Assignment 12 Follow Up

Will Doyle

4/9/2019

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.1.0          v purrr   0.3.2
## v tibble  2.1.1          v dplyr   0.8.0.1
## v tidyr   0.8.3.9000     v stringr 1.3.1
## v readr   1.3.1          v forcats 0.3.0

## -- Conflicts --------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
library(stats)
library(flexclust)
```

```
## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4
```

```
library(ggplot2)
library(LICORS)
library(knitr)
library(modelr)
```

```
ev<-read_xlsx("turkiyestudentevaluation_generic.xlsx")
```

```
names(ev)<-tolower(names(ev))

myvars<-paste0("q",1:10)

ev_full<-ev

ev<-ev%>%select_at(.vars = myvars)
```

Checking on Number of Clusters Needed

```
# Test to see how many clusters are needed
c_test <- stepFlexclust(ev, k = 2:7, nrep = 20)
```

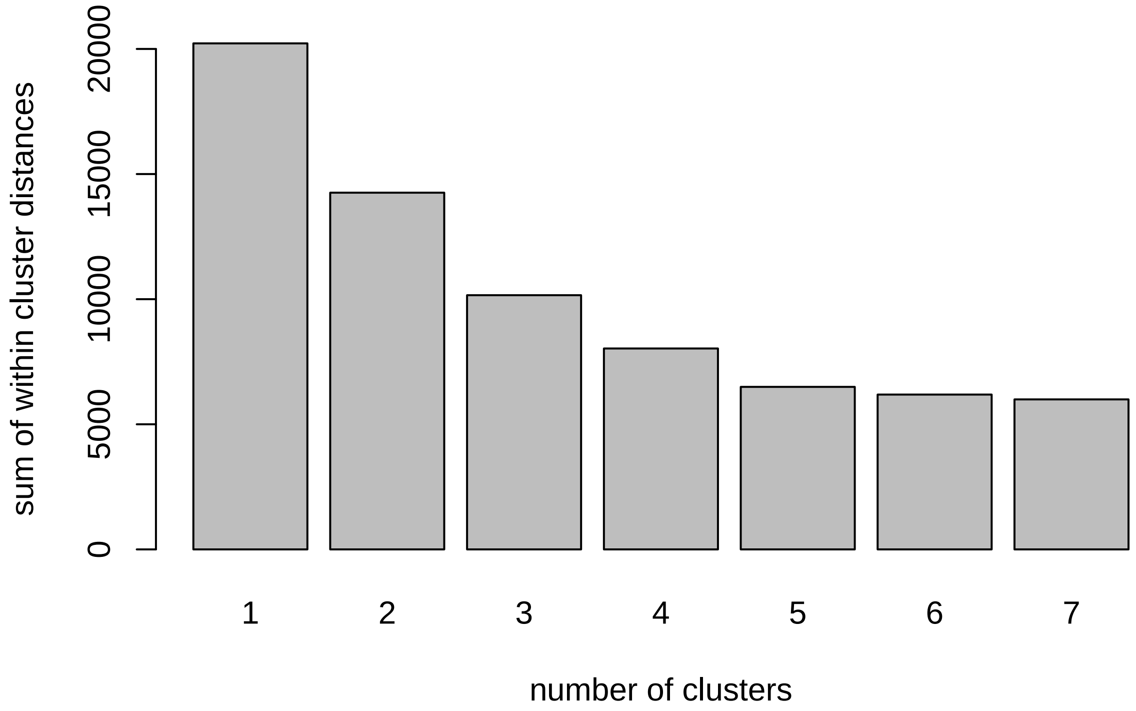
```
## 2 :
## 3 :
## 4 :
## 5 :
## 6 :
```

```
## 7 :
```

```
c_test
```

```
## stepFlexclust object of family 'kmeans'
##
## call:
## stepFlexclust(x = ev, k = 2:7, nrep = 20)
##
##   iter converged   distsum
## 1   NA        NA 20223.039
## 2    5      TRUE 14255.484
## 3   10      TRUE 10157.736
## 4   10      TRUE  8029.395
## 5   11      TRUE  6494.809
## 6   25      TRUE  6187.715
## 7   26      TRUE  5994.677
```

```
plot(c_test)
```



It looks like 3 clusters should work

```
c1<-kmeanspp(ev,k=3,start="random",iter.max=1000,nstart = 50)
table(c1$cluster)
```

```
##
##    1    2    3
## 2080 1467 2273
```

```
ev$cluster<-c1$cluster

table(ev$cluster)
```

```
##
##    1    2    3
## 2080 1467 2273
```

Summarizing Clusters

```
ev%>%
  group_by(cluster)%>%
  summarize_at(.vars=myvars,.funs = "mean")
```

```
## # A tibble: 3 x 11
##   cluster    q1    q2    q3    q4    q5    q6    q7    q8    q9   q10
##     <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1       1  4.25  4.35  4.36  4.36  4.38  4.36  4.35  4.34  4.36  4.37
## 2       2  1.37  1.43  1.63  1.46  1.44  1.47  1.42  1.43  1.63  1.42
## 3       3  2.73  2.97  3.09  2.96  3.02  3.02  2.95  2.89  3.06  3.00
```

Basic idea: there are three clusters of students: the happy ones, the unhappy ones, and the “meh” ones.

```
ev<-ev%>%
  mutate(cluster=fct_recode(as_factor(as.character(cluster)),
                            "Unhappy"="2",
                            "Happy"="1",
                            "Meh"="3"))
table(ev$cluster)
```

```
## 
##     Meh   Happy Unhappy 
##    2273    2080    1467
```

Relationship of clusters with course difficulty

```
ev_full<-ev_full%>%select(difficulty,attendance)

ev<-ev%>%bind_cols(ev_full)

mod1<-lm(difficulty~as.factor(cluster)+attendance,data=ev)

summary(mod1)
```

```
## 
## Call:
## lm(formula = difficulty ~ as.factor(cluster) + attendance, data = ev)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.7882 -1.0496 -0.1873  0.8127  2.9504 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 2.18732    0.03152  69.386  < 2e-16 ***
## as.factor(cluster)Happy    -0.13772    0.03684  -3.739 0.000187 ***
## as.factor(cluster)Unhappy -0.09994    0.04082  -2.448 0.014398 *  
## attendance                  0.40021    0.01090  36.719  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.212 on 5816 degrees of freedom
## Multiple R-squared:  0.1929, Adjusted R-squared:  0.1924
## F-statistic: 463.2 on 3 and 5816 DF,  p-value: < 2.2e-16
```