Florida COVID-19 Trends and Insights

Milestone 2

Anna Harvey

Spring 2021

https://github.com/anhar421/Portfolio

Which Domain?

The domain of COVID-19 science data has been aggressively investigated by data scientists over this past year. In terms of data analysis, general trends and basic information have been distributed in order to inform based on pure statistics. The deeper impact of what those numbers mean from a medical standpoint need to be assessed by medical professionals. Often, what we have seen in this area is that citizen data scientists or medical data scientists have analyzed the data which then is distributed to multiple venues. At this point, medical scientists or medical data scientists interpret the data so that the medical significance is made clear. The Covid Tracking Project (https://covidtracking.com) is a great example of this.

Which Data?

This has been my main struggle this week. By this point I had hoped to have some EDA completed but I have been hung up by my main data sources missing codebooks. The Florida Department of Health has been unresponsive to emails, so I have also tried contacting them through social media and hope to hear back at some point soon. If I do not hear back, I will probably use the data from The Covid Tracking Project. Unfortunately, that data does not include a county-by-county breakdown so I will need to reframe my questions if that is the case.

Research Questions? Benefits? Why analyze these data?

An interesting piece of information I have found is that according to the Florida DOH dashboard, "comparison of county data is not possible because case data are not adjusted by population." (https://www.arcgis.com/apps/opsdashboard/index.html#/8d0de33f260d444c852a615dc7837c86) Understandably, larger populated counties will have more cases and deaths just based on probability. I would likely need to standardize the case data by population to fully understand data comparisons between counties. I still believe this is a valid and important analysis for someone to conduct, as county-by-county comparisons have not been done yet. Theoretically, demographic differences between counties, particularly regarding age distribution, income levels, and political affiliations, would affect the amount of COVID-19 spread in a community.

What Method?

With my current setbacks, I may need to settle for accomplishing EDA for this data. I would still like to attempt some correlation calculations as well, although that seems like it might be difficult to accomplish. Creating some quality visualizations would be a helpful tool for communicating information to the public. I want to analyze the data in both raw form and standardized for population. Raw data visualizations and analysis are important for communicating the actual information. Standardized visualizations and analysis are important for understand the differences between the counties. I will likely use Python for most, if not all, of the analysis and Python and/or PowerBI for visualizations.

Potential Issues?

The main issue is that I need to find a database that has a codebook. I also need to look into methods to standardize the data based on population. I know that I want to compare positive case numbers and number of deaths. If I can get a codebook for the FLDOH data, the accessible data is only available for that current day in CSV. I believe that the API available should have information for each day over the past year of collecting data. That could be an issue as working with APIs is still difficult for me. I may attempt to reframe my research questions to work with the available data in a different way.