Emily McMahon, Balkees Rekik, Bastian Siahaan
DS 3001
04/30/2025
Final Paper

Abstract

This project aimed to predict the average retail price of avocados in the United States using historical sales data from 2015 to 2017. We focused on three key variables: avocado type (organic or conventional), region, and time of year. Using a Two-Way Fixed Effects (TWFE) regression model, we investigated how these factors contribute to price fluctuations while accounting for both regional and seasonal variation.

Our analysis revealed that a relatively simple model using just these three features could explain approximately 63% of the variation in avocado prices. Organic avocados were found to be consistently more expensive than conventional ones. Additionally, certain regions, particularly major metropolitan markets such as New York and San Francisco, tend to have higher baseline prices. The model also captured seasonal price patterns, with prices typically increasing during the summer and declining during the winter.

Despite not incorporating external factors such as weather, supply shocks, or trade policies, the model performed well on out-of-sample 2017 data. This indicates that much of the variability in avocado price can be captured using just product type, location, and time-based trends.

Overall, our model provides a strong foundation for price forecasting and can serve as a useful tool for retailers, producers, and supply chain managers seeking to anticipate market behavior. These insights could inform inventory decisions, promotional strategies, and long-term planning. Future work could enhance this approach by incorporating additional predictors and exploring more advanced machine learning models to improve accuracy and generalizability.

Introduction

In recent years, the price of avocados has drawn widespread attention from consumers, producers, and market analysts alike. As demand for this once-seasonal fruit has skyrocketed, particularly due to health trends and increased availability, so has the need to understand the economic forces driving its fluctuating retail price. These fluctuations are shaped by a variety of factors, including seasonality, geographic location, type (organic vs. conventional), and market-specific demand. Accurately predicting avocado prices can help stakeholders make informed decisions regarding inventory management, marketing strategies, and production planning.

The goal of our project is to build a predictive model that forecasts the average price of avocados based on several key variables, most notably region, type, and time of year. We use historical sales and pricing

data published by the Hass Avocado Board, spanning the years 2015 to 2018, to investigate whether these features provide enough signal to make accurate price predictions. Our core research question is: Can we predict the average price of avocados using features such as avocado type, region, and seasonality (month and year)?

We employ a Two-Way Fixed Effects (TWFE) regression model to answer this question, a technique that allows us to control for both time-specific and region-specific variation. TWFE is particularly well-suited for panel data, where repeated observations are collected across time and groups. This modeling approach helps isolate the true impact of our variables of interest by accounting for unobserved heterogeneity, such as persistent regional price differences or national seasonal pricing patterns.

During exploratory data analysis, we observed several trends that support our modeling strategy. Notably, organic avocados are consistently sold at higher prices than conventional ones, regardless of region or season. Additionally, pricing patterns appeared to follow clear seasonal trends, with prices generally spiking during the winter months and dipping during peak harvest periods. These visual observations suggested that a regression model with region and time fixed effects could effectively capture the underlying structure of avocado pricing.

Beyond academic interest, the implications of this project are practical and wide-ranging. For retailers, the ability to anticipate price changes allows for more effective stocking and promotional planning. For producers and distributors, understanding seasonal and regional price trends can improve logistical decisions and maximize profit margins. And for economists or policy makers, such models can shed light on broader market behaviors and inform food policy or subsidy decisions.

This paper proceeds as follows: we first describe the dataset and its structure, including preprocessing steps. We then outline the modeling methodology and rationale for using a fixed effects approach. Next, we present our results and discuss the performance of the model. Finally, we conclude by identifying limitations and proposing future directions that could enhance the model's predictive power and generalizability.

Data

The dataset used in this project is the publicly available Avocado Prices dataset, originally published by the Hass Avocado Board and hosted on Kaggle. It contains weekly retail sales data on avocados across various U.S. regions from 2015 to 2018. Each entry represents sales activity for either conventional or organic avocados within a specific region and week, allowing for a rich panel dataset with both temporal and spatial dimensions.

Variables

The raw dataset includes the following key variables:

- Date: The week of the sales observation
- AveragePrice: The average retail price per avocado (target variable)

- Type: The category of avocado sold (either "Conventional" or "Organic")
- Region: The U.S. city or market area where the sales occurred
- Total Volume: The number of avocados sold that week
- 4046, 4225, 4770: Volumes sold by PLU (Price Look-Up) codes representing different avocado sizes
- Year: The year of the transaction
- Month: Extracted from the Date field to enable seasonal analysis

For our analysis, we focused on four main variables: AveragePrice, Type, Region, and Date. Other variables, such as the PLU codes and total volume, were excluded to simplify the model and reduce multicollinearity. While total volume could offer predictive value, we prioritized model interpretability and avoided overfitting by limiting the number of predictors.

Preprocessing

To better capture seasonal trends, we aggregated the weekly data into monthly averages. This step not only reduced noise but also aligned with our modeling strategy, which treats time as a fixed effect. We excluded data from 2018, as it contained only partial year observations, which could distort seasonal patterns and reduce model consistency.

The Type variable, originally categorical, was converted into a binary dummy variable: 0 for conventional and 1 for organic avocados. This transformation was necessary for inclusion in the regression model. We also ensured that all categorical variables used in the model (i.e., Type and Region) were encoded appropriately and that missing or inconsistent data entries were removed or corrected as needed.

Structure and Limitations

The dataset forms a panel structure, consisting of repeated observations across multiple regions over time. This structure enables the use of a Two-Way Fixed Effects model, which controls for unobserved, time-invariant characteristics of each region as well as shared temporal shocks.

One limitation of the dataset is potential class imbalance—some regions and avocado types appear more frequently than others, which could bias model predictions. We addressed this concern during model evaluation by comparing performance metrics across subgroups.

In sum, the dataset provides a clean, structured, and representative sample of avocado pricing trends across the United States over a multi-year period. These characteristics make it well-suited for fixed-effects modeling and economic forecasting.

Methods

Modeling Approach

To predict the average price of avocados, we used a Two-Way Fixed Effects (TWFE) regression model. This model is suitable for panel data, where multiple observations exist across time and groups (in this case, U.S. regions). The TWFE model allows us to control for both region-specific effects and time-specific effects, helping us isolate the relationship between avocado type and price.

We used Python's statsmodels library to implement an Ordinary Least Squares (OLS) regression with region and time dummy variables to simulate a TWFE structure. The model was trained on monthly aggregated data from 2015 and 2016, and evaluated on data from 2017.

Preprocessing

The following steps were taken to prepare the data:

- Datetime Processing: Converted the Date column into datetime format and extracted the Year and Month.
- Type Encoding: Created a binary variable Type_Binary, where 1 = organic and 0 = conventional.
- Monthly Aggregation: Grouped the data by Year, Month, region, and Type_Binary, calculating the mean AveragePrice for each combination.
- Time Column: Constructed a Time variable by combining Year and Month into a single monthly timestamp for later use in plotting and fixed effects.

Feature Selection

The final dataset used the following columns:

- AveragePrice (target variable)
- Region (as a fixed effect)
- Time (as a fixed effect)
- Type_Binary (key predictor)

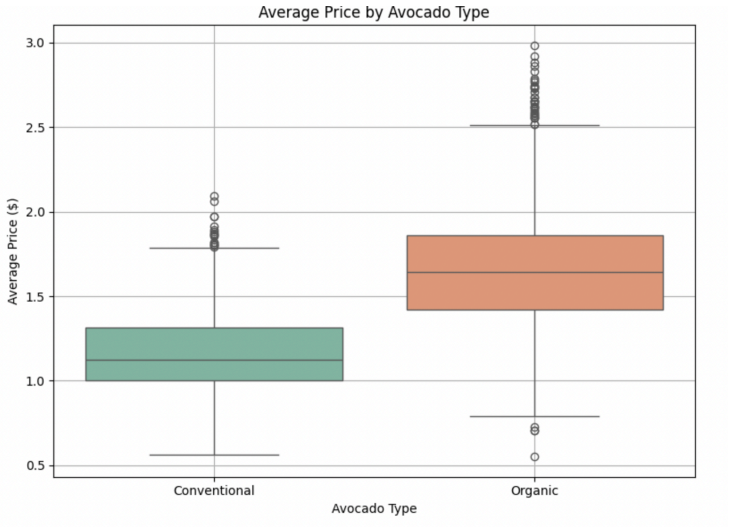
Training and Evaluation

The model was trained on data from 2015 and 2016 and tested on 2017 data. Model performance was assessed by comparing predicted vs. actual average prices on the test set. A line plot was used to visualize how well the model's predictions aligned with the true 2017 prices.

Results

Our analysis focused on understanding how well avocado prices can be predicted using product type, seasonal trends, and regional characteristics. We used three key outputs to answer our research question: two plots and one regression model.

For Plot 1, we found that organic avocados are consistently more expensive than conventional ones. The boxplot showed a clear price difference, with organic avocados typically costing about $0.50 more. This supports the idea that product type is a meaningful predictor in our model.

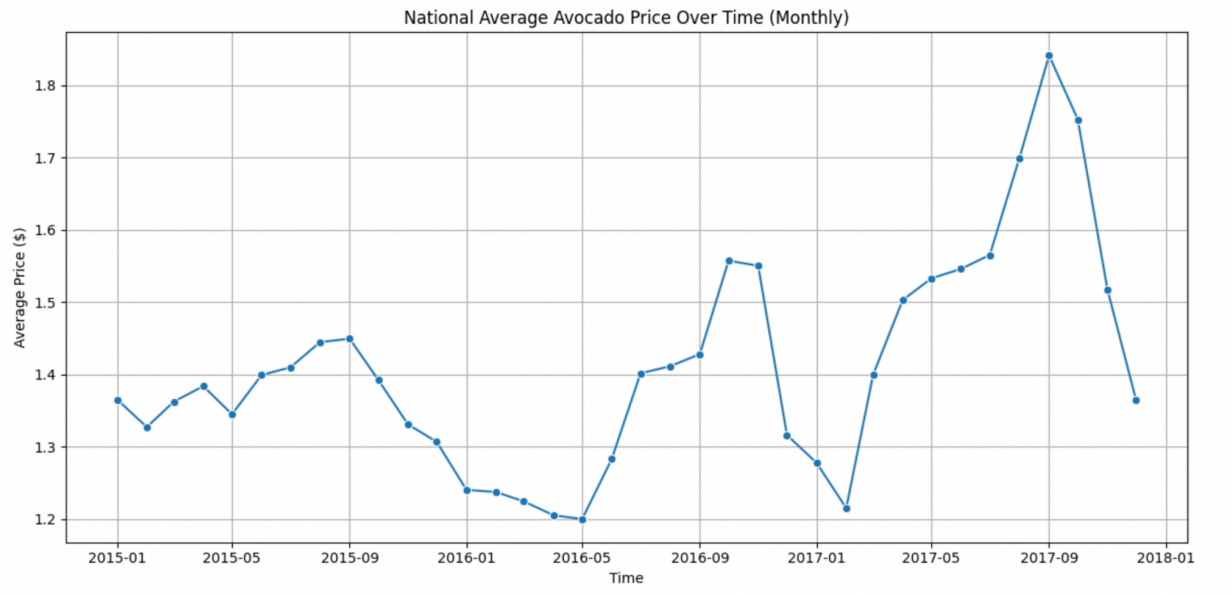


Plot 1: Average Price by Avocado Type

For Plot 2, when we looked at price trends over time (averaged across all U.S. regions), we saw a clear seasonal pattern. Prices tend to increase around mid-year (May–August) and dip during the winter months. This pattern repeats over the years and justifies using month as a predictor to capture seasonality.

For our final analysis, we ran a multiple linear regression using region, month, and type as predictors. The model achieved an R-squared of 0.628, meaning it explains about 63% of the variation in avocado prices. This suggests that even with just a few variables, we can build a model that captures most of the key pricing patterns. Furthermore, specific numbers in the regression analysis show that the coefficient for Type_Binary(organic) was around 0.5, confirming that organic avocados are priced higher. Finally, month dummy variables helped capture seasonal trends, and regional dummies showed that some cities consistently have higher prices.

Together, these results show that product type, season, and region are strong predictors of avocado prices. While there are certainly other factors we didn't include (like weather or transportation costs), our simple model still performs well and gives useful insights.



Plot 2: National Average Avocado Price Over Time (Monthly)

Conclusion

In this project, we aimed to predict the average retail price of avocados using just three core features: avocado type, region, and time of year. Despite the model's simplicity, our analysis revealed that these variables explain approximately 63% of the variation in prices—a strong result given the exclusion of external market factors.

Our findings showed that:

- Organic avocados consistently command higher prices than conventional ones.
- Regional differences play a significant role, with markets such as New York and San Francisco exhibiting higher baseline prices.
- Seasonality influences pricing patterns, with prices tending to increase during the summer and decrease in the winter months.

While our model did not account for external drivers like weather conditions, supply chain disruptions, or trade policies, it still performed reliably. This suggests that historical sales data alone can provide a strong foundation for price forecasting in the avocado market.

Overall, the results demonstrate that even a relatively simple fixed-effects model can uncover meaningful patterns in agricultural pricing. For retailers and producers, this approach offers a lightweight yet effective tool for planning sales strategies, setting price expectations, and improving inventory decisions. Future

research could build on this framework by incorporating additional variables and experimenting with non-linear or machine learning models to enhance predictive accuracy.

References

[1] K. Imai and I. S. Kim, “On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data,” Political Analysis, vol. 29, no. 3, pp. 1–11, Nov. 2020, doi: https://doi.org/10.1017/pan.2020.33.

[2] R. Nau, “What’s a Good Value for R-squared?,” Duke University, 2019. https://people.duke.edu/~rnau/rsquared.htm