Sp Jain school of global management

STATISTICAL ANALYSIS ON CAR MILEAGE

Detailed Project Report:

Made By:

Ananya Singh

BDS 2021-2024

Content:

INTRODUCTION

A car's mileage is the total number of miles (1.61 km) driven in a certain period of time. For this model - It is the number of miles a car can travel in a gallon (3.78 liters) of oil. There is a difference in the mileage of a car between highway and city driving.

For starters, city driving is harsher on your engine than highway travel. Because it must start and stop more frequently than highways (traffic jams, lights, pedestrians, etc.)On the other hand, As long as you're not driving at a speed much faster than your vehicle is built to manage, this is easier on your engine than city driving.This is gentler on your engine than city driving since your engine has to do less work to maintain the pace your car is travelling.

The goal of this project is to predict a car's city mileage based on the car's characteristics and highway mileage.

The attributes of the data set we chose are as follows:

Target variable :

MPG_City - Miler per Gallon in the city

Attributes:

1. Make
2. Model
3. Type
4. Origin
5. Drive Train
6. MSRP-(manufacturer's suggested retail price)
7. Invoice
8. Engine size
9. Cylinders
10. Horse Power
11. MPG_Highway
12. Weight
13. Wheelbase
14. Length

NOTE: R Studio was used to build this project, and all clips are from R.

EXPLORATORY DATA ANALYSIS

Sample data:

```
> head(carsdf,10)
   Make                  Model   Type Origin DriveTrain    MSRP Invoice EngineSize Cylinders Horsepower MPG_City MPG_Highway Weight Wheelbase Length
1  Acura                    MDX    SUV   Asia        All $36,945 $33,337        3.5         6        265       17          23   4451       106    189
2  Acura          RSX Type S 2dr  Sedan   Asia      Front $23,820 $21,761        2.0         4        200       24          31   2778       101    172
3  Acura                TSX 4dr  Sedan   Asia      Front $26,990 $24,647        2.4         4        200       22          29   3230       105    183
4  Acura                 TL 4dr  Sedan   Asia      Front $33,195 $30,299        3.2         6        270       20          28   3575       108    186
5  Acura              3.5 RL 4dr  Sedan   Asia      Front $43,755 $39,014        3.5         6        225       18          24   3880       115    197
6  Acura 3.5 RL w/Navigation 4dr  Sedan   Asia      Front $46,100 $41,100        3.5         6        225       18          24   3893       115    197
7  Acura  NSX coupe 2dr manual S Sports   Asia       Rear $89,765 $79,978        3.2         6        290       17          24   3153       100    174
8   Audi             A4 1.8T 4dr  Sedan Europe      Front $25,940 $23,508        1.8         4        170       22          31   3252       104    179
9   Audi  A41.8T convertible 2dr  Sedan Europe      Front $35,940 $32,506        1.8         4        170       23          30   3638       105    180
10  Audi             A4 3.0 4dr  Sedan Europe      Front $31,840 $28,846        3.0         6        220       20          28   3462       104    179
```

Summarization of the data:

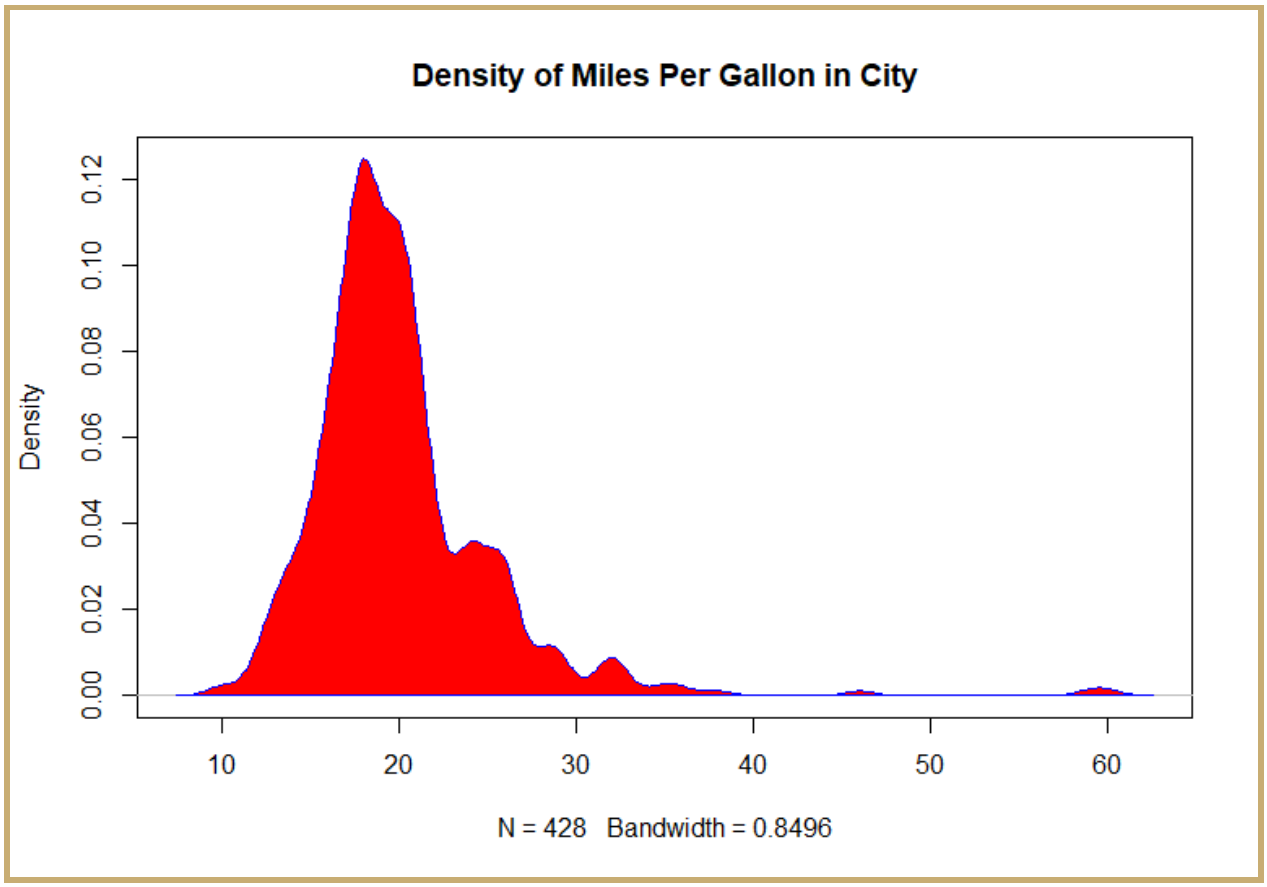
```
     Make               Model              Type              Origin           DriveTrain       
 Length:428         Length:428         Length:428         Length:428         Length:428        
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  


     MSRP              Invoice          EngineSize      Cylinders        Horsepower       MPG_City    
 Length:428         Length:428         Min.   :1.300   Min.   : 3.000   Min.   : 73.0   Min.   :10.00  
 Class :character   Class :character   1st Qu.:2.375   1st Qu.: 4.000   1st Qu.:165.0   1st Qu.:17.00  
 Mode  :character   Mode  :character   Median :3.000   Median : 6.000   Median :210.0   Median :19.00  
                                       Mean   :3.197   Mean   : 5.799   Mean   :215.9   Mean   :20.06  
                                       3rd Qu.:3.900   3rd Qu.: 6.000   3rd Qu.:255.0   3rd Qu.:21.25  
                                       Max.   :8.300   Max.   :12.000   Max.   :500.0   Max.   :60.00  
  MPG_Highway        Weight       Wheelbase          Length     
 Min.   :12.00   Min.   :1850   Min.   : 89.0   Min.   :143.0  
 1st Qu.:24.00   1st Qu.:3104   1st Qu.:103.0   1st Qu.:178.0  
 Median :26.00   Median :3474   Median :107.0   Median :187.0  
 Mean   :26.84   Mean   :3578   Mean   :108.2   Mean   :186.4  
 3rd Qu.:29.00   3rd Qu.:3978   3rd Qu.:112.0   3rd Qu.:194.0  
 Max.   :66.00   Max.   :7190   Max.   :144.0   Max.   :238.0  
```

This is a summary of all of our dataset's columns. The mean and median for all attributes are fairly close, indicating that there are very few outliers that need not be deleted.

There are also no null values.

```
> sum(is.na(carsdf))
[1] 0
```

Density of MPG in city

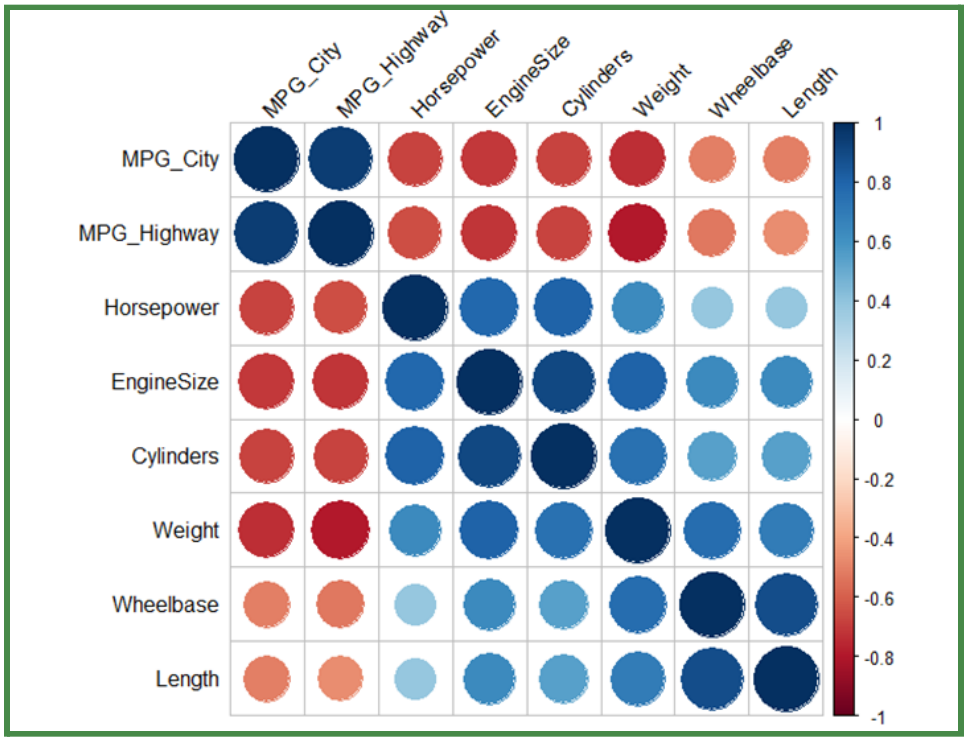


To help us better grasp each column, we can see more information and sample rows in the graphic below.

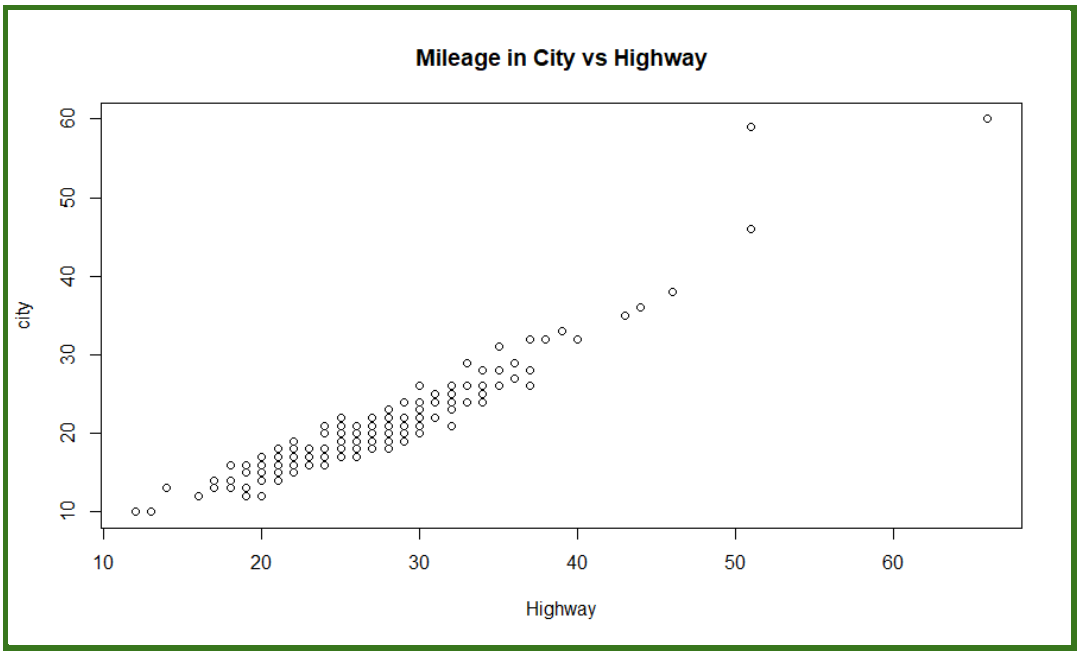
```
> str(carsdf)
'data.frame':   428 obs. of  15 variables:
 $ Make        : chr  "Acura" "Acura" "Acura" "Acura" ...
 $ Model       : chr  "MDX" "RSX Type S 2dr" "TSX 4dr" "TL 4dr" ...
 $ Type        : chr  "SUV" "Sedan" "Sedan" "Sedan" ...
 $ Origin      : chr  "Asia" "Asia" "Asia" "Asia" ...
 $ DriveTrain : chr  "All" "Front" "Front" "Front" ...
 $ MSRP        : chr  "$36,945" "$23,820" "$26,990" "$33,195" ...
 $ Invoice     : chr  "$33,337" "$21,761" "$24,647" "$30,299" ...
 $ EngineSize : num  3.5 2 2.4 3.2 3.5 3.5 3.2 1.8 1.8 3 ...
 $ Cylinders   : int  6 4 4 6 6 6 6 4 4 6 ...
 $ Horsepower : int  265 200 200 270 225 225 290 170 170 220 ...
 $ MPG_City    : int  17 24 22 20 18 18 17 22 23 20 ...
 $ MPG_Highway: int  23 31 29 28 24 24 24 31 30 28 ...
 $ Weight      : int  4451 2778 3230 3575 3880 3893 3153 3252 3638 3462 ...
 $ Wheelbase   : int  106 101 105 108 115 115 100 104 105 104 ...
 $ Length      : int  189 172 183 186 197 197 174 179 180 179 ...
```

CORRELATION MATRIX

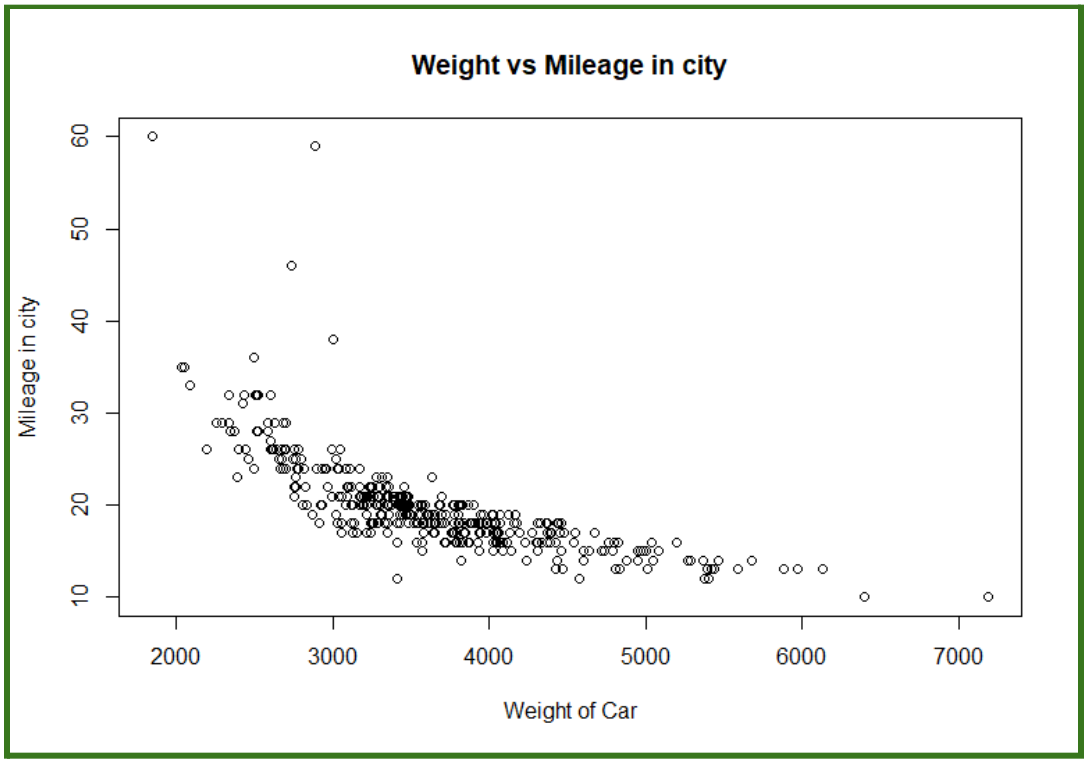
As seen in the above clip, there are seven columns with character data, one with number data, and seven with integer data. The model will not use all of the columns since they do not effect the mileage in the same way or by a substantial amount; this may be verified using a graphical depiction of the - ***correlation matrix.***



Positive correlations are shown in blue, while negative correlations are shown in red. The correlation coefficients are proportional to the colour intensity and circle size. As can be seen, MPG_City and MPG Highway and MPG_City and Weight have close relationships.



Scatterplot showing positive relationship between mileage on highway and city



The above graph shows the negative relationship between weight and mileage.

SPLITTING THE DATA

```
> set.seed(310)
>
> C_index <- sample(x = nrow(carsdf), size = nrow(carsdf)*0.70)
>
>
> C_train <- carsdf[C_index,]
> C_test <- carsdf[-C_index,]
>
```

We've set the seed to ensure that the model produces the same results every time.

The training data makes up 70% of the total data, while the test data makes up the remaining 30%.

MODEL BUILDING

We can now use the information from the EDA and our training data set to create a model that is appropriate for our test dataset.

A multiple linear regression model will be used because there are several explanatory factors.

Initially lets set our **X variables** as:

1. Type
2. DriveTrain
3. EngineSize
4. Cylinder
5. Horsepower
6. Weight
7. Wheelbase
8. Length

Columns removed are **Make, Model , MSRP, Origin, MPG_Highway and invoice.**

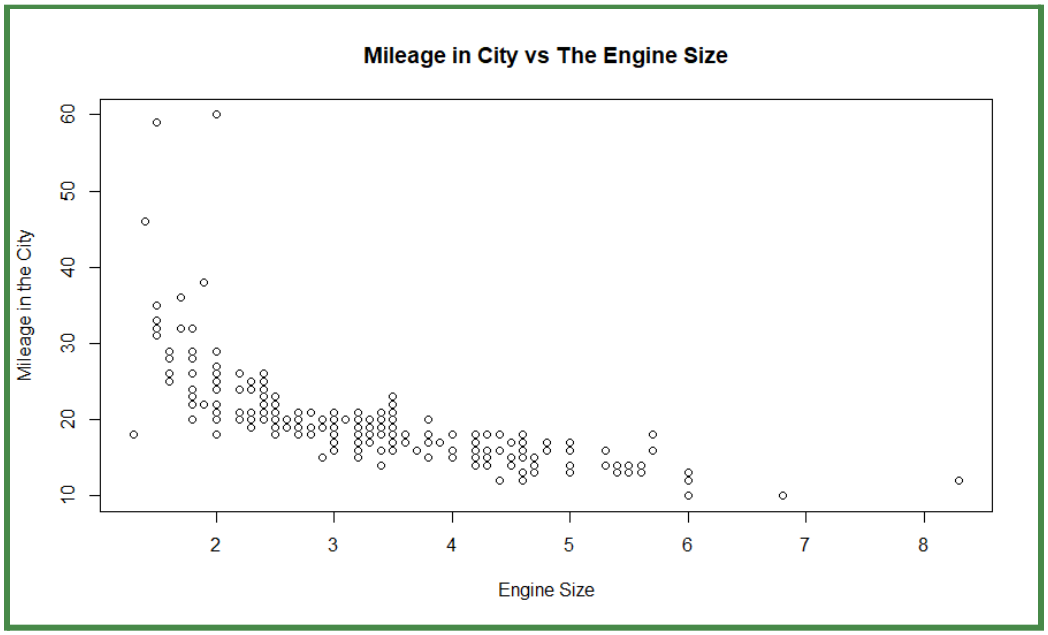
The make, model, and country of origin were eliminated because the car's details were judged to be more important. MSRP was removed because retail price has no bearing on mileage, and invoice was also removed for the same reason. If we include MPG_Highway, the accuracy of the model will increase. However, it won't be an efficient model because the values are close to each other(because of which we get strong correlation) but do not effect each other.

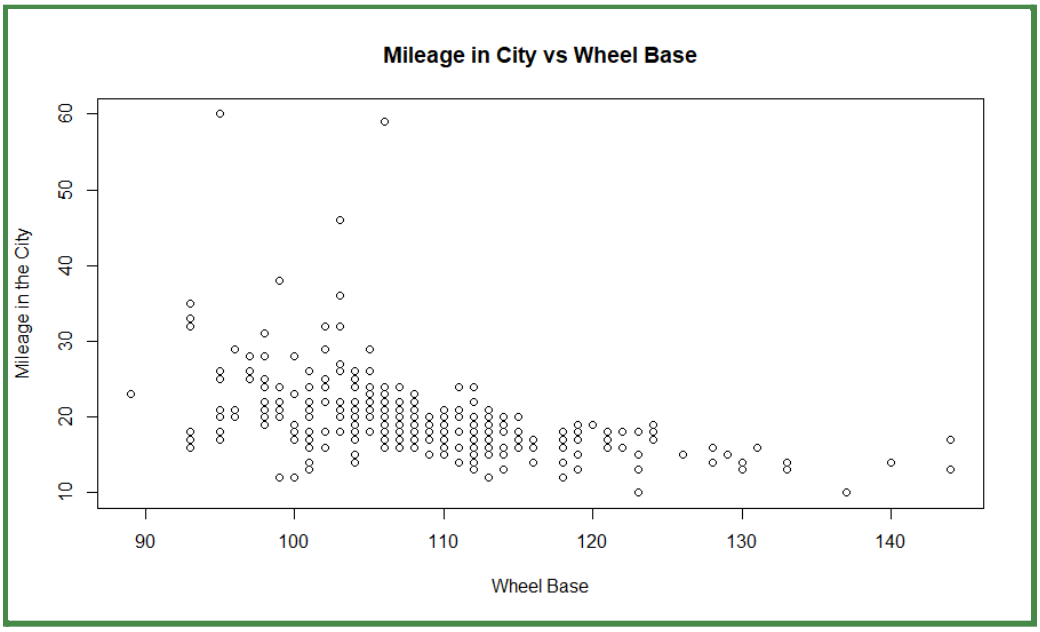
```
Call:
lm(formula = MPG_City ~ Type + DriveTrain + EngineSize + Cylinders +
    Horsepower + Weight + Wheelbase + Length, data = carsdf)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0101 -1.3593 -0.1001  0.9396 13.4296

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     65.538559   2.531074  25.894  < 2e-16 ***
TypeSedan      -28.437684   1.239146 -22.949  < 2e-16 ***
TypeSports     -29.779222   1.319592 -22.567  < 2e-16 ***
TypeSUV        -29.426102   1.297800 -22.674  < 2e-16 ***
TypeTruck      -29.152848   1.351282 -21.574  < 2e-16 ***
TypeWagon      -28.348305   1.293641 -21.914  < 2e-16 ***
DriveTrainFront  1.098024   0.324153   3.387 0.000773 ***
DriveTrainRear   0.049160   0.368932   0.133 0.894061
EngineSize      -0.078390   0.274653  -0.285 0.775470
Cylinders       -0.215163   0.178678  -1.204 0.229204
Horsepower      -0.013484   0.002917  -4.622 5.08e-06 ***
Weight          -0.002888   0.000370  -7.805 4.91e-14 ***
Wheelbase        0.057748   0.033688   1.714 0.087240 .
Length          -0.048349   0.017226  -2.807 0.005241 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.101 on 414 degrees of freedom
Multiple R-squared:  0.844,     Adjusted R-squared:  0.8391
F-statistic: 172.3 on 13 and 414 DF,  p-value: < 2.2e-16
```

With a R squared value of 84 percent and a P value less than 0.05, the first model is a reasonably good fit. Not all columns, however, have a significant P-value.Namely - DriveTrainRear(a subset of DriveTrain column), EngineSize, Cylinders, Wheelbase.

From the above mentioned attributes, DriveTrainRear cannot be removed as DriveTrainFront does have a significant P value. additionally , EngineSize and Wheelbase had a high correlation with MPG_City.

Using scatterplots to visualise this relationship -


Mileage in City vs The Engine Size
Mileage in the City
10
20
30
40
50
60
2
3
4
5
6
7
8
Engine Size


Mileage in City vs Wheel Base
Mileage in the City
10
20
30
40
50
60
90
100
110
120
130
140
Wheel Base

After building several different models , we get the following values for adjusted R-Squared and Residual Standard Error.

Model	1	2	3	4	5	6
R squared	83.9	83.3	83.8	83.6	83.8	83.7
Residual Standard Error	2.101	2.13	2.106	2.11	2.104	2.11

After running comparative analysis on all the models,

```
> compare_performance(lm1,lm2,lm3,lm4,lm5,lm6,rank=TRUE)
# Comparison of Model Performance Indices

Name | Model |    R2 | R2 (adj.) |  RMSE | Sigma | AIC weights | BIC weights | Performance-Score
-----------------------------------------------------------------------------------------------------
lm5  |    lm | 0.843 |     0.839 | 2.074 | 2.104 |       0.401 |       0.249 |            80.98%
lm1  |    lm | 0.844 |     0.839 | 2.066 | 2.101 |       0.259 |       0.003 |            77.49%
lm6  |    lm | 0.841 |     0.838 | 2.084 | 2.111 |       0.151 |       0.715 |            67.74%
lm3  |    lm | 0.843 |     0.838 | 2.074 | 2.106 |       0.155 |       0.013 |            63.44%
lm4  |    lm | 0.841 |     0.837 | 2.086 | 2.116 |       0.034 |       0.021 |            40.42%
lm2  |    lm | 0.837 |     0.833 | 2.109 | 2.139 |     < 0.001 |     < 0.001 |             0.00%
```

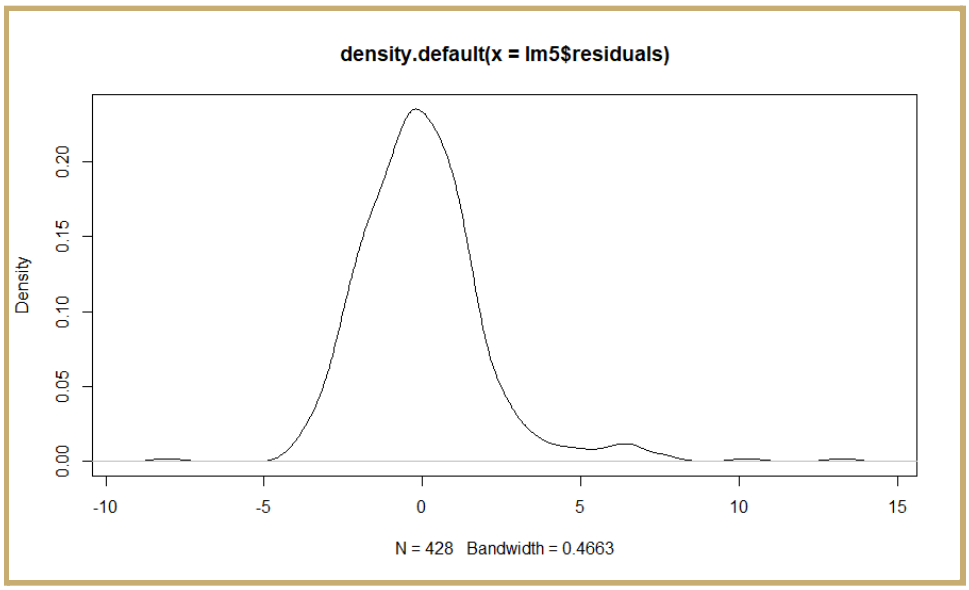
Model number 5 has the **best fit**,

- Firstly it has a good adjusted R squared value paired with Residual standard error
- Secondly, all the columns have significant p values except one column(Drive Train Rear).

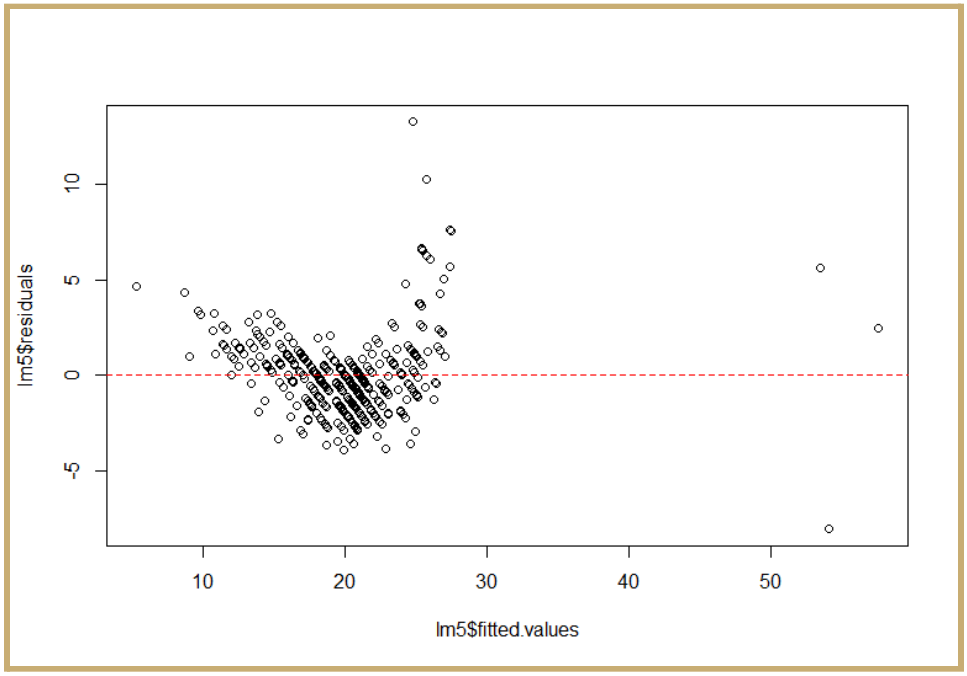
```
Residuals:
    Min      1Q  Median      3Q     Max
-8.0549 -1.3468 -0.1661  0.9858 13.2624

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.788e+01  2.089e+00  32.490  < 2e-16 ***
TypeSedan       -2.860e+01  1.236e+00 -23.133  < 2e-16 ***
TypeSports      -3.016e+01  1.303e+00 -23.156  < 2e-16 ***
TypeSUV         -2.967e+01  1.292e+00 -22.972  < 2e-16 ***
TypeTruck       -2.908e+01  1.349e+00 -21.550  < 2e-16 ***
TypeWagon       -2.856e+01  1.288e+00 -22.172  < 2e-16 ***
Cylinders       -2.701e-01  1.375e-01  -1.964 0.050250 .
Horsepower      -1.371e-02  2.850e-03  -4.809 2.12e-06 ***
Weight          -2.633e-03  3.308e-04  -7.961 1.65e-14 ***
DriveTrainFront  1.134e+00  3.237e-01   3.503 0.000509 ***
DriveTrainRear   1.681e-01  3.629e-01   0.463 0.643575
Length          -3.089e-02  1.301e-02  -2.374 0.018057 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.104 on 416 degrees of freedom
Multiple R-squared:  0.8429,    Adjusted R-squared:  0.8387
F-statistic: 202.9 on 11 and 416 DF,  p-value: < 2.2e-16
```

PREDICTING MODEL & RESIDUAL PLOTS

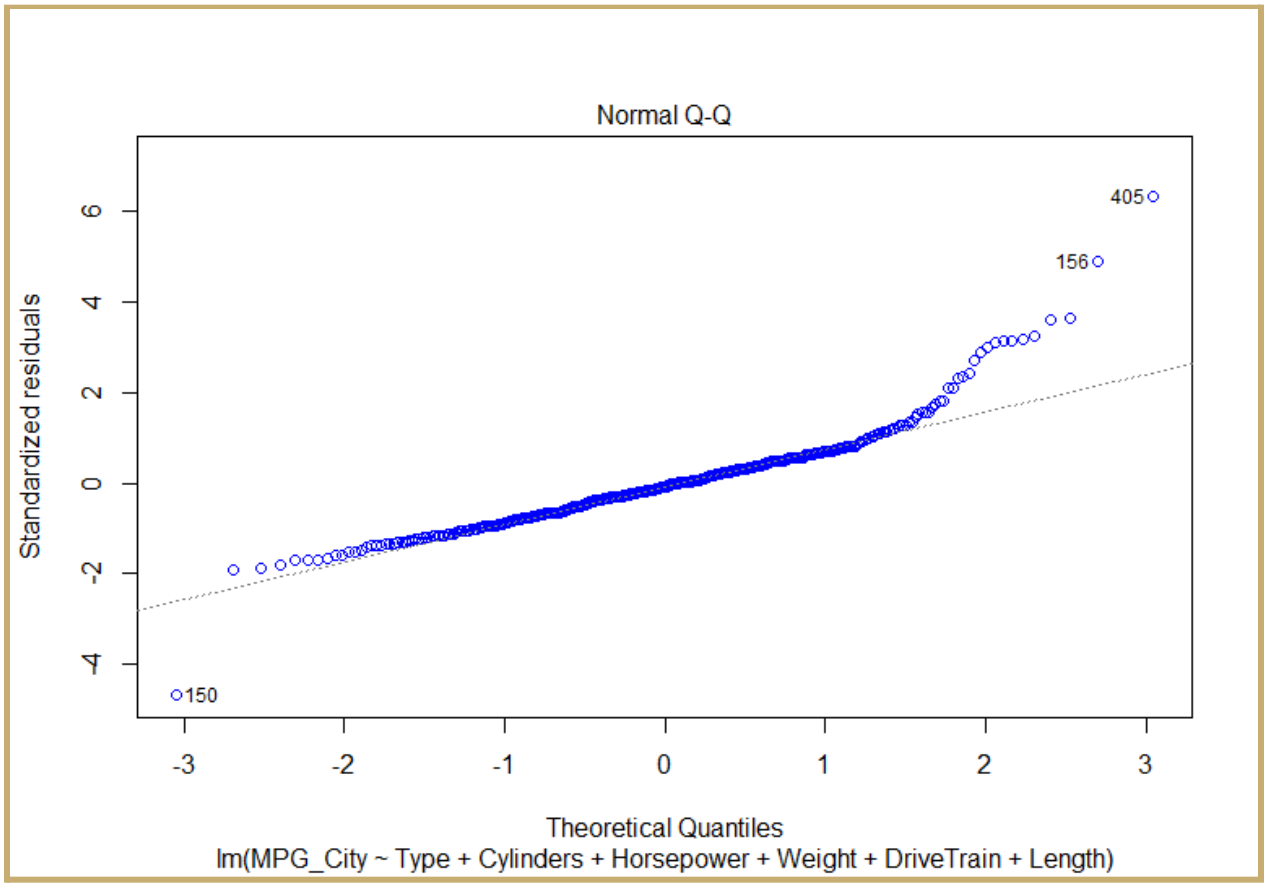
After predicting values using model 5, we use a residuals plot to analyse our model.



The residuals follow a normal distrubution



As can be seen from above plot, the residuals do not follow any pattern , i.e. a multiple linear regression was a correct choice. This can also be confirmed with a QQ plot.



The Q-Q plot reveals that the residuals follow a nearly straight line, indicating that they are normally distributed. There are some outliers that stick out among the residuals.As a whole, a good alignment of residuals are observed around the line of best fit.

HYPOTHESIS TESTING

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

$$ag. \quad H_1 : \beta_j \neq 0, \; for \; atleast \; one \; j.$$

```
> anova(lm5)
Analysis of Variance Table

Response: MPG_City
            Df Sum Sq Mean Sq  F value    Pr(>F)
Type         5 5348.3  1069.7 241.6945 < 2.2e-16 ***
Cylinders    1 3456.6  3456.6 781.0358 < 2.2e-16 ***
Horsepower   1  376.2   376.2  85.0136 < 2.2e-16 ***
Weight       1  613.3   613.3 138.5826 < 2.2e-16 ***
DriveTrain   2   55.9    27.9   6.3114  0.001994 **
Length       1   24.9    24.9   5.6352  0.018057 *
Residuals  416 1841.1     4.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be observed from the anova table all p values are significant therefore , null hypothesis is rejected and the model can be used for prediction.

RESULTS

MODEL INFO:
Observations: 428
Dependent Variable: MPG_City
Type: OLS linear regression

MODEL FIT:
$F(11,416) = 202.85$, $p = 0.00$
$R^2 = 0.84$
Adj. $R^2 = 0.84$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	67.88	2.09	32.49	0.00
TypeSedan	-28.60	1.24	-23.13	0.00
TypeSports	-30.16	1.30	-23.16	0.00
TypeSUV	-29.67	1.29	-22.97	0.00
TypeTruck	-29.08	1.35	-21.55	0.00
TypeWagon	-28.56	1.29	-22.17	0.00
Cylinders	-0.27	0.14	-1.96	0.05
Horsepower	-0.01	0.00	-4.81	0.00
Weight	-0.00	0.00	-7.96	0.00
DriveTrainFront	1.13	0.32	3.50	0.00
DriveTrainRear	0.17	0.36	0.46	0.64
Length	-0.03	0.01	-2.37	0.02

This model is statistically significant with significant p - values and f- statistic. This model has **84% accuracy** to predict the mileage of cars in the city.

ACKNOWLEDGEMENTS

I would like to express gratitude to my professor Suchismita ma'am and our institute - SP jain school of Global management for providing me the opportunity, guidance and resources to make this project possible.

REFERENCES

1. https://www.lexingtontoyota.com/blog/gas-mileage-different-city-highway/
2. Data- https://drive.google.com/file/d/1dCW_f496NXEfKx6sFdquxfDc8eZU0jKb/view?usp=sharing