CUSTOMER SEGMENTATION IN R

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.

Dataset: CSV Format

https://drive.google.com/file/d/17XAg12DTD4XxN9ORwjH6I6wXaocf5qk5/view?usp=sharing

Implementation:

#Importing data set from csv file

```
customer_data=read.csv("Mall_Customers.csv")

str(customer_data)

names(customer_data)

head(customer_data)

summary(customer_data)

#sd of all names
sd(customer_data$Age)
summary(customer_data$Annual.Income..k..)
sd(customer_data$Annual.Income..k..)
summary(customer_data$Age)

##bar plot visualization for column names

a=table(customer_data$Gender)
```

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

##conclusion = no of females are higher than males

##pie chart to observer the ratio of male and female

```
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
install.packages("plotrix")
library(plotrix)
pie3D(a,labels=lbs,main="Pie Chart Depicting Ratio of Female and Male")
```

#conlcusion that percentage of female is 56% whereas male is as 44 %

#now vizulization of age of the people

```
summary(customer_data$Age)
#plotting a histogram
hist(customer_data$Age,
    col="blue",
    main="Histogram to Show Count of Age Class",
    xlab="Age Class",
    ylab="Frequency",
    labels=TRUE)
```

#boxplot

```
boxplot(customer_data$Age,
       col="blue",
       main="Boxplot for Descriptive Analysis of Age")
```

#col=max age is b/w 30 and 35 the min age is 18 and max is 70

#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")
```

con=min income is 15 and max is 137 averaage income of 70

##spending score of the customes

```
summary(customer_data$Spending.Score..1.100.)


boxplot(customer_data$Spending.Score..1.100.,
      horizontal=TRUE,
      col="red",
      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="#6600cc",
    labels=TRUE)
```

##applying k mean clustring algorithm

#for k using the elbo method

```
install.packages("tidyverse")
library(purrr)
set.seed(123)
```

 # function to calculate total intra-cluster sum of square

```
customer_data
customer_data[,3:5]

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")
```

##Average Silhouette Method

```
library(cluster)
install.packages("gridExtra")
library(gridExtra)
library(grid)
```

 # when k=2

```
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")))
```

 # when k=3

```
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")))
```

 # when k=4

```
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")))
```

#when k=5

```
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")))
```

 # when k=6

```
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")))
```

 # 7when k=

```
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")))
```

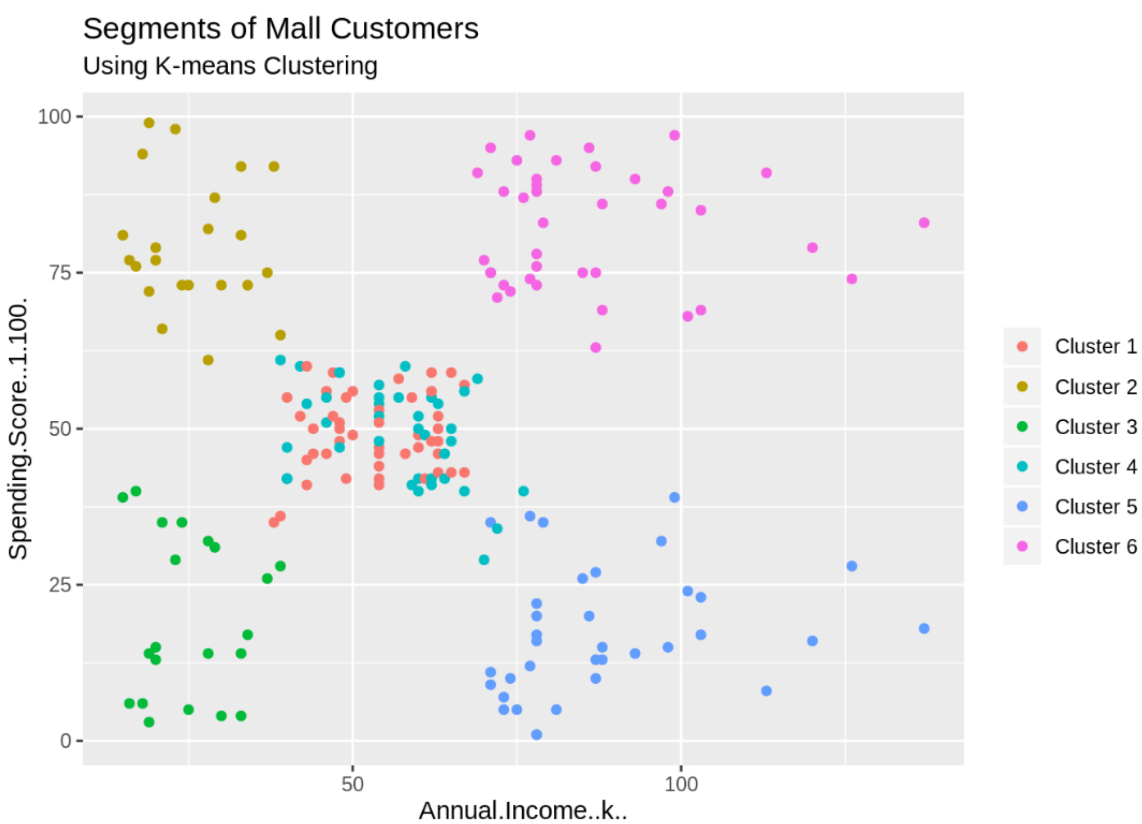
 # when k=

```
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")))
```

```
# now vizuaising the cluster
install.packages("NbClust")
library(NbClust)
install.packages("colorspace")
install.packages("factoextra")
library(factoextra)

fviz_nbclust(customer_data[,3:5], kmeans, method = "silhouette")



##Gap Statistic Method
set.seed(125)
stat_gap <- clusGap(customer_data[,3:5], FUN = kmeans, nstart = 25,
                K.max = 10, B = 50)
fviz_gap_stat(stat_gap)


######now taking 6 as optimal clusters

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


#Visualizing the Clustering Results using the First Two Principle Components
pcclust=prcomp(customer_data[,3:5],scale=FALSE) #principal component analysis
summary(pcclust)

pcclust$rotation[,1:2]

## annual income and spending score

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")


#spending score adn age
```

```
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)))
```

<u>Spendings vs Annual Income</u>



OBSERVATIONS

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.*

Conclusion

implemented the customer segmentation model using a class of machine learning known as unsupervised learning. Specifically, made use of a clustering algorithm called K-means clustering and analysed and visualized the data and then proceeded to implement our algorithm