Final Project Executive Summary

Determining what the average price of avocados should be is crucial to companies such as Calavo Growers Inc., the largest avocado producer in the United States. Pricing an avocado too high will result in a limited number customers buying the product. Especially if there are specific reasons nobody is buying the avocado, such as its size, type, or overcharging in a specific region of the country. Pricing it too low will result in overselling the product and thus not making a profit on its true value.

In this analysis, a dataset containing over 18,000 rows of avocados sold in the United States between 2015 and 2018 were examined. The raw dataset has 13 variables that were used to predict the response variable "AveragePrice".

Three different methods were used to find which predictors had the greatest influence on price. Multiple linear regression was used to obtain the best model based on a number of factors including the adjusted r-squared, residual sum of squares, BIC, and mallows CP. All four of these measurements showed that the model containing every predictor was the best fit. The mean squared error (MSE) of each model is displayed below in Figure 1. After plotting the linear model, there appeared to be some outliers. While these were not extreme outliers, utilizing methods such as robust regression with huber and tukey weights were examined to see if they had any impact on the overall accuracy of the model. Minimizing these outliers proved not to be effective as it resulted in a higher cv_{10} than ordinary least squares regression.

The data was then tested using a decision tree method called boosting. This method was preferred as certain models are weighted according to their performance. It was also chosen because all types of decision trees are immune to outliers. After iterating through 1000 trees with a shrinkage parameter of 0.001 and an interaction depth of 3, the variable "type" had the greatest influence on the average price of avocados. Interestingly this method found that a number of variables (Total.Volume, X4225, X4770, Small.Bags, XLarge.Bags) had no impact on avocado prices. However, this model produced the highest cv_{10} , so it was not used.

The linear model produced the lowest cv_{10} of 0.029, compared to robust regression with huber weights (cv_{10}=0.0301), and boosting (cv_{10}=0.048). Double cross validation confirmed that the model utilizing every predictor produced the lowest cv_{10} .

Looking at the summary of the final model, the variable "type" does appear to have the strongest positive association on average price. Figure 2 below shows that organic avocados do in fact cost more than conventional avocados on average. The summary also shows that most months appear to be positively associated with avocado price, with the later months (September - November) averaging the highest price. This makes sense as most avocados are out of season in the United States by late fall.[1]

In conclusion this dataset is challenging to analyze given its limited number of predictors. Some predictors also proved to be completely unhelpful. There does appear to be correlation among certain variables and further exploration possibly utilizing penalized regression or generalized least squares may be better suited for this dataset.

[1] "Since the burgeoning of the international avocado trade, U.S. avocado production is highest from April to July, when imports from Mexico abate somewhat." *Economic Research Service.*

Figure 1:

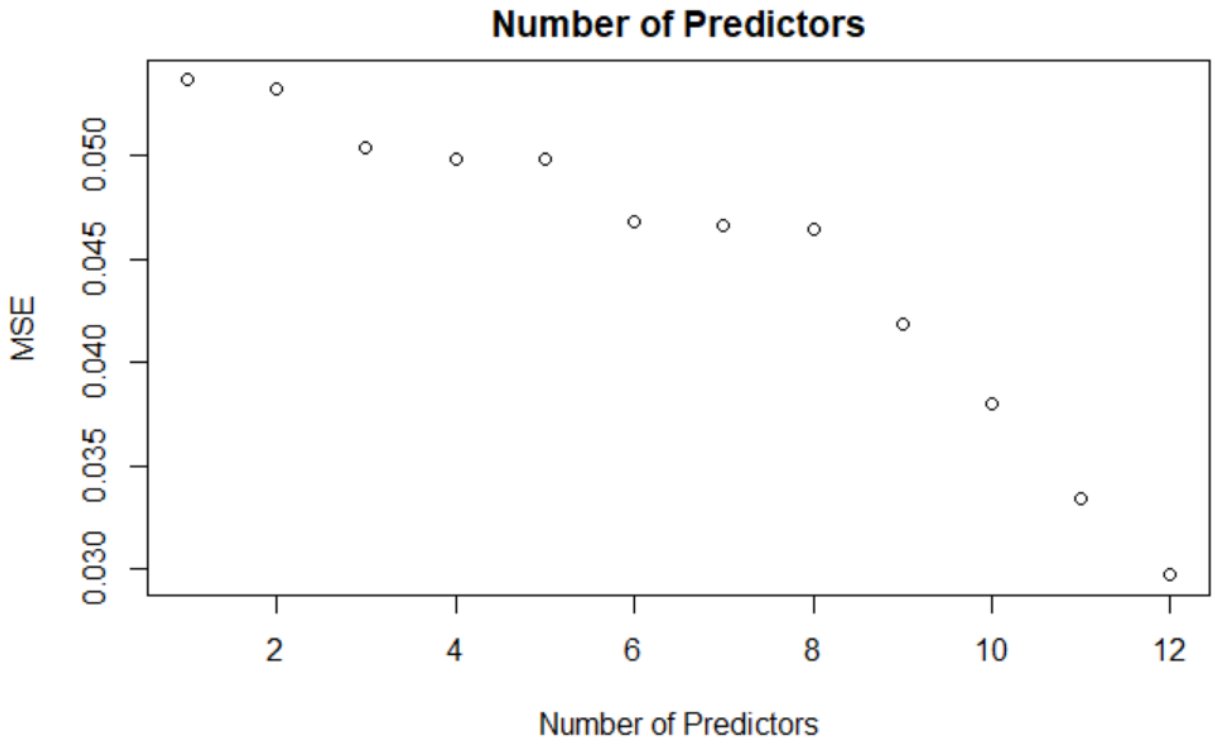
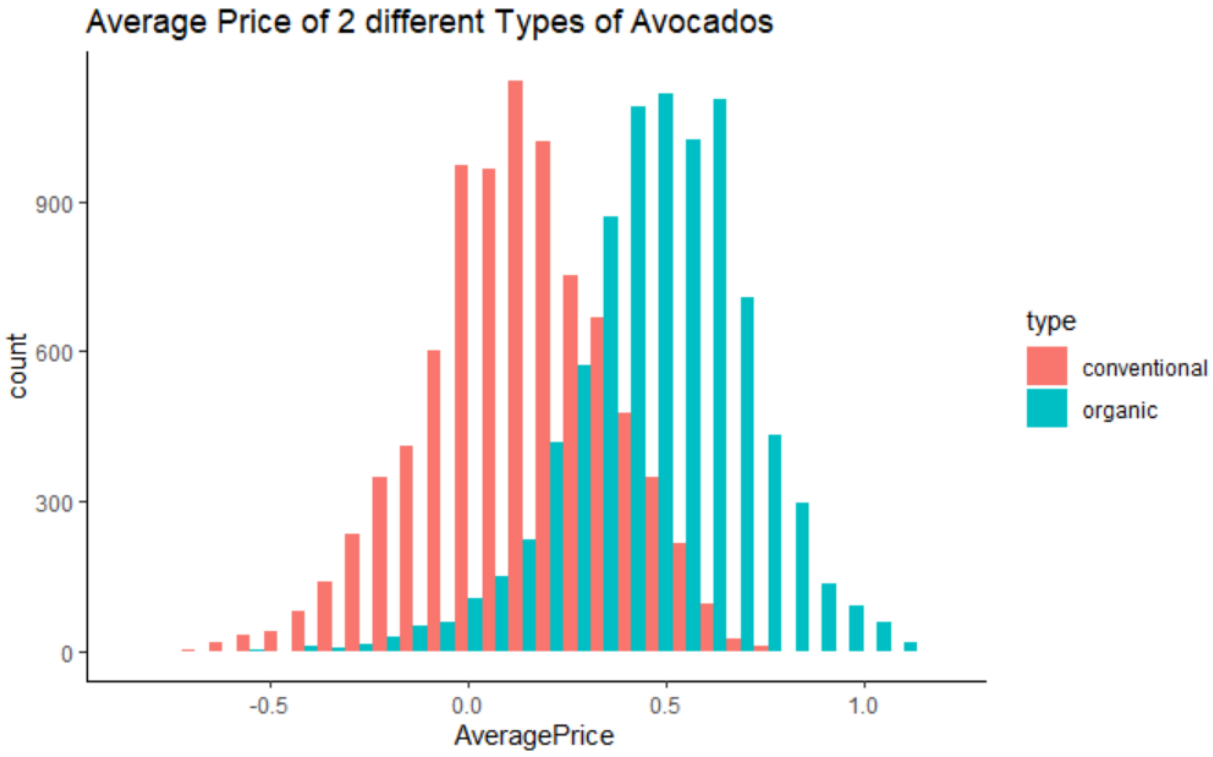


Figure 2:



Work Cited

"ICYMI... Since the Burgeoning of the International Avocado Trade, U.S. Avocado Production Is Highest from April to July, When Imports from Mexico Abate Somewhat." *USDA ERS - Chart Detail*, www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=93750.