Effects of Social Determinants of Health on Covid Infection and Death Rates

Temitope Oshinowo
SML 310
11/28/2020

I. Overview

Intro:

Since its classification as a pandemic in March 2020, the Covid-19 pandemic has dramatically changed the way we live, work, and learn. For this reason, it is important to explore as much as we can regarding the topic, especially as it relates to the rates of cases and deaths. Such an understanding is critical as it will allow healthcare workers and policymakers to create better interventions in mitigating the effects of the pandemic.

One topic which has been heavily studied is the effects of preexisting conditions on Covid severity (in terms of symptoms) and outcomes (death rate). Such research by the CDC has found that conditions such as diabetes, obesity, and high blood pressure can worsen Covid outcomes. These findings are extremely valuable as they promote patient education and suggest that people should be aware of their risk factors. Nevertheless, it is important to explore other avenues which may be affecting Covid prevalence and death rates.

In addition to highlighting the negative effect of preexisting conditions, this pandemic has illustrated how social determinants of health affect Covid outcomes. In this project, I defined social determinants of health as the economic and social conditions which affect a person's living environment and health status. Social determinants are often the result of human interference or achievement; examples of these include residential environment, occupation, or income levels. With these in mind, I will use these social determinants to study Covid prevalence and outcomes.

Main Question:

From this context, the question arises: What, if any, are the socially deterministic drivers of Covid prevalence and mortality? How do these drivers affect Covid infection and death rates?

Results:

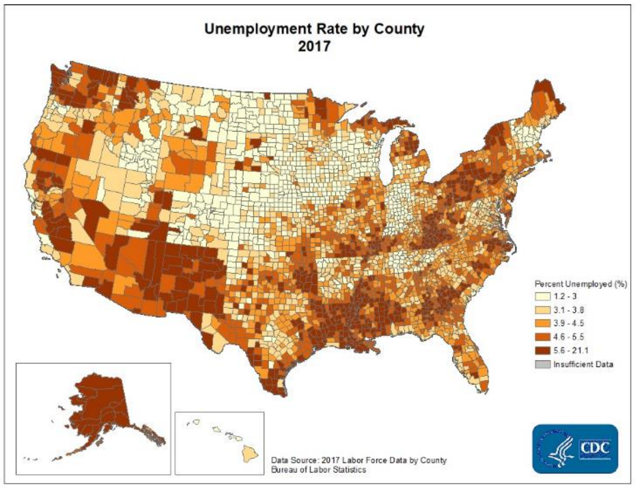
This study utilized a decision tree classifier and random forest model to better understand Covid trends as they relate to the social determinants studied. Both classifiers illustrated that of the social determinants studied, education level and health insurance were the most important features in understanding the data.

When looking at Covid cases and deaths, it is evident that these features are loosely correlated with outcome: counties with a higher proportion of people without a high school diploma are more likely to have higher Covid cases and deaths, and the same holds true for counties with a higher proportion of people without health insurance. However, analysis of the most advantaged and disadvantaged counties by education and insurance rates found that this disparity was not statistically significant. Despite this, analysis also found that there is more variability in outcomes for disadvantaged counties compared to more advantaged counties. From this, it is evident that while social deterministic factors may not directly be statistically responsible for increasing Covid cases or death rates (as compared to more biological data), there are minor relationships between these features and outcome. While Covid effects due to such social determinants may not play out on the national level with all 3000+ counties, the effects of such determinants are more present in disadvantaged communities.

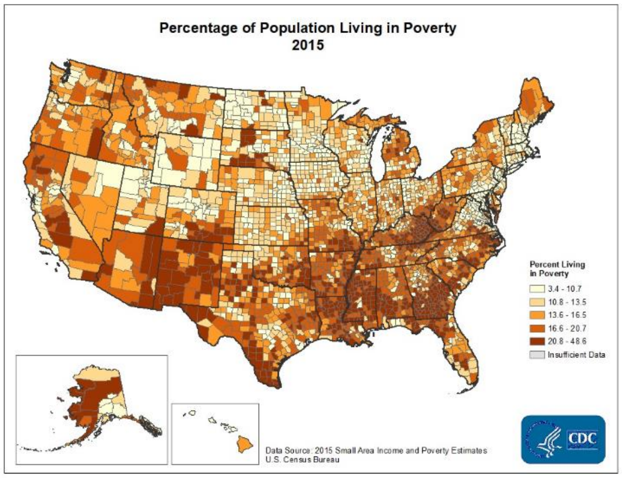
II. Related Work

Before undertaking my project, I first conducted a search to discover projects which have elements like what I wanted to do. This search led me to three key sources.

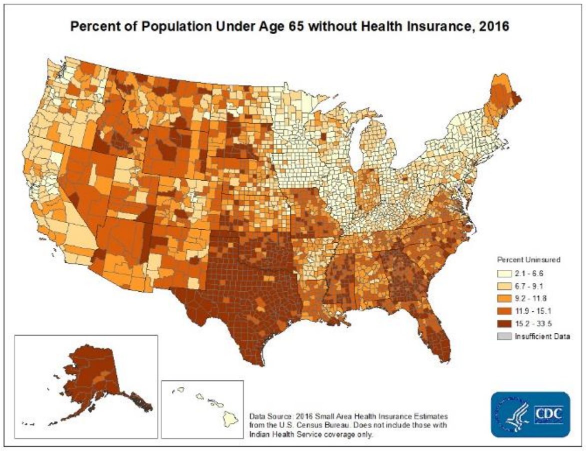
Because I was interested in studying Covid disparities across counties, I was interested in finding projects that also utilized county level data. This search led me to a project which studied various rates of occurrences per county and visualized them using county-level FIPS (Federal Information Processing Standards) codes. While my project examines Covid cases (per 1,000) and deaths (per 1,000) per county, and their relationship to social determinants, this Plotly project examines things such as: percent change in alcohol use disorders, diabetes prevalence, and rates of interpersonal violence— all per county. Observing this project was valuable as it allowed me to better understand how to view and interpret county level data.



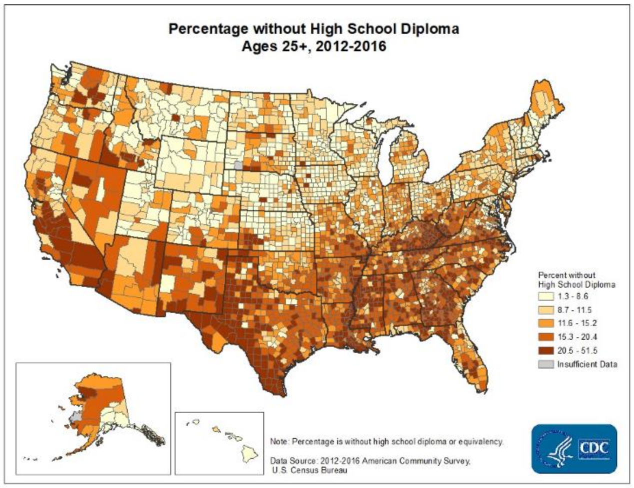
Data Source: 2017 Labor Force Data by County
Bureau of Labor Statistics



Data Source: 2015 Small Area Income and Poverty Estimates
U.S. Census Bureau



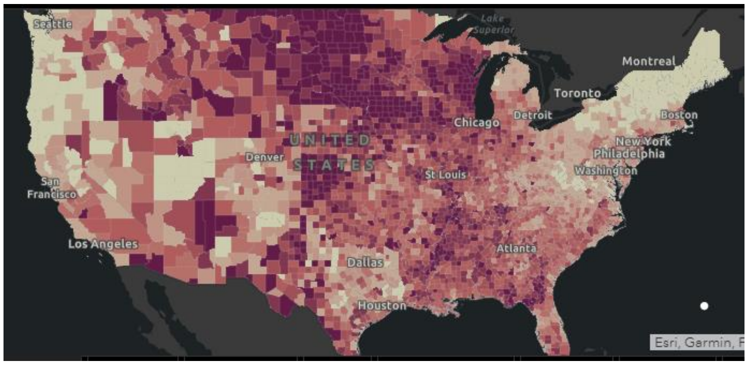
Data Source: 2016 Small Area Health Insurance Estimates from the U.S. Census Bureau. Does not include those with Indian Health Service coverage only.



Note: Percentage is without high school diploma or equivalency.

Data Source: 2012-2016 American Community Survey,
U.S. Census Bureau

In looking at the social determinants of health, the CDC compiled information on such determinants on the county level (shown above). These heatmaps will be helpful when comparing visually if there are any relationships between determinant prevalence and cases and deaths prevalence. From the heatmaps, we can already visually see that there appears to be a relationship between unemployment rate and % living in poverty. In my analysis, take a more quantitative approach in looking at such relationships.



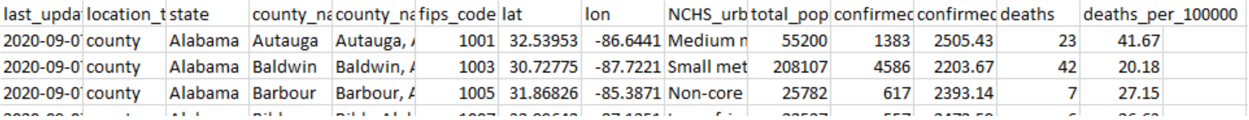
Another work related to this project specifically examines Covid cases on a county level. Johns Hopkins University's (JHU) COVID-19 Dashboard is updated once a day, and includes information on confirmed cases, number of deaths, and fatality rates (among other variables) by county (for the United States) and per country (worldwide). This work is relevant because it contains some of the parameters my model will also utilize (such as number of deaths or total population). While JHU's heatmap observes absolute cases per county, I will observe cases per 1,000 in population (to account for the fact that different counties have differences in their population and compare in a more consistent manner across counties). I will also expand on this work by observing how Covid cases and deaths (per 1,000 in population) relate to some of the social determinants in the CDC dataset .

III. Relevant Data

OSEM: Obtain Data

The first step in the OSEM data science life cycle is to obtain the data. Because I am interested in observing the relationship between socioeconomic factors and Covid cases and deaths, I obtained data related to these criteria. I pulled my data from six key sources:

1. Johns Hopkins Coronavirus Case Tracker (2020)

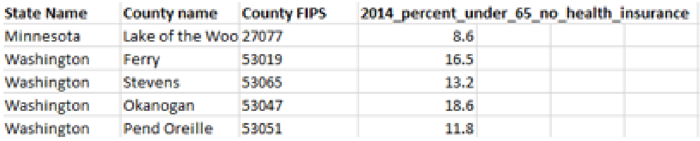


last_upda	location_t	state	county_na	county_na	fips_code	lat	lon	NCHS_urb	total_pop	confirmec	confirmec	deaths	deaths_per_100000
2020-09-0	county	Alabama	Autauga	Autauga, /	1001	32.53953	-86.6441	Medium n	55200	1383	2505.43	23	41.67
2020-09-0	county	Alabama	Baldwin	Baldwin, /	1003	30.72775	-87.7221	Small met	208107	4586	2203.67	42	20.18
2020-09-0	county	Alabama	Barbour	Barbour, /	1005	31.86826	-85.3871	Non-core	25782	617	2393.14	7	27.15

As described above, the JHU Coronavirus Case Tracker contains county-level data of Covid incidence and death rate, from March 2020 to September 7, 2020. This data is appropriate for my project because it includes population information-based counties. From this set I primarily utilized the confirmed cases per 100,000 (converting them to 1,000), and deaths per 100,000 (converting them to 1,000) as tools to build/assess my model (predicting the likelihood that someone will die from Covid and assessing this model and its classifications with the actual outcomes).

In terms of data processing, this data is already clean. Because it included the values, I would be modeling based on (confirmed cases and deaths) as well as the FIPS codes per county, I considered this as my main dataset and after cleaning other sets, merged them into this one (as those may have extra values or may be missing values from this set)

2. CDC Social Determinants of Health (2014-2018)



State Name	County name	County FIPS	2014_percent_under_65_no_health_insurance
Minnesota	Lake of the Woo	27077	8.6
Washington	Ferry	53019	16.5
Washington	Stevens	53065	13.2
Washington	Okanogan	53047	18.6
Washington	Pend Oreille	53051	11.8

As described above, the CDC Social Determinants of Health datasets include information State and county names, FIPS codes, and the statistics for each county. These statistics include the Percent of People without a high school diploma, percent of people without health insurance, percent of people living in poverty, and unemployment rate per county. I hypothesize that as each of these percentages increase, the number of Covid deaths per 1000 will also increase, as research by the CDC has shown that such factors negatively impact other health outcomes (such as heart health). In terms of cleaning this data, the data itself is already clean. I will then merge each of the files with the main dataset using FIPS code.

3. Per-Capita Income by County (2018)

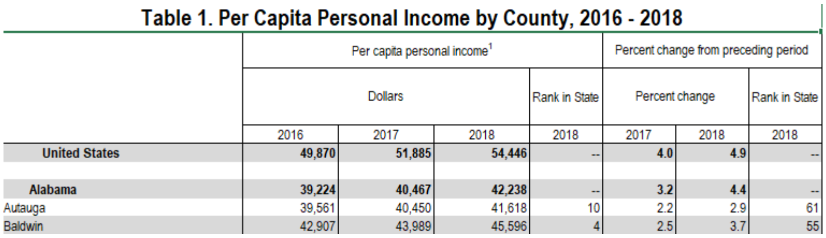


Table 1. Per Capita Personal Income by County, 2016 - 2018

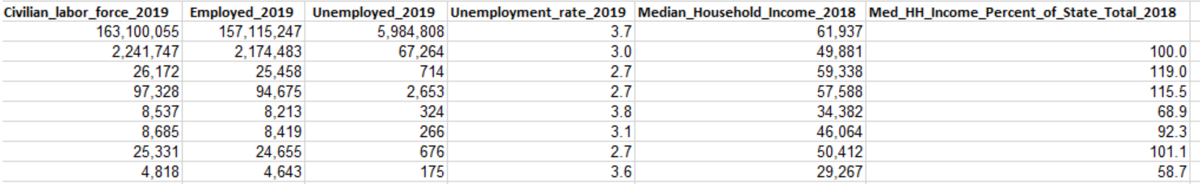
	Per capita personal income[1]				Percent change from preceding period		
	Dollars			Rank in State	Percent change		Rank in State
	2016	2017	2018	2018	2017	2018	2018
United States	**49,870**	**51,885**	**54,446**	--	**4.0**	**4.9**	--
Alabama	**39,224**	**40,467**	**42,238**	--	**3.2**	**4.4**	--
Autauga	39,561	40,450	41,618	10	2.2	2.9	61
Baldwin	42,907	43,989	45,596	4	2.5	3.7	55

This data by the Bureau of Economic Analysis contains information on the per capita personal income per county. This information is valuable because I plan to use average income level as one of the predictors of Covid mortality. I predicted that counties with a lower average income will have a greater rate of Covid deaths.

While the most recent data on the county level is from 2018, I assessed the information to still be recent. Because the income changes from 2016-2018 are minimal (the percent changes are all < 0.10), the change from 2018-2020 should be similar. I believe that higher income counties from 2018 will still have higher levels of income in 2020 (with lower income counties from 2018 still having lower levels of income in 2020), making this dataset still relevant to my work.

For the most part, this data was already clean, however, I formatted headings so that they could easily be merged into   (especially when joining this dataset to the Johns Hopkins Dataset by FIPS code.

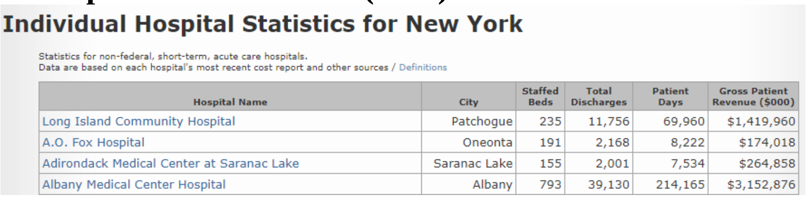
4. Median Household Income (2018)



Civilian_labor_force_2019	Employed_2019	Unemployed_2019	Unemployment_rate_2019	Median_Household_Income_2018	Med_HH_Income_Percent_of_State_Total_2018
163,100,055	157,115,247	5,984,808	3.7	61,937	
2,241,747	2,174,483	67,264	3.0	49,881	100.0
26,172	25,458	714	2.7	59,338	119.0
97,328	94,675	2,653	2.7	57,588	115.5
8,537	8,213	324	3.8	34,382	68.9
8,685	8,419	266	3.1	46,064	92.3
25,331	24,655	676	2.7	50,412	101.1
4,818	4,643	175	3.6	29,267	58.7

In addition to observing per-capita income, I will observe median household income, as per-capita income may be affected by outlier towns within a given county. I pulled this information from the United States Department of Agriculture Economic Research Service. I am particularly interested in median household income for 2018, so I will filter out values from other years and merge this with the main data using FIPS codes.

5. Hospital Beds Per State (2018)



Individual Hospital Statistics for New York

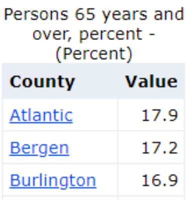
Statistics for non-federal, short-term, acute care hospitals.
Data are based on each hospital's most recent cost report and other sources / Definitions

Hospital Name	City	Staffed Beds	Total Discharges	Patient Days	Gross Patient Revenue ($000)
Long Island Community Hospital	Patchogue	235	11,756	69,960	$1,419,960
A.O. Fox Hospital	Oneonta	191	2,168	8,222	$174,018
Adirondack Medical Center at Saranac Lake	Saranac Lake	155	2,001	7,534	$264,858
Albany Medical Center Hospital	Albany	793	39,130	214,165	$3,152,876

This data by the American Hospital Association contains information on the hospital name, city location, number of beds in the hospital, and gross patient revenue per hospital. All this information is compiled on a state-by-state level. This information is valuable because I used county hospital bed count per 1,000 as one of the predictors of Covid mortality. I predicted that areas with a lower hospital bed count (as a function of the population) would have a greater rate of Covid deaths.

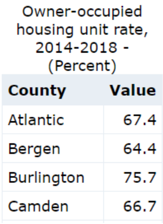
Although this data is from 2018, I suspect that the changes from 2018 to 2019 should be minimal (2019 reports were delayed due to the Covid-19 pandemic). In terms of the OSEMN process, this data required much scrubbing. Since the information was presented on a state-by-state level, in Excel I merged the tables per state into one table for the United States as a whole. Since the information on hospitals is presented by city, I looked up the corresponding FIPS codes per city and then summed for the number of beds per unique FIPS code. I then joined this dataset (FIPS codes and bed count) to the Johns Hopkins dataset using the FIPS code. During my analysis, this was helpful, as I could then compute statistics such as the number of beds per 1000 people per county.

6. Percentage of Population Over 65, % of Homeowners, % Female by County (2018)



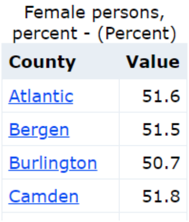
Persons 65 years and over, percent - (Percent)

County	Value
Atlantic	17.9
Bergen	17.2
Burlington	16.9



Owner-occupied housing unit rate, 2014-2018 - (Percent)

County	Value
Atlantic	67.4
Bergen	64.4
Burlington	75.7
Camden	66.7



Female persons, percent - (Percent)

County	Value
Atlantic	51.6
Bergen	51.5
Burlington	50.7
Camden	51.8

This data from the U.S. Census Bureau, Population Estimates Program contains information on the county name and percentage of persons 65 years old or above. The same holds true for Owner-Occupied Housing Unit Rate (percent of homeowners living in their house) and percent of female persons. All this information is compiled on a state-by-state level. This information is valuable because although age and percent female are biological (rather than socioeconomic) markers, it will be interesting to observe how the model assesses biological vs. socioeconomic features. Based on what science has reported about Covid, I predict that counties with a greater percentage of persons 65 or above will have a greater rate of Covid deaths. Likewise, I predict that counties with a greater percentage of females (less males) will have lower rates of Covid deaths. Finally, I predict that areas with a lower owner-occupied housing rate will have higher rates of Covid cases and deaths (as lower home ownership suggests cities, with housing alternatives such as apartments, clustering people together and promoting infection transmission.

Based on the OSEMN process, this data will require some scrubbing. Because the information is presented on a state-by-state level, I will compile the tables per state into one table for the United States as a whole. I will then join this dataset to the Johns Hopkins dataset using state and county names (as there are no FIPS codes here).

OSEM: Scrub Data

After obtaining my data, and scrubbing the necessary files as described. I then put the cleaned excel sheets into one file. The excel file "main" consisted of the five key sheets which considered in my analysis.

covid_income_sheet: Information on county name, state name, county FIPS code, county type (urban, suburban, rural, etc.), total population, number of confirmed cases, number of deaths, deaths per 100,000, per-capita income information, % over 65 years old, % owner occupied home, and % female (all last updated in 2018, except for Covid incidence and Covid deaths, which are from 2020)

beds_sheet: number of beds across all hospitals per county (last updated 2018, 2019 information delayed due to pandemic)

edu_health_sheet: % of people over age 25 with no high school diploma (per county, last updated in 2016)

unemployment_sheet: unemployment rate per county (last updated 2017)

poverty_sheet: % of population living under the poverty line (per county, last updated 2015)

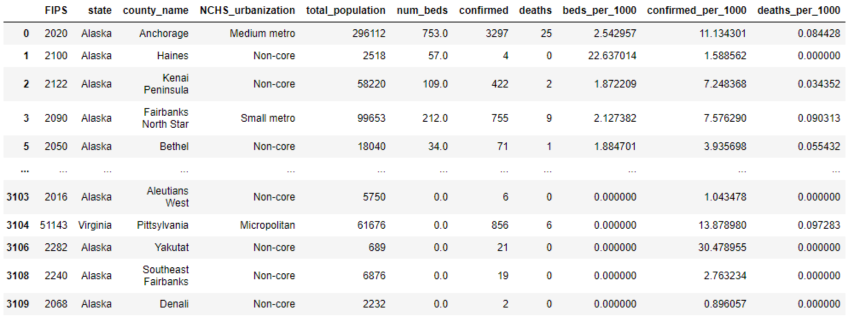
median_income_sheet: median household income (per county, last updated 2018)

When observing the rates of change for certain data pieces (such as per-capita income) it was evident that values do not change much across a few years. Therefore, I will proceed with the assumption that the data from the 2014-2018 surveys are relevant to 2020.

After importing the sheets, I read them into a pandas dataframe and proceeded to convert NaN values for beds to 0 (as those counties did not have beds, so the bed count is zero). Using the number of cases, and total population columns, I then calculated the cases per 1000 using the formula: cases per 1000 = total cases in a countytotal population in a county 1,000. I also performed similar calculations for beds per 1000 and deaths per 1000. I then added these three columns per county to my dataframe.

Because my classification models predicted the quartiles of cases_per_1000 and deaths_per_1000 as discrete variables, I added 2 new columns with the actual quartiles for each variable. These columns essentially illustrate the county's quartile for case rate and death rate.

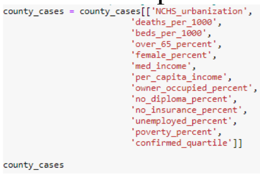
I then renamed and reorganized some columns for clarity purposes before filtering out rows with NaN values, creating a complete dataset. This operation resulted in the number of counties reducing from 3114 to 3105, so overall, most of the data was represented across all the sheets used. I then saved my resulting dataset as a csv for future reference. Below is a snip of the final main dataframe:



	FIPS	state	county_name	NCHS_urbanization	total_population	num_beds	confirmed	deaths	beds_per_1000	confirmed_per_1000	deaths_per_1000
0	2020	Alaska	Anchorage	Medium metro	296112	753.0	3297	25	2.542957	11.134301	0.084428
1	2100	Alaska	Haines	Non-core	2518	57.0	4	0	22.637014	1.588562	0.000000
2	2122	Alaska	Kenai Peninsula	Non-core	58220	109.0	422	2	1.872209	7.248368	0.034352
3	2090	Alaska	Fairbanks North Star	Small metro	99653	212.0	755	9	2.127382	7.576290	0.090313
5	2050	Alaska	Bethel	Non-core	18040	34.0	71	1	1.884701	3.935698	0.055432
...	...	...	...	...	...	...	...	...	...	...	...
3103	2016	Alaska	Aleutians West	Non-core	5750	0.0	6	0	0.000000	1.043478	0.000000
3104	51143	Virginia	Pittsylvania	Micropolitan	61676	0.0	856	6	0.000000	13.878980	0.097283
3106	2282	Alaska	Yakutat	Non-core	689	0.0	21	0	0.000000	30.478955	0.000000
3108	2240	Alaska	Southeast Fairbanks	Non-core	6876	0.0	19	0	0.000000	2.763234	0.000000
3109	2068	Alaska	Denali	Non-core	2232	0.0	2	0	0.000000	0.896057	0.000000

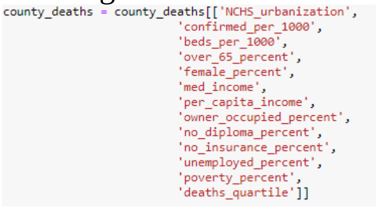
Because I wanted to analyze both cases and deaths per 1000, I split up my analysis into 2 parts for the rest of this project. I first made two copies of my main dataframe, filtered for desired model features, and performed a 70-30 test train split. The training sets' X values consisted of 2173 counties and 12 features, while test sets' X values had 932 counties and 13 features. In examining cases, I predicted the quartiles for cases per 1000, while in examining deaths I predicted the quartiles for deaths per 1000.

In my cases_quartiles training and test sets I had these variables (with 'confirmed_quartile' being the value I was predicting):



```
county_cases = county_cases[['NCHS_urbanization',
                             'deaths_per_1000',
                             'beds_per_1000',
                             'over_65_percent',
                             'female_percent',
                             'med_income',
                             'per_capita_income',
                             'owner_occupied_percent',
                             'no_diploma_percent',
                             'no_insurance_percent',
                             'unemployed_percent',
                             'poverty_percent',
                             'confirmed_quartile']]

county_cases
```

While in my death_quartiles training and test sets I had these variables (with 'deaths_quartile' being the value I was predicting):



```
county_deaths = county_deaths[['NCHS_urbanization',
                               'confirmed_per_1000',
                               'beds_per_1000',
                               'over_65_percent',
                               'female_percent',
                               'med_income',
                               'per_capita_income',
                               'owner_occupied_percent',
                               'no_diploma_percent',
                               'no_insurance_percent',
                               'unemployed_percent',
                               'poverty_percent',
                               'deaths_quartile']]
```

OSEM: Explore Data

Before running the models and making these predictions however, I conducted some exploratory data analysis. For visualization purposes, I focused on observing confirmed cases per 1000 and deaths per 1000 rather than the quartiles, as such graphics provided more meaningful information (quartiles were more valuable for classification purposes and confirmed cases per 1000 and deaths per 1000 were not included in the actual models, as that would cause overfitting). I will first discuss my findings when looking at confirmed cases, and then discuss findings when observing deaths.

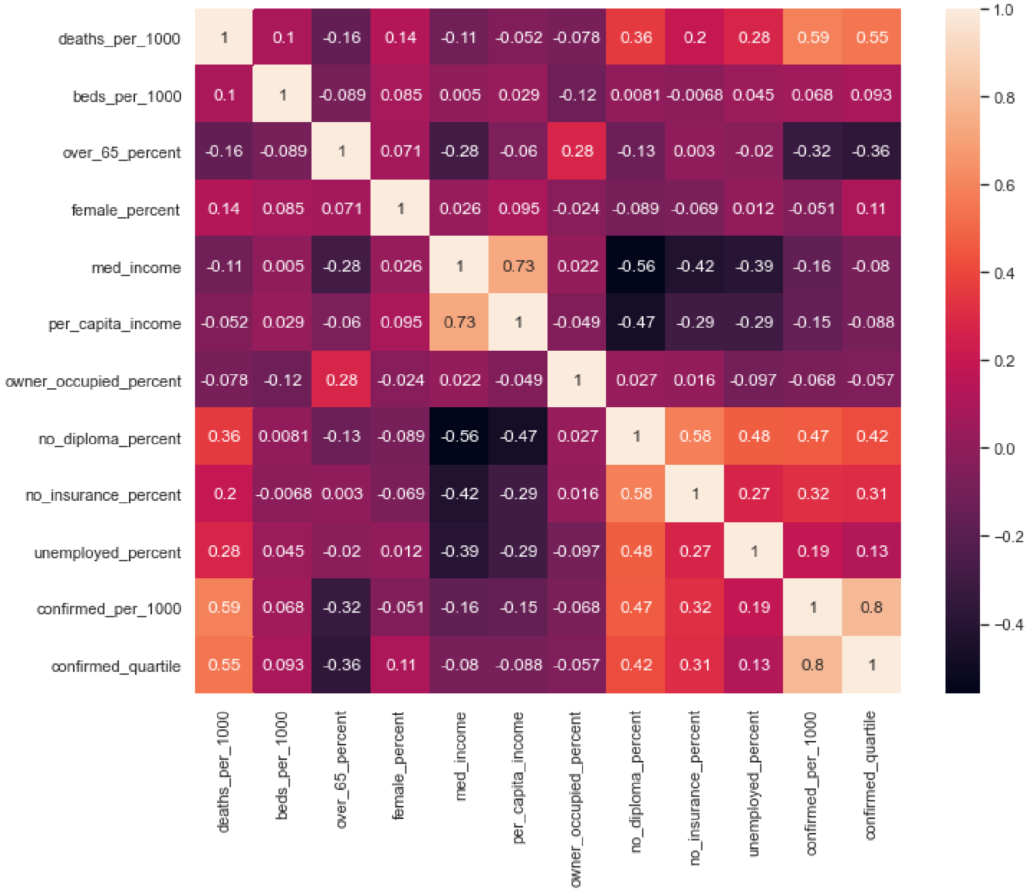
Confirmed Cases:

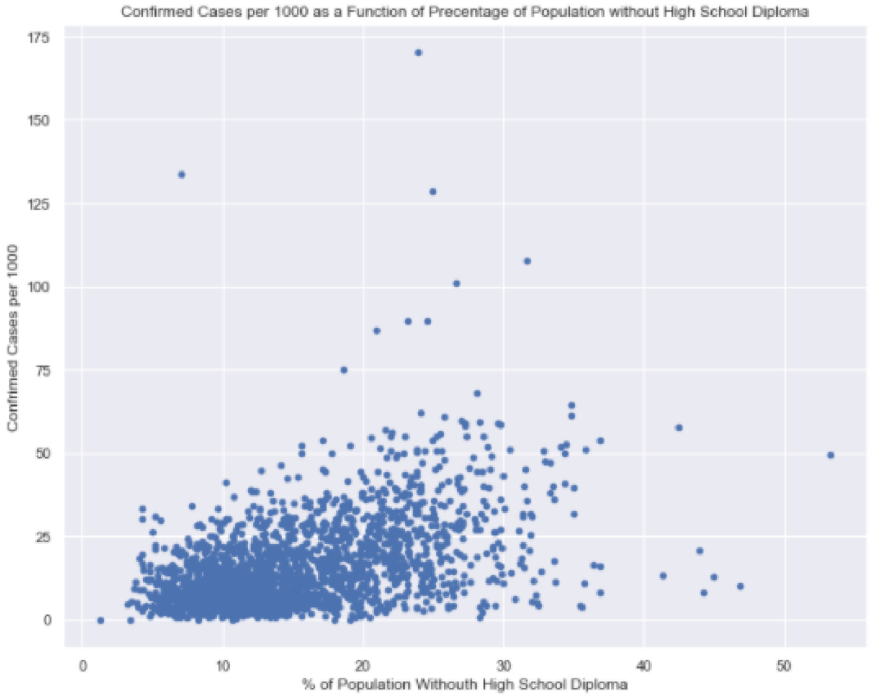
I first used seaborn to generate the default pair plot (see Supplementary Figure 1a. FigS1a) and visually note any interesting trends. Then I looked more specifically at the row with trends related to confirmed cases per 1000 (Fig. S2a). From these graphs, I noticed that in general:

- There appears to be a positive correlation between Covid cases per 1000 and % without a high school diploma, % lacking health insurance, having more people under the poverty level, and having a higher unemployment rate.
- There appears to be a negative correlation between income (both median household and per-capita) and education level, and a negative correlation between income and Covid incidence (lower income counties seem to have higher incidences of Covid).
- Finally, there appears to be a negative correlation between the percentage of people over age 65 and Covid incidence (counties with more of an elderly population seem to have lower incidences of Covid). This was surprising to me, as I would expect the opposite. However, since we are looking at confirmed cases (and not all cases) that may explain this discrepancy.

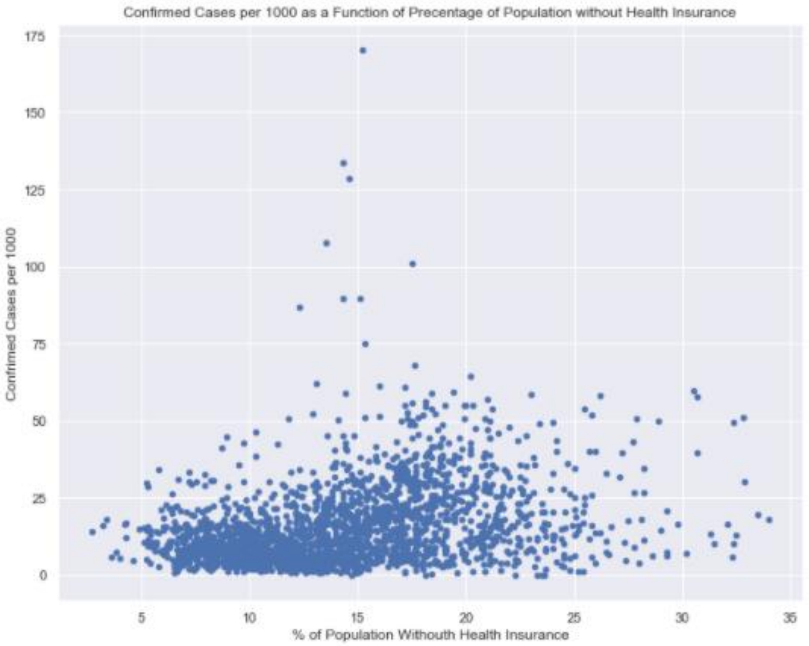
To further investigate my visual hypotheses, I computed the correlation matrix and then looked specifically at features which stood out to me.

Fig 1: Correlation Coefficients When Analyzing Confirmed Cases

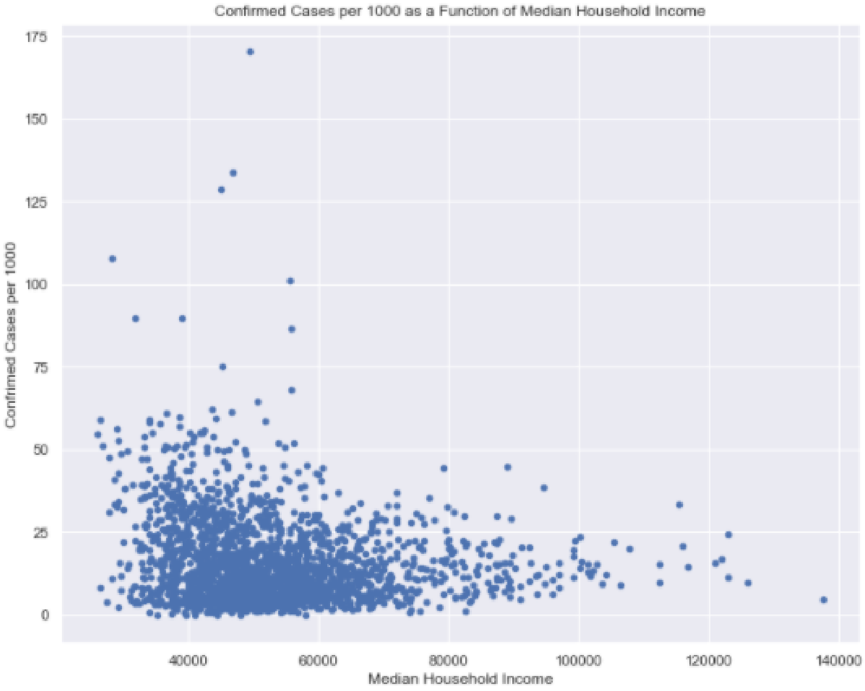




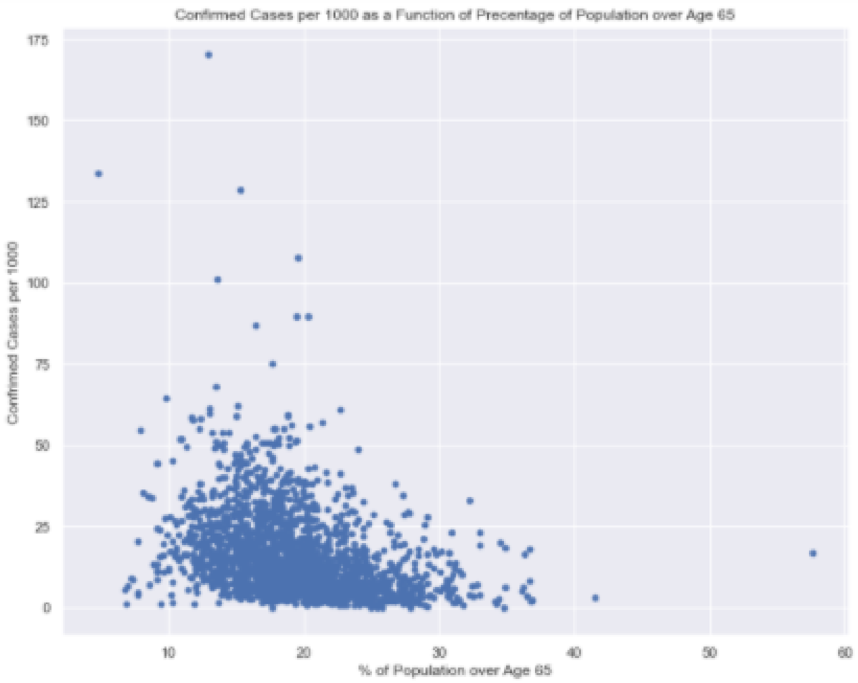
With a correlation coefficient of 0.47, this is intersting, as it demonstrates that across counties in the US, there is a moderate positive correlation between % without a high school diploma and Covid rates per 1000.



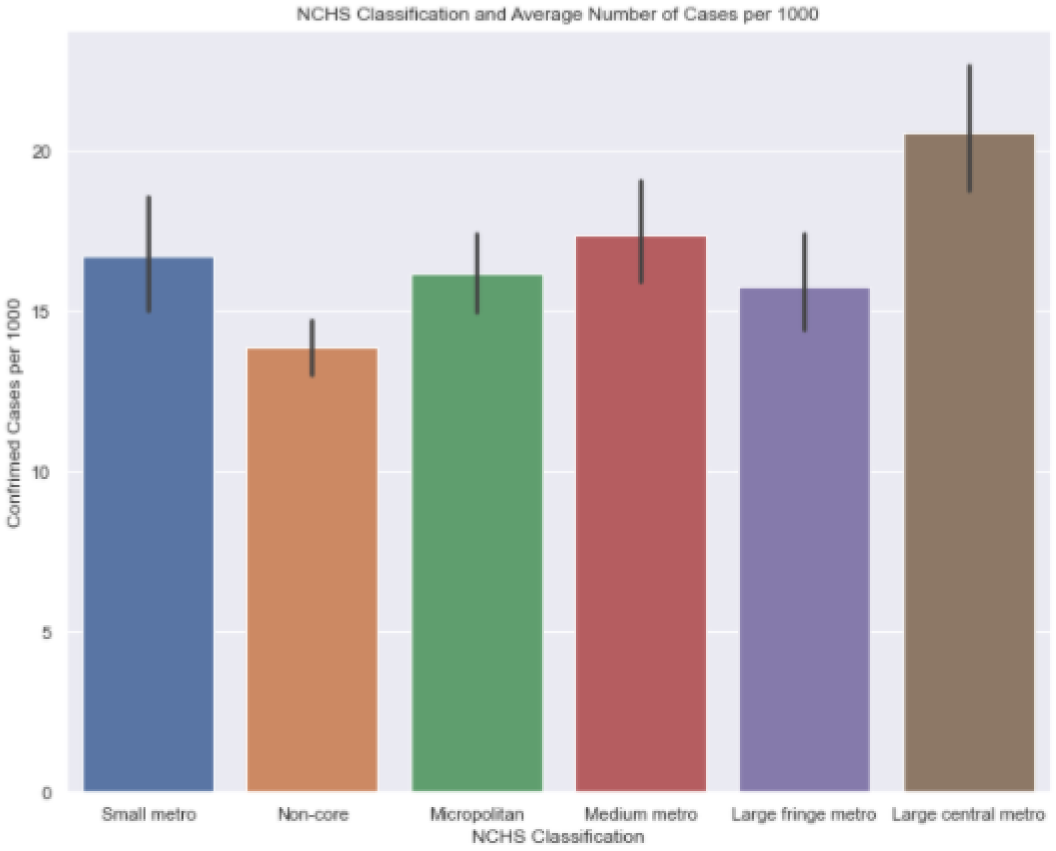
With a correlation coefficient of 0.32 this demonstrates that across counties in the US, there is a slight positive correlation between % lacking health insurance and Covid rates per 1000.



With a correlation coefficient of -0.16 this demonstrates that across counties in the US, there is a minor negative correlation between median household income and Covid rates per 1000.



With a correlation coefficient of -0.32 this demonstrates that across counties in the US, there is a slight negative correlation between having an older population and Covid rates per 1000. This was surprising to me, as I expected older population to have hihger rates of confirmed cases. However, the data shows that this is not necessarily the case.



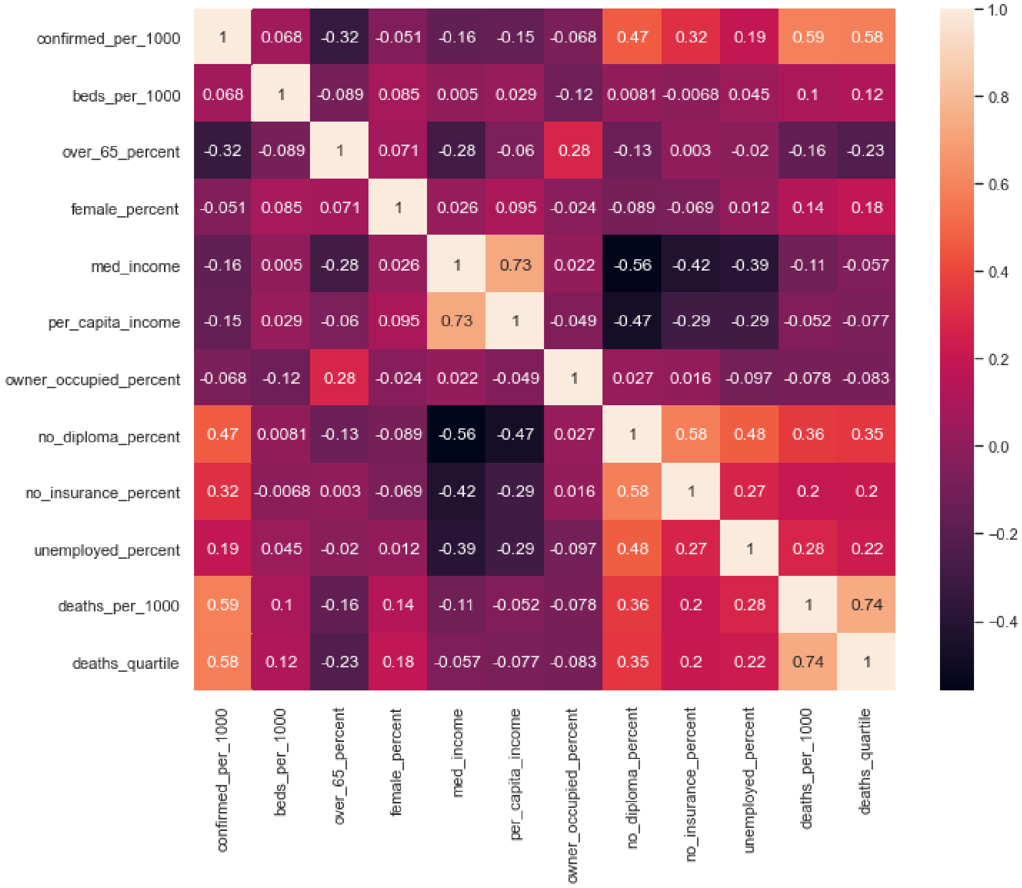
Large central metro regions have significantly more Covid cases compared to non-core rural areas.

Then I Look at Deaths:

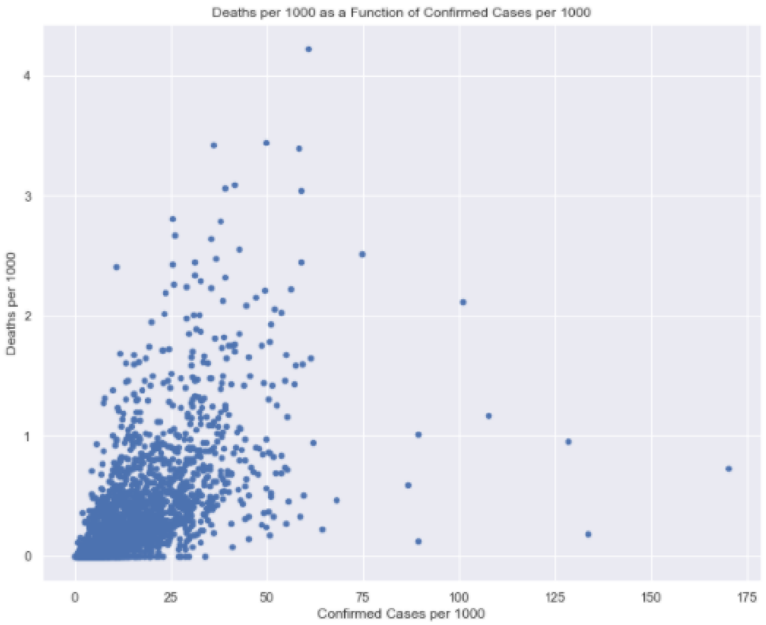
Based on the default pair plot (FigS2a) and the row with trends related to deaths per 1000, one can see that:

- Deaths per 1000 appears to have a strong positive correlation with confirmed per 1000. This makes sense since one must have the infection before dying of it.
- There appears to be a positive correlation between Covid deaths per 1000 and education levels. There also appears to be a correlation between Covid deaths per 1000 and lacking health insurance. Finally, the data also suggests a positive correlation between the percentage of females in a county and the Covid deaths per 1000. However, this may be because of the few outliers of places that are abundantly male (and happen to have low Covid death rates).
- There appears to be a negative correlation between income (both median household and per-capita) and Covid death (lower income counties seem to have higher rates of Covid death). There may be a minor negative correlation between beds per 1000 and Covid death rate (more beds per 1000 is associated with more Covid deaths per 1000).

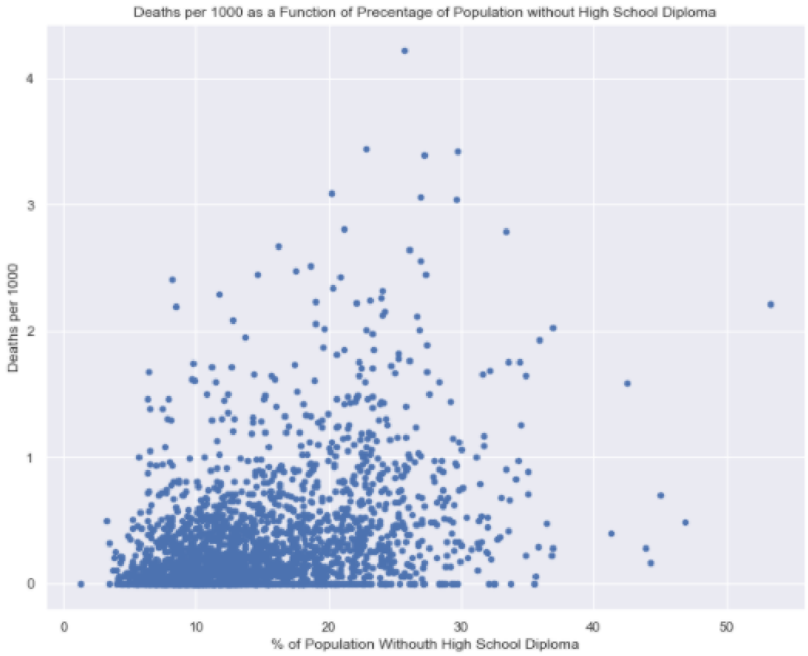
Fig 2: Correlation Coefficients When Analyzing Deaths



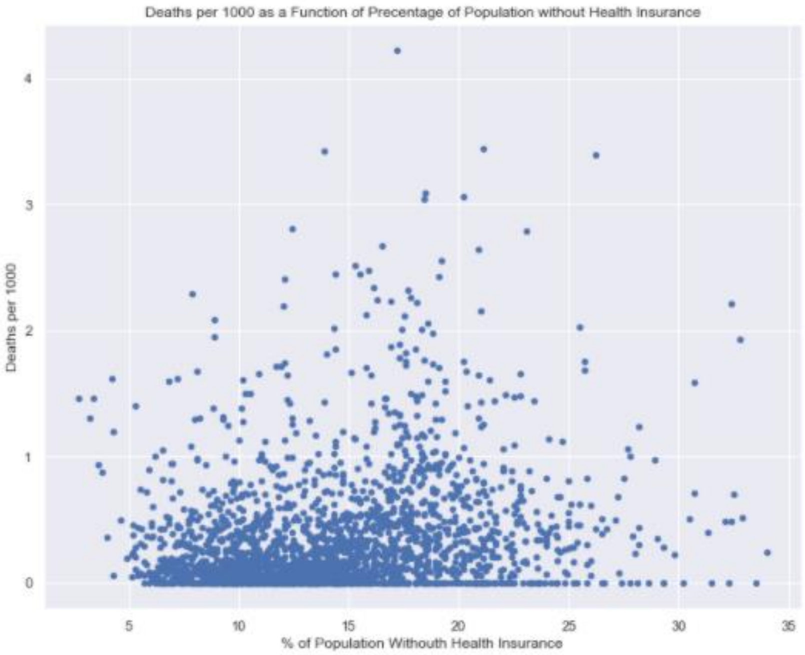
	confirmed_per_1000	beds_per_1000	over_65_percent	female_percent	med_income	per_capita_income	owner_occupied_percent	no_diploma_percent	no_insurance_percent	unemployed_percent	deaths_per_1000	deaths_quartile
confirmed_per_1000	1	0.068	-0.32	-0.051	-0.16	-0.15	-0.068	0.47	0.32	0.19	0.59	0.58
beds_per_1000	0.068	1	-0.089	0.085	0.005	0.029	-0.12	0.0081	-0.0068	0.045	0.1	0.12
over_65_percent	-0.32	-0.089	1	0.071	-0.28	-0.06	0.28	-0.13	0.003	-0.02	-0.16	-0.23
female_percent	-0.051	0.085	0.071	1	0.026	0.095	-0.024	-0.089	-0.069	0.012	0.14	0.18
med_income	-0.16	0.005	-0.28	0.026	1	0.73	0.022	-0.56	-0.42	-0.39	-0.11	-0.057
per_capita_income	-0.15	0.029	-0.06	0.095	0.73	1	-0.049	-0.47	-0.29	-0.29	-0.052	-0.077
owner_occupied_percent	-0.068	-0.12	0.28	-0.024	0.022	-0.049	1	0.027	0.016	-0.097	-0.078	-0.083
no_diploma_percent	0.47	0.0081	-0.13	-0.089	-0.56	-0.47	0.027	1	0.58	0.48	0.36	0.35
no_insurance_percent	0.32	-0.0068	0.003	-0.069	-0.42	-0.29	0.016	0.58	1	0.27	0.2	0.2
unemployed_percent	0.19	0.045	-0.02	0.012	-0.39	-0.29	-0.097	0.48	0.27	1	0.28	0.22
deaths_per_1000	0.59	0.1	-0.16	0.14	-0.11	-0.052	-0.078	0.36	0.2	0.28	1	0.74
deaths_quartile	0.58	0.12	-0.23	0.18	-0.057	-0.077	-0.083	0.35	0.2	0.22	0.74	1



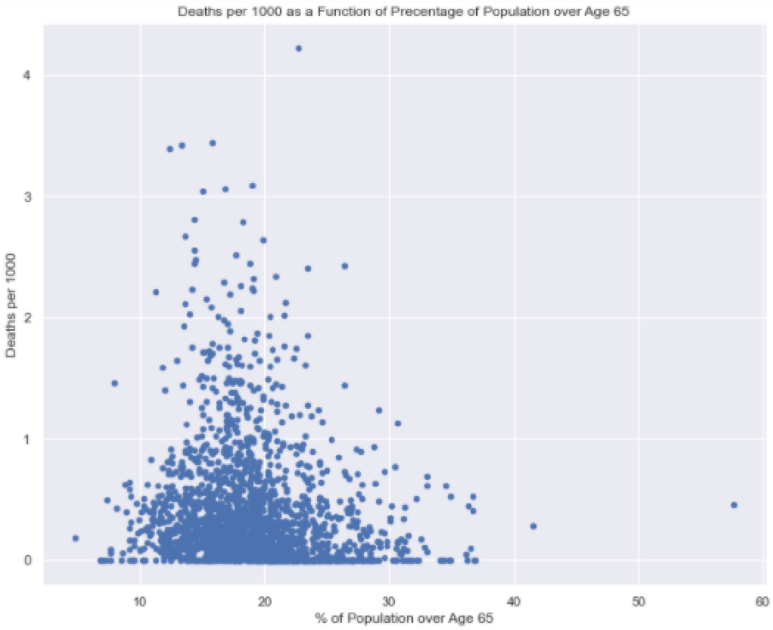
With a correlation coefficient of 0.59 this demonstrates that across counties in the US, there is a strong positive correlation between confirmed cases per 1000 and Covid deaths per 1000.



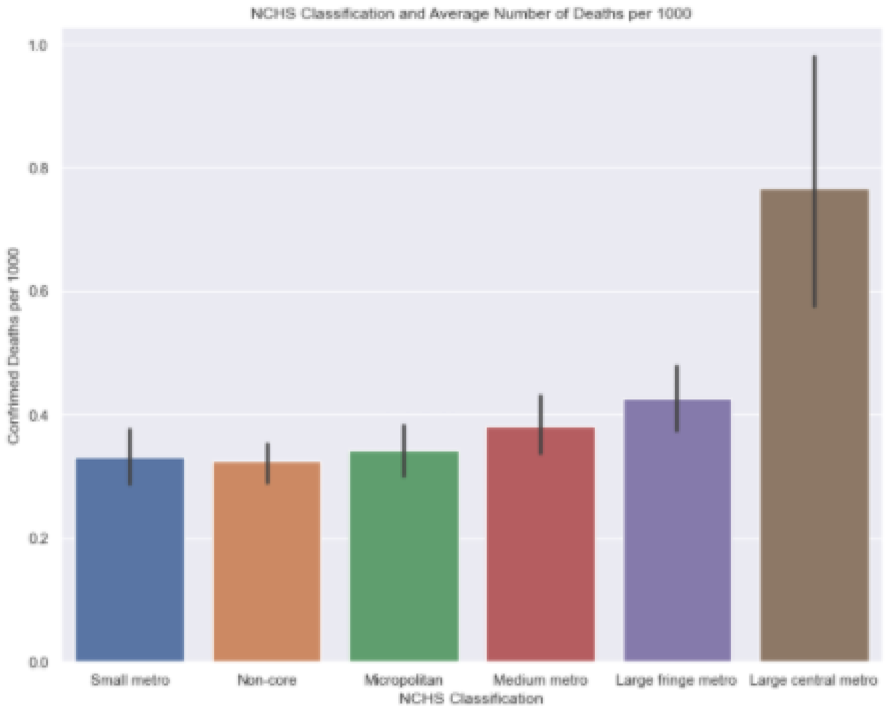
With a correlation coefficient of 0.36, this demonstrates that across counties in the US, there is a slight positive correlation between % without a high school diploma and Covid deaths per 1000.



With a correlation coefficient of 0.20 this demonstrates that across counties in the US, there is a slight positive correlation between % lacking health insurance and Covid deaths per 1000.



With a correlation coefficient of -0.16 this demonstrates that across counties in the US, there is a slight negative correlation between having an older population and Covid rates per 1000. As with the cases data, this is again surprising, as one would expect that having an older population would relate to have higher rates of Covid deaths cases. However, the data shows that this is not necessarily the case.



Large central metro regions have significantly more Covid deaths compared to all other regions.

IV. Analysis/Modeling

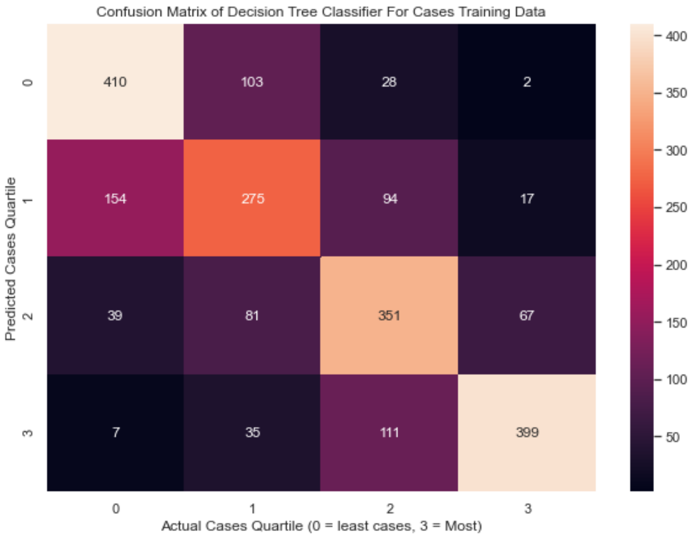
OSEM: Model Data

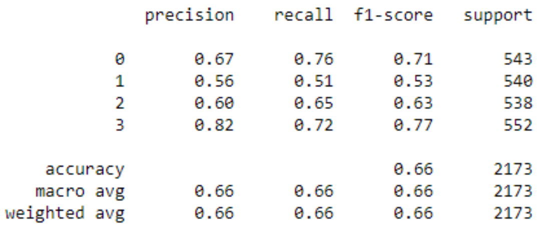
To assess the explain ability of my data, I decided to create 2 types of models for both the confirmed cases quartiles and deaths quartiles. I then compared the accuracies between models and within classes (how well the models were at labeling each quartile). To examine consistency across models and efficacy at classifying the data. Here, I also examined how the models assess feature importance, as that would provide greater insight into which social determinants may be driving Covid case and death rates.

Decision Tree Predicting Cases Quartile :

I first used Scikit-learn’s Decision Tree Classifier (max_depth = 6) to predict case quartiles. On the training set, the model was moderately accurate at 66%. These findings are supported by the classifications in the confusion matrix. With greater classification accuracy on the first and fourth quartiles (as demonstrated by higher precision, recall, and F1 score) and lower accuracy in the middle two quartiles (as demonstrated by lower precision, recall, and F1 score).

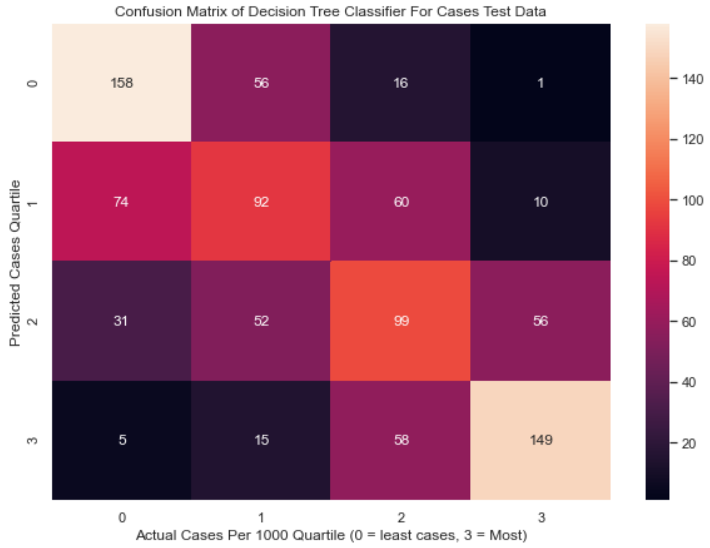
Fig 3a: Training Cases Decision Tree Classifier

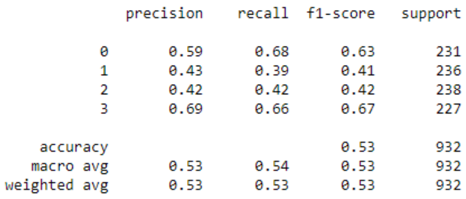




```
              precision    recall  f1-score   support

           0       0.67      0.76      0.71       543
           1       0.56      0.51      0.53       540
           2       0.60      0.65      0.63       538
           3       0.82      0.72      0.77       552

    accuracy                           0.66      2173
   macro avg       0.66      0.66      0.66      2173
weighted avg       0.66      0.66      0.66      2173
```

Fig 3b: Test Cases Decision Tree Classifier



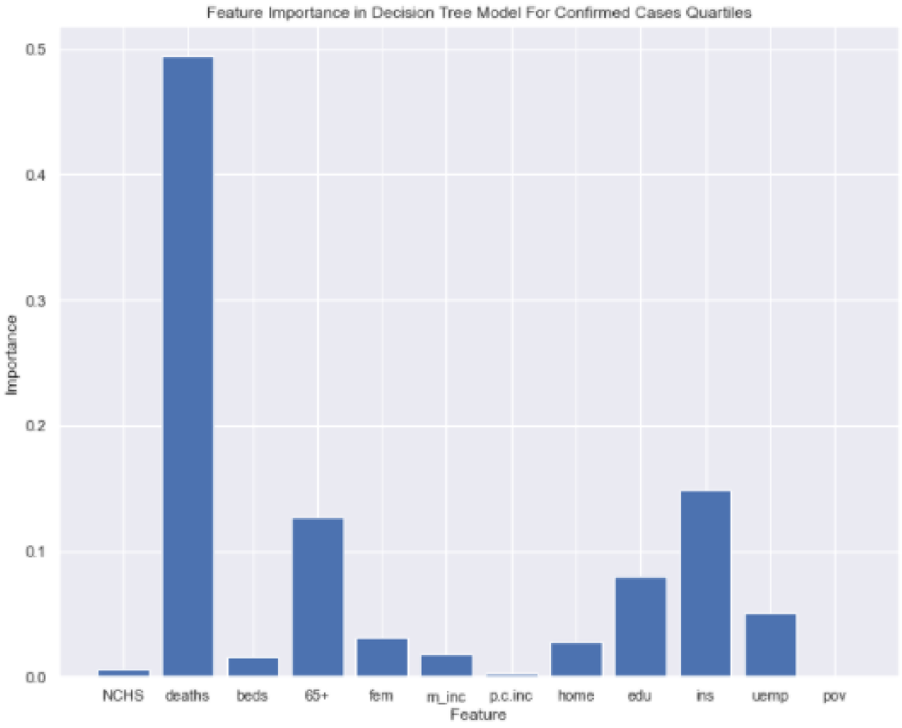


```
              precision    recall  f1-score   support

           0       0.59      0.68      0.63       231
           1       0.43      0.39      0.41       236
           2       0.42      0.42      0.42       238
           3       0.69      0.66      0.67       227

    accuracy                           0.53       932
   macro avg       0.53      0.54      0.53       932
weighted avg       0.53      0.53      0.53       932
```

I then ran the model on the test set, and here accuracy dropped to barely above chance at 53%. First and fourth quartiles had a higher F1 scores (63% and 67%) while the middle two quartiles had lower F1 scores (possibly due to the classifier's inability to distinguish items with these labels from one another, as shown in the confusion matrix).

In terms of figure importance:

Fig 3c: Feature Importance Decision Tree Classifier

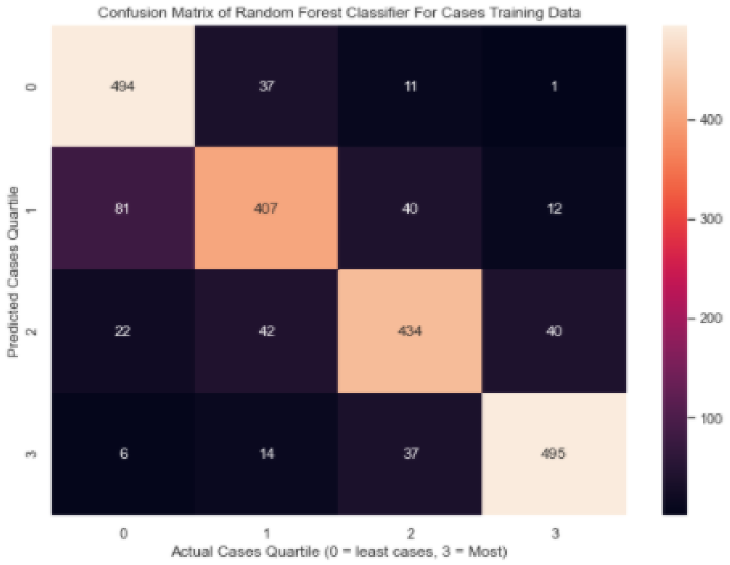


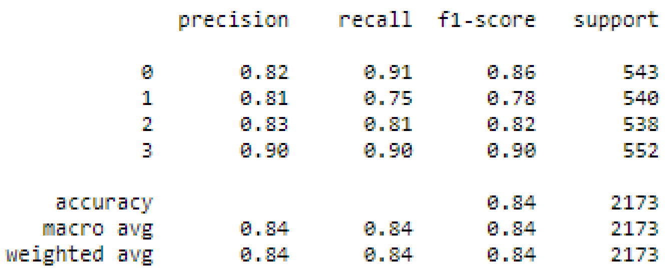
This model emphasizes deaths per 1000, percentage of population over 65, lack of health insurance, education, and unemployment.

Random Forest Predicting Cases Quartile :

To compare between models, I then used Scikit-learn’s Random Forest Classifier (n_estimators=20, max_depth=8, random_state=0) to predict case quartiles. On the training set, the model was highly accurate at 84%. These findings are confirmed by the classifications in the confusion matrix. The model has greater classification accuracy on the first and fourth quartiles (as demonstrated by higher precision, recall, and F1 score) and lower accuracy in the middle two quartiles (as demonstrated by lower precision, recall, and F1 score).

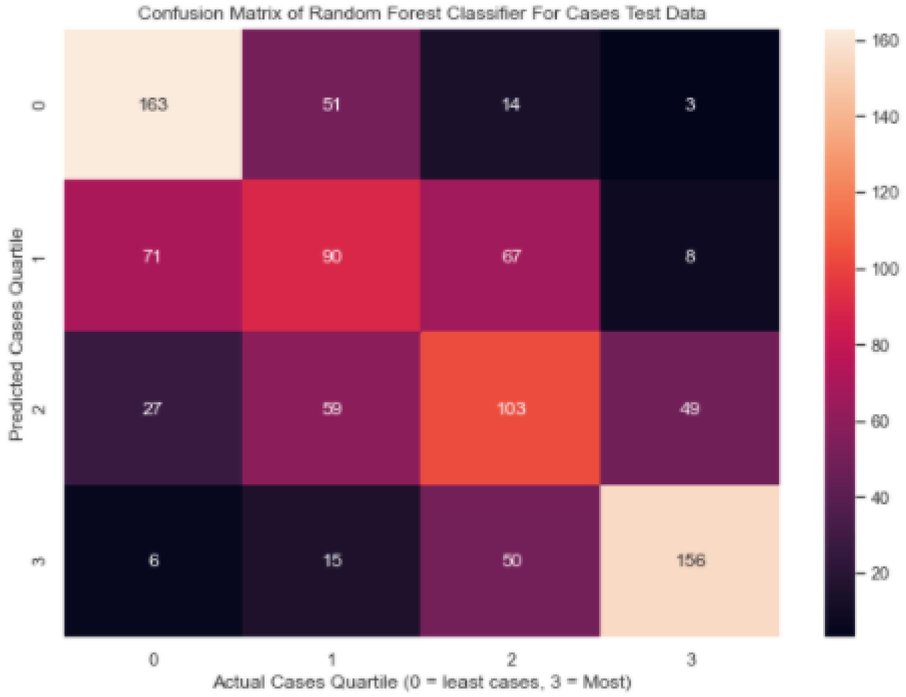
Fig 4a: Training Cases Random Forest Classifier

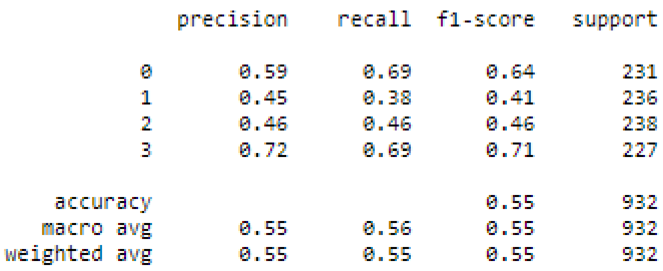




```
              precision    recall  f1-score   support

           0       0.82      0.91      0.86       543
           1       0.81      0.75      0.78       540
           2       0.83      0.81      0.82       538
           3       0.90      0.90      0.90       552

    accuracy                           0.84      2173
   macro avg       0.84      0.84      0.84      2173
weighted avg       0.84      0.84      0.84      2173
```

Fig 4b: Test Cases Random Forest Classifier

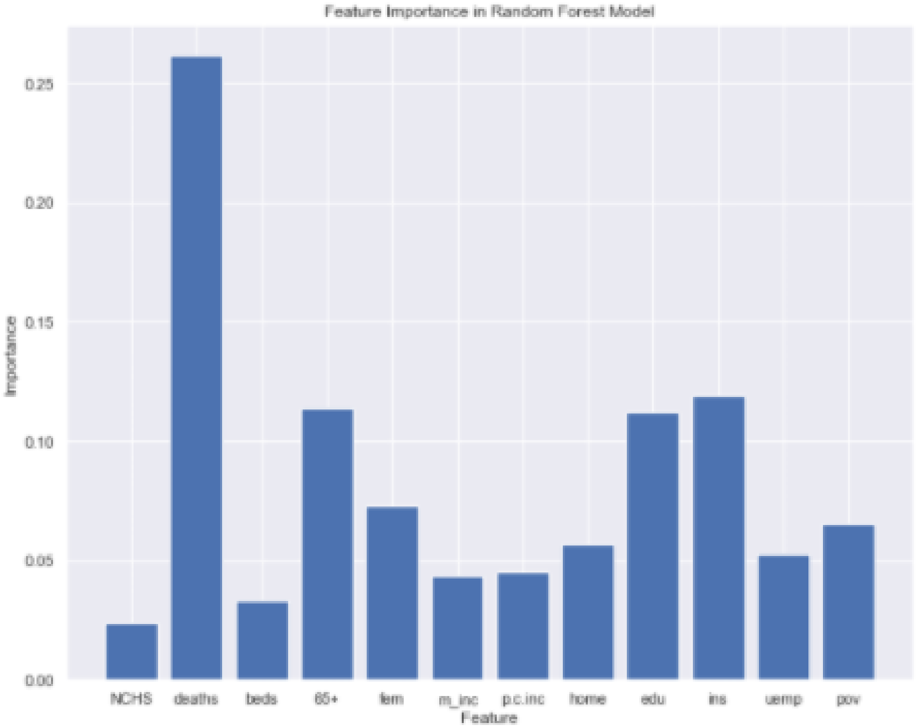




```
              precision    recall  f1-score   support

           0       0.59      0.69      0.64       231
           1       0.45      0.38      0.41       236
           2       0.46      0.46      0.46       238
           3       0.72      0.69      0.71       227

    accuracy                           0.55       932
   macro avg       0.55      0.56      0.55       932
weighted avg       0.55      0.55      0.55       932
```

I then ran the model on the test set, and here accuracy dropped to barely above chance at 55%. First and fourth quartiles had higher F1 scores (64% and 71%) while the middle two quartiles had lower F1 scores (possibly due to the classifier's inability to distinguish items with these labels from one another, as shown in the confusion matrix).

Fig 4c: Feature Importance Random Forest

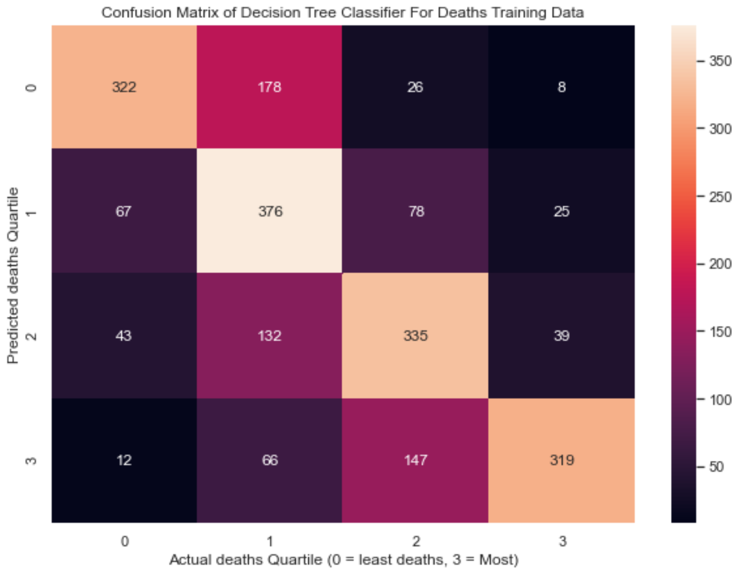


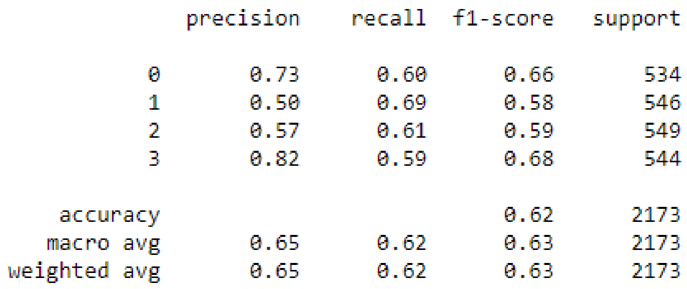
This model emphasizes deaths per 1000, lack of health insurance, and education, and percentage of population above age 65.

<u>Decision Tree Predicting Deaths Quartile :</u>

To make death quartile predictions, I then went back to Decision Tree Classifier (max_depth = 6) to predict death quartiles. On the training set, the model was moderately accurate at 62%. With greater classification accuracy on the first and fourth quartiles (as demonstrated by higher precision and F1 score) and lower accuracy in the middle two quartiles (as demonstrated by lower precision and F1 score).

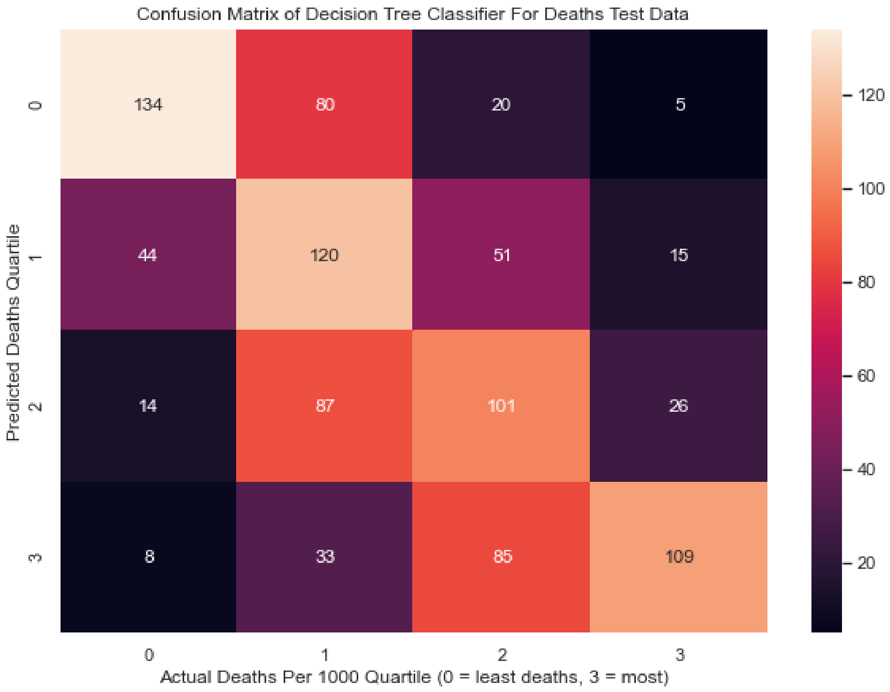
Fig 5a: Training Deaths Decision Tree Classifier

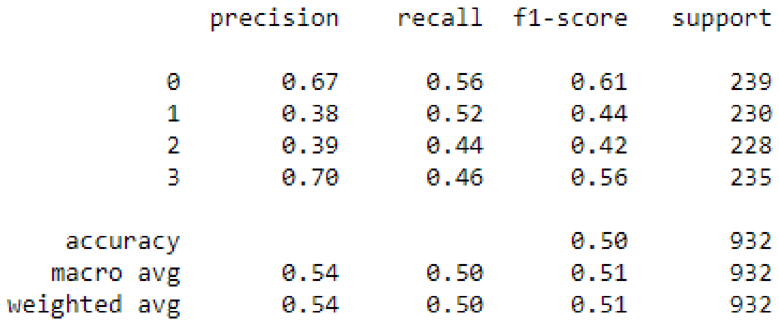




```
              precision    recall  f1-score   support

           0       0.73      0.60      0.66       534
           1       0.50      0.69      0.58       546
           2       0.57      0.61      0.59       549
           3       0.82      0.59      0.68       544

    accuracy                           0.62      2173
   macro avg       0.65      0.62      0.63      2173
weighted avg       0.65      0.62      0.63      2173
```

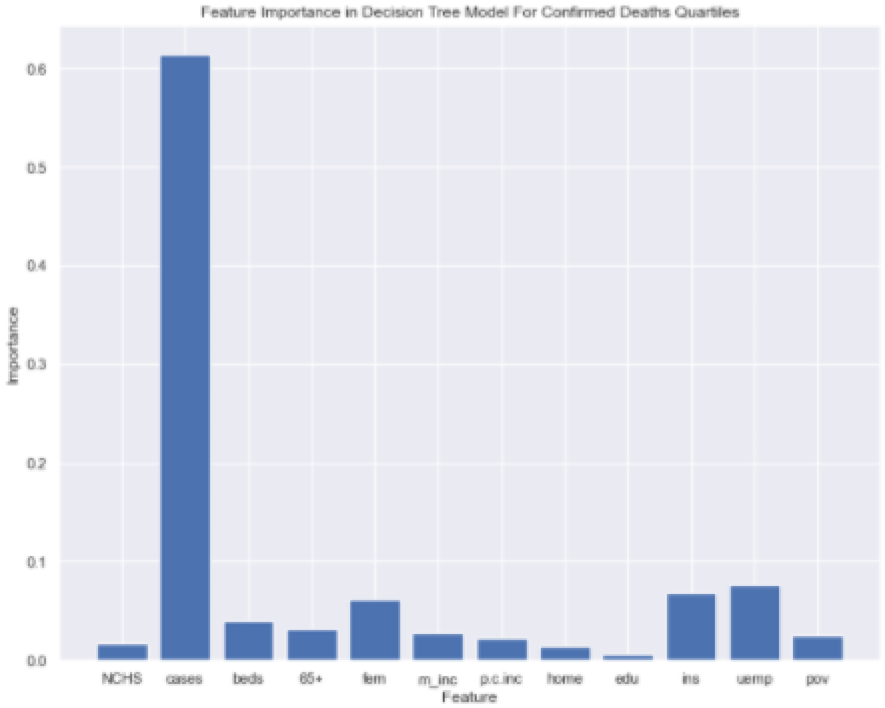
Fig 5b: Test Deaths Decision Tree Classifier





```
              precision    recall  f1-score   support

           0       0.67      0.56      0.61       239
           1       0.38      0.52      0.44       230
           2       0.39      0.44      0.42       228
           3       0.70      0.46      0.56       235

    accuracy                           0.50       932
   macro avg       0.54      0.50      0.51       932
weighted avg       0.54      0.50      0.51       932
```

I then ran the model on the test set, and here accuracy dropped to barely above chance at 50%. First and fourth quartiles had slightly higher F1 scores (61% and 56%) while the middle two quartiles had lower F1 scores (possibly due to the classifier's inability to distinguish items with these labels from one another, as shown in the confusion matrix).

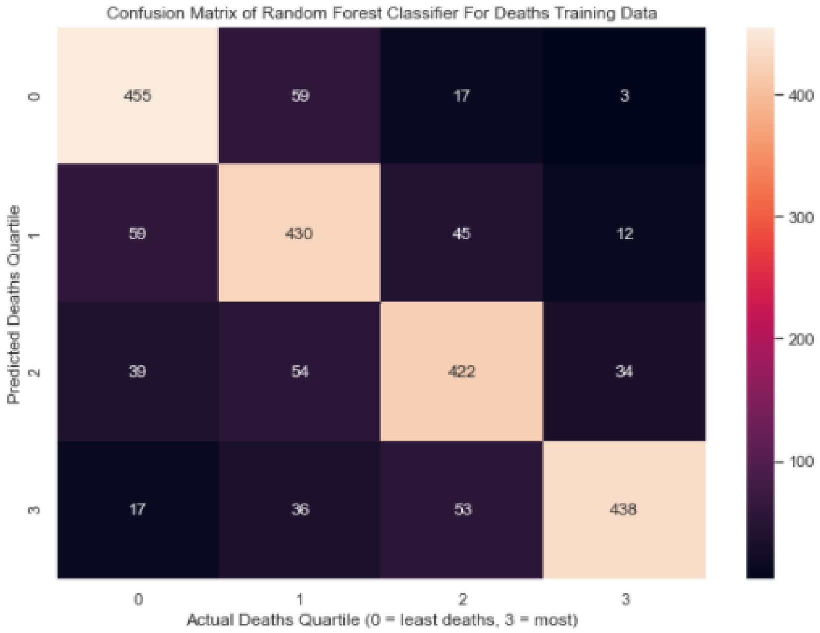


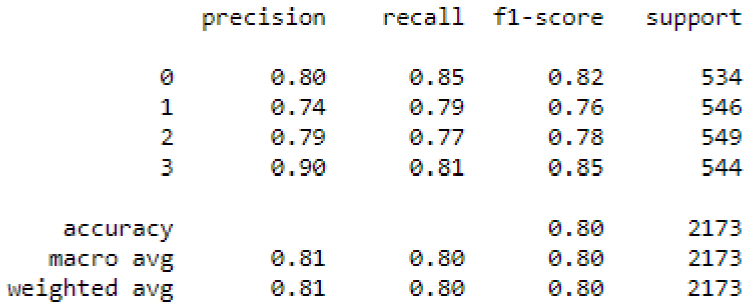
This model emphasizes cases per 1000, lack of health insurance, and unemployment rate.

Random Forest Predicting Deaths Quartile :

To compare between models when looking at deaths, I then used Scikit-learn's Random Forest Classifier (n_estimators=20, max_depth=8, random_state=0) to predict death quartiles. On the training set, the model was highly accurate at 80%. With greater classification accuracy on the first and fourth quartiles (as demonstrated by higher precision, recall, and F1 score) and lower accuracy in the middle two quartiles (as demonstrated by lower precision, recall, and F1 score). This trend is also shown in the confusion matrix.

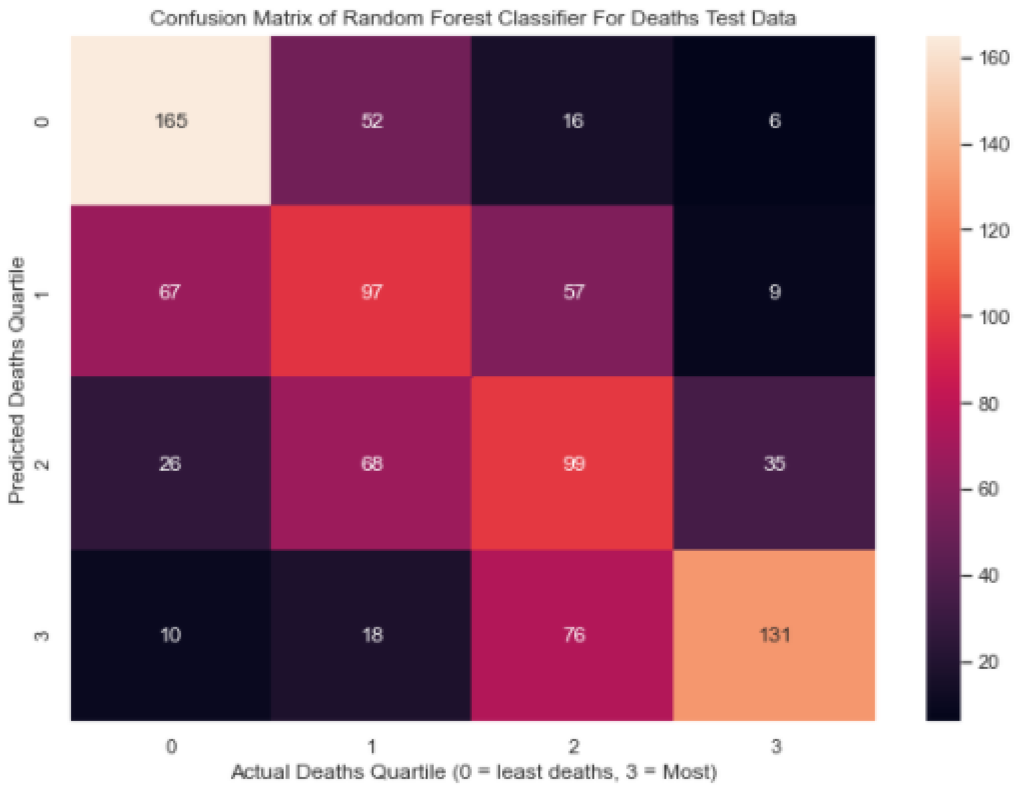
Fig 6a: Training Deaths Random Forest Classifier

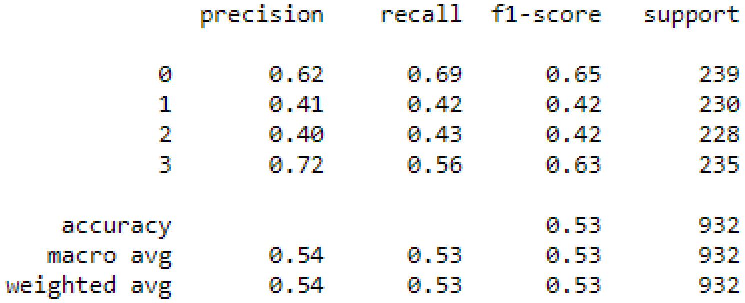




```
              precision    recall  f1-score   support

           0       0.80      0.85      0.82       534
           1       0.74      0.79      0.76       546
           2       0.79      0.77      0.78       549
           3       0.90      0.81      0.85       544

    accuracy                           0.80      2173
   macro avg       0.81      0.80      0.80      2173
weighted avg       0.81      0.80      0.80      2173
```

Fig 6b: Test Deaths Random Forest Classifier

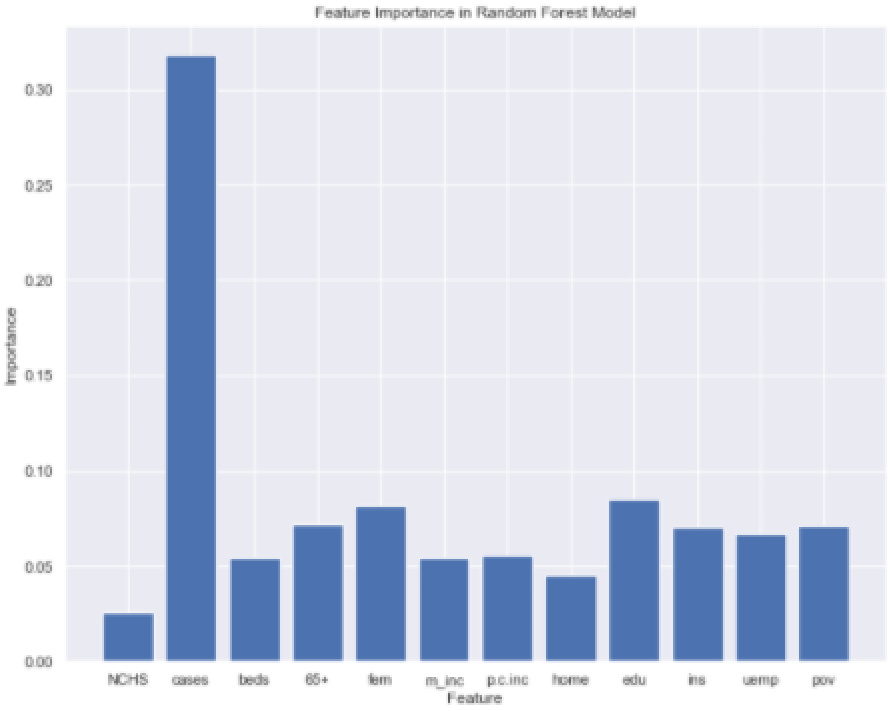




```
              precision    recall  f1-score   support

           0       0.62      0.69      0.65       239
           1       0.41      0.42      0.42       230
           2       0.40      0.43      0.42       228
           3       0.72      0.56      0.63       235

    accuracy                           0.53       932
   macro avg       0.54      0.53      0.53       932
weighted avg       0.54      0.53      0.53       932
```

I then ran the model on the test set, and here accuracy dropped to barely above chance at 53%. First and fourth quartiles had slightly higher F1 scores (65% and 63%) while the middle two quartiles had lower F1 scores (possibly due to the classifier's inability to distinguish items with these labels from one another, as shown in the confusion matrix).

Fig 6c: Feature Importance Random Forest Classifier



This model emphasizes cases per 1000, % of population without a high school diploma, and percentage of population that is female.

When looking at the 2 models, (decision trees and random forest), it is evident that the training set has better accuracy (in predicting the prevalence of Covid deaths per county) compared to the test set (in some cases, the test set was just above chance when classifying counties into the appropriate quartiles). However, accuracies for the test set were relatively high when observing the top and bottom quartiles (above 60% with all models). For this reason, my analysis will consider the top and bottom quartiles when looking at Covid deaths disparities.

It is also clear that in addition to Covid cases, features such as % of people over age 65 (negatively correlated), % of people without a high school diploma (positively correlated), % of people without health insurance (positively correlated), and % of people unemployed (positively correlated) are important when examining prevalence of Covid deaths.

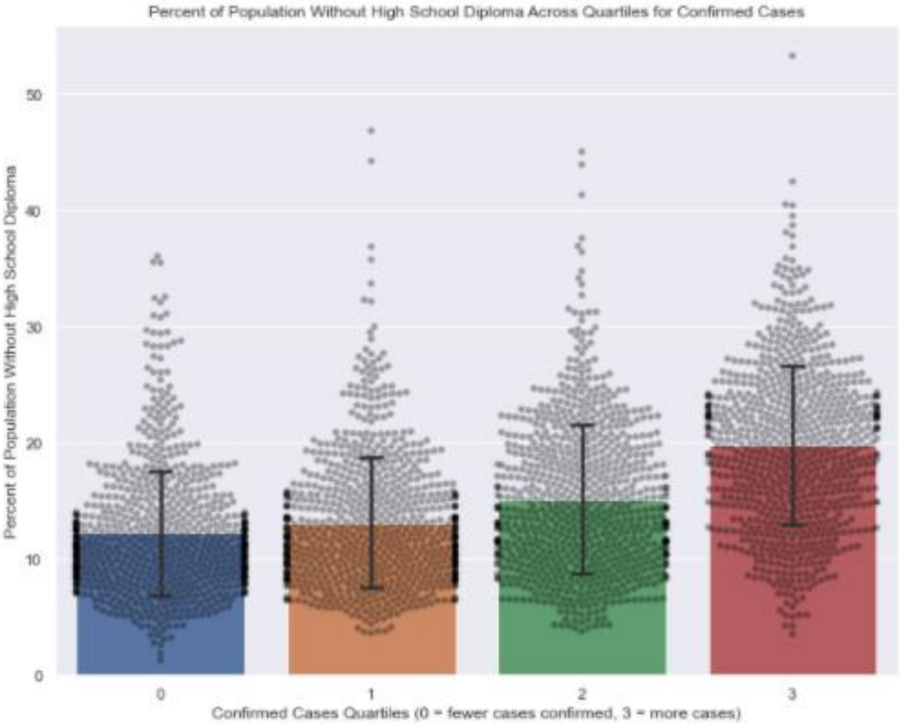
The random forest model had the best test accuracy for both confirmed cases quartile and deaths quartile, this model is good because it provides close predictions in the top and fourth quartiles and gives features which help to explain the differences, we are seeing between the Covid cases and death across the quartiles. Such features include % of people without a high school diploma (positively correlated with higher incidence of Covid cases and death) and % of people without health insurance (also positively correlated with higher incidence of Covid cases and death).

V. Results

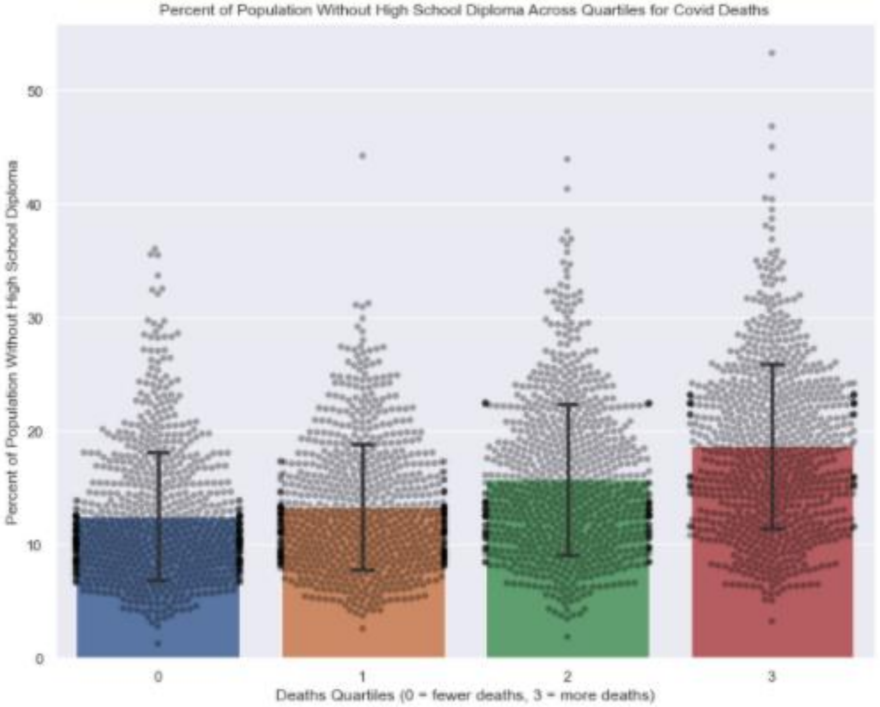
Looking at Averages Across Quartiles:

Based on the random forest classifiers modeling confirmed Covid cases and deaths, it is evident that features such as percent of people without a high school diploma, percent of people without health insurance, and percent of people living below the poverty line are the primary socioeconomic drivers of the differences within the data. For this reason, I looked at the average values of these features between the top and bottom quartiles of confirmed cases and deaths.

Figure 7: Looking at Averages Across Quartiles, % Without High School Diploma



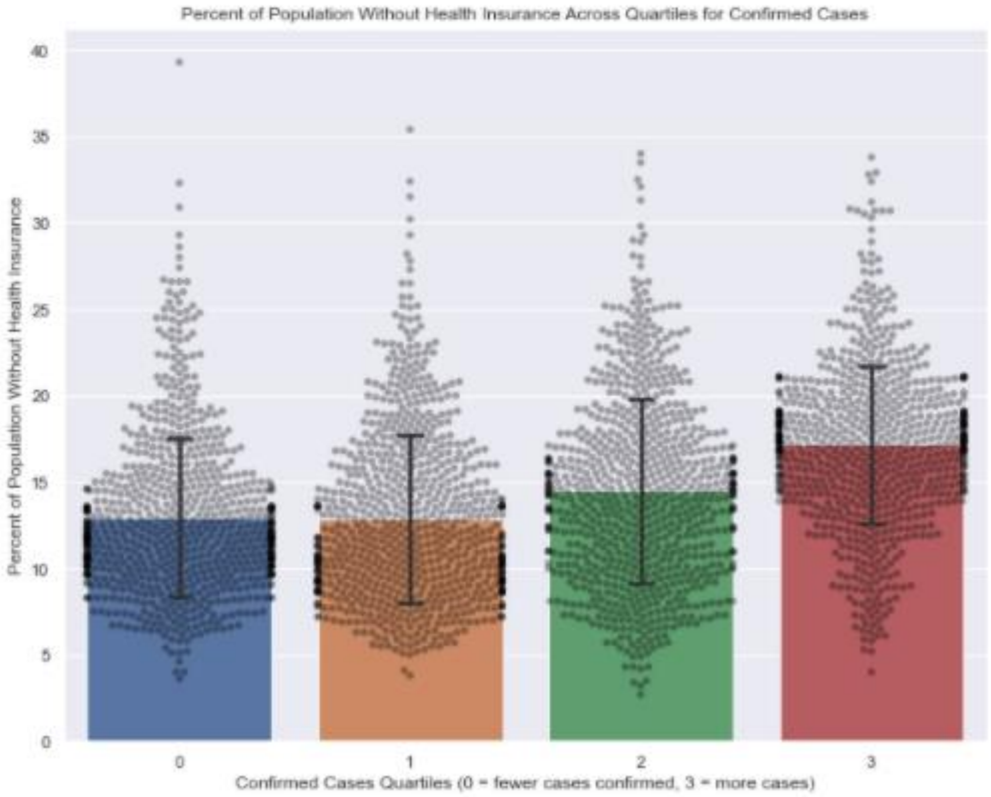
The mean percentage of people without a high school diploma in the bottom quartile (lowest Covid Cases): 12.22+/- 5.38
The mean percentage of people without a high school diploma in the top quartile (highest Covid Cases): 19.79 +/- 6.87



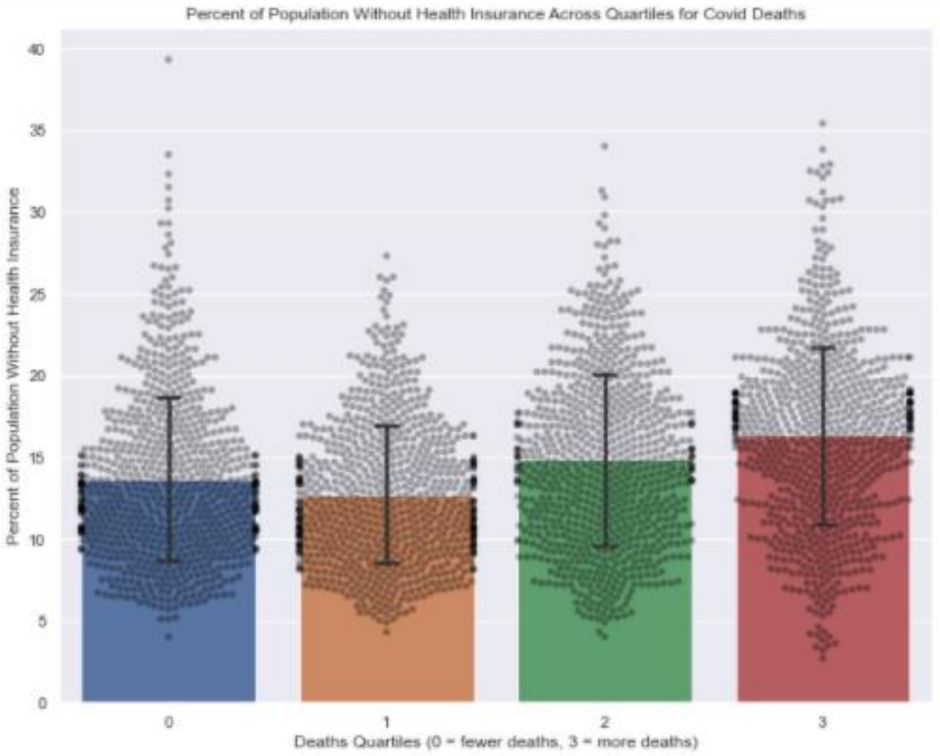
The mean percentage of people without a high school diploma in the bottom quartile (lowest Covid deaths) of Covid death rates is: 12.44 +/- 5.65
The mean percentage of without a high school diploma in the top quartile (highest Covid deaths) of Covid death rates is: 18.69 +/- 7.23

From this data (the graphs as well as the means and standard deviations) we can see that as % of people without a high school diploma increases, Covid incidence and deaths also increase. This trend is also supported by the distribution of % no diploma (scatterplot) across the 4 outcome buckets. However, analysis demonstrated that because of the standard deviations, differences across these buckets were not significant.

Figure 8: Looking at Averages Across Quartiles, % Without Health Insurance



The mean percentage of people without health insurance in the bottom quartile (lowest Covid Cases) of Covid case rates is: 12.99+/- 4.60
The mean percentage of without health insurance in the top quartile (highest Covid Cases) of Covid case rates is: 17.17 +/- 4.58



The mean percentage of people without health insurance in the bottom quartile (lowest Covid deaths) of Covid death rate is: 13.68 +/- 5.01
The mean percentage of without health insurance in the top quartile (highest Covid deaths) of Covid death rate is: 16.30 +/- 5.42

From this data (the graphs as well as the means and standard deviations) we can see that as % of people without a health insurance increases, COVID-19 incidence and deaths also increase. This trend is also supported by the distribution of % no insurance (scatterplot) across the 4 outcome buckets, however analysis demonstrated that because of the standard deviations, differences across these buckets were not significant.

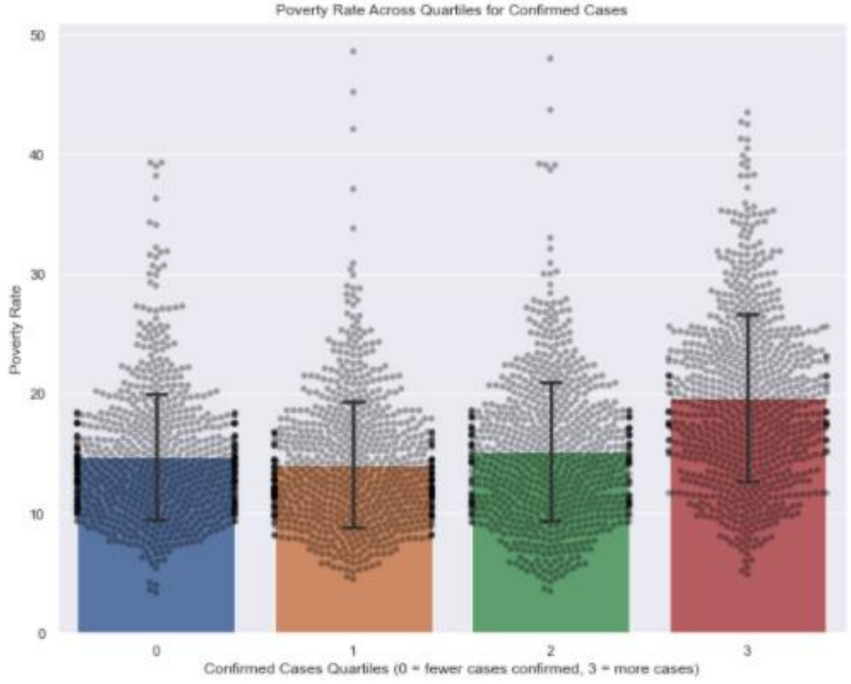
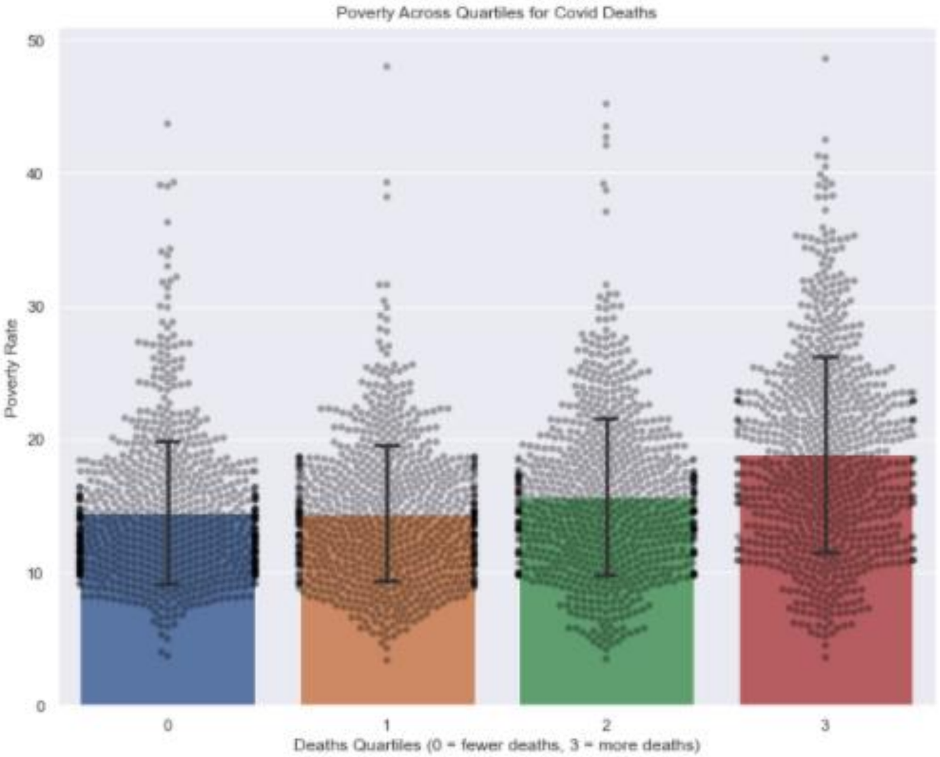


Figure 9: Looking at Averages Across Quartiles, Poverty Rates

The mean percentage of people below the poverty line in the bottom quartile (lowest Covid Cases) of Covid case rates is: 14.74 +/- 5.25
The mean percentage of below the poverty line in the top quartile (highest Covid Cases) of Covid case rates is: 19.65 +/- 6.99



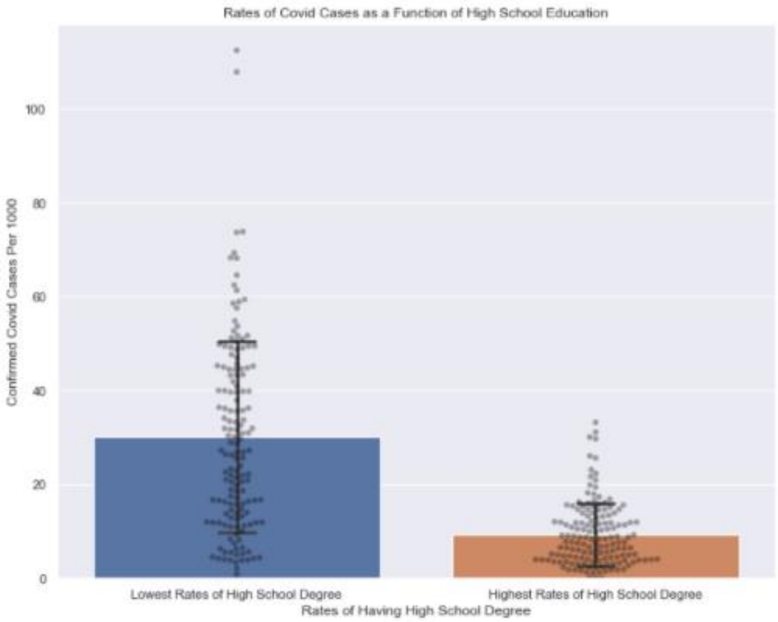
The mean percentage of people below the poverty line in the bottom quartile (lowest Covid deaths) of Covid death rates is: 14.52 +/- 5.38
The mean percentage of below the poverty line in the top quartile (highest Covid deaths) of Covid death rates is: 18.90 +/- 7.35

From this data (the graphs as well as the means and standard deviations) we can see that as % of people below the poverty line increases, Covid incidence and deaths also increase. This trend is also supported by the distribution of % of people below the poverty line across the 4 outcome buckets (as seen in the scatterplot). However, analysis demonstrated that because of the standard deviations, differences across these buckets were not significant.

Taken together, these plots reveal that when looking across Covid outcome and Covid death quartiles, a wide variety of the population is affected by the pandemic to similar degrees. This is demonstrated by the overlapping error bars even between the top and bottom quartiles when looking at features which may help explain the differences in outcome. These not significant results are most likely due to the pandemic's widespread effects, targeting various locations with different qualities, as represented by the large range in dots within an individual outcome bucket (quartile). From this, it is evident that the disparities we are seeing (differences between the top and bottom quartile) are not due to most of the data, but rather due to outliers. Since outliers drive disparities, I will now look at the effects of outliers on Covid cases and deaths.

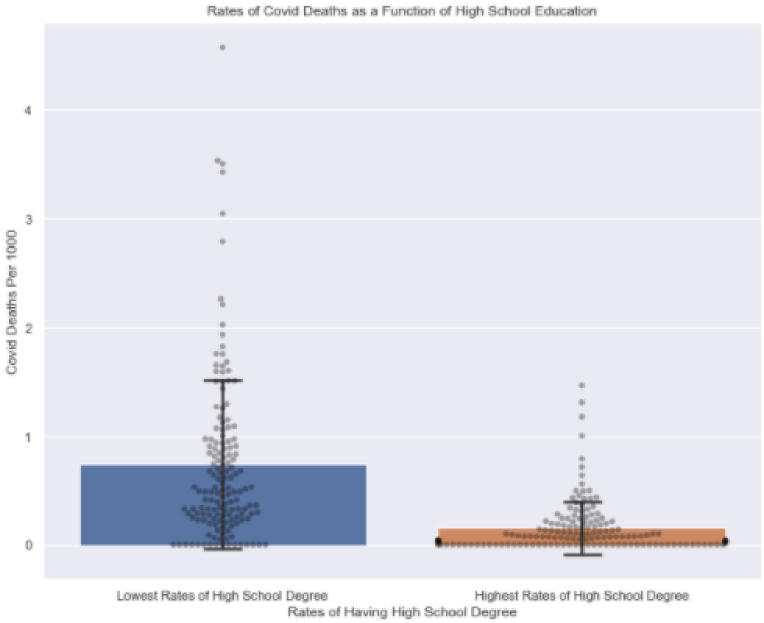
I thought it would be interesting to examine the top 150 (top 5%) and bottom 150 counties (bottom 5%) across a variety of socioeconomic features and observe the differences in confirmed cases and deaths across these 2 groups.

Disparities: % of Population Without High School Diploma and Covid Cases & Deaths

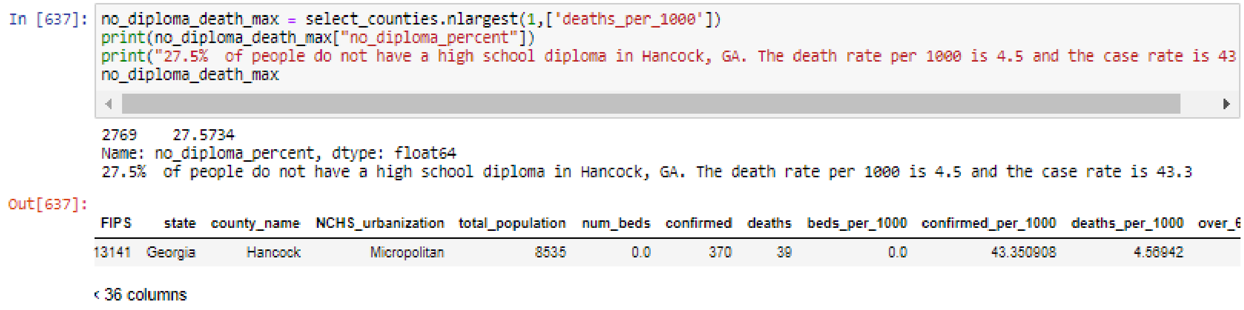


The mean number of Covid Cases per 1000 among the group with the lowest rates of having a high school degree is: 30.01 +/- 20.39
The mean number of Covid Cases per 1000 among the group with the highest rates of having a high school degree is: 9.31 +/- 6.68

The mean number of Covid Deaths per 1000 among the group with the lowest rates of having a high school degree is: 0.74 +/- 0.77
The mean number of Covid Deaths per 1000 among the group with the highest rates of having a high school degree is: 0.15 +/- 0.24



What is the point at the very top on the left?



In [637]:
```
no_diploma_death_max = select_counties.nlargest(1,['deaths_per_1000'])
print(no_diploma_death_max["no_diploma_percent"])
print("27.5%  of people do not have a high school diploma in Hancock, GA. The death rate per 1000 is 4.5 and the case rate is 43
no_diploma_death_max
```

```
2769    27.5734
Name: no_diploma_percent, dtype: float64
27.5%  of people do not have a high school diploma in Hancock, GA. The death rate per 1000 is 4.5 and the case rate is 43.3
```

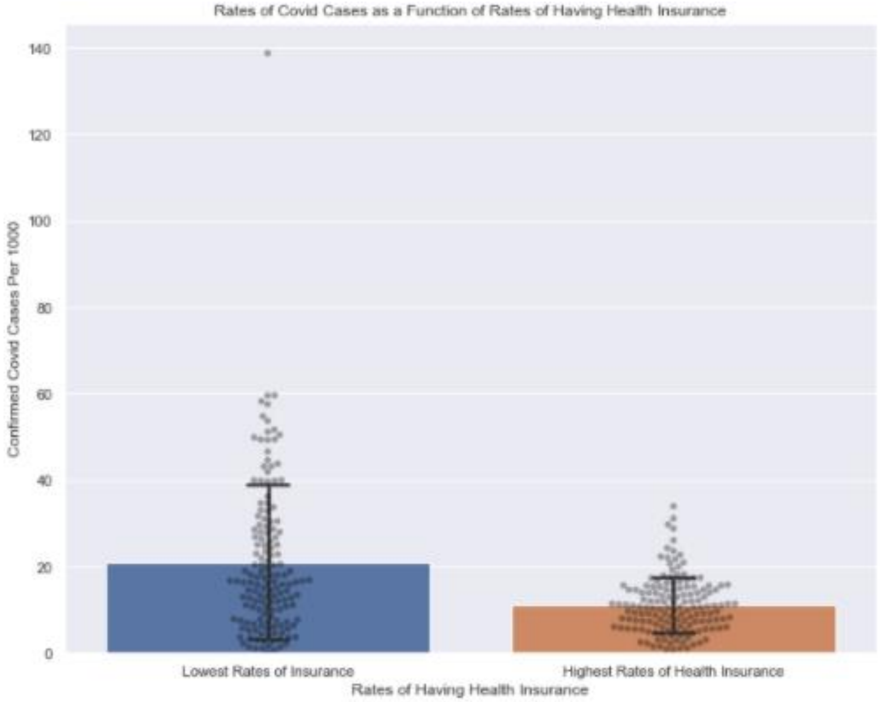
Out[637]:

FIPS	state	county_name	NCHS_urbanization	total_population	num_beds	confirmed	deaths	beds_per_1000	confirmed_per_1000	deaths_per_1000	over_6
13141	Georgia	Hancock	Micropolitan	8535	0.0	370	39	0.0	43.350908	4.56942	

‹ 36 columns

Once again, the data (the graphs as well as the means and standard deviations) shows that while having a high school education is associated with reduced rates of Covid cases and deaths, these differences are not significant, as error bars and ranges overlap. However, we start to see interesting data points, such as Hancock, GA, which has the lowest rate of high school diplomas and high rates of Covid cases and death.

Disparities: % of People without Health Insurance and Covid Cases & Deaths

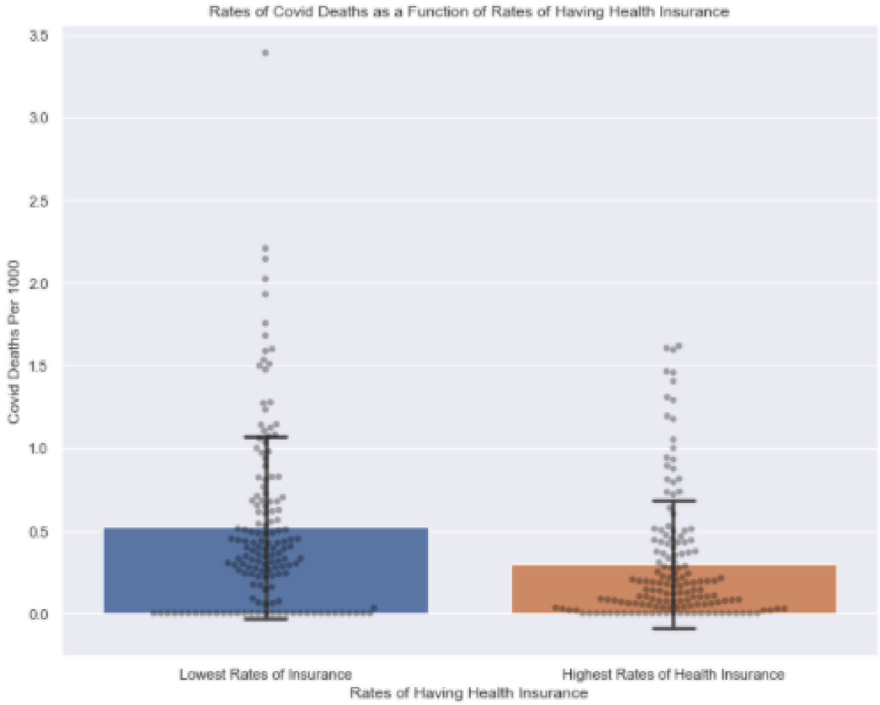


The mean number of Covid Cases per 1000 among the group with the lowest rates of having health insurance is: 20.99 +/- 17.99

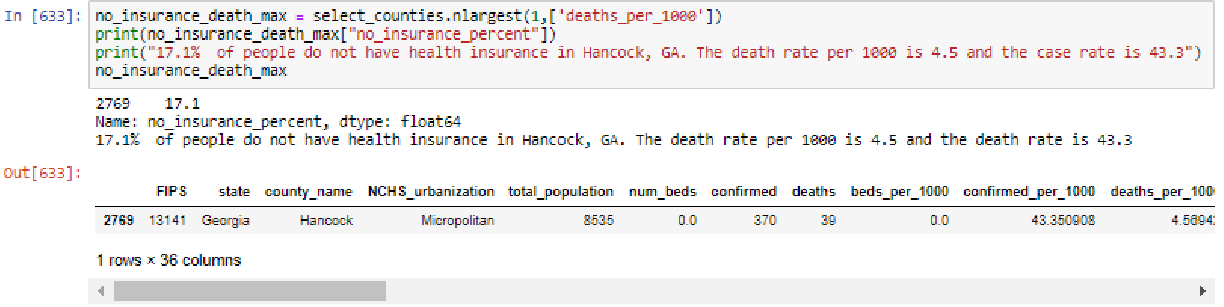
The mean number of Covid Cases per 1000 among the group with the highest rates of having health insurance is: 11.05 +/- 6.42

The mean number of Covid Deaths per 1000 among the group with the lowest rates of having health insurance is: 0.52+/- 0.55

The mean number of Covid Deaths per 1000 among the group with the highest rates of having health insurance is: 0.30 +/- 0.39



What is the point at the very top on the left?



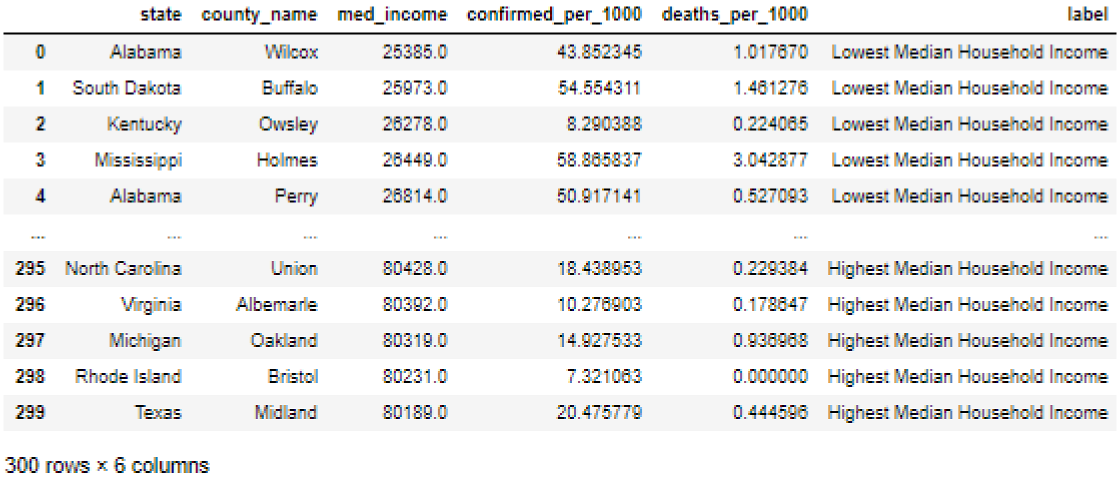
```
In [633]: no_insurance_death_max = select_counties.nlargest(1,['deaths_per_1000'])
          print(no_insurance_death_max["no_insurance_percent"])
          print("17.1%  of people do not have health insurance in Hancock, GA. The death rate per 1000 is 4.5 and the case rate is 43.3")
          no_insurance_death_max

          2769    17.1
          Name: no_insurance_percent, dtype: float64
          17.1%  of people do not have health insurance in Hancock, GA. The death rate per 1000 is 4.5 and the death rate is 43.3

Out[633]:
```

	FIPS	state	county_name	NCHS_urbanization	total_population	num_beds	confirmed	deaths	beds_per_1000	confirmed_per_1000	deaths_per_100
2769	13141	Georgia	Hancock	Micropolitan	8535	0.0	370	39	0.0	43.350908	4.5694

1 rows × 36 columns

Once again, the data (the graphs as well as the means and standard deviations) shows that while having a high school education is associated with reduced rates of Covid cases and deaths, these differences are not significant, as error bars and ranges overlap. In terms of looking at counties, we see that Hancock, GA, which was in the bottom 10% for high school graduates, is also in the bottom 10% for having health insurance (while having high levels of cases and deaths).

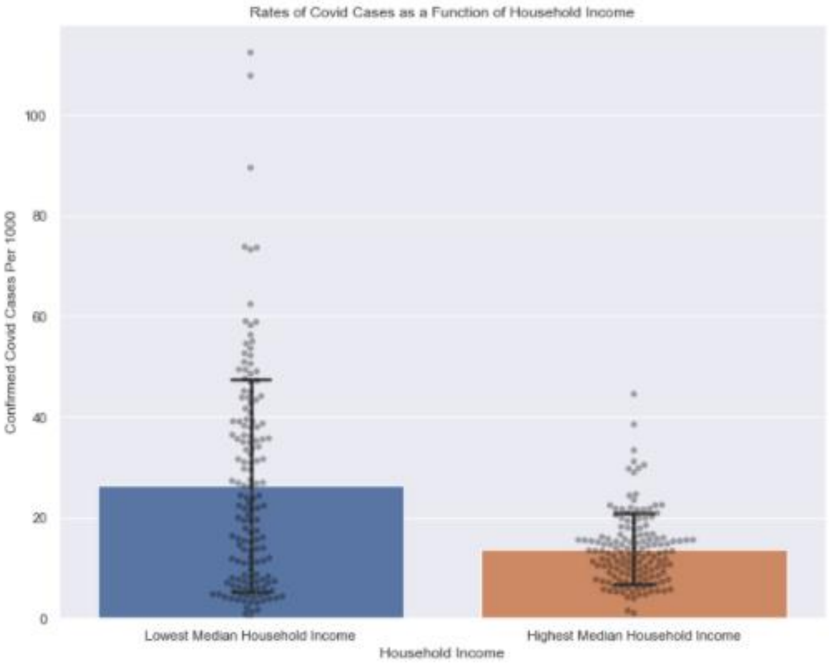
Disparities: Median Household Income and Covid Cases & Deaths



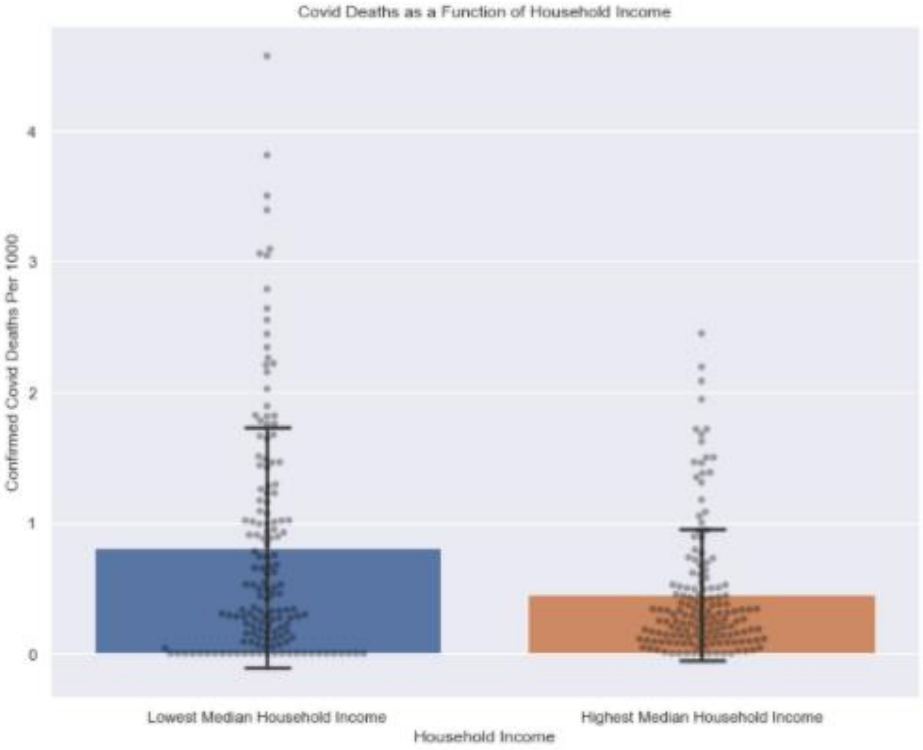
	state	county_name	med_income	confirmed_per_1000	deaths_per_1000	label
0	Alabama	Wilcox	25385.0	43.852345	1.017670	Lowest Median Household Income
1	South Dakota	Buffalo	25973.0	54.554311	1.461276	Lowest Median Household Income
2	Kentucky	Owsley	26278.0	8.290388	0.224065	Lowest Median Household Income
3	Mississippi	Holmes	26449.0	58.865837	3.042877	Lowest Median Household Income
4	Alabama	Perry	26814.0	50.917141	0.527093	Lowest Median Household Income
...	...	...	...	...	...	...
295	North Carolina	Union	80428.0	18.438953	0.229384	Highest Median Household Income
296	Virginia	Albemarle	80392.0	10.276903	0.178647	Highest Median Household Income
297	Michigan	Oakland	80319.0	14.927533	0.936968	Highest Median Household Income
298	Rhode Island	Bristol	80231.0	7.321063	0.000000	Highest Median Household Income
299	Texas	Midland	80189.0	20.475779	0.444596	Highest Median Household Income

300 rows × 6 columns

I was interested in looking at this feature since it is inherently disparate

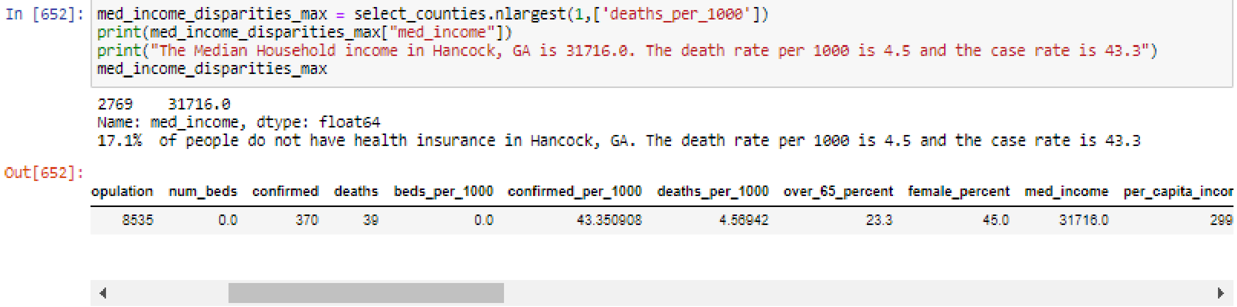


The mean number of Covid Cases per 1000 among the group with the lowest Median Household Income is: 26.39 +/- 21.19
The mean number of Covid Cases per 1000 among the group with the highest Median Household Income is: 13.79 +/- 7.05



The mean number of Covid Deaths per 1000 among the group with the lowest Median Household Income is: 0.81+/- 0.92
The mean number of Covid Deaths per 1000 among the group with the highest Median Household Income is: 0.45 +/- 0.50

What is the point at the very top on the left?



```
In [652]: med_income_disparities_max = select_counties.nlargest(1,['deaths_per_1000'])
          print(med_income_disparities_max["med_income"])
          print("The Median Household income in Hancock, GA is 31716.0. The death rate per 1000 is 4.5 and the case rate is 43.3")
          med_income_disparities_max

          2769    31716.0
          Name: med_income, dtype: float64
          17.1%  of people do not have health insurance in Hancock, GA. The death rate per 1000 is 4.5 and the case rate is 43.3

Out[652]:
```

opulation	num_beds	confirmed	deaths	beds_per_1000	confirmed_per_1000	deaths_per_1000	over_65_percent	female_percent	med_income	per_capita_incor
8535	0.0	370	39	0.0	43.350908	4.56942	23.3	45.0	31716.0	299

Once again, the data (the graphs as well as the means and standard deviations) shows that while having a higher median household income is associated with reduced rates of Covid cases and deaths, these differences are not significant, as error bars and ranges overlap. In terms of looking at counties, we see that Hancock, GA, which was in the bottom 10% for high school graduates and having health insurance (while having high levels of cases and deaths) is also in the bottom 10% for median household income, suggesting that in these factors are related to each other, and may be negatively affecting Covid outcomes in the county.

V. Discussion

From the results, it is evident that while disparities such as lack of a high school diploma and lack of health insurance have a slight positive correlation with Covid-19 cases and deaths, these correlations do not produce significant results when comparing the most advantaged and most disadvantaged groups (in terms of Covid cases and outcomes). Nevertheless, it is evident that these social determinants do play some role. When observing disparities in determinants observed, there was more variability (greater standard deviations and therefore greater variance) in Covid outcomes for disadvantaged groups compared to the most advantaged groups. While these social determinants did not illustrate significant differences in Covid outcomes across all 3000+ counties they certainly play a role in bringing Covid harm to already disadvantaged communities (as seen with counties like Hancock, GA). Future work must continue to address such determinants of health to ensure that the present disparities do not become further widened because of the pandemic. Such is especially true for high school education, as the current pandemic has forced students to engage in remote learning. Because this determinant has already displayed a slight (but not significant) relationship to Covid outcome, we must endure that the education gap does not widen due to a lack of resources to learn.

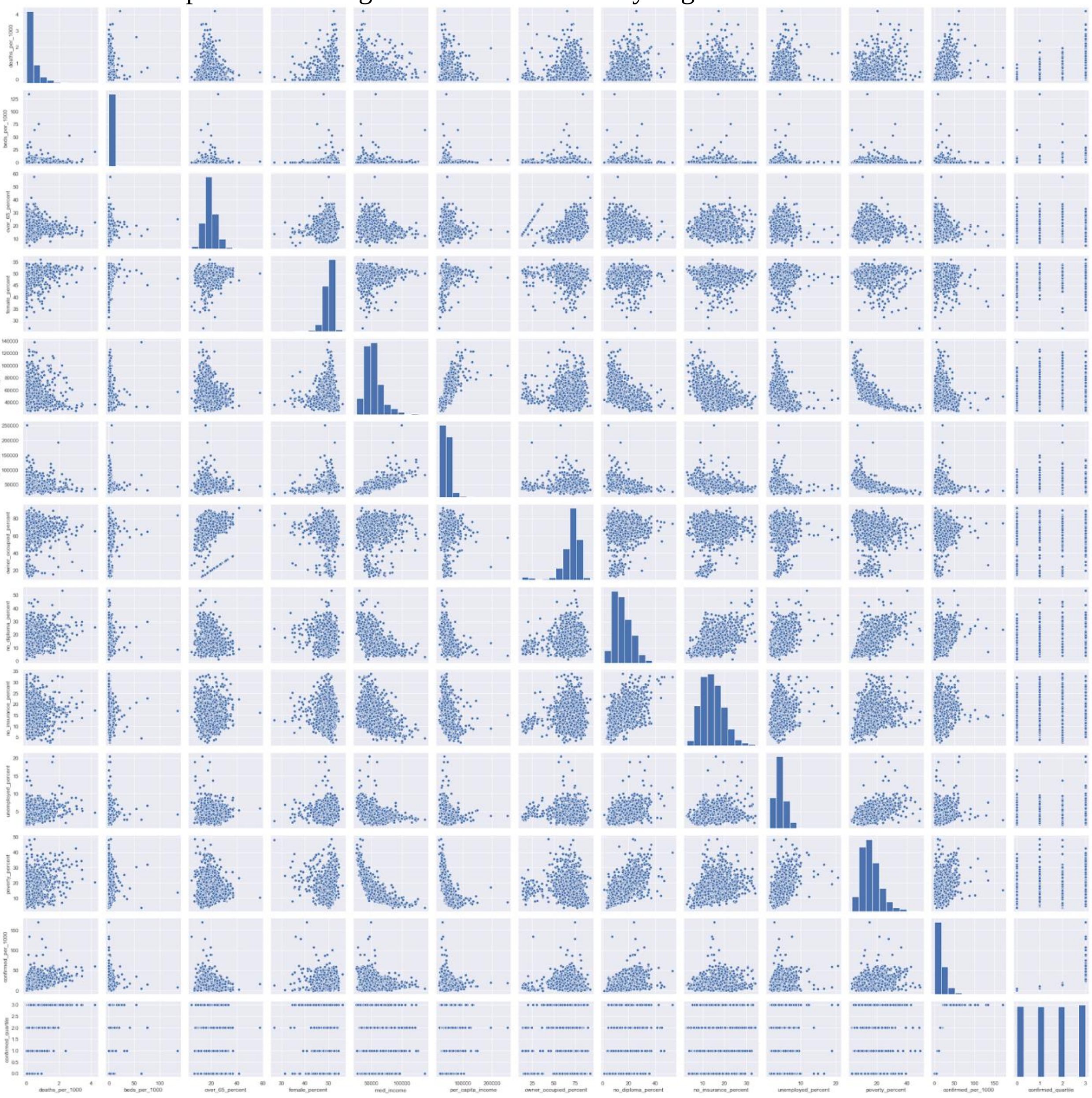
As the Covid situation is ongoing, there is still much to learn and there are many avenues for future investigation. First, this project observed Covid deaths and cases from March to September 7, 2020. It is now December 2020, and within a few months, we have already seen spikes in new cases and deaths. Although these deaths are unfortunately, future studies should work with the most recent data when it becomes available to better understand how social determinants may influence Covid outcomes longitudinally.

Another possible avenue of research is to observe cities of similar size and population density to understand how policy may affect outcomes. My project illustrated that those in a large metropolitan area are significantly affected by Covid cases and deaths (so while city environment is a social determinant, it is not a health disparity). Therefore , it would be interesting to observe how demographically similar cities fare during a pandemic because of policy intervention (or lack thereof).

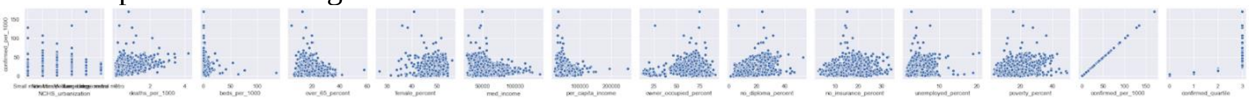
Ultimately this project illustrated how several non-biological factors relate to Covid outcomes. It determined that while factors such as education and having health insurance are not strongly directly related to Covid outcomes nationwide, they may still work to harm already vulnerable populations.

Supplementary Figures

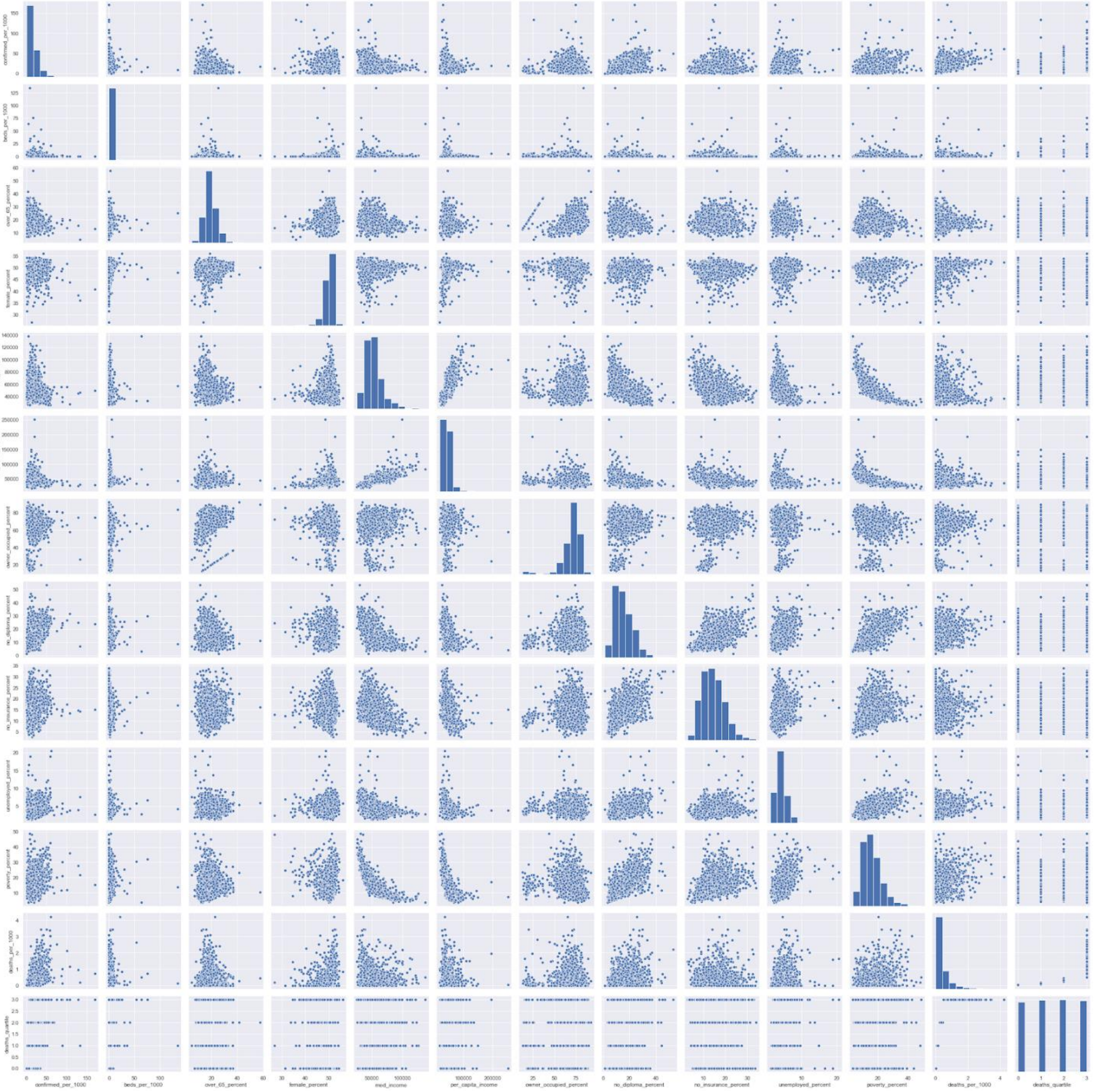
S1a: Default Pairplot for Looking at Features when Analyzing Confirmed Cases Per 1000



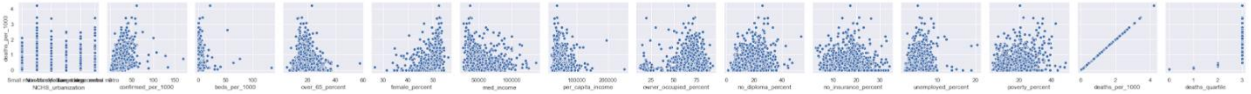
S1b: Pairplot Just Looking at Cases Per 1000 Row



S2a: Default Pairplot for Looking at Features when Analyzing Deaths Per 1000

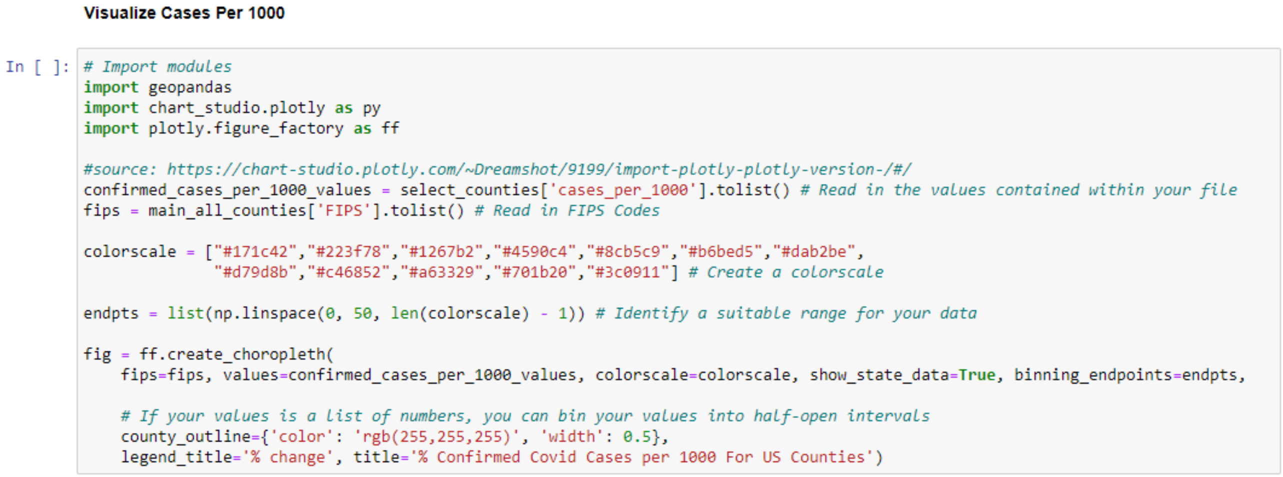


S2b: Pairplot Just Looking at Deaths Per 1000 Row



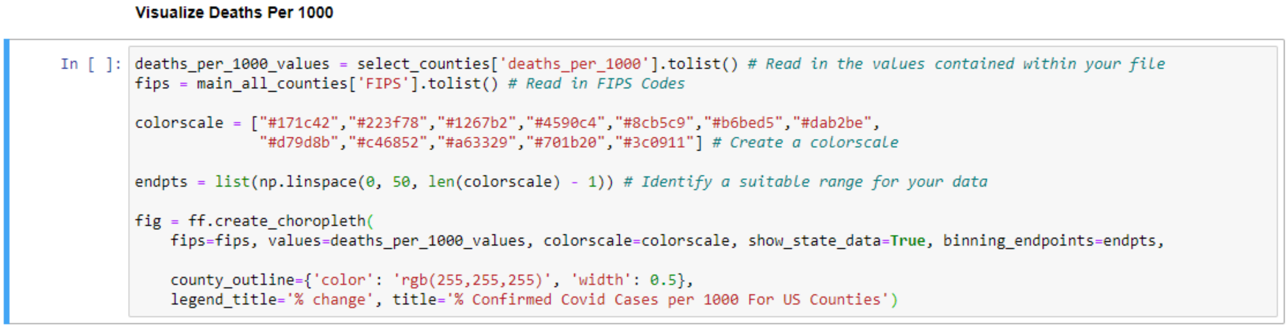
Note:

I was unable to properly input geopandas and its dependencies in niether Jupyter Notebook nor CoLab, so here is the code I would have run to produce my desired plots.



Visualize Cases Per 1000

```
In [ ]: # Import modules
        import geopandas
        import chart_studio.plotly as py
        import plotly.figure_factory as ff

        #source: https://chart-studio.plotly.com/~Dreamshot/9199/import-plotly-plotly-version-/#/
        confirmed_cases_per_1000_values = select_counties['cases_per_1000'].tolist() # Read in the values contained within your file
        fips = main_all_counties['FIPS'].tolist() # Read in FIPS Codes

        colorscale = ["#171c42","#223f78","#1267b2","#4590c4","#8cb5c9","#b6bed5","#dab2be",
                      "#d79d8b","#c46852","#a63329","#701b20","#3c0911"] # Create a colorscale

        endpts = list(np.linspace(0, 50, len(colorscale) - 1)) # Identify a suitable range for your data

        fig = ff.create_choropleth(
            fips=fips, values=confirmed_cases_per_1000_values, colorscale=colorscale, show_state_data=True, binning_endpoints=endpts,

            # If your values is a list of numbers, you can bin your values into half-open intervals
            county_outline={'color': 'rgb(255,255,255)', 'width': 0.5},
            legend_title='% change', title='% Confirmed Covid Cases per 1000 For US Counties')
```



Visualize Deaths Per 1000

```
In [ ]: deaths_per_1000_values = select_counties['deaths_per_1000'].tolist() # Read in the values contained within your file
        fips = main_all_counties['FIPS'].tolist() # Read in FIPS Codes

        colorscale = ["#171c42","#223f78","#1267b2","#4590c4","#8cb5c9","#b6bed5","#dab2be",
                      "#d79d8b","#c46852","#a63329","#701b20","#3c0911"] # Create a colorscale

        endpts = list(np.linspace(0, 50, len(colorscale) - 1)) # Identify a suitable range for your data

        fig = ff.create_choropleth(
            fips=fips, values=deaths_per_1000_values, colorscale=colorscale, show_state_data=True, binning_endpoints=endpts,

            county_outline={'color': 'rgb(255,255,255)', 'width': 0.5},
            legend_title='% change', title='% Confirmed Covid Cases per 1000 For US Counties')
```

I hypothesize that these plots would have looked similar to the JHU plots from the beginning. Although that graph did not plot per 1000 people, I belie that because Covid heavily affects cities, areas with higher rates of Covid prevalence in that graph would have higher cases per 1000 and deaths per 1000 in my graphs.

Works Cited

"American Hospital Directory: Hospital Statistics by State." American Hospital Directory, 29 May 2020. Web. Retrieved 20 Sep 2020. <https://www.ahd.com/state_statistics.html>.

"Certain Medical Conditions and Risk for Severe Covid-19 Illness." *CDC.gov*, U.S. Department of Health & Human Services, 11 Sep 2020. Web. Retrieved 26 Sep 2020. <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html>.

"County-level Data Sets." *Economic Research Service,* United States Department of Agriculture. 13 May 2020. Web. Retrieved 26 Sep 2020. <https://www.ers.usda.gov/data-products/county-level-data-sets/>.

"COVID-19 United States Cases by County." *JHU.edu,* 26 Sep 2020. Web. Retrieved 26 Sep 2020. <https://coronavirus.jhu.edu/us-map>.

"Health Care Maps—Health Insurance" *CDC.gov*, U.S. Department of Health & Human Services, 14 Nov 2020. Web. Retrieved 26 Nov 2020. <https://www.cdc.gov/dhdsp/maps/sd_insurance.htm>.

"Johns Hopkins COVID-19 Case Tracker." *DataWorld.com*, 7 Sep 2020. Web. Retrieved 7 Sep 2020. <https://data.world/associatedpress/johns-hopkins-coronavirus-case-tracker>.

"Personal Income by County, Metro, and Other Areas." *Bureau of Economic Analysis*, United States Department of Commerce, 14 Nov 2019. Web. Retrieved 20 Sep 2020. <https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas>.

"Socioenvironmental Maps—High School Education" *CDC.gov*, U.S. Department of Health & Human Services, 14 Nov 2020. Web. Retrieved 26 Nov 2020. <https://www.cdc.gov/dhdsp/maps/sd_high_school.htm>.

"Socioenvironmental Maps—Poverty" *CDC.gov*, U.S. Department of Health & Human

Services, 14 Nov 2020. Web. Retrieved 26 Nov 2020.

<https://www.cdc.gov/dhdsp/maps/sd_poverty.htm>.

“Socioenvironmental Maps—Unemployment” *CDC.gov,* U.S. Department of Health & Human Services, 14 Nov 2020. Web. Retrieved 26 Nov 2020.

<https://www.cdc.gov/dhdsp/maps/sd_unemployment.htm>.

“United States - Owner Occupied Housing Unit Rate.” *IndexMundi.com,* 2019. Web. Retrieved 25 Nov 2020.

<https://www.indexmundi.com/facts/united-states/quick-facts/all-states/homeownership-rate>.

“United States - Population 65 Years and Over, Percent by State.” *IndexMundi.com,* 2019. Web. Retrieved 26 Sep 2020.

<https://www.indexmundi.com/facts/united-states/quick-facts/all-states/percent-of-population-65-and-over#table>.

“United States - Population Female, Percent by State.” *IndexMundi.com,* 2019. Web. Retrieved 25 Nov 2020.

<https://www.indexmundi.com/facts/united-states/quick-facts/all-states/female-population-percentage#table>.

“What Is a FIPS Code? County-Level Charts in Python.” *Medium.com*, 1 May 2018. Web. Retrieved 7 Sep 2020.

<medium.com/plotly/what-is-a-fips-code-county-level-charts-in-python-4eff383a4cf6>.

This paper represents my own work in accordance with University Policy.

-Temitope Oshinowo