|  |  |
| --- | --- |
| Close-up of semiconductor elements | Analysis of COVID-19 Across US States  Keyline divider beneath title  **Michael Zaghi**  **Jason Cao**  **Course:** CMPT 353 Computational Data Science  **Term:** Summer 2020  **Instructor:** Greg Baker  **Date:** August 13, 2020 |

# Problem Statement

COVID-19 is an infectious diesis causing fever, cough, shortness of breath and other cold-like symptoms. While the majority of cases are mild to moderate in severity, those individuals who are elderly or have underlying health conditions are more susceptible to the virus. As of August 7th, 2020, there are 4.95M confirmed cases and 161K deaths in the United States alone (*citation here*). The goal of our analysis is twofold. First, analyze underlying socioeconomic and racial factors in order to more broadly understand the state of the pandemic in the US and to identify potential driving forces behind the death rate. Next (How we refined the idea), we use this descriptive analysis as a framework to build a model which can have some success in predicting future deaths in the US.

# The Data

###### Sources and Collection Method

We used three different sources of data for analysis. The *COVID Tracking Project* dataset is used for daily death and race data groups by state. For socioeconomic data, we used a dataset available off *Kaggle* which aggregates *John Hopkins* US Death data with ACS (American Community Survey) US Census data. Lastly, we collected US temperature related data from the *NOAA* (National Centers for Environmental Information) database. All data was gathered in csv format. Links to the original data can be found under the *References* section.

###### Cleaning

An initial cleanup script is run in order to transform the collected data into a format that is more suitable for analysis. In general, this means selecting, reformatting, merging, and grouping the three datasets so that they can act as the ‘base data’ for any other manipulations that may need to be done. If needed, this allows any analysis script to further modify the data more quickly and efficiently. These base datasets can be found in the *output* folder as *daily\_cases\_with\_race.csv, total\_cases\_with\_race.csv,* and *total\_cases with\_socioecon.csv*. (Explain how you cleaned the weather data here)

# Descriptive Analysis

###### Choropleth Maps

Choropleth maps are a good way to get a broad understanding of the geographical characteristics of the virus. *Figure 1* shows total deaths for all 50 US states. It is clear that New York and neighboring States have taken the brunt of the impact. On the west coast, California has the greatest number of deaths. Meanwhile, Middle America (the more rural US interior) has been significantly less impacted by the virus.

*Deaths by Population Over a*nd *Deaths by Population Density* (*Figures 2* and *3* respectively), use relative measures which may allow us to make more insightful comparisons across States. The IFR (infection fatality rate) is significantly greater for individuals over the age of 65 (5.6%) compared to the next category down which is 50-64 (0.14%) (*citation here*). *Figure 2* gives a better measure of how States are performing compared to absolute deaths. One State that differentiates itself here is Louisiana since that State has a relatively small population over 60 compared to their total deaths. We will arrive back at the results of *Figure 3* once we can show that population density is a useful measure.

###### Linear Regression of Socioeconomic Factors

The goal behind this analysis was to identify aggregate level drivers of the US death rate which could be used for prediction. We collected several interesting independent variables such as *Household Income Mean Earnings*, *Total Households with More than One Occupant per Room*, *Age Over 60 in the Civilian Labor Force*, and the *CDC SVI index* (a measure of socioeconomic, minority, language, and disability status). The only meaningful predictor of death we were able to uncover that had a linear relationship was *Population Density* (see *Figure 4*), which had an R^2 of 0.575 and a *p-value* less 0.05.

Now that *Population Density* has been established as statistically significant, we can see how a low ratio of death to population density is desirable. Looking at *Figure 3,* we can see that New York performs relatively better under this measure, and that Arizona performs the worst even though there was no indication of this using the two previous statistics.

# Race Analysis

###### Correlation Matrix

By observing the correlation coefficients between different racial categories, we may be able to identify some general patterns or trends that explain how each group is impacted by the virus. Note that, currently, the only states reporting cases and deaths for every group is CA, CO, and WA. Therefore, any information drawn from this data may not be representative of the entire US population. As expected, *Figure 7* shows that there is a positive correlation between cases and deaths for all groups. Probably the most notable pattern is the strong positive correlation between Black and White in all categories (cases on cases, death on death, and cases on death). Another interesting pattern is the relatively week or negative correlation between, for example, AIAN and every other group.

These general trends could give some indication that, for example, the White and Black communities are tightly coupled in a way that leads to a similar infection rate. Likewise, you could theorize that the AIAN is the most segregated, which may explain their cases and deaths do not correlate with any other group. Of course, further analysis would need to be done to validate these claims, but it is an interesting starting point for further analysis.

###### Daily Death Inference

For this part of the analysis, we were interested in the question “Is the mean daily death rate for each race different when weighted by their respective population?”. We used an ANOVA test to answer this question (our alternative hypothesis). Again, we had the problem of insufficient data and had to broaden our scope from the correlation analysis to AK, CA, CO, GA, IL, LA, MN, NC, and WA by dropping AIAN, NHPI and Multiracial. Of course, this is not representative of the entire US population. Some challenges in working with the data included negative adjustments to daily deaths, distinguishing between days with no reporting and zero deaths, and transforming the data so that it is normal enough to use the ANOVA test.

These issues were handled in the follow ways: the few negative adjustments that occurred were left in because excluding them would deviate from the true mean, the distribution was heavily right-skewed, so normalizing using a min-max scaler and taking the logarithm of the data fixed this issue, days with zero deaths were removed because it became evident by looking at the transformed distribution that very few days had actual zero deaths. The end result of running the ANOVA test was that the p-value was less than 0.05, allowing us to conclude that there was a difference in the means between groups. We were able to further reduce this information by utilizing Tukey’s HSD test to compare pairs. The results are shown in *Figure 8*. You can see that the difference between all groups is statistically significant, with Asian having the highest death rate for the mentioned states.

### Glossary of Terms

**AK** - Alaska

**CA** - California

**CO** - Colorado

**GA** - Georgia

**IL** - Illinois

**LA** - Louisiana

**MN** - Minnesota

**NC** - North Carolina

**WA** - Washington

**AIAN** - Native American and Alaskan Native

**NHPI** - Native Hawaiian and Pacific Islander

**Repository**

git@csil-git1.cs.surrey.sfu.ca:mzaghi/cmpt-353-project-covid19.git

### References

<https://covidtracking.com/>

<https://www.kaggle.com/jtourkis/us-county-level-acs-features-for-covid-analysis>

<https://www.acsh.org/news/2020/06/23/coronavirus-covid-deaths-us-age-race-14863>

### Appendix 1A







