Customer Segmentation Using K-means algorithm

Abstract

The management of customer service can be considered as a key to achieving revenue growth and profitability in today's fast-moving world of marketing from product-oriented to customer-oriented. Customer behavior knowledge can assist marketing managers in reevaluating existing customer tactics and planning to improve and increase the use of the most effective strategies. B2B or business customers are more complicated, have a more difficult buying procedure, and have a higher sales value. Business marketers, on the other hand, want to work with fewer but larger customers than final consumer marketers. Because a business transaction necessitates more decision-making and professional buying effort than a consumer purchase, maintaining a productive relationship with business customers is critical. Most customer segmentation methods based on customer value fail to account for the factor of time and the trend of value changes. Because there are so many potential customers who are unsure of what to buy and what not to buy, today's business is based on new concepts. The businesses themselves are unable to diagnose their target potential customers. This is where machine learning comes in, various algorithms are used to detect hidden patterns in the data in order to make better decisions.

Customer segmentation is the practice of dividing a customer base into many groups of people who are similar in various aspects significant to marketing, such as gender, age, interests, and other spending habits. Companies that use customer segmentation believe that each client has unique needs that require a tailored marketing strategy to satisfy. Companies want to obtain a better understanding of the customers they're after. As a result, their goal must be explicit, and it must be adjusted to meet the needs of each and every individual customer. Furthermore, by analyzing the data acquired, businesses can gain a better grasp of client preferences as well as the needs for identifying profitable segments. This allows them to more effectively strategize their marketing strategies while reducing the chance of their investment being jeopardized. Customer segmentation is a process that is depending on a number of factors. Data on demographics, geography, economic position, and behavioral tendencies are all important factors in establishing the company's approach to distinct sectors.

The customer segmentation uses the clustering technique to determine which consumer segment to target. The clustering algorithm we have employed in this project is the K-means algorithm, which is a partitioning algorithm for segmenting clients based on comparable criteria. The elbow approach is used to determine the best clusters.

Literature review

Because of the intense rivalry in the business sector, businesses have had to improve their profitability and business throughout time by satisfying client requests and attracting new customers based on their wants. Client identification and meeting individual customer wants is a difficult task. This is due to the fact that clients differ in terms of their wants, desires, preferences, and so on. Customer segmentation, as opposed to a "one-size-fits-all" strategy, separates customers into groups with similar features or behavioral characteristics. Customer segmentation, according to, is a strategy for splitting the market into homogeneous groups. The data utilised in the customer segmentation technique, which divides customers into groups, is

based on a variety of characteristics including demographics, regional circumstances, economic situations, and behavioral tendencies.

The customer segmentation strategy enables a company to make better use of its marketing spending and obtain a competitive advantage. Displaying a superior understanding of the customer's requirements, it also aids a company in improving marketing efficiency, budgeting for marketing, recognizing new market prospects, developing a stronger brand strategy, and measuring client retention.

[1] Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry

In this paper Prasanta Bandyopadhyay proposed two different clustering models to segment 700032 customers by considering their RFM values. They detected that the current customer segmentation which built by just considering customers' expense is not sufficient. Companies need to understand the customers' data better in all aspects. Detecting similarities and differences among customers, predicting their behaviours, proposing better options and opportunities to customers became very important for customer-company engagement. Segmenting the customers according to their data became vital in this context. RFM (recency, frequency and monetary) values have been used for many years to identify which customers valuable for the company, which customers need promotional activities, etc. Data-mining tools and techniques widely have been used by organizations and individuals to analysis their stored data. Clustering, which one of the tasks of data mining has been used to group people, objects, etc.

[2] Mall customer segmentation using clustering algorithms:

M.A. Ishantha USE/2017/OCT/0045 done research on data set that is about behaviour of the customers having credit card using many unsupervised algorithms. The dataset which contains 8950 transactions or information about account that belong to customers. And also, she has found how many clusters can distinguish the customers according to their transactions or behaviours". The methodology they have followed is K-Means clustering, Minibatch KMeans Clustering Algorithm, Hierarchical Clustering Segmentation and Elbow Method.

[3] Approaches to Clustering in Customer Segmentation: Techniques and approaches

Shreya Tripathi, Aditya Bhardwaj, Poovammal have gone through the various approaches and techniques to clustering in the segmentation process. They have explained what is the customer relationship management, necessities and importance of a customer segmentation in various industries. They have found which is giving the max optimal customers among the all-clustering algorithms like K-Means Clustering, Elbow method and Hierarchical Clustering-. Agglomerative, divisive with optimization. And also, they have done
Visualization of the formation of clusters in the studied dataset with the help of a dendrogram

[4] Comparisons between data clustering algorithms

Osama Abu Abbas have done the mathematical implementation on clustering algorithms with a sample dataset. And also, he explored how algorithms are implemented. He chosen four different clustering algorithms to investigate, study, and compare them. The algorithms he has chosen are: K-means, Self-Organization (SOM), Hierarchal clustering algorithms and Expectation Maximization (EM) clustering algorithm. He listed the reason why he has chosen

the particular algorithms to compare, study etc. He has done all the work to find the optimal clusters from each algorithm finally he explained how algorithms are compared. He said this paper intended to compare between some data clustering algorithms. Through his search he was unable to find any study attempts to compare between the four clustering algorithms under investigation.

[5] Concept Decompositions for Large Sparse Text Data Using Clustering:

I. Dhillon and D. Modha said that It is of tremendous practical relevance to apply machine learning and statistical algorithms such as clustering, classification, principal component analysis, and discriminant analysis to text data sets. In this paper they mainly focus on clustering of text data sets. The structure of the clusters created by the spherical k-means algorithm when applied to text data sets is the first focus. With the goal of acquiring unique insights into the distribution of sparse text data in high-dimensional environments. Such structural insights are a necessary first step toward their second goal, which is to investigate the tight linkages between clustering with the spherical k-means algorithm and the problem of matrix approximation for word-by-document matrices. They've also looked into massive document collection vector space models. These models are extremely high-dimensional and sparse, posing computational and statistical issues not seen in low-dimensional dense data. The spherical k-means algorithm looks for clusters with strong coherence. They discovered that average cluster coherence is poor, implying that each thought vector is surrounded by a big gap in high-dimensional space. Furthermore, they discovered that at various resolutions, the average intra- and inter-cluster architectures were identical. The only significant distinction is the gradual separation of intra-cluster and inter-cluster structure.

[6] Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms

In this paper A. Ansari and A. Riasi, to cluster the customers of the steel sector, they have combined fuzzy c-means clustering and genetic algorithms. The LRFM (length, recency, frequency, monetary value) model variables were used to separate the customers into two groups. Data from 120 consumers was collected and standardised in order to do the clustering. The information comprised four separate variables: the length of the relationship, the recentness of the trade, the frequency of the trade, and the monetary worth of the trade. GAFuzzy Clustering software was used to accomplish the fuzzy clustering. Clustering is done using a combination of fuzzy c-means clustering and a genetic algorithm in this software. Finally, the customers were divided into two clusters. To compare the efficiency of combined algorithms (Fuzzy c-means and genetic algorithms) means square error (MSE) and run time error were used. When compared to the average values of these parameters for all consumers, customers in the first cluster had a longer relationship, more recent trade, and more frequency of trade, but a lower monetary value.

[7] Identifying patients in target customer segments using a two-stage clusteringclassification approach: A hospital-based assessment:

Identifying patients in a Target Customer Segment (TCS) is critical for determining demand for health care services and allocating resources properly. The goal of this research is to develop a two-stage clustering-classification model by combining the RFM attribute and the K-means algorithm to cluster TCS patients, and then combining the global discretization method and rough set theory to classify hospitalized departments and optimize health care services. To

evaluate the proposed model's performance, a dataset from a representative hospital (dubbed Hospital-A) was collected from a database from an empirical study in Taiwan that included 183,947 samples classified by 44 variables in 2008. The suggested model was compared to three techniques: Decision Tree, Naive Bayes, and Multilayer Perceptron, with the empirical results indicating that it has a high likelihood of being accurate. The knowledge-based rules that are developed give important information for maximizing resource use.

[8] A Two-Phase Clustering Method for Intelligent Customer Segmentation

M. Namvar, M. Gholamian and S. KhakAbi states that many studies have looked at the use of data mining technology in customer segmentation and found it to be successful. However, in the majority of cases, it is done utilising customer data from a unique perspective rather than a systematic process that considers all stages of CRM. Using data mining technologies, they have developed a new customer segmentation strategy based on RFM, demographic, and LTV data. There are two stages to the new consumer segmentation method. Customers are first clustered into distinct segments based on their RFM using K-means clustering. Second, each cluster is partitioned into new clusters using demographic data. Finally, a customer profile is constructed using LTV. They have applied this approach to a dataset from an Iranian bank, yielding some valuable management recommendations and measures.

The method they have followed was based on a two-phase clustering model using the kmeans algorithm. Finally existing customers were split into nine groups based on their shared transactional behavior and features when the strategy was used to our case study (in the banking business). Marketers could evaluate the profiles of clients in each category to develop strategies for each group.

[9] Application of data mining techniques in customer relationship management: A literature review and classification:

E. Ngai, L. Xiu and D. Chau remarked despite the relevance of data mining techniques in customer relationship management (CRM), a complete literature evaluation and classification scheme for them are lacking. This was the first scholarly review of the application of data mining techniques to CRM that has been identified. It offers an academic database of literature from 2000 to 2006, comprising 24 journals, as well as a classification scheme for the articles. A total of 900 papers were found and assessed for their direct relation to CRM data mining methodologies. Following that, 87 articles were chosen, examined, and categorised. Each of the 87 articles was divided into four categories: customer identification, customer attraction, customer retention, and customer development, as well as seven data mining tasks (Association, Classification, Clustering, Forecasting, Regression, Sequence Discovery and Visualization). Based on the main subject of the papers, they were further divided into nine sub-categories of CRM elements using various data mining approaches.

[10] An efficient k-means clustering algorithm: analysis and implementation:

He offered a set of n data points in d-dimensional space R/sup d/ and an integer k in k-means clustering, and the aim was to find a set of k centers in Rd to minimize the mean squared distance between each data point and its nearest center. Lloyd's (1982) technique is a prominent k-means clustering heuristic. He described the filtering algorithm, a simple and efficient implementation of Lloyd's k-means clustering technique. This algorithm is simple to build, as it only uses a kd-tree as a primary data structure. These models are extremely highdimensional

and sparse, posing computational and statistical issues not seen in lowdimensional dense data. The spherical k-means algorithm looks for clusters with strong coherence. They discovered that average cluster coherence is poor, implying that each thought vector is surrounded by a big gap in high-dimensional space. Furthermore, they discovered that at various resolutions, the average intra- and inter-cluster architectures were identical. He established the filtering algorithm's practical efficiency in two methods. First, he gave a datasensitive study of the method's running time, which indicates that as the spacing between clusters rises, the process runs faster. Second, he provided a variety of empirical analyses based on both synthetically created data and genuine application data sets.

Proposed Work

We proposed that we will build a cluster model using K-means clustering algorithm, which is an unsupervised technique, to cluster a data set regarding customer behaviour at mall. Our major goal is to determine the number of clusters in which we can distinguish customers based on their spending ratings from the mall dataset.

- For future operations/marketing projects, to group clients based on shared purchase behaviour.
- To blend the greatest mathematical, visual, programming, and business methods into a meaningful analysis that can be comprehended across a wide range of settings and disciplines.
- The purpose of customer segmentation is to determine how to relate to customers in each segment in order to optimise each customer's value to the company.

Methodology K-means Clustering: -

The first step in applying the k-means clustering algorithm is to specify the number of clusters (k) that we want in the final result. The algorithm begins by randomly selecting k objects from the dataset to act as the cluster's initial centres. The cluster means, also known as centroids, are the selected objects. The closest centroid is then assigned to the remaining items. The object's centroid is determined by the Euclidean Distance between it and the cluster mean. This is referred to as "cluster assignment." After the assignment is finished, the algorithm calculates a new mean value for each cluster in the data. The observations are reviewed to see if they are closer to a different cluster once the centres have been recalculated. The objects are reassigned using the updated cluster mean. This is repeated multiple times until the cluster assignments are finalised. The clusters found in the current iteration are identical to those found in the previous iteration.

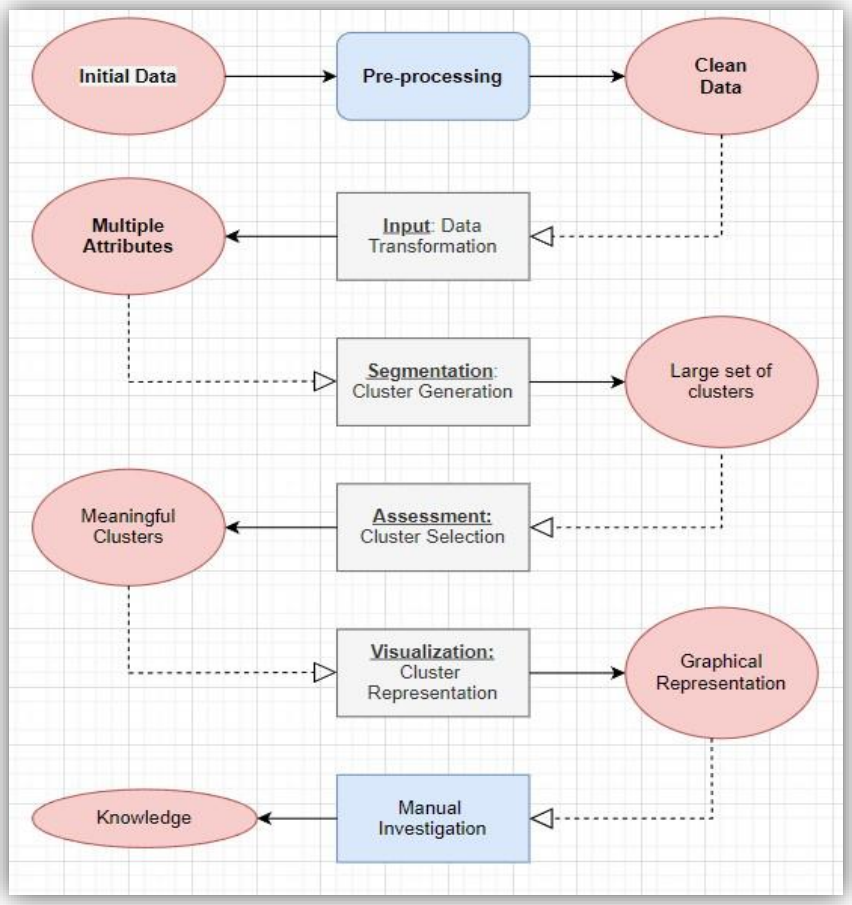


Figure 1: Architecture

As the diagram suggest the proposed system will involve a beginning with extraction of data. Followed by the pre-processing of the data, here we used the data set of Mall Customers. Cleaning the data and segmentation has high level of importance for more appropriate results. Later on, we enter the stage of clustering and analysis wherein we will use ggplot2 library in R studio for reading the data in related variables and plotting trips across different parameters such age, gender, Annual Income etc. Final stage consists of Analyzing the clustered data with proper graphs interpretation. In this project we have used K-means clustering algorithm to segment the mall customers data.

We must indicate the number of clusters to use while working with clusters. We'd like to make use of the most clusters possible. There are three popular ways for determining the best clusters —

Elbow method

The basic purpose of cluster partitioning algorithms such as k-means is to define clusters with the least amount of intra-cluster variation.

k=1...k, minimize (sum W(Ck))

W(Ck) denotes intra-cluster variation, and Ck denotes the kth cluster. The compactness of the clustering boundary can be assessed by measuring the total intra-cluster variation. After that, we can define the best clusters as follows:

To begin, we calculate the clustering procedure for a variety of k values. This can be accomplished by varying the number of clusters in k from one to ten. The total intra-cluster sum

of squares is then calculated (iss). Then we plot iss dependent on how many k clusters there are. This graph depicts the required number of clusters in our model. The optimum number of clusters is indicated by the placement of a bend or a knee in the plot. **Silhouette method**
We have assessed the quality of our clustering operation using the average silhouette method. We have used this to see how well the data object fits into the cluster. We can tell if we have successful clustering if we have a large average silhouette width. For varying k values, the average silhouette method derives the mean of silhouette observations. The average silhouette over significant values for k clusters can be maximised with the ideal number of k clusters. The average silhouette width can be calculated using the kmean function and the silhouette function from the cluster package.
 Gap statistic method
The Gap Statistic Method was published in 2001 by Stanford University scholars R. Tibshirani, G. Walther, and T. Hastie. This method can be applied to any clustering method, such as K-means, hierarchical clustering, and so on. The gap statistic can be used to compare the total intra-cluster variation for various values of k with their predicted values under the null reference data distribution. The sample dataset can be generated using Monte Carlo simulations. We can determine the range between min(xi) and max(xj) for each variable in the dataset, via which we may output values uniformly from interval lower bound to upper bound. We have used the clusGap function to compute the gap statistics method, which returns the gap statistic as well as the standard error for a given output.

In this project, we tried to find the optimal clusters using above mentioned three methods with key factors like age, Gender, Annual Income etc. from a mall dataset which we have taken from the Kaggle. It helps Mall to enhance their business by finding the behavioral patterns among the different customers. And also, it will help them to improve their selling strategies by understating the customer buying strategies.

General View of Mall Dataset
The data set has 2000 customers information which can be described the customer behaviour.
Features
CUSTOMER ID: It is as identification of the customer. Each customer will be having a unique id to differentiate the one customer from among
GENDER: Gender of each customer (The purchase patterns can be different from male and female)
AGE: Age of the customer (Certain age group of customers can have the similar purchasing behavioral patterns)
ANNUAL INCOME (k$): Annual Income of the customers (The purchase patterns also depend upon the income of the customers)
SPENDING SCORE (1-100): The mall gave the spending score based on the customer purchase behavior.
Implementation

- R Studio: In this project, we have used R to visualize the data, plot some graphical interpretation, as well as for analyzing the data and cluster the data. RStudio is an integrated development environment (IDE) for R. It includes a console, syntax, highlighting editor that supports direct code execution, as well as tools for plotting, history, and debugging and workspace management.

- Anaconda Jupyter Notebook: We used Jupyter notebook to build models in python and cluster the customers. Jupyter Notebook allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

Implemented Modules

- Data Preparation
- Data Visualization
- Data Analysis
- Feature Engineering
    - Filling NA, Drop useless column
- Algorithms for Segmentation (Clustering)
    - K-means clustering Algorithm
- Methods to optimize the clusters
    - Elbow Method
    - Silhouette method
    - Gap Static Method

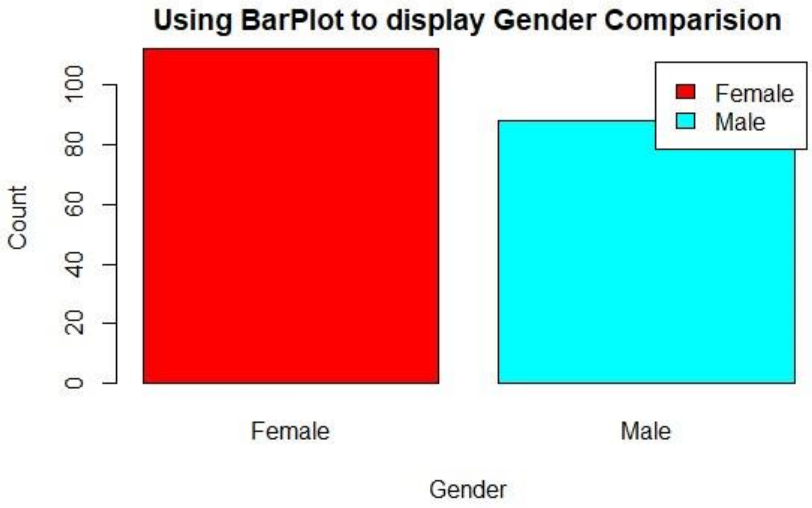
Reason to use Unsupervised Learning Algorithms.

Unsupervised Learning, unlike Supervised Learning, has no target variable and just independent variables. The information is unlabeled. Unsupervised learning aims to model the data's underlying structure or distribution in order to gain a better understanding of it. For segmentation, we're going to look at a dataset of Mall customers. There is no feature for customers' labels. That is to say, we do not have any information regarding the qualities of our customers. We're going to try clustering clients using machine learning methods to find similarities.
In the new marketing disciplines, customer segmentation plays a crucial role for businesses. Because of the costs, businesses must reach the proper target audiences with the right tactics.

Results and Discussion

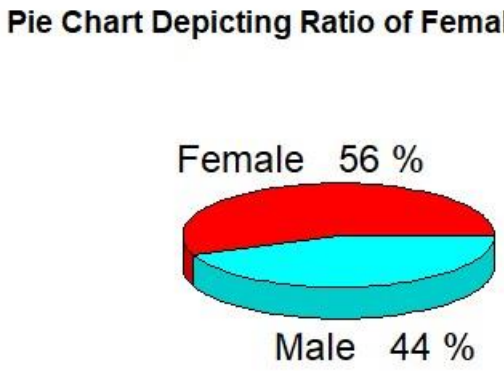
Customer Gender Visualization

- A barplot showing the gender distribution across our customer_data dataset.



Inference: From the above barplot, we observe that the number of females is higher than the males.

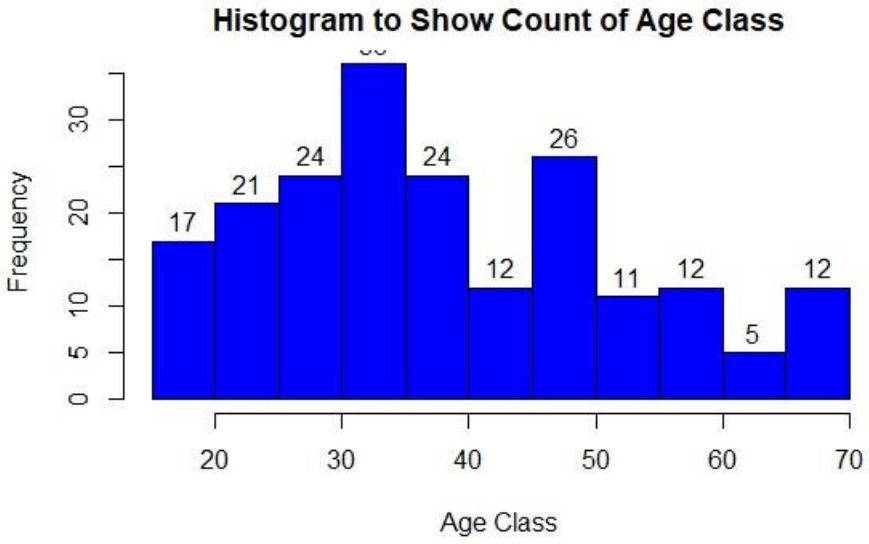
- A pie chart to observe the ratio of male and female distribution.

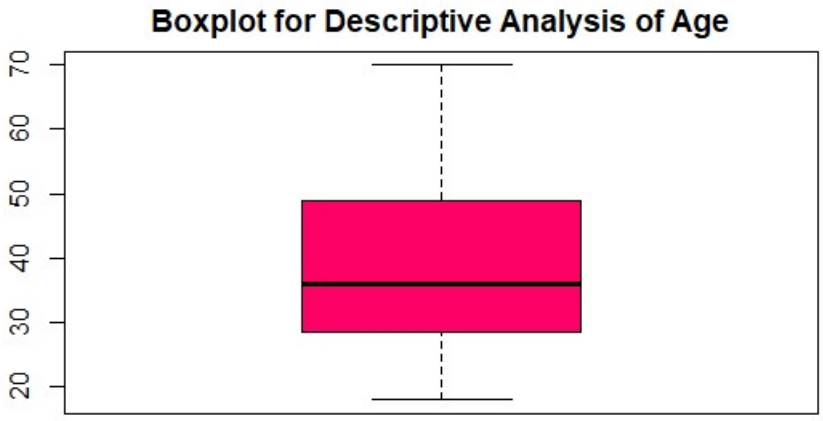


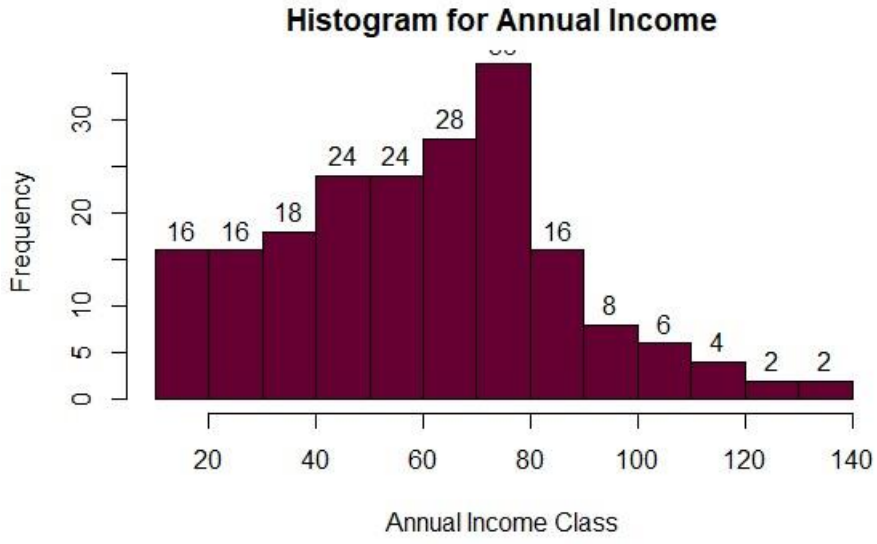
Inference: From the above graph, we conclude that the percentage of females is **56%**, whereas the percentage of male in the customer dataset is **44%**.

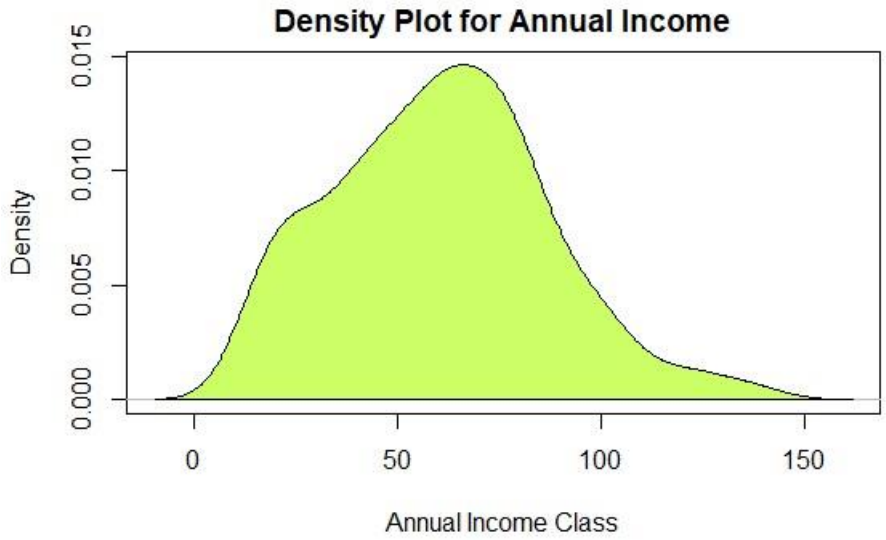
Visualization of Age Distribution

A histogram to view the distribution to plot the frequency of customer ages.


Histogram to Show Count of Age Class
Frequency
0
5
10
20
30
17
21
24
24
12
26
11
12
5
12
20
30
40
50
60
70
Age Class


Boxplot for Descriptive Analysis of Age
20
30
40
50
60
70

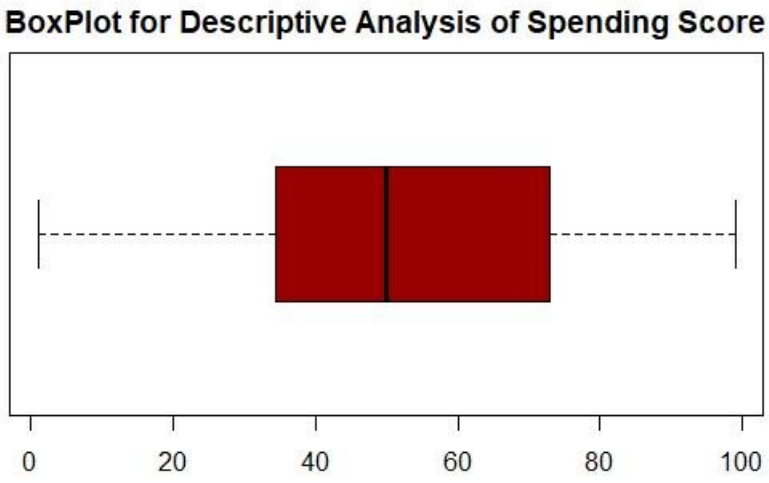


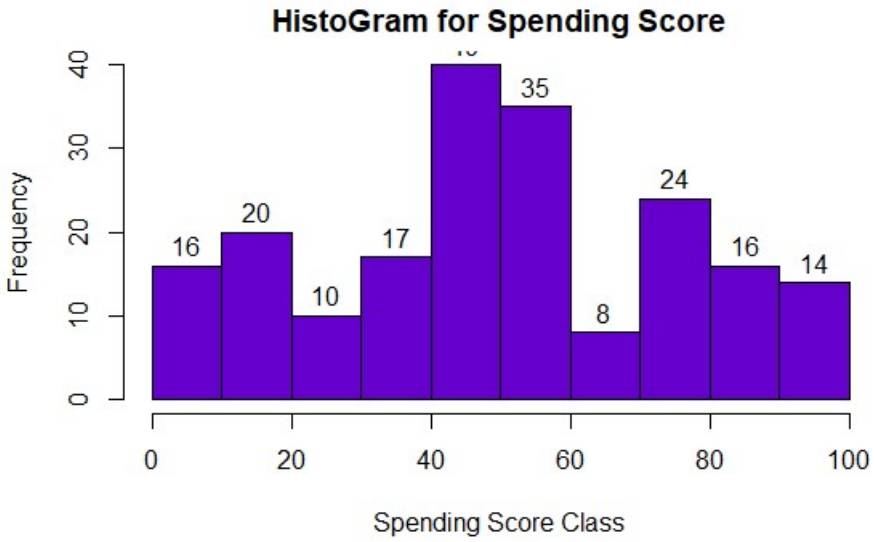


Inference: From the above two visualizations, we conclude that the maximum customer ages are between 30 and 35. The minimum age of customers is 18, whereas, the maximum age is 70.

Analysis of the Annual Income of the Customers

A histogram and density plot to analyse the annual income of the customers.



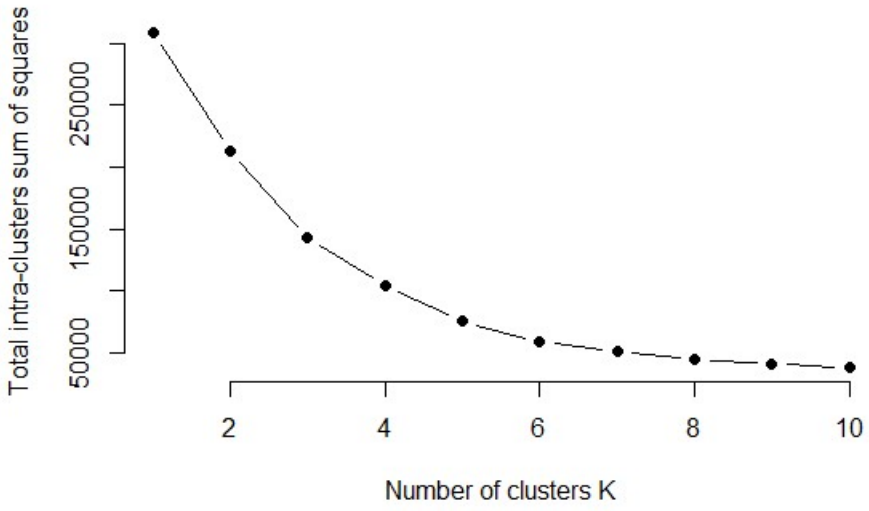


Inference: From the above descriptive analysis, we conclude that the minimum annual income of the customers is 15 and the maximum income is 137. People earning an average income of 70 have the highest frequency count in our histogram distribution. The average salary of all the customers is 60.56. In the Kernel Density Plot that we displayed above, we observe that the annual income has a normal distribution.

Analyzing Spending Score of the Customers

Inference: The minimum spending score is 1, maximum is 99 and the average is 50.20. We can see Descriptive Analysis of Spending Score is that Min is 1, Max is 99 and avg. is 50.20. From the histogram, we conclude that customers between class 40 and 50 have the highest spending score among all the classes.

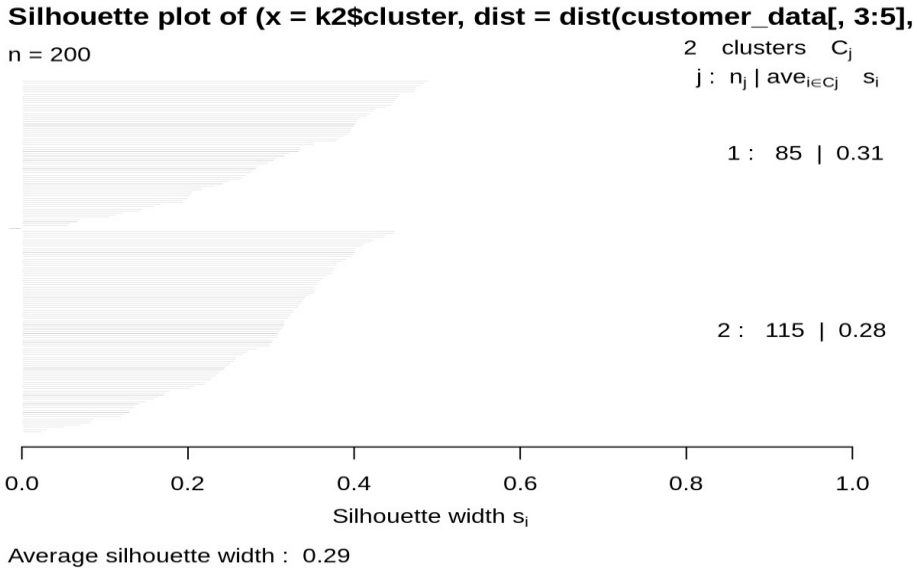
Determining Optimal Clusters 1)Elbow Method



Inference: From the above graph, we conclude that 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot.

2)Average Silhouette Method

Using the silhouette function in the cluster package, we computed the average silhouette width using the kmean function. Here, the optimal cluster will possess highest average. **When n=2**



When n=3

Average silhouette width : 0.38

When n=4

Average silhouette width : 0.41

When n=5

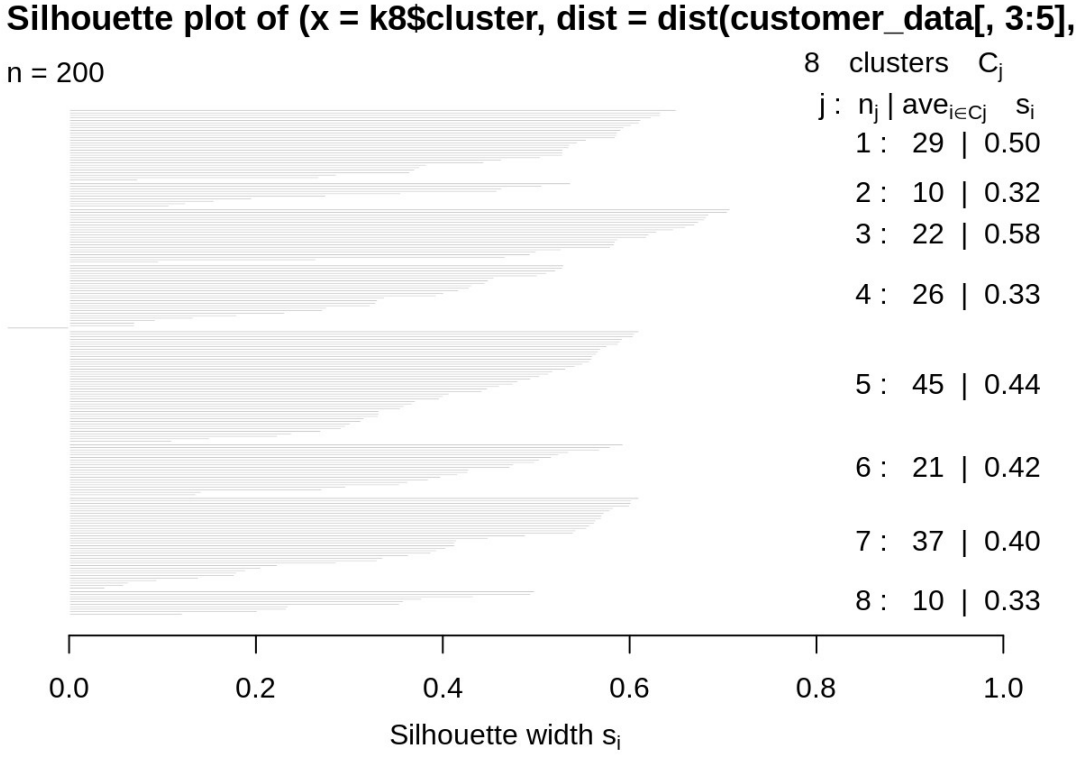
Average silhouette width : 0.44

When n=6

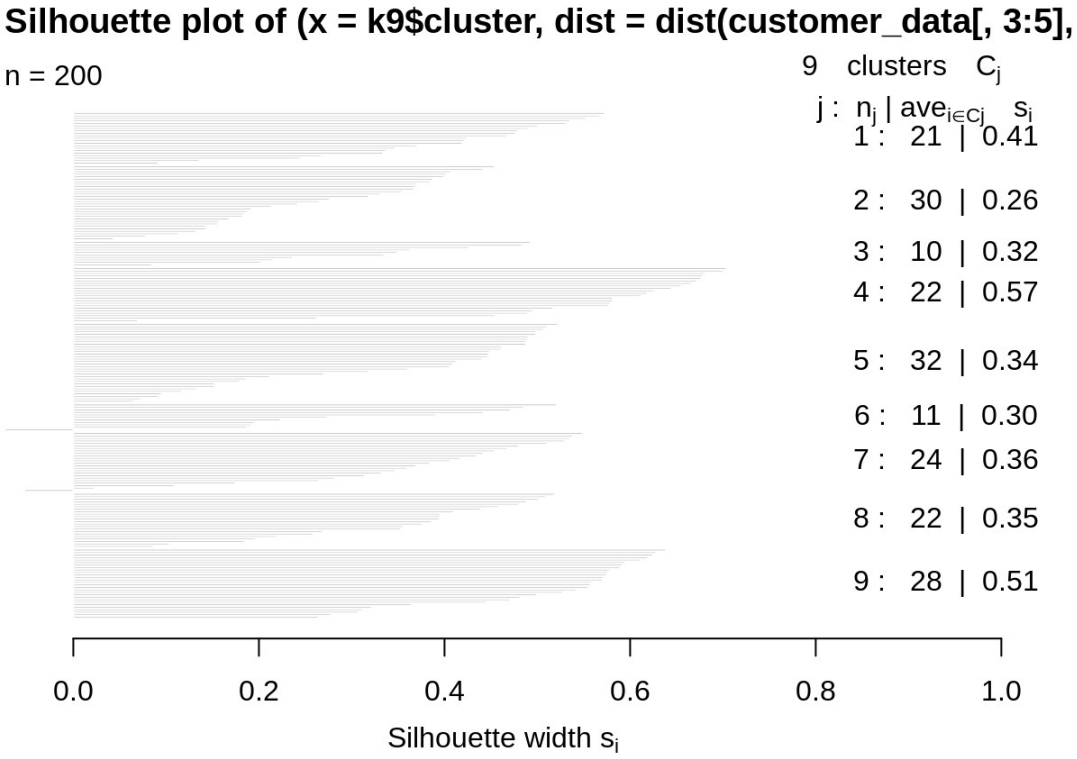
Average silhouette width : 0.45

When n=7

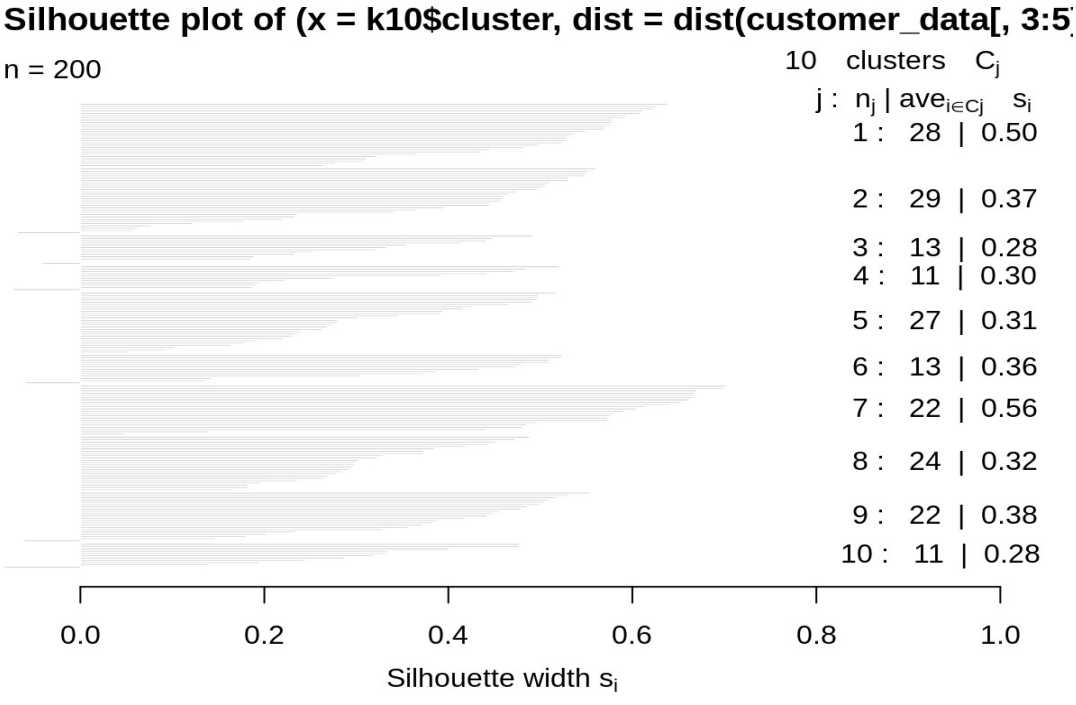
When n=8



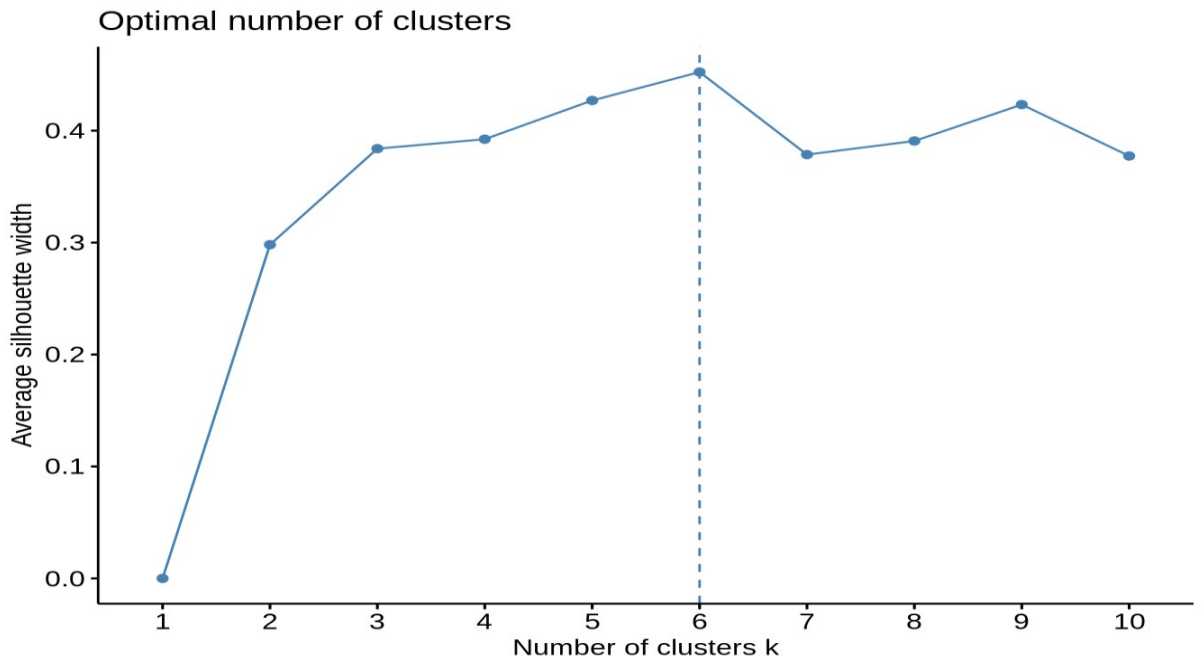
When n=9



When n=10

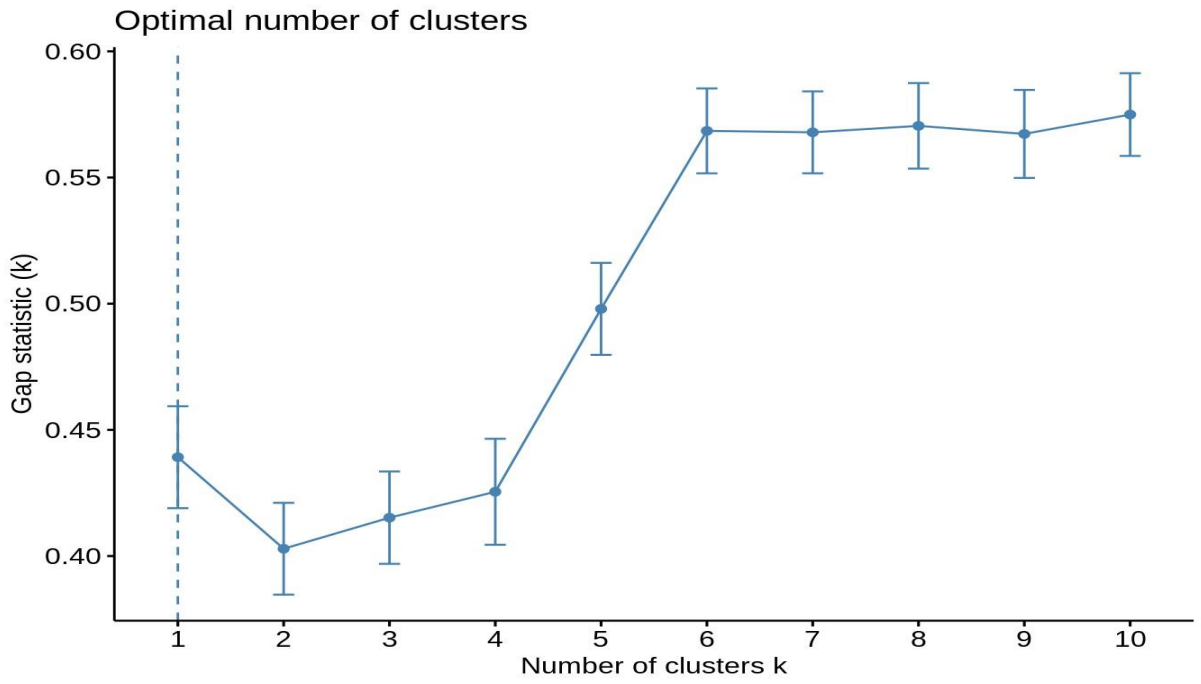


Now, we made use of the fviz_nbclust() function to determine and visualize the optimal number of clusters as follows –



3)Gap Statistic Method

For computing the gap statistics method, we had utilize the clusGap function for providing gap statistic as well as standard error for a given output.



K-means clustering with 6 clusters of sizes 45, 21, 35, 39, 38, 22

Cluster means:

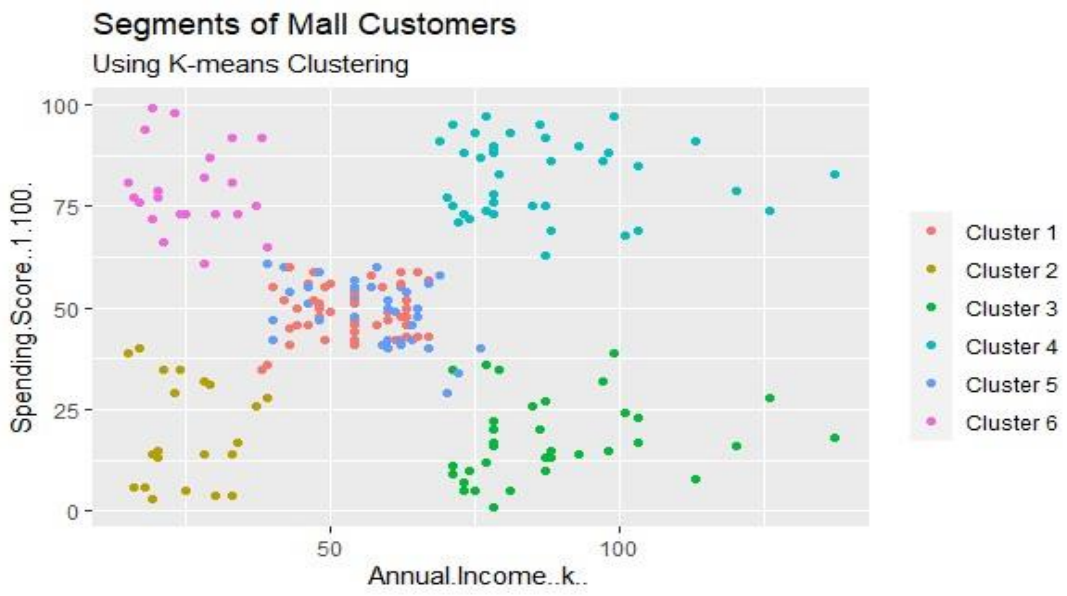
```
      Age Annual.Income..k.. Spending.Score..1.100.
1 56.15556         53.37778          49.08889
2 44.14286         25.14286          19.52381
3 41.68571         88.22857          17.28571
4 32.69231         86.53846          82.12821
5 27.00000         56.65789          49.13158
6 25.27273         25.72727          79.36364

Clustering vector:
  [1] 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6
 [33] 2 6 2 6 2 6 2 6 1 6 1 5 2 6 1 5 5 5 1 5 5 1 1 1 1 1 5 1 1 5 1 1
 [65] 1 5 1 1 5 5 1 1 1 1 1 5 1 5 5 1 1 5 1 1 5 1 1 5 5 1 1 5 1 5 5 5
 [97] 1 5 1 5 5 1 1 5 1 5 1 1 1 1 1 5 5 5 5 5 1 1 1 1 5 5 5 4 5 4 3 4
[129] 3 4 3 4 5 4 3 4 3 4 3 4 3 4 5 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
[161] 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
[193] 3 4 3 4 3 4 3 4

Within cluster sum of squares by cluster:
[1]  8062.133  7732.381 16690.857 13972.359  7742.895  4099.818
 (between_SS / total_SS =  81.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

In the output of our kmeans operation, we observe a list with several key information. From this, we conclude the useful information being –

- **cluster –** This is a vector of several integers that denote the cluster which has an allocation of each point.
- **totss –** This represents the total sum of squares.
- **centers –** Matrix comprising of several cluster centers
- **withinss –** This is a vector representing the intra-cluster sum of squares having one component per cluster.
- **tot.withinss –** This denotes the total intra-cluster sum of squares.
- **betweenss –** This is the sum of between-cluster squares.
- **size –** The total number of points that each cluster holds.

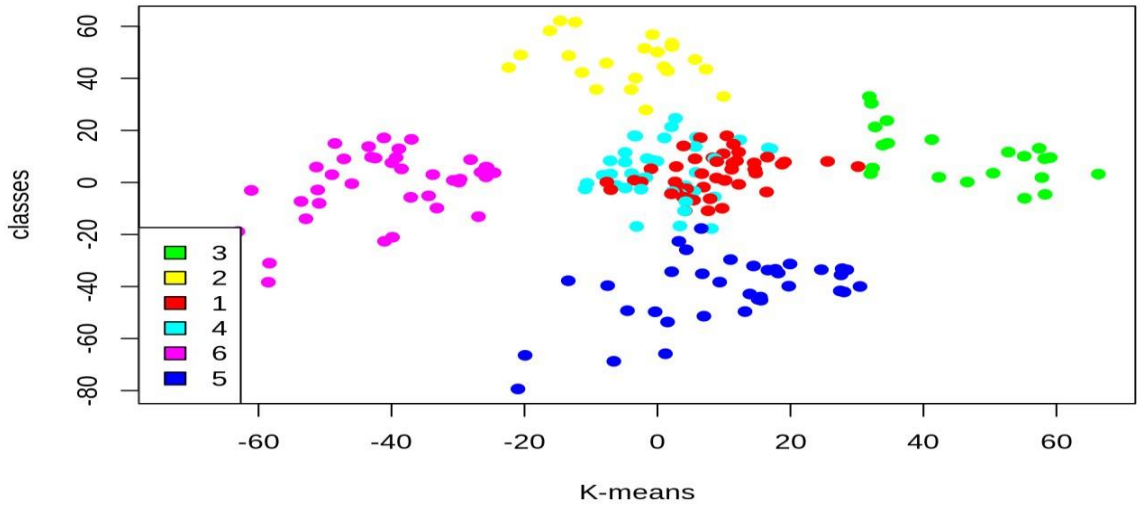
Visualizing the Clustering Results using the First Two Principal Components

```
Importance of components:
                            PC1      PC2      PC3
Standard deviation     26.4625 26.1597 12.9317
Proportion of Variance  0.4512  0.4410  0.1078
Cumulative Proportion   0.4512  0.8922  1.0000
                               PC1         PC2
Age                     0.1889742 -0.1309652
Annual.Income..k..     -0.5886410 -0.8083757
Spending.Score..1.100. -0.7859965  0.5739136
```



From the above visualization, we observe that there is a distribution of 6 clusters as follows –

Cluster 6 and 4 – These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary.
Cluster 1 – This cluster represents the customer_data having a high annual income as well as a high annual spend.
Cluster 3 – This cluster denotes the customer_data with low annual income as well as low yearly spend of income.
Cluster 2 – This cluster denotes a high annual income and low yearly spend.
Cluster 5 – This cluster represents a low annual income but its high yearly expenditure.



Cluster 4 and 1 – These two clusters consist of customers with medium PCA1 and medium PCA2 score.
Cluster 6 – This cluster represents customers having a high PCA2 and a low PCA1.
Cluster 5 – In this cluster, there are customers with a medium PCA1 and a low PCA2 score.
Cluster 3 – This cluster comprises of customers with a high PCA1 income and a high PCA2.
Cluster 2 – This comprises of customers with a high PCA2 and a medium annual spend of income.

With the help of clustering, we can understand the variables much better, prompting us to take careful decisions. With the identification of customers, companies can release products and services that target customers based on several parameters like income, age, spending patterns, etc. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation.