millennial_clustering

suqiang_song

February 17, 2018

install.packages("factoextra")

```
library(factoextra) # clustering algorithms & visualization
```

```
## Loading required package: ggplot2
```

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

—- Read in the data —-

```
data <- read.table('C:/Data/millennial')
col <- read.table('C:/Data/millennial_column')
colnames(data) <- col$V2
```

—- Sample/separate data —-

```
n <- dim(data)[1]
p <- dim(data)[2]

visit <- which( 1:p %% 2 == 0 )
spend <- which( 1:p %% 2 == 1 )[-1]
# visit
data_v <- data[,visit]
# spend
data_s <- data[,spend]
# visit & spend
data_vs <- data
# part of columns to identify the quality of cluster
visit_vars<-c("google_visit","apple_visit","jcrew_visit","itunes_visit")
spend_vars<-c("google_aveSpend","apple_aveSpend","jcrew_aveSpend","itunes_aveSp
end")
visit_spend_vars<-c("google_visit","apple_visit","jcrew_visit","itunes_visi
t","google_aveSpend","apple_aveSpend","jcrew_aveSpend","itunes_aveSpend")
# scale the data sets
scaleDataV=scale(data_v[,-1])
scaleDataS=scale(data_s[,-1])
scaleDataVS=scale(data_vs[,-1])
```

define the UDF wssplot which helps to find best K

```
wssplot <- function(data, nc=15, seed=123){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")}
```

define the UDF for cluster

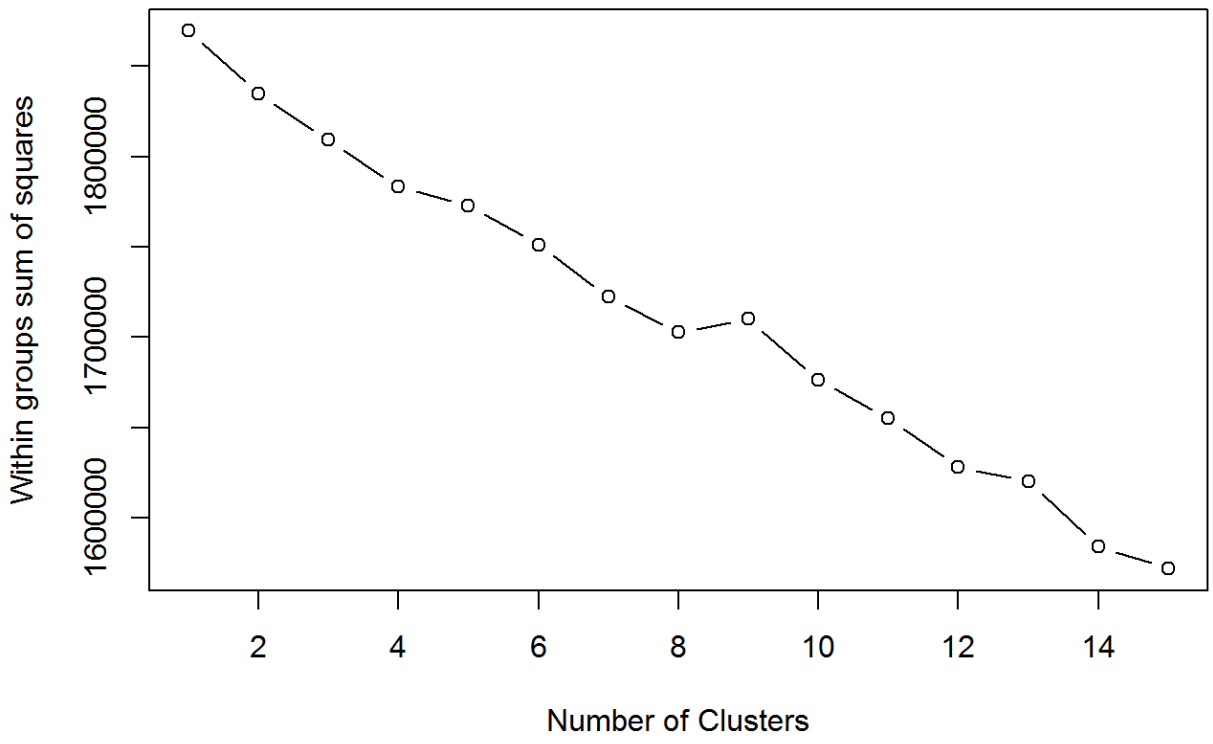
```
doCluster <- function(scaleData,bestk=10,iterMax=500,nstart=15) {
  return (kmeans(scaleData, bestk,iterMax,nstart))
}
```

define the UDF findBiggestCluster

```
findBiggestCluster <- function(rawData,scaleData,bestk=10,iterMax=500,nstart=1
5) {
  k_out <- doCluster(scaleData,bestk,iterMax,nstart)
  table(k_out$cluster)
  neg = which.max(k_out$size)
  mil = !is.element(k_out$cluster, neg)
  total = apply(scaleData, 1, sum)
  tapply(total, mil, summary)
  list(k_out=k_out,d_out=rawData[is.element(k_out$cluster, (1:bestk)[-neg]),1],
neg=neg)
}
```

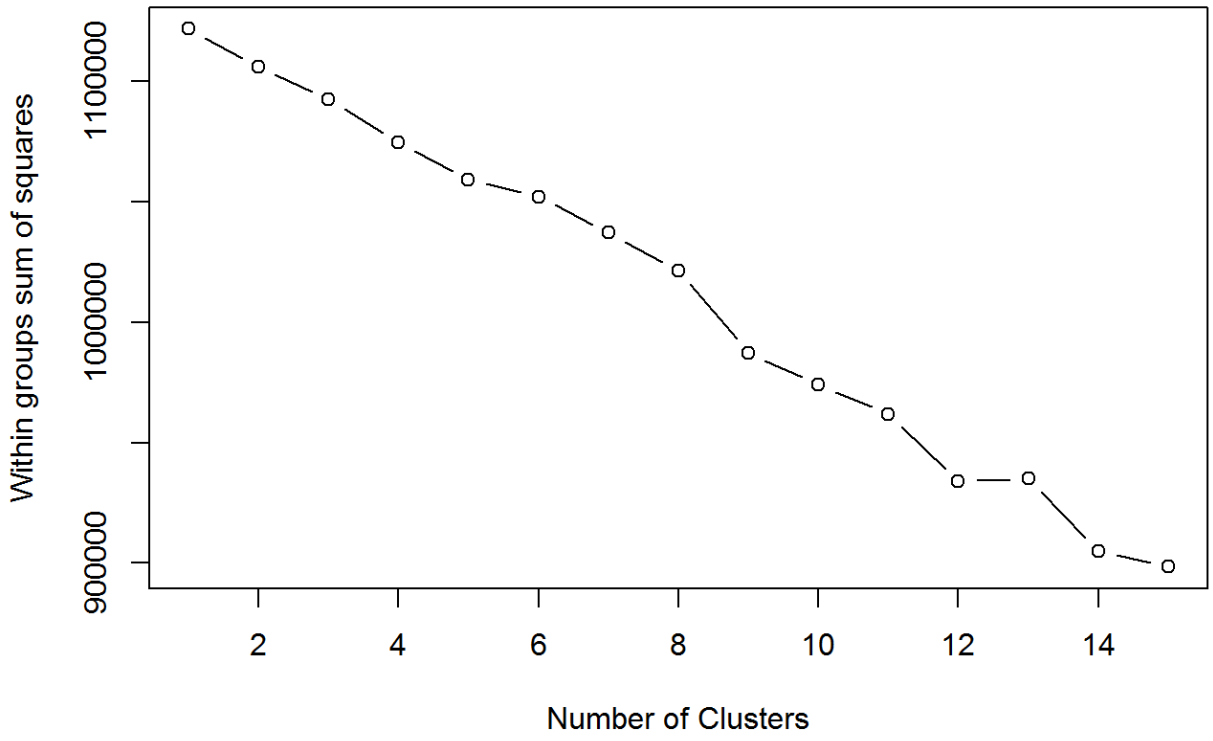
=== clustering with user visit ===

```
# find the best k from SSE
wssv1 <- wssplot(scaleDataV)
```

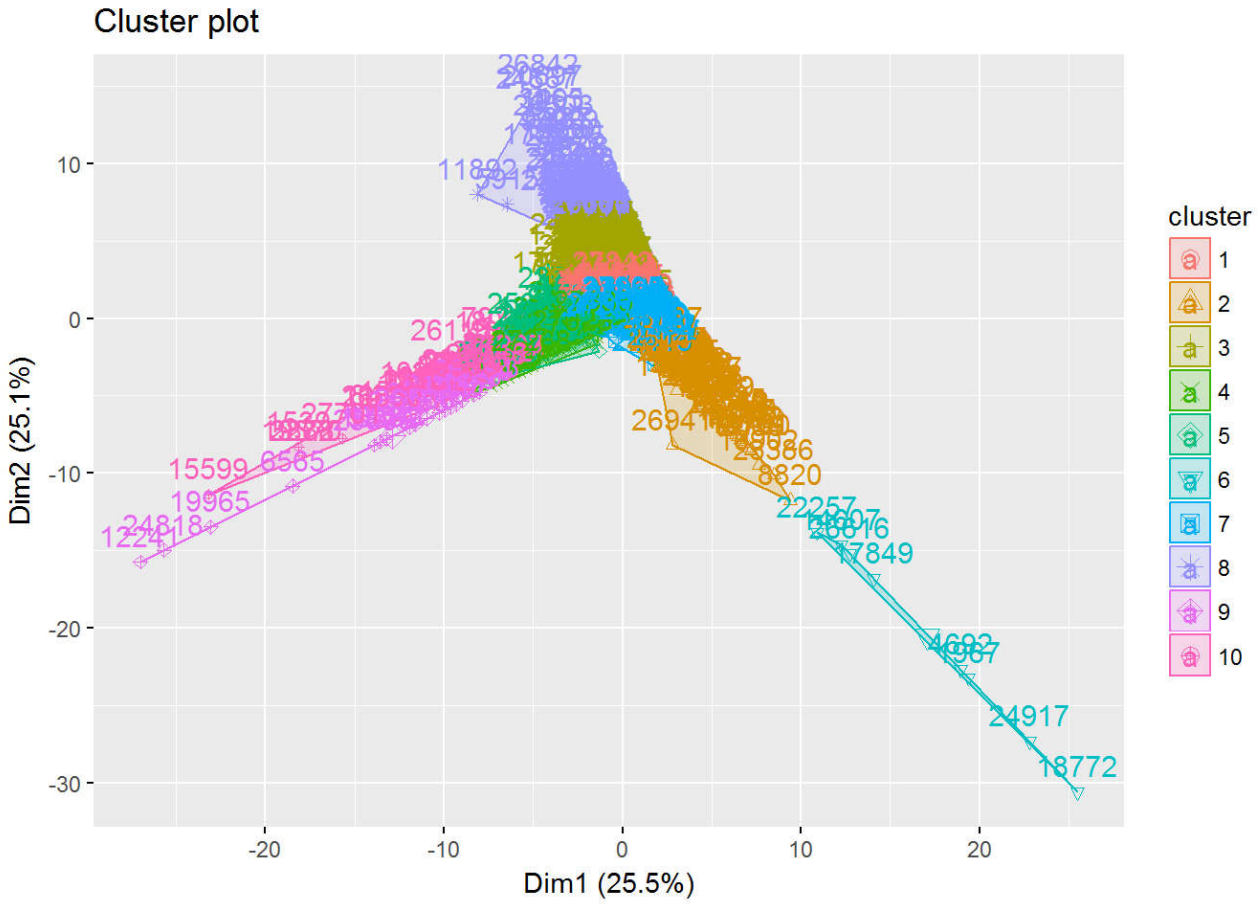


```
#best_v
best_v1 <- 12
millennial_v1 <- findBiggestCluster(data,scaleDataV,best_v1)
#View(millennial_v1$d_out)

# further break down the biggest one
row <- is.element(millennial_v1$k_out$cluster, millennial_v1$neg)
col <- which(apply(data_v[row, ], 2, sum) != 0)
data_v2 <- data_v[row, col]
scaleDataV2 <- scale(data_v2[,-1])
#View(data2_v)
wssv2 <- wssplot(scaleDataV2)
```



```
#best2_v
best_v2 <- 10
millennial_v2 <- findBiggestCluster(data,scaleDataV2,best_v2)
user_v2 = data[row, 1]
millennial_v_final <- user_v2[ is.element(millennial_v2$k_out$cluster, (1:best_
v2)[-millennial_v2$neg])]
#View(millennial_v_final)
#visual the clusters, but use few columns
data_par_v <- subset(scaleDataV2,select = visit_vars)
#View(data_par_v)
k_v_par <- doCluster(data_par_v,best_v2)
fviz_cluster(k_v_par, data = data_par_v)
```



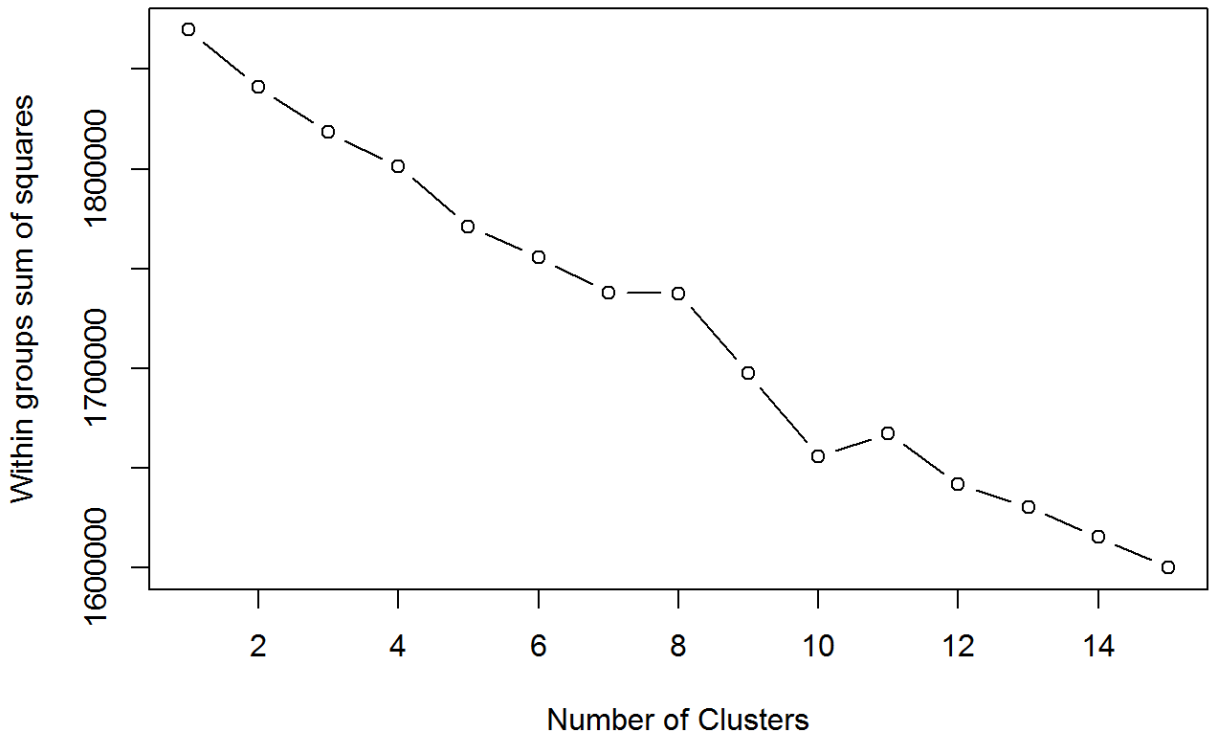

```
# save the final result to csv
write.csv(millennial_v_final, "C:/Data/millennial_v_final.csv")
```

=== clustering with user avgSpend ===

```
# find the best k from SSE
wsss1 <- wssplot(scaleDataS)
```

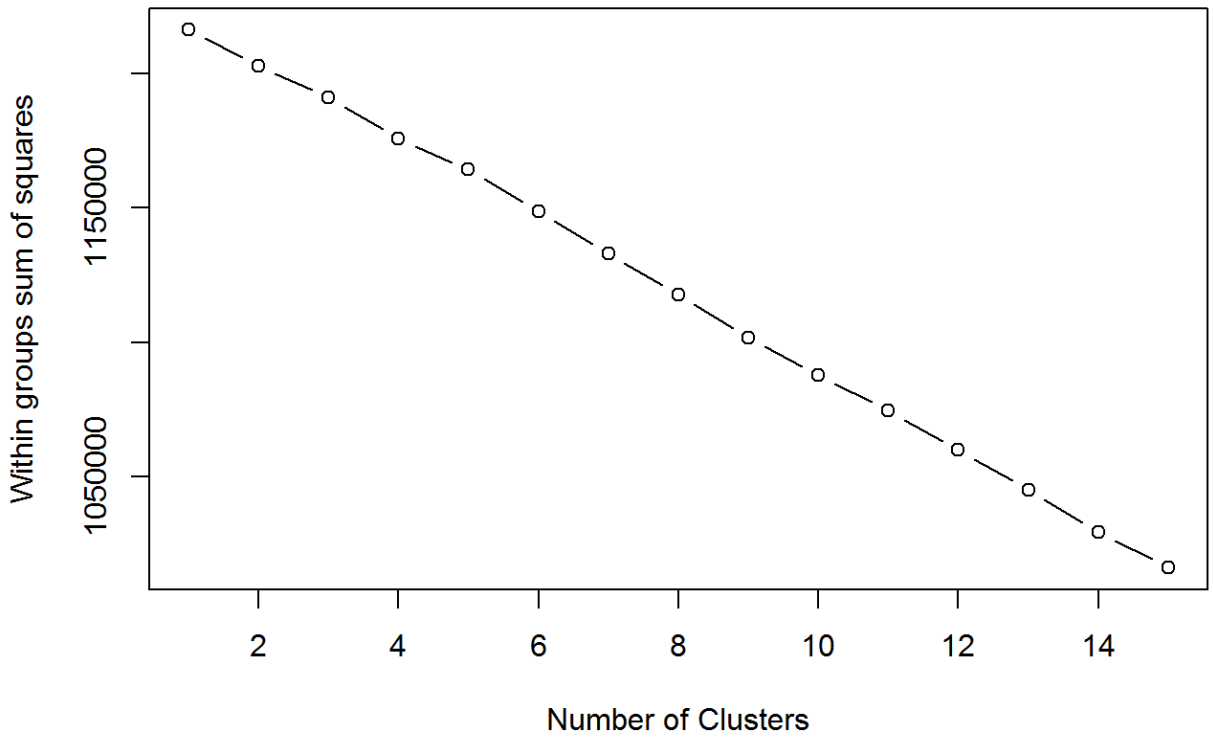
```
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
```

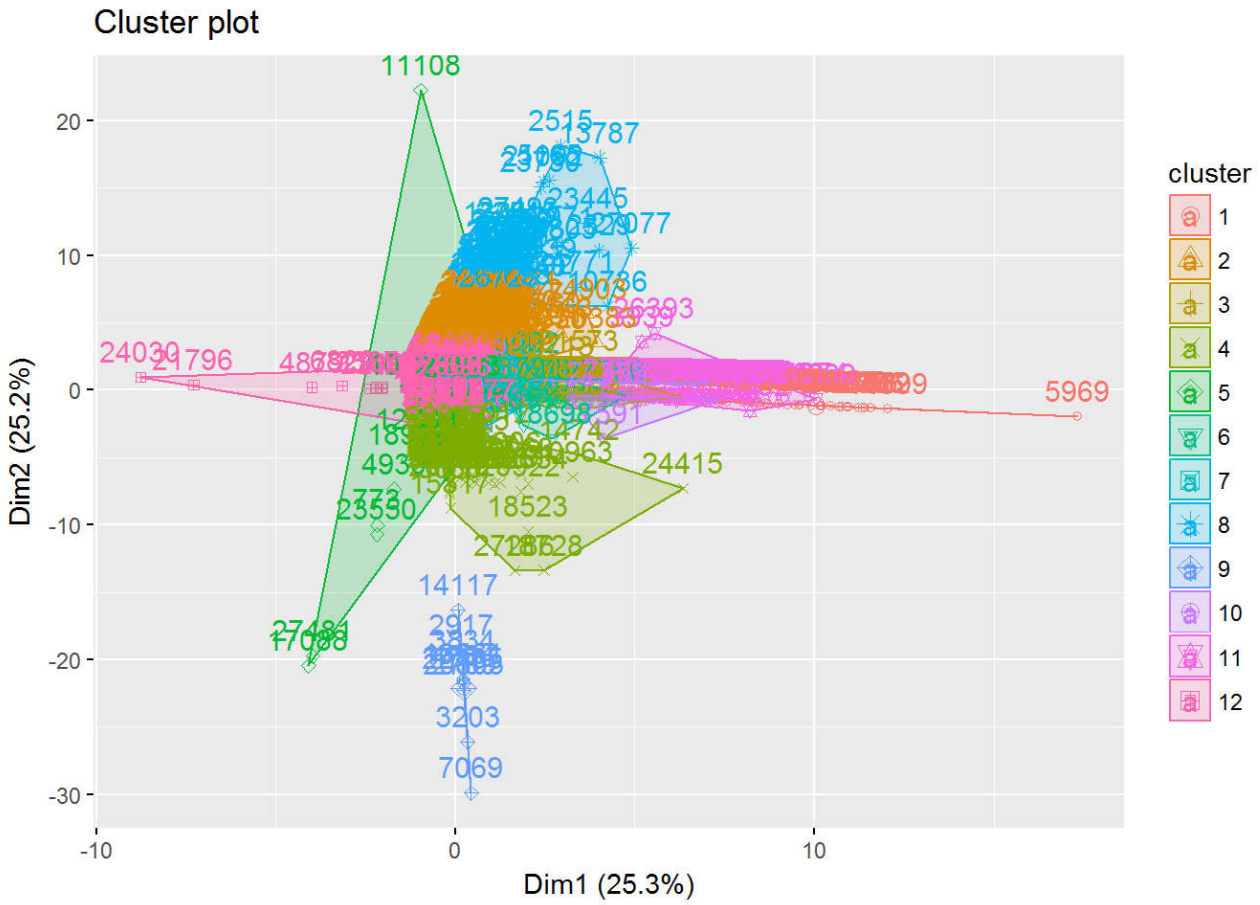



```
#best_s
best_s1 <- 11
millennial_s1 <- findBiggestCluster(data,scaleDataS,best_s1)

# further break down the biggest one
row <- is.element(millennial_s1$k_out$cluster, millennial_s1$neg)
col <- which(apply(data_s[row, ], 2, sum) != 0)
data_s2 <- data_s[row, col]
scaleDataS2 <- scale(data_s2[,-1])
wsss2 <- wssplot(scaleDataS2)
```




```
#best_s2
best_s2 <- 12
millennial_s2 <- findBiggestCluster(data,scaleDataS2,best_s2)
user_s2 = data[row, 1]
millennial_s_final <- user_s2[ is.element(millennial_s2$k_out$cluster, (1:best_
s2)[-millennial_s2$neg])]
#View(millennial_s_final)
#visual the clusters, but use few columns
data_par_s <- subset(scaleDataS2,select = spend_vars)
#View(data_par_s)
k_s_par <- doCluster(data_par_s,best_s2)
fviz_cluster(k_s_par, data = data_par_s)
```

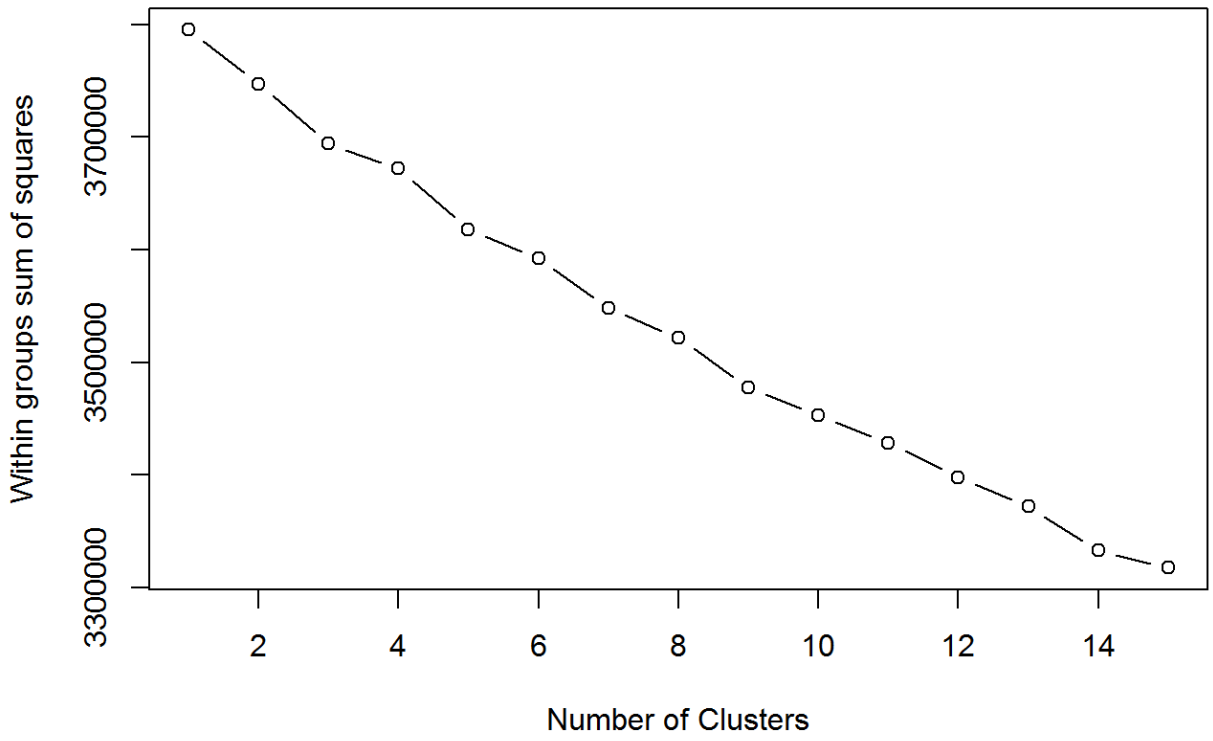



```
# save the final result to csv
write.csv(millennial_s_final, "C:/Data/millennial_s_final.csv")
```

=== clustering with user visit avgSpend ===

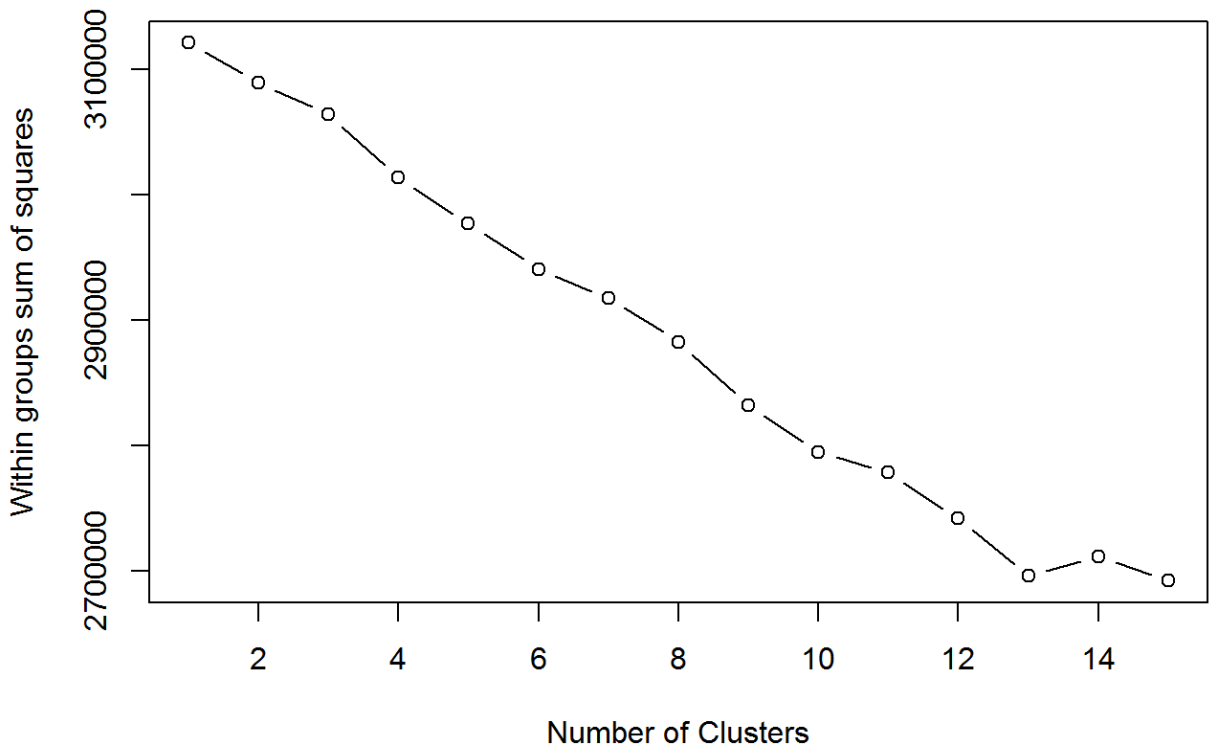
```
# find the best k from SSE
wssvs1 <- wssplot(scaleDataVS)
```

```
## Warning: did not converge in 10 iterations
```

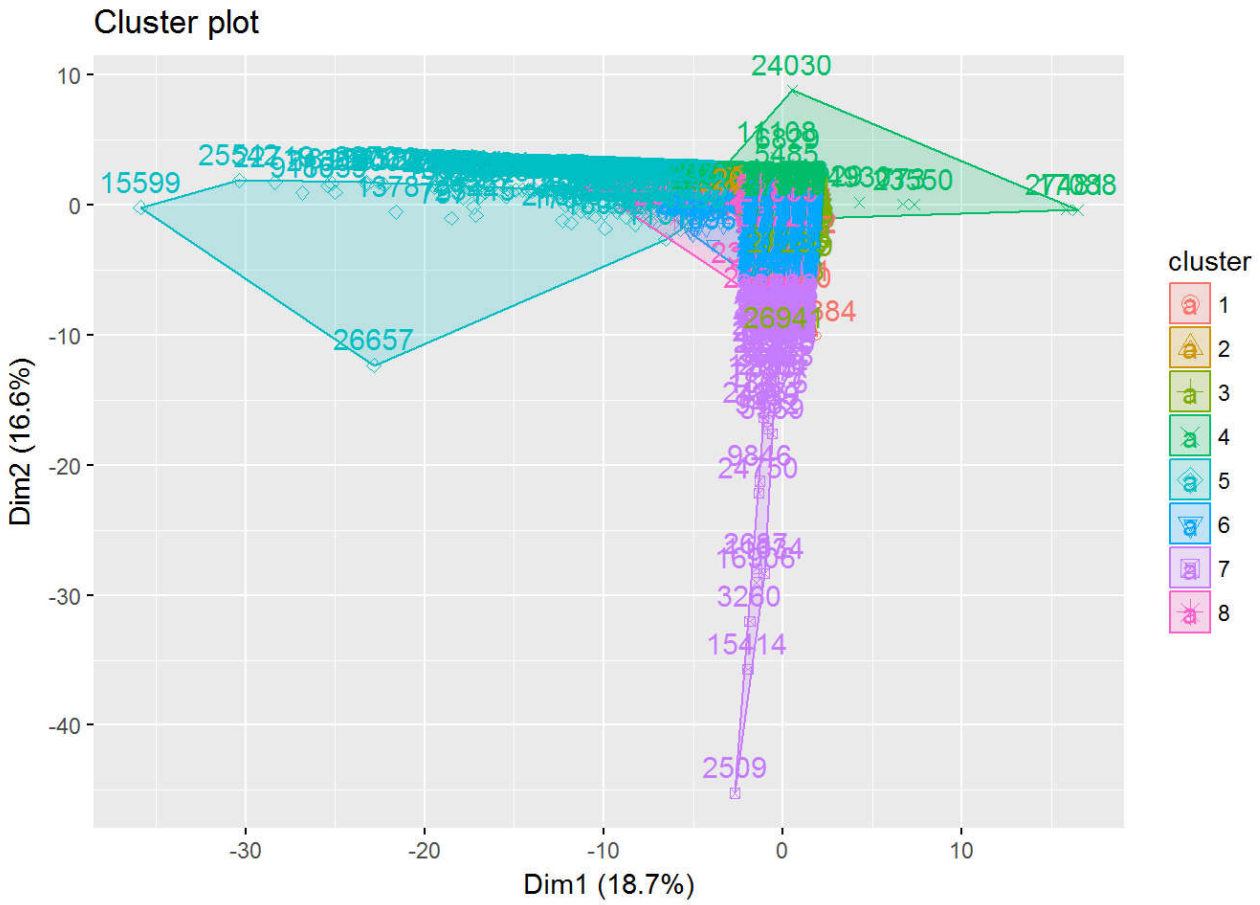



```
#best_vs
best_vs1 <- 10
millennial_vs1 <- findBiggestCluster(data,scaleDataVS,best_vs1)

# further break down the biggest one
row <- is.element(millennial_vs1$k_out$cluster, millennial_vs1$neg)
col <- which(apply(data_vs[row, ], 2, sum) != 0)
data_vs2 <- data_vs[row, col]
scaleDataVS2 <- scale(data_vs2[,-1])
wssvs2 <- wssplot(scaleDataVS2)
```




```
#best_vs2
best_vs2 <-
millennial_vs2 <- findBiggestCluster(data,scaleDataVS2,best_vs2)
user_vs2 = data[row, 1]
millennial_vs_final <- user_vs2[ is.element(millennial_vs2$k_out$cluster, (1:be
st_vs2)[-millennial_vs2$neg])]
#View(millennial_s_final)
#visual the clusters, but use few columns
data_par_vs <- subset(scaleDataVS2,select = visit_spend_vars)
#View(data_par_vs)
k_vs_par <- doCluster(data_par_vs,best_vs2)
fviz_cluster(k_vs_par, data = data_par_vs)
```




```
# save the final result to csv
write.csv(millennial_vs_final, "C:/Data/millennial_vs_final.csv")
```