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 22, 2024

This report analyzes product pricing data to examine trends, price changes, and the weak correlation between current price and unit price. The findings suggest that factors beyond unit price, such as demand or branding, significantly influence pricing strategies. Understanding these patterns highlights the complexity of marketplace pricing decisions for businesses and consumers.

Table of contents

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

1 Introduction

This study uses data obtained from Project Hammer by Jacob Filipp (Filipp 2024), a large-scale dataset capturing product prices and other relevant attributes. The analysis is conducted using Structured Query Language (SQL) (DB Browser for SQLite Core Team 2024) to process and clean the data, while the statistical programming language R (R Core Team 2023) is employed to generate tables and graphs. The analysis reveals a very weak negative correlation between current price and unit price, suggesting that unit cost has little direct impact on product pricing.

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. Section 4 discusses key issues in the analysis, including correlation vs. causation, missing data, and sources of bias. Finally, Section 5 concludes the report by summarizing the findings gained from the analysis, emphasizing the limitations due to data quality issues, and offering recommendations for future research.

2 Data

2.1 Overview

This analysis uses data from Project Hammer, curated by Jacob Filipp (Filipp 2024), which includes key pricing variables such as current price, old price, and price per unit. These variables enable the calculation of basic statistics, price changes over time, and potential relationships between unit pricing and overall price. The dataset provides a snapshot of product prices from multiple vendors, capturing diverse pricing strategies and behaviors.

While alternative datasets, such as retail pricing indices, exist, they often lack the granularity and cross-vendor scope of Project Hammer. This dataset's unique combination of breadth and detail makes it particularly valuable for exploring how pricing varies across products and vendors.

2.2 Measurement

In this analysis, we focus on three key variables related to product pricing: current price (current_price), old price (old_price), and price per unit (price_per_unit). Each variable represents an aspect of pricing that reflects real-world phenomena, allowing us to investigate trends, changes, and potential correlations in the data.

- Current Price (current_price): This variable represents the most recent recorded price of each product. It reflects the final price at which the product is available to consumers at the time of data collection. This price may be influenced by various factors, such as market demand, brand value, seasonal promotions, and vendor-specific pricing strategies. By analyzing current_price, we can observe the immediate pricing environment for each product.
- Old Price (old_price): The old price is the previous recorded price for each product. This variable enables us to calculate the percentage change in price over time, offering views into price trends and volatility. Tracking these historical prices provides a dynamic view of pricing strategies and changes over time.
- Price per Unit (price_per_unit): This variable captures the unit price of each product, calculated based on quantity or weight. Theoretically, price_per_unit could be associated with the overall current_price, as products with higher unit costs may be priced higher. However, the analysis reveals a weak correlation, suggesting that unit price is not a sole determinant of final pricing. Understanding price_per_unit helps to contextualize price setting relative to quantity or other intrinsic measures of value.

2.3 Data Cleaning and Constructed Variables

The dataset underwent high-level cleaning to ensure accuracy and consistency in the analysis. Key cleaning steps included filtering out entries with missing or invalid values in current_price and old_price, standardizing units in price_per_unit where applicable, and constructing new variables like price_difference to measure the absolute difference between current_price and old_price.

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.

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

- Avg Curr Price: The average current price of products is 6.76, which serves as an indicator of the typical price point for items at the latest recorded time. This low average suggests that many products in the dataset are priced affordably.
- Max Curr Price: The highest recorded current price is 479.99, representing premium or high-value items within the dataset. This outlier highlights that some products deviate significantly from the average, perhaps due to unique market factors or product characteristics.
- Min Curr Price: The minimum current price is 0, indicating that some items may be listed without a price or potentially available for free. This finding points to data entries that could represent out-of-stock items or promotional offerings.
- Avg Old Price: The average old price is 8.51, slightly higher than the average current price, suggesting an overall trend of price reduction. This could be due to seasonal sales, discounts, or adjustments in pricing strategies over time.
- Max Old Price: The maximum old price recorded is 999.5, suggesting a significantly high historical price for certain items, potentially luxury or specialized products. This contrasts sharply with the typical price range and emphasizes the dataset's pricing diversity.
- Min Old Price: The minimum old price is also 0, similar to current price, indicating historical listings of free items or products with missing data.
- Avg Price Diff: The average price difference is -1.75, reflecting a minor but noticeable decrease in prices over time. This result aligns with the trend observed in Avg Old Price and Avg Curr Price, indicating that many products have seen price reductions.
- Max Price Diff: The largest positive price difference is 77.06, representing the maximum recorded price increase for a product. This finding highlights cases where product prices have surged, potentially due to increased demand or supply constraints.
- Min Price Diff: The largest negative price difference is -979.53, showing a significant price drop for some products. This could be due to clearance sales, product downgrades, or significant adjustments in pricing strategies.

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.

- Overall Trends: Most products exhibit minimal or moderate price changes, indicating that the majority of price adjustments remain within a typical range. This is consistent with the dataset's overall trend of slight price reductions.
- Extreme Values: A subset of products shows significant spikes in price change, with some exceeding 2000%. These extreme values may be attributed to unique circumstances such as vendor re-pricing, data entry errors, or seasonal promotions. Such deviations highlight the diverse pricing strategies employed by different vendors and the variability inherent in observational pricing data.
- Variability Across Products: The plot reveals that price changes are not uniform across products. While some products demonstrate steady pricing, others experience dramatic increases or decreases, suggesting differences in market positioning, demand fluctuations, or vendor-specific pricing strategies.

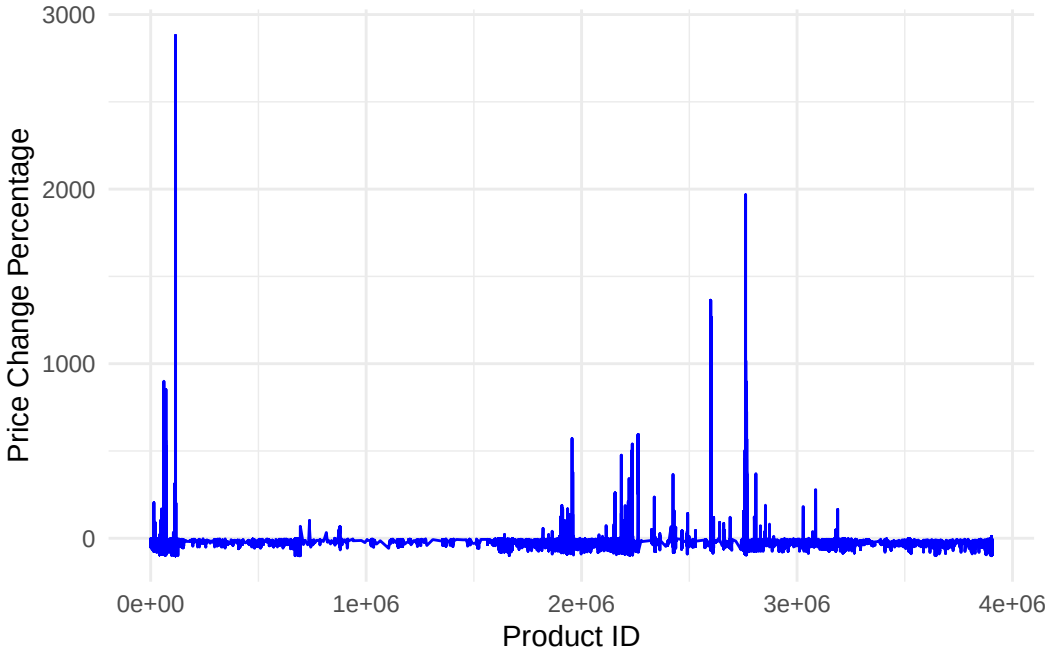

Figure 1: Price Change Percentage by Product ID

3.3 Correlation Analysis

The analysis of the correlation between current price and price per unit reveals an approximate value of -0.03256, indicating a very weak negative correlation. This weak association suggests that changes in price per unit have almost no linear relationship with current price, implying that price per unit alone is not a strong determinant of a product's overall pricing in this dataset.

3.3.1 Interpretation of Correlation Results

- Weak Correlation: With a correlation of -0.03256, the relationship between current price and price per unit is negligible. This implies that the unit price does not substantially impact the final selling price for most products, possibly due to varying factors such as product category, vendor-specific pricing strategies, and demand fluctuations.
- Potential Influencing Factors: The weak correlation suggests that other factors, not captured by unit price alone, likely play a more prominent role in price determination. This could include brand value, product uniqueness, seasonal demand, or targeted marketing strategies that influence current price independently of price per unit.

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 highlights product pricing trends and variability while emphasizing the limitations of observational data. The weak correlation between current price and price per unit suggests that other factors, such as brand or demand, play a greater role in pricing. Missing data and biases further limit the dataset's reliability.

Future research should use more comprehensive datasets with consistent reporting, longitudinal tracking, and reduced missing values to better understand the 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.