Predicting Per-Unit Prices of Hass Avocados

Yunyao Zhu

3/1/2021

Introduction

In 2017, an Australian property developer claimed that millennials were spending too much money on avocado toast instead of saving for their first home (CNNMoney 2017). While the millennial avocado toast stereotype is by no means convincing, avocados are indeed a favored fruit in the U.S. According to a market report (USAID 2014), the U.S. is the world's biggest importer of avocados. As more health benefits about avocados are discovered and promoted, demand surged even more and so did the price. The goal of this project is to investigate whether we can reasonably predict the price of avocados based on past prices and geographical data.

Existing Literature

A study (Evans 2009) investigated the various factors that might influence the avocado market. The researchers predicted that the avocado prices would likely decrease in the 2009-2010 season due to an increase in supplies and a probable decrease in demand caused by the financial crisis. Since the study was conducted over ten years ago, it seems be interesting to examine if the avocado market has changed over the years and if we can still predict the avocado prices given its supposed volatility in recent years.

Data

The dataset used in this analysis is the publicly accessible Kaggle Avocado Prices dataset, which credited the Hass Avocado Board for the collection and release of the data. This dataset contains 13 variables encompassing the per-unit prices, total volumes, regions, and sizes of Hass avocados, a cultivar of avocados. Data from the start of 2015 to the end of the first quarter of 2018 are available. Each entry represents one observation from one region in the U.S. during one week. There are over 18,000 entries in total. Details of data encoding are provided in the appendix.

To enrich this dataset, we join it with selected macroeconomic data. Specifically, we included the unadjusted monthly unemployment rate and average earnings data available on the website of the U.S. Bureau of Labor Statistics. We also included the unadjusted monthly consumer price index for all items in the U.S. (assuming the index 2015 = 100) from the economic data provided by the Federal Reserve Bank.

Research Goal

Predicting the per-unit prices of Hass avocados Specifically, we want to use the data from 2015 to the first quarter of 2017 to predict the avocado prices in the following year. Doing so allows us to compare the predicted prices with the recorded prices and evaluate the accuracy of the predictions. Predicting one year of avocado prices enables us to examine any seasonal variations in pricing.

Exploratory Data Analysis

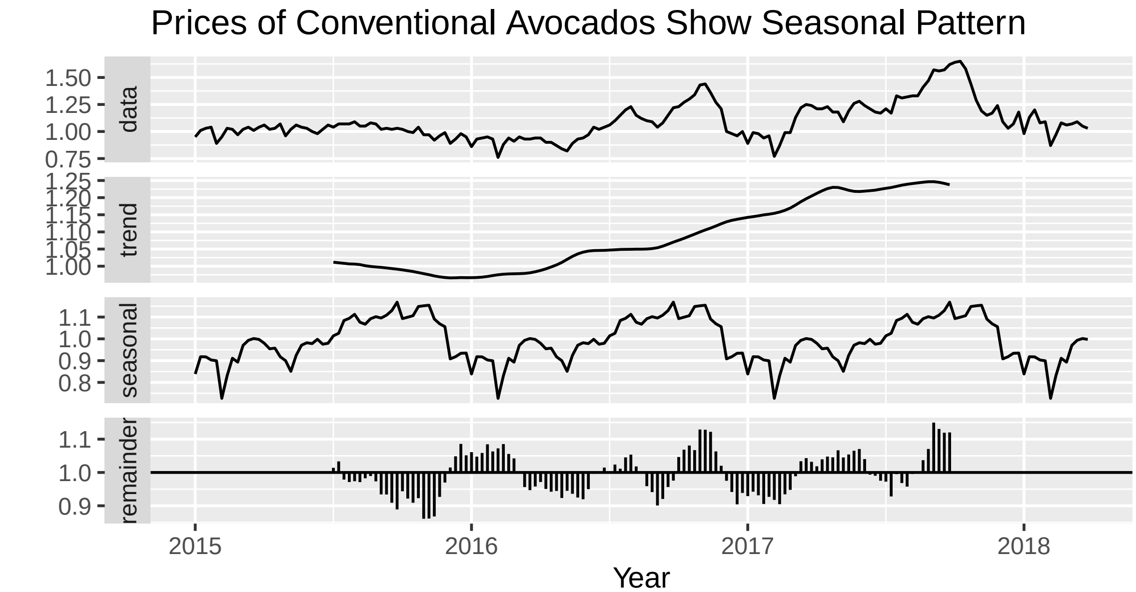

Figure 1: Time Series Plot of Per-unit Prices of Conventional Avocados in the U.S.

In Figure 1, the per-unit price of the conventional avocados increases yearly and exhibit seasonal trends. Prices peak in the third quarter and drop to the lowest in the first quarter. Organic avocados show very similar results.

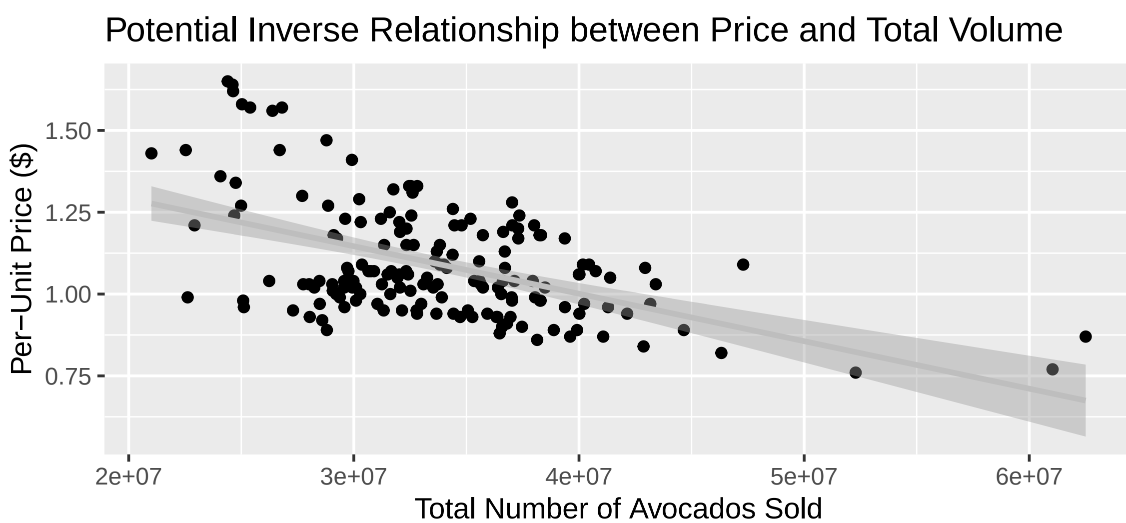

Figure 2: Scatter Plot of Per-Unit Price v.s. Total Volume of Avocados Sold

In Figure 2, an inverse relationship seems to exist between the per-unit price and the total sales volume of the conventional avocados.

Methodology

Since the avocado prices data in our dataset is a sequence of data collected successively in weeks, it can be interpreted as a time series. Thus, we conduct a time series forecasting to predict the per-unit avocado prices from the first quarter of 2017 to the first quarter of 2018.

One common type of time series forecasting model is the AutoRegressive Integrated Moving Average (ARIMA) model. It rests on the idea that the previous values in the time series can be used to predict future values (Prabhakaran 2021).

Referencing Figure 1, we noticed a seasonal pattern in our time series. In order to take this seasonal pattern into consideration when we predict the avocado prices, we choose to use a seasonal ARIMA (SARIMA), which includes seasonal terms (specified below) in addition to the ARIMA model. Since we also observed an inverse relationship between per-unit prices and total volumes, we include the total volume of avocados as an exogenous variable in our SARIMA model.

Weekly Model Specification

Seasonal ARIMA $(p, d, q) \times (P, D, Q)_s = (1, 0, 0) \times (0, 1, 1)_{52}$ (Rundel 2017a):

$$(1 - \phi_1 L)(1 - L^{52})y_t = \delta + (1 + \Theta_1 L)w_t$$

The parameter p is the number of time lags for the autoregressive (AR) model, d is the degree of differencing, and q is the order of the moving-average (MA) model. The uppercase P, D, Q denote the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model (SAS 2021). The values of the parameters are determined using principles from the Box–Jenkins method (Rundel 2017b).

Monthly Model Specification

Seasonal ARIMA $(p, d, q) \times (P, D, Q)_s = (2, 0, 0) \times (1, 1, 1)_{12}$:

$$(1 - \Phi_1 L^{12})(1 - \phi_1 L - \phi_2 L^2)(1 - L^{12})y_t = \delta + (1 + \Theta_1 L)w_t$$

We also explore a similar model with monthly avocado prices (by taking the average over the weekly prices in a given month) and exogenous variables including the monthly total volume of avocados, unemployment rate, average earnings, and consumer price index.

Results

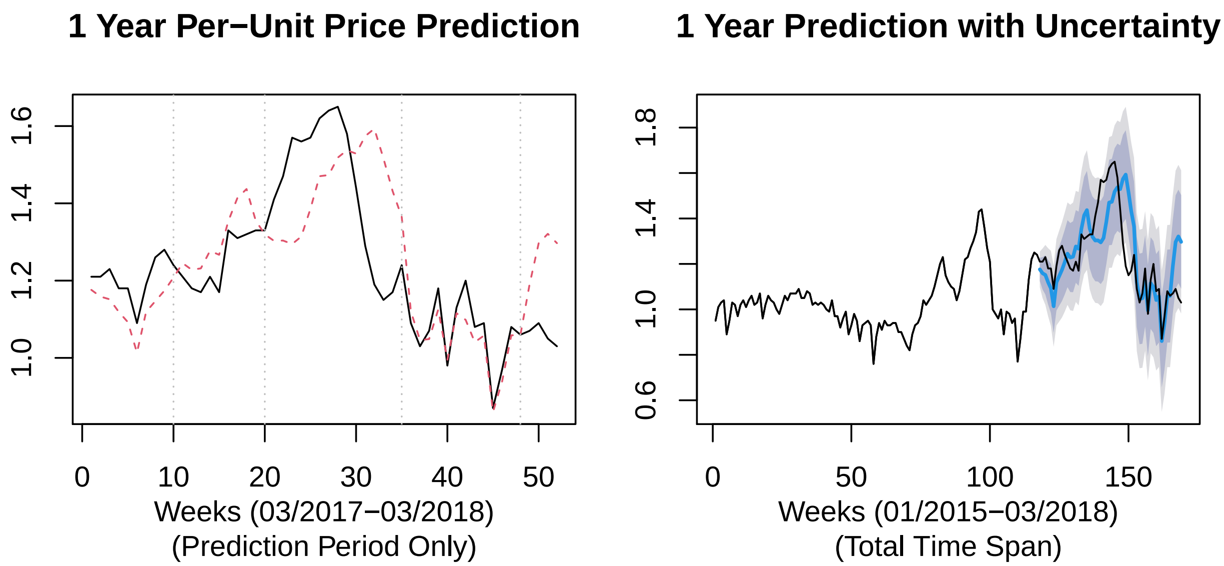

Figure 3: Predicting the Weekly Per-Unit Conventional Avocado Prices from 2017-2018

The predicted per-unit prices (in dotted red line) for weeks 0-10 and weeks 35-48 are relatively close to the observed prices (differ by less than \$0.1). The predicted prices for weeks 25-35 exhibit a similar pattern to the actual prices for weeks 20-30, i.e. the predictions seem to lag by roughly 5 weeks.

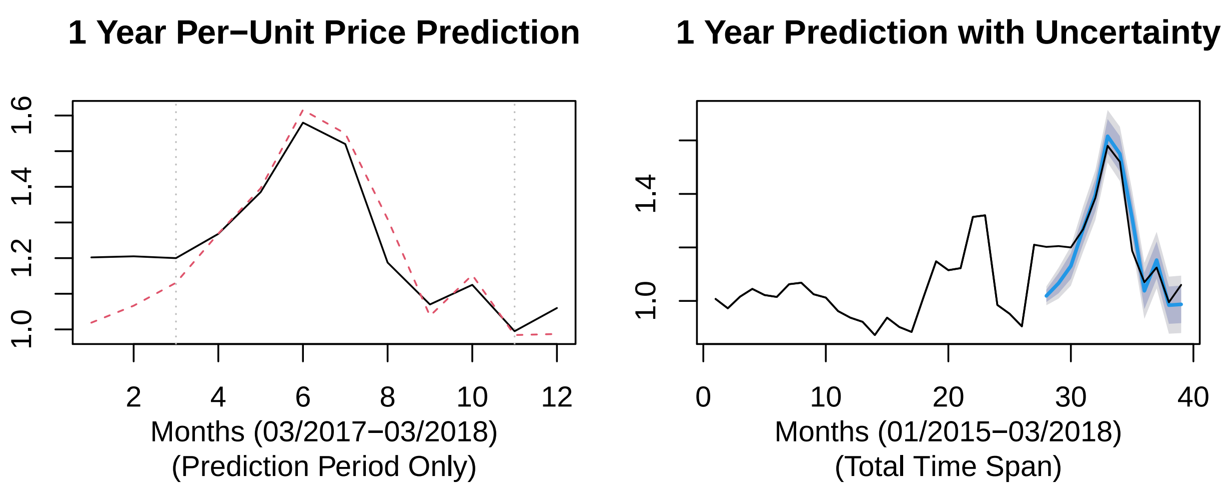

Figure 4: Predicting the Monthly Per-Unit Conventional Avocado Prices from 2017-2018

In months 3-11, the predicted per-unit prices differ no more than \$0.1 from the observed prices. In contrast, the predicted prices for months 1-3 and 11-12 show greater differences, though never exceeding \$0.2. The lag observed in the weekly model seems to disappear.

Appendix

Seasonal ARIMA Residuals

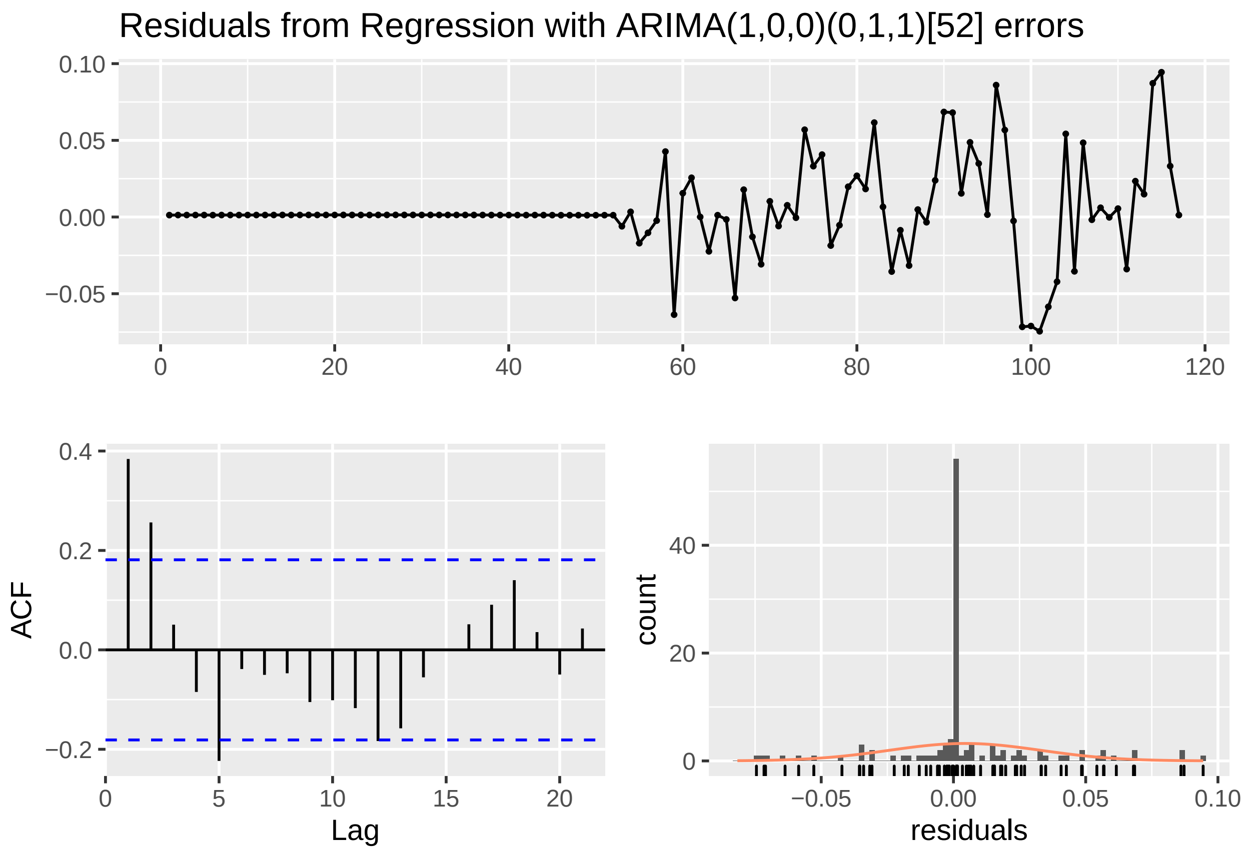

Figure 5: ARIMA Residuals based on Weekly Data

```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,0,0)(0,1,1)[52] errors
## Q* = 36.598, df = 7, p-value = 5.589e-06
##
## Model df: 3.   Total lags used: 10
```

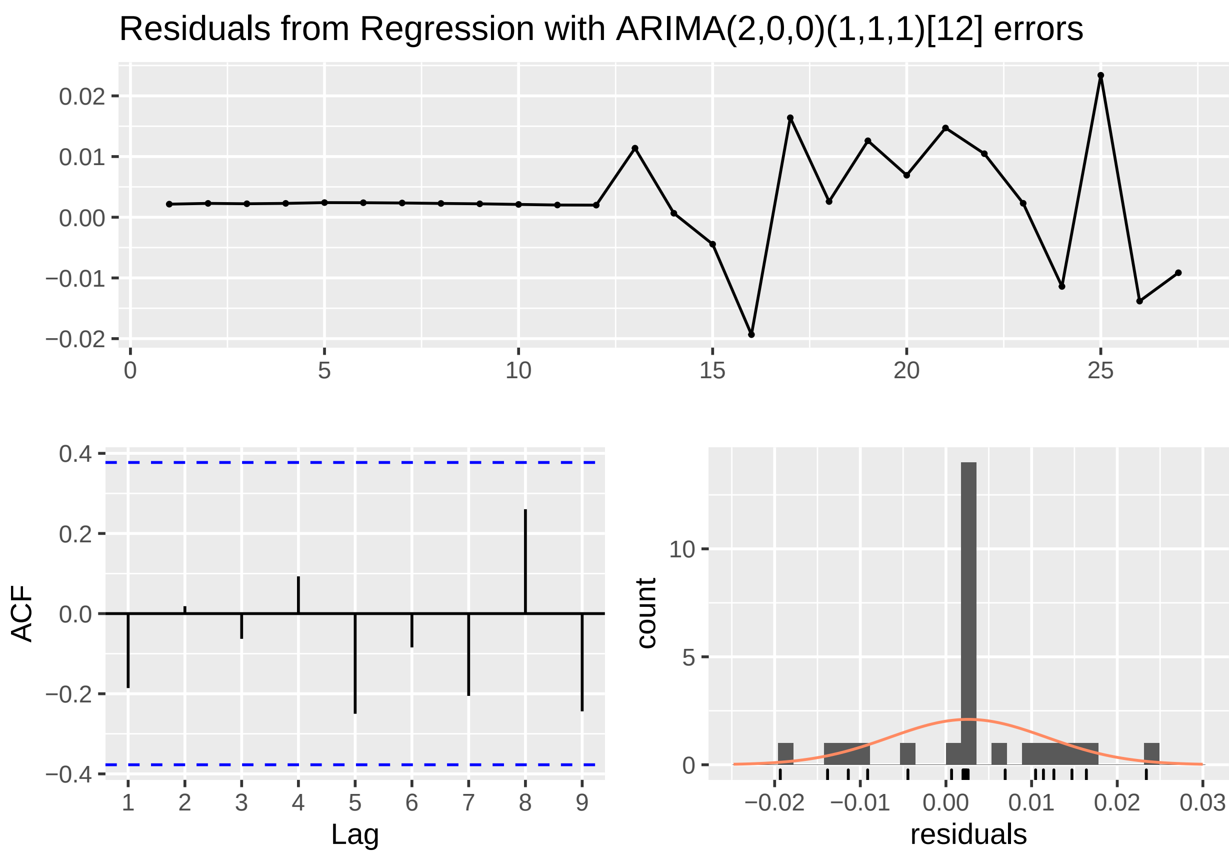

Figure 6: ARIMA Residuals based on Monthly Data

```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(2,0,0)(1,1,1)[12] errors
## Q* = 8.4112, df = 3, p-value = 0.03824
##
## Model df: 5.   Total lags used: 8
```

Bibliography

CNNMoney. 2017. "Millionaire to Millennials: Lay Off the Avocado Toast If You Want a House." *Cable News Network*.

Evans, Sikavas, Edward A. & Nalampang. 2009. "Forecasting Price Trends in the U.s. Avocado (Persea Americana Mill.) Market." *Food Distribution Research Society, Vol. 40(2), Pages 1-10*. https://doi.org/10.1016/S0140-6736(17)32812-X.

Lamstein, Ari. 2021. "Mapping Census Bureau Data in R with Choroplethr." *United States Census Bureau*. https://www.census.gov/data/academy/courses/choroplethr.html.

Prabhakaran, Selva. 2021. "ARIMA Model – Complete Guide to Time Series Forecasting in Python." *Machine Learning Plus*. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python.

Rundel, Colin. 2017a. "Lecture 11 - Seasonal Arima." *Duke Statistical Science Department*. http://www2.stat.duke.edu/~cr173/Sta444_Sp17/slides/Lec11.pdf.

———. 2017b. *Duke Statistical Science Department*. http://www2.stat.duke.edu/~cr173/Sta444_Sp17/slides/Lec9.pdf.

SAS. 2021. "Notation for Arima Models." *SAS/ETS(R) 9.3 User's Guide*. https://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_tffordet_sect016.htm.

USAID. 2014. "The Us Market for Avocado." *ACCESO*. https://pdf.usaid.gov/pdf_docs/PA00KP28.pdf.