Data & Methods

Aarushi

2023-01-08

Data & Methods

This project sources two datasets in order to investigate the relationship between certain economic and social variables American Community Survey (ACS) and COVID-19 data in the DC, Maryland, and Virginia. The American Community Survey (ACS) is acquired from https://api.census.gov/data/key_signup.html and COVID-19 Data is acquired from https://apidocs.covidactnow.org.

Our final combined data set, with both ACS and COVID-19 data, has 158 total observations with 13 variables of interest.

The American Community Survey is a perpetual survey run by the United States Census Bureau that provides the United States government, and its population, access to information surrounding different economic and social characteristics such as jobs, income, and education on a yearly basis. The ACS contacts over 3.5 million households every year for participation in this survey, where all individuals contacted are legally obligated to answer all the questions in the survey. The sample is selected at random by the Census Bureau, where no address is chosen more than once every 5 years.

Data surrounding the spread and impact of COVID-19 in the United States was sourced from COVID ActNow. This organization is a collection of scientists, pandemic experts, engineers, and public health experts who collect COVID-19 data in the United States, and then subsequently assess its quality. Their works supports federal, state, and local government agencies and their data sets are considered some of the most-trusted regarding the COVID-19 pandemic.

In determining the variables of interest that we wanted to investigate from the American Community Survey, we were most interested in the number of working individuals aged 16 years and older that commute to work, the total mean household earnings, the number of individuals that have health insurance coverage, and the percent of working individuals aged 16 years and older that are employed in an essential worker capacity. The specific variables and their official ACS codes are below:

- DP03_0018 Estimate, **Commuting to Work**, Workers 16 years and over
- DP03_0065 Estimate, Income and Benefits (in 2021 inflation adjusted dollars), **Total households, With earnings, Mean earnings** (dollars)
- DP03_0095 Estimate, **Health Insurance Coverage**, Civilian noninstitutionalized population
- DP03_0042P Percent, Industry, Civilian **employed** population 16 years and over, **Educational services, and health care and social assistance**

With the variables selected from the ACS, we first started out with time series COVID-19 data for Maryland, DC, and Virginia. We only selected the year 2021 for COVID-19 data in order to match the ACS data from 2021. We were able to get a daily cumulative case and death count from COVID ActNow, but in order to ensure that our analysis was on the same level as ACS data, we created two new variables – average cases from 2021 and average deaths from 2021. The specific variables of interest from this data set:

- Average Cases
- Average Deaths

For the purpose of building a model to assess the significance of the relationship between multiple economic and social indicators and the spread of COVID in DC, Maryland, and Virginia, we aimed to combine both data sets to prepare it for analysis. We were able to load in both Census ACS data and COVID-19 data from COVID ActNow through their respective APIs, and then subsequently cleaned the source data.

For cleaning the COVID-19 data, we parsed out DC, Maryland, Virginia from our complete time series data from COVID ActNow, then created a seperate data set with just these three geographic locations, and computed average cases and deaths for each county during 2021. Then we merged both data sets to combine variables from both sets.

We heavily relied on the `tidyverse` and `janitor` packages for cleaning. `Tidycensus` was extensively used to pull direct data from the ACS. Finally, we used the the `keyring`, `httr`, and `jsonlite` packages for API imports from the respective sources.

Since this project aims at investigating the relationship, and its subsequent significance, between certain economic and social indicators and COVID-19 average cases and deaths in different locales within DC, Maryland, and Virginia. Because of this goal, a linear model is the best fit to further analyze our research question.