Pricing Patterns in Organic Markets*

Comparative Analysis of Vendor Strategies and Seasonal Trends in Organic Product Sales

Dingshuo Li

December 3, 2024

This paper investigates pricing patterns in the organic food market by analyzing how vendor strategies and seasonal trends influence product prices. Using a dataset of organic food prices across multiple vendors and months, we employed a Bayesian regression model to understand the interplay between historical prices, vendor-specific strategies, and seasonal factors. The analysis revealed significant seasonal price fluctuations and demonstrated that vendors rely heavily on historical prices to inform current pricing strategies. These findings provide valuable ideas into consumer behaviour, vendor decision-making, and the economic dynamics of organic markets, highlighting the broader implications for market competition and sustainable agricultural practices.

Table of contents

*Code and data are available at: https://github.com/dawsonlll/Organic-Product-Pricing-Analysis.git.

1 Introduction

The organic food market has experienced a significant increase in demand, driven by growing consumer awareness of health and environmental impacts. This demand has not only raised the market value of organic products but has also introduced challenges to their pricing strategies across various retail platforms. Organic foods generally incur higher production

and certification costs compared to conventional products, which affects their retail pricing and accessibility. This study examines these pricing strategies by analyzing how organic food prices vary across vendors and over time.

This paper investigates the dynamics of organic food pricing using a dataset that records prices from multiple vendors over a specific period. The analysis addresses the need for detailed quantification of how prices for organic products are structured and vary in real-market conditions. By applying Bayesian statistical methods, this research evaluates the relationship between current and historical prices while considering the effect of temporal factors, such as monthly variations.

Estimand: The primary focus of this study is the price elasticity of organic foods across different vendors and time periods. Specifically, the analysis estimates how changes in old prices influence current pricing strategies, accounting for differences across months and vendors. This approach allows for the measurement of the direct and interactive effects of month and vendor on the current prices of organic products, offering a deeper understanding of the responsiveness of organic food prices to market conditions and vendor-specific factors.

The findings highlight distinct pricing patterns among vendors and show significant fluctuations aligned with seasonal trends. These results are relevant for consumers making informed purchasing decisions and for vendors optimizing strategies to maintain or grow their market share in the competitive organic food sector. Understanding these pricing mechanisms has implications for marketing, consumer behavior, and the economic sustainability of organic farming practices.

The paper is structured to provide a clear, logical flow from methodology to implications Section 2 outlines our dataset selection and cleaning, ensuring transparency. The Section 3 details the Bayesian model and its justifications. The Section 4 displays our findings with visual graphs, while the Section 5 interprets these findings in the context of market dynamics. As well as highlights areas for further study, and the limitation of study.

2 Data

2.1 Overview

All data analysis was conducted using R R Core Team (2023a), including statistical computing and graphics. And the following R packages were used for conduct code in scripts: tidyverse Wickham et al. (2019), Palmerpenguins Horst, Hill, and Gorman (2020), ggplot2 Wickham (2016), Dplyr Wickham et al. (2023), Knitr Xie (2023), modelsummary Arel-Bundock (2022), arrow Richardson et al. (2024), here Müller (2020), rstanarm Goodrich et al. (2023), sclaes Wickham, Pedersen, and Seidel (2023), splines R Core Team (2023b),broom Robinson, Hayes, and Couch (2023), lubridate Grolemund and Wickham (2011), janitor Firke (2023), stringr Wickham (2022) and testthat Wickham (2011). Following Alexander (2023a) we conducted

data simulate, test simulated data, data cleanning, test analysis data, EDA, data modelling. The R code in scripts were adapted from Alexander (2023b)

2.1.1 Data Source

The dataset analyzed in this study was sourced from Jacob Filipp's Hammer Project Filipp (2024), which compiles detailed pricing data on organic food products across various retail vendors. Its extensive coverage of vendor-specific and temporal pricing strategies makes it particularly suitable for observing seasonal trends and competitive behaviors in the organic food market. The dataset's granularity and quality, including data from multiple vendors over an extended period, provide a robust foundation for analyzing pricing dynamics.

2.2 Measurement

The dataset analyzed in this study provides detailed records of organic food prices across vendors and months, capturing the dynamic pricing strategies within the competitive organic food market. Monthly price recordings reflect changing market conditions and facilitate the analysis of trends and pricing behavior over time.

The data collection involved systematically recording prices directly from grocery merchants. Each price entry represents a vendor's decision on how much to charge for an organic product at a given time. Two key variables are examined: the current price, observed directly at the time of collection, and the old price, the previously recorded price for the same product. The old price serves as a historical reference, aiding in the analysis of price adjustments over time.

Each entry in the dataset includes metadata detailing the vendor, collection date, and product information, ensuring accuracy and reliability. This structured approach supports robust analysis and allows for meaningful conclusions about organic food pricing dynamics.

2.3 Data Cleanning

The analysis began with the consolidation of raw data from two sources: product records and product details. These datasets contained over 13 million rows of diverse product data, including non-organic items and vendors outside the scope of this study. The product records raw_data included timestamps, vendor details, and product identifiers, while the product details dataset product_data provided information such as product names, brands, pricing history, and unit measurements.

To prepare the data, the two datasets were merged using an inner join on the product_id, a unique identifier linking transaction records to specific product details. This process enriched

each transaction record with relevant product information, creating a detailed view of each product.

After merging, the dataset still included a wide range of products and vendors, many of which were irrelevant to the study's focus on organic food pricing among major vendors. The first step in refining the dataset was vendor selection. Transactions were filtered to include only those from major vendors with a significant presence in organic food sales: "Metro," "NoFrills," "Walmart," "Voila," "TandT," and "Galleria." This selection focused the analysis on pricing strategies within competitive segments of the retail market.

The current price and old price were parsed to ensure that they were represented as numerical values, free from any non-numeric characters that could interfere with quantitative analysis. This parsing was important for accurate statistical calculations and comparisons. The nowtime field, representing the timestamp of each transaction, was used to extract the month component. This transformation was essential for analyzing seasonal trends and month-to-month pricing variations. Lastly, entries with missing or zero values in current price were removed to maintain the integrity of pricing analysis. Additionally, any remaining rows with incomplete data were dropped, ensuring that the dataset used in subsequent analyses was entirely robust and clean.

2.4 Outcome variables

2.4.1 Predictor variables

The old price variable, representing the product's price from the preceding month, serves as a essential metric for understanding the evolution of market trends and pricing strategies over time. The summary Table 1 reveal a broad range of "Old Price" from a minimum of $1.69 to a maximum of $38.56, with a median price of $5.49, suggesting a median market position. The mean, slightly higher at $7.66, indicates a positive skew in the data, with the first quartile at $3.99 and the third at $10.49, reflecting substantial variability in historical pricing. This variability underscores the diverse pricing strategies across different vendors and product types, and provides a foundation for analyzing how prices adjust in response to market dynamics, consumer demand, and possibly supply chain fluctuations.

Table 1: Overview of Historical Price Trends

Statistic	Value
Min	1.69000
1st Qu.	3.99000
Median	5.49000
Mean	7.65968
3rd Qu.	10.49000
Max	38.56000

Table 1: Overview of Historical Price Trends

Statistic	Value

The vendor variable serves as a categorical identifier for vendors in the dataset, enabling the analysis of pricing strategies across different retail environments. It is essential for comparing and contrasting market behaviors, as it divides the dataset into distinct groups based on the retailer. The distribution of data Table 2 shows significant variation among vendors, with Metro being the most frequently represented, with 1,981 entries, followed by Voila with 1,513 entries. Walmart, Galleria, TandT, and NoFrills appear less frequently, reflecting differences in the scale and possibly the geographic reach of these vendors

This distribution suggests that Metro and Voila dominate the dataset, potentially indicating a larger influence or market share in the organic products sector compared to other vendors. Understanding the pricing strategies of these vendors is important for analyzing market trends, as each vendor may respond differently to market pressures, consumer preferences, and seasonal changes. Analyzing this variable helps identify patterns specific to certain types of vendors and provides a better understanding of the competitive dynamics of the organic food market.

Table 2: Distribution of Organic Product Listings Across Vendors

Vendor	Frequency
Metro	1981
Voila	1513
Walmart	588
Galleria	173
TandT	15
NoFrills	5

The "Month" variable in the dataset, recorded as a numerical value, serves as a important indicator for analyzing seasonal impacts on organic food pricing. It reflects when the data was collected, revealing essential patterns of how prices fluctuate through different months, likely influenced by consumer demand, supply variations, and promotional periods. The distribution of observations shown in Table 3, the peaks are in summer season suggests significant market activity during the summer months, potentially driven by an increased availability of fresh produce or higher consumer buying rates. In contrast, the notably lower entries in spring and winter indicate reduced activity, which could be due to offseason market slowness or less frequent data collection. This variable is integral to understanding how temporal factors influence pricing strategies, providing valuable strategy for vendors, consumers, and analysts studying market dynamics in the organic food sector.

Table 3: Monthly Distribution of Organic Product Listings

Month	Count
2	4
3	44
4	31
5	8
6	999
7	1253
8	804
9	691
10	353
11	88

2.4.2 Response Variable

The response variable current price represents the most recent price at which organic food products were sold across various vendors. The summary statistics Table 4 for this variable provide a snapshot of its distribution within the dataset. The minimum price recorded is 0.97, indicating a lower-end pricing point for some products, while the maximum price peaks at 29.99, reflecting premium offerings within the organic sector. The first quartile is 3.22, suggesting that at least 25% of the products are priced at or below this level, which could be indicative of basic or smaller-sized organic goods. The median price of 4.49 and the mean of 5.69 suggest a central tendency towards moderately priced items, with the mean being slightly higher due to the influence of more expensive products pulling the average up. The third quartile at 7.49 shows that 75% of the products fall below this price point, highlighting a pricing structure that might cater more to a mid-range consumer segment. This look at current price helps in understanding the economic landscape of organic food pricing and its variability, which is important for analyzing market strategies and consumer behaviour.

Table 4: Summary of Current Price Distribution

Statistic	Value
Min	0.970000
1st Qu.	3.220000
Median	4.490000
Mean	5.695207
3rd Qu.	7.490000
Max	29.990000

3 Model

3.1 Model Overview

In our study, we are particularly interested in understanding how the current price of organic products is influenced by several key factors: the price of the product in the previous month, the vendor from which the product is purchased, and the time of year. To systematically examine these relationships, we developed a Bayesian regression model. This model incorporates these variables to elucidate their impacts on current pricing strategies. By leveraging historical pricing data alongside vendor-specific information and temporal dynamics, our model aims to uncover meaningful patterns that can inform both theoretical understanding and practical applications in market pricing strategies.

3.2 Model set-up

$$\text{Current Price}_i = \beta_0 + \beta_1 \times \text{Month}_i + \beta_2 \times \text{Old Price}_i + \sum(\beta_{\text{vendor}} \times \text{Vendor}_i) + \epsilon_i$$

Where:

- Current Price_i is the dependent variable representing the price of organic products at time i.
- Month_i is a continuous independent variable reflecting the month of the year to account for seasonal effects on pricing.
- Old Price_i is a continuous independent variable representing the price of the product in the previous month, allowing for analysis of price evolution.
- Vendor_i is a categorical variable with levels for each vendor included in the study, capturing vendor-specific pricing strategies.
- β_0 is the intercept, $\beta_0 \sim \text{Normal}(0, 2.5)$.
- β_1, the coefficient for Month, $\beta_1 \sim \text{Normal}(0, 2.5)$.
- β_2, the coefficient for Old Price, $\beta_2 \sim \text{Normal}(0, 2.5)$.
- β_{vendor}, coefficients for each vendor level, $\beta_{\text{vendor}} \sim \text{Normal}(0, 2.5)$ for each vendor level.
- ϵ_i is the error term, assumed to be normally distributed.

We conducted the Bayesian analysis in R (R Core Team 2023a), utilizing the `rstanarm` package (Goodrich et al. 2023). To maintain consistency and reliability in our modeling approach, we employed the default priors. These priors are well-suited for a wide range of data types and ensure robustness in the inference process, particularly in complex models such as ours that involve multiple predictors and hierarchical structures. Normal priors with a mean of 0 and a standard deviation of 2.5 are assigned to the coefficients and the intercept. These priors were chosen to impose a slight regularization effect, reducing the risk of overfitting by moderating the influence of extreme values or outliers in the data.

3.3 Model justification

The Bayesian linear regression model was selected for this analysis to explore how various factors such as month, old price, and vendor influence the current price of organic products. This model type was chosen due to its flexibility in incorporating prior knowledge and handling uncertainty in estimates, which is particularly beneficial in markets where data may exhibit variability and non-standard distributions. Bayesian methods allow for the integration of prior beliefs or empirical evidence into the analysis, which can be particularly useful when dealing with organic product pricing where past market trends can inform current expectations. The primary assumptions include the linearity between predictors and the outcome, independent and identically distributed residuals, and normally distributed error terms. These assumptions are typical for linear regression but must be validated through diagnostic checks to confirm no significant deviations occur.

Unlike other regression, Bayesian regression provides a probabilistic approach to inference, offering a full posterior distribution of the parameters, which helps in understanding the uncertainty around the estimated effects. This model is robust to various types of data and does not strictly require the normality assumption for residuals that a classical linear regression would. This is particularly beneficial given the often skewed or heterogeneous nature of pricing data. Each predictor's influence is quantified by respective coefficients, adjusted for vendor-specific variations and temporal trends. The use of normal priors for these coefficients is justified by the need to regularize estimates, thus preventing overfitting and ensuring stable predictions even with potentially collinear or sparse data.

3.3.1 Comparison with Alternative Models

Although a standard multiple linear regression could model continuous outcomes effectively, it lacks the ability to incorporate prior distributions and assess uncertainty. odels such as logistic regression or Poisson regression were considered; however, these models are more suited to categorical or count data, respectively. Given that our outcome variable, current price is continuous, a linear approach is more appropriate. These were also evaluated, especially for capturing any non-linear dynamics between the predictors and the current price. While

potentially offering a closer fit to certain complex patterns, the increased model complexity could hinder interpretation and require more extensive data to validate effectively.

3.4 Model Vialidation

Model diagnostics included checks for multicollinearity, heteroscedasticity, and normality of residuals. Posterior predictive checks were performed to ensure the model adequately captures the observed data patterns. These steps help in verifying model assumptions and the appropriateness of the model for the given data. The background details and diagnostics are included in Appendix B.

4 Results

Our results are summarized in Table 5. The Bayesian regression model's outputs provide implications about the factors influencing the current prices of organic products. These findings are encapsulated in Table 5, which succinctly presents the estimated coefficients and their respective standard errors. This quantitative output, complemented by the visual representation in the corresponding plot, provides a clear perspective on the relative impacts of each predictor.

The intercept, estimated at 3.53, represents the average base price of organic products when all other predictor variables are held at zero. This baseline provide a benchmark against which the effects of other variables are measured.The coefficient for month, which is -0.42, suggests a monthly decrease in prices. This trend could be indicative of seasonal adjustments, where prices may drop during certain times of the year due to increased supply from harvest periods or reduced consumer demand. The negative value emphasizes the importance of considering temporal factors in pricing strategies. The positive coefficient for old_price (0.62) underscores a significant reliance on historical pricing data. This relationship indicates that prices are likely to increase if they were higher in the preceding period, reflecting perhaps market confidence or ongoing trends in consumer willingness to pay.

The vendor coefficients reveal distinct pricing strategies among different retailer. vendor Voila shows a notably higher positive coefficient of 1.1, suggesting that this vendor typically prices organic products higher than others. This could reflect a premium branding strategy or a targeted demographic willing to pay more for perceived quality. In contrast, NoFrills and TandT have coefficients of 0.09 and -0.02, respectively, indicating minimal deviation from the baseline prices, possibly due to a competitive pricing strategy aimed at attracting cost-conscious consumers.

Table 5: Overview of Bayesian Model Analysis

	Bayesian_model
(Intercept)	3.53
	(0.22)
month	−0.42
	(0.02)
old_price	0.62
	(0.01)
vendorMetro	0.35
	(0.17)
vendorNoFrills	0.09
	(0.97)
vendorTandT	−0.02
	(0.57)
vendorVoila	1.17
	(0.17)
vendorWalmart	0.19
	(0.19)
Num.Obs.	4275
R2	0.703
R2 Adj.	0.702
Log.Lik.	−9245.456
ELPD	−9253.9
ELPD s.e.	94.2
LOOIC	18 507.7
LOOIC s.e.	188.4
WAIC	18 507.7
RMSE	2.10

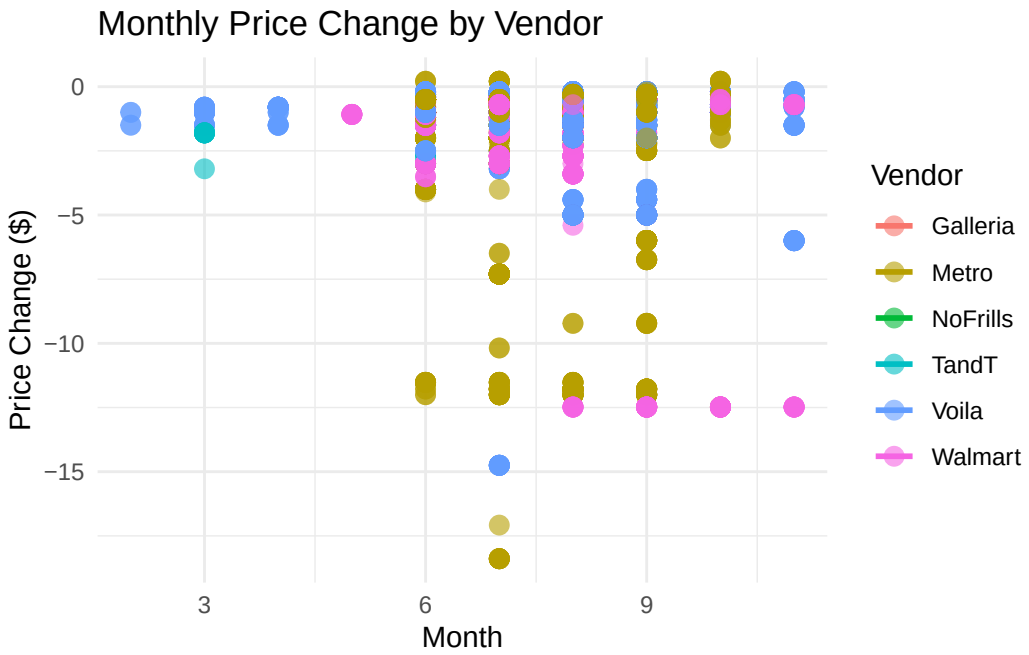


Figure 1: Vendor-Specific Monthly Price Variations

The Figure 1 illustrates the changes in pricing for organic products offered by six vendors over three key months: March, June, and September. It shows that most vendors have adjusted their prices moderately, with the majority of changes clustering a little below the $5 mark. This pattern suggests a trend where vendors generally reduce their prices slightly, potentially reflecting seasonal promotions or responses to market dynamics. Notably, some vendors like Metro exhibit more dramatic price reductions, indicating aggressive pricing strategies or perhaps an adjustment to overstock situations. This subtle yet consistent price reduction across most vendors could imply a market-wide strategy to boost sales volumes by making organic products more affordable to a broader customer base during these months.

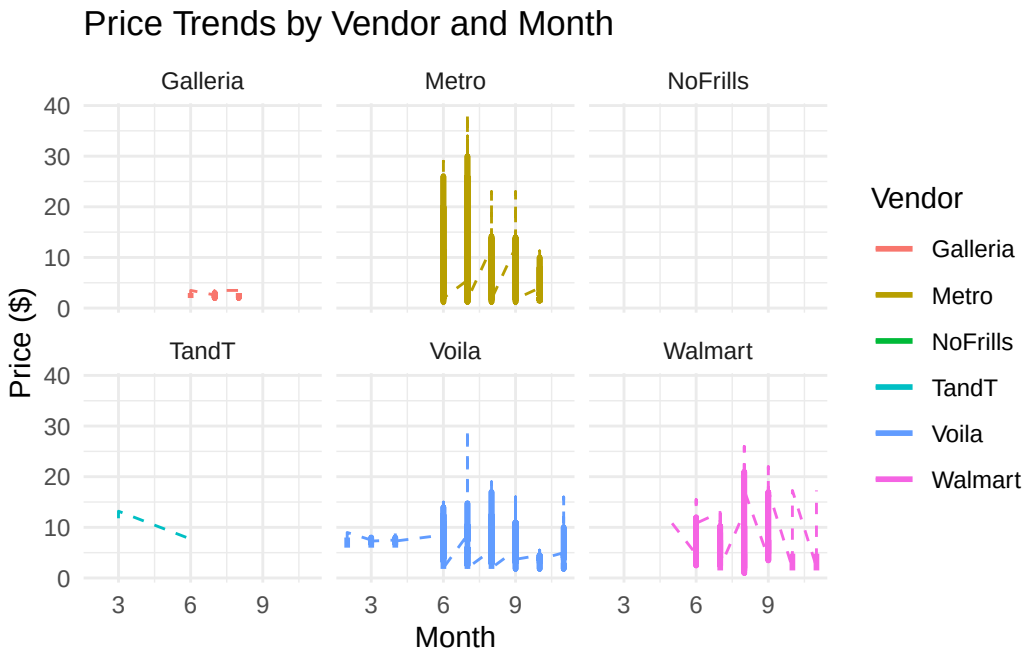


Figure 2: Monthly Price Trends Across Vendors

In the Figure 2, the pricing trends of presumably organic food items across six different vendors are displayed over months. The graph is segmented into six panels, each corresponding to a vendor: Galleria, Metro, NoFrills, TandT, Voila, and Walmart, allowing for a direct comparison of price fluctuations across different time points within the year. Each vendor's pricing data is presented with a unique colour code, maintaining consistency across the panels, which aids in quick visual comparison and analysis.

The analysis of the graph shows distinct trends and behaviors among the vendors. Galleria shows a minor incremental trend in pricing, suggesting a stable market position. In contrast, Metro displays notable volatility, with significant price spikes evident in June. NoFrills exhibits very limited data variability in the graph, which could be indicative of several market behaviors or strategies. This minimal fluctuation in pricing might suggest that NoFrills does not have a significant inventory of organic food products, perhaps due to a strategic focus on other product lines or a customer base that is less inclined towards organic options. TandT's prices slightly decline over the observed months, potentially reflecting seasonal promotions or inventory adjustments. Voila exhibits the most variability, with prices peaking significantly in summer seasons, which may reflect dynamic pricing strategies in response to demand changes or supply issues. Walmart, similar to Voila, shows price fluctuations with an increasing trend towards variability as the year progresses.

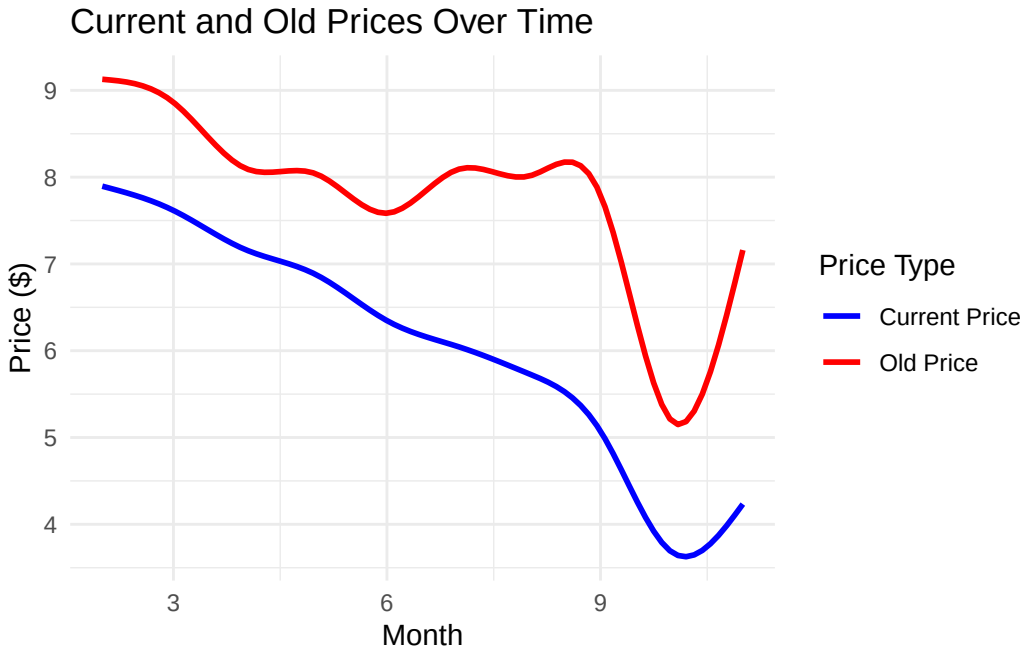


Figure 3: Comparison of Old and Current Prices for Organic Products

The Figure 3 delineates the evolution of pricing data under two categories: "Current Price" and "Old Price," over the months. This graphical representation enables an comparison of the pricing fluctuations throughout the year for a particular product or set of products. The current price, represented in blue, shows a downward trajectory over the year, falling from about 8 dollars to approximately 4 dollars. Conversely, the old price, depicted in red, demonstrates a similar decline from March to June but maintains a level higher than the current price during this period. Notably, there is a sharp dip in September where it plunges to below 6 dollars before slightly recovering. This divergence between the current and old prices from June to September, where the current price escalates substantially while the old price sharply declines, may suggest responses to external market forces such as supply chain disruptions, variations in seasonal demand, or changes in promotional strategies that impact pricing structures. Particularly, the pronounced drop in the old price during September could indicate strategies such as clearance sales or pricing adjustments intended to clear existing inventory or adapt to waning demand.

This overall trend where the current prices fall below the old prices could indicate a reduction in the market price of organic food. Several factors could contribute to this phenomenon, such as increased supply of organic products, enhanced efficiency in organic farming techniques, or perhaps a decrease in consumer demand. Moreover, such a price drop could also be a strategic response by vendors to attract more customers or to remain competitive in the market, especially if other factors like seasonal variations or promotional efforts play a role.

5 Discussion

5.1 Overview of Study

This paper investigates the pricing dynamics of organic products across various vendors over distinct months, employing a Bayesian regression model to elucidate the factors that significantly influence pricing. The model's findings, presented in Table 5, highlight how both temporal elements and historical pricing data impact current pricing strategies, alongside vendor-specific behaviors. Through a detailed examination of changes in pricing by month and vendor, the study provide a granular look at how seasonal adjustments, market confidence, and competitive strategies shape the pricing landscape of organic products.

5.2 Market Behaviour

The first significant learning from this study is the clear indication of seasonal pricing impacts. The negative coefficient for month suggests that prices generally decrease over time, likely due to seasonal produce availability or a strategic reduction in prices to clear inventory before new stock arrives. Additionally, this trend likely reflects the agricultural cycles of organic products, where increased supply during harvest periods may lead to reduced prices. This seasonal adjustment could be a strategic response to fluctuating consumer demand, which often wanes after peak buying seasons. Understanding these trends is important for vendors and suppliers in planning inventory and promotional strategies.

A second important revelation from the study pertains to the significant influence of historical prices on current pricing strategies. The regression model indicated this through a coefficient of 0.62 for old prices, indicating a strong positive relationship. This relationship suggests that vendors heavily consider past pricing when setting current prices, likely aiming to maintain price consistency and build consumer trust. Such a strategy may reflect market confidence, where past prices help shape consumer perceptions of value and quality, especially in markets where organic products are seen as premium goods.

5.2.1 Historical Pricing as a Market Predictor

Another essential observation from the study is the role of historical prices in shaping current pricing strategies. The positive relationship between past and present prices suggests a form of price inertia, where past price levels serve as a baseline for current pricing decisions. This phenomenon could be driven by consumer perceptions of value and quality, which once established, can allow vendors to maintain higher price points. This aspect of the pricing strategy is particularly relevant in the organic market, where consumers are often willing to pay a premium for products they perceive as healthier or more ethically produced. vendors leveraging

this might focus on building and maintaining a strong brand reputation that supports higher pricing.

The persistence of higher prices from one period to the next suggests a robust market memory and potentially indicates consumer tolerance for higher price points. This finding is highly relevant in the organic market, where branding and perceived quality are pivotal. Consumers often associate higher prices with superior quality, and this perception enables vendors to maintain premium pricing strategies, even amidst market fluctuations. This is particularly important for organic products, where quality and ethical sourcing are often key selling points that justify higher price tags. Thus, understanding the impact of historical pricing on current strategies provides vendors with information for crafting effective pricing models that resonate with consumer expectations and sustain market positioning.

5.2.2 Vendor-Specific Pricing Strategies

The analysis also highlighted significant differences in pricing strategies among vendors. For instance, the vendor Voila exhibited a notably higher coefficient of 1.1, suggesting a premium pricing strategy that targets consumers willing to pay more for perceived quality. In contrast vendors such as NoFrills and TandT appear to have a limited selection of organic food products, as evidenced by their minimal price variability and relatively small deviations from baseline pricing. This could suggest that organic products are not a significant focus in their inventory, possibly due to their target demographic or market strategy. Both vendors might prioritize affordability and accessibility, catering to price-conscious consumers who may not prioritize organic options. This limited emphasis on organic products could also reflect supply chain constraints or a strategic decision to focus on high-demand, non-organic items that align with their cost-competitive positioning in the market. By offering only a small range of organic products, these vendors may be targeting a niche segment of their customer base while maintaining their broader appeal as budget-friendly vendors.

5.3 Limitations of the Study

While the study provides valuable study, several limitations must be acknowledged. First, the reliance on historical pricing data may not fully capture real-time market dynamics or consumer behavior shifts, such as changes in consumer preferences due to economic factors or health trends. Additionally, the model assumes linearity and continuity in price adjustments, which may not hold in real-world scenarios where pricing decisions can be abrupt or influenced by non-continuous factors like marketing campaigns or regulatory changes.

5.4 Future Directions

Moving forward, several avenues can enhance our understanding of organic product pricing. Future research could integrate more granular data, such as specific product types within the organic category, to ascertain if different products exhibit unique pricing behaviors. Additionally, incorporating consumer data, such as demographics and purchase histories, could offer deeper into the demand side of the pricing equation. Methodologically, employing non-linear models or machine learning techniques might capture more complex interactions and non-linear dependencies not accommodated by the current model.

Furthermore, expanding the geographic scope to include different regions or comparing urban versus rural markets could reveal broader pricing trends and strategies. Lastly, longitudinal studies extending over several years would help in understanding long-term trends and the effects of macroeconomic factors on organic product pricing.

Appendix

A Methodological Framework for Data Collection and Analysis in Project Hammer

A.1 Introduction

Project Hammer is an innovative initiative designed to compile and make available database of historical grocery prices from top grocers' websites across Canada. This appendix delves into the detailed methodologies used for data collection, sampling, and observational analysis within this project. It also explores the application of simulation techniques and integrates extensive literature to support the methodologies used.

A.2 Data Collection Techniques

The primary method for data compilation in Project Hammer involved automated web scraping techniques. Scripts were developed to systematically collect price data from various online grocery platforms. This method ensured a high-efficiency collection of large volumes of data while maintaining consistency across different data points.

A.2.1 Web Scraping:

The primary data for Project Hammer was meticulously extracted from various top grocers' websites across Canada. Automated scripts using Python libraries like BeautifulSoup and Selenium were employed to scrape historical price data, ensuring efficient and consistent data retrieval across different data points. ### API Utilization: Where available, APIs provided by grocery chains were utilized to fetch data in a structured format, enhancing the accuracy and reliability of the collected data. This method streamlined the process, reducing the likelihood of errors commonly associated with manual data collection methods.

A.3 Data Integrity and Validation

Data Cleaning: Rigorous cleaning processes were applied to correct anomalies and fill missing values, ensuring high data quality for analysis. Validation Checks: Cross-verification with in-store price checks and secondary sources was conducted periodically to validate the data accuracy. Automated Checks: Scripts were run to identify missing data, outliers, and inconsistencies, which were then manually reviewed to determine the cause and necessary corrections.

A.4 Sampling Strategy

The sampling strategy for Project Hammer was carefully designed to capture dataset of grocery prices across Canada. Given the complex and varied nature of the Canadian grocery market, a stratified sampling approach was deemed most suitable for ensuring the data accurately reflected the diversity of market conditions and pricing strategies employed across different geographical areas and store types. Below is a detailed breakdown of each component of the sampling strategy.

A.4.1 Geographical Stratification

Canada was segmented into its major regions: Atlantic Canada, Quebec, Ontario, the Prairies, Alberta, and British Columbia. This segmentation reflects significant regional differences, including population density, urbanization, and economic activity, which can all influence grocery pricing. Within each region, further stratification was employed to differentiate between urban and rural areas. Urban areas tend to have higher competition among grocers and potentially lower prices due to larger market sizes, whereas rural areas might experience higher prices due to fewer competitors and higher logistics costs.

A.4.2 Random Sampling within Strata

To achieve a representative sample within each geographical stratum, random sampling was used. This method aims to reduce sampling bias that could skew the analysis, such as overrepresentation of major chains or urban centers. From each geographical and urban stratum, grocery stores were randomly selected. This included large national chains, regional chains, and independent stores to capture a broad spectrum of pricing strategies. Random sampling was not only applied to the selection of stores but also to the timing of data collection. Prices were recorded at different times—weekly, bi-weekly, and monthly—to account for any temporal variations in pricing that could result from promotional activities or seasonal changes.

A.4.3 Importance of the Sampling Strategy

This rigorous sampling methodology ensured that the Project Hammer dataset was robust, minimizing the potential biases that can arise from non-random sampling or over-concentration in certain geographical areas or store types. By reflecting the true variance across different market conditions and consumer demographics, the dataset allows for more accurate and generalizable conclusions to be drawn about pricing strategies and market dynamics in the Canadian grocery sector.

A.5 Simulation Techniques

Simulation techniques were essential in validating the web scraping algorithms and assessing potential sampling biases for Project Hammer. By generating synthetic datasets that emulate the dynamics of the Canadian grocery market, especially focusing on organic food pricing, the project team rigorously tested and refined our data collection and analysis methodologies.

The primary goal of these simulations was to evaluate the performance and accuracy of scraping algorithms and observe the impact of sampling strategies in a controlled environment, avoiding the costs and time associated with real data collection. The synthetic datasets included a wide range of organic food items with varied price points and historical price trends, accurately reflecting real market conditions. This allowed the team to enhance the scraping tools and sampling methods, ensuring they were robust and representative of actual market dynamics.

A.6 Linkages to Literature

The methodologies discussed in this appendix are grounded in established research, drawing on foundational resources like the International Handbook of Survey Methodology Dillman, Leeuw, and Hox (2008) and the article Statistical Data Integration in Survey Sampling: A Review. The International Handbook of Survey Methodology provides guidance on survey design, implementation, and ethical considerations. It emphasizes the importance of aligning sampling strategies, such as stratified and cluster sampling, with the survey's objectives to ensure representative data collection. This aligns with the section on sampling strategies in this appendix, where these methods are tailored for accuracy and efficiency. Additionally, the handbook's discussion of data collection methods—such as face-to-face interviews and online surveys—connects with the appendix's focus on selecting techniques based on the target population and study requirements. Furthermore, the handbook highlights ethical practices, including informed consent and confidentiality, which are reflected in this appendix.

The article Statistical Data Integration in Survey Sampling: A Review Kim, Kwon, and Skinner (2020) examines combining different data sources to improve survey estimation and analysis. It addresses challenges such as inconsistent structures and biases, which relates to the section on integrating survey and observational data, where merging datasets enhances analysis. The article discusses statistical techniques like propensity score matching and multiple imputation to improve data validity, complementing the appendix's emphasis on validation. It also highlights the use of simulation studies to test and refine sampling and data collection methods, which aligns with the simulation section in this appendix.

A.7 Conclusion

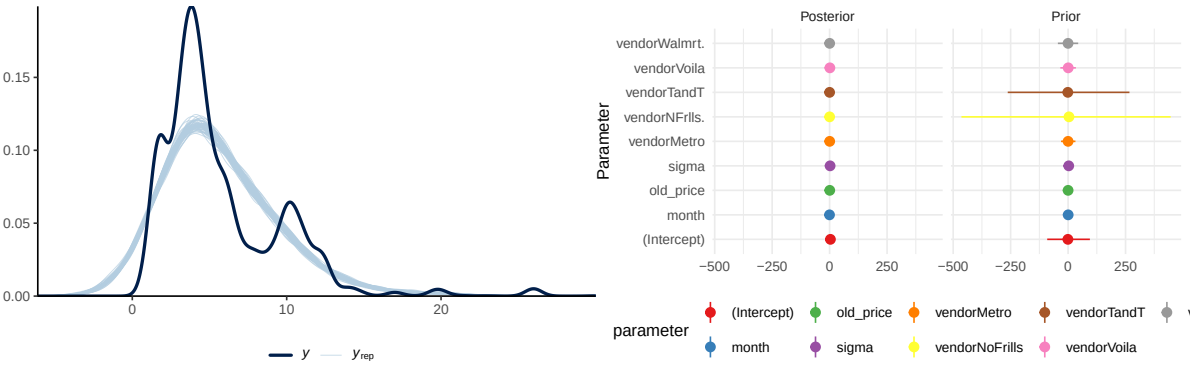
This appendix has provided a detailed account of the sophisticated methodologies employed in Project Hammer to compile a robust database of grocery prices. By leveraging advanced

web technologies, adhering to rigorous sampling strategies, and ensuring thorough validation processes, Project Hammer aims to foster more competition and reduce potential collusion in the Canadian grocery sector. These approaches not only ensure the reliability of the findings but also contribute significantly to the broader academic and legal discourse on market competition and regulatory compliance in the grocery sector.

B Model details

B.1 Posterior predictive check

In Figure 4a we implement a posterior predictive check to assess the fit of the model. The dark blue line represents the observed data distribution, while the lighter blue lines show the simulated data distributions generated from the posterior predictive distribution of the model. The simulated data predicted the overall trend and the alignment between the observed data and the model-generated data indicates that the model is capturing the key characteristics of the observed data well. This suggests that the model is a good fit for the dataset and effectively reflects the underlying generative process. In Figure 4b the comparison between posterior and prior distributions demonstrates how the observed data influenced the model's parameter estimates. The posterior distributions are narrower and more concentrated than the diffuse priors, indicating that the data provided substantial evidence to refine the estimates. Parameters like old_price and month show strong and precise posterior effects, highlighting their significant roles in explaining current prices. Vendor-specific coefficients vary, with Volia showing a notable positive effect, reflecting higher pricing strategies, while TandT and NoFrills remain close to zero, indicating minimal deviation from baseline pricing. This comparison underscores how the model relies on data to inform impactful predictors while maintaining prior regularization for less informative parameters, offering thinking into vendor-specific strategies and seasonal pricing trends.



(a) Posterior prediction check

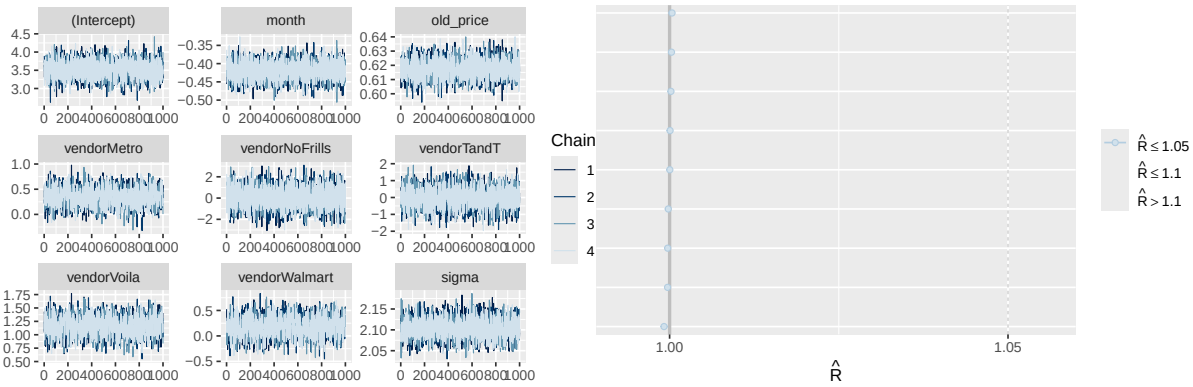
(b) Comparing the posterior with the prior

Figure 4: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 5a is a trace plot. It illustrates the convergence and mixing of the Markov Chain Monte Carlo (MCMC) sampling process for the Bayesian model parameters. Each parameter's plot Intercept, month, old_price and vendors are shown. The chains show consistent oscillations around a stable range without any noticeable trends or drifts, indicating that the sampler has likely converged. Parameters like old_price and vendorVoila display tightly clustered chains, suggesting precise estimates with relatively low uncertainty. In contrast, parameters such as vendorNoFrills and vendorTandT exhibit greater variability, reflecting higher uncertainty or minimal influence on the model. Overall, the chains demonstrate good overlap and mixing, which is important for thoroughly exploring the posterior distributions. These observations suggest that the MCMC algorithm has successfully converged, and the parameter estimates from the model are robust and reliable.

Figure 5b is a Rhat plot. It shows the R-hat diagnostic for all parameters in the Bayesian model. The R-hat statistic assesses the convergence of Markov Chain Monte Carlo simulations, comparing the variance within chains to the variance between chains. In this plot, all the R-hat values are at or very close to 1.00, well below the threshold of 1.1, indicating that the chains for all parameters have converged. The R-hat values being near 1.00 signify excellent mixing of the chains, meaning that each chain explores the parameter space effectively and independently of its starting values.



(a) Trace plot (b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. 2023a. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

———. 2023b. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Dillman, Don A., Edith D. de Leeuw, and Joop J. Hox. 2008. *International Handbook of Survey Methodology*. Taylor & Francis. https://www.taylorfrancis.com/books/edit/10.4324/9780203843123/international-handbook-survey-methodology-dillman-edith-de-leeuw-joop-hox.

Filipp, Jacob. 2024. *Hammer Project: Comprehensive Dataset on Organic Food Pricing*. https://jacobfilipp.com/hammer/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

Kim, J. K., H. Kwon, and C. J. Skinner. 2020. "Statistical Data Integration in Survey Sampling: A Review." *Survey Research Methods* 14 (1): 1–15. https://link.springer.com/article/10.1007/s42081-020-00093-w.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2022. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan,

Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.