Introductory Data science and R Programming

Capstone Project

By

Rekha Vellanki

Capstone Project – Brief Description

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. The analysis could show correlations between these factors on fuel usage and help predict fuel usage based on these. Also show the influence of fuel usage on environment impacts based on air pollution score and GHG emissions.

Capstone Project

- Problem Statement :
  - Understand the various factors influencing the fuel efficiency of a car.
  - Study the influence of fuel usage on environment impacts based on air pollution score and GHG emissions

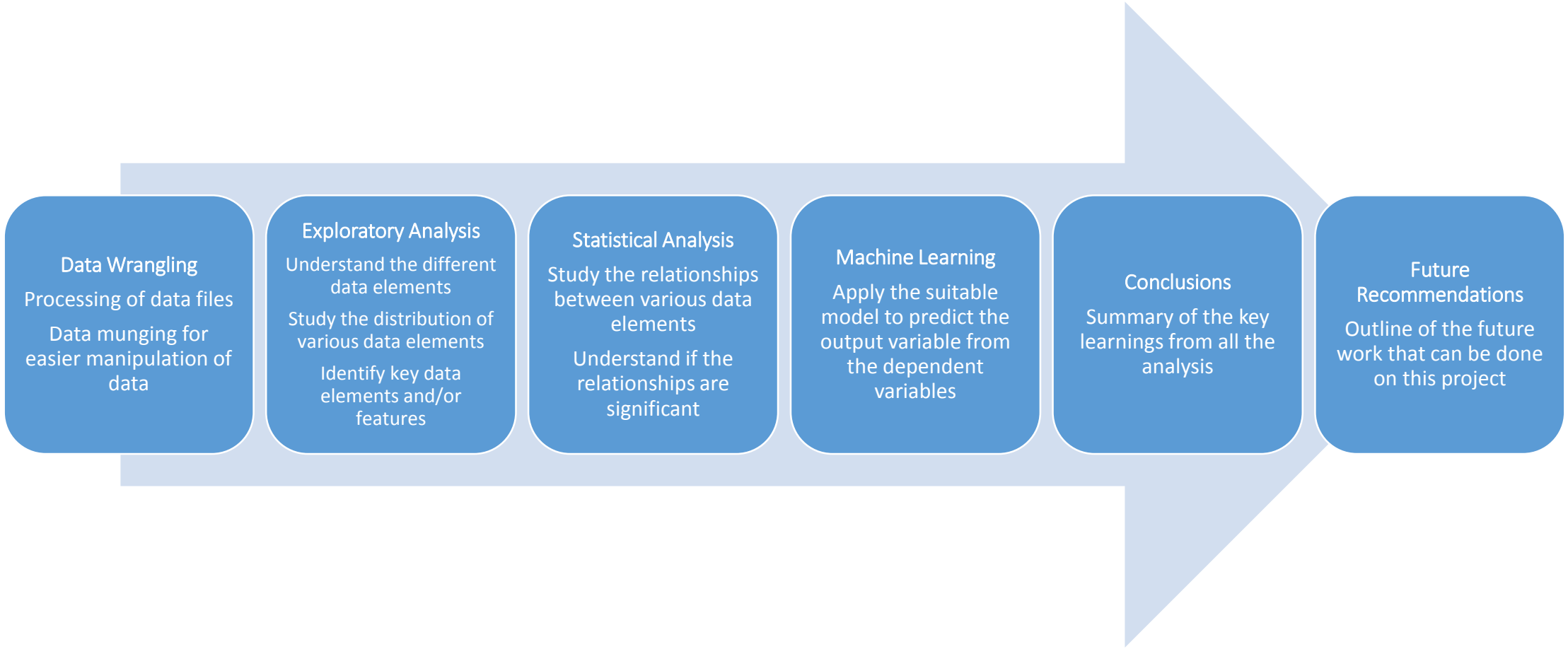
- Client and uses :

  Potential car buyers, car owners wanting to cut fuel costs, car manufactures, environmental enthusiasts. Helps all of them to understand the benefits of efficient fuel usage on environment and come up with more fuel efficient options

- Acquiring the Data :

  The data can be downloaded from the following site :
  https://www.fueleconomy.gov/feg/download.shtml

Outline Approach



Dataset - Key Variable names and description

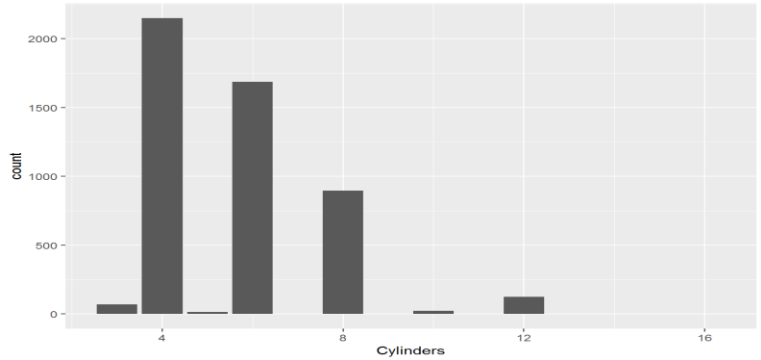
Data Element	Description
CityFuel	Fuel usage measured in miles per gallon (mpg)
Mfrname	Manufacturer name
Cylinders	Number of cylinders (4, 6, 8 etc.)
EngDispl	Engine weight/volume
Carline	Car line type (two seater, midsize, compact etc.)
DriveDesc	Drive type (4 wheel, 2 wheel front, 2 wheel rear)
Gears	Number of Gears (1 to 10)
Transmission	Type of transmission (automated, manual etc.)
GHGRating	Green House Emissions Rating

Data Wrangling

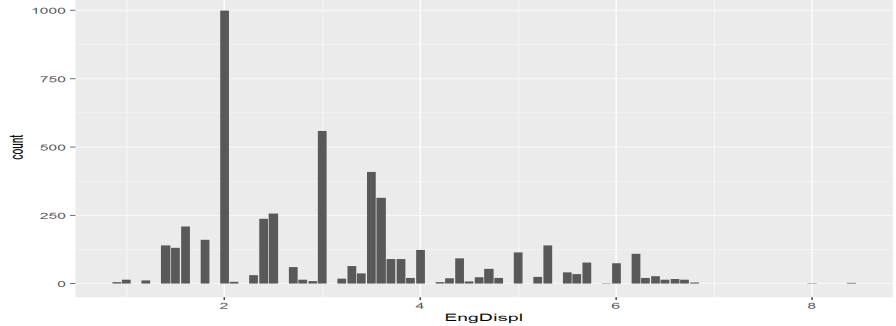
- Extracted the data files from the site
  https://www.fueleconomy.gov/feg/download.shtml
- Identified the data files needed and loaded into data frames
- Empty values of variables with some missing data are replaced with NA for easier data manipulation
- Removed data rows with all empty values
- Discarded the columns not needed for analysis
- Renamed the column names for easy read and data manipulation

Exploratory Analysis (Distribution of key variables)

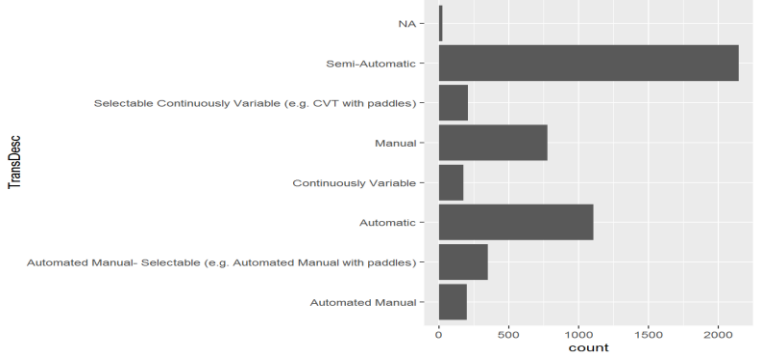
Most cars have 4,6,8 cylinders, highest being 4 cylinder cars



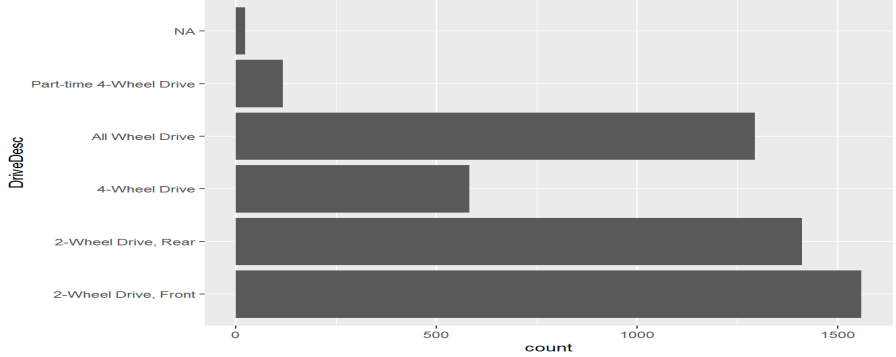
Majority of cars Engine displacement ranging from 2 to 3.5. Higher the cylinders, higher is the engine displacement in general



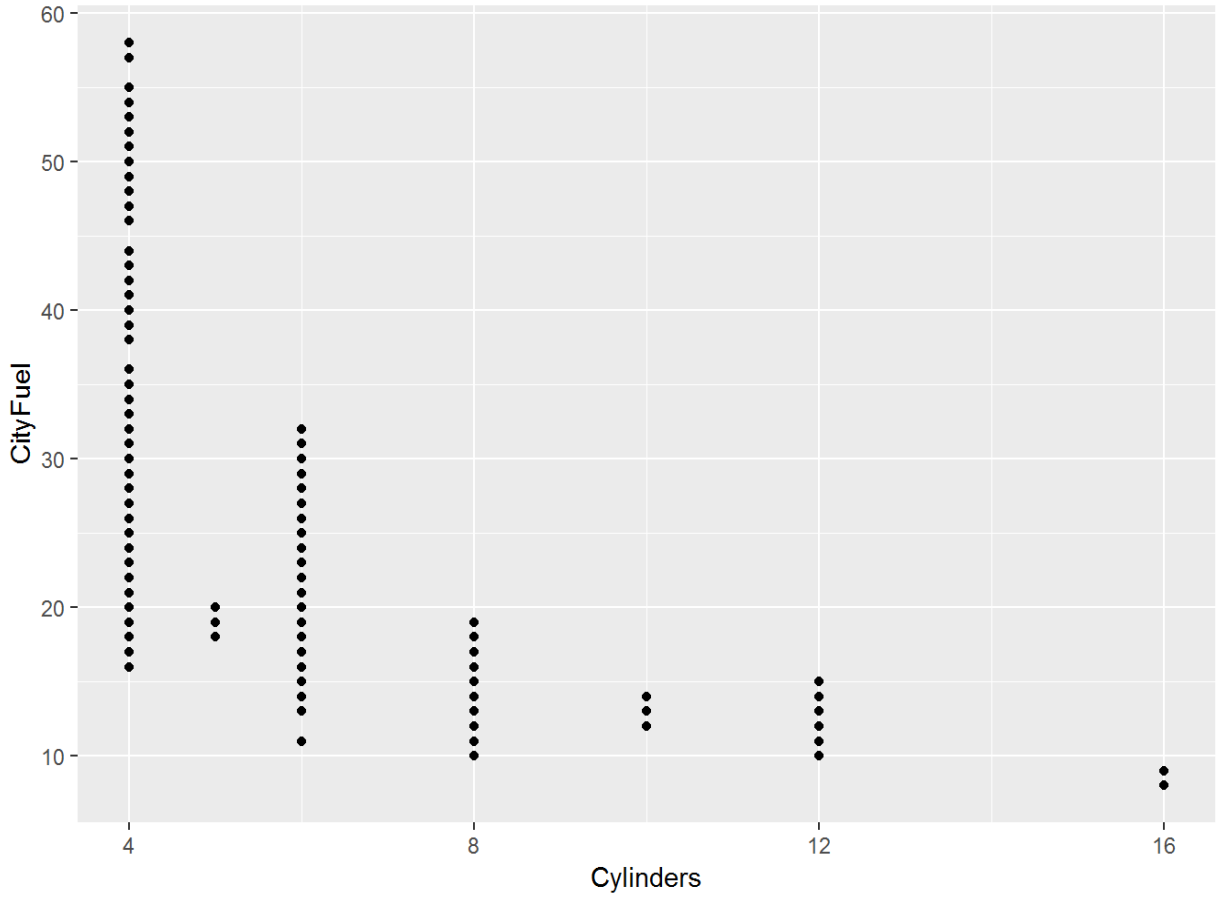
Semi automatic cars are higher in number followed by automatic and manual



2 wheel drive and All wheel drive are higher than 4 wheel drive cars



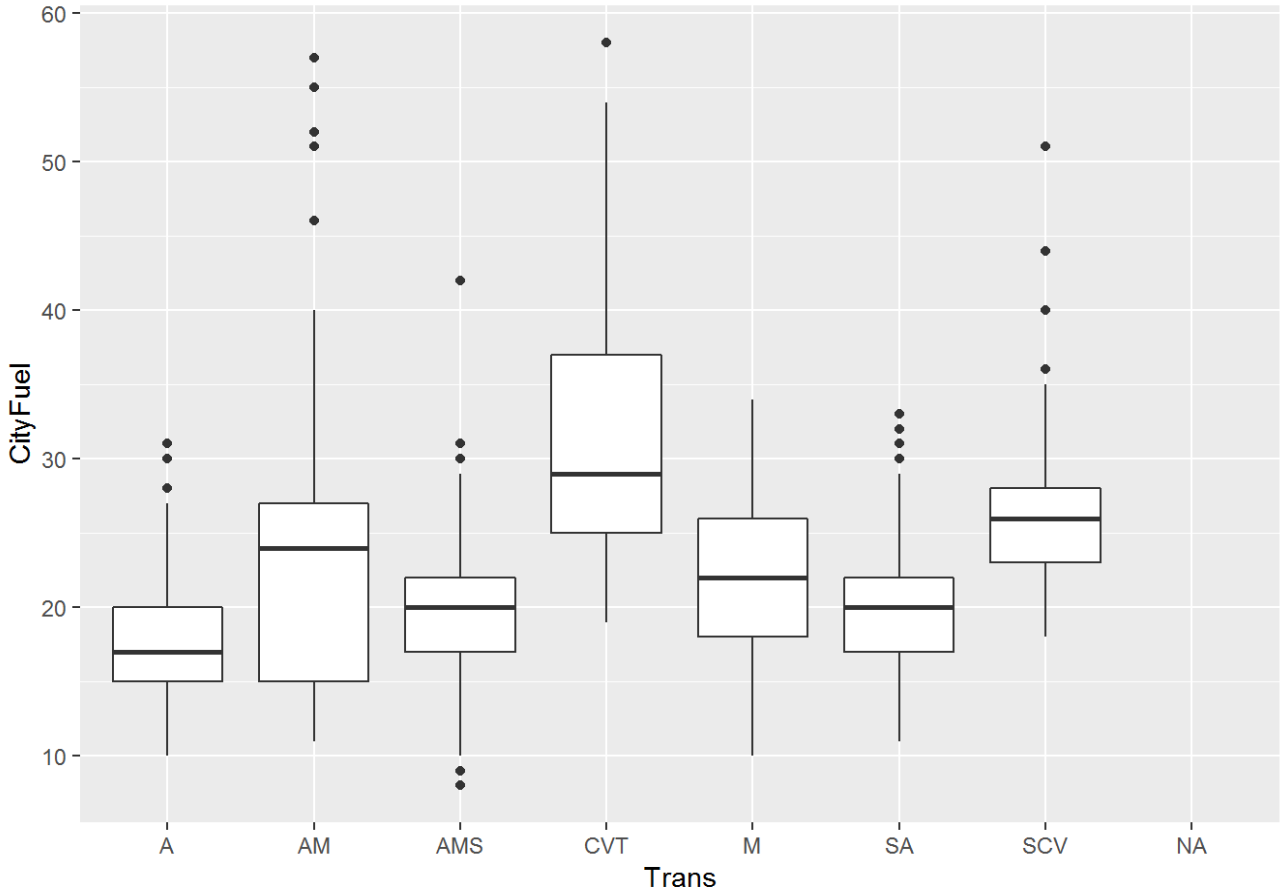
Exploratory Analysis



Impact of Cylinders on fuel efficiency

- 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.

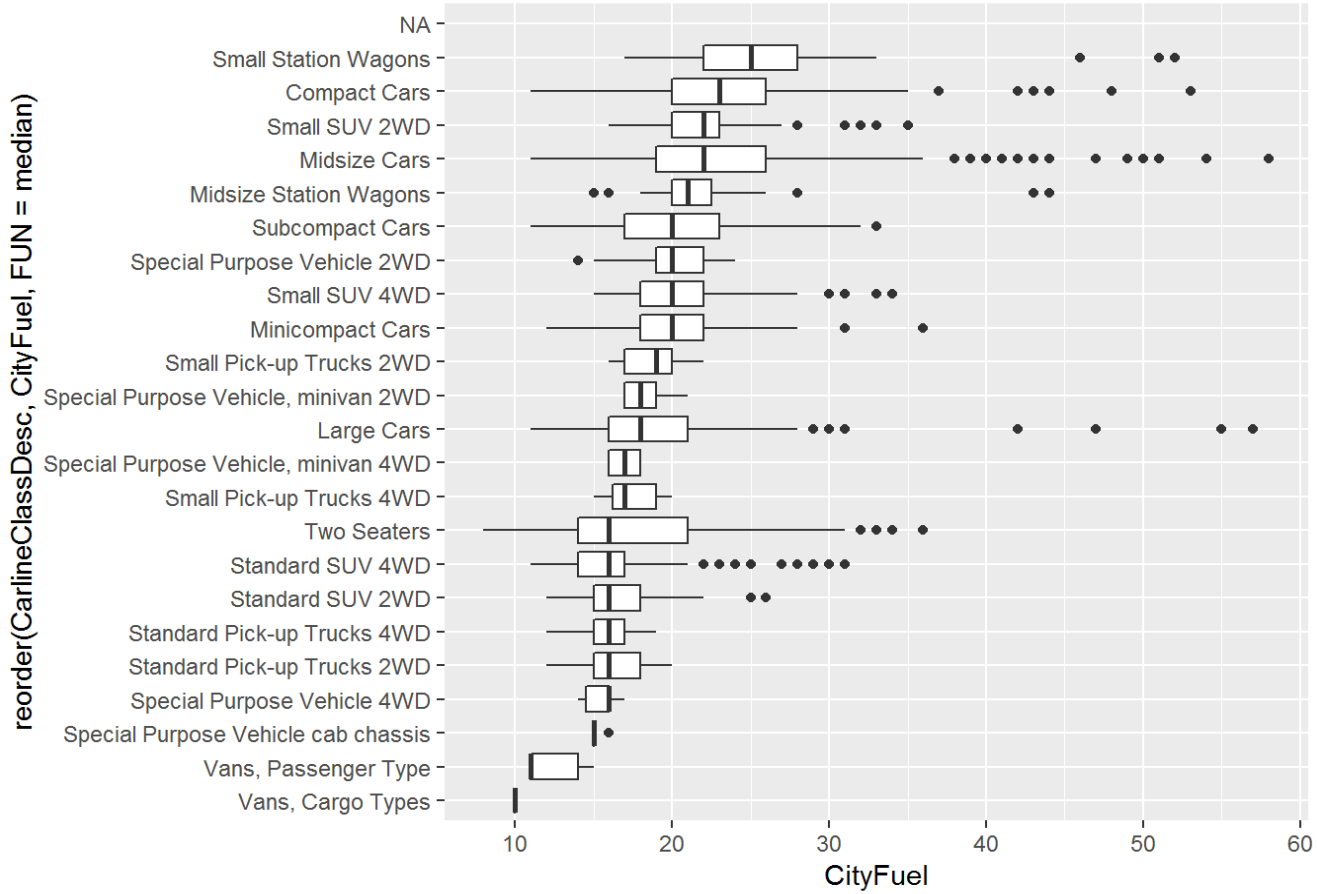
Exploratory Analysis



Impact of Transmission on City Fuel

- CVT cars have better fuel consumption (mpg) , better than manual. 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.

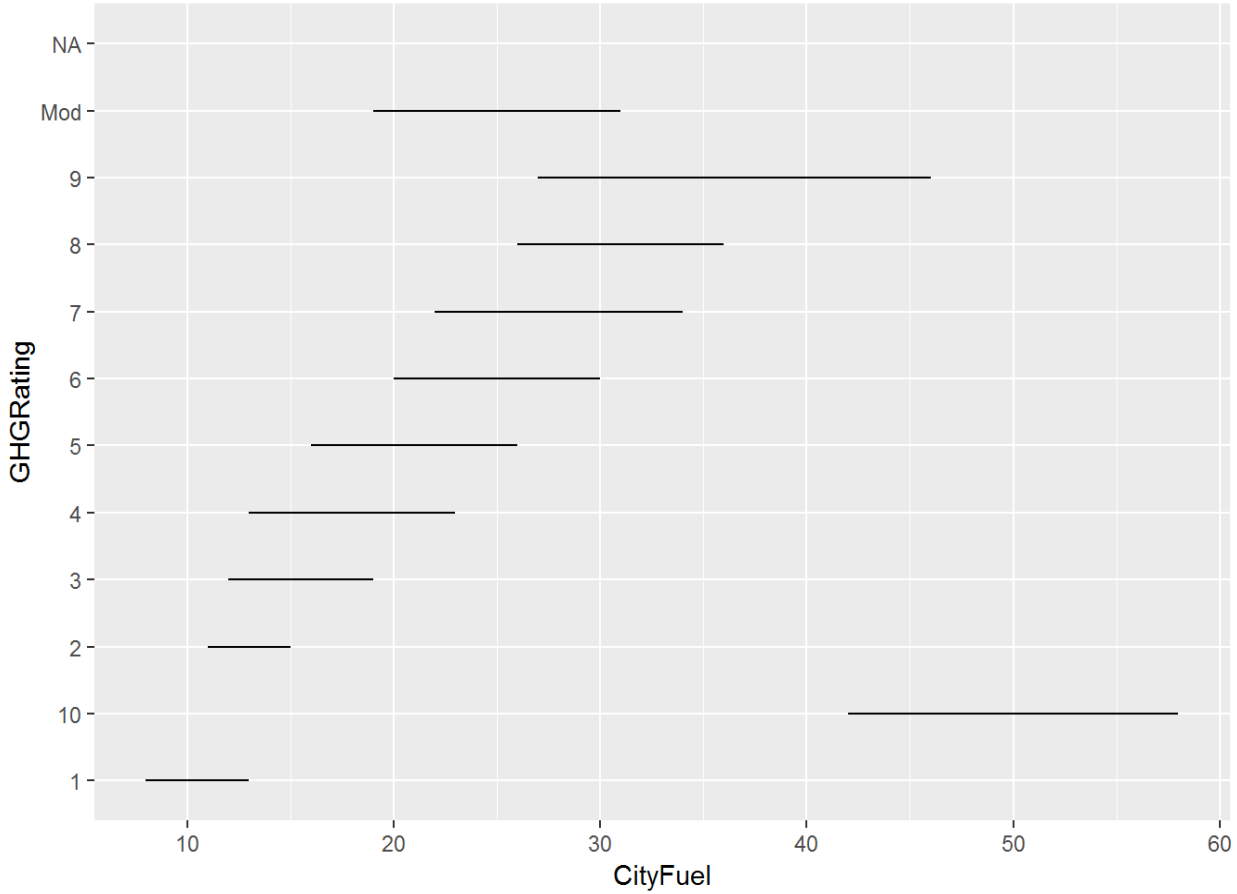
Exploratory Analysis



Impact of Carlines on fuel efficiency

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

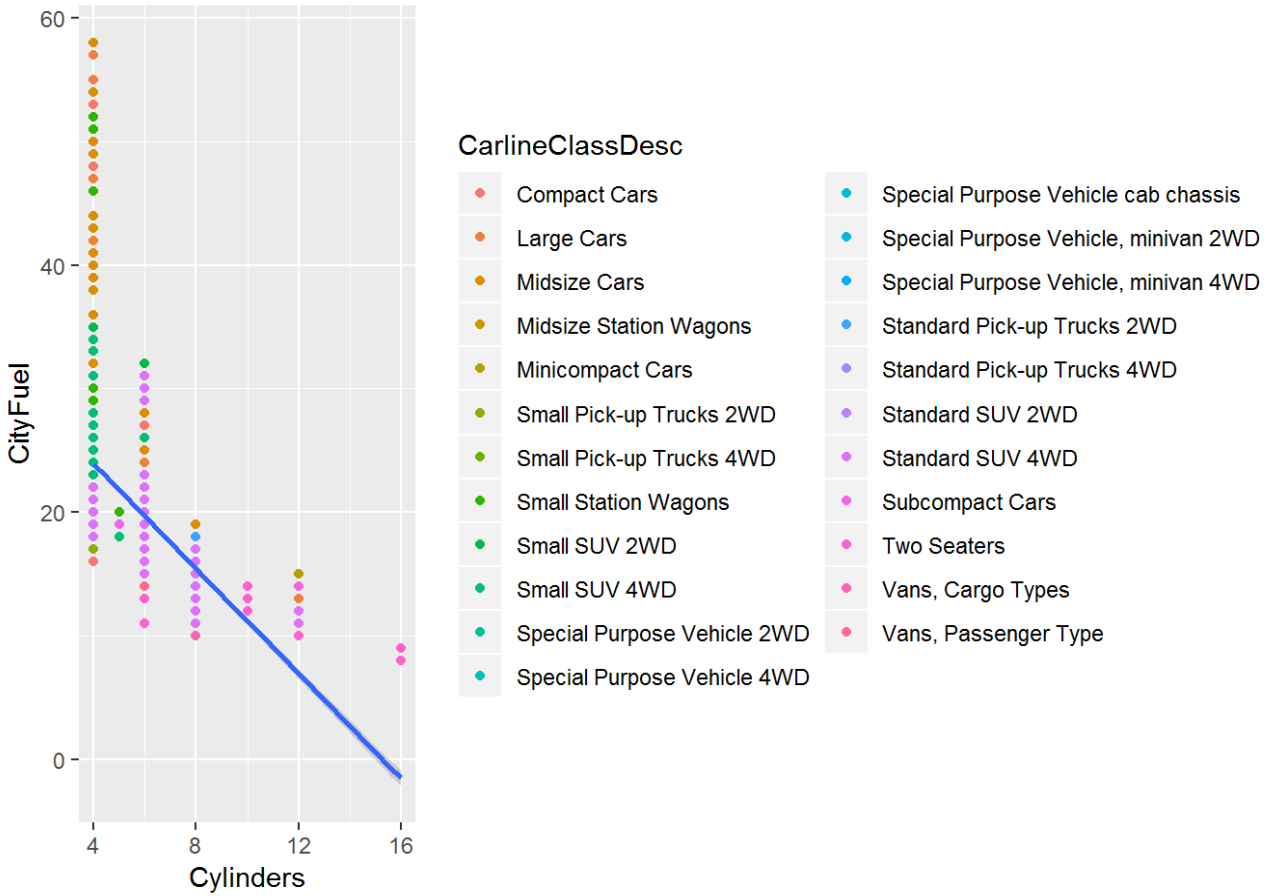
Exploratory Analysis



Fuel efficiency and GHG rating

- (GHG 1-10 , 10 being good rating) Higher fuel efficiency is associated with better GHG emissions and better GHG rating

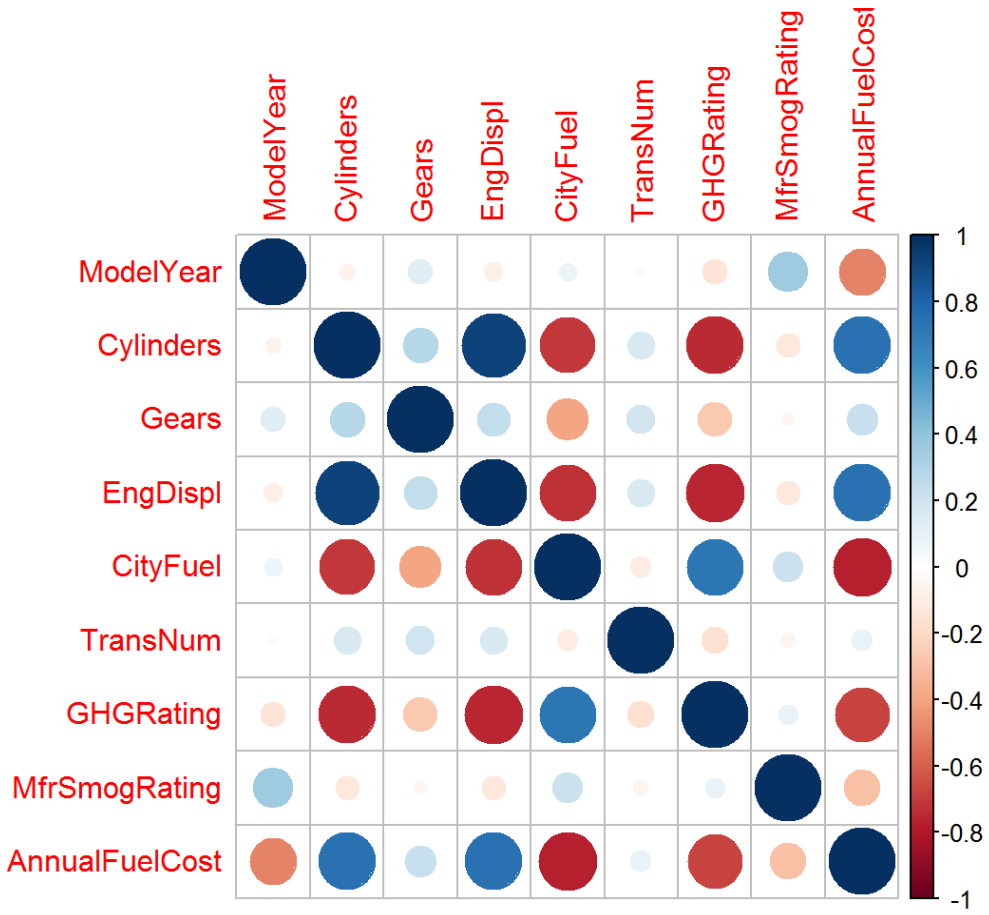
Statistical Analysis (Scatter plot with regression line)



City Fuel and Cylinders

- The scatter plot is showing the visual comparison of City fuel (mpg) and cylinders. The regression line shows 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

Statistical Analysis (Correlation plot)



Key Observations

- Cylinders having moderate negative correlation with fuel consumption , i.e higher the number of cylinders, lower the fuel efficiency.
- 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 negative 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).

Machine Learning (Linear Regression Concepts)

Key terms of linear model

- **Multiple R-squared** : R-squared is a statistical measure of how close the data are to the fitted regression line. 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. Higher the t-value, the better
- **P-val**: 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.

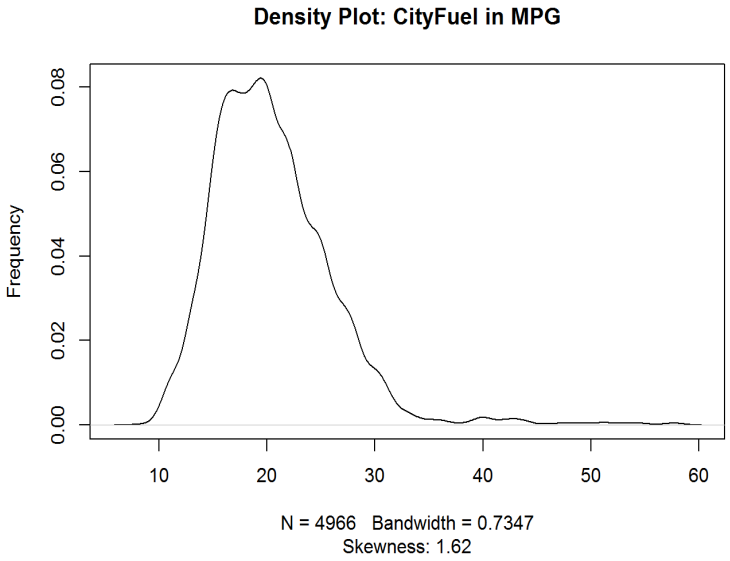
Linear Regression Plots

- Plots run to see if model fits the data best
- **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 plot shows if residuals have non linear pattern.
- **Normal Q-Q -** This plot shows if residuals are normally distributed.
- **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).
- **Residuals vs Leverage -** This plot helps to find influential outliers if any.

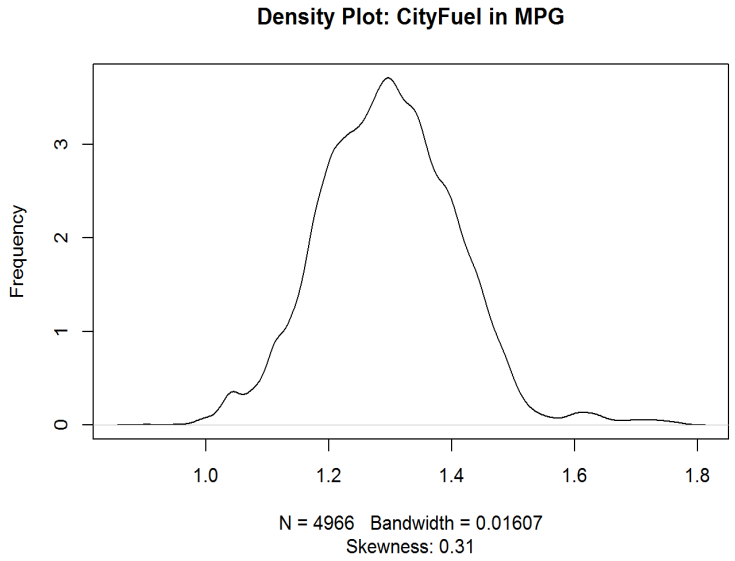
Gvlma package

- Gvlma **i**s a package that runs global validation of Linear models assumptions. This is run as a double check on the linear regression 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 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).

Machine Learning (Distribution Curve)



The distribution curve of city fuel after log transformation shows normal curve.

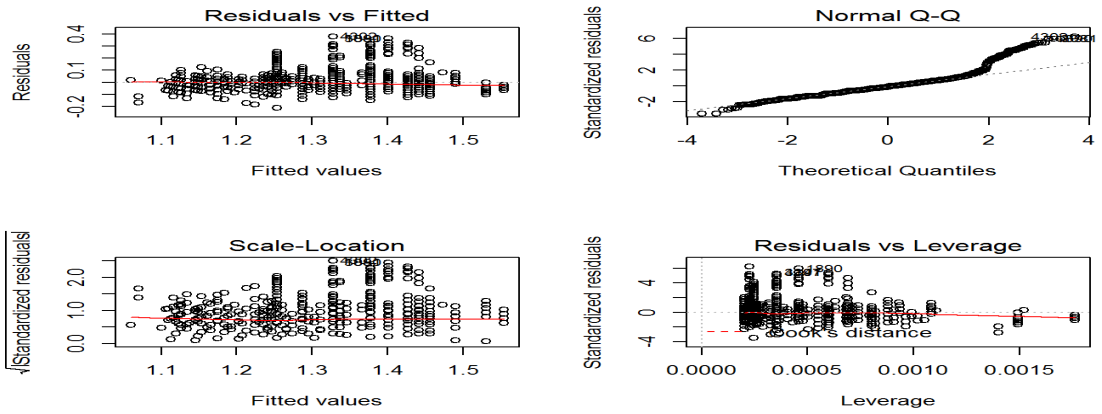


Linear regression (Results of single linear regression (City fuel ~ Engine Displacement))

	Values	Explanation
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

Conclusion : Linear regression results show the model coefficients are significant and the impact of Engine displacement on city fuel is not by chance.
However the plots and gvlma package shows that not all assumptions are met. Hence the model is not a good fit to the dataset.

Residual vs Fitted plot shows some non linearity and normal curve is not met.



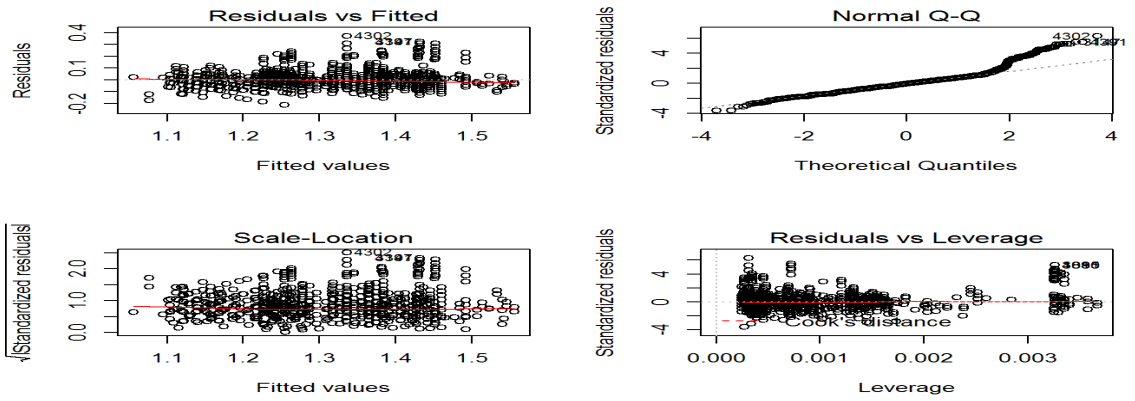
Assumption	Satisfied or not	Explanation
Global Stat	Assumptions not satisfied	Linearity not met
Skewness	Assumptions not satisfied	Not normal distribution
kurtosis	Assumptions not satisfied	Distribution is kurtotic, tail heavy dataset
Link function	Assumptions not satisfied	Dependent variables not continuous
Heteroscedasticity	Assumptions acceptable	Variance of model residuals is constant

Linear regression (Results of multiple linear regression (City fuel ~ Engine Displacement + Transmission + Gears))

	Values	Explanation
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 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

Conclusion : Linear regression results show the model coefficients are significant and the impact of Engine displacement , gears and transmission together on city fuel is not by chance.
However the plots and gvlma package shows that not all assumptions are met. Hence this model is better but still not a good fit to the dataset.

Residual vs Fitted plot shows some non linearity and normal curve is not met.



Assumption	Satisfied or not	Explanation
Global Stat	Assumptions not satisfied	Linearity not met
Skewness	Assumptions not satisfied	Not normal distribution
kurtosis	Assumptions not satisfied	Distribution is kurtotic, tail heavy dataset
Link function	Assumptions acceptable	Dependent variables are all continuous
Heteroscedasticity	Assumptions acceptable	Variance of model residuals is constant

Summary

Exploratory Analysis

- 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.

Machine Learning

- Linear regression model was chosen to show the impact of selected features on City Fuel. The features includes Cylinders, Engine displacement, transmission type. 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 that do not make assumptions about the distribution of the data are useful. The nonparametric correlations do not assume that the data are linearly related. Kendall-Theil-Siegel regression and/or Quantile regression is a very flexible approach, and can be used for a linear regression. 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. 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.

Acknowledgements

- Mentor Branko Kovac
  - For his valuable guidance and constant motivation that gave me confidence in completing the course
  - He is very knowledgeable and our calls were very informative and always set me the right path

- Springboard community –
  - For all the information exchanges between students and/or community manager Guy Maskall that served as second set of guidance

- Springboard Student Guides –
  - Have provided prompt responses to all questions