R-Basics.R

Shraddha Somani

```
# R Basics
# Q1) Compute the 5 Number Summary of all the vairables. Do any of the variables exhibit some skewness? Determine if any values could be declared missing, and then convert the values to missing NA. Rerun the 5Number Summary.
ToyotaPrices <- read.csv("D:/DADM/Assignment/ToyotaPrices.csv")
summary(ToyotaPrices)
```

```
##        Id             Price          Age_08_04       Mfg_Month
##  Min.   :   1.0   Min.   : 4350   Min.   : 1.00   Min.   : 1.000
##  1st Qu.: 361.8   1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 3.000
##  Median : 721.5   Median : 9900   Median :61.00   Median : 5.000
##  Mean   : 721.6   Mean   :10731   Mean   :55.95   Mean   : 5.549
##  3rd Qu.:1081.2   3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 8.000
##  Max.   :1442.0   Max.   :32500   Max.   :80.00   Max.   :12.000
##     Mfg_Year          KM               HP          Automatic
##  Min.   :1998   Min.   :     1   Min.   : 69.0   Min.   :0.00000
##  1st Qu.:1998   1st Qu.: 43000   1st Qu.: 90.0   1st Qu.:0.00000
##  Median :1999   Median : 63390   Median :110.0   Median :0.00000
##  Mean   :2000   Mean   : 68533   Mean   :101.5   Mean   :0.05571
##  3rd Qu.:2001   3rd Qu.: 87021   3rd Qu.:110.0   3rd Qu.:0.00000
##  Max.   :2004   Max.   :243000   Max.   :192.0   Max.   :1.00000
##        cc            Doors          Cylinders      Gears
##  Min.   : 1300   Min.   :2.000   Min.   :4   Min.   :3.000
##  1st Qu.: 1400   1st Qu.:3.000   1st Qu.:4   1st Qu.:5.000
##  Median : 1600   Median :4.000   Median :4   Median :5.000
##  Mean   : 1577   Mean   :4.033   Mean   :4   Mean   :5.026
##  3rd Qu.: 1600   3rd Qu.:5.000   3rd Qu.:4   3rd Qu.:5.000
##  Max.   :16000   Max.   :5.000   Max.   :4   Max.   :6.000
##  Quarterly_Tax        Weight      Mfr_Guarantee     BOVAG_Guarantee
##  Min.   : 19.00   Min.   :1000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 69.00   1st Qu.:1040   1st Qu.:0.0000   1st Qu.:1.0000
##  Median : 85.00   Median :1070   Median :0.0000   Median :1.0000
##  Mean   : 87.12   Mean   :1072   Mean   :0.4095   Mean   :0.8955
##  3rd Qu.: 85.00   3rd Qu.:1085   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :283.00   Max.   :1615   Max.   :1.0000   Max.   :1.0000
##  Guarantee_Period       ABS            Airbag_1         Airbag_2
##  Min.   : 3.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median : 3.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   : 3.815   Mean   :0.8134   Mean   :0.9708   Mean   :0.7228
##  3rd Qu.: 3.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :36.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      Airco          Automatic_airco   Boardcomputer      CD_Player
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.00000   Median :0.0000   Median :0.0000
##  Mean   :0.5084   Mean   :0.05641   Mean   :0.2946   Mean   :0.2187
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##   Central_Lock    Powered_Windows Power_Steering       Radio
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.000   Median :1.0000   Median :0.0000
##  Mean   :0.5801   Mean   :0.562   Mean   :0.9777   Mean   :0.1462
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##    Mistlamps      Sport_Model     Backseat_Divider  Metallic_Rim
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :0.257   Mean   :0.3001   Mean   :0.7702   Mean   :0.2047
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Radio_cassette      Tow_Bar
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1455   Mean   :0.2779
##  3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
```

```
library(e1071)
skewness(ToyotaPrices$Id)
```

```
## [1] 0.0007873344
```

```
skewness(ToyotaPrices$Price)
```

```
## [1] 1.700327
```

```
skewness(ToyotaPrices$Age_08_04)
```

```
## [1] -0.8249756
```

```
skewness(ToyotaPrices$Mfg_Month)
```

```
## [1] 0.2900542
```

```
skewness(ToyotaPrices$Mfg_Year)
```

```
## [1] 0.9094007
```

```
skewness(ToyotaPrices$KM)
```

```
## [1] 1.013791
```

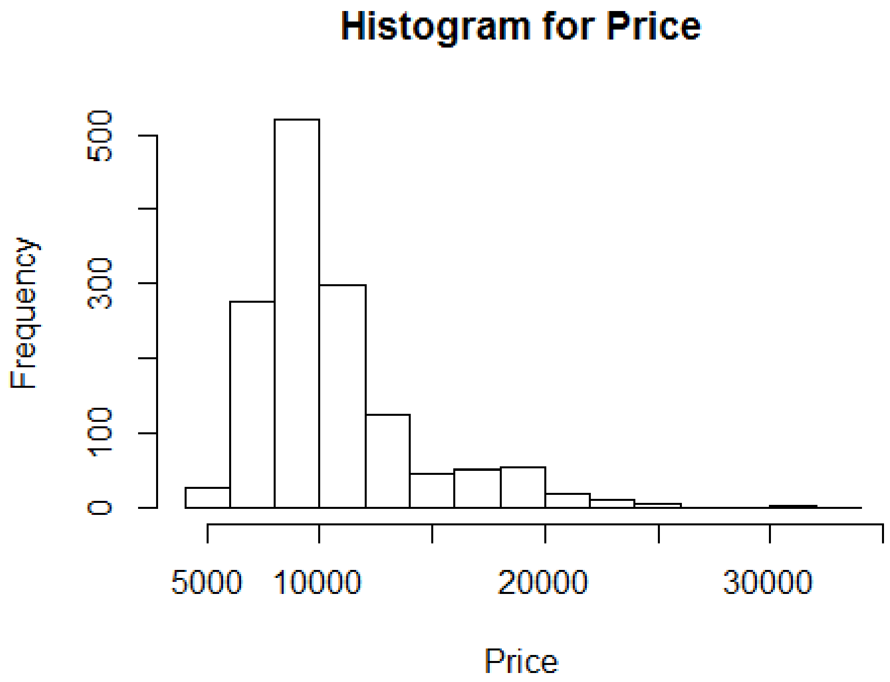
```
skewness(ToyotaPrices$HP)

## [1] 0.9538397

skewness(ToyotaPrices$Automatic)

## [1] 3.870099

skewness(ToyotaPrices$cc)

## [1] 27.37451

skewness(ToyotaPrices$Doors)

## [1] -0.07623547

skewness(ToyotaPrices$Cylinders)

## [1] NaN

skewness(ToyotaPrices$Gears)

## [1] 2.27919

skewness(ToyotaPrices$Quarterly_Tax)

## [1] 1.98967

skewness(ToyotaPrices$Weight)

## [1] 3.102148

skewness(ToyotaPrices$Mfr_Guarantee)

## [1] 0.367818

skewness(ToyotaPrices$BOVAG_Guarantee)

## [1] -2.583797

skewness(ToyotaPrices$Guarantee_Period)

## [1] 5.826243

skewness(ToyotaPrices$ABS)

## [1] -1.606941

skewness(ToyotaPrices$Airbag_1)

## [1] -5.581705

skewness(ToyotaPrices$Airbag_2)

## [1] -0.9946859
```

```
skewness(ToyotaPrices$Airco)

## [1] -0.03339594

skewness(ToyotaPrices$Automatic_airco)

## [1] 3.841523

skewness(ToyotaPrices$Boardcomputer)

## [1] 0.9003747

skewness(ToyotaPrices$CD_Player)

## [1] 1.359867

skewness(ToyotaPrices$Central_Lock)

## [1] -0.324185

skewness(ToyotaPrices$Powered_Windows)

## [1] -0.2495767

skewness(ToyotaPrices$Power_Steering)

## [1] -6.46609

skewness(ToyotaPrices$Radio)

## [1] 2.000253

skewness(ToyotaPrices$Mistlamps)

## [1] 1.111236

skewness(ToyotaPrices$Sport_Model)

## [1] 0.8712372

skewness(ToyotaPrices$Backseat_Divider)

## [1] -1.283138

skewness(ToyotaPrices$Metallic_Rim)

## [1] 1.461959

skewness(ToyotaPrices$Radio_cassette)

## [1] 2.008161

skewness(ToyotaPrices$Tow_Bar)

## [1] 0.9908115
```

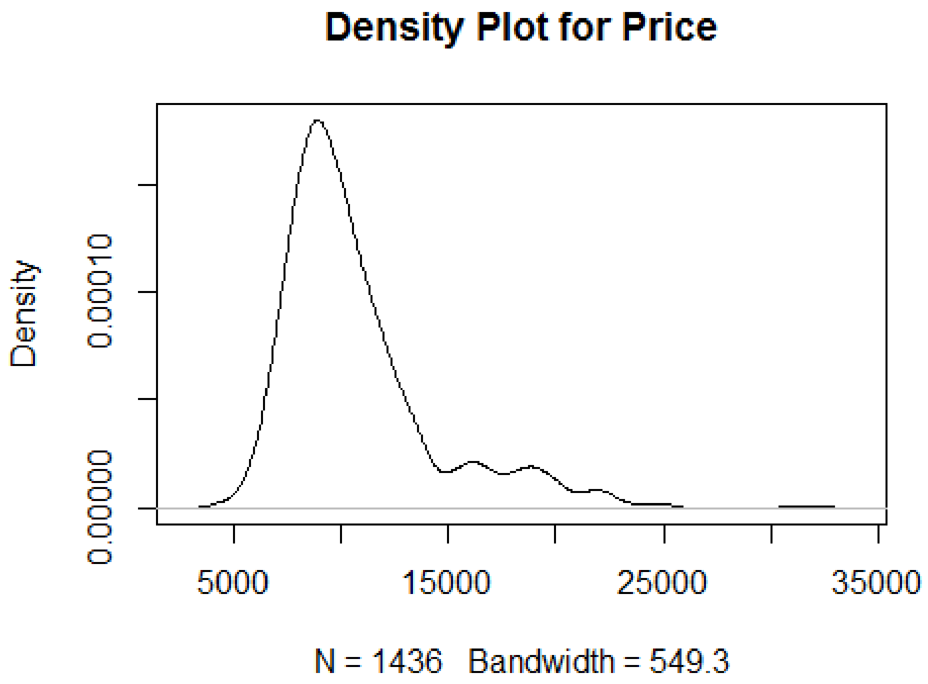
```r
# Analysis:
# - Skewness is the measure of symmetry.
# - The following varibles exhibit Positive Skewness - ID; Price; Mfg_Month;
Mfg_Year; KM; HP; Automatic; cc; Gears; Quarterly_Tax; Weight; Mfr_Gurantee;
Gurantee_Period; Automatic_airco; Boardcomputer; CD_Player; Radio; Mistlamps;
Sport_Model; Metallic_Rim; Radio_cassette and Tow_Bar.
# - Positive Skewness is also called as Left Skew.
# - The following variables exhibit Negative Skewness - Age_08_04; Doors;
BOVAG_Gurantee; ABS; Airbag_1; Airbag_2; Airco; Central_Lock; Power_Windows;
Power_Steering and Backseat_Divider
# - Negative Skewness is also called as Right Skew.
# - No Skew : Cylinder
# - is.na(ToyotaPrices)         #returns TRUE is any values are missing
# - There are no values that can be declared as missing values


# Q2) Convert categorical variables to factor. After doing the conversions
rerun the 5 Number Summary. Do any of the factor variables have "unbalanced"
counts; ie, more of one kind than another? Unbalanced counts would tend to
weaken the strength of a factor to predict the price of a Toyota.
ToyotaPrices$Automatic = factor(ToyotaPrices$Automatic)
ToyotaPrices$Doors = factor(ToyotaPrices$Doors)
ToyotaPrices$Cylinders = factor(ToyotaPrices$Cylinders)
ToyotaPrices$Gears = factor(ToyotaPrices$Gears)
ToyotaPrices$Mfr_Guarantee = factor(ToyotaPrices$Mfr_Guarantee)
ToyotaPrices$BOVAG_Guarantee = factor(ToyotaPrices$BOVAG_Guarantee)
ToyotaPrices$ABS = factor(ToyotaPrices$ABS)
ToyotaPrices$Airbag_1 = factor(ToyotaPrices$Airbag_1)
ToyotaPrices$Airbag_2 = factor(ToyotaPrices$Airbag_2)
ToyotaPrices$Airco = factor(ToyotaPrices$Airco)
ToyotaPrices$Automatic_airco = factor(ToyotaPrices$Automatic_airco)
ToyotaPrices$Boardcomputer = factor(ToyotaPrices$Boardcomputer)
ToyotaPrices$CD_Player = factor(ToyotaPrices$CD_Player)
ToyotaPrices$Central_Lock = factor(ToyotaPrices$Central_Lock)
ToyotaPrices$Powered_Windows = factor(ToyotaPrices$Powered_Windows)
ToyotaPrices$Power_Steering = factor(ToyotaPrices$Power_Steering)
ToyotaPrices$Radio = factor(ToyotaPrices$Radio)
ToyotaPrices$Mistlamps = factor(ToyotaPrices$Mistlamps)
ToyotaPrices$Sport_Model = factor(ToyotaPrices$Sport_Model)
ToyotaPrices$Backseat_Divider = factor(ToyotaPrices$Backseat_Divider)
ToyotaPrices$Metallic_Rim = factor(ToyotaPrices$Metallic_Rim)
ToyotaPrices$Radio_cassette = factor(ToyotaPrices$Radio_cassette)
ToyotaPrices$Tow_Bar = factor(ToyotaPrices$Tow_Bar)
summary(ToyotaPrices)
```

```
##        Id             Price         Age_08_04       Mfg_Month
##  Min.   :   1.0   Min.   : 4350   Min.   : 1.00   Min.   : 1.000
##  1st Qu.: 361.8   1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 3.000
##  Median : 721.5   Median : 9900   Median :61.00   Median : 5.000
##  Mean   : 721.6   Mean   :10731   Mean   :55.95   Mean   : 5.549
##  3rd Qu.:1081.2   3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 8.000
##  Max.   :1442.0   Max.   :32500   Max.   :80.00   Max.   :12.000
##     Mfg_Year          KM               HP        Automatic       cc
##  Min.   :1998   Min.   :     1   Min.   : 69.0   0:1356    Min.   : 1300
##  1st Qu.:1998   1st Qu.: 43000   1st Qu.: 90.0   1:  80    1st Qu.: 1400
##  Median :1999   Median : 63390   Median :110.0             Median : 1600
##  Mean   :2000   Mean   : 68533   Mean   :101.5             Mean   : 1577
##  3rd Qu.:2001   3rd Qu.: 87021   3rd Qu.:110.0             3rd Qu.: 1600
##  Max.   :2004   Max.   :243000   Max.   :192.0             Max.   :16000
##  Doors   Cylinders Gears    Quarterly_Tax        Weight     Mfr_Guarantee
##  2:  2   4:1436    3:   2   Min.   : 19.00   Min.   :1000   0:848
##  3:622             4:   1   1st Qu.: 69.00   1st Qu.:1040   1:588
##  4:138             5:1390   Median : 85.00   Median :1070
##  5:674             6:  43   Mean   : 87.12   Mean   :1072
##                             3rd Qu.: 85.00   3rd Qu.:1085
##                             Max.   :283.00   Max.   :1615
##  BOVAG_Guarantee Guarantee_Period ABS      Airbag_1 Airbag_2 Airco
##  0: 150          Min.   : 3.000   0: 268   0:  42   0: 398   0:706
##  1:1286          1st Qu.: 3.000   1:1168   1:1394   1:1038   1:730
##                  Median : 3.000
##                  Mean   : 3.815
##                  3rd Qu.: 3.000
##                  Max.   :36.000
##  Automatic_airco Boardcomputer CD_Player Central_Lock Powered_Windows
##  0:1355          0:1013        0:1122    0:603        0:629
##  1:  81          1: 423        1: 314    1:833        1:807
##
##
##
##
##  Power_Steering Radio    Mistlamps Sport_Model Backseat_Divider
##  0:  32         0:1226   0:1067    0:1005      0: 330
##  1:1404         1: 210   1: 369    1: 431      1:1106
##
##
##
##
##  Metallic_Rim Radio_cassette Tow_Bar
##  0:1142       0:1227         0:1037
##  1: 294       1: 209         1: 399
##
##
##
##

# Analysis:
# - Unbalanced count are present in the following variables:
# - Automatic; Mrf_Guarantee; BOVAG_Guarantee; ABS; Airbag_1; Airbag_2;
Airco; Automatic_airco; Boardcomputer; CD_Player; Central_Lock;
Powered_Windows; Power_Steering; Radio; Mistlamps; Sport_Model;
```

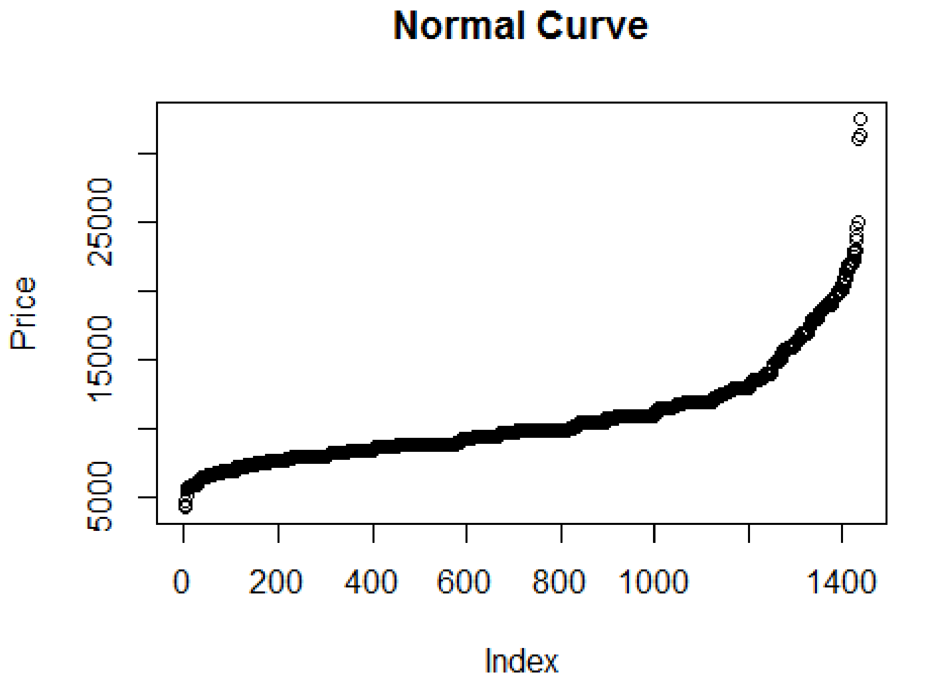
```
Backseat_Divider; Metallic_Rim; Radio_Cassette and Tow_Bar


# Q3) Explore the distribution of Price. Prepare the necessary plots, such as
histogram, density plot, sort plot, QQplot. Is the variable normal? Is the
variable skewed? Are there any clusters?
require(ggplot2)

## Loading required package: ggplot2

# Histogram
hist(ToyotaPrices$Price, main = "Histogram for Price", xlab = "Price")
```



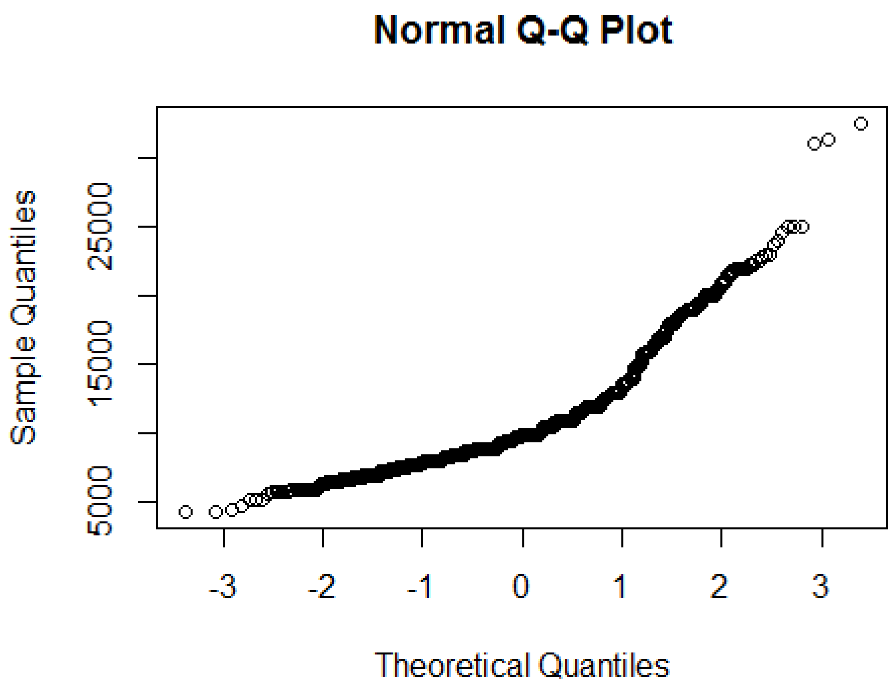
```
# Density Plot
plot(density(ToyotaPrices$Price), main = "Density Plot for Price")
```



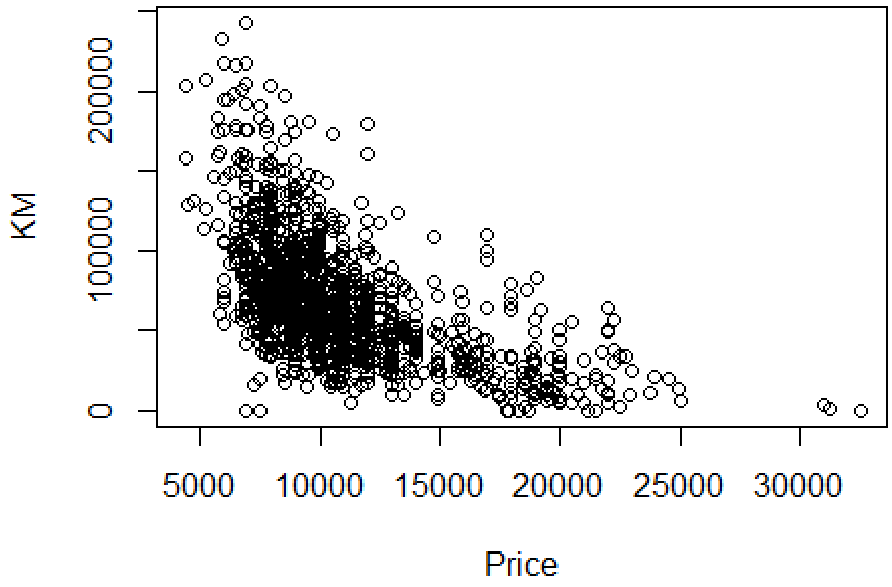
```r
# Sort Plot
plot(sort(ToyotaPrices$Price), main = "Normal Curve", ylab = "Price")
```



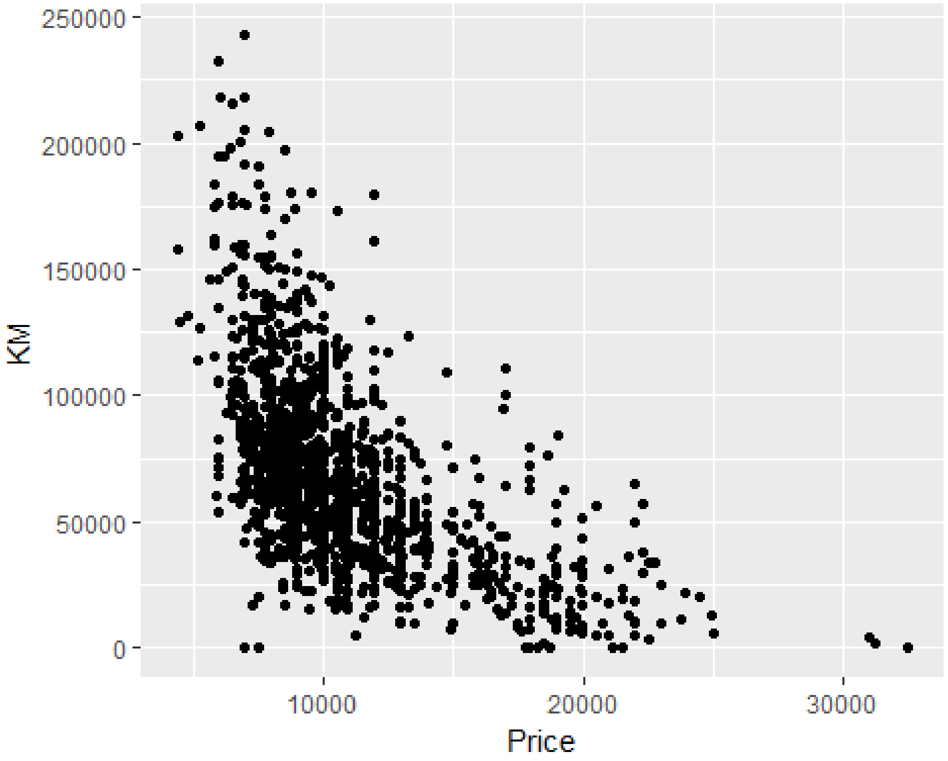
```
# QQPlot
qqnorm(ToyotaPrices$Price)
```



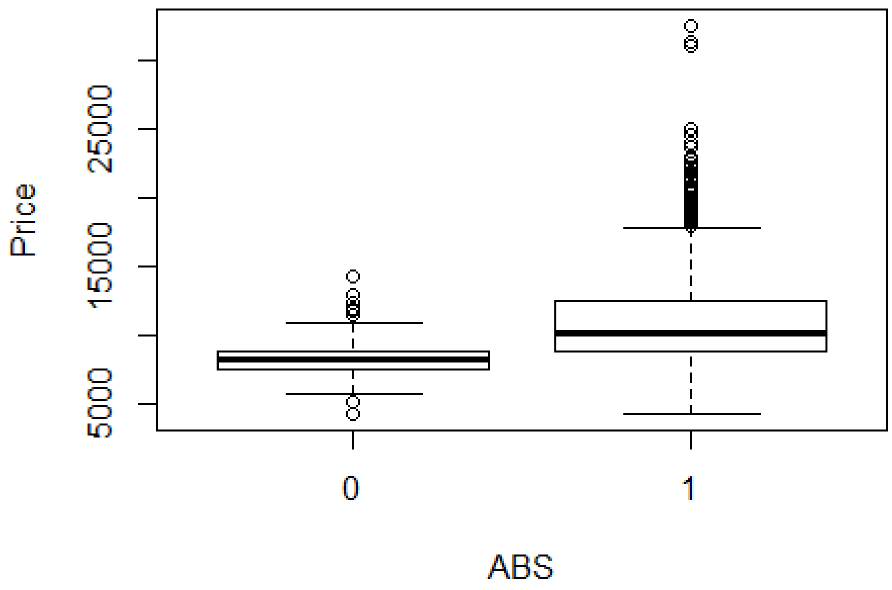
```
# Analysis:
# - The variable is normal.
# - The variable "Price" is positively skewed since the distribution is
concentrated on the left side of the figure.
# - There are  two clusters present.


# Q4) Produce the scatterplot of Price versus the number of KM (kilometers).
Use both the plot() and the qplot() functions. Does the relations look like a
line or a curve?
# plot()
plot(ToyotaPrices$Price, ToyotaPrices$KM, xlab = "Price", ylab = "KM")
```



```r
# qplot()
qplot(ToyotaPrices$Price, ToyotaPrices$KM, xlab = "Price", ylab = "KM")
```



```
# Analysis:
# - The relation looks like a curve.


# Q5) BoxWhisker plot of Price versus ABS
boxplot(Price ~ ABS, data = ToyotaPrices, xlab = "ABS", ylab = "Price")
```



```
# Analysis:
# - Yes, automobiles with anti-locking breaks tend to have a higher price.
# - Yes, there are outliers present in ABS.


# Q6) Compute the correlation between Price and KM. Is it a strong or weak
correlation? Is it positive or negative? If it is positive, what does it
mean? Or If it is negative, what does it mean?
cor(ToyotaPrices$Price, ToyotaPrices$KM)

## [1] -0.5699602

# Analysis:
# - The correlation between Price and KM is negative.
# - It is a weak correlation because as the number of kilometer decreases,
the price increases.
```