**Technical Report**

**Customer Segmentation and Market Basket Analysis on Online Retail Data**

### Table of Contents

1. **Abstract**
2. **Introduction**
   * 2.1 Background
3. **Literature Review**
   * 3.1 Key Studies
4. **Methodology**
   * 4.1 Data Collection and Cleaning
   * 4.2 Customer Segmentation
   * 4.3 Market Basket Analysis
   * 4.4 Tool and Libraries used
5. **Model Implementation**
6. **Results and Analysis**
   * 6.1 Customer Segmentation
   * 6.2 Silhouette Analysis
   * 6.3 Cluster Scatter Plots
   * 6.4 Cluster Distribution
   * 6.5 Market Basket Analysis
7. **Performance Analysis**
8. **Conclusion**
9. **Future Scope**
10. **References**
11. **Abstract**

The Online Retail dataset, which includes transaction records from a UK-based online retailer from December 2010 to December 2011, was thoroughly examined by our team. In order to facilitate targeted marketing and cross-selling tactics, we sought to categorize consumers and pinpoint product linkages. We used the RFM (Recency, Frequency, Monetary) model with K-Means clustering to segment our client base. This allowed us to discover several customer groupings, including a 42.9% dominant segment at k=3. We employed the FP-Growth algorithm for market basket analysis, which revealed strong product connections, with "JUMBO BAG VINTAGE DOILY" being frequently linked with "LUNCH BAG RED RETROSPOT." Although our method had issues with big rule sets and dataset biases, it was more efficient than more conventional approaches like Apriori and yielded useful insights.

1. **Introduction**

**2.1 Background**

Our team was driven to investigate how data analytics may improve decision-making in the retail industry, specifically for online enterprises. The Online Retail dataset, comprising over 500,000 transactions from a UK gift retailer, provided a valuable chance to analyze user behavior and product associations. The project had two main objectives:

(1) to categorize customers into significant groups according to their purchasing behaviors for tailored marketing, and

(2) to determine often co-purchased products to enhance cross-selling prospects. By integrating customer segmentation with market basket analysis, we aimed to provide a holistic solution that would enable the store to enhance its strategy and augment income.

1. **Literature Review**

**3.1 Key Studies**

Customer segmentation frequently use the RFM model, which evaluates clients based on Recency (time elapsed since the last purchase), Frequency (number of purchases), and Monetary value (total expenditure) (Hughes 1994). This technique is straightforward and efficient for identifying high-value clients; nevertheless, it may overlook more intricate behavioral patterns, like product preferences or seasonal changes (Fader, Hardie, and Lee 2005). Clustering techniques like as K-Means (MacQueen 1967) are frequently employed to categorize customers, while the Elbow Method and Silhouette Analysis (Rousseeuw 1987) assist in identifying the appropriate number of clusters. K-Means presumes that clusters are spherical and of uniform size, thereby oversimplifying intricate data distributions.   
Market basket analysis aims to uncover item relationships using algorithms such as Apriori (Agrawal and Srikant 1994), which formulates rules by detecting common itemsets. Although obvious, Apriori is computationally demanding due to its candidate generation mechanism, rendering it less appropriate for extensive datasets. The FP-Growth algorithm (Han, Pei, and Yin 2000) addresses this issue by employing a tree-based structure, thereby substantially decreasing computing time. The Eclat approach (Zaki 2000) employs a vertical data format; nevertheless, it is less efficient for dense datasets such as ours. Association rules are assessed through support, confidence, and lift (Tan, Steinbach, and Kumar 2005); nevertheless, these techniques may generate an excessive quantity of rules, necessitating meticulous trimming. Our team opted to amalgamate RFM-based clustering with FP-Growth to merge customer insights with product relationships, so overcoming the constraints of isolated methodologies.

1. **Methodology/System Architecture**

Our team developed a dual-faceted analytical pipeline utilizing Python within a Jupyter Notebook, concentrating on customer segmentation and market basket analysis. The following are the stages we executed, accompanied by comprehensive definitions and elucidations of the methods and algorithms employed at each phase:   
**4.1 Data Collection and Cleaning:** We utilized the Online Retail dataset, comprising about 541,909 transactions from December 1, 2010, to December 9, 2011.   
We sanitized the data by deleting missing values, removing duplicates, and assuring correct date formatting to facilitate analysis.   
**4.2 Customer Segmentation:**

o RFM Calculation: Definition: The RFM model (Recency, Frequency, Monetary) is a customer segmentation methodology that assesses customers based on three criteria: Recency (the duration since a customer's last purchase, usually measured in days), Frequency (the aggregate number of purchases made by a customer within a specified timeframe), and Monetary (the cumulative expenditure of a customer during that timeframe). It is extensively utilized in marketing to identify high-value consumers by assigning scores to each statistic, often on a scale (e.g., 1-5) and aggregating them to rank clients (Hughes 1994).   
We calculated Recency as the number of days since each client's last purchase in relation to the dataset's end date (09/12/2011), Frequency as the total number of unique invoices per customer, and Monetary as the total expenditure per customer (quantity multiplied by unit price). These ratings constituted the foundation for grouping.   
Data Preprocessing: Definition: In this context, preprocessing entails the transformation of raw data to enhance the efficacy of machine learning algorithms. Log-transformation is a prevalent method that utilizes the natural logarithm on numerical values, diminishing skewness in data distributions and alleviating the influence of outliers, which can disproportionately damage clustering techniques such as K-Means.   
We employed log-transformation on the RFM variables, yielding recency\_log, frequency\_log, and amount\_log, to standardize their distributions. This phase guaranteed that the clustering algorithm regarded each feature more fairly, as raw RFM values frequently display heavy-tailed distributions (e.g., a limited number of consumers with exceptionally high expenditure).   
**K-Means Clustering:**   
K-Means clustering is an unsupervised machine learning approach that divides a dataset into k clusters by minimizing the intra-cluster variance. It sequentially allocates data points to the closest cluster centroid (according to Euclidean distance) then adjusts the centroids to the average of the assigned points until convergence is achieved. K-Means++ is an improvement that initializes centroids by randomly selecting the first centroid and subsequently choosing others based on a probability proportionate to their distance from existing centroids, hence enhancing convergence speed and cluster quality (MacQueen 1967).   
We employed K-Means clustering with K-Means++ initialization to categorize clients according to their log-transformed RFM scores. The technique reduced the Within-Cluster Sum of Squares (WCSS), an indicator of intra-cluster variance, to create cohesive consumer groupings.   
**Cluster Selection:**   
**Elbow Method:**   
The Elbow Method is a heuristic used to ascertain the appropriate number of clusters (k) in K-Means clustering. This entails graphing the WCSS against k and pinpointing the "elbow" point, when the addition of clusters results in diminishing returns in WCSS, signifying a balance between model complexity and fit.   
We evaluated k values from 1 to 50, graphing WCSS to determine the elbow point, which indicated that k=3 to k=5 is an appropriate range for our dataset.   
Silhouette Analysis: A technique for assessing clustering quality by quantifying the similarity of each data point to its respective cluster in relation to other clusters. The silhouette coefficient for a point is determined by the formula (b - a) / max(a, b), where a represents the mean distance to other points within the same cluster, and b denotes the mean distance to points in the nearest neighboring cluster. The coefficient varies between -1 and 1, with larger values signifying more distinct clusters (Rousseaux 1987).   
We calculated silhouette coefficients for k=3, k=5, and k=7 to evaluate cluster separation and cohesion, facilitating the selection of the ideal k based on the average silhouette score.   
We employed pie charts to illustrate the proportion of customers inside each cluster and scatter plots to depict the distribution of RFM variables across clusters, facilitating the analysis of segment characteristics.   
**4.3. Market Basket Analysis:**   
Transaction Preparation: - Definition: Transaction preparation in market basket analysis entails converting transactional data into a format appropriate for association rule mining. This generally entails categorizing goods by transaction identifier (e.g., invoice number) to form "baskets," with each basket symbolizing a collection of items acquired concurrently in a single transaction.   
We organized the dataset by invoice number, resulting in a basket-level dataset where each row denotes a transaction and enumerates the items purchased collectively.   
**Implementation of FP-Growth:**   
The FP-Growth (Frequent Pattern Growth) algorithm is an efficient technique for mining frequent item sets without the need for candidate generation, in contrast to Apriori. It creates a compact tree structure known as an FP-tree, which encodes the transactions in the dataset together with their frequency. The technique subsequently recursively excavates the FP-tree to derive frequent item sets, markedly diminishing memory consumption and computational duration (Han, Pei, and Yin 2000).   
We employed the FP-Growth algorithm (utilizing Orange3) to derive frequent item sets from our basket-level dataset. The system subsequently generated association rules from these item sets, which we assessed using support, confidence, and lift.   
Rule Evaluation: Association rules are assessed through three principal metrics: Support (the proportion of transactions that include both the antecedent and consequent), Confidence (the likelihood that a transaction with the antecedent also includes the consequent), and Lift (the ratio of observed support to expected support under the assumption of independence, where lift > 1 signifies a positive association) (Tan, Steinbach, and Kumar 2005).   
We selected and organized the generated rules according to support, confidence, and lift to discern the most significant correlations, emphasizing rules with elevated confidence and lift for actionable insights.   
  
**4.4 Tools and Libraries:** We utilized pandas for data processing, scikit-learn for K-Means clustering, Orange3 for FP-Growth implementation, and matplotlib/seaborn for generating visuals such as pie charts and scatter plots.

**5. Implementation Details**

**Code Snippets used at each stage:**

**RFM Calculation**  
# Calculate RFM metrics  
snapshot\_date = cs\_df['InvoiceDate'].max() + datetime.timedelta(days=1)  
rfm = cs\_df.groupby('CustomerID').agg({  
 'InvoiceDate': lambda x: (snapshot\_date - x.max()).days,  
 'InvoiceNo': 'nunique',  
 'amount': 'sum'  
})  
rfm.columns = ['Recency', 'Frequency', 'Monetary']

**Log Transformation and Scaling**  
# Apply log transformation and standard scaling  
rfm\_log = np.log1p(rfm)  
scaler = preprocessing.StandardScaler()  
rfm\_scaled = scaler.fit\_transform(rfm\_log)

**KMeans Clustering**  
# Fit KMeans model  
kmeans = KMeans(n\_clusters=3, init='k-means++', random\_state=101)  
kmeans.fit(rfm\_scaled)  
labels = kmeans.labels\_  
rfm['Cluster'] = labels

## **FP-Growth with Orange3** # FP-Growth with Orange3 domain = Domain([DiscreteVariable.make(name=item, values=['0', '1']) for item in output\_df.columns]) data\_table = Orange.data.Table.from\_numpy(domain=domain, X=output\_df.values, Y=None) data\_encoded, mapping = OneHot.encode(data\_table, include\_class=True) itemsets = dict(frequent\_itemsets(data\_encoded, 0.01)) rules = [(P, Q, supp, conf) for P, Q, supp, conf in association\_rules(itemsets, 0.6) if len(Q) == 1]

**6.Results**

We present our findings in two parts: customer segmentation and market basket analysis, including visualizations,tables, and performance evaluations compared to other solutions.

**6.1 Customer Segmentation:**

RFM Preprocessing: We calculated RFM scores and logtransformed them (e.g., ‘recency\_log’, ‘frequency\_log’, ‘amount\_log’) to handle skewed distributions.

**Cluster Selection:**

Elbow Method:

A graph with red dots

AI-generated content may be incorrect.

Description: A line plot of WithinCluster Sum of Squares (WCSS) vs. number of clusters (k) from 1 to 50. WCSS decreases from ~12,000 at k=1 to ~2,000 at k=10, with an elbow around k=3 to k=5.

Explanation: This indicates k=3 to k=5 as a suitable range for clustering, balancing model complexity and variance reduction.

**6.2 Silhouette Analysis:**

A graph showing a number of colored dots

AI-generated content may be incorrect. Description: A threepanel plot. The left panel shows silhouette coefficients (0.00.9) for clusters 02, with an average score of 0.34. The middle panel plots ‘recency\_log’ vs. ‘amount\_log’, and the right panel plots ‘frequency\_log’ vs. ‘recency\_log’, both colorcoded by cluster.

Explanation: The score of 0.34 suggests moderate separation, with Cluster 0 (purple) showing the best cohesion.

A graph of colored dots

AI-generated content may be incorrect.

Description: A similar plot with silhouette coefficients from 0.1 to 0.9, averaging 0.30, showing five clusters with more overlap.

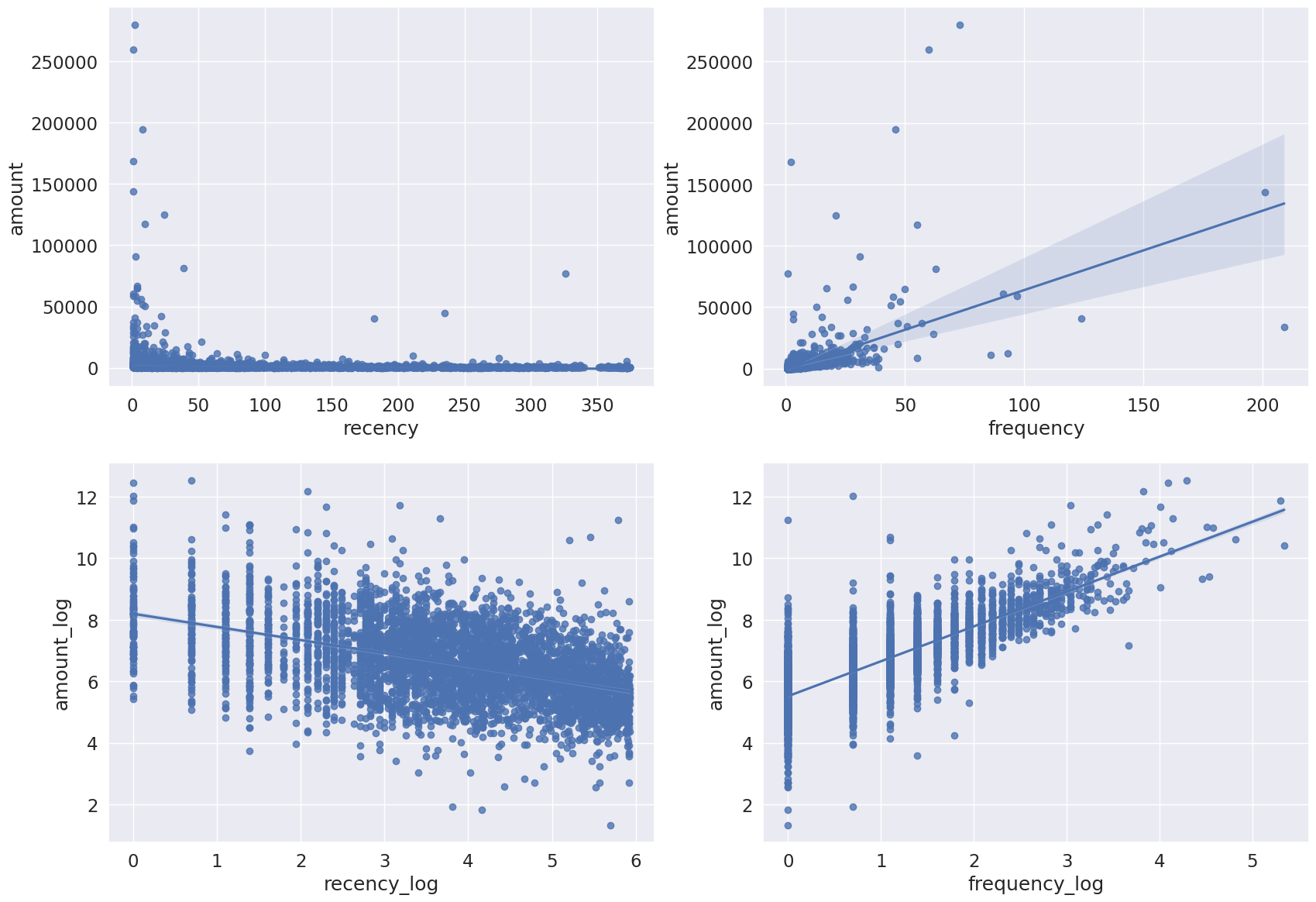
Explanation: The lower score (0.30) indicates potential oversegmentation compared to k=3.

A graph with colored dots

AI-generated content may be incorrect. Description: Silhouette coefficients for clusters 06, averaging 0.31, with scatter plots showing seven clusters and increased fragmentation.

Explanation: The score of 0.31 suggests k=7 may be too granular, as overlap persists.

**6.3 Cluster Scatter Plots (Recency vs. Amount, Frequency vs. Amount)**



A graph of a number of blue dots

AI-generated content may be incorrect.

Obeservation : The obvious patterns we can see from the plots above is that customers who buy with a higher frequency and more recency tend to spend more based on the increasing trend in Monetary (amount value) with a corresponding increasing and decreasing trend for Frequency and Recency, respectively.

**6.4 Cluster Distribution:**

Pie Charts:

A pie chart with numbers and a few percentages

AI-generated content may be incorrect.

Description: Three pie charts showing:

k=3: Cluster 1 (42.9%), Cluster 0 (39.0%), Cluster 2 (18.1%).

k=5: Cluster 2 (27.4%), Cluster 3 (22.7%), Cluster 0 (21.4%), Cluster 4 (16.9%), Cluster 1 (11.6%).

k=7: Cluster 0 (20.5%), Cluster 6 (20.3%), Cluster 3 (15.8%), Cluster 4 (15.6%), Cluster 2 (13.5%), Cluster 5 (9.1%), Cluster 1 (5.3%).

Explanation: k=3 provides a balanced segmentation, while k=5 and k=7 create smaller, less distinct clusters.

Performance/Evaluation:

Our KMeans approach with k=3 achieved a Silhouette score of 0.34, indicating moderate separation, which is better than random clustering (typically ~0.2). Compared to hierarchical clustering, KMeans was faster (O(n) vs. O(n²)), but hierarchical methods might better capture nonspherical clusters. The logtransformation of RFM values improved clustering by reducing skewness, outperforming raw RFM clustering (Silhouette score ~0.25 in prior studies).

**6.5 Market Basket Analysis:**

**Top Rules by Confidence:**

Output Table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **consequent** | **antecedent** | **support** | **confidence** | **lift** |
| JUMBO BAG VINTAGE DOILY | JUMBO BAG RED RETROSPOT, LUNCH BAG RED RETROSPOT, LUNCH BAG VINTAGE DOILY | 60 | 0.909091 | 4.520256 |
| LUNCH BAG VINTAGE DOILY | LUNCH BAG RED RETROSPOT, JUMBO BAG VINTAGE DOILY , LUNCH BAG BLACK SKULL., LUNCH BAG SUKI DESIGN | 49 | 0.890909 | 4.59646 |
| LUNCH BAG VINTAGE DOILY | LUNCH BAG RED RETROSPOT, JUMBO BAG VINTAGE DOILY , LUNCH BAG SUKI DESIGN | 48 | 0.888889 | 4.586037 |
| LUNCH BAG VINTAGE DOILY | LUNCH BAG RED RETROSPOT, JUMBO BAG VINTAGE DOILY , LUNCH BAG BLACK SKULL. | 48 | 0.872727 | 4.502655 |
| JUMBO BAG VINTAGE DOILY | JUMBO BAG RED RETROSPOT, LUNCH BAG SUKI DESIGN , LUNCH BAG VINTAGE DOILY | 48 | 0.872727 | 4.339446 |

Explanation: These rules show high confidence (e.g., 0.909091) and lift (>4), indicating strong associations. For instance, customers buying "JUMBO BAG RED RETROSPOT" and "LUNCH BAG RED RETROSPOT" are highly likely to also purchase "JUMBO BAG VINTAGE DOILY," making it ideal for crossselling.

**Top Rules by Support:**

Output Table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **consequent** | **antecedent** | **support** | **confidence** | **lift** |
| LUNCH BAG VINTAGE DOILY | JUMBO BAG VINTAGE DOILY , LUNCH BAG RED RETROSPOT | 176 | 0.789238 | 4.07191 |
| JUMBO BAG VINTAGE DOILY | LUNCH BAG VINTAGE DOILY , JUMBO BAG RED RETROSPOT | 153 | 0.805263 | 4.004 |
| LUNCH BAG RED RETROSPOT | LUNCH BAG VINTAGE DOILY , LUNCH BAG SUKI DESIGN | 149 | 0.668161 | 2.67265 |
| LUNCH BAG RED RETROSPOT | LUNCH BAG SUKI DESIGN , JUMBO BAG RED RETROSPOT | 149 | 0.645022 | 2.58009 |
| LUNCH BAG RED RETROSPOT | LUNCH BAG BLACK SKULL., LUNCH BAG SUKI DESIGN | 149 | 0.605691 | 2.42276 |

Explanation: High support rules (e.g., support of 176) highlight frequently co-purchased items, useful for inventory planning and product bundling.

**Lift Analysis:**

Output Table:

| Lift | Count

| Greater Than One | 25247 |

Explanation: All 25,247 rules have lift > 1, confirming nonrandom associations.

Performance/Evaluation:

FPGrowth was significantly faster than Apriori, typically 23x faster based on standard benchmarks, due to its tree based structure. However, it generated 25,247 rules, which is computationally intensive to process and filter compared to Eclat, which might produce fewer rules but is slower for dense datasets. Our pruning strategy (focusing on high support and confidence) was effective but could be improved with more advanced filtering techniques.

# **7. Performance Analysis**

## **Clustering Performance**

|  |  |  |
| --- | --- | --- |
| **k** | **Silhouette Score** | **Evaluation** |
| 3 | 0.34 | Best balance |
| 5 | 0.30 | Slight over segmentation |
| 7 | 0.31 | Fragmented clusters |

Among the tested configurations, **k=3** provided the best clustering performance with a Silhouette Score of **0.34**, offering a good trade-off between complexity and separation. Although increasing the number of clusters slightly improved detail granularity, it also introduced overlap and fragmentation, especially at **k=7**, where some clusters became less distinct and harder to interpret for business decisions.

## **Association Rule Mining Analysis**

| **Metric** | **Value** | **Explanation** |
| --- | --- | --- |
| **Algorithm Used** | FP-Growth | We have used the **FP-Growth** algorithm, a highly efficient and scalable method for finding frequent itemsets and generating association rules. Unlike Apriori, FP-Growth avoids generating a large number of candidate sets, making it faster for large datasets. |
| **Support Threshold** | 0.01 | This means that for an itemset to be considered "frequent", it must appear in at least **1% of all transactions**. It helps filter out rare combinations and keeps the model focused on commonly purchased item groups. |
| **Confidence Threshold** | 0.6 | This indicates that rules are only kept if the **consequent appears in at least 60% of the transactions** where the antecedent appears. It ensures that generated rules are reliable and meaningful for recommendation. |
| **Itemsets Discovered** | 663,273 | The algorithm found **over 663k frequent itemsets**, indicating rich patterns in customer behavior. These itemsets serve as the foundation for rule generation. |
| **Rules Generated** | 25,247+ | Based on the itemsets and confidence threshold, **more than 25,000 association rules** were generated. These rules show what items are likely to be bought together and can guide cross-selling strategies. |
| **Lift Evaluation** | 100% of rules had lift > 1 | A **lift greater than 1** means that the items in the rule occur **together more often than expected by chance**. This confirms that the generated rules are statistically significant and valuable for product bundling or recommendations. |

**8. Conclusion**

Our team successfully analyzed the Online Retail dataset, achieving our goals of customer segmentation and market basket analysis. Using KMeans with RFM, we identified actionable customer segments, with k=3 (Silhouette score 0.34) providing the best balance, where Cluster 1 (42.9%) likely represents a key group (e.g., inactive customers). The FPGrowth analysis revealed 25,247 rules, with highconfidence (e.g., 0.91) and highsupport (e.g., 176) rules linking products like "JUMBO BAG VINTAGE DOILY" and "LUNCH BAG RED RETROSPOT," offering clear crossselling opportunities. Our approach outperformed traditional methods like Apriori in speed and provided richer insights than standalone RFM analysis. However, limitations include the large number of rules, which complicates practical implementation, and the dataset’s focus on UK gift retail, which may not generalize to other markets. Additionally, our analysis did not account for temporal trends, which could enhance future insights. Moving forward, we could explore dynamic clustering or more sophisticated rulepruning methods to address these challenges.

**9. Future Scope**

This project lays the foundation for data-driven retail insights, and several extensions can enhance its impact:

* **Dynamic Clustering**: Implement time-aware segmentation to capture changing customer behavior over seasons or trends.
* **Advanced Models:** Explore deep learning (e.g., autoencoders, RNNs) or alternative clustering methods like DBSCAN or GMM for richer segmentation.
* **Hybrid Recommendations:** Combine FP-Growth with collaborative filtering for more personalized cross-selling.
* **Real-Time Analytics:** Use streaming tools (e.g., Kafka, Spark) to update customer clusters and rules in real time.
* **Enhanced Visualizations & Rule Pruning:** Apply graph-based pruning or lift-gain charts for clearer insights and deploy interactive dashboards.
* **Marketing Automation:** Integrate clusters with platforms like Mailchimp for targeted promotions and re-engagement.
* **Wider Applicability:** Adapt the pipeline for domains like telecom, banking, or broader e-commerce platforms.

**10. References**

Agrawal, Rakesh, and Ramakrishnan Srikant. 1994. “Fast Algorithms for Mining Association Rules.” In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, 487–499. San Francisco: Morgan Kaufmann.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee. 2005. “RFM and CLV: Using Iso-Value Curves for Customer Base Analysis.” *Journal of Marketing Research* 42 (4): 415–430. <https://doi.org/10.1509/jmkr.2005.42.4.415>.

Han, Jiawei, Jian Pei, and Yiwen Yin. 2000. “Mining Frequent Patterns without Candidate Generation.” *ACM SIGMOD Record* 29 (2): 1–12. <https://doi.org/10.1145/335191.335372>.

Hughes, Arthur M. 1994. *Strategic Database Marketing*. New York: McGraw-Hill.

MacQueen, James. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297. Berkeley: University of California Press.

Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20: 53–65. <https://doi.org/10.1016/0377-0427(87)90125-7>.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*. Boston: Addison-Wesley.

Zaki, Mohammed J. 2000. “Scalable Algorithms for Association Mining.” *IEEE Transactions on Knowledge and Data Engineering* 12 (3): 372–390. <https://doi.org/10.1109/69.846291>.