

DATA SCIENCE IN CUSTOMER PROFILING AND RETAIL ENHANCING E-COMMERCE EXCELLENCE - CUSTOMER SEGMENTATION AND INTELLIGENT PRODUCT RECOMMENDATION

1. Introduction

Data mining and customer-centric marketing are being quickly adopted by a large number of small online retailers and beginners to the online retail sector. They frequently lack the technical understanding and skill necessary to accomplish this, though. An online retailer's use of data mining tools for customer-centric business information is the subject of a case study presented in this project. The main objective of this study is to assist the business in better understanding its target audience so they can execute customer-centric marketing strategies with more efficiency. In this project I have effectively used the k-means clustering algorithm to divide its customer base into relevant groups by applying the Recency, Frequency, and Monetary model. This has made it possible to identify the main characteristics of each segment's customers with clarity.

Online retail sales have increased dramatically during the past 10 years, showing a significant shift in customer behavior regarding financial services and shopping. The significant increase in UK internet spending from \$50 billion in 2011 to 2000 provides proof of the transformation that e-commerce is changing the retail sector.

When compared to shopping in actual stores, online shopping offers numerous advantages. Online merchants may customize their approach to every customer with features like real-time tracking, customized experiences, and comprehensive customer profiles. By using data, they can remain ahead of the competition and increase profitability.

This method is based on important business factors that are focused around understanding customer behavior and preferences:

- Tracking how users interact with product websites and recording the frequency and sequence of visits.
- Identifying valuable and less valuable clients and the characteristics that set them apart.
- Analyzing consumer involvement and loyalty patterns.
- Identifying patterns in consumer behavior, such as orders and linked products.
- Putting consumer responses to advertising initiatives.
- Examining sales trends from a variety of angles, including products, places, and periods of time.

Online retailers commonly utilize data mining techniques to address these problems. Big businesses like Tesco, Amazon, and Walmart have integrated data mining into their operations to enable advanced customer-focused marketing strategies. Still, startups and smaller companies sometimes lack the technological expertise needed to fully utilize these strategies. The case study presented in this project shows how data mining may be used to provide customer-centric business insight for an online store. The objective is to improve the

marketing efforts of a small business that is relatively new to the online retail industry by better understanding their customer base. I may use the k-means clustering technique and decision tree induction to separate clients into distinct groups by applying the RFM model. I can use this to pinpoint particular consumer groups and their distinctive characteristics.

SAS Enterprise Guide and SAS Enterprise Miner are used to do the analysis in a systematic way. Pre-processing the data comes first, then segmentation and grouping. Every stage is covered in detail and offers insightful information. Actionable suggestions for customer-focused marketing and further data analysis are the final products.

2. Background

This project centers on a UK-based online retailer. This registered non-store business has been in operation since 1981 and involves about 80 people. At first, they mostly relied on phone orders and direct mail catalogs to market their distinctive gifts for every occasion. They did, however, take a big turn two years ago when they launched their own website and completely embraced the internet platform. Since then, they have been successful in drawing in large numbers of customers from all across the UK and Europe, collecting significant quantities of customer data in the process. They also use Amazon.co.uk as a platform for product marketing and sales.

[Table 1](#) lists the eight variables which make up the merchant's customer transaction dataset, which includes all transactions made between 2010 and 2011. For the business, CustomerID is an essential variable since it makes it possible to identify and track individual customers, which allows for a thorough analysis in the current study.

Variable name	Data type	Description; typical values and meanings
Invoice	Nominal	Invoice number; a 6-digit integral number uniquely assigned to each transaction
StockCode	Nominal	Product (item) code; a 5-digit integral number uniquely assigned to each distinct product
Description	Nominal	Product (item) name
Quantity	Numeric	The quantities of each product (item) per transaction
UnitPrice	Numeric	Product price per unit in sterling; \$45.23
InvoiceDate	Numeric	The day and time when each transaction was generated; 31/05/2011 15:59
CustomerID	Nominal	Delivery address postcode, mainly for consumers from the UK; SE1 0AA
Country	Nominal	Delivery address country; England

Table 1 Variables in the customer transaction dataset (4381 instances)

The company's first trial research aimed to collect useful consumer data by examining purchases made between January 1, 2011, and December 31, 2011. There were 22,190 legitimate transactions during this time period associated with 4,381 different postcodes. The

dataset contains 406,830 record rows as a result of these transactions, each of which reflects a distinct item that was bought. Every postcode had five transactions on average, meaning that consumers were making purchases from the online store around every two months. Moreover, only United Kingdom consumers were included in the investigation.

3. Methods

3.1 Data pre-processing

The initial steps involve data cleaning and preprocessing to ensure the quality of the dataset. This includes handling missing values, excluding irrelevant stock codes, and adjusting stock on hand:

- There are 135080 missing values and no null values in the dataset. But CustomerID is a key attribute in determining the RFM so I will omit the missing rows from our dataset.
- In order to perform analysis, I need to split the InvoiceDate into Day, Month, Year and Hour. Hence, I will first convert it to character and split the InvoiceDate records into weekOfDay, hourOfDay, month and year. This makes it possible to treat separate transactions done by the same customer on the same day but at different times.
- The dataset should be arranged according to postcode, and three important aggregated variables should be produced: monetary, frequency, and recency.

To implement the RFM analysis, I need to further process the data set in by the following steps:

- Find the most recent date for each ID, to get the Recency data
- Calculate the quantity of transactions of a customer till present date, to get the Frequency data
- Sum of Total Sales is the Monetary data.

I have successfully produced a target dataset for the study after completing these procedures. The dataset was originally in MS Excel format; however, I transformed it into the final SAS format.

3.2 RFM Calculation

Recency, frequency, and monetary value are the three main characteristics that RFM analysis takes into account when analyzing consumer behavior. Businesses are able to customize their marketing efforts by using such parameters to divide their customer base into separate groups.

I have calculated the RFM based on least number of purchases made through Customer ID (Postal code) and used mean and median to find out the number of purchases done by a customer ID and I have used them to calculate RFM and I have also calculated the first and third quarters from a year of purchases and finally the maximum number of purchases made by the customer ID. So, all of these parameters have been taken into consideration while

calculating the RFM which is shown in [Figure 1](#). Below are the results that have been received after the calculation.

CustomerID	Frequency	Monetary	Recency
Min. :12346	Min. : 1.00	Min. : 0.0	Min. : 0.0
1st Qu.:13813	1st Qu.: 17.00	1st Qu.: 0.0	1st Qu.:115.0
Median :15300	Median : 42.00	Median : 200.2	Median :253.0
Mean :15300	Mean : 93.05	Mean : 522.8	Mean :225.3
3rd Qu.:16778	3rd Qu.: 102.00	3rd Qu.: 604.8	3rd Qu.:331.0
Max. :18287	Max. :7983.00	Max. :26626.8	Max. :373.0

Figure 1 Calculated Recency, Frequency and Monetary

3.3Other Observations

3.3.1 Transactions By Year Analysis

By comparing the total number of transactions in year 2010 and 2011, I found out that people's consumption habits changed a lot. People became interested in shopping online. Below is the bar graph representation [Figure 2](#) to visualize the difference a lot better way.

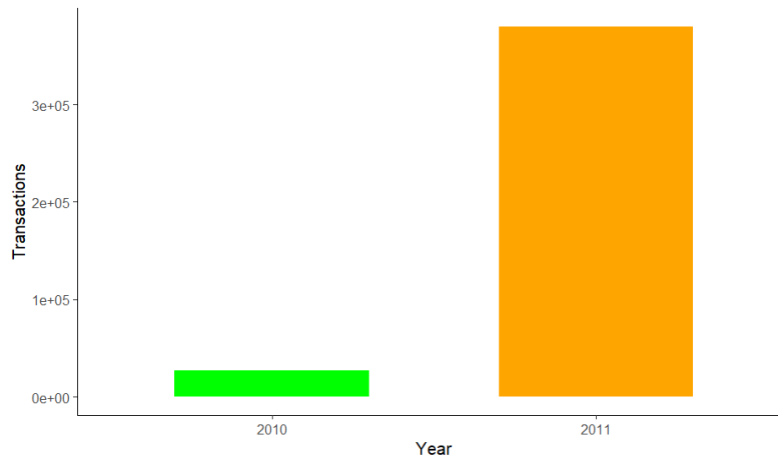


Figure 2 2010 vs 2011

3.3.2 Transactions By hour of the Day Analysis

The graph explains that between 10am till 3pm most of the orders are placed online. To find out the results in more visualizing way I have plotted bar graph [Figure 3](#) and these are the results I have got.

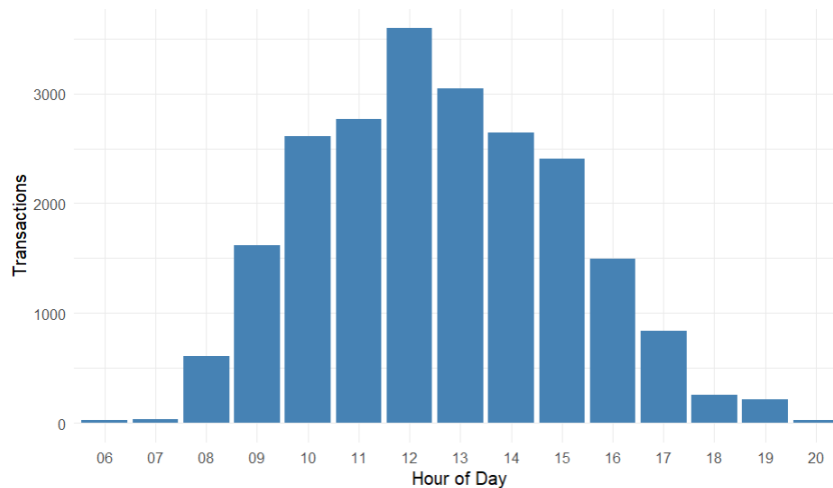


Figure 3 Transactions by Hour of Day

3.4 Clustering

3.4.1 K-Means Clustering

When I collected the target information, I had a clear objective in mind to find out if customers could be categorized according to their frequency, recency, and monetary values. I can use the k-means clustering algorithm, which is easily implemented with the help of the Cluster node, to do this.

I will use two most popular methods to find an optimal number of clusters which are Elbow Method and Silhouette Method. After using elbow method to find the number of clusters below is the output received

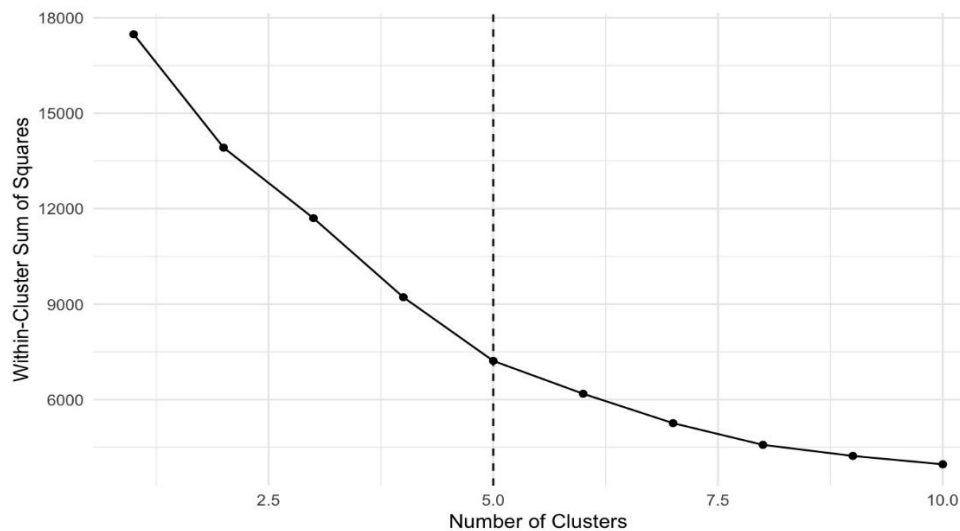


Figure 4 Elbow plot for K-means Clustering

The graph in [Figure 4](#) starts to bend at Cluster 5, hence I can determine that K=5 is the Optimal Cluster.

When I used Silhouette Method to calculate the number of clusters below are the number of clusters I have got as output.

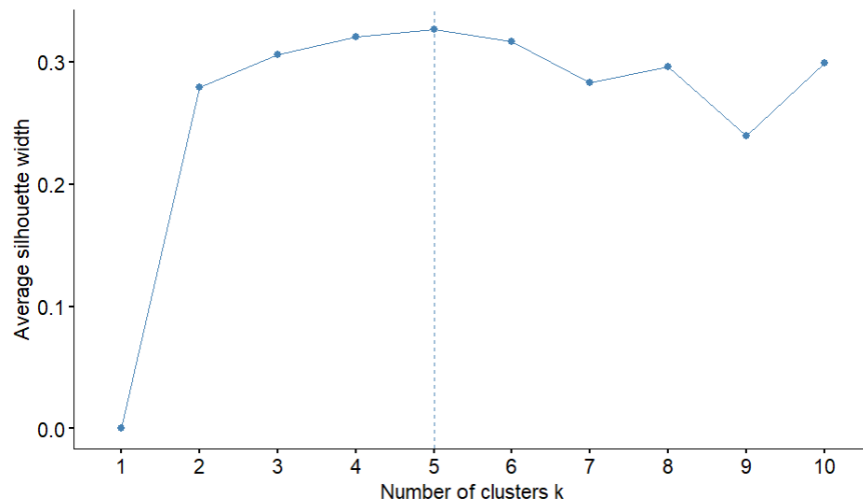


Figure 5 Optimal number of clusters

From [Figure 5](#) we can visualize that k=5 is the Optimal number of Cluster and k=8 is the next best. I will visualize k-means clusters using k=5 for better understanding.

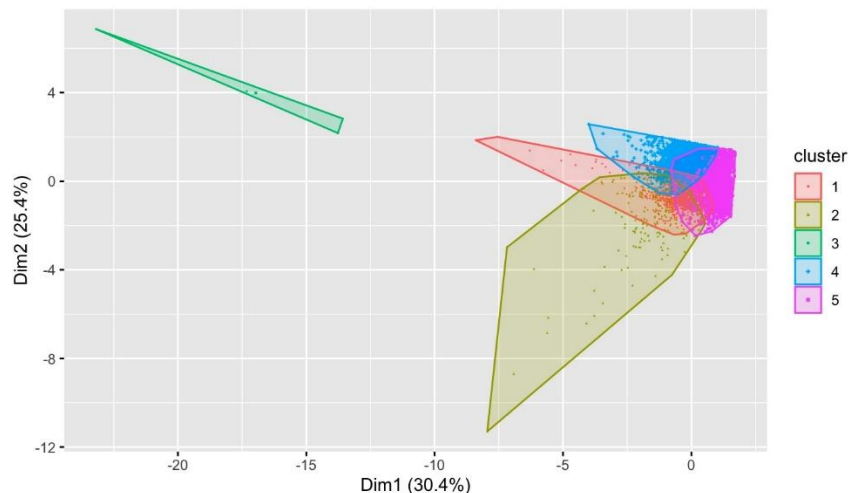


Figure 6 Optimal number of Cluster k=5

After observing the elbow graph results and Silhouette plot results, we have concluded to take $K = 5$ as our preferred clusters by observing [Figure 6](#)

3.4.2 Agglomerative hierarchical clustering

I have also Agglomerative hierarchical clustering as well to be more precise as this can help me in Customer Segmentation, Assortment Planning, Visual Merchandising, Market Basket Analysis, Store Layout Optimization, Customer Lifetime Value Prediction and more.

In data mining and statistics, a technique called agglomerative hierarchical clustering is used to group together comparable objects. Different linking criteria are used in agglomerative

clustering to determine the distance between clusters. I have used four linking criteria to determine which criteria is best for the retail business data, below are the results of the linkage criteria shown in [Figure 7](#):

```
average: 0.9932029
single: 0.9830277
complete: 0.9958161
ward: 0.9977612
```

Figure 7 Linkage criteria

The `agnes$ac` value gets the agglomerative coefficient, which measures the amount of clustering structure found (values closer to 1 suggest strong clustering structure). We see that all the four linkage methods are quite similar and close to 1, but Ward's method gives the best result. Also, in general, Complete and Ward's linkage are preferred over others.

To make sure I choose the right linkage criteria I have also plotted dendrograms between complete linkage and ward linkage.

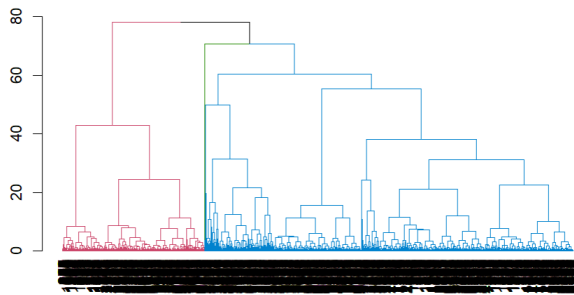


Figure 8 Ward linkage

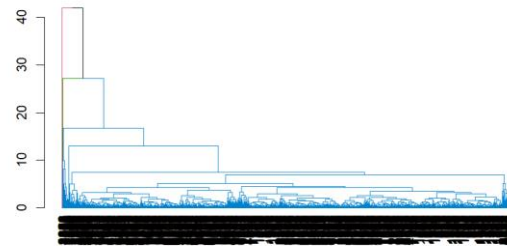


Figure 9 Complete linkage

We observe that the Complete linkage creates clusters for each outlier and thus creates 2 clusters each for 2 outliers as shown in [Figure 8](#) which would not provide good result. We will go with Ward's method as the way better results than complete linkage as shown in [Figure 8](#).

3.4.3 Dunn's Index test

After performing two different clustering methods now I am performing Dunns Index test to identify which is best suitable clustering method for the retail business data I have chosen.

```

####{r}
# Compute Dunn index for K-means clustering (k = 3)
dunn_km = dunn(clusters = k3$cluster, Data = RFM_scaled)
dunn_km
####

```

```
[1] 0.001585576
```

```

####{r}
# Perform hierarchical clustering using Ward's method and cut dendrogram into k = 3 clusters
memb_ward = cutree(hc_list$ward.D2, k = 3)
dunn_ward <- dunn(clusters = memb_ward, Data = RFM_scaled)
dunn_ward
####

```

```
[1] 0.003844301
```

Figure 10 Dunn index test results

We see that the Dunn's Index for hierarchical is higher than k-means clustering indicating hierarchical gives better Clustering results in [Figure 10](#).

3.5 Association Rule Mining

Association rules are mined to discover patterns and relationships between products. The Apriori algorithm is used to identify associations with high confidence and support, providing valuable insights into product co-occurrences.

```

as itemMatrix in sparse format with
23169 rows (elements/itemsets/transactions) and
8689 columns (items) and a density of 0.002037598

most frequent items:
WHITE HANGING HEART T-LIGHT HOLDER          REGENCY CAKESTAND 3 TIER
JUMBO BAG RED RETROSPOT                      1953                      1871
1721
(Party) PARTY BUNTING                        LUNCH BAG RED RETROSPOT
(Other)                                     1448                      1373
401834

```

Figure 11 Top 5 items bought frequently

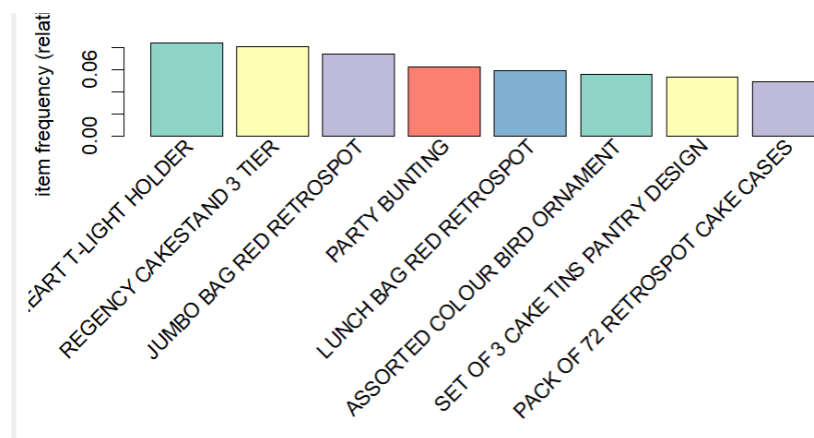
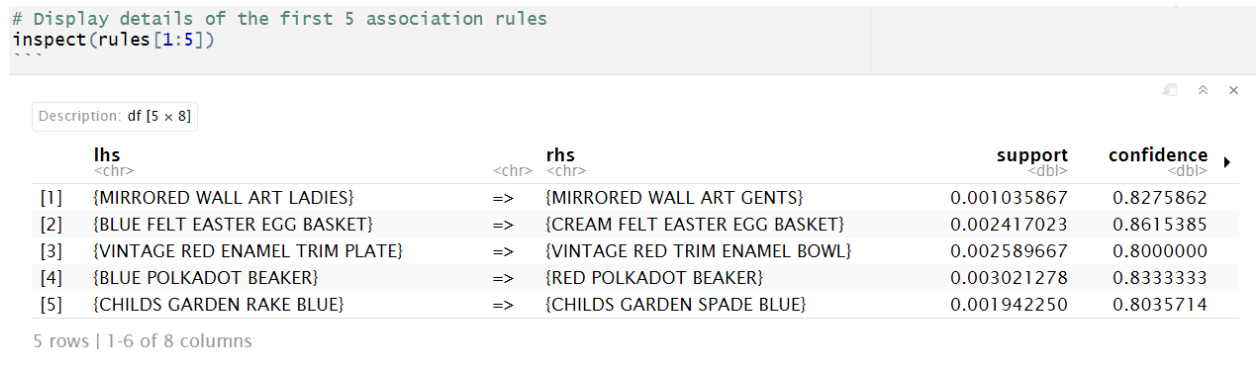


Figure 12 Most 8 Frequently purchased items

According to the plot, the items with the highest sales are 'WHITE HANGING HEART T-LIGHT HOLDER' and 'REGENCY CAKESTAND 3 TIER'. To boost the sales of 'SET OF 3 CAKE TINS PANTRY DESIGN', the retailer could place it near the 'REGENCY CAKESTAND 3 TIER' you can observe these results from both [Figure 11](#) and [Figure 12](#).

3.5.1 Apriori algorithm

One popular method used in association rule learning and data mining is the Apriori algorithm. Finding frequent item sets and significant relationships between items in large datasets is its primary goal. This method finds products that are commonly purchased together, which makes it very useful in retail business analysis.

```
# Display details of the first 5 association rules
inspect(rules[1:5])
```


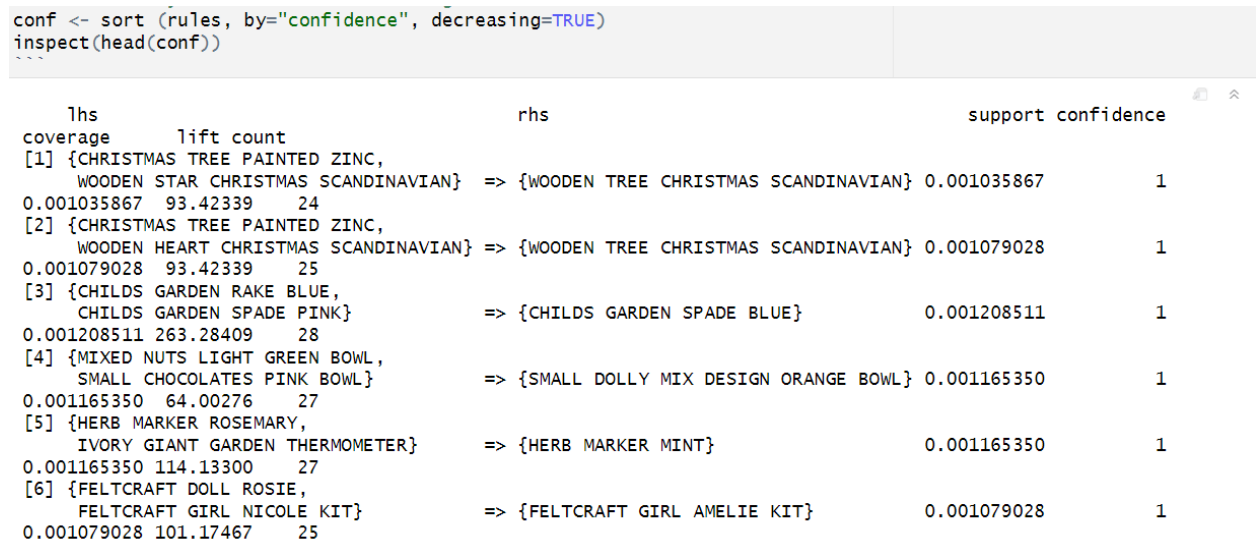
	lhs <chr>	rhs <chr>	support <dbl>	confidence <dbl>
[1]	{MIRRORED WALL ART LADIES}	=> {MIRRORED WALL ART GENTS}	0.001035867	0.8275862
[2]	{BLUE FELT EASTER EGG BASKET}	=> {CREAM FELT EASTER EGG BASKET}	0.002417023	0.8615385
[3]	{VINTAGE RED ENAMEL TRIM PLATE}	=> {VINTAGE RED TRIM ENAMEL BOWL}	0.002589667	0.8000000
[4]	{BLUE POLKADOT BEAKER}	=> {RED POLKADOT BEAKER}	0.003021278	0.8333333
[5]	{CHILDS GARDEN RAKE BLUE}	=> {CHILDS GARDEN SPADE BLUE}	0.001942250	0.8035714

5 rows | 1-6 of 8 columns


```

Figure 13 First 5 association rules

According Apriori, if a customer buys BLUE FELT EASTER EGG BASKET, there is a high probability of 86% that they will also buy CREAM FELT EASTER EGG BASKET as per [Figure 13](#).

```
conf <- sort (rules, by="confidence", decreasing=TRUE)
inspect(head(conf))
```


|     | lhs<br>coverage lift count                                         | rhs                                     | support     | confidence |
|-----|--------------------------------------------------------------------|-----------------------------------------|-------------|------------|
| [1] | {CHRISTMAS TREE PAINTED ZINC, WOODEN STAR CHRISTMAS SCANDINAVIAN}  | => {WOODEN TREE CHRISTMAS SCANDINAVIAN} | 0.001035867 | 1          |
| [2] | {CHRISTMAS TREE PAINTED ZINC, WOODEN HEART CHRISTMAS SCANDINAVIAN} | => {WOODEN TREE CHRISTMAS SCANDINAVIAN} | 0.001079028 | 1          |
| [3] | {CHILDS GARDEN RAKE BLUE, CHILDS GARDEN SPADE PINK}                | => {CHILDS GARDEN SPADE BLUE}           | 0.001208511 | 1          |
| [4] | {MIXED NUTS LIGHT GREEN BOWL, SMALL CHOCOLATES PINK BOWL}          | => {SMALL DOLLY MIX DESIGN ORANGE BOWL} | 0.001165350 | 1          |
| [5] | {HERB MARKER ROSEMARY, IVORY GIANT GARDEN THERMOMETER}             | => {HERB MARKER MINT}                   | 0.001165350 | 1          |
| [6] | {FELTCRAFT DOLL ROSIE, FELTCRAFT GIRL NICOLE KIT}                  | => {FELTCRAFT GIRL AMELIE KIT}          | 0.001079028 | 1          |


```

Figure 14 Sorted by confidence level

When the confidence is 1, it means that whenever the items on the left-hand side (LHS) are purchased, the items on the right-hand side (RHS) are also purchased 100% of the time.

According to the information provided in the output, we can perform the following analysis: All customers who purchased 'CHRISTMAS TREE PAINTED ZINC' and 'WOODEN STAR CHRISTMAS SCANDINAVIAN' also purchased 'WOODEN TREE CHRISTMAS SCANDINAVIAN'. The lift value in the rule 1 is significantly high, indicating that the occurrence of the initial three items has a substantial influence on the confidence value in [Figure 14](#).

What were the other items purchased by customers who bought the green Regency tea plate?

inspect(rules_tea)

Description: df [1 x 8]

	lhs <chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>
[1]	{REGENCY TEA PLATE GREEN}	=> {REGENCY TEA PLATE ROSES}	0.0108766	0.8289474	0.01312098

1 row | 1-7 of 8 columns

Figure 15 Sorted with the word 'tea'

82.9% of the time, customers who purchased REGENCY TEA PLATE GREEN also bought REGENCY TEA PLATE ROSES as per [Figure 15](#).

3.6 Conclusion

In conclusion, this project combines customer segmentation and association rule mining to provide e-commerce platforms with actionable insights. By understanding customer behavior and product associations, retailers can enhance the overall shopping experience, increase sales, and build lasting customer relationships.

Word Count: 2005

4. References

[1] UCI Machine Learning Repository - Online Retail Data Set,
<https://archive.ics.uci.edu/ml/datasets/online+retail>