COVID-19 Analysis

Principal Investigators: Project Group 29
Wally Estenson (westenson@wisc.edu), Michael Marotta (mmarotta@wisc.edu)

1. Abstract

The novel COVID-19 virus has drastically altered the lives of the human race in ways most deemed unimaginable. In the United States alone, over 1 million cases have been detected and upwards of 60,000 people have succumbed to the virus [1]. Thus, we decided to delve into an analysis of COVID-19 deaths at the county level in the United States. We wanted to determine what variables lead to higher deaths in counties across the U.S.. After our analysis, we found that race, internet/electronics, and age/health-care status were directly correlated with COVID-19 deaths. However, we concluded that our regression model could not be used to accurately predict deaths.

2. Analysis

2.1. Data

In order to conduct our analysis on COVID-19 related deaths in the United States, we utilized several reliable data sets. Our primary data source is the Johns Hopkins COVID-19 github repository [3]. This source is largely recognized as the most comprehensive and widely used data set in the world. It pulls directly from WHO and CDC data sources. Our other main data source is the US Census [4] which provides a vast array of demographic data by county level in the US.

2.1.1. Data Cleaning and Inclusion

For the scope of our project, we utilized the COVID-19 data by U.S. county. We decided to focus mainly on deaths due to the lack of testing and the high likelihood that current case estimates are gross misrepresentations of actual infections. We also decided to solely focus on the U.S. to narrow the scope of our project, avoid discrepancies in data between countries, and take advantage of widely available demographic data provided by the US Census.

Our next major decision on data inclusion came when we discovered how much of an outlier New York County was in our data. New York County is clearly the epicenter of the U.S. coronavirus outbreak, but after considering the goals of our project, we deemed it necessary to exclude the county from our analysis. Our primary objectives are to understand and predict which variables make counties more susceptible to the virus's spread. Due to the extreme nature of New York, New York's situation, it would not be helpful to include it in our analysis. On a similar note, we decided to exclude all counties with zero deaths. They may have provided us some insights into what prevents the spread of the virus, but in our preliminary analysis, we found that they simply muddled the data and were not useful in our prediction models.

2.2. Preliminary Analysis

During our preliminary analysis of the data, we explored direct relationships between the US Census data that we gathered and the COVID-19 data. For our comparisons, we largely utilized the metric of "Deaths per 100,000" people in order to normalize population differences between counties. We believe

this metric to be the most useful in analyzing the impact on counties, because it is directly related to hospital capacities within counties. In other words, absolute deaths is not as helpful as a metric because larger counties will likely experience more deaths but also likely have a proportionally higher hospital capacity. The impact of this decision to focus on deaths per 100,000 can be seen at a high level by comparing Figure 1 and Figure 2. Counties with smaller absolute deaths but a high number of cases per capita, like Orleans, are brought into the picture as a result.

Figure 1: Deaths Per County

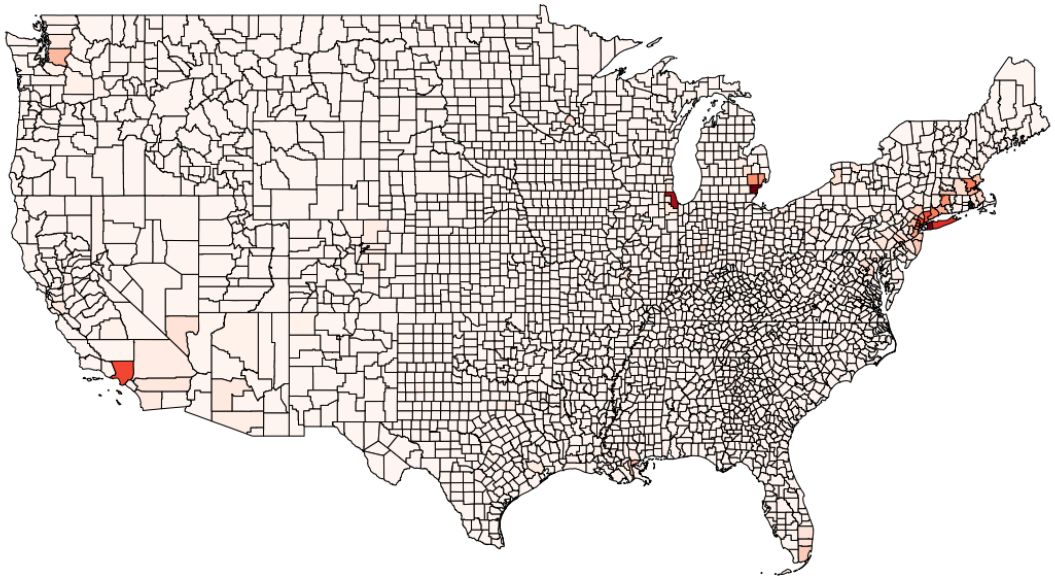
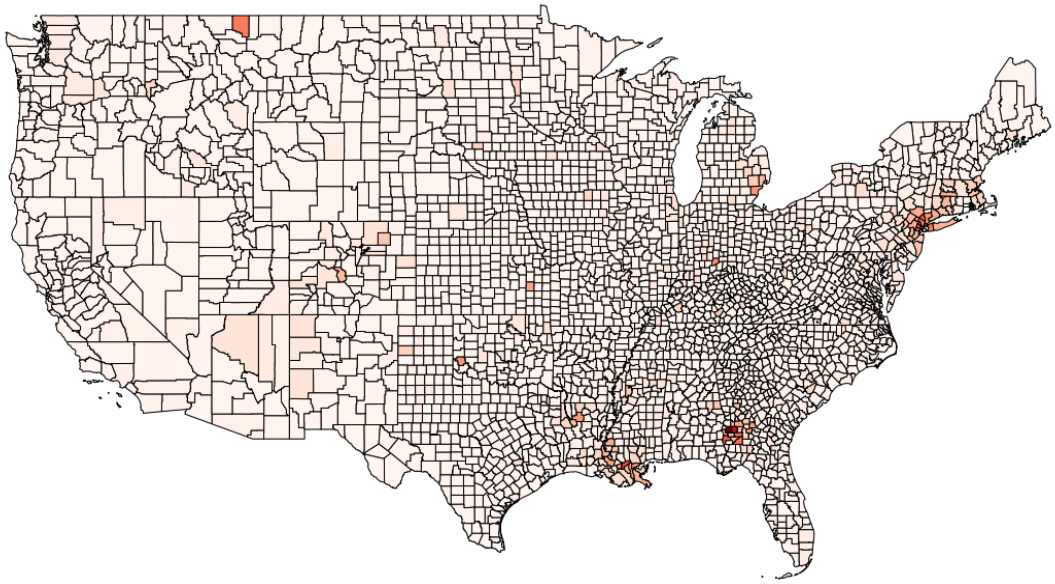


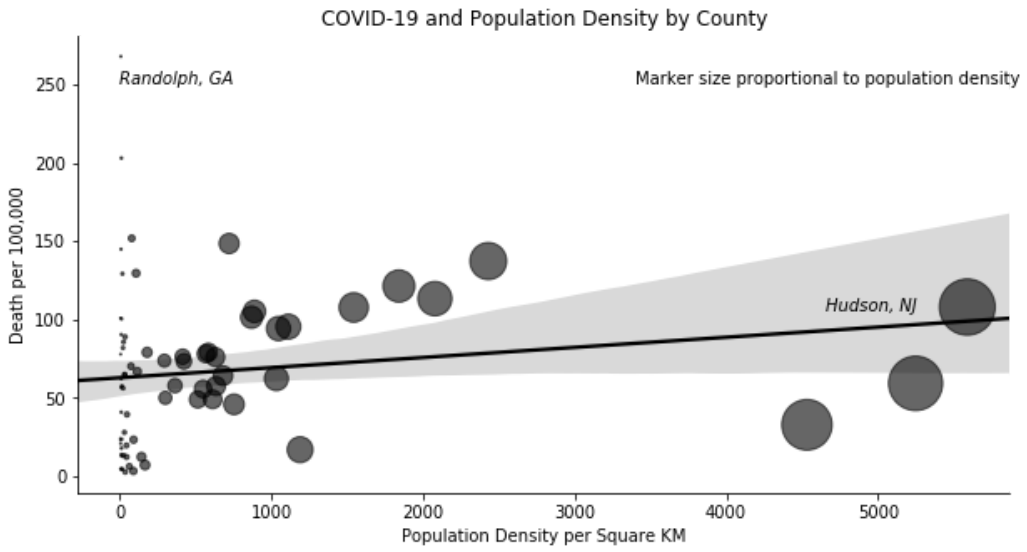
Figure 2: Deaths Per 100,000 People by County



2.2.1. Density

The first metric we studied was population density. Figure 3 below shows that there may be less of a relationship than one would expect. Although there does seem to be some positive correlation, when comparing with normalized death counts, the relationship is smaller than we expected.

Figure 3



2.2.2. Race

Next, we studied the relationship between race and COVID-19 deaths. Our results were quite striking. We found that there is a clear negative correlation between the percentage of white population and deaths per 100,000 people [Figure 4]. On a similar note, we found a positive relationship between counties with higher black populations and deaths. It was also interesting to learn that the black counties most impacted tended to have lower population densities [Figure 5]. The relationships of Asian and Latino/Hispanic communities were not as clear [Figure 6, Figure 7].

Figure 4

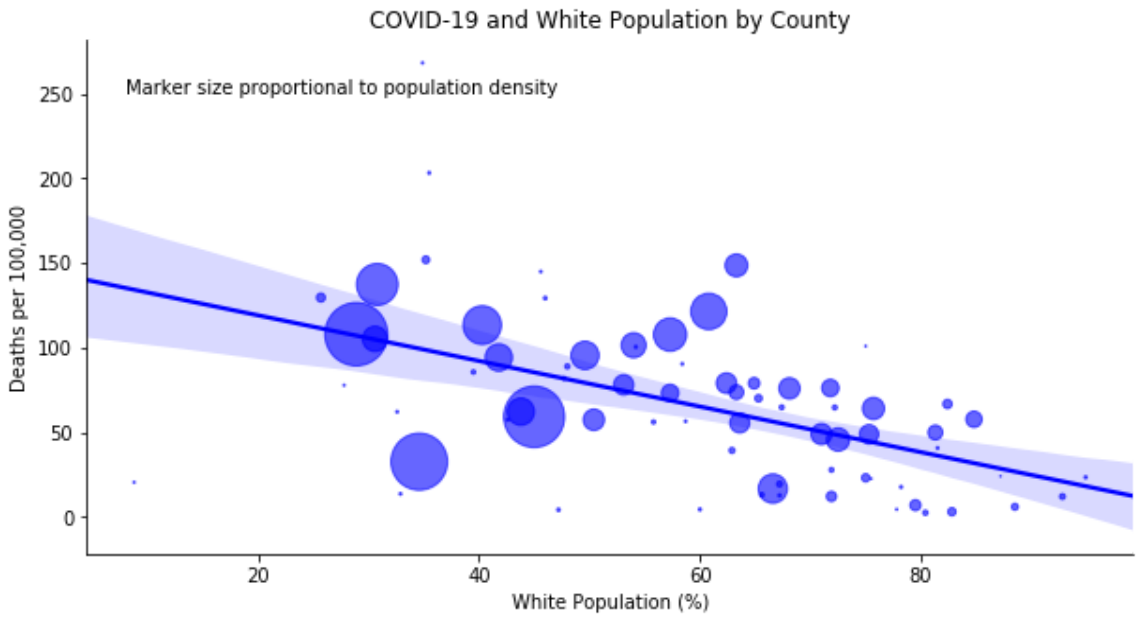


Figure 5

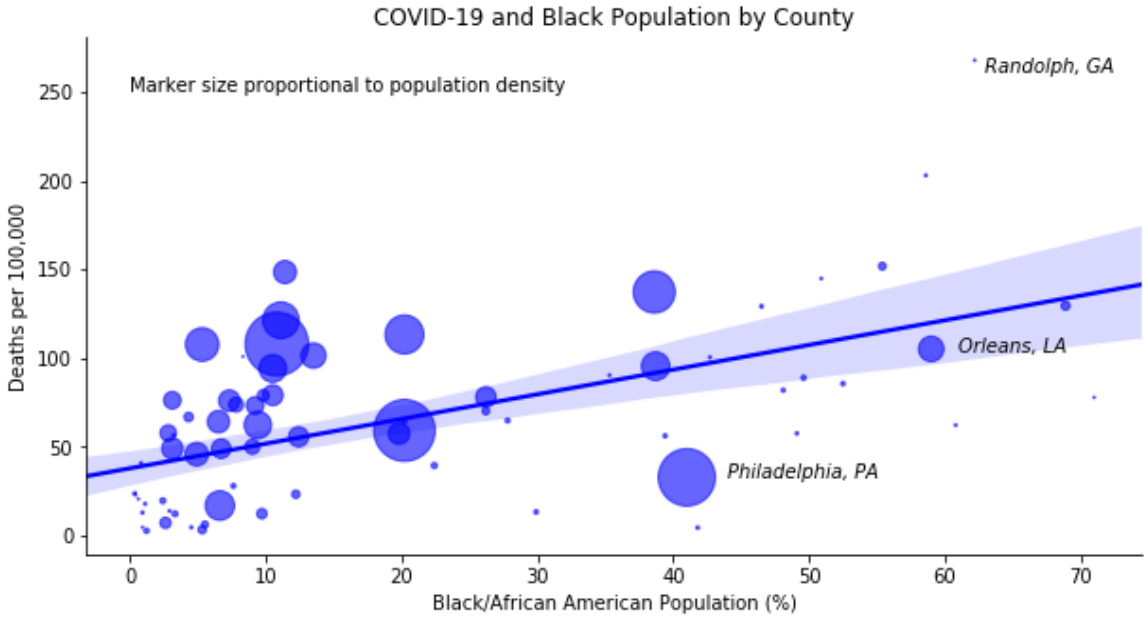


Figure 6

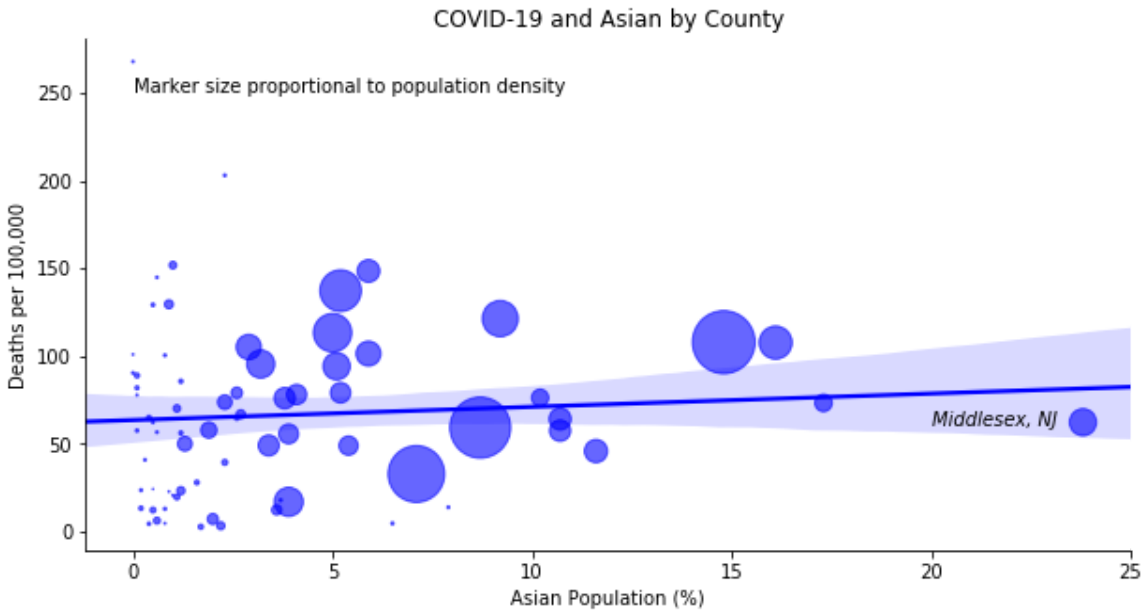
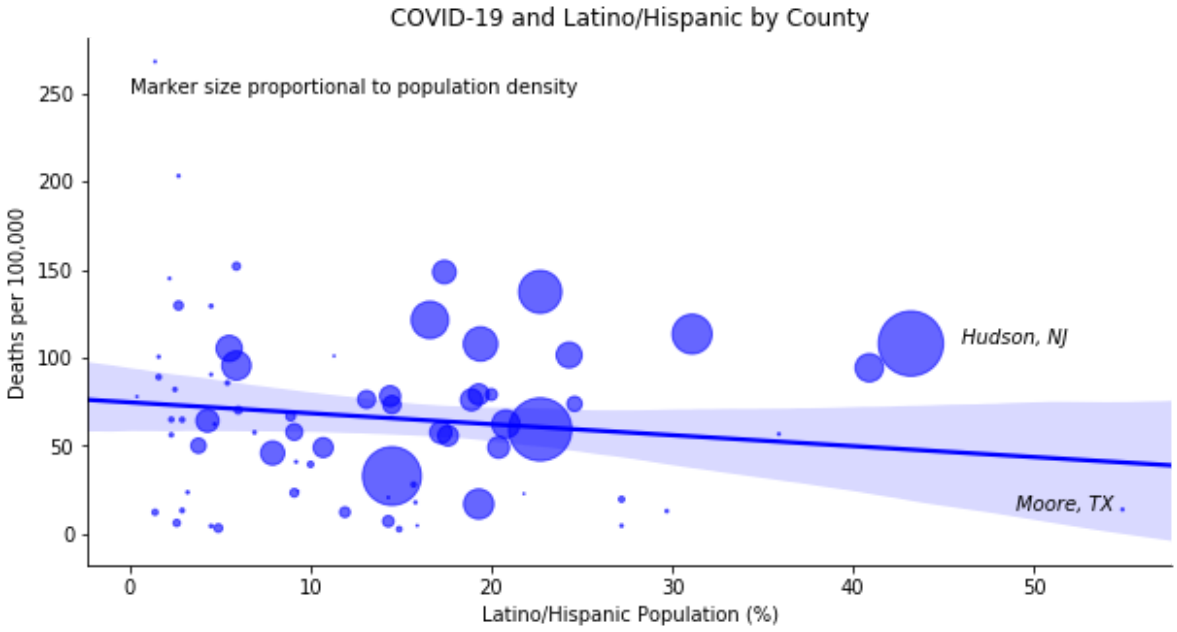


Figure 7



2.2.3. Housing

Before our analysis, we expected poorer counties to be disproportionately impacted by the virus. However, this was not the case. We found no relationship with these variables [Figure 8, Figure 9].

Figure 8

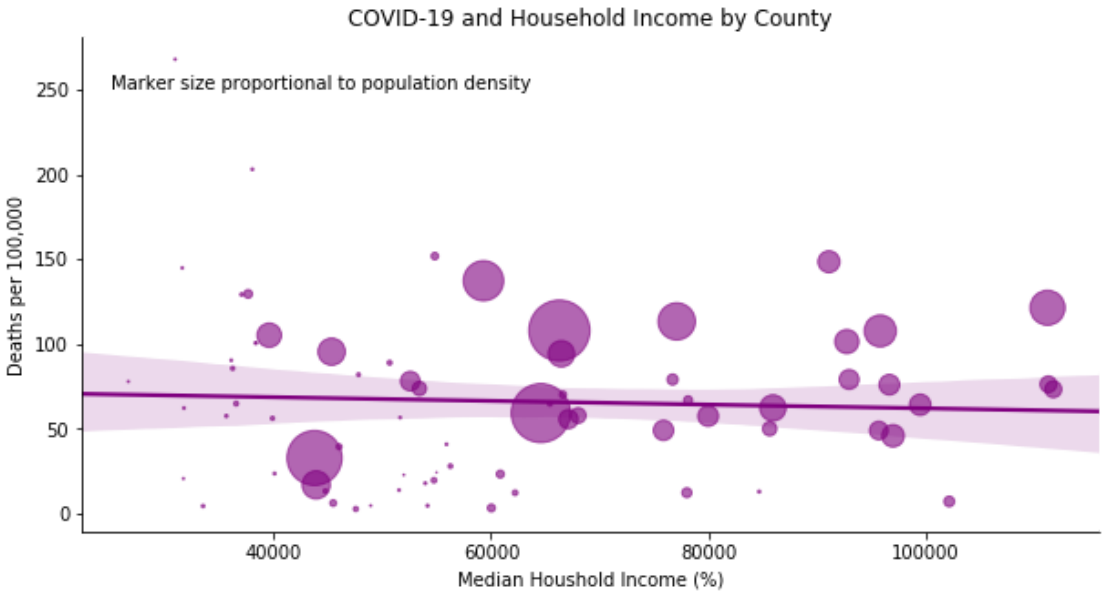
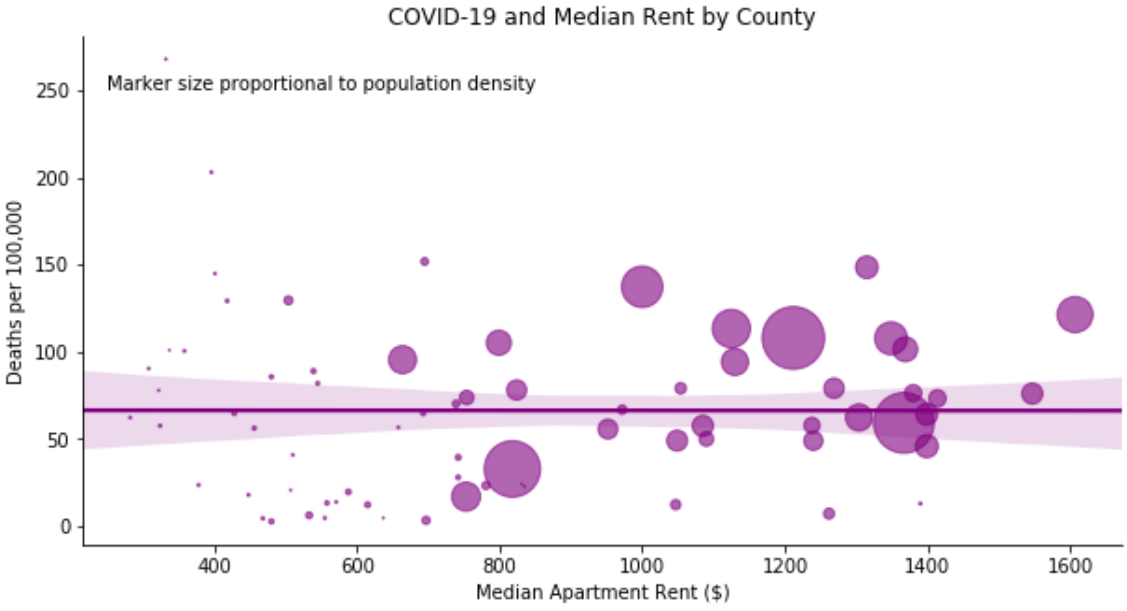


Figure 9



2.2.4. Internet and Electronics

We found an interesting relationship between counties lacking internet and increased normalized death rates [Figure 10]. We hypothesize that this may be due to lack of awareness by individuals in the community. We found a similar relationship with household electronics [Figure 11].

Figure 10

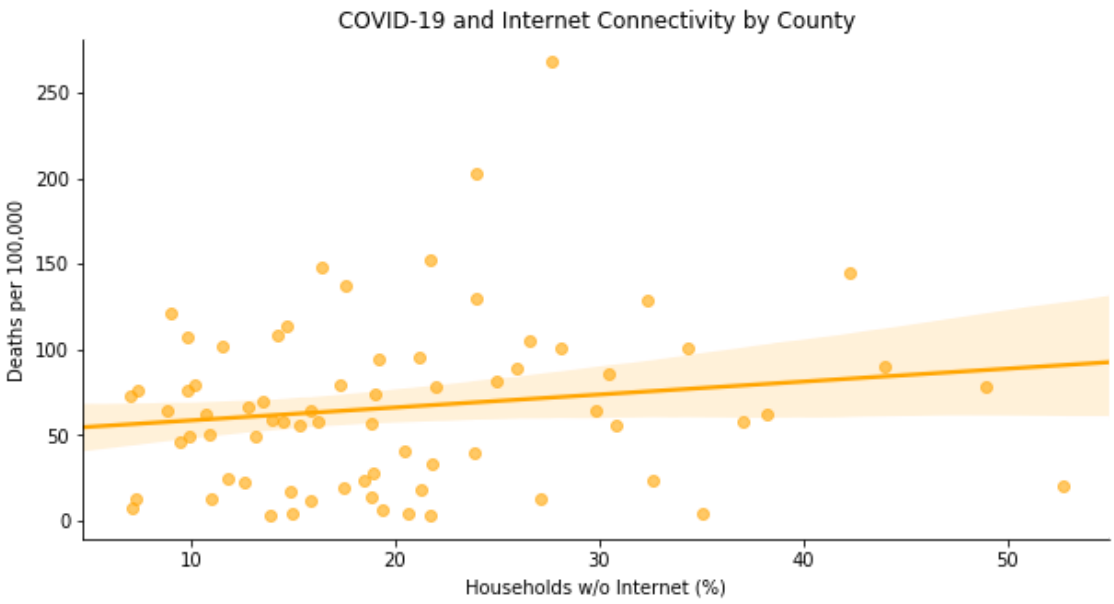
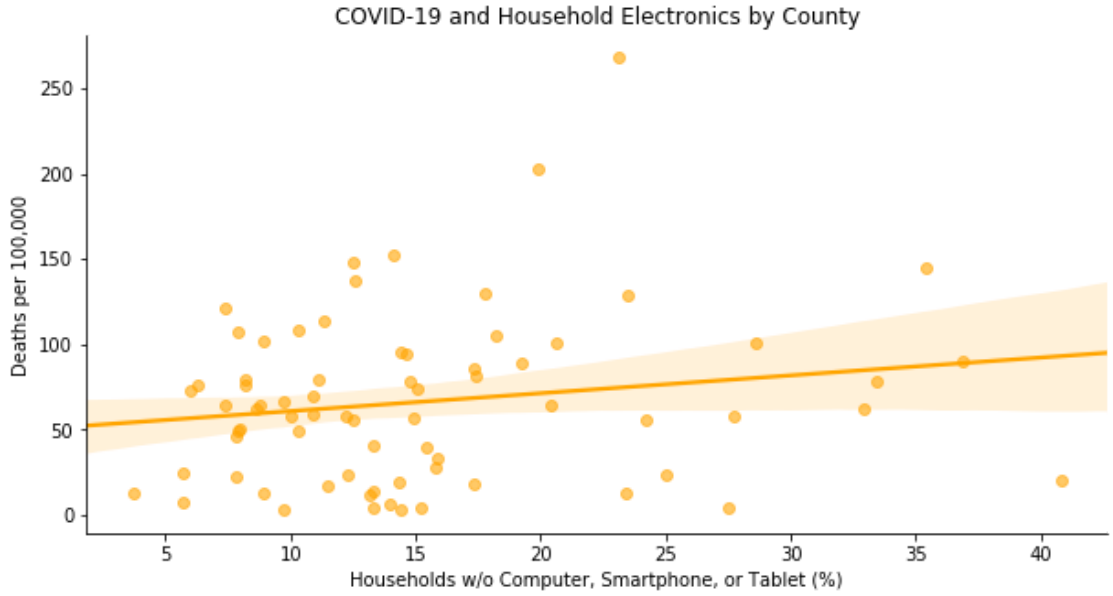


Figure 11



2.2.5. Health Care and Age

In the final part of our preliminary analysis, we found that a higher percentage of older people within communities had led to higher deaths [Figure 12]. This was expected as it is currently a highly studied subject. Next, we looked at health-care coverage. We first looked at the relationship with "Deaths per 100,00" and found that it was positive. However, we looked further into "Death Rates" [Figure 13] and found an even stronger relationship between the variables, especially for individuals between the ages of 19 and 34 [Figure 14]. We expect that those without health insurance may be less healthy than those with insurance and/or less willing to seek help in the early stages of an illness, leading to greater chance of death.

Figure 12

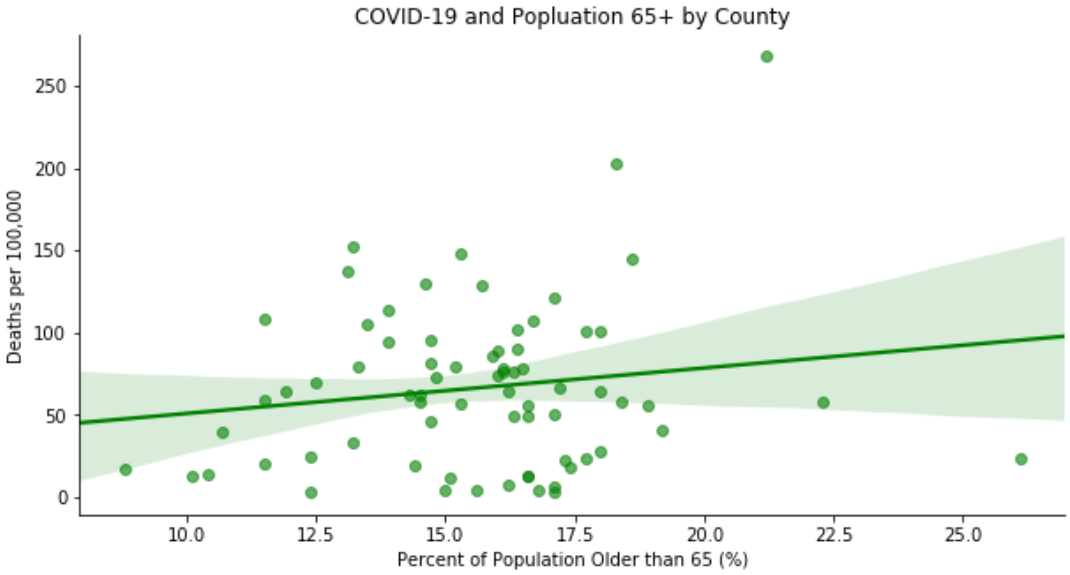


Figure 13

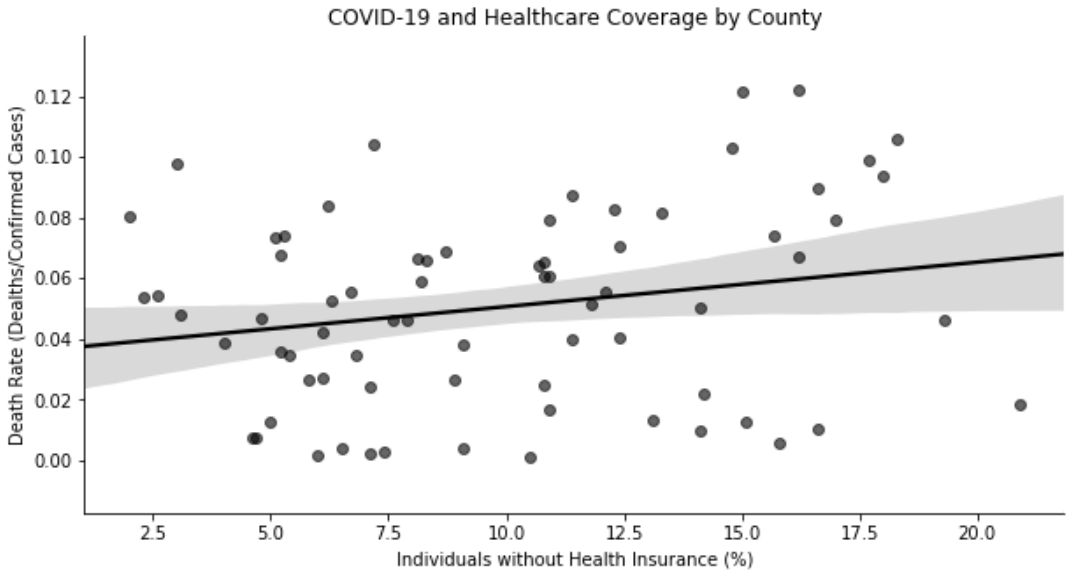
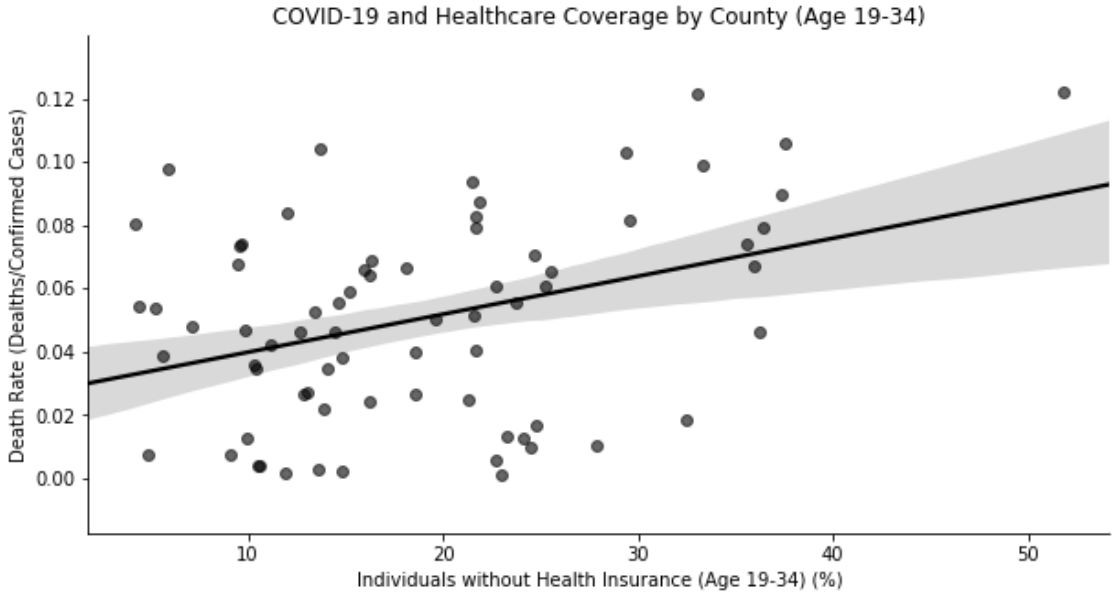


Figure 14



2.3. Regression Analysis

After analyzing the graphs above, we decided to run a regression analysis to see if we could accurately predict the number of deaths in a county. We removed collinear variables that number of deaths may be dependent such as number of cases and death rate. As seen in Table 1, the following variables led to lower Deaths in the county due to negative coefficient values: Males, PCT65, Occupied_Housing_Units, Home_Ownership_Rate, Median_Rent, and Pct_NoCoverage. The rest of the variables were positive. When we use a statistical significance level of .05, only the following variables seem to be significant: Population, Pct_Asian, Housing_Units, Occupied_Housing_Units, and Density.

Table 1

OLS Regression Results

Dep. Variable:	Deaths	R-squared:	0.759
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	47.08
Date:	Fri, 01 May 2020	Prob (F-statistic):	3.71e-76
Time:	15:19:07	Log-Likelihood:	-1870.3
No. Observations:	304	AIC:	3781.
Df Residuals:	284	BIC:	3855.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-233.0475	176.960	-1.317	0.189	-581.367	115.272
Population	0.0018	0.000	8.310	0.000	0.001	0.002
Males	-0.0007	0.001	-0.522	0.602	-0.003	0.002
Females	0.0025	0.001	1.849	0.065	-0.000	0.005
PCT65	-5.5308	5.364	-1.031	0.303	-16.088	5.027
Median_Age	6.1991	4.539	1.366	0.173	-2.734	15.133
Pct_White	1.1687	1.268	0.922	0.357	-1.326	3.664
Pct_Black	0.7529	1.166	0.646	0.519	-1.543	3.048
Pct_Asian	8.6190	4.195	2.055	0.041	0.362	16.876
Pct_Latino	1.3692	1.573	0.871	0.385	-1.726	4.465
Household_Income	0.0005	0.001	0.332	0.740	-0.002	0.003
Housing_Units	0.0027	0.001	3.001	0.003	0.001	0.004
Occupied_Housing_Units	-0.0092	0.001	-7.435	0.000	-0.012	-0.007
Home_Owernship_Rate	-1.6573	1.498	-1.106	0.269	-4.606	1.291
Median_Rent	-0.0791	0.092	-0.855	0.393	-0.261	0.103
Median_Home_Value	0.0004	0.000	1.867	0.063	-1.96e-05	0.001
Pct_NoComp	1.8072	3.668	0.493	0.623	-5.413	9.028
Pct_NoInternet	1.0208	3.112	0.328	0.743	-5.105	7.147
Pct_NoCoverage	-2.9230	4.360	-0.670	0.503	-11.505	5.659
Pct_NoCoverage_19to34	1.2197	2.070	0.589	0.556	-2.854	5.294
Density	0.0464	0.016	2.824	0.005	0.014	0.079

Omnibus:	215.300	Durbin-Watson:	2.089
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6157.719
Skew:	2.441	Prob(JB):	0.00
Kurtosis:	24.501	Cond. No.	5.18e+15

The next regression that we ran had the same independent and dependent variable as above, but instead of removing counties that had less than 15 deaths per 100,000 people, we removed counties that had cases lower than 750 cases per 100,000 people to see if we'd have better results. As seen in Table 2**,** we got different coefficient values and a higher R-squared value. The following variables led to lower Deaths in the county due to negative coefficient values: Pct_White, Occupied_Housing_Units, Home_Ownership_Rate, Median_Rent, Pct_NoComp, and Pct_NoCoverage_19to34. When we use a statistical significance level of .05, only the following variables seem to be significant: Population and Occupied_Housing_Units.

Table 2

OLS Regression Results

Dep. Variable:	Deaths	R-squared:	0.858
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	20.00
Date:	Fri, 01 May 2020	Prob (F-statistic):	8.69e-20
Time:	17:36:34	Log-Likelihood:	-525.92
No. Observations:	83	AIC:	1092.
Df Residuals:	63	BIC:	1140.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-269.0064	452.809	-0.594	0.555	-1173.872	635.859
Population	0.0022	0.000	4.451	0.000	0.001	0.003
Males	0.0015	0.004	0.411	0.683	-0.006	0.009
Females	0.0007	0.004	0.172	0.864	-0.007	0.008
PCT65	3.8954	15.283	0.255	0.800	-26.644	34.435
Median_Age	4.8250	12.541	0.385	0.702	-20.236	29.886
Pct_White	-0.0504	3.093	-0.016	0.987	-6.231	6.130
Pct_Black	1.8182	2.604	0.698	0.488	-3.385	7.021
Pct_Asian	0.1272	7.393	0.017	0.986	-14.647	14.901
Pct_Latino	1.4887	3.173	0.469	0.641	-4.852	7.830
Household_Income	0.0047	0.004	1.199	0.235	-0.003	0.012
Housing_Units	0.0023	0.002	1.192	0.238	-0.002	0.006
Occupied_Housing_Units	-0.0095	0.003	-2.917	0.005	-0.016	-0.003
Home_Owernship_Rate	-3.9246	3.652	-1.075	0.287	-11.222	3.373
Median_Rent	-0.3113	0.245	-1.270	0.209	-0.801	0.179
Median_Home_Value	0.0004	0.000	0.896	0.373	-0.000	0.001
Pct_NoComp	-10.1507	10.573	-0.960	0.341	-31.279	10.978
Pct_NoInternet	9.9999	9.700	1.031	0.307	-9.384	29.383
Pct_NoCoverage	8.0486	12.181	0.661	0.511	-16.293	32.390
Pct_NoCoverage_19to34	-2.1251	5.511	-0.386	0.701	-13.137	8.887
Density	0.0176	0.030	0.585	0.560	-0.043	0.078

Omnibus:	19.011	Durbin-Watson:	2.110
Prob(Omnibus):	0.000	Jarque-Bera (JB):	108.810
Skew:	-0.195	Prob(JB):	2.36e-24
Kurtosis:	8.596	Cond. No.	1.52e+16

2.3.1. Regression Analysis Limitations

Based on our results from our preliminary analysis and general information known about the virus, the second regression seems to be more accurate. For example, we have seen that minorities have been disproportionately affected by the virus. In addition, the older you are, the more likely you are to die from the virus. This is best represented from the regression in Table 2.

Although our regressions have relatively high R-squared values, the results are not very promising. Most of the variables are not statistically significant to relate to the rest of the total population. We hypothesise that there are many different factors not included in the scope of this report that impact COVID-19 death counts and case rates. One example of this is the differing policy taken up by States. We expect this to have been a significant factor in the spread of the virus so far but this variable was not included in our analysis.

3. Conclusions and directions for future research

COVID-19 has altered many things in our lives but has highlighted the importance of data analytics to "see the unseen." This report attempted to determine which demographics lended to higher mortalities in counties in the United States. Some interesting findings were based on population, race, internet/electronics, and age/health-care status. Surprisingly, we did not find a strong relationship between population density and deaths; there was a weak positive relationship [Figure 3]. We found a clear negative correlation between the percentage of white population and deaths per 100,000 people [Figure 4] and a positive relationship between counties with higher black populations and deaths [Figure 5]. In addition, we found that counties lacking internet access had increased death rates [Figure 10]. This may be due to an absence of knowledge about the virus due to lack of connectedness to the world. Further, we found that those without health insurance tended to have higher death rates. This may be because those without health insurance are less willing to seek help in the early stages of an illness, leading to greater chance of death. The results of our regressions were not as promising as we hoped, which is likely due to the countless other variables that affect the spread of a virus not included in our analysis. Going forward, it would be very interesting to study how the regression changes as the data continues to be updated. Finally, when this pandemic has come to its close, we can compare how our results fared by comparing current data and future data.

4. References

[1] CDC
https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

[2] NYT
https://www.nytimes.com/2020/04/05/us/coronavirus-deaths-undercount.html

[3] Johns Hopkins
https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv

[4] US Census Demographics
https://covid19-uscensus.hub.arcgis.com/

[4.1]General Population
https://covid19-uscensus.hub.arcgis.com/datasets/acs-total-population-county?geometry=-131.696%2C-0.672%2C132.327%2C76.524

[4.2]Housing
https://covid19-uscensus.hub.arcgis.com/datasets/acs-highlights-population-housing-basics-county-centroids?geometry=110.373%2C26.199%2C17.385%2C65.984

[4.3]Internet Connectivity
https://covid19-uscensus.hub.arcgis.com/datasets/acs-internet-connectivity-county?geometry=-133.190%2C28.795%2C133.821%2C67.148

[4.4]Health Insurance Coverage
https://covid19-uscensus.hub.arcgis.com/datasets/acs-health-insurance-coverage-county?geometry=-133.190%2C28.795%2C133.821%2C67.148

[5] US Census County Geometry
https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html