



Enhancing Customer Segmentation in Retail through Machine Learning Models and Targeted Marketing Techniques

Master Thesis

Author: Jacobo Valderrama Rovira

Advisor: Cedric Verbeeck

Programme: Data Analytics & Artificial Intelligence

Date: 16/06/2025

EDHEC Business School does not express approval or disapproval concerning the opinions given in this paper which are the sole responsibility of the author.

I certify that: the Master Project being submitted for examination is my own research, the data and results presented are genuine and actually obtained by me during the conduct of the research and that I have properly acknowledged using parenthetical indications and a full bibliography all areas where I have drawn on the work, ideas and results of others and that the master project has not been presented to any other examination committee before and has not been published before.

Abstract

This thesis investigates customer segmentation with traditional marketing techniques and clustering algorithms, focused on RFM, RFMD, and CLV features. Clustering algorithms included K-Means, Hierarchical Clustering, DBSCAN, and Fuzzy C-Means which were evaluated and compared using metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Results show that K-Means offers a good balance between performance, simplicity, and efficiency and was able to identify 5 distinct customer segments. Notably, RFMD proved to be more effective than the traditional RFM model by adding greater segmentation clarity which made it easier when providing actionable recommendations to each segment. CLV was incorporated in the analysis to better understand customer value and guide resource allocation, despite some limitations due to missing parameters in the dataset it was helpful in providing recommendations. Overall, this research shows that combining machine learning with traditional marketing techniques can be beneficial in providing actionable customer insights.

Contents

1	Introduction	4
1.1	Background of the study	4
1.2	Problem Statement	4
1.3	Objectives	5
1.4	Research Questions	5
1.5	Thesis Structure Overview	6
2	Literature Review	6
2.1	Customer Segmentation in Marketing	6
2.2	Customer Lifetime Value (CLV)	7

2.3	Market Basket Analysis (MBA)	9
2.4	RFM and RFMD Models	9
2.5	Machine Learning for Segmentation	10
2.5.1	K-Means	10
2.5.2	Fuzzy C-Means	11
2.5.3	Hierarchical Clustering (HCA)	12
2.5.4	DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	13
2.6	Evaluation Metrics for Clustering	13
2.7	Gaps in Literature and Contribution	16
3	Methodology	17
3.1	Use of AI Assistance	17
3.2	Research Design	17
3.3	Dataset Description	17
3.3.1	Overview of the Dataset	17
3.3.2	Variable Descriptions	18
3.4	Segmentation Pipeline Overview	18
3.5	Data Preprocessing	18
3.6	RFM and RFMD Feature Construction	19
3.6.1	Exploratory Data Analysis	20
3.7	Applied Segmentation Techniques	22
3.8	Cluster Evaluation and Model Comparison	22
3.8.1	Stability and Validation of Clustering Results	23
3.9	Customer Lifetime Value Estimation	23
4	Results	24
4.1	RFM and RFMD Feature Construction	24
4.2	Applied Segmentation Techniques	31
4.3	Cluster Evaluation	34
4.3.1	Evaluation Metrics for Clustering Parameter Selection	34
4.3.2	Metrics Used to Compare Clustering Models	36
4.4	Model Comparison	38
4.4.1	K-Means Results (RFMD)	38
4.4.2	Hierarchical Clustering Results (RFMD)	40
4.4.3	Fuzzy C-Means Results (RFMD)	41
4.4.4	DBSCAN Results (RFMD)	42
4.5	CLV Distribution and Cluster Sizes by Segmentation Method	44

5 Discussion	46
5.1 Insights from RFM and RFMD Feature Distributions	46
5.2 Segmentation Outcomes	48
5.2.1 K-Means Clustering (RFM and RFMD)	48
5.2.2 Hierarchical Clustering (RFM and RFMD)	49
5.2.3 Fuzzy C-Means Clustering (RFM and RFMD)	50
5.2.4 DBSCAN Clustering (RFM and RFMD)	51
5.3 Model Evaluation and Comparison	51
5.4 Method Complementarity and Practical Insights	52
5.5 Customer Lifetime Value by Cluster	52
5.6 Limitations	55
6 Recommendations	55
7 Conclusion	57
7.1 Summary of Findings	57
7.2 Contributions to the Field	58
7.3 Future Work	58
8 Annexes	65

1 Introduction

1.1 Background of the study

Nowadays, with the sudden increase in competition and the emergence of new tools, businesses have to use different methods to attract new customers and at the same time keep the customers they already have loyal to them to improve revenue. The majority of companies around the world offer very similar products, which is why it is important to have a competitive advantage that can help them thrive. Customers right now are similar to the ones 100 years ago or even much further back, the only difference is the number of options they can choose from, which is why the term customer segmentation has become so popular, but what is customer segmentation? It's a tool that helps divide customers into smaller groups, the way it does it is by grouping these customers by some characteristics that are shared between them, this way targeting becomes much more personalized and in the most part the one-size-fits-all is forgotten.

Customer segmentation can use any type of data to group customers based on who these customers are, these include: Demographics, psychographics, and firmographics. The other possibility is to group customers based on what they do, this means to focus on how they behave with a product, how much do they spend, what do they spend the most money on, how long have they been considered customers, etc. Qualtrics (2020). Examples of companies using customer segmentation is one that is widely used by most of the people, Netflix used machine learning and artificial intelligence to analyze peoples behavior when they use the service to recommend new series and movies. Another example is Amazon, which analyzes the products most requested by a customer as well as the recently bought products to recommend new ones that the customer might like.

1.2 Problem Statement

Businesses that don't implement segmentation or at least use the more basic/traditional segmentation cannot make a profit given today's market conditions. The main reason for using dynamic customer segmentation is to limit and reduce loss of customers and therefore loss of revenue. One of the metrics that is heavily impacted with customer segmentation is Churn. It refers to the number of customers that stopped using a particular product/service during a specific time for any given reason Prabadevi et al. (2023), and churn rate is just the number of people who stopped using the given product/service compared to the numbers of total customers. Churns are very important because many studies have proven that it is less expensive to invest on current customers than to acquire new ones. According to Ali Cubdy (as cited in The Wharton

School (2022)), it costs six-to-seven times more to acquire new customers than it does to retain current ones. She notes that 5% increase retention can lead to high levels of profitability and potentially 96% increase in profits. The main goal of this study is to show how the different techniques can be used to use customer segmentation effectively and give appropriate recommendations.

1.3 Objectives

The main idea is to use marketing techniques and customer segmentation with and without machine learning models to group customers by different characteristics and find the most efficient methods. At the end is to give recommendations.

1. Apply Machine Learning techniques (K-Means, Hierarchical Clustering, DBSCAN) for customer segmentation based on RFM, RFMD and CLV features.
2. Evaluate and compare the performance of the different clustering algorithms using the appropriate evaluation metrics (Silhouette Score, Davies-Bouldin Index, etc.).
3. Interpret the characteristics of the customer segments found.
4. Map marketing actions and recommendations (such as loyalty programs, win-back campaigns, and upselling) to each identified segment.
5. Address data challenges by developing an approach to handle unknown customers and cancelled orders effectively during the segmentation process.

1.4 Research Questions

1. Which machine learning algorithms (K-Means, Hierarchical Clustering, GMM, DBSCAN) provide the most meaningful customer segments based on RFM, RFMD, and CLV features?
2. How do the identified customer segments differ in terms of purchasing behavior, engagement, and value?
3. What marketing strategies can be developed based on the characteristics of each customer segment to improve customer retention, loyalty, and cross-selling opportunities?
4. How should unknown customers and canceled orders be treated during segmentation to ensure meaningful and actionable insights?

1.5 Thesis Structure Overview

This thesis is organized into seven chapters:

- **Chapter 1 – Introduction:** Presents the background, problem statement, research objectives, research questions, scope, and structure of the thesis.
- **Chapter 2 – Literature Review:** Reviews existing work on customer segmentation methods, including traditional approaches, machine learning techniques, Customer Lifetime Value (CLV), Market Basket Analysis (MBA), evaluation metrics, and strategic mapping. Gaps in the literature and the contribution of this work are also discussed.
- **Chapter 3 – Methodology:** Describes the research design, dataset characteristics, data preprocessing steps, applied segmentation techniques, cluster evaluation, and the marketing strategy mapping process.
- **Chapter 4 – Results:** Provides a detailed analysis of the customer dataset, presents the segmentation results, evaluates the performance of different models, and shares insights from the Market Basket Analysis.
- **Chapter 5 – Discussion:** Interprets the segmentation outcomes, proposes strategic marketing actions based on the findings, and addresses the study's limitations.
- **Chapter 6 – Conclusion:** Summarizes the main findings, outlines the contributions of the research, and suggests directions for future work.
- **Chapter 7 – Self-Reflection:** Reflects on the technical and personal learning experience throughout the research project.

2 Literature Review

2.1 Customer Segmentation in Marketing

Globalization is commonly related to something that has been happening at least for the past 100 years but in reality it traces to more than 2000 years ago. In fact the origins from the term itself had an origin in the first century BC when luxury products produced in China started to appear in other parts of the world. Later on spice routes were introduced thanks to Islamic merchants and the list goes on until reaching today. Trading of goods in different places and to different people has been happening even before it was called trading and it something that will continue on foreverVanham

(2019). Human evolution has made the human being more capable of learning new things, using new technologies and along with that new behaviors. Modern business are completely different from what it used to be and from what will be, the emergence of new technologies and desires has lead to people and business using new technologies to keep customer satisfied. Using traditional market segmentation also known as mass marketing its not being used anymore because focusing on particular segments its more profitableAlves Gomes and Meisen (2023).

According to Cambridge University Press (n.d.), mass marketing can be described as the use of marketing to target a very large number of people. It basically works with the one-size-fits-all approach different from the niche marketing in which marketing campaigns are designed for specific people based on common characteristics. Not because niche is better at targeting specific people it means mass marketing is not used anymore, on the contrary each has a particular use. Marketing is commonly found in billboards, print ads and television ads and its role is to increase brand awareness and reduce costsMasterClass (2021).

On the other hand, customer segmentation focuses on classifying customers into specific groups that share characteristics. In other words its just refining the general campaigns for each group, this is based on using elements that interest everyone. Its an approach used in both B2B and B2C with some small variations for each, for example in the case of B2C its based on demographics and needs while in B2B more on products purchased in the past and industry. To group customers is important to have the appropriate data which is why its important to recollect different kinds of information from customers. Among the most common segmentation grouping methods include geographic segmentation, demographic segmentation, psychographic segmentation, behavioral segmentation, technographic segmentation, needs-based segmentation, values-based segmentation, etc. This are only the most important categorization techniques but in reality this can be done with almost any type of data ultimately depending on business needsSurveyMonkey Inc. (2024).

2.2 Customer Lifetime Value (CLV)

Implementing techniques to improve customer loyalty is in every business interest, and therefore it is important to know which tactics to use. Customer Lifetime Value(CLV) helps companies understand their customers from a monetary standpoint. It repre-

sents the amount of money spend by a customer from their first to their last purchase. According to Wharton marketing professor David Reibstein(as cited in The Wharton School (2022)), the probability to sell to existing customers is fourteen times greater than to a new customer. Customer lifetime Value is calculated with average customer value(average value spent by the customer) times the average customer lifespan.

A research conducted by Siti Monalisa (2019) examined LWC Company, one of the biggest pain distributors in Indonesia. The company works with both companies and customers, meaning it operates with B2B and B2C business models. In the study they analyzed the use of of marketing strategies across their customers and what were the means to assigning this and to which customers. They analyze their transactional data using the RFM method and generate clusters based on this information, after that they used the following formula to compute the value of CLV:

$$C_j = WR \cdot CR_j + WF \cdot CF_j + WM \cdot CM_j$$

Where each variable is represented below:

- C_j: CLV rating of a particular customer
- CR_j: Normalization of Recency(R) from the cluster
- CF_j: Normalization of Frequency(F) from the cluster
- CM_j: Normalization of Monetary(M) from the cluster
- WR : Weight associated to Recency(R)(In an scenario where variables are equally important it equals 0.33)
- WF : Weight associated to Frequency(F)(In an scenario where variables are equally important it equals 0.33)
- WM : Weight associated to Monetary(M)(In an scenario where variables are equally important it equals 0.33)

They consider the values of RFM as well as a weight associated to each variable which is determined by the importance of each variable, it ultimately depends what variable needs to be analyzed further. In the end, customers with high recency, high frequency and high monetary value are those with the highest CLV.

Building on this, another study by Jasek et al. (2018) aimed to use primary empirical research of CLV and then compare it with articles that also use CLV based on

real data. They used RFM to compute the variable different from other theoretical studies where they used more complex formulas and then MAE to compare the different results. The study was based on several medium to large online retail stores from Czech republic and slovakia including information from at least two years old and one thousand unique customers. After comparing the values for MAE and sensitivity of each result the research concluded by giving some options on how to use the CLV. This was divided in an individual customer level, by clusters of customers and customer base levels.

2.3 Market Basket Analysis (MBA)

Market Basket Analysis is a widely used method to discover customer purchasing patterns. It relies on the idea of analyzing customers' purchase histories to identify which products are more likely to be bought together. Currently, MBA includes two types: Predictive Market Basket Analysis and Differential Market Basket Analysis. Both focus on grouping items and cross-selling opportunities, but the differential type also analyzes purchases across different stores TechTarget Contributor ([2023](#)).

Hoque et al. ([2024](#)) conducted a study using Kaggle datasets to analyze transactions from grocery and retail stores. They combined marketing techniques like RFM and K-Means clustering to study customer behavior patterns and then applied MBA to complement their analysis by identifying product groups that triggered additional purchases. For the grocery dataset, they found that purchasing almost any product led to buying whole milk. In the retail dataset, almost all products led to the purchase of a completely different product, showing MBA's precision when combined with K-Means.

MBA is also valuable for building recommendation models in marketing environments. For example, Maraghi et al. ([2020](#)) analyzed CRM strategies for supermarkets to identify the most profitable customers and their product demand patterns. They segmented customers, assigned categories, and implemented tailored marketing plans based on their needs. Similarly, Paranalavithana et al. ([2021](#)) applied segmentation and MBA on a UK e-commerce dataset. They used a three-level segmentation (high-, medium-, and low-value customers) and showed how MBA can improve customer loyalty and optimize stock management.

2.4 RFM and RFMD Models

RFM remains one of the most widely used customer segmentation tools due to its simplicity and reliance on transactional data. The acronym stands for Recency (how

recently a customer made a purchase), Frequency (how often), and Monetary (how much they spend) Stormi et al. (2020).

Introduced over 30 years ago, RFM has evolved and been combined with other methods to become more dynamic. Despite its strengths, RFM does not incorporate demographic or other complementary data, which can limit insights. Customers are grouped by scoring each RFM dimension on a scale—such as 1 to 3 (yielding 27 segments) or 1 to 5 (yielding 125 segments)—depending on business needs. Recent efforts to enhance RFM include integration with time series analysis, recommendation systems, clustering algorithms like K-Means or C-Means, and considering product variety, leading to extensions like RFMD or RFMV models.

For example, Wulansari and Heikal (2024) analyzed Indonesia’s top three e-commerce sites (Shopee, Tokopedia, Lazada) using RFM. Customers were segmented into High, Medium, and Low groups based on RFM scores, focusing on those with uniform scores across all three dimensions (e.g., high-high-high). They identified six distinct clusters with varying behaviors across platforms, highlighting the need for targeted strategies.

In a UK-based e-commerce study, Christy et al. (2021) categorized customers into five levels based on RFM scores, where 5 indicates the best customers with high recency, frequency, and monetary values, and 1 represents lost or soon-to-churn customers. Machine learning algorithms further grouped similar customers, and while results were promising, the study recommended integrating product-level data and additional ML models to improve recommendations.

2.5 Machine Learning for Segmentation

This section reviews the most frequently used unsupervised clustering algorithms, focusing on the state-of-the-art techniques and the metrics commonly used to evaluate their performance.

2.5.1 K-Means

K-Means is among the most popular clustering methods mentioned in the literature. According to IBM (n.d.), K-Means is an unsupervised machine learning algorithm used to group data points based on similarity. Each data point belongs to exactly one cluster. The algorithm iteratively minimizes the sum of squared distances between data points and their assigned cluster centroids.

The process starts by choosing the number of clusters, K , which directly affects the granularity of the grouping: a larger K results in more clusters with more similar data points, while a smaller K yields fewer, broader clusters. Once K centroids are randomly initialized, each data point is assigned to its nearest centroid based on Euclidean distance. The centroids are then recalculated as the mean of their assigned points. This assignment-update cycle repeats until the centroids stabilize and clusters satisfy two key conditions: points within a cluster are similar, and clusters are distinct from one another Piech (n.d.). K-Means is widely used in diverse sectors, including retail and healthcare.

Abdullah et al. (2022) applied K-Means to cluster Indonesian provinces based on COVID-19 confirmed, death, and recovered cases in 2019, supporting targeted policy implementation. Cluster validity was ensured using silhouette, elbow, and gap statistics, resulting in three meaningful clusters. Janardhanan and Muthalagu (2020) studied 45 stores with 81 products each to identify profitable products and customer purchase behavior, enhancing market profit. Similarly, Xiahou and Harada (2022) used K-Means to segment customers of an e-commerce site into loyal and churn-risk groups, aiming to develop retention strategies.

2.5.2 Fuzzy C-Means

While K-Means assigns each data point to exactly one cluster, Fuzzy C-Means (introduced in the 1980s alongside fuzzy set theory by Ruspini) allows points to belong to multiple clusters with varying degrees of membership. This approach better captures overlapping clusters by associating membership probabilities rather than strict assignments—for example, a point could be 85% in cluster 1 and 15% in cluster 2 Ghosh and Dubey (2013).

Uddin et al. (2024) demonstrated the importance of understanding digital native customers by applying Fuzzy C-Means to a Kaggle dataset of students from 2006-2009. They formed four clusters aligned with predominant topic interests. In another study, Idowu et al. (2019) combined RFM analysis with Fuzzy C-Means to classify customers into five groups:

1. High Spending customers
2. Average Spending customers
3. Less Spending, Less Recent customers
4. Less Spending, Recent customers

5. Least Spending, Recent customers

For comparison, K-Means clusters were:

1. High Spending customers
2. Less Spending, Recent customers
3. Average Spending, Recent customers
4. Less Spending, Less Recent customers
5. Average Spending, Less Recent customers

Fuzzy C-Means is advantageous for capturing customers' overlapping behaviors, enabling businesses to apply similar or hybrid strategies to different segments.

2.5.3 Hierarchical Clustering (HCA)

Hierarchical clustering is another unsupervised method, commonly used in biology, image analysis, and social sciences. Like K-Means, it clusters based on distances but uses linkage criteria to form a hierarchy of clusters, often visualized via dendograms Noble (2024). The main linkage methods are:

- **Single linkage:** minimum distance between points of two clusters
- **Complete linkage:** maximum distance between points of two clusters
- **Average linkage:** average distance between points of two clusters

HCA has two approaches:

- **Bottom-up (agglomerative):** starts with individual points and merges clusters until all data form one cluster.
- **Top-down (divisive):** starts with one cluster containing all data, splitting iteratively into smaller clusters.

Choice depends on the goal, but both are computationally intensive compared to other methods.

Abdulhafedh (2021) applied HCA to segment credit card customers, comparing it with K-Means. Using dendograms and average linkage, they identified three major clusters. While K-Means produced better overall results, HCA helped estimate the optimal number of clusters, which K-Means requires as input. Gomes and Meisen (2023) reviewed 105 customer segmentation studies in e-commerce, finding K-Means was predominant, whereas HCA appeared only in five studies. Although less efficient for large-scale data, HCA remains useful in less demanding environments where interpretability is key.

2.5.4 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that identifies clusters of closely packed points and labels points in low-density regions as noise or outliers GeeksforGeeks contributors (2025) and Yenigün (2024). Unlike K-Means or HCA, DBSCAN does not require specifying the number of clusters beforehand and can detect arbitrarily shaped clusters.

DBSCAN categorizes points as:

- **Core points:** points with at least MinPts neighbors within radius ϵ
- **Border points:** points near core points but with fewer neighbors
- **Noise points:** points not belonging to any cluster

The algorithm works as follows:

1. Identify core points by counting neighbors within ϵ .
2. For each unassigned core point, create a new cluster and recursively add all density-connected points.
3. Density connectivity ensures points are linked through chains of core points.
4. Label all remaining unassigned points as noise.

Kachroo (2023) applied DBSCAN to an e-commerce dataset, finding it effective at handling noise but challenging to choose the number of clusters. They proposed hybridizing with Fuzzy C-Means to improve cluster selection. Paramita and Hariguna (2024) compared K-Means and DBSCAN on a US e-commerce customer segmentation dataset. K-Means yielded more balanced clusters, while DBSCAN identified smaller scattered clusters and filtered irrelevant points. The authors concluded that algorithm choice depends on data characteristics: DBSCAN excels when customers appear highly similar and noise removal is critical.

2.6 Evaluation Metrics for Clustering

When it comes down to the evaluation of machine learning algorithms, it is important to select the right metrics. Some metrics are more compatible with certain algorithms than others. In the following part we will cover the most used clustering metrics and explain how they work.

Clustering metrics include:

1. **Silhouette Score:** Measures if clusters are well defined or in other words if data-points within them are part or not. It is measured from -1 to 1 where -1 means the point isn't part of the cluster, 0 means the point is between 2 clusters and 1 that the point is in the right cluster.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (1)$$

where:

- $a(i)$ is the average distance from point i to all other data points in the same cluster.
- $b(i)$ is the smallest average distance from point i to all points in a different cluster (i.e., the nearest cluster that i is not a part of).

2. **Davies-Bouldin Index(DBI):** It measures how close the data-points are within each other. For instance, low values are better because it means clusters are more compact meanwhile higher values mean clusters are not tight and data-points share clusters.

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{R_{ii} + R_{jj}}{R_{ij}} \right) \quad (2)$$

where:

- k is the total number of clusters,
- R_{ii} is the compactness (intra-cluster distance) of cluster i ,
- R_{jj} is the compactness of cluster j ,
- R_{ij} is the dissimilarity (inter-cluster distance) between cluster i and cluster j .

3. **Calinski-Harabasz Index (Variance Ratio Criterion):** Measures tightness of clusters and cluster separation. The higher the value means data-points within clusters are close to each other and far away from other clusters.

Calinski–Harabasz Index (CH Index)

The Calinski–Harabasz index is defined as:

$$CH = \frac{B}{W} \times \frac{K-1}{N-K} \quad (3)$$

where:

- B is the between-cluster sum of squares,
- W is the within-cluster sum of squares,
- N is the total number of data points,
- K is the number of clusters.

Between-group Sum of Squares (B)

$$B = \sum_{k=1}^K n_k \cdot \|C_k - C\|^2 \quad (4)$$

where:

- n_k is the number of observations in cluster k ,
- C_k is the centroid of cluster k ,
- C is the centroid of the entire dataset.

Within-group Sum of Squares (W)

$$W = \sum_{k=1}^K \sum_{i=1}^{n_k} \|X_i^k - C_k\|^2 \quad (5)$$

where:

- X_i^k is the i -th observation in cluster k ,
- C_k is the centroid of cluster k ,
- n_k is the number of observations in cluster k .

GeeksforGeeks ([n.d.](#))

On the other hand there are some visualization techniques such as Dendograms and the elbow method that help to choose the appropriate number of clusters needed for a dataset.

1. **Elbow method:** Visual representation in which the WCSS (Within-Cluster Sum of Squares) for each value of k is computed and then plotted. The place in the graph where the slope changes rapidly will be the optimal number of clusters GeeksforGeeks ([2025a](#)).
2. **Dendrogram:** It shows how individual data-points or clusters merge together:

In the study by Paramita and Hariguna ([2024](#)) they compared and analyzed the use of clustering techniques such as K-Means and DBSCAN in an e-commerce context.

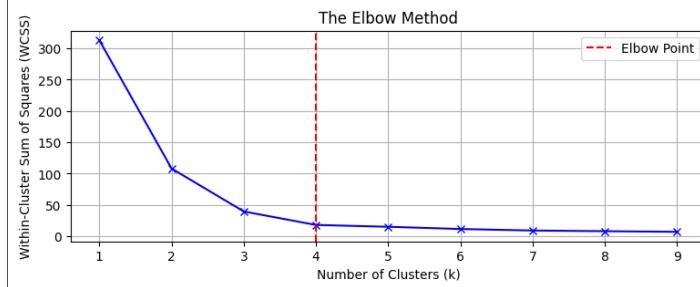


Figure 1: Illustration of the Elbow Method in K-Means Clustering GeeksforGeeks (2025a)

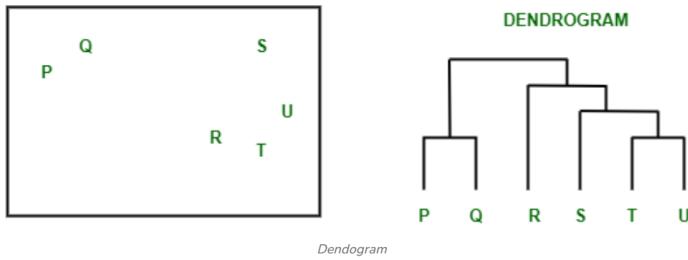


Figure 2: Dendrogram illustrating Hierarchical Clustering GeeksforGeeks (2025b)

Due to the contrasting ways in which each method operates, they used the silhouette score, Davies-Bouldin index and Calinski-Harabasz score to determine which of the two methods performed better. In the end each metric has similar conclusions but none are exclusively important. The use of each depends on the task; in this article K-Means presented better results overall but DBSCAN had an alarmingly high value of Calinski-Harabasz score which can be translated to better separation of clusters which corresponds due to its high capacity of separating actual points from noise.

2.7 Gaps in Literature and Contribution

Despite the continuous growth in the world due to globalization, businesses have to innovate everyday to dominate in today's market. Due to this, customer segmentation has become something everyone has to do to survive. Even though technology has been growing so much, several gaps remain in literature. The majority of the studies evaluate the use of individual clustering techniques, specially K-Means, but only some compare it with other techniques, this makes it difficult to determine which technique proves more useful in certain contexts.

Moreover, much of the current studies revolve around segmentation using RFM(Recency, Frequency, Monetary), but few fail to include much newer methods such as RFMD(Recency, Frequency, Monetary, Diversity), additionally only some include things like Customer

Lifetime Value analysis and even more Market Basket Analysis to get proper results. Additionally, the studies only include the analysis of the clustering techniques but fail to provide further analysis and recommendations. To address these issues, this study will focus on comparing different clustering techniques with distinct metrics to compare their results. It will also include things like RFM, RFMD, CLV and MBA and will finish by providing general recommendations to the most important segments found.

3 Methodology

3.1 Use of AI Assistance

A significant part of the code used to generate the results in thesis was developed with the assistance of ChatGPT. While I wrote the majority of the code, ChatGPT helped me troubleshoot many problems throughout the process. For Fuzzy C-Means and DBSCAN clustering, ChatGPT provided the full function. By the end of the project. ChatGPT was used to improve readability and the overall efficiency by organizing the workflow better, and suggest optimizations. Some of the prompts used with ChatGPT during the development of this project can be found in the Appendix section.

3.2 Research Design

This research will follow a quantitative approach focused on analyzing customer segmentation in the retail industry and then giving recommendations. The idea is to use customers' purchasing behavior and marketing techniques like RFM and RFMD and machine learning algorithms to complement the clustering process.

By the end of the project, the goal is to have chosen a machine learning model that is suited for the task considering the requirements, and to provide recommendations that go along with the results of each cluster.

3.3 Dataset Description

3.3.1 Overview of the Dataset

The dataset used in this study comes from the UCI Machine Learning Repository, as described in D. Chen (2015). It is based on a UK-registered, non-store online retail business. The data contains transactional records occurring between 01/12/2010 and 09/12/2011. The company deals in various types of products and transactions. The dataset consists of 541,909 instances accompanied by 8 features.

3.3.2 Variable Descriptions

The following variable descriptions are adapted from D. Chen (2015):

- **InvoiceNo:** A 6-digit integer uniquely assigned to each transaction. If the code starts with the letter 'C', it indicates a cancellation.
- **StockCode:** A 5-digit integer uniquely assigned to each distinct product.
- **Description:** A brief description of the product.
- **Quantity:** The quantity of each product (item) per transaction.
- **InvoiceDate:** The date and time when each transaction was generated.
- **UnitPrice:** The price per unit of product, in pounds sterling.
- **CustomerID:** A 5-digit integer uniquely assigned to each customer.
- **Country:** The name of the country where each customer resides.

3.4 Segmentation Pipeline Overview

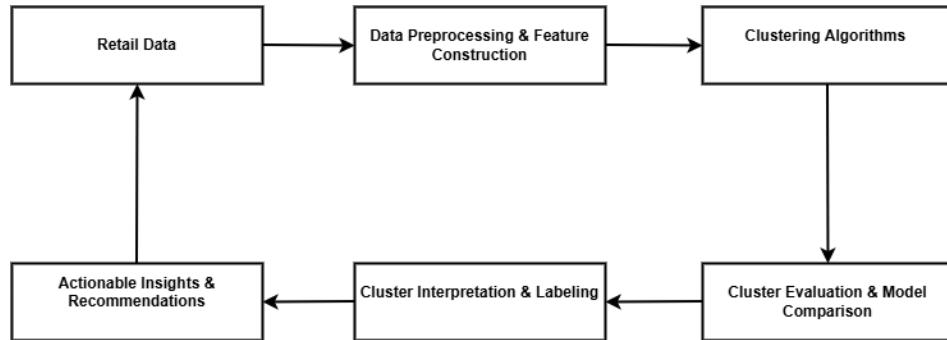


Figure 3: Overview of the customer segmentation pipeline, illustrating the key stages from data Preprocessing and feature construction through clustering, evaluation, model comparison, and actionable insights.

3.5 Data Preprocessing

In the first step, after trying to import the dataset with pandas we realized that it was taking too much time and for that reason we used the "Polars" library to import the dataset and then convert that to pandas to facilitate operations.

In the next step, we deleted the column "`__UNNAMED_8`" which got created when using the polars library and then converting it back to a pandas data-frame. Next step

was to delete all records whose `InvoiceNo` was null, this was done because the purpose of the study is to use segmentation techniques and then give recommendations and it doesn't make sense to take into consideration invalid transactions.

Next was to address the issues of null or invalid descriptions for some products. The first step was identifying the most frequent description with the same `StockCode`, then for every row with a missing description, we selected the most common description associated with the same `StockCode`. After checking the data-frame for more null values, we decided to remove other null descriptions because they didn't have a repeating `StockCode` associated with them.

When checking the new dataframe after the previous cleaning steps, the only column with missing values was `CustomerID`, with 134,578 entries, which corresponds to 24.83% of the data. Since marketing techniques such as RFM or RFMD require identifiable customers, the presence of missing `CustomerIDs` makes it impossible to track individual behavior over time. For this reason, all records without a valid `CustomerID` were excluded from the analysis. Nevertheless, we created and assigned artificial identifiers (e.g., `UNK+id`), which can be used if there is a need to analyze purchasing patterns or implement techniques such as Market Basket Analysis in future studies.

In the next part some special cases of non-useful descriptions were considered. We made a list with the patterns found in the descriptions that didn't help with the segmentation task such as: '?', '??', '???' , '?missing', '???missing', '?sold as sets?', '??missing', '??', '???lost', '????damages????', '????missing'. Then we disposed the rows where these patterns appeared.

In the next part we added four new columns to the dataset. These included one for the total price which was computed by multiplying the values of the columns "quantity" and "unit price". The other three columns refer to the date decomposed by year, month and day. This were included to analyze possible seasonal trends.

3.6 RFM and RFMD Feature Construction

RFM basetable was created using 2011-12-09 12:50:00 as the reference date. Recency corresponds to the number of days since the customers' most recent purchase, frequency is the number of unique transactions made by the customer, and Monetary is the total value spent by that customer (sum of all purchases). To include another version of RFM which is RFMD, the variable D (Diversity) was added by counting the number

of unique products a single customer bought.

Table 1: Summary Statistics for RFMD Features

Feature	Mean	Median	Std
Recency	92.0	50.0	100.0
Frequency	4.0	2.0	8.0
Monetary	2054.0	674.0	8988.0
Diversity	61.0	35.0	85.0

3.6.1 Exploratory Data Analysis

It's important to mention that after importing the dataset and analyzing all the features, it was found that 134,102 values had unknown CustomerID, which corresponds to 24.79% of the data. In the upcoming sections it will be mentioned how these unknown customers were treated. In the following part, there will be information that will help analyze the problem further. To start off, four columns were added to the dataset: "Year", "Month", "Day", and "Total_Price". The idea behind this is to analyze seasonality in a simpler way and also the cost by customer. This was computed using the value for each product times the quantity of the products. A column **churn** was added which was computed by looking at the transactions made by customers and if they hadn't bought anything for the past 90 days they were considered as churn.

After analyzing the numeric variables it can be seen that there are 531,860 values. Regarding the quantity, on average customers order 10 units of any product and the lowest value is negative, which corresponds with a cancellation. In the case of the UnitPrice, products cost 3.4 pounds on average and the minimum cost is 0, which corresponds to the customer getting something for free, and the most expensive product is 8142 pounds. In the case of Total_Price, the average value spent by customers is 19.6 pounds, the minimum is 0, and the maximum a customer spent was 168,469.

	Quantity	UnitPrice	Total_Price
count	531860.000000	531860.000000	531860.000000
mean	10.250000	3.440000	19.600000
min	-9600.000000	0.000000	-0.000000
25%	1.000000	1.250000	3.750000
50%	3.000000	2.080000	9.900000
75%	10.000000	4.130000	17.700000
max	80995.000000	8142.750000	168469.600000

Table 2: Descriptive Statistics of the Dataset

The majority of the customers originate from United Kingdom, while significantly fewer come from other countries such as Germany and France (see Figure 4).

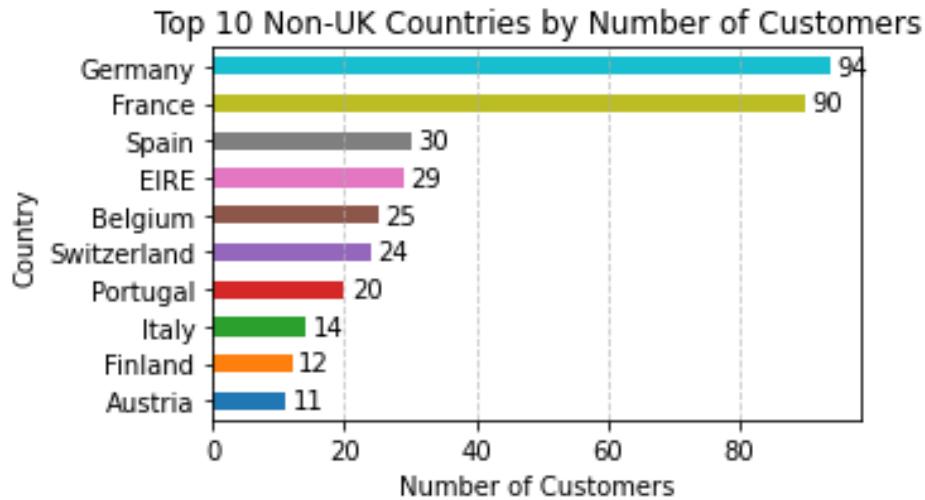


Figure 4: Top 10 countries with highest number of customers excluding UK

Another important metric to understand the behavior of customers in the dataset is to analyze from which countries the majority of unique invoices come from. Just like the number of customers by country, UK has the highest number of unique invoices, followed by Germany with significantly less and France (see Figure 5).

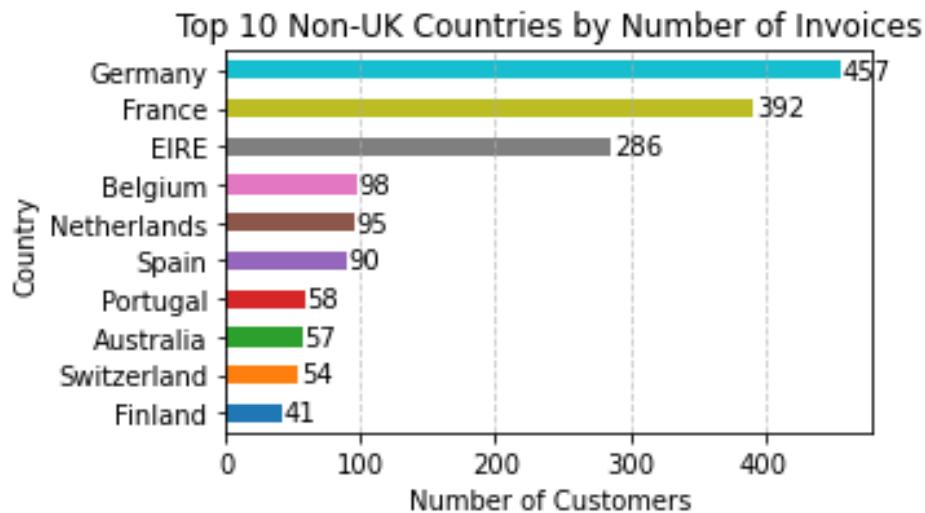


Figure 5: Top 10 countries with highest number of unique InvoiceNo excluding UK

3.7 Applied Segmentation Techniques

In terms of segmentation techniques, 4 algorithms were used, these are: K-Means, Fuzzy C-Means, Hierarchical using Ward, and DBSCAN. The idea behind choosing these techniques is to use the algorithms that have been most widely used in similar studies and then compare them to see which presents the most reliable results at the minimum computational cost. All algorithms were implemented two times, one using the RFM table and the other using the RFMD table to see if there were significant differences with the algorithms and data provided.

These methods were chosen because of their easy implementations and wide applicability. K-Means, Hierarchical (Ward), and DBSCAN use hard cluster membership while Fuzzy C-Means is soft; this means the methods assign data-points to one cluster at a time and in the other data-points can be part of different clusters. Another reason for choosing these methods was the shapes that work with them: K-Means and Fuzzy C-Means work best when data-points have a spherical shape, Hierarchical (Ward) and DBSCAN cover random shapes. Additionally, some of these methods handle outliers much better which helps with the dataset.

3.8 Cluster Evaluation and Model Comparison

Cluster evaluation involved assessing the clustering algorithms mentioned before, which include: K-Means, fuzzy C-Means, hierarchical clustering (Ward), and DBSCAN. In the case of K-Means and hierarchical clustering (Ward), we used metrics such as WCSS (Inertia) and silhouette score to select the best k (number of clusters), for fuzzy C-Means we focused on the fuzzy partition coefficient (FPC) and the fuzzy entropy to select k (number of clusters); it is important to mention that due to its functioning there is a way to assign hard labels to data points, by doing this it is possible to use silhouette score too. Finally, for DBSCAN, we varied the parameters epsilon and min_samples to choose the number of clusters formed and the outlier detection. Each method was used with the RFM and RFMD data to understand if product diversity affected the results obtained. In the end, PCA was used to visualize the clusters in 2D.

To choose the most appropriate model for a task like this, we focused on using metrics like the Calinski-Harabasz Index, Davies-Bouldin Index, and Between-Cluster Sum of Squares (BSS) which can be used in methods like K-Means, fuzzy C-Means, hierarchical clustering (Ward). The idea was to choose the best metrics for each dataset (RFM or RFMD) and then choose the model accordingly.

3.8.1 Stability and Validation of Clustering Results

Choosing the right parameters for each method is an important part because it provides a type of internal validation, or in other words, helps us verify the results obtained and if they make sense. Machine learning methods are divided into two: supervised and unsupervised methods. In this case, the methods applied are classified as unsupervised, and given this, they cannot be validated using cross-validation. In the case of K-Means and hierarchical clustering, we used the silhouette score to determine the optimal number of k (clusters). For this, we plotted the silhouette scores obtained for different numbers of k (from 2–14) and chose the local maximum. Usually, we take the global maximum, but given that in this case this value equals two, we decided to move with the local to improve segmentation and make it more real considering the context.

In the case of fuzzy C-Means, given the nature of the algorithm, the silhouette score doesn't make that much sense, so to choose the optimal number of c (clusters) we focused on the Fuzzy Partition Coefficient (FPC) and Fuzzy Entropy. We used a method referred to as grid-search in which we use different values of the parameters to find the best combination of parameters that gives the best metrics, in this case the two previously mentioned. We ranged from 2–14 values of c , $m \in \{1.5, 2.0, 2.5\}$ which explain how hard or soft the clusters become, $\epsilon \in \{0.001, 0.005\}$ which explain the changes made, the lower the value the better, maxiter $\in \{500, 1000\}$ refers to the number of steps the model uses before stopping.

Finally, for DBSCAN, which doesn't take into consideration the value of clusters k to form groups, we only used a small grid search with ϵ , which is the distance between two elements to be considered neighbors. For this, we varied the values between 0.1 and 2 with steps of 0.1, and the number of samples was varied between 2 and 10. For DBSCAN we used silhouette score to assess the success of the method. All the graphs corresponding to each method were graphed using the best parameters.

3.9 Customer Lifetime Value Estimation

The Customer Lifetime Value (CLV) was calculated using the Discounted Cash Flow. The formula used is:

$$CLV = \sum_{t=1}^T \frac{R - C}{(1 + d)^t} - CAC$$

where:

- R is the estimated annual revenue per customer, computed as Monetary/ T ,
- C is the annual cost, estimated as $R \times$ cost ratio,
- d is the discount rate (10%),
- CAC is the customer acquisition cost (£100),
- T is the assumed customer lifespan (3 years).

4 Results

4.1 RFM and RFMD Feature Construction

To analyze customer behavior, we constructed a table based on RFMD features. The analysis was divided to discuss first RFM (Recency, Frequency and Monetary) and then RFMD which is essentially the same but includes another feature which is diversity, measuring also the amount of unique products bought by each customer.

First of all, we constructed a box-plot for each of the four features to see the distribution just as shown in Figure 6. Additionally, Table 3 helps to represent the exact values of the plots found in Figure 6.

As shown in Table 3, "Recency" feature is right skewed. This indicates that while the average customer buys products every 92 days, half of the customers purchase every 50 days or less. This behavior illustrates that this e-commerce site has a mix of recent buyers alongside customers with longer recency. Figure 6 shows that most customers made a purchase between 25 and 150 days ago, while a few customers haven't bought anything in the past 300 days. It's also important to mention that there aren't any customers with recency between 150 and 325 days.

As shown in Table 3, "Frequency" feature is right skewed. This shows that while the average customer bought at least 4 times in the given period, half of the customers bought 2 times or less. This behavior illustrates that the customer base consists of one-time buyers alongside others who have made multiple purchases. Figure 6 shows that most customers made between 0 and 60 purchases in the given period, while a small percentage bought more than 200 times.

As shown in Table 3, "Monetary" feature is right skewed, meaning most customers spend less than £2054.0. The median value of £674.0 shows that half of the customers

spend this amount or less, a small number of customers could be classified as high spenders and this group is responsible for raising the average. Figure 6 shows most purchases fall below £50,000, with a few outliers spending more than £250,000.

As shown in Table 3, "Diversity" feature is right skewed. This means that while the average customer bought at least 62 different products, half of the customers bought 35 or fewer. This behavior illustrates that most customers tend to buy a high number of distinct products, while a smaller segment purchases even more diverse products. Figure 6 shows most customers purchase between 0 and 150 distinct products while a number of outliers purchase more than 250 products.

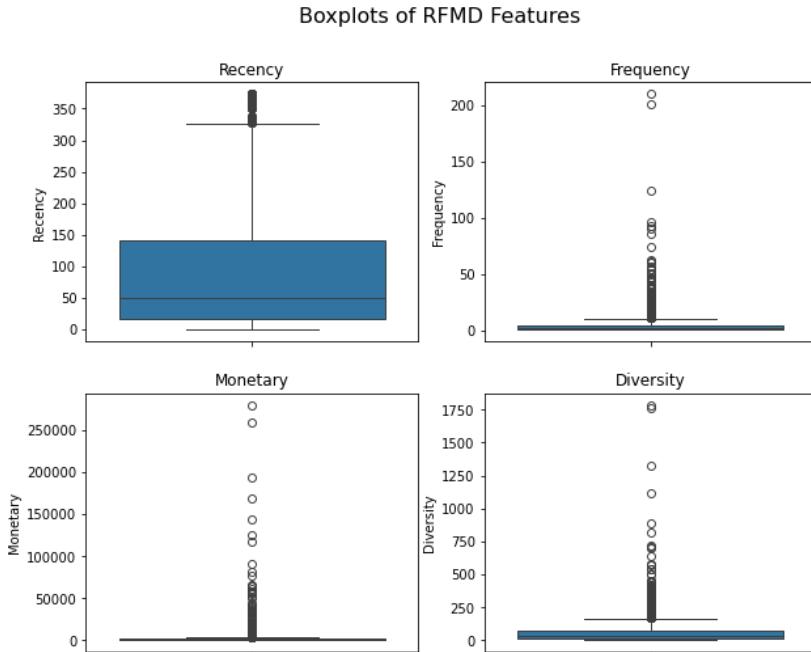


Figure 6: Box-plots for Recency, Frequency, Monetary, and Diversity

Table 3: Summary Statistics for RFMD Features

Feature	Mean	Median	Std
Recency	92.0	50.0	100.0
Frequency	4.0	2.0	8.0
Monetary	2054.0	674.0	8988.0
Diversity	61.0	35.0	85.0

Another graph used to display customer segments based on RFM is the segmentation heatmap, which helps understand customer behavior patterns by categorizing customers according to different feature levels. In this case, the graphs include RFM

features and the number of customers in each segment. As mentioned before, these graphs were constructed by transforming each feature into a scale from 1 to 5, where 1 represents the lowest level and 5 the highest.

Figure 7 divides customers into 25 segments, where each group is defined by a combination of Recency, Frequency, and Monetary levels, and also shows the number of customers in each group. According to the scale, the graph shows that the most valuable customers have high levels in all features (e.g., R=5, F=5, M=5), while customers with a high churn risk exhibit the opposite pattern (e.g., R=1, F=1, M=1). Additionally, it is important to consider the middle segments, typically represented by medium levels (e.g., R=3, F=3, M=3), which correspond to average customers.

These heatmaps allow companies to focus their efforts on a small portion of customers by providing personalized recommendations for each segment. In this case, the best customers have the following characteristics: R=5, F=5, and M=4.8, with 439 customers in this group, representing 10.2% of the total customer base. Concerning the average customers, they have characteristics R=3, F=3, and M=2.9, comprising 186 customers or 4.29%, which is notably smaller than the best or the worst segments. On the other hand, the lowest segment—associated with customers at risk of churn—has R=1, F=1, and M=1.7, with 363 customers, accounting for 8.36% of the customer base.

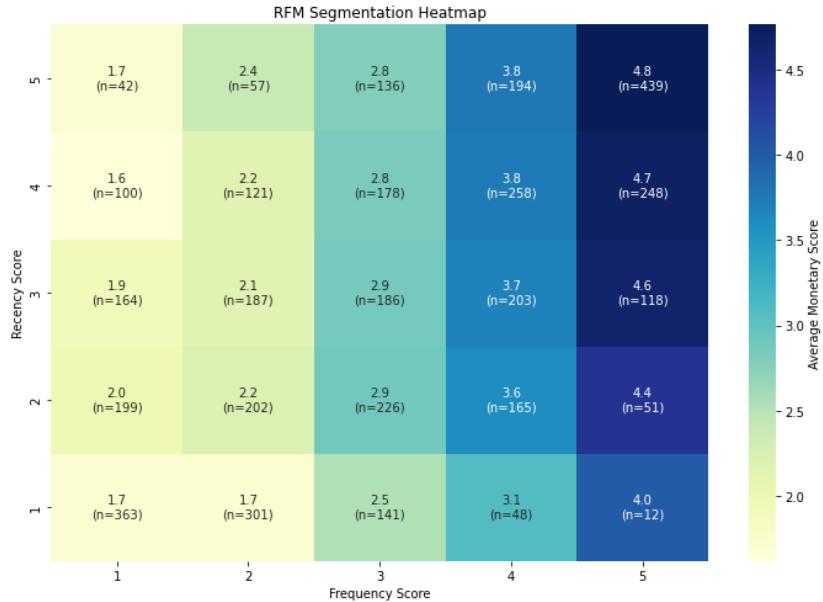


Figure 7: RFM Heatmap – Avg. Monetary by R F (+Cust. Count)

In the following section, we describe the RFMD heatmaps, which were constructed similarly to the RFM heatmaps but with an important distinction: five separate graphs

were created, each representing a different level of Diversity (D) from 1 to 5. Each graph displays the distribution of customers across Recency, Frequency, Monetary, and the specific level of Diversity.

Figure 8 presents the RFMD heatmap for customers with the lowest diversity level (D=1). This graph reveals that the best customers in this segment have high feature values of R=5, F=5, and M=4.4. Notably, there are only 10 customers in this top segment, representing 1.07% of the customer base within the D=1 group. The average customers in this segment are characterized by R=3, F=3, and M=1.7, comprising 21 customers or 2.27% of the D=1 base. Finally, the segment with customers at highest risk of churn is represented by R=1, F=1, and M=1.3, which includes 174 customers, accounting for 18.79% of the D=1 customer base. It is important to highlight that the average customers and the least loyal customers in this group have similar spending levels.

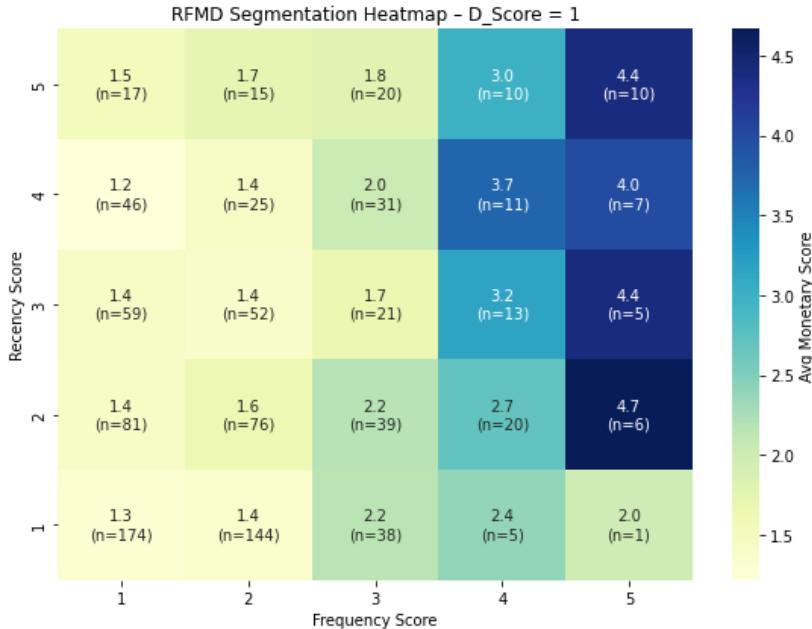


Figure 8: RFMD Heatmap–D=1(Avg. Monetary Score Customer Count)

Figure 9 displays the RFMD heatmap for customers with Diversity level 2 (D=2). The best customers in this segment are characterized by high feature values of R=5, F=5, and M=4.4. It is important to note that there are only 10 customers in this top segment, representing 1.22% of the customer base within the D=2 group. The average customers are characterized by R=3, F=3, and M=1.7, comprising 35 customers or 4.27% of the D=2 base. Finally, the segment with the highest churn risk is defined by R=1, F=1, and M=1.3, which includes 174 customers and accounts for 14.30% of the

customer base in this diversity level. It is also worth mentioning that the average and least loyal customers have similar spending levels in this segment.

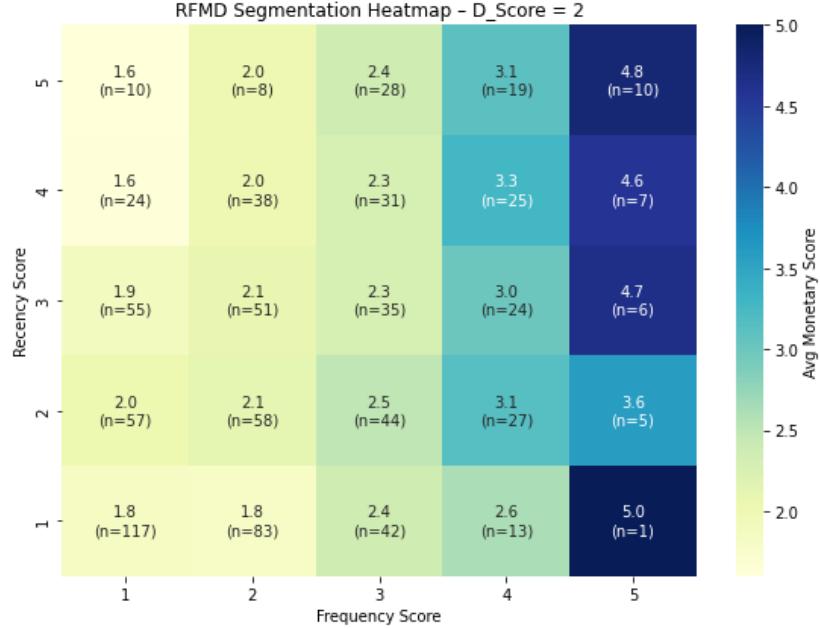


Figure 9: RFMD Heatmap–D=2(Avg. Monetary Score Customer Count)

Figure 10 presents the RFMD heatmap for customers with Diversity level 3 (D=3). The best customers in this segment are characterized by high values of R=5, F=5, and M=4.6. There are 38 customers in this top segment, representing 4.41% of the customer base within the D=3 group. The average customers are characterized by R=3, F=3, and M=3.1, comprising 55 customers or 6.38% of the D=3 base. Lastly, the segment with the highest churn risk is defined by R=1, F=1, and M=2.4, which includes 46 customers and accounts for 5.33% of the customer base in this diversity level. It is important to note that the spending differences among these three main groups are significant.

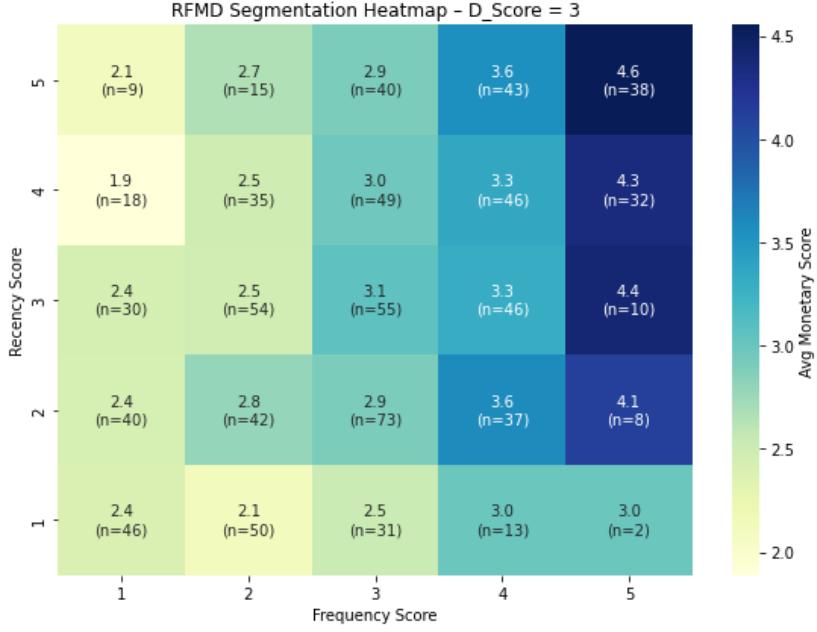


Figure 10: RFMD Heatmap–D=3(Avg. Monetary Score Customer Count)

Figure 11 presents the RFMD heatmap for customers with Diversity level 4 (D=4). The best customers in this segment are characterized by high values of R=5, F=5, and M=4.6. This segment includes 93 customers, representing 10.73% of the customer base within the D=4 group. The average customers show characteristics of R=3, F=3, and M=3.0, with 48 customers accounting for 5.54% of the D=4 base. Finally, the segment at highest risk of churn is defined by R=1, F=1, and M=2.5, consisting of 24 customers or 2.77% of the segment. It is important to note that spending differences among these three significant groups vary notably.

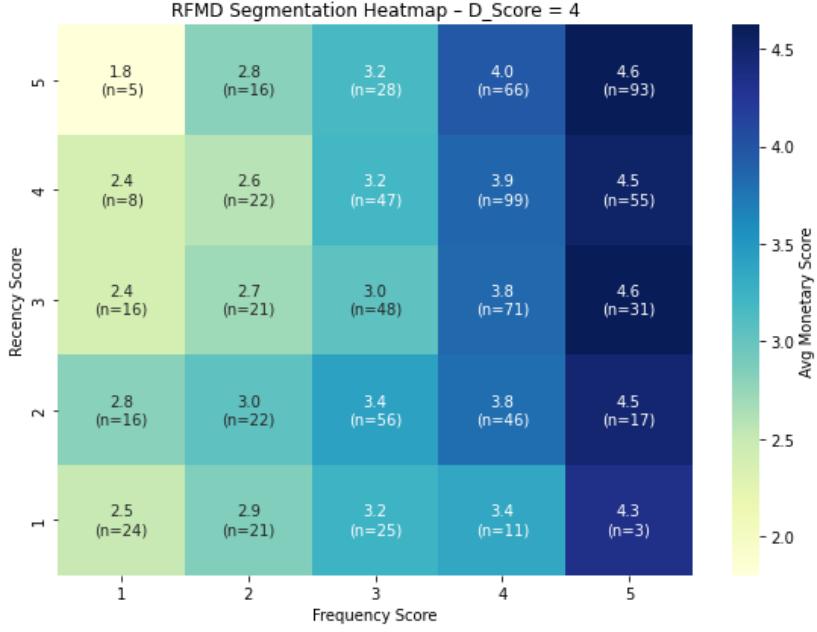


Figure 11: RFMD Heatmap–D=4(Avg. Monetary Score Customer Count))

Figure 12 shows the RFMD heatmap for customers with Diversity level 5 (D=5). The best customers in this segment have high values of R=5, F=5, and M=4.8, comprising 288 customers, which accounts for 33.26% of the customer base in the D=5 group. The average customers are characterized by R=3, F=3, and M=3.9, with 27 customers representing 3.12% of the segment. Interestingly, the segment with R=1, F=1, and M=4.5 includes only 2 customers (0.23%), but these least loyal customers have higher spending than the average customers and nearly match the most loyal customers in monetary value. It is also noteworthy that customers purchasing a wide variety of distinct products generally spend more than the overall average, except for the subgroup with low frequency and high recency, which spends less.

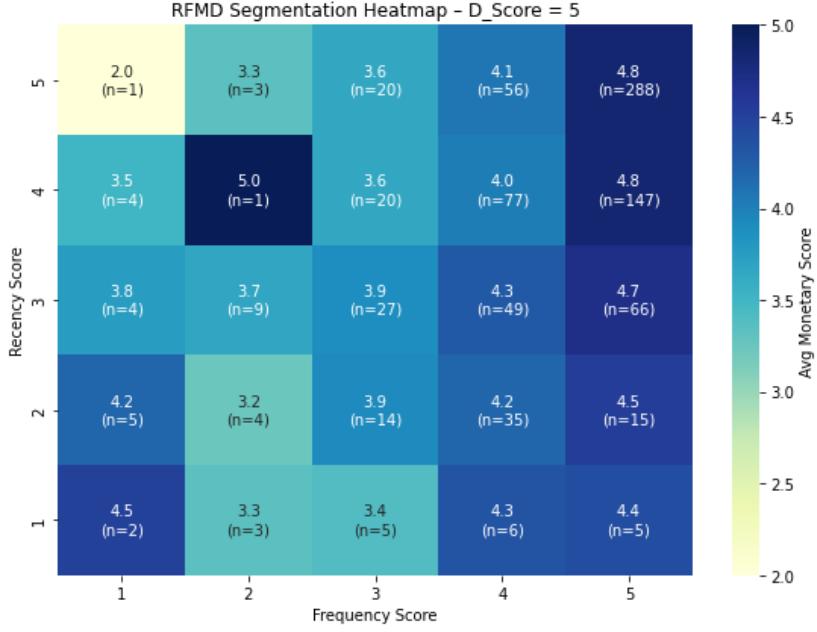


Figure 12: RFMD Heatmap–D=5(Avg. Monetary Score Customer Count)

4.2 Applied Segmentation Techniques

As mentioned before, this research project aims to compare clustering algorithms and ultimately select the one most appropriate for the task, balancing computational efficiency with accuracy. The following algorithms were tested: K-Means, Fuzzy C-Means, Hierarchical Clustering (Ward), and DBSCAN. These algorithms were implemented using two feature sets: RFM (Recency, Frequency, Monetary) and RFMD (Recency, Frequency, Monetary, Diversity). This was done to identify if including product diversity leads to significant differences in the clustering results.

The visualization of the clusters was performed by applying the dimensionality reduction technique Principal Component Analysis (PCA). It is important to note that PCA was only applied during the plotting phase to reduce the number of features and enable 2D visualization. It was not used in the preprocessing phase to reduce dimensionality before clustering.

K-Means was the first method applied to the feature sets. It works by partitioning data into k clusters through minimizing within-cluster variance, making clusters as compact as possible. The algorithm creates a centroid for each cluster and assigns data points to the nearest centroid, iterating this process until convergence. Figure 13 illustrates the clusters obtained by applying K-Means on the RFMD dataset.

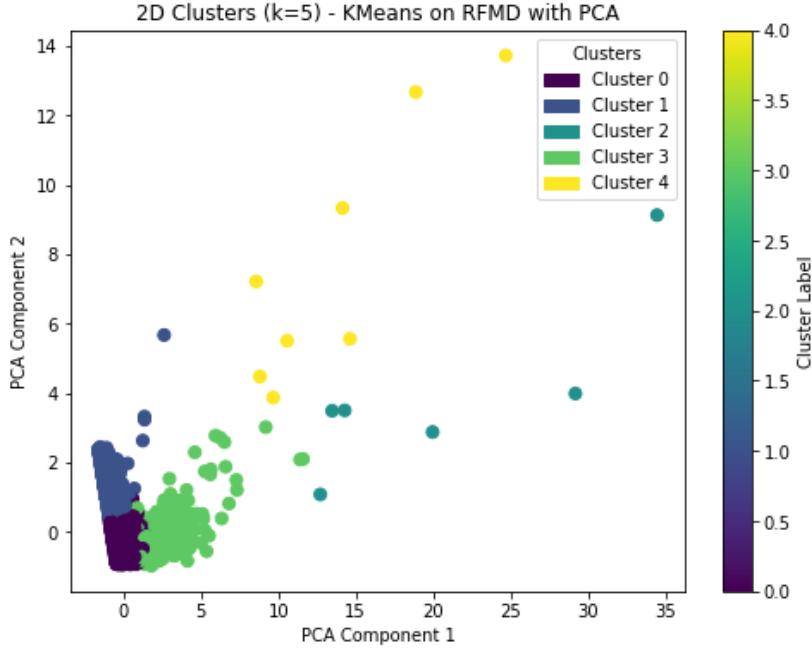


Figure 13: K-Means Clustering with 4 Clusters on RFMD (PCA 2D View)

Hierarchical clustering is a method where clusters are formed by progressively merging two clusters at a time until only one large cluster remains. This approach is valuable because it allows the user to observe the order in which clusters are combined. In this research, the clustering process follows Ward's method, which forms clusters by minimizing the total within-cluster variance. Figure 14 shows the dendrogram, a graph illustrating the merging of clusters and their hierarchical structure based on the RFMD features. The corresponding dendrogram for RFM is provided in the appendix. Figure 15 displays the clustering results obtained using the RFMD features, while the RFM cluster diagram can be found in the appendix.

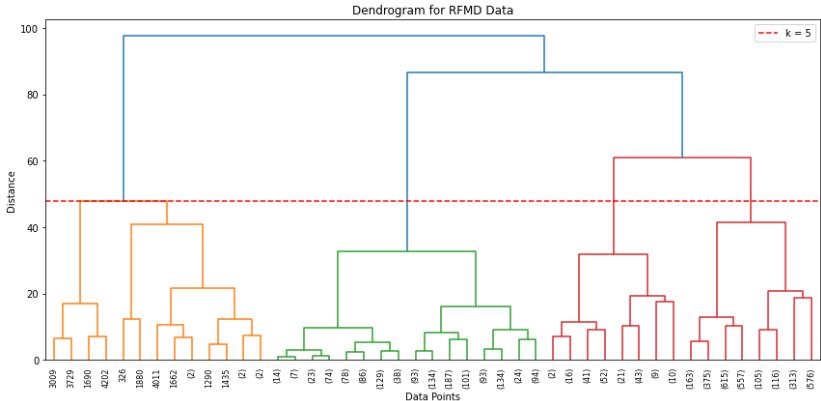


Figure 14: Dendrogram of Hierarchical Clustering (Ward Method) on RFMD Data

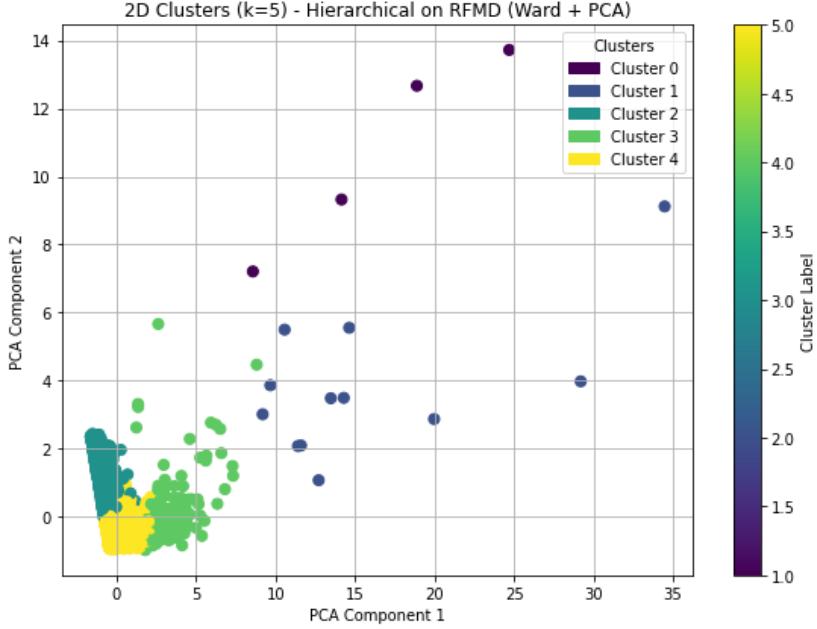


Figure 15: Hierarchical Clustering (Ward) with 4 Clusters on RFMD (PCA 2D View)

Fuzzy C-Means is a clustering method where data points are assigned to multiple clusters simultaneously, each with a degree of membership. This method is based on the idea that, in real-world scenarios, data points rarely belong entirely to a single group but often share similar attributes across groups. Figure 16 displays the clustering results using RFMD features for both hard and soft labels, with the RFM cluster diagram shown in the appendix.

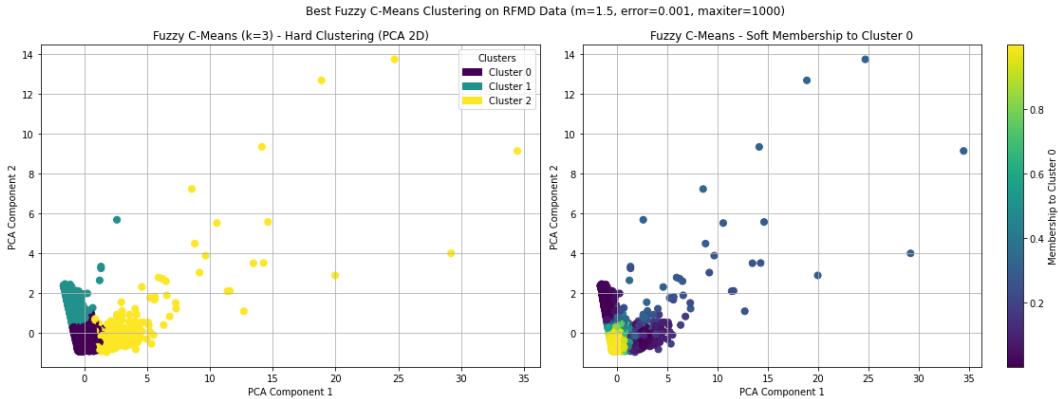


Figure 16: Fuzzy C-Means Clustering on RFMD: Hard vs Soft Assignments (PCA 2D)

Density-based clustering, also known as DBSCAN, is a clustering algorithm that forms clusters based on a specified radius (epsilon) and a minimum number of neighboring points. Points not meeting these criteria are classified as noise. Figure 17 displays the clustering results using RFMD features, with the RFM cluster diagram shown in the appendix.

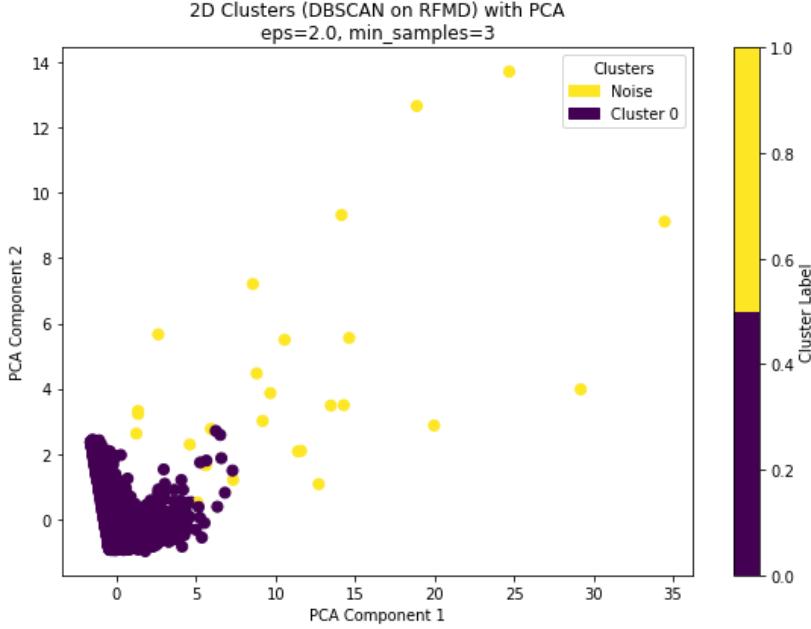


Figure 17: DBSCAN Clustering on RFMD (PCA 2D View)

4.3 Cluster Evaluation

4.3.1 Evaluation Metrics for Clustering Parameter Selection

In the next part we will focus on the metrics used to determine the best parameters in each clustering method. It is important to mention that all the figures shown will be referring to RFMD features, while corresponding results for RFM features can be found in the appendix.

First, we discuss the metrics used to evaluate K-Means and Hierarchical Clustering (Ward). For K-Means, we focused on the elbow method and the silhouette score to choose the optimal number of clusters. For Hierarchical Clustering (Ward), the dendrogram was also used to support the decision along with these two metrics. Figure 42 displays a total of four graphs; in the top row, from left to right, we visualize the K-Means elbow plot and the K-Means silhouette score. In the bottom row, the same plots are shown but for Hierarchical Clustering (Ward). Each graph includes a vertical line indicating the optimal number of clusters k chosen. The selected number of clusters for both methods is five. Additionally, Figure 43 shows the dendrogram for Hierarchical Clustering (Ward), where a horizontal line cuts the dendrogram at the level balancing the metrics, which also corresponds to five clusters. Corresponding graphs for RFM can be found in the appendix.

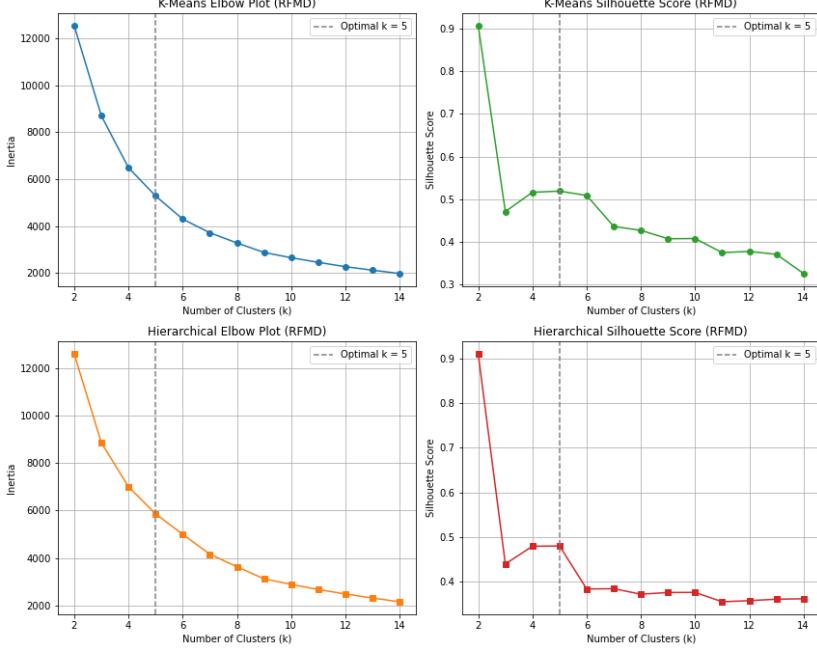


Figure 18: Elbow method and silhouette score K-Means and Hierarchical clustering(RFMD)

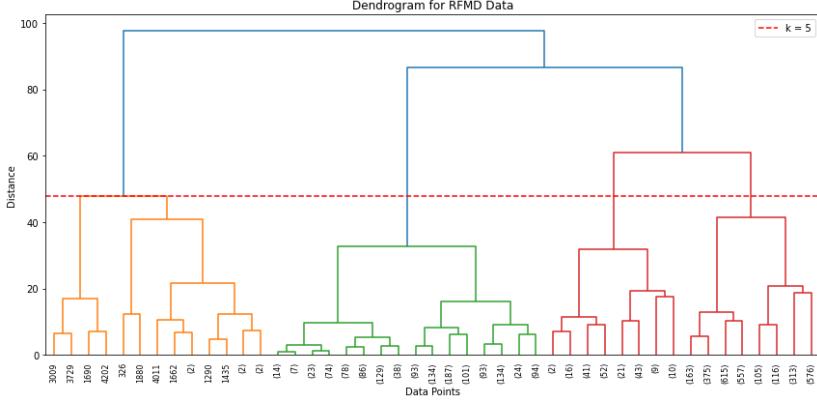


Figure 19: Dendrogram Hierarchical Clustering (RFMD)

In the case of Fuzzy C-Means, since it relies on a different logic, it only makes sense to use the Fuzzy Partition Coefficient (FPC) and Entropy as evaluation metrics. To obtain the best results, we performed a grid search to vary the algorithm’s parameters and then applied the best parameters to find the optimal number of clusters k . The final parameter values obtained were $m = 1.5$, error = 0.001, and max iterations = 1000. Based on these metrics and by plotting the FPC and Entropy graphs, we determined that the optimal number of clusters is three, as shown in Figure 44. Corresponding figures for RFM can be found in the appendix.

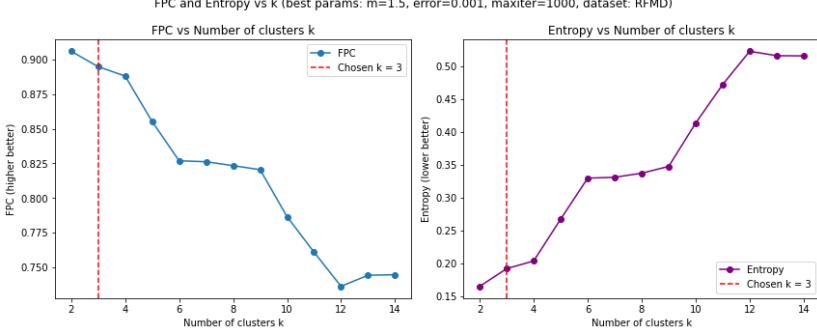


Figure 20: FPC and entropy Fuzzy C-Means (RFMD)

For DBSCAN, the metric used to determine the optimal number of neighbors (`minPts`) required for a point to be considered part of a cluster or noise is the k-distance plot. Figure 21 displays the k-distance plot, which indicates that the optimal value for k is two.

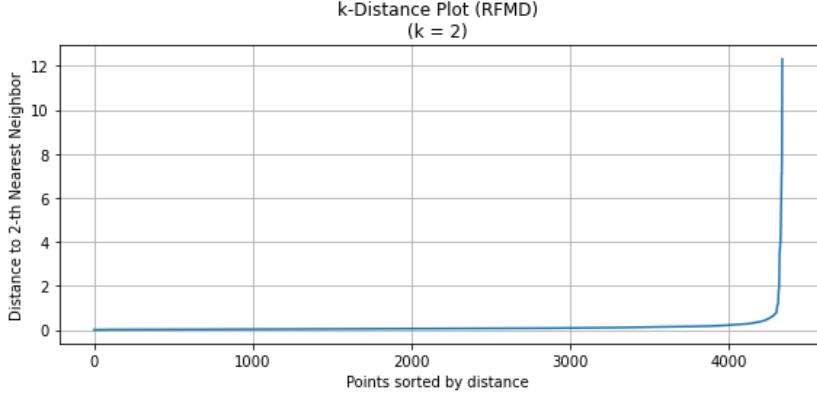


Figure 21: k-distance plot (RFMD)

RFM results can be found in the appendix for reference.

4.3.2 Metrics Used to Compare Clustering Models

Table 4 and Table 5 present the best metric results for each of the clustering methods. In the first table, we observe the metrics for K-Means, Hierarchical Clustering (Ward), and Fuzzy C-Means. The table shows that K-Means achieves the highest silhouette score, followed by Fuzzy C-Means, and then Hierarchical Clustering. For the Calinski-Harabasz index, K-Means again holds the highest value, followed by Hierarchical Clustering and then Fuzzy C-Means. Regarding the Davies-Bouldin Index (DBI), Fuzzy C-Means records the highest (worst) score, followed by K-Means and finally Hierarchical Clustering.

Additionally, it can be noted that both K-Means and Hierarchical Clustering resulted in $k = 5$ clusters, while Fuzzy C-Means opted for $k = 3$ clusters. In the case

of DBSCAN, only the silhouette score of 0.88 is reported, achieved with an ϵ value of 2 and a minimum samples parameter of 3. The results for RFM can be found in the appendix.

Table 4: Clustering Performance Metrics for RFMD (excluding DBSCAN)

Method	Silhouette	Calinski-Harabasz	Davies-Bouldin	Optimal k
K-Means	0.5189	2477.40	0.8601	5
H. Clustering (Ward)	0.4801	2124.56	0.8510	5
Fuzzy C-Means	0.5092	1526.98	0.9401	3

Table 5: DBSCAN Performance Metrics for RFMD

Method	Epsilon (eps)	Min Samples	Silhouette
DBSCAN	2.00	3	0.8807

In the next part, we present two bar graphs corresponding to the Calinski-Harabasz index and the Davies-Bouldin Index (DBI). Figure 22 illustrates the comparison of Calinski-Harabasz scores for K-Means, Hierarchical Clustering (Ward), and Fuzzy C-Means. Similarly, Figure 23 shows the comparison of DBI values for the same clustering methods. The corresponding figures for RFM are available in the appendix.

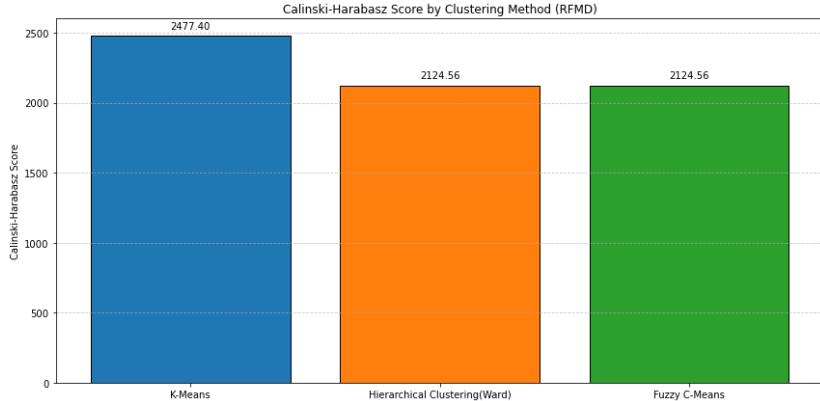


Figure 22: Calinski-Harabasz scores for K-Means, Hierarchical, and Fuzzy C-Means (RFMD)

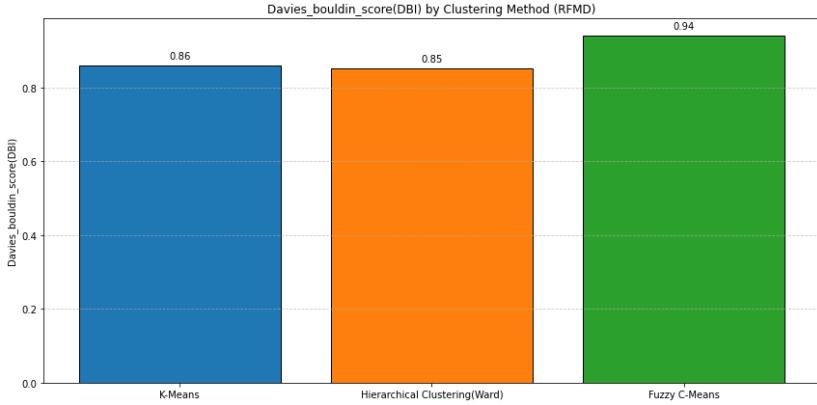


Figure 23: DBI scores for K-Means, Hierarchical, and Fuzzy C-Means (RFMD)

4.4 Model Comparison

4.4.1 K-Means Results (RFMD)

As mentioned before, PCA was used to reduce the number of features to two dimensions for visualizing the clusters. The axes of the graph correspond to Principal Component 1 and Principal Component 2. In the top corner, there is a legend indicating the number of clusters and their corresponding colors. Additionally, a vertical color bar helps identify cluster labels according to the colors. In the following section, we present the scatter plot of clusters for RFMD, while the corresponding plot for RFM is provided in the appendix for reference.

As shown in Figure 24, each point represents a customer, and the color indicates the cluster assignment. The graph displays five distinct clusters. Among these, clusters 0, 1, and 3 are located close to each other, with some overlap between customers, whereas clusters 2 and 4 are more dispersed and lack tightness. It is evident that the clusters do not form clearly defined shapes.

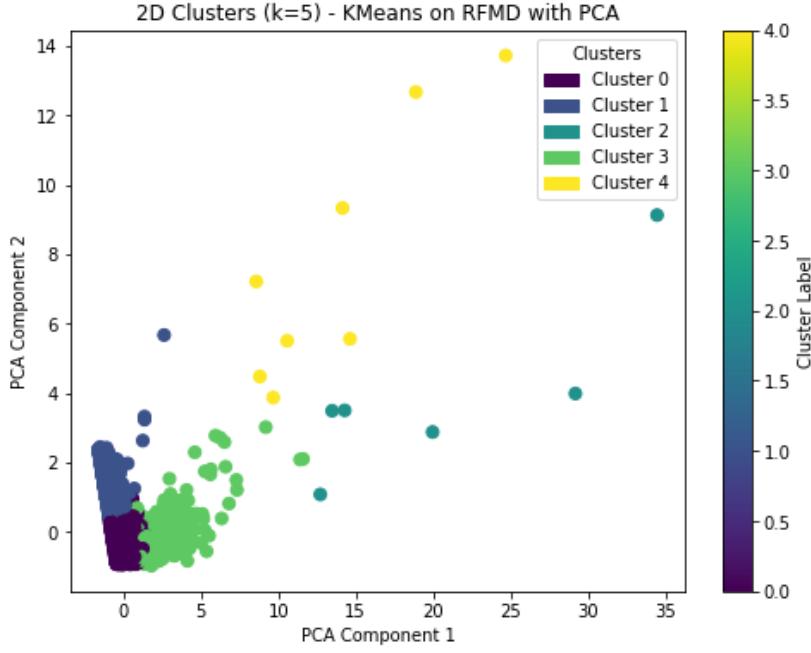


Figure 24: K-Means clusters visualized using PCA (RFMD features)

Just as mentioned before, PCA was used to plot the clusters in two dimensions. After applying PCA, the explained variance by the two principal components was found to be approximately 78%. Additionally, as shown in Figure 25, Frequency and Diversity are the features that contribute the most to Principal Component 1 (PC1), although Recency and Monetary also play a significant role in its explanation. On the other hand, Principal Component 2 (PC2) is mainly explained by Recency and Monetary, while Frequency and Diversity have a much smaller impact on this component.

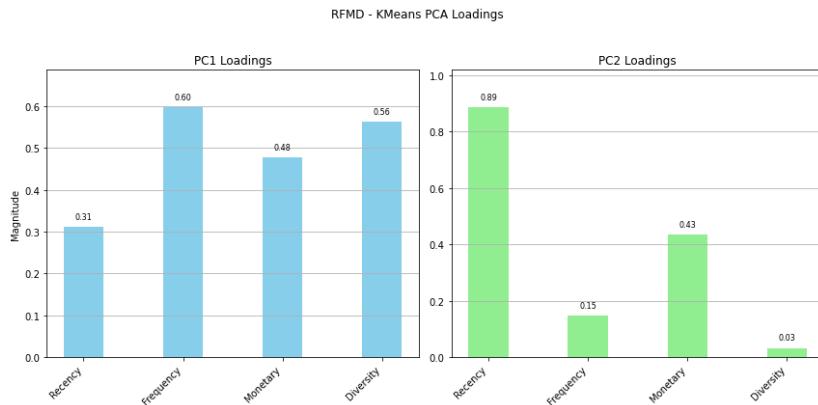


Figure 25: PCA component loadings of RFMD features used in K-Means clustering

4.4.2 Hierarchical Clustering Results (RFMD)

As mentioned before, PCA was used to reduce the number of features to two dimensions for visualization purposes. The axes of the graph correspond to Principal Component 1 and Principal Component 2. In the top corner, there is a legend indicating the cluster numbers and their corresponding colors; additionally, a vertical bar helps identify the cluster labels by color. In the following section, we present the scatter plot of clusters for RFMD, while the corresponding plot for RFM can be found in the appendix for reference.

As shown in Figure 26, each point represents a customer and the color indicates the cluster to which they belong. The graph displays 5 distinct clusters. Among these, clusters 2, 3, and 4 are located close to each other, with some overlap between customers, whereas clusters 0 and 1 are more spread out across the figure, showing less tightness. It is evident that there is no clear or distinct shape defining the clusters.

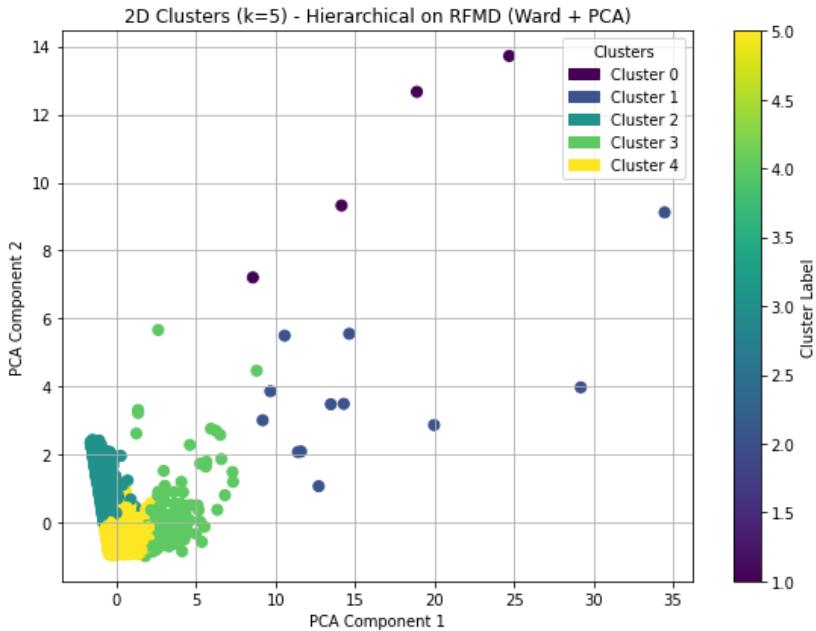


Figure 26: Hierarchical clusters visualized using PCA (RFMD features)

Just as mentioned before, PCA was used to plot the clusters in two dimensions. After applying PCA, the explained variance by the principal components is approximately 78%. Additionally, as can be seen in Figure 27, frequency, monetary, and diversity are the features that contribute the most to PC1, while recency plays a less important role. On the other hand, PC2 is mainly explained by recency, followed by monetary, whereas frequency and diversity have little influence—it's important to note that diversity has a value close to zero in PC2.

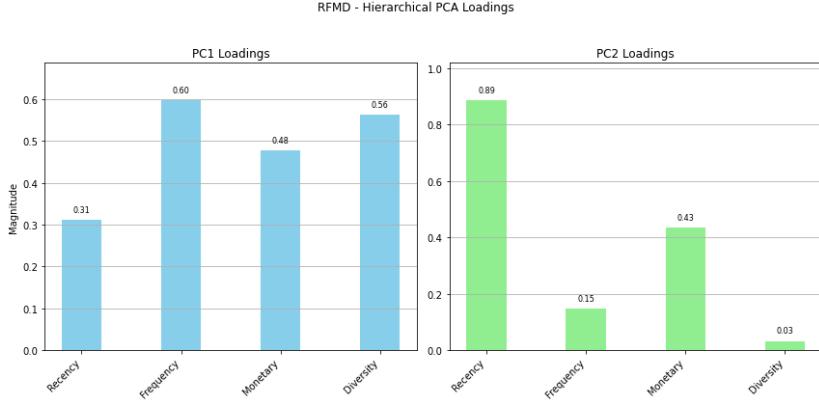


Figure 27: PCA component loadings of RFMD features used in Hierarchical clustering

4.4.3 Fuzzy C-Means Results (RFMD)

As it was mentioned before, PCA was used to reduce the number of features into two and view the clusters graph in two dimensions. The axes of the graph correspond to principal component 1 and principal component 2 and in the top corner there is a legend for the number of clusters and their color; additionally, there is a vertical bar that helps identify the cluster label according to the colors. In the next part we will show scatter-graphs of clusters for RFMD, and the one for RFM can be found in the appendix for reference.

As shown in Figure 28, each point represents a customer and the color indicates the cluster they are assigned to. The graph shows 3 distinct clusters which are very close to one another with no clear boundaries. While clusters 1 and 2 are concentrated in the lower left corner, cluster 0 is next to them but some of the customers in that cluster are also spread across the graph. In Figure 28, the graph on the right represents the soft labels of fuzzy c-means. In this case, points are assigned to multiple clusters simultaneously, and depending on the color intensity of the points, more than half of the customers belong partially to more than one cluster.

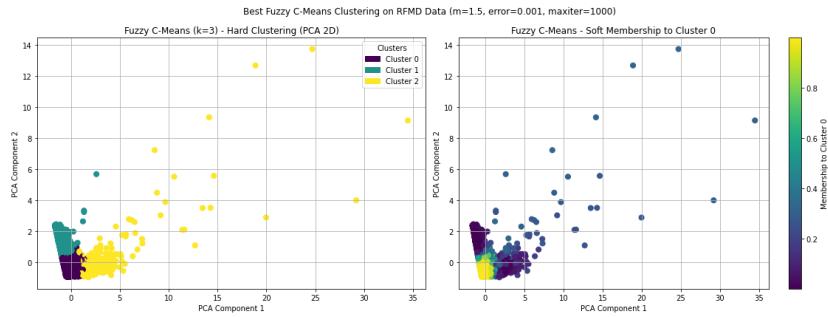


Figure 28: Fuzzy C-Means clusters visualized using PCA (RFMD features)

Just as it was mentioned before, PCA was used to plot the clusters in two dimensions. After doing this, we obtain that the explained variance by the principal components is approximately 78%. Additionally, as can be seen from Figure 29, frequency, monetary, and diversity are the features that explain most of PC1, while recency is less important. On the other hand, PC2 is mainly explained by recency followed by monetary, whereas frequency and diversity play a minor role; it is important to mention that diversity has a value close to zero in PC2.

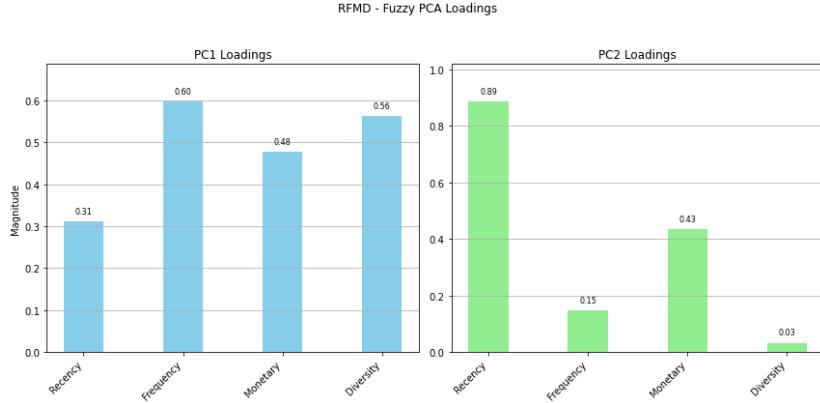


Figure 29: PCA component loadings of RFMD features used in Fuzzy C-Means

4.4.4 DBSCAN Results (RFMD)

As it was mentioned before, PCA was used to reduce the number of features into two and view the clusters graph in two dimensions. The axes of the graph correspond to principal component 1 and principal component 2 and in the top corner there is a legend for the number of clusters and their color; additionally, there is a vertical bar that helps identify the cluster label according to the colors. In the next part we will show scatter-graphs of clusters for RFMD, and the one for RFM can be found in the appendix for reference.

In the case of DBSCAN, as a method completely different from the others, some points are classified into clusters while others are classified as noise. As shown in Figure 30, the majority of customers were classified into a single cluster located near (0,0) without presenting a distinct shape. Additionally, customers not close to one another were classified as noise given the parameters used.

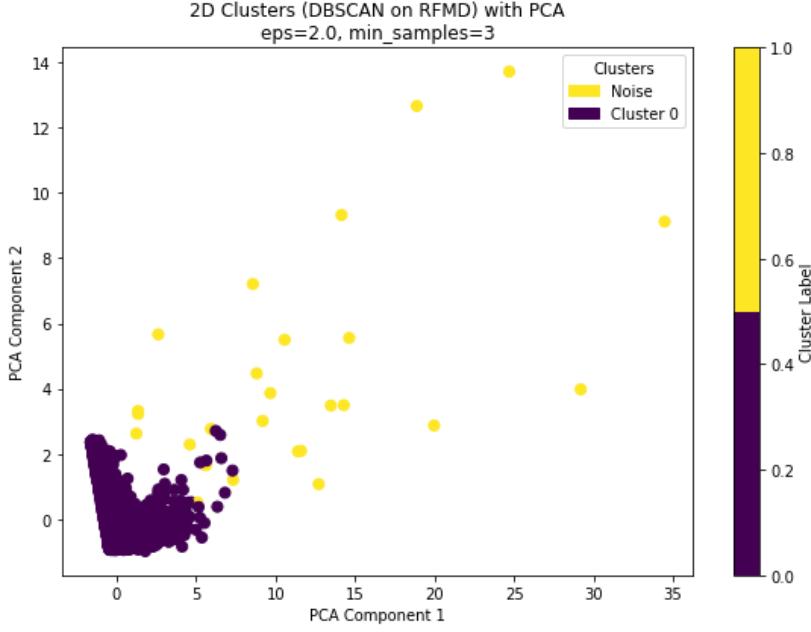


Figure 30: DBSCAN clusters visualized using PCA (RFMD features)

Just as it was mentioned before, PCA was used to plot the clusters in two dimensions. After doing this, we obtain that the explained variance obtained by the principal components is approximately 85%. Additionally, as can be seen from Figure 31, frequency, monetary, and diversity are the features that explain most of PC1, while recency is less important. On the other hand, PC2 is explained mainly by recency followed by monetary, while frequency and diversity play a minor role; it is important to mention that diversity has a value close to zero.

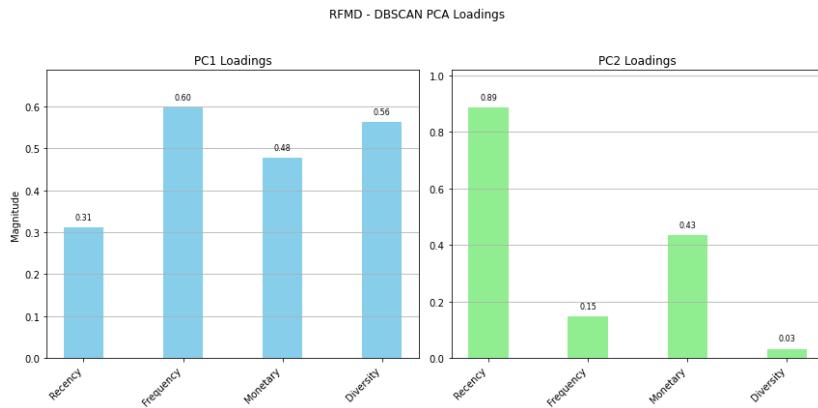


Figure 31: PCA component loadings of RFMD features used in DBSCAN

4.5 CLV Distribution and Cluster Sizes by Segmentation Method

Customer Lifetime Value (CLV) was computed using the formula described in the methodology. In the following section, we present graphs illustrating the number of customers per cluster for each clustering method based on the RFMD dataset, alongside box-plots displaying the CLV distribution within each cluster per method. The results for RFM are included in the appendix; however, only RFMD results are shown here as they provide a clearer picture.

Figure 54 displays four bar charts representing the number of customers per cluster for each method. This visualization provides insight into the relative importance of each cluster in terms of customer count, aiding reliable business decision-making. For K-Means, cluster 0 contains the largest number of customers, followed by clusters 1 and 3. Clusters 2 and 4 represent only a small fraction of the customer base. In contrast, Hierarchical clustering shows that cluster 5 has a similar size to K-Means cluster 1, while clusters 3 and 4 also contain significant customer counts; clusters 1 and 2 have near-zero customers. Fuzzy C-Means, with only three clusters, shows no clusters with negligible sizes; clusters 0 is the most representative in terms in the number of customers, while clusters 1 and 2 are less representative. Lastly, DBSCAN assigns the majority of customers to cluster 0, with the remainder classified as noise or ambiguous points. Corresponding bar plots for RFM features can be found in the appendix.

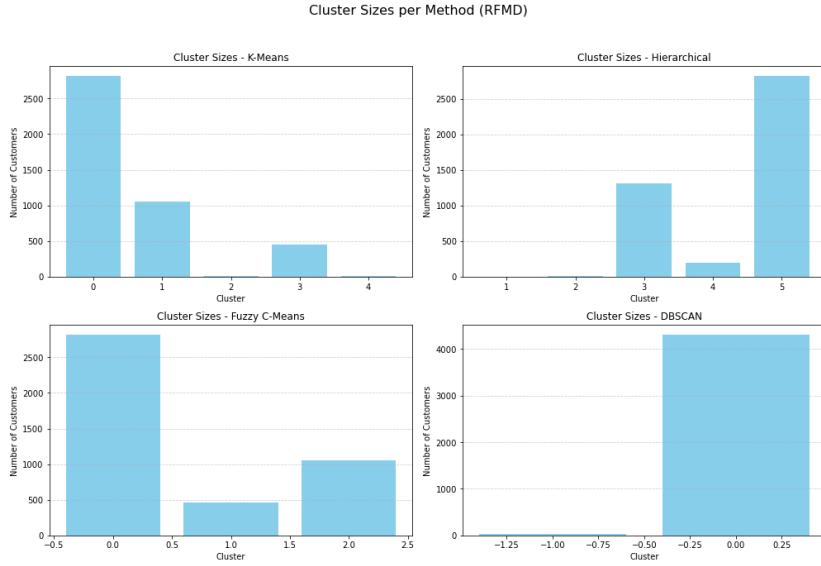


Figure 32: Number of Customers per Cluster per Method (RFMD)

Figure 33 shows the CLV distribution for each cluster. These box-plots assist in determining actionable strategies tailored to each cluster. It is important to note that wider boxes indicate greater value dispersion and a larger interquartile range (IQR), while the absence or narrowness of boxes suggests homogeneity in CLV values. For

K-Means (5 clusters labeled 0 to 4), cluster 4 exhibits the widest box, indicating diverse CLV values ranging between 75,000 and 140,000. Next is cluster 2 which shows a small box near the value of 2500 which suggests a set of varied values of CLV. Clusters 0 and 1 have narrow or nearly absent boxes, implying similar low CLV values among their customers. Finally, cluster 3 has a small box above 0 which mean points following which suggests a great number of outliers. In Hierarchical clustering, clusters 1 and 2 also show notable dispersion with those in cluster having higher values in CLV than in cluster 2, whereas clusters 3 to 5 demonstrate less variability, with cluster 4 showing a small variability in CLV close 500. For Fuzzy C-Means (3 clusters), none display substantial boxes, suggesting homogeneous low CLV values; however, cluster 2 contains more outliers those being customers with exceptionally high values of CLV. For DBSCAN, customers classified as noise (cluster -1) show a wider and higher CLV range compared to cluster 0, whose members have more similar CLV values closer to 0. RFM box-plots are provided in the appendix.

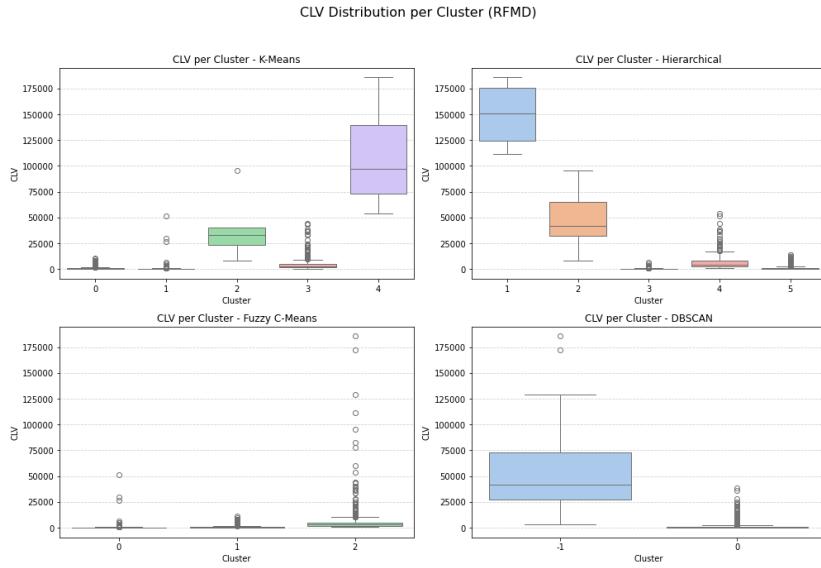


Figure 33: CLV Distribution per Cluster per Method (RFMD)

Table 6 summarizes the average CLV and the number of customers per cluster for each clustering method. This overview assists in understanding the financial value and representation of each segment. Notably, some methods yield high average CLV values reaching 150,000 for certain clusters, while others report average CLV values closer to 300.

Table 6: Average CLV and Number of Customers per Cluster (RFMD)

Method	Cluster	Average CLV	Number of Customers
K-Means	0	668.96	2817
	1	312.97	1052
	2	38,616.06	6
	3	4,731.63	456
	4	109,095.02	8
Hierarchical	1	149,589.69	4
	2	47,723.54	12
	3	249.92	1309
	4	7,561.80	194
	5	890.29	2820
Fuzzy C-Means	0	674.95	2820
	1	6,984.99	464
	2	314.12	1055
DBSCAN	-1	56,759.32	26
	0	927.44	4313

5 Discussion

5.1 Insights from RFM and RFMD Feature Distributions

The RFM and RFMD methodologies are widely used today due to their simplicity and the valuable insights they provide. They enable analysis of customer behavior across various markets by examining how customers interact with the business, facilitating segmentation and identification of beneficial patterns.

Figure 6 presents box-plots illustrating customer behavior with respect to each feature. In this case, Recency shows a large interquartile range (IQR), indicating high variance in how recently customers have made purchases. Most customers have recency values between 25 and 150 days, with some outliers having much older recency values. Conversely, Frequency, Monetary, and Diversity display much narrower boxes, suggesting low variability and that most customers have similar spending habits, purchase frequencies, and product diversity. However, Diversity reveals a subgroup of customers with variability ranging approximately from 0 to 80 products, while the rest purchase similar numbers of products. These observations align well with the descriptive statistics summarized in Table 3.

To complement this analysis, Figure 34 shows histograms for each feature. It is evident that Recency has a broader distribution compared to the other features, reflecting a more dispersed customer behavior regarding how recently purchases were made. This

pattern is expected in an e-commerce context, where a high proportion of customers are one-time buyers who spend significantly but rarely purchase diverse products or return frequently.

Beyond univariate distributions, we examine the joint distribution of RFMD features through heatmaps. While box-plots and histograms provide an overview of individual feature behavior, heatmaps allow visualization of customer distribution across combined feature scores, alongside the number of customers in each segment. These heatmaps use scaled RFMD scores (from 1 to 5) to segment customers, resembling a Pareto-like distribution that helps identify key groups such as loyal customers, low spenders, and average customers. Such segmentation facilitates targeted marketing strategies that can boost retention and maximize business impact.

Figure 7 shows the distribution of customers across RFM scores. Among the 4,339 customers segmented into 25 groups, those with high Recency and Frequency scores also exhibit the highest average Monetary scores (around 4.7–4.8). This indicates that recent and frequent buyers tend to be high spenders, which aligns with intuitive expectations. Conversely, customers with low Recency and Frequency scores (1–2) tend to have the lowest monetary scores (approximately 1.6–2), representing likely one-time buyers or low engagement customers.

Notably, a small group of 12 customers shows high Frequency but low Recency scores combined with high Monetary values—these customers are prime candidates for reactivation campaigns as they have potential for renewed engagement. Similarly, customers with low Frequency but high Recency scores and low spending may benefit from targeted discounts or promotional emails. Overall, the steady increase in monetary value with rising Frequency and Recency underscores the importance of understanding these relationships for effective customer management.

Figures 8 through 12 illustrate how customer segmentation evolves as the Diversity score increases from 1 to 5. At low diversity ($D=1$), the heatmap resembles the original RFM distribution, but with a key difference: whereas the standard RFM heatmap shows most customers clustered in high-value segments (high Recency, Frequency, and Monetary), here the majority fall into the lowest-value segment, with the fewest customers in the highest-value group.

Focusing on extremes, the number of customers in the lowest RFM group decreases drastically from 174 at $D=1$ to only 2 at $D=5$ —a 98.9% reduction. Conversely, the count of VIP customers surges from 10 at $D=1$ to 288 at $D=5$, representing a 2780% increase. Across all 25 segments, monetary scores generally rise, with only one segment

scoring below 3. This trend suggests that incorporating Diversity effectively encourages greater customer engagement and spending, highlighting its value as an additional segmentation dimension in this business context.

5.2 Segmentation Outcomes

5.2.1 K-Means Clustering (RFM and RFMD)

As previously mentioned, Principal Component Analysis (PCA) was applied to the RFM and RFMD datasets to reduce dimensionality and facilitate visualization of customer features in two dimensions. It was also important to analyze the loadings of each principal component to understand which features most strongly influenced them. Initial insights from the RFM model are provided in the appendix in Figures 46 (clusters) and 47 (component loadings).

Figure 46 shows that after evaluating clustering metrics, K-Means with 4 clusters was selected. The first principal component (PC1) is mainly influenced by Frequency and Monetary value, with Recency contributing less—reflecting the general or average customer behavior. The second principal component (PC2) is dominated by Recency and moderately by Monetary value, capturing customer activity levels and responsiveness to promotions or new products.

Cluster interpretations are as follows: - Cluster 2 (green) groups customers with medium to high Recency, Frequency, and Monetary scores, including outliers with very high values—these correspond to loyal or VIP customers. - Cluster 3 (yellow), near the origin, shows low values across all features, indicating low-engagement or potentially churned customers. - Cluster 1 (blue) has higher Recency than cluster 3 but low Frequency and Monetary, suggesting one-time buyers. - Cluster 0 (purple) exhibits higher Frequency and Monetary but low Recency, representing average customers possibly influenced by special conditions.

While RFM-based clustering provided useful segmentation, incorporating Diversity (RFMD) enhances customer understanding. In the RFMD PCA (Figure ??), PC1 is mainly influenced by Frequency, Diversity, and Monetary, reflecting general customer purchasing behavior, while PC2 remains dominated by Recency and Monetary, indicating customer activity. This combination allows clustering customers not only by how often and how much they buy but also by how many different products they purchase.

K-Means with 5 clusters was implemented on RFMD features (Figure 24), with the following interpretations: - Cluster 0 (purple), near the origin, represents customers low in all features—likely one-time or disengaged buyers, possibly churned. - Cluster 1 (blue) has increasing Recency but low other features, corresponding to recent new customers. - Cluster 3 (green) includes customers with higher purchase counts, mone-

tary values, and product diversity but lower Recency—previously active but now less engaged. - Clusters 2 (dark green) and 4 (yellow) represent average customers with potential to become VIPs.

Given this diversity, tailored strategies per cluster are essential. For example, Cluster 4 customers should be engaged via loyalty programs and exclusive offers to maintain satisfaction. Cluster 3 customers may benefit from special offers and early access to premium features to encourage VIP status. Cluster 2 could respond well to incentives promoting more frequent or higher-value purchases, though resource allocation here should be balanced. Clusters 0 and 1 (purple and blue) contain low-engagement or churned customers; evaluating their Customer Lifetime Value (CLV) is critical to decide if reactivation efforts are cost-effective.

Overall, adding Diversity as a feature improved segmentation granularity and revealed a fifth meaningful cluster, aiding better business decisions that benefit both company and customers. Next, hierarchical clustering results will be examined to compare segmentation approaches.

5.2.2 Hierarchical Clustering (RFM and RFMD)

Similarly, hierarchical clustering using Ward's linkage was applied to RFMD features to uncover customer segments. Ward's method was chosen for its tendency to produce compact, spherical clusters and for its popularity in market segmentation. Unlike K-Means, hierarchical clustering produces a dendrogram illustrating the order in which customers and clusters merge, facilitating visual interpretation. The method was applied to both RFM and RFMD features to assess differences in cluster structures.

For RFM, Figure 49 in the appendix shows feature influences on principal components. PC1 is mainly influenced by Frequency and Monetary, while PC2 is dominated by Recency and Monetary. Thus, customers scoring high on PC1 resemble average buyers, whereas those with high PC2 scores tend to purchase infrequently but spend substantially each time. Interestingly, despite including Diversity, RFMD clustering showed a similar cluster organization.

Hierarchical clustering (Ward) identified 5 main clusters (Figure 48) with the following characteristics: - Cluster 4 (yellow), near the origin, contains customers low in all features, typically one-time buyers or disengaged customers who have churned or are at risk. - Cluster 2 (dark green) represents new customers with increasing Recency but low Frequency and Monetary scores. - Cluster 3 (green) comprises customers with higher purchase volume and spending but low Recency, suggesting seasonal or occasional buyers. - Cluster 0 (purple) includes customers developing loyalty, reflected in moderate feature increases and wider spread. - Cluster 1 (blue) shows high values in

all features, commonly loyal or VIP customers, though with notable cluster spread and some extreme outliers.

For RFMD features, the clustering structure resembles the RFM results (Figure 27), which is notable given the inclusion of Diversity. PC1 is influenced mainly by Frequency, Diversity, and Monetary, reflecting general purchasing patterns, while PC2 is governed by Recency and Monetary, capturing customer activity. Understanding these helps tailor retention strategies.

Despite differences in clustering methods, the resulting groups show strong consistency in structure and customer profiles, reinforcing cluster validity. The inclusion of Diversity as a fourth feature improves segmentation quality, offering richer insights for targeted marketing. For cluster references, see Figure 24.

5.2.3 Fuzzy C-Means Clustering (RFM and RFMD)

One of the other clustering algorithms used was Fuzzy C-Means, which differs fundamentally from K-Means and Hierarchical Clustering. The key idea behind this method is that data points can belong to multiple clusters simultaneously with varying degrees of membership. A modification exists to treat this method like K-Means, assigning hard labels so that each data point belongs to only one cluster. Given the multiple parameters to tune, we performed an iterative grid search to maximize the fuzzy partition coefficient and minimize fuzzy entropy. The resulting optimal parameters were: for RFM, $m = 1.5$, error = 0.005, and number of iterations = 500; and for RFMD, $m = 1.5$, error = 0.001, and number of iterations = 1000.

Similar to previous methods, we compared RFM and RFMD models. PCA was applied to reduce dimensionality and visualize the clusters in 2D (Figure 38). The left graph shows hard clustering labels, while the right depicts soft memberships. To interpret these clusters, it is important to understand the principal components as shown in Figure 51, where PC1 is influenced more by Frequency and Monetary, and PC2 by Recency.

Figure 38 shows three clusters: clusters 0 and 2 are close and condensed, while cluster 1 is more spread out. Overlapping clusters suggest similar RFM characteristics. Cluster 0 (lower left near origin) denotes customers who were loyal but are starting to decline. Cluster 1 (central) represents customers with moderate behavior across features, including some outliers with high metrics. Cluster 2 has increased Recency but low Frequency and Monetary, typically one-time buyers. The soft membership plot confirms that cluster 0's members have high membership values, supporting correct

assignment. Notably, no cluster shows high metrics associated with VIP customers, indicating a lack of successful loyalty strategies and an overall decline in engagement.

For RFMD features (Figure 16), PC1 is influenced by Frequency, Monetary, and Diversity, while PC2 is mainly influenced by Recency (Figure 29). The optimal number of clusters remained 3. Cluster 0 near the origin has very low values across features, representing both new and tenured customers who have not been effectively engaged. Cluster 1 shows higher Recency but similar values otherwise, associated with infrequent, one-time buyers. Cluster 2 shows improved behavior with higher Frequency and Diversity but still low Recency, indicating seasonal buyers and a small subset starting to exhibit VIP-like behavior. Soft membership boundaries are less distinct here than in RFM, suggesting that Diversity adds complexity and allows customers to belong partially to multiple segments based on their product variety.

5.2.4 DBSCAN Clustering (RFM and RFMD)

The last method tested was DBSCAN, which clusters dense regions and labels points in sparse areas as noise. Noise does not imply unimportance but rather that these customers are too unique to belong to any cluster. Parameter selection was challenging; a grid search identified optimal values for RFM as $\epsilon = 1.8$ and minimum samples = 3. Figure 52 visualizes the clusters using principal components (Figure 53), where PC1 is dominated by Frequency and Monetary, and PC2 by Recency. DBSCAN grouped most customers into a single large cluster near the origin, representing disengaged customers with low frequency, monetary value, and recency. Customers with moderate to high features were classified as noise, reflecting their unique behaviors that prevent clustering.

The RFMD results were similar (Figure 30 and Figure 31), with cluster 0 showing low engagement and higher proportions of noise points. This suggests that adding Diversity affects customer behavior and cluster membership. High-value customers were often labeled as noise, implying the need for further analysis (e.g., CLV) to determine if these "noisy" customers are worth retaining. DBSCAN's lack of clear segmentation limits its direct business utility, but it can be useful as a preprocessing step to isolate core customer groups before applying more detailed segmentation algorithms.

5.3 Model Evaluation and Comparison

To evaluate the clustering methods, we focused on Silhouette score, Calinski-Harabasz Index, and Davies-Bouldin Index (DBI). Tables 8 and 9 summarize RFM clustering

metrics. DBSCAN achieved the highest Silhouette score, indicating well-defined clusters. However, for metrics not computable for DBSCAN, K-Means performed best on Calinski-Harabasz and Hierarchical Clustering (Ward linkage) on DBI. This is supported by Figures 40 and 41. Overall, hierarchical clustering yielded the strongest results.

Tables 4 and 5 provide the corresponding metrics for RFMD. DBSCAN again dominated the Silhouette score, while K-Means led on Calinski-Harabasz, with fuzzy C-Means having the lowest and DBI best scores. Taking all metrics into account, K-Means stands out as the best performer, alongside hierarchical clustering with its consistent five-cluster solution, allowing richer segmentation. These results confirm that the RFMD approach outperforms traditional RFM by offering a more detailed understanding of customer behavior, which supports better business decision-making (see Figures 22 and 23).

5.4 Method Complementarity and Practical Insights

While the objective was to identify the most suitable clustering method, the results demonstrate that no single approach suffices for comprehensive customer segmentation. K-Means and Hierarchical Clustering (Ward linkage) produced similar, stable results with five meaningful clusters, reinforcing the presence of distinct customer segments. In contrast, Fuzzy C-Means revealed only three clusters, and DBSCAN found essentially one cluster plus noise without clear shape or interpretability.

Given the limitations of each method, a hybrid approach is advisable. For example, DBSCAN can first isolate unique or outlier customers as noise, while K-Means or Hierarchical Clustering can then segment the core customers more precisely. This layered strategy enables efficient resource allocation and targeted marketing tailored to the most valuable customer segments, ultimately improving retention and profitability.

5.5 Customer Lifetime Value by Cluster

It is important to acknowledge that this CLV model is far from perfect. It integrates a set of assumptions to make use of a discounted cash flow CLV formula. It sets $T=3$ as the expected lifespan of a customer, costs were taken as 20% of revenue, discount rate of 10% and an estimated customer acquisition cost of £100. The reason for choosing these values is just because

For RFM features, Table 10 presents the average CLV and number of customers per cluster. For instance, Cluster 1 in K-Means (1,062 customers, avg. CLV £217.06),

Cluster 3 in Hierarchical clustering (949 customers, avg. CLV £203.27), and Cluster 2 in Fuzzy C-Means (1,095 customers, avg. CLV £323.15) present the lowest values of CLV. This suggests that these are lost/disengaged customers that stopped buying products. It's alarming because the number of customers represents a high percentage in each cluster. It's often important to attack the problem from the root and not lose this amount of customers, which can often be changed into average customers.

Conversely, some clusters exhibit high average CLV values, often reaching close to £110,000. For example, Cluster 2 in K-Means has an average CLV exceeding £84,345.74 (13 customers), Clusters 1 and 2 in Hierarchical clustering reach £34,145.29 and £109,095.02 (7 and 8 customers respectively), Cluster 0 in Fuzzy C-Means reaches £63,442.26 (21 customers), and the noise cluster (-1) in DBSCAN reaches £64,323.26 (19 customers). Given the small number of customers in these clusters and their associated CLV, these groups likely consist of high-value individuals or companies purchasing in large volumes. While these segments highlight a significant share of business income, caution is needed because it can be dangerous to give them special care considering the number of customers. Sometimes, just because a customer brings the most amount of money doesn't mean all resources should be focused on them.

When evaluating customer importance from a business perspective, the average or standard customer provides the most reliable insight. This customer type, characterized by intermittent purchases, may represent significant revenue or not, depending on various factors. Cluster 0 in K-Means (211 customers, avg. CLV £8,158.49), Cluster 4 in Hierarchical clustering (175 customers, avg. CLV £9,200.05), Cluster 1 in Fuzzy C-Means (3,223 customers, avg. CLV £1,175.82), and Cluster 0 in DBSCAN (4,320 customers, avg. CLV £984.64) embody this average customer behavior with moderate value and high numbers. Notably, the average CLV in DBSCAN is higher because it consists of just one cluster grouping multiple customer profiles identified separately by other methods. Focusing retention and engagement efforts on these customers is essential, as many have the potential to become VIP clients with appropriate strategies.

Table 6 shows the average CLV and number of customers per cluster for RFMD features. Cluster 1 in K-Means (1,052 customers, avg. CLV £312.97), Cluster 3 and 5 in Hierarchical clustering (1,309 customers, avg. CLV £249.92 and 2,820 customers, avg. CLV £890.29), and Cluster 0 and 2 in Fuzzy C-Means (2,820 customers, avg. CLV £674.95 and 1,055 customers, avg. CLV £314.12), and Cluster 0 in DBSCAN (4,313 customers, avg. CLV £927.44). It's important to analyze the factors that influence the customers in these clusters because there are a lot of customers with low values

of CLV. Most of these customers represent lost/disengaged customers. Conversion of these customers into more average and stable ones is vital to generating higher revenue.

Several RFMD clusters present high average CLV values. For instance, Cluster 2 and 4 in K-Means have CLV values of £38,616.06 (6 customers) and £109,095.02 (8 customers), Cluster 1 and 2 in Hierarchical clustering reach CLV values of £149,589.69 (4 customers) and £47,723.54 (12 customers), and the noise cluster (-1) in DBSCAN presented average CLV of £56,759.32 (26 customers). Although the number of customers in these clusters is very little, they represent high-value individuals. Sometimes it can be just people buying expensive products, but sometimes are companies buying thousands of products which lead to these high CLV values. Again, it's important to understand what path to follow with these segments. It's important to not overspend in the measures implemented to keep these customers happy. In the majority of the cases, these customers are that happy that just by using the same tactics they will continue to be engaged.

Then there can be found the average customers, these are clusters in which customers have average behaviors. They buy products from time to time and are often influenced by various factors. Cluster 3 in K-Means (456 customers, avg. CLV £4,731.63), Cluster 4 in Hierarchical clustering (194 customers, avg. CLV £7,561.80), Cluster 0 in Fuzzy C-Means (2,820 customers, avg. CLV £6,984.99). In the majority of the cases, average clusters tend to concentrate the vast majority of the customers, but sometimes they just show average CLV values. Also, given that these values fluctuate, it's difficult to come up with the appropriate measures to keep these customers engaged in the business. Usually it's good to try different methods at the same time, but it's important to focus a lot on these—the high number of customers that can be easily engaged or disengaged.

Overall, RFMD clusters tend to show some groups with higher average CLV when compared to RFM. The inclusion of Diversity (D) in RFMD provides a deeper understanding of customer behavior that basic RFM might overlook. When choosing segmentation strategies, it is crucial to align with the business's main objectives and consider multiple clustering methods to uncover hidden segments. Ultimately, comprehending CLV facilitates more effective resource allocation. Although DBSCAN doesn't give much information on the clusters, it can be used to uncover hidden patterns.

5.6 Limitations

Despite favorable outcomes, several limitations must be acknowledged. First, the dataset lacked detailed customer information such as gender, age, and coupon or discount usage, which are valuable for tailoring more precise recommendations. Moreover, access to accurate customer acquisition cost (CAC) data would improve the precision of the CLV calculation. Consequently, the CLV values derived in this study are approximations rather than exact figures but remain useful for general insights.

Additionally, the modeling approach was basic in nature. For example, in hierarchical clustering, not all linkage methods were explored, which could have yielded improved segmentation results. Future work could focus on expanding feature sets, refining CLV models, and exploring additional clustering techniques or linkage methods to enhance accuracy and applicability.

6 Recommendations

Now that we have analyzed the results, here are some practical steps to improve customer retention and boost overall value. After examining the clustering methods, it is worth mentioning that although the goal was to identify the best algorithm for this project, the findings suggest that a hybrid approach can often yield more accurate and actionable results.

K-Means was selected as the best overall method. While its results were not perfect, it consistently showed solid performance across metrics such as Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index, and produced a representative number of clusters. This makes K-Means a reliable choice for customer segmentation. Hierarchical clustering also showed promising results, but its higher computational demands may make K-Means preferable for larger datasets.

Based on the customer profiles within each cluster and the CLV analysis, the following recommendations are made. Pareto analysis was implemented to help with the allocation of resources. As shown in Table ??, which describes the total CLV per cluster and Pareto analysis, cluster 3 and cluster 0 account for 72.5% of the total CLV, while cluster 4, 1 and 2 only account for 25.8% of the revenue. It's important to mention that given that cluster 4 has a high average, it has a small number of valuable customers.

Table 7: Total CLV and Pareto Analysis per Cluster (RFMD)

Cluster	Avg. CLV	Customers	Total CLV	% of Total	Cumulative %
3	4,731.63	456	2,156,623.28	38.7%	38.7%
0	668.96	2,817	1,883,681.52	33.8%	72.5%
4	109,095.02	8	872,760.16	15.7%	88.2%
1	312.97	1,052	329,032.44	5.9%	94.1%
2	38,616.06	6	231,696.36	4.2%	98.3%

- **Cluster 0 (Churned or Disengaged Customers):** It's important to use available resources in a meaningful way because customers in this cluster account for a big part of the overall revenue. Only use special offers on customers who show potential.
 - **Reminder messages:** Send emails with special offers and messages only if they haven't churned.
 - **Special offers:** Give disengaged customers special offers like discounts or free expedited shipping.
 - **Share new features:** Give them new things added to the site and let them know about newly highly purchased products which go along with their interests.
 - **Referral programs:** Word of mouth is the most powerful marketing tool, so offer them discounts or other advantages for each friend they refer. Give them codes that if entered provide them with advantages automatically.
 - **Behavior triggers:** Only reach out if they've visited the site recently or clicked on a message. No need to spend effort on fully inactive customers.
- **Cluster 1 (New Customers):**
 - **Welcome messages:** Give a small introduction on who the brand is and their values (advantages).
 - **First purchase:** Give discounts on similar products viewed. Give them a small survey in their account so that the site recommends products interesting to them.
 - **Referral programs:** Word of mouth is the most powerful marketing tool, so offer them discounts or other advantages for each friend they refer. Give them codes that if entered provide them with advantages automatically.
 - **Get familiar offers:** Suggest bestsellers or curated bundles so they find something they like without much effort.

- **Clusters 2 and 4 (Old Active Customers with VIP Potential):** Don't represent high values of CLV or revenue, so spend resources accordingly. Try to implement buy-back campaigns that don't require giving them high discounts.
 - **Loyalty programs:** Introduce loyalty programs to customers with extremely high CLV.
 - **Events:** Give them access to special events and offers to show them their importance.
 - **Referral programs:** Word of mouth is the most powerful marketing tool, so offer them discounts or other advantages for each friend they refer. Give them codes that if entered provide them with advantages automatically.
 - **Feedback:** Include them in decisions like the addition of certain features.
 - **Product nudges:** Push products from categories they don't usually buy from to increase variety.
- **Cluster 3 (Recently Reactivated Customers):** Represents high contribution to total CLV and revenue.
 - **Welcome messages:** Highlight the importance of them coming back and that we value them (e.g., "Thanks for coming back!").
 - **Loyalty programs:** Give them access to loyalty programs and explain the benefits of joining them.
 - **Offers:** Give them new discounts which will make them want to buy again.
 - **Ask why they returned:** Add a short optional pop-up or message asking what brought them back. Helps future campaigns.
 - **Time-limited perks:** Give them something like "10% off your next order if placed within 5 days" to keep them active.

7 Conclusion

7.1 Summary of Findings

This thesis explored customer segmentation using clustering algorithms applied to RFM and RFMD features, complemented by Customer Lifetime Value (CLV) analysis. The primary objective was to identify meaningful customer groups, deliver actionable marketing recommendations, and assess the effectiveness of different clustering methods to recommend one suitable for similar segmentation tasks.

The evaluated techniques included K-Means, Hierarchical Clustering (Ward linkage), Fuzzy C-Means (hard and soft), and DBSCAN. Performance was compared using metrics such as Calinski-Harabasz Index, Silhouette Score, and Davies-Bouldin Index. K-Means provided the best balance of performance, interpretability, and computational efficiency. While Hierarchical Clustering showed comparable results, its higher computational cost reduces practicality for large datasets. Fuzzy C-Means and DBSCAN exhibited some strengths but were less accurate overall.

All methods initially suggested two as the optimal number of clusters. However, given the business context, a larger number of clusters was chosen as it provided more meaningful customer segmentation. Ultimately, both K-Means and Hierarchical Clustering identified five distinct customer segments, each characterized by unique purchasing behaviors and engagement patterns. Tailored marketing actions were proposed to improve retention and profitability.

CLV was incorporated into the analysis, but the lack of certain parameters—such as precise costs, discount rates, and customer acquisition costs—made it so that we had to make assumptions based on the industry, which limited the accuracy of the CLV estimates.

In conclusion, the research showed that incorporating Diversity (D) into RFM features (resulting in RFMD) leads to more refined customer segmentation. Combining marketing frameworks with machine learning algorithms enhances the practical value of segmentation efforts.

7.2 Contributions to the Field

This research contributes to customer segmentation by introducing the use of the RFMD model, which remains underutilized due to its limited application in prior studies. Additionally, it demonstrates the value of comparing multiple clustering algorithms to reinforce segmentation validity and consistency. Finally, it provides actionable marketing recommendations tailored to the clusters identified, bridging the gap between analysis and business strategy.

7.3 Future Work

Future work could incorporate Market Basket Analysis (MBA) to improve recommendation quality. Currently, recommendations are primarily based on CLV, which is

limited by missing acquisition cost data and resulting CLV inaccuracies. MBA would enable the development of more automated, data-driven recommendation systems akin to those used by leading companies like Amazon, leveraging AI to deliver personalized product suggestions based on purchase histories.

Additionally, there can be included more complicated clustering algorithms which give better results overall.

References

- Abbas, W., Usman, M., & Qamar, U. (2022). Churn prediction of customers in a retail business using exploratory data analysis. *2022 International Conference on Frontiers of Information Technology (FIT)*, 130–135. <https://doi.org/10.1109/FIT57066.2022.00033>
- Abdulhafedh, A. (2021). Incorporating k-means, hierarchical clustering and pca in customer segmentation. *Journal of City and Development*, 3(1), 12–30.
- Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2022). The application of k-means clustering for province clustering in indonesia of the risk of the covid-19 pandemic based on covid-19 data. *Quality & Quantity*, 56(3), 1283–1291.
- Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 21(3), 527–570. <https://doi.org/10.1007/s10257-023-00640-4>
- Andriana, A. D., & Mardiani, G. T. (2025). Customer segmentation analysis in crm framework using rfm methods. *AIP Conference Proceedings*, 3200(1).
- Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, 12(1), 17–30.
- Blattberg, R. C., Getz, G., & Thomas, J. (2001). *Customer equity: Building and managing relationships as valuable assets*. Harvard Business School Press.
- Cambridge University Press. (n.d.). *Mass marketing*. Retrieved April 30, 2025, from <https://dictionary.cambridge.org/us/dictionary/english/mass-marketing>
- Chen, D. (2015). *Online retail* [UCI Machine Learning Repository]. <https://doi.org/10.24432/C5BW33>
- Chen, M. (2024). *Qu'est-ce que le big data ?* Retrieved January 12, 2025, from <https://www.oracle.com/fr/big-data/what-is-big-data/>
- Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). Rfm ranking – an effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, 33(10), 1251–1257. <https://doi.org/10.1016/j.jksuci.2018.09.004>
- Fader, P. S., Hardie, B. G., & Lee, K. L. (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>
- First Page Sage. (2025). *Average cac for ecommerce companies: 2025 edition*. Retrieved May 28, 2025, from <https://firstpagesage.com/reports/average-cac-for-ecommerce-companies/>

- GeeksforGeeks. (2021). *Ml: Fuzzy clustering*. Retrieved May 16, 2025, from <https://www.geeksforgeeks.org/ml-fuzzy-clustering/>
- GeeksforGeeks. (2023). *Dbscan clustering in ml: Density-based clustering*. Retrieved May 16, 2025, from <https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/>
- GeeksforGeeks. (2025a). *Elbow method for optimal value of k in kmeans*. Retrieved May 6, 2025, from <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>
- GeeksforGeeks. (2025b). *Hierarchical clustering in machine learning*. Retrieved May 6, 2025, from <https://www.geeksforgeeks.org/hierarchical-clustering/>
- GeeksforGeeks. (n.d.). *Clustering metrics*. Retrieved May 6, 2025, from <https://www.geeksforgeeks.org/clustering-metrics/>
- GeeksforGeeks contributors. (2025, January). *Dbscan clustering in ml — density based clustering*. Retrieved May 6, 2025, from <https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/>
- Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications*, 4(4).
- Gladys, N., Baesens, B., & Croux, C. (2009). Modeling churn and clv: Using a data mining approach. *International Journal of Bank Marketing*, 27(4), 274–291. <https://doi.org/10.1108/02652320910968344>
- Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 21(3), 527–570.
- Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162, 1–12. <https://doi.org/https://doi.org/10.1016/j.infsof.2023.107268>
- Google. (n.d.). *Algoritmos de agrupamiento en clústeres*. Retrieved January 12, 2025, from <https://developers.google.com/machine-learning/clustering/clustering-algorithms?hl=es-419>
- Hoque, E. M. J., Islam, M. S., & Mohtasim, S. A. (2024). Optimizing decision-making through customer-centric market basket analysis. *Journal of Operational and Strategic Analytics*, 2(2), 72–83.
- IBM. (n.d.). *K-means clustering: What it is and how it works* [Retrieved May 3, 2025]. <https://www.ibm.com/fr-fr/think/topics/k-means-clustering>
- Idowu, S., Annam, A., Rangarajan, E., & Kattukottai, S. (2019). Customer segmentation based on rfm model using k-means, hierarchical and fuzzy c-means clustering algorithms [no. August 2019]. *Jerarchical y Fuzzy C-Means*.

- Investopedia. (n.d.). *Market segmentation* [Investopedia]. Retrieved April 30, 2025, from <https://www.investopedia.com/terms/m/marketsegmentation.asp>
- Janardhanan, S., & Muthalagu, R. (2020). Market segmentation for profit maximization using machine learning algorithms. *Journal of Physics: Conference Series*, 1706(1), 012160.
- Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. *Informatics*, 5(1), 2.
- Kachroo, V. (2023). Customer segmentation and profiling for e-commerce using dbSCAN and fuzzy c-means. *Proceedings on Engineering*, 5(3), 539–544.
- Lewaaelhamd, I. (2024). Customer segmentation using machine learning model: An application of rfm analysis. *Journal of Data Science and Intelligent Systems*, 2(1), 29–36.
- Lim, T. (2021). K-means clustering-based market basket analysis: Uk online e-commerce retailer. *ICIT*, 126–131.
- Ling, L. S., & Weiling, C. T. (2025a). Enhancing segmentation: A comparative study of clustering methods. *IEEE Access*.
- Ling, L. S., & Weiling, C. T. (2025b). Enhancing segmentation: A comparative study of clustering methods. *IEEE Access*, 13, 47418–47439. <https://doi.org/10.1109/ACCESS.2025.3550339>
- Maraghi, M., Adibi, M. A., & Mehdizadeh, E. (2020). Using rfm model and market basket analysis for segmenting customers and assigning marketing strategies to resulted segments. *Journal of Applied Intelligent Systems and Information Sciences*, 1(1), 35–43.
- MasterClass. (2021). *Mass marketing explained: Definition, strategies, and examples*. Retrieved April 30, 2025, from <https://www.masterclass.com/articles/mass-marketing>
- Molaei, R., Abbasimehr, H., & Rahsepar Fard, K. (2025). Proposing a new framework based on the rfm model and multivariate time series for customer segmentation and behavior analysis: A case study of a food industry company. *Sciences and Techniques of Information Management*.
- Musthofa Pradana, H. H. (2021). Maximizing strategy improvement in mall customer segmentation using k-means clustering. *Journal of Applied Data Sciences*, 2(1), 19–25. <https://doi.org/10.47738/jads.v2i1.18>
- Nair, H., & Acharya, A. (2006). *Customer lifetime value: Models and applications* (tech. rep.). UCLA Anderson School of Management. Retrieved May 28, 2025, from [https://www.anderson.ucla.edu/documents/areas/fac/marketing/JSR2006\(0\).pdf](https://www.anderson.ucla.edu/documents/areas/fac/marketing/JSR2006(0).pdf)

- Noble, J. (2024). *What is hierarchical clustering?* Retrieved May 5, 2025, from <https://www.ibm.com/think/topics/hierarchical-clustering>
- Nugraha, A., Effendi, Y. A., Nicholas, N., Tao, Z., Afifuddin, M., & Nuzulita, N. (2025). K-means clustering interpretation using recency, frequency, and monetary factor for retail customers segmentation. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 23(2), 435–446.
- Paramita, A. S., & Hariguna, T. (2024). Comparison of k-means and dbscan algorithms for customer segmentation in e-commerce. *Journal of Digital Market and Digital Currency*, 1(1), 43–62.
- Paranavithana, I. R., Rupasinghe, T. D., & Prior, D. D. (2021). Unsupervised learning and market basket analysis in market segmentation. *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering*.
- Patterson-Cross, R. B., Levine, A. J., & Menon, V. (2021). Selecting single cell clustering parameter values using subsampling-based robustness metrics. *BMC Bioinformatics*, 22(1), 39. <https://doi.org/10.1186/s12859-021-03957-4>
- Pfeifer, P. E., & Carraway, R. L. (2000). Modeling customer relationships as markov chains. *Journal of Interactive Marketing*, 14(2), 43–55. [https://doi.org/10.1002/\(SICI\)1520-6653\(200021\)14:2<43::AID-DIR4>3.0.CO;2-W](https://doi.org/10.1002/(SICI)1520-6653(200021)14:2<43::AID-DIR4>3.0.CO;2-W)
- Piech, C. (n.d.). *K-means* [Based on a handout by Andrew Ng. Retrieved May 3, 2025]. <https://stanford.edu/~cpiech/cs221/handouts/kmeans.html>
- Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, 4, 145–154.
- Qualtrics. (2020). *Customer segmentation analysis: Definition & methods - qualtrics* [Qualtrics]. Retrieved April 26, 2025, from <https://www.qualtrics.com/experience-management/brand/customer-segmentation/>
- Qualtrics. (2022). *What is customer lifetime value (clv) and how can you increase it?* Retrieved May 5, 2025, from <https://www.qualtrics.com/experience-management/customer/customer-lifetime-value/>
- Ramkumar, G., Bhuvaneswari, J., Venugopal, S., Kumar, S., Ramasamy, C. K., & Karthick, R. (2025). Enhancing customer segmentation: Rfm analysis and k-means clustering implementation. In *Hybrid and advanced technologies* (pp. 70–76). CRC Press.
- SAP. (n.d.). *¿qué es machine learning?* Retrieved January 12, 2025, from <https://www.sap.com/latinamerica/products/artificial-intelligence/what-is-machine-learning.html>

- Scikit-learn developers. (2024). *Dbscan — scikit-learn 1.3.0 documentation*. Retrieved May 16, 2025, from <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>
- scikit-learn developers. (2024). *Sklearn.cluster.dbscan — scikit-learn 1.4.2 documentation*. scikit-learn. Retrieved May 6, 2025, from <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>
- Selim, M. (2022). The effect of customer analytics on customer churn. *Unpublished*.
- Sharma, P. (2020, October). *Quick guide to evaluation metrics for supervised and unsupervised machine learning*. Retrieved May 6, 2025, from <https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/>
- Siti Monalisa, R. N., Putri Nadya. (2019). Analysis for customer lifetime value categorization with rfm model [The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia]. *Procedia Computer Science*, 161, 834–840. <https://doi.org/https://doi.org/10.1016/j.procs.2019.11.190>
- Sitorus, E. R., & Nugraha, I. (2025). Customer segmentation analysis with rfm model (recency, frequency, monetary) and k-means clustering: Case study of bottled water sales at pt xyz. *Jurnal Serambi Engineering*, 10(2).
- Stormi, K., Lindholm, A., Laine, T., & Korhonen, T. (2020). Rfm customer analysis for product-oriented services and service business development: An interventionist case study of two machinery manufacturers. *Journal of Management and Governance*, 24(3), 623–653.
- SurveyMonkey Inc. (2024). *Customer segmentation: A complete guide*. Retrieved April 30, 2025, from <https://www.surveymonkey.com/market-research/resources/the-complete-guide-to-customer-segmentation/>
- TechTarget Contributor. (2023). *Market basket analysis*. Retrieved May 3, 2025, from <https://www.techtarget.com/searchcustomerexperience/definition/market-basket-analysis>
- The Wharton School. (2022). *Customer lifetime value: What it is and why it matters*. Retrieved May 5, 2025, from <https://online.wharton.upenn.edu/blog/why-customer-lifetime-value-matters/>
- Uddin, M. A., Talukder, M. A., Ahmed, M. R., Khraisat, A., Alazab, A., Islam, M. M., Aryal, S., & Jibon, F. A. (2024). Data-driven strategies for digital native market segmentation using clustering. *International Journal of Cognitive Computing in Engineering*, 5, 178–191.
- Vanham, P. (2019, January). *A brief history of globalization*. Retrieved April 30, 2025, from <https://www.weforum.org/stories/2019/01/how-globalization-4-0-fits-into-the-history-of-globalization/>

- Walter, Y. (2024, April). *Mastering modern marketing strategies for today's business landscape* [marketing strategies]. Retrieved April 30, 2025, from <https://www.forbes.com/councils/forbesbusinesscouncil/2024/04/17/mastering-modern-marketing-strategies-for-todays-business-landscape/>
- Wulansari, S., & Heikal, J. (2024). Analysis of customer segmentation in the top three most visited e-commerce platforms in indonesia in 2023 using rfm model and clustering techniques. *Jurnal Scientia*, 13(03), 1164–1174.
- Xiahou, X., & Harada, Y. (2022). B2c e-commerce customer churn prediction based on k-means and svm. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(2), 458–475. <https://doi.org/10.3390/jtaer17020024>
- Yankelovich, D., & Meer, D. (2006). Rediscovering market segmentation [segmentation]. *Harvard Business Review*. Retrieved April 30, 2025, from <https://hbr.org/2006/02/rediscovering-market-segmentation>
- Yenigün, O. (2024, March). *Dbscan clustering algorithm demystified*. Retrieved May 6, 2025, from <https://builtin.com/articles/dbSCAN>

8 Annexes

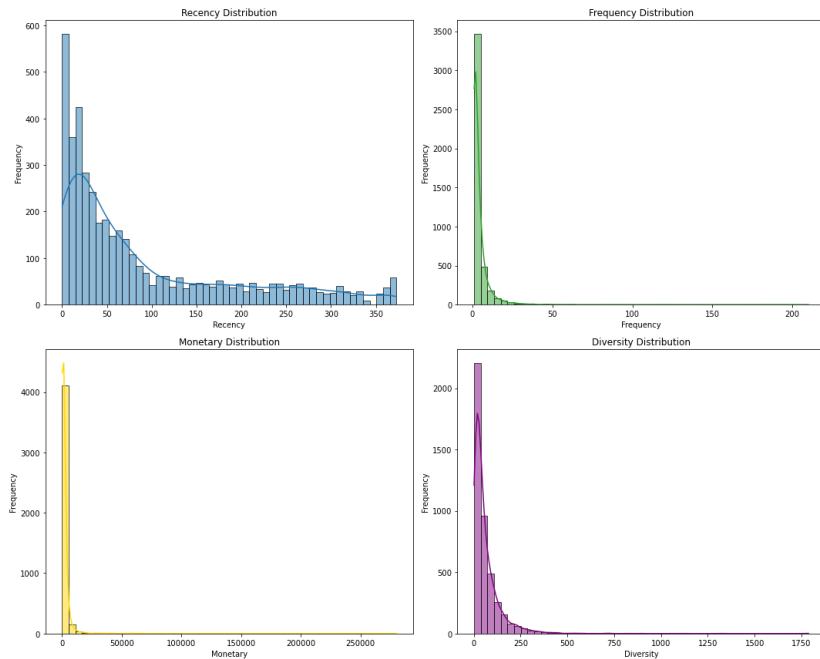


Figure 34: Histograms for Recency, Frequency, Monetary, and Diversity

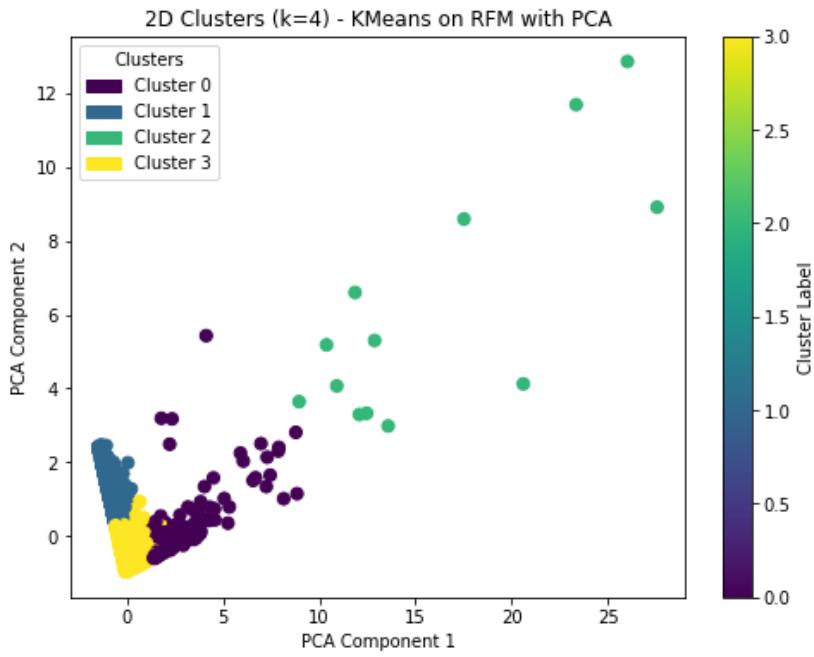


Figure 35: K-Means Clustering with 4 Clusters on RFM (PCA 2D View)

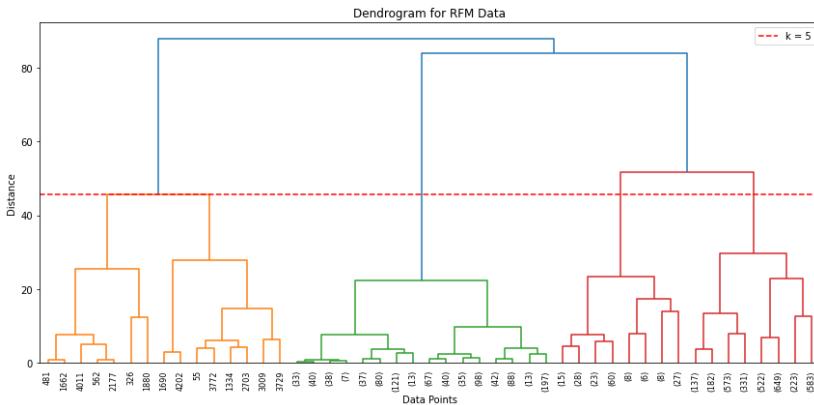


Figure 36: Dendrogram of Hierarchical Clustering (Ward Method) on RFM Data

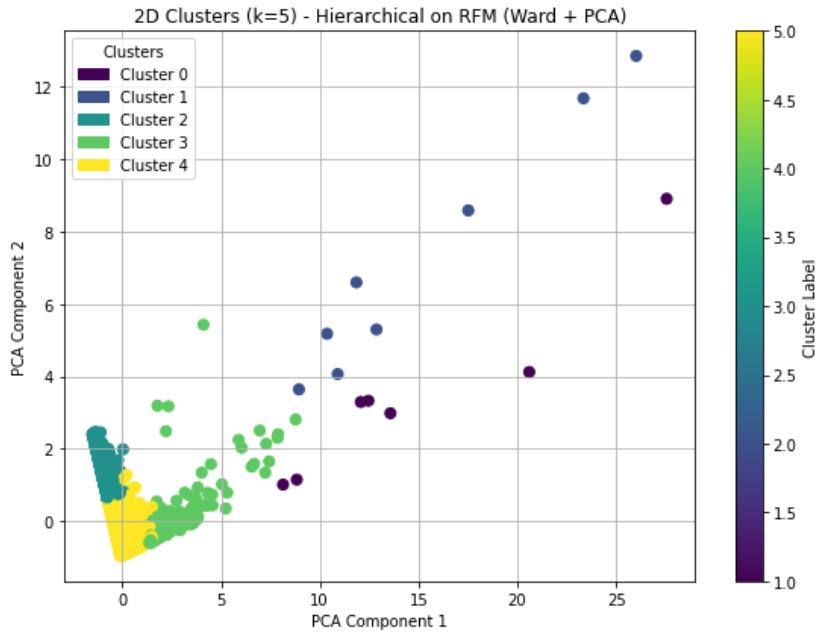


Figure 37: Hierarchical Clustering (Ward) with 4 Clusters on RFM (PCA 2D View)

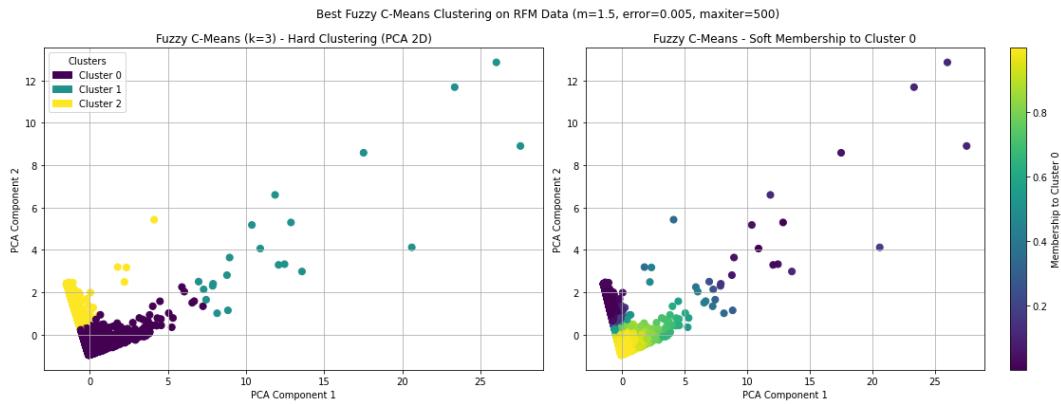


Figure 38: Fuzzy C-Means Clustering on RFM: Hard vs Soft Assignments (PCA 2D)

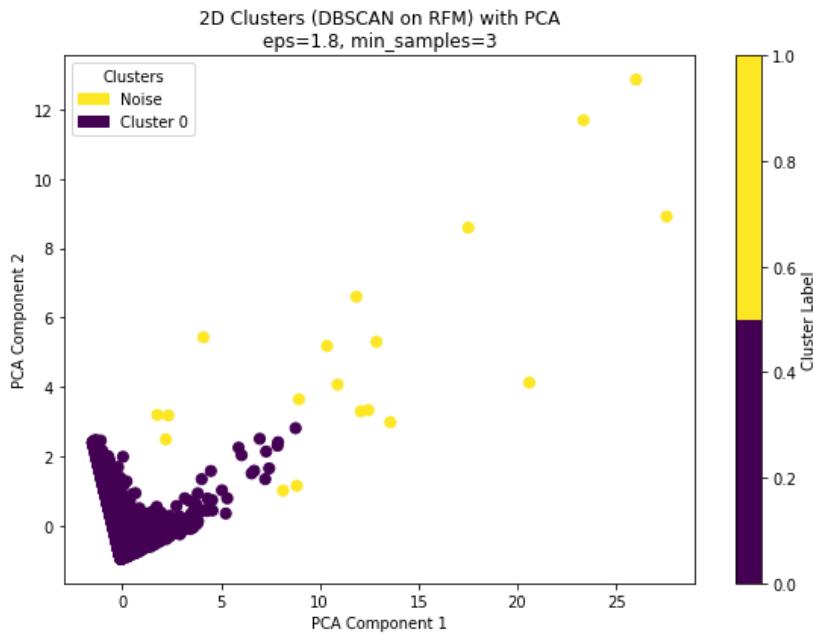


Figure 39: DBSCAN Clustering on RFM (PCA 2D View)

Table 8: Clustering Performance Metrics for RFM (excluding DBSCAN)

Method	Silhouette	Calinski-Harabasz	Davies-Bouldin	Optimal k
K-Means	0.6161	3149.99	0.7524	4
H. Clustering (Ward)	0.6151	2938.83	0.8068	4
Fuzzy C-Means	0.5905	3002.45	0.700	3

Table 9: DBSCAN Performance Metrics for RFM

Method	Epsilon (eps)	Min Samples	Silhouette
DBSCAN	1.80	3	0.9063

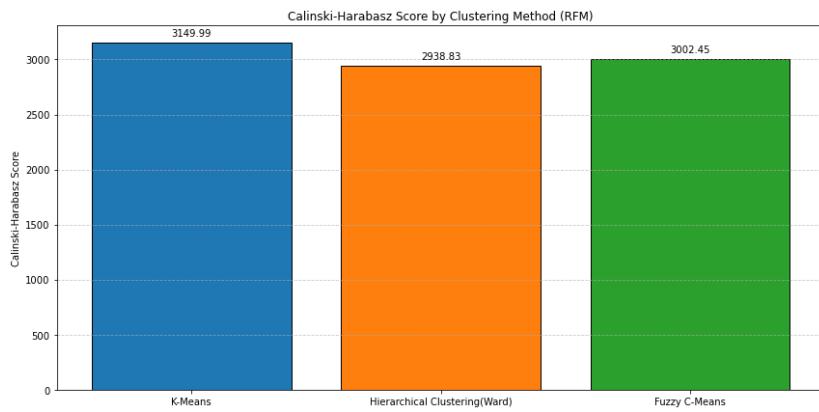


Figure 40: Calinski-Harabasz scores for K-Means, Hierarchical, and Fuzzy C-Means (RFM)

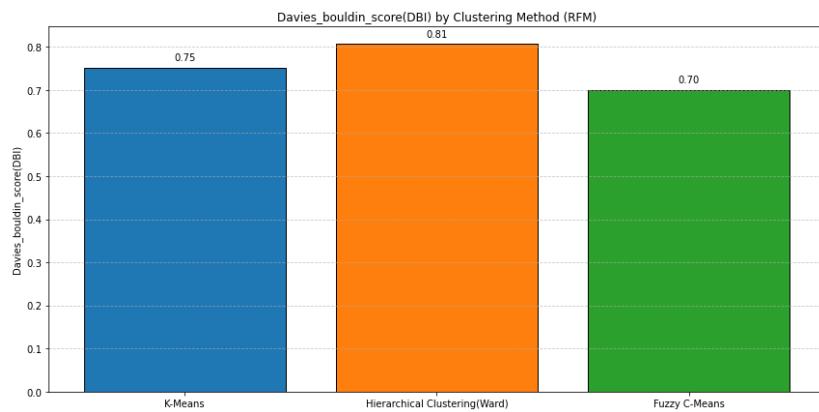


Figure 41: DBI scores for K-Means, Hierarchical, and Fuzzy C-Means (RFM)

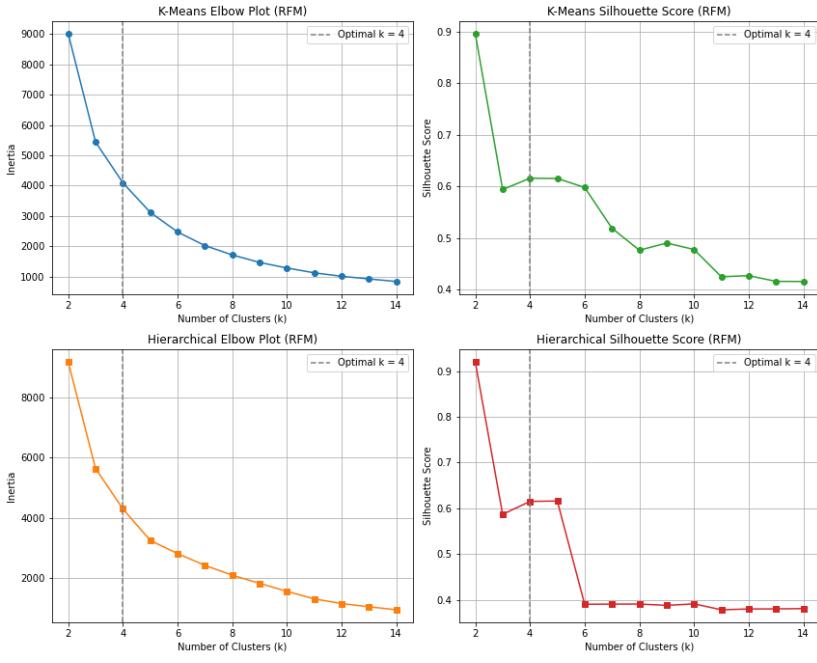


Figure 42: Elbow method and silhouette score K-Means and Hierarchical clustering(RFM)

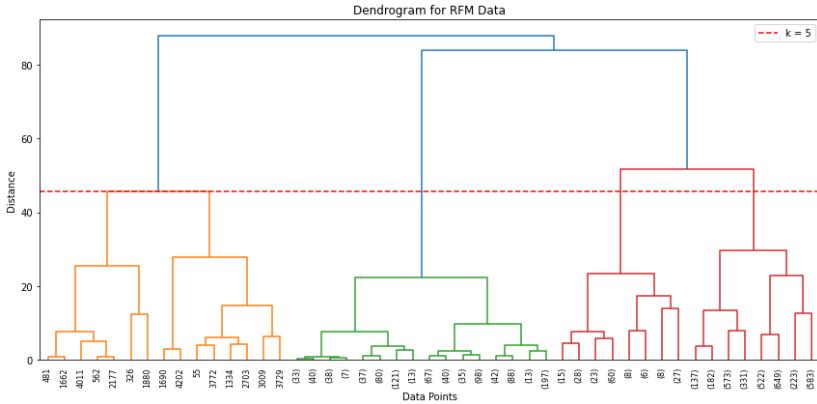


Figure 43: Dendrogram Hierarchical Clustering (RFM)

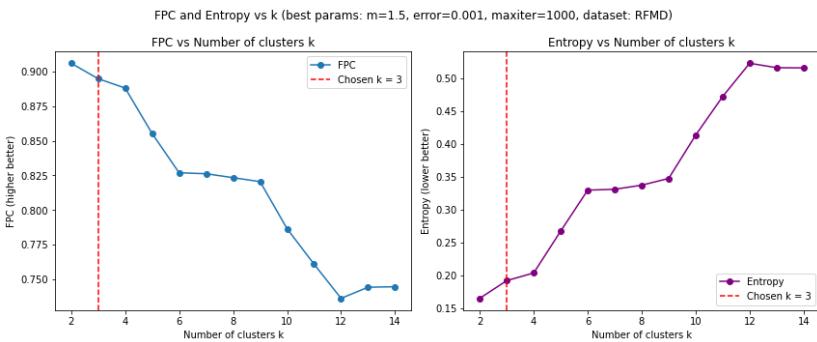


Figure 44: FPC and entropy Fuzzy C-Means (RFM)

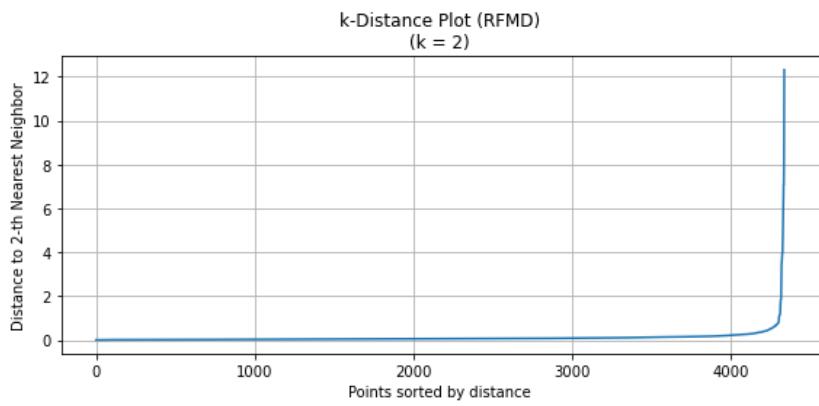


Figure 45: k-distance plot (RFM)

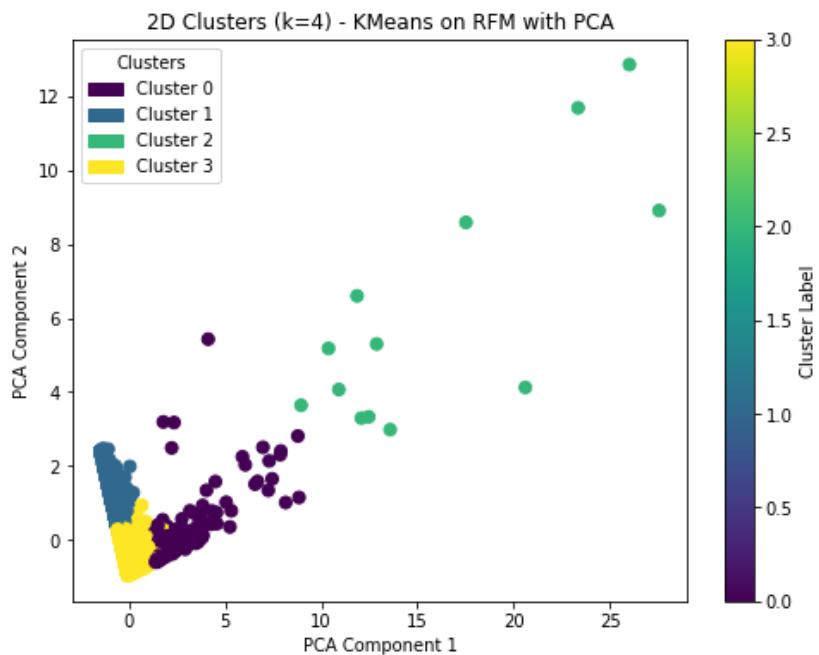


Figure 46: K-Means clusters visualized using PCA (RFM features)

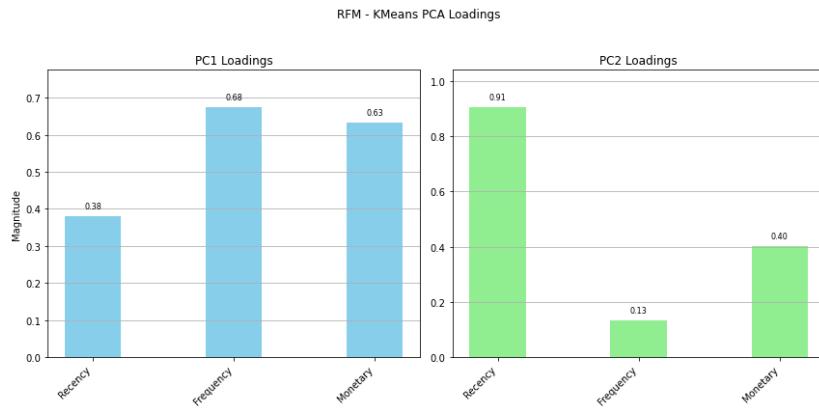


Figure 47: PCA component loadings of RFM features used in K-Means clustering

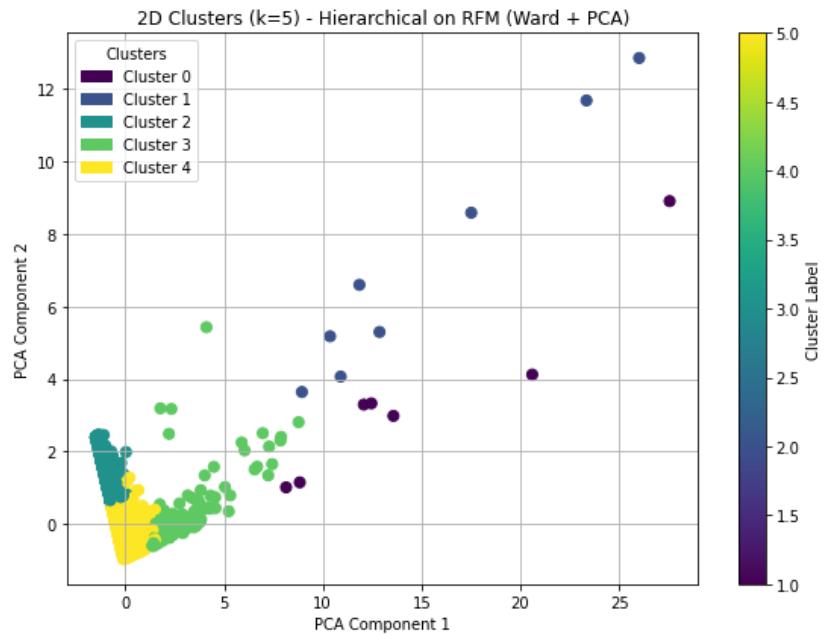


Figure 48: Hierarchical clusters visualized using PCA (RFM features)

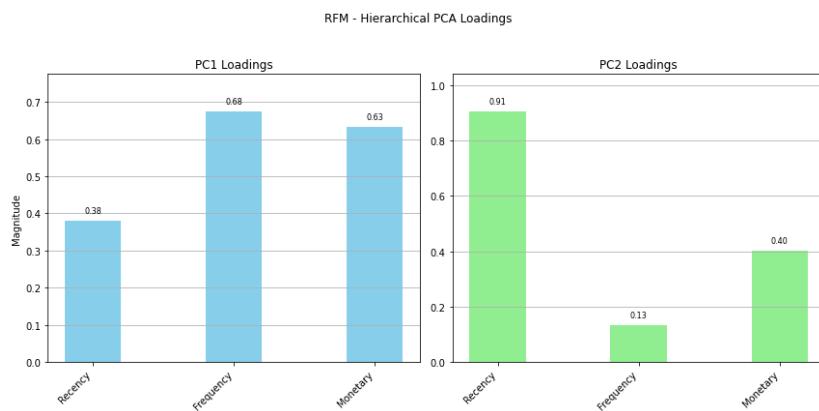


Figure 49: PCA component loadings of RFM features used in Hierarchical clustering

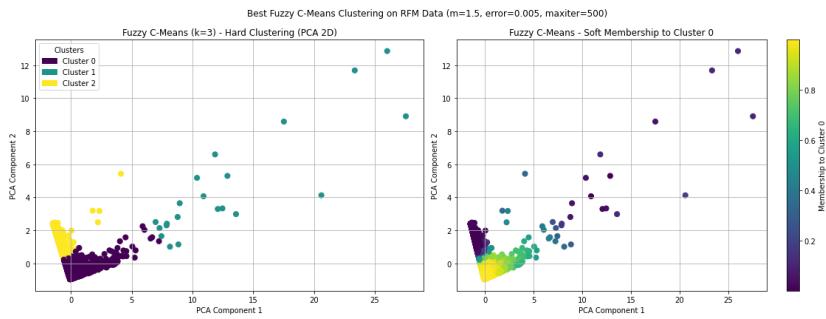


Figure 50: Fuzzy C-Means clusters visualized using PCA (RFM features)

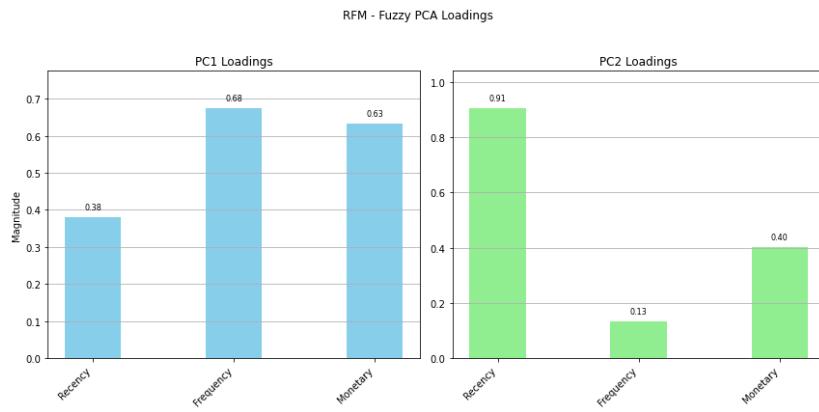


Figure 51: PCA component loadings of RFM features used in Fuzzy C-Means

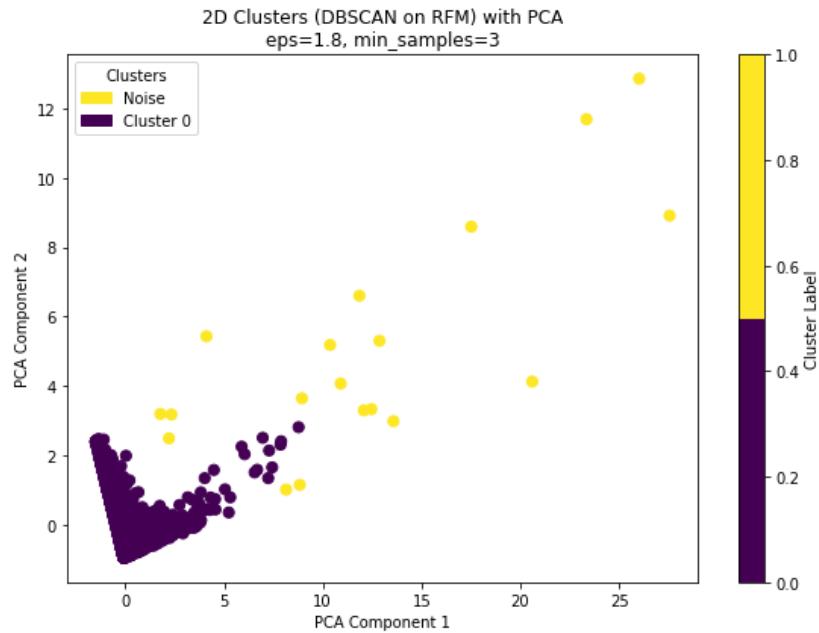


Figure 52: DBSCAN clusters visualized using PCA (RFM features)

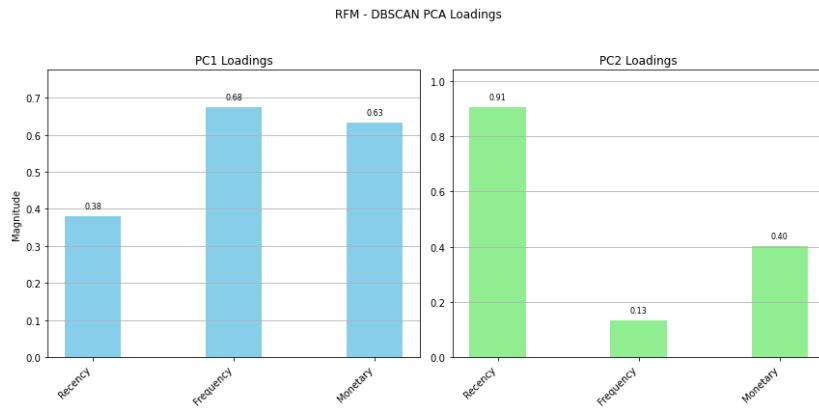


Figure 53: PCA component loadings of RFM features used in DBSCAN

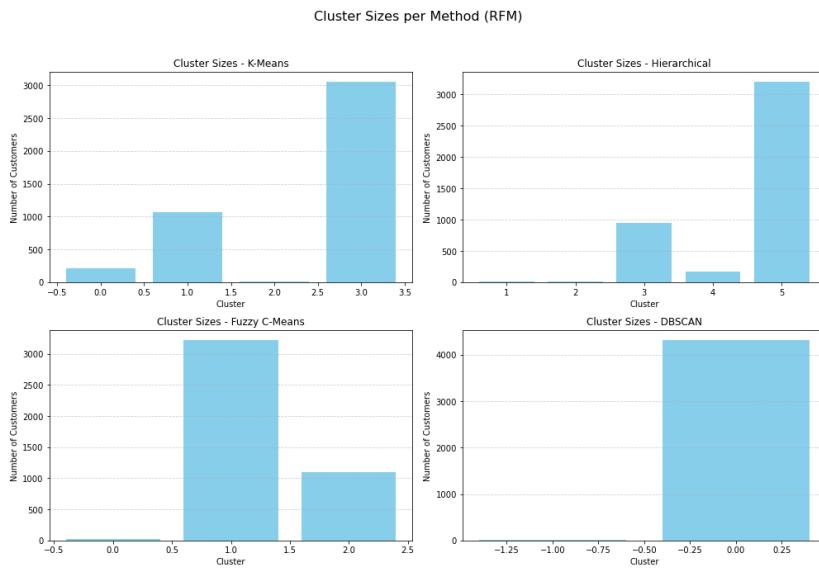


Figure 54: Number of Customers per Cluster per Method (RFM)

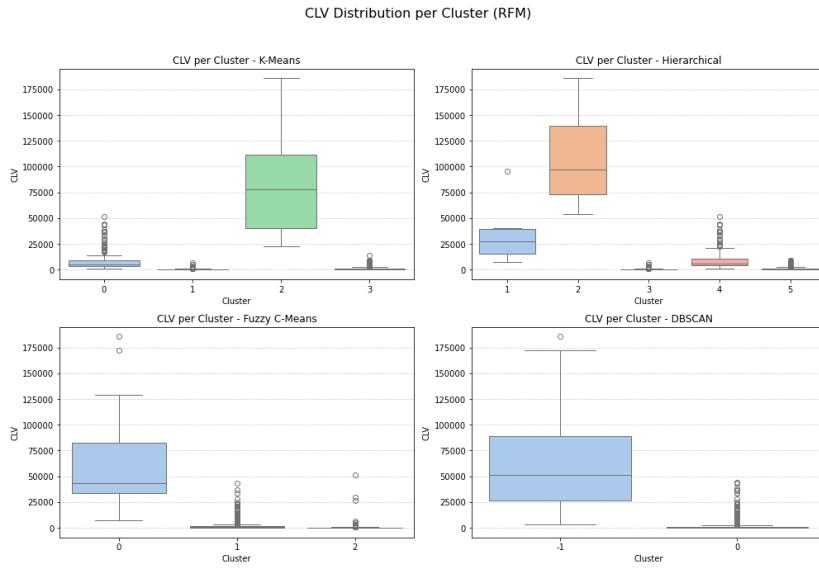


Figure 55: CLV Distribution per Cluster per Method (RFM)

Table 10: Average CLV and Number of Customers per Cluster (RFM)

Method	Cluster	Average CLV	Number of Customers
K-Means	0	8,158.49	211
	1	217.06	1062
	2	84,345.74	13
	3	795.07	3053
Hierarchical	1	34,145.29	7
	2	109,095.02	8
	3	203.27	949
	4	9,200.05	175
	5	800.35	3200
Fuzzy C-Means	0	63,442.26	21
	1	1,175.82	3223
	2	323.15	1095
DBSCAN	-1	64,323.26	19
	0	984.64	4320

Appendix: Sample Prompts Used with ChatGPT

Below is a list of example prompts used during the coding and analysis phases of this thesis:

1. How can I optimize this Python code for better performance with large datasets?

2. Can you help me debug this function that processes the RFMD table? It's throwing an error.
3. What are common pitfalls when working with Polars dataframes, and how do I avoid them?
4. How can I improve the readability and structure of this clustering pipeline code?
5. What debugging strategies should I use to track down a bug in my clustering results?
6. How can I create dynamic, interactive plots in Python to visualize cluster distributions?
7. Can you help me improve the appearance of my bar plots for cluster sizes?
8. What libraries can I use to make my CLV boxplots more insightful and user-friendly?
9. How do I add labels and legends dynamically to matplotlib plots based on cluster data?
10. Can you suggest ways to automate the plotting process for multiple clustering methods?
11. Can you provide a Python implementation of Fuzzy C-Means clustering that works with my dataset?
12. How do I adapt the Fuzzy C-Means algorithm from a web example to my specific RFMD data?
13. Can you help me implement DBSCAN clustering and handle noise points properly?
14. What are best practices for tuning DBSCAN parameters on my customer data?
15. How can I integrate Fuzzy C-Means and DBSCAN results with other clustering outputs for comparison?
16. How do I convert my clustering scripts into reusable Python functions?
17. Can you help me write a function that takes a dataset and clustering parameters, and returns cluster labels?
18. What's a clean way to organize clustering functions and plotting utilities in a single Python file?

19. How can I make my cluster analysis pipeline modular so I can easily switch methods?
20. Can you help me automate evaluation metrics calculation within a clustering function?
21. What's the best way to document and comment my clustering code for reproducibility?
22. How can I efficiently handle missing or inconsistent data before clustering?
23. Can you help me merge cluster labels from different algorithms into a combined table?
24. How do I export clustering results and summary statistics into clean tables for LaTeX reports?
25. Can you help me write a script that summarizes customer counts and average CLV per cluster?
26. I'm going to paste part of my thesis. I want you to correct only basic grammar, punctuation, and spelling mistakes, but keep my original tone and sentence structure. Do not rewrite or rephrase my sentences, and don't make them sound like AI or formal academic language. If something is a bit awkward but understandable, leave it as it is — I want it to feel like something I wrote.

I'm writing this in LaTeX, so you might see LaTeX syntax like % for percentages or commands. Do not modify any LaTeX commands or formatting.