Describing Canadian Grocer Bread Pricing*

Luka Tosic

December 4, 2024

This paper creates a descrpitive model of bread price for certain bread items found at major Canadian grocers. Analysis was conducted to find the most cost effective and cheapest bread options across grocers over multiple months; with _____ from _____ found as the most cost effective and _____ from _____ the cheapest in totality across time. 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 expensive 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 the major grocery vendors in Canada: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart and Save-On-Foods. This paper will utilize linear regression modeling 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

1.4 Why it matters paragraph

The need here is for Canadian's 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 endeavor that becomes more necessary as groceries continue to grow into more of an expense.

The rationale behind analyzing bread prices is that they are common use 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.

1.5 Telegraphing paragraph: The remainder of this paper is structured as follows.

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 what 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 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 procss 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 grocer's 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 labeled 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

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 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.

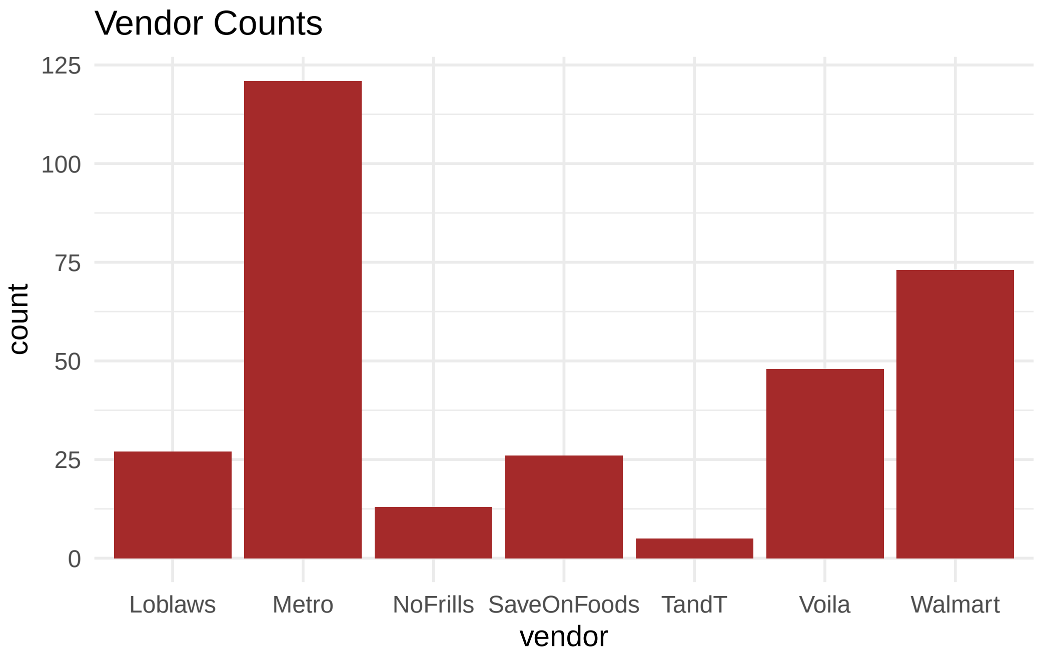
2.5.2 Mass - Units

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.

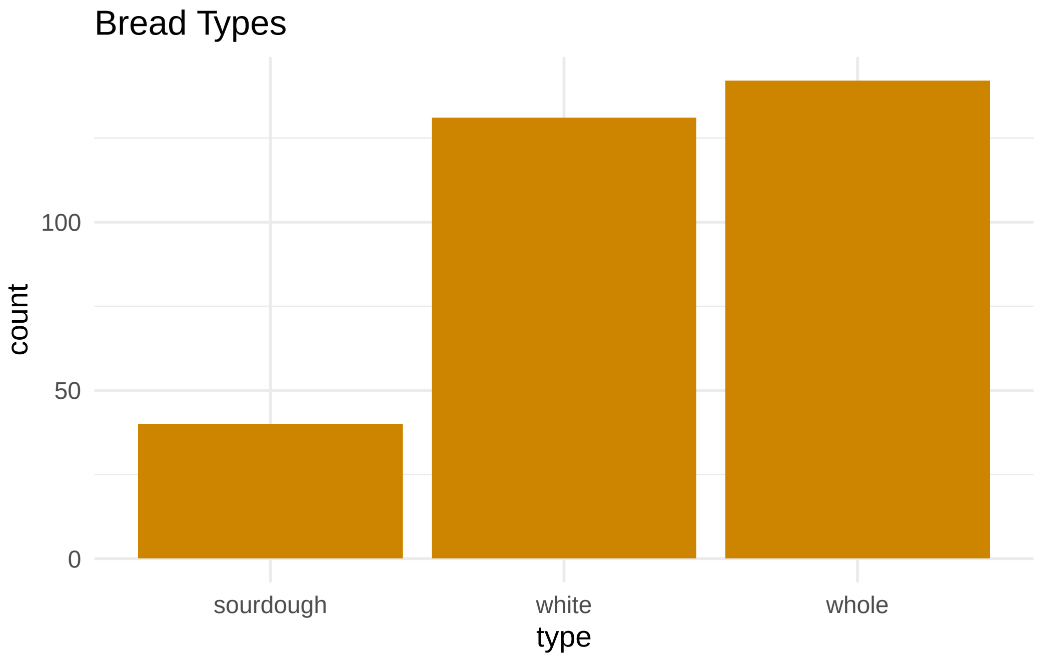
The below model is a barplot of the number of occurences of each vendor.



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. Metro dominates the other vendors in bread varieties

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".

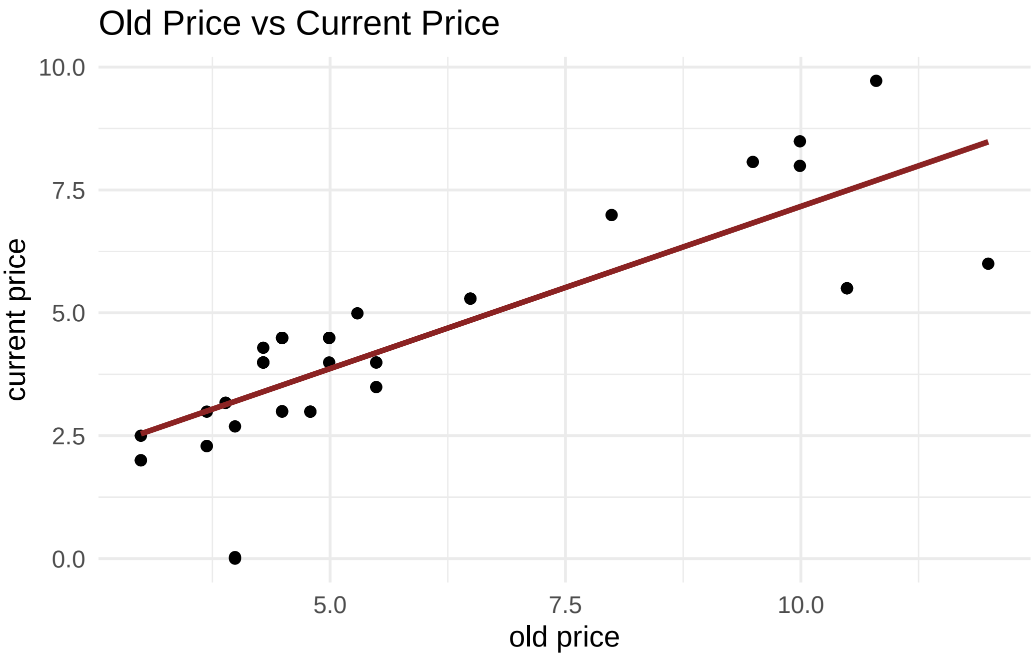


Though white bread may be thought of as the most common bread type, the data says otherwise. Likely, whole wheat and whole grain bread occupys many different categories as a label, so people may not see all whole types of bread as one. As well, sourdough is far less common which is not surprising to anyone who has shopped for bread at a grocery store.

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.



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

> price modeled as a statistical relationship of bread type, on sale status, vendor, and weight in grams, seperate model of old price/unit vs current price/unit that should be a primary model for discussion and results, how much does value increase or decrease with a sale for bread. Maximizing buying opportunity with sale

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.

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.

basic description of model results again. Go into

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

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 valid, that is to say, is following the assumptions of linear regression.

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 nessecities in Canadian grocery markets.

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

A Surveys, Sampling, and Observational Data

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/.