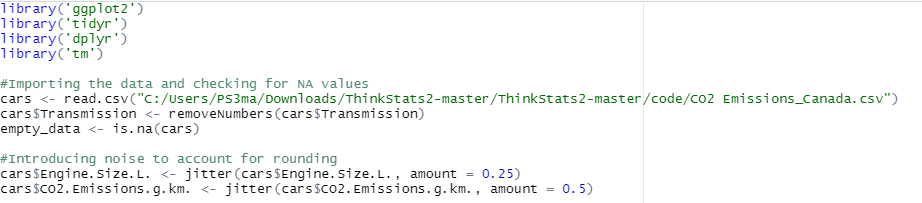
**Importing the data**

First, I will import my required libraries then import my dataset. The Transmissions vector contains the transmissions type and gear number, but I will not be using the gear number, so I will use the function removeNumbers() from the tm library to remove all numeric values from the vector.

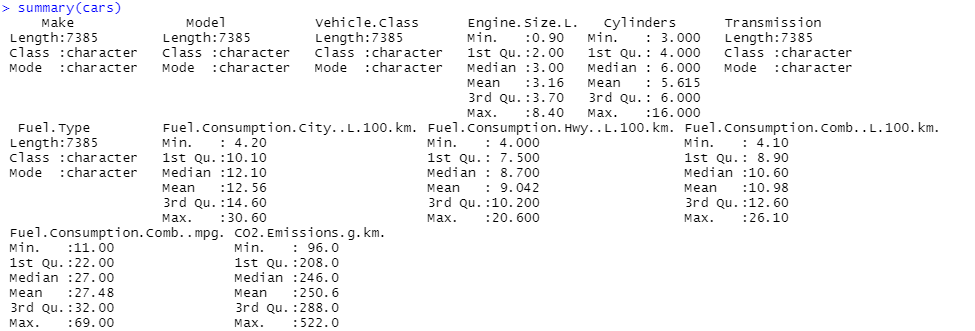
The dataset was rated a 10.0 in usability on Kaggle; this means that it should be free of any missing data. I confirmed this using the is.na() function, which returned a vector with all elements FALSE.

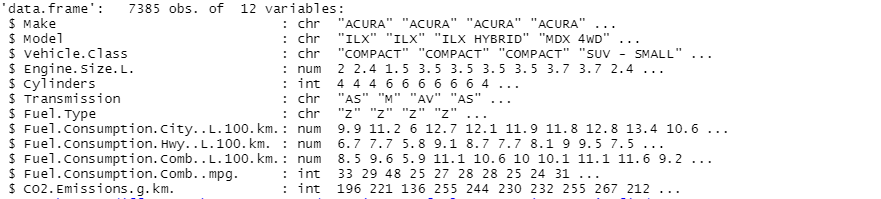
I then introduce jittering to some of the rounded continuous variables to account for data lost in rounding.



**Condensed form of the data**

Below is the summary statistics of the data frame as well as the data type of each column





**Uncovering new information that is not self-evident**

Right now, it is not evident whether there is a difference in fuel consumption and CO2 emissions between all the fuel types. It is also not evident whether diesel fuel pollutes more than other fuels, or what car companies are responsible for the bulk of CO2 emissions.

For the former, I will plot a scatter plot of CO2 emissions versus combined fuel consumption with points color-coded by fuel type. If the difference seen is small, I will run a two-sample t-test on CO2 emissions of diesel versus gasoline (the most common fuel type in the dataset) and fuel consumptions of the two gasoline types.

To find out what car companies are hurting the planet the most, a simple box plot of CO2 emissions versus manufacturers can be constructed.

**Different ways to look at this data**

One way to look at this data would be to gauge the effect of each individual variable on CO2 emissions and fuel consumption. More specifically, I could see how transmission and fuel type, commonly overlooked engine parameters, will affect fuel consumption and CO2 emissions.

**How I will slice and dice the data**

I will slice the data by each one of its variables (except for model) to see which variables have the most effect on CO2 emissions and also combined fuel consumption.

**Plots and table to help me illustrate my findings**

I will use histograms and bar charts to find the most common engine type, scatter plots to illustrate trends between continuous variables, and box plots to illustrate trends between continuous and categorical variables. It looks like box plots are the only way to illustrate trends between categorical and continuous variables if the categories are very numerous.

I will also use a correlation matrix to show the correlation of all continuous variables at once to save time (correlation type will depend on the trend seen in the scatter plots).

**Machine Learning**

I will use a multiple regression model to predict CO2 emissions from my variables. I will use another multiple regression model to predict combined fuel consumption from the same variables.