Canadian Grocery Price Analysis on 200G Miss Vickie's Original Recipe Chips of Eight Main Grocery Vendors*

Walmart the Lowest, T&T the Highest, and Mid-tier Range for Others

Angel Xu Yunkai Gu Yitong Wang

November 14, 2024

This paper analyzes observational data on grocery retailer behavior, specifically focusing on price changes over time and their relationship to product availability and promotions. Data manipulation is conducted entirely in SQL to ensure transparency and reproducibility, with R employed for visualization and statistical exploration. Besides the fact that T&T consistently maintains the highest prices across the months and Walmart maintains the lowest, Galleria and Loblaws exhibit stable pricing in a mid-tier range, while Metro and SaveOnFoods display more noticeable month-to-month price changes. Our findings offer insights into grocery pricing dynamics and promotion strategies, helping public have a deeper understand of grocery market.

Table of contents

*Code and data are available at: https://github.com/Kylie309/Canadian-grocery-price-analysis.

1 Introduction

Retail analytics relies heavily on understanding consumer behavior to optimize inventory, pricing, and promotions. In the grocery sector, retailers' decisions regarding price adjustments, stock availability, and promotional tactics significantly influence market dynamics. Through observational data, this study analyzes how grocers adjust prices over time in response to various factors, including seasonal demand, supply chain changes, and competition. Specifically, we examine these pricing patterns of 200G Miss Vickie's Original Recipe Chips, and their impact on stock levels and promotions to understand grocers' strategic behavior in a dynamic market.

This paper emphasizes reproducibility and clarity throughout our analysis. We utilize SQL exclusively for data manipulation to ensure a transparent and structured workflow, while R provides the framework for detailed visualization and statistical analysis. Our discussion addresses three fundamental challenges in observational studies: distinguishing correlation from causation, managing missing data, and identifying sources of bias. These considerations are vital for accurate data interpretation, particularly as grocery pricing and availability decisions are often influenced by external factors that observational data alone cannot fully capture.

In the results section, our analysis highlights distinctive pricing strategies and trends among various grocery vendors over a six-month period. T&T consistently maintains the highest prices across the months. Conversely, Walmart's prices remain the lowest, showcasing a commitment to a low-cost strategy. Other vendors, such as Galleria and Loblaws, exhibit stable pricing in a mid-tier range with minor fluctuations. Meanwhile, Metro and SaveOnFoods display more noticeable month-to-month price changes, which may reflect seasonal promotions or competitive adjustments. This diversity in pricing behavior highlights the varied market positions and strategies of each vendor.

The results of the study could be important and significant because it provides valuable insights into the strategic behavior of grocery retailers, helping to establish reasonable and optimized pricing and promotion strategies. By understanding how grocers adjust prices in response to dates, and even underlying factors such as seasonal demand and supply chain changes, this research also helps decision-making process for consumers by presenting dynamic patterns of grocery prices. This study provides actionable conclusions for retailers and buyers, contributing to discussions on pricing transparency and broader economic implications of grocery pricing.

The remainder of this paper is structured as follows. Section 2 introduces the overview of the data (Section 2.1) and measurement (Section 2.2). Section 3 presents the result of our analysis, and Section 4 discusses the study in three perspectives: the differences between correlation and causation (Section 4.1), potential reasons of missing data (Section 4.2), and sources of bias (Section 4.3).

2 Data

2.1 Overview

The raw data for the study was initially obtained from Project Hammer (Filipp 2024). Following Alexander (2023), this study considers the grocery prices of 200G Miss Vickie's Original Recipe Chips from eight main grocery vendors by dates.

The analyses presented in this paper were conducted using R programming language (R Core Team 2023) and database software SQLite (Consortium 2024). The `tidyverse` packages (Wickham et al. 2019) were used in the process of data simulation. We use `testthat` package (Wickham 2011), `tidyverse` (Wickham et al. 2019) and `dplyr` (Wickham et al. 2023) to develop the test for structure and format of simulation and analysis data. Data cleaning progress used database software SQLite (Consortium 2024). Then, results were presented by graphs using `ggplot2` package (Wickham 2016), `lubridate` package (Grolemund and Wickham 2011) and `janitor` package (Firke 2023).

2.2 Measurement

The dataset used in this analysis translates real-world grocery prices from multiple vendors into structured data, providing a digital snapshot of pricing behaviors across a six-month period. The data collection process likely involved systematic recording of product prices, vendor identities, and timestamps for each entry. This process captures dynamic pricing, where price adjustments reflect factors such as supply and demand, seasonal shifts, promotional events, and market competition. By recording prices over time, the dataset enables analysis of both long-term trends and short-term fluctuations.

To convert this real-world phenomenon into data entries, individual prices were likely sourced from point-of-sale systems, online listings, or periodic manual recordings, standardizing each entry with essential attributes: current_price, vendor, product_name, nowtime, potentially brand and product_id to identify unique products. Each price entry reflects a snapshot of the vendor's pricing strategy on a given day, encompassing both stable pricing practices and variable strategies, such as discounts or premium adjustments.

In the digital dataset, timestamps are formatted to represent exact date and time, allowing each record to be linked to a specific month and facilitating monthly trend analysis. Vendors are categorized by name, enabling direct comparison of pricing strategies across different brands. This structured approach to data collection allows for the investigation of broader market behaviors and pricing trends, as each data point becomes an entry reflecting the intersection of real-world economic factors and vendor pricing decisions.

This transformation from real-world grocery pricing to a structured digital dataset—captures the nuances of competitive pricing in the grocery industry and offers insights into how vendors respond to external pressures. Through this data, we can quantitatively examine the varied approaches each vendor takes in navigating market demands, cost management, and customer appeal.

3 Results

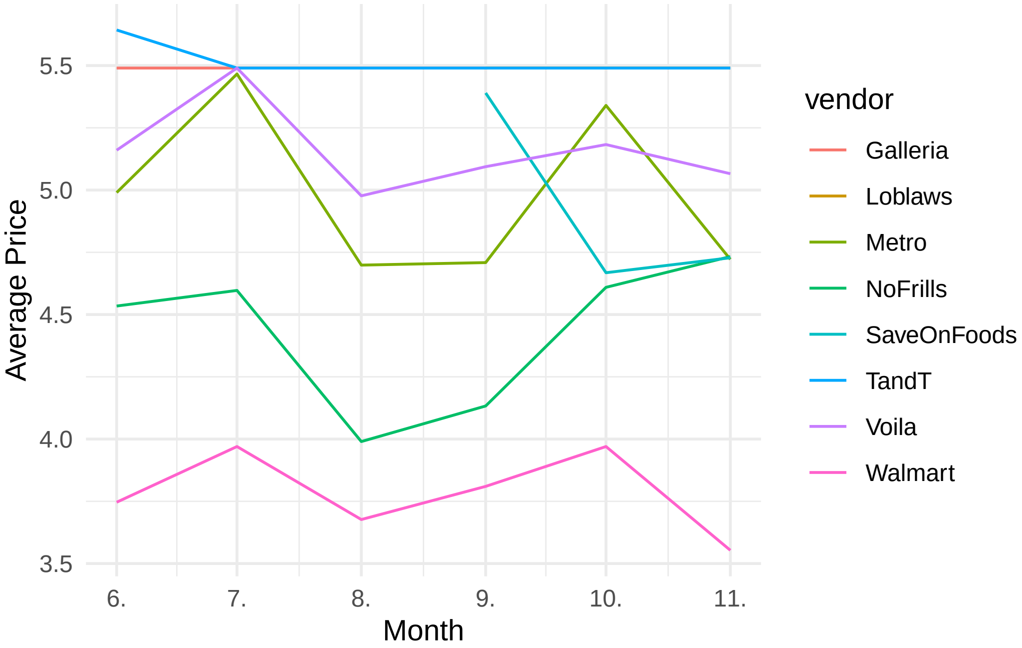

Figure 1: Monthly Average Price Trend by Vendor

Figure 1 illustrates the average price trends for different vendors over a six-month period from June to November. Examining each vendor's pricing behavior reveals distinctive patterns and possible underlying strategies.

T&T stands out with its consistent pricing, as shown by the blue line, which remains virtually unchanged across the months. This stability suggests a consistent pricing approach, potentially due to fixed-cost items or a stable supply chain that allows T&T to maintain predictable prices. In contrast, Metro and SaveOnFoods exhibit significant price fluctuations. Metro's prices drop sharply from June, reaching their lowest in August, before climbing again in October and then dropping in November. SaveOnFoods also shows notable variation, particularly with a decrease around September followed by an increase in October. Such trends might be a result of seasonal promotions, temporary adjustments, or competitive pricing changes.

Walmart's prices, represented by the pink line, consistently remain lower than those of other vendors, staying below 4.5 throughout the period. This aligns with Walmart's well-known low-cost pricing strategy, which emphasizes affordability. Voila, represented by the purple line, shows moderate variability, with a price dip in July and August and a slight increase in September, suggesting seasonal adjustments or strategic price changes to remain competitive.

Galleria and Loblaws demonstrate a more stable trend with only minor fluctuations. Galleria's prices generally hover around the 5.0, indicating steady pricing without major seasonal or promotional impacts. Loblaws shows slight movement but maintains a close range around 5.0, suggesting it follows a consistent pricing strategy with occasional adjustments.

The variations in pricing across these vendors could be explained by several factors. For instance, the pronounced fluctuations observed for Metro and SaveOnFoods may result from promotional events, discounts, or seasonal sales aimed at boosting demand. T&T's flat trend and Walmart's consistently low prices might reflect stable supply chains or standardized pricing policies that align with each brand's strategy. Vendors like Galleria, Loblaws, and Voila may be adjusting prices in response to market dynamics or competitor actions.

Figure 2 provides a different perspective compared to the previous time trend plot by focusing on the distribution of individual prices for each vendor within each month. Unlike the time trend plot, which emphasizes average monthly trends and highlights general pricing movements over time, this scatter plot reveals more granular details about the price points vendors set each month. This plot allows us to observe the range and dispersion of prices across vendors on a monthly basis, offering insights into each vendor's price stability, variability, and positioning within each discrete month.

Figure 2 reveals that T&T maintains a consistently high price level around 6 for every month, indicating a stable premium pricing strategy. This stability suggests that T&T does not engage in monthly price adjustments or promotions, instead choosing to keep its prices fixed at a high level. This could reflect a strategy aimed at higher-end consumers, emphasizing quality and consistency.

On the other hand, Walmart maintains its prices around 3 consistently across all months, as seen from the uniformly positioned pink points. Walmart's approach to monthly pricing is

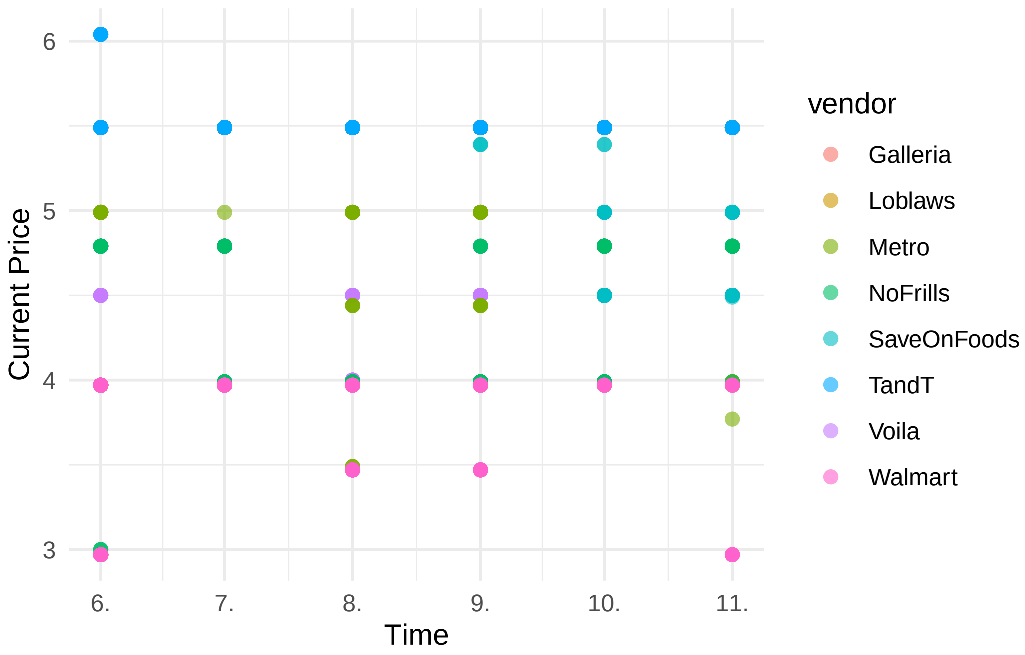

Figure 2: Scatter Plot of Current Price by Vendor

similarly stable, but at the opposite end of the price spectrum compared to T&T. This stability at a lower price range reinforces Walmart's commitment to a low-cost pricing strategy, aiming to attract price-sensitive consumers with minimal monthly price changes.

For vendors like Galleria and Loblaws, moderate price points are shown in the range of 4 to 5 each month, with small variations. This pattern suggests a mid-tier pricing strategy that is somewhat responsive to monthly market conditions. However, the limited movement in prices for these vendors indicates that they also prefer consistency, likely to maintain a balance between competitive pricing and stable revenue without drastic month-to-month changes.

Some vendors, including Metro, NoFrills, and SaveOnFoods, display slight price shifts between months, showing variability different from T&T or Walmart. Metro's prices appear to fluctuate slightly month-to-month, indicating a more flexible approach where prices are adjusted based on demand or competitive factors within each month. This variation may suggest a strategy aimed at responding to specific monthly market dynamics or consumer purchasing behavior. SaveOnFoods and NoFrills also show scattered points around their monthly price range, reflecting a similar but less pronounced tendency toward price flexibility.

Voila stands out with prices that generally fall between Walmart's low and the middle levels of Galleria and Loblaws. Its prices vary across months within a mid-to-low range, suggesting a niche or value-oriented strategy. Voila's monthly prices do not follow a rigid pattern, as seen with T&T or Walmart, and may instead adjust based on short-term market factors, reflecting a more adaptive monthly pricing strategy.

Figure 2 highlights that while some vendors, like T&T and Walmart, maintain consistent price points every month, others exhibit moderate to slight variations within each month. The clustering of prices in different ranges reflects distinct pricing strategies across vendors within individual months. In contrast to Figure 1, which shows overarching trends and general monthly averages, Figure 2 emphasizes the nuances of each vendor's pricing flow and flexibility, revealing whether they adopt static or adaptive pricing within each month.

This analysis suggests that vendors follow diverse approaches, with some preferring fixed monthly pricing and others adjusting to capture variations in consumer demand or competition. This insight underscores the importance of evaluating individual price points to understand vendors' strategies at a more detailed monthly level, capturing nuances that an average trend might overlook.

4 Discussion

4.1 Causation vs Correlation

In observational data analysis, correlation and causation are crucial yet distinct concepts.

- Correlation refers to a statistical relationship between two variables, where changes in one are associated with changes in another. This association, however, does not imply that one variable causes the other to change; for instance, while we might observe a correlation between price reductions and increased sales, this alone doesn't confirm that price changes directly drive sales.
- Causation implies a direct cause-effect relationship, where changes in one variable directly influence the other. Establishing causation requires controlled conditions or advanced statistical methods to rule out confounding factors.

Take example as observational data, like our grocery dataset, allows us to identify correlations but not causation, due to potential external influences such as seasonality or competition. Recognizing this distinction is essential in our study to avoid overstating the findings and to ensure that our conclusions about grocers' pricing strategies are realistic and data-supported.

4.2 Missing Data

In the "Project Hammer" initiative (Filipp 2024), Jacob Filipp provided a dataset of grocery prices from various Canadian retailers. While the project aims to provide comprehensive pricing information, several factors may contribute to missing data:

4.2.1 Data Collection Methodology

According to Jacob Filipp, the data provided is “from a screen-scrape of website UI” (Filipp 2024), which is a potential reason why there is missing information that might get from the internal APIs that power grocers’ websites.

The changes in website structure would disrupt the scraping process and result in missing or incomplete data. If there exist some mechanisms that prevent automated scraping on the websites, the data for certain products or retailers would be missing. Additionally, potential failures of some scraping tools due to technological issues may also lead to loss of data.

4.2.2 Temporal Issues

The data is collected on a irregular data collection frequency, i.e. it is not collected every single day. For example, some dates may have been missed due to technical issues in the scraping schedule. This could create temporal gaps in the dataset, which may be particularly problematic for analysis requiring daily price comparisons. Also, not every vendor may have been consistently included in the data collection process.

Specifically, according to Jacob Filipp (Filipp 2024), “the prices were gathered only for a ‘small basket of products’ between Feb. 28 and July. 10/11”, and prices for a much larger variety of products were gathered later on.

This indicated that data of prices for some products were not included during certain time interval, but were then recorded in the dataset, leading to NA values appearing inside the raw data.

4.2.3 Variations across vendors

After browsing through the raw data, we could see that different vendors have different naming methods for the same product. The dataset also included a variety of Miss Vickie’s chips with different sizes and packaging (e.g. 200g, 59g, 275g), and the dataset does not always specify which size is being priced at each retailer in the “units” column.

The cleaning process overtaken by this study cleaned the raw dataset for identifying the 200G Miss Vickie’s Original Recipe Chips by checking both the product name and brand name columns of the raw data. However, there may occur issues regarding dropping valuable information due to the confusing names recorded.

The cleaning process then continues by filtering out other sizes of chips according to the “units” column. However, the column also encounters issues of missing information and error messages (e.g. containing data of prices per unit in the column instead of weights), and may lead to problems after cleaning them.

4.3 Source Bias

4.3.1 Concept Explanation

Source bias, in the context of data analysis, refers to systematic errors or distortions that arise from the way data is collected, the sample used, or the limitations of the data source itself. This type of bias can affect the validity and reliability of findings, as it may lead to overrepresentation or underrepresentation of specific trends, populations, or behaviors within the dataset. Source bias can stem from several factors, including the selection of subjects, the timing and location of data collection, and the design of the study. In retail and price trend analysis, source bias might mean that certain vendors, price points, or products are disproportionately represented, potentially skewing conclusions drawn from the data.

4.3.2 Bias on Price Trend

As the data used in our analysis is accessed from Jacob Flipp's Hammar project (Filipp 2024), the subset of grocery vendors and their prices may not represent the full market dynamics comprehensively. In turn, this bias may lead to impact on price patterns interpretation across vendors. Included vendors of Walmart, T&T, and Loblaws are incomplete since there may be other competing vendors not represented here. This incompleteness limits the generalizability of our findings.

Additionally, if the data collection process is biased toward particular types of products, geographic regions, or times of year, this could impact observed price trends. For example, prices might appear more stable or variable than they actually are if certain promotional periods or seasonal items are overrepresented or underrepresented.

4.3.3 Bias on Vendor Pricing Model

Another potential source of bias in this analysis is the consistency and completeness of data for each vendor. Vendors with fewer data points might appear to have more stable prices merely due to a lack of data that captures their monthly fluctuations. Conversely, vendors with a comprehensive set of monthly records may appear more variable, even if their actual pricing strategy is stable.

For example, vendors like T&T and Walmart have more complete and consistent monthly records, capturing price points across all observed months from June to November. This level of data completeness allows for a more accurate assessment of their pricing strategies and trends.

However, vendors such as NoFrills and SaveOnFoods have fewer data records, which might make their prices appear artificially stable or unchanging. With fewer records, there's a higher likelihood that price fluctuations within each month or across different months are missed. This could create a misleading impression that these vendors maintain consistent pricing when, in reality, there may be more variation that the dataset does not capture.

This uneven representation can introduce bias in interpreting which vendors have consistent or adaptive pricing models.

4.3.4 Original Data Collection Bias

The raw dataset from Jacob Filipp's Hammer project (Filipp 2024) inherently carries the limitations and potential biases associated with the original data collection methods used in that project. Without additional details on the data collection methodology, such as sampling criteria, the geographic scope of vendors, or the diversity of product categories, it will be challenging to ascertain the full extent of the source bias. However, acknowledging these potential biases is crucial for interpreting the results accurately, as it reminds us that findings may not be universally applicable to all grocery vendors or regions. This understanding of source bias highlights the importance of complementing this dataset with additional sources or broader market data if more generalized conclusions are desired.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Consortium, SQLite. 2024. "SQLite: A c Library That Implements a SQL Database Engine." https://www.sqlite.org/.

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

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

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