Grocery Prices by Vendor*

Sophia Brothers Bruce Zhang Ruizi Liu Deyi Kong
Yuechen Zhang Yingke He

November 14, 2024

This analysis explores price trends across different vendors for grocery products using observational data. SQL was used for data preparation while R was employed for visualization. Key points in this paper include the analysis of price distributions by vendor, challenges in interpreting observational data, and discussions on data limitations such as missing values and potential biases.

Table of contents

1 Introduction **2**

2 Data **2**

3 Measurement and Data Preparation **2**

4 Discussion **4**
4.1 **Loblaws and Metro are Pricier than other Grocery Chains** 4
4.2 **Correlation vs. Causation** 4
4.3 **Missing Data** 4
4.4 **Sources of Bias** 4

5 Conclusion **5**

References **6**

*Code and data are available at: https://github.com/sophiabrothers1/groceryprices.

1 Introduction

The grocery market offers dynamic pricing across various vendors and product types, making it essential to understand patterns for consumer insights and economic analysis. This study examines grocery prices by vendor to find average price trends, using SQL for data extraction and R (R Core Team 2023) for visualization.

2 Data

This data was obtained from an online source of grocery prices related data Filipp (2024). The analysis made use of R Programming language R Core Team (2023), including packages dplyr Wickham, François, et al. (2023), tibble, Wickham, Vaughan, et al. (2023), and ggplot2 Wickham et al. (2023). The dataset displays the price of different products for different vendors. The vendors include a series of grocery chains including Metro, Loblaws, Walmart, T&T, and a few more.

The dataset includes:

- Vendor: One of seven major grocery vendors.

- Product Name: The name of the product

- Current Price: Price at the time of data collection.

3 Measurement and Data Preparation

Data manipulation and aggregation were conducted in SQL, calculating average product prices by vendor. Figure 1 shows that for products that are carried by >= 3 vendors, Metro has the most expensive price 28609 times. Loblaws follows shortly behind at 27561.

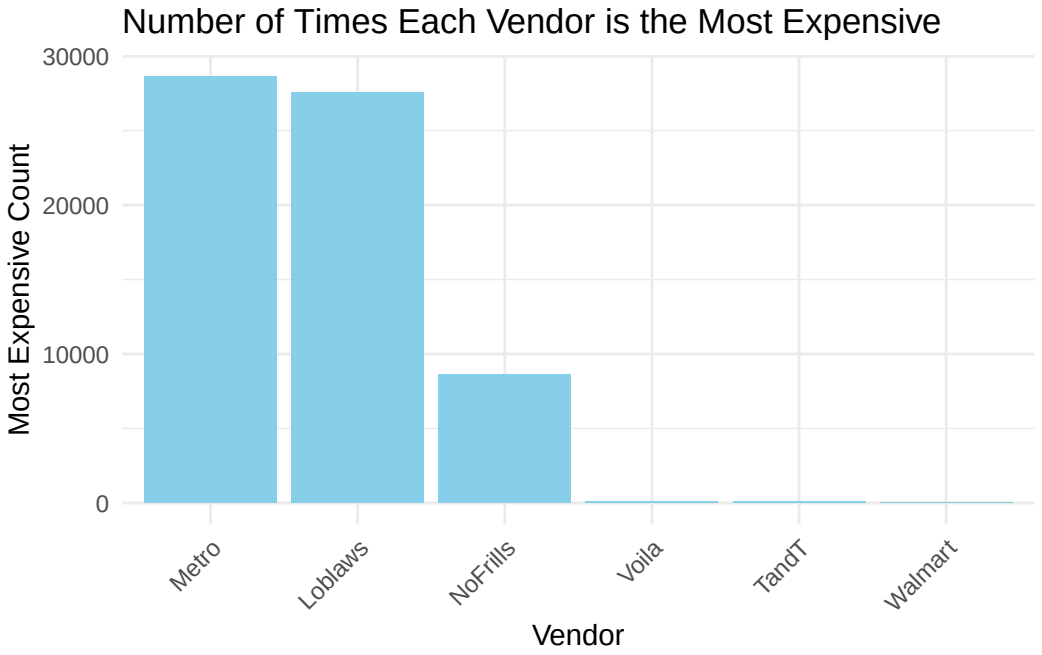

Figure 1: This bar graph shows the number of times a product was found to be the most expensive at a specific vendor. Products must be cross-carried by three vendors to count.

4 Discussion

This study highlights observable trends in grocery prices by vendor but also underscores limitations due to the nature of observational data.

4.1 Loblaws and Metro are Pricier than other Grocery Chains

Loblaws and Metro are more expensive than other grocery chains based on the number of products that are most expensive at these chains. This can demonstrate real-world trends of pricing, as these chains are large and often located in urbanized regions. However, we cannot conclude a causation between location, size, and price. Many limitations of this paper highlighted below show that further research is needed to make meaningful, causal conclusions.

4.2 Correlation vs. Causation

While we observe differences in pricing across vendors, these differences do not imply causative factors without further experimental controls. In order to conclude causation, it is important to have an experimental design that involves the manipulation of variables. This dataset is obtained based on observed price trends and does not invovle variable manipulation. Similarly, vendors may set prices based on external market trends or supply chain factors not captured in this dataset. Further data collection and analysis including variable manipulation must be done in order to conclude causation.

4.3 Missing Data

Missing data is a common challenge in grocery datasets. It is unlikely that every aspect of each data point is collected perfectly, especially in a grocery store setting where the main purpose is to sell products instead of data collection. For example, some vendors may not report product prices consistently, leading to potential biases in the average prices. Handling missing data effectively, such as by imputation or exclusion, is essential for future analyses.

4.4 Sources of Bias

Several potential biases may influence these results. Vendor pricing strategies vary based on store policies, regional preferences, or consumer behavior. The pricing may be influenced by the abundance of the grocery chain in specific locations. For example, more urbanized regions may have more larger grocery chain stores and also have higher pricing compared to rural areas. Furthermore, online-only data may exclude regional in-store promotions and other timnely changes in pricing, limiting the generalizability of our findings.

5 Conclusion

The data in this analysis present a general idea of the pricing in grocery chains, summarizing the information on which chains have the most expensive pricing. The analysis breaks down the number of most expensive products by grocery chain, offering a broad overview of which chains have the highest occurence of being the most expensive. Future research could expand on this analysis by incorporating additional data on consumer preferences and regional factors.

References

Filipp, Jacob. 2024. "Hammer: A Toolkit for Quantitative Data Processing." https://jacobfilipp.com/hammer/.

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. *ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.

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.

Wickham, Hadley, Davis Vaughan, et al. 2023. *tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.