Data Findings

After examining the four CSV files, there are a few points worth noting.

Quality of the data:

Most of the features in many rows have null values. Consider brand_code in the brands csv, which has 25 null values, and the same feature brand_code in receipt_items csv has more than 200000 null values.

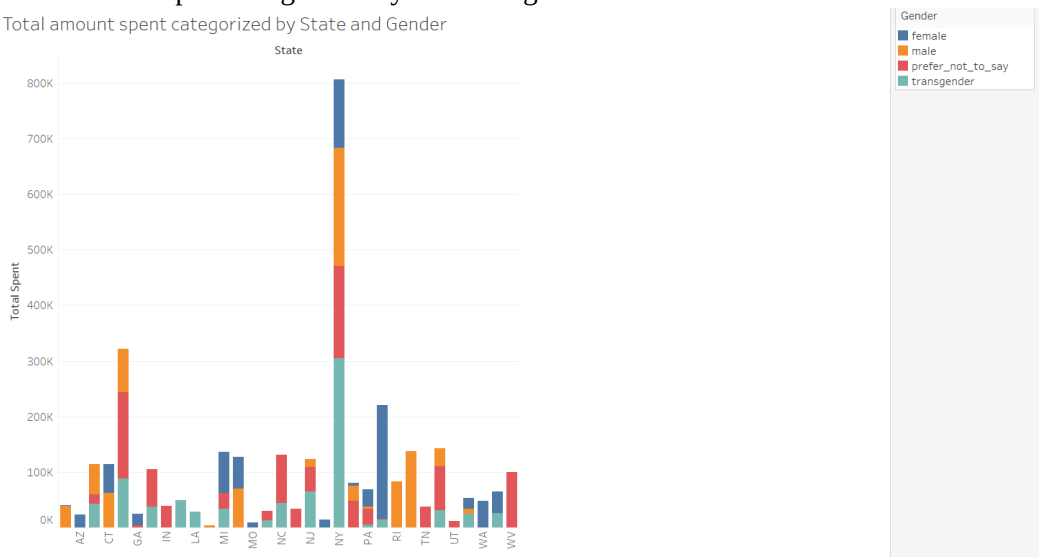
These numerous null values could skew the results. Since a great deal of them are missing, we cannot determine whether the results we obtained are accurate or not. So, due to this, business judgments may be flawed.

Although the majority of the features clearly state what they are intended to do, some do not. It would be easier to use all the data and make business decisions if the data description was present.

Observations:

I have noticed some interesting facts about the given data. I used tableau to create visualizations.

1. Total amount spent categorized by state and gender

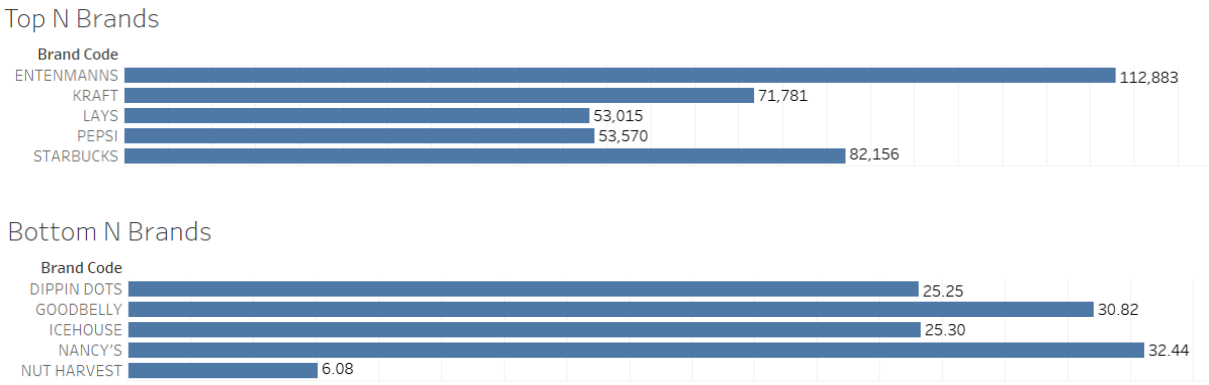


The entire amount spent is shown in the graph above, broken down by state and gender. As seen in the top right corner, the various colors represent the various genders while the vertical columns represent the amount spent, while. We can observe that New York spent a lot overall, with transgender people making the most contribution. Maryland has the least amount spent.

2. Total final price spent in each year

The total annual spending is shown in the image above. Each year, the amount spent grows progressively. Due to insufficient data, it is lower in 2023 and the years before 2020. This can demonstrate how findings might be completely skewed as a result of missing data.

3. Top/Bottom Brands Based on Spending

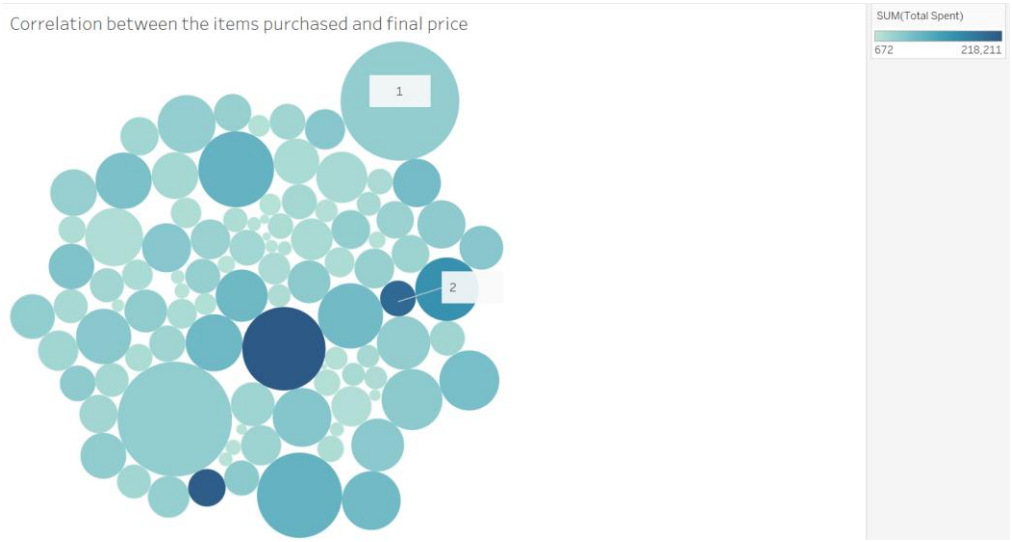


Based on the total dollars spent, the top five and lowest five brands are shown in this graph. This interactive dashboard was made with Tableau, and it allows users to sort brands by total spending to identify the top or bottom n .
As seen from the chart, Entenmanns has the highest dollars spent and Nut Harvest has the lowest dollars spent.

4. Quantity Purchased Based on the Price

The quantity bought in relation to price is displayed in the histogram above. The graphic above does not display the total bins. We can observe that the chart is skewed to the left since a cheaper item has been bought more frequently.

5. Correlation between the Items Purchased and the Total amount spent by user ID



In the image above, the size of the bubbles represents the number of products bought, while the color represents the total amount paid. The chart up top is drawn for unique user id. The amount of money spent does not always increase with the number of goods brought.

Conclusion:

We can make crucial business decisions with the aid of such aforementioned graphics. The construction of a warehouse is one potential business case. If we want to build a warehouse, we can find the ideal site, the products that should be filled more frequently, or the brands that need to be prioritized. Similarly, if sales in a certain area are down, we can close a few warehouses.