# Multiple Linear Regression

**Instructions**

Please share your answers filled in this word document. Submit code files wherever applicable.

Please ensure you update all the details:

**Name: Prajay B. Urkude**

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

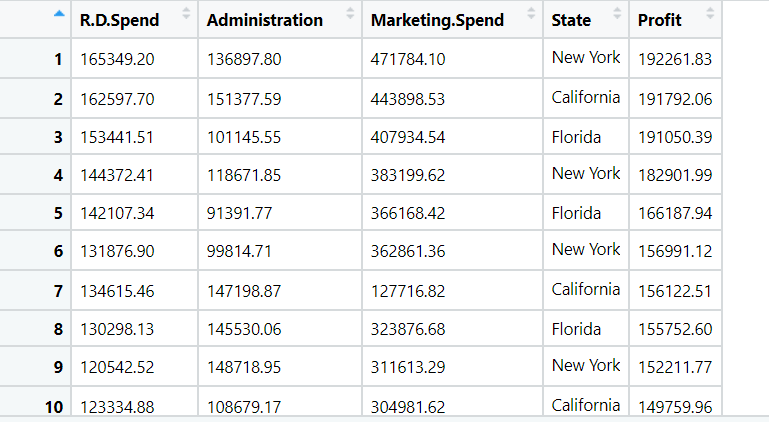
Please ensure you update all the details:

**Name: Prajay B. Urkude Batch ID:** \_16092021

**Topic: Multilinear Regression**

**Problem Statements: -**

1. An analytics company has been tasked with the crucial job of finding out what factors affect a startup company and if it will be profitable or not. For this, they have collected some historical data and would like to apply multilinear regression to derive brief insights into their data. Predict profit, given different attributes for various startup companies.



Ans: **Business Objective:**

To build model to predict the profit affecting the different attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| R&D Spend | Spend on research and development | Quantitative, Ratio | Relevant |
| Administration | Spend on administration | Quantitative, Ratio | Relevant |
| Marketing Spend | Spend on marketing | Quantitative, Ratio | Relevant |
| State | State name | Quantitative, Ratio | Relevant |
| Profit | Profit in the business | Quantitative, Ratio | Relevant |

Steps For the multilinear Regression:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* **Data Visualization**

Histogram: To check the distribution of the data

Distribution plot: To check the data is normally distributed or not but in this we get the bell shaped curve.

Boxplot: To check the outliers present in the dataset.

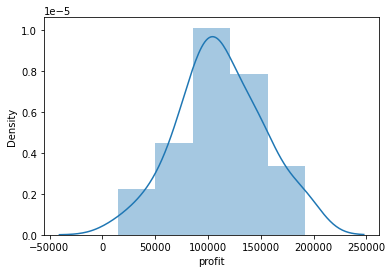
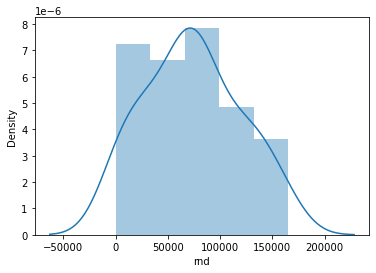
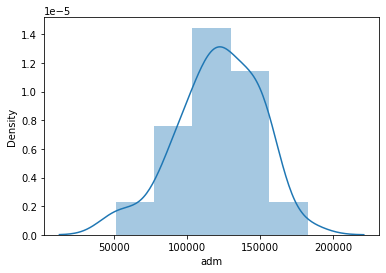
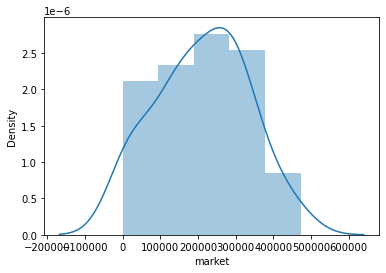
Countplot: It is used to check the count of each unique features each column

Normal Q\_Q plot: To check the data is normally distributes or not

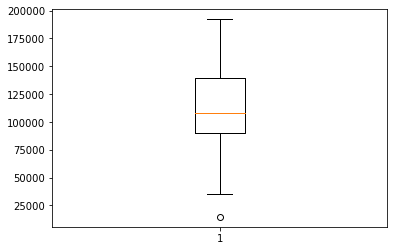
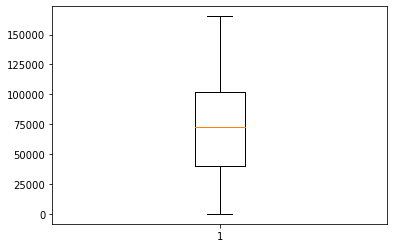
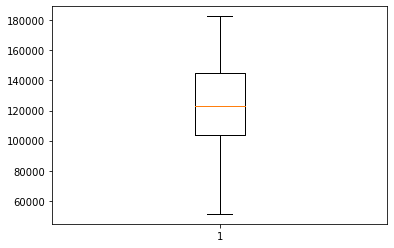
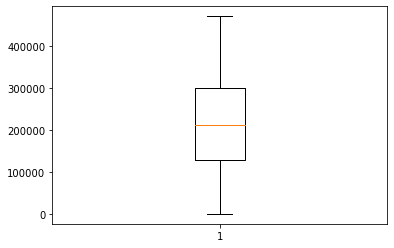
Scatter Plot: Scatter Plot is used for the bivariate visualization.

Paiplot : Pairplot is the combination of scatter plot between different variables and the histogram, It is used to find the correlation of the independent variables with the dependent variables and to find out the collinearity between different independent variables.

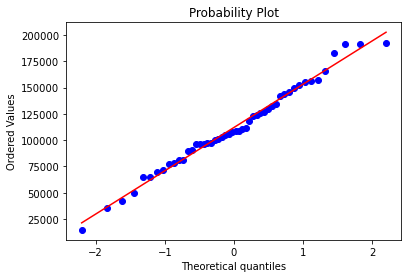
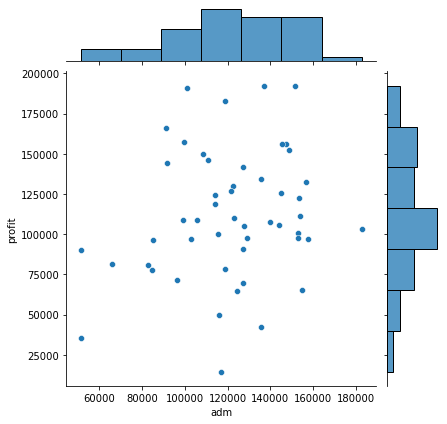
ProbPlot ; probplot is used to find the relationship between the two different variables.



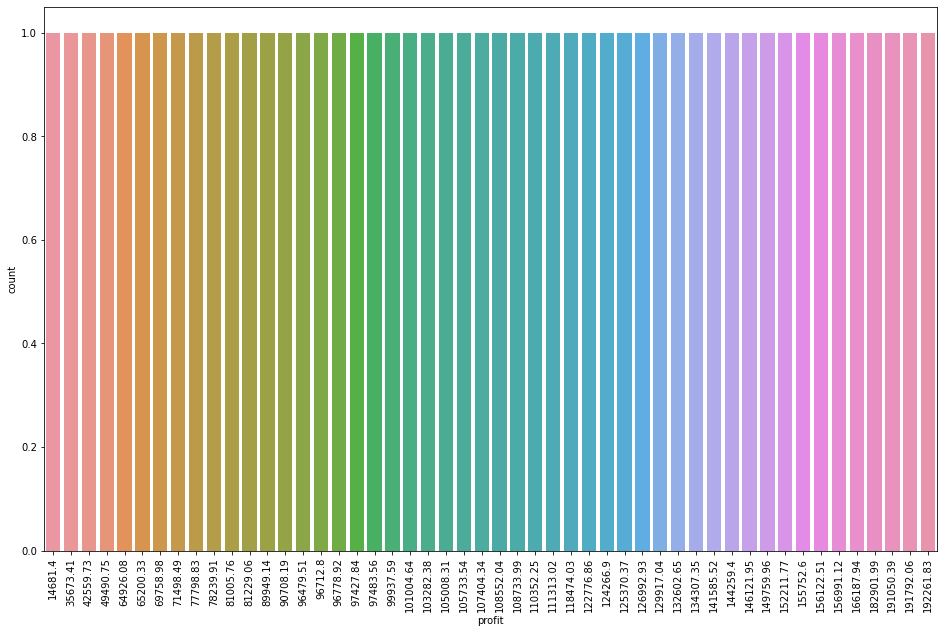
The first graph is almost normally distributed, 2nd graph slightly normally distributes, 3rd graph is also normally distributed and 4th graph is also slightly normally distributes.



From the above plot we can see that in 1st boxplot the outlier present and in other boxplots outliers is not present. So we have remove that outlier by using **winsorization** Technique.

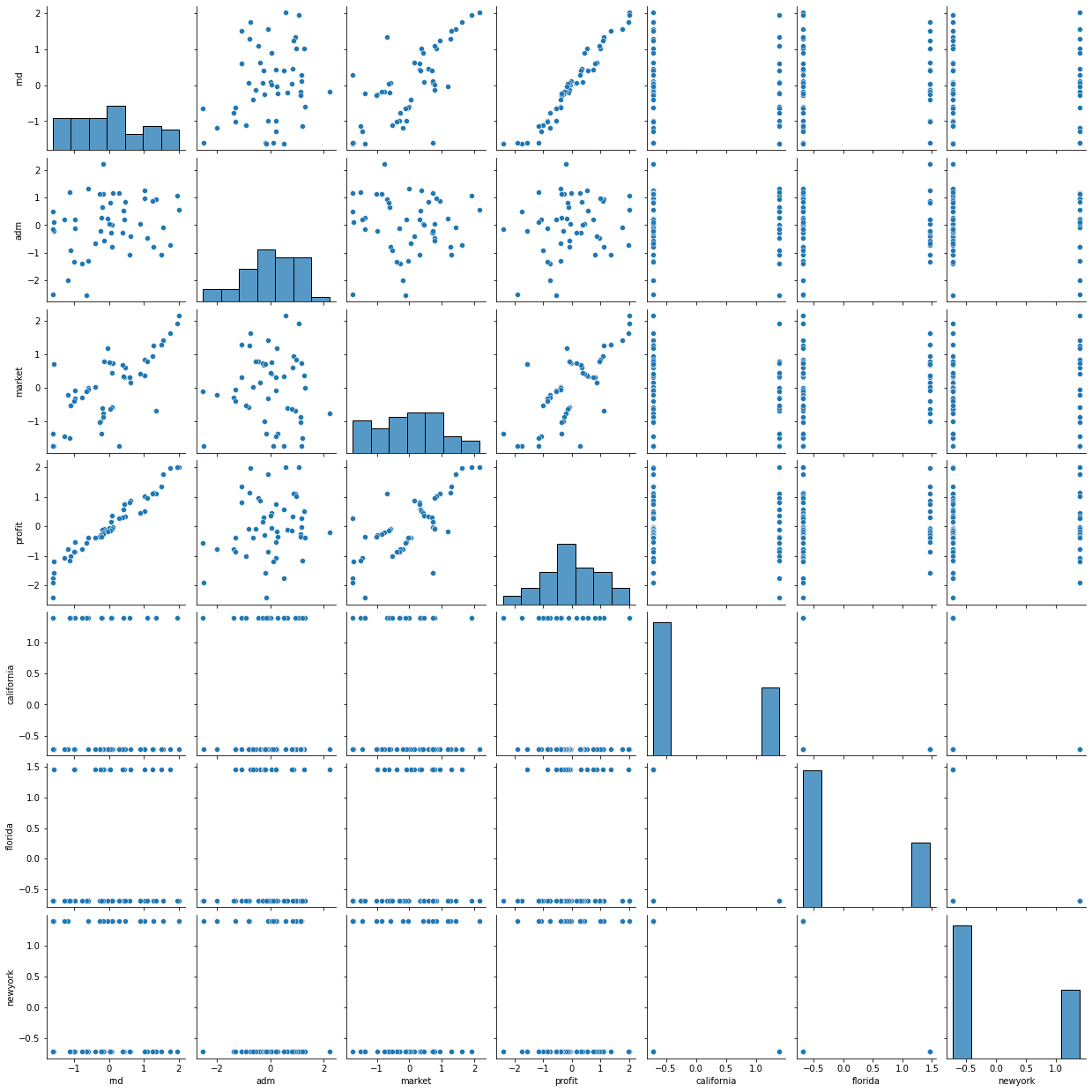
This is the Q-Q plot for the profit, and we can see that data is linear so the data is normally distributed. We check this for rest of the variables also.

This is the jointplot which is the combination of univariate visualization and bivariate visualization.

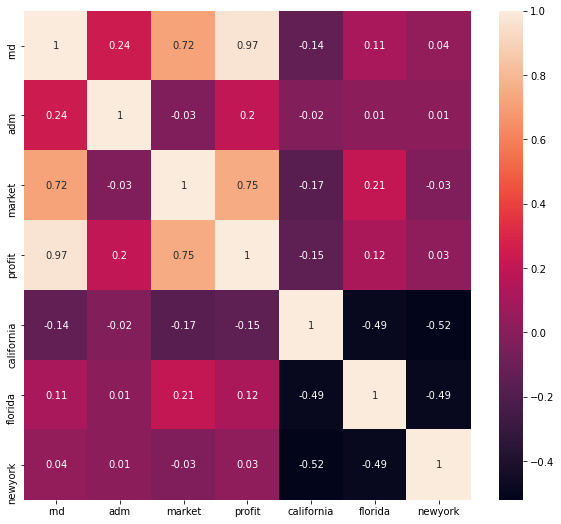
Here we can see the scatter plot between the adm and profit and the histogram for adm and profit also from which we can see the data is normally distributed or not and the correlation between two variables. And like this we plot the jointplot for all the variables.

The above graph is the count plot which shows the counts of the each unique values of the particular variable.

* Standardize the data as to bring the magnitude of the data at the same level.



The above graph is the pair plot which shows the correlation between independent variables and dependent variables and in between different independent variables .



The above graph is the heatmap which shows the correlation matrix from which we can see the correlation between different variables based on the coefficient of correlation value.

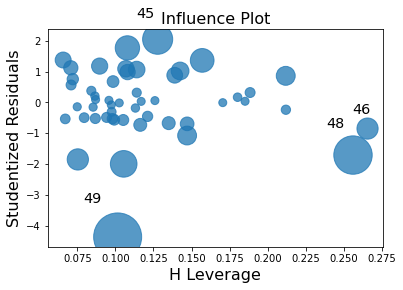
# From the heatmap we can say that RND spend and profit have the strong correlation.

# marketing\_spend and profit having the moderate correlation

# Administration and profit having the very less correlation

# Market and rnd having the moderate collinearity

# states column does not correlated with profit so we don’t consider state column for analysis

* Preparing multilinear regression model by considering all the variables and view the summary.
* From the summary we have to see the p value of the coefficient of each variable which should be less that 0.05 and the R2 value should be > 0.85.
*  Checking whether data has any influential values and it is checked by Influence Index Plots

From the graph we can see that index 46, 48and 49 is showing high influence so we can exclude that entire rows and then fit the model and check the summary for better accuracy.( to bring the p values below 0.05)

* If still the model is not fitting properly and the p values is not coming below 0.05 then we have to check the collinearity between independent variables by calculating Variance Inflation factor (VIF) for all the variables and as the R2 value increases the VIF value increases.

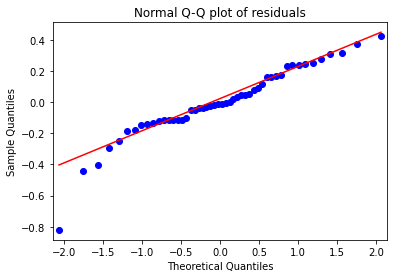
Assumption: VIF > 10 = collinearity

As we can see that VIF for state column is infinite so we confirm that collinearity does not exist between them

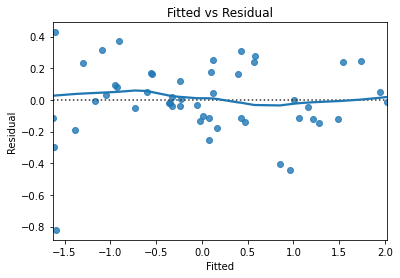
From the above correlation table and VIF values we can say that the States and Administration are not significant variables for predicting the Profit values.

We will build a model using R&D Spend(rnd) and Marketing Spend(market)

* Creating the model by using rnd and marketing variables and again check the p value and r2 values.
* Prediction to be done on the entire dataset.
* Draw the Q-Q plot to check the normality of errors

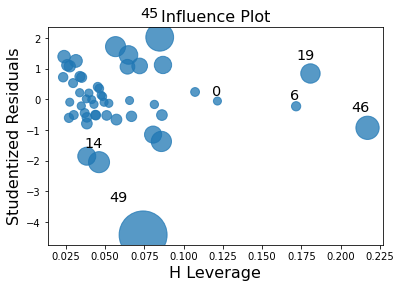


From the graph we can see that the errors are normally distributed.

* Draw the residual plot to check the homoscedasticity between the errors.

There is no pattern in the Residual plot, so we can say there is a linear relation and no error variance

* detecting influencers, outliers by plotting the influential plot.

We can see that the 49th observation is an Influencer point and has more distance than other data points.

We will delete this data point to further increase our accuracy.

* Build the Model. For building the model we split the data into train data and test data and in the 80:20 ration and then fit the final model on the train dataset and evaluate on the test dataset and calculate the RMSE value for the test datasets.
* Again, evaluate the model on the train datasets and calculate the root mean squared values (RMSE) value for the train datasets and compare the values between train datasets and test datasets ands finalized the model. As the rmse value for both the training dataset and test dataset are almost equal then we can say that the model evaluated perfectly.

1. Perform multilinear regression with price as the output variable and document the different RMSE values.



Ans: **Business Objective:**

To build model to predict the price of the computer best on the other attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| price | Price of the computer | Quantitative, Ratio | Relevant |
| speed | Speed of the computer i,e, processor | Quantitative, Ratio | Relevant |
| hd | hd resolution | Quantitative, Ratio | Relevant |
| ram | Ram of the computer | Quantitative, Ratio | Relevant |
| screen | Screen size of the computer | Quantitative, Ratio | Relevant |
| cd | Is cd can be fit or not | Qualitative, nominal | Relevant |
| multi | Is computer multipurpose or not | Qualitative, nominal | Relevant |
| premium | Is computer premium or not | Qualitative, nominal | Relevant |
| ads | Alternate data stream | Quantitative, Ratio | Relevant |
| trend | Is computer trending or not | Quantitative, Nominal | Relevant |

Steps For the multilinear Regression:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* **Data Visualization**

Histogram: To check the distribution of the data

Distribution plot: To check the data is normally distributed or not but in this we get the bell shaped curve.

Boxplot: To check the outliers present in the dataset.

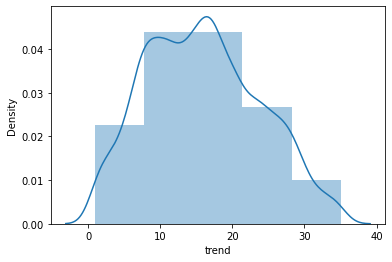
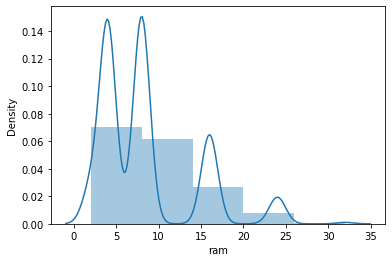
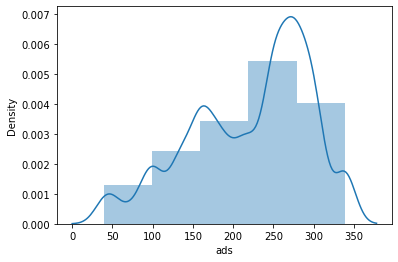
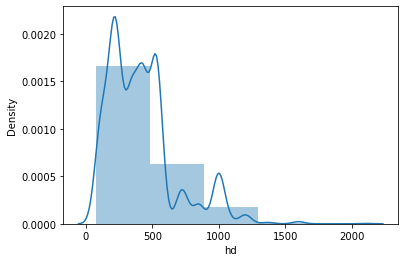
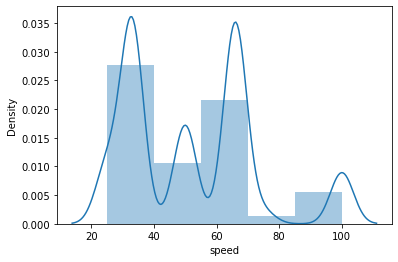
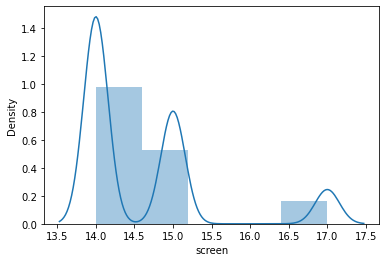
Countplot: It is used to check the count of each unique features each column

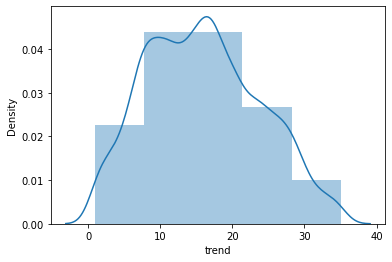
Normal Q\_Q plot: To check the data is normally distributes or not

Scatter Plot: Scatter Plot is used for the bivariate visualization.

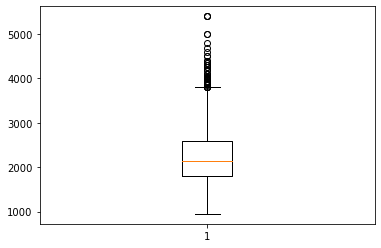
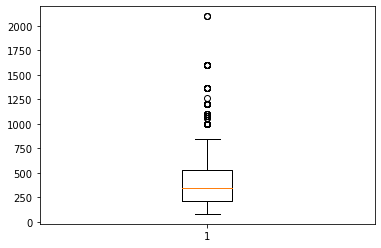
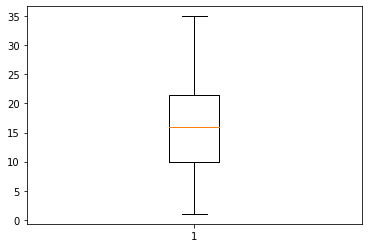
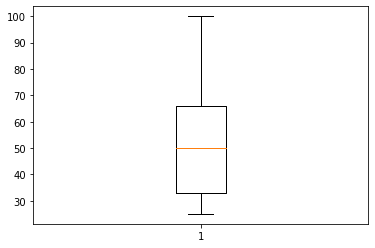
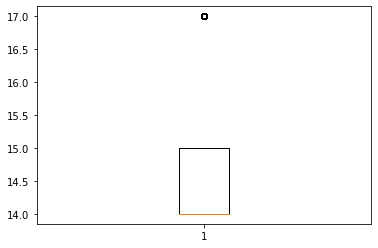
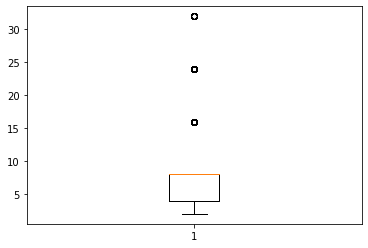
Paiplot : Pairplot is the combination of scatter plot between different variables and the histogram, It is used to find the correlation of the independent variables with the dependent variables and to find out the collinearity between different independent variables.

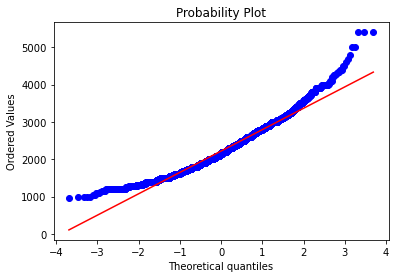
ProbPlot ; probplot is used to find the relationship between the two different variables.



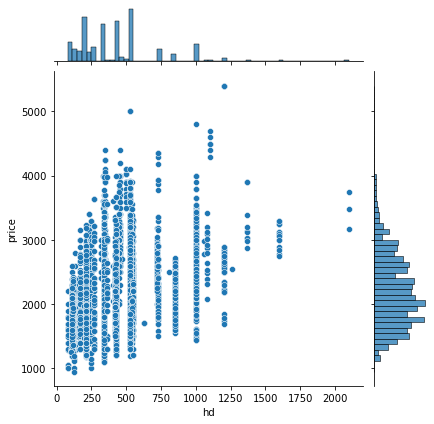


The first and last graph is almost normally distributed, 2nd ,3rd,4th , 5th graph is not normally distributed and is right skewed and 6thrd graph is also not normally distributed and has left skewed.



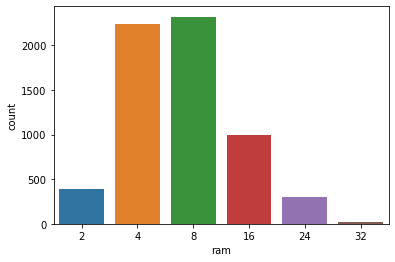
From the above plot we can see that in almost all the boxplots outliers are present so we treat the outliers by using winsorization technique.

This is the Q-Q plot for the price, and we can see that data is linear so the data is normally distributed. We check this for rest of the variables also.

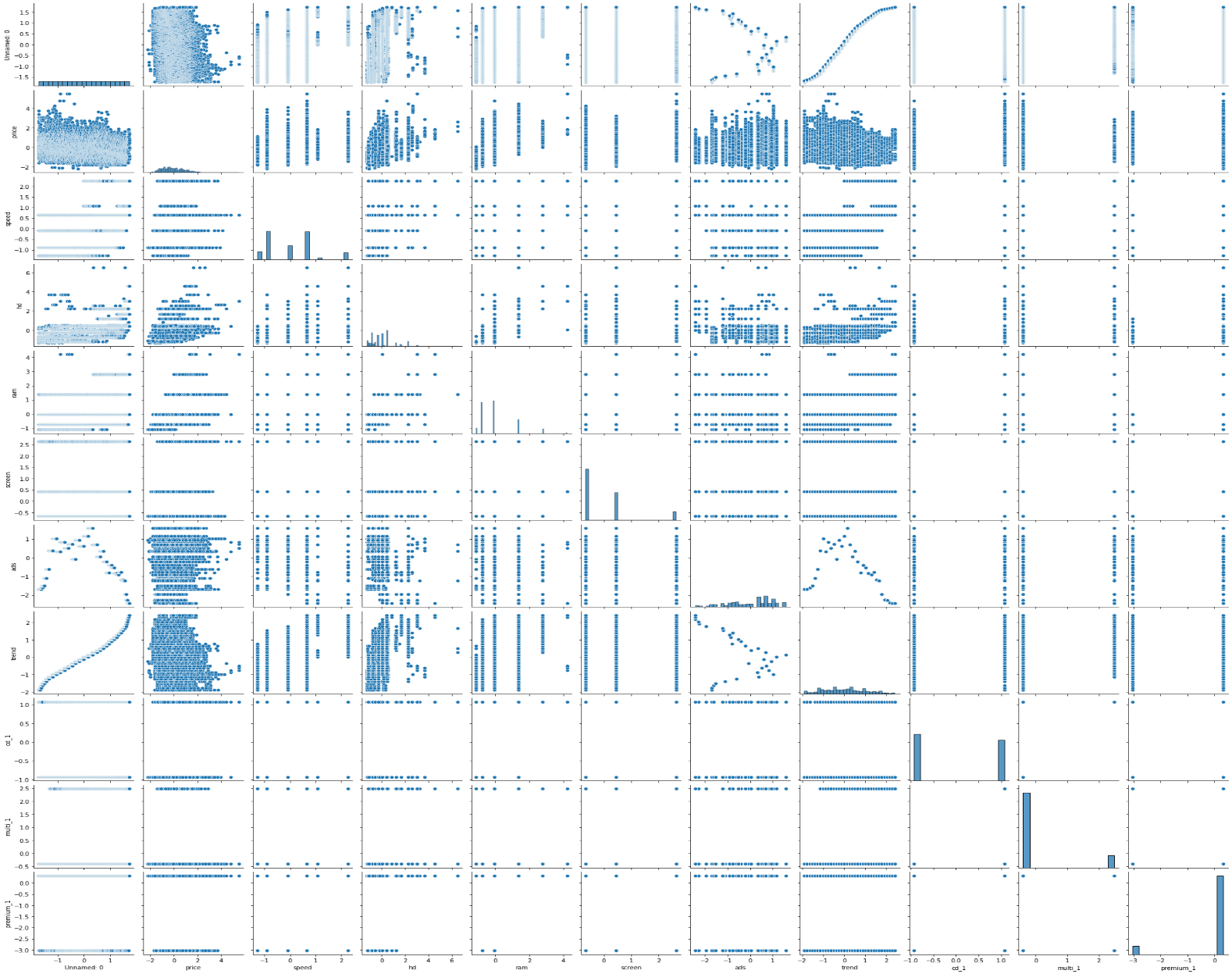


This is the joint plot which is the combination of univariate visualization and bivariate visualization.

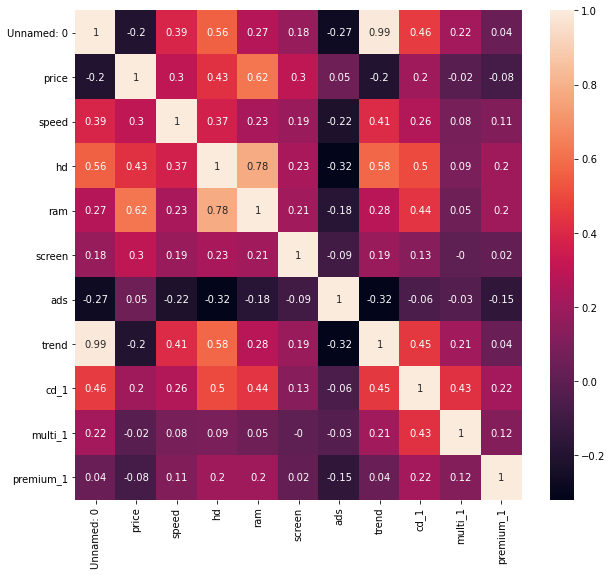
Here we can see the scatter plot between hd and price and the histogram for hd and price also from which we can see the data is normally distributed or not and the correlation between two variables. And like this we plot the joint plot for all the variables.



The above graph is the count plot which shows the counts of the each unique values of the particular variable.

* Standardize the data as to bring the magnitude of the data at the same level.

The above graph is the pair plot which shows the correlation between independent variables and dependent variables and in between different independent variables.



The above graph is the heatmap which shows the correlation matrix from which we can see the correlation between different variables based on the coefficient of correlation value.

# From the table we can say that ram and hd are moderately correlated

# Preparing model considering all the variables Preparing multilinear regression model by considering all the variables and view the summary.

* From the summary we have to see the p value of the coefficient of each variable which should be less that 0.05 and the R2 value should be > 0.85.
* Checking whether data has any influential values and it is checked by Influence Index Plots

From the graph we can see that index 46, 48and 49 is showing high influence so we can exclude that entire row and then fit the model and check the summary for better accuracy. (to bring the p values below 0.05)

* If still the model is not fitting properly and the p values is not coming below 0.05 then we have to check the collinearity between independent variables by calculating Variance Inflation factor (VIF) for all the variables and as the R2 value increases the VIF value increases.

Assumption: VIF > 10 = collinearity

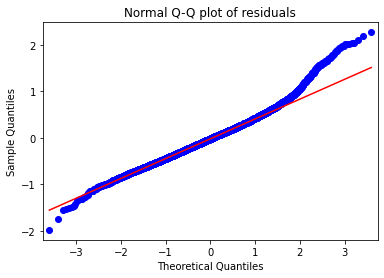
As we can see that VIF for state column is infinite so we confirm that collinearity does not exist between them

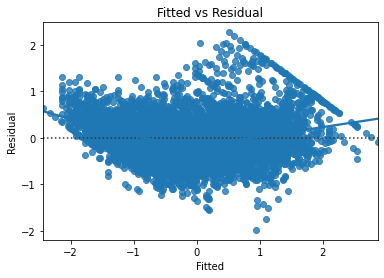
From the above correlation table and VIF values we can say that the States and Administration are not significant variables for predicting the Profit values.

We will build a model using R&D Spend(rnd) and Marketing Spend(market)

* Creating the model by using rnd and marketing variables and again check the p value and r2 values.
* Prediction to be done on the entire dataset.
* Draw the Q-Q plot to check the normality of errors

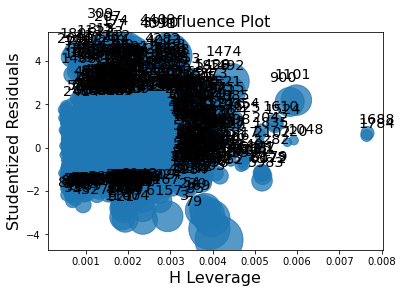
From the graph we can see that the errors are normally distributed.



* Draw the residual plot to check the homoscedasticity between the errors.

There is no pattern in the Residual plot, so we can say there is a linear relation and no error variance

* detecting influencers, outliers by plotting the influential plot.



# We can see that the 5560, 4477, 3783 th observation is an Influencer point and has more distance than other data points. We will delete this data point to further increase our accuracy.

* Build the Model. For building the model we split the data into train data and test data and in the 80:20 ration and then fit the final model on the train dataset and evaluate on the test dataset and calculate the RMSE value for the test datasets.
* Again, evaluate the model on the train datasets and calculate the root mean squared values (RMSE) value for the train datasets and compare the values between train datasets and test datasets ands finalized the model. As the rmse value for both the training dataset and test dataset are almost equal then we can say that the model evaluated perfectly.

1. An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars. For this, they would like an automated model which can predict the price of the car once the user inputs the required factors. Help the business achieve their objective by applying multilinear regression on the given dataset. Please use the below columns for the analysis purpose: price, age\_08\_04, KM, HP, cc, Doors, Gears, Quarterly\_Tax, and Weight.



: **Business Objective:**

To build model to predict the price of the car based on the other attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| Id | Price of the computer | Quantitative, Nominal | Irelevant |
| Model | Model of the car | Qualitative, nominal | Irelevant |
| Price | Offer Price in euros | Quantitative, Ratio | Relevant |
| Age\_08\_04 | Age in months as in August 2004 | Quantitative, Ratio | Relevant |
| Mfg\_Month | Manufacturing month (1-12) | Quantitative, Nominal | Irelevant |
| Mfg\_Year | Manufacturing Year | Qualitative, nominal | Irelevant |
| KM | Accumulated Kilometers on odometer | Quantitative, Ratio | Relevant |
| Fuel\_Type | Fuel Type (Petrol, Diesel, CNG) | Qualitative, nominal | Relevant |
| HP | Horse Power | Quantitative, Ratio | Relevant |
| Met\_Color | Metallic Color? (Yes=1, No=0) | Quantitative, Nominal | Irelevant |
| Color | Color (Blue, Red, Grey, Silver, Black, etc.) |  | Irelevant |
| Automatic | Automatic ( (Yes=1, No=0) |  | Irelevant |
| Cc | Cylinder Volume in cubic centimeters | Quantitative, Ratio | Irelevant |
| Doors | Number of doors | Quantitative, Ratio | Relevant |
| Cylinders | - Number of cylinders | Quantitative, Ratio | Relevant |
| Gears | Number of gear positions | Quantitative, Ratio | Relevant |
| Quarterly\_Tax | Quarterly road tax in euros | Quantitative, Ratio | Relevant |
| Weight | Weight in Kilograms | Quantitative, Ratio | Irelevant |
| Mfr\_Guarantee | Within Manufacturer's Guarantee period (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| BOVAG\_Guarantee | BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Guarantee\_Period | Guarantee period in months | Qualitative,Nominal | Irelevant |
| ABS | Anti-Lock Brake System (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Airbag\_1 | Driver\_Airbag (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Airbag\_2 | Passenger Airbag (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Airco | Airconditioning (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Automatic\_airco | Automatic Airconditioning (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Boardcomputer | Boardcomputer (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| CD\_Player | CD Player (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Central\_Lock | Central Lock (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Powered\_Windows | Powered Windows (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Power\_Steering | Power Steering (Yes=1, No=0 | Qualitative,Nominal | Irelevant |
| Radio | Radio (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Mistlamps | Mistlamps (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Sport\_Model | Sports3 Model (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Backseat\_Divider | Backseat Divider (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Metallic\_Rim | Metallic Rim (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Radio\_cassette | Radio Cassette (Yes=1, No=0) | Qualitative,Nominal | Irelevant |
| Tow\_Bar | Tow Bar (Yes=1, No=0) | Qualitative,Nominal | Irelevant |

Steps For the multilinear Regression:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* **Data Visualization**

Histogram: To check the distribution of the data

Distribution plot: To check the data is normally distributed or not but in this we get the bell shaped curve.

Boxplot: To check the outliers present in the dataset.

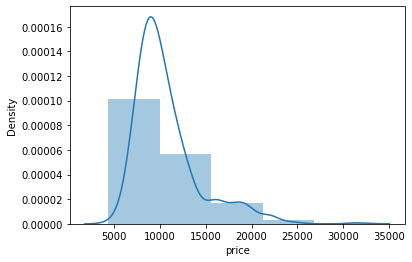
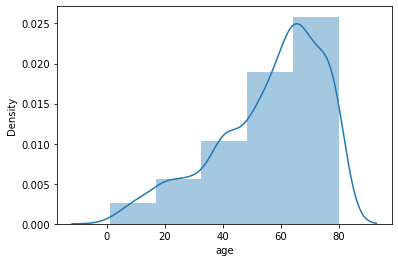
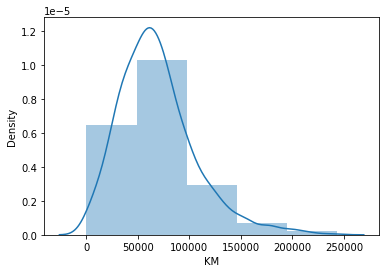
Countplot: It is used to check the count of each unique features each column

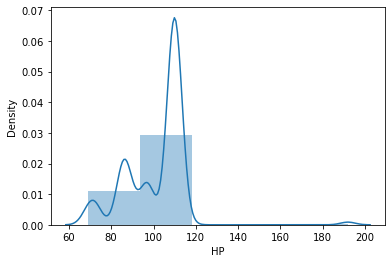
Normal Q\_Q plot: To check the data is normally distributes or not

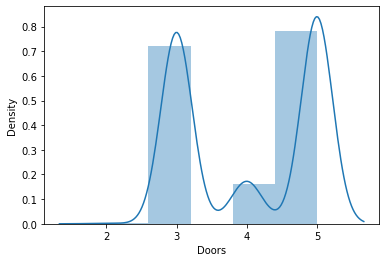
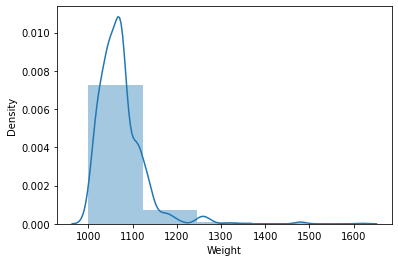
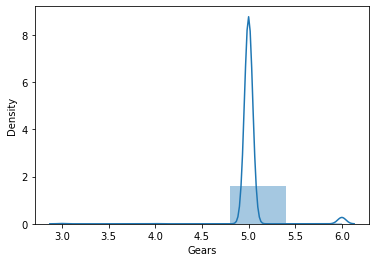
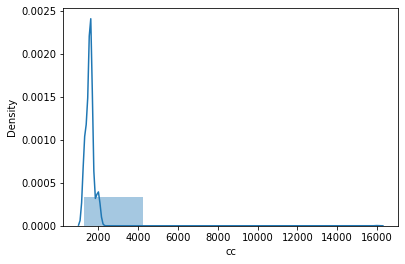
Scatter Plot: Scatter Plot is used for the bivariate visualization.

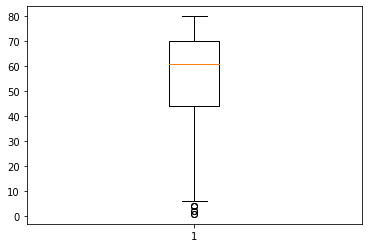
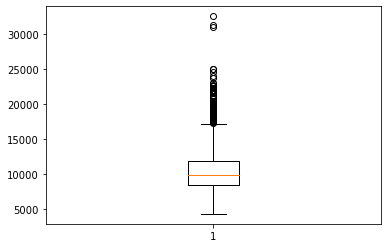
Paiplot : Pairplot is the combination of scatter plot between different variables and the histogram, It is used to find the correlation of the independent variables with the dependent variables and to find out the collinearity between different independent variables.

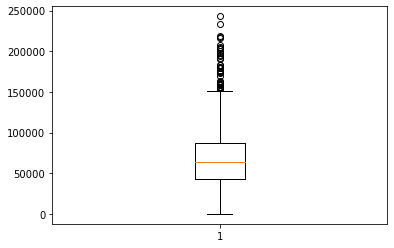
ProbPlot ; probplot is used to find the relationship between the two different variables.

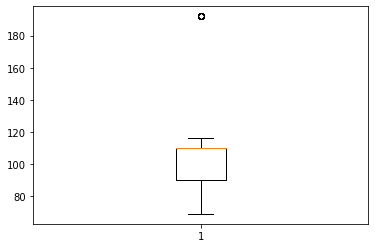


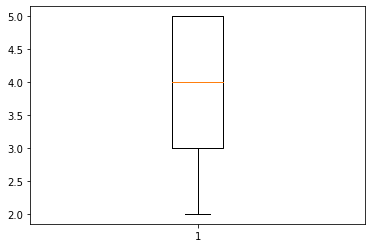
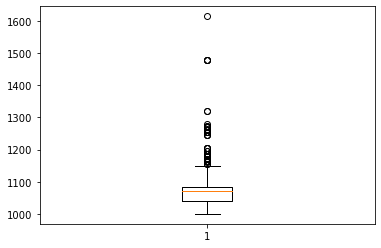
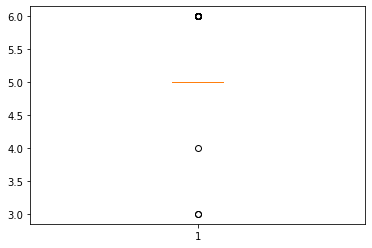
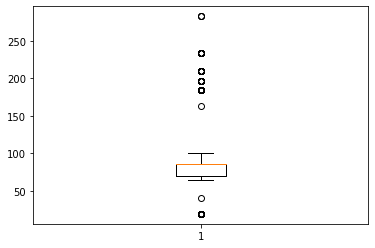
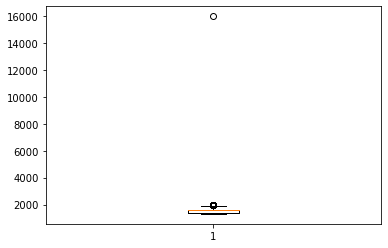
****

None of the variables having the normal distribution.

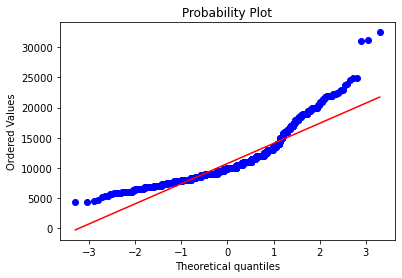








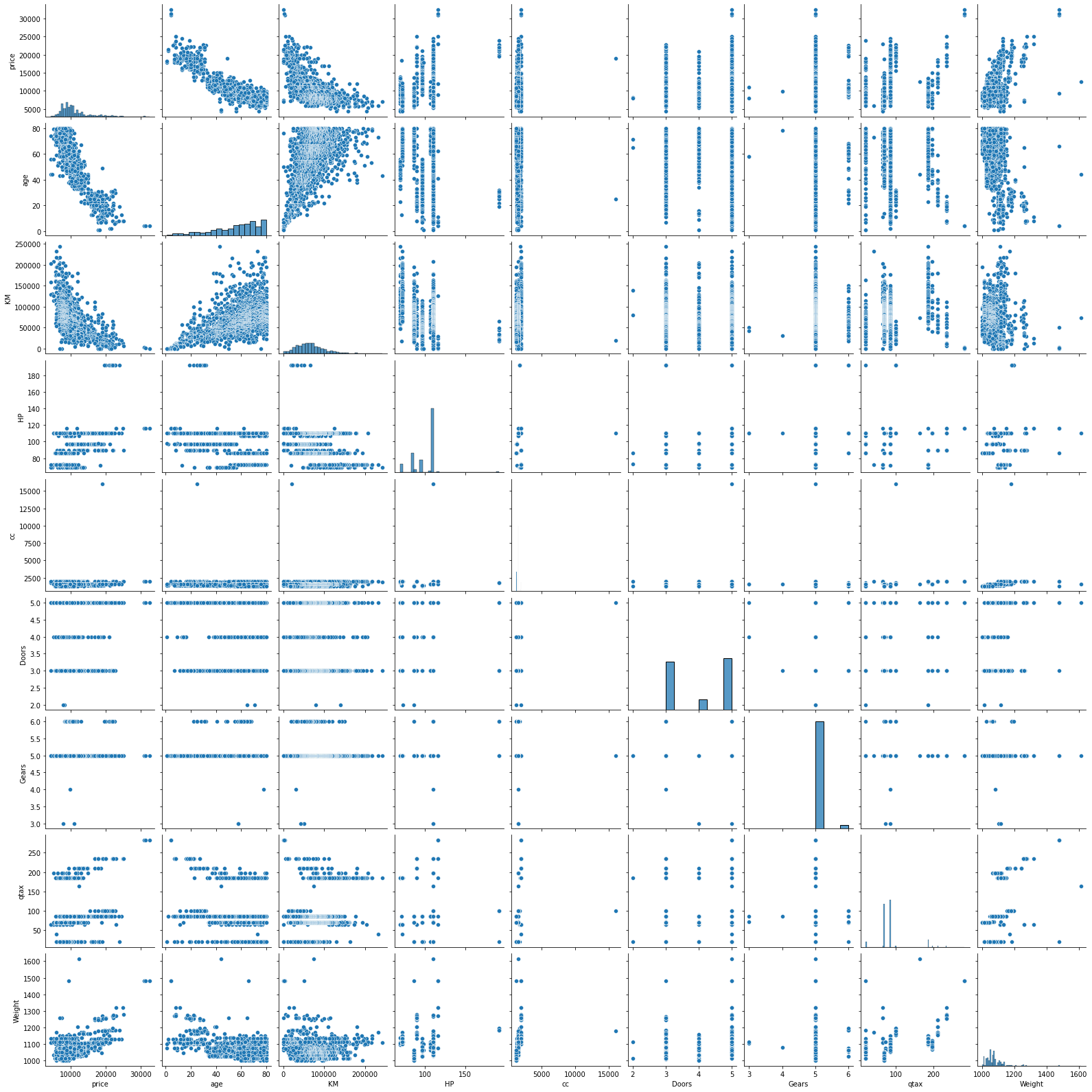
From the above plot we can see that in almost all the boxplots’ outliers are present so we treat the outliers by using winsorization technique.



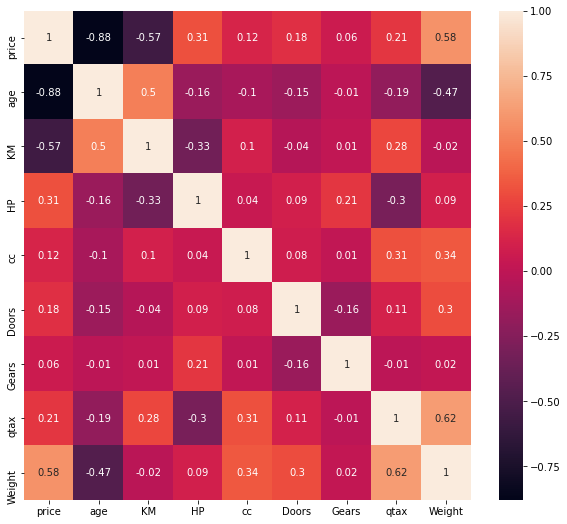
This is the Q-Q plot for the price, and we can see that data is linear so the data is normally distributed. We check this for rest of the variables also.

The above graph is the count plot which shows the counts of the each unique values of the particular variable.

* Standardize the data as to bring the magnitude of the data at the same level.



The above graph is the pair plot which shows the correlation between independent variables and dependent variables and in between different independent variables.



The above graph is the heatmap which shows the correlation matrix from which we can see the correlation between different variables based on the coefficient of correlation value.

# From the pairplot we can say that age is strongly correlated with price

# KM, Weight are moderately correlated with price

# age and Km have the some collinearity

# cc and qtax has some collinearity and cc and weight has moderate collinearity

# qtax and weight has moderate collinearity From the summary we have to see the p value of the coefficient of each variable which should be less that 0.05 and the R2 value should be > 0.85.

* Checking whether data has any influential values and it is checked by Influence Index Plots

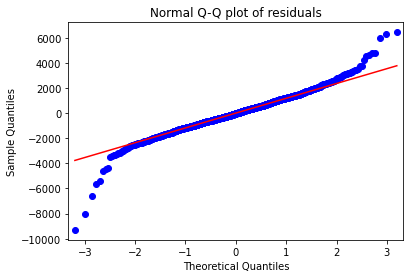
From the graph we can see that index 955 and 990 is showing high influence so we can exclude that entire row and then fit the model and check the summary for better accuracy. (To bring the p values below 0.05)

* If still the model is not fitting properly and the p values is not coming below 0.05 then we have to check the collinearity between independent variables by calculating Variance Inflation factor (VIF) for all the variables and as the R2 value increases the VIF value increases.

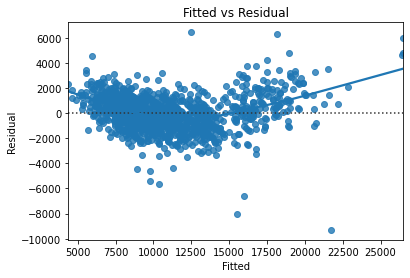
Assumption: VIF > 10 = collinearity

* Creating the model by using age, KM ,HP, cc, Doors, Gears, qtax, Weight,' again check the p value and r2 values.
* Prediction to be done on the entire dataset.
* Draw the Q-Q plot to check the normality of errors

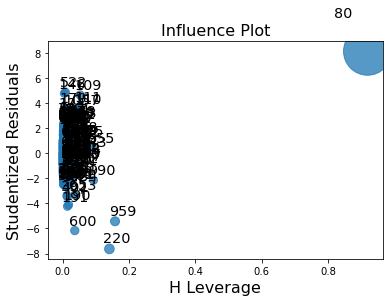
From the graph we can see that the errors are normally distributed.



* Draw the residual plot to check the homoscedasticity between the errors.



There is no pattern in the Residual plot, so we can say there is a linear relation and no error variance

* detecting influencers, outliers by plotting the influential plot.

# We can see that the 5560, 4477, 3783 th observation is an Influencer point and has more distance than other data points. We will delete this data point to further increase our accuracy.

* Build the Model. For building the model we split the data into train data and test data and in the 80:20 ration and then fit the final model on the train dataset and evaluate on the test dataset and calculate the RMSE value for the test datasets.
* Again, evaluate the model on the train datasets and calculate the root mean squared values (RMSE) value for the train datasets and compare the values between train datasets and test datasets ands finalized the model. As the rmse value for both the training dataset and test dataset are almost equal then we can say that the model evaluated perfectly.

1. With the growing consumption of avocados in the USA, a freelance company would like to do some analysis on the patterns of consumption in different cities and would like to come up with a prediction model for the price of avocados. For this to be implemented, build a prediction model using multilinear regression and provide your insights on it.

Snapshot of the dataset is given below: -

A close up of a piece of paper

Description automatically generated

Ans: **Business Objective:**

To build model to predict the price of the avocado based on the other attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| AveragePrice | the average price of a single avocado | Quantitative, Ratio | Relevant |
| Total\_Volume | Total number of avocados sold | Quantitative, Ratio | Relevant |
| tot\_ava1 | total number of small avocados sold | Quantitative, Ratio | Relevant |
| tot\_ava2 | total number of large avocados sold | Quantitative, Ratio | Relevant |
| tot\_ava3 | total number of xtra large avocados sold | Quantitative, Ratio | Relevant |
| Total\_Bags | Is cd can be fit or not | Qualitative, nominal | Relevant |
| Small\_Bags | Total number of Small Bags sold | Qualitative, nominal | Relevant |
| Large\_Bags | Total number of Large Bags sold | Qualitative, nominal | Relevant |
| XLarge Bags | Total number of XLarge Bags sold | Quantitative, Ratio | Relevant |
| type | conventional or organic | Qualitative, Nominal | Relevant |
| year | the year | Quantitative, Nominal | Relevant |
| region | the city or region of the observation | Qualitative, Nominal | Relevant |

Steps For the multilinear Regression:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel :- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* **Data Visualization**

Histogram: To check the distribution of the data

Distribution plot: To check the data is normally distributed or not but in this we get the bell shaped curve.

Boxplot: To check the outliers present in the dataset.

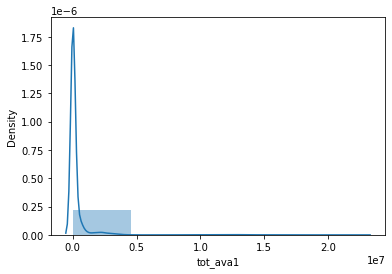
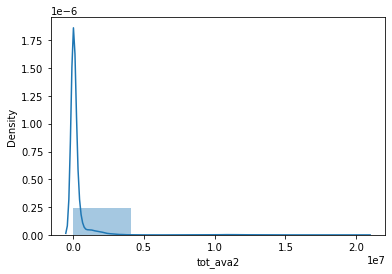
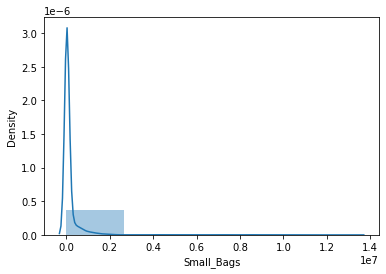
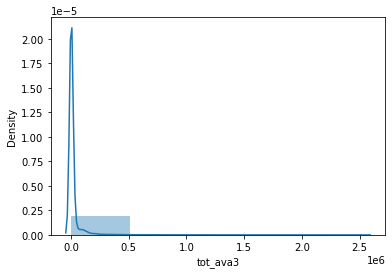
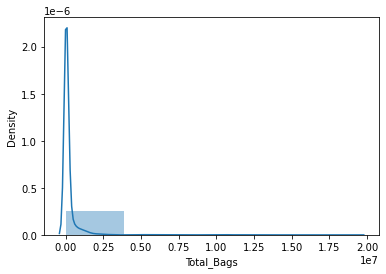
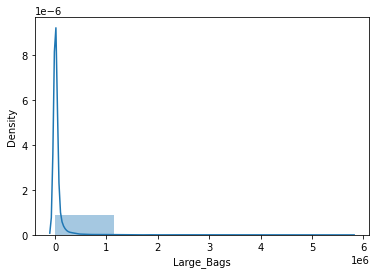
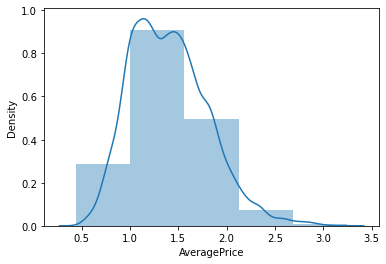
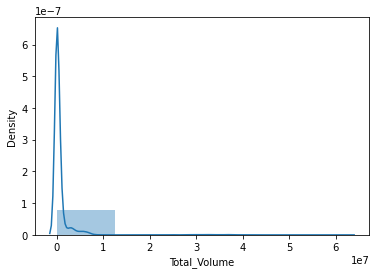
Countplot: It is used to check the count of each unique features each column

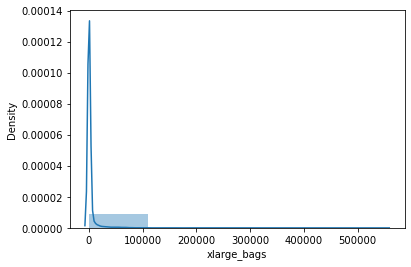
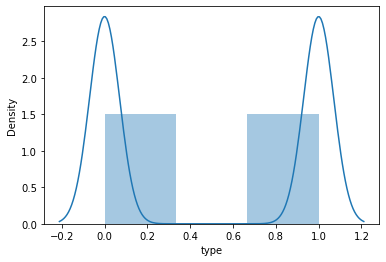
Normal Q\_Q plot: To check the data is normally distributes or not

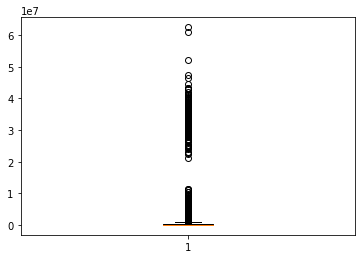
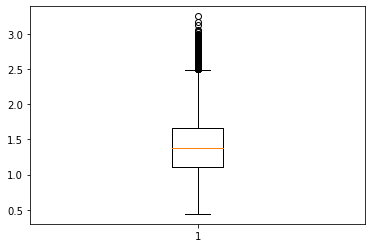
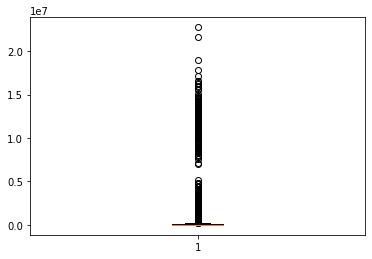
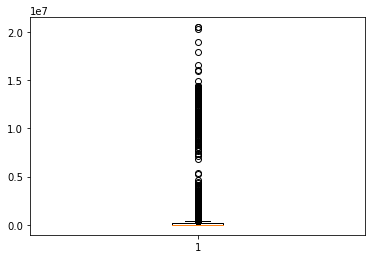
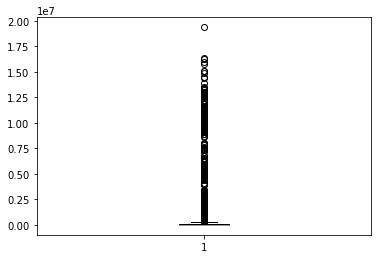
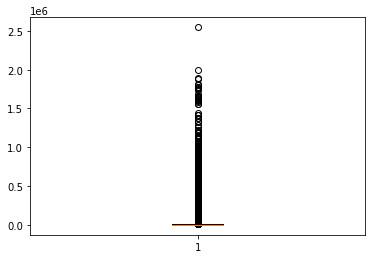
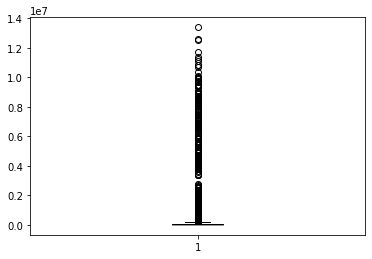
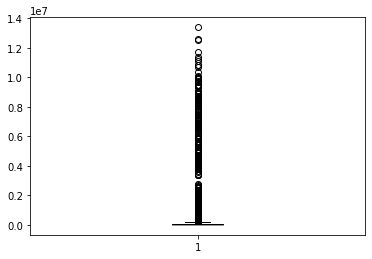
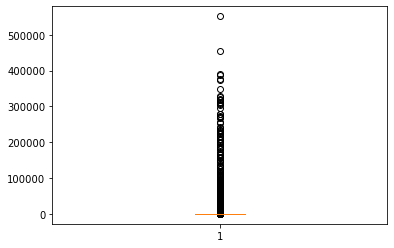
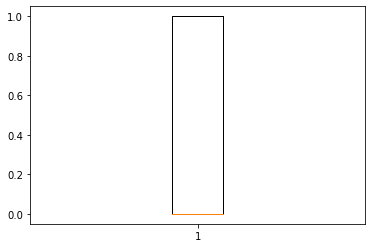
Scatter Plot: Scatter Plot is used for the bivariate visualization.

Paiplot : Pairplot is the combination of scatter plot between different variables and the histogram, It is used to find the correlation of the independent variables with the dependent variables and to find out the collinearity between different independent variables.

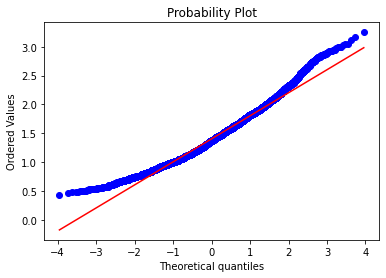
ProbPlot ; probplot is used to find the relationship between the two different variables.



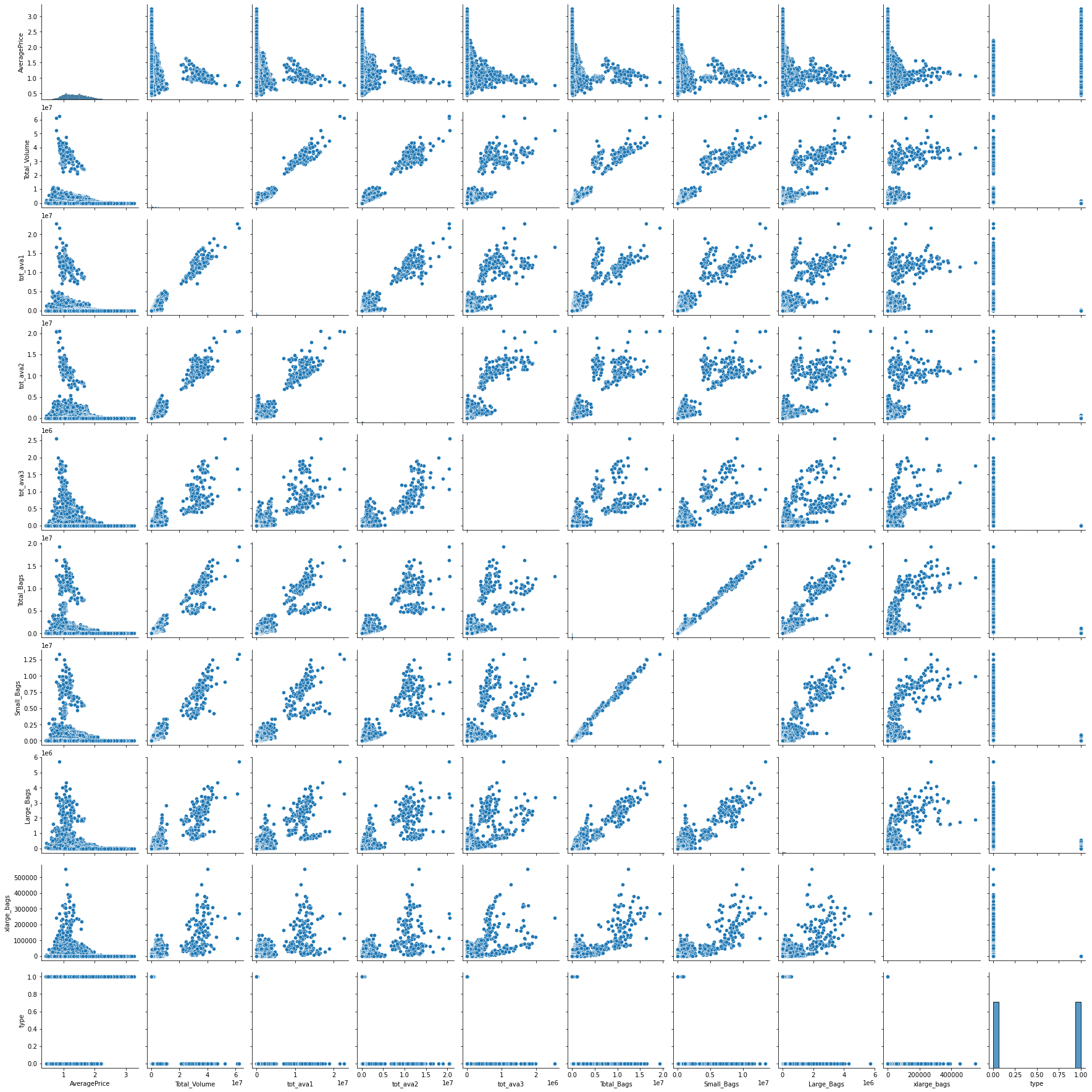
The first and last graph is almost normally distributed, 2nd ,3rd,4th , 5th graph is not normally distributed and is right skewed an d 6thrd graph is als o not normally distributed and has left skewe  d.

From the above plot we can see that in almost all the boxplots except last one outliers are present so we treat the outliers by using winsorization technique

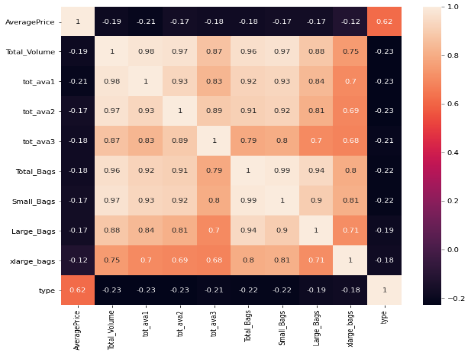


This is the Q-Q plot for the price, and we can see that data is linear so the data is normally distributed. We check this for rest of the variables also.

* Standardize the data as to bring the magnitude of the data at the same level.



The above graph is the pair plot which shows the correlation between independent variables and dependent variables and in between different independent variables.



The above graph is the heatmap which shows the correlation matrix from which we can see the correlation between different variables based on the coefficient of correlation value.

# the collinearity between the all the independent variable is high

# Total\_Volume has higher collinearity with tot\_ava2, Total\_Bags, Small\_bags

# total bag and small bags has the highest colliniearity. From the summary we have to see the p value of the coefficient of each variable which should be less that 0.05 and the R2 value should be > 0.85.

Checking whether data has any influential values and it is checked by Influence Index Plot.

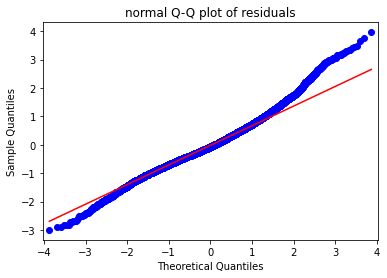
* If still the model is not fitting properly and the p values is not coming below 0.05 then we have to check the collinearity between independent variables by calculating Variance Inflation factor (VIF) for all the variables and as the R2 value increases the VIF value increases.

Assumption: VIF > 10 = collinearity

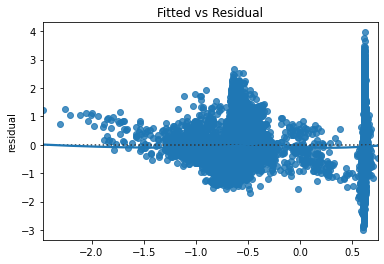
total\_volume and total\_bags has the highest VIF values and we see that the colinearity between total volume and total bags is very high significant variables for predicting the Profit values.

We will build a model using Total\_Volume, tot\_ava1 , tot\_ava2 ,,tot\_ava3, Small\_Bags, Large\_Bags, xlarge\_bags ,type

* Creating the model by using rnd and marketing variables and again check the p value and r2 values.
* Prediction to be done on the entire dataset.
* Draw the Q-Q plot to check the normality of errors



From the graph we can see that the errors are normally distributed.



* Draw the residual plot to check the homoscedasticity between the errors.

There is no pattern in the Residual plot, so we can say there is a linear relation and no error variance

* detecting influencers, outliers by plotting the influential plot.
* Build the Model. For building the model we split the data into train data and test data and in the 80:20 ration and then fit the final model on the train dataset and evaluate on the test dataset and calculate the RMSE value for the test datasets.
* Again, evaluate the model on the train datasets and calculate the root mean squared values (RMSE) value for the train datasets and compare the values between train datasets and test datasets ands finalized the model. As the rmse value for both the training dataset and test dataset are almost equal then we can say that the model evaluated perfectly.