ETL Project: General Car Info

Team Members: Ronald Ho, Michael Trias

For this project, we were able to combine 2 different data sets that related to car info. Our datasets were obtained from two different sources on Kaggle. The “cars\_database.json” file included performance information such as acceleration and maximum speed, while the “cars\_database.csv” included more general information such as origin, MSRP, and MPG. As an initial step, we read both of these data sets into Jupyter notebook for data cleanup. Not all of the columns from each data set were desired so we removed the columns we didn’t need. We decided that in order to match these 2 data sets together we would need to filter out matching “Makes” and “Models” that weren’t in both data sets. We had to do more cleanup when it came to matching the “Make”. Some of the data for the “Make” included additional info such as different editions, how many doors, etc. so we cleaned this up but running a for loop through the model list and splitting up the values using “.split()”. From here we took the first value using the first index and put that into a new list. We then updated our dataframe with this new list. To condense the json dataset we used a “.isin()” function to remove any rows that weren’t present in both datasets. In order to prevent any problems if different “Makes” had similar named “Models” we created a new “PK” column to act as a primary key. This column value concatenated the make and model to give a unique value. There were instances where certain values were not available, so we had to use “.dropna” several times. Once our data transformation was complete, we then loaded these tables into Postgres using SQLAlchemy to make the connection. From here we were able to join both datasets using “PK” as the primary key and perform various queries. For more details, please refer to our Github Repo.