Market Basket Analysis Using Apriori Algorithm

Anhelina Naliuka 024712312B
Sofiia Parkhomets 024712322C

1. Project Objective and Motivation

The main goal of our project was to explore and analyze market basket data from a dataset of online stores using the Apriori algorithm. Specifically, we sought to identify product associations in the transactions of shoppers from the four top-selling countries: the United Kingdom (UK), France, Germany, and Ireland (EIRE).

Our analysis was driven by a desire to identify localized shopping patterns and country-specific product associations. Using associative rule mining, we wanted to extract useful insights that could improve marketing strategies, optimize product placement, identify cross-selling opportunities, and simplify regional targeting.

In addition, conducting such analysis in different geographical regions allows us to compare consumer behavior in different countries. The Apriori algorithm was chosen due to its efficiency in generating frequent itemsets and associative rules from transactional data, which is fully consistent with the structure of the retail dataset used in this study.

We used the Online Retail dataset from the UCI Machine Learning Repository. This dataset contains transactional data from a UK-based online retailer. It includes over 500,000 records of purchases made between December 2010 and December 2011. Each entry represents a product purchase in a specific invoice, with attributes such as “InvoiceNo”, “StockCode”, “Description”, “Quantity”, “InvoiceDate”, “UnitPrice”, “CustomerID”, and “Country”.

Before proceeding with the analysis, we performed a detailed data quality check to understand the structure and limitations of the dataset. We found that there are many missing values in the dataset, especially in the “CustomerID” and “Description” columns, as well as many duplicate rows. In addition, the “Description” column is a serious problem due to the large number of inconsistent and non-standardized product names (e.g., typos, different case variants, redundant terms).

We start with data preprocessing to clean and structure the dataset, and then apply the Apriori algorithm. We then visualize the strongest rules using heatmaps and provide interpretations and recommendations based on our findings.

2. Data Preprocessing

Data preprocessing is one of the most important steps in any data analysis process. In our project, it was particularly important because of the low initial quality of the dataset. The goal of this stage was to clean, standardize, and structure the data to effectively apply the Apriori algorithm and extract meaningful associative rules.

We started with basic quality checks and general cleaning operations. We left only the main columns needed for analysis: “InvoiceNo”, “Description”, ‘CustomerID’ and “Country”, removing unrelated or redundant fields to improve processing efficiency.

After that, we identified and removed rows with missing values in either “Description” or “CustomerID”, as these fields are essential for understanding what was purchased and by whom. Additionally, we excluded canceled transactions (identified by the letter 'C' in the “InvoiceNo”), as they do not represent successful sales and would distort results.

The “Description” column presented significant challenges due to inconsistent formatting, numerous typos, stopwords, and varying forms of the same product name. To address this:

- We converted all product names to lowercase for consistency.
- We applied fuzzy string matching to group and standardize similar product names.
- We identified key issues such as:
    - stopwords,
    - different forms of the same word,
    - unnecessary punctuation,
    - repeated spaces or numbers (e.g. “set of 6” or “SET6”),
    - and general textual noise.

To resolve these:

- We removed all non-letter characters and extra spaces.
- We deleted common stopwords and frequently appearing filler words.
- We used NLTK's lemmatization techniques to reduce words to their base form (e.g., “bags” → “bag”).
- We built and applied a custom dictionary of corrections to handle specific cases not resolved by automated methods.

This multi-stage preprocessing resulted in a much more consistent and cleaner version of the dataset, allowing us to present the same product under one standardized name, significantly improving the quality of associative rule mining.

3. Association Rule Mining using Apriori Algorithm

To discover patterns in customer purchasing behavior, as was previously said, we applied the Apriori algorithm, a classic algorithm in data mining used for association rule learning. We selected it because of interpretability (output rules are easy to understand), support and confidence metrics, which allow us to filter useful patterns, and suitability for structured data.

The Apriori algorithm operates in two main steps:

1. Frequent itemset generation — identify all product combinations that appear together in transactions with a frequency above a defined threshold (min_support).
2. Association rule generation — from these frequent itemsets, generate rules that imply certain product combinations given others, using min_confidence as a filtering criterion.

We implemented the algorithm separately for each of the four countries with the highest sales volume: the United Kingdom, France, Germany, and EIRE. The goal was to identify localized purchasing patterns and understand how customer behavior varies across regions.

Steps:

1. The Apriori algorithm requires transactions to be represented as lists of items. We grouped the data by "InvoiceNo" to form individual transactions, where each transaction is a list of purchased items. For example:
   Invoice 536365 -> ['white hanging heart t-light holder', 'red woolly hottie white heart.']

2. We used the TransactionEncoder from mlxtend to convert transactions into a binary matrix format, where each column represents a product, and each row is a transaction. This format is required for the Apriori algorithm to compute itemset support efficiently.
3. We applied the `apriori` function to generate frequent itemsets with a minimum support of 2%.
4. We used `association_rules` to generate rules based on a minimum confidence of 50%.
5. Duplicate or semantically identical rules were filtered out to reduce redundancy.

Example output: for France, one of the top rules was:

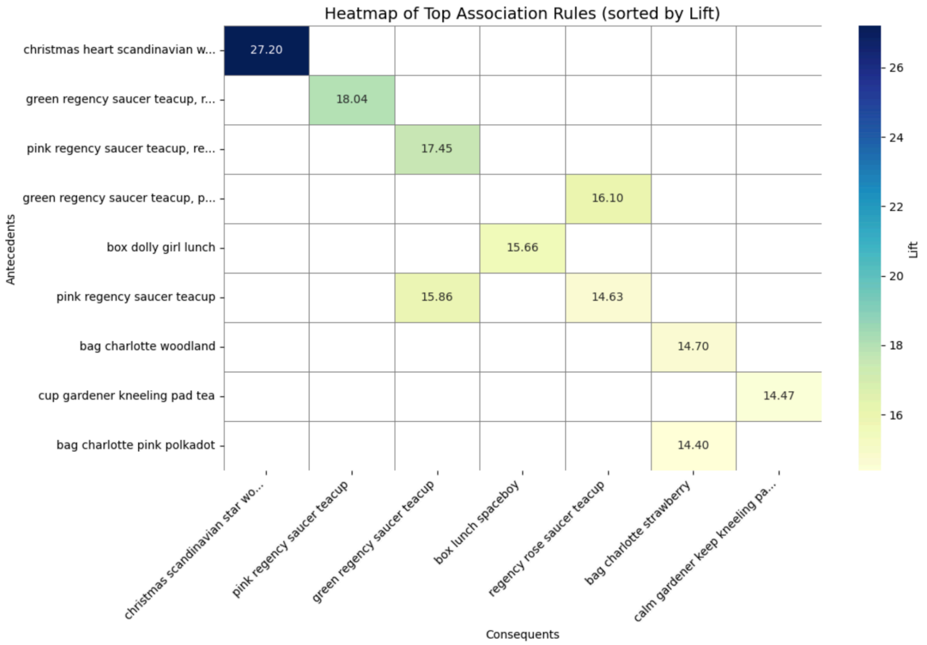
(acapulco mat recycled red, acapulco mat recycled turquoise) → acapulco lavender mat recycled (lift = 34.84)

Interpretation: When customers buy the red and turquoise versions of the acapulco mat, there is a very strong chance they will also buy the lavender one. This kind of insight can be used to optimize product placement and promotional bundling.

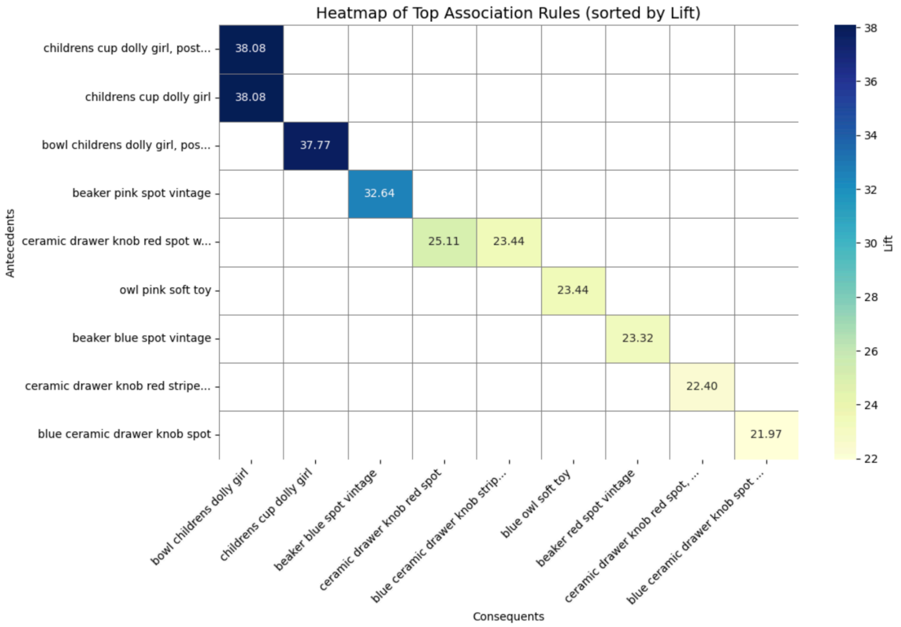
4. Visualization

Once the Apriori algorithm was applied and association rules were generated for each of the four countries, the next essential step was to visualize the results. Visual representation of association rules enhances interpretability, especially when dealing with large rule sets. It allows us to spot strong relationships quickly and make decisions based on visual insights rather than raw numbers.

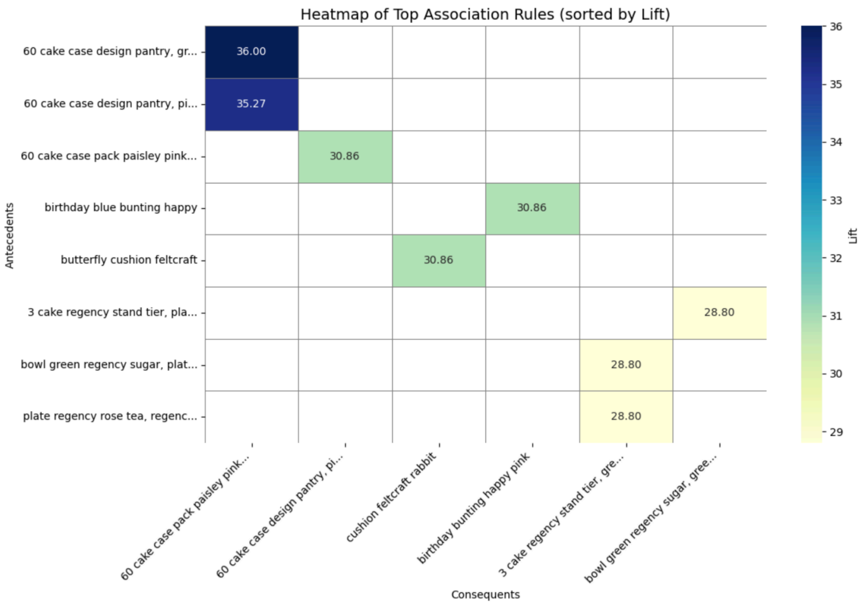
Heatmaps were chosen as the primary visualization tool because they allow for a compact yet powerful comparison of multiple association rules. Each cell in the heatmap represents a rule, with the antecedent on the y-axis and the consequent on the x-axis. The color intensity corresponds to the lift value.



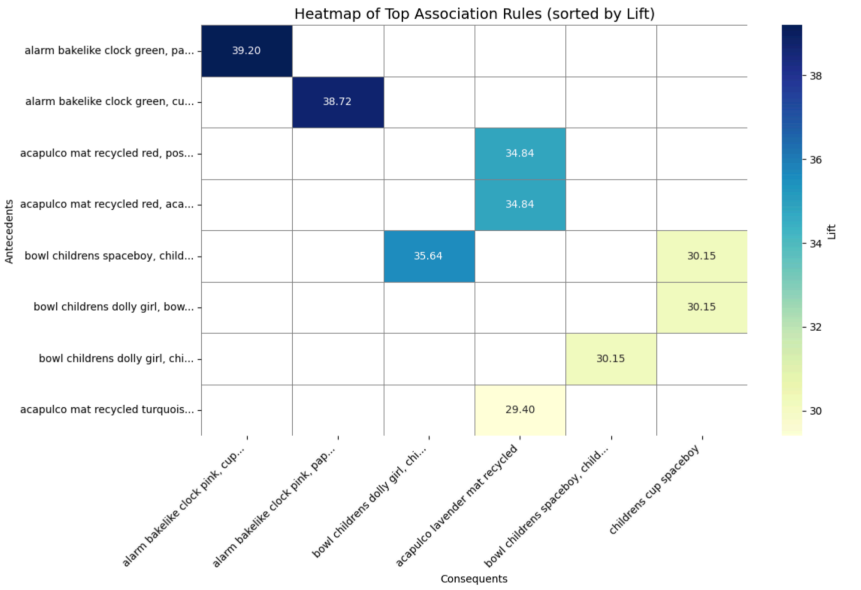
Heatmap for United Kingdom



Heatmap for Germany



Heatmap for France



Heatmap for Ireland

Key Insights from visualization and apropi rules:

United Kingdom. The UK dataset gave the largest number of association rules due to its higher transaction volume. Frequent combinations included home decor items such as candle holders, trays, and bunting. For example, customers who bought "white hanging heart t-light holder" often also purchased "heart of wicker small" and "red woolly hottie white heart." These items are stylistically similar, suggesting that UK consumers tend to buy coordinated home accessories.

France. In France, we identified associations involving colorful kitchen and garden items, such as "acapulco mat recycled red," "acapulco mat recycled turquoise," and "acapulco lavender mat recycled." These rules indicate a preference for vibrant, eco-friendly products that are often purchased together, perhaps as a set or matching decor.

Germany. German transactions showed a tendency toward functional and practical items. Combinations like "jam making set with jars" and "jam making set with labels" were common, suggesting a market for DIY or kitchen-related goods. This might reflect a cultural inclination towards crafting and home preparation.

Ireland. In Ireland, strong associations included festive and seasonal items such as "birthday bunting happy blue" and "birthday bunting happy pink." These patterns reveal a consumer interest in celebratory decorations, potentially aligning with frequent small-scale events or parties.

Although all four countries have similar interests in home goods and decor, their preferences differ in style and purpose:

- The UK market is broader and includes a wide range of decorative items.
- France shows a distinct preference for colorful and possibly eco-conscious items.
- Germany leans toward functional, DIY-related purchases.
- Ireland emphasizes festive and seasonal goods.

These findings highlight the importance of localizing product bundles and tailoring marketing strategies to specific regional tastes. For example, a promotional package with relevant home decor items may be more effective in the UK, while DIY kits may appeal more to German audiences.

5. Conclusions

Through the analysis, we were able to identify relevant patterns in customers' purchasing behavior that can be useful for marketing strategies, inventory optimization, and personalized recommendations.

Apriori's algorithm helped identify strong associations between frequently purchased items, highlighting sets of products that are often bought together. For example, products such as recycled rugs or sets of decorative bows often form strong rules with high lift values, suggesting localization and consistency of buying habits across countries. The comparative analysis allowed us to explore country-specific preferences and better understand regional consumer behavior.

Despite meaningful results, the project also revealed several limitations in the dataset. Our recommendations for improving the dataset quality are as follows:

1. Separate successful and cancelled transactions.

2. Avoid using negative values for price and quantity: even if an order was cancelled, it is better to store transaction status explicitly rather than using negative values, which can complicate preprocessing and mislead algorithms.

3. Implement data validation: basic validation mechanisms such as spell-checking, standard naming conventions, and input constraints could drastically reduce noise in the data and prevent typos or inconsistent formatting.

4. Avoid missing key information: enforcing completeness for essential fields like CustomerID and Description would significantly increase the reliability of the dataset and prevent unnecessary data loss during preprocessing.

5. Use consistent product naming: avoid duplicated entries for the same product caused by minor textual differences (e.g., upper/lowercase, typos, or extra whitespace). Implementing standardized naming or product IDs would solve this issue.

Implementing these recommendations will make future analyses more accurate, less time-consuming, and easier to scale. Overall, we are satisfied with the results obtained and knowledge gained during the project, especially in the areas of data preprocessing, application of algorithms and interpretation of consumer information.

P.s. Work distribution:

Sofiia Parkhomets (024712322C): Data quality and preparation, Data preprocessing, Conclusions & Recomendations

Angelina Naliuka (024712312B): Apriori Algorythm, Visualization, Conclusions & Recomendations