# Assignment 5: Association Analysis on Purchases

Ryan Fischbach
Dr. Khuri

November 13th, 2020

## 1 The Dataset

The purchases dataset contains 9,835 samples. Each sample is a transaction ranging from 1 to 32 items. The total number of items purchased was 43,367 with an average of 4.41 items per transaction.

## 2 Preprocessing the Data

To enable the best association analysis of this data, several steps were taken.

1. Duplicates samples were not removed because transactions with the same items purchased might not indicate that it was the same transaction. With no primary key, there is no way to tell if a transaction is a duplicate.

2. Null/NA values were identified and mitigated. In this case, NA values in rows were eliminated and the transactions were turned into a Pandas DataFrame of lists.

3. Mlxtend's Transaction Encoder was applied to turn the transactions into a sparse dataframe where each row was a sample and each column was an item that was purchased. A 1 in a column indicates that the item was purchased for the transaction and a 0 indicated it wasn't purchased.

## 3 Data Exploration

After data preprocessing, data exploration was performed to better understand the data.

The transaction encoded matrix was analyzed first. The density of the matrix was 0.026 (proportion of nonzero matrix cells).

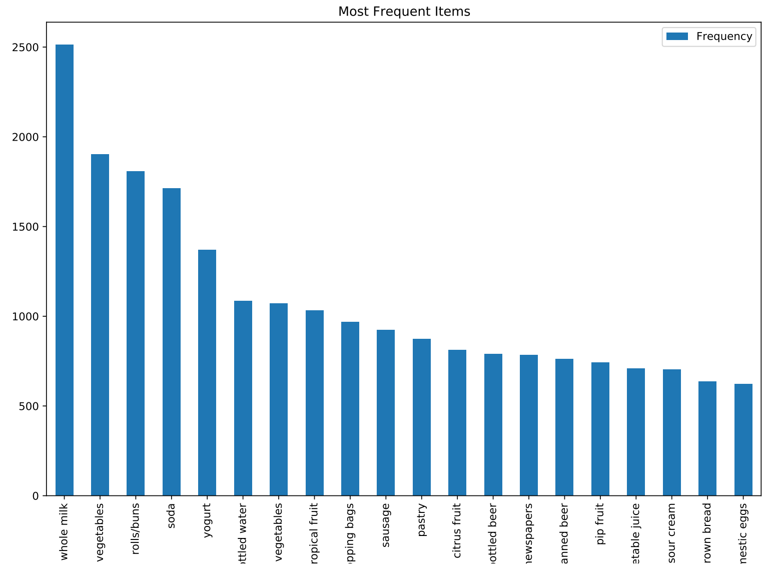

Figure 1: *A bar plot showing the most frequently purchased items from the transactions. The x axis shows each item with a corresponding frequency on the y axis. Whole milk was purchased the most (roughly 2500) times while domestic eggs were purchased the least (roughly 750 times).*

# 4 Association Analysis

The Mlxtend package was used to generate frequent itemsets and then generate rules. Rule mining gave insight into beneficial insight for the store.

## 4.1 Frequent Itemset Generation

Frequent itemsets were created via Mlxtend using a minimum support value. Support indicates the proportion of time a itemset appeared in all transactions. Minimum support was investigated and 0.01 produced the best results.

## 4.2 Rule Generation

Once itemsets were generated, rules were generated. Rules are a relationship between two itemsets where X implies Y. The confidence of a rule measures how often itemset Y appears in transactions with X. Thus, a higher confidence implies a more potent rule.

Mlxtend allows for rule mining with a minimum confidence value specified. The minimum confidence value was investigated and a minimum confidence of

0.25 generated the best results.

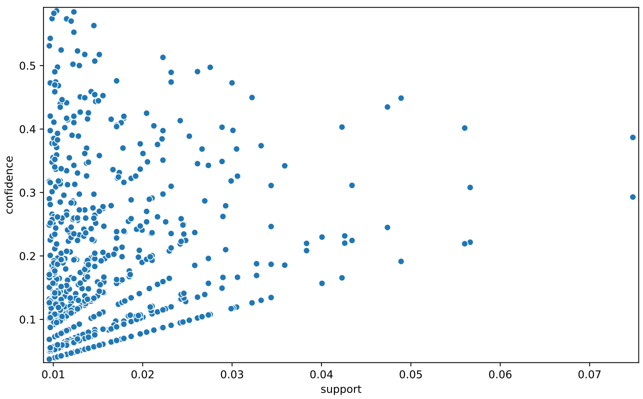

Figure 2: *The relationship between a rule's confidence and support was investigated, with support on the x-axis and confidence on the y-axis. There seems to be a weak positive linear relationship between the two.*

# 5 Results and Recommendations

With the minimum support and confidence thresh holds set to 0.01 and 0.25 respectively, the following rules were generated (sorted from highest confidence to lowest). Below are the 10 rules with highest confidence.

| | antecedents | consequents | support | confidence |
|---|---|---|---|---|
| 0 | (citrus fruit, root vegetables) | (other vegetables) | 0.010371 | 0.586207 |
| 1 | (tropical fruit, root vegetables) | (other vegetables) | 0.012303 | 0.584541 |
| 2 | (yogurt, curd) | (whole milk) | 0.010066 | 0.582353 |
| 3 | (other vegetables, butter) | (whole milk) | 0.011490 | 0.573604 |
| 4 | (tropical fruit, root vegetables) | (whole milk) | 0.011998 | 0.570048 |
| 5 | (yogurt, root vegetables) | (whole milk) | 0.014540 | 0.562992 |
| 6 | (domestic eggs, other vegetables) | (whole milk) | 0.012303 | 0.552511 |
| 7 | (yogurt, whipped/sour cream) | (whole milk) | 0.010880 | 0.524510 |
| 8 | (rolls/buns, root vegetables) | (whole milk) | 0.012710 | 0.523013 |
| 9 | (pip fruit, other vegetables) | (whole milk) | 0.013523 | 0.517510 |

These rules can provide insight into what patterns emerge when buying certain items. For example, other vegetables are bought 58 percent of the time citrus fruit and root vegetables are purchased. From these rules, a few suggestions for the store managers.

1. Root vegetables appear to be consistently in the antecedent itemsets, suggesting that items are consistently purchased with the itemset containing root vegetables. A sale on root vegetables may incentivize people to buy the consequent itemset with root vegetables, yielding more items per transaction.

2. Whole milk is a consequent for 8 of the top 10 rules, each with a confidence over 0.5. This suggests that whole milk is frequently purchased with these items. Whole milk should be moved closer to the antecedent itemsets to facilitate easier purchasing for those who buy the itemsets together. This carries the additional benefit of people who have not purchased these itemsets together having an easier time doing so.

3. Fruit, vegetables, and dairy make up the majority of this list. This suggests that these items are purchased together with high confidence. Placing the vegetable, fruit, and dairy locations close to each other could result in more frequent purchases of multiple items.

4. Rules consisting of the same items could have different confidences. If a confidence of one rule with an itemset as a consequent has a higher confidence than that itemset as a antecedent, then that itemset should be seen as a consequent. A strategy to utilize this is to print coupons for a consequent item at checkout if a person buys the antecedent itemset. This will have them likely purchase the consequent itemset if they haven't before or have them continue with buying the consequent by reinforcing their behavior. This could allow buyers to "tack on" that additional consequent item because of the deal. This would establish behavior and have them buy that consequent more frequently.

# 6 Citations

[1] Purchases.csv
[2] Tan et al. 2005. Introduction to Data Mining, 88 pages.
[3] Mlxtend. 2020. http://rasbt.github.io/mlxtend/
[4] Pandas. 2020. https://pandas.pydata.org
[5] Matplotlib. 2020. https://matplotlib.org
[6] Seaborn. 2020. https://seaborn.pydata.org
[7] Scikit learn. 2020. https://scikit-learn.org/stable/
[8] Natalia Khuri. 2020. lecture-7, 80 pages.
[9] Natalia Khuri. 2020. lecture-8.