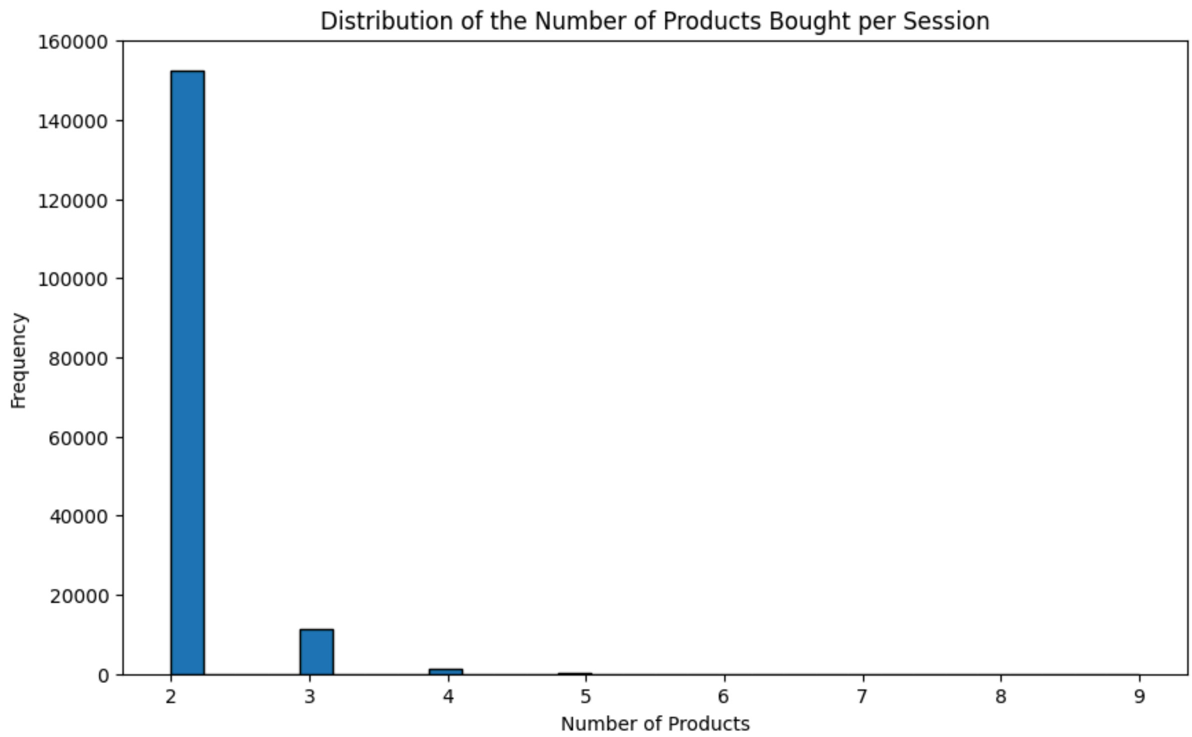
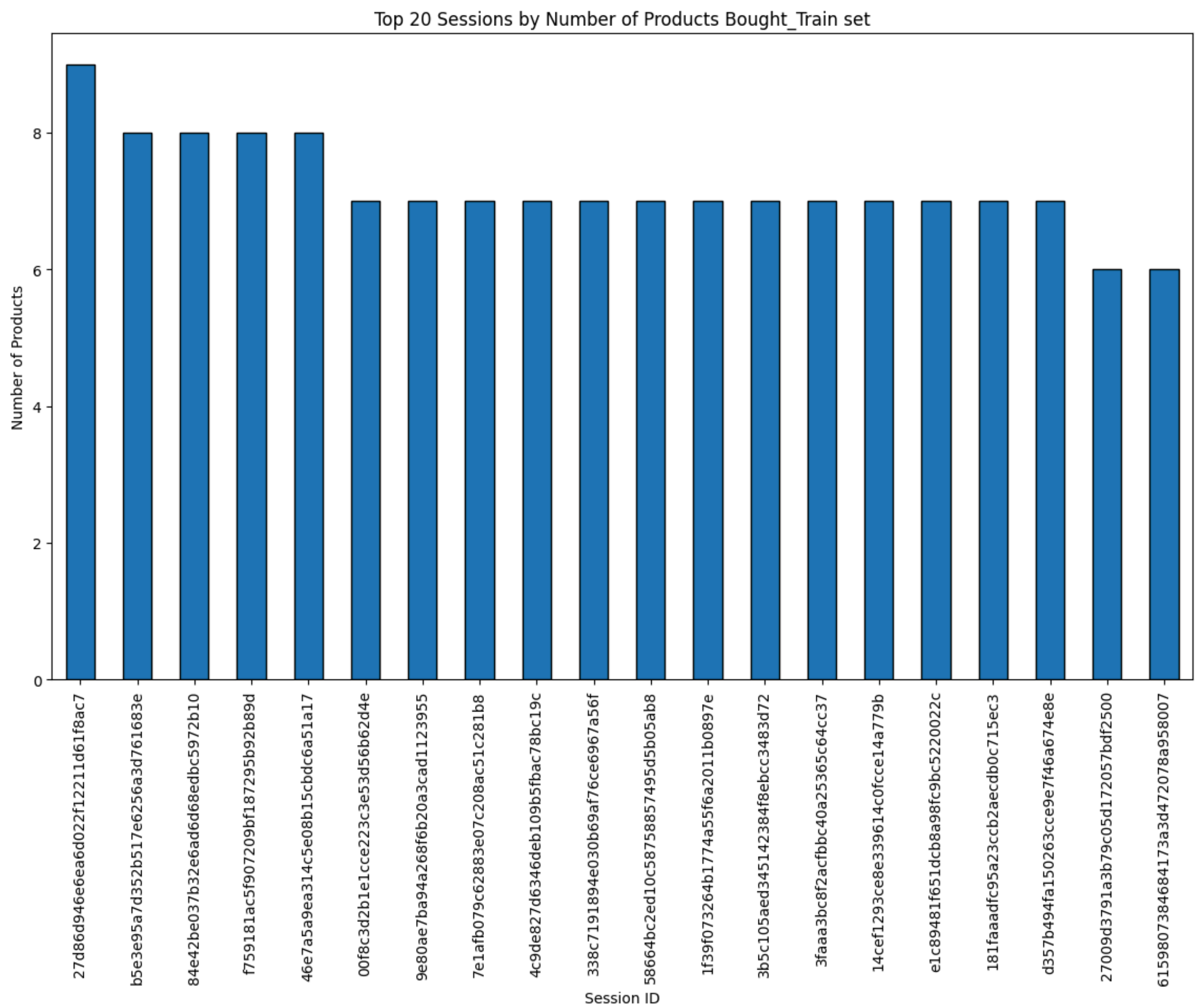
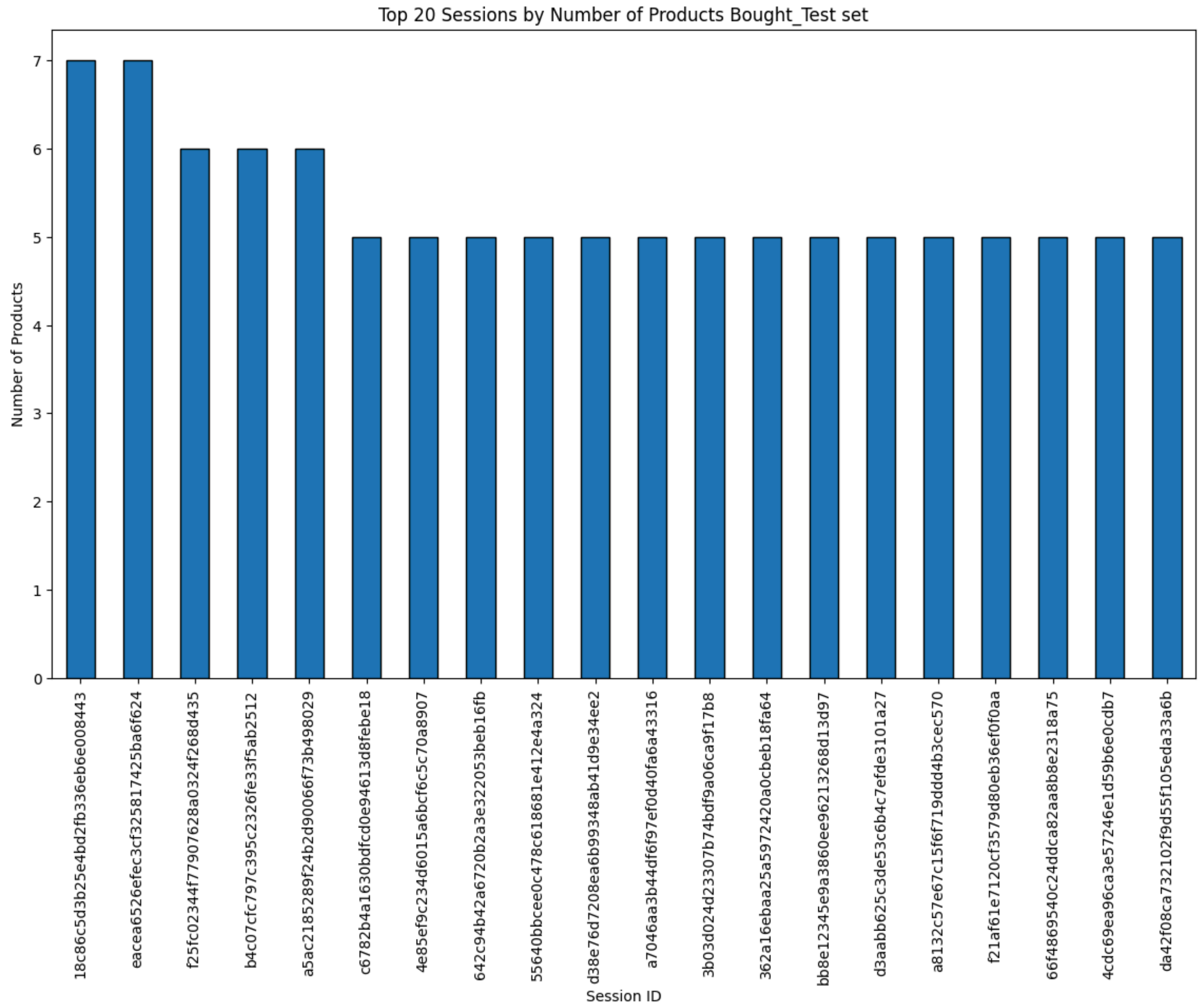
EDA done:



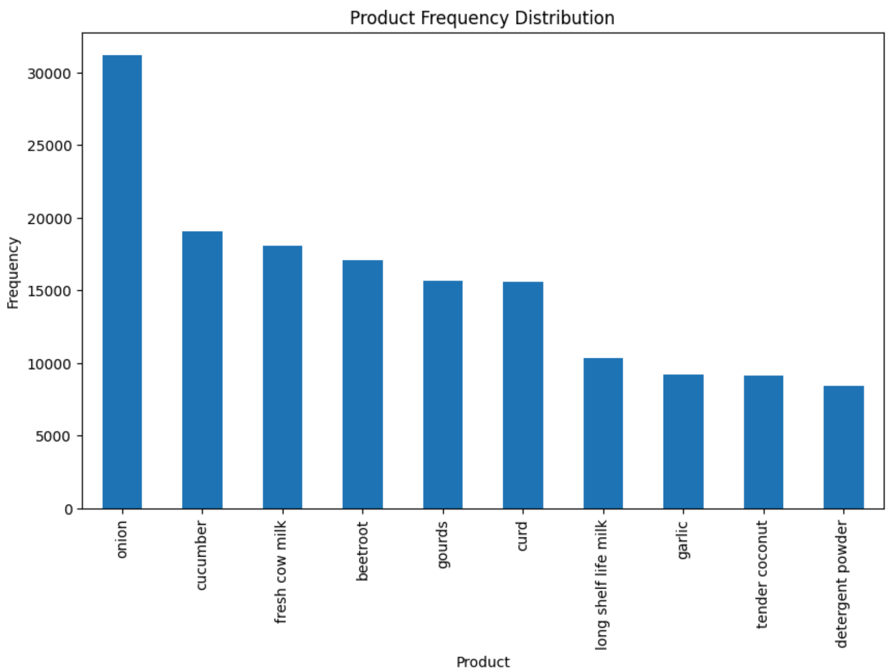
Distribution of the Number of Products Bought per Session
Frequency
0
20000
40000
60000
80000
100000
120000
140000
160000
Number of Products
2
3
4
5
6
7
8
9



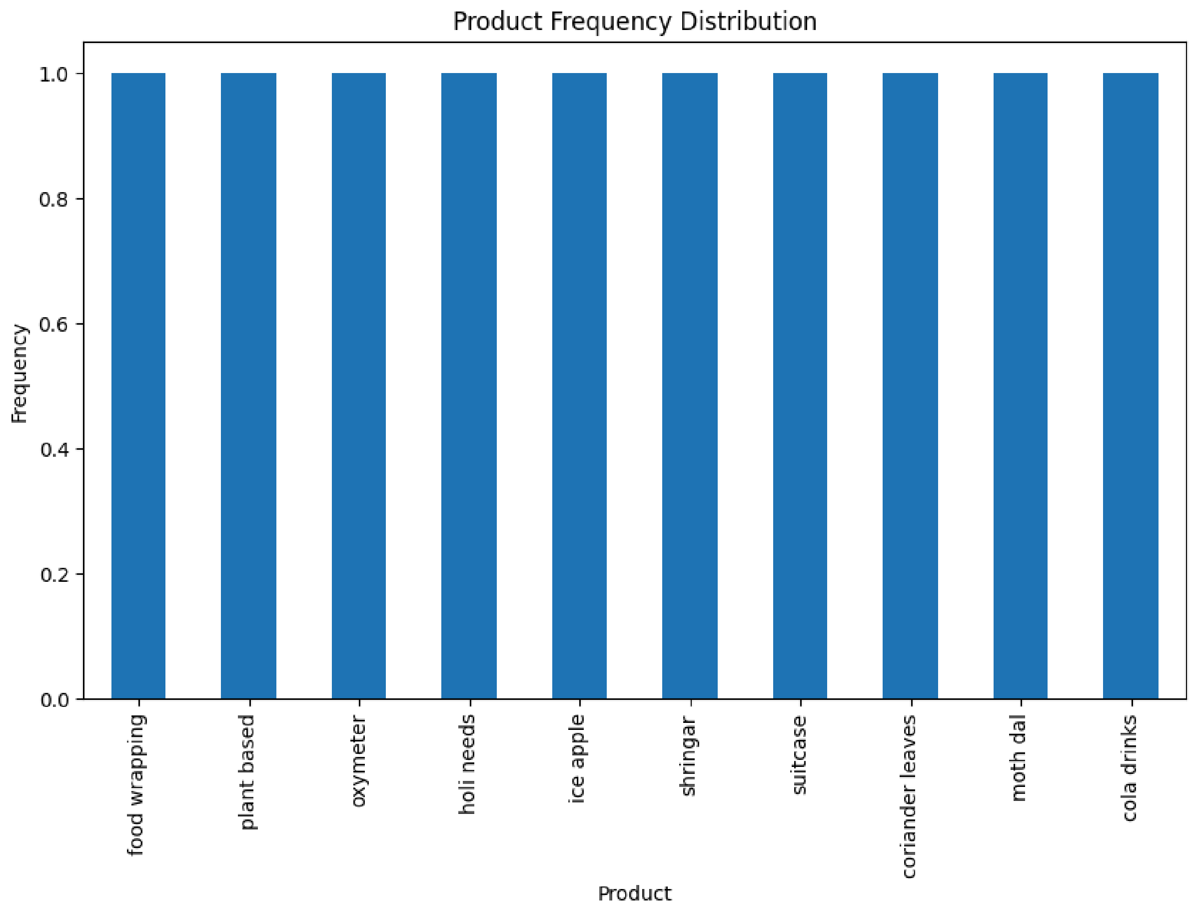
Top 20 Sessions by Number of Products Bought_Train set
Number of Products
0
2
4
6
8
Session ID
27d86d946e6ea6d022f12211d61f8ac7
b5e3e95a7d352b517e6256a3d761683e
84e42be037b32e6ad6d68edbc5972b10
f759181ac5f907209bf187295b92b89d
46e7a5a9ea314c5e08b15cbdc6a51a17
00f8c3d2b1e1cce223c3e53d56b62d4e
9e80ae7ba94a268f6b20a3cad1123955
7e1afb079c62883e07c208ac51c281b8
4c9de827d6346deb109b5fbac78bc19c
338c7191894e030b69af76ce6967a56f
58664bc2ed10c58758857495d5b05ab8
1f39f073264b1774a55f6a2011b0897e
3b5c105aed345142384f8ebcc3483d72
3faaa3bc8f2acfbbc40a25365c64cc37
14cef1293ce8e339614c0fcce14a779b
e1c89481f651dcb8a98fc9bc5220022c
181faaadfc95a23ccb2aecdb0c715ec3
d357b494fa150263cce9e7f46a674e8e
27009d3791a3b79c05d172057bdf2500
61598073846841 73a3d472078a958007



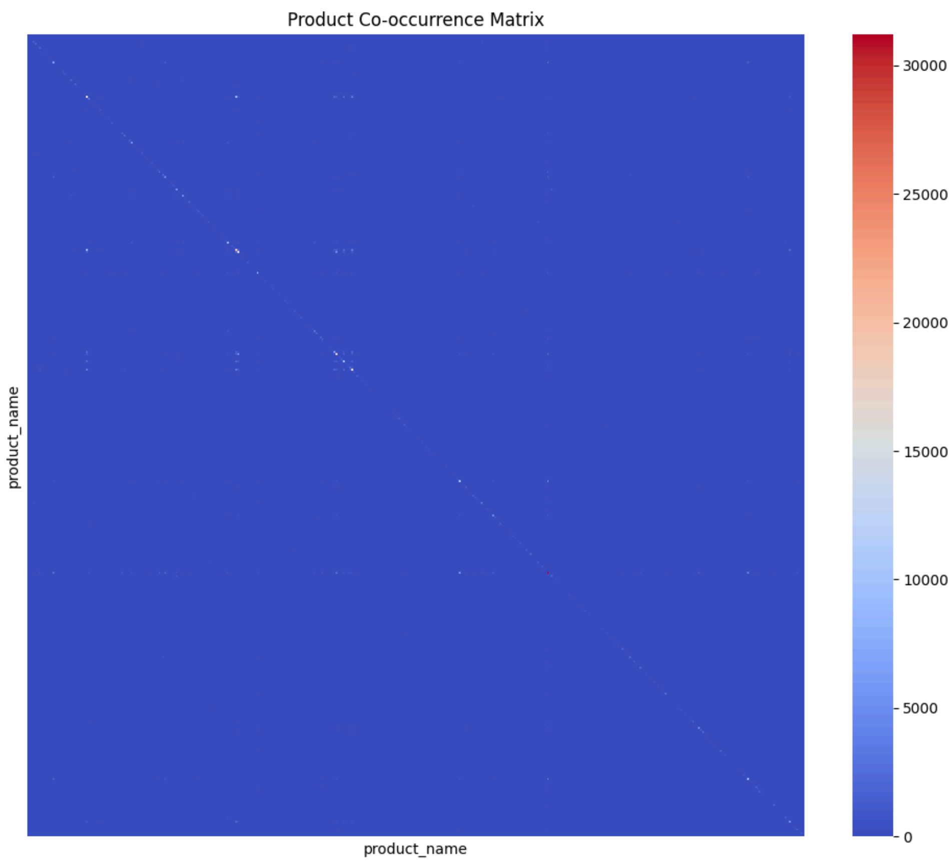
Top 20 products by frequency:



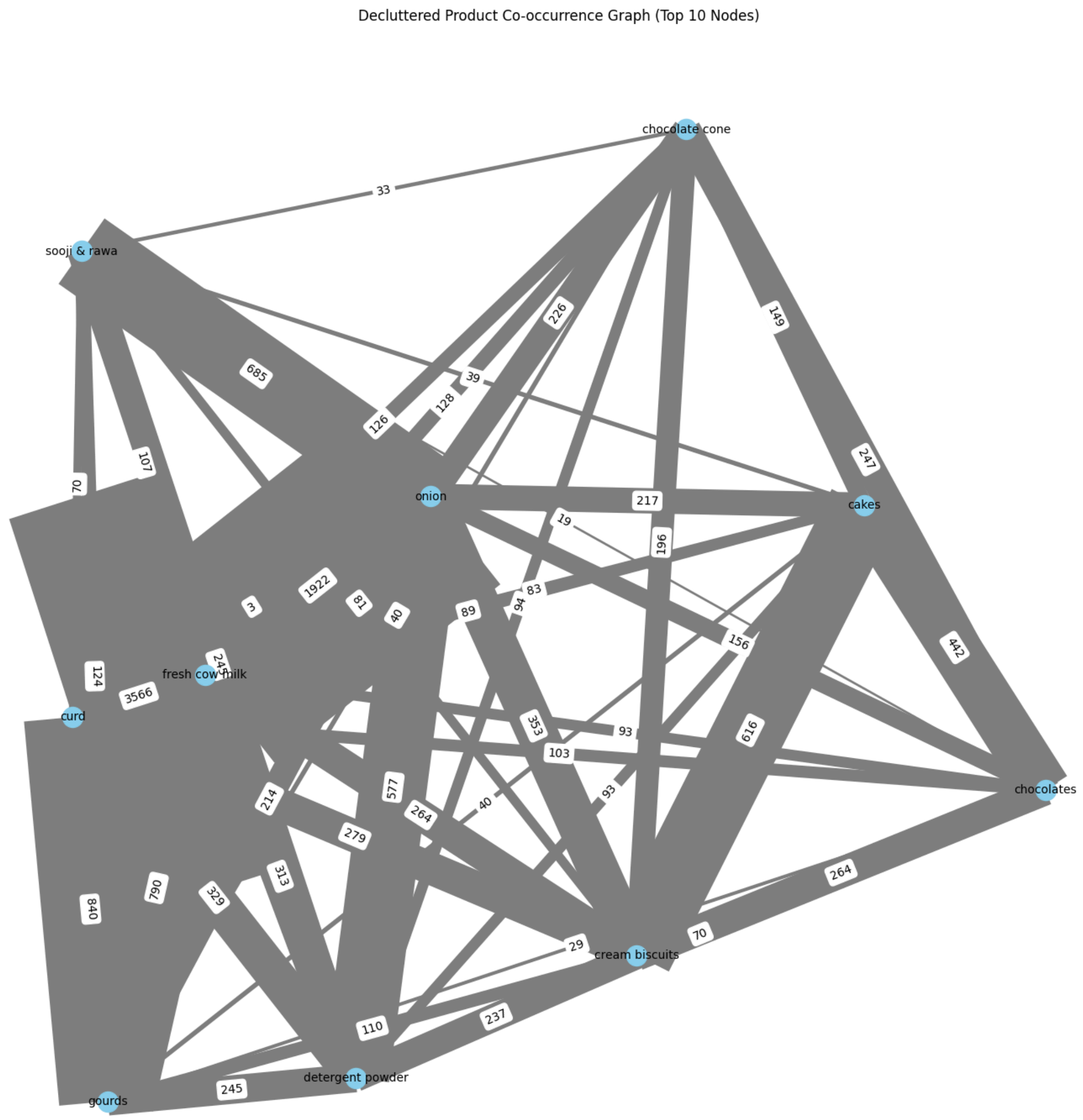
Last 20 products by frequency:



Product Co-occurrence Matrix:



Graphical relationship in the data (thicker edges exhibit higher co-occurrence):



Unique Insights from EDA:

1. The products have little co-occurrence with each other
2. The products that very frequently bought may be given as generic recommendations because of their higher co-occurence in the cart (For example, there may not be any contextual reason for onion to be correlated to other items in the cart, but it is a part of almost every user's cart as a daily item)
3. The mode of the number of products per session is 2, indicating that most users buy 2 products per session. Also, we can see customers buying a maximum of 9 products per session. Thus, a co-occurrence based approach can be used to recommend items based on products that frequently occur in the same cart with items in the current cart.

Specifics of model training:

I tried out 3 approaches:

Association Rules:

1. Data Preparation: The input is a transaction dataset (mentioned as basket in the code) where each transaction contains a list of items bought together.
2. Model Training:
   - Frequent Itemsets Generation: Use algorithms like Apriori or FP-Growth to find frequent itemsets (sets of products that appear together in transactions above a specified support threshold).
   - Rule Generation: Generate association rules from the frequent itemsets. Each rule has an antecedent (if) and a consequent (then) with metrics like support, confidence, and lift.
3. Output: A set of association rules that can be used for making recommendations.

Graph-Based Recommendations using Jaccard Coefficient:

1. Data Preparation: The input is a transaction dataset where each transaction contains a list of items bought together.
2. Model Training:
   Graph Construction: Built a graph where nodes represent products and edges represent co-occurrence of products in transactions. The weight of an edge is the number of times the two products appear together.
3. Output: A graph with nodes and weighted edges representing product co-occurrences.

Results:

Using Association Rules:

```
current_cart = ['banana', 'onion']
recommended_products = recommend_products(rules, current_cart)
print(f"Recommended products: {recommended_products}")
```

```
Recommended products: ['baby fruits', 'long shelf life milk', 'tender coconut']
```

Using Graph based recommendations:

```
current_cart = ['banana', 'onion']
```

```
Recommended products: ['curd', 'fresh cow milk', 'cucumber', 'long
shelf life milk', 'gourds', 'beetroot', 'tender coconut', 'grapes',
'baby fruits', 'detergent powder']
```

The various assumptions are being made with the problem statement and the proposed solution:

Assumptions in the PS:

1. Products frequently bought together are complementary. (As illustrated earlier in the case of onion, it might not be the case)
2. Past purchasing behaviour is indicative of future behaviour.
3. Preferences are static and do not depend on time/season etc.

Assumptions in the solution:

Association Rules:

1. Items that frequently appear together in transactions have a meaningful relationship.
2. The chosen thresholds for support and confidence are appropriate for identifying significant itemsets and rules.
3. Relationships between items are considered binary (bought together or not), without considering the order or quantity. (Because customers may buy items in a particular order; co-occurence of particular items may be season based; if customers are buying a particular product in large quantity, taking quantity into account gives better predictions than predicting based on single purchase)

Graph- based recommendations:

1. That the chosen similarity metric (e.g., Jaccard Coefficient) accurately captures the relationship between items.
2. That the items co-occurring in transactions can be represented as a graph with meaningful edges.
3. That we can ignore the contextual relationship between the items and solely focus on the frequency of co-occurrence.

Scope for further improvement given more data fields:

Can break down customer pull into: **Pre-sales -> During buying -> Post buying**
Therefore the following data can be very useful in each of the stages

1. User Data

- **User Preferences**: Explicit preferences such as liked/disliked products or categories.
- **Purchase History**: Historical data on past purchases can help identify long-term patterns and preferences.

- **Behavioural Data**: Clickstream data, browsing history, and time spent on different products or categories.

2. Product Data

- **Product Attributes**: Detailed attributes like brand, category, price, size, colour, ingredients, etc.
- **Product Descriptions**: Textual descriptions, reviews, and ratings

3. Contextual Data

- **Time of Purchase**: Time of day, day of the week, seasonality, etc., to capture temporal patterns.
- **Location**: Geographic location of the purchase, which can influence product preferences.

4. Transactional Data

- **Basket Data**: Detailed information about what products are bought together, including quantities.
- **Price Sensitivity**: Information on how changes in price affect purchase behaviour.
- **Discounts and Promotions**: Impact of sales and promotions on purchasing decisions.

6. Feedback and Ratings

- **Explicit Feedback**: Ratings and reviews provided by users.
- **Implicit Feedback**: Data inferred from user interactions, such as time spent on a product page or frequency of visits.

Code: Zepto_assignment_21EC39023.ipynb

Steps to Interpret the Graph

1. **Identify Frequently Co-occurring Products**:
   - Look for thick edges. These indicate pairs of products that are frequently bought together.
   - For example, "whole wheat breads" and its neighboring products have thick edges, indicating they are often purchased together.
2. **Clustered Groups**:
   - Products that are closely connected and form clusters likely belong to similar categories or are complementary items.
   - For instance, "veg nuggets" and "almond" have several connections, indicating these are part of a frequently purchased set of products.
3. **Outliers**:

    - Nodes that are connected with fewer edges or are more isolated may represent products that are not commonly purchased with others.
    - Products like "sooji rawa" and "sanitary pads" are less interconnected, suggesting they are less commonly bought with a wide variety of other products.
4. **Product Popularity**:
    - Nodes with a high degree (many connections) indicate popular products that are frequently bought with many other items.
    - "veg nuggets" and "whole wheat breads" seem to be highly connected, suggesting they are popular items in many sessions.
5. **Edge Weights**:
    - The numbers on the edges provide a quantitative measure of co-occurrence frequency. Higher numbers indicate more frequent joint purchases.
    - An edge weight of 1150 between two products signifies very frequent co-purchases compared to an edge with a weight of 6.

Additional Analysis

- **Central Products**: Identify products that serve as hubs connecting various other products.
- **Category Analysis**: Analyze if certain categories of products (e.g., dairy, snacks) are more interconnected.

Example Interpretation

- **Highly Connected Products**:
    - "veg nuggets" is a central node with numerous thick edges, indicating it is a commonly bought product with various other items.
    - "whole wheat breads" also shows numerous connections, suggesting it's a staple in many shopping sessions.
- **Less Connected Products**:
    - "sooji rawa" appears less connected, indicating it is bought less frequently with other products.
- **Frequent Pairs**:
    - Products connected with a weight of 1150 (if there are any) show an extremely high frequency of being bought together, suggesting a strong complementary relationship.

Further Steps

- To gain deeper insights, you could analyze subgraphs focusing on specific categories of products.
- You could also use this graph to identify potential opportunities for cross-selling by looking at products that are frequently bought together.