**Final Project, Deliverable 3**

**Sophia Tang Hart**

**Data Visualization**

**Summer 2021**

**Professor Ali El-Annan**

**How COVID19 Effects Different Race**

**OBJECTIVE:**

COVID19 has affected everyone in the world. The disease is extremely contagious and serious. It started in Wuhan, Hubei Province in China. It was announced by the World Health Organization on Dec 31, 2019, that it would spread to the rest of the world if we don’t act quickly in the time window. Unfortunately, it did spread all around the world. 182M cases and 4M deaths happened worldwide (as of July 2021). In the United States, nearly 33.6M cases were found and 604K deaths happened.

Different races are affected by COVID19 by different degrees. It’s been known that African Americans and Latinos have more cases than Whites and Asians. What about American Indians (i.e. The First Nation)? Can we use data to show how different the different races are impacted? What race is the most disadvantage and needs attention? Secondly, it would be interesting to see, is the difference due to genetic make-up or due to socioeconomics? In this project, I will illustrate COVID19 cases and deaths by race using open source data from the COVID Tracking Project [1]. Then I will correlate the impact difference with household incomes by race.

**METHODOLOGY:**

My illustration is composed of three plots: (A) Cases by race in the United States, (B) Death by race in the US, (C) Household income by race in the US. Each race is represented by different colors (Asian-orange, Black-greenish-brown, Hispanic-green, American Indian-blue, and White-pink) to make the colors interesting. Legends on the x-axis is blue, and labels on x-axis and y-axis are red. The titles for the plots are black. This red-blue-black combination makes them each stand-out.

The metrics for Plots A and B are proportion, which is the number of cases for a race normalized by the total population of the corresponding race. Proportion is chosen instead of the number of cases, because each state in the US has different amount of people for each race. For example, California has 2M Black people and 5.7M Asian. If there are 150K Black people who contracted COVID19 and 285K Asians who contracted COVID19, there is 7.5% of the Black people cases and 5% of the Asian people cases. Therefore, there is higher proportion of Black cases, although 150K is less than 285K. I think proportion is more relevant than the absolute number. Proportion is also chosen for COVID19 deaths.

To find proportions, I found two datasets. Dataset #1 is COVID19 cases and deaths for each of the 52 states in the United States for different races as of March 7th 2021. The tracking of cases by states stop on this day; this is the latest day from the data[1]. This data can be found in COVIDdata\_racial.csv file. Dataset #2 is population of different races for each of the 52 states[2]. This data can be found in populationbyRace2019.csv file. Proportion is calculated as dataset #1 divided by dataset # 2 for each of the states. Plots A and B are boxplots for the states, side-by-side with difference races. I wrote a Python code in Jupyter Notebook to read in the two .csv files. The code can be found in SophiaHart\_COVID\_FinalProj.jpynb. Then I computed the proportions as described above, and exported the proportion values into another csv file, COVID\_CasesDeathsRatios\_byRaceState.csv. When I was coding, I used the word ratio, although statistically speaking, I should call it proportion which I used for this final report. This applies to the R code as well.

Then I made all the illustration by writing an R code in RStudio with the ggplot2 library. The code can be found in COVID\_Race\_part2.r. Because of the way my data is structured, I needed to use the tidyr and dplyr libraries to pivot the data. Then I piped (%>%) the data into reshape\_cases and reshape\_deaths. geom\_boxplot() and geom\_col() were called to make the kinds of plots I wanted. Theme() was also applied to make the plots with more aesthetic details. I made three plots: cases\_plot, deaths\_plot and income\_plot. Then I used ggarrange of the ggpubr library to put these three plots into one figure: plots A, B and C. Finally, the figure is saved into a file, COVIDbyRace.png.

My audience is policy makers and politicians, because they hold the power to change policies and allocate money. If my data and illustration show there is racial socioeconomic injustices and it’s not due to biological make-up, we will need the policy makers or politicians to pay attention to it and allocate funding to help to the ethnic groups that are in lower margin. I think this is the most effective way to change our socioeconomic system and the future of America.

**DATA:**

Data is taken from The Covid Tracking Project [1]. The tracking began on April 12, 2020 and ended on March 7, 2021. The original data contains tracking for 52 states (including Washington) every day. Therefore, there are lots of data and it’s time series. The original dataset can be found in COVIDdata\_Racial.csv. In the Jupyter Notebook, I extracted the accumulated data from only one time point (March 7th, 2021) for all the states. The original dataset contains 54 columns, but I selected only 13 columns that contains information I needed for illustration purpose and those that did not have too much missing data. The 13 columns are Date, State, Cases for Total, White, Black, Asian, American Indian and Hispanic, and the same races for Deaths. I removed the states which have any NaN value. This left me with 35 states to analyze.

The population by race data can be found in populationbyRace2019.csv file. The names of the states in the original data are full names (e.g. California), while the state names for COVIDdata\_Racial.csv is abbreviated (e.g. CA). So I had to create another column in populationbyRace2019.csv with state abbreviation to match the COVID dataset for string search. I wrote two “for” loops in Python for a string comparison to match state name and then calculate proportion (= number of cases for the race / total population of the race for that state.)

For the R program, the reshape\_deaths data have outliers for American Indian especially on the upper bound. I removed outliers using the filter function and took between Q1-1.5\*IQR and Q3+0.75\*IQR. IQR stands for interquartile range. Usually, the upper bound for outliers are Q3+1.5\*IQR. However, the outliers for American Indian are high in value and it make the boxplots for other races condensed. That’s why I removed more outliers of American Indian and used a 0.75 factor instead of 1.5.

**CONCLUSION:**

Looking at Plot A of the illustration in COVID19byRace.png, I noticed that American Indian (ie. The First Nation) and Hispanics have the most proportion of cases. Following by Black, then White, then Asian. It’s been brought to attention that Hispanic and Black community are vulnerable to COVID19 cases, but little has been mentioned about The First Nation and they have the most proportion of cases. Plot B shows that White, Black and First Nation have similar COVID19 death proportion. I believe death proportion is mostly due to health condition or how strong the immune system responds to COVID19 after contracting the disease. Death proportion is not the factor of the socioeconomic system. However COVID19 cases is a reflection of the socioeconomic system. This is because if The First Nations or Hispanics are poor, they have to live in small houses in close proximity with other people. Also, The First Nation people live in tribal group and the Hispanics tend to have larger families. These make social distancing impossible. Recall from Step 1 “Why outbreaks like coronavirus spread exponentially, and how to ‘flatten the curve’” *The Washington Post*, the illustration shows that social distancing is the key to prevent the spread of COVID19. Not having the social distancing in place, more of the First Nation and the Hispanics contract COVID19.

I also noticed that cases proportion is directly correlated to household income (Plot A and Plot C in COVID19byRace.png). The First Nation people has the lowest household income (35K) and one of the highest COVID19 cases (median of 9%). While the Asians have the highest household income (98K) and the lowest COVID19 cases (median of 5%). I am Asian so I am partially happy that I am in the better situation, probably because Asians work very hard. However, I believe as a nation, we need to pay attention to the The First Nation people and open up opportunity and resources to these people. Nonetheless, The First Nation people are the first people who lived in America and the I heard that European immigrants drove them out. I understand it’s human instinct to want better resources for ourselves, but we could still share and help out our brothers and sisters of our country. There is a popular saying, “Giving is better than taking”. For example, the government can allocate funding for scholarships for students who are First Nation descents, for incentives if employers hire First Nation people, for assistant housings and food stamps to First Nation families.

It’s a big task to gather COVID19 data from all states in the nation with racial data incorporated and I am glad that I was able to use the data for this project. One thing I would have done differently is some of the missing data. For example, there are many data from the multiracial group that are missing, so I have to not include this group so that there is no NaN values for analysis. My children are Asian and White mixed, so I don’t want people like my children to not be represented in this illustration.

Although the country ordered us to social distance like the illustration in Step 1 shows, not every people would be able to social distance because of socioeconomic reasons. People who are at the margin have to live in close proximity with other people and they cannot stay-at-home and be isolated. People who are homeless don’t even have a place to shelter-in-place. The audience of my project is the government officials or policy makers. Scientifically, social distancing like the illustration in Step 1 makes sense but the problem is more complicated. We need the government to change the socioeconomic structure of the nation, so that the First Nation, Hispanics and Black people are not impacted drastically by infectious disease like COVID19.

**APPENDIX:**

1. COVID19 data retrieved from:

<https://covidtracking.com/data>

2. Population by race (2019) data retrieved from:

https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?dataView=1&currentTimeframe=0&selectedRows=%7B%22states%22:%7B%22all%22:%7B%7D%7D,%22wrapups%22:%7B%22united-states%22:%7B%7D%7D%7D&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

3. Household Income by race (2019) data retrieved from:

<https://www.epi.org/blog/racial-disparities-in-income-and-poverty-remain-largely-unchanged-amid-strong-income-growth-in-2019/>