Analyzing changes in Canadian grocery prices*

Abdullah Motasim Elizabeth Luong

November 14, 2024

This paper investigates grocery pricing trends across Canada using SQL analysis on the 'Project Hammer' dataset, which contains detailed vendor-specific price data. Focusing on statistical relationships and potential pricing biases, we explore factors driving price variations across different regions and vendors. A Bayesian regression model is applied to understand correlations, accounting for limitations and highlighting areas where causation cannot be assumed. Findings reveal competitive dynamics within Canadian grocery markets, implications for consumers, and recommendations for future research into market and pricing behavior across vendors.

1 Introduction

The rising cost of groceries in Canada has drawn significant public attention, impacting consumer budgets and prompting regulatory interest. While inflation and supply disruptions are often cited as contributors, the specifics of how major Canadian grocery vendors set and adjust prices remain underexplored. This study addresses this gap by analyzing data from Project Hammer, a data set that tracks grocery prices across vendors including Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. Through examining these data, we aim to uncover factors driving grocery price dynamics and assess the degree to which vendor-specific pricing may reflect competitive or non-competitive practices.

Using a Bayesian regression model, our analysis identifies prior prices, vendor identity, and recent price changes as significant determinants of current grocery prices. The consistency in historical pricing indicates relative stability across product categories, while unique vendor patterns suggest differing market strategies, likely shaped by operational costs and customer demographics. Our findings have practical implications for both consumers and policymakers:

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

consumers can use vendor-specific insights to make informed purchasing decisions, while policymakers can leverage this analysis to support regulatory strategies that promote competition and price transparency.

The paper is structured as follows: Section 2 discusses the data types included in the raw data, the cleaning process for the data, and the reason for selecting the data set we did. Section 3 discusses model specification and justification for utilizing a Bayesian linear regression model. Section 4 presents the trends and correlations between different variables utilizing tabular and graphical means. **?@sec-discussion** discusses the results of Section 4 going into detail on what the simulation results can tell us about grocery prices in Canada as well as discussing missing data and sources of bias.

2 Data

2.1 Overview

For this study, we utilize data from Project Hammer, a Canadian initiative designed to monitor grocery prices across major retailers in Canada. The Project Hammer dataset was collected from eight prominent Canadian grocery vendors—Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods—between February 28, 2024, and the latest available data load in the database [citeProjectHammerData]. This dataset enables us to investigate price fluctuations, identify pricing trends, and explore potential competitive or collusive patterns within the Canadian grocery sector. Project Hammer aims to support regulatory efforts by providing data on price transparency, making it suitable for academic research and potential legal analysis. Following data processing and cleaning procedures outlined in Section 3.2, the dataset was structured to allow an analysis of historical price changes and vendor-specific price patterns. The dataset contains 1,996,969 rows and 5 columns, with variables capturing product (product name), vendor_name (name of the grocery retailer), price_current (current product price at the time of recording), price_old (previous recorded price for the product), and price_difference (difference between the current and old prices). Price data is recorded in Canadian dollars and captures a broad range of grocery items from various categories, including fresh produce, dairy, pantry staples, and household items. To ensure the quality and consistency of the data, we focused on removing missing values and imputing outlier prices when necessary. For this analysis, we excluded entries with extreme price fluctuations beyond three standard deviations, as these could represent temporary discounts or data recording anomalies. This cleaning process provides a balanced and accurate reflection of grocery pricing, allowing for robust statistical analysis on vendor-specific trends and average price comparisons across product types. By narrowing the scope to recent data, Project Hammer's dataset captures real-time fluctuations in grocery pricing, providing timely insights into potential market dynamics and consumer price impacts. The analysis allows for future investigation of temporal price trends, vendor-based price differences, and, more broadly, consumer affordability in a market facing rising cost-of-living pressures.

2.2 Measurement

Each entry in the data set represents a captured moment in time where prices, products, and vendor information were recorded. The measurement process involved regular data collection from the online pricing systems of the following prominent grocery chains: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. Each entry captures variables such as: product name, vendor name, and both current and previous prices. The Project Hammer initiative leveraged web scraping technology to ensure a consistent and standardized entry format for all retailers, documenting prices at regular intervals. This approach ensures that price points reflect real-time data rather than historical estimates or annual averages, providing a close approximation of consumer experiences. The data set records the price of each grocery item in Canadian dollars, including items from broad categories such as fresh produce, dairy, pantry goods, and household essentials, to present a holistic view of grocery costs.

Essentially, the real world phenomenon we observed was the website of each vendor with a listing of their products for that week, we turned this phenomenon into an entry within out data set with the use of a screen-scrape of the website UI. This means a HTTP request to load the page was sent to the desired vendors website on a specific day, than the returned HTML was parsed to extract specific content such as text, images, prices, etc. These captured features were than utilized to fill out the corresponding columns within the data set.

3 Model

Our modeling approach seeks to explore and quantify the relationship between previous grocery prices, vendor identities, and price differences in the current prices observed within Canadian grocery stores. This analysis employs a Bayesian linear regression model implemented via the `stan _glm` function in the `rstanarm` package to examine how factors such as historical prices, vendor differences, and observed price changes impact current prices for various grocery items.

In this model, `price_current` serves as the response variable, while `price_old`, `vendor_name`, and `price_difference` act as predictor variables. The linear regression model assumes a Gaussian distribution for the response variable `price_current`, allowing for a straightforward interpretation of the estimated parameters.

3.1 Model set-up

The model includes the following predictor variables:

- Previous Price (`price_old`): The price of the product in a previous time period.

- Vendor (`vendor_name`): A categorical variable representing the grocery store chain selling the product, capturing vendor-specific pricing differences.
- Price Difference (`price_difference`): The difference between the current and previous price, which may indicate market or vendor-specific pricing adjustments.

The model can be represented mathematically as follows:

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{Previous Price}_i + \beta_2 \cdot \text{Vendor}_i + \beta_3 \cdot \text{Price Difference}_i \\
\epsilon_i &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

Where:

- β_0 is the intercept term, representing the baseline estimate of $price_{current}$
- β_1, β_2, and β_3 are the coefficients representing the effects of $price_{old}$, $vendor_{name}$ and $price_{difference}$ on $price_{current}$.
- σ^2represents the variance of the error term, capturing unexplained variability in current prices.

The model is executed in `R` (R Core Team 2023) using the `rstanarm` package (Goodrich et al. 2022), with priors set to regularize the estimates and prevent overfitting. Specifically, we use a normal prior for the coefficients, centered at zero with moderate variance, to ensure stable estimates without overly restrictive assumptions.

3.1.1 Model justification

The selection of a linear regression model for this analysis is based on the continuous nature of `price_current`, which allows us to quantify how previous pricing, vendor, and price changes influence current pricing. Existing economic theories suggest that previous prices and historical pricing patterns are often predictive of current pricing, especially in competitive retail markets where price sensitivity and vendor-specific strategies play significant roles. The inclusion of `price_old` as a predictor aligns with time-series economic models where past data points inform future values.

Additionally, vendor differences, represented by `vendor_name`, capture potential competitive dynamics in the Canadian grocery market. Vendors may adopt distinct pricing strategies or respond differently to market conditions, and the inclusion of this categorical variable enables an analysis of these patterns across various major grocery chains. The `price_difference`

variable, which represents recent price changes, provides an understanding of market adjustments and reflects fluctuations potentially influenced by external factors, such as supply chain constraints or inflation.

The Bayesian approach was selected for its flexibility in incorporating prior information and its robustness in handling uncertainty within small or moderate sample sizes. Furthermore, Bayesian methods allow us to capture the posterior distributions of the parameters, which can be used to evaluate the strength and credibility of each predictor's effect on current prices. This approach also aligns with the objectives of Project Hammer by facilitating a transparent and probabilistically rigorous analysis of grocery pricing trends.

In summary, this model helps explain the primary factors of current grocery prices in Canadian stores, accounting for both historical prices and vendor effects. By utilizing Bayesian framework, this analysis provides a clear quantification of the influence of each predictor while allowing for robust estimation and interpretability in the context of economic and competitive pricing strategies.

4 Results

The results of our analysis of the Project Hammer dataset explain important findings about pricing trends across major Canadian grocery vendors. By examining factors such as prior pricing (`price_old`), vendor identities (`vendor_name`), and recent price changes (`price_difference`), we assess how these variables affect current prices (`price_current`). The following sections summarize key findings, supported by visualizations that highlight trends and relationships within the data.

4.1 Vendor-Specific Price Trends

We begin by examining average current prices (`price_current`) across different grocery vendors to identify potential pricing variations in the Canadian grocery market.

4.2 Price Difference Analysis

We further analyze `price_difference`, which captures the change between the current price (`price_current`) and the previous price (`price_old`). This variable provides insights into recent price adjustments across vendors.

4.3 Relationship Between Previous and Current Prices

We assess the relationship between previous prices (`price_old`) and current prices (`price_current`) to understand price consistency and potential price elasticity across products.

4.4 Distribution of Price Differences

To further explore pricing dynamics, we examine the distribution of `price_difference`, capturing both price increases and decreases across the dataset.

4.5 Model Evaluation

To evaluate our model's performance, we assess the predictive accuracy of the Bayesian linear regression model on the test data. The model's Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) provide insights into the model's ability to predict `price_current` based on `price_old`, `vendor_name`, and `price_difference`.

Table 1: Model Evaluation Metrics. The table displays the MAE and RMSE for the model, reflecting the average and squared prediction errors, respectively. Lower values indicate better predictive accuracy, suggesting that the model effectively captures the relationship between previous prices, vendor identities, and price differences.

These results demonstrate that vendor identity and previous prices are significant predictors of current prices in the Canadian grocery sector, with implications for understanding competitive strategies and consumer price sensitivity in the market.

5 Discussion {sec-discussion}

5.1 Interpretation of Grocery Pricing Dynamics

This study investigates the primary factors influencing current grocery prices in Canada by examining the Project Hammer dataset. Our Bayesian regression model reveals that prior prices (`price_old`), vendor identity (`vendor_name`), and recent price differences (`price_difference`) play significant roles in determining current prices. The influence of previous prices indicates price continuity across time, suggesting that grocery items tend to exhibit relatively stable pricing patterns with minor adjustments. Vendor-specific effects highlight the role of retailer pricing strategies, with some vendors maintaining consistently higher or lower prices. These vendor variations may reflect differences in market positioning, operational costs, or competitive strategies, impacting price consistency across the grocery sector.

Additionally, the presence of both positive and negative values in the `price_difference` variable illustrates the prevalence of both price increases and decreases, underscoring the dynamic nature of grocery pricing in response to external market forces. External factors, such as supply chain disruptions, inflation, and seasonal demand fluctuations, likely contribute to the observed variability. This model highlights the complexity of pricing dynamics and the importance of historical pricing and vendor-specific factors in shaping current grocery prices across major Canadian retailers.

5.2 Implications for Competition and Consumer Costs

The findings from Project Hammer have broader implications for both competition in the grocery sector and consumer expenses. Vendor-specific price differences suggest that some retailers adopt competitive pricing strategies to attract cost-sensitive consumers, while others may leverage brand loyalty or perceived quality to justify higher prices. In this context, price-sensitive consumers may benefit from exploring pricing variations across vendors to find the best deals, while retailers may face pressure to adjust their prices to remain competitive.

The stability of historical prices as a predictor of current prices also highlights potential concerns about price rigidity in certain product categories. This rigidity may limit the effectiveness of competition, as vendors might rely on historical price baselines rather than actively adjusting prices in response to market demand. Policymakers interested in promoting competition in the grocery sector could consider encouraging more frequent pricing updates or transparency in pricing strategies, particularly for staple items that heavily impact household budgets.

5.3 Limitations of Data and Model

This analysis has several limitations that may impact the accuracy and comprehensiveness of our findings. First, the dataset is limited to major grocery chains, which may not capture price variations across smaller or regional stores that could offer different pricing structures. Additionally, the lack of granular product-level data limits our ability to analyze category-specific trends (e.g., dairy versus produce) or regional differences within the same vendor.

The model's focus on historical prices, vendor identity, and price differences also excludes potential external factors that may influence pricing, such as seasonal demand, promotions, or supply chain constraints. Including additional variables, such as the frequency of price adjustments or regional economic indicators, could enhance the model's predictive power and provide a more comprehensive view of grocery pricing dynamics.

5.4 Future Research Direction

Future research could build on this work by incorporating a wider range of variables, including product-specific attributes, seasonal factors, and promotional data, to capture a broader

spectrum of influences on grocery prices. Exploring how demographic factors, such as income levels in different regions, correlate with price patterns across vendors could offer insights into the socio-economic impacts of grocery pricing on different communities.

Increasing the model's granularity to analyze price dynamics at the regional or product-category level would provide a deeper understanding of how specific items are priced within the same store or across regions. Additionally, expanding the analysis to include other sectors, such as online grocery prices or specialty stores, could provide a more holistic view of the Canadian grocery market.

Incorporating machine learning techniques could also improve the model's adaptability to rapidly changing market conditions, allowing for more nuanced forecasts of pricing trends. This approach could be particularly useful in the face of ongoing challenges such as inflation and supply chain disruptions, offering a valuable tool for policymakers and consumers interested in understanding and mitigating grocery costs in Canada.

Appendix

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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