Predicting Grocery Price Trends Using Bayesian Models*

An analysis of price patterns across Canadian grocery vendors to identify the most beneficial vendor for consumers

Tina Kim

December 1, 2024

This study uses Bayesian statistical models to analyze grocery price trends across eight Canadian vendors by examining current and historical grocery price data. Our findings show that vendor-specific factors and product type significantly influence price changes. This research provides actionable insights for consumers and vendors looking to optimize their grocery purchasing and pricing strategies while adhering to the 2019 Canada's Food Guide, especially in light of the growing demand for affordable, sustainable food choices.

1 Introduction

Overview paragraph

The 2019 Canada's Food Guide (CFG) (Health Canada 2019) emphasizes the importance of balanced nutrition, advocating for a diet rich in vegetables, fruits, whole grains, and protein foods sourced from both plants and animals. It encourages reducing saturated fats, added sugars, and highly processed foods while recommending unsweetened, lower-fat milk or fortified plant-based alternatives. Notably, the 2019 CFG introduces greater flexibility compared to previous editions by integrating dairy products into the broader protein foods category, reflecting a more inclusive approach to nutrition.

However, adhering to the 2019 CFG has become increasingly challenging due to rising grocery costs. A basket of groceries that cost $1,000 five years ago now costs approximately $1,296 in 2024 (Royal Bank of Canada 2024). This escalation forces consumers to make more informed and difficult choices, often weighing the affordability of less healthy but cheaper options against the higher costs of healthier alternatives.

*Code and data are available at: https://github.com/thk421/Canadian-Grocery-Prices.git

To address this issue, our study explores grocery pricing trends across eight Canadian vendors: Metro, Save-On-Foods, Galleria, T&T Supermarket, Voilà, Walmart, No Frills, and Loblaws. By examining predictors such as product categories, vendor pricing, and temporal trends, this research provides critical insights into the dynamics of grocery pricing. Additionally, Bayesian statistical methods are employed to model future price changes, offering valuable tools for consumers and policymakers navigating the growing challenges of maintaining healthy, cost-effective diets.

Estimand paragraph

The key estimand of this study is the effect of vendor choice, product category, and temporal trends on grocery pricing across Canada. Specifically, we aim to estimate how much more or less consumers might pay for certain product categories when purchasing from different vendors, considering the ongoing trends in grocery pricing and the role that seasonal and economic shifts play. Through this analysis, we seek to estimate the probability of future price decreases or increases based on these factors.

Another key estimand is the lowest total cost of constructing a week's worth of meals using commonly available ingredients that adhere to the 2019 CFG, tailored for each vendor. All analysis is conducted using a standardized price measure to ensure consistency, comparability, and fairness across products, units, vendors, and contexts. Since prices are often reported in various units (e.g., kilograms, grams, liters, or milliliters), standardizing them to a common unit (such as grams, milliliters, or individual items) eliminates variability caused by differing measurement scales, enabling accurate and direct comparisons. This comparison extends beyond affordability to also consider quality, reliability, transparency, and appeal to calculate an overall score measure. Quality is assessed by comparing if the product is organic or not. Reliability is based on stock availability. Transparency is measured through quiet price decreases. Lastly, best seller or sale promotions are integrated for appeal. This composite metric provides a comprehensive rating for each vendor to offer tailored recommendations, helping consumers identify the vendor that best aligns with their specific needs and priorities.

Results paragraph

The analysis reveals that vendor choice plays a significant role in determining grocery prices, with large national vendors generally offering lower prices compared to local or regional options. Furthermore, we observe a clear seasonal trend in grocery pricing, with prices typically rising in the winter months and dropping during certain sales periods. The model also indicates that certain product categories, particularly organic and plant-based foods, are subject to higher price volatility, while staple items such as grains and protein foods remain relatively stable.

Why it matters paragraph

Understanding how pricing trends vary across different vendors and product categories is crucial for consumers seeking to optimize their grocery spending while adhering to the dietary guidelines outlined by Canada's Food Guide. This research not only informs consumer decision-making but also provides valuable insights for policymakers and vendors aiming to

address food accessibility and affordability in Canada. By modeling future price trends, we can better anticipate and prepare for fluctuations in grocery costs, helping people in Canada make healthier, more cost-effective food choices.

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2 describes the data used in the analysis, Section 3 details the Bayesian model and assumptions, Section 4 presents the results, and Section 5 offers a discussion of the findings and suggests areas for future research.

2 Data

2.1 Overview

The "Canadian Grocery Price Data" dataset was sourced from Project Hammer (Filipp 2024), and a statistical analysis was performed using the R programming language (R Core Team 2023). The primary objective of this analysis was to clean, explore, and visualize the data to identify which grocery vendor currently and in the future offers the most cost-effective options for individuals adhering to the 2019 CFG. Various R libraries were utilized throughout the process. The tidyverse library (Wickham et al. 2019) provides a comprehensive set of tools for data manipulation, cleaning, and visualization, enabling efficient handling of datasets. The lubridate library (Grolemund and Wickham 2011) simplifies working with date-time objects while arrow (Richardson et al. 2024) is used to handle large datasets efficiently, especially for columnar formats like Parquet. For Bayesian modeling, rstanarm (Goodrich et al. 2022) and brms (Bürkner 2017) offer powerful tools for fitting and interpreting multilevel models while bayesplot (J and T 2024) aids in the graphical exploration of Bayesian model results, enhancing the interpretation of complex models. A summary table showcasing 10 sample observations from the cleaned dataset is presented in Table 1. The table was divided into Part 1 and Part 2 for better readability, but the original dataset is formatted as one continuous table.

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

In this analysis, the phenomenon of interest is the cost of constructing a meal plan based on the 2019 CFG by going through grocery prices of different products from various vendors. The original Canadian Grocery Price Data from Project Hammer used in this analysis came in two files, one containing metadata and product details, and the other containing time-series price data. The files were merged into one dataset and select variables were chosen to construct the analysis data for this paper. Detailed cleaning steps for the analysis data can be seen in Section B.

Table 1: Sample Canadian Grocery Price Data

(a) Sample Canadian Grocery Price Data - Part 1

nowtime	current_price	vendor	product_name	price_decrease
2024-08-30 10:11:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-08-31 12:27:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-01 11:36:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-02 11:13:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-03 14:06:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-04 10:38:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-05 10:58:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-06 13:24:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-07 09:45:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-08 10:20:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA
2024-09-09 13:35:00	5.79	Loblaws	Egg Creations! Whole Eggs, Original	NA

(b) Sample Canadian Grocery Price Data - Part 2

stock_status	is_sale	is_best	is_organic	food_category	price_per_standard_unit
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79
in_stock	FALSE	FALSE	FALSE	Protein Foods	5.79

After the cleaning process, we are left with a dataset containing several attributes for our analysis such as vendor, product, price, and food category. Through these steps, we systematically translate complex real-world phenomena into structured, analyzable data that can be used to answer research questions about grocery pricing dynamics.

2.3 Outcome variables

Price per Unit - The primary outcome variable in our study is the price per unit of each product. This outcome variable captures the key phenomenon we are interested in: how much consumers are paying per unit of a product, after accounting for varying packaging sizes and vendor pricing strategies. It is calculated by dividing the current price by the unit size (in grams, milliliters, or items) for each product. This allows us to standardize prices across different products that may be sold in different quantities. The calculation of price per unit is crucial for comparing products sold in various units of measurement and understanding the underlying cost dynamics of each product.

Overall Vendor Score - Another outcome variable is the overall vendor score, which takes into account multiple criteria to evaluate vendors. The highest weight is put on affordability which is measured by the average price per unit of groceries from each vendor. Some additional factors taken into account are whether products are organic (quality), are in stock (reliability), are on sale or are best sellers (appeal), and if there was a quiet price decrease (transparency).

Future Price Trends - This outcome variable represents the predicted trajectory of grocery prices over time, estimated using a Bayesian statistical model. The model incorporates historical pricing data, vendor-specific trends, and product categories to generate probabilistic forecasts. By quantifying uncertainty, the model provides a nuanced view of potential future price fluctuations, helping consumers anticipate changes and make informed decisions.

2.4 Predictor variables

Temporal Information - Timestamp (nowtime): This refers to the time when the data was collected. It enables the identification of trends over time, such as the seasonal impact on stock or sales. This element is critical when applying Bayesian inference to predict future price trends. From Figure 3, we see that majority of the data was collected after October, which might bring some limitations that will be discussed further into this paper.

Vendor Characterisics - Vendor Name (vendor): There are 8 vendors in this analysis: Voila, Loblaws, NoFrills, Metro, Walmart, TandT, SaveonFoods, and Galleria. This variable is crucial since we are comparing results between these vendors. For instance, we can factor

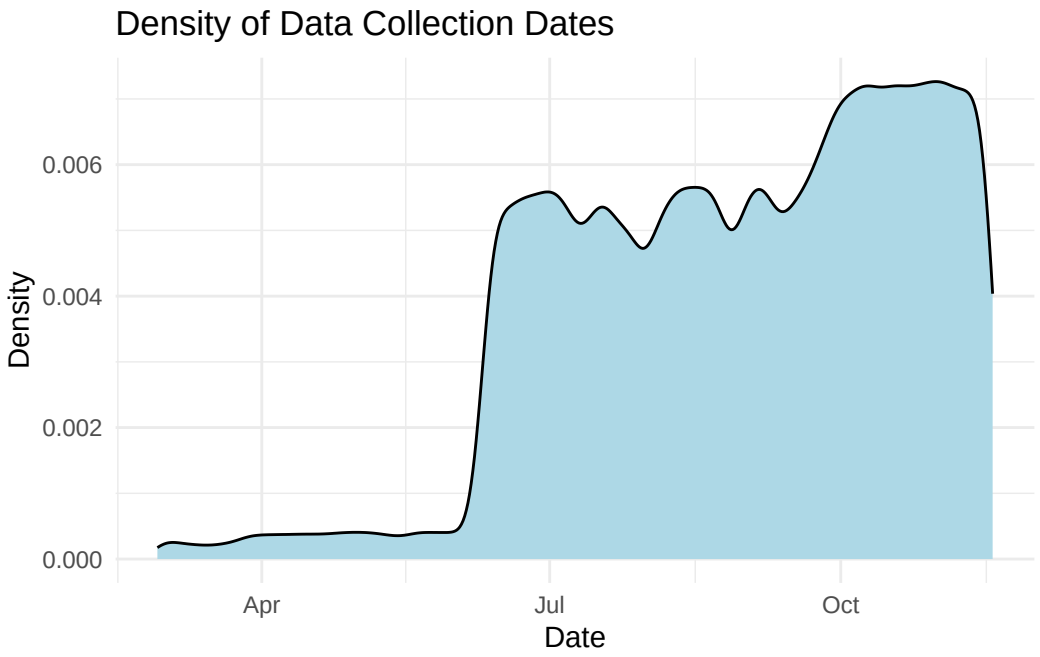

Figure 1: **?(caption)**

in the fact that Voila has approximately 4 times the number of products compared to Galleria as seen in the table below.

Table 2: Summary table for vendors

Vendor	Product Count
Voila	84834
Loblaws	82462
NoFrills	75425
Metro	73039
Walmart	56757
TandT	25248
SaveOnFoods	22312
Galleria	21960

Product Classification - Food Category (food_category): Classification as "Fruits & Vegetables," "Protein Foods," or "Grain Products" affects weighted cost calculations. Looking at the summary table below, we see that the overall number of protein foods far outweigh grain products.

Table 3: Summary table for food category

Food Category	Product Count
Protein Foods	261206
Fruits & Vegetables	114473
Grain Products	66358

Product Price Information - Current Price (current_price): This is the most recent price of a product (in CAD) used when measuring weighted costs and quiet decreases. The table below presents the price statistics for a product range, including the minimum price (0.2), maximum price (65.99), mean price (7.03), and standard deviation (5.20). The standard deviation of 5.20 is relatively high compared to the mean price of 7.03, suggesting significant variability in prices.

Table 4: Summary table for current price

Min Price	Max Price	Mean Price	Standard Deviation
0.2	65.99	7.02672	5.198359

- Quiet Price Decrease (price_decrease): Tracks subtle pricing adjustments by identifying instances where a product's price decreased without being labeled as a sale. This outcome sheds light on subtle pricing strategies vendors might employ to remain competitive without overt promotional activities. It provides insights into whether vendors prioritize transparent discounting (sales) or discreet price reductions. As seen in **?@fig-price-decrease**, Metro has the most number of products with decreased prices that were not on sale.

Stock and Promotional Indicators - Stock Status (stock_status): The proportion of products available (in stock) for each vendor reflects the reliability of a vendor in meeting consumer demands. Vendors with higher average stock levels may be perceived as more dependable by consumers. According to **?@fig-stock-status**, Loblaws has the most Fruits & Vegetables, Galleria has the most Grain Products, and Metro has the most Protein Foods.

- On Sale (is_sale): When a product is on sale and is explicitly labeled, we can use this predictor for identifying explicit promotions that can benefit consumers. In **?@fig-on-sale**, we measure the average number of products explicitly labeled as "on sale" per vendor, but we notice that only Voila has this label.

- Best Seller Status (is_best): The "Best Seller" designation is a key predictor for stock prioritization, promotional strategies, and pricing trends. As shown in **?@fig-best-seller**, Walmart is the only vendor that explicitly uses this label, a crucial observation to consider when developing our statistical model.

- Organic Product Status (is_organic): Whether a product is "Organic" serves as a critical predictor for those who want higher quality products while risking affordability. We compare the availability of high-quality products across vendors and categories by looking for product names that include "fresh" or "organic". In Figure 8, we see that Metro and Voila have the widest range of organic products. These products tend to be more pricey than normal products, which is a trade-off for people who want to invest in better quality over affordability.

3 Model

The model is expressed as:

$$\text{price\_per\_standard\_unit}_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

where:
- μ_i is the predicted mean price for product (i).
- σ^2 represents the residual variance.

The mean structure is defined as:

$$\mu_i = \beta_0 + \beta_1 \text{nowtime}_i + \beta_2 \text{vendor}_i + \beta_3 \text{food\_category}_i + \beta_4 \text{current\_price}_i + \epsilon_i$$

where:

- (_0): Intercept representing the baseline price.
- (nowtime_i): Continuous variable indicating the timestamp of data collection.
- (vendor_i): Categorical variable with levels for the eight vendors: Voila, Loblaws, NoFrills, Metro, Walmart, TandT, SaveonFoods, and Galleria.
- (food_category_i): Categorical variable indicating "Fruits & Vegetables," "Protein Foods," or "Grain Products."
- (current_price_i): Continuous variable for the most recent price of a product.
- (_i): Random error term capturing unexplained variability.

3.1 Priors

The Bayesian model incorporates weakly informative priors to reflect plausible ranges for parameters:
- $\beta_0 \sim \mathcal{N}(0, 10)$: Baseline prices are expected to fall within typical ranges.
- $\beta_1, \beta_2, \beta_3, \beta_4 \sim \mathcal{N}(0, 5)$: Moderate effect sizes are prioritized, but larger effects are not ruled out.
- $\sigma \sim$ Student-$t(3, 0, 10)$: Allows flexibility for variance in product pricing.

3.1.1 Model Justification

The selection of predictors included or excluded in the Bayesian model was guided by their relevance, interpretability, and alignment with the research objectives. Below is a detailed breakdown:

3.1.1.1 Included Predictors

1. **Temporal Information**:
    - `nowtime`: Captures potential seasonal trends or time-dependent price fluctuations, ensuring the model can account for temporal variability in pricing.
2. **Vendor Characteristics**:
    - `vendor`: Reflects inter-vendor differences, which are critical for evaluating affordability across vendors.
3. **Product Classification**:
    - `food_category`: Ensures the model incorporates the 2019 Canadian Food Guide's proportional weightings for different food categories, linking the analysis to dietary guidelines.
4. **Product Price Information**:
    - `current_price`: Serves as the baseline price, providing essential context for predicting future prices.

These features were chosen based on insights from exploratory data analysis, ensuring that the model is interpretable and reflects the underlying structure of the dataset.

3.1.1.2 Excluded Predictors

1. **Product Price Information**:

   - `price_decrease`: Highlights subtle pricing strategies, such as quiet reductions, which may not be explicitly marked as sales.

   Reason for Exclusion: There may be collinearity with predictors that already account for price change such as **current_price**, and may overfit due to being highly vendor-specific.

2. **Stock and Promotional Indicators**:

   - `stock_status`: Represents product availability and vendor reliability.
   - `is_sale`: Captures the effect of explicit promotions.
   - `is_best`: Evaluates the influence of "Best Seller" labels.

   Reason for Exclusion:
   These variables were excluded because they are highly vendor-specific and exhibit limited variability across vendors or over time. Including them could introduce vendor-specific biases and reduce the generalizability of the model across multiple vendors.

3. **Organic Product Status**:

   - `is_organic`: Represents consumer preferences for organic products.

   Reason for Exclusion:
   The impact of this variable is often reflected in other covariates, such as `food_category` or `vendor`. Including it could lead to redundancy and multicollinearity, complicating the model without providing significant additional insights.

This selection ensures the model is both parsimonious and robust, focusing on variables with the highest explanatory power while avoiding potential confounders or redundant predictors.

3.2 Assumptions

- The outcome variable, **price_per_standard_unit**, is normally distributed around the predicted mean.
- Predictors have linear effects on the outcome.
- Residual errors are independent and identically distributed.
- Vendor and product effects are hierarchical, reflecting their nested relationships.

3.3 Implementation and Validation

The model was implemented in R using the `brms` package. Validation was done by visually checking posterior predictive checks to assess the model's fit.

3.4 Limitations

The model assumes a consistent distribution of predictors across vendors and time, which may not hold in sparse data. Additionally, the linearity assumption may oversimplify complex relationships. Future iterations could explore interaction terms or non-linear effects to enhance flexibility.

3.5 Alternative Models Considered

Linear regression without Bayesian inference was considered but failed to incorporate uncertainty adequately. A Bayesian logistic model was evaluated for binary outcomes like **is_sale**, **is_best**, and **is_organic** but could not capture the continuous nature of the primary outcome. The chosen model balances complexity and interpretability while addressing hierarchical dependencies.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix: Sampling Methdology Overview - Weaknesses and Potential Improvements

B Appendix: Additional Data Details

B.1 Data Cleaning Steps

To prepare the dataset for analysis, we conducted comprehensive data cleaning. First, product and price datasets were merged based on product identifiers to create a unified dataset. Rows with missing critical information, such as unit or price, were removed. To standardize product categorization, keywords were defined for three primary food categories: fruits and vegetables, protein foods, and grain products. Products matching these keywords were assigned to their respective categories, while others were excluded.

Unit standardization was implemented to ensure consistency in price comparisons. Units were cleaned, converted to lowercase, and mapped to standardized units (e.g., grams, milliliters). A conversion table was applied to normalize all prices to a standard per-unit measure. Additional flags were created to capture stock status, sale indicators, best-seller labels, and organic products based on product descriptions.

Unnecessary columns, such as URLs, brands, and redundant identifiers, were dropped to streamline the dataset. Prices were then adjusted to calculate standardized price-per-unit values, enabling direct comparisons across products and vendors. Rows with unrecognized units or missing standardized price information were filtered out. Finally, duplicate entries were removed to ensure data integrity, and the cleaned dataset was sorted by vendor for further analysis.

C Appendix: Additional Model Details

C.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

Figure 2: **?(caption)**

C.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algorithm

Figure 3: **?(caption)**

References

Bürkner, Paul-Christian. 2017. "Brms: An r Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Filipp, Jacob. 2024. "Canadian Grocery Price Data." https://jacobfilipp.com/hammer/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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/.

Health Canada. 2019. *Canada's Food Guide*. https://food-guide.canada.ca.

J, Gabry, and Mahr T. 2024. "Bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

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.

Royal Bank of Canada. 2024. "Where Did My Money Go?: The Rising Cost of Groceries in Canada." Available at https://www.rbcroyalbank.com/en-ca/my-money-matters/debt-and-stress-relief/struggling-to-make-ends-meet/where-did-my-loonies-go-the-rising-cost-of-groceries-in-canada/.

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.