# CUSTOMER SEGMENTATION FOR SHOPPING MALL A PROJECT REPORT

*for*

# CAPSTONE PROJECT

*in*

# M.Tech. (Integrated) Software Engineering

*by*

# VIGNESH.P(18MIS0277)

*Under the Guidance of*

# Dr. KARTHIKEYAN D

Associate Professor, SITE

Guide signature



# School of Information Technology and Engineering

April, 2023

**DECLARATION BY THE CANDIDATE**

I hereby declare that the project report entitled **“CUSTOMER SEGMENTATION FOR SHOPPING MALL”** submitted by me to Vellore Institute of Technology; Vellore in partial fulfillment of the requirement for the award of the course **CAPSTONE PROJECT** is a record of bonafide project work carried out by me under the guidance of **Dr. KARTHIKEYAN. D .** I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place: Vellore Signature

Date: 4.4.2023 P. Vignesh



# School of Information Technology & Engineering [SITE]

**CERTIFICATE**

This is to certify that the project report entitled **“CUSTOMER SEGMENTATION FOR SHOPPING MALL”** submitted by **VIGNESH.P(18MIS0277)** to Vellore Institute of Technology, Vellore in partial fulfillment of the requirement for the award of the course **CAPSTONE PROJECT** is a record of bonafide work carried out by them under my guidance.

# Dr. Karthikeyan.D GUIDE

**Associate Professor, SITE**

# CUSTOMER SEGMENTATION FOR SHOPPING MALL

**Abstract**

Customer segmentation is a separation of a market into multiple distinct groups of consumers who share similar characteristics. Machine learning is achieving and aiming through every aspect that tends to develop different models. The shopping malls can use this machine learning models to manage their customers data to target on the right customers which tends to give more profits. This not only will increase sales however also makes the complexes economical. Some basic information regarding your customers like customer ID, age, gender, annual income and spending score are used. during this project, whenever you wish to seek out your best client. Customer segmentation is that the ideal methodology. In this machine learning project, we will make use of K-means clustering which is the essential algorithm for clustering unlabeled dataset. This data set was created for the purpose of learning only the customer segmentation concepts, also known as market basket analysis. A wide variety of analyzes will be created in this section. However, each case will be searched and machine learning algorithms will be used.

We are going to use R studio for working of the project.

**Keywords** – DWT, DCT, SVD, Alpha Blending, MSE, PSNR

# INTRODUCTION

Shopping Mall complexes are always indulged in the race of increasing their customers and hence want to make huge profits by increasing them. To achieve the task they desired we apply machine learning technique to many stores to increase the profits. It is amazing for us to realize about the fact that how machine learning can be used to aid in ambitions like that. The shopping Malls make huge profits by using their customer entry data to develop Machine Learning models to target the right customers for productivity in the sales. This not only increases the sales of the management but also makes the shopping malls efficient enoughto manage their productivity. A supermarket mall had some membership cards that consist of

some basic data about their customers such as the Customer ID, age, gender, annual income and purchasing score.

# BACKGROUND

By using the clustering techniques, the companies can identify the several segments of customers by allowing them to target the potential user for the shopping malls. By the data set that was created we can learn the customer segmentation concepts for shopping complexes that are also known as market basket analysis of data.

# Literature Survey

1. TITLE: Customer Segmentation Using K-Means Clustering Advantages of Customer Segmentation K-Means Clustering Algorithm.

PUBLISHED YEAR: JUNE 2021

This data set is created only for the learning purpose of the customer segmentation concepts , also known as Market Basket Analysis . I will demonstrate this by using Unsupervised Machine Learning Technique (K-Means Clustering Algorithm) in the simplest form.

1. TITLE: Using data mining techniques in customer segmentation AUTHORS: ZIAFAT.H & SHAKERI.M

PUBLISHED YEAR: JULY 2020

Data mining plays important role in marketing and is quite new. Data mining algorithms are powerful but cannot effectively work without the active support of business experts. Business knowledge can help and enrich the data mining results.

1. TITLE: Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis

AUTHORS: SARVARI, P. A., USTUNDAG, A., & TAKCI, H. YEAR: JUNE 2021

In this paper impacts of RFM and demographic attributes have been challenged in order to enrich factors that lend comprehension to customer segmentation. The prediction of customer behaviors is a strategically important and difficult issue because of the high variance and wide range of customer orders and preferences.

1. TITLE: Customer Segmentation Using Data Warehouse and Neural Networks AUTHORS: DAJANA ĆORIĆ, KATARINA ĆURKO, ZVONKO MERKAŠ.

YEAR: JULY 2019

This paper shows how data warehouse and neural networks can be useful in the process of predicting customer segments. Data warehouse can collect and transform millions of records for comprehensive analysis and provide tools that firms can and should use to understand the customer behavior.

1. TITLE: Customer Segmentation Using Clustering and Data Mining Techniques AUTHORS: KISHANA R. KASHWAN, MEMBER, IACSIT, AND C. M.

VELU

YEAR: FEB 2021

The comprehensive report of k-means clustering technique and SPSS Tool to develop a real time and online system for a particular super

market to predict sales in various annual seasonal cycles. The model developed was an intelligent tool which received inputs directly from sales data records and automatically updated segmentation statistics at the end of day’s business. An ANOVA analysis was also carried out to test the stability of the clusters.

1. TITLE: A financial data mining model for extracting customer behaviour AUTHORS: MARK K.Y. MAK, GEORGE T.S AND S.L. TING

YEAR: MAY 2021

It develops an intelligent Financial Data Mining Model (FDMM) for extracting customer behaviour in the financial industry, so as to increase the availability of decision support data and hence increase customer satisfaction. The proposed financial model first clusters the customers into several sectors, and then finds the correlation among these sectors.

1. TITLE: Online Auction Customer Segmentation Using a Neural Network Model

AUTHORS: CHAN, C. C. H YEAR: MAR 2021

In this environment, sellers and buyers have difficulty in knowing each other’s behaviors. Unpredictable results could happen to bring a lot of negative impacts to bidders. To realize the behaviour of customers, here SOM neural network model is used to cluster customer data into homogenous groups. By this we can segment the customers.

1. TITLE: K-means clustering via Principal Component Analysis. AUTHORS: C. Ding and X.-F. He

YEAR: MAY 2020

Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction. K-means clustering is a commonly used data clustering for performing

unsupervised learning tasks. Traditional data reduction perspective derives PCA as the best set of bilinear approximations.

1. TITLE: A segmentation study of cinema consumers based on values and lifestyle Authors: asuncion diaz

Year: mar 18

The objective of this study is to analyse the segments of cinemagoers in a shopping centre based on their values and lifestyle. Hierarchical segmentation techniques are used to identify different groups of consumers. Specifically, four segments are obtained from a sample of 391 participants, and the variation among the segments in the frequency of leisure activities in the shopping centre is analysed

1. Title: CONSUMER CULTURAL STUDIES Yaer:December 2019

Authors: Edurado norman

The challenging retail environment requires a need to manage shopping malls effectively to understand the attributes that attract shopping mall visitors to visit shopping malls. The purpose of the study aimed to determine shopping mall visitors' perceptions or ratings towards shopping mall attributes they consider when choosing which shopping mall to visit.

1. Title: Determining shopping mall visitors' perceptions on mall attributes Author: siphon saleto

The data were collected at shopping centre in the capital city of South Africa, City of Tshwane. A descriptive analysis method was used to analyze the quantitative data. The findings of the study revealed that the shopping mall visitors' ranked adequate parking availability high. This study contributes to the current literature and provides valuable information to South African retailers and shopping mall developers, with regard to marketing communications and marketing strategies that aim to attract shopping mall visitors. Suggestions for future research are provided.

1. Title: Examining shopping mall consumer decision-making styles, satisfaction and purchase intention

Author: seyed ali alavi Year: mar 2020

Understanding consumer decision-making (CDM) styles is essential for market segmentation, positioning and crafting marketing strategies within a market. Few studies have examined the structural relationship among decision-making styles that consumers exhibit during mall

shopping, level of satisfaction and purchase intention. The purpose of this study was to examine CDM styles as the antecedents and predictors of level of satisfaction and purchase intention.

1. Title: Segmenting Young Indian Impulsive Shoppers:

Authors: sanjrrv prashar Year: mar 2021

There exist five different shopper segments on the basis of the antecedent factors. This research also established a relationship between income levels and the motives behind impulsive shopping. Arguably, to the best of our knowledge, this is the first empirical study to examine and segment young Indian shoppers on the basis of their impulsiveness in buying.

1. Title: Clustering Internet Shoppers: An Empirical Finding from Indonesia. Author: m.mujiya mulkah

Year: may 2019

The first five segments are considered to be more likely to purchase products online, i.e., the online shoppers; while the rest three are the non-online shoppers. The ANOVA test confirmed that the eight segments were appropriate since it created more differentiated and consistent clusters.

1. Title: Performance-enhanced rough $$k$$-means clustering algorithm Author: sivaguru

Year: jan 2021

Customer segmentation (CS) is the most critical application in the field of customer relationship management that primarily depends on clustering algorithms. Rough k-means (RKM) clustering algorithm is widely adopted in the literature for achieving CS objective. However, the RKM has certain limitations that prevent its successful application to CS. First, it is sensitive to random initial cluster centers.

Done by Gokul.k 18MIS0139

1. Title: Quality improvement initiatives based on customer and service provider perspectives in shopping malls

Author: Fatma pakdil Year: april 2018

Design/methodology/approach QFD is used to determine and close the gap between the most important customer needs and expectations and the opinions of service providers using a unique platform. Findings On customer side, the highest relative weight was given to “prompt response to customer concerns”, “not being crowded and loud”, “providing services for disabled customers”, and “security of mall” customer expectations.

1. Title: Experience Economy and the Management of Shopping Centers: The Role of Entertainment

Author: savelli Year: jan 2019

This chapter applies the experience logic perspective to the retail industry by analyzing the role and the management of entertainment strategies in the shopping center format. The purpose is twofold: (i) proposing a conceptual classification of entertainment based on the existing literature; (ii) examining the influences of entertainment strategies on shopping centers’ market performances to provide suggestions in regards to the effectiveness of such strategies.

1. Title: Shopping Motives, Mall Attractiveness, and Visiting Patterns in Shopping Malls in the Middle East: A Segmentation Approach

Author: m.h.koksal Year: mar 2019

Middle Eastern consumers change their shopping behaviors in line with developments in global markets. Large-scale Western-style malls with various well-known stores, shops, and cafes and restaurants have become shopping and attraction spaces. The purpose of the study is to segment Lebanese customers based on mall shopping motives. The data in the study were collected through a structured questionnaire distributed in the main shopping malls of the Lebanese capital, Beirut. A total of 300 mall customers were interviewed at mall exits.

1. Title: The importance of distance and attraction in patronizing a shopping mall. Author: pioter kwatiok

Year: Dec 2020

Analysis reveals that the performance of buying and social activities factors had a significant impact on the frequency of visits, while the amount of time spent per visit was significantly affected only by the social activities factor. Furthermore, mall size is more important than distance to the mall. Finally, gender differences in shoppers’ mall preferences and behaviors were reported.

1. Title: Profiling Shopping Mall Customers during hard times Year: april 2019

Author: Shopping malls are facing times of hardship due to the recent economic crisis, the maturity of this retailing format and the increasing competition of the electronic commerce. This study aims to examine and profile shopping mall customer segments, since this context of hardship may have altered the previous mall customer segmentations.

1. Title: A segmentation of Turkish consumers based on their motives to visit shopping centres Year: mar 19

Author: banu paksoy

This paper aims to contribute to the literature by providing a segmentation of Turkish consumers based on their motives for visiting shopping centres. First, we identified the motives that Turkish consumers had for visiting shopping centres and then we used those motives to segment consumers. Data were collected through a survey from 390 participants living in the six largest cities in Turkey.

1. Title: Integration of materialism with shopping motivations: Motivations based profile of Indian mall shoppers

Author: Devinder pal singh Year: oct 2018

The paper aims to investigate materialism as one of the retail shopping motives along with utilitarian/hedonic motivations in the Indian context. It aims to identify the key shopping motivations, which explain the shopping value in the context of malls. Furthermore, it intends to develop a shopping motivations-based typology of Indian mall shoppers, and to profile the motivational and demographic characteristics of the discerned segments. Design/methodology/approach The data were collected through a mall intercept survey.

1. Title: Role of the eco-natural environment as an alternative attractiveness factor in malls Author: Leonardo

Year: june 2019

The mall industry in Latin America has grown rapidly in the past decade, offering diverse proposals oriented to improving the attractiveness of this commerce format; along this line, despite the fact that several studies have analyzed variables from an ecological perspective based on the relationship of the physical space with the consumers .

1. Title: The role of culture and purchasing power parity in shaping mall-shoppers’ profiles

Author: shaked giloba Year: jan 2020

Global mall managers and retailers need to recognize and address variations among groups of shoppers, particularly how they vary between countries, to optimize their global operations. Despite many international mall-shopper studies, only few have compared countries using uniform constructs and descriptors.

1. Title: A Data Analytics Approach to Online Tourists’ Reviews Evaluation

Author: evripides Year: jan 2020

Segmentation has come into its own as a marketing tool and though less common in service research, some research documents its merit; for example, Grove et al. segmented tourists by their perceptions of the relative importance of various delivery.

1. title: The impact of an exciting store environment on consumer pleasure and shopping intentions

Author: Jonas homquist Year: mar 2021

In this paper, we reinvestigate whether a stimulating store environment is beneficial or whether it could have a negative effect on consumers. Consistent with previous studies we find that the answer lies in the consumer's motivational orientation: a stimulating in-store experience has a positive effect on pleasantness and shopping intentions for consumers with a recreational motive while at the same time having a negative effect on both pleasantness and shopping intentions for task-oriented consumers.

1. Title: The critical factors of shopping malls in urban complexes Author: Eddie c.m. hui

Year: feb 2020

Then cluster analysis is applied to divide the customers into three segments, showing the importance of each factor to different customer segments. Furthermore, correspondence analysis is conducted to investigate the relationship between customer segments and customer characteristics (gender, occupation, age and income).

1. Title: Segmentation of clearance sales shoppers based on store attributes Author: hemalatha

Year: mar 2020

Clearance sales are commonly used by retailers selling season goods. These special sales events are designed to increase store traffic and sales. Clearance sales act as stimulus to consumers who are likely to elicit a positive response.

1. Title: Customer-Oriented Benefit Segmentation: An Integrated Approach Author: Mohammed hasan aghdaie

Year: feb 2019

Segmentation is a common and important task for most of marketing departments. Besides, other marketing decisions are influenced by market segmentation results. Benefit segmentation is one of the best approaches for market segmentation among others. In this paper, we proposed a novel hybrid benefit segmentation approach which applied two-stage clustering.

1. Title: Online Auction Customer Segmentation Using a Neural

Network Model AUTHORS: CHAN, C. C. H

Year: feb 2021

In this environment, sellers and buyers have difficulty in knowing each other’s behaviors. Unpredictable results could happen to bring a lot of negative impacts to bidders. To realize the behaviour of customers, here SOM neural network model is used to cluster customer data into homogenous groups.

# 1. DATASET DESCRIPTION & SAMPLE DATA

1. The count is 200 means we have records of 200 customers with us.
2. The minimum age of customer in our data is 18 yrs and maximum age is 70. The mean here is 38 and median is 36.Here Mean>Median means our data has high outliers i.e. more of youngsters prefer to go malls.
3. The minimum annual income of customer is 15k$ and maximum is 137k$.T. The mean and median here is 60k$ and 61k$ respectively.
4. Spending Score is something you assign to the customer based on your defined parameters like customer behaviour and purchasing data. Here the minimum spending score assigned is 1 and maximum ranges till 99.Both mean and median is 50.

# 2.PROPOSED ALGORITHM WITH FLOWCHART:

Here in our proposed system we will see the descriptive analysis of our data and we are going to implement the K-means clustering Algorithm to extract the information from the related customer annual spending behaviors. In this project we use K-means clustering for clustering the unlabeled dataset. K-means clustering is a simple and a fast algorithm to implement and it has computational advantages in terms of the implementation of larger datasets. In the customer Segmentation process we are gone divide the customers into divisions relevant to their marketing such as the gender, age, interests and etc. the customer segmentation has a several key differentiators divided into customer groups such as the data related demographics, geographic, economic status and behavioural patterns are going to play an important role in determining process of the various segments.

K-Means Clustering algorithm to cluster the data is used in the system proposed and to implement the K- means clustering we have to look the elbow method that is a interpretation and validation of the consistency within the cluster analysis from the number of clusters in the data set that has been used to implement the system. Here in this process the clusters obtained in previous iterations are the one used in the next iterations.

K-means clustering usage method: Specify the cluster’s number that we are going to create. An object as initial cluster or mean is used. The algorithm uses K objects at random from the dataset. Euclidean Distance between object and centroid is observed. Through calculation K- clusters update the centroid and p is the number of variables in the observations in k-th cluster. Iterative minimization technique is used to sum up the squares. With the default value in the R software, we are going to use the maximum iterations.

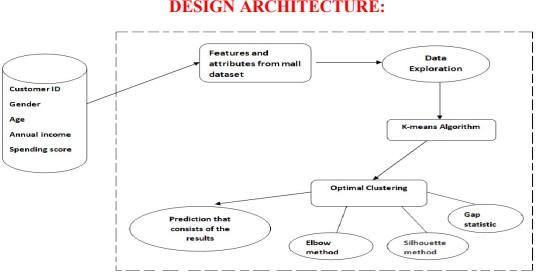
11

Here we are going to Determine the Optimal Clusters so we need the three popular methods:-

Elbow method Silhouette method Gap Static

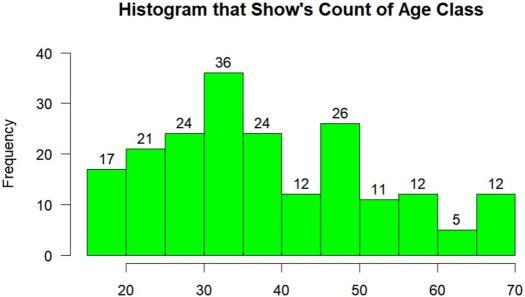


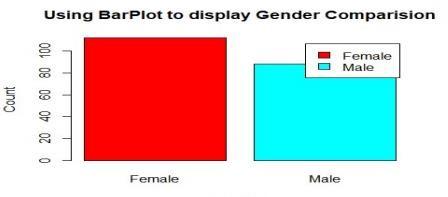
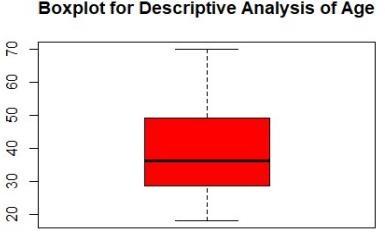
In the proposed system we are going to implement these K-means clustering and Optimal clustering techniques in the R software. Our system is going to show the proper customer segmentation in the shopping mall and we are going to get proper information regarding the customers so that we can plan the mall according to them.

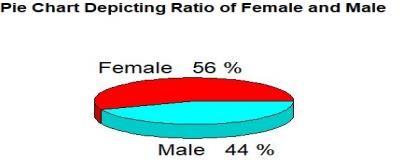


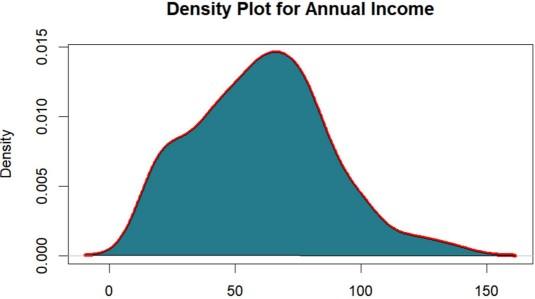
# EXPERIMENTS RESULTS

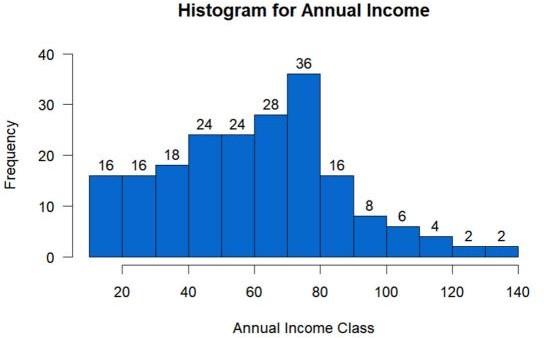
.

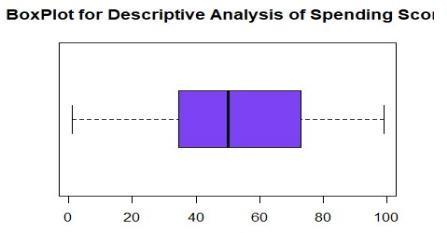


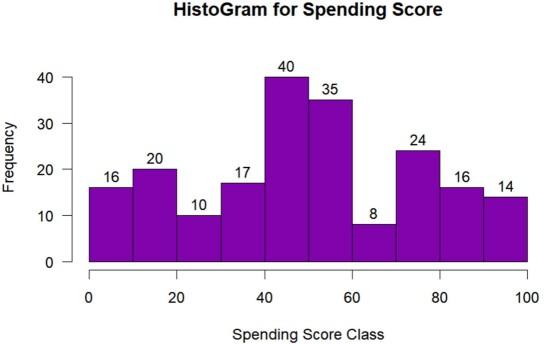


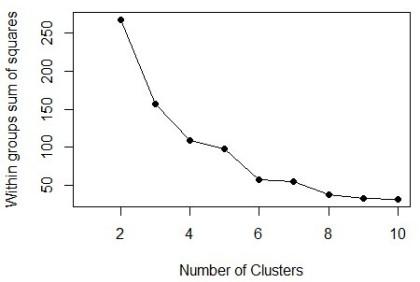


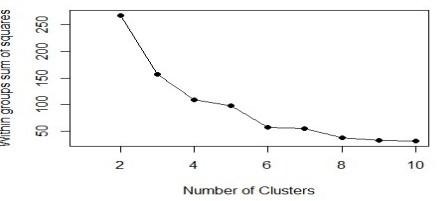


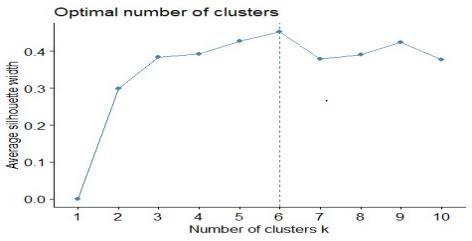
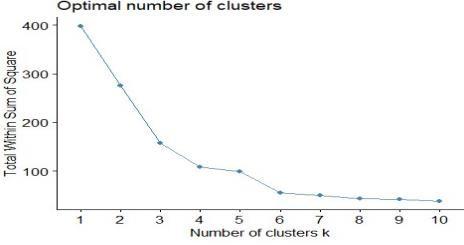


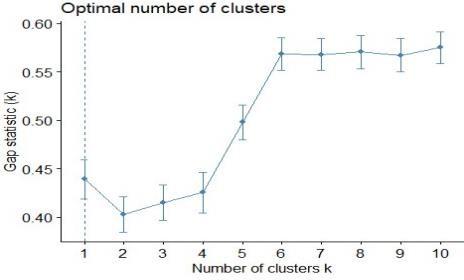


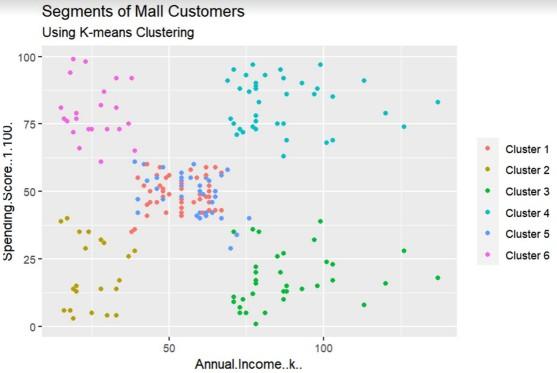


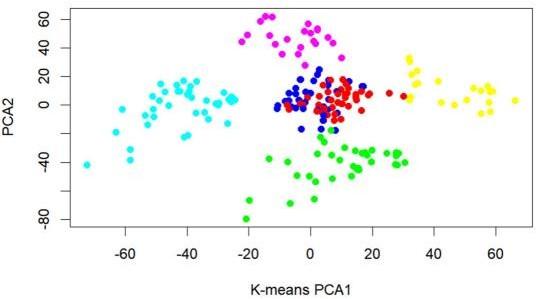












# COMPARATIVE STUDY / RESULTS AND DISCUSSION

**Elbow method:**

The main goal for K-mean method is for the cluster partitioning and to define that the intra-cluster variation that is always minimum between the clusters.

We can evaluate the compactness of the cluster boundary by the measurement of total intracluster variation. we are going to calculate the clustering algorithm for

many values of the K. then only we calculate the total intra-cluster sum of squares. Then we are going to get the output consists og the plots that denote the appropriate number of clusters required.

Silhouette method:

With the help of this method we are going to determine the how the data object is there within the cluster. If we are going to obtain a high average silhouette width, then it means we are having good clustering in our system and the proposed method. Silhouette method is used to calculate the mean of the observations for the different K- values with the optimal number of clusters we are going to maximize the significant values ofK-value clusters. We can also use the fviz\_nb cluster function to determine the optimal number of clusters used in the system.

# 3.CONCLUSION AND FUTURE WORK

With the info provided higher than, we have a tendency to might take some analysis outline which will be used for promoting set up as follows: We might see from the EDA half that the feminine customers share (56%) is slightly more than male customers (44%), with this data we have a tendency to might targeting the male customers additional for promoting campaign or promotions than feminine customers although the proportion completely different isn't too huge. This case we have a tendency to can also select male client target showing wisdom with combined factors and age. Cluster one & three generally has low outlay scores, despite their financial gain levels, and usually the purchaser square measure higher than 40s. With that information, we have a tendency to might bear in mind to analysis and adding some

brands that square measure fashionable among customers at those ages, and running campaigns to focus on them with the correct merchandise.

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. We went through the customer segmentation model.

We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. We analyzed and visualized the data and then proceeded to implement our algorithm.

# 3.REFERENCES

Ziafat, H., & Shakeri, M. (2019). Using data mining techniques in customer segmentation.

Journal of Engineering Research and Applications, 4(9), 70- 79

1. Ćorić, D., Ćurko, K., & Merkaš, Z. (2019). Customer Segmentation Using Data Warehouse and Neural Networks. WSEAS Transactions on Business and Economics, ISSN/E-ISSN: 1109-9526/2224-2899, Volume 12, 2015, Art.# 17, 12(17), 186-197.
2. Kishana R. Kashwan, Member, IACSIT, C.M.Velu, “CustomerSegmentation using clustering and Data Mining Techniques”, International Journal of Computer Theory and Engineering, Vol.5, No.6, December 2013.
3. Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes.
4. Mark K.Y.Mak,George T.S.Ho,S.L.Ting,“A Financial Data Mining Model for extracting Customer Behaviour”, INTECH open access publisher, 23 July 2021. C. Ding and X.-F. He, “K-means Clustering via Principal Component Analysis,” in Proc. Int'l Conf. Machine Learning (ICML), 2021, pp. 225- 232.
5. Chan, C. C. H. (2021). Online auction customer segmentation using a neural network model. International Journal of Applied Science and Engineering, 3(2), 101-109.
6. Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes.
7. Ş. Ozan, "A case study on customer segmentation by using machine learning methods", 2018 International Artificial Intelligence and Data Processing Symposium (IDAP), Sep 2018.
8. C. Ding and X.-F. He, “K-means Clustering via Principal Component Analysis,” in Proc. Int'l Conf. Machine Learning (ICML), 2020, pp. 225-232.
9. Chan, C. C. H. (2020). Online auction customer segmentation using a neural network model. International Journal of Applied Science and Engineering, 3(2), 101-109.

# Appendix

importing\_dataset.R #importing customer data set

customer\_data=read.csv("/Mall\_Customers.csv") dim(customer\_data)

## [1] 200 5

str(customer\_data)

## 'data.frame': 200 obs. of 5 variables:

## $ CustomerID : int 1 2 3 4 5 6 7 8 9 10 ...

## $ Genre : chr "Male" "Male" "Female" "Female" ... ## $ Age : int 19 21 20 23 31 22 35 23 64 30 ...

12

## $ Annual.Income..k.. : int 15 15 16 16 17 17 18 18 19 19 ...

## $ Spending.Score..1.100.: int 39 81 6 77 40 76 6 94 3 72 ... names(customer\_data)

## [1] "CustomerID" "Genre" "Age"

## [4] "Annual.Income..k.." "Spending.Score..1.100." #displaying data set values

head(customer\_data)

## CustomerID Genre Age Annual.Income..k.. Spending.Score..1.100. ## 1 1 Male 19 15 39

## 2 2 Male 21 15 81

## 3 3 Female 20 16 6

## 4 4 Female 23 16 77

## 5 5 Female 31 17 40

## 6 6 Female 22 17 76 tail(customer\_data)

## CustomerID Genre Age Annual.Income..k.. Spending.Score..1.100. ## 195 195 Female 47 120 16

## 196 196 Female 35 120 79

## 197 197 Female 45 126 28

## 198 198 Male 32 126 74

## 199 199 Male 32 137 18

## 200 200 Male 30 137 83

summary(customer\_data)

## CustomerID Genre Age Annual.Income..k..

## Min. : 1.00 Length:200 Min. :18.00 Min. : 15.00

## 1st Qu.: 50.75 Class :character 1st Qu.:28.75 1st Qu.: 41.50

## Median :100.50 Mode :character Median :36.00 Median : 61.50 ## Mean :100.50 Mean :38.85 Mean : 60.56

## 3rd Qu.:150.25 3rd Qu.:49.00 3rd Qu.: 78.00 ## Max. :200.00 Max. :70.00 Max. :137.00

## Spending.Score..1.100. ## Min. : 1.00

## 1st Qu.:34.75

## Median :50.00

## Mean :50.20 ## 3rd Qu.:73.00 ## Max. :99.00

# caluculating standard deviation sd(customer\_data$Age)

## [1] 13.96901

sd(customer\_data$Annual.Income..k..) ## [1] 26.26472

sd(customer\_data$Spending.Score..1.100.) ## [1] 25.82352

#visualization of age distribution customer\_data=read.csv("/Mall\_Customers.csv")

hist(customer\_data$Age, col="green",

main="Histogram that Show's Count of Age Class", xlab="Age",

ylab="Frequency", labels=TRUE)

#boxplot boxplot(customer\_data$Age, col="red",

main="Boxplot for Descriptive Analysis of Age")

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.

#gender visualization a=table(customer\_data$Genre)

barplot(a,main="Using BarPlot to display Gender Comparision", ylab="Count",

xlab="Gender", col=rainbow(2), legend=rownames(a)) pct=round(a/sum(a)\*100)

lbs=paste(c("Female","Male")," ",pct,"%",sep=" ") library(plotrix)

pie3D(a,labels=lbs,

main="Pie Chart Depicting Ratio of Female and Male")

#creating a visualizations to analyze the annual income of the customers. customer\_data=read.csv("/Mall\_Customers.csv") summary(customer\_data$Annual.Income..k..)

## Min. 1st Qu. Median Mean 3rd Qu. Max. ## 15.00 41.50 61.50 60.56 78.00 137.00

#ploting a histogram and then proceed to examine this data using a density

hist(customer\_data$Annual.Income..k.., col="#0867cc",

main="Histogram for Annual Income", xlab="Annual Income Class", ylab="Frequency",

labels=TRUE)

#density plot(density(customer\_data$Annual.Income..k..), col="yellow",

main="Density Plot for Annual Income", xlab="Annual Income Class", ylab="Density")

polygon(density(customer\_data$Annual.Income..k..), col="#257b8a")

Analyzing-Spending-Score-of-the-Customers.R customer\_data=read.csv("/Mall\_Customers.csv") summary(customer\_data$Spending.Score..1.100.) ## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 1.00 34.75 50.00 50.20 73.00 99.00

#boxplot boxplot(customer\_data$Spending.Score..1.100., horizontal=TRUE,

col="#7e42f5",

main="BoxPlot for Descriptive Analysis of Spending Score")

#histogram hist(customer\_data$Spending.Score..1.100., main="HistoGram for Spending Score", xlab="Spending Score Class", ylab="Frequency",

col="#8103ab", labels=TRUE)

elbow-method.R:

customer\_data=read.csv("/Mall\_Customers.csv") cust.sc<-scale(customer\_data[,c(4,5)])

#Finding best K for K mean using Elbow Method: wss <- function(data, maxCluster = 10) {

# Initialize within sum of squares

SSw <- (nrow(data) - 1) \* sum(apply(data, 2, var)) SSw <- vector()

for (i in 2:maxCluster) {

SSw[i] <- sum(kmeans(data, centers = i)$withinss)

}

plot(1:maxCluster, SSw, type = "o", xlab = "Number of Clusters", ylab

="Within groups sum of squares", pch=19)

}

set.seed(100) wss(cust.sc)

cust.sc<-scale(customer\_data[,c(4,5)])

#Finding best K for K mean using Elbow Method: wss <- function(data, maxCluster = 10) {

# Initialize within sum of squares

SSw <- (nrow(data) - 1) \* sum(apply(data, 2, var)) SSw <- vector()

for (i in 2:maxCluster) {

SSw[i] <- sum(kmeans(data, centers = i)$withinss)

}

plot(1:maxCluster, SSw, type = "o", xlab = "Number of Clusters", ylab

="Within groups sum of squares", pch=19)

}

set.seed(100) wss(cust.sc)

# loading required packages

library(factoextra) 23

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at <https://goo.gl/ve3WBa>

library(NbClust)

# Elbow method using fviz\_nbclust set.seed(123)

fviz\_nbclust(cust.sc, kmeans, method = "wss") Average Silhouette Method:

With the help of the average silhouette method, we can measure the quality of our clustering operation. With this, we can determine how well within the cluster is the data object. If we obtain a high average silhouette width, it means that we have good clustering. The average silhouette method calculates the mean of silhouette observations for different k values. With the optimal number of k clusters, one can maximize the average silhouette over significant values for k clusters.

Using the silhouette function in the cluster package, we can compute the average silhouette width using the kmean function. Here, the optimal cluster will possess highest average.

library(cluster) library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(NbClust)

fviz\_nbclust(customer\_data[,3:5], kmeans, method = "silhouette")

Gap Statistic Method:

# Gap Static Method library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(NbClust) library(FunCluster)

## Loading required package: Hmisc ## Loading required package: lattice ## Loading required package: survival ## Loading required package: Formula ##

## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base': ##

## format.pval, units

## Loading required package: cluster 26

##

## Attaching package: 'FunCluster'

## The following object is masked from 'package:ggplot2': ##

## annotate set.seed(123)

stat\_gap <- clusGap (customer\_data[,3:5], FUN = kmeans, nstart = 25, K.max = 10, B = 50)

print(stat\_gap, method = "firstmax")

## Clustering Gap statistic ["clusGap"] from call:

## clusGap(x = customer\_data[, 3:5], FUNcluster = kmeans, K.max = 10, B = 50, nstart = 25)

## B=50 simulated reference sets, k = 1..10; spaceH0="scaledPCA" ## --> Number of clusters (method 'firstmax'): 1

## logW E.logW gap SE.sim

## [1,] 7.829990 8.269156 0.4391660 0.02019110

## [2,] 7.625794 8.028688 0.4028939 0.01819356

## [3,] 7.417921 7.833125 0.4152046 0.01830737

## [4,] 7.256540 7.682002 0.4254622 0.02100410

## [5,] 7.104745 7.602685 0.4979402 0.01822540

## [6,] 6.965334 7.533818 0.5684832 0.01682332

## [7,] 6.903828 7.471733 0.5679041 0.01623561

## [8,] 6.847482 7.417950 0.5704682 0.01694642

## [9,] 6.803046 7.370316 0.5672698 0.01744320

## [10,] 6.749955 7.324903 0.5749475 0.01638668

fviz\_gap\_stat(stat\_gap)

k6<-

kmeans(customer\_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd") k6

## 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 2 6 2 6 2 6

2

## [38] 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 1 5 1 1 5 5 1 1 1

1

28

## [75] 1 5 1 5 5 1 1 5 1 1 5 1 1 5 5 1 1 5 1 5 5 5 1 5 1 5 5 1 1 5 1 5 1 1 1 1

1

## [112] 5 5 5 5 5 1 1 1 1 5 5 5 4 5 4 3 4 3 4 3 4 5 4 3 4 3 4 3 4 3 4 5 4 3 4 3

4

## [149] 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 3 4 3 4

3

## [186] 4 3 4 3 4 3 4 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" "tot.withinss" ## [6] "betweenss" "size" "iter" "ifault"

Visualizing the Clustering Results using the First Two Principle Components

library(cluster) library(gridExtra) library(grid) library(ggplot2) 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")

kCols=function(vec){cols=rainbow (length (unique (vec))) return (cols[as.numeric(as.factor(vec))])}

digCluster<-k6$cluster; dignm<-as.character(digCluster); # K-means clusters

plot(pcclust$x[,1:2], col =kCols(digCluster),pch =19,xlab ="K- means",ylab="classes")

legend("bottomleft",unique(dignm),fill=unique(kCols(digCluster)))