Name: Arbaz Khan
Student ID: 23100243
Instructor: Shaheena Bashir

Research Question

How is a car's price affected by company, year, horsepower, number of cylinders, number of doors, miles per gallon on the highway and miles per gallon in the city?

Abstract

People are buying vehicles every day, and the price of these changes every. So, we need to have some model to detect the cost of the vehicles. In this, I made a car pricing model which will predict the car price based on the company, year, horsepower, number of cylinders, number of doors, miles per gallon on the highway and miles per gallon in the city. I know there are many other factors affecting the prices of cars, but I am using these seven factors because they are the most important thing a buyer sees before buying a vehicle. Every buyer wants a good combination of these factors. So, we can estimate the price using these factors.

Introduction

My dataset is from Kaggle and represents the data of cars sold in the USA between 1990 and 2018. It has 11914 instances of data and 16 attributes. I have selected the company, year, horsepower, number of cylinders, number of doors, miles per gallon on the highway, miles per gallon in the city, and the car's price for my model. I removed the other attributes from the dataset as they were not as crucial as these are. The cost of the car is my response variable, and year, horsepower, number of cylinders, number of doors, miles per gallon on the highway and miles per gallon in the city are the predictors for my model.

1) getwd()
2) setwd('C:/Users/2018n/Desktop/Statistics (MATH 231)/Arbaz Khan/Project')

3) cars_data <- read.csv("data.csv", header=TRUE, stringsAsFactors = FALSE)
4) class(cars_data)
5) head(cars_data)
6) tail(cars_data)
7) cars_subdata <- subset(cars_data, select = c("Year", "Engine.HP", "Engine.Cylinders", "Number.of.Doors", "highway.MPG", "city.mpg", "Popularity", "MSRP"))
8) head(cars_subdata)
9) tail(cars_subdata)

Data Cleaning

After getting the sub-dataset for my model, I renamed the columns, so we have meaningful names and know which column represents which feature. In my dataset, the company is a categorical variable with 48 categories. I select the top five occurring categories to understand and interpret my model quickly; otherwise, using all 48 types will result in a complex model, and it won't be easy to understand and interpret that mode. After selecting the top five categories, we have 4185 instances of the data in our dataset. I counted the NAN values in each column and found that the Horsepower has 7 NAN values, and the total NAN values in our data are 28. I removed the rows with NAN values because Horsepower, the Number of cylinders and number of doors are essential features to predict the price, and NAN values in these columns can cause problems in our model. Then I changed the index of row so we can have indexes without any gaps between. There were 4185 instances of the data before removing the NAN values and 4157 after removing the NAN values.

1) colnames(cars_subdata)[2] <- 'HorsePower'
2) colnames(cars_subdata)[3] <- 'Cylinders'
3) colnames(cars_subdata)[4] <- 'Doors'
4) colnames(cars_subdata)[5] <- 'MPGHighway'
5) colnames(cars_subdata)[6] <- 'MPGCity'
6) colnames(cars_subdata)[8] <- 'Price'

```
7) head(cars_subdata)
8) tail(cars_subdata)
9) cars_subdata <- subset(cars_subdata, Company == 'Chevrolet' | Company == 'Ford' |
   Company == 'Volkswagen' | Company == 'Toyota' | Company == 'Dodge')
10) rownames(cars_subdata) <- NULL
11) head(cars_subdata)
12) tail(cars_subdata)
13) sum(is.na(cars_subdata$Year))
14) sum(is.na(cars_subdata$Horsepower))
15) sum(is.na(cars_subdata$Cylinders))
16) sum(is.na(cars_subdata$Doors))
17) sum(is.na(cars_subdata$MPGHighway))
18) sum(is.na(cars_subdata$MPGCity))
19) sum(is.na(cars_subdata$Popularity))
20) sum(is.na(cars_subdata$Price))
21) sum(is.na(cars_subdata))
22) nrow(cars_subdata)
23) cars_subdata <- na.omit(cars_subdata)
24) nrow(cars_subdata)
25) rownames(cars_subdata) <- NULL
26) head(cars_subdata)
27) tail(cars_subdata)
```

Data Exploration and Visual Analysis

I have plotted the summaries and boxplots for all the variables. We can see all the potential outliers, first and 3rd quartiles for the dataset for each column.

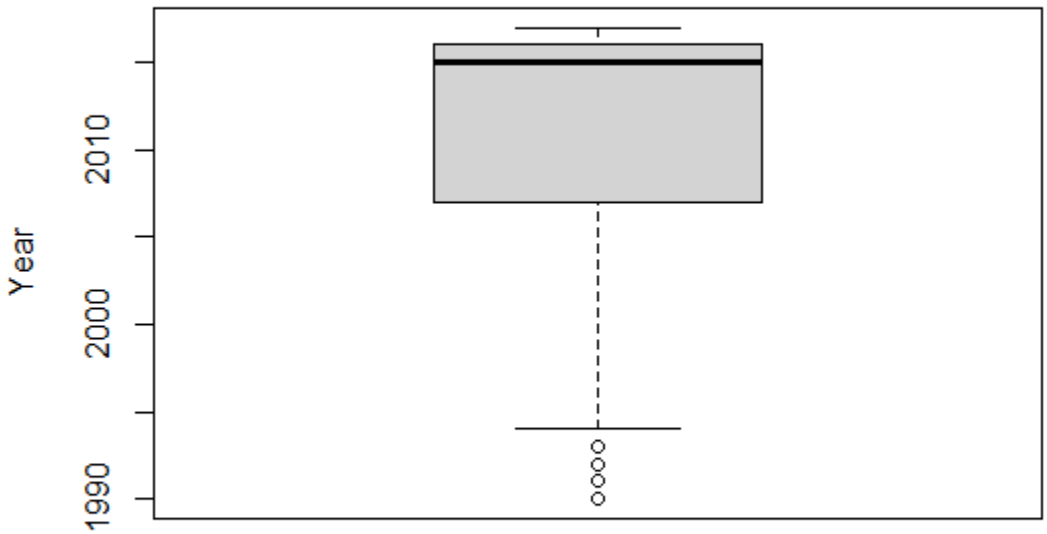
i) Year

We can see from the box that most of the data belongs from 2007 to 2018. We can also see that the median of the year is 2015, which means 50% of the information is collected before 2015 and 50% after 2015. We can see only four observation lies from 1990 to 1994. These are the potential outliers, but we cannot remove them because these can provide essential information in our prediction of the cars price, which was made in the 1990s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1990	2007	2015	2010	2016	2017

IQR = 2016 – 2007 = 9

Lower Limit = 2007 – 1.5 * 9 = 1993.5 ≈ 1994



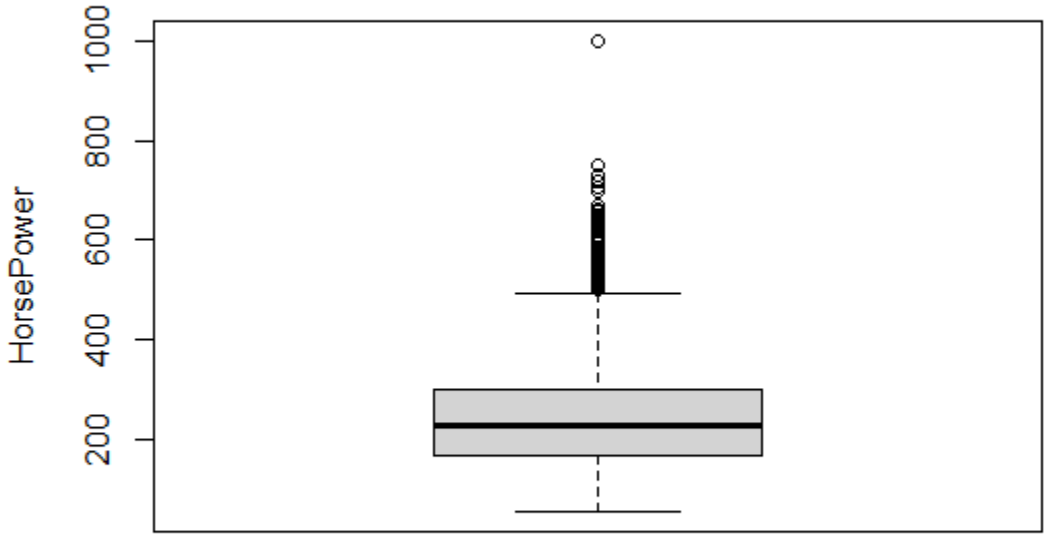
ii) Horsepower

We can see from the box pot that many observations are above the upper limit for potential outliers.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
55.0	170.0	227.0	249.5	300.0	1001.0

IQR = 300.0 – 170.0 = 130

Upper Limit = 300 + 1.5 * 130 = 495



iii) Cylinders

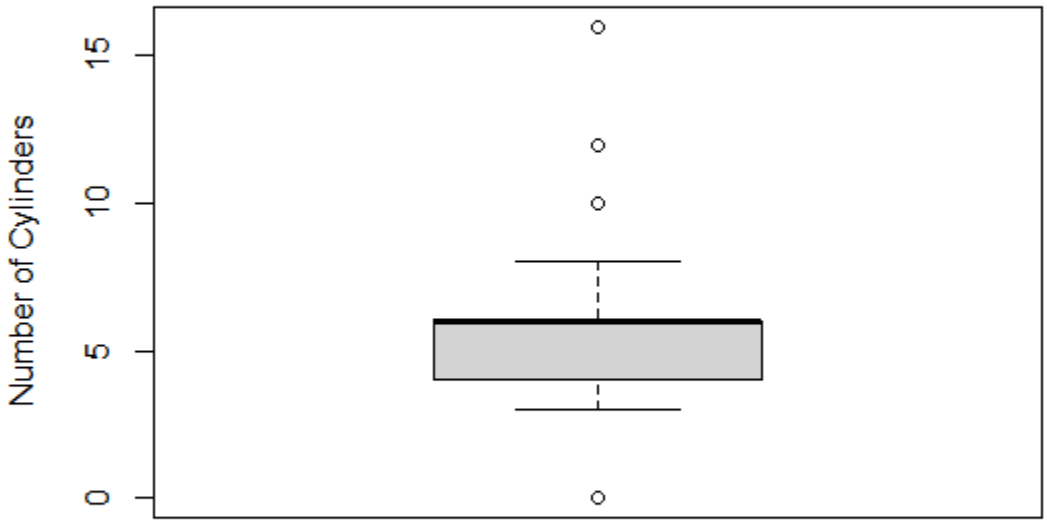
We can see from the box plot there are few potential outliers above the upper limit and few observation below the lower limit.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	4.00	6.00	5.65	6.00	16.00

IQR = 6 – 4 = 2

Lower Limit = 4 – 1.5 * 2 = 7

Upper Limit = 6 + 1.5 * 2 = 9

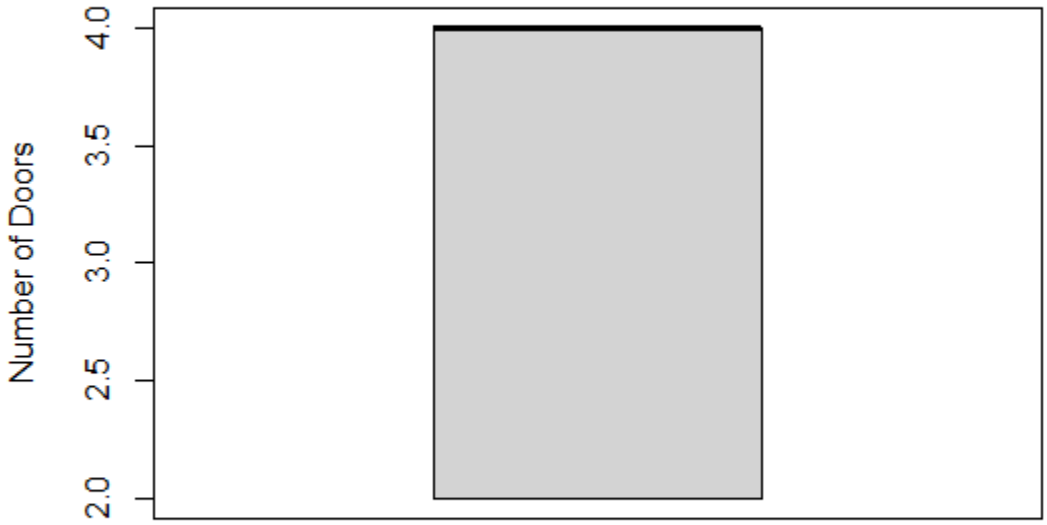


iv) Doors

We can see the median of the data is 4, so we say at least 50% of the cars in our dataset have four doors. We can see that 1st quartile is 2, and 2 is a minimum value

for the number of doors, so we can say that at least 25% of the cars in our data set have 2 doors.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	2.000	4.000	3.433	4.000	4.000



v) Miles per gallon on Highway

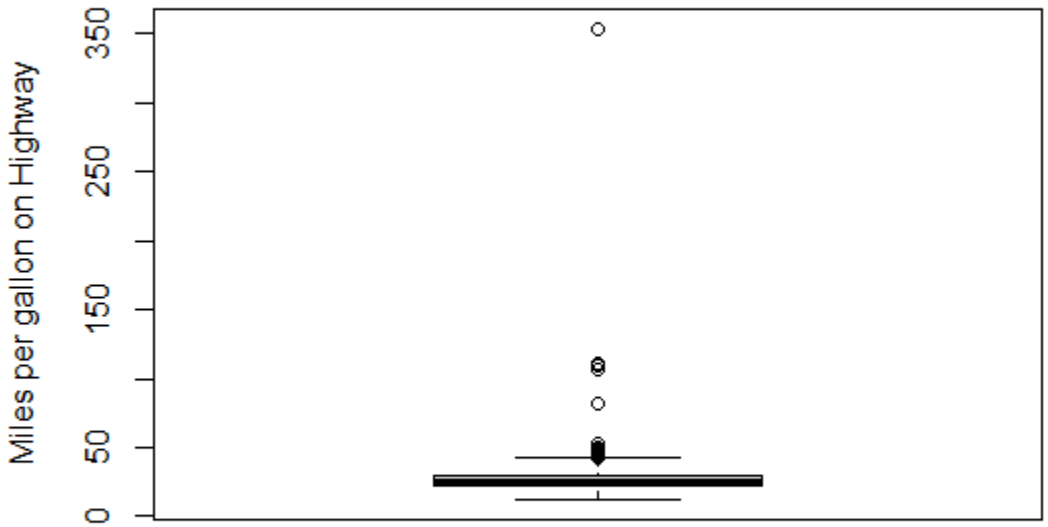
Few observations are very far from the other, so this could be an outlier. There are some observations which are also the upper limit and potential outliers.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.00	22.00	26.00	26.32	30.00	354.00

IQR = 30 – 22 = 8

Lower Limit = 22 – 1.5 * 8 = 10

Upper Limit = 30 + 1.5 * 8 = 42



vi) Miles per gallon in City

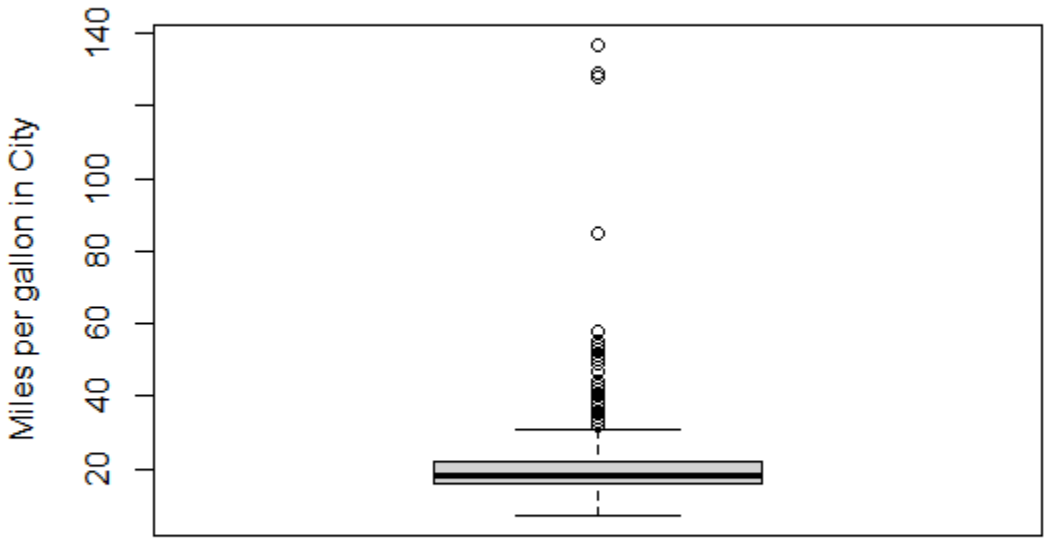
Some observations are very far from other observations and above the upper limit.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.00	16.00	18.00	19.33	22.00	137.00

IQR = 22.00 – 16.00 = 6

Upper Limit = 22.00 + 1.5 * 6 = 31

Lower Limit = 16.00 - 1.5 * 6 = 7



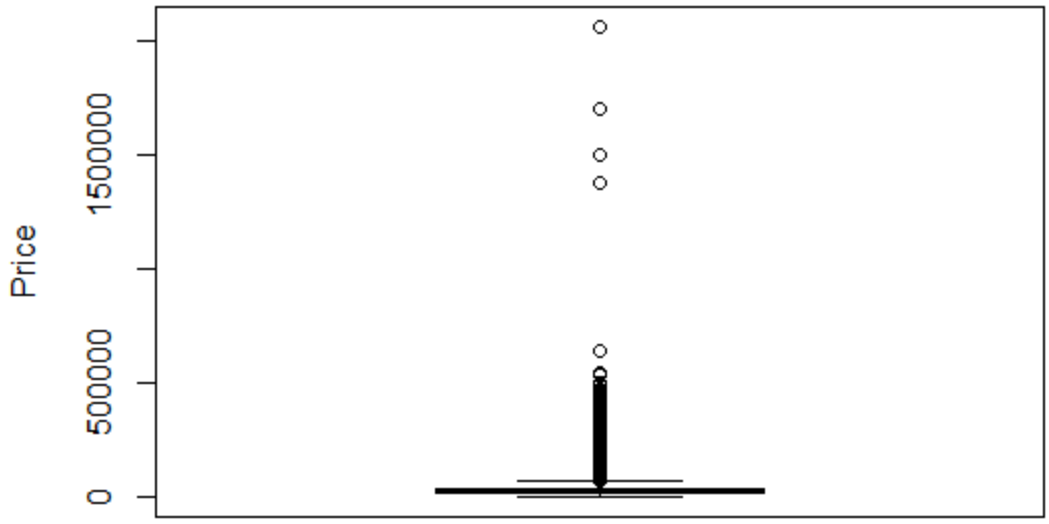
vii) Price

There are some observations which are very far from other observations and also above from the upper.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2000	20990	29960	40554	42200	2065902

IQR = 42200 – 20990 = 21210

Upper Limit = 42200 + 1.5 * 21210 = 74015



1) summary(cars_subdata$Year)
2) boxplot(cars_subdata$Year, ylab='Year')
3) summary(cars_subdata$HorsePower)
4) boxplot(cars_subdata$HorsePower, ylab='HorsePower')
5) summary(cars_subdata$Cylinders)
6) boxplot(cars_subdata$Cylinders, ylab='Cylinders')
7) summary(cars_subdata$Doors)
8) boxplot(cars_subdata$Doors, ylab='Doors')
9) summary(cars_subdata$MPGHighway)
10) boxplot(cars_subdata$MPGHighway, ylab='MPGHighway')
11) summary(cars_subdata$MPGCity)
12) boxplot(cars_subdata$MPGCity, ylab='MPGCity')
13) summary(cars_subdata$Popularity)
14) boxplot(cars_subdata$Price, ylab='Price')

Method

My response variable has continuous values so, I used multiple linear regression to predict the price, which is the dependent variable for my model. I used six independent variables to predict the value of the car price in my model, which are year, horsepower, number of cylinders in the car, number of doors of the car, miles per gallon on the highway, and miles per gallon in city and popularity of the car. Following is the general form for my model.

$$h(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_6 x_6 + \beta_7 x_7 + \varepsilon \text{ where } \varepsilon \sim N(0, 1)$$

In this, h(x) is our repose variable, and x1, x2, x3, x4, x5, x6, and x7 are the independent variables of my model, which represent the company, year, horsepower, number of cylinders in the car, number of doors of the car, miles per gallon on the highway, and miles per gallon in city and β1, β2, β3, β4, β5, β6 and β7 are their weights respectively. These weights determine the contribution of their respective independent variables to predict the value of the price. ε is the random error of data.

Results and Analysis

Call:

lm(formula = Price ~ Company + Year + HorsePower + Cylinders +

Doors + MPGHighway + MPGCity, data = cars_subdata)

Residuals:

Min	Q1	Median	Q3	Max
-26062	-4343	-211	3831	93523

Coefficients:

Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	-1.484e+06	4.767e+04	-31.130	< 2e-16 ***
CompanyDodge	-1.829e+03	3.796e+02	-4.817	1.51e-06 ***
CompanyFord	-2.899e+02	3.393e+02	-0.854	0.39290
CompanyToyota	-9.800e+02	3.667e+02	-2.672	0.00757 **
CompanyVolkswagen	4.154e+03	3.742e+02	11.100	< 2e-16 ***
Year	7.377e+02	2.395e+01	30.805	< 2e-16 ***
HorsePower	1.158e+02	2.425e+00	47.732	< 2e-16 ***
Cylinders	-4.018e+02	1.559e+02	-2.577	0.00999 **
Doors	1.096e+03	1.458e+02	7.513	7.02e-14 ***
MPGHighway	-2.830e+02	4.765e+01	-5.939	3.10e-09 ***
MPGCity	3.506e+02	4.251e+01	8.248	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7414 on 4146 degrees of freedom

Multiple R-squared: 0.7827,   Adjusted R-squared: 0.7821

F-statistic: 1493 on 10 and 4146 DF, p-value: < 2.2e-16

Anova Table

Anova Table (Type II tests)

Response: Price

	Sum Sq	Df	F value	Pr(>F)
Company	1.2214e+10	4	55.5557	< 2.2e-16 ***
Year	5.2158e+10	1	948.9371	< 2.2e-16 ***
HorsePower	1.2523e+11	1	2278.3342	< 2.2e-16 ***
Cylinders	3.6512e+08	1	6.6428	0.00999 **
Doors	3.1028e+09	1	56.4511	7.017e-14 ***
MPGHighway	1.9389e+09	1	35.2750	3.098e-09 ***
MPGCity	3.7390e+09	1	68.0256	< 2.2e-16 ***
Residuals	2.2788e+11	4146		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

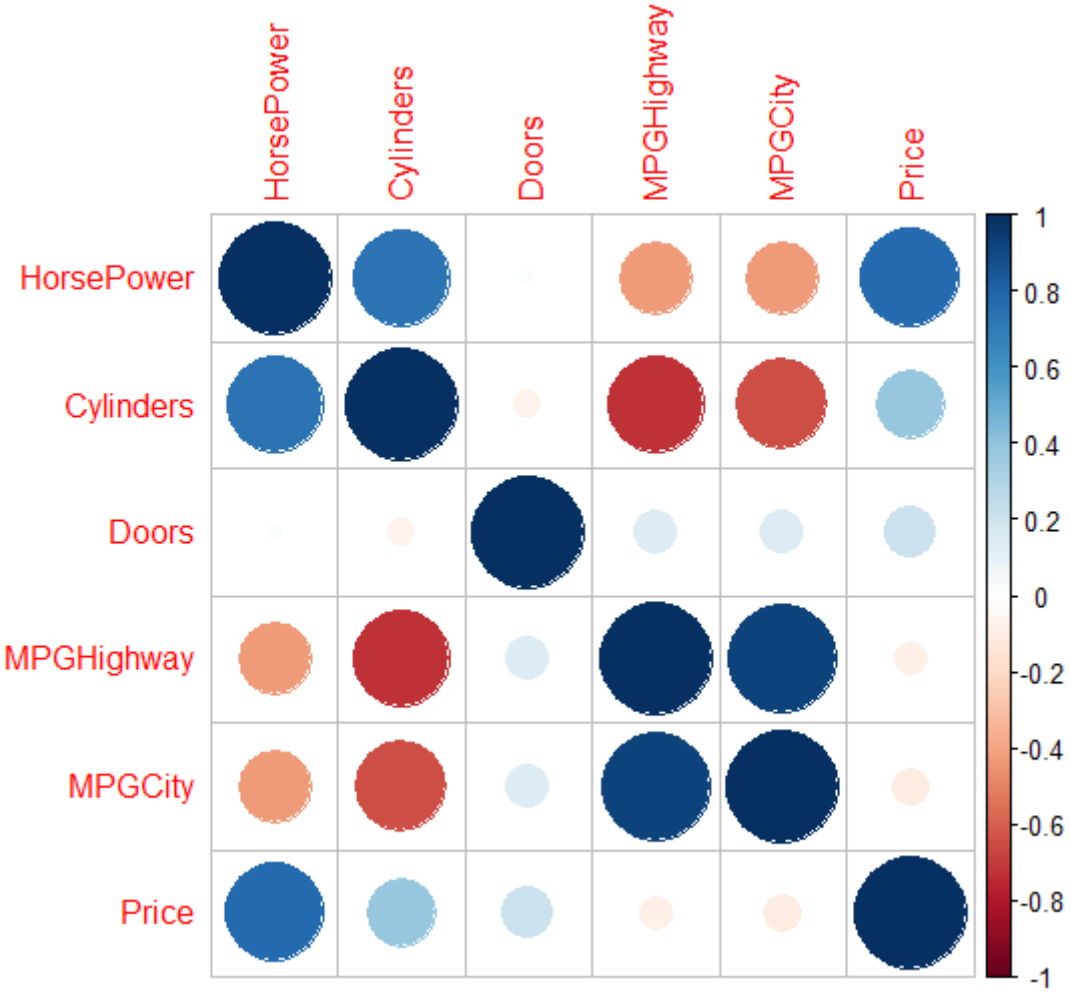
VIFG

Since most of the values of VIFG are greater than 1 and lesser than 5, we can conclude that there definitely is a moderate correlation amongst our variables, but it is not severe enough to warrant corrective measures in our regression model. The VIFG of MPGHighway and MPGCity are greater than 5 and therefore our results maybe a little biased or inaccurate due to it.

	GVIF	Df	GVIF^(1/(2*Df))
Company	1.591340	4	1.059791
Year	2.596138	1	1.611254
HorsePower	4.130172	1	2.032282
Cylinders	4.656383	1	2.157865
Doors	1.293739	1	1.137427
MPGHighway	10.368952	1	3.220086
MPGCity	7.892216	1	2.809309

Correlation Matrix

As we can see from the matrix that the Miles per gallon is highly correlated with miles per gallon in the that is they have higher values of VIFG than 5. So, we can remove one other variables in the model instead of using both.



Results Using Miles per gallon on Highway instead of both

$h(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_6 x_6 + \varepsilon$ where $\varepsilon \sim N(0, 1)$

Above is our updated equation for our new model and everything is same so we now will have new estimates for this model. We can write the new equations using these updated values.

Call:

lm(formula = Price ~ Company + Year + HorsePower + Cylinders +

Doors + MPGHighway, data = cars_subdata)

Residuals:

Min	Q1	Median	Q3	Max
-25255	-4322	-317	3950	93372

Coefficients:

Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	-1.454e+06	4.791e+04	-30.342	< 2e-16 ***
CompanyDodge	-1.788e+03	3.827e+02	-4.672	3.08e-06 ***
CompanyFord	-6.819e+01	3.409e+02	-0.200	0.8415
CompanyToyota	-2.511e+02	3.588e+02	-0.700	0.4841
CompanyVolkswagen	3.724e+03	3.735e+02	9.969	< 2e-16 ***
Year	7.213e+02	2.406e+01	29.983	< 2e-16 ***
HorsePower	1.140e+02	2.435e+00	46.816	< 2e-16 ***
Cylinders	-2.570e+02	1.561e+02	-1.646	0.0998 .
Doors	1.125e+03	1.469e+02	7.658	2.33e-14 ***
MPGHighway	5.954e+01	2.354e+01	2.529	0.0115 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7473 on 4147 degrees of freedom

Multiple R-squared: 0.7791, Adjusted R-squared: 0.7786

F-statistic: 1625 on 9 and 4147 DF, p-value: < 2.2e-16

Updated Anova Table

Anova Table (Type II tests)

Response: Price

	Sum Sq	Df	F value	Pr(>F)
Company	9.7960e+09	4	43.8474	< 2.2e-16 ***
Year	5.0210e+10	1	898.9690	< 2.2e-16 ***
HorsePower	1.2241e+11	1	2191.7054	< 2.2e-16 ***
Cylinders	1.5137e+08	1	2.7102	0.09979 .
Doors	3.2758e+09	1	58.6504	2.329e-14 ***
MPGHighway	3.5721e+08	1	6.3956	0.01148 *
Residuals	2.3162e+11	4147		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Updated VIFG

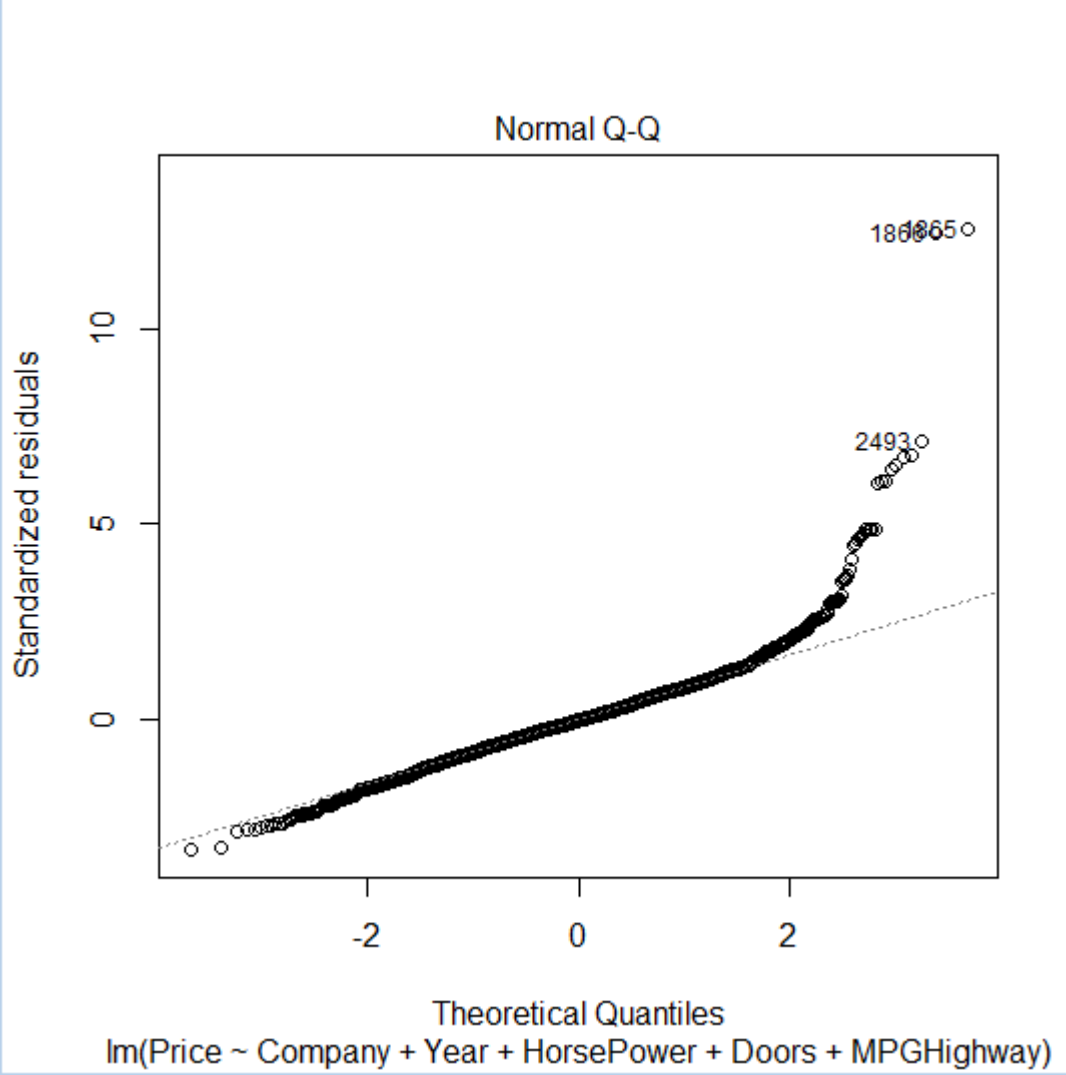
Now since all the above values are greater than 1 and lesser than 5, we can conclude that there definitely is a moderate correlation amongst our variables, but it is not severe enough to warrant corrective measures in our regression model.

	GVIF	Df	GVIF^(1/(2*Df))

Company	1.402370	4	1.043177
Year	2.578223	1	1.605685
HorsePower	4.098768	1	2.024541
Cylinders	4.597384	1	2.144151
Doors	1.292946	1	1.137078
MPGHighway	2.491464	1	1.578437

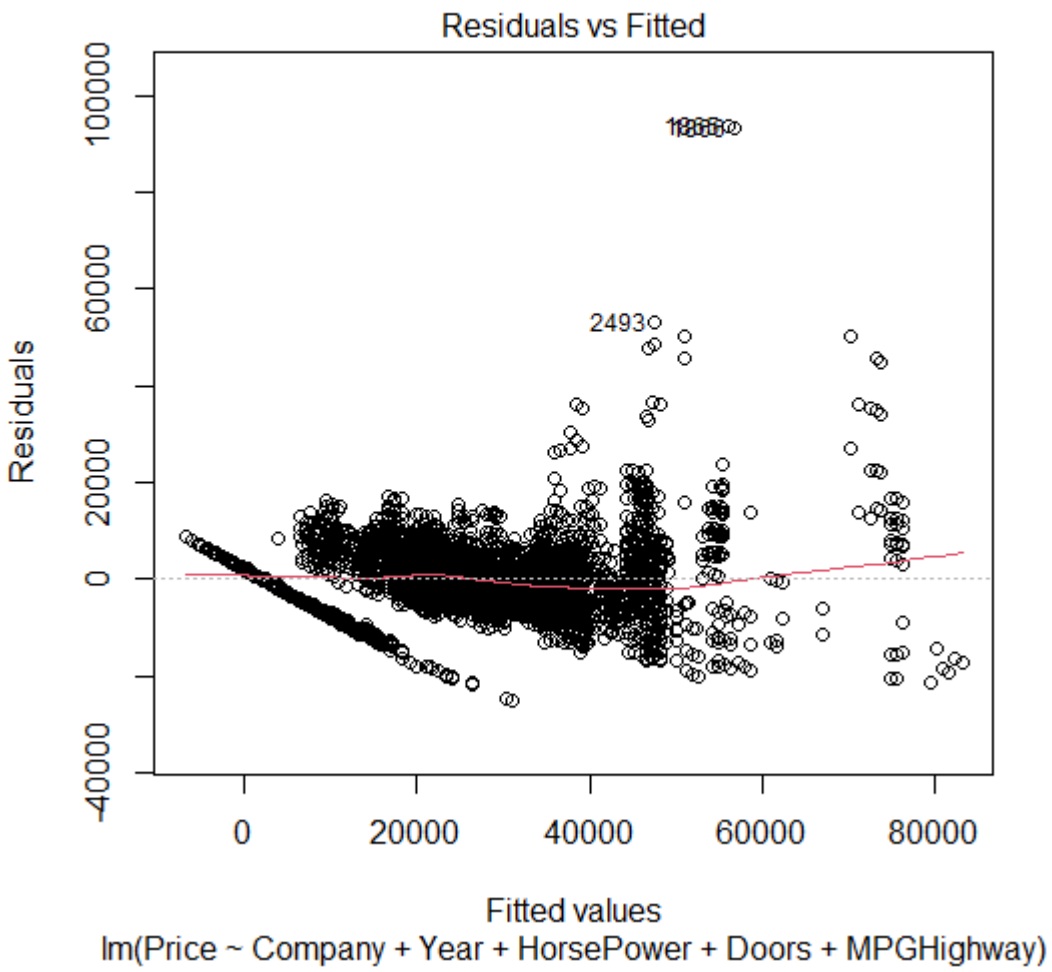
QQ Plot

Even though there are some observations that are far from the lines, and these are the potential outliers of the data. However, most of the data lies on 45 degrees line so it is safe to assume that our data is normally distributed.

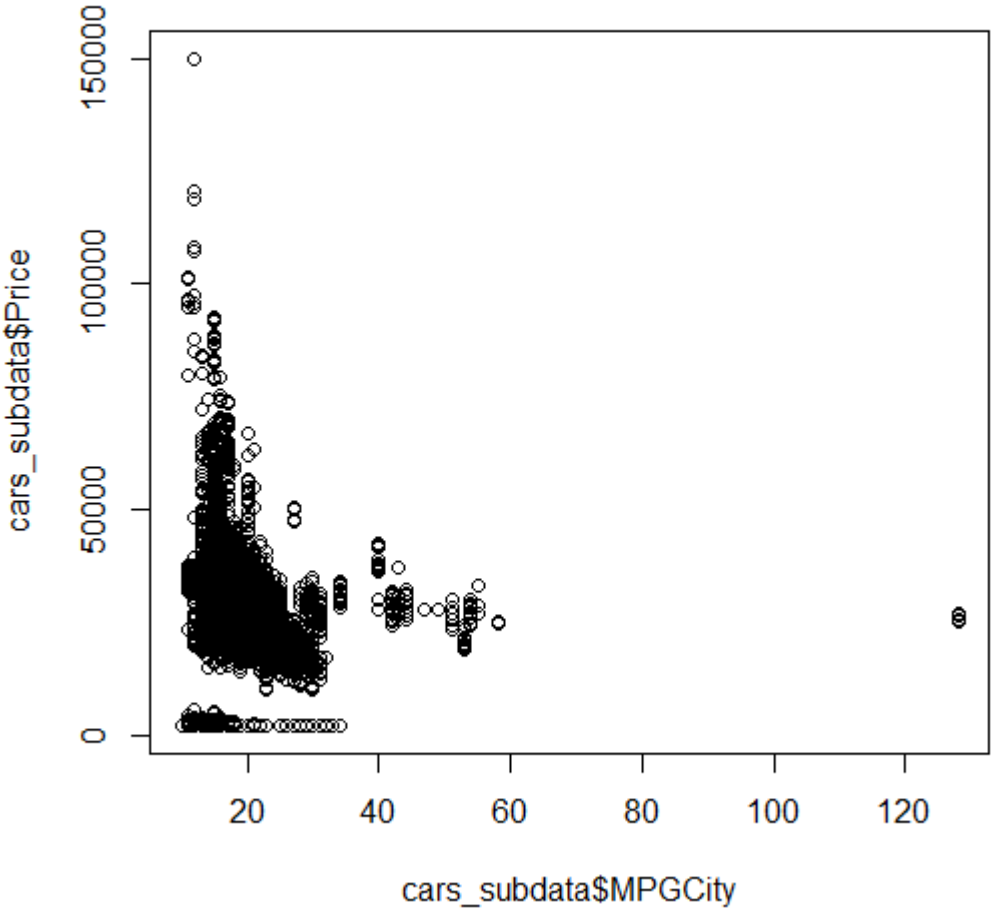


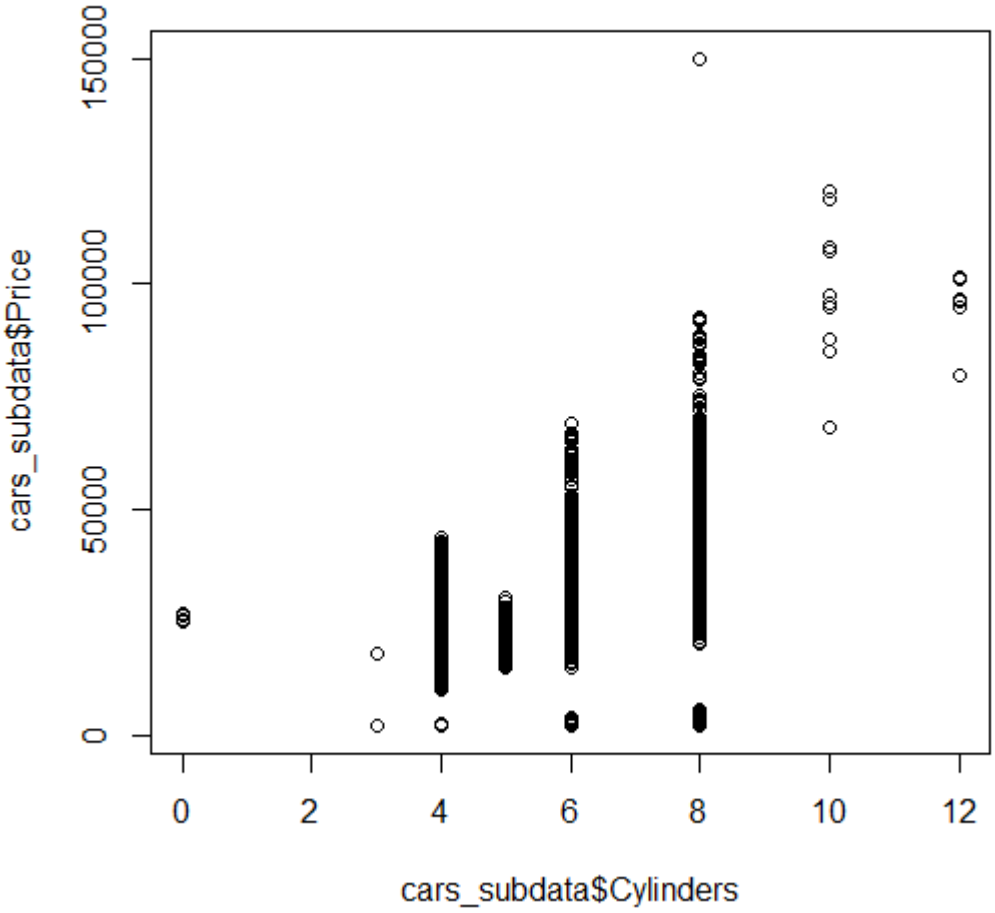
Residuals vs Fitted Graph

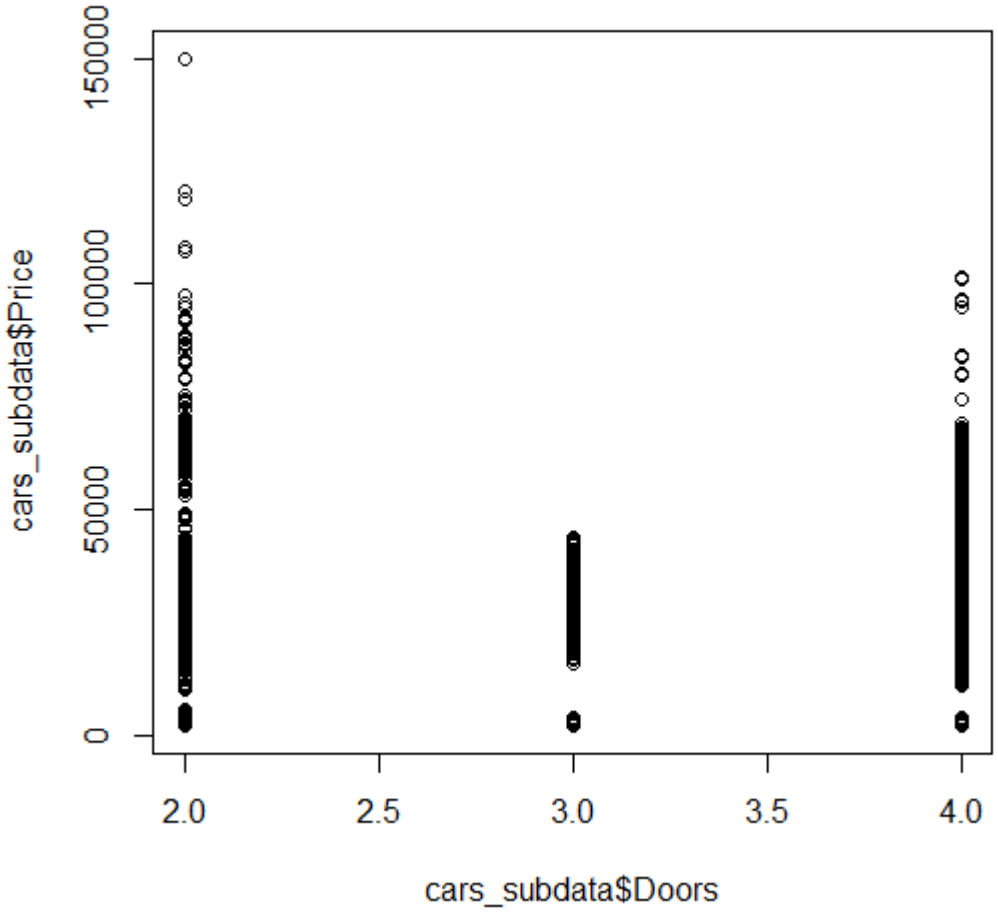
A pattern in the Residuals vs Fitted Values plot may indicate violation of one or more assumptions of the OLS regression model. However, since our plot looks like a random scatter about the line y = 0 showing no specific pattern and constant variance, this then naturally follows that there is no violation of any sort of assumption in our analysis.

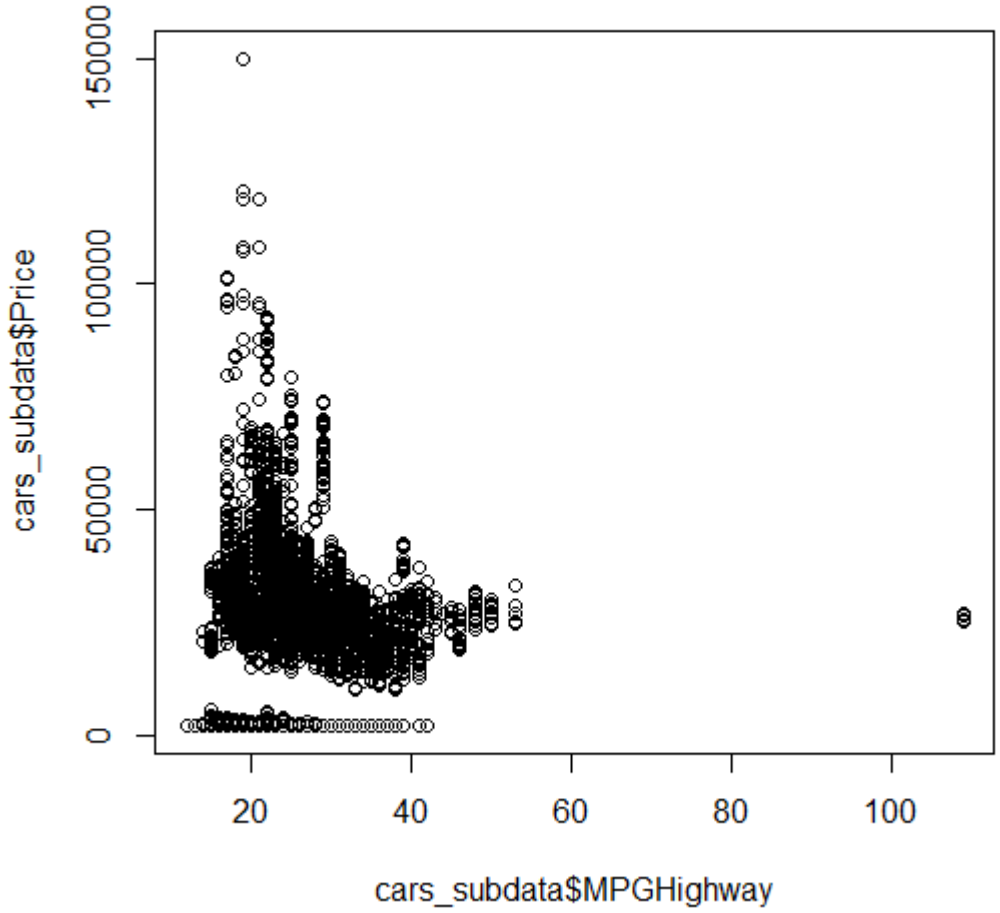


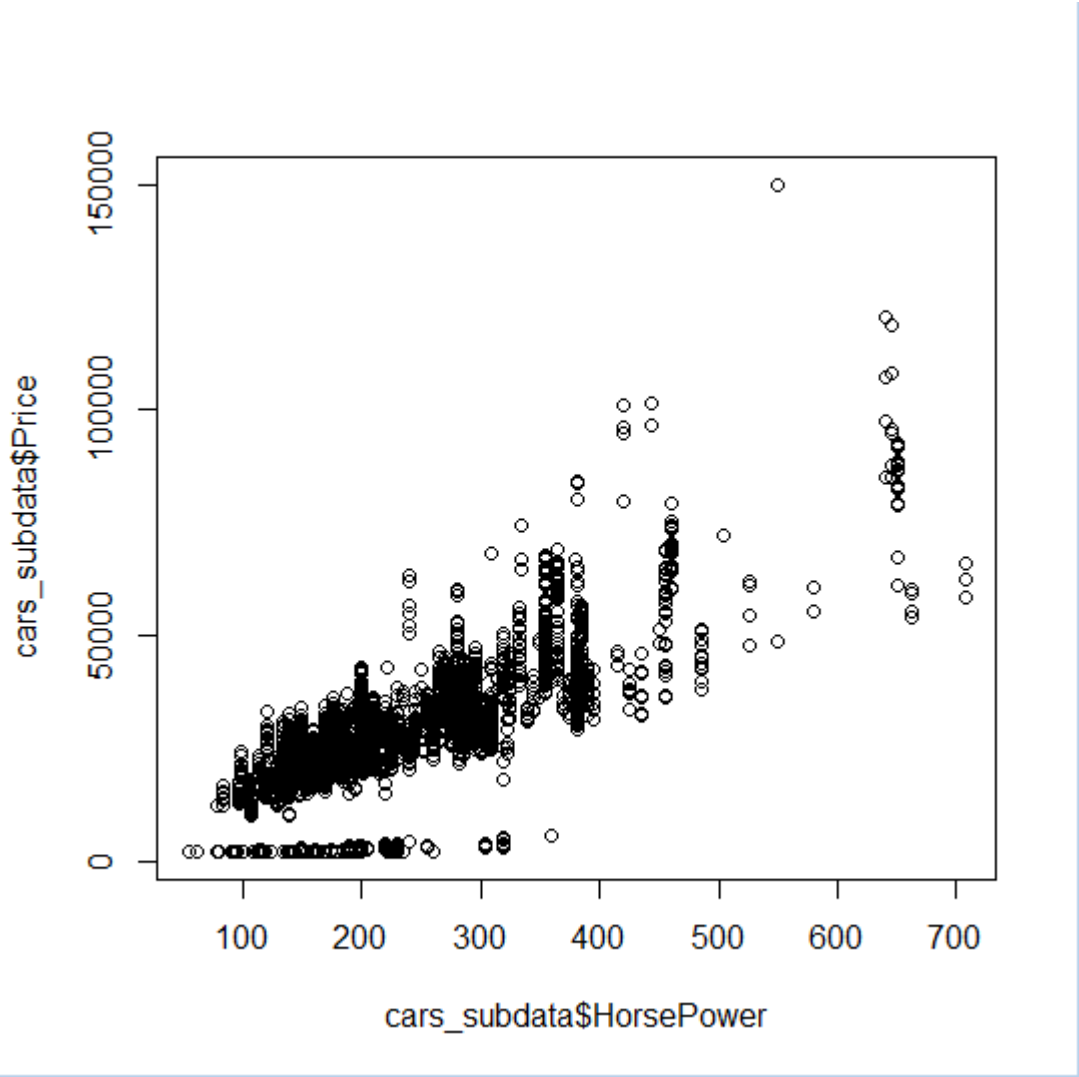
Scatter Plots


cars_subdata$Price
0
50000
100000
150000
20
40
60
80
100
120
cars_subdata$MPGCity


cars_subdata$Price
0
50000
100000
150000
0
2
4
6
8
10
12
cars_subdata$Cylinders


cars_subdata$Price
0
50000
100000
150000
2.0
2.5
3.0
3.5
4.0
cars_subdata$Doors


cars_subdata$Price
150000
100000
50000
0
20
40
60
80
100
cars_subdata$MPGHighway



Summary of Model

$h(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_6 x_6 + \varepsilon$ where $\varepsilon \sim N(0, 1)$

My general model has eight weights, and I estimated this weight using multiple linear model. We have different value for all weight which describes the change in the price of car if change respective independent variable by 1 unit and keep everything constant. β_1 show the weight of company and company is a categorical variable and it has five categories as I discussed earlier. We have five different values for this weight which will change the category of the variable will change to predict the cars price.

Chevrolet, Ford, Volkswagen, Toyota, and Dodge are the five categories of company variable. The value of β_1 will be 0, -6.819e+01, 3.724e+03, -2.511e+02 and -1.788e+03 respectively. Value for β0, β2, β3, β4, β5, β6 and β7 will be -1.454e+06, 7.213e+02, 1.140e+02, -2.570e+02, 1.125e+03, and 5.954e+01and respectively. So, we can write our model equations as follows.

i) For Chevrolet:

h(x) = -1.454e+06 + 7.213e+02 * x2 + 1.140e+02 * x3 - 2.570e+02 * x4 + 1.125e+03 * x5 + 5.954e+01 * x6

ii) For Ford:

h(x) = -1.454e+06 - 6.819e+01 * x1 + 7.213e+02 * x2 + 1.140e+02 * x3 - 2.570e+02 * x4 + 1.125e+03 * x5 + 5.954e+01 * x6

iii) For Volkswagen:

h(x) = -1.454e+06 + 3.724e+03 * x1 + 7.213e+02 * x2 + 1.140e+02 * x3 - 2.570e+02 * x4 + 1.125e+03 * x5 + 5.954e+01 * x6

iv) For Toyota:

h(x) = -1.454e+06 - 2.511e+02* x1 + 7.213e+02 * x2 + 1.140e+02 * x3 - 2.570e+02 * x4 + 1.125e+03 * x5 + 5.954e+01 * x6

v) For Dodge:

h(x) = -1.454e+06 - 1.788e+03* x1 + 7.213e+02 * x2 + 1.140e+02 * x3 - 2.570e+02 * x4 + 1.125e+03 * x5 + 5.954e+01 * x6

Conclusion

As shown in the above equations, car's price is affected by company, year, horsepower, number of cylinders, number of doors, miles per gallon on the highway and miles per gallon in the city. The magnitude of their impact is shown by the respective Betas of these variables. It can also be observed in the above tables that all our variables are significant as all the variables have their P values less than 0.05. Since the adjusted R square has value of 0.7821, this means that 78.21 % variability in our model is explained by our independent variables. This means that our model significantly predicts the prices of the car.

Appendix

```
getwd()

setwd('C:/Users/2018n/Desktop/Statistics (MATH 231)/Arbaz Khan/Project')

cars_data <- read.csv("data.csv", header=TRUE, stringsAsFactors = FALSE)

class(cars_data)

head(cars_data)

tail(cars_data)

cars_subdata <- subset(cars_data, select = c("Make", "Year"coor, "Engine.HP",
"Engine.Cylinders", "Number.of.Doors", "highway.MPG", "city.mpg", "MSRP"))
```

```
head(cars_subdata)

tail(cars_subdata)

colnames(cars_subdata)[1] <- 'Company'

colnames(cars_subdata)[3] <- 'HorsePower'

colnames(cars_subdata)[4] <- 'Cylinders'

colnames(cars_subdata)[5] <- 'Doors'

colnames(cars_subdata)[6] <- 'MPGHighway'

colnames(cars_subdata)[7] <- 'MPGCity'

colnames(cars_subdata)[8] <- 'Price'

head(cars_subdata)

tail(cars_subdata)



cars_subdata <- subset(cars_subdata, Company == 'Chevrolet' | Company == 'Ford' | Company == 'Volkswagen' | Company == 'Toyota' | Company == 'Dodge')

rownames(cars_subdata) <- NULL

head(cars_subdata)

tail(cars_subdata)



sum(is.na(cars_subdata$Company))
```

```
sum(is.na(cars_subdata$Year))

sum(is.na(cars_subdata$Horsepower))

sum(is.na(cars_subdata$Cylinders))

sum(is.na(cars_subdata$Doors))

sum(is.na(cars_subdata$MPGHighway))

sum(is.na(cars_subdata$MPGCity))

sum(is.na(cars_subdata$Price))

sum(is.na(cars_subdata))

nrow(cars_subdata)

cars_subdata <- na.omit(cars_subdata)

nrow(cars_subdata)

rownames(cars_subdata) <- NULL

head(cars_subdata)

tail(cars_subdata)



summary(cars_subdata$Year)

boxplot(cars_subdata$Year, ylab='Year')

summary(cars_subdata$HorsePower)

boxplot(cars_subdata$HorsePower, ylab='HorsePower')
```

```
summary(cars_subdata$Cylinders)

boxplot(cars_subdata$Cylinders, ylab='Cylinders')

summary(cars_subdata$Doors)

boxplot(cars_subdata$Doors, ylab='Doors')

summary(cars_subdata$MPGHighway)

boxplot(cars_subdata$MPGHighway, ylab='MPGHighway')

summary(cars_subdata$MPGCity)

boxplot(cars_subdata$MPGCity, ylab='MPGCity')

summary(cars_subdata$Price)

boxplot(cars_subdata$Price, ylab='Price')


model <- lm(Price ~ Company + Year + HorsePower + Doors + MPGHighway, data =
cars_subdata)

summary(model)

plot(model)


plot(cars_subdata$HorsePower, cars_subdata$Price)

plot(cars_subdata$Year, cars_subdata$Price)

plot(cars_subdata$Cylinders, cars_subdata$Price)
```

```
plot(cars_subdata$Doors, cars_subdata$Price)

plot(cars_subdata$MPGHighway, cars_subdata$Price)

plot(cars_subdata$MPGCity, cars_subdata$Price)
```