ETL Project Overview

Data Sets Used

* <https://www.kaggle.com/sametevik/car-sales>
* <https://www.kaggle.com/murtio/car-noise-specification>

Extraction

Both data sets were formatted as CSV files. Naturally Pandas csv\_read function was used to bring the data into python.

Transform

The goal of this ETL process was to bring both of these data sets into a SQL database so the tables could be joined on the manufacturer and model of the car. The car noise data had to be transformed the most to achieve this; the first step was to drop duplicate model varieties since the car sales data did not include the year of each car thus unfortunately making it impossible to join these tables on model, make, and year. The next step was to concatenate the ‘brand’ and ‘model’ columns of the car noise data, set the result as the index, and drop the columns themselves. Next, the data types of the car noise table were examined for any oddities, and we saw that the ‘year’ column was unnecessarily a float, so we converted that column to integers. With the car noise data sufficiently transformed and cleaned, we moved onto the car sales data; the first step was to drop unnecessary columns including: resale value, power/performance factor, width, length, and latest launch. Then we performed the same concatenation on this data’s ‘model’ and ‘manufacturer’ columns, set the result as the index, and dropped the columns themselves. Lastly, we converted any columns in the data that were floats to integers where possible.

Load

The first step of the loading process was to create our tables in PostgreSQL. SQL was chosen since from the beginning we wanted these tables to be clean and able to be joined on the index. We didn’t want any empty values in the tables. So, we made the ‘Manufacturer\_Model’ index in both data sets the primary key and foreign key for car sales and car noise respectively. Then we created and labeled all the columns of the table appropriately. Lastly, we created and connected our engine in python, and proceeded to insert the data frames into our SQL tables.