Trends in Observational Product Inventory Data for Canadian Grocers

Divya Gupta

14 November, 2024

This paper provides an overview of product inventory trends for Canadian grocers using an observational dataset provided by Jacob Flipp/Project Hammer. Through a preliminary SQL analysis, I examine the distribution of vendors and brands at popular Canadian grocery chains and report on patterns of missing data. R is used to help visualise these findings through bar charts to further understand the data. This paper then discusses the unequal distribution of market share among a handful of vendors and brands, as well as discussing potential biases in data, challenges related to correlation versus causation, and does a deep dive into the missing data from the data set and its implications on our observations.

Introduction

There are numerous grocery chains across Canada, however, there are some that are seemingly more popular and widespread than others. Observational studies on the market distribution of these grocery chain stores can help provide insight into price distribution, market share occupation, vendor preferences, consumer brand preferences, and product availability across the country. This study provides a basic exploration of the dataset from Jacob Flipp/Project Hammer that aims to drive more competition and reduce collusion in the Canadian grocery sector. This is done by compiling a database of historical grocery prices from top grocers' websites. This dataset contains information regarding the various grocery vendors, product brands, products and prices across Canada. This study uses this database to understand vendor dominance, distribution of brands in terms of market share, availability and popularity, and availability of data on this topic.

Data & Measurement

R was used for preliminary exploratory data analysis in order to understand the nature of the data set and construct visualizations (Figures 1 & 2). SQL was used for data manipulation to better understand the missing values, vendor distribution, and data completeness in this dataset.
The dataset consists of over 120,000 records of variables such as "vendor", "product_name", "brand", "units" and "upc."

Results

SQL was used to generate information about the total number of unique vendors and brands (which can be seen in Figures 1 & 2), identify missing values in the dataset (identified at the "upc" variable) and understand how products are distributed across various vendors and brands.

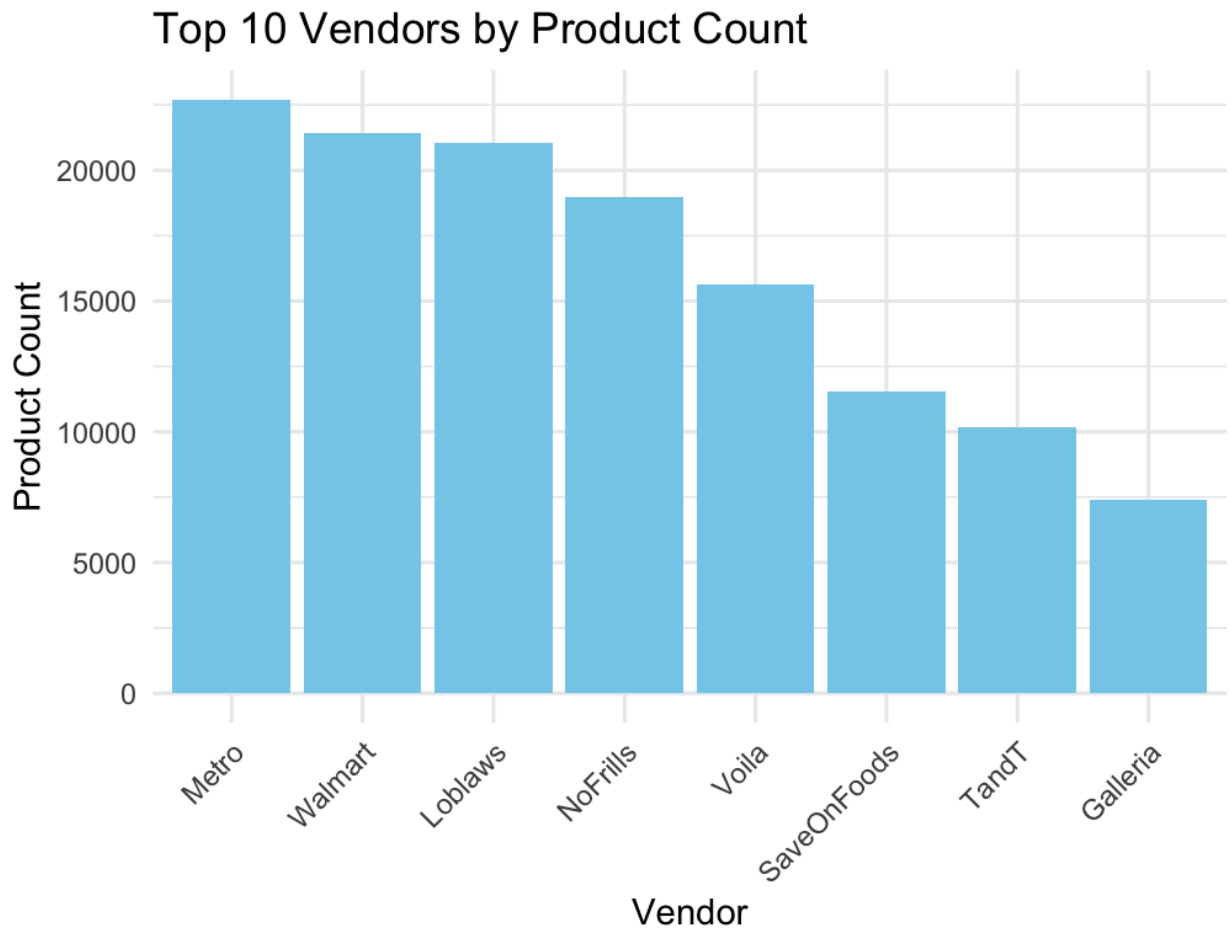

Figure 1. Vendor Distribution by Product Count

Figure 1 shows the distribution of the top ten vendors of the dataset by product count. Here, we can see that Metro, Walmart and Loblaws carry the largest amount/variation of products which suggests that these three grocery vendors have the widest selection of products. Galleria and TandT show the least amount of product variation, which can be explained by the fact that they are largely Asian-focused grocery stores, focusing on serving as a pan-Asian grocery store instead of a one-stop-shop for all kinds of products.

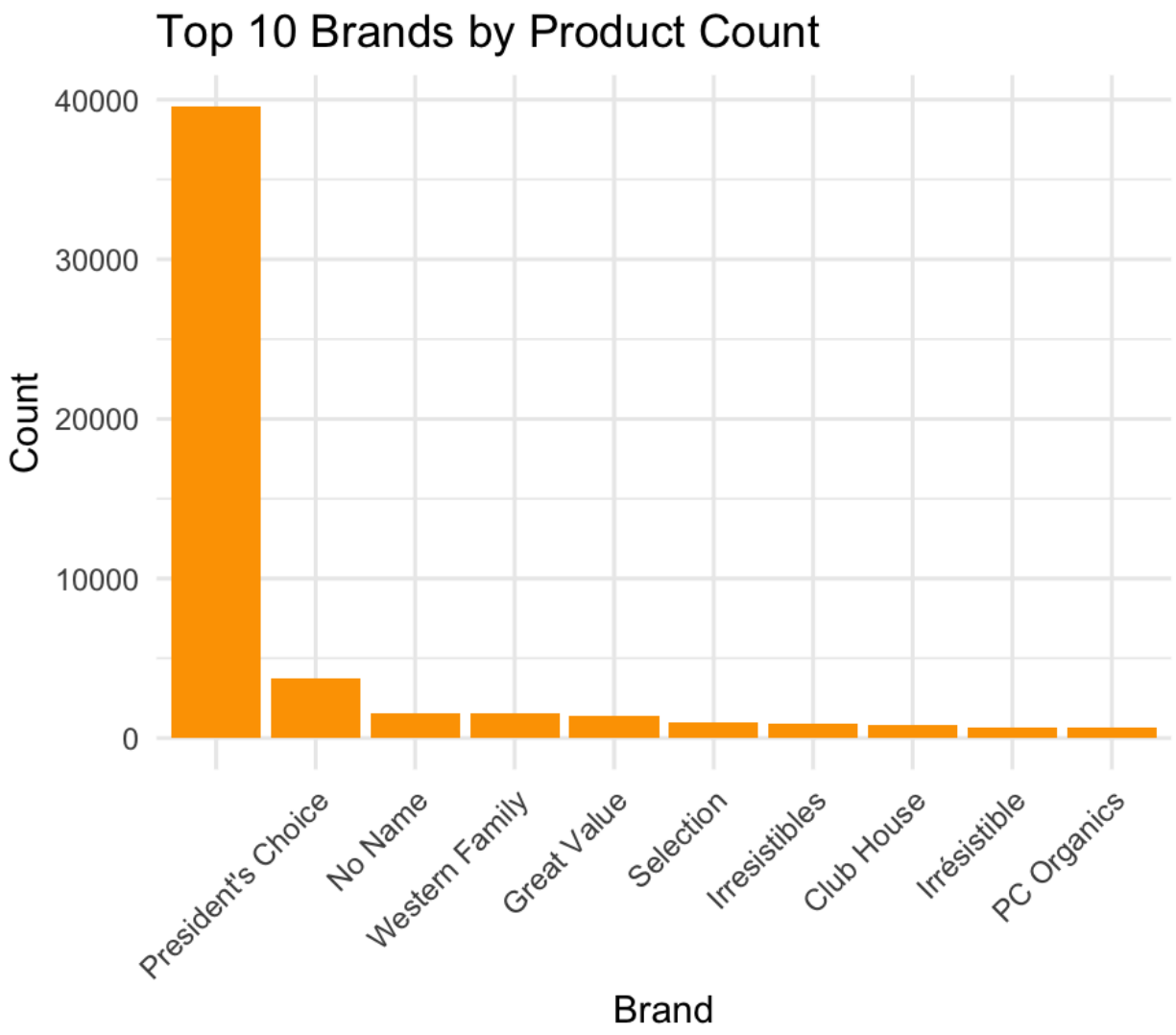

Figure 2. Brand Distribution by Product Count

Figure 2 shows the distribution of various brands across grocers by the number of products. This chart shows an overwhelming majority of President's Choice products dominating the market, with a high skewness towards President's Choice, which is then followed up by No Name. This illustrates that President's Choice has a wide market presence likely due to the high variation in their products and direct affiliations with vendors such as Loblaws and Shoppers Drug Mart.

Missing Values

The data analysis illustrated that there was a significant portion of the dataset that had incomplete entries, specifically for variables such as "brand" and "units", which can skew the results of our data exploration. Further, "upc", which stands for Universal Product Code, was overwhelmingly missing despite being a valid and crucial piece of information. This can limit the cross-referencing of products and may lead to double counting for certain products,

especially while generating an inventory, if not properly logged. This may also be a barrier to a dataset comparison over a period of time as there would be no way to monitor product usage and movement over time without ensuring that the products are not being double counted.

Units

While various units of measurement are included, the data has a higher number of products with units as “grams” or “kilograms” which allows us to extract key findings regarding the kind of products mostly stocked at grocery vendors. However, this is not regulated/consistent across all grocery stores, which can lead to problems with standardization.

Discussion

Correlation versus Causation

It is important to be aware of real world trends while analysing this dataset. While our visualizations (Figure 2) show an overwhelming skewness towards President’s Choice as the brand with the most product count, it does not necessarily reflect the brand’s market share, consumer trends, or consumer loyalty. It is likely that the dataset is extracted from vendors with affiliations with President’s Choice - such as Loblaws or Shoppers Drug Mart. Moreover, an overwhelming amount of product availability does not mean that it is also consumed or favored by consumers at the same amount. On the other hand, a lot of smaller grocery vendors, such as local stores, were not reported in the data set - leading to incomplete information about product distribution and availability across Canada.

Missing Data

Missing data about the “upc” can significantly impact an analysis about product consumption and consumer trends as it runs the risk of double counting products. This can also lead to discrepancies such as the skewness towards President’s Choice from Figure 2. The dataset does have comprehensive information about “brands” and “product_name” which allows us to use the full extent of the dataset without cleaning out too many unusable value points.

Sources of Bias

It is essential to consider the context through which this dataset was collected before drawing any conclusions about grocery vendor and product trends in Canada. Many smaller or culturally-forward grocers were not included in this dataset, leading to an incomplete picture of the kinds of products available. Further, various products are seasonal or only regionally-available, which bars us from drawing logical, pan-Canada conclusions about product availability. The high concentration of products from brands like President’s Choice also indicates a potential bias towards data collection from particular sources that may favor President’s Choice or may overrepresent its abundance in the market. Moreover, many brands or grocers don’t have their data readily available, managed, or structured in a way such that it can

be clearly drawn from when conducting an observational study, which may lead to an incomplete understanding of the market structure.

Conclusion

This study provides a high-level insight into the Canadian market of grocery vendors, with specific information about how the market is distributed amongst various vendors and brands. Key vendors such as Metro, Walmart, and Loblaws seem to lead in the variety of products available. On the other hand, President's Choice brand is leading in product variety at these vendors. More work needs to be done to better understand consumer interaction and preferences with these product availability trends in the market.

References

1. **Data Source**: https://jacobfilipp.com/hammer/

2. **Software and Tools**:
   - [Quarto](https://quarto.org/) for document preparation.
   - SQL for data manipulation.
   - R and Python for visualization.