National College of Ireland

# Cluster-Based Customer Segmentation for Targeted Marketing in Retail Using Unsupervised Learning Algorithms

MSc Research Project

MSc In Data Analytics

Leela Sri Venkata Sai Narayana Vakacherla

Student ID: X23246383

School of Computing

National College of Ireland

Supervisor: Bharath Agarwal

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student Name:** Leela Sri Venkata Sai Narayana Vakacherla

**Student ID:** X23246383

**Programme:** Msc In Data Analytics **Year:** 2025-2026

**Module:** Research Practicum

**Supervisor:** Bharath Agarwal

**Submission Due Date:** 28/01/2026

**Project Title:** **Cluster-Based Customer Segmentation for Targeted Marketing in Retail Using Unsupervised Learning Algorithms**

**Word Count:** ***6266***

**Page Count** ……..…….

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** ……
……

**Date:** ……
……

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on a computer. | □ |

Assignments that are submitted to the Programme Coordinator's Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Cluster-Based Customer Segmentation for Targeted Marketing in Retail Using Unsupervised Learning Algorithms

Leela Sri Venkata Sai Narayana Vakacharla

X23246383

**Abstract**

The research uses unsupervised machine learning to categorise retail customers with Recency, Frequency, and Monetary (RFM) data. A Kaggle marketing dataset was cleaned, transformed and modelled with K-Means, DBSCAN, Hierarchical clustering and Gaussian Mixture Model. The models were assessed using a multi-metric model such as the Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index. The results indicate that K-Means has the most differentiated and consistent clusters that are entrenched with the highest scores of separations, and DBSCAN is not successful because of the density constraints. The study is valuable as it bridges the gaps in the single-metric assessment and proves the power of deep segmentation as the source of targeted marketing and customer relationship options.

# 1. Introduction

## 1.1 Introduction to the Research Topic

Customer segmentation is one of the key aspects of contemporary retail strategy since it assists companies in learning about behavioural variations among consumers. Nonetheless, the ancient segmentation techniques fail to comprehend the huge and intricate data. Whilst RFM (Recency, Frequency, Monetary) analysis is an effective behavioural summary, it will not necessarily give insight into meaningful groups in the absence of more advanced methods of analysis (Akande et al., 2024). This has created a need to explore further machine learning-based methods of clustering, which are capable of detecting latent trends and facilitating individualised marketing at scale. The paper is thus research on ways in which unsupervised algorithms may be used to enhance the quality of segmentation in retail settings. The strong segmentation is a direct benefit to retail decision-making, enhancing the targeting of the loyalty programme, offering personalised offers and reinforcing customer retention initiatives. It also assists in determining churn-prone groups so that retailers can commit marketing resources in a more effective and efficient way.

## 1.2 Problem Statement

Customers are becoming more important to retailers, and it is clear that noisy transactions, behavioural heterogeneity and weaknesses of traditional analytical tools continue to complicate the process of accurately segmenting customers. Most of the current segmentation practices have simplified rules or single metric-based assessments, and this leaves the customer groups unstable or incomplete. Further, various clustering algorithms yield different results in regard to the structure of the data. It is not clear which algorithm is most appropriate in RFM-based segmentation. Thus, a comparative study of algorithms, K-Means, DBSCAN, Hierarchical, and GMM, as well as their application on several performance indicators to determine which method offers the most credible segmentation to apply to retail marketing tasks, is required (Ufeli et al., 2025).

## 1.3 Research Aim, Questions, and Objectives

***Aim***

The goal of the study is to find out which of these unsupervised clustering algorithms, K-Means, DBSCAN and Hierarchical clustering results in the improvement of customer segmentation for retail marketing using RFM.

***Questions***

1. What is the comparison between DBSCAN, K-Means, and Hierarchical Clustering algorithms when it comes to segmentation of retail customers based on “RFM (Recency, Frequency, Monetary)” variables?
2. What are the strengths and weaknesses of using multiple performance metrics, such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, for evaluating clustering algorithms in retail customer segmentation?
3. What are the advantages of using customer segmentation models, which are founded on unsupervised learning algorithms, to support targeted marketing in the retail industry?

***Objectives***

- To address gaps in existing literature, where clustering models are often evaluated using a single metric, by applying multi-metric validation.
- To compare the performance of K-Means, DBSCAN, and Hierarchical Clustering using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.
- To recommend the most effective clustering algorithm for customer segmentation in retail marketing.
- To improve the precision of customer segmentation by utilising RFM features for dynamic segmentation models.

## 1.4 Justification and Significance of the Study

The majority of available literature assesses the models of clustering through a single measure that does not provide good reliability. Not so many of them compare several algorithms, and even fewer say why some of the models perform well in RFM data. Clustering results are rarely related to the actual marketing behaviour of research. This research fills this gap by using multi-metric assessment and segmentation interpretation as applied in practice. The results will help improve retail marketing because businesses will be able to streamline their marketing strategies on the basis of more precise data-driven customer insights, which will increase engagement and profitability.

## 1.5 Research Scope and Limitations

The study works with the Kaggle retail marketing dataset in order to produce the RFM features and implement four clustering algorithms are K-Means, DBSCAN, Hierarchical, and GMM. Silhouette, Davies-Bouldin, and Calinski-Harabasz indices are used to determine the model performance. Findings show that K-Means gives the most stable and well-separated customer groups. It is narrowed down to the multi-metric validation of these three algorithms. Certain weaknesses are that the data quality can be problematic, such as the absence of values, and that the research is based on a single dataset, which can influence the extrapolation of the results to other retail settings.

## 1.6 Overview of the Dissertation Structure

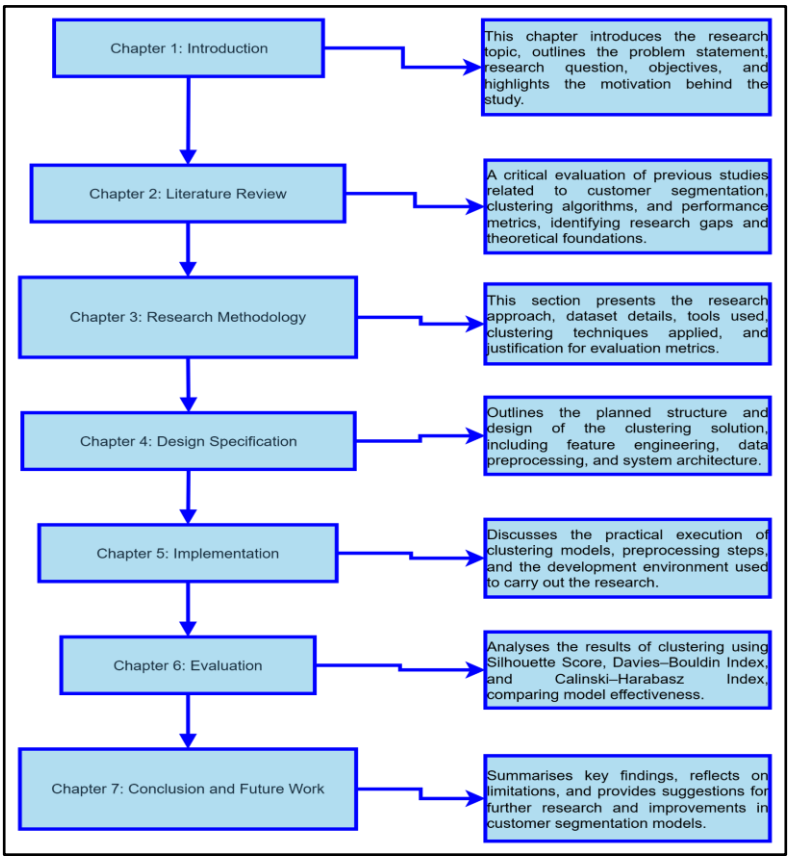



**Figure 1: Structure of Dissertation**

# 2. Related Work

## 2.1 Introduction

This literature review of customer segmentation in retail items, from AI-driven clustering to traditional demographic models as the pivot of consideration. This research looks at unsupervised learning, dynamic frameworks based on RFM, clustering results and retail advertisement strategies.

## 2.2 Development of customer segmentation on Retail Marketing

The concept of customer segmentation has gone beyond the demographic frameworks and turned into its more dynamic and data-driven alternatives made possible by artificial intelligence and big data. Early demographic segmentation was restrictive when consideration of behavioural difference and heterogeneity was needed, and transaction-based models emerged as dynamic emerged. By examining vast behavioural and transactional patterns rather than using predefinitions, Cordova (2024) revealed the improved accuracy of segmentation with the help of big data analytics. Even though the study noted an enhanced targeting, it was premised on high-level computing provisions and the non-uniformity of information quality that restricted its application in the real world.

As an extension of behavioural segmentation, Putri et al. (2024) used K-Means clustering on the variables of RFM and labelled six significant behavioural clusters. The results of their work proved that the RFM-based clustering can support such interventions as those of loyalty enhancement and reactivation of customers. Nevertheless, the research used solely one internal measure, that was limiting the strength of the results and limited generalisability in retail-relevant situations. This indicates a common loophole in the ongoing research work: segmentation models have not been subjected to complete performance analysis since they are commonly tested with the help of a single measure.

In another perspective, Tavor et al. (2023) investigated segmentation through the lenses of the economic and behavioural view, through modelling the impact of loyal and deal-prone customers on levels of pricing and revenue. Though conceptually sound, the model was not empirically validated and did not consist of machine learning approaches, which did not make it useful in modern retail analytics.

Through these studies, however, it becomes evident that retail segmentation has shifted from descriptive demographic modellers to behaviourally driven and AI-enhanced studies. Nevertheless, there is still the issue

of no comparative algorithms assessment and multi-metric validation. This supports the necessity of the proposed research, where the different clustering algorithms are studied based on a variety of internal measures to present a more informative perspective of the performance of the segmentation in retail contexts.

## 2.3 Unsupervised Machine Learning Methods and International Metrics of Evaluation (Customer Segmentation)

Unsupervised learning has played a great role in enhancing retail segmentation by enabling the analysts to find out naturally known groups of customers without any classification. As it was evidenced by Hicham and Karim (2022), several algorithms, namely DBSCAN, K-Means, Minibatch K-Means, and Mean Shift, are useful individually, although an ensemble clustering method proves beneficial. Using ARI, NMI, Dunn Index, and Silhouette Score, their performance was good with varying customer behaviours. The ensemble model was, however, resource-heavy and computationally intensive and demanded large datasets and heavy parameter tuning; thus, it could not be applied widely in retail applications.

Compared to K-Means, Rodrigues et al. (2025) simulated DBSCAN on retail transaction data and demonstrated that the first one was better at modelling irregular cluster shapes and noises that can be observed. Nevertheless, K-Means worked more effectively during the homogenous distribution of customer behaviour and formed more cohesion and separation. Their results emphasised that the performance of the algorithm is strongly contingent on the structure of data, supporting the necessity to test the models under different conditions. It, however, relied on only two algorithms and two metrics, with open questions as to whether there is diversity in performance between larger sets of algorithms.

Furthering this claim, John et al. (2023) measured five algorithms: K-Means, DBSCAN, Agglomerative Clustering, GMM, and BIRCH with respect to RFM-based retail data and various validation measures. Their findings indicated that GMM performed better than the other models since it dealt with overlapping customer behaviours. However, the research recognised the sensitivity to the data dimensions and computing costs. The present comparison allows emphasising the significance of incorporating a probabilistic model, such as GMM, in segmentation studies.

In these studies, there exist three gaps in methodology; most of them tend to concentrate on the performance of algorithms in isolation, instead of the comparative analysis. Some of them are based on one validation measure, which restricts the strength of inferences. There are not many studies that clarify why some algorithms are better than others, and there are gaps in the interpretation. These gaps are explicitly filled in

the current dissertation through a systematic comparison of four algorithms and three internal metrics, which is a more holistic and theory-based evaluation framework.

## 2.4 Dynamic Segmentation Models: RFM-Based and Practical Implications in Customer Segmentation

Historically, RFM segmentation has offered a basic and intuitive foundation for evaluating customer value; however, recent studies have pointed out the weaknesses of the segmentation technique in terms of assessing time-based behaviour and customer emotion. To resolve these flaws, Wang et al. (2024) proposed the LRFMS model, which combines satisfaction and time-related trends based on time-series clustering algorithms like DTW-D and SBD. Their model was more valid than the conventional RFM but needed intense computational power and massive data sets, thus being difficult to use in real-time. Here we can see the conflict between the sophistication of the model and its operation in the retail arena.

On the same note, Gomes and Meisen (2023) came up with an adaptive RFM-SOM that recalculated the customer profile with time-changing behaviour. Their approach was the identification of behavioural changes and real-time decision-making. SOMs were, however, posing interpretation challenges as they represented non-linear; therefore, they are sensitive to initial parameters, which decreases their usability to both managers.

A hybrid approach was used by Griva et al. (2022), who input behavioural and geographic segmentation with K-Means and hierarchical clustering. This made it a multidimensional tool and helped a better interpret and strategic value of logistics and targeted marketing. However, the approach was not scalable and needed retailer-specific data that reduced generalisability.

In all of these studies, RFM is a pertinent instrument used in segmentation, yet there are several loopholes: Most developed RFM models require extensive computation, and this makes their practical implementation difficult. There are not many studies that combine RFM and various clustering algorithms to compare the stability of these methods. Not many studies support the dynamic RFM models with the help of multi-metric appraisal. These convictions warrant the methodological decisions in this dissertation, in which several clustering algorithms are deployed on RFM features and assessed through various measures in the effort to attain reproducible and understandable retail segmentation.

## 2.5 Integrating Clustering Results into Targeted Marketing Strategies in Retail

Even though clustering is a common tool to detect customer groups, less research directly relates segmentation results to practical marketing alternatives. Wang et al. (2025) investigated this connection with deep embedded clustering with the use of BERT-LDA to extract customer insight from texts. Their analysis enhanced segment understandability but caused difficulties associated with computing cost and non-transparent neural structures. This restricted its relevance to the typical retail space, where sophisticated analytics systems were not in place.

Lizybetinova et al (2020) explored the managerial approach to segmentation and demonstrated that data-driven segmentation was significantly more effective in companies, as campaigns responded better and the customers were more satisfied. Their mixed-method design provided useful information but relied more on a managerial interpretation, thereby lowering predictive strength. Their results indicate the difference between algorithmic accuracy and real-life marketing.

Song (2025) suggested a hybrid PCA agglomeration–ka ke means direction that provided directly distinct offices and converted them straight into marketing moves like incentives of loyalty and diverse differentiation. Even though the research had practical implement ability, it failed to compare the model performance with other algorithms, and hence, this raises the question of the best segmentation technique.

Taken together, these articles attest to the importance of segmentation in personalised marketing, retention of customers, and maximisation of profits. However, they always demonstrate lapses in: connecting the performance of the algorithm to feasible marketing actions, comparing the results of several clustering systems in a systematic fashion, and refining the notion of using multi-metric evaluation in marketing-focused segmentation studies. These problems are solved by the current study, which finds behaviourally different clusters of customers through the application of several algorithms and the explanation of how these subsets may be used to provide desired retail interventions.

## 2.6 Theoretical Underpinning

Theory of Planned Behaviour (TPB), which was developed by Ajzen (1991), is a theory that explains how, as an individual, an awareness and intent to behave is considered based on the attitude, subjective norm,s and the perceived behavioural control. TPP works towards a behavioural ground as regards to the retail segmentation, the ground is able to assist in understanding why customers have varying degrees of purchase

recency, frequency, and spending. The attitude of customers also affects their readiness to pay more or make purchases at a regular time; subjective norms determine loyalty and responsiveness to promotion; and perceived control has an impact on the possibility of making a repeat purchase, and this influences the recency. The relation of TPB to RFM-based segmentation can be used as a way of explaining the psychological motivation behind the cluster pattern. This integration makes sure that segmentation is not only statistics-based, but also theories, which are based on how consumers make decisions.

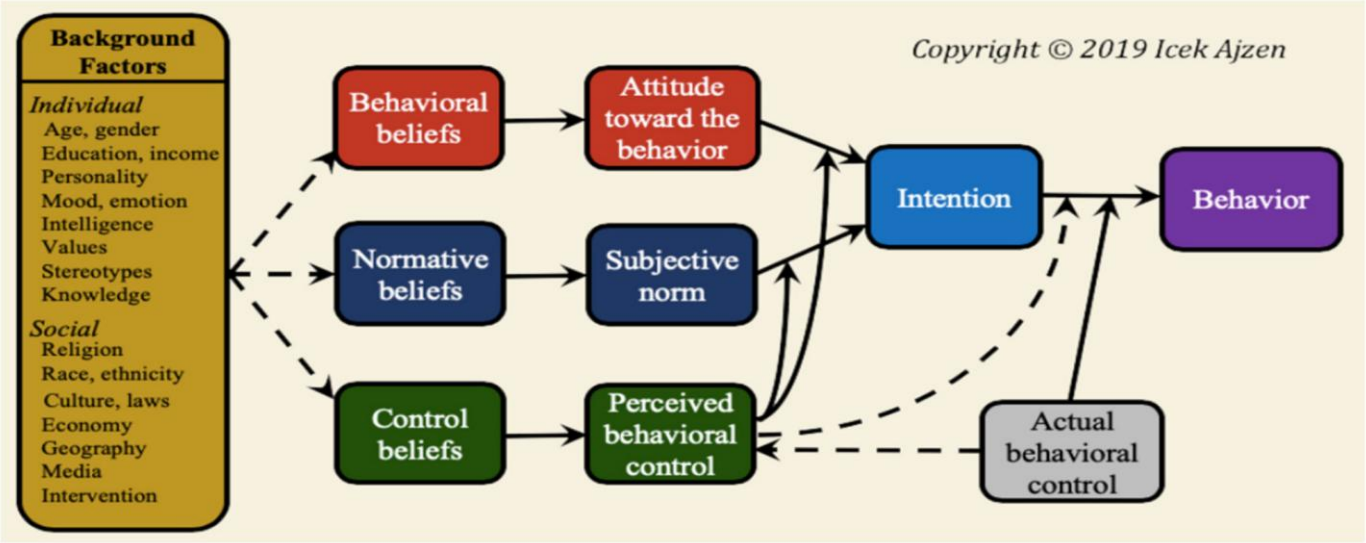



**Figure 2: Framework of Theory of Planned Behaviour (TPB)**

(Source: Alhamad and Donyai, 2021)

## 2.7 Literature Gaps

In the studied articles, there were a number of significant gaps that were not addressed, which provided a good reason to conduct the current dissertation. First, most of the previous literature depends on only one clustering algorithm, most of which is K-Means, without analysing the behaviour of various algorithms when applied to the same set of RFM data. This limits knowledge in algorithm appropriateness, particularly because the retail data can have non-linear, noisy, and overlapping patterns. Second, the majority of studies assess the segmentation with single or two internal evaluation measures, including Silhouette. The present research bridges this gap since it employs a three-metric evaluation framework (Silhouette, Davies-Bouldin, Calinski-Harabasz) to present a broader assessment.

Third, other researchers like Putri et al. (2024) and Rodrigues et al. (2025) indicate more clusters but do not provide the reasons why algorithms would have worked better mathematically. This dissertation is helpful because it allows the analysis of model behaviour in comparison to RFM structure, density uniformity, feature scaling, and cluster geometry, which allows to enhance a better theoretical interpretation. Not many show the impacts of resulting clusters as the driver of loyalty programmes, retention strategies, personalised offers, or campaign targeting. This paper fills that gap by connecting the cluster profiles and marketing insights that can be acted upon.

## 2.8 LR Matrix

| Study | Dataset | Algorithm(S) | Metrics Used | Key Finding | Gap Identified |
|---|---|---|---|---|---|
| Cordova (2024) | E-commerce behavioural logs | Descriptive analytics | None (exploratory) | Big data facilitates dynamic segmentation of behaviour that is not based on demographics. | No quantitative validation, no algorithmic comparison, Lacks clustering. |
| Putri et al. (2024) | The giant retail transaction storage. | K-Means | Silhouette | The six valuable RFM groups: medium cohesion. | Bases itself on a single algorithm and measure; not robust and does not have a comparative assessment. |
| Tavor et al. (2023) | Theory data sets of pricing. | Economic segmentation model. | Revenue modelling | Segmentation has an effect on the best prices. | None done empirically; poor association with machine learning methods. |

| | | | | | |
|---|---|---|---|---|---|
| Hicham and Karim (2022) | 35,000 Moroccan retail customers | Minimizing distance ( DBSCAN: DBSCAN, K-Means: K-Means, MiniBatch: MiniBatch, Mean Shifts: Mean Shifts) | ARI, NMI, DI, SC | The ensemble performed better than the individual algorithms. | Very costly to calculate; it does not provide a business interpretation of clusters. |
| Rodrigues et al. (2025) | Retail transaction data | K-Means, DBSCAN | Silhouette, DBI | DBSCAN works on irregular shapes of clusters. | None is multi-metrically validated; of variable data structure performances. |

***Table 1: LR Matrix***

## 2.9 Conceptual Framework

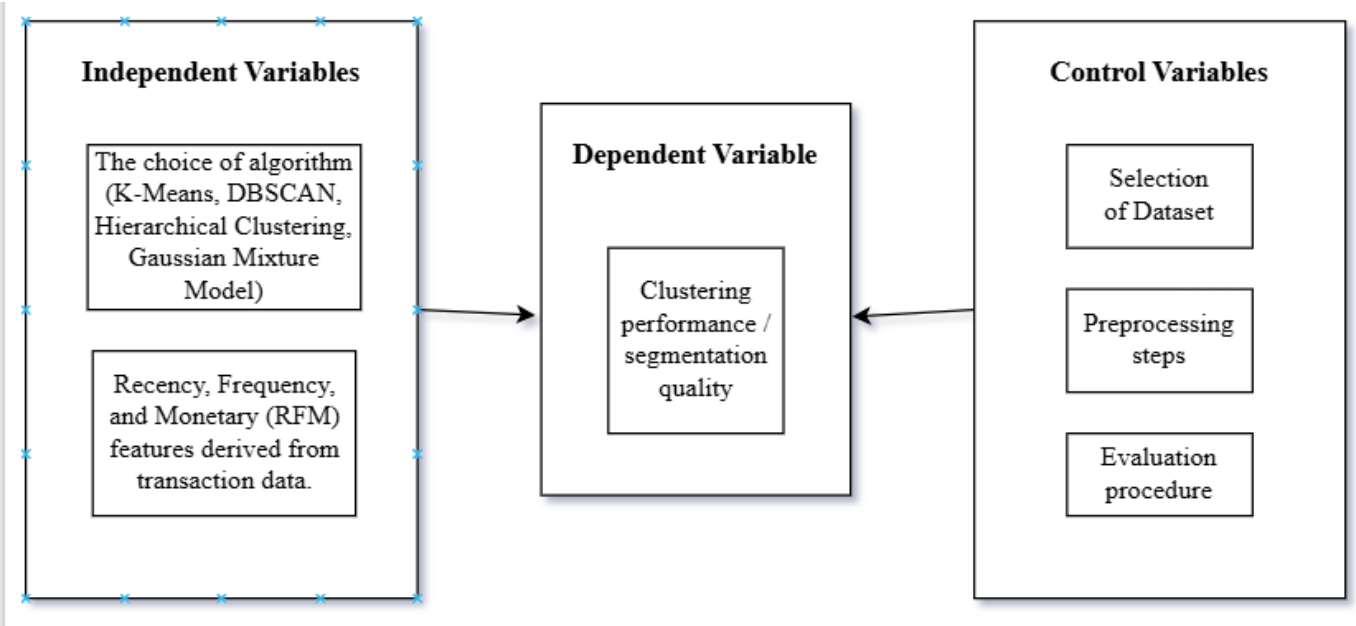



**Figure 3: Conceptual Framework**

## 2.10 Chapter Summary

The literature review presented here shows how customer segmentation has evolved from demographic-based to AI-based through clustering models. Studies have shown advances in unsupervised learning, RFM-based dynamic segmentation, and marketing integration. However, gaps exist in validation across multiple metrics, generalisability of models, and actionable retail strategy formulation.

# 3. Methodology

## 3.1 Introduction

The chapter outlines the methodology that was used to realise the research objective of assessing the effectiveness of clustering algorithms in the customer segmentation of retail marketing. It entails a systematic outline of philosophical position, research design, study design, data gathering and data analysis processes. The research utilises the quantitative approach, which takes the form of secondary data retrieved from the Kaggle retail dataset. K-Means, DBSCAN, Hierarchical, and Gaussian Mixture are unsupervised learning algorithms that are used to create data-driven customer groups. The evaluation is done based on Silhouette, Davies Bouldin and Calinski Harabasz indices. This methodological framework guarantees transparency, reproducibility and validity, which facilitates proper determination of the best clustering methods to be used to improve retail marketing segmentation.

## 3.2 Research Philosophy

The current research study is supported by a ***positivist*** philosophy since it aims at studying the behaviour of customers on the basis of quantifiable and empirical trends in transactional data. Positivism suits the approach because clustering algorithms rely on numerical evidence, which can be seen and not on subjective interpretation. The philosophy of replicable evaluation by standard scores, in the form of Silhouette, Davies-Bouldin, and Calinski-Harabasz, is supported. It guarantees an objective, scientific basis for model comparison.

## 3.3 Research Approach

This research adheres to the deductive research methodology based on the theoretical and algorithmic premises. As Hall, Hall, and Shaw (2022) put it, deduction is the process of testing previously developed theories or models using empirical data to prove the results. Such a strategy will enable the research to

confirm the presence of already identified algorithm properties, including K-Means being effective in a spherical cluster, in an RFM structure. The deductive approach should be adopted because it tests the preconceived assumptions using the empirical experimentation method.

## 3.4 Research Design

The study utilises the quantitative design, and, according to Ghanad (2023), the design is based on measurable data and statistical approaches to produce objective findings. The quantitative research design is adopted to make it possible to make systematic statistical comparisons between algorithms. Numerical preprocessing, evaluation based on metrics, as well as reproducible experimentation are supported by this design. It is also in line with the objective of the study to determine the most efficient algorithm of customer segmentation in regard to quantifiable measures of performance across multiple measures.

## 3.5 Data Collection

The research relies on the secondary data in the Kaggle dataset Customer Segmentation: Clustering developed by Kapoor (2021). This data includes information about retail transactions. The variables include InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. These uncooked variables have been converted into Recency, Frequency and monetary (RFM) scores to measure customer buying behaviour. Data cleaning will be used to eliminate duplicates, missing values, and put the date formats in a uniform format. Scaling has been applied to normalise the dataset to have an equal input to machine learning models. This dataset has been chosen since it offers a full picture representation of the purchasing trends of retailers, which can be used in the segmentation of algorithms by means of unsupervised learning. The ethical compliance, reliability and replicability are ensured due to the use of publicly available, anonymised data. The size and diversity of the dataset will be suitable for implementing clustering algorithms like K-Means, DBSCAN, Hierarchical, and Gaussian Mixture Models to attain the required accuracy in customer segmentation.

## 3.6 Justification of Selecting Machine Learning Models

| Model | Parameters | Findings from Literature | Justification for Use in This Study |
|---|---|---|---|

| | | | |
|---|---|---|---|
| **K-Means Clustering** | k = 4 | Ran et al. (2021) developed a superior K-Means algorithm that has a noise-handling mechanism to determine hotspots in urban areas more precisely. Their research was more precise in clustering, and more computationally efficient than the conventional K-Means, and thus their robustness in recognising dense regions of data even in the presence of noise interference. | K-Means has been chosen due to its simplicity, scalability and interpretability, which makes K-Means suitable for retail segmentation. The algorithm is effective in decomposing large RFM-based data into homogeneous clusters to easily decide on the customer segments, including the high-value, loyal, and dormant shopper categories. It has a low computational cost, which helps to implement it practically in retail systems of the real world. |
| **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** | eps = 0.3, min_samples = 5 | Bushra and Yi (2021) emphasised that DBSCAN was the most successful in terms of working with clusters of any form and reasonably identifying outliers compared to other density-based algorithms. Their comparative analysis over benchmark datasets showed to have a higher noise tolerance and also flexibility to detect irregular data distributions. | DBSCAN has been adopted in the study because it is able to deal with unevenly distributed and noisy data of customer transactions. It is also adaptive to the natural structure of the retail dataset, as it does not need predefined numbers of clusters, unlike K-Means. Its strength enhances the precision in the segmentation of different customer behaviours. |

| | | | |
|---|---|---|---|
| **Hierarchical Clustering** | linkage = 'ward' | Li et al. (2022) proposed a novel similarity metric for ensemble agglomerative hierarchical clustering-based data mining to enhance the stability as well as the cluster cohesion performance. Their results showed better interpretability and accuracy in groupings of multi-level data than in single-linkage. | Hierarchical clustering has been used to do visual exploration of customer relationships using dendrograms. It allows seeing the built-in group structures- handy in differentiating premium, occasional and inactive customers. It is interpretable, which is complementary to quantitative results, presenting a detailed segmentation structure. |
| **Gaussian Mixture Model (GMM)** | covariance_type = 'full' | A soft computing-based GMM framework was proposed by Gogebakan (2021), which enhanced probabilistic estimation as well as minimised misclassification in boundary areas among overlapping clusters. The model was more successful in attaining cohesion and separation among mixed datasets in regard to the distribution of features. | GMM has been incorporated to cater to overlapping customer behaviours in which the boundaries between clusters are not clear. Its probabilistic nature makes it more flexible, and a soft assignment of customers to segments is possible. It can be used especially in marketing situations where individualised suggestions and cross-segmentation are demanded. |

***Table 2: Justification for Selecting Machine Learning Models***

(Source: Self-acquired)

## 3.7 Data Analysis and Software Tools

The analysis of the data was done in a systematic pipeline. The first step was to convert cleaned transactions into RFM variables, which were behavioural attributes. The values were scaled with MinMaxScaler to avoid the dominance of big magnitude features. Dimensional reduction involved PCA, which enhanced the visual perception of structural cluster information. Four clustering algorithms have then been used with the scikit-learn libraries. Silhouette, Davies-Bouldin, and Calinski-Harabasz indices were used to evaluate their outputs

and thus offer multi-metric validation. Python was primarily used as the analysis environment with the aid of Pandas, NumPy, Matplotlib, and Seaborn to manipulate and visualise.

Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index are used to assess the quality of clusters, which guarantees the multi-metric validation method. The decrease in the dimensions through Principal Component Analysis (PCA) allows visualisation in a 2D format to facilitate a better interpretation of whether clusters can be separated. This analytic procedure has achieved computational efficiency, transparency, as well as reproducibility. The application of Python and its open-source libraries has offered accuracy, scalability, and a solid framework to test the performance of the algorithm in customer segmentation in retail marketing applications.

## 3.8 Ethical Consideration

This study has adhered to the ethical requirements stipulated in the Irish Data Protection Act 2000(electronic Irish Statute Book, 2000). The Kaggle dataset employed is open, as well as anonymised, and does not contain any personal identifiers. Data has been applied in academic purposes, with integrity, transparency and reproducibility. All the algorithms and visualisations are responsibly implemented with open-source software. Correct referencing and credit of authors have been ensured to avoid plagiarism and comply with the ethical standards of research and data management.

## 3.9 Chapter Summary

The chapter has outlined the research methodology, which included philosophical, theoretical and analytical frameworks used to meet the study objectives. Quantitative research design has been used based on a positivist philosophy and deductive approach. RFM-based feature engineering and analysis of the secondary data stored in Kaggle have been conducted in Python with the Scikit-learn library. Four clustering algorithms (K-Means, DBSCAN, Hierarchical, and Gaussian Mixture clustering algorithms) have been deployed and tested based on multi-metric validation. There has been a preservation of ethics and data integrity, thus transparency and academic credibility are upheld.

# 4. Design Specification

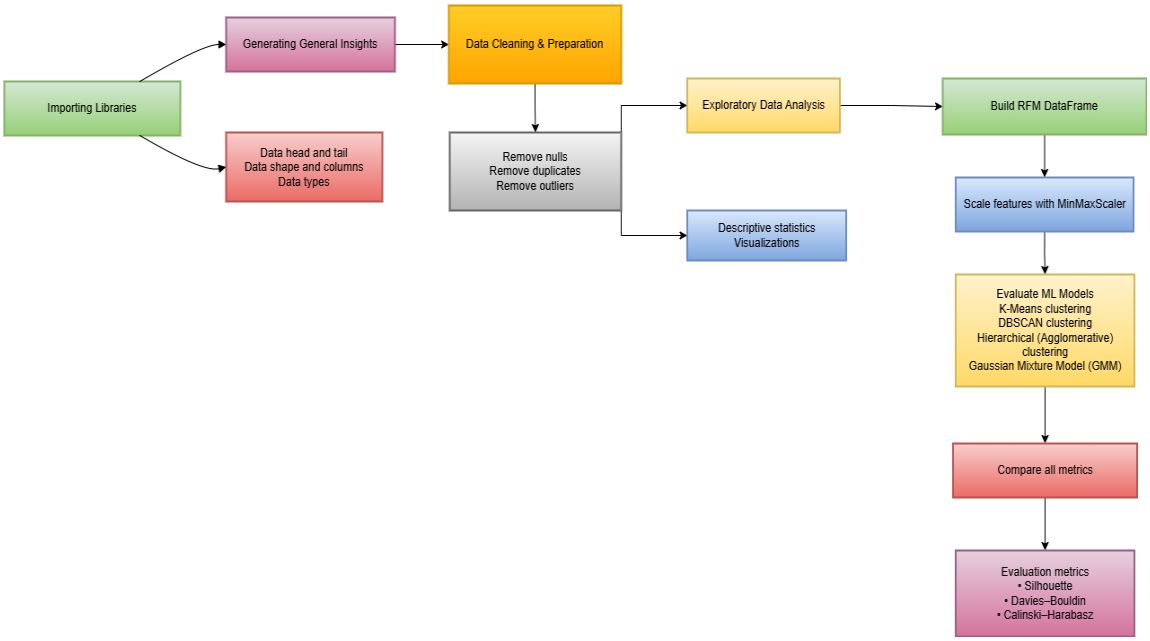



***Figure 4: Design Specification***

The flow chart illustrates a customer segmentation end-to-end analytics process based on purchase behaviour. It starts with the importation of libraries that offer statistical, plotting, and machine-learning libraries later to be used. The next step is that the analyst creates general insights, such as previewing the head and tail of the data, verifying shape, enumerating columns, and inspecting data types. Such checks are made to ascertain that, before any transformation, there are assumptions concerning variables. The data cleaning and preparation are then entered via the pipeline. In this case, missing values are imputed, duplicate data is dropped, and extreme outliers are addressed to ensure downstream models are not biased. Exploratory data analysis is done after cleaning. EDA includes descriptive statistics and visualisations to learn about distributions, relationships, and what may be wrong with it, and ensure that some level of cleaning actually gave interpretable data.

In the presence of reliable data, the process creates an RFM DataFrame, retaining three behaviour signals, viz. Recency, Frequency, and Monetary value with indexing by customer. These features vary in scale, so to ensure that no one variable dominates, MinMaxScaler normalises them to a 0-1 range. Different clustering algorithms, including K-Means, DBSCAN, hierarchical (agglomerative) and Gaussian Mixture Models, are tested in the modelling phase to provide support to the shapes and densities of clusters of varying sizes. The

results of every model are evaluated on the quantitative level through internal validation measures, including the Silhouette (cohesion and separation), Davies-Bouldin (similarity of clustering, lower is better), and Calinski-Harabasz (between-cluster dispersion, higher is better). One of the steps is a comparison step, which ranks models based on the measures in order to find out the most credible segmentation.

# 5. Implementation

The implementation phase converted the methodology scheme into a systematic analysis procedure which aimed to pre-process the data, design behavioural attributes, and use customer clustering algorithms. Instead of specifying the operations of each code line, the description of the conceptual implementation of each step is presented in this section, and the complete Python scripts and results can be found in the appendices to be transparent in terms of technical aspects.

The first step was to import required analytical, visualisation and machine-learning libraries which facilitate the management, numerical and modelling of data. Kaggles' data were then loaded and initially underwent some structural analysis to ensure that it was of the appropriate size, format and variable structure. Preparation of data then came in with systematic cleaning processes to ensure the integrity and reliability of the data to be used in modelling. Median substitution was used to impute missing income values to maintain distribution properties, and the customer registration date was converted to an identical data type. Duplications and other discrepancies that remained were eliminated in order to have a uniform analytical foundation.

The next stage was the outlier detection and treatment. To determine the outliers that would bias the clustering outcomes, the study used the Interquartile Range (IQR) for the key behavioural variables, Income, Recency, Frequency, and Monetary. The values were limited at the 1st and 99th percentiles instead of eliminating outliers, which could be important high-value behaviours. This guaranteed the stability of the subsequent algorithms, preserving the overall behavioural structure.

It was followed by feature engineering of transactional data into Recency, Frequency, and Monetary (RFM) measures, which formed a behaviour-based dataset to use in segmentation. Since RFM variables are scaled on different scales, they have been normalised by the MinMax scaling in order to obtain fair distance values in algorithms like DBSCAN or K-Means.

Principal Component Analysis (PCA) dimensionality reduction model was used to project the three variables of RFM into two principal components. This enabled the clusters to be interpreted visually more clearly and minimise noise without much loss of variance. The scaled data were then executed with four clustering models, which include K-Means, DBSCAN, Agglomerative Hierarchical Clustering and Gaussian Mixture Models. Their outputs were tested in terms of the Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index to generate a multi-metric test which enhanced the model transparency and comparability. Complete code, visualisations and intermediate results are provided in the appendices; hence it is reproducible, with the main dissertation being conceptually oriented.

# 6. Evaluation

## 6.1 Model Evaluation and Comparison

| Model | Silhouette Score | Davies–Bouldin Index | Calinski–Harabasz Index |
|---|---|---|---|
| **K-Means** | 0.3854 | 0.9490 | 1929.85 |
| **Hierarchical** | 0.3483 | 1.0083 | 1625.45 |
| **GMM** | 0.3437 | 0.9823 | 1561.66 |

***Table 3: Model Comparison Metrics***

Appendix 5, Elbow Method plot depicts a reduction in inertia with an increase in the number of clusters, but there is an evident bend at the value of k=4, representing the optimal number of clusters. The PCA scatter plot displays the K-Means output, and in this case, four different clusters can be identified, distinctions of the customer Recency, Frequency, and Monetary behaviour, which are clearly separated in the plot of the dimensions after reducing the data.

Appendix 6, DBSCAN plot indicates that the algorithm marked virtually all points as one cluster (orange), and there was a small amount of noise (blue). This implies that DBSCAN was unable to find significant density-based clusters in the RFM data, probably because the uniform distribution of points following scaling resulted in density differences that were too low to be separated.

Appendix 7, a hierarchical clustering plot, demonstrates that there are four clusters that are well-formed following PCA reduction. All colours show different customer groups according to Recency, Frequency, and Monetary behaviour. The groups can be seen as distinct, which means that their differences are meaningful, as the evaluation measures provide moderate Silhouette scores and high Calinski-Harabasz values.

Appendix 8, following the GMM clustering visual, indicates that there are four customer groupings that have been created as a result of the reduction of PCA. The GMFs generate gentler, probability-based clusters, which represent any overlapping practices of Recency, Frequency and Monetary behaviour. The groups are kept far apart, and the measures of evaluation reveal that the segmentation is stable and the Silhouette scores are moderate, and the Calinski-Harabasz performance is relatively high.

Appendix 9, This is a comparison table that summarises the performance of clustering using the most important evaluation metrics in three models. K-Means is the most successful as it has the highest Silhouette score (0.385), which implies better separation between clusters. It is also compact and distinct in clusters, Davies-Bouldin (0.949) and strong in cluster structure, Calinski-Harabasz (1929). On the one hand, hierarchical and GMM models are moderately performing, which exhibit a slightly poorer division and compactness. DBSCAN was not used as it yielded a single effective cluster for noise removal.

## 6.2 Findings of the Analysis

The analysis indicates distinct customer behaviour patterns according to their purchasing recency, frequency, and expenditure capabilities. Following the cleaning and transformation of the dataset, the RFM features were highly varied among the customers, indicating that there are meaningful differences in their purchase date in the past, the frequency at which they spend and the amount of money that they spend. This was manually tested using visual analysis, and it was found that certain customers were adding more than others.

The results of clustering show that customers naturally tend to create different behavioural groups. K-Means was also better than other algorithms since RFM features generate cluster structures which are approximately spherical on scaling, and thus Euclidean distance is an effective measure. The MinMax scaling would eliminate the issue of density variability, making DBSCAN treat most of the points as a single cluster. GMM creates clusters that are probabilistic and therefore do not give boundaries. Hierarchical clustering is quite susceptible to linkage techniques, as well as noise and decreases the consistency of clusters. Therefore, K-Means is the one that is best in line with the mathematical characteristics of the transformed RFM space.

The better K-Means clusters help in taking retail actions through a clear differentiation of high-value, loyal, at-risk, and low-engagement customers. The segments promote focused loyalty programmes, personalised promotions, predicting churn and marketing campaigns with budget constraints. The accurate, data-driven decisions will allow retailers to focus on profitable segments, re-engage the idle customers, and increase the customer lifetime value.

## 6.3 Discussion

The results of this research support the tendencies mentioned in the previous literature, and the limitations deal with central blank areas concerning multi-metric validation and inter-algorithm comparison. In line with Akande et al. (2024) and Putri et al. (2024), it is revealed that K-Means still provides constant and understandable clusters of customers and proves to be beneficial in terms of RFM-based retail marketing. Nevertheless, in contrast to the research that employed a single measure, this study employed a multi-metric method of evaluation, closing the methodological gap found in the existing studies (Cordova, 2024; Hicham and Karim, 2022). The somewhat poorer performance of Hierarchy and GMM models is consistent with the finding of Rodrigues et al. (2025) that the efficacy of algorithms is based on the structure of data and not the thoroughness of the theoretical complexity. The lack of any meaningful clusters in DBSCAN also confirms the results of John et al. (2023), who remarked that it is sensitive to the choice of parameters and noise distribution. The discussion confirms that the small size of the dataset, single source origin, and homogenous nature of the structure of the element of RFM reduce generalisability. The results are consistent with Putri et al. (2024) and Akande et al. (2024), who find that K-Means is stable and contrast with Rodrigues et al. (2025), who discovered better DBSCAN performance when irregular densities are present. Results were dependent on parameter sensitivity, especially in DBSCAN and GMM. The technology of the multi-velocity validation metrics increases the levels of reliability in comparison to the past single-metric tests, presenting a more stringent evaluation of segmentation.

# 7. Conclusion and Future Work

## 7.1 Conclusion

The research concludes that customer segments can be identified effectively based on RFM-based clustering, and each of the research questions is answered explicitly. First, the research established that machine-learning methods can significantly categorise customers as a set of distinct behavioural patterns. Second, the

comparison of the algorithms revealed that K-Means generated the most consistent clusters, whereas DBSCAN was inappropriate because of the even density. Thirdly, the use of multi-metric assessment enhanced the strength of the results as compared to the case of single-metric studies. Lastly, the outputs of the segmentation showed to have a definite marketing value that aids in targeted promotions, retentions, and optimisation of loyalty. The main contributions made are a strict multi-metric validation framework, a four-algorithm comparison, and useful information regarding the segment characteristics. These results demonstrate the benefits of data-driven segmentation in the retail setting to improve decision-making and competitiveness.

## 7.2 Strengths and Weaknesses

Some of the major strengths of the present study include the use of a multi-metric validation framework that makes the clustering results reliable in comparison with other studies that used only one metric. Four variants of algorithms, which are K-Means, DBSCAN, Hierarchical and GMM, provide a deeper methodological approach and allow making a comprehensive comparison. The data was cleansed, scaled, and PCA, which guaranteed the quality of the data and solid visualisation, which further supported the soundness of the results. Nevertheless, the research has drawbacks as well. Generalisation of results to other retail situations is limited by the fact that only one Kaggle dataset is used. The poor performance of DBSCAN highlights the sensitivity to parameter values and homogeneous distributions of data. Also, clustering results rely on RFM features but not demographic or behavioural variables, which may improve the quality of segmentation.

## 7.2 Recommendation for Future Work

Future studies ought to expand this research by including several retail datasets from various industries to enhance generalisability and achieve the consistency of the clustering results under diverse customer behaviours. Segmentation would be deeper with the addition of other characteristics other than RFM, including demographics, product types, customer lifetime value, or sentiment analysis of reviews. Additional research is needed on more sophisticated clustering algorithms, such as ensemble clustering and deep learning-based systems, and systems that combine supervised and unsupervised approaches. DBSCAN and GMM could be reinforced using such tools as parameter optimisation, which could enhance the performance of these methods, like grid search or Bayesian optimisation. Tying the results of segmentation to the actual marketing tests in the real world, A/B testing or monitoring campaign responses, would legitimise the potential practical effectiveness of clusters and make them more useful to strategic decision-making. The

hyperparameter optimisation methods, like grid search or Bayesian tuning, should be included in future studies to increase the model precision. The segmentation could be validated by CRM integration and actual customer behaviours that would enhance external reliability. Also, it would be more practical to A/B test campaign segments and correlate the results with the KPIs like conversion, retention, and customer lifetime value.

# References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), pp.179–211. doi:https://doi.org/10.1016/0749-5978(91)90020-T.

Akande, O.N., Akande, H.B., Asani, E.O. and Dautare, B.T. (2024). Customer Segmentation through RFM Analysis and K-means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy. pp.1–8. doi:https://doi.org/10.1109/seb4sdg60871.2024.10630052.

Alhamad, H. and Donyai, P. (2021). The Validity of the Theory of Planned Behaviour for Understanding People's Beliefs and Intentions toward Reusing Medicines. *Pharmacy*, [online] 9(1), p.58. doi:https://doi.org/10.3390/pharmacy9010058.

Book (eISB), electronic I.S. (2000). *electronic Irish Statute Book (eISB)*. [online] www.irishstatutebook.ie. Available at: https://www.irishstatutebook.ie/eli/2000/act/28/enacted/en/html [Accessed 24 Oct. 2025].

Bushra, A.A. and Yi, G. (2021). Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. *IEEE Access*, 9, pp.87918–87935. doi:https://doi.org/10.1109/access.2021.3089036.

Carr, B.J., Clesse, S., García-Bellido, J., Hawkins, M.R.S. and Kühnel, F. (2024). Observational evidence for primordial black holes: A positivist perspective. *Physics Reports*, 1054, pp.1–68. doi:https://doi.org/10.1016/j.physrep.2023.11.005.

Cordova, R. (2024). CUSTOMER SEGMENTATION IN THE ONLINE RETAIL INDUSTRY USING BIG DATA ANALYTICS. *Journal of Theoretical and Applied Information Technology*, [online] 102(22). Available at: https://www.jatit.org/volumes/Vol102No22/25Vol102No22.pdf.

Ghanad, A. (2023). An Overview of Quantitative Research Methods. *International Journal of Multidisciplinary Research and Analysis*, [online] 6(8), pp.3794–3803. doi:https://doi.org/10.47191/ijmra/v6-i8-52.

Gogebakan, M. (2021). A Novel Approach for Gaussian Mixture Model Clustering Based on Soft Computing Method. *IEEE Access*, 9, pp.159987–160003. doi:https://doi.org/10.1109/access.2021.3130066.

Gomes, M.A. and Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, [online] 21(21), pp.527–570. doi:https://doi.org/10.1007/s10257-023-00640-4.

Griva, A., Zampou, E., Stavrou, V., Papakiriakopoulos, D. and Doukidis, G. (2022). A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data. *Journal of Decision Systems*, 33(1), pp.1–29. doi:https://doi.org/10.1080/12460125.2022.2151071.

Hall, J.R., Hall, S.S. and Shaw, E.H. (2022). A Deductive Approach to a Systematic Review of Entrepreneurship Literature. *Management Review Quarterly*, 73(1), pp.1–30. doi:https://doi.org/10.1007/s11301-022-00266-9.

Hicham, N. and Karim, S. (2022). Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering. *International Journal of Advanced Computer Science and Applications*, 13(10). doi:https://doi.org/10.14569/ijacsa.2022.0131016.

John, J., Shobayo, O. and Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, [online] 2(4), pp.809–823. doi:https://doi.org/10.3390/analytics2040042.

Kapoor, K. (2021). *Customer Segmentation: Clustering* 🛍️🛒🛒. [online] kaggle.com. Available at: https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering/input [Accessed 24 Oct. 2025].

Li, T., Rezaeipanah, A. and Tag El Din, E.M. (2022). An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *Journal of King Saud University - Computer and Information Sciences*, [online] 34(6, Part B), pp.3828–3842. doi:https://doi.org/10.1016/j.jksuci.2022.04.010.

Ližbetinová, L., Štarchoň, P., Lorincová, S., Weberová, D. and Průša, P. (2020). Application of Cluster Analysis in Marketing Communications in Small and Medium-Sized Enterprises: An Empirical Study in the Slovak Republic. *Sustainability*, 11(8), p.2302. doi:https://doi.org/10.3390/su11082302.

Putri, Y., Aldo, D. and Ilham, W. (2024). Retail Marketing Strategy Optimization: Customer Segmentation with Artificial Intelligence Integration and K-Means Clustering. *sinkron*, 8(4), pp.2155–2163. doi:https://doi.org/10.33395/sinkron.v8i4.14000.

Ran, X., Zhou, X., Lei, M., Tepsan, W. and Deng, W. (2021). A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots. *Applied Sciences*, 11(23), p.11202. doi:https://doi.org/10.3390/app112311202.

Rodrigues, G.N., Mir, M.D.Nazmul.H., Bhuiyan, M.D.Shahriar.M., Rafi, M.D.A.L., Hoque, A.S.M.Morshedul., Maua, J. and Mridha, M.F. (2025). NLP-driven customer segmentation: A comprehensive review of methods and applications in personalized marketing. *Data Science and Management*. doi:https://doi.org/10.1016/j.dsm.2025.09.002.

Song, Y. (2025). Data-Driven Customer Segmentation and Marketing Strategies in Grocery Retail. *SHS Web of Conferences*, [online] 218, p.02011. doi:https://doi.org/10.1051/shsconf/202521802011.

Tavor, T., Gonen, L.D. and Spiegel, U. (2023). Customer Segmentation as a Revenue Generator for Profit Purposes. *Mathematics*, [online] 11(21). doi:https://doi.org/10.3390/math11214425.

Ufeli, C.P., Sattar, M.U., Hasan, R. and Mahmood, S. (2025). Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms. *Information*, 16(6), p.441. doi:https://doi.org/10.3390/info16060441.

Wang, M., Meng, T., Gu, X., Wang, D., Wang, R. and Zhao, R. (2025). Deep segmentation of retail customers based on improved DEC and multimodal semantic representation. *Alexandria Engineering Journal*, [online] 130, pp.1–10. doi:https://doi.org/10.1016/j.aej.2025.09.012.

Wang, S., Sun, L. and Yu, Y. (2024). A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. *Scientific Reports*, 14(1). doi:https://doi.org/10.1038/s41598-024-68621-2.

# Appendices

## Appendix 1: Import Libraries and dataset

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans, DBSCAN, AgglomerativeClustering
from sklearn.mixture import GaussianMixture
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_score
import warnings
warnings.filterwarnings('ignore')
```

```
## Loading of the dataset
```

```
df = pd.read_csv("marketing_campaign.csv", sep='\t')
df.head(10)
```

***Figure 5: Import Libraries and the dataset***

## Appendix 2: Data Cleaning

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

```
df.isna().sum()
```

```
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                24
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Z_CostContact          0
Z_Revenue              0
Response               0
dtype: int64
```

```
# Imputing Income with median
df["Income"] = df["Income"].fillna(df["Income"].median())
```

```
# Converting the date field
df["Dt_Customer"] = pd.to_datetime(df["Dt_Customer"], dayfirst=True, errors="coerce")
```

```
# Checking for duplicate rows
duplicate_count = df.duplicated().sum()
print(f"Number of duplicate rows: {duplicate_count}")
```

```
Number of duplicate rows: 0
```

```
 # Re-checking the missing values after cleaning
print("\nMissing values after cleaning:")
print(df.isna().sum())
```

```
Missing values after cleaning:
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                 0
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Z_CostContact          0
Z_Revenue              0
Response               0
dtype: int64
```

```
##  Detect outliers using IQR method

num_cols = ["Income", "Recency", "Frequency", "Monetary"]

for col in num_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    outliers = ((df[col] < lower) | (df[col] > upper)).sum()
    print(f"{col}: {outliers} potential outliers")
# Visual checking of outliers
import matplotlib.pyplot as plt
plt.figure(figsize=(12,6))
sns.boxplot(data=df[num_cols])
plt.title("Boxplot to Visualize Outliers in Key Variables")
plt.show()
```

```
Income: 8 potential outliers
Recency: 0 potential outliers
Frequency: 2 potential outliers
Monetary: 3 potential outliers
```

```
# Defining a function to cap outliers at 1st and 99th percentiles
def cap_outliers(df, cols):
    for col in cols:
        lower = df[col].quantile(0.01)
        upper = df[col].quantile(0.99)
        df[col] = np.where(df[col] < lower, lower, df[col])
        df[col] = np.where(df[col] > upper, upper, df[col])
        print(f"{col}: capped between {lower:.2f} and {upper:.2f}")
    return df

# Applying capping to affected numerical columns
outlier_cols = ["Income", "Frequency", "Monetary"]
df = cap_outliers(df, outlier_cols)

# Rechecking the outlier presence visually
import matplotlib.pyplot as plt
plt.figure(figsize=(12,6))
sns.boxplot(data=df[outlier_cols])
plt.title("Boxplot After Outlier Capping")
plt.show()
```

```
Income: capped between 7705.92 and 94437.68
Frequency: capped between 4.00 and 32.00
Monetary: capped between 13.00 and 2126.00
```

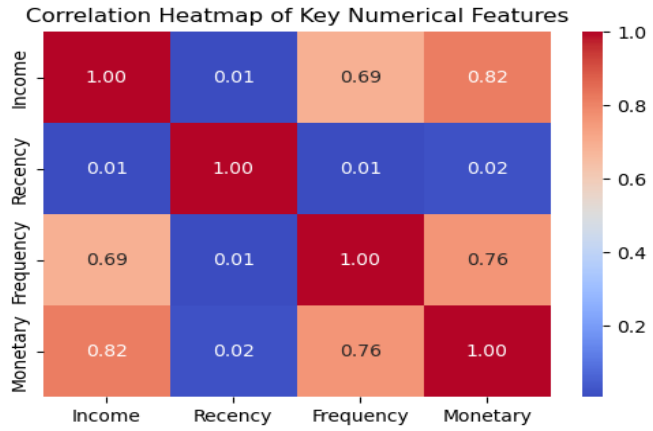
***Table 3: Data Cleaning***

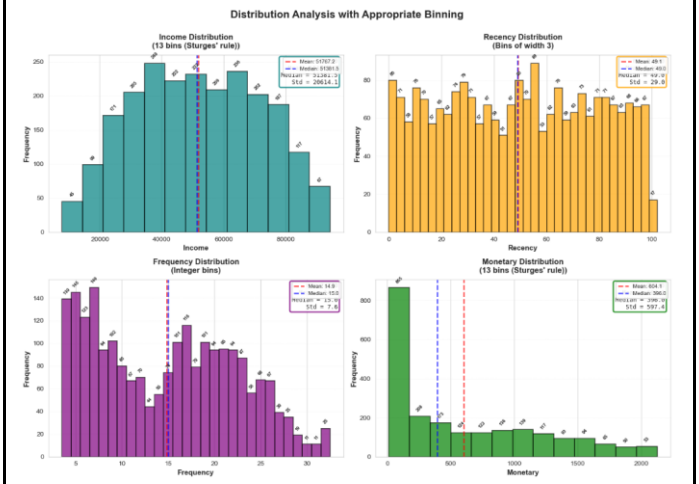
# Appendix 3: EDA

```
df.describe()
```

| | ID | Year_Birth | Income | Kidhome | Teenhome | Dt_Customer | Recency | MntWines | MntFruits | MntMeatProducts | ... | AcceptedCmp4 | AcceptedCmp5 | AcceptedCmp1 | AcceptedCmp2 | Complain | Z_CostContact | Z_Revenue | Response | Frequency | Monetary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 2240.000000 | 2240.000000 | 2240.000000 | 2240.000000 | 2240.000000 | 2240 | 2240.000000 | 2240.000000 | 2240.000000 | 2240.000000 | ... | 2240.000000 | 2240.000000 | 2240.000000 | 2240.000000 | 2240.000000 | 2240.0 | 2240.0 | 2240.000000 | 2240.000000 | 2240.000000 |
| mean | 5592.159821 | 1968.805804 | 51767.195893 | 0.444196 | 0.506250 | 2013-07-10 10:01:42.857142784 | 49.109375 | 303.935714 | 26.302232 | 166.950000 | ... | 0.074554 | 0.072768 | 0.064286 | 0.013393 | 0.009375 | 3.0 | 11.0 | 0.149107 | 14.854911 | 604.068304 |
| min | 0.000000 | 1893.000000 | 7705.920000 | 0.000000 | 0.000000 | 2012-07-30 00:00:00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 3.0 | 11.0 | 0.000000 | 4.000000 | 13.000000 |
| 25% | 2828.250000 | 1959.000000 | 35538.750000 | 0.000000 | 0.000000 | 2013-01-16 00:00:00 | 24.000000 | 23.750000 | 1.000000 | 16.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 3.0 | 11.0 | 0.000000 | 8.000000 | 68.750000 |
| 50% | 5458.500000 | 1970.000000 | 51381.500000 | 0.000000 | 0.000000 | 2013-07-08 12:00:00 | 49.000000 | 173.500000 | 8.000000 | 67.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 3.0 | 11.0 | 0.000000 | 15.000000 | 396.000000 |
| 75% | 8427.750000 | 1977.000000 | 68289.750000 | 1.000000 | 1.000000 | 2013-12-30 06:00:00 | 74.000000 | 504.250000 | 33.000000 | 232.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 3.0 | 11.0 | 0.000000 | 21.000000 | 1045.500000 |
| max | 11191.000000 | 1996.000000 | 94437.680000 | 2.000000 | 2.000000 | 2014-06-29 00:00:00 | 99.000000 | 1493.000000 | 199.000000 | 1725.000000 | ... | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 3.0 | 11.0 | 1.000000 | 32.000000 | 2126.000000 |
| std | 3246.662198 | 11.984069 | 20614.112834 | 0.538398 | 0.544538 | NaN | 28.962453 | 336.597393 | 39.773434 | 225.715373 | ... | 0.262728 | 0.259813 | 0.245316 | 0.114976 | 0.096391 | 0.0 | 0.0 | 0.356274 | 7.591392 | 597.352982 |

8 rows × 29 columns



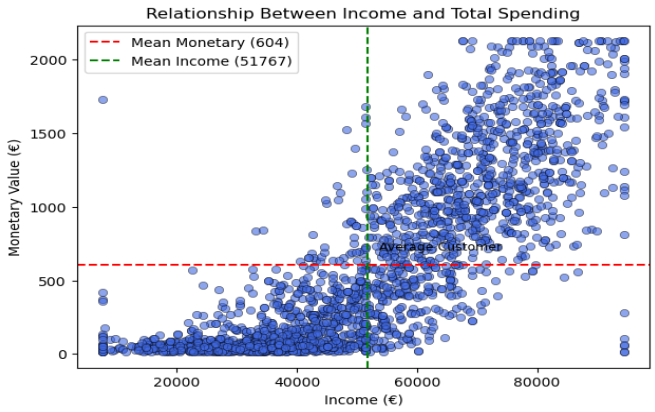




***Table 4: EDA***

# Appendix 4: Feature scaling

```
## Build RFM DataFrame
```

```
# Building the RFM DataFrame from the cleaned dataset
rfm = df[["ID", "Recency", "Frequency", "Monetary"]].copy()
rfm = rfm.set_index("ID")

print("RFM DataFrame preview:")
display(rfm.head())
```

RFM DataFrame preview:

| ID | Recency | Frequency | Monetary |
|---|---|---|---|
| **5524** | 58 | 25.0 | 1617.0 |
| **2174** | 38 | 6.0 | 27.0 |
| **4141** | 26 | 21.0 | 776.0 |
| **6182** | 26 | 8.0 | 53.0 |
| **5324** | 94 | 19.0 | 422.0 |

```
## Scale features with MinMaxScaler
```

```
# Scaling numerical values to avoid dominance by large magnitudes
scaler = MinMaxScaler()
rfm_scaled = scaler.fit_transform(rfm)

rfm_scaled_df = pd.DataFrame(
    rfm_scaled, columns=["Recency", "Frequency", "Monetary"], index=rfm.index
)

print("Scaled RFM data:")
display(rfm_scaled_df.head())
```

Scaled RFM data:

| ID | Recency | Frequency | Monetary |
|---|---|---|---|
| **5524** | 0.585859 | 0.750000 | 0.759110 |
| **2174** | 0.383838 | 0.071429 | 0.006626 |
| **4141** | 0.262626 | 0.607143 | 0.361098 |
| **6182** | 0.262626 | 0.142857 | 0.018930 |
| **5324** | 0.949495 | 0.535714 | 0.193564 |

```
## PCA for visualisation
```

```
pca = PCA(n_components=2, random_state=42)
rfm_pca = pca.fit_transform(rfm_scaled_df)

rfm_pca_df = pd.DataFrame(rfm_pca, columns=["PC1", "PC2"], index=rfm.index)
print("Explained variance ratio:", pca.explained_variance_ratio_)
rfm_pca_df.head()
```

Explained variance ratio: [0.56514594 0.35792132]

| ID | PC1 | PC2 |
|---|---|---|
| **5524** | 0.599746 | 0.072011 |
| **2174** | -0.419016 | -0.098703 |
| **4141** | 0.202574 | -0.241114 |
| **6182** | -0.364786 | -0.222193 |
| **5324** | 0.053221 | 0.449891 |

```
## Defining the Helper function for evaluation
```

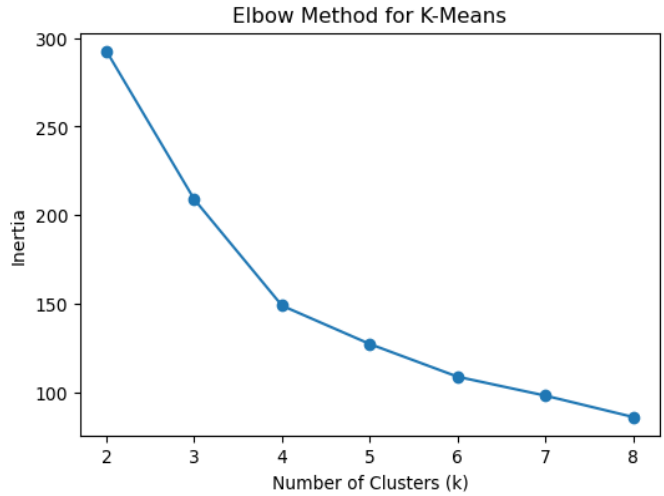
```
def evaluate_clustering(X, labels, name="Model"):
    unique_labels = set(labels)
    if len(unique_labels) <= 1:
        print(f"{name}: Not enough clusters to evaluate.")
        return None

    if -1 in unique_labels:
        mask = labels != -1
        X_eval, labels_eval = X[mask], labels[mask]
        if len(set(labels_eval)) <= 1:
            print(f"{name}: Only noise or single cluster after excluding noise.")
            return None
    else:
        X_eval, labels_eval = X, labels

    sil = silhouette_score(X_eval, labels_eval)
    db = davies_bouldin_score(X_eval, labels_eval)
    ch = calinski_harabasz_score(X_eval, labels_eval)
    print(f"{name} → Silhouette: {sil:.4f}, Davies-Bouldin: {db:.4f}, Calinski-Harabasz: {ch:.2f}")
    return {"model": name, "silhouette": sil, "davies_bouldin": db, "calinski_harabasz": ch}
```

***Table 5: Feature scaling***

# Appendix 5: K-Means clustering



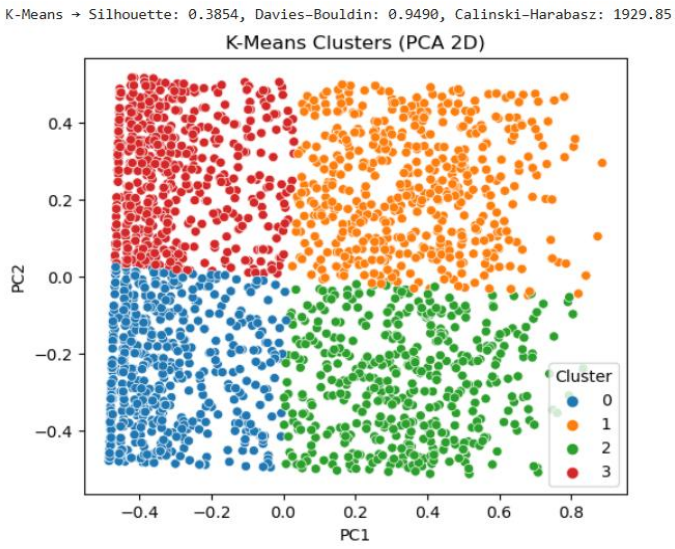




***Table 6: K-Means clustering***

## Appendix 6: DBSCAN clustering

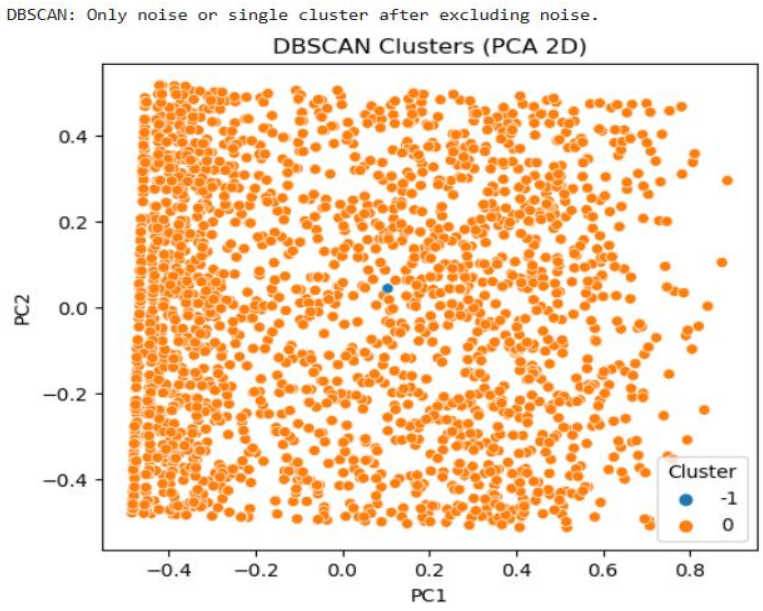



***Figure 6: DBSCAN clustering***

## Appendix 7: Hierarchical (Agglomerative) clustering

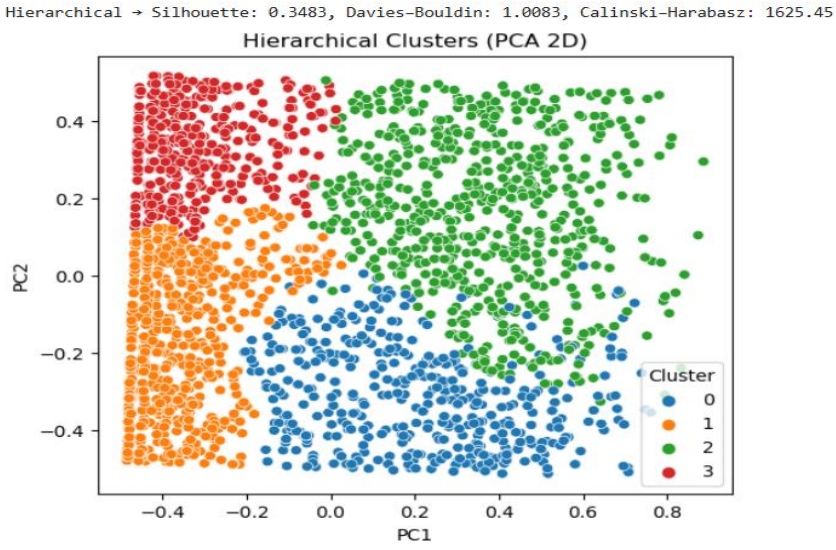



***Figure 7: Hierarchical (Agglomerative) clustering***

## Appendix 8: Gaussian Mixture Model (GMM)

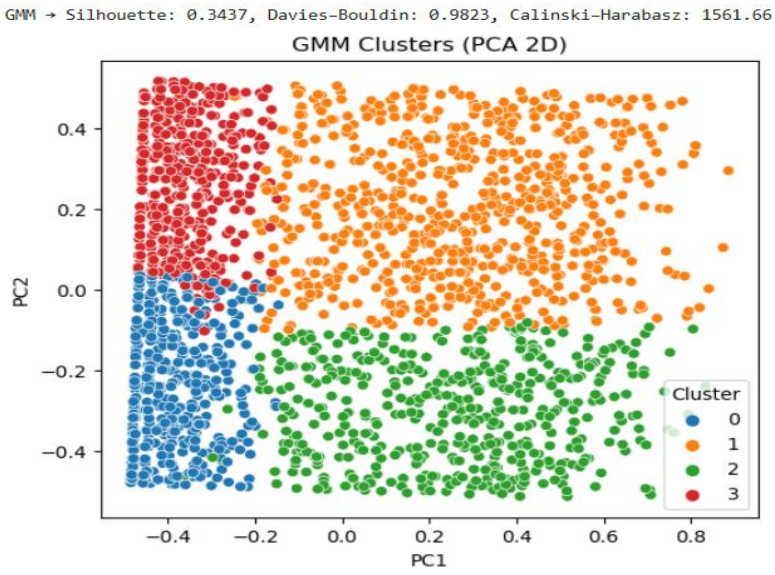



***Figure 8: Gaussian Mixture Model (GMM)***

## Appendix 9: Comparing the model metrics

```
K-Means → Silhouette: 0.3854, Davies-Bouldin: 0.9490, Calinski-Harabasz: 1929.85
DBSCAN: Only noise or single cluster after excluding noise.
Hierarchical → Silhouette: 0.3483, Davies-Bouldin: 1.0083, Calinski-Harabasz: 1625.45
GMM → Silhouette: 0.3437, Davies-Bouldin: 0.9823, Calinski-Harabasz: 1561.66

Model Comparison Metrics:
```

| | model | silhouette | davies_bouldin | calinski_harabasz |
|---|---|---|---|---|
| **0** | K-Means | 0.385440 | 0.949004 | 1929.850431 |
| **1** | Hierarchical | 0.348310 | 1.008338 | 1625.446913 |
| **2** | GMM | 0.343681 | 0.982294 | 1561.664941 |

***Figure 9: Comparing the model metrics***