

**AVOCADO PROJECT**



Submitted by:

CHETHANA M

Batch ID : 1843

DataTrained

**INTRODUCTION**

**Avocado is a fruit consumed by people heavily in the United States and now a days it is being consumed most of the countries due to its health benefits.**

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.

**Technical Goals :**

* Predicting Average price of the Avocado using Regressors
* Predicting Region of the Avocado using Classifiers

Our main aim today is to make a model which can give us a good prediction on the Average price and Region of the Avocado based on other variables. We are going to use Linear Regression, Support vector regressor, Decision Tree Regressor, K Neighbors Regressor, Ridge and Lasso, Random forest classifier, Decision Tree Classifier, K Neighbors Classifier to build the different models for this dataset and see which model gives us a good accuracy.

**Steps Involved in this Project :**

1. Understanding the Problem

2. Data Analysis

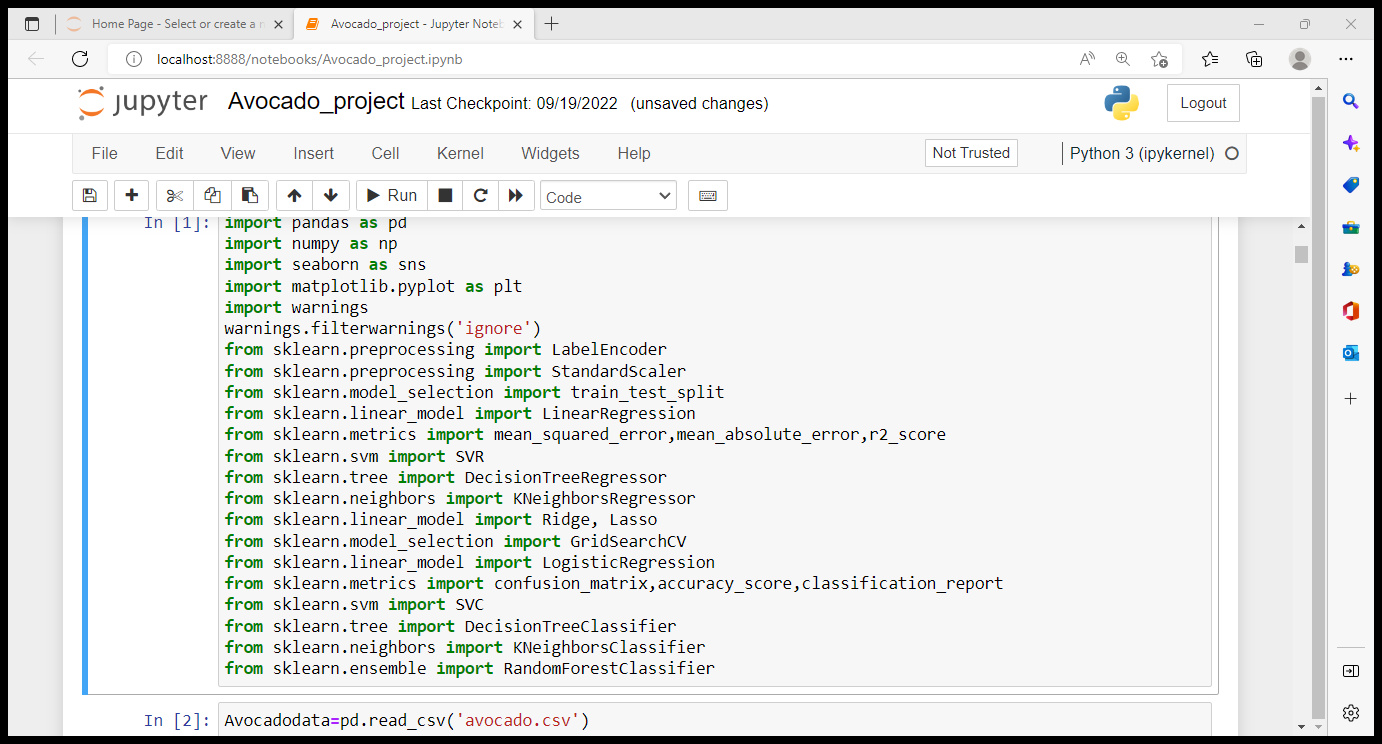
3. EDA Concluding Remark

4. Pre-Processing

5. Building Machine Learning Models and finalizing the best model

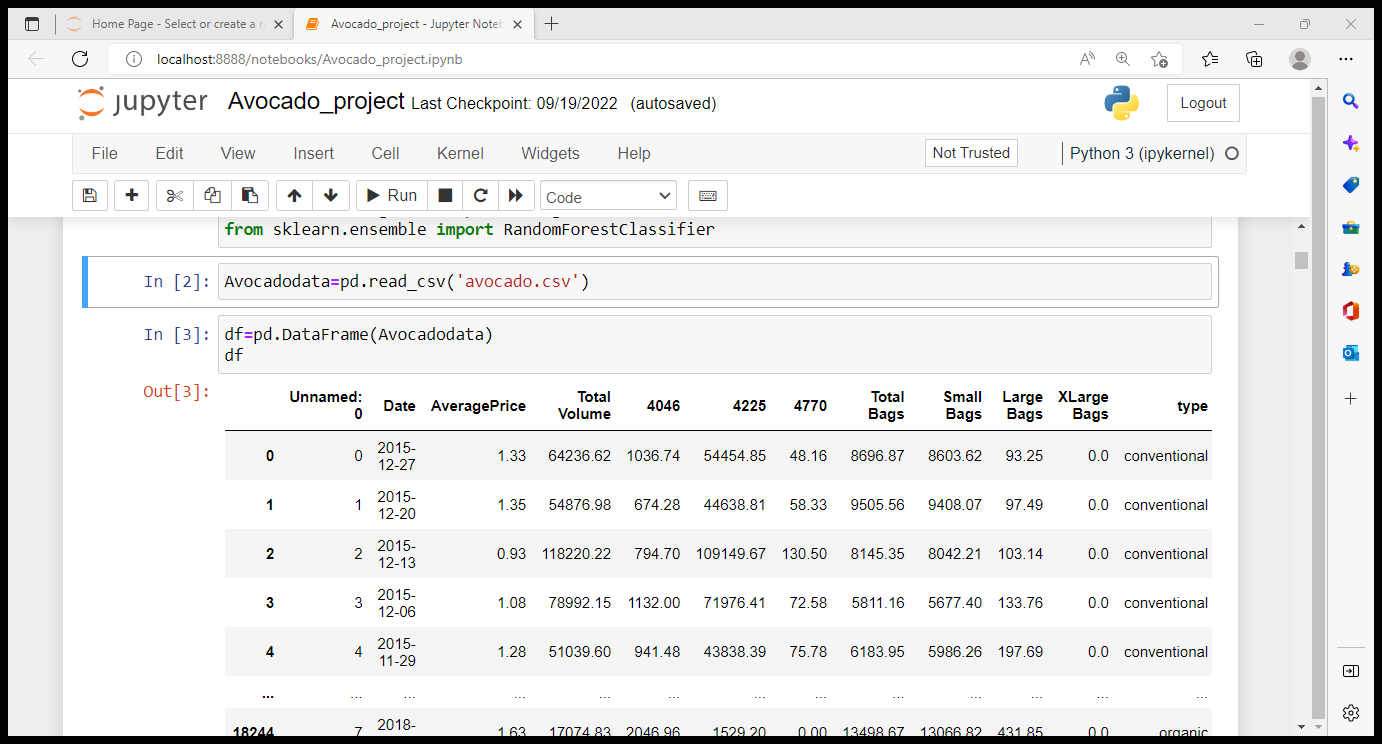
6. Concluding Remarks

**Loading the Libraries :**

****

**EDA and Visualization**

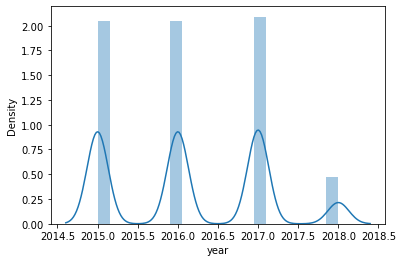
**Loading the data :**



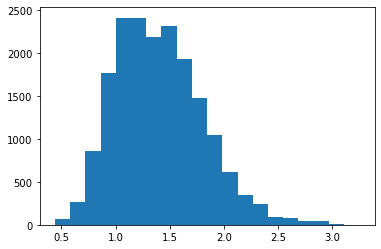
**Checking for features of dataset :**

The data set has 18,249 rows and 14 variables. The variables in the data set and their dtypes are,

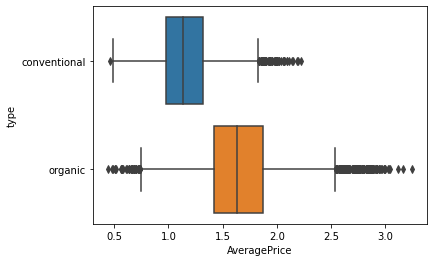
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Dtype | Description | Missing values | Outliers | Skew |
| 1.Unnamed | int64 | Which is just an index given in dataset, we can drop it in further steps | - | - | - |
| 2.Date | object | The date of the observation | No | Yes | No |
| 3.AveragePrice | float64 | the average price of a single avocado | No | Yes | No |
| 4.Total Volume | float64 | Total number of avocados sold | No | Yes | Yes |
| 5.4046 | float64 | Total number of avocados with PLU 4046 sold | No | Yes | Yes |
| 6.4225 | float64 | Total number of avocados with PLU 4225 sold | No | Yes | Yes |
| 7.4770 | float64 | Total number of avocados with PLU 4770 sold | No | Yes | Yes |
| 8.Total Bags | float64 | Total No of bags sold | No | Yes | Yes |
| 9.Small Bags | float64 | Total No of small bags sold | No | Yes | Yes |
| 10.Large Bags | float64 | Total No of Large bags sold | No | Yes | Yes |
| 11.XLarge Bags | float64 | Total No of XLarge bags sold | No | Yes | Yes |
| 12.Type | object | Type of Avocados | No | No | No |
| 13.Year | int64 | The year in which Avocado was sold | No | No | No |
| 14.Region | object | Region the Avocado originated from | No | No | No |



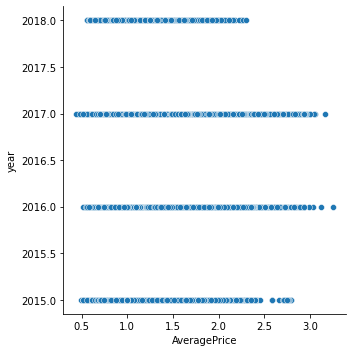
Above graph tells that the number of avocados sold in the year 2015, 2016, 2017 are almost equal in number and 2018 records are lesser in number.



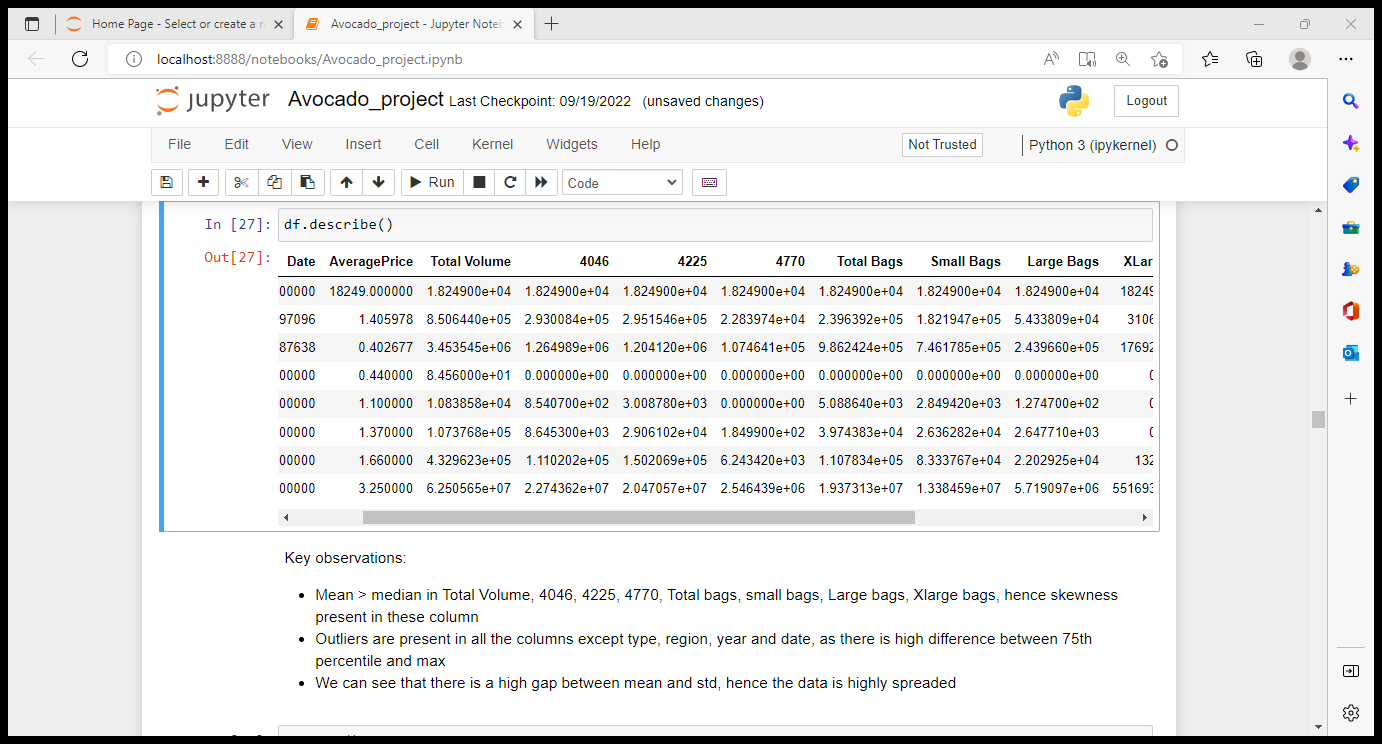
From the above plot we can observe that maximum price of avocado ranges from 1 to 1.6



From the above plot we can observe that the price is more for organic avocado and less for conventional avocado.

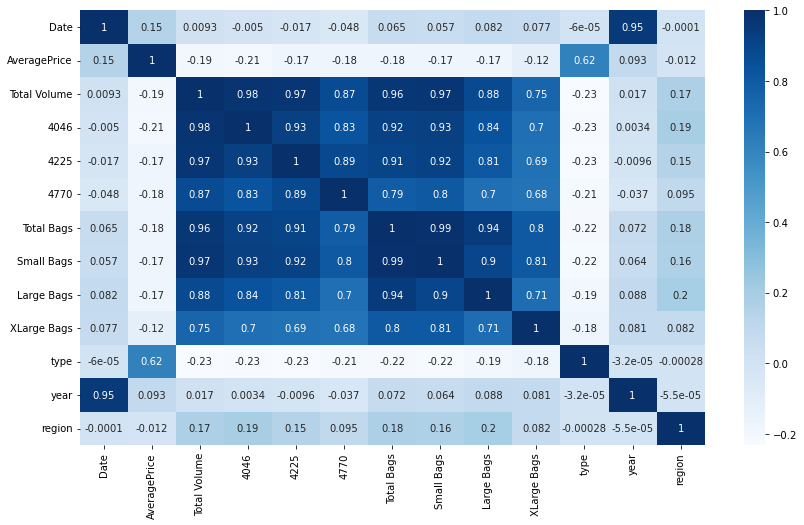


From the above plot we can observe that the AveragePrice reaches highest value in the year 2017 and 2016 than in 2015 and 2018.



From the above describe function we can observe that,

* Mean > median (50th Percentile ) in Total Volume, 4046, 4225, 4770, Total bags, small bags, Large bags, Xlarge bags hence skewness present in these variables.
* Outliers are present in all the columns except type, region, year and date as there is high difference between 75th Percentile and max.
* We can see that there is a high gap between mean and std in case of all the variables, hence the data is highly spreaded.

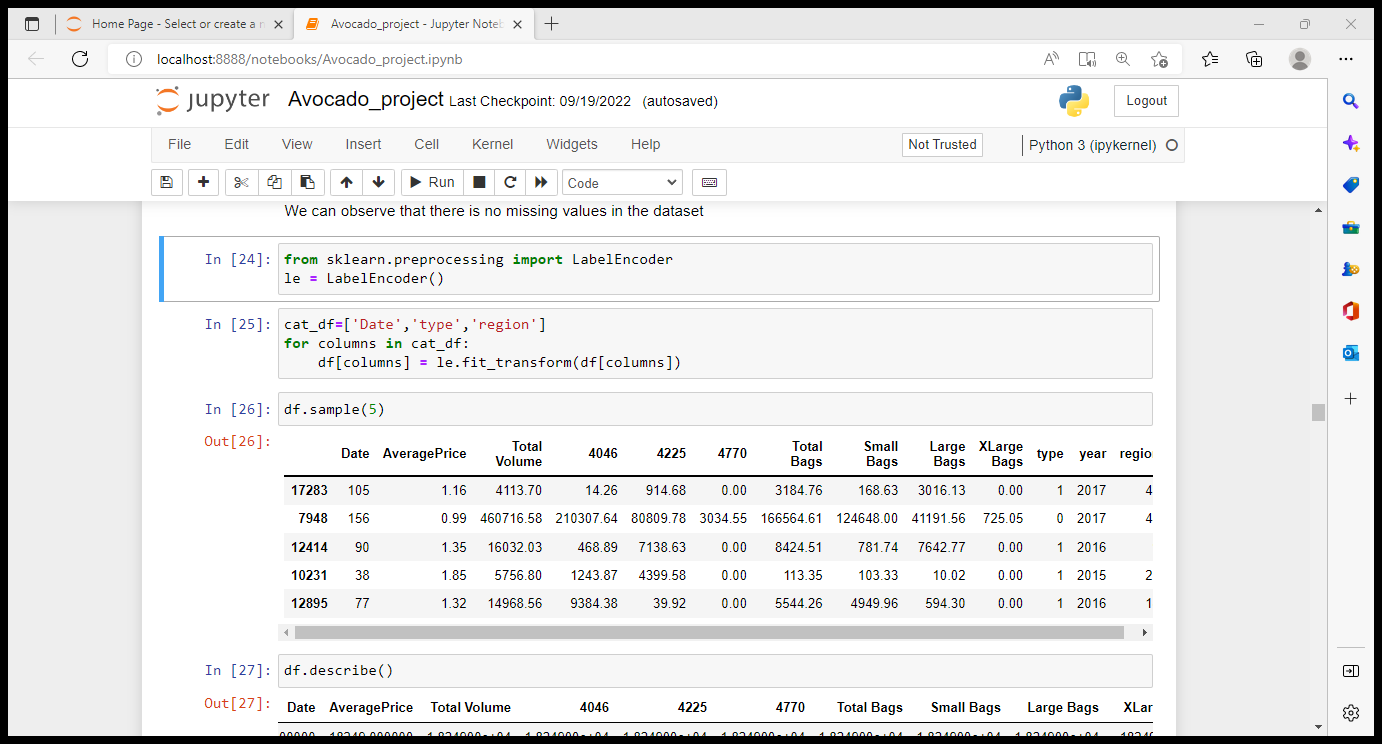


Key Observations from the above plot :

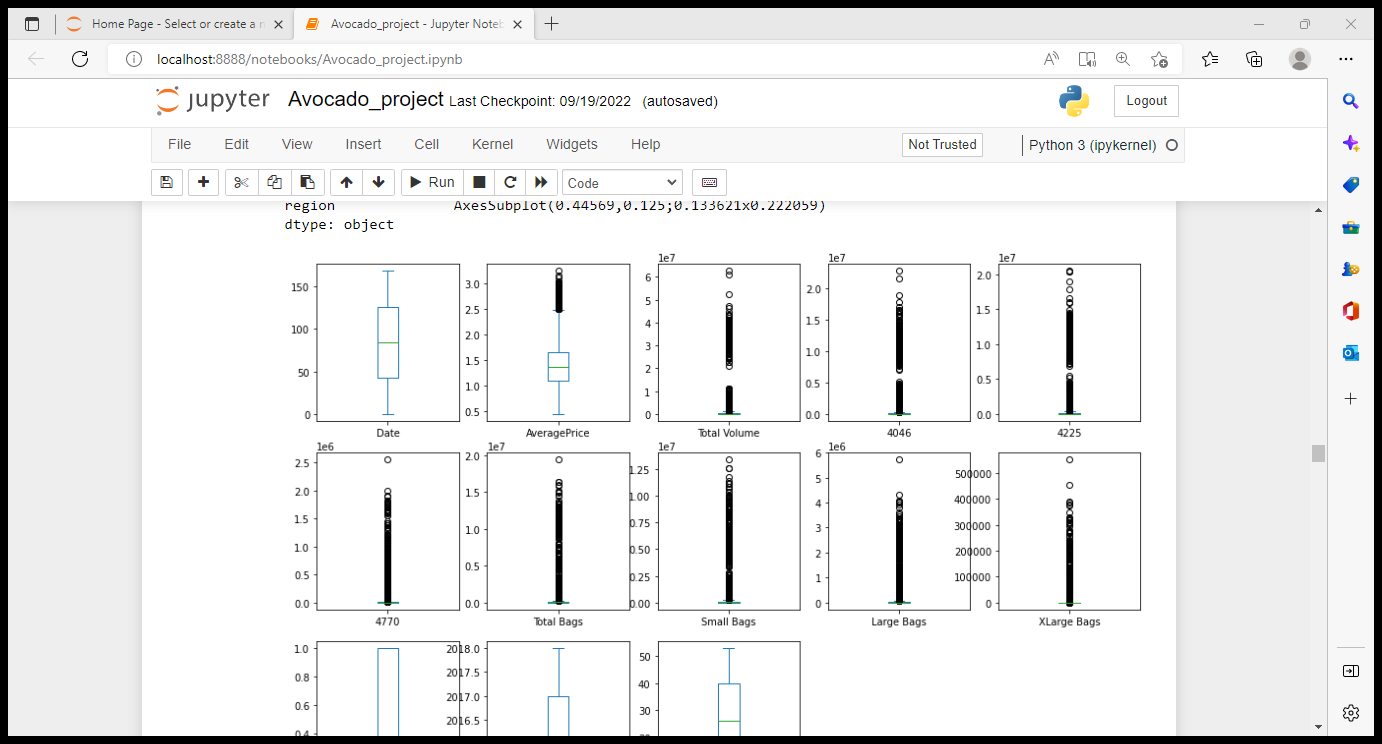
* Average Price is high correlation only with type and has negative correlation with other variables except region.
* Total bags has a good correlation with small bags and total volume.
* Total volume and 4046 are highly correlated.
* Region has positive correlation with XLarge Bags, Small Bags, Large Bags, Total Bags, 4046, 4225, 4770 and Total Volume.

As we discussed before now we can encode object type data to integer data using Label encoder, remove Outliers using Z-score method and remove skewness using 'yeo-johnson' method.

Encoding the Object type data using Label Encoder :

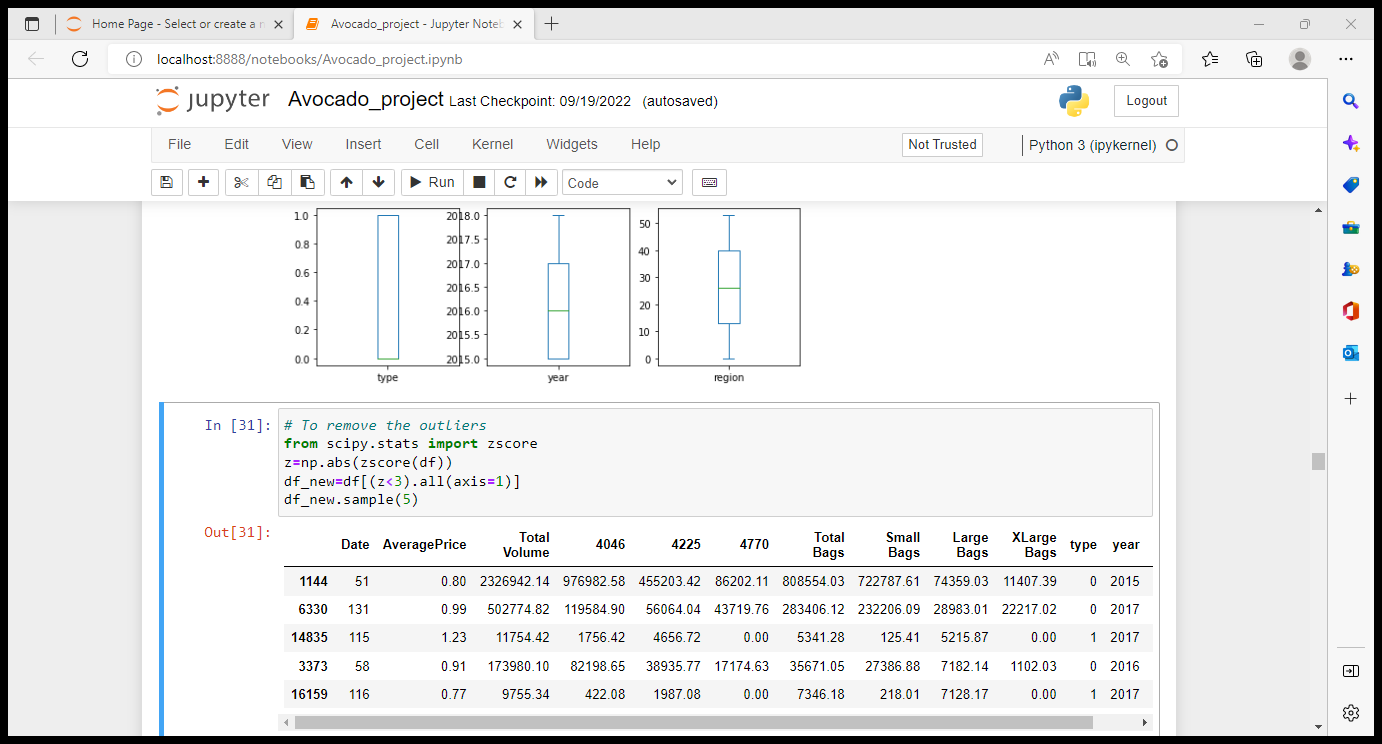


Plotting box plot for all the variables and checking for presence of Outliers in the data :



From the above plots it is clear that Outliers are present in case of Average Price, Total Volume, 4046, 4225, 4770, Total Bags, Small Bags, Large Bags, Xlarge Bags.

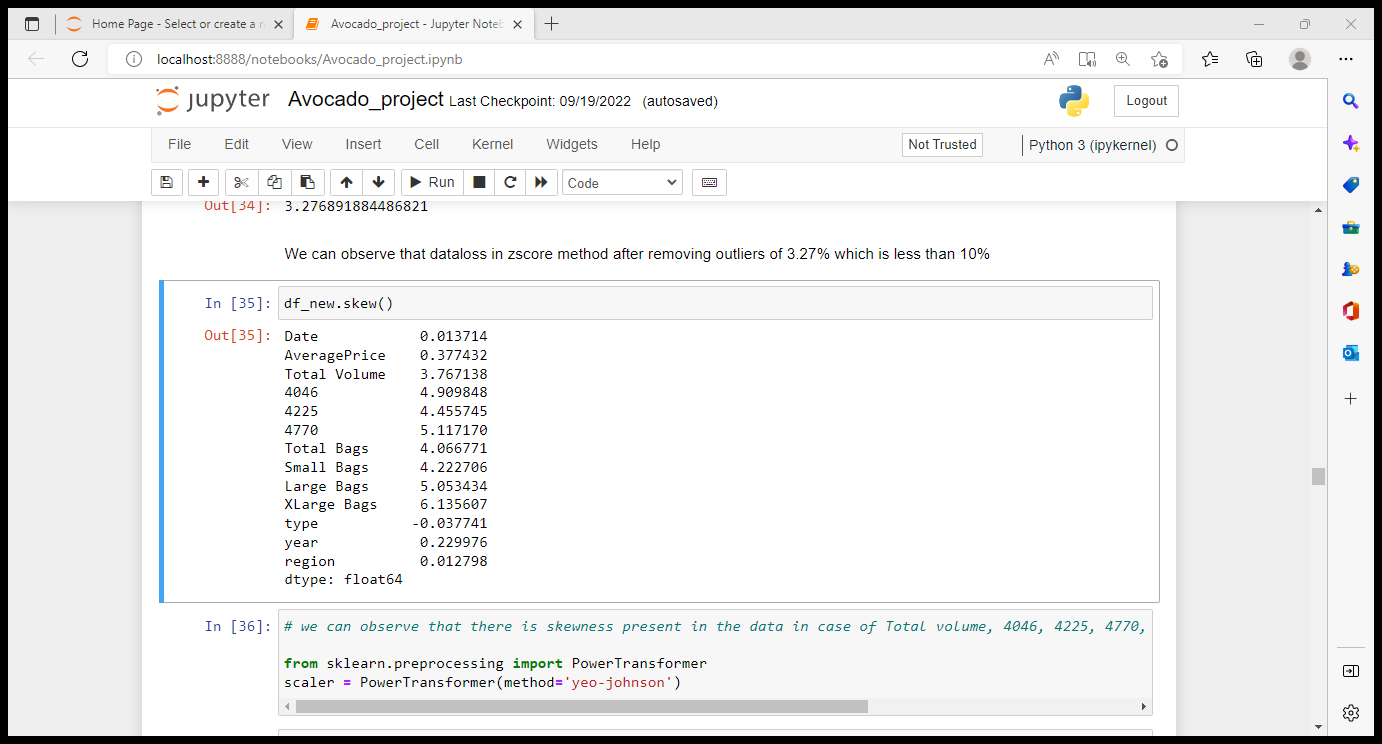
Removing Outliers using Z-score methos :



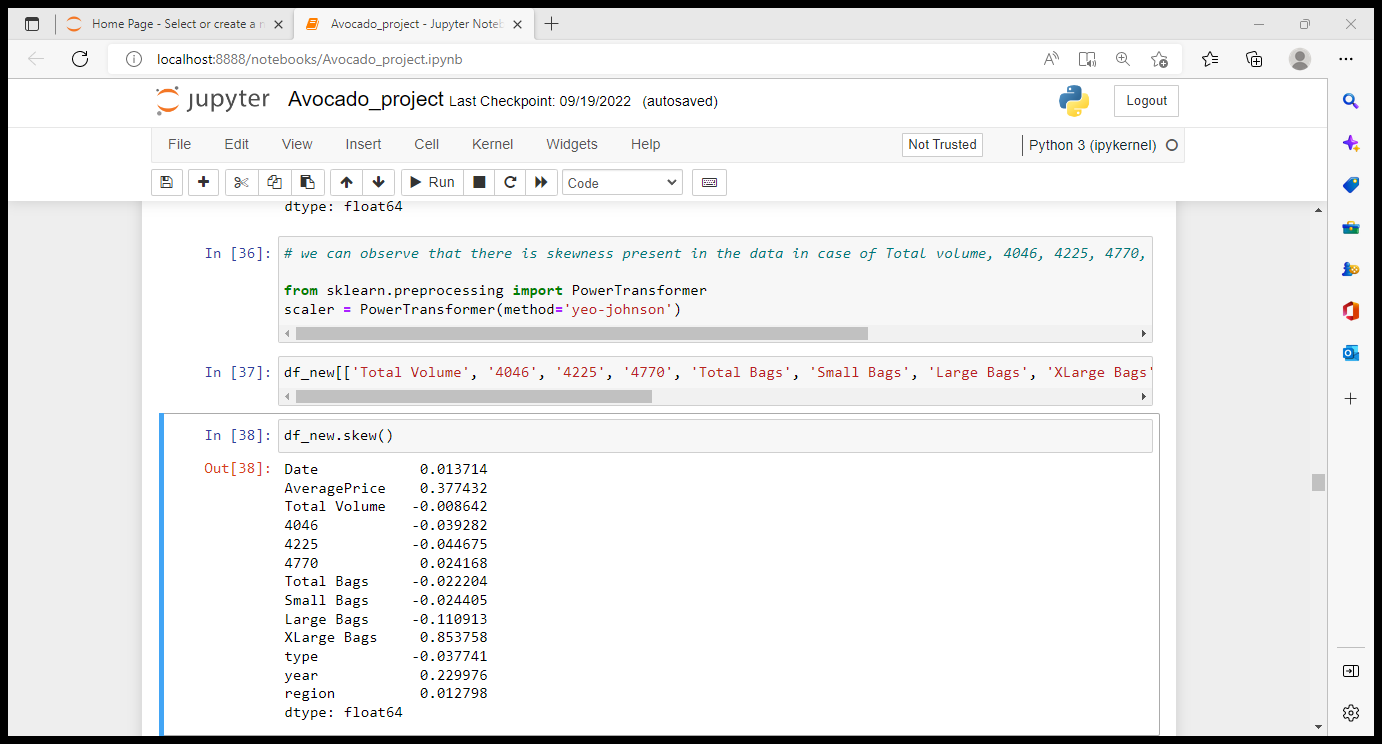
Once we remove the outliers from the data we can check for the percentage of data lost while removing outliers.

The shape of new dataset after removing outliers is 17651 rows and 13 columns, hence the percentage of lost data is 3.27% which is less than 10% and tells that it is appropriate to remove the outliers from the dataset.

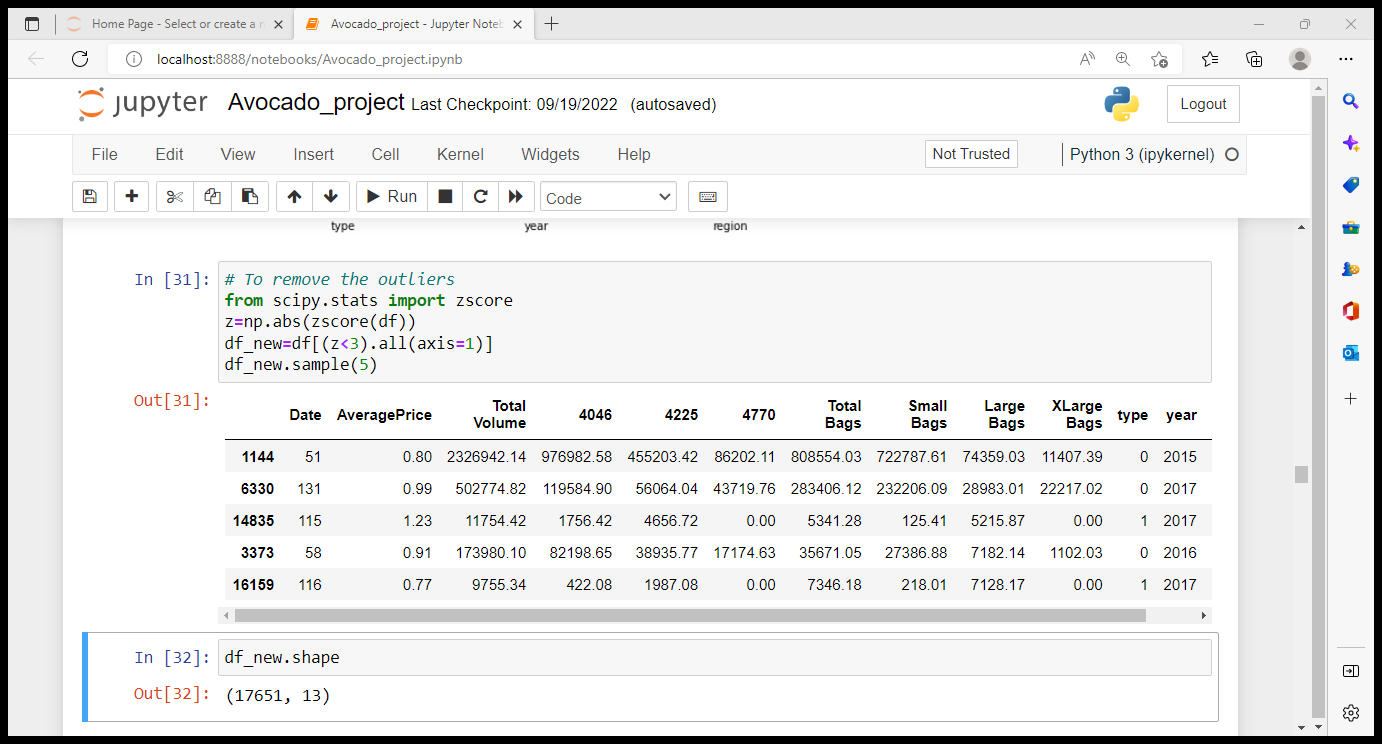
Checking for skewness in the data :



So we can observe that skewness is present in the variables like Total Volume, 4046, 4225, 4770, Total Bags, Small Bags, Large Bags, Xlarge Bags and we can remove this skewness usning **yeo-johnson** method.

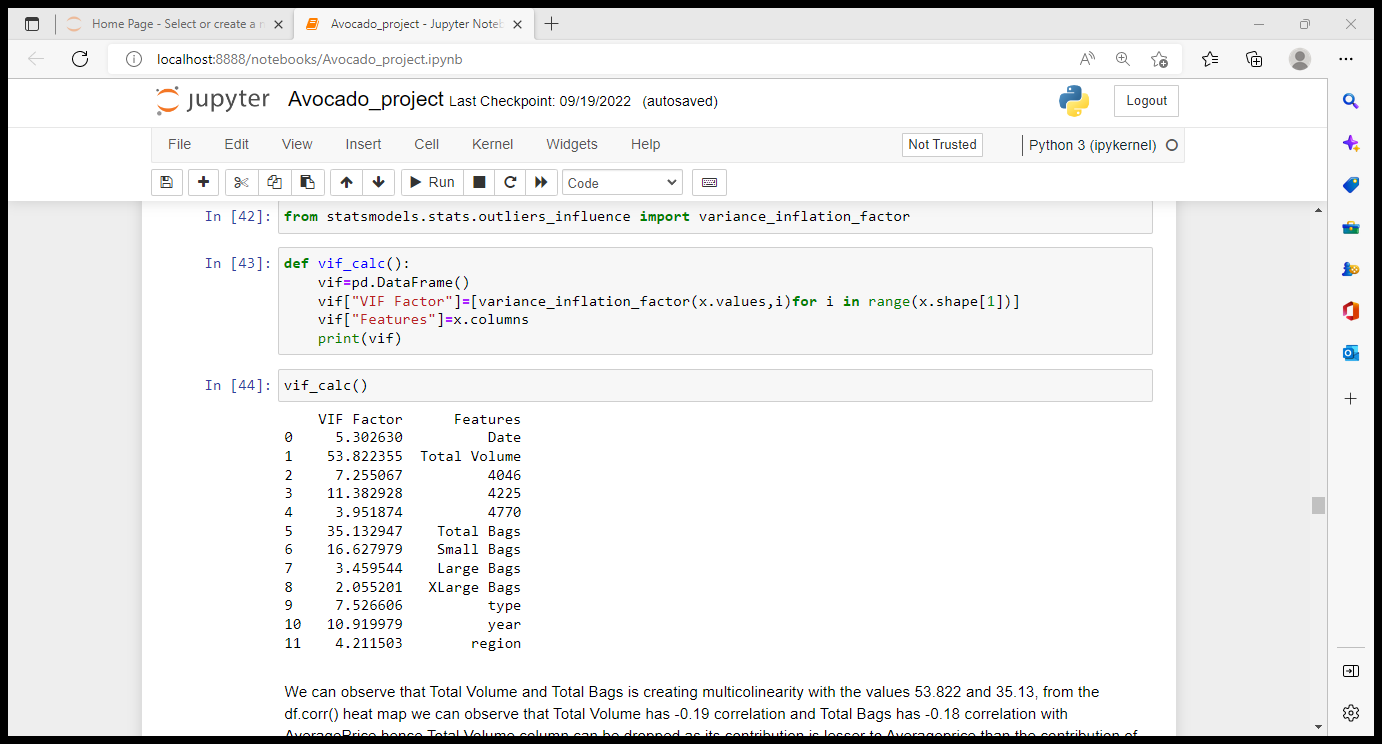


And hence we can observe from the above picture that skewness is removed from the data. After all the cleaning process we can see the cleansed data as below.

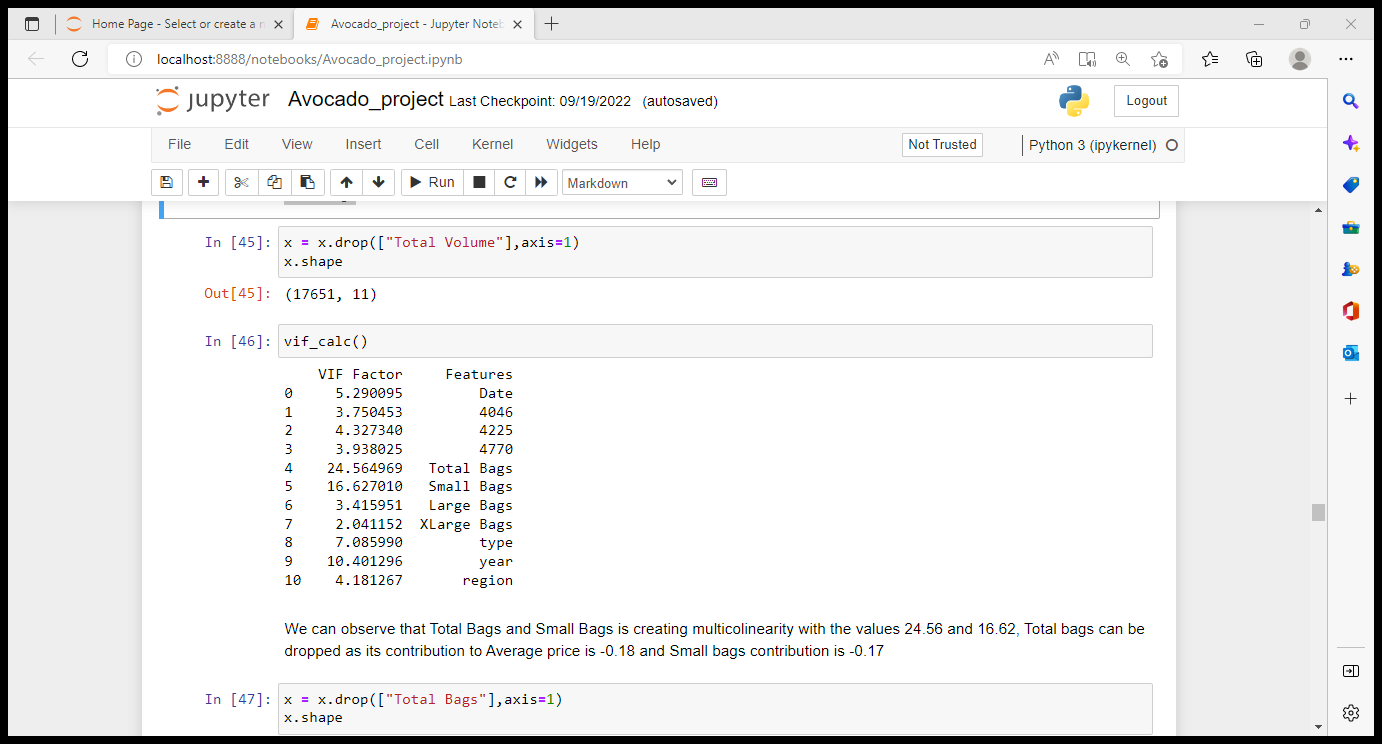


# Avocado AveragePrice Prediction using regressors

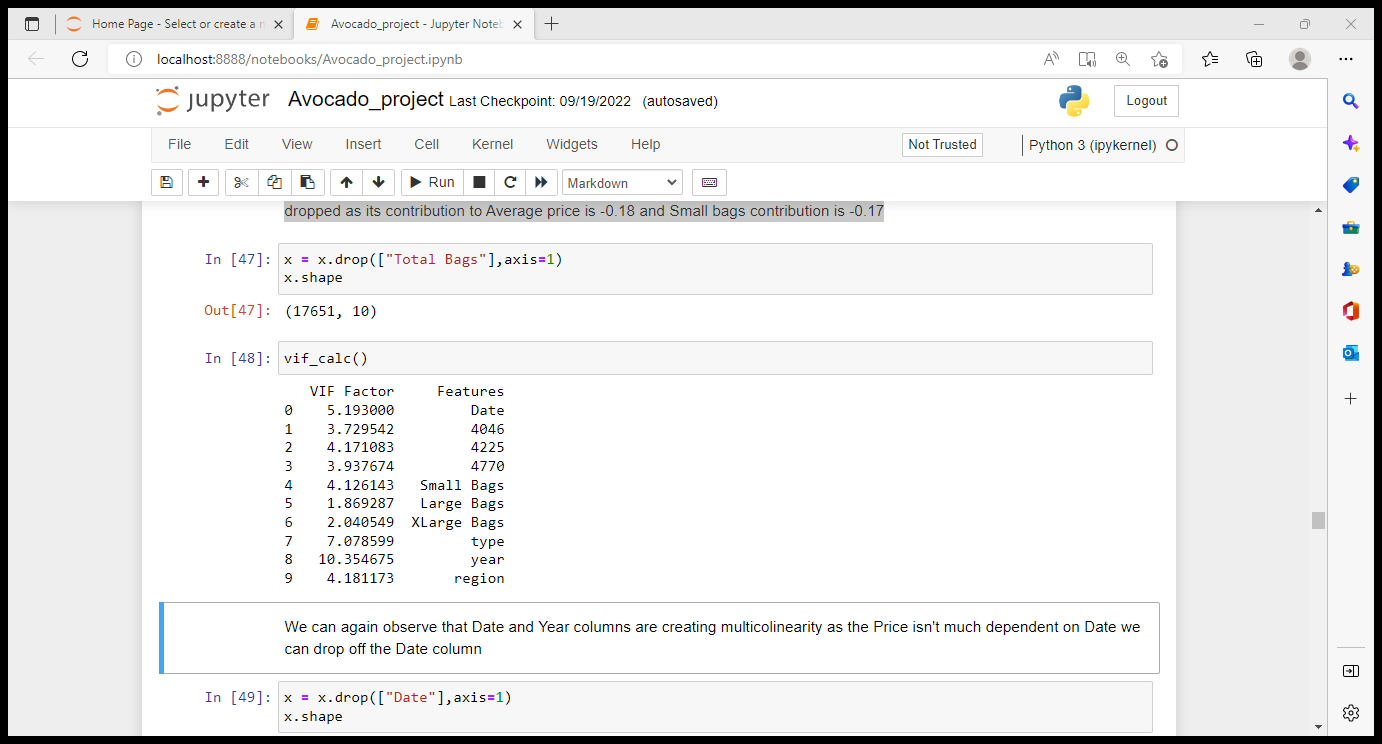
We can now split the input and target variables and check for multicollinearity present in the data.



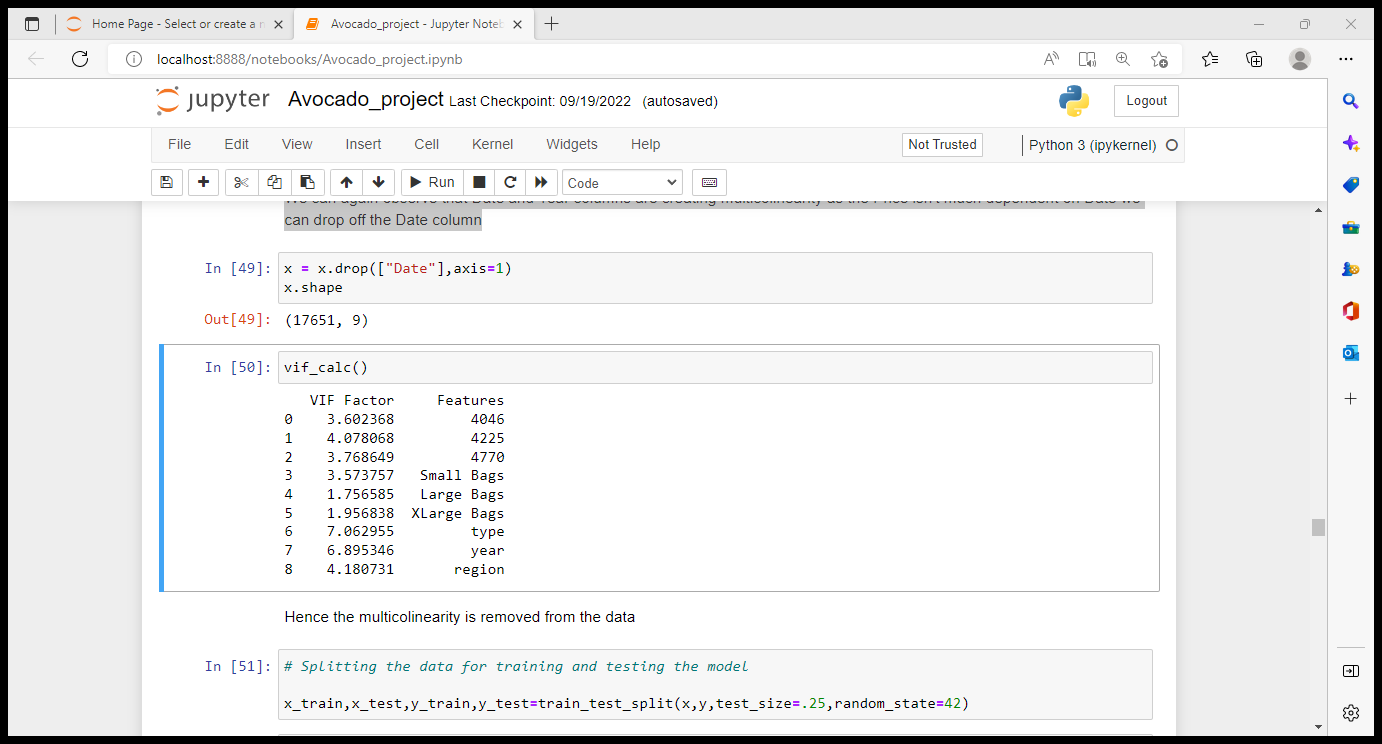
We can observe that Total Volume and Total Bags is creating multicollinearity with the values 53.822 and 35.13, from the df.corr() heat map we can observe that Total Volume has -0.19 correlation and Total Bags has -0.18 correlation with Average Price hence Total Volume column can be dropped as its contribution is lesser to Averageprice than the contribution of Total Bags.



We can observe that Total Bags and Small Bags is creating multicollinearity with the values 24.56 and 16.62, Total bags can be dropped as its contribution to Average price is -0.18 and Small bags contribution is -0.17.

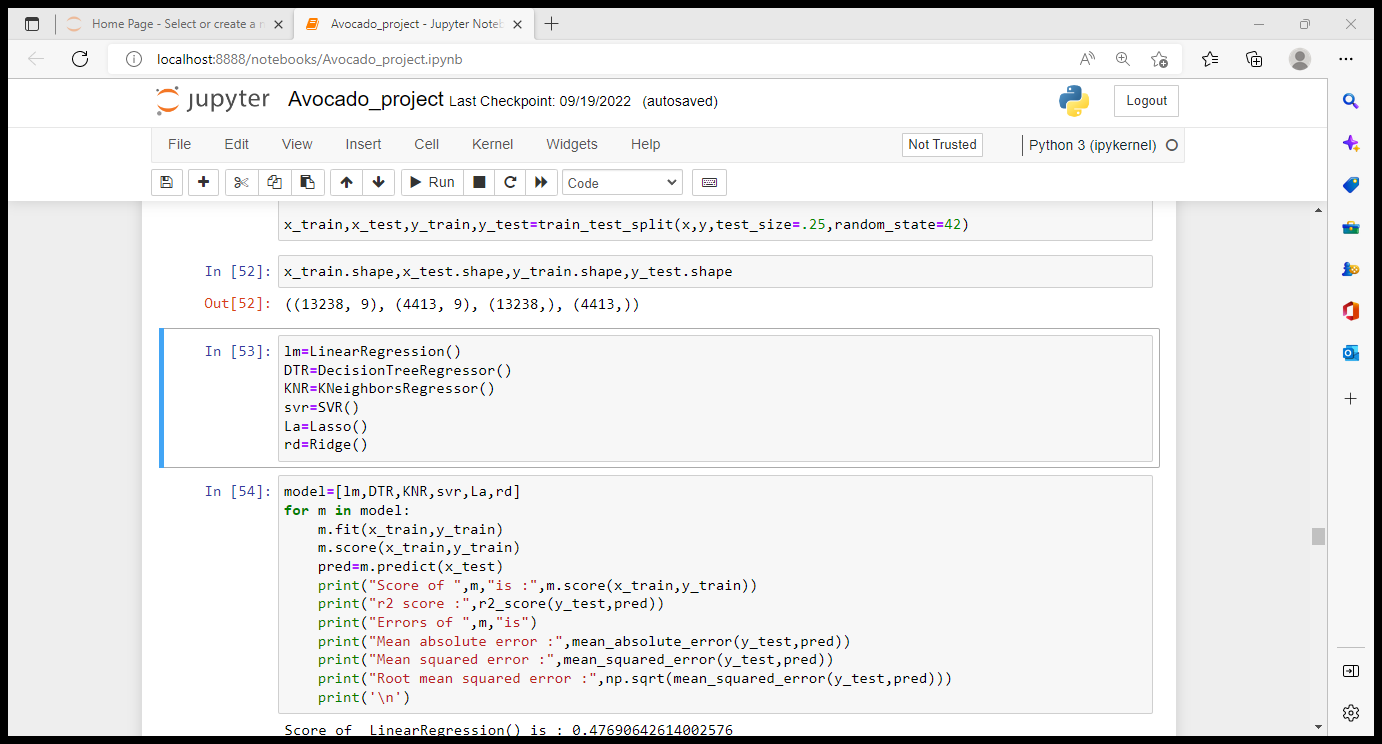


We can again observe that Date and Year columns are creating multicollinearity as the Price isn't much dependent on Date we can drop off the Date column.



Hence the multicollinearity is removed from the dataset and we are left with 9 columns as input variables and AveragePrice remains as our target variable.

Splitting the data for training and testing the models to predict AveragePrice :



The model Score and r2 score od the above built models are as follows,

Score of LinearRegression() is : 0.47690642614002576

r2 score : 0.4690412634055062

Score of DecisionTreeRegressor() is : 1.0

r2 score : 0.6872103909929783

Score of KNeighborsRegressor() is : 0.8941862860604416

r2 score : 0.8290833823087831

Score of SVR() is : 0.0397605353460464

r2 score : 0.040570823280558166

Score of Lasso() is : 0.0

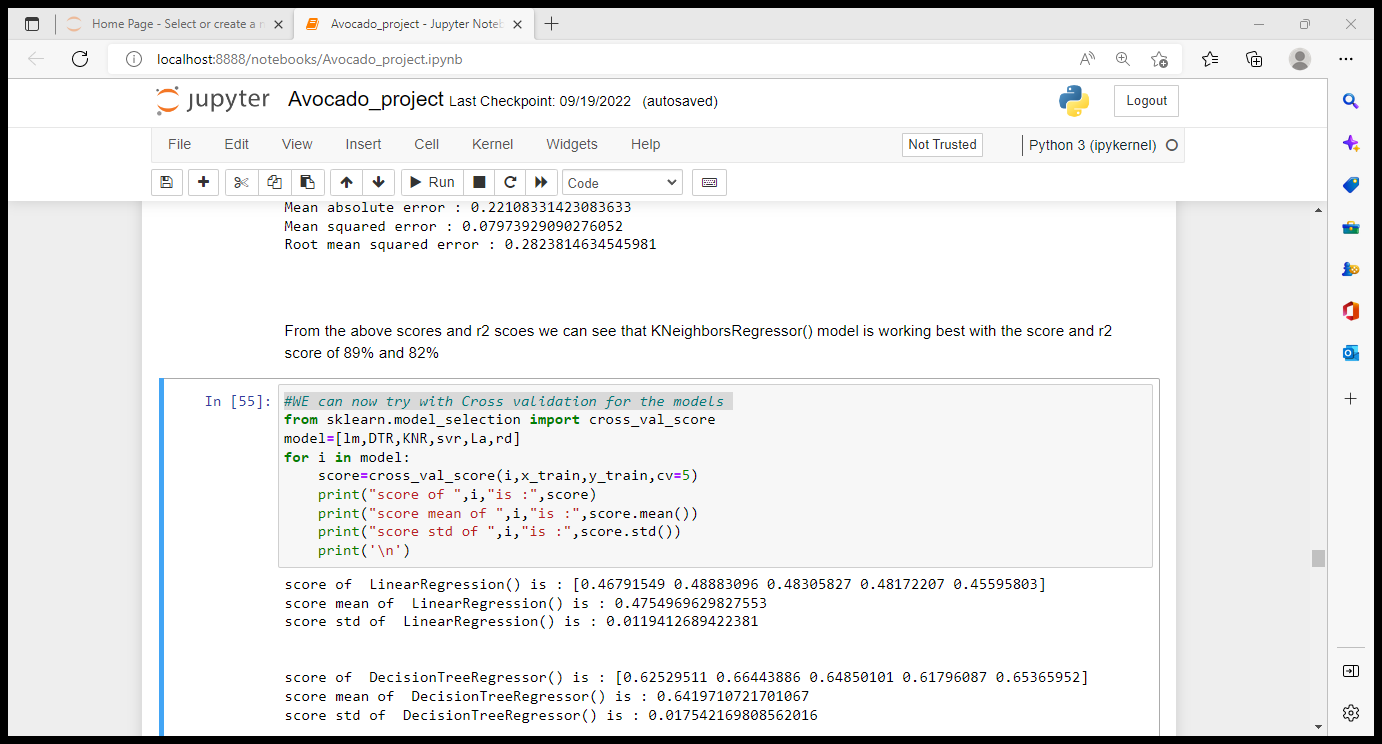
r2 score : -5.600800817040508e-05

Score of Ridge() is : 0.4769063251976897

r2 score : 0.4690435598876084

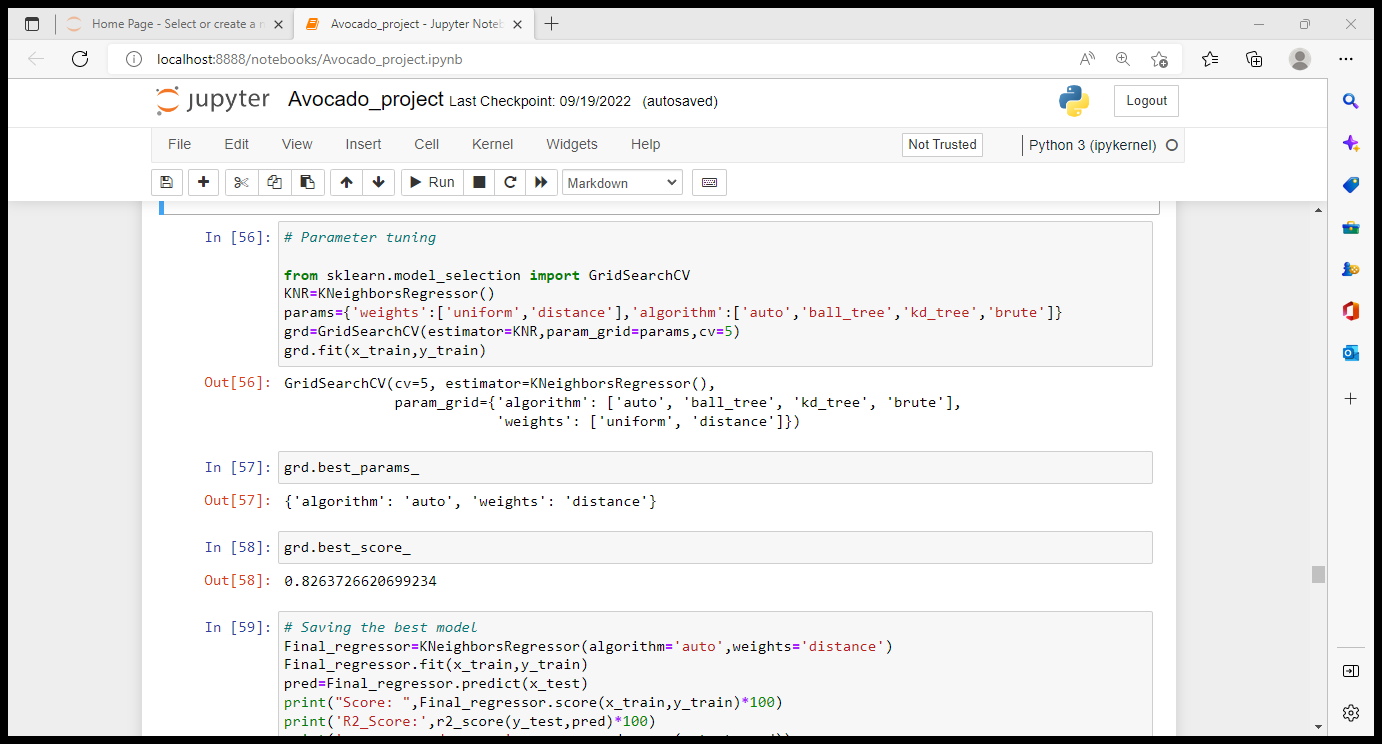
From the above scores and r2 scores we can see that KNeighborsRegressor() model is working best with the score and r2 score of 89% and 82%.

We can now try with Cross validation for the models



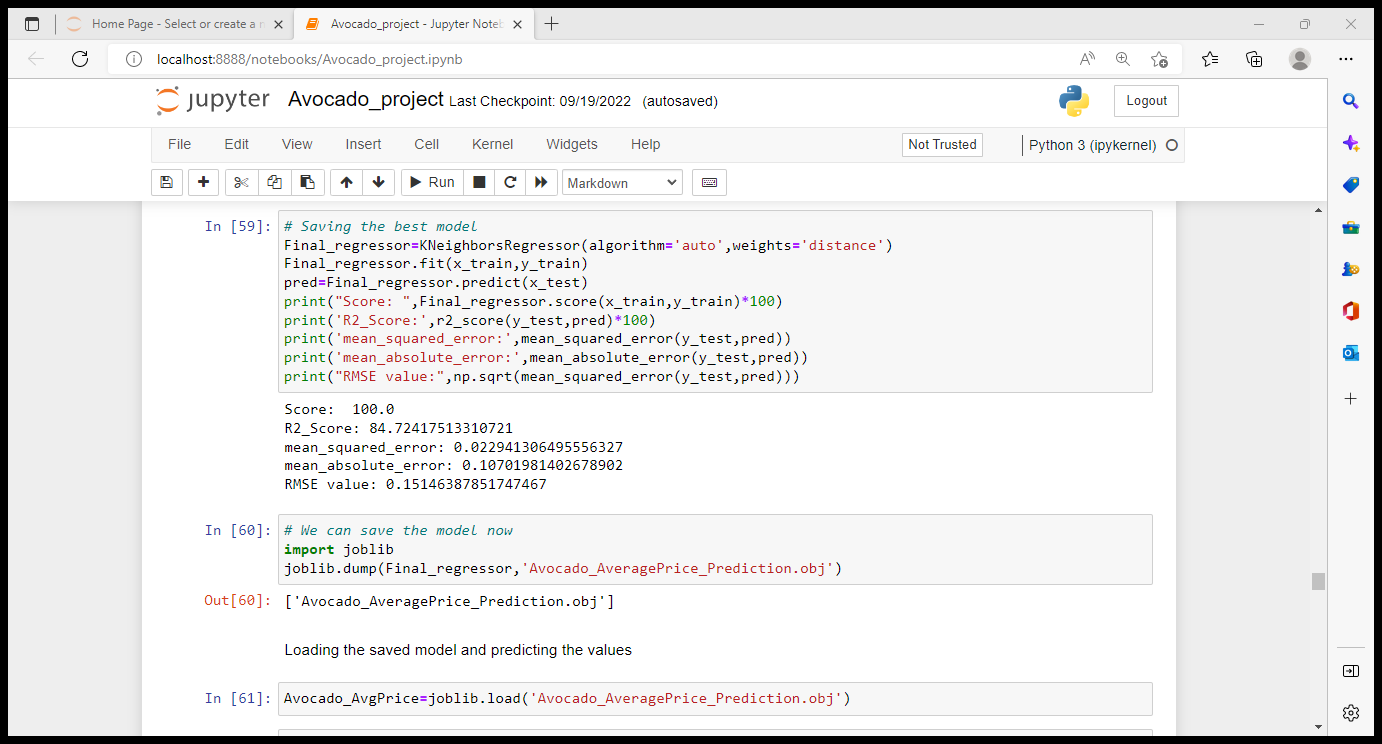
Even with CV we can see that KNeighborsRegressor() model is performing well with mean score of 80%.

Parameter Tuning for KNeighbors Regressor :

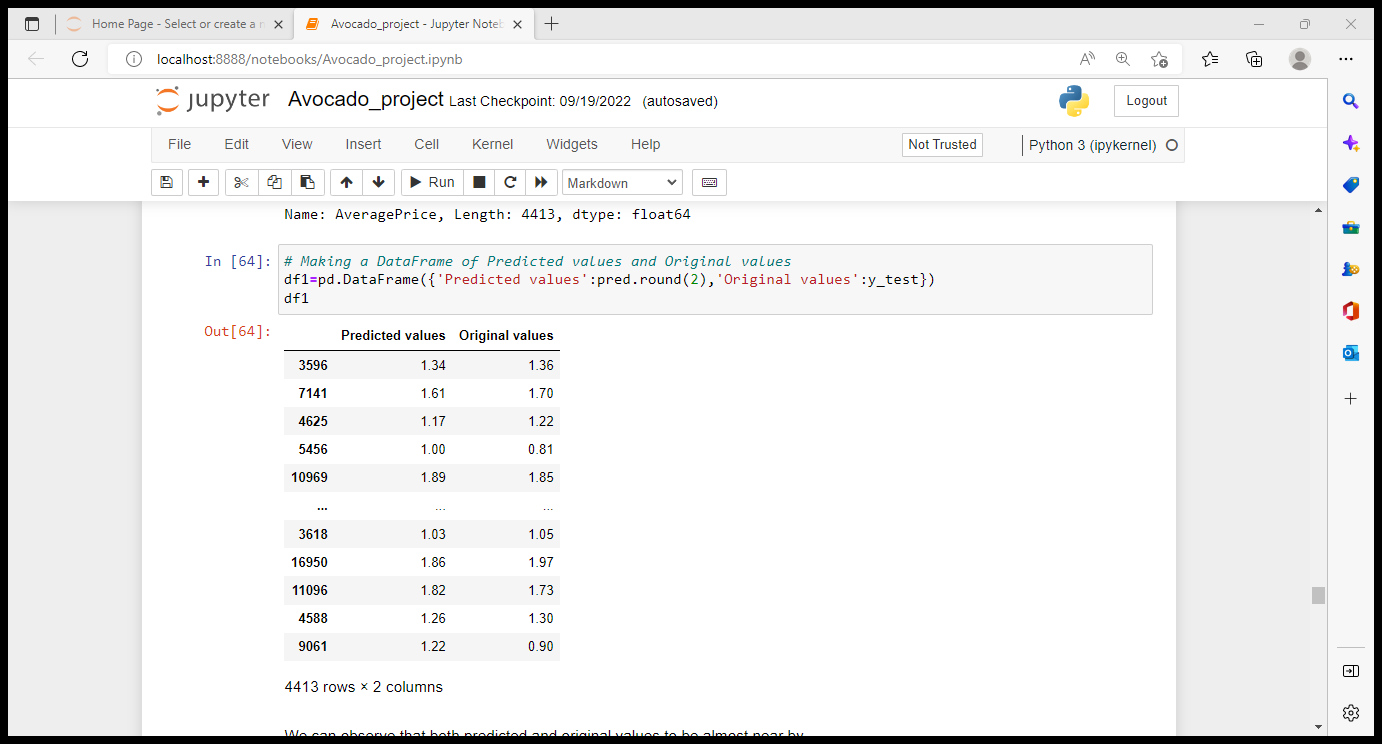


We can observe that KNeighbors regressor is giving the best score of 82.63% with the parameters algorithm = ‘auto’ and weights = ‘distance’.

Saving the best model :



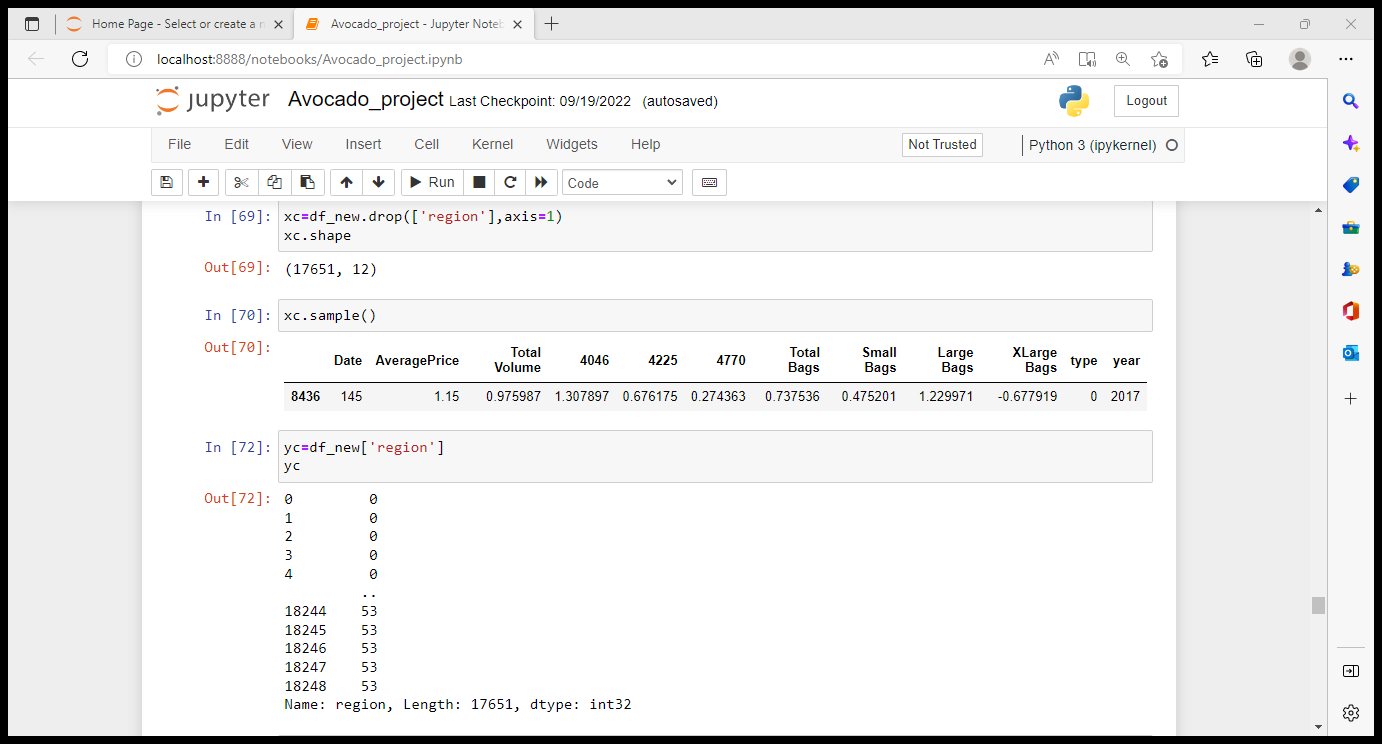
Checking for predicted outputs :



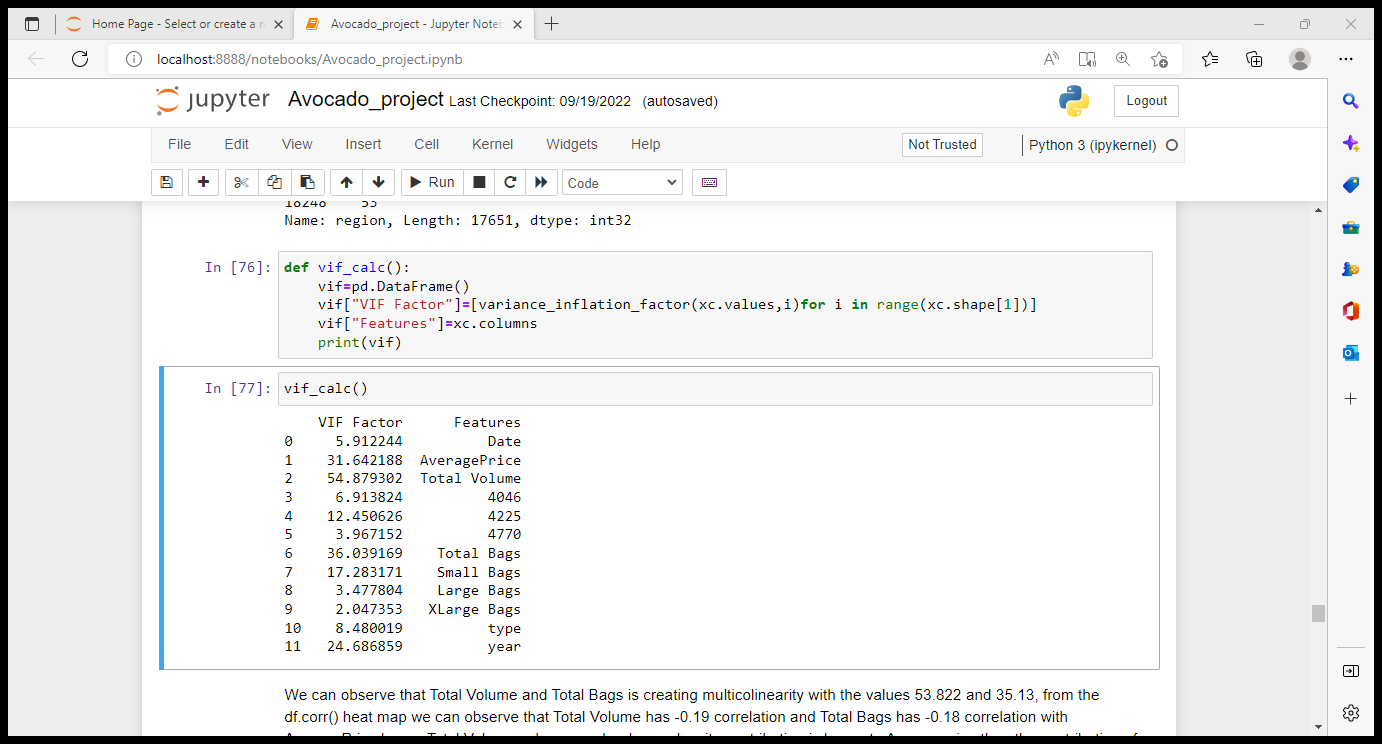
# Avocado Region prediction using classifiers

We can now split the target variable and input variables to predict Region and check for presence of multicollinearity in the data.

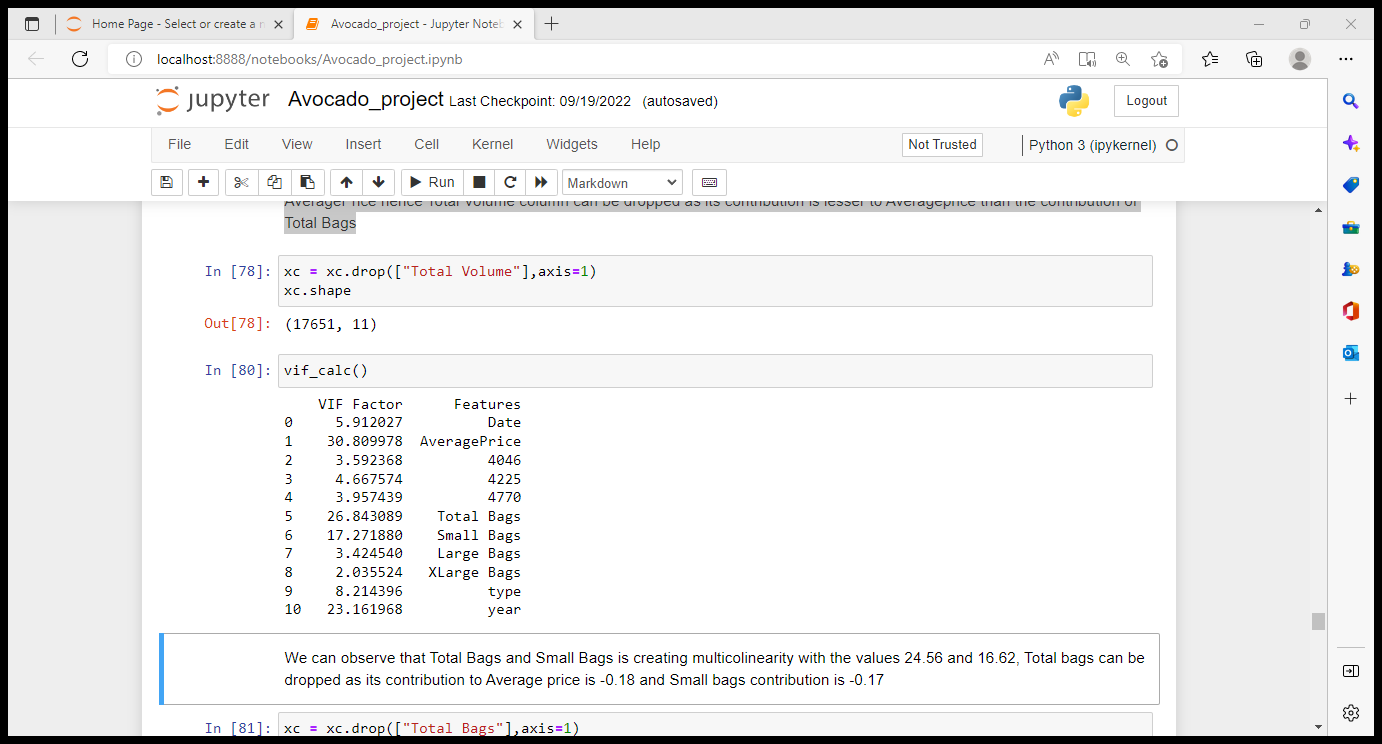
Spliting the data :



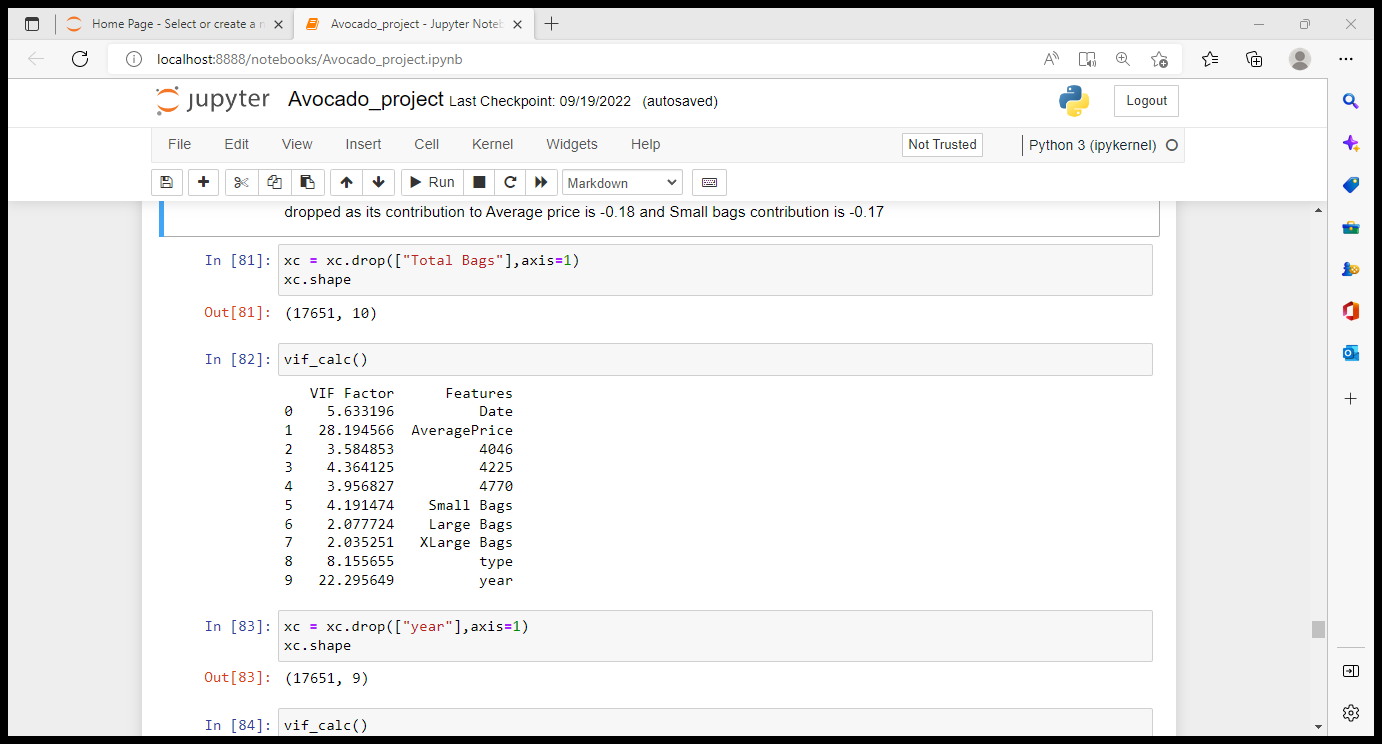
Checking for multicollinearituy :



We can observe that Total Volume and Total Bags is creating multicollinearity with the values 53.822 and 35.13, from the df.corr() heat map we can observe that Total Volume has -0.19 correlation and Total Bags has -0.18 correlation with AveragePrice hence Total Volume column can be dropped as its contribution is lesser to Averageprice than the contribution of Total Bags.

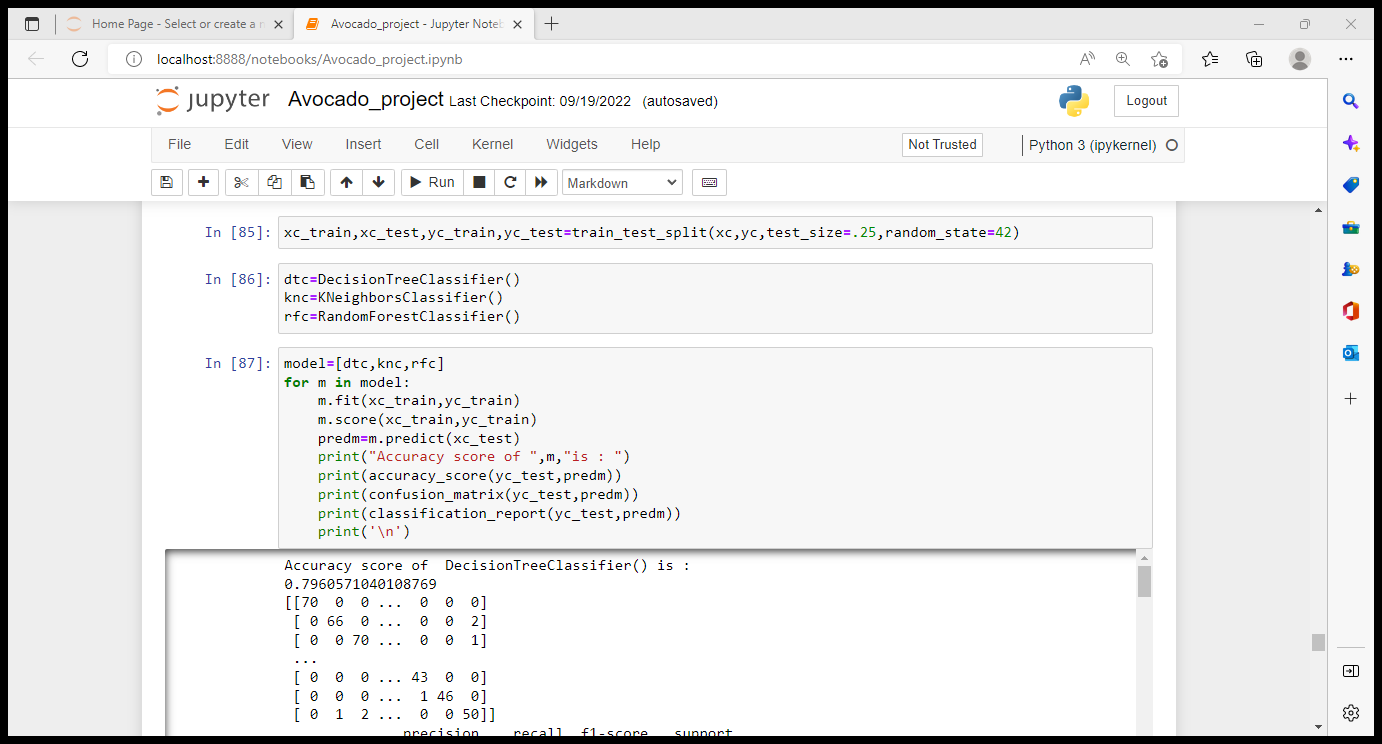


We can observe that Total Bags and Small Bags is creating multicollinearity with the values 24.56 and 16.62, Total bags can be dropped as its contribution to Average price is -0.18 and Small bags contribution is -0.17.



We can observe that AveragePrice and year are creating multicollinearity with the values 28.19 and 22.29 respectively, As we can observe from df.corr() the contribution of variable year is lesser to target variable Region than that of AveragePrice, hence we can drop Year at this stage and observe that all trhe multicollinearity is removed from the dataset.

Splitting train and test data and building Classifier models for predicting Region :



Accuracy score of DecisionTreeClassifier() is :

0.7960571040108769

[[70 0 0 ... 0 0 0]

[ 0 66 0 ... 0 0 2]

[ 0 0 70 ... 0 0 1]

...

[ 0 0 0 ... 43 0 0]

[ 0 0 0 ... 1 46 0]

[ 0 1 2 ... 0 0 50]]

Accuracy score of KNeighborsClassifier() is :

0.2821210061182869

[[57 0 0 ... 0 0 0]

[ 0 30 0 ... 0 0 0]

[ 0 0 62 ... 0 0 0]

...

[ 0 2 1 ... 20 1 0]

[ 0 1 0 ... 0 20 0]

[ 0 4 9 ... 0 0 11]]

Accuracy score of RandomForestClassifier() is :

0.9032404260140494

[[79 0 0 ... 0 0 0]

[ 0 75 0 ... 0 0 1]

[ 0 0 80 ... 0 0 0]

...

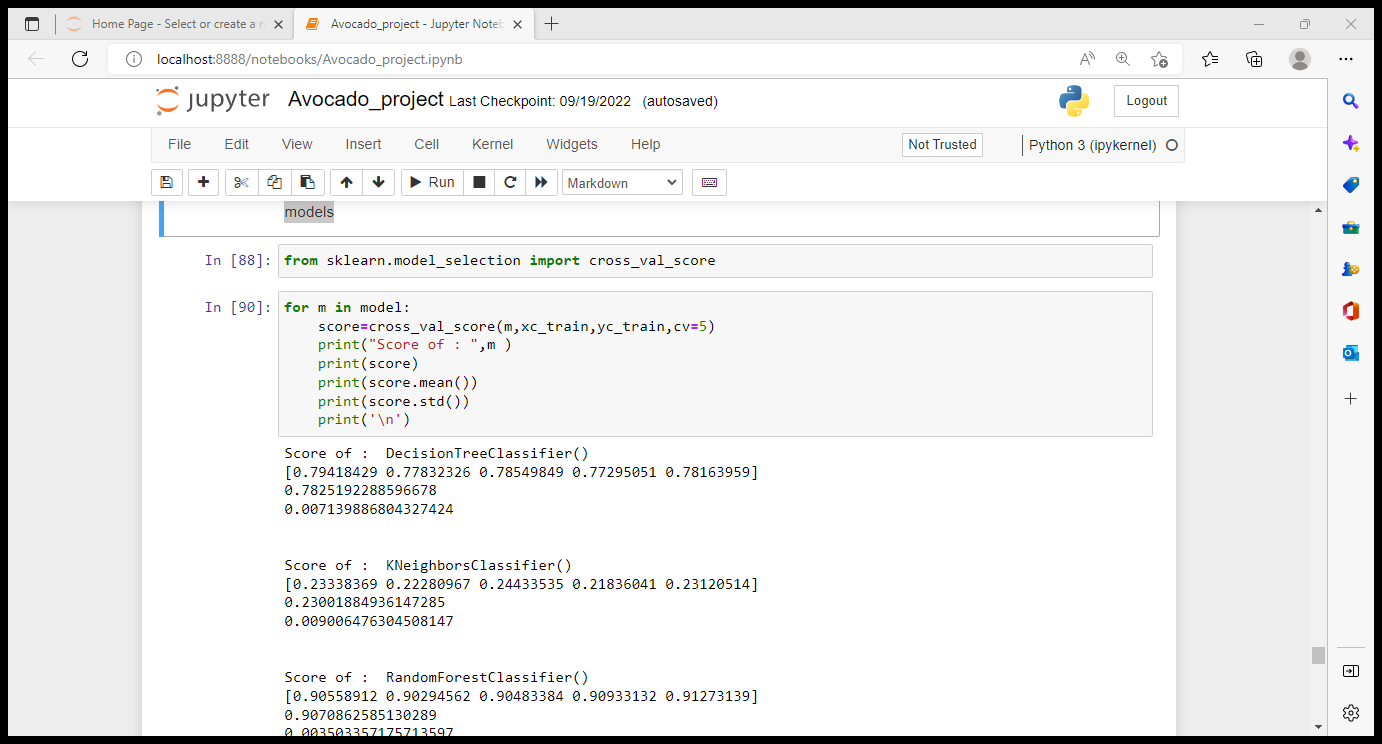
[ 0 0 0 ... 44 0 0]

[ 0 0 0 ... 1 47 0]

[ 0 0 0 ... 0 0 65]]

From the above matrix we can consider RandomForestClassifier() to be the best model as its accuracy score is 90.32% and higher compared to other models.

We can now check for the performance of above models with cross validation.



Score of : DecisionTreeClassifier()

[0.79418429 0.77832326 0.78549849 0.77295051 0.78163959]

0.7825192288596678

0.007139886804327424

Score of : KNeighborsClassifier()

[0.23338369 0.22280967 0.24433535 0.21836041 0.23120514]

0.23001884936147285

0.009006476304508147

Score of : RandomForestClassifier()

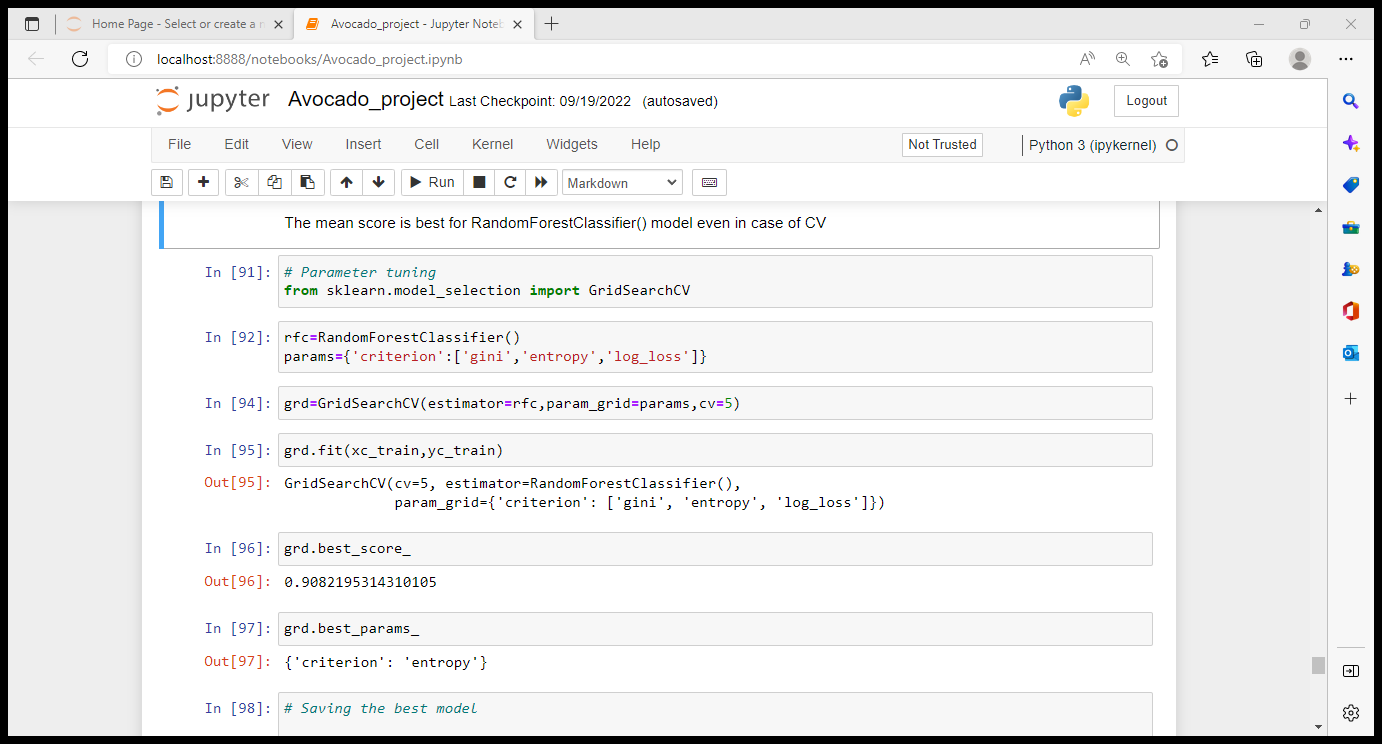
[0.90558912 0.90294562 0.90483384 0.90933132 0.91273139]

0.9070862585130289

0.003503357175713597

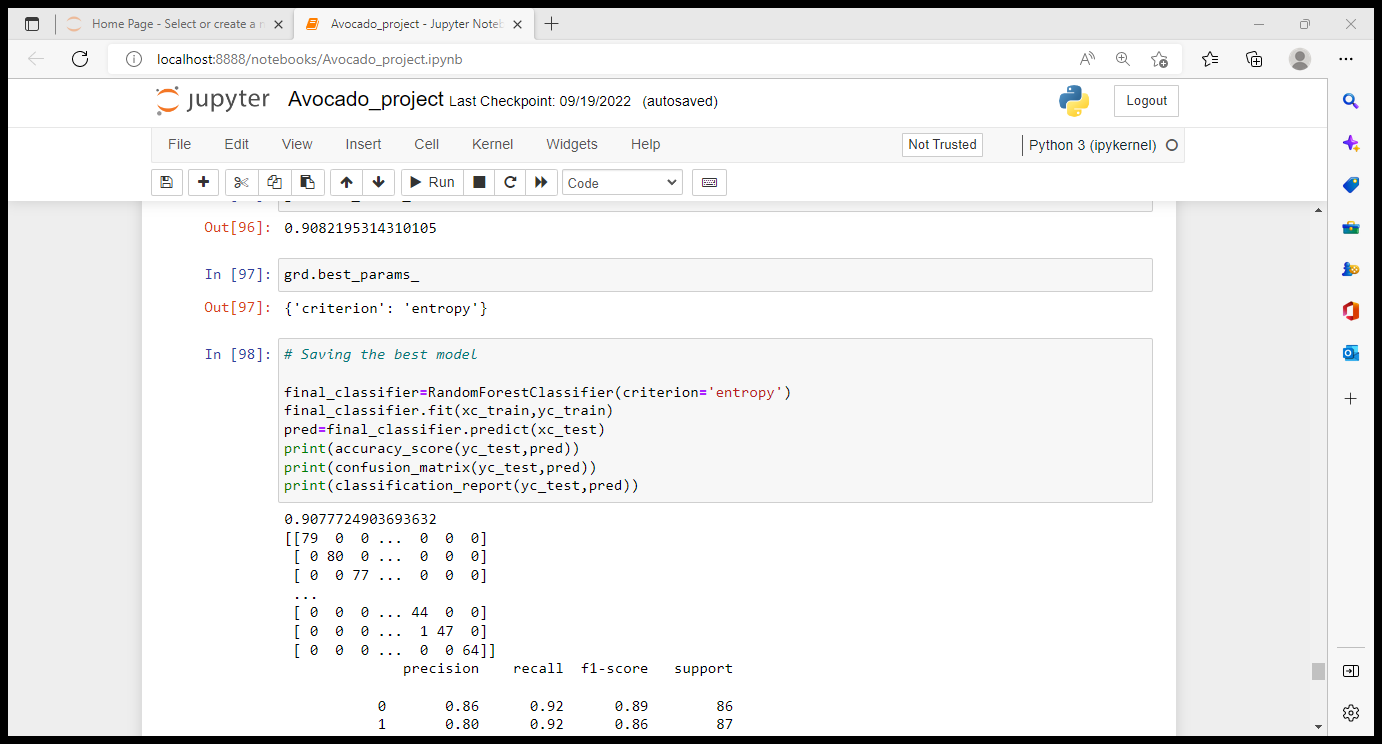
The mean score is best for RandomForestClassifier() model with 90% mean score even in case of Cross validation.

Parameter tuning for RandomForestClassifier() :

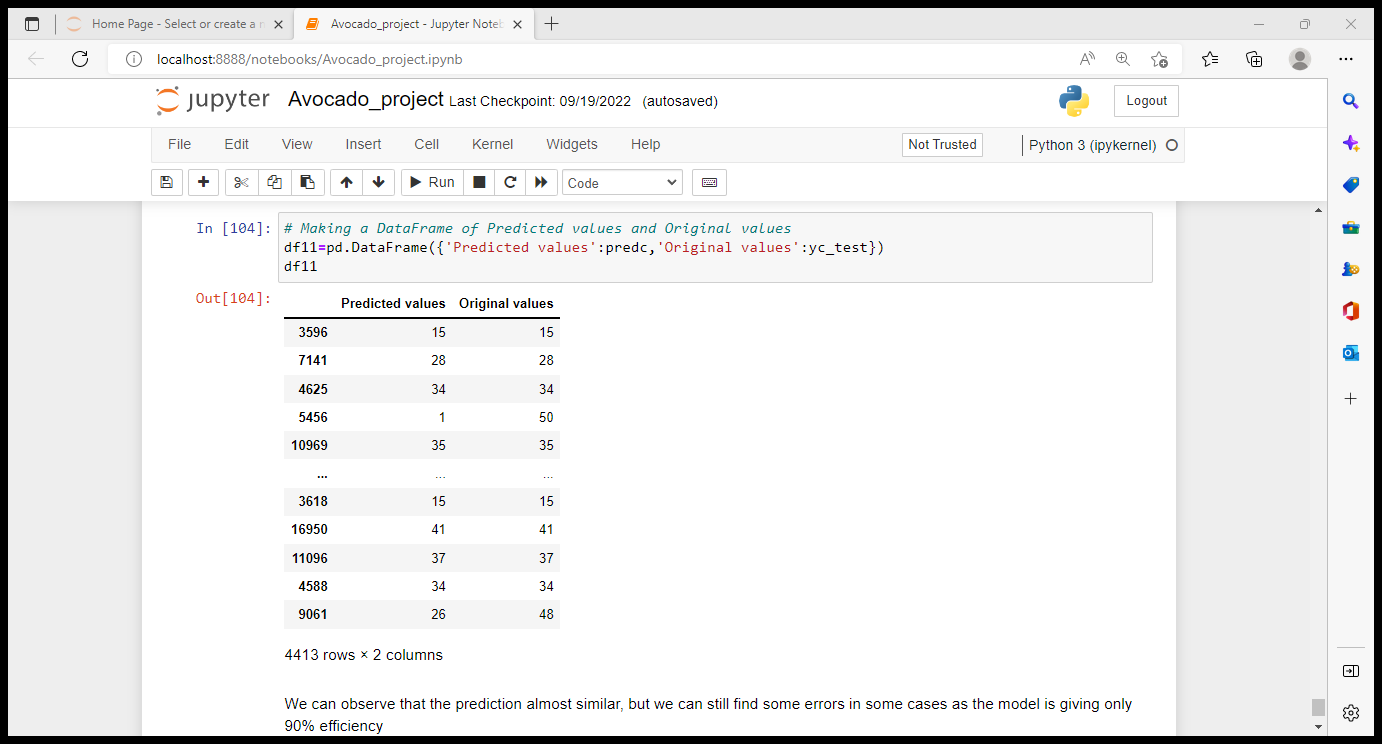


We can see that Random forest Classifier is giving best score as 90.8% with the ‘criterion’ = ‘entropy’

Saving the best model to predict region :



Checking the original and Predicted values :



**Conclusion**

The data we had to build was with the shape of 18249 rows and 14 columns, As we studied the data in detailed using EDA and Visualization techniques, we understood that we had to perform several operations such as Encoding, Removing outliers, removing skewness in the data. Hence going further we used **Label encoder** to encode the data, Used **Z-score** method to remove the outliers and **Yeo-johnson** method to remove the skewness.

After performing all the process mentioned above the cleansed data we had was 1765 rows and 13 columns.

Using the cleansed data we further checked for multicollinearity in the data and had to drop off few variables then built several models like Linear Regression, Decision Tree Regressor, K Neighbors Regressor, Support vector Regressor, Lasso and Ridge for predicting Average price of the Avocado. From the above built models we can observe that in case of Average price prediction we chose **KNeighbors Regressor** as the best model as its **score is 100%** with **‘algorithm’ = ‘auto’** and **‘weights’ = ‘distance’** and **r2 score is 84.72%.**

In case of predicting region of the Avocado we built different models such as Decision Tree classifier, K Neighbors Classifier and Random Forest Classifier and we found **Random Forest classifier** with **‘criterion’ = ‘entropy’** to be the best model as it is giving a **score of 90.77%**, which is best compared to other models. In the predicted values of region we can observe some errors as the model is giving only 90.7% accuracy and rest 9.3% less accuracy is responsible for the occurrence of the error.