Why Do Some Counties in the US Suffer More from COVID-19?

Rachel Adenekan
Mechanical Engineering Department
Stanford University
Stanford, California
adenekan@stanford.edu

Usman Khaliq
Mechanical Engineering Department
Stanford University
Stanford, California
usmank@stanford.edu

***Abstract*—Coronavirus (COVID-19) has affected most of the United States. The burden that COVID-19 poses to varies greatly across states and particularly across different counties. Various factors including population density, racial composition, income disparities, healthcare capacity, are cited as possible sources of this variation. However, many studies have not released detailed information on how much these various features contribute to possible COVID-19 burden as indicated by case and death statistics. In this paper, both a qualitative and quantitative analysis are completed in order to better understand the features that result in high COVID-19 burden. Upon developing and analyzing several linear regression models, we find that race, specifically percentage of Blacks in a county, is the single-most important predictor of COVID-19 burden. After race, population density was the most significant predictor.**

***Index Terms*—COVID-19, coronavirus, risk factors, counties**

I. Introduction

With more than 2 million confirmed COVID-19 cases[1], and more than 115,000 deaths, the United States has the unenviable record of witnessing the world's biggest COVID-19 outbreak. However, the spread of the virus has not been distributed equally throughout the country. States such as New York, California, Seattle and New Jersey witnessed the initial peaks in infections and deaths, other states in the south, such as Florida, Texas, Georgia and others are now facing an upward trend in the number of infections and deaths. Apart from the divergences in the spread of the virus in different parts of the country, another concerning trend that is becoming more prominent is that it seems like there are racial and ethnic disparities in terms of which populations are getting disproportionately affected by COVID-19. In this paper, we want to explore which factors are responsible for possible COVID-19 burden, as indicated by the number of cases and deaths in different counties in the US.

II. Motivation

A growing body of research is highlighting how COVID-19 is disproportionately affecting certain populations in the United States[2]. Research from the Center for Disease Control(CDC) has shown that there is a disproportionate burden of illness and deaths among racial and ethnic minority groups. According to the CDC, in New York City, 33 percentage of hospitalized patients were Black though they only comprise 18 percent in the general population[3]. Research from ProPublica has shown that the COVID-19 outbreak in Wisconsin was concentrated in Milwaukee, where 39 percentage of the population is Black.[4] Similarly, in Michigan, where 14 percent of the population is Black, Blacks make up 35 percent of total COVID-19 cases and 40 percent of COVID-19 related deaths. These statistics suggest that racial and structural inequalities that exist in the healthcare system in the US may be causing disparities in COVID-19 burden. Our aim was to further explore which features are contributing to the spread of COVID-19 and the number of deaths at the county level in the US, with the aim of contributing to conversations on how some of these systemic inequalities might be fixed.

III. Related Work

A number of tools and dashboards are being developed that aim to quantify a ”risk score” to different counties based on several factors. For instance, the COVID-19 Vulnerability Index Map[5] developed by Conduent displays a ”risk score” for each county in the US based on trends in reported COVID-19 and deaths, clinical risk factors, and social and economic determinants. However, for most of these tools, including the one mentioned above, the methodology on how this score was calculated isn't publicly available. Furthermore, several counties that were predicted as low risk, have very high case and death numbers. This indicates that there is likely some error in the method of developing ”risk score.” For this paper, we wanted to explore publicly available datasets and share our methods for investigating which features might be contributing to the difference in the burden of COVID-19 in different counties in the US.

IV. Datasets

For this paper, we explored the following variables from these datasets[6]:

- Population Density from the US Census Bureau Population Distribution in Counties 2018, US Census Bureau County Land Areas 2011
- Socioeconomic status from the US Census Bureau Income and Benefits in Counties 2018
- Race and ethnicity data from the US Census Bureau Race and Ethnicity Distribution in Counties 2018
- Healthcare Capacity at County Level from the Definitive Healthcare: US Hospital Beds June 2020
- Total COVID-19 Case and Death Count at US County Level from the New York Times

V. Methods

We conducted a qualitative exploratory data analysis to examine trends between county-level COVID-19 cases and deaths in relationship to population density, socioeconomic status, race, and healthcare capacity. For each of these features, we created county-level chloropleth maps for the feature and compared them to county-level chloropleth maps of COVID-19 case (and/or) numbers. To gain more quantitative insights, we then developed linear regression models that take the features described above as inputs and outputs COVID-19 cases (or deaths). The goal was to understand which features are more important in predicting COVID-19 case and death numbers. We developed two models, one to predict COVID-19 cases per 100,000 people and one to predict COVID-19 deaths per 100,000 people, as indicated by the equations below:

$$Y_1 = a_0 + a_1 * x_1 + a_2 * x_2 + a_3 * x_3 \tag{1}$$

$$Y_2 = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * x_4 \tag{2}$$

where:

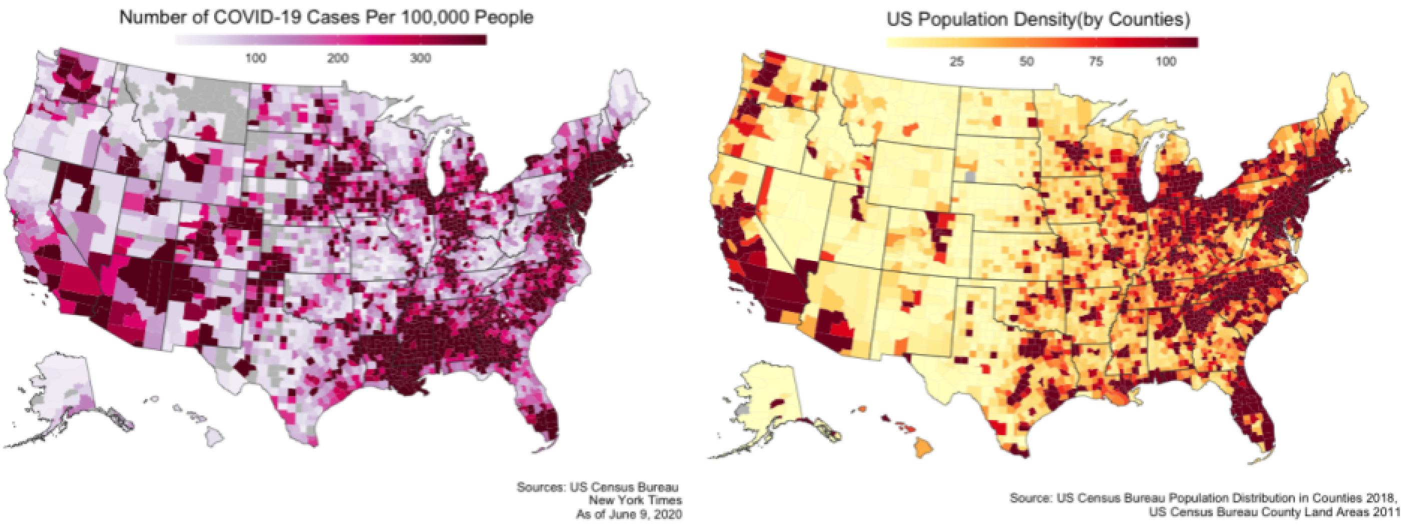
Y_1 = COVID-19 cases per 100,000 people
Y_2 = COVID-19 deaths per 100,000 people
a = feature coefficients corresponding to COVID-19 cases
b = feature coefficients corresponding to COVID-19 deaths
x_1 = Population density
x_2 = Percentage of population on food benefits (socioeconomics)
x_3 = Black population percentage (race)
x_4 = Number of people per hospital bed (healthcare capacity)

VI. Results

A. Exploratory Data Analysis

From our exploratory data analysis, we observed the following trends:

- Although there seems to be a correlation between the population density in counties on the east coast and the number of COVID-19 cases per 100,000 as well as the number of deaths due to COVID-19 per 100,000 people, we can see that this relationship does not hold strongly in some of the other states (see Figure 1 and Figure 2). For instance, in the southern states of Mississippi, Georgia, Alabama and Louisiana, we can see that although the population density in the counties is in the mid range, these states still registered a high outbreak of COVID-19 cases. Similarly, we can also see a similar cluster of outbreaks in some counties in Arizona, Utah and Colorado, although the corresponding counties have very low population densities. These trends hint at the fact that there are other factors apart from population density that are contributing to the spread of COVID-19 in different counties in the US.
- When we visually compared the case load of COVID-19 against the percentage of households in the county that were on food benefits, we observed an interesting pattern(see Figure 3). In some states in the South such as Mississippi, Georgia, Alabama and Louisiana, we can see that counties that had a higher percentage of households on food benefits also recorded higher rates of confirmed COVID-19 cases per 100,000 people. However, this pattern did not hold that well in other states that also had a high percentage of households on food benefits. For instance, some counties in West Virgina, Kentucky, Oregon and some parts of Missouri had a high percentage of households on food benefits, but they reported moderate levels of COVID-19 spread. This is an interesting trend, and it hints at the fact that there are other factors beyond just the level of poverty, as indicated by food insecurity, that contribute to the spread of COVID-19.
- We then compared the case load of COVID-19 against the percentage of black population in the countries, which led to a striking observation(see Figure 4). In the southern states of Louisiana, Mississippi, Alabama, and Georgia, we can see that in counties where more than 40 percentage of the people were Black, the number of COVID-19 cases was higher than that in other counties that had a lower percentage of people that were Black. We can also observe this trend in other states such as North and South Carolina and in some counties in Virginia. This observation strongly hints at the fact that COVID-19 is disproportionately affecting blacks, and is spreading at a higher rate in counties which have a higher proportion of blacks.
- Finally, we compared the case load of COVID-19 against the number of people per hospital bed in the counties, to observe whether the virus was spreading more rapidly in counties with overburdened healthcare systems(see Figure 5). We did not see any observable trends that supported this point of view. We note that this feature also had the most amount of missing data for counties. This may have influenced the lack of an observable trend.



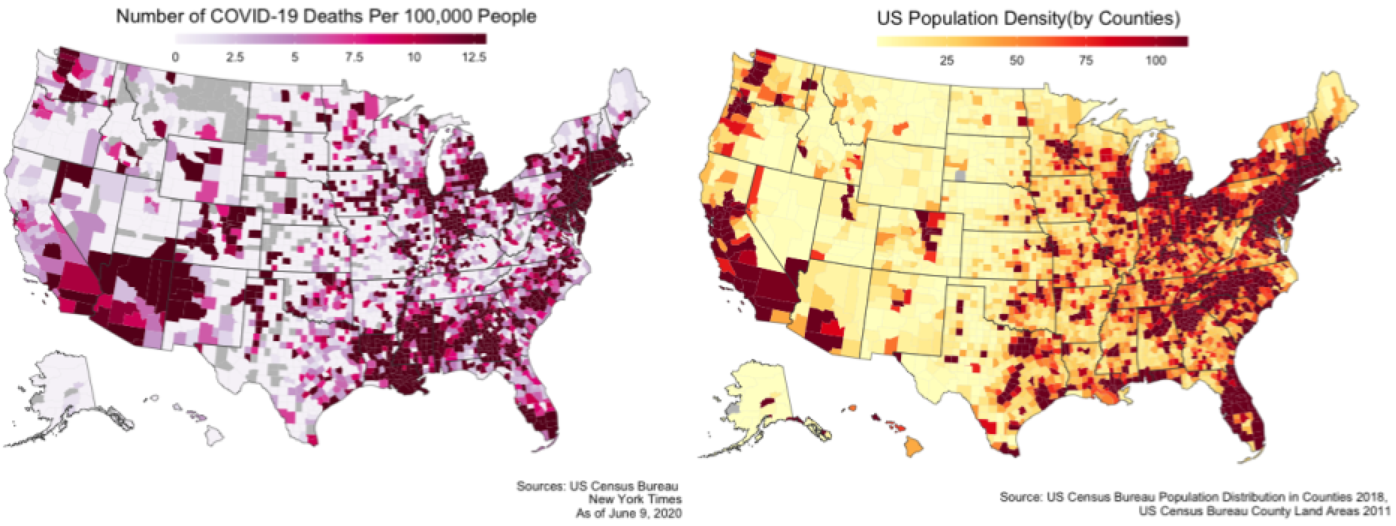
Number of COVID-19 Cases Per 100,000 People

Sources: US Census Bureau
New York Times
As of June 9, 2020

US Population Density(by Counties)

Source: US Census Bureau Population Distribution in Counties 2018,
US Census Bureau County Land Areas 2011

Fig. 1. Chloropleth maps comparing population density to COVID-19 Cases Per 100,000 People. Population Density alone does not appear to explain the disparities in county-level COVID-19 cases.



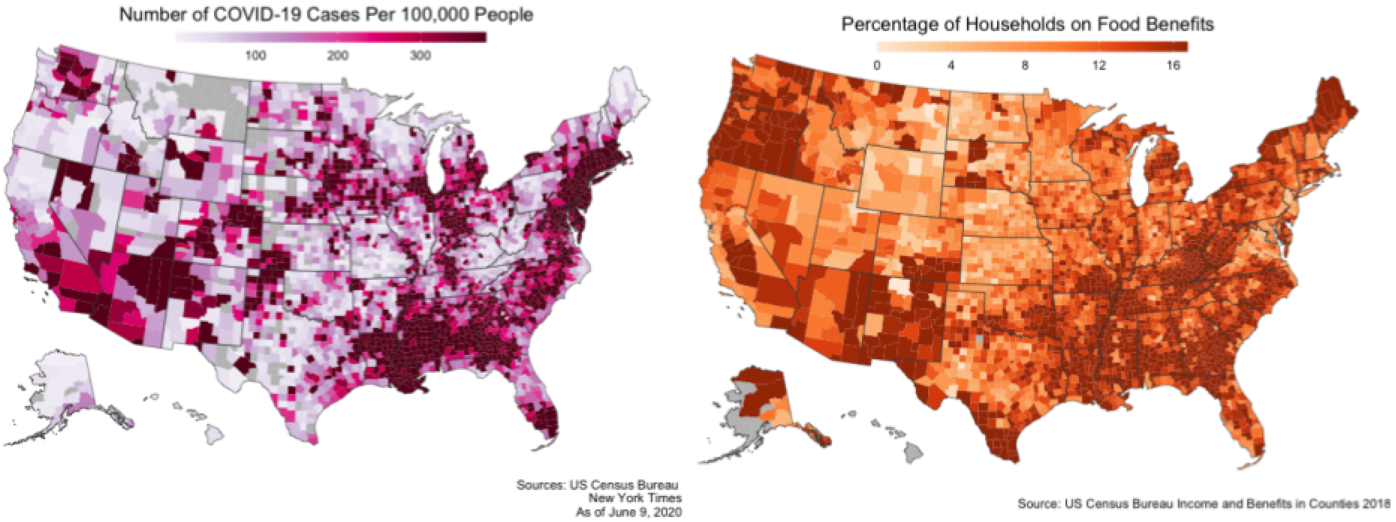
Number of COVID-19 Deaths Per 100,000 People

Sources: US Census Bureau
New York Times
As of June 9, 2020

US Population Density(by Counties)

Source: US Census Bureau Population Distribution in Counties 2018,
US Census Bureau County Land Areas 2011

Fig. 2. Chloropleth maps comparing population density to COVID-19 Deaths Per 100,000 People. Population Density alone does not appear to explain the disparities in county-level COVID-19 deaths.



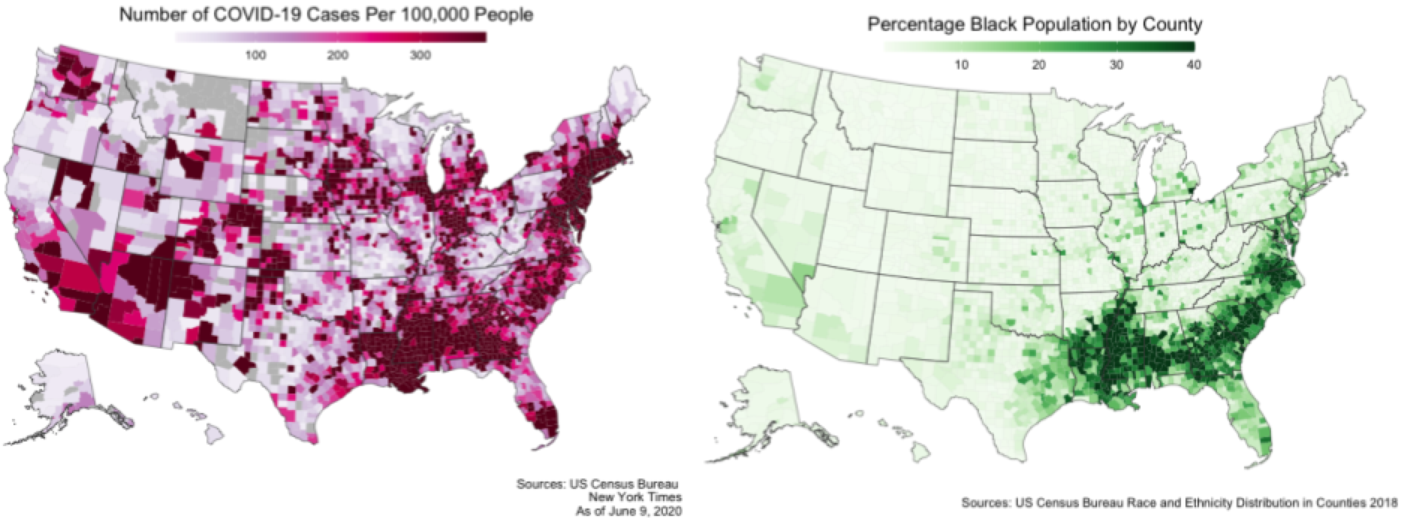
Number of COVID-19 Cases Per 100,000 People

Sources: US Census Bureau
New York Times
As of June 9, 2020

Percentage of Households on Food Benefits

Source: US Census Bureau Income and Benefits in Counties 2018

Fig. 3. Chloropleth maps comparing percentage of households on food benefits to COVID-19 Cases Per 100,000 People.



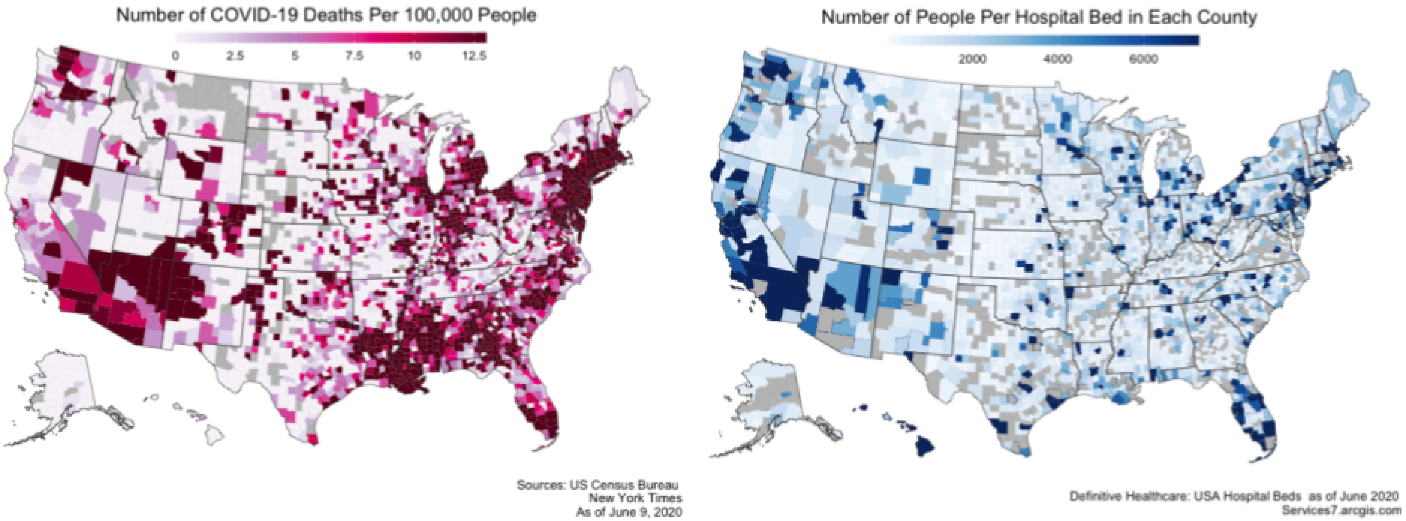
Number of COVID-19 Cases Per 100,000 People

Sources: US Census Bureau
New York Times
As of June 9, 2020

Percentage Black Population by County

Sources: US Census Bureau Race and Ethnicity Distribution in Counties 2018

Fig. 4. Chloropleth maps comparing percentage of population that is Black to COVID-19 Cases Per 100,000 People.



Number of COVID-19 Deaths Per 100,000 People

Sources: US Census Bureau
New York Times
As of June 9, 2020

Number of People Per Hospital Bed in Each County

Definitive Healthcare: USA Hospital Beds as of June 2020
Services7.arcgis.com

Fig. 5. Chloropleth maps comparing number of people per hospital bed to COVID-19 Cases Per 100,000 People.

B. Linear Regression Model

All of our linear regression models has pretty poor fits to the data as indicated by our low R^2 values. However, we were still able to gain insights regarding which features are more correlated with higher COVID-19 case and death numbers.

From our linear regression models that were performed over all the US counties (see Table 1 and Table 2), we observed the following:

Percentage of blacks in the population is the most predictive feature in identifying high case and death numbers. The higher this percentage, the higher the case and death numbers. Population density was the next most predictive feature. Intuitively higher populations densities were associated with higher case and death numbers. Surprisingly, neither socioeconomic data, as indicated by dependence on food benefit income, nor healthcare capacity, as indicated by number of people per hospital bed, were significant in predicting COVID-19 case or death burden in the counties.

In an effort to possibly improve our model fit, we then decided to perform linear regressions over all counties in particular regions of the United States, as opposed to the entire United Sates. The two regions we chose to investigate through these region specific models were the Southern US counties (see Table 3 and Table 4) and the East Coast US counties (see Table 5 and Table 6). While this improved model fit in some cases, the model fit was still pretty poor as evidenced by the low R^2 values. However the results were quite consistent. For both the Southern and East Coast US Counties, race, specifically percentage of Blacks in the county, was the most significant coefficient in predicting COVID-19 case numbers.

Some differences however did emerge. In the Southern counties, percentage of households on food benefit income was the most significant predictor after race, of case number. Population density was not significant in predicting case numbers (see Table 3). However, population density was significant in predicting deaths in Southern counties (see Table 4).

In the East Coast US counties, population density was again the most significant feature after race in predicting case numbers (see Table 5). However, for East Coast counties death load, population density was the most significant feature in predicting deaths. It is the only condition for which, race was not the most predictive feature. However, race was still very significant. In fact, it was the second most predictive feature for predicting East Coast counties death numbers (see Table 6).

VII. Conclusions and Future Work

Although there is still much to be learnt about the spread of COVID-19 across the world and how it spreads differently in various regions, exploratory data analysis on the spread of the virus within counties in the US, along with the various other factors that could have contributed to the spread of the virus in these counties, strongly suggests that race is a very important factor in determining which counties are more affected. A simple linear regression model that determined which features were important in predicting the case burden of COVID-19 also identified race as the single biggest factor, specifically the percentage of blacks in the counties. This is a striking observation, since it is pointing at how the public health system in the US appears to be racially biased. Without undertaking long overdue systemic changes to the healthcare system that can ensure equitable access and quality of care for all people, regardless of their race, it seems like COVID-19 will continue to disproportionately affect minorities, particularly Blacks.

There are several future directions that can be taken from our work. Instead of using the number of confirmed cases and deaths from COVID-19 as a proxy for determining the case load of the virus as a result of different factors, the predicted growth rate in the number of confirmed COVID-19 cases can be used as the outcome variable instead to see whether the linear model gives a better fit. Also, the factors that we have identified in this paper, such as the proportion of blacks in the counties and the number of families on food benefits, appear to be correlated with each other. In any future work on the impact of these factors on the spread of COVID-19, it would be important to isolate these factors and to independently measure what impact these factors have on the spread of the virus.

Acknowledgments

We thank Professor Edward Chang and Saahil Jain for their continuous support. Their constructive feedback on our methods is especially appreciated.

VIII. References

[1] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis; published online Feb 19. https://doi.org/10.1016/S1473-3099(20)30120-1.

[2] “American Inequality Meets Covid-19.” The Economist, The Economist Newspaper, www.economist.com/united-states/2020/04/18/american-inequality-meets-covid-19.

[3] “COVID-19 in Racial and Ethnic Minority Groups.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 4 June 2020, www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/racial-ethnic-minorities.html.

[4] Johnson, Akilah, and Talia Buford. “Early Data Shows African Americans Have Contracted and Died of Coronavirus at an Alarming Rate.” ProPublica, www.propublica.org/article/early-data-shows-african-americans-have-contracted-and-died-of-coronavirus-at-an-alarming-rate.

[5] “Locating At-Risk Populations.” COVID, www.covid19atrisk.org/vulnerability.html.

[6] “Census COVID-19 Data Hub.” COVID, covid19.census.gov/datasets/income-and-benefits-in-2018-adjusted-dollars-counties-1/data?geometry=107.191

TABLE I
COEFFICIENT ANALYSIS-ALL COUNTIES, CASES PER 100,000 PEOPLE

Feature	**Coefficients**	**p-value**
Population Density	0.13	5.36e^{-14***}
Percentage of Population-Black	11.52	$< 2e^{-16***}$
Percentage of Households on Food Benefits	-0.48	0.82

*Indicates statistical significance. All models had low R^2 values. Adjusted R^2 value = 0.095

TABLE II
COEFFICIENT ANALYSIS-ALL COUNTIES, DEATHS PER 100,000 PEOPLE

Feature	**Coefficients**	**p-value**
Population Density	0.0088	$2.12e^{-14***}$
Percentage of Population-Black	0.5908	$< 2e^{-16***}$
Percentage of Households on Food Benefits	-0.0165	0.873
People per Hospital Bed	-0.0001	0.88

*Indicates statistical significance. All models had low R^2 values. Adjusted R^2 value = 0.14

TABLE III
COEFFICIENT ANALYSIS-SOUTHERN COUNTIES, CASES PER 100,000 PEOPLE

Feature	**Coefficients**	**p-value**
Population Density	-0.05897	0.7070
Percentage of Population-Black	8.98438	2.13e^{-5***}
Percentage of Households on Food Benefits	17.57031	0.0187*

*Indicates statistical significance. All models had low R^2 values. Adjusted R^2 value = 0.08

TABLE IV
COEFFICIENT ANALYSIS-SOUTHERN COUNTIES, DEATHS PER 100,000 PEOPLE

Feature	**Coefficients**	**p-value**
Population Density	-0.16	0.0130*
Percentage of Population-Black	0.7045	$< 2e^{-16***}$
Percentage of Households on Food Benefits	-0.4073	0.1821
People per Hospital Bed	0.000799	0.6608

*Indicates statistical significance. All models had low R^2 values. Adjusted R^2 value = 0.18

TABLE V
COEFFICIENT ANALYSIS-EAST COAST COUNTIES, CASES PER 100,000 PEOPLE

Feature	**Coefficients**	**p-value**
Population Density	0.18	$7.05e^{-5***}$
Percentage of Population-Black	59.95	$1.19e^{-9***}$
Percentage of Households on Food Benefits	-12.51	0.235

*Indicates statistical significance. All models had low R^2 values. Adjusted R^2 value = 0.47

TABLE VI
COEFFICIENT ANALYSIS-EAST COAST COUNTIES, DEATHS PER 100,000 PEOPLE

Feature	**Coefficients**	**p-value**
Population Density	0.027642	$3.62e^{-7***}$
Percentage of Population-Black	3.074685	$8.53e^{-5***}$
Percentage of Households on Food Benefits	0.643719	0.380
People per Hospital Bed	-0.002	0.833

*Indicates statistical significance. All models had low R^2 values. Adjusted R^2 value = 0.18