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 2, 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
2024-11-04 10:56:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-05 11:36:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-06 10:25:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-07 09:17:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-08 10:43:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-09 11:26:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-10 10:35:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-11 11:02:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-12 10:08:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-13 10:37:00	4.99	Loblaws	Coloured Carrots	NA
2024-11-14 09:30:00	4.99	Loblaws	Coloured Carrots	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	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99
in_stock	FALSE	FALSE	FALSE	Fruits & Vegetables	4.99

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 can be seen in Table 2, where the current price of the 'Portion Of Frozen Atlantic Salmon' product is standardized from 2.99 to 0.00299 to match the standardized unit, while the other products in the table remain the same because they are already standardized. 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.

Table 2: Summary of price per unit across product categories and vendors

product_name	current_price	price_per_standard_unit
Oikos 2% Greek Yogurt Plain 750 g	7.49	7.49000
Portion Of Frozen Atlantic Salmon	2.99	0.00299
Wu Xian Zhai Bbq Flavor Dired Tofu (108g)	2.52	2.52000
Gold Egg Omega 3 White Eggs Grade A Medium 6 Count	3.59	3.59000
Pierogie, Spinach/Broccoli/Cheddar	7.99	7.99000

Overall Vendor Score

- Another outcome variable is the overall vendor score as seen in Figure 1, 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

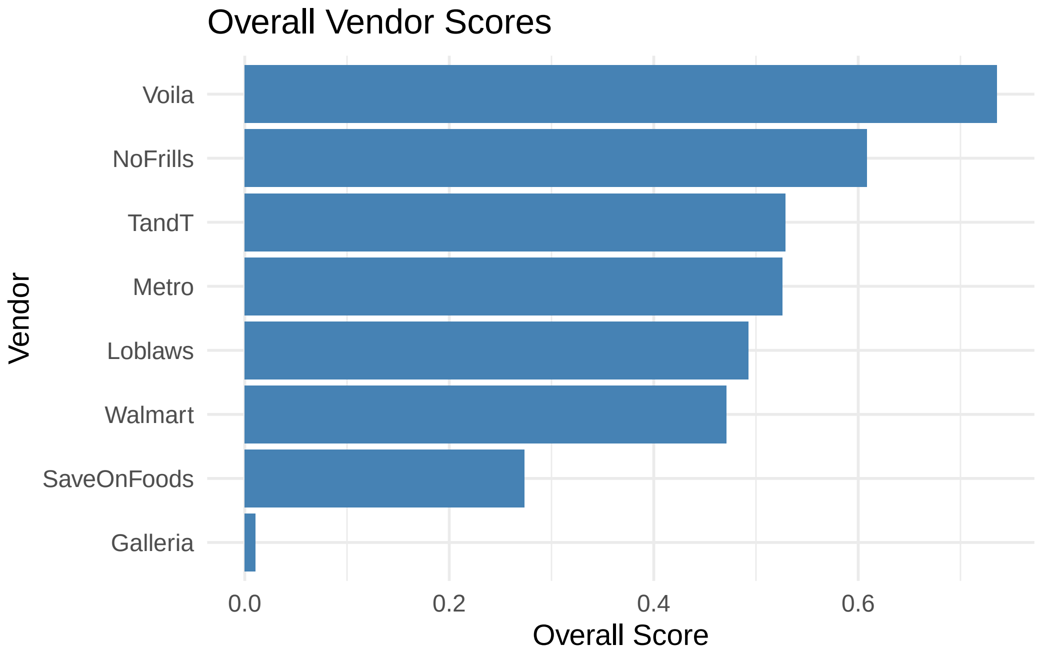

Figure 1: Summary of overall vendor scores

- This outcome variable as seen in Figure 2 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

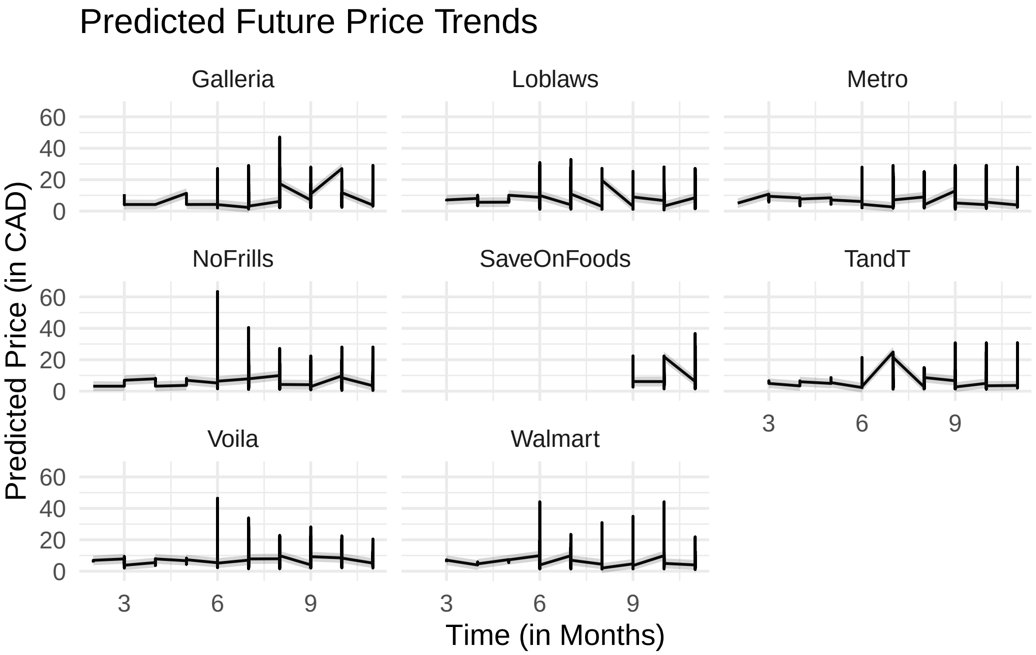

Figure 2: Predicted Future Price Trends

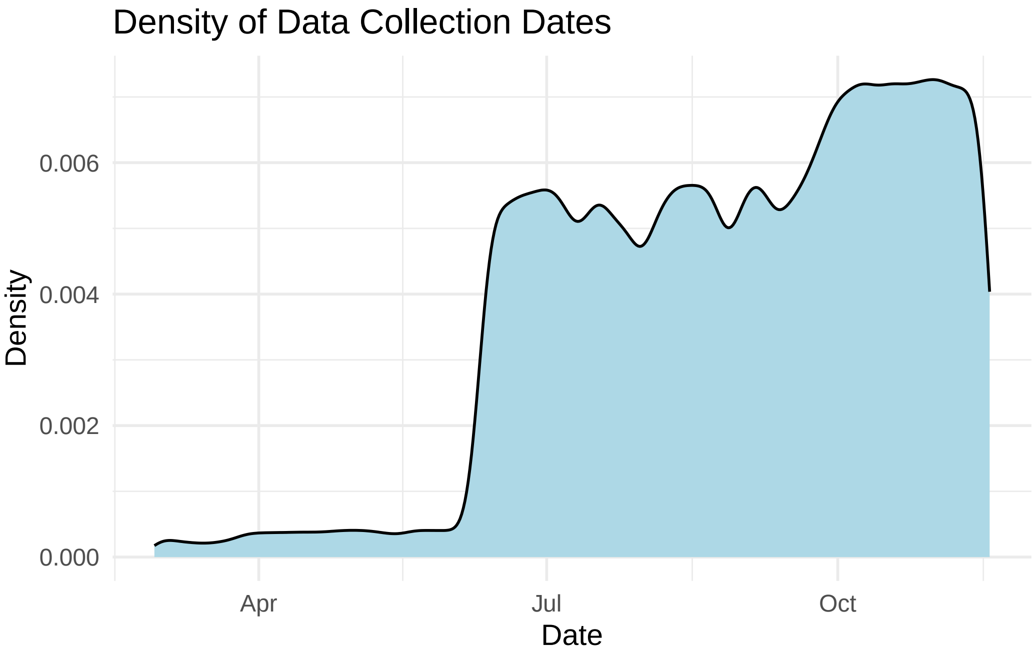

Figure 3: Nowtime Density Plot

are comparing results between these vendors. For instance, we can factor in the fact that Voila has approximately 4 times the number of products compared to Galleria as seen in the table below.

Table 3: 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 4: 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 5: 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 Figure 4, Metro has the most number of products with decreased prices that were not on sale.

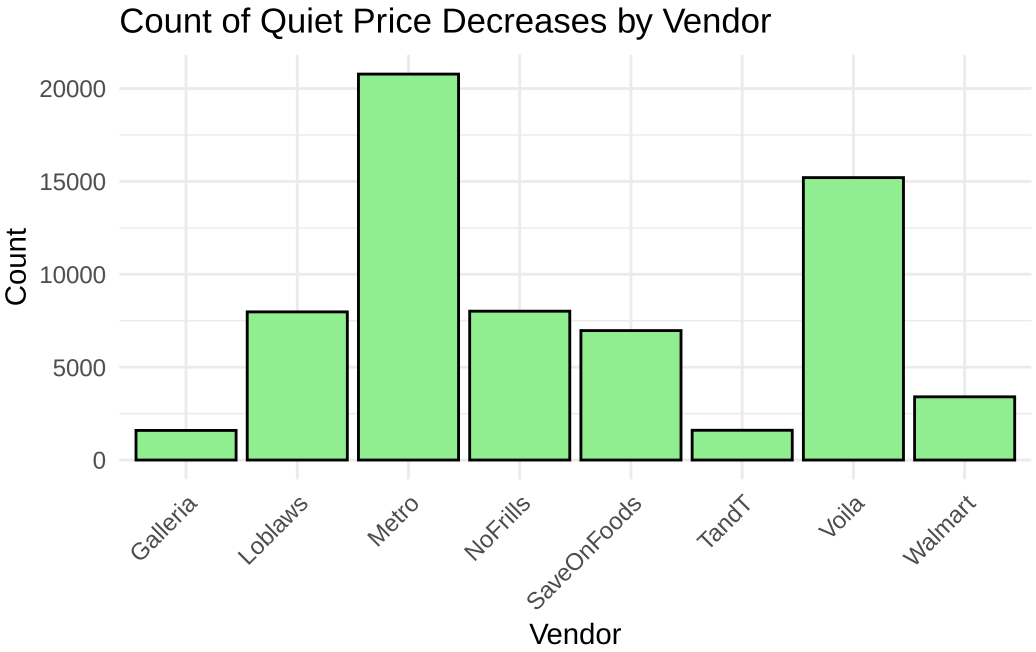

Figure 4: Count of quiet price decreases by vendor

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 Figure 5, 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 Figure 6, we

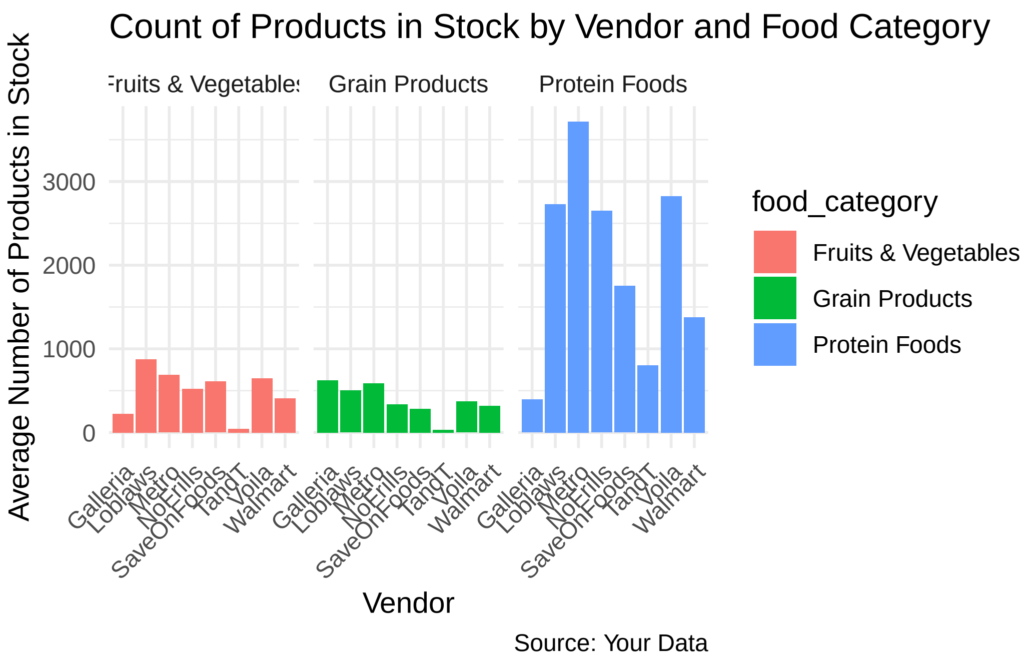

Figure 5: Count of products in stock by vendor and food category

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 Figure 7, 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.

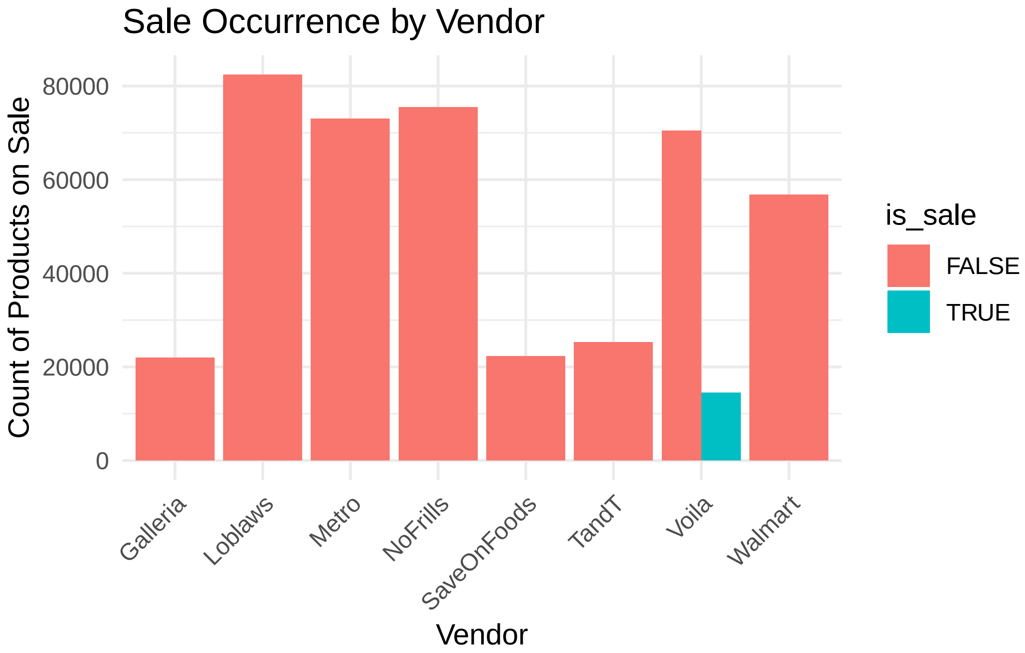

Figure 6: Sale occurrence by vendor

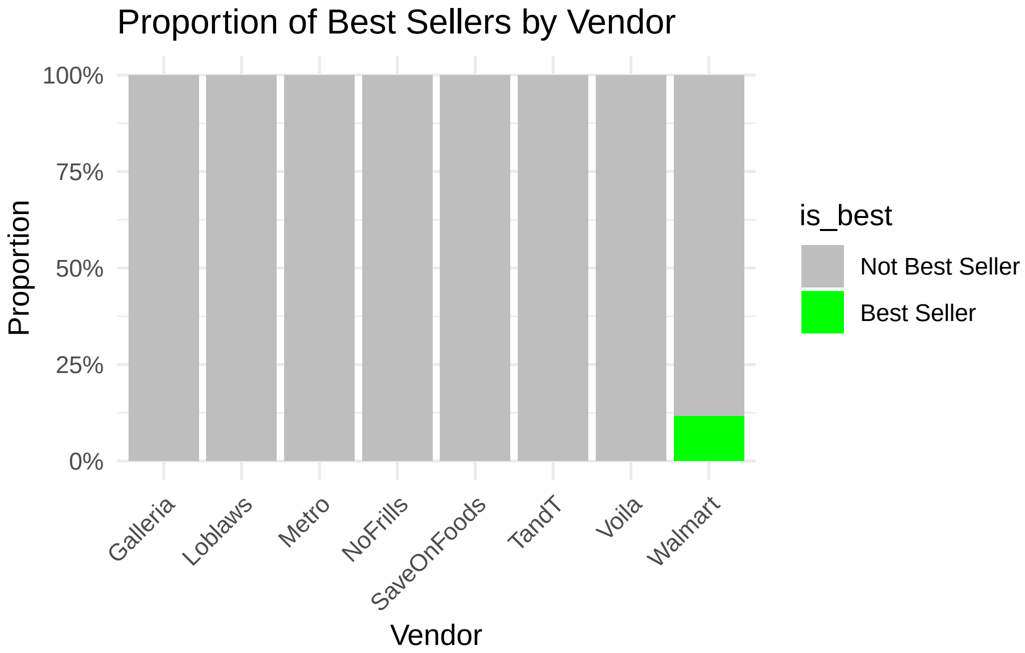

Figure 7: Proportion of best sellers by vendor

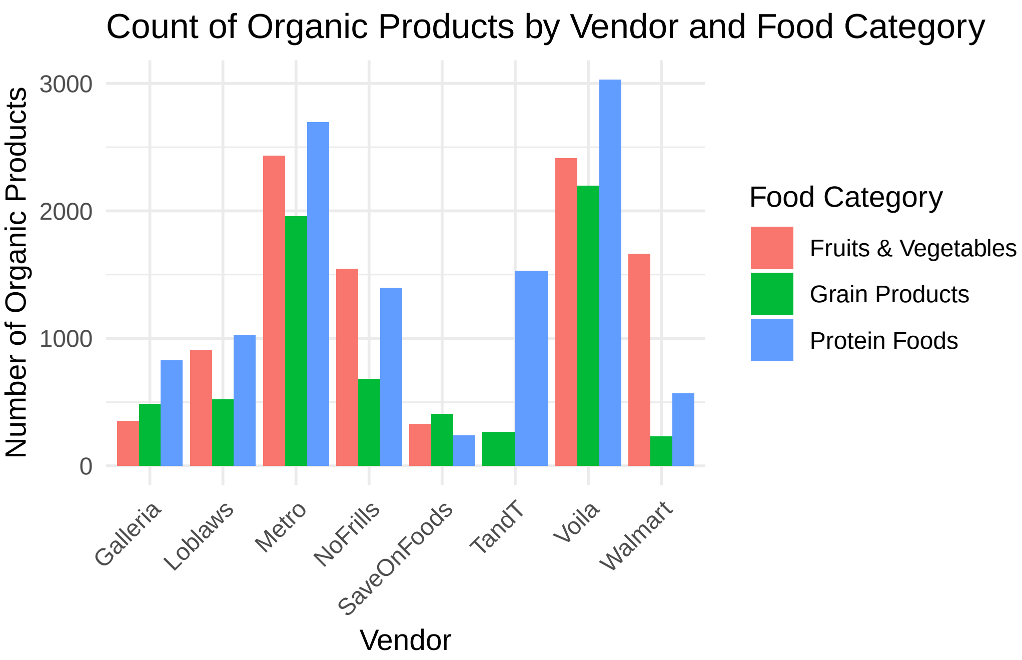

Figure 8: Count of organic products by vendor and food category

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:

- β_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.

- ϵ_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.
- $\sigma \sim \text{Student-}t(3, 0, 10)$: 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. A detailed breakdown can be seen in Section C.

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

The average weekly meal plan cost by vendor and food category in Figure 9 shows in order which vendors overall are the most cost-effective to the least cost-effective. By looking at the different colors that represent the food catgory, it can also be compared which food group is more cost-effective in which store.

The overall best vendor can be seen in Figure 1, and the predicted future price trends per vendor is shown in Figure 2.

Our Bayesian model results are summarized in Table 6. It provides the parameter estimates, predictive metrics, and significance levels to support the model. Further discussion for these results can be found in Section 5.2.

5 Discussion

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.

5.1 Paper Summary

what was done in this paper?

Table 6: Bayesian model for predicted future price trends summary

	(1)
b__current__price	0.955
b__food__categoryGrainProducts	0.065
b__food__categoryProteinFoods	0.178
b__vendorLoblaws	−0.013
b__vendorMetro	0.006
b__vendorNoFrills	−0.030
b__vendorSaveOnFoods	−0.033
b__vendorTandT	−0.167
b__vendorVoila	−0.044
b__vendorWalmart	−0.074
b__month	0.008
sigma	1.596
Num.Obs.	4947
R2	0.903
R2 Adj.	0.903
R2 Marg.	0.903
ICC	0.0
ELPD	−9361.9
ELPD s.e.	338.1
LOOIC	18 723.8
LOOIC s.e.	676.2
WAIC	18 729.2
RMSE	1.59
r2.adjusted.marginal	0.903769892933777

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

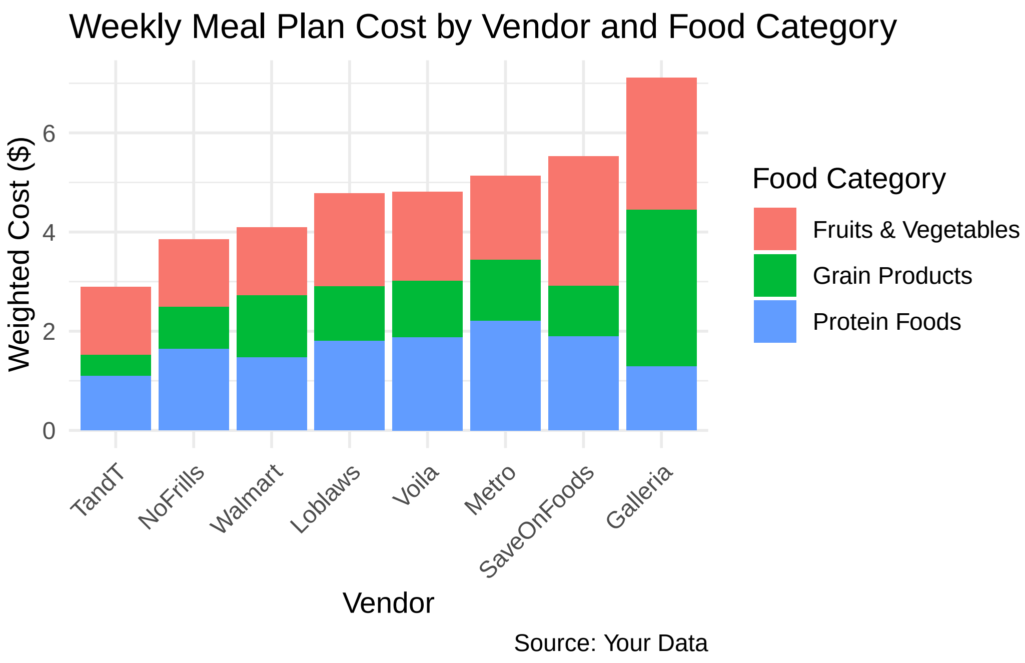

Figure 9: Average weekly meal plan cost by vendor and food category

5.2 Bayesian Model Explanation

5.2.1 Key Terms

- b_current_price (0.955): The regression coefficient for current_price. This indicates that for each unit increase in current_price, the future price is expected to increase by 0.955 units, holding other variables constant.
- b_food_categoryGrainProducts (0.065): Grain products are associated with a slight positive increase in the predicted future price relative to the baseline category (likely "Fruits & Vegetables").
- b_food_categoryProteinFoods (0.178): Protein foods are associated with a larger increase in the predicted future price compared to the baseline category.
- b_vendor (e.g., -0.013 for Loblaws):* These coefficients represent the effect of each vendor relative to the reference vendor. For instance, Loblaws has a slight negative impact on future prices compared to the baseline vendor.
- b_month (0.008): The slight positive effect of the month variable suggests minor upward trends in prices over time.

- sigma (1.596): This represents the residual standard deviation, indicating the variability of the outcome variable (future price) that is not explained by the predictors.
- R^2 (0.903): This high value indicates that 90.3% of the variance in the future price is explained by the predictors in the model.
- R^2 Adj. (0.903): Adjusted R^2 accounts for the number of predictors and indicates a similarly strong fit.
- R^2 Marg. (0.903): Marginal R^2 reflects the variance explained by fixed effects only.
- ICC (0.0): The Intraclass Correlation Coefficient suggests no substantial variance at the grouping level (e.g., product name in this case).

5.2.2 Predictive Metrics

- ELPD (-9361.9): The Expected Log Predictive Density is a measure of predictive accuracy, where higher (less negative) values indicate better predictions.
- ELPD s.e. (338.1): The standard error for ELPD, showing uncertainty around this estimate.
- LOOIC (18,723.8): The Leave-One-Out Information Criterion measures model fit; smaller values are better. It is often used for model comparison.
- LOOIC s.e. (676.2): The standard error of the LOOIC estimate.
- WAIC (18,729.2): The Watanabe-Akaike Information Criterion is another metric for model fit, similar to LOOIC. Again, smaller values are better.
- RMSE (1.59): The Root Mean Square Error quantifies the average error of the model's predictions, with lower values indicating better fit.

5.2.3 Bayesian Model Results Summary

The high R^2 and R^2 Marg. values suggest the model explains most of the variability in future prices. Amongst the predictors, **current_price** is the strongest predictor, with nearly a 1:1 relationship to future prices. Other variables like **food categories** and **vendors** have smaller but notable effects. Metrics like LOOIC and RMSE show that the model performs well, but further investigation into predictive accuracy and residual patterns could refine interpretations.

5.2.4 Predicted Future Price Trends

According to the predicted future price trends in Figure 2, Galleria exhibits relatively stable predicted price trends, with slight variations over time. The predicted prices stay consistent, without significant spikes, suggesting steady pricing strategies. Loblaws shows stable prices with slight fluctuations but no dramatic increases or decreases. This implies consistent pricing practices with minor seasonal or product-related variations. Metro displays a similar pattern of stability to Loblaws but with marginally larger variations. The periodicity of the vertical bars may indicate uncertainty in the predictions during certain months. NoFrills has a noticeable spike mid-period, followed by a return to stability. This indicates a potential pricing anomaly or predicted fluctuation in a specific month before returning to normal pricing levels. SaveOnFoods presents consistent prices with some minor upward shifts, suggesting a gradual increase or variability in certain product categories. However, sporadic trends can be seen where there is an abrupt jump between months, which could reflect noise in the data. T&T has a slight peak mid-period, followed by stabilization. This may reflect a seasonal effect or a one-time price fluctuation for specific products. Voila shows a small spike mid-period but generally maintains consistent prices. The predicted pricing pattern mirrors a temporary fluctuation, likely tied to vendor or product-specific dynamics. Walmart demonstrates the most volatility in its price trends, with noticeable spikes. This may highlight significant uncertainty in Walmart's pricing patterns or irregularities caused by promotional events or external market forces. The vertical lines represent the uncertainty intervals (credible intervals) of the Bayesian predictions. Wider intervals imply greater uncertainty about the predicted prices for that month. These intervals can stem from limited data availability for specific vendors or products, variability in historical pricing trends, or external factors influencing the model's predictive accuracy.

5.3 Overall Best Vendor

Choosing the best vendor based on the predicted future price trends depends on the specific priorities of consumers, such as affordability, price stability, and certainty. For affordability, NoFrills appears to have the lowest predicted prices overall, despite a noticeable spike in one period. Its consistent low prices outside this anomaly make it the most affordable option for price-sensitive consumers. Walmart also shows relatively low predicted prices but with more variability and uncertainty than NoFrills. For price stability, Loblaws and Metro exhibit the most stable price trends, with minimal fluctuations or uncertainty over time. For consumers prioritizing predictability in costs, these vendors may be the best choice. Galleria, Loblaws, and Metro have narrower credible intervals (vertical bars), indicating higher confidence in the model's predictions for their prices. This suggests these vendors may have more predictable pricing strategies, making them reliable options. On the other hand, Walmart and NoFrills have wider credible intervals, reflecting more uncertainty in their pricing. This could indicate a higher risk of price volatility. Thus, based on the bayesian model, we conclude that NoFrills emerges as the best option when considering affordability, as its prices are consistently low with only one significant spike. Loblaws or Metro could be the best option for consumers who value price stability and predictability over absolute cost.

Taking into account the Overall Vendor Score from Figure 1, we can now make a more comprehensive evaluation. We observe that Voila achieves the highest rating, while Galleria ranks the lowest. These scores were primarily driven by affordability, which accounted for 50% of the total rating. Other factors contributing to the score included reliability (20%), measured by the vendor's consistency in keeping products in stock; sale transparency (10%), reflecting whether the vendor avoids quiet price increases; product appeal (10%), with a focus on the availability of organic products; and promotional efforts (10%), which evaluated how clearly products were marketed as being on sale. Despite its higher prices compared to others, Voila has the highest overall vendor score. This high score suggests that it performs well across multiple factors like reliability, transparency, product appeal, and promotional efforts. These qualities could make it the best choice for consumers who prioritize a combination of affordability and additional factors like product quality, transparency, and sales efforts. Voila seems to provide strong value beyond just price, appealing to consumers seeking a well-rounded vendor. Although NoFrills has the second-best overall vendor score, it is the most affordable in terms of predicted prices. This makes it an excellent choice for consumers primarily focused on low cost. However, it ranks slightly lower in the overall score because factors like transparency and promotional efforts are less emphasized compared to Voila.

Given the importance of both affordability and vendor quality (including reliability, appeal, and transparency), Voila should be considered the best vendor. Despite higher prices compared to NoFrills, Voila offers a balanced combination of affordable pricing and superior service, making it the most well-rounded option overall. It also has narrower credible intervals in the predicted price trends compared to NoFrills, making it a more stable choice.

5.4 Weaknesses

While this study provides valuable insights into the dynamics of grocery pricing across different vendors, there are several limitations worth considering. First, the model's reliance on historical pricing data and vendor-specific trends may not fully capture the complexity of real-world pricing fluctuations, which are influenced by a variety of external factors such as market shifts, supply chain disruptions, or changes in consumer behavior. Furthermore, the data collected might not show accurate prediction of price tends throughout the year because most of the data collected as seen in Figure 3 is from after July. As a result, while the predictions offer valuable insights, they may not always align with actual price trends in the future.

Second, the analysis primarily focused on a few key variables (such as current price, vendor, and food category), and may have overlooked other potentially important factors. For example, changes in consumer preferences, seasonality, and macroeconomic conditions could have significant impacts on prices and may not have been fully integrated into the model. This represents a gap that could be explored in future studies.

Additionally, the dataset was limited to a specific time frame (2019-2020) and geographic region (Canada), meaning that the findings may not be generalizable to other countries or time periods. To improve the model's applicability, future research could expand the dataset to include a broader range of time periods, geographic regions, and product categories.

5.5 Next Steps

To address these weaknesses, there are several avenues for future research. First, incorporating more granular data, such as real-time pricing or additional product attributes (e.g., packaging size, brand loyalty, or consumer ratings), could provide a more comprehensive understanding of price dynamics and vendor performance. Expanding the model to include additional variables could lead to more accurate predictions and help better capture the real-world complexities of grocery pricing.

Second, future studies could explore the impact of external factors such as supply chain disruptions, inflation, or environmental sustainability initiatives on grocery prices. This would involve integrating external data sources to complement the existing dataset and help contextualize the price trends observed in this study.

Lastly, conducting similar analyses across different countries or regions would allow for a comparative assessment of pricing strategies and vendor performance, providing valuable insights for international policy makers, researchers, and consumers. This could also facilitate the development of more robust, generalized pricing models that can be applied across various markets.

A Appendix: Sampling Methdology Overview - Weaknesses and Potential Improvements

The Canadian Grocery Price Data sourced from Project Hammer provides valuable insights into grocery pricing trends across vendors, product categories, and regions. The data was obtained through a screen scrape of the website's user interface, which is why it lacks certain details that would typically be available through the internal APIs powering the grocers' websites. However, the sampling methodology underlying this dataset presents several weaknesses that could impact the robustness and generalizability of findings.

A.1 Weaknesses

1. The dataset was limited to a specific time frame (2019-2020) and geographic region (Canada), meaning that the findings may not be generalizable to other countries or time periods. Specifically, data is collected starting from February 28, 2024, but these were for smaller baskets of products. Majority of the usable data with a variety of products starts from July, potentially missing seasonal variations or short-term fluctuations in prices caused by external factors such as supply chain disruptions, weather events, or promotions.
2. The dataset relies on product identifiers and descriptions to standardize items across vendors. Inconsistent labeling or categorization may lead to mismatches or incomplete comparisons, particularly for highly variable products like produce.
3. There may also be missing data for certain days for certain vendors when specific extracts failed. Instances of missing unit information, product details, or pricing history reduce the completeness of the dataset and necessitate assumptions or imputation that may introduce errors.

A.2 Improvements

While the current sampling methodology provides valuable insights, several improvements can enhance the dataset's robustness, accuracy, and overall usefulness for analysis:

1. To capture a more comprehensive picture of pricing trends, data collection should span a longer time frame, ideally including multiple seasons and holidays. This would help account for seasonal variations, promotional periods, and other temporal factors influencing grocery prices. A broader time range would allow for a better understanding of price fluctuations throughout the year, enabling more accurate predictions and trend analysis.

2. Implement a more refined product matching system, potentially by using standardized product IDs or barcodes across vendors. Additionally, categorization should be standardized across all products to ensure accurate comparisons, particularly for perishable goods like produce. This would reduce mismatches and make comparisons between vendors more reliable, ensuring that price trends reflect genuine product similarities rather than differences in labeling or categorization.

3. Enhance the data collection process by improving the scraping method to reduce data loss. Using a more robust scraping framework that retries failed extractions or integrates better error handling can reduce instances of missing data. This would improve the completeness of the dataset, ensuring a more reliable foundation for analysis and reducing the need for imputation that might introduce errors. Where possible, supplement the screen-scraped data with data sourced from the internal APIs of the grocers' websites. Internal APIs tend to offer more accurate, comprehensive, and real-time data, including product details, stock levels, and historical price trends. This would increase the accuracy of the dataset and provide additional insights into product availability, pricing trends, and promotions, leading to more precise analyses.

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 Included Predictors in the Bayesian Model

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.

C.2 Excluded Predictors in the Bayesian Model

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.

C.3 Posterior predictive check

In Figure 10, we implement a posterior predictive check for the bayesian model of our analysis. This shows an overlay of the observed data density (y) and the posterior predictive densities (y_rep) drawn from the model. The dark line represents the observed data density (y), while the lighter lines represent the posterior predictive densities (y_rep) simulated from the model. The predictive densities align well with the observed density in the bulk of the distribution (around 0 to 10), indicating that the model is capturing the central tendency of the data reasonably well. There appears to be a slight discrepancy in the right tail of the distribution (beyond ~10). The observed density drops more steeply than most of the predictive densities, suggesting that the model may be slightly overestimating the probability of higher prices. Overall, the model fits the main part of the data distribution adequately and the slight deviation in the tail may be acceptable, but if predicting high prices accurately is critical, the model might need more refinement by adding more predictors.

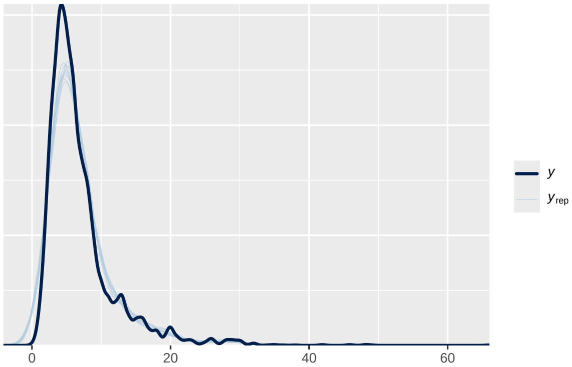

Figure 10: Examining how the model fits, and is affected by, the data

C.4 Diagnostics

Table 7 shows us the summary of the bayesian model. Since the Rhat value for all parameters are 1, and both Bulk_ESS and Tail_ESS for all parameters are larger than 100, we can conclude that there is excellent convergence in the bayesian model.

Table 7: Checking the convergence of the bayesian model using Rhat and ESS

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	0.19	0.16	-0.13	0.50	1	1588.19	1860.70
current_price	0.95	0.00	0.95	0.96	1	3694.60	2033.85
food_categoryGrainProducts	0.07	0.07	-0.08	0.21	1	4267.51	2201.70
food_categoryProteinFoods	0.18	0.06	0.06	0.29	1	3295.59	2037.78
vendorLoblaws	-0.01	0.12	-0.25	0.22	1	827.80	1510.94
vendorMetro	0.01	0.12	-0.23	0.24	1	903.17	1529.25
vendorNoFrills	-0.03	0.12	-0.27	0.21	1	874.52	1538.65
vendorSaveOnFoods	-0.03	0.16	-0.33	0.29	1	1383.23	1586.80
vendorTandT	-0.17	0.14	-0.44	0.11	1	1238.61	1904.83
vendorVoila	-0.05	0.12	-0.28	0.20	1	1004.14	1585.05
vendorWalmart	-0.08	0.12	-0.32	0.17	1	999.61	1693.41
month	0.01	0.01	-0.02	0.03	1	4264.70	2151.50

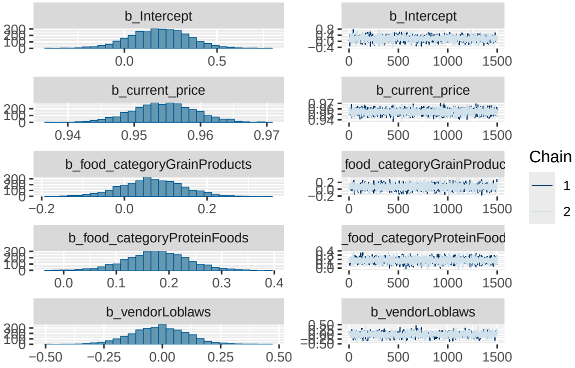

Figure 11: Checking the convergence of the bayesian model using trace plot

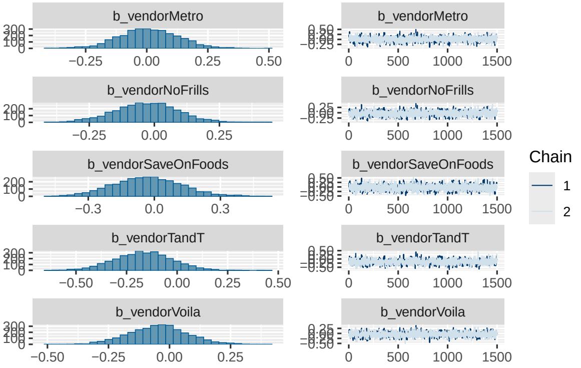

Figure 12: Checking the convergence of the bayesian model using trace plot

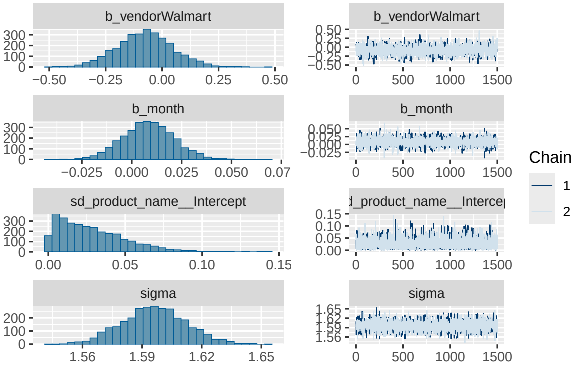

Figure 13: Checking the convergence of the bayesian model using trace plot

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.