Data Mining

Assignment 2

Building a Clean Data Warehouse for Market Basket Discovery

Objective: To transition from a "dirty" unstructured raw dataset to a professionally modeled Data Warehouse (Star Schema) and use that structure to perform Data Mining (Association Rule Mining) to discover product purchase patterns.

Data Source

Dataset to use: Retail Store Sales (Dirty for Data Cleaning)

Link: https://www.kaggle.com/datasets/ahmedmohamed2003/retail-store-sales-dirty-for-data-cleaning

This dataset is selected because it mimics the "Handling diverse, unstructured data formats" challenge found in real-world projects. It contains intentional errors, missing values, and inconsistent formatting in columns like Transaction ID, Item, and Price.

Assignment Tasks & Requirements

Part A (*30 points*): Data Cleaning (Extraction & Transformation)

Before the data is "mining-ready," you must build a preprocessing pipeline.

- **Standardization:** Handle inconsistent date formats and case-sensitivity in product names (e.g., "coffee" vs "Coffee").
- **Missing Values:** Identify and resolve null values in item names or price, etc. If the null value in a record cannot be resolved, remove it from your dataset.
- **Business Logic Validation:** Ensure Total Spent equals Quantity × Price. Flag or fix rows that fail this integrity check.

Part B (*40 points*): Data Warehouse Modeling (The Star Schema)

Design and implement a Star Schema to store your cleaned data. This structure must enable efficient querying for business insights.

- **Sales Table:** Create a Sales table containing measurable data (Quantity, Price, Total, etc.).
- **Other Tables:** Create Product, Customer, and Date tables. The Date table pre-calculates attributes like Is_Weekend, Month_Name, Quarter, and Fiscal_Year. You

then simply use a foreign key from Sales table to Date table. This will help in analyzing sales such as whether a specific product is sold more at weekends, etc.

- **Requirements:** Tables must be linked via Foreign Keys.
- **Note:** Table design and choosing right attributes for each table are part of the assignment requirements.

Part C (*30 points*): Finding Associations (Market Basket Analysis)

Once the data is in your warehouse, perform a Data Mining task to find product associations.

- Identify which products are frequently purchased together in the same transaction.
    - **Note:** TransactionID in this dataset is used as a RowID and does not represent an actual transaction. To determine which products were bought together, consider CustomerID, Transaction Date, and Location (In-store / Online). You may assume a customer performs only one online or in-store purchase on a single day.
- Do you see the same association between products in different months?
    - **Note:** The dataset includes data from January 2022 until December 2025 (48 months). Perform the association calculations by considering each month separately.

Deliverables

You must submit a package containing:

1. **Code:** Python or SQL code showing the cleaning and normalization steps. Also includes comments in your code to show how you handled the null attributes.
2. **Architecture Diagram:** A visual representation of your Star Schema (Database tables).
3. **The Warehouse:** A screenshot of the cleaned data.
4. **Mining Report:** A summary of the top 5 product associations found, including the differences you detected in different months.