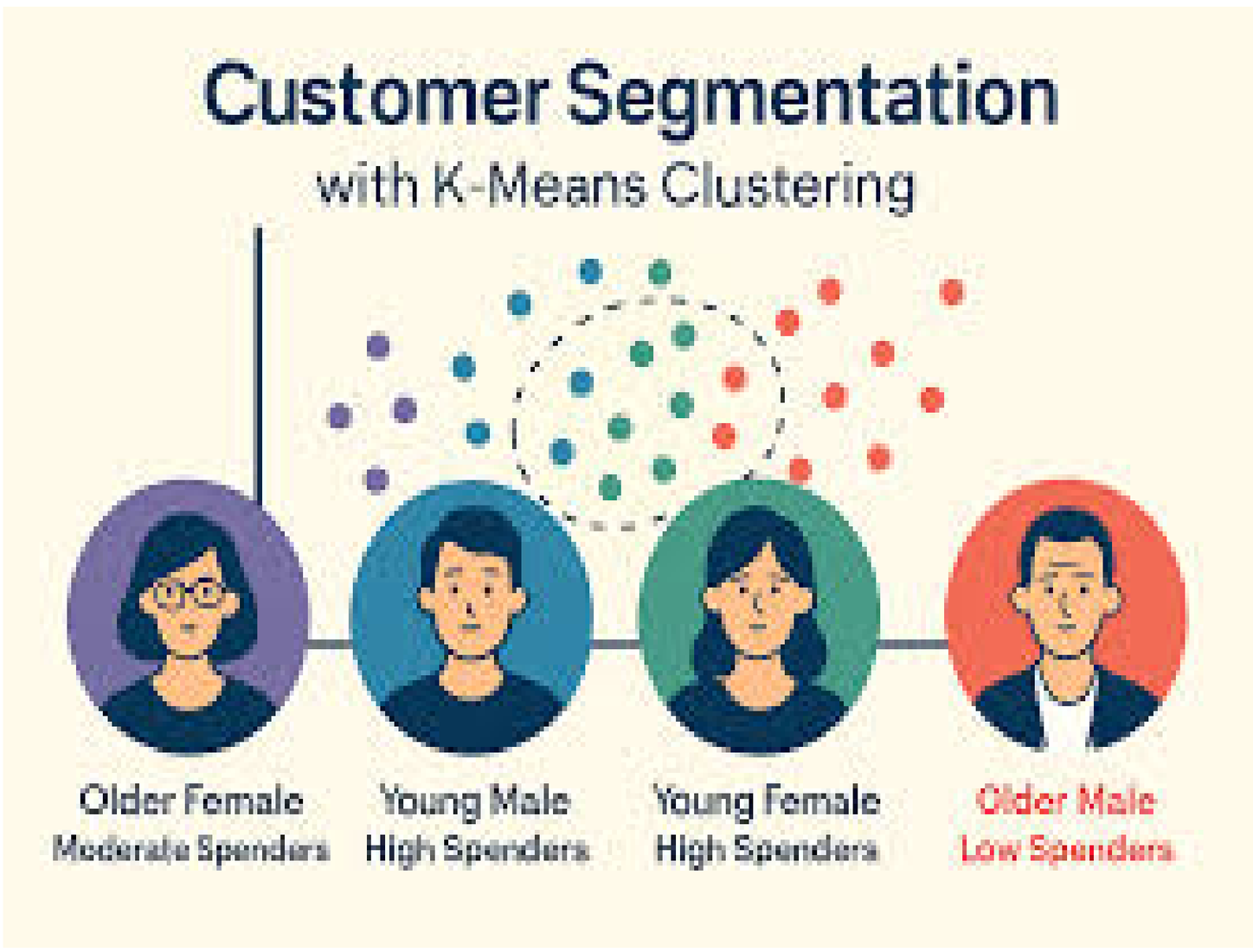
ASSIGNMENT--4

AIM: Customer Segmentation k-means&PCA

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 machine learning project, we will make use of K-means clustering which is the essential algorithm for clustering unlabeled dataset. Before ahead in this project, learn what actually customer segmentation is.



What is 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.

What is 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.

Dataset:
The dataset is aquired from kaggle and the link is given below :

https://www.kaggle.com/nelakurthisudheer/mall-customer-segmentation

The dataset consists of following five features of 200 customers:

CustomerID: Unique ID assigned to the customer

Gender: Gender of the customer

Age: Age of the customer

Annual Income (k$): Annual Income of the customer

Spending Score (1-100): Score assigned by the mall based on customer behavior and spending nature.

Steps for implementation
Import all neccessary packages
import ----- from ------
import -----
Data Exploration
customer_data=read.csv("/home/dataflair/Mall_Customers.csv")

```
str(customer_data)
names(customer_data)

head(customer_data)
summary(customer_data$Age)
Statistical Analysis
sd(customer_data$Age)
summary(customer_data$Annual.Income..k..)
sd(customer_data$Annual.Income..k..)
summary(customer_data$Age)
Visualizations
Bar Plot

a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))


Pie Chart
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")
Histogram
hist(customer_data$Age,
      col="blue",
      main="Histogram to Show Count of Age Class",
      xlab="Age Class",
      ylab="Frequency",
      labels=TRUE)
Box Plot
boxplot(customer_data$Age,
        col="ff0066",
        main="Boxplot for Descriptive Analysis of Age")
Analysis
Analyzing the annual income of the customers through the Histogram
summary(customer_data$Annual.Income..k..)
hist(customer_data$Annual.Income..k..,
   col="#660033",
   main="Histogram for Annual Income",
```

```
    xlab="Annual Income Class",
    ylab="Frequency",
    labels=TRUE)
Density Plot
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="#ccff66")
Analyzing Spending Score of the Customers with the help of BoxPlot
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(customer_data$Spending.Score..1.100.,
      horizontal=TRUE,
      col="#990000",
      main="BoxPlot for Descriptive Analysis of Spending Score")
```

K means Algorithm :

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.

Determining Optimal Clusters

While working with clusters, you need to specify the number of clusters to use. You would like to utilize the optimal number of clusters. To help you in determining the optimal clusters, there are three popular methods –

Elbow method The main goal behind cluster partitioning methods like k-means is to define the clusters such that the intra-cluster variation stays minimum.

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

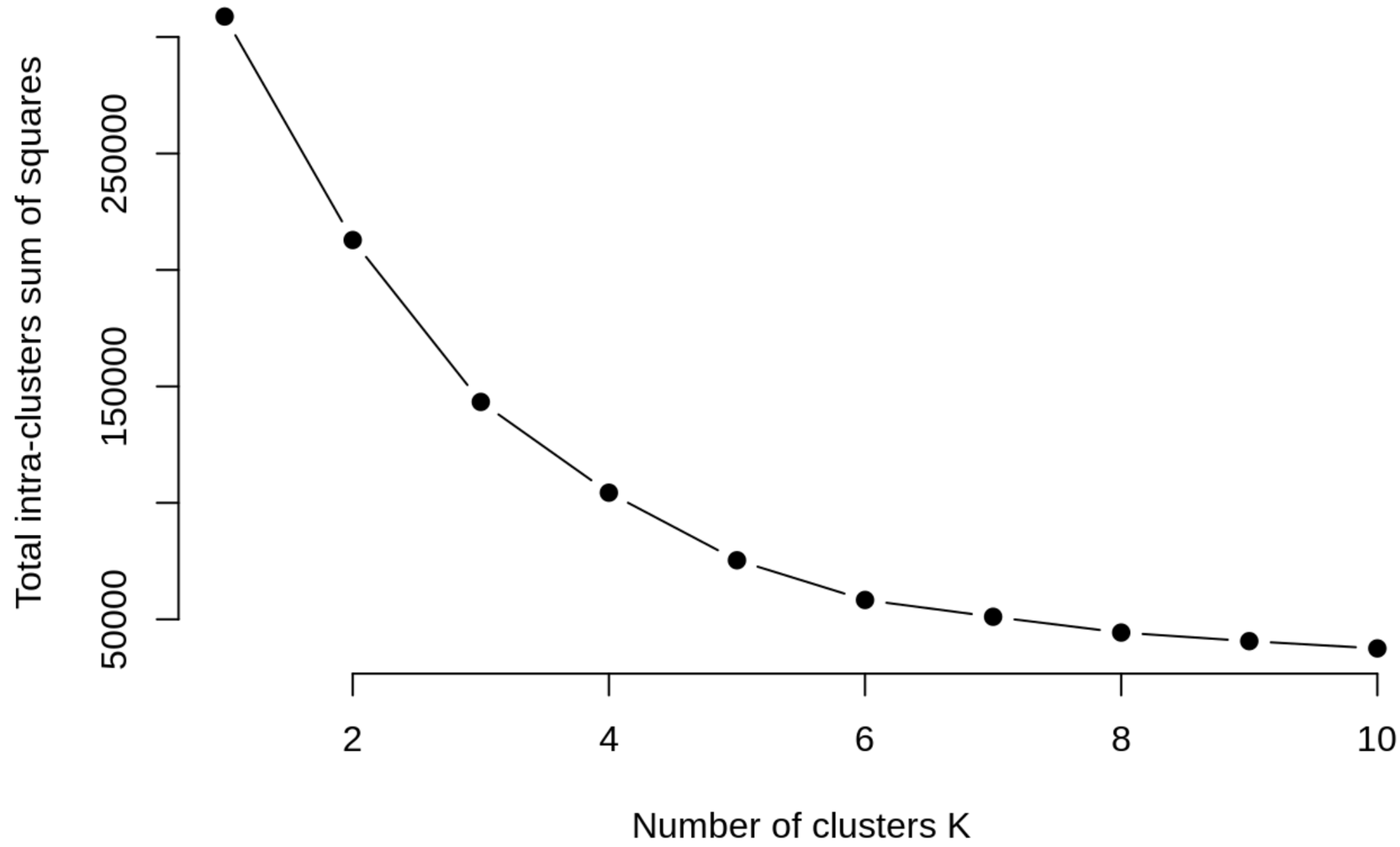
Where Ck represents the kth cluster and W(Ck) denotes the intra-cluster variation. With the measurement of the total intra-cluster variation, one can evaluate the compactness of the clustering boundary. We can then proceed to define the optimal clusters as follows –

First, 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.

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

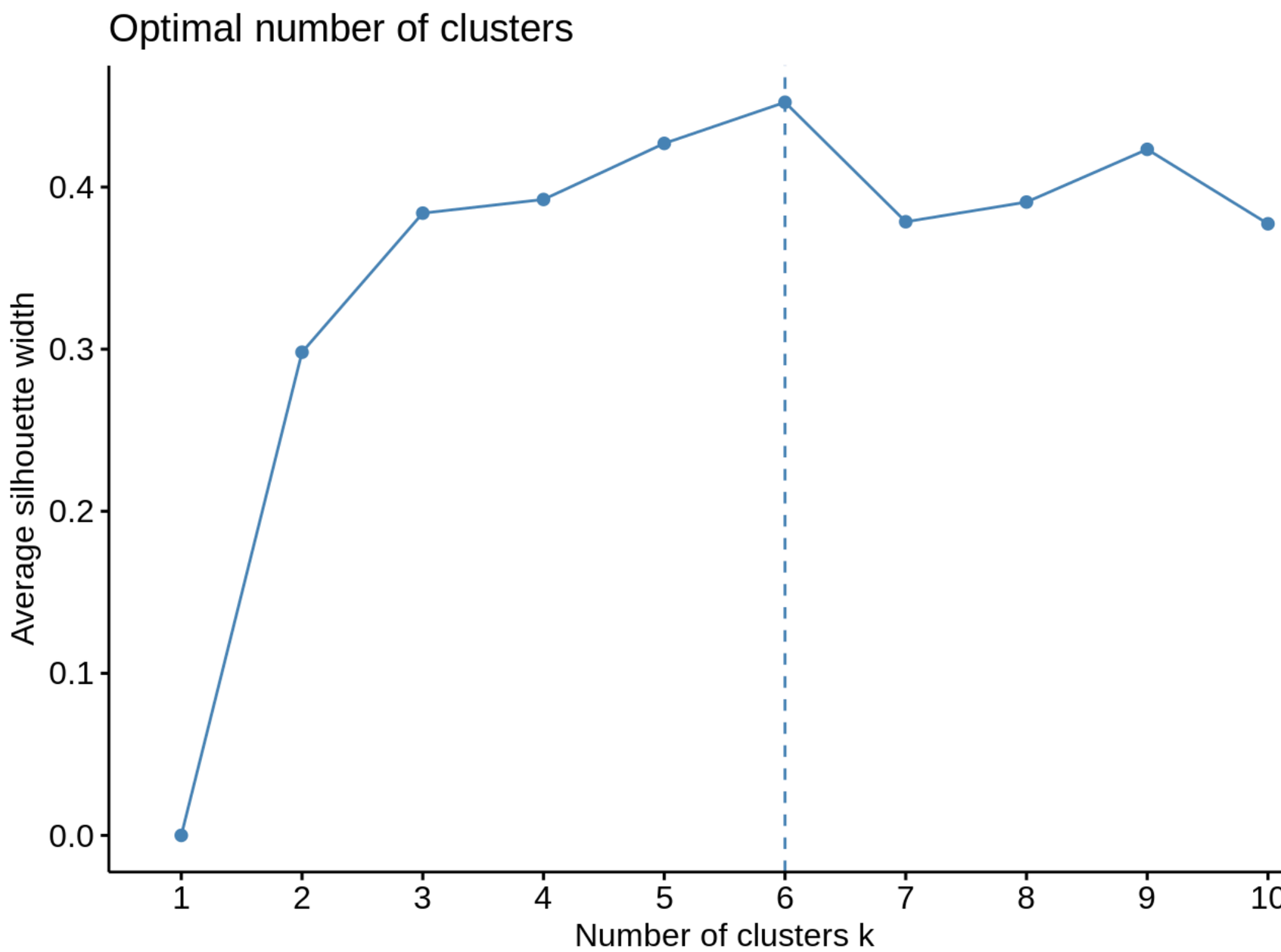


from the above graph, we conclude that 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot.

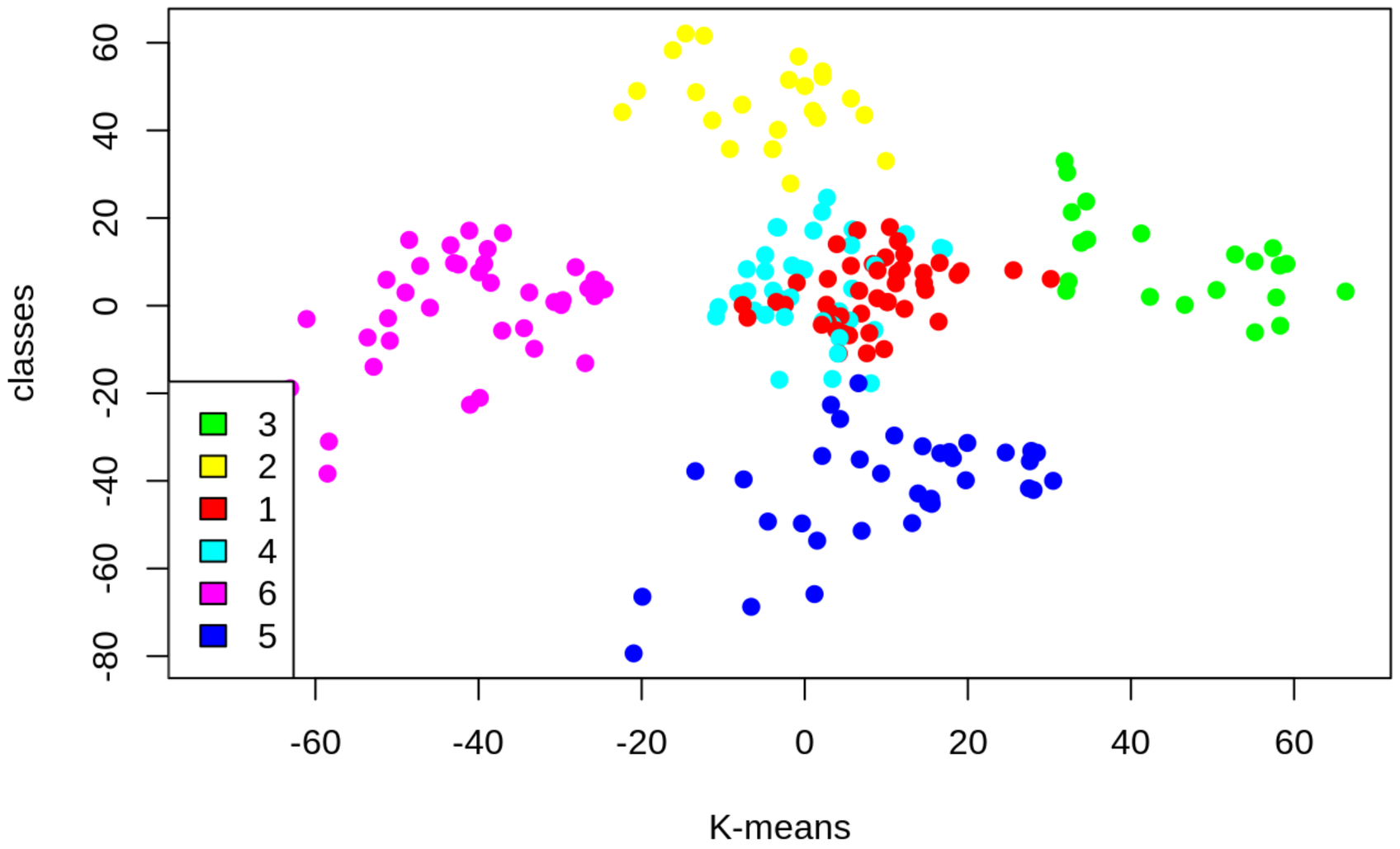
Average Silhouette method With the help of the average silhouette method, we can measure the quality of our clustering operation. With this, we can determine how well within the cluster is the data object. If we obtain a high average silhouette width, it means that we have good clustering. The average 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(gridExtra)
library(grid)
```

```
k2<-kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
s2<-plot(silhouette(k2$cluster,dist(customer_data[,3:5],"euclidean")))
```



By Using these three methods in k-means clustering we have to find out which is giving the best minimum number optimal clusters.



From the above segemented graph:

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 ----------
This analysis should provide a solid base for discussion with relevant business stakeholders.
Normally I would present my client with a variety of customer profiles based on different combinations of customer features and formulate my own data-driven recommendations.

However, it is ultimately down to them to decide how many groups they want settle for and what characteristics each segment should have.

#Reference
#
https://diegousai.io/2019/09/steps-and-considerations-to-run-a-successful-segmentation/