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)
}
```

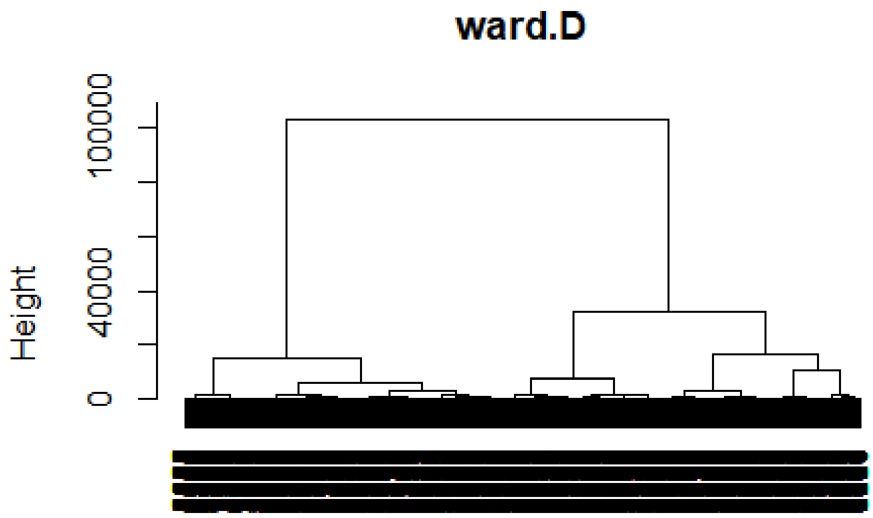
Creating Data

```
train_feat = features(train)
test_feat = features(test)

hc_ward=hclust(dist(train_feat), method="ward.D")
```

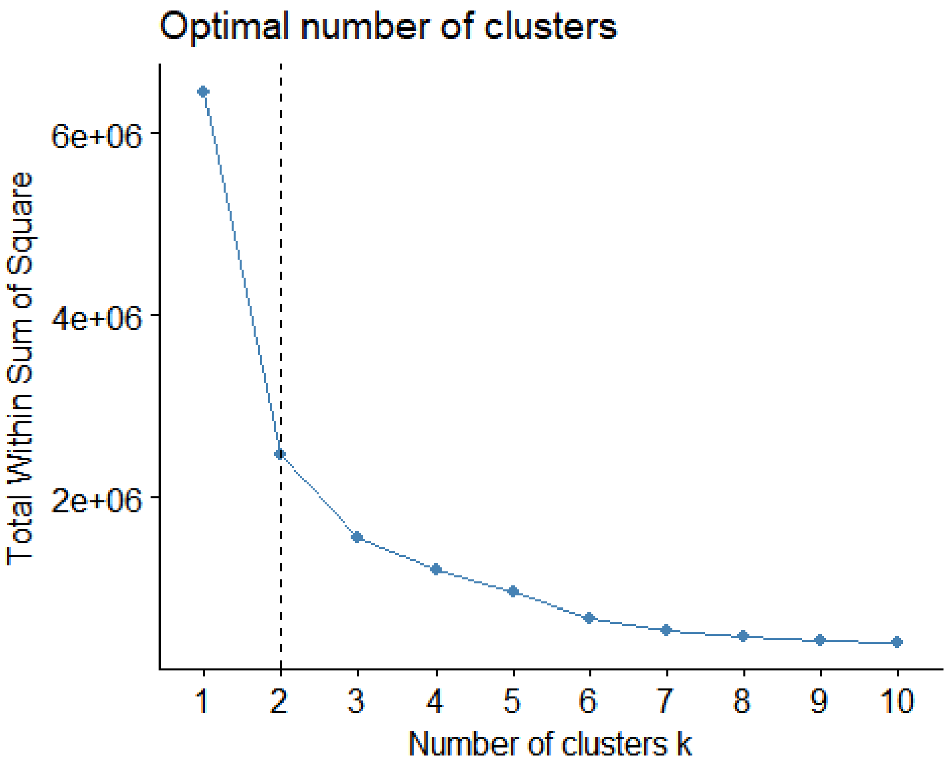
Questions related to type of Dissimilarity measure to use?

```
plot(hc_ward,main="ward.D", xlab="", sub="", cex=.9)
```



Here we can see only 2 clusters.

```r
fviz_nbclust(train_feat, hcut, method = "wss",hc_method = "ward.D", main =
"Ward.D") +
  geom_vline(xintercept = 2, linetype = 2)
```


Optimal number of clusters
Total Within Sum of Square
6e+06
4e+06
2e+06
Number of clusters k