ARTICLE ON AVOCADO PROJECT

1.Problem Definition

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV.

The table below represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados.

Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags.

The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (e.g. greenskins) are not included in this table.

Input Columns:-

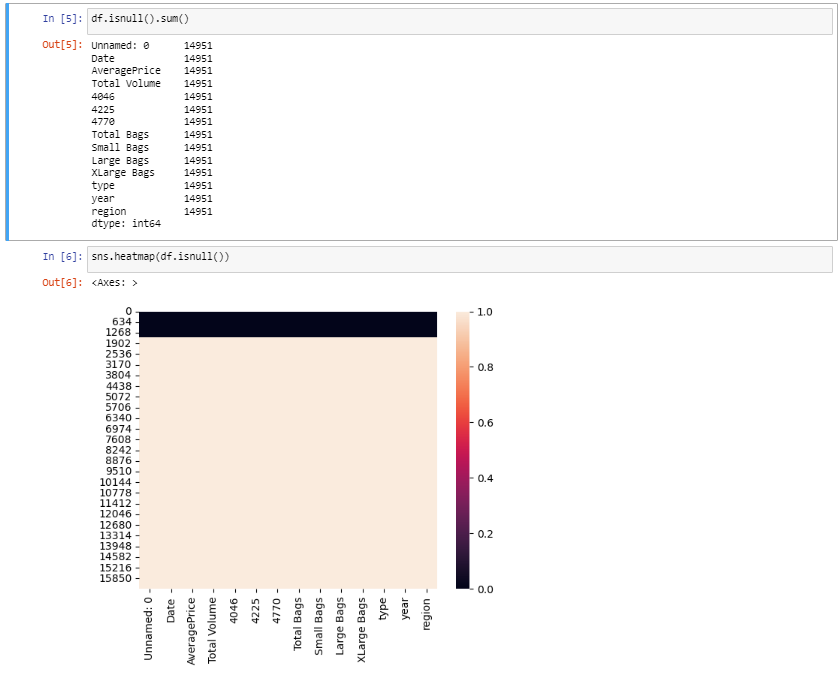
* Date - The date of the observation
* AveragePrice - the average price of a single avocado
* type - conventional or organic
* year - the year
* Region - the city or region of the observation
* Total Volume - Total number of avocados sold
* 4046 - Total number of avocados with PLU 4046 sold
* 4225 - Total number of avocados with PLU 4225 sold
* 4770 - Total number of avocados with PLU 4770 sold

Output Columns:-

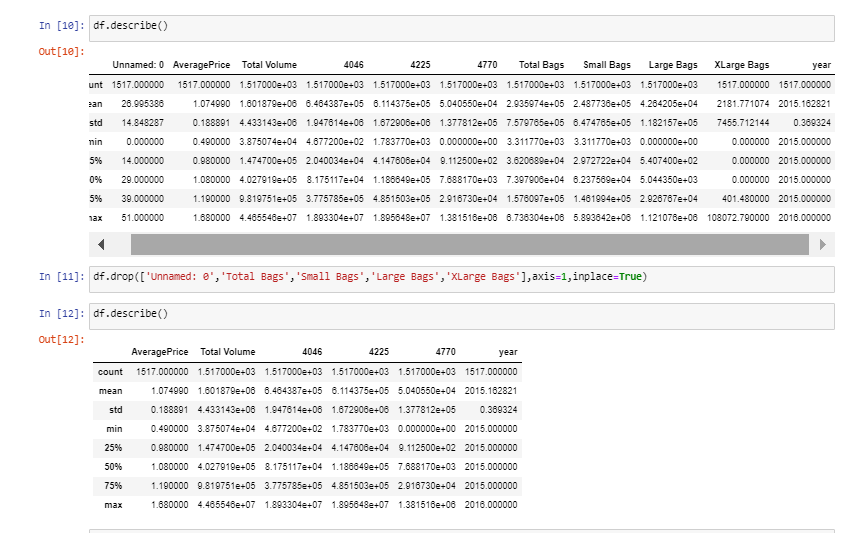
* Average price
* Region

2.Data Analysis

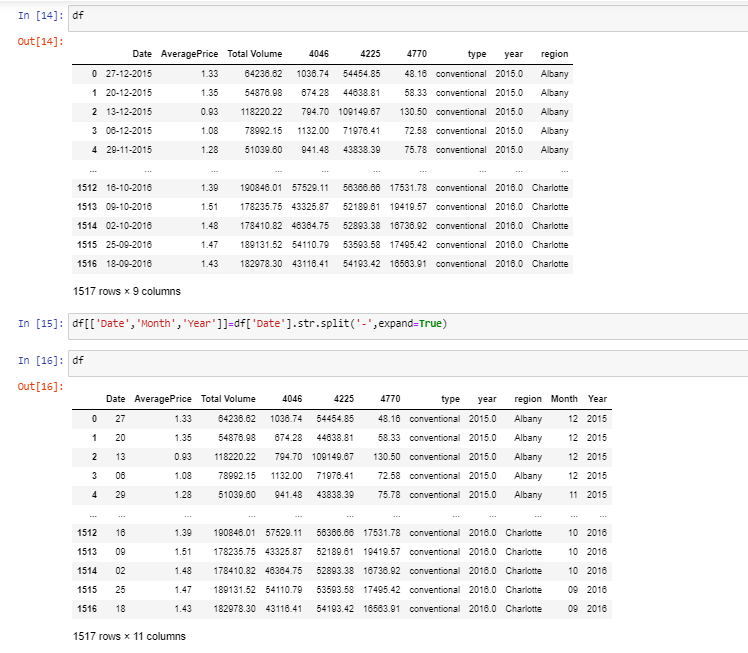
* There is a huge amount of Null values present in Data Structure.

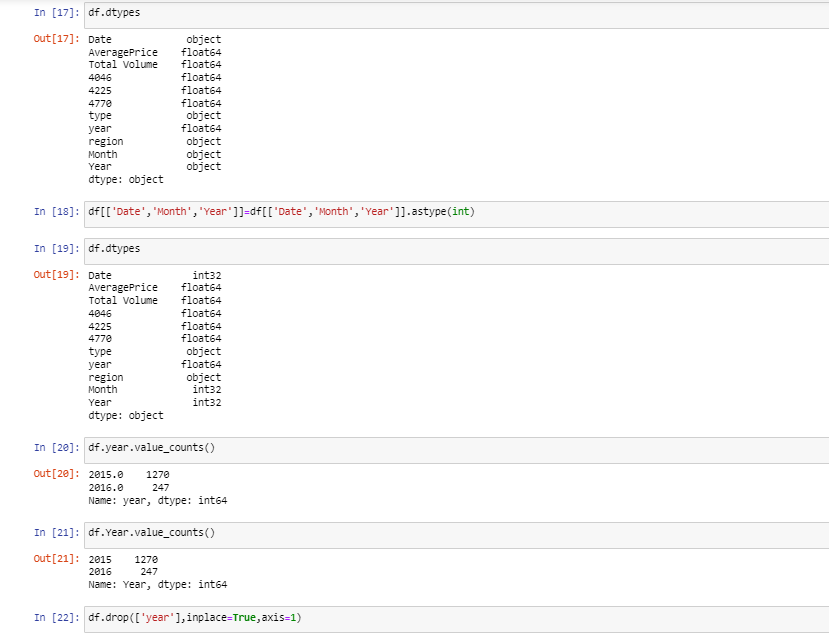


* Some extra columns having no connection or less are removed.



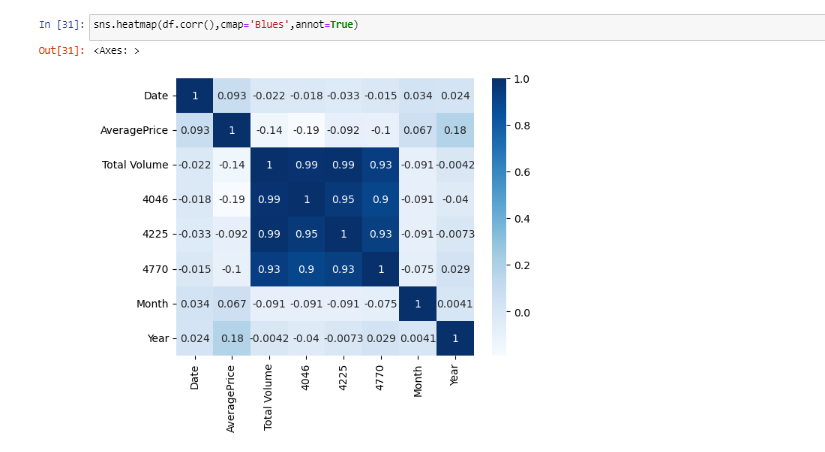
* Date column is being splitted with columns date,month and year. Year column is matched with the already present Date column for clarification and can be removed the initially present one.

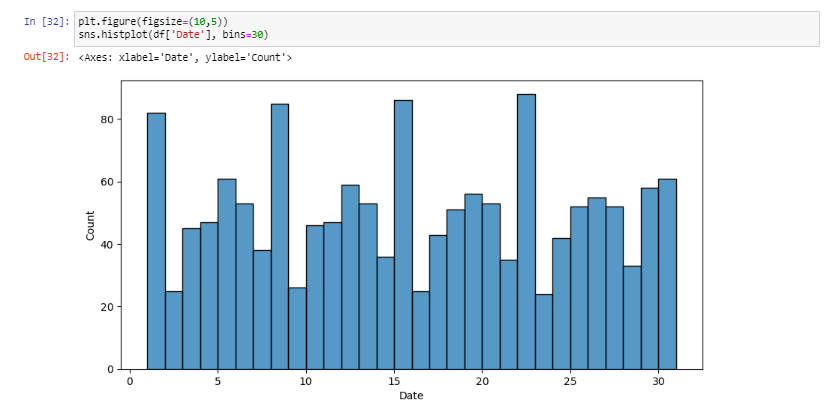


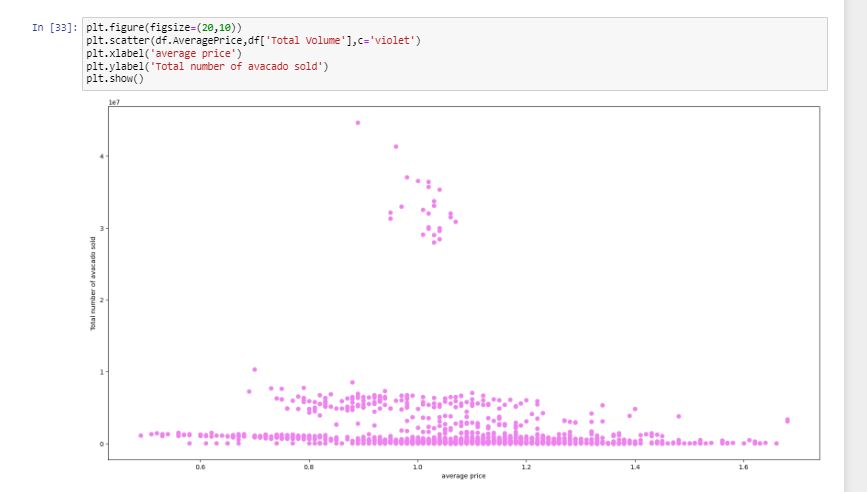


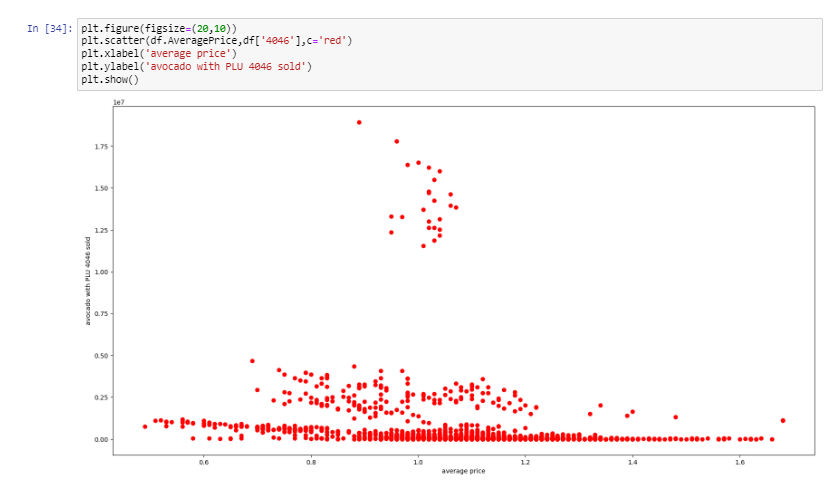
3.EDA Concluding Remarks

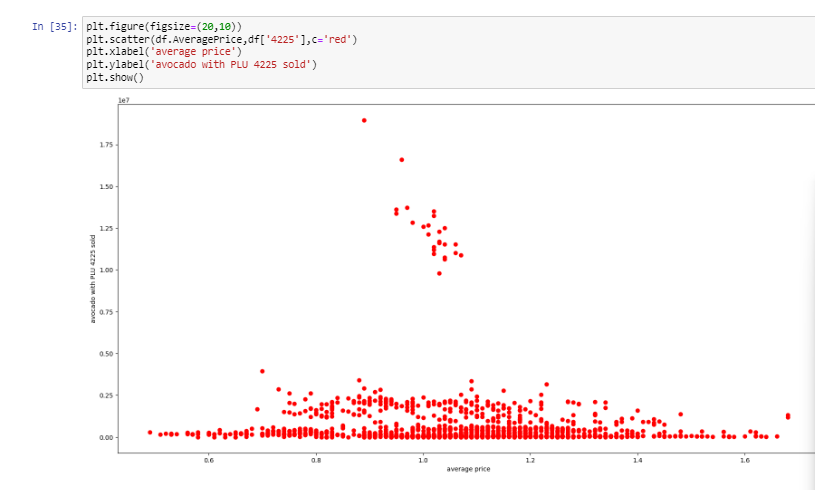
By using EDA we can get details about correlation and also we can check univariate,bivariate and multivariate analysis.

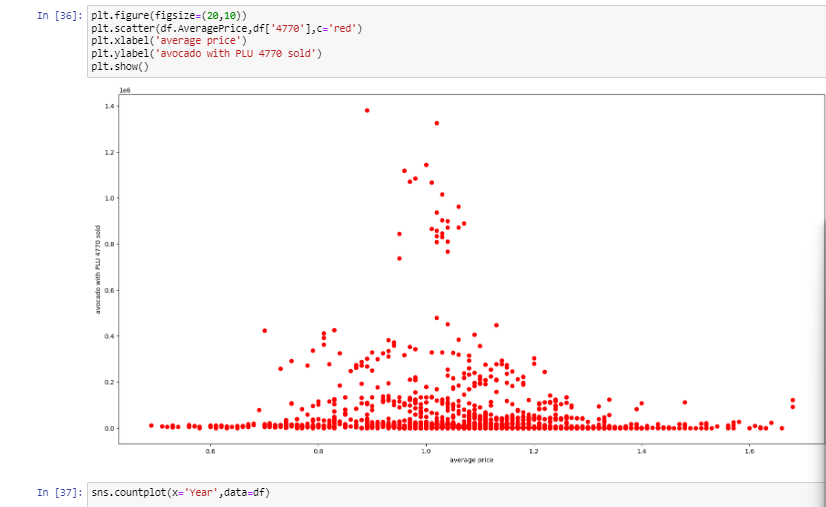


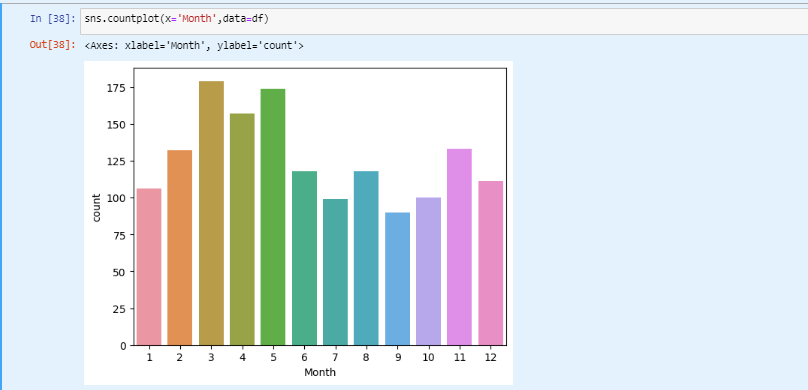








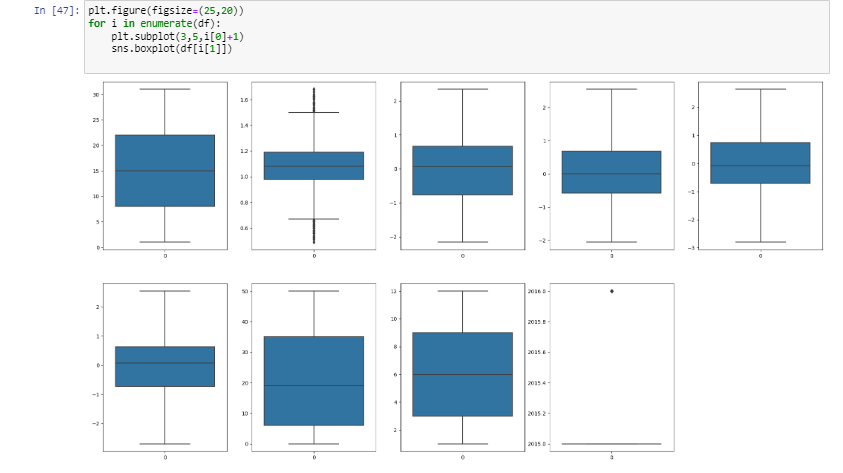


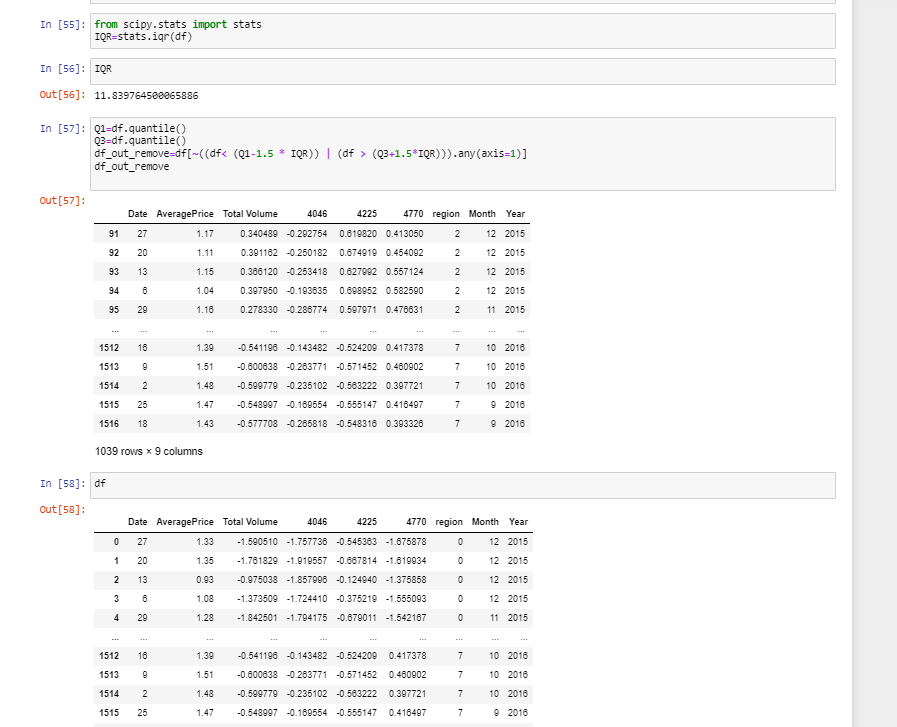


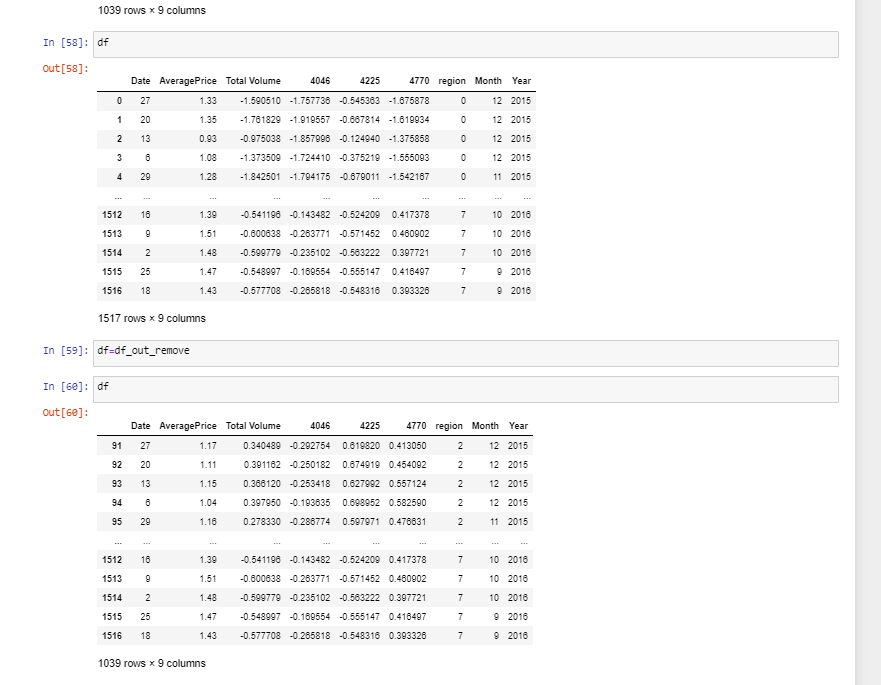
4.Preprocessing Pipeline

For data pre-processing we have to check skewness,outliers (if present).And then after we have to remove them using power\_transform,zscore or quantile method.



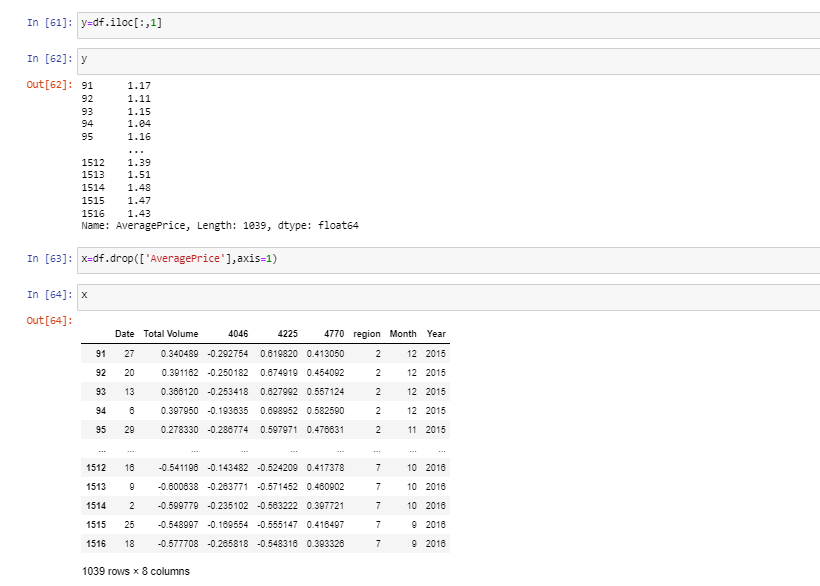


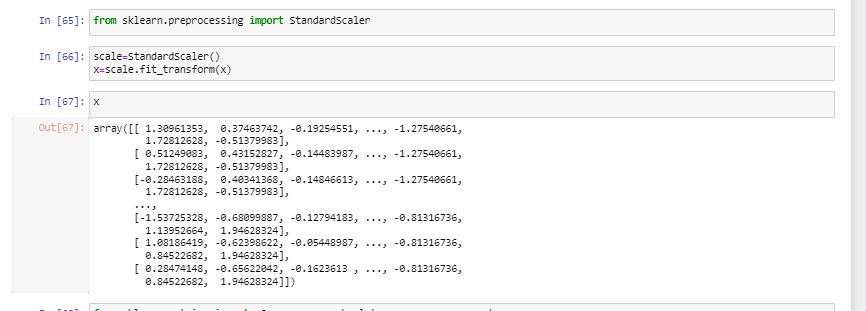




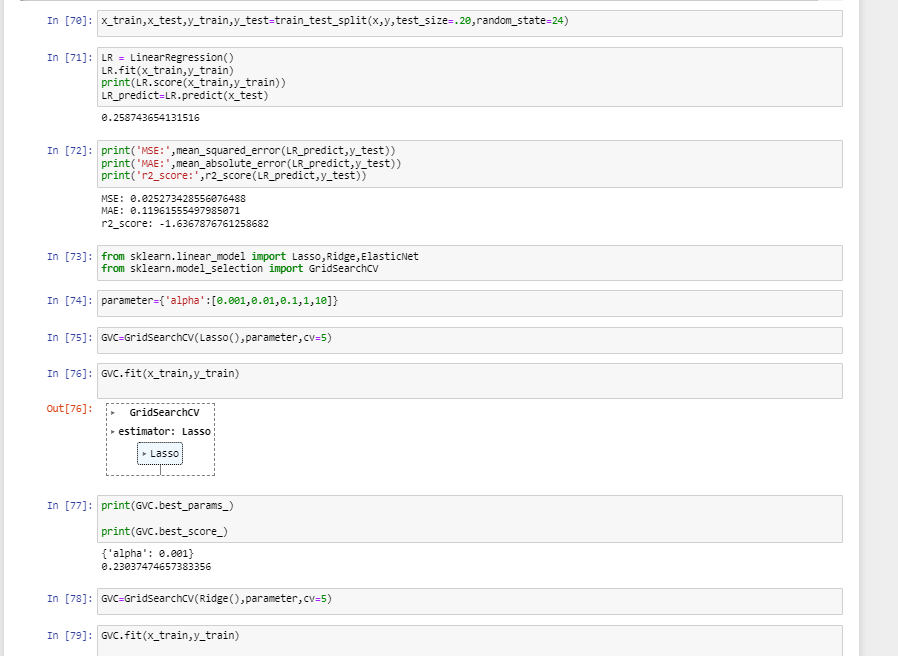
5.  Building Machine Learning Models

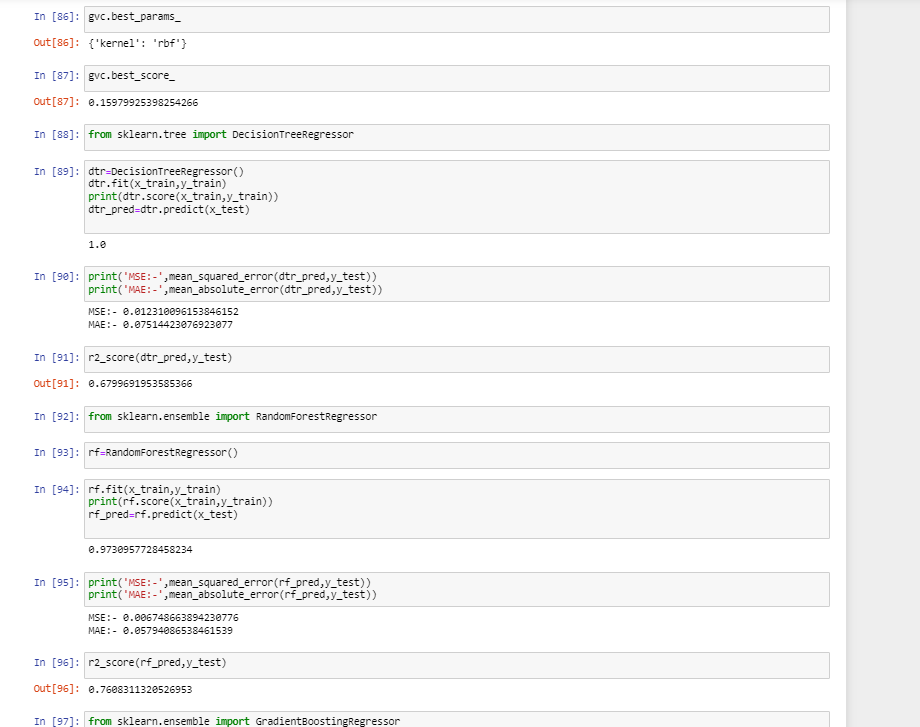
For making ML model we have to standardize the df using standard scaler after separating the df with input data and output data.

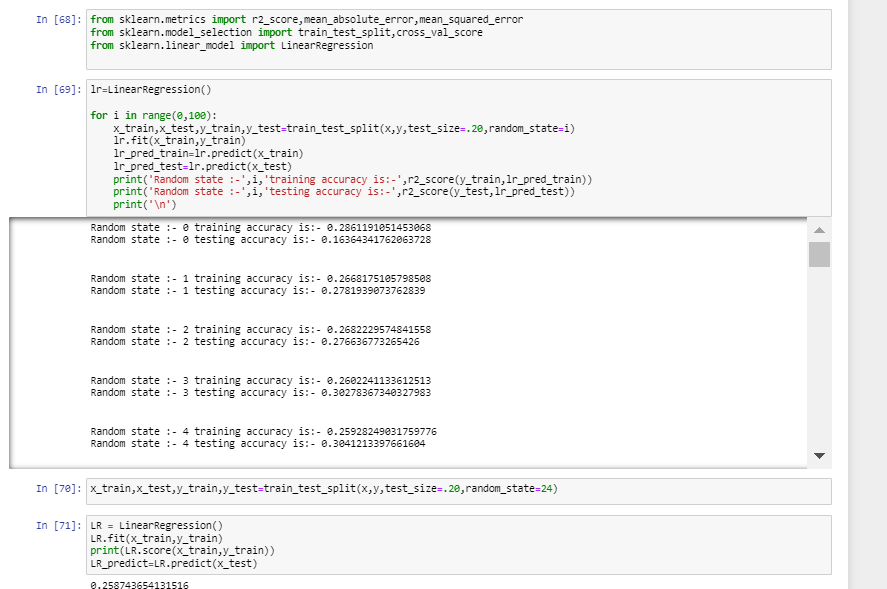




Firstly our 1st target variable is Average price, so We have to find the best Random state for better accuracy of model.



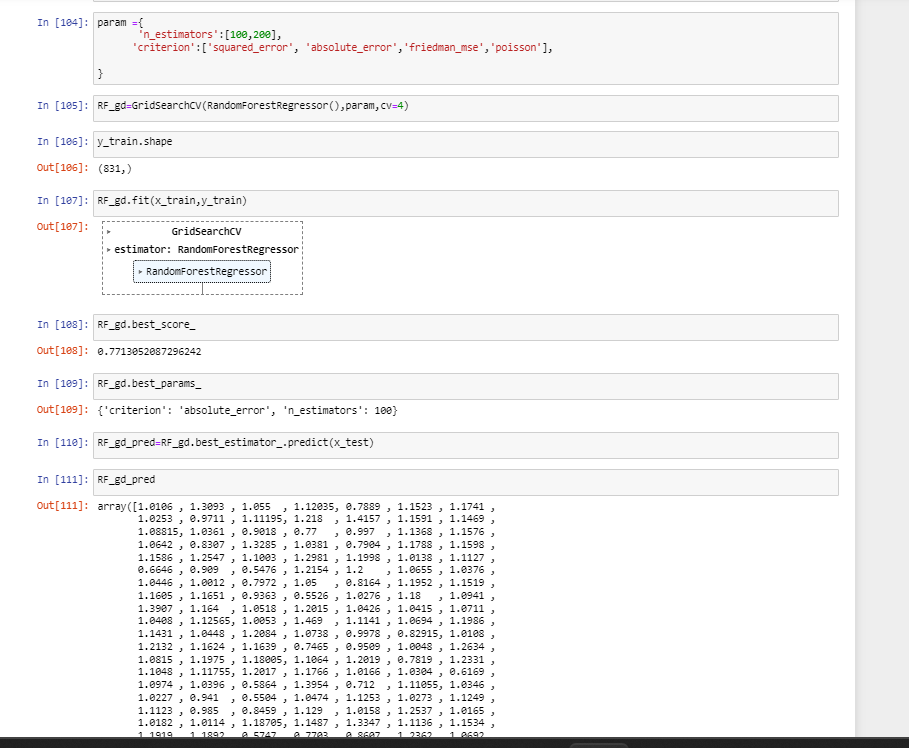


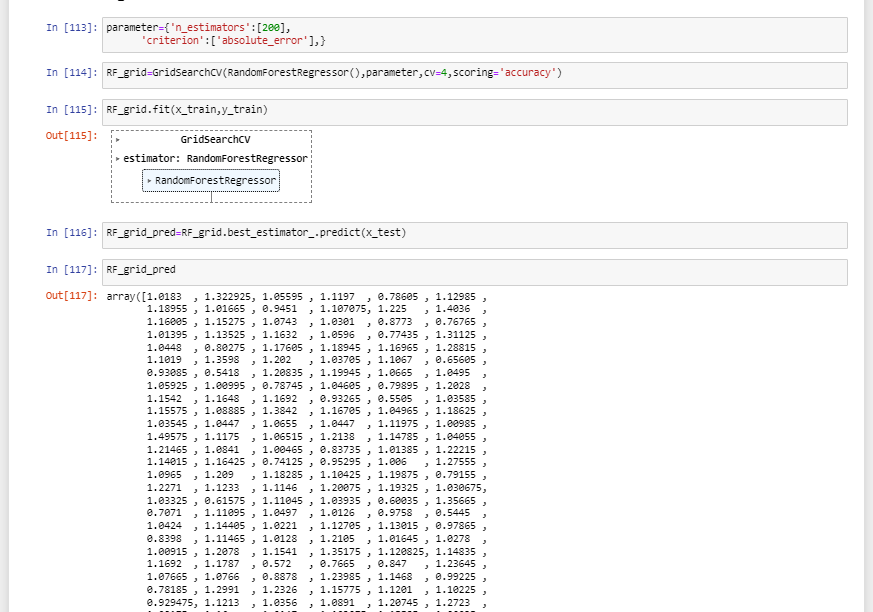


Then cross validate with less error and max r2\_score.

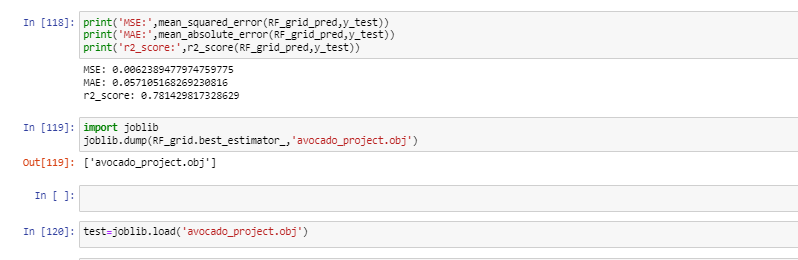


Choosing Random Forest Regressor and tuning with Grid Search CV for best score and parameter.





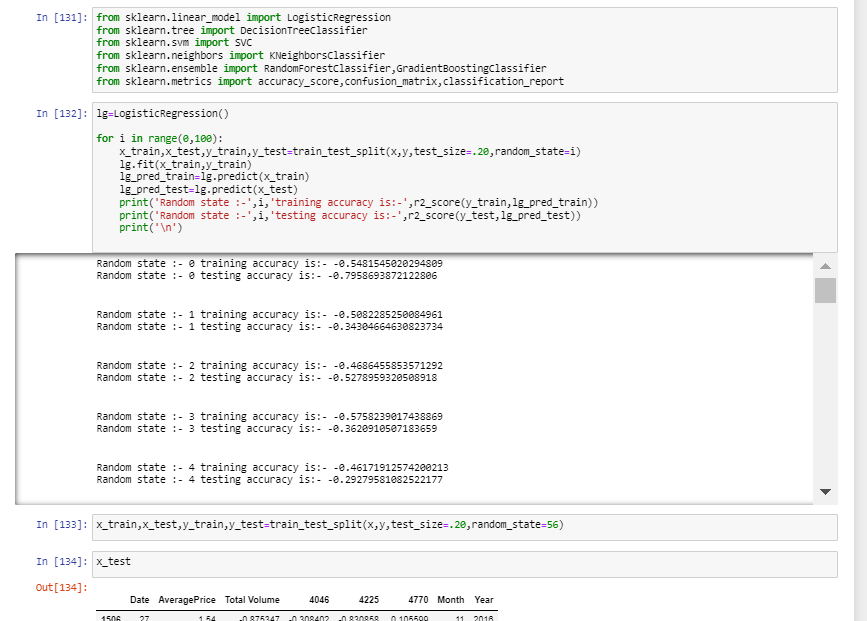
Saving the model using joblib.



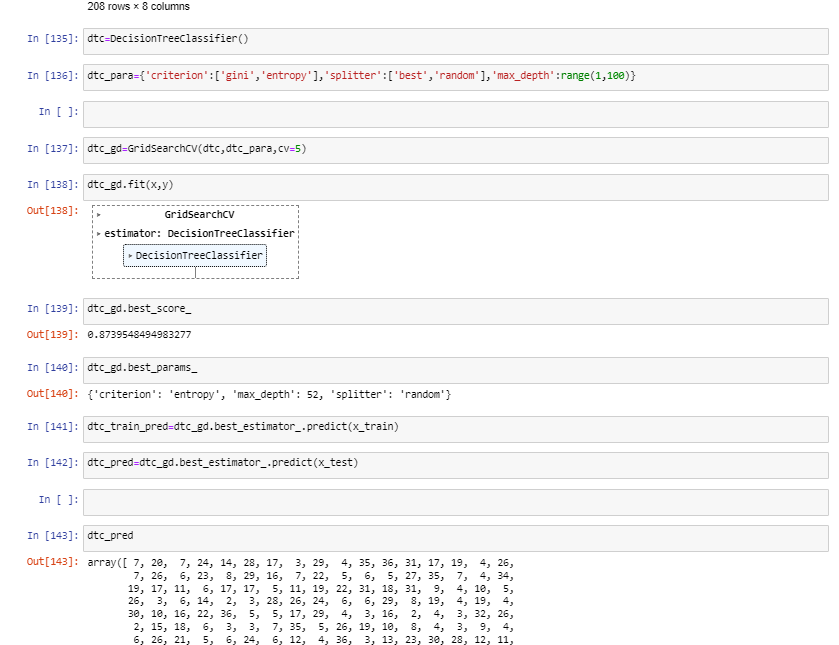
Classification Part:-

Our next outcome is region.

Like regression we have to split x and y variable and find the random state for better accuracy.



Decesion Tree Regressor having best accuracy. So choosing this model .



6.Concluding Remarks:-

* **We have to use matplotlib,seaborn for Data visualization.**
* **Boxplot,zscore,quantile for checking outliers and remove the same.**
* **From analysis I get to know about the negative and positive correlations among the columns.**
* **I checked the models like Linear, Decision Tree and other applicable and required models to fine tune the predictions.**
* **I came to know through analysis which model will be work with better accuracy with the help of low residual and RMSE scores, accuracy score.**