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. One might want to compare these things in order to determine which county would be the best fit to move to.

Extract *(your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc) Delete once complete.)*

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

*Carl sources (delete this header once complete)*

Transform (*what data cleaning or transformation was required. Delete once complete.)*

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.

*Carl Cleanup - (delete once complete)*

Load (*the final database, tables/collections, and why this was chosen.* *Delete once complete.)*

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. In Python, a connection was created to the Postgres database and the dataframes were converted to SQL. To verify this process worked, the three tables in Postgres were joined on county and the data successfully imported. This process was chosen due to the ease of verification through Postgres.