How much will customers pay for their Avocados?

EDA and Time Series Forecasting of Avocado prices

DS 5110

Introduction to Data Management and Processing

Authors:

Amey Shankar Basangoudar

Priyesh Priyesh

Riddhi Narayan

Saachi Chandrashekhar

EDA and Time Series Forecasting of Avocado prices

Authors: Amey Shankar Basangoudar, Priyesh Priyesh, Riddhi Narayan, Saachi Chandrashekhar

Summary:

The United States is one of the largest consumers of avocados in the world. But still the consumption of avocados in the US is lower than that of other products such as apples, bananas, lemons, cranberries, etc. Thus there is a huge growth potential in the volume of consumption of avocados. Moreover, most of the avocados sold in the US are not produced here. They are mainly exported from Mexico and Chile. Hence the price of avocados is subject to continuous change in the agriculture market.

The project aims to perform a thorough analysis of the volume of avocados sold and the sale of avocados in the country by recognising patterns and performing time series analysis on the data. The dataset that we used consists of 13 attributes such as 'Date', 'Average Price', 'Region', 'Total volume' etc. The dataset was downloaded in May of 2018 from the Hass Avocado Board. The retail scan data is directly sourced from the cash in the retailer's cash registers, i.e, the cash generated from the sale of Hass Avocados.

This project has two main goals. Firstly to build a predictive model that predicts the Average Price of Avocados. The prediction is done using Linear Regression and Random Forest Regression. The efficacy of both models is compared using coefficient of determination (R squared values), root mean squared error (RMSE) and Mean Absolute Values (MAE). Next we perform time series analysis to check the trends with respect to time and seasons and accordingly perform forecasting using ARIMA.

Methods:

The first step was to import all the libraries. Some of the important libraries used throughout the project were tidyr, ggplot2, dplyr, and modelr. The dataset was loaded into RStudio using read_csv() and thoroughly checked for missing and null values. After no null or missing values were found, the dataset was converted into a tidy format. The dataset originally had values as column names. This is not a tidy format and hence the columns were converted into values using the tidyr function pivot_longer(). Figure-1 shows the variables present in the tidied dataset.

avocado.csv

1) Date
2) Average Price
3) Total Volume
4) Total Bags
5) Type
6) Year
7) Region
8) '4046'
9) '4225'
10) '4770'
11) Small Bags
12) Large Bags
13) XLarge Bags

Figure-1: Variables in the dataset

'4046', '4225' and '4770' are the PLU types. PLU stands for Price Lookup Code. The PLU number refers to the avocado produced based on commodity, variety and size group. PLU 4046 refers to non-organic small/medium Hass Avocados (~3 - 5 oz), PLU 4225 refers to non - organic large Hass Avocados (~ 8 - 10 oz) and PLU 4770 refers to non-organic extra large Hass Avocados (~ 10 - 15 oz). The avocados were sold in three kinds of bags namely, small bags, large bags and extra large bags. Figure-2 shows the volume of avocados sold in each kind of bag between the years 2015 - 2018.

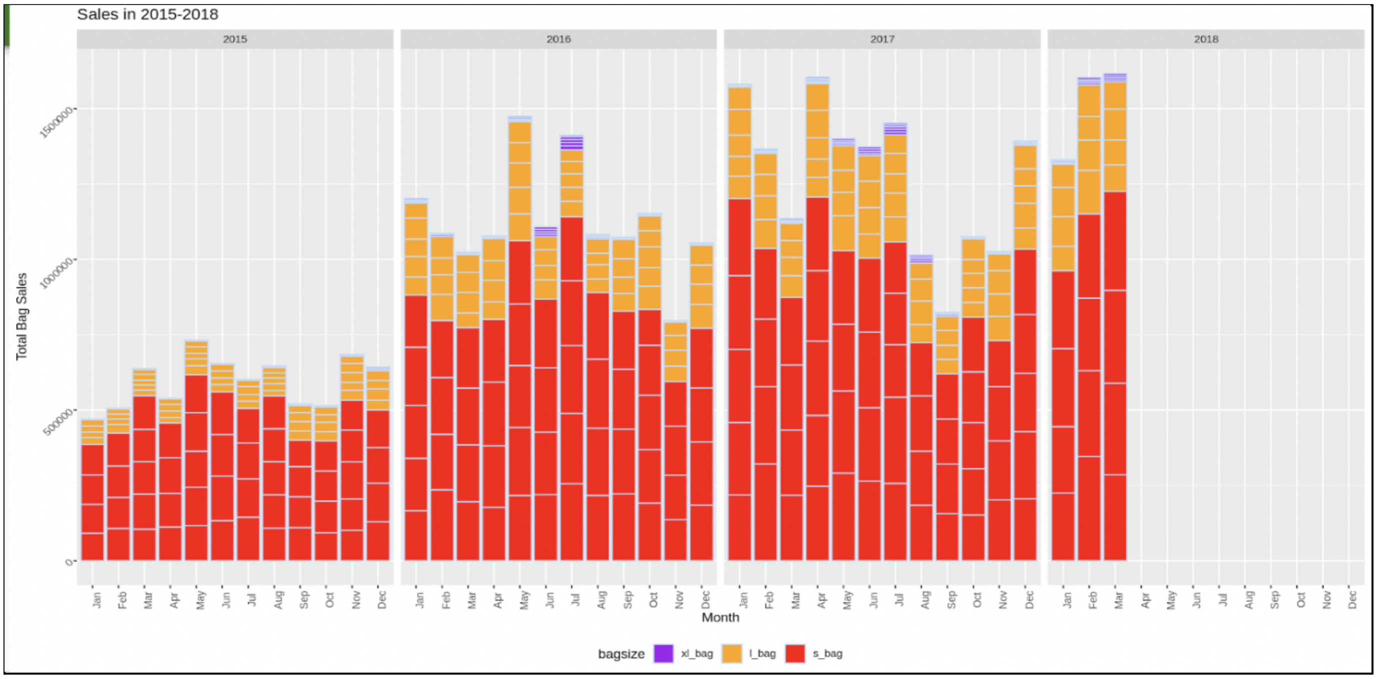


Figure-2: Volume of avocados sold in each kind of bag from 2015-2018

Next exploratory data analysis was performed on the tidied dataset. First it was found that the dataset had two types of avocados namely, organic and conventional. It was seen that the average price of organic avocados was larger than that of conventional avocados. Also the sale of avocados across a lot of regions in the United States was visualized. It was found that California had the highest sales in terms of volume of avocados sold and Syracuse had the least (refer Appendix Fig 2). It was also seen that Hartford had the highest average price of avocados whereas Houston had the lowest (refer Appendix Fig 3). For the city of Boston, it was found that people preferred buying conventional avocados over organic owing to the cheaper rates of the conventional variety (refer Appendix Fig 4 and 5). Figure-3 shows the choropleth for the change in average price of avocados across the United States in the years 2015 - 2018. The grey areas are where data was not available.

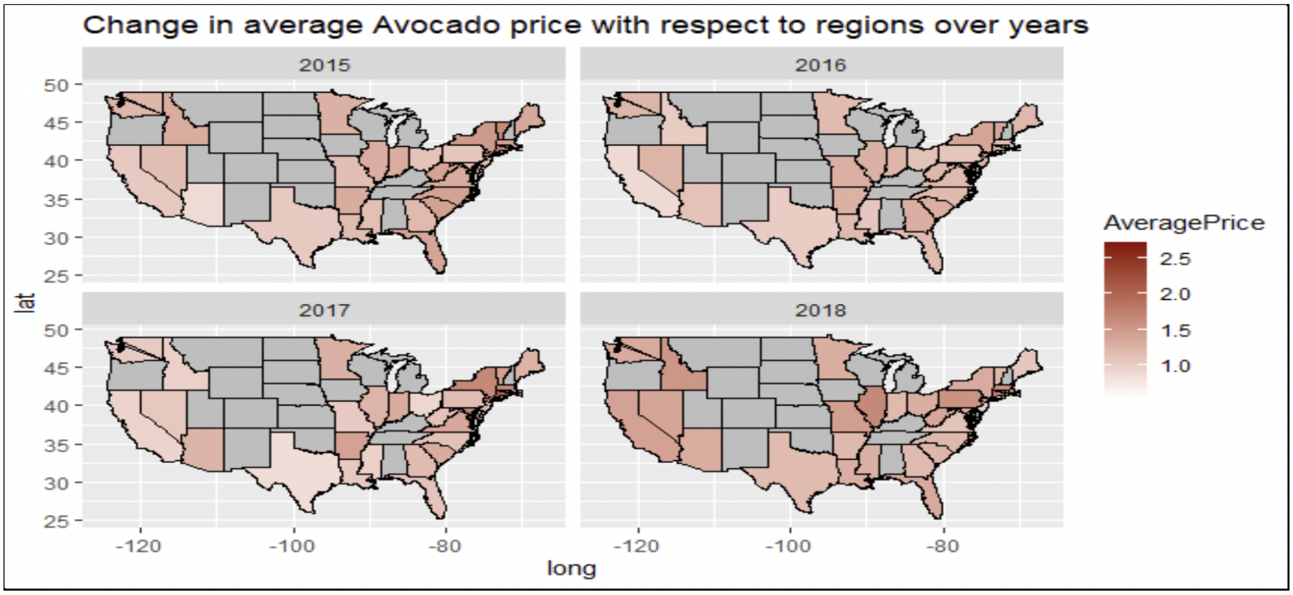


Figure-3: Chloropleth showing change in avocado prices from 2015-2018

The average price of avocados in the dataset are seen to vary with respect to time. Hence it is important to perform time series analysis to check for trends and seasonality. Figure-4 shows the time series analysis plots. The trend shows that change in average price is non-linear, it shows a decrease in the average price in 2015, followed by an increase in the trend which tends to flatten up by 2018. Some repeated crests and troughs are seen in the plot, which shows that the average price value changes with some repeated pattern wherein some months have higher prices than others. This justifies that the average avocado prices are seasonal. The seasonality in average price could justify the downtrend in prices in 2015. Prices are highest in the fall season and then keep falling as winter approaches (refer Appendix Fig 6).

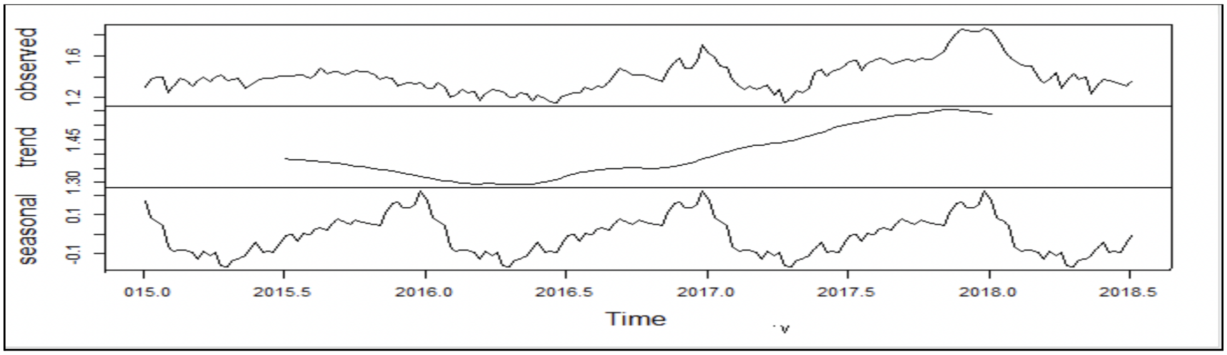


Figure-4: Time series analysis

For the modelling Linear Regression and Random Forest Regression were used. Linear Regression is a model that assumes a linear relationship between the input variables (x) and the single output variable (y). The most linear fit for the model is found and that fit is used to predict the output. Random Forest operates by constructing a multitude of decision trees at training time and outputs the mean prediction of the individual trees. At each level the tree will split based on which feature is giving the most information.

For linear regression the best linear fit is found. The best fit is the one that can predict the avocado price accurately. For that, each variable is plotted against the average price and transformations are applied wherever required. Next, residuals and QQ plots are used to find the best variables. Also stepwise AIC and RMSE functions are used to see which variables should be added to the fit.

Finally we perform time series forecasting using ARIMA. The initial step was to split the dataset in accordance with the two types of avocados, namely - conventional and organic and check if the data for each of these two subsets was stationary. Through our time series analysis, we can confirm that our data has seasonality.

Time series forecasting cannot work with seasonal data and thus, we perform the first differencing method in order to remove seasonality and make our data stationary. In order to validate whether the two subsets of the data are stationary, the KPSS unit root test is performed. The p-values generated from this test for the two subsets are 0.1 (>0.05). In addition, the residual plots show that all the lags lie between the two confidence intervals and proves that the residuals are not autocorrelated. Through these results, we can confirm that the data is stationary and ready to be used for our ARIMA model.

Results:

- **Linear Regression:**

In the linear regression model, it is seen that there is a significant difference in the actual and predicted values, some price values even going to negatives. Figure-5 shows the stepwise RMSE curve that shows the final fit of the linear regression model. The variables used to fit the model were - Type, region, volume, PLU type, bag_size.

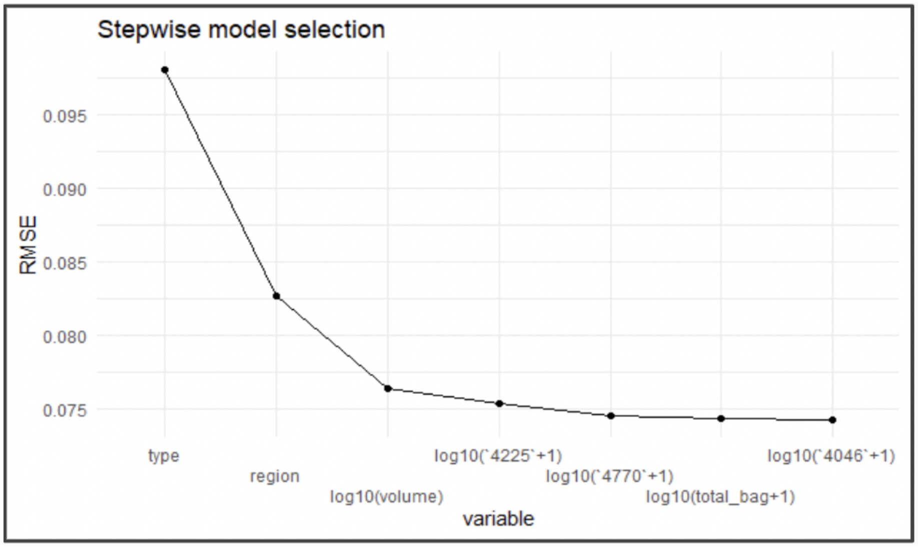
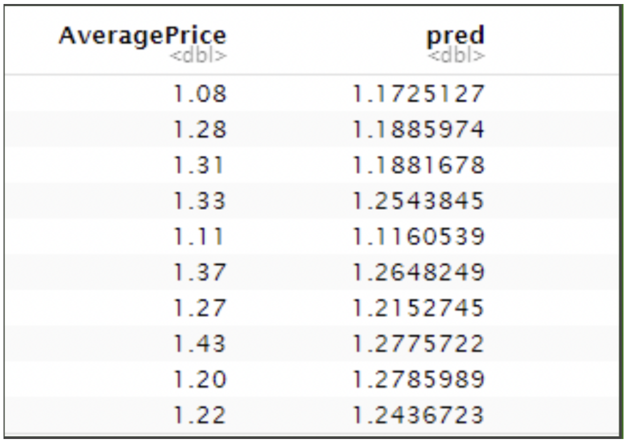


Figure-5: Stepwise model selection using RMSE

- **Random Forest Regression:**

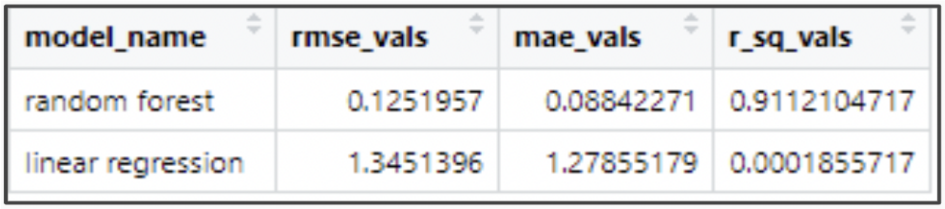
Random Forest model was able to give us a minimum difference in the actual and predicted values and we can say that the model is able to discover more complex dependencies and can predict better when the relationship of the variables are a little messy and are not as linear. Figure-6 shows the Actual Average price of avocados and the predicted average price using Random Forest..



AveragePrice <dbl>	pred <dbl>
1.08	1.1725127
1.28	1.1885974
1.31	1.1881678
1.33	1.2543845
1.11	1.1160539
1.37	1.2648249
1.27	1.2152745
1.43	1.2775722
1.20	1.2785989
1.22	1.2436723

Figure-6: Actual and Predicted Average Price of avocados

For a better understanding of which of the two models is better, we calculate the RMSE, MAE and R-squared values for both models. For a good model, we know that RMSE, MAE should be less whereas the R-squared value should be high. Figure-7 shows the computed values of RMSE, MAE and R-squared for both linear regression and random forest regression. It is clearly seen that the RMSE and MAE value for the random forest model is much lesser than that of linear regression and the R-squared value is higher for random forest. Hence we can say that random forest would give us the best result from the two when we want to predict avocado price.



model_name	rmse_vals	mae_vals	r_sq_vals
random forest	0.1251957	0.08842271	0.9112104717
linear regression	1.3451396	1.27855179	0.0001855717

Figure-7: RMSE, MAE and R-Squared values of both models

- **ARIMA:**

Two separate ARIMA models are used for the two time series objects created. The first- for the 'conventional' type of avocados and the second for the 'organic' type of avocados. Since the data was transformed using the first difference, the models are configured such that the seasonal (d) and non-seasonal (D) parameters are set to 1. Forecasting is performed for the next 3 years, i.e through 2018-2021, and hence, the (h) parameter is set to 36. Figure-8 shows the forecasted average prices of avocados for conventional and organic types of avocados.

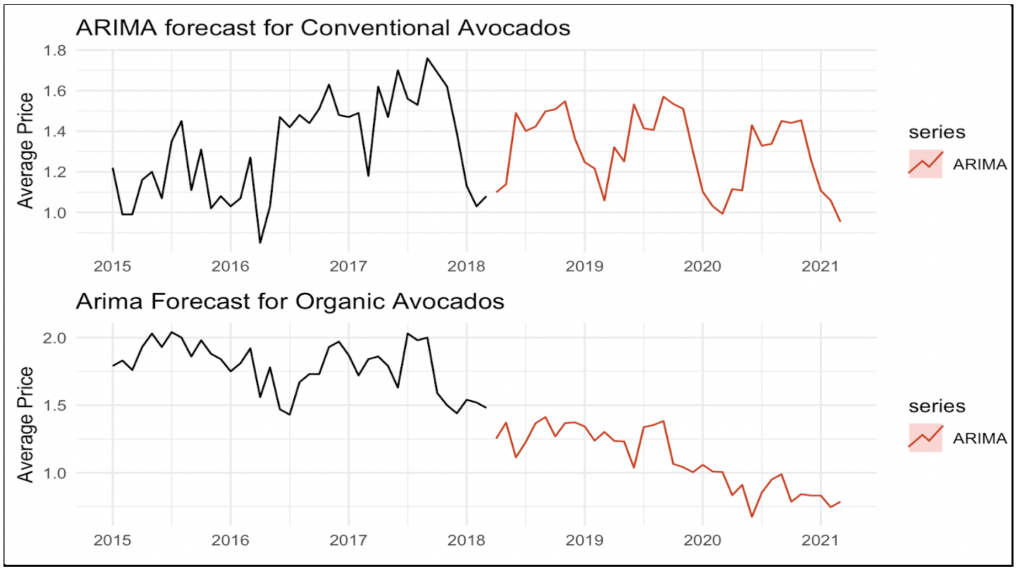


Figure-8: ARIMA Forecasting for Conventional and Organic Avocados

The conventional avocados seem to deviate between the average price ranges of $1- $1.6, with a jagged trend through the years. This trend may be due to the fluctuating demands for such avocados. Certain seasons might not favour the growth of good quality avocados which might result in lesser demands during these times.

The organic avocados observe a downward trend, with their price ranges being below $1.5. One of the main reasons for this downward trend may be because of the increase in volume of organic avocados sold over the years 2015-2018. It thus makes sense for the model to be predicting lesser prices because cheaper rates facilitate a higher demand for such avocados.

Discussion:

From the project it was found that predicting the average avocado price can be seen as a regression problem and as a time series problem where Random Forest and ARIMA models gave us promising results. It can be inferred that the price of avocados has seasonal changes, where the prices are the highest in Fall and gradually decrease thereafter.

The models can be used as a baseline to try and figure out where in the country should the prices be increased/decreased depending on the consumption level of consumers in each area to help the avocado sellers and farmers better.

An idea for expanding the project would be to try more modelling algorithms to see if there’s any other model which could give a better result than Random Forest. Another improvement to the project would be to find the average price of avocados for the missing regions in the dataset and then plot a comprehensive choropleth for the entire United States.

Statement of Contribution:

Amey Shankar Basangoudar - Data Tidying and EDA
Priyesh Priyesh - Time series Analysis
Riddhi Narayan - Time series Forecasting using ARIMA
Saachi Chandrashekhar - Modelling using Linear Regression and Random Forest

References:

[1] Dataset: https://www.kaggle.com/neuromusic/avocado-prices

[2] Modelling: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[3] Time series forecasting: https://otexts.com/fpp2/arima-r.html

Appendix:

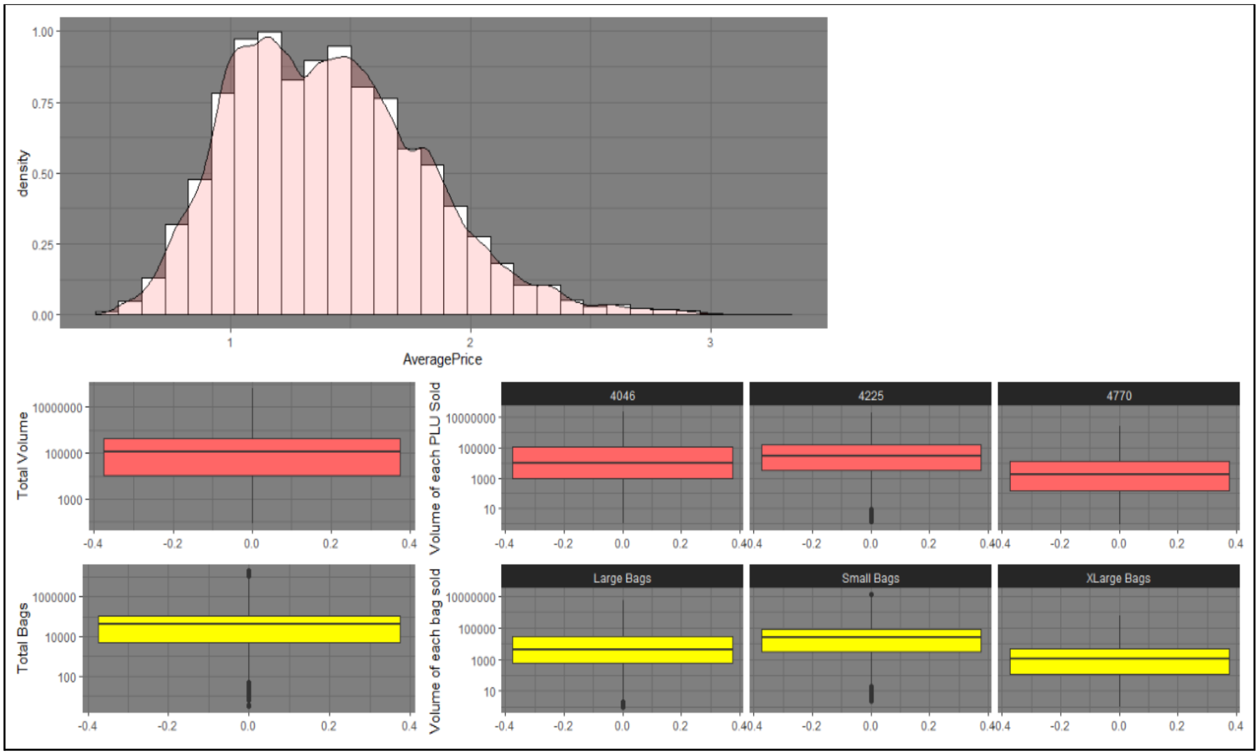


Fig 1: Distribution of dataset variables

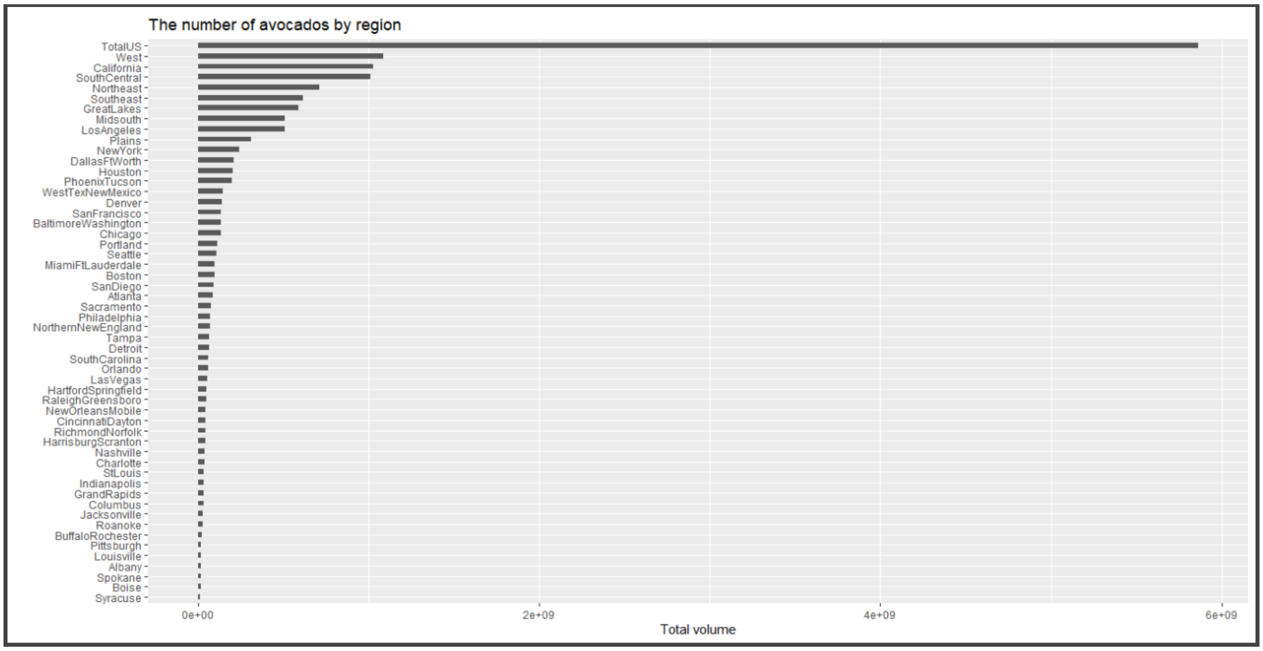


Fig 2: Avocado volume per region

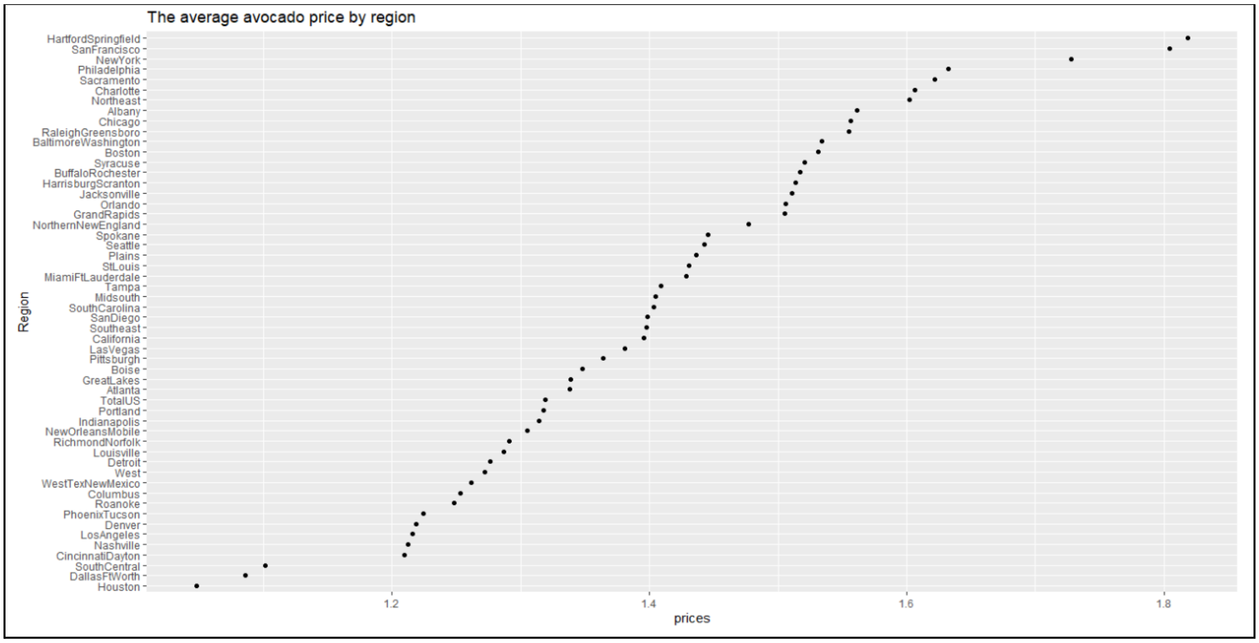


Fig 3: Avocado price per region

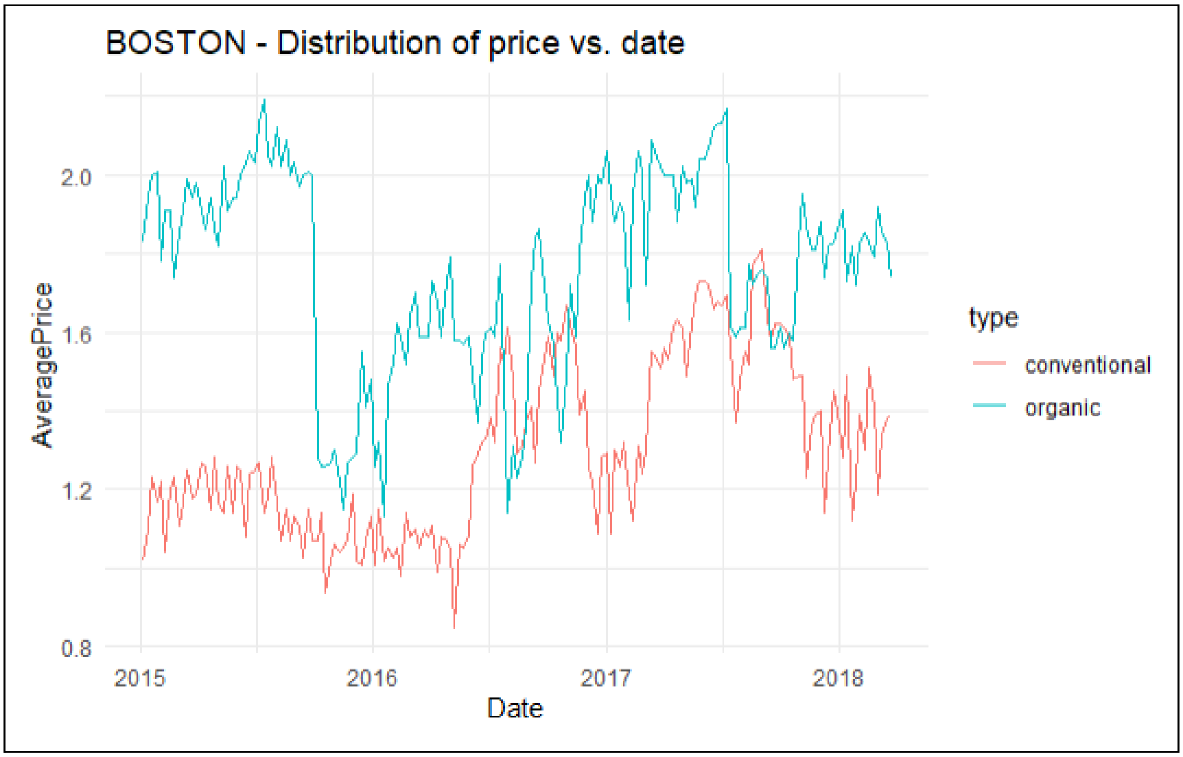


Fig 4: BOSTON - Avocado price distribution over time

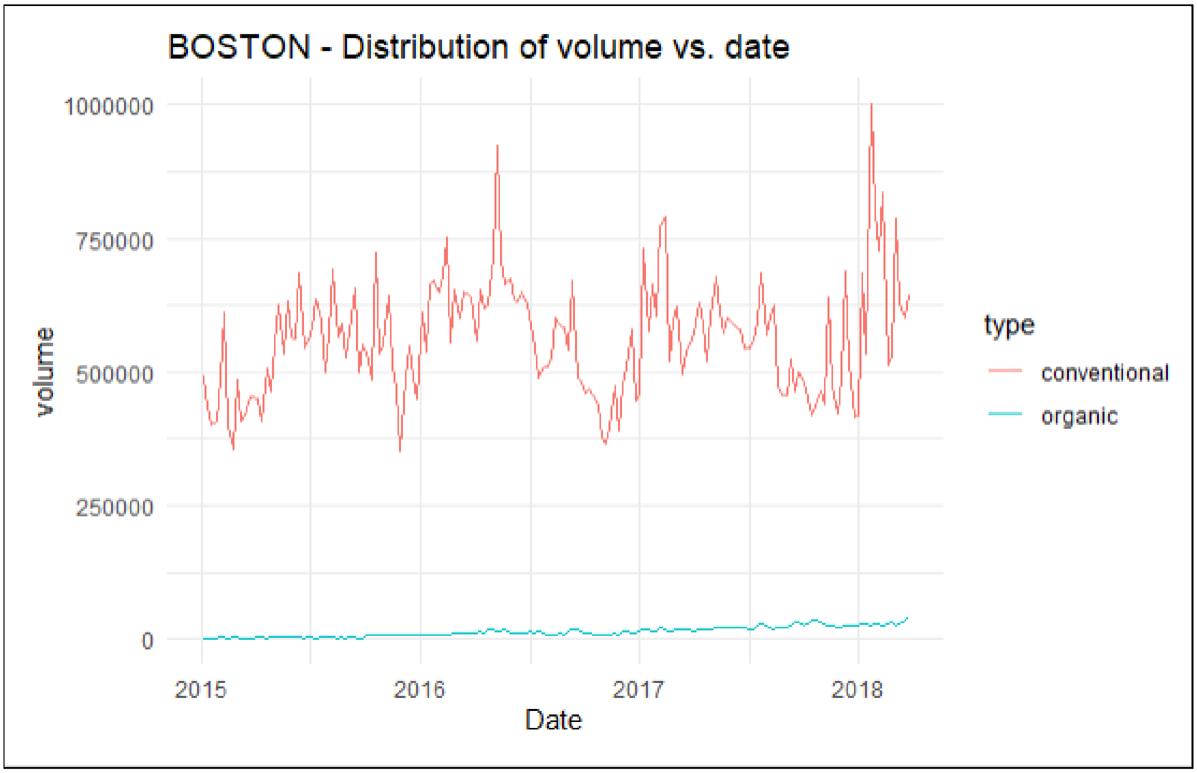


Fig 5: BOSTON - Avocado consumption over time

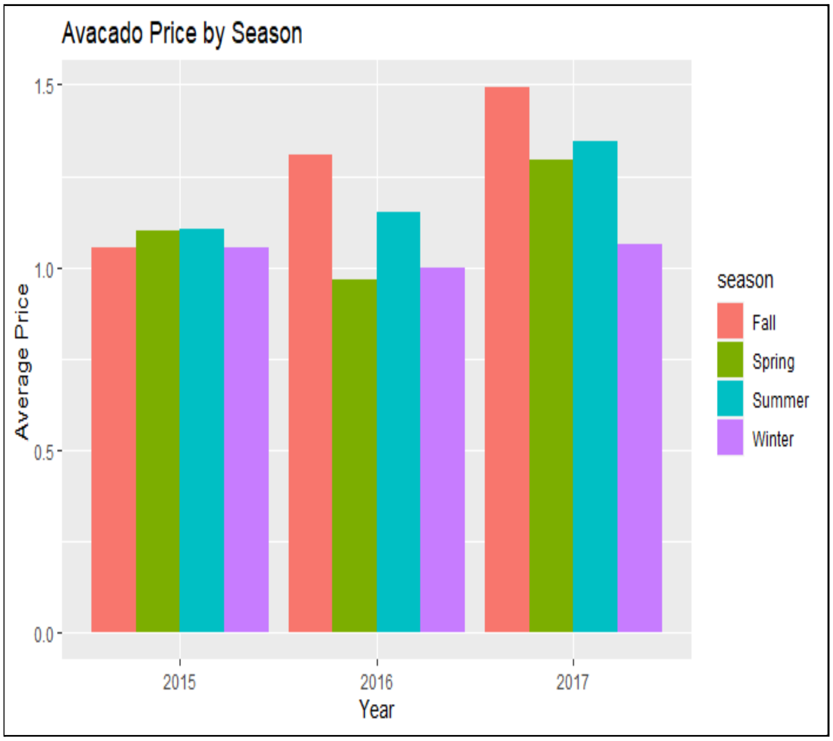


Fig 6: Avocado price by season