HUDK4050: Class Activity 6

Ningyao Xu

10/16/2018

Data Management

```
#Load data
DF1 <- read.csv("HUDK405019-clustering.csv", header = TRUE)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#rownames == First name + Last name
for (i in c(1,2,15,16))
{DF1[,i] = as.character(DF1[,i])}
DF1$name <- paste(DF1$First.Name, DF1$Last.Name)
rownames(DF1) <- DF1$name
DF1 <- DF1[,3:16]


#Delete those who write latitude and longtitude twice in the survey
list <- NULL
for (i in 1:nrow(DF1))
  { if (DF1[i,13] == DF1[i,14] )
list <- c(list,i)}
DF1 <- DF1[-list,]


# reverse those who put latitude and longtitude in wrong order
a <-NULL
b<- NULL
reverse <- grep("E",DF1[,13])
for (i in reverse)
  { a =  DF1[i,13]
    b =  DF1[i,14]
    DF1[i,13] <- b
    DF1[i,14] <- a}

#Find the signal and delete all the things after the signal
#"Â° is how my DELL shows "°", I have no idea why it shows this way
for (j in c(13:14)){
for (i in 1:nrow(DF1))
```

```
  { if (grepl("Â°", DF1[i,j]) )
  { psn <- as.numeric(regexpr("Â°", DF1[i,j]))
    DF1[i,j] <-  substr(DF1[i,j], 1, psn-1)}}}

#If you are using mac, use the following one
for (j in c(13:14)){
for (i in 1:nrow(DF1))
  { if (grepl("\\D", DF1[i,j]) )
  { psn <- as.numeric(regexpr("\\D", DF1[i,j]))
    DF1[i,j] <-  substr(DF1[i,j], 1, psn-1)}}}

#Delete all the space, alphabet from the data and turn all the data into numeric
 for (i in c(1:11,13,14))
{ DF1[,i]= gsub("[[:alpha:]]", "", DF1[,i])
  DF1[,i]= gsub(" ", "", DF1[,i])
  DF1[,i] = as.numeric(DF1[,i])}
```

```
## Warning: NAs introduced by coercion
```

```
# Omit all the NAs from the data
DF1 <- na.omit(DF1)

DF2 <- data.frame(select_if(DF1,is.numeric))
#Convert the index numbers of the data fram into the student names.

#Wrangle data using dplyr to include only the numerical values.

#Scale the data so that no variable has undue influence

DF2 <- scale(DF2)
```

Find lattitudes & longitudes for cities

```
#Unfortunately Google has restricted access to the Googgple Maps API so the code below no longer works.

#install.packages("ggmap")
#install.packages("rgdal")
#library(ggmap)
#library(tmaptools)

#Request lattitude and longitude from Google Maps API
#DF2 <- geocode(as.character(DF2$Q1_1), output = "latlon", source = "dsk")
```

Now we will run the K-means clustering algorithm we talked about in class. 1) The algorithm starts by randomly choosing some starting values 2) Associates all observations near to those values with them 3) Calculates the mean of those clusters of values 4) Selects the observation closest to the mean of the cluster 5) Re-associates all observations closest to this observation 6) Continues this process until the clusters are no longer changing

Notice that in this case we have 10 variables and in class we only had 2. It is impossible to vizualise this process with 10 variables.

Also, we need to choose the number of clusters we think are in the data. We will start with 4.

```r
fit <- kmeans(DF2, 4)

#We have created an object called "fit" that contains all the details of our clustering including which

#We can access the list of clusters by typing "fit$cluster", the top row corresponds to the original or

fit$cluster
```

```
##       Timothy Lee        jiahao guo Leonardo Restrepo        Xinke Song
##                 3                 2                 4                 2
##        Zixuan  Ma          Yiwei Qi        XINYI ZHOU       XIAOJUE LIU
##                 2                 3                 1                 2
##      Minruo  Wang         Anqi Duan      Chengxuan Hu    CHAOXIONG CHEN
##                 3                 3                 2                 4
##           Ling Ai      Joellyn Heng       Ruiqi  Wang          BOZI JIN
##                 3                 2                 2                 4
##        Qiyang Lin          Yiwen Ma        Ziyuan Guo       Shijie Shao
##                 2                 2                 2                 2
##        Eudora Niu     Jiancong Shen          Yijia Wu           XI YANG
##                 3                 2                 2                 2
##        Beibei Cao        Chenyu Yan     LINGLING MIAO     Maho  Hayashi
##                 2                 3                 3                 4
##        Suwon Jung      Xiaowen Chen         Jiali Jin        Lintong Li
##                 4                 2                 3                 2
##        Ningyao Xu   Zhongyuan Zhang           Yaqi Lu       Yujun Zhang
##                 3                 2                 4                 2
##      Xudian Zhang          Jie Chen          Han Wang
##                 2                 2                 4
```

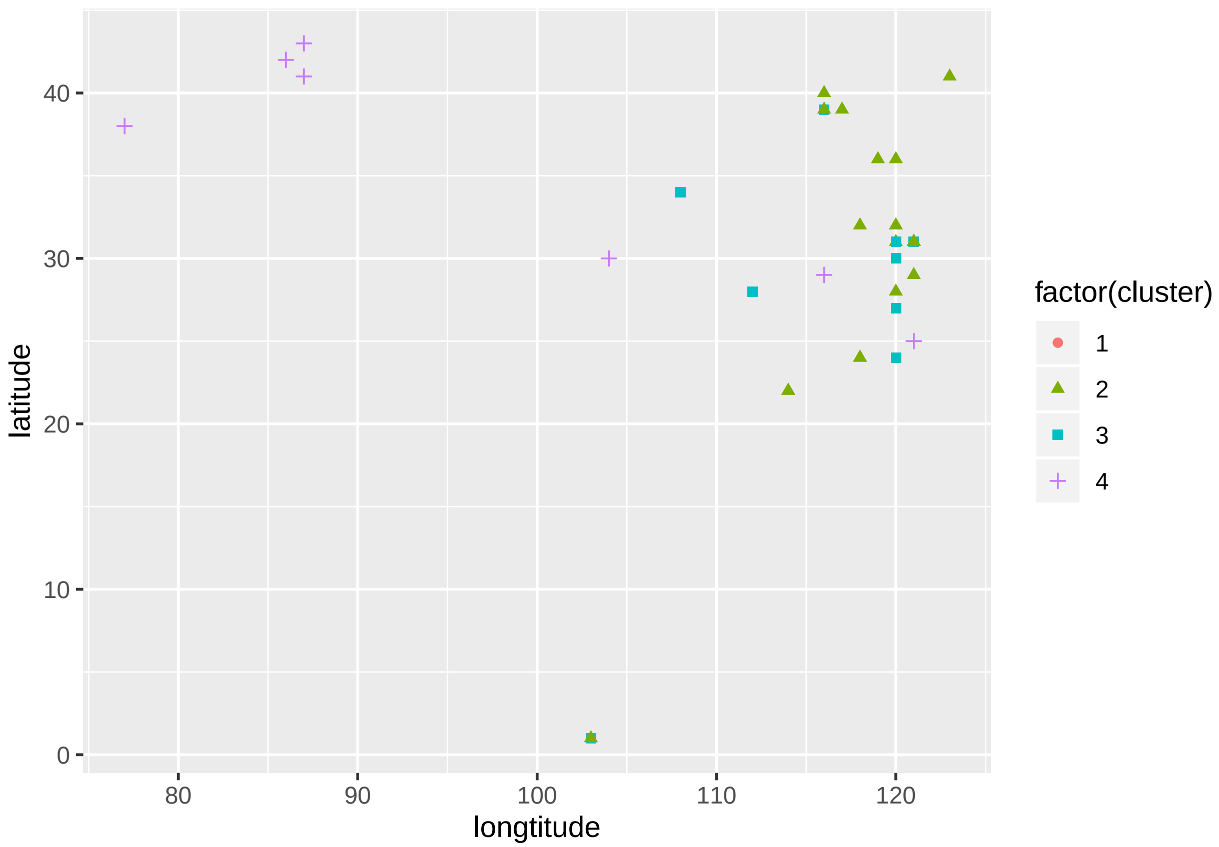
```r
#We can also attach these clusters to te original dataframe by using the "data.frame" command to create

DF3 <- data.frame(DF2, fit$cluster)

#Have a look at the DF3 dataframe. Lets change the names of the variables to make it more convenient wi
```

Visualize your clusters in ggplot

```r
#Create a scatterplot that plots location of each student and colors the points according to their clus
DF4 <- data.frame(DF1[,13],DF1[,14],fit$cluster)
names(DF4)  <- c("latitude", "longtitude","cluster")
attach(DF4)
library(ggplot2)
ggplot(DF4, aes(x =  longtitude, y =latitude, pch = factor(cluster))) +
    geom_point(aes(color = factor(cluster)))
```



Can you group students from the classes data set in Assignment 2 using K-modes?

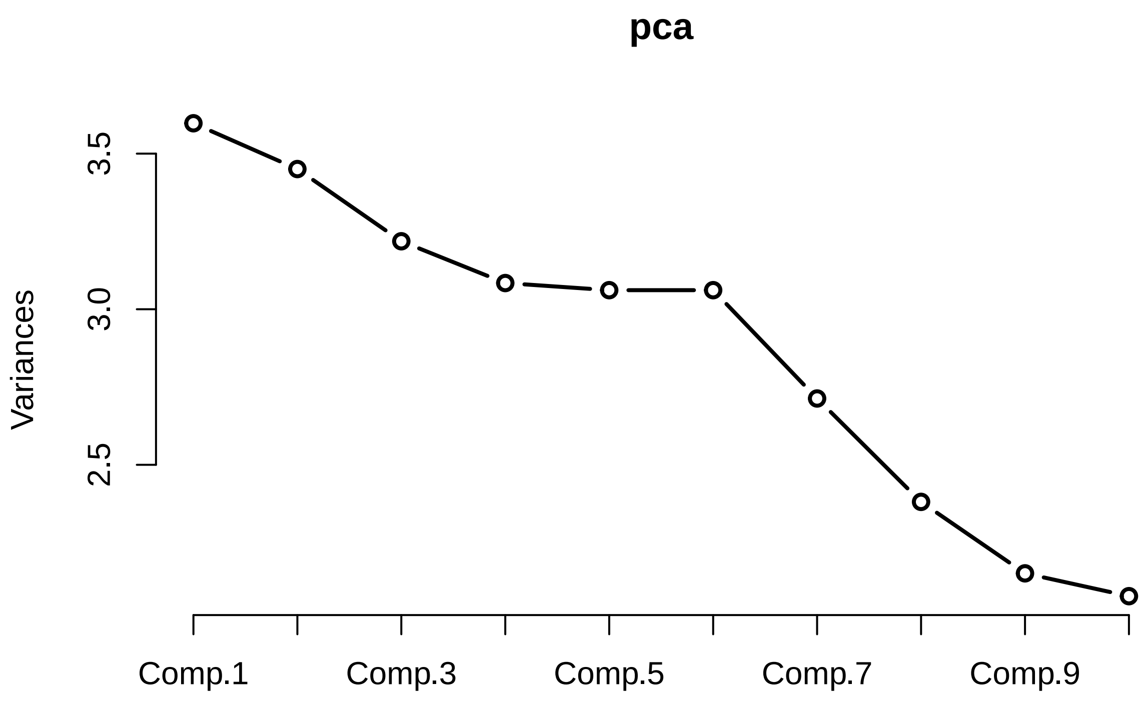
```
DT1 <- read.csv("hudk4050-classes.csv",header = TRUE)
DT1$Name <- paste(DT1$First.Name, DT1$Last.Name)
DT2_dirty <- DT1[,3:9]
DT3 <- DT2_dirty %>% gather(classnum, classcode, `Class.1`, `Class.2`, `Class.3`, `Class.4`, `Class.5`,
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
DT3$classcode = gsub(" ", "", DT3$classcode)
DT3 <- DT3 %>% filter(classcode != "HUDK4050") %>% filter(Name != "ZIMO CHEN")
DT3$Count = 1
DT3 <- DT3[which(DT3$classcode != ""),]
DT4 <- DT3 %>% spread(classcode,Count)
row.names(DT4) = DT4$Name
DT4$Name <- NULL
DT4 = ifelse(is.na(DT4), 0, 1)
DT5 = as.matrix(DT4)
DT5 <- scale(DT5)
library(MASS)
```

```
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
```

```
set.seed(123)
pca=princomp(DT5[,1:50],cor=T)
screeplot(pca,type="line",lwd=2)
```



```
#According to this plot, maybe we should try cluster 6 groups.
set.seed(123)
fit2 <- kmeans(DT5,6)
fit2$size
```

```
## [1]  1  1 30  2 13  3
```

```
cluster <- data.frame(fit2$cluster)
colnames(cluster) <- c("cluster")
cluster
```

```
##                    cluster
## Alysandra Zhang          3
## Anqi Duan                3
## Artemas Wang             3
## Beibei Cao               6
## Bernell Downer           3
## chaoxiong chen           3
## Chengxuan Hu             5
## Chenyu Yan               3
## Christine Odenath       3
## David Pearce            3
## Di Mao                  5
## Eudora Xinyi Niu        4
## HAN GE                  3
## Han Wang                3
## INDIRA BATAYEVA         3
## jiahao guo              5
## Jiancong Shen           3
## Jie Chen                3
## Jingru Zhang            5
## Joellyn Heng            2
## Leonardo Restrepo       3
## Ling Ai                 3
## LINGLING MIAO           5
## Lintong Li              3
## Luyi Dai                5
## Maho Hayashi            1
## Minruo Wang             3
## Ningyao Xu              5
## Qiyang Lin              3
## Ruiqi Wang              5
## Shijie Shao             5
## Suwon Jung              3
## Timothy Lee             3
## Wanruo Zhang            3
## XI YANG                 5
## Xiaowen Chen            3
## xinyi zhou              3
## XUDIAN ZHANG            5
## YAQI LU                 6
## Yawei Zhu               3
## Yigao Liu               4
## Yiwei Qi                3
## Yiwen Ma                5
## yixiao li               6
## Yixuan Zhu              3
## Yiyi Xie                3
## Yujun Zhang             3
## Zhaozhuo Zheng          3
## Zhongyuan Zhang         3
## Ziyuan Guo              5
```

```r
# Just to check, I am in the Applied Statistics program and people in the group 5 are exactly those in
```