## Data Science Project – Customer Segmentation using Machine Learning in R

customer\_data <- read.csv("Mall\_Customers.csv")

str(customer\_data)

names(customer\_data)

head(customer\_data)

summary(customer\_data$Age)

sd(customer\_data$Age)

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

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

summary(customer\_data$Age)

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

a=table(customer\_data$Genre)

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

ylab="Count",

xlab="Genre",

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")

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)

boxplot(customer\_data$Age,

col="orange",

main="Boxplot for Descriptive Analysis of Age")

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)

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")

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")

hist(customer\_data$Spending.Score..1.100.,

main="HistoGram for Spending Score",

xlab="Spending Score Class",

ylab="Frequency",

col="#6600cc",

labels=TRUE)

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")

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")))

k3<-kmeans(customer\_data[,3:5],3,iter.max=100,nstart=50,algorithm="Lloyd")

s3<-plot(silhouette(k3$cluster,dist(customer\_data[,3:5],"euclidean")))

k4<-kmeans(customer\_data[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")

s4<-plot(silhouette(k4$cluster,dist(customer\_data[,3:5],"euclidean")))

k5<-kmeans(customer\_data[,3:5],5,iter.max=100,nstart=50,algorithm="Lloyd")

s5<-plot(silhouette(k5$cluster,dist(customer\_data[,3:5],"euclidean")))

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

s6<-plot(silhouette(k6$cluster,dist(customer\_data[,3:5],"euclidean")))

k7<-kmeans(customer\_data[,3:5],7,iter.max=100,nstart=50,algorithm="Lloyd")

s7<-plot(silhouette(k7$cluster,dist(customer\_data[,3:5],"euclidean")))

k8<-kmeans(customer\_data[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")

s8<-plot(silhouette(k8$cluster,dist(customer\_data[,3:5],"euclidean")))

k9<-kmeans(customer\_data[,3:5],9,iter.max=100,nstart=50,algorithm="Lloyd")

s9<-plot(silhouette(k9$cluster,dist(customer\_data[,3:5],"euclidean")))

k10<-kmeans(customer\_data[,3:5],10,iter.max=100,nstart=50,algorithm="Lloyd")

s10<-plot(silhouette(k10$cluster,dist(customer\_data[,3:5],"euclidean")))

library(NbClust)

library(factoextra)

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

set.seed(125)

stat\_gap <- clusGap(customer\_data[,3:5], FUN = kmeans, nstart = 25,

K.max = 10, B = 50)

fviz\_gap\_stat(stat\_gap)

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

k6

pcclust=prcomp(customer\_data[,3:5],scale=FALSE) #principal component analysis

summary(pcclust)

pcclust$rotation[,1:2]

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")

ggplot(customer\_data, 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); # 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)))