Assignment 1

Exploratory data analysis

(Dataset: Car sales, Data Source: Kaggle.com)

Submitted by,

Ponkothandaraman.S

Submitted to,

Dr.V.Bhuvaneshwari

Assumptions :

- The dataset explains the factors of a car sale.
- Horsepower is the most important thing on a car, it is the thing that defines the value of the car. If the car have more horsepower then the car have more value.
- We assume that horsepower plays a vital role in the sales and the price of the vehicle.
- Fuel efficiency is the key feature of a vehicle, it may also affect the sales of the vehicle
- We will be performing EDA to verify these assumptions.

```
library(MASS)
library(dplyr)

library(plyr)

library(lattice)
library(ggplot2)
library(tidyverse)

df1 <- read.csv('Car_sales.csv')
head(df1)

##   Manufacturer   Model Sales_in_thousands X__year_resale_value
Vehicle_type
## 1        Acura Integra             16.919               16.360
Passenger
## 2        Acura      TL             39.384               19.875
Passenger
## 3        Acura      CL             14.114               18.225
Passenger
## 4        Acura      RL              8.588               29.725
Passenger
## 5         Audi      A4             20.397               22.255
Passenger
## 6         Audi      A6             18.780               23.555
Passenger
##   Price_in_thousands Engine_size Horsepower Wheelbase Width Length
Curb_weight
## 1              21.50         1.8        140     101.2  67.3  172.4
2.639
## 2              28.40         3.2        225     108.1  70.3  192.9
3.517
## 3                 NA         3.2        225     106.9  70.6  192.0
3.470
## 4              42.00         3.5        210     114.6  71.4  196.6
3.850
## 5              23.99         1.8        150     102.6  68.2  178.0
2.998
## 6              33.95         2.8        200     108.7  76.1  192.0
3.561
##   Fuel_capacity Fuel_efficiency Latest_Launch Power_perf_factor
## 1          13.2              28      2/2/2012          58.28015
## 2          17.2              25      6/3/2011          91.37078
## 3          17.2              26      1/4/2012                NA
## 4          18.0              22     3/10/2011          91.38978
## 5          16.4              27     10/8/2011          62.77764
## 6          18.5              22      8/9/2011          84.56511
```

```
#Descriptive statistics
summary(df1)

##  Manufacturer          Model           Sales_in_thousands
X__year_resale_value
##  Length:157         Length:157         Min.   :  0.11     Min.   : 5.16
##  Class :character   Class :character   1st Qu.: 14.11     1st Qu.:11.26
##  Mode  :character   Mode  :character   Median : 29.45     Median :14.18
##                                        Mean   : 53.00     Mean   :18.07
##                                        3rd Qu.: 67.96     3rd Qu.:19.88
##                                        Max.   :540.56     Max.   :67.55
##                                                           NA's   :36
##  Vehicle_type       Price_in_thousands  Engine_size      Horsepower
##  Length:157         Min.   : 9.235     Min.   :1.000   Min.   : 55.0
##  Class :character   1st Qu.:18.017     1st Qu.:2.300   1st Qu.:149.5
##  Mode  :character   Median :22.799     Median :3.000   Median :177.5
##                     Mean   :27.391     Mean   :3.061   Mean   :185.9
##                     3rd Qu.:31.948     3rd Qu.:3.575   3rd Qu.:215.0
##                     Max.   :85.500     Max.   :8.000   Max.   :450.0
##                     NA's   :2          NA's   :1       NA's   :1
##    Wheelbase         Width           Length       Curb_weight
##  Min.   : 92.6   Min.   :62.60   Min.   :149.4   Min.   :1.895
##  1st Qu.:103.0   1st Qu.:68.40   1st Qu.:177.6   1st Qu.:2.971
##  Median :107.0   Median :70.55   Median :187.9   Median :3.342
##  Mean   :107.5   Mean   :71.15   Mean   :187.3   Mean   :3.378
##  3rd Qu.:112.2   3rd Qu.:73.42   3rd Qu.:196.1   3rd Qu.:3.800
##  Max.   :138.7   Max.   :79.90   Max.   :224.5   Max.   :5.572
##  NA's   :1       NA's   :1       NA's   :1       NA's   :2
##  Fuel_capacity   Fuel_efficiency Latest_Launch      Power_perf_factor
##  Min.   :10.30   Min.   :15.00   Length:157         Min.   : 23.28
##  1st Qu.:15.80   1st Qu.:21.00   Class :character   1st Qu.: 60.41
##  Median :17.20   Median :24.00   Mode  :character   Median : 72.03
##  Mean   :17.95   Mean   :23.84                      Mean   : 77.04
##  3rd Qu.:19.57   3rd Qu.:26.00                      3rd Qu.: 89.41
##  Max.   :32.00   Max.   :45.00                      Max.   :188.14
##  NA's   :1       NA's   :3                          NA's   :2

str(df1)

## 'data.frame':    157 obs. of  16 variables:
##  $ Manufacturer        : chr  "Acura" "Acura" "Acura" "Acura" ...
##  $ Model               : chr  "Integra" "TL" "CL" "RL" ...
##  $ Sales_in_thousands  : num  16.92 39.38 14.11 8.59 20.4 ...
##  $ X__year_resale_value: num  16.4 19.9 18.2 29.7 22.3 ...
##  $ Vehicle_type        : chr  "Passenger" "Passenger" "Passenger"
"Passenger" ...
##  $ Price_in_thousands  : num  21.5 28.4 NA 42 24 ...
##  $ Engine_size         : num  1.8 3.2 3.2 3.5 1.8 2.8 4.2 2.5 2.8 2.8 ...
##  $ Horsepower          : int  140 225 225 210 150 200 310 170 193 193 ...
##  $ Wheelbase           : num  101 108 107 115 103 ...
```

```
##  $ Width              : num  67.3 70.3 70.6 71.4 68.2 76.1 74 68.4 68.5
70.9 ...
##  $ Length             : num  172 193 192 197 178 ...
##  $ Curb_weight        : num  2.64 3.52 3.47 3.85 3 ...
##  $ Fuel_capacity      : num  13.2 17.2 17.2 18 16.4 18.5 23.7 16.6 16.6
18.5 ...
##  $ Fuel_efficiency    : int  28 25 26 22 27 22 21 26 24 25 ...
##  $ Latest_Launch      : chr  "2/2/2012" "6/3/2011" "1/4/2012" "3/10/2011"
...
##  $ Power_perf_factor  : num  58.3 91.4 NA 91.4 62.8 ...
```

```
#Checking null values
colSums(is.na(df1))
```

```
##         Manufacturer                Model   Sales_in_thousands
##                    0                    0                    0
## X__year_resale_value         Vehicle_type   Price_in_thousands
##                   36                    0                    2
##          Engine_size           Horsepower            Wheelbase
##                    1                    1                    1
##                Width               Length          Curb_weight
##                    1                    1                    2
##        Fuel_capacity      Fuel_efficiency        Latest_Launch
##                    1                    3                    0
##    Power_perf_factor
##                    2
```

```
#removing null values
df1 <- na.omit(df1)
colSums(is.na(df1))
```

```
##         Manufacturer                Model   Sales_in_thousands
##                    0                    0                    0
## X__year_resale_value         Vehicle_type   Price_in_thousands
##                    0                    0                    0
##          Engine_size           Horsepower            Wheelbase
##                    0                    0                    0
##                Width               Length          Curb_weight
##                    0                    0                    0
##        Fuel_capacity      Fuel_efficiency        Latest_Launch
##                    0                    0                    0
##    Power_perf_factor
##                    0
```

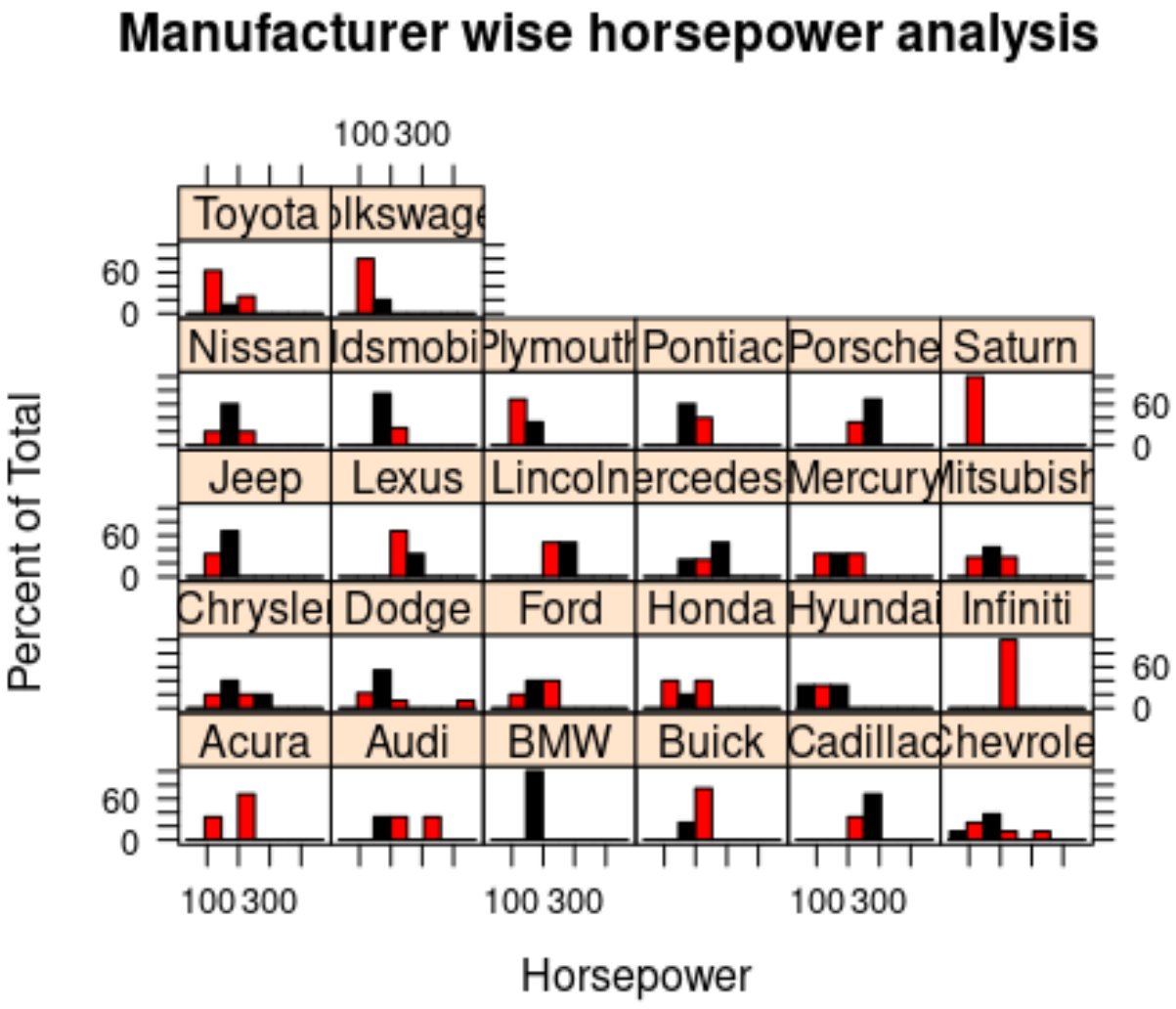
```
summary(df1)
```

```
##  Manufacturer          Model           Sales_in_thousands
X__year_resale_value
##  Length:117         Length:117         Min.   :  0.11     Min.   : 5.16
##  Class :character   Class :character   1st Qu.: 16.77     1st Qu.:11.24
##  Mode  :character   Mode  :character   Median : 32.30     Median :14.01
##                                        Mean   : 59.11     Mean   :18.03
##                                        3rd Qu.: 76.03     3rd Qu.:19.88
##                                        Max.   :540.56     Max.   :67.55
##  Vehicle_type       Price_in_thousands  Engine_size      Horsepower
##  Length:117         Min.   : 9.235     Min.   :1.000   Min.   : 55.0
##  Class :character   1st Qu.:16.980     1st Qu.:2.200   1st Qu.:140.0
##  Mode  :character   Median :21.665     Median :3.000   Median :175.0
##                     Mean   :25.969     Mean   :3.049   Mean   :181.3
##                     3rd Qu.:29.465     3rd Qu.:3.800   3rd Qu.:210.0
##                     Max.   :82.600     Max.   :8.000   Max.   :450.0
##    Wheelbase         Width           Length       Curb_weight
##  Min.   : 92.6   Min.   :62.60   Min.   :149.4   Min.   :1.895
##  1st Qu.:102.4   1st Qu.:68.50   1st Qu.:177.5   1st Qu.:2.911
##  Median :107.0   Median :70.40   Median :187.8   Median :3.340
##  Mean   :107.3   Mean   :71.19   Mean   :187.7   Mean   :3.324
##  3rd Qu.:111.6   3rd Qu.:73.60   3rd Qu.:196.5   3rd Qu.:3.823
##  Max.   :138.7   Max.   :79.30   Max.   :224.5   Max.   :5.115
##  Fuel_capacity   Fuel_efficiency Latest_Launch      Power_perf_factor
##  Min.   :10.30   Min.   :15.00   Length:117         Min.   : 23.28
##  1st Qu.:15.30   1st Qu.:22.00   Class :character   1st Qu.: 55.30
##  Median :17.20   Median :24.00   Mode  :character   Median : 70.66
##  Mean   :17.81   Mean   :24.12                      Mean   : 74.93
##  3rd Qu.:19.80   3rd Qu.:26.00                      3rd Qu.: 85.83
##  Max.   :32.00   Max.   :45.00                      Max.   :188.14
```

```
#Changing the specific variables to categorical
df1$Manufacturer=as.factor(df1$Manufacturer)
df1$Vehicle_type=as.factor(df1$Vehicle_type)
str(df1)
```

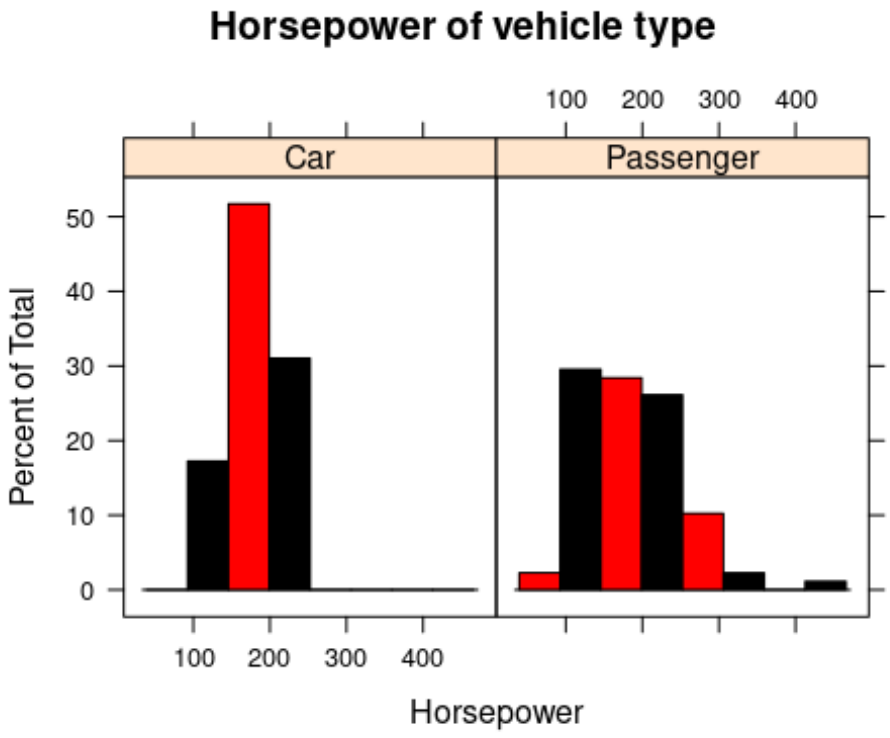
```
## 'data.frame':    117 obs. of  16 variables:
##  $ Manufacturer        : Factor w/ 26 levels "Acura","Audi",..: 1 1 1 2 2
2 3 3 4 4 ...
##  $ Model               : chr  "Integra" "TL" "RL" "A4" ...
##  $ Sales_in_thousands  : num  16.92 39.38 8.59 20.4 18.78 ...
##  $ X__year_resale_value: num  16.4 19.9 29.7 22.3 23.6 ...
##  $ Vehicle_type        : Factor w/ 2 levels "Car","Passenger": 2 2 2 2 2 2
2 2 2 2 ...
##  $ Price_in_thousands  : num  21.5 28.4 42 24 34 ...
##  $ Engine_size         : num  1.8 3.2 3.5 1.8 2.8 4.2 2.8 2.8 3.1 3.8 ...
##  $ Horsepower          : int  140 225 210 150 200 310 193 193 175 240 ...
##  $ Wheelbase           : num  101 108 115 103 109 ...
##  $ Width               : num  67.3 70.3 71.4 68.2 76.1 74 68.5 70.9 72.7
```

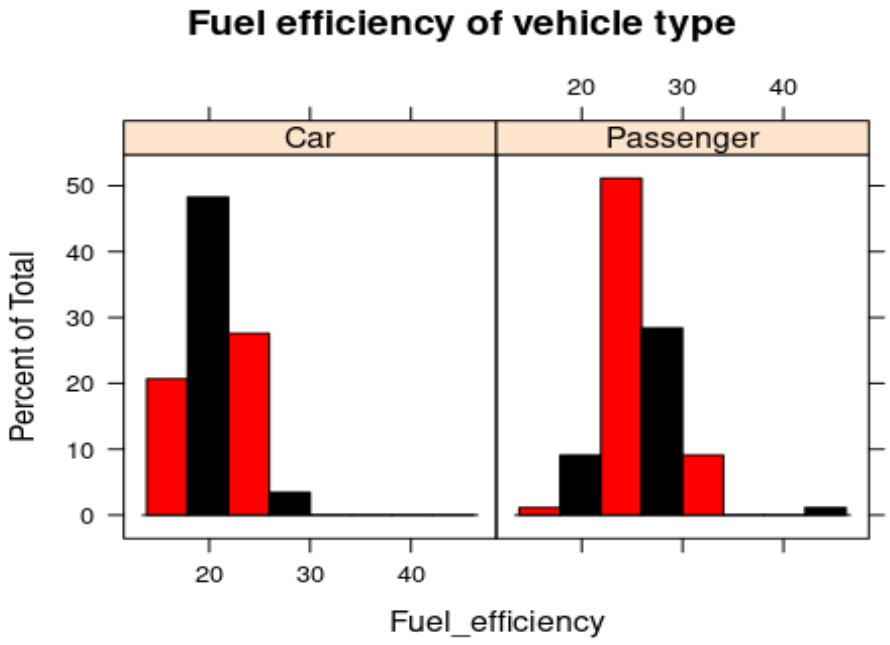
```
72.7 ...
##  $ Length              : num  172 193 197 178 192 ...
##  $ Curb_weight         : num  2.64 3.52 3.85 3 3.56 ...
##  $ Fuel_capacity       : num  13.2 17.2 18 16.4 18.5 23.7 16.6 18.5 17.5
17.5 ...
##  $ Fuel_efficiency     : int  28 25 22 27 22 21 24 25 25 23 ...
##  $ Latest_Launch       : chr  "2/2/2012" "6/3/2011" "3/10/2011"
"10/8/2011" ...
##  $ Power_perf_factor   : num  58.3 91.4 91.4 62.8 84.6 ...
##  - attr(*, "na.action")= 'omit' Named int [1:40] 3 8 16 19 28 34 35 39 45
51 ...
##   ..- attr(*, "names")= chr [1:40] "3" "8" "16" "19" ...

#Histogram analysis
histogram(~Horsepower|Manufacturer,data=df1,col=c("black","red"),main="Manufa
cturer wise horsepower analysis")
```



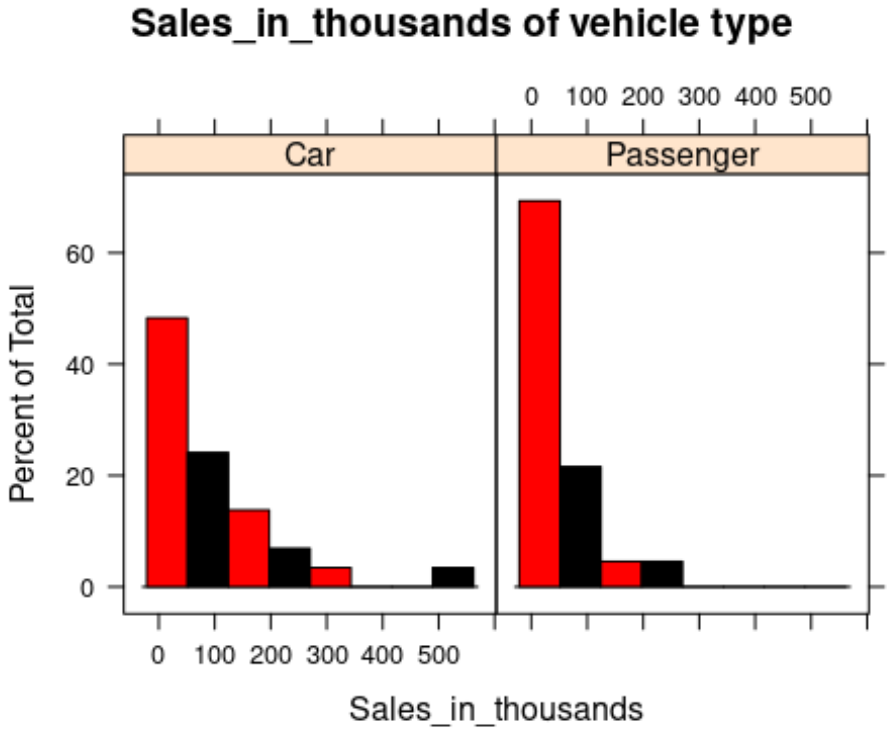
```
histogram(~Fuel_efficiency|Vehicle_type,data=df1,col=c("red","black"),main="F
uel efficiency of vehicle type")
```

```
histogram(~Horsepower|Vehicle_type,data=df1,col=c("red","black"),main="Horsep
ower of vehicle type")
```



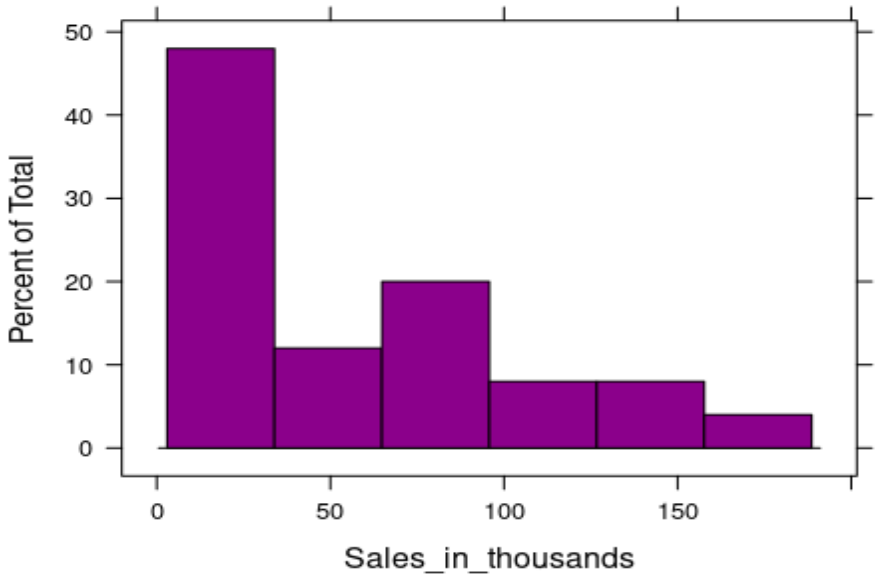


```
histogram(~Sales_in_thousands|Vehicle_type,data=df1,col=c("red","black"),main
="Sales_in_thousands of vehicle type")
```

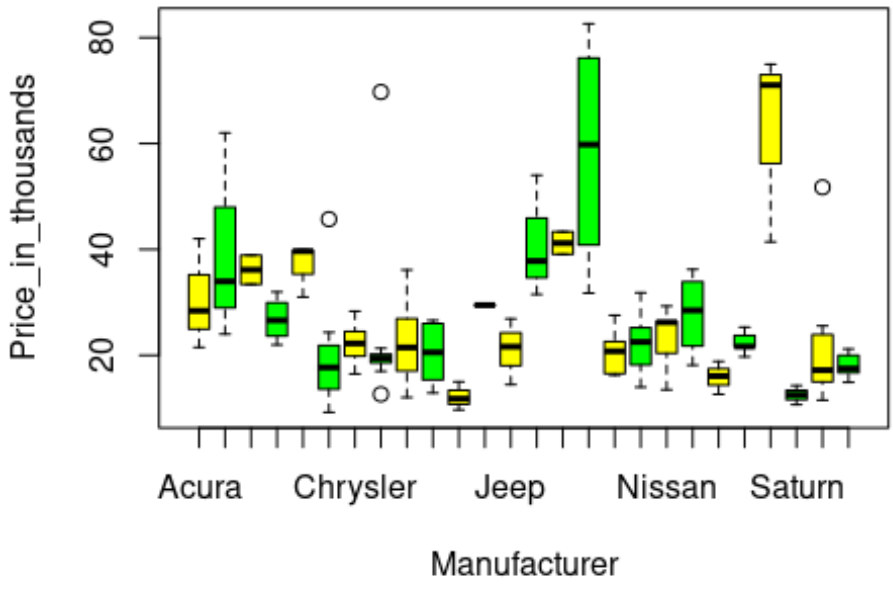



```
#Subsetting the vehicle type attribute to Car
df2<-subset(df1,Vehicle_type== 'Car' & Sales_in_thousands >
480,select=c(Vehicle_type,Sales_in_thousands))
head(df2)

##    Vehicle_type Sales_in_thousands
## 57          Car            540.561

df3<-subset(df1,Vehicle_type== 'Car' & Sales_in_thousands <
200,select=c(Vehicle_type,Sales_in_thousands))
head(df3)

##    Vehicle_type Sales_in_thousands
## 42          Car             16.767
## 43          Car             31.038
## 44          Car            111.313
## 46          Car            181.749
## 54          Car            155.787
## 55          Car            125.338
```

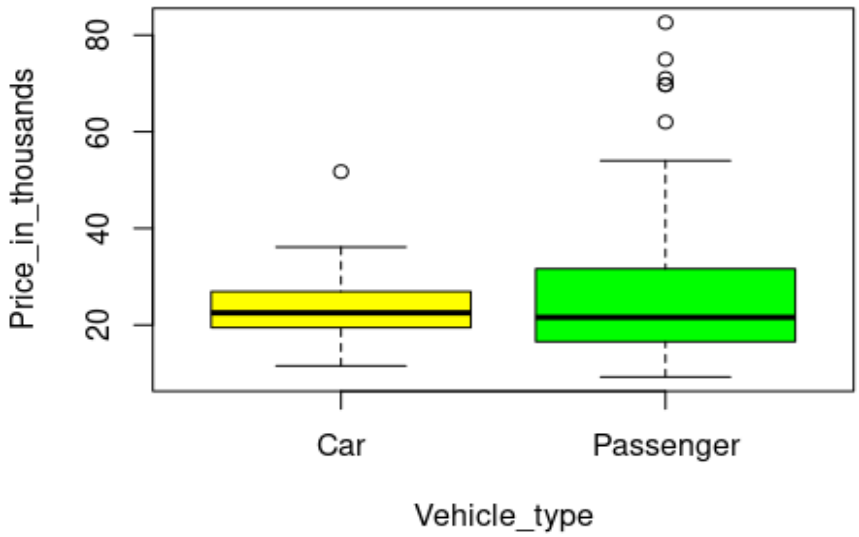
```
histogram(~Sales_in_thousands,data=df3,col="darkmagenta")
```



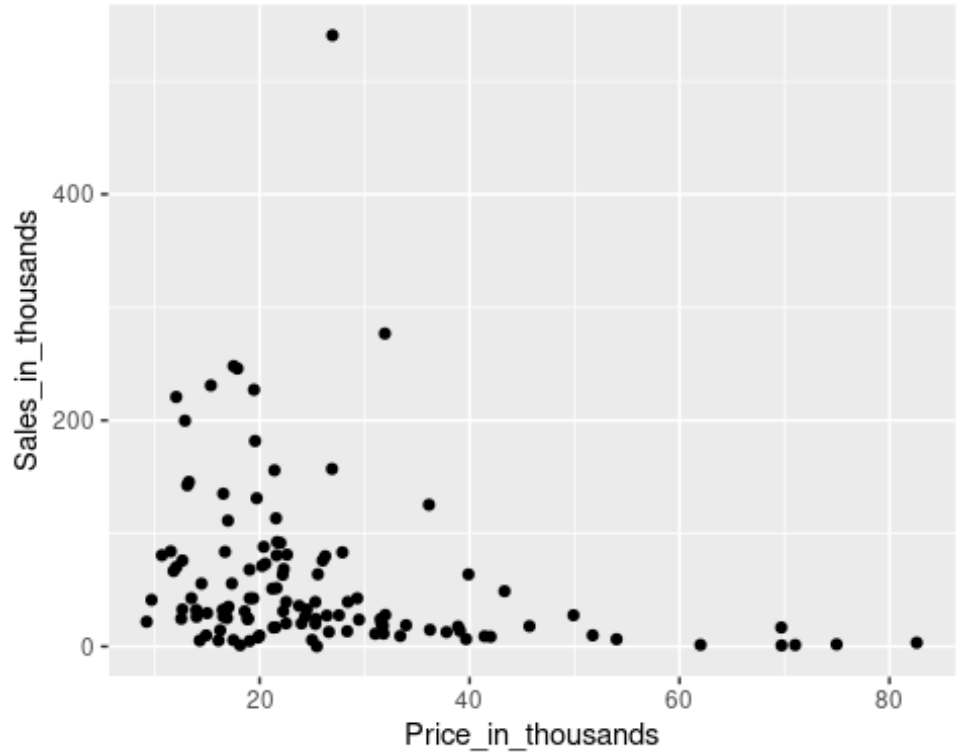

```
#Boxplot analysis
boxplot(Price_in_thousands~Manufacturer,data=df1,col=c("yellow","green"))
```



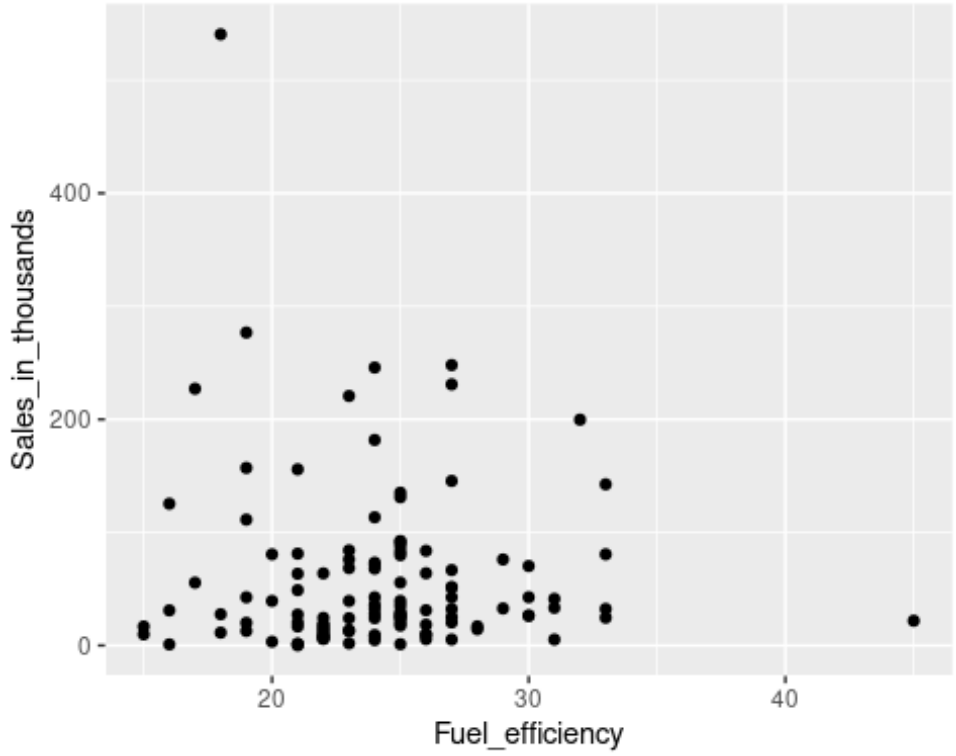
```
boxplot(Price_in_thousands~Vehicle_type,data=df1,col=c("yellow","green"))
```



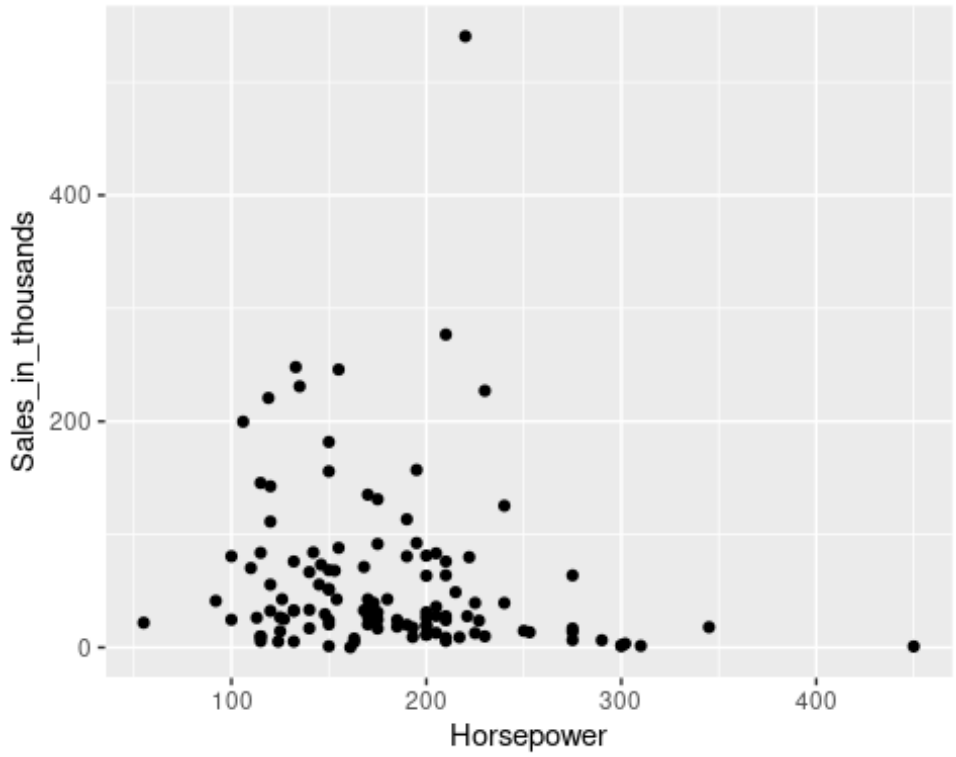

```
#Scatterplot analysis
ggplot(df1, aes(x=Price_in_thousands, y=Sales_in_thousands)) + geom_point()
```



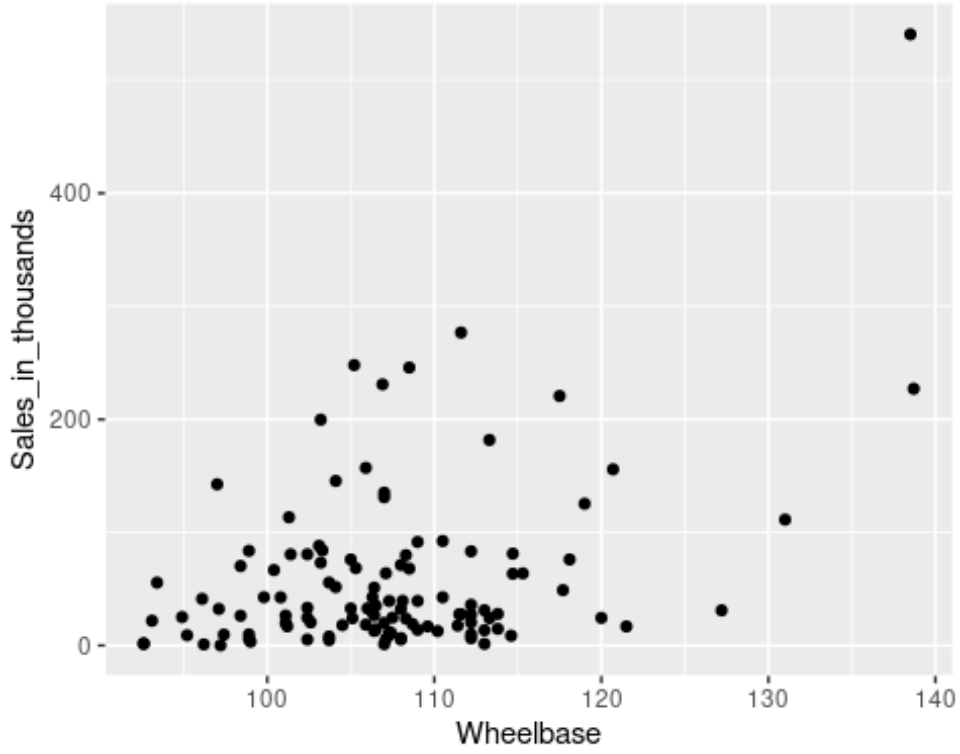
```
ggplot(df1, aes(x=Fuel_efficiency, y=Sales_in_thousands)) + geom_point()
```




```
ggplot(df1, aes(x=Horsepower, y=Sales_in_thousands)) + geom_point()
```



```
ggplot(df1, aes(x=Wheelbase, y=Sales_in_thousands)) + geom_point()
```



Attribute description :

- The “Car_sales” dataset have 16 features and 117 observations.
- There is no categorical variable in the dataset ,it is filled with numerical, integer and character data types.
- There are 26 manufacturers , Sales of the cars is scaled in thousands, there two vehicle types : Cars, passengers.
- And there are attributes like Horsepower , Fuel efficiency, Fuel capacity etc..,
- There are attributes based on resale value and latest launch.

Inference

Histogram:

- We can start the analysis with the 'Horsepower' attribute. Dodge is the only company producing car/cars with horsepower more than 350.Chevrolet mostly produce the cars with horsepower less 150, they also produce some cars having horsepower more than 300.All the BMW cars have the horsepower between 100-150.
- The fuel efficiency of passenger vehicle type is higher than of car vehicle type. There is an observation with fuel efficiency more than 40.Car vehicle type has the maximum fuel efficiency of 30.
- Passenger cars are available with horsepower over 100-350 while car vehicle type's maximum horse power is 250.
- Car vehicle type have higher sales than passenger vehicle type.

Boxplot:

- Most of the boxes in price-manufacturer boxplot are positively skewed. Outliers are detected in Chevrolet, Dodge and Toyota.
- In the price-vehicle boxplot 'cars' have no skewness, while the 'passenger' box was positively skewed. Outliers were detected in both of their upper bound.

Scatterplot:

- We compared sales with the horsepower, there is no correlation in it.
- Then we created boxplot for sales with price, horsepower and wheelbase. All of those boxplots exhibited no correlation.

Insights

- The cars produced by dodge and Chevrolet have high horsepower.
- All the other manufacturers producing cars with moderate horsepower.
- Cars vehicle type have high sales than passenger vehicle type.
- Passenger vehicle type high fuel efficiency than cars vehicle type.
- Most of the Passenger type vehicle price higher than median price value.
- Price of the car is independent of all the other features.
- Porsche produces more cars with lower than their median price value.
- Surprisingly passenger car manufacturers are producing vehicles with horsepower more than 400.
- Some of the passenger cars have high fuel efficiency.
- But none of these affect that mush of the price and sales of the car.
- Dodge car has the highest horsepower in the whole dataset.
- BMW and infiniti have a standard range of cars with horsepower 100-200 and 200-300HP respectively.