Homework 7 Solution

Mohsen Nabian

7/09/2015

Using a high-dimensional dataset of your choice, perform a factor analysis and clustering and interpret the results. You may use, for instance, the datasets inside the psych package, such as bfi (25 personality items thought to boil down to a few core personality types) or iqitems (14 scores that are thought to boil down to a few core mental skills), or anything else you can find. (Load the data using, for instance, data(bfi) after loading the psych package; you may need to clean it a bit first with na.omit() to remove the observations with na items, or else impute those missing items.) For the factor analysis, you may use any of the methods covered in the lesson - they should all produce similar results, though princomp and prcomp might be simplest. You don't have to interpret everything, say, fa() outputs, which is a lot of stuff - easier to use str() to examine the output of your function and find the quantities you want.

Q1) After running your factor analysis or PCA, be sure to discuss and interpret your output:

1. Examine the factor eigenvalues or variances (or the sdev or standard deviations as reported by prcomp or princomp, which you then need to square to get the variances). Plot these in a scree plot and use the "elbow" test to guess how many factors one should retain. What proportion of the total variance does your subset of variables explain?
2. Examine the loadings of the factors on the variables (sometimes called the "rotation" in the function output) - ie, the projection of the factors on the variables - focusing on just the first one or two factors. Sort the variables by their loadings, and try to interpret what the first one or two factors "mean." This may require looking more carefully into the dataset to understand exactly what each of the variables were measuring. You can find more about the data in the psych package using ?psych or visiting http://personality-project.org/ .

Solution:

Factor analysis proceadure:

1)Standardize the variables 2)Create the covariance matrix 3)Find the eigenvectors and eigenvalues for that matrix. 4)Choose how many of them we want to keep and analyze.

```
library(psych)  #have some data as well as some machine learning functions like fa()
```

```
## Warning: package 'psych' was built under R version 3.1.3
```

```
data(bfi)
data(bfi.dictionary)
bfi.dictionary
```

```
##          ItemLabel                                    Item     Giant3
## A1           q_146  Am indifferent to the feelings of others.   Cohesion
## A2          q_1162          Inquire about others' well-being.   Cohesion
## A3          q_1206               Know how to comfort others.   Cohesion
## A4          q_1364                             Love children.   Cohesion
## A5          q_1419                  Make people feel at ease.   Cohesion
## C1           q_124                   Am exacting in my work.  Stability
## C2           q_530         Continue until everything is perfect.  Stability
```

```
## C3            q_619            Do things according to a plan.  Stability
## C4            q_626           Do things in a half-way manner.  Stability
## C5           q_1949                            Waste my time.  Stability
## E1            q_712                         Don't talk a lot. Plasticity
## E2            q_901      Find it difficult to approach others. Plasticity
## E3           q_1205            Know how to captivate people. Plasticity
## E4           q_1410                      Make friends easily. Plasticity
## E5           q_1768                             Take charge. Plasticity
## N1            q_952                        Get angry easily.  Stability
## N2            q_974                    Get irritated easily.  Stability
## N3           q_1099               Have frequent mood swings.  Stability
## N4           q_1479                          Often feel blue.  Stability
## N5           q_1505                           Panic easily.  Stability
## O1            q_128                         Am full of ideas. Plasticity
## O2            q_316          Avoid difficult reading material. Plasticity
## O3            q_492  Carry the conversation to a higher level. Plasticity
## O4           q_1738           Spend time reflecting on things. Plasticity
## O5           q_1964       Will not probe deeply into a subject. Plasticity
## gender       gender                          males=1, females=2       <NA>
## education education in HS, fin HS, coll,  coll grad , grad deg       <NA>
## age             age                                age in years       <NA>
##                          Big6        Little12 Keying IPIP100
## A1              Agreeableness      Compassion     -1    B5:A
## A2              Agreeableness      Compassion      1    B5:A
## A3              Agreeableness      Compassion      1    B5:A
## A4              Agreeableness      Compassion      1    B5:A
## A5              Agreeableness      Compassion      1    B5:A
## C1          Conscientiousness     Orderliness      1    B5:C
## C2          Conscientiousness     Orderliness      1    B5:C
## C3          Conscientiousness     Orderliness      1    B5:C
## C4          Conscientiousness Industriousness     -1    B5:C
## C5          Conscientiousness Industriousness     -1    B5:C
## E1               Extraversion     Sociability     -1    B5:E
## E2               Extraversion     Sociability     -1    B5:E
## E3               Extraversion   Assertiveness      1    B5:E
## E4               Extraversion     Sociability      1    B5:E
## E5               Extraversion   Assertiveness      1    B5:E
## N1         Emotional Stability         Balance     -1    B5:N
## N2         Emotional Stability         Balance     -1    B5:N
## N3         Emotional Stability         Balance     -1    B5:N
## N4         Emotional Stability         Balance     -1    B5:N
## N5         Emotional Stability         Balance     -1    B5:N
## O1                   Openness       Intellect      1    B5:O
## O2                   Openness       Intellect     -1    B5:O
## O3                   Openness       Intellect      1    B5:O
## O4                   Openness        Openness      1    B5:O
## O5                   Openness        Openness     -1    B5:O
## gender                   <NA>            <NA>     NA    <NA>
## education                <NA>            <NA>     NA    <NA>
## age                      <NA>            <NA>     NA    <NA>
```

```
bfi_data_not_scaled <- na.omit(bfi)
varnames<-names(bfi_data_not_scaled)    #keeping the names to use later
```

But we need to linearly scale the data into the same range. To do this, I write the following function.

```
linMap <- function(DF, from, to)   #linear mapping
{

  for (i in 1:ncol(DF))
  {
    x<-DF[,i]
    DF[,i]<-((x - min(x)) / max(x - min(x))) * (to - from) + from

  }

  return(DF)
}
```

I will scale data linearly in [0,100]

```
bfi_data<-linMap(bfi_data_not_scaled,0,100)  #normalized in the scale of -100, 100

names(bfi_data)<-varnames
dim(bfi_data)
```

```
## [1] 2236   28
```

```
head(bfi_data)
```

```
##         A1  A2  A3  A4  A5  C1  C2  C3 C4  C5 E1  E2  E3  E4  E5  N1 N2 N3
## 61623 100 100  80 100  80 100 100 100  0  40 20   0 100  80 100  40 80 20
## 61629  60  40   0  80   0  40  20  60 20  60 40 100  60  20   0 100 40 20
## 61634  60  60  80 100  80  60  40  80 40  20  0  40  20  80  60  40 40 60
## 61640  60  80  20  20   0  80  80  80 20  20 40  60  40 100  80  20 60 20
## 61661   0  80 100  80 100  60  40  20 60  80 20   0  20  80  20  20 20 20
## 61664  20 100  80 100  80  40  80 100 40 100 20  20  60 100 100  60 60 60
##         N4  N5  O1 O2 O3  O4 O5 gender education      age
## 61623  20  40  60 40 80 100  0    100        50 21.68675
## 61629 100  60  40 20 60  80 40      0        25 19.27711
## 61634  20  40  80 40 80 100 40      0         0 21.68675
## 61640  20  40  80 20 80  80 80      0         0 16.86747
## 61661  20  20 100  0 80  80 20      0       100 78.31325
## 61664 100 100 100  0 80 100  0    100        25 28.91566
```

```
tail(bfi_data)
```

```
##       A1 A2 A3  A4 A5  C1  C2 C3  C4  C5 E1 E2  E3  E4  E5 N1  N2  N3  N4
## 67541 60 60 80   0 20  40  80 40 100  60 60 60  60  40  60 60  60 100 100
## 67544 80 80 80 100 80 100 100 20  60  80 80  0  80 100  60 80  60  80  40
## 67547 40 60 40   0 40  80  60 80  40  60 40 80  20  40   0 80 100  80  80
## 67556 20 40 80  20 80  80  80 80   0   0 20 20 100  40 100 40  60  40  40
## 67559 80 20 20  60 60  80  80 80  20 100 20 20  60  80  60 80  80 100  60
## 67560 20 40  0  60 20  80  80 40  40  40 40 40   0  20  20  0  20  20   0
##         N5  O1 O2  O3 O4 O5 gender education      age
```

```
## 67541  60  80 60  80 20 60    100         50 22.89157
## 67544  60  60 80  80 60 60    100         50 22.89157
## 67547 100 100  0  60 80 20    100         75 25.30120
## 67556   0  80  0 100 60 40    100         75 31.32530
## 67559   0  80 20  80 80  0      0         75 33.73494
## 67560   0  40  0  40 80  0    100         75 56.62651
```

now factor analysis:

```r
eigenm <- eigen(cov(bfi_data))      # Calculating Eigen Values and Eigne Vectors
varnames<-names(bfi_data)

################ First Factor
eigen1 <- eigenm$vectors[,1]
factor1<-data.frame(varnames[order(eigen1)],eigen1[order(eigen1)]) # making a data frame putting data i
names(factor1)<-c("variable","coeff")
head(factor1)
```

```
##   variable      coeff
## 1       E4 -0.2721838
## 2       A5 -0.2075926
## 3       E3 -0.2035585
## 4       E5 -0.1995416
## 5       A3 -0.1857474
## 6       A4 -0.1830021
```

```r
################ Second Factor
eigen2 <- eigenm$vectors[,2]
factor2<-data.frame(varnames[order(eigen2)],eigen2[order(eigen2)])
names(factor2)<-c("variable","coeff")
head(factor2)
```

```
##   variable      coeff
## 1   gender -0.6592427
## 2       N5 -0.3143186
## 3       N3 -0.3085925
## 4       N2 -0.2750784
## 5       N1 -0.2686966
## 6       A3 -0.1634220
```

```r
################ Third Factor
eigen3 <- eigenm$vectors[,3]
factor3<-data.frame(varnames[order(eigen3)],eigen3[order(eigen3)])
names(factor3)<-c("variable","coeff")
head(factor3)
```

```
##   variable      coeff
## 1       N1 -0.2933251
## 2       N2 -0.2587770
## 3       E3 -0.2537076
## 4       O3 -0.2457958
## 5       N3 -0.2444709
## 6       E5 -0.2220712
```

```
################ Forth Factor
eigen4 <- eigenm$vectors[,4]
factor4<-data.frame(varnames[order(eigen4)],eigen4[order(eigen4)])
names(factor4)<-c("variable","coeff")
head(factor4)
```

```
##   variable      coeff
## 1       C4 -0.3708560
## 2       C5 -0.3699444
## 3       O2 -0.2769823
## 4       E4 -0.2440007
## 5       O5 -0.1705853
## 6       A5 -0.1259859
```

```
#####################
################ 5th Factor
eigen5 <- eigenm$vectors[,5]
factor5<-data.frame(varnames[order(eigen5)],eigen5[order(eigen5)])
names(factor5)<-c("variable","coeff")
head(factor5)
```

```
##   variable      coeff
## 1       O2 -0.5046488
## 2       O5 -0.3904196
## 3       A4 -0.2396141
## 4       A1 -0.2247871
## 5       C3 -0.2003373
## 6       C2 -0.1876155
```

```
################ 6th Factor
eigen6 <- eigenm$vectors[,6]
factor6<-data.frame(varnames[order(eigen6)],eigen6[order(eigen6)])
names(factor6)<-c("variable","coeff")
head(factor6)
```

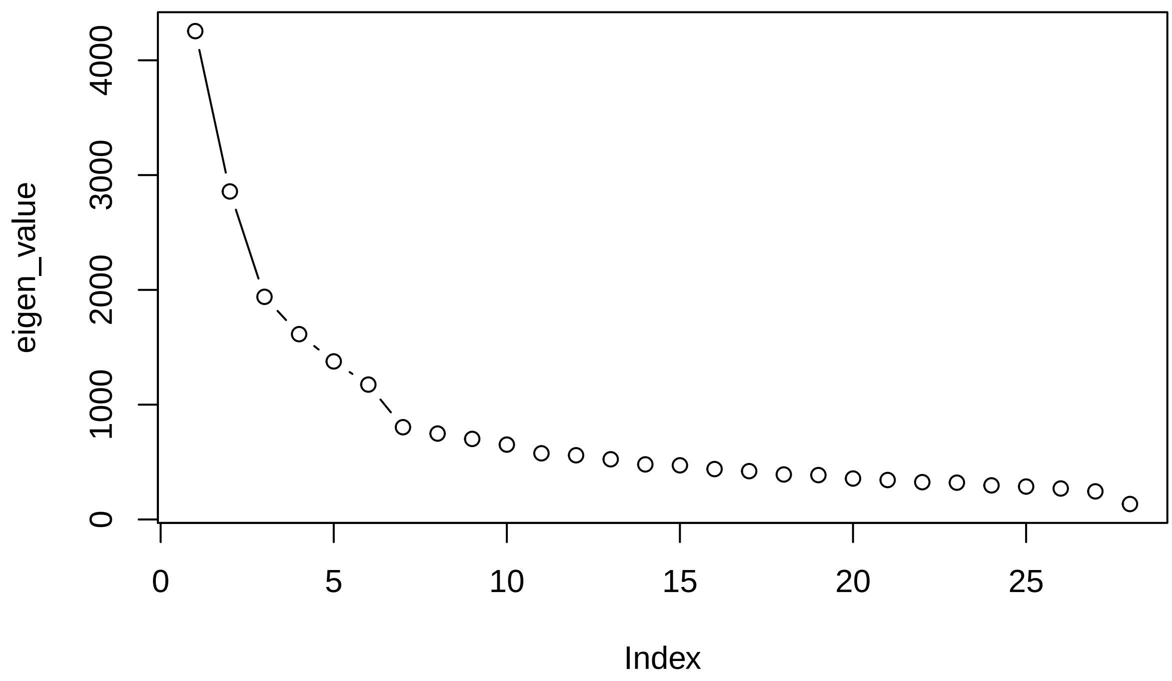
```
##   variable      coeff
## 1       A4 -0.3234662
## 2       A3 -0.3149289
## 3       E1 -0.2940488
## 4       A2 -0.2649651
## 5       E2 -0.2559279
## 6       A5 -0.2480322
```

```
################ 7th Factor
eigen7 <- eigenm$vectors[,7]
factor7<-data.frame(varnames[order(eigen7)],eigen4[order(eigen7)])
names(factor7)<-c("variable","coeff")
head(factor7)
```

```
##   variable       coeff
## 1       A1  0.04250819
```

```
## 2       E3 -0.08992742
## 3       E1  0.30371424
## 4       O1  0.08650949
## 5       C4 -0.37085600
## 6       O3  0.05534729
```

```r
#####################
############Eigen Values
eigen_value<-eigenm$values
plot(eigen_value,type="b")
```



So based on the "Elbow" rule, we would pick 7 factors as our prinipals and assume the rest as noises.

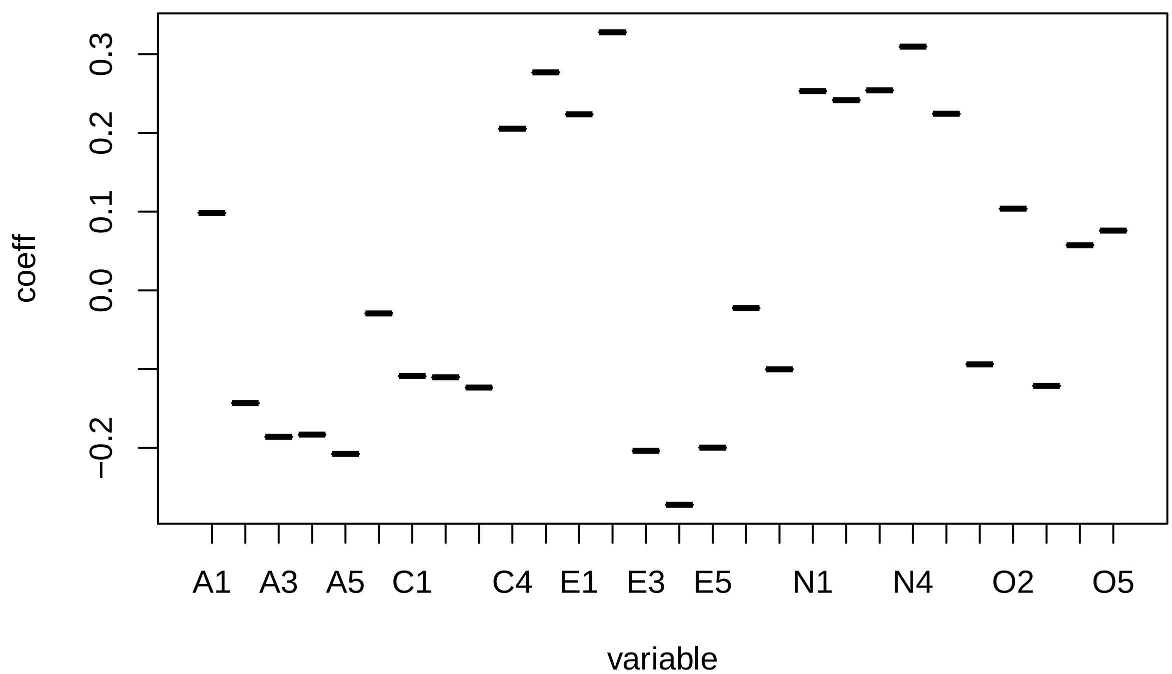
Looking further in factor 1:

```r
factor1
```

```
##     variable       coeff
## 1         E4 -0.27218378
## 2         A5 -0.20759257
## 3         E3 -0.20355852
## 4         E5 -0.19954159
## 5         A3 -0.18574741
## 6         A4 -0.18300208
## 7         A2 -0.14326681
## 8         C3 -0.12329704
## 9         O3 -0.12112937
```

```
## 10        C2 -0.11032735
## 11        C1 -0.10893571
## 12    gender -0.10020384
## 13        O1 -0.09392283
## 14       age -0.02921922
## 15 education -0.02266203
## 16        O4  0.05721164
## 17        O5  0.07596257
## 18        A1  0.09848966
## 19        O2  0.10380727
## 20        C4  0.20530128
## 21        E1  0.22361057
## 22        N5  0.22431097
## 23        N2  0.24159433
## 24        N1  0.25312128
## 25        N3  0.25408639
## 26        C5  0.27689672
## 27        N4  0.30950581
## 28        E2  0.32780915
```

```r
plot(factor1)
```

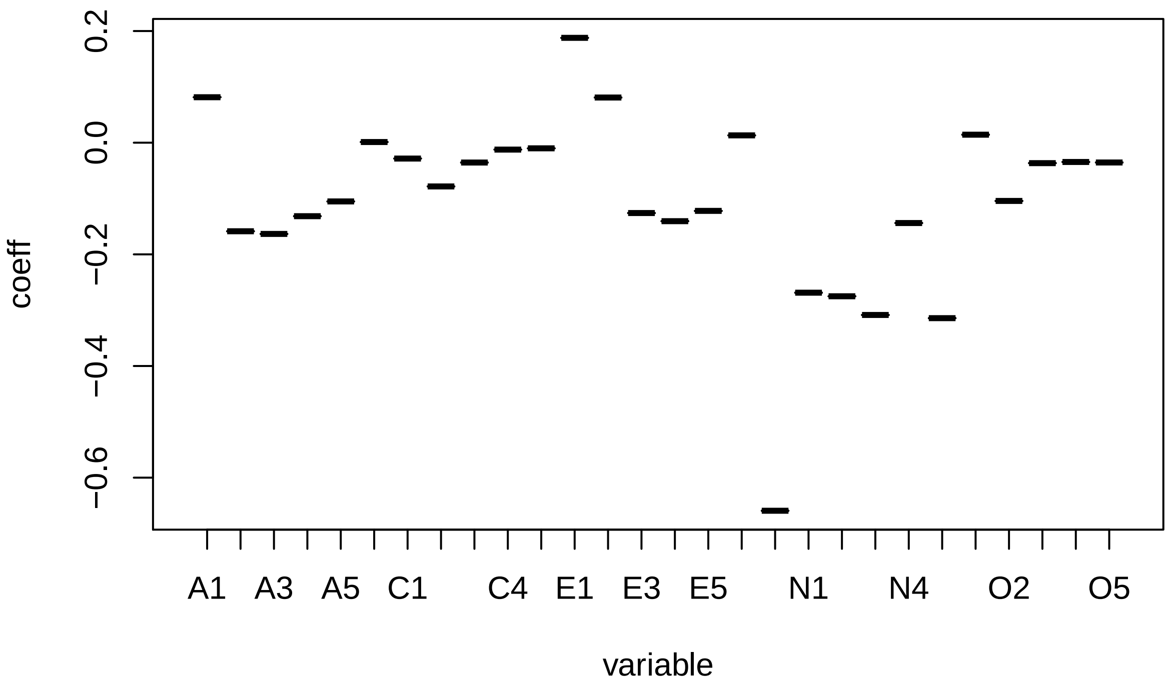


Factor 1: High in E4 and E2 That means peoples personality is dominantly dependent on the factor that wether they are socail and make friends or they are not good at socializing.

```r
factor2
```

```
##     variable       coeff
## 1     gender -0.65924268
## 2         N5 -0.31431864
## 3         N3 -0.30859246
## 4         N2 -0.27507843
## 5         N1 -0.26869658
## 6         A3 -0.16342203
## 7         A2 -0.15862877
## 8         N4 -0.14392256
## 9         E4 -0.14059361
## 10        A4 -0.13166272
## 11        E3 -0.12600292
## 12        E5 -0.12216063
## 13        A5 -0.10517137
## 14        O2 -0.10433846
## 15        C2 -0.07835635
## 16        O3 -0.03663671
## 17        C3 -0.03555257
## 18        O5 -0.03547111
## 19        O4 -0.03453805
## 20        C1 -0.02842378
## 21        C4 -0.01236368
## 22        C5 -0.01010810
## 23       age  0.00118689
## 24 education  0.01312702
## 25        O1  0.01434098
## 26        E2  0.08088405
## 27        A1  0.08142249
## 28        E1  0.18778631
```

```r
plot(factor2)
```



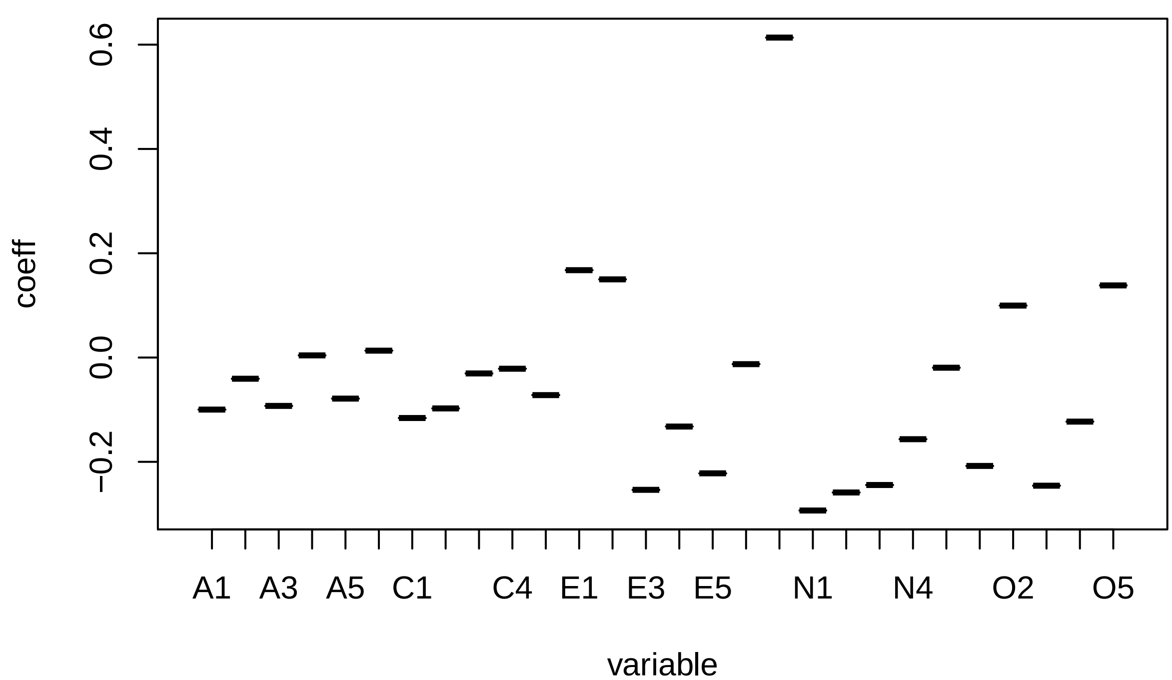
Factor 2: Gender in this facor plays a dominant role. That means gender could paly a deteministic role in people's personality characterizations.

```
factor3
```

```
##     variable        coeff
## 1         N1 -0.293325055
## 2         N2 -0.258776994
## 3         E3 -0.253707551
## 4         O3 -0.245795750
## 5         N3 -0.244470911
## 6         E5 -0.222071158
## 7         O1 -0.207912284
## 8         N4 -0.156591934
## 9         E4 -0.132344701
## 10        O4 -0.122899773
## 11        C1 -0.115965720
## 12        A1 -0.099819083
## 13        C2 -0.097603684
## 14        A3 -0.092751085
## 15        A5 -0.078776571
## 16        C5 -0.072086491
## 17        A2 -0.040686075
## 18        C3 -0.030476161
## 19        C4 -0.021489283
## 20        N5 -0.019560622
## 21 education -0.012670655
```

```
## 22         A4  0.004155159
## 23        age  0.013197381
## 24         O2  0.099608351
## 25         O5  0.138333522
## 26         E2  0.149957329
## 27         E1  0.167500003
## 28     gender  0.613544477
```

```
plot(factor3)
```



Factor3: In this factor also gender plays the dominant role.

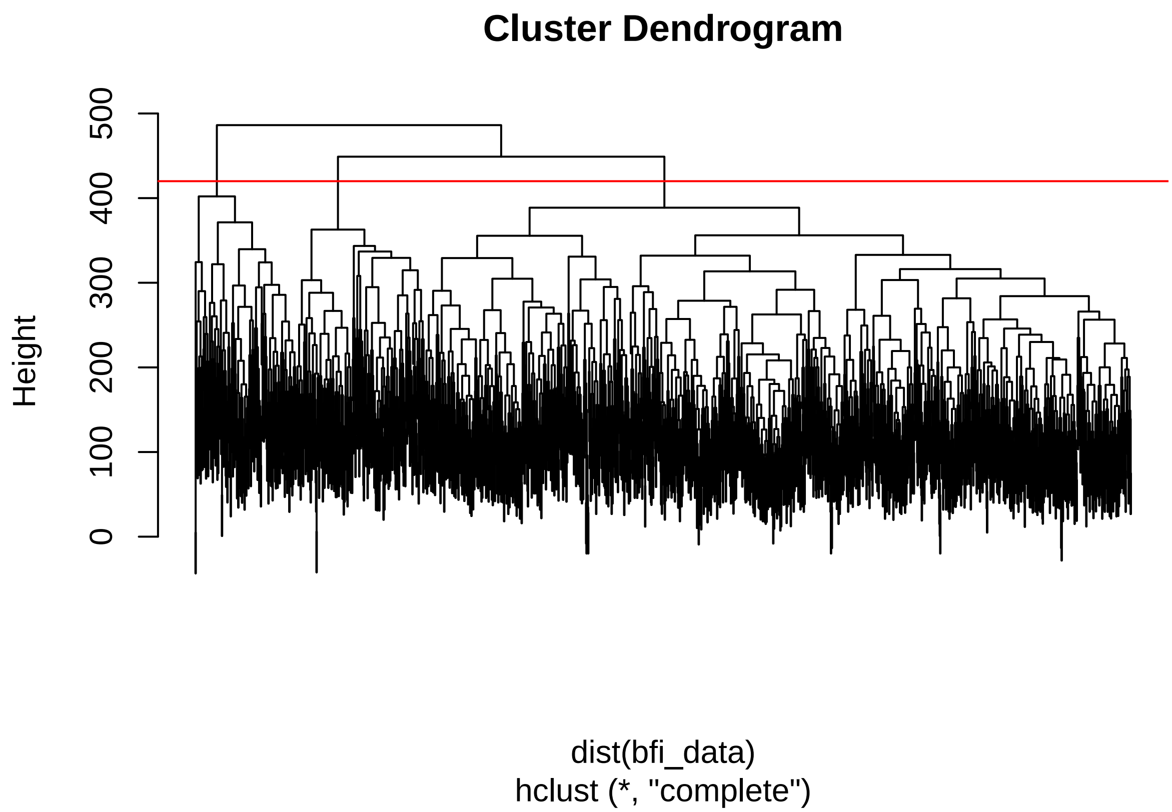
Summary: In this analysis we might say peoples personality Questionnaire highly revolves around the gender and social skills questions and these are key questions.

Q2) Next perform a cluster analysis of the same data. 3. First use k-means and examine the centers of the first two or three clusters. How are they similar to and different from the factor loadings of the first couple factors? 4. Next use hierarchical clustering. Print the dendrogram, and use that to guide your choice of the number of clusters. Use cutree to generate a list of which clusters each observation belongs to. Aggregate the data by cluster and then examine those centers (the aggregate means) as you did in (3). Can you interpret all of them meaningfully using the methods from (3) to look at the centers? 5. From the factor and cluster analysis, what can you say more generally about what you have learned about your data?

```
hout <- hclust(dist(bfi_data),method="complete") #or method="average" or...
plot(hout,labels=FALSE)

#To get the cluster assignments, we just apply cutree to the hclust output,
```

```
#choose either the height to cut it at, or the number of clusters:
#as.vector(cutree(hout,h=420)) # cut the plot at height=21
abline(a=420,b=0,col="red")
```



According to the plot,having 3 clusters seems to be reasonable. Lets do the K-means with 3 clusters:

```
set.seed(100)

kout <- kmeans(bfi_data,centers=3,nstart=30)  #means 30 times change the random initialization of the c
                                              # and choose the best one that has minimum

centroids <- kout$centers

topvars_centroid1 <- centroids[1,order(centroids[1,])]
topvars_centroid2 <- centroids[2,order(centroids[2,])]
topvars_centroid3 <- centroids[3,order(centroids[3,])]



topvars_centroid1
```

```
##        C4       A1       N1       E2       N4       O5       N5
##  17.03617 20.18670 20.77013 22.98716 23.54726 24.52742 25.01750
##        O2       E1       N3       C5       N2      age education
##  26.60443 26.79113 27.60793 27.67795 33.93232 34.08500  54.78413
```

```
##         E3        C3        C2        O4        O3        C1        E5
##   71.50525  73.11552  74.00233  75.70595  76.07935  77.08285  78.90315
##         O1        A3        E4        A5        A2        A4    gender
##   82.12369  83.50058  83.92065  84.17736  85.39090  85.50758  85.76429
```

```
topvars_centroid2
```

```
##     gender         O5        age         N5         O2         A1
##  0.1631321 29.3637847 30.3779555 32.9200653 34.0946166 36.5089723
##         C4         N1         N3         N4         N2         E1
## 36.8026101 41.1092985 43.0668842 50.0489396 50.2446982 51.3539967
##         E2         E3  education         C5         E4         E5
## 53.1158238 53.6704731 54.2006525 55.4649266 58.8254486 60.6851550
##         C3         C2         A3         A5         A4         A2
## 62.6753670 63.2300163 63.3278956 63.4584013 65.9380098 66.9494290
##         C1         O3         O1         O4
## 68.4828711 68.9722675 78.0097879 81.4029364
```

```
topvars_centroid3
```

```
##         A1       age        O5        C4        O2        E1        E3
##   27.91123  30.79053  34.02089  39.11227  41.56658  43.91645  52.76762
## education        N1        E2        C5        E4        N5        N4
##   55.25457  55.27415  55.61358  56.31854  59.32115  59.63446  60.60052
##         C3        E5        N3        O3        A5        C2        N2
##   62.03655  62.74151  63.00261  63.02872  64.33420  65.16971  66.94517
##         C1        A3        O1        A4        A2        O4    gender
##   67.36292  67.78068  68.79896  70.46997  74.72585  80.65274  99.86945
```

cluster 1 and 3 are specified for men. and cluster 2 is for women. according to cluster 1, A4, A2 A5 and E4,A3 are high for men. That means these men are: 1)loving children 2)Inquire about others' well-being 3)Make people feel at ease 4)Make friends easily 5)Know how to comfort others These Men are so social and caring about others.

however, O4 and O1 are high in women. These women are: 1) Spending time reflecting on things 2) full of ideas apparently alot of these women are passionate about their future, thinking alot and not so social.

The 3rd cluster which is again for men:

They are high in O4 and significantly less in A4,A2 and A5 These means that these men are 1)full of ideas 2) Less social These men are more serious about the life.

Summary:

Factor analysis and Cluster analysis in this study provides the fact that people's personality could be divided majorly based on their gender and social skills. Those having the same sex and social skills would be having almost the same other personality characteristics. Also cluster analysis, and the hierarchy plot, demonstrated us that socialability in men are forming their other characteristics which is not the case in women as much.