**Market Basket Analysis (MBA)**

**Abstract:** This project applies **Market Basket Analysis (MBA)** to a publicly available dataset from an **online retail** store, located in the United Kingdom. The dataset includes transactions recorded between December 2010 and December 2011, capturing invoices, product descriptions, quantities purchased, and customer IDs. The primary objective of this project is to uncover patterns in customer purchasing behavior, identify frequent itemsets, and generate association rules to help the retailer optimize product placement, cross-selling strategies, and inventory management.

**Data Understanding:**

**The dataset includes information on invoices and sales with the following key features:**

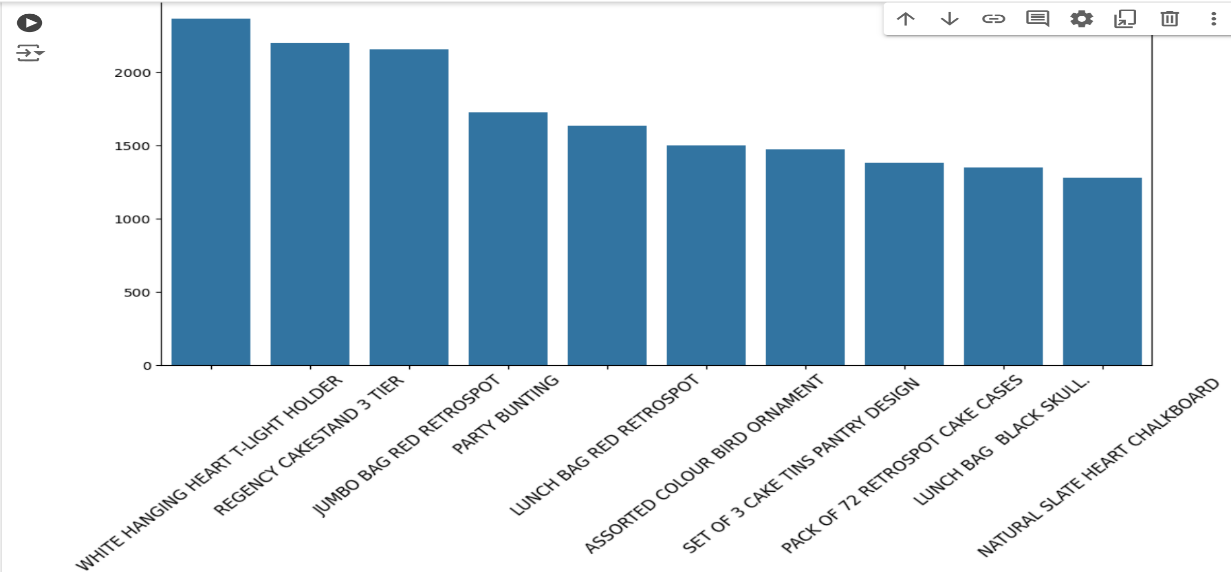
1. **InvoiceNo**: Unique identifier for each sale transaction, allowing for invoice tracking.
2. **StockCode**: Product identifier in inventory, useful for monitoring sales performance.
3. **Description**: Details about the sold products, aiding in product categorization and customer preference analysis.
4. **Quantity:** Number of units sold, providing insights into sales volume.
5. **InvoiceDate:** Date and time of the transaction, useful for analyzing sales trends over time.
6. **UnitPrice:** Price per unit sold, important for calculating total revenue.
7. **CustomerID:** Unique identifier for each customer, enabling analysis of purchasing behavior.
8. **Country:** Country of sale, helpful for understanding geographical market dynamics.

**Data Processing Steps:**

1. **Data Loading:** Load the dataset into a DataFrame from an external file
2. **Data Cleaning:** Remove missing values, duplicate invoices, and return/cancellation transactions.
3. **Data Transformation:** Convert the data into a binary matrix using One-Hot Encoding, where each product is a column.
4. **Data Filtering: Retain** only transactions with at least two distinct products for meaningful analysis.
5. **Frequent Itemset Extraction:** Apply the Apriori algorithm to find frequent itemsets based on a minimum support threshold.
6. **Association Rule Extraction:** Generate association rules using metrics like confidence and lift to understand product relationships.

**Model Choice:** Apriori Algorithm

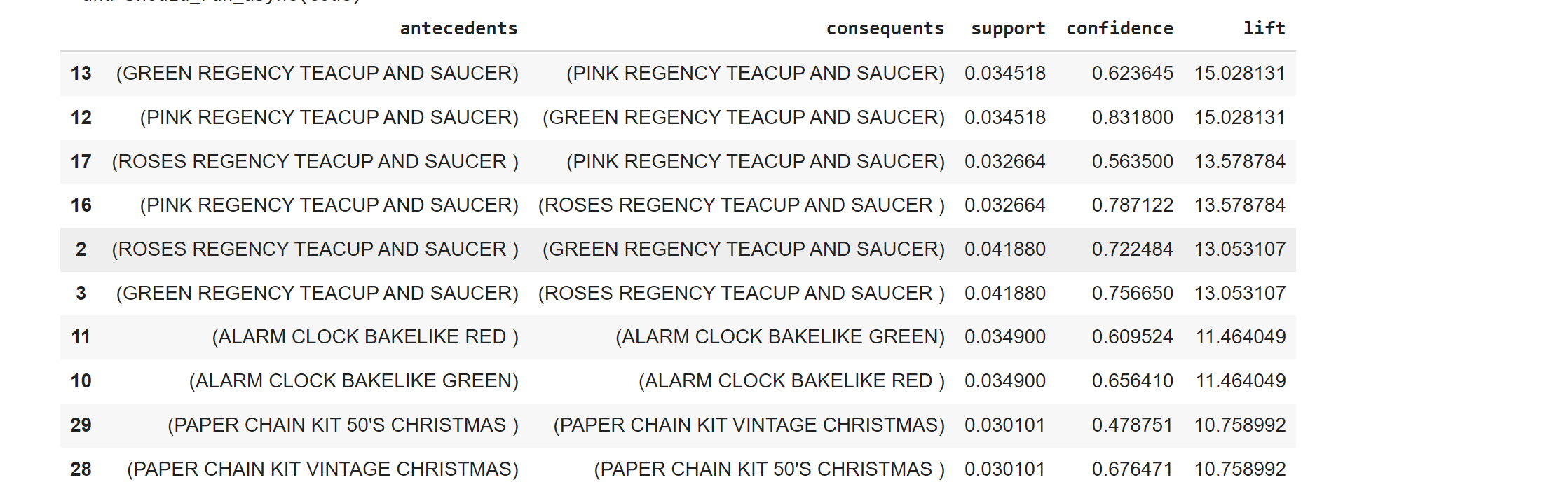
**Here we visualize the 10 best selling products:**

****

**The bar plot representing the top 10 best-selling products based on the number of sales for each item. Here's the interpretation of the results:**

1. The highest-selling product is WHITE HANGING HEART T-LIGHT HOLDER, with approximately 2000 sales.
2. It is closely followed by REGENCY CAKESTAND 3 TIER, which has almost the same number of sales, making them nearly equal in terms of popularity.
3. Other products like JUMBO BAG RED RETROSPOT, PARTY BUNTING, and LUNCH BAG RED RETROSPOT also have high sales but fall slightly behind the top two.

**The Top 10 Rules based on lift:**

****

**Insights Gained:**

1. **Strong Cross-Purchasing Patterns:**

The associations between GREEN REGENCY TEACUP AND SAUCER and PINK REGENCY TEACUP AND SAUCER demonstrate a robust cross-purchasing pattern, with a lift of 15.03. This suggests that customers who buy one are highly inclined to purchase the other, indicating a strong preference for these complementary items. Retailers can leverage this insight by promoting these teacups as a matching set to boost sales.

1. **High Customer Preference for Sets:**

The significant lift values (e.g., 13.58 between **ROSES REGENCY TEACUP AND SAUCER and PINK REGENCY TEACUP AND SAUCER**) indicate that customers often choose to purchase multiple styles within the same category. This suggests a strong inclination towards having coordinated tea sets, highlighting the potential for effective marketing strategies that focus on themed promotions or bundled offers**.**

1. **Complementary Product Marketing:**

The findings show a consistent trend in purchasing complementary products, particularly within the **REGENCY TEACUP AND SAUCER** range. For example, the relationships among the green, pink, and roses teacups indicate an opportunity for retailers to create attractive promotional displays or targeted marketing campaigns that highlight these complementary products together.

1. **Color Variation Appeal:**

The high confidence in rules involving variations of the **ALARM CLOCK BAKELIKE models (e.g., ALARM CLOCK BAKELIKE RED and GREEN)** suggests that customers are often drawn to variations of the same product. This insight can guide retailers in their recommendation strategies, encouraging cross-selling of different color options to satisfy customer preferences for variety.

1. **Seasonal Product Synergy:**

The relationships between the **PAPER CHAIN KIT 50'S CHRISTMAS and PAPER CHAIN KIT VINTAGE CHRISTMAS** highlight a notable synergy among seasonal products, with a lift of 10.76. This indicates that customers frequently buy multiple Christmas-themed items together, presenting an excellent opportunity for seasonal bundling strategies to maximize sales during the holiday period.