Predicting Avocado Sales
Arnaud Laprais

Executive Summary

The main question of interest for this report is: Can we accurately predict weekly Hass avocado sales in San Francisco based on historical data? Of secondary interest is whether or not Hass avocados experience an off-season, and what effect this has on said prediction. We present an exploratory analysis of the data that suggests a slight off-season for organic Hass avocados as well as the absence of any such off-season for conventional avocados. We also present a Bayesian prediction strategy through the use of Jeffreys prior and a Poisson model. The presented strategy produces estimates with, on average, 20% deviations from observed values, usually in the form of over-estimation, which we deem an encouraging result. Finally, we discuss how this prediction strategy can be refined for more nuanced prediction outcomes.

Introduction

As is discussed in Arpaia, Hass avocados are cultivated nearly year-round in California, with most of harvesting occurring between February - August, and the Chilean market presumably making up for any dearth of avocados during the winter months. However, because of the increased costs and risks associated with international shipping, California retailers may be specifically interested in buying the correct amount of avocados to sell during the winter months (Sept-Jan). For this reason, we consider the predictive framework of off-season sales considering a subset of the data collected during these winter months. However, since this “off-season” is not a true off-season, we also consider the predictive framework provided by the full dataset, and compare the two approaches.

Data Characteristics

This data is pulled from the Kaggle “Avocado Price” dataset. This observational data, originally sourced from the Hass Avocado Board, covers weekly retail Avocado data for United States regions. This dataset includes weekly price data, weekly units sold (for organic and conventional avocados), and of course dates for all the sales. Since we are interested in predicting number of sales, we keep the “Date”, “Total.Volume”, “type”, and “region” columns.

The data covers sales from Jan 2015 to March 2018. We also choose to focus our attention on the San Francisco region, but this analysis could be repeated for a larger regional center. As discussed previously, we consider sales occurring between the months of Sept-Jan for the “off-season” analysis, and consider all months of the year for the “full-season” analysis. For the “in-season” data subsets we simply consider all observations that occurred in non-“off-season” months. We also divide our data along “organic” and “conventional” distinctions, and consider the separate predictions for each of these avocado types. The “full” and “off-season” sets consist of 169 weeks and 70 weeks of sales data respectively for both conventional and organic avocado types.

In our model, we include data only from years 2015-2017, and use 2018 observations for model validation. However, to provide a fully up-to-date summary of the data, we provide exploratory information on all available data (years 2015-2018).