Proposal: Avocado Price

Kristine Dinh*
San Diego State University
STAT 794: Statistical Communication in Data Science

October 29, 2020

Background

We are interested in predicting the prices of avocados in different regions. Avocado lovers wish to purchases good avocados with lower prices to make avocado toast for breakfast or guacamole to dip with chips. However, there are no way of know which avocado has good price and where should ones buy avocado at.

Specific Aims

- Study the relationship between avocado characteristics, geography and average price.
- Predict average avocado price in 2020 and compare the predicted and empirical avocado prices to access the quality of the model.
- Determine if the model is good to forecast average avocado price for next year, 2021.

Data

This data set was obtained from Kaggle. It is originally from the Hass Avocado Broad (HAB) website. This dataset has historical data of avocado prices and characteristics. The dataset contains two time series columns, and 10 predictors, as shown in table below. PLU stands for Price Look-Up code. Other variables are self-explanatory.

Variable	Definition
Date	The date of the avocado observed
Year	Year of observation
Type	Conventional or organic
Region	The city or region of the avocado
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
Total Bags	Total number of bags sold
Small Bags	Total number of small bags sold
Large Bags	Total number of large bags sold
X-large Bags	Total number of extra large bags sold

*Email: kdinh@sdsu.edu

The data is from 2015 to 2020 with 54 distinct geographical regions of where the avocados are from. Avocado lovers want to build a model that can predict the price of avocado.The train set of the model will be using avocado observed between 2015 and 2019. All predictors of avocado prices in 2020 are provided in the dataset along with the price of avocado to be the test set. If the model can generate predicted price similar to empirical avocado price, avocado lovers will be more likely to use this model to predict the price of avocado for future years. With the predicted avocado prices, individuals who eat avocado toast and guacamole would have a better understanding of what kind of avocado and where the avocado come from would decrease or increases the price.

Methods

To serve the purpose of this data analysis, we will build the model using a multiple linear regression with 11 possible predictors. The analysis will start with a data exploratory analysis to determine the relationship between avocado prices and each of the independent variable. After the exploratory analysis process, we will develop the model using the significant variables. During this process, we want to select the best variables for our model using different criterion including AIC, BIC, MSE, and R-square. After selecting the best model, we would check for the model quality by running diagnostics on the model to assess the normality of the residuals, influential points, multicollinearity, etc..

Broader Impacts

This multiple linear regression model will be the main tool to point out the relationship between average avocado prices and other avocado characteristics and locations. We will predict the prices of avocado in 2020 and compare the predicted price to the empirical price. If the errors of the predicted and empirical avocado prices are small, we can use this model to project the prices of next year. Finally, we will talk about the limitation of this dataset and suggest future ideas on this dataset to improve and implement the model. Of course, all report on this project will be shared with details documentations and reproducible R codes.