COVID-19 and Potential Impact on Minorities

Marianne Boerenko, Rebecca Gonzalez, Arun Kara

INTRODUCTION

According to the CDC, *general health and social imbalances put some members of minority groups at increased risk of getting COVID-19*. Given the ongoing pandemic and parallel social movements, we sought to determine if COVID-19 poses a greater risk to minorities in America. While this project focuses directly on the infection rate of the coronavirus on minorities, it does not address mitigating factors such as underlying health concerns, lack of health care or other socioeconomic factors such as income or lack of transportation in low income areas with drive-thru testing sites (Godoy). Additionally, we found inconsistent reporting on COVID-19 infection and death rates by race. Given the inconsistencies, we used the U.S. Census Bureau's definition of race for our hypothesis and analysis and did not consider ethnicity *(Race..).*

DATA SOURCES

We used several different datasets to compile the data frame to use for our study. The first dataset came from the COVID Tracking Project ("The COVID..."). This is a volunteer organization started by a couple of journalists at *The Atlantic* to organize their research. In April, they partnered with the Center for Anti-Racist Research to study the impact of the coronavirus on minorities. This dataset uses reporting from the states rather than the CDC due to the disparity between the state reporting and the CDC. As the state and CDC reporting gets more in sync, they may start using the CDC numbers.

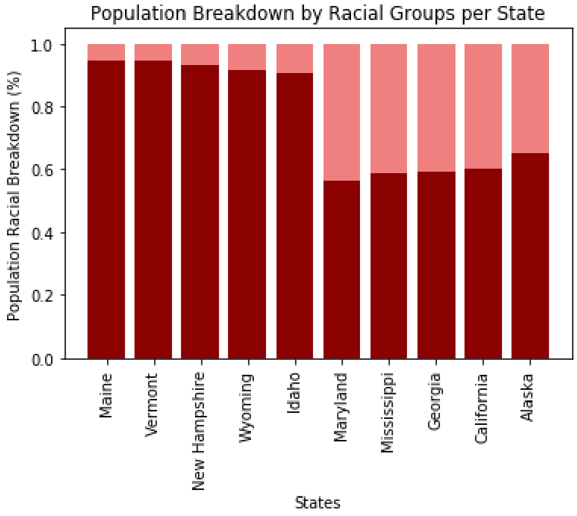
In addition to the COVID Tracking Project information, we pulled the population of each state from the US Census and state race demographics from the World Population Review. While we do not know much about the authors behind the World Population Review, they cite the source of their data as the United Nations – Department of Economic and Social Affairs and the US Census. After analyzing the data, we were fairly confident that the race breakdown data from World Population Review was consistent with our other data sets.

DATE CLEANING AND SETUP

Once all of the data was compiled, we took several steps to clean the data and make it usable for our study. The COVID Tracking Project includes both race and ethnicity. We chose to concentrate our study to race so the ethnicity columns were removed. We removed any states that had more "NaN" fields than actual data. For example, some states report only total cases and do not break it down by race. In addition, we had to convert the states in one data set from abbreviations to full names using a dictionary so they could be merged with the other data sets. Finally, we had to replace the "NaN" fields with 0s and ensure all values were integers and not strings to allow us to combine all race data into one column. Due to timing and lack of experience, for the purposes of this project we lumped race into two broad categories: “Caucasian” and “Minorities”. Once consolidated, we reorganized the columns of the data frame to make it easier to read.

To begin our data exploration, we sorted the data frame by the percentage of Cauasians in the population and used the top five and bottom five states for our analysis. The share of the population identified as Caucasian is depicted in dark red while minorities are depicted in coral. The top five states with the highest percentage of Caucasians are Maine, Vermont, New Hampshire, Wyoming, and Idaho while the states with the highest percentage of minorities are Maryland, Mississippi, Georgia, California, and Alaska. This provided us with a general breakdown of our sample.

FIGURE 01: RACIAL DEMOGRAPHIC BREAKDOWN

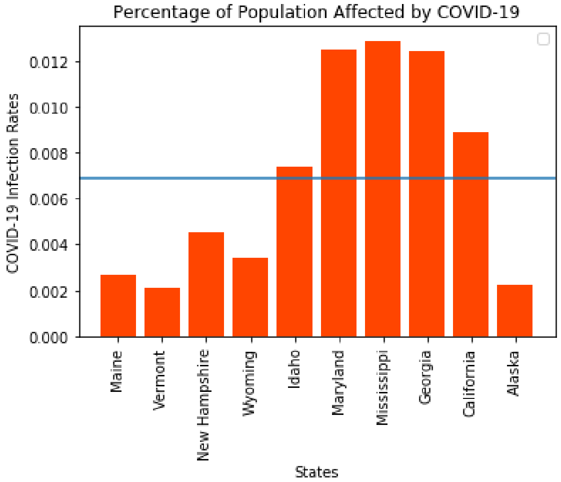


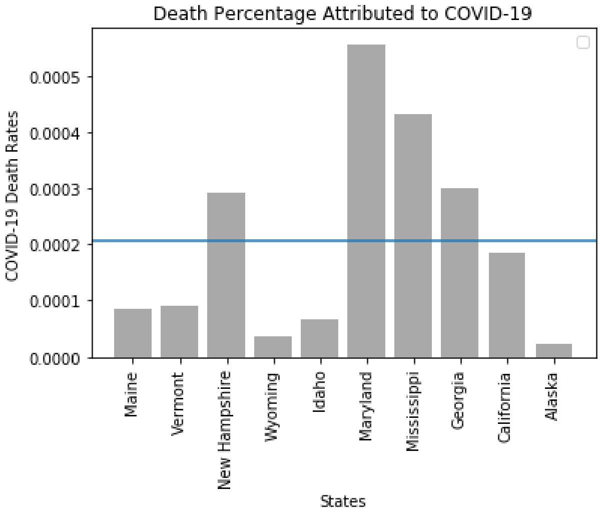
RESEARCH QUESTIONS

Question 1: Do states with larger minority groups experience higher COVID-19 infection rates?

The next step was to determine what percentage of the population has been infected with COVID-19 in order to compare infection rates across the states. We chose not to look at total counts, but rather the percentage of the population affected (infection rates) to normalize the comparison. . We displayed it in a bar graph which shows the percentage of the population affected by COVID-19 as a whole, just to give us a starting point. The graph below shows the highest percentage of infection being Mississippi, with just over 1.2%, and the lowest percentage of infected being Vermont with just under 2%.

FIGURE 02: COVID-19 INFECTION RATES BY STATE (LEFT)
FIGURE 03: COVID-19 DEATHbRATES BY STATE (LEFT)

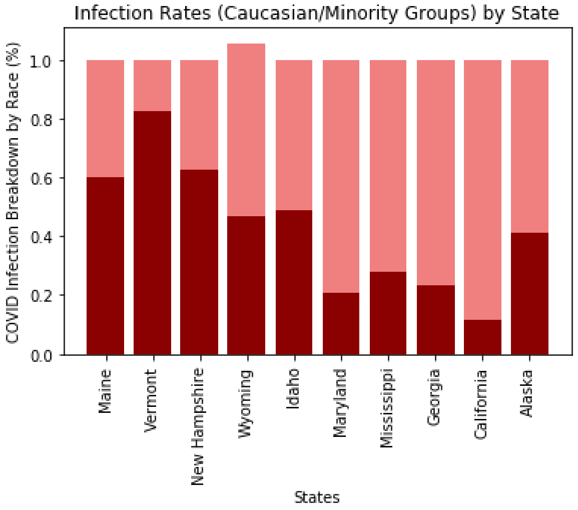


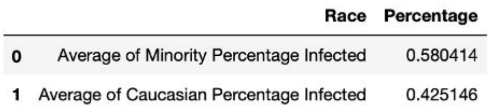


We were also curious as to whether COVID-19 reported death rates mirrored the trend of infection rates and plotted those in a bar chart for similar comparison. The death rates followed the trend since states with a higher percentage of infections also reported a higher death rate.

Question 2: What is the racial breakdown of COVID-19 infections?

Based on our research hypothesis, our main question was *do states with larger minority groups experience higher COVID-19 infection rates and/or death rates?* Due to the disparity in available information, we decided to go with a sample of the top five states with the highest Caucasian percentage, and the top five states with the highest minority percentage. In this graph, "Infection Rates of Caucasian vs Minority Groups by State", you are able the percentage of minority groups affected in light coral versus the percentage of Caucasians infected in dark red. Looking at the sample mean of infection rates, both show that there is a higher percentage of minorities that get infected by COVID-19 overall.



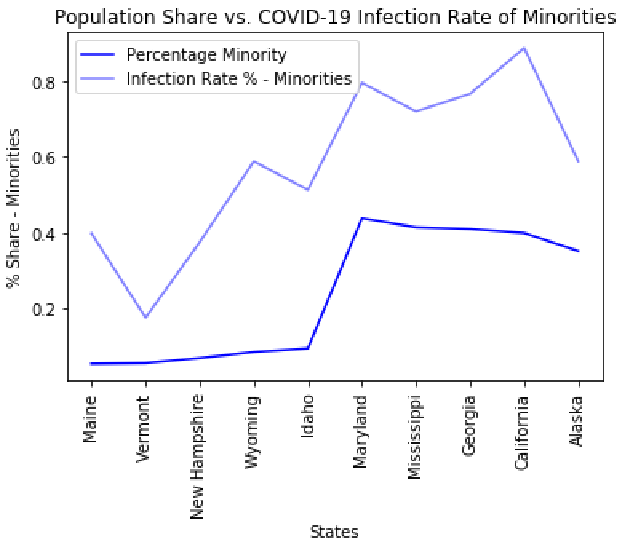


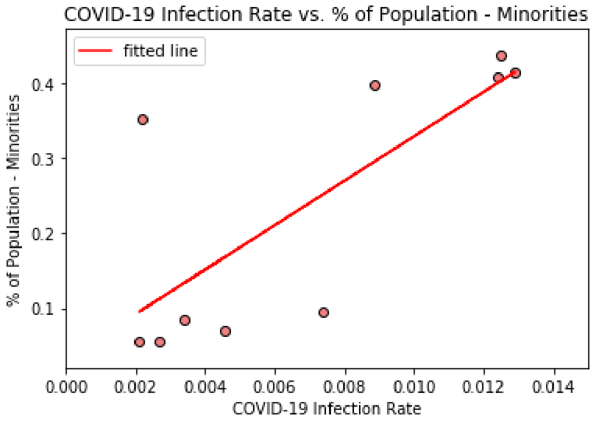
	Race	Percentage
0	Average of Minority Percentage Infected	0.580414
1	Average of Caucasian Percentage Infected	0.425146

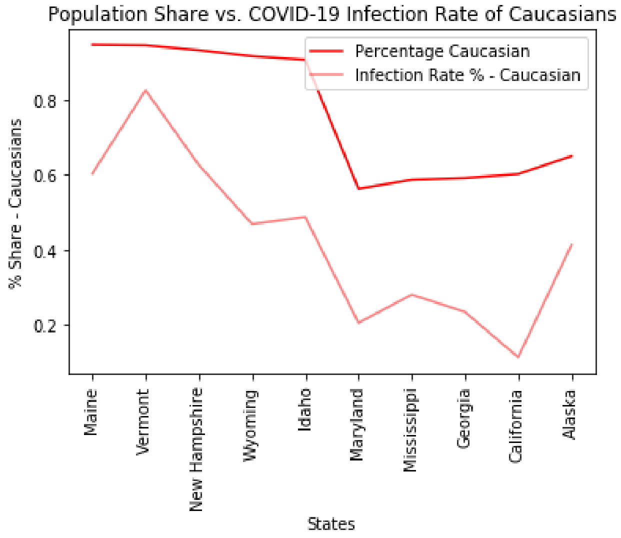
FIGURE 04: RACIAL BREAKDOWN OF COVID-19 INFECTION RATES

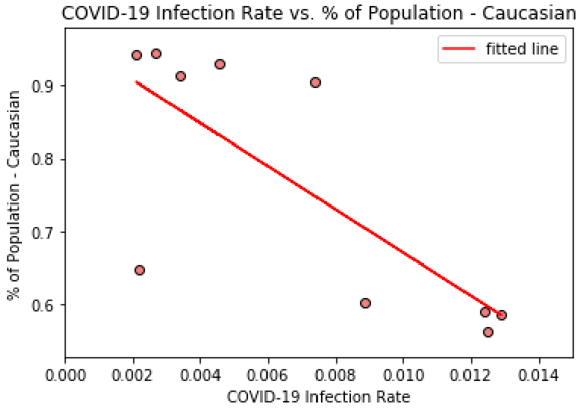
Question 3: Do COVID-19 infection rates reflect the state's racial breakdown?

Now that we have our infected population broken down by states as well as the percentage of minorities versus the percentage of Caucasians infected. To see if there was a positive correlation between the % of minorities in a state and the percent of those minorities getting infected we created a scatter plot with a linear regression. From the graph below you can see that as the percentage of minority groups rose in a state, so did the percentage of COVID-19 infections.









To determine whether we should accept or reject the null hypothesis, we ran a t-stat test on the number of cases for minorities versus the number of cases for Caucasians. Our results for the T-Stat was 2.576 which is falling on the far-right side of a normal distribution curve. As well as a P-Value of 0.019, we decided to go with a significance level of .05 which we felt comfortable with and both outcomes prove that minorities have a higher chance of getting infected versus Caucasians. Looking at the p-value it is less than our significance level, so we would reject the null hypothesis and accept the alternative hypothesis with sufficient evidence. As well as looking at the T-stat since our level of significance is 2.576 which is outside of our z-score for our significance level of 1.645. Therefore, we can reject the null hypothesis, and accept the alternative hypothesis with sufficient evidence.

CONCLUSION AND NEXT STEPS

As you can see from our findings, we have shown it possible to accept our alternative hypothesis that states with a larger share of minorities also report a higher COVID-19 infection rate compared to predominantly Caucasian states. We know we over-simplified our data, but we have sufficient evidence to warrant further study by expanding our data set to all states reporting COVID-19 positive cases to determine if our hypothesis still holds true. Additionally, we would explore other possible contributing factors that could explain the correlation between the population share of minorities and COVID-19 infection rates such as income per capita, underlying health conditions, healthcare access and COVID-19 testing procedures and accessibility. Additional non-socioeconomic factors to consider include state lockdown measures and tourism hot spots or international airport hubs or ports.

LESSONS LEARNED

After further analysis, post-project completion, we realized our intent for the project was good-natured but our execution faulty. Given the chance to do the project again, the first step would be to ensure our data sources enable us to execute the desired level of analysis. Ideally, each race category as defined by the US Census Bureau would be uniquely categorized, but the summation of non-Caucasian race groups for this project is sufficient. However, “unknown” cases of race should be counted separately to avoid misrepresenting data rather than a part of minority groups. States with null or missing infection/test counts could be replaced with 0 to account for missing value since it won’t necessarily skew the data, but rather highlight potential problems in the applicability and legitimacy of our results. Additionally, a more accurate population comparison to determine possible minority disparities in COVID-19 would be COVID-19 positive cases versus total number of tests. This normalizes the comparison of racial demographics between the groups and can later be compared to the overall state population racial composition.

After data cleanup, the data frame would be separated by our two race groups (minorities vs. Caucasian/non-minority) and sorted from highest to lowest by infection rate (# of COVID-19 positive results / # of COVID-19 tests). From there, we would calculate the summary statistics table and determine the interquartile range (IQR) and chart the data. Box plots would easily showcase any outliers in our dataframe and the variance would show how much the data varies from the mean infection rate. Further analysis would then lead us to create a scatter plot of the percentage of COVID-19 positive cases against the % of the population per 100,000 by racial group for each state (or at least a sample of 30). A multiple linear regression would be performed to determine if there is any relationship for either minorities or Caucasians. The correlation coefficient would then show whether the data has a positive relationship (r-squared closer to 1) or negative relationship (r-squared closer to -1). An anova test (or t-test) would be used to indicate whether the variables are a good fit and if we should accept or reject the null hypothesis.

At this point, it would be interesting to explore which states showed the highest infection rates and if they share any similar characteristics. Factors such as income per capita, overall population health, type of labor industries per state, and even international airports could be explored to determine if there is any type of relationship. Identical analysis efforts could then be performed with the death counts reported for COVID-19 to determine if there is a similar trend. Our results would then lead to thought-provoking and difficult dialogue concerning potential minority disparity in the United States and possible measures to better contain COVID-19 and/or address contributing factors. Although we recognize our statistical analysis and data approach was over-simplified, we learned a lot and look forward to implementing our newfound knowledge in the next project.

PRESENTATION LINK: https://my.visme.co/view/pvg8zxyv-owplnmm077zz2zd6

WORKS CITED

“Assessment of New CDC COVID-19 Data Reporting.” *The COVID Tracking Project*, covidtracking.com/cdc-paper/

Godoy, Maria, and Daniel Wood. “What Do Coronavirus Racial Disparities Look Like State By State?” *NPR*, NPR, 30 May 2020

Menon, Pradeep. “Data Science Simplified Part 3: Hypothesis Testing.” *Medium*, Towards Data Science, 5 Aug. 2017, towardsdatascience.com/data-science-simplified-hypothesis-testing-56e180ef2f71.

Race and Ethnicity - Census.gov. (n.d.). Retrieved July 25, 2020, from https://www.census.gov/mso/www/training/pdf/race-ethnicity-onepager.pdf

“State Testing Data by Race.” *Johns Hopkins Coronavirus Resource Center*, coronavirus.jhu.edu/data/racial-data-transparency.

“The COVID Racial Data Tracker.” *The COVID Tracking Project*, covidtracking.com/race.