Consider only the below columns and prepare a prediction model for predicting Price.

Corolla<-Corolla[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]

#importing the data

library(readr)

data <- read.csv(file.choose())

View(data)

str(data)

data1 <- data[,c(3,4,7,9,13,14,16,17,18)]##The busieness objective is to predict the price with this variables

View(data1)

attach(data1)

##First moment business model ##

summary(data1)

Price Age\_08\_04 KM HP cc Doors Gears

Min. : 4350 Min. : 1.00 Min. : 1 Min. : 69.0 Min. : 1300 Min. :2.000 Min. :3.000

1st Qu.: 8450 1st Qu.:44.00 1st Qu.: 43000 1st Qu.: 90.0 1st Qu.: 1400 1st Qu.:3.000 1st Qu.:5.000

Median : 9900 Median :61.00 Median : 63390 Median :110.0 Median : 1600 Median :4.000 Median :5.000

Mean :10731 Mean :55.95 Mean : 68533 Mean :101.5 Mean : 1577 Mean :4.033 Mean :5.026

3rd Qu.:11950 3rd Qu.:70.00 3rd Qu.: 87021 3rd Qu.:110.0 3rd Qu.: 1600 3rd Qu.:5.000 3rd Qu.:5.000

Max. :32500 Max. :80.00 Max. :243000 Max. :192.0 Max. :16000 Max. :5.000 Max. :6.000

Quarterly\_Tax Weight

Min. : 19.00 Min. :1000

1st Qu.: 69.00 1st Qu.:1040

Median : 85.00 Median :1070

Mean : 87.12 Mean :1072

3rd Qu.: 85.00 3rd Qu.:1085

Max. :283.00 Max. :1615

>##second moment business model

var(HP)##224.432the data points are spread from the mean

var(Age\_08\_04)##345.9596 the data points are spread from the mean

var(cc)## 180104.1 the data points are far away from the mean

var(Doors)## 0.9075927 the data points are near to mean

var(KM)##1406733707 data point are spread from the mean

var(Gears)##0.03553619 data point are near to mean

var(Quarterly\_Tax)##1619.563 data point are spread from the mean

var(Weight)##2771.088 data point are spred from mean

sd(HP)##14.98 speed variable is 14.98 times deviated from the mean

sd(Age\_08\_04)##18.599 variable has the higest HP compare other variables in the data set

sd(cc)##424.3868 data is spread from mean

sd(Doors)##0.9526766 the data is more normalized

sd(KM)##37506 the data is spread from the mean

sd(Gears)##0.1885104 the data is less spread from the mean

##Third moment business model

skewness(HP)##0.9548 moderately skewed right

skewness(Age\_08\_04)## -0.8258 negatively skewed left

skewness(cc)##27.40313 storngly skewed right

skewness(Doors)##-0.07631517 negatively skewed left

skewness(KM)##1.014851 positively skewed right

skewness(Gears)##2.281573 positively skewed right

##fourth moment business model

kurtosis(HP)##11.80152 positive kurtosis means the data set has heavy tails on either side

kurtosis(Age\_08\_04)##2.919 the distribution is normal

kurtosis(cc)##930.4696 heavy tails on either side

kurtosis(Doors)##1.27581 skinner tails

kurtosis(KM)##4.67 heavier tails on either side

kurtosis(Gears)##40.568 heavier tails on either side

## Fifth moment model

plot(cc,Price)## the cc does not effect the price that much

plot(KM,Price)## the km effects the price in negative way

plot(Age\_08\_04,Price)## As the age increases they are going for less expensive one

plot(Gears,Price)## no effect on the price

plot(HP,Price)

plot(Quarterly\_Tax,Price)

##TO check for correlation between variable

pairs(data1)

## model buliding##

model <- lm(Price~.,data = data1)

summary(model)

##multiple r square value .8638 ##

##but here the cc and doors are insignificant

model1 <- lm(Price~cc)

summary(model1)## cc became significant rsquare=0.015

model2 <- lm(Price~Doors)

summary(model2)## Doors became significant rsquare=0.03435

model3 <- lm(Price~Doors+cc)

summary(model3)## both became significant but rsquare is 0.04688

influence.measures(model3)

library(psych)

library(caret)

library(car)

influenceIndexPlot(model3)

influencePlot(model3)## we can clear see 81 observation is influencing the data set model

# Delete influentails records and build the model

model4 <- lm(Price ~ ., data = data1[-c(81),])

summary(model4)

## Doors are still insignificant

car::vif(model4)

car::avPlots(model4)

## by seeing the plot we can say that there is no influence of door variable in the model so we will remove it

finalmodel <- lm(data1$Price~ data1$HP+data1$cc+data1$Age\_08\_04+data1$Quarterly\_Tax+data1$KM+data1$Weight+data1$Gears,data =data1[-c(81,110,112),] )

summary(finalmodel)

##rsquare-.8638

pre <- predict(finalmodel)

pre

plot(finalmodel)

qqPlot(finalmodel)



