---
title: "Customer Segmentation"
author: "Aditi Teriar"
date: "10/31/2020"
output: pdf_document

```{r setup, include=FALSE}

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

```
      legend=rownames(a))

install.packages(plotrix)

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="ff0066",
      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)))



knitr::opts_chunk$set(echo = TRUE)
```
```
```