# **AVACADO PROJECT**

Problem Statement:

**Avocado is a fruit consumed by people heavily in the United States.**

### Content

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

Weekly retail scan data for National retail volume (units) and price for 2018 is demonstrated in the table below. The retailers’ cash register supplies this data which is grounded on actual retail sales of Hass avocados.

The table below reflects an increased, multi-outlet retail data set since 2013. Multi-outlet reporting comprises an aggregation the following channels: grocery, club, mass, dollar, drug and military. The Average Price (of avocados) in the table reflects as per unit (per avocado) cost, even when multiple units (avocados) are sold in bags.

The table containing Product Lookup codes (PLU’s) are only for Hass avocados. Other variants 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

1. Problem Definition:

Avocado price data comprises of year wise analysis from 2015 to 2018. The extracted and downloaded data from “Avocado project” covers the .csv files ranging from average prices, types (conventional or organic) to cities and regions where avocados were sold. The goal is to predict the average price which is continuous in nature of the different type of avocado and using the region that in which region they are lying.

# 2. Data Analysis:

# Data Preparation and Cleaning

* Reading the CSV file and doing initial statistical analysis (shape, values etc)
* Data Pre-processing: Reading the unique values for each column and removing those which won’t be significant in the analysis further.
* Create a new data frame to proceed with the analysis further

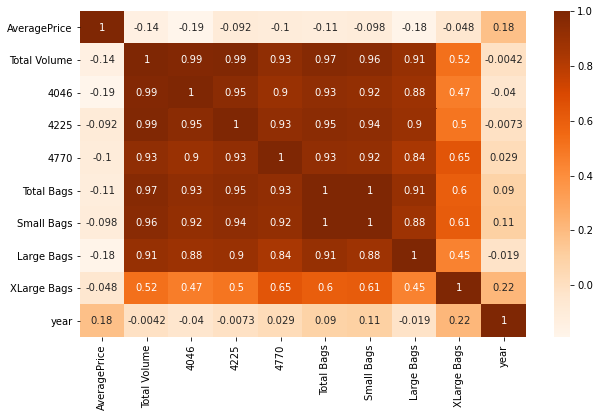
Dataset contains

* 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

This project is based on a hypothetical dataset downloaded from Avacado project which has 1,517 data points (rows) and 35 features (columns) describing each fruit background and characteristics. Machine Learning models can help to understand and determine how these factors relate to workforce attrition

I had a hypothesis of demand and supply e.g., if the consumed volumes are higher, then the prices would be lower. Noteworthy, the scatter plot I mentioned, using matplotlib in Python libraries clearly depicts that it seems there is a trend for that direction as well. The Pearson correlation coefficient showed a small negative correlation between the average price and average volume consumption. Thus, there is a coalition between demand and supply, but it cannot explain the price structure and the factors linked to it. Having some outliers on the right side of the plot where some cities had the highest prices while the consumed volumes are limited.

The correlation between different features of the dataset showed that employees with low satisfaction level are in left. The correlation heat map is shown below:



Although the correlation matrix does not indicate any high degree of correlation with the dependent variable, it does serve us with a clear view of all the factors.

* All the features in the heatmap above are not correlated with the **Average Price column**. Instead of that most of them are correlated with each other. So as for now I am bit worried because that might not aid to obtain a good model. Let’s give it a chance and see.
* First, we have to perform some Feature Engineering on the **categorical Features: region and type**

# 3. EDA Concluding Remark.

* Search patterns of data through visualization and trace the hidden trends from collected data.
* Visualisation of data using both matplotlib and seaborn library.
* Finding relationships between features using bar graphs, histograms, box plots, heatmap.
* Analysing both the numerical and the categorical columns separately.

### \* Data Loading and Description

* This data was downloaded and provided by Data trained academy website Avocado Board website in May of 2018 & compiled into a single CSV.
* Represents weekly 2018 retail scan data for National retail volume (units) and price.
* The dataset comprises of **1517 observations of 35 columns**. Below is a table showing names of all the columns and their description.
* The unclear numerical variables terminology is explained in the next section:

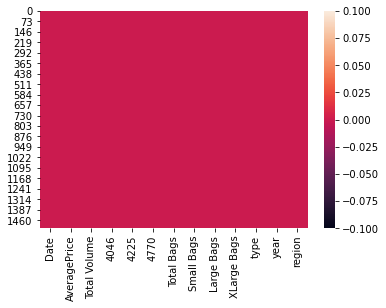
|  |  |
| --- | --- |
| **Features** | **Description** |
| Average Price | Its just a useless index feature that will be removed later |
| ‘Total Volume’ | Total sales volume of avocados |
| ‘4046’ | Total sales volume of Small/Medium Hass Avocado |
| ‘4225’ | Total sales volume of Large Hass Avocado |
| ‘4770’ | Total sales volume of Extra-Large Hass Avocado |
| ‘Total Bags’ | Total number of Bags sold |
| ‘Small Bags’ | Total number of Small Bags sold |
| ‘Large Bags’ | Total number of Large Bags sold |
| ‘XLarge Bags’ | Total number of XLarge Bags sold |

Here type will be the target variable. The dataset is well organised with no missing values Target class is imbalance.



Here heat map contains the null values of the Dataset.

As per my findings, there are no null values in the data set below because the red colour is distributed equally correspond to each column.

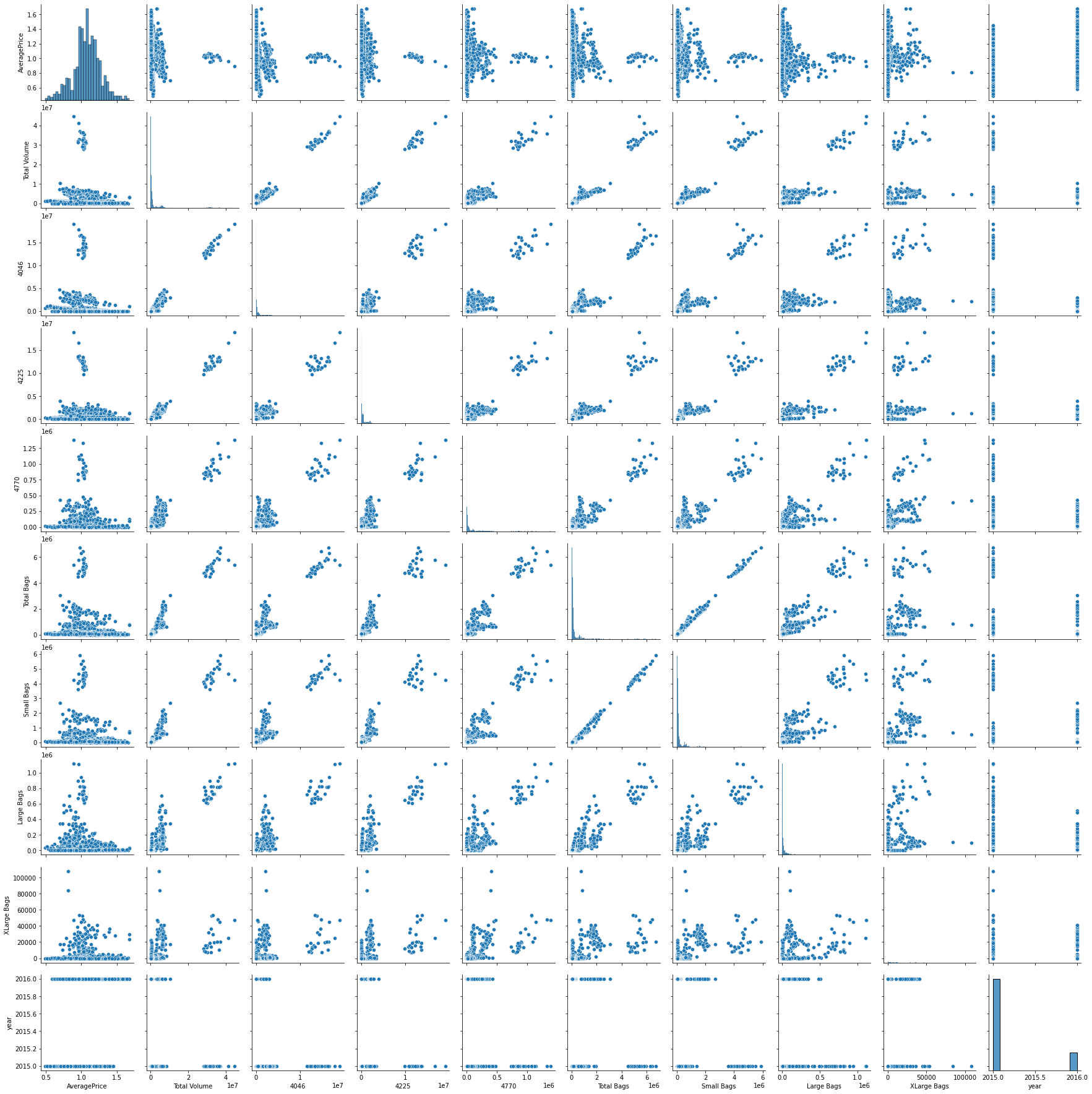


sns. pairplot(df)

Remove the missing values.

drop the nagativitycorrelated columns.

remove the outliers.



# 4. Pre-Processing Pipeline:

Encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.

This is essential to perform since the machine learning algorithms only work on the numerical data. So, the necessity to convert the categorical column into a numerical one is must.

For each of the transformations in Python that has a fit\_transform() method , we can wrap them up in an actual pipeline that executes them all in order and even go back and view attributes of each of the transformations. Additionally, you can set parameters for each transformation and the syntax for that is in the link I just shared. If anything, using this pipeline has just cleaned up my code a lot and organized my thinking better.

Sklearn provides a very efficient tool for **encoding** the levels of categorical features into numeric values. **Label Encoder encode labels** with a value between 0 and n\_classes-1 where n is the number of distinct **labels**. If a **label** repeats it assigns the same value to as assigned earlier .Convert Region and Type into numeric value by using encoder.

Data needs to be pre-processed because machine learning models are efficient in reading numbers than words. Using label encoding, categorical data can be replaced with numbers. Below code is to display all categorical data

**from** **sklearn.preprocessing** **import** Label Encoder

LE=Label Encoder()

df["year”] =LE.fit\_transform(df["year"])

Using the above label encoding method, categorical data can be replaced with number.

# 5. Building Machine Learning Models.

* Let’s apply our model which is going to be the **Linear Regression because our Target variable 'AveragePrice' is continuous**.
* Let's now begin to train out regression model We will need to first split up our data into an **X array that contains the target variable**, and a **y array with the target variable**.

x=df. drop(["AveragePrice"], axis=1)

y=df["AveragePrice"]

x\_train, x\_test, y\_train, y\_test=train\_test\_split (x, y, random\_state=50, test\_size=0.2)

Here we can use Regression method to build the models

1. Decision Tree Regressor
2. Logistic Regression
3. Random Forest Regressor

* I had done this prediction by taking average price as an output variable which is continuance in nature so that why I’m using the regression technique.
* While calculating the best random state the 80 is best state which provides the highest R2 score value for this model.
* I can find the best param using GridSeachCV and then I used these param for that model.
* There are following matrices which I find, and which are providing the best score.
* I also performed the scatter plot graph and it depicts that the actual value and predicted values are very close to each other, so the line is the best fit line.

Now, I am finding the score by taking region as a y value with the help of Regressor method because the region data is categorical in nature. So I am importing the Regressor model and their matrices.

The highest final model of the dataset: In the end, we can conclude that utilizing data science on Average Price is very much beneficial to the business as we can tag each fruit with the averages price and bring new ideas for customized Avacado retention for each group.

According to the Regressor report, the accuracy of the model is 63%. However its recall is lower at 15% of positive cases. The DecisionTreeRegressor model is providing excellent results, however the purpose of the problem is to identify fruits that are likely to be sold. That is why accuracy then becomes a very important measure as it gives clear picture of values that are identified correctly.

Decision Tree Classifier has emerged as the final winning model with 63% and the highest. This could be the highest possible score achieved with the inherent limitations in the dataset. Therefore, we can conclude that Decision Tree Regressor is the best model.

Machine learning models are as good as the data to feed it, and more data would strengthen the model. For example, in this dataset, the feature ‘Performance Rating’ has been restricted to scores of 3 and 4 only. More insights could be generated if the full spectrum of performance ratings is included. In the real-life situation, getting the right data is often more challenging than the analytics itself.

# 6. Concluding Remarks.

* With the help of notebook I learnt how **EDA** can be carried out using **Pandas and other plotting libraries**.
* Also, I have seen making use of packages like **matplotlib, plotly and seaborn** to develop better insights about the data.
* I have also seen how **preproceesing** helps in dealing with **missing values and irregualities** present in the data. I also learnt **how to create new features** which will in turn help us to better predict the survival.
* I also make use of **pandas profiling** feature to generate an html report containing all the information of the various features present in the dataset.
* I have seen the impact of columns like **type, year/date** on the **Average price increase/decrease rate.**
* All of the performed analysis helped me a lot to know the **features on which price is highly positively and negatively coorelated with.**
* This project helped me to gain insights and how I should go with the flow, which model to choose first and go step by step to attain results with good accuracy. Also I get to know **where to apply Linear, Decision Tree and other applicable and required models to fine tune the predictions**.
* Random Forest Regressor model predicts the average price more accurately than linear regression model.
* In this project, the trend and periodity of avocado price and sales volume time series as well as their association is analysed. We extracted monthly and annual patterns from the spectrum density analysis and also determined the trend of price variation from the spectrum decomposition, which is not constantly increasing but shows a decreasing trend in recent years. Additionally, we applied a regression on the price and sales volume time series and discovered a negative correlation between the two time series. This is consistent with our empirical knowledge.
* The visualisation we were aiming to achieve at the beginning of this project has now become attainable.