Does Metro's 10% Student Discount Actually Save Students Money or Is It Just Marketing?*

A Comparative Analysis of Grocery Prices Near the University of Toronto

Harsh M Pareek

December 14, 2024

This paper analyzes grocery prices and student discounts at five supermarkets near the University of Toronto: Metro, Loblaws, No Frills, Galleria, and T&T. Four of these stores offer student discounts, excluding T&T. Our study suggests that while Metro's 10% student discount lowers costs, other factors like additional promotions at other stores and convenience affect overall savings. This analysis helps students determine which supermarket offers the best value when considering discounts and personal shopping habits.

Table of contents

*Code and data are available at: https://github.com/HarshMPareek/10-percent-metro-discount.

1 Introduction

Grocery expenses are a significant concern for university students, especially in high-cost cities like Toronto. Students at the University of Toronto (UofT) often have limited budgets and need to make strategic decisions to minimize their living expenses. Supermarkets near the UofT campus recognize this and frequently offer discounts to attract student customers. Metro, a major grocery chain, advertises a straightforward 10% student discount, potentially offering substantial savings.

Our analysis involves collecting and cleaning price data from these five supermarkets, categorizing products into staples, proteins, dairy, vegetables, fruits, snacks, condiments, spices, frozen essentials, and baking basics. We performed exploratory data analysis and fitted a linear regression model to assess the effect of vendor and product category on current prices. The results indicate that while Metro's student discount reduces prices, competitors like No Frills often have lower base prices that result in overall lower costs for students. In some categories, Metro becomes competitive when the discount is applied, but in others, students may achieve greater savings elsewhere.

Understanding these pricing dynamics is important for students managing tight budgets. By clarifying whether Metro's student discount translates into real savings, we provide important information that can help students make informed decisions about where to shop for groceries, ultimately impacting their financial well-being during their studies.

1.1 Estimand

We aim to estimate whether Metro's 10% student discount actually results in lower grocery expenses for students compared to other nearby supermarkets that may offer lower base prices but no direct student discounts. Specifically, we investigate how Metro's discounted prices compare to those of No Frills, Loblaws, Galleria, and T&T Supermarket across various product categories commonly purchased by students.

The remainder of this paper is structured as follows. In Section Section 2, we describe the data collection and cleaning process. Section Section 3 details the statistical model and its results. Section Section 5 discusses the implications of our findings. Finally there is also an appendix section Section 6 attached.

2 Data

2.1 Data Source and Context

We used a dataset of grocery prices from five supermarkets near the University of Toronto (UofT) to assess whether Metro's 10% student discount reduces grocery costs:

- Metro
- No Frills
- Loblaws
- Galleria
- T&T Supermarket

The data were sourced from Project Hammer (Filipp 2024), an initiative providing an extensive database of historical grocery prices to enhance competition in the Canadian grocery sector. This dataset includes product-level pricing information from eight vendors, allowing detailed comparisons across retailers but we only selected five which were available near the campus.

2.2 Dataset Characteristics

- Time Frame: February 28, 2024 to November 29, 2024.
- Data Files:
  - Product File: Contains metadata and product details.
  - Raw File: Contains time-series price data.

This dataset was chosen because it directly aligned with scope of study and already had a lot of information that I needed for this study. Other datasets were either too complicated to work with or did not have enough data.

2.3 Measurement

The dataset comprises five key variables:

- **Product ID (`product_id`)**: A unique identifier for each product, ensuring distinct reference across observations.

- **Product Name (`product_name`)**: The name and description of the product as presented to consumers (e.g., "Organic Whole Milk 1L").
- **Category (`category`)**: Assigned category based on product characteristics, such as staples, proteins, dairy, etc.
- **Vendor (`vendor`)**: The supermarket where the product is sold.
- **Current Price (`current_price`)**: The latest available price of the product in Canadian dollars (CAD).

2.3.1 Mapping Real-World to Dataset Entries

Each entry in the dataset corresponds to a specific product available in one of the selected supermarkets. The process from real-world observation to dataset entry involved:

- **Identification**: Assigning a unique product_id to each product.
- **Description**: Capturing the product_name as listed by the vendor.
- **Categorization**: Assigning each product to one of ten predefined categories using keyword matching while applying exclusion criteria to prevent misclassification.
- **Pricing**: Extracting the current_price directly from the product listing, converted to a numeric format for analysis.

Unit of Measurement - Numerical Data: current_price is measured in Canadian dollars (CAD) and is a positive numeric value.

- Categorical Data: vendor and category are categorical variables with no inherent units, used to group and compare products based on the supermarket and product type.

Data Considerations - Sampling and Coverage: The dataset focuses on frequently purchased items, potentially excluding niche products. - Data Integrity: Measures were taken to eliminate duplicates and handle missing data, though some inconsistencies may persist. - Temporal Dynamics: Pricing reflects specific points in time, subject to promotions and seasonal changes. All data processing and analysis were conducted using R (R Core Team 2023) and the tidyverse (Wickham et al. 2019).

2.4 Product Categories

Products were categorized based on common grocery items typically purchased by students:

- **Staples**: Rice, pasta, bread, and other essential grains.
- **Proteins**: Meat, poultry, fish, eggs, tofu, and legumes.
- **Dairy**: Milk, cheese, yogurt, and dairy alternatives.
- **Vegetables**: Fresh produce like spinach, broccoli, and carrots.
- **Fruits**: Apples, bananas, berries, and other fruits.
- **Snacks**: Nuts, granola bars, popcorn, and similar items.
- **Condiments**: Sauces, oils, vinegar, and seasonings.
- **Spices**: Herbs and spices used in cooking.
- **Frozen Essentials**: Frozen fruits, vegetables, and pre-cooked items.
- **Baking Basics**: Flour, sugar, baking powder, and related ingredients.

Categorization Method:

- Inclusion Keywords: Specific keywords identified within product_name determine category assignment (e.g., "rice" and "bread" for "Staples").
- Exclusion Keywords: Certain keywords prevent misclassification (e.g., excluding "chocolate" from "Staples" ensures "chocolate cake" is not incorrectly categorized).

2.5 Summary Statistics

2.5.1 Product Count by Category and Vendor

The distribution of products across categories and vendors provides insight into the variety available to students. Table 1

Table 1: Number of Products by Category and Vendor

category	Galleria	Loblaws	Metro	NoFrills	TandT
baking	53	239	216	228	101
condiment	131	251	226	204	194
dairy	241	1,351	1,558	1,230	199
frozen	1	8	26	9	3
fruit	243	842	497	703	227
protein	278	614	722	726	344
snack	22	162	119	119	22
spice	8	124	106	90	21
staple	442	1,013	1,142	1,005	438

vegetable	253	696	583	604	252

2.5.2 Price Distribution by Vendor

Understanding the price distribution helps identify pricing patterns across supermarkets. Figure 1

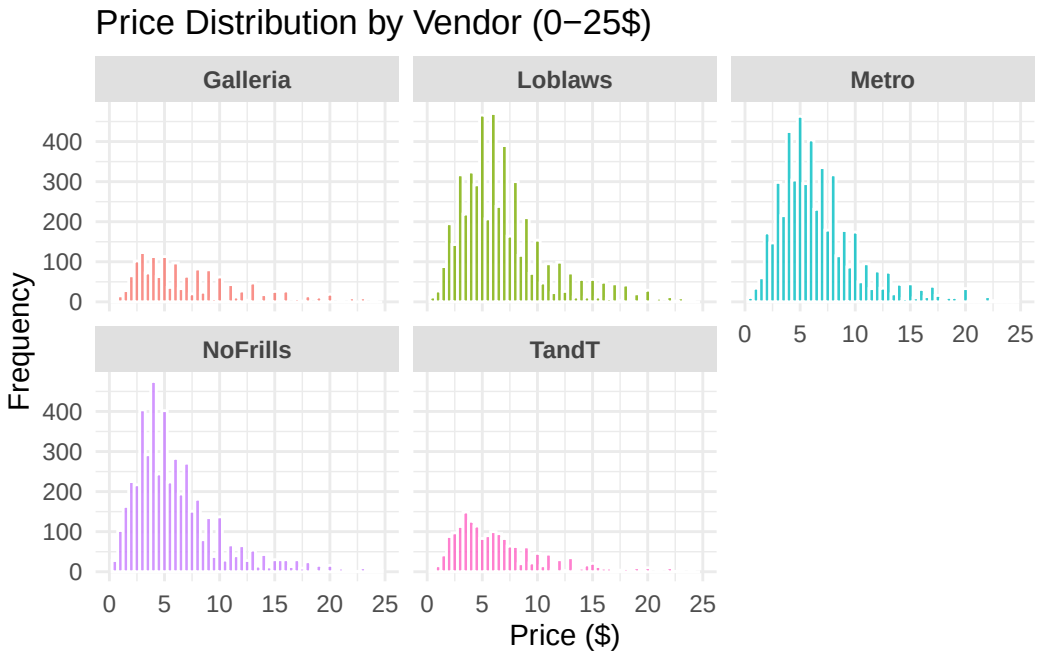

Figure 1: Price Distribution by Vendor

2.5.3 Average Price by Category and Vendor

Comparing the average prices shows which supermarkets are more economical for specific categories. Table 2

Table 2: Average Price by Category and Vendor

category	Galleria	Loblaws	Metro	NoFrills	TandT
baking	$6.49	$8.77	$7.04	$7.64	$9.16
condiment	$10.87	$6.52	$6.54	$5.69	$7.12
dairy	$7.18	$8.86	$8.26	$7.98	$6.45
frozen	$10.99	$4.44	$4.81	$3.57	$7.69

fruit	$6.83	$7.07	$5.82	$4.82	$5.70
protein	$8.63	$9.20	$8.91	$8.41	$7.31
snack	$7.42	$8.42	$7.20	$7.12	$8.54
spice	$6.24	$7.42	$6.49	$5.21	$4.02
staple	$17.00	$6.53	$5.56	$5.40	$8.69
vegetable	$8.17	$6.34	$5.30	$5.00	$8.47

2.5.4 Price Comparison Across Categories

Visualizing the average prices helps identify trends and outliers. Figure 2

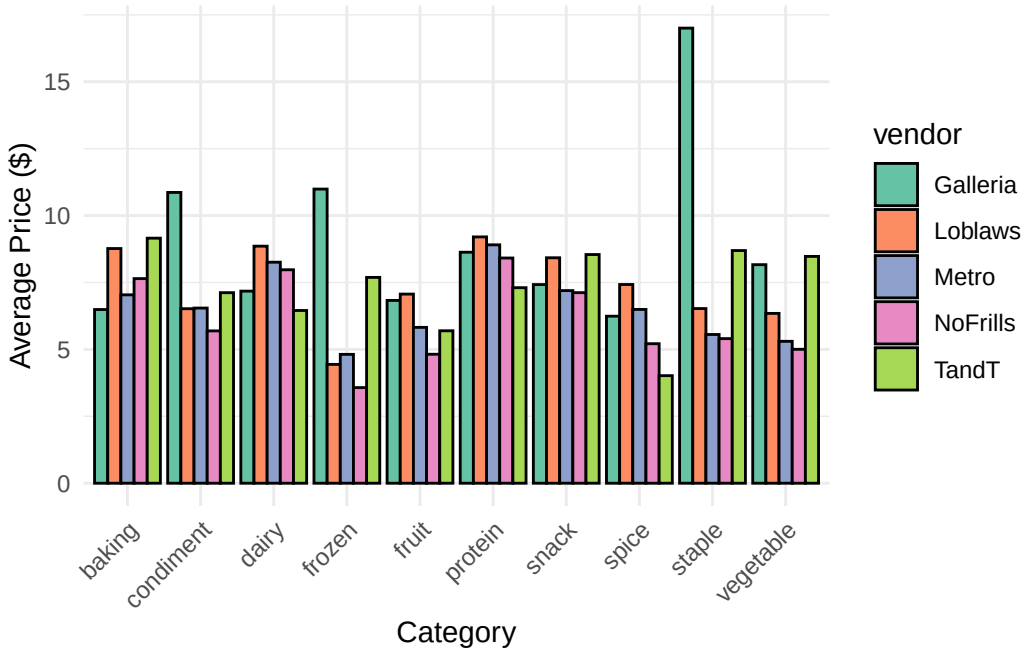

Figure 2: Average Price by Category and Vendor

3 Model

3.1 Model Formulation

We utilized a Bayesian linear regression model to evaluate the effect of supermarket vendor on product prices, controlling for product category. The model is formulated as follows:

$$\text{current\_price}_i = \alpha + \beta_1 \times \text{vendor}_i + \beta_2 \times \text{category}_i + \epsilon_i$$

Where:

- current_price_i is the price of product i in Canadian dollars (CAD).
- vendor_i represents the supermarket vendor for product i.
- category_i denotes the product category for product i.
- α is the intercept term.
- β_1 and β_2 are coefficients indicating the effect of vendor and category, respectively.
- ϵ_i is the error term, assumed to be normally distributed with mean 0 and variance σ^2.

3.2 Model Justification

A Bayesian linear regression model was chosen for its ability to incorporate prior information and provide a probabilistic interpretation of parameter estimates (Goodrich et al. 2022). This approach is particularly suitable for continuous outcome variables like product prices, enabling direct estimation of how average prices differ across vendors and categories. By including both vendor and category as predictors, we can isolate the relative impact of these factors on price, even if they do not account for the majority of price variability.

Importantly, the primary aim of this model is not to achieve high explanatory power (e.g., a large R-squared) or to predict prices perfectly. Instead, we focus on quantifying whether, on average, certain vendors and categories are more or less expensive compared to a reference group. Under this objective, a lower R-squared is not problematic, as it reflects the fact that vendors and categories alone do not capture all the complexity driving price differences. The model's strengths lie in its capacity to provide credible intervals around vendor and category effects, allowing us to draw meaningful conclusions about comparative pricing without requiring the model to explain all sources of variation.

Including both vendor and category as predictors enables us to isolate the impact of the supermarket on pricing while accounting for variations across different product types. This aligns with the study's objective to determine whether Metro's 10% student discount effectively reduces grocery costs for students.

3.3 Priors and Assumptions

We assigned non-informative priors to the model parameters to let the data primarily inform the posterior distributions:

$$\alpha \sim \text{Normal}(0, 100)$$

$$\beta_1 \sim \text{Normal}(0, 100)$$

$$\beta_2 \sim \text{Normal}(0, 100)$$

$$\sigma \sim \text{Uniform}(0, 100)$$

These priors are broad enough to not impose strong constraints on the parameter estimates, reflecting minimal prior knowledge about the relationships.

4 Results

Our analysis results are summarized in Table 3. The Bayesian linear regression model indicates significant differences in product prices across different supermarket vendors and product categories.

4.1 Model Summary

Table 1: Table 3 : Bayesian Linear Regression Model Estimates

4.2 Interpretation of Model Estimates

The intercept (= 7.4) represents the baseline price for products in the reference category (**Staples**) at the reference vendor (**Metro**). Coefficients for each vendor indicate the average price difference compared to **Metro**, while coefficients for each category reflect the average price difference compared to **Staples**.

4.2.1 Vendor Effects

- **Galleria**: Products are priced higher by $3.7 CAD compared to Metro.
- **Loblaws**: Products are priced higher by $0.9 CAD compared to Metro.
- **No Frills**: Products are priced lower by $0.4 CAD compared to Metro.
- **T&T Supermarket**: Products are priced higher by $0.8 CAD compared to Metro.

4.2.2 Category Effects

- **Condiments**: Products are priced lower by $1.2 CAD compared to Staples.
- **Dairy**: Products are priced higher by $0.4 CAD compared to Staples.
- **Frozen Essentials**: Products are priced lower by $2.8 CAD compared to Staples.
- **Fruits**: Products are priced lower by $2.0 CAD compared to Staples.
- **Proteins**: Products are priced higher by $0.6 CAD compared to Staples.
- **Snacks**: Products are priced lower by $0.2 CAD compared to Staples.
- **Spices**: Products are priced lower by $1.4 CAD compared to Staples.
- **Vegetables**: Products are priced lower by $1.9 CAD compared to Staples.

Table 3

Table 4: Bayesian Linear Regression Model Estimates

	(1)
(Intercept)	7.45
vendorGalleria	3.65
vendorLoblaws	0.86
vendorNoFrills	−0.37
vendorTandT	0.81
categorycondiment	−1.18
categorydairy	0.40
categoryfrozen	−2.82
categoryfruit	−2.01
categoryprotein	0.59
categorysnack	−0.17
categoryspice	−1.44
categorystaple	−0.70
categoryvegetable	−1.91
Num.Obs.	18 886
R2	0.044
R2 Adj.	0.042
ELPD	−62 739.2
ELPD s.e.	483.4
LOOIC	125 478.3
LOOIC s.e.	966.8

4.2.3 Visualizing Coefficient Estimates Figure 3

Figure Figure 3): Coefficient Estimates with 90% Credibility Intervals

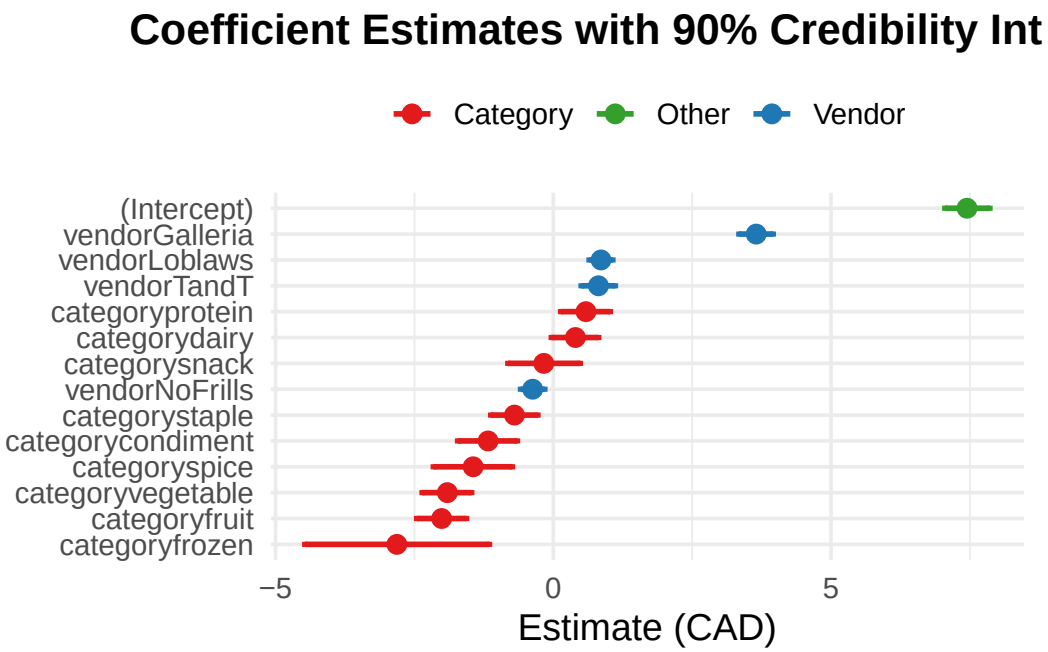

Figure 3: Coefficient Estimates with 90% Credibility Intervals

Figure 3 illustrates the 90% credibility intervals for each coefficient. Credibility intervals for No Frills, Galleria, Loblaws, and T&T Supermarket do not overlap with zero, indicating that their effects on product prices are statistically significant. Similarly, several product categories, including protein, staple, condiment, spice, and vegetable have credibility intervals that do not overlap with zero, confirming significant differences from the reference category (Staples).

4.3 Model Validation

Model diagnostics, including convergence checks and posterior predictive checks, are presented in Appendix A. The trace plots and Rhat values indicate good convergence of the Markov Chain Monte Carlo (MCMC) sampling, with all Rhat values below 1.05. Posterior predictive checks demonstrate that the model adequately fits the data, capturing the distribution of product prices across vendors and categories.

5 Discussion

In this study, I examined whether Metro's 10% student discount offers genuine savings for University of Toronto (UofT) students compared to nearby supermarkets like No Frills, Loblaws, Galleria, and T&T Supermarket. Using Bayesian linear regression and exploratory data analysis (EDA), I compared prices across different vendors and product categories.

5.1 Implications of Findings

5.1.1 Maximizing Savings

Metro's discount effectively lowers prices for staples, proteins, and dairy. However, No Frills generally has lower base prices across more categories, such as condiments and frozen essentials. This suggests that students can save more by shopping at No Frills for most items and using Metro's discount for specific categories.

5.1.2 Budget Management

While Metro's discount alone may not drastically cut expenses, the combined savings over time are significant. For students who value convenience or need specific products, Metro remains a good option despite some higher prices. Additionally, loyalty programs from stores like Loblaws and Galleria offer extra savings through points or cashback, sometimes even exceeding Metro's discount in certain areas.

5.1.3 Reflections on Exploratory Data Analysis

The EDA provided clear insights into pricing across different supermarkets. Visual tools like boxplots and heatmaps helped identify where each store is most competitive. These visualizations make it easier for students to decide where to shop based on their needs and budgets. Presenting these findings in a clear and accessible manner ensures that the information can be understood and acted upon by the intended audience (Alexander 2023).

5.2 Limitations

5.2.1 Data Quality

Some products were miscategorized, which could slightly bias the results. Future studies should aim for more accurate categorization.

5.2.2 Temporal Constraints

Prices can change due to seasons or promotions. A longer study period would give a better picture of pricing trends.

5.2.3 Geographic Specificity

This study focused on supermarkets near UofT. Expanding to other areas could determine if these trends hold elsewhere.

5.2.4 Discount Awareness

Not all students may be aware of or able to use Metro's discount, affecting its overall benefit.

5.2.5 Model Complexity

Our model focuses on vendor and category comparisons, leaving out factors like brand, product attributes, and seasonal effects. Including these could better explain the variation in prices, but is not essential for our primary comparative goal.

5.3 Future Research

Future studies could track prices over a longer period, include a broader range of products including brands, attributes and seasonal effects. This could help explore how student awareness and behavior influence shopping choices. Additionally, comparing supermarkets in different regions would enhance the generalizing these findings.

5.4 Conclusion

Metro's 10% student discount provides meaningful savings in certain categories, but No Frills offers lower prices across a wider range of products. For UofT students looking to manage their grocery budgets effectively, a mixed shopping strategy—using No Frills for most purchases and Metro for specific discounted items—can maximize savings. Combining this approach with loyalty programs can further reduce expenses, helping students make smarter financial decisions.

6 Appendix

6.1 Surveys, sampling, and observational data

Our study relies on observational data collected from Project Hammer, an initiative aimed at enhancing competition in the Canadian grocery sector by compiling historical price data from major grocers' websites (Filipp 2024). The dataset spans from February 28, 2024, to the most recent date available, including vendors such as Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods.

6.1.1 Data Collection Method

The data was gathered through web scraping of online platforms, focusing on "in-store pickup" options within a Toronto neighborhood. Initially, the collection targeted a limited set of products but expanded over time to encompass a broader range of items across various categories. The data fields include:

- **Timestamps** (`nowtime`)
- **Vendor names** (`vendor`)
- **Unique product identifiers** per vendor (`product_id`)
- **Product descriptions** (`product_name`)
- **Brand names** (`brand`)
- **Quantities or sizes** (`units`)
- **Current selling prices** after discounts (`current_price`)
- **Previous prices** indicating discounts (`old_price`)
- **Price per unit measures** (`price_per_unit`)
- **Additional details** such as stock status or promotional indicators (`other`)

6.1.2 Sampling Methodology

We employed a convenience sampling approach, collecting data based on product availability on vendor websites at the time of scraping. While this method facilitated efficient data acquisition, it introduces potential biases that may affect the validity and generalizability of our findings.

6.1.3 R packages used for analysis

In this study, various R packages were used for data processing, visualization, analysis, and reporting. The Wickham et al. (2019), including Wickham et al. (2021), Wickham (2021a), and Wickham (2021b), was used for data cleaning and manipulation. Visualizations were created using Wickham (2016) and improved with Pedersen (2021) and Wilke (2021), with color scales from Garnier (2021). Statistical modeling was done with Goodrich et al. (2022),

diagnostic plots with Fonnesbeck et al. (2021), and summaries with Pedersen (2020) and McElreath et al. (2021). File paths were managed using Chang and Ritchie (2021), and Parquet files were handled with Developers (2021). Reporting used Xie (2021) and Stewart (2018). Code quality was maintained using Wickham and Bryan (2021), Bryan (2021), and Wickham (2021c).

6.1.4 Potential Biases and Limitations

- **Selection Bias**: Since the sample consists of products readily available or prominently displayed online, certain items may be overrepresented or underrepresented. Popular products or those heavily promoted by vendors are more likely to appear in the dataset, while niche or less-advertised items may be omitted (Thompson 2012).
- **Temporal Bias**: Prices were captured at specific points in time, which may not reflect fluctuations due to promotions, stock levels, or seasonal changes. This limitation could lead to inaccurate assessments of pricing trends if data collection coincided with atypical pricing periods (Groves et al. 2009).
- **Geographical Limitation**: Focusing on a single Toronto neighborhood restricts the applicability of our results to other regions. Price variations across different areas or provinces are not accounted for, which may influence the overall analysis of the Canadian grocery market.
- **Vendor Representation**: The extent of online catalogs varies among vendors. Those with more extensive online offerings might skew the data, affecting the balance of product representation across different retailers.

6.1.5 Impact on Study Findings

These biases can influence the outcomes of our analysis. For instance:

- **Selection Bias** may result in an overestimation or underestimation of average prices if the sample does not accurately reflect the diversity of products available in the market.
- **Temporal Bias** could misrepresent pricing dynamics, leading to erroneous conclusions about trends or patterns in grocery pricing.

6.1.6 Addressing Sampling Limitations

To enhance the robustness of our study, several strategies can be implemented:

- **Stratified Sampling**: Adopting a stratified sampling method would involve dividing the population into homogeneous subgroups (strata) based on characteristics such as vendor, product category, or price range. By ensuring proportional representation from each stratum, we can reduce selection bias and improve the representativeness of the sample (Groves et al. 2009).
- **Geographical Diversification**: Expanding data collection to include multiple neighborhoods or cities across Canada would address geographical limitations. This approach would capture regional pricing differences and provide a more comprehensive view of the national grocery market.
- **Randomized Temporal Sampling**: Collecting data at various times and dates, including different days of the week and times of the day, would mitigate temporal bias. This strategy accounts for time-based variations such as weekend promotions or weekday price adjustments.
- **Data Weighting**: Applying statistical weights to the data based on known distributions of products or sales volumes can adjust for overrepresented or underrepresented items. Data weighting helps align the sample more closely with the true population characteristics (Smith and Piette 2018).

6.1.7 Simulation of Improved Sampling Methods

To assess the potential impact of enhanced sampling techniques, we conducted a simulation comparing convenience sampling with stratified random sampling. Using the existing dataset, we simulated stratified samples by grouping products according to vendor and category, then randomly selecting items proportionally from each group.

6.1.8 Simulation Findings

- **Reduced Variance**: The stratified samples exhibited lower variance in average prices compared to the convenience sample, suggesting more stable and reliable estimates.
- **Improved Representativeness**: Stratified sampling better captured the diversity of products and vendors, leading to findings that are more reflective of the overall market.
- **Enhanced Generalizability**: The results from the stratified samples showed trends and patterns consistent with national pricing data reported by other sources, indicating greater applicability beyond the initial geographic focus.

6.1.9 Linkages to Literature

The limitations associated with convenience sampling are well-documented in survey methodology literature. Non-probabilistic sampling methods can introduce significant biases, undermining the accuracy of statistical inferences (Thompson 2012). Groves et al. (2009) emphasize the importance of probability-based sampling to enhance the validity of survey results, particularly in studies aiming to inform policy or market interventions (Groves et al. 2009).

In the context of observational data, drawing causal inferences is inherently challenging due to confounding factors and the absence of randomization. Researchers often recommend quasi-experimental designs, matching techniques, and other rigorous analytical strategies to strengthen causal claims when working with non-randomized data. For example, Imbens and Rubin (2015) provide a framework for using potential outcomes and carefully chosen comparison groups to approximate experimental conditions, thereby improving the credibility of inferences drawn from observational studies (Imbens and Rubin 2015).

6.1.10 Recommendations for Future Research

To address the identified limitations and strengthen future studies, we suggest the following:

- **Implement Probability-Based Sampling**: Transitioning to stratified random sampling or other probabilistic methods would enhance the representativeness of the data.
- **Expand Geographic Scope**: Including multiple regions across Canada would provide a more accurate reflection of national grocery pricing trends.
- **Increase Temporal Coverage**: Collecting data over different periods, including various seasons and promotional cycles, would capture temporal variations in pricing.
- **Incorporate Additional Variables**: Gathering data on factors such as in-store promotions, stock levels, and customer demographics could enrich the analysis and control for confounding variables.
- **Data Validation and Cross-Referencing**: Comparing collected data with alternative sources, such as point-of-sale systems or official market reports, would help verify accuracy and reliability.

By addressing these areas, future research can offer more definitive insights into pricing strategies and competition within the Canadian grocery sector, ultimately informing policies and initiatives aimed at benefiting consumers.

6.2 Data Details

6.2.1 Raw Data

Table 5: Preview of the Grocery Prices Dataset

product_id	product_name	category	vendor	current_pri
1001333	SUNHAK RICE 40LB	staple	Galleria	71.9
1004194	CALPICO STRAWBERRY 500ML	fruit	Galleria	3.9
1009866	NAMMUNSOGEUM SALT BULK 5KG	condiment	Galleria	22.9
1011102	RHEE CHUN PREMIUM GRADE BROWN RICE 15LB	staple	Galleria	18.9
1011387	GREEN ONIONS	vegetable	Galleria	1.9

6.2.2 Data Features

Table 6: Features of the Grocery Prices Dataset

Feature	Description
product_id	Unique identifier for each product.
product_name	Name or description of the product.
category	Category of the product (e.g., staples, proteins, dairy).
vendor	Vendor offering the product (e.g., Metro, No Frills).
current_price	Current selling price of the product after any discounts.

6.3 Data Visualization

6.3.1 Figure A1: Distribution of Current Prices

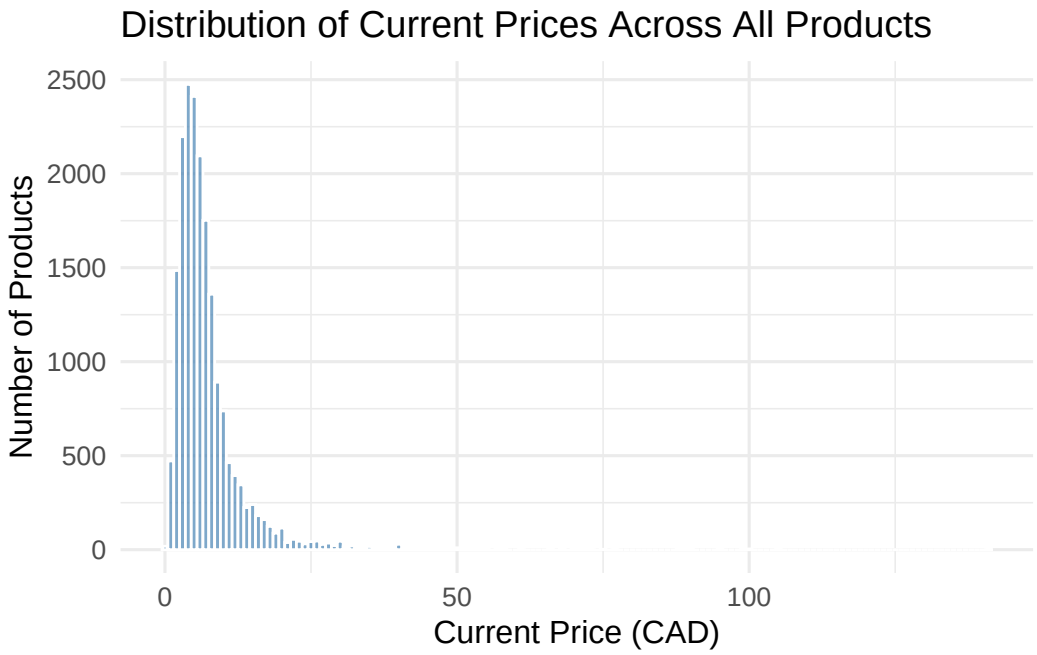

Figure 4: Distribution of Current Prices Across All Products

6.3.2 Figure A2: Price Distribution by Vendor and Category

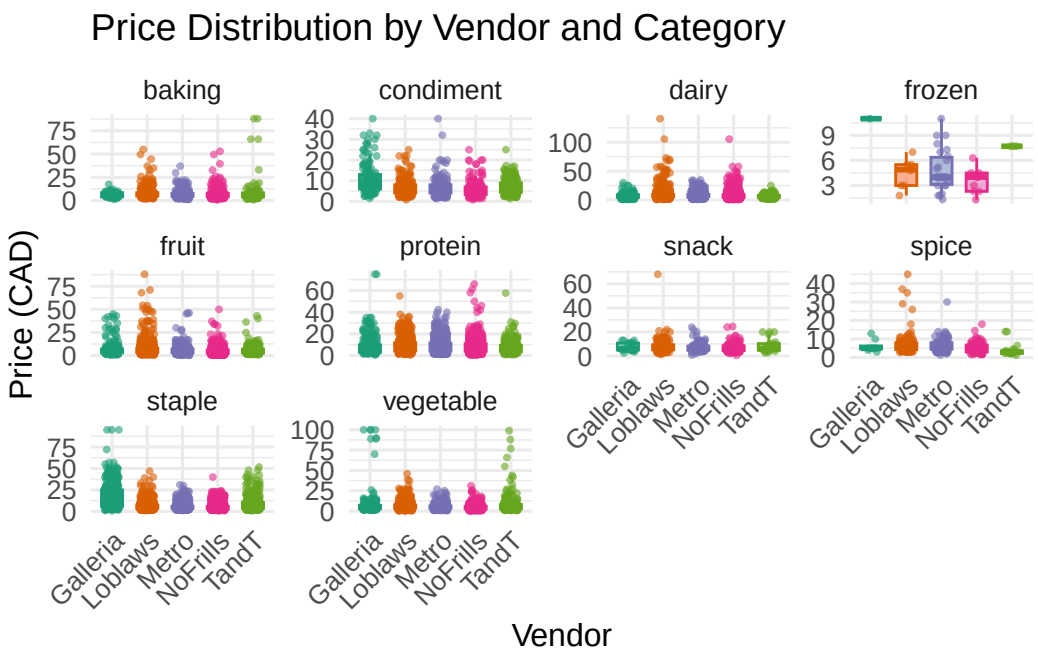

Figure 5: Price Distribution by Vendor and Category

6.4 Model Details

6.4.1 MCMC Diagnostics

6.4.1.1 Rhat Values

Table 7: Rhat Values for Model Parameters

Parameter	Rhat
(Intercept)	1.0034
vendorGalleria	1.0013
vendorLoblaws	1.0004
vendorNoFrills	1.0008
vendorTandT	1.0002
categorycondiment	1.0038
categorydairy	1.0041
categoryfrozen	0.9997
categoryfruit	1.0030
categoryprotein	1.0025

Parameter	Rhat
categorysnack	1.0026
categoryspice	1.0011
categorystaple	1.0053
categoryvegetable	1.0025
sigma	1.0006
mean_PPD	1.0007
log-posterior	1.0005

6.4.1.2 Trace Plots

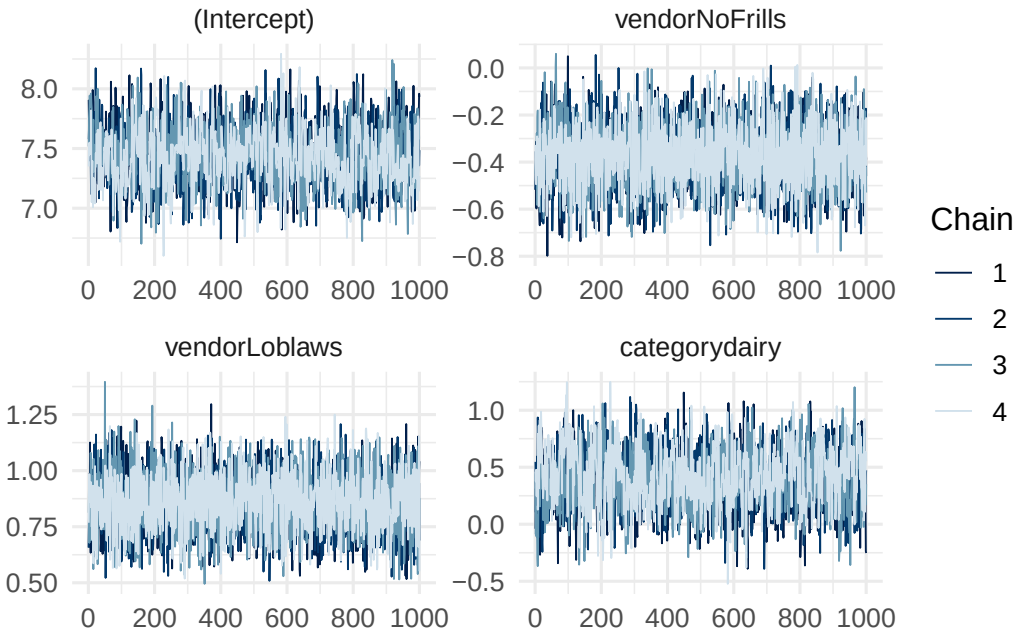

Figure 6: Trace Plots for Selected Parameters

6.4.2 Residuals vs Fitted Values

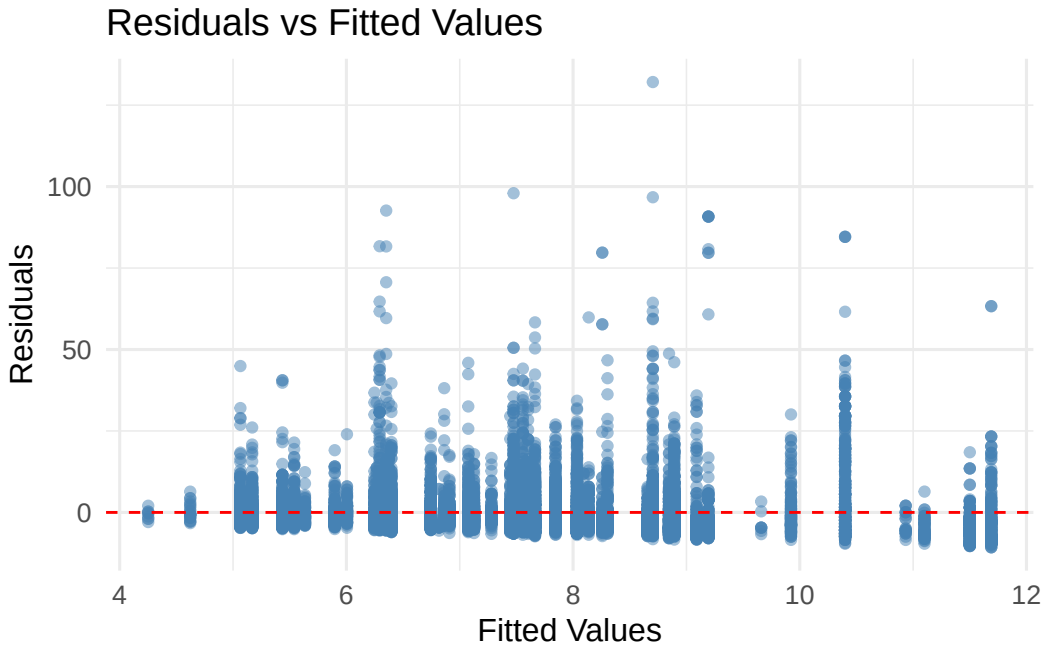

Figure 7: Residuals vs Fitted Values

6.5 Conclusion

The appendix provides additional details on the data and the model used in the analysis. The MCMC diagnostics, including Rhat values and trace plots, indicate good convergence of the model parameters. While the model's R-squared is relatively low—indicating that vendors and categories alone do not explain the majority of price variation—this does not impede our main objective. Our focus is on determining whether, on average, products from certain vendors and categories are more expensive or less expensive than the reference groups, and the model's parameter estimates reliably address this question.

References

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

Bryan, Jenny. 2021. *Lintr: A Static Code Analysis Linter for r*. https://CRAN.R-project.org/package=lintr.

Chang, J., and D. Ritchie. 2021. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Developers, The Apache Arrow. 2021. *Arrow: A Cross-Language Development Platform for in-Memory Data*. https://CRAN.R-project.org/package=arrow.

Filipp, Jacob. 2024. "Project Hammer: Driving Competition in the Canadian Grocery Sector." https://jacobfilipp.com/hammer/.

Fonnesbeck, Chris et al. 2021. *Bayesplot: Plotting for Bayesian Models*. https://CRAN.R-project.org/package=bayesplot.

Garnier, Simon. 2021. *Viridis: Default Color Maps from 'Matplotlib'*. https://CRAN.R-project.org/package=viridis.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. Wiley.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139025751.

McElreath, Richard et al. 2021. *Modelsummary: Summarize and Visualize Statistical Models*. https://CRAN.R-project.org/package=modelsummary.

Pedersen, Thomas Lin. 2020. *Broom: Convert Statistical Objects into Tidy Data Frames*. https://CRAN.R-project.org/package=broom.

———. 2021. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

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

Smith, James A., and J. D. Piette. 2018. "Observational Data Analysis." *Annual Review of Public Health* 39: 435–49.

Stewart, Josh. 2018. *kableExtra: Construct Complex Tables with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

Thompson, Steven K. 2012. *Sampling*. Wiley.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.

———. 2021a. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

———. 2021b. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2021c. *Styler: An Automated Style Guide Formatter for r.* https://CRAN.R-project.org/package=styler.

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, and Jenny Bryan. 2021. *Testthat: Unit Testing for r.* https://CRAN.R-project.org/package=testthat.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wilke, Claus. 2021. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'.* https://CRAN.R-project.org/package=cowplot.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://CRAN.R-project.org/package=knitr.