library(ggplot2)

library(factoextra)

library(plotly)

#Reading Input data ---> DataTable.csv file from a local directoy

DataTable <- read.csv("D://dc++//workspace 2//DataTable.csv",header = T,sep = ",")

#Date is in factor variable and others are in int variables so converting them

DataTable$Date <- as.Date(DataTable$Date, format = "%m/%d/%y")

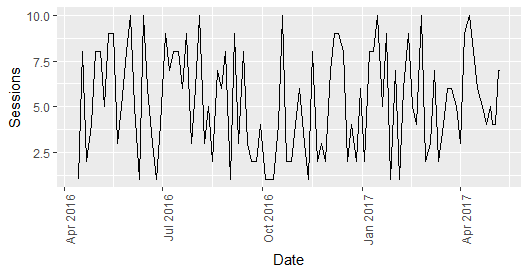
DataTable$Sessions <- as.numeric(DataTable$Sessions)

DataTable$Page.Views <- as.numeric(DataTable$Page.Views)

DataTable$Avg..Session.Duration <- as.numeric(DataTable$Avg..Session.Duration)

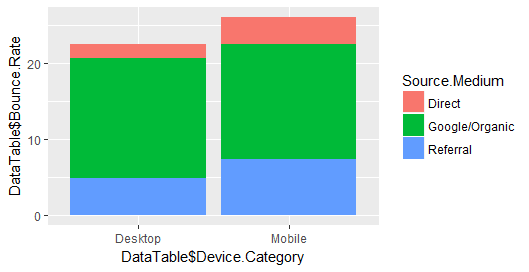
#line plot to check variations of sessions with respect to date

ggplot(DataTable,aes(x=Date,y=Sessions,group=1))+geom\_line()+theme(axis.text.x = element\_text(angle = 90, hjust = 1))



#how bounce rate is dependent varies for device category with respect to source media

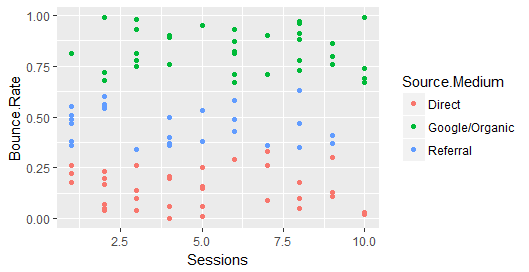
ggplot(data = DataTable,aes(x = DataTable$Device.Category,y=DataTable$Bounce.Rate))+geom\_bar(aes(fill=Source.Medium),stat = "identity")



#Shows that organic/google searches has highest bounce rate and direct visit has less bounce rate

#to differentiate no of clusters by viewing it

ggplot(DataTable,aes(Sessions,Bounce.Rate))+geom\_point(aes(col=Source.Medium))



#we can see three clusters by eye visualization but just to make sure we are running elbow algorithm to find no of clusters

#using elbow method to get no of clusters

#elbow method maynnot always give the correct no of clusters 'Beware!!'

mydata <- scale(DataTable[,c(8,12)])

wss[1] <- (nrow(mydata)-1)\*sum(apply(mydata,2,var))

#apply(x,1,...)means rows

#apply(x,2,...)means columns

#sum up the variance of each row and mutiply (nrow(mydata)-1)

for (i in 2:15) {

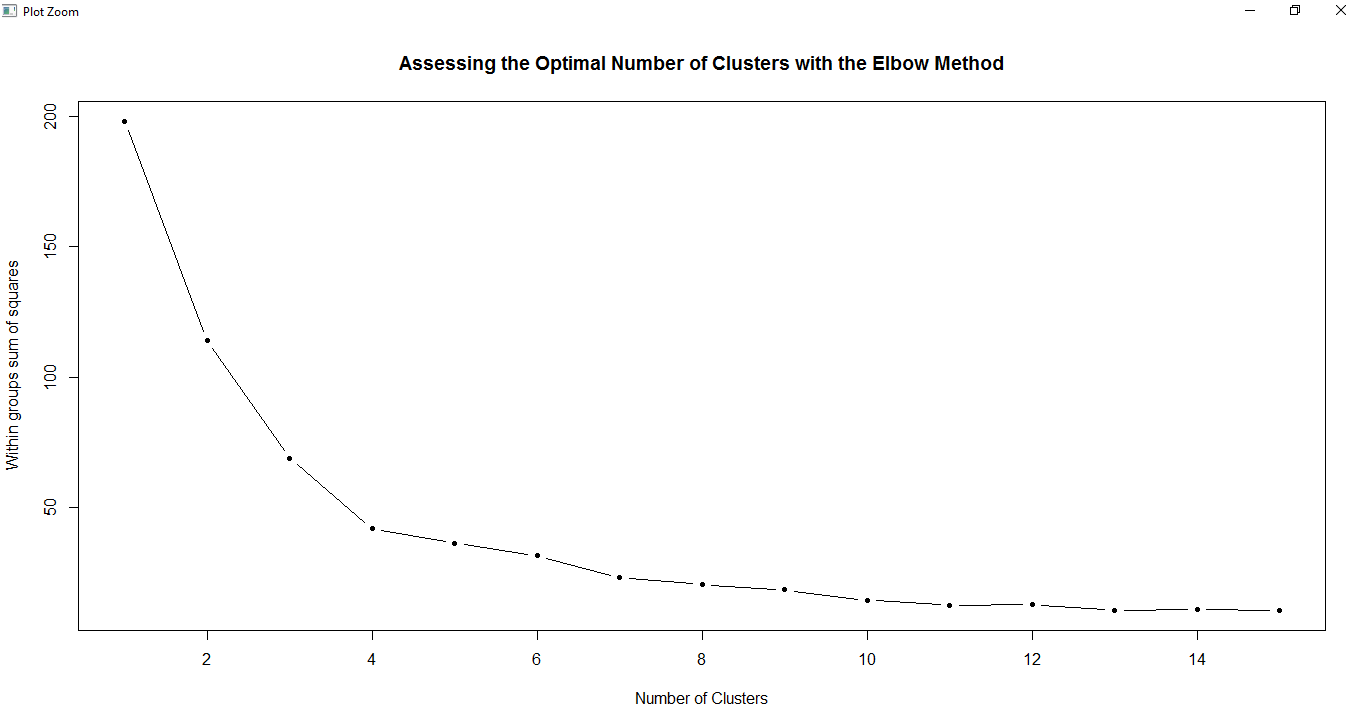
wss[i] <- sum(kmeans(mydata,centers=i)$withinss)

}

#calculate the SSE of each clustering

plot(x = 1:15, y = wss, type="b", xlab="Number of Clusters",

ylab="Within groups sum of squares",

 main="Assessing the Optimal Number of Clusters with the Elbow Method",pch=20)

#kmeans clustering performed on sessions and bounce rate to have clusters of Sourcemedia

Kcluster <- kmeans(scale(DataTable[,c(8,12)]), 3, nstart = 1)

table(Kcluster$cluster,DataTable$Source.Medium)

#output of table

# Direct Google/Organic Referral

# 1 19 0 9

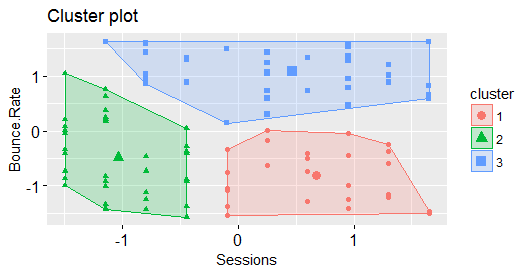
# 2 17 3 15

# 3 0 34 3

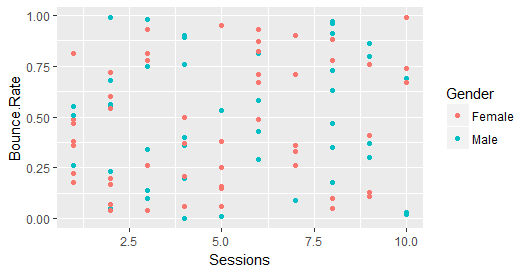
#visualizing the clusters

fviz\_cluster(Kcluster,data=(DataTable[,c(8,12)]),geom="point")

#the previous point plot now divided into 3 clusters ☺



ggplot(DataTable,aes(Sessions,Bounce.Rate))+geom\_point(aes(col=Gender))

#plot shows that no clustering for Gender wrt to sessions and bouncerate

#Similarly there is no pattern or corelation for sessions and (pageviews,bouncerate) for (gender,Customer.Type,country)

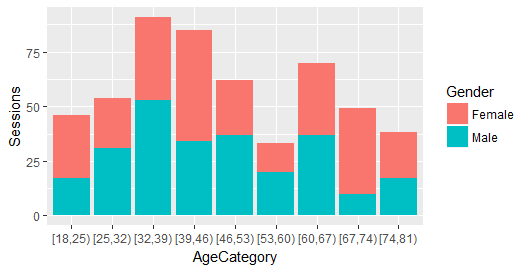
#catgorical Age and their interests

#Making integeral age to categorical age

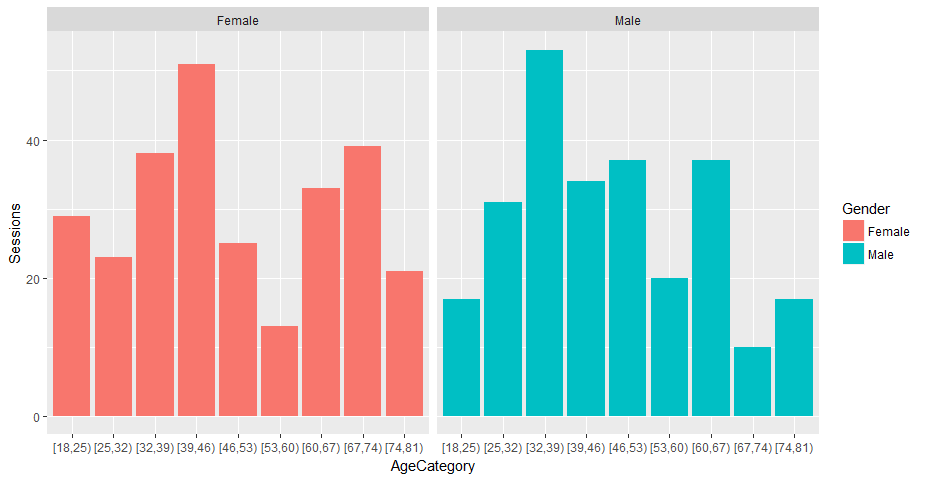
AgeCategory <- cut(DataTable$Age,breaks = c(18,25,32,39,46,53,60,67,74,81),right = F)

newDataTable <- data.frame(DataTable,AgeCategory)

ggplot(data = newDataTable,aes(AgeCategory,Sessions))+geom\_bar(stat="identity",aes(fill=Gender))

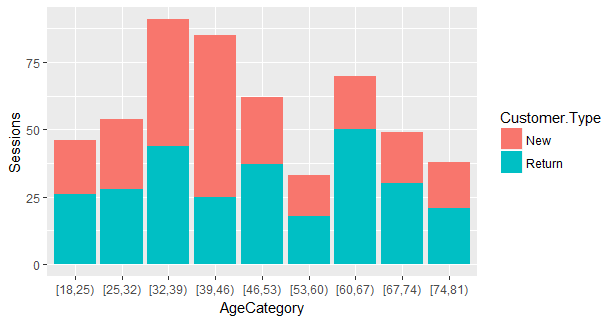


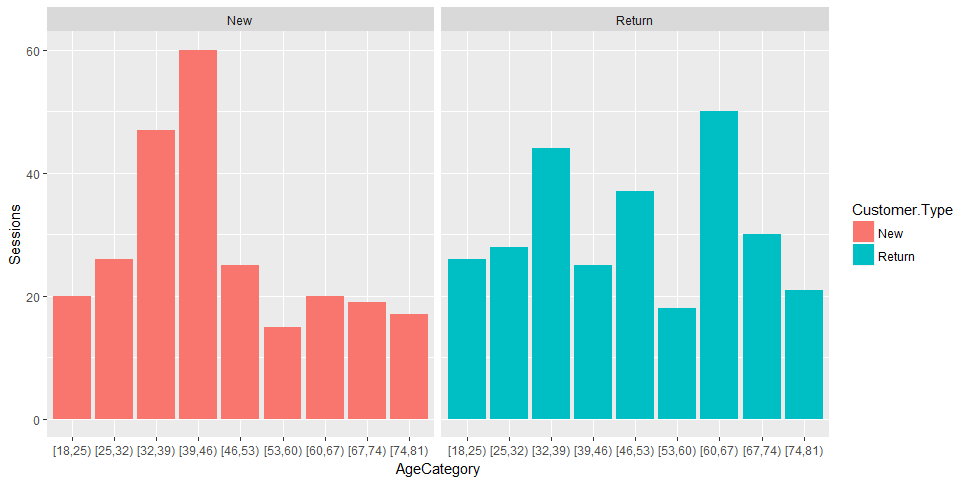
#Age category and gender separated

ggplot(data = newDataTable,aes(AgeCategory,Sessions))+geom\_bar(stat="identity",aes(fill=Gender))+facet\_grid(.~Gender)

ggplot(data = newDataTable,aes(AgeCategory,Sessions))+geom\_bar(stat="identity",aes(fill=Customer.Type))

#Age category and new and return type of customers

ggplot(data = newDataTable,aes(AgeCategory,Sessions))+geom\_bar(stat="identity",aes(fill=Customer.Type))+facet\_grid(.~Customer.Type)

#Age category and type of customers separated