My title[*]

My subtitle if needed

First author Another author

November 26, 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

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (**shelter?**)…. Following (**tellingstories?**), we consider…

Overview text

[*] Code and data are available at: https://github.com/RohanAlexander/starter_folder.

2.2 Measurement

The measurement process involves transforming real-world phenomena into data that can be analyzed. In this case, the real-world phenomenon is the sale of potatoes by various vendors over time. Each transaction recorded in the raw dataset represents an instance of this phenomenon.

To transition from real-world sales events to a structured dataset, we began by recording each transaction, which included details like the product name and brand, prices, and the vendor information. This research focus on the single product-potato. Therefore, `product_name` was used to identify the specific types of potatoes (yellow and white), allowing us to focus on these items exclusively. The `current_price` and `old_price` fields represent the financial aspect of these sales, capturing both the price at the time of the transaction and the historical price for comparison.The `month` variable was derived from the transaction timestamp (`nowtime`) to provide a temporal context for analysis, allowing us to observe trends and patterns over different periods.

By filtering and cleaning the raw data, the dataset reflects the aspects of potato sales that are of interest for this study. This process of measurement transforms abstract sales activities into quantifiable data points that can be used for statistical analysis, providing insights into vendor behavior and price dynamics over time.

2.3 Cleaning Process

The Canadian Grocery Price Data (Filipp (2024)) used in this analysis was sourced from the official website and consists of two datasets: raw sales data and product-specific information. Project Hammer aims to drive more competition and reduce collusion in the Canadian grocery sector. The dataset includes sales data from eight vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. The available dates range from February 28, 2024, to the latest data load.

Initially, two raw datasets were downloaded from the Project Hammer website. These datasets were merged to create a comprehensive dataset containing all relevant columns, including `nowtime`, `vendor`, `product_id`, `product_name`, `brand`, `units`, `current_price`, `old_price`, `price_per_unit`, and `other`.

The cleaning process involved several key steps to refine the dataset for analysis. First, we selected only the columns relevant to our study shown in Table 1. The month of each transaction was extracted from the nowtime field to provide temporal context for the sales data.

Table 1: Summary of Selected Columns

Column	Description
nowtime	Timestamp indicating when the data was gathered
vendor	One of the 7 grocery vendors
current_price	Price at time of extract
old_price	An 'old' struck-out price, indicating a previous sale price
product_name	Product name, may include brand and units

Next, we filtered the dataset to include only the products of interest—yellow and white potatoes—by searching for these keywords in the product_name column, ensuring our analysis was focused specifically on these items. Additionally, any entries with missing values (NA) were removed to maintain data quality. Finally, the nowtime column was dropped after extracting the month, as it was no longer necessary for the analysis. Table 2 gives an preview of the cleaned dataset we will use in the following sections.

Table 2: Preview of Cleaned Data on Potato Sales

vendor	current_price	old_price	product_name	month
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.50	5.99	Small White Potatoes	10
Loblaws	4.50	5.99	Small White Potatoes	10

These steps resulted in a cleaned dataset containing the essential information needed for the analysis of potato sales trends.

2.4 Variables of Interest

The cleaned dataset contains 685 rows, representing individual products sold, along with vendor information and price details. The summary statistics of the cleaned data indicate that the mean of current price is 2.811, and the mean of old price is 3.785. Figure 1 and Table 3 show that the cleaned dataset only contains five vendors instead of the eight in the raw dataset, and the month is from June to November. This is expected because data collection from February 28 to July 10/11 focused on a smaller set of products by the description of the raw dataset. After July, more products were added, and some data may be missing for certain vendors or days when extraction failed.

Table 3: Overview of Analysis Data

current_price	old_price	month
Min. :0.940	Min. :1.490	Min. : 6.000
1st Qu.:2.000	1st Qu.:2.490	1st Qu.: 8.000
Median :2.290	Median :2.490	Median :10.000
Mean :2.811	Mean :3.785	Mean : 9.064
3rd Qu.:3.990	3rd Qu.:5.990	3rd Qu.:10.000
Max. :7.990	Max. :8.990	Max. :11.000

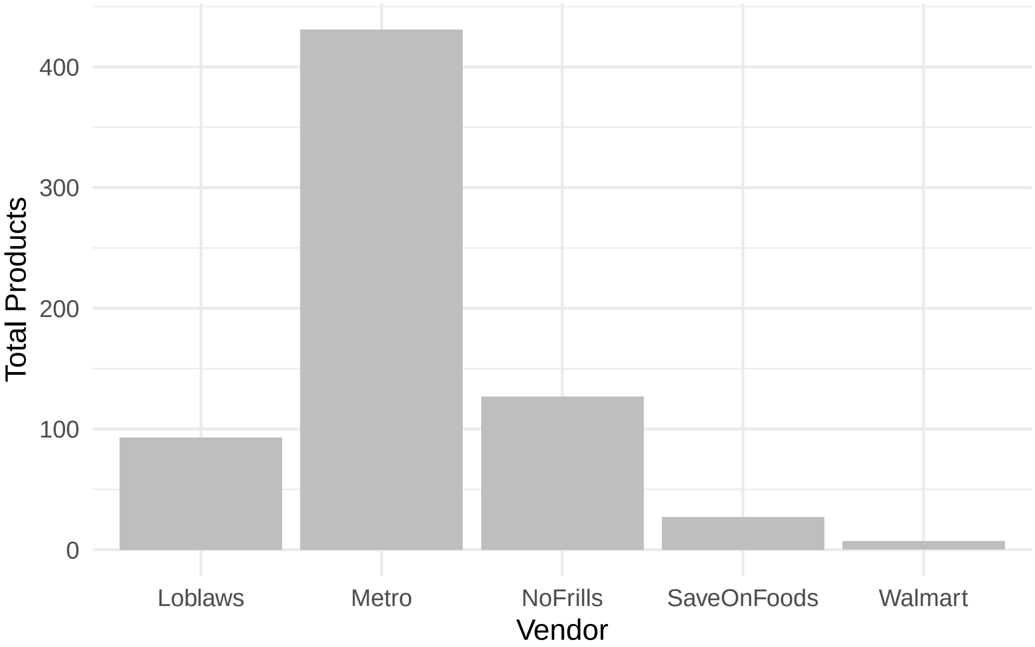

Figure 1: Total Number of Products by Vendor

2.5 Other dataset

A dataset on monthly average retail prices for selected products (Government of Canada (2024)) was identified but ultimately not utilized. The dataset only provides average prices, which limits its ability to capture price fluctuations over time. Furthermore, it lacks vendor-specific information, which is essential for analyzing the impact of different vendors on product pricing.

3 Model

The goal of our analysis was to understand the relationship between current potato prices and influencing factors, specifically `month`, `old_price`, and `vendor`. To accomplish this, we developed a Bayesian regression model to capture the underlying dynamics that contribute to changes in pricing.

Our variables of interest will be the predictor variables for the model.The Bayesian model uses the following predictors:

- `month`: Represents the month during which the data was collected. The values range from June to November (i.e., months 6 to 11).
- `old_price`: Represents the previous price of the product, providing insight into historical pricing trends. It is used to understand how past prices influence current pricing.
- `vendor`: Represents the specific grocery vendor selling the product. The dataset includes five vendors, including Loblaws, Metro, No Frills,Save-On-Foods and Walmart.

3.1 Model set-up

To predict the current price of a product, we assume a linear relationship between the outcome variable (`current_price`) and our predictor variables (`month`, `old_price`, and `vendor`). We define our model as a Bayesian regression model as follows:

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \times \text{month}_i + \beta_2 \times \text{old\_price}_i + \beta_3 \times \text{vendor}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

In this model, y_i represents the current price for product i, modeled as a normal distribution with mean μ_i and standard deviation σ. The predictor variables month_i, old_price_i, and vendor_i represent the month of data collection, the old price of the product, and the vendor, respectively. The coefficients β_0, β_1, β_2, and β_3 are assigned normal priors with a mean of 0 and a standard deviation of 2.5, while σ is assigned an exponential prior with rate 1. Choosing a relatively loose prior allows the coefficients to fluctuate within a moderate range while avoiding overly restrictive assumptions. A normal distribution with a standard deviation of 2.5 can capture most plausible true effects but avoids assigning excessively high probabilities to very large coefficients. This takes into account that the influence of predictors (such as month, old price, and vendor) on the response variable (current price) is generally limited or moderate.

The Bayesian model offers advantages, such as the ability to incorporate prior information and quantify parameter uncertainty more comprehensively. We ran the model using the `rstanarm` package (Goodrich et al. 2024) in R (R Core Team 2023). Model diagnostics, including trace plots, Rhat values, and posterior predictive checks, can be found in the Appendix (Section A).

3.1.1 Model justification

The multiple linear regression model, with the same predictor variables as the Bayesian model, is also suitable for this analysis. First, the model demonstrates strong explanatory power for the data, as indicated by high R-squared and adjusted R-squared values. Additionally, most predictor variables are statistically significant, allowing for a clear interpretation of their influence on `current_price`. Moreover, the conclusions about the significance of the variables are consistent between the linear and Bayesian models, reinforcing the reliability of the results. These analyses and diagnostics can also be found in the Appendix (Section A).

However, we ultimately chose the Bayesian approach for its flexibility in incorporating prior knowledge and providing a posterior distribution of the model parameters, which offers a more nuanced understanding of the uncertainty. The Bayesian model is particularly useful when dealing with smaller datasets or high variability. The cleaned data we used has only around 600 observations, which makes the Bayesian model better suited for our study compared to an ordinary least squares model.

4 Results

Our results are summarized in Table 4.

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.

Table 4: Explanatory model of current price based on month, old price, and vendor

	Bayesian regression model for current price
(Intercept)	0.31
	(0.16)
month7	0.08
	(0.10)
month8	0.00
	(0.09)
month9	0.05
	(0.09)
month10	0.34
	(0.09)
month11	0.52
	(0.10)
old_price	0.54
	(0.02)
vendorMetro	0.35
	(0.11)
vendorNoFrills	−0.34
	(0.09)
vendorSaveOnFoods	2.56
	(0.14)
vendorWalmart	−1.75
	(0.24)
Num.Obs.	685
R2	0.856
R2 Adj.	0.854
Log.Lik.	−607.781
ELPD	−618.4
ELPD s.e.	35.4
LOOIC	1236.8
LOOIC s.e.	70.7
WAIC	1236.8
RMSE	0.58

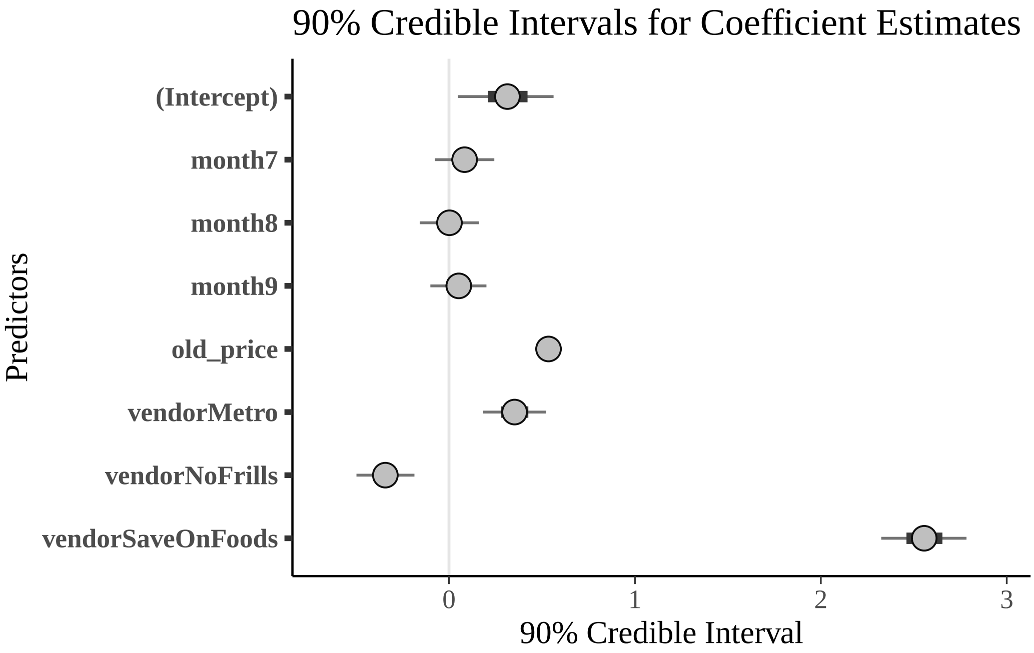

Figure 2

5.2 Second discussion point

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

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

B Additional data details

C Model details

C.1 Posterior predictive check

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

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

C.2 Diagnostics

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

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

References

Filipp, Jacob. 2024. *Project Hammer*. https://jacobfilipp.com/hammer/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Government of Canada. 2024. *Monthly Average Retail Prices for Selected Products*. https://open.canada.ca/data/en/dataset/8015bcc6-401d-4927-a447-bb35d5dfcc91/resource/5ab003e5-fea3-40b8-8c10-8ac8988bfc93.

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