Analysis of Grocery Pricing Data*

Daniel Xu

2024-11-14

Abstract

This paper explores a dataset of products from various vendors, focusing on pricing dynamics across different brands. Key goals include identifying price patterns, examining potential correlations, and discussing statistical issues like correlation vs. causation, missing data, and bias sources.

Introduction

The competitive landscape of product sales requires a keen understanding of price dynamics. This study examines a dataset with columns for vendor, product details, pricing, and time to analyze how these factors influence each other. The discussion includes an analysis of correlations, missing data challenges, and sources of bias. This paper uses (Team 2023),(Müller and Bryan 2023) , (Lionel Henry Hadley Wickham Romain François and Müller 2023), (Lionel Henry Hadley Wickham Winston Chang 2023) ,(Wei and Simko 2023) ,(International Organization for Standardization 2016) to do data analysis, and data from (Filipp 2023).

Data and Measurement

The dataset includes: - **vendor**: Vendor name - **product_name**: Product name - **nowtime**: Timestamp of data record - **brand**: Product brand - **total_records**: Number of records for the product - **min_price**: Minimum price - **max_price**: Maximum price - **price_difference**: Difference between max and min price - **avg_price**: Average price across records

*Code and data are available at: https://github.com/possibleburger2/.

Data preprocessing and summary statistics are shown below.

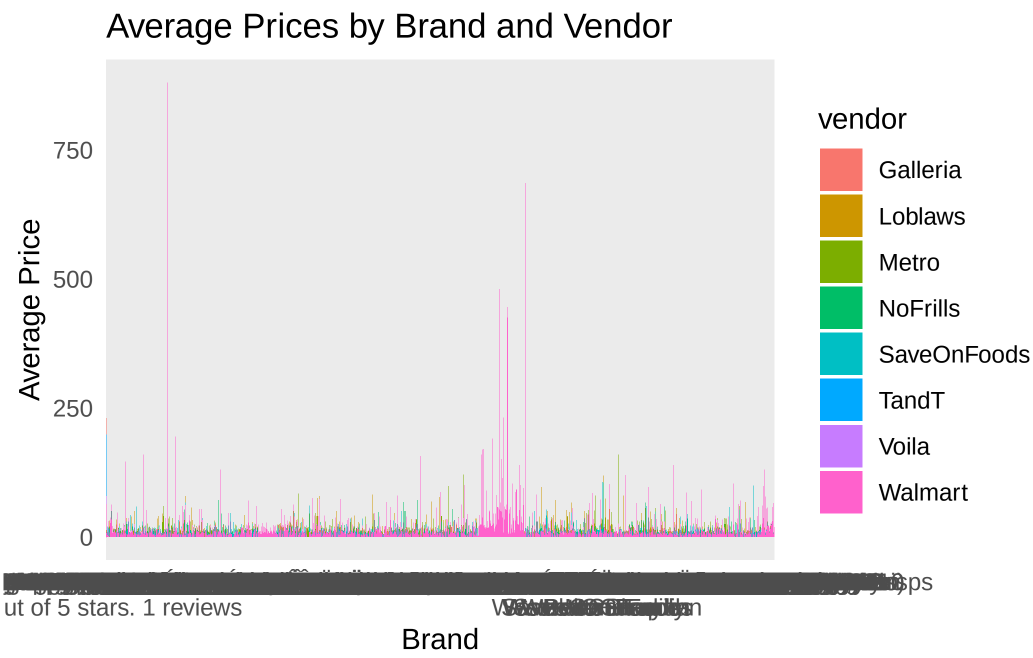
```r
# Load necessary packages
library(dplyr)
library(ggplot2)
library(corrplot)
library(here)

# Load data
data <- read.csv(here("data/cleaned_data/final_data.csv"))

# Summary statistics
summary(data)
```

```
    vendor          product_name         product_id        nowtime         
 Length:115641      Length:115641      Min.   :      3   Length:115641     
 Class :character   Class :character   1st Qu.: 713869   Class :character  
 Mode  :character   Mode  :character   Median :1704983   Mode  :character  
                                       Mean   :1910434                     
                                       3rd Qu.:2892878                     
                                       Max.   :4076381                     
    brand           total_records      min_price          max_price        
 Length:115641      Min.   :    1.0   Length:115641      Length:115641     
 Class :character   1st Qu.:   47.0   Class :character   Class :character  
 Mode  :character   Median :   94.0   Mode  :character   Mode  :character  
                    Mean   :  101.7                                        
                    3rd Qu.:  149.0                                        
                    Max.   :91881.0                                        
 price_difference    avg_price      
 Min.   :-125.30   Min.   :  0.000  
 1st Qu.:   0.00   1st Qu.:  3.745  
 Median :   0.20   Median :  5.506  
 Mean   :   0.92   Mean   :  7.437  
 3rd Qu.:   1.30   3rd Qu.:  8.308  
 Max.   :3502.98   Max.   :880.050  
 NA's   :83        NA's   :83       
```

Results Price Trends by Vendor and Brand

This section examines average prices by vendor and brand.

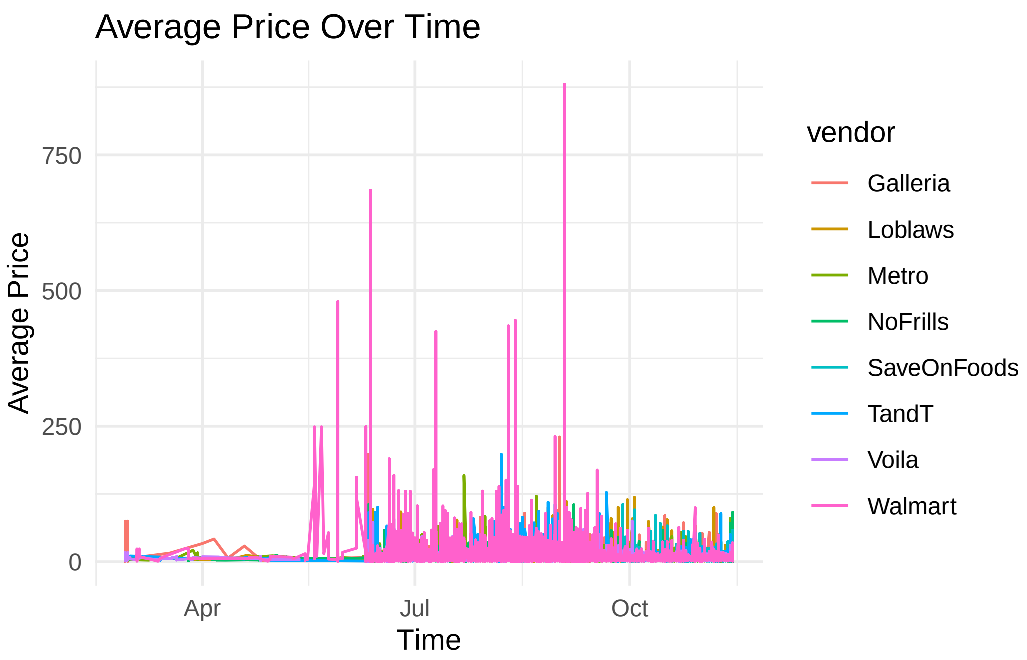
```r
# Average price by brand or vendor
ggplot(data, aes(x = brand, y = avg_price, fill = vendor)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average Prices by Brand and Vendor",
       x = "Brand",
       y = "Average Price") +
  theme_minimal()
```



Average Price Over Time

This line plot shows average price changes over time.

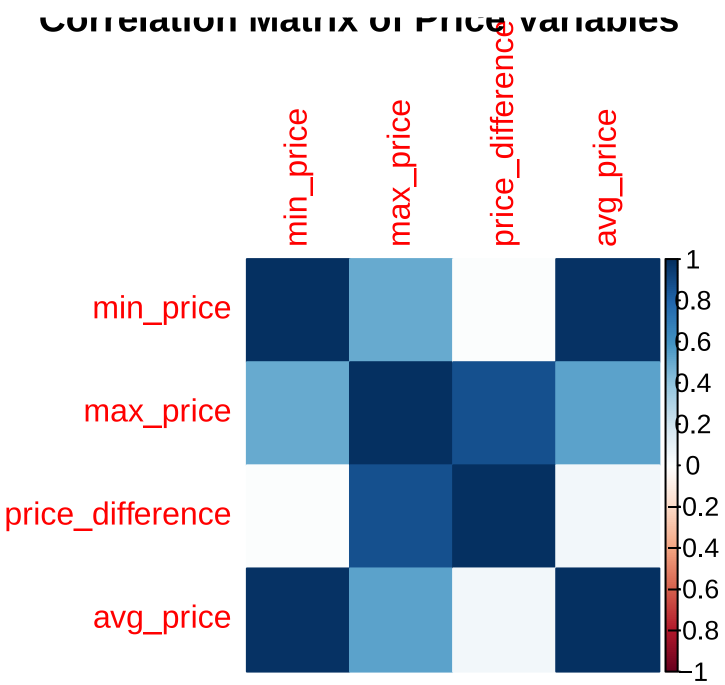
```r
# Line plot of avg_price over time
data$nowtime <- as.Date(data$nowtime)  # Convert to Date if needed
ggplot(data, aes(x = nowtime, y = avg_price, color = vendor)) +
  geom_line() +
  labs(title = "Average Price Over Time",
       x = "Time",
       y = "Average Price") +
  theme_minimal()
```



Correlation Matrix

A heatmap of correlations among numerical variables.

```
# Correlation matrix for price-related variables
price_data <- data %>% select(min_price, max_price, price_difference, avg_price)
price_data <- price_data %>%
  mutate(across(everything(), ~ as.numeric(as.character(.))))
cor_matrix <- cor(price_data, use = "complete.obs")
corrplot(cor_matrix, method = "color", title = "Correlation Matrix of Price Variables")
```



Discussion 1. Correlation vs. Causation

While the correlation matrix provides insights into potential relationships among price variables, we cannot infer causation. Higher correlations between min_price and avg_price may suggest pricing trends but do not indicate a causal effect. 2. Missing Data

Missing values in certain columns, such as max_price or avg_price, may skew results. Imputation techniques like mean substitution or data exclusion were considered to handle these issues. 3. Sources of Bias

Possible sources of bias include selection bias if only certain vendors are represented, temporal bias if data collection was concentrated around specific periods, and brand bias where established brands might skew average prices upwards. Conclusion

The analysis highlights notable price patterns and correlations among vendors and brands. Handling issues such as missing data and bias is crucial to ensure valid insights. Future studies should aim to gather more comprehensive data across a wider range of vendors and time frames.

Filipp, Jacob. 2023. "Hammer Data Science Blog." 2023. https://jacobfilipp.com/hammer/.

Hadley Wickham, Lionel Henry, Romain François, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Hadley Wickham, Lionel Henry, Winston Chang. 2023. *Ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.

International Organization for Standardization. 2016. "ISO/IEC 9075: Information Technology - Database Languages - SQL." https://www.iso.org/standard/63555.html.

Müller, Kirill, and Jennifer Bryan. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Team, R Core. 2023. "R: A Language and Environment for Statistical Computing." https://www.r-project.org/.

Wei, Taiyun, and Viliam Simko. 2023. *Corrplot: Visualization of a Correlation Matrix*. https://CRAN.R-project.org/package=corrplot.