Analysis and Prediction of Common Fruit Prices Per Pound Based on Bayesian Models*

Seasonal and Vendor Trends Reveal Predictable Patterns in Cost Dynamics

Mingjin Zhan

December 3, 2024

This study explores the factors influencing the price per pound of common fruits, such as apples, bananas, melons, and oranges, across different vendors and time periods. Using a Bayesian regression model, we analyzed how historical prices, seasonal trends, vendor practices, and fruit categories contribute to current pricing. Our findings indicate that historical prices and seasonal changes are the strongest predictors, while vendor-specific pricing strategies also play a significant role. By providing actionable insights into pricing patterns, this research can guide vendors in optimizing pricing strategies and help consumers make informed purchasing decisions.

Table of contents

1 Introduction **2**
- 1.1 Estimand . . . 3

2 Data **4**
- 2.1 Overview . . . 4
- 2.2 Measurement . . . 4
  - 2.2.1 Raw Dataset Variables . . . 5
  - 2.2.2 Cleaned Dataset Variables . . . 5
  - 2.2.3 Data Cleaning Process . . . 6
  - 2.2.4 Final Dataset . . . 6
- 2.3 Outcome variables . . . 6

*Code and data are available at: https://github.com/Re9shepp/Analysis-and-Prediction-of-Common-Fruit-Prices-Per-Pound-Based-on-Bayesian-Models.git

2.4 Predictor variables . . . 8
2.4.1 Summary of Predictor Variables . . . 8
2.4.2 Old Price Per Lb . . . 8
2.4.3 Vendor . . . 9
2.4.4 Month . . . 10
2.4.5 Fruit Category . . . 11

3 Model **13**
3.1 Model set-up . . . 13
3.2 Model justification . . . 14
3.2.1 Defining and Justifying Reasonable Priors for the Bayesian Model . . . 15

4 Results **16**
4.1 Estimates of Predictors . . . 16
4.2 Residuals Distribution . . . 17
4.3 Predicted Price Per Pound by Month, Fruit Category, and Vendor . . . 18

5 Discussion **19**
5.1 Summary of Findings . . . 19
5.2 Limitation . . . 20
5.2.1 Temporal Scope . . . 20
5.2.2 Lack of External Factors . . . 20
5.2.3 Absence of Interaction Effects . . . 21
5.2.4 Limitations of Prior Distributions . . . 21
5.2.5 Regional Generalizability . . . 21
5.3 Future Research . . . 21
5.3.1 Extending Temporal Coverage . . . 21
5.3.2 Incorporating External Factors . . . 21
5.3.3 Including Interaction Effects . . . 22
5.3.4 Optimizing Prior Distributions . . . 22
5.3.5 Cross-Regional Validation . . . 22

A Appendix **23**
A.1 Appendix: Survey Design . . . 23
A.1.1 Questionnaire . . . 23

References **26**

1 Introduction

The cost of fruits is a fundamental aspect of consumer expenditure and a key indicator of market dynamics in the grocery industry. Understanding the factors that influence fruit prices has

broad implications, ranging from optimizing consumer purchasing decisions to improving retailer pricing strategies and informing agricultural production. Despite the apparent simplicity of pricing, it is a complex phenomenon influenced by historical trends, seasonal fluctuations, vendor-specific policies, and the types of fruits being sold. This study seeks to unravel these complexities by employing a Bayesian modeling framework to predict the current price per pound of common fruits, including apples, bananas, melons, and oranges.

Existing literature has extensively explored price dynamics in food markets, but there remains a gap in applying probabilistic approaches that explicitly incorporate uncertainty and leverage rich datasets for predictive insights. Furthermore, studies rarely examine the interplay of historical prices, seasonal effects, and vendor-specific strategies in a unified model. To address this gap, our research investigates fruit prices using a comprehensive dataset of over 1,600 entries collected from multiple vendors, spanning six months. By employing Bayesian regression, this study integrates prior information with observed data to generate robust predictions, contributing a unique probabilistic perspective to the field of pricing analysis.

Our findings reveal notable patterns in fruit pricing. Historical prices strongly influence current prices, while seasonal variations are evident across fruit categories. Vendors adopt distinct pricing strategies, with melons consistently priced higher than apples and oranges, reflecting differences in production costs and demand patterns. Additionally, the Bayesian framework enhances our understanding of price uncertainty, offering actionable insights for stakeholders.

This research is significant because it not only advances methodological approaches by employing Bayesian techniques but also provides practical implications for consumers, producers, retailers, and policymakers. For consumers, the results highlight optimal times to purchase specific fruits. Producers can leverage the insights to plan production and distribution schedules. Retailers gain a deeper understanding of pricing strategies to enhance competitiveness, and policymakers can use the findings to stabilize prices and ensure market fairness.

The remainder of this paper is structured as follows: Section 2 details the data, including its sources, cleaning process, and key variables. Section 3 describes the model setup and justifies the choice of priors. Section 4 presents the results, highlighting significant predictors and their implications. Section 5 discusses the findings, outlines limitations, and proposes directions for future research.

1.1 Estimand

In this study, our estimand is the expected value of the current price per pound for common fruits (such as apples, bananas, melons, and oranges).

2 Data

2.1 Overview

In this analysis, we utilized R (R Core Team 2023) to examine fruit pricing data, focusing on factors influencing current price per pound across different vendors, fruit categories, and time periods. Our dataset, sourced from local vendors, provides a comprehensive view of pricing dynamics and trends from June to November.

Several R packages were crucial for our data manipulation, modeling, and visualization processes. The tidyverse and dplyr packages were instrumental in data transformation and summarization (Wickham, François, et al. 2023), offering efficient workflows for cleaning and structuring the dataset. The ggplot2 package was employed to create informative visualizations that captured key trends and relationships within the data (Wickham, Chang, et al. 2023). For Bayesian modeling, we used the rstanarm package, which enabled us to implement sophisticated regression models and incorporate prior distributions into our analyses (Goodrich et al. 2023). The modelsummary package streamlined the presentation of model outputs, ensuring clarity and coherence in reporting results (Arel-Bundock 2023). Additionally, the kableExtra package was used to produce customizable tables, enhancing the visual appeal and readability of our findings (Zhang 2023).

To manage files and directories effectively, we relied on the here package, which helped establish reproducible paths to data and output files (Müller and Bryan 2023). The arrow package facilitated efficient storage and retrieval of large datasets in parquet format, ensuring smooth data handling throughout the project (Richardson, Ooms, et al. 2023). Finally, the bayesplot package allowed us to evaluate and visualize Bayesian model diagnostics and posterior distributions, ensuring the validity and interpretability of our modeling results (Gabry et al. 2023).

Our analysis focused on predicting the current price per pound for different fruit categories. To ensure data reliability, we implemented rigorous cleaning and transformation steps, filtering out anomalies and aligning variables for consistency. This structured approach enabled us to uncover valuable insights into vendor-specific and seasonal pricing patterns, providing actionable findings for pricing strategy optimization.

2.2 Measurement

In this section, we outline the variables included in both the raw dataset and the cleaned dataset used in the analysis. The raw dataset variables were sourced from Hammer (Filipp 2024), while the cleaned dataset variables were derived from the data cleaning process using the R script.

At the core of this study is the concept of understanding how various factors influence the pricing of common fruits, such as apples, bananas, melons, and oranges, across different vendors

and timeframes. Our aim was to answer specific questions: What drives price changes? How do different vendors set their prices? And what patterns emerge across fruit categories? To answer these questions, we needed a dataset that captured detailed information about fruit prices, vendor practices, and temporal trends.

Below, we provide a detailed description of each variable, its origin, and its transformation.

2.2.1 Raw Dataset Variables

- **Nowtime:** Timestamp indicating when the data was gathered. Used to track data collection times and align price trends temporally.
- **Vendor:** One of the seven grocery vendors, representing the retail sources of the data.
- **Product Id:** An internal ID for a product, unique within a vendor, enabling tracking of specific products over time. IDs are not consistent across vendors.
- **Product Name:** The name of the product, which may include brand information and unit descriptors.
- **Brand:** The brand name of the product. This field may be blank for some vendors.
- **Units:** The unit of measurement (e.g., grams, kg, or the number of items in a package). This field may be blank for certain products or vendors.
- **Current Price:** The product's price at the time of extraction.
- **Old Price:** A struck-out "old" price indicating a sale or price drop. Helps distinguish advertised sales from "quiet" price reductions.
- **Price Per Unit:** The vendor-listed price per unit. This value may not match the calculation of current price divided by units and needs verification.
- **Other:** Additional details from the product listing, such as stock status ("Out of stock"), promotional information ("SALE"), or conditions ("$5.00 MIN 2").

This raw data was reflective of real-world phenomena — fluctuations in grocery prices driven by sales, seasonality, supply chain dynamics, and vendor-specific policies.

2.2.2 Cleaned Dataset Variables

The cleaned dataset builds upon the raw data, standardizing and constructing new variables to facilitate analysis:

- **Current Price Per Lb:** The current price per pound of the product, calculated by standardizing current price and units to a consistent per-pound metric.
- **Old Price Per Lb:** The historical price per pound, derived similarly to current price per lb.
- **Month:** The month of data collection, extracted from nowtime, to analyze seasonal price trends.
- **Vendor:** Retained from the raw data but cleaned for consistent labeling across records.

- **Fruit Category:** A new categorical variable identifying the type of fruit (e.g., Apple, Banana, Melon, Orange), constructed based on product name.

2.2.3 Data Cleaning Process

The raw data was processed to create the cleaned dataset using the following steps:

1. Standardization: Converting all price metrics to a per-pound basis.
2. Categorization: Mapping product names to fruit categories.
3. Imputation: Handling missing or inconsistent data entries.
4. Outlier Removal: Identifying and excluding extreme outliers in price data.
5. Variable Construction: Deriving new variables such as month and predicted_price_per_lb.

2.2.4 Final Dataset

The resulting dataset is a structured and clean representation of the real-world pricing phenomena we set out to study. It includes key variables like vendor, fruit category, timestamp, and calculated metrics such as price per pound. This transformation from raw data to structured information bridges the gap between a conceptual idea and actionable insights, enabling a rigorous analysis of fruit pricing dynamics.

2.3 Outcome variables

The primary outcome variable in the analysis is the current price per pound. This variable represents the price of fruits after accounting for various influencing factors such as historical prices, vendors, months, and fruit categories. The study focuses on understanding and predicting this value to identify key trends and determinants of fruit pricing.

Figure 1 illustrates the frequency distribution of current price per pound, showing how often different price ranges occur in the dataset. The majority of prices are concentrated between $1 and $2 per pound, with the highest frequency observed just above $1. This indicates that most products are priced in this affordable range. A secondary cluster is visible between $2 and $3 per pound, although the frequency is significantly lower. Additionally, there are smaller clusters of prices in the range of $4 to $6 per pound, suggesting that higher-priced products are less common.

The distribution is right-skewed, with the majority of observations concentrated in the lower price range, below $2, and progressively fewer products as the price increases. The high concentration of products in the $1-$2 range suggests affordability is a key pricing strategy, likely targeting price-sensitive consumers. The smaller clusters at higher price ranges may

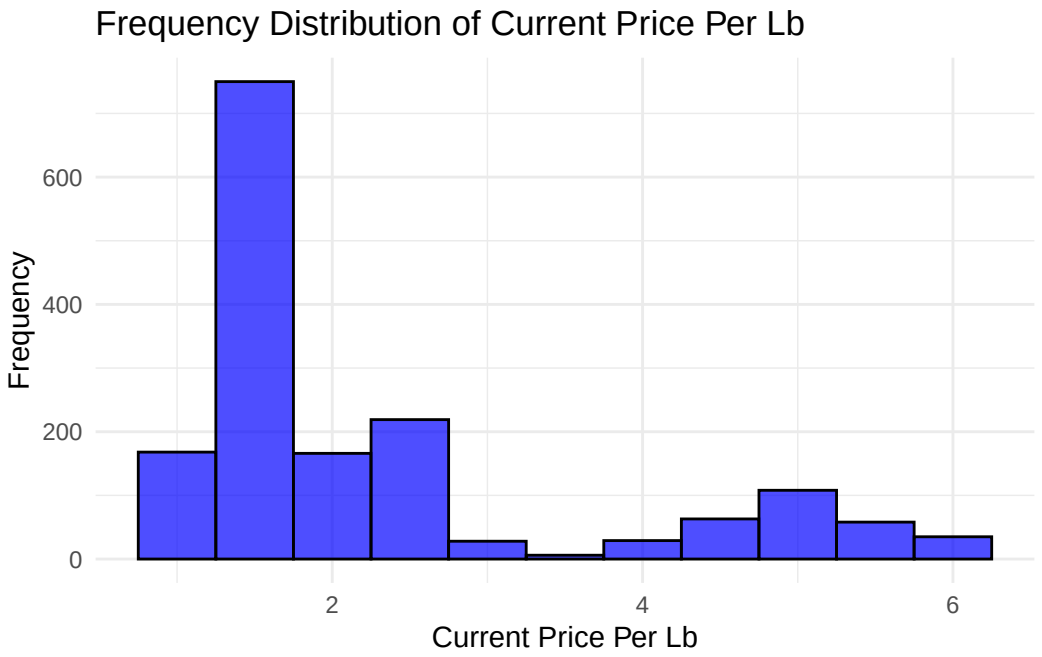

Figure 1: Frequency Distribution of Current Price Per Lb

reflect premium or specialty products, which are less frequently priced or purchased. Overall, this distribution highlights the dominance of affordable pricing in the dataset while indicating the presence of a smaller segment of higher-priced items.

Table 1: Summary Statistics for Current Price Per Pound

Mean	Median	Standard Deviation	Minimum	Maximum
2.33	1.68	1.36	0.99	6.21

Table 1 summarizes the distribution of current prices per pound in the dataset, highlighting key statistical measures. The mean price is $2.33, representing the average cost per pound across all observations. The median price is $1.68, indicating the midpoint of the price distribution. The fact that the median is lower than the mean suggests that the distribution may be skewed, with some higher prices driving the average upwards. The standard deviation of $1.36 reflects the level of variability in the prices, with higher values indicating more dispersed price points. The minimum price is $0.99, representing the least expensive items, while the maximum price reaches $6.21, highlighting the most expensive products in the dataset.

2.4 Predictor variables

The predictor variables in the analysis include several key factors that influence the current price per pound. One of the most important predictors is the old price per pound, which reflects the historical price of the product before any discounts or sales. This variable plays a critical role in understanding pricing trends, as a strong positive correlation is observed between the old price and the current price. Another significant predictor is the month, which represents the time period from June to November. This variable captures the impact of seasonal or temporal changes on fruit pricing, shedding light on how prices vary throughout the year.

The vendor is also a crucial predictor, categorizing the retailers into Metro and Voila in this study. This variable helps to identify how pricing strategies differ across vendors, reflecting the unique approaches taken by each. Lastly, the fruit category serves as an essential predictor, dividing products into four groups: Apple, Orange, Banana, and Melon. This variable demonstrates the varying effects of fruit type on pricing, with some categories showing distinct patterns and trends. Together, these predictors provide a comprehensive framework for analyzing the factors that shape fruit pricing dynamics.

2.4.1 Summary of Predictor Variables

- **Old Price Per Lb:** The price per pound before the sale or discount was applied
- **Vendor:** The retailer or supplier from which the product is sold, one of Voila, Metro.
- **Month:** The pricing period for this product ranges from June to November.
- **Fruit Category:** This product belongs to one of the following fruit categories: Apple, Orange, Banana, or Melon.

2.4.2 Old Price Per Lb

Figure 2 illustrates the relationship between old price per pound (x-axis) and current price per pound (y-axis). Overall, within the x-axis range of 0 to approximately 6, prices exhibit a clear positive correlation, where an increase in old price per pound corresponds to an increase in current price per pound. This indicates that products with higher historical prices generally maintain relatively higher current prices.

At the same time, data points are significantly clustered within the range of 0 to 5 for old price per pound and 0 to 6 for current price per pound, suggesting that most products are distributed within this price range. However, when the old price per pound exceeds 10, data points become more scattered, and the relationship between old and current prices is no longer consistent. Some products with old prices exceeding 15 per pound maintain current prices of

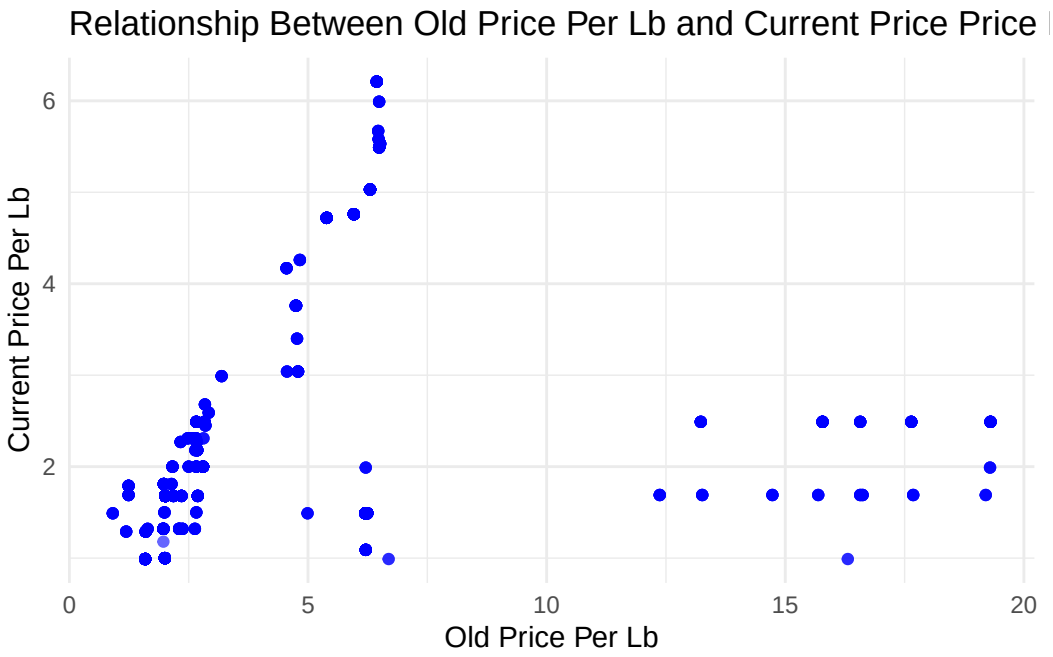

Figure 2: Relationship between old-price-per-lb and current-price-per-lb

around 1 to 2 per pound, possibly reflecting the impact of price discounts or changes in market demand.

Moreover, at the upper end of the y-axis, where current prices approach 5 to 6 per pound, prices demonstrate a "plateau" effect. Even as old prices per pound increase further, there is very limited growth in current prices. This may be due to pricing caps in the market or a decline in demand for such high-priced products.

In summary, for products with moderate old prices, such as those between 0 and 5, current prices seem to adjust proportionally, likely driven by steady demand and competitive pricing strategies. For products with higher old prices, however, there may be more aggressive discounting strategies or relatively stable current prices, reflecting a possible shift in pricing strategies by sellers.

2.4.3 Vendor

Figure 3 illustrates the distribution of current price per pound across different vendors, comparing two vendors: Metro and Voila. Overall, the median current prices per pound for Metro and Voila are very similar, indicating that the central pricing trends for the two vendors are aligned. The green box represents the interquartile range (IQR), which contains 50% of the

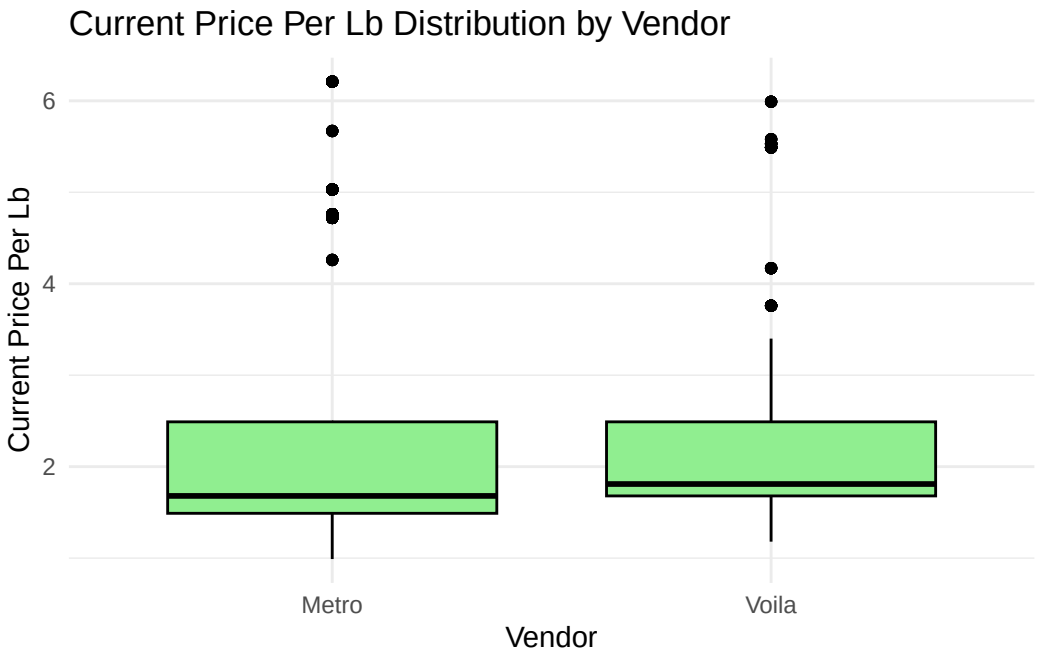

Figure 3: Relationship between vendor and current-price-per-lb

data. The narrow IQR for both vendors suggests that their pricing structures are relatively stable and consistent.

In terms of price range, Voila exhibits a slightly broader range compared to Metro, indicating that Voila's product prices are somewhat more variable. Regarding outliers, both vendors have some high-price outliers above the upper whisker, representing products priced significantly higher than most items, which could be premium or specialty products. The plot shows that Metro has slightly more outliers than Voila, which might suggest that Metro occasionally features higher-priced items in its catalog.

In summary, the price distributions for the two vendors are very similar, with no significant differences in central pricing trends or overall price ranges.

2.4.4 Month

Figure 3 illustrates the monthly variation in current price per pound, showing the price distribution across five months: June, July, August, October, and November. From the chart, we can observe that the median price per pound varies slightly across the months, with August having a significantly higher median compared to the others. In contrast, June, July, October, and November have relatively similar median prices. August also exhibits the widest interquartile range, indicating greater price fluctuations during this month, with a wider range of both

Figure 4: Relationship between month and current-price-per-lb

higher-priced and lower-priced products. The other months have narrower IQRs, reflecting a more concentrated price distribution.

Additionally, all months feature outliers above the upper whisker, representing products priced significantly higher than most others. August and November have more outliers, suggesting that these months may have some products with exceptionally high prices. In contrast, June and October show more consistent pricing with fewer outliers.

Overall, August demonstrates the greatest price variability, which may reflect seasonal demand, promotional activities, or a broader variety of products. The other months exhibit relatively stable pricing, indicating a more balanced market environment.

2.4.5 Fruit Category

Figure 5 displays the distribution of current price per pound by fruit category, comparing four types of fruit: Apple, Banana, Melon, and Orange. Apples and Oranges have the lowest median prices per pound, indicating that these fruits are generally more affordable compared to Bananas and Melons. Among the categories, Bananas have a higher median price than Apples and Oranges, while Melons exhibit the highest median price. The interquartile range for Apples and Oranges is relatively narrow, suggesting consistent pricing with minimal variability. In

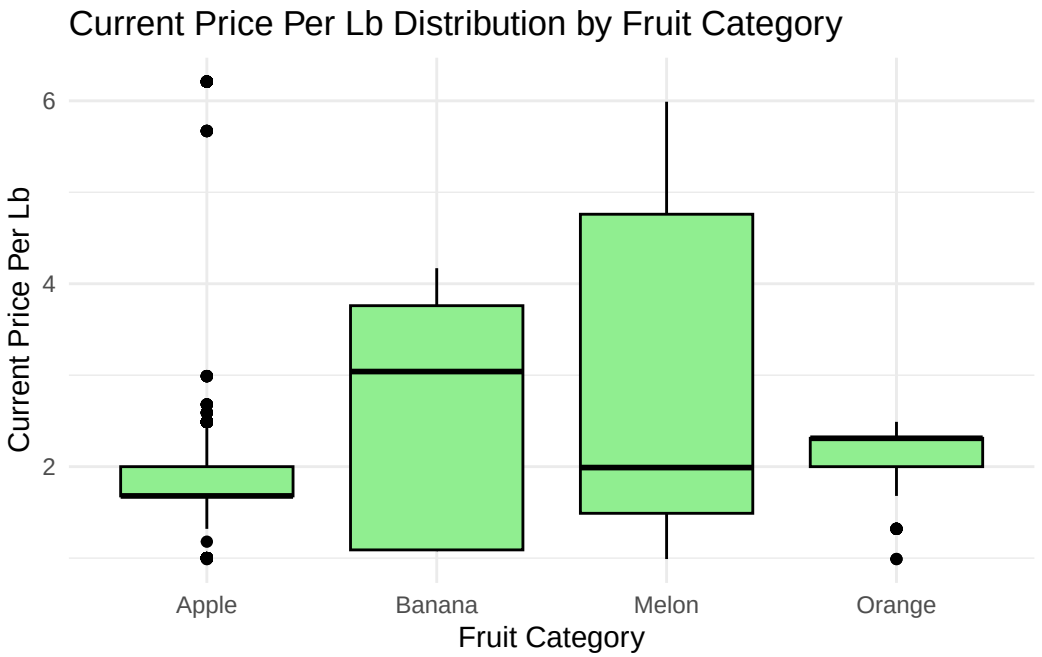

Figure 5: Relationship between fruit and current-price-per-lb

contrast, Bananas have a wider IQR, and Melons show the widest IQR, reflecting significant fluctuations in price per pound.

Outliers are present in the data. Apples and Oranges have some outliers below the lower whisker, representing unusually low-priced products. On the other hand, Melons have outliers above the upper whisker, which likely correspond to premium or specialty products with exceptionally high prices. Overall, Melons exhibit the greatest variability in pricing, likely due to differences in size, type, or quality, while Apples and Oranges are priced more consistently, likely reflecting their status as more common or standard fruit options.

In summary, Apples and Oranges are consistently affordable, making them accessible to a broader range of consumers. Bananas and Melons, on the other hand, display greater price variability, possibly influenced by factors such as seasonality, product variety, or regional differences. The high-priced outliers for Melons suggest the presence of occasional premium items, while the low-priced outliers for Apples and Oranges may indicate discounts or bulk pricing.

3 Model

The goal of this section is to address the inherent complexities and variations present in fruit pricing data to build a robust predictive model. The key challenge lies in achieving an optimal balance between model complexity and fit, ensuring that the model accurately captures the dynamics of fruit pricing patterns while avoiding overfitting. To this end, we evaluated multiple model specifications to identify the one that best meets our forecasting objectives.

We chose to use variables such as "old price per pound," "month," "vendor," and "fruit category" instead of potentially less relevant or subjective factors. For instance, including historical pricing as "old price per pound" provides a reliable foundation for understanding current price trends, while factors like "fruit category" and "vendor" offer meaningful segmentation for better predictions. These variables allow the model to focus on quantifiable and interpretable drivers of pricing differences, reducing potential noise and bias.

Additionally, we systematically explored the effects of these variables by adding complexity incrementally. For instance, starting with historical prices as the sole predictor, we gradually introduced temporal (month), categorical (fruit category), and vendor-specific variables to refine the model's explanatory power. This systematic approach ensures that the model captures meaningful relationships without unnecessary complexity.

By systematically comparing model specifications with different combinations of variables, we aim to identify a predictive framework that balances accuracy and generalizability, ultimately providing reliable forecasts for fruit pricing trends.

3.1 Model set-up

We aim to model the current price per pound of a product based on factors including the old price per pound, month, vendor, and fruit category. Using a Bayesian framework, this model provides probabilistic estimates for how each factor individually impacts the current price, offering a comprehensive understanding of the determinants of product pricing while accounting for uncertainty in the estimates.

$$y_i = \beta_0 + \beta_1 \cdot \text{OldPricePerLb}_i + \beta_2 \cdot \text{Month}_i + \beta_3 \cdot \text{Vendor}_i + \beta_4 \cdot \text{FruitCategory}_i + \epsilon_i$$

Where:

- y_i: The current price per pound for product i.
- β_0: Intercept term, representing the baseline predicted price when all independent variables are at their reference levels.
  Prior: $\beta_0 \sim \mathcal{N}(0, 2.5)$

- β_1: Effect of the old price per pound on the current price.
  Prior: $\beta_1 \sim \mathcal{N}(0, 2.5)$
- β_2: Effect of the month on the current price.
  Prior: $\beta_2 \sim \mathcal{N}(0, 2.5)$
- β_3: Effect of the vendor on the current price.
  Prior: $\beta_3 \sim \mathcal{N}(0, 2.5)$
- β_4: Effect of the fruit category on the current price.
  Prior: $\beta_4 \sim \mathcal{N}(0, 2.5)$
- ϵ_i: Residual term, assumed to follow a normal distribution with zero mean.
  Prior for residual standard deviation (σ): $\sigma \sim$ Exponential(1)

3.2 Model justification

Table 2: Model Comparison: Frequentist and Bayesian Metrics

Model	Variables	R^2	Adjusted R^2	AIC	BIC	WAIC	LOO	RMSE
Model 1 (Frequentist)	Old Price Per Lb	0.08434	0.08378	5487.941	5504.130	NA	NA	1.30036
Model 2 (Frequentist)	Old Price Per Lb, Month	0.09945	0.09835	5462.809	5484.394	NA	NA	1.28958
Model 3 (Frequentist)	Old Price Per Lb, Month, Vendor	0.11358	0.11195	5439.033	5466.014	NA	NA	1.27942
Model 4 (Frequentist)	Old Price Per Lb, Month, Vendor, Fruit Category	0.29034	0.28771	5082.526	5125.697	NA	NA	1.14478
Model 5 (Bayesian)	Old Price Per Lb, Month, Vendor, Fruit Category	NA	NA	NA	NA	-2541.461	-2541.468	1.14478

Table 2 compares the performance of five models: Models 1–4 are frequentist models, while Model 5 is a Bayesian model. Among the frequentist models, Model 4 performs the best. Its R^2 is 0.29034, and its Adjusted R^2 is 0.28771, indicating that it explains the largest proportion of variance in the dependent variable. From Model 1 to Model 4, the inclusion of predictors such as `Month`, `Vendor`, and `Fruit Category` leads to a gradual improvement in R^2 values.

Model 4 also has the lowest AIC (5082.526) and BIC (5125.697) among the frequentist models, suggesting it achieves the best balance between model fit and complexity. Additionally, Model 4 has the lowest RMSE (1.14478), demonstrating that it provides the most accurate predictions among the frequentist models.

For the Bayesian Model 5, its WAIC (-2541.461) and LOO (-2541.468) are significantly lower than the AIC and BIC values for the frequentist models. These metrics indicate that the Bayesian model outperforms the frequentist models in predictive accuracy and its ability to generalize to new data. Furthermore, Model 5's RMSE is the same as Model 4's, at 1.14478, showing similar performance on observed data. However, the Bayesian model provides the additional advantage of quantifying uncertainty in predictions, making it more robust for predictive tasks.

In summary, while Model 4 is the best frequentist model, Model 5 (Bayesian) performs better in terms of predictive accuracy, as reflected by its significantly lower WAIC and LOO values. Therefore, for predicting the future prices of common fruits, Model 5 (Bayesian) is the best choice.

3.2.1 Defining and Justifying Reasonable Priors for the Bayesian Model

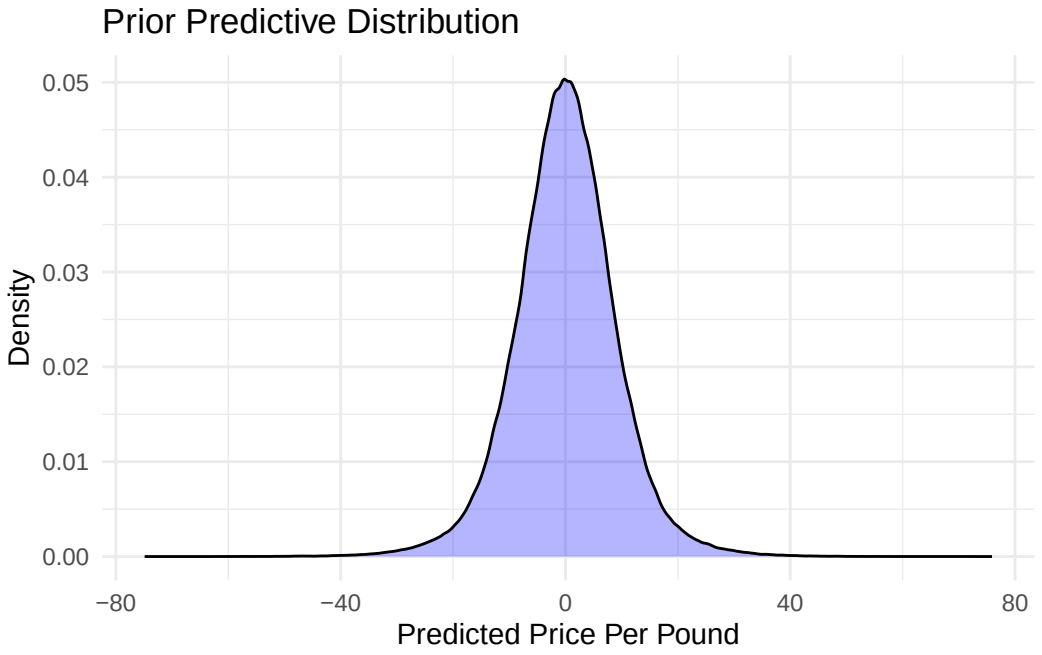

Figure 6: Prior Predictive Distribution

To validate the reasonableness of the priors in the Bayesian model, we generated prior predictive distributions and analyzed them through visualization. The priors in the model include

a normal distribution (N(0, 2.5)) for the intercept and coefficients, and an exponential distribution (Exponential(1)) for the residual standard deviation (( )). These priors are weakly informative, allowing sufficient flexibility without imposing excessive subjective assumptions.

The generated prior predictive distribution, as shown Figure 6, is a symmetric normal distribution centered around 0, with a wide range covering both negative and positive values. This distribution indicates that, in the absence of data constraints, the model allows for extreme scenarios, such as negative or very high prices. However, this does not imply that the model is flawed or unreasonable; rather, it reflects the mathematical properties of the normal distribution, which has a range extending from negative infinity to positive infinity. At this stage, the predictions are solely based on the priors, without the influence of actual data constraints.

Although actual prices cannot be negative, the negative values in the prior predictive distribution highlight the nature of weakly informative priors: they provide the model with considerable flexibility in the absence of sufficient data, allowing the data to take the lead in subsequent inferences. During the actual model fitting process, the data will constrain the priors, updating the posterior distribution and eliminating the unrealistic negative price values.

4 Results

4.1 Estimates of Predictors

Figure 7 illustrates the estimated coefficients of the predictors in the regression model for the current price per pound, along with their uncertainty intervals. The horizontal axis represents the estimated coefficient values, while the vertical axis lists the predictor variables. Each point indicates the estimated coefficient, and the horizontal lines represent the 95% confidence intervals, reflecting the uncertainty in the estimates.

The intercept's estimated value is approximately 1.5, indicating that when all other predictors are at their baseline levels, the baseline current price per pound is $1.5. The coefficient for past price (old_price_per_lb) is significantly positive, suggesting that higher past prices are associated with higher current prices, highlighting the strong influence of historical pricing on current pricing. The coefficient for month is negative, indicating that as the months progress, there may be a slight downward trend in the price per pound.

The coefficient for vendor Voila is about 0.6, suggesting that prices from Voila are slightly higher compared to the baseline vendor. The impact of different fruit categories on pricing is also notable: the coefficient for banana (fruit_categoryBanana) is positive, indicating that banana prices are higher than the reference category; melon (fruit_categoryMelon) has the largest coefficient, showing that melon prices are significantly higher than other categories; while orange (fruit_categoryOrange) has a coefficient close to zero, indicating no significant difference compared to the reference category.

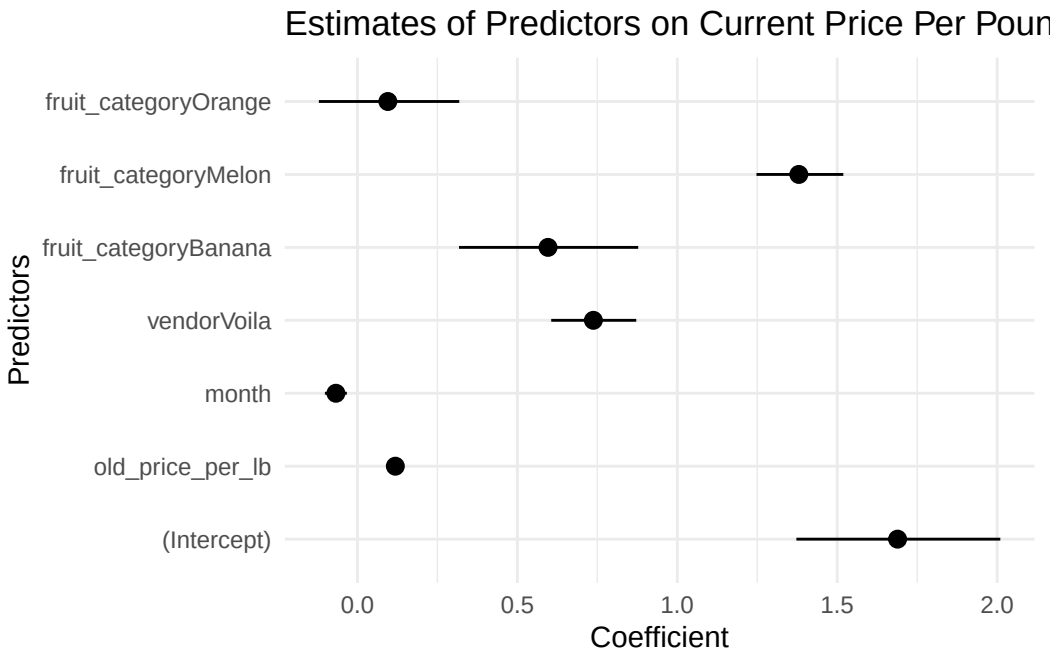

Figure 7: Estimates of Predictors on Current Price Per Pound

In summary, different fruit categories, vendors, and past prices have a significant impact on current prices, while the influence of the month is relatively minor.

4.2 Residuals Distribution

Figure 8 shows the distribution of residuals from the predictive model, which measures the difference between the observed and predicted prices per pound. Residuals are plotted on the x-axis, while their frequency is displayed on the y-axis.

The residuals are mostly centered around 0, indicating that the model's predictions are reasonably accurate for a majority of the data points. The highest frequency of residuals lies close to zero, highlighting the model's good fit to the observed data. However, there is some variability in the distribution, with residuals extending both to the left and right of zero.

On the positive side, a small cluster of residuals around 2.5 suggests that the model has underestimated the observed values for a few data points. Similarly, residuals spread slightly to the left of zero, indicating instances where the model overestimated prices.

Overall, the residual distribution is fairly symmetrical and concentrated near zero, suggesting that the model does not exhibit significant bias in its predictions.

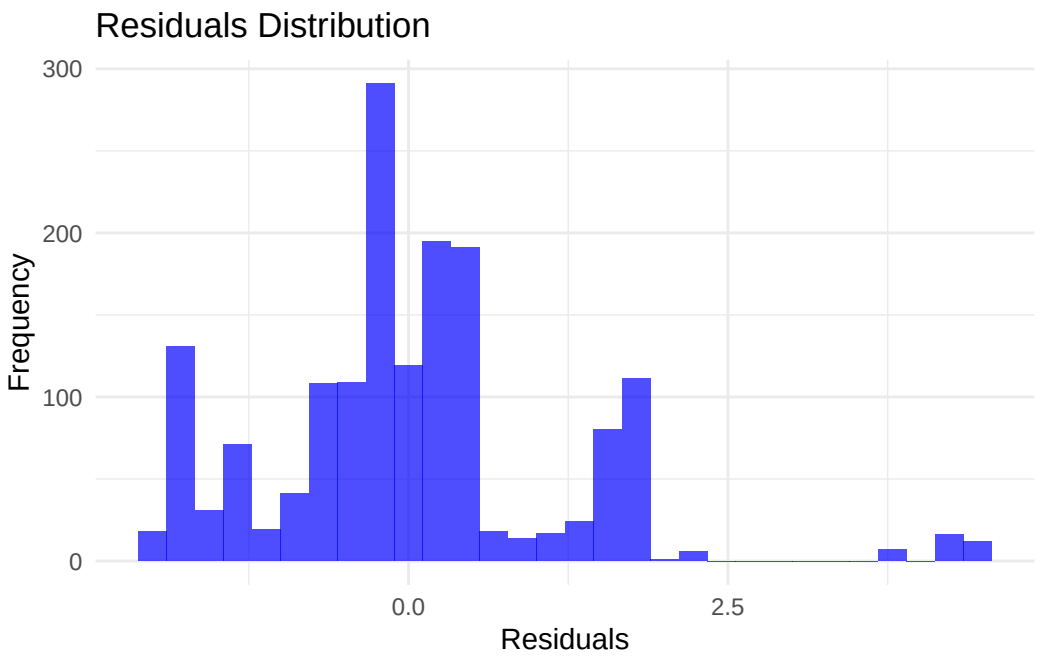

Figure 8: Residuals Distribution

4.3 Predicted Price Per Pound by Month, Fruit Category, and Vendor

Figure 9 illustrates the predicted price per pound for different fruit categories (Apple, Banana, Melon, Orange) across the months of June to November (6 to 11) for two vendors, Metro and Voila. The data is displayed in two facets, one for each vendor, to facilitate a clear comparison of trends.

For the Metro vendor, melons consistently have the highest predicted price per pound, maintaining a stable trend throughout the observed months. Bananas follow with relatively stable prices that remain below those of melons. Oranges exhibit a steady price trend in the lower range, while apples show significant variation, with a sharp price increase in months 10 and 11.

In contrast, the Voila vendor also sees melons as the most expensive fruit category, though their prices gradually decline over the months. Bananas maintain a middle price range, showing a consistent decrease from month 6 to month 11. Oranges and apples are the most affordable categories, with oranges slightly declining in price over time and apples showing minimal variation.

Across both vendors, melons stand out as the most expensive category, likely due to higher production costs or seasonal demand. Apples and oranges remain the least expensive, rarely

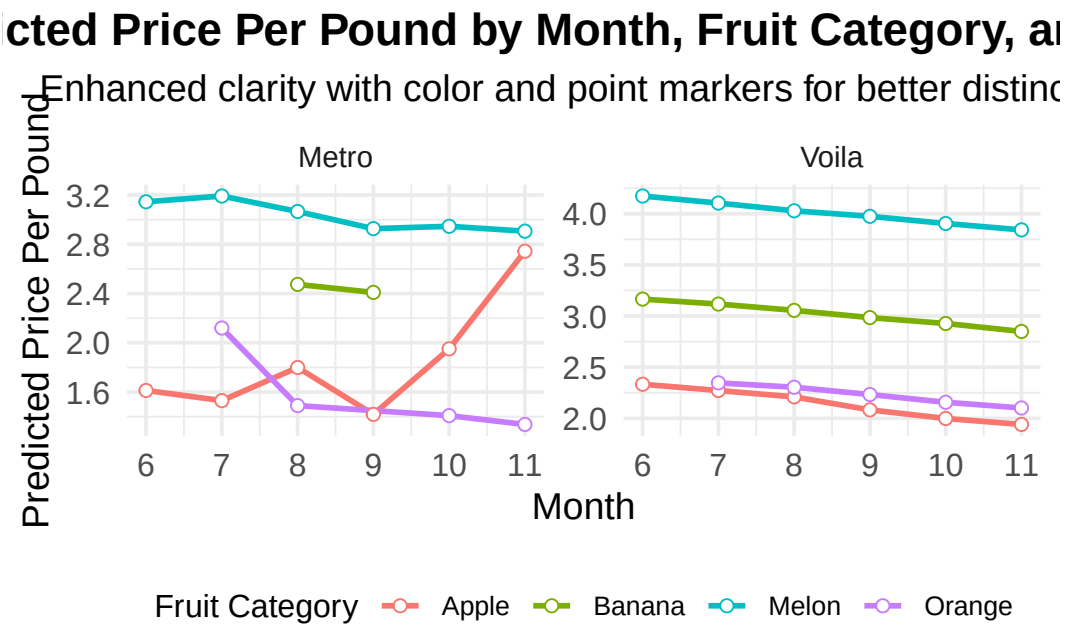

Figure 9: Predicted Price Per Pound by Month, Fruit Category, and Vendor

exceeding $2 per pound. Notably, Metro displays more variability in apple prices, while Voila exhibits a gradual decline across most categories.

5 Discussion

5.1 Summary of Findings

This study employs Bayesian regression models to reveal the factors influencing fruit prices, offering significant insights for consumers, producers, retailers, and policymakers. For consumers, the research provides price transparency, enabling them to better plan their purchasing decisions. By understanding the impact of historical prices, seasonal fluctuations, and supplier differences on current prices, consumers can choose to buy fruits from specific suppliers or during periods of lower prices, thus saving money. Additionally, consumers can optimize their dietary choices based on the pricing characteristics of different fruit categories and leverage the findings to avoid price exploitation due to information asymmetry.

For producers, the study provides guidance on optimizing production plans. The analysis of historical prices and seasonal trends offers market-oriented insights, helping producers adjust planting areas and harvest schedules to maximize profits. Moreover, understanding retailers'

pricing strategies during different months can enhance producers' market competitiveness and improve their supply strategies.

Retailers can use the findings to design precise pricing strategies. For instance, they can adjust inventory and promotional efforts for high-margin fruits to increase profitability and attract more customers. Additionally, the analysis of price differences among suppliers and fruit categories can help retailers optimize procurement and logistics processes, reducing costs and improving supply chain efficiency.

For policymakers, the study offers valuable support in several areas. First, it can inform the design of food price monitoring and regulation policies to stabilize market supply and protect consumer rights. Second, by identifying price differences and market dynamics among suppliers, the research provides a basis for developing fair competition regulations. Lastly, the results can support targeted food subsidies for low-income groups, ensuring they can afford a healthy diet.

From a societal perspective, this study promotes transparency in the food market, helping stakeholders make more rational and optimized decisions. In the future, as analytical methods advance, such as incorporating consumer behavior data and macroeconomic indicators, this research can expand to include other food categories, contributing further to improving the efficiency and fairness of the overall food system.

5.2 Limitation

Despite its contributions, the study has several limitations that warrant attention for future research:

5.2.1 Temporal Scope

The dataset only spans from June to November, omitting the remaining months of the year. This limited coverage excludes potential seasonal trends and significant pricing dynamics that may occur during unobserved periods, such as holiday-related demand surges or long-term trends.

5.2.2 Lack of External Factors

The model does not incorporate external factors that could significantly influence pricing, such as economic conditions (e.g., inflation rates, consumer confidence indices), weather patterns (e.g., extreme weather events, precipitation), or marketing strategies (e.g., discounts, advertising campaigns). These factors may have a direct or indirect impact on vendor pricing and consumer purchasing behaviors.

5.2.3 Absence of Interaction Effects

Interaction terms were not included in the model, limiting insights into how combinations of variables jointly influence pricing. For example, interactions between vendors and fruit categories could reveal vendor-specific pricing strategies for certain fruits. Similarly, interactions between seasons and fruit categories could capture seasonal price variations for specific fruits.

5.2.4 Limitations of Prior Distributions

The Bayesian model employed weakly informative priors, allowing flexibility in predictions. However, the current priors (e.g., normal distributions) were not restricted to positive values, which could result in unrealistic negative price predictions. While posterior distributions corrected this issue with sufficient data, such negative priors might raise concerns about the model's validity, especially in data-sparse scenarios.

5.2.5 Regional Generalizability

The dataset's regional specificity may limit the generalizability of the findings to other markets. Consumer preferences, market structures, and climate conditions vary significantly across regions, potentially leading to different pricing dynamics. The current model may not apply universally without adaptation.

5.3 Future Research

To address these limitations and expand the scope of the study, future research could focus on the following directions:

5.3.1 Extending Temporal Coverage

Future studies should collect datasets covering the entire year to explore complete seasonal patterns, long-term trends, and the effects of specific events (e.g., holiday promotions and non-seasonal supply-demand changes).

5.3.2 Incorporating External Factors

Adding economic, climate, and marketing variables, such as inflation, weather anomalies, and promotional activities, could improve the model's explanatory power and predictive accuracy. These factors would provide a more holistic understanding of pricing dynamics.

5.3.3 Including Interaction Effects

Future models could incorporate interaction terms, such as those between vendors and fruit categories or between seasons and fruit categories. These terms could uncover more complex relationships, such as vendor-specific pricing strategies for premium fruits during peak seasons.

5.3.4 Optimizing Prior Distributions

Employing stricter prior distributions in Bayesian models, such as truncated normal or log-normal distributions, would ensure realistic positive price predictions, further enhancing the model's reliability and practical applicability.

5.3.5 Cross-Regional Validation

Applying the model to different regions or countries could test its robustness and adaptability. Comparing price dynamics across various markets would uncover global commonalities and regional differences, enriching the understanding of fruit pricing mechanisms.

A Appendix

A.1 Appendix: Survey Design

To complement the observational data and provide a comprehensive understanding of fruit pricing dynamics, we designed a consumer survey to gather insights into purchasing behaviors, price perceptions, and vendor preferences. This survey aims to address key research questions: What factors influence consumers' fruit purchasing decisions, such as price, freshness, and brand? How do consumers perceive the pricing and quality of different vendors? Are consumers aware of seasonal variations in fruit prices?

The target audience includes adults aged 18 and above who regularly purchase fruits, focusing on the geographic regions covered by the observational data. We employed a stratified random sampling approach to ensure representation across different income levels, age groups, and regions.

The survey questionnaire is divided into four sections. The demographics section collects information on age, household income, and geographic location. The purchasing behavior section explores the frequency of fruit purchases and the impact of factors like price and freshness on purchasing decisions. The vendor perception section evaluates consumer preferences for specific vendors (e.g., Metro, Voila, Galleria, and Walmart) and rates these vendors based on price, quality, and variety. Finally, the price perception section asks respondents about price differences across vendors, their awareness of seasonal price variations, and their expectations for reasonable prices for fruits such as apples, bananas, melons, and oranges.

The survey data will serve multiple purposes. First, it will validate the accuracy of observational data by comparing consumer-reported price perceptions with actual vendor prices. Second, it will identify key factors influencing purchasing behaviors and vendor selection. Lastly, it will analyze consumer awareness of seasonal price variations and its alignment with observed trends. By incorporating consumer perspectives, this survey provides valuable context for the pricing models, enhancing the practical implications of the study's findings.

A.1.1 Questionnaire

A.1.1.1 Section 1: Demographics

1. Age:

- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 and above

2. Household Income:

- Less than $25,000
- $25,000–$49,999
- $50,000–$74,999
- $75,000–$99,999
- $100,000–$149,999
- $150,000 and above

3. Geographic Location:

- City/Town: _____
- State/Province: _____

A.1.1.2 Section 2: Purchasing Behavior

4. How often do you purchase fruits?

- Daily
- Weekly
- Bi-weekly
- Monthly
- Rarely

5. What factors influence your fruit purchasing decisions? (Select all that apply)

- Price
- Freshness
- Brand
- Vendor reputation
- Seasonal availability
- Other (please specify): _____

6. Do you prioritize organic fruits over non-organic?

- Yes
- No
- Sometimes

A.1.1.3 Section 3: Vendor Perception

7. Which vendor do you prefer for purchasing fruits?

- Metro
- Voila
- Galleria
- Walmart
- Other (please specify): _____

8. Rate the following vendors based on your experience (1 = Poor, 5 = Excellent):

- Metro: [1] [2] [3] [4] [5]
- Voila: [1] [2] [3] [4] [5]
- Galleria: [1] [2] [3] [4] [5]
- Walmart: [1] [2] [3] [4] [5]

9. What do you value most in a vendor? (Select all that apply)

- Price
- Quality
- Variety
- Convenience

A.1.1.4 Section 4: Price Perception

10. Do you notice significant price differences between vendors?

- Yes
- No
- Unsure

11. Do you observe seasonal variations in fruit prices?

- Yes
- No
- Unsure

12. What do you consider a reasonable price per pound for the following fruits?

- Apples: _____
- Bananas: _____
- Melons: _____
- Oranges: _____

References

Arel-Bundock, Vincent. 2023. "Modelsummary: Create, Edit, and Display Summary Tables for Statistical Models in r." https://vincentarelbundock.github.io/modelsummary/.

Filipp, Jacob. 2024. "Hammer: Grocery Price Dataset." https://jacobfilipp.com/hammer/.

Gabry, Jonah et al. 2023. "Bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

Goodrich, Ben, Jonah Gabry, Imad Ali, Sam Brilleman, and Josh Pritikin. 2023. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill, and Jennifer Bryan. 2023. "Here: A Simpler Way to Find Your Files." https://here.r-lib.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Jeroen Ooms, et al. 2023. "Arrow: Integration to Apache Arrow Library." https://arrow.apache.org/docs/r/.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2023. "Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics." https://ggplot2.tidyverse.org/.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. "Dplyr: A Grammar of Data Manipulation." https://dplyr.tidyverse.org/.

Zhang, Haoyu. 2023. "kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax." https://haozhu233.github.io/kableExtra/.