Prediction System on price and consumption of Avocados

Yunze Qiu UNI: yq2310
Yiran Lin UNI: yl4628
Stacy Lai UNI: sl4450

Abstract

Avocados have become popular in today's market. As a result, the avocado industry is increasing at a quick pace and more and more farmers are growing this product. In this project, the goal is to build a system which can predict the price and consumption of avocados in certain markets as well as to generate a visual data report by analyzing the marketing data of avocados to help farmers make the best profits in sales of avocados.

We trained 6 different regression models using 18,000 historical data of avocados' prices and consumptions and compared their performance on the test set. Finally, we picked the SVM model which outperformed other models to build the prediction system.

The prediction system can predict the price, consumption, and revenue giving users' inputs including the type of avocado, the month of the year, and the region where they want to make the sales. The prediction system would also give a suggestion on whether to do the sale this month or one month earlier or later.

In addition, we also create a visualization report of all historical data where users can see how price and consumptions change according to the region and time.

1. Introduction

Increasing trend of avocado has hit a record high. According to Statista(2021), the global value of avocado has increased from 12.82 billion dollars in 2019 to 14.33 billion dollars in 2021, and the growth is projected to continue increasing at a steady rate until 2025 to 17.91 billion dollars. This rapid growth of consumption levels and retail sales is due to heavily industry-funded research of marketing campaigns changing the image of avocado as one of the superfoods with exceptional nutrient density(Carman 2019).

In this paper, our aims are to build a system to predict the price as well as the consumption level of avocados in the U.S. by utilizing machine learning algorithms to train a model from the existing dataset that we find on Kaggle.

2. Related Works

Previous work has been done predicting the sales of avocado using the weather data(Rincon-Patino, Juan & Lasso 2018). The paper utilized machine learning algorithms including Linear Regression, SVM(Support Vector Machine), Multilayer Perceptron, and Multivariate Regression to generate the Prediction Model and generalize an application to help sellers in the process of demand planning and estimate the sales revenue.

Another research has been done toward revealing the avocado buying trend in the U.S.(Jones, Keyse, Melgoza, Perez, Qamar, Villalpando, & Woo 2021). They used SAP analytics cloud(SAC) to present marketing insights of consumer buying behavior, including the most sought-after type of avocado as well as how preferences vary due to seasonal trends using a dashboard to visualize the data.

Based on the previous work that we found, we hope to bridge the gap between those two articles where we can build a system that allows users to estimate the price and consumption level of avocado as well as a visual data report analyzing the consumer purchasing behavior in the U.S.

3. Data

3.1 Dataset

The data we are using for this project is an existing dataset named ***Avocado Prices*** which is publicly available on Kaggle. This dataset records the historical data on avocado prices and sales volume in multiple US markets and was created 3 years ago by collecting data directly from the Hass Avocado Board website and then compiled into a single CSV.

volume: The size of the dataset is about 2 MB, and there are 18,200 records with 14 features. Part of the features is given below.

- `Date` - The date of the observation
- `AveragePrice` - the average price of a single avocado
- `type` - conventional or organic
- `year` - the year
- `Region` - the city or region of the observation
- `Total Volume` - Total number of avocados sold
- `4046` - Total number of avocados with PLU 4046 sold
- `4225` - Total number of avocados with PLU 4225 sold
- `4770` - Total number of avocados with PLU 4770 sold

Figure 1: List of relevant features

Among these features that we are concerned about, Date, type, year and region are categorical or discrete data while AveragePrice, TotalVolume, 4046, 4225, and 4770 are numerical and continuous features.

velocity: Although the dataset was created 3 years ago, the data itself was generated very quickly. In the dataset, the records of prices and

sales are created on a 1-week basis with more than 50 different markets across the US.

variety: The dataset does not contain different types or formats of data as it is an existing dataset in the CSV file.

3.2 Feature Selection

The dataset would be used in two parts of our project. In the first part, the dataset would be used to build the historical data visualization system. In another part, the dataset would be used to train machine learning models to implement a prediction system on avocado's price, consumption and gross revenue.

The dataset is a CSV file therefore we decided to use **pandas** in python to process our data.

The 14 features are not all useful to the goal of our project so we firstly filter out some irrelevant features manually like the index of data and the number of avocados in bags sold.

The new data now contains 9 features after manual selection including **Date, AveragePrice, Type, Year, Region, 4046, 4225, 4770**. The last three features are the total number of avocados sold with the specific product lookup codes for three different types of avocados.

The Date feature is also converted into day and month.

The data with these selected features would be analyzed to be used in the data visualization system.

3.3 Data Processing

As mentioned above, the dataset would also be used as training data for our machine learning models to predict the avocado's price and consumption. Therefore, data processing is necessary to make the data usable to train the model.

Instead of using all features we selected from 3.2, we decided to drop the last three features which are the total numbers of avocados sold with the specific product lookup codes. The reason for removing the product lookup codes is the target user of the prediction system is farmers who grow avocados while the product lookup codes are not decided by them but the supermarkets according to commodity, variety, and size group of the avocados. Therefore, the user's inputs should not include such information.

Apart from removing the three product lookup codes in the dataset, we also remove the year and day from the Date feature and only keep the month feature. The logic behind this operation is that the data is actually not recorded day by day but periodically in a 7-days routine. Therefore, the data itself is not sufficient to train the model

to predict with a specific day of a year. In addition, while the dataset contains historical data of price and consumption of avocados in a large variety of regions, the dataset only contains records in three years which is from 2015 to 2018. Therefore, instead of using year as a feature in training, we decided to use month which provides more variety and makes more sense to user's concerns as we expect users who are farmers care more about which month in a year instead of which year to sell the avocados is the best option considering the farmers would always sell their products every year.

After filtering and converting the features again, we still need to perform processings to the dataset as the dataset contains both continuous and discrete features. To make all features compatible with each other for training purposes, we use a one-hot vector to encode the categorical data.

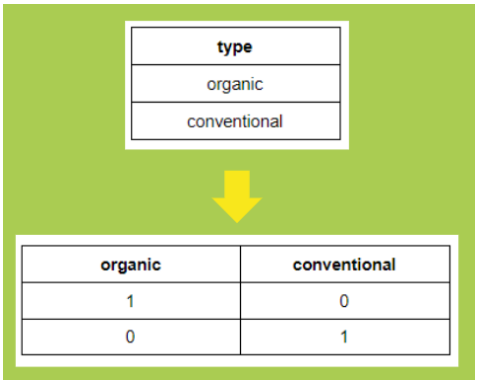


Figure 2: an example of one-hot vector

As mentioned above, there are type, region, and month that are categorical features in the dataset. After applying a one-hot encoder, the type which was either 'organic' or 'conventional' is converted into a vector with two columns. With the same process, the month feature is converted into a vector with 12 columns that each column represents a specific month. There are 54 regions in the original dataset so that region feature is converted into a vector with 54 entries.

4. Methods

4.1 Data Visualization

Instead of showing users the complex and boring numbers, we use charts and marks to visualize the dataset. Different types of charts can give users different information to understand the data. For example, the line chart can help users understand the changing trend of the data in a given period of time, the pie chart can help users see how much space each category has taken of the total share and a bar chart can show how different data performs compared to each other.

4.2 Regression & Machine learning

The prediction system is to predict the price or consumption of avocados based on the user's inputs of avocado's type, the month, and the region of the sales. Therefore, we are trying to

find the corresponding price/consumptions given these input features. Therefore, it can be defined as a regression problem.

Regression analysis is to find the relationships between independent variables and a response variable using statistical methods.

To implement the prediction system on avocado's price and consumption, we decided to use machine learning techniques to train a model that can take the user's input and predict results according to the inputs.

As the target of our system is to predict a value that is a regression problem, there are several regression models we can choose from. In this project, we tried 6 different regression models which are Linear Regression, Ridge Regression, Bayes Ridge, Random Forest, Support Vector Machine, and Multilayer Perceptron Regressor.

4.3 Model Selection & Evaluation

To pick out the best model for our project, we made evaluations on each model based on three metrics, Coefficient of Determination, Mean Absolute Error, and Mean Absolute Percentage Error. The coefficient of determination which is usually referred to as R2 is a score to show how well the model fits the data. The Mean Absolute Percentage Error and Mean Absolute Error are used to show how well the model performs in predicting prices and sales of avocados as well as to compare different models so that we can choose the best one.

In the case of overfitting, the data is separated into a training dataset and a validation (testing) dataset. The training dataset contains 80% of original data while the left data would be used to measure the performance of each model.

5. Experiments

5.1 Model Evaluation Results

The processed dataset is shuffled and split into a training set and a testing set. For each set, we set apart the features for training and the target values for reaching. As the goal of our project is to predict the prices and consumptions, prices and sales are picked out as the y-targets while other features are kept as X features.

The results in predicting the price of all six models in three evaluation metrics are shown in the table below. Linear Regression, Bayes Ridge, and Ridge Regression performed quite similarly to each other while Support Vector Machine and Multilayer Perceptron slightly outperform other models.

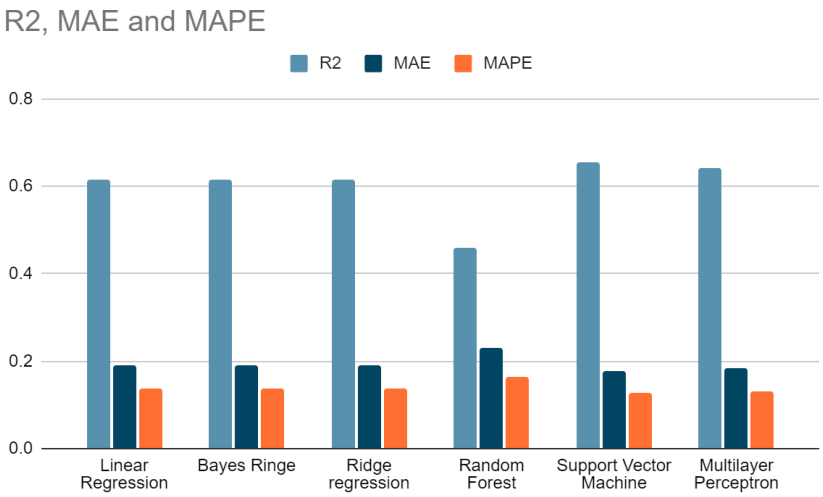


Figure 3: bar chart showing performance of 6 models

5.2 Failed experiments

Although Multilayer Perceptron also shows a good performance in predicting the price of avocados, it cannot converge in epochs (1000 epochs) when training using the consumption (total volume) as a target value. We have tried to alter the architecture of the layers using more layers and different activation functions. However, it still cannot converge even if we tune these hyper parameters.

This failure somehow alerts us that the consumption data may not be really relevant to the features we select. Indeed, the other models

```
evaluateModel(lr_model_2,'linear','consumption',X_test,y_consumption_test)
evaluateModel(R_reg_2,'Ridge','consumption',X_test,y_consumption_test)
evaluateModel(nb_model_2,'BayesRinge','consumption',X_test,y_consumption_test)
evaluateModel(rf_regr_2,'RandomForest','consumption',X_test,y_consumption_test)
evaluateModel(svm_regr_2,'SVM Regression','consumption',X_test,y_consumption_test)
```

```
The r2 score of linear for consumption prediction on test dataset is 0.5698930576909838
The MAE (mean absolute error) of linear for consumption prediction on test dataset is 1105618.176216096
The MAPE (mean absolute percentage error) of linear for consumption prediction on test dataset is 0.8035807137914431
The r2 score of Ridge for consumption prediction on test dataset is 0.5694501964539065
The MAE (mean absolute error) of Ridge for consumption prediction on test dataset is 1100362.8164143749
The MAPE (mean absolute percentage error) of Ridge for consumption prediction on test dataset is 0.8085814466052975
The r2 score of BayesRinge for consumption prediction on test dataset is 0.5690599329775681
The MAE (mean absolute error) of BayesRinge for consumption prediction on test dataset is 1100479.1940773618
The MAPE (mean absolute percentage error) of BayesRinge for consumption prediction on test dataset is 0.8092834038055075
The r2 score of RandomForest for consumption prediction on test dataset is 0.9460111661440372
The MAE (mean absolute error) of RandomForest for consumption prediction on test dataset is 455121.5845625224
The MAPE (mean absolute percentage error) of RandomForest for consumption prediction on test dataset is 0.8285730595079043
The r2 score of SVM Regression for consumption prediction on test dataset is -0.04735361617603795
The MAE (mean absolute error) of SVM Regression for consumption prediction on test dataset is 1069024.3837820285
The MAPE (mean absolute percentage error) of SVM Regression for consumption prediction on test dataset is 9.7853118540531
```

figure 4: The results of models in predicting consumptions

also perform worse in predicting the consumptions of avocados compared to their performance in predicting prices.

Therefore, we decided to still use SVM for predicting the consumptions of avocados which is the best performing model out of all 6 models.

6. System

6.1 Overview

The entire system consists of two sub-systems, a historical data visualization system and a prediction system on avocado's price, consumptions, and gross revenue based on the user's inputs.

The historical data visualization system is used to show users all historical data of prices and consumption of avocados. For convenience and better understanding, we visualize these data using different charts and marks which can provide users with enough information to check how the price and consumption of avocados vary in the past.

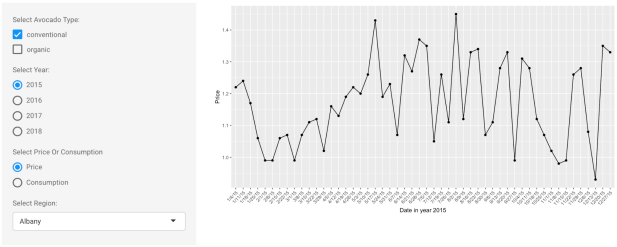
Another system is implemented to predict the avocado's price, consumption, and gross revenue given users' inputs. The gross revenue is calculated based on the predicted price and consumptions. The inputs include the type of the avocados to sell, the month the avocados are to sell, and the target market. This system can help users to determine whether it is the best time or market to sell avocados and what type of avocados can make the greatest profit.

6.2 Historical Data Visualization System

Data visualization was coded using Shiny App and it consists of four different charts. Users would see a checkbox bar on the left hand side of the webpage which is for users to control the features.

6.2.1 Historical Price/Consumption Line Chart

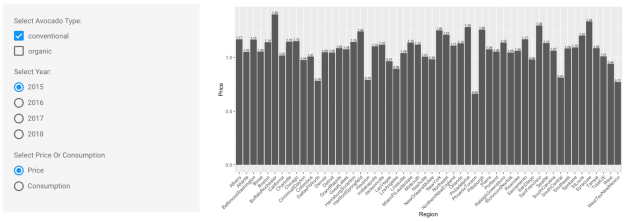
A line graph on date(x-axis) and price/consumption(y-axis) (for the full image, see appendix 1).



Four features - avocado types, years, selection on price or consumption, region - can be altered. Users could see the trend of price/consumption of a certain type of avocado in the selected year and region.

6.2.2 Regional Price/Consumption Bar Chart

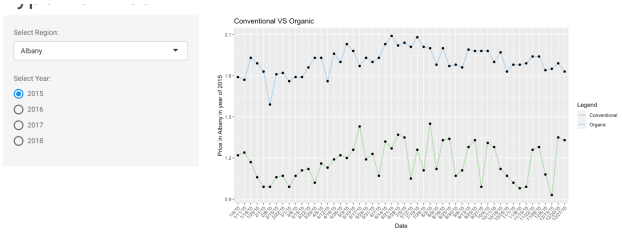
A bar chart on region(x-axis) and price/consumption(y-axis) (for the full image, see appendix 2).



Three features - avocado types, years, selection on price or consumption- can be altered. Users would see the difference of price/consumption among all regions for certain types of avocado in the selected year.

6.2.3 Price Comparison Between Avocado Types

A line chart with two lines (conventional and organic) on the date(x-axis) and price(y-axis) (for the full image, see appendix 3).

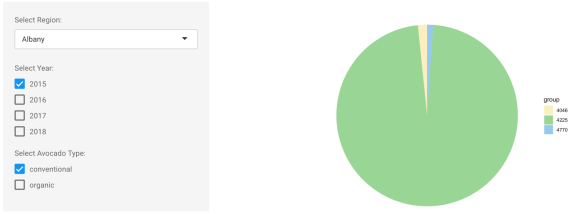


Regions and years can be altered. Users would see the price gap between conventional and organic avocado in the selected year and region.

6.2.4 Market Shares of Various Size of Avocado

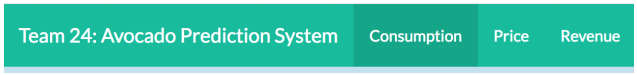
A pie chart of three different sizes of avocado(4046(S/M), 4225(L), 4770(XL))(for the full image, see appendix 4). Three features - region, years, avocado types- can be altered.

Users would see the market share of different sizes of avocado straightforwardly.



6.3 Prediction System

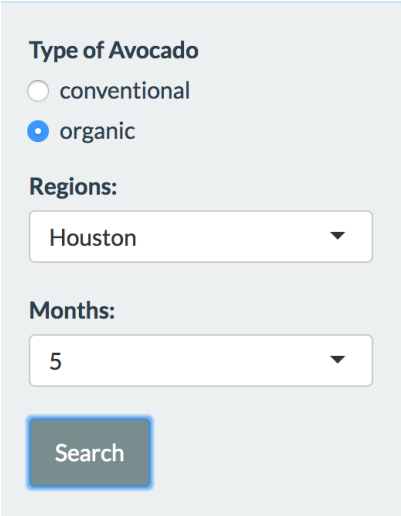
When the user enters the interface, the user is able to choose if they want to see the prediction of consumption, or price, or revenue of Avocado from the top panel.



For each page, users are able to select the type of avocado they are interested in 2 categories: conventional or organic, type or select a list of regions, and the month they want to predict.

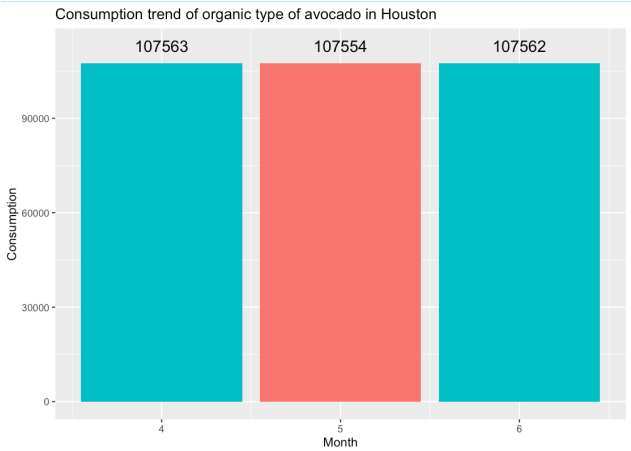
After the user inputs all the information and clicks the Search tab, for example, shown below,

the system will output a bar chart on the right.



6.3.1 Consumption Prediction page

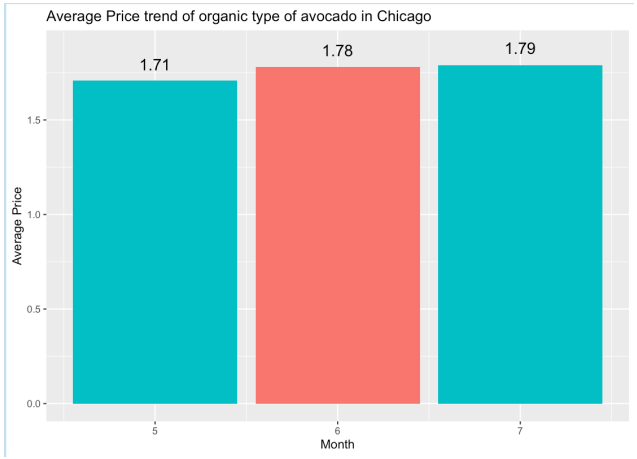
On the consumption prediction page, after the user has input the information, the system will output a bar chart displaying the predicted consumption amount for the month selected, in this case, is May(highlighted in orange), as well as the month previous(April) and the month after(June). Below is an example of the predicted consumption trend of organic types of avocado in Houston.



The y-axis represents the predicted consumption amount and the x-axis represents the month. Numbers on top of each bar chart are the value of the predicted consumption level for each month which helps the user to read.

6.3.2 Price Prediction page

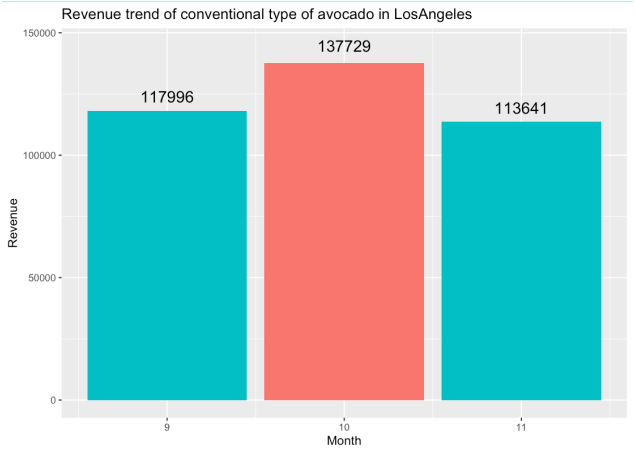
On the price prediction page, after the user has input the information, the system will output a bar chart displaying the predicted average price for the month selected, for example in June(highlighted in orange from below chart), as well as the month previous(May) and the month after(July). Below is an example of the predicted average price trend of organic types of avocado in Chicago.



The y-axis represents the predicted average price and the x-axis represents the month. Numbers on top of each bar chart are the value of the predicted average price for each month.

6.3.3 Revenue Prediction page

On the revenue prediction page, after the user has input the information, the system will output a bar chart displaying the predicted revenue for the month selected, for example in October(highlighted in orange from below chart), as well as the month previous(September) and the month after(November). Below is an example of the predicted revenue trend of the conventional type of avocado in Los Angeles.



The y-axis represents the predicted revenue and the x-axis represents the month. Numbers on top of each bar chart are the value of the predicted revenue for each month.

7. Conclusion

7.1 Results

In this project, we accomplished two goals. First, we analyzed the historical data on avocado prices and sales volume in multiple US markets using R Shiny to build a visualization system. In the visualization system, we created four different interactive charts which give users a better understanding of the overall history about prices and consumptions of avocados, the difference in prices and consumptions of avocados in different markets, the difference in prices and consumptions between organic and conventional avocados as well as the market shares of various avocados based on the product lookup codes.

Apart from the visualization system, we also create a system to predict prices, consumptions, and gross revenues in sales of avocados. The prediction system can predict the results for a specific type of avocado at a selected region in a month of the year and it can also show the predictions for one month earlier and later apart from the month the user chooses which we believe can help farmers make better decisions on when and where to sell their avocados.

Last but not least, we have learned a lot in this project. We practiced our coding in implementing the systems and familiarizing ourselves with different software. We also learned more about how to do literature reviews, academic research as well as writing reports, doing representations, and teamwork.

7.2 Future extensions

The systems can be further extended with more data and more functionalities. Instead of using the existing dataset, we can acquire data directly from HASS AVOCADO BOARD which records more detailed data of Hass avocado weekly. Therefore, our model of prediction system can be updated with the latest data every week as well as the historical data visualization. Apart from more data, the functionalities of the system can be extended as well, our prediction model can be able to calculate the net revenue difference by using data of transporting fresh products from one region to another. In addition, our prediction system can automatically find the best solution according to the user's need either in the nearest selling regions or earliest time to sell by using functions to calculate the priorities of needs and the revenue.

References

[1] Evans, E. A., & Ballen, F. H. (2015). An econometric demand model for Florida green-skin avocados. HortTechnology, 25(3), 405-411.

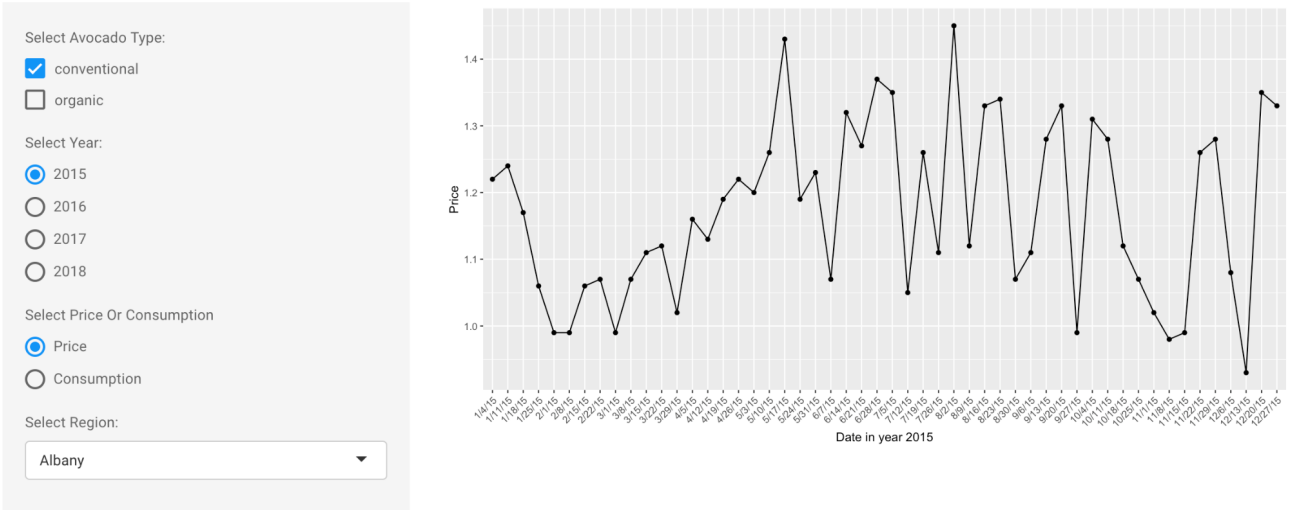
[2] Hoy F. Carman. 2019. "The Story Behind Avocados' Rise to Prominence in the United States." ARE Update 22(5): 9-11. University of California Giannini Foundation of Agricultural Economics. https://giannini.ucop.edu/filer/file/1560199276/19171/.

[3] Jones, V., Keyse, K., Melgoza, A., Perez, K., Qamar, T., Villalpando, J., & Woo, J. (2021). Avocado Buying Trends in the United States Using SAC. arXiv preprint arXiv:2104.04649.

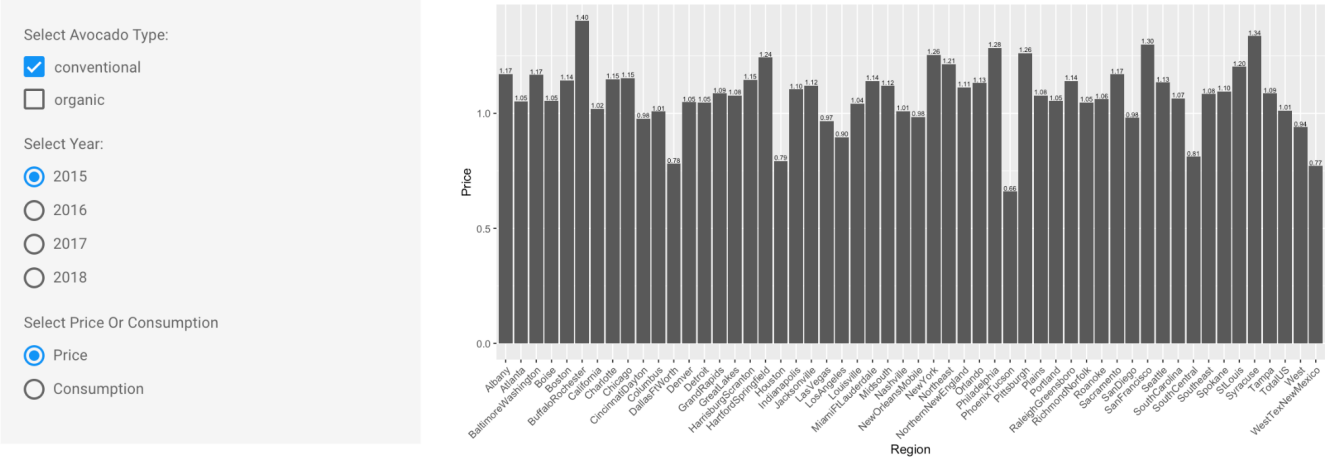
[4] Rincon-Patino, Juan & Lasso, Emmanuel & Corrales, Juan. (2018). Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data. Sustainability. 10. 3498. 10.3390/su10103498.

[5] Statista. (2021, March 2). Global avocado market value 2019–2025. https://www.statista.com/statistics/931183/global-avocado-market-value/

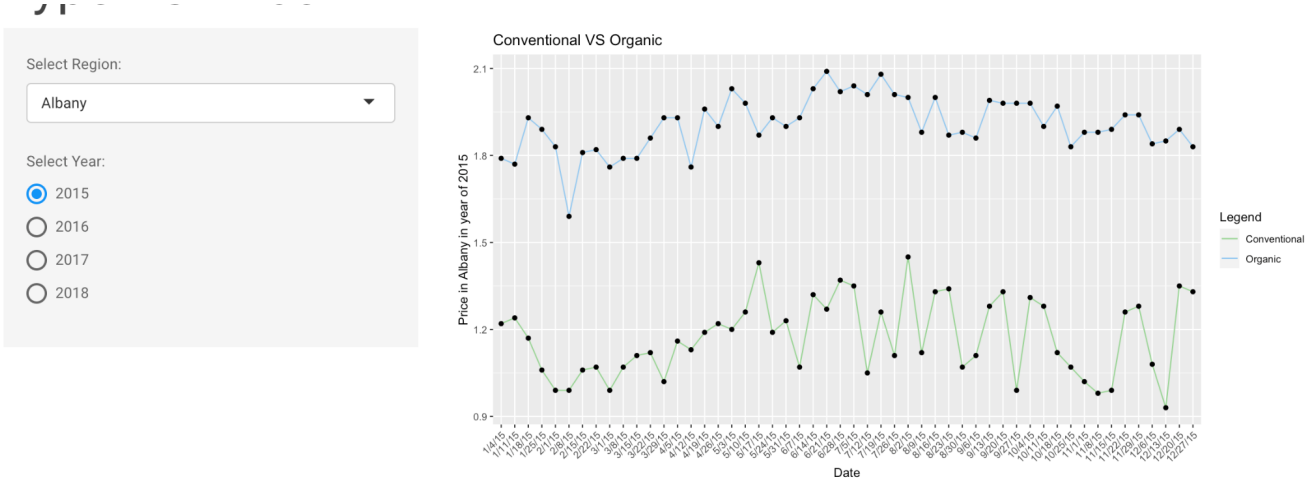
Appendix



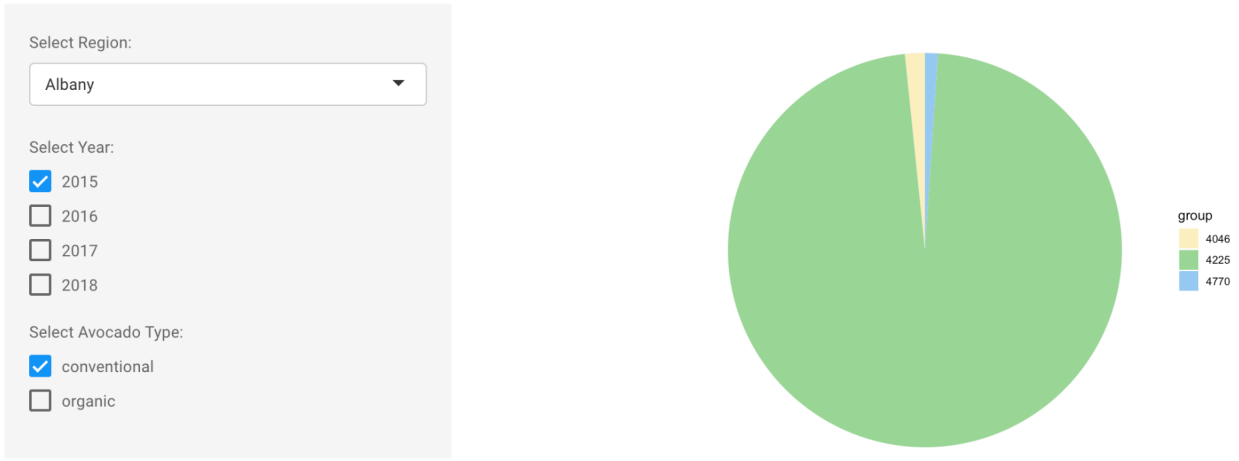
appendix 1: line chart of overall historical data



appendix 2: bar chart of overall historical data



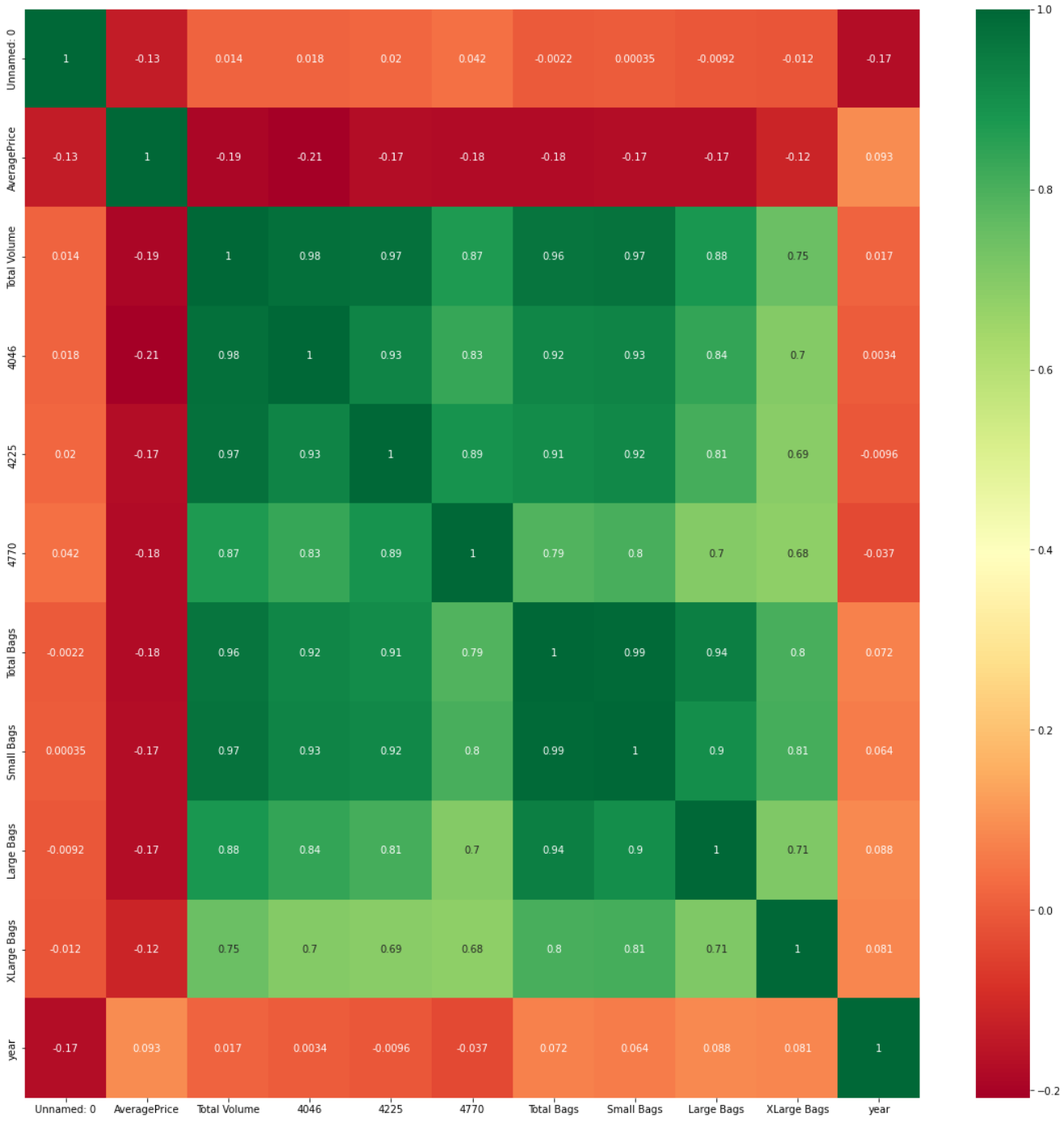
appendix 3: line chart showing difference between two types of avocados



appendix 4: pie charts showing sales taken by different sizes of avocados

	Linear Regression	Bayes Ridge	Ridge Regression	Random Forest	Support Vector Machine	Multilayer Perceptron
R2	0.6164	0.6162	0.6164	0.4587	0.6566	0.6417
MAE	0.1909	0.1909	0.1908	0.2293	0.1762	0.1851
MAPE	0.1362	0.1362	0.1362	0.1651	0.1278	0.1299

appendix 5: performance of 6 models in predicting prices



appendix 6: heat map to show correlations between data features