PROJECT TITLE

AVACADO PROJECT

Submitted by:

MOHAMMED ZAHED N

**ACKNOWLEDGMENT**

Below are some of the sources from which code snippets have been helpful during the project completion

**Refernces:**

1. <https://www.scikitlearn.org>

2. <https://www.askpython.com>

3. <https://www.stackoverflow.com>

4. <https://www.geeksforgeeks.org>

**INTRODUCTION**

* Business Problem Framing/Problem Definition

Avocado is a fruit consumed by people heavily in the United States, data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV using the same data along with machine learning we have to create a model that best predicts the average price of the Avocado.

.

* Conceptual Background of the Domain Problem

 This 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. green skins) are not included in this table.

Some relevant columns in the dataset:

* Date**- The date of the observation**
* Average Price**- 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**

**Data Analysis**

* Mathematical/ Analytical Modelling of the Problem

Many Statistical models were used also mathematical models necessary, Some of the used models are listed below.

Statistical Model:

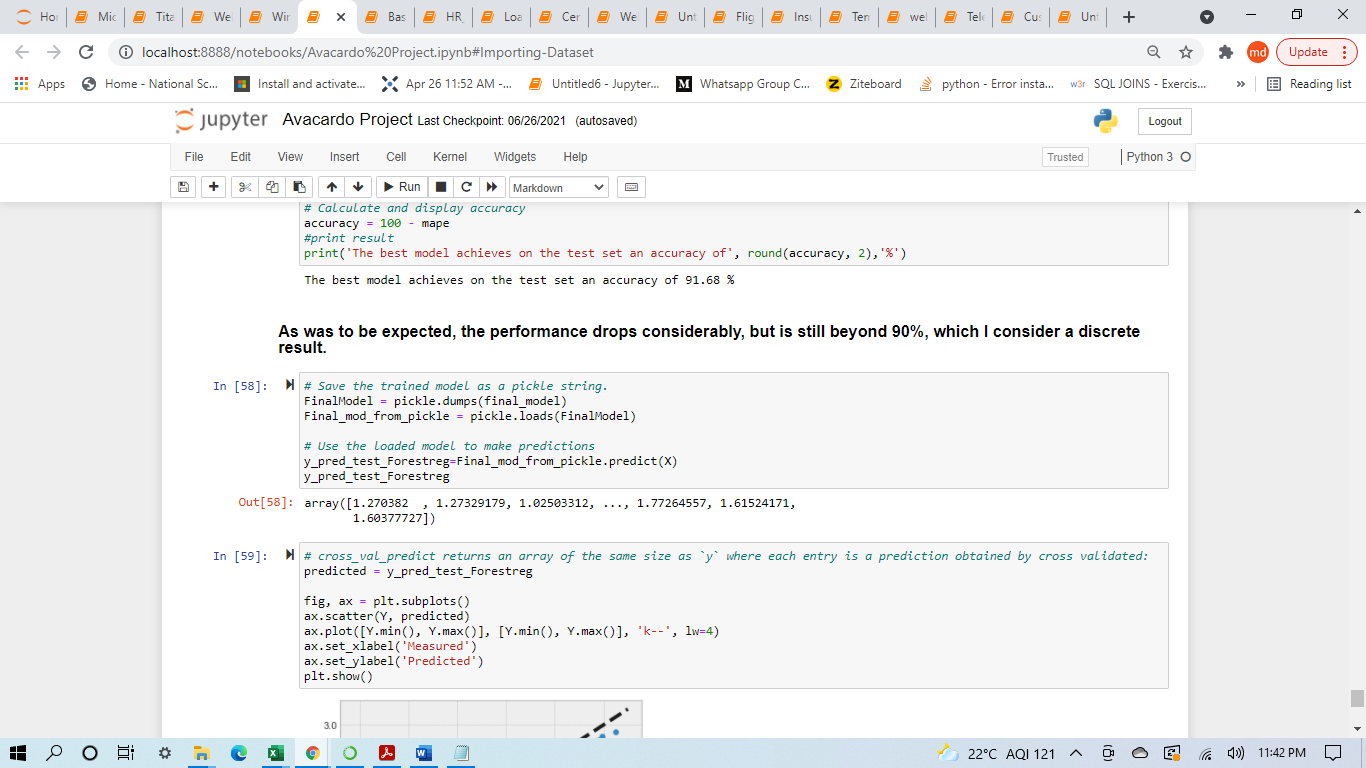
1. Pandas Profiling
2. Label Encoders
3. Co-relation metrics
4. Linear Regressor
5. Random Forest Regressor
6. Decision Tree Regressor

* Data Sources and their formats

Most of the Code snippets and their formats are taken from below site

<https://www.geeksforgeeks.org>

Code on how to save a model using pickle library



* Data Preprocessing Done

1. Renaming PLU Codes(4046, 4225,4770,) into meaning full column names for better Understanding

2. Checking Missing Values

3. Data Description

4. Checking Data Imbalance For Avacado Type

5. Dropping Unnamed column as it does not contain any much significant data in it

6. Converting Categorical Columns Such as Type Column

7. Applying Label Encoder on Region Column

* Data Inputs- Logic- Output Relationships

Most of the data in the dataset was already Numerical except for few columns after data cleaning and processing sent to model as independent features and run Test & Train along with dependent feature as output variable.

**Independent Features->Dependent Feature->Model(Prediction)**

State the set of assumptions (if any) related to the problem under consideration

Renaming these PLU codes

'4046':'HASS\_Avacado\_small sold',

'4225':'HASS\_Avacado\_Large sold',

'4770':'HASS\_Avacado\_XLarge sold'

* Hardware and Software Requirements and Tools Used

List of hardware and software requirements along with the tools, libraries and packages used.

**Softwares:**

1. Ms Word: Ms Word for documentation purpose

2. Ms Excel: To view Dataset and to perform basic subtraction on columns for comparion

3. Jupiter Notebook(Anaconda):To run python code for building suitable model for prediction.

**Hardware:**

OS: Windows 10

RAM: 4GB

HDD: 1TB

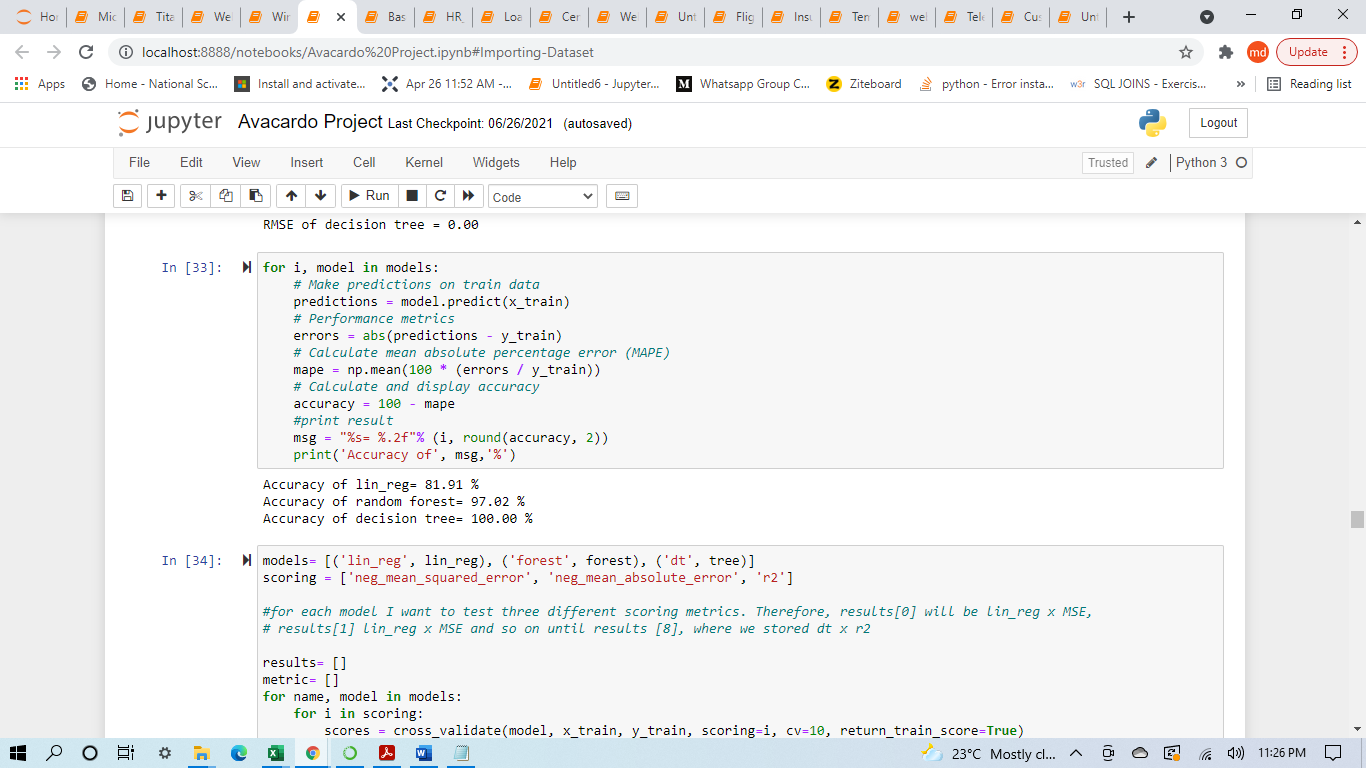
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Loading the data into Jupiter Notebook
* Checking the basic details (Null Value, Datatype, Shape etc.)
* Identifying the target and independent features and perform EDA (Univariate, Bivariate and Multivariate analysis) using Data Visualization and Statistical approach accordingly.
* Performing data cleaning, outliers handling, missing value imputation
* Building model
* Evaluating the built model for accuracy score
* Performing hyperparameter tunning to enhance the performance
* Evaluating the model again
* Make prediction
* Testing of Identified Approaches (Algorithms)

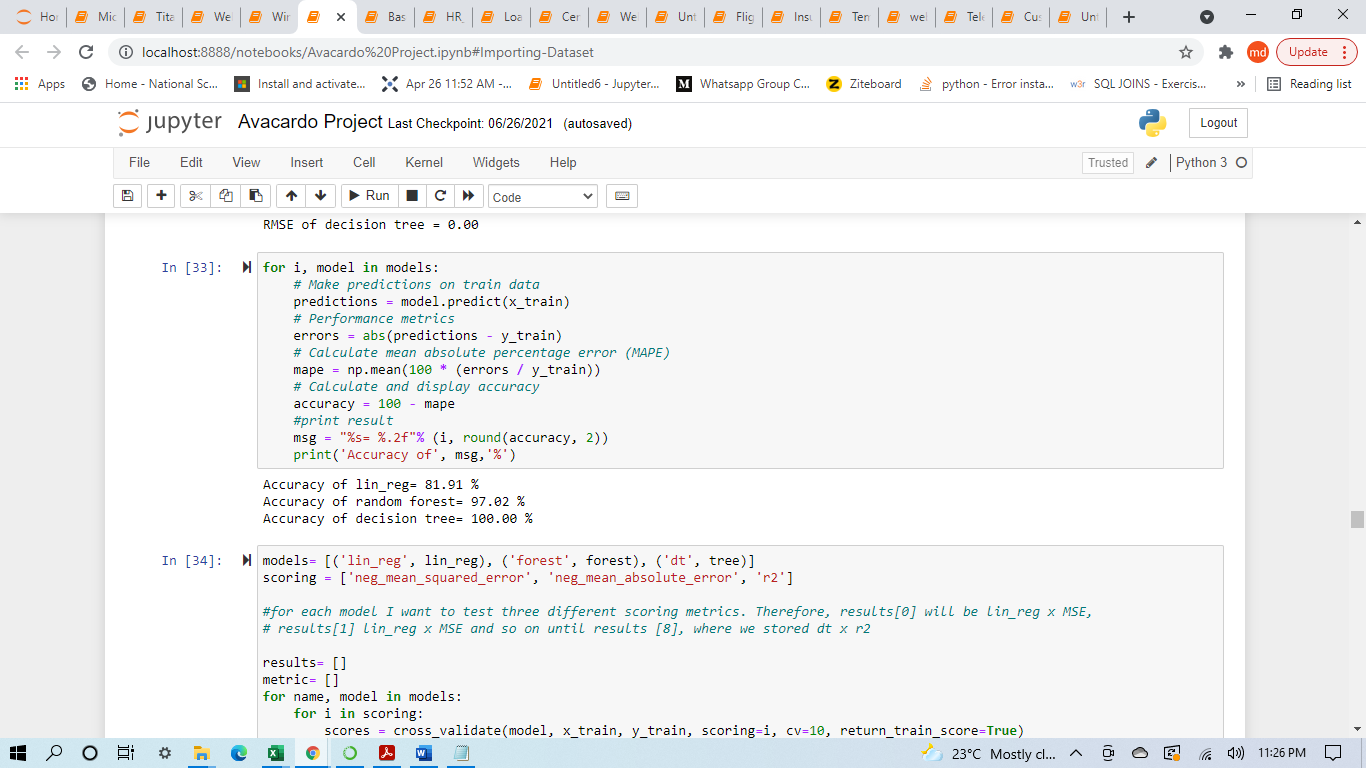
1. Decision Tree Regressor
2. Random Forest Regressor
3. Support Vector Regressor

These were the algorithms that are used for training and testing purpose.

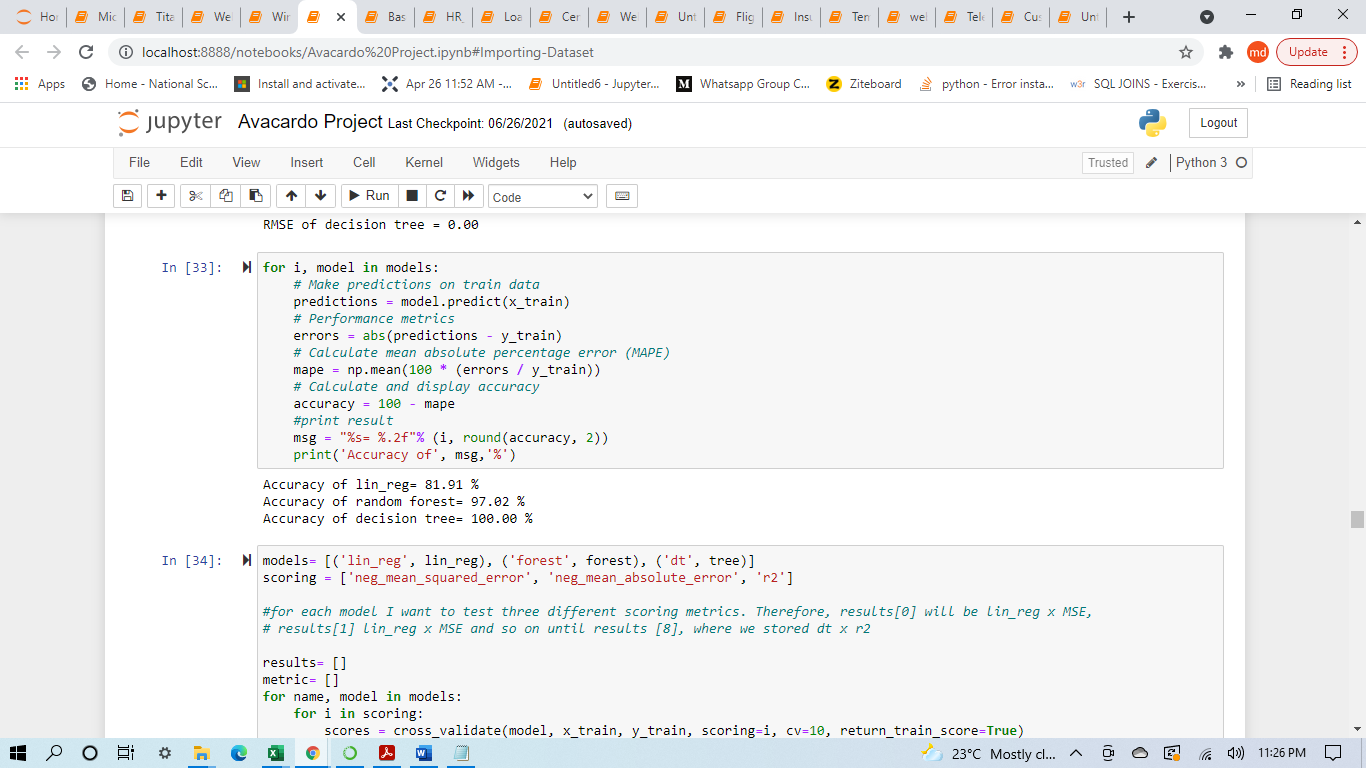
* Run and evaluate selected models
* Support Vector Regressor: The Accuracy found by using this Model is least which is 81.91%.



* Decision Tree Regressor: The Accuracy found by using this Model is 100%.



* Random Forest Regressor: The Accuracy found by using this Model is 97.02%.



* Key Metrics for success in solving problem under consideration

**Hyper Parameter Tuning** was used to maximise the output of the model but there was not much significant change observed compared to normal accuracy achieved from the selected Model.

In conclusion, the metrics for our best models on the training set after our hyperparameter fine-tuning are the following:

**Grid-search:**

**RMSE: 0.0**

**accuracy (MAPE): 100.0 %**

**Randomized Search:**

**RMSE: 0.19**

**accuracy: 93.06 %**

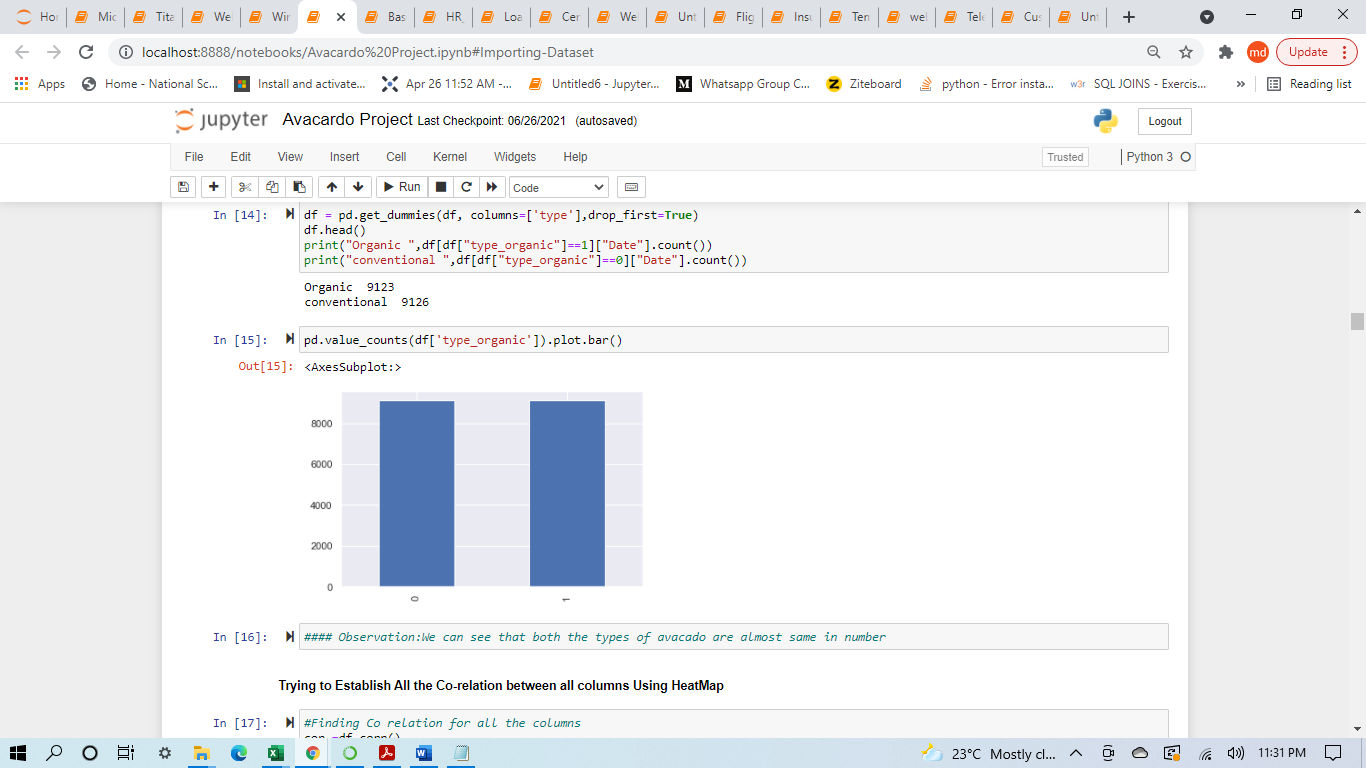
This of course does not mean that the randomized search is inherently inferior to grid-search.

Its just that in this notebook I used the information provided by the former to refine the latter.

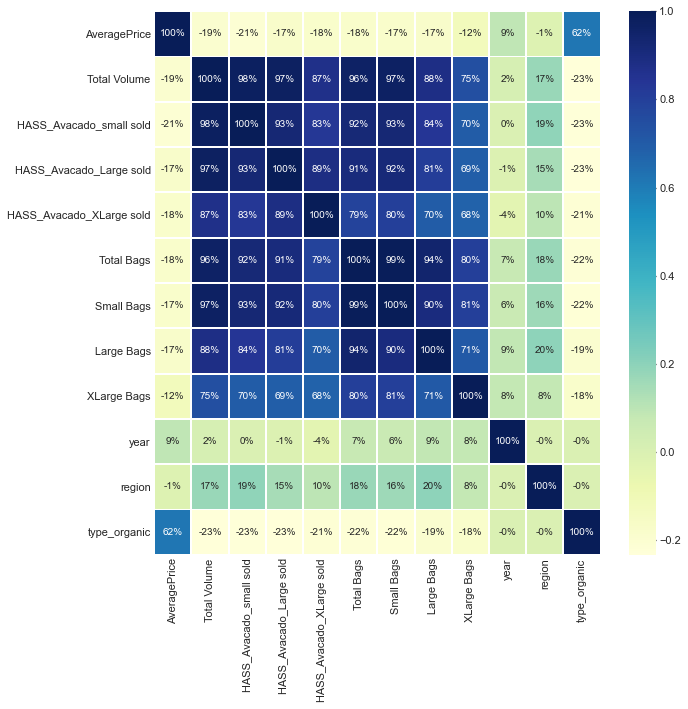
In any case, this time I will consider the model obtained by the Grid search as the best one.

* Visualizations

1. This is Count plot which tells that the dataset is almost balanced.



1. This is Heat Map for plotting co relation between all features



**Observation:** We Can see That Average Price has stronger Co relation with Year and Type of Avacado

There is a high correlation between pairs:

HASS\_Avacado\_small & Total Volume (0.98)

HASS\_Avacado\_Large & Total Volume (0.97)

HASS\_Avacado\_XLarge & Total Volume (0.87)

Total Bags & Total Volume (0.96)

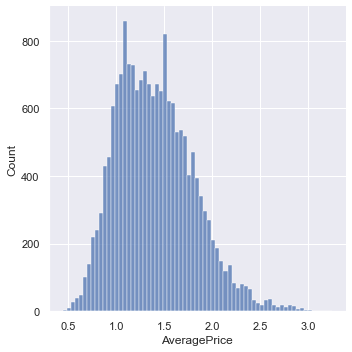
Small Bags & Total Bags (0.99) etc

HASS\_Avacado\_small avocados are the most preferred/sold type in the US and customers tend to buy those avocados as bulk, not bag.

Retailers want to increase the sales of bagged avocados instead of bulks. They think this is more advantageous for them.

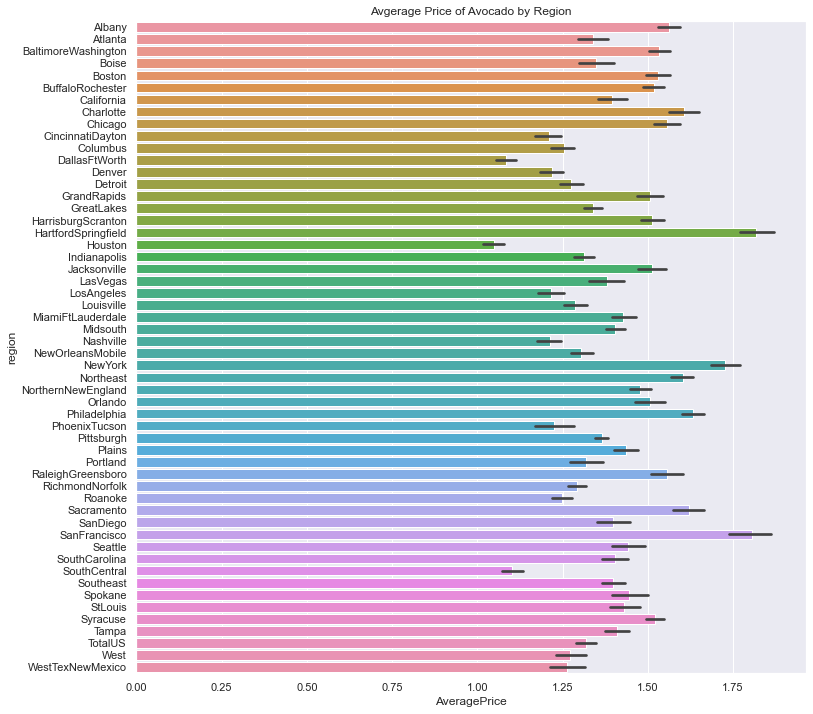
Total Bags variable has a very strong correlation with Total Volume (Total Sales) and Small Bags, so we can conclude that most of the bagged sales comes from the small bags.

1. displot: Starting With Distribution of Average price.



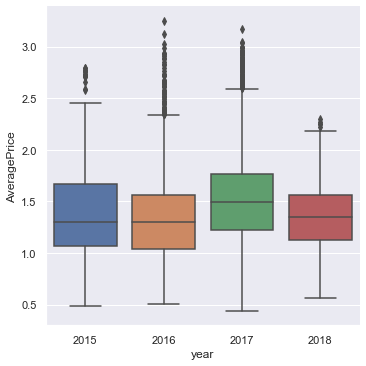
**Observation:** Most of the Average Price Lies Range between **0.8-1.6**

1. Barplot: Price Variation of Average Price from Various Regions shown using Bar Graph



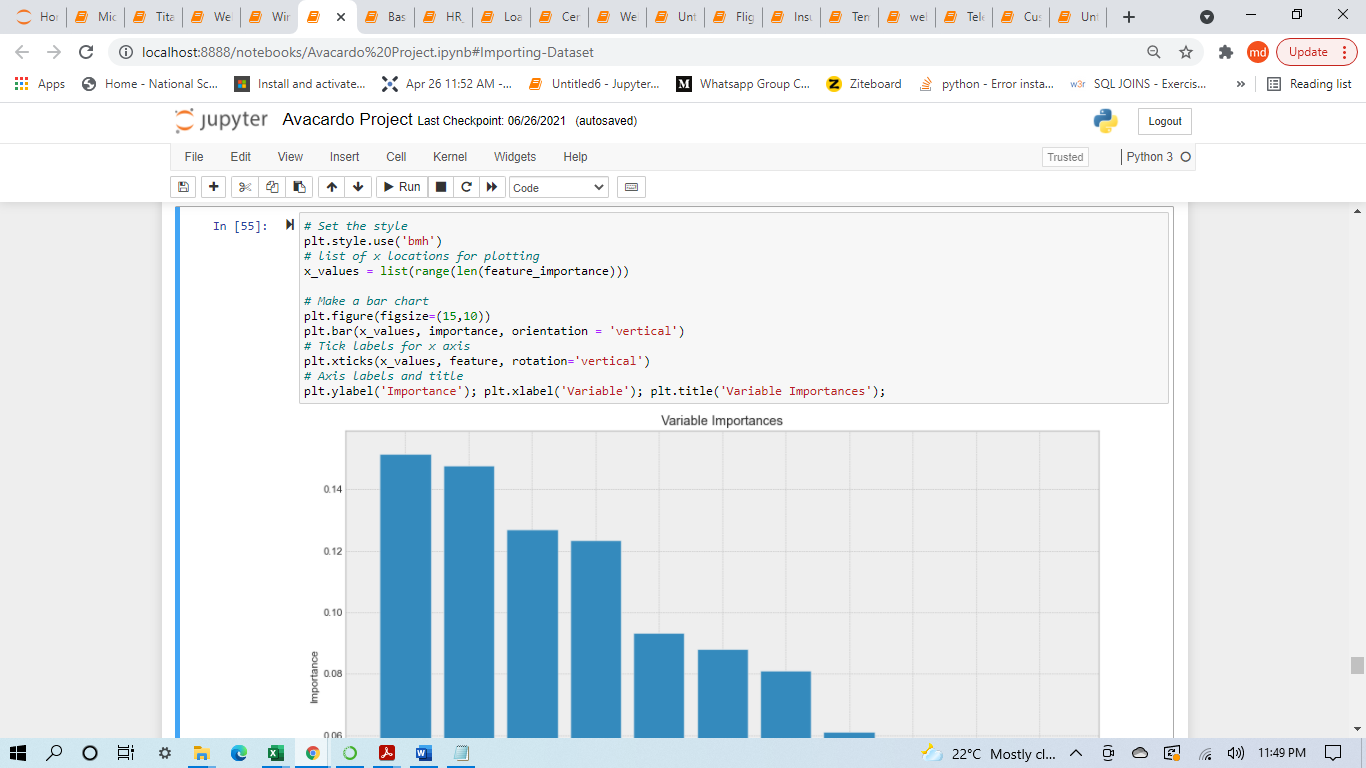
**Observation:** With HartfordSpringfield being highest and Houston being lowest price to get cheap Avocado, AveragePrice varies with region. Price of avocado varies from one region to another

1. Boxplot: Boxplot of Average Price



**Observation:** Average Price is High in the year 2017 compared to the rest of the years.

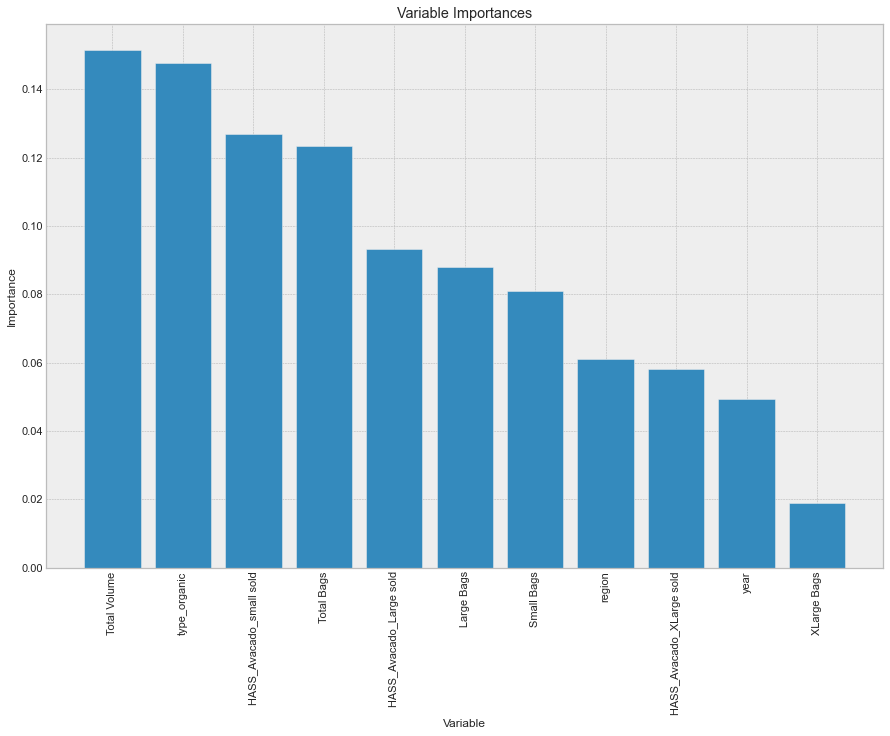
* Feature Importance



**Observation:** A future aim may be to cut the less relevant features (lets say we can drop large bags in terms of importance),

estimate a new model and compare it with the old ones.

I reckon it would lose predictive power, but on the other hand it would improve in terms of training speed.

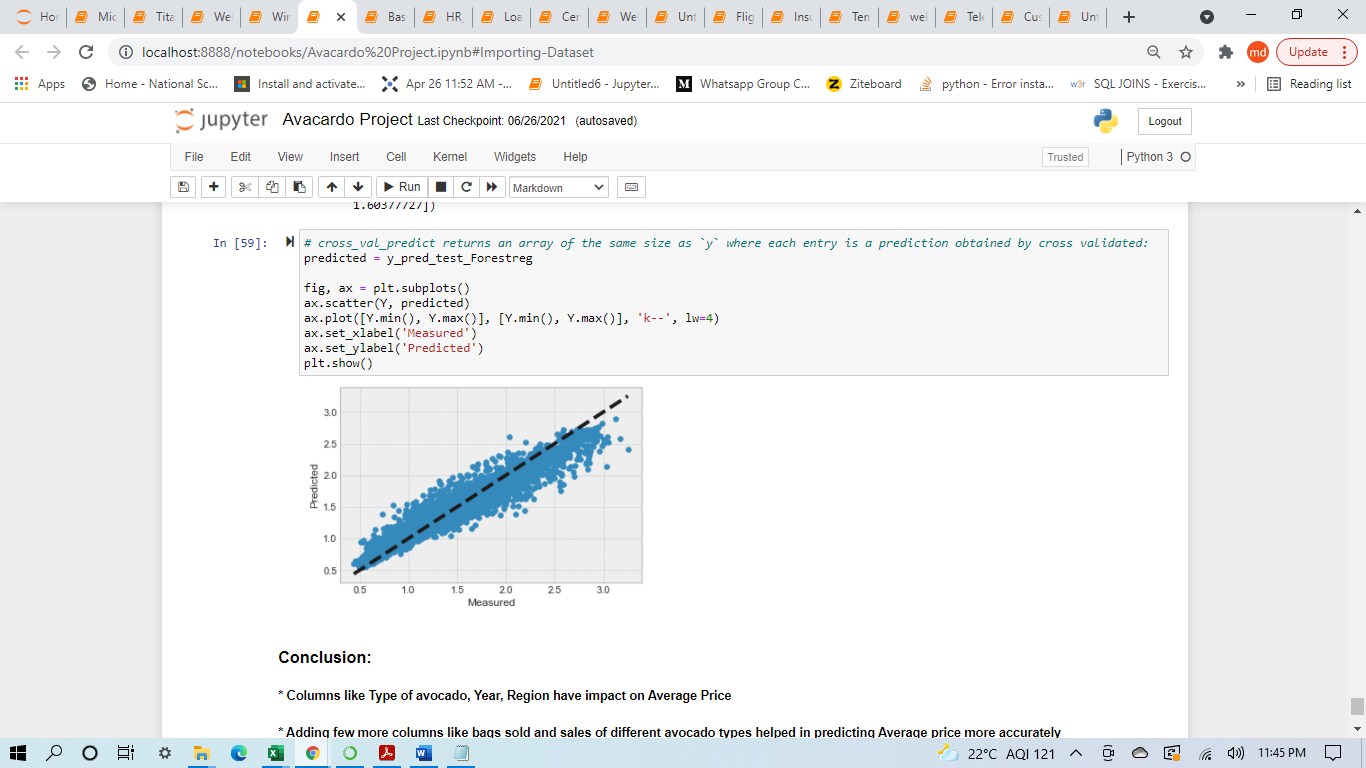


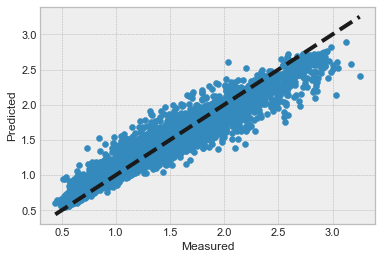
**In conclusion, these are my final considerations on the model:**

1. The best feature to reliably predict the AveragePrice of a Avacado is Type of Avacado. After Type\_of\_Avacado that there is a group of 6 features including: HASS\_Avacado\_small,Total Volume,Total bags large bags and HASS\_Avacado\_large.

2. The features such as region and year are respectively on 8th and 10th in this ranking. XLarge\_Bags is the less relevant feature in this cluster.

* Predicted vs Actual





**CONCLUSION**

* Key Findings and Conclusions of the Study

1. Columns like Type of avocado, Year, Region have impact on Average Price

2. Adding few more columns like bags sold and sales of different avocado types helped in predicting Average price more accurately

3. The Actual vs Predicted plot clearly indicates the predicted values are almost linear hence performance of model is considerably Good

4. Random Forest Regressor model predicts the average price more accurately than linear regression model.