Assessing the Multifaceted Influences on Current Coffee Prices*

Coffee Pricing in Canadian Grocery Stores: Current Prices Strongly Shaped by Historical Trends

Yi Tang

November 29, 2024

This paper investigates the factors influencing coffee product prices in Canadian grocery stores, focusing on historical prices, vendor differences, and seasonal trends. Using a Bayesian multiple linear regression model, the study reveals that historical prices have the most significant positive impact on current prices, while vendor differences and seasonal variations also play measurable roles. These findings highlight how past prices serve as a benchmark for current pricing strategies, emphasizing the importance of historical data in retail decision-making. The results contribute to understanding the dynamics of coffee pricing and provide insights for optimizing retail pricing strategies and future research directions.

Table of contents

*Code and data are available at: https://github.com/YiTang2/Canadian_Grocery_Analysis.git

1 Introduction

Consuming goods is probably the most common thing in everyone's life. Consumers are also most concerned about the price of goods and whether there is a price premium. What is the basis for pricing, the purchasing power of consumers? Or the cost of each intermediate goods, etc. This paper combines the product data of Canadian grocery stores with the raw data, and uses the current price as the outcome predictor to find the factors that may affect it. And I selected the most intuitive potential influencing factor - old price - for analysis. I also selected two vendors located in eastern and western Canada: Metro and SaveOnFoods, and different months to analyze possible seasonal trends.

In the analysis of coffee products pricing in Canadian grocery stores, the estimand is the effect of variables like historical prices (`old_price`), month of the year (`month`), and specific vendors (`vendor`), such as Metro and SaveOnFoods, on the current prices of coffee products. This paper estimates this effect using a Bayesian MLR model applied to collected dataset, which is to quantify how these factors influence the pricing of coffee in the market.

The finding of this paper is that the old price of coffee has a more significant positive impact on the current price among so many predictors. Based on this, my guess is that suppliers rely on past price data as a benchmark for setting new prices, which may to maintain the consistency in price levels expected by customers. Also due to the lack of data on coffee products in certain months, the analysis of seasonal monthly pattern effect cannot be established. This is very critical and important because it can reflect from a basic level the topic that this paper is most concerned about, which is what will affect the current price.

The paper is structured as following: Section 2 describes the data used for analysis, Section 3 describes how to set up, justify and validate the model, **?@sec-result** tells the finding of the data and model, Section 5 discusses the implication, potential problems, and future expectations.

2 Data

2.1 Overview

Price of each month's coffee of different vendors data is provided by(**citedata?**). This dataset records detailed sales about fast-moving consumer goods (FMCG) sold by various vendors, including volia, T&T, Loblaws, SaveOnFoods, Galleria, Metro, NoFrills and Walmart. It is also includes product-level details, such as the product name, current price, historical price (`old_price`), and the corresponding units and price per unit. The data also captures time-specific observations(2024-2-28 to 2024-6-22), with timestamps (nowtime) that can be used to analyze trends over days or months.

In order to simulate data, test simulated data, clean data, test cleaned data, exploratory data analysis and model data, we used R programming language (R Core Team 2023) to analyze the data and plot the graphs. Specific libraries that assisted the analysis include `tidyverse` (**citetidyverse?**), `palmerpenguins` (Horst, Hill, and Gorman 2020), `knitr` (**citeknitr?**), `arrow` (**citearrow?**), `ggplot2` (**citeggplot2?**), `dplyr` (**citedplyr?**), `here` (**citehere?**), `kableExtra` (**citekableExtra?**), `gridExtra`(**citegridExtra?**), `modelsummary`(**citemodelsummary?**), `rstanarm`(**citerstanarm?**).

The inspiration for my data processing came from my desire to study what factors would affect the current price of coffee products from two vendors in different regions of Canada, such as the current price of coffee products from two vendors, Metro and SaveOnFoods. The following variables are the data I selected after cleaning the data:

- `vendor`: The retailer selling the product in Canada.
- `old_price`: The historical price of the product, showing previous pricing or discounts.
- `product_name`: The specific product being sold, providing product-level insights.
- `current_price`: The price of the product at the time of observation.

New variable extracted and transformed from raw data:

- `month`: The month of data collection, extracted from `nowtime`.

Since the variable nowtime only records 4 months, it is considered a lack of Long-Term Trends, which means it's difficult to identify long-term pricing or demand patterns by using short data periods. So I only extracted a new variable—month from date of nowtime, which can simplify temporal analysis and identify trends, such as seasonal monthly pattern with price changes or demand patterns. It allows grouping data for monthly aggregation and supporting seasonality-focused insights or forecasting models.

To provide an preview of the coffee pricing with all potential factors that might affect it. Here, Table 1 simply reveals the variation between current price and old price in June for Metro's coffee products.

Table 1: Sample of Analysis Data Showing Products Sold by Both Vendors

vendor	product_name	current_price	old_price	month
Metro	Non-Dairy Vanilla Flavoured Latte Coffee Cream	7.49	8.99	6
Metro	Vanilla And Caramel Flavoured K-Cup® Coffee Capsules	9.99	12.99	6
Metro	Classic Black K-Cup® Coffee Capsules	9.99	12.99	6
Metro	Cold Brew Unsweetened Iced Coffee	7.49	7.99	6
Metro	Limited Edition Coffee Whitener, Coffee Mate	4.99	6.99	6
Metro	Italian Blend Dark Roast K-Cup Coffee Pods	6.49	6.99	6
Metro	Classic Roast Ground Coffee	8.99	12.49	6
Metro	Medium Roast Decafreinated K-Cup Coffee Pods, Pike P...	22.99	26.99	6
Metro	Medium Roast House Blend K-Cup Coffee Pods, Organic	6.49	6.99	6
Metro	Classic Decaf Ground Coffee	8.99	12.99	6

2.2 Measurement

The dataset from Hammer represents real-world retail activities, capturing product details, vendor listings, and price updates. When vendors update product information, such as pricing or availability, Hammer collects and structures this information into the dataset.

Key fields include vendor, product name, current price, old price, and nowtime. The data is collected through scraping and structured to enable analysis of retail trends, pricing strategies, and market dynamics over time. Each entry serves as a snapshot of a product's presence in the market at a specific time, allowing for focused analyses, like tracking price trends for specific products (e.g., coffee).

2.3 Data Visualization

Figure 1 shows a notable distribution with significant concentrations around 10 and 20 units. This suggests a pricing tier system where certain types of coffee products might be grouped by price points due to quality, brand, or other market factors. The dual peaks could indicate two main categories of coffee products, possibly differentiated by premium versus regular products.

According to Figure 2, the distribution is mainly focused between 5 and 20 units, pointing to a past market strategy where products were clustered around these prices. The consistency in this price range might reflect a stable market before any recent pricing adjustments influenced by external factors like supplier changes or inflation.

In Figure 3, the frequency of products available from Metro significantly surpasses that from SaveOnFoods, suggesting that Metro has a larger share of the market or a wider variety of

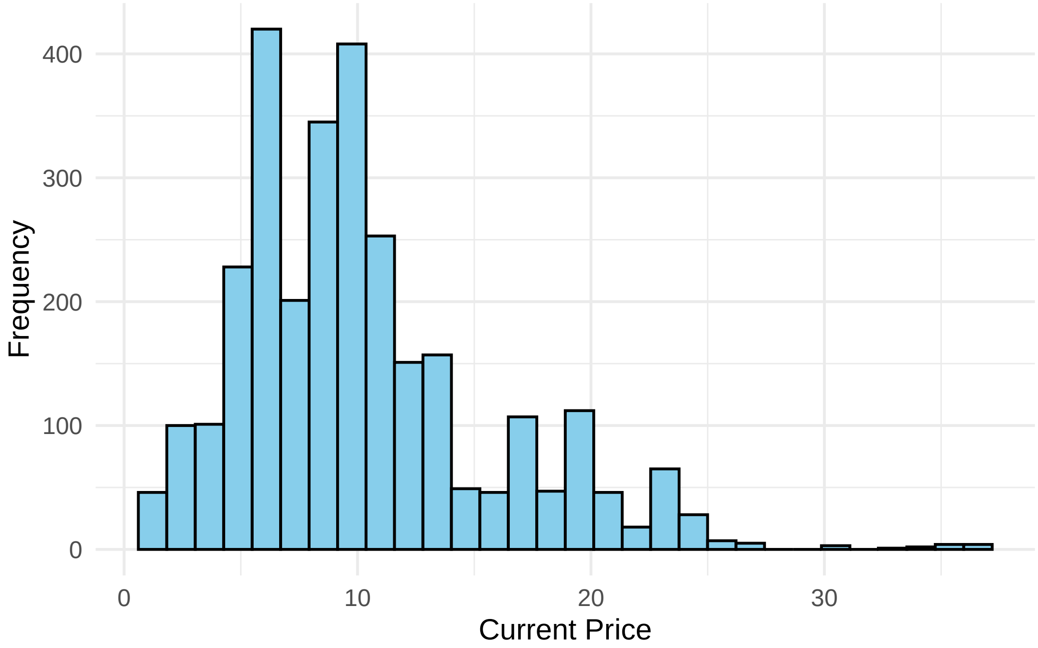


Figure 1: Histogram of Current Price

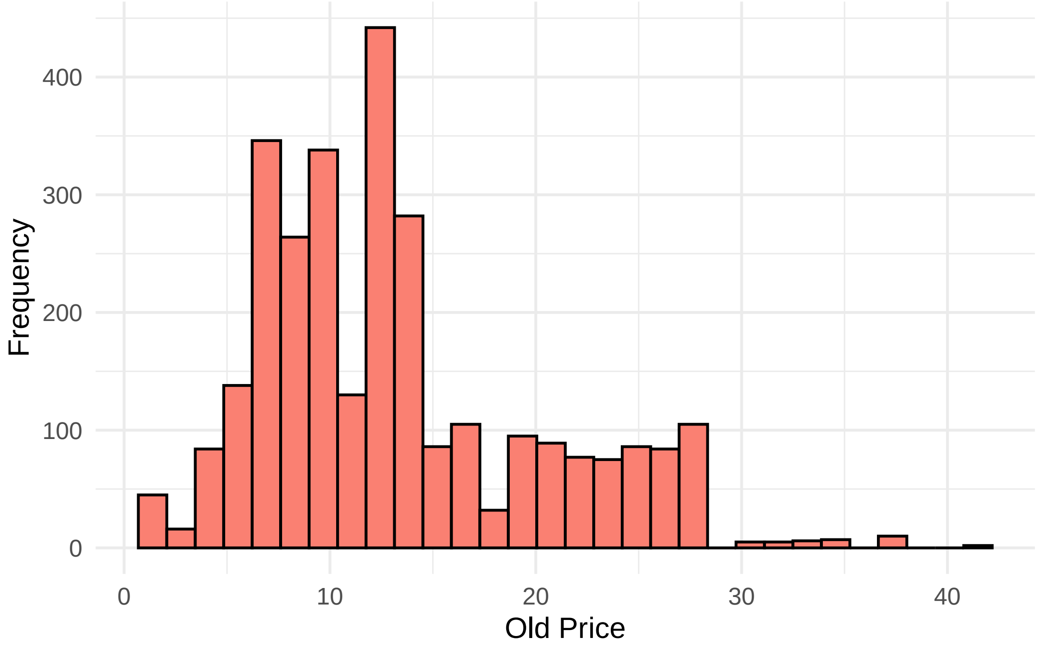


Figure 2: Histogram of Old Price

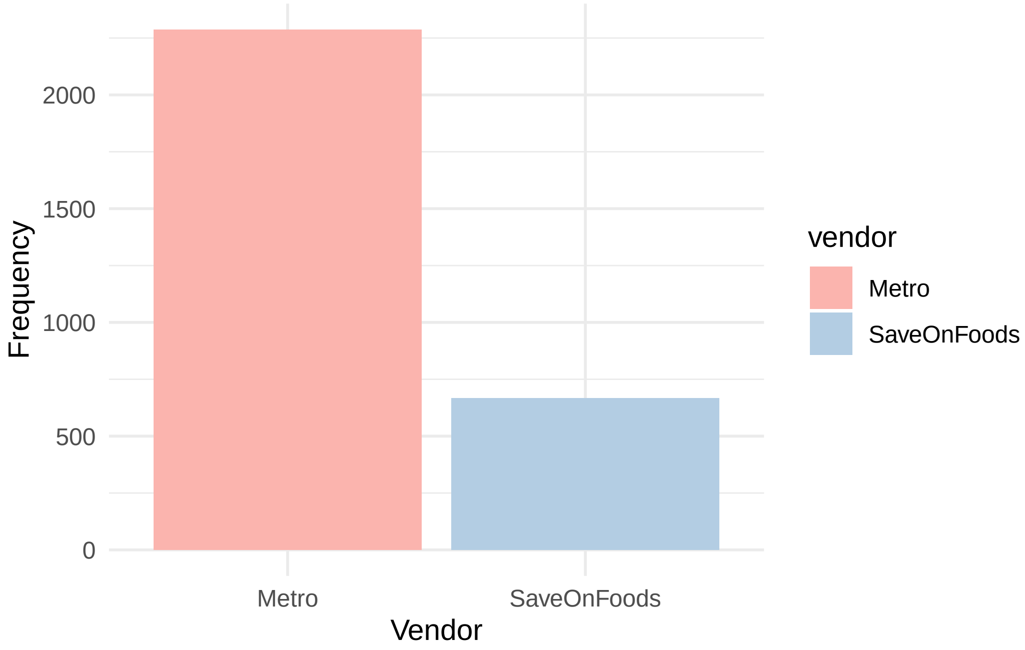


Figure 3: Bar Plot of Old Price

coffee products. This dominance in product offerings could provide Metro with a competitive edge in attracting a broader customer base.

The positive relationship shown in Figure 4 here indicates that current prices are influenced by their historical prices, maintaining a proportional increase or decrease. This trend suggests a pricing policy that adjusts prices based on previous benchmarks while taking into account factors like cost adjustments or market demand.

Figure 3 highlights how each vendor prices their products within the market. It shows that both vendors offer a wide range of prices, yet the spread and density of the data points may indicate Metro's pricing strategy targets both lower and upper market segments, whereas SaveOnFoods might be focusing on a specific niche.

3 Model

The goal of our Bayesian multiple linear regression is to investigate the factors that influence the current price of coffee in our dataset. Specifically, we try7 to understand how historical pricing, vendor differences, and seasonal monthly pattern affect current coffee prices.

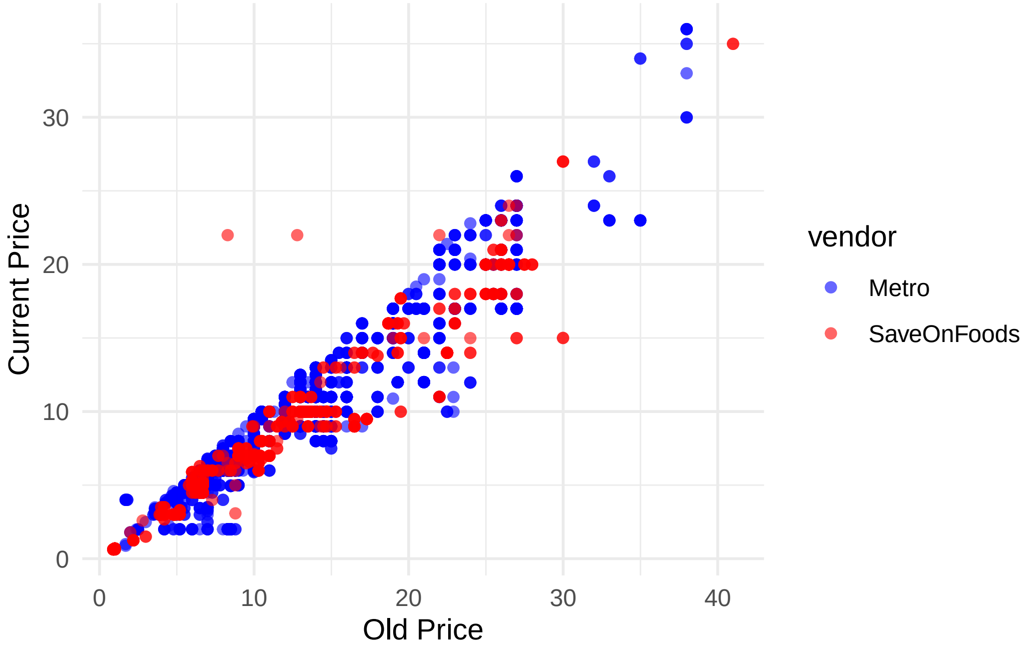


Figure 4: Scatterplot of Current Price vs Old Price

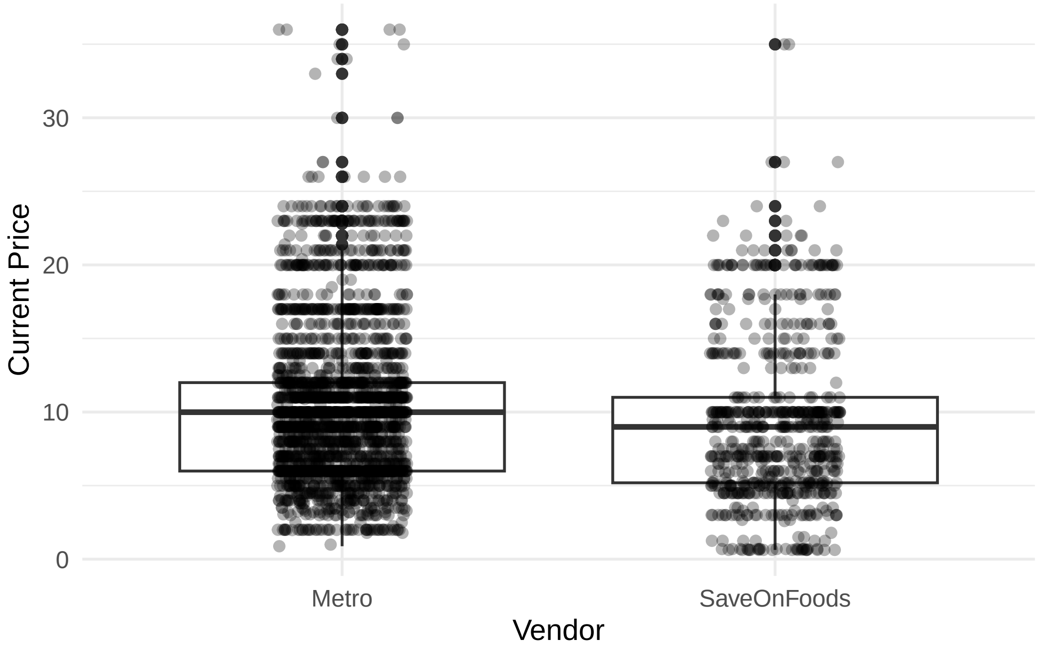


Figure 5: Boxplot of Current Price by Vendor

3.1 Model set-up

Define y_i as the current price of coffee for the i-th observation in the dataset. The predictors include:

- x_{1i}, the old price of the coffee,
- x_{2i}, dummy variable for the vendor, where:
  $x_{2i} = 1$: Vendor is "SaveOnFoods"; $x_{2i} = 0$: Vendor is "Metro",
- x_{3i}, the numeric month variable.

The model is formulated as follows:

$$\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma), && (1)\\
\mu_i &= \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i}, && (2)\\
\alpha &\sim \text{Normal}(0, 2.5), && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5), && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5), && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5), && (6)\\
\sigma &\sim \text{Exponential}(1). && (7)
\end{aligned}$$

This model describes the relationship between the current price of coffee (y_i) and three predictors: the old price of coffee (x_{1i}), a categorical vendor variable (x_{2i}) indicating whether the vendor is "Metro" or "SaveOnFoods," and a numeric variable for the month (x_{3i}). The response variable (y_i) is modeled as normally distributed with mean μ_i and standard deviation σ. The mean μ_i is defined as a linear combination of these predictors, with coefficients β_1, β_2, and β_3, and an intercept α. Prior distributions for the parameters are specified, including normal priors for α and the coefficients, and an exponential prior for σ. Intercept α represents the baseline mean current price for Metro if $x_{2i} = 1$; otherwise when $x_{2i} = 0$, it represents the mean current price for SaveOnFoods. Also, when old price and month is equal to 0, the intercept is not meaningful. Coefficient β_1 captures how changes in the old price affect the current price. Coefficient β_2 measures the difference in the mean coffee price between SaveOnFoods ($x_{2i} = 1$) and Metro ($x_{2i} = 0$). Coefficient β_3 reflects how the month influences current pricing, potentially capturing seasonal effects.

To implement this Bayesian model, we use the `rstanarm` package (**citerstanarm?**) in `R` (R Core Team 2023).

3.2 Model justification

The Bayesian Multiple Linear Regression (MLR) model is a suitable choice for analyzing the relationship between `current_price` (the dependent variable) and the predictors in the dataset. The dependent variable is continuous, and the Bayesian framework assumes a normal distribution for the response, which aligns well with the nature of coffee prices. This model captures the linear relationships between `old_price` (continuous), `vendor` (categorical, represented as a dummy variable), and `month` (numeric). These predictors are assumed to have additive effects on the response, which fits the linear regression framework. Logistic regression is used when the outcome variable is binary (e.g. 0 or 1). However, in our dataset, the dependent variable, `current_price`, is continuous. Since logistic regression cannot model continuous outcomes, it is unsuitable for this analysis. Also, poisson or negative binomial regression is typically applied when the response variable represents count data (e.g., the number of events occurring in a fixed period). `current_price` does not represent counts but rather continuous pricing data. Thus, these models do not align with the nature of the dependent variable.

3.3 Model validation

Figure 6 shows that our model accurately captures the central tendency of the data, though there are some deviations in the tail ends suggesting that the fit could be improved for extreme values. Additionally, the parameter estimation comparison chart highlights that most parameter estimates closely align with their priors, indicating a strong influence of prior settings on the estimates, especially under limited data. This is particularly evident with the vendorSaveOnFoods parameter, where its posterior distribution significantly diverges from others, hinting at potential anomalies in data sources or unique behaviors that warrant further investigation.

Figure 8 illustrate that our Bayesian model parameters are converging and demonstrating stability, essential for robust statistical inference. The slight oscillation of the intercept around -0.5 indicates minor variability. Meanwhile, the old price coefficient shows remarkable consistency at approximately 0.77, underlining dependable estimates. The month parameter's minor fluctuations suggest a subtle yet consistent temporal effect. The vendorSaveOnFoods coefficient consistently remains near -0.6, indicating a persistently negative price influence compared to Metro. Finally, sigma's stability around 1.75 ensures the model's error variability is well-accounted for. These observations collectively affirm that the model's parameters are effectively calibrated, offering a reliable foundation for understanding the influences on coffee prices.

In Figure 9 analysis of coffee product pricing in Canada, the use of the $\hat{R}$ values to assess model convergence reveals that all parameters have $\hat{R}$ values below 1.05, indicating excellent convergence of the model. This result validates the reliability of our model in estimating coffee prices and ensures the robustness of the analysis outcomes. It allows us to trust the model outputs, providing a solid foundation for further strategic decision-making and market analysis.

4 Results

Table 2 indicates a robust fit with a high R^2 of 0.900, suggesting a strong explanatory power of the model regarding the variance in coffee product prices. The model's parameters show that the old price of coffee (coefficient = 0.77) has a significant positive influence on the current price, suggesting that past pricing trends are good predictors of current pricing strategies. Vendor impact, specifically SaveOnFoods, shows a negative association with current price (coefficient = -0.61), indicating that coffee products from SaveOnFoods tend to be cheaper compared to Metro. Additionally, the month coefficient (0.06) implies a slight monthly variation in coffee pricing. The model's predictive performance is validated by low WAIC and LOOIC scores, and a small RMSE of 1.75, enhancing confidence in the reliability of its predictions.

Figure 7 llustrates the posterior distributions for the parameters in your Bayesian regression model, analyzing factors influencing coffee prices. The intercept shows a slight positive baseline effect. The old price of coffee strongly and positively affects the current price, indicating a direct relationship. The month variable shows a minimal and variable impact. SaveOnFoods, as a vendor, is associated with lower prices compared to Metro. The sigma parameter, indicating the model's error variance, shows a sharp and precise estimation, suggesting consistent variability in the data explained by the model.

5 Discussion

5.1 Overall findings and Implication

This research conducted a detailed examination of coffee product pricing across major Canadian grocery store chains, specifically Metro and SaveOnFoods. By utilizing a regression model, the study uncover how historical price data (by month) can predict current coffee prices. This approach not only helped identify pricing trends but also provided a systematic framework to analyze the economic factors influencing those trends.

The findings highlight a dynamic interaction between past and present pricing strategies, underscoring the influence of historical prices on current market behaviors. This analysis reveals that historical prices are strong predictors of current prices, suggesting that vendors likely use past data as a benchmark for future pricing decisions. This could imply a strategy focused on maintaining market stability and optimizing profit margins, which is crucial in a competitive retail environment.

5.2 Additional Insights from the Data Analysis

The coffee products prices are influenced by more than just historical trends; they also respond to a range of external economic factors. The study highlights how variations in coffee prices align with broader economic indicators like inflation rates and consumer demand. This insight is crucial for understanding the dynamic nature of the coffee market, suggesting that vendors adjust their pricing strategies not only based on past price behavior but also in reaction to economic conditions. This deeper understanding of price adjustments helps illustrate the broader economic landscape within which these businesses operate, emphasizing the strategic decisions vendors make in response to economic pressures. This analysis significantly enriches our understanding of market behaviors and the economic strategies that drive vendor actions in the competitive grocery sector.

5.3 Shortcomes and Future outlook

5.3.1 Limitation

This study integrates two distinct raw datasets to construct the analysis dataset, which presents unique challenges due to varying variable names and formats. The process involved carefully aligning these variables to ensure consistency and accuracy in the dataset used for modeling. This integration underscores the importance of meticulous data preparation in ensuring the reliability of the results obtained from statistical analyses.

A significant limitation of this study is the absence of geographic data for the vendors. Although our analysis provides insights into the pricing strategies of Metro and SaveOnFoods, we lack the data to explore how geographical location influences pricing strategies across Canada. Metro and SaveOnFoods operate in different regions of Canada, potentially the Eastern and Western parts, and it is plausible that regional economic conditions, competition, and consumer demographics could affect pricing strategies. The absence of this data restricts our ability to analyze and understand regional pricing dynamics comprehensively. lack of month: only have data of coffee prouduct pricing in June to November.

5.3.2 Weaknesses and next steps

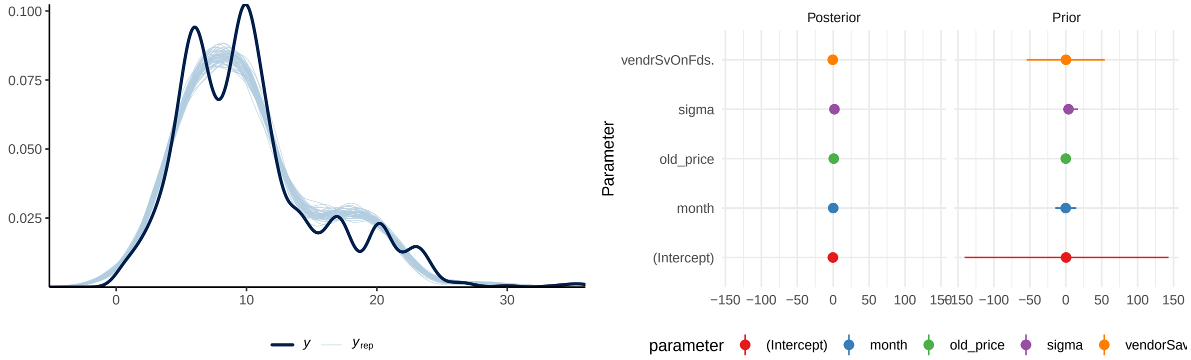
While the model provides valuable insights, there are limitations due to the dataset only covering a short period (June to November). This limits the ability to analyze long-term trends or seasonal impacts beyond this timeframe. Additionally, the model assumes linear relationships among variables, which may not fully capture more complex dynamics in price fluctuations.

Future research could expand the timeframe of data collection to include multiple years to better understand long-term trends and seasonal variations. Incorporating additional variables such as promotional activities, competitor prices, and economic indicators could also enrich the analysis. Further, exploring non-linear models or machine learning approaches may provide deeper insights into the pricing strategies of different vendors. To address these limitations, future research should focus on acquiring and integrating geographic data into the analysis.

This would allow for a more detailed examination of how location-specific factors influence pricing, enhancing our understanding of regional market behaviors. Exploring geographical influences on pricing can reveal targeted strategies that vendors might use to cater to local consumer preferences or to respond to regional competition. Incorporating such data would significantly enrich the analytical framework and potentially yield insights that could inform more localized or region-specific business strategies for these vendors.

Appendix

A Model details



(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 6: PPcheck & Posterior vs Prior

Table 2: Model summary of Coffee product pricing

	Coffee product pricing
(Intercept)	−0.41
old_price	0.77
month	0.06
vendorSaveOnFoods	−0.61
Num.Obs.	2954
R2	0.900
R2 Adj.	0.900
Log.Lik.	−5837.981
ELPD	−5843.3
ELPD s.e.	71.3
LOOIC	11 686.6
LOOIC s.e.	142.6
WAIC	11 686.6
RMSE	1.75

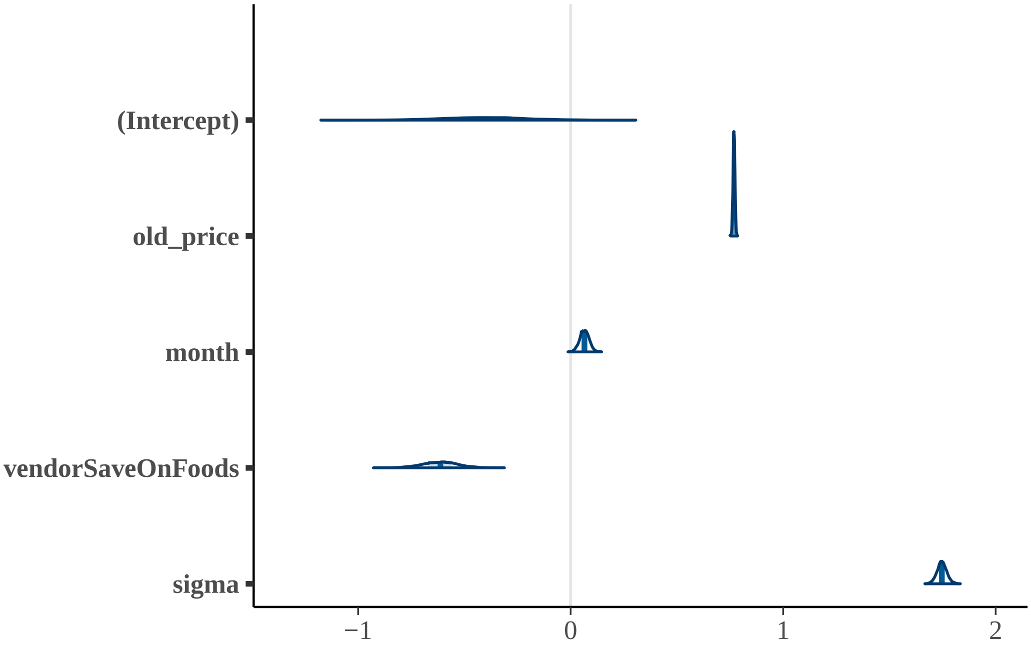


Figure 7: credibility interval

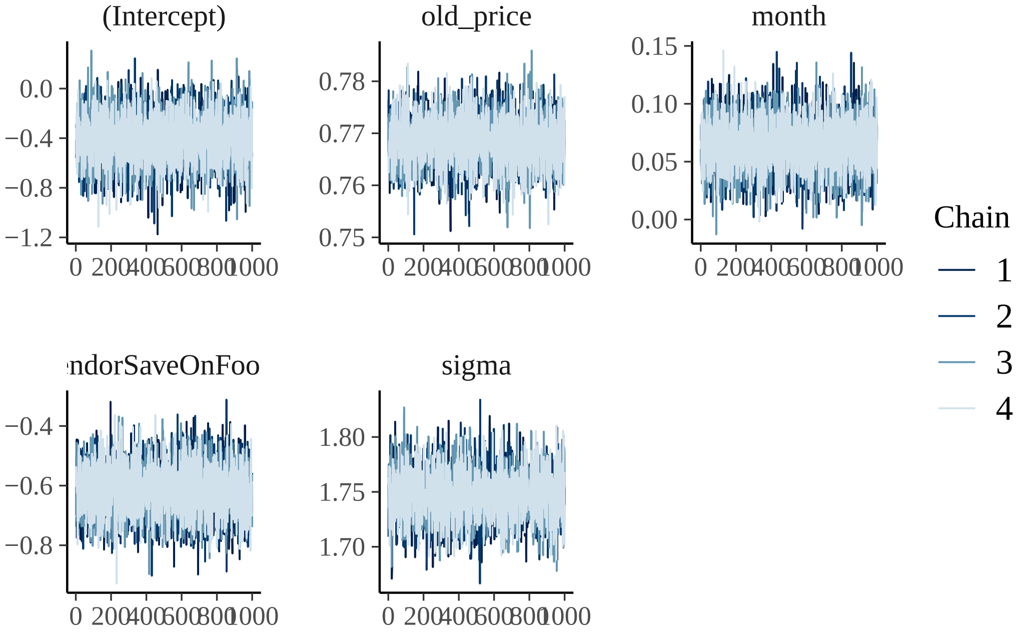


Figure 8: Trace plot

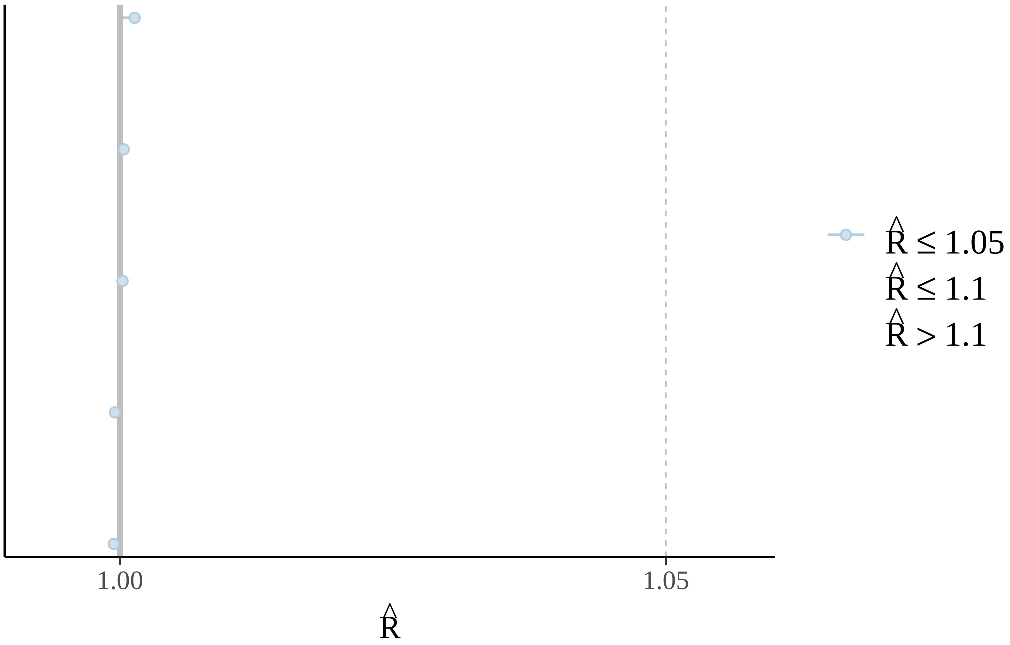


Figure 9: R-hat plot

References

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

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