CMPN451 – Big Data

Team 20 Project Document

TEAM MEMBERS:

YOUSSEF ALAA MOSTAFA	1180092
MOHAMED KHALED SHAMS	1180552
MOSTAFA MOHAMED SABRY	1162211
MUSTAFA KHALED ABD AL-BARR	1180126

Table of Contents

Problem Description

An online business wants to understand its customer’s buying behavior and segment them to create targeted marketing campaigns. We plan to use market basket analysis and RFM analysis to achieve this.

The business has data on its customers' purchase history, including the items purchased, the date of purchase, and the amount spent.

The main objectives of this project are:

1. To identify items that are commonly purchased together and recommend cross-selling opportunities.
2. To segment customers based on their RFM scores and create targeted marketing campaigns to increase sales and customer loyalty and retention.

Dataset Description

The dataset used contains online retail transactions data from a UK-based and registered, non-store online retail company between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware and many customers are wholesalers.

The dataset has 8 columns:

- Invoice: Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.

- InvoiceDate: Invice date and time. Numeric. The day and time when a transaction was generated.
- UnitPrice: Unit price. Numeric. Product price per unit in sterling (Â£).
- Customer ID: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal. The name of the country where a customer resides.

The dataset has 1,067,371 rows and is 60 MB in size1. The dataset can be used for customer segmentation, market basket analysis, and other data mining tasks.

Pipeline

- Data Preprocessing
    - Identifying Irrelevant Columns
    - Data Cleaning
    - Handling of missing data
    - Dealing with outliers
- Market Basket Analysis
- RFM Analysis

Data Preprocessing

Identifying irrelevant columns:

Since we use all the columns for Market Basket Analysis and RFM analysis, there are no irrelevant columns.

Cleaning and handling missing data:

Firstly, the dataset is preprocessed by removing duplicate values to ensure accuracy in the subsequent analysis. The next step involves investigating the presence of missing values (NAs), which are found in both the "Description" and "Customer ID" columns. As the Customer ID is crucial for customer segmentation, imputation techniques are not feasible since they can introduce bias into the results.

To address this, we examined whether each invoice was unique for each customer ID. Subsequently, we checked if any invoice was related to multiple Customer IDs, including NaN. In this case, we attempted to substitute the NaN values with another Customer ID associated with the same invoice. However, this approach was not successful since all invoices were connected to only one value, including NaN.

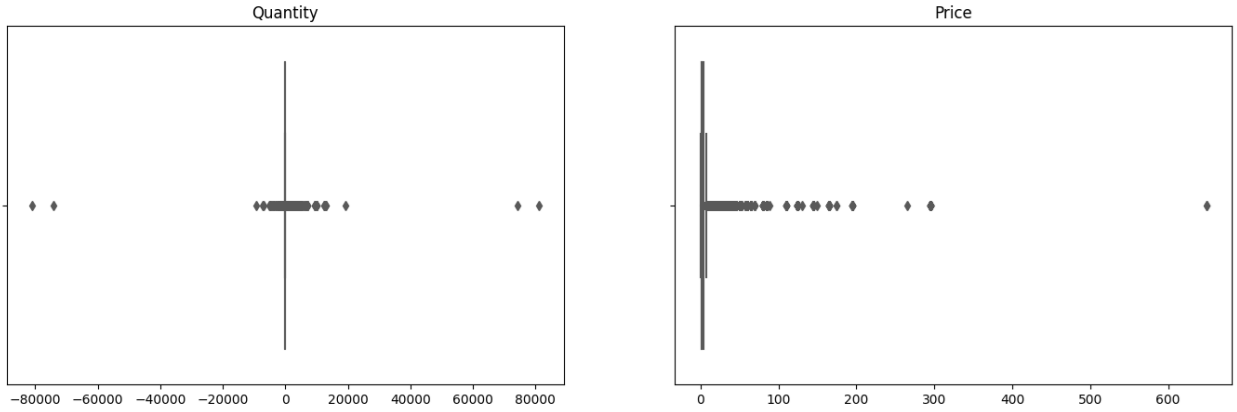
Therefore, we decided to remove the records with NaN Customer ID values, which also removed records with missing Description values and invalid prices. This data cleaning process ensures that the subsequent analysis is based on a clean and reliable dataset.

Additionally, after checking for some insights, we removed the Cancellation transactions as we are investigating the original products and transactions.

Finally, we conducted a thorough exploration of the dataset to identify any potential data quality issues or anomalies. During this process, we observed that some transactions were marked as "Cancellation" transactions, which indicated that the original transaction was canceled. Since our analysis aimed to investigate customer behavior and product preferences based on original transactions, we decided to remove these cancellation transactions from the dataset.

Dealing with outlier data

To identify outlier data in the dataset, we employed Tukey's method, which uses the outer fence bounds (3*IQR) to detect extreme values. Our analysis found that around 2.25% of the "Price" values and 4.76% of the "Quantity" values are outliers.



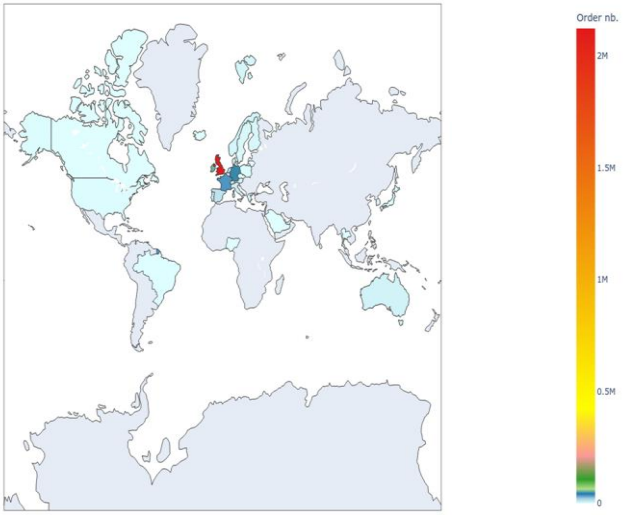
As it is a relatively large percentage of the dataset, we decided not to remove them since we are not utilizing any predictive or iterative clustering models that may be adversely affected by outliers. Instead, we will keep the outlier values in the dataset.

We didn’t need to handle Numerical and Categorial data and we didn’t need to handle the wide ranges of numerical variables as we don’t need to utilize supervised or unsupervised learning algorithms.

Data insights during preprocessing

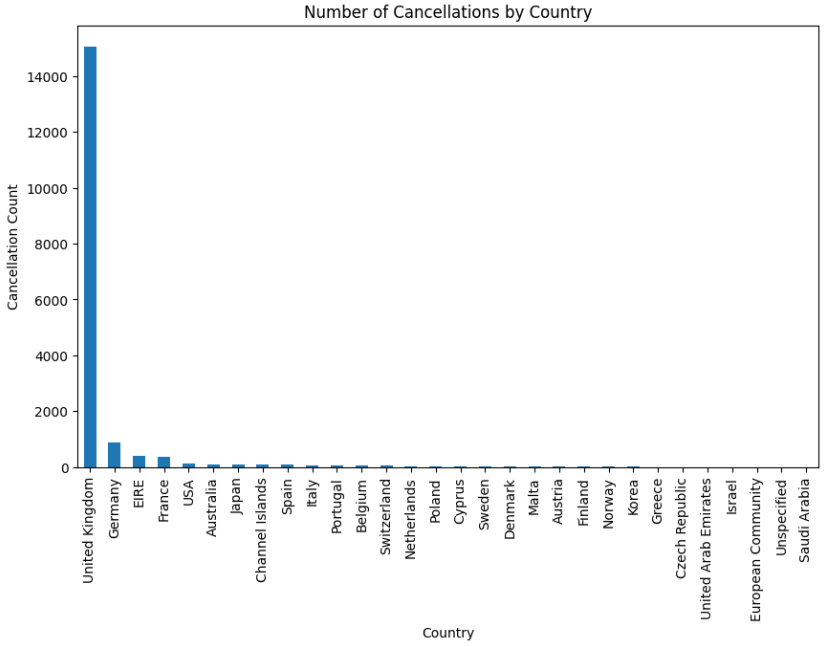
We found out the business gets the most orders from the United Kingdom, followed by Germany and France.

We found out that the business gets the most value per order also from the United Kingdom, followed by Ireland and Germany.

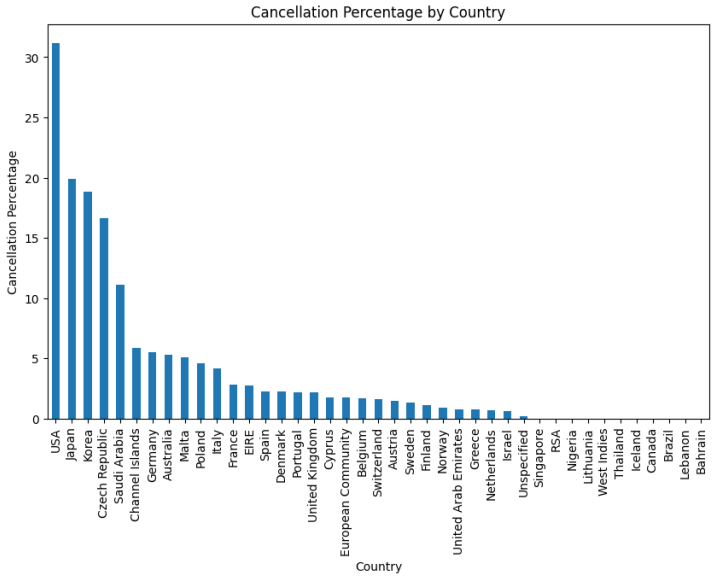


We found out that there are Cancelation transactions inside the dataset where the “Invoice” values that starts with ‘C’ are Cancelation invoices for existing transactions inside the dataset. These invoices are about 16.6% of the total number of invoices.

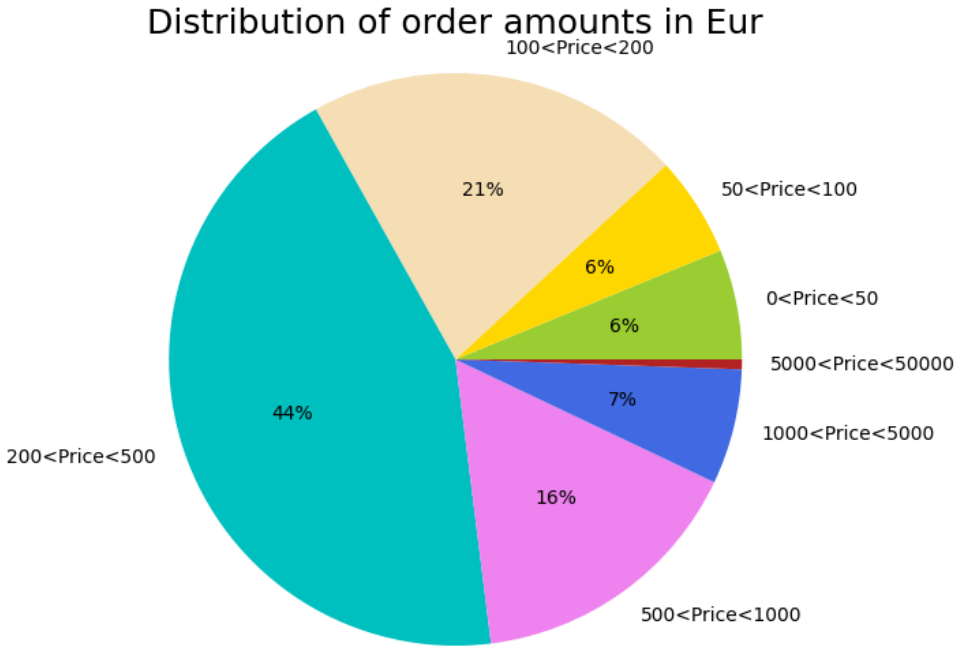
The United Kingdom is the country with the largest number of cancelation invoices exceeding 14000 cancelation invoices, followed by Germany with less than a 1000 cancelation transaction.



We measured the percentage of cancellation invoices relative to the total number of invoices for each country, where USA has the highest percentage exceeding 30% followed by Japan with 20% then Korea with slightly less than 20%.



Investigating the distribution of the purchase amount of orders showed us that a significant proportion of orders involve high-value purchases, as approximately 65% of purchases result in prices exceeding £200.



Market Basket Analysis

Market basket analysis is a data mining technique that identifies the relationships and associations between items frequently purchased together by customers. It is commonly used in retail and e-commerce to analyze transactional data and uncover patterns in customer behavior, such as cross-selling opportunities, product recommendations, and inventory management.

FP-growth (Frequent Pattern-growth) is a popular algorithm used in association rule mining, a subset of market basket analysis. FP-growth algorithm generates frequent itemsets (sets of items that frequently appear together in transactions) and association rules from transactional data. It utilizes a compressed representation of the database and employs a divide-and-conquer strategy to reduce the computational time required to generate frequent itemsets and association rules.

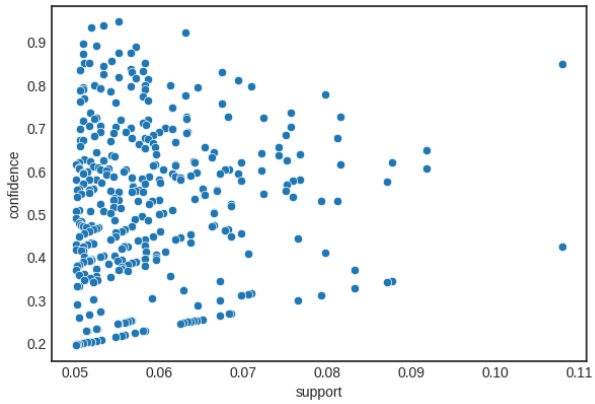
FP-growth algorithm is particularly useful when dealing with large and complex datasets, where traditional Apriori algorithm (another popular association rule mining algorithm) may suffer from scalability issues. Additionally, FP-growth algorithm can also generate association rules with higher confidence levels compared to Apriori algorithm, thereby improving the quality of the analysis results.

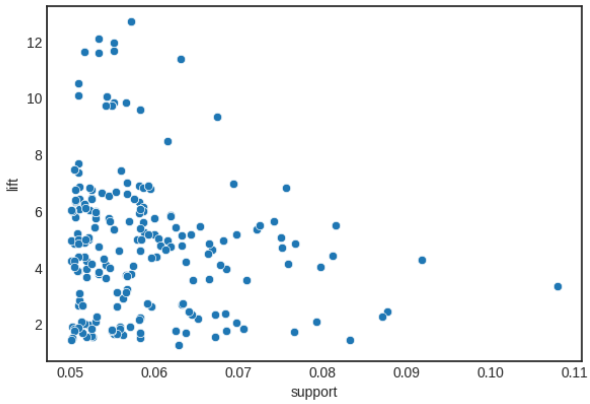
So we elected to use FP-growth to find the items frequently bought together for the whole dataset, then for every one of the top countries to check if they need special consideration, (like special discounts, …etc)

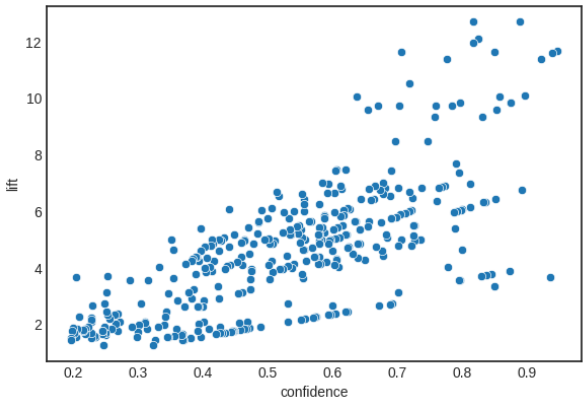
The following are the top association rules for the whole dataset:

	antecedent	consequent	confidence	lift	support
0	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[GREEN REGENCY TEACUP AND SAUCER]	0.947214	11.671672	0.055185
1	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[GREEN REGENCY TEACUP AND SAUCER]	0.939759	11.579810	0.053306
2	[CANDLEHOLDER PINK HANGING HEART]	[WHITE HANGING HEART T-LIGHT HOLDER]	0.935385	3.674367	0.051939
3	[PINK REGENCY TEACUP AND SAUCER]	[GREEN REGENCY TEACUP AND SAUCER]	0.922500	11.367142	0.063045
4	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[ROSES REGENCY TEACUP AND SAUCER]	0.897590	10.083678	0.050914
5	[LUNCH BAG WOODLAND, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.892442	6.748659	0.052452
6	[POPPY'S PLAYHOUSE BEDROOM]	[POPPY'S PLAYHOUSE KITCHEN]	0.888594	12.685224	0.057236
7	[GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...	[ROSES REGENCY TEACUP AND SAUCER]	0.875661	9.837324	0.056552
8	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[ROSES REGENCY TEACUP AND SAUCER]	0.875339	9.833700	0.055185
9	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.873900	3.892647	0.050914
10	[TOILET METAL SIGN]	[BATHROOM METAL SIGN]	0.857143	10.053822	0.054331
11	[PINK REGENCY TEACUP AND SAUCER]	[ROSES REGENCY TEACUP AND SAUCER]	0.852500	9.577126	0.058261
12	[LUNCH BAG RED RETROSPOT, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.851852	6.441717	0.051085
13	[BLUE HAPPY BIRTHDAY BUNTING]	[PINK HAPPY BIRTHDAY BUNTING]	0.850704	11.633579	0.051597
14	[RED HANGING HEART T-LIGHT HOLDER]	[WHITE HANGING HEART T-LIGHT HOLDER]	0.849462	3.336848	0.107979
15	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.845528	3.766269	0.053306
16	[GREEN REGENCY TEACUP AND SAUCER, ROSES REGENC...	[REGENCY CAKESTAND 3 TIER]	0.837975	3.732622	0.056552
17	[LUNCH BAG WOODLAND, LUNCH BAG BLACK SKULL]	[LUNCH BAG SPACEBOY DESIGN]	0.836158	6.323041	0.050572
18	[LUNCH BAG SUKI DESIGN, LUNCH BAG BLACK SKULL]	[LUNCH BAG SPACEBOY DESIGN]	0.833333	6.301680	0.058090
19	[GREEN REGENCY TEACUP AND SAUCER]	[ROSES REGENCY TEACUP AND SAUCER]	0.831579	9.342095	0.067487
20	[PINK REGENCY TEACUP AND SAUCER]	[REGENCY CAKESTAND 3 TIER]	0.830000	3.697100	0.056723
21	[GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...	[PINK REGENCY TEACUP AND SAUCER]	0.825397	12.077619	0.053306
22	[GREEN REGENCY TEACUP AND SAUCER, ROSES REGENC...	[PINK REGENCY TEACUP AND SAUCER]	0.817722	11.965310	0.055185

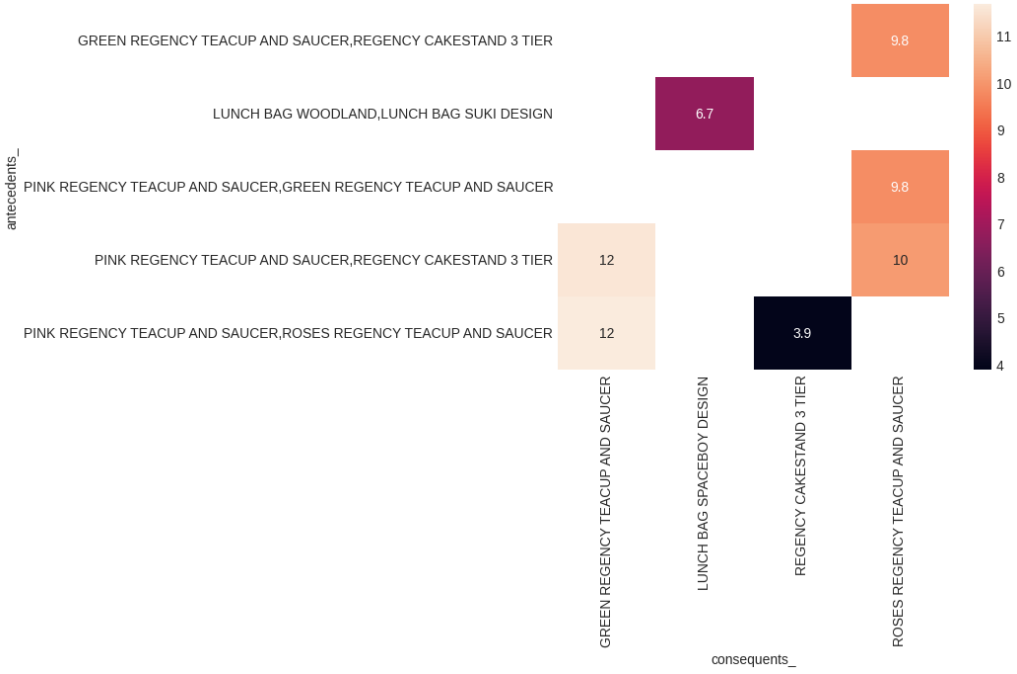
The following are scatter plots for all association rules:



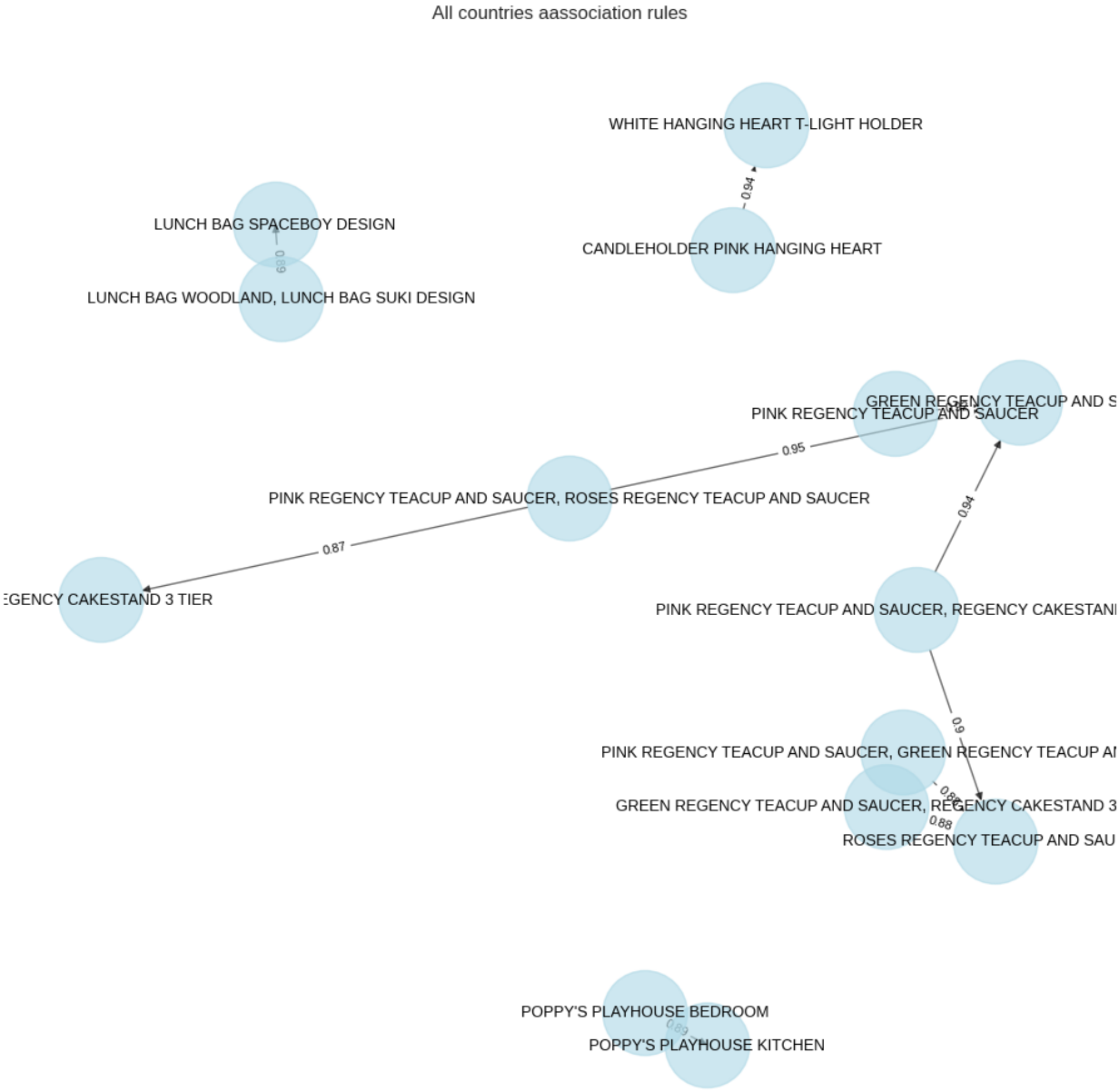




The following is a lift heatmap for the top 10 association rules:



The following is the network graph of the connection between the items in the top 10 association rules:



The following are the association rules for the top 2 countries:

Country: United Kingdom
Top 10 Association Rules (minimum number of items = 3):

	antecedent	consequent	confidence	lift	support
0	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[GREEN REGENCY TEACUP AND SAUCER]	0.947214	11.671672	0.055185
1	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[GREEN REGENCY TEACUP AND SAUCER]	0.939759	11.579810	0.053306
2	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[ROSES REGENCY TEACUP AND SAUCER]	0.897590	10.083678	0.050914
3	[LUNCH BAG WOODLAND, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.892442	6.748659	0.052452
4	[GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...	[ROSES REGENCY TEACUP AND SAUCER]	0.875661	9.837324	0.056552
5	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[ROSES REGENCY TEACUP AND SAUCER]	0.875339	9.833700	0.055185
6	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.873900	3.892647	0.050914
7	[LUNCH BAG RED RETROSPOT, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.851852	6.441717	0.051085
8	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.845528	3.766269	0.053306
9	[GREEN REGENCY TEACUP AND SAUCER, ROSES REGENC...	[REGENCY CAKESTAND 3 TIER]	0.837975	3.732622	0.056552

Country: United Kingdom
Top 10 Association Rules (maximum number of items = 2):

	antecedent	consequent	confidence	lift	support
0	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[GREEN REGENCY TEACUP AND SAUCER]	0.947214	11.671672	0.055185
1	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[GREEN REGENCY TEACUP AND SAUCER]	0.939759	11.579810	0.053306
2	[CANDLEHOLDER PINK HANGING HEART]	[WHITE HANGING HEART T-LIGHT HOLDER]	0.935385	3.674367	0.051939
3	[PINK REGENCY TEACUP AND SAUCER]	[GREEN REGENCY TEACUP AND SAUCER]	0.922500	11.367142	0.063045
4	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[ROSES REGENCY TEACUP AND SAUCER]	0.897590	10.083678	0.050914
5	[LUNCH BAG WOODLAND, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.892442	6.748659	0.052452
6	[POPPY'S PLAYHOUSE BEDROOM]	[POPPY'S PLAYHOUSE KITCHEN]	0.888594	12.685224	0.057236
7	[GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...	[ROSES REGENCY TEACUP AND SAUCER]	0.875661	9.837324	0.056552
8	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[ROSES REGENCY TEACUP AND SAUCER]	0.875339	9.833700	0.055185
9	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.873900	3.892647	0.050914

Country: Ireland
Top 10 Association Rules (minimum number of items = 3):

	antecedent	consequent	confidence	lift	support
0	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[GREEN REGENCY TEACUP AND SAUCER]	0.947214	11.671672	0.055185
1	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[GREEN REGENCY TEACUP AND SAUCER]	0.939759	11.579810	0.053306
2	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[ROSES REGENCY TEACUP AND SAUCER]	0.897590	10.083678	0.050914
3	[LUNCH BAG WOODLAND, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.892442	6.748659	0.052452
4	[GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...	[ROSES REGENCY TEACUP AND SAUCER]	0.875661	9.837324	0.056552
5	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[ROSES REGENCY TEACUP AND SAUCER]	0.875339	9.833700	0.055185
6	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.873900	3.892647	0.050914
7	[LUNCH BAG RED RETROSPOT, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.851852	6.441717	0.051085
8	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.845528	3.766269	0.053306
9	[GREEN REGENCY TEACUP AND SAUCER, ROSES REGENC...	[REGENCY CAKESTAND 3 TIER]	0.837975	3.732622	0.056552

Country: Ireland
Top 10 Association Rules (maximum number of items = 2):

	antecedent	consequent	confidence	lift	support
0	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[GREEN REGENCY TEACUP AND SAUCER]	0.947214	11.671672	0.055185
1	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[GREEN REGENCY TEACUP AND SAUCER]	0.939759	11.579810	0.053306
2	[CANDLEHOLDER PINK HANGING HEART]	[WHITE HANGING HEART T-LIGHT HOLDER]	0.935385	3.674367	0.051939
3	[PINK REGENCY TEACUP AND SAUCER]	[GREEN REGENCY TEACUP AND SAUCER]	0.922500	11.367142	0.063045
4	[PINK REGENCY TEACUP AND SAUCER, REGENCY CAKES...	[ROSES REGENCY TEACUP AND SAUCER]	0.897590	10.083678	0.050914
5	[LUNCH BAG WOODLAND, LUNCH BAG SUKI DESIGN]	[LUNCH BAG SPACEBOY DESIGN]	0.892442	6.748659	0.052452
6	[POPPY'S PLAYHOUSE BEDROOM]	[POPPY'S PLAYHOUSE KITCHEN]	0.888594	12.685224	0.057236
7	[GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...	[ROSES REGENCY TEACUP AND SAUCER]	0.875661	9.837324	0.056552
8	[PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	[ROSES REGENCY TEACUP AND SAUCER]	0.875339	9.833700	0.055185
9	[PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	[REGENCY CAKESTAND 3 TIER]	0.873900	3.892647	0.050914

RFM Analysis

RFM analysis is a customer segmentation technique that is commonly used in marketing and sales to segment customers based on their purchasing behavior. RFM stands for Recency, Frequency, and Monetary Value, which are three key metrics used to segment customers.

Recency refers to the amount of time that has passed since a customer's last purchase. Customers who have made a purchase recently are considered to be more valuable than those who have not made a purchase in a long time.

Frequency refers to the number of purchases that a customer has made within a given time period. Customers who make frequent purchases are considered to be more valuable than those who make fewer purchases.

Monetary Value refers to the amount of money that a customer has spent on purchases within a given time period. Customers who spend more

money are considered to be more valuable than those who spend less money.

RFM analysis is a powerful tool for customer segmentation because it allows businesses to identify their most valuable customers and create targeted marketing and sales strategies to retain those customers and encourage them to make additional purchases.

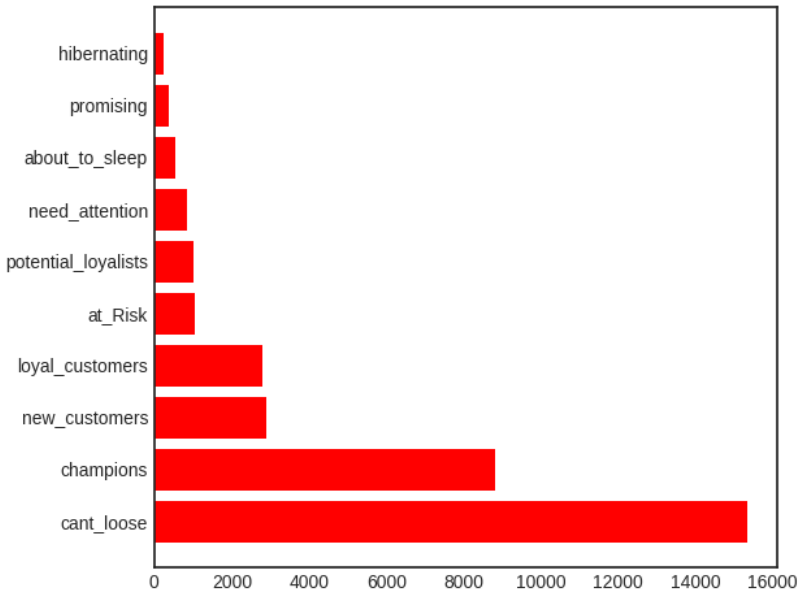
The clusters in RFM Analysis are the following:

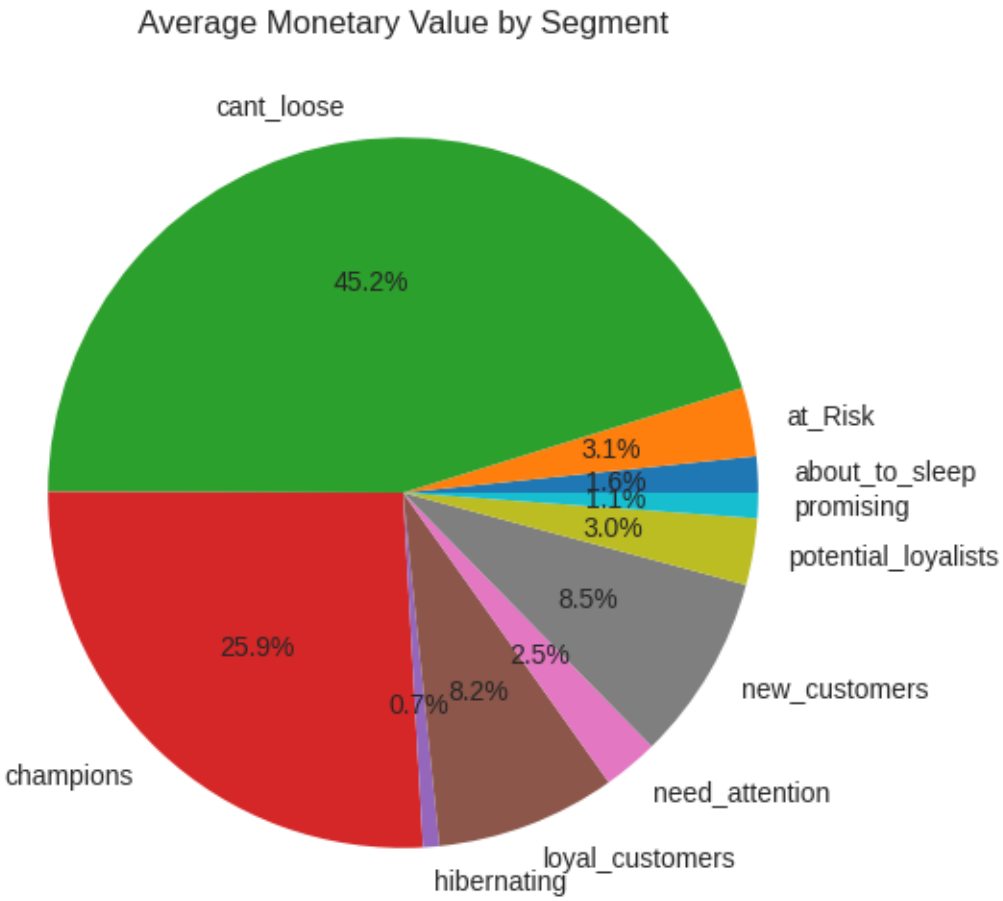
- **Champions**: These are your most valuable customers who have bought from you recently, frequently, and in large amounts. They are highly engaged with your business and are likely to be advocates for your brand.
- **Loyal Customers**: These customers are highly engaged and make frequent purchases, but their monetary value is lower than champions. They are likely to be repeat buyers and loyal to your brand.
- **Potential Loyalists**: These customers have made more recent purchases but may not have been consistent with their buying

frequency or monetary value. They have the potential to become loyal customers with a little encouragement.

- **Recent Customers**: These customers have made recent purchases but may not have returned frequently or spent a significant amount of money. They are still getting to know your brand and need to be engaged with relevant messaging and offers.
- **Promising**: These customers have made a high-value purchase in the past, but haven't made a purchase recently. They need to be re-engaged with your brand through targeted messaging and offers.
- **Needs Attention**: These customers haven't made a purchase in a while and have low purchase frequency and monetary value. They are at risk of being lost and need to be re-engaged with targeted messaging and offers.
- **About to Sleep**: These customers have low purchase frequency and monetary value but have made purchases more recently. They are at risk of becoming inactive and need to be re-engaged with targeted messaging and offers.
- **At Risk**: These customers have made purchases in the past but haven't been active recently. They are at risk of being lost and need to be re-engaged with targeted messaging and offers.
- **Cant Lose Them**: These customers used to be high-value customers but haven't made a purchase recently. They are still engaged with your brand and should be re-engaged with targeted messaging and offers to prevent churn.
- **Hibernating**: These customers have made no purchases in a while and have low purchase frequency and monetary value. They are at high risk of being lost and need to be re-engaged with targeted messaging and offers.

Here are the potential clusters:





Future Works

Classifying customers using the defined classes from the first purchase using SVM.