Vendor Price Correlation Analysis

Jacob Gilbert Aman Rana

November 14, 2024

Abstract

This study analyzes the correlations in product pricing between two major vendors using data from Project Hammer, with a focus on correlation versus causation, missing data, and potential biases. The findings aim to better understand competitive behaviors and pricing strategies. We find a weak positive correlation between the prices of 1L of Heinz Ketchup at Loblaws and Voila, suggesting some similarity in pricing trends.

Table of contents

Introduction

The aim of this study is to analyze the correlations in product pricing between eight major vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods, using data from Filipp (2024). By exploring the relationship between the pricing of ketchup (a common household good), we hope to uncover patterns that might help explain competitive behaviors, pricing strategies, and the impact of broader economic factors on retail prices.

This paper specifically focuses on assessing the correlations between the prices of ketchup between two vendors, attempting to discern whether a relationship exists and how strong that relationship is. While correlation can often hint at similar trends or suggest price competition, it is important to exercise caution when drawing conclusions about the underlying causes. For this reason, we dedicate a sub-section to the distinction between correlation and causation, highlighting the dangers of misinterpretation. Additionallly, since we are only looking at two vendors, we discuss the potential biases and limitations of this analysis.

We find a weak positive correlation between the prices of 1L of Heinz Ketchup at Loblaws and Voila. This suggests that while there is some similarity in pricing trends, the relationship is weak, barring on non-existent.

This analysis uses SQlite (SQLite Development Team 2024) for data manipulation and Python (Zelle 2024) to display the correlation.

Data

We begin by creating a csv for our downstream parsing. Observing the database we find 1L of Heinz ketchup has ids 716683 and 1704661. We filter the database of products to just ketchup, drop rows which have a sale price, and then save the table to a csv file. This data is then aggregated to a daily price level by taking average prices across a day. We forward-fill data, assuming that prices are the same until an observation is added into the database.

An example of the data is shown below:

Table 1: Table showing a sample of the vendor-pricing timeseries data

vendor date	Loblaws	Voila
2024-09-23	5.99	6.49
2024-09-24	5.99	6.49
2024-09-25	5.99	6.49
2024-09-26	5.99	6.49
2024-09-27	5.99	6.49

Results

The correlation coefficent when looking at the price of 1L of Heinz Ketchup between Loblaws and Voila is 0.21, this is a weak positive correlation.

Discussion

Our results show a weak positive correlation but there are many limitations, first of the data itself, then of the possible interpretations of the result. The subset of the data we observe is a timeseries of only 52 observations. This is a small sample, and a small period of time. Considering pricing regimes exist, and the complexity of factors that affect price, we cannot generalize this result to a larger timespan. Furthermore, the data is only from two vendors, and we cannot generalize this result to other vendors considering the sparsity of our data.

Below we go into detail about broader limitations of this kind of analysis.

Correlation v Causation

Correlation is a measure of the relationship between two variables, such as milk prices across different vendors. While a high correlation can indicate that prices move in similar ways, it does not imply that one vendor's pricing directly causes changes in another's. Factors like regional supply chain disruptions, seasonal demand, or broader market trends could all lead to similar pricing patterns without any direct causal link between vendors.

To understand causation, one must look beyond mere correlation and consider controlled experiments or advanced econometric models that account for external influences. In this analysis, while we identify price similarities, attributing them to causation would be misleading without deeper scrutiny. One method of examining causation would be difference-in-differences analysis to identify the effect a treatment (such as an increase in competitors' pricing) may have. Thus, it is essential to recognize that correlation highlights relationships but does not inherently explain the reasons behind them.

Missing data

In the context of this study, one significant challenge is the absence of data from other potential sources of the same products, such as smaller local vendors or specialized online retailers. These sources can have a substantial impact on overall market dynamics, yet they are not represented in the current dataset, potentially leading to incomplete or biased insights.

Another important consideration is the absence of competitor pricing data. Competitors not included in the dataset could influence the pricing strategies of the vendors analyzed.

The exclusion of these other sources means that our analysis may overemphasize the influence of the major vendors included in the dataset. Prices at smaller vendors may differ significantly due to factors such as lower overhead costs, niche market positioning, or different supply chains. The absence of this data could lead to a skewed understanding of price correlations, making it seem as though the major vendors are more interconnected than they might be in a broader context.

To address these gaps, future research should consider incorporating a wider variety of data sources to provide a more comprehensive view of the market and improve the robustness of the conclusions drawn.

Sources of Bias

Main biases are likely to arise from the missing data mentioned previously. By focusing solely on major vendors like Voila, T&T, Loblaws, and others, we may inadvertently ignore market forces that influence smaller competitors, leading to an incomplete picture of the overall market dynamics. This selection bias can create a distorted view of price competition and consumer behavior. The unawareness of relationships with competitors allows more selection bias also.
One source of bias could be from unobserved influences on price and sales. For example, advertising campaigns of discounts that influence prices in particular stores. This could lead to observing a downturn in prices in one store that is not fitting with the other stores. This could give our correlation coefficient unreliable results.

Future Work

An extension of this work would be to run crossectional regressions on price with time and product classification fixed effects. This may give us a better approximation of the factors effecting price. Correlation in the residuals of these regressions could give us a better idea of the relationship between vendors.

References

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

SQLite Development Team. 2024. "SQLite Documentation." https://www.sqlite.org/docs.html.

Zelle, John M. 2024. *Python Programming: An Introduction to Computer Science*. Franklin, Beedle & Associates Inc.