final\_clust

JAY SHAH

March 16, 2018

### Setting working Directory to Data lOcation

library(data.table)  
library(factoextra)

## Loading required package: ggplot2

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

train = fread("E:/USA/Projects/Research/R\_code/w6/train\_clust.csv",data.table = T)  
train = train[,-1]  
test = fread("E:/USA/Projects/Research/R\_code/w6/test\_clust.csv",data.table = T)  
test = test[,-1]

Here I have built Custom Function to create Features

features = function(data){  
 newdata = NULL  
 mean\_speed = as.data.frame( rep(0,dim(data)[1]))  
 mean\_acc\_lot =as.data.frame( rep(0,dim(data)[1]))  
 mean\_acc\_lan = as.data.frame(rep(0,dim(data)[1]))  
 sd\_speed = as.data.frame(rep(0,dim(data)[1]))  
 sd\_acc\_lot = as.data.frame(rep(0,dim(data)[1]))  
 sd\_acc\_lat = as.data.frame(rep(0,dim(data)[1]))  
 max\_speed = as.data.frame(rep(0,dim(data)[1]))  
 max\_acc\_lot = as.data.frame(rep(0,dim(data)[1]))  
 max\_acc\_lat = as.data.frame(rep(0,dim(data)[1]))  
 min\_speed = as.data.frame(rep(0,dim(data)[1]))  
 min\_acc\_lot = as.data.frame(rep(0,dim(data)[1]))  
 min\_acc\_lat = as.data.frame(rep(0,dim(data)[1]))  
 for (i in c(1:dim(data)[1])) {  
 mean\_speed[i,] = mean(unlist(data[i,4:64]))  
 mean\_acc\_lot[i,] = mean(unlist(data[i , 65:125]))  
 mean\_acc\_lan[i,] = mean(unlist(data[i, 126:186]))  
 sd\_speed[i,] = sd((unlist(data[ i,4:64])))  
 sd\_acc\_lot[i,] = sd((unlist(data[i , 65:125])))  
 sd\_acc\_lat[i,] = sd((unlist(data[i , 126:186])))  
 max\_speed[i,] = max((unlist(data[ i,4:64])))  
 max\_acc\_lot[i,] = max((unlist(data[i , 65:125])))  
 max\_acc\_lat[i,] = max((unlist(data[i , 126:186])))  
 min\_speed[i,] = min((unlist(data[ i,4:64])))  
 min\_acc\_lot[i,] = min((unlist(data[i , 65:125])))  
 min\_acc\_lat[i,] = min((unlist(data[i , 126:186])))  
 }  
 newdata =as.data.table(cbind(mean\_speed,mean\_acc\_lot,mean\_acc\_lan, sd\_speed,sd\_acc\_lot,sd\_acc\_lat,  
 max\_speed,max\_acc\_lot,max\_acc\_lat,min\_speed,mean\_acc\_lot,mean\_acc\_lan))  
 colnames(newdata) = c("mean\_speed","mean\_acc\_lot","mean\_acc\_lan", "sd\_speed","sd\_acc\_lot","sd\_acc\_lat","max\_speed",  
 "max\_acc\_lot","max\_acc\_lat","min\_speed", "mean\_acc\_lot","mean\_acc\_lan")  
 return(newdata)  
}

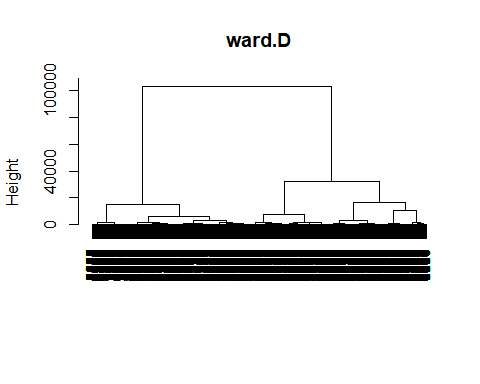
### Creating Data

train\_feat = features(train)  
test\_feat = features(test)

hc\_ward=hclust(dist(train\_feat), method="ward.D")

### Questions related to type of Dissimilarity measure to use?

plot(hc\_ward,main="ward.D", xlab="", sub="", cex=.9)

 Here we can see only 2 clusters.

fviz\_nbclust(train\_feat, hcut, method = "wss",hc\_method = "ward.D", main = "Ward.D") +  
 geom\_vline(xintercept = 2, linetype = 2)

