Analysis of Product and Brand Usage Data from Hammer Data*

Ruiyang Wang

2024-11-15

Abstract

This analysis examines observational data from the Hammer dataset, focusing on trends in average price over time, brand distribution, and unit usage frequency. Using SQL for data manipulation and R for data visualization, this study identifies significant patterns in price fluctuations, brand popularity, and unit distribution, shedding light on potential sources of bias, the impact of missing data, and insights into correlation versus causation.

Introduction

The data for this analysis originates from Hammer's publicly available dataset ([https://jacobfilipp.com/hammer/
The objective of this study is to explore trends in pricing, brand representation, and unit usage distribution. Through this investigation, we aim to understand the factors that could potentially influence these patterns. This paper employs SQL for initial data preparation and R for data visualization, with each key finding presented in a corresponding chart. The dataset consists of three key components, each focusing on a different aspect of the data:

1. **Average Price Over Time**: This data segment captures the average product prices across different time points, enabling a time-series analysis of price fluctuations.
2. **Brand-Based Unit Usage**: This part of the data illustrates the frequency of unit usage among different brands, highlighting which brands are more prominent in the dataset.
3. **Unit Usage Frequency**: This dataset segment shows the usage frequency of various units, providing insight into common measurements and packaging formats for products.

*Code and data are available at: [https://github.com/Ruiyang-Wang/STA304_Ref6.git).

Each of these data segments provides a unique perspective, allowing us to explore trends in pricing, brand representation, and unit distribution.

Measurement

For the analysis, each dataset was processed to focus on the most relevant trends. The "Average Price Over Time" dataset was used to plot price trends throughout the year. The "Brand-Based Unit Usage" dataset was filtered to include only the top 10 brands by unit count, focusing on brands with significant representation. The "Unit Usage Frequency" dataset was limited to the top 20 units, providing insights into the most commonly used units.

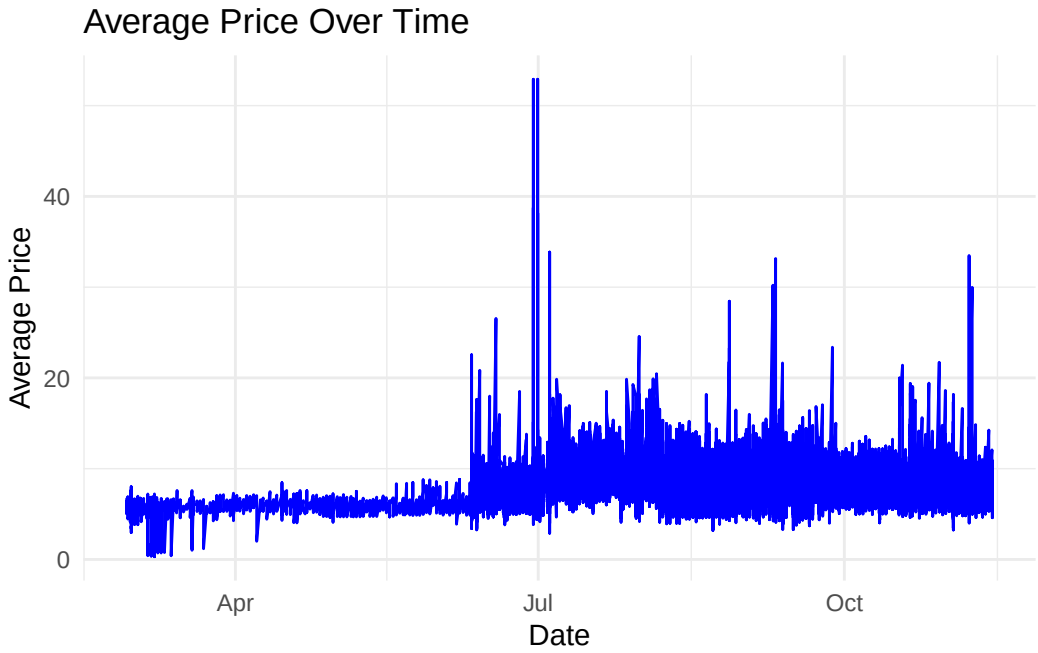
Results

Average Price Over Time

The first analysis focuses on the trend of average product prices over time. This trend provides insights into how prices vary throughout the year, which may reflect seasonal trends, market demands, or other external factors.

[Insert Average Price Over Time Plot Here]

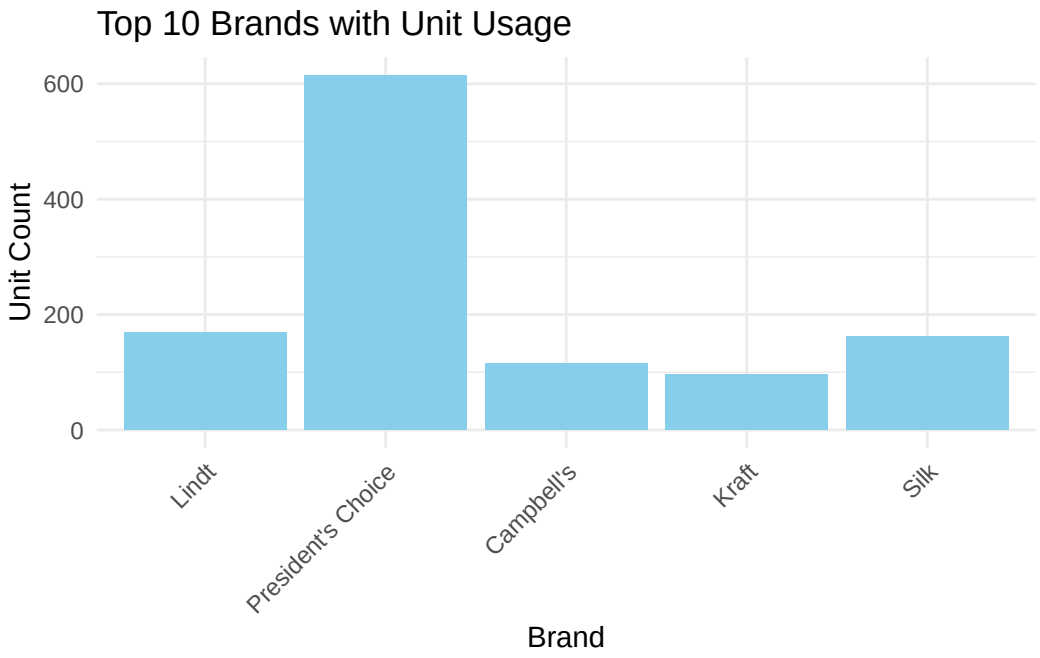
The plot shows a noticeable spike in average prices around June to July, with smaller fluctuations occurring throughout the rest of the year. This sharp increase during mid-year could indicate seasonal demand surges, supply chain constraints, or other economic influences. Further investigation into potential contributing factors, such as sales events or production cycles, could help explain these observed patterns.



The plot shows a noticeable spike in average prices around June to July, with smaller fluctuations occurring throughout the rest of the year. This sharp increase during mid-year could indicate seasonal demand surges, supply chain constraints, or other economic influences. Further investigation into potential contributing factors, such as sales events or production cycles, could help explain these observed patterns.

Brand-Based Unit Usage

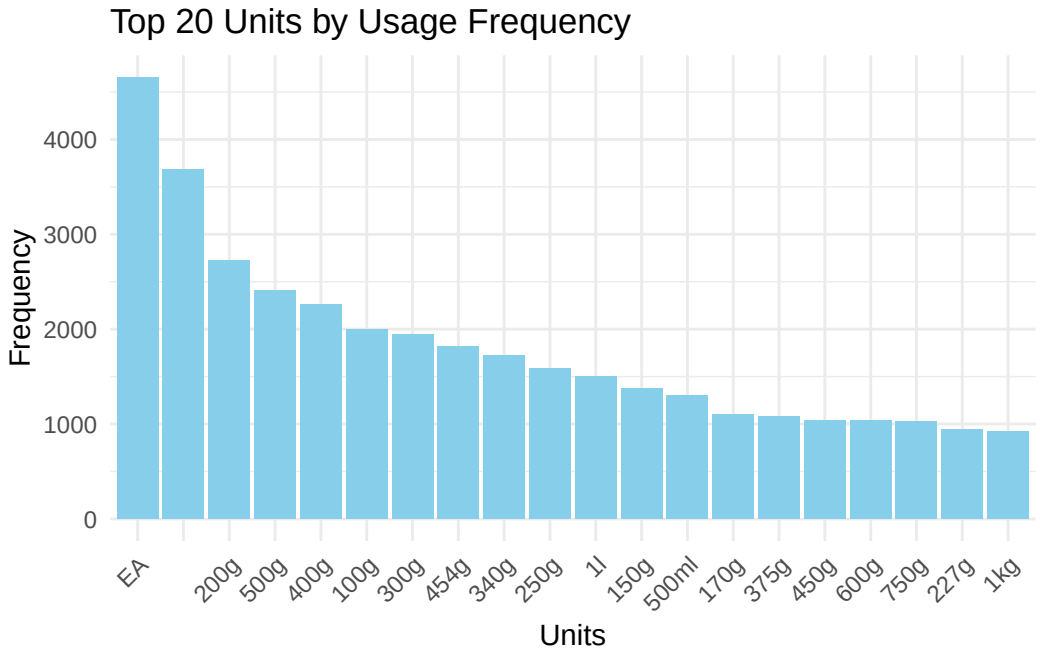
The second analysis examines the distribution of unit usage across different brands. Understanding brand representation is essential, as brands with higher unit counts may indicate a stronger market presence or a broader product selection within the dataset.



The chart reveals that "President's Choice" has a significantly higher unit count compared to other brands, suggesting a dominant position in the dataset. This disparity may point to a larger market share or a more extensive product line for "President's Choice." However, this concentration of data towards certain brands may introduce bias, limiting the generalizability of findings if these brands are overrepresented.

Unit Usage Frequency

The final analysis investigates the frequency distribution of various units used across products. This analysis provides a look into the common measurements or packaging formats within the dataset, which can be essential for understanding product packaging trends.



Ignoring "EA" and "blank" unit, the chart shows that the most frequently used units are now weight-based measurements, with "200g," "500g," and "400g" appearing at the top. These units suggest standardized packaging sizes, indicating that many products in the dataset are packaged in specific weight increments. The predominance of these units likely reflects typical packaging norms for certain types of goods, such as food items or household products, which are often sold in these weights.

This pattern allows us to better understand common product formats and offers insight into market trends in packaging preferences. Excluding "EA" and blank values gives a more focused view of measured units, highlighting the most common package sizes without the influence of single-item counts.

Discussion

Correlation vs. Causation

The "Average Price Over Time" analysis shows that average prices tend to rise around mid-year, especially in June and July. While this could suggest a link between these months and higher demand, it is not certain that this is the main cause. Other factors, such as changes in production costs, supply chain issues, or even special promotions, might also be influencing prices during these months. To truly understand why prices go up in mid-year, it would be helpful to examine other variables that could be driving these trends.

Missing Data

In the brand distribution analysis, some brands appear to have little or no data attached to them, especially brands with lower representation. This lack of data could skew the results, as it places more emphasis on brands with more available information, like "President's Choice." This means that certain brands might look more significant in the dataset than they really are in the market. A more complete dataset, with balanced data across brands, would help provide a clearer picture of true brand distribution.

Sources of Bias

The data also shows some potential biases, particularly in how brands and units are represented. In the "Brand-Based Unit Usage" chart, the strong presence of "President's Choice" could mean that trends specific to this brand are overemphasized, rather than showing a true market-wide pattern. Additionally, the original "Unit Usage Frequency" analysis had "EA" as the most common unit, which might focus too much on individually packaged products. By excluding "EA" and empty values in the revised analysis, we get a clearer picture of commonly used measured units, like "200g" and "500g." These units suggest standardized packaging sizes, which may better represent broader trends. It's important to consider these biases, as they may limit how well the findings apply beyond this dataset. Ensuring more balanced brand and unit representation would help to reduce these biases.

Conclusion

This analysis of the Hammer dataset provides some useful insights into patterns in pricing, brand representation, and unit distribution. We found that prices tend to spike in mid-year, especially around June and July. Certain brands, like "President's Choice," dominate the brand distribution, and specific units such as "200g" and "500g" appear most often when we exclude "EA" and blanks.

However, there are some limitations to these results. Missing data for certain brands could make it harder to see the full picture, and the strong presence of a few brands could skew the overall findings. Future studies could add more datasets to verify these results and explore other reasons for the price fluctuations we observed. Also, addressing the missing data and balancing the representation of brands and units would improve the quality and reliability of the analysis.