.

Exponential Relationship between Population and Confirmed Cases

Our two maps and the plot between longitude and cases revealed that there is an underlying relationship between population count and confirmed cases, and that the eastern half of the U.S. contains the most amount of counties. Therefore, we visualized the relationship between county population density and total number of confirmed cases. According to the graph visual, the number of confirmed county cases grew exponentially with the population, and the number of cases for most counties lie in the range 0 - 200,000. An explanation as to why the relationship between population and COVID-19 cases is exponential may be that infectious diseases spread faster in crowded cities because the rate of transmission is high, and overcrowding exacerbates sanitation issues.

Challenges

Refining project scope

One of the challenges we faced along the way was making changes and adding limits to the overall scope of our project itself. Our initial thoughts for this project was to try looking at the relationship between health insurance percentages and vaccine rates, in addition to the various COVID-19 rates.

Unable to investigate vaccine data

However, the vaccine process was at a very early stage at the time of data collection, and therefore it was difficult to find data in relation to the vaccine processes. Being early on in the vaccine process also meant that the process was only at the beginning phases. From that knowledge, we were able to conclude that the majority of people being vaccinated are mainly healthcare workers and first responders, which would make it difficult to gain any clear and interesting observations for the relationship between the vaccine process and health insurance at the time.

Limitations

Influence of outside factors

A limitation that we also had to keep in mind was that there could have been various outside factors that affect the datasets themselves or become the cause of the relationships we discovered. Some outside features include population differences, testing frequencies, and asymptomatic people. Aside from health insurances, There could have also been other factors that influenced COVID-19 rates such as age demographics, lock down periods, or climate differences. Although health insurance coverage is something that reflects such determinants, it does broaden out the perspectives to a very expansive state. Thus we could have found additional correlation relationships between specific factors of health insurance and various COVID-19 rates.

Outdated health insurance data

The most recent data we could find for county health insurance was from 2018. Because of this, we initially wanted to try adding predictions to the original data as the COVID-19 data we were intending to use was for March 2021. However, we were unable to get the relevant predictions in time and instead used the most recently updated dataset for our plots and analysis.

Next Steps

If given more time, we would like to investigate the effects of other social determinants of health on COVID-19 cases. Other interesting determinants include population demographics, COVID-19 waves, lock down periods, and climate differences. We may also dive deeper into population statistics since our analysis revealed that the number of confirmed cases increase exponentially with county population density. We hope that our findings contribute to a better understanding on how social determinants affect population health and pandemic outcomes.