Project Pre-Analysis Plan

Emily McMahon, Balkees Rekik, Bastian Siahaan
DS 3001 03/25/2025

Introduction

The price of avocados has become a topic of interest for both consumers and producers due to the fruit's rising popularity and its sensitivity to seasonal, regional, and market dynamics. Prices can fluctuate dramatically based on harvest cycles, consumer demand, and distribution logistics. Understanding what drives these price changes is crucial for retailers, farmers, and supply chain managers. By analyzing historical data on avocado sales and pricing across U.S. regions, we can identify trends and develop predictive tools that help forecast future prices more accurately. This leads us to our research question, "Can we predict the average price of avocados using features such as type (organic or conventional), region, total volume sold, and time of year?"

The goal of this study is to build a predictive model for average avocado prices using available features in the dataset, such as region, type, sales volume, and seasonal indicators like month and year. A successful model could help producers optimize pricing strategies, assist retailers with inventory planning, and allow economists to better understand produce pricing dynamics in the U.S. market. By anticipating price fluctuations, stakeholders can make more informed decisions, reduce waste, and respond to market changes more effectively.

Data Overview

The dataset comes from Kaggle's "Avocado Prices" dataset, originally published by the Hass Avocado Board. It contains weekly retail data on avocado sales across multiple U.S. regions, spanning from 2015 to 2018. The key variables include:

- Date: The week of observation
- AveragePrice: The average price of avocados per unit
- Type: Avocado Type (Either Conventional or Organic)
- Region: US metro area or market
- Total Volume: Total number of avocados sold in that week
- 4046, 4225, 4770: Volume sold by PLU(Price Look-Up) codes representing different avocado sizes
- Year: The year of the transaction
- Month: Extracted from the Date field for seasonal analysis

Each observation represents a weekly summary of avocado sales in a particular region and for a specific avocado type. For example, a single row might describe the total volume of organic avocados sold in Chicago during the third week of 2017, along with their average price.

Approach and Methodology

To predict future avocado prices, we will use a Two-Way Fixed Effects (TWFE) model, which is a supervised learning approach because it is trained on labeled data where the dependent variable (avocado price) is known. Specifically, it is a regression model since it predicts a continuous outcome—price. The model will be trained on data from 2015 and 2016 and tested on data from 2017. The TWFE model applies Ordinary Least Squares (OLS) regression while incorporating fixed effects to control for unobserved heterogeneity and eliminate bias from factors like seasonality and the differences between conventional and organic avocados. This model is particularly appropriate for our dataset because we observed clear seasonality during our Exploratory Data Analysis (EDA) and a significant price gap between conventional and organic avocados.

A common regression equation for TWFE is:

$$Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it} \quad [1]$$

Where:

- Y_{it} represents the dependent variable
- α_i represents unit-specific fixed effects
- γ_t represents time-specific fixed effects
- β represents the coefficient on X_{it}—the binary treatment indicator—indicating the effect of the treatment or condition
- ϵ_{it} represents the error term.

In the context of our project, the TWFE model could be formulated:

$$\text{AveragePrice}_{it} = \alpha_i + \gamma_t + \beta \cdot \text{Type}_{it} + \epsilon_{it}$$

Where:

- AveragePrice is the dependent variable
- α_i is the unit-specific fixed effect for Region
- γ_t is the time-specific fixed effect for Date
- β is the coefficient of the Type variable
- Type is the independent variable

During preprocessing, we will limit our dataset to those four variables:

- Date: The week of observation
- AveragePrice: The average price of avocados per unit
- Type: Avocado Type (Either Conventional or Organic)
- Region: US metro area or market

We will then aggregate the Date variable from weekly to monthly data, as seasonality shifts are more pronounced at the monthly level. Additionally, we will create a binary dummy variable for our Type variable using indicator 0 for conventional and 1 for organic. This is necessary, since regression models cannot handle categorical variables directly without conversion into numerical formats. Afterwards, we will split the data into years 2015-2016 for training and 2017 for testing. Since we only have about 5 months of data in 2018, we opt not to use it.

After Preprocessing, we will use Python's linearmodels library to estimate a Two-Way Fixed Effects model for our training data and apply the model to the training data.

Success Criteria

To evaluate the success of our model, we will examine the R^2 value produced by our model. R^2 values closer to 1 indicate that the model does a better job explaining the variance in the dependent variable (avocado price), as influenced by the independent variables (type and region). The closer R^2 is to 1, the better predictors type, region, and seasonality are in our model. We will conclude that our model is successful if we achieve an R^2 value of 0.7. We expect a high value because we observe a visual relationship between our dependent and independent values. However, we recognize that our model may not be able to predict the annual upward trend in price that we observe in our data [2].

Potential Challenges and Mitigation Strategies

- **Anticipated Weaknesses**:
    - Multicollinearity: some predictors in our data set might be correlated with unobserved variables or each other, which can distort coefficient estimates and make interpretation difficult.
    - Omitted variable bias: by limiting the model to only three features (region, type, and time), we may be excluding important predictors which could explain additional variance in price.
    - Data imbalance: certain regions or avocado types may appear more frequently in the training set, potentially biasing predictions toward those subgroups.

- **Mitigation Plan**: Describe how you plan to address these challenges if they arise (e.g., cross-validation, regularization).
    - We will calculate the Variance Inflation Factor (VIF) for predictors to assess multicollinearity. If VIF is high, we may consider dropping or combining related variables
    - Although our core model focuses on region, type, and date, we may experiment with re-introducing volume as a control variable in later iterations if model performance is weak.
    - Well evaluate not only R^2, but also Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to ensure predictive quality and consistency.

Results Presentation

- **Expected Results**: we expect our Two- Way Fixed Effects regression model to reveal meaningful insights into how seasonality, region, and avocado type affect price fluctuations.
    - **Insights from Regression Coefficients**

      - Type: we anticipate a positive and statistically significant coefficient for organic avocados, indicating that organic avocados are consistently priced higher than conventional ones.

      - Region Fixed Effects: the model will capture differences in average price between U.S. regions. We expect certain high-cost markets (e.g., New York) to have higher baseline prices than others.

      - Time Fixed Effects: by aggregating by month, we expect to observe seasonal price trends, such as higher prices during winter months and dips during harvest seasons.

Overall , we expect an R^2 value above 0.7, indicating that type, region, and seasonality together explain a large proportion of the variance in avocado prices.

- **Presentation Format**: To demonstrate our model's performance and predictions:
    - We will produce a line graph that overlays the predicted avocado prices for 2017 on top of the actual 2017 prices, showing how accurately our model captures real-world price patterns.
    - We will also present a summary table of regression results, including coefficients, p-values, and R^2 score.

These elements will allow us to clearly communicate the effectiveness of our model and the influence of each predictor.

Conclusion

Our preliminary analysis suggests that organic avocados may be priced higher than conventional ones, and that region and seasonality could significantly influence pricing. Once validated through our TWFE model, these insights could help retailers optimize inventory and pricing strategies, assist producers in planning harvests and distribution, and support economists and policymakers in assessing market stability. Understanding these factors would enable better decision-making across the supply chain.

To improve price predictions, future research should incorporate additional variables, such as weather conditions, fuel costs, and trade policies—factors we currently lack data for. Exploring more advanced machine learning models could also improve predictive accuracy, while

extending the dataset beyond 2018 would enhance robustness. Refining time series analysis could better capture short-term fluctuations, and studying demand elasticity would shed light on consumer price sensitivity. Implementing these improvements will strengthen predictive capabilities and offer greater applications for industry stakeholders.

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