Avocado Case study

(ML-1 Project)

Ankita Bhanushali

Pursuing M.Sc. Statistics
Data Analyst, Kantar

INSAID, March 2020 GCD Cohort.

About Dataset

The avocado, a tree likely originating from south-central Mexico, is classified as a member of the flowering plant family Lauraceae. The fruit of the plant, also called an avocado, is botanically a large berry containing a single large seed. The table below represents weekly 2018 retail scan data for National retail volume (units) and price.

The most important variables are the following:

- 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

PLU stands for product lookup codes. The 3 different PLUs are simply just 3 different kinds of Hass avocados. Other variations (such as greenskins) are not included. The data comes from US observations, so prices are in dollars and regions are strictly American. (‘Size’ of avocado relates to the total weight in ounces or an estimate of the shape)

Measurements of avocado sizes ^

- 4046 - Total number of avocados with PLU 4046 sold (Hass Variety, Small Size, Size 60 and smaller)
- 4225 - Total number of avocados with PLU 4225 sold (Hass Variety, Large Size, typically Size 40-48)
- 4770 - Total number of avocados with PLU 4770 sold (Hass Variety, All Sizes, Size 36 and larger)

Table of Contents

01 Problem Statement

02 Exploratory Data Analysis

03 Model Evaluation

04 Conclusion

01

Problem Statement

Problem statement

The goal is to build a model to predict the Average price of Avocados which is continuous in nature of the different types of avocados

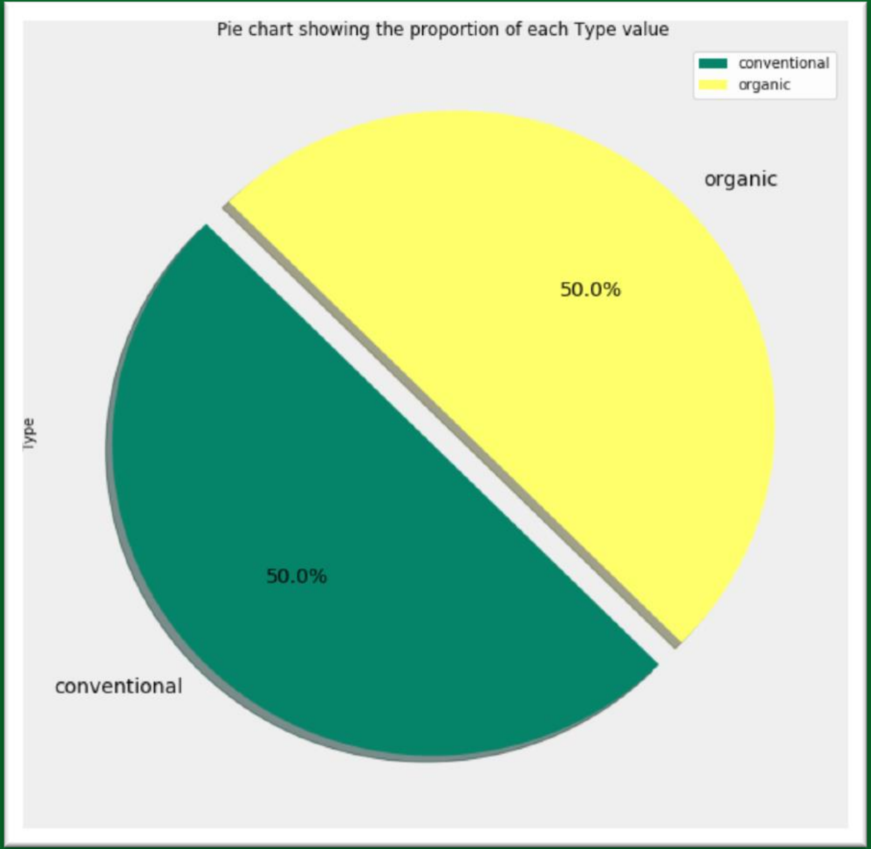
Exploratory Data Analysis

02

EDA

By Type:- Organic VS Conventional

Pie chart indicates both the types have equal proportion i.e. 50%.

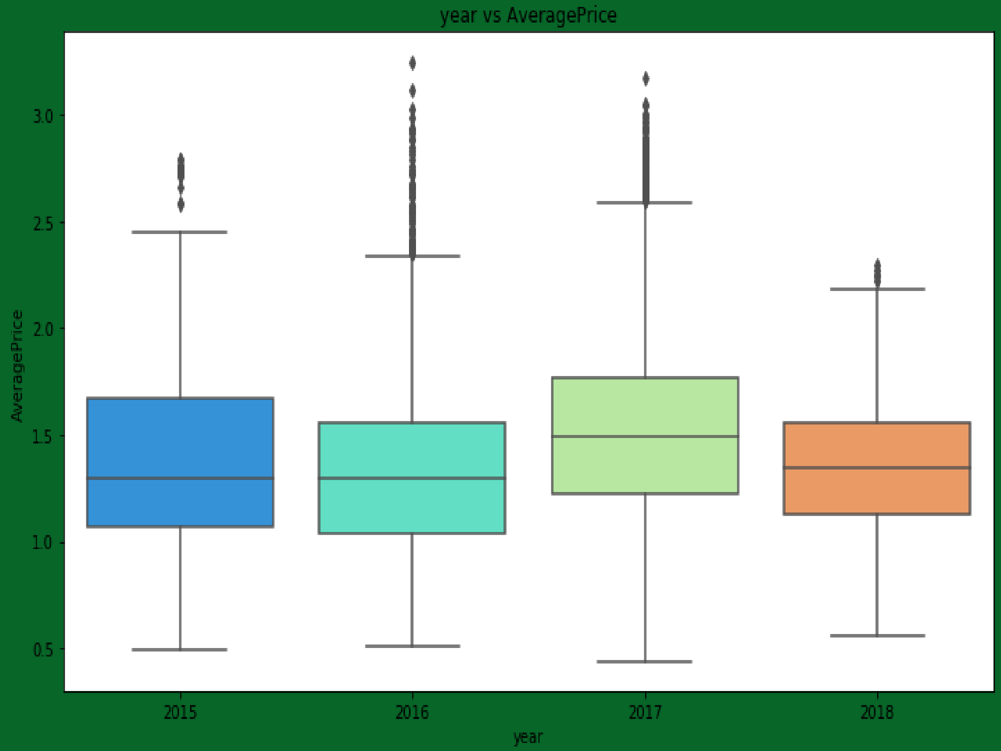


EDA

By year:- Average Price

- Looking at trended data from 2015 to 2018, year 2017 has Maximum Price.

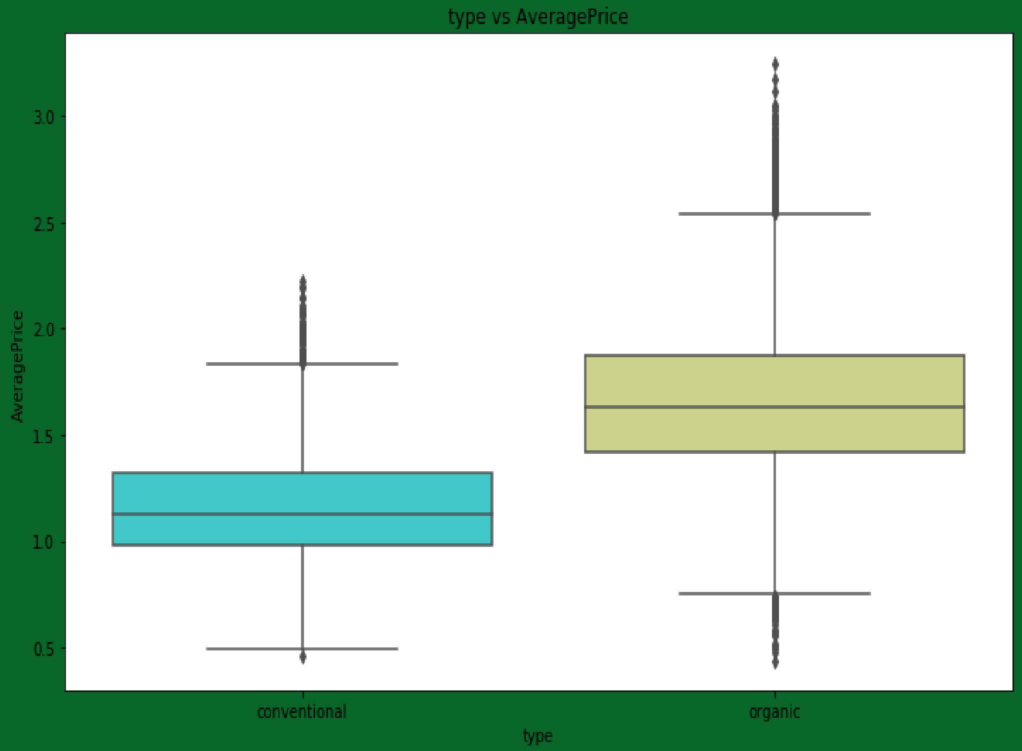
- Rest of year has less difference.



EDA

By type:-Average price

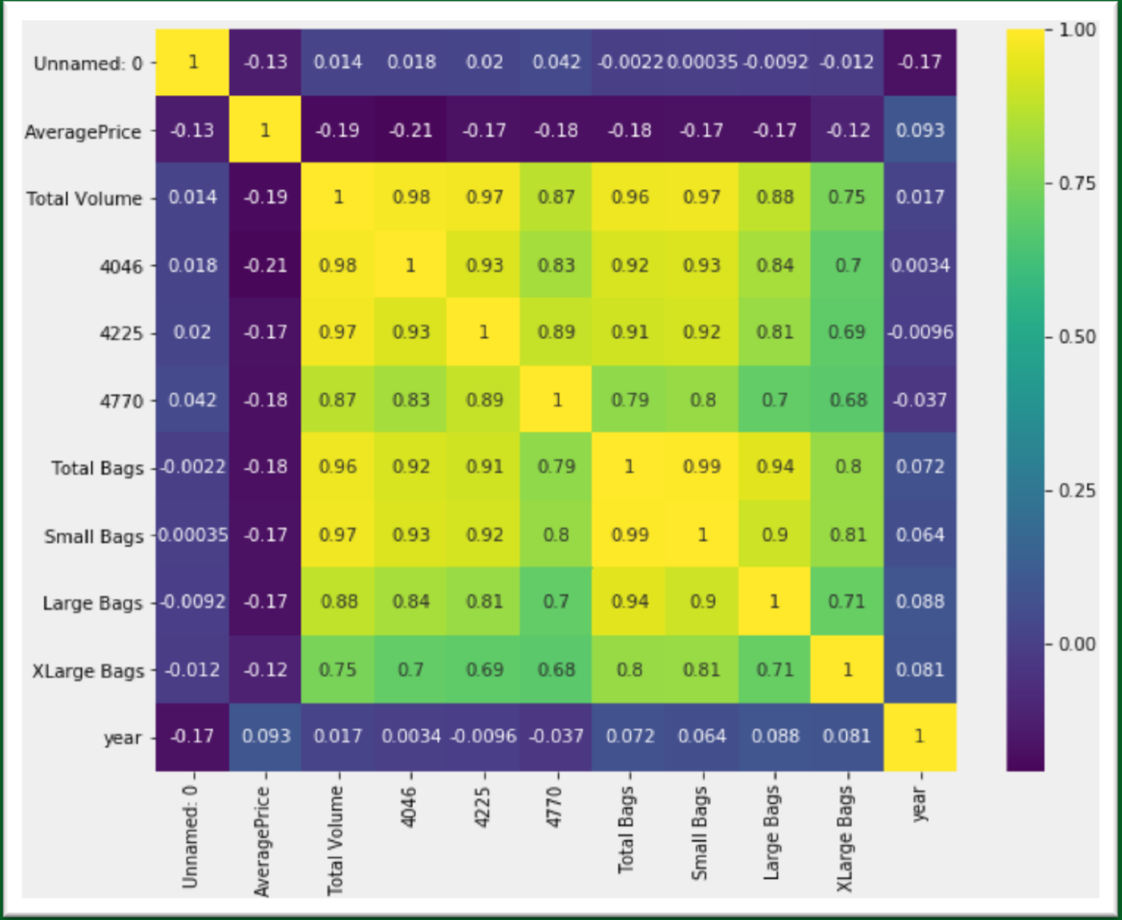
While comparing types of avocado with respect to average price, organic has highest average price.



EDA

Correlation Matrix

Variables are highly correlated in nature.

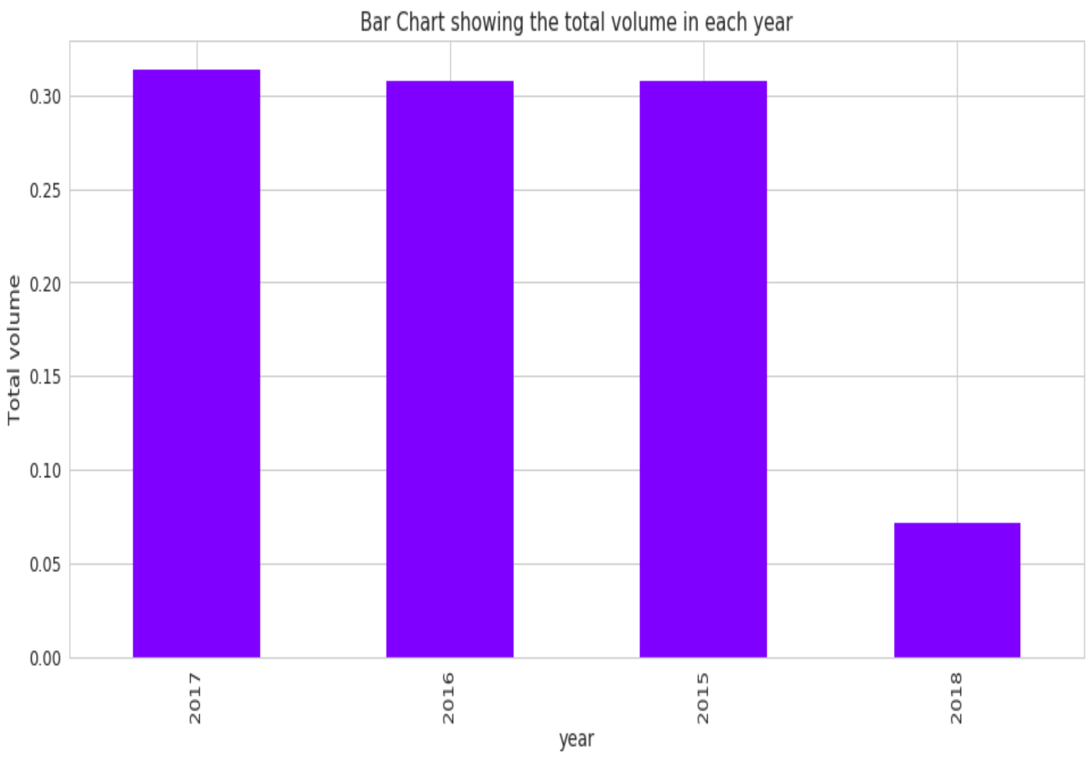


	Unnamed: 0	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	year
Unnamed: 0	1	-0.13	0.014	0.018	0.02	0.042	-0.0022	0.00035	-0.0092	-0.012	-0.17
AveragePrice	-0.13	1	-0.19	-0.21	-0.17	-0.18	-0.18	-0.17	-0.17	-0.12	0.093
Total Volume	0.014	-0.19	1	0.98	0.97	0.87	0.96	0.97	0.88	0.75	0.017
4046	0.018	-0.21	0.98	1	0.93	0.83	0.92	0.93	0.84	0.7	0.0034
4225	0.02	-0.17	0.97	0.93	1	0.89	0.91	0.92	0.81	0.69	-0.0096
4770	0.042	-0.18	0.87	0.83	0.89	1	0.79	0.8	0.7	0.68	-0.037
Total Bags	-0.0022	-0.18	0.96	0.92	0.91	0.79	1	0.99	0.94	0.8	0.072
Small Bags	0.00035	-0.17	0.97	0.93	0.92	0.8	0.99	1	0.9	0.81	0.064
Large Bags	-0.0092	-0.17	0.88	0.84	0.81	0.7	0.94	0.9	1	0.71	0.088
XLarge Bags	-0.012	-0.12	0.75	0.7	0.69	0.68	0.8	0.81	0.71	1	0.081
year	-0.17	0.093	0.017	0.0034	-0.0096	-0.037	0.072	0.064	0.088	0.081	1

EDA

By year:- Sale

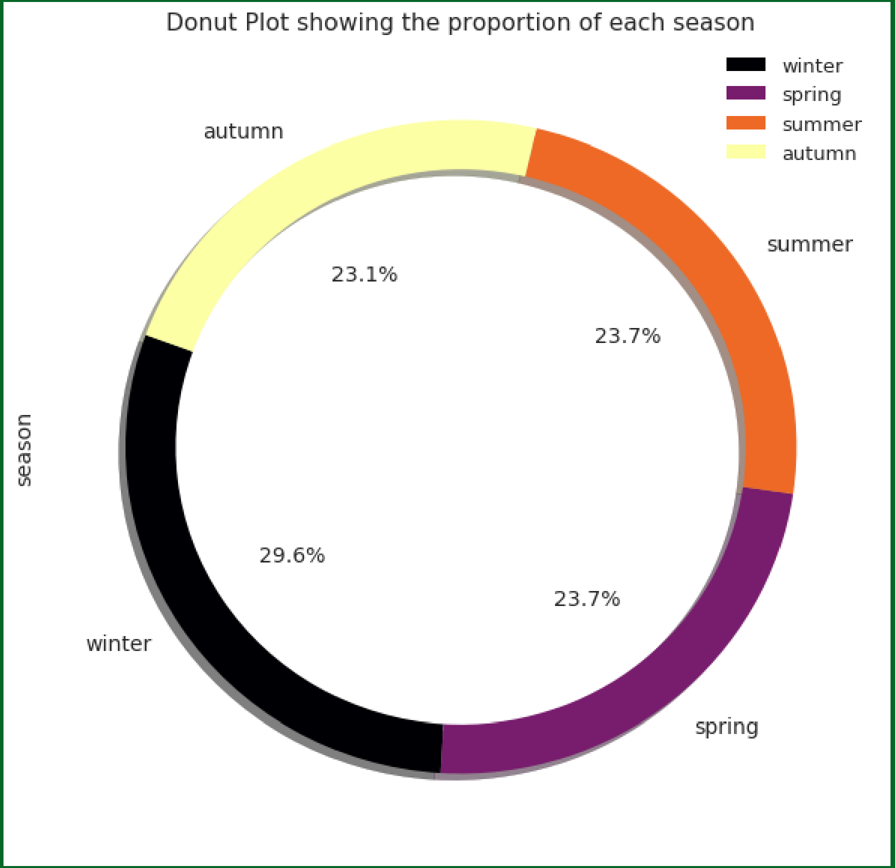
More number of avocado sold in the year 2017 compare to rest of the year.



EDA

By season:- Sell

- Little higher sell in winter compare to rest of the season.

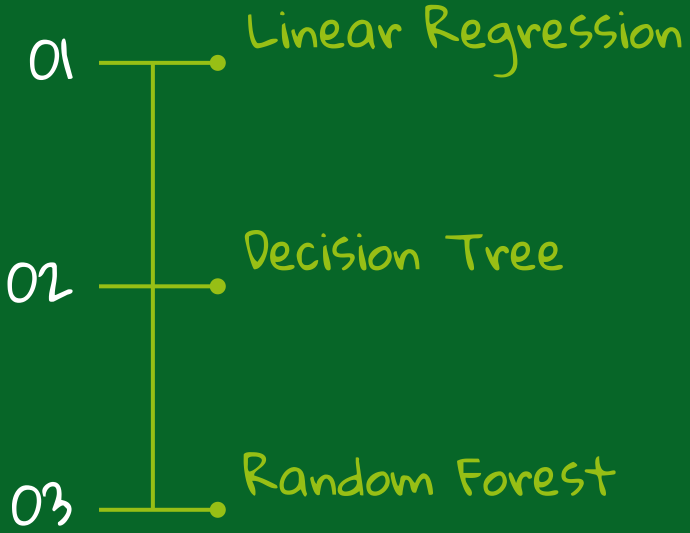


“Seasonality” is derived variable using **Month** information from the variable: **Date**

03

Model Evaluation

Modeling Techniques



01 Linear Regression

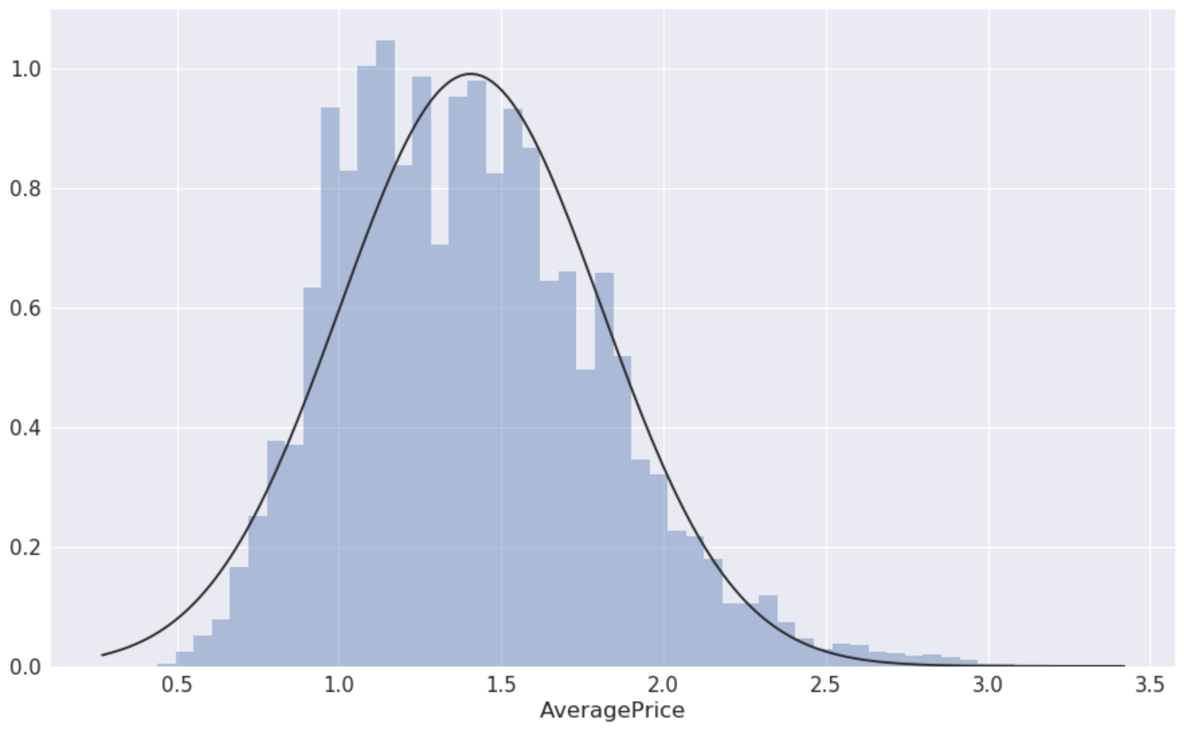
02 Decision Tree

03 Random Forest

Above techniques are used considering **Target variable(TV) -** Average price of avocado is continuous in nature.

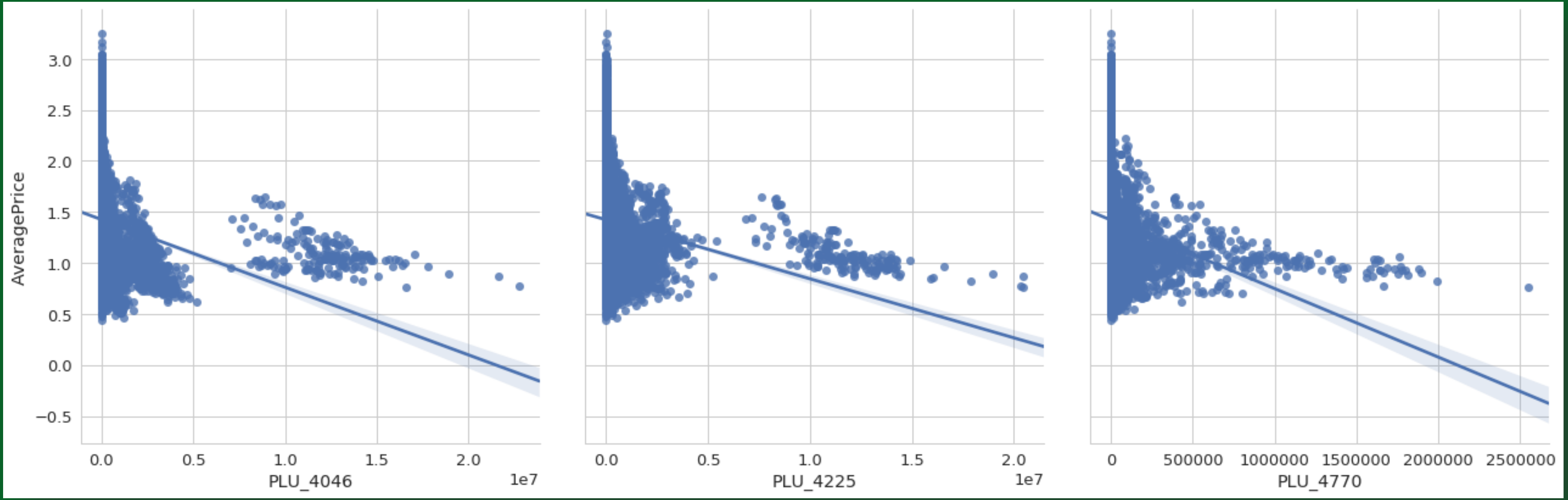
Assumption 1 : Normality

Average Price (Target Variable) is Normally distributed.



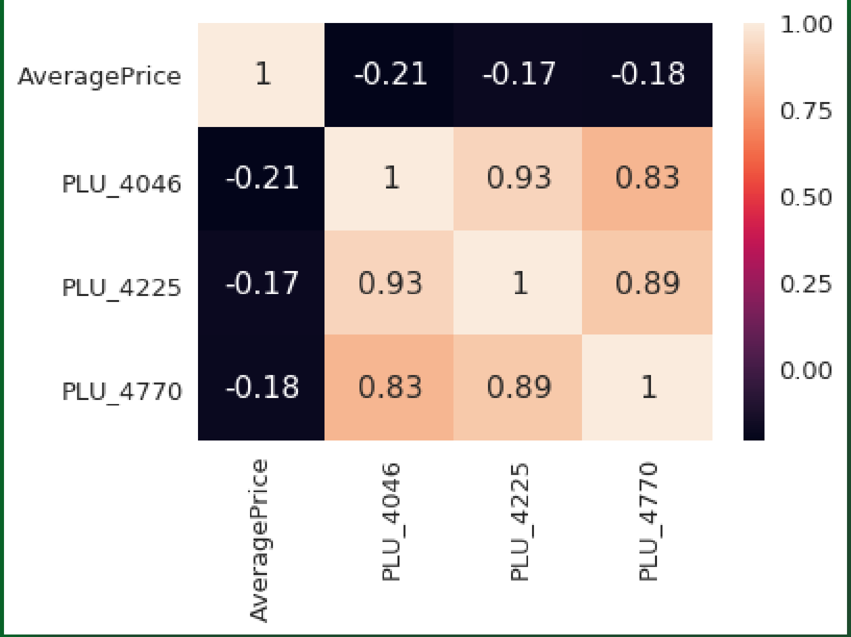
Assumption 2 : Linear Relationship

There is linear relation between Explanatory Variables i.e. size of Avocado & Target variables i.e. Average Price.



Assumption 3 : Multicollinearity

There is no Multicollinearity.
Average price & explanatory variables i.e. size of avocado are not significantly correlated.



	AveragePrice	PLU_4046	PLU_4225	PLU_4770
AveragePrice	1	-0.21	-0.17	-0.18
PLU_4046	-0.21	1	0.93	0.83
PLU_4225	-0.17	0.93	1	0.89
PLU_4770	-0.18	0.83	0.89	1

Model Evaluation

	Errors in different Models		
	Linear regression	Decision Tree	Random Forest
MAE	0.1936	0.1180	0.0878
MSE	0.0648	0.0336	0.0161
RMSE	0.2546	0.1833	**0.1270**

Looking at RMSE, best Fit model for the Avocado data set is **Random Forest Regressor**.

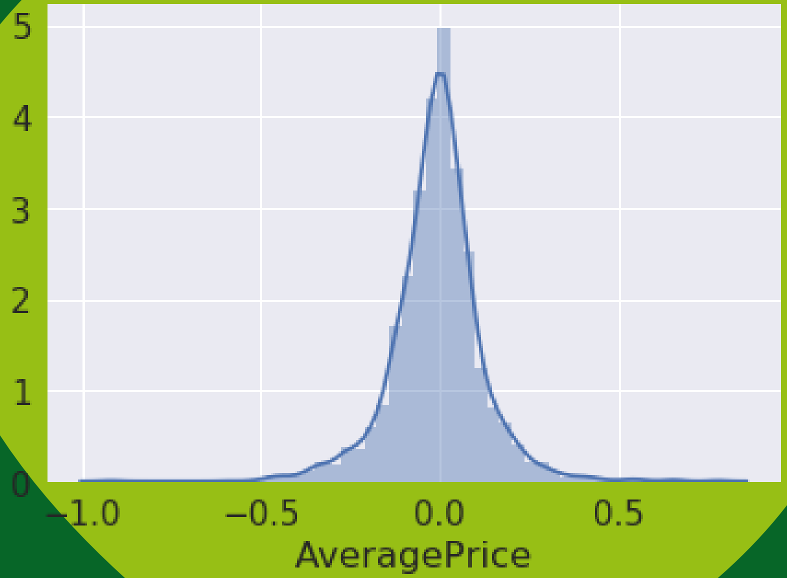
Conclusion

04

Conclusion

- Our residuals looked to be normally distributed and that's really a good sign which means that our model was a correct choice for the data.

- Random Forest Regressor over performed Linear Regression, and Decision Tree Regressor with an RMSE of 0.1270.



Resources

THANK
YOU