# Market Basket Analysis

Chia-Chiao Liu

## A. Introduction

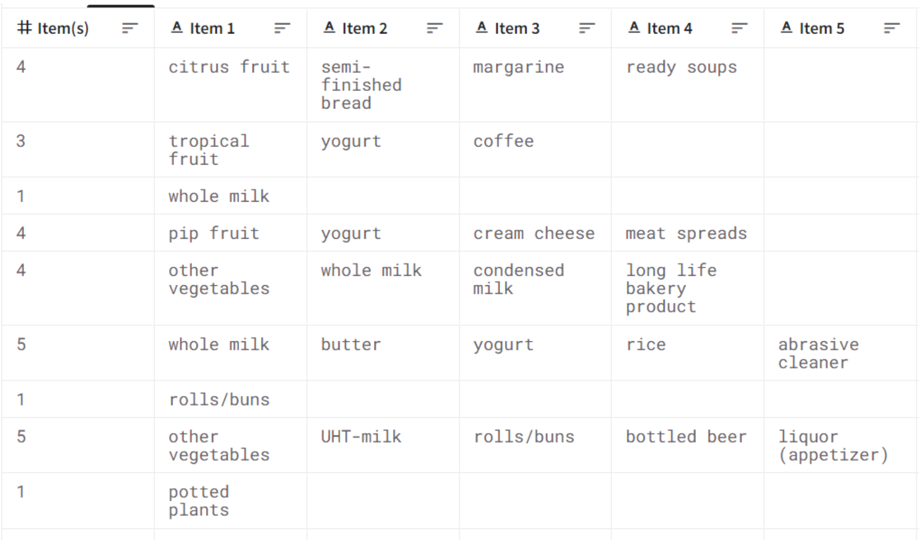
In 2003, a large grocery retailer in the United States called Giant Eagle used data mining to analyze customer purchasing behavior and optimize its marketing strategies [1]. One key approach they used was to create targeted promotions based on individual customer purchase histories, rather than generic offers that were sent to all customers.

As a result of these efforts, Giant Eagle was able to increase its sales by 8.6% in just three months, which translated to an additional $60 million in revenue. This was achieved through a combination of targeted promotions and improvements in supply chain efficiency that resulted from better insights into customer demand.

Based on this experience, the use of association rule learning is proposed to analyze customer transaction data and identify popular shopping combinations. This approach can help retailers gain a deeper understanding of customer shopping habits and market trends, allowing them to make informed decisions such as optimizing the placement of products and launching targeted sales promotions to boost revenue.

## B. Dataset

Using The Groceries Market Basket Dataset from Kaggle [2], a total of 9835 transactions and 169 products are included. The first column is the total number of purchases for this transaction, and the remaining columns are the list of purchased items.

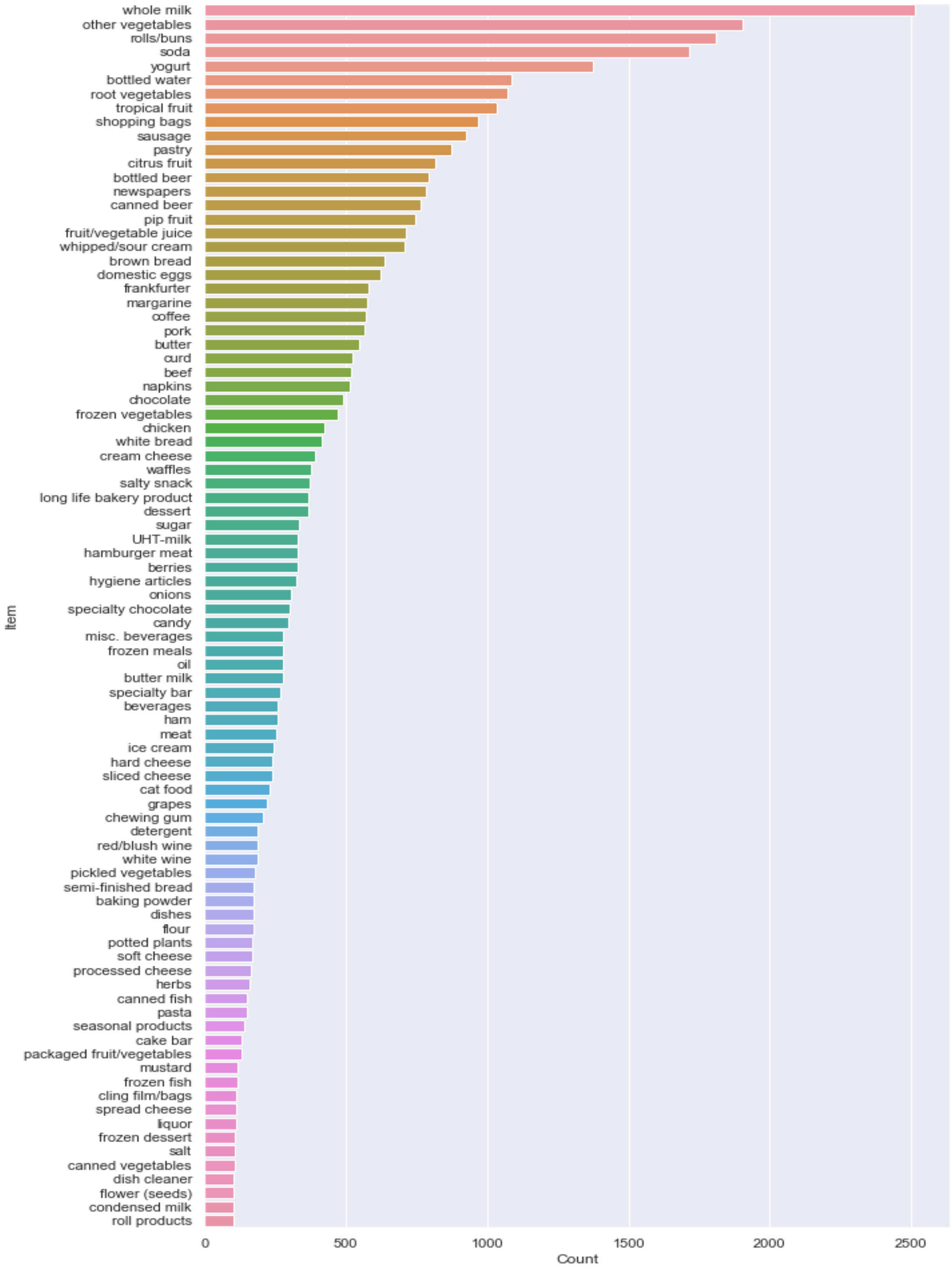


| # Item(s) | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 |
|---|---|---|---|---|---|
| 4 | citrus fruit | semi-finished bread | margarine | ready soups | |
| 3 | tropical fruit | yogurt | coffee | | |
| 1 | whole milk | | | | |
| 4 | pip fruit | yogurt | cream cheese | meat spreads | |
| 4 | other vegetables | whole milk | condensed milk | long life bakery product | |
| 5 | whole milk | butter | yogurt | rice | abrasive cleaner |
| 1 | rolls/buns | | | | |
| 5 | other vegetables | UHT-milk | rolls/buns | bottled beer | liquor (appetizer) |
| 1 | potted plants | | | | |

# C. Method

## a. Preprocessing

In the total of 9835 data, there are some products that rarely appear. If they are also included in the association rule learning, the denominator of the support will be raised, which is not conducive to finding popular products with high relevance. Thus, at the beginning of data processing, products with less than 99 purchases are deleted, the rest items are shown in the figure below.



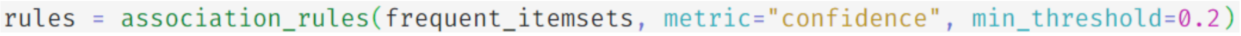
A total of 7311 transaction records and 88 commodities remain, which is about 74% of the original dataset, and the remaining transaction records are converted into one-hot vectors for association rule learning.

## b. Association Rule Learning

Association rule learning is a common data mining technique used to discover hidden Frequency Itemset and association rules in large datasets. Therefore, this project uses mlxtend [3], a well-known data science suite for Python, to implement common algorithms such as Apriori Algorithm and Association Rules.

The data pre-processing only retains the transaction data consisting of items purchased over 100 times, so based on the probability of two items purchased at the same time will also retain only those that exceed over 100 times. Experimentally found that doubling the threshold value can be more effective, so the minimum support is set to be (100/7311) ^ 2 * 2 ≈ 0.0037.

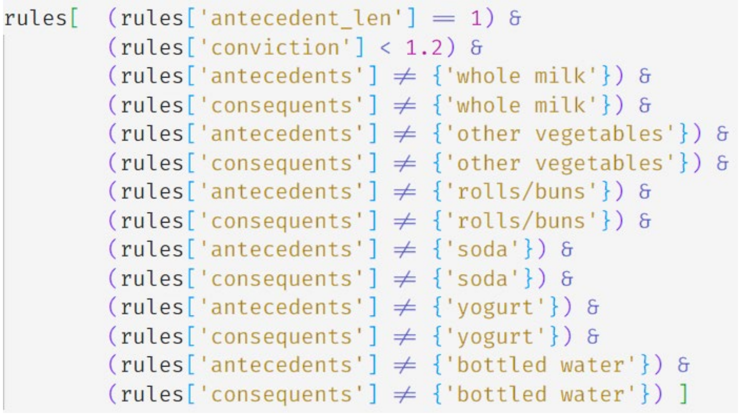
Another common metric for association rule learning is confidence, which represents the probability of buying a good A given that another good B is purchased. Set confidence(A, B) to be 0.2.



```
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.2)
```

The biggest problem with using only confidence to explore product association is that when two products A and B are both popular, then confidence(A, B) will be very high and unable to reflect the independence of these two products bought at the same time. Therefore, the rule conviction(A, B) < 1.2 is added which can be interpreted as the ratio of the expected frequency that A occurs without B should be less than 1.2.

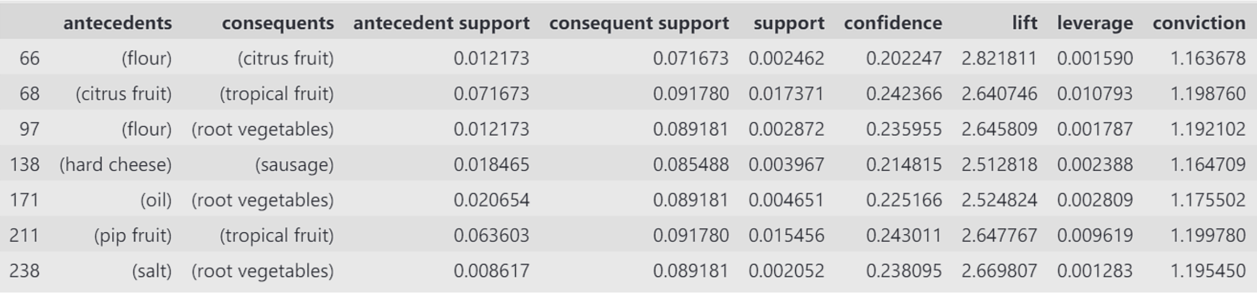
Since whole milk, rolls/buns, soda, yogurt, and bottled water are already popular items, these five items will be removed from the itemset. Other vegetables are also popular according to statistics, but this category does not reflect customers' buying tendencies, so it is also removed.



```
rules[  (rules['antecedent_len'] == 1) &
        (rules['conviction'] < 1.2) &
        (rules['antecedents'] ≠ {'whole milk'}) &
        (rules['consequents'] ≠ {'whole milk'}) &
        (rules['antecedents'] ≠ {'other vegetables'}) &
        (rules['consequents'] ≠ {'other vegetables'}) &
        (rules['antecedents'] ≠ {'rolls/buns'}) &
        (rules['consequents'] ≠ {'rolls/buns'}) &
        (rules['antecedents'] ≠ {'soda'}) &
        (rules['consequents'] ≠ {'soda'}) &
        (rules['antecedents'] ≠ {'yogurt'}) &
        (rules['consequents'] ≠ {'yogurt'}) &
        (rules['antecedents'] ≠ {'bottled water'}) &
        (rules['consequents'] ≠ {'bottled water'}) ]
```

## D. Result

After the rules were added, only seven itemsets remain, as shown in the following figure.



| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|---|---|---|---|---|---|---|---|---|---|
| 66 | (flour) | (citrus fruit) | 0.012173 | 0.071673 | 0.002462 | 0.202247 | 2.821811 | 0.001590 | 1.163678 |
| 68 | (citrus fruit) | (tropical fruit) | 0.071673 | 0.091780 | 0.017371 | 0.242366 | 2.640746 | 0.010793 | 1.198760 |
| 97 | (flour) | (root vegetables) | 0.012173 | 0.089181 | 0.002872 | 0.235955 | 2.645809 | 0.001787 | 1.192102 |
| 138 | (hard cheese) | (sausage) | 0.018465 | 0.085488 | 0.003967 | 0.214815 | 2.512818 | 0.002388 | 1.164709 |
| 171 | (oil) | (root vegetables) | 0.020654 | 0.089181 | 0.004651 | 0.225166 | 2.524824 | 0.002809 | 1.175502 |
| 211 | (pip fruit) | (tropical fruit) | 0.063603 | 0.091780 | 0.015456 | 0.243011 | 2.647767 | 0.009619 | 1.199780 |
| 238 | (salt) | (root vegetables) | 0.008617 | 0.089181 | 0.002052 | 0.238095 | 2.669807 | 0.001283 | 1.195450 |

## E. Conclusion

The preprocessing has revealed that whole milk, rolls/buns, soda, yogurt, and bottled water are the most popular items with a count greater than 1000. Therefore, retailers are suggested to prominently display these items in the store, or even place them in multiple locations throughout the store to attract more purchases.

Furthermore, through association rule learning, there are seven products related to each other and can be grouped into four product placements or promotional combinations. These suggestions are as follows:

- No. 68, 211: Citrus fruit, pip fruit, and tropical fruit should be placed together to encourage sales.
- No. 97, 171, 238: Root vegetables, flour, oil, and salt should be placed together to encourage sales.
- No. 66: Flour and citrus fruit can be sold together to encourage purchases.
- No.138: hard cheese and sausage can be sold together to encourage purchases.

## F. Reference

**[1] Kumar, V., & Reinartz, W. (2006). Customer Relationship Management: A Database Approach. John Wiley & Sons.**

**[2] https://www.kaggle.com/datasets/irfanasrullah/groceries**

**[3] http://rasbt.github.io/mlxtend/**