Group Members: Brandon Helvey, Bryan Loy, Walter Stern

Introduction

The purpose of this project was to prepare data for an analysis of the relationship between COVID-19 cases and deaths and air pollution. The specific focus was on fine particulate matter (PM2.5) since previous research has reported a relationship between exposure to this particulate matter and COVID-19 death rates (pre-print of research provided here: <https://www.medrxiv.org/content/10.1101/2020.04.05.20054502v2>).

Extract

Data for this hypothetical analysis were obtained from two sources. The first was from the project “Johns Hopkins COVID-19 Case Tracker” on the website data.world. We used the dataset titled “2\_cases\_and\_deats\_by\_county\_timeseries.csv” (i.e., the COVID data) which included new counts of cases and deaths by county by day. Data.world included python code that could be used to import the dataset into a pandas dataframe. We used this code to read in the csv from data.world.

The second dataset was obtained through a request to the United States Environmental Protection Agency (EPA) Air Quality System (AQS) API. In reviewing the documentation we found that PM2.5 is measured every hour per county on every third day using multiple different measurement techniques. Our group determined we only wanted the 24 hour average of PM2.5 using the most recent measurement technique, which began in 2012. We wrote a for loop in python to only request this type of measurement, which allowed us to obtain PM2.5 counts per county per day. Our python code allowed us to create a dataframe of pollutant data and then export to a csv.

Transform

In reading documentation on the COVID data, we could see that several of the columns were superfluous. UID was a unique identifier for a location, as assigned by Johns Hopkins, so it was deleted. Documentation also stated that sometimes the fips-code variable was used to identify exceptions, such as the Grand Princess cruise ship. In sorting the data we could see that the fips code was 99999 in these cases. Other sources suggested the highest “real” fips code we should ever see is 78030 (<https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697>). Based on this documentation, we filtered the data to only include fips codes lower than 78030. This yielded a max fips code of 56045, which corresponds to Weston, Wyoming. Next, the dataset was sorted by date to determine if any data prior to 1/22/2020 were included. It was not, so no dates were deleted. We also inspected other columns in the dataset using the describe function to be sure the data included realistic values.

Once the COVID data were cleaned, a csv file was exported. Python code was used to read this csv and the csv from the EPA AQS API request into a pandas dataframe. Columns were dropped that we not needed for a hypothetical correlation analysis between air pollution and COVID-19 cases and deaths. An inner join was used to merge the two dataframes on state, county, and date. We used an inner join to drop rows from the data where there was not a matching date. This occurred a number of times as the EPA AQS only reported PM2.5 counts ~every 3 days, whereas the COVID-19 day had a row of updated cases and deaths for every day.

Load

The last lines of code in the final\_merge.ipynb file load the pandas dataframe into an in-memory SQLite database.

Steps to Replicate our ETL Process

1. Run code within the file called covid\_cleanup.ipynb. The end result of this is an exported csv called covid\_part1.csv.
2. Run code within the pollution\_cleanup.ipynb file. The end result of this code is an exported csv file called pollutant\_data.csv.
3. Run the final\_merge.ipynb file. This will read in both of the previously created csv files and merge them. The last line of code within this file creates an in-memory SQLite database that is ready for data analysis.