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")]

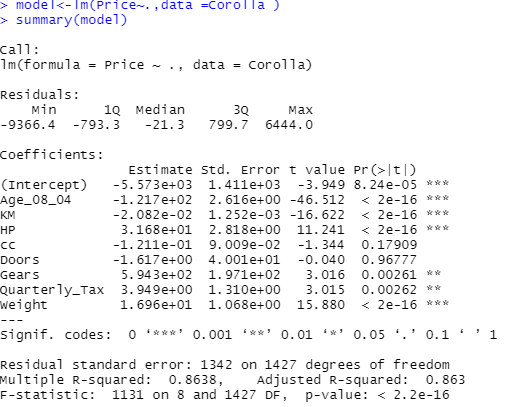
BUSSINESS PROBLEM: To predict price

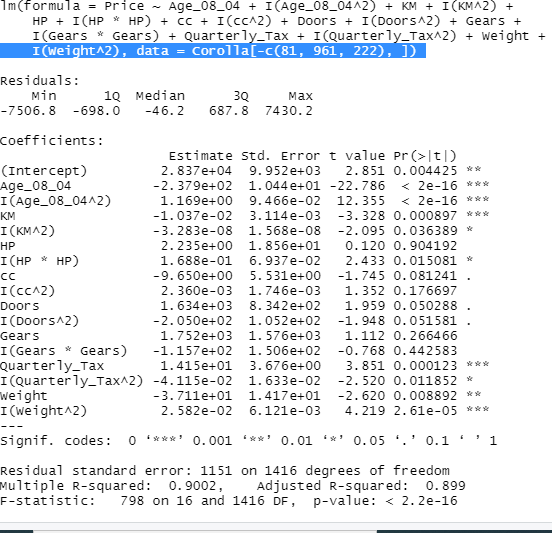
|  |  |  |
| --- | --- | --- |
| PRICE:  HISTOGRAM | Data is right skewed | |
| BOX-PLOT | Outliers are present in upper extreme | |
| QQPLOT | Data shows normal distribution after transformation | |
| skewness(Corolla$Price)  [1] 1.702105  > kurtosis(Corolla$Price)  [1] 6.720604 | | right skewed because value is greater than 0  leptokurtic as data is more distributed in  tails |

|  |  |  |
| --- | --- | --- |
| HP:  HISTOGRAM | Data is left skewed | |
| BOX-PLOT | Outlier is present at upper extreme | |
| QQPLOT | Data is normally distributed | |
| skewness(Corolla$HP)  [1] 0.9548369  > kurtosis(Corolla$HP)  [1] 11.80152 | | right skewed because value is less than 0  leptokurtic, due to its peakedness and data is more distributed in  tails |

MULTI-VARIET ANALYSIS

|  |
| --- |
|  |
| cor(Corolla)     |  | | --- | | Price Age\_08\_04 KM HP cc  Price 1.00000000 -0.876590497 -0.56996016 0.31498983 0.12638920  Age\_08\_04 -0.87659050 1.000000000 0.50567218 -0.15662202 -0.09808374  KM -0.56996016 0.505672180 1.00000000 -0.33353795 0.10268289  HP 0.31498983 -0.156622020 -0.33353795 1.00000000 0.03585580  cc 0.12638920 -0.098083739 0.10268289 0.03585580 1.00000000  Doors 0.18532555 -0.148359215 -0.03619661 0.09242450 0.07990330  Gears 0.06310386 -0.005363947 0.01502333 0.20947715 0.01462935  Quarterly\_Tax 0.21919691 -0.198430508 0.27816470 -0.29843172 0.30699580  Weight 0.58119759 -0.470253184 -0.02859846 0.08961406 0.33563740  Doors Gears Quarterly\_Tax Weight  Price 0.18532555 0.063103857 0.219196911 0.58119759  Age\_08\_04 -0.14835921 -0.005363947 -0.198430508 -0.47025318  KM -0.03619661 0.015023328 0.278164697 -0.02859846  HP 0.09242450 0.209477146 -0.298431717 0.08961406  cc 0.07990330 0.014629352 0.306995798 0.33563740  Doors 1.00000000 -0.160141430 0.109363225 0.30261764  Gears -0.16014143 1.000000000 -0.005451955 0.02061328  Quarterly\_Tax 0.10936323 -0.005451955 1.000000000 0.62613373  Weight 0.30261764 0.020613284 0.626133733 1.00000000 | |  | | |  | | --- | |  | | |
| cor2pcor(cor(Corolla))  [,1] [,2] [,3] [,4] [,5] [,6]  [1,] 1.000000000 -0.776238352 -0.402745405 0.28521314 -0.03556185 -0.001069746  [2,] -0.776238352 1.000000000 0.002383081 0.24531845 -0.02014628 -0.002800916  [3,] -0.402745405 0.002383081 1.000000000 -0.06039653 0.05108725 0.026724172  [4,] 0.285213137 0.245318454 -0.060396533 1.00000000 0.09871851 0.068175272  [5,] -0.035561846 -0.020146283 0.051087249 0.09871851 1.00000000 -0.016060377  [6,] -0.001069746 -0.002800916 0.026724172 0.06817527 -0.01606038 1.000000000  [7,] 0.079586710 0.051074865 0.100506331 0.20769268 -0.01198838 -0.189249333  [8,] 0.079548117 0.015830863 0.261673195 -0.38254954 0.12380803 -0.074825415  [9,] 0.387523482 0.094746528 0.187502181 0.12427899 0.16043171 0.231960007  [,7] [,8] [,9]  [1,] 0.07958671 0.07954812 0.38752348  [2,] 0.05107486 0.01583086 0.09474653  [3,] 0.10050633 0.26167319 0.18750218  [4,] 0.20769268 -0.38254954 0.12427899  [5,] -0.01198838 0.12380803 0.16043171  [6,] -0.18924933 -0.07482541 0.23196001  [7,] 1.00000000 0.03732241 -0.02325832  [8,] 0.03732241 1.00000000 0.51026027  [9,] -0.02325832 0.51026027 1.00000000 |
| INFERENCE:  MODEL-BUILDING |





LINE ASSUMPTIONS of model2

|  |  |
| --- | --- |
|  | Residuals are dependent of each other,  (no auto-correlation )  And residuals are symmetrically distributed, hence homoscedasticity in nature |
|  | Residuals are normally distributed |
|  | VIF values are <10,hence the input variables are independent of each other  ( No multi-collinearity) |
|  | Residuals have equal variance |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MODEL | MULTIPL R2 | ADJUSTED R2 | TRAIN ERROR | TEST ERROR | INFERENCE |
| Model1 | 0.863 | 0.863 | 1361.7 | 1361.7 | Right fit model |
| Model2 | 0.90 | 0.89 | 1159.7 | 1159.7 | Right fit model |

PYTHON CODE

import pandas as pd

cor=pd.read\_csv("C:/Users/USER/Desktop/ToyotaCorolla.csv",encoding= 'unicode\_escape')

dataset=cor[['Price','Age\_08\_04','KM','HP','Doors','cc','Gears','Quarterly\_Tax','Weight']]

dataset=pd.DataFrame(dataset)

import numpy as np

import scipy

from scipy import stats

import statsmodels.formula.api as smf

import matplotlib.pylab as plt

import seaborn as sns

import statsmodels.api as sm

########univariet analysis

plt.hist(dataset.Price)

plt.boxplot(dataset.Price)

#######multi variet analysis########

cor\_cof=np.corrcoef(dataset)

sns.pairplot(dataset)

#######model

model=smf.ols('Price~Age\_08\_04+KM+HP+Doors+cc+Gears+Quarterly\_Tax+Weight',data=dataset).fit()

model.summary()

##multiple r2 =0.86 and adjusted r2=0.86

fig = sm.graphics.influence\_plot(model)

model=smf.ols('Price~Age\_08\_04+I(Age\_08\_04^2)+KM+(KM^2)+HP+(HP^2)+Doors+(Doors^2)+cc+I(cc^2)+Gears+I(Gears^2)+Quarterly\_Tax+I(Quarterly\_Tax^2)+Weight+I(Weight^2)',data=dataset1).fit()

model.summary()

##########train and test##########

from sklearn.model\_selection import train\_test\_split

dataset1\_train,dataset1\_test=train\_test\_split(dataset1,test\_size=0.2)

model=smf.ols('Price~Age\_08\_04+I(Age\_08\_04^2)+KM+(KM^2)+HP+(HP^2)+Doors+(Doors^2)+cc+I(cc^2)+Gears+I(Gears^2)+Quarterly\_Tax+I(Quarterly\_Tax^2)+Weight+I(Weight^2)',data=dataset1\_train).fit()

pred=model.predict(dataset1\_test)

#######multuple r2 0.89 and adjustedr2=0.89

err=pred-dataset1\_test.Price

err\_sqr=err\*err

err\_mean=np.mean(err\_sqr)

err\_sqrt=np.sqrt(err\_mean)

err\_sqrt

pred=model.predict(dataset1\_train)

err=pred-dataset1\_train.Price

err\_sqr=err\*err

err\_mean=np.mean(err\_sqr)

err\_sqrt=np.sqrt(err\_mean)

err\_sqrt

##########train error=1179.6681 and test error=1126.40

###right fit model