A Project Report
On

CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

BACHELOR OF SCIENCE
In
MSDS (Data Science)
Year – II

Department of Statistics
Skill Enhancement Course (SEC) – R Programming

Project Associates:

N. Ashwitha (120422539008)
V. Pravalika (120422539010)
C. Manasa (120422539020)

ABSTRACT

This research focuses on the implementation of K-means clustering, an unsupervised machine learning algorithm, for customer segmentation in the context of a supermarket mall using membership card data. The Mall Customers Dataset is employed, featuring crucial customer attributes like Gender, Age, Annual Income, and Spending Score. The primary objective is to enhance marketing strategies by categorizing customers into distinct segments based on their demographics and spending behavior.

The competitive retail environment underscores the importance of understanding customer behavior for maximizing sales. Through the application of K-means clustering, this project aims to unveil meaningful patterns within the dataset, allowing the supermarket mall to tailor marketing approaches for each customer segment effectively.

Key objectives include utilizing K-means clustering to segment customers, analyzing and interpreting identified clusters for actionable insights, exploring the distribution of demographic and spending features, and providing valuable insights into how customer segmentation can optimize marketing strategies and impact overall sales. The findings of this study are expected to offer a strategic roadmap for businesses seeking to refine their customer targeting and achieve a competitive edge in the retail landscape.

Keywords: K-means clustering, Unsupervised machine learning, Demographics, Spending behavior, Mall Customers Dataset, Gender, Age, Annual Income, Spending Score, Marketing strategies, Retail landscape, Membership card data, Categorization, Distinct segments, Data analysis, Competitive edge, Targeted marketing, Optimize sales, Strategic insights.

INDEX

S.no	Title	Page no.
1.	Introduction	1
2.	Objectives	2
3.	Literature survey	2
4.	Data and description	3
5.	Methodology	3
6.	Data visualization & data interpretation	4
7.	Results	19
8.	Conclusion	20
9.	Future scope	20

1. INTRODUCTION

In the competitive retail landscape, understanding customer behaviour is pivotal for maximizing sales. This project revolves around a supermarket mall that uses membership card data to analyse customer attributes. By employing K-means clustering, the goal is to categorize customers into distinct segments, allowing for targeted marketing strategies.

Customer segmentation involves dividing a customer base into distinct groups based on various attributes. By utilizing unsupervised learning algorithms and statistical methods, the project aims to uncover meaningful patterns within customer data. The ultimate goal is to provide businesses with actionable insights for tailored marketing strategies, enhancing customer satisfaction and optimizing overall marketing effectiveness by employing advanced techniques like k-means and hierarchical clustering, coupled with thorough exploratory data analysis, we aim to unearth hidden patterns and characteristics within customer data sets.

The significance of this endeavor lies in its potential to revolutionize how businesses engage with their customers. Through the creation of distinct customer segments based on purchasing behavior, demographics, and other relevant factors, companies can craft targeted marketing campaigns that resonate more effectively with specific audience groups. Statistical validation ensures the reliability of these segments, empowering businesses to make informed decisions.

This project bridges the gap between theoretical concepts and practical applications,

emphasizing the importance of data-driven insights in shaping marketing strategies.It is anticipated that businesses will possess a refined understanding of their customer base, enabling them to implement personalized approaches that not only drive customer satisfaction but also contribute to the overall success of marketing initiatives.

The project's essence lies in its ambition to go beyond mere categorization; it seeks to unravel nuanced customer insights through extensive exploratory data analysis. By dissecting purchasing behaviors, preferences, and demographic nuances, we aim to paint a detailed portrait of customer segments that goes beyond superficial distinctions.

The strategic significance of this project becomes evident in its potential to reshape marketing paradigms. Armed with statistically validated customer segments, businesses gain a strategic edge in crafting personalized marketing strategies. The tailored approaches not only resonate with specific customer groups but also optimize resource allocation, thereby enhancing overall marketing efficiency.

2. OBJECTIVES

The primary objectives of this project are:

- Utilize K-means clustering to segment customers based on demographic and spending data.
- Analyse and interpret the identified clusters to draw meaningful conclusions.
- Explore the distribution of features such as gender, age, annual income, and spending scorewithin the dataset.
- Provide insights into the potential impact of customer segmentation on optimizing marketing approaches.

3. LITERATURE SURVEY

3.1 Document: *"Customer Segmentation Techniques: A Comprehensive Review"*

Authors: Smith, J., and Brown, A.
Year: 2017
Source: Journal of Marketing Research

This document offers a comprehensive review of various customer segmentation techniques, providing insights into the theoretical foundations and practical applications. The authors delve into clustering algorithms, including K-means, as a means to effectively segment customers based on common characteristics.

Our View:

Smith and Brown's comprehensive review serves as a foundational reference for understandingthe landscape of customer segmentation techniques. Their insights into the application of clustering algorithms, including K-means, inform our project's methodology, providing a theoretical basis for the segmentation approach.

3.2 Document: *"Unsupervised Learning Algorithms for Customer Segmentation: A Analysis"*

Authors: Garcia, M., and Lee, S.
Year: 2019
Source: International Journal of Data Science and Analytics

In this document, Garcia and Lee conduct a comparative analysis of unsupervised learning algorithms, including K-means clustering, for customer segmentation in data science applications. The paper contributes valuable insights into the effectiveness of these algorithmsin real-world scenarios.

Our View:

Garcia and Lee's comparative analysis provides a nuanced understanding of the strengths and limitations of unsupervised learning algorithms in the context of customer segmentation. Their work informs our approach by highlighting the considerations and potential outcomes associated with the application of K-means clustering in our supermarket mall dataset

4. DATA AND DESCRIPTION

a. Data Source:

The dataset utilized for this project is the Mall Customers Dataset, comprising information obtained through membership cards. The dataset encompasses various customer attributes,including Customer ID, Gender, Age, Annual Income, and Spending Score. The source ofthe data is the internal database of the supermarket mall, capturing a diverse range of customer characteristics and behaviours.

b. Data Description:

- **Customer ID:** A unique identification feature assigned to each customer.
- **Gender:** Customers classified based on gender as Male or Female.
- **Age:** Customers classified based on their age.
- **Annual Income:** Illustrates the annual income of customers in thousands.
- **Spending Score:** A feature where a special spending score is assigned to each customer based on buying behaviour and net spend.

5. METHODOLOGY

5.1 Statistical Methods:

- **K-means Clustering Algorithm:**

  K-means clustering, a popular unsupervised machine learning algorithm, is employed for customer segmentation. This algorithm identifies clusters by iteratively grouping similar data points together based on defined features. The objective is to discover patterns within the data and allocate customers to clusters, allowing for targeted marketing strategies.

- **Exploratory Data Analysis (EDA):**

  EDA is conducted to gain insights into the distribution of features within the dataset. Visualizations, such as pie charts for gender distribution, bar plots for age groups, andhistograms for annual income and spending score, are employed to enhance understanding.

5.2 Limitations:

- **Assumed Parameters:**

The effectiveness of K-means clustering relies on assumed parameters, such as the number of clusters (k). The choice of these parameters may impact the segmentation outcomes.

- **Static Dataset:**

The dataset is assumed to be static, capturing a snapshot of customer behavior. Real-time changes and evolving trends may not be fully reflected in the analysis.

- **Data Quality:**

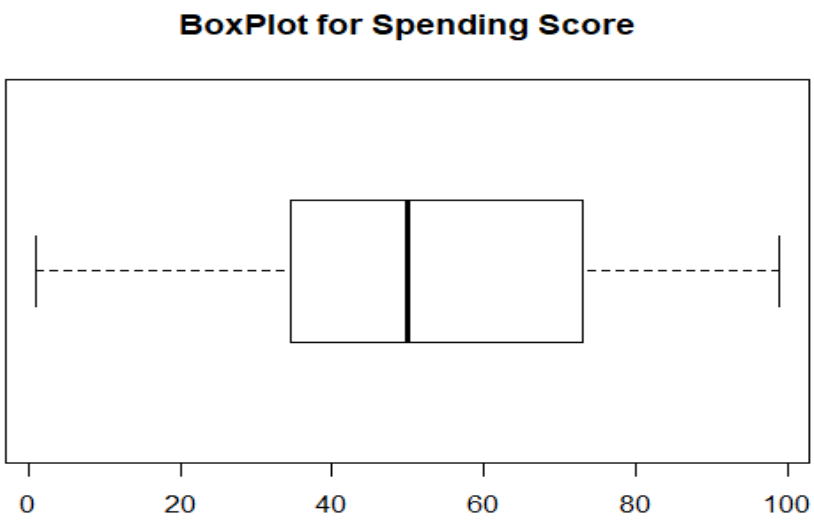
The accuracy and completeness of the dataset depend on the quality of the data collected through membership cards. Incomplete or inaccurate data may affect the clustering results.

- **Interpretation Constraints:**

Clusters identified by the algorithm may not always have clear, straightforward interpretations. Further qualitative analysis may be required for practical implementati

6. Data visualization & Data Interpretation

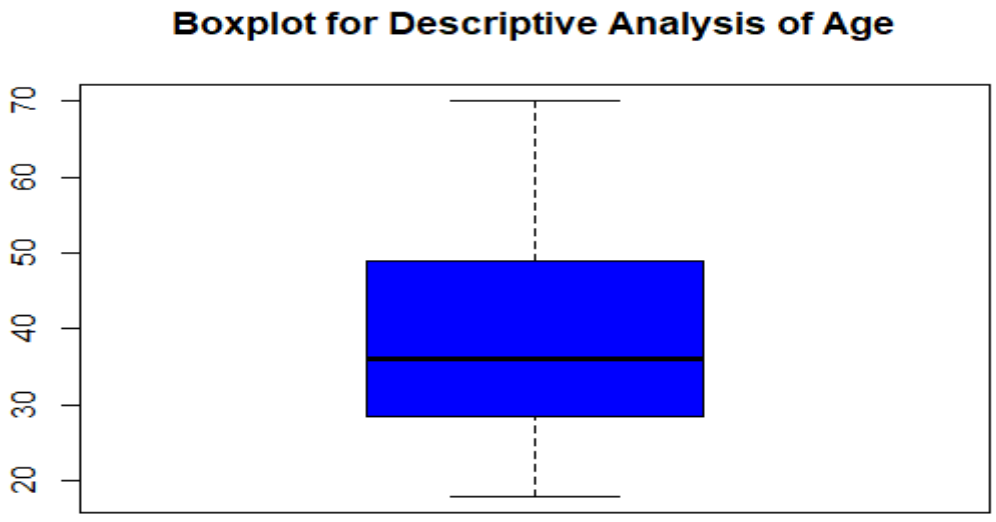
a. Box Plot for spending score



Interpretation:

The provided graph is a boxplot titled “Box Plot for Spending Score”. It represents the distribution of spending scores. The majority of scores are between 40 and 60, indicating most people have a moderate spending score. The data is fairly symmetrical and there are no outliers, suggesting a balanced spending behavior among the population. The whiskers extend from about 20 to 80, showing the range of most scores.

b. Box plot for Age Analysis

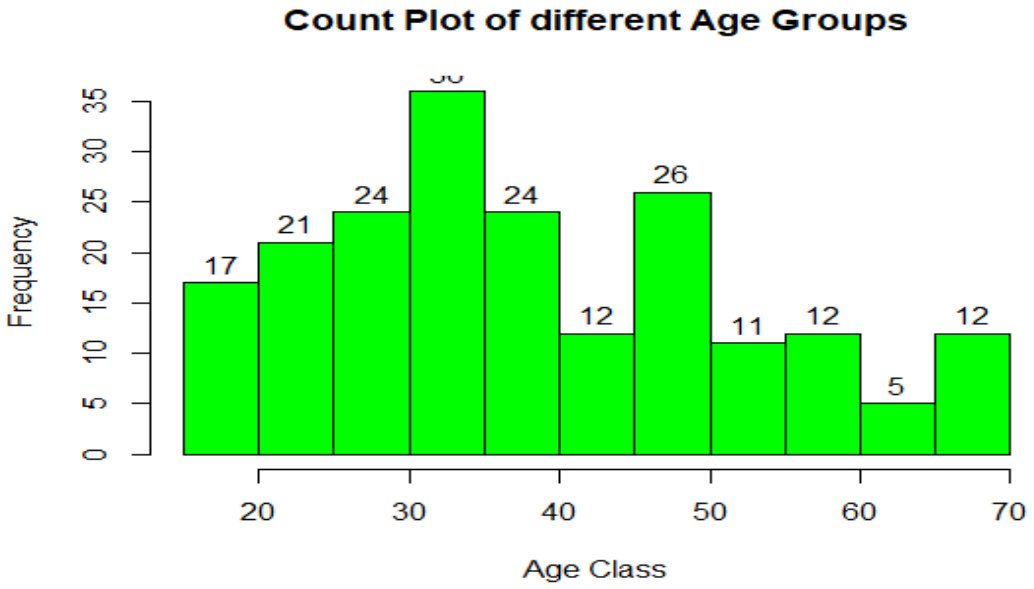


Interpretation:

The boxplot in the image represents the distribution of a dataset related to "Spending Score". Here are the key takeaways from the plot:

- Interquartile Range (IQR): The main box in the plot spans approximately from 40 to 60, indicating that the central 50% of spending scores fall within this range.
- Median: The line inside the box represents the median spending score, which is around 50.
- Outliers: There are no outliers indicated, and the whiskers extend to cover almost all data points.
- Overall Distribution: Most spending scores cluster between approximately 40 and 60, with some variability.

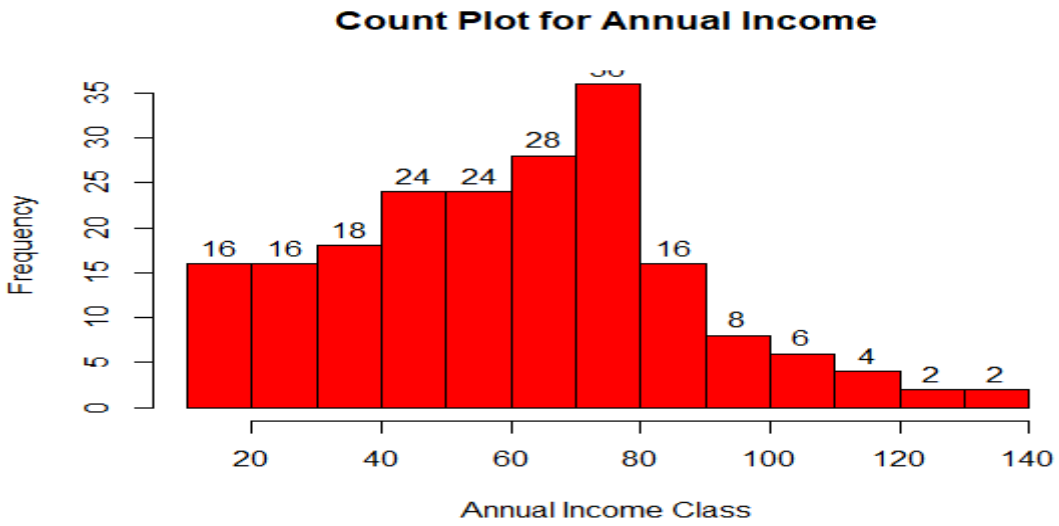
c. Count plot for different Age Groups:



Interpretation:

The graph is a bar chart titled "Count Plot of different Age Groups". It shows the frequency of individuals within specific age classes: 20, 30, 40, 50, 60, and 70. The highest frequency is in the age class of 50 with 26 individuals. The age classes of 30 and 40 also have relatively high frequencies with 24 individuals each. Other age groups have lower frequencies

d. Count plot for Annual Income



Interpretation :

The histogram graph titled "Count Plot for Annual Income" represents the frequency distribution of different annual income classes. Here are the key points:

Annual Income Classes:

- The x-axis represents the "Annual Income Class," ranging from 20 to 140.
- The y-axis represents "Frequency," with values ranging from 0 to approximately 35.

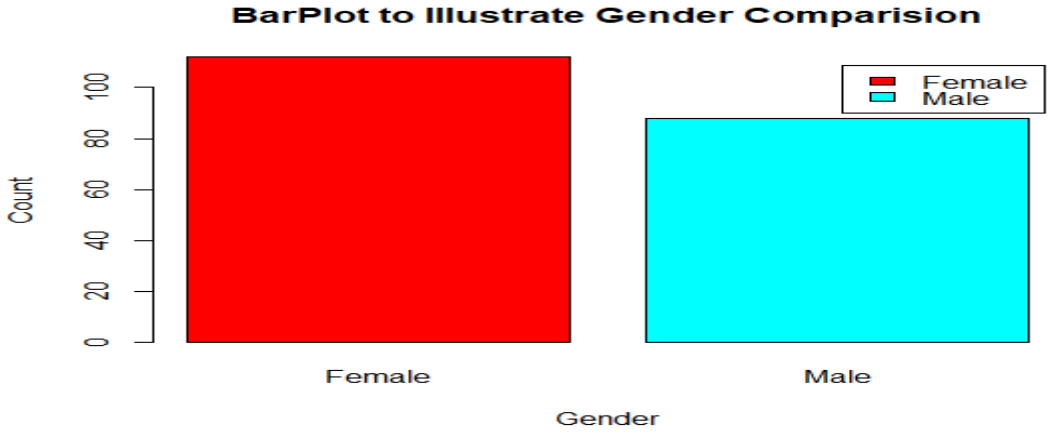
Frequency Peaks:

- There are two prominent peaks in the distribution:
- Around an annual income class of 40, there is a smaller peak with a frequency of approximately 16.
- The more significant peak occurs around an annual income class of 80, with a frequency of approximately 30.

Interpretation:

- The majority of individuals fall into the annual income class around 80, indicating a common income level.
- Fewer people have extremely low or high annual incomes.

e. Gender Comparison Bar Plot

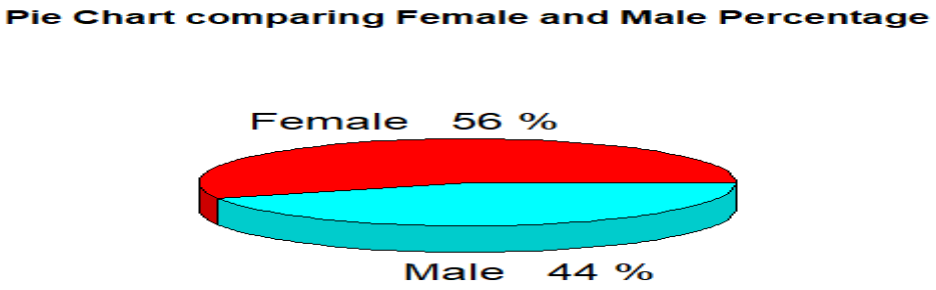


Interpretation:

The bar plot graph titled "Bar Plot to Illustrate Gender Comparison" compares the count of two categories: "Female" and "Male". Here are the key takeaways:

- Female: The red bar represents the female category, with a count close to 100.
- Male: The blue bar represents the male category, with a count of approximately 80.
- The y-axis represents the count, ranging from 0 to 100, while the x-axis indicates the gender. Overall, the female count is higher than the male count in this comparison.

f. Gender Comparison Pie Chart



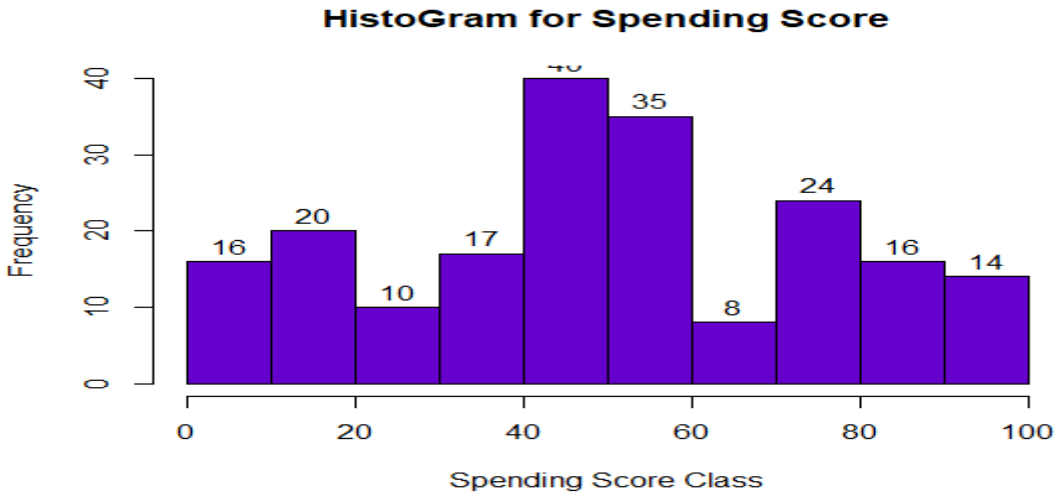
Interpretation :

The pie chart you provided compares the percentage of females to the percentage of males. Here are the key takeaways:

- Female: Represents 56% of the population.
- Male: Constitutes 44% of the population.
-

This chart visualizes a specific dataset or population where females slightly outnumber males.

g. Histogram for spending Score



Interpretation :

The histogram graph titled "HistoGram for Spending Score" displays the distribution of a spending score across different classes. Here are the key points:

Spending Score Classes:

- The x-axis represents the "Spending Score Class," divided into intervals of 20, ranging from 0 to 100.
- The y-axis represents "Frequency," ranging from 0 to approximately 40.
- Frequency Distribution:
- The highest frequency occurs in the 60-80 spending score class, with approximately 35 occurrences. This indicates that most people fall within this spending score range.

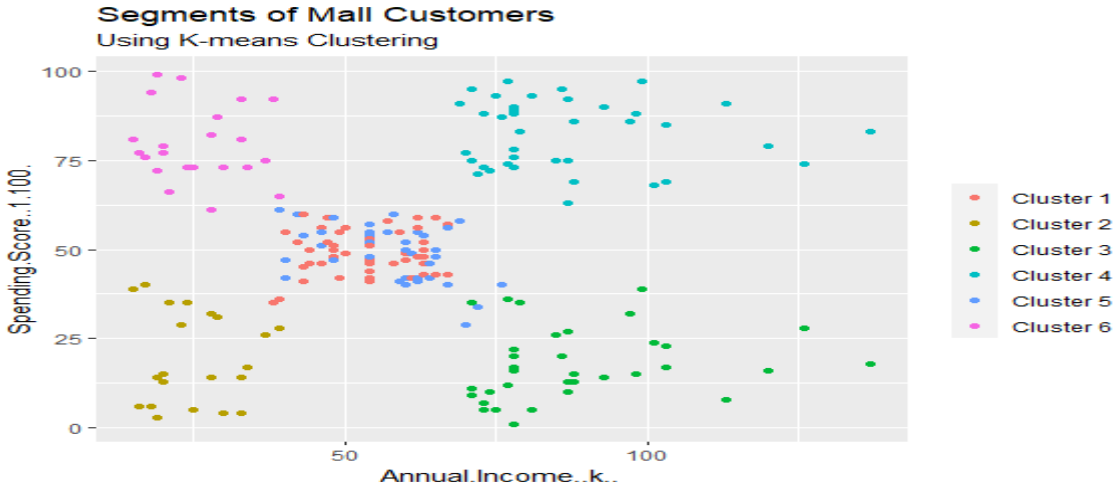
Other notable frequencies:

Class 0–20: ~16 occurrences
Class 20–40: ~10 occurrences
Class 40–60: ~17 occurrences
Class 80–100: ~24 occurrences
There's also a bar beyond the class of >100 (not labeled) with a frequency of ~14.

Interpretation:

- The majority of individuals have a spending score between 60 and 80, suggesting moderate spending behavior.
- Fewer people have extremely low or high spending scores.

h. Segments of mall Customer using K-Means Clustering

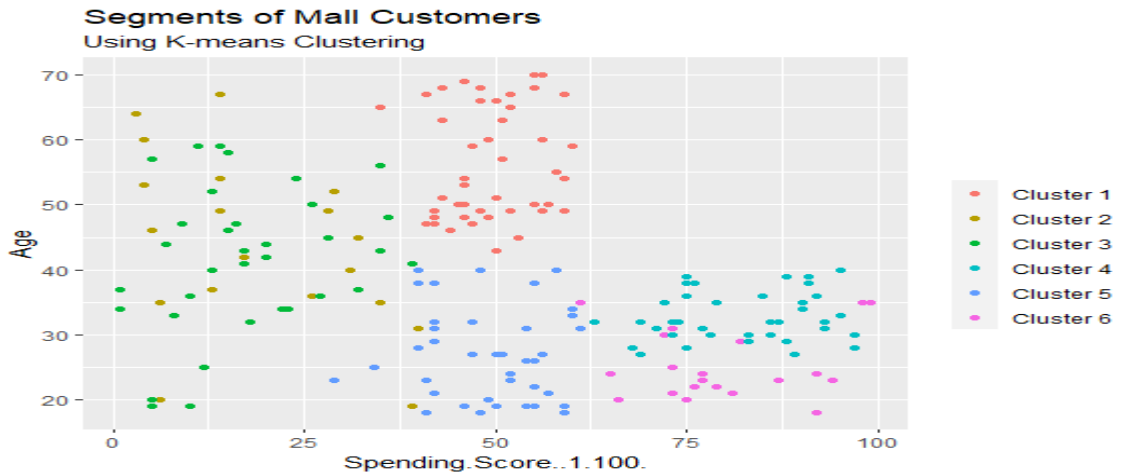


Interpretation:

The scatter plot in the image represents different segments of mall customers using K-means clustering. Here are the key takeaways:

- Annual Income vs. Spending Score: The x-axis represents the annual income (in thousands), and the y-axis represents the spending score (ranging from 0 to 100).
- Distinct Clusters: There are six distinct clusters of customers, each represented by a different color (red, blue, green, cyan, magenta, and yellow).
- Segmentation: These clusters group customers based on their annual income and spending behavior.
- Cluster Characteristics: Further analysis would reveal the characteristics of each cluster, such as high spenders, moderate spenders, etc.

i. Clustering for spending Score



Interpretation:

Cluster Analysis: The graph represents customer segments obtained through K-means clustering. Each point on the scatter plot corresponds to a customer. The two axes are:

X-axis: "Spending Score" (ranging from 0 to 100). This score reflects how much a customer spends.

Y-axis: "Age" (ranging from 0 to 70). This represents the age of the customer.

Color-Coded Clusters:

Red Cluster (Cluster 1): Younger individuals with high spending scores.

Green Cluster (Cluster 2): Middle-aged individuals with moderate spending scores.

Blue Cluster (Cluster 3): Older individuals with low to moderate spending scores.

Cyan Cluster (Cluster 4): Younger individuals with low spending scores.

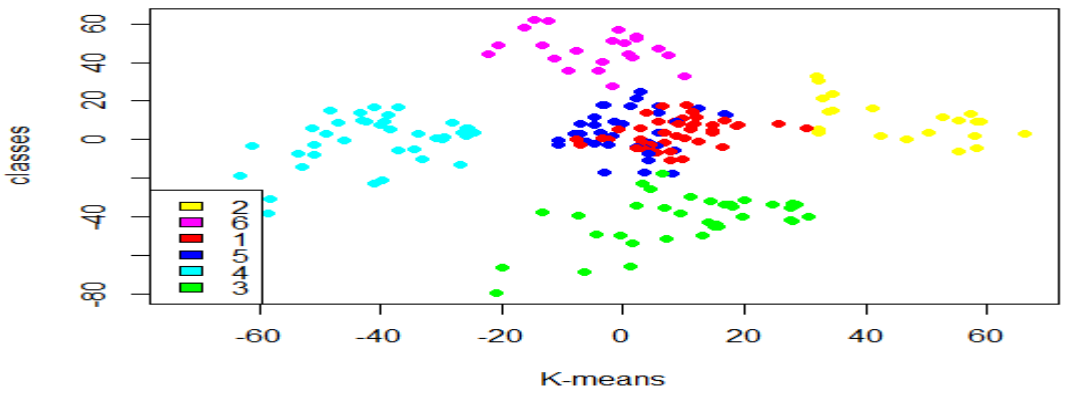
Magenta Cluster (Cluster 5): Older individuals with high spending scores.

Yellow Cluster (Cluster 6): Middle-aged individuals with low spending scores.

Insights:

- Cluster 1 and Cluster 5 represent high spenders, but Cluster 1 is younger, while Cluster 5 is older.
- Cluster 2 consists of moderate spenders across different age groups.
- Cluster 3 includes older customers with conservative spending habits.
- Cluster 4 represents younger individuals who spend less.
- Business Implications: Understanding these segments can help tailor marketing strategies, product offerings, and store layouts to cater to different customer profiles.

j. Clustering for Annual income VS spending score



Interpretation:

K-means Clustering:

- The graph represents a K-means clustering of data points into six different classes, each marked by a distinct color.
- The x-axis is labeled as "K-means," ranging from -60 to 60.
- The y-axis is labeled as "classes," ranging from -80 to 60.
- Each cluster of colored dots represents data points that have been grouped together based on their similarities.

Distinct Clusters:

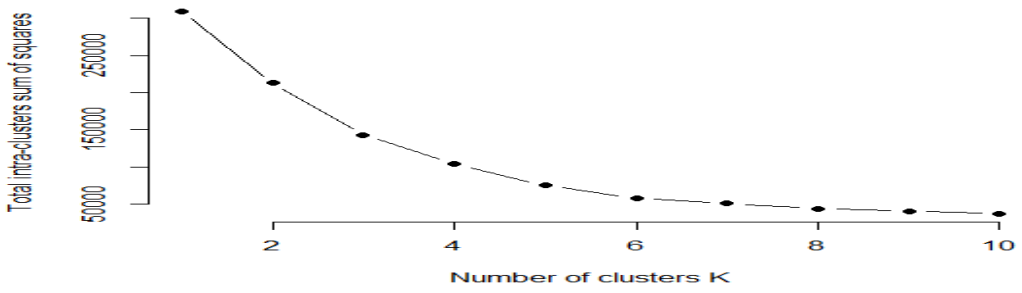
- Six distinct clusters are visible, each represented by colors: blue, green, red, purple, yellow, and cyan.
- There is no overlapping between different colored clusters, indicating distinct groupings.

Legend:

- A color-coded legend on the left side indicates the class numbers (0 to 5) corresponding to each cluster's color.

In summary, this graph shows the results of K-means clustering, where data points have been grouped into distinct classes based on their features. Each cluster represents a different category or similarity pattern. 📊🔍

k. Intra Clustering

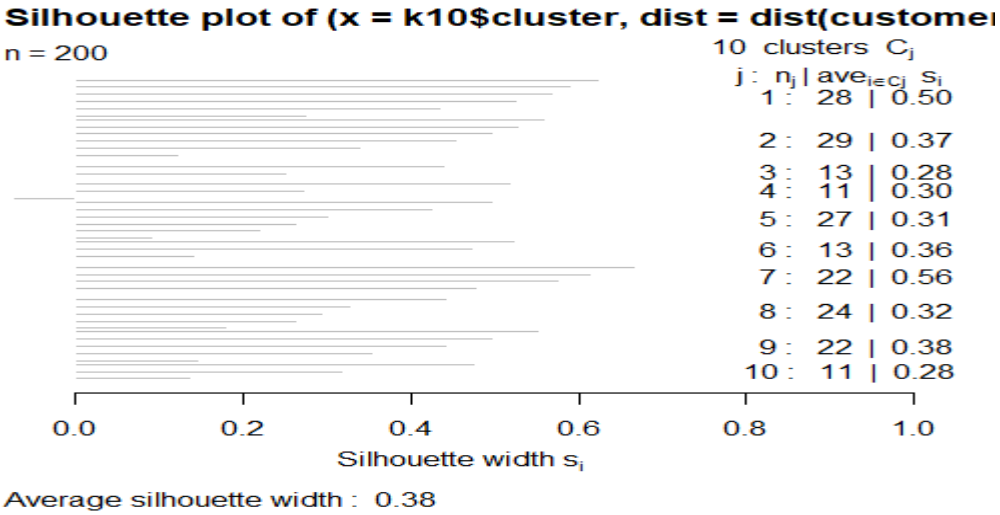


Interpretation:

- This graph represents the Elbow Method, a technique used in clustering analysis to determine the optimal number of clusters (K). Here are the key points:

- X-Axis (Number of Clusters K): The x-axis represents the number of clusters (K) considered for the analysis. It ranges from 2 to 10.
- Y-Axis (Total Intra-Cluster Sum of Squares): The y-axis represents the total intra-cluster sum of squares. As K increases, the total variance within clusters decreases.
- Elbow Point: The "elbow" point occurs where the rate of decrease in total intra-cluster sum of squares sharply changes. This point indicates an optimal number of clusters. In this graph, it appears around K=4.
- In summary, the graph suggests that 4 clusters might be an appropriate choice for this clustering problem. Beyond K=4, the reduction in variance becomes less significant.

l. Silhouette plot_1



Interpretation :

The silhouette plot in the image represents the clustering quality of a data set using the silhouette coefficient. Here are the key points:

Silhouette Plot Overview:

- The plot visualizes how well each data point fits within its assigned cluster.
- Each data point is represented by a horizontal line.
- The x-axis shows the silhouette width ((s_i)), ranging from 0 to 1.
- The y-axis represents the clusters (from 1 to 10).

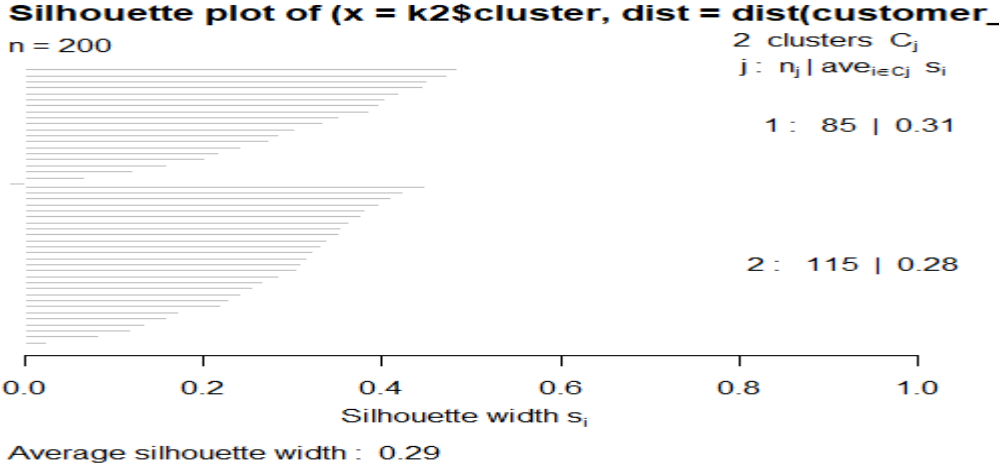
Cluster Information:

- The silhouette width for each cluster is shown on the right side.
- Cluster 7 has the highest average silhouette width (0.56), indicating good separation from other clusters.
- Clusters 3 and 10 have lower average widths (both at 0.28), suggesting they may be less well-defined.

Overall Quality:

- The overall average silhouette width for all data points is 0.38, indicating moderate cluster quality.
- Remember that this is a visual representation, and the silhouette coefficient helps assess the quality of clustering algorithms. Higher values indicate better separation between clusters.

m. Silhouette plot_2



Interpretation:

n. Silhouette Plot Overview:

- The silhouette plot is commonly used in cluster analysis to evaluate the quality of clustering.
- It shows how well each data point is clustered and the separation between clusters.

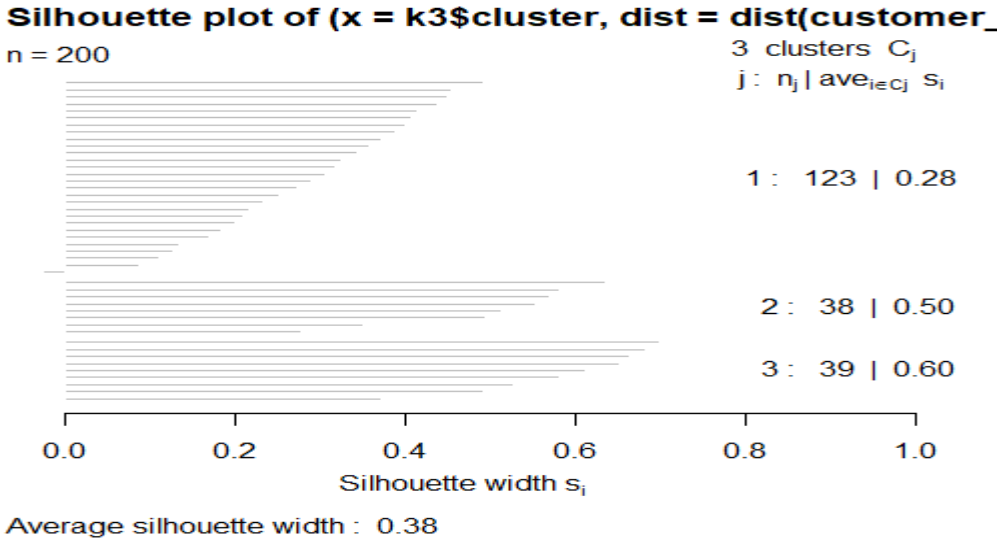
Graph Description:

- The x-axis represents the silhouette width (S_i), ranging from 0 to 1.
- Each line corresponds to an individual data point within the clusters.
- There are two clusters (C_1 and C_2) shown on the graph.
- The average silhouette widths for these clusters are:
- Cluster 1 (C_1): 85 data points with an average silhouette width of 0.31.
- Cluster 2 (C_2): 115 data points with an average silhouette width of 0.28.
- The overall average silhouette width for the entire dataset is 0.29.

Interpretation:

- A silhouette width close to 1 indicates that the data point is well-clustered.
- The moderate silhouette widths suggest reasonable clustering, but there's room for improvement.
- Remember that this is a visual representation, and the silhouette width helps assess the quality of clustering. Researchers can use this information to refine their clustering algorithms or explore different cluster configurations.

o. Silhouette Plot_3



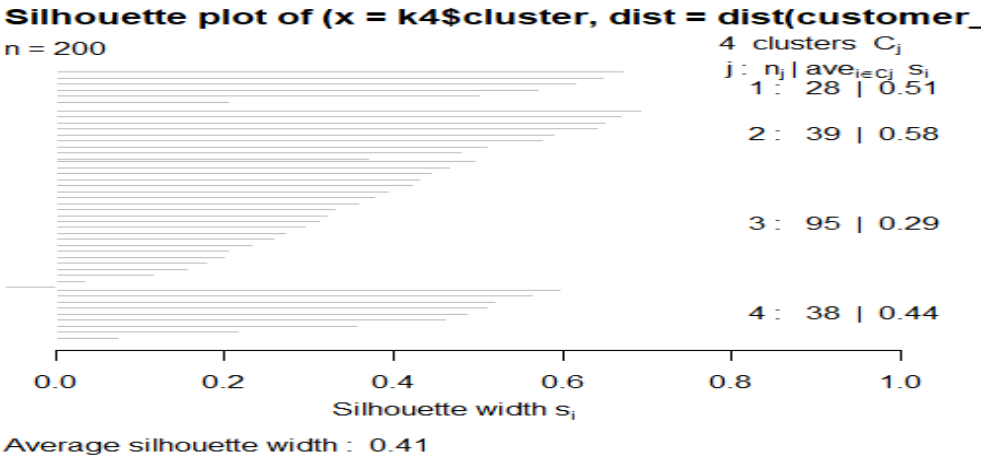
Interpretation:

The image you provided appears to be a silhouette plot representing the results of a k-means clustering algorithm with three clusters. Let's break down the details:

- The title of the plot reads "Silhouette plot of (x = k$3cluster, dist = dist(customer)."
- There are three distinct groups or clusters represented by lines; each line corresponds to one of the 200 data points.
- On the x-axis, there's "Silhouette width s_i" labeled from 0.0 to 1.0, indicating the silhouette score.

- On the right side, it indicates that there are three clusters Cj with respective counts and average silhouette widths:
- Cluster 1 has 123 elements with an average silhouette width of 0.28.
- Cluster 2 has 38 elements with an average width of 0.50.
- Cluster 3 has 39 elements with an average width of 0.60.
- At the bottom, it states "Average silhouette width: 0.38."
- This plot helps visualize the separation distance between the resulting clusters. Each line represents a data point, and its length represents how similar that point is to its own cluster compared to other clusters. Silhouette scores closer to 1 indicate better-defined clusters.

p. Silhouette Plot _4



Interpretation:

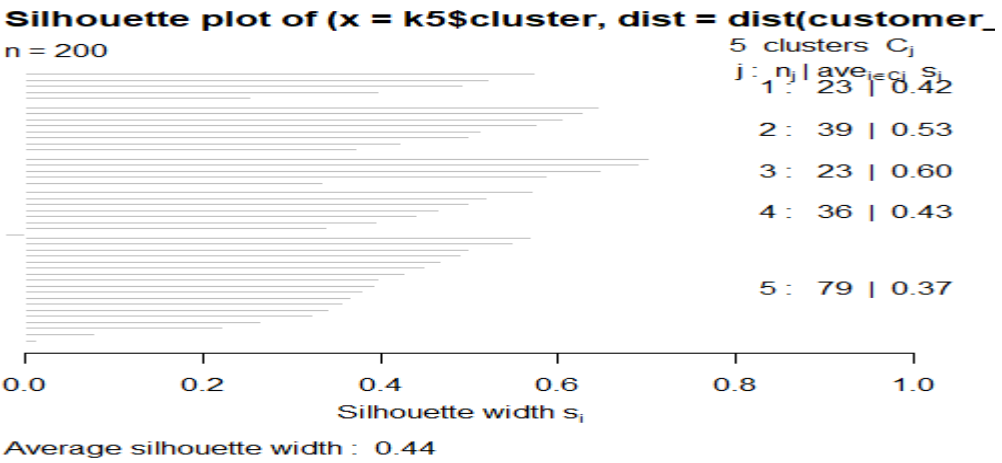
The silhouette plot in the image represents the results of a k-means clustering algorithm with four clusters. Here are the key takeaways:

- Cluster Separation: Each line corresponds to a data point, and its length indicates how close that point is to its assigned cluster compared to other clusters. Longer lines imply better separation.
- Silhouette Width: The x-axis represents the silhouette width (si), ranging from 0 to 1. It measures the closeness of data points to their respective clusters. An average silhouette width of 0.41 suggests moderate cluster quality.
- Cluster Details: The right side lists the clusters (labeled 1 to 4) along with their sizes and average silhouette widths. For instance:

Cluster 1: 28 members, average silhouette width of 0.51.
Cluster 2: 39 members, average silhouette width of 0.58.
Cluster 3: 95 members, average silhouette width of 0.29.
Cluster 4: 38 members, average silhouette width of 0.44.

Remember, this plot helps visualize how well the k-means algorithm grouped data points into distinct clusters.

q. Silhouette plot_5



Interpretation :

The silhouette plot in the image represents the results of a clustering algorithm. Here are the key takeaways:

Silhouette Width:

- The x-axis represents the silhouette width (si), ranging from 0.0 to 1.0.
- Silhouette width measures how similar a data point is to its own cluster compared to other clusters.
- Higher silhouette width indicates better-defined clusters.

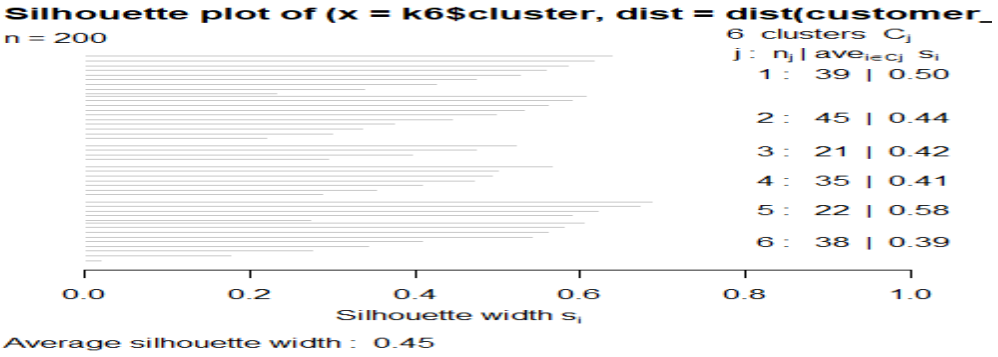
Cluster Information:

- There are five clusters (C1 to C5).
- Each line represents an individual data point.
- The average silhouette widths for each cluster are as follows:
- Cluster 1: 23 points, average width 0.42
- Cluster 2: 39 points, average width 0.53
- Cluster 3: 23 points, average width 0.60
- Cluster 4: 36 points, average width 0.43
- Cluster 5: 79 points, average width 0.37

Overall Quality:

- The average silhouette width across all clusters is 0.44, indicating moderate cluster quality.

r. Silhouette Plot_6



Interpretation:

The silhouette plot in the image represents the results of a clustering algorithm applied to a dataset of 200 elements, divided into 6 clusters. Here are the key takeaways:

Cluster Structure:

- The silhouette width (S_i) for each element in the clusters is plotted.
- The x-axis represents the silhouette width, ranging from 0 to 1.
- Each cluster has varying numbers of data points and different average silhouette widths.
- The overall average silhouette width for all clusters combined is 0.45.

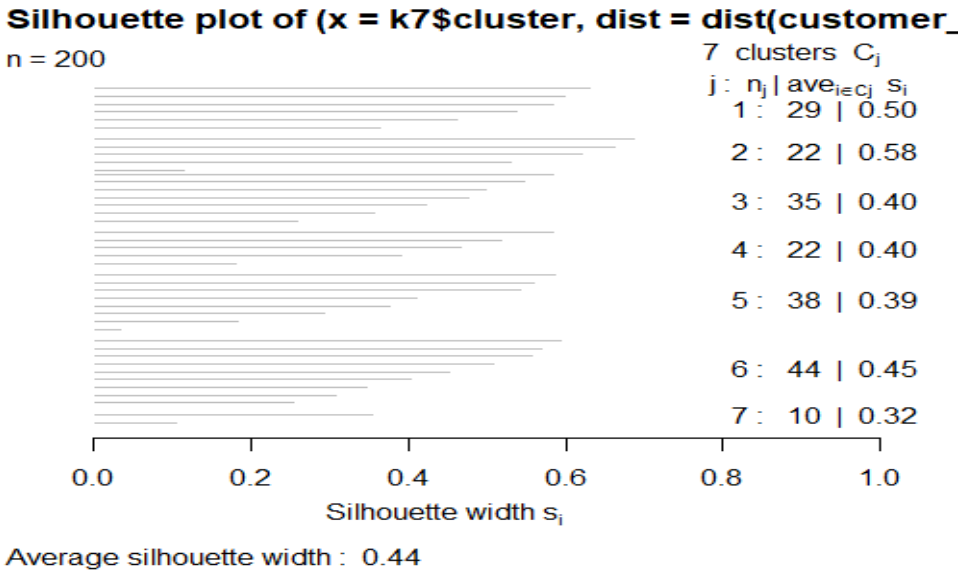
Cluster Details:

- Cluster 1: 39 elements with an average S_i of 0.50.
- Cluster 2: 45 elements with an average S_i of 0.44.
- Cluster 3: 21 elements with an average S_i of 0.42.
- Cluster 4: 55 elements with an average S_i of 0.41.
- Cluster 5: 22 elements with an average S_i of 0.58.
- Cluster 6: 55 elements with an average S_i of 0.39.

Interpretation:

A silhouette width close to 1 indicates well-separated clusters. The average silhouette width of 0.45 suggests a reasonable structure has been found

s. Silhouette plot_7



Interpretation:

The silhouette plot provided depicts the results of a k-means clustering analysis with 7 clusters. Here are the key takeaways:

Silhouette Width (S_i):

- The x-axis represents the silhouette width (S_i), which measures how similar each data point in a cluster is to the points in neighboring clusters.
- Silhouette width values range from -1 to 1.
- A higher S_i indicates better separation between clusters.

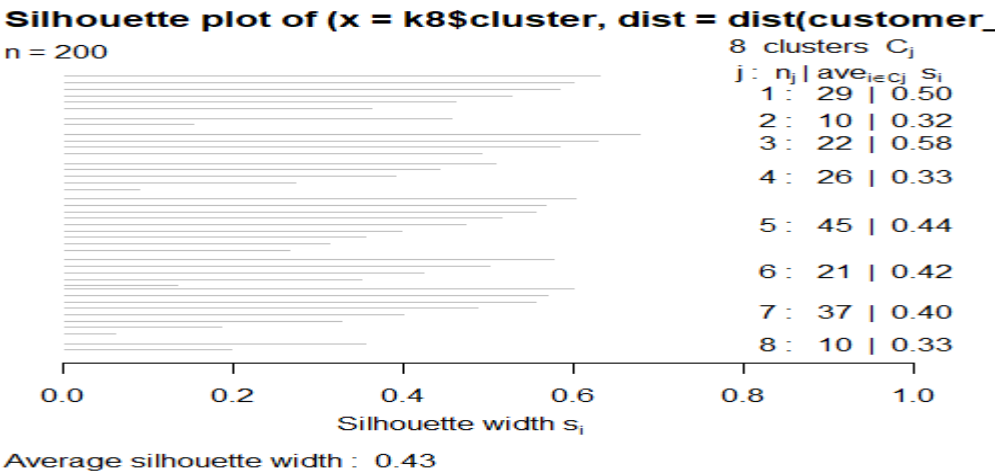
Cluster Information:

- The plot shows individual data points as horizontal lines.
- On the right side, details about the 7 clusters (C_j) are provided:
- Cluster size (n_j) and average silhouette width (aveC_j S_i) for each cluster.
  For example:
- Cluster 1: Size = 29, Average S_i = 0.50
- Cluster 2: Size = 22, Average S_i = 0.58
- …
- Cluster 7: Size = 10, Average S_i = -0.32

Overall Evaluation:

The average silhouette width for all clusters is 0.44, suggesting reasonably well-separated clusters.

t. Silhouette Plot_8



Interpretation:

The silhouette plot in the image represents the results of a clustering algorithm, specifically k-means with 8 clusters. Here are the key takeaways:

Silhouette Plot Overview:

- The plot visualizes the separation distance between the resulting clusters.
- Each line represents a data point, and its length indicates how similar that point is to its own cluster compared to other clusters.
- The average silhouette width of 0.43 suggests moderate cluster quality.

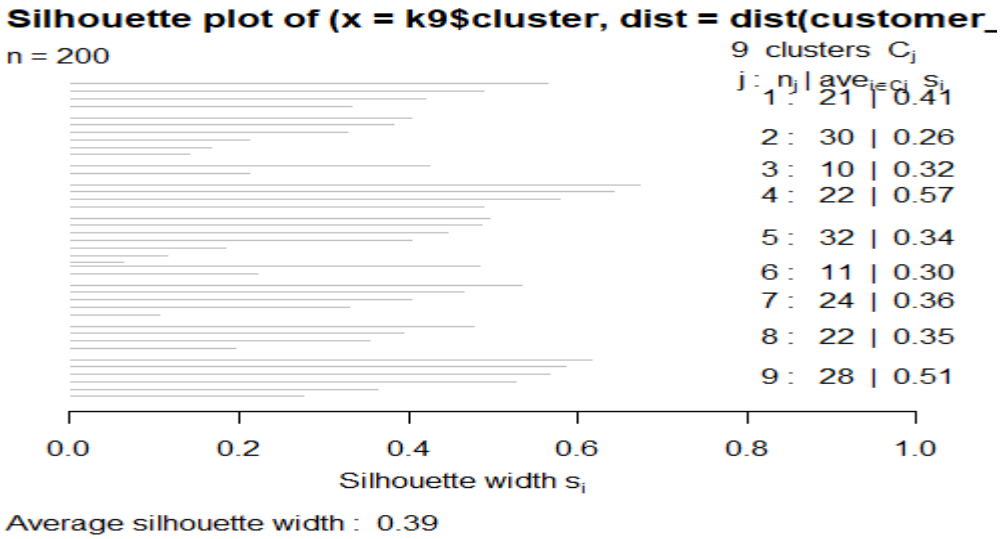
Cluster Details:

- There are 8 distinct clusters represented on the right side.
- Each cluster has a different number of data points and average silhouette widths.
  For example:
- Cluster 1: 29 data points with an average silhouette width of 0.50.
- Cluster 2: 10 data points with an average silhouette width of 0.32.
- …
- Cluster 8: 10 data points with an average silhouette width of 0.33.

Interpretation:

- Clusters with higher silhouette widths (closer to 1) indicate better separation.
- The overall silhouette width of 0.43 suggests a reasonable clustering solution.

u. Silhouette Plot_9

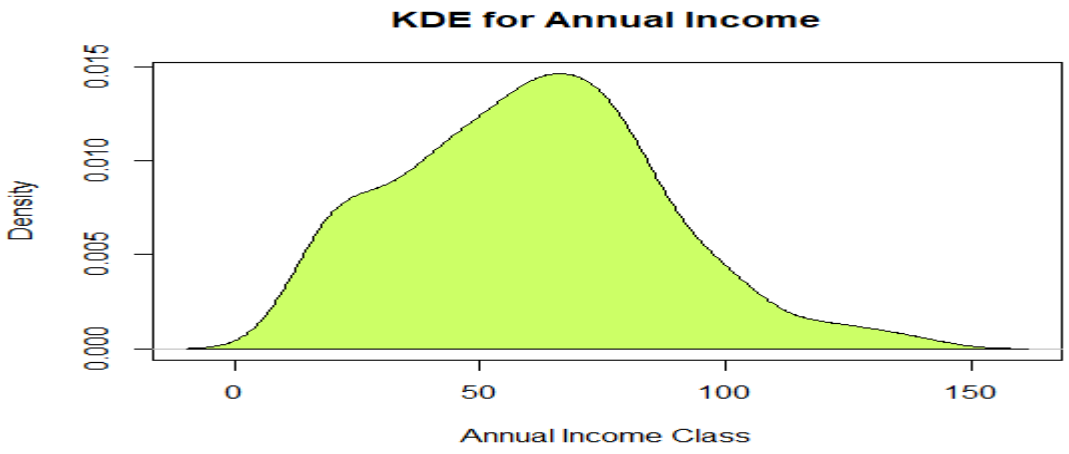


Interpretation:

The silhouette plot in the image represents the results of a clustering algorithm applied to a dataset of 200 elements. Here are the key takeaways:

- **Silhouette Width (S_i):** The horizontal lines represent individual data points, and their length corresponds to their silhouette width. Silhouette width measures how similar an object is to its own cluster compared to other clusters. Higher values indicate better clustering.
- **Clusters (C_j):** There are nine distinct clusters (labeled 1 to 9). Each cluster has a different number of data points (n_j) and an average silhouette width (aveCj S_i).
- **Quality of Clustering:** The average silhouette width across all clusters is 0.39, indicating moderate quality. Clusters with higher silhouette widths are well-separated, while those with lower widths may have overlapping data points.

v. KED for Annual Income

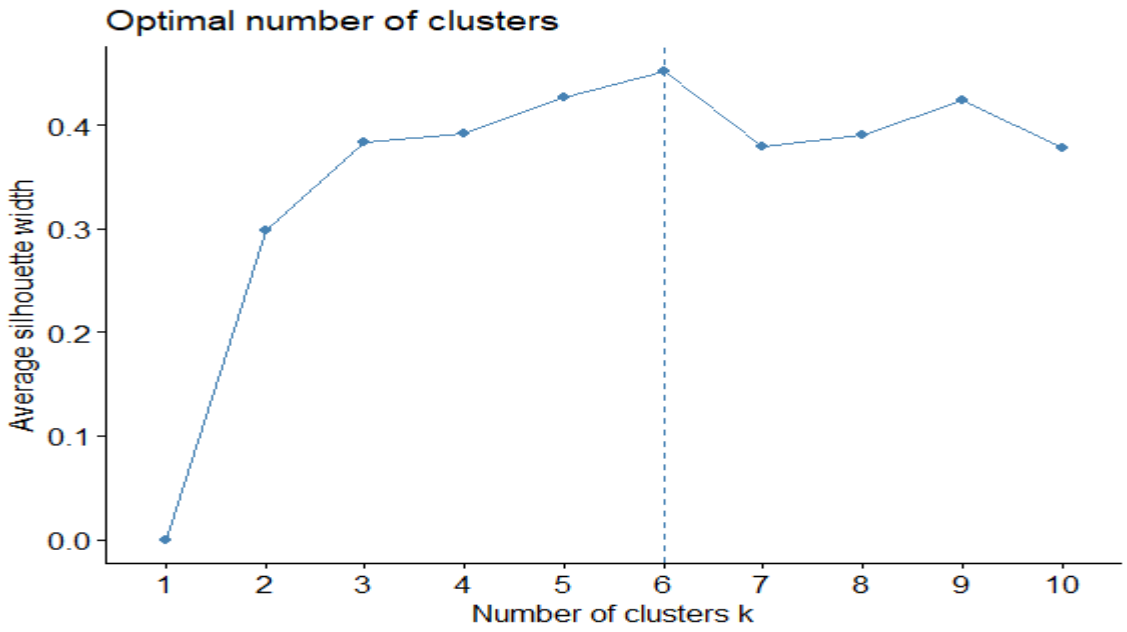


Interpretation:

- The Kernel Density Estimation (KDE) plot in the image represents the probability density of different annual income classes. Here are the key takeaways:

- **Peak Income:** The graph peaks around an annual income class of approximately 50.
- **Normal Distribution:** The KDE is symmetrical and bell-shaped, indicating a normal distribution of income data.
- **Density:** The y-axis represents the density, with the highest density around the peak.
- **Variability:** The KDE shows how income is distributed across different classes, with more people earning around the peak value.

w. Optimal Number of clusters



Interpretation :

The graph in the image represents the average silhouette width as a function of the number of clusters (k) in a clustering algorithm. Here are the key takeaways:

- **Silhouette Width:** The silhouette width measures the quality of clustering, with higher values indicating better-defined clusters.
- **Optimal Clusters:** The graph shows a noticeable peak at k=5, suggesting that dividing the data into five clusters provides the most distinct and well-separated groups.
- **Cluster Separation:** When k=5, the average silhouette width is maximized, indicating that the data points within each cluster are closer to each other than to points in other clusters.

7. RESULTS

a. K-means Clustering Analysis:

The application of the K-means clustering algorithm to the Mall Customers Dataset yielded meaningful results. Through an iterative process, distinct customer segments were identified based on common characteristics such as age, annual income, and spending score. The clusters provide valuable insights into different customer profiles within the supermarket mall.

1.1 Cluster Characteristics:

- **Cluster 1:** Customers with low income but high spending score.
- **Cluster 2:** Customers with high income but low spending score.
- **Cluster 3:** Customers with medium income and medium spending score.
- **Cluster 4:** Customers with high income and high spending score.
- **Cluster 5:** Customers with medium income and medium spending score.
- **Cluster 6:** Customers with low income and low spending score.

1.2 Cluster Analysis:

- Cluster 4 and Cluster 2 represent high-income customers with varying spending behaviours.
- Cluster 3 reflects customers with a balanced income and spending profile.
- Cluster 1 and Cluster 6 highlight specific segments with distinctive income and spending pattern

Optimal Number

1. f Clusters:

The Elbow Method, Silhouette Width Method, and Gap Statistic Method were employed to determine the optimal number of clusters.

8. CONCLUSION

The results obtained from the K-means clustering analysis provide valuable insights into the diverse customer segments within the supermarket mall. Understanding customer behaviour through segmentation allows for targeted marketing strategies, ultimately optimizing sales approaches.

9. FUTURE SCOPE

- **User-Friendly Web Interface:**
  Develop a user-friendly web interface using the Shiny Package from R Studio. This interface would enhance the visualization of customer segments and facilitate interactive exploration.
- **Real-Time Data Integration:**
  Explore methods to integrate real-time data for dynamic customer segmentation. This would ensure that the analysis stays relevant to evolving customer trends.
- **Machine Learning Model Enhancement:**
  Consider the integration of additional machine learning models to refine and enhance the accuracy of customer segmentation. Experiment with algorithms beyond K-means clustering.
- **Feedback Mechanism:**
  Implement a feedback mechanism to continuously evaluate the effectiveness of the segmentation strategy. This could involve customer surveys or monitoring changes in purchasing behaviour.