My title*

My subtitle if needed

Ziheng Zhong

November 24, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

2 Data

This project is motivated and guided by Rohan Alexander and his book (Alexander 2023). Data used in this paper was cleaned, analyzed and modeled with the programming language R (R Core Team 2023). Also with support of additional packages in R: `readr` (**citeReadr?**), `ggplot2` (**citeGgplot2?**), `tidyverse` (**citeTidyverse?**), `dplyr` (**citeDplyr?**), `here` (**citeHere?**), `knitr` (**citeKnitr?**), `kableExtra` (**citeKableExtra?**), `rstanarm` (**citeRstanarm?**), `modelsummary` (**citeModelsummary?**),

*Code and data are available at: https://github.com/iJustinn/Canadian_Grocery_Analysis.git.

2.1 Source

The dataset used for this research was sourced from Hammer, a publicly available repository designed to provide pricing data from various retail chains. The dataset comprises detailed information on product attributes, such as product name, vendor, current price, old price, units available, and additional details like the month of certain price. The dataset's broader context is centered on consumer pricing trends across different vendors, helping to understand how factors like vendor type, previous pricing, and seasonal changes influence current prices. This information is pivotal for examining market behaviors and assessing pricing strategies in retail environments.

The dataset variables include several categorical and numerical components. Key variables such as 'vendor', which identifies the retailer (e.g., Walmart, Galleria), and 'product_name', which details the item, are crucial for understanding distribution channels. Variables like 'current_price' and 'old_price' offer insights into pricing dynamics, enabling analyses of price changes over time. Additionally, temporal variables such as 'month' allow for the identification of seasonal variations in pricing. Summary statistics, along with graphs for each of these variables, help illustrate their distributions and relationships. Graphs have been included to show the frequency distribution for vendors, trends in price changes, and comparisons between old and current prices. Relationships among variables like 'vendor' and 'current_price' have been highlighted to give a comprehensive view of the dataset.

There were other potential datasets available for this analysis, such as proprietary retail data sources or other consumer purchasing databases. However, they were not selected due to restrictions on data availability, licensing requirements, or lack of detail in specific product-level attributes. The Hammer dataset was chosen as it provides granular pricing data that is crucial for understanding product-level trends across multiple vendors. The dataset also allowed for the construction of new variables, such as the calculated 'price_per_unit', which enabled more meaningful comparisons between products. High-level data cleaning included handling missing values in the 'old_price' variable by replacing them with the 'current_price', ensuring the analysis was consistent and complete.

2.2 Measurement

The process of measurement in this research involves translating real-world phenomena into quantifiable data entries within the dataset. For instance, consumer purchasing behavior, influenced by factors such as vendor type, pricing history, and seasonal changes, has been captured through variables like 'vendor', 'current_price', and 'month'. The 'vendor' variable serves as an identifier of the source of the product, which helps to provide insights into how different retail environments may affect pricing strategies. The 'current_price' and 'old_price' variables reflect pricing trends and allow us to measure the temporal changes in consumer costs, thereby linking economic activities to specific data points.

The temporal dimension, captured through the 'month' variable, enables us to measure the impact of seasonal patterns on pricing. Constructed variables such as 'price_per_unit' provide a standardized measurement that allows comparisons across different products, accounting for varying package sizes or quantities. This transformation from abstract consumer behaviors and retail dynamics into structured data ensures that the analysis remains rooted in real-world market phenomena, thereby enabling a more nuanced understanding of pricing behavior and trends across different vendors.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

2.3.1 Current Price

The outcome variable in this study is current_price. This variable represents the price of the product at the time of data collection and is central to understanding the effects of various factors on product pricing. By modeling current_price, we aim to explore how factors such as historical pricing, vendor type, and seasonal influences impact current market prices. This variable is crucial for determining the pricing dynamics employed by different vendors and how those strategies affect consumer costs.

To provide a comprehensive view of the outcome variable, Here's a table for easier understanding (Table 1). This table provides the minimum, maximum, mean, median, and standard deviation of the current_price variable, offering insights into the overall distribution and variability in product prices.

Table 1: Summary statistics for the outcome variable (current_price)

Statistic	Value
Min	0.770000
Max	23.980000
Mean	9.111933
Median	8.580000
Standard Deviation	4.327384

In addition to the summary table, the distribution of current_price is visualized through a histogram (Figure 1), highlighting common price ranges and the overall spread. This visualization helps to identify any skewness or clustering in the data, which may indicate specific pricing patterns. Together, these tables and graphs provide a comprehensive understanding

of the outcome variable, offering a detailed view of how product prices vary across different vendors and over time.

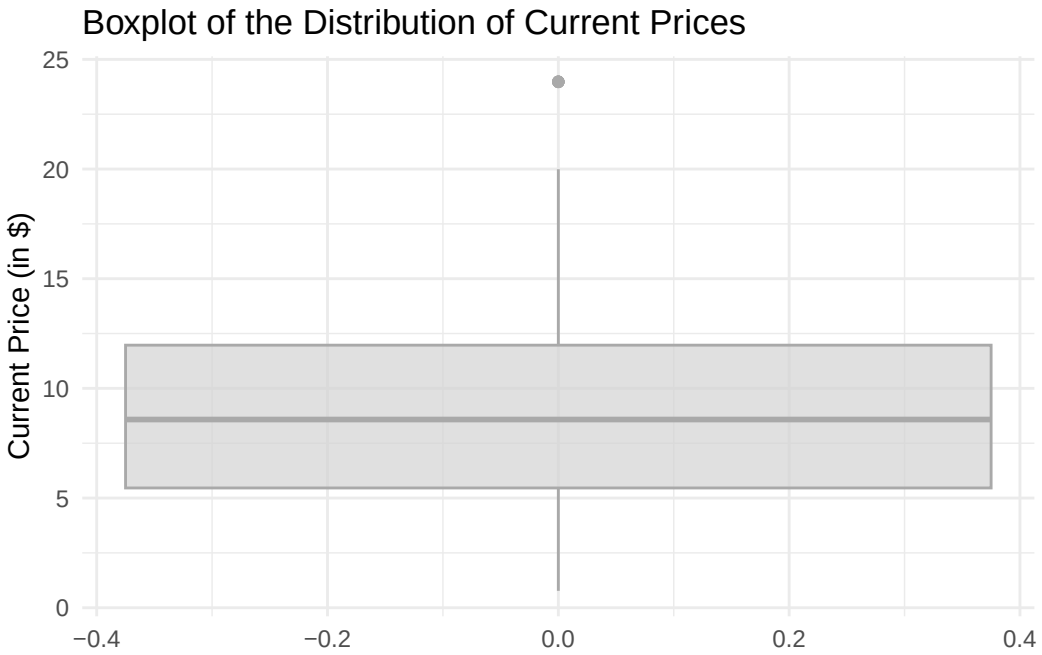

Figure 1: Boxplot of the distribution of current prices

The boxplot (Figure 1) presented above provides a detailed summary of the distribution of current_price for the products in the dataset. Boxplots are useful visualization tools for understanding the spread, central tendency, and overall variability of continuous data. In this case, the boxplot helps in identifying key features such as the median price, the interquartile range (IQR), and any potential outliers that may exist in the pricing data.

The median, represented by the horizontal line within the box, provides an indication of the central price of products in the dataset, highlighting where the bulk of prices are situated. The box represents the IQR, showing where the middle 50% of data points lie, thus giving insights into price concentration. The whiskers, extending from the box, represent the range within which most prices fall, excluding any extreme values or outliers. No outliers were visibly identified in this boxplot, which suggests that most product prices are consistently within a particular range.

The flat shape of the boxplot with a wide range of the IQR indicates low variability in current product prices. The relatively small difference between the lower and upper bounds of the box implies that the pricing of the products is consistent, with most prices concentrated around the median. This consistency may suggest that vendors are using similar pricing strategies or that there is limited differentiation in the types of products sold across vendors. Overall, the

boxplot is a valuable tool for summarizing and communicating the underlying distribution of current_price, which aids in understanding the dynamics of pricing across different vendors.

2.4 Predictor variables

Add graphs, tables and text.

2.4.1 Month

The month variable is used as a predictor to capture potential seasonal influences on pricing. By including month, we can observe if specific times of the year are associated with higher or lower prices. This is particularly important for identifying temporal patterns in pricing, which can be influenced by factors such as holidays, sales events, or seasonal demand changes. Table (Table 2) provides a summary of the observations for each month, showing how frequently each month appears in the dataset. This information can help highlight any periods with higher or lower data collection frequency, which could influence the interpretation of seasonal pricing effects.

Table 2: Summary statistics for the predictor variable (month)

Month	Count
6	433
7	514
8	624
9	535
10	578
11	342

2.4.2 Old Price

old_price serves as another key predictor variable, representing the historical price of a product. This variable helps gauge the effect of historical pricing on current pricing strategies, indicating whether a product has undergone discounts or price hikes. Table (Table 3) provides a summary of key statistics for old_price, including minimum, maximum, mean, median, and standard deviation. Understanding the distribution of old_price helps us assess whether historical prices significantly differ from current prices, and whether pricing adjustments (e.g., discounts) have been applied uniformly across products. The relationship between old_price and current_price is visually explored in Chart 4, which helps understand the influence of past pricing decisions on current prices.

Table 3: Summary statistics for the predictor variable (old_price)

Statistic	Value
Min	1.87000
Max	25.99000
Mean	11.48035
Median	10.97000
Standard Deviation	5.32401

2.4.3 Vendor

The vendor variable is an important categorical predictor that differentiates pricing behavior across different retail chains. It allows us to analyze how pricing strategies vary among vendors such as Walmart and Galleria. Table (Table 4) shows the count of observations for each vendor, allowing us to compare how frequently data was collected for each retail chain. This helps in understanding the representation of each vendor within the dataset and assessing whether certain vendors may have a stronger influence on the overall analysis. Chart 2, which shows the price difference by vendor, provides insights into how different vendors adjust their pricing strategies, offering valuable comparisons among them. Additionally, Chart 3 presents the average current_price over time, broken down by vendor, providing a clearer picture of how vendor-based strategies influence pricing trends.

Table 4: Count of observations for each vendor

Vendor	Count
TandT	1330
Walmart	1696

3 Model

Background details and diagnostics are included in Appendix B.

3.1 Model overview

To model the current price of beef, y_i, at time i, we use a Bayesian linear regression model implemented with the `stan_glm` function in the R package `rstanarm`. The response variable, y_i, represents the **current price of beef** in dollars. Our predictors include x_1, x_2, and x_3, which represent the **month**, **old price**, and **vendor**, respectively. Each component of the model is defined and justified below.

3.2 Model set-up

Define y_i as the current price of beef at time i, measured in dollars. Let x_1, x_2, and x_3 represent the predictors: x_1 is the month (categorical variable), x_2 is the old price, and x_3 is the vendor (categorical variable).

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \tag{2}$$

In this model: - α represents the **intercept** term, capturing the baseline price of beef. - β_1 corresponds to the **effect of the month** on the current price, allowing us to capture **seasonal variations** that may impact beef pricing. The month is treated as a categorical variable, acknowledging that different months can bring different demand or supply effects due to holidays, weather, or other seasonal factors. - β_2 represents the **effect of the old price** on the current price. This allows us to account for price inertia or trends where past pricing influences current pricing. - β_3 represents the **effect of the vendor**, also treated as a categorical variable. Different vendors may have distinct pricing strategies, and including vendor as a categorical predictor helps us capture this vendor-specific pricing variation.

We assume normal priors for the model coefficients and intercept, with mean 0 and standard deviation 2.5:

$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{6}$$

These priors are chosen to be **weakly informative**, allowing the data to speak for itself while still providing a reasonable range for the coefficients based on prior expectations. Specifically, a standard deviation of 2.5 reflects our expectation that most reasonable effects should fall within a plausible range without being overly restrictive.

For the residual standard deviation, σ, we assume an **Exponential(1)** prior:

$$\sigma \sim \text{Exponential}(1) \tag{7}$$

This prior reflects our belief that the standard deviation should be positive and allows flexibility, while preferring smaller values over larger ones, consistent with the expectation of modest variability around the mean prediction.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.3 Model justification

The choice of predictors is well-balanced in terms of complexity and appropriateness for this scenario. By including **month** and **vendor** as categorical variables, we acknowledge the significant impact that both seasonal changes and vendor-specific factors can have on the current price of beef. The inclusion of the **old price** reflects the idea of price inertia, where historical prices are likely to influence current pricing trends.

The model is neither **overly simplistic** nor **unnecessarily complex**: it captures important predictors without adding extraneous complexity that could lead to overfitting. **Month** and **vendor** are treated as categorical variables rather than grouping by more arbitrary categories, which maintains the expressiveness of our model while being parsimonious.

The model is implemented using `stan_glm` from the `rstanarm` package, a high-level interface to Stan for Bayesian modeling. **Stan** provides powerful sampling algorithms, making it an appropriate tool for fitting our Bayesian model efficiently, ensuring **robust convergence** and accurate posterior estimation.

3.3.1 Model Assumptions and Limitations

- **Linearity**: The model assumes a linear relationship between the predictors and the response variable. This may be a limitation if the true relationship is nonlinear.
- **Normality of Residuals**: The residuals are assumed to follow a normal distribution. If this assumption does not hold, it could bias the model predictions.
- **Priors**: We used weakly informative priors to allow flexibility. However, if we had prior knowledge of specific pricing patterns, more informative priors could be used for better inference.

3.3.2 Model Validation and Diagnostics

- **Posterior Predictive Checks**: We performed posterior predictive checks to verify that the model adequately captures the variation in the data. These checks involve comparing simulated data from the posterior distribution to the observed data to check for any discrepancies.
- **Convergence**: Model convergence was assessed using **trace plots** and the **R-hat statistic**, with all values being close to 1, indicating successful convergence.

Table 5

	First model
(Intercept)	−0.37
	(0.14)
month	−0.01
	(0.01)
old_price	0.81
	(0.00)
vendorWalmart	0.56
	(0.04)
Num.Obs.	3026
R2	0.945
R2 Adj.	0.944
Log.Lik.	−4349.672
ELPD	−4354.7
ELPD s.e.	58.7
LOOIC	8709.5
LOOIC s.e.	117.3
WAIC	8709.4
RMSE	1.02

- **Alternative Models**: We also considered fitting simpler linear models that did not include the vendor or seasonal components. These models showed significantly lower predictive performance, indicating the importance of including both the vendor and seasonal factors for accuracy.

In summary, the chosen model provides a comprehensive yet interpretable understanding of how different factors influence the current price of beef. The Bayesian approach allows us to incorporate uncertainty and prior beliefs, while our specific choice of priors and predictors ensures that the model is well-suited for the available data and research questions.

4 Results

Our results are summarized in Table 5.

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.

Appendix

A Additional data details

B Model details

B.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

B.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 algo-
rithm

References

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

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

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