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 analyzes product pricing data to understand trends in current prices, changes over time, and the relationship between current price and unit price. The analysis reveals a very weak correlation between current price and unit price, indicating that other factors likely play a more significant role in determining product prices. This finding suggests that pricing strategies may be influenced by variables not captured in simple price-per-unit calculations. Understanding these patterns is essential for businesses and consumers alike, as it highlights the complexity of pricing decisions in the marketplace.

Table of contents

1 Introduction **2**

2 Data **3**

2.1 Overview . . . 3

2.2 Measurement . . . 3

2.3 Data Cleaning and Constructed Variables . . . 4

3 Results **4**

3.1 Basic Statistics . . . 4

3.2 Price Change Analysis . . . 5

3.3 Correlation Analysis . . . 6

3.3.1 Interpretation of Correlation Results . . . 7

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

4 Discussion 7
4.1 Correlation vs. Causation 7
4.2 Missing Data 7
4.3 Sources of Bias 7

5 Conclusion 8

References 9

1 Introduction

Understanding how product prices are determined is essential for businesses, consumers, and policymakers alike. Pricing decisions impact consumer behavior, market competition, and ultimately the profitability of businesses. Despite the importance of pricing, the specific factors that influence product prices are often complex and may vary widely across industries and vendors. This report seeks to contribute to this understanding by analyzing observational price data from multiple vendors, specifically examining the relationship between current product price and unit price, as well as trends in price changes over time.

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 primary objective of this analysis is to identify price trends, calculate changes in price over time, and investigate whether there exists a meaningful correlation between current price and price per unit. A notable finding of this analysis is the extremely weak negative correlation between these two variables, suggesting that unit price has little to no direct impact on the overall product price.

The importance of this finding lies in its implications for understanding pricing behavior. The weak correlation indicates that other factors—such as brand reputation, demand fluctuations, and vendor-specific pricing strategies—may play a more significant role in determining product prices than unit cost alone. This insight is crucial for stakeholders who may otherwise rely heavily on unit price as an indicator of pricing strategy, highlighting the need for more nuanced approaches to price setting and market analysis.

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 to improve understanding of product pricing determinants.

2 Data

2.1 Overview

The dataset for this analysis originates from Project Hammer, curated by Jacob Filipp (Filipp 2024), and contains key pricing information for various products across multiple vendors. The data includes critical variables such as current price, old price, and price per unit, allowing us to calculate basic statistics, examine price changes over time, and assess potential relationships between unit pricing and overall product price. This dataset offers a diverse view of pricing strategies by covering a range of products, vendors, and pricing behaviors, providing a valuable resource for understanding real-world pricing dynamics.

The data collected from various vendors captures a snapshot of product prices at specific points in time. Each entry represents an observation of a product's pricing across these key variables, providing insights into both static and dynamic pricing practices. In terms of broader context, datasets that track longitudinal price changes across diverse vendors and product categories are rare, making this dataset a unique resource for exploring how pricing varies by vendor and product type. While alternative datasets could include retail pricing indices or specific vendor pricing records, they often lack the granularity or cross-vendor scope of Project Hammer, limiting their applicability for this type of multi-faceted analysis.

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 insights into price trends and volatility. For example, a reduction in old price compared to current price could indicate a promotional discount, while an increase might suggest adjustments due to supply constraints or increased demand. 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.

To ensure data integrity, the dataset was meticulously filtered and cleaned, particularly focusing on the current price and old price columns to remove any missing or invalid entries. This cleaning process involved handling null values and standardizing units where necessary to maintain consistency across entries. Such preprocessing steps allow for a more accurate and meaningful analysis by reducing noise and focusing on valid, comparable data points. Each entry in the dataset thus represents a refined record of a product's price at two different time points, along with its unit price, translating real-world pricing phenomena into a structured format suitable for analysis.

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.

- 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.

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

- 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.

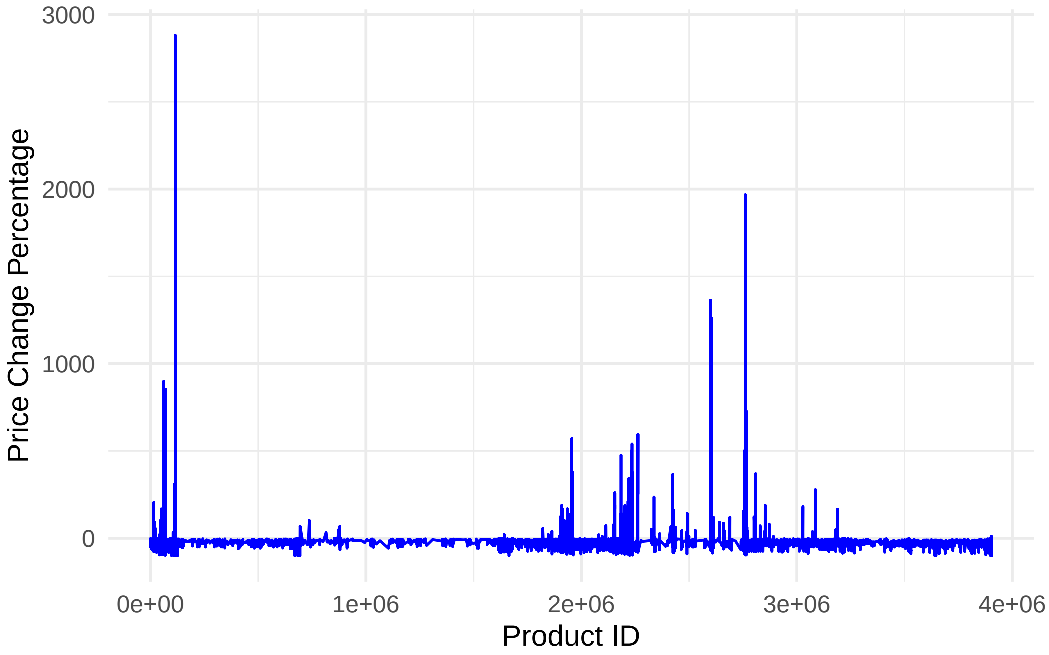

Figure 1: Price Change Percentage by Product ID

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 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.