**ETL Project Report**

Group Members

-Henry Poe

-Juli Chalk

-Sebastian Lindegaard

Extract

Our first data set contains information about used car sales across the United States. We found this on kaggle and downloaded it as a csv. The other data set that we decided to include in our project is household income data sorted by zip code, a column that was present in our first data set as well. All of the household income data was gotten from the US census website and also downloaded as a csv.

Transform

Income Data

The first step in cleaning up the census data was to get rid of all of the columns we didn’t want. Then it was a matter of renaming columns to more quickly understandable titles and making sure they were identical to our sql column names. Lastly we had to remove extra text that the census data included in the zip code to make it line up with our other data set.

Car Sales Data

First we deleted the unnecessary columns for our data set. Once down to the columns we deemed as necessary, we dropped n/a columns and columns that had ridiculous mileages(above 600k). We then renamed the columns using underscores to work in pgAdmin in the future. The data was also containing 2020 data as well as 2019, but we only wanted 2019 so we filtered the data down to only cars sold in 2019.

Load

On Postgres, we created tables with the proper column names to match our cleaned data. We then made a connection between PANDAS and Postgres. After establishing a connection, we were able to import our clean data to the SQL database.