**Capstone Project – Data Wrangling Report**

The purpose of EPA’s fuel economy estimates is to provide a reliable basis for comparing vehicles. Most vehicles in this guide (other than plug-in hybrids) have three fuel economy estimates: A city estimate, highway estimate and a combination of city and and highway driving. This data provides annual fuel cost estimates, rounded to the nearest $50 for each vehicle across different makes and models. The data also includes CO2 emissions, and smog ratings that helps evaluate environmental affects based on fuel consumption of vehicles.

The data files are provided by year. Each individual data file is down loaded , saved as csv files to facilitate reading into data frames.

The csv files are loaded into data frames. And in this step, the empty values of variables with some missing data are replaced with NA/None for easier data manipulation later.

The initial data frame for one year datafile consisted of :

'data.frame': 1263 obs. of 311 variables:

Now, the data frame is cleaned by removing any variables with all empty values. After this step, the data frame for one year datafile consisted of :

'data.frame': 1263 obs. of 127 variables:

Repeated the process for data files for 4 years.

Merged the data frames into one final data frame. The data files for different years varied in the number of variables. Used the function rbind.all.columns function to merge the data frames. It is useful to combine data frames with different number of columns. The data frame now consists of :

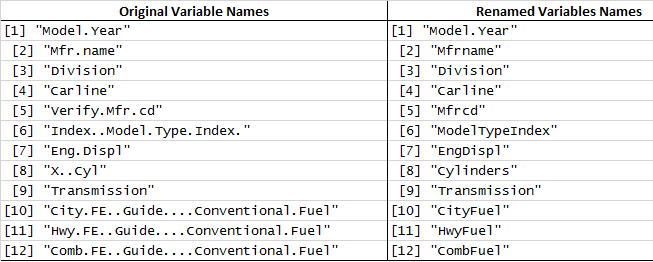
'data.frame': 4990 obs. of 129 variables:

As we can see, there are way too many variables (129). Observed the data frame and discarded the columns not

needed for analysis. Now the dataset consists of :

'data.frame': 4990 obs. of 63 variables:

The column names in the original data file are very long and with spaces and special characters. Renamed the column names for easy read and data manipulation. Some of the variables shown below :



The next step is to perform some exploratory analysis to understand the fuel usage by various types of cars and to understand influence of various variables on fuel usage.

