**Customer Segmentation Classification**

The project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**Manjunadha Nagasai Tummala Mohith Sashank Namburu Thanishq Chakkirala**

**(AP20110010184) (AP20110010146) (AP20110010171)**

**A picture containing text

Description automatically generated**

Under the Guidance of

**Dr. Sriramulu Bojjagani**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**November, 2023**

# Certificate

Date: 19-Jan-24

This is to certify that the work present in this Project entitled “**Customer  Segmentation Classification**” has been carried out by **Manjunadha nagasai Tummala, Mohith Sashank Namburu, Thanishq Chakkirala** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in the **School of Engineering and Sciences**.

**Supervisor**

(Signature)

Dr. Sriramulu Bojjagani

Assistant Professor,

Department of Computer Science Engineering.

SRM University AP.

# Acknowledgments

In making this project, and completing it successfully, I had to get help and guidance from my instructor. I am grateful that I was given the opportunity to make this project, which has enhanced my knowledge in so many aspects.

I would like to show my gratitude to my course instructor, Dr. Sriramulu Bojjagani for giving me great guidance for this project. I wish to extend my special thanks to my teammates who have always been supportive and guided me throughout this project.

# Table of Contents

[Certificate i](#_Toc105799088)

[Acknowledgments ii](#_Toc105799089)

[Table of Contents iii](#_Toc105799090)

[Abstract iv](#_Toc105799091)

[List of Figures v](#_Toc105799092)

[1. Introduction 1](#_Toc105799099)

[1.1 Customer Segmentation 1](#_Toc105799100)

1.2  [R Programming 2](#_Toc105799101)

[2. Methodology 3](#_Toc105799102)

[2.1 Reading Events From Dataset 3](#_Toc105799104)

[2.2 Customer Gender Visualization 3](#_Toc105799105)

2.3 Visualization of age distribution………………………………………………………………...4

[3. Discussion 9](#_Toc105799106)

[4. Conclusion 11](#_Toc105799107)

[5. Future Work 12](#_Toc105799108)

[References 13](#_Toc105799109)

# Abstract

This Project focuses on the critical application of unsupervised learning known as Customer Segmentation. Leveraging advanced clustering techniques, specifically the k-means algorithm, the project aims to unveil distinct customer segments within unlabeled datasets. The utilization of k-means clustering becomes paramount in this context, as it enables companies to discern various customer segments effectively. By identifying these segments, businesses can strategically target and engage with their potential user base, thereby optimizing marketing efforts and enhancing overall customer satisfaction.

# List of Figures

Figure 1 : Customer Segmentation…………………………………………………2

Figure 2 : overview of customers dataset………………………………………….3

Figure 3 : Gender Comparison………………………………………………….…..4

Figure 4 : Gender Ratio………………………………………………………………4

Figure 5 : Age Distribution………………………………………………………..…5

Figure 6 : Annual Income……………………………………………………………6

Figure 7 : Clusters…………………………………………………………………….8

Figure 8 : Visualization……………………………………………………….……...9

Figure 9: K-Means Visualization…………………………………………………....10

# Introduction

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this project, we will make use of k-mean Clustering which is the essential algorithm for clustering unlabelled dataset. In the pursuit of identifying the ideal methodology for discovering your best customer, we delve into the significant realm of Customer Segmentation, a key application of unsupervised learning. Leveraging clustering techniques, particularly the k-means algorithm, we aim to uncover distinct customer segments within unlabelled datasets. This project marks a practical implementation of customer segmentation in R, harnessing the power of machine learning to strategically target and engage with different customer segments. By integrating unsupervised learning techniques into the customer segmentation process, we enhance our ability to optimize marketing strategies and elevate overall customer satisfaction.

## Customer Segmentation

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

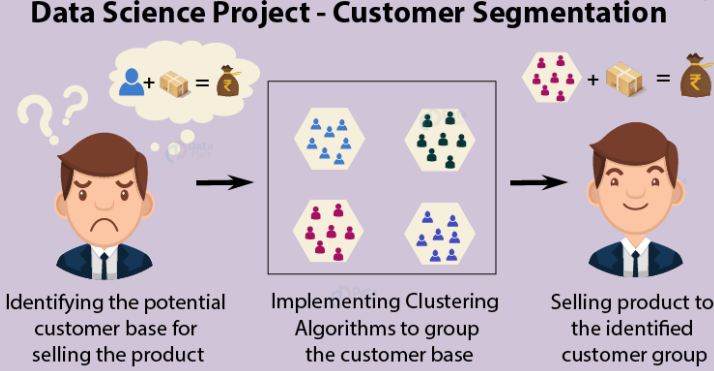


Figure1 : Customer Segmentation

## R Programming

R was specifically designed for statistical analysis, which makes it highly suitable for data science applications. Although the learning curve for programming with R can be steep, especially for people without prior programming experience, the tools now available for carrying out text analysis in R make it easy to perform powerful, cutting-edge text analytics using only a few simple commands. One of the keys to R’s explosive growth has been its densely populated collection of extension software libraries, known in R terminology as packages, supplied and maintained by R’s extensive user community. Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package. The best known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published. Text analysis in particular has become well established in R. There is a vast collection of dedicated text processing and text analysis packages, from low-level string operations to advanced text modeling techniques such as fitting Latent Dirichlet Allocation models, R provides it all. One of the main advantages of performing text analysis in R is that it is often possible, and relatively easy, to switch between different packages or to combine them. Recent efforts among the R text analysis developers’ community are designed to promote this interoperability to maximize flexibility and choice among users. As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features.

# Methodology

In the first step of this data science project, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it.



## READING EVENTS FROM MALL\_CUSTOMERS.CSV

Before going to customer segmentation analysis, the first step is to read the data for performing analysis on. The data is saved in dataset named as Mall\_Customers.csv. This dataset contains 400 record of various type of customers. The events saved in dataset are unstructured. To perform analysis, reading of data set is done using command “read.csv”.

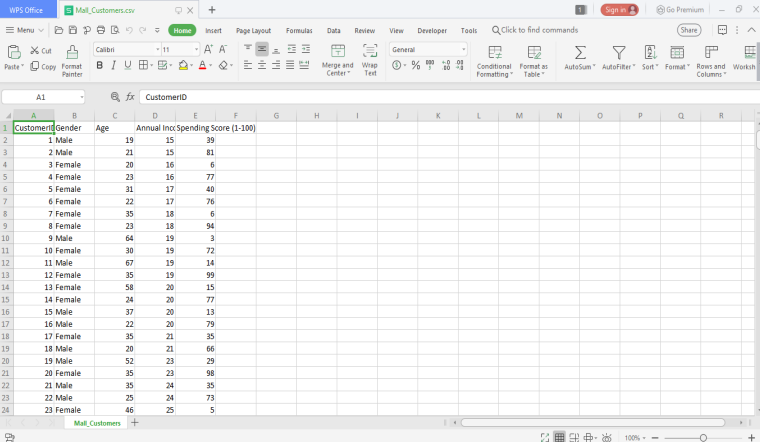


Figure 2: overview of customers dataset

## Customer Gender Visualization:

In this, we will create a barplot and a piechart to show the gender distribution across our customer\_data dataset. A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable. R uses the function barplot() to create bar charts. R can draw both vertical and Horizontal bars in the bar chart. In bar chart each of the bars can be given different colors.

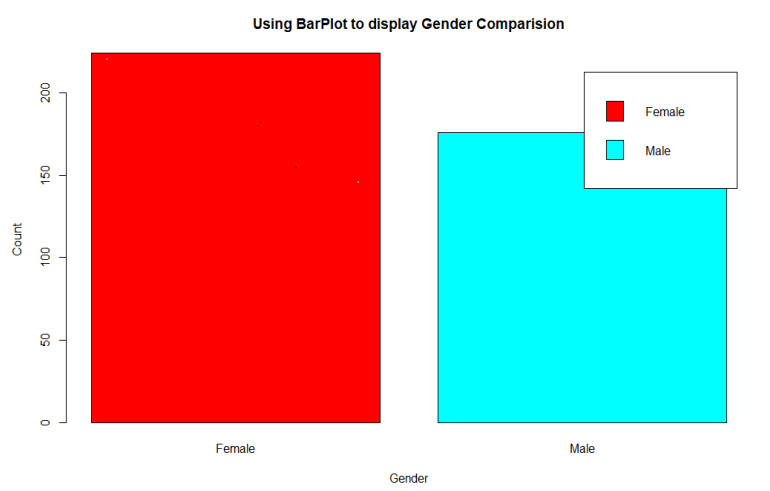


Figure 3: Gender Comparison

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

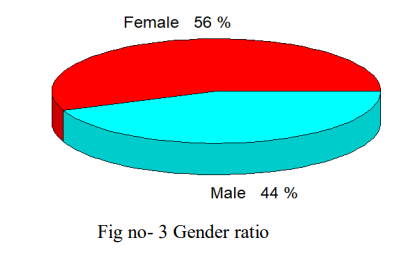


Figure 4: Gender Ratio

## Visualization of Age Distribution

Let us plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable.

Code:

summary(customer\_data$Age)

hist(customer\_data$Age,

col="blue",

main="Histogram to Show Count of Age Class",

xlab="Age Class",

ylab="Frequency", labels=TRUE)

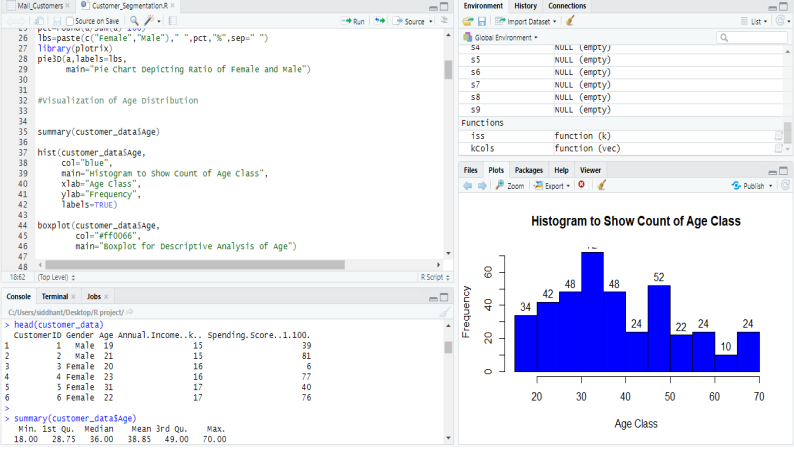


Figure 5: Age Distribution

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:

In this section of the R project, we will create visualizations to analyze the annual income of the customers. We will plot a histogram and then we will proceed to examine this data using a density plot.

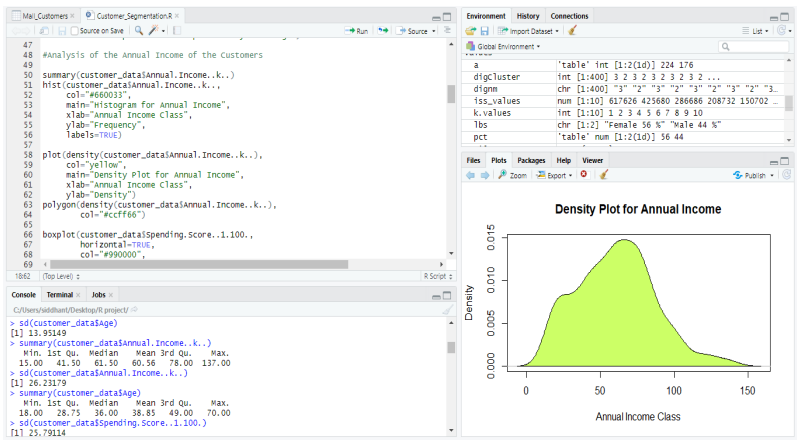


Figure 6: Annual Income

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.

## K-means Algorithm

While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that we wish to produce in the final output. The algorithm starts by selecting k objects from dataset randomly that will serve as the initial centers for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as “cluster assignment”. When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centers, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration.

Summing up the K-means clustering –

• We specify the number of clusters that we need to create.

• The algorithm selects k objects at random from the dataset. This object is the initial cluster or mean.

• The closest centroid obtains the assignment of a new observation. We base this assignment on the Euclidean Distance between object and the centroid.

• k clusters in the data points update the centroid through calculation of the new mean values present in all the data points of the cluster. The kth cluster’s centroid has a length of p that contains means of all variables for observations in the k-th cluster. We denote the number of variables with p.

• Iterative minimization of the total within the sum of squares. Then through the iterative minimization of the total sum of the square, the assignment stop wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations.

we calculate the clustering algorithm for several values of k. This can be done by creating a variation within k from 1 to 10 clusters. We then calculate the total intra-cluster sum of square (iss). Then, we proceed to plot iss based on the number of k clusters. This plot denotes the appropriate number of clusters required in our model. In the plot, the location of a bend or a knee is the indication of the optimum number of clusters. Let us implement this in R as follows –

Code:

library(purrr)

set.seed(123)

# function to calculate total intra-cluster sum of square

iss <- function(k) {

kmeans(customer\_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withinss

}

k.values <- 1:10

iss\_values <- map\_dbl(k.values, iss)

plot(k.values, iss\_values,

type="b", pch = 19, frame = FALSE,

xlab="Number of clusters K",

ylab="Total intra-clusters sum of squares")

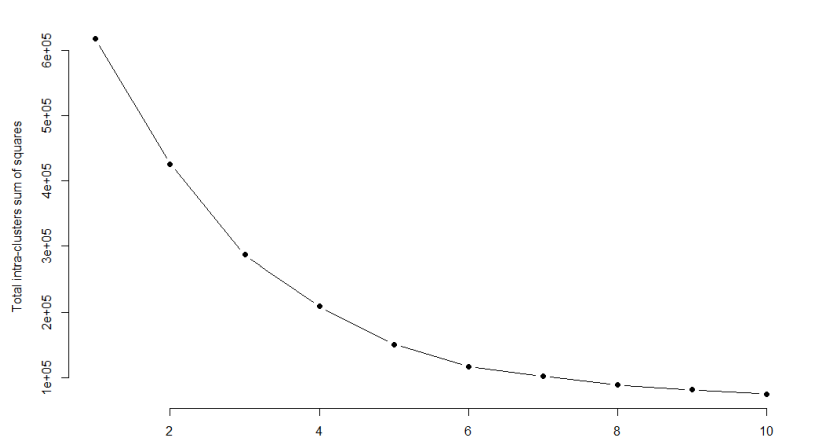


Figure 7: Clusters

# Discussion

**Visualizing the Clustering Results using the First Two Principle Components:**

A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called ‘markers’ connected by straight line segments. It is a basic type of chart common in many fields. Used across many fields, this type of graph can be quite helpful in depicting the changes in values over time. We are going to use ggplot for depicting the line plot.

**Code:**

set.seed(1)

ggplot(customer\_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +

geom\_point(stat = “identity”, aes(color = as.factor(k6$cluster))) +

scale\_color\_discrete(name=” “,

breaks=c(“1”, “2”, “3”, “4”, “5”,”6”),

labels=c(“Cluster 1”, “Cluster 2”, “Cluster 3”, “Cluster 4”, “Cluster 5”,”Cluster 6”))

ggtitle(“Segments of Mall Customers”, subtitle = “Using K-means Clustering”)

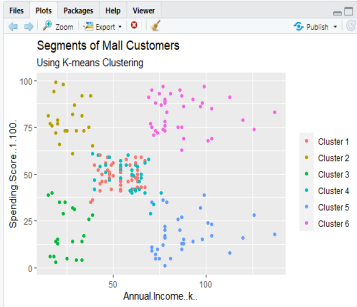


Figure 8: Visualization

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.

K-Means Visualization:

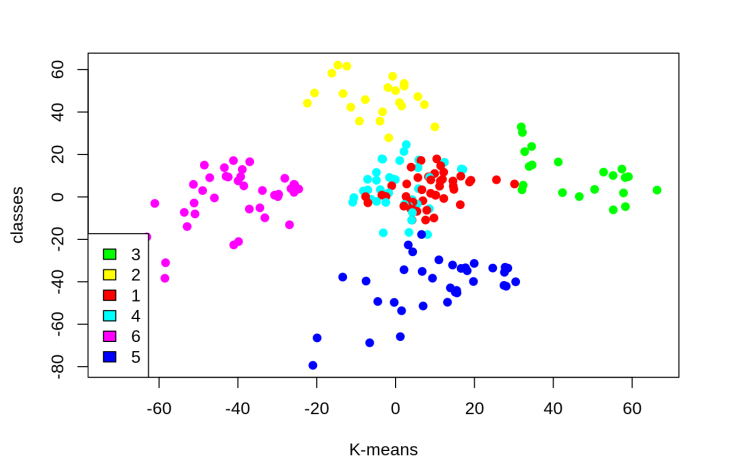


Figure 9: K-Means Visualization

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.

# Conclusion

In this project, we embarked on a comprehensive exploration of the customer segmentation model, leveraging the capabilities of unsupervised machine learning. Our focus extended beyond traditional applications, as we successfully adapted the model to various industries, including e-commerce, finance and banking, healthcare, and tourism. The robustness and scalability of our approach were evident as we navigated the intricacies of diverse datasets and customer behaviors within each sector.

The development of our customer segmentation model involved a meticulous process, prominently featuring the application of K-means clustering, a powerful unsupervised learning algorithm. We began by rigorously analyzing and visualizing the data, gaining valuable insights into customer patterns and behaviors. Subsequently, our implementation of the K-means clustering algorithm allowed us to categorize customers effectively, offering a foundation for targeted strategies and personalized services.

# Future Work

Looking forward, there are a few things we can do to make our research even better. Firstly, we might want to try our methods on bigger sets of data to make sure they work well in lots of different situations. Also, we could check if using different computer programs gives us better results. It's like trying out different tools to see which one works best. Talking to more people who know a lot about this topic and getting their ideas could be really helpful. If we want to make our research more useful in the real world, it's a good idea to ask people who might use it what they think and if they have any ideas to make it better. So, there's still a bunch of things we can do to make our research more helpful for everyone.

For future research, there are exciting avenues to explore. Firstly, delving into the intricacies of industry-specific features and customer behaviors within e-commerce, finance, healthcare, and tourism can refine our segmentation model. Investigating how external factors, such as economic trends or technological advancements, influence customer segments in these sectors could provide a more nuanced understanding. Additionally, exploring the integration of real-time data and machine learning algorithms to adapt to evolving customer behaviors in these dynamic industries represents a promising direction. Collaborating with industry experts and stakeholders to gather feedback on the practical implications of our segmentation approach in these domains can contribute to its real-world applicability. Furthermore, investigating the ethical considerations surrounding the use of customer data in sensitive sectors like healthcare and finance is crucial for ensuring responsible and secure implementation. This future work aims to not only broaden the scope of our research but also to enhance its relevance and impact in diverse and vital sectors.

# References

1. T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.
2. V. L. Narayana, S. Sirisha, G. Divya, N. L. S. Pooja and S. A. Nouf, "Mall Customer Segmentation Using Machine Learning," 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022, pp. 1280-1288, doi: 10.1109/ICEARS53579.2022.9752447.
3. Mulyo I, Heikal J. Customer Clustering Using The K-Means Clustering Algorithm in Shopping Mall in Indonesia. MAJ [Internet]. 4Jan.2023 [cited 13Nov.2023];11(4):365-71.
4. M. F. Alam, R. Singh and S. Katiyar, "Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 94-98, doi: 10.1109/ICAC3N53548.2021.9725644.