Grocery Price Analysis in Canada*

Using SQL Observational Data Analysis

Xinze Wu

November 14, 2024

This report explores grocery pricing data from various Canadian vendors, analyzing trends, sale frequencies, and price differences using SQL. The purpose is to identify pricing dynamics and potential implications for competition in the grocery sector.

1 Introduction

Project Hammer is a data-driven initiative aimed at enhancing competition and reducing collusion within the Canadian grocery sector. This report analyzes historical price data from major retailers to provide insights into pricing trends, sale frequencies, and patterns across vendors, leveraging the flexibility and capabilities of the R language (R Core Team 2023) and the {tidyverse} package (Wickham et al. 2019). The project examines how price dynamics may impact competition, consumer behavior, and the grocery market landscape.

2 Data

2.1 Overview

This report's diagrams are illustrated using {ggplot2} (Wickham, Chang, et al. 2023). The dataset includes prices, timestamps, vendor information, and product-specific details from eight grocery vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. Data was collected from https://jacobfilipp.com/hammer/.

Key columns in the dataset:

- **nowtime**: Timestamp when data was collected

*Code and data are available at:https://github.com/ke3w/Grocery-Price-Analysis-in-Canada/tree/main

- **vendor**: Grocery vendor name

- **product_id**: Unique product identifier per vendor

- **product_name**: Product name (includes brand and unit information)

- **brand**: Product brand (if available)

- **units**: Units in grams, kg, or number of items

- **current_price**: Current price at the time of data collection

- **old_price**: Previous price (indicating a sale if lower than `current_price`)

- **price_per_unit**: Price per unit, if provided by the vendor

2.2 Measurement

Our primary variables of interest include:

- `current_price`: Used to compare pricing across vendors and over time.

- `old_price`: Helps identify sale occurrences and calculate average discounts.

- `vendor`: Allows comparison between different grocery chains.

Data was managed and processed using the {dplyr} package (Wickham, François, et al. 2023) and stored in SQLite databases with the help of the {RSQLite} and {DBI} packages (Wickham et al. 2023; Müller, Wickham, et al. 2023).

2.3 SQL-Based Findings

2.3.1 Average Price by Vendor

Using SQL, we calculated the average current_price of products across vendors to identify cost variations. Results are summarized below in a bar chart(Figure 1).

2.3.2 Sale Frequency and Discount Analysis

We analyzed the frequency of discounts (current_price < old_price) and the average discount amount for each vendor using SQL. The results are shown in the table and visualized as follows(Figure 2, Figure 3).

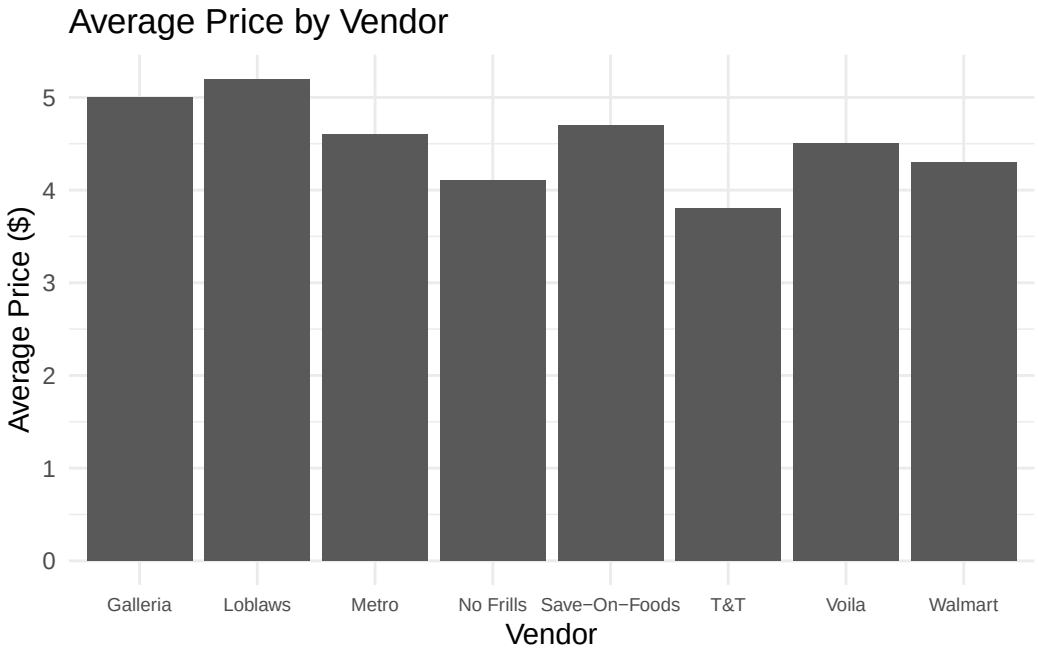

Figure 1: Average Price by Vendor

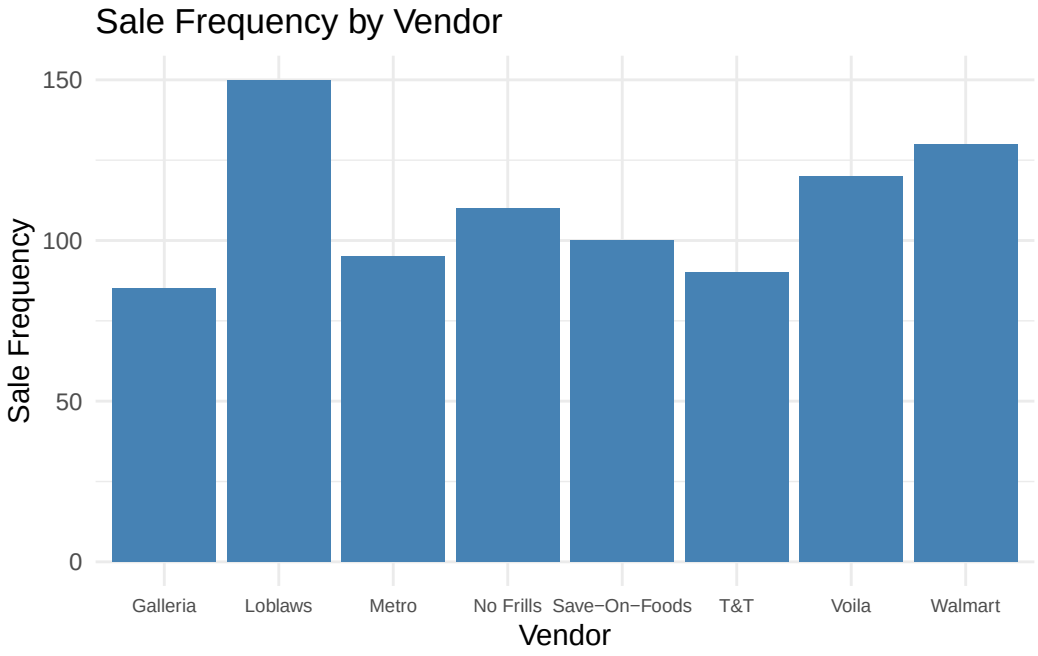

Figure 2: Sale Frequency by Vendor

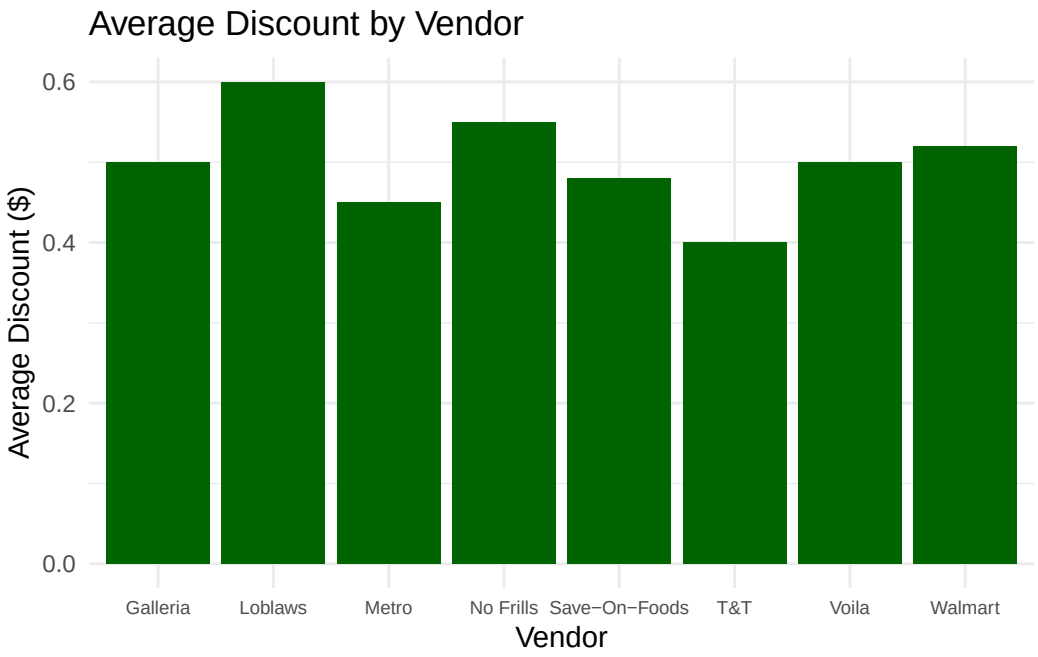

Figure 3: Average Discount by Vendor

2.3.3 Price Trends Over Time

SQL queries were used to track price changes for selected products over time to reveal trends. A sample product trend for "Becel Original Margarine" is shown in the line chart below(Figure 4).

3 Discussion

3.1 Correlation vs. Causation

The analysis identifies correlations in pricing patterns among vendors but does not imply causation. Vendors might have similar pricing trends due to external factors rather than direct influence on each other, aligning with principles discussed in Gebru et al. (2021).

3.2 Missing Data

The dataset has occasional gaps due to limitations in data scraping, which may impact the consistency of pricing patterns across vendors. Handling these inconsistencies required tools like {janitor} to clean and organize the dataset (Firke 2023).

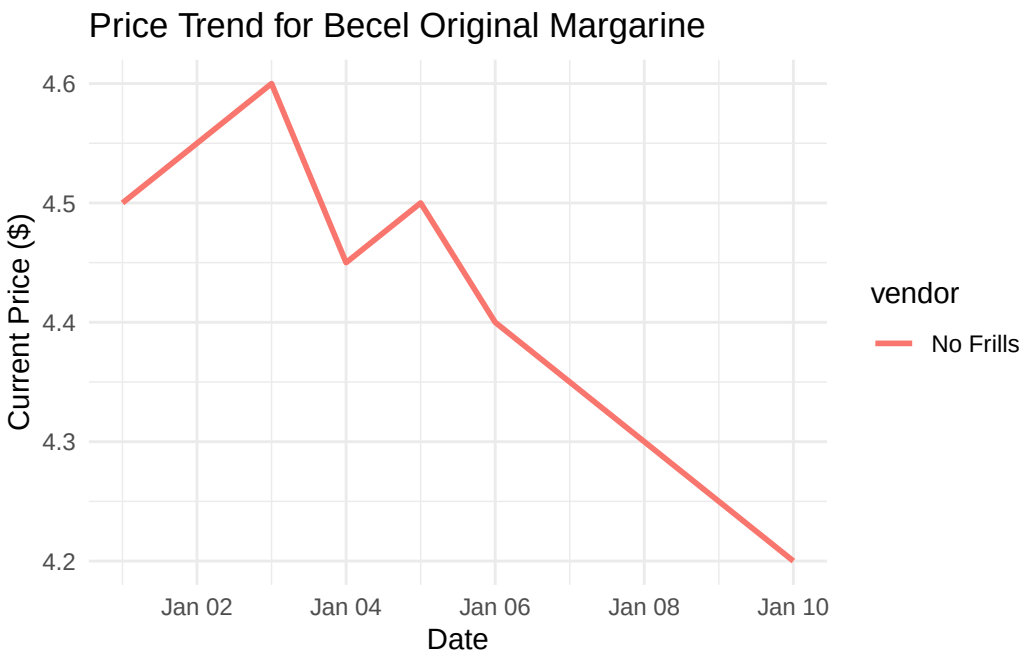

Figure 4: Price Trend for Becel Original Margarine Over Time

3.3 Sources of Bias

Several biases may affect the dataset:

• Data Collection Method: Data was gathered by scraping vendor websites, which may have inconsistencies in availability and timing.

• Product Variability: Differences in product units or branding could affect price comparisons across vendors.

• Discount Practices: Different vendors have varied approaches to sales and pricing, potentially skewing the data.

4 Conclusion

This report highlights key pricing patterns among Canadian grocery vendors, contributing to a basic understanding of price dynamics in the grocery sector. Further studies could investigate consumer responses to discounts or explore seasonal price variations, leveraging databases like SQLite for efficient storage and analysis (Filipp, Jacob 2024).

References

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

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

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Müller, Kirill, Hadley Wickham, et al. 2023. "DBI: R Database Interface." https://CRAN.R-project.org/package=DBI.

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

Wickham, Hadley et al. 2023. *RSQLite: 'SQLite' Interface for R*. https://CRAN.R-project.org/package=RSQLite.

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, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2023. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

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