# Introduction to Data Science and R programming

# Capstone Project Final Submission

## Capstone Project :

### Problem to solve

Cars are used by all aspects of a growing society. And, one of the most important factors to consider while purchasing a car is the fuel efficiency of the car. The premise of this analysis is to understand the various factors influencing the fuel efficiency of a car. Some of the factors of interest are engine cylinders, vehicle class, transmission type, highway vs city miles, drive axle etc. These factors themselves are important considerations for most buyers and so it is important to understand their influence on fuel efficiency. Fuel efficient cars not only benefit the buyers in terms of reducing fuel costs but they also serve to benefit environment by reducing the air pollution. The analysis would show correlations between these factors on fuel usage and help predict fuel usage based on these. Also would study the influence of fuel usage on environment impacts based on air pollution score and GHG emissions.

Goal of this project is to perform exploratory analysis on fuel economy based on the factors influencing the mile per gallon of different cars. Also, predict the fuel usage based on the various factors influencing fuel usage

### Client and why do they care about the problem

Potential car buyers, car owners wanting to cut fuel costs, environmental enthusiasts. Also, car manufacturers can benefit from looking at the benefits on environment and come up with more fuel efficient options.

### Data that will be used and how to acquire this data

This dataset contains the fuel consumption (mile per gallon) data for different cars along with other features of the car like vehicle class, cylinders, axle type, transmission type, highway and city miles. The data can be downloaded from the following site :

https://www.fueleconomy.gov/feg/download.shtml

Download Fuel Economy Data

www.fueleconomy.gov

Download Fuel Economy Data. Fuel economy data are the result of vehicle testing done at the Environmental Protection Agency's National Vehicle and Fuel Emissions Laboratory in Ann Arbor, Michigan, and by vehicle manufacturers with oversight by EPA.

### Outline approach to solving this problem

Extract the data from the above data site. Perform the required data wrangling to get the data into R required format. Perform exploratory data analysis on the data to understand basic information like which cars and which models in different car brands are more fuel efficient. Then dive into details of various features of a car influencing the fuel consumption. Predict how the fuel economy can be improved by increase in purchase of fuel efficient cars. And, the benefits on environment with cut down in fuel consumption.

## Dataset and Data Wrangling

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 effects based on fuel consumption of vehicles.

Identified the data files needed. The data files are provided by year. Each individual data file is downloaded , 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 1263 observations with 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 1263 observations with 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 consisted of 4990 observations with 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. Basically eliminated columns that do not provide value to analysis. (eg: most of the data is null, columns providing descriptions, unit description columns etc) Now the dataset consisted of 4990 observations with 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. Few Column renames are as follows :

| Original variable names | Renames variable names |
|---|---|
| Model.Year | ModelYear |
| Mfr.Name | MfrName |
| Division | Division |
| Carline | Carline |
| Verify.Mrf.Cd | MfrCd |
| Index..Model.Type.Index | ModelTypeIndex |
| Eng.Displ | EngDispl |
| X..Cyl | Cylinders |
| Transmission | Transmission |
| City.FE..Guide….Conventional.Fuel | CityFuel |
| Hwy.FE..Guide….Conventional.Fuel | HwyFuel |
| Comb.FE..Guide….Conventional.Fuel | CombFuel |

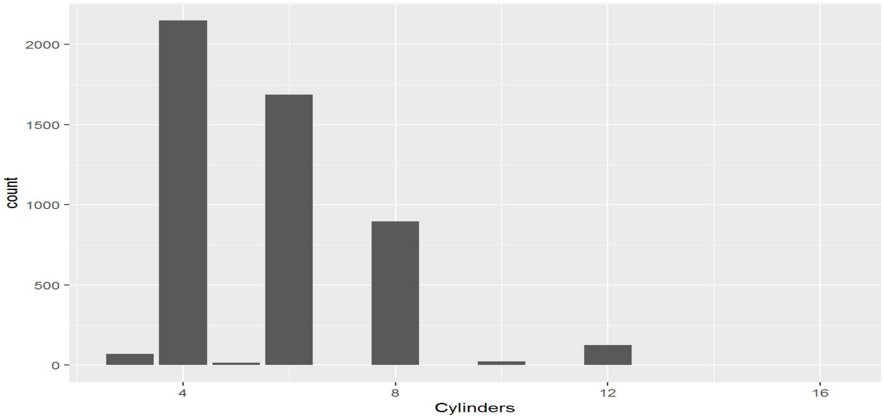
The next step is to perform some exploratory analysis to understand the fuel usage by various types of cars and to understand influence of various variables on fuel usage.

## Exploratory and statistical Analysis :

First, examined for distribution of different variables in the data set. Different plots (line, bar, histogram, box) were looked at and finally kept the plots that are easy to read and analyze. As a next step, studied the influence of various variables on fuel consumption (mpg). The key variables studied for this analysis are cylinders, engine displacement, transmission, and drive systems. There are other variables like viscosity, air aspiration methods, valves/valves timing etc in the data set. For the scope of this analysis, looked at variables that are usually looked by consumers while purchase of an vehicle. Examined the fuel consumption across different carlines and manufacturers. Correlation plots of key categorical variables was studied. Examined few scatter plots with linear regression to understand the correlation.

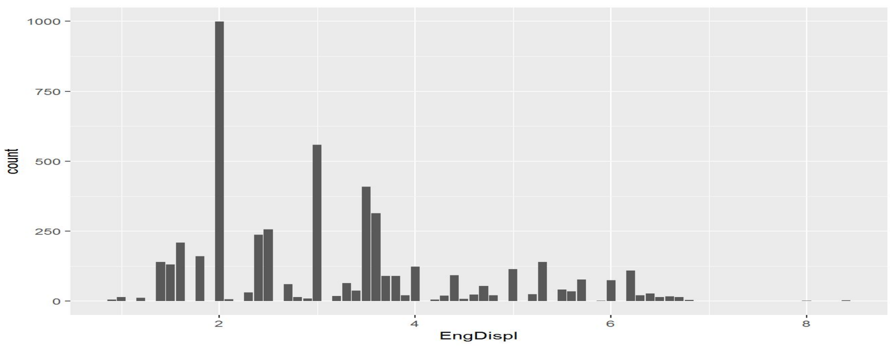
### Cylinders

Most cars have 4,6,8 cylinders, highest being 4 cylinder cars. As this data set is all about fuel usage, interesting to see how the number of cylinders are in correlation with fuel consumption



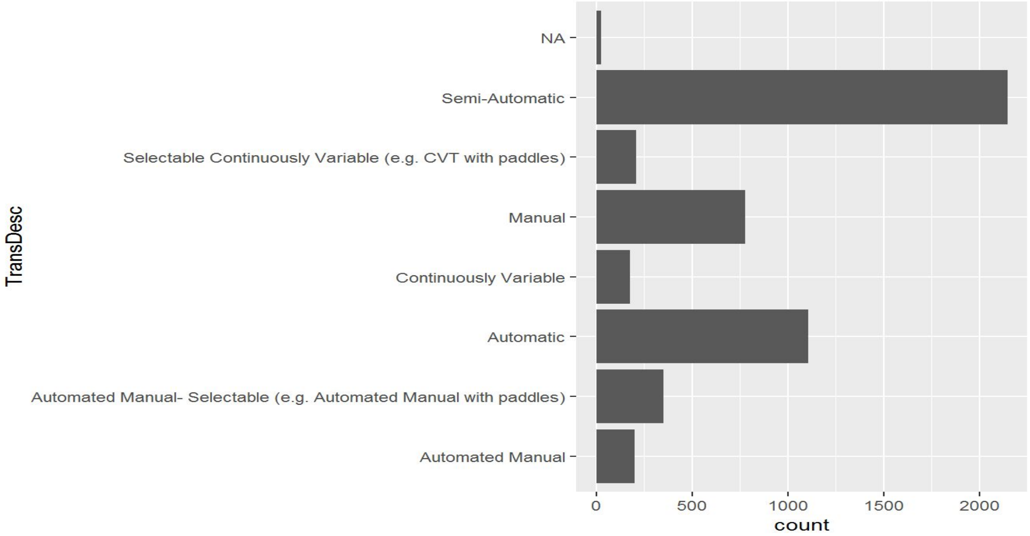
### Engine Displacement

Majority of cars Engine displacement ranging from 2 to 3.5. Higher the cylinders, higher is the engine displacement in general. So expecting cylinders and displacement are correlated to fuel consumption in similar way.



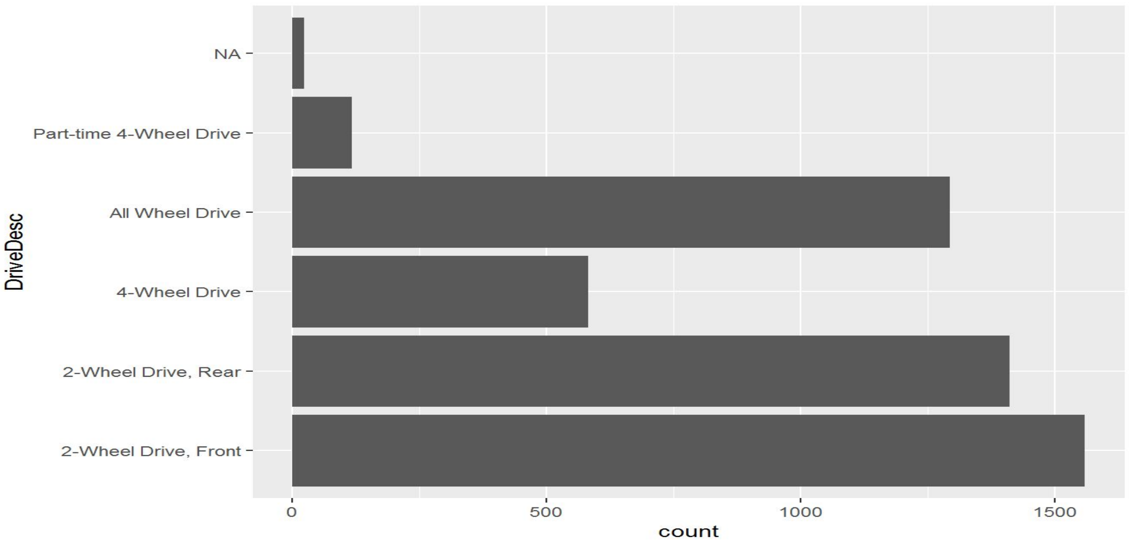
## Transmission

Semi automatic cars are higher in number followed by automatic and manual. The transmission impact on fuel consumption will be studied to understand their correlation.



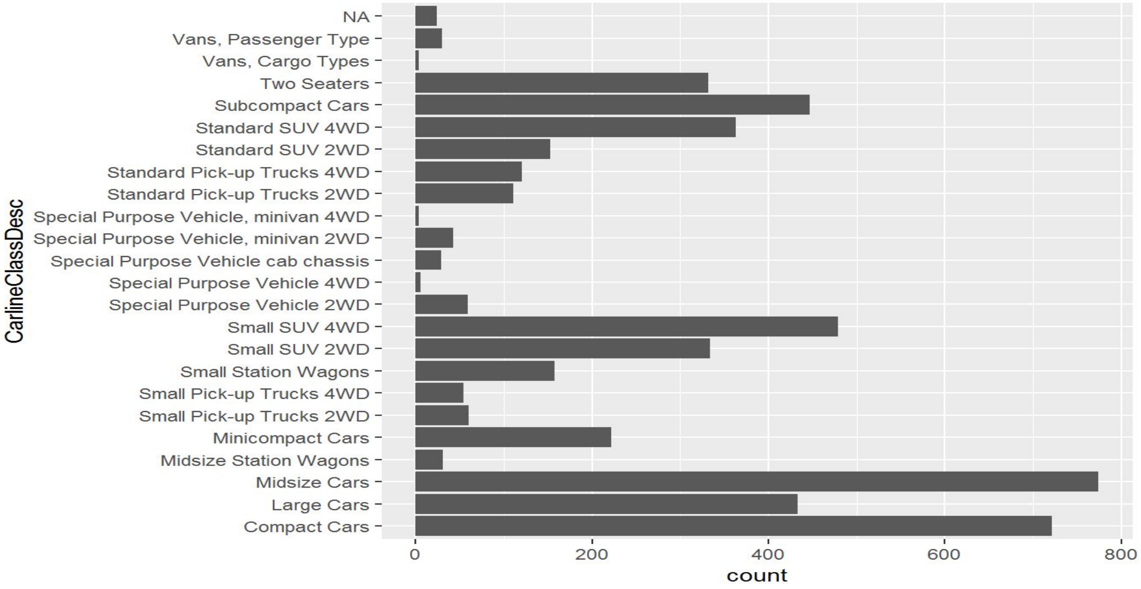
## Drives

2 wheel drive and All wheel drive are higher than 4 wheel drive cars. Is 2 wheel drive more popular because of the higher mpg it gives ? And the drive system distribution within the carlines need to be looked at to completely understand the impact on fuel efficiency.



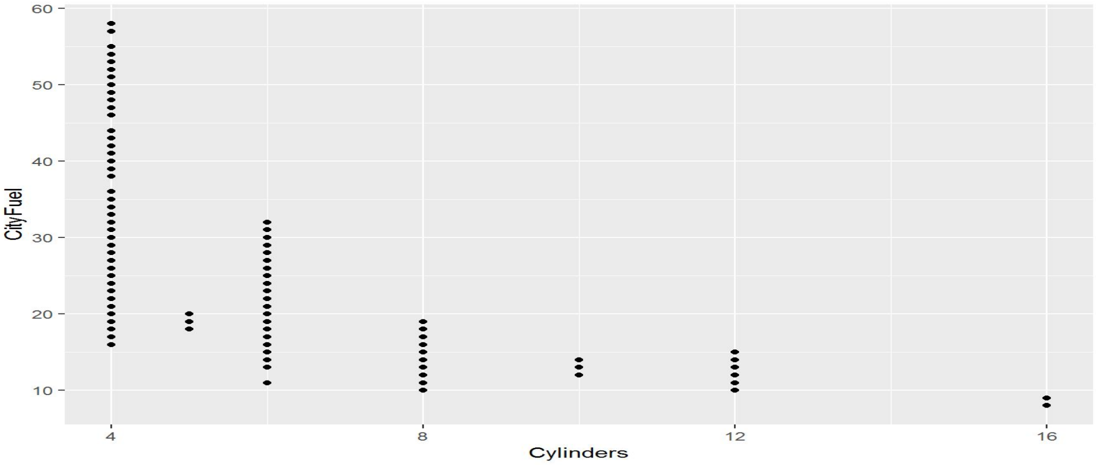
## Carlines

Midsize and compact cars are the most driven cars. Next in line are subcompact and small suv's. There could be several factors like vehicle price, family size, daily distance travelled etc influencing this distribution but will look at how the fuel consumption varies between the car lines.



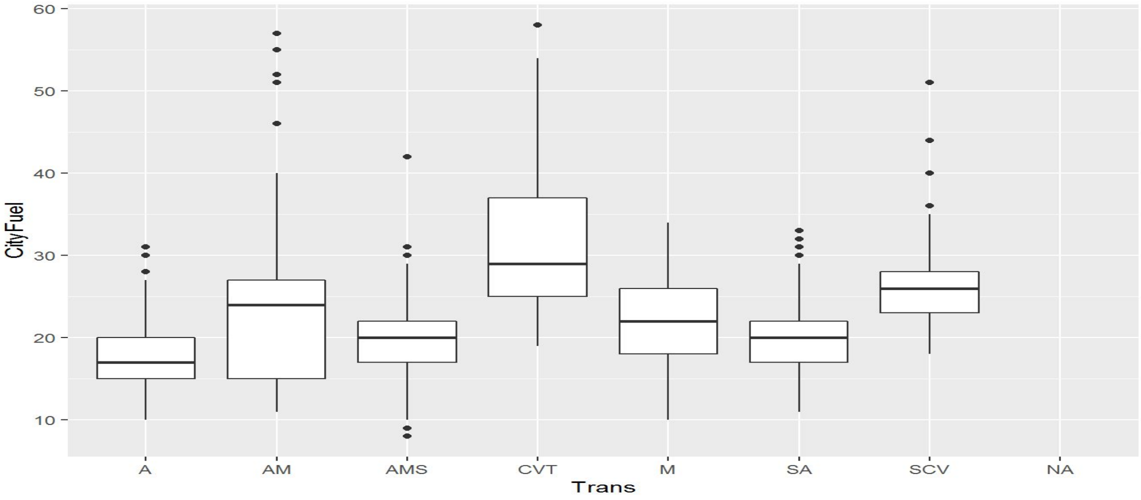
# Relation of variables on fuel consumption

The plot shows that as the number of cylinders increases, the city fuel efficiency decreases. In other words, lower the cylinders, the better is the fuel consumption. So it looks like there is a negative correlation on fuel efficiency with increase in number of cylinders. To understand the extent of significance , statistical analysis is run on these 2 variables and plots are shown in the later part.



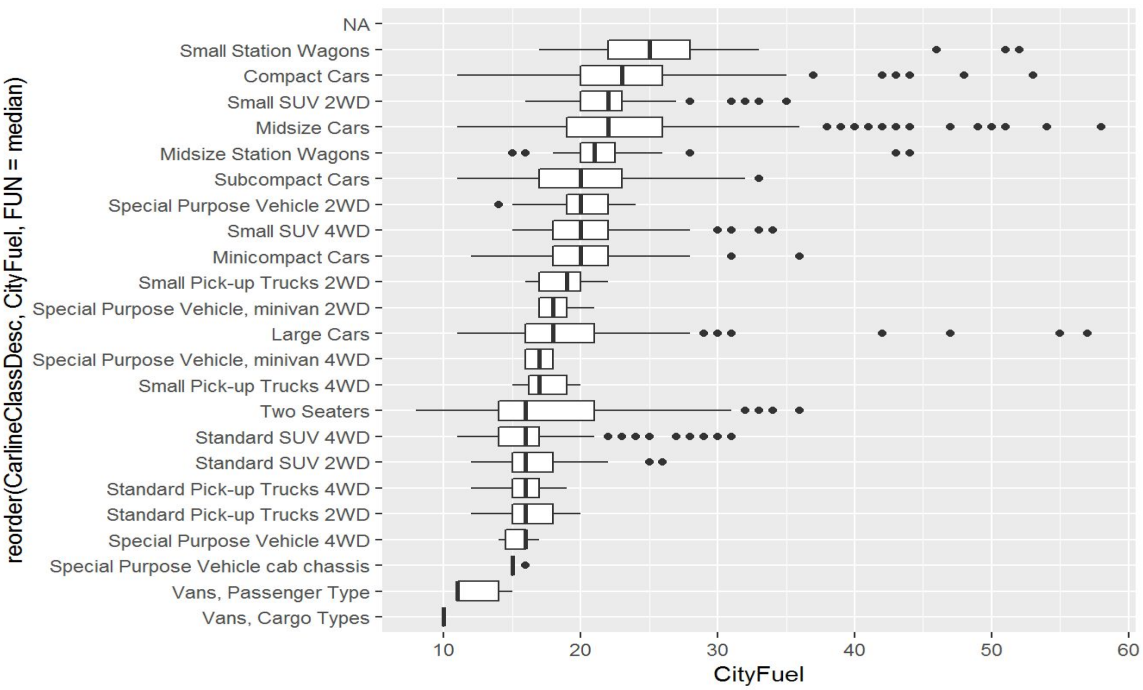
### Transmission type on cityfuel

CVT cars have better fuel consumption (mpg) , better than manual. In the above bar charts, we have seen the manual cars are used lower than semi automatic and automatic. The low fuel efficiency in manual can be one of the reasons, they are driven less.



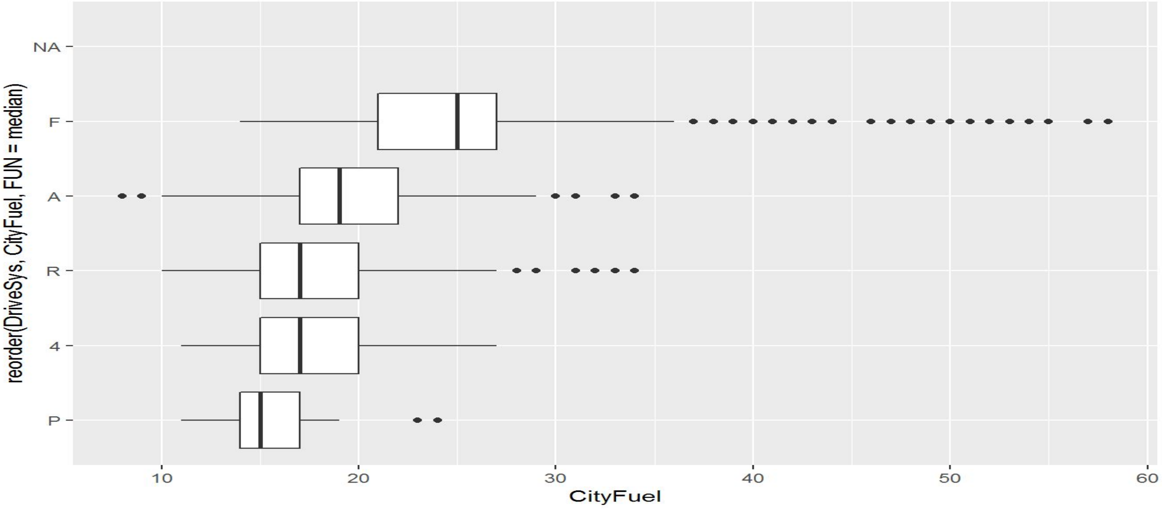
### Carlineclassdesc and cityfuel

Order of better fuel consumption, from highest to lowest - Small station wagons, compact cars, small suvs, mini cars, suvs and trucks. So, small station wagons followed by compact cars give better miles per gallon.



### DriveSys on cityfuel

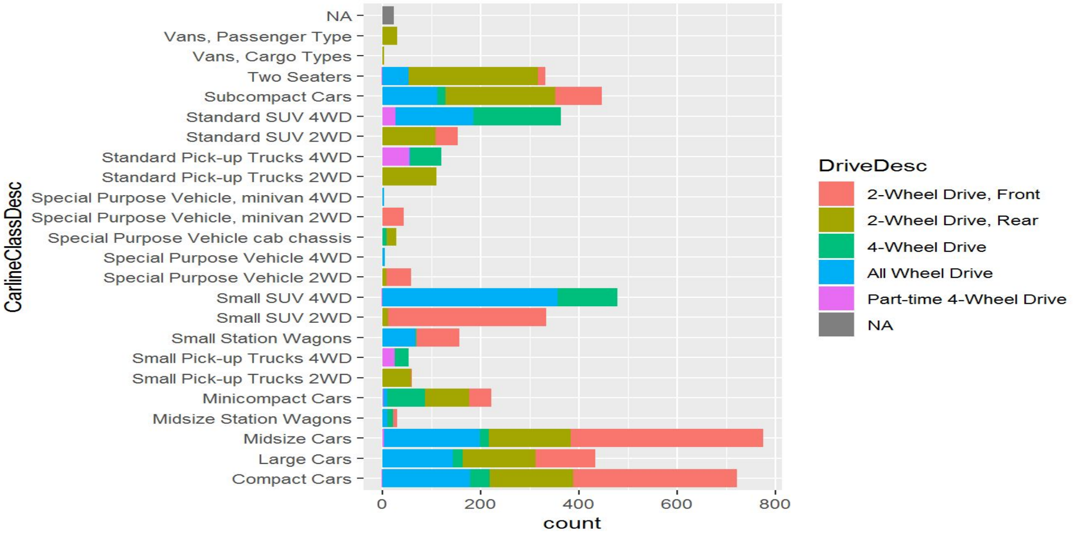
2 wheel front drive cars have better mpg than other drive systems. All wheel drive vehicles are the next fuel efficient. So, to get better fuel usage, the better option is a 2 wheel front drive.



## Multivariable distribution

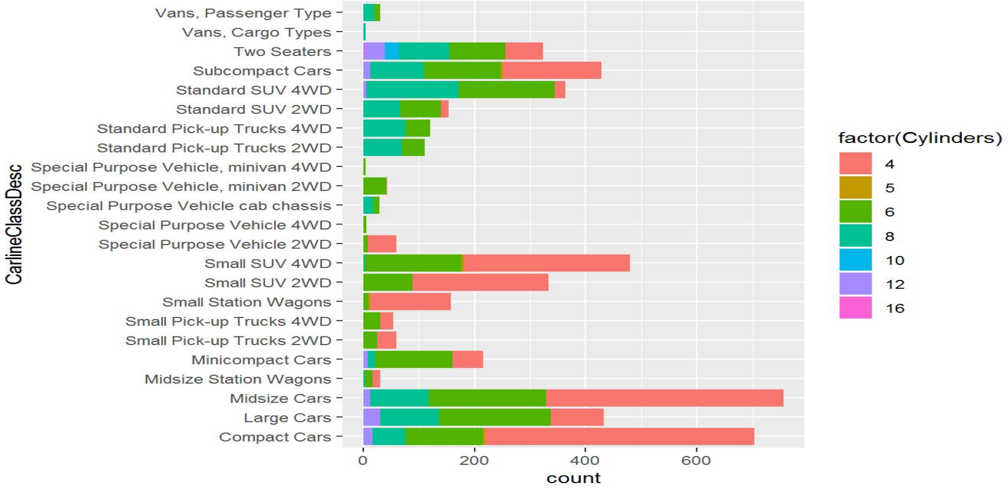
### Carline and DriveDesc

Below plot shows the distribution of drive system within each car line. As you saw above, 2 wheel drive front has higher fuel efficiency. And, compact, small suv’s and midsize are more fuel efficient cars. The plot below shows that the fuel efficient car lines comprise of more 2 wheel drive front drive systems.



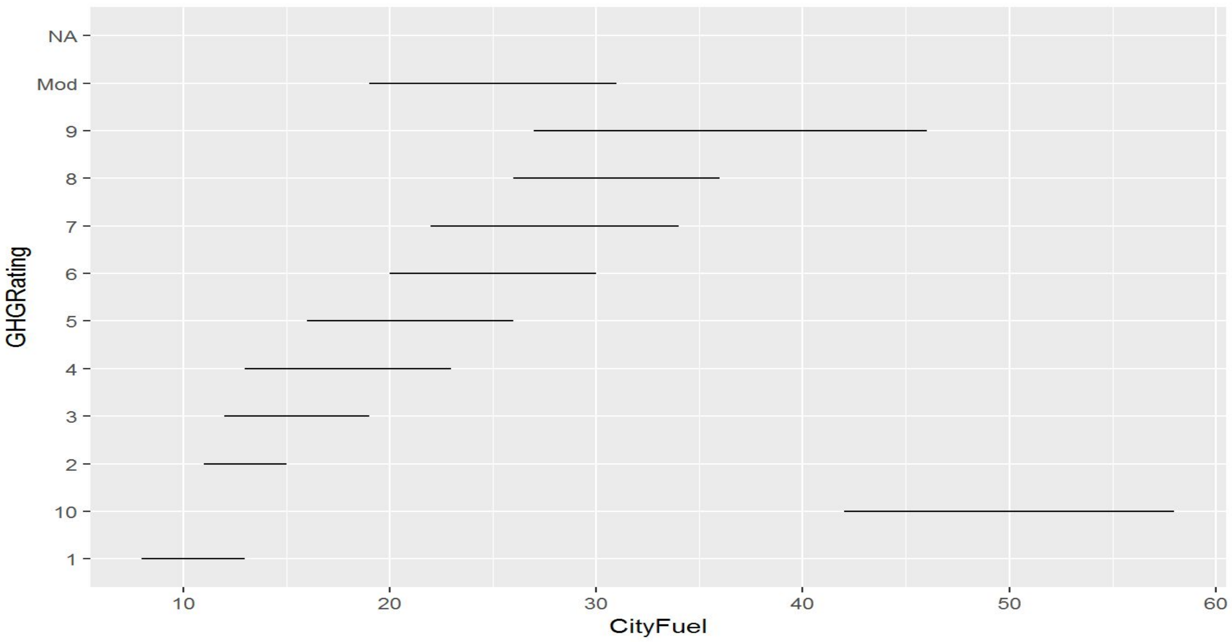
## Carline and Cylinders

Below plot shows the distribution of Cylinders within each car line. As you saw above, lower number of cylinders has higher fuel efficiency. And, compact, small suv's and midsize are more fuel efficient cars. The plot below shows that distribution of 4 cylinders are higher in compact, midsize and small suv's.



## Fuel consumption and green house rating

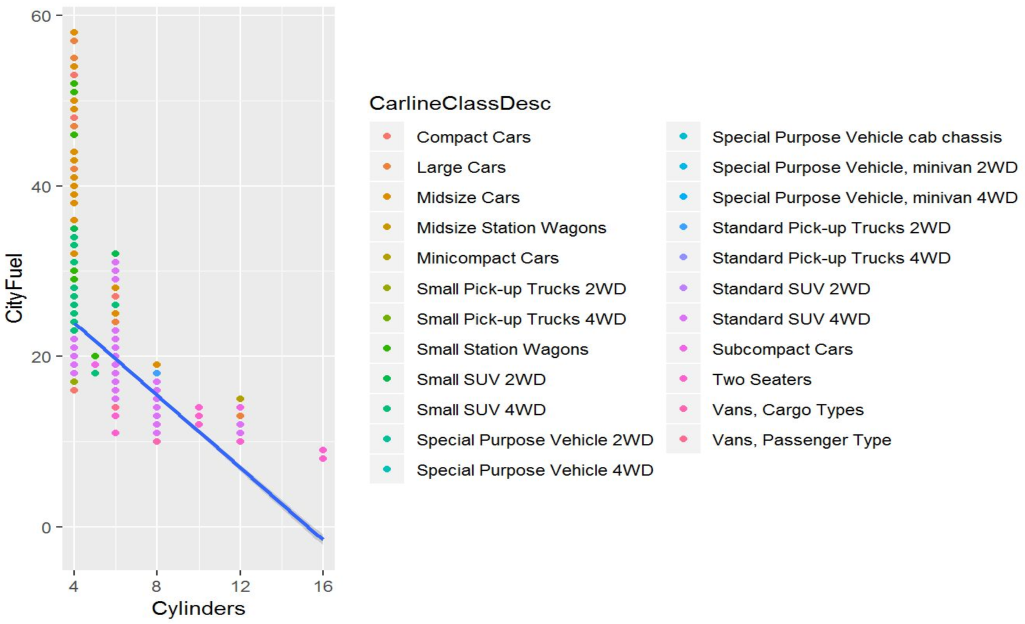
(GHG 1-10 , 10 being good rating) Higher fuel efficiency is associated with better GHG emissions and better GHG rating. So cars with higher miles per gallon are more environmental friendly.



## Statistical Analysis

### Linear regression of cylinders and Cityfuel

A scatter plot gives a comparison between 2 variables. Below is a scatter plot showing the visual comparison of City fuel (mpg) and cylinders. The below scatter plot also use a regression line to see the decline of fuel efficiency (mpg) from the 4 cylinder vehicles to the 8-12 cylinder vehicles. The line shows how mpg decreases with increasing number of cylinders. The car lines that have high fuel efficiency are compact, subcompact and small suv’s with 4 cylinders.

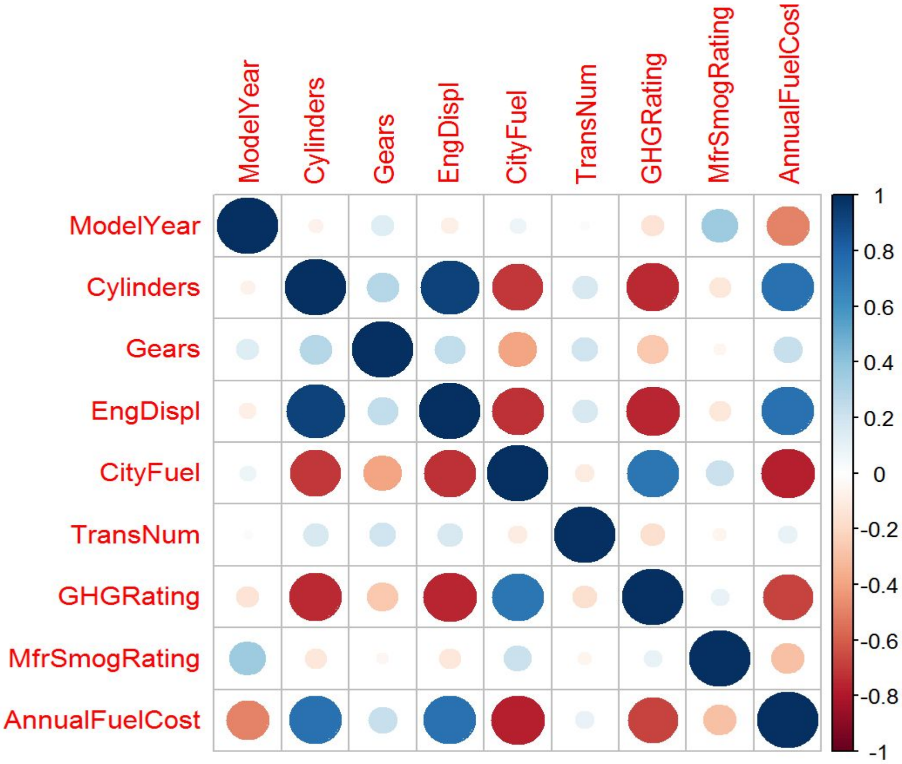


### Finding correlations between different variables.

Key observations are cylinders having moderate -ve correlation with fuel consumption , i.e higher the number of cylinders, lower the fuel efficiency. So, cars with lower number of cylinders give more mile per gallon.
Cylinders and engine displacement show positive correlation. So, Engine displacement correlation to city fuel is similar to Cylinders. Higher the engine displacement, lower is the fuel efficiency.
City fuel and gears show moderate negative correlation. Higher the number of gears, lower the fuel efficiency.
city fuel and annual fuel cost show moderate to good -ve correlation. The fuel cost spent on cars with higher mile per gallon is lower than fuel cost spent on cars with lower miles per gallon.
City fuel and GHG rating show moderate positive correlation. Higher the mile per gallon, higher the GHG rating (lower the emissions). Higher GHG rating is better for the environment.

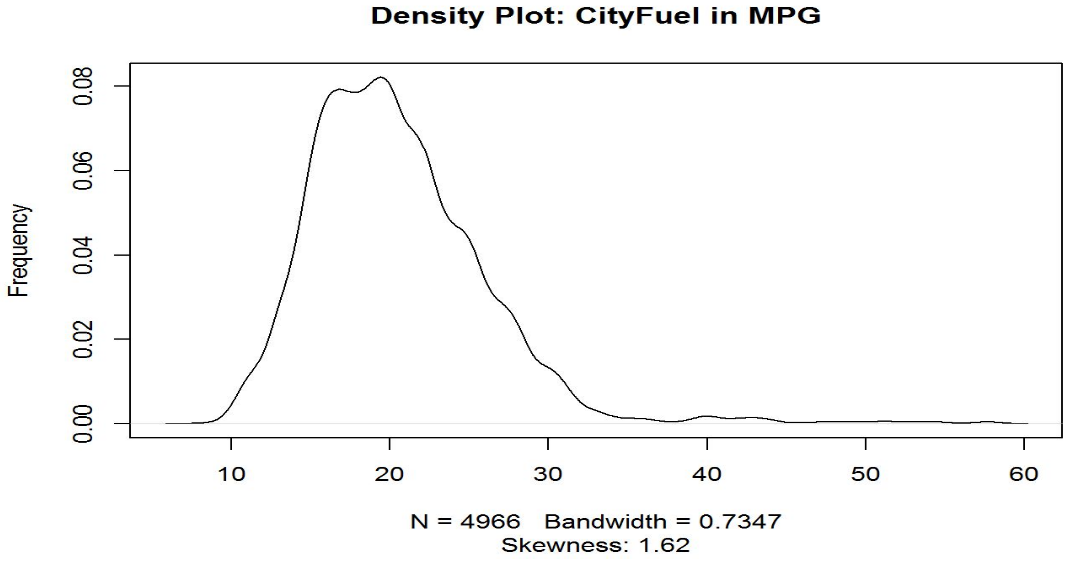
### Correlation between City Fuel and Engine Displacement

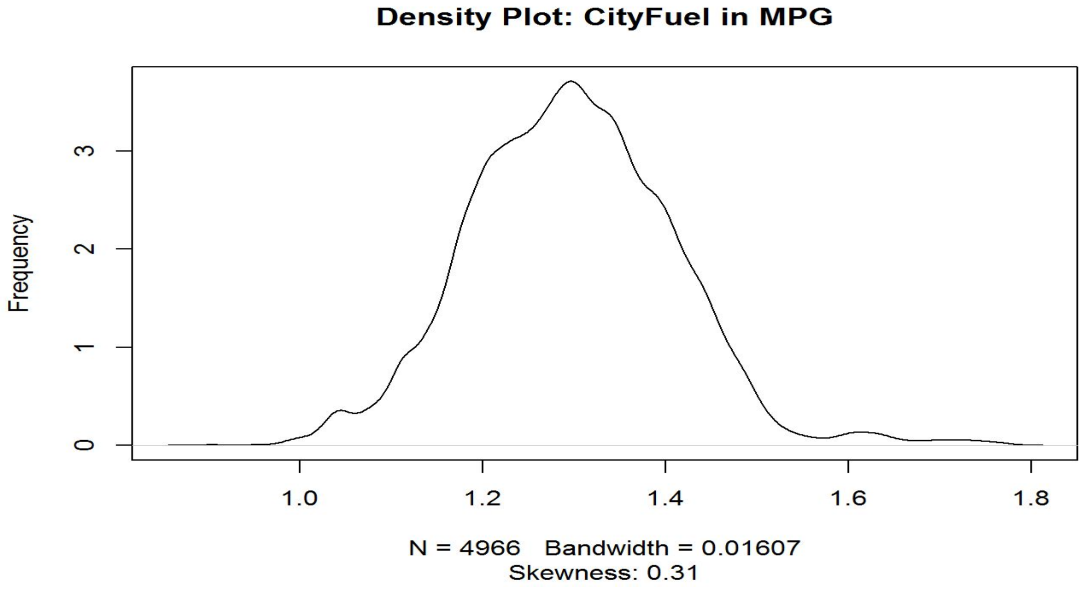
The engine displacement (eng size, , eng volume in liters) is one of the features in manufacturer configurations. It is highly negatively correlated with City Fuel - the higher the eng size, the lower is the fuel efficiency
Correlation :  -0.8364356


ModelYear
Cylinders
Gears
EngDispl
CityFuel
TransNum
GHGRating
MfrSmogRating
AnnualFuelCost
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

### Linear Model of all features on City Fuel (measured in miles per gallon (mpg))

The model is to predict the impact of multiple variables on city fuel (mpg). The features included in the model are engine displacement, transmission type and gears. Before running the model, checked to see if the city fuel distribution is normal. The distribution is not normal and skewed positively to right. The distribution curve of city fuel after log transformation shows normal curve. Scaled all the numerical variables of the data set. Ran the single linear regression first.





## Results from the Linear regression models

### Single Linear Regression , Engine Displacement on City Fuel

Having seen the correlation between engine size and city fuel, a linear regression model was run to look at how the significance is. The significant codes indicate that the feature Engine Displacement is very likely to have an impact on the dependent variable. But we need to see how significant it is. As the output shows, the chosen feature engine displacement explains 70% observed variance on the dependent variable city fuel.

Key terms of linear model explained :

**Multiple R-squared** : R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.It is the percentage of the response variable variation that is explained by a linear model.

**Adjusted R-squared** : Adjusted R-squared adjusts the statistic based on the number of independent variables in the model.

**T-value** : A larger *t-value* indicates that it is less likely that the coefficient is not equal to zero purely by chance. So, higher the t-value, the better

**P-value** : If p-value is low, the coefficient values are significant. If The value is < 0.05, the results are significant at 95% Lower the p-value, the better.

Results of single linear model (city fuel ~ engine displacement)

| | | |
|---|---|---|
| Multiple R-squared | 0.6996 | This explains 70% of observed variance on the City fuel |
| Adjusted R-squared | 0.6996 | As it single linear model, the value is same as multiple-r-squared. |
| T-value | 107.5 | In this case, T-value is higher indicating the correlation exist between engine displacement and City fuel and its not by chance. |
| P-value | 2e-16 | . Shows the results are significant |

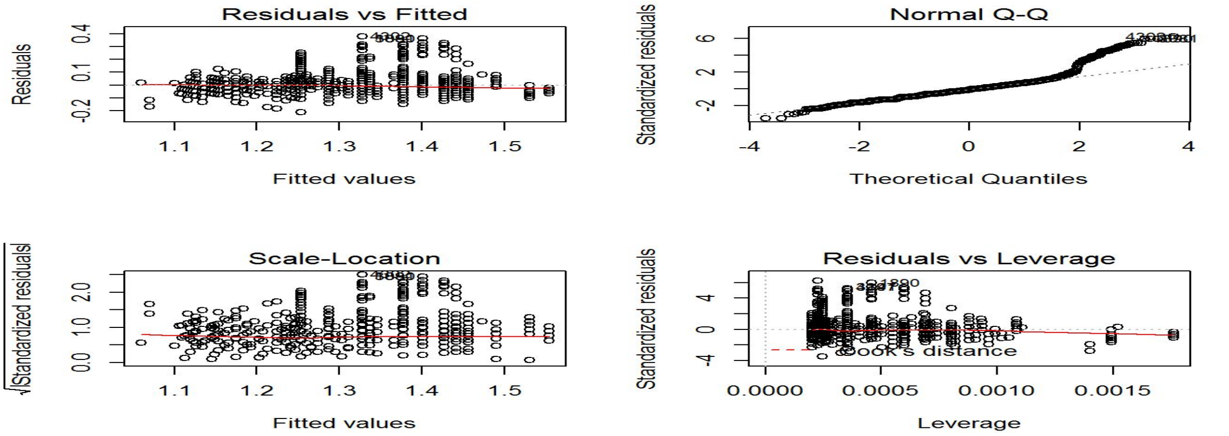
The below plots show if the model works well for data.

**Residuals vs Fitted** - Residuals are leftover of the outcome variable after fitting a model to data and they could reveal unexplained patterns in the data by the fitted model. This information, can be used to check if linear regression assumptions are met. This plot shows if residuals have non linear pattern. The plot shows that there is some non linearity with residuals spread beyond the horizontal line on the upper half.   The model does not show 100% linearity.

**Normal Q-Q -** This plot shows if residuals are normally distributed. The residuals are lined well on the straight dashed line but after mid way, they deviate from the normal curve.

**Scale-Location -** It’s also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predicting features. This is to check the assumption of equal variance (homoscedasticity. Most of the points are spread equally around the horizontal line but there is spread of points away from horizontal line in the upper half.

**Residuals vs Leverage -** This plot helps to find influential outliers if any. The outlying values are spotted at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. The cases outside of a dashed line, Cook’s distance. The regression results will be altered if we exclude those cases. There seems to be no influential outliers.



Gvlma results for single linear model. Gvlma is a package that runs global validation of Linear models assumptions. This is run as a double check on the above plots.
Global Stat-  Indicates if the relationships between predictors and outcome variable are linear.
Skewness - Indicates if there is normality or if the distribution is skewed positively or negatively
Kurtosis- Indicates if distribution is kurtotic (highly peaked or very shallowly peaked),
Link Function- Indicates if the Is your dependent variables are truly continuous, or categorical.
Heteroscedasticity- Indicates if the variance of the model residuals is constant across the range of X (assumption of homoscedasticity).

| Assumption | Satisfied or not | Explanation |
|---|---|---|
| Global Stat | Assumptions not satisfied | Linearity not met |
| Skewness | Assumptions not satisfied | Normal curve is skewed |
| Kurtosis | Assumptions not satisfied | Distribution in kurtotic, dataset is tail heavy |
| Link function | Assumptions not satisfied | Dependent variable are not continuous |
| Heteroscedasticity | Assumptions acceptable | Variance of model residuals not constant |

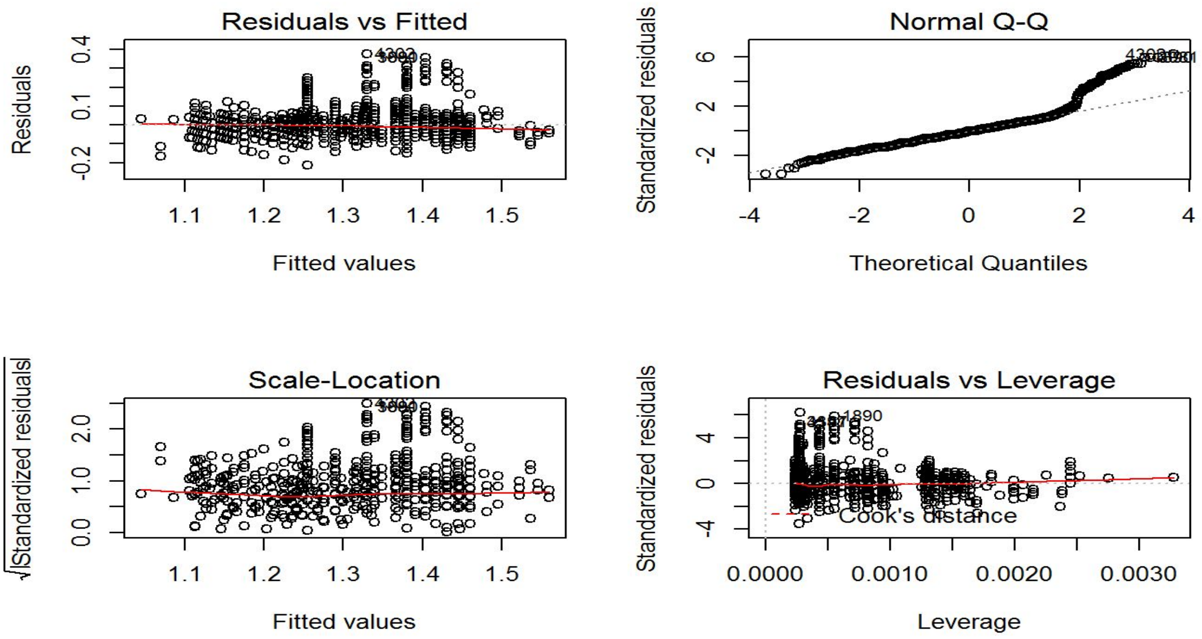
### Multiple Regression model , Engine Displacement and Transmission on City Fuel

One of the features of manufacturer configuration is transmission type. Created a two level categorical variable for transmission type and ran the model with engine displacement and transmission to see the impact on City fuel. As the output shows, both features show they have very likely to have impact on City Fuel. And, compared to the single linear model with just engine size, this model is same as the other give that residuals remained almost same..
Assumptions for this model as well show slight non linearity and normality is not met.

Results of multiple linear model (city fuel ~ engine displacement + transmission)

| | | |
|---|---|---|
| Multiple R-squared | 0.7016 | This explains 70% of observed variance on the City fuel |
| Adjusted R-squared | 0.7015 | This value is slightly higher than value from single linear model (0.69) which indicates this model is only slightly better by adding transmission (transmission coefficient shows only 13% impact on city fuel) |
| T-value | 106.4/5.72 | In this case, T-value is higher indicating the correlation exists between selected features and City fuel and its not by chance. |
| P-value | 2.26e-16 | . Shows the results are significant |

The residual vs fitted plot below shows some non linearity and Q-plot does not meet normality. There seems to be no influential outliers in the residuals vs leverage plot.



Gvlma results for multiple regression model are same as for single linear model with only the homoscedasticity assumption met.

| Assumption | Satisfied or not | Explanation |
|---|---|---|
| Global Stat | Assumptions not satisfied | Linearity not met |
| Skewness | Assumptions not satisfied | Normal curve is skewed |
| Kurtosis | Assumptions not satisfied | Distribution in kurtotic, dataset is tail heavy |
| Link function | Assumptions not satisfied | Dependent variable are not continuous |
| Heteroscedasticity | Assumptions acceptable | Variance of model residuals not constant |

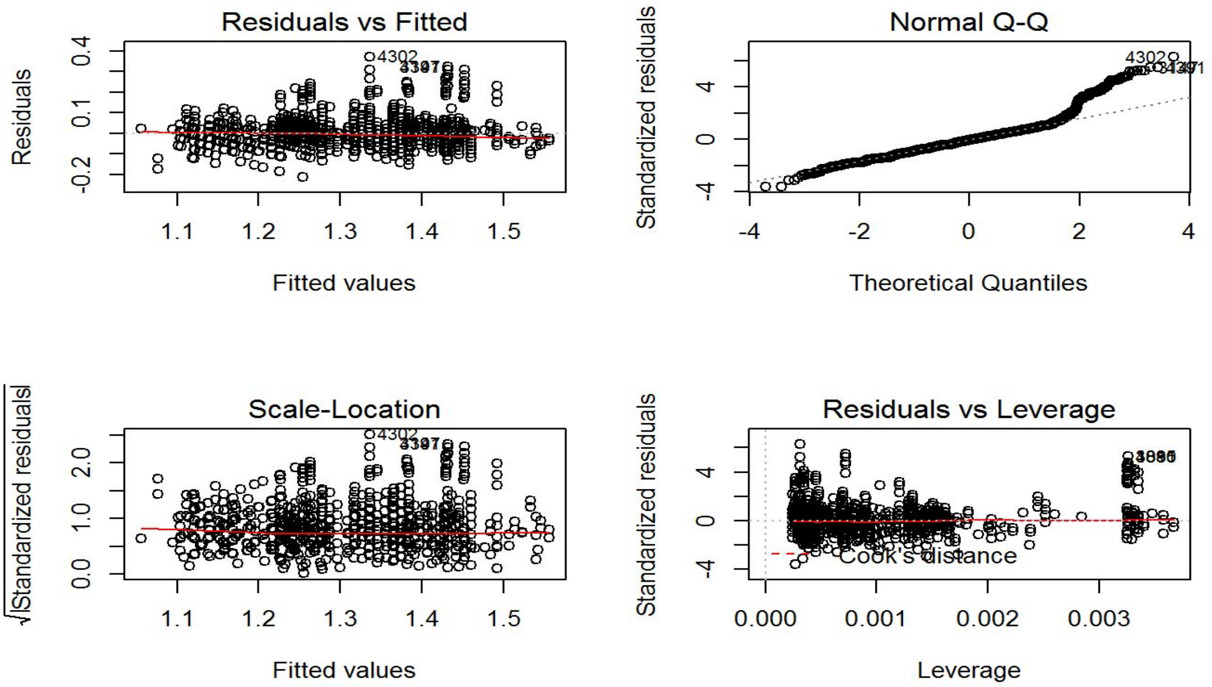
### Multiple Regression model , Engine Displacement , Transmission and Gears on City Fuel

In this model, included Gears as feature to see the impact on City Fuel. This model shows the highest residual and adjusted residuals , 0.72, showing it is the best model out of the 3 to predict City fuel.
From an assumptions standpoint, this model has homoscedastic residuals (homoscedasticity met) and also shows the features are truly continuous. (link function assumption met) Although , we still see some non linearity and skewness, not meeting the assumption of normality.

Results of multiple linear model (city fuel ~ engine displacement + transmission + gears)

| | | |
|---|---|---|
| Multiple R-squared | 0.7158 | This explains 72% of observed variance on the City fuel |
| Adjusted R-squared | 0.7156 | This value is slightly higher than value from both single linear model (0.69) and multiple with transmission (0.71) which indicates this model is the best among the three models. |
| T-value | 101.4/8.06/15.7 | In this case, T-value is higher indicating the correlation exists between selected features and City fuel and its not by chance. |
| P-value | 2.2e-16 | . Shows the results are significant |

The residual vs fitted plot below shows some non linearity and Q-plot does not meet normality. There seems to be no influential outliers in the residuals vs leverage plot.



Gvlma results for this multiple regression model are the best with 2 assumptions met. The variance of residuals is constant and the chosen features are continuous.

| Assumption | Satisfied or not | Explanation |
|---|---|---|
| Global Stat | Assumptions not satisfied | Linearity not met |
| Skewness | Assumptions not satisfied | Normal curve is skewed |
| Kurtosis | Assumptions not satisfied | Distribution in kurtotic, dataset is tail heavy |
| Link function | Assumptions acceptable | Dependent variable are continuous |
| Heteroscedasticity | Assumptions acceptable | Variance of model residuals not constant |

# Summary of Data Analysis

The exploratory data analysis shows the distribution of variables like cylinders, engine displacement, drive systems, transmission, carlines etc  in the data set. Also within each of the carlines ,the distribution of these variables was studied. And the relation of these variables on the fuel efficiency was studied. Some key observations on the relation of the variables  are :

- Lower the cylinders, the better is the fuel consumption. So , there is a negative correlation on fuel efficiency with increase in number of cylinders.
- Cylinders and Engine Displacement share multicollinearity relationship.
- Higher the number of gears, lower the fuel efficiency.
- Automatic cars have better fuel consumption (mpg) , better than manual.
- Small station wagons followed by compact cars give better miles per gallon. The distribution of 4 cylinders are higher in compact, midsize and small suvs. Mini cars, suvs and trucks come next to compact cars in fuel efficiency, suvs and trucks being the least fuel efficient.
- 2 wheel front drive cars have better mpg than other drive systems. The fuel efficient car lines comprise of more 2 wheel drive front drive systems.
- Higher fuel economy is associated with better GHG emissions and better GHG rating. So cars with higher miles per gallon are more environmental friendly.

Model used and Prescription :

Linear regression model was chosen to show the impact of selected features on City Fuel. The manufacture configuration includes Cylinders, Engine displacement, transmission type. So selected these as the features for the dependent variable City Fuel. Cylinders was excluded in the model, as it shows multicollinearity with Engine displacement. Gears show moderate correlation with city fuel, so added gears as another feature in multiple regression

- All 3 features run together in the model showed significant and better impact (72%) on city fuel, compared to each of them individually. The model would have been a good fit if all the assumptions of linear model are met. But the model shows some non linearity and skewness in normality. Although the model shows significance, we cannot use the model to predict fuel usage as it did not meet all assumptions of linear model.

Future Recommendations :

- The linear model did not meet all assumptions. So looking into nonparametric approaches are useful. These tests do not make assumptions of about the distribution of the data. The nonparametric correlations do not assume that the data are linearly related. Kendall-Theil-Siegel regression fits a best fit line between two variables. It is robust to outliers. Quantile regression is a very flexible approach, and can be used for a linear regression. It is robust to outliers in the y variable.  The model can then be used to predict the fuel usage with the identified car features.

- The dataset has data pertaining to greenhouse emissions, CO2 emissions and cost and savings from fuel usage. Also, the exploratory analysis showed that the fuel efficient cars have lower emissions and GHG rating is higher (higher the rating,  better it is on environment). Using this, the analysis can be extended to predict the impact of city fuel usage (along with its features, engine displacement, transmission, gears) on these emissions/savings. This can be attached to the carlines to serve as guideline for consumers when making the purchase.