Analysis of Canadian Grocery Price from February to November 2024*

Tracking Price Fluctuations and Stability Across Major Canadian Grocery Vendors

Jiwon Choi

November 14, 2024

This paper analyzes grocery price trends across eight major Canadian vendors, using data collected from Project Hammer to provide insights into price volatility and vendor-specific behaviors over time. By examining timestamped pricing records, we identify significant price fluctuations for specific vendors, particularly Metro and Voila, during certain periods, suggesting varied responses to market pressures. These findings reveal patterns of price stability and volatility that differ between vendors, highlighting potential competitive dynamics and the impact on Canadian consumers. This analysis contributes to a broader understanding of grocery pricing behavior, offering valuable information for consumers, policymakers, and regulators concerned with market fairness and affordability.

1 Introduction

The rising cost of groceries has become a pressing issue in Canada, impacting consumers nationwide and drawing increased scrutiny from policymakers, economists, and the public. With inflation and supply chain challenges affecting essential goods, understanding grocery price dynamics has taken on new importance. Project Hammer addresses this need by compiling and analyzing Canadian grocery price data across multiple vendors, aiming to provide insights into pricing trends, vendor behaviors, and potential drivers of price changes. This analysis is particularly relevant as it creates a historical record that can inform competition policy, consumer advocacy, and market regulation efforts, offering stakeholders a data-driven approach to address the complexities of the grocery sector.

This paper focuses on exploring grocery pricing data collected from eight major Canadian grocery vendors, including Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and

*Code and data are available at: https://github.com/jwonc4602/Canadian-Grocery-Price.

Table 1: Sample of Cleaned Grocery Data

[!h]

nowtime	current_price	old_price	price_per_unit	vendor	product_name	units	product_id
2024-06-27 09:39:00	3.99	5.29	$0.72/100ml	Voila	iÖGO nanö Lactose Free Drinkable Yogurt Apple-Grape 1% 6 x 93 ml	6 x 93ml	120047
2024-06-28 11:11:00	3.99	5.29	$0.72/100ml	Voila	iÖGO nanö Lactose Free Drinkable Yogurt Apple-Grape 1% 6 x 93 ml	6 x 93ml	120047
2024-06-29 10:39:00	3.99	5.29	$0.72/100ml	Voila	iÖGO nanö Lactose Free Drinkable Yogurt Apple-Grape 1% 6 x 93 ml	6 x 93ml	120047
2024-06-30 09:58:00	3.99	5.29	$0.72/100ml	Voila	iÖGO nanö Lactose Free Drinkable Yogurt Apple-Grape 1% 6 x 93 ml	6 x 93ml	120047
2024-07-01 10:13:00	3.99	5.29	$0.72/100ml	Voila	iÖGO nanö Lactose Free Drinkable Yogurt Apple-Grape 1% 6 x 93 ml	6 x 93ml	120047
2024-07-02 10:15:00	3.99	5.29	$0.72/100ml	Voila	iÖGO nanö Lactose Free Drinkable Yogurt Apple-Grape 1% 6 x 93 ml	6 x 93ml	120047
2024-07-03 17:54:00	3.99	5.29	$0.72/100ml	Voila	iÖGO nanö Lactose Free Drinkable Yogurt Apple-Grape 1% 6 x 93 ml	6 x 93ml	120047

Save-On-Foods. Through systematic data collection, this project captures variables such as timestamped price records, vendor identifiers, product details, and price-per-unit values, enabling detailed trend analysis over time. Unlike traditional economic datasets that often lack granularity, Project Hammer's data is structured to reflect real-time pricing fluctuations, sale events, and unit-based cost comparisons across vendors. This paper uses the dataset to investigate how prices have evolved across vendors over time and identifies specific periods of high volatility, providing insights into possible reasons behind these fluctuations.

Despite the wealth of data, analyzing grocery pricing trends presents challenges due to issues like missing data, bias, and distinguishing correlation from causation. This paper addresses these challenges by implementing statistical and data processing methods that improve data quality and enable reliable analysis. However, a clear gap remains: while existing studies provide broad overviews of grocery inflation, few have documented vendor-specific pricing behaviors and patterns over time. This paper fills that gap by focusing on vendor-level data and identifying periods where specific vendors exhibit unique pricing trends, which can highlight competition dynamics and consumer impact. The analysis also considers methodological limitations, such as potential biases in data extraction and missing data handling, to ensure transparency and rigor.

The findings reveal notable differences among vendors, with certain periods showing significant price spikes for specific vendors like Metro and Voila, suggesting susceptibility to market pressures or unique pricing strategies. By comparing these fluctuations across vendors, the paper highlights variations in pricing stability, with some vendors maintaining steady prices while others exhibit considerable volatility. These findings underscore the importance of tracking grocery prices over time and contribute valuable insights to discussions around market fairness, competition, and regulatory intervention. The paper is structured as follows: it begins with a detailed description of the dataset and methodology, followed by an exploration of key findings on price trends and volatility (Section 2), and concludes with a discussion on the implications for consumers, policymakers, and future research directions (Section 3). This structure provides a comprehensive overview of Project Hammer's findings and their relevance in today's economic landscape.

2 Data

This dataset, part of "Project Hammer," compiles Canadian grocery price data from February 28, 2024, to the most recent update (Filipp 2024). It focuses on pricing information from eight major vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. The project aims to improve competition in the Canadian grocery sector by creating a historical record of grocery prices accessible for analysis and potential legal inquiries. The data was compiled and examined using the R statistical programming software (**r?**) and SQL (SQLite Development Team 2024), supplemented by various tools such as `tidyverse` (Wickham et al. 2019), `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2023), `gridExtra` (Auguié 2017), `grid` (R Core Team 2023), `knitr` (Xie 2014), and `here` (Müller and Bryan 2020).

The dataset is organized into two primary files. The first, **Product Metadata**, provides detailed information about each product, including fields such as `product_id`, `product_name`, `brand`, and `units`. These fields enable the tracking and identification of unique products across different vendors, though it is important to note that `product_id` is specific to each vendor. The second file, **Raw Price Data**, contains time-series price data, capturing variables such as `nowtime` (timestamp), `current_price` (price at the time of capture), `old_price` (indicative of a sale price), `price_per_unit`, and additional status details like "Out of Stock" or "Sale." Together, these files support comprehensive tracking and analysis of grocery prices across time and vendors.

Each primary variable in the dataset offers unique insights essential for analyzing grocery price trends. The **nowtime** variable records the timestamp of each data extraction, providing a temporal reference that is vital for tracking price changes and understanding trends over specific periods. The **vendor** variable identifies one of the eight participating grocers, allowing for cross-vendor comparisons and trend analysis. Each product is assigned a **product_id** unique to the vendor, which enables historical tracking of individual items but may vary across different versions of the dataset. **Product_name** contains the product's name and sometimes the brand, acting as a primary identifier for user recognition. Additionally, **brand** explicitly identifies the product's brand when available, though some records may leave this field blank, requiring brand inferences based on product names.

The **units** variable indicates the measurement unit (such as grams or count) for each product, enabling standardization in price-per-unit comparisons across vendors. **Current_price** reflects the active price at the time of data capture, providing a focal point for monitoring inflation, price stability, and discounts. Meanwhile, **old_price**, a historic or "struck-out" price, denotes a sale event when present, allowing for analysis of promotional pricing frequency and impact. Finally, **price_per_unit** shows the per-unit cost as displayed on vendor websites, though it sometimes deviates from the calculated unit price, indicating potential discrepancies in vendor calculations. These variables collectively facilitate an in-depth examination of pricing behavior and trends in the Canadian grocery market.

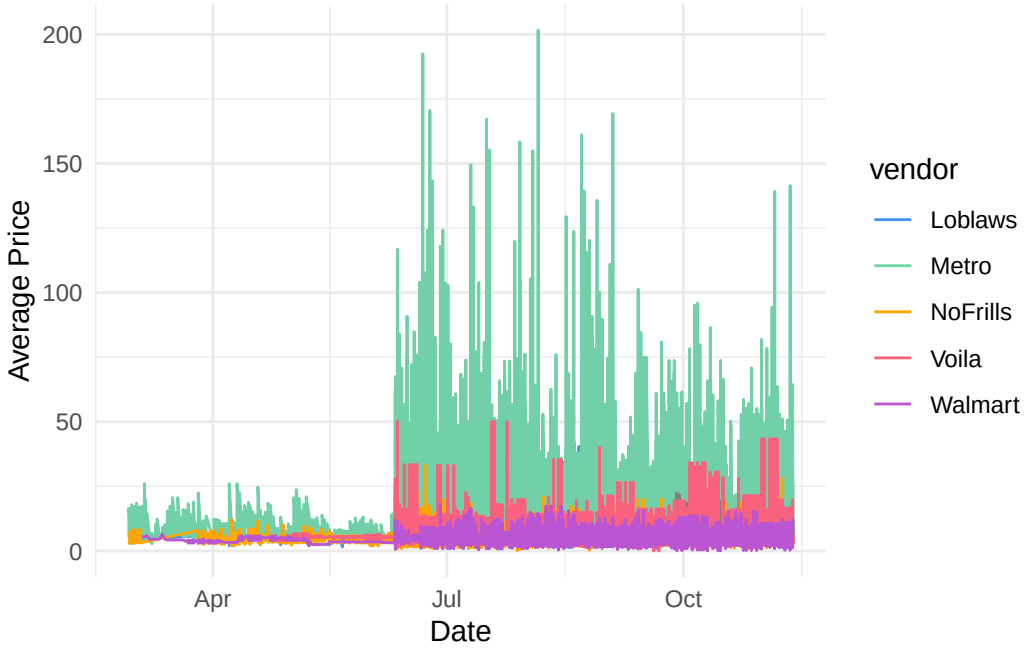

Figure 1: Average Pricing Trend for All Items over Time

Figure 1 displays the average price trends for various vendors over the course of the year, with distinct colors representing each vendor (Loblaws, Metro, NoFrills, Voila, and Walmart). Early in the year, from March to June, prices across all vendors are relatively stable, with only minor fluctuations around a low price range. However, from July onward, there is a marked increase in price volatility, particularly for Metro (green) and Voila (red).

Metro sees significant price spikes in late July and early October, reaching the highest average prices among all vendors during these periods. Voila also experiences noticeable price increases around the same periods, though not quite as high as Metro's peaks. In contrast, Walmart (purple) and NoFrills (orange) show more consistent pricing with fewer dramatic shifts, remaining generally lower than Metro and Voila in the latter half of the year. Loblaws (blue) has moderate fluctuations but does not reach the same extremes as Metro or Voila. This divergence suggests that Metro and Voila may be more susceptible to pricing fluctuations due to factors like demand surges or supply constraints, while Walmart and NoFrills maintain a more stable pricing strategy.

2.1 Measurement

In Project Hammer, ensuring that measurements align with the intended estimand—capturing genuine grocery prices across vendors—is essential. Valid measurements relate closely to the actual prices consumers encounter, reflecting real costs rather than promotional or artificially

adjusted figures. Standard metrics, like prices and timestamps, are relatively straightforward. However, constructed metrics, such as `price_per_unit`, introduce complexity, as vendor-provided per-unit prices sometimes differ from computed values (e.g., `current_price` divided by `units`). These discrepancies, potentially due to vendor rounding practices, require transparent reporting to prevent bias and ensure that our data reliably reflects true pricing strategies.

Reliability also plays a critical role, given the dynamic nature of online data extraction. For instance, duplicate entries on the same day for a product can arise when it is promoted across multiple categories, or due to automated listings. Such variations necessitate checks to confirm consistency in `current_price` and `old_price`, capturing sales conditions accurately. These measurement decisions echo broader challenges found in complex constructs, where context—such as promotional timing and vendor-specific factors—can influence interpretation. By rigorously capturing timestamp (`nowtime`) and vendor identifiers, the dataset provides a reliable and contextualized view of grocery pricing in the Canadian market.

3 Discussion

3.1 Correlation vs. Causation

Correlation refers to a statistical relationship between two variables, where changes in one variable are associated with changes in the other. However, correlation alone does not imply causation, meaning that just because two variables move together, it doesn't mean that one variable causes the other to change. For instance, ice cream sales may increase alongside drowning incidents during the summer months, but this doesn't imply that ice cream consumption causes drowning; rather, both are influenced by the warmer weather. In data analysis, identifying correlations can be useful for recognizing patterns, but it's crucial to avoid jumping to conclusions about cause-and-effect relationships without further investigation.

Causation, on the other hand, indicates that one event directly affects another. Establishing causation requires more rigorous methods, such as controlled experiments or advanced statistical techniques, to rule out other factors and isolate the effect of one variable on another. Failing to distinguish between correlation and causation can lead to misleading conclusions, which is particularly problematic in fields like healthcare, economics, and social science. Analysts must be cautious and use appropriate methods to determine causality, as decisions based on misinterpreting correlations as causative can have significant real-world consequences.

3.2 Missing Data

Missing data occurs when observations or values are absent from a dataset, which can arise for various reasons, such as data entry errors, non-responses in surveys, or system failures. Missing data can reduce the quality of analysis and compromise the results, as it limits the

completeness and accuracy of the information available. Different patterns of missing data, such as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), have unique implications on how data analysis should proceed. Ignoring missing data or handling it inappropriately can lead to biased results and undermine the validity of the analysis.

To address missing data, analysts may use various techniques, including deletion methods, imputation, or advanced modeling approaches. Simple deletion of records with missing values can reduce data size, potentially introducing bias if the missing data is systematic rather than random. Imputation methods, such as mean imputation, regression, or machine learning-based techniques, aim to fill in missing values with plausible estimates, allowing for a more comprehensive analysis. Selecting the right method for handling missing data depends on the nature and extent of the missingness and the goals of the analysis, with more sophisticated methods often yielding more reliable outcomes in cases of extensive or non-random missing data.

3.3 Sources of Bias

Bias in data analysis refers to systematic errors that can lead to inaccurate or misleading results, often skewing findings in a particular direction. There are various sources of bias, including sampling bias, where the sample is not representative of the population; measurement bias, where data collection methods systematically favor certain outcomes; and confirmation bias, where researchers may subconsciously look for evidence that supports their hypotheses. Each type of bias can compromise the validity of results and is particularly detrimental in fields that rely on objective evidence, such as scientific research and policy-making.

Identifying and addressing bias is essential for conducting credible and ethical research. Techniques to minimize bias include random sampling, blinding, and careful design of data collection instruments. Additionally, transparency in reporting methodologies and potential limitations allows others to assess and interpret the results with an understanding of any inherent biases. Recognizing and mitigating bias is crucial for data analysts, as biased results can lead to misguided decisions, reinforce stereotypes, or create misinformation that impacts public understanding and trust.

References

Auguié, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

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

Müller, Kirill, and Jennifer Bryan. 2020. *Here: A Simpler Way to Find Your Files*. https://cran.r-project.org/web/packages/here/index.html.

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

SQLite Development Team. 2024. *SQLite Documentation*. https://www.sqlite.org/docs.html.

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

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, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.