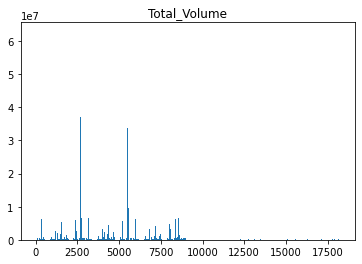
**Business Problem:** With the growing consumption of avocados in the USA, a freelance company would like to do some analysis on the patterns of consumption in different cities and would like to come up with a prediction model for the price of avocados. We need to implement, and build a prediction model using multilinear regression and give insights.

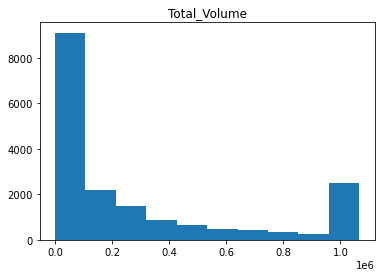
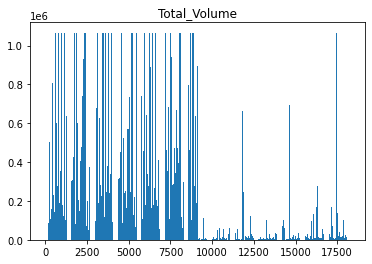
|  |  |  |
| --- | --- | --- |
| **Description** | **Type** | **Relevance** |
| Average price | Numeric | Price of avocado |
| Total Volume | Numeric | How many kgs |
| tot\_ava1 | Numeric | At the location1 available avocados |
| tot\_ava2 | Numeric | At the location2 available avocados |
| tot\_ava3 | Numeric | At the location3 available avocados |
| Total\_Bags | Numeric | Number of bags available |
| Small\_Bags | Numeric | Number of small bags |
| Large\_Bags | Numeric | Number of large bags |
| Xlarge\_Bags | Numeric | Number of large bags |
| Type | Categorical | What catageoric avacados |
| Year | Numeric | The date avocados produce |
| Region | Categorical | County |

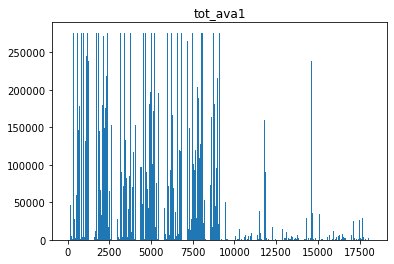
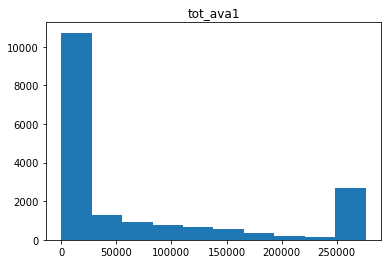
**Data Pre-Processing:**

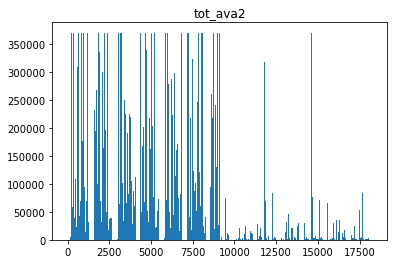
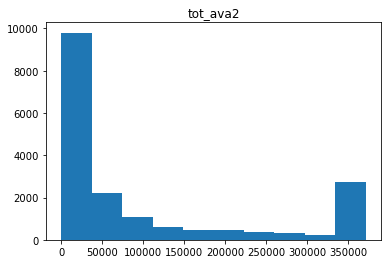
* In the data we didn’t get any null values and both columns are in integer type
* We get outliers in the Total\_Volume, tot\_ava2,\_ava1, tot\_ava3, Small\_Bags, and Large\_Bags, Xlarge\_Bags data by using the Winsorization technique we compressed the data into normal.
* We didn’t have duplicates and missing value
* For the state column we perform the label encoding technique converting data into numeric form.

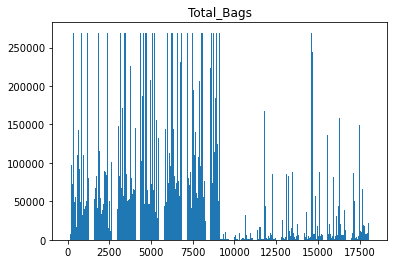
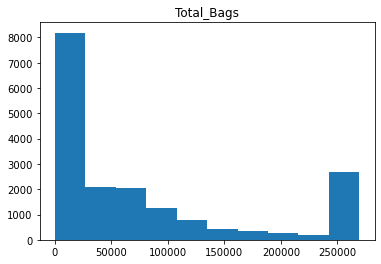
**Exploratory Data Analysis (EDA):**

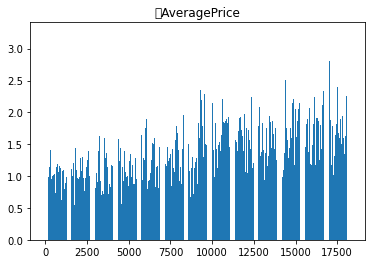
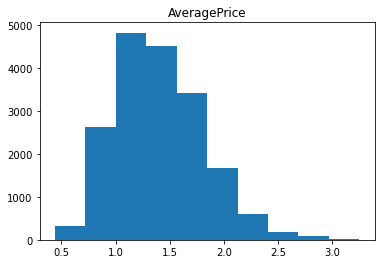
 

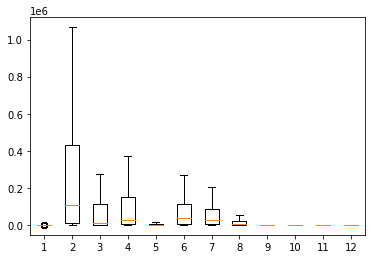
 



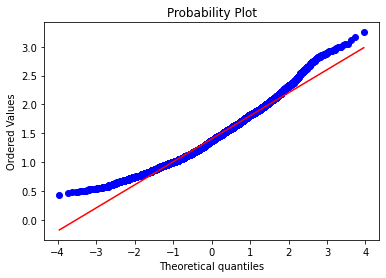
 

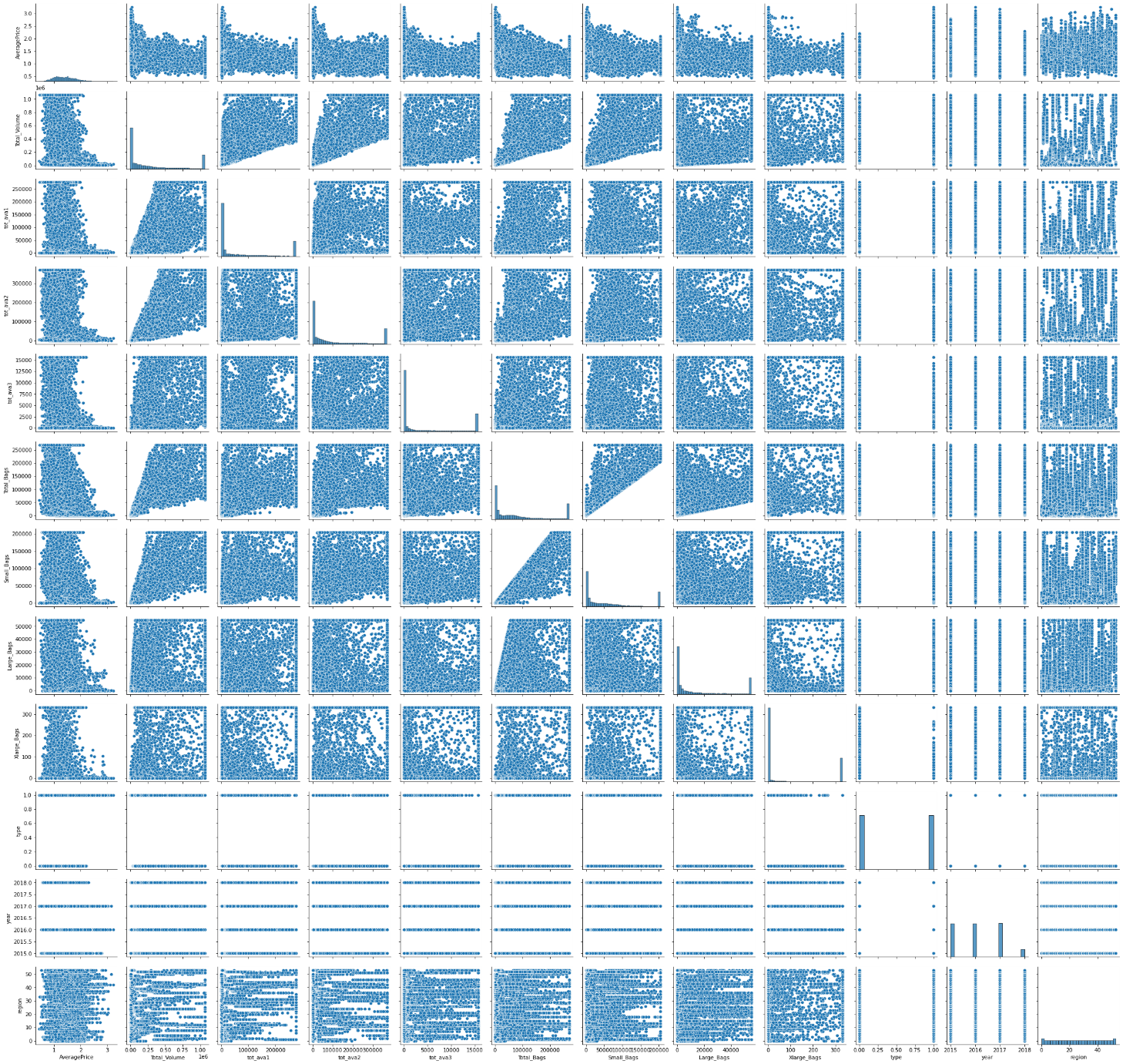
 

* By observing all the bar plots we see which avocados have how many bags and which avocados sold high and which were low all we can see is in these.
* Comes to avg price we can see which 17000 avocados was the high price was happed that one compares to others.
* From the above Histogram we can observe skewness column-wise is that left skew or right skew. Skewness is a measure of symmetry or the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

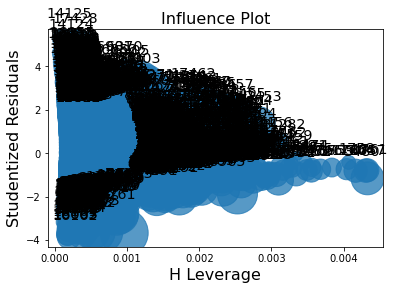


* Finally we found our data without outliers.
* A straight, diagonal line means that you have normally distributed data. If the line is skewed to the left or right, it means that you do not have normally distributed data.

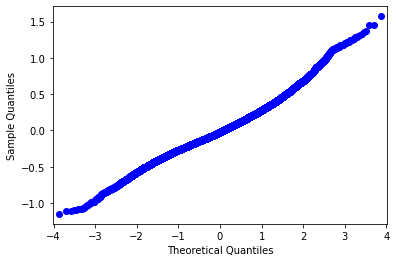
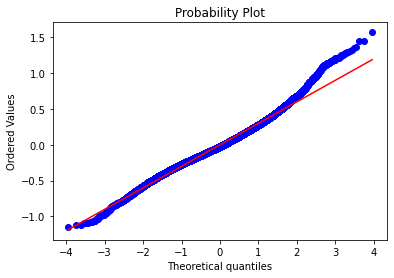




* This Pair plot explains the relation between one column of data to another column of data. If it was linearly distributed they have e good co-relation



* Let’s develop the model building. Price is the output and the remaining columns as the input.
* For model1 we get the results as R-sqr:44% accuracy and intercepting value is -87. The P-value was <0.5 which means we are getting a good result. By using Influence plot indexes, we can remove any of the far away from the remaining indexes. Then we checked with VIF(variance influence factor), a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. in this step, if any of the column factors was > 10 we have to ignore that column and process with the remaining columns. Like total volume, VIF has 39, and tot ava2 has VIF 12 so we removed that column.
* After removing the VIF again we develop the model. Now we are applying the data we got getting the results as R-sqr:43% accuracy and intercepting value is -76. The P-value was <0.5 which means we are getting a good result and RMSE was 0.30 which was a low error.
* Now we are applying log for input data we got getting the results as R-sqr:37% accuracy and intercepting value is 1311. The P-value was <0.5 which means we are getting a good result and RMSE was 0.31.
* Let’s build the Exponential transformation for the data. After building the data getting the results as R-sqr:45% accuracy and intercepting value is -50. The P-value was <0.5 which means we are getting a good result and RMSE was 1.1.
* Compared to all the model1 given good results all the p-values are < 0.5 and its accuracy was 43% and with less error of 0.3. So we can process this as a final model.



* This is the final model residuals Q-Q plot and probability plot. Q–Q plot, "Quantile-Quantile" plot Normal probability plot, and Q–Q plot against the standard normal distribution.



* When conducting a residual analysis, a "residuals versus fits plot" is the most frequently created plot. It is a scatter plot of residuals on the y-axis and fitted values (estimated responses) on the x-axis. The plot is used to detect non-linearity, unequal error variances, and outliers.
* Now we have trained our data with 80% and tested the data with 20%. In this process, we were given a random state of 0 to get consistent results from our data. For the test, the prediction result was 0.302 RMSE value and the train prediction was 0.301.