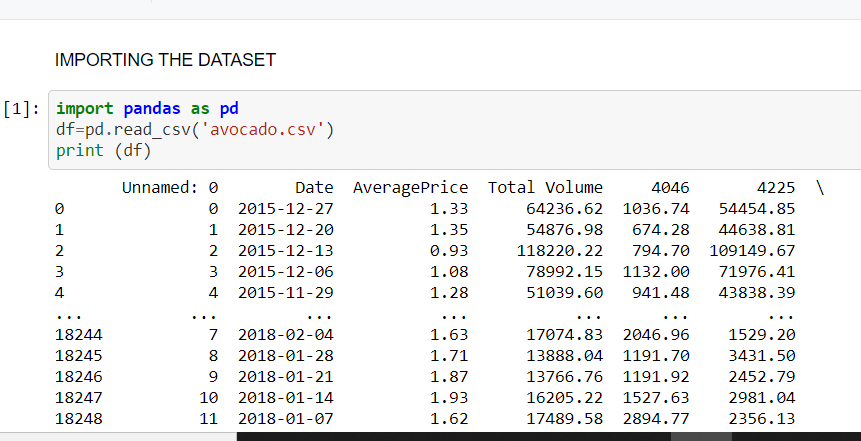
PROBLEM DEFINITION:

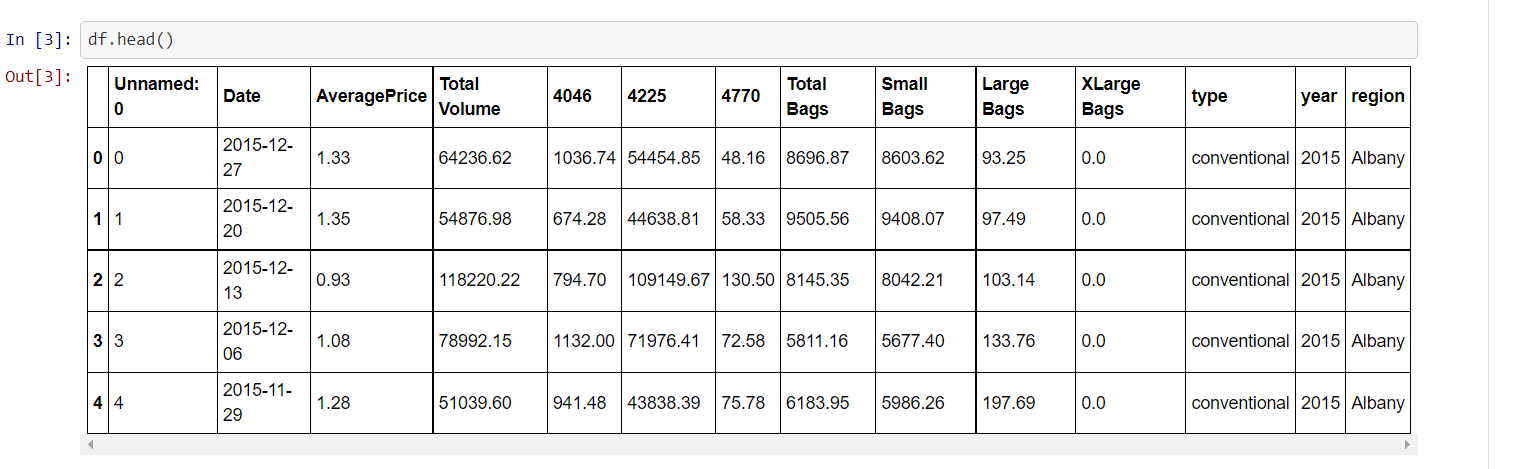
Our goal is to predict the average price with respect to size, region,volume,type etc.

Importing the data set

 I am importing pandas library and importing the csv file.

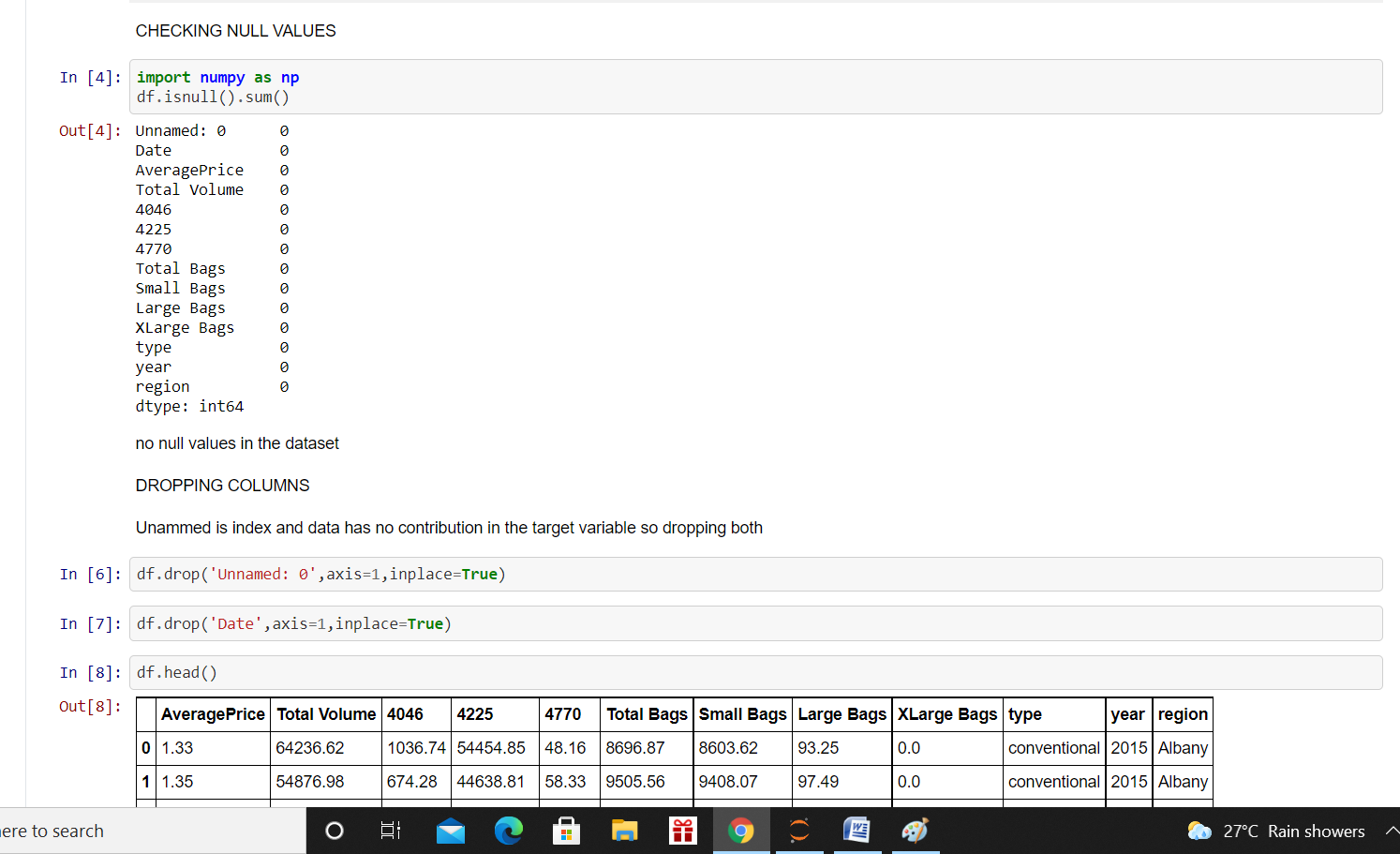
I am using other various librarys and importing them in various aspects which help me in the processes of EDA or MODEL BUILDING. Libraies such as numpy,

LabelEncoder, Matplotlib, seaborn, minmax scaler, zscore, linear regression,R2 score,grid search cv, random regressor etc are used for data visualization, data processing,model building,encoding,scaling and model building.

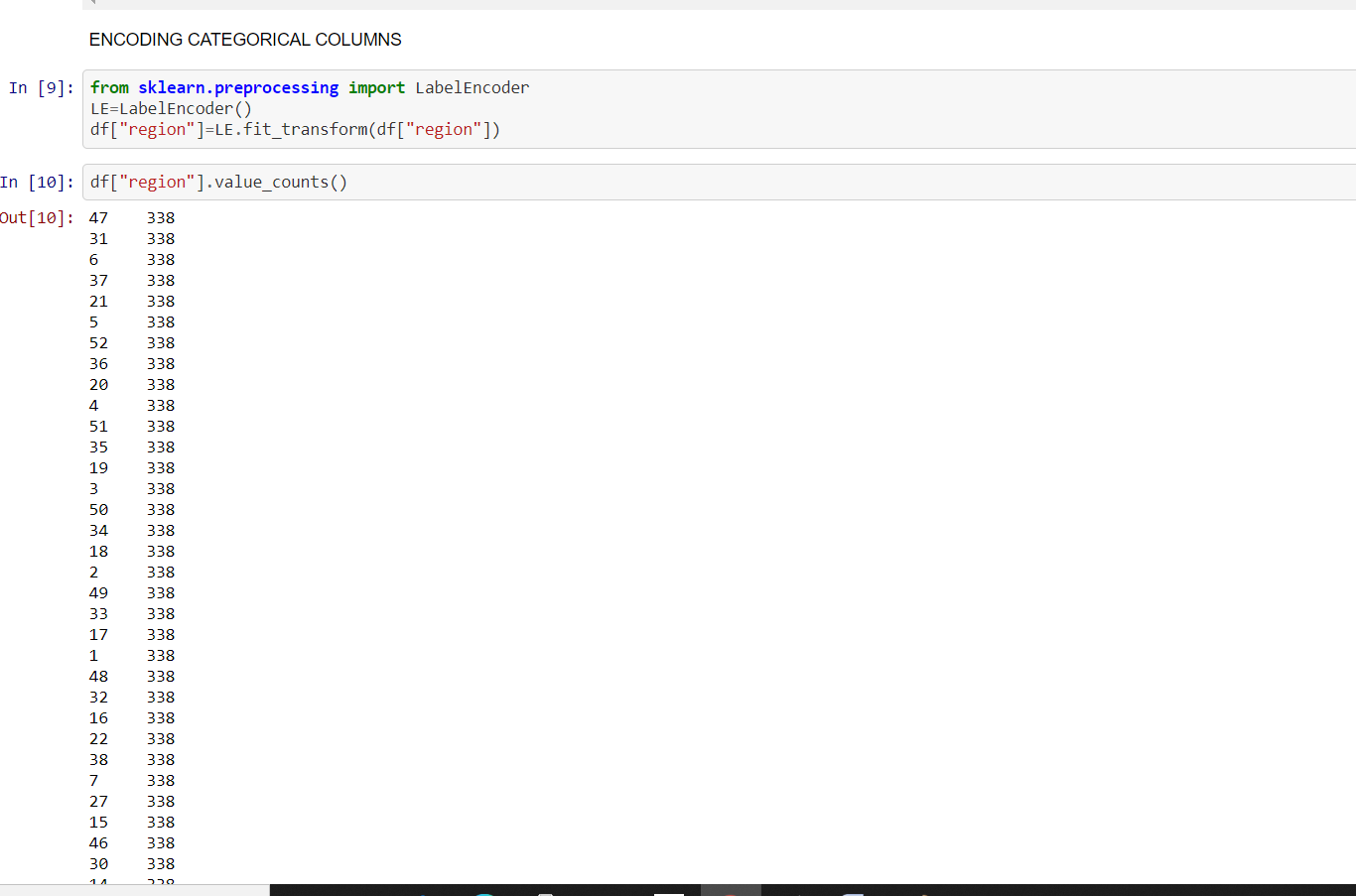


Here I am loading the data into a variable i,e-df and processing the first 5 rows most of the columns are float in nature and ‘sex’ column is categorical.

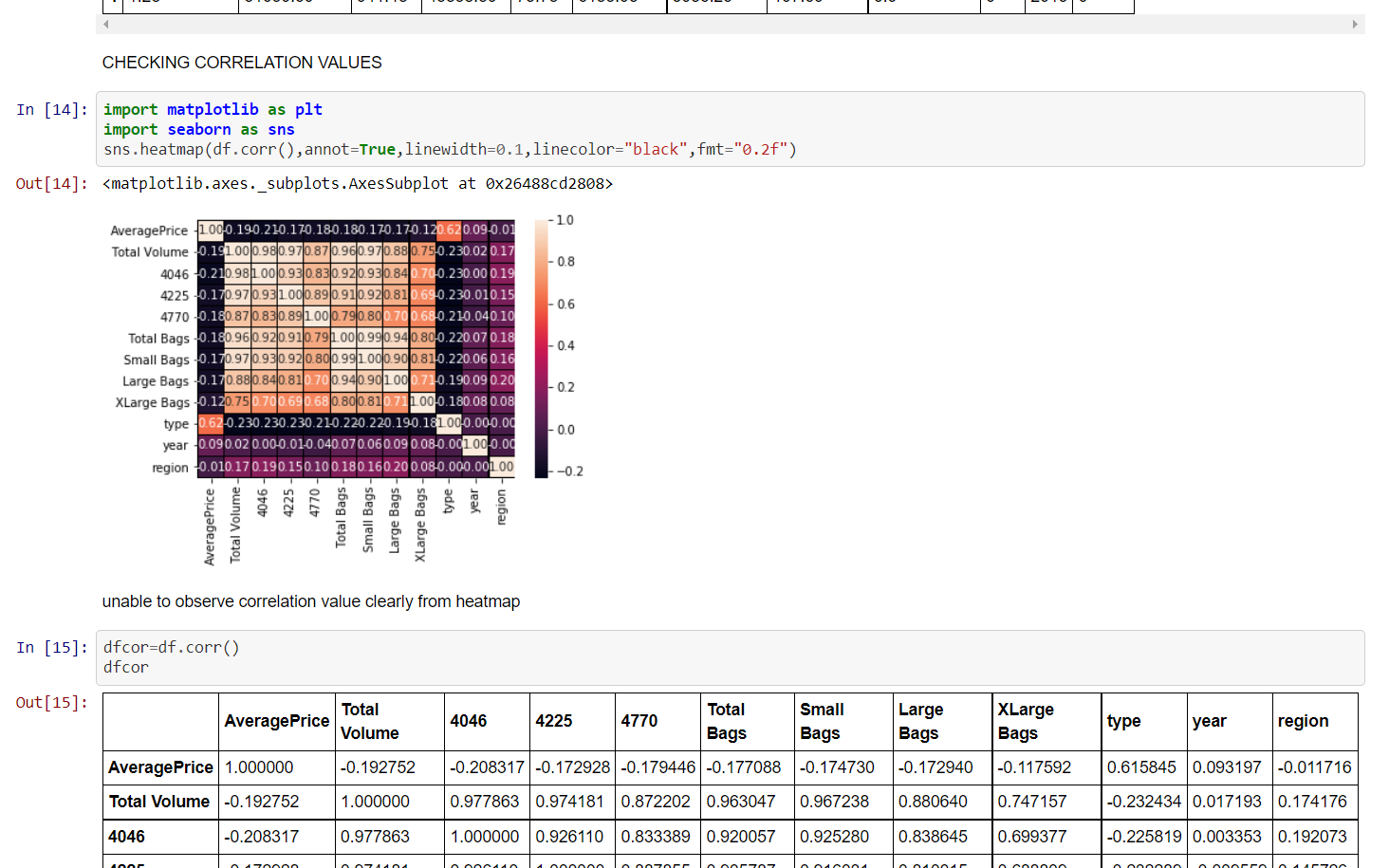
DATA ANALYSIS:



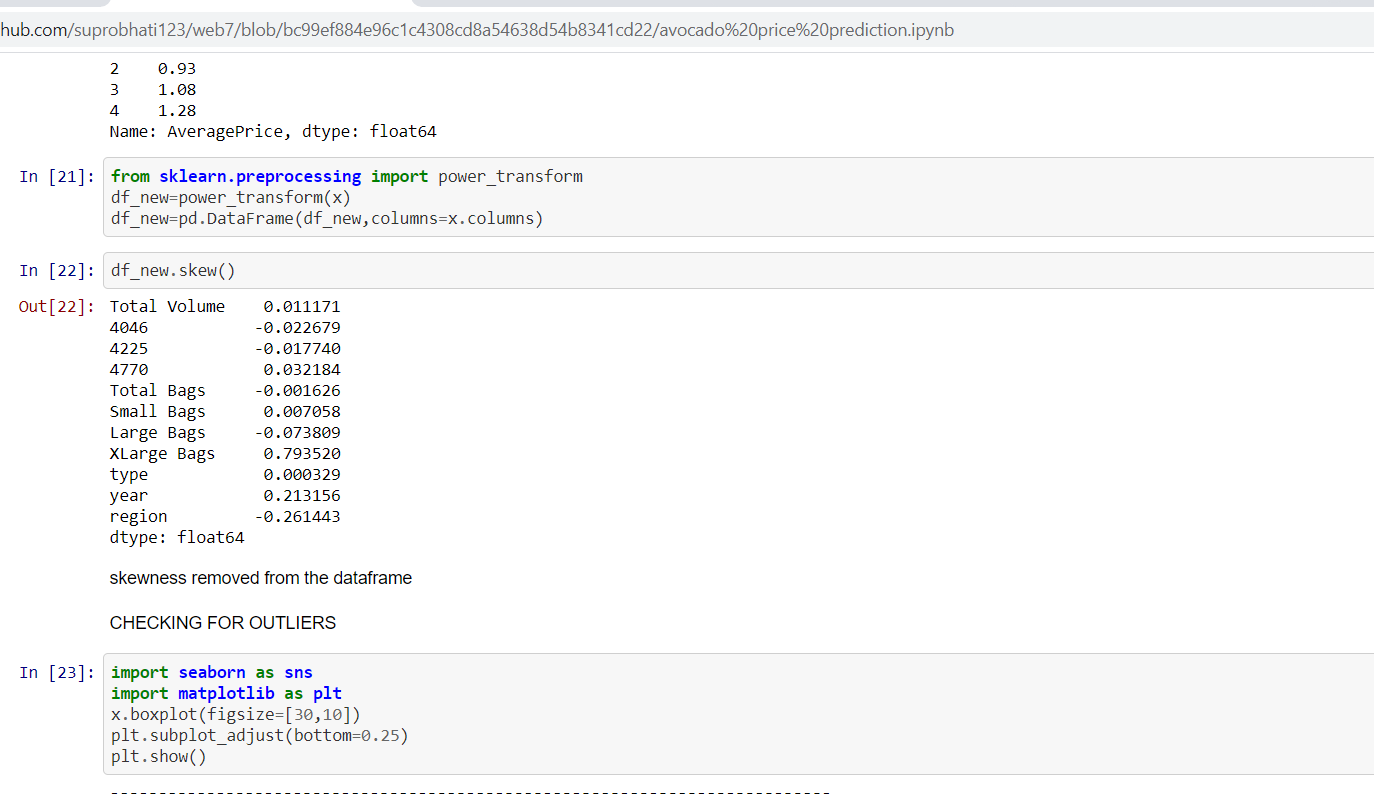
I have checked the null values of all the columns. As there is no null value, I have dropped the unnamed and Date column as they have very less contribution in determining the average price.also I am checking the shape of the data i,e 18249 rows and 14 columns.

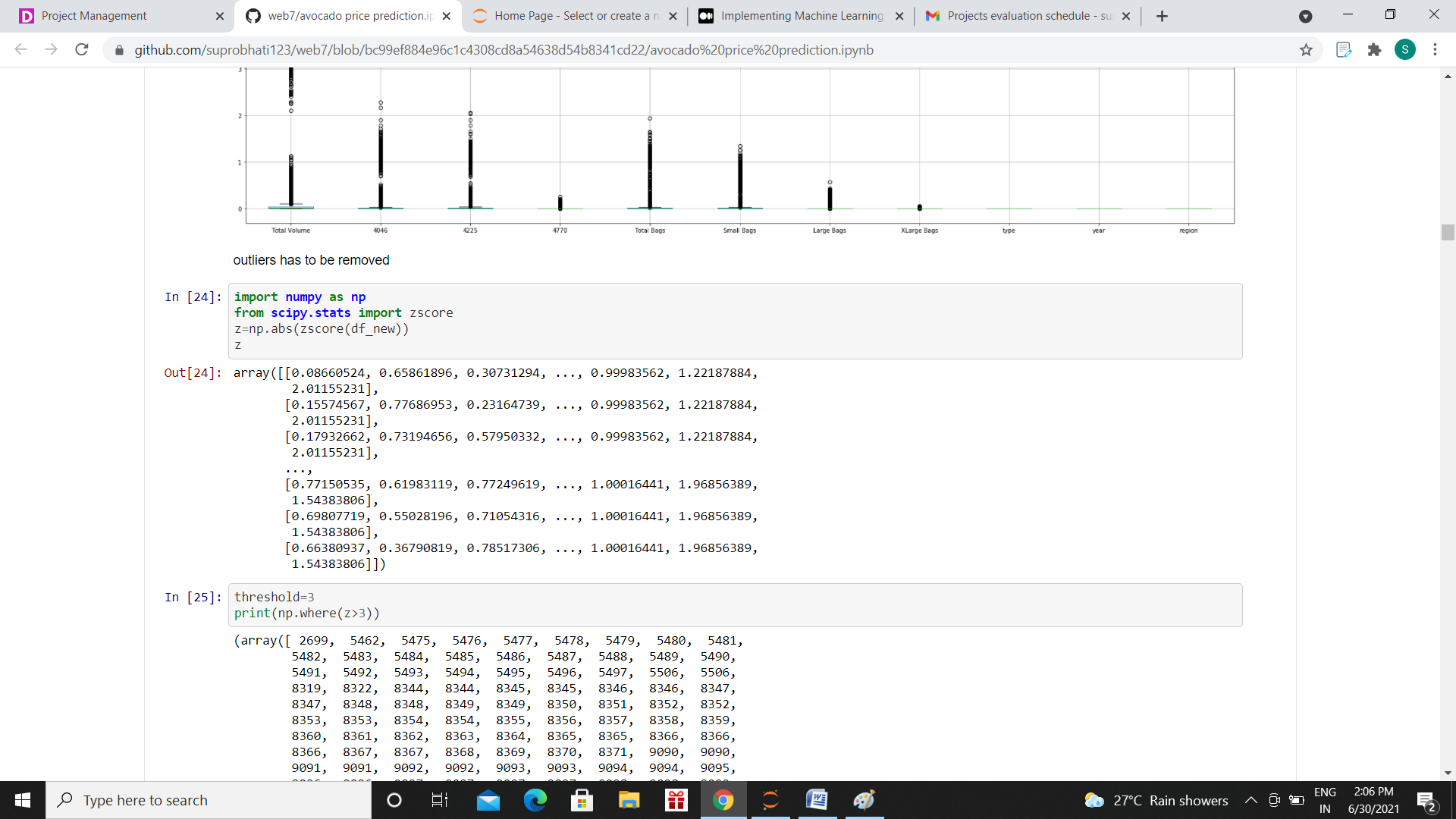


Using label encoder I have encoded the region and type columns as they both are categorical.



I have checked the correlation of different variables using heatmap and corr function find out the variables which has strong correlation with the target variable.

Then using power transform I have removed the skewness and checked for outliers.

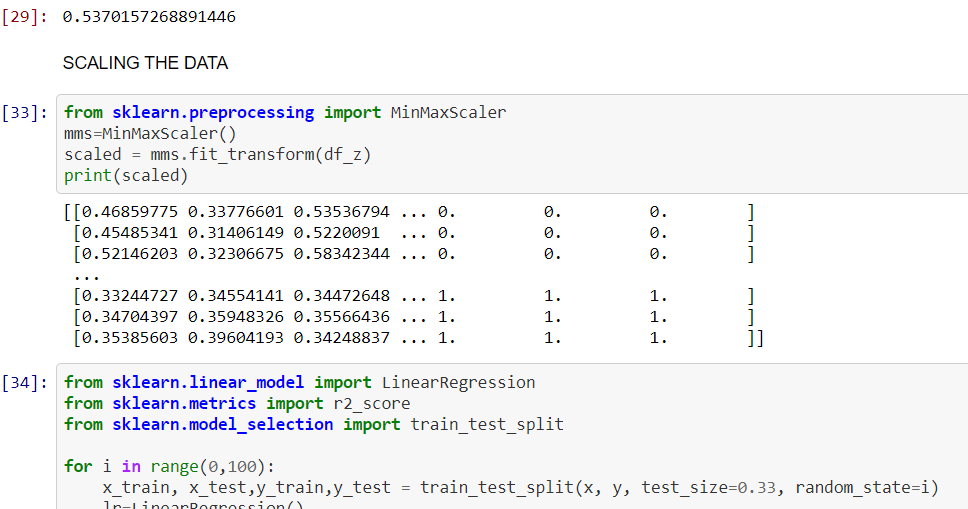


Importing z score value we have removed the outlier part.

EDA CONCLUDING REMARKS:

From the EDA part we have seen that most columns are float in nature and total volume, ,total bags, small bags, large bags, xlarge bags has strong correlation with target variable. The data is skewd so we have to remove the skewness. It has outliers that’s why we are importing z score to remove it. The data loss while removing the outlier is very less.

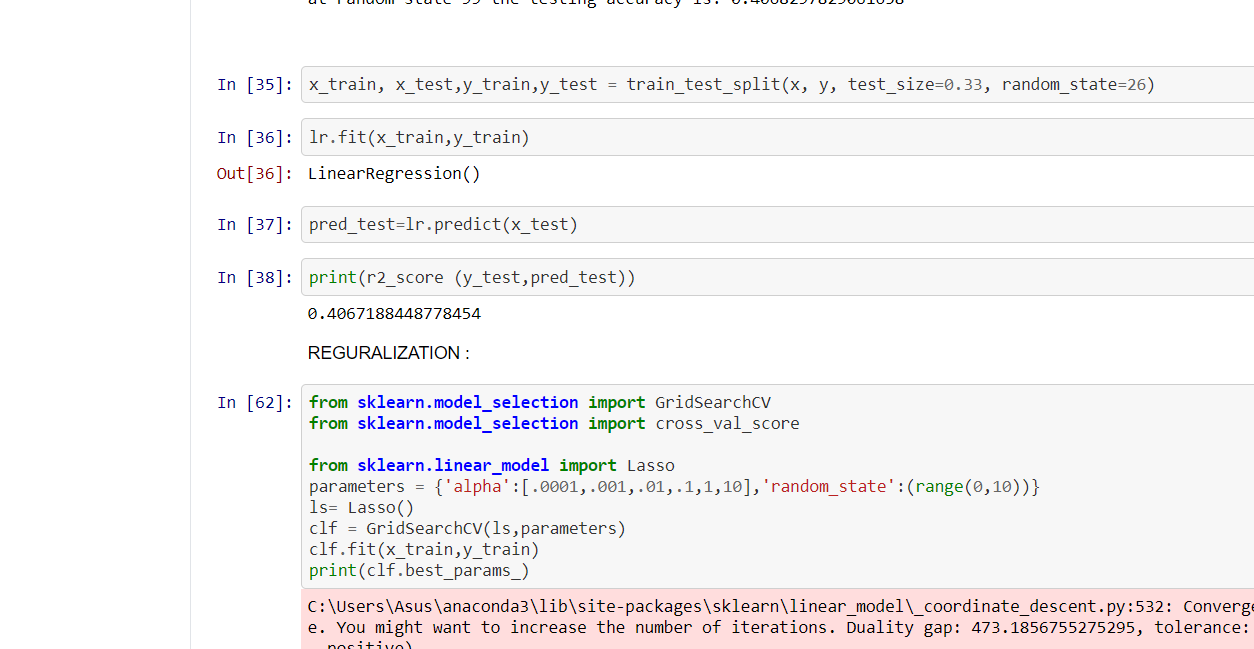
DATA PREPROCESSING:

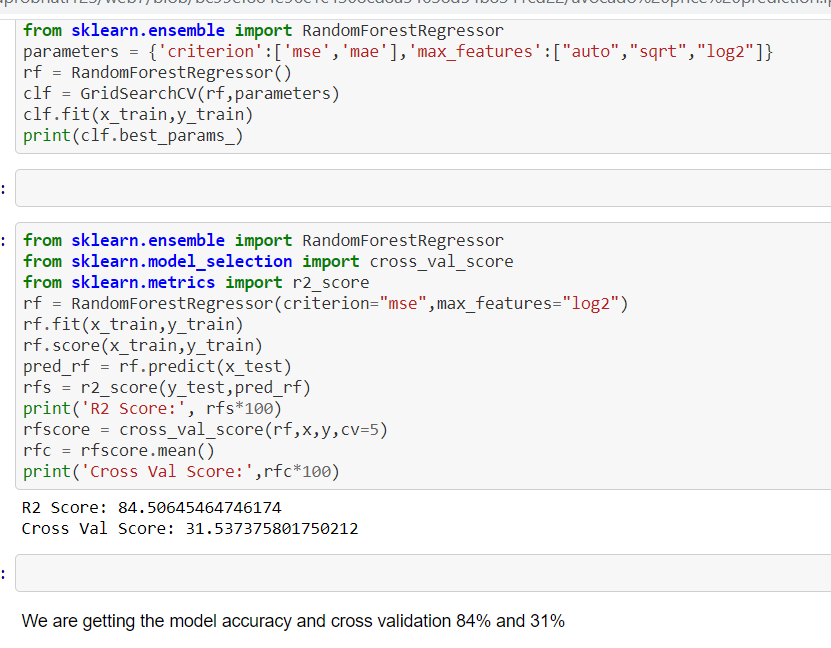


We are scaling the data using min max scaler and then splitting into x and y variable with respect to target variable we are diving it into train and test data. Then choosing the best random state we are processing further processing it.

BUILDING MACHIENE LEARNING MODEL:

We have created models linear regression, lasso regularization,random forest etc using the best random state and parameters. We noticed that the R2 score value of linear regression and lasso is about 40% and 41% .Thats why we are choosing random forest model in which R2 score value is about 84% and croos validation score 31%.





CONCLUSION: Following are the EDA process and model building part of my project. As random forest model is giving us best accuracy to determine the average price we are using the model to predict the data.Thus we can determine the avocado average price.