**Dataset Overview**

The **Online Retail Dataset** is a transactional dataset collected from an online retail business based in the United Kingdom. It captures customer purchases made between **December 2010 and December 2011**. The dataset provides a comprehensive view of customer behavior and purchasing patterns, making it suitable for tasks like exploratory data analysis, customer segmentation, sales forecasting, and basket analysis.

**Key Features**

* **InvoiceNo**: A unique identifier for each transaction (Invoice).
* **StockCode**: A unique product code for each item.
* **Description**: A textual description of the product.
* **Quantity**: The number of products purchased in each transaction.
* **InvoiceDate**: The date and time of the transaction.
* **UnitPrice**: Price per unit of the product.
* **CustomerID**: A unique identifier for each customer (some transactions may lack this).

**Characteristics**

* **Size**: Approximately 540,000 rows and 8 columns.
* **Data Type**: Mixed types, including numerical, categorical and date time
* **Geographical Scope**: Includes transactions from customers in over 30 countries, primarily from the UK.

**Feature Engineering**

* Created a new column **TotalAmount** to calculate the total value of each transaction **(Quantity x UnitPrice)**.
* Extracted time-based features (**Month, Weekday**, and **Hour**)from the **InvoiceDate** column to facilitate trend analysis.

**Exploratory Data Analysis (EDA) Description**

* **Descriptive Statistics:** 
  + **Summarized numerical data to understand distributions, averages, and variations in features like Quantity, UnitPrice and TotalAmount**
* **Visualizations:** 
  + **Created histograms to analyze the distribution of TotalAmount and identify potential outliers.**
  + **Identified the top 10 most popular products based on the quantity sold using bar charts.**
  + **Visualized sales trends across countries to highlight geographical differences in revenue.**
* **Trend Analysis:** 
  + **Explored customer purchasing behavior over time, focusing on monthly, daily, and hourly patterns.**

**Customer Segmentation (Clustering)**

In the **Customer Segmentation (Clustering)** step, we grouped customers into distinct segments based on their purchasing behavior to better understand their characteristics and preferences.

1. **Feature Selection**:
   * Aggregated customer data to calculate metrics such as total spending **(TotalAmount)**, the number of unique transactions **(InvoiceDate)**, and total products purchased **(Quantity**).
2. **Data Normalization**:
   * Standardized the selected features to ensure fair clustering by eliminating the effects of scale differences.
3. **K-Means Clustering**:
   * Applied the K-Means algorithm to classify customers into three segments based on their purchasing behavior.
4. **Visualization**:
   * Visualized the customer clusters to identify patterns in spending and transactional habits.

**Basket Analysis (Association Rules)**

In this step, we used the **Apriori algorithm** to identify frequent itemsets from the online retail dataset and generate **association rules** based on these itemsets. The key goal was to identify which items tend to be purchased together frequently, using **lift** as the metric for strong associations.

**Key Findings:**

* The **support** of the association rules represents the frequency with which the itemsets appear in the dataset. The support values for the rules found ranged from **2.3% to 3.2%**, indicating that the associated items are purchased together in a significant portion of the transactions.
* The **lift** values for the association rules ranged from **16.6** to **20.4**, which are significantly greater than 1. A lift value greater than 1 indicates a positive relationship between the items, meaning that the items are more likely to be purchased together than by random chance. The high lift values suggest strong and interesting relationships between the pairs of items.

**Sample Association Rules:**

1. **(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY TEACUP AND SAUCER)** – Support: 0.027, Lift: 20.36
2. **(PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)** – Support: 0.0257, Lift: 17.11
3. **(GARDENERS KNEELING PAD CUP OF TEA, GARDENERS KNEELING PAD)** – Support: 0.027, Lift: 16.63

These rules highlight specific item combinations that have a strong tendency to be bought together. For example, **PINK REGENCY TEACUP AND SAUCER** is frequently purchased with both **GREEN REGENCY TEACUP AND SAUCER** and **ROSES REGENCY TEACUP AND SAUCER**, showing clear patterns of related products in the dataset.