Price Prediction and Machine Learning Analysis

Avocado

Group：Yi Lu, Sophia Zhang, Stella Zhou

CONTENTS

Economic Impact

- Avocados are a coveted commodity, especially for millennials
- Avocado producers, distributers, and consumers all rely on fair and accurate price
- Can optimize the supply chain based on demand and revenue predictions

Social Impact

- Resource allocation for farms and supply chain
- Important export for low income, unstable countries
- People’s jobs and likelihoods in critical regions depend on the supply and demand

Dataset Overview

Variables
Description

1 Variables

Avocado Price

Average price, $USD

Avocado Type

Organic vs Conventional

Total Volume

Total number of avocados sold

Bag Type

Volume of avocadoes sold in Small, Large, XL bags

PLU Code

Avocados sold per code '4046', '4225', '4770’

Region

City or Region in the US

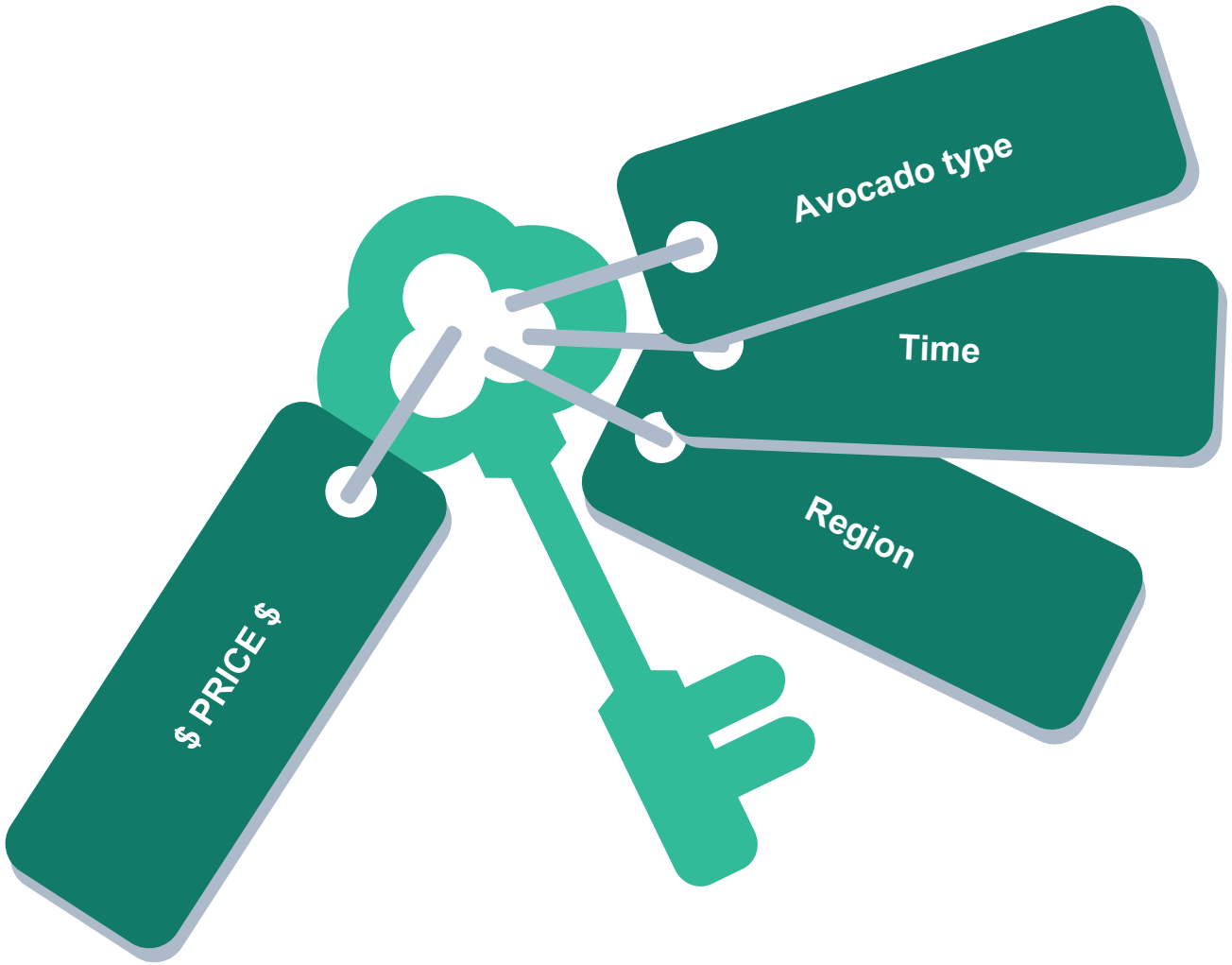
Date

Date of observation

Total Bags Sold

Total number of avocados sold in bags

1 Key Focus



Price Prediction

Which variables predominantly impact price and by how much?

Avocado Type

Which avocado types are the most prominent, are there differences to distinguish the types?

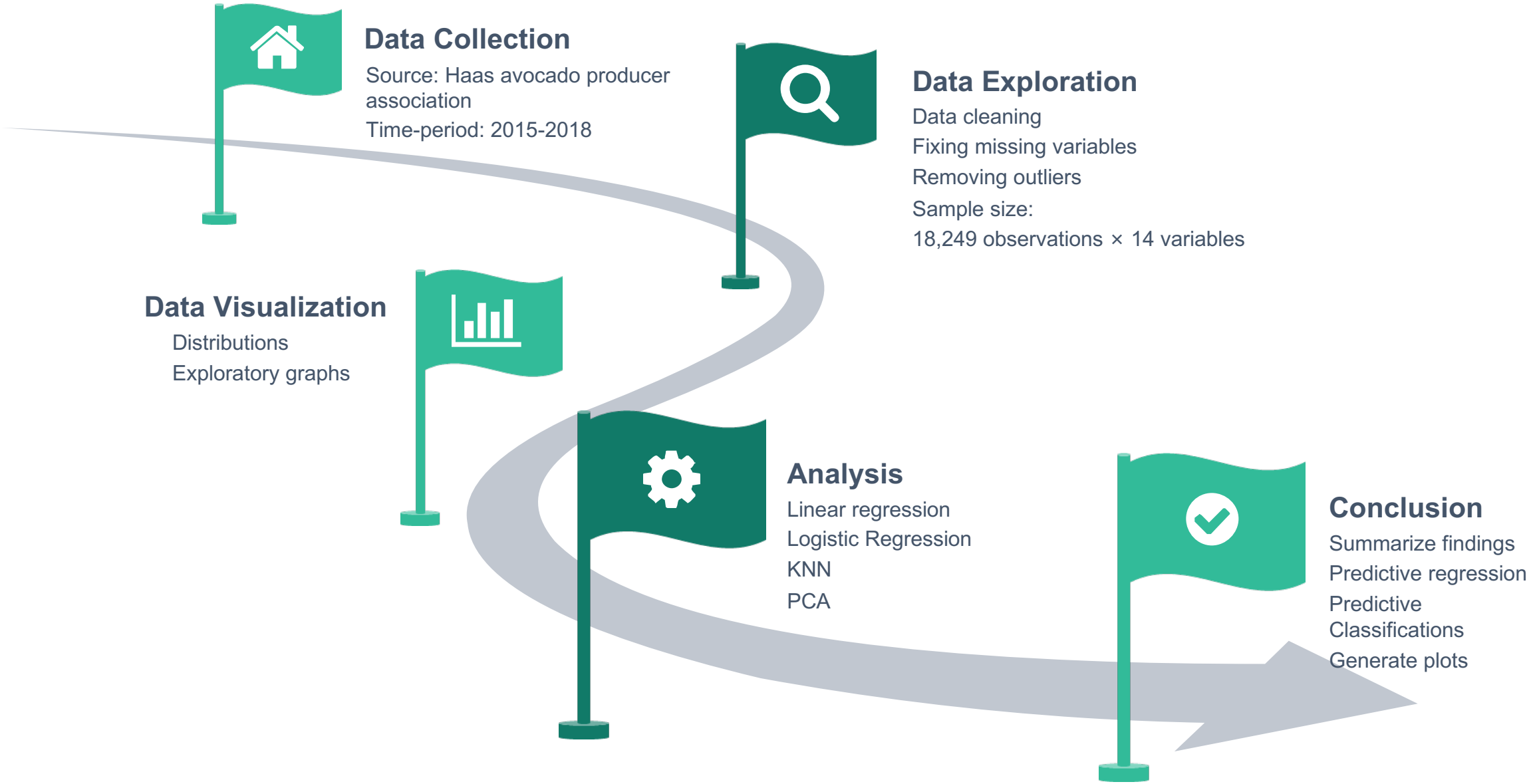
Time

Does the price of avocados change seasonally or by year?

Region

Are there difference in price and consumption in different regions of the US?

1 Methodology



Data Collection

Source: Haas avocado producer association

Time-period: 2015-2018

Data Exploration

Data cleaning

Fixing missing variables

Removing outliers

Sample size:

18,249 observations × 14 variables

Data Visualization

Distributions

Exploratory graphs

Analysis

Linear regression

Logistic Regression

KNN

PCA

Conclusion

Summarize findings

Predictive regression

Predictive Classifications

Generate plots

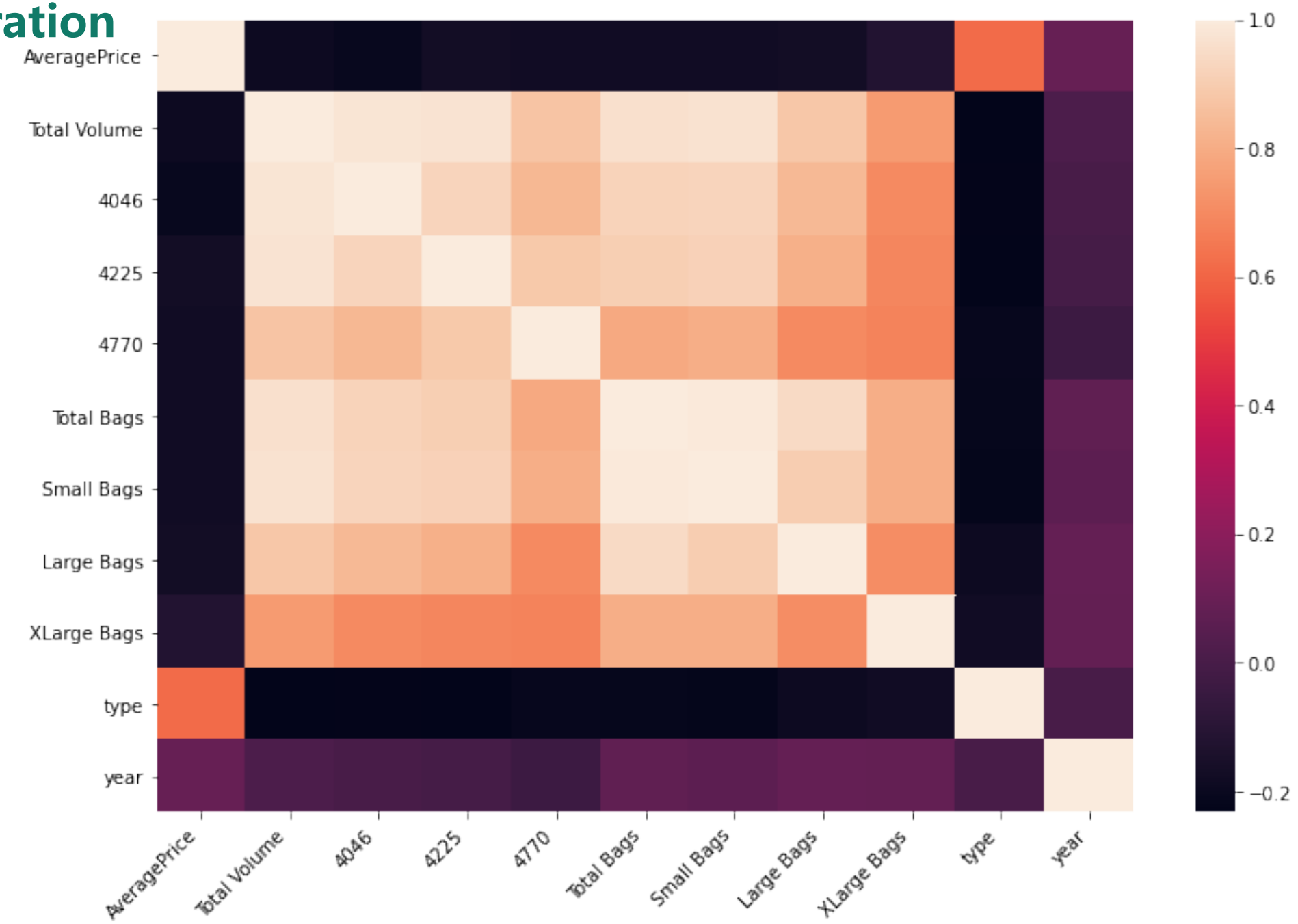
Data Visualization

Exploration &
Trends

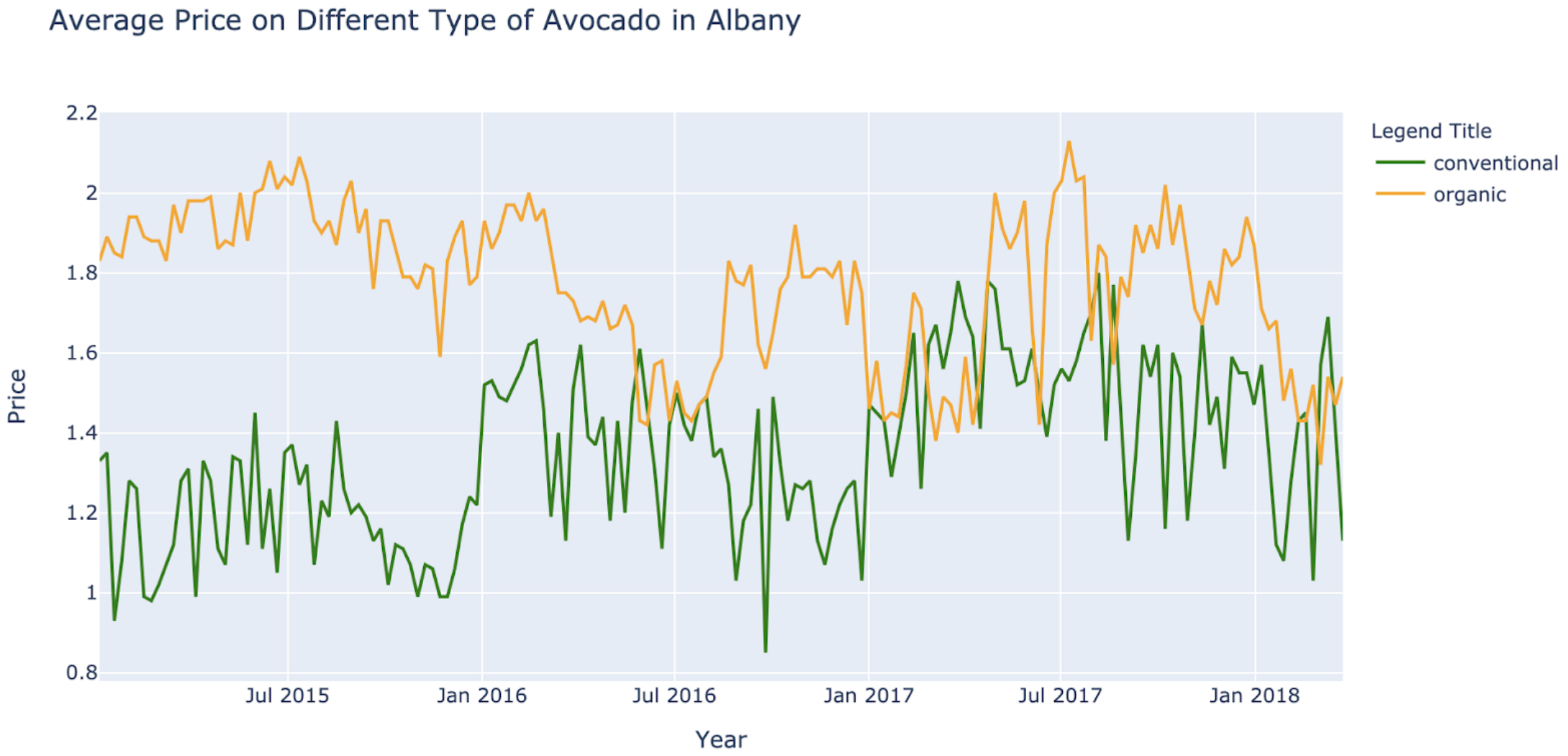
2 Data Exploration

Average price is generally **negatively correlated or no correlation** to other variables.

Price is only strongly **positively correlated** with the **type of avocado**, organic or conventional

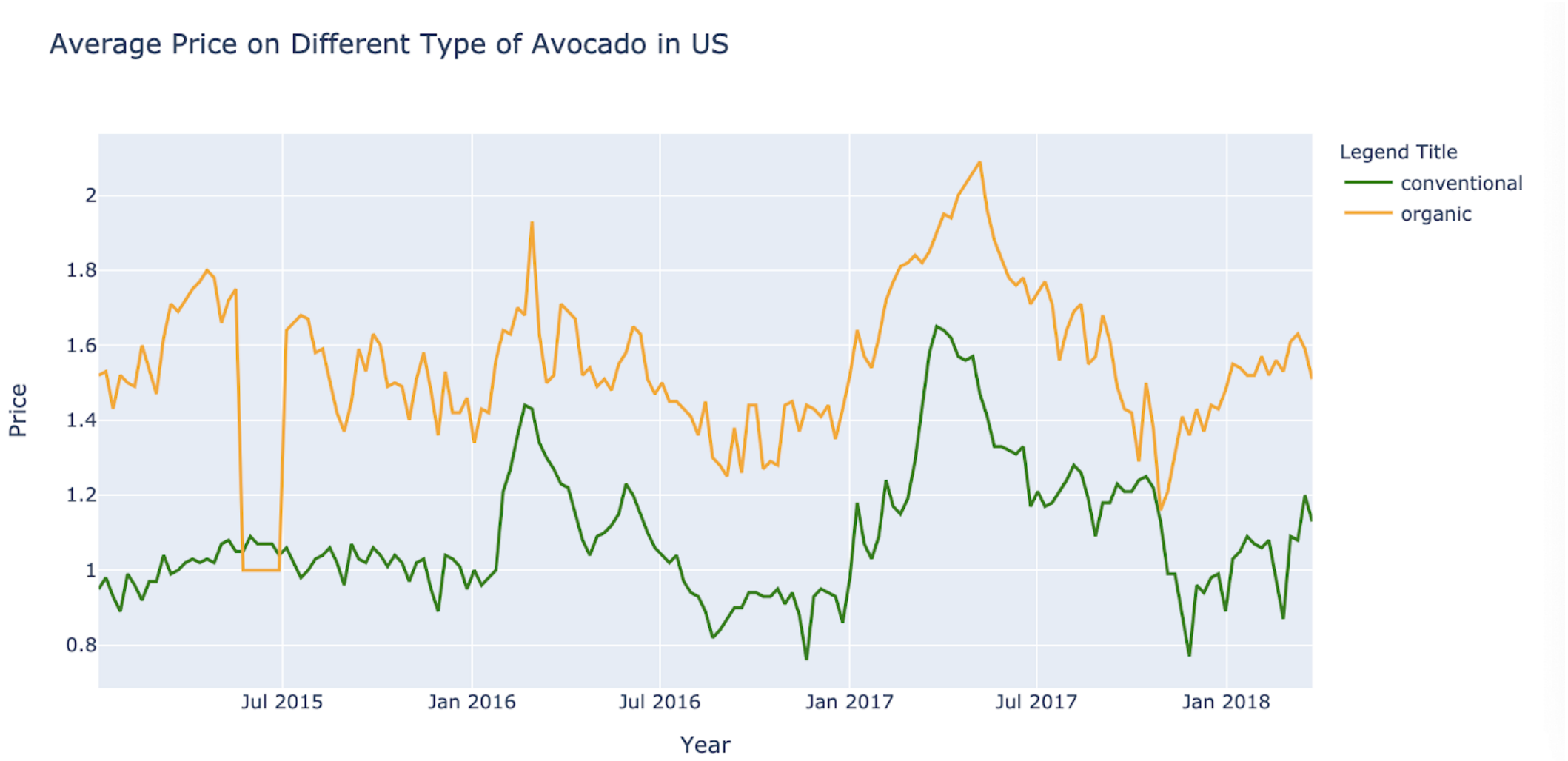


2 Data Exploration



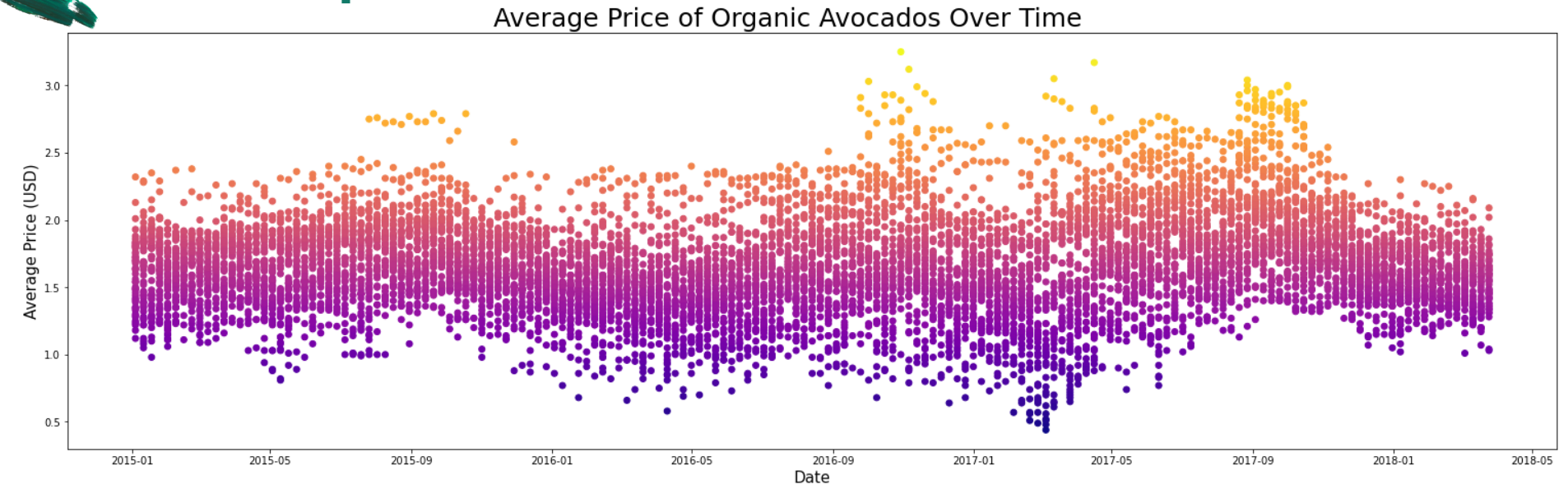
Over time, price in Albany between the different types of avocados are converging. Organic avocados are getting cheaper over time, and conventional avocados are becoming more expensive over time.

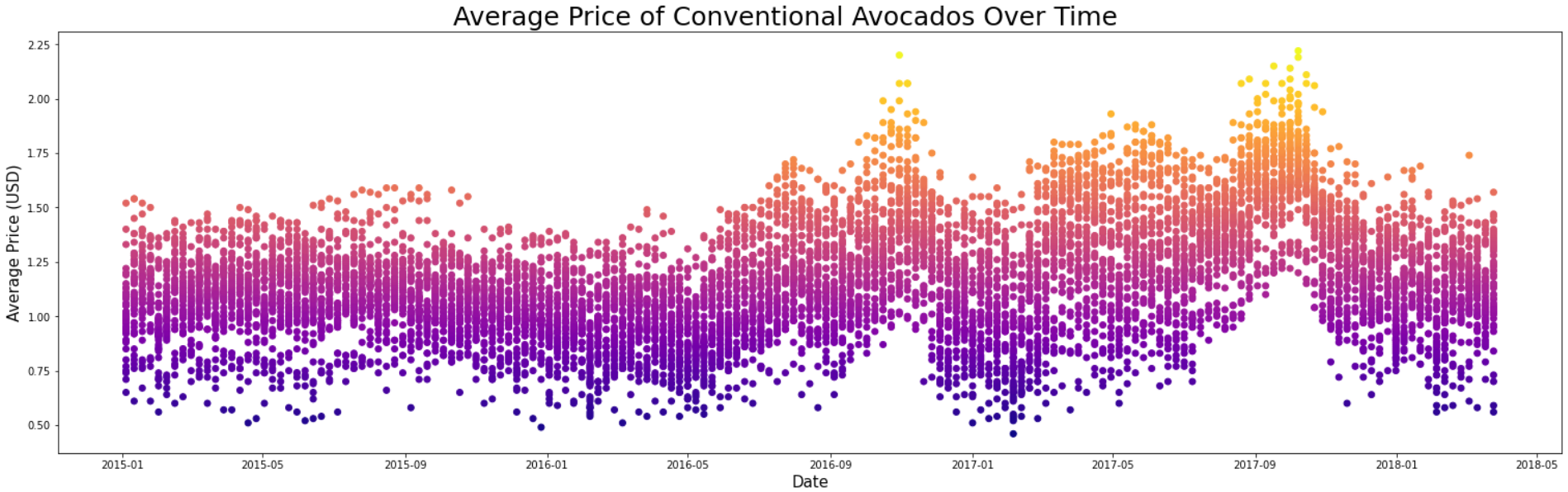
2 Data Exploration



Over time, price in the US between the different types of avocados show a cyclic trend. Both organic and conventional avocados show a decrease in price in the summer months, and increases in the winter months.

2 Data Exploration





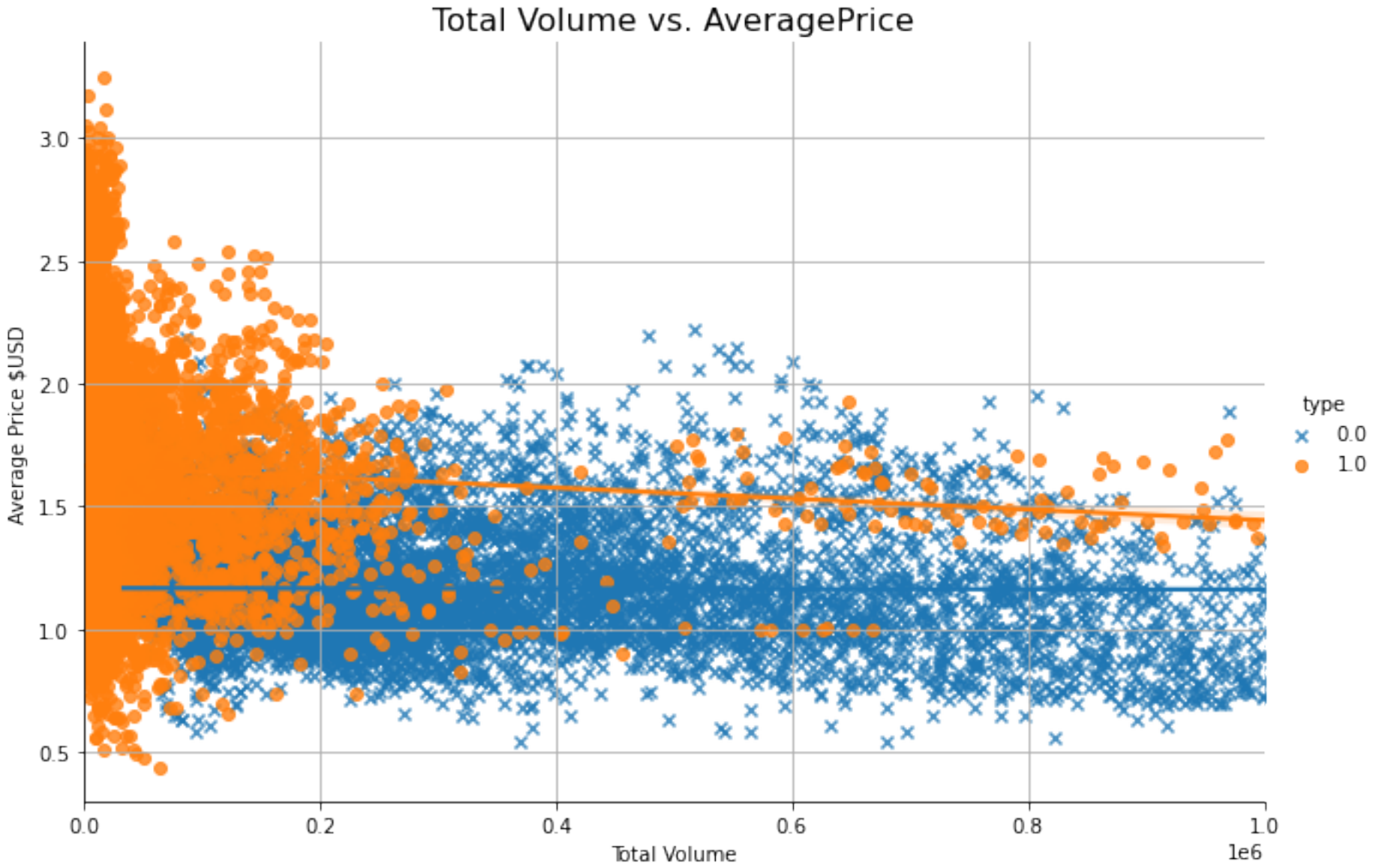
Organic avocados show more steady price trend compared to conventional avocados, which show sharp decrease and increases in price in certain periods from 2016-2018.

Every point on this graph represents the average price of a different region at a specific date in time. Region variations are greatest in the fall.

2 Data Exploration

Average **price for conventional** avocados does **not decrease** relative to **total volume**

Price decreases as volume increases for **organic** avocados as expected in a normal supply-demand relationship. Indicating a possible **scarcity** condition.



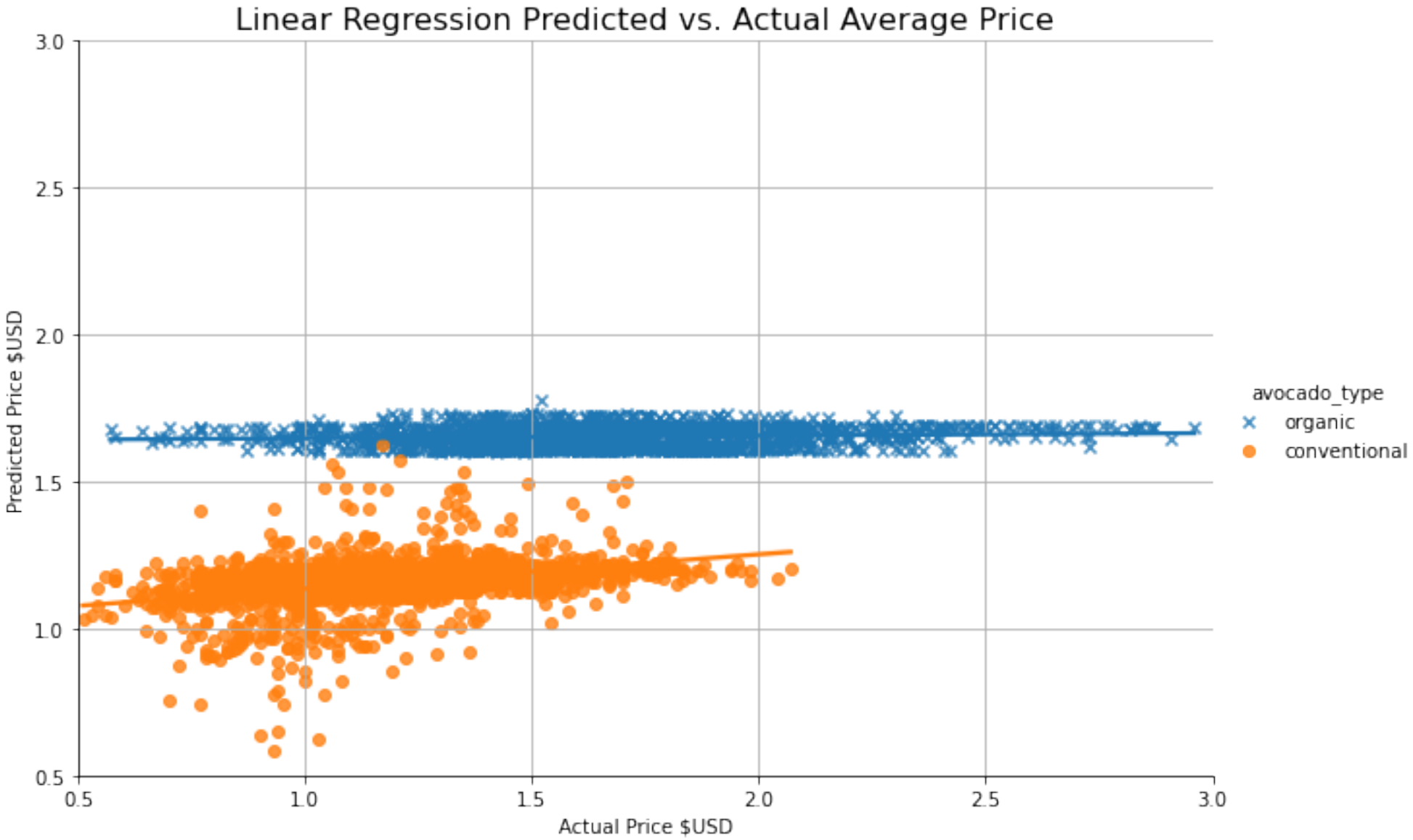
Regression

3 Linear Regression

Predictions for Average **price for conventional** avocados is typically lower than actual price, especially at higher actual prices.

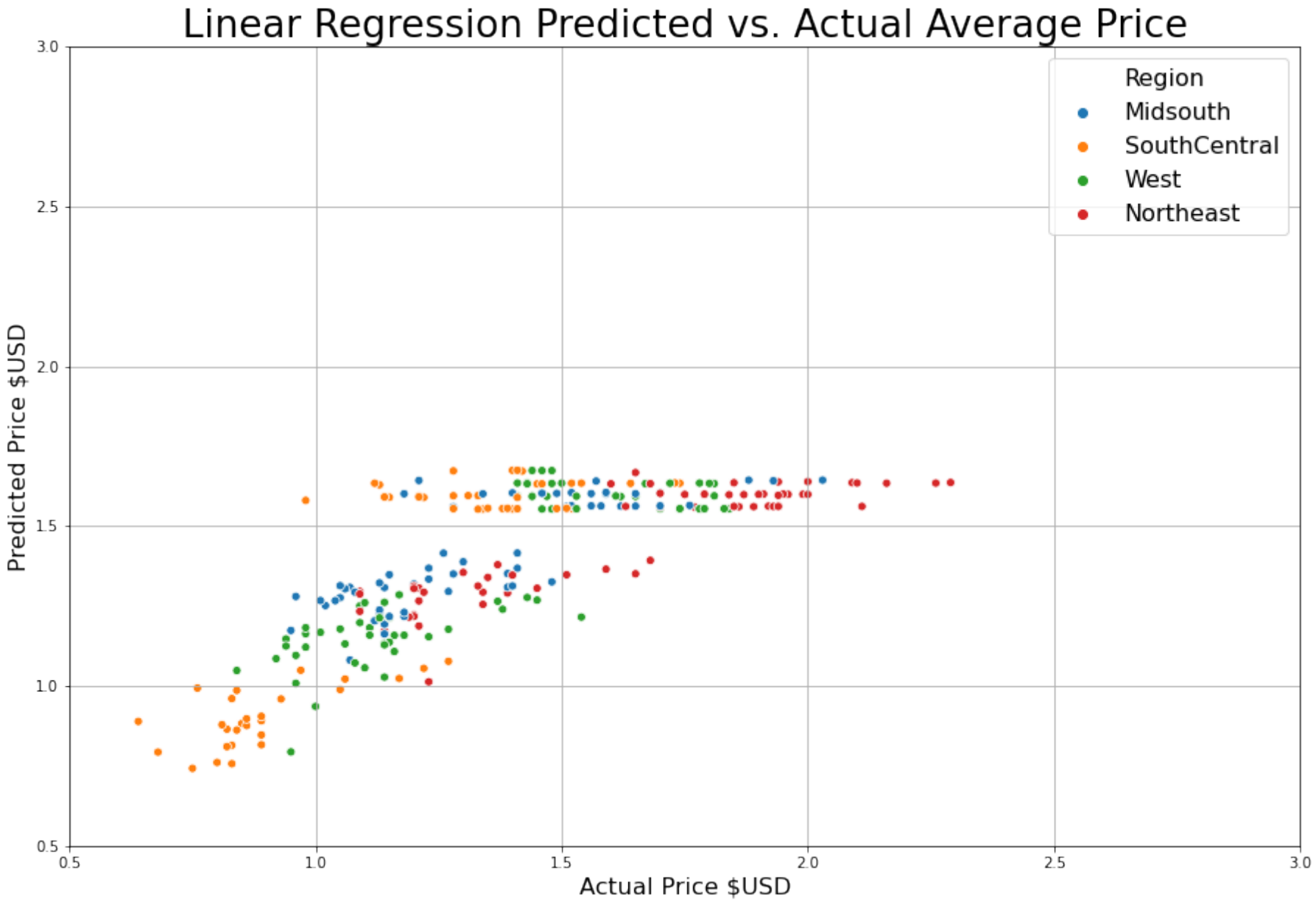
Predictions for **organic** avocados is not very good, always predicts the same values.

Accuracy on test data = 41%



3 Linear Regression with Region

Predictions for Average price for **Northeast is highest.**
Predictions for **South Central** is **lowest**.

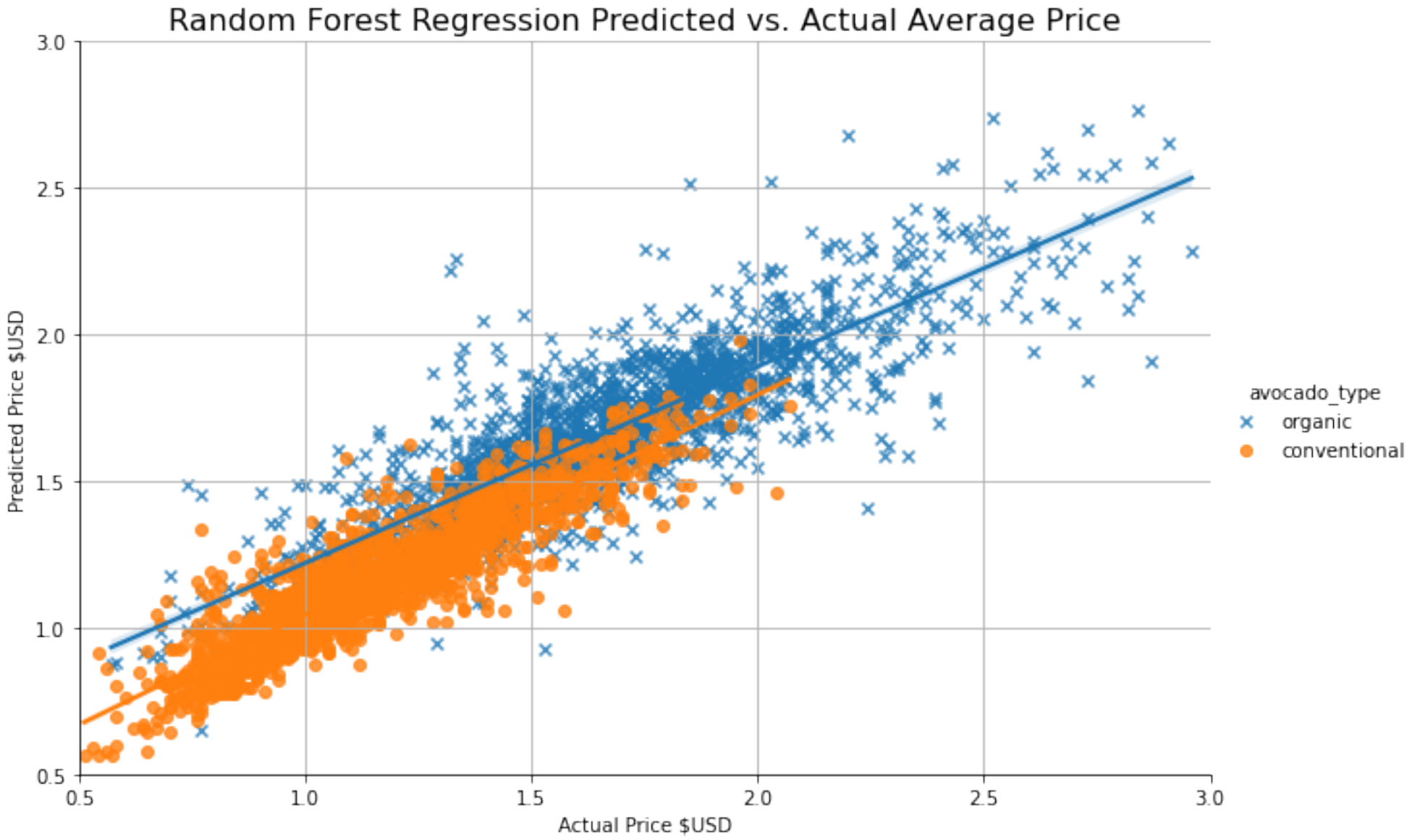


3 Random Forest Regression

Predictions using random forest regression is much **better than linear regression**

Accuracy on test data = 85%

Increased 107% from linear regression

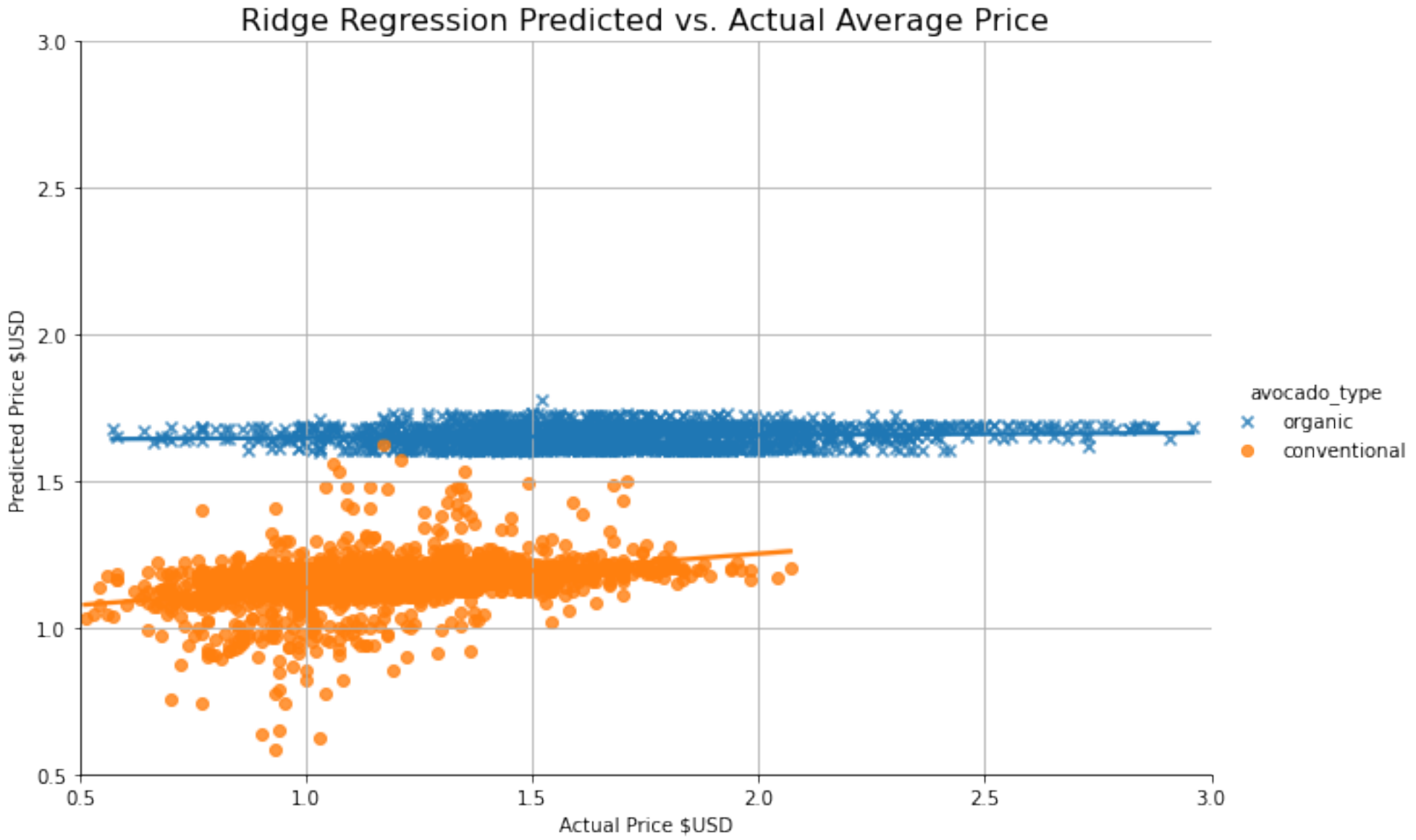


3 Ridge Regression

Predictions using ridge regression is **similar to linear regression**

Accuracy on test data = 41%

No change from linear regression

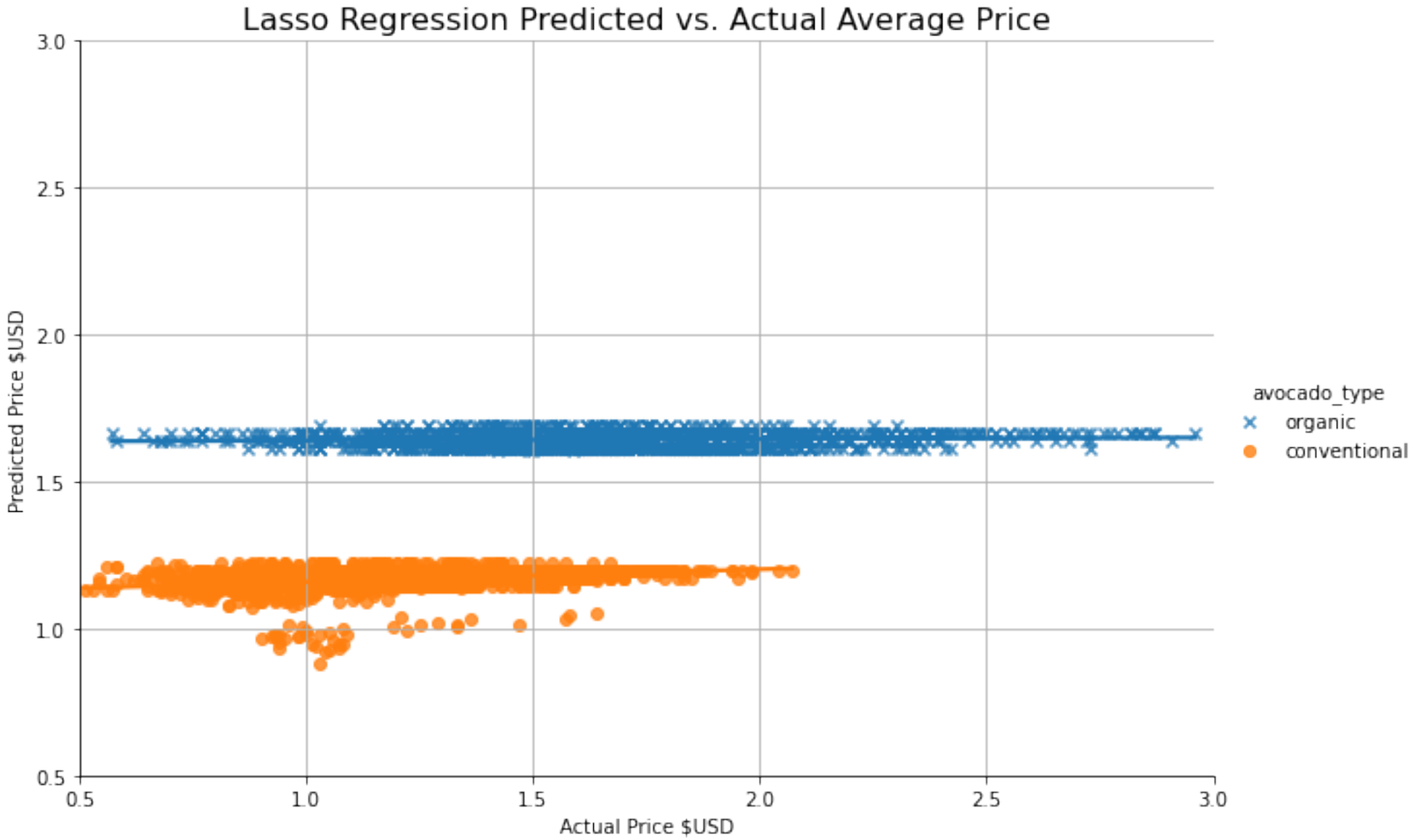


3 Lasso Regression

Predictions using lasso regression is **worse than linear regression**

Accuracy on test data = 40%

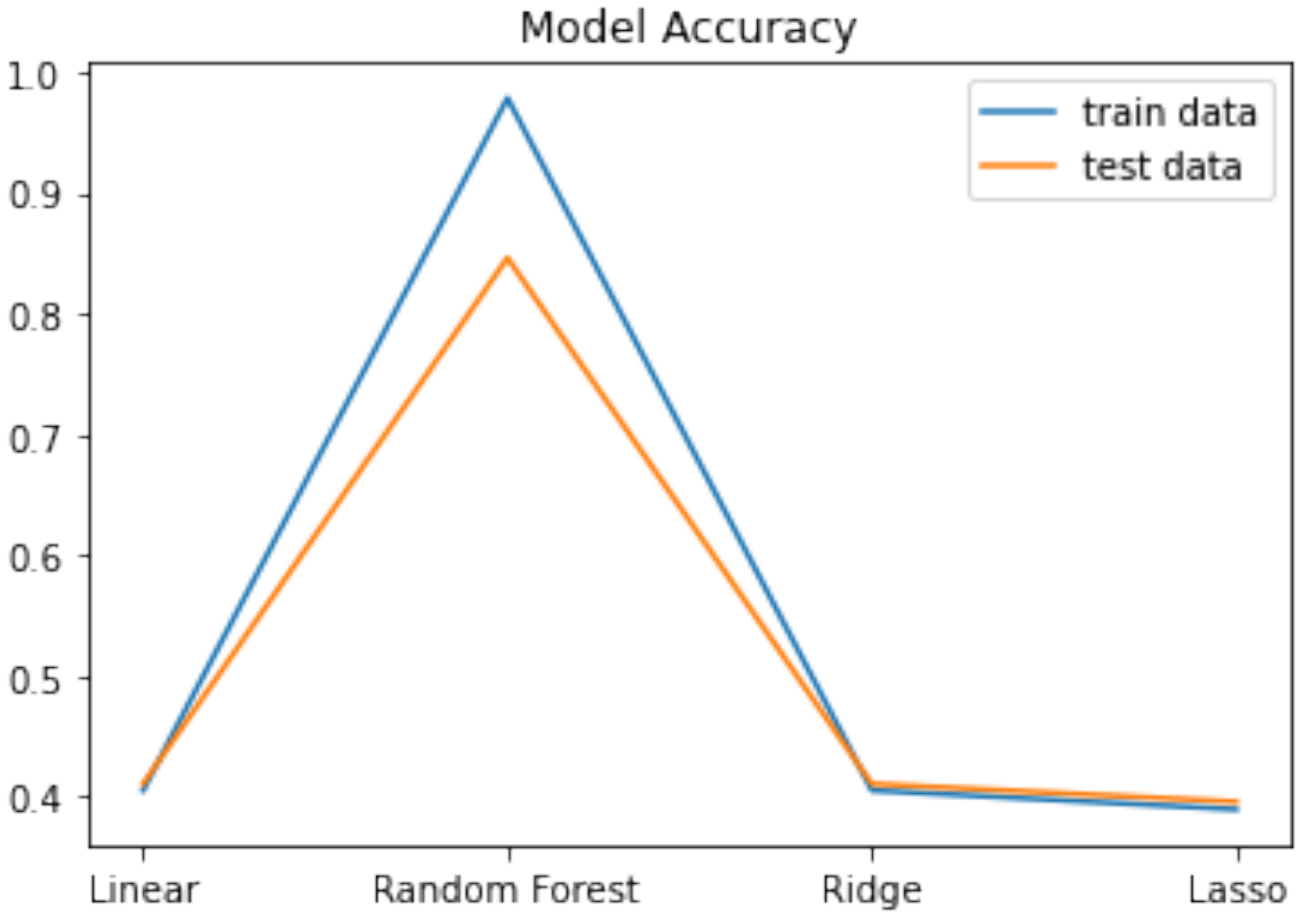
Decreased 1% from linear regression



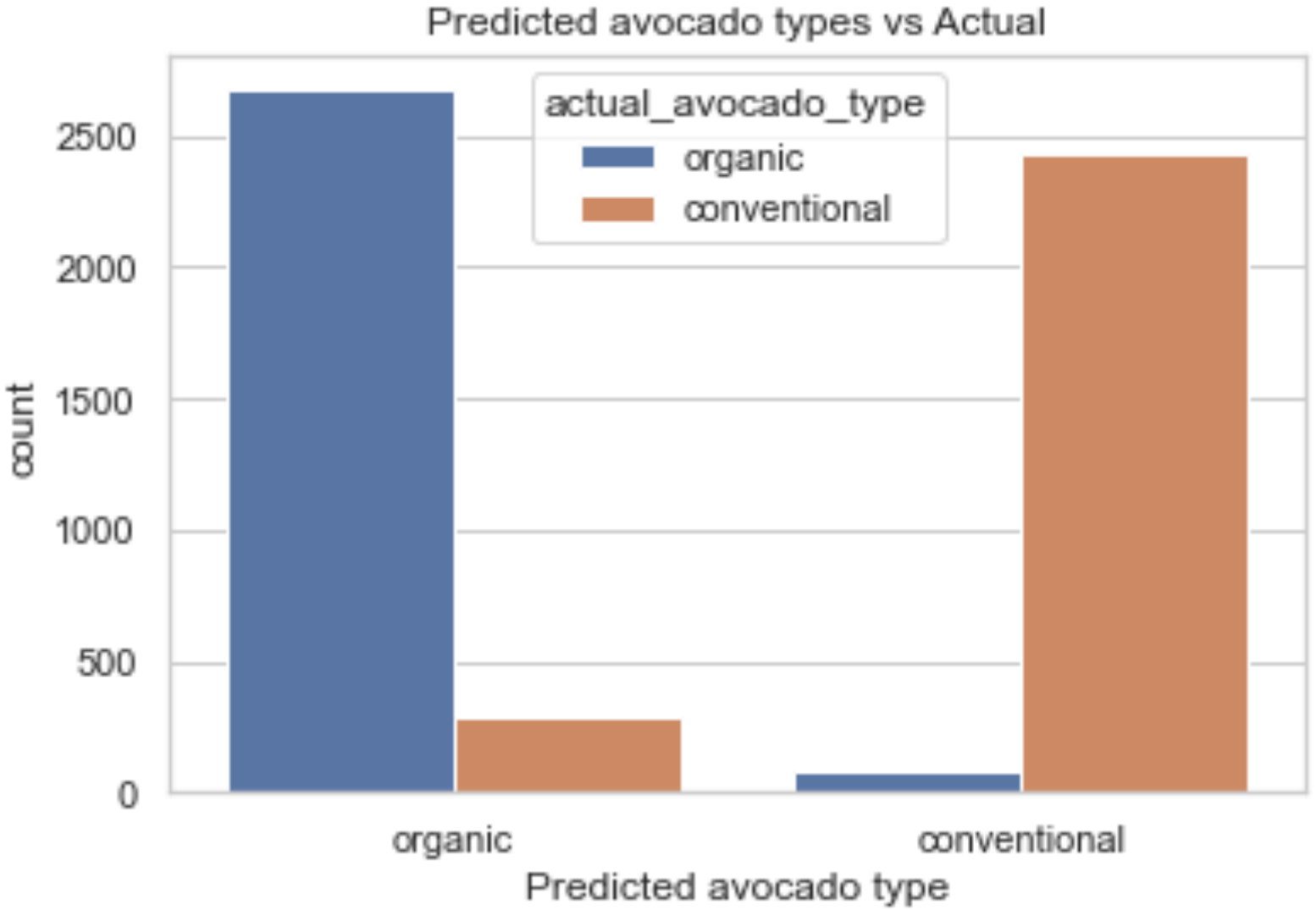
3 Regression Comparison

Best regression for price prediction is Random Forest

All others show similar accuracy

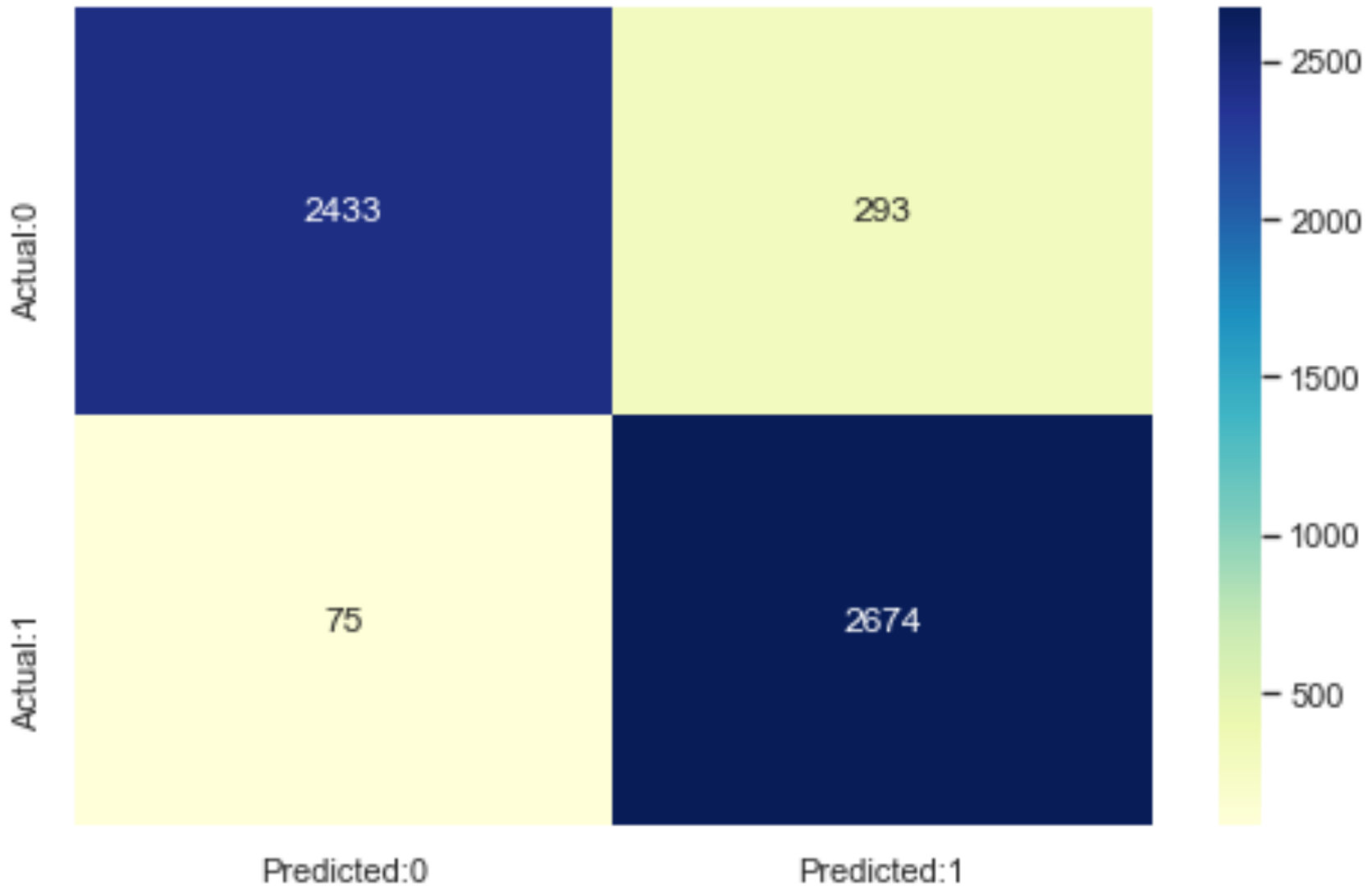


3 Logistic Regression



Prediction error is higher for organic avocados than conventional

3 Logistic Regression



Prediction of Avocado Type using Average Price, PLU code, number of bags sold, and year:

The accuracy of the model = 93%

Specificity or True Negative Rate = 89 %

Sensitivity or True Positive Rate = 97%

K-nearest Neighbor

4 K-nearest Neighbour

Type

KNN on Avocado Type

Month

KNN on Month

Year

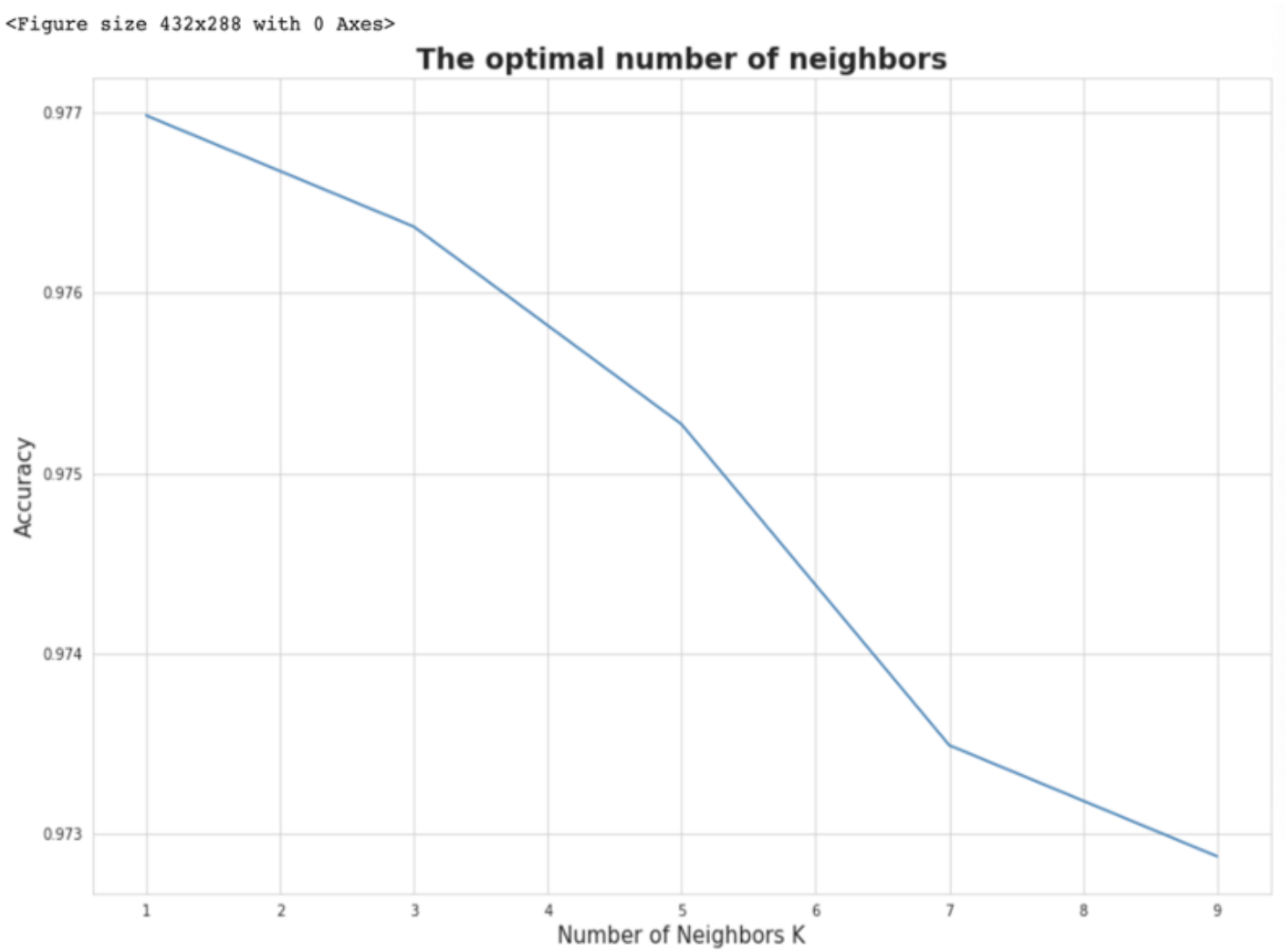
KNN on Year

4 K-nearest Neighbour - Avocado Type

Use KNN to cluster types of Avocado using Average Price, PLU code and number of bags sold

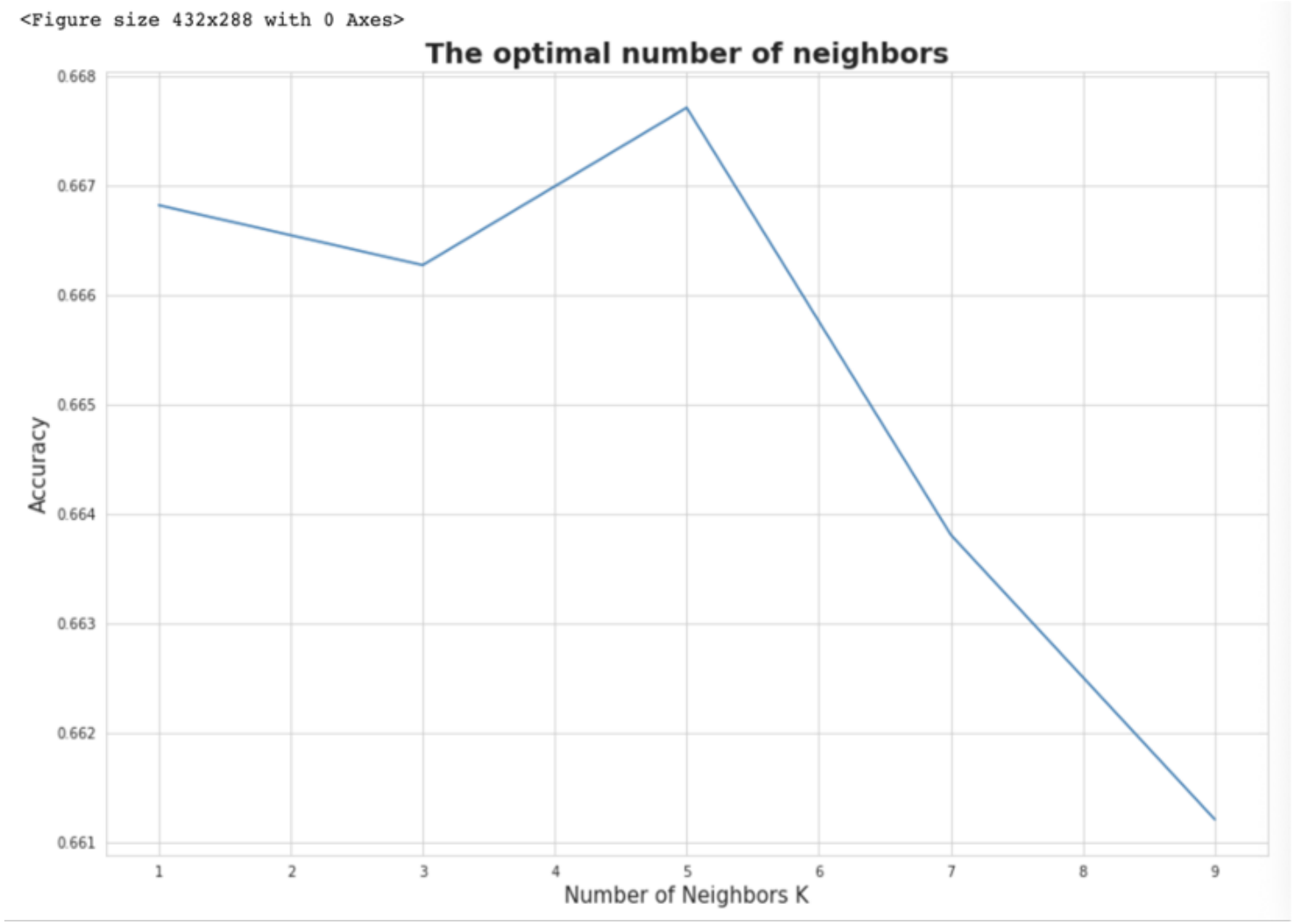
Best accuracy of our model is equal to 97.70% with the number of neighbors being 1

Confusion Matrix:
[[1740 35]
 [51 1824]]



4 K-nearest Neighbour - Year

<Figure size 432x288 with 0 Axes>



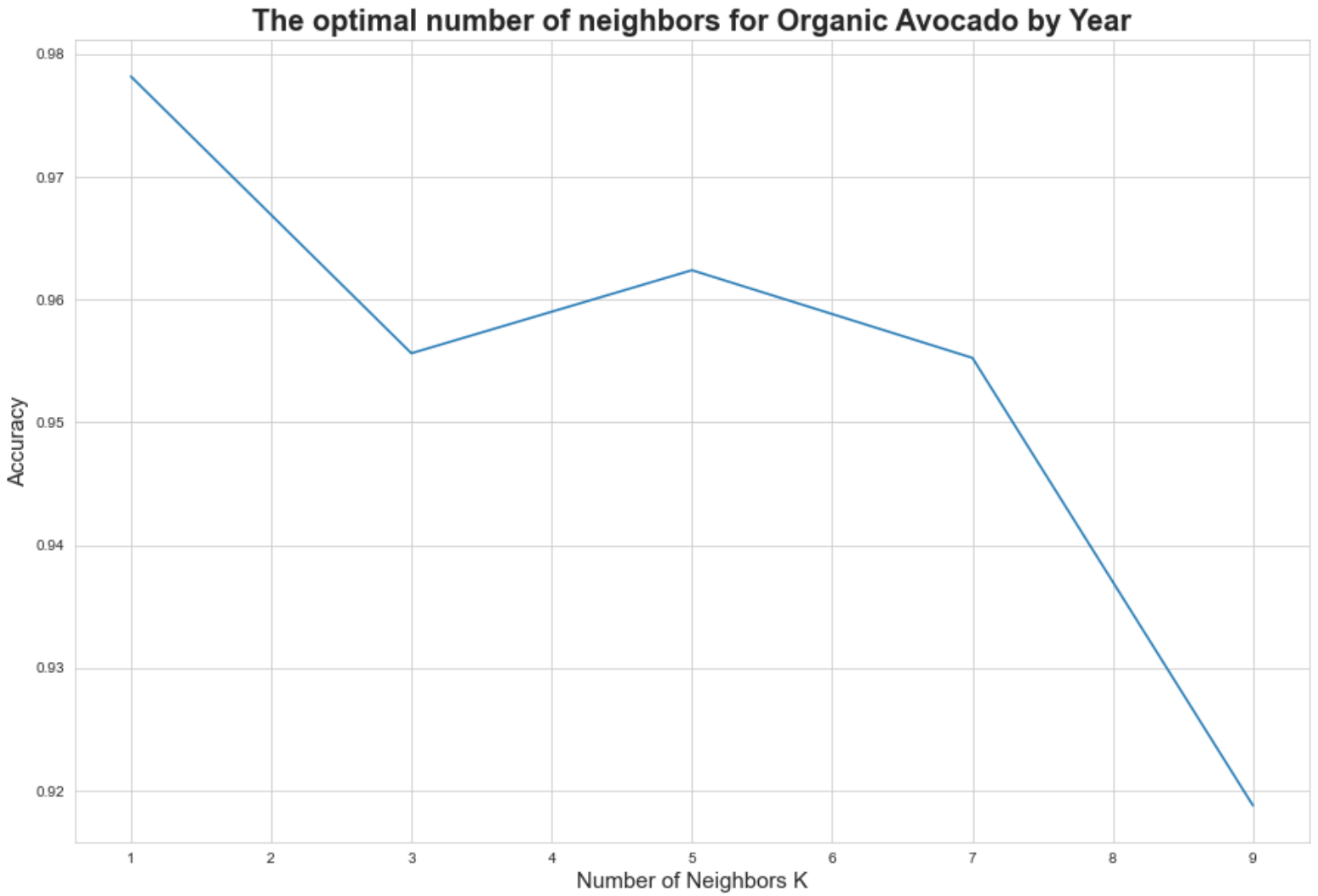
Use KNN to cluster year of Avocado being sold using Average Price, PLU code and number of bags sold

Best accuracy of our model is equal to 66.78% with the number of neighbors being 5

Confusion Matrix:

```
[[960  105   45    7
 233  702  176   21]
 126  255  750   44]
   8   34  115   69]
```

4 K-nearest Neighbour - Year

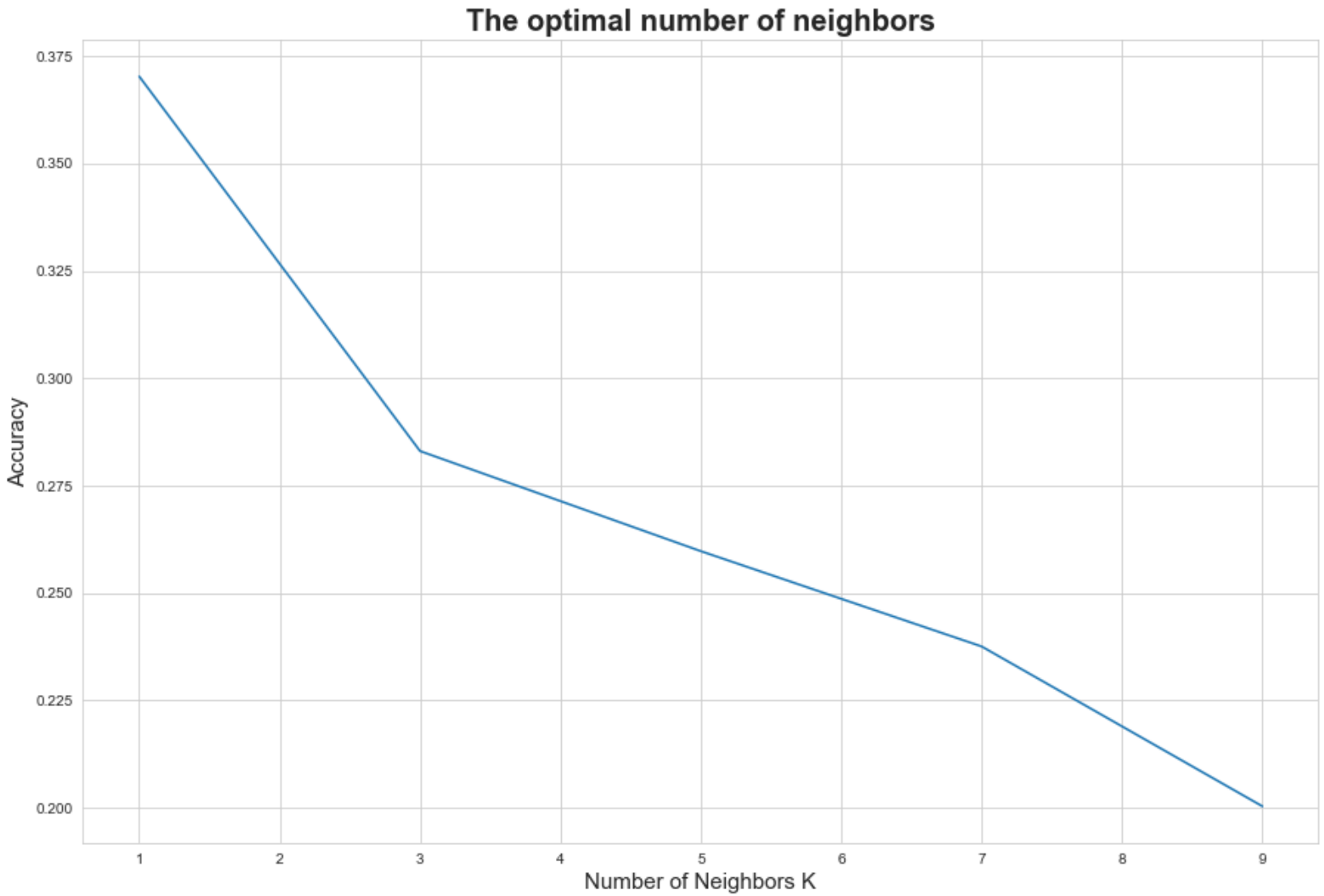


Use KNN to cluster year of Organic Avocado being sold using Average Price, PLU code and number of bags sold

Best accuracy of our model is equal to 97.84% with the number of neighbors being 1

Confusion Matrix:
[[8 0 0 0]
[1 11 0 0]
[0 2 10 0]
 [0 0 0 2]]

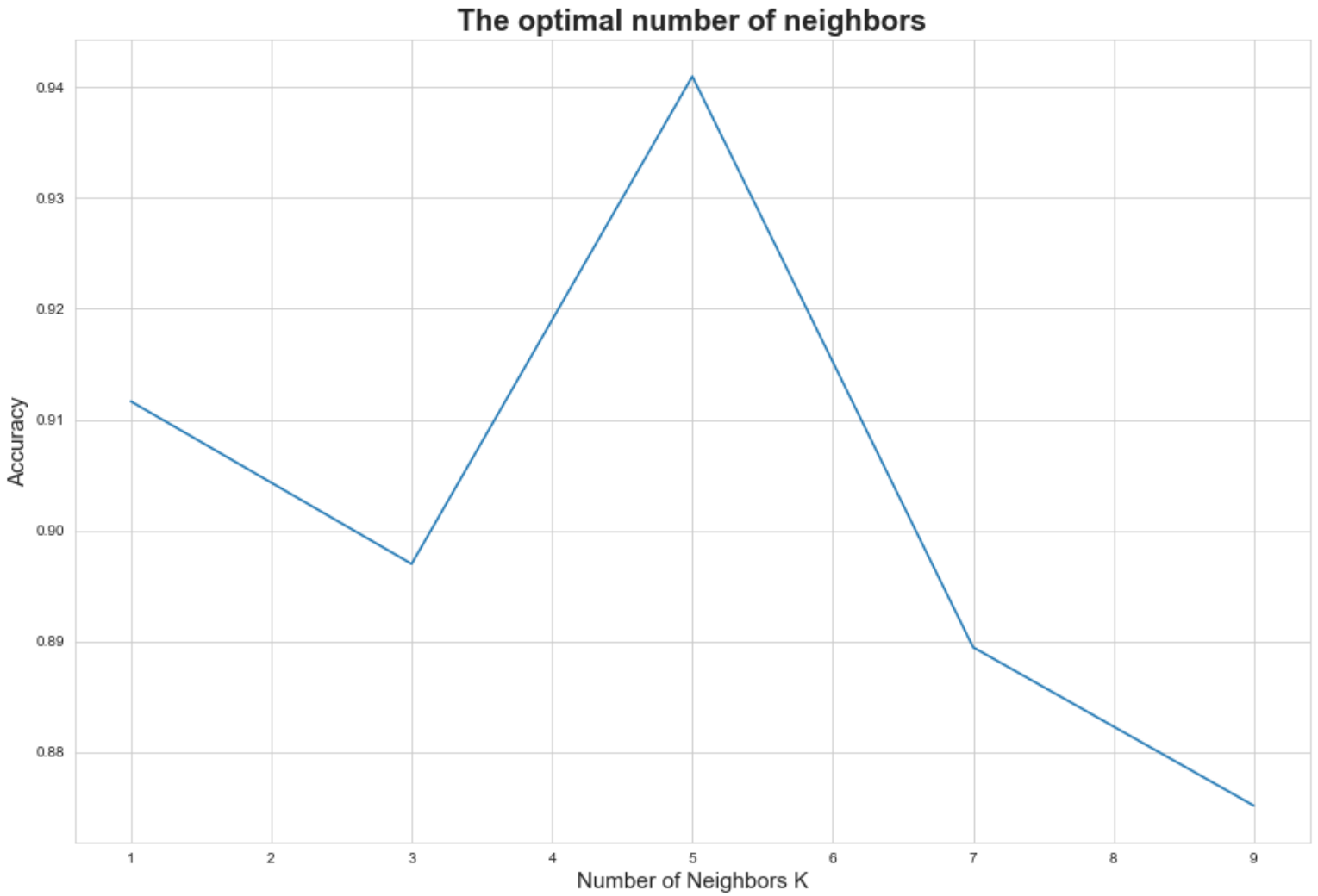
4 K-nearest Neighbour - Month



Use KNN to cluster month of Organic Avocado being sold using Average Price, PLU code and number of bags sold

Best accuracy of our model is equal to 36.69% with the number of neighbors being 1

4 K-nearest Neighbour - Month



Use KNN to cluster months Jan to June and July to Dec of Organic Avocado being sold using Average Price, PLU code and number of bags sold

Best accuracy of our model is equal to 94.12% with the number of neighbors being 5

PCA on Regions

5 PCA on Regions

- Total of 54 regions
- we choose 2015 year as our PCA dataset
- each regions has 104 observations based on time series

```
df['year'].value_counts()
```

```
2017    5722
2016    5616
2015    5615
2018    1296
Name: year, dtype: int64
```

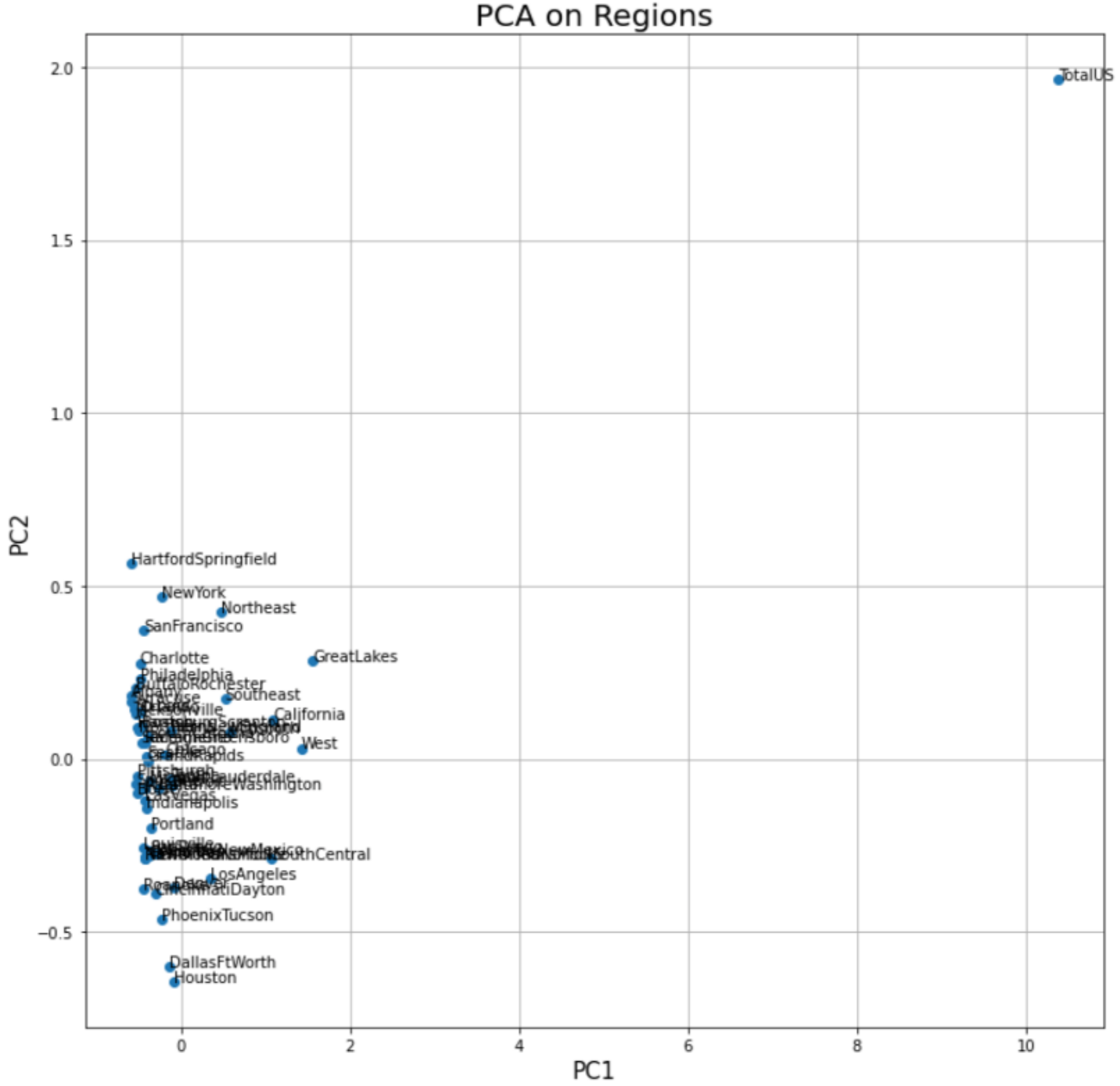
5 PCA on Regions

	region	Principal Component 1	Principal Component 2
0	Albany	-0.351671	-0.883940
1	Albany	-0.361979	-0.847585
2	Albany	-0.178556	-1.617896
3	Albany	-0.251251	-1.344296
4	Albany	-0.337681	-0.977282

```
pca.explained_variance_ratio_
```

```
array([0.65632642, 0.20453237])
```

- **explained_variance_ratio_:** Percentage of variance explained by each of the selected components.


PCA on Regions
PC1
PC2
TotalUS
HartfordSpringfield
NewYork
Northeast
SanFrancisco
Charlotte
GreatLakes
Southeast
California
West
Portland
LosAngeles
PhoenixTucson
DallasFtWorth
Houston
2.0
1.5
1.0
0.5
0.0
−0.5
0
2
4
6
8
10

5 PCA on Regions

	region	PCA0	PCA1	PCA2
0	Albany	-0.351671	-0.883940	0.040763
1	Albany	-0.361979	-0.847585	0.041121
2	Albany	-0.178556	-1.617896	0.049452
3	Albany	-0.251251	-1.344296	0.049502
4	Albany	-0.337681	-0.977282	0.045302

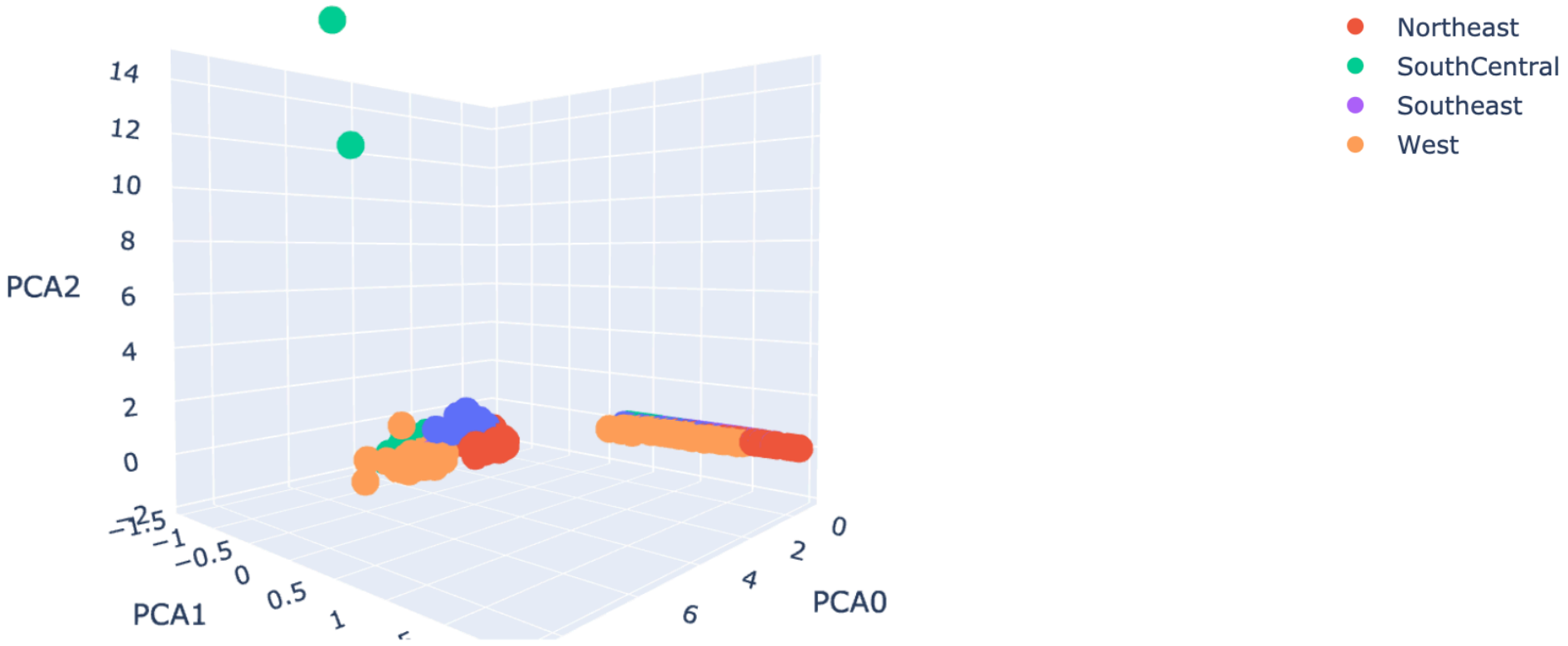
- Let's try to add one more PC and cluster all the data for one region

```
pca.explained_variance_ratio_
```

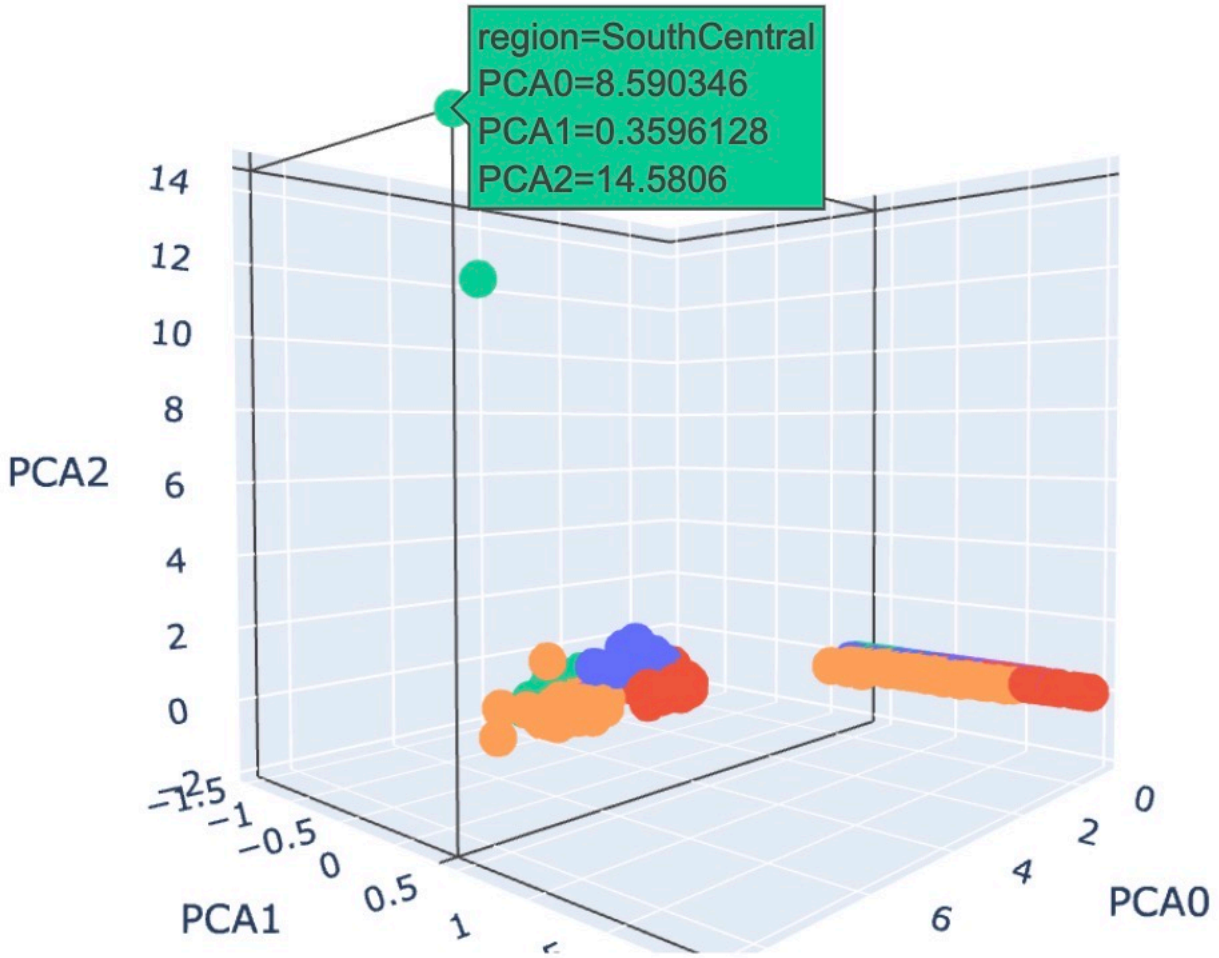
```
array([0.65632642, 0.20453237, 0.07453736])
```

5 PCA on Regions

3D Scatter Plot



5 PCA on Regions



6 Lessons Learned

Lesson 1

Avocado price vs time

Price varies by month and can be separated into two groups. Data from January to June forms a cluster and data from July to December forms a cluster.

The southcentral region of the US show distinct differences to the rest of the US. The prices in this region is much lower. This may be due to it being located near avocado production zones.

Lesson 2

Regional Differences

Lesson 3

Avocado type factor

Avocado type is a strong factor variable. Predictions should be applied separately based on type of avocado. Otherwise, the train dataset are noisy and predictions have very low accuracy.

Random forest regression best predicts numerical price. For classifying avocado type, KNN has higher accuracy than Logistic Regression. PCA suitable for seeing patterns in regions.

Lesson 4

Model Predictions

6 Future Applications

Application 1

Avocado price vs time

Investigate factors that influence price from January to June that distinguishes it from July to December. Can use these factors to influence the price.

The southcentral region may be used as a baseline point for price predictions, as it is consistently lower in price from other regions. Can use it to look at minimum expected price.

Application 2

Regional Differences

Application 3

Avocado type factor

Different approaches to price prediction or classification should be applied to organic and conventional avocados. They should be treated as entirely different commodities from a business perspective.

More variables are needed to accurately predict price, the variables in this dataset are primarily for volume of avocados sold and location.

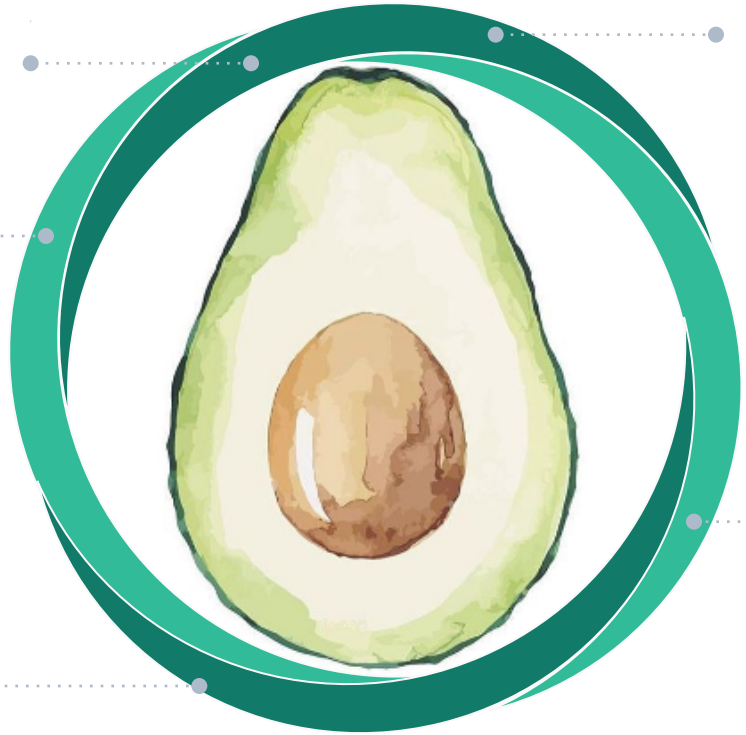
Application 4

Model & Algorithms

6 Conclusions

Regression
Random forest model is best for price prediction

Seasonal Impacts
Prices are high in the fall and less expensive in the spring



Regional Impacts
Wide range of prices over different regions. South and southwest tend to have lower prices.

Organic is Scarce
Organic avocados have large price fluctuations and low supply

Classification
Logistic regression and KNN show similar accuracy for classifying type of avocado.

PCA highlights regional differences


THANKS