Product Price Analysis Using Observational Data*

Analysis of Price Trends, Minimal Correlation with Unit Pricing, and Potential Data Biases

Yizhe Chen Bo Tang Qizhou Xie

November 15, 2024

This report provides an analysis of observational product price data, focusing on current price statistics, price changes, and the weak correlation observed between current price and unit price. The findings indicate that there is minimal linear relationship between these two variables, suggesting that other factors may play a more substantial role in determining product prices. In the discussion, we address the implications of this weak correlation, the impact of missing data on price change calculations, and potential sources of bias that could influence the analysis.

Table of contents

*Code and data are available at: https://github.com/YizheChenUT/Product_Prices.git.

1 Introduction

This report presents an analysis of product price data obtained from the Project Hammer – Jacob Filipp (Filipp 2024). We use the Structured Query Language (SQL) (DB Browser for SQLite Core Team 2024) to process and analyze and apply the statistical programming language R (R Core Team 2023) to make tables and graphs. The primary objective of this analysis is to explore price trends, calculate price changes over time, and examine the potential correlation between current price and price per unit. The results show a very weak negative correlation between these two variables, indicating that changes in price per unit have little to no linear relationship with current price. This observational data, sourced from multiple vendors, provides insights into pricing behavior but also presents challenges, such as missing data and potential sources of bias, which could impact the reliability of the findings.

The remainder of this report is structured as follows. Section 2 provides an overview of the data and details the measurement and calculation of key variables. Section 3 presents the main findings. Finally, Section 4 discusses key issues in the analysis, including correlation vs. causation, missing data, and sources of bias.

2 Data

2.1 Overview

The dataset used in this analysis from the Project Hammer – Jacob Filipp (Filipp 2024) includes key columns such as current price, old price, and price per unit. These columns allow us to calculate basic statistics on product pricing, examine price changes over time, and investigate any potential relationship between unit pricing and overall price. The data was collected from various vendors, providing a range of pricing practices and helping to create a comprehensive view of pricing strategies.

2.2 Measurement

Key variables in this analysis include:

- current price(current_price): The most recent recorded price of the product.

- old price(old_price): The previous price recorded for each product, which is used to calculate the percentage change in price over time.
- price per unit(price_per_unit): The unit price of each product, which could theoretically be associated with the overall current price.

The dataset was filtered and cleaned to remove missing or invalid entries, particularly in the current price and old price columns, to ensure data integrity. This cleaning process included handling missing values and standardizing units where necessary, allowing for a more accurate and meaningful analysis of price trends and correlations.

3 Results

3.1 Basic Statistics

Table 1 presents basic statistics on current price and old price, as well as the calculated price difference between the two. These statistics provide an overview of the price distribution and variations observed within the dataset.

- Avg Curr Price: The average current price of products, calculated as 6.76, indicating the typical price point in the latest records.
- Max Curr Price: The maximum current price recorded, which is 479.99, showing the highest price among the products.
- Min Curr Price: The minimum current price, observed as 0, suggesting there are items without a recorded price or products available for free.
- Avg Old Price: The average old price, which is 8.51, slightly higher than the current average, indicating a general price reduction trend.
- Max Old Price: The highest old price recorded, at 999.5, suggesting a significant price for certain items in the past.
- Min Old Price: The minimum old price, also 0, reflecting products with no historical price or free items.
- Avg Price Diff: The average price difference, calculated as -1.75, which indicates a minor overall decrease in prices over time.
- Max Price Diff: The largest positive price difference of 77.06, showing the maximum price increase for a product.
- Min Price Diff: The largest negative price difference of -979.53, representing a significant drop for some products.

Table 1: Basic statistics of current and old prices

Avg Curr Price	Max Curr Price	Min Curr Price	Avg Old Price	Max Old Price	Min Old Price	Avg Price Diff	Max Price Diff	Min Price Diff
6.755484	479.99	0	8.505236	999.5	0	-1.749752	77.06	-979.53

3.2 Price Change Analysis

Figure 1 shows the percentage change in price for each product where an old price is available. The figure was generated using the tidyverse package (Wickham et al. 2019) in R, which allowed for clear visualization of price change patterns across a large number of products. This line plot illustrates the degree of price variation across different products, with each point on the x-axis representing a unique Product ID and the y-axis representing the corresponding Price Change Percentage.

In the plot, most products show minimal or moderate price changes, while a few products exhibit significant spikes, with some price changes exceeding 2000%. These extreme values may indicate unique circumstances, such as product re-pricing or error corrections. This variability highlights the potential influence of vendor pricing strategies, product demand fluctuations, or seasonal adjustments.

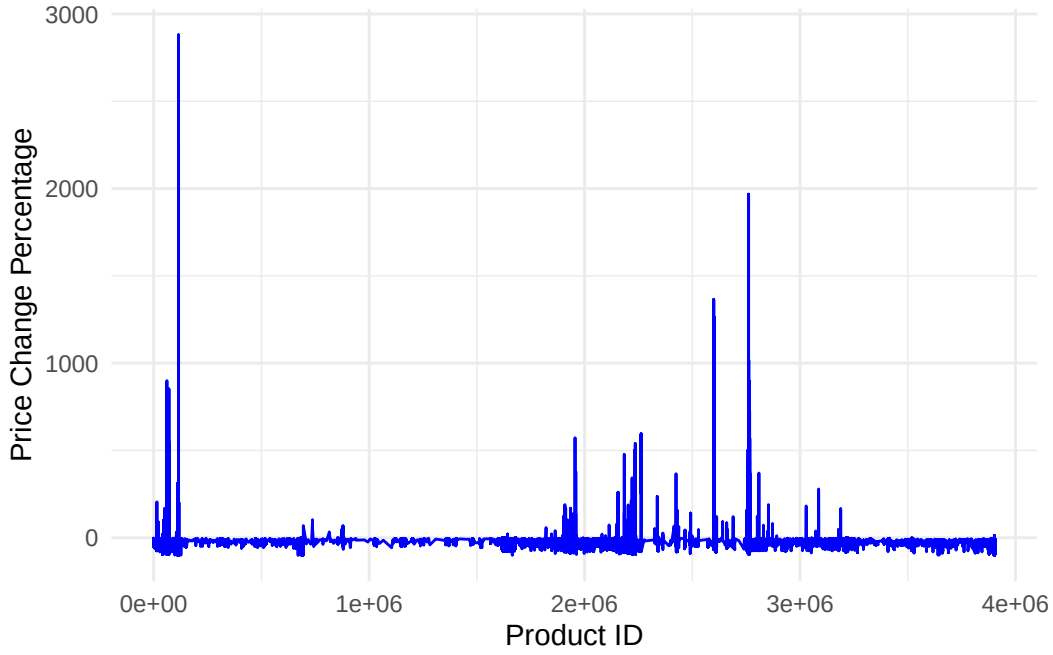

Figure 1: Line Plot of Price Change Percentage by Product ID

3.3 Correlation Analysis

The correlation between current price and price per unit is approximately -0.03256. This indicates a very weak negative correlation, suggesting that there is almost no linear relationship between the current price of the products and the price per unit. In other words, changes in current price are not significantly associated with changes in price per unit based on this observational data.

This weak correlation highlights that while price per unit might theoretically influence current price, other factors could play a more substantial role in determining the product's pricing.

4 Discussion

4.1 Correlation vs. Causation

The correlation analysis between current price and price per unit yielded a very weak negative correlation of approximately -0.03256. This value indicates an almost negligible linear relationship between these two variables, suggesting that fluctuations in current price do not strongly align with changes in price per unit. While this finding might imply that price per unit does not have a significant impact on the current price, it is important to remember that correlation does not imply causation. External factors, such as brand reputation, product demand, or vendor pricing strategies, could play a more substantial role in determining prices. Thus, we cannot conclude that changes in price per unit would directly cause any notable change in current price.

4.2 Missing Data

The dataset includes missing values, particularly in the old price column. These missing data points could potentially skew the analysis by reducing the sample size for certain calculations, such as price change percentage. Missing old price values might indicate discontinued products or errors during data collection. This gap in the data could affect the accuracy of the price change analysis and may introduce bias, as it might lead to an underrepresentation of certain types of products.

4.3 Sources of Bias

Several potential sources of bias in this observational dataset could affect the analysis:

- Selection Bias: The data might over-represent specific vendors or product categories, potentially skewing the findings. For instance, if one vendor has more products listed than others, this could influence the overall average prices or price trends.

- Temporal Bias: Prices fluctuate over time due to factors like seasonal demand, promotions, and economic conditions. Since this dataset represents observational data from specific points in time, it may not accurately capture the longer-term trends or seasonality.
- Measurement Bias: Variability in how vendors report price per unit or current price could lead to inconsistencies. For example, some vendors may round prices or calculate price per unit differently, leading to slight variations that are not reflective of actual market conditions.

5 Conclusion

This analysis provides insights into product pricing trends and variability within the dataset, but it also underscores the importance of context when interpreting observational data. The correlation analysis showed a very weak relationship between current price and price per unit, suggesting that other factors are likely more influential in determining product prices. Additionally, missing data and potential sources of bias—such as selection, temporal, and measurement biases—highlight the limitations inherent in this dataset.

In conclusion, while the data offers valuable observations, these findings should be interpreted cautiously. Future analyses could benefit from a more comprehensive dataset that includes consistent reporting practices across vendors, longitudinal data to capture temporal trends, and efforts to minimize missing values. This would provide a more robust foundation for understanding the true determinants of product pricing.

References

DB Browser for SQLite Core Team. 2024. "DB Browser for SQLite." *DB Browser for SQLite*. https://sqlitebrowser.org/.

Filipp, Jacob. 2024. "Project Hammer." *Canadian Grocery Price Data*. 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/.

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.