ASSOCIATION RULES

The Objective of this assignment is to introduce students to rule mining techniques, particularly focusing on market basket analysis and provide hands on experience.

Dataset:
Use the Online retail dataset to apply the association rules.

Data Preprocessing:
Pre-process the dataset to ensure it is suitable for Association rules, this may include handling missing values, removing duplicates, and converting the data to appropriate format.
Answer:

Code used : online_retail.py

Data Preprocessing
Objective:
Before applying Association Rule Mining, the dataset must be cleaned and structured into a suitable format (transactions).

Steps taken:
1. **Dataset structure:**
    - The Online Retail dataset consisted of **7500 rows**.
    - Each row represented a **transaction (basket)**, containing a comma-separated list of items purchased together.

    Example rows:
    burgers, meatballs, eggs
    chutney
    turkey, avocado
    mineral water, milk, energy bar, whole wheat rice, green tea
    low fat yogurt
2. **Tokenization of items:**
    - Each transaction string was split into a list of individual products.
    - Example:
      "burgers, meatballs, eggs" → ['burgers', 'meatballs', 'eggs']
3. **Removing duplicates within a basket:**
    - If an item appeared more than once in the same basket, duplicates were removed.
    - This ensured each transaction is a set of unique products.
4. **One-hot encoding:**
    - To apply Apriori, we converted the dataset into a **binary matrix**:
        - Rows = transactions
        - Columns = products
        - Values = 1 if product is present, 0 otherwise.

    Example encoded table (first 5 rows):

burgers	meatballs	eggs	chutney	turkey	avocado	mineral water	milk	green tea	low fat yogurt
1	1	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0

burgers	meatballs	eggs	chutney	turkey	avocado	mineral water	milk	green tea	low fat yogurt
0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	0	0	0	1

This preprocessing step prepared the dataset for **frequent itemset mining** and the generation of association rules.

Association Rule Mining:

- **Implement an Apriori algorithm using tool like python with libraries such as Pandas and Mlxtend etc.**

Answer:

Code used : frequent.py

What this code does:

- **If basket_one_hot.csv exists at given folder it loads that and computes item supports directly.**
- **Otherwise it expects a transactions Python variable (list of lists) in memory and computes supports from that.**
- **Filters items with support ≥ MIN_SUPPORT (default 0.05).**
- **Prints and saves the result to frequent_items_single.csv.**

- **Apply association rule mining techniques to the pre-processed dataset to discover interesting relationships between products purchased together.**

Answer:

Code used : pairwise_rules.py

- **pairwise_rules.csv with columns:**
  - **antecedent, consequent, support, confidence, lift, pair_count**
- **The code only considers pairs derived from items that meet MIN_SUPPORT (so it's efficient).**
- **Rules are sorted by lift (descending) then confidence.**

Quick tips

- **Raise MIN_SUPPORT if you want fewer, stronger rules. Lower it to explore more rare combos (but expect explosion in pairs).**
- **Set MIN_CONFIDENCE = 0.2 (or whatever you like) to filter weak implications.**
- **Lift > 1 means positive association; >1.3–1.5 is often interesting in retail contexts but context matters.**

- **Set appropriate threshold for support, confidence and lift to extract meaning full rules.**

Answer :

Choosing Thresholds for Association Rule Mining

1. **Support (≥ 0.05 or 5%)**
    - Support measures how often an itemset appears across all transactions.
    - A **too low threshold** will generate thousands of rules, most of which are spurious (noise).
    - A **too high threshold** may miss interesting but less frequent patterns.
    - In our dataset (7500 transactions), we chose **5% minimum support**, meaning the product pair must appear in at least ~375 transactions to be considered.
    - This strikes a balance: captures popular patterns without overwhelming the analysis.
2. **Confidence (≥ 0.2 or 20%)**
    - Confidence measures the probability of buying the consequent given the antecedent.
    - A rule like *spaghetti → mineral water* with **confidence 34%** means that 34% of people who bought spaghetti also bought mineral water.
    - We set **20% confidence threshold** to ensure rules represent reasonably strong conditional relationships (not just random co-occurrences).
3. **Lift (> 1.2)**
    - Lift tells us how much more likely items occur together compared to being independent.
    - A lift of **1.0** means no real association (just chance).
    - We kept only rules with **lift > 1.2**, meaning the relationship is at least 20% stronger than random expectation.
    - Example: *spaghetti → mineral water* had **lift ≈ 1.44**, a strong positive association.

Final Thresholds Used

- **Support ≥ 5%**
- **Confidence ≥ 20%**
- **Lift > 1.2**

These thresholds helped us extract meaningful and interpretable rules (like “mineral water is a hub product bought with spaghetti and chocolate”) while filtering out trivial or misleading ones.

Analysis and Interpretation:

- **Analyse the generated rules to identify interesting patterns and relationships between the products.**
-

Answer :

Analysis of Generated Rules

After applying Apriori with thresholds (Support ≥ 5%, Confidence ≥ 20%, Lift > 1.2), several interesting product relationships were discovered:

1. **Mineral Water as a “hub” product**
    - *Spaghetti → Mineral Water*
        - Support ≈ 5.9% | Confidence ≈ 34% | Lift ≈ 1.44
    - Customers buying spaghetti are much more likely to also purchase mineral water. This suggests meal-planning behavior (pasta dishes + beverages).
    - The high lift confirms this is not just because mineral water is popular, but because it co-occurs disproportionately often with spaghetti.
    - *Chocolate → Mineral Water*
        - Support ≈ 5.3% | Confidence ≈ 32% | Lift ≈ 1.35
    - A strong tendency for customers buying chocolate to also buy mineral water. This could reflect impulse purchases of sweets alongside drinks.

2. **Eggs with Mineral Water**
   - *Eggs → Mineral Water*
     - Support ≈ 5.1% | Confidence ≈ 28% | Lift ≈ 1.19
   - Eggs often appear in larger grocery baskets, and mineral water seems to be a consistent companion product. This suggests mineral water is a "default add-on" in shopping trips.
3. **Patterns of Cross-Selling**
   - Mineral water is central to multiple strong rules.
   - It plays the role of a **gateway item**, linking with both meal staples (spaghetti, eggs) and indulgences (chocolate).
   - Retailers could use this by creating **combo offers**:
     - "Spaghetti + Mineral Water" as a family meal bundle.
     - "Chocolate + Mineral Water" as a quick snack/drink combo.
4. **Insights on Customer Behavior**
   - **Planned Meals:** Customers buying ingredients like spaghetti also add beverages (water), showing meal-based shopping.
   - **Impulse/Convenience:** Pairings like chocolate with water hint at small indulgent purchases bundled with essentials.
   - **Mineral Water's Anchor Role:** Since mineral water co-occurs across very different categories, it acts as a common denominator in grocery baskets.

Conclusion from Analysis

- Mineral water consistently appears in strong association rules, making it the most influential product in the dataset.
- The discovered patterns can be directly used in **promotion strategies, store placement (e.g., keeping mineral water near staples), and combo discounts** to increase sales.

- **Interpret the results and provide insights into customer purchasing behaviour based on the discovered rules.**

Answer:

Interpretation and Customer Insights

The association rules reveal clear trends in customer purchasing behaviour:

1. **Mineral Water as a Core Basket Item**
   - Mineral water shows up in multiple strong rules with spaghetti, chocolate, and eggs.
   - This indicates it is a **default companion product**, suggesting customers frequently add water to their baskets regardless of the main purchase.
   - Customer mindset: "If I'm shopping anyway, I might as well stock up on water."
2. **Meal-Oriented Shopping**
   - The rule *Spaghetti → Mineral Water* suggests that customers buying pasta are often meal planning, and beverages are a natural complement.
   - This indicates **planned, recipe-driven shopping trips**, where items are purchased together to complete a meal.
3. **Snack/Impulse Behaviour**
   - The *Chocolate → Mineral Water* rule reflects small indulgence purchases paired with drinks.
   - This points to **impulse buying behaviour**, where customers add a drink to go with a snack (or vice versa).
4. **Cross-Selling Opportunities**

- Since mineral water connects with diverse categories (staples like eggs, indulgences like chocolate, and meal ingredients like spaghetti), it acts as a **cross-category anchor product**.
- Business implication: bundle offers and product placement can leverage this — e.g., placing mineral water near pasta shelves or near confectionery aisles to stimulate additional purchases.

Key Customer Insights

- Customers treat **mineral water as a staple add-on**, often purchased alongside very different categories.
- **Meal planners** buy complementary products together (pasta + water).
- **Impulse buyers** tend to pair indulgences (chocolate) with essentials (water).
- Retailers can design **combo deals, shelf placement strategies, and targeted promotions** around these associations to increase sales and enhance customer satisfaction.

Interview Questions:

1. **What is lift and why is it important in Association rules?**
   a. Lift = (Confidence of A→B) / (Support of B).
   b. It measures how much more likely A and B occur together than if they were independent.
   c. Lift > 1 means a positive association. It's important because high confidence alone might be misleading if the consequent is just a very popular item.

2. **What is support and Confidence. How do you calculate them?**
   a. Support(A→B) = Probability(A and B occur together) = count(A∪B) / total transactions.
   b. Confidence(A→B) = Probability(B occurs given A) = Support(A∪B) / Support(A).

3. **What are some limitations or challenges of Association rules mining?**

   - Generates a huge number of rules, many of which are not meaningful.
   - Choosing thresholds for support, confidence, lift is subjective.
   - Computationally expensive for large datasets (explodes with combinations).
   - Doesn't consider time/order (solved by sequence mining).

Here's the full breakdown of **Association Rule Mining** assignment with the Online Retail dataset:

Data Preprocessing

- The dataset was structured as transactions (each row = a shopping basket).
- I split the string items into lists and converted them into a one-hot encoded table (each column = product, each row = 1 if purchased).
- Removed duplicates automatically during encoding. Missing values were not an issue here.

Frequent Itemsets (Support ≥ 5%)

Top single items:

- Mineral water: **23.8%**
- Eggs: **17.9%**
- Spaghetti: **17.4%**
- French fries: **17.1%**
- Chocolate: **16.4%**

Association Rules (pairs only, sorted by Lift)

Antecedent	Consequent	Support	Confidence	Lift
Spaghetti → Mineral water	5.9%	34.3%	**1.44**	
Mineral water → Spaghetti	5.9%	25.1%	**1.44**	
Mineral water → Chocolate	5.3%	22.1%	**1.35**	
Chocolate → Mineral water	5.3%	32.1%	**1.35**	
Eggs → Mineral water	5.1%	28.3%	1.19	

Interpretation:

- **Spaghetti & Mineral Water** are strongly associated (lift > 1.4). This means they co-occur more often than chance.
- **Chocolate & Mineral Water** also show a meaningful relationship.
- Mineral water acts as a “hub” product — commonly bought with many items.

Analysis & Insights

1. **Mineral water** is a frequent anchor product. Customers who buy it tend to also buy spaghetti or chocolate.
2. **Spaghetti + Mineral Water** could signal planned meals (pasta dishes with water).
3. Retailers could design **combo offers** around these associations to boost cross-sales.