My title*

My subtitle if needed

Ziheng Zhong

November 24, 2024

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

Table of contents

---

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

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?**),

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

The outcome variable for this study, current_price, serves as the key focus of analysis in understanding product pricing dynamics. This section provides a structured overview of the current_price variable, including its distribution and any observed trends, supported by both summary statistics and visualizations. The aim is to offer a clear understanding of how product prices vary and what factors may influence those variations.

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

The predictor variables in this study—month, old_price, and vendor—each play an important role in explaining variations in current_price. This section outlines how these variables potentially influence pricing, with a focus on capturing temporal patterns, historical effects, and vendor-specific pricing strategies. The summary tables and visual analyses help contextualize the impact of each predictor on the outcome, providing a broad understanding of the pricing dynamics within the dataset.

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.
- **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

4.1 Charts

Figure 2 (Figure 2) presents the distribution of current beef prices in the dataset, offering insights into the range and frequency of beef pricing. The horizontal axis represents the price in dollars, ranging from 0 to approximately 25, while the vertical axis indicates the frequency, or the number of observations, within each price range. The grey bars illustrate the frequency distribution, showing how beef prices are spread across different price levels.

The histogram indicates that the most frequent price range is between $5 and $10, with a notable peak in the frequency around $7 to $8. This suggests that a significant proportion of beef products are priced within this range, making it a central pricing cluster in the market. Additionally, there are smaller peaks around $13 and $15, hinting at secondary groupings where beef prices are relatively common. The declining frequency beyond $15 suggests fewer beef products are priced in the higher range, indicating that higher prices are less typical in this dataset.

The overlaid red line represents a smoothed density curve that helps to visualize the general shape and distribution trend of the prices. The curve follows the histogram closely, showing a skewed distribution with a concentration towards the lower prices, and tails off as prices increase. This pattern suggests that while lower-priced beef products are more common, higher prices are less frequent, potentially indicating a general preference or affordability constraint
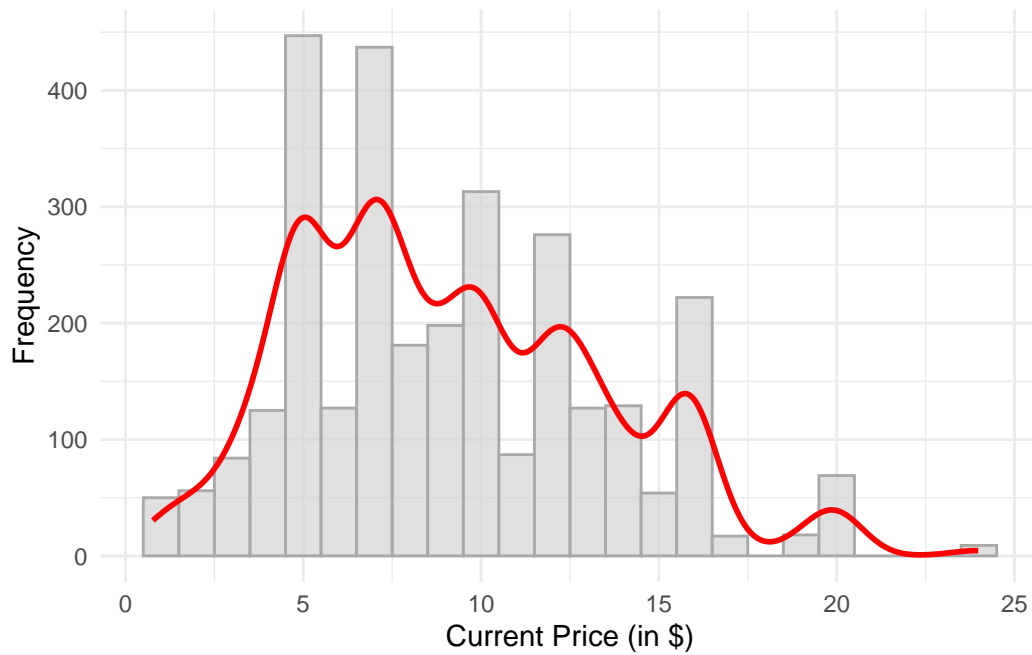
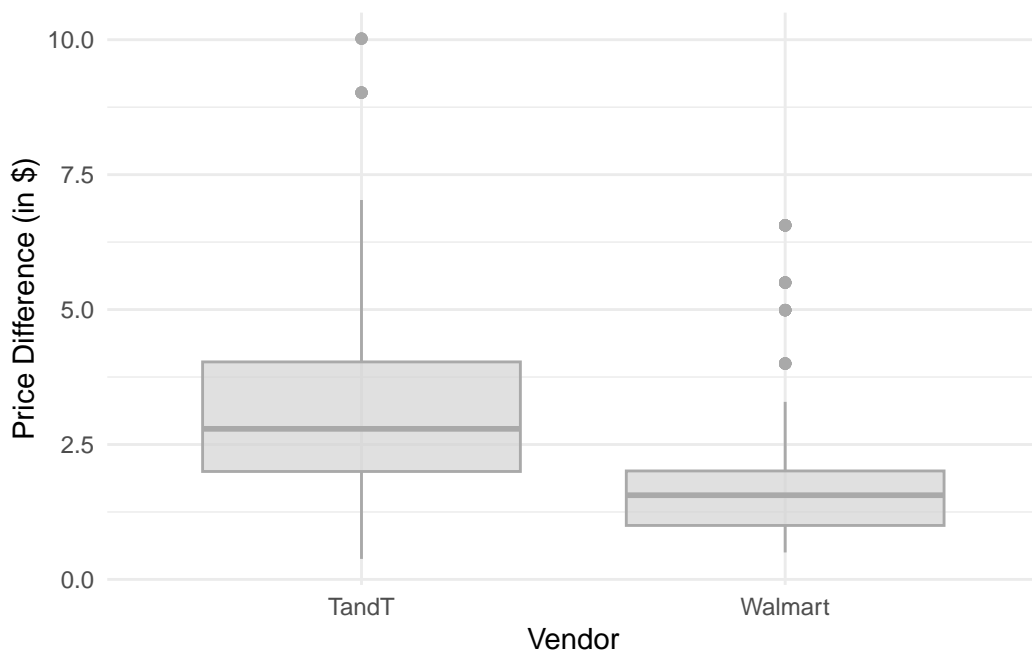Figure 2: Distribution of Current Prices

Figure 3: Price Difference by Vendor

for lower-cost beef among consumers. The presence of multiple peaks could also imply different pricing tiers, perhaps influenced by factors such as beef quality or specific product types.

Figure 3 (Figure 3) illustrates the distribution of price differences between beef products offered by two vendors, T&T and Walmart, providing an insight into the variability of pricing across these retailers. The y-axis represents the price difference in dollars, while the x-axis labels the two vendors under comparison. The boxplots visually depict the range, median, and variability in the price differences for each vendor.

For T&T, the boxplot shows a wider range of price differences, with the interquartile range (IQR) extending from roughly $1.5 to $4, indicating a moderate to substantial variability in price differences among the products they sell. The median price difference is around $2.5, suggesting that many of the products have a price difference near this value. However, T&T also shows significant variability with the presence of two outliers beyond $7.5 and $10, indicating that a few beef products exhibit much larger price differences compared to the rest of the data.

In contrast, Walmart exhibits a narrower spread of price differences, as indicated by the more compact boxplot. The IQR for Walmart lies approximately between $1 and $2.5, and the median price difference is lower compared to T&T, close to $1.5. This suggests that Walmart's beef products tend to have smaller and less variable price differences, indicating more consistent pricing practices. There are also a few outliers above $4, but these are less extreme compared to T&T's outliers.

Overall, the boxplots suggest that T&T has more variability in their pricing, possibly reflecting differences in product types or marketing strategies, whereas Walmart tends to maintain a narrower price range with fewer significant deviations. This disparity may indicate that Walmart focuses on more standardized pricing practices, which could influence consumer choice based on price predictability and consistency.

Figure 4 (Figure 4) shows the trend of average beef prices over time, providing insight into fluctuations in the market from July to November. The y-axis represents the average current price in $, while the x-axis represents the date, with labels indicating the progression through the months.

The line chart indicates that average beef prices experienced significant variability throughout the observed period, with several sharp peaks and troughs. In early July, prices were relatively high, reaching a peak of approximately $13, but this was followed by a rapid decline, dropping to around $7 by the end of the month. The fluctuations continue through August, where we see notable volatility, including several spikes back to higher average prices of over $10. Such fluctuations suggest sensitivity in pricing, possibly influenced by seasonal changes, supply chain disruptions, or market demand dynamics.

Entering September and October, there is a pronounced drop in prices, with average prices occasionally dipping below $6, reflecting a period of reduced pricing. Following this low, prices exhibit recovery and some stabilization, hovering around $8 to $9 from October into early
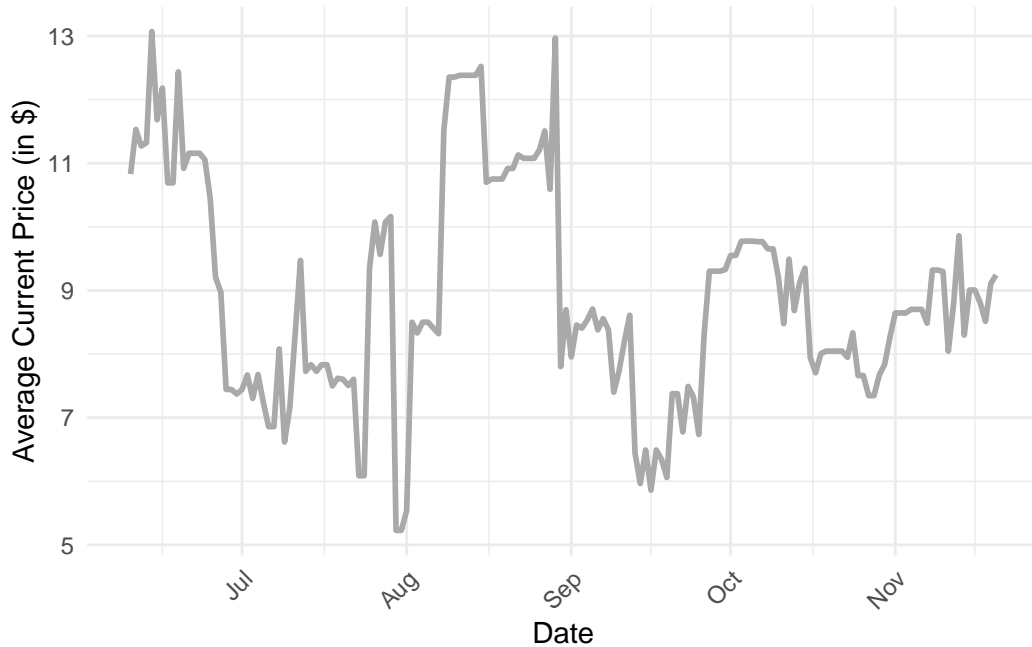
Figure 4: Average Current Price Over Time

November, albeit with some ongoing small fluctuations. The chart's volatility indicates that beef prices are influenced by various factors throughout these months, with possible implications from supply conditions, retailer pricing strategies, or changing consumer demand. This temporal analysis of beef pricing is critical for understanding market stability and identifying periods of significant price changes that may warrant further investigation.

Figure 5 (Figure 5) visualizes the relationship between the old price and current price of beef, providing insight into how pricing has evolved over time for various products. The scatter plot includes individual points that represent different observations of beef prices, with the x-axis indicating the old price (in $) and the y-axis representing the current price (in $). The grey trend line overlaid on the plot illustrates the overall trend in the data.

The plot indicates a strong positive correlation between old prices and current prices. As the old price increases, the current price also tends to increase, suggesting consistency in price adjustments over time. The data points are generally clustered around the trend line, implying a stable relationship with only minor deviations, which highlights that most beef products have seen proportional changes in their pricing. This could indicate the influence of factors such as inflation or consistent pricing practices across vendors.

However, there are some noticeable variations, with data points deviating above and below the trend line. These variations suggest that certain products have experienced different rates of price change, potentially reflecting changes in supply costs, varying consumer demand, or
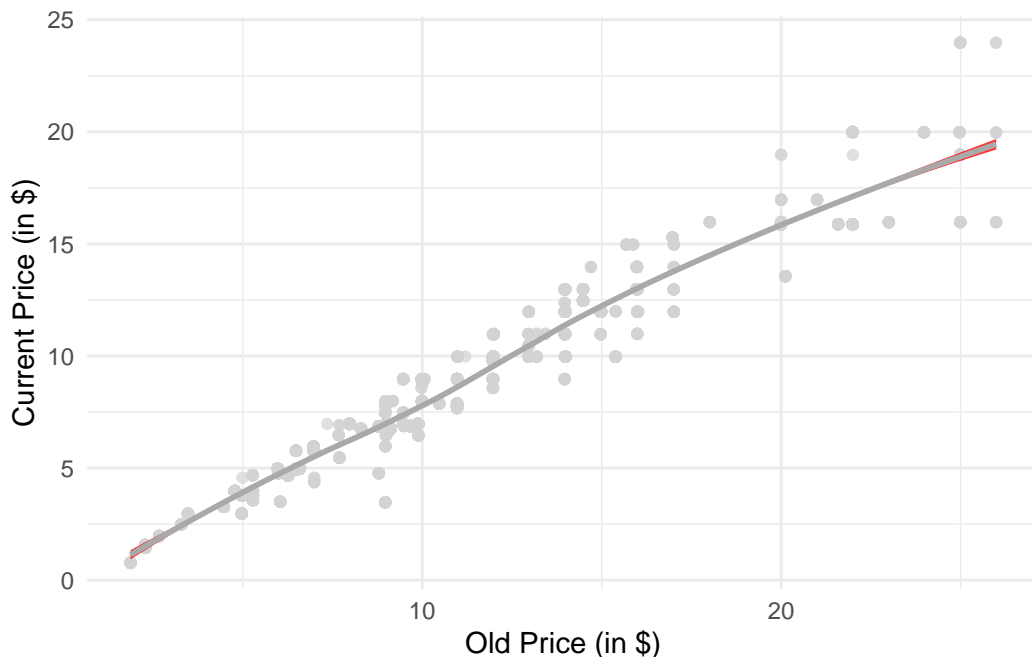
Figure 5: Distribution of Current vs. Old Price of Beef

differences in retailer pricing strategies. For instance, products above the trend line indicate higher current prices compared to their historical pricing, possibly pointing towards a higher demand or limited availability. Conversely, points below the trend line indicate products that may have seen reductions or smaller increments in pricing. Overall, the plot reveals a proportional relationship between historical and current beef prices, with a trend towards increasing prices that is likely influenced by broader market forces.

Figure 6 (Figure 6) illustrates the distribution of unit prices for beef, measured in $ per 100 grams. The x-axis represents the unit price, while the y-axis indicates the frequency of occurrences for each price range. The histogram, represented by grey bars, shows how often each unit price appears in the dataset, and the red curve provides a smoothed density estimate to visualize the general trend.

The most striking feature of this figure is the concentration of unit prices near zero, with a significant spike in the frequency at this price point. This suggests that a substantial number of observations in the dataset have a very low or nominal unit price, which could indicate promotions, discounted products, or even data entry inconsistencies that require further examination. After this initial peak, the frequency drops sharply, followed by a few smaller peaks across higher unit price ranges, particularly around $1, $3, and $6 per 100 grams. These smaller peaks indicate the presence of additional groupings of unit prices, potentially corresponding to different product categories, quality grades, or pricing tiers.
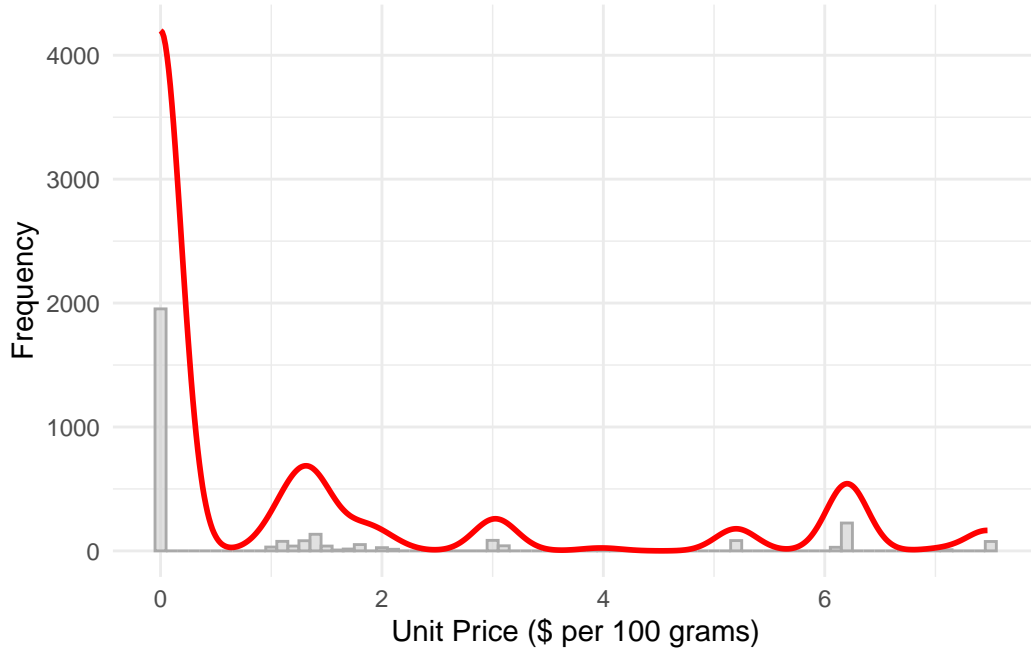
Figure 6: Distribution of Unit Prices

The overall distribution is heavily right-skewed, suggesting that the majority of beef products have lower unit prices, while higher prices are less common and scattered. This right-skewness might be indicative of pricing strategies that cater to a wide consumer base, where most products are priced affordably, with only a few premium items available at significantly higher unit prices. The presence of several distinct peaks, rather than a single uniform distribution, suggests price differentiation that could be driven by various factors, such as beef cut type, quality, or retailer-specific pricing strategies. This distribution pattern highlights the diverse pricing landscape for beef products in the dataset, suggesting the existence of both low-cost and premium segments in the market.

4.2 Model

Table 5 (Table 5) summarizes the results from the Bayesian model applied to analyze beef prices, including key parameter estimates and model diagnostics. The model provides insight into the relationship between the current price of beef and various predictors such as the old price, vendor, and the time of observation (represented by the "month" variable).

The estimated intercept of -0.37 suggests that, after controlling for the predictors in the model, the baseline level of the current price is negative, though this value must be interpreted with caution in terms of its contextual implications. The "month" coefficient is -0.01, indicating a

Table 5

	Beef model (Bayesian)
(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

slight negative association between the time variable and current price, although the magnitude of the effect is relatively small. This suggests that as time progresses, there is a minor decrease in the current price, potentially reflecting seasonal effects or gradual changes in market dynamics. The "old_price" coefficient is 0.81, showing a strong positive relationship with the current price, meaning that for each unit increase in the old price, the current price increases by a comparable margin. This result implies a consistent upward adjustment in beef pricing over time, reinforcing the trends seen in previous charts. Finally, the "vendorWalmart" coefficient is 0.56, indicating that beef sold by Walmart tends to be priced higher, on average, than that sold by the baseline vendor (T&T), which aligns with our earlier findings on vendor price differences.

In terms of model fit, the R-squared (R^2) value of 0.945 and the adjusted R-squared value of 0.944 indicate a high degree of explained variance, suggesting that the model is capable of capturing the main factors influencing beef prices. The negative log-likelihood value (-4349.672) and expected log pointwise predictive density (ELPD) of -4354.7 provide additional metrics on model fit, while other diagnostic values, such as the leave-one-out information criterion (LOOIC) and Widely Applicable Information Criterion (WAIC), both around 8709, provide an indication of model performance for out-of-sample predictive accuracy. The root mean square error (RMSE) is 1.02, suggesting the model's predictions have a relatively small average deviation from the actual values, indicative of good predictive performance.

These results together illustrate the strong relationship between past pricing and current pricing, as well as vendor influence on beef pricing. The Bayesian approach has provided robust parameter estimates while accounting for uncertainty, as reflected in the provided standard errors. Overall, the model appears to perform well, with high explained variance and reasonable prediction accuracy, making it a valuable tool for understanding and forecasting beef price trends.

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

A.1 Vendor Choice

A.2 Price per unit

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 algorithm

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/.