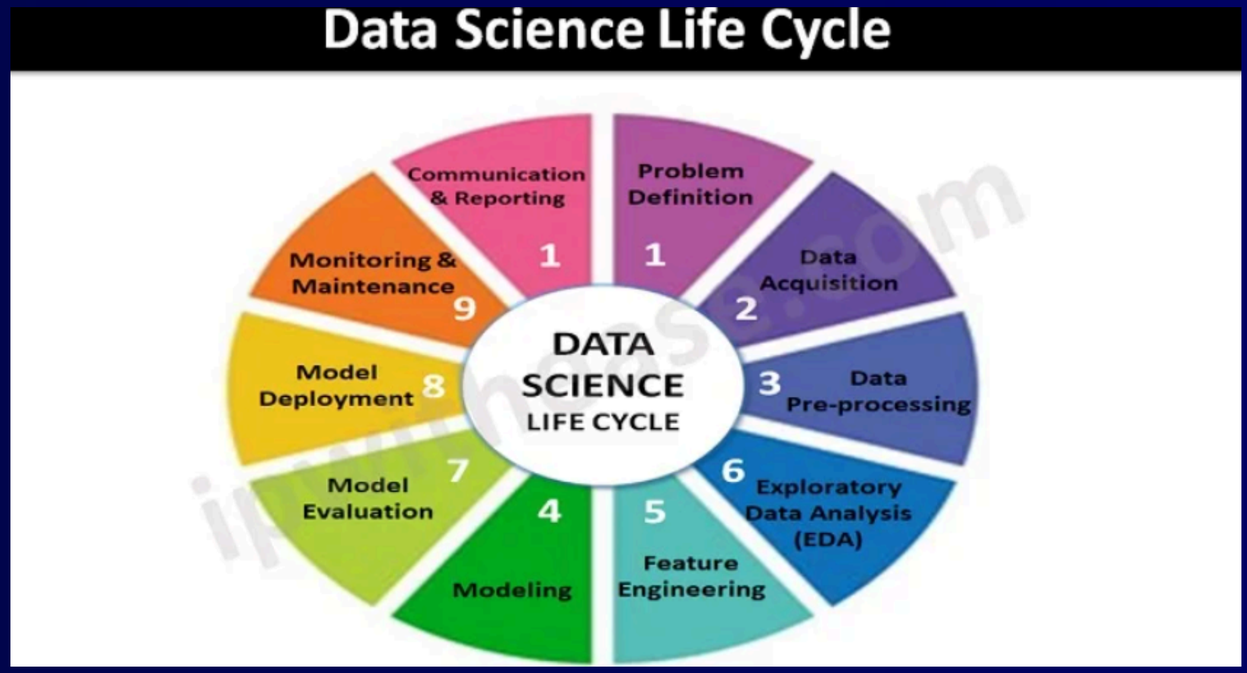
CORNERSTONE PROJECT :

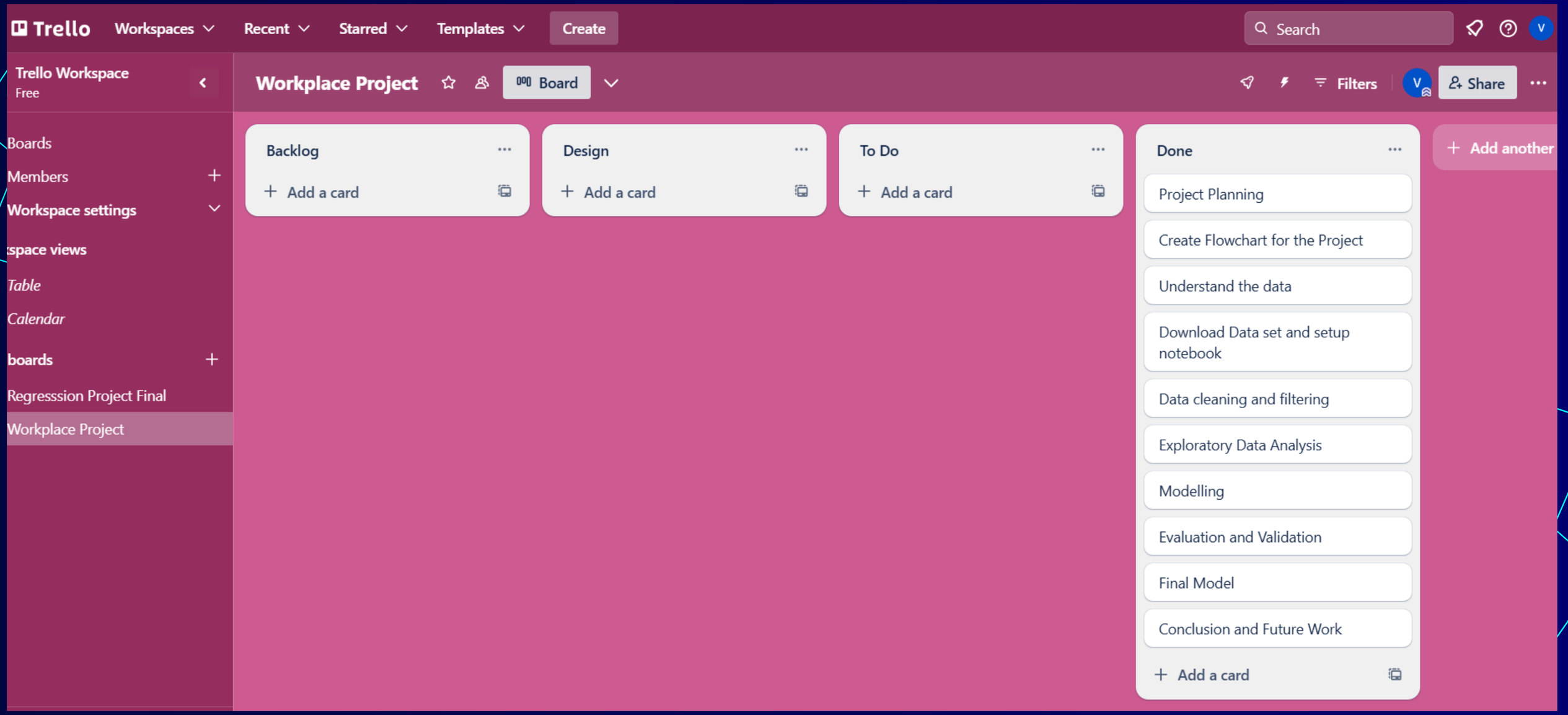
ANALYSIS OF AVOCADO PRICES AND SALES VOLUME 2015-2023, IN THE USA

PROJECT OVERVIEW

- **Objective of the Project: The aim of this project is to analyze the average prices and volume of avocado's sold per region in the USA.**
- **Data Source: The data was downloaded from Kaggle and details the volume and average selling prices of avocados across the USA over 5 years(2015 - 2023). The data has been updated till December 3, 2023.Data provided here was collected from Hass Avocado Board - Category Data.**
- **Key Questions or Hypotheses: Null hypothesis , the Total Volume of avocados consumed is negatively proportional to the average price of avocados. The goal is to determine which regression Models can best be used to predict future avocado prices.**
-



PROJECT PLANNING : TRELLO BOARD



DATA DESCRPITION & COLLECTION :

- **Purpose : The data was downloaded from Kaggle and details the volume and average selling prices of avocados across the USA over 5 years(2015 - 2023). The data has been updated till December 3, 2023.Data provided here was collected from Hass Avocado Board - Category Data.**

- **Note regarding the regions mentioned in dataset "region" includes both regions and key locations which are cities or sub-regions. The values for locations do not add up to that of regions. The following are regions, as described here (another description of the dataset) California, West, Plains, South Central, Southeast, Midsouth, Great Lakes, Northeast.**

- **Details: The dataset consists of 53415 rows (observations) and 12 columns (features).**

DATA CLEANING AND FILTERING

- **Purpose: Prepare the data for analysis by cleaning and filtering.**
- **Details: Best practice is to Include steps for handling missing values, removing outliers, correcting errors, and possibly reducing the data (filtering based on certain criteria or features).**

Summary of Data :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53415 entries, 0 to 53414
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Date          53415 non-null  object
 1   AveragePrice  53415 non-null  float64
 2   TotalVolume   53415 non-null  float64
 3   plu4046       53415 non-null  float64
 4   plu4225       53415 non-null  float64
 5   plu4770       53415 non-null  float64
 6   TotalBags     53415 non-null  float64
 7   SmallBags     41025 non-null  float64
 8   LargeBags     41025 non-null  float64
 9   XLargeBags    41025 non-null  float64
 10  type          53415 non-null  object
 11  region        53415 non-null  object
dtypes: float64(9), object(3)
memory usage: 4.9+ MB
```

Steps I used in the data cleaning process is as follows :

- **Check for null values and replace with a zero**
- **Check for duplicate rows , there we no duplicate rows in my data**

EXPLORATORY DATA ANALYSIS (EDA)

- **Purpose: Explore and visualize the data to uncover patterns, trends, and relationships.**
- **Details: Use statistics and visualizations to explore the data. This may include histograms, box plots, scatter plots, and correlation matrices. Discuss any significant findings.**

```
# Convert the DataFrame to a table format
print(tabulate(summary, headers='keys', tablefmt='grid'))
```

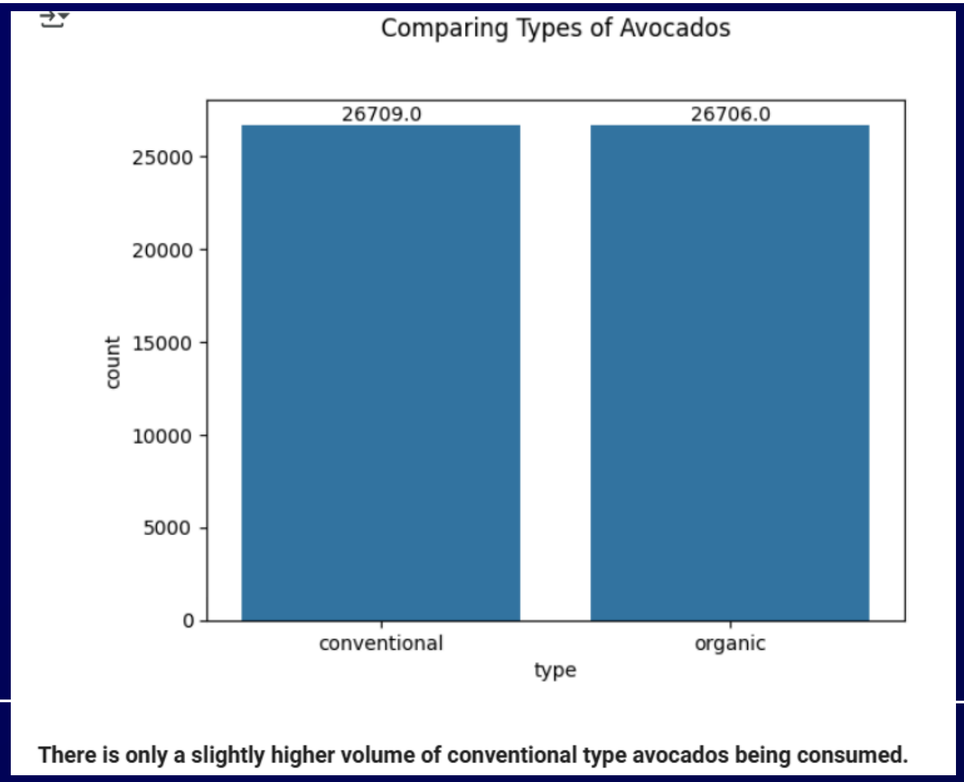
```
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
|       |  AveragePrice |   TotalVolume |       plu4046 |       plu4225 |       plu4770 |    TotalBags |     SmallBags |     LargeBags | XLargeBags |
+=======+===============+===============+===============+===============+===============+===============+===============+===============+============+
| count |         53415 |         53415 |         53415 |         53415 |         53415 |         53415 |         41025 |         41025 |      41025 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
| mean  |         1.429 |        869447 |        298271 |        222217 |         20532 |        217508 |        103922 |       23313.2 |    2731.81 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
| std   |         0.393 |     3.54527e+06 |     1.30767e+06 |        955462 |        104098 |        867695 |        569261 |        149662 |    22589.1 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
| min   |          0.44 |         84.56 |             0 |             0 |             0 |             0 |             0 |             0 |          0 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
| 25%   |         1.119 |       16264.7 |       694.725 |        2120.8 |             0 |       7846.52 |             0 |             0 |          0 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
| 50%   |           1.4 |        120352 |       14580.6 |       17516.6 |         90.05 |       36953.1 |        694.58 |             0 |          0 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
| 75%   |          1.69 |        454238 |        128792 |       93515.6 |       3599.74 |        111015 |         37953 |       2814.92 |          0 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
| max   |         3.441 |     6.10345e+07 |     2.54472e+07 |     2.04706e+07 |     2.86003e+06 |     1.62983e+07 |     1.25672e+07 |     4.32423e+06 |     679587 |
+-------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+---------------+------------+
```

Results : The average price across the dataset is $1.42 , the average total volume consumed is 8.59 pounds.

type	count
conventional	26709
organic	26706

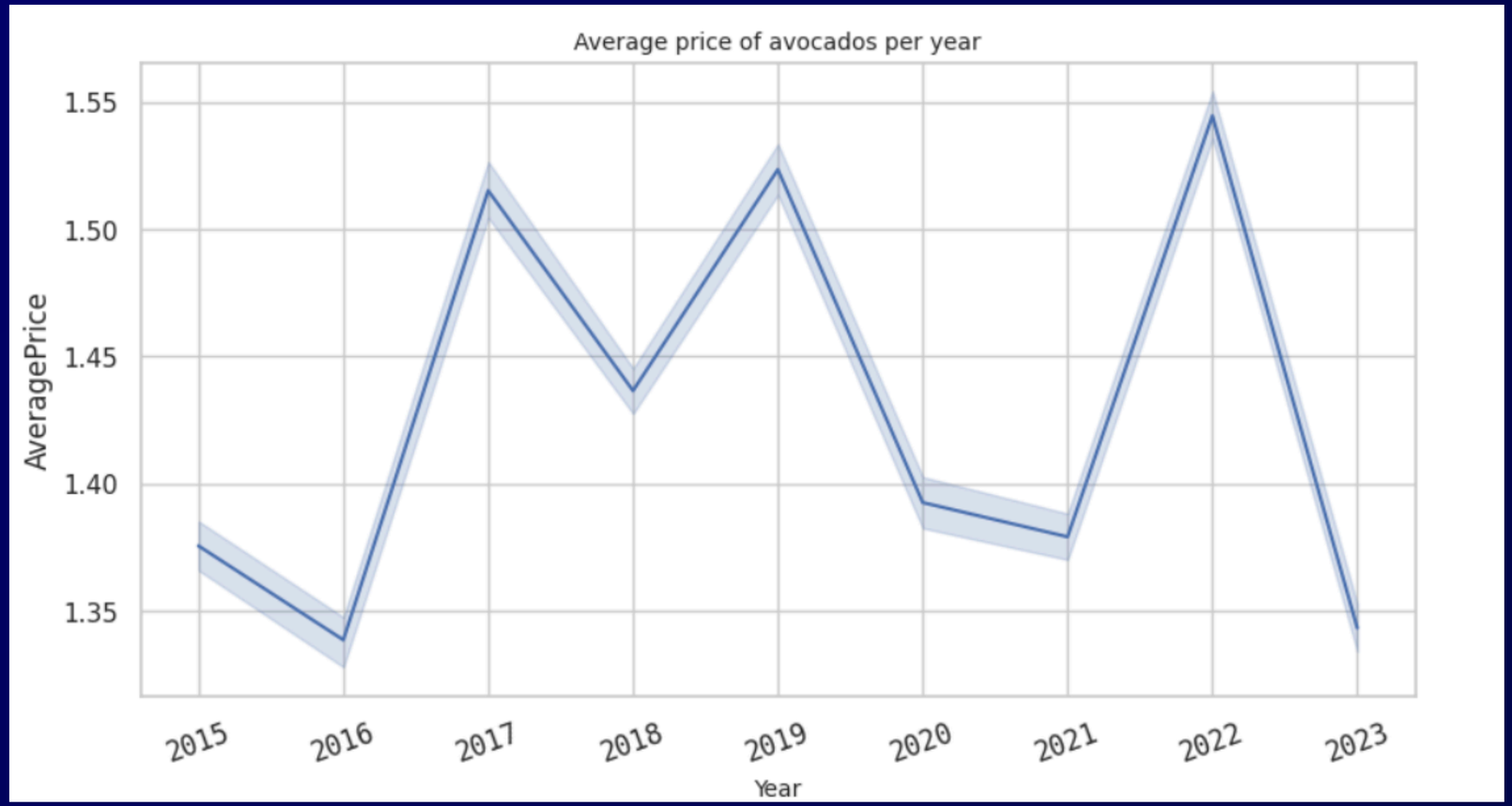
dtype: int64

Results : Based on the data provided, it's evident that there slightly more conventional avocados being consumed than organic.



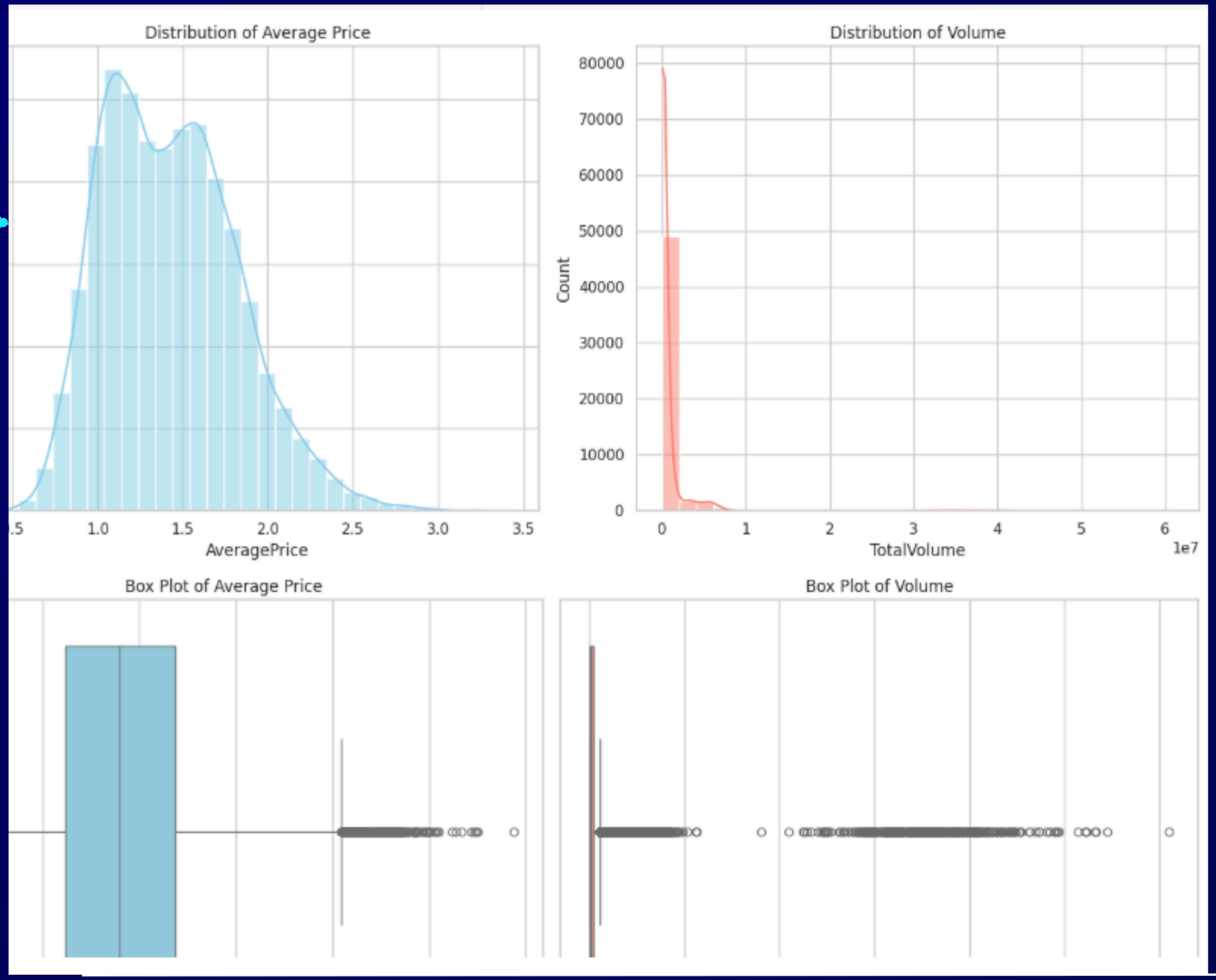
There is only a slightly higher volume of conventional type avocados being consumed.

EDA : TREND ANALYSIS OF THE AVERAGE PRICES OF AVOCADO'S ACROSS THE US FROM 2015 TO 2023



THE PRICES OF AVOCADOS ARE SPORADIC , POSSIBLE DUE TO HARVEST CONDITIONS , CLIMATE CHANGES AND EXTERNAL FACTORS THAT INFLUENCE GROWING OF THIS TREE. NOTE THAT AVOCADO TREES NEED A LOT OF WATER TO GROW SO THERE COULD HAVE BEEN SPORADIC CONDITION TO PRODUCE A SPORADIC TREND ANALYSIS LIKE THIS.

HISTOGRAMS & BOX PLOTS



EXPLANATION OF PLOTS:

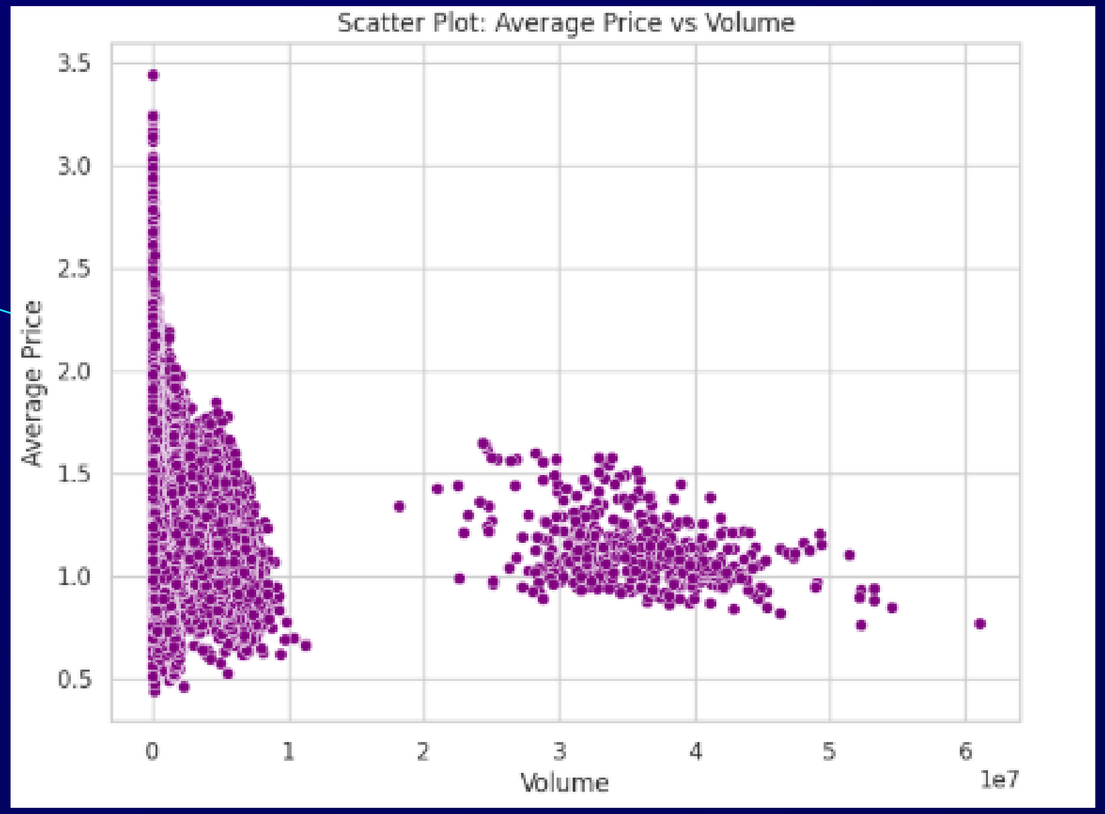
HISTOGRAMS:

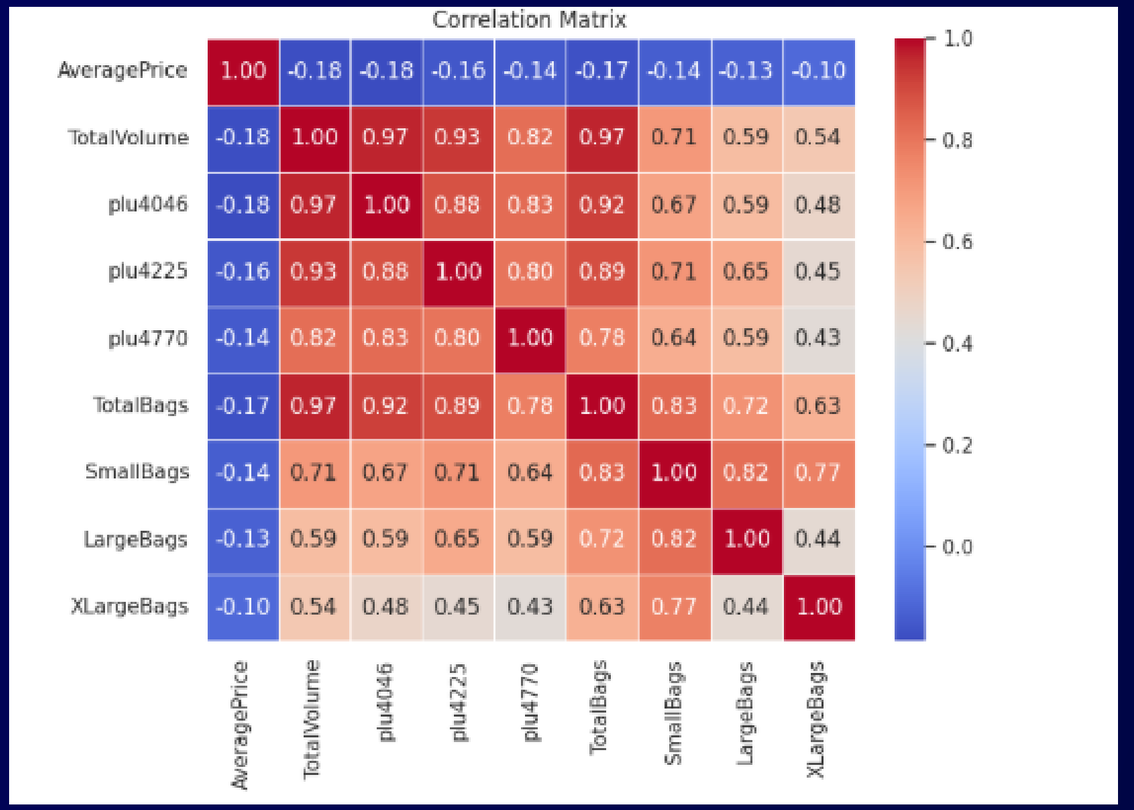
THE HISTOGRAMS FOR AVERAGEPRICE AND VOLUME SHOW THE DISTRIBUTION OF THESE TWO VARIABLES. FOR EXAMPLE: IF THE DISTRIBUTION OF AVERAGEPRICE IS SKEWED, IT MAY SUGGEST THAT AVOCADO PRICES ARE NOT UNIFORMLY DISTRIBUTED AND MIGHT HAVE PERIODS OF HIGHER PRICES. THE VOLUME HISTOGRAM CAN REVEAL HOW MUCH AVOCADO IS SOLD OVER TIME AND WHETHER THERE ARE PERIODS OF HIGHER OR LOWER SALES.

BOX PLOTS:

THE BOX PLOTS FOR AVERAGEPRICE AND VOLUME PROVIDE SUMMARY STATISTICS SUCH AS THE MEDIAN, INTERQUARTILE RANGE (IQR), AND ANY POTENTIAL OUTLIERS. IN THE CASE OF AVERAGEPRICE, IF THERE ARE A LOT OF OUTLIERS, IT COULD INDICATE SUDDEN PRICE SPIKES OR SIGNIFICANT VARIATIONS IN PRICE. FOR VOLUME, IF THE BOX PLOT SHOWS A WIDE SPREAD OR OUTLIERS, THIS COULD POINT TO PERIODS WITH HIGH SALES VOLUMES (LIKE CERTAIN SEASONS OR EVENTS).

SCATTER PLOT || CORRELATION MATRIX





SCATTER PLOT:

THE SCATTER PLOT BETWEEN VOLUME AND AVERAGEPRICE VISUALIZES ANY RELATIONSHIP BETWEEN THE TWO VARIABLES. IF THERE'S A TREND OR PATTERN (E.G., A POSITIVE OR NEGATIVE SLOPE), IT MIGHT SUGGEST A CORRELATION BETWEEN THE AMOUNT OF VOLUME SOLD AND THE PRICE. IF THERE'S NO CLEAR PATTERN, IT MIGHT INDICATE THAT VOLUME AND PRICE DO NOT HAVE A DIRECT RELATIONSHIP.

CORRELATION MATRIX:

THE CORRELATION MATRIX HEATMAP SHOWS THE PAIRWISE CORRELATIONS BETWEEN NUMERIC COLUMNS IN YOUR DATASET. CORRELATION VALUES RANGE FROM -1 TO 1: 1 INDICATES A PERFECT POSITIVE CORRELATION. -1 INDICATES A PERFECT NEGATIVE CORRELATION. 0 INDICATES NO LINEAR CORRELATION. THIS MATRIX CAN HELP YOU IDENTIFY POTENTIAL RELATIONSHIPS BETWEEN VARIABLES (E.G., BETWEEN AVERAGEPRICE AND VOLUME, OR BETWEEN AVERAGEPRICE AND YEAR).

FEATURE SELECTION

```
X = df.iloc[:,2:-1]
X
```

	TotalVolume	plu4046	plu4225	plu4770	TotalBags	SmallBags	LargeBags	XLargeBags	type
0	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.00	conventional
1	1373.95	57.42	153.88	0.00	1162.65	1162.65	0.00	0.00	organic
2	435021.49	364302.39	23821.16	82.15	46815.79	16707.15	30108.64	0.00	conventional
3	3846.69	1500.15	938.35	0.00	1408.19	1071.35	336.84	0.00	organic
4	788025.06	53987.31	552906.04	39995.03	141136.68	137146.07	3990.61	0.00	conventional
...	...	...	...	...	...	...	...	...	...
41020	232183.87	10823.86	16260.84	0.00	70369.80	66539.04	235.22	0.00	organic
41021	638610.43	210569.31	65265.76	7866.48	242168.31	235281.79	3072.02	1083.57	conventional
41022	21567.76	629.09	266.09	0.00	17391.15	17189.55	0.00	0.00	organic
41023	70289.47	16251.89	1464.33	6.10	21416.50	18595.17	1656.51	532.21	conventional
41024	2211.57	220.56	18.48	0.00	628.82	581.57	0.00	0.00	organic

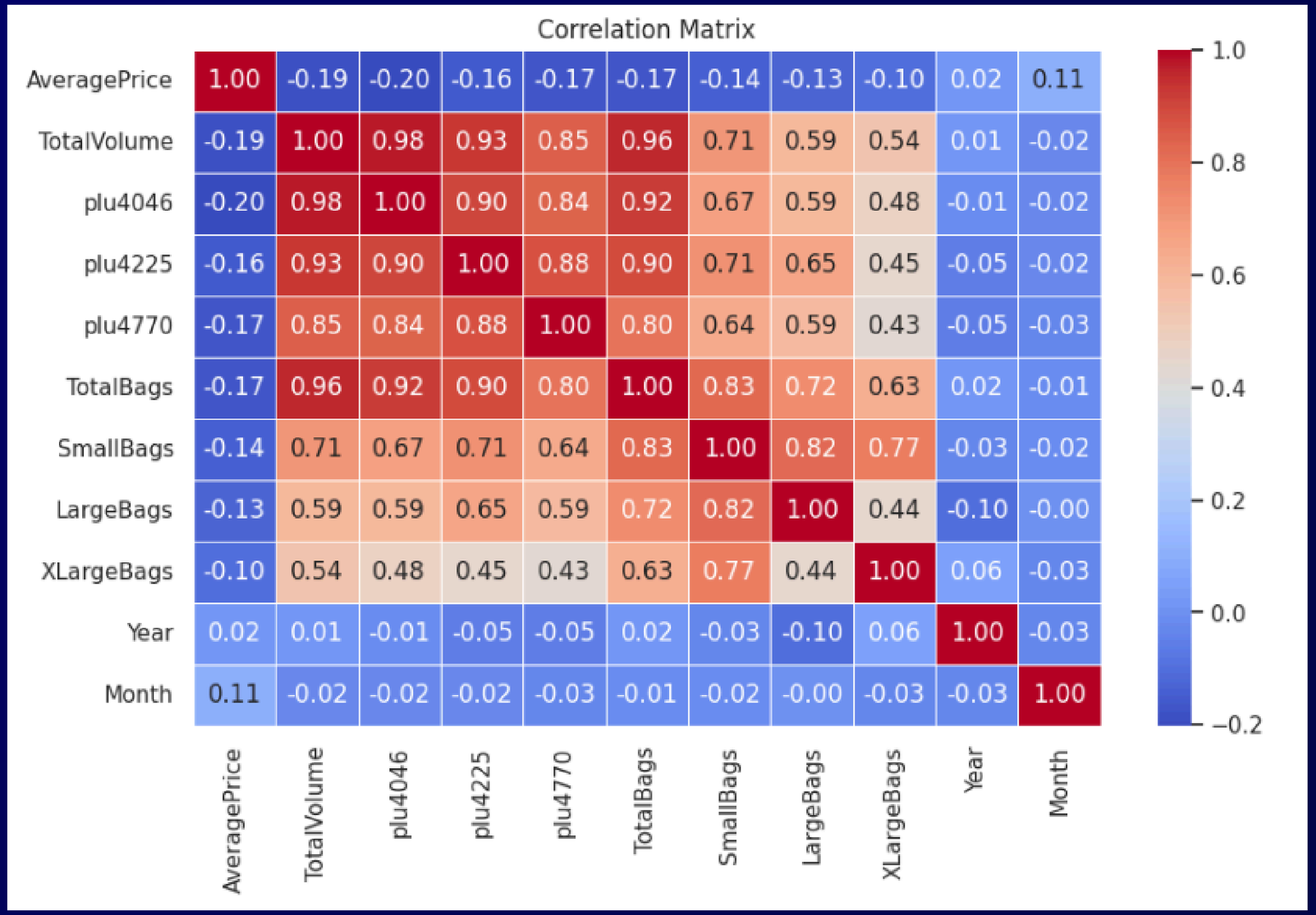
41025 rows × 9 columns

```
[25] y = df.iloc[:,1]
     y
```

	AveragePrice
0	1.220000
1	1.790000
2	1.000000
3	1.760000
4	1.080000

```
# Selecting relevant features
# we'll use: 'Year', 'Month', ' Volume', and 'Total Bags' as features
X = df[['Year', 'Month', 'TotalVolume', 'TotalBags']]
y = df['AveragePrice']
```

MODELLING : CORRELATION MATRIX



MODEL PERFORMANCE COMPARISON

```
Linear Regression - MAE: 0.31416447029050826, MSE: 0.14847069872444593, R2: 0.05141564527613007
Random Forest - MAE: 0.1570916628135929, MSE: 0.04742995198559569, R2: 0.6969684201302102
Gradient Boosting - MAE: 0.20616918345801874, MSE: 0.06940695478199625, R2: 0.5565566001853244
```

MODEL PERFORMANCE COMPARISON

- MAE (MEAN ABSOLUTE ERROR): MEASURES THE AVERAGE ABSOLUTE DIFFERENCES BETWEEN THE PREDICTED VALUES AND ACTUAL VALUES. A LOWER MAE IS BETTER.

- MSE (MEAN SQUARED ERROR): MEASURES THE AVERAGE OF THE SQUARED DIFFERENCES BETWEEN PREDICTED AND ACTUAL VALUES. A LOWER MSE IS BETTER, BUT IT PENALIZES LARGER ERRORS MORE HEAVILY.

- R^2 (R-SQUARED): MEASURES HOW WELL THE MODEL EXPLAINS THE VARIANCE IN THE TARGET VARIABLE. A HIGHER R^2 VALUE INDICATES A BETTER FIT, WITH 1 MEANING A PERFECT FIT AND 0 MEANING THE MODEL DOESN'T EXPLAIN ANY VARIANCE.

1. LINEAR REGRESSION MAE: 0.3142 — THIS INDICATES THE AVERAGE ERROR IN PRICE PREDICTION IS AROUND 0.31. MSE: 0.1485 — A RELATIVELY LOW MSE, BUT THE MODEL IS STILL UNDERPERFORMING COMPARED TO OTHERS. R^2: 0.0514 — THIS INDICATES THAT ONLY ABOUT 5.14% OF THE VARIANCE IN THE DATA IS EXPLAINED BY THE MODEL. THIS IS VERY LOW, SUGGESTING THAT THE LINEAR REGRESSION MODEL IS NOT A GOOD FIT FOR THIS DATA.

2. RANDOM FOREST REGRESSION MAE: 0.1571 — THE RANDOM FOREST MODEL HAS A MUCH LOWER MAE COMPARED TO LINEAR REGRESSION, WHICH INDICATES IT IS MAKING MORE ACCURATE PREDICTIONS. MSE: 0.0474 — THE MSE IS SIGNIFICANTLY LOWER THAN THAT OF LINEAR REGRESSION, SHOWING THAT THE RANDOM FOREST MODEL IS BETTER AT MINIMIZING LARGER PREDICTION ERRORS. R^2: 0.6970 — THIS IS A SUBSTANTIAL IMPROVEMENT OVER LINEAR REGRESSION, WITH ALMOST 70% OF THE VARIANCE IN THE TARGET VARIABLE BEING EXPLAINED. THE RANDOM FOREST MODEL DOES A MUCH BETTER JOB OF CAPTURING COMPLEX RELATIONSHIPS IN THE DATA.

3. GRADIENT BOOSTING REGRESSION MAE: 0.2062 — THE MAE FOR GRADIENT BOOSTING IS HIGHER THAN RANDOM FOREST BUT STILL BETTER THAN LINEAR REGRESSION. MSE: 0.0694 — THE MSE FOR GRADIENT BOOSTING IS ALSO HIGHER THAN RANDOM FOREST, BUT LOWER THAN LINEAR REGRESSION. R^2: 0.5566 — WHILE NOT AS HIGH AS RANDOM FOREST, THE R^2 VALUE IS STILL FAIRLY GOOD AND INDICATES THAT ABOUT 55.66% OF THE VARIANCE IS EXPLAINED BY THE GRADIENT BOOSTING MODEL.

CONCLUSION

- BEST MODEL (BY R^2 AND MSE): THE RANDOM FOREST MODEL IS CLEARLY THE BEST PERFORMER. IT HAS THE HIGHEST R^2 (0.6970) AND THE LOWEST MSE (0.0474), INDICATING THAT IT DOES THE BEST JOB OF BOTH FITTING THE DATA AND MINIMIZING PREDICTION ERRORS. GRADIENT BOOSTING PERFORMS BETTER THAN LINEAR REGRESSION IN TERMS OF MAE AND MSE, BUT IT DOESN'T OUTPERFORM RANDOM FOREST. LINEAR REGRESSION PERFORMS THE WORST ACROSS ALL METRICS.

- ITS LOW R^2 SUGGESTS IT CANNOT ADEQUATELY MODEL THE COMPLEX RELATIONSHIPS IN THE DATA. THIS IS TYPICAL FOR LINEAR MODELS WHEN THE DATA HAS NON-LINEAR RELATIONSHIPS. RECOMMENDATIONS: BASED ON THE RESULTS, RANDOM FOREST WOULD BE THE BEST MODEL TO DEPLOY FOR PREDICTING THE AVOCADO PRICE, AS IT PERFORMS THE BEST IN TERMS OF BOTH ERROR METRICS AND R^2.

- GRADIENT BOOSTING COULD ALSO BE A GOOD ALTERNATIVE, ESPECIALLY IF WE WERE TO FINE-TUNE THE MODEL WITH HYPERPARAMETER OPTIMIZATION (E.G., ADJUSTING N_ESTIMATORS, LEARNING_RATE, ETC.). LINEAR REGRESSION IS NOT RECOMMENDED UNLESS THE PROBLEM REQUIRES A SIMPLE, INTERPRETABLE MODEL. GIVEN THE LOW R^2, IT'S UNLIKELY TO CAPTURE ALL THE COMPLEXITY IN THE DATASET.

FINAL MODEL CONCLUSION

THE RANDOM FOREST REGRESSOR IS THE FINAL MODEL CHOSEN FOR PREDICTING AVOCADO PRICES DUE TO ITS SUPERIOR PERFORMANCE IN TERMS OF MAE, MSE, AND R^2. IT IS HIGHLY EFFECTIVE IN CAPTURING COMPLEX RELATIONSHIPS BETWEEN FEATURES, MAKING IT THE IDEAL MODEL FOR THIS PREDICTION TASK.

WHILE GRADIENT BOOSTING ALSO PERFORMED WELL, RANDOM FOREST PROVIDED A MORE ROBUST AND ACCURATE SOLUTION.

CONCLUSION AND FUTURE WORK

THIS PROJECT FOCUSED ON PREDICTING AVOCADO PRICES USING VARIOUS REGRESSION MODELS. WE EXPLORED MULTIPLE MACHINE LEARNING TECHNIQUES, INCLUDING LINEAR REGRESSION, RANDOM FOREST REGRESSOR, AND GRADIENT BOOSTING REGRESSOR. THE GOAL WAS TO IDENTIFY THE MODEL THAT BEST CAPTURED THE COMPLEXITY OF THE AVOCADO PRICE DATASET.

THE RESULTS SHOWED THAT:
LINEAR REGRESSION PERFORMED THE WORST ACROSS ALL METRICS, WITH A LOW R^2 VALUE OF 0.0514, INDICATING THAT IT WAS UNABLE TO EFFECTIVELY CAPTURE THE RELATIONSHIPS BETWEEN THE FEATURES AND THE TARGET VARIABLE.
GRADIENT BOOSTING SHOWED BETTER PERFORMANCE THAN LINEAR REGRESSION, WITH A R^2 VALUE OF 0.5566. HOWEVER, IT STILL FELL SHORT COMPARED TO RANDOM FOREST.

THE RANDOM FOREST REGRESSOR EMERGED AS THE BEST-PERFORMING MODEL. IT EXPLAINED ABOUT 69.7% OF THE VARIANCE IN AVOCADO PRICES ($R^2 = 0.6970$), AND IT EXHIBITED LOWER MAE (0.1571) AND MSE (0.0474) COMPARED TO THE OTHER MODELS.

KEY INSIGHTS GAINED:

RANDOM FOREST WAS PARTICULARLY EFFECTIVE DUE TO ITS NON-LINEAR NATURE, ABILITY TO HANDLE COMPLEX INTERACTIONS BETWEEN FEATURES, AND ROBUSTNESS AGAINST OVERFITTING. THE DATASET EXHIBITED NON-LINEAR RELATIONSHIPS BETWEEN FEATURES (E.G., VOLUME, REGION, AND TIME), WHICH MADE LINEAR MODELS LIKE LINEAR REGRESSION UNSUITABLE. THE R^2 VALUE OF 69.7% FOR RANDOM FOREST INDICATES A GOOD PREDICTIVE PERFORMANCE, THOUGH THERE'S STILL ROOM FOR IMPROVEMENT IN EXPLAINING MORE OF THE VARIANCE IN AVOCADO PRICES.

THANK YOU!