National College of Ireland

Tailoring Customer Engagement: Advanced Segmentation for Growth in Garden Business

MSc Research Project
Data Analytics

Punya Nandakumar Kedambadi
Student ID: X23134658

School of Computing
National College of Ireland

Supervisor: Mr. Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Punya Nandakumar Kedambadi
Student ID:	X23134658
Programme:	Data Analytics **Year:** 2024-2025
Module:	MSc Research Project
Supervisor:	Hicham Rifai
Submission Due Date:	29/01/2025
Project Title:	Tailoring Customer Engagement: Advanced Segmentation for Growth in the Garden Business
Word Count:	7663 **Page Count** 22

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: Punya Nandakumar Kedambadi

Date: 29/01/2025

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 computer.	□

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

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

Tailoring Customer Engagement: Advanced Segmentation for Growth in the Garden Business

Punya Nandakumar Kedambadi
X23134658

Abstract

The garden industry is comprising of the seasonal sales patterns and eco-conscious customers, faces challenges to segment the customers who are diverse in their characteristics. This research explores the advanced data mining techniques like K-means, DBSCAN, Hierarchical Clustering, and GMM algorithms integrated with RFM analysis to address the challenges. Using the customer purchasing details of the company "My Dream Garden", the study focuses to evaluate the performance of the algorithms using Silhouette Score, Elbow Method, and Davies-Bouldin Index. The research described DBSCAN method as the most effective to segment the garden customer showcasing the highest Silhouette Score (0.588) confirming that the algorithm can handle noise and irregular cluster shapes. GMM exhibits a score of 0.502 which is a probability clustering while the Hierarchical clustering exhibits a score of 0.477. Despite the popularity of the K-means, this method underperformed for the garden industry to segment the customers with the lowest score of 0.278 struggling hard with the non-linear data. RFM analysis skilfully categorises the customers to "Best Customers" and "At Risk Customers" underlining the need to apply strategies on personalised marketing, customisation and to retain the customer. While DBSCAN proved robust but its reliance on manual parameter tuning and the study's focus on transactional data reveal opportunities for refinement. This research provides a robust framework for customer segmentation in the gardening sector, empowering businesses with insights to enhance customer satisfaction, loyalty, and revenue.

1 Introduction

The Garden industry is experiencing a rebirth as people are connecting back with nature and trying to focus on having a sustainable lifestyle. By 2028, the advancements in the technology have made humans free from their busy working life which allowed them to spend more time on their personal interest and hobbies as envisioned by British economist (John Maynard Keynes; 1932). There is growing popularity for gardening products that led to the creation of more and more companies related to the gardening. In addition, companies nowadays rely on data to optimise their process in general to identify potential business opportunities. In relation to gardening, the preference of customers is huge which makes business segmentation challenging. In addition, a major challenge for this sector is to retain customers and enhance the customer satisfaction to flourish amongst its competitors. Moreover, as highlighted by (Ben et al.; 2015), how the perceptions of consumers are on terms like "eco-friendly" and "sustainable living" plays an important role on the purchase of the products. Marketing campaigns which highlight the benefits on the environment are more beneficial to attract the eco conscious consumers. This will create an opportunity for the

garden business but also gives them a challenge on how to well define, communicate about their products and earn consumer trust.

As per the current business strategies, customer segmentation is an important approach as it allows them to divide the customers into distinct groups who have similar characteristics or pattern. Existing literatures on the customer segmentation covered various industries like retail, ecommerce where the common approaches used was K-means and RFM analysis. Applying these techniques are difficult on this gardening sector due to its seasonal patterns. Traditionally K-means was taken as the primary method for clustering as mentioned in (Kasem et al.; 2023) and (Sarkar et al.; 2024). But also, few of the research have used new techniques like DBSCAN, GMM to be more precise on the clusters. While this technique is widely explored across sectors a very little exploration is done on the garden sector which is the highlight of this study. To address this gap, the paper is aiming to use data mining techniques K-means Clustering, Hierarchical Clustering, DBSCAN (Density-Based Spatial /Clustering of Applications with Noise), GMM (Gaussian Mixture Models), along with RFM (Recency, Frequency, Monetary) analysis to separate the customers and understand the preference so that the marketing and the sales strategies can be made on the gardening business.

This research provides practical insights to the marketing team and sales team along with the owners of the business. Traditional marketing ideas lack the flexibility to address the personalisation's the customers ask. This can be overcome by making use of the data mining and machine learning techniques. This will act as a literature for the segmentation process by using the clustering techniques which is not widely done and gives information about how the customers can be targeted based on the cluster grouping and their characteristics, personalising the marketing strategies, and increasing sales. Segmentation can help to increase the revenue of the company and increase the loyalty among the customers.

Overall, the study is focused on the below research question: How effectively can data mining techniques, such as K-means, DBSCAN, and GMM, along with RFM analysis be applied in the garden business to segment customers, understand their preferences, and predict sales trends?

Major objectives related to the research question are:

- Perform a state-of-the-art literature review about the implementations on other sectors.
- Cleaning and Pre-processing of the data is done, and the data is collected from the My Dream Garden company after receiving a consent to use the dataset.
- Converting the categorical variables and doing Min-Max scaling to numeric features.
- Calculating the RFM scores and then segmenting the customers based on the metric.
- Performing PCA for dimensionality reduction and then segment the customer using the data mining techniques as per the research question.
- The Elbow and Silhouette methods will be used to assess the optimal number of clusters.

This report is structured as follows: Literature review gives an overview of customer segmentation and its applications on various sectors but trying to cover a gap in research on garden business. The methodology demonstrates the pre-processing steps, RFM analysis, clustering techniques and its evaluation criteria for this study. The result sections explain about the findings obtained from each of the clustering methods and concluding which is better for segmenting garden business customers. Finally in discussion and conclusion sections the practical implications, future work and the potential benefits are suggested.

2 Related Work

In the present data driven marketing scenarios, customer segmentation is the first step to understanding the engagement of the customers effectively. This section delves into the related works conducted by various researchers since 2013 in the field of customer segmentation by focusing on all the different sectors other than the garden business. Relevant studies and methodologies are documented systematically and analysed critically so that the base for the present research is built. It's also been observed that, most relied on demographic and behavioural characteristics to segment the customers. Both traditional and advanced data mining techniques which are applied by the other researchers are examined in detail in the following sectors to address the specific challenges and provide solutions to ensure consumer loyalty and trust.

2.1 Eco-friendly, geographic, and psychographic factors for segmentation

A thorough analysis of the consumer's perceptions of terms like "eco-friendly" and "sustainable" revealed that consumers do value and care for the environment, but sometimes these terms are used without clear definitions by the businesses, leading to the creation of doubt and losing trust among the customers. This can harm the business even if the product is genuine as suggested by (Ben et al.; 2015). This study highlights that the business should be able to put out clear message to the customers about environmental benefits to maintain the trust for years according to an online survey of the US and Canadian consumers. However, the primary limitation of those studies is that it lacks practical usage to how to maintain the consistency and forcing the businesses to handle the issues on their own.

(Andreas et al.; 2013), performed segmentation focusing on the behavioural and demographic factors by taking the data obtained by face-to-face interviews of the Horticulture consumers in the Southern Germany. This research enabled to design the region-specific marketing campaigns and understand the regional preferences. The study lacked the integration of using data mining methods as the dataset was too small. Psychographic segmentation for the gardening sector divided the customers into "Native Averse", "Native Curious", and "Native Enthusiasts". Segmentation was able to align with the environment values so that the eco conscious consumers are targeted as suggested by (Alicia et al.; 2024) and the data is collected by the online survey of the US consumers. This method is very much suitable for the niche market where the groups are small, and consumers are known to the businesses but has limitations like scalability and cannot be applied on broader markets.

The summary of the eco-friendly, geographic, and psychographic factors for segmentation is shown in the below Table 1.

Table 1: Eco-friendly, geographic, and psychographic factors for segmentation

Author(s)	Methods Used	Strengths	Weaknesses
(Ben et al.; 2015)	Consumer perception	Build consumer trust	Inconsistent usage of the environmental terms
(Andreas et al.; 2013)	Regional segmentation	Create regional specific market campaign	Lack the usage of advanced data mining methods
(Alicia et al.; 2024)	Psychographic segmentation and behaviours	Reachable to the niche small markets	Limited to specific customer segment

2.2 Traditional and advanced customer segmentation approaches

RFM analysis is a widely used technique used for evaluating the customers based on their buying behaviour. A scoring method is developed to evaluate scores of Recency, Frequency, and Monetary. Finally, the scores of all three variables are consolidated as RFM score as described by (Haiying et al.; 2010). (Joy et al.; 2018), proposed the RFM ranking method that could make the segmentation of customers simple. Also, the authors strongly believed that retaining the customers is more important than finding new customers. After the rankings, K-means for choosing the initial centroid and Fuzzy C methods was implemented on the online retail transactional dataset. Further K-means clustering was widely used for the segmenting customers as mentioned by (Sujatha et al.; 2023) and they have grouped the customers based on their worth to the company. However, this method does not work well with unstructured or non-linear data.

(Israa; 2023), utilised the transaction dataset of online retail in the UK and RFM analysis was implemented along with its comparison with K-means and DBSCAN clustering which divided the customers into six distinct clusters. These methods helped in customising the marketing strategies however, its effective only on transactional data alone and failed to adapt for specific industries and time sensitive markets. Similarly, (Saxena et al.; 2024) implemented unsupervised algorithms like K-means, affinity propagation, and DBSCAN on the mall customer data. The idea was to make use of the basic and complicated ideas in segmenting the customers and is not well suited for the companies who have some constraints on their resources.

Building on this research, (Rahul et al.; 2021) analysed the UK's e-commerce dataset and develop segmentation model using RFM and K-means methods. Clusters were divided into "super customers", "Intermediate customers", and those "on the verge of churning" and a web-based application was developed for the business analysts so that they can analyse their own data's and visualise them for valuable insights. This innovation allows e-commerce start-ups and analysts to target customer clusters effectively, reinforcing the importance of RFM based segmentation in improving customer relationships and strengthen business performance. (Sarkar et al.; 2024) explored on the efficiency of K-means clustering and

provided persuasive evidence that K-means algorithm along with RFM is better for customer segmentation. Consumers were grouped based on their similarities and then personalised marketing can be done to enhance growth.

Multicriteria decision-making with a combination of psychographic and behavioural data, which can increase the accuracy of the segmentation process, was introduced by (Felipe et al.; 2024), and the data was obtained from a multinational healthcare company. RFM method along with the Global Local Net Flow Sorting (GLNF sorting) algorithm was implemented and compared against the K-means method. This helped the businesses where the customer values and their lifestyle are main criteria to help them grow in their business. Although it was performing well by giving precise results, it was suitable only for small scale industries and failed in handling the complex business with more technical expertise.

A detailed review of these customer segmentation methods and their applications across all industries was conducted by (Saumendra et al.; 2022). They had specified that industry specific research and its solutions should be provided to overcome the challenges of scalability and data quality. Their research was useful when relating to the gardening sector as they mentioned that both the transactional and the behavioural data should be considered for the segmentation process to overcome the challenges. To handle these limitations, (Chongkolnee et al.; 2023) made an extension for the hierarchical clustering by combining RFM analysis with Formal Concept Analysis (FCA) which was very much useful for the companies who need the precise segmentation. However, this approach was limited due to its scalability and wide range of applicability. Similarly (Areeba et al.; 2024) implemented the hierarchical clustering on the mall dataset to enhance the shopping experience. This paper acted as powerful tool to provide insights in the retail sector and was well suited for the larger datasets.

(Kasem et al.; 2023) performed the RFM analysis for a digital start up data and K-means algorithm was coupled with Elbow method, Silhouette coefficient, and Gap Statistics method to identify the specific clusters and acted as a robust method. By analysing these clusters meaningful connections can be made to increase the growth of the company. It's an AI-driven approach, and as this approach needs more computational requirements, it's not suitable for small and medium-scale industries. The evolution of big data has transformed the customer segmentation and (Xiaotong et al.; 2024), integrated this approach with Support Vector Machines (SVM) and clustering algorithms. This could be easily implemented on the large and different datasets, but it lacked the practicality.

(Sumallika et al. (2024) checked the quality of the clusters and optimisation using the Elbow and Silhouette Analysis. The main agenda was to check which of the techniques, when combined with K-means, would give the better optimal clusters and the data used was collected from the health care. Both the methods gave the same results, but they were not ideal for the complicated examples.

The summary of the traditional and advanced customer segmentation approaches is shown in the below Table 2.

Table 2: Traditional and advanced customer segmentation approaches

Author(s)	Methods Used	Strengths	Weaknesses
(Joy et al.; 2018)	RFM, K-means, Fuzzy C-means	Simplifies segmentation and emphasizes customer retention	Concentrated on transaction data only
(Sujatha et al.; 2023)	K-means	Groups customers based on their worth to the company	Not suited for unstructured or non-linear data
(Israa; 2023)	RFM, K-means, DBSCAN	Enables customized marketing strategies	Concentrated on transaction data only
(Saxena et al.; 2024)	K-means, Affinity Propagation, DBSCAN	Uses advanced unsupervised algorithm	Unsuitable for resource constrained companies
(Rahul et al.; 2021)	RFM, K-means	Useful for business analysts	Highly dependent on psychographic and contextual analysis
(Sarkar et al.; 2024)	K-means with RFM	Increased efficiency and personalized marketing	Limited applicability in industries with complex customer data
(Felipe et al.; 2024)	RFM with GLNF Sorting	Multi-criteria Approach	Required technical expertise
(Saumendra et al.; 2022)	Literature Review	Highlighting all the industry specific segmentations	No empirical validation
(Chongkolnee et al.; 2023)	RFM with FCA	Precise segmentation	Scalability, applicability
(Areeba et al.; 2024)	Hierarchical Clustering	Useful for retail sectors	lacked broader generalization
(Kasem et al.; 2023)	RFM, K-means, Elbow, Silhouette, Gap Statistics	AI-driven approach	Only for large scale industries
(Xiaotong et al.; 2024)	SVM, Clustering Algorithms	Scalable for large datasets	Not suited for real world scenarios
(Sumallika et al.; 2024)	K-means with Elbow and Silhouette Analysis	Analise cluster quality	Not ideal for complex datasets

2.3 Challenges

Some of the challenges identified related to the garden industry are:

- As highlighted by (Chongkolnee et al.; 2023), they are not suited for a seasonal and diverse customer base, which is so common in the garden industry. The misalignment fails to provide effective marketing strategies.

- Algorithms like K-means and DBSCAN (Saxena et al., 2024; Israa, 2023) struggling with time-sensitive gardening datasets where valuable insights can be lost if not acted upon actively.

- Miscommunication to the customers about environmental values can make the customers break trust as mentioned by (Ben et al.; 2015) and this can hinder the reputation of the company.

- Most of them require computational resources along with technical expertise and retaining the customers is majorly important for this industry.

2.4 Summary

From the overall review of the literature, the traditional methods like RFM and K-means clustering are the most effective across various industries but failed to handle several of challenges related to the gardening sector. Seasonal sales pattern, eco-conscious consumers and the consistent pressure on the personalised marketing makes the industry more challenging. Although algorithms like DBSCAN and GLNF sorting methods can handle the behavioural and psychographic factors, they require technical expertise. They are not feasible for small and medium scale industries as well. All these limitations urge the necessity to perform customer segmentation within the gardening industry by combining all the factors including transactional, behavioural, and psychographic data. Also, there is no research which is specific to the garden industry as lot of focus is on the retail and e-commerce. The paper will try to incorporate robust clustering techniques and validating by the Elbow, Silhouette, and Davies-Bouldin Index. This approach will not only fill the gaps which is observed in this review but also helps the customers to trust this garden industry which will help the company to grow. Overall, it empowers the garden business by helping them to survive in the competitive and environment focused marketplace.

3 Research Methodology

The present study has used CRISP DM methodology to present the customer segmentation process in the garden business using R Studio and the complete procedure is systematically represented in below Figure 1.

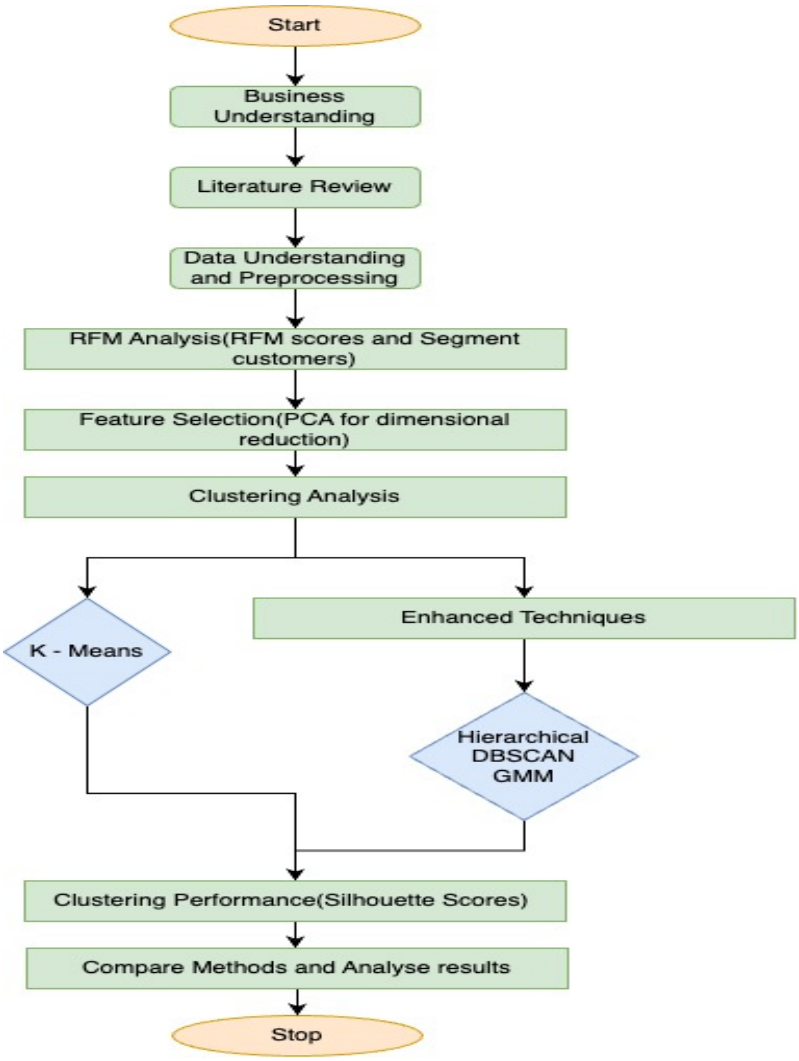

Figure 1: Customer segmentation in garden business.

3.1 Data Acquisition

For this research the data is collected from the company, My Dream Garden located in Bengaluru, India to proceed with the analysis. As the data is available from a non-public site, consent is taken from the CEO on the dataset usage. Obtained data consists of the customer and their purchase details of the gardening packages at that store between the year 2018 to 2024 with an entry of 6569 customers. Mainly the data identifies the customers as NEW or OLD to identify the repeat customers. Dataset overall consists of the demographic (age, marital status, location), purchase behavioural factors (type of customer, product, preferred category), along with the transactional data (price, total value of sale). The data is then further processed using the R programming language and R studio. Most of the variables found in the data are categorical and integer values.

3.2 Data Pre-processing and Transformation

In prior to applying the clustering and segmentation algorithms, the dataset must be extensively pre-processed and transformed so that the quality and scalability of the

algorithms are measured correctly. Below are the important steps which is involved in cleaning the dataset.

3.2.1 Detecting missing/duplicate values and their handling

There were no missing values found in this data and if in case there was any missing data it would have been imputed using the mean/median method for numerical/categorical variables. Eight duplicate entries were removed to make sure that the data used for the model building have unique customer values and now the dataset had 6561 observations.

3.2.2 Exploratory analysis

Visualisations are done so that we understand the data in a pictorial format and get basic insights along with identifying key patterns.

- Plot of monthly and yearly sales trend
  From Figure 2, it is observed that both the monthly and yearly sales increase steadily over time. The peak value of the monthly sales confirms that there are seasonal sales which may be due to the promotional and marketing event and during 2023 the company has observed the maximum sales. However, from 2024, the declining sales has started to raise over the months. From 2022, there is gradual decline of the sales across years, which may be due to the lesser customer retention factors, and till 2024, we can see that the graph is declining. These graphs highlight that the industry should maintain innovative ideas to maintain growth sustainably.

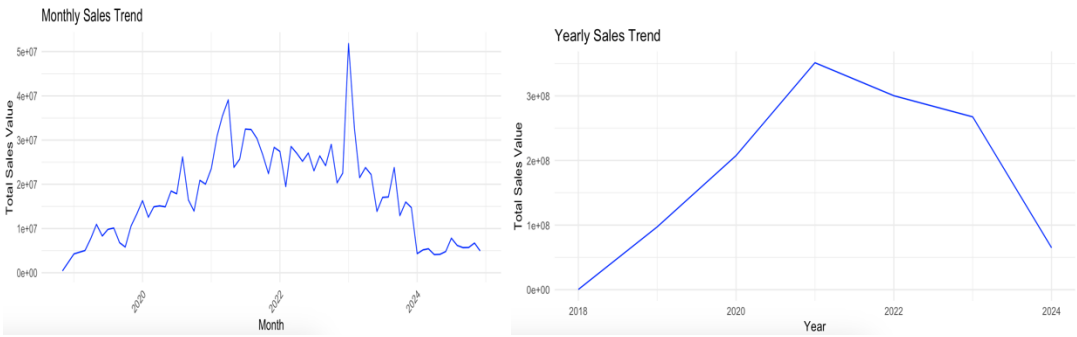

Figure 2: Monthly and yearly sales trend.

- Plot of top 10 products and quantity sold
  Figure 3 depicts the top 10 products by total quantity sold over the years and as per graph “Basic Terrace Garden Package” is the most popular product with the highest quantity of 1324 units sold followed by “Customised Basic”, “Landscaping” and so on. Packages like “Edible Flower Garden Package”, “Basic Package and Customised Balcony”, and “Rooftop Garden Package” also illustrate sales values. This graph says that companies must focus on personalisation and terrace specific products. Also provides insights to the inventory team to work more on the higher performing categories and business strategies.

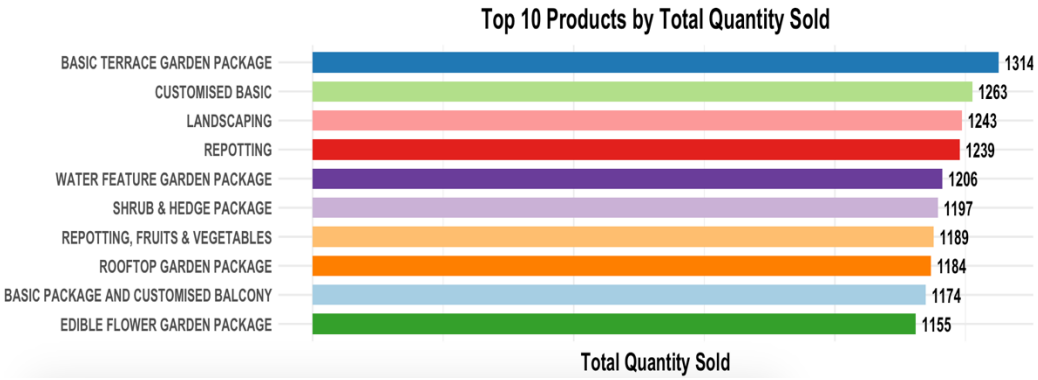

Figure 3: Top 10 products and quantity sold.

- Plot of total sales by type of customer and marital status
  Figure 4 illustrates the distribution of total sales value based on customer type and marital status. By the type of customer, its observed that there is balanced distribution of the OLD and NEW customers. Graph states that the company has maintained consistency, trust and retained the old customers along with attracting the new customers which is very crucial for business growth. It is very clear that there is no such influence based on the marital status as all the people are interested towards the greenery and eco-friendly living.

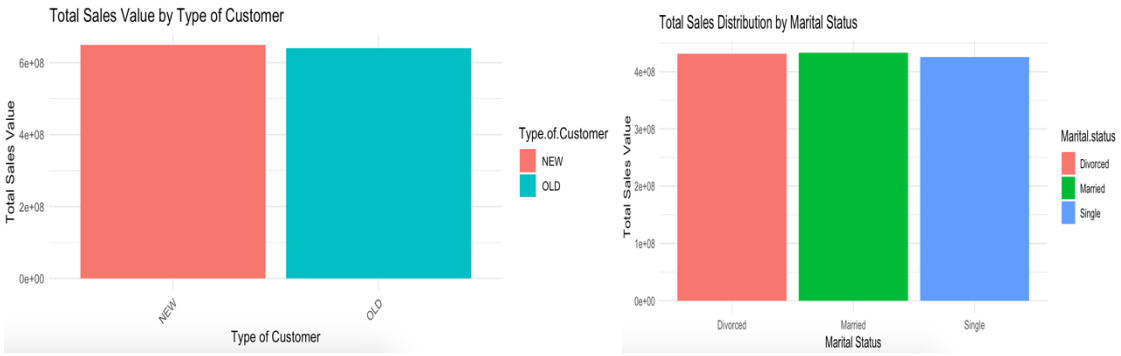

Figure 4: Total sales by type of customer and marital status.

- Plot of sales distribution by age group
  Figure 5 illustrates the distribution of total sales across different age groups. Boxplot is used to represent the median sales value across all age groups and there is no difference in their spending behaviour. Age groups 40–49, and 60–69 show a slightly higher values in the median sales indicating that this age group might be more inclined towards gardening or have higher purchasing capacity. We can observe that there are outliers present which might be due to bulk purchasing orders done by set of customers and removing these outliers are necessary as they tend to hinder the algorithm performance.

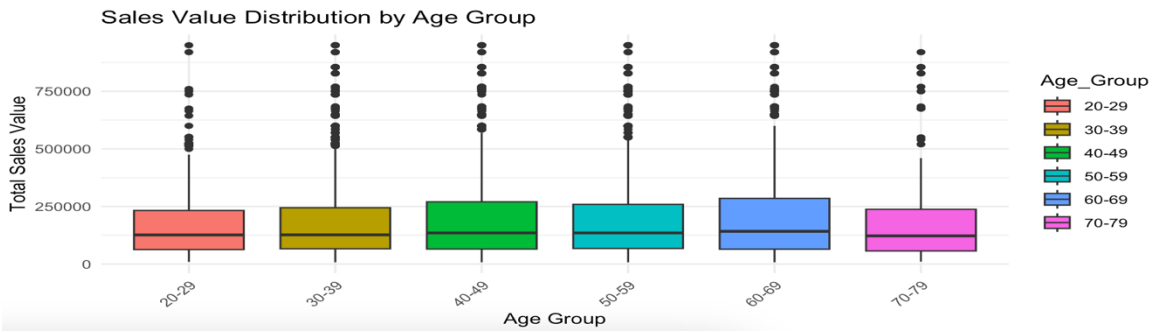

Figure 5: Sales distribution by age group.

3.2.3 Outlier detection and handling

The categorical variables are converted into factorial as the algorithms effectively work on the numerical or factorial variables. As shown in Figure 5, the outliers must be removed so that the integrity of data is preserved, and the skewing of the algorithms results are prevented, and this becomes the most crucial step on the data pre-processing. For this research, we have used both the combination of the Z-score method and the Interquartile Range (IQR) method to rectify and remove so that the global and local outliers are simplified. As per Z-score method the data points greater than +3 or less than -3 are considered as outliers and are eliminated so that they do not have any influence on the model building. Once these are removed, IQR method is calculated again and data points outside the range are removed as well.

3.2.4 Feature normalisation

Making sure that all the variables do contribute to the analysis of the research question, Min-Max Scaling is applied on the numeric columns where they are scaled between 0 and 1. As per this study, normalisation step is beneficial as the clustering algorithms and machine learning models depend upon calculating the distances between data points.

3.3 RFM analysis

The RFM (Recency, Frequency, and Monetary) analysis measures the three important purchase-related variables which highly influence the future purchases of the customers, as mentioned by (Birant, D; 2024) and acts as a powerful tool to segment the customers based on transaction history. Once the customers are segmented, different marketing strategies can be applied on each of the segments for the growth of the business and maintaining trust with the customers. The three main metrics are:

- Recency: How recently was the purchase made by customer.
- Frequency: How often customer purchases or number of transactions made.
- Monetary: Cumulative total of money spent by a particular customer.

RFM values are calculated by grouping the customer ids and each of the metric is assigned a score between 1 to 5 using the ntile() [1] function which is sort of very rough rank, which breaks the input vector into "n" buckets. Customers are then ranked into five groups based on RFM scores where higher values are considered better for F and M scores and lower values are better for R scores and an aggregated RFM score is calculated. Customer with higher scores are the most valuable customers for the business and are segmented into 4 segments as following:

- Best Customers (RFM Score ≥ 9): High value and loyal customers.
- Loyal Customers (RFM Score 6-8): Regular customers with consistent purchases.
- Potential Customers (RFM Score 4-5): Mid-range customers with potential growth.
- At Risk Customers (RFM Score < 4): Customers who may be at risk of churning.

Figure 6 provides a visual representation on how the customers are segmented based on the purchasing habits. The chart explains that the "Best Customers" have higher values of RFM and constitute to the bigger group highlighting that they are the most engaging and valuable customers. The "Loyal Customers" are also consistent in their purchase but are not frequent as that of best customers. Meanwhile, the "Potential Customers" have slighter lower value indicating that targeted marketing strategies and more engagements need to be done on this set of customers. Lastly, the "At Risk Customers" are the smallest portion and are the least engaged customers who need more attention and maximum satisfaction should be provided so that the company does not lose the customers.

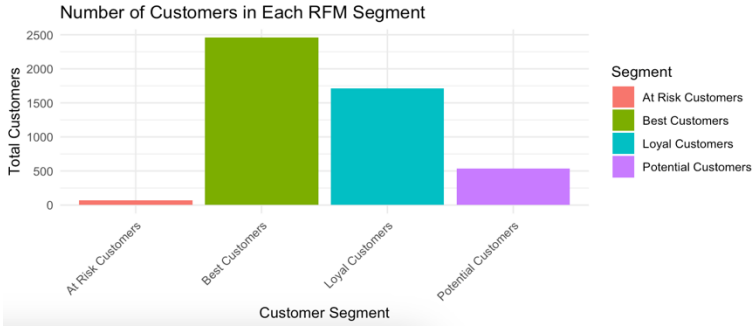

Figure 6: Number of customers distributed in each segment.

Figure 7(Left Chart) illustrates the average RFM values for each segment. "At Risk Customers" show high recency value saying they are not the frequent engagers in the purchase and have very low values of frequency and monetary values. The "Best Customers" show lowest recency value amongst all saying they are the most engaging, spending, and valuable customers. "Loyal Customers" and "Potential Customers" demonstrate moderate levels across all metrics, suggesting they are consistent but not as valuable as the best customers.

1 https://search.r-project.org/CRAN/refmans/dplyr/html/ntile.html

Figure 7(Right Chart) illustrates average monetary value for each customer segment where the “Best Customers” contribute to the highest revenue for the company, making them the most valuable group to retain and reward. “Loyal Customers” have consistent spending habits and help to maintain stable revenue for the company. “Potential Customers” show moderate monetary value highlighting that more opportunities and engagement needs to be done on these customers and in contrast, “At Risk Customers” exhibit the lowest monetary contribution and planning must be done to revive their purchasing activity.

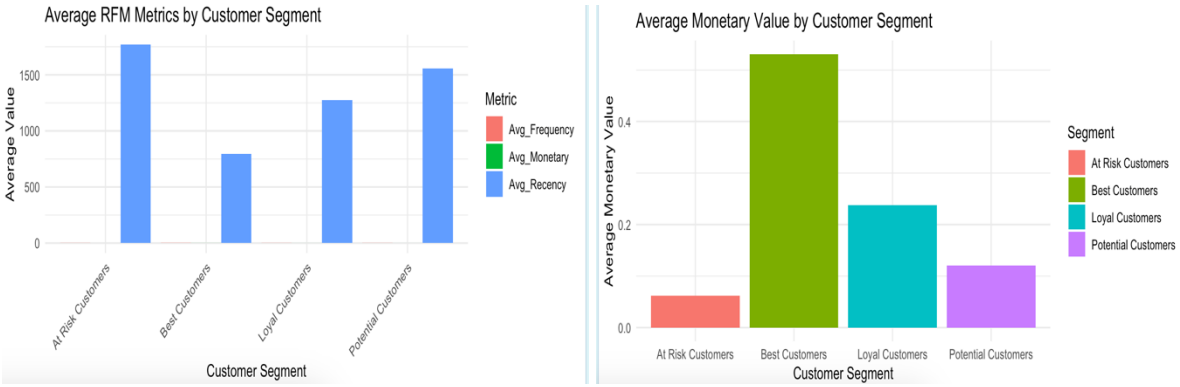

Figure 7: Average RFM metrics and monetary value based on the customer segment.

3.3.1 Feature selection

Principal Component Analysis (PCA) was applied to the numeric features of the dataset to reduce dimensionality while retaining most of the information and all the features were standardised to maintain the features on same scale. As shown in Figure 8, a variance explained plot was generated to visualize the proportion of variance captured by each principal component (PC). The components which collectively contribute to the 90% of the total variance are retained so that there is lesser redundancy. Later the transformed dataset is comprised of the selected components which is performed by eliminating noise and reducing computational complexity without losing the important data which contributes well to the algorithm efficiency.

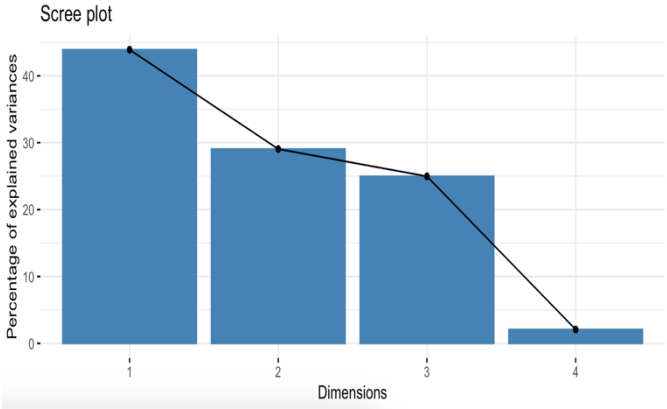

Figure 8: Scree with the dimensions.

4 Design Specification

The overall architecture of the customer segmentation in the garden business is divided into three main phases, which are explained in detail below:

- Data Layer
  This layer is responsible to acquire, store and pre-process the customer data along with their purchase history. The data was sourced from My Dream Garden, Bengaluru, India, with permission from the organization. Pre-processing steps was performed using R Studio as it is mainly used for the statistical computing making it highly efficient to do any transformations and identify outliers. With packages like dplyr, tidyr, and data.table, the tasks of cleaning the data is made simpler and these steps are important to prepare the dataset ready for model building. All pre-processed data was stored on cloud or local platforms for scalability and accessibility.

- Application Layer
  This layer focuses on the customer segmentation and clustering processes using RFM analysis. Advanced clustering methods such as K-means, DBSCAN, Hierarchical Clustering, and GMM, were applied to segment customers further. These clustering models were validated using evaluation metrics like the Elbow Method, Silhouette Analysis, and Davies-Bouldin Index to determine optimal cluster numbers and assess model performance. R packages (factoextra, cluster) were utilized for efficient model building and evaluation.

- Presentation Layer
  This layer concentrates on obtaining the analytical results by the aid of visualisations and reporting tools. The algorithms are displayed through scatterplots, dendrograms, and Silhouette plots to assess and compare cluster quality. Tools like ggplot2 and factoextra are utilized for clear, aesthetic visualizations, while interactive dashboards can be implemented using Tableau or Power BI to empower marketing and sales teams with insights tailored to customer segments, enabling personalized campaigns and optimized inventory planning.

5 Implementation

As clustering is an unsupervised learning technique, the dataset is not divided into training and testing subsets. These algorithms make use of the fundamental patterns which is required for the groupings so that there is need to predefine the variables as target variable. The entire dataset is used in cluster formation and validation. This approach eliminates the need for data splitting as the main goal is not to predict but to identify meaningful patterns by making use of the already existing patterns.

5.1 K-means Clustering

K-means clustering is the basic and widely used iterative approach to divide the dataset into meaningful clusters relying upon the similarity of the features. The method assists in minimising the intra-cluster variance. The dataset which underwent the pre-processing steps along with PCA analysis was utilised and two methods was used to obtain the optimal number of clusters.

- Elbow Method was used to evaluate the Total Within-Cluster Sum of Squares (WSS) for different numbers of clusters (K) in Figure 9.
- Silhouette Analysis was used to check the quality of the clusters for different values of K in Figure 9.

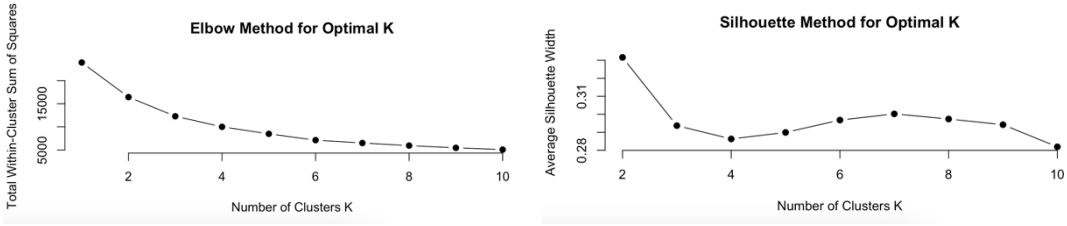

Figure 9: Elbow and Silhouette for optimal K.

Both these methods identified the K value to be 6 as shown in Figure 10, and using this optimal K=6, final K-means clustering algorithm is implemented where each data point is assigned to the cluster with nearest centroid, and this happens iteratively until the centroids converge. Also, the Davies-Bouldin Index, gave the lowest DBI score with K=6 confirming that the clusters are well separated.

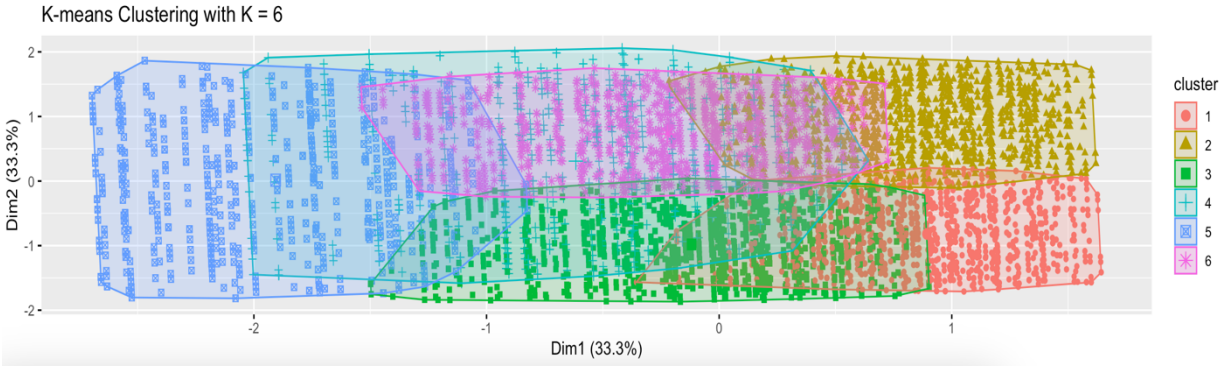

Figure 10: K-means clustering with K=6.

5.2 Hierarchical Clustering

Hierarchical clustering was implemented to classify the groups by merging or splitting the clusters and this is one of the advanced clustering techniques. A distance matrix was calculated on the PCA transformed data which quantifies the pairwise dissimilarities between data points. Ward's method was used to perform the clustering as it reduces the variances within the clusters during each step of merging. This method helps to ensure that the clusters

formed are compact and homogenous. The clusters are visually represented using Dendrogram which clearly divided the six hierarchical clusters using the K=6 maintaining the consistency with K-means.

5.3 DBSCAN Clustering

DBSCAN is used to detect the dense clusters by handling the noise within the dataset and the clusters are separated as per their lower density. The Epsilon (eps) and minPts (minimum points) are tuned over iterations and these parameters help to identify the Silhouette scores. As the algorithm concentrate on the denser clusters, only the core points are used in the cluster formation while the border points for each cluster are assigned as per their proximity to the core points. A value of minPts=5 was chosen and outliers or the points having noise are excluded in this method. DBSCAN method provides extra flexibility as it can identify the arbitrary-shaped clusters and the clusters formed by DBSCAN algorithm are illustrated in Figure 11 and showed clear and well-separated clusters.

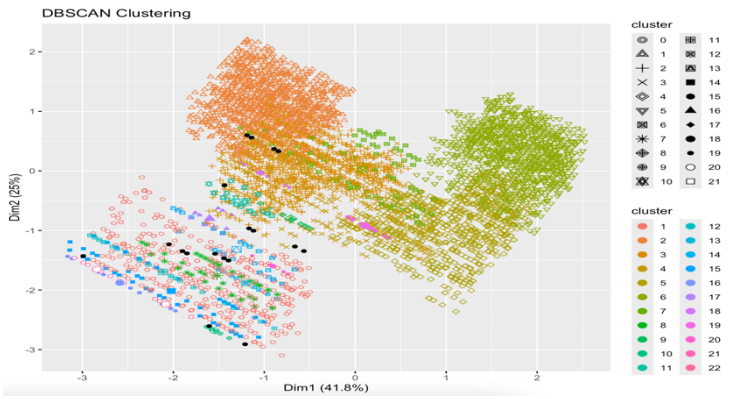

Figure 11: DBSCAN clustering.

5.4 GMM Clustering

GMM is a probabilistic clustering approach that assumes the data points are created by combining the several Gaussian distributions with unknown parameters. The number of components (G=6) was pre-defined based on the previous analysis such as the Elbow and Silhouette methods. The Mclust package is used to estimate the parameters of the Gaussian distributions and assigns data points to the most probable component. Each data point is assigned to a cluster based on the highest posterior probability of belonging to a Gaussian component. This ensures a soft clustering approach, where probabilities of cluster membership are computed and can be plotted using fviz_mclust to represent the clustering results as shown in Figure 12. As the method assumes Gaussian distributions, the resulting clusters tend to look like ellipse.

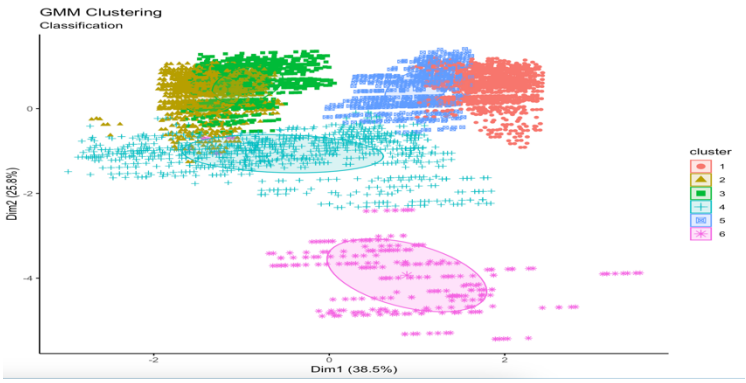

Figure 12: GMM clustering.

6 Evaluation

The section dives into the comprehensive evaluation of the results obtained from the four algorithms—K-means, Hierarchical Clustering, DBSCAN, and GMM. The analysis is mainly focused on the Silhouette scores and are visually represented with the Silhouette plots for each of the algorithms. This evaluation emphasises that the results are aligned with the research objective concentrating on both statistical and practical implications for the results.

6.1 K-means Clustering

Clustering was rerun with K values of 2, 4, and 5 to confirm robustness and consistently validated that the optimal value of K is 6. Post clustering, each cluster was profiled to get some meaningful insights and to identify their unique characteristics, such as average age, frequency, monetary value, and recency metrics. Silhouette score is calculated with K=6 for K-means clustering and the average Silhouette score is 0.2784522.

As depicted in Figure 13 the Silhouette width ranges between -1 to 1 and helps with measuring the separation and cohesion of the clusters with each other. Positive values closer to 1 indicate that the points are tightly grouped. The red dashed line highlights the average Silhouette score for K=6 suggesting good separation. Also, all the points have positive Silhouette widths indicating that the quality of the clusters are satisfactory. Overall, this Silhouette analysis confirms that K=6 provides fair clustering, aligning with other methods like the Elbow Method and Davies-Bouldin Index.

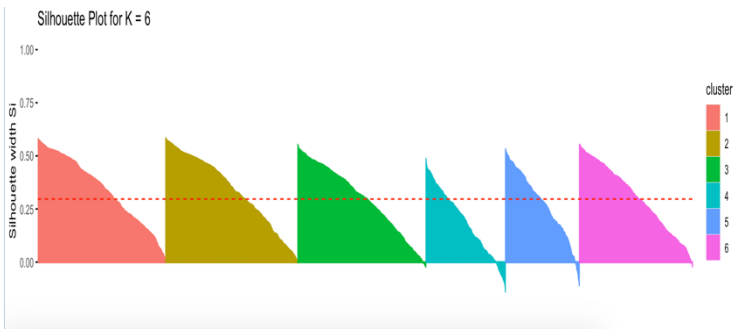

Figure 13: Average Silhouette Score for K means.

6.2 Hierarchical clustering

Silhouette score is calculated with K=6 for Hierarchical clustering and the average Silhouette score is 0.4766735 and is clearly represented by the red dash line as shown in Figure 14. The six clusters formed by cutting the dendrogram indicated there is a reasonable solution to the clustering and there is relationship between the clusters at each level. This method does not require a predefined number of clusters when compared to K-means.

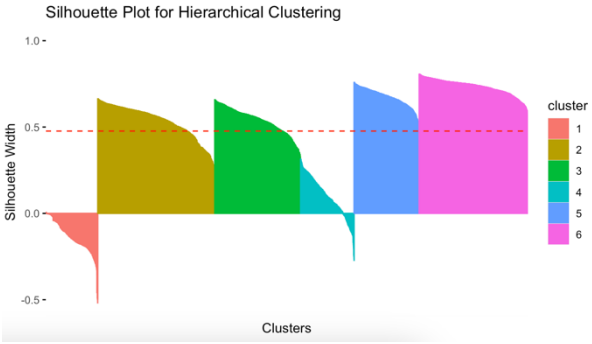

Figure 14: Average Silhouette Score for Hierarchical clustering.

6.3 DBSCAN Clustering

DBSCAN produced the highest scores, indicating well-separated clusters with clear boundaries. The performance of this clustering comprises of both the qualitative and quantitative factors as it focuses on the dense regions of data. Points outside the dense regions are labelled as noise. Silhouette score is calculated excluding the noise for DBSCAN clustering and the average Silhouette score is 0.5881333 and is highlighted using the red dashed line shown in Figure 15. The results are too sensitive for the correct choice of eps and minPts as incorrect choosing of parameters could lead to over clustering or under clustering.

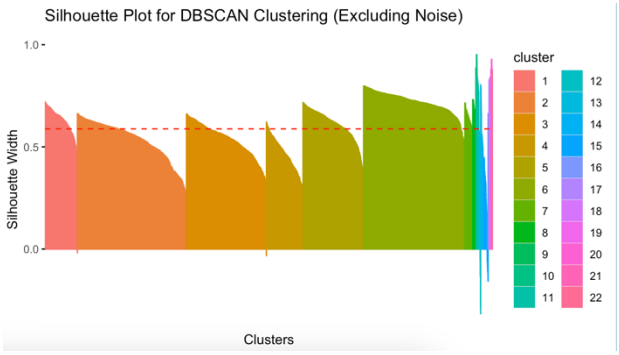

Figure 15: Average Silhouette Score for DBSCAN clustering.

6.4 GMM Clustering

The evaluation of the GMM clustering mostly focus on the quality and coherence values of the clusters. The Silhouette score for GMM clustering was calculated to see how well each

data point will fit inside the assigned clusters when comparing them with the other clusters and the average Silhouette score is calculated to be 0.5021256 as highlighted in the Figure 16 by a red dashed line. The score obtained is moderate and say that some of the clusters do overlap with each other. Silhouette score showed that GMM's clusters were slightly less well-defined compared to DBSCAN, which excelled at handling noise and irregular shapes.

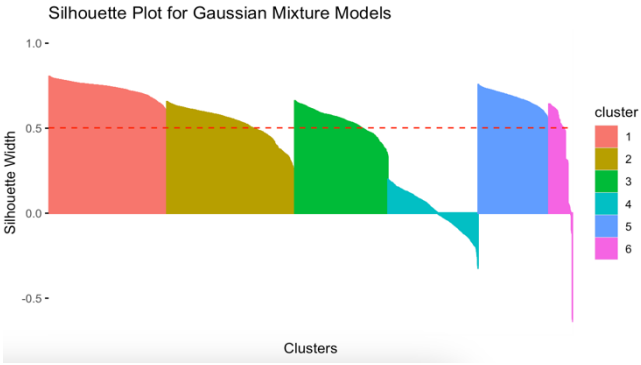

Figure 16: Average Silhouette Score for GMM clustering.

6.5 Comparison between models

The bar graph in Figure 16 demonstrates a comparison of the average Silhouette scores across four clustering algorithms: DBSCAN, GMM, Hierarchical Clustering, and K-means. Amongst all, DBSCAN achieves the highest average Silhouette score of 0.588 and it is the better cluster algorithm to segment the customers in garden business. GMM and Hierarchical Clustering follow with scores of 0.502 and 0.477, respectively, reflecting relatively good performance. In contrast, K-means scores the lowest at 0.278, suggesting it is not feasible to segment the customers effectively in this business. This comparative analysis emphasises the strength of DBSCAN to provide valuable insights based on the clusters and is beneficial in the garden business.

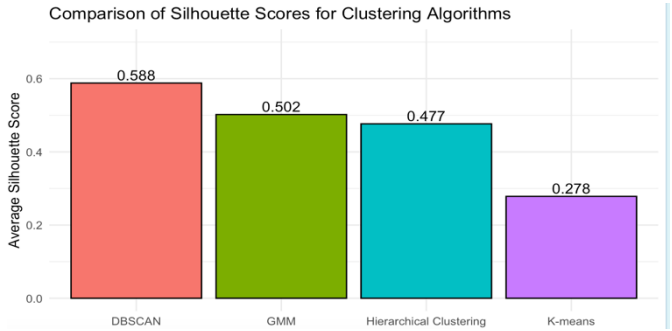

Figure 16: Comparison.

6.6 Discussion

The evaluation of the clustering algorithms helped to get insights regarding the strengths and weaknesses where the algorithms are applied to segment the customers in garden business. DBSCAN showed the highest Silhouette score and for this research, DBSCAN method is the most effective approach exhibiting the capability to handle the irregular cluster shapes and

noise which is so common in the real-world dataset. But DBSCAN completely relies on the parameter autotuning which affects the generalizability across other datasets and indicates the necessity to perform automated or adapted tuning mechanisms to increase the robustness of the algorithms. GMM assumes the Gaussian distribution of data and showed good performance but the clusters overlapping and the algorithm relying on the assumptions based on probability makes it less effective than DBSCAN. GMM assumptions are beneficial only for soft clustering but its less effective with gardening sector as the customers show complex purchasing behaviours.

Hierarchical Clustering is useful in understanding the relationship between the clusters as they are hierarchically distributed but showed moderate Silhouette score. This method requires more computational cost, is very sensitive to noise and not productive on larger datasets. Also, the algorithms lack flexibility to define the clusters after the merging and splitting process making it very less adaptive when compared to DBSCAN. K-means being the most used algorithms due to its simplicity, efficiency, showed the lowest Silhouette score and struggled to separate the clusters. The algorithms were dependent on spherical cluster assumptions and sensitive to the centroid, made it less applicable for the garden dataset. It could not handle noise and not applicable on the larger dataset. As per the literature review, K-means and RFM methods are the most dominant methods in all the other industries and thus, the research highlighted the necessity to explore DBSCAN for gardening sector. Customer diversity and seasonal patterns make this dataset different and hence requires an algorithm which is more flexible and robust where K-means completed struggled. GMM and Hierarchical Clustering showed better performance but were limited to handle noise suggesting that hybrid or ensemble methods can be implemented which can enhance the effectives of the clustering algorithms. Additional metrics like Calinski-Harabasz Index or Dunn Index can be calculated to further validate the results of the clustering algorithms.

7 Conclusion and Future Work

The research explored advanced data mining techniques—K-means, DBSCAN, Hierarchical Clustering, and GMM—combined with RFM analysis to segment customers in the gardening industry where it had seasonal sales patterns and eco conscious customers. The analysis of the “My Dream Garden” data demonstrated that DBSCAN algorithm outperformed the other techniques by obtaining the highest Silhouette score of 0.588, making it a best possible solution to identify the separated clusters and reduce noise. GMM, an approach based on probability also performed well with Silhouette score of 0.502, while Hierarchical Clustering also provided valuable insights with Silhouette score of 0.477 and K-means performed the least with Silhouette score of 0.278. The RFM analysis was able to identify and group the customers like “Best Customers” and “At Risk Customers” suggesting the company to concentrate more on personalised marketing so that the customers are retained. While the study demonstrated the strengths of advanced clustering methods, limitations like DBSCAN's sensitivity to parameter selection and the reliance on transactional data indicate the need for further refinement.

By combining the RFM scores with the clustering algorithms a robust framework can be built to enhance customer segmentation to help the business increase revenue and provide

customer satisfaction. Additionally, while PCA successfully reduced dimensionality, some data attributes may have been excluded, potentially affecting clustering precision. Overall, this study underscores the strengths of advanced clustering techniques in addressing the unique challenges of the gardening industry. By segmenting customers effectively and providing actionable insights, the research contributes significantly to the development of personalized marketing strategies and optimized resource allocation.

Weather patterns can be included to enrich the customer segmentation insights. Automating the DBSCAN parameters and using the hybrid models with combination DBSCAN, GMM and Hierarchical Clustering can be explored to improve the performances of the algorithms. Interactive dashboards can be implemented to visualise the clusters and enable real time segmentation to empower the business and help them grow in their business and gain customer trust.

8 Acknowledgement

The author would like to appreciate Mr. Hicham Rifai for the valuable assistance towards the completion of this work. His expertise helped in honing both the technical and report writing fronts. Also, would like to thank My Dream Garden for providing the dataset for this research.

References

Alicia, L.R., Ariana, P.T., Bridget, K.B. and Sue, B. (2024). Unwrapping the Native Plant Black Box: Consumer Perceptions and Segments for Target Marketing Strategies, *Hort Technology* 34(3):361-371.

Andreas, G. and Klaus, M. (2013). Segmentation of Customers of Horticultural Non-food Products in Southern Germany, *German Journal of Agricultural Economics* 62(3):192-202.

Areeba, A., Laiba, K., Muhammad, Z.H., Muhammad, Z.H, Muzzamil, M. and Aqsa, K. (2024). Customer Segmentation Using Hierarchical Clustering, *IEEE 9th International Conference for Convergence in Technology (I2CT)*.

Ben, C., Hayk, K. and Bridget, K.B. (2015). Consumer Perceptions of Eco-friendly and Sustainable Terms, *Agricultural and Resource Economics Review* 44(1):21-34.

Chongkolnee, R., Pakwan, R., Arthit, I. and Khanchit, C. (2023). RFM model customer segmentation based on hierarchical approach using FCA, *Expert Systems with Applications* 237(5):121449.

Birant, D. (2011). Data Mining Using RFM Analysis, *Knowledge-Oriented Applications in Data Mining*, Prof. Kimito Funatsu (Ed.), ISBN: 978-9.

Felipe, B., Marina, S. and Maroto, C. (2024). Multiple criteria decision support system for customer segmentation using a sorting outranking method, *Expert Systems with Applications*.

Haiying, M. and Yu, Guo. (2010). Customer Segmentation Study of College Students Based on the RFM, *The International Conference on E-Business and E-Government*, Guangzhou, China, Proceedings.

Israa, L. (2023). Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis, *Journal of Data Science and Intelligent Systems* 2(1).

John Maynard Keynes. (1932). Economic Possibilities for our Grandchildren, *Essays in Persuasion, New York: Harcourt Brace*: 358-373.

Joy, C., Arumugum, U., Priyatharsini, L. and Neyaa, A. (2018). RFM Ranking – An Effective Approach to Customer Segmentation, *Journal of King Saud University - Computer and Information Sciences* 33(2).

Kasem, M.S., Hamada, M. and Eddin, I.T. (2023). Customer profiling, segmentation, and sales prediction using AI in direct marketing, *Neural Computing & Applications* 36: 4995–5005.

Rahul, S., Laxmiputra, S. and Saraswati, J. (2021). Customer Segmentation using RFM Model and K-Means Clustering, *International Journal of Scientific Research in Science and Technology*.

Sarkar, M., Puja, A.R. and Chowdhury, F.R. (2024). Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis, *Journal of Business and Management Studies* pp. 54-60.

Saumendra, D. and Janmenjoy, N. (2022). Customer Segmentation via Data Mining Techniques: State-of-the-Art Review, *Computational Intelligence in Data Mining* pp. 489–507.

Saxena, A., Agarwal, A., Pandey, B.K. and Digvijay, P. (2024). Examination of the Criticality of Customer Segmentation Using Unsupervised Learning Methods, *Circular Economy and Sustainability 4*, 1447–1460.

Sujatha, V., Shaik, N., Taduvai, S.L. and Sravya, M. (2023). Customer Segmentation Using K-Means Clustering, *Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems* pp. 457-466.

Sumallika, T., Alekya, V., Raju, P.V.M., Raja Rao, M.V.L.N., Gnnana Shiney, D.E. and Vijaya Sudha, M. (2024). Exploring Optimal Cluster Quality in Health Care Data (HCD): Comparative Analysis utilizing k-means Elbow and Silhouette Analysis, *International Journal of Chemical and Biochemical Sciences* (IJCBS), 25(16) (2024): 48-60.

Xiaotong, L. and Young, S.L. (2024). Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm, *Journal of Cases on Information Technology* (1):1-16.