HW12_Sampathirao_A

Anvita Sampathirao

8/13/2019

```
#install.packages("psych")
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.6.1
```

```
data(bfi)
bfi1 <- bfi[,1:25]
bfi1 <- data.frame(bfi1)
bfi2 <- na.omit(bfi1)
names(bfi2)
```

```
##  [1] "A1" "A2" "A3" "A4" "A5" "C1" "C2" "C3" "C4" "C5" "E1" "E2" "E3" "E4"
## [15] "E5" "N1" "N2" "N3" "N4" "N5" "O1" "O2" "O3" "O4" "O5"
```

```
bfi2 <- scale(bfi2)
dim(bfi2)
```

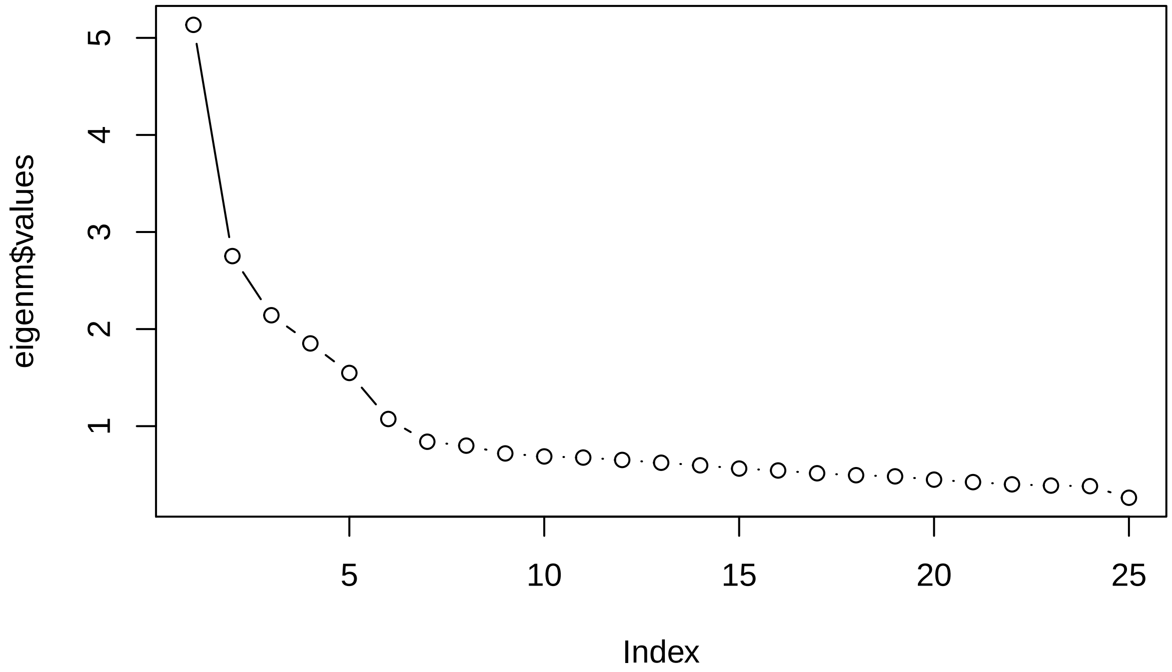
```
## [1] 2436   25
```

#1

```
covm <- cov(bfi2)
eigenm <- eigen(covm)
eigen1 <- eigenm$vectors[,1]
eigen1
```

```
##  [1] -0.11020184  0.21890238  0.24587240  0.20180378  0.27036361
##  [6]  0.16582935  0.15916572  0.15453267 -0.21742086 -0.23117092
## [11] -0.19552510 -0.28337131  0.24841978  0.27470650  0.24401690
## [16] -0.19287599 -0.18741480 -0.18313719 -0.24160999 -0.16286763
## [21]  0.15948516 -0.09673327  0.19150231 -0.02955231 -0.09941144
```

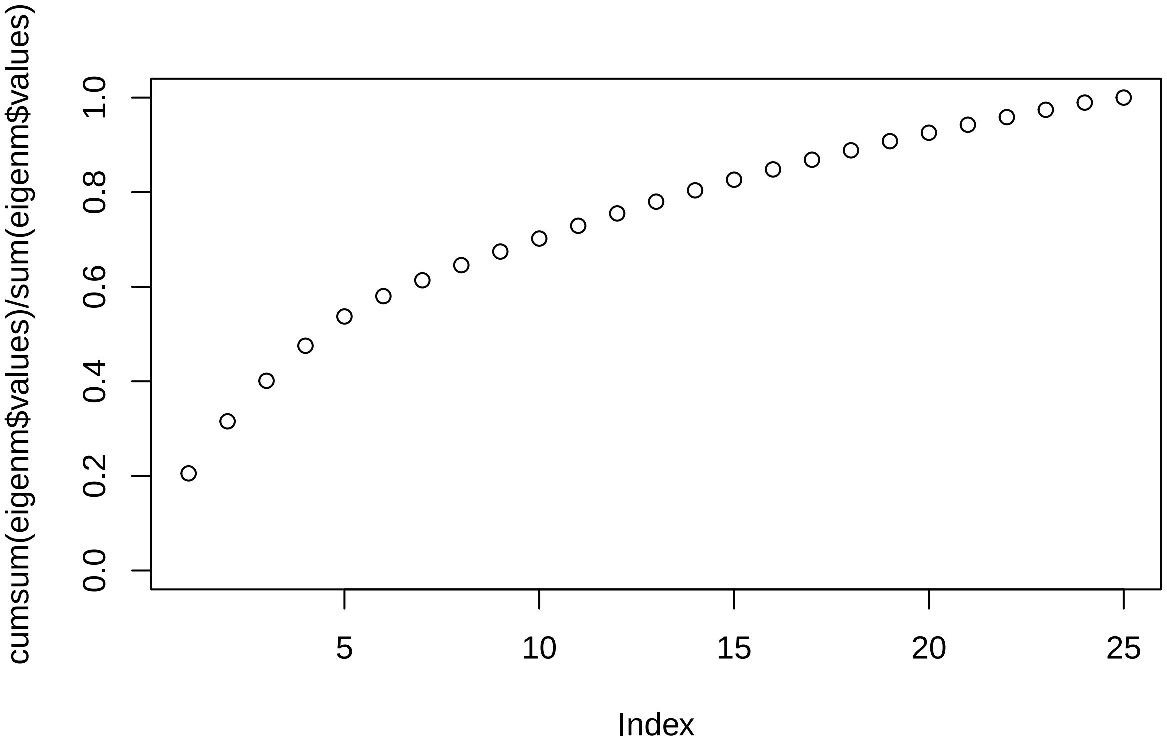
```
plot(eigenm$values, type = "b")
```



As per the elbow test, it looks like 5 factors can be retained.

#2

```
plot(cumsum(eigenm$values)/sum(eigenm$values), ylim=c(0,1))
```



From the plot, 4 factors are needed to explain 50% of the total variance.

#3

```
#install.packages("GPArotation")
fact <- fa(bfi2, nfactors = 2)
```

```
## Loading required namespace: GPArotation
```

```
fact1 <- fact$loadings[,1]
fact1[order(fact1)]
```

```
##          E2          E1          C5          C4          A1          O5 
## -0.52757969 -0.44042558 -0.30664324 -0.30441315 -0.19132123 -0.18779463 
##          N4          O2          N5          N1          N2          N3 
## -0.17928388 -0.11625500 -0.02895752  0.02746048  0.03353453  0.05109512 
##          O4          C3          C1          C2          O1          A4 
##  0.09560663  0.28282287  0.34629040  0.36665443  0.38092907  0.41014682 
##          O3          A2          A5          E4          E5          A3 
##  0.48829177  0.54312988  0.58587301  0.58982630  0.60087501  0.60760897 
##          E3 
##  0.64126300
```

```
fact2 <- fact$loadings[,2]
fact2[order(fact2)]
```

```
##            E4            C3            A5            A4            C1
## -0.0821545514 -0.0767611374 -0.0696489386 -0.0623811621 -0.0191903873
##            E1            C2            O5            O1            A3
## -0.0006471381  0.0345683940  0.0348966895  0.0474700009  0.0552157977
##            E5            A2            A1            O3            E3
##  0.0594313765  0.0667876425  0.0733206191  0.0906447152  0.0911005036
##            O2            E2            O4            C4            C5
##  0.1350364437  0.2039751238  0.2490161137  0.2686508437  0.3242850281
##            N5            N4            N2            N3            N1
##  0.5258609569  0.6018685664  0.7492696668  0.7608824955  0.7619744166
```

As we can see from the above results:

The factors look like the principle component which best summarize the data and most likely to have less variability with the data. When we look at the loadings on one end, we can observe that variables suggest extrovertedness and on the complete opposite end are variables suggesting a negative or anxious emotional state, which are 2 opposite states of mind. The factors underlying these variables may be optimism and pessimism.

#4

```r
#2 centers
kout2 <- kmeans(bfi2, centers = 2, nstart = 25)
centroids2 <- kout2$centers
topvars_centroid21 <- centroids2[1,order(centroids2[1,])]
topvars_centroid22 <- centroids2[2,order(centroids2[2,])]
tail(topvars_centroid21)
```

```
##        A2        E5        E3        A3        E4        A5
## 0.3737496 0.3835171 0.4215278 0.4229396 0.4685683 0.4749490
```

```r
tail(topvars_centroid22)
```

```
##        E1        N1        C4        C5        N4        E2
## 0.4037985 0.4117615 0.4298565 0.4509334 0.5035935 0.5800236
```

```r
#3 centers
kout3 <- kmeans(bfi2, centers = 3, nstart = 25)
centroids3 <- kout3$centers
topvars_centroid31 <- centroids3[1,order(centroids3[1,])]
topvars_centroid32 <- centroids3[2,order(centroids3[2,])]
topvars_centroid33 <- centroids3[3,order(centroids3[3,])]
tail(topvars_centroid31)
```

```
##        A3        N4        N5        N2        N1        N3
## 0.3145348 0.5212833 0.5710454 0.7258122 0.7478177 0.7593344
```

```r
tail(topvars_centroid32)
```

```
##        E5        A4        A3        E3        A5        E4
## 0.3367581 0.3447249 0.3566305 0.3776597 0.4649855 0.4820678
```

```
tail(topvars_centroid33)
```

```
##        A1        N4        C4        C5        E1        E2
## 0.2436845 0.3896889 0.4251611 0.4420349 0.6547507 0.7151784
```

```
kout2$tot.withinss
```

```
## [1] 52766.5
```

```
kout3$tot.withinss
```

```
## [1] 49592.21
```

In both analysis we are trying to bucket the data into certain groups that may be unknown in the beginning. Depending on the groups formed, we can draw inferences on the nature of the groups or identify characteristics which make them fall under the same group.

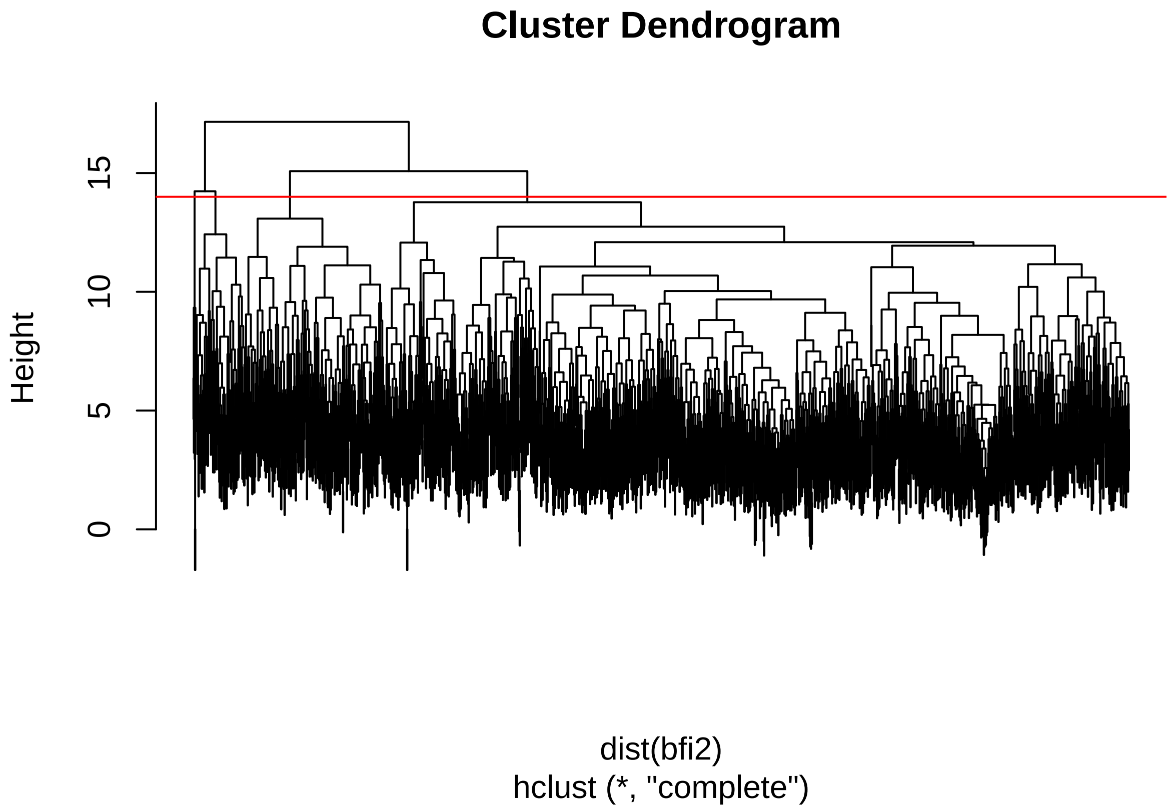
The groups identified in factor analysis comprised of all variables included in the study, differing in their contributions (weightages) towards optimism and pessimism groups (for eg: Optimism would be 99% be defined by positive personality type and 0.7% neutral personality type and 0.3% negative personality type)

The groups identified using cluster analysis strongly restricted the participation of variables to a certain number depending on the proximity to the cluster centroids (like cluster 1 would have only E1, N1, C4, C5, N4 and E2 suggesting a grumpy personality type)

However,the classification using cluster analysis seems to group individuals by the personality types into 1) Extroverted 2) Empathizing 3) Neurotic

#5

```
hout <- hclust(dist(bfi2),method="complete")
plot(hout, labels = FALSE)
abline(a=14,b=0,col="red") #Setting a dissimilarity level of 14
```



Selecting 4 clusters to proceed with as we see that above our threshold line, 4 clusters are coming together to form the hierarchy.

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.6.1
```

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

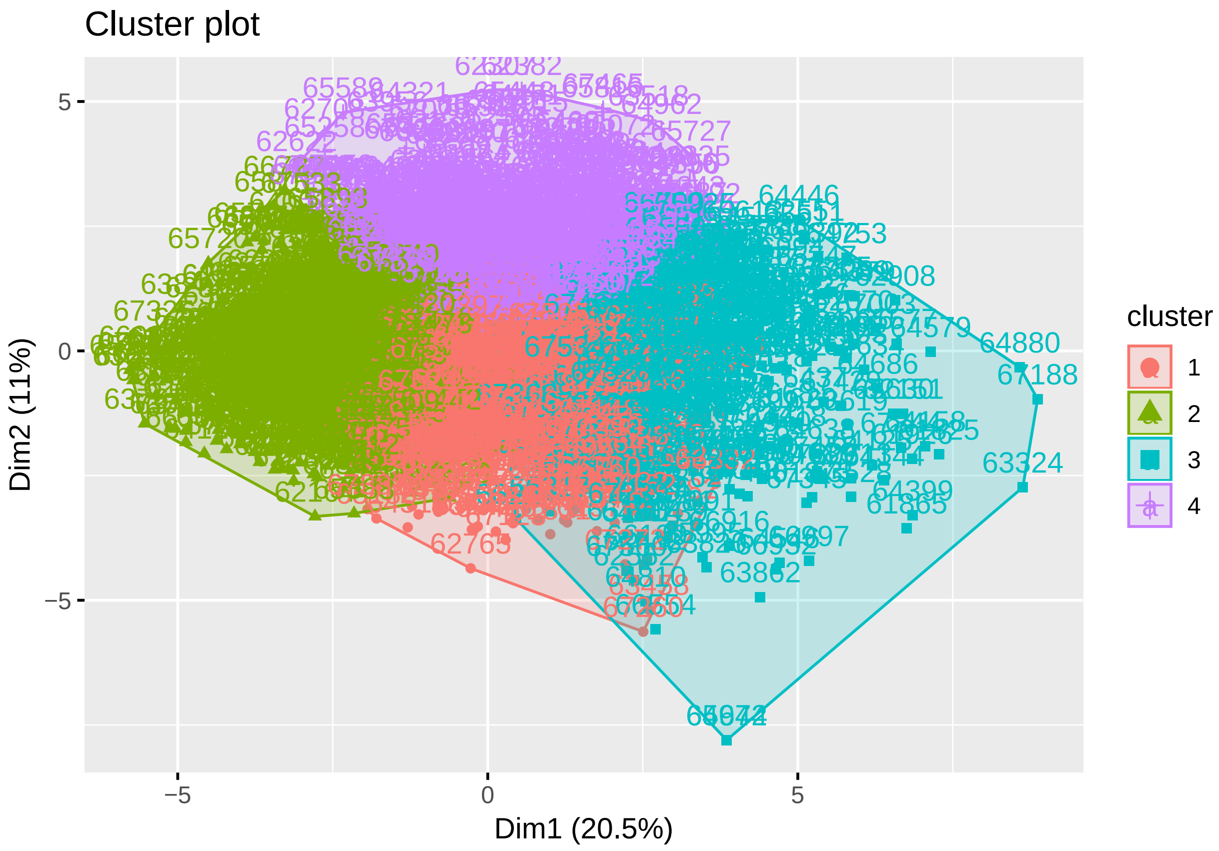
```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

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

```r
kout4 <- kmeans(bfi2, centers = 4, nstart = 25)
fviz_cluster(kout4, data=bfi2)
```



```r
centroids4 <- kout4$centers
topvars_centroid41 <- centroids4[1,order(centroids4[1,])]
topvars_centroid42 <- centroids4[2,order(centroids4[2,])]
topvars_centroid43 <- centroids4[3,order(centroids4[3,])]
topvars_centroid44 <- centroids4[4,order(centroids4[4,])]
tail(topvars_centroid41)
```

```
##         E4         E1         A4         C4         O2         O5
## 0.09549329 0.11039421 0.11243959 0.16006379 0.36162521 0.37766893
```

```r
tail(topvars_centroid42)
```

```
##        A3        O3        E4        E5        A5        E3
## 0.4588839 0.4821680 0.5625301 0.5626976 0.5711970 0.5812381
```

```r
tail(topvars_centroid43)
```

```
##        N2        C4        C5        N4        E1        E2
## 0.3862504 0.4769460 0.5542490 0.6638581 0.8266996 0.9838609
```

```r
tail(topvars_centroid44)
```

```
##        O3        N4        N5        N2        N1        N3
## 0.3621027 0.6183617 0.6377154 0.8142317 0.8312324 0.8406266
```

#6

We learnt from factor analysis that there are primarily 2 extreme groups (positive and negative types/ optimism and pessimism type) under which the personality types were lying. There was demarcation achieved from factor analysis. However, from cluster analysis we learnt that there are groups that lie at the juncture of these 2 extremes and given their central points, varied a lot from the extreme points- pessimism and optimism. These groups lie in the middle of the 2 extreme groups and are a better descriptors of the data. There are always going to be personality types which need not lie in the extreme ends of the spectrum, tehey often turn out to be neutral or rather balanced types.