Capstone Data Wrangling

Tesla Battery Degradation

The work described in this report is contained in the tesla-battery-degradation.ipynb notebook of the same repository.

The data set was downloaded from Google drive in a form of MaxRange Tesla Battery Survey.xlsx file. The goal was to use python to import and inspect the data for separate spreadsheets from different locations which include Asia & Europe, USA, Canada, and UK, isolate relevant variables, determine the correlation between the target variable (remaining range in %) and other variables, organize the data frames for different locations into one data frame, and resolve any missing, invalid, or corrupted rows.

Import packages, .xlsx file, and inspect the data

After the raw data set was imported from the Google Drive, the initial inspection of the excel file established that there were 4 spreadsheets of interest from different locations. Each location was imported separately by parsing each spreadsheet into a data frame. Each location data frame was cleaned of empty rows which contained only NaN values. All the separate data frames were merged into one master data frame. The resulting master data frame `tesla_survey` contains 1156 observations and 54 columns which represent answers provided by Tesla drivers and related to characteristics of their cars.

Organize and re-name columns

Before merging the location data frames, the column names were changed because the original header was too descriptive. All 54 columns were renamed to merge the data frames. The last five columns with names that include `book_keeping…` were eliminated because they had 0 non-null values. Columns which contain information about chart data used by the owner of the file were eliminated too. Some of the columns represent additional optional information which users could provide if they had any recent trips with overnight charging. Majority of the users did not answer these questions (92.5%). Therefore, these columns were eliminated. As a result, we have 36 columns which contain required, calculated, and optional information about Tesla battery performance. Some of the columns for Asia & Europe as well as Canada had inconsistent units for mileage and mileage per day because in the original data frame, SI units were used. These columns were transformed to miles and miles/day for `mileage` and `mileage_per_day`, respectively.

Evaluate NULL values

Because we identified the empty rows in each location data frame before merging, we did not have to perform any extensive investigation of the master data frame. There are 4 observations (rows) that have no information about manufacturing date which is why the milage per day was not calculated for them. However, they still have information about the total mileage and remaining range which means that we can still use these values in the future analysis.