Cars Predicted value

Prats Jamison

2016-11-15

install.packages(knitr) library(knitr)

Calculating with R

Before we start we will load necessary libraries

```
library(knitr)
library(leaps)#Exhaustive search for the best of variables in x for predicting y
library(e1071)#skewness and kurotosis
library(moments)
```

```
##
## Attaching package: 'moments'
```

```
## The following objects are masked from 'package:e1071':
##
##     kurtosis, moment, skewness
```

```
library(broom)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.3.2
```

```
## ------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.3.2
```

```
library(plm)
```

```
## Warning: package 'plm' was built under R version 3.3.2
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'plm'
```

```
## The following object is masked from 'package:dplyr':
##
##     between
```

```
library(lattice)
library(ggplot2)
```

To run the analysis , we need to load the data in to R

```
#Read Data
Toyotacor <- read.csv("ToyotaCorolla.csv")
attach(Toyotacor)
summary(Toyotacor)
```

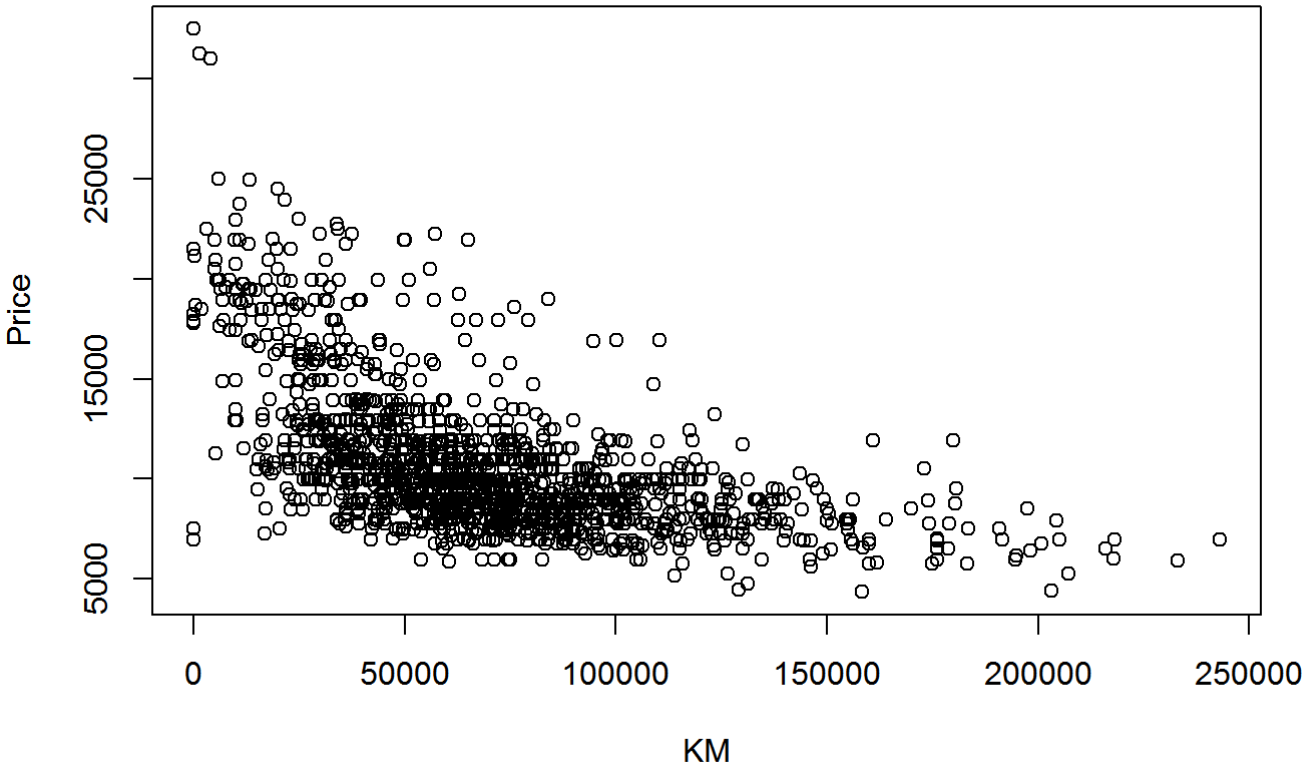
```
##      Price            Age              KM            FuelType
##  Min.   : 4350   Min.   : 1.00   Min.   :     1   Min.   :0.0000
##  1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000   1st Qu.:1.0000
##  Median : 9900   Median :61.00   Median : 63390   Median :1.0000
##  Mean   :10731   Mean   :55.95   Mean   : 68533   Mean   :0.9039
##  3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021   3rd Qu.:1.0000
##  Max.   :32500   Max.   :80.00   Max.   :243000   Max.   :2.0000
##        HP           MetColor        Automatic            CC
##  Min.   : 69.0   Min.   :0.0000   Min.   :0.00000   Min.   :1300
##  1st Qu.: 90.0   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1400
##  Median :110.0   Median :1.0000   Median :0.00000   Median :1600
##  Mean   :101.5   Mean   :0.6748   Mean   :0.05571   Mean   :1567
##  3rd Qu.:110.0   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1600
##  Max.   :192.0   Max.   :1.0000   Max.   :1.00000   Max.   :2000
##      Doors           Weight
##  Min.   :2.000   Min.   :1000
##  1st Qu.:3.000   1st Qu.:1040
##  Median :4.000   Median :1070
##  Mean   :4.033   Mean   :1072
##  3rd Qu.:5.000   3rd Qu.:1085
##  Max.   :5.000   Max.   :1615
```

Familiarize yourself with th data

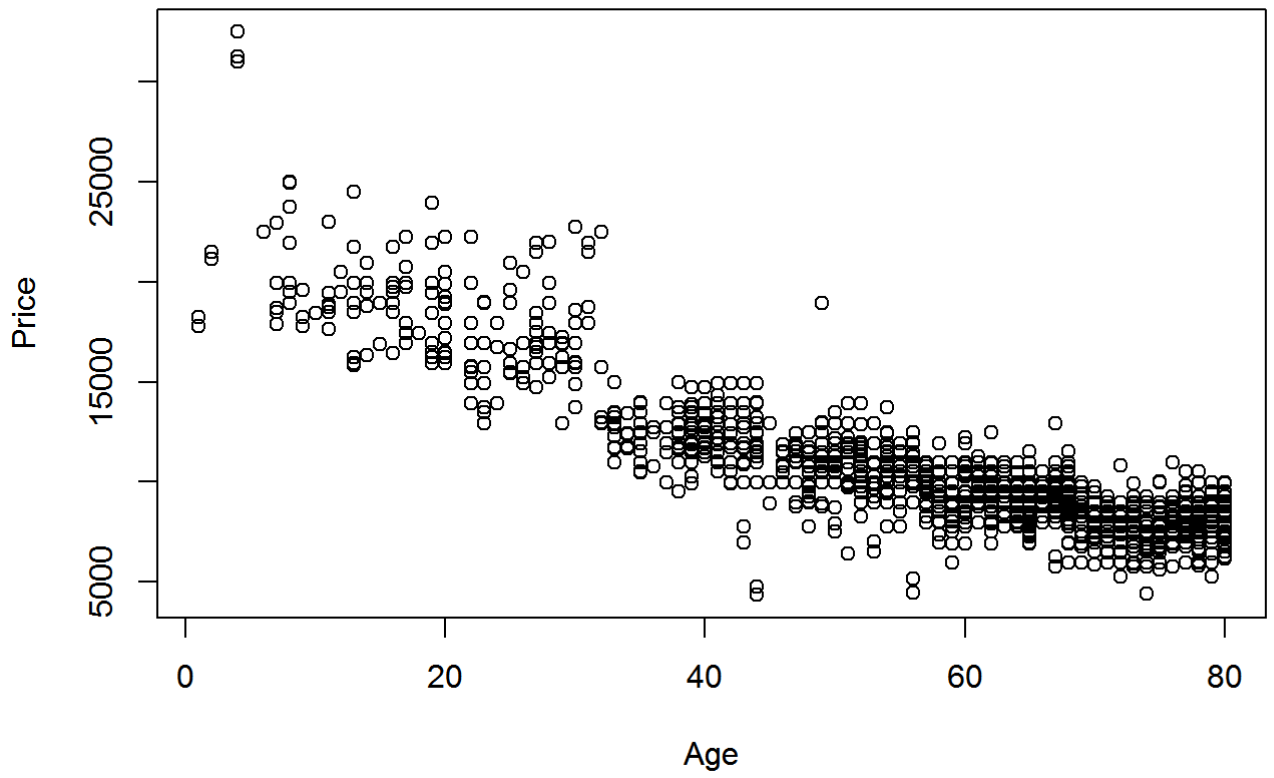
```
# Variable definition
# MPG - Miles per galon
# GPM - Galons per miles
# WT - Weight
# DIS - Dicplacement
# NC - Number of cylinders
# HP - Horsepower
# ACC - acceleration (0-60mph) in seconds
# ET - V-type engine (0) or straight (1)

head(Toyotacor)
```

```
##   Price Age    KM FuelType HP MetColor Automatic   CC Doors Weight
## 1 13500  23 46986        0 90        1         0 2000     3   1165
## 2 13750  23 72937        0 90        1         0 2000     3   1165
## 3 13950  24 41711        0 90        1         0 2000     3   1165
## 4 14950  26 48000        0 90        0         0 2000     3   1165
## 5 13750  30 38500        0 90        0         0 2000     3   1170
## 6 12950  32 61000        0 90        0         0 2000     3   1170
```

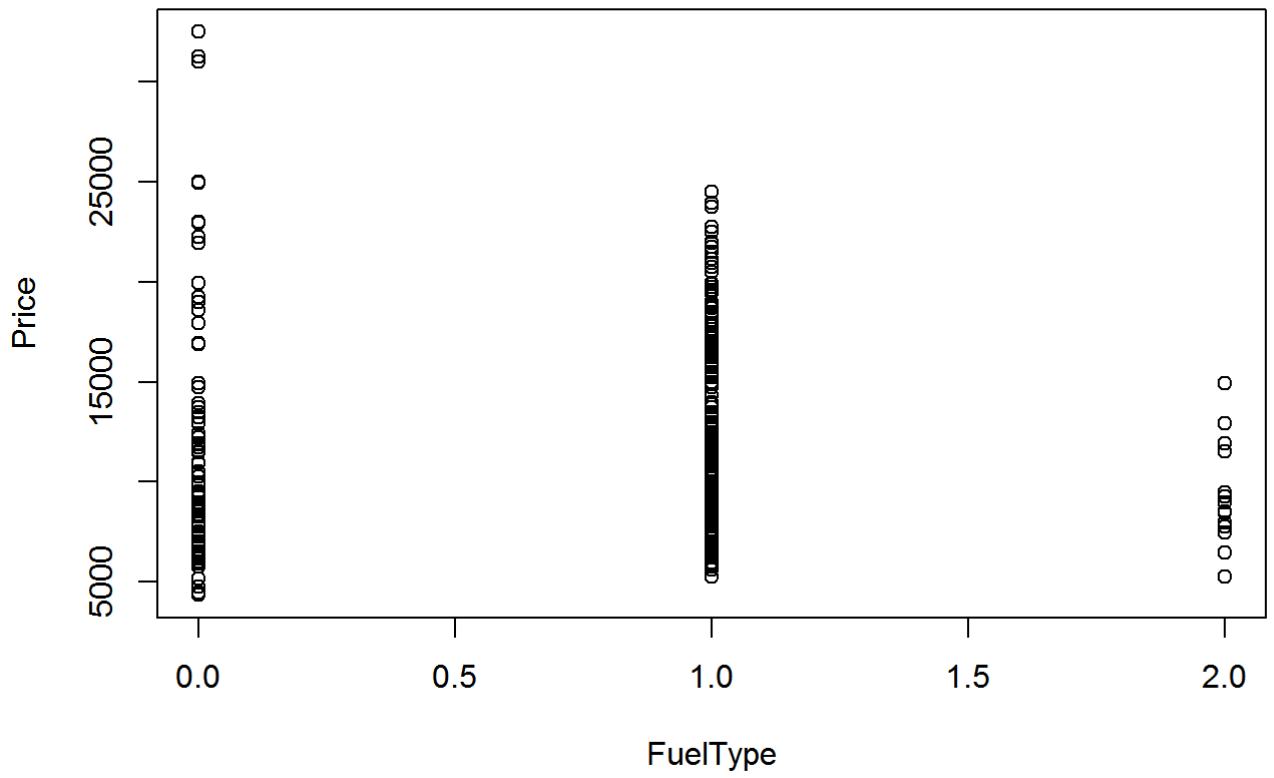
```
 # Become familiar with the data
plot(Price ~ KM, data = Toyotacor)
```



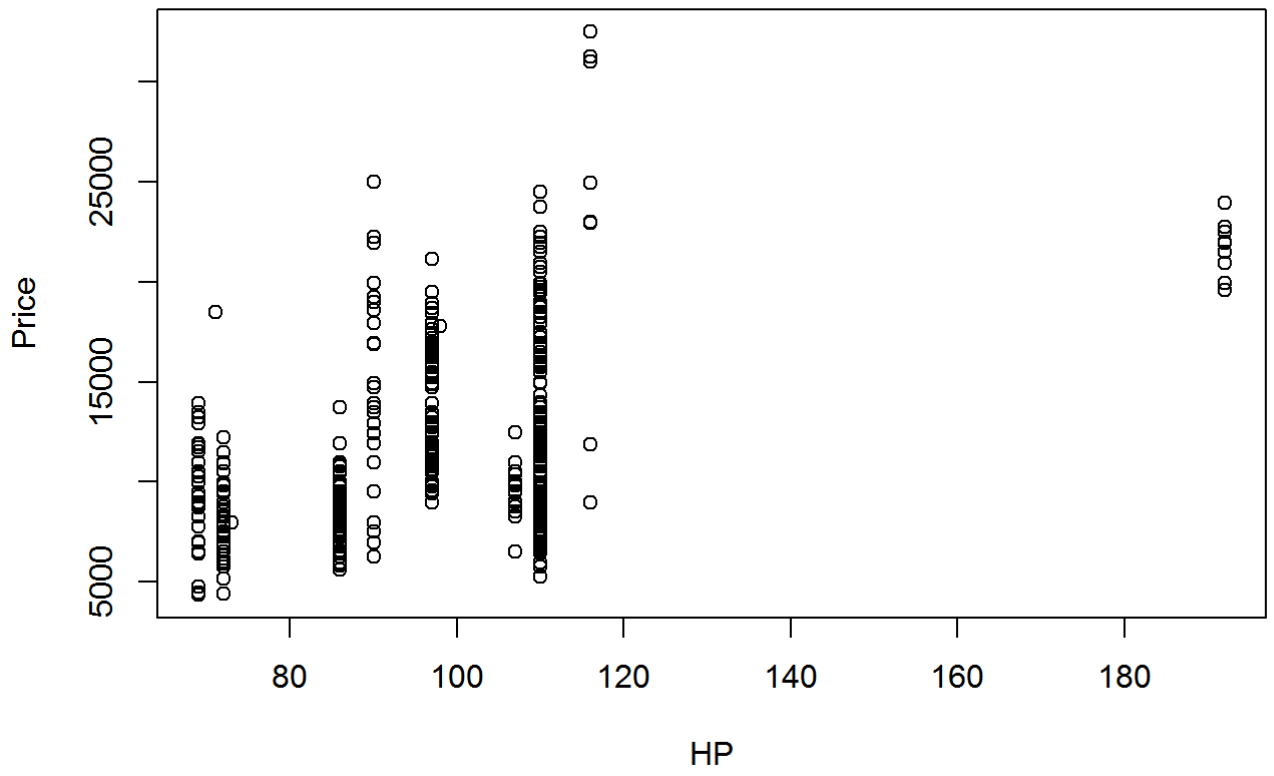

```
plot(Price ~ Age, data = Toyotacor)
```



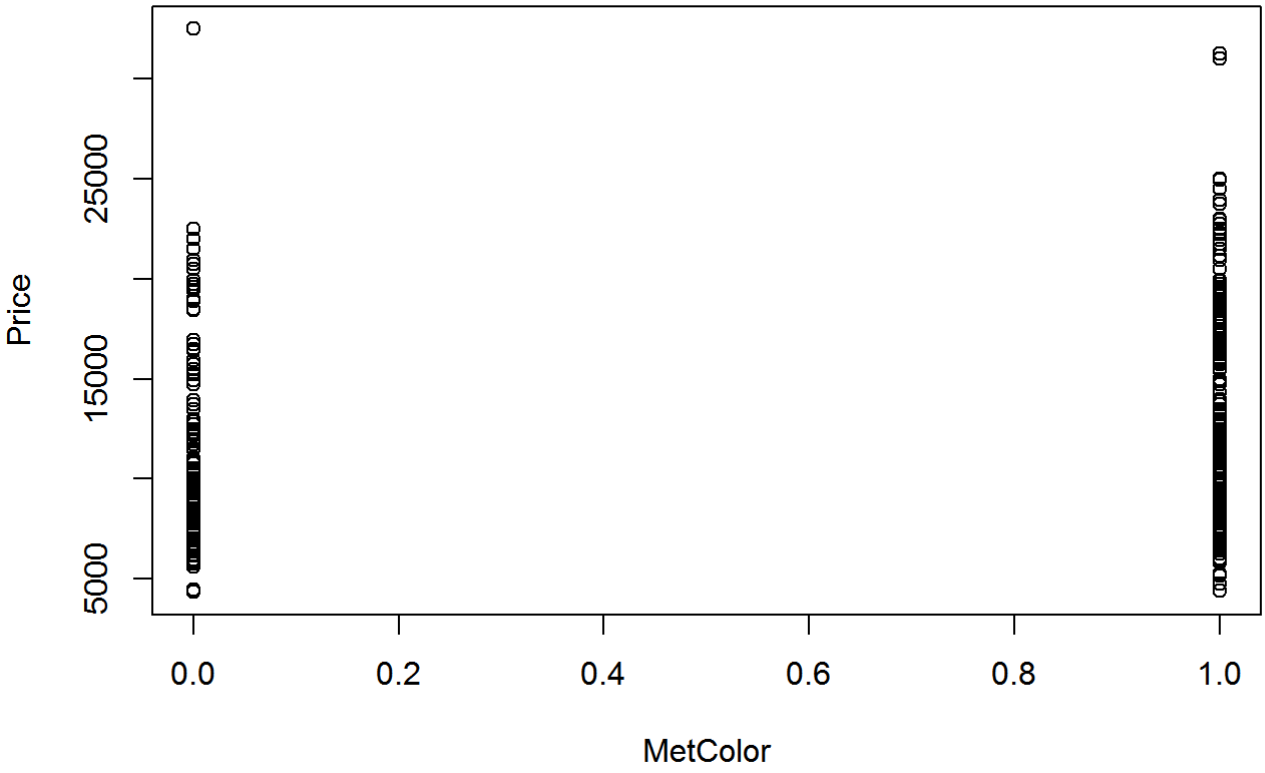

```
plot(Price ~ FuelType, data = Toyotacor)
```



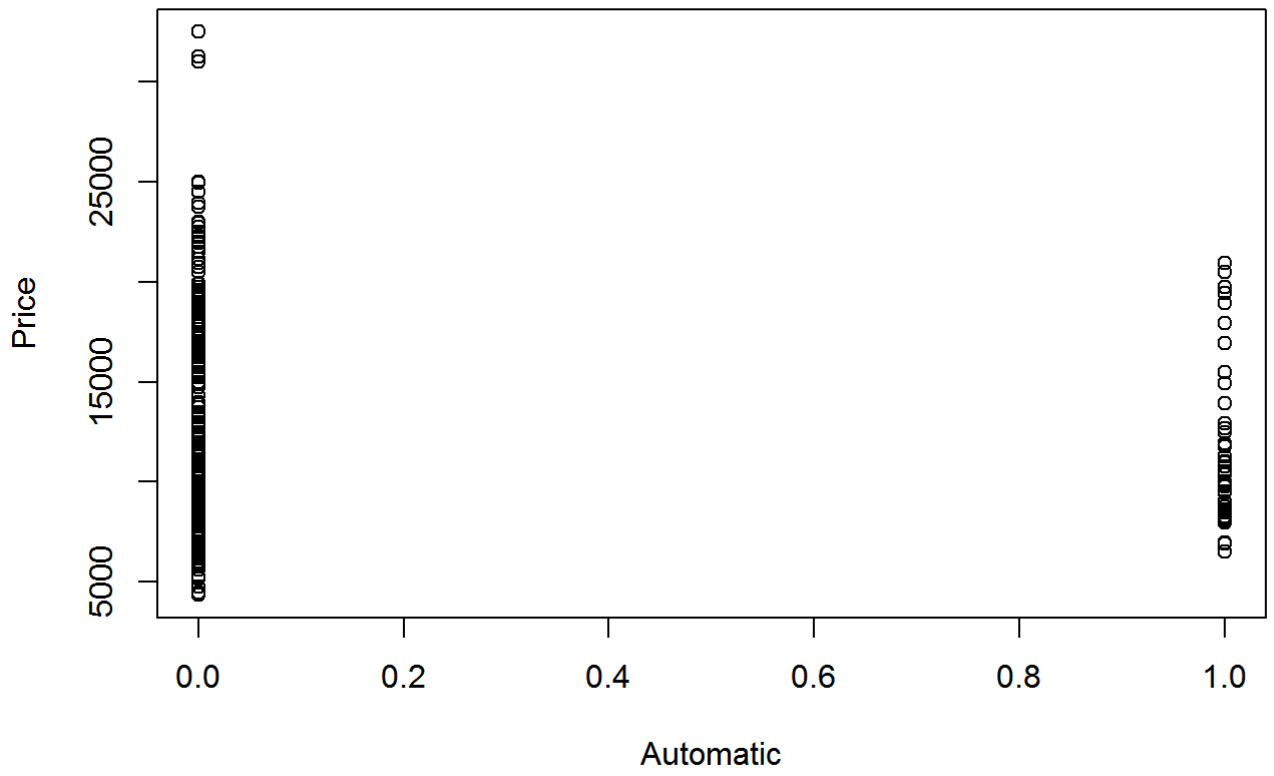

```
plot(Price ~ HP, data = Toyotacor)
```



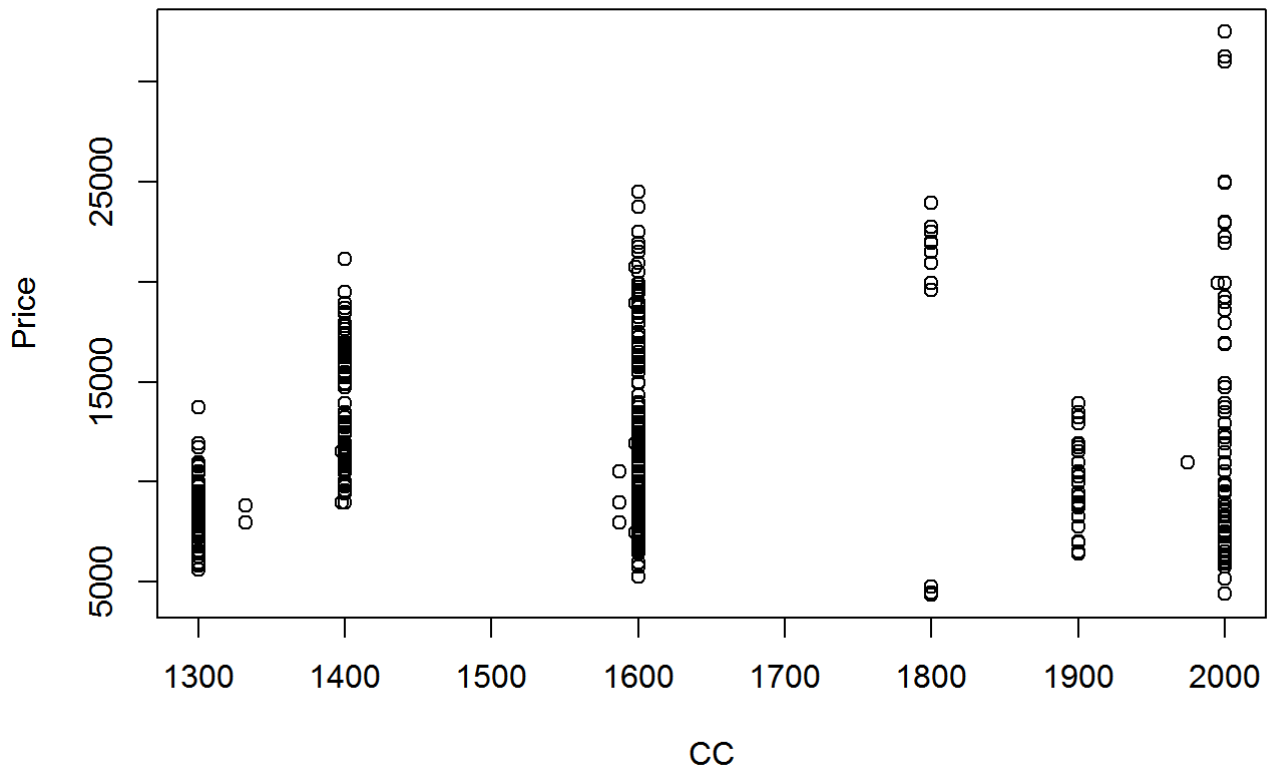

```
plot(Price ~ MetColor, data = Toyotacor)
```



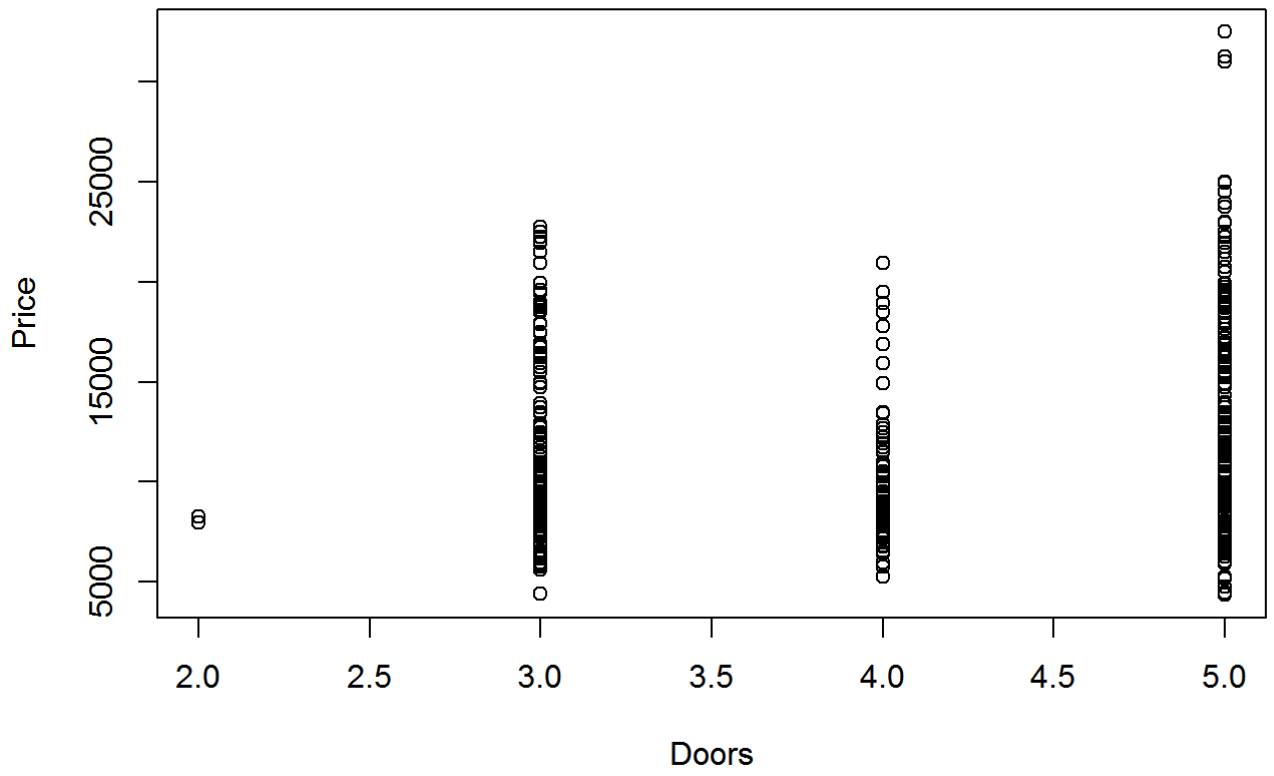

```
plot(Price ~ Automatic, data = Toyotacor)
```



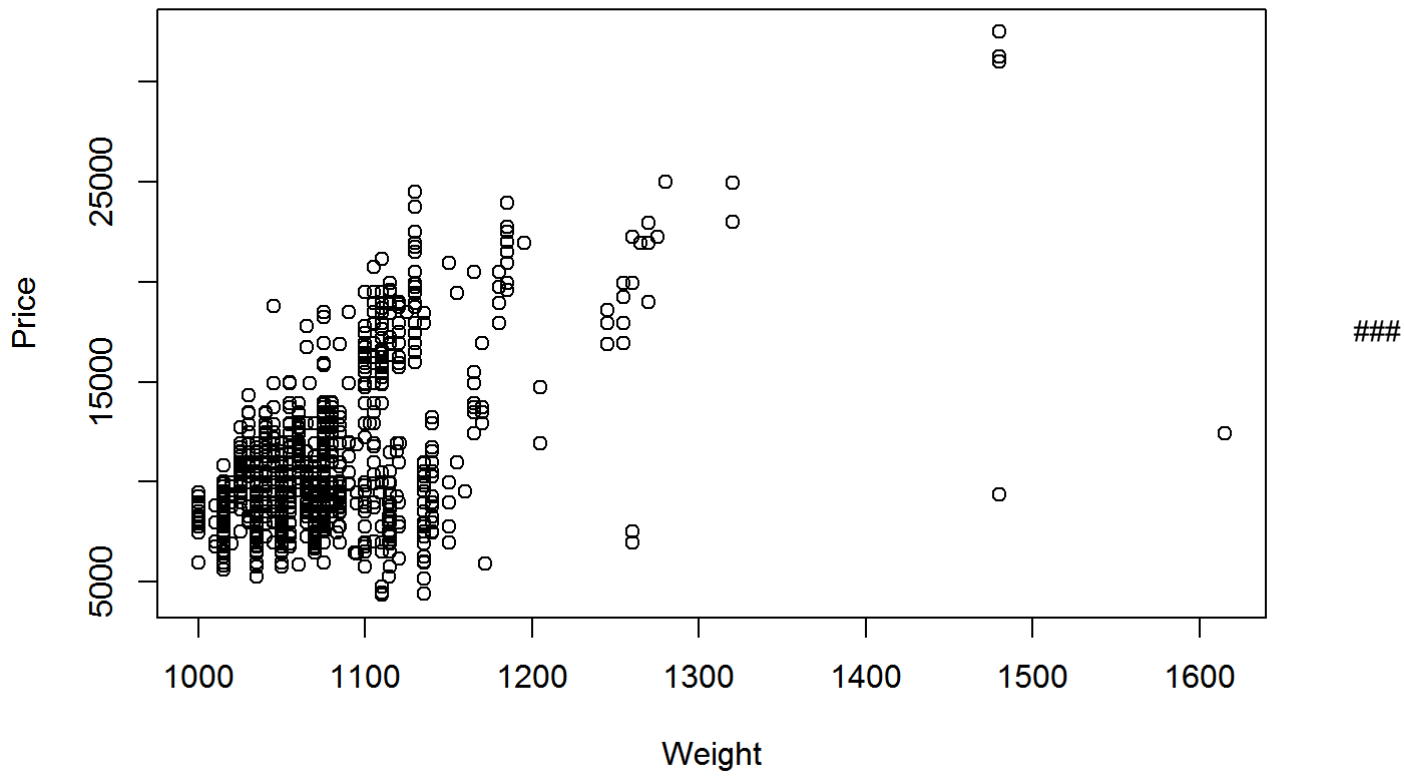

```
plot(Price ~ CC, data = Toyotacor)
```




```
plot(Price ~ Doors, data = Toyotacor)
```




```
plot(Price ~ Weight, data = Toyotacor)
```



###

Prepare for analysis

Descriptive statistics are statistics that quantitatively describe or summarise features of a collection of information

Linear Regression

Linear regression models are probably the most common used technique in data and buisiness analytics. They can be very powerful but one has to remember their limitations and constrains. The most important limitation is, that they should only be used to predict values within the range of the test data set. Here are are predicting the values for Toyota Coralla.

```
options(max.print = 10)
# show results of first Analysis
Toyotacor.m1 <- lm(Price ~ ., data = Toyotacor)
summary(Toyotacor.m1) # Show results of the first model
```

```
##
## Call:
## lm(formula = Price ~ ., data = Toyotacor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11209.6   -748.0      8.9    735.9   6374.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.358e+03  1.154e+03  -2.043   0.0412 *
## Age         -1.226e+02  2.589e+00 -47.336  < 2e-16 ***
##  [ reached getOption("max.print") -- omitted 8 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1317 on 1426 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8681
## F-statistic:  1051 on 9 and 1426 DF,  p-value: < 2.2e-16
```

```
#Confident Interval
confint(Toyotacor.m1)
```

```
##                     2.5 %        97.5 %
## (Intercept) -4.622093e+03   -94.3881168
## Age         -1.276522e+02  -117.4931773
## KM          -1.818931e-02    -0.0131468
## FuelType    -2.044477e+03 -1064.7406133
## HP           4.477963e+01    60.8020798
##  [ reached getOption("max.print") -- omitted 5 rows ]
```

```
Toyotacor.m1.confint <- confint(Toyotacor.m1)
## Check with a correlation matrix if predictor variables are themselves related
Toyotacor.D= as.data.frame(matrix(Toyotacor))[1,]
TyCo.col <- cor(data.frame(lapply(Toyotacor.D, rank)))
head(print(TyCo.col))
```

```
##                                                                 c.1217.5..1228..1239.5..1261.
5..1228..1183..1315..1356..1409..
## c.1217.5..1228..1239.5..1261.5..1228..1183..1315..1356..1409..
                               1
```

```
##                                                                 c.1217.5..1228..1239.5..1261.
5..1228..1183..1315..1356..1409..
## c.1217.5..1228..1239.5..1261.5..1228..1183..1315..1356..1409..
                               1
```

Calculate regression - Model 1

```
options(max.print = 10)
#calculate regression - Model 1
x <- Toyotacor[, 2:10] # independent variable
head(x)
```

```
##   Age    KM FuelType HP MetColor Automatic   CC Doors Weight
## 1  23 46986        0 90        1         0 2000     3   1165
##  [ reached getOption("max.print") -- omitted 5 rows ]
```

```
y <- Toyotacor[,1] # dependent variable
head(y)
```

```
## [1] 13500 13750 13950 14950 13750 12950
```

```
#model selection
Toyotacor.out <- summary(regsubsets(x,y, nbest = 2, nvmax = ncol(x)))
Toyotacor.regtab <- cbind(Toyotacor.out$which, Toyotacor.out$rsq, Toyotacor.out$adjr2, Toyotaco
r.out$cp)

colnames(Toyotacor.regtab) <- c("(Intercept)", "Age", "KM", "FuelType", "HP", "MetColor", "Trans
mission", "CC" , "Doors", "Weight", "R-Sq", "R-Sq(adj)", "Cp")
print(Toyotacor.regtab) # pValue is < 0.05, so we reject null hypothesis
```

```
##   (Intercept) Age KM FuelType HP MetColor Transmission CC Doors Weight
##        R-Sq R-Sq(adj)          Cp
##  [ reached getOption("max.print") -- omitted 17 rows ]
```

```
head.matrix(Toyotacor.regtab)
```

```
##   (Intercept) Age KM FuelType HP MetColor Transmission CC Doors Weight
##        R-Sq R-Sq(adj)        Cp
##  [ reached getOption("max.print") -- omitted 6 rows ]
```

Regression analysis with the given variable 7 expect km and windows and also find the car cost using these variables

```
options(max.print = 10)
#create second model
Toyotacor.m2 <- lm(Price~ Age + FuelType+HP+MetColor+ Automatic+ CC + Doors, data = Toyotacor) #
just with weight
Toyotacor.m2.summary <- summary(Toyotacor.m2)
head(print(Toyotacor.m2.summary))
```

```
##
## Call:
## lm(formula = Price ~ Age + FuelType + HP + MetColor + Automatic +
##     CC + Doors, data = Toyotacor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7743.0  -917.8    -2.5   845.8 10889.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18013.1737   653.1245  27.580  < 2e-16 ***
## Age          -158.7395     2.2938 -69.203  < 2e-16 ***
##  [ reached getOption("max.print") -- omitted 6 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1534 on 1428 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8211
## F-statistic: 941.8 on 7 and 1428 DF,  p-value: < 2.2e-16
```

```
## $call
## lm(formula = Price ~ Age + FuelType + HP + MetColor + Automatic +
##     CC + Doors, data = Toyotacor)
##
## $terms
## Price ~ Age + FuelType + HP + MetColor + Automatic + CC + Doors
## attr(,"variables")
## list(Price, Age, FuelType, HP, MetColor, Automatic, CC, Doors)
## attr(,"factors")
##           Age FuelType HP MetColor Automatic CC Doors
## Price       0        0  0        0         0  0     0
##  [ reached getOption("max.print") -- omitted 7 rows ]
## attr(,"term.labels")
## [1] "Age"       "FuelType"  "HP"        "MetColor"  "Automatic" "CC"
## [7] "Doors"
## attr(,"order")
## [1] 1 1 1 1 1 1 1
## attr(,"intercept")
## [1] 1
## attr(,"response")
## [1] 1
## attr(,".Environment")
## <environment: R_GlobalEnv>
## attr(,"predvars")
## list(Price, Age, FuelType, HP, MetColor, Automatic, CC, Doors)
## attr(,"dataClasses")
##     Price       Age  FuelType        HP  MetColor Automatic        CC
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##     Doors
## "numeric"
##
## $residuals
##          1          2          3          4          5          6
## -2706.5038 -2456.5038 -2097.7644  -710.7171 -1275.7593 -1758.2803
##          7          8          9         10
##  1328.4540  3504.6724   203.2553 -1803.2219
##  [ reached getOption("max.print") -- omitted 1426 entries ]
##
## $coefficients
##                 Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept) 18013.173675 653.1245226  27.5799990 1.388601e-134
## Age          -158.739466   2.2938141 -69.2032842  0.000000e+00
##  [ reached getOption("max.print") -- omitted 6 rows ]
##
## $aliased
## (Intercept)         Age    FuelType          HP    MetColor   Automatic
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##          CC       Doors
##       FALSE       FALSE
##
## $sigma
## [1] 1534.134
```

```
#Model 2 Conf Intervals
Toyotacor.m2.confint <- confint(Toyotacor.m2)
head(print(Toyotacor.m2.confint))
```

```
##                    2.5 %       97.5 %
## (Intercept) 16731.987223 19294.360128
## Age          -163.239073  -154.239859
## FuelType    -3535.549866 -2431.821782
## HP            71.408754    88.950692
## MetColor     -101.680116   240.816723
##  [ reached getOption("max.print") -- omitted 3 rows ]
```

```
##                   2.5 %      97.5 %
## (Intercept) 16731.98722 19294.36013
## Age          -163.23907  -154.23986
## FuelType    -3535.54987 -2431.82178
## HP            71.40875    88.95069
## MetColor     -101.68012   240.81672
##  [ reached getOption("max.print") -- omitted 1 row ]
```

```
Toyotacor.m2 <- lm(Price~ Age + FuelType+HP+MetColor+ Automatic+ CC + Doors, data = Toyotacor)
head(Toyotacor.m2)
```

```
## $coefficients
##  (Intercept)          Age     FuelType           HP     MetColor
## 18013.173675  -158.739466 -2983.685824    80.179723    69.568303
##    Automatic           CC        Doors
##  1059.890897    -3.000606   186.601905
##
## $residuals
##          1          2          3          4          5          6
## -2706.5038 -2456.5038 -2097.7644  -710.7171 -1275.7593 -1758.2803
##          7          8          9         10
##  1328.4540  3504.6724   203.2553 -1803.2219
##  [ reached getOption("max.print") -- omitted 1426 entries ]
##
## $effects
##  (Intercept)          Age     FuelType           HP     MetColor
## -406640.2022 -120438.6902    2387.9590  -27864.2806     985.7555
##    Automatic           CC        Doors
##    9004.1549   10173.4959    6572.8321     492.1586   -1621.2331
##  [ reached getOption("max.print") -- omitted 1426 entries ]
##
## $rank
## [1] 8
##
## $fitted.values
##        1        2        3        4        5        6        7        8
## 16206.50 16206.50 16047.76 15660.72 15025.76 14708.28 15571.55 15095.33
##        9       10
## 21296.74 14753.22
##  [ reached getOption("max.print") -- omitted 1426 entries ]
##
## $assign
## [1] 0 1 2 3 4 5 6 7
```

```
#Assigning values
given.Toyotacor <- data.frame(Age=12, FuelType=1, HP=185, MetColor=1, Automatic=0, CC=2000, Door
s=4)
predicted.price <- predict(Toyotacor.m2,given.Toyotacor)
as.data.frame(print(predicted.price))# Predicted Car Value
```

```
##        1
## 22772.63
```

```
##   print(predicted.price)
## 1              22772.63
```

```
options(max.print = 10)
#predicted Values/ Residuals
Toyotaco_hat <- fitted(Toyotacor.m1)# predicted values
print(Toyotaco_hat)
```

```
##        1        2        3        4        5        6        7        8
## 16382.87 15976.27 16342.95 15943.64 15707.16 15109.49 16825.94 16751.58
##        9       10
## 21203.72 13925.19
##  [ reached getOption("max.print") -- omitted 1426 entries ]
```

```
Toyota_resid <- residuals(Toyotacor.m1) # residuals
print(Toyota_resid)
```

```
##           1           2           3           4           5           6
## -2882.87323 -2226.27163 -2392.94951  -993.63561 -1957.16437 -2159.48783
##           7           8           9          10
##    74.05566  1848.42081   296.28065  -975.19203
##  [ reached getOption("max.print") -- omitted 1426 entries ]
```

So now that we have two models, the question would be, which one is better? In order to answer this question, we need to cross validated the combinaion of each model. let us start with the first model.

cross- validation (leave one out) for the model on all six regressors

```
options(max.print = 10)
n <- length(Toyotacor$Price)
diff <- dim(n)
percdiff <- dim(n)
for (k in 1:n) {
  train1 <- c(1:n)
  train <- train1[train1 !=k ]
  m2 <- lm(Price~ ., data = Toyotacor[train,])
  pred <- predict(m2, newdat = Toyotacor[-train,])
  obs <- Toyotacor$Price[-train]
  diff[k] <- obs - pred
  percdiff[k] <- abs(diff[k]) / obs
}
Toyotacor.m2.me <- mean(diff)
Toyotacor.m2.rmse <- sqrt(mean(diff**2))
Toyotacor.m2.mape <- 100*(mean(percdiff))

Toyotacor.m2.me
```

```
## [1] -1.842833
```

```
Toyotacor.m2.rmse
```

```
## [1] 1350.015
```

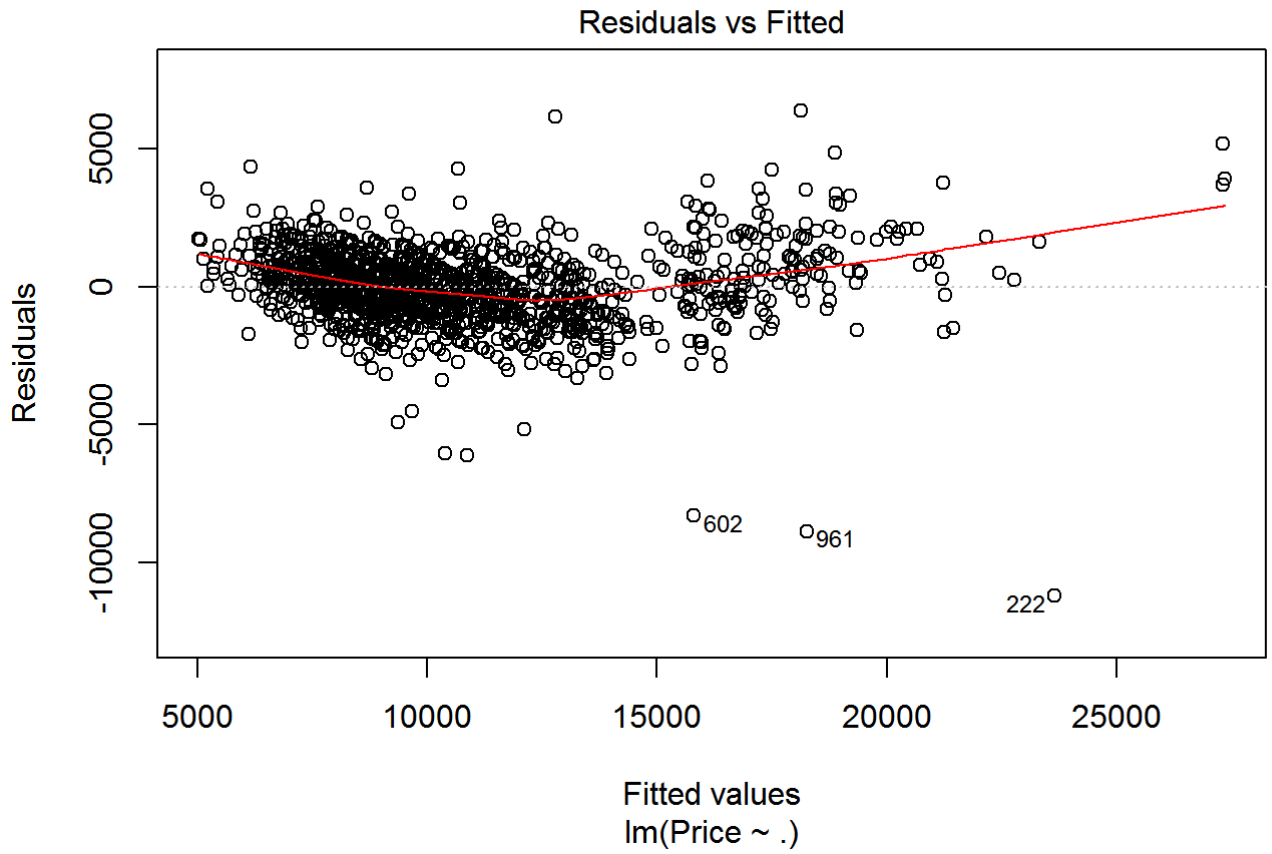
```
Toyotacor.m2.mape
```

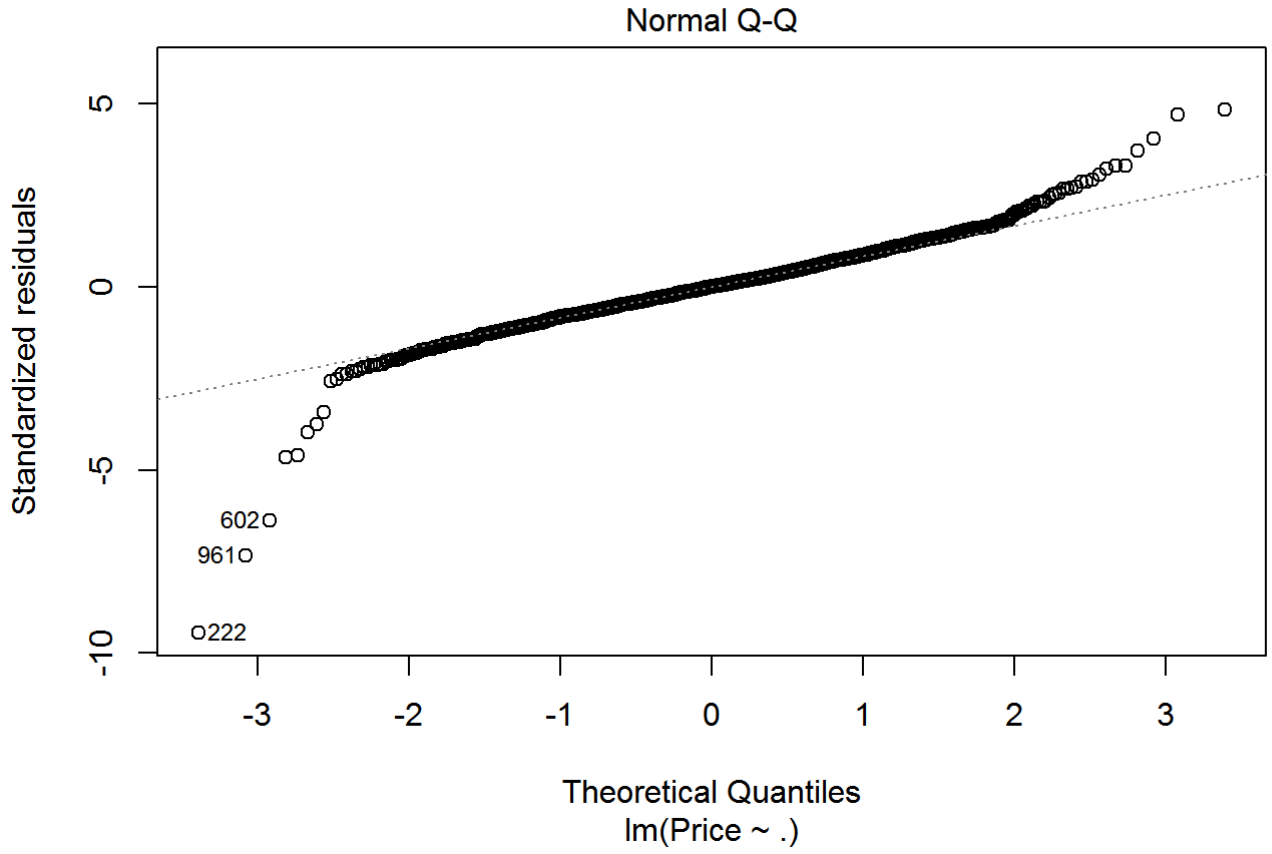
```
## [1] 9.53237
```

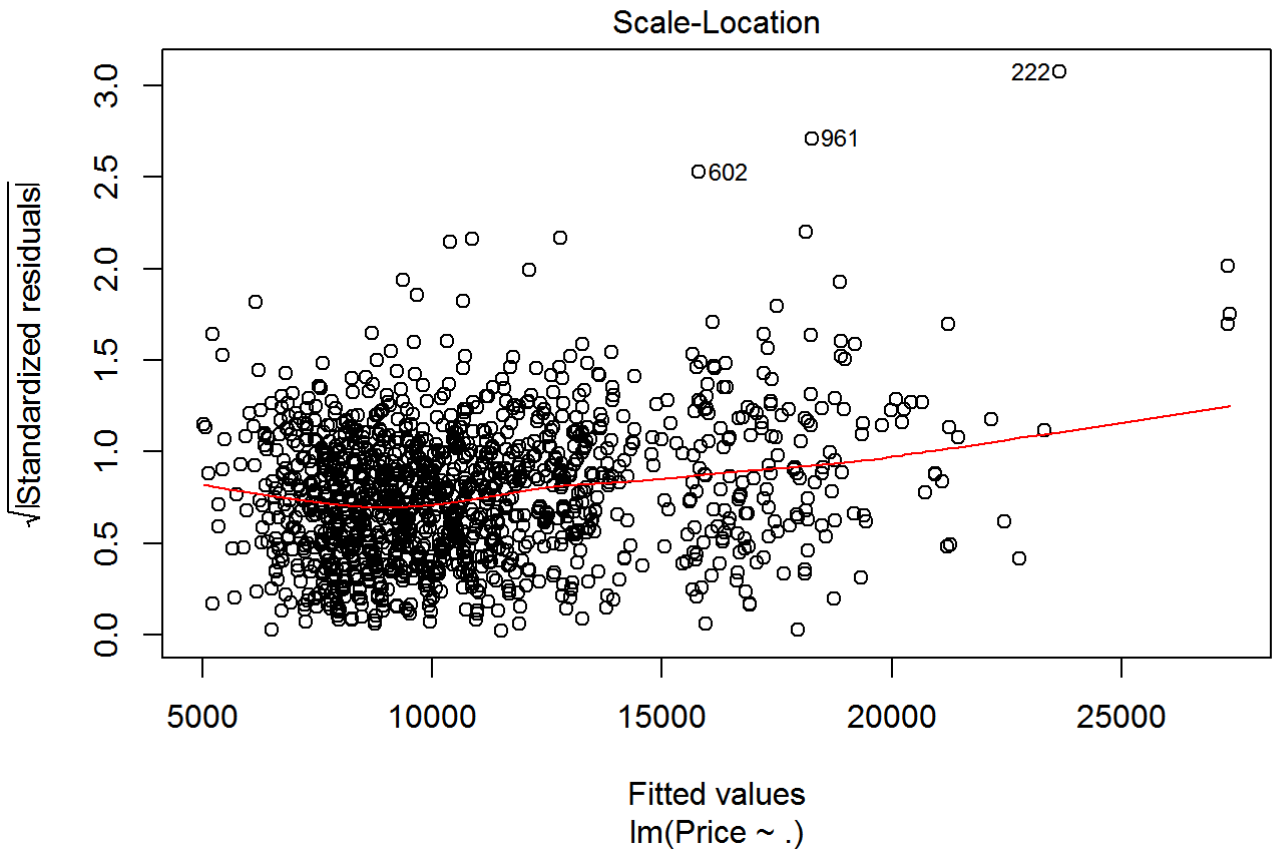
```
predicted.price
```

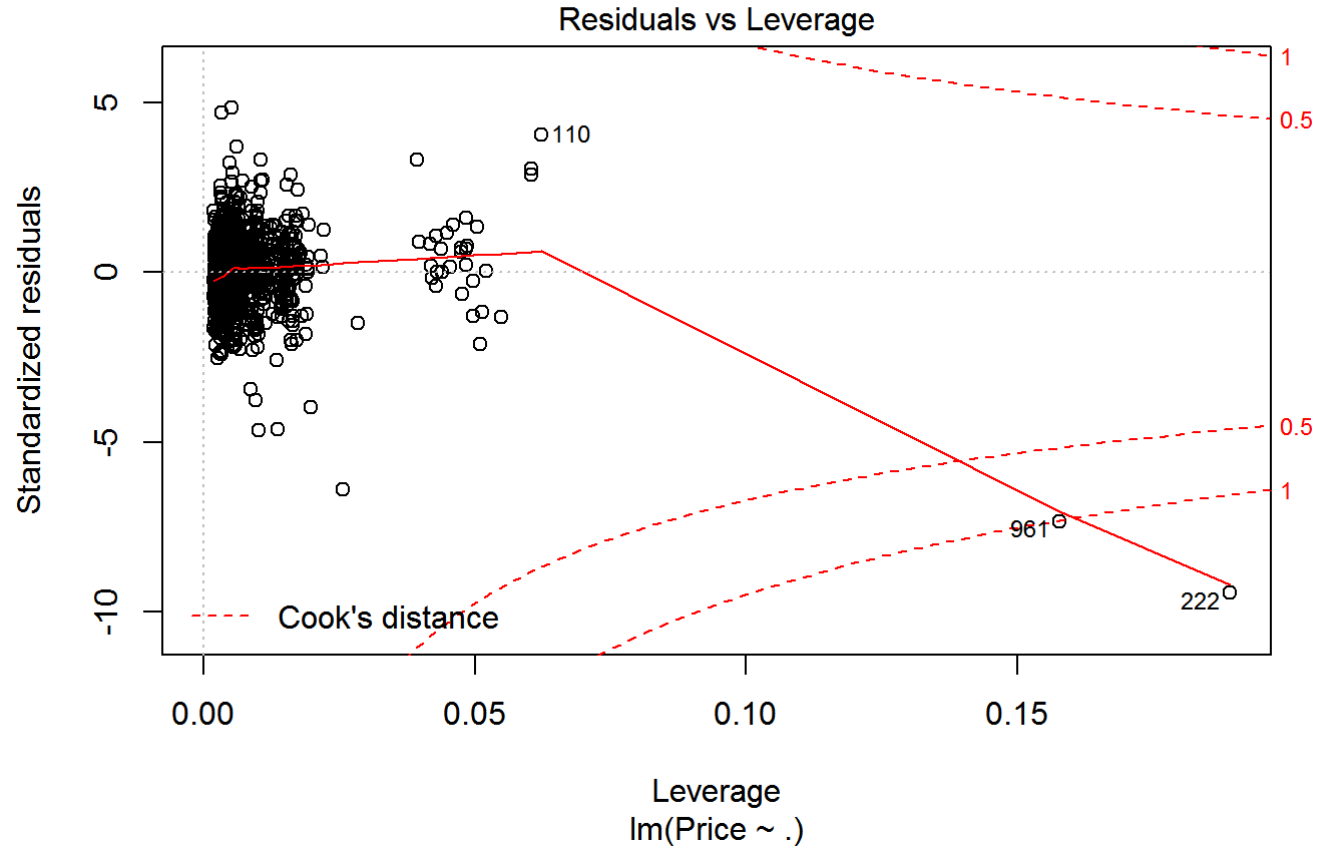
```
##        1
## 22772.63
```

```
#check if assumptions are met...
plot(Toyotacor.m1)
```


Residuals vs Fitted
Residuals
5000
0
-5000
-10000
602
961
222
5000
10000
15000
20000
25000
Fitted values
lm(Price ~ .)


Normal Q-Q
Standardized residuals
5
0
-5
-10
602
961
222
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Price ~ .)


Scale-Location
√|Standardized residuals|
222
961
602
0.0
0.5
1.0
1.5
2.0
2.5
3.0
5000
10000
15000
20000
25000
Fitted values
lm(Price ~ .)


Residuals vs Leverage
Standardized residuals
110
961
222
Cook's distance
1
0.5
0.5
1
5
0
-5
-10
0.00
0.05
0.10
0.15
Leverage
lm(Price ~ .)

```
ggsave("HW3 Graph/ToyotaCorlin.pdf")
```

```
## Saving 7 x 5 in image
```