Data Bootcamp ETL Project

12-22-2018

Robert Voss

Joshua Steinmetz

The extraction phase of our project used two separate CSV files we discovered, one being population data and the other collected pollution data. Both of these datasets are comprised of data spanning years.

Our first dataset used came from a “county facts” CSV file found in a 2016 primary election data file found on Kaggle. The data relevant to our final project used looked for in this county facts file was state population from the years 2010 and 2014.

Our second file, also found on Kaggle, collected data concerned with pollution. This was a very large dataset collecting levels of four pollutants detected at numerous stations in nearly every state.

The extraction phase was carried out using Python tools. Our first manipulation of these datasets was dropping data irrelevant to our project of comparing population and pollution data.

Using pandas and numpy we created a dataframe of each state and it’s abbreviation, population in 2010, population in 2014, and the percent of population change between each of those years. Columns were organized and renamed then using sqlalchemy and pymysql were then prepared to be viewed in a SQL database.

Our pollution file was very large, nearly 1.7 million rows total. This was managed over two separate workbooks to work more efficiently. First irrelevant data was dropped and columns were renamed to represent state, date data was recorded, and four separate levels of four different pollutants recorded by state. Using datetime we created tables averaging Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone levels recorded at each state in 2010 and then at each state in 2014. These tables were merged to display each state’s average of each pollutant in both years. This information was also prepared to be used in a SQL database.

The final workbook used the same information used in the paragraph above. Using datetime we wrote script that counted the number of days an unsafe level of each pollutant was recorded and organized the results by state and year. Only unsafe levels of nitrogen dioxide and ozone were recorded. The tables were merged to organize our data into a final form of state and then number of unsafe pollutant levels for 2010 and 2014 recorded in each state. This final dataframe was also prepared to use in a SQL database like each one before.

We loaded these three python workbooks into a SQL database using sqlalchemy and pymysql to be stored. These three dataframes were joined as one final SQL database.