Price Forecast of National Avocado Retail

Haoning Deng

Junfei Hao

Zidi Wang

Ziying Yan

Danli Hu

Yinan QI

05/09/2020

Honor Pledge:

We pledge on our honor that we have not given or received any unauthorized assistance on this assignment.

1 Executive Summary

We worked on the data containing weekly retail scan data for national avocado retail volume and price. Target of this project is two-folded. We aim at predicting the future price of avocados depending on the variables we have, and providing retail stores with managerial advice. Two models were applied in the project: Random Forest and XGBoost. For further improvement, we applied time series analysis. According to estimation, there is an overall downward trend of average avocado price in the future 12 month, with 44.44% market will have a downward trend of avocado price while 55.56% of the markets are likely to enjoy an increase in avocado price. Based on the analysis, we came up with some recommendations. We recommend that retailers improve the display and cross display, offer additional size and packages, and be flexible with the source of avocado importers. Since avocado is becoming widely recognized as associated with good health. We recommend that retailers incrementally increase the proportion of organic avocado and highlight avocado's nutritional benefits for children, women's health, and bodybuilding.

2 Background Introduction

Avocado is a stone fruit with a creamy texture that grows in tropical climates. The mania of avocado has increased year by year and there is a growing market of this kind of fruit in America. The avocado dataset contains weekly from 2015 to 2018 retail scan data for National retail volume and price. The retail scan data comes directly from retailers' cash registers based on actual retail sales of Hass avocados. The dataset consists of 14 variables and 18249 observations, detailed information of variables are listed in **Table 1.** To notice that the "Average Price" in the table reflects a per unit cost, even when multiple avocados are sold in bags. The Product Lookup codes (PLU's) in the table are only for Hass avocados. There is no missing value in this dataset.

3 Descriptive Statistics

Type

There are two types of avocado in this dataset: conventional and organic. Below is the change of average price and total volumes of the two types from Jan 2015 to March 2018:

***1）Average price*:** **Exhibit 1** indicates that organic avocado has a higher average price of $1.654, and the average price for conventional avocado is $1.158. The time series chart from 2015 to 2018 shows the turbulence price through years.

***2) Total volume*:** **Exhibit 2** and **Exhibit 3** describe the change of total volumes of conventional and organic avocado. According to the chart, the average amount of conventional avocado sold for a specific day is 35 times larger than organic avocado. Besides, the chart shows that the peak of volume sold for both types appear between December and June. Corresponding to the conclusion of average price described above, we can infer that the lower the price is, the higher the sales is. The chart also shows the increasing popularity of organic avocado during these years, total volume of organic avocado kept growing at a high rate.

Region

In the dataset are 55 unique area names, among which 42 are city level areas and 13 are regional level areas **(Table2)** . As to total sales, Western, Northeastern and Mid-south account for the largest amount, indicating the popularity of avocado in those areas. Northeastern has the highest average price, while the price of avocado is relatively low at the middle area **(Exhibit 4 and 5)**.

Size & Bags

Hass avocado is the most commonly known variety of avocado. The Hass variety accounts for the majority of avocados sold in the U.S.

The most commonly sold sizes of fresh Hass avocado can be identified by their Product Lookup code or PLU or sticker. Introduction of PLU code can be seen at **Table 3**. 4770 represents the largest avocado among the three, while 4225 codes for the medium size avocado and 4046 represents the smallest size. Avocados are also available for sale in netted bags. Three bag sizes are included in this dataset-small bags, large bags and extra large bags. We can see from Chart X that the sales of 4046 and 4225 are neck to neck while the sales of 4770 are far behind **(Exhibit 6)**. The most popular bag size is small bag, followed by large bag and XLarge bag **(Exhibit 7)**.

Price

Exhibit 8 illustrates the fluctuation of average avocado price from 2015 to 2018. In general, average price increased from $1.37 in 2015 to $1.51 in 2017 during this period. Price in 2018 showed a slight drop because of the limited records for only 3 months. **Exhibit 9** shows the average price change in month for two types of avocado. Price of avocado peaks from August to October, and drops to the lowest in February.

Correlation table

According to the correlation table (**Exhibit 10**), there is a high correlation between small hass & total volume (0.89), total bags & total volume (0.87), and small bags & total bags (0.96). Small Hass avocados are the most preferred type in the US and customers tend to buy those avocados as bulk, not bags. They think this is more advantageous for them. Total Bags variable has a very high correlation with Total Volume and Small Bags, so we can say that most of the bagged sales comes from the small bags.

4 Methodology

Data Preparation

For the time series dataset, we separated the "Date" variable to the "year", and "month". Also, we tried to find some price fluctuation patterns in the year. Based on **Exhibit 9**, we found both prices for organic and conventional avocado have seasonal fluctuation. Therefore, we created 4 dummy variables, "spring" (from February to April), "summer" (from May to July), "autumn"(from August to October), and "winter" (from November to January next year).

For the category variable "region", we have two different methods to deal with it. One was one-hot encoding and the other is label-encoding for decreasing the risk of feature sparsity.

Model building

We chose random forest and XGBoost as our prediction model.

1) Random Forest

Random Forest is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces overfitting problems and also reduces the variance and therefore improves the accuracy. Random forest can deal with multicollinearity problems better than other models.

2) XGBoost

We chose XGBoost model, which has in-built regularization to prevent the model from overfitting. Its ability to parallel processing makes it faster than the Gradient Boost model. At the same time, it allows users to run a cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run.

Since XGBoost supports both one-hot encoding and label-encoding datasets very well, we chose label encoding dataset which makes training and prediction more efficient and accurate.

Model Optimization:

1) 5-fold Cross Validation

We applied the cross-validation technique to determine the parameters of both models in order to test which parameter can optimize the model performance. In the 5-fold cross-validation, the original sample was randomly partitioned into 5 equal sized subsamples. From the 5 subsamples, a single subsample was retained as the validation data for testing the model, and the remaining 4 subsamples were used as training data. The cross-validation process was then repeated 5 times.

2) Parameter Adjustment (Grid Search)

To perform hyperparameter optimization for the Random Forest model and XGBoost model, we used the grid search, which is an exhaustive search over specified parameter values for an estimator.

- ***Random Forest***

For this random forest model, the parameters we decided to tune are max_features, min_samples_leaf, min_sample_split and n_estimators, which combined play an important role in the final prediction accuracy. We tuned one parameter for each step and applied the parameter value yielding the best result into the next step for another parameter tuning. This tuning method, compared with grid search that evaluates all combinations we define, is much more practical and saves more computation time. The combination of parameters we decided on was: max_features=41, min_samples_leaf=2, min_sample_split=2 and n_estimators=1000.

- ***XGBoost***

Since there are many parameters that should be tuned, the entire grid search is time-consuming and inaccurate. We chose to tune parameters step by step. Before tuning, we used the default value of all the parameters. And in each step, we decided one or two specific parameters and applied them to the next step. After deciding the basic num_boost_round to 20, we applied grid search to the model, based on the result, we finally adjusted num_boost_round to 100, “gamma” to 0, “max_depth” to 10, “min_child_weight” to 5, and “eta” to 0.09.

5 Results and findings

Important Features & Model selection

For the XGBoost model, the most important feature which influences the model result is “4225”, followed by “region” and “4046” **(Exhibit 11)**; For the Random Forest model, the most important feature which influences the model result is “Xlarge Bags”, followed by “Total Volume”, “Small bags” and “4046” **(Exhibit 12)**. Combining the feature importance of two models, avocado “4046” is the commonly most important variable to affect the average price.

Model Selection

We applied 10-fold cross validation in the training dataset to evaluate these two optimized models’ performance. Considering two aspects, we chose RMSE and standard deviation as accuracy and stability evaluation respectively. We stored the RMSE score in each cross validation and calculated the standard deviation for each model.

Based on the chart and box-plot **(Exhibit 13)**, the average of RMSE from random forest is 0.1233, and the standard deviation is 0.0035. For the XGBoost model, the average of RMSE is 0.1200, the standard deviation is 0.0033. It is clear that both models performed well and were stable. To determine the better one, we decided to apply the XGBoost model to predict the testing dataset. The RMSE for the final prediction on the testing dataset is 0.117, which is satisfying for us.

6 Time Series Analysis

Preprocessing

We applied time series forecasting with Prophet in Python and plotted the result with Exploratory because there seems to be a seasonal pattern in the average price. First, we prepared the dataset for time series prediction. We dropped the unnamed column and rename undefined columns 'Unnamed: 0', then we rename "4046" as "Small Hass", "4225" as "Large Hass", and "4770" as "XLarge Hass". Next, we converted the "Date" column to "datetime" type by using the "DatetimeIndex" method. We specified the uncertainty interval to 95% and predicted the price of avocado for the future 12 months We asked the "make_future_dataframe" method to generate timestamps and stored the results according to future dates in the "ds" column.

Result and Analysis

For the entire market, we forecasted the average price of the next 10 weeks and next 52 weeks (1 year). To assess the performance of the model, we did cross-validation to assess prediction performance on a horizon of 365 days, starting with 730 days of training data in the first cutoff and then making predictions every 180 days. We observed a mean RMSE of 0.1382 for the whole market **(Exhibit 14)**. Indicated by **Exhibit 15,** there is estimated a decrease in price for the next year. Next, we forecasted the trend of price for every existing market, the result is shown in **Exhibit 16**. We discovered that 44.44% market will have a downward trend of avocado price while 55.56% of the markets are likely to enjoy an increase in avocado price. As we can see from **Exhibit 17**, cities with decrease in price intensively locate in Northeast and South West, while Northwest, Great Lakes and Middle South's market tend to have a decrease in price.

7 Recommendation

From analysis of the scan data for national retail volume (units) and price, we derived recommendations for retailers in terms of purchasing, stocking and displaying.

1) Source

According to USDA, Mexico accounts for 87% of the total volume and 88% of the total value of U.S. avocado imports. However, the fact that avocado price rise and fall on border threats might lead to the potential shortage of supplement and risk of price changes at a future point in time. We recommend that retailers diversify their avocado import sources in order to lower the risk of price fluctuation.

- Avocado ripens off the tree in autumn. We recommend that retailers trade off between the prices and transportation fee, consider shifting between avocado exporters in northern and southern hemispheres.

2) Type

According to **Exhibit 18**, more customers inclined to purchase organic avocados. We recommend that retailers incrementally increase the proportion of organic avocados. Table ABC illustrates that people in different cities and regions' preference over avocado sizes, bags and types varies, we recommend retailers research on the reasons that lead to the difference and prepare the products according to customer's favor.

3) Bag & Bulk

We noticed that some region's grocery stores sell a limited variety of sizes and types of avocado. For example, the volume sold of the PLU 4046 in New York is very low **(Exhibit 19),** and the volume sold of the Extra Large bags in New York is 0 **(Exhibit 20),** which might result from the fact that groceries in New York didn't prepare PLU 4046 and Extra Large bags at all and people had no other choice. However, we believe selling multiple sizes and packages will increase sales incrementally.

We recommend having different sizes of avocados displayed in the same place, even placing a promotional size or bag of avocados together, which will stimulate the sale of secondary displays.
- Specifically, we found that PLU 4046 is much more important than other sizes both in the random forest model and XGBoost model **(Exhibit 11 & 12)**. Therefore, we recommend retailers should pay more attention to the sale of PLU 4046 to gain more profit.

4) Price

According to our prediction of price for the future 12 month. 44.44% of the market in our dataset will have a decrease in avocado price while 55.56% are likely to have an increase. Getting the price right is crucial. We recommend running a pricing market research study and investigating the reason behind avocado price fluctuation.

Since avocado price is affected by many more factors than those included in the model, we recommend that beyond the model, grocery store take into account weather condition which affect the growth and harvest of avocado, policy related to trading with external avocado producer, the change of people's lifestyle, and their potential shift in attitude towards avocado.

5) Promotion

In the past, taste has been the main reason why consumers purchase avocados, today nutritional benefits become the top reason. It is becoming widely recognized that consumers associate avocados with good health. We recommend that retailers highlight avocado's nutritional benefits for children, women's health, and bodybuilding.

6) Display & Cross-Display

We recommend that grocery stores pair avocados with tomatoes, onions, and other items that appear in popular recipes along with avocado a lot. To inspire shoppers of new recipes or remind them of the traditional avocado recipe could motivate the sales. As people's aware of organic avocado rise, we recommend that retailers consider highlighting organic avocado when displaying.

8 Limitation

1) We don't have information telling where the avocado is produced. The origin of the avocado might shed light to some of the patterns we discover. Take New York as an example. The current data indicates that New Yorkers bought more #4225 Avocado than all the other cities, but there is a possibility that avocado bought by New York's avocado retailers are mostly #4225. Secondly, in the dataset we only have information about avocado sold in bags, which limits our interpretation. We believe that it's important to measure pruit for sale in units.

2) Three sizes of bags are involved, but we don't have a precise definition as to how large or how heavy the bags are, or how many avocados are in the bags.

3) There are potential limitations for Random Forest Model: After introducing more parameters into the model, the "best" parameter value gained in the last step of tuning might not still be the one that gives the lowest test-RMSE. Also, efficiency is a problem. It would take much more time for the model to compute if we use grid search to fit best parameters automatically.

Appendix

Table1

AveragePrice	The average price of a single avocado
Conventional	Type of avocado, equals 1 if the avocado is conventional
Organic	Type of avocado, equals 1 if the avocado is organic
Region	The city or region of the observation
Total volumes	Total number of avocados sold
X4046	Total number of avocados with PLU 4046 sold
X4225	Total number of avocados with PLU 4225 sold
X4770	Total number of avocados with PLU 4770 sold
Year	The year of the observation
Month	The month of the observation
Day	The day of the observation
Small.Bags	/
Large.Bags	/
ELarge.Bags	/

Table2 City Level and Regional Level areas

City Level	Regional Level
California, GreatLakes, SouthCarolina, West TexNewMexico, NorthernNewEngland, Plains, Midsouth, Northeast, SouthCentra, Southeast, West, TotalUS	Albany, Atlanta, BaltimoreWashington, Boise , Boston, BuffaloRochester, Charlotte, Chicago, CincinnatiDayton, Columbus, Dallas-FtWorth, Denver, Detroit, GrandRapids, Harrisburg-Scranton, Hartford-Springfield, Houston, Indianapolis, Jacksonville, LasVegas, Los Angeles, Louisville, Miami-FtLauderdale, Nashville, New Orleans Mobile, NewYork, Orlando, Philadelphia, PhoenixTucson, Pittsburgh, Portland, Raleigh-Greensboro, Richmond-Norfolk, Roanoke, Sacramento, SanDiego, San Francisco,Seattle, Spokane StLouis, Syracuse, Tampa
13	42

Table 3 PLU Code of Avocado

product	size	PLU code
Small/Medium Hass Avocado	(~3-5oz avocado)	#4046
Large Hass Avocado	(~8-10oz avocado)	#4225
Extra Large Hass Avocado	(~10-15oz avocado)	#4770

Exhibit 1 Average Price of Avocado Through Year

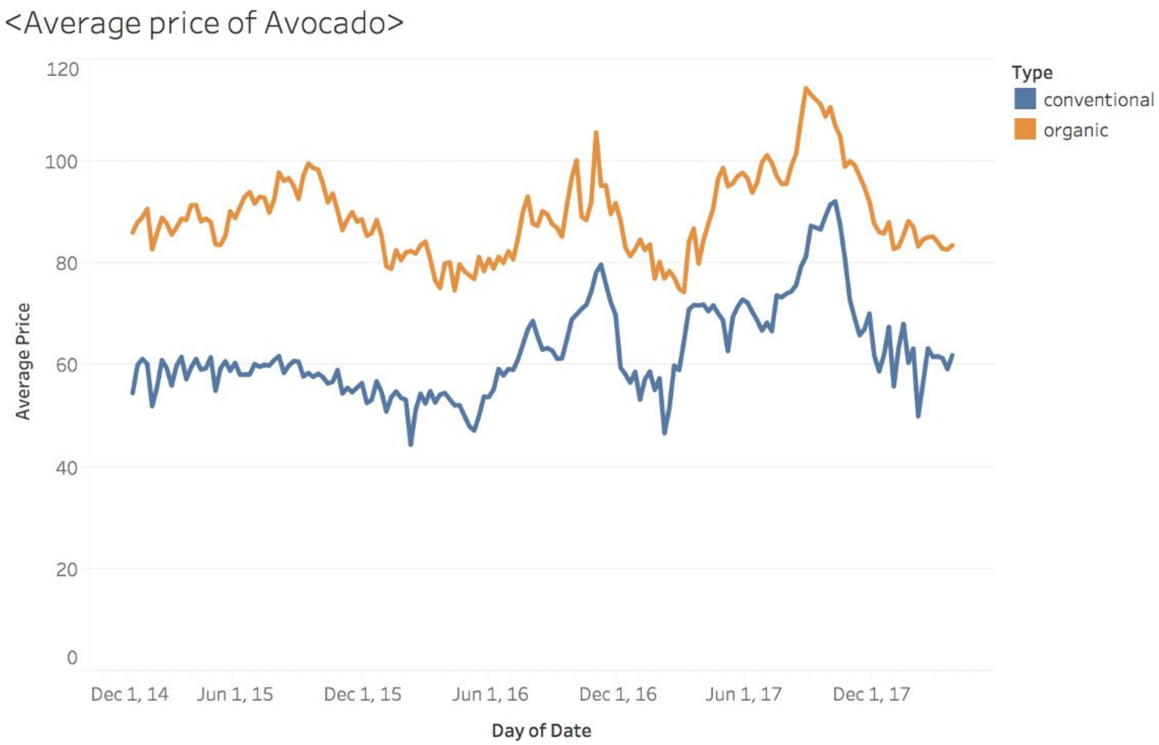


Exhibit 2 Average Volume of Avocado Sold Through Year (Conventional)

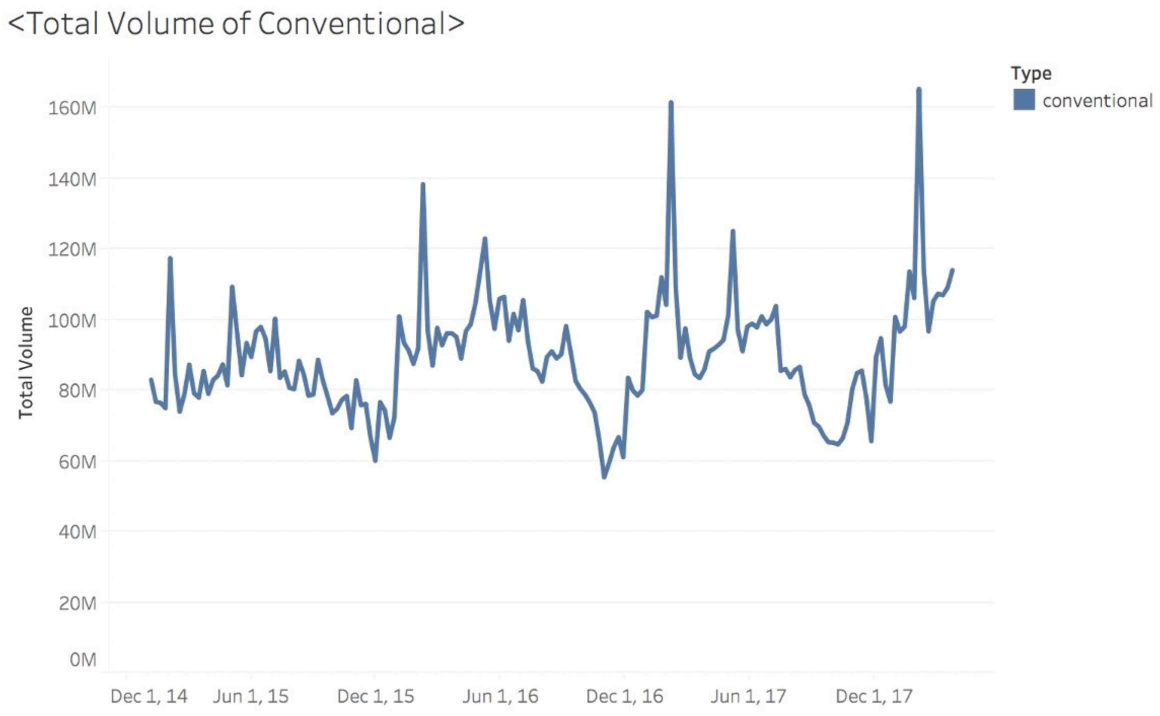


Exhibit 3 Average Volume of Avocado Sold Through Year (Organic)

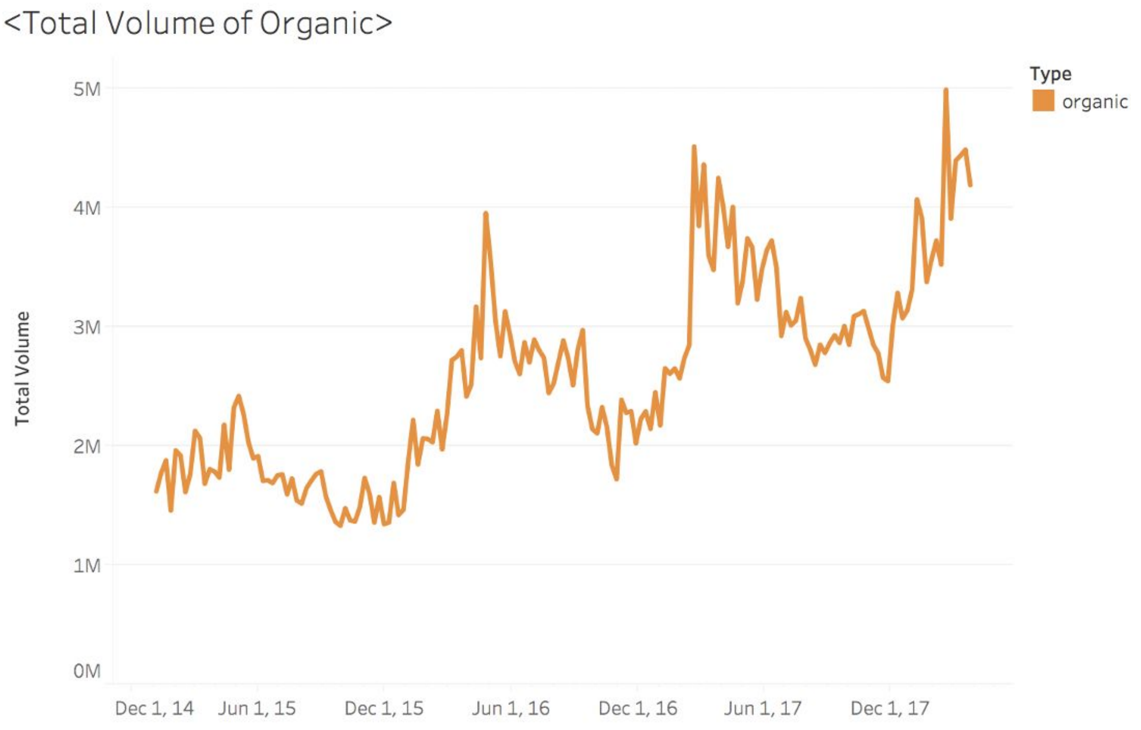


Exhibit 4 Heat map of Total Sales

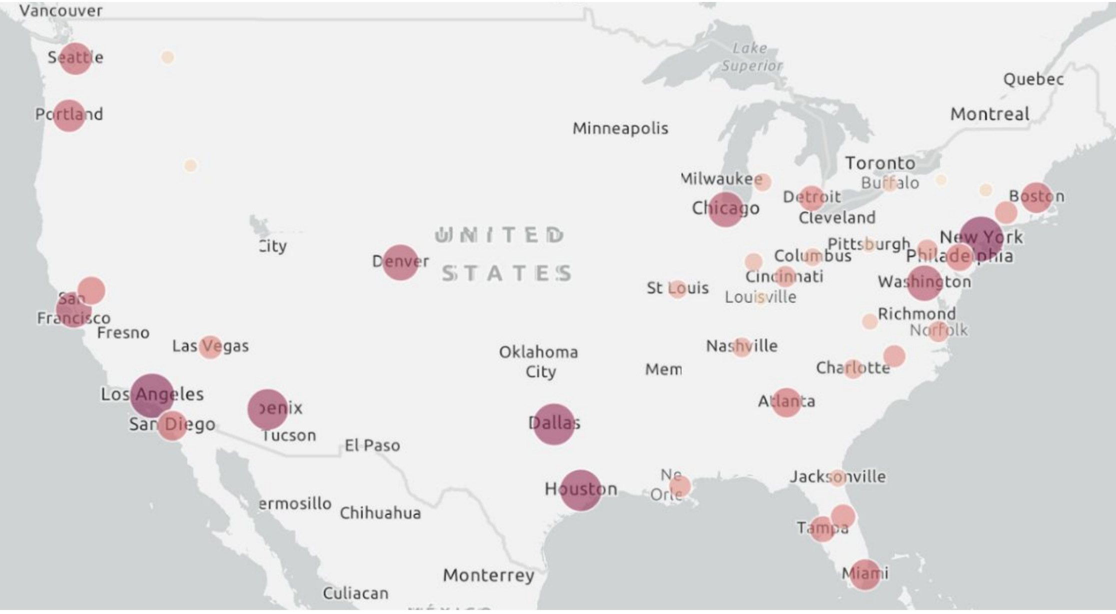


Exhibit 5 Heat map of avocado price

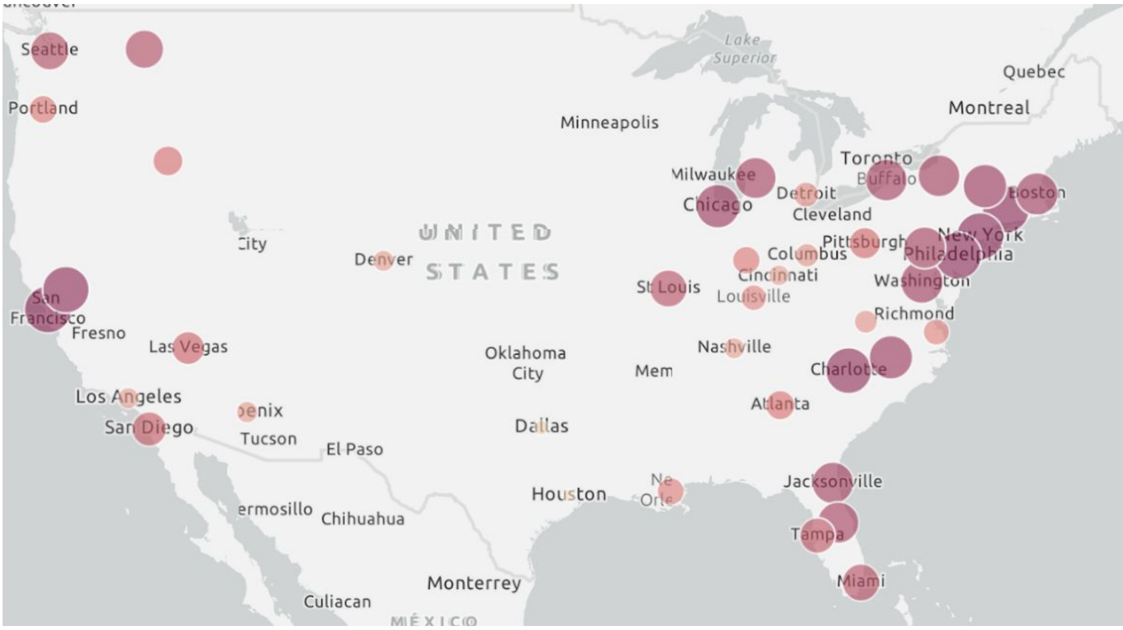


Exhibit 6 Total volume for different types

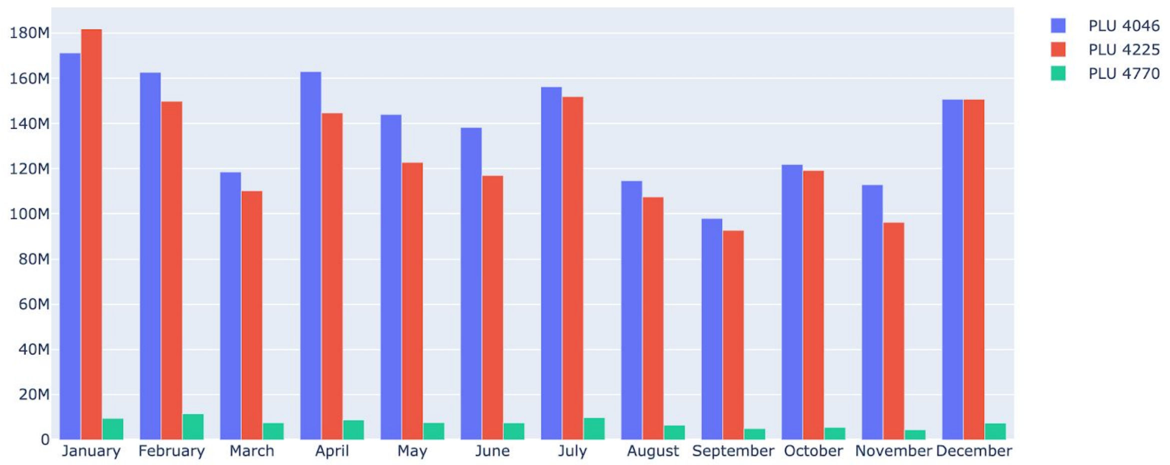


Exhibit 7 Total volume for different size of bags

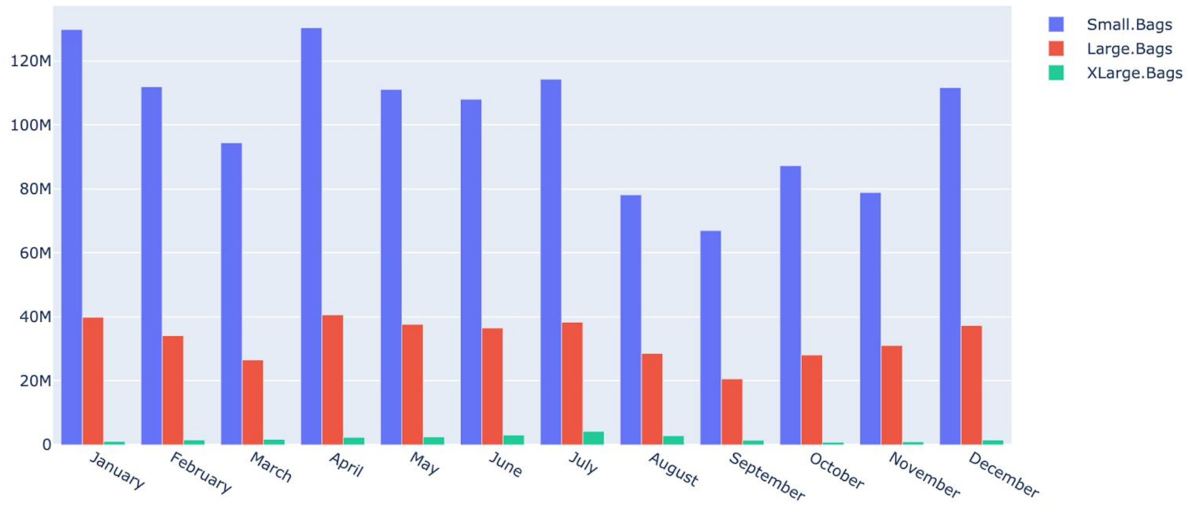


Exhibit 8 Average price of Avocado

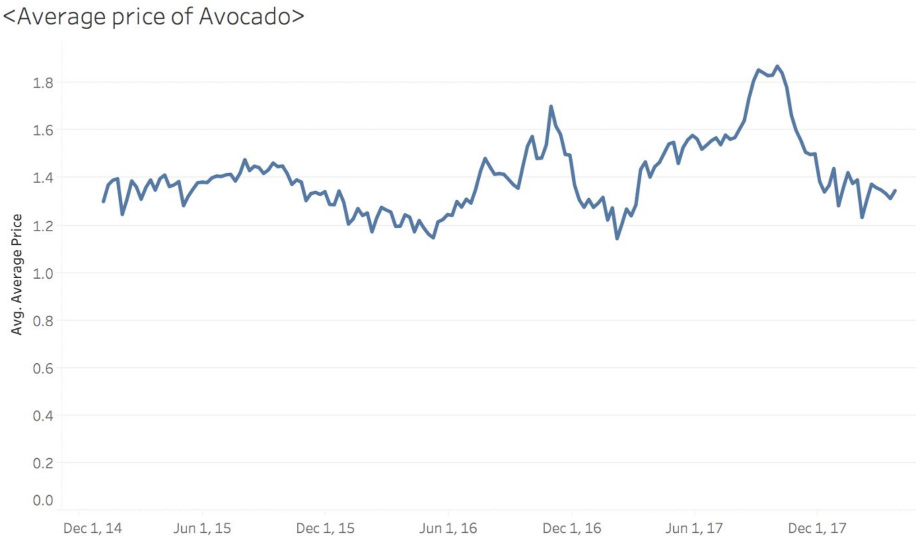


Exhibit 9 Average price change of avocado in months

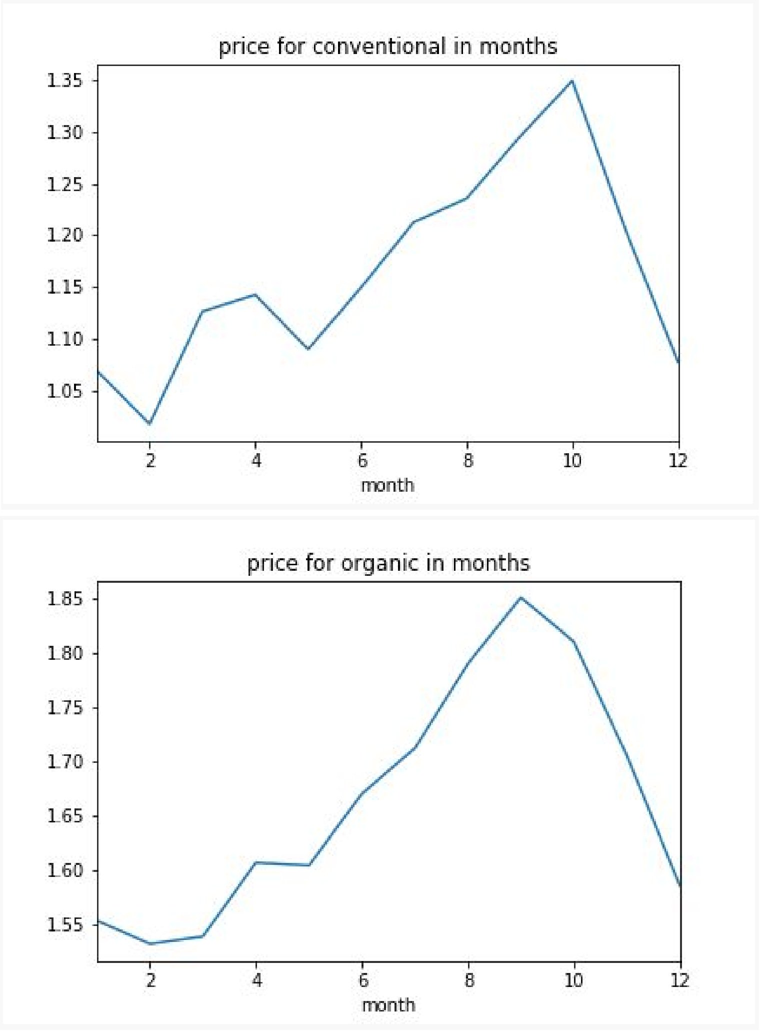
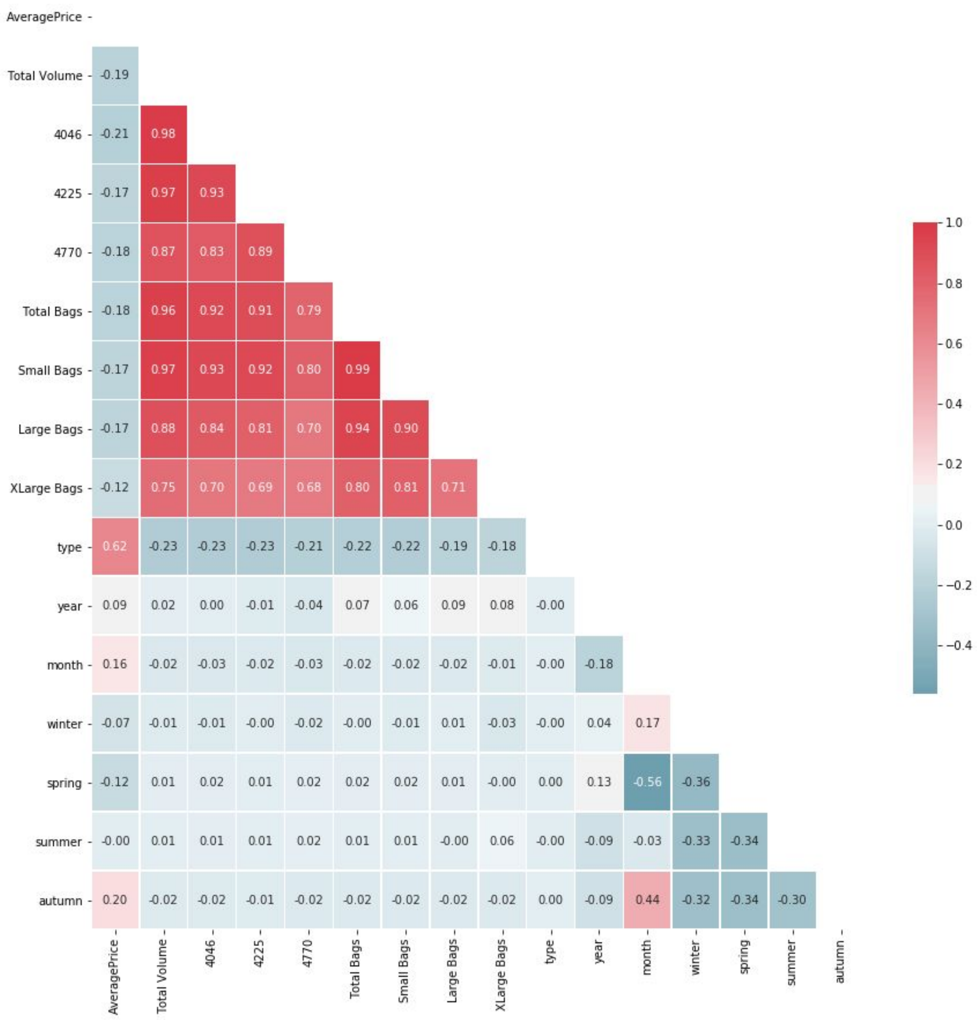


Exhibit 10 Correlation Table



	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	month	winter	spring	summer	autumn
AveragePrice																
Total Volume	-0.19															
4046	-0.21	0.98														
4225	-0.17	0.97	0.93													
4770	-0.18	0.87	0.83	0.89												
Total Bags	-0.18	0.96	0.92	0.91	0.79											
Small Bags	-0.17	0.97	0.93	0.92	0.80	0.99										
Large Bags	-0.17	0.88	0.84	0.81	0.70	0.94	0.90									
XLarge Bags	-0.12	0.75	0.70	0.69	0.68	0.80	0.81	0.71								
type	0.62	-0.23	-0.23	-0.23	-0.21	-0.22	-0.22	-0.19	-0.18							
year	0.09	0.02	0.00	-0.01	-0.04	0.07	0.06	0.09	0.08	-0.00						
month	0.16	-0.02	-0.03	-0.02	-0.03	-0.02	-0.02	-0.02	-0.01	-0.00	-0.18					
winter	-0.07	-0.01	-0.01	-0.00	-0.02	-0.00	-0.01	0.01	-0.03	-0.00	0.04	0.17				
spring	-0.12	0.01	0.02	0.01	0.02	0.02	0.02	0.01	-0.00	0.00	0.13	-0.56	-0.36			
summer	-0.00	0.01	0.01	0.01	0.02	0.01	0.01	-0.00	0.06	-0.00	-0.09	-0.03	-0.33	-0.34		
autumn	0.20	-0.02	-0.02	-0.01	-0.02	-0.02	-0.02	-0.02	-0.02	0.00	-0.09	0.44	-0.32	-0.34	-0.30	

Exhibit 11 Feature Importance of XGBOOST

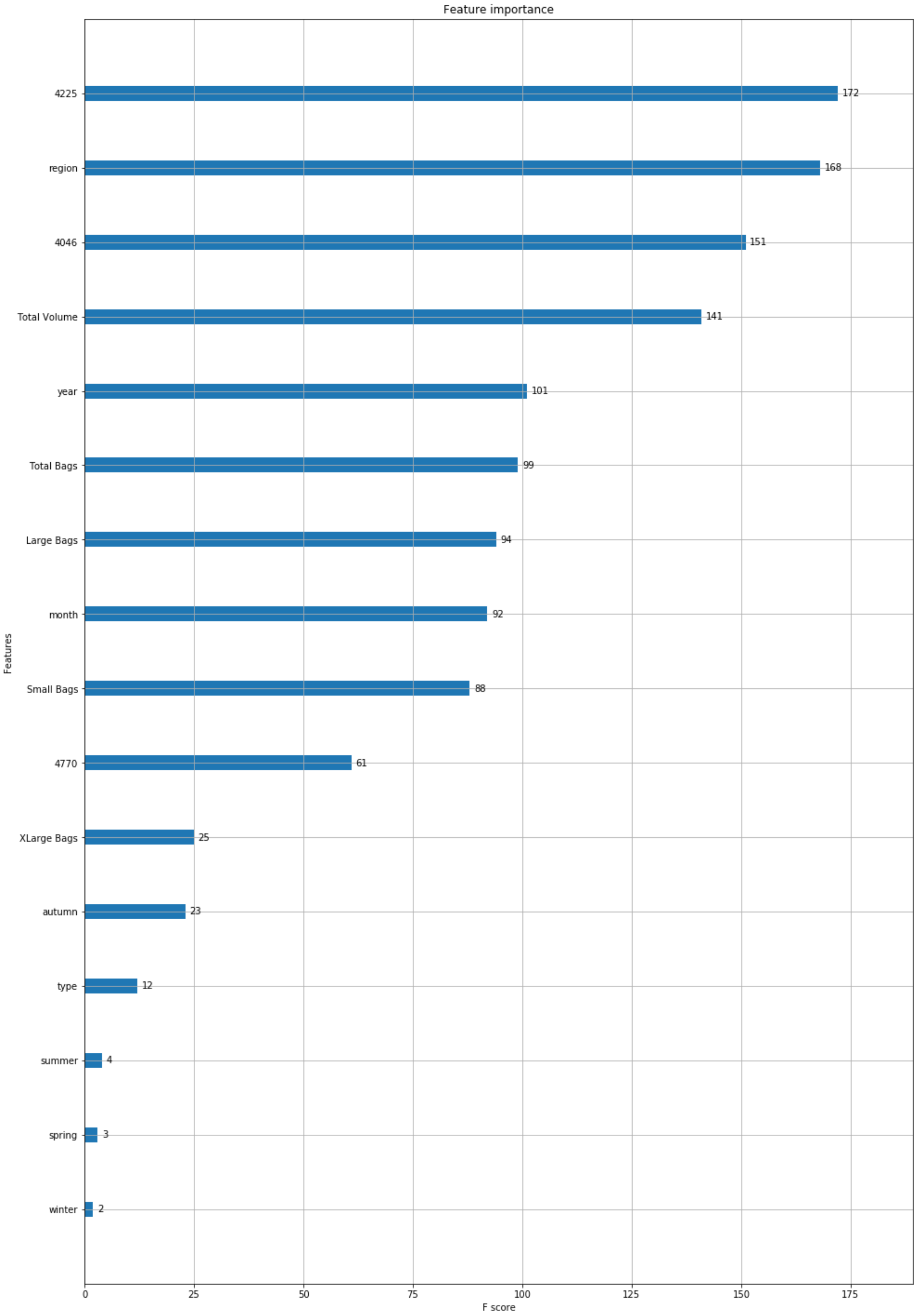


Exhibit 12 Feature importance for Random Forest

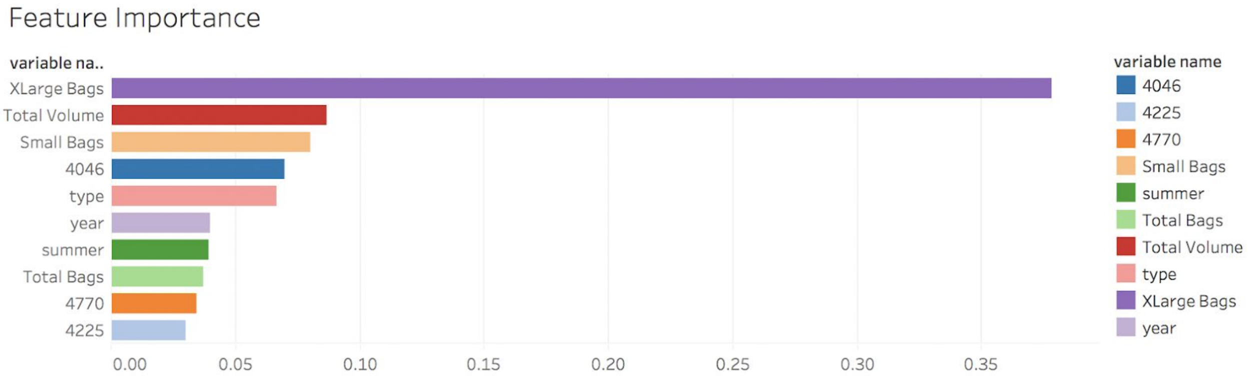
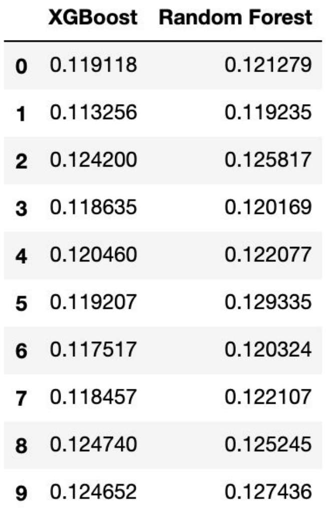


Exhibit 13 RMSE Results Comparison of Two Models



	XGBoost	Random Forest
0	0.119118	0.121279
1	0.113256	0.119235
2	0.124200	0.125817
3	0.118635	0.120169
4	0.120460	0.122077
5	0.119207	0.129335
6	0.117517	0.120324
7	0.118457	0.122107
8	0.124740	0.125245
9	0.124652	0.127436

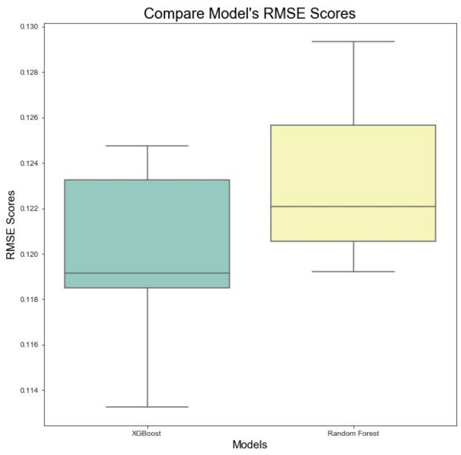


Exhibit 14 The performance matrix for Time Series

	horizon	mse	rmse	mae	mape	mdape	coverage
44	337 days	0.004362	0.066044	0.051736	0.040163	0.040659	0.8
45	344 days	0.005889	0.076737	0.066675	0.05129	0.041656	0.8
46	351 days	0.00543	0.073688	0.064758	0.048328	0.041656	1
47	358 days	0.009002	0.094879	0.082893	0.062092	0.065509	0.8
48	365 days	0.01178	0.108536	0.10441	0.078099	0.085986	0.8

Exhibit 15 Price forecasting for the next 52 weeks of the whole market (1 year)

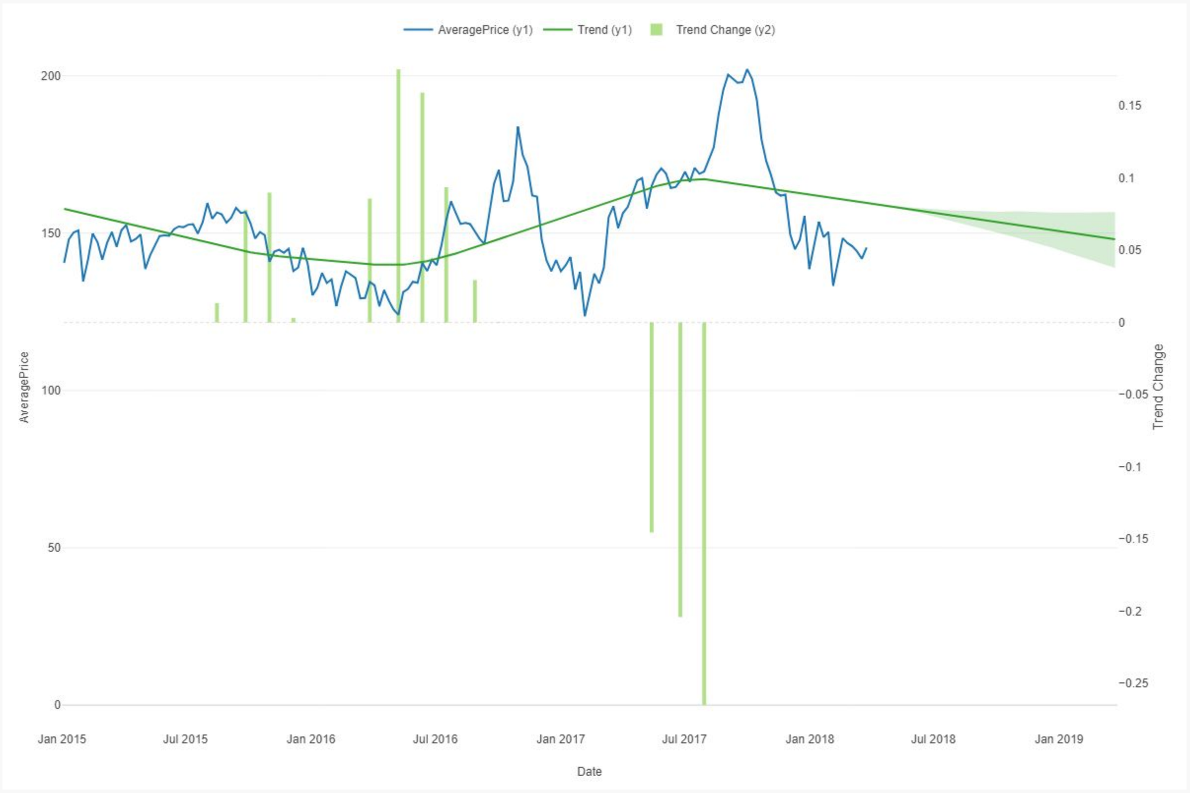
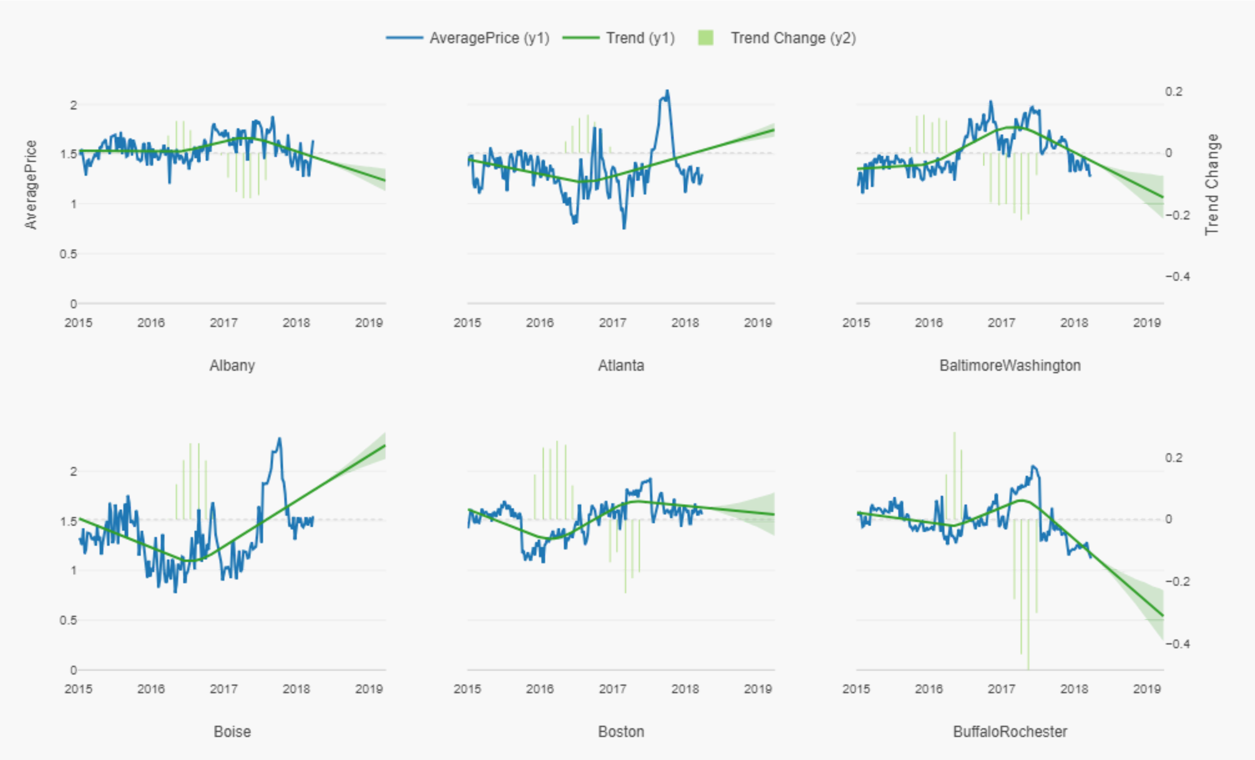
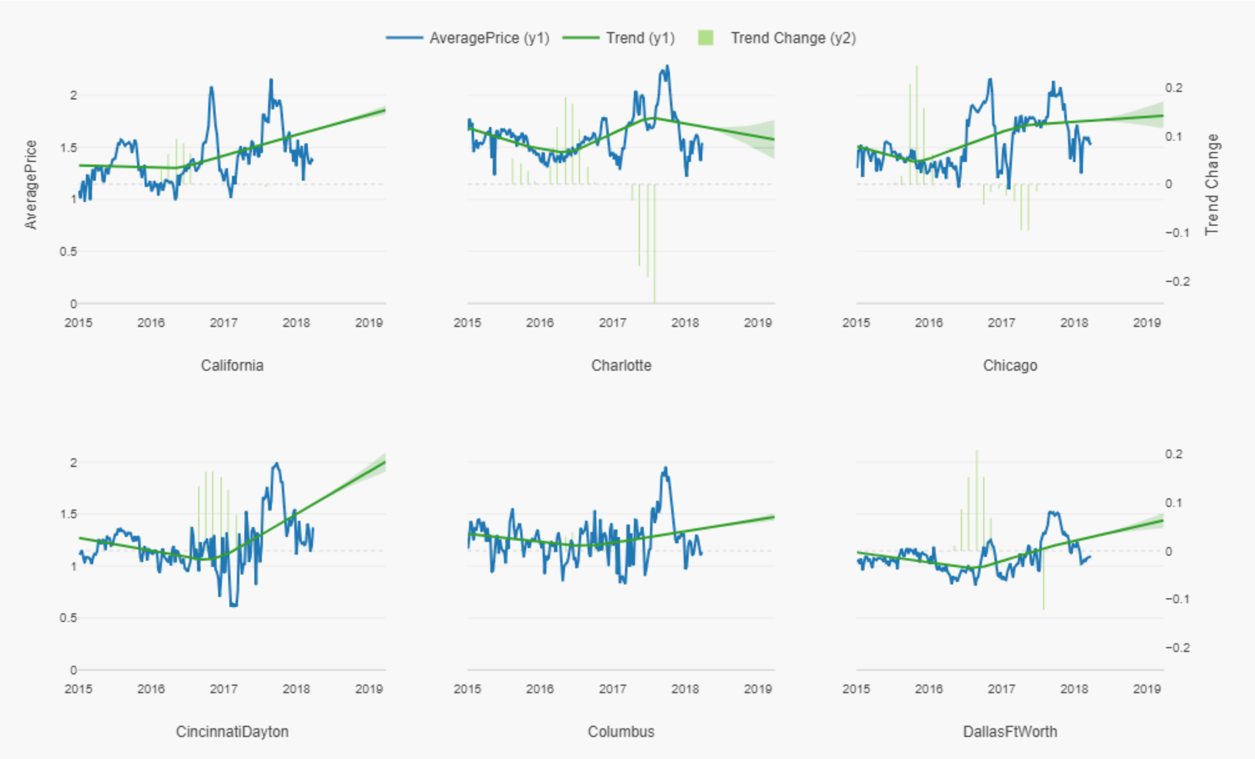
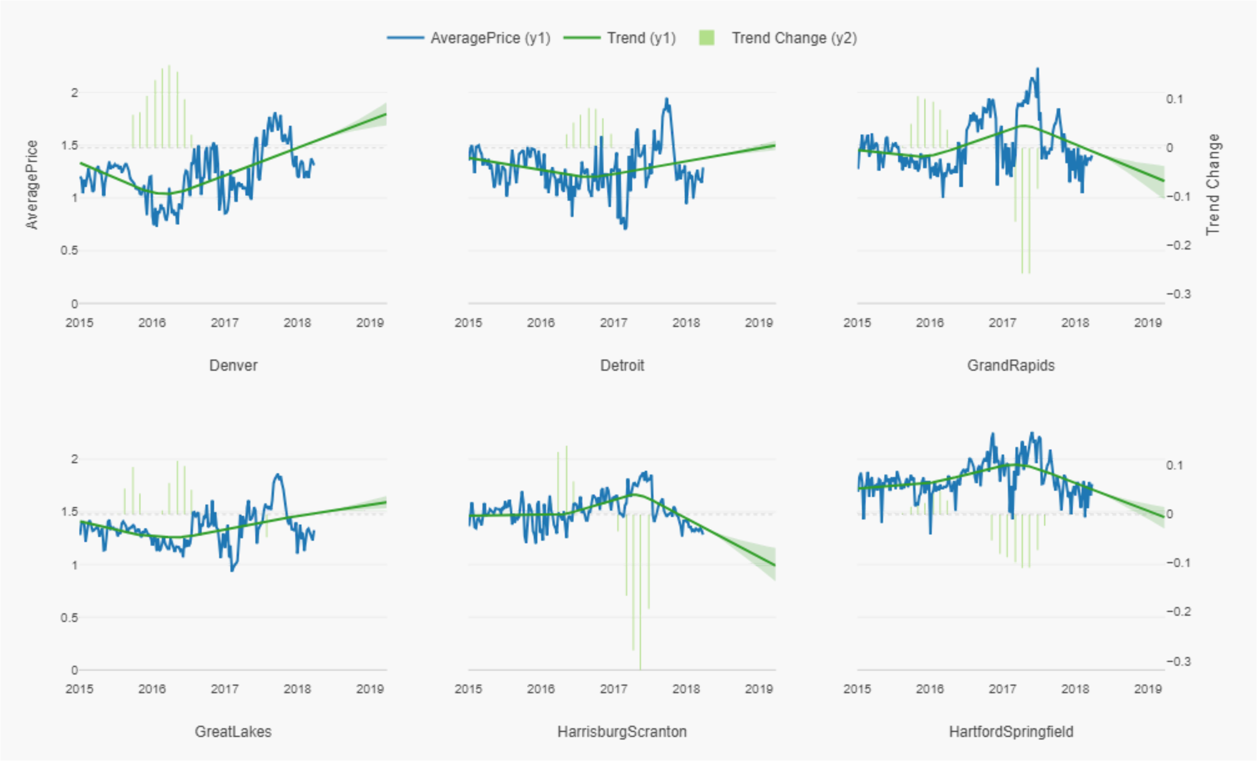
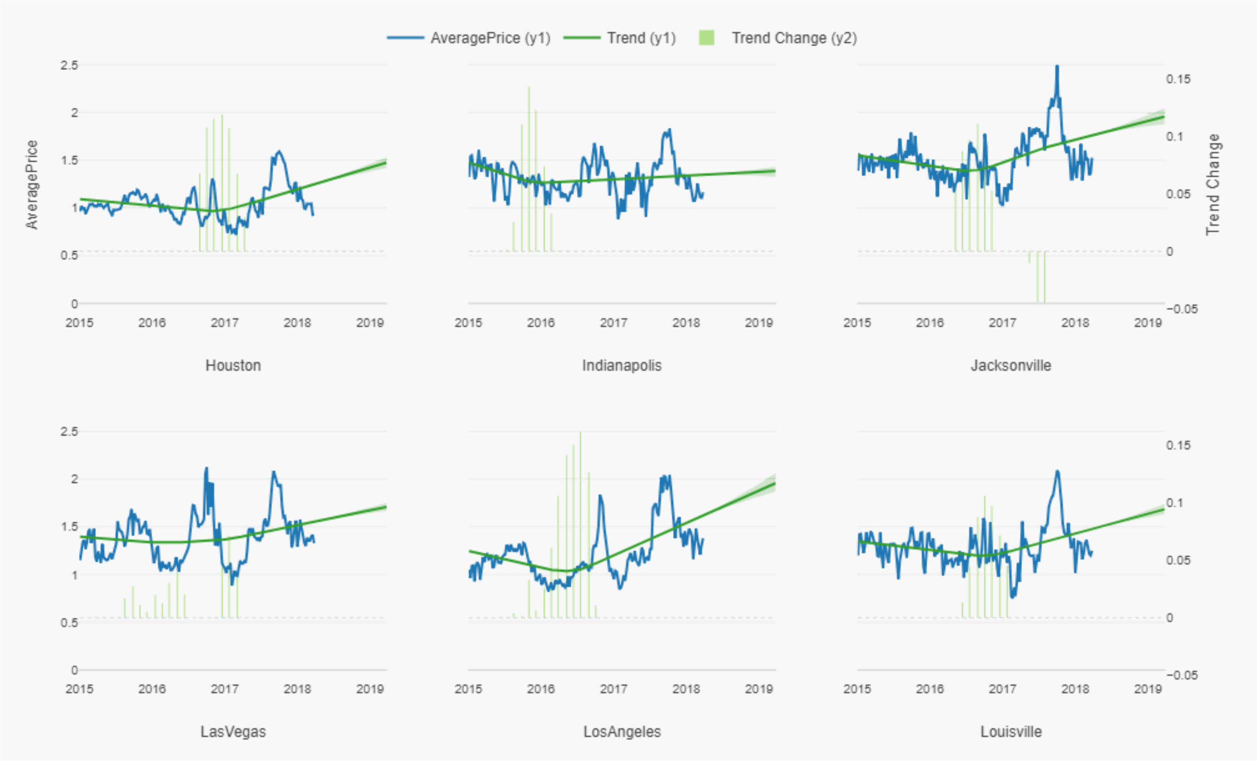


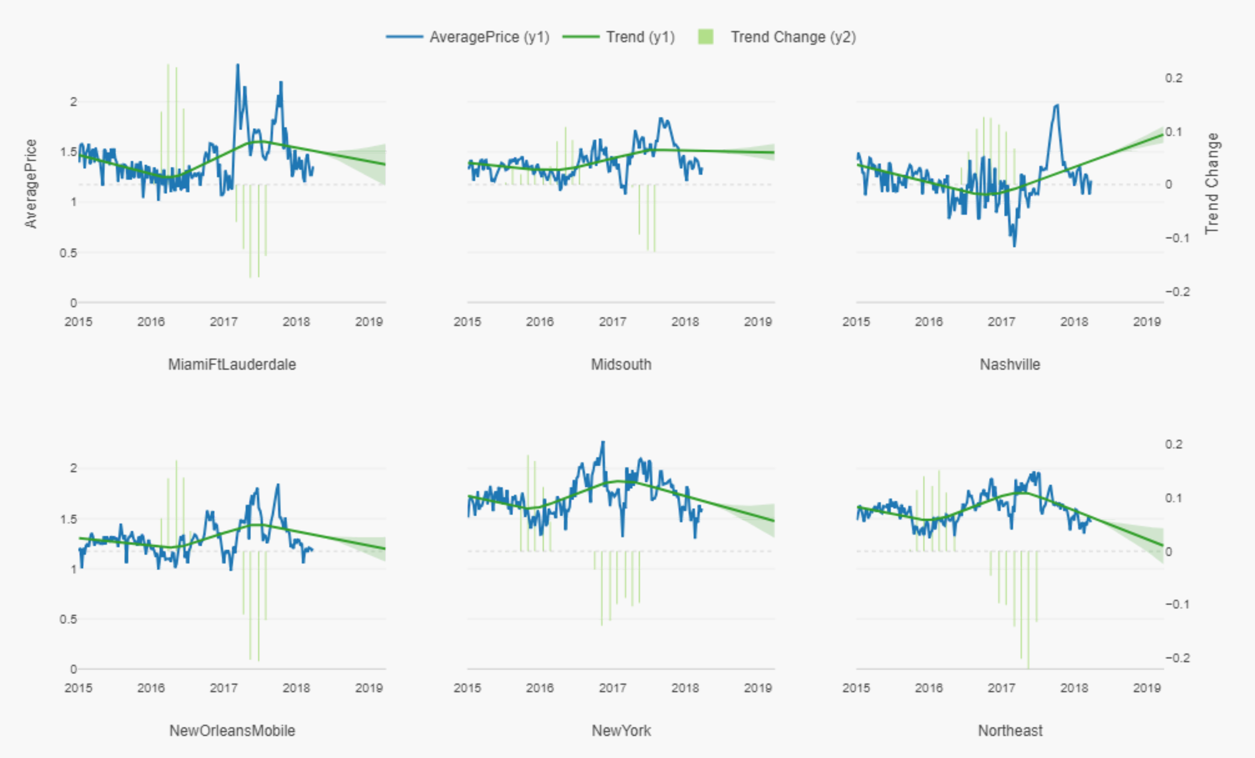
Exhibit 16 Trend forecasting for each individual market

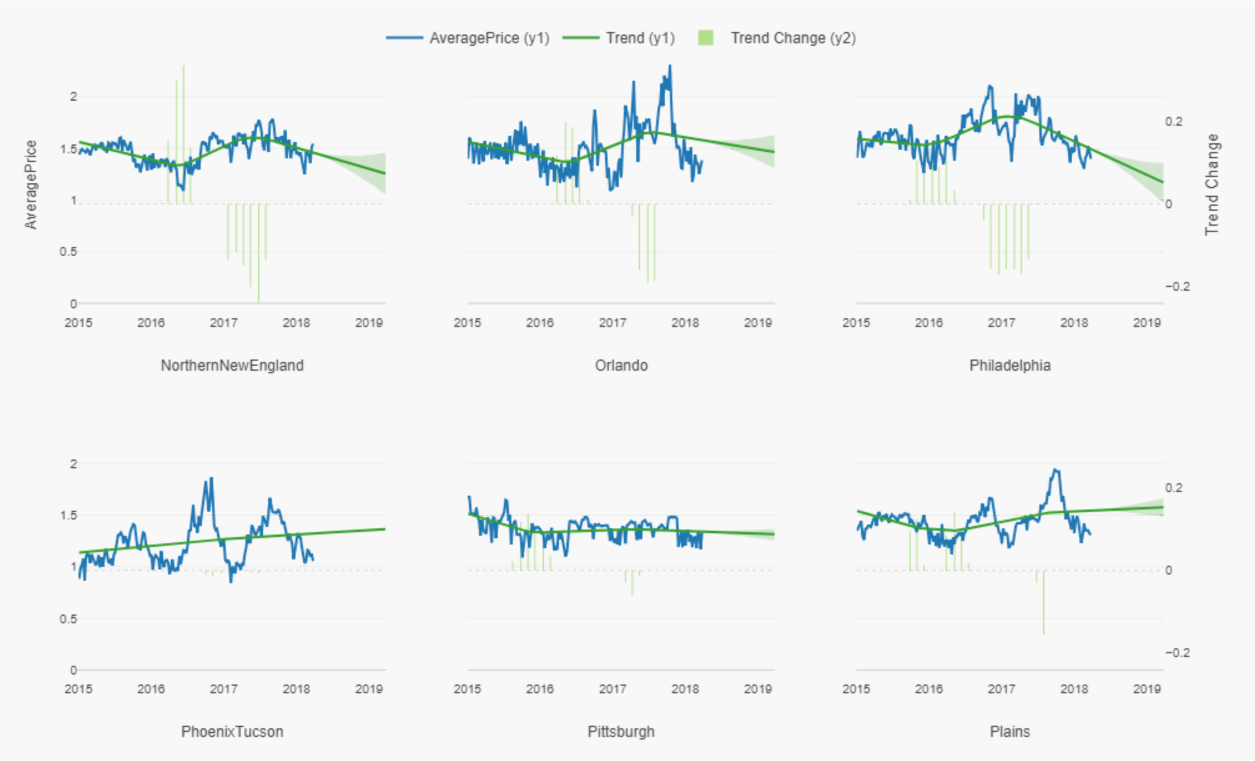


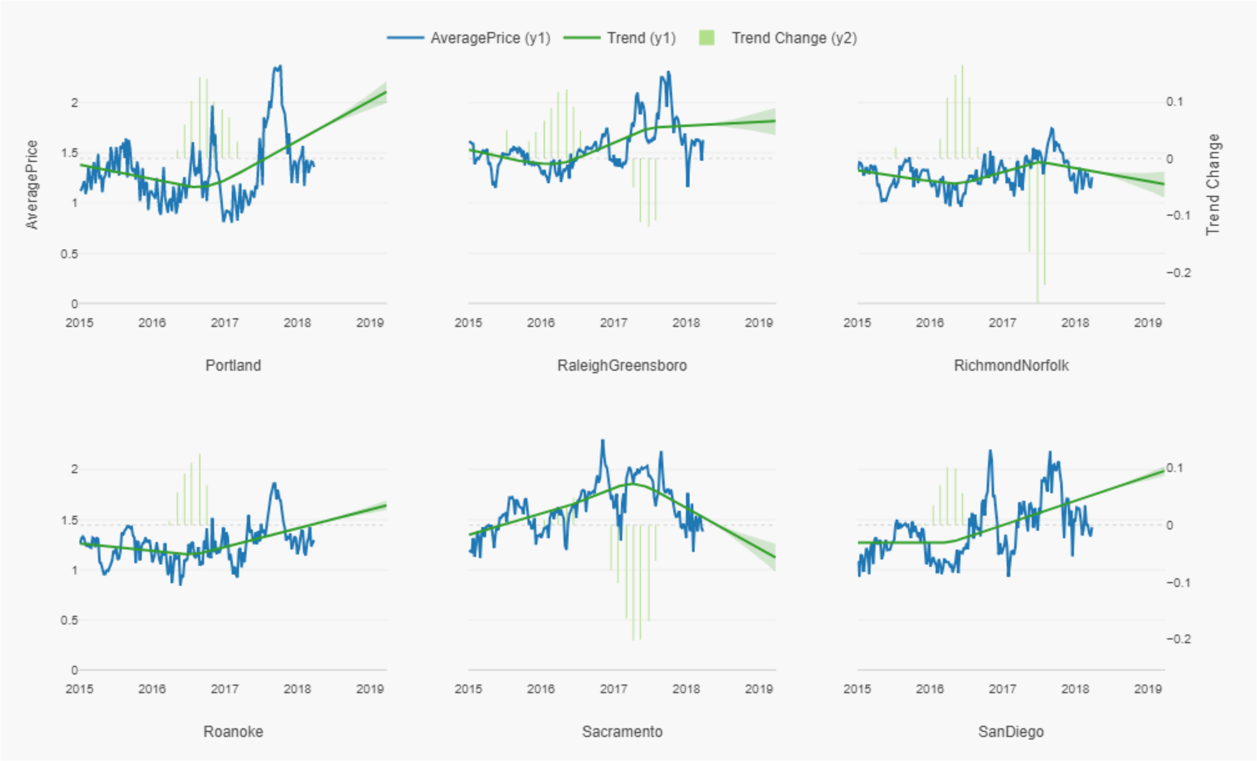


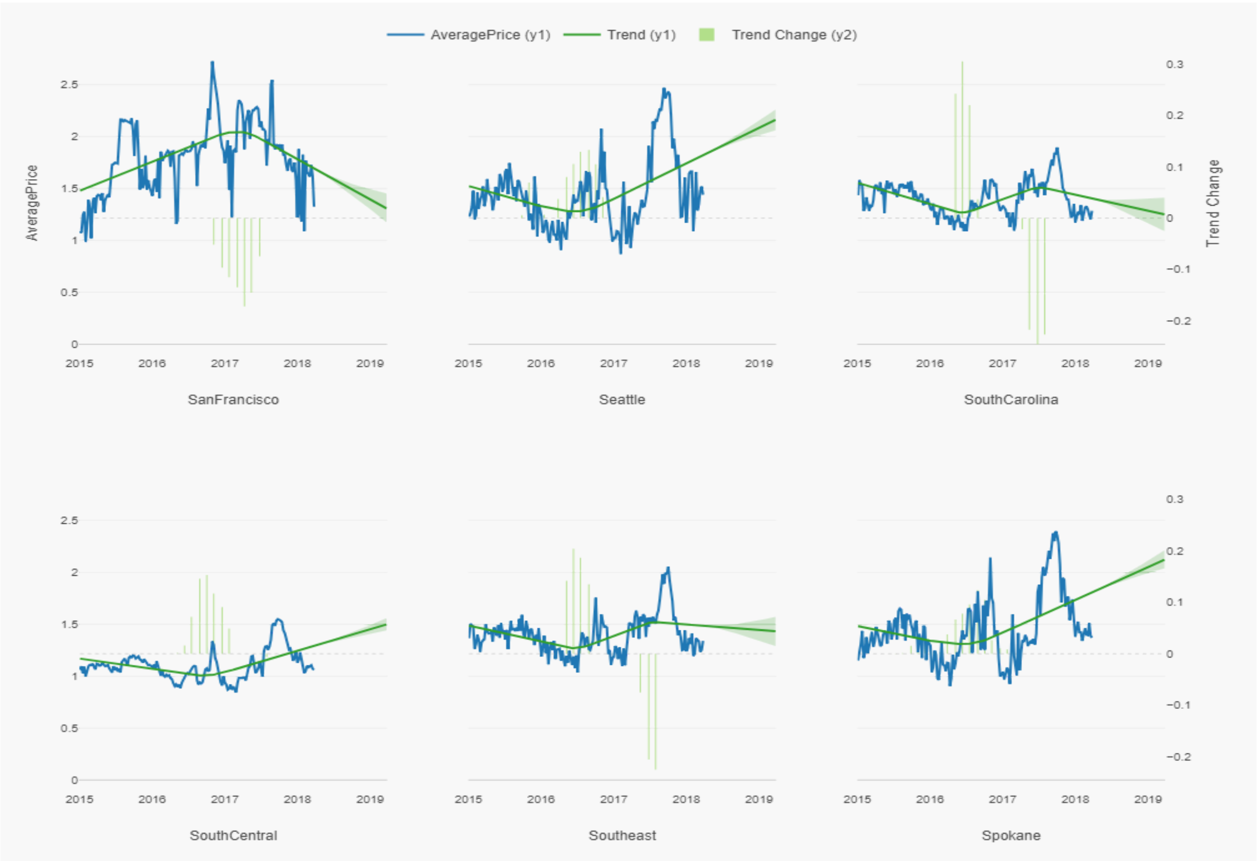

AveragePrice (y1)
Trend (y1)
🟩 Trend Change (y2)
AveragePrice
Trend Change
0
0.5
1
1.5
2
0.1
0
−0.1
−0.2
−0.3
2015
2016
2017
2018
2019
Denver
Detroit
GrandRapids
GreatLakes
HarrisburgScranton
HartfordSpringfield


AveragePrice (y1)
Trend (y1)
🟩 Trend Change (y2)
AveragePrice
Trend Change
0
0.5
1
1.5
2
2.5
0.15
0.1
0.05
0
−0.05
2015
2016
2017
2018
2019
Houston
Indianapolis
Jacksonville
LasVegas
LosAngeles
Louisville


AveragePrice (y1)
Trend (y1)
🟩 Trend Change (y2)
AveragePrice
Trend Change
0
0.5
1
1.5
2
0.2
0.1
0
−0.1
−0.2
2015
2016
2017
2018
2019
MiamiFtLauderdale
Midsouth
Nashville
NewOrleansMobile
NewYork
Northeast


AveragePrice (y1)
Trend (y1)
🟩 Trend Change (y2)
AveragePrice
Trend Change
0
0.5
1
1.5
2
0.2
0
−0.2
2015
2016
2017
2018
2019
NorthernNewEngland
Orlando
Philadelphia
PhoenixTucson
Pittsburgh
Plains


AveragePrice (y1)
Trend (y1)
🟩 Trend Change (y2)
AveragePrice
Trend Change
0
0.5
1
1.5
2
0.1
0
−0.1
−0.2
2015
2016
2017
2018
2019
Portland
RaleighGreensboro
RichmondNorfolk
Roanoke
Sacramento
SanDiego


AveragePrice (y1)
Trend (y1)
🟩 Trend Change (y2)
AveragePrice
Trend Change
0
0.5
1
1.5
2
2.5
0.3
0.2
0.1
0
−0.1
−0.2
2015
2016
2017
2018
2019
SanFrancisco
Seattle
SouthCarolina
SouthCentral
Southeast
Spokane

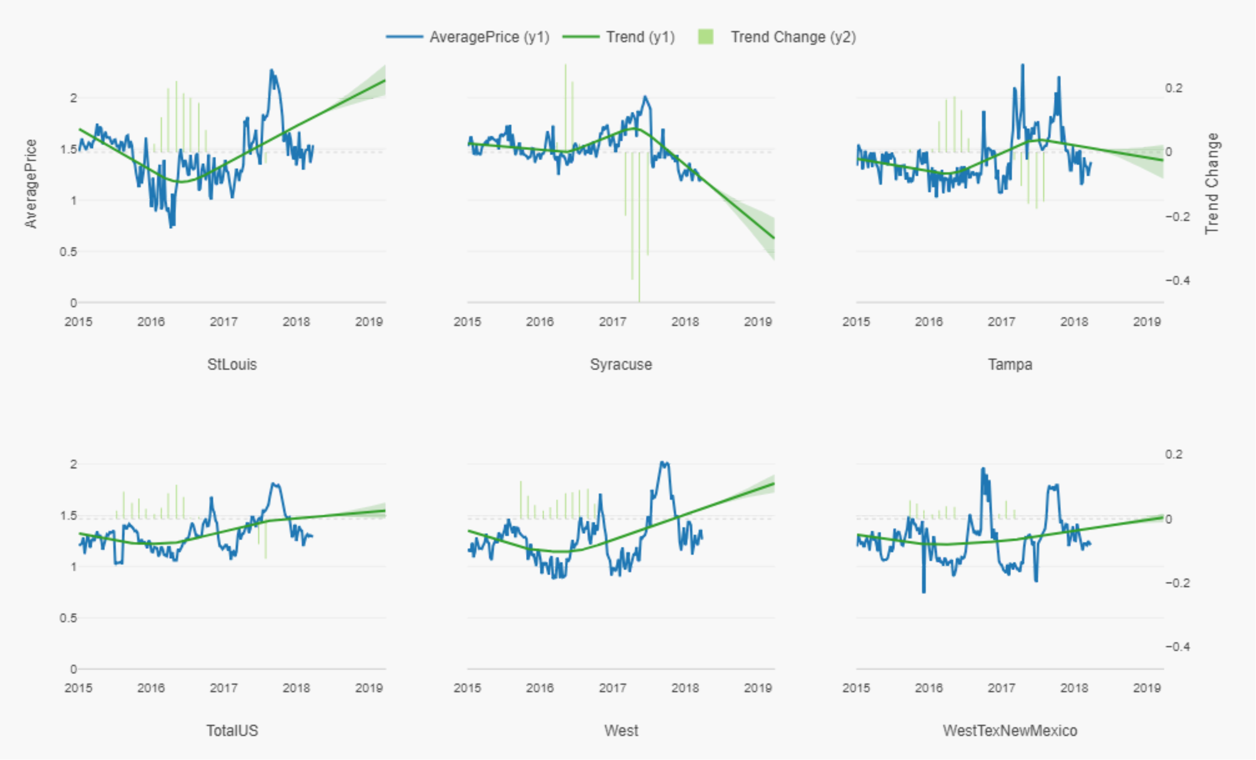

AveragePrice (y1)
Trend (y1)
Trend Change (y2)
AveragePrice
Trend Change
2
1.5
1
0.5
0
0.2
0
−0.2
−0.4
2015
2016
2017
2018
2019
StLouis
Syracuse
Tampa
TotalUS
West
WestTexNewMexico

Exhibit 17 Prediction of average avocado price for city-level market

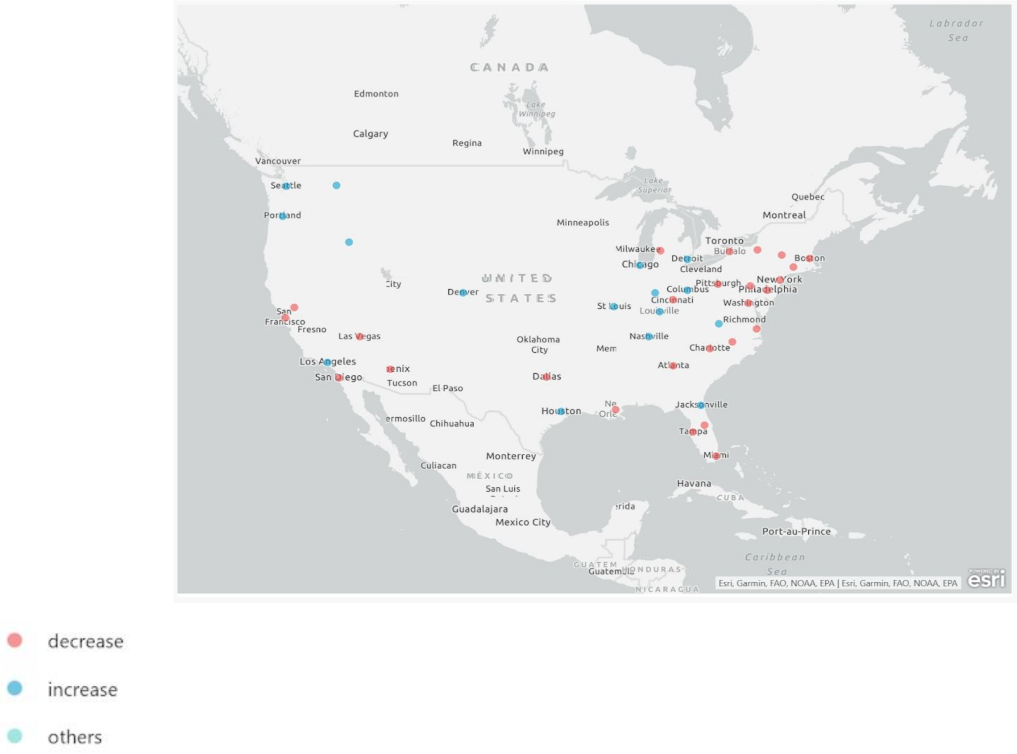


Exhibit 18 The Sale of Organic Avocado Tendency

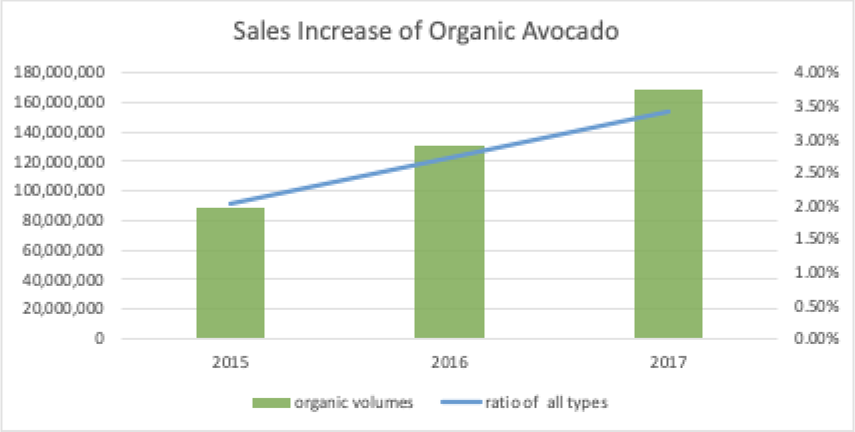
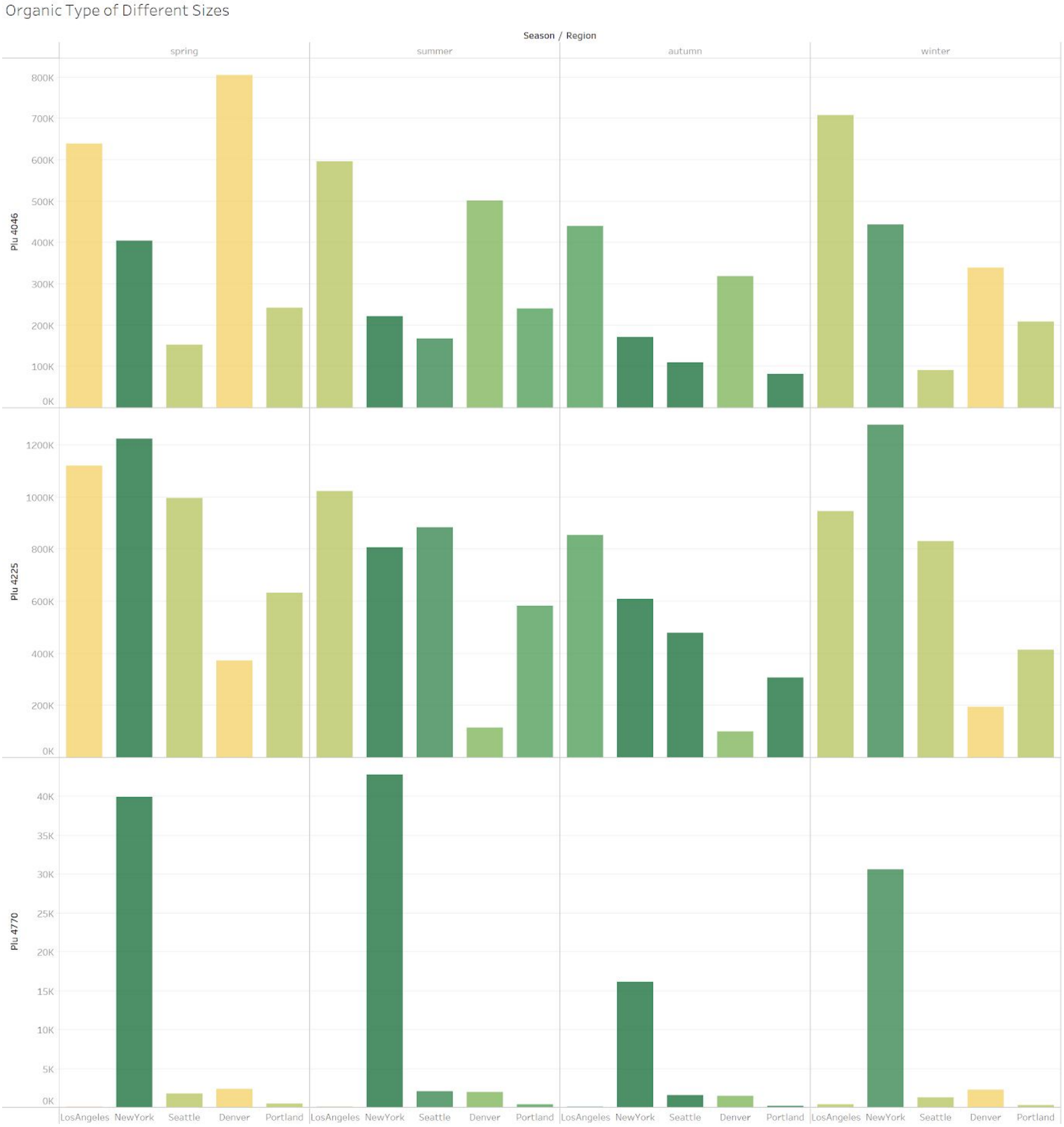


Exhibit 19 Average Price of Organic Avocado of Different Sizes Change Through Season

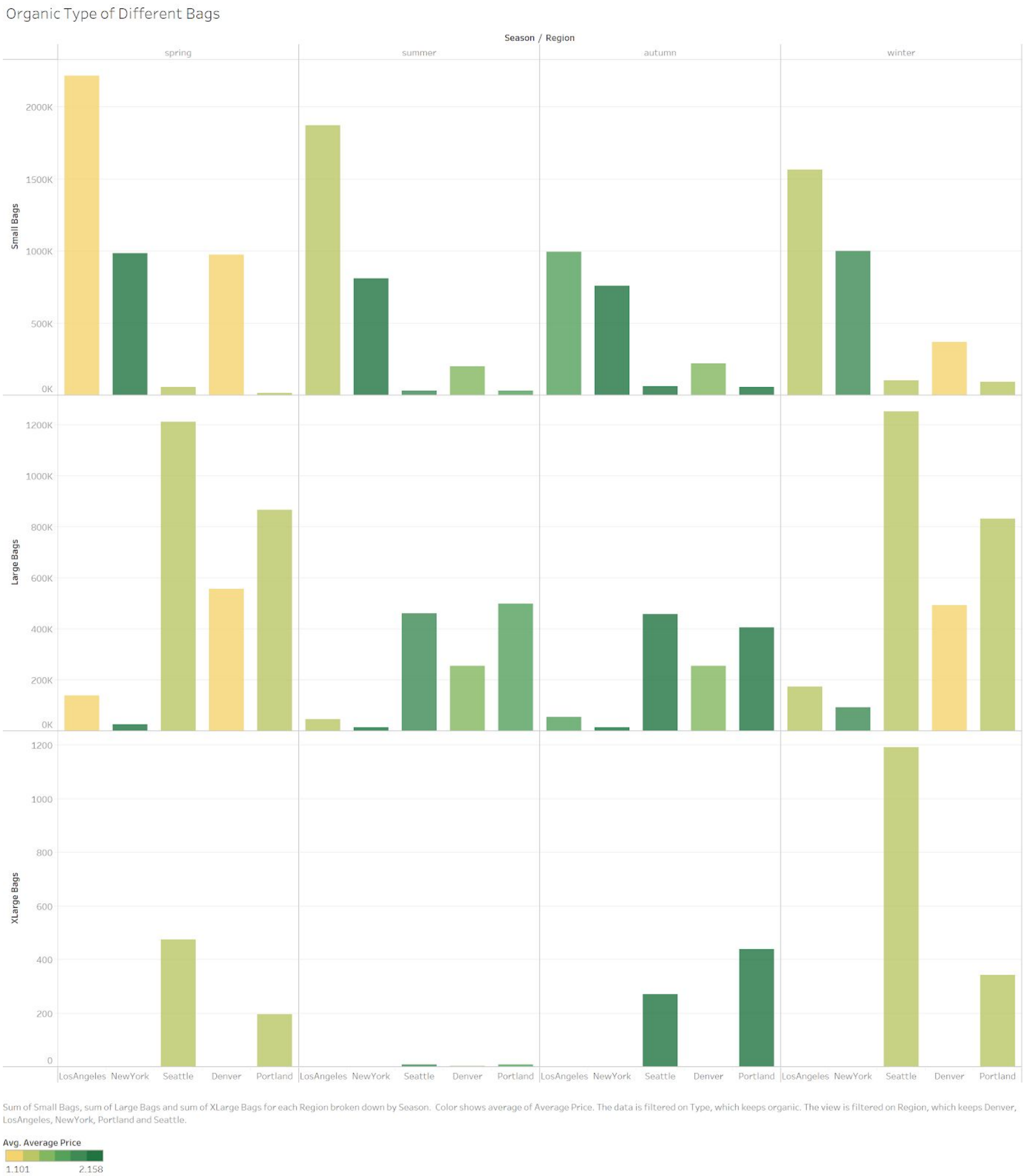


Sum of Plu 4046, sum of Plu 4225 and sum of Plu 4770 for each Region broken down by Season. Color shows average of Average Price. The data is filtered on Type, which keeps organic. The view is filtered on Region, which keeps Denver, LosAngeles, NewYork, Portland and Seattle.

Avg. Average Price

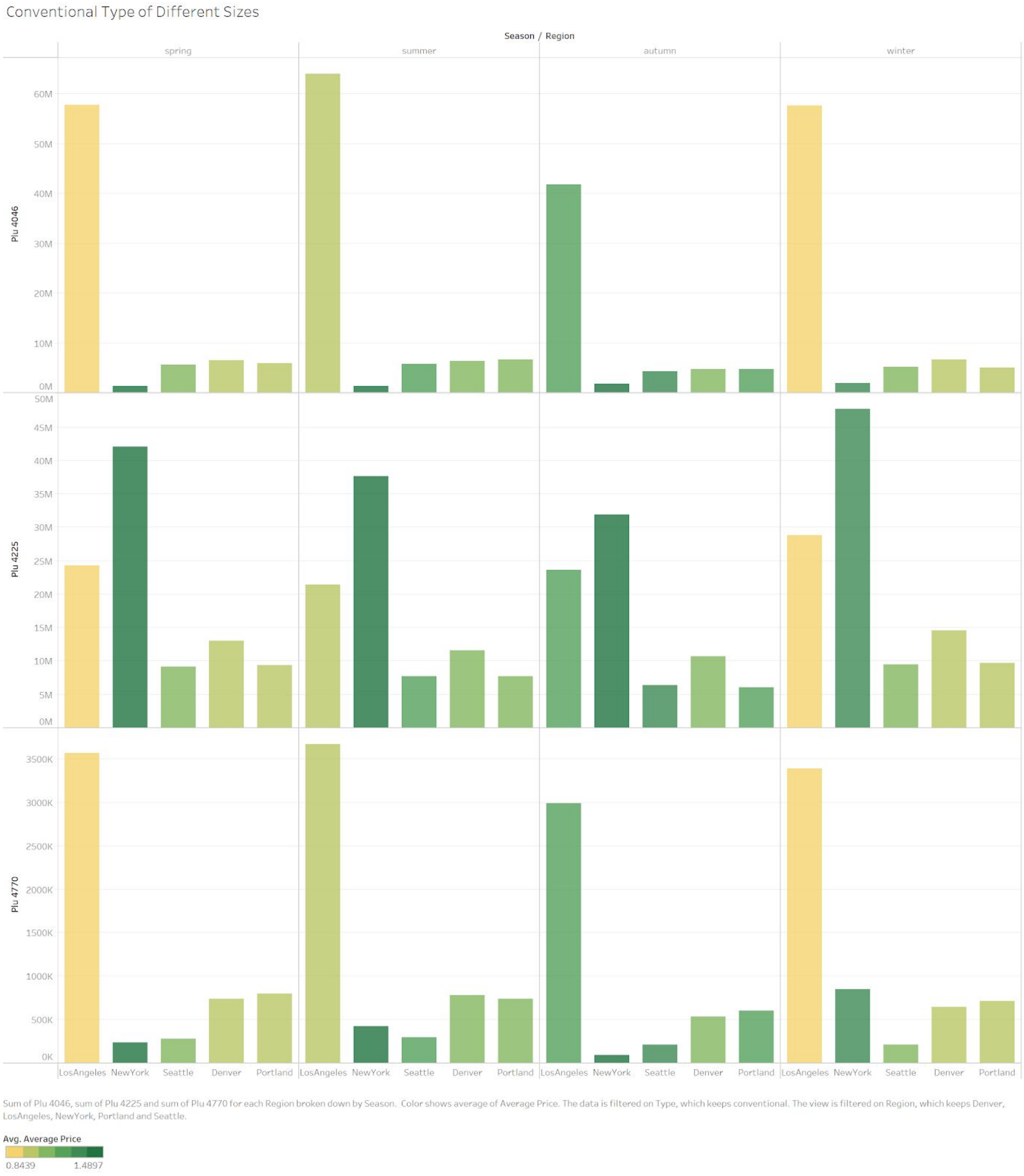
1.101 2.158

Exhibit 20 Average Price of Organic Avocado of Different Bags Change Through Season



Sum of Small Bags, sum of Large Bags and sum of XLarge Bags for each Region broken down by Season. Color shows average of Average Price. The data is filtered on Type, which keeps organic. The view is filtered on Region, which keeps Denver, LosAngeles, NewYork, Portland and Seattle.

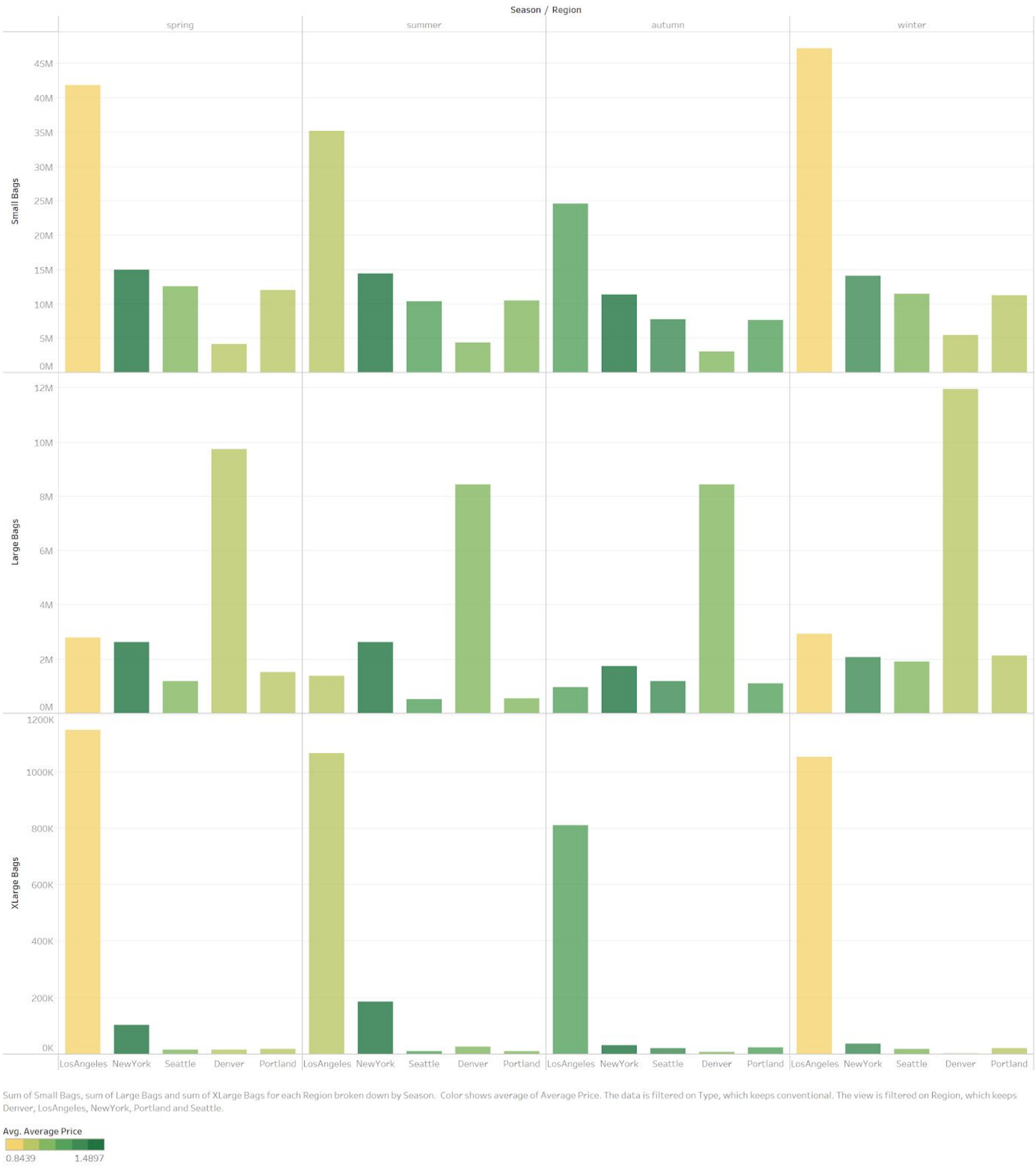
Average Price of Conventional Avocado of Different Sizes Change Through Season


Conventional Type of Different Sizes

Sum of Plu 4046, sum of Plu 4225 and sum of Plu 4770 for each Region broken down by Season. Color shows average of Average Price. The data is filtered on Type, which keeps conventional. The view is filtered on Region, which keeps Denver, LosAngeles, NewYork, Portland and Seattle.

Average Price of Conventional Avocado of Different Bags Change Through Season

Conventional Type of Different Bags



Sum of Small Bags, sum of Large Bags and sum of XLarge Bags for each Region broken down by Season. Color shows average of Average Price. The data is filtered on Type, which keeps conventional. The view is filtered on Region, which keeps Denver, LosAngeles, NewYork, Portland and Seattle.