Describing Canadian Grocer Bread Pricing*

Luka Tosic

December 14, 2024

This paper creates a descriptive model of bread price for certain bread items found at major Canadian grocers. Analysis was conducted to find how much influence bread mass, the vendor it was purchased at, and the type of bread purchased had on the mean price of bread. From the findings, we predict that bread purchased at Metro has a higher chance of being more expensive than from other vendors, and bread from T&T has a higher chance of being less expensive than from other vendors. The findings of this paper give Canadian shoppers the information to understand some of the factors influencing bread prices, and what vendors to seek out or avoid for cost-effective buying options. Further analysis would look at other mainstay and essential grocery items such as milk, eggs and rice.

1 Introduction

1.1 Overview paragraph

Groceries are a major expense in most Canadian households (), and as inflation grows and wages stagnate (), budgeting for groceries becomes increasingly more important. This paper aims to understand and describe the pricing trends in the cheapest bread and milk options among the eight of major grocery vendors in Canada: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart and Save-On-Foods. This paper will utilize linear regression modelling to create descriptive models of the mean price for bread based on previous pricing trends, the vendor, and other predictors. The motivation behind analyzing bread items is their ubiquity in the diets of Canadians.

*Code and data are available at: https://github.com/LukaTosic09/final-paper.git

1.2 Estimand

The estimand that this paper aims to estimate through linear regression is the mean bread item price per unit by vendor. As stated, a multiple linear regression model will be utilized as an estimator for this estimate. Linear regression naturally captures the variability or noise of a statistically linear relationship and attempts to explain this variation through a functional relationship. Through a mix of regression methods, a best model will be fit for each vendor.

1.3 Results paragraph

In short, Metro's branding carries a slight factor into bread price, with their and Loblaw's increasing mean bread price the most. It was also found that Voila and TandT bread increased mean bread price the least. Another finding was that whole wheat and especially white bread decreased mean bread price compared to sourdough.

1.4 Why it matters paragraph

The need here is for Canadians to gauge broadly the affordability of groceries across common vendors. The hope is that the results will be meaningful enough that potential shoppers would be able to know where to go for more affordable groceries, an endeavour that becomes more necessary as groceries continue to grow into more of an expense.

The rationale behind analyzing bread prices is that they are commonly used grocery items in many Canadian households. Moreover, bread is one of the simplest and most symbolic foods eaten in the Western world, and is a fitting starting point when doing pricing analysis on groceries.

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

The data section outlines the predictor variables chosen to fit the models and their shape. It provides context on the dataset as a whole, as well as how it was procured and turned into data. It also covers the checks done on the data and aspects of the data cleaning.

The model section goes over the process by which the model was fit, the checks on regression assumptions and how violations were minimized. It also covers tests on the statistical significance of the model and its goodness. Broadly, it covers all elements of the final model's creation.

The results section showcases the results of the model fitting process and the information it offers us.

The discussion section contains an in-depth look at the results and interprets them to come to conclusions about the estimand, as well as other practical conclusions that may be drawn.

It also touches on the limitations of the model, what it cannot be used to tell us, and how the data and model can be refined to gain further information.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to perform our analysis and produce our model, in addition to Quarto to create this paper. Our data (Filipp 2024) is taken from the Project Hammer website, by Jacob Filipp. From Jacob himself on his website, "Project Hammer aims to drive more competition and reduce collusion in the Canadian grocery sector." Following Alexander (2023), we consider the five-step process of telling a story with data: we plan an endpoint, simulate data for it, acquire and explore real data, explore and model data, and finally share and discuss our results.

2.2 Measurement

The observations in the dataset are all individual grocery items from one of the eight major vendors mentioned above. According to (Filipp 2024), the dataset is updated with observations taken by screen scraping the website UIs of the vendors' grocery store websites. As such, all grocery information is already made publicly available on the internet by the vendors. Another point made by (Filipp 2024) is that the prices recorded are made by setting an "in-store pickup" option for some Toronto neighbourhood. This is relevant to note as what neighbourhood each entry's price refers to is not included in the data.

The API scrapes website UI, so all the information it gathers is what the grocers make available on their websites; and moreover, not every grocer's website includes the same variables in the same way (or at all), so some pre-processing would have been required to get the observations into the dataset in the same format. Broadly, each observation is a text encoding of the information that a vendor has displayed on their website about a grocery item that they are selling, taken at a specific time from a specific vendor. Since data entries take note of date and time to the second, there are many instances of the same grocery item from the same vendor, recorded at multiple different points in time. This characteristic of the dataset makes it possible to conduct price analysis over time on grocery items, and to take note of a changing status of an item, such as "out of stock", or "sale" in the other column, if any such applicable information is listed on the vendor's page for that item.

2.3 Data Cleaning

The original raw dataset has over 12 million observations and needed to be joined with the product dataset that contains categorical information on each of the roughly 130,000 unique grocery items. While the product data had information such as item id, it more importantly contained detailed information about what each id was, and what vendor it was from, whereas the raw data had the current and old prices, the price per unit, units, and time of data collection for many multitudes of instances of the same id, or the same grocery item.

The id or product_id columns were of particular importance, not only because they allowed for the joining of the two initial datasets, but also because multiple instances of the same grocery item with different time and/or date values in the nowtime column could be grouped or viewed as one with the product_id column. The product_name column can achieve the same result, and is useful for searching, but product_id was used in most instances to parse through or select in code as a short one to six-digit id is shorter than any product name.

The product_name column was used to filter out any grocery items that were not bread, including any bread-adjacent items like bread mix or breadcrumbs. As well, the product_name string was used to filter for only bread items that were of the type "white", "whole", or "sourdough", after which all bread items received one of those values in the newly created bread_type column. The reason behind limiting the data to only bread items of those types is that it allowed us to remove any observations that were labelled as bread, but were not what one would typically consider when looking to buy a loaf, such as fruit bread which is more of a dessert. Additionally, this cut the number of observations down from roughly 160,000 to roughly 50,000, which is much more manageable to work with.

Next, items whose units were not in grams were removed, so that the units column could more easily be converted to a numeric, and also to filter out any items that were large packs of several bulk bread items, as such observations would be more difficult to work with, and may not follow a typical pricing relationship.

As a final cleaning step, a new dataset was created combining all instances of the same product into one observation, with a mean price of all instances, to be used for the primary linear model as it avoids repeat observations.

2.4 Outcome variables

2.4.1 Response Variable

Our response variable for the linear model is current price in Canadian dollars. It was chosen over price per unit so that mass from the units column could be used as a predictor in the model. If the bread item had a recorded price on sale, that is the value populating this column. The ((**scatterplot-Old-Price-vs-Current-Price?**)) figure shows the relationship between current and old price.

2.5 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.

2.5.1 Nowtime

The first variable of interest is nowtime. This variable is not used in the primary descriptive model, but is instead used to create a plot of price per unit versus time for several bread items. The values stored inside the variable are a date and time of the form yyyy-mm-dd hh:mm:ss. The reason behind not including the variable in the primary model is that we expect an increase of close to 0 dollars for a unit increase in nowtime, even if the variable is transformed to partition time into far larger slices than seconds, that is to say, we do not expect it to affect the mean price of a bread item as the current price value for any nowtime does not change for most items, and for the ones that it does change, it is generally for a very small number of observations of the same product.

2.5.2 Mass - Units

The mass of a grocery item is encoded in the units column. This column importantly does not force values to a specific measurement unit, and in fact, many items have categorical information in this column.

2.5.3 Vendor

The vendor variable is important as it allows us to make conclusions about bread pricing across the various major Canadian vendors. It is a categorical variable with values “Voila”, “T&T”, “Loblaws”, “No Frills”, “Metro”, “Galleria”, “Walmart” and “Save-On-Foods”. This corresponds to 7 levels in the model.

The below model is a barplot of the number of occurences of each vendor.

Galleria carries no bread items marked in grams, that are a sort of white, whole wheat, or sourdough bread. This is unsurprising considering it is an East Asian vendor, where bread is not a staple food. We also see Metro dominates the other vendors in total bread varieties, with only Walmart being able to compete in total variety. Additionally, we see that Walmart carries many more varieties of white bread proportional to the total number of bread items it carries in comparison with other vendors. Broadly, we can say that most vendors carry roughly the same proportion of white and whole bread items, about a 3/7 each, with sourdough accounting for a third to a quarter of what white or whole bread account for individually.

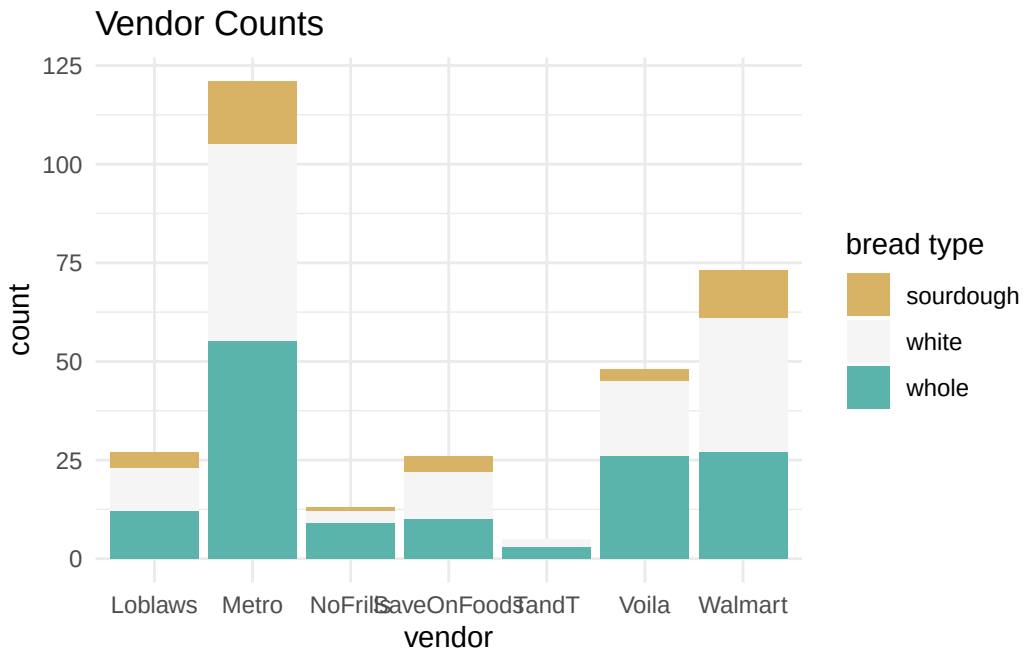

Figure 1: fig-vendor-counts

2.5.4 Bread Type

This is a categorical variable with three levels, the derivation of which from the raw data has been explained above. It contains the levels "white, sourdough, and whole".

Though white bread may be commonly thought of as the most common bread type, the data says otherwise. The likely reason is that whole wheat and whole grain bread occupy many different categories as a label, so we may not consider all bread products that have the word "whole" in their name as belonging to one category. We also see that sourdough is far less common than the other two, with around a third as many entries in the data as whole wheat bread, matching our previous plot.

2.5.5 On Sale - current vs old price

The other column includes information we are mostly uninterested in, expect the term "SALE" which is always present if and only if the old_price column has a numeric value populating it. The old price of an item then is the price prior to it going on sale.

Below is a scatterplot of the linear model with old price as a predictor and current price as a response. The points are bread items that have had their price recorded on and off sale.

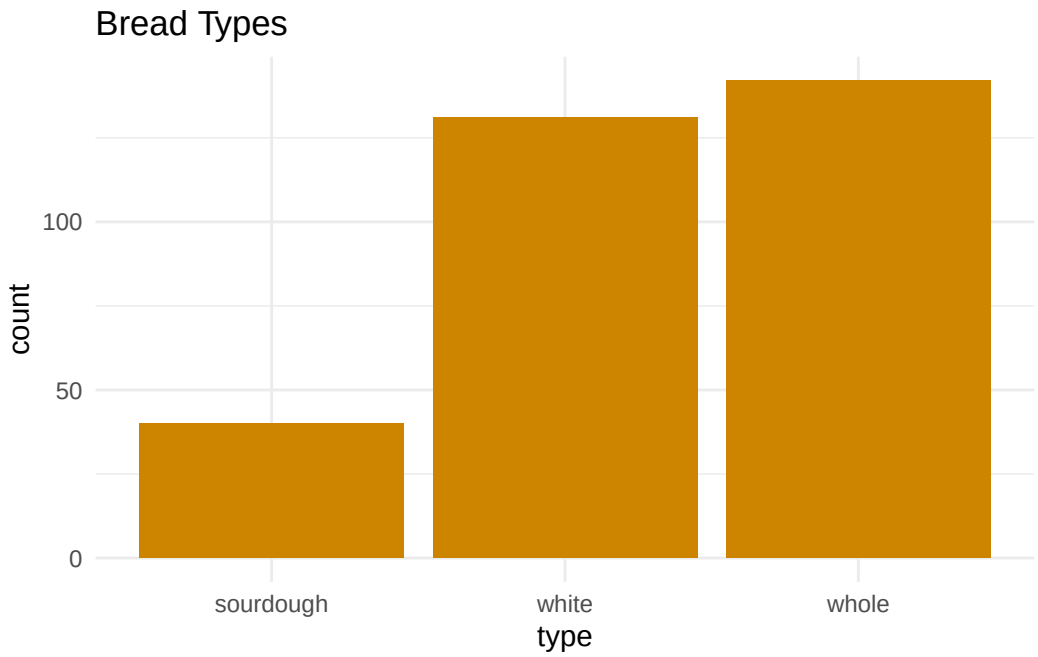

Figure 2: fig-bread-counts

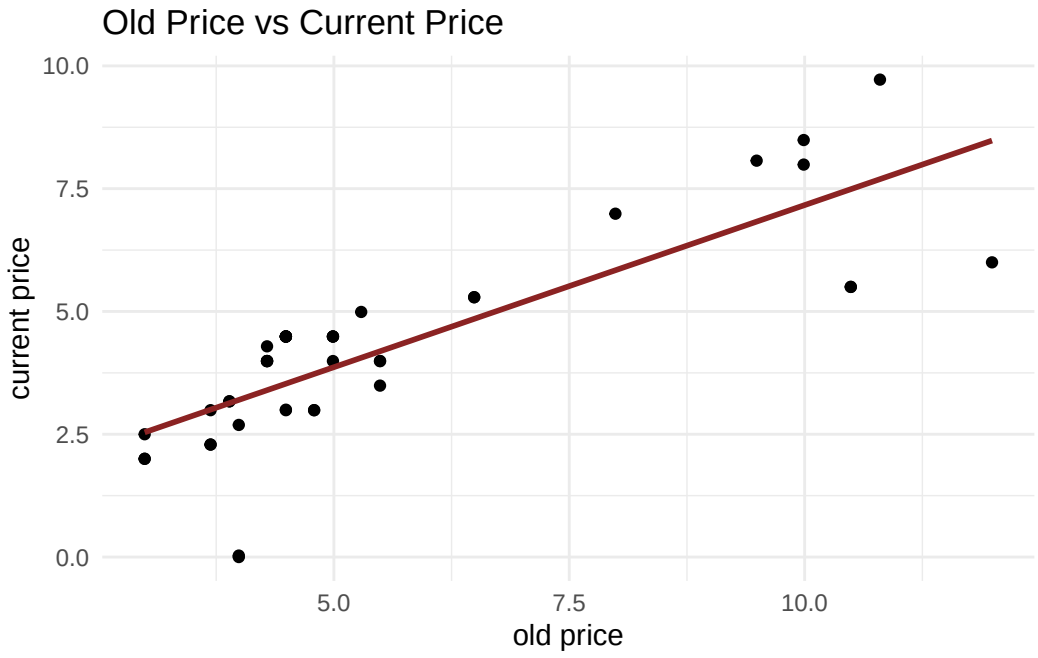

Figure 3: fig-price

The result is unsurprising as sales generally reduce price by a percentage, meaning that we can expect a there to be a positive linear relationship between old and current price.

3 Model

The goal of our modelling strategy is to create a descriptive linear regression model that captures to a degree found sufficient, the true relationship between the response variable and the predictor variables. In this case, we compare two models. A more complicates one having the mean price of bread in dollars as its response with predictors of mass in grams, bread type, the vendor selling the bread, and the pre-sale price in the event that the bread item is on sale compared to a simpler model where we are attempting to explain the same response variable with only vendor and mass as predictors.

3.1 Model set-up

The descriptive linear regression model we will use follows the typical form $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ to estimate the functional relationship $\mathbb{E}[Y|X_1, X_2, X_3, X_4] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ that captures the the trend of the statistical relationship $Y = X\beta + \varepsilon$ given in matrix form.

X_1 is mass in grams, X_2 is vendor, X_3 is bread type, and X_4 is old price.

We run the model in R (R Core Team 2023).

4 Results

Our results are summarized in Table 1.

what does this show broadly and briefly. Use 302 terminology

The baseline level of the indicator variable for vendor is Loblaws. The model summary indicates that all t-tests on predictors have p-value such that we should reject the null hypothesis that their coefficients are zero, meaning that our predictors are statistically significant, with the exceptions of NoFrills and TandT as indicators; however, we would still keep these in our model to preserve the context of the analysis. The baseline for bread type is sourdough

We can say that we expect a mean bread price of Loblaws sourdough loaf to have a price of approximately 5 dollars when at a mass of zero grams, and we expect an approximately -0.003 dollar increase in mean price for a 1 gram increase in mass. Similar results can be read for each other category of bread and vendor. The negative coefficient on total_units seems contradictory, and suggests that we have reason to consider another model due to the fact that

Table 1: Explanatory model of bread price based on vendor, mass, bread type, and old price

	First model
(Intercept)	3.173
	(1.544)
total_units	−0.002
	(0.002)
vendorMetro	−0.347
	(1.301)
vendorSaveOnFoods	−0.372
	(1.374)
vendorTandT	−0.202
	(1.546)
vendorVoila	−1.474
	(1.381)
vendorWalmart	1.286
	(1.701)
bread_typewhite	−0.653
	(0.699)
bread_typewhole	−0.375
	(0.648)
old_price	0.583
	(0.082)
Num.Obs.	44
R2	0.724
R2 Adj.	0.651
AIC	148.1
BIC	167.7
Log.Lik.	−63.041
RMSE	1.01

we are unlikely to predict a decrease in price of bread for higher mass. As such, refitting or transforming the relationship to produce a more contextually sound model is advisible.

From the coefficients of the vendor types, we see that Metro and the baseline Loblaws bread increas mean price the most, and Voila and TandT increase it the least, with Walmart and NoFrills also not increasing it as drastically as the two premium vendors. Ironically SaveOnFoods indicates that one would not particularily be saving on bread when shopping with them. We also see that whole wheat and especially white bread decrease mean bread price compared to sourdough.

5 Discussion

5.1 Bread Affordability

The results mostly speak for themselves. Having a negative coefficient for total_units means we expect mean price to decrease as total_units increase, which is contradictory. This would warrant a refitting of the model, but that is not a world we live in. As long as this model exists and is being used, what can we learn about price and mass of bread? Well, the conditional nature of regression makes it difficult to say that that is the true relationship between the total_units predictor and price response, I can guarantee that isolates, they would have a different relationship, but because of the way this model is fit and because of the specific data that is observed, we do not see a relationship we would expect. A suspicion I have is that this is due to the fact that most of the observatins have a mass of 400-500g, so there is not much deviation in the total_units, nor is there in price itself. As a result, the other predictors explain the variation in price better when all fit into one model.

The vendor type predictors work out quite nicely, with there being a fairly expected order of coefficient magnitudes for each. The only surprises for me are SaveOnFoods and NoFrills, which are generally more affordable options, but not according to the data we had the model fit.

5.2 Bread Types

Unsurprisingly, sourdough is the most expensive, followed by whole wheat, and followed by white bread. The gap between whole wheat and white bread is not large though, which is also not a surprise.

5.3 Real Value of Sales

Based on **?@fig-price**, old sale price has a near linear relationship with price on sale. Nice to know that vendors are generally sticking to percentage based sales.

5.4 Weaknesses and next steps

A weakness of the model is that it is checked with an ANOVA test of overall significance for statistical significance. Moreover, no residual plots are checked to determine if the linear model is following the assumptions of linear regression. A significant weakness of the modelling strategy is that no alternative models were considered. A mean price decreasing with higher mass of bread does not make sense in the context of the data. As such, all results should be considered with a grain of salt, and no definite conclusions should be taken from this analysis.

Additionally, it is important to recognize that the Canadian economy is currently in tumult, with grocery prices still being unstable [] for over a year. This analysis did not consider any prices pre 2023 and makes no conclusions about the changing trends in grocery pricing.

A next step would be performing a similar analysis for other staple groceries such as eggs, rice, and milk. with such analysis, a more comprehensive picture can be had of grocery pricing for necessities in Canadian grocery markets. Despite the fact that Canadians do as a whole consume large amounts of Bread per capita, many Canadians have rice as a consistent carb in their diet, with the country consuming more rice per capita than any other North American nation according to []. In the same vain, eggs and milk are both dietary staples that make up much of Canadians' diets, and are also unaccounted for in this analysis of grocery purchasing power.

A thorough analysis of Project Hammer's data could produce similar findings about these other key grocery store items.

A Surveys, Sampling, and Observational Data

This appendix will discuss aspects of Project Hammer's (Filipp 2024) sampling, data scraping and data collection in order to provide criticsm and assess how conclusions can be made from the data.

Recall from the measurement section that each observation of this dataset is a capturing of some of the information displayed on a vendor's website of a specfic grocery item. When this information is turned into an observation, the time at which it was captured was noted, as well as the current price of the item, an old price if present on the page, the the size in some units, the product name, brand, and vendor, as well as a column called other where certain relevant information present on the webpage is stored. Each unique grocery item is given an id to match it to other observations of the same item captured at different times.

A.1 Ethical Considerations

To create an entry into the full dataset, an API credited by Project Hammer to (**groceryapp?**) is used to scrape the website UI of the major vendors mentioned in this paper. As per the HACKING.md file in the (**groceryapp?**) github repository, it is clear that the creator(s) of this API web-scraping tool did not gain permission to scrape the websites of these vendors. Project Hammer's own page says that part of their mission is to "reduce collusion in the Canadian grocery sector", indicating that we should not be surprised at the fact that (Filipp 2024) is not concerned with abiding by the traditional codes of ethics around data scraping as outlined in chapter 7 of (Alexander 2023). My perspective on this is that it is largely not an issue seeing as such a project does not harm any individuals within these large corporations or expose them in any way, nor does it interfere with these businesses' ability to conduct their business. Based purely on principle, while it would be ideal to gain permission in all instances that a scraper is used, why should independent developers care when their mission goal is already opposed to these businesses in the first place? After all, (Filipp 2024) is trying to provide clarity about the grocery market in Canada, something that corporations like Walmart are actively having their own developers make difficult to do by creating anti-botting measures according to (**groceryapp?**). How much of Walmart's anti-botting practices are driven by attempting to thwart independent do-gooders trying to unmask their pricing structures, versus trying to protect their internl networks from cyber attack can't be known, though I would wager it is both.

A.2 Technical and Practical Considerations

The first thing one notices about the full raw data created is its sheer size. 0.63GB for the hammer-4-raw dataset of each observation that must be joined with the 0.1GB raw dataset of each unique grocery item is massive. The raw dataset has over 12 million observations as of

November 23rd 2024, with its size constantly bloating as new additions to the dataset are mad and no indication on the Project Hammer website if certain observations from a cutoff time are deprecated. It is not indicated anywhere, but it seems from observation of the nowtime column of the data that new observations for any given grocery item are added one to two times per day, though it is not consistent. Some part of this is likely due to the specifics of the API scraper, though nothing is mentioned in the documentation on the repository.

This large filesize poses a challenge that is linked acutely to the nature of the dataset. This data allows its users to create time analysis of grocery item prices, to this end, the specifics of the grocery items must be recorded at various points in time as data entries. This makes the data very dense, but also difficult to parse through, and potentially difficult to render (RStudio crashed not just on my machine, but a much stronger one I borrowed as I tried to clean the data). This tradeoff is a necessary one on some level for such data to be able to exist. Project Hammer could have the dataset formatted such that each observation is a single grocery item, with a vector of price and time pairs, but this solution is inelegant for a number of reasons; it would require additional work from Project Hammer, the onus of which should not necessarily be one, it would require observations to be modified each time the scraper updates, and csv files cannot store vectors of values as true vectors but as characters instead which would then require additional cleaning and processing on the user end. I believe that (Filipp 2024) made the best decision possible for storing this information about the world as actual entries in the dataset for the purpose of time based analysis.

Another consideration to be made is how to deal with other relevant information about each grocery item. The other column exists to highlight this, which is a rather crude method since it is seemingly only every used to indicate one thing in a character string. For example, an item on sale will have "SALE" in the other column, and will thus have its old_price column populated. If an item is given in unusual units on the website, the other column may indicate "pack of 8 individually packaged _____", and the units column would thus be populated by a string like "8pack" or something similar rather than a mass in grams or whatever other type of unit. The other column does serve to highlight special or unusual observations. The reason behind such a column needing to exist is that not all observations are created equal as a consequence of them being observations of webpages. Each vendor uses its own website with its own method of displaying item information, and not every page includes the same information. The scraper sometimes cannot find the price listed of an item to capture, which is a key value to have for any observation. There is no particular conclusion to be made here, but rather a discussion to be had about how observations are produced by a web scraper such as (**groceryapp?**). It also highlights a key aspect of turning real world phenomena, in this case webpages about groceries, into data; that being that not all information can be cleanly or easily translated as we see with the other column and units just having to be units because not everything is or can be listed in grams. It also makes us aware of missing data, either because the data is truly missing, or because we cannot collect it such as the webpage-UI scraper not being able to identify an item's price.

References

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

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

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