Report- 1

“A Study on the Groceries Market Basket Dataset using Association Rule Mining”

Group Name: ***MarketInsights***

Shri Abhiraami Thangavel, Rohan Reddy Galipur, Greg Gibson, Henry Yang, Mounika Yallamandhala

Table of Contents

Introduction:

The Market Basket Analysis will involve discovering patterns in customer purchasing behavior to enhance marketing and sales strategies. This study will focus on using the Apriori algorithm for association rule mining to analyze the Groceries Market Basket Dataset. The primary objective will be to uncover frequent itemsets and association rules that can provide insights into purchasing patterns and associations between different grocery items. By doing so, we aim to understand customer preferences and suggest strategies to optimize product placements, promotions, and inventory management.

Data:

The dataset comprises 9,835 transactions from customers shopping for groceries, involving 169 unique items. Individual transactions are represented as a list.

Link to Kaggle Dataset: https://www.kaggle.com/datasets/irfanasrullah/groceries/data

Methodology:

Data Preprocessing

After importing the data we found it has a shape of 9835 by 33, representing 9835 transactions with a possible 33 items. We also found that all 9835 transitions had at least one item. So we did not need to drop any transactions

Exploratory Data Analysis

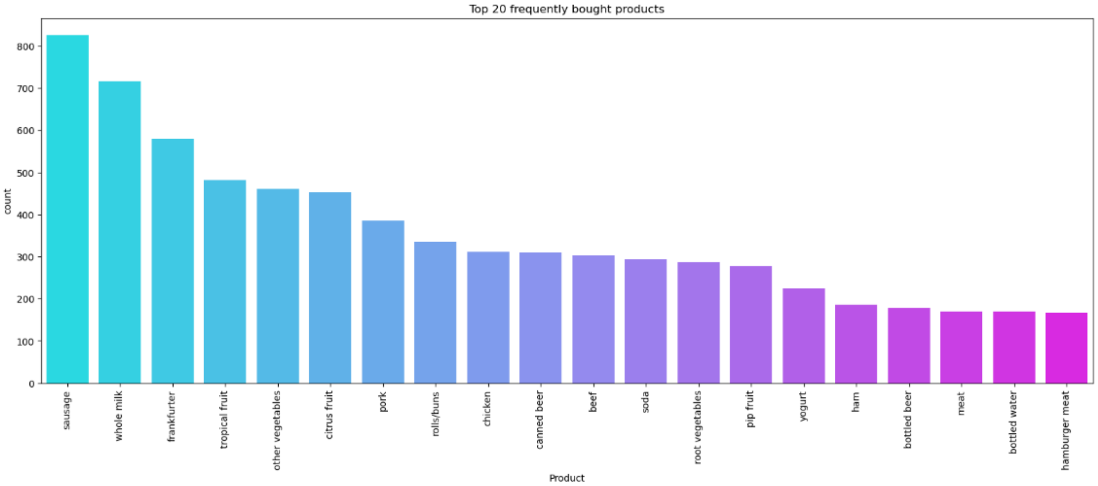
We used a variety of techniques to gain summary statistics in the early stages of analysis. Starting with a basic mean we found the average number of items per transaction was 4.4. Additionally we found that 50% of transactions have 3 items or less. The max number of items in a transaction was 32.

Frequently bought products

Frequently bought items are products or goods that consumers often purchase together in the same transaction, while standalone purchases are items that customers typically buy individually, without necessarily being related to other products in the same purchase.

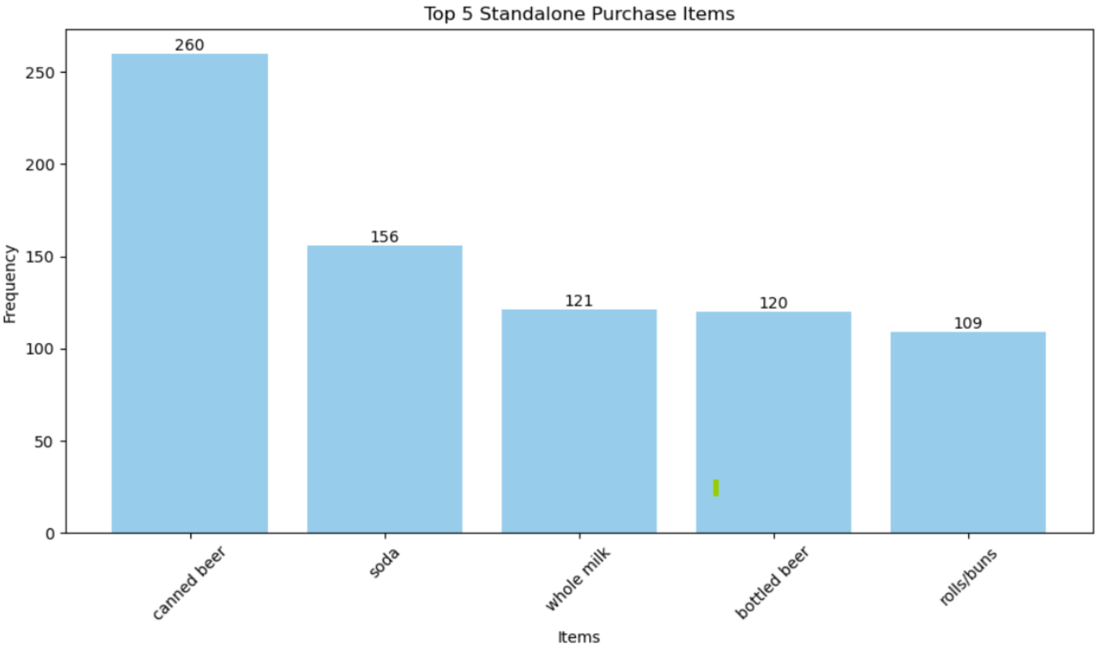
Frequently bought items are also known as "bundled" or "complementary" products. These are products that are often used in conjunction with each other or are commonly bought together due to their compatibility, usefulness, or as part of a set. For example, if you go to a hardware store to buy a new sink, you might also frequently buy a faucet and some plumbing fittings at the same time because these items are often needed together for a plumbing project.

We have identified frequently purchased products as shown in the graph below:

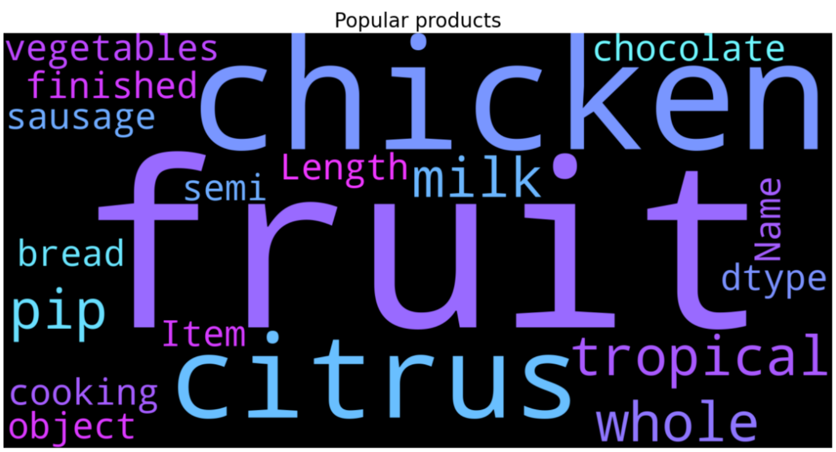


Standalone purchase products:
Standalone purchases, on the other hand, are products that are typically purchased individually without any strong association with other products. These items are not necessarily related or dependent on each other for their use. For instance, buying a book, a toaster, or a shirt are examples of standalone purchases because they are not usually bundled with other products.

Top 5 Standalone purchases are shown in the graph below:



We also made a wordcloud of the most common products from text mining the dataset.



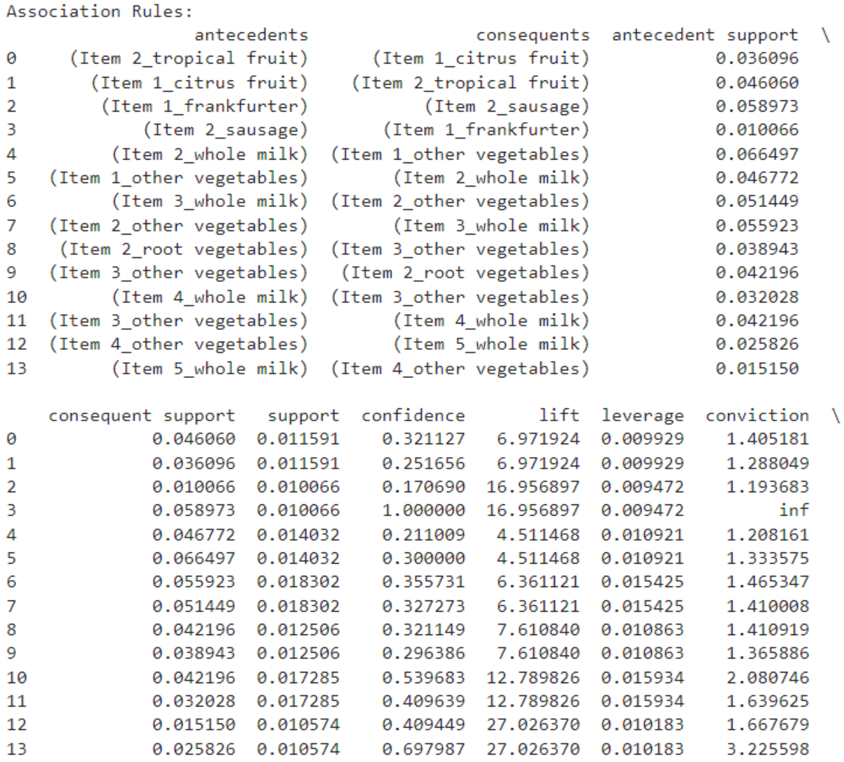
Testing and Training:

The entire dataset is utilized for training the Apriori algorithm.

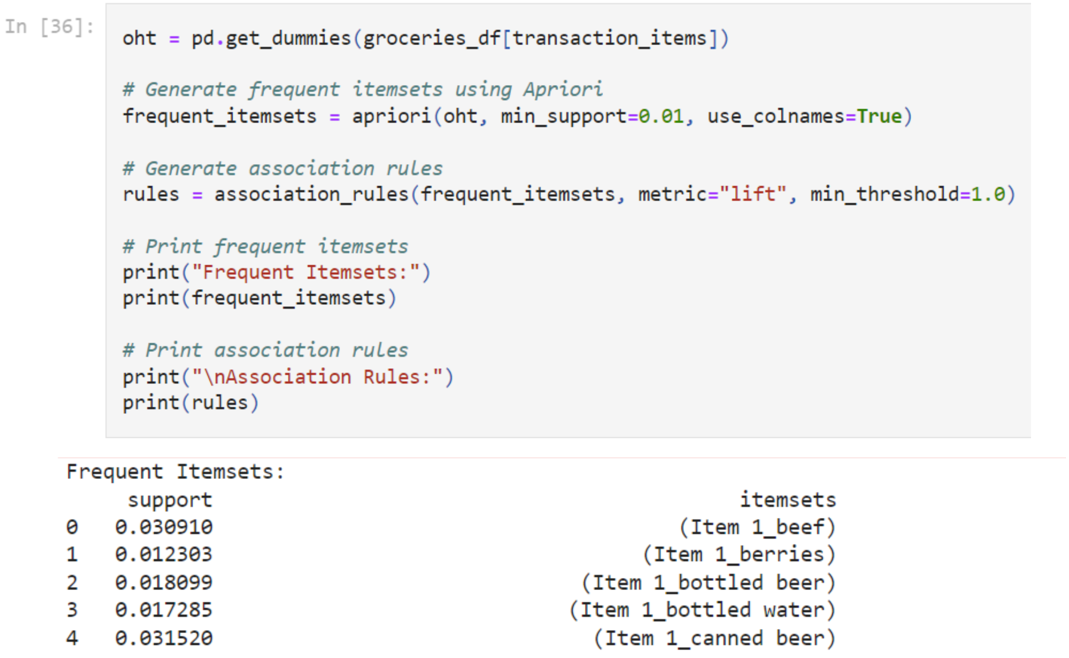
Constructing the Apriori Algorithm:

After feature engineering the transactions into dummy variables. We generated the list of frequent items and created the first association rules.

We found the following association rules:



```
Association Rules:
                  antecedents                 consequents  antecedent support  \
0     (Item 2_tropical fruit)     (Item 1_citrus fruit)            0.036096
1       (Item 1_citrus fruit)   (Item 2_tropical fruit)            0.046060
2        (Item 1_frankfurter)          (Item 2_sausage)            0.058973
3            (Item 2_sausage)      (Item 1_frankfurter)            0.010066
4         (Item 2_whole milk)  (Item 1_other vegetables)            0.066497
5   (Item 1_other vegetables)        (Item 2_whole milk)            0.046772
6         (Item 3_whole milk)  (Item 2_other vegetables)            0.051449
7   (Item 2_other vegetables)        (Item 3_whole milk)            0.055923
8    (Item 2_root vegetables)  (Item 3_other vegetables)            0.038943
9   (Item 3_other vegetables)   (Item 2_root vegetables)            0.042196
10        (Item 4_whole milk)  (Item 3_other vegetables)            0.032028
11  (Item 3_other vegetables)        (Item 4_whole milk)            0.042196
12  (Item 4_other vegetables)        (Item 5_whole milk)            0.025826
13        (Item 5_whole milk)  (Item 4_other vegetables)            0.015150

    consequent support   support  confidence       lift  leverage  conviction  \
0             0.046060  0.011591    0.321127   6.971924  0.009929    1.405181
1             0.036096  0.011591    0.251656   6.971924  0.009929    1.288049
2             0.010066  0.010066    0.170690  16.956897  0.009472    1.193683
3             0.058973  0.010066    1.000000  16.956897  0.009472         inf
4             0.046772  0.014032    0.211009   4.511468  0.010921    1.208161
5             0.066497  0.014032    0.300000   4.511468  0.010921    1.333575
6             0.055923  0.018302    0.355731   6.361121  0.015425    1.465347
7             0.051449  0.018302    0.327273   6.361121  0.015425    1.410008
8             0.042196  0.012506    0.321149   7.610840  0.010863    1.410919
9             0.038943  0.012506    0.296386   7.610840  0.010863    1.365886
10            0.042196  0.017285    0.539683  12.789826  0.015934    2.080746
11            0.032028  0.017285    0.409639  12.789826  0.015934    1.639625
12            0.015150  0.010574    0.409449  27.026370  0.010183    1.667679
13            0.025826  0.010574    0.697987  27.026370  0.010183    3.225598
```



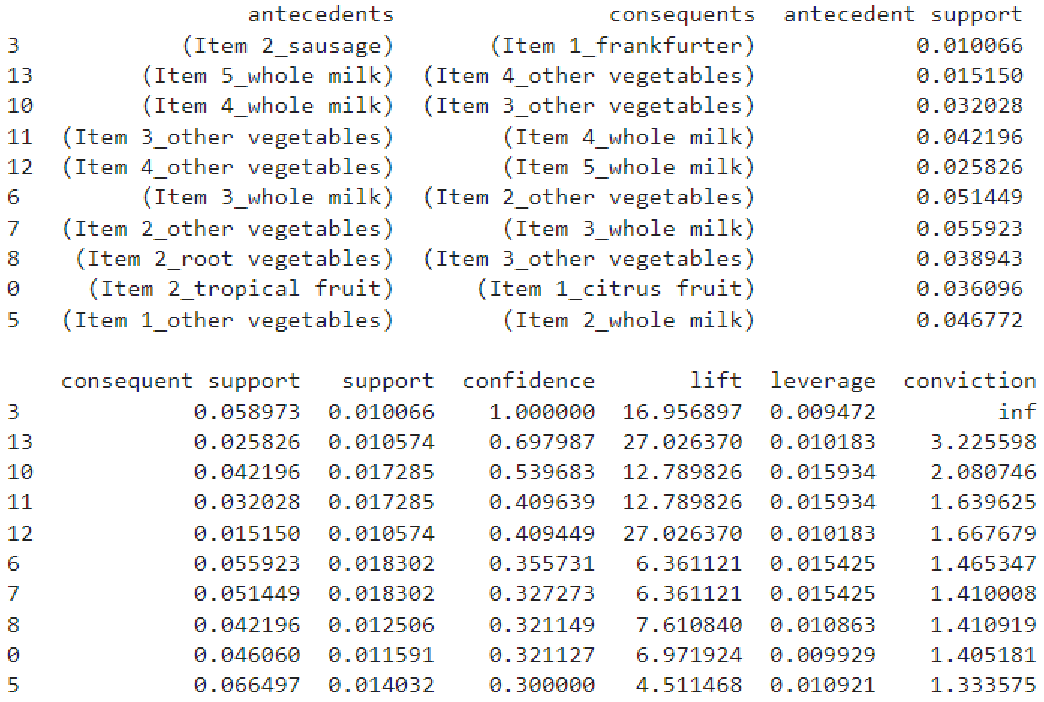
```
In [36]:
oht = pd.get_dummies(groceries_df[transaction_items])

# Generate frequent itemsets using Apriori
frequent_itemsets = apriori(oht, min_support=0.01, use_colnames=True)

# Generate association rules
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.0)

# Print frequent itemsets
print("Frequent Itemsets:")
print(frequent_itemsets)

# Print association rules
print("\nAssociation Rules:")
print(rules)
```

```
Frequent Itemsets:
     support                                           itemsets
0   0.030910                                      (Item 1_beef)
1   0.012303                                   (Item 1_berries)
2   0.018099                              (Item 1_bottled beer)
3   0.017285                             (Item 1_bottled water)
4   0.031520                               (Item 1_canned beer)
```

Analysis of Association rules:

We have sorted the associations obtained using metrics “Confidence” and “Lift” and obtained the following top 10 association rules.

Top 10 Association Rules:



```
                   antecedents                 consequents  antecedent support
3             (Item 2_sausage)        (Item 1_frankfurter)            0.010066
13         (Item 5_whole milk)  (Item 4_other vegetables)            0.015150
10         (Item 4_whole milk)  (Item 3_other vegetables)            0.032028
11  (Item 3_other vegetables)         (Item 4_whole milk)            0.042196
12  (Item 4_other vegetables)         (Item 5_whole milk)            0.025826
6          (Item 3_whole milk)  (Item 2_other vegetables)            0.051449
7   (Item 2_other vegetables)         (Item 3_whole milk)            0.055923
8    (Item 2_root vegetables)  (Item 3_other vegetables)            0.038943
0     (Item 2_tropical fruit)       (Item 1_citrus fruit)            0.036096
5   (Item 1_other vegetables)         (Item 2_whole milk)            0.046772

    consequent support   support  confidence       lift  leverage  conviction
3             0.058973  0.010066    1.000000  16.956897  0.009472         inf
13            0.025826  0.010574    0.697987  27.026370  0.010183    3.225598
10            0.042196  0.017285    0.539683  12.789826  0.015934    2.080746
11            0.032028  0.017285    0.409639  12.789826  0.015934    1.639625
12            0.015150  0.010574    0.409449  27.026370  0.010183    1.667679
6             0.055923  0.018302    0.355731   6.361121  0.015425    1.465347
7             0.051449  0.018302    0.327273   6.361121  0.015425    1.410008
8             0.042196  0.012506    0.321149   7.610840  0.010863    1.410919
0             0.046060  0.011591    0.321127   6.971924  0.009929    1.405181
5             0.066497  0.014032    0.300000   4.511468  0.010921    1.333575
```

Difficulties Faced:

The group has encountered a few difficulties so far in our analysis. Choosing minimum support threshold and slow computing time to name a few. We are working on strategies to address these as well as continuing to refine the analysis under the Action Items section. Another difficulty is lack of metadata, in a real world scenario you would more then likely have some supporting data such as payment method, store zip code or transaction time that could further aid our analysis and show macro trends. Beer is more likely to be purchased on Fridays than Mondays for example. In this dataset we only have the items in the transaction.

Action Items:

- Further refine our Apriori model to get more accurate results
- Explore the possibility of supplemental metadata

Looking Forward:

- Determine the Apriori pruning strategy
- How to handle spare data