# 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 : DHEERAJ MISHRA**

**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: DHEERAJ MISHRA Batch ID:**  DS\_01072021

**Topic: Multilinear Regression**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Treatment.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**
   2. **Perform Multi linear regression model and check for VIF, AvPlots, Influence Index Plots.**
   3. **Train and Test the data and compare RMSE values. Tabulate R-Squared and RMSE values for different models in the documentation and provide an explanation.**
   4. **Briefly explain the model output in the documentation.**
   5. **Tune the model and improve its accuracy.**
3. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statements: -**

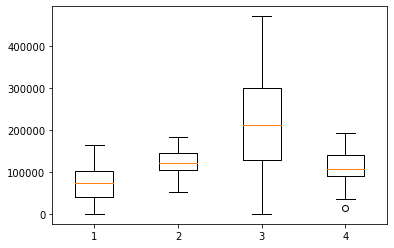
1. An analytics company has been tasked with the crucial job of finding out what factors affect a startup company and if it will be profitable or not. For this, they have collected some historical data and would like to apply multilinear regression to derive brief insights into their data. Predict profit,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

Maximize relationship between multivariables

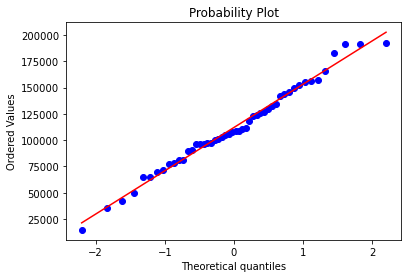
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| R &D Spend | Research and Development team | Continuous | Relevant |
| Administration | Administrative team | Continuous | Relevant |
| Marketing.spend | Marketing spend | Continuous | Relevant |
| State | State name | Discrete | Relevant |
| Profit | Total profit | Continuous | Relevant |

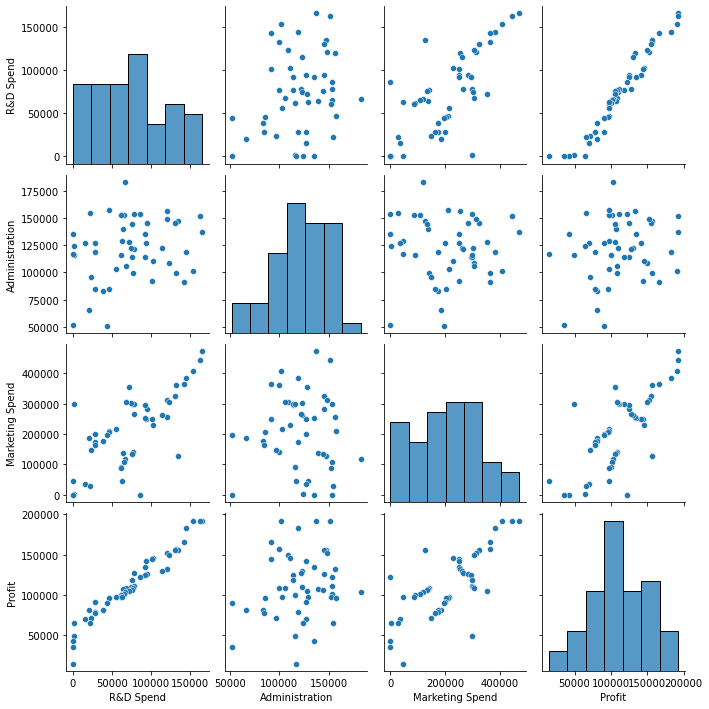
1. DATA CLEANSING :-
2. Dataset consists of 5 colums and 50 rows
3. Duplicate row does not exists
4. All data types are of form int64 and object64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers does not present
8. Dummy column applied for categorical column
9. Corelation coefficients calculated
10. Normalization is used for scaling data
11. There is collinearity problem
12. Some are positively skewed and some are negatively skewed
13. Columns name changed
14. From QQ plot data is normal
15. EDA:-
16. From box plot



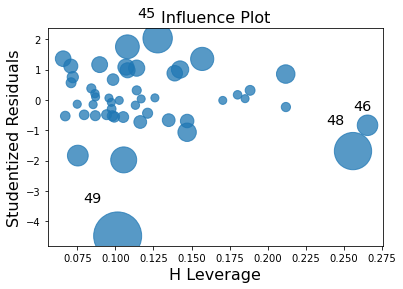
1. From QQ plot



1. From pair plot



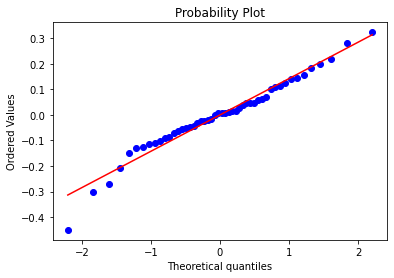
1. MODEL BUILDING:-
2. Model builded but P value is high
3. There is collinearity problem
4. Influence index calculated



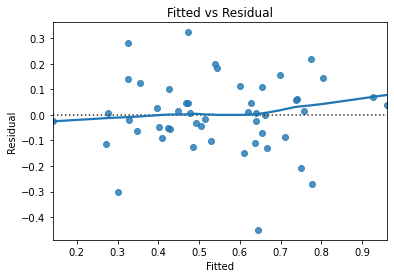
1. Dropping 49 but problem of collinearity not solved
2. Calculating VIF

|  |  |
| --- | --- |
| VARIABLES | VIF |
| PROFIT | 2 20.305593 |
| R& D | 2 19.604937 |
| ADMISTRATION | 2 1.184921 |
| P MARKETING SPEND | 2 2.552858 |
| STATE FLORIDA | 1.387751 |
| STATE NEWYORK | 1.335066 |

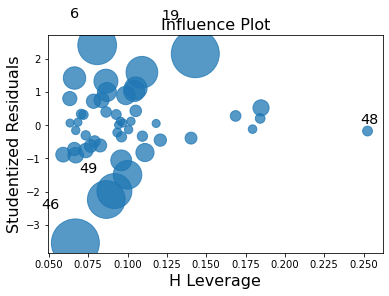
1. Dropping R& D as it has high VIF
2. QQ plot for residual



1. For linearity



1. Influence index for final



OUTPUT:-

1. Best fit model with R2 = 0.613
2. Intercept = 0.1180
3. Administration coefficient = 0.2396
4. Marketing spend coefficient = 0.6661
5. State newyork coefficient = 0.0218
6. State florida coefficient = -0.0096
7. Splitting data to train 80% and test 20%
8. Train accuracy :- 0.09232783330324601
9. Test accuracy:- 0.15031668333814854
10. It is right fit model
11. BENEFITS :-

From above information we can predict for profit against all input variables .

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



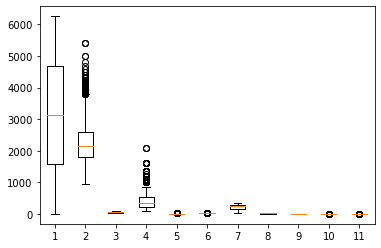
1. BUSINESS OBJECTIVE:-

Maximize relationship between multivariables

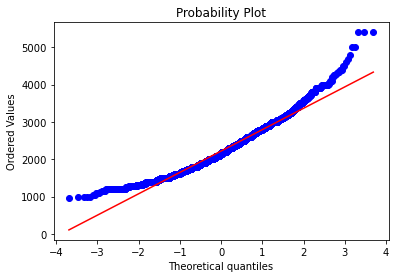
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Index | Index no | Discrete | Not Relevant |
| price | Price of comp | Continuous | Relevant |
| Speed | Speed of computer | Discrete | Relevant |
| Hd | High definition range | Discrete | Relevant |
| Ram | Ram of computer | Discrete | Relevant |
| Screen | Size of screen | Discrete | Relevant |
| Cd | Cd present or not | Binary | Relevant |
| Multi | Multi operational | Binary | Relevant |
| Premium | Premium category | Binary | Relevant |
| ads | Advertisement | Discrete | Relevant |
| trend | Trend of computer | Binary | Relevant |

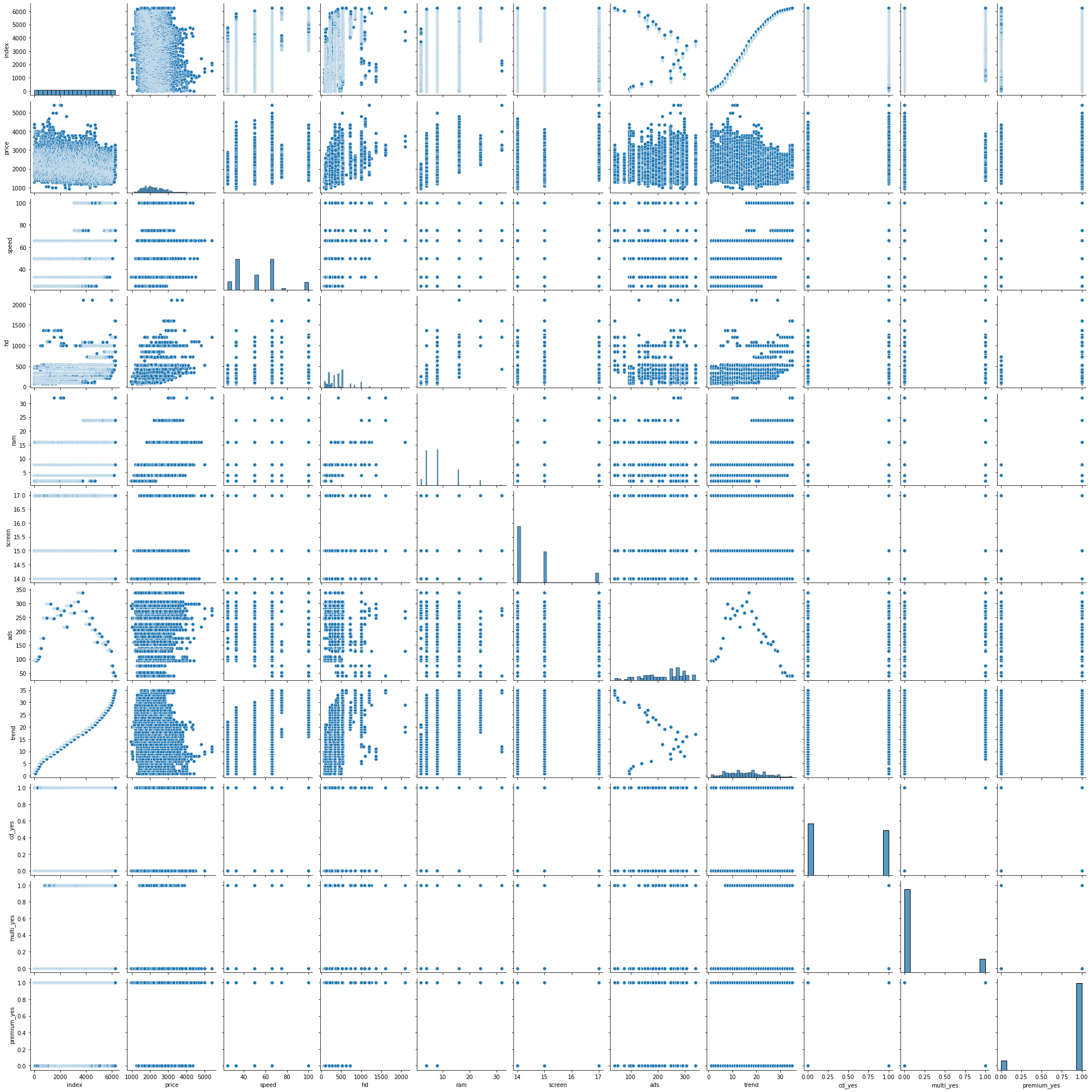
1. DATA CLEANSING :-
2. Dataset consists of 6259 colums and 11 rows
3. Dropping index column
4. Duplicate row does not exists
5. All data types are of form int64 and object
6. No null values found in each column
7. From describe function mean , median and standard deviation obtained
8. Outliers does present and retained
9. Dummy column applied for categorical column
10. Corelation coefficients calculated
11. Normalization is used for scaling data
12. There is notcollinearity problem
13. Some are positively skewed and some are negatively skewed
14. Columns name changed
15. From QQ plot data is normal
16. EDA:-
17. From box plot



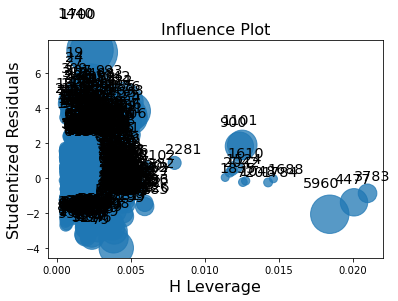
1. From QQ plot



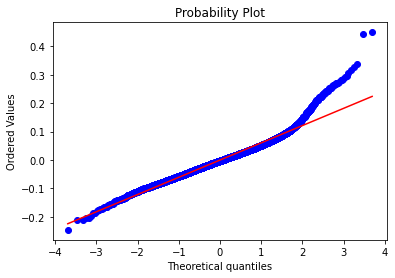
1. From pair plot



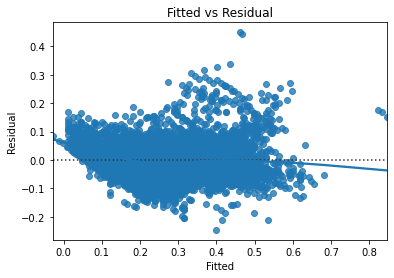
1. MODEL BUILDING:-
2. Model builded
3. There is no collinearity problem
4. Influence index calculated



1. QQ plot for residual



1. For linearity



OUTPUT:-

1. Best fit model with R2 = 0.776
2. Intercept = 0.3254
3. speed coefficient = 0.1571
4. hd coefficient = 0.3549
5. screen coefficient = 0.0830
6. ram coefficient = 0.3253
7. ads coefficient = 0.0443
8. trend coefficient = -0.3962
9. cd coefficient = 0.0137
10. multi coefficient = 0.0234
11. premium coefficient = -0.1144
12. Splitting data to train 80% and test 20%
13. Train accuracy :- 0.06148209501742068
14. Test accuracy:- 0.0632336045026483
15. It is right fit model
16. BENEFITS :-

From above information we can predict for price of computer against all input variables

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.



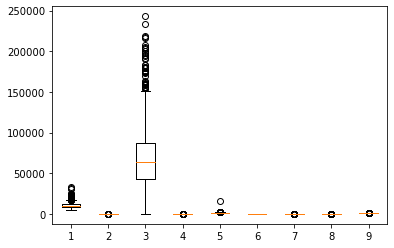
1. BUSINESS OBJECTIVE:-

Maximize relationship between multivariables

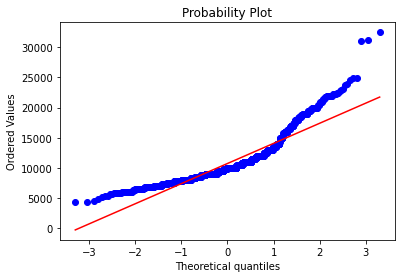
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Price | Car price | Continuous | Relevant |
| Age | Cars age | Discrete | Relevant |
| KM | Kilometer travel | Discrete | Relevant |
| HP | Horse power | Discrete | Relevant |
| Cc | Cc of car | Discrete | Relevant |
| Doors | No of doors | Discrete | Relevant |
| Gears | No of gears | Discrete | Relevant |
| Qt | Quarterly tax | Discrete | Relevant |
| Weight | Weight of car | Continuous | Relevant |

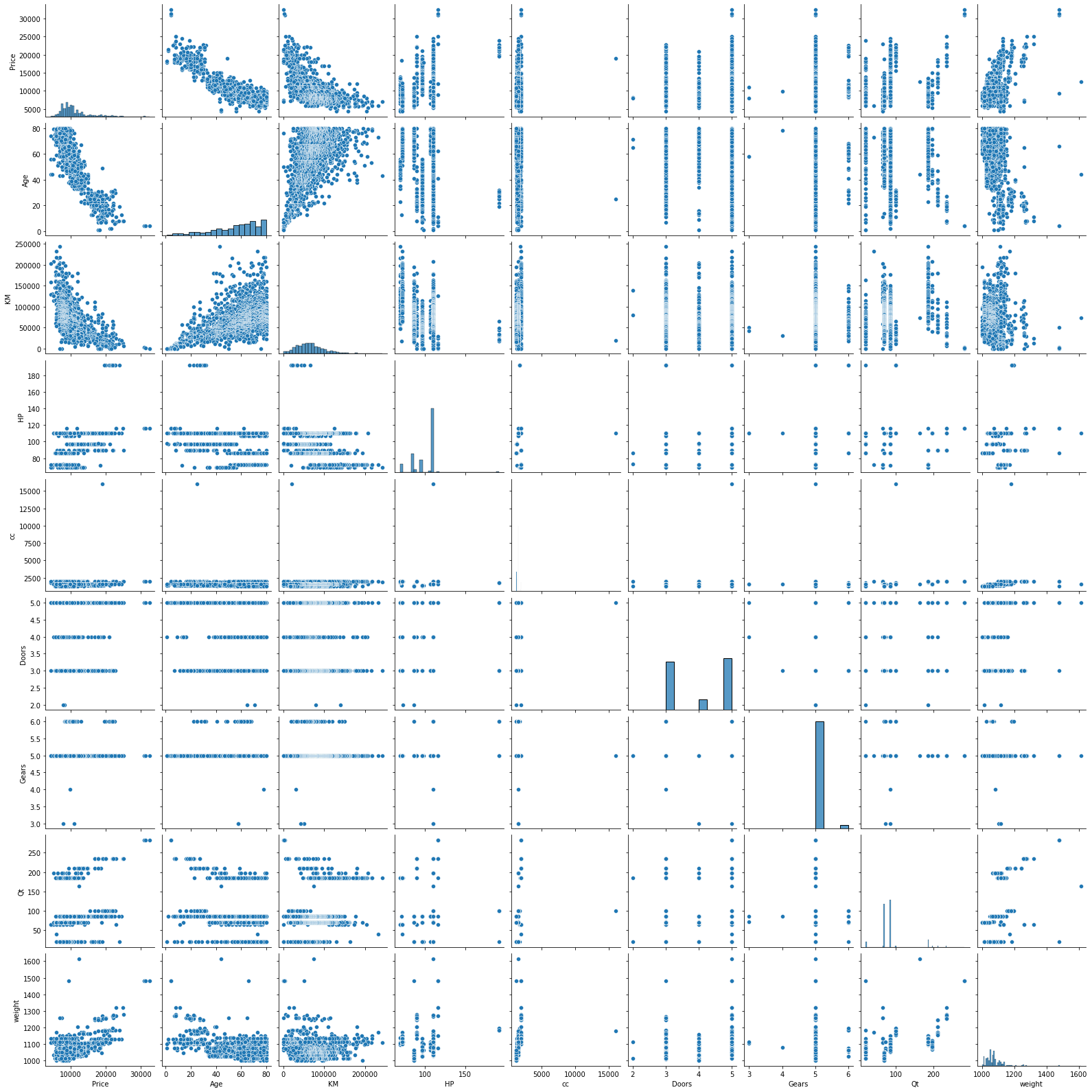
1. DATA CLEANSING :-
2. Dataset consists of 1435 colums and 37 rows
3. Nine columns considered
4. Duplicate row exists and removed
5. All data types are of form int64
6. No null values found in each column
7. From describe function mean , median and standard deviation obtained
8. Outliers present and retained
9. Corelation coefficients calculated
10. Normalization is used for scaling data
11. There is collinearity problem
12. Some are positively skewed and some are negatively skewed
13. From QQ plot data is normal
14. EDA:-
15. From box plot



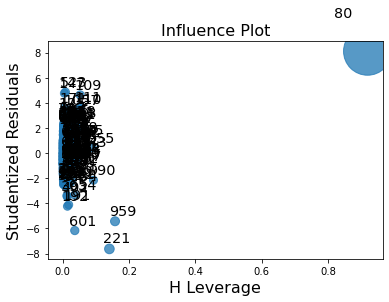
1. From QQ plot



1. From pair plot



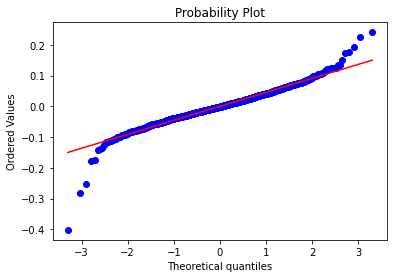
1. MODEL BUILDING:-
2. Model builded but P value is high
3. There is collinearity problem
4. Influence index calculated



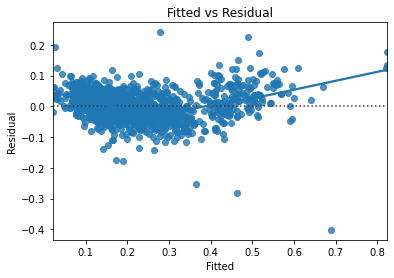
1. Dropping 80 but problem of collinearity not solved
2. Calculating VIF

|  |  |
| --- | --- |
| VARIABLES | VIF |
| Price | 2 7.273382 |
| Age | 2 4.727868 |
| Km | 2 2.095945 |
| P hp | 2 1.543725 |
| cc | 1.164654 |
| Doors | 1.155518 |
| Gears | 1.105944 |
| Qt | 2.309641 |
| weight | 2.919142 |

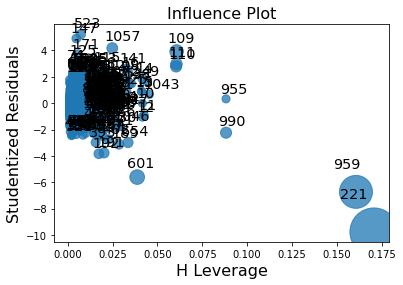
1. Since all vif is low so considending influenced data and removing doors
2. QQ plot for residual



1. For linearity



1. Influence index for final



OUTPUT:-

1. Best fit model with R2 = 0.868
2. Intercept = 0.3712
3. Age coefficient = -0.3382
4. km coefficient = -0.1541
5. hp coefficient = 0.1706
6. cc coefficient = -1.3005
7. grears coefficient = 0.0589
8. Qt coefficient = 0.0841
9. weight coefficient = 0.4340
10. Splitting data to train 80% and test 20%
11. Train accuracy :- 0.049216435651826
12. Test accuracy:- 0.04599425076788963
13. It is right fit model
14. BENEFITS :-

From above information we can predict for price of a car for all input variables .

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

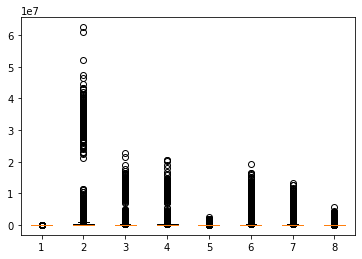
1. BUSINESS OBJECTIVE:-

Maximize relationship between multivariables

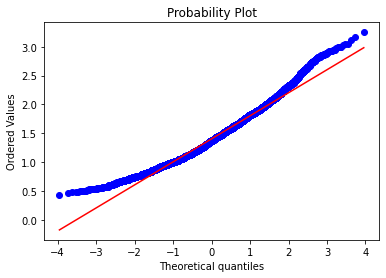
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| AveragePrice | Average price | Continuous | Relevant |
| Total\_Volume | Total volume | Continuous | Relevant |
| tot\_ava1 | Total average 1 | Continuous | Relevant |
| tot\_ava2 | Total average 2 | Continuous | Relevant |
| tot\_ava3 | Total average 3 | Continuous | Relevant |
| Total\_Bags | Total bags | Continuous | Relevant |
| small\_Bags | Small bags | Continuous | Relevant |
| large\_Bags | Large bags | Continuous | Relevant |
| XLarge Bags | Extra large bags | Continuous | Relevant |
| type | Type of avocado | Discrete | Not relevant |
| Year | Year of avocado | Discrete | Not relevant |
| region | Region of avocado | Discrete | Not relevant |

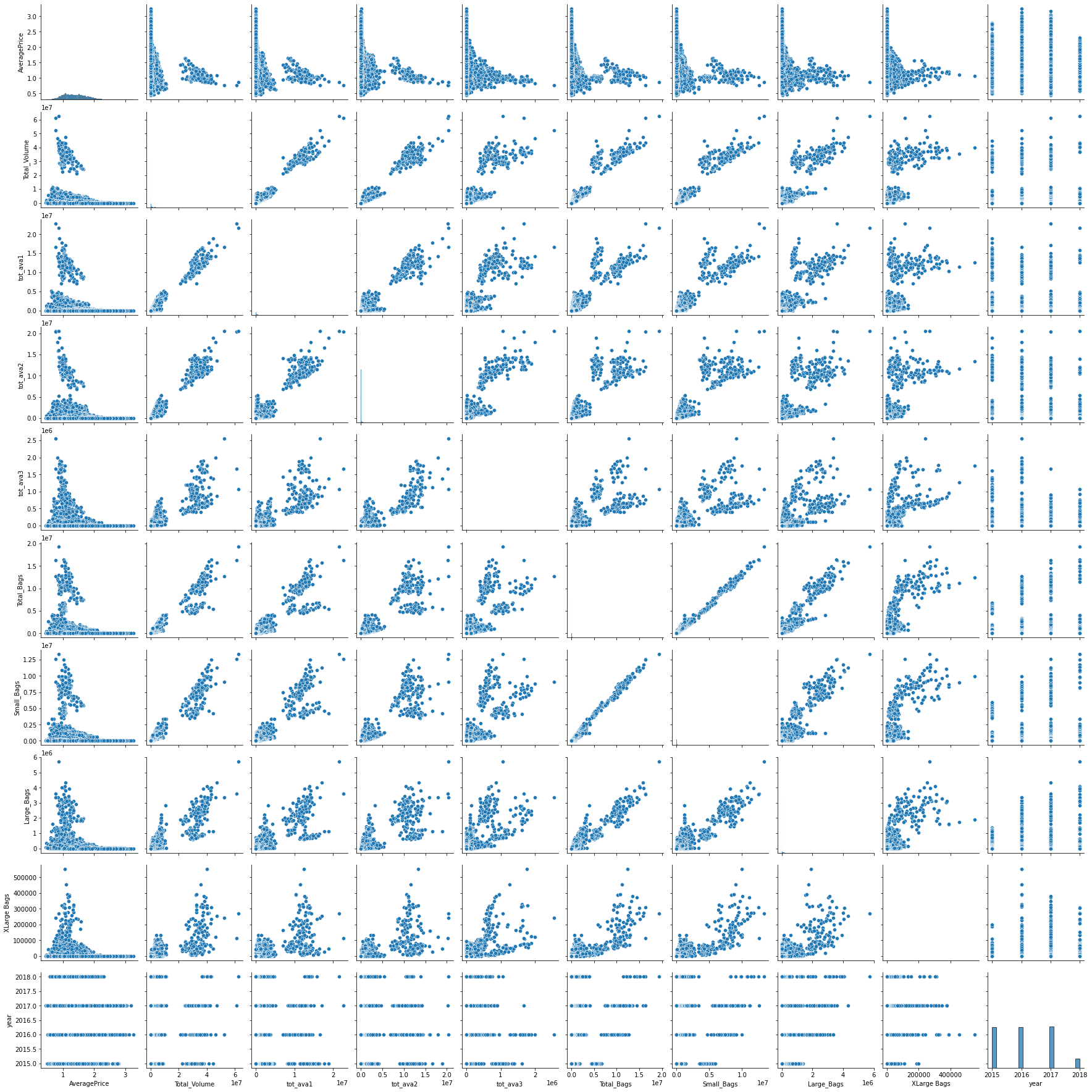
1. DATA CLEANSING :-
2. Dataset consists of 18249 colums and 12 rows
3. Nine columns considered
4. Duplicate row does not exists
5. All data types are of form float64 and object
6. No null values found in each column
7. From describe function mean , median and standard deviation obtained
8. Outliers present and retained
9. Corelation coefficients calculated
10. Normalization is used for scaling data
11. There is collinearity problem
12. All are positively skewed
13. From QQ plot data is normal
14. EDA:-
15. From box plot



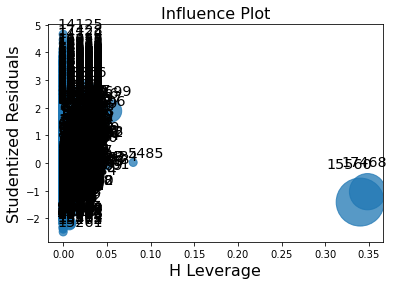
1. From QQ plot



1. From pair plot



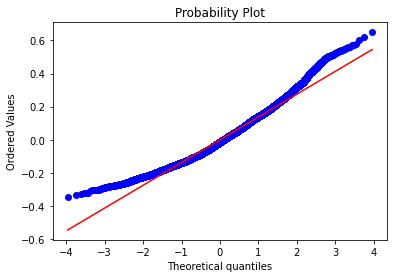
1. MODEL BUILDING:-
2. Model builded but P value is high
3. There is collinearity problem
4. Influence index calculated



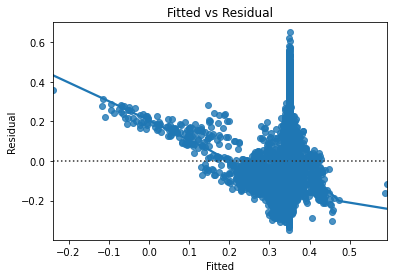
1. Dropping 17468 but problem of collinearity not solved
2. Calculating VIF

|  |  |
| --- | --- |
| VARIABLES | VIF |
| AveragePrice | 1.054387e+00 |
| Total volume | 2 4.907221e+09 |
| Total average 1 | 2 6.583781e+08 |
| P Total average 2 | 2 5.965439e+08 |
| Total average 3 | 4.751604e+06 |
| Total bags | 2.370316e+14 |
| S Small bags | 1.364727e+14 |
| Large bags | 1.448103e+13 |
| Xlarge bags | 7.620239e+10 |

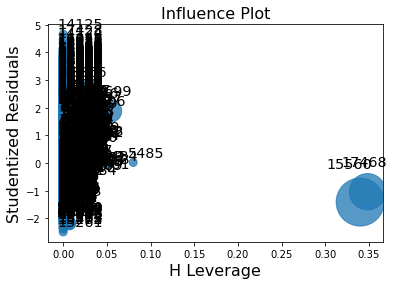
1. Since total bags having high vif i so dropping it
2. QQ plot for residual



1. For linearity



1. Influence index for final



OUTPUT:-

1. Best fit model with R2 = 0.052
2. Intercept = 0.3550
3. tv coefficient = 3802.6937
4. ta1 coefficient = -1384.5448
5. ta2 coefficient = -1244.8421
6. ta3 coefficient = -155.3551
7. sb coefficient = -814.1585
8. lb coefficient = -348.1631
9. xlb coefficient = -33.2852
10. Splitting data to train 80% and test 20%
11. Train accuracy :- 0.1401319525566205
12. Test accuracy:- 0.1373225053789519
13. It is right fit model
14. BENEFITS :-

From above information we can predict for average price of avocado for all input variables .