REPORT

Supervised by:- Prof. Amit Mitra

TITLE

Retail Revelations: Transforming Online Sales with customer segmentation and Basket Analysis

Submitted by:-

Anoop Patel (231080017)

Chandan Kumar Singh (231080030)

Ishan Sharma (231080045)

Kritika Jamini (231080052)

Laxmi Agarwal (231080053)

Content:-

Abstract

1. Introduction
   - Overview of Customer Segmentation and Market Basket Analysis
   - Objectives of the Analysis
2. Source of Data
3. Aims and Objectives
   - Key Goals of the Analysis
4. Exploratory Data Analysis (EDA)
   - Analysis of Top Customers and Products by Sales Amount
   - Sales Distribution Patterns
5. Customer Segmentation
   - RFM Model for Customer Value Assessment
     - Recency, Frequency, and Monetary Analysis
   - K-Means Clustering for Customer Groups
     - Determining Optimal Clusters
     - Silhouette Analysis
8. Market Basket Analysis
   - Association Rule Mining
     - Support, Confidence, and Lift Definitions
   - Apriori, Eclat, and FP-Growth Algorithms
   - Dataset Transformation for Transactional Analysis
   - Dataset Pruning for Frequently Purchased Items

- Identification of Key Items Contributing to Sales
9. Association Rule Mining: Support and Confidence Thresholds
- Setting Minimum Support and Confidence
- Results of Rule Generation
10. Exploration of Generated Rules
- Top Rules by Confidence
- Top Rules by Support
- Lift Analysis
11. Conclusion
- Summary of Findings
- Key Results and Strategic Implications
12. References
- https://archive.ics.uci.edu/dataset/352/online+retail

Abstract

This report investigates online retail transactions to uncover patterns in customer behavior through customer segmentation and market basket analysis. By employing clustering techniques and association rule mining, we identified frequent item sets and meaningful co-purchasing trends that reveal customer preferences and shopping patterns. The analysis highlights a few high-value customers and top-performing products that significantly drive sales. Through RFM analysis and K-means clustering, we developed customer segments that support targeted marketing strategies, while association rule mining identified potential cross-selling opportunities. These insights are valuable for strategic decision-making in inventory management, personalized marketing, and cross-selling.

SOURCE OF DATA

The Online Retail dataset captures transactions from a UK-based, non-store online retailer, spanning from December 1, 2010, to December 9, 2011. This retailer specializes in unique gifts for various occasions, with a significant portion of its clientele comprising wholesale buyers.

INTRODUCTION

Customer segmentation and market basket analysis are two key techniques in data analysis and marketing, particularly valuable for understanding customer behavior and personalizing marketing strategies. Here's an overview of each:

Customer segmentation is the process of dividing a customer base into distinct groups, or segments, based on shared characteristics. These segments allow businesses to target different groups with tailored marketing, products, and services.

Techniques used:

1.Clustering Algorithms (K-means, hierarchical clustering): Used to group customers based on similarity in features.

The goal of K-means is to partition the dataset into K clusters so that each data point belongs to the cluster with the nearest mean, which serves as the cluster's center or "centroid."

2.RFM Analysis (Recency, Frequency, Monetary): Analyzes the time since the last purchase, how often purchases are made, and how much is spent to identify loyal, new, or at-risk customers.

➢ Recency (R)

Recency refers to how recently a customer made a purchase. The idea is that customers who have bought something more recently are more likely to engage with the brand again.

➢ Frequency (F)

Frequency indicates how often a customer makes purchases over a specific period. Customers who buy more frequently are generally considered more loyal and valuable.

➢ Monetary (M)

Monetary refers to the total amount a customer has spent over a specific period. High-spending customers are often more valuable and worth retaining.

Market Basket Analysis, often associated with association rule learning, is a technique used to understand products frequently bought together. It reveals relationships between items, which is valuable in cross-selling, upselling, and shelf arrangement in retail.

Key Concepts:

- **Association Rules**: Rules that describe the likelihood of certain items being purchased together, represented as "If X, then Y."
    - **Support**: The frequency with which an itemset appears in the dataset.
    - **Confidence**: The probability that a customer who bought one item will buy another.
    - **Lift**: The likelihood of co-occurrence of items beyond random chance.

Techniques:

Apriori Algorithm: One of the most common algorithms, which identifies frequent itemsets and derives association rules.

AIMS AND OBJECTIVES:

Exploratory Data Analysis is done to get the following:

- Percentage of sales amount for the the Top 10 customers.
- Percentage of sales amount for top 10 products.
- Percentage of sales amount for top 50 products.

Since our dataset is limited to the sales records, and didn´t include anothers information about our customers, we will use a **RFM**,***Recency, Frequency and Monetary Value**, based model of customer value for finding our customer segments. The RFM model will take the transactions of a customer and calculate three important informational attributes about each customer.

Exploratory Data Analysis

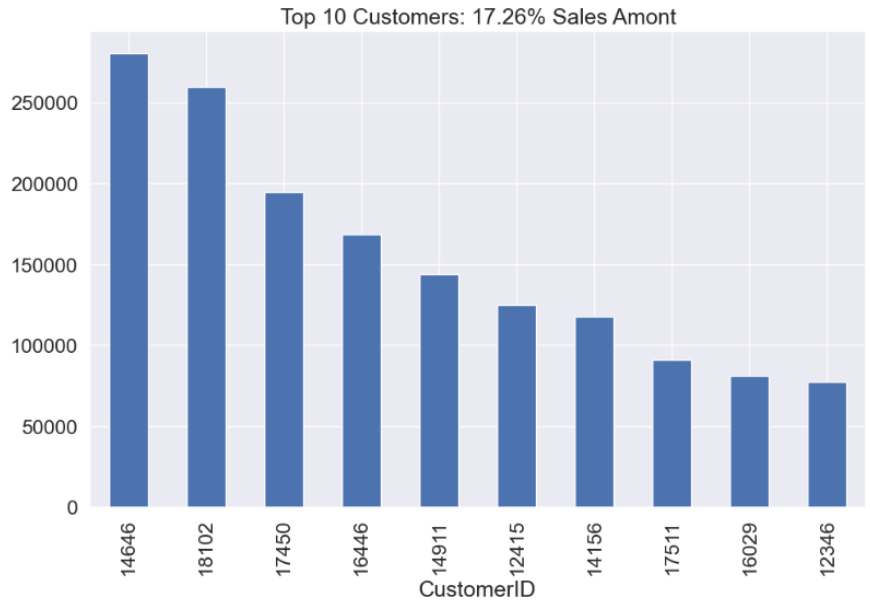
In the Exploratory Data Analysis (EDA) phase, the dataset was meticulously examined to uncover essential patterns, relationships, and data distributions that inform downstream modeling decisions. Initial assessments included descriptive statistics to capture fundamental metrics like mean, median, standard deviation, and distribution skewness. Visualizations, including histograms, box plots, and scatter matrices, were utilized to identify potential outliers, assess normality, and observe relationships among key variables.

Percentage of sales amount for the the Top 10 customers.

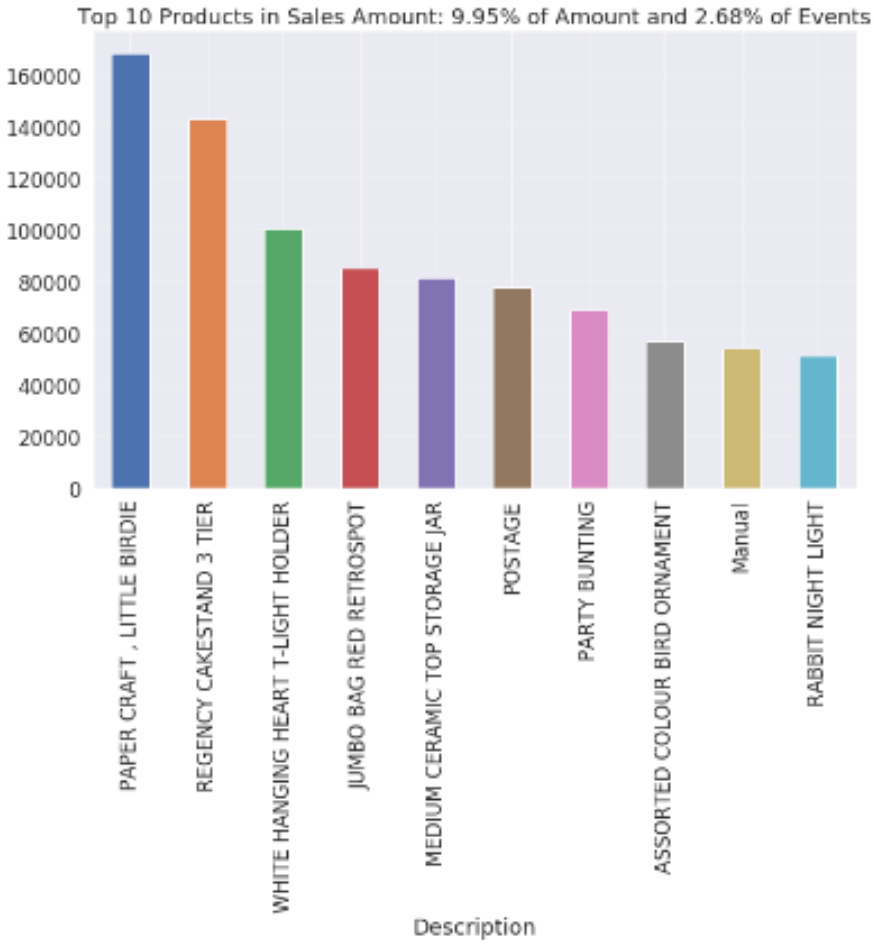
This bar chart presents the top 10 customers by sales amount, representing 17.26% of the total sales.

i. The top 10 customers contribute to a significant portion (17.26%) of total sales, indicating a degree of reliance on a small group of high-value customers. This suggests that retaining these customers is essential for maintaining sales levels.
ii. Customers with IDs 14646, 18102, and 17450 stand out, with each contributing over 150,000 in sales.
iii. Given that these top customers contribute a substantial share of sales, there may be opportunities to increase sales from lower-contributing customers outside this top group. Implementing loyalty programs, personalized

marketing, or upselling strategies could help expand the revenue base beyond these top 10 customers.



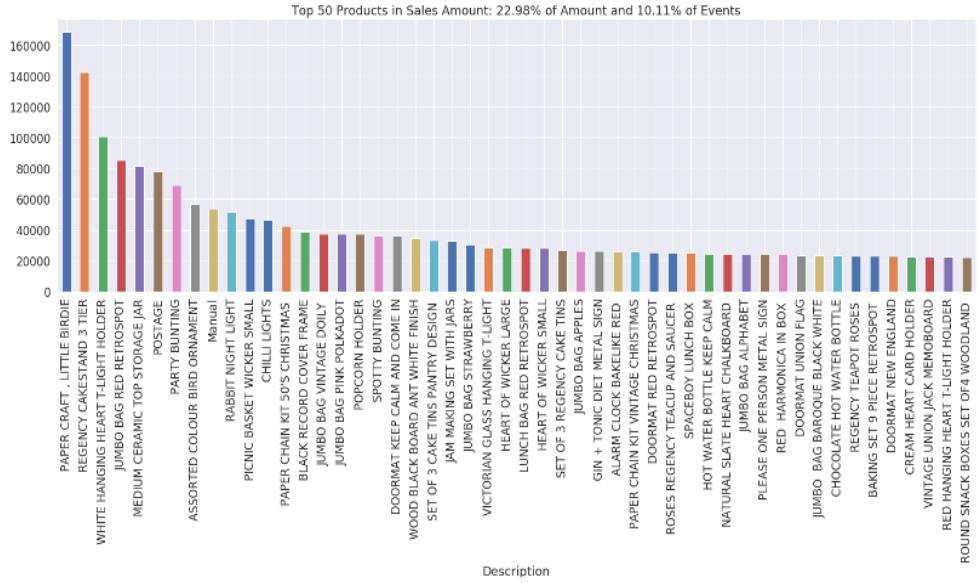
Percentage of sales amount for top 10 products.



i. The product "PAPER CRAFT, LITTLE BIRDIE" significantly outperforms the others in terms of sales amount. It's responsible for a substantial portion of the total sales, as indicated by the 9.95% contribution to the total amount.

ii. The top few products, particularly the top two, contribute a significant portion of the total sales. This suggests that a small number of products drive a large portion of the revenue.

iii. The chart includes a mix of products from various categories, including home decor, storage, and party supplies. This suggests a diverse product range.



Percentage of sales amount for top 50 products.

i. The top few products, particularly the top two, contribute a significant portion of the total sales amount (22.98%) and events (10.11%). This indicates that a small number of products drive a large portion of the revenue and customer activity.

ii. The chart showcases a diverse range of products, including home decor, kitchenware, party supplies, and more. This suggests a broad customer base with varied interests.

iii. The sales distribution follows a long-tail pattern, with a few top-performing products and a large number of products with lower sales.

Customer Segmentation

Customer segmentation is the practice of dividing an organization's customer base into various segments based on specific customer attributes. This approach is based on identifying differences in customer behavior and patterns.

Key Objectives and Benefits of Customer Segmentation:

Increased Revenue: A primary goal of customer segmentation is to enhance revenue.

Customer Insight: Understanding customers is essential in business, often called the principle of "knowing your customer." Segmentation allows for a detailed analysis of customer profiles.

Targeted Marketing: A clear benefit of customer segmentation is the ability to effectively direct marketing efforts. By identifying different customer segments, a company can create targeted campaigns suited to each group, which boosts the likelihood of campaign success.

Product Placement Optimization: Effective customer segmentation can guide the development of new products or bundle offerings that meet the needs of specific segments.

Clustering:

One common technique for customer segmentation is clustering, an unsupervised machine-learning method. This involves gathering as much relevant data about customers as possible, grouping them based on common characteristics, and analyzing each group to understand their distinct traits.

Exploratory Data Analysis:

Exploratory data analysis (EDA) is another useful method for identifying customer segments. Typically performed by analysts with in-depth knowledge of the product and customer domain, EDA is flexible and allows for the prioritization of key decision points.

RFM Model for Customer Value:

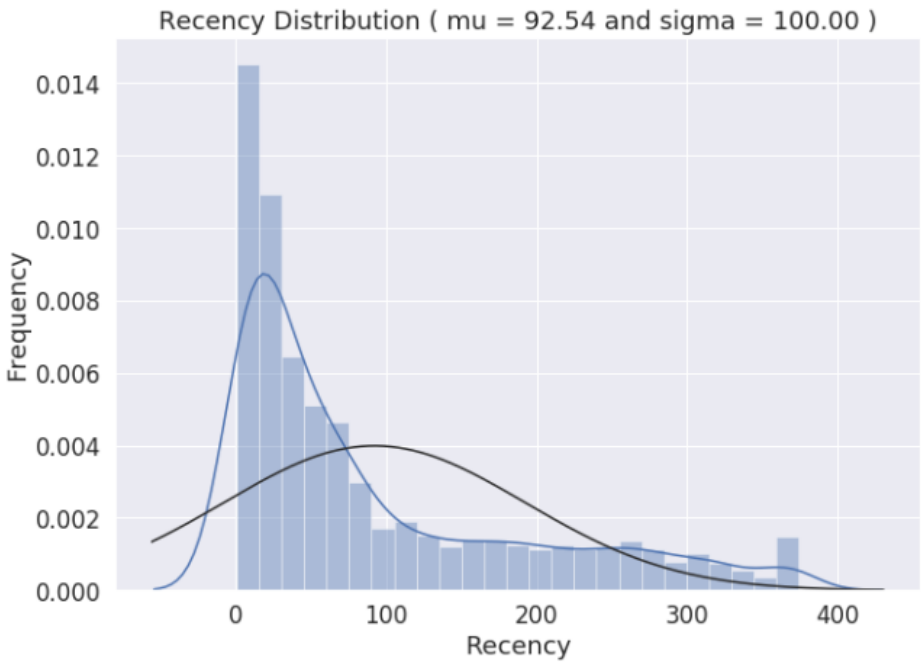
Given the limitations of our dataset to only sales records, we employ the RFM model—Recency, Frequency, and Monetary value—to evaluate customer value and identify customer segments. This model assesses customers based on:

- **Recency**: The number of days since a customer's last purchase.
- **Frequency**: The frequency of a customer's transactions.
- **Monetary Value**: The total monetary value of a customer's transactions.

Recency Analysis:

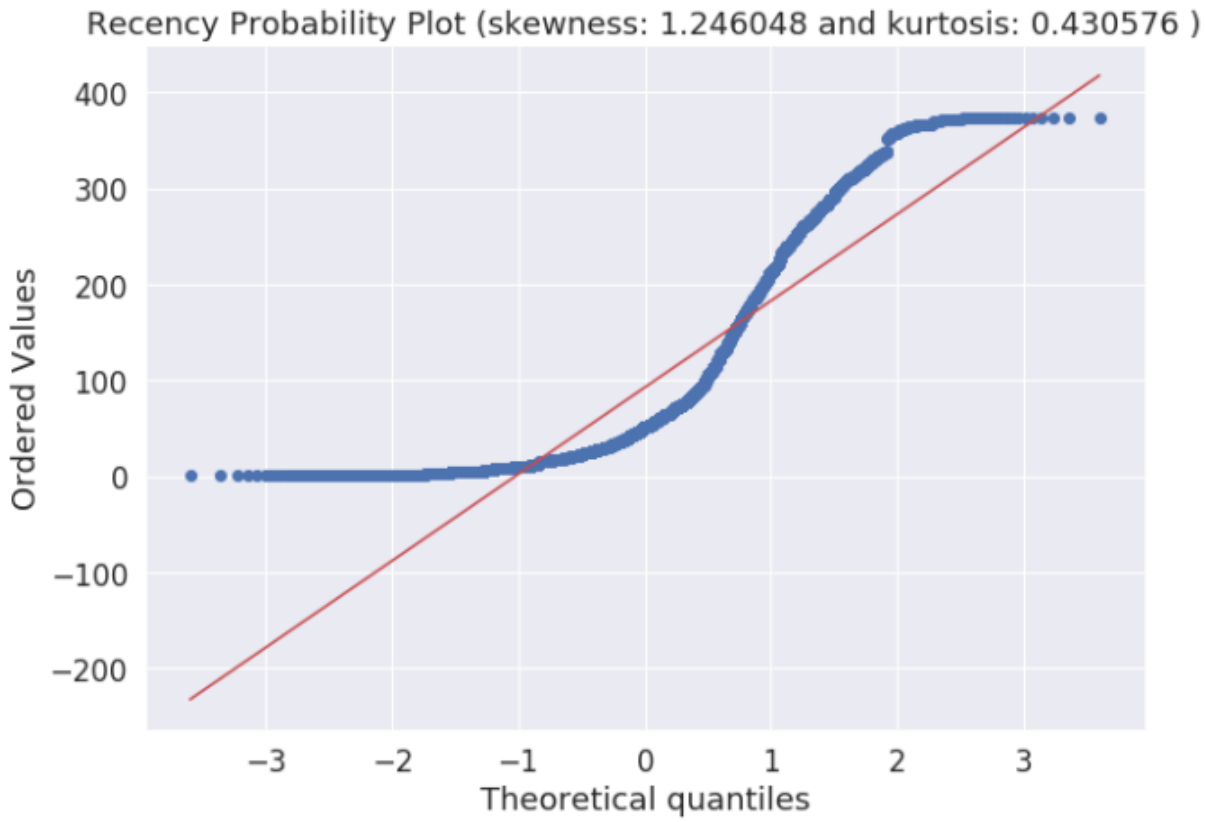
To generate the recency feature, a reference date is set. We then calculate the number of days before this reference date (December 10, 2011, 12:50:00) that a customer last made a purchase.

	count	mean	std	min	25%	50%	75%	max
CustomerID	4338.0	15300.408022	1721.808492	12346.0	13813.25	15299.5	16778.75	18287.0
recency	4338.0	92.536422	100.014169	1.0	18.00	51.0	142.00	374.0

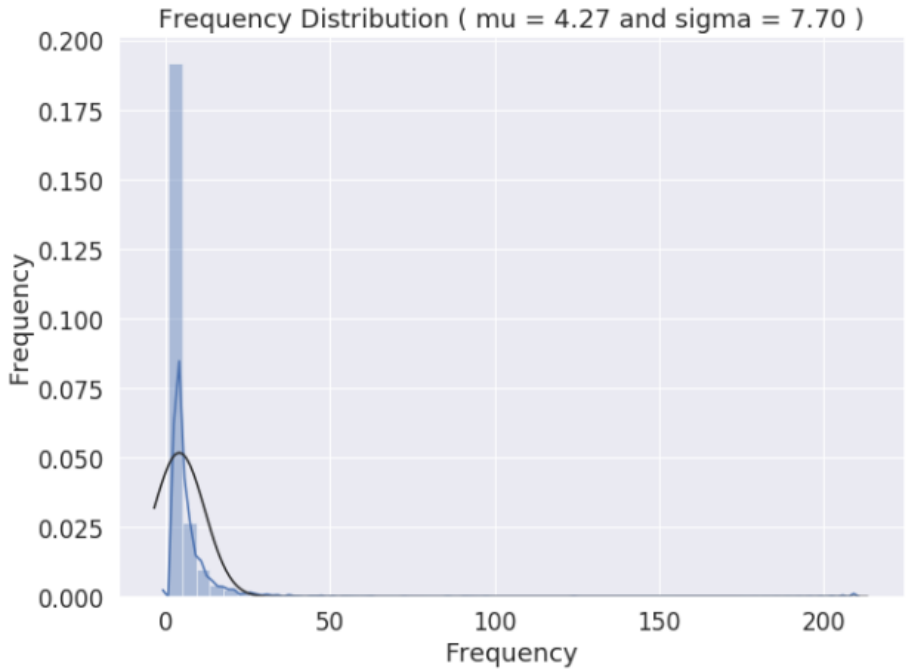


Distribution Plot: The recency distribution (number of days since the last purchase) shows a peak near the lower end, with a long tail extending to the right.

This right-skewed pattern suggests that most customers have made recent purchases, but a small number have not returned in a long time.

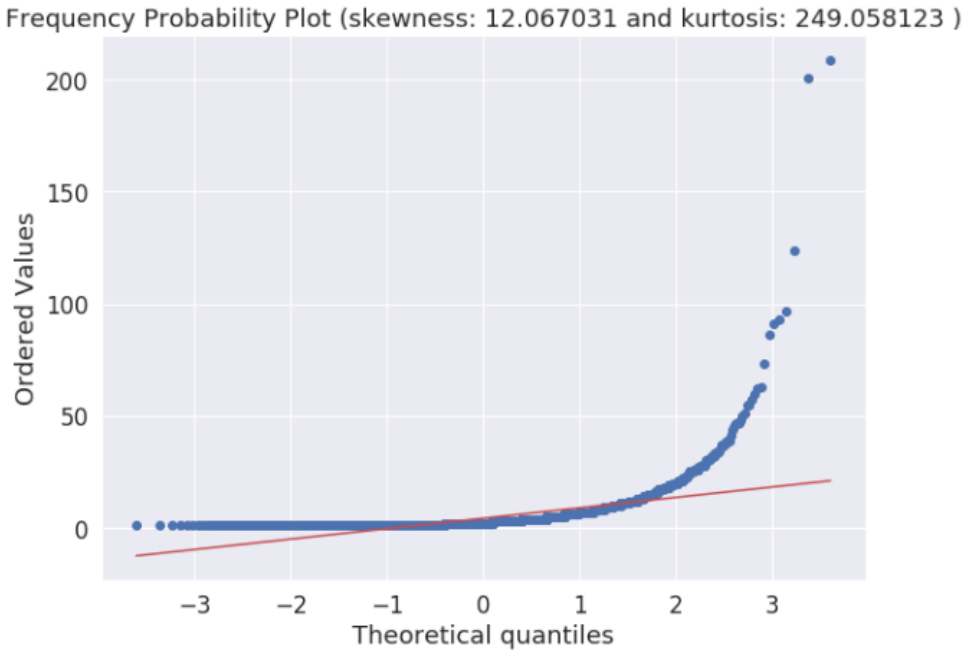


Probability Plot: The QQ plot for recency indicates that the data does not follow a normal distribution, as it deviates from the diagonal red line. The positive skewness value (1.25) confirms a lack of symmetry, with a right tail longer than the left. The positive kurtosis (0.43) suggests a heavy-tailed distribution with some outliers, indicating that a few customers have unusually high recency values.



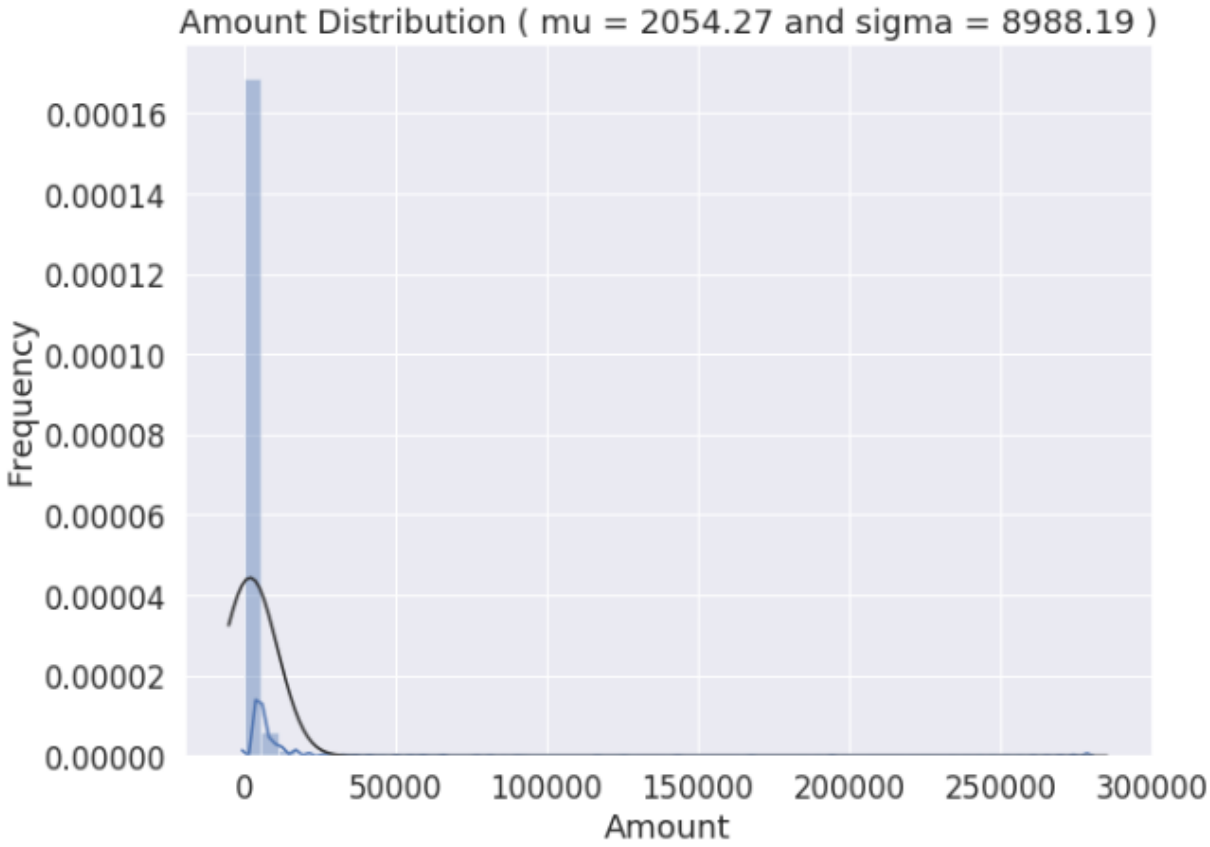
Frequency Analysis:

Distribution Plot: The frequency distribution plot shows that most customers have a low transaction frequency, with a peak close to zero and a steep drop-off, followed by a right tail. This positive skew indicates that only a few customers make purchases frequently, while most have a low transaction frequency..

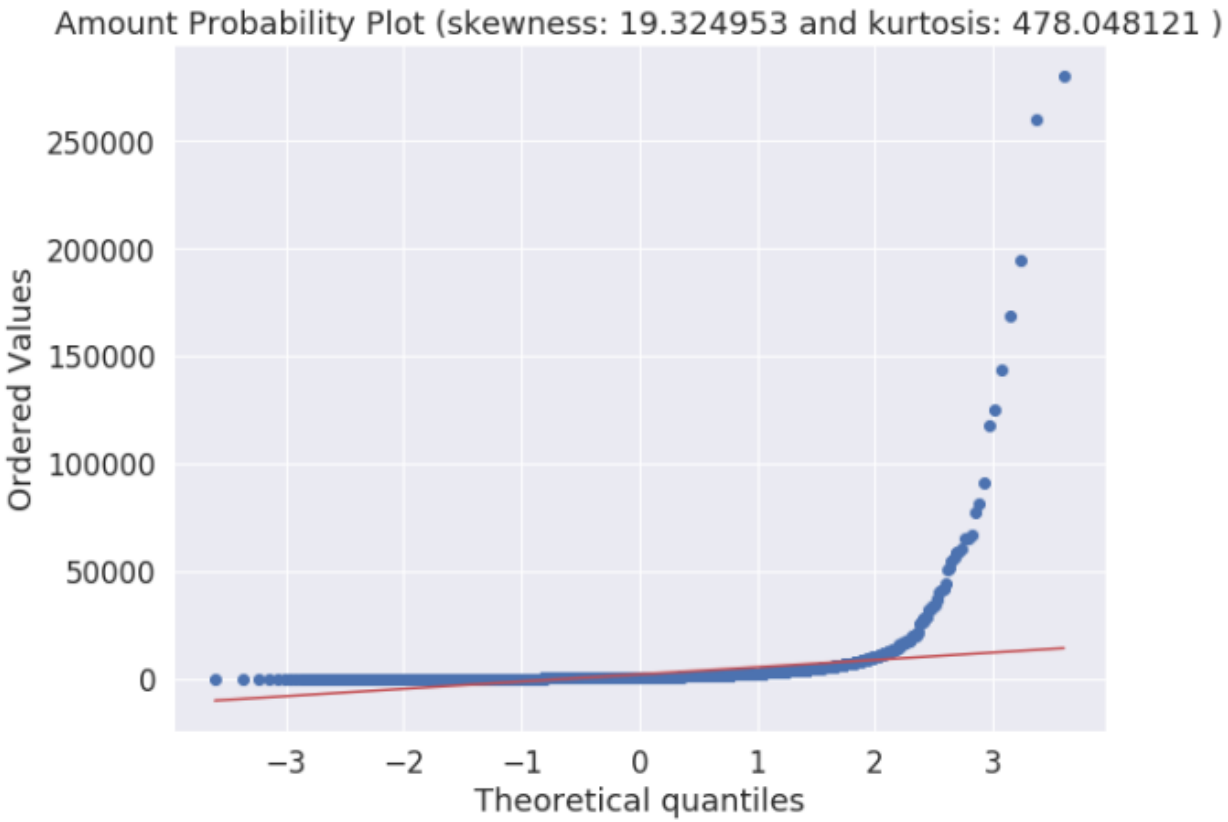


Probability Plot: In the QQ plot, the frequency values also deviate from the normal distribution line. The significant positive skewness (12.1) demonstrates a high level of asymmetry, meaning most frequency values are clustered at the lower end. The high kurtosis value (249) reveals a heavy-tailed distribution with a considerable number of outliers, showing that some customers have very high purchase frequencies compared to the rest.

Monetary Value Analysis:



Distribution Plot: The monetary value distribution plot has a peak near zero with a long right tail, indicating that while most customers spend small amounts, a few customers have made large purchases. This positively skewed distribution suggests that a minority of customers contribute disproportionately to the total revenue.



Probability Plot: The QQ plot confirms that monetary values are not normally distributed, with deviations from the red line, especially on the right. The high positive skewness (19.3) indicates a significant right skew, while the very high kurtosis (478) implies a heavily-tailed distribution, reflecting numerous outliers with exceptionally high spending.

Summary of Interpretations:

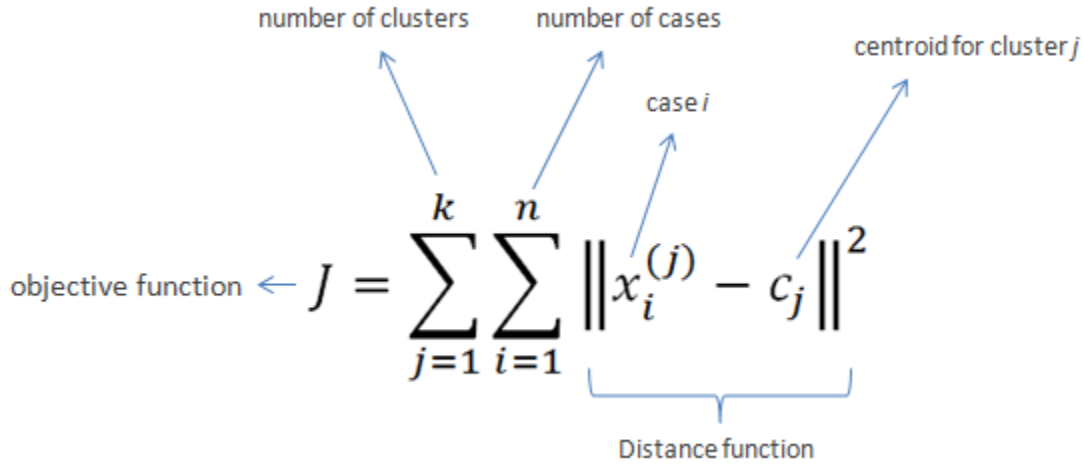
- All three features—recency, frequency, and monetary value—display right-skewed distributions with heavy tails.
- **Skewness** values in all cases are positive, which shows that the distributions lean toward lower values with long right tails.
- **Kurtosis** values are high, particularly for frequency and monetary value, indicating the presence of outliers and extreme values.

These distributions indicate that a small number of customers are responsible for high frequency and high spending, while most have lower values in these metrics. This pattern is typical in customer segmentation and RFM analysis, where a inority of high-value customers contribute significantly more than the average customer.

K-Means Clustering

The K-means clustering belongs to the partition based\centroid based hard clustering family of algorithms, a family of algorithms where each sample in a dataset is assigned to exactly one cluster.

Based on this Euclidean distance metric, we can describe the k-means algorithm as a simple optimization problem, an iterative approach for minimizing the within-cluster sum of squared errors (SSE), which is sometimes also called cluster inertia. So, the objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



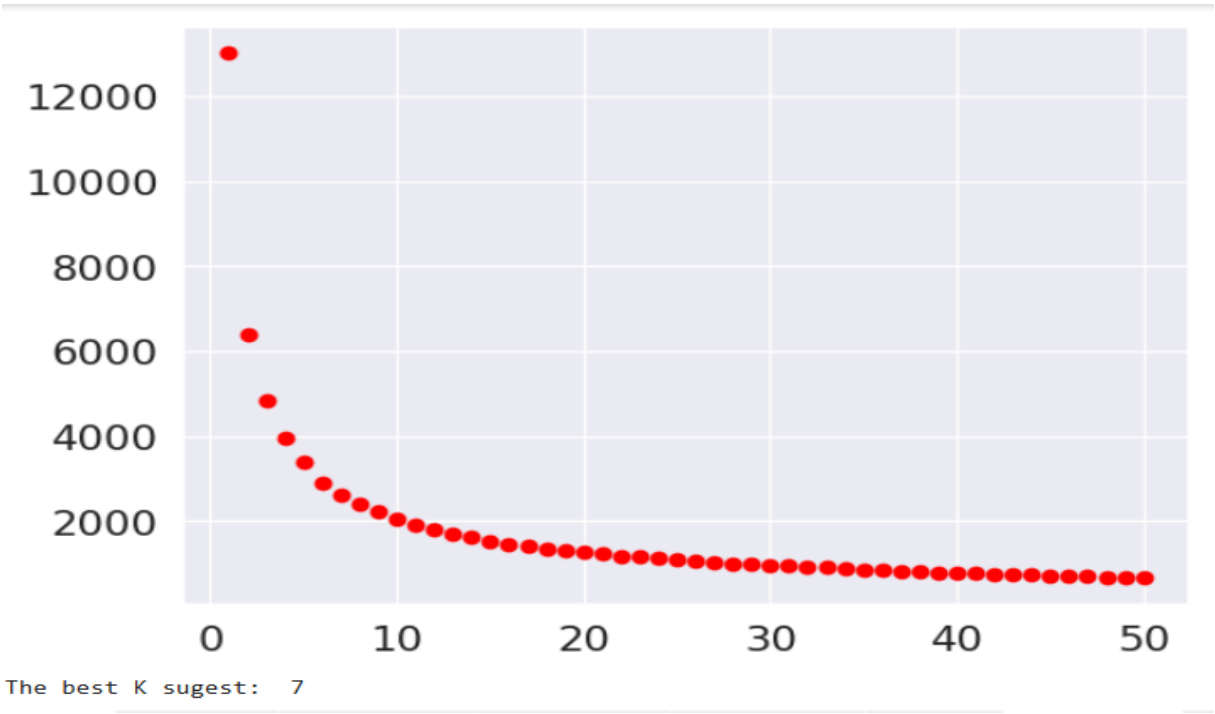
The steps that happen in the K-means algorithm for partitioning the data are as given follows:

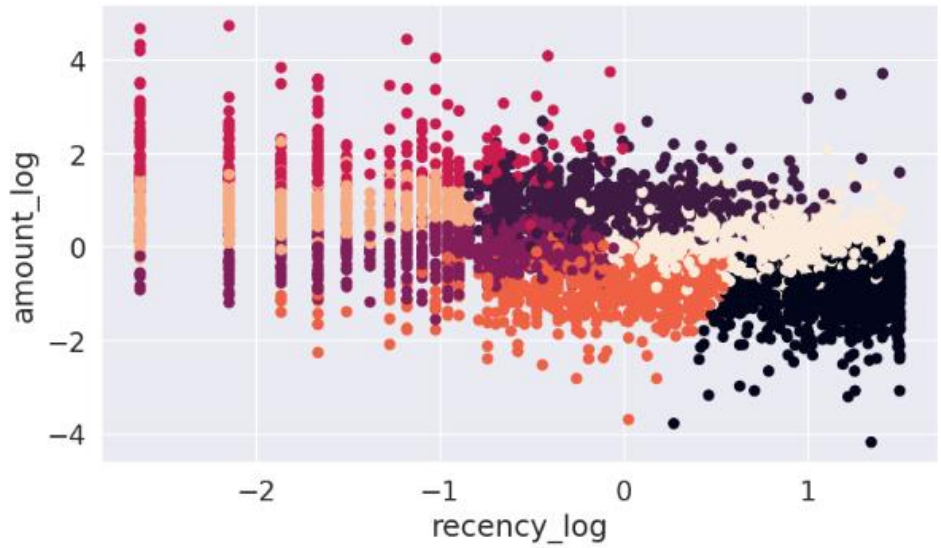
1. The algorithm starts with random point initializations of the required number of centers. The “K” in K-means stands for the number of clusters.
2. In the next step, each of the data point is assigned to the center closest to it. The distance metric used in K-means clustering is normal Euclidian distance.
3. Once the data points are assigned, the centers are recalculated by averaging the dimensions of the points belonging to the cluster.
4. The process is repeated with new centers until we reach a point where the assignments become stable. In this case, the algorithm terminates.

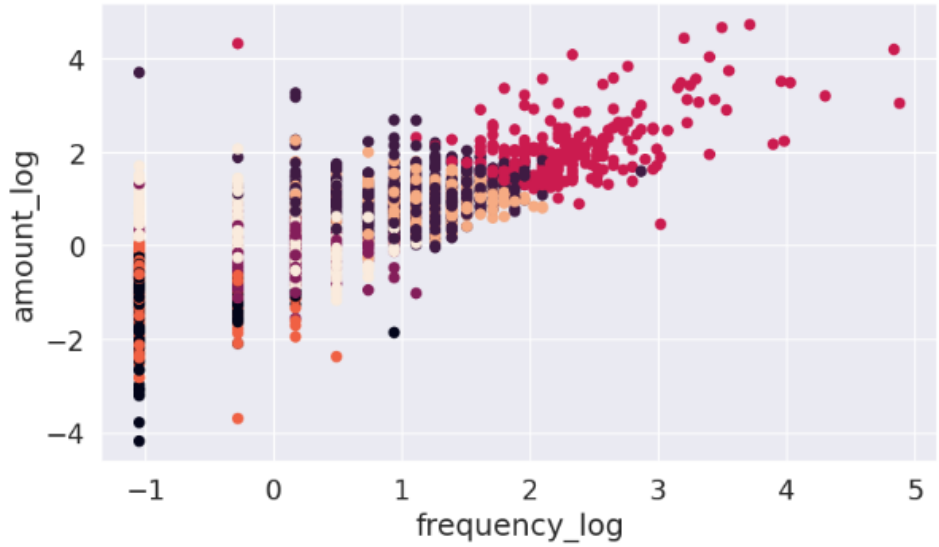
The Elbow Method to determine “K”

Using the elbow method to find the optimal number of clusters. The idea behind the elbow method is to identify the value of k where the distortion begins to increase most rapidly. If k increases, the distortion will decrease, because the samples will be closer to the centroids they are assigned to.

This method looks at the percentage of variance explained as a function of the number of clusters. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified. Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test. A slight variation of this method plots the curvature of the within group variance.







Note that by the Elbow method from a K equal to 3 we already observed low rates of gain in the decay of the distortions with the decrease of K reaching the limit of 10% with the K equal to 7. With this in mind, we will begin to evaluate the options more deeply with 3, and 7, starting with the silhouette analysis.

Silhouette analysis on K-Means clustering

Silhouette analysis can be used to study the separation distance between the resulting clusters, as a strategy to quantifying the quality of clustering via graphical tool to plot a measure of how tightly grouped the samples in the clusters are. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually.

It can also be applied to clustering algorithms other than k-means

Silhouette coefficients has a range of [-1, 1], it calculated by:

1. Calculate the cluster cohesion a(i)as the average distance between a sample x(i) and all other points in the same cluster.
2. Calculate the cluster separation b(i) from the next closest cluster as the average distance between the sample x(i) and all samples in the nearest cluster.
3. Calculate the silhouette s(i) as the difference between cluster cohesion and separation divided by the greater of the two, as shown here:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Which can be also written as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

Where:

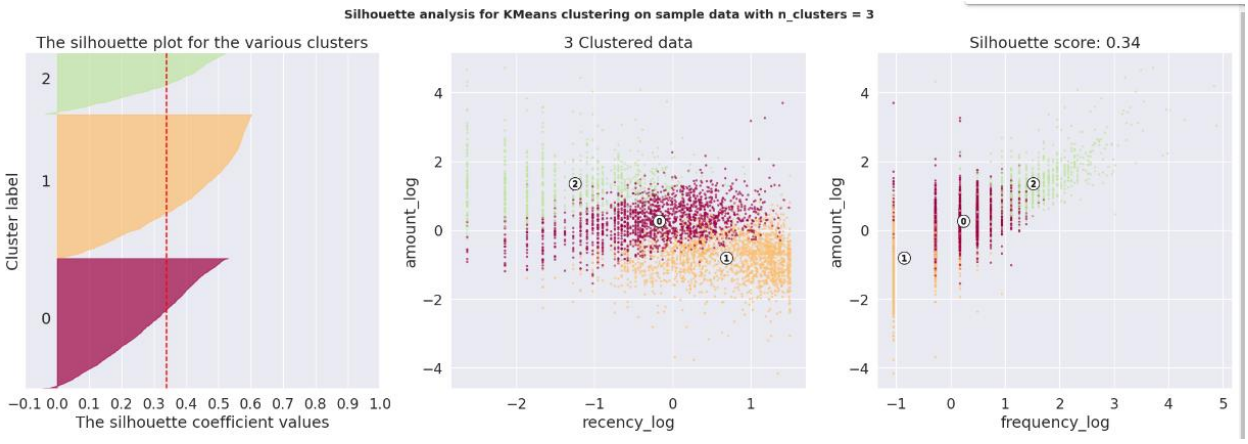
- If near +1, it indicate that the sample is far away from the neighboring clusters.
- a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.
- If most objects have a high value, then the clustering configuration is appropriate.
- If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

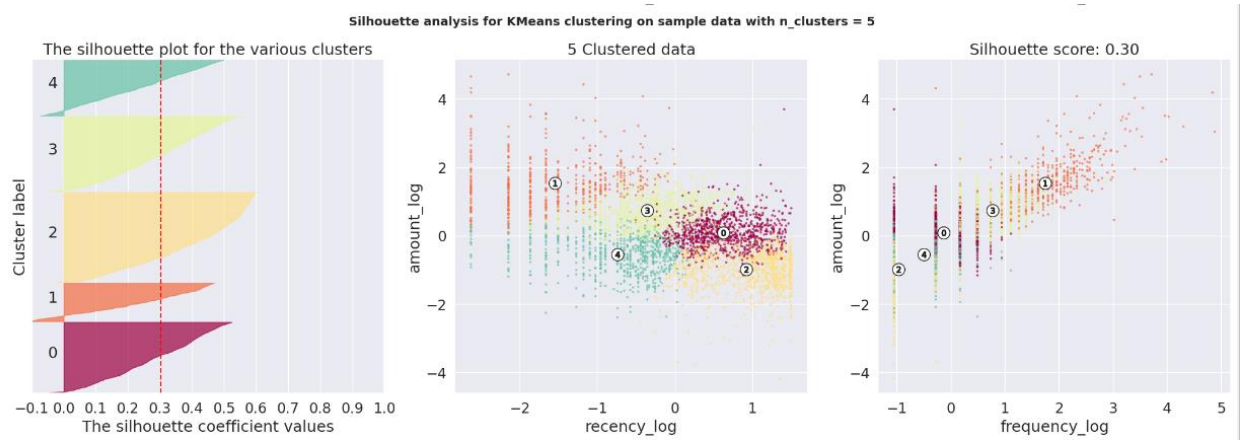
- A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters
- Negative values indicate that those samples might have been assigned to the wrong cluster.

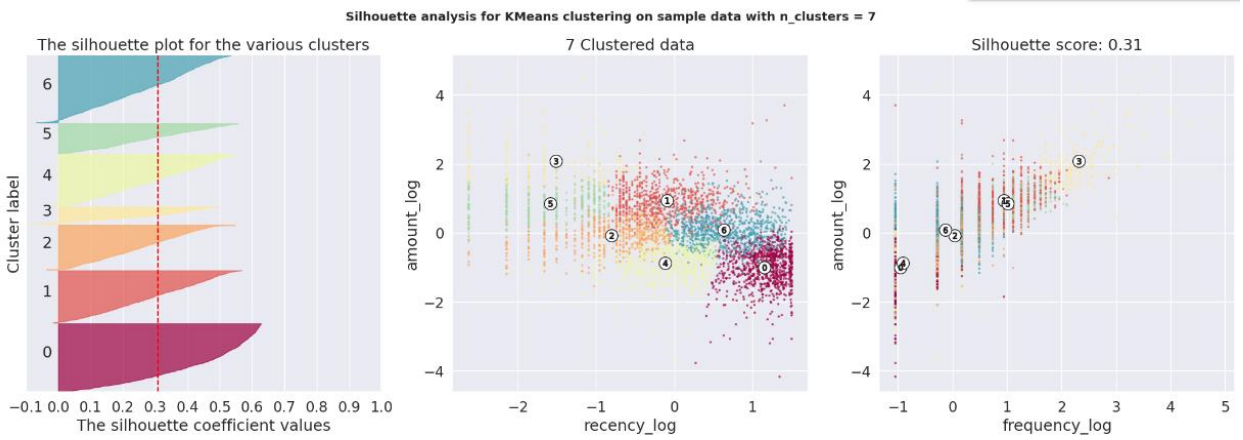
The silhouette plot can shows a bad K clusters pick for the given data due to the presence of clusters with below average silhouette scores and also due to wide fluctuations in the size of the silhouette plots. A good k clusters can found when all the plots are more or less of similar thickness and hence are of similar sizes.

Although we have to keep in mind that in several cases and scenarios, sometimes we may have to drop the mathematical explanation given by the algorithm and look at the business relevance of the results obtained.

Let's see below how our data perform for each K clusters groups (3, 5 and 7) in the silhouette score of each cluster, along with the center of each of the cluster discovered in the scatter plots, by amount_log vs recency_log and vs frequency_log.







When we look at the results of the clustering process, we can infer some interesting insights:

- First notice that all K clusters options is valid, because they don't have presence of clusters with below average silhouette scores.
- In the other hand, all options had a some wide fluctuations in the size of the silhouette plots.

So, the best choice may lie on the option that gives us a simpler business explanation and at the same time target customers in focus groups with sizes closer to the desired.

Clusters Center:

Let's look at the cluster center values after returning them to normal values from the log and scaled version.

```
for 3 clusters the silhouette score is 0.34
Centers of each cluster:
        amount     recency  frequency
0   261.952265  116.604917   1.190876
1  3967.994380    7.236580  10.044493
2  1006.914317   33.819966   3.152227
--------------------------------------------------
for 5 clusters the silhouette score is 0.31
Centers of each cluster:
        amount     recency  frequency
0   213.876290  159.060239   1.088129
1  5708.668108    4.285608  13.677542
2  1929.872406   22.442129   5.413014
3   372.314665   14.590855   1.665686
4   863.093356  100.092666   2.395562
--------------------------------------------------
for 7 clusters the silhouette score is 0.31
Centers of each cluster:
         amount     recency  frequency
0    809.713152  107.590047   2.277095
1   2115.751105    4.436558   6.395614
2    239.805507   36.372861   1.132543
3    667.345658   13.698858   2.663541
4    205.016462  225.462781   1.082459
5   2414.804796   38.026754   6.003854
6  10182.351681    4.961015  20.687947
--------------------------------------------------
```

Clusters Insights:

With the plots and the center in the correct units, let's see some insights by each clusters groups:

In the three-cluster:

- The tree clusters appears have a good stark differences in the Monetary value of the customer, we will confirm this by a box plot.
- Cluster 1 is the cluster of high value customer who shops frequently and is certainly an important segment for each business.
- In the similar way we obtain customer groups with low and medium spends in clusters with labels 0 and 2, respectively.

- Frequency and Recency correlate perfectly to the Monetary value based on the trend (High Monetary-Low Recency-High Frequency).

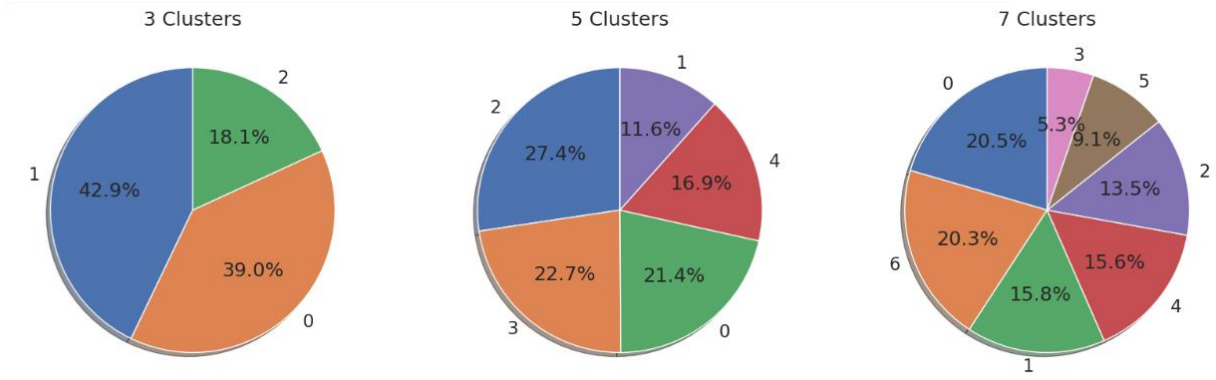
In the five-cluster:

- Note that clusters 0 and 1 are very similar to their cluster in the configuration with only 3 clusters.
- The cluster 1 appears more robust on the affirmation of those who shop often and with high amount.
- The cluster 2 are those who have a decent spend but are not as frequent as the cluster 1
- The cluster 4 purchases medium amounts, with a relatively low frequency and not very recent
- The cluster 3 makes low-cost purchases, with a relatively low frequency, but above 1, and made their last purchase more recently. This group of customers probably response to price discounts and can be subject to loyalty promotions to try increase the medium-ticket, strategy that can be better defined when we analyzing the market basket.
- The silhouette score matrix says that the five cluster segments are less optimal then the three cluster segments.

In the five-cluster:

- Definitely cluster 6 defines those who shop often and with high amount.
- Clusters 1 and 5 show good spending and good frequency, only deferring in how recent were their last purchases, where 5 is older, which suggests an active action to sell to group 5 as soon as possible and another to 1 seeking to raise its frequency.
- Cluster 0 presents the fourth best purchase and a reasonable frequency, but this is a long time without buying. This group should be sensible to promotions and activations, so that they do not get lost and make their next purchase.
- Cluster 5 is similar to 0, but has made its purchases more recently and has a slightly better periodicity. Then actions must be taken to raise their frequency and reduce the chances of them migrating to cluster 0 by staying longer without purchasing products.

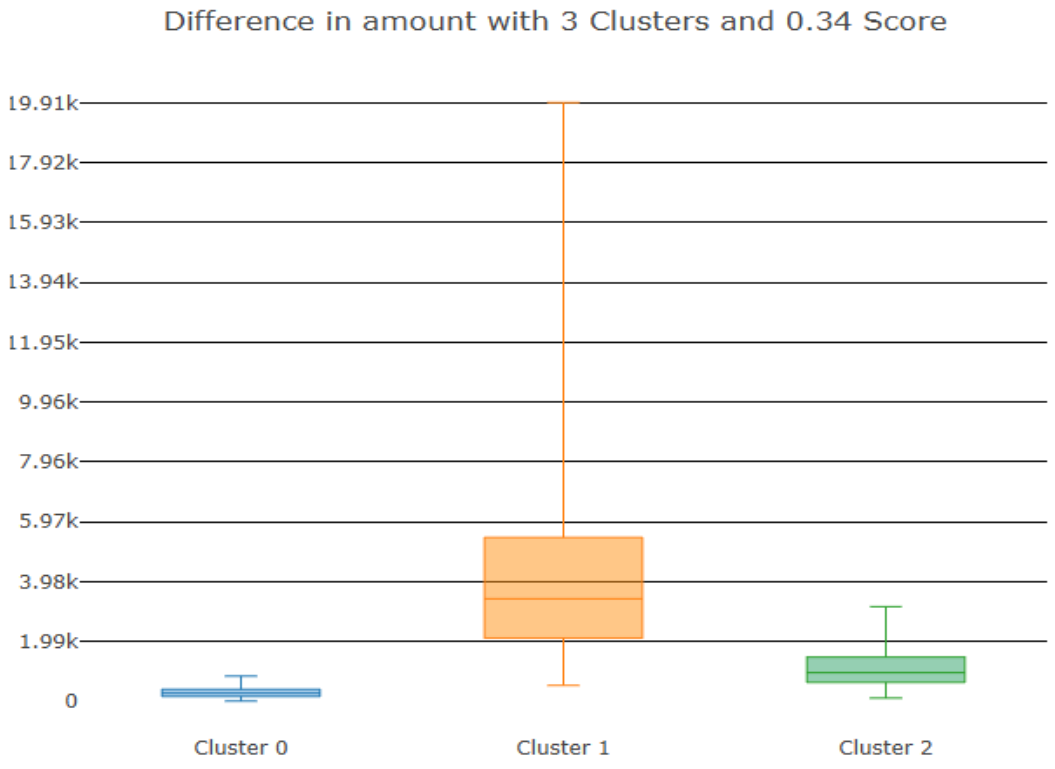
Drill Down Clusters: To further drill down on this point and find out the quality of these difference, we can label our data with the corresponding cluster label and then visualize these differences.

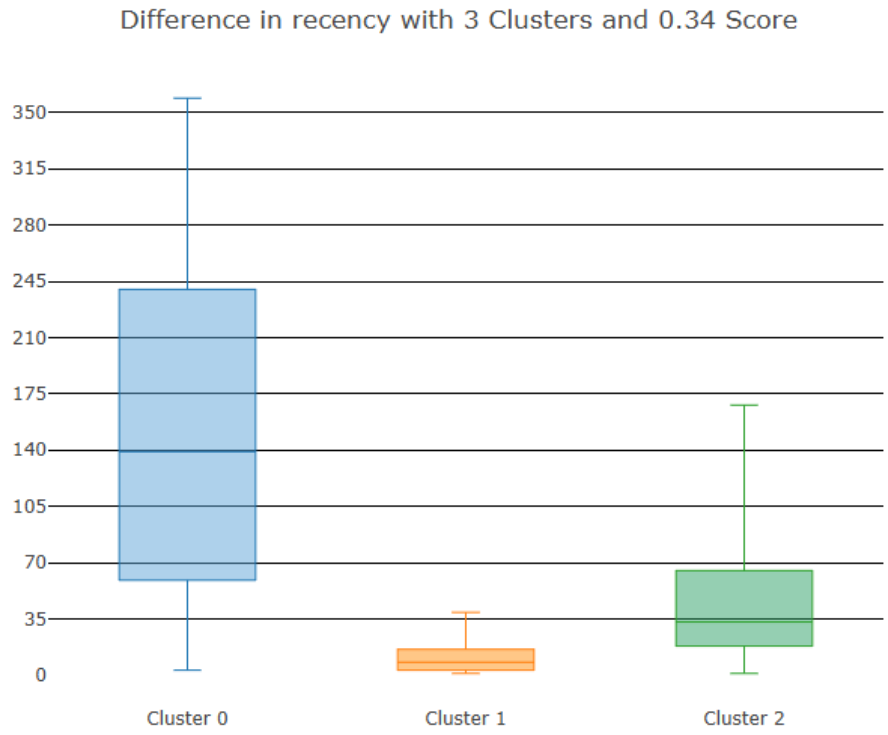


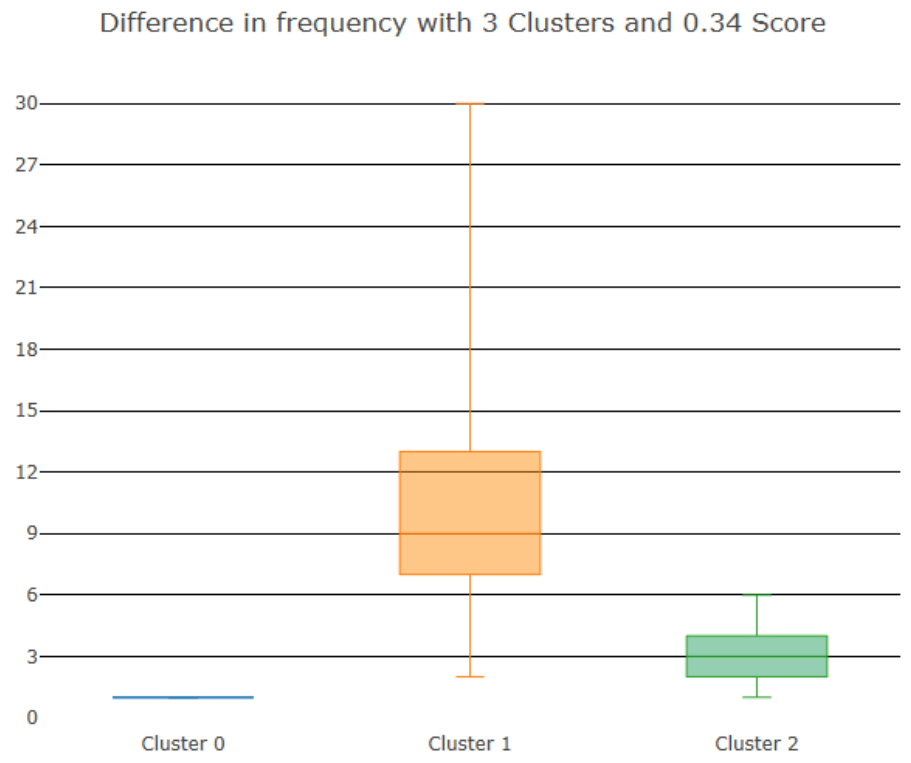
By visualizing each cluster, we aimed to highlight the key differences in central values and spread, offering insights into typical behaviors and characteristics across segments.

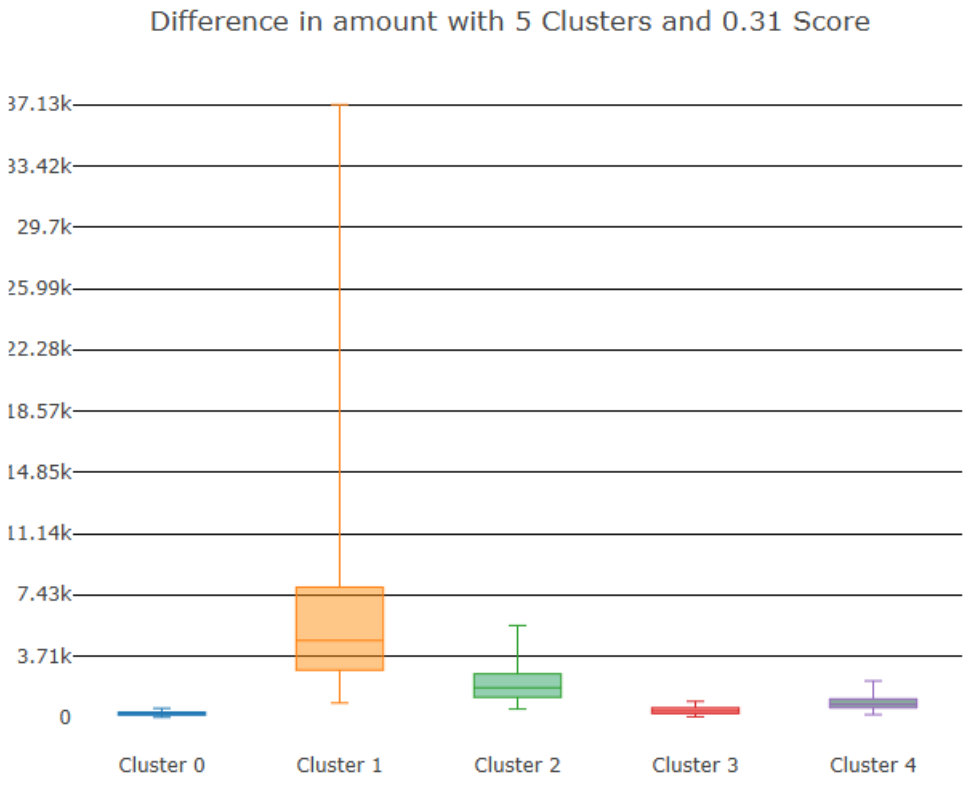
To achieve this, we focused on the majority of data points within each cluster, setting a threshold at the 95th percentile. This approach was chosen to minimize the impact of extreme values, which can distort observations, especially given that the data contains only positive values. By excluding these outliers, we ensured that the analysis would provide a more accurate view of each cluster's core tendencies, reflecting the typical behavior within each segment rather than being skewed by exceptional cases.

Using boxplots, we compared clusters based on central tendency, spread, and interquartile range, which revealed differences in median values and variability. This helped us gain a comparative view of each cluster's typical characteristics, enhancing our understanding of each segment's profile. Through this method, we were able to extract more reliable insights, making it possible to identify common traits and trends within clusters without interference from outliers.

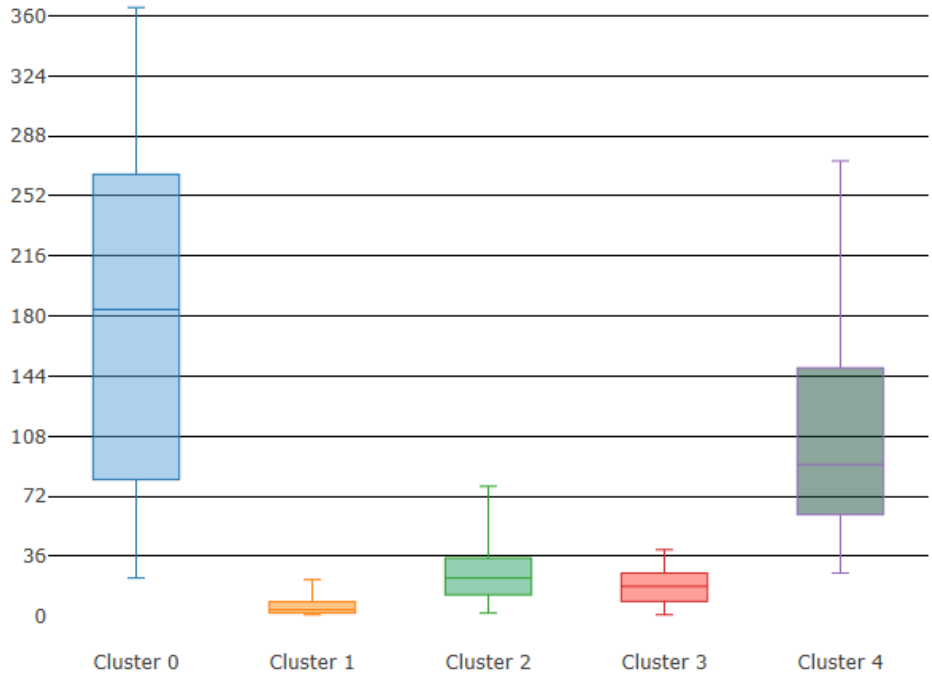

Difference in amount with 3 Clusters and 0.34 Score
19.91k
17.92k
15.93k
13.94k
11.95k
9.96k
7.96k
5.97k
3.98k
1.99k
0
Cluster 0
Cluster 1
Cluster 2


Difference in recency with 3 Clusters and 0.34 Score
350
315
280
245
210
175
140
105
70
35
0
Cluster 0
Cluster 1
Cluster 2

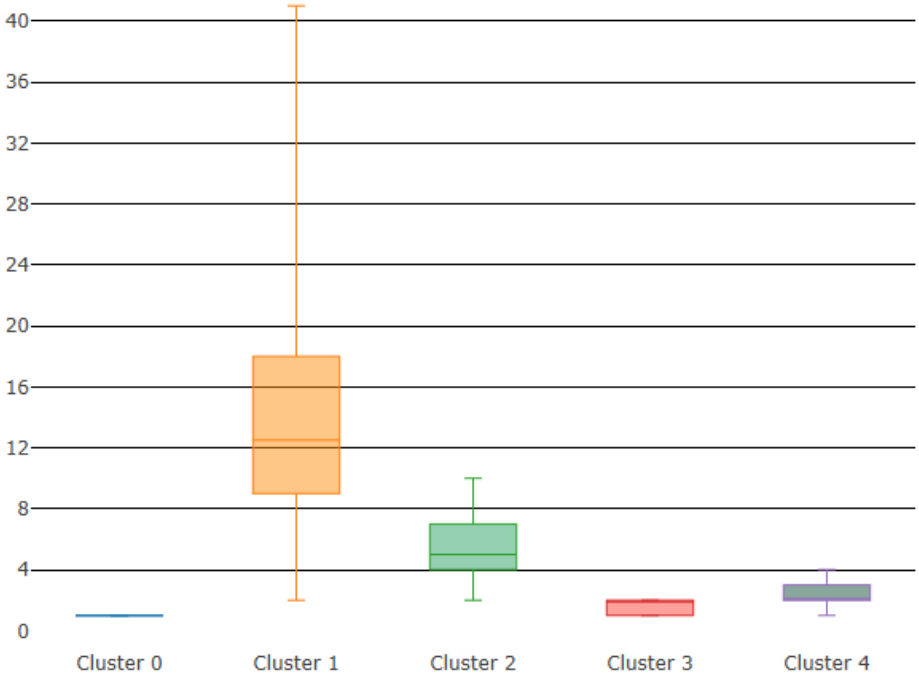

Difference in frequency with 3 Clusters and 0.34 Score
30
27
24
21
18
15
12
9
6
3
0
Cluster 0
Cluster 1
Cluster 2

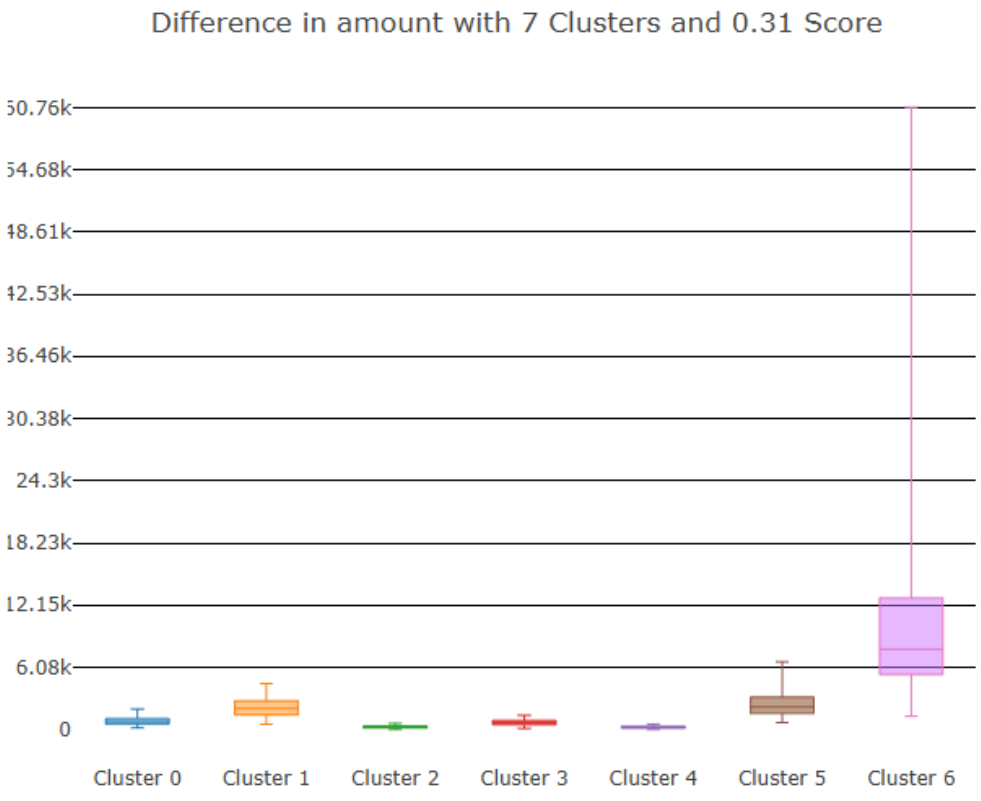

Difference in amount with 5 Clusters and 0.31 Score
37.13k
33.42k
29.7k
25.99k
22.28k
18.57k
14.85k
11.14k
7.43k
3.71k
0
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4

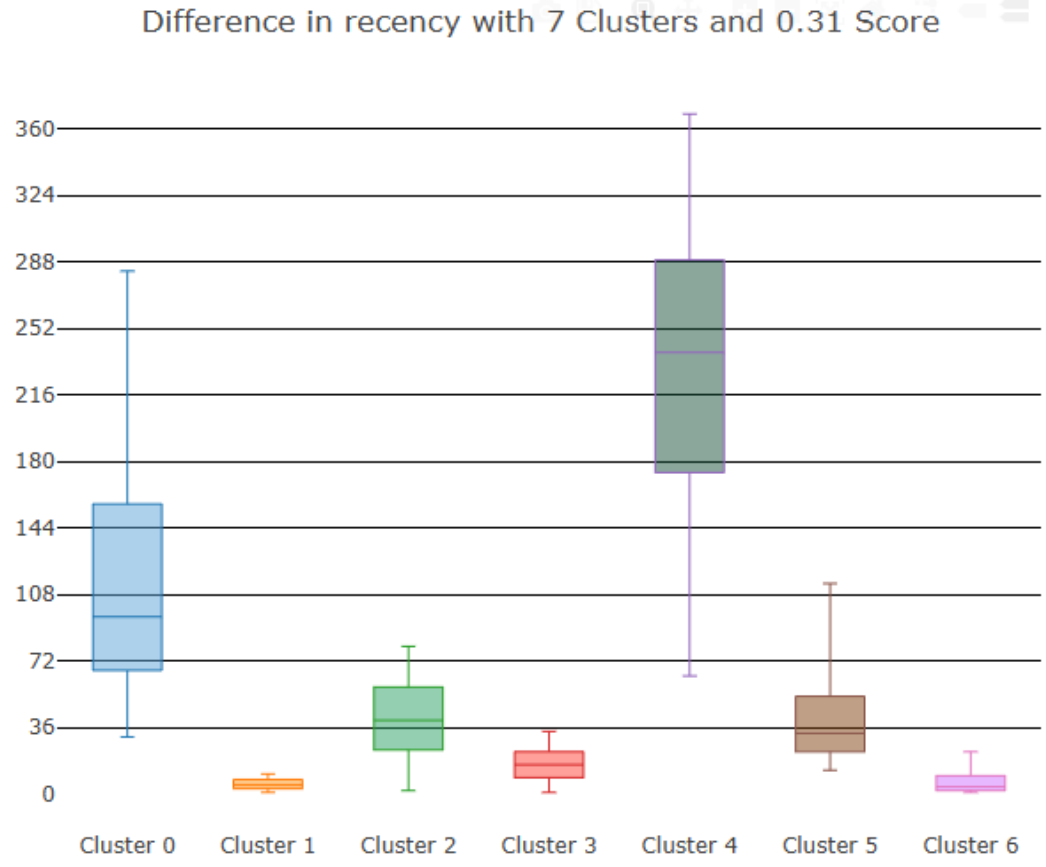
Difference in recency with 5 Clusters and 0.31 Score

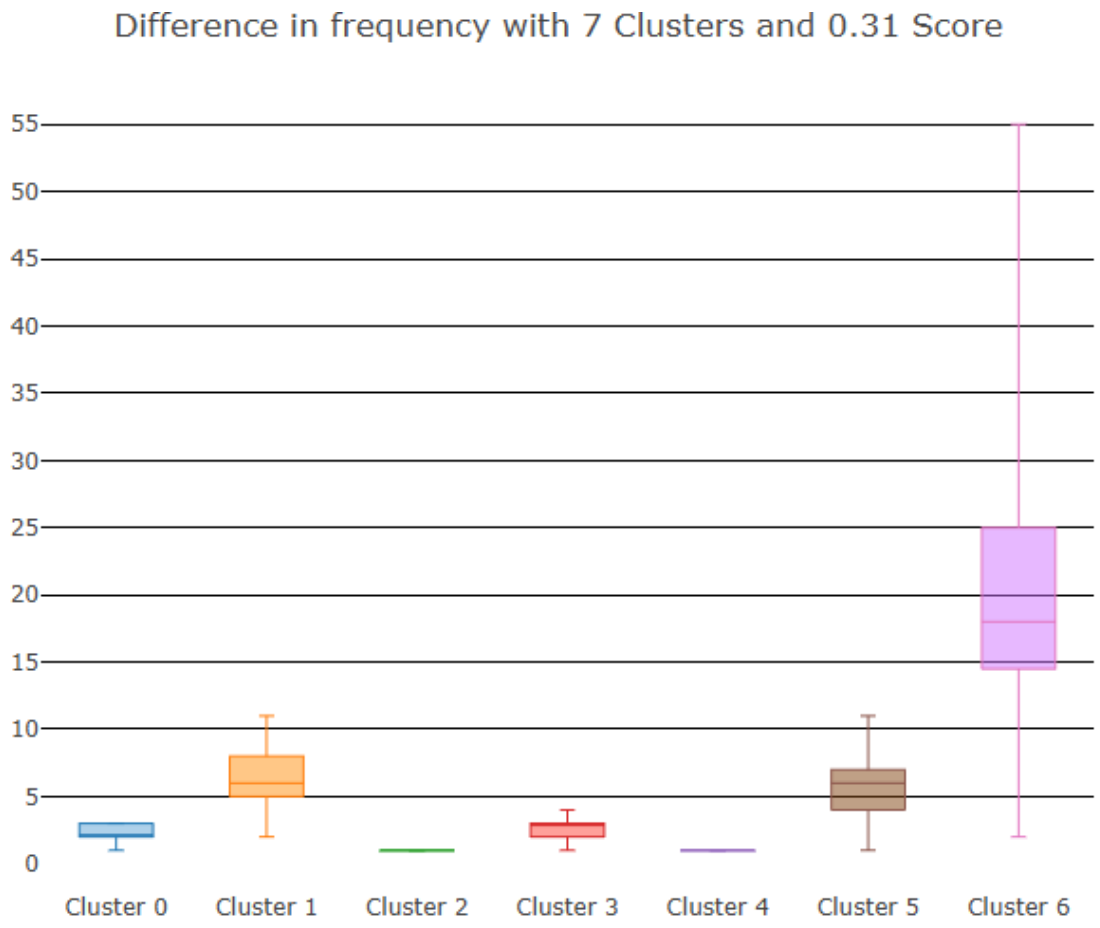


Difference in frequency with 5 Clusters and 0.31 Score




Difference in amount with 7 Clusters and 0.31 Score
50.76k
54.68k
48.61k
42.53k
36.46k
30.38k
24.3k
18.23k
12.15k
6.08k
0
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6


Difference in recency with 7 Clusters and 0.31 Score
360
324
288
252
216
180
144
108
72
36
0
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6



Next Steps in the Segmentation:

To enhance discovery and can further improve the quality of clustering by adding relevant features, other customer information and purchases details may be included in this dataset.

For example, but not limited to:

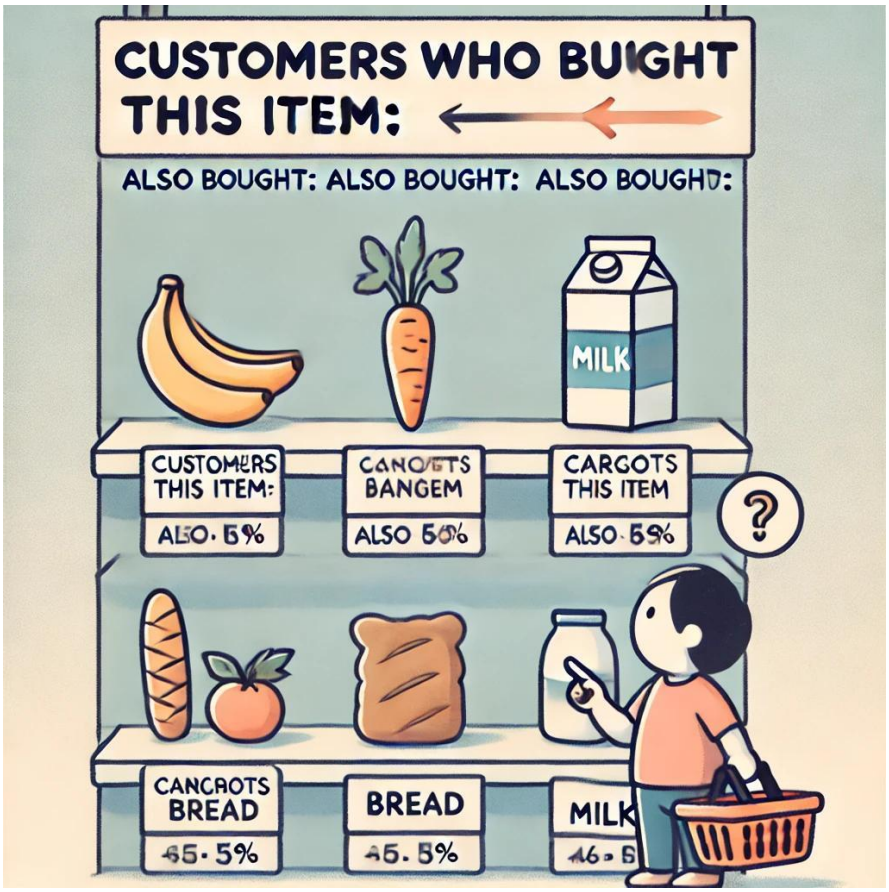
- New indicators, such as customer relationship time, based on the date of the first purchase of the client
- whether the customer is from abroad or not
- some group or category of product to be obtained through the SKUs
- External data vendors and use it, and so on.

Cross Selling

Cross-selling is a strategy that aims to encourage customers to buy additional products by leveraging insights into their shopping habits and general purchasing trends. By analyzing these patterns, retailers can identify products that align well with a customer's preferences, making suggested items more appealing and relevant. Often, these recommended items are bundled with attractive discounts or promotions, making it likely that customers will opt for the bundle over just the original product.

This approach involves examining transaction data to identify complementary products that may suit a customer's needs, presenting these suggestions in hopes of enhancing both customer satisfaction and sales for the retailer. In this section, we explore association rule mining, a robust technique ideal for identifying these relationships, and apply market basket analysis to our retail transactions dataset to uncover potential cross-selling opportunities.

Market Basket Analysis with Association Rule-Mining



Association rule mining is founded on the idea that customer purchasing behavior follows certain patterns, which can be leveraged to recommend additional products to customers in the future.

This technique, known as association rule learning, is a rule-based machine learning method that uncovers relationships between variables in large datasets. By identifying significant patterns or "strong rules" within the data, this approach uses measures of interestingness to filter valuable insights. As more data is analyzed, the method continues to generate new rules, expanding its understanding of customer behavior. With a sufficiently large dataset, the goal is to help a machine replicate the human brain's ability to extract features and identify associations from uncategorized data.

An association rule usually has the structure like below:

$$\{\text{butter, bread}\} \rightarrow \{\text{milk}\}$$

This rule can be read in the obvious manner that when the customer bought items on the left of the rule he is likely to buy the item on the right. Following are some vital concepts pertaining to association rule-mining.

- **Itemset**: Is just a collection of one or more items that occur together in a transaction. For example, here {milk, bread} is example of an itemset.
- **Support**: is defined as number of times an itemset appears in the dataset. The support of ***X*** with respect to ***T*** is defined as the proportion of transactions ***t*** in the dataset which contains the itemset ***X***. Mathematically it is defined as:

$$\operatorname{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

- **Confidence**: Confidence is an indication of how often the rule has been found to be true. It is a measure of the times the number of times a rule is found to exist in the dataset. For a rule which states {beer → diaper} the confidence is mathematically defined as:

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$$

- **Lift**: Lift of the rule is defined as the ratio of observed support to the support expected in the case the elements of the rule were independent. For the previous set of transactions if the rule is defined as {X → Y}, then the lift of the rule is defined as:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cap Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

  - If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.

  - If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.

    - If the lift is < 1, that lets us know the items are substitute to each other. This means that presence of one item has negative effect on presence of otheritem and vice versa.

- **Frequent itemset**: Frequent itemsets are itemsets whose support is greater than a user defined support threshold.
- **Conviction**: Is the ratio of the expected frequency that item X occurs without a item Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions. The conviction of a rule is defined as:

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$$

Algorithms in Association Rule Mining

Several well-known algorithms, such as Apriori, Eclat, and FP-Growth, are commonly used for mining frequent itemsets. However, these algorithms alone are not sufficient to complete association rule mining; an additional step is needed to generate rules from the frequent itemsets identified.

A primary challenge in association rule-mining algorithms is the generation of frequent itemsets. If a transactional dataset contains (k) unique items, there could be up to (2^k) possible itemsets, making the task computationally intensive.

1. Apriori Algorithm

The Apriori algorithm employs a breadth-first search approach to count the support of itemsets and uses a candidate generation function that leverages the downward closure property of support. The algorithm first generates candidate itemsets and then proceeds to identify frequent itemsets among them. For a dataset with approximately 100 unique items, the number of possible itemsets is enormous, making Apriori computationally expensive and sometimes impractical.

2. Eclat Algorithm

The Eclat (Equivalence Class Transformation) algorithm, on the other hand, uses a depth-first search technique based on set intersection. It is well-suited for both sequential and parallel processing due to its ability to work with locality-enhancing properties, which makes it more efficient for some applications than Apriori.

3. FP-Growth Algorithm

FP-Growth (Frequent Pattern Growth) improves upon the Apriori algorithm by eliminating the need to generate all candidate itemsets. Instead, it uses a divide-and-conquer approach with a specialized data structure called the FP-tree to find frequent itemsets without generating every possible itemset. The main steps of FP-Growth are:

1. First Pass: The algorithm scans the transactional dataset to count the occurrences of each item (or attribute-value pair) and stores this information in a 'header table.'

2. Second Pass: It constructs the FP-tree structure by adding transactions as instances of frequent itemsets. Items in each transaction are sorted in descending order by frequency, which enables efficient processing. Items that do not meet the minimum support threshold are discarded,

and if many transactions share common items, the FP-tree compresses them near the root.

3. Conditional Dataset Creation: The FP-tree is divided into multiple conditional datasets, each associated with a frequent pattern. This allows for efficient pattern mining in each subset.

4. Pattern Mining: Patterns in each subset are mined by recursively concatenating shorter patterns into longer ones, making the process highly efficient. The algorithm recursively processes this compressed version of the dataset, growing large itemsets directly instead of generating and testing candidate itemsets against the full dataset.

5. Recursive Growth: The recursive growth begins from the bottom of the header table (where the longest branches are) by finding all instances that match the given condition. A new FP-tree is created, with counts projected from the original tree based on the set of instances conditional on a particular attribute. Each node receives the cumulative count of its children. The recursion stops when no individual items conditional on the attribute meet the minimum support threshold, and the algorithm continues to process the remaining header items.

After completing the recursive process, the algorithm identifies all large itemsets that meet the minimum support threshold, which are then used to generate association rules.

Building the Transaction Dataset

To apply algorithms such as Apriori, Eclat, and FP-Growth, we first need to transform our dataset into a transaction table. In this format, each unique product sold is represented by a column. For each transaction (or sales event), the table records a value of 1 if the product was sold in that transaction and 0 if it was not.

This binary representation allows us to analyze sales patterns and determine which products are frequently purchased together, laying the foundation for generating association rules that are useful for cross-selling and recommendation strategies.

Dataset Pruning for Frequently Purchased Items

In our earlier EDA, we observed that a small number of items contribute to the majority of our sales. To refine our dataset accordingly, we've created the `prune_dataset` function, which allows us to reduce the dataset size based on specific criteria. This function can perform three types of pruning:

1. Pruning Based on Total Sales Percentage: Using the parameter `total_sales_perc`, we can select a subset of items that account for a specified percentage of total sales, with a default value of 50% or 0.5.

2. Pruning Based on Item Rank: Alternatively, we can prune the dataset by specifying the starting and ending ranks for the items we want to retain.

3.Pruning Based on Specific Features: By passing a list of columns to the `TopCols` parameter, we can filter the dataset to include only items with those specific features.

Additionally, the function defaults to transactions containing at least two items, as transactions with a single item do not align with the concept of association rule mining.

```
Total of Sales Amount by the Top 15 Products in Sales Events (Invoice): 778377.21
Number of Sales Events: 4664
Number of Products: 15
```

	item_name	item_count
0	WHITE HANGING HEART T-LIGHT HOLDER	1978
1	REGENCY CAKESTAND 3 TIER	1703
2	JUMBO BAG RED RETROSPOT	1600
3	PARTY BUNTING	1379
4	ASSORTED COLOUR BIRD ORNAMENT	1375
5	LUNCH BAG RED RETROSPOT	1289
6	SET OF 3 CAKE TINS PANTRY DESIGN	1146
7	POSTAGE	1099
8	JUMBO BAG VINTAGE DOILY	1080
9	LUNCH BAG BLACK SKULL.	1052
10	LUNCH BAG SUKI DESIGN	1043
11	POPCORN HOLDER	1035
12	PACK OF 72 RETROSPOT CAKE CASES	1029
13	SPOTTY BUNTING	1009
14	LUNCH BAG VINTAGE DOILY	1006

Identification of Key Items Contributing to Sales

Through our analysis, we identified that **15 items are responsible for 8.73% of the total sales volume**. Additionally, approximately **5% of all transactions, totaling 4,664 events**, include these items alongside others in the same purchase.

This insight forms the basis for the next stage of our process, where we will transform this selected subset of data into the required tabular structure for further analysis. This restructuring is essential to ensure the data is compatible with the intended analytical methodologies, such as association rule mining, and allows us to focus on high-impact items in the dataset.

Association Rule Mining: Support and Confidence

Step 1: Setting Minimum Support

To identify frequently occurring itemsets, we set a minimum support threshold of 1% (0.01). This value specifies that only items appearing in at least 1% of all transactions will be included in the rule-mining process. Based on this threshold, 663,273 itemsets were generated. This high volume underscores the numerous item combinations possible even with a low support threshold.

Step 2: Setting Minimum Confidence

To refine the itemsets further, we defined a minimum confidence level of 60%. Confidence measures the likelihood that an item will appear in a transaction given the presence of another item. By setting a higher confidence threshold, we filtered out less significant associations, focusing on more reliable co-occurrence patterns.

Results of Rule Generation

Applying the defined support and confidence levels yielded a raw data frame containing 25,247 association rules. These rules were subsequently pruned to retain only the most meaningful associations, with calculations of support, confidence, coverage, strength, lift, and leverage for each rule.

Exploration of Generated Rules

The strongest rules were identified based on the following criteria:

	consequent	antecedent	support	confidence	lift
20	JUMBO BAG VINTAGE DOILY	JUMBO BAG RED RETROSPOT, LUNCH BAG RED RETROSPOT, LUNCH BAG VINTAGE DOILY	60	0.909091	4.520256
131	LUNCH BAG VINTAGE DOILY	LUNCH BAG RED RETROSPOT, JUMBO BAG VINTAGE DOILY , LUNCH BAG BLACK SKULL., LUNCH BAG SUKI DESIGN	49	0.890909	4.596460
132	LUNCH BAG VINTAGE DOILY	LUNCH BAG RED RETROSPOT, JUMBO BAG VINTAGE DOILY , LUNCH BAG SUKI DESIGN	48	0.888889	4.586037
129	LUNCH BAG VINTAGE DOILY	LUNCH BAG RED RETROSPOT, JUMBO BAG VINTAGE DOILY , LUNCH BAG BLACK SKULL.	48	0.872727	4.502655
21	JUMBO BAG VINTAGE DOILY	JUMBO BAG RED RETROSPOT, LUNCH BAG SUKI DESIGN , LUNCH BAG VINTAGE DOILY	48	0.872727	4.339446

1.Highest Confidence: Rules with the highest confidence included items like "JUMBO BAG VINTAGE DOILY" and "LUNCH BAG VINTAGE DOILY," often paired with other items in high-confidence transactions. These rules highlight a strong likelihood of co-purchasing among certain items, with confidence levels above 0.87.

	consequent	antecedent	support	confidence	lift
121	LUNCH BAG VINTAGE DOILY	JUMBO BAG VINTAGE DOILY , LUNCH BAG RED RETROSPOT	176	0.789238	4.071908
28	JUMBO BAG VINTAGE DOILY	LUNCH BAG VINTAGE DOILY , JUMBO BAG RED RETROSPOT	153	0.805263	4.003995
90	LUNCH BAG RED RETROSPOT	LUNCH BAG VINTAGE DOILY , LUNCH BAG SUKI DESIGN	149	0.668161	2.672646
82	LUNCH BAG RED RETROSPOT	LUNCH BAG SUKI DESIGN , JUMBO BAG RED RETROSPOT	149	0.645022	2.580087
76	LUNCH BAG RED RETROSPOT	LUNCH BAG BLACK SKULL., LUNCH BAG SUKI DESIGN	149	0.605691	2.422764

2. Highest Support: The top rules by support further confirmed frequently co-purchased items, with support values indicating robust occurrence patterns across the dataset. Items such as "LUNCH BAG VINTAGE DOILY" and "JUMBO BAG VINTAGE DOILY" consistently appeared together, emphasizing their popularity.

Lift Analysis

Lift is a measure that indicates the strength of an association between items. A lift value greater than 1 suggests that the antecedent and consequent items are not independent, meaning they frequently co-occur beyond random chance. In this analysis, all 25,247 rules had lift values above 1, underscoring significant relationships and co-purchasing trends within the dataset.

Some concluding points :--

This report documented the end-to-end process of identifying meaningful item associations within a transactional dataset. By applying a 1% support threshold and a 60% confidence level, we identified numerous high-confidence itemsets, with strong lift values indicating

reliable associations. These insights provide a foundation for further strategic analysis, enabling more effective decision-making around inventory management, marketing, and targeted promotions.

As we have seen, the generation of rules is a very simple process, but very computationally expensive, since it grows exponentially with the increase of the set of items.

Overall, we seek the proper balance between support and confidence leading to a reasonable number of strong rules.

In the other hand, if the goal is to identify rare but with high confidence patterns, we should proceed as previously, by establishing a low level of support and a higher level of confidence, which leads to a large number of rules.

With this in mind, the rules with low support and high confidence would then be our target for further study and than outlining of strategies to raise cross selling.

Conclusion

1. Identification of High-Value Customers and Products

- The top 10 customers contributed 17.26% of total sales, emphasizing the importance of customer retention strategies targeting high-value clients.

- Key products, particularly the top two items, accounted for a significant portion of sales, collectively contributing 22.98% of total sales. This highlights the potential for focused promotions on best-selling items.

2. Customer Segmentation Insights

-RFM Analysis: segmented customers based on recency, frequency, and monetary value, revealing patterns such as:

-High-value customers with recent purchases, frequent transactions, and high spending.

- A long-tail distribution showed that while a few customers made frequent, high-value purchases, most had lower transaction volumes.

- K-means clustering grouped customers into segments with distinct behaviors. For example:

- Cluster 1 represented frequent, high-spending customers, identified as priority targets.

- Cluster 0 included infrequent, low-spending customers, suggesting potential for loyalty programs to increase engagement.

3. Association Rule Mining and Market Basket Analysis

- Using support and confidence thresholds of 1% and 60%, respectively, we generated 25,247 association rules to identify frequently co-purchased items.

- Items such as "JUMBO BAG VINTAGE DOILY" and "LUNCH BAG VINTAGE DOILY" emerged as frequent associations, with confidence values above 0.87 and lift values greater than 4, indicating strong potential for cross-selling.

- The findings suggest bundling complementary items to boost sales, as these high-confidence, high-lift rules highlight items often bought together.

4. Overall Findings and Strategic Implications

- The insights provide a data-driven foundation for inventory management, targeted marketing, and promotional strategies, focusing on high-value customers and popular product bundles.

- These strategies are projected to enhance customer engagement, drive sales, and improve overall revenue through tailored offerings and optimized product placement.

By combining RFM segmentation with market basket analysis, the report offers actionable strategies for increasing customer loyalty and optimizing sales by understanding and leveraging key patterns in purchasing behavior.

References

https://archive.ics.uci.edu/dataset/352/online+retail

https://en.wikipedia.org/wiki/Association_rule_learning

https://en.wikipedia.org/wiki/Market_segmentation