Car Analysis Report

For this project our team selected a number of datasets from Kaggle which were related to car sales, available used cars, and car data. Each of these datasets approached the topic of car sales from a slightly different angle. The purpose of our work in this project was to collect all of this data on the subject of car sales for potential future analysis.

Dataset 1: Vehicle Data

This dataset includes information for a few hundred different types of cars. This file is basically an encyclopedia of information for each car including information such as: engine size, horsepower, and car size. This file can be considered a reference that can inform the sales data in the other two files.

Dataset 2: Japan Used Cars Data

This dataset contains information for all of the used cars available for sale in Japan. (year, source). It contains the usual car make and model information and various parameters about the state of the car from its mileage to the level of wear on the car.

Dataset 3: Belarus Used Cars Data

This data contains information about all of the used cars for sale in Belarus (Eastern European country). It contains information similar to the one above about the cars from their mileage to their overall wear.

ETL

For each dataset the columns were reviewed for relevance and for information contained.

If a column contained information whose source or values were undocumented it was removed from the analysis. For example, the Belaruss Used Cars Data contained a large number of columns whose titles contained “Feature” followed by a number. The dataset on Kaggle did not describe either where this information originally came from and what these values represented. As a result they were removed from the dataset here for simplicity.

Other column exclusions were made if the data was duplicated in another column. For example, a table with a column for fuel type and engine type Hypothetically a hybrid engine type would result in a different value from the fuel type, but none of our datasets included hybrid cars. If a future used was interested in further analysis of engine type this column would need to be returned.

Each table was checked for complete value counts meaning that each row contained information in each cell. Any rows that were found to contain missing values were examined or removed.

Finally, each table was reviewed and the column names were standardized into the same format removing spaces and adding capital letters as needed.

Future Analysis

This gathering of data in our PostgreSQL database provides an opportunity for a large number of analyses related to cars and used car sales. A user could examine the types of cars available in each of these distinct regions to find any trends or overlap. They could examine what the most common types of cars that are available for sale have in common. For example, do they typically have small engines or are they predominantly Toyota?

While this could provide useful information for someone buying a car (if they were in either of our disparate geographical regions, it would be most useful examining trends between the regions and evaluating what the types of cars predominantly for sale in each area have to say about the people who live there. As stated in our presentation the car is one of the most significant pieces of equity a family owns and as a result the car market is a valuable indicator of economic wellbeing. This dataset on its own and potentially bolstered with more used car datasets from across the world could provide interesting current and historical insight into each region.