Grocery Prices by Vendor*

Sophia Brothers

November 13, 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

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.

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

2 Data

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.

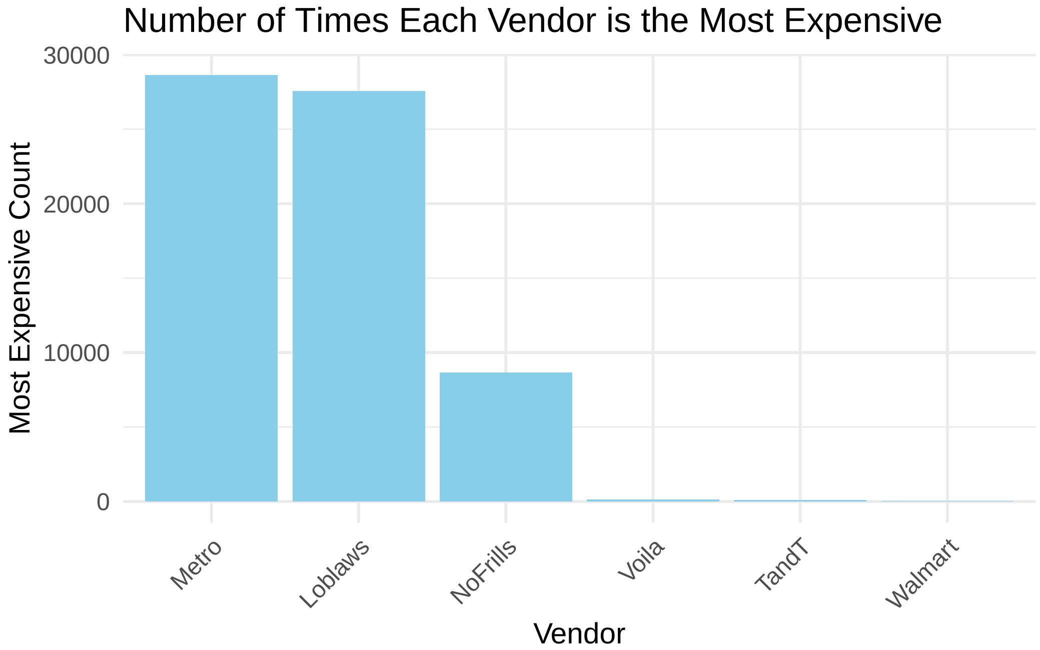

Figure 1: This bar graph shows the number of times a product was found to be the most expensive at a specific vendor. Products must be cross-carried by three vendors to count.

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

write discussion here

4.2 Correlation vs. Causation

While we observe differences in pricing across vendors, these differences do not imply causative factors without further experimental controls. Vendors may set prices based on external market trends or supply chain factors not captured in this dataset.

4.3 Missing Data

Missing data is a common challenge in grocery datasets. 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. Furthermore, online-only data may exclude regional in-store promotions, limiting the generalizability of our findings.

5 Conclusion

Future research could expand on this analysis by incorporating additional data on consumer preferences and regional factors.

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