Certain Bundles of Products Tend to Get Marked-Down More During Sales.*

We compare each product's price data for Loblaws and Walmart over time.

Talia Fabregas Lexi Knight Fatimah Yunusa
Aliza Mithwani Arav Agarwal

November 14, 2024

This paper analyzes grocery pricing data to identify patterns in markdown frequency and magnitude across vendors. We found that certain product bundles are consistently discounted more often or to a greater extent, suggesting strategic pricing trends within the sector. These findings help us find out how vendors approach sales, providing new and established vendors with guidance on optimizing markdowns. This analysis also helps policymakers understand the economic impacts of pricing strategies on households, especially during periods of economic change.

Table of contents

*Code and data are available at: [https://github.com/alizamithwani/Grocery-Prices.git).

1 Introduction

This paper uses data from Jacob Phillip's website for Project Hammer @ which aims to reduce collusions and increase competitiveness in the Canadian Grocery sector. The dataset reveals information about the prices over time as observed from the timestamps of when the data was collected, which can be used to deduce how frequently items went on sale. The old (before-sale) price as displayed on the product can be compared to the new price to determine how much each product has been marked-down by for each vendor.

This is an important topic as it helps new vendors understand which products they should be marking-down for sales and the reasoning behind this. Governments can also make better deductions and predictions about which types of households are most affected during times of deflation or when prices fall based on the typical basket of good they purchase and which frequently marked-down product(s) are included in this. Pricing strategy is a complex topic and could vary between vendors, but if the same few items are consistently marked-down during sales, there must be a deeper-rooted reasoning behind it. We collected and analyzed data on product pricing over time from Project Hammer. We measured the markdown frequency and the magnitude of the markdown as a percentage of the original price.

In this paper, the primary estimand is the average markdown frequency and markdown magnitude for product bundles across different vendors in the Canadian grocery sector, particularly Loblaws. Specifically, we aim to estimate the probability and extent to which certain groups of products experience price reductions during sales events and try to see if there is a pattern. . The estimand contains two components: (1) the likelihood (frequency) that a product bundle is marked down by each vendor within a given timeframe, and (2) the average percentage decrease in price (magnitude) for these markdowns. By examining these two elements, we can quantify and compare vendor behaviors in discounting strategies, ultimately identifying whether specific products or bundles are systematically discounted more frequently or to a greater extent across vendors.

The remainder of this paper is structured as follows: Section 2 presents the methods used to analyze grocery prices , followed by Section 3, which details our findings. Section 4 offers a discussion of the findings, limitations and suggestions for future research directions.

2 Data

2.1 Overview

This dataset includes grocery price information gathered from various Toronto vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. Prices were collected for products under the "in store pickup" option in a Toronto neighborhood. The data was screen-scraped directly from website user interfaces, so some information that may be accessible via internal APIs is missing. Initially, data collection focused on a "small basket" of products, from February 28 through early July, after which it expanded to cover a larger variety of items. This structured dataset allows us to track price changes over time across multiple vendors, offering a foundation for comparative analysis of pricing behaviors and promotions.

However, the dataset does have some limitations due to its collection method. Notably, data for certain vendors or specific days may be absent if extraction processes failed, leading to potential gaps in the dataset. Additionally, product identification relies on vendor-specific internal IDs that may change with data refreshes, limiting the continuity of analysis over multiple data versions. These IDs remain consistent within a single vendor for any given download, though they are not universal across vendors. Understanding these limitations is important for accurate interpretation, especially when comparing price behaviors across different vendors.

We used R (R Core Team (2023)) to write this paper, the `ggplot2` package from Wickham (2016) to produce the visualizations, and the `arrow` package from Richardson et al. (2024) to save and access our cleaned data. We used embedded SQL in R and the `RSQLite` package from Müller et al. (2024) to clean our raw data.

2.2 Measurement

The data is structured with columns capturing details such as timestamps, vendor information, product ID, product name, brand, units, and pricing details. Timestamps mark when prices were gathered, enabling time-series analysis of price trends. Product names often include brands or units, while separate columns provide brand names and units explicitly, though these may be blank for some vendors. Each record includes a "current price" at the time of extraction, an "old price" indicating a struck-out former price to signal sales, and a price per unit, though this unit price may occasionally differ from an actual division of price by units.

Each product entry is linked to a unique ID within the vendor's system, allowing consistent price tracking over time. However, these IDs are not stable across dataset versions, meaning that new downloads may introduce different IDs, complicating long-term tracking. In cases where brands are represented differently across vendors (e.g., "Miss Vickies" with variations

in apostrophe placement), SQL queries are used to account for these inconsistencies, ensuring comprehensive analysis across vendor representations.

2.3 Outcome variables

The primary outcome variables in this study are the frequency and magnitude of markdowns across products for each vendor. The "current price" and "old price" columns capture direct price changes, providing insights into markdown events. The presence of an "old price" generally signifies a sale, allowing for differentiation between explicit sales events and subtler price fluctuations. By observing trends in these variables, it's possible to quantify markdown frequency (how often a product is marked down) and magnitude (the amount by which prices are reduced) across different vendors.

This setup also allows for exploring whether specific product bundles are subject to more frequent or substantial markdowns. By examining various combinations of products and vendors, patterns in promotional pricing can be identified, contributing to a broader understanding of sales strategies across the grocery market. This analysis is particularly valuable for identifying whether vendors follow similar or different markdown practices, highlighting competitive pricing behaviors within Toronto's grocery industry.

2.4 Predictor variables

In addressing the question of whether certain product bundles tend to be marked down more frequently or at greater magnitudes during sales, the dataset includes several key predictor variables. The primary predictor is the vendor variable, indicating the grocery store from which the data was collected (e.g., Walmart Canada, T&T, Loblaws). Vendor differences are likely to influence pricing and markdown strategies, as each vendor may have distinct promotional cycles and strategies for specific product bundles. Additional predictor variables include product_name and brand, capturing product characteristics that may impact the likelihood and extent of markdowns. Since product names can include brand and unit details, these variables help to differentiate products at a granular level, allowing for more accurate analysis of markdown trends across distinct bundles.

Another set of predictor variables critical to this analysis includes price attributes like current_price, old_price, and price_per_unit. The current_price represents the price at the time of data collection, while old_price (if available) indicates a prior, often higher, price point that signals a markdown. By comparing current_price and old_price, we can quantify markdown magnitudes for each product and bundle. The price_per_unit variable provides a standardized measure, allowing comparisons of prices on a per-unit basis, which is especially valuable for analyzing bundle-level markdowns across different vendors. Together, these predictors enable a robust investigation into the frequency and magnitude of markdowns across grocery vendors and product bundles.

3 Results

4 Discussion

4.1 Correlation vs. Causation

Understanding the distinction between correlation and causation is essential for accurate analysis and responsible decision-making. Correlation indicates an association between two variables that tend to move together, but it does not imply that one causes the other. Causation, however, is a cause-and-effect relationship where changes in one variable directly lead to changes in another. Establishing causation often requires rigorous methods, such as controlled experiments, to rule out other influencing factors and confirm the direct link between variables.

Misinterpreting correlation as causation can lead to incorrect conclusions and ineffective policies or actions. By recognizing that correlation does not automatically mean causation, researchers and analysts can avoid drawing unwarranted inferences and instead base conclusions on solid, evidence-backed relationships. Distinguishing between these concepts ensures more accurate data interpretation and more effective decisions across various fields.

In our study, a critical point of discussion is the distinction between correlation and causation in product markdown trends. While patterns observed in the markdown frequency and magnitude may correlate with certain product types or vendor behaviors, establishing a causal relationship would require additional controls and experimental design. Our observational approach limits our ability to definitively conclude that any identified trends directly cause certain markdown behaviors. Future studies involving controlled experiments or quasi-experimental designs could more accurately attribute causative factors to these pricing trends.

4.2 Missing Data

Missing data posed some limitations in our analysis, primarily due to inconsistencies in data extraction across different vendors and time points. For instance, gaps in data for certain vendors or days may have introduced bias in our findings, affecting the accuracy of markdown frequency estimations. Addressing missing data through imputation methods or focusing on vendors with consistent data availability could improve robustness. Future efforts should also consider enhancing data collection methods to reduce missing information and provide a more complete view of pricing trends. We removed observations where old_price and sale columns were N/A, as this means these items never went on sale. We only analyzed and compared 2 vendors: Loblaws and Walmart. This is for simplicity and since product IDs are not consistent across vendors.

4.3 Sources of Bias

Our analysis is subject to several potential sources of bias, including selection and vendor-specific biases. The selection of specific vendors and products may not represent the entire grocery sector, and variations in vendor-specific data may reflect internal policies rather than broader market trends. Additionally, our reliance on screen-scraped data limits our control over certain product identifiers, which may vary between downloads, leading to inconsistencies in long-term tracking. Addressing these biases in future studies by expanding the vendor pool and ensuring consistent data capture would help to mitigate these concerns.

Furthermore given that the analysis is limited to items that have been on sale and data is restricted to only two vendors, this could introduce selection and representational biases that affect the generalizability of the findings. Since only sale items are included, the data may not fully represent the broader pricing strategies of vendors, as it excludes regularly priced items that could reveal different patterns in vendor behavior. By focusing exclusively on sale items, we might overestimate markdown frequency or trends that do not apply to the entire inventory. Additionally, restricting the dataset to two vendors may introduce vendor-specific biases; these vendors' sales strategies may not reflect the strategies used across the wider market. This selection may create an analysis that, while accurate for these two vendors, does not generalize well across other vendors in Toronto or elsewhere.

4.4 Weaknesses and next steps

The study's primary limitations include data collection constraints and the inability to account for all external factors influencing markdown decisions, such as seasonality or vendor inventory levels. Next steps could involve gathering data over extended periods and across multiple regions to capture a more representative view of grocery pricing strategies. Additionally, exploring machine learning models that account for external variables like demand forecasting or vendor inventory would provide more nuanced insights into pricing dynamics. Expanding the dataset and refining the model can strengthen the reliability of the study's conclusions.

In future studies, expanding to a broader selection of vendors and incorporating items regardless of sale status would mitigate these biases and offer a more representative insight into the grocery sector's pricing behavior.

References

Müller, Kirill, Hadley Wickham, David A. James, and Seth Falcon. 2024. *RSQLite: SQLite Interface for r*. https://CRAN.R-project.org/package=RSQLite.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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