**Implementation of Machine Learning on Avacado Data set:**

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

The data 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.

**Inspiration /Label**

Some relevant columns in the dataset:

* 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

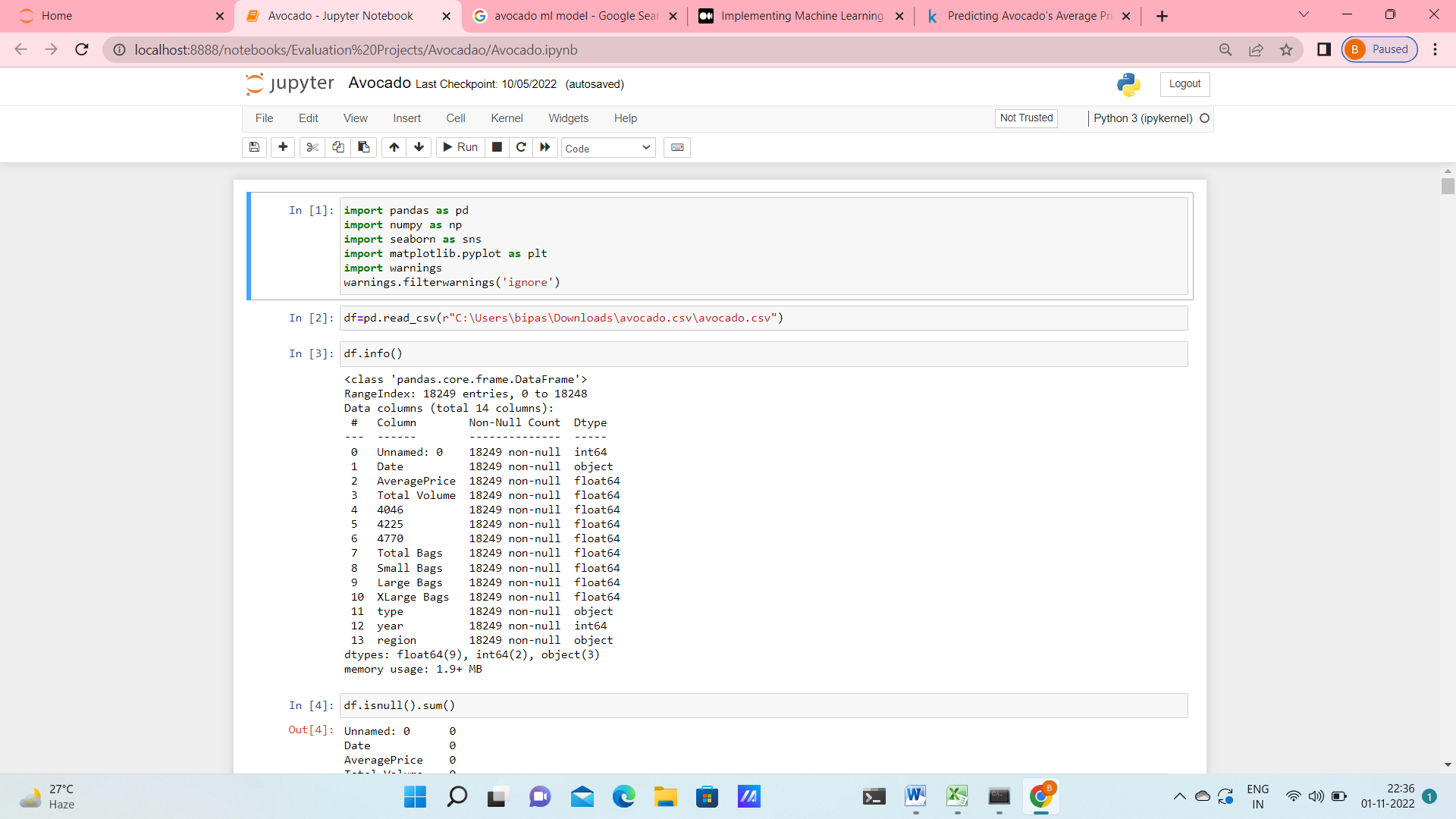
Problem Statement:

The dataset can be seen in two angles to find the region and find the average price .

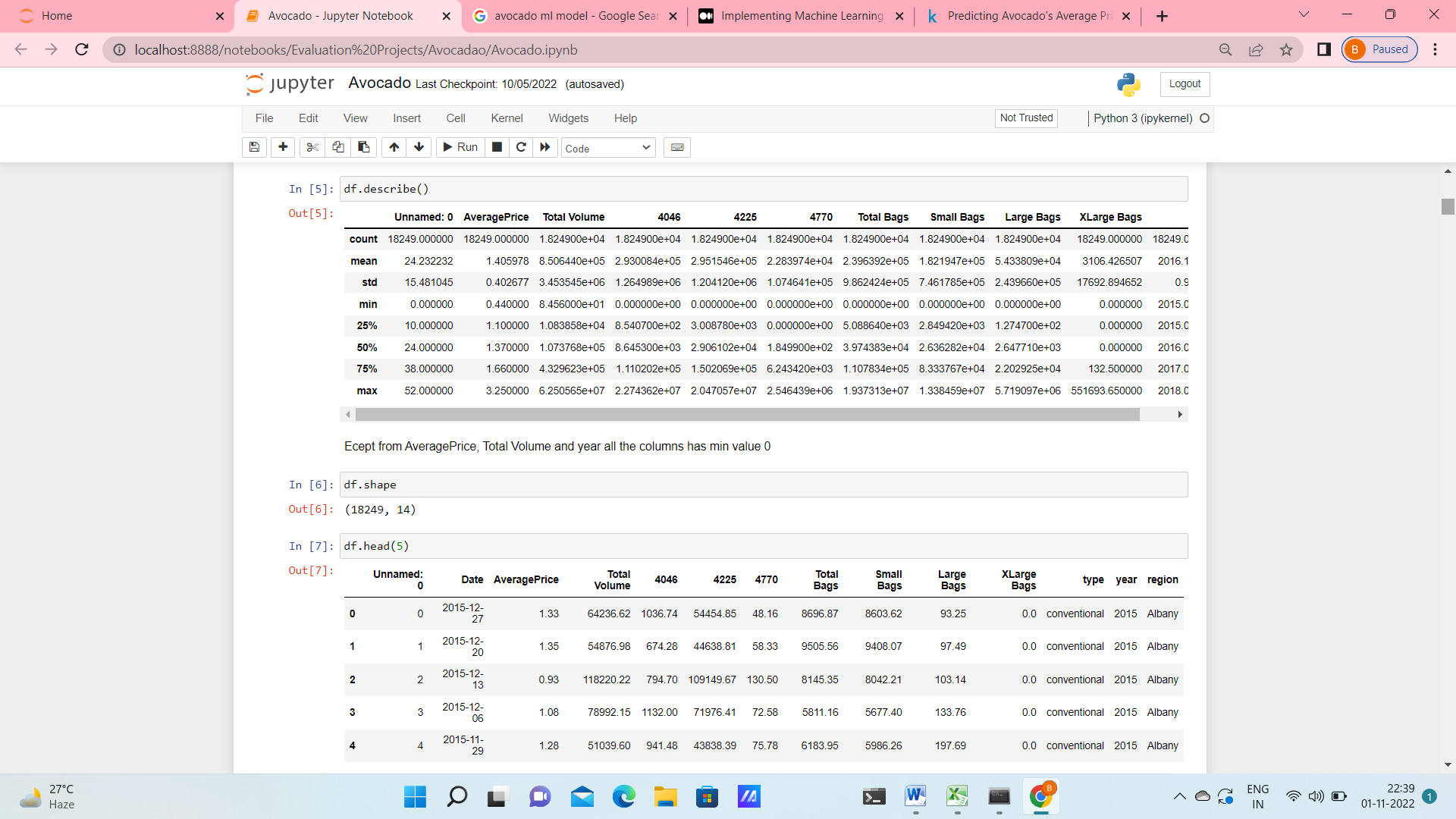
Task: One of Classification and other of Regression

**Data Analysis and Pre-processing :**

At first I shall import basic libraries that are necessary.



Then the data set is loaded into the dataframe “df” & the basic informations about the data are extracted.



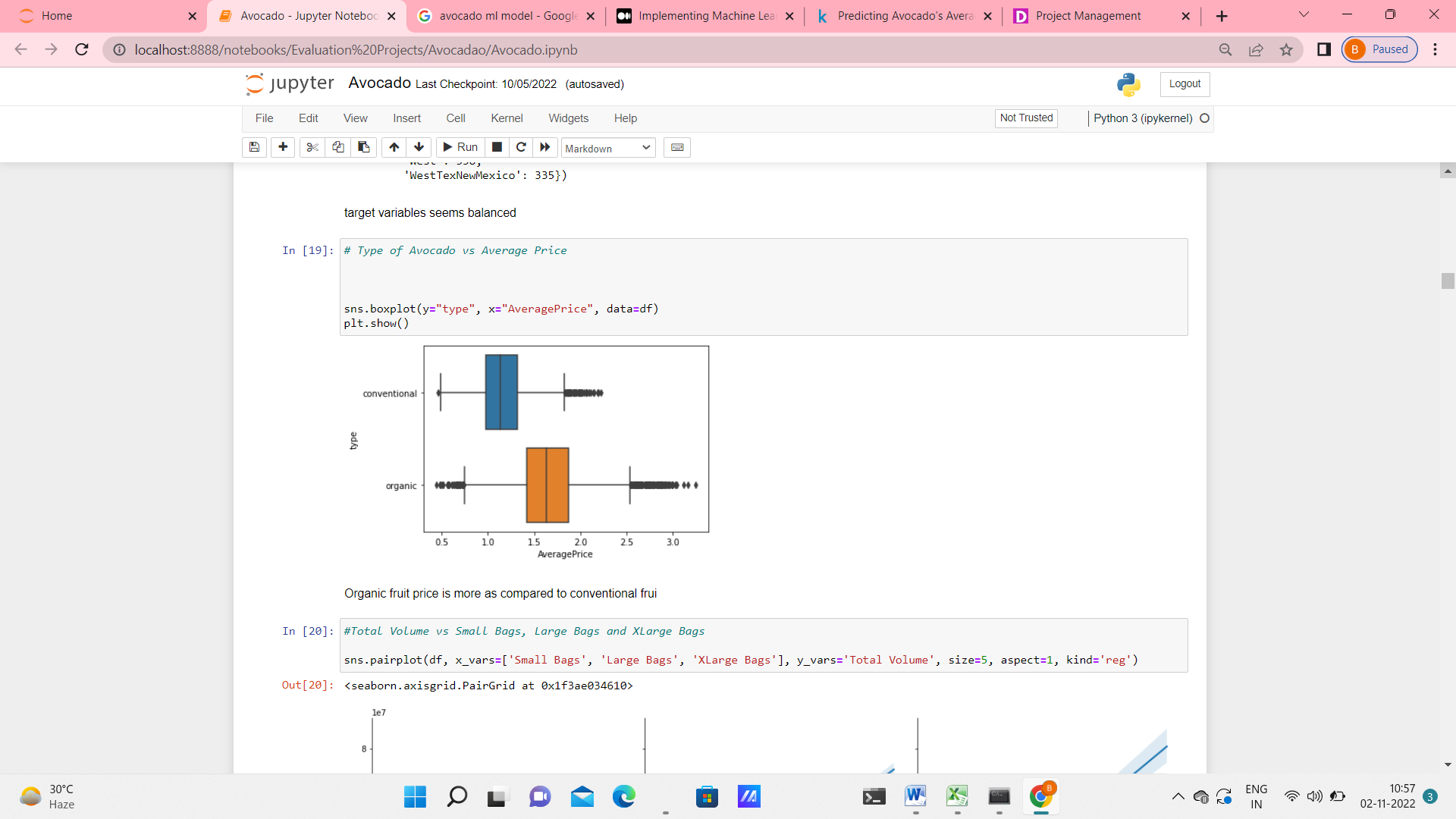
**EDA :**

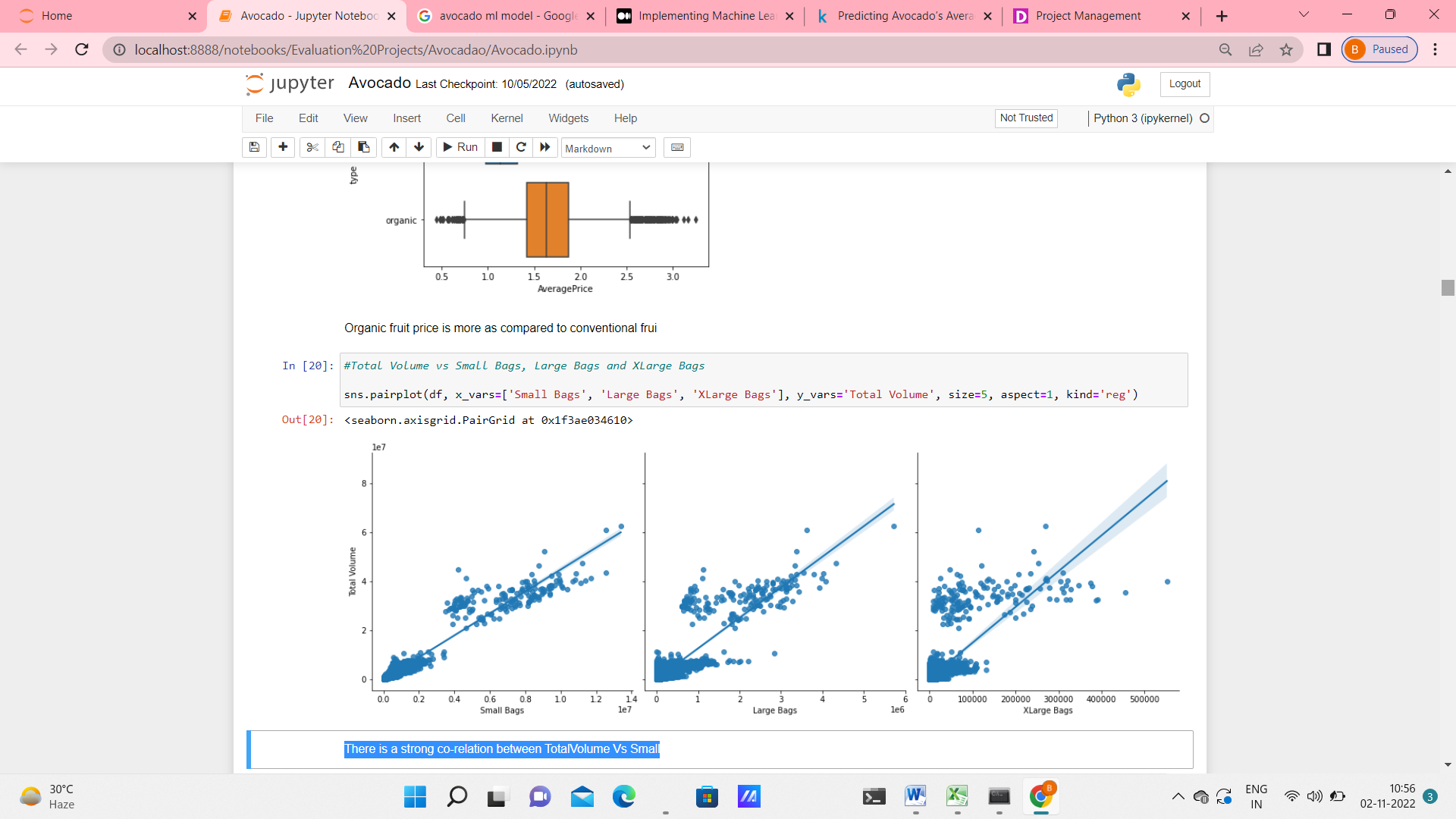
The index column ‘Unnamed: 0’ has no significant values. So I have deleted this column.

I have also found that ‘region’ is a multiclass variable in which all the classes are balanced.

It was found that the dataset has 18249 rows and 14 columns. No null values were present in the dataset. Most of the columns are of integer and float types, only 3 of the columns are of object type.

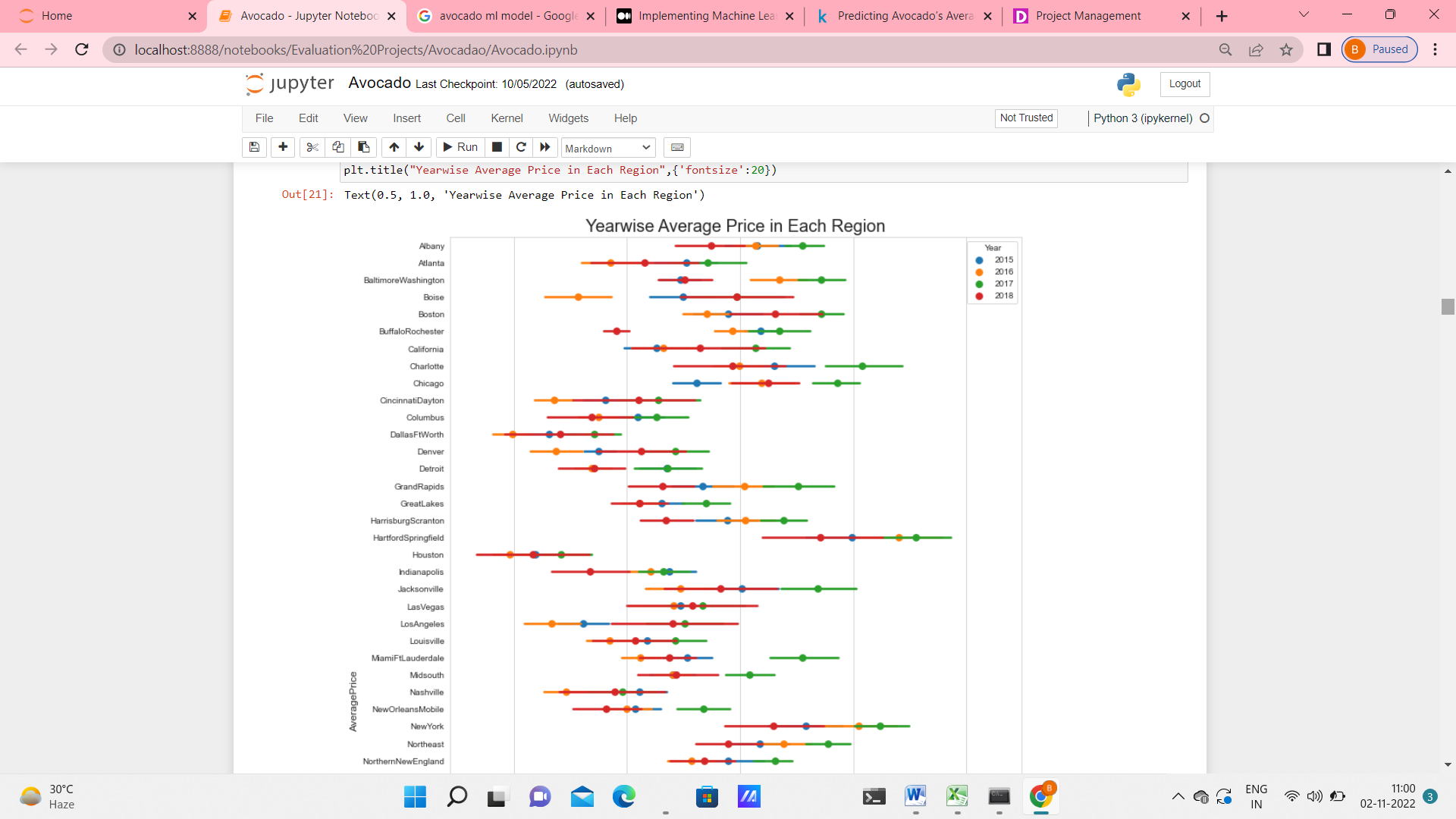
It was seen that the organic fruit price is more compared to conventional fruit price.

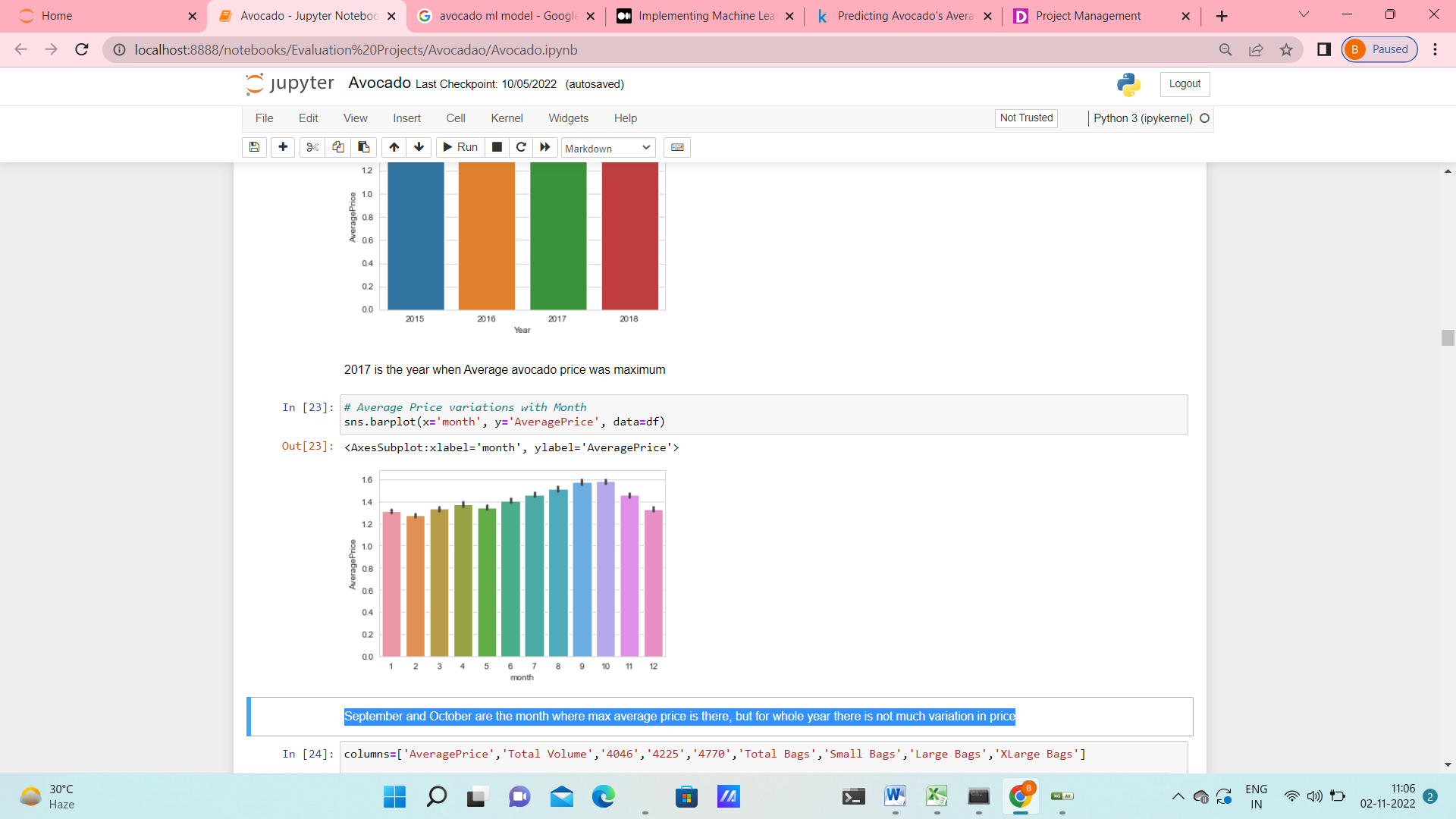




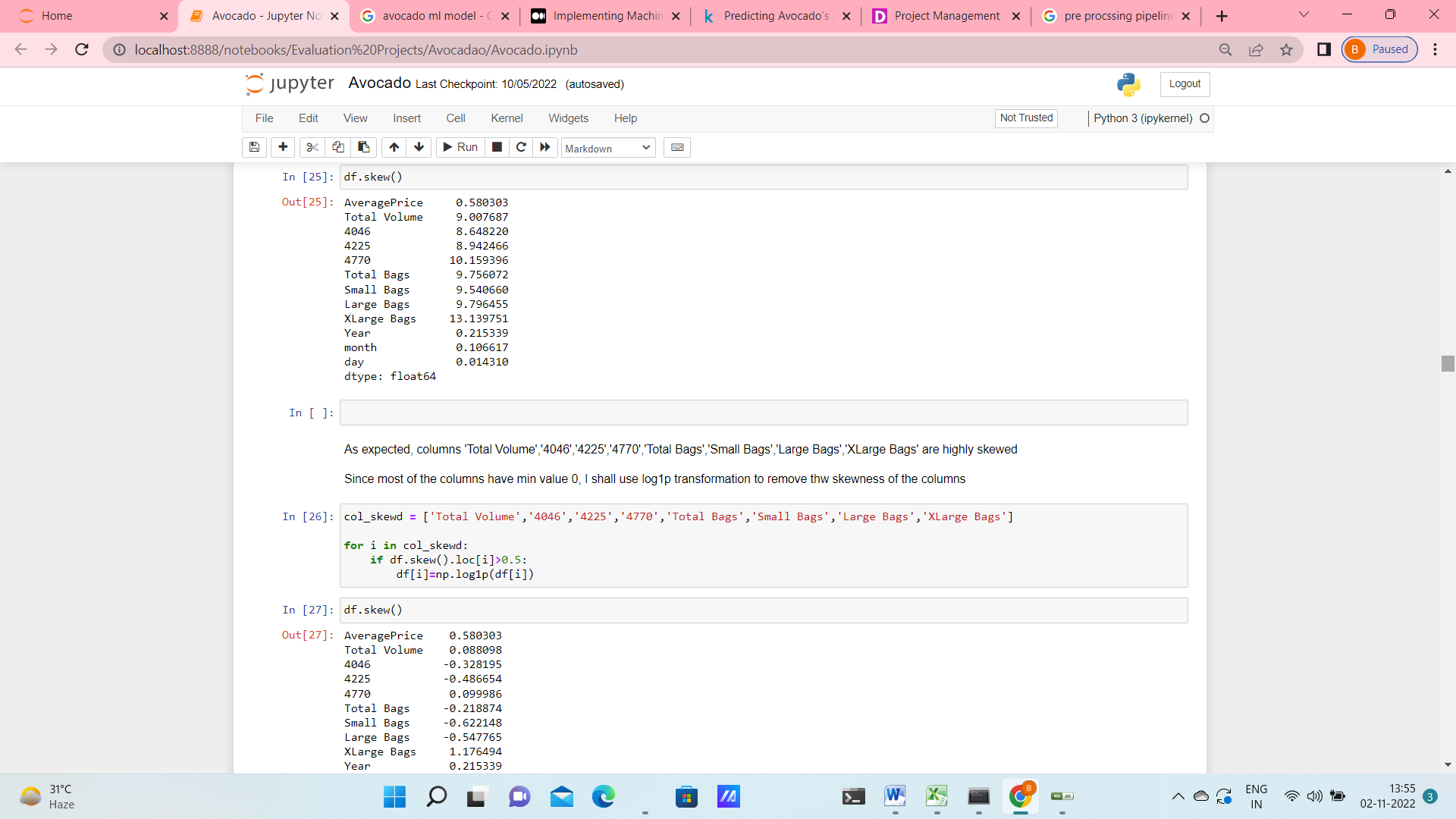
There is a strong co-relation between TotalVolume Vs Small bags.

I have also checked the year wise average price in each region & found that price was highest in 2017, and September and October are the month where max average price is there, but for whole year there is not much variation in price.





The distribution of features has been seen by distribution plots. I have also checked the skewness of the features. It is seen that most of the features are right skewed.

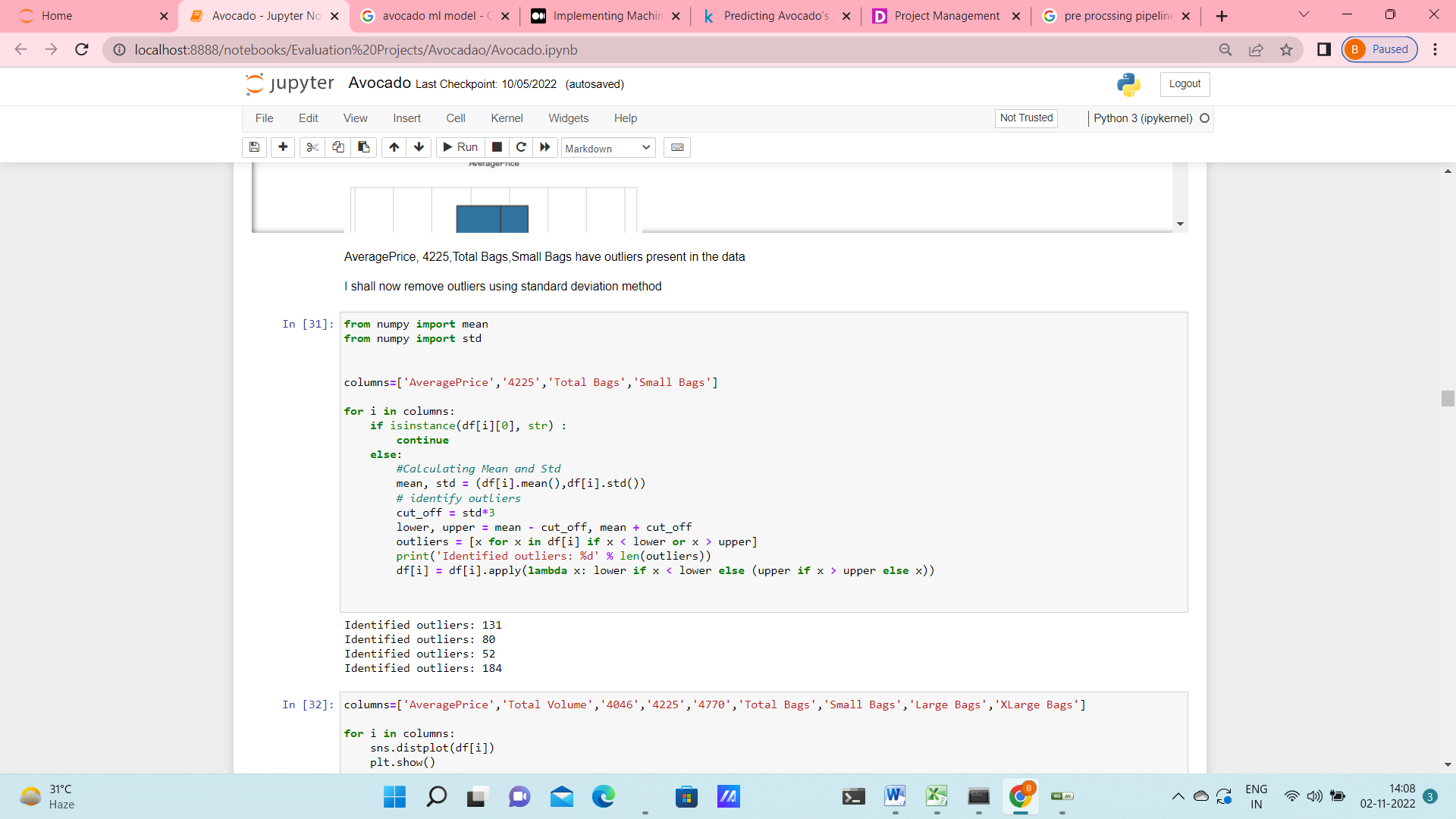


The skewness has been removed by **log1p transformation**.

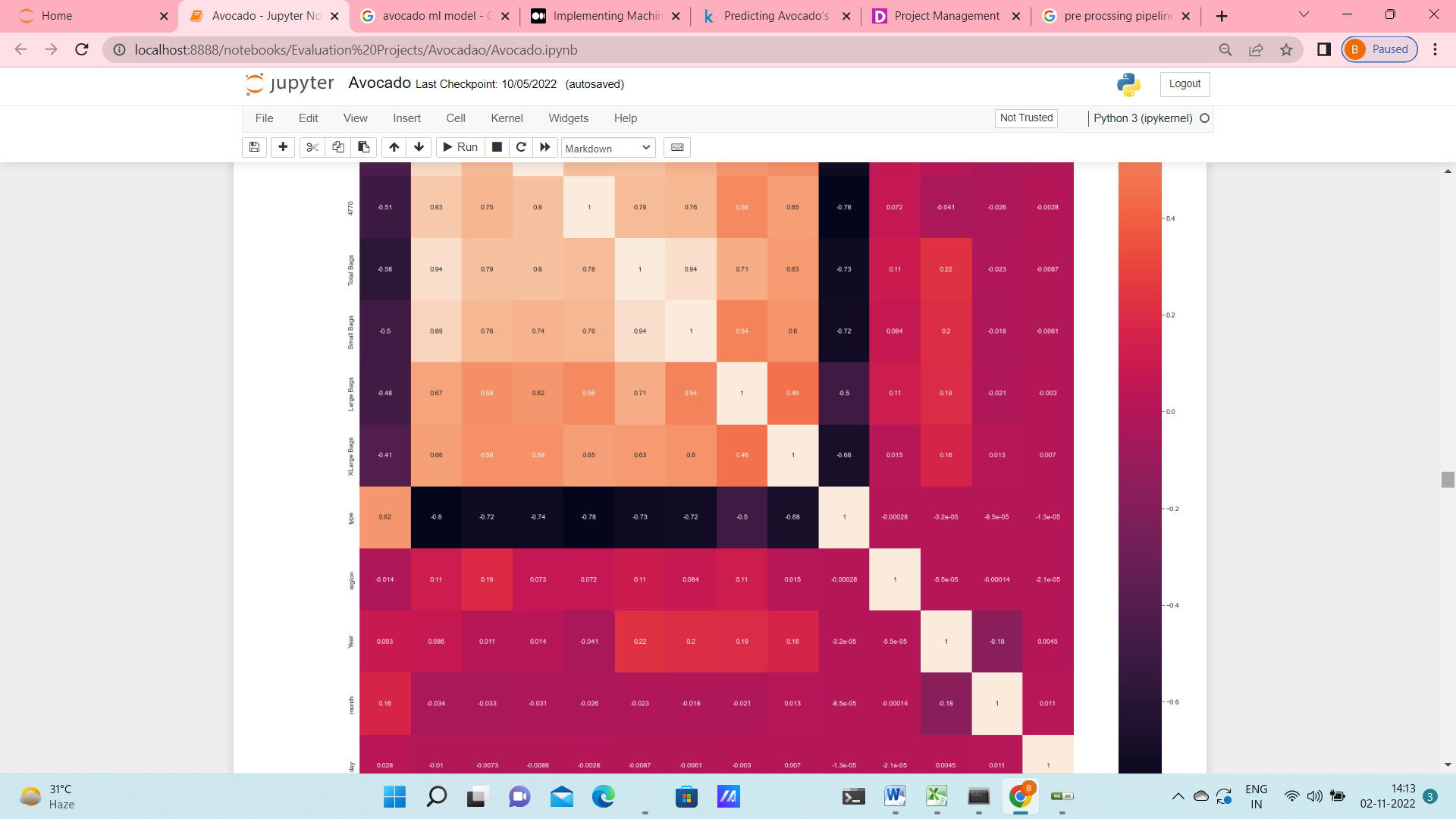
Only the skewness of the column ‘XLarge Bags’ had to be removed by applying **power\_transformation** further.

In the next step I checked for outliers and removed the outliers using 3 **standard deviation** as cut off.

In order to convert the categorical data ('type','region') into numeric I am using **Label Encoder.**



To check correlation between the features I have used pair plot & heat map.



From the heat map it seems like many of the features are correlated.

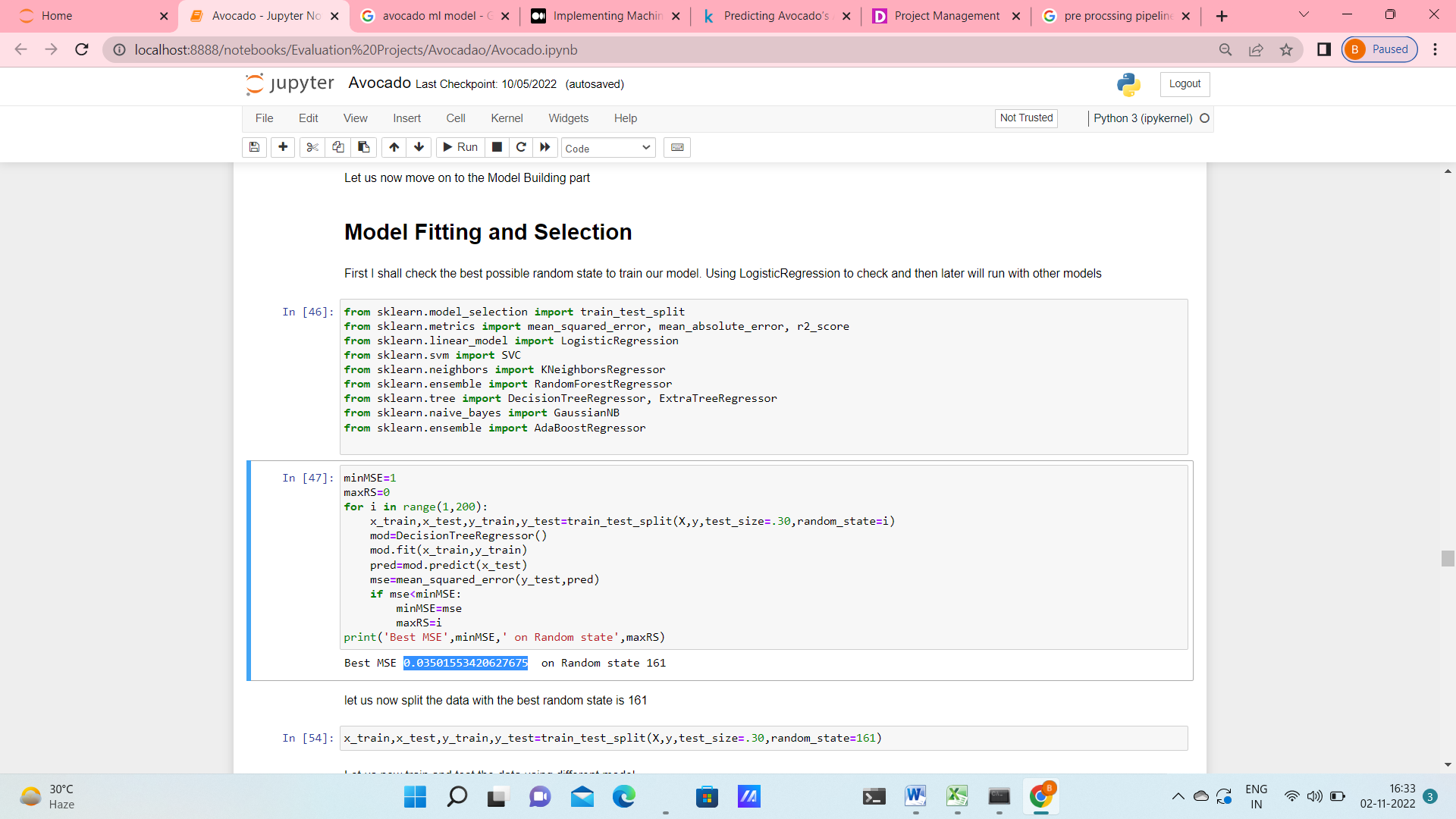
So, to check further I have used **VIF method**, after standardizing the features using **standard scalar** **transformation**. "Total Volume" & "Total Bags" have been found to have high correlation, so I have removed those columns from the data.

Thus our data seems to be clean now. With no skewness & outliers present. Feature reduction have been taken into account. Now we can move on to model building phase. In this phase we have to build one classification ( to predict the region) and another regression (to predict the average price) model.

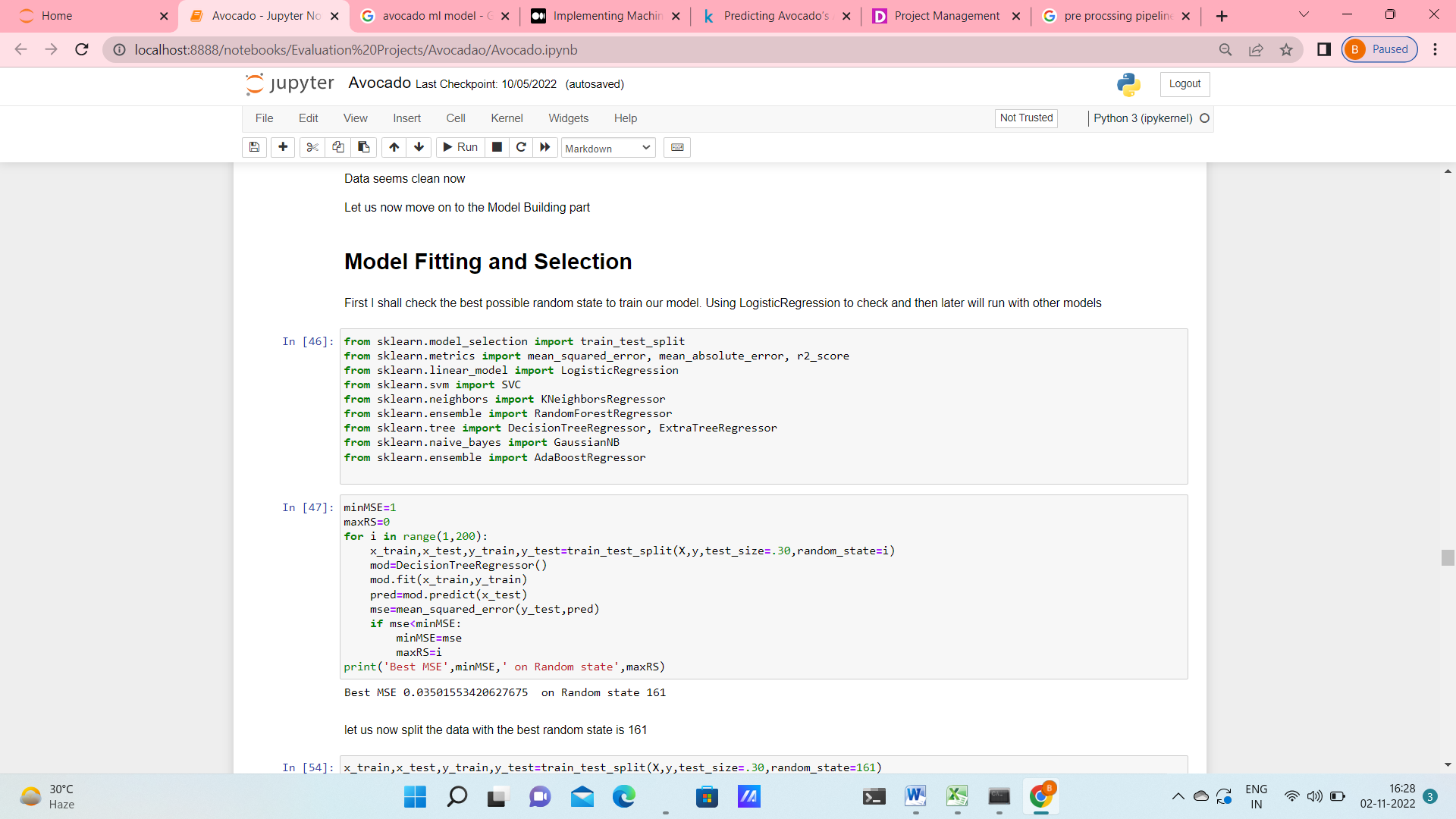
**Building Machine Learning Models:**

**Predicting Average Price using Regression**

At first I have imported all the necessarily libraries, metrices & models.



Keeping ‘Average Price’ as ‘y’ & rest of the features as ‘x’ I first split the data using a for loop over random state values to get the best MSE value & the corresponding value of random state.



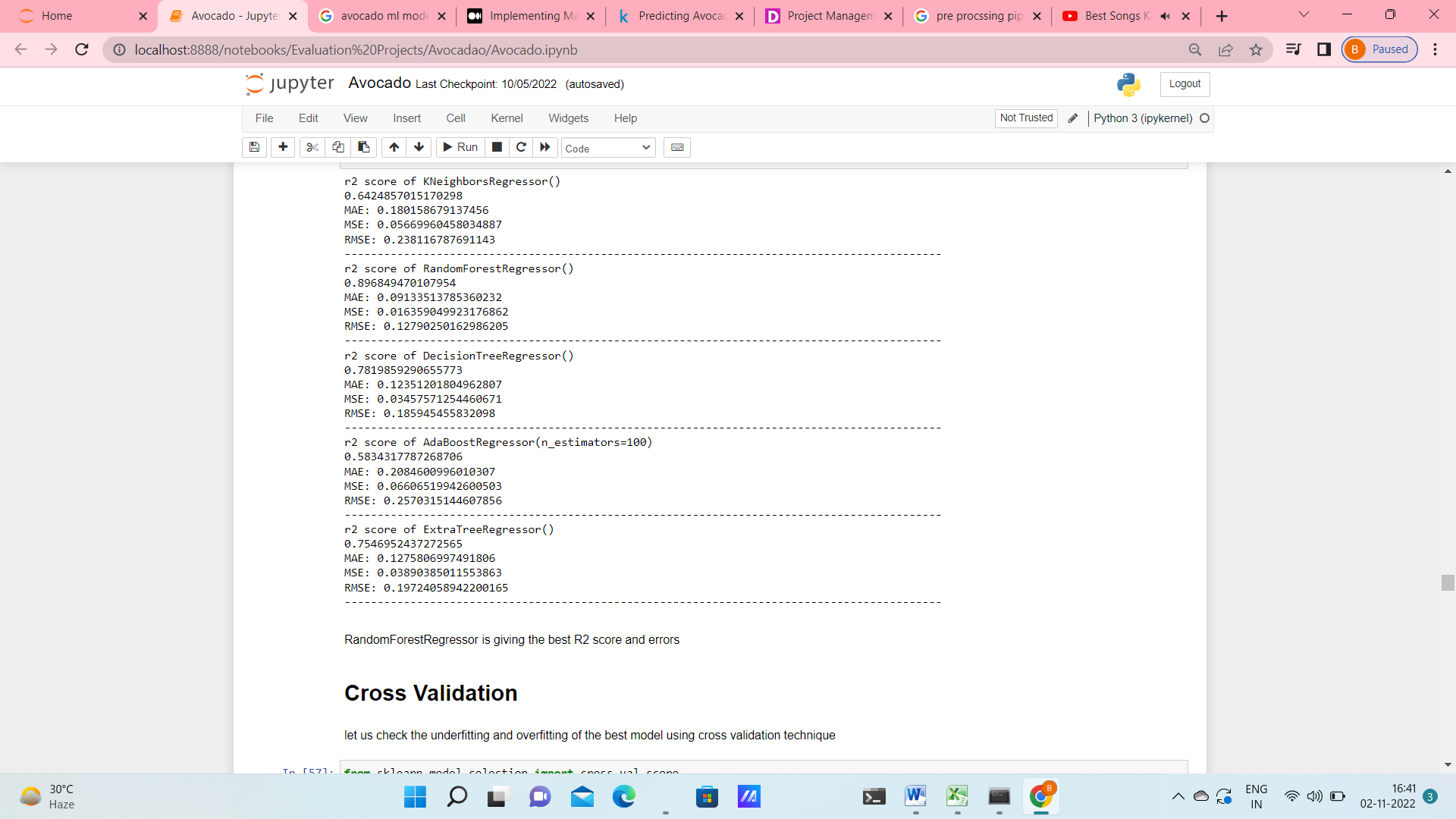
The best value of MSE thus obtained is 0.03501553420627675 and the corresponding random state value 161

With this value of random state I now split the dataset into train & test set. Next I have tried to fit the data with 8 different models e.g.

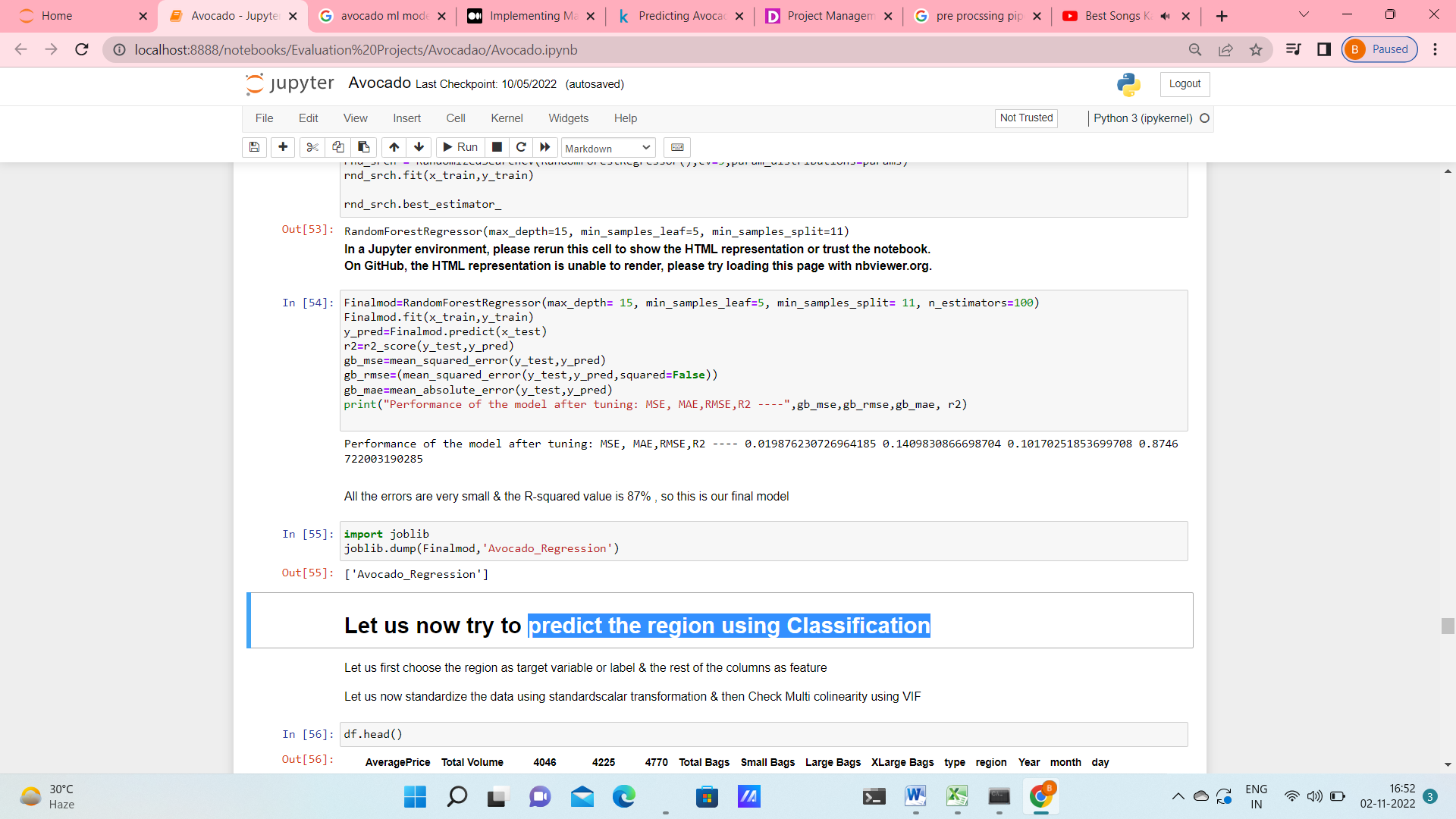
LogisticRegression(), SVC(probability=True), KNeighborsRegressor(n\_neighbors=5),

RandomForestRegressor(n\_estimators=100), DecisionTreeRegressor(), GaussianNB(),

AdaBoostRegressor(n\_estimators=100), ExtraTreeRegressor() & the corresponding metrices are checked (MSE, RMSE, MAE and r2 score) for each of the model. In this step Randomforest is leading with r2 score of 0.896.



Next I tried to Cross validate using negative root mean square value as the metrics & see that ADB and KNN are giving less difference between test & CV score. However the test r2 score of both these models are much less than RFC. So, I still stick with RFC & try to hypertune the model in the next step.

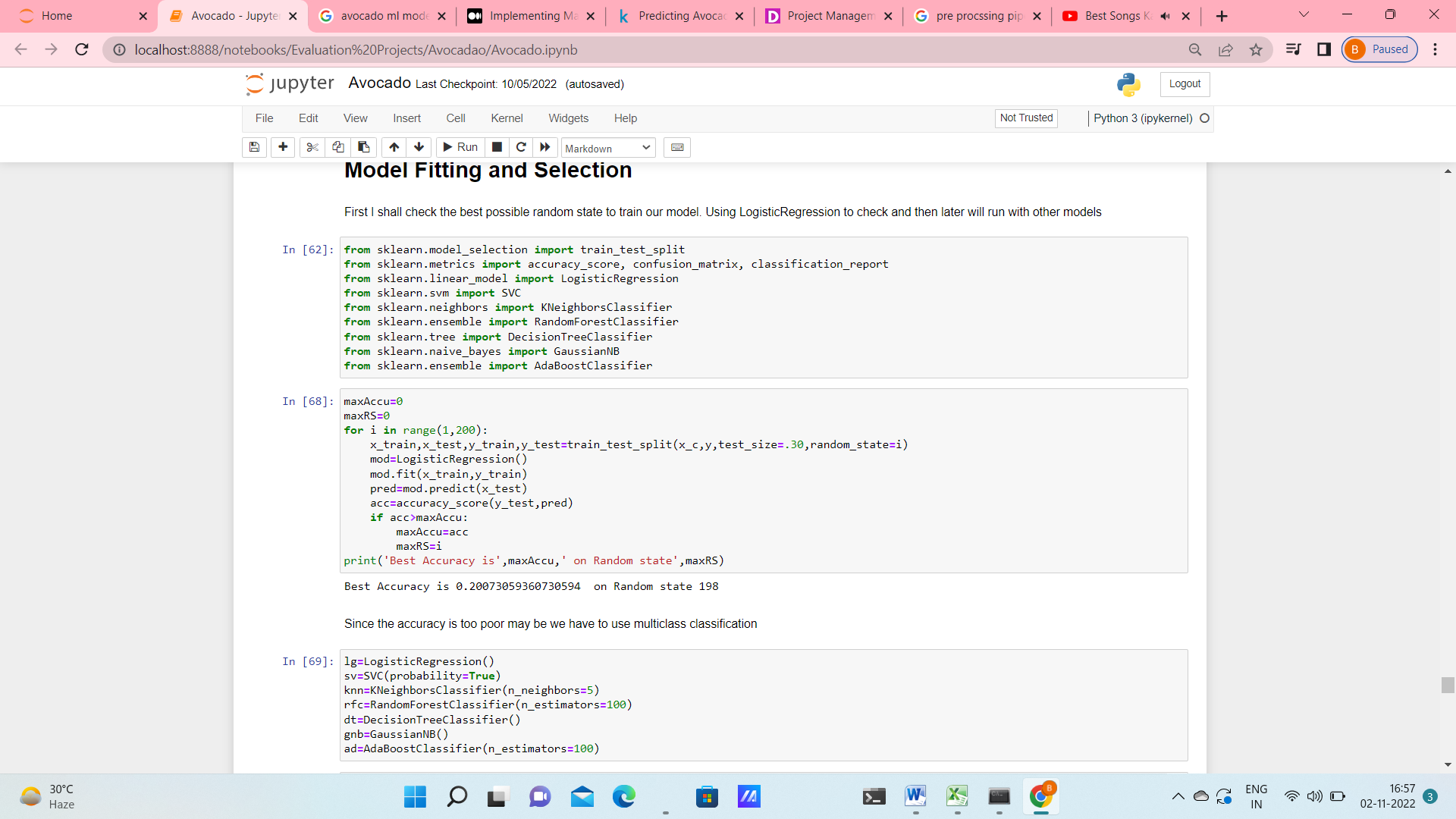


After tuning the parameters with GridSearchCV and RandomizedSearchCV I get the best r2 score of 87% with RFC with MSE 0.02. So this is a good prediction rate.

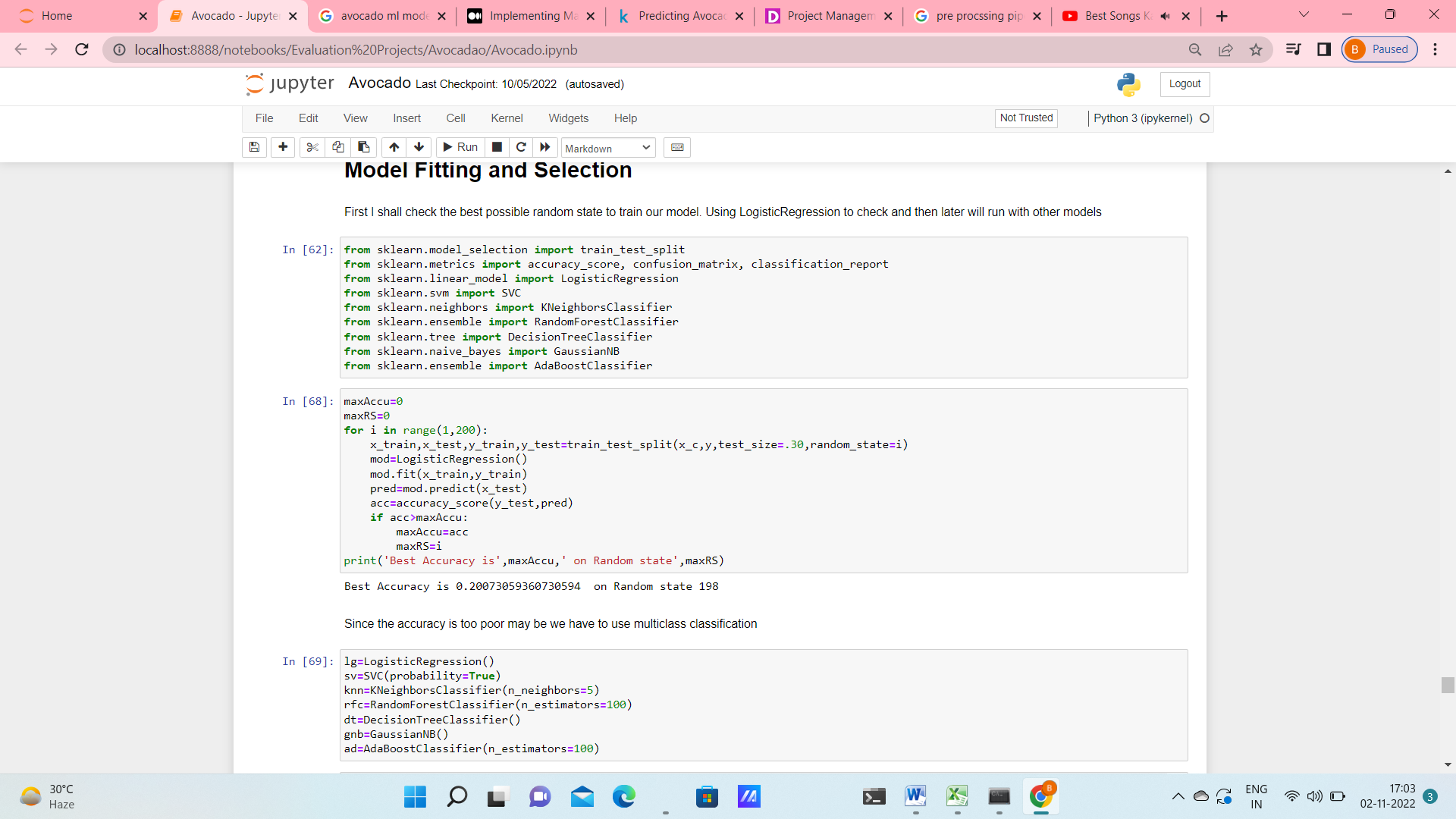
**Predicting the region using Classification:**

For this step I choose ‘region’ as y & rest of the features as ‘x’ & then standardized the features in x using standard scalar transformation. VIF was applied to remove correlated features

Again at first I have imported all the necessarily libraries, metrics & models.



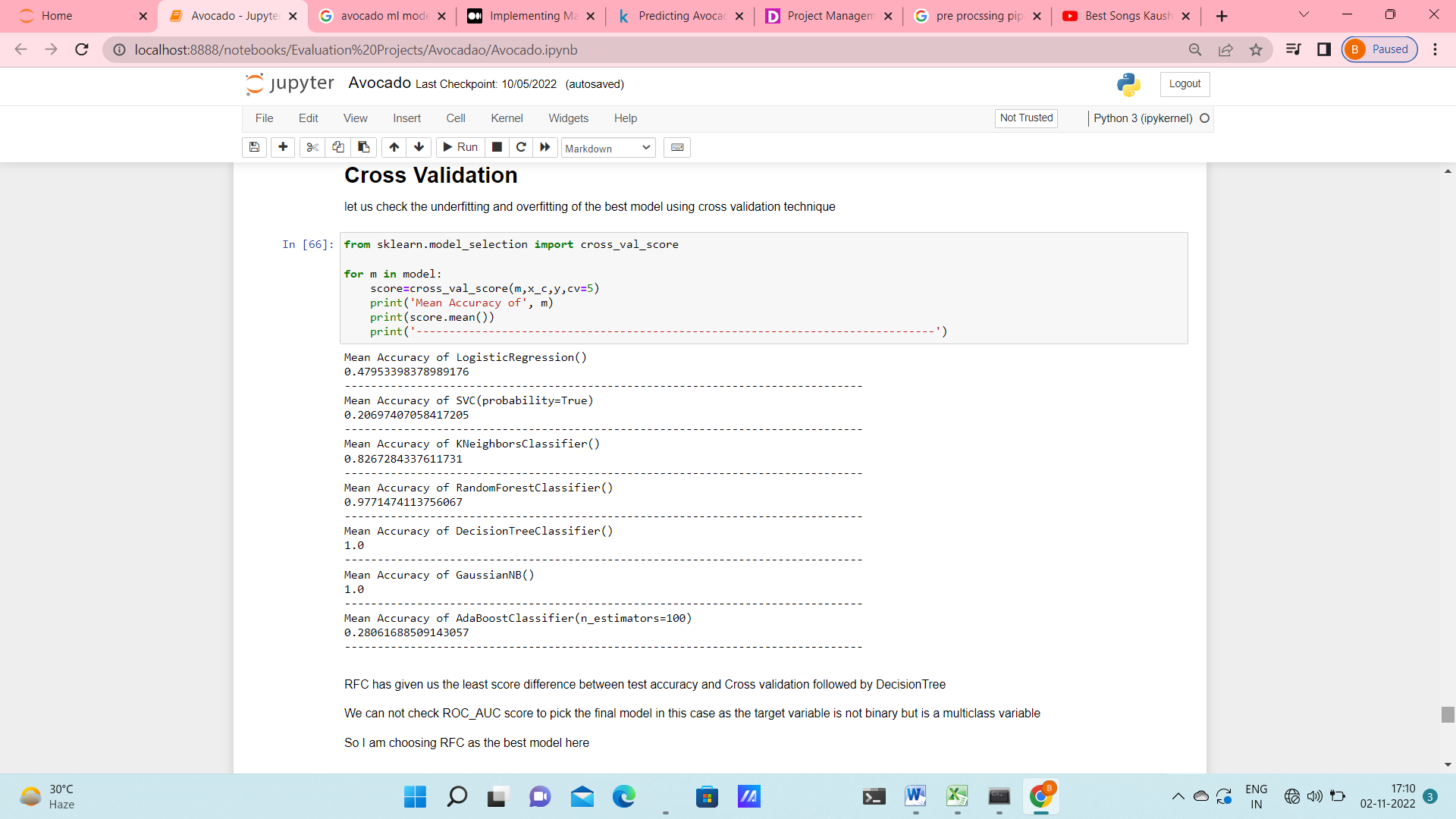
Data was splitted using a for loop over random state values to get the best accuracy value & the corresponding value of random state.



With thus obtained random state value I split the dataset into train & test sets.

I have checked the test accuracy & cv score for 7 different classification models & have seen that most are performing poorly except from RandomForest and DecisionTree.

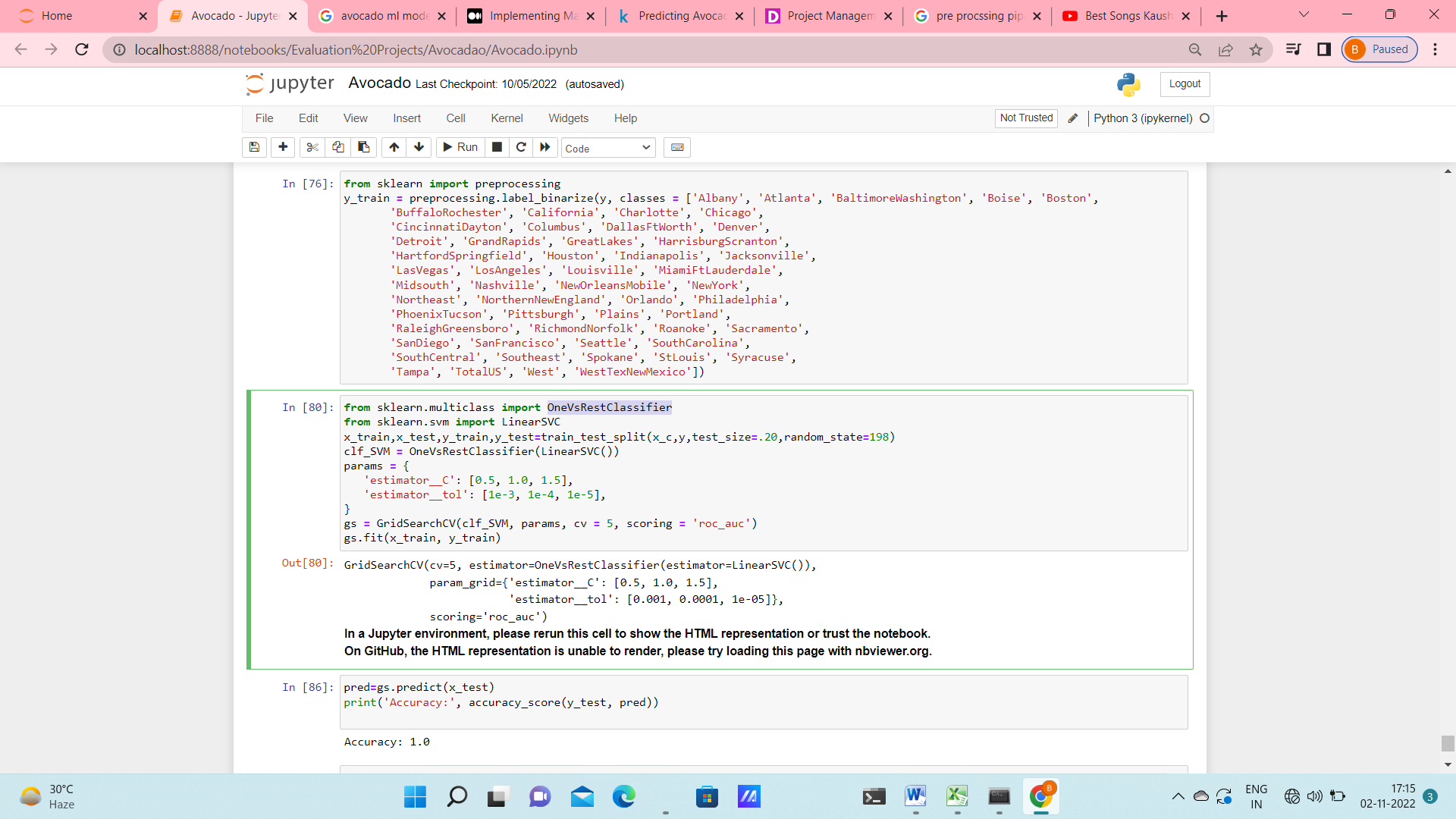
RFC was the best model as it is giving the max accuracy (91%).



Again after checking with Cross Validation It was seen that RFC was giving the least difference of score between test & CV score.

ROC score could not be checked in this model as ‘region’ is a multi class variable. So finally I used RandomizedSearchCV to hypertune the parameters with RFC as the estimator. After tuning the accuracy was improved to be 100%.

Yet again, I wanted to check with a multiclass classifier, the result of prediction. For that I have followed the steps below.



At first I have binarized the label ‘region’ by using preprocessing.label\_binarize. OneVsRestClassifier(LinearSVC()) classifier was used to classify the label. I have also used GridSearhCV to tune the parameters. In this case also, with the best set of parameters I got an accuracy score of 100% on the test set.

Thus the model performance is considerably good for both train as well test data.

**Concluding Remarks:**

1. The dataset initially had a skewness, outliers, categorical variables and correlation among variables.
2. This problems were treated adequately to have a cleaner dataset.
3. Both classification ( to predict the region) and regression ( to predict the average price) models were built.
4. Both models performs good on train & test sets.
5. The regression model gives us r2 score of 87%
6. The classification model gives better performance with accuracy score of 100%