My title*

My subtitle if needed

Cher Ning-Li

December 3, 2024

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

1 Introduction

#TODO: Overview paragraph

The estimand that is targetted within this paper's analysis is the price of a dozen eggs across the 8 top Canadian grocers.

#TODO: Results paragraph

#TODO: Why it matters paragraph

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

2 Data

2.1 Overview

The dataset used for this paper's analysis is titled "Canadian Grocery Price Data" and was obtained from Project Hammer (Filipp 2024), which compiles grocery store price listings across the 8 major Canadian Grocers — Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. The dataset contains data starting from February 28, 2024 and is updated regularly with the addition of new entries. Only the entries up until November 25, 2024 are considered within this paper.

*Code and data are available at: (https://github.com/cher-ning/cndgrocers)[https://github.com/cher-ning/cndgrocers].

The dataset was downloaded as two separate csv files titled `hammer-4-product.csv` and `hammer-4-raw.csv`. The `product` table includes information about specific products that are available across the different grocers, with unique product ID numbers to identify each listing at each grocer. The `raw` table includes the price and more time-specific information regarding each product's listing. The two tables were joined by matching the product IDs to create the full dataset, a comprehensive table where each entry holds all the necessary product and listing information.

The relevant variables of interest to us are: `nowtime`, `vendor`, `current_price`, `old_price`, and `product_name`, as explained below:

- **nowtime**: Date and time of when the data was collected
- **vendor**: One of the 8 Canadian grocery vendors
- **current_price**: Price at time of extract
- **old_price**: An 'old' struck-out price, indicating the item's regular pre-sale price
- **product_name**: Product name, may include brand and/or units

For ease of analysis, these additional variables were derived:

- **month**: The month of when a data entry was collected; extracted from `nowtime`
- **prev_month_avg**: The average price of a dozen eggs across all vendors during the previous month

All entries with empty values in any of these key variables were dropped. Further details regarding the data cleaning process and a glimpse of the first 5 rows of the final cleaned dataset can be found in Section A.1.

Under the guidance of Alexander (2023), the R programming language (R Core Team 2023) was used for analysis of this dataset. The package `tidyverse` (Wickham et al. 2019) was used to simulate data and test the simulated data before analysis. Packages `tidyverse` (Wickham et al. 2019) and `arrow` (Richardson et al. 2024) were utilized to clean the full raw dataset, which were then also used alongside packages `testthat` (Wickham 2011) and `here` (Müller 2020) to test the cleaned dataset. Lastly, packages `tidyverse` (Wickham et al. 2019), `rstanarm` (Goodrich et al. 2022), `arrow` (Richardson et al. 2024), and `here` (Müller 2020) were used to build predictive models using the cleaned dataset.

2.2 Measurement

The observations in the dataset are scraped from the grocer's official website's interface, which means it only contains the information that is publically listed and available. The recorded prices for each item are the listed price for the "in store pick up" option, with the target pickup neighbourhood being a neighbourhood in Toronto. Certain values and listings may be missed, since the internal APIs that power the grocer's websites are not accessed and specific extracts may error for certain vendors on particular days for unforeseeable reasons. Starting from July

11, the targeted variety of grocery items for data collection was greatly increased, meaning that there may be products which are missing pricing information before then.

2.3 Variables of Interest

The outcome variable that we are interested in understanding is the `current_price`, representing the sale price of a dozen eggs. The predictor variables are `month`, `vendor`, `old_price`, and `prev_month_avg`.

Figure 1 shows a boxplot of the sale pricing of a dozen eggs across the 8 vendors. From this graph, we can observe that T&T has the lowest prices overall and Save On Foods the highest. There are also differences in the range of prices offered, with T&T and Metro having much wider price ranges compared to Galleria and Save On Foods, which have highly consistent sale pricing.

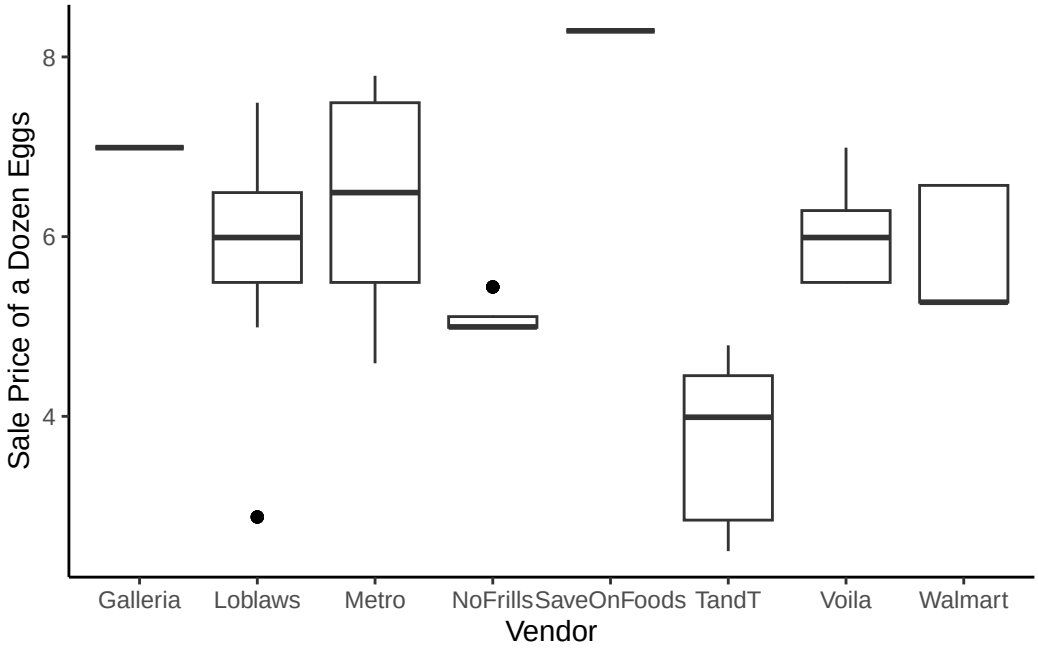

Figure 1: Average Sale Price of a Dozen Eggs Across Vendors

Figure 2 shows the number of sales provided on egg dozens, showing a large variation between vendors. Metro notably had the most egg discounts, and Galleria the least.

Next, Figure 3 shows the average sale price of a dozen eggs across different months. It can be observed that price dips slightly during the summer, reaching its lowest in July, before increasing again. Differences throughout months are typically gradual and do not fluctuate very much.

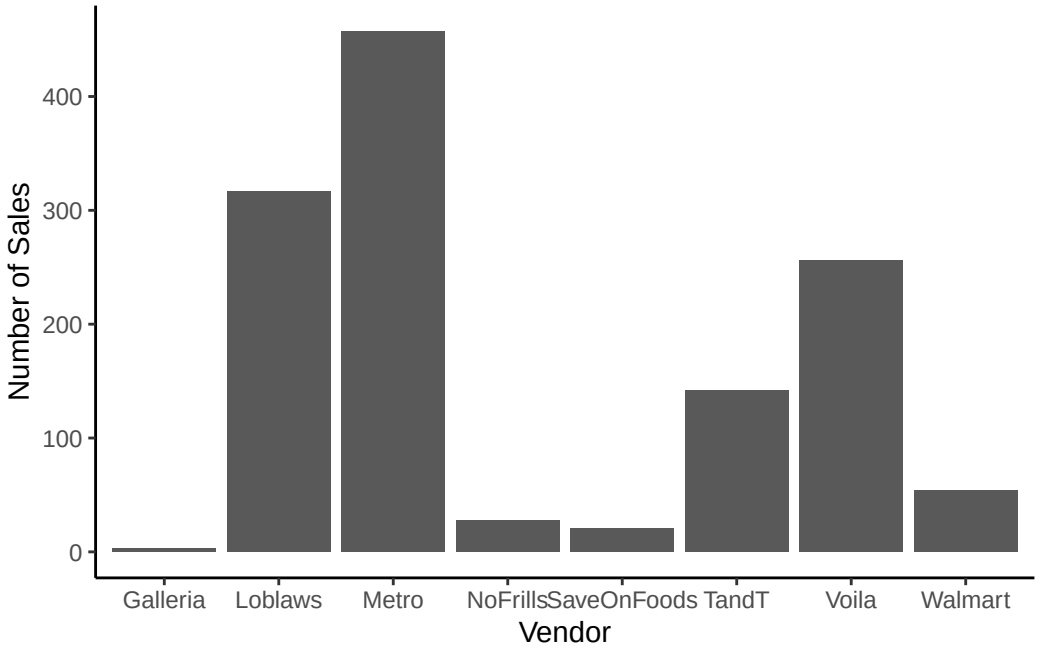

Figure 2: Number of Discounts Provided By Each Vendor On Eggs Between March and November 2024

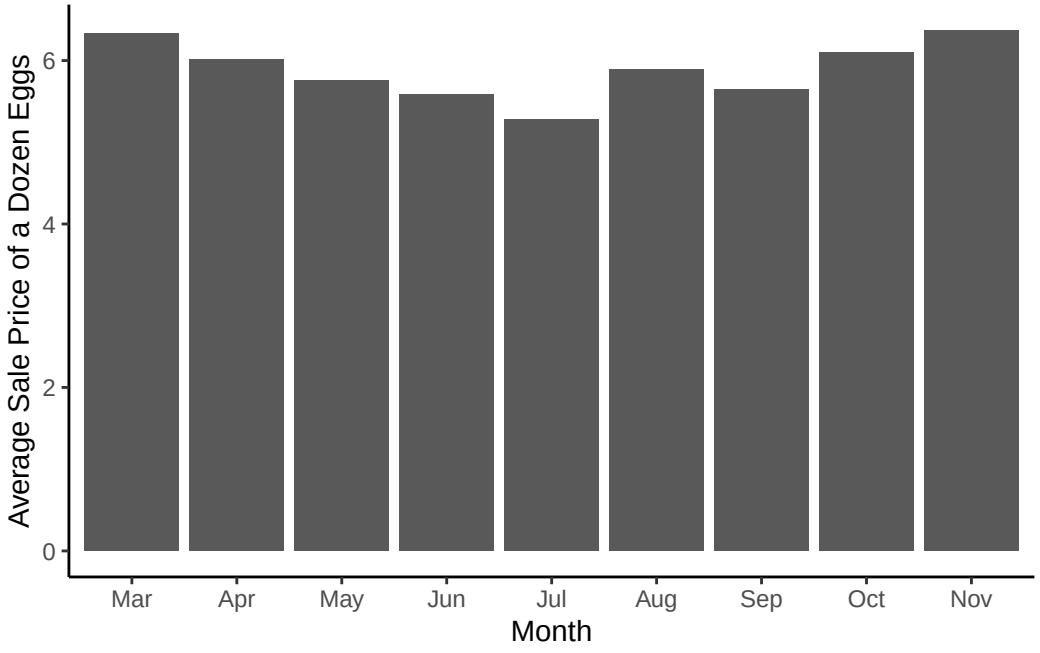

Figure 3: Average Sale Price of a Dozen Eggs Across Months March to November

Figure 4 shows the positive and relatively linear relationship between the original presale price and discounted price of a dozen eggs across all 8 vendors. The dotted line has a slope of 1 and indicates how prices would look if discounted prices are equal to presale prices. All data points are on or below the dotted line, indicating that discounted prices are always equal to or less than the original presale price. The blue line, representing the line of best fit, has a positive but smaller slope than the dotted line, indicating that sale prices increase at a slower rate than presale prices. In other words, the greater an item's presale price, the greater the price difference would be. This confirms our intuitions about how sale events occur in the real world, as sales are typically calculated as a percentage of the original price.

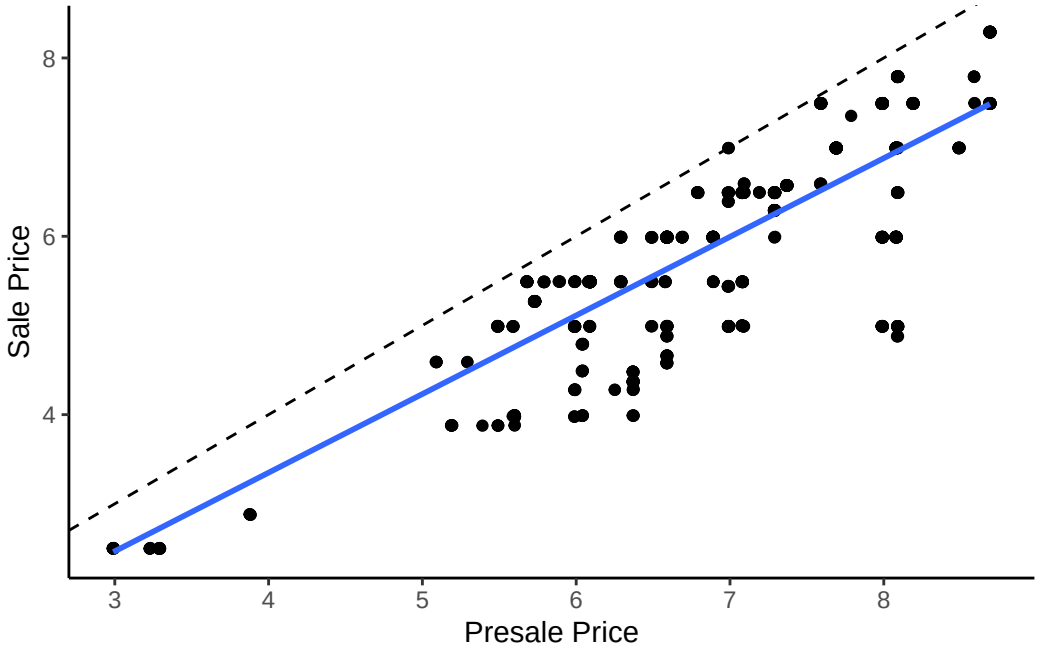

Figure 4: Presale and Sale Prices of a Dozen Eggs

Figure 5 shows the relationship between the current sale price of a dozen eggs compared to the previous month's average price. Similar to Figure 4, the dotted line has a slope of 1 and indicates variables growing at an equal rate whereas the blue line represents the line of best fit. Here, in contrast to Figure 4, the relationship appears much weaker with a greater scatter and deviation of values. The line of best fit has a positive slope that is less than 1, indicating that current sale prices increase when previous month's prices do, but at a slower rate. That is, a higher price average in the previous month is correlated with higher prices in the current month, but not to a high degree.

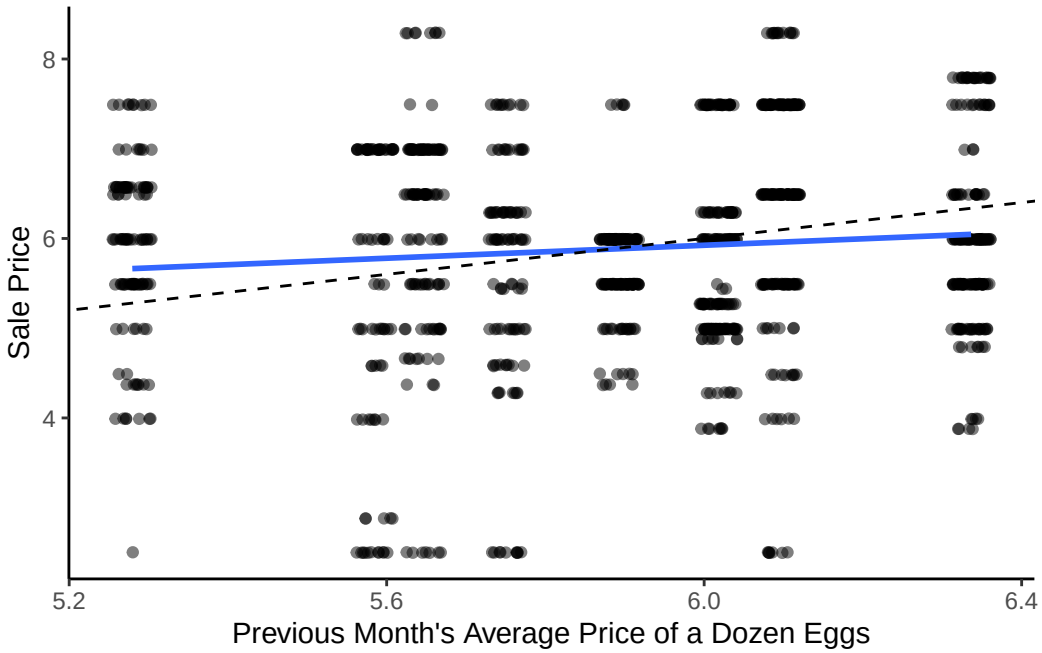

Figure 5: Sale Price of a Dozen Eggs vs Previous Month's Average Price

2.4 Other Datasets

There are other food pricing related datasets available for use, such as Statistics Canada's "Average Retail Food Prices" (Statistics Canada 2024). Though it does also contain average price information for a dozen eggs, this dataset was not utilized within this paper's analysis in the end because it only contains monthly averages that are also not divided by vendor. As well, the averages in this dataset are generalized across the entire provinces, in contrast to Project Hammer's dataset which is centred on Toronto prices (Filipp 2024). This could introduce inaccuracies into our model, as the wider geographical range may be failing to capture Toronto's comparatively higher cost of living, compared to other Ontario regions.

3 Model

The goal of our modeling is to understand how the predictor variables — month, vendor, presale price, and previous month's average sale price — affects the response variable of current price. For this paper, a multiple linear regression model will be utilized.

We define Y_i as the current price, and initialize the full model as: $Y_i = \beta_0 + \beta_1 month + \beta_2 vendor + \beta_3 old\_price * prev\_month\_avg$

We create the model in R (R Core Team 2023), with help from package `tidyverse` (Wickham et al. 2019).

3.0.1 Model Weaknesses

The first weakness of this model is that it assumes a linear relationship between the discounted price and the list of predictor variables. This oversimplication could be missing many aspects of their true relationship if it is nonlinear. As well, there could be real-world events such as economic changes or policy changes which could have significant impact on pricing, but this model would fail to capture these phenomena.

Additionally, due to the targeting of dozen egg items within this analysis, it has also caused our dataset to be quite small with less than two thousand data points. Though this decision to focus in on one type of item helps narrow down the scope of analysis and increase the generalizability of findings, the small dataset size may simultaneously have the opposite effect in decreasing accuracy.

To use this model to predict future egg discount prices, there would also be an added assumption of causality between the predictor variables and the response variable. However, this dataset is only observational data, and therefore it is insufficient to conclude causality from our current analysis. More details regarding the impact and context of this assumption can be found in Section B.1.

Diagnostics that show the model's weaknesses can be found in Section B.2.

4 Results

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. - maybe should try to specify the prev_month_avg down to also selection by vendor

Appendix

A Additional data details

A.1 Data-Cleaning

After combining the two tables, `raw` and `product`, by matching each entry's product ID, the full dataset with $12,753,964$ entries was created. Then, entries with "egg" or "eggs", and "dozen" or "12" appearing in their product name or units were selected for. All February entries were omitted from analysis due to data collection starting on 28 February, 2024, which means there is a lack of sufficient data for that month to make any reliable conclusions from.

All unannounced price changes, including increases or decreases, have only a `current_price` value and empty `old_price` values, so all such entries were filtered out to narrow down analysis onto explicitly advertised sale events only. Then, the `current_price` and `old_price` columns were filtered to ensure all values were non-negative, as these entries are evidently the result of measurement error during data scraping.

After cleaning the dataset and adding in the derived variables, the final dataset of $1,278$ entries is what analysis will be conducted on. The first 5 rows from this analysis data is shown in Figure 6.

current_price	old_price	vendor	product_name	month	prev_month_avg
4.99	5.49	Loblaws	Naturegg Nest Laid White Eggs, Large	May	6.018973
7.49	7.99	Loblaws	Naturegg Omega Plus Solar Free Range Eggs, Large	May	6.018973
4.28	5.99	TandT	Naturegg Large Omega-3 Brown Eggs (12s)	May	6.018973
5.99	6.89	Metro	Large White Free Run Omega-3 Eggs	May	6.018973
4.99	5.49	Loblaws	Naturegg Nest Laid White Eggs, Large	May	6.018973

Figure 6: Sample of Analysis Data

Examining how the model fits, and is affected
by, the data

Figure 7: **?(caption)**

B Model details

B.1 Observational Data

The dataset used is a compilation of price listings collected from grocery store websites directly, meaning that it is only observational data that was compiled without conducting of an experiment. This means that caution needs to be taken when making conclusions of causation from this data, because we can only be certain of correlation unless thorough tests are conducted. Our conclusions can possibly be affected by two common misconceptions, Simpson's and Berkson's paradoxes.

Simpson's paradox is when a subset of data presents a relationship that is different from when the full dataset is considered due to lack of consideration for a confounder variable, which is the true cause of a relationship. Berkson's paradox is when a dataset is so specifically subsetted that patterns found are different from the full dataset (Alexander 2023).

Within this paper's context, both paradoxes are at risk of occurring because our analysis is conducted on the subset of only egg dozen listings. Therefore, when making conclusions, it is important to remember that the patterns of correlation between variables observed within our subset of egg prices cannot be generalized to the full dataset of all grocery store listings. For example, it is possible that though in our analysis, T&T offers the lowest discounted prices Figure 1 for eggs, they could be offering the highest prices when averaged across all product listings.

In the same vein, the lack of significant price change around September could also be because of our sampling for egg items. Due to the school semester starting in September, intuition tells us that there are likely many sales going on for school supplies. Therefore, if we did not limit our search to egg dozens only, it is possible that September could show the lowest average prices instead of July, as shown in Figure 3.

This is one weakness inherent to using observational data, that there are many unknown or uncontrolled variables which could impact our analysis. Many errors can occur from making unsupported claims that are generalized beyond the specific subset where analysis was conducted, so care must be taken when interpretting model outcomes.

B.2 Diagnostics

Checking the convergence of the MCMC algorithm

Figure 8: **?(caption)**

References

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

Filipp, Jacob. 2024. *Canadian Grocery Price Data*. https://jacobfilipp.com/hammer/#FAQ.

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

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Statistics Canada. 2024. *Average Retail Food Prices Data Visualization Tool*. https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2024005-eng.htm.

Wickham, Hadley. 2011. "testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.