Overview and Data

The goal of this project is to apply association rule mining to uncover purchasing patterns in the Groceries dataset. Each transaction represents all items purchased by a given member on a particular date. I implemented the Apriori algorithm in Python with a minimum support threshold of 0.005. After an update from the TA, I did not enforce the original minimum confidence of 0.3, because no rules at this support level reach that value. Instead, I generated rules from the frequent itemsets and ranked them by interestingness, defined as the absolute difference between confidence and the prior probability of the consequent.

Implementation (Apriori)

My code is organized in a single script, project3.py. The dataset is loaded with pandas and converted into a list of transactions by grouping the CSV rows by (Member_number, Date). Each transaction is the list of itemDescription values for that shopping trip.

For frequent itemset mining, I implemented the standard Apriori algorithm:

- Frequent 1-itemsets: I scan all transactions, count how many baskets contain each distinct item (using sets to avoid double counting within the same basket), divide by the total number of transactions to obtain support, and keep items with support ≥ 0.005.
- Higher-order itemsets ($k \geq 2$): I generate candidate k-itemsets by joining pairs of frequent $(k-1)$-itemsets that differ by one item. I then apply the Apriori property: a candidate is kept only if all its $(k-1)$-subsets are frequent. For these candidates, I rescan the transactions, count occurrences, compute support, and keep those above the threshold. This process repeats for increasing k until no new frequent itemsets are found.

For association rules, I use all frequent itemsets of size at least 2. For each itemset I and each non-empty proper subset $X \subset I$, I create a rule $X \rightarrow Y$, where $Y = I \setminus X$. I compute:

- $\text{support}(X \rightarrow Y) = \text{support}(X \cup Y)$
- $\text{confidence}(X \rightarrow Y) = \text{support}(X \cup Y) / \text{support}(X)$
- $\text{prior}(Y) = \text{support}(Y)$
- $\text{interestingness}(X \rightarrow Y) = |\text{confidence} - \text{prior}|$

I did not implement the PCY algorithm, so there is no PCY comparison in this report.

Results: Frequent Itemsets and Top Rules

After grouping, there are 14,963 transactions. With min_support = 0.005, Apriori finds 126 frequent itemsets, and the largest frequent itemsets have size 2; there are no frequent 3-itemsets or larger.

The top 10 frequent itemsets (all single items) and their supports are: 1. {whole milk} – 0.1579, 2. {other vegetables} – 0.1221, 3. {rolls/buns} – 0.1100, 4. {soda} – 0.0971 ,5. {yogurt} – 0.0859, 6.{root vegetables} – 0.0696 7. {tropical fruit} – 0.0678 8. {bottled water} – 0.0607 9.{sausage} – 0.0603 10.{citrus fruit} – 0.0531

From the frequent itemsets, my implementation generates 74 association rules. Because no rule reaches confidence 0.3, I rank rules by interestingness. The top 5 rules are:

1. root vegetables → whole milk
    a. support = 0.0076, confidence = 0.109, prior(whole milk) = 0.158, |conf − prior| = 0.049

2. root vegetables → other vegetables
    a. support = 0.0053, confidence = 0.076, prior(other vegetables) = 0.122, |conf − prior| = 0.046
3. bottled water → whole milk
    a. support = 0.0072, confidence = 0.118, prior(whole milk) = 0.158, |conf − prior| = 0.040
4. soda → whole milk
    a. support = 0.0116, confidence = 0.120, prior(whole milk) = 0.158, |conf − prior| = 0.038
5. tropical fruit → whole milk
    a. support = 0.0082, confidence = 0.121, prior(whole milk) = 0.158, |conf − prior| = 0.037

Interpretation and Discussion

The frequent itemsets show the most commonly purchased products across all baskets, with whole milk, vegetables, rolls/buns, and soda appearing in a large fraction of transactions. However, even though these items are popular individually, their combinations do not reach high support once we consider triplets; adding a third item quickly reduces support below 0.005, which explains why there are no frequent 3-itemsets.

The association rules reveal that, in this dataset, buying one common item is not a strong predictor of also buying another common item. The confidences of the top rules (around 0.076–0.121) are all lower than or comparable to the baseline (prior) probabilities of their consequents. For example, in the rule root vegetables → whole milk, whole milk's prior is about 0.158, while the conditional probability of whole milk given root vegetables is only about 0.109. This suggests that customers who buy root vegetables are actually slightly less likely than average to buy whole milk in the same trip. Similar patterns hold for the rules with bottled water, soda, and tropical fruit as antecedents.

Because of this, confidence alone is not very informative in this setting; the interestingness metric, |confidence − prior|, is more useful for highlighting deviations from the global purchasing rate. The top rules are "interesting" in the sense that the antecedent makes the consequent somewhat less likely than its overall frequency would suggest, indicating mild negative associations rather than strong positive ones.

I did not include a visualization in this report, but a simple bar chart of the supports of the top frequent pairs, or a network diagram where nodes are items and edges represent frequent pairs colored by interestingness, could further illustrate these relationships. Overall, Apriori on the Groceries dataset with min_support = 0.005 recovers sensible frequent items and a small set of mildly interesting associations, but no strong high-confidence rules.