Customer Segmentation

Subash chander swaminathan

Import Dataset

customer\_df = read.csv("Mall\_Customers.csv")  
str(customer\_df)

## 'data.frame': 200 obs. of 5 variables:  
## $ CustomerID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Gender : chr "Male" "Male" "Female" "Female" ...  
## $ Age : int 19 21 20 23 31 22 35 23 64 30 ...  
## $ 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\_df)

## [1] "CustomerID" "Gender" "Age"   
## [4] "Annual.Income..k.." "Spending.Score..1.100."

head(customer\_df)

## CustomerID Gender 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

Customers Age

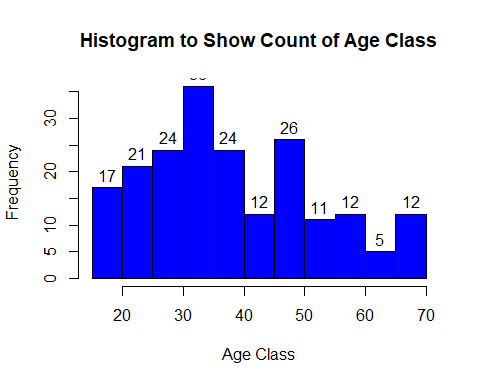
summary(customer\_df$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 28.75 36.00 38.85 49.00 70.00

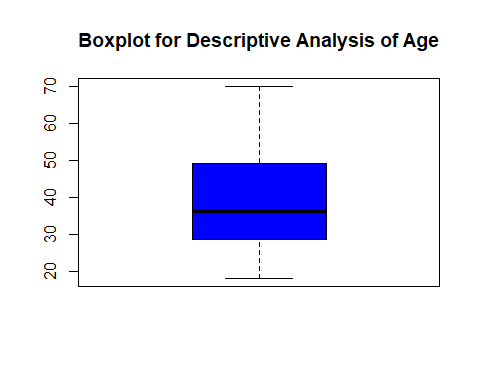
sd(customer\_df$Age)

## [1] 13.96901

hist(customer\_df$Age,  
 col="blue",  
 main="Histogram to Show Count of Age Class",  
 xlab="Age Class",  
 ylab="Frequency",  
 labels=TRUE)



boxplot(customer\_df$Age,  
 col="blue",  
 main="Boxplot for Descriptive Analysis of Age")



Customers Annual Income

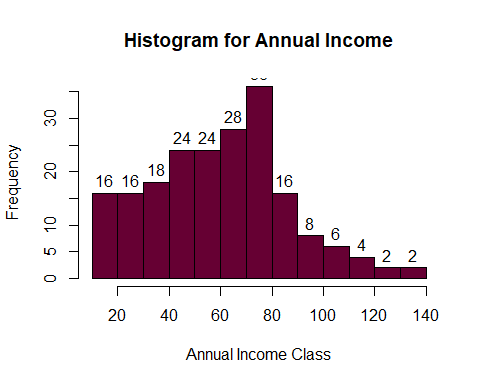
summary(customer\_df$Annual.Income..k..)

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

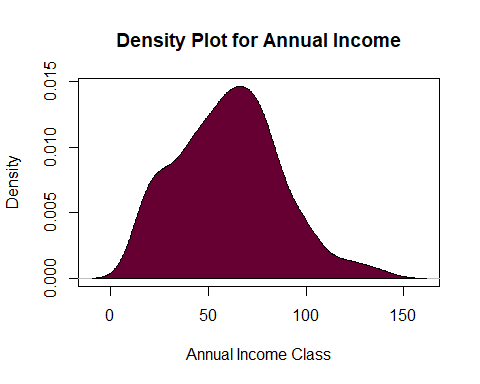
sd(customer\_df$Annual.Income..k..)

## [1] 26.26472

hist(customer\_df$Annual.Income..k..,  
 col="#660033",  
 main="Histogram for Annual Income",  
 xlab="Annual Income Class",  
 ylab="Frequency",  
 labels=TRUE)



plot(density(customer\_df$Annual.Income..k..),  
 col="#660033",  
 main="Density Plot for Annual Income",  
 xlab="Annual Income Class",  
 ylab="Density")  
polygon(density(customer\_df$Annual.Income..k..),  
 col="#660033")



Customers spending score

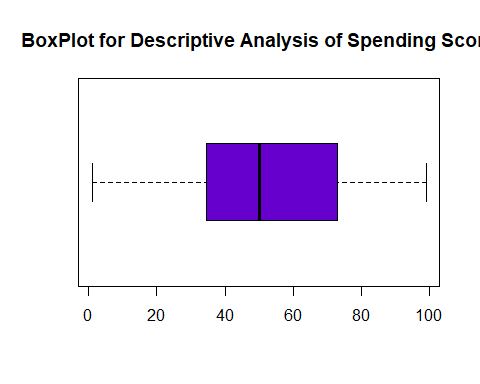
summary(customer\_df$Spending.Score..1.100.)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 34.75 50.00 50.20 73.00 99.00

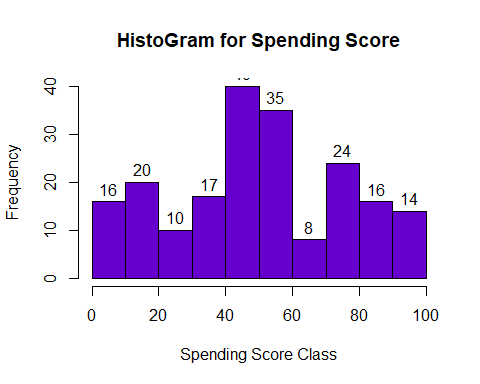
sd(customer\_df$Spending.Score..1.100.)

## [1] 25.82352

boxplot(customer\_df$Spending.Score..1.100.,  
 horizontal=TRUE,  
 col="#6600cc",  
 main="BoxPlot for Descriptive Analysis of Spending Score")

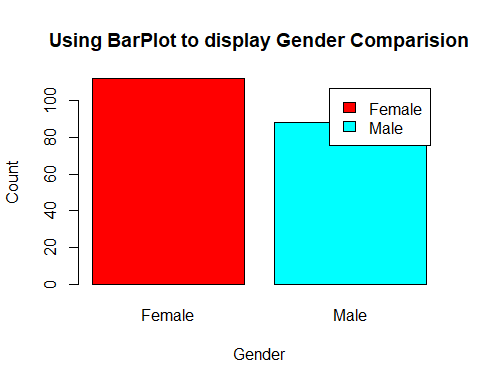


hist(customer\_df$Spending.Score..1.100.,  
 main="HistoGram for Spending Score",  
 xlab="Spending Score Class",  
 ylab="Frequency",  
 col="#6600cc",  
 labels=TRUE)



Number of Customer based on Gender

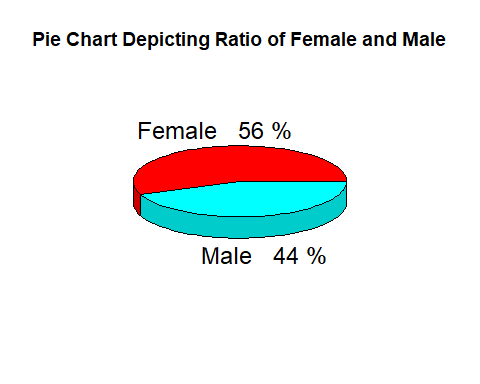
a=table(customer\_df$Gender)  
barplot(a,main="Using BarPlot to display Gender Comparision",  
 ylab="Count",  
 xlab="Gender",  
 col=rainbow(2),  
 legend=rownames(a)  
)



pct=round(a/sum(a)\*100)  
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")  
library(plotrix)

## Warning: package 'plotrix' was built under R version 4.0.3

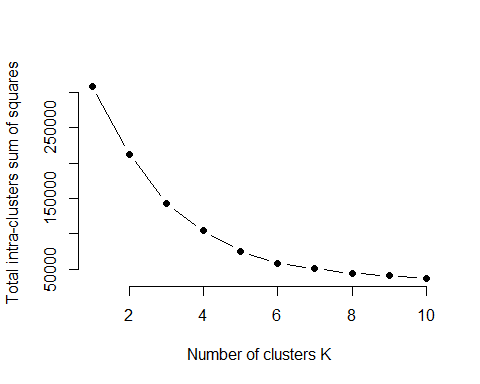
pie3D(a,labels=lbs,  
 main="Pie Chart Depicting Ratio of Female and Male")

 Cluster model to segment customers

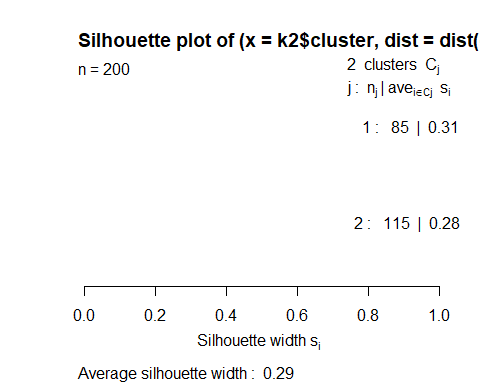
library(purrr)  
set.seed(1234)

Function to calculate total intra-cluster sum of square

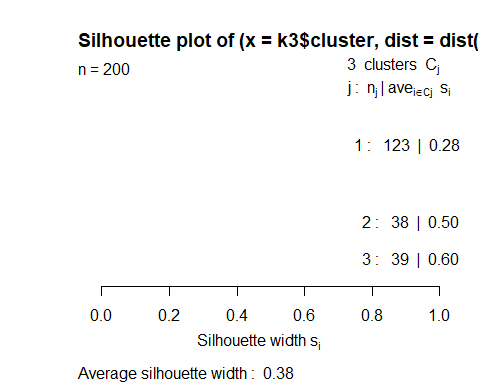
iss <- function(k) {  
 kmeans(customer\_df[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withinss  
}  
k.values <- 1:10  
  
iss\_values <- map\_dbl(k.values, iss)  
  
plot(k.values, iss\_values,  
 type="b", pch = 19, frame = FALSE,   
 xlab="Number of clusters K",  
 ylab="Total intra-clusters sum of squares")

 6 cluster combination produced better result that others

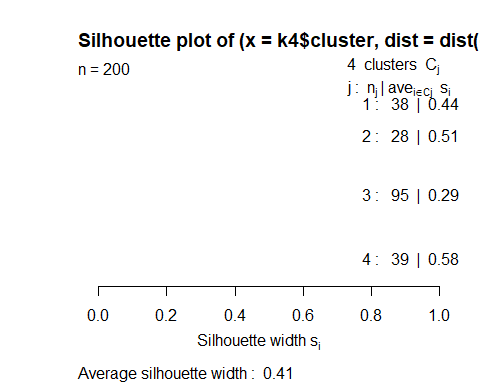
library(cluster)   
library(gridExtra)  
library(grid)  
  
  
k2<-kmeans(customer\_df[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")  
s2<-plot(silhouette(k2$cluster,dist(customer\_df[,3:5],"euclidean")))



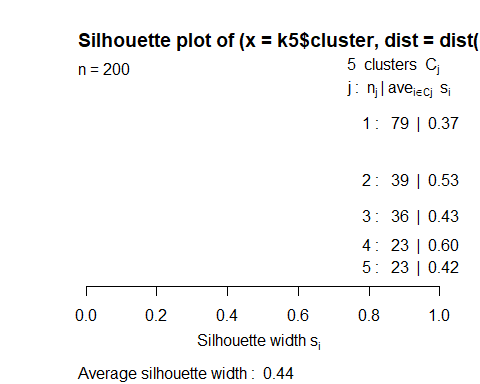
k3<-kmeans(customer\_df[,3:5],3,iter.max=100,nstart=50,algorithm="Lloyd")  
s3<-plot(silhouette(k3$cluster,dist(customer\_df[,3:5],"euclidean")))



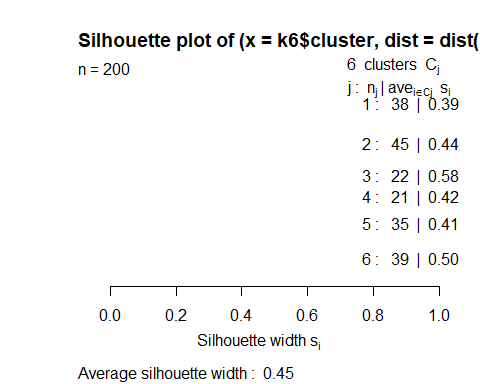
k4<-kmeans(customer\_df[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")  
s4<-plot(silhouette(k4$cluster,dist(customer\_df[,3:5],"euclidean")))



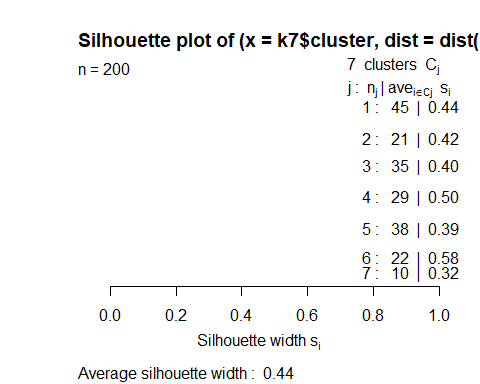
k5<-kmeans(customer\_df[,3:5],5,iter.max=100,nstart=50,algorithm="Lloyd")  
s5<-plot(silhouette(k5$cluster,dist(customer\_df[,3:5],"euclidean")))



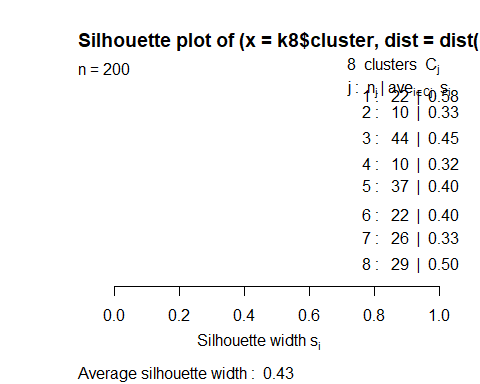
k6<-kmeans(customer\_df[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")  
s6<-plot(silhouette(k6$cluster,dist(customer\_df[,3:5],"euclidean")))



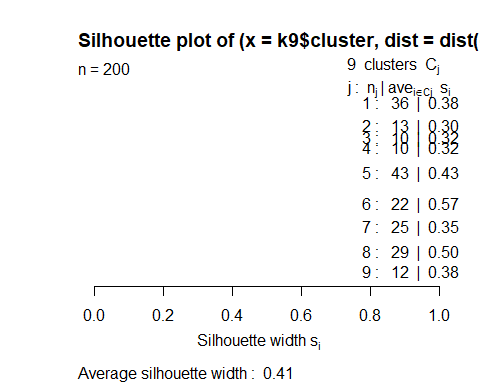
k7<-kmeans(customer\_df[,3:5],7,iter.max=100,nstart=50,algorithm="Lloyd")  
s7<-plot(silhouette(k7$cluster,dist(customer\_df[,3:5],"euclidean")))



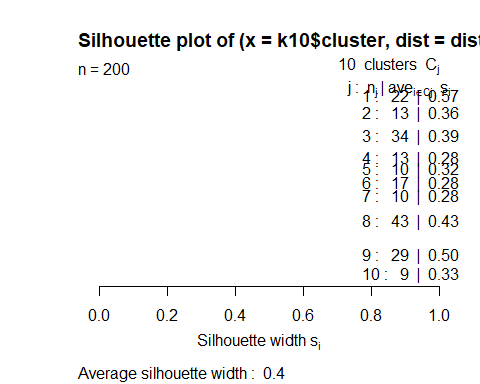
k8<-kmeans(customer\_df[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")  
s8<-plot(silhouette(k8$cluster,dist(customer\_df[,3:5],"euclidean")))



k9<-kmeans(customer\_df[,3:5],9,iter.max=100,nstart=50,algorithm="Lloyd")  
s9<-plot(silhouette(k9$cluster,dist(customer\_df[,3:5],"euclidean")))



k10<-kmeans(customer\_df[,3:5],10,iter.max=100,nstart=50,algorithm="Lloyd")  
s10<-plot(silhouette(k10$cluster,dist(customer\_df[,3:5],"euclidean")))

 used fviz\_nbclust() function to determine and visualize the optimal number of clusters

library(NbClust)

## Warning: package 'NbClust' was built under R version 4.0.3

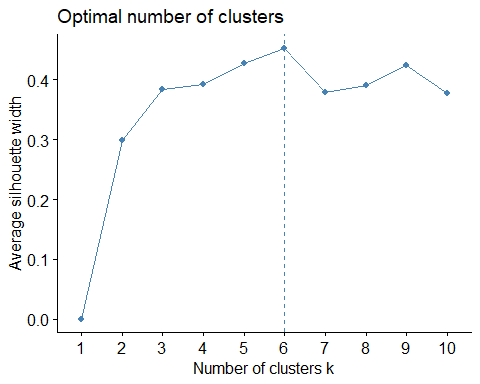
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.0.3

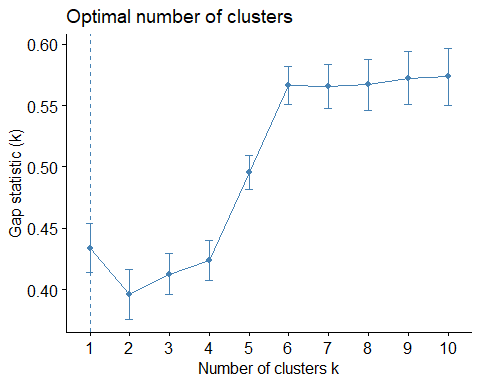
## Loading required package: ggplot2

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

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

 Used the clusGap() function for providing gap statistic as well as standard error

set.seed(1234)  
stat\_gap <- clusGap(customer\_df[,3:5], FUN = kmeans, nstart = 25,  
 K.max = 10, B = 50)  
fviz\_gap\_stat(stat\_gap)



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

## K-means clustering with 6 clusters of sizes 45, 22, 21, 38, 35, 39  
##   
## Cluster means:  
## Age Annual.Income..k.. Spending.Score..1.100.  
## 1 56.15556 53.37778 49.08889  
## 2 25.27273 25.72727 79.36364  
## 3 44.14286 25.14286 19.52381  
## 4 27.00000 56.65789 49.13158  
## 5 41.68571 88.22857 17.28571  
## 6 32.69231 86.53846 82.12821  
##   
## Clustering vector:  
## [1] 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3  
## [38] 2 3 2 1 2 1 4 3 2 1 4 4 4 1 4 4 1 1 1 1 1 4 1 1 4 1 1 1 4 1 1 4 4 1 1 1 1  
## [75] 1 4 1 4 4 1 1 4 1 1 4 1 1 4 4 1 1 4 1 4 4 4 1 4 1 4 4 1 1 4 1 4 1 1 1 1 1  
## [112] 4 4 4 4 4 1 1 1 1 4 4 4 6 4 6 5 6 5 6 5 6 4 6 5 6 5 6 5 6 5 6 4 6 5 6 5 6  
## [149] 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5  
## [186] 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6  
##   
## Within cluster sum of squares by cluster:  
## [1] 8062.133 4099.818 7732.381 7742.895 16690.857 13972.359  
## (between\_SS / total\_SS = 81.1 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

Principal component analysis

pcclust=prcomp(customer\_df[,3:5],scale=FALSE)  
summary(pcclust)

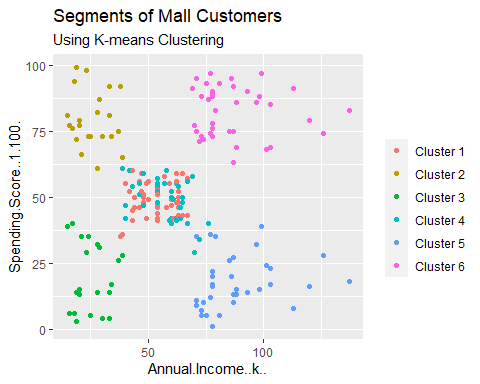
## Importance of components:  
## PC1 PC2 PC3  
## Standard deviation 26.4625 26.1597 12.9317  
## Proportion of Variance 0.4512 0.4410 0.1078  
## Cumulative Proportion 0.4512 0.8922 1.0000

pcclust$rotation[,1:2]

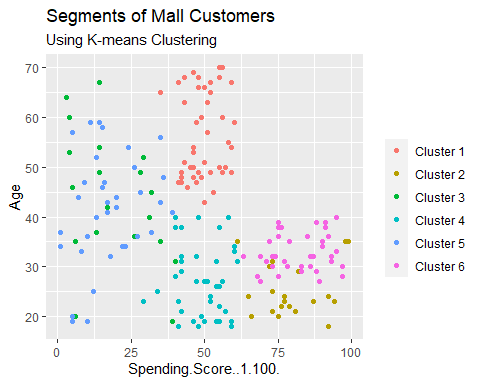
## PC1 PC2  
## Age 0.1889742 -0.1309652  
## Annual.Income..k.. -0.5886410 -0.8083757  
## Spending.Score..1.100. -0.7859965 0.5739136

Visuvalized the clusters

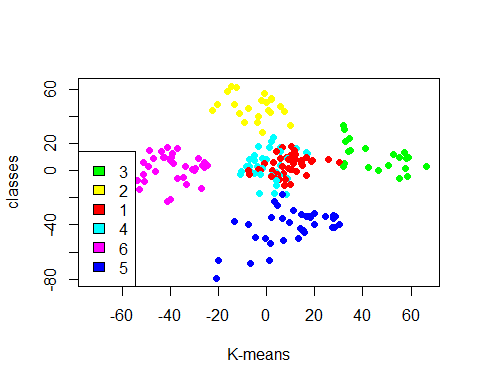
set.seed(1234)  
ggplot(customer\_df, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +   
 geom\_point(stat = "identity", aes(color = as.factor(k6$cluster))) +  
 scale\_color\_discrete(name=" ",  
 breaks=c("1", "2", "3", "4", "5","6"),  
 labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5","Cluster 6")) +  
 ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")



ggplot(customer\_df, aes(x =Spending.Score..1.100., y =Age)) +   
 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);   
  
plot(pcclust$x[,1:2], col =kCols(digCluster),pch =19,xlab ="K-means",ylab="classes")  
legend("bottomleft",unique(dignm),fill=unique(kCols(digCluster)))

 Cluster 1 and 4 – 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.