Avocado Average Price Prediction

Moh. Ridwan

1. Summary

Business Background

PT X is a company engaged in selling avocados in America.

Problems Statements

How to determine the average avocado price?

Objective

Create a model that can predict the average price of avocado

Proposed Solutions

PT X can make a regression prediction model to predict the average avocado price

Result:

The regression model using the XGBRegressor algorithm is more suitable in this case. The model has a coefficient of determination (R-square) of 88.27% with an RMSE Test of 0.138 and an RMSE Train of 0.106.

Business Benefit

- PT X can predict the average price of avocado so that it can determine the average price that is effective in selling avocados (the price is not too cheap or too expensive)

2. Data Overview

Data

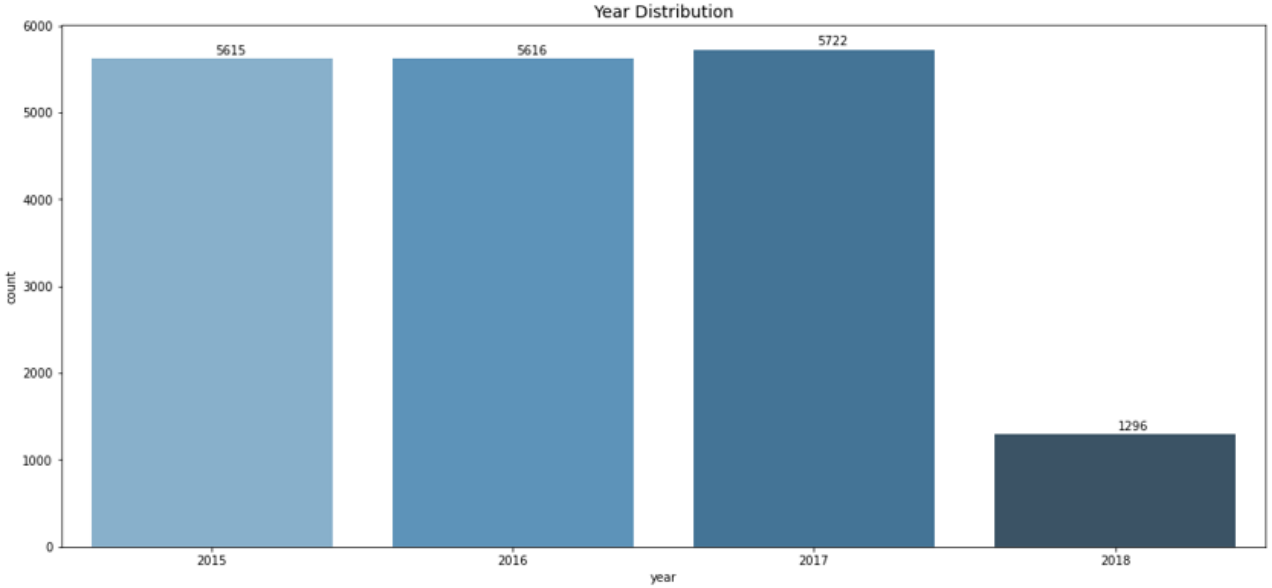
	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18244	7	2018-02-04	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	organic	2018	WestTexNewMexico
18245	8	2018-01-28	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	organic	2018	WestTexNewMexico
18246	9	2018-01-21	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	organic	2018	WestTexNewMexico
18247	10	2018-01-14	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	organic	2018	WestTexNewMexico
18248	11	2018-01-07	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	organic	2018	WestTexNewMexico

18249 rows × 14 columns

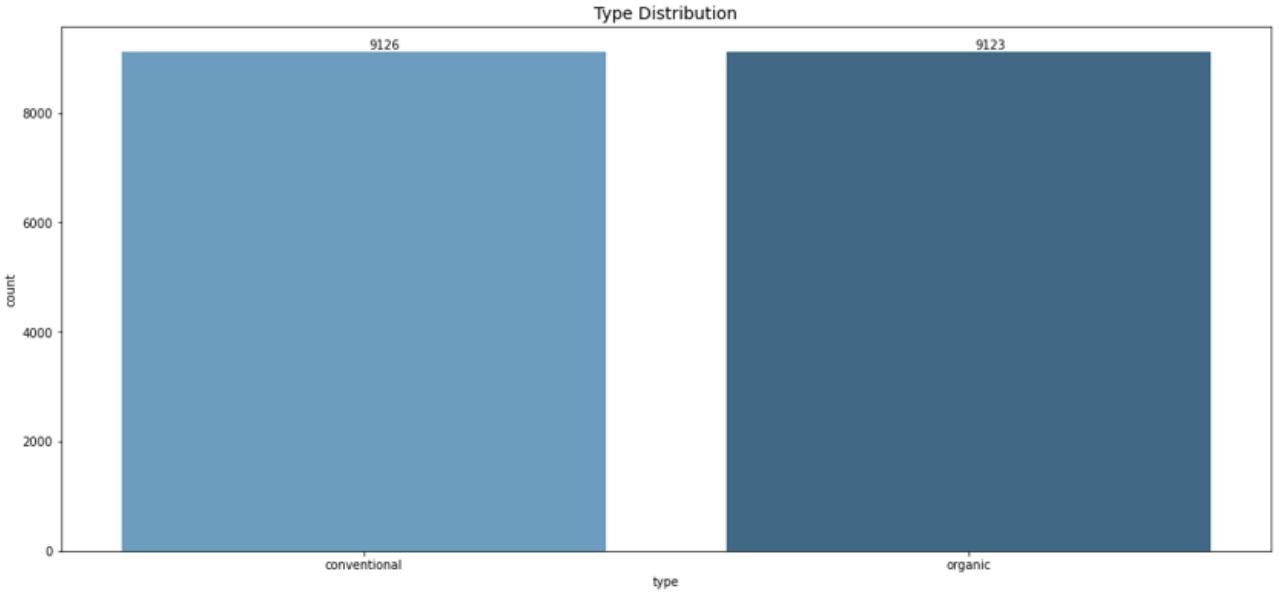
Information about data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18249 entries, 0 to 18248
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    18249 non-null  int64
 1   Date          18249 non-null  object
 2   AveragePrice  18249 non-null  float64
 3   Total Volume  18249 non-null  float64
 4   4046          18249 non-null  float64
 5   4225          18249 non-null  float64
 6   4770          18249 non-null  float64
 7   Total Bags    18249 non-null  float64
 8   Small Bags    18249 non-null  float64
 9   Large Bags    18249 non-null  float64
 10  XLarge Bags   18249 non-null  float64
 11  type          18249 non-null  object
 12  year          18249 non-null  int64
 13  region        18249 non-null  object
dtypes: float64(9), int64(2), object(3)
memory usage: 1.9+ MB
```

2. Exploratory Data Analysis

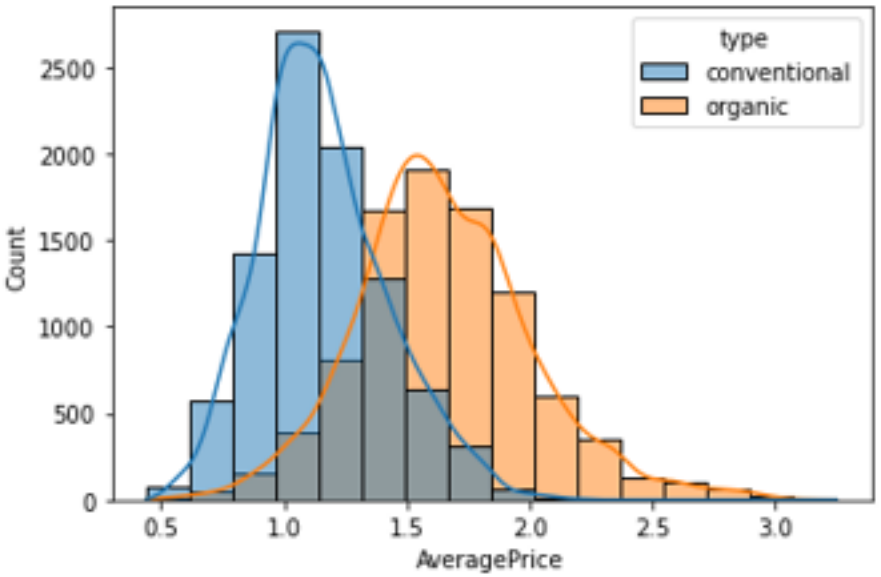


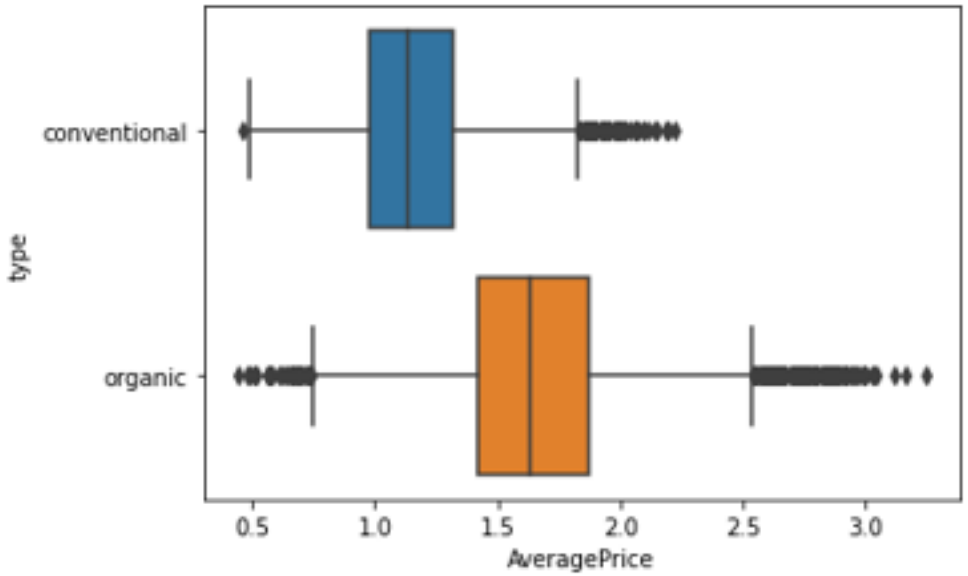
The data located in the dataset consists of 4 years of data, namely 2015, 2016, 2017, and 2018



There are two types of avocados sold, namely organic and conventional avocados. The number of the two types of avocado in the dataset is balanced

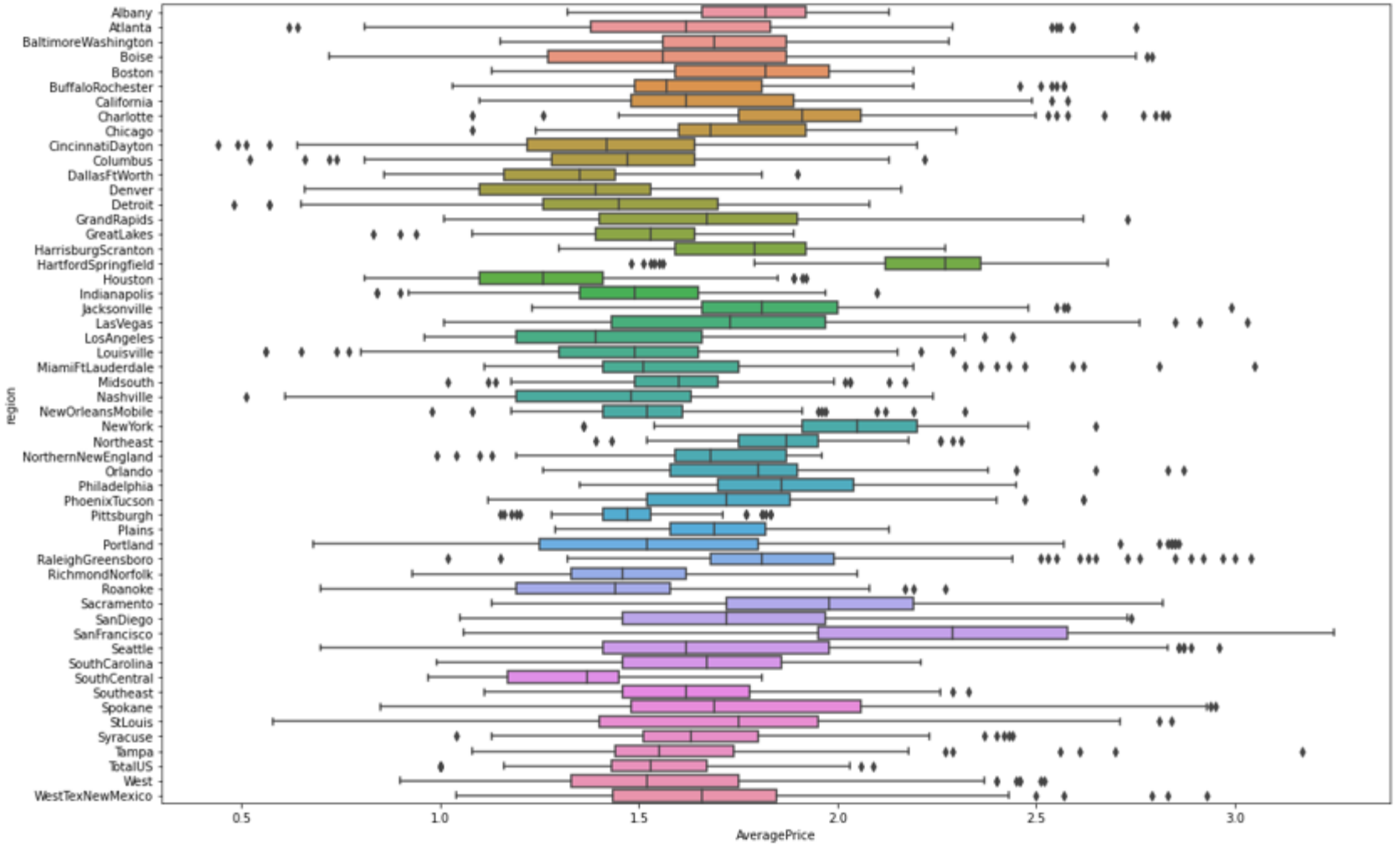
2. Exploratory Data Analysis





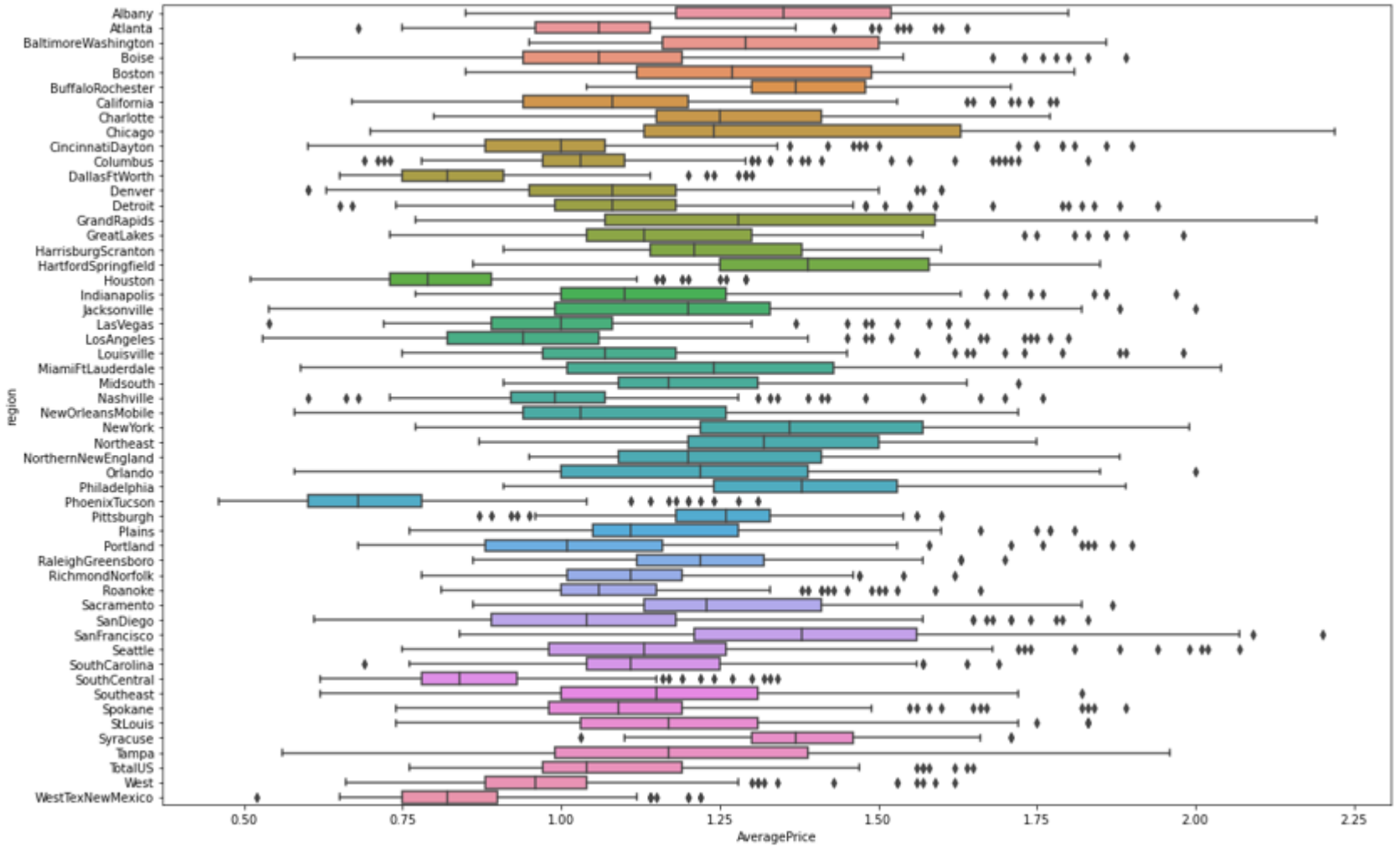
The largest average of the average avocado price is on organic avocado, this is in accordance with the price of organic avocado which are more expensive than conventional avocado

2. Exploratory Data Analysis



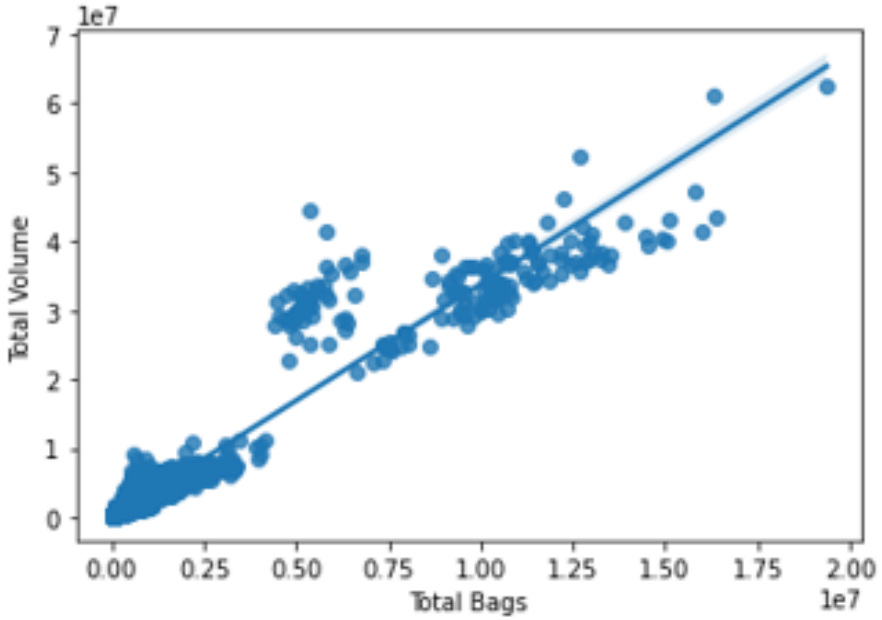
The largest average of the average organic avocado prices is in Hartford Springfield and San Francisco

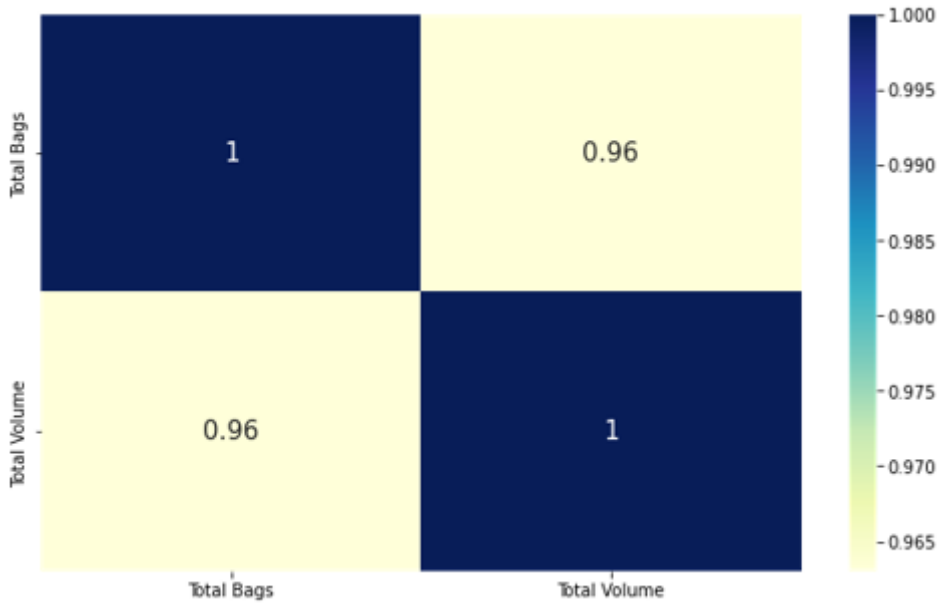
2. Exploratory Data Analysis



The largest average of the average confentional avocado prices is in Hartford Albany, Buffalo Rochester, Hartford Springfield, New York, Philadelphia, San Francisco, and Syracus

2. Exploratory Data Analysis





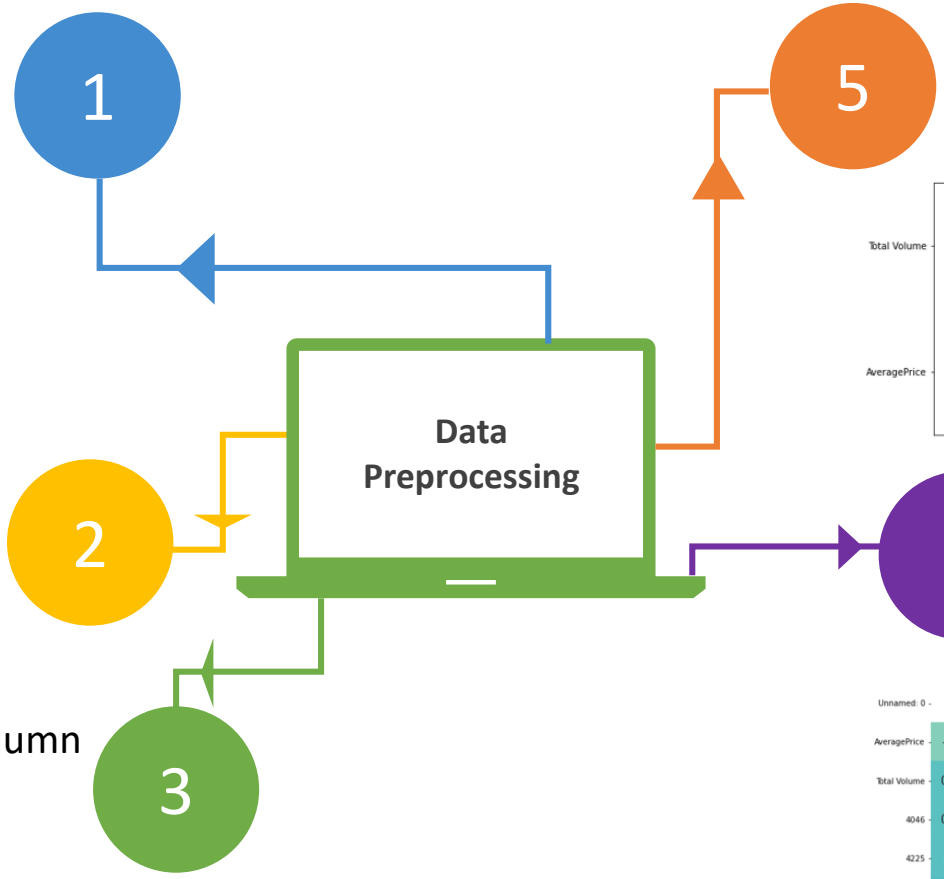
If the total number of avocados sold is increasing, the number of bags needed will also increase

3. Data Preprocessing

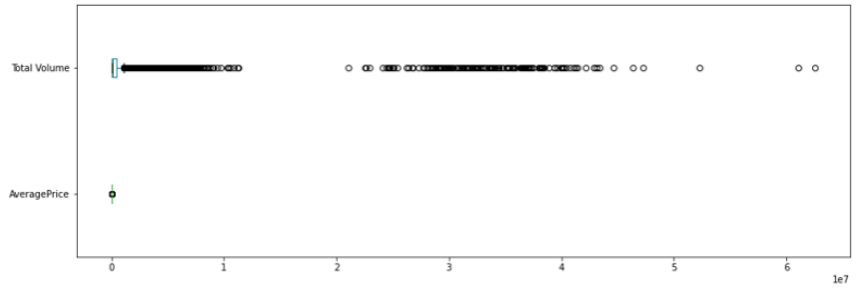
Check number of null value

```
Unnamed: 0      0
Date            0
AveragePrice    0
Total Volume    0
4046            0
4225            0
4770            0
Total Bags      0
Small Bags      0
Large Bags      0
XLarge Bags     0
type            0
year            0
region          0
dtype: int64
```

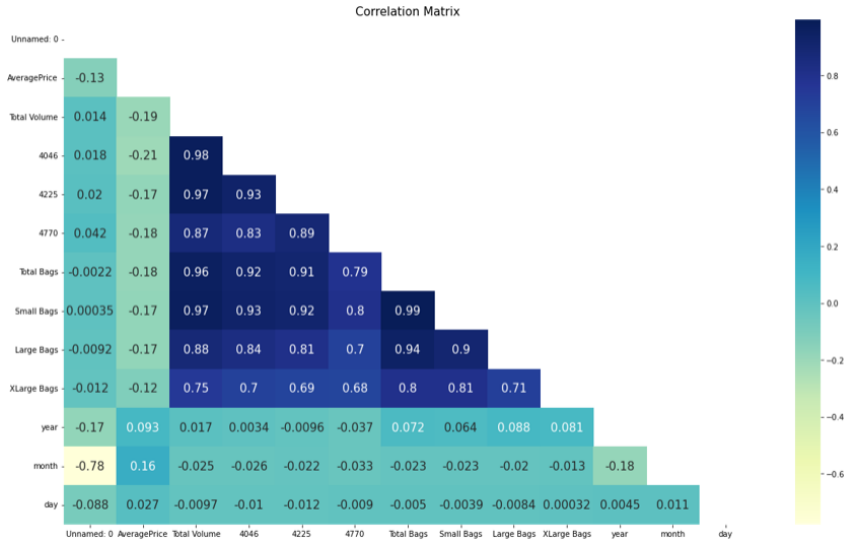
Check for duplicate values



Check outliers using boxplot



Check correlation and remove highly correlated features



Ceck data type and change type of Date column and make new column (day and month)

```
Unnamed: 0       int64
Date            object
AveragePrice   float64
Total Volume   float64
4046           float64
4225           float64
4770           float64
Total Bags     float64
Small Bags     float64
Large Bags     float64
XLarge Bags    float64
type            object
year             int64
region          object
dtype: object
```

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region	month	day
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany	12	27
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany	12	20
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany	12	13
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany	12	6
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany	11	29

4. Modeling and Evaluation

Scaling continous column

Define target and features

Split the data (train and test)

Linear Regression

R-square = 58.05 %
RMSE Train = 0.2571788700500718
RMSE Test = 0.26139793632303066

Random Forest Regressor

R-square = 89.98 %
RMSE Train = 0.04705813951291984 4
RMSE Test = 0.1277425504891751

XGBRegressor

R-square = 88.27 %
RMSE Train = 0.10597661077104431
RMSE Test = 0.13822386697258787

Gradreint Boosting Regressor

R-square = 68.80 %
RMSE Train = 0.21949502386361477
RMSE Test = 0.2254194962002768

In this case, the XGBRegressor model is a fit model compared to the others

	AveragePrice	AveragePrice_Prediction
9181	1.48	1.722112
1013	1.05	1.064347
14625	1.27	1.327191
15234	2.15	2.048163
18247	1.93	1.563677
...	...	...
10657	2.07	1.900100
17490	1.41	1.393710
6634	1.76	1.762157
10947	1.69	1.702125
13532	1.61	1.564079

3650 rows × 2 columns

THANK YOU