Certain Bundles of Products Tend to Get Marked-Down More During Sales.*

We compare each product's price data for each vendor over time.

Talia Fabregas Lexi Knight Fatimah Yunusa
Aliza Mithwani

November 14, 2024

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

Table of contents

*Code and data are available at: [https://github.com/alizamithwani/Grocery-Prices.git).

1 Introduction

This paper uses data from Jacob Phillip's website for Project Hammer which aims to reduce collusions and increase competitiveness in the Canadian Grocery sector. The dataset reveals information about the prices over time as observed from the timestamps of when the data was collected, which can be used to deduce how frequently items went on sale. The old (before-sale) price as displayed on the product can be compared to the new price to determine how much each product has been marked-down by for each vendor.

Estimand paragraph

Results paragraph

This is an important topic as it helps new vendors understand which products they should be marking-down for sales and the reasoning behind this. Governments can also make better deductions and predictions about which types of households are most affected during times of deflation or when prices falL based on the typical basket of good they purchase and which frequently marked-down product(s) are included in this. Pricing strategy is a complex topic and could vary between vendors, but if the same few items are consistently marked-down during sales, there must be a deeper-rooted reasoning behind it.

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

2 Data

2.1 Overview

This dataset includes grocery price information gathered from various Toronto vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. Prices were collected for products under the "in store pickup" option in a Toronto neighborhood. The data was screen-scraped directly from website user interfaces, so some information that may be accessible via internal APIs is missing. Initially, data collection focused on a "small

basket" of products, from February 28 through early July, after which it expanded to cover a larger variety of items. This structured dataset allows us to track price changes over time across multiple vendors, offering a foundation for comparative analysis of pricing behaviors and promotions.

However, the dataset does have some limitations due to its collection method. Notably, data for certain vendors or specific days may be absent if extraction processes failed, leading to potential gaps in the dataset. Additionally, product identification relies on vendor-specific internal IDs that may change with data refreshes, limiting the continuity of analysis over multiple data versions. These IDs remain consistent within a single vendor for any given download, though they are not universal across vendors. Understanding these limitations is important for accurate interpretation, especially when comparing price behaviors across different vendors.

2.2 Measurement

The data is structured with columns capturing details such as timestamps, vendor information, product ID, product name, brand, units, and pricing details. Timestamps mark when prices were gathered, enabling time-series analysis of price trends. Product names often include brands or units, while separate columns provide brand names and units explicitly, though these may be blank for some vendors. Each record includes a "current price" at the time of extraction, an "old price" indicating a struck-out former price to signal sales, and a price per unit, though this unit price may occasionally differ from an actual division of price by units.

Each product entry is linked to a unique ID within the vendor's system, allowing consistent price tracking over time. However, these IDs are not stable across dataset versions, meaning that new downloads may introduce different IDs, complicating long-term tracking. In cases where brands are represented differently across vendors (e.g., "Miss Vickies" with variations in apostrophe placement), SQL queries are used to account for these inconsistencies, ensuring comprehensive analysis across vendor representations.

2.3 Outcome variables

The primary outcome variables in this study are the frequency and magnitude of markdowns across products for each vendor. The "current price" and "old price" columns capture direct price changes, providing insights into markdown events. The presence of an "old price" generally signifies a sale, allowing for differentiation between explicit sales events and subtler price fluctuations. By observing trends in these variables, it's possible to quantify markdown frequency (how often a product is marked down) and magnitude (the amount by which prices are reduced) across different vendors.

This setup also allows for exploring whether specific product bundles are subject to more frequent or substantial markdowns. By examining various combinations of products and vendors,

patterns in promotional pricing can be identified, contributing to a broader understanding of sales strategies across the grocery market. This analysis is particularly valuable for identifying whether vendors follow similar or different markdown practices, highlighting competitive pricing behaviors within Toronto's grocery industry.

Some of our data is of penguins (Figure 1), from Horst, Hill, and Gorman (2020).

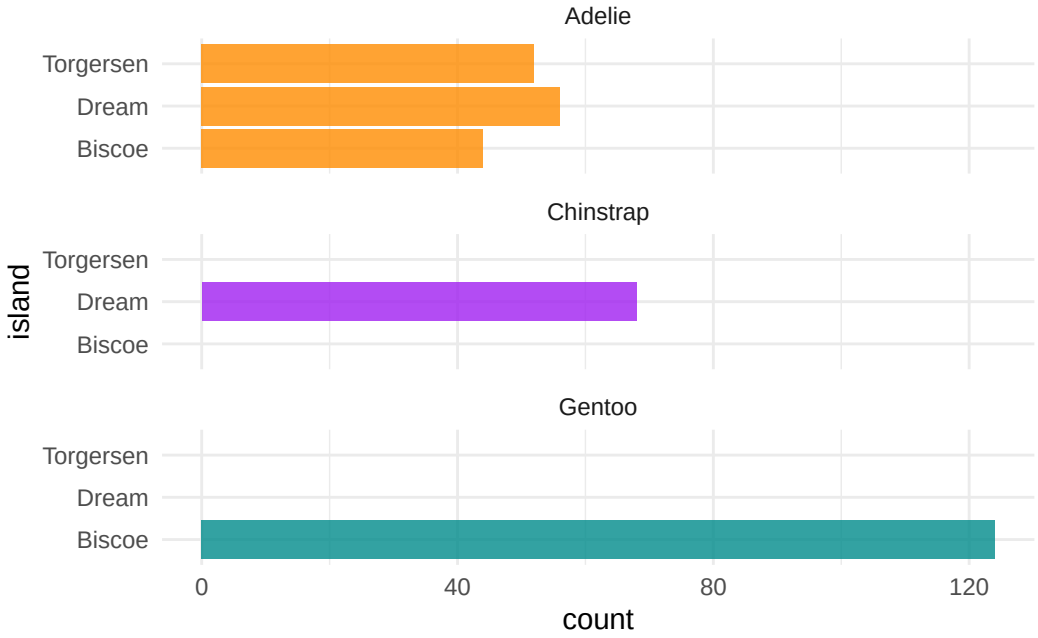


Figure 1: Bills of penguins

Talk more about it.

And also planes (Figure 2). (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.

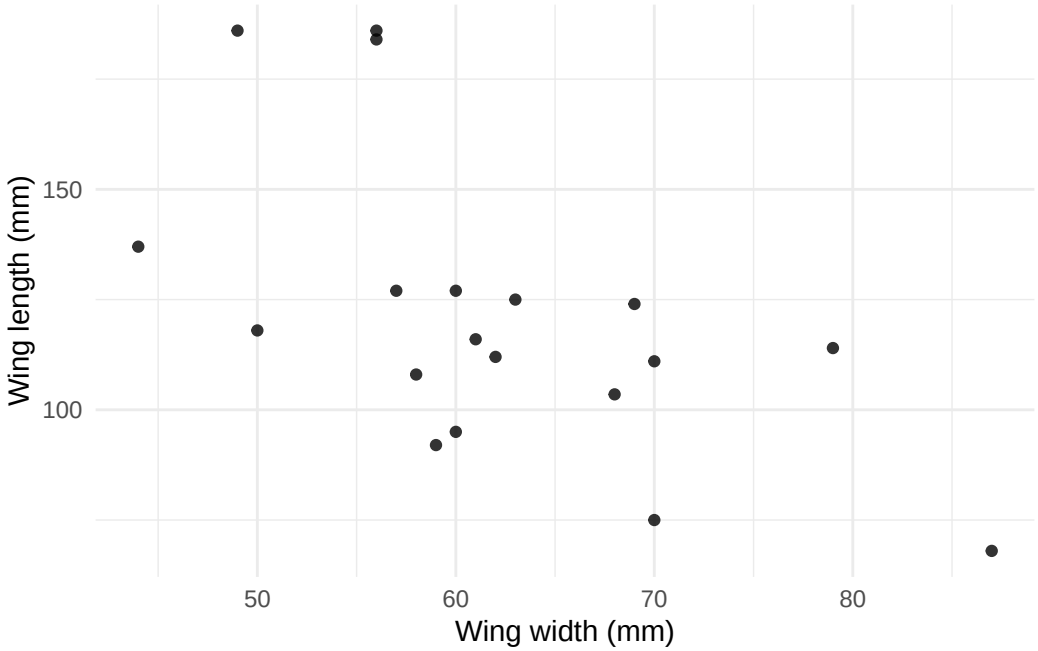


Figure 2: Relationship between wing length and width

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

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`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 1.

5 Discussion

5.1 Correlation vs. Causation

Understanding the distinction between correlation and causation is essential for accurate analysis and responsible decision-making. Correlation indicates an association between two variables that tend to move together, but it does not imply that one causes the other. Causation, however, is a cause-and-effect relationship where changes in one variable directly lead to changes in another. Establishing causation often requires rigorous methods, such as controlled experiments, to rule out other influencing factors and confirm the direct link between variables.

Misinterpreting correlation as causation can lead to incorrect conclusions and ineffective policies or actions. By recognizing that correlation does not automatically mean causation, researchers and analysts can avoid drawing unwarranted inferences and instead base conclusions on solid, evidence-backed relationships. Distinguishing between these concepts ensures more accurate data interpretation and more effective decisions across various fields.

5.2 Missing Data

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

5.3 Sources of Bias

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

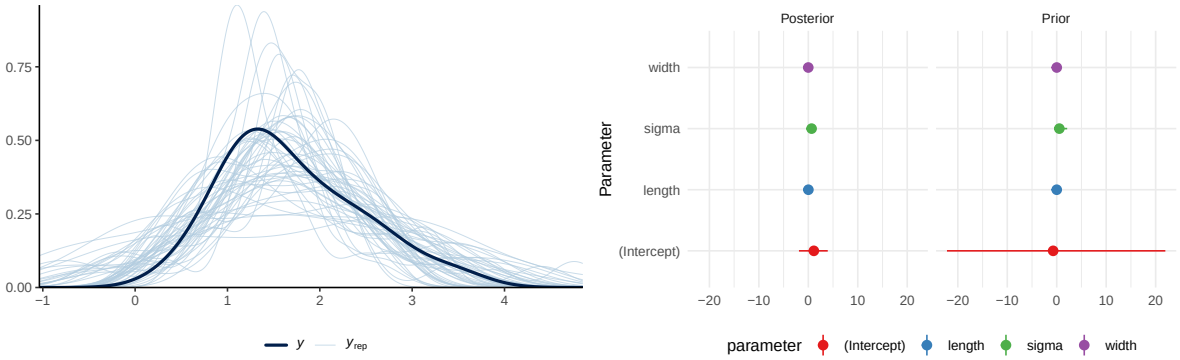
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 3a we implement a posterior predictive check. This shows...

In Figure 3b we compare the posterior with the prior. This shows...



(a) Posterior prediction check

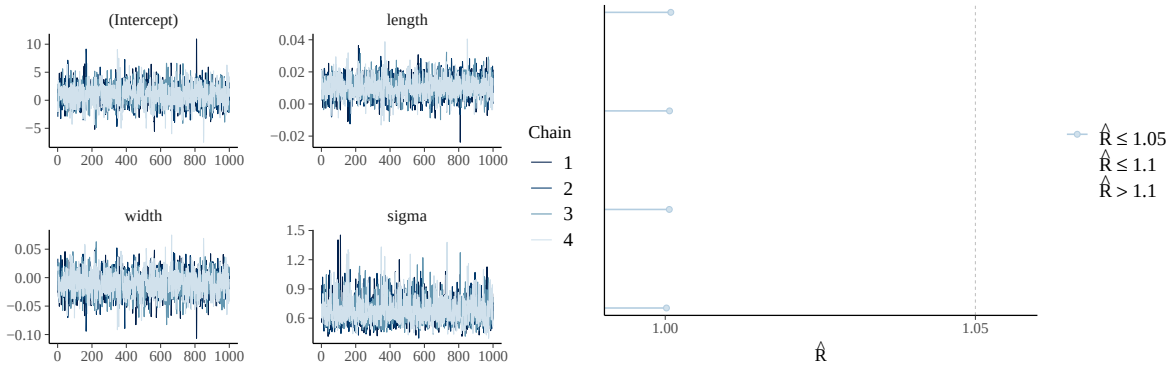
(b) Comparing the posterior with the prior

Figure 3: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 4a is a trace plot. It shows... This suggests...

Figure 4b is a Rhat plot. It shows... This suggests...



(a) Trace plot (b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

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/.