MultipleLinearRegression-Basics.R

Shraddha Somani

```
# Multiple Linear Regression
# Load the ToyotaPrices dataset
ToyotaPrices <- read.csv("D:/DADM/Assignment/ToyotaPrices.csv")
names(ToyotaPrices)
```

```
##  [1] "Id"               "Price"            "Age_08_04"
##  [4] "Mfg_Month"        "Mfg_Year"         "KM"
##  [7] "HP"               "Automatic"        "cc"
## [10] "Doors"            "Cylinders"        "Gears"
## [13] "Quarterly_Tax"    "Weight"           "Mfr_Guarantee"
## [16] "BOVAG_Guarantee"  "Guarantee_Period" "ABS"
## [19] "Airbag_1"         "Airbag_2"         "Airco"
## [22] "Automatic_airco"  "Boardcomputer"    "CD_Player"
## [25] "Central_Lock"     "Powered_Windows"  "Power_Steering"
## [28] "Radio"            "Mistlamps"        "Sport_Model"
## [31] "Backseat_Divider" "Metallic_Rim"     "Radio_cassette"
## [34] "Tow_Bar"
```

```
# Q1) Obtain the summary table of all the variables in myData_PKWT. From
inspection of the median and the mean, do any of the variables show skewness?
Which ones?
ToyotaPrices_PKWT = subset(ToyotaPrices, select = c(Price, KM, Weight,
Tow_Bar))
head(ToyotaPrices_PKWT)
```

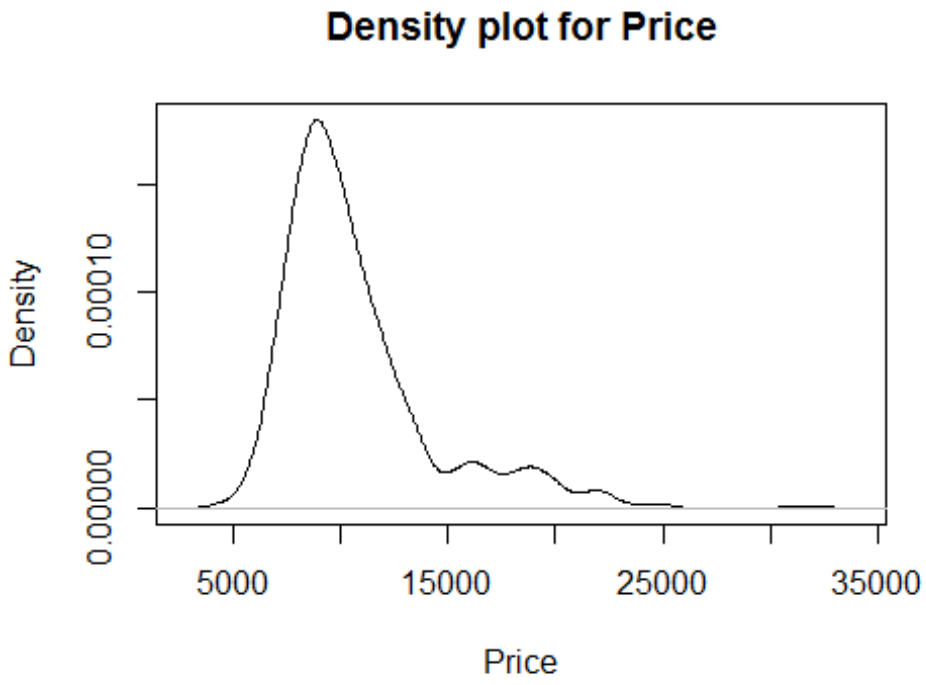
```
##   Price    KM Weight Tow_Bar
## 1 13500 46986   1165       0
## 2 13750 72937   1165       0
## 3 13950 41711   1165       0
## 4 14950 48000   1165       0
## 5 13750 38500   1170       0
## 6 12950 61000   1170       0
```

```
summary(ToyotaPrices_PKWT)
```

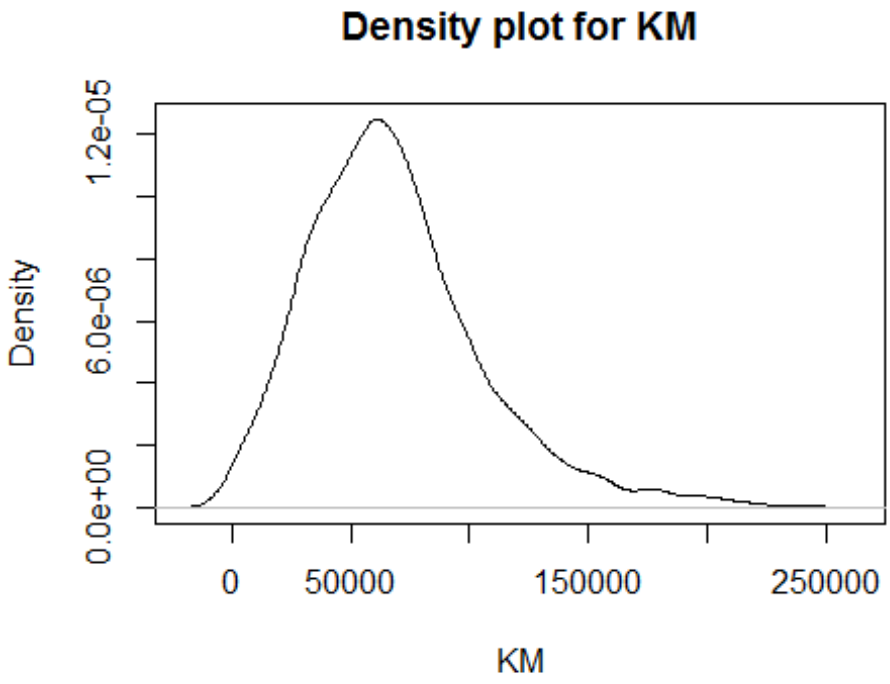
```
##      Price             KM             Weight        Tow_Bar
##  Min.   : 4350   Min.   :     1   Min.   :1000   Min.   :0.0000
##  1st Qu.: 8450   1st Qu.: 43000   1st Qu.:1040   1st Qu.:0.0000
##  Median : 9900   Median : 63390   Median :1070   Median :0.0000
##  Mean   :10731   Mean   : 68533   Mean   :1072   Mean   :0.2779
##  3rd Qu.:11950   3rd Qu.: 87021   3rd Qu.:1085   3rd Qu.:1.0000
##  Max.   :32500   Max.   :243000   Max.   :1615   Max.   :1.0000
```

```
# Answer:
# - If Mean and Median are unequal, skewness is present.
```

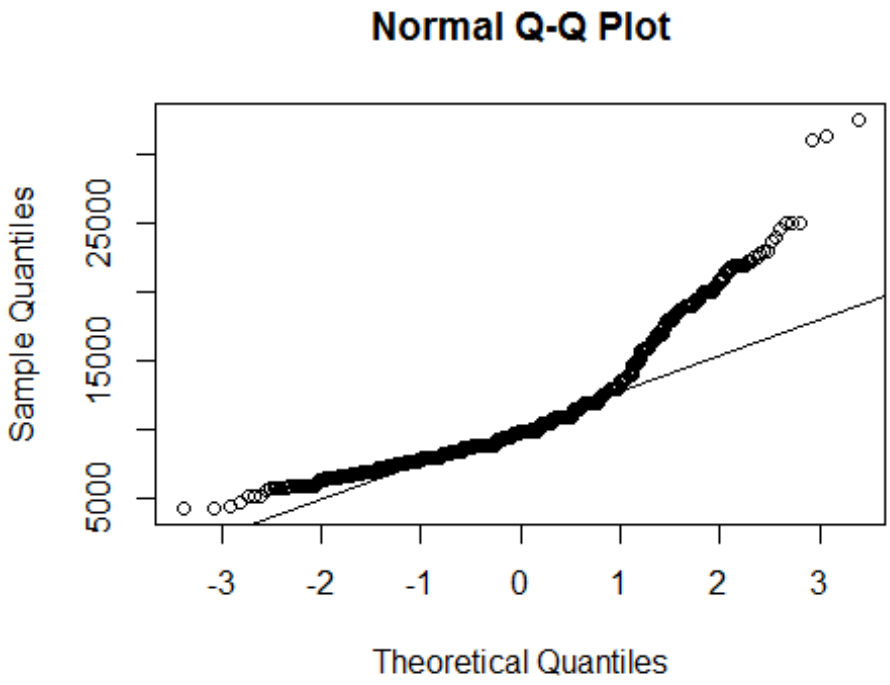
```r
# - Skewed - Price and KM
# - Not Skwed - Weight
# - In Tow_Bar, there is no concept of skewness because it has binary values.


# Q2) Obtain density plots and normal probability QQ-Plots of Price and KM.
Fron the patterns in these graphs, are any of the variables skewed? Which are
and which are not? Are any variables normally distributed? Which are and
which are not?
# Density Plot - Price
plot(density(ToyotaPrices_PKWT$Price), xlab = "Price", main = "Density plot
for Price")
```



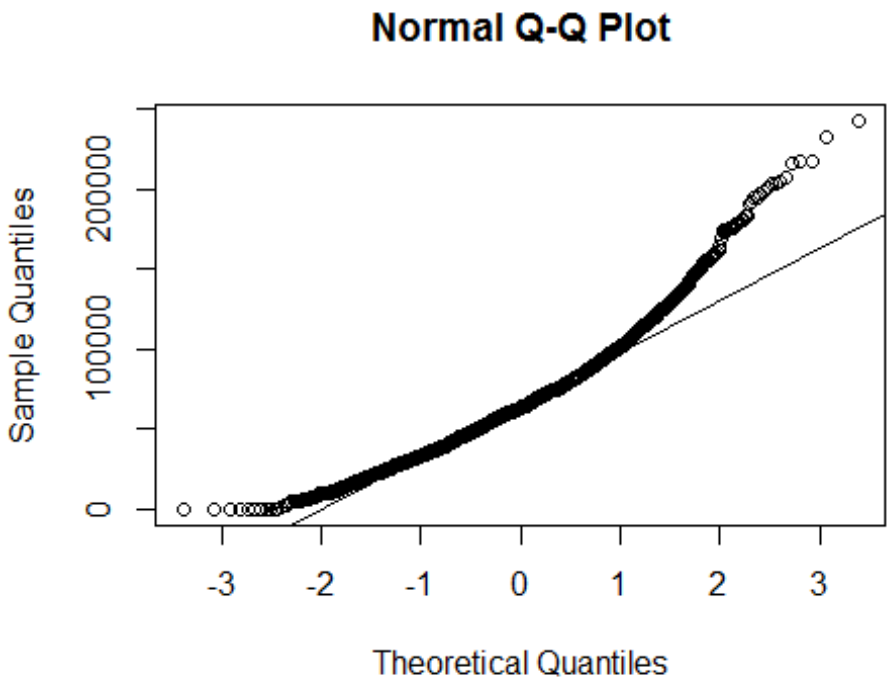

```r
# Density Plot - KM
plot(density(ToyotaPrices_PKWT$KM), xlab = "KM", main = "Density plot for
KM")
```




```
# Normal QQ Plot - Price
qqnorm(ToyotaPrices_PKWT$Price)
qqline(ToyotaPrices_PKWT$Price)
```



```r
# Normal QQ Plot - KM
qqnorm(ToyotaPrices_PKWT$KM)
qqline(ToyotaPrices_PKWT$KM)
```



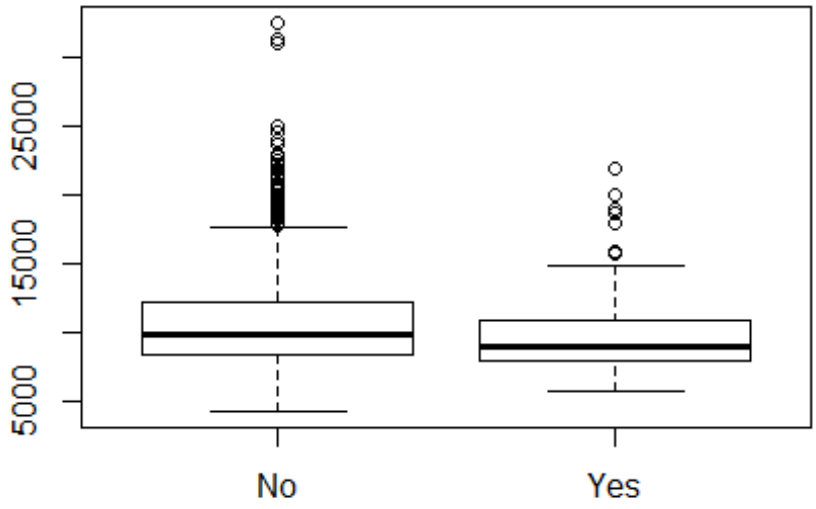

```r
# Answer:
# - Skewness is present in both Price and KM.
# - Price is not normally distributed. KM is normally distributed.


# Q3) Convert Tow_Bar to a factor with yes, no levels. Show the results.
ToyotaPrices_PKWT$Tow_Bar = factor(ToyotaPrices_PKWT$Tow_Bar)
summary(ToyotaPrices_PKWT$Tow_Bar)
```

```
##    0    1
## 1037  399
```

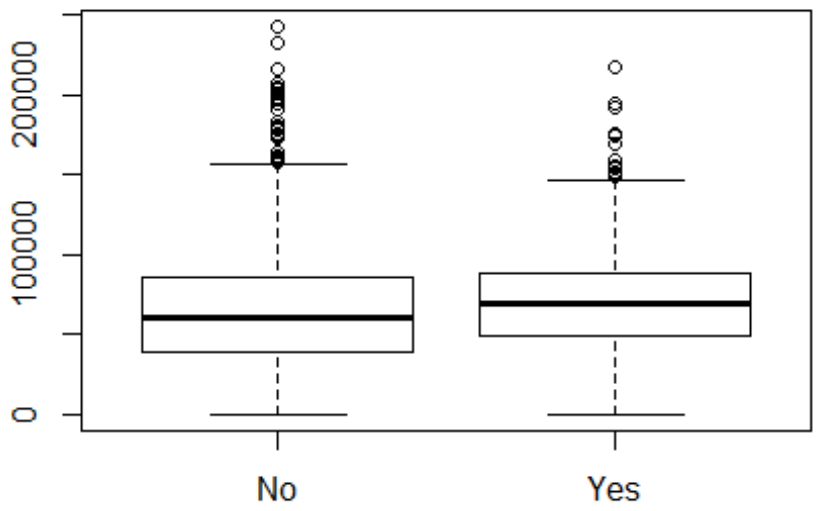
```r
levels(ToyotaPrices_PKWT$Tow_Bar) = c("No", "Yes")
summary(ToyotaPrices_PKWT)
```

```
##      Price             KM            Weight     Tow_Bar
##  Min.   : 4350   Min.   :     1   Min.   :1000   No :1037
##  1st Qu.: 8450   1st Qu.: 43000   1st Qu.:1040   Yes: 399
##  Median : 9900   Median : 63390   Median :1070
##  Mean   :10731   Mean   : 68533   Mean   :1072
##  3rd Qu.:11950   3rd Qu.: 87021   3rd Qu.:1085
##  Max.   :32500   Max.   :243000   Max.   :1615
```

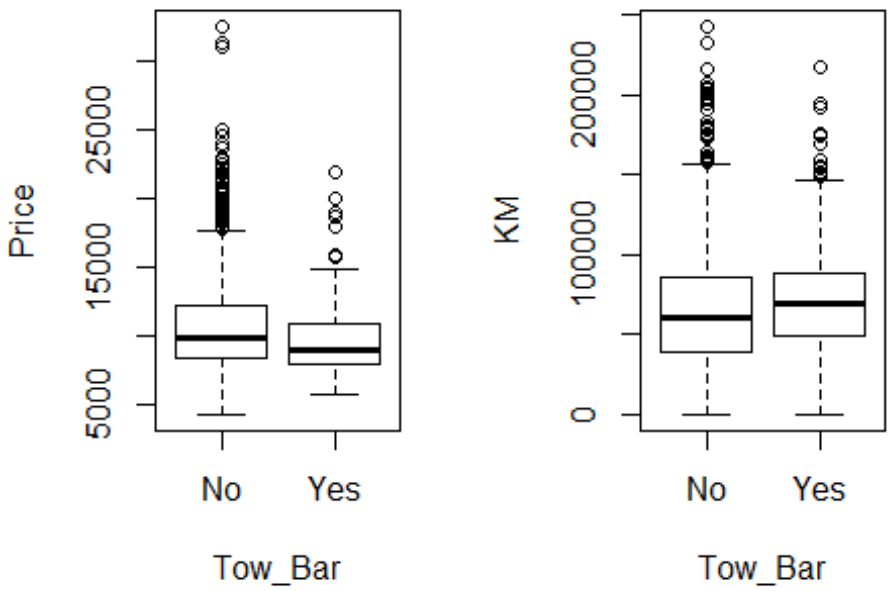
```r
# Q4) Obtain a boxplot of Price versus Tow_Bar. How are the two boxplots
different? Does Tow_Bar appear to predict Price?
boxplot(Price ~ Tow_Bar, data = ToyotaPrices_PKWT)
```



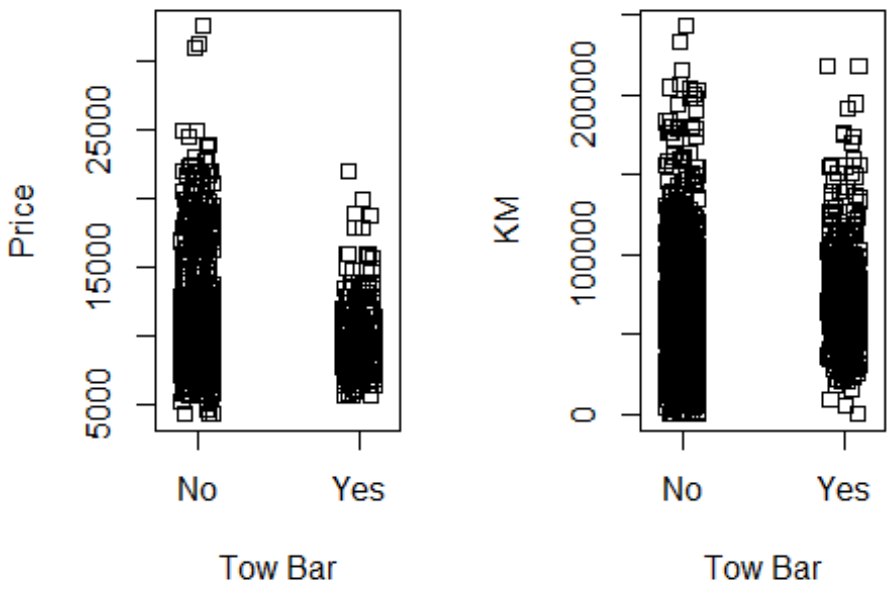

```r
# Answer:
# - The two boxplots are different in terms of outliers.
# - No, Tow_Bar does not appear to predict the price


# Q5) Obtain a boxplot of KM versus Tow_Bar. How are the two boxplots
different? Does Tow_Bar appear to predict KM?
boxplot(KM ~ Tow_Bar, data = ToyotaPrices_PKWT)
```



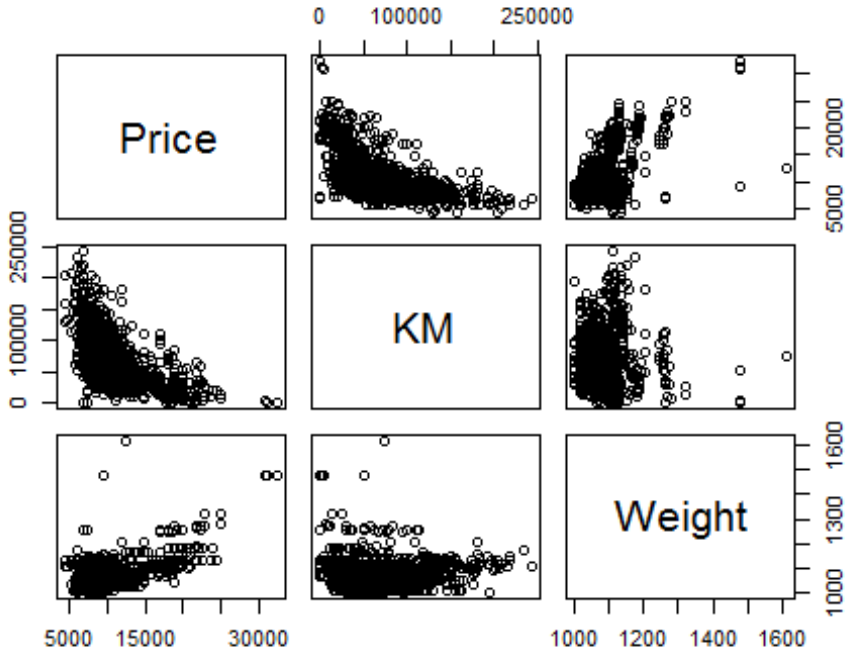

```
# Answer:
# - The two boxplots are different in terms of outliers.
# - No, Tow_Bar does not appear to predict KM.


# Q6) Can you explain why we the direction of prediction on price?
x = par(mfrow = c(1,2))
plot(Price ~ Tow_Bar, data = ToyotaPrices_PKWT)
plot(KM ~ Tow_Bar, data = ToyotaPrices_PKWT)
```



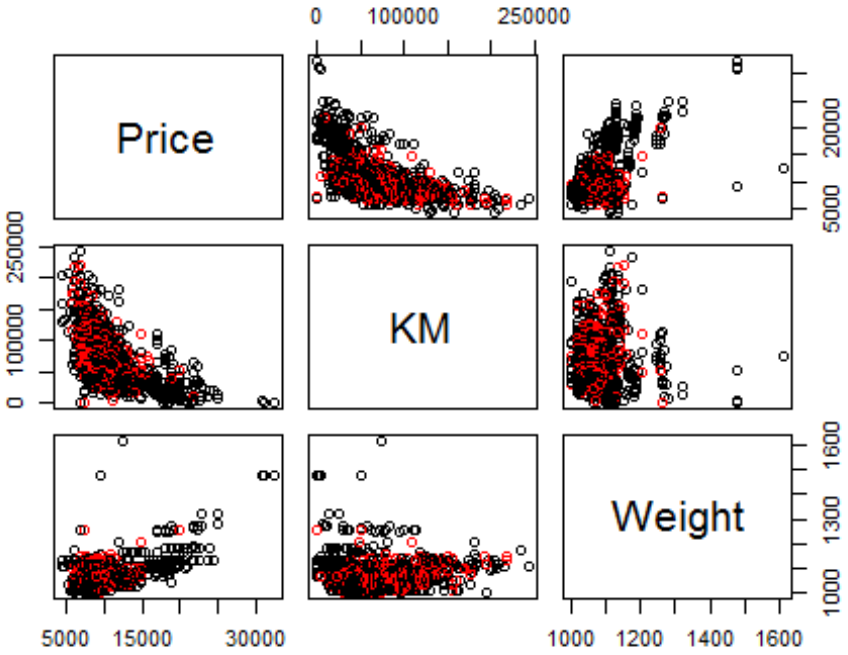

```
par(x)
y = par(mfrow = c(1,2))
stripchart(Price ~ Tow_Bar, data = ToyotaPrices_PKWT, method = "jitter",
vertical = TRUE, xlab="Tow Bar")
stripchart(KM ~ Tow_Bar, data = ToyotaPrices_PKWT, method = "jitter",
vertical = TRUE, xlab="Tow Bar")
```




```r
par(y)
# Answer:
# - We can see that the price of the car is less when the tow_bar is absent.


# Q7) Obtain a scatterplot matrix of Price, KM and Weight. Discuss the plot.
What sort of function would likely fit the expected value function of Price?
Does KM and Weight appear to be redundant? Are their any ouliers in the
plots; if so what are they?
pairs(~ Price + KM + Weight, data = ToyotaPrices_PKWT)
```

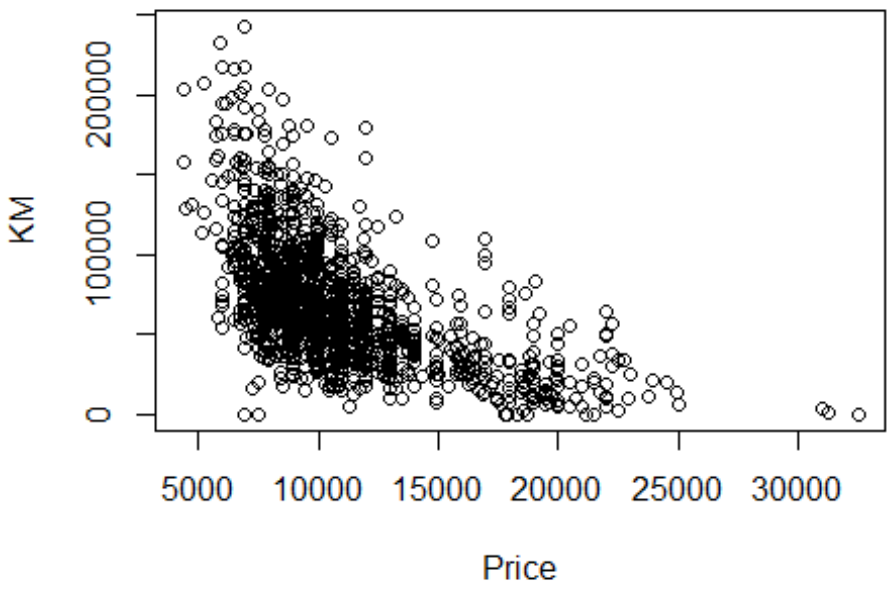



```
# Answer:
# - From the graph it is clear that, the price of car is less when the number
of KM travelled by the car is more.
# - There is no clear relationship present bewteen Price and Weight.
# - Yes, KM and Weight appear to be redundant.
# - Yes, there are outliers present in all the three variables.


# Q8) Obtain a scatterplot matrix of Price, KM, and Weight with the points
colored by the levels of Tow_Bar. Discuss the plot. Does it appear that the
relation between Price and KM is the same or different for cars with or
without a tow bar? I.e., are there any clear relationship visible that appear
to be different for groups of cars with or without a tow bar?
pairs(~ Price + KM + Weight, data = ToyotaPrices_PKWT, col =
ToyotaPrices_PKWT$Tow_Bar)
```




```
# Answer:
# - It appears that the relation between Price and KM is the same for cars
with or without a tow bar.
# - There is no clear relationship visible.


# Q9) Fit Price against KM and Weight and Tow_Bar.
fit = lm(Price ~ KM + Weight + Tow_Bar, ToyotaPrices_PKWT)
summary(fit)

##
## Call:
## lm(formula = Price ~ KM + Weight + Tow_Bar, data = ToyotaPrices_PKWT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19077.1  -1248.8    -38.2   1230.7   8795.0
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.678e+04  1.168e+03 -22.930  < 2e-16 ***
## KM          -5.288e-02  1.515e-03 -34.910  < 2e-16 ***
## Weight       3.853e+01  1.079e+00  35.726  < 2e-16 ***
## Tow_BarYes  -6.835e+02  1.271e+02  -5.378  8.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2144 on 1432 degrees of freedom
## Multiple R-squared:  0.6513, Adjusted R-squared:  0.6505
## F-statistic: 891.4 on 3 and 1432 DF,  p-value: < 2.2e-16

# Q10) Discuss the residual five-number summary. Do the residuals appear to
be skewed?
summary(fit$residuals)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -19080.00  -1249.00    -38.23      0.00   1231.00   8795.00

# Answer:
# - The residuals appear to be a non-parametric summary of their
distribution.
# - The residuals appears to be skewed little on the left


# Q11) Discuss the intercept coefficient. What does it tell us?
coef(fit)

##   (Intercept)            KM        Weight    Tow_BarYes
## -2.677787e+04 -5.288276e-02  3.853090e+01 -6.835050e+02

# Answer:
# - Negative coefficient indicates that they are inversely proptional to each
other.
# - KM is negatively corelated i.e with increase in KM there is a  decrease
in Price


# Q12) Discuss the signs of the slope coeffcients. Do they make sense?
# Answer:
# - A positive sign of the correlation coefficient indicates that as the
value of one variable increases, the value of the other variable also
increases;
# - A negative correlation coefficient indicates that as the value of one
variable increases, the other decreases


# Q13) How does the price of the car change as KM increases? Does the price
go up or down? How much? Does this make sense?
plot(ToyotaPrices_PKWT$Price, ToyotaPrices_PKWT$KM, xlab = "Price", ylab =
"KM")
```




```
# Answer:
# - Greater the value of KM, less is the Price of car.


# Q14) How does the price of the car change as Weight increases? Does the
price go up or down? How much? Does this make sense?
plot(ToyotaPrices_PKWT$Price, ToyotaPrices_PKWT$Weight, xlab = "Price", ylab
= "Weight")
# Answer:
# - As the weight increases there is no change in the price.

# Q15) What is the Euro-price difference between Toyotas with and without the
Tow_Bar automobile accessory for cars with the same KM and Weight? This does
this value make sense?
# Answer:
# - There is not much price difference of automobiles with and without
Tow_bar with the same KM and weight.
# - The value does not make sense.


# Q16) Obtain R^2. Is is a measure of the Goodness-of-Fit of the model. Is
this value indicate a good fitting model, or not?
deviance = deviance(fit)
deviance

## [1] 6583099567
```

```r
y = ToyotaPrices_PKWT$Price
totalss = sum((y-mean(y))^2)
totalss
```

```
## [1] 18877241464
```

```r
1 - deviance/totalss
```

```
## [1] 0.6512679
```

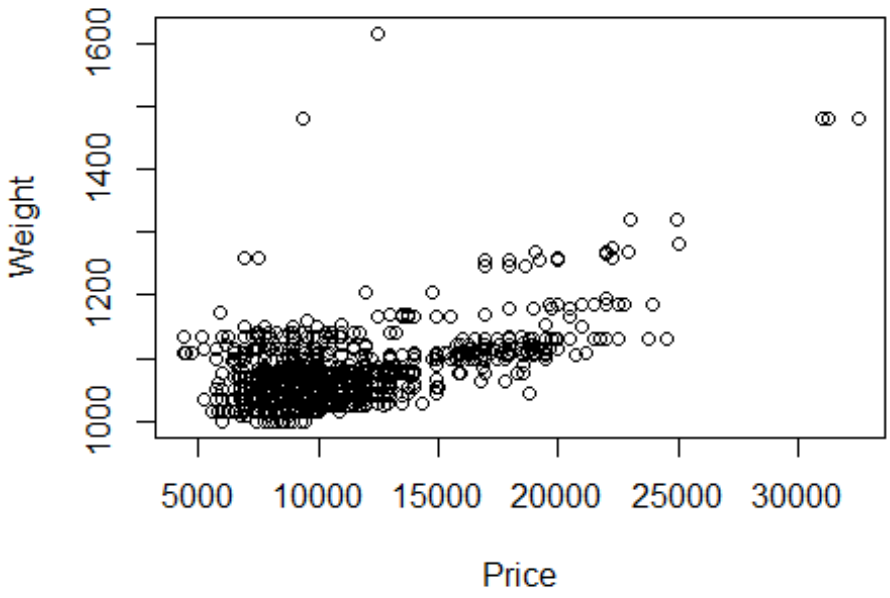
```r
summary(fit)$r.square
```

```
## [1] 0.6512679
```

```r
# Answer:
# - This value indicates good fitting model.


# Q17) R^2 indicates the correlation between the Price observations and their
fitted values. Obtain r, the Pearson correlation between  Price and the
Fitted Values, then square it. Verify that this value is equal to the R^2
value found in the model summary.
c(summary(fit)$r.square, cor(fitted.values(fit), ToyotaPrices_PKWT$Price)^2)
```

```
## [1] 0.6512679 0.6512679
```

```r
# Answer:
# - Both r and R^2 have the same value.


# Q18) Obtain the Fit Plot, with a 45 degree diagonal line. Do the fitted
values from this model predict the actual prices well?
library(ggplot2)
```




```
qplot(fit$fitted.value, Price, data=ToyotaPrices_PKWT) +geom_abline(intercept
= 0, slope = 1, color="green") +ggtitle("Fit Plot")
```