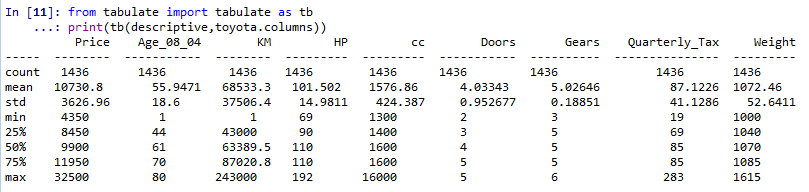
**Case Study of Toyota Corolla**

**Problem Statement:**

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")]

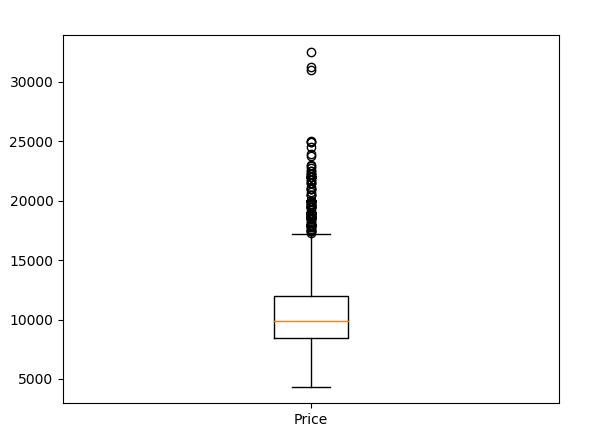
**Exploratory data analysis:**



The above is the descriptive statistics of the Toyota dataset. There are 1436 observations and 9 variables in the data frame.

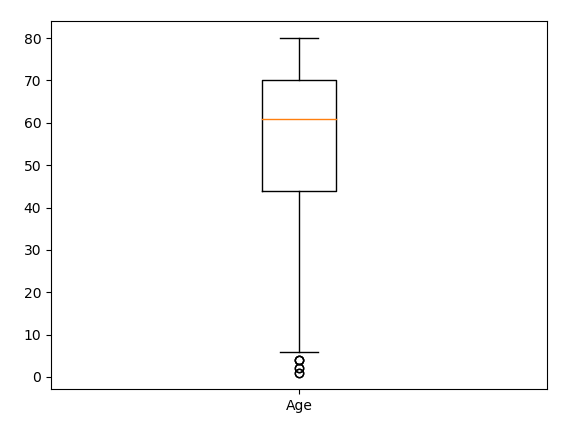
The following are the boxplots of the variables:

Price:

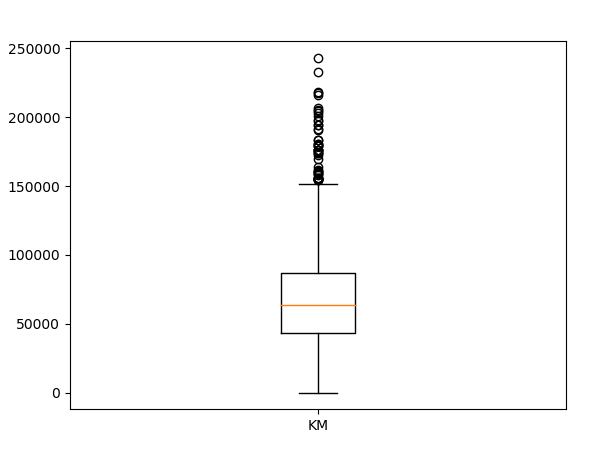


The price above the upper wiskers has many outliers indicating that the price of certain cars are much above the median.

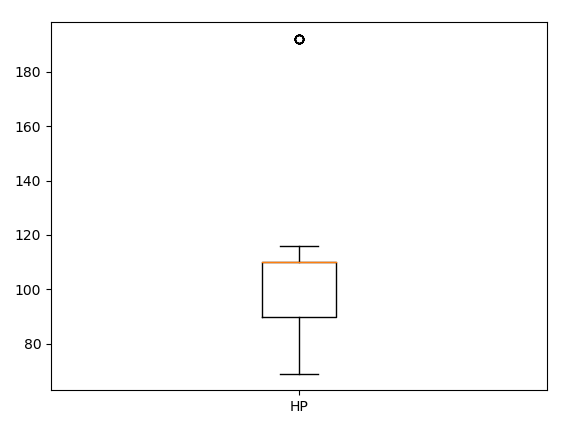
Age:



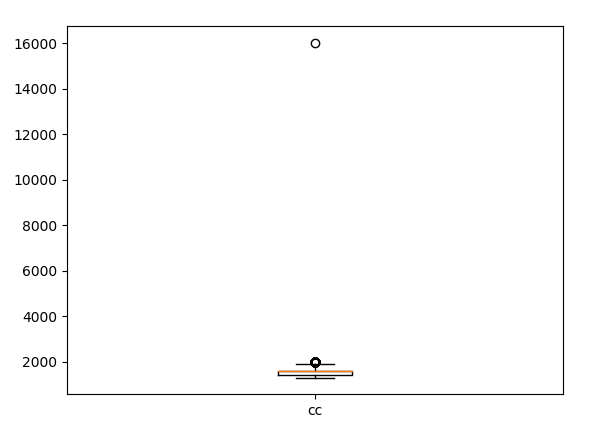
KM:



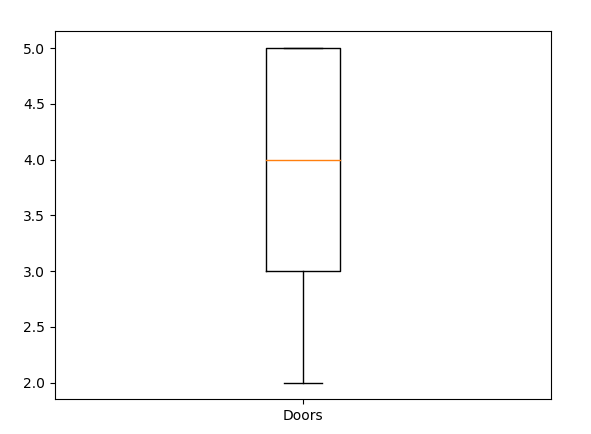
HP:



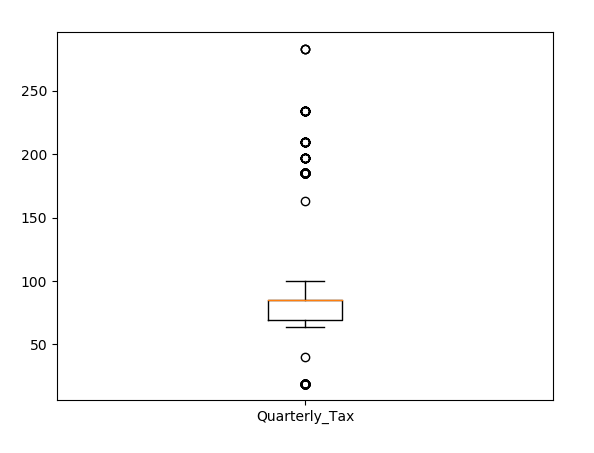
CC:



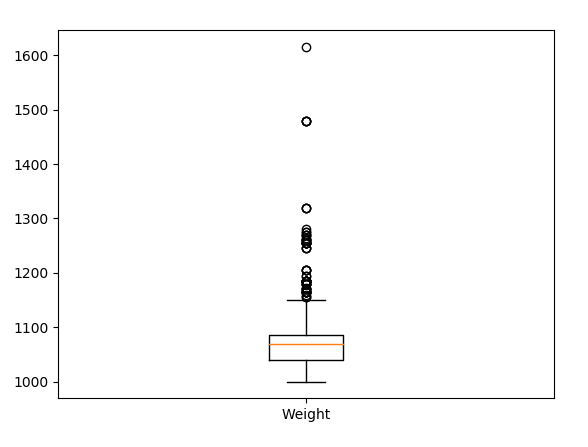
Doors:



Quarterly Tax:

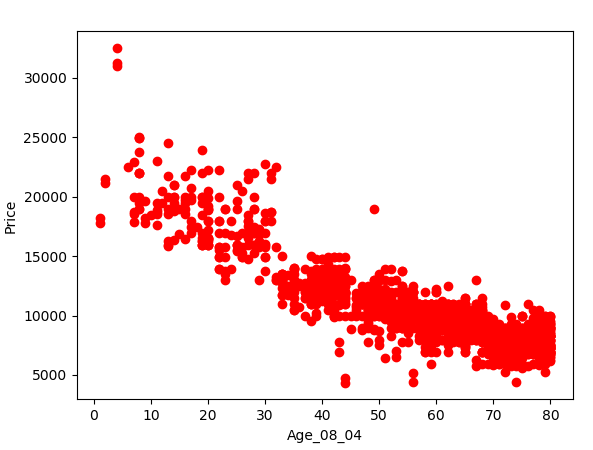


Weight:



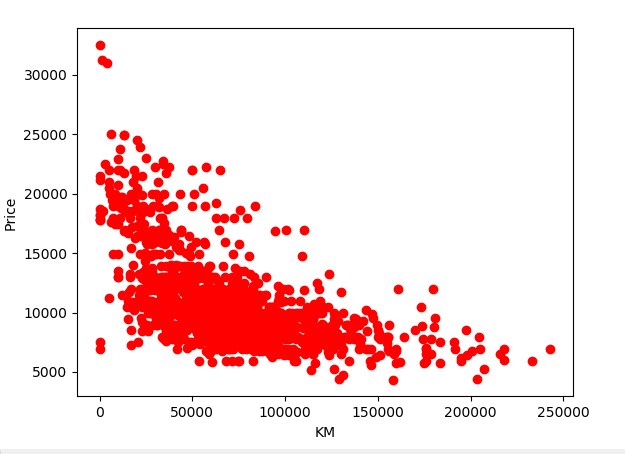
Scatter Plots:

Price Vs Age:



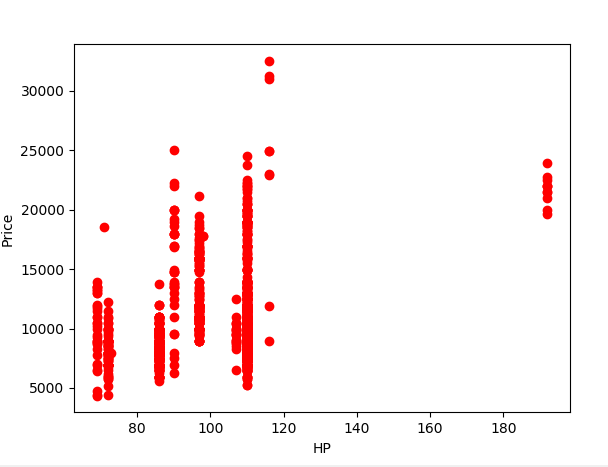
The above graph indicates that as the age of the car increases the price decreases.

Price Vs KM

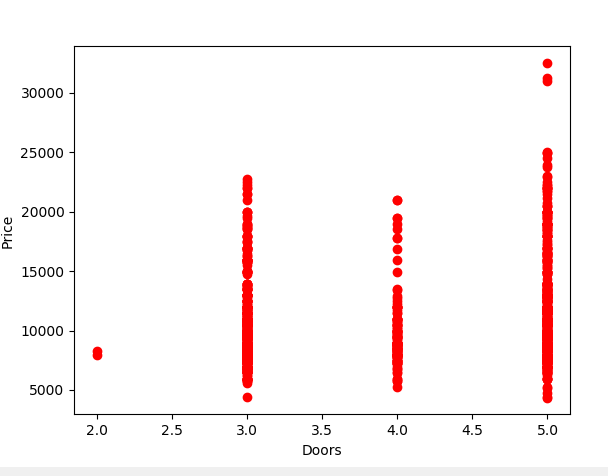


The above graph indicates that as the Kilometres run by the car increases the price decreases.

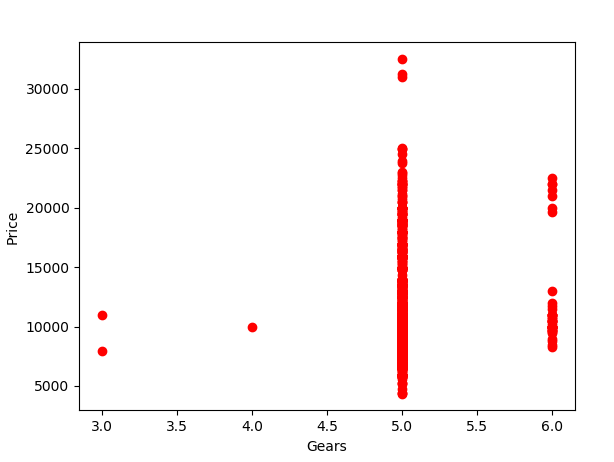
Price Vs HP



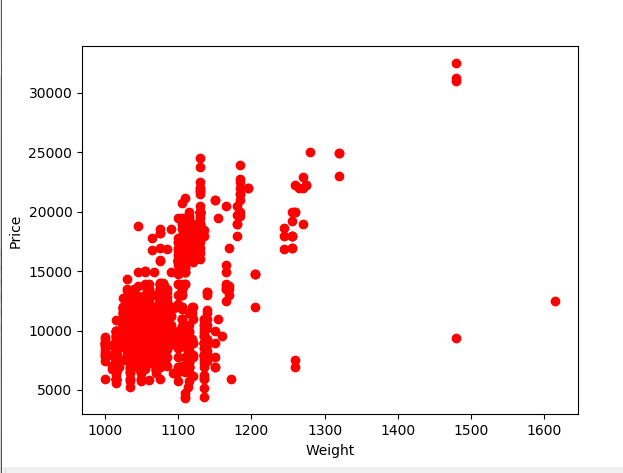
Price Vs Doors



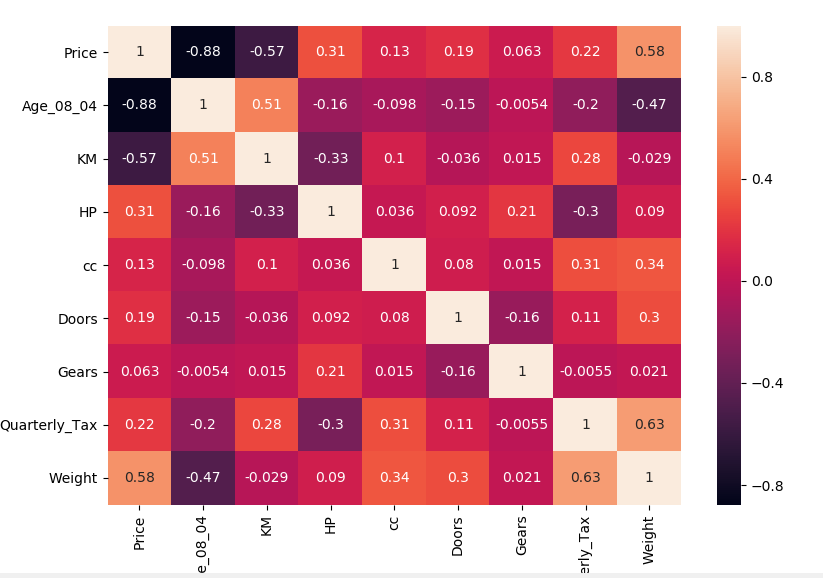
Price Vs Gears:



Price Vs Weight:



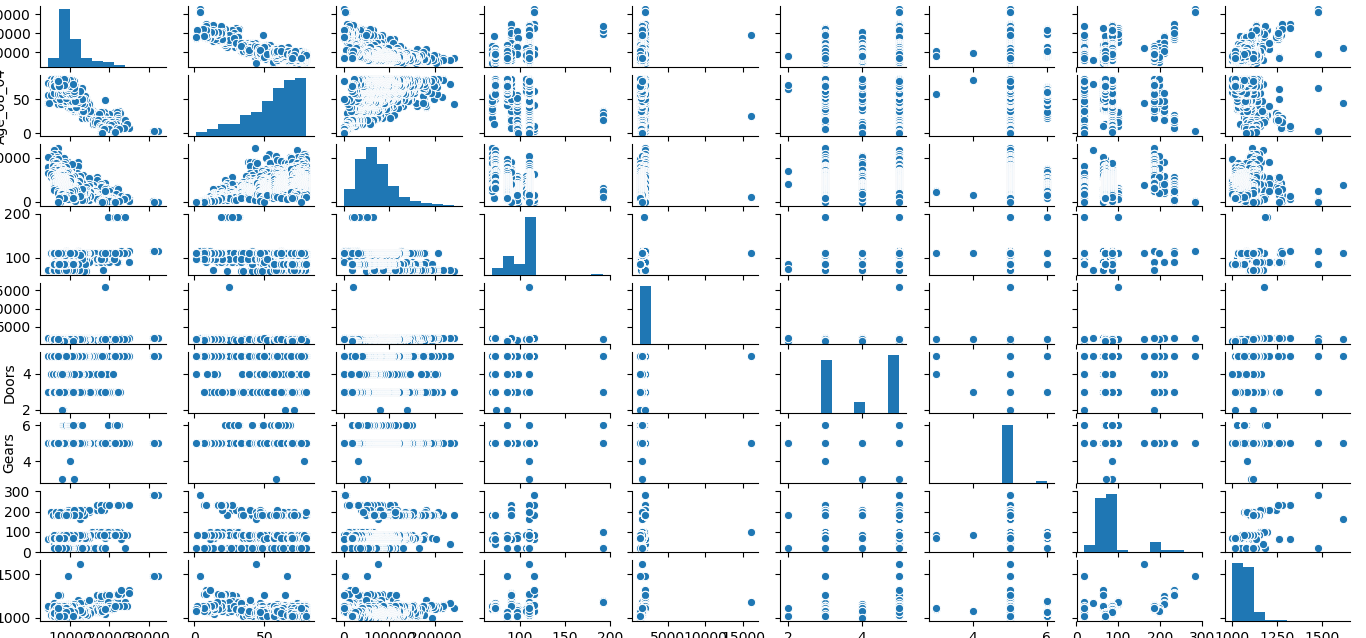
The heat map of the correlation between different variables



From the above heat map we can estimate the following:

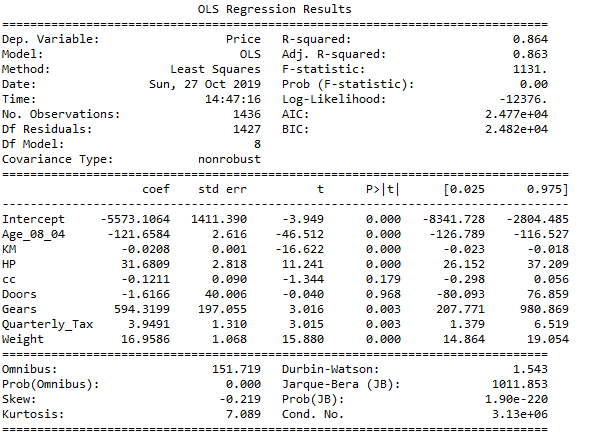
* High negative correlation between Price and Age
* Almost 60% negative correlation between Price and KM
* The correlation between independent variables are no significant enough to consider Multi Co linearity problem in the given dataframe.

The following is the pairplot of the dataframe:



Building the regression model:

The following is the model considering all the variables



From the above model summary.

* The variables or the features “Doors” and “CC” are statistically insignificant for the outcome variable of price.
* The R-squared and Adjusted R-squared values of .864 and 0.863 indicates that the model is very good.
* The F-statistics value of less than 0.05 also indicates the good health of the model.

The above models have insignificant variables that needs to be treated.

# preparing model based only on Doors

ml\_d=smf.ols('Price~Doors',data = toyota).fit()

ml\_d.summary()

R-squared: 0.034

Adj. R-squared: 0.034

# Preparing model based only on cc

ml\_cc=smf.ols('Price~cc',data = toyota).fit()

ml\_cc.summary()

R-squared: 0.016

Adj. R-squared: 0.015

# Preparing model based only on Doors & cc

ml\_dc=smf.ols('Price~Doors+cc',data = toyota).fit()

ml\_dc.summary()

R-squared: 0.047

Adj. R-squared: 0.046

# Preparing model based only on Age\_08\_04,Doors & cc

ml\_ac=smf.ols('Price~Age\_08\_04+Doors+cc',data = toyota).fit()

ml\_ac.summary()

R-squared: 0.773

Adj. R-squared: 0.772

**The variable age has a high statistical significance in the outcome of the output variable “Price” and hence there is an improvement in the overall R-squared and Adjusted R-squared value of the model**

# Preparing model based only on KM,Doors & cc

ml\_kdc=smf.ols('Price~KM+Doors+cc',data = toyota).fit()

ml\_kdc.summary()

R-squared: 0.382

Adj. R-squared: 0.381

# Preparing model based only on HP,Doors & cc

ml\_hdc=smf.ols('Price~HP+Doors+cc',data = toyota).fit()

ml\_hdc.summary()

R-squared: 0.135

Adj. R-squared: 0.133

# Preparing model based only on Gears,Doors & cc

ml\_gdc=smf.ols('Price~Gears+Doors+cc',data = toyota).fit()

ml\_gdc.summary()

R-squared: 0.055

Adj. R-squared: 0.053

# Preparing model based only on Quarterly\_Tax,Doors & cc

ml\_qdc=smf.ols('Price~Quarterly\_Tax+Doors+cc',data = toyota).fit()

ml\_qdc.summary()

R-squared: 0.077

Adj. R-squared: 0.075

# Preparing model based only on Weight,Doors & cc

ml\_wdc=smf.ols('Price~Weight+Doors+cc',data = toyota).fit()

ml\_wdc.summary()

R-squared: 0.343

Adj. R-squared: 0.342

# Preparing model based only on Weight & Doors

ml\_wd=smf.ols('Price~Weight+Doors',data = toyota).fit()

ml\_wd.summary()

R-squared: 0.338

Adj. R-squared: 0.337

# Preparing model based only on Weight & cc

ml\_wc=smf.ols('Price~Weight+cc',data = toyota).fit()

ml\_wc.summary()

R-squared: 0.343

Adj. R-squared: 0.342

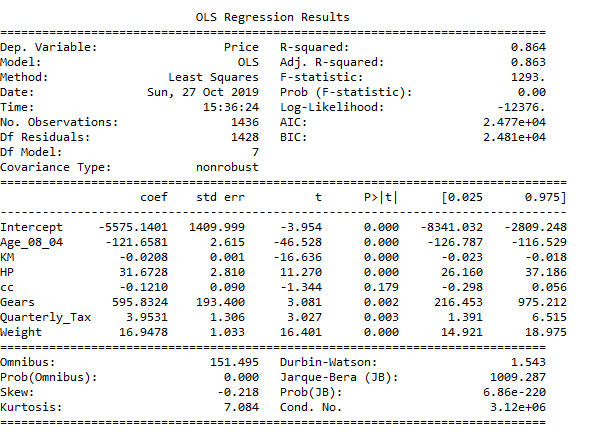
#Dropping variable "doors" since it is insignificant to the outcome

ml1\_v2 = smf.ols('Price~Age\_08\_04+KM+HP+cc+Gears+Quarterly\_Tax+Weight',data=toyota).fit() # regression model

ml1\_v2.summary()

R-squared: 0.864

Adj. R-squared: 0.863

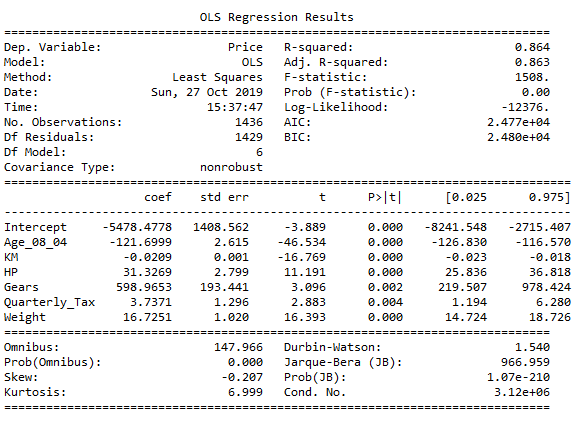


In the above model summary. The variable cc is statistically insignificant to the outcome.

#Dropping variable "cc" since it is insignificant to the outcome

ml1\_v3 = smf.ols('Price~Age\_08\_04+KM+HP+Gears+Quarterly\_Tax+Weight',data=toyota).fit() # regression model

ml1\_v3.summary()

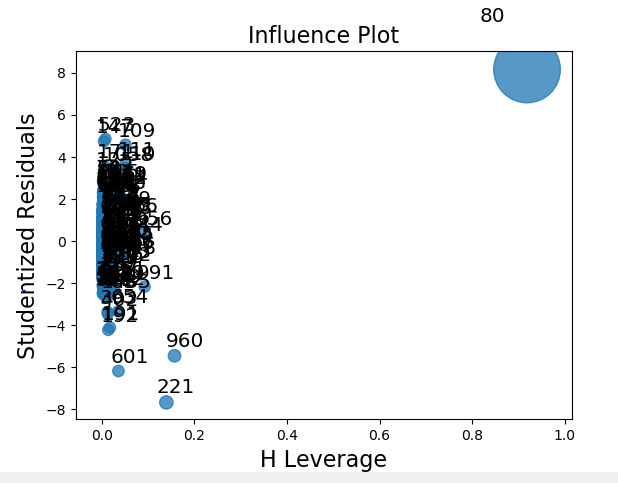


ml1\_v3.params

#The model ml1\_v3 produces a model of R-sqaured value of 0.864

Checking whether data has any influential values

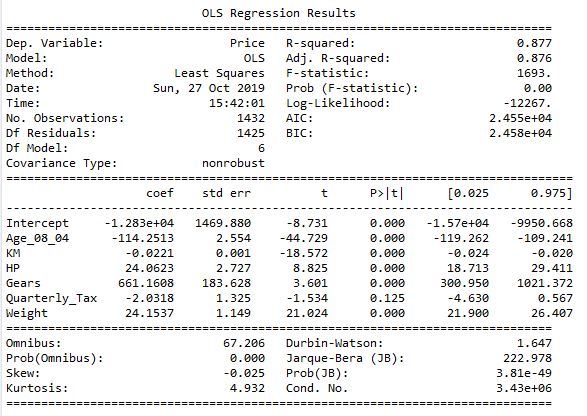
# influence index plots



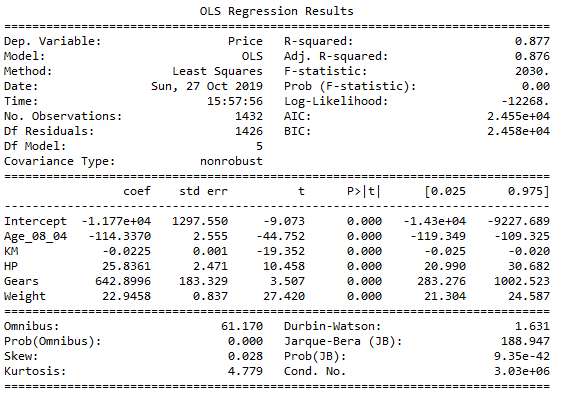
The Observations 80, 221, 601 and 960 is highly influential to the outcome of Price.

Building the final model after dropping the influential points.

We get the following results:



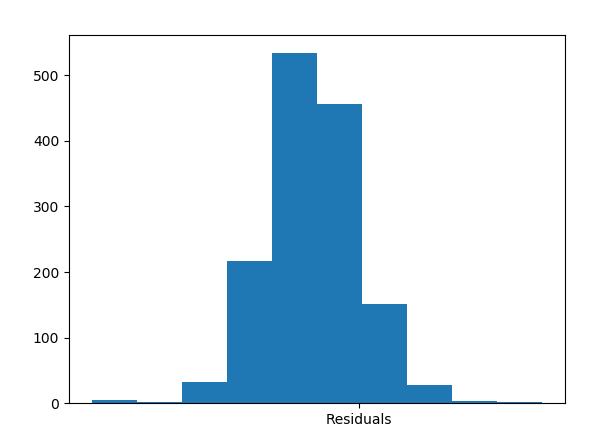
From the above model summary: The variable Quarterly Tax is an insignificant variable and rebuilding the model without quarterly tax variable`.



* The model is very good since there are no insignificant variables with high p-values and the R-squared & Adjusted R-squared values are high.

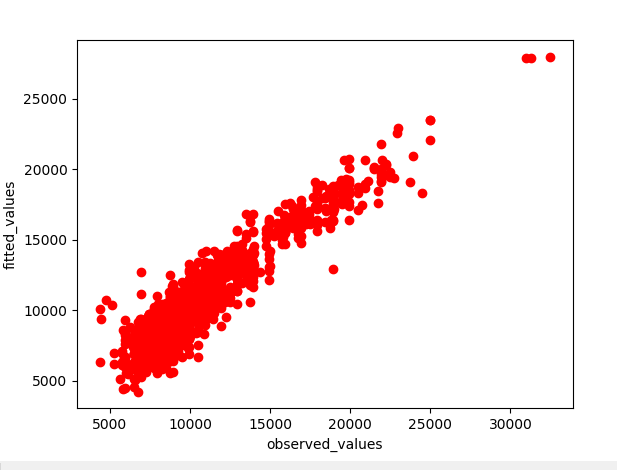
**Residual Analysis:**

Checking the normality of the residuals or errors:



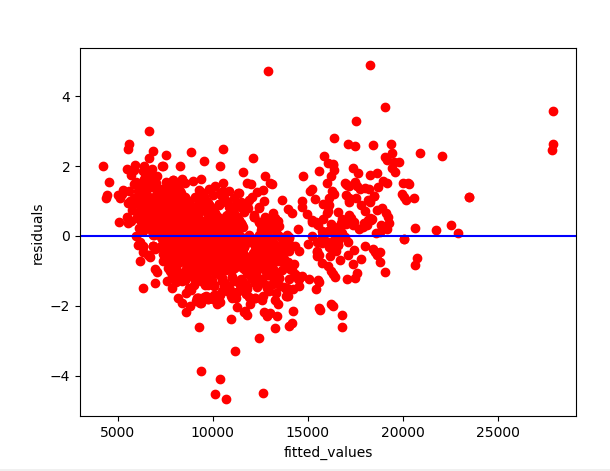
There is almost normal graph of the residuals.

**Checking for Fitted Vs Observed Values**



The above graph shows the conformance of the fitted and observed values

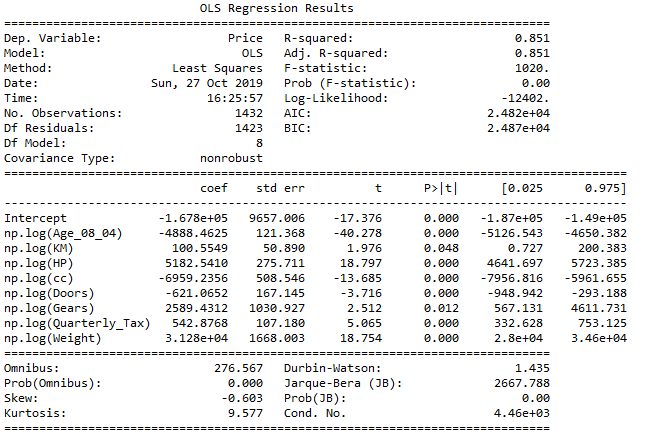
Checking for Residuals Vs Fitted values



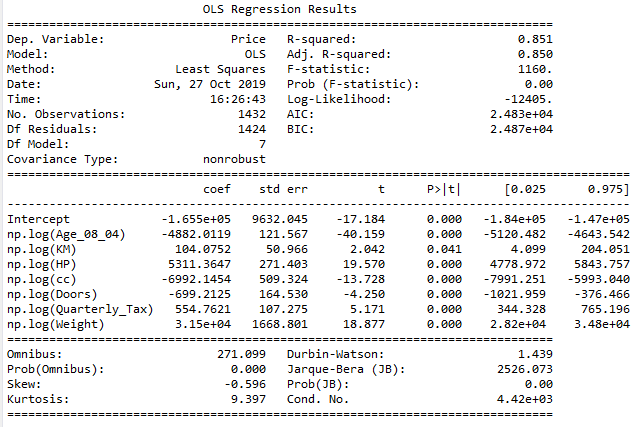
The above graph shows that there is uneven variance between the residuals and the fitted values. Hence the outcome is Heteroscedastic in nature.

To treat Heteroscedasticity, we transform the model to achieve Homoscedastic outcome.

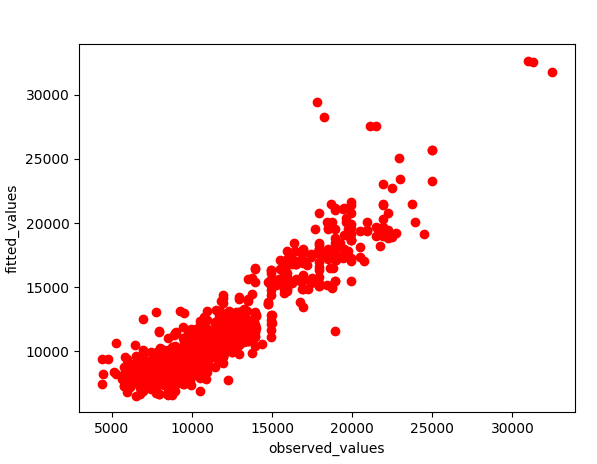
**Applying Log Transformation:**

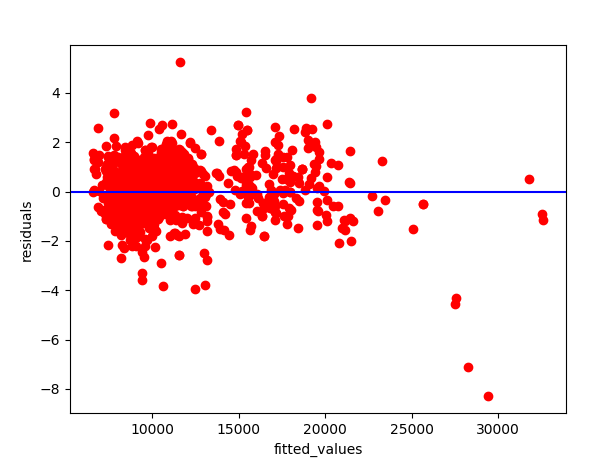


The variable “Gears” is insignificant. Hence removing the variable and rebuilding the model:



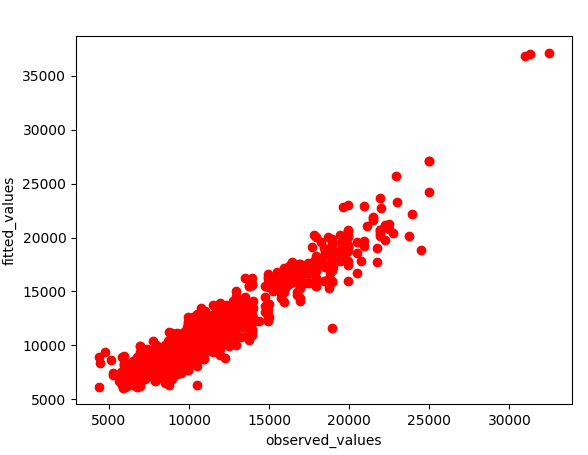
The above model provides a good R-squared values and there are no insignificant variables:



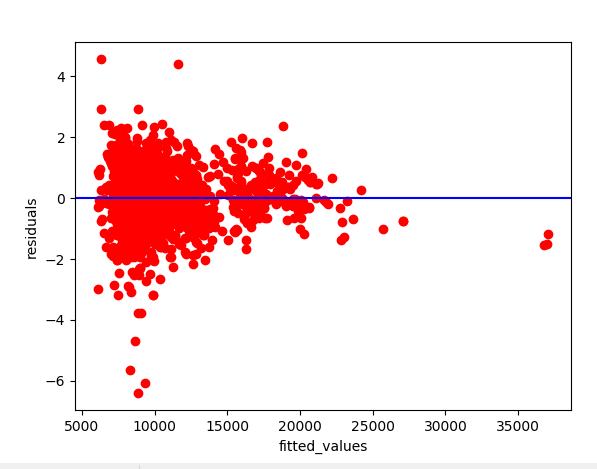


From the above graph, The log transformed model still consists of Heteroscedasticity.

**We apply exponential transformation to the model:**

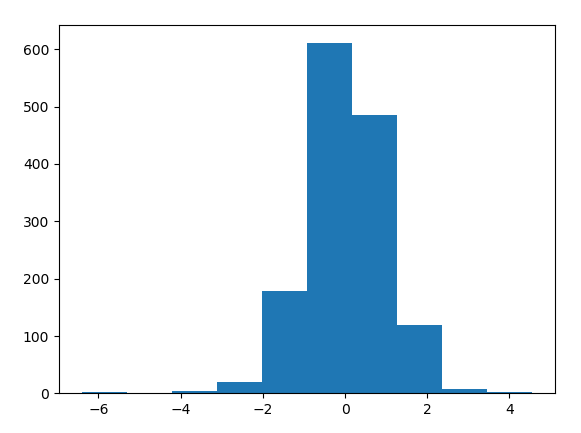


The exponential model still maintains good linearity with the fitted values



Exponential transformation removes the remaining heteroscedasticity.

Normality of exponential model residuals:



Implementing the above model in the prediction model of regression