Analyzing Changes in Canadian Grocery Prices*

Abdullah Motasim Elizabeth Luong

November 21, 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 dataset 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/Luongel1/project_hammer.

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 dataset 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. Section 5 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 can be found on the project hammer website (Filipp 2024). This dataset enables us to investigate price fluctuations, identify pricing trends, and explore potential competitive or collusive patterns within the Canadian grocery sector.

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.

Data cleaning involved removing missing values and excluding extreme price fluctuations (beyond three standard deviations) to ensure quality and consistency. This process provides a reliable basis for statistical analysis of vendor-specific trends and price comparisons.

By focusing on recent data, the dataset captures real-time grocery price dynamics, offering insights into market trends, vendor-based differences, and consumer affordability amid rising living costs.

2.2 Measurement

Each dataset entry represents a snapshot of prices, products, and vendor details collected through regular web scraping from major grocery chains, including Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. Key variables include product name, vendor, current price, and previous price, all recorded in Canadian dollars.

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 dataset 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 dataset 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 dataset.

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

A linear regression model was chosen for this analysis due to the continuous nature of price_current, enabling us to quantify how previous pricing, vendor, and price changes affect current prices. Economic theories suggest that past prices and pricing patterns are predictive of current prices, particularly in competitive retail markets. Including price_old as a predictor aligns with time-series models, where past data informs future values.

The vendor_name variable captures competitive dynamics, as different vendors may adopt distinct pricing strategies. Price_difference reflects recent price changes, offering insight into market adjustments influenced by factors like supply chain issues or inflation.

A Bayesian approach was selected for its flexibility, handling uncertainty and small sample sizes while incorporating prior information. This method allows us to evaluate the strength of each predictor's impact on current prices and aligns with Project Hammer's goal of providing a transparent, probabilistically rigorous analysis of grocery pricing trends.

In summary, this model explains the key factors influencing grocery prices in Canada, using Bayesian analysis to quantify the impact of historical prices and vendor-specific 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.

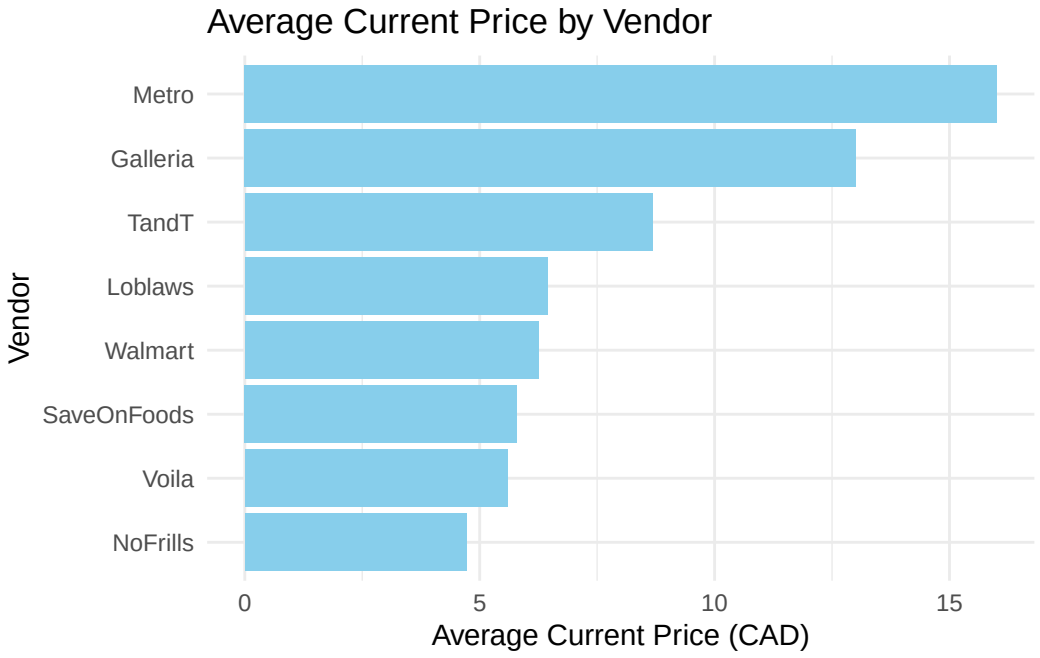

Figure 1: This bar plot shows the average price of products sold by each grocery vendor in Canadian dollars (CAD). Variations in pricing suggest differences in market positioning or pricing strategies among vendors, with some retailers consistently offering higher or lower prices across product categories.

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.

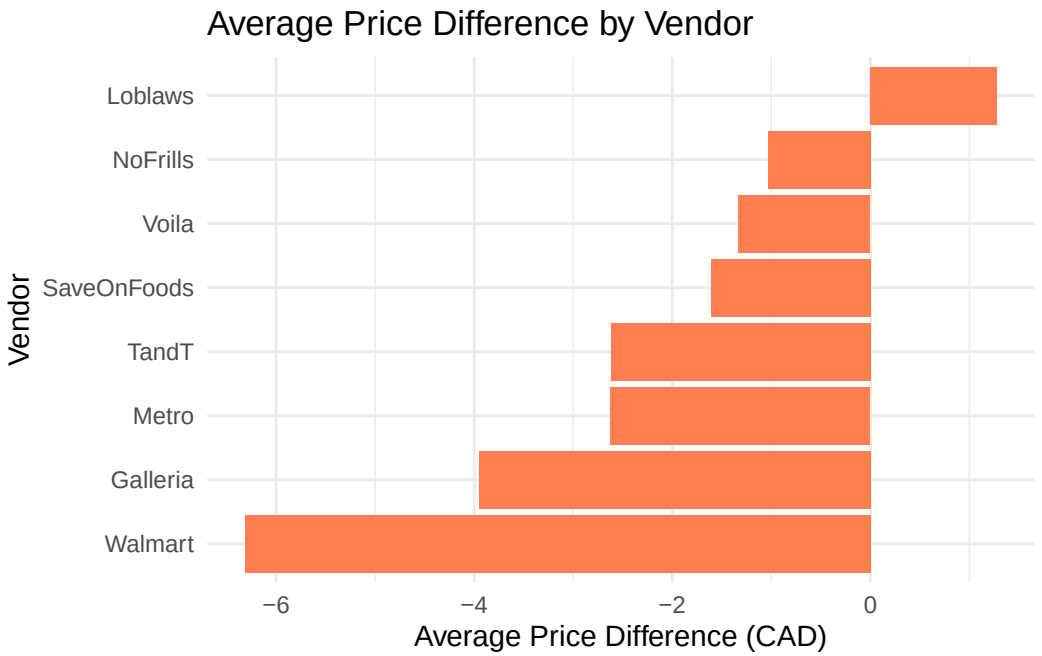

Figure 2: This bar plot shows the average price change per vendor, highlighting retailers with the most significant pricing adjustments. Positive differences suggest recent price increases, while negative differences indicate reductions, reflecting possible discounting or competitive pricing strategies.

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.

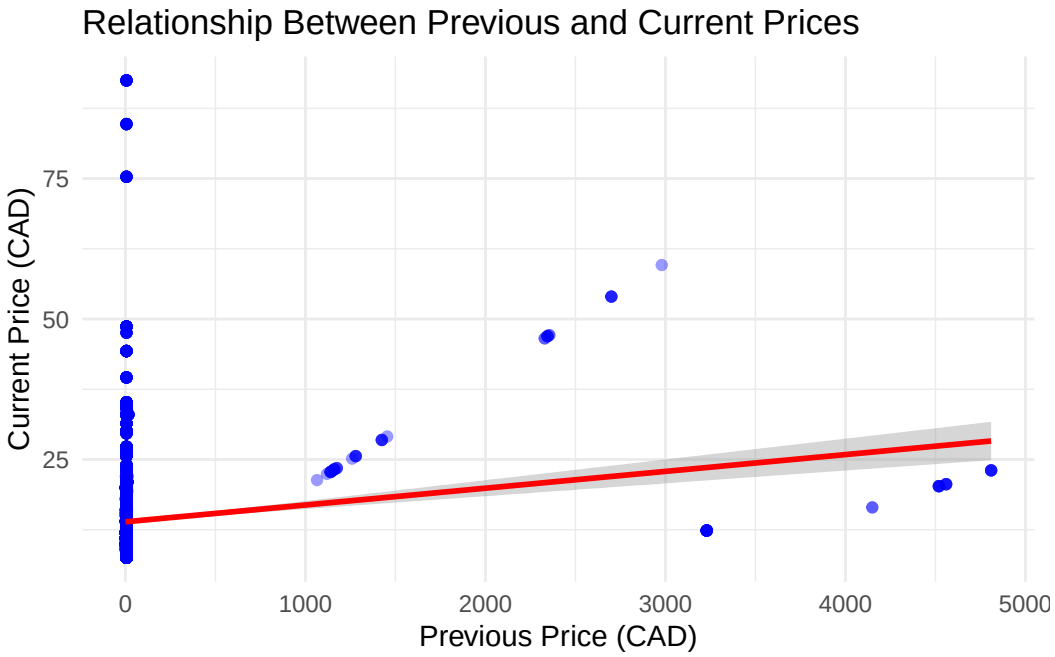

Figure 3: This scatter plot with a linear regression line shows the correlation between previous and current prices. A positive relationship would indicate that higher previous prices predict higher current prices, implying potential price stability or gradual price changes in the grocery market.

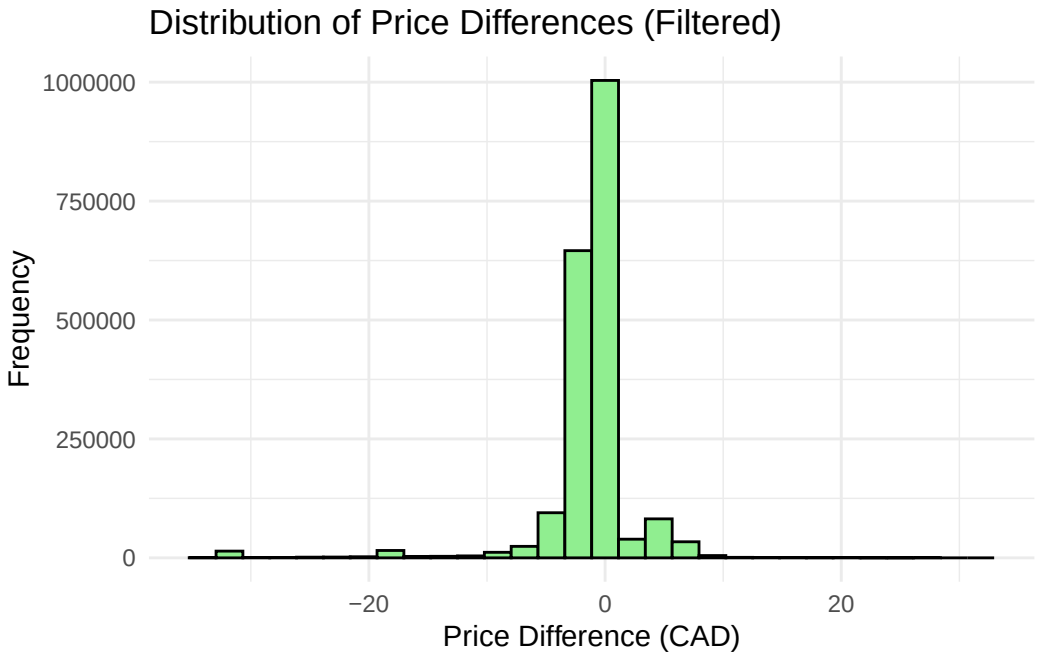

Figure 4: This scatter plot with a linear regression line shows the correlation between previous and current prices. A positive relationship would indicate that higher previous prices predict higher current prices, implying potential price stability or gradual price changes in the grocery market.

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`.

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

5.1 Correlation vs. Causation

This analysis utilizes Bayesian regression model to identify factors correlated with current grocery prices, including prior prices (`price_old`), vendor identity (`vendor_name`), and recent price differences (`price_difference`). While the model shows that historical prices and vendor-specific effects are strongly associated with current pricing, it's essential to interpret these relationships cautiously in terms of causation.

The correlation between previous and current prices likely reflects price continuity rather than a causal mechanism. For example, stable pricing over time may arise from vendor policies or industry norms rather than from any inherent property of the product or market structure. Similarly, vendor identity correlates with price differences, but this does not imply that being a particular vendor causes higher or lower prices. Instead, vendor-specific prices likely reflect underlying operational costs, competitive positioning, or branding strategies. Caution is warranted in inferring causation, as unobserved factors—such as cost structures or consumer loyalty—may influence both vendor identity and pricing patterns.

The presence of both price increases and decreases (reflected in `price_difference`) further highlights the dynamic nature of grocery pricing. While these changes are correlated with the final prices observed, it would be speculative to conclude that specific price fluctuations directly cause current price levels. External factors like inflation, seasonal demand, and supply chain disruptions may drive these observed correlations, underscoring the need for a cautious interpretation of causality.

5.2 Missing Data

The Project Hammer dataset contains some limitations in data coverage, which may impact the completeness and accuracy of our findings. Specifically, the data does not include smaller or regional grocery stores, which may have distinct pricing strategies and could offer lower or

more competitive prices in certain areas. The absence of these smaller players may lead to an incomplete picture of the overall grocery market, as price variations across different types of stores are not captured.

Additionally, missing data on certain product categories or regions within vendors limits the granularity of the analysis. For example, without category-specific data (e.g., dairy, produce, or packaged goods), it's challenging to examine whether price dynamics vary significantly across product types. This lack of detail may obscure category-level trends, which could be particularly relevant for staple items that heavily influence household spending. Future data collection efforts could focus on capturing a wider range of store types and product details to provide a more representative dataset.

Finally, missing information on factors like seasonal influences, promotions, or regional economic conditions means that some relevant variables are absent from the model. These factors could help explain variations in grocery pricing, particularly for products sensitive to external factors (e.g., fresh produce affected by seasonality). Addressing these data gaps could significantly improve the model's ability to capture the full range of influences on grocery prices.

5.3 Sources of Bias

The Project Hammer dataset and the corresponding analysis may contain several sources of bias that could affect the interpretation of results. One potential source of bias stems from the focus on major grocery chains, which may not accurately represent pricing dynamics across smaller, independent stores. Since major vendors typically have broader operational reach and standardized pricing strategies, the findings may be biased toward the practices of these larger corporations, potentially overestimating price stability or ignoring competitive pricing seen in smaller markets.

Another possible bias comes from the data's retrospective nature, focusing on historical pricing as a predictor of current prices. This approach may inherently favor price continuity and overlook the role of short-term market disruptions, promotions, or external shocks. For instance, a focus on historical prices could understate the impact of sudden events (like supply chain breakdowns or economic crises) that may cause sharp price changes. This form of selection bias could lead to an overly conservative estimate of price volatility in the model.

Furthermore, vendor-specific effects could introduce bias if certain vendors follow distinct pricing philosophies, such as premium or discount positioning. If these vendors have unique market positions that attract particular types of consumers, the model may reflect not only price dynamics but also underlying consumer segmentation. This brand-driven bias could distort interpretations, particularly if certain vendors consistently maintain higher or lower prices due to factors outside competitive dynamics (e.g., brand loyalty or perceived product quality).

References

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

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/.