META SCIFOR TECHNOLOGIES, BANGALORE.

AI INTERN

MINOR PROJECT-1
REPORT

(AVOCADO PROJECT)

By,

M.Rachel

CONTENT

1) PROBLEM STATEMENT

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

The table below 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.

FEATURES:

- Date - The date of the observation
- type - Conventional or Organic
- year - year
- 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

Target:

The problem statement can been analysed as two different problems such as Regression and Classification problems.

For Regression target :

- Average Price - The Average price of a single avocado

For Classification target:

- Region - the city or region of the observation

2) DATA EXPLORATION

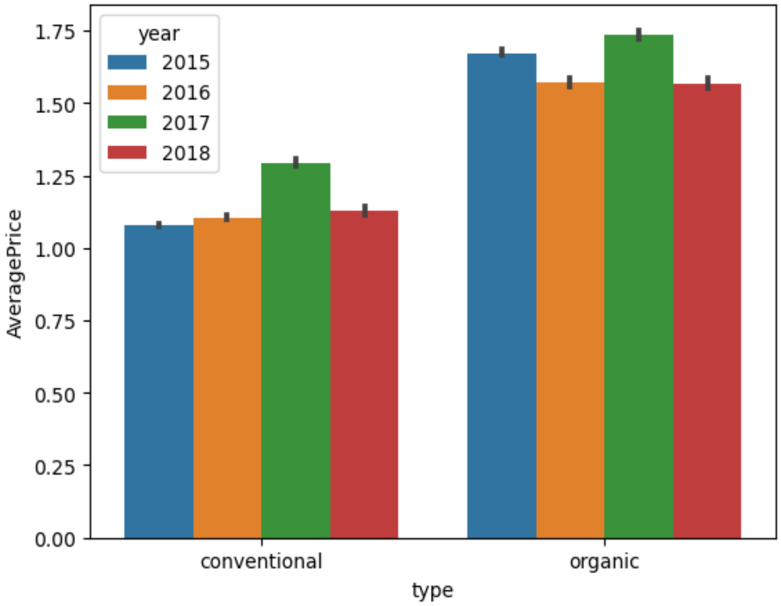
- Importing the Necessary Libraries such as Numpy, Pandas, Data Visualization libraries such as Matplotlib, Seaborn and Machine Learning Libraries such as Scikit-learn, Linear Regression for Regression task, Logistic Regression for Classification task, train_test_split for splitting the dataset as training set and testing set for Model Training, and Lastly Evaluation metrics such as accuracy score, r2 score, Error metrics such as MAE, MSE, RMSE.

- Load the dataset using pandas, explore the data such as load first 5 rows of the data, checking the information of the data which shows the datatypes and null values.

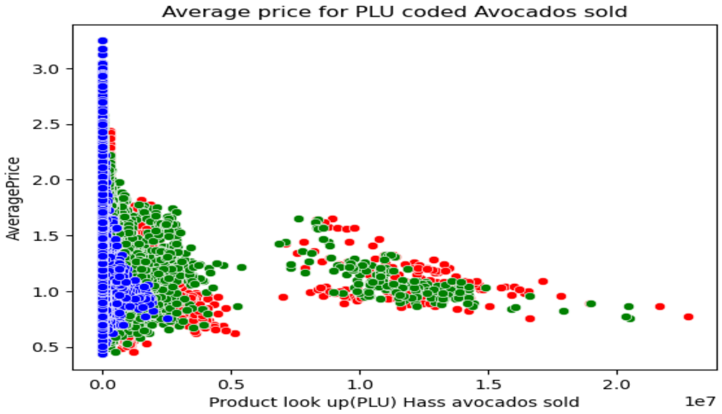
- Since there are null values in the data check and treat the many 0 values with their mean and treat the outliers but before lets visualize the data distribution and plotting with different features for better understanding of data.

3) DATA VISUALIZATION

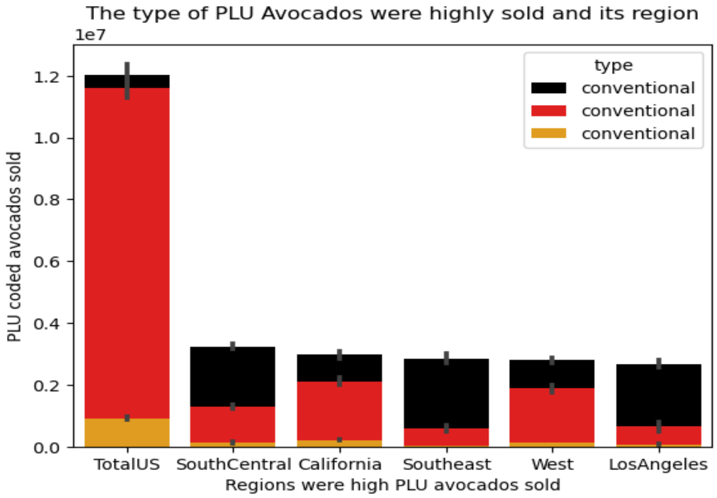
- Visualize type of Avocado with Average price with comparison of year



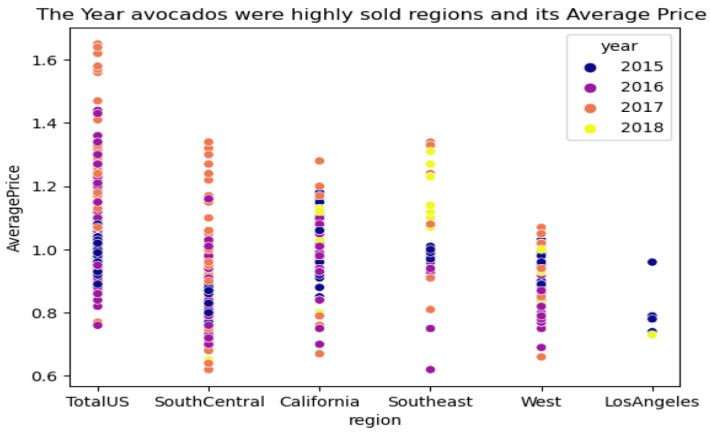
- visualize the PLU Hass Avocados with their Average price



- visualize high PLU coded avocados sold with region and difference with type of avocados



- visualize highly sold avocados region and their average price with their year



4) DATA PREPROCESSING

- Label Encoding all the Categorical Features
- Plotting the distribution plot for checking all features to be Normal distribution.
- Plotting the boxplot for visualizing the outliers.
- Through quantile 25% (q1) and quantile 75%(q3) we calculate the IQR range and treat the outliers.
- Reindexing the data after treating the outliers of the features.
- Checking for Multicollinearity problem through correlation matrix, out of all features Total bags features has high correlation around 0.82 so we drop that feature.
- Now we the data set is ready for Model training.

5) MODEL BUILDING OR TRAINING

- We have done Model Training as 2 Tasks
  1) **Regression Task**
  2) **Classification Task**

1) Regression Task

- Splitted the dataset as X for Independent features and y for target variable as Average price.
- Transforming the X features into scaled features through Standardization.
- Using train test split function Splitting the x scaled and y as training set and testing set with the parameters.
- Now for Model Training for the regression task we use Linear Regression Machine Learning Algorithm, fitting the model with training sets.
- Now predict the model with r2 training score as 0.34 and test score as 0.32 and check the overfitting issue with error metrics.
- For proving the model training is not overfitted we have used regularization method such as L1 and L2 form Lasso and Ridge.

2) Classification Task

- Splitted the dataset as X for Independent features and y for target variable as region.
- Transforming the x features into scaled features through Standardization.
- Using train test split function Splitting the x scaled and y as training set and testing set with the parameters.

- Now for Model Training for the classification task we use Logistic Regression Machine Learning Algorithm, fitting the model with training sets.
- Now predict the model with accuracy score with training as 40% and test score as 39% and the model is not overfitted by Classification metrics by precision, recall scores.

6) EVALUATION METRICS

Regression Problem:

R2 Score:

1) Training score - 0.34
2) Test score - 0.32

Error Metrics:

1) MAE- 0.24
2) MSE- 0.10
3) RMSE- 0.32

Classification Problem:

Accuracy score:

1. Training score - 40%
2. Test score - 39%