Canadian_Grocery_Coffee_Price_Analysis*

Jin Zhang

November 27, 2024

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

Table of contents

*Code and data are available at: https://github.com/KrystalJin1/Canadian_Grocery_Coffee_Price.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

2.1 Overview

Price of each month's coffee of different vendors data is provided by(**citedata?**). This dataset records detailed sales about fast-moving consumer goods (FMCG) sold by various vendors, including volia, T&T, Loblaws, SaveOnFoods, Galleria, Metro, NoFrills and Walmart. It is also includes product-level details, such as the product name, current price, historical price (old price), and the corresponding units and price per unit. The data also captures time-specific observations(2024-2-28 to 2024-6-22), with timestamps (nowtime) that can be used to analyze trends over days or months.

In order to simulate data, test simulated data, clean data, test cleaned data, exploratory data analysis and model data, we used R programming language (R Core Team 2023) to analyze the data and plot the graphs. Specific libraries that assisted the analysis include `tidyverse` (**tidyverse?**), `palmerpenguins` (**citepalmerpenguins?**), `knitr` (**citeknitr?**), `arrow` (**citearrow?**), `ggplot2` (**citeggplot2?**), `dplyr` (**citedplyrx?**), `here` (**citehere?**), `kableExtra` (**citekableExtra?**), `gridExtra`(**citegridExtra?**), `modelsummary`(**citemodelsummary?**), `rstanarm`(**citerstanarm?**).

The inspiration for my data processing came from my desire to study what factors would affect the current price of coffee products from two vendors in different regions of Canada, such as the current price of coffee products from two vendors, Metro and SaveOnFoods. The following variables are the data I selected after cleaning the data:

- vendor: The retailer selling the product in Canada.
- old_price: The historical price of the product, showing previous pricing or discounts.
- product_name: The specific product being sold, providing product-level insights.
- current_price: The price of the product at the time of observation.

New variable extracted and transformed from raw data:

- month: The month of data collection, extracted from `nowtime`.

Since the variable nowtime only records 4 months, it is considered a lack of Long-Term Trends, which means it's difficult to identify long-term pricing or demand patterns by using short data periods. So I only extracted a new variable—month from date of nowtime, which can simplify temporal analysis and identify trends, such as seasonal price changes or demand patterns. It allows grouping data for monthly aggregation and supporting seasonality-focused insights or forecasting models.

To provide an preview of the coffee pricing with all potential factors that might affect it. Here, Table 1 simply reveals the variation between current price and old price in June for Metro's coffee products.

Table 1: Sample of Analysis Data Showing Products Sold by Both Vendors

vendor	product_name	current_price	old_price	month
Metro	Non-Dairy Vanilla Flavoured Latte Coffee Cream	7.49	8.99	6
Metro	Vanilla And Caramel Flavoured K-Cup® Coffee Capsules	9.99	12.99	6
Metro	Classic Black K-Cup® Coffee Capsules	9.99	12.99	6
Metro	Cold Brew Unsweetened Iced Coffee	7.49	7.99	6
Metro	Limited Edition Coffee Whitener, Coffee Mate	4.99	6.99	6
Metro	Italian Blend Dark Roast K-Cup Coffee Pods	6.49	6.99	6
Metro	Classic Roast Ground Coffee	8.99	12.49	6
Metro	Medium Roast Decafreinated K-Cup Coffee Pods, Pike P...	22.99	26.99	6
Metro	Medium Roast House Blend K-Cup Coffee Pods, Organic	6.49	6.99	6
Metro	Classic Decaf Ground Coffee	8.99	12.99	6

2.2 Measurement

The dataset from Hammer represents real-world retail activities, capturing product details, vendor listings, and price updates. When vendors update product information, such as pricing or availability, Hammer collects and structures this information into the dataset.

Key fields include vendor, product_name, current_price, old_price, and nowtime. The data is collected through scraping and structured to enable analysis of retail trends, pricing strategies, and market dynamics over time. Each entry serves as a snapshot of a product's presence in the market at a specific time, allowing for focused analyses, like tracking price trends for specific products (e.g., coffee).

2.3 Data Visualization

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

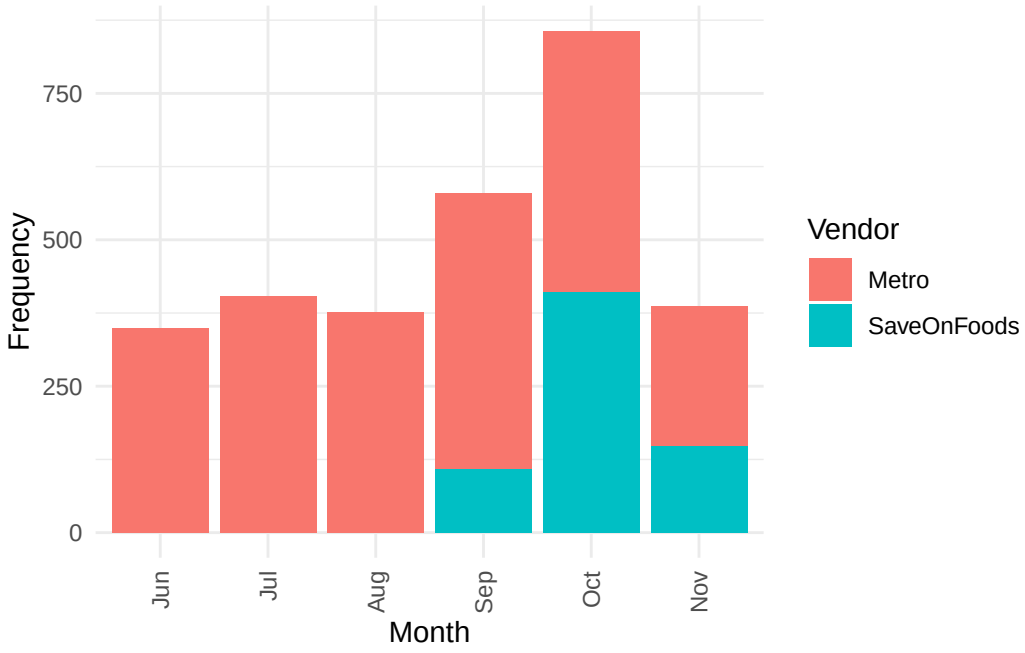


Figure 1: Monthly Distribution of Coffee Products by Vendor

Figure 2: Comparison of Current and Old Prices for Coffee by Vendor

Table 2: Statistics summary of the cleaned coffee pruducts pricing

vendor	product_name	current_price	old_price	month
Length:2954	Length:2954	Min. : 0.63	Min. : 0.89	Min. : 6.000
Class :character	Class :character	1st Qu.: 5.99	1st Qu.: 8.29	1st Qu.: 7.000
Mode :character	Mode :character	Median : 9.49	Median :12.49	Median : 9.000
NA	NA	Mean :10.21	Mean :13.26	Mean : 8.796
NA	NA	3rd Qu.:12.00	3rd Qu.:15.99	3rd Qu.:10.000
NA	NA	Max. :35.99	Max. :40.99	Max. :11.000

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our Bayesian multiple linear regression is to investigate the factors that influence the current price of coffee in our dataset. Specifically, we aim to understand how historical pricing, vendor differences, and seasonal patterns affect current coffee prices.

3.1 Model set-up

Define y_i as the current price of coffee for the i-th observation in the dataset. The predictors include:

- x_{1i}, the old price of the coffee,
- x_{2i}, dummy variable for the vendor, where:
  $x_{2i} = 1$: Vendor is "SaveOnFoods"; $x_{2i} = 0$: Vendor is "Metro",
- x_{3i}, the numeric month variable.

The model is formulated as follows:

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma), && (1)\\
\mu_i &= \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i}, && (2)\\
\alpha &\sim \text{Normal}(0, 2.5), && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5), && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5), && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5), && (6)\\
\sigma &\sim \text{Exponential}(1). && (7)
\end{aligned}
$$

This model describes the relationship between the current price of coffee (y_i) and three predictors: the old price of coffee (x_{1i}), a categorical vendor variable (x_{2i}) indicating whether the vendor is "Metro" or "SaveOnFoods," and a numeric variable for the month (x_{3i}). The response variable (y_i) is modeled as normally distributed with mean μ_i and standard deviation σ. The mean μ_i is defined as a linear combination of these predictors, with coefficients β_1, β_2, and β_3, and an intercept α. Prior distributions for the parameters are specified, including normal priors for α and the coefficients, and an exponential prior for σ. Intercept α represents the baseline mean current price for Metro if $x_{2i} = 1$; otherwise when $x_{2i} = 0$, it represents the mean current price for SaveOnFoods. Also, when old price and month is equal to 0, the intercept is not meaningful. Coefficient β_1 captures how changes in the old price affect the current price. Coefficient β_2 measures the difference in the mean coffee price between SaveOnFoods ($x_{2i} = 1$) and Metro ($x_{2i} = 0$). Coefficient β_3 reflects how the month influences current pricing, potentially capturing seasonal effects.

To implement this Bayesian model, we use the `rstanarm` package (Goodrich et al. 2022) in `R` (R Core Team 2023), with its default priors..

3.1.1 Model justification

The Bayesian Multiple Linear Regression (MLR) model is a suitable choice for analyzing the relationship between current_price (the dependent variable) and the predictors in the dataset. The dependent variable is continuous, and the Bayesian framework assumes a normal distribution for the response, which aligns well with the nature of coffee prices. This model captures the linear relationships between old_price (continuous), vendor (categorical, represented as a dummy variable), and month (numeric). These predictors are assumed to have additive effects on the response, which fits the linear regression framework. Logistic regression is used when the outcome variable is binary (e.g. 0 or 1). However, in our dataset, the dependent variable, current_price, is continuous. Since logistic regression cannot model continuous outcomes, it is unsuitable for this analysis. Also, poisson or negative binomial regression is typically applied when the response variable represents count data (e.g., the number of events occurring in a

fixed period). current_price does not represent counts but rather continuous pricing data. Thus, these models do not align with the nature of the dependent variable.

3.1.2 Model validation

4 Results

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

5 Discussion

5.1 First discussion point

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

5.2 Second discussion point

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

5.3 Third discussion point

###Limitation lack of month: only have data of coffee prouduct pricing in June to November. data

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Table 3

	Coffee_product_pricing
(Intercept)	−0.41
old_price	0.77
month	0.06
vendorSaveOnFoods	−0.61
Num.Obs.	2954
R2	0.900
R2 Adj.	0.900
Log.Lik.	−5838.034
ELPD	−5843.5
ELPD s.e.	71.3
LOOIC	11 687.0
LOOIC s.e.	142.5
WAIC	11 687.0
RMSE	1.75

Model summary of Coffee product pricing

Appendix

.1 Model details

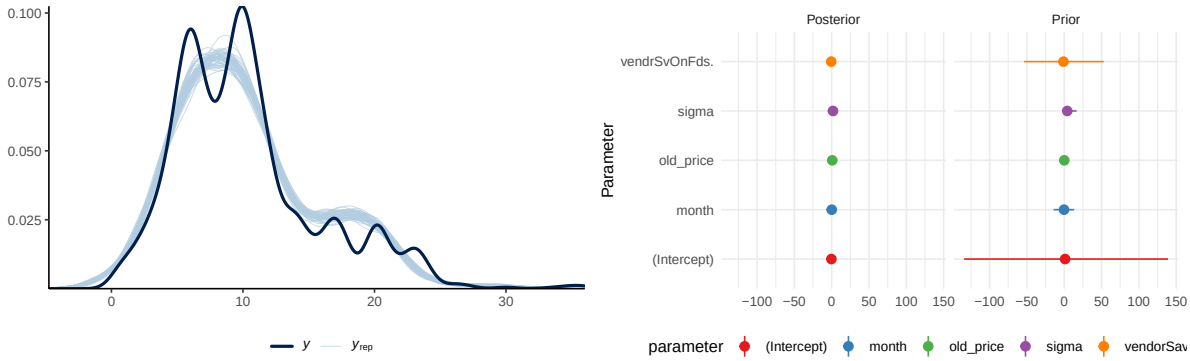
A Additional data details

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…



(a) Posterior prediction check

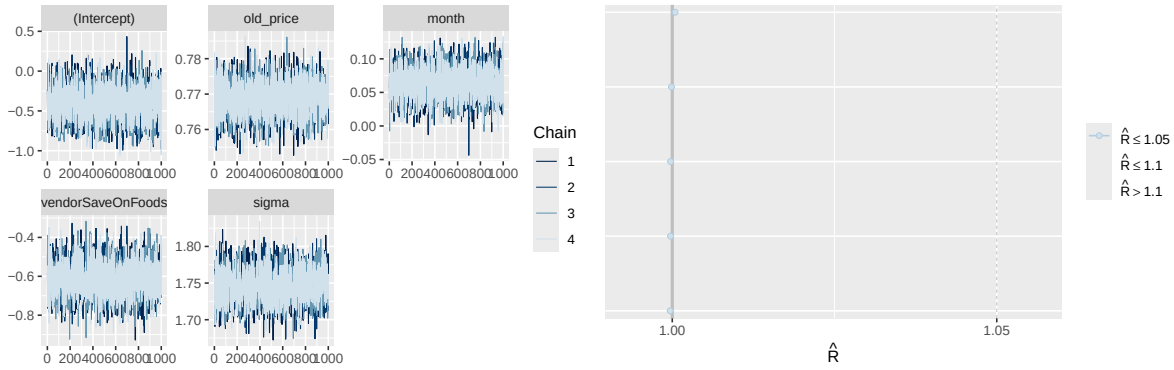
(b) Comparing the posterior with the prior

Figure 3: 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...



(a) Trace plot

(b) Rhat

Figure 4: Trace and R-hat plot

References

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

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