Katie Manzoline

Carl Reim

Cohort 6

ETL Project Technical Report

Our ETL data focused on comparing crime rates, household income, and car accidents for all counties in California between 2016 and 2018. There are several reasons why one might want to combine these statistics. One could use the information to determine desirable locations to move to. A government planner may look to see if there is any relationship between income, crime and the accident rates on roads. This information could help prioritize the implementation of safety projects along roadways.

Extract

The crimes data source was downloaded via CSV from the California Open Justice website (<https://openjustice.doj.ca.gov/data>).

The accident data was downloaded from Kaggle.com, and includes all traffic accidents in the United States between 2016 and 2019. Originally, the dataset was extracted from Bing and MapQuest APIs. (<https://www.kaggle.com/sobhanmoosavi/us-accidents>)

The income data was sourced from the Census Bureau’s data portal (<https://data.census.gov>). It includes median household income for each county in California, among other related income figures. This data was produced by the American Community Survey data in 2018 as a five-year average.

Transform

The crimes data originally downloaded was for years 1985-2018. The data was reduced to only years 2016-2018. The original data listed many types of crimes. For this project, the data was reduced to only the columns that were found relevant (County, Year, Number of Vehicle Theft, and Number of Violent Crimes). All counties in data set contained the name of the county plus the word “county”. In order to combine with the other data sets, the word “county” was dropped. The cleaned crimes data was exported as a new CSV to combine with the other datasets. Column names were further changed in order to match Postgres table names to prepare for import.

The accident data originally contained over 3 million records and was stored in a very large CSV file. The first step was to filter the dataset down to include only crashes in California and only crashes occurring between 2016 and 2018. Several columns were dropped, the only ones that were kept were: ID and County. The columns were originally renamed and then the crashes were summarized within each county. Then, the resulting “cleaned” dataframe was extracted as a new CSV. This CSV was then added into a new jupyter notebook with the goal of immediately importing it into Postgres. However, some more operations were necessary to reset the index and remove all uppercase letters from the column names. *Note, the originally downloaded csv was too large to be handled by Git. Therefore, it was stored outside of the Git repository, while the “cleaned”* CSV *was exported within the repository. You can view the original csv by extracting it from the Kaggle link above.*

The income data was brought into a dataframe and immediately reduced to only two columns: “Geographic Area Name” and “Estimate!!Households!!Median income (dollars)”. The columns were renamed and the Geographic Area Name column was stripped of the words “County” and “California”. The data was then exported into a cleaned CSV, and reimported into the final jupyter notebook. Here, the columns were removed of capital letters to more easily be imported into Postgres.

Load

The following three tables were created in Postgres database etl\_project: crime, income, and accidents. Each of these tables contained the same columns from cleaned data CSVs. Using the SQLAlchemy Python module, a connection was created to the Postgres database and the dataframes were converted to into database tables. To verify this process worked, the three tables in Postgres were joined on county and the data successfully imported. A Postgres database was chosen due the table-based nature of the project, and to give us more practice using a SQL-database which is used at our places of employment.