ETL Challenge

Extract

For data, we chose to get the data from the City of Austin’s Police Department. Specifically Traffic Fatalities from both 2016 and 2017. This file came in the form of CSV and JSON.

We were interested in analyzing this data because it would help us determine if a specific area had a statistically higher incidence of traffic fatalities.

Transform

After we were able to get the data, we loaded it into a data frame through pandas. For the CSV, we dropped irrelevant columns and cleaned data by taking out columns with no data. Loading the JSON file proved to be more difficult, we had to drop the first couple of irrelevant columns, then we cleaned up column names. Modifying the column names was difficult because dictionary keys are immutable objects. So instead of modifying the key, we reassigned the value to a new key and removing the old key. Once we finally modified the keys, we dropped some further irrelevant columns then the data was ready to be analyzed.

Load:

Used SQLAlchemy to connect to the ETL database, create the tables and load data. Using SQLAlchemy to create tables and load data reduced the overhead of manually creating the tables, errors encountered on handling the data types in addition to making the process time efficient

Steps to reproduce data Load

- Create the database named ETL in pgAdmin.

- Execute the Extract and Transform Code.ipynb