DSSH 6301 - HW 11 Solutions

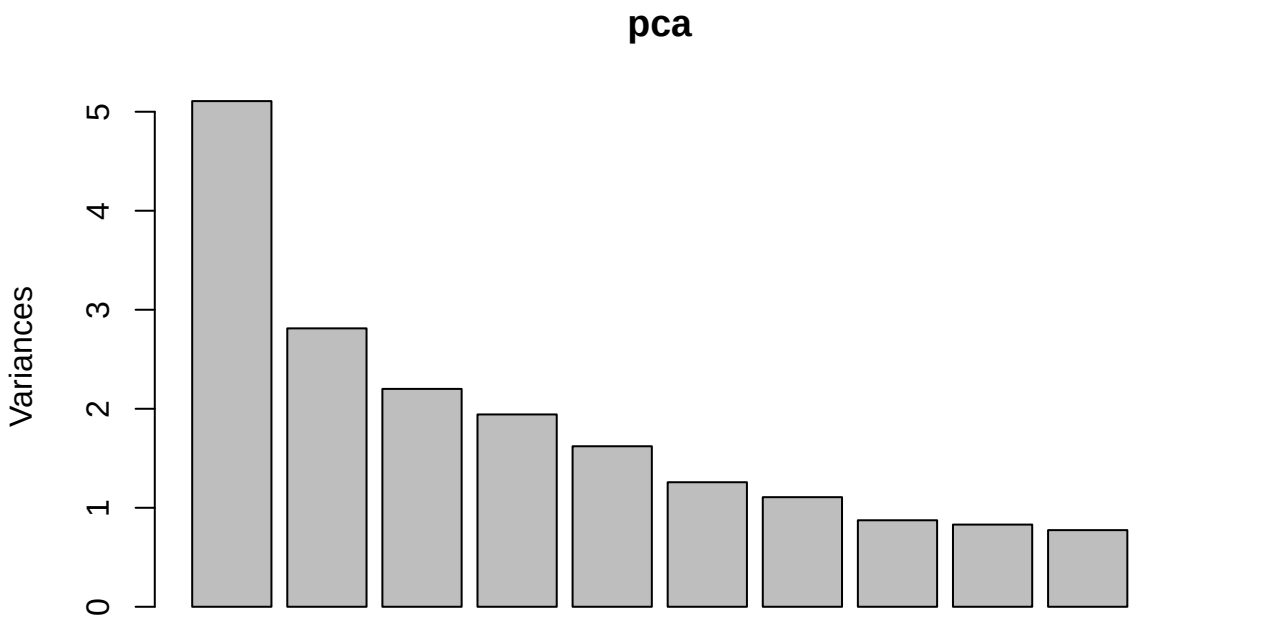
Load the data.

```
data(bfi, package="psych")
head(bfi)
```

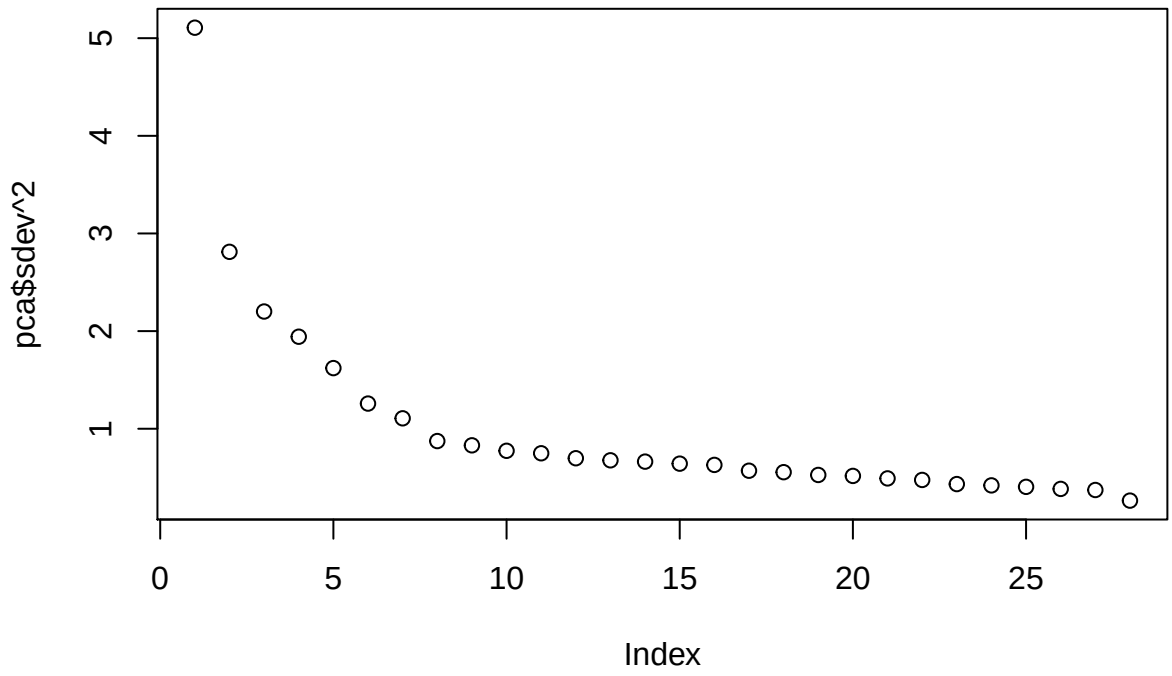
```
##       A1 A2 A3 A4 A5 C1 C2 C3 C4 C5 E1 E2 E3 E4 E5 N1 N2 N3 N4 N5 O1 O2 O3
## 61617  2  4  3  4  4  2  3  3  4  4  3  3  3  4  4  3  4  2  2  3  3  6  3
## 61618  2  4  5  2  5  5  4  4  3  4  1  1  6  4  3  3  3  3  5  5  4  2  4
## 61620  5  4  5  4  4  4  5  4  2  5  2  4  4  4  5  4  5  4  2  3  4  2  5
## 61621  4  4  6  5  5  4  4  3  5  5  5  3  4  4  4  2  5  2  4  1  3  3  4
## 61622  2  3  3  4  5  4  4  5  3  2  2  2  5  4  5  2  3  4  4  3  3  3  4
## 61623  6  6  5  6  5  6  6  6  1  3  2  1  6  5  6  3  5  2  2  3  4  3  5
##       O4 O5 gender education age
## 61617  4  3      1        NA  16
## 61618  3  3      2        NA  18
## 61620  5  2      2        NA  17
## 61621  3  5      2        NA  17
## 61622  3  3      1        NA  17
## 61623  6  1      2         3  21
```

Problem 1

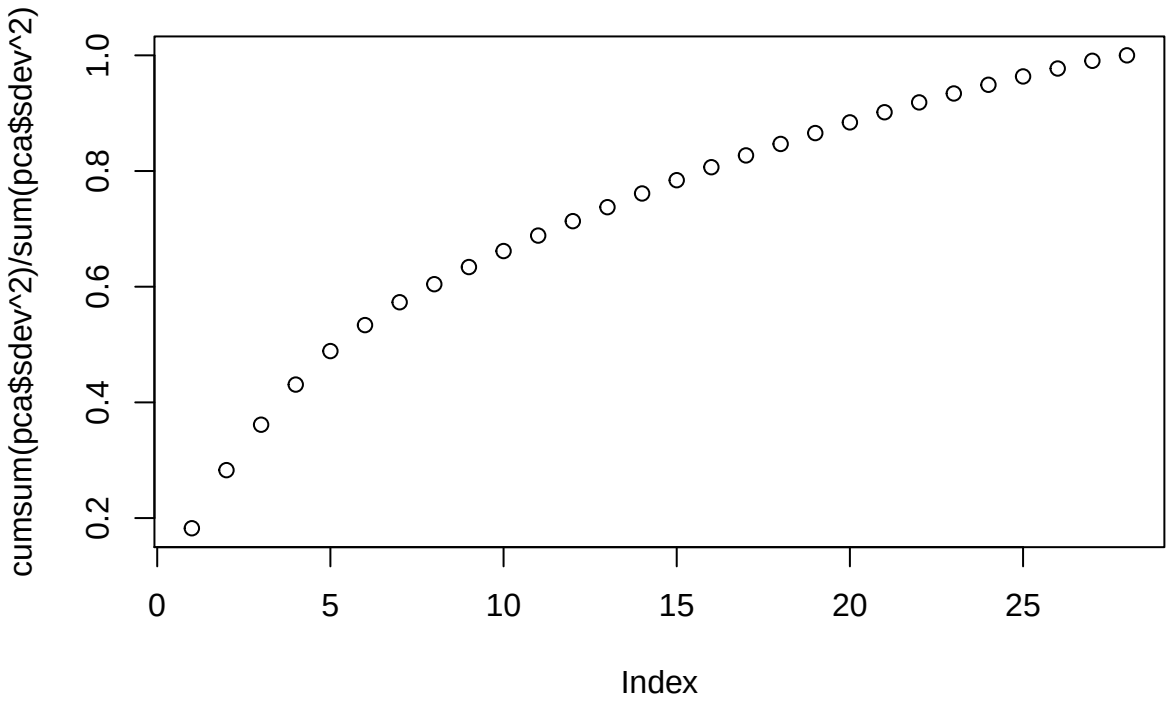
```
# First we clean up the data and rescale it for easier comparison between variables.
bfi <- na.omit(bfi)
bfi <- scale(bfi)
bfi <- as.data.frame(bfi)

pca <- prcomp(bfi)

# Two ways to create a scree plot
screeplot(pca)
```




```
plot(pca$sdev^2)
```




```
plot(cumsum(pca$sdev^2)/sum(pca$sdev^2))
```



For these data, there seems to be a bit of an elbow at about factor 7. Including the first 7 factors explains almost 60% of the total variance.

Problem 2

```
pc1_rotation <- pca$rotation[, 1]
sort(pc1_rotation)
```

```
##          E2          N4          C5          C4          E1          N1
## -0.28568502 -0.24234436 -0.22914087 -0.21755803 -0.19933450 -0.18661712
##          N2          N3          N5          A1          O5          O2
## -0.18053429 -0.17839176 -0.16249313 -0.10986654 -0.09890417 -0.09554183
##          O4   education      gender         age          C2          O1
## -0.03975386  0.03177401  0.05394314  0.07686110  0.15360272  0.15698527
##          C3          C1          O3          A4          A2          A3
##  0.15759463  0.15954124  0.18566786  0.19659188  0.21633627  0.24419609
##          E5          E3          A5          E4
##  0.24891167  0.25260715  0.26649341  0.27781216
```

Looking at the variable codings, we can see that attributes that have to deal with extroverts are given a large positive score (e.g. E4 - makes friends easily) and those that are more associated with introverts are given a large negative score (e.g. E2 - difficult to approach others). It appears this factor could be used to separate people into these two categories.

```
pc1_rotation <- pca$rotation[, 2]
sort(pc1_rotation)
```

```
##          N3          N1          N2          N5          N4          E3
```

```
## -0.40626717 -0.39152283 -0.39042626 -0.31991366 -0.27351771 -0.20578109
##          A2          A3          E5          O4      gender          O3
## -0.20556735 -0.20484516 -0.17580509 -0.16950430 -0.16263249 -0.15673712
##          A5          E4          C2          C5          A4          O1
## -0.12265691 -0.11747312 -0.11204932 -0.10870632 -0.09395125 -0.09335734
##          C4          O2          C1          C3          O5   education
## -0.09237747 -0.08661006 -0.06757389 -0.02590981  0.01091388  0.01317862
##          E2          A1         age          E1
##  0.02524469  0.03305166  0.04267018  0.13773994
```

The second factor by contrast seems to be moody vs withdrawn based on the highest positive and negative loading variables.

Problem 3

```
kout <- kmeans(bfi, centers=2, nstart=25)

centroids <- kout$centers

topvars_centroid1 <- sort(centroids[1, ])
topvars_centroid2 <- sort(centroids[2, ])

tail(topvars_centroid1)
```

```
##        N1        C4        E1        C5        N4        E2
## 0.4176598 0.4296434 0.4339739 0.4544198 0.5178946 0.6184365
```

```
tail(topvars_centroid2)
```

```
##        A2        E5        A3        E3        A5        E4
## 0.3526705 0.3746481 0.4078066 0.4181671 0.4501740 0.4568777
```

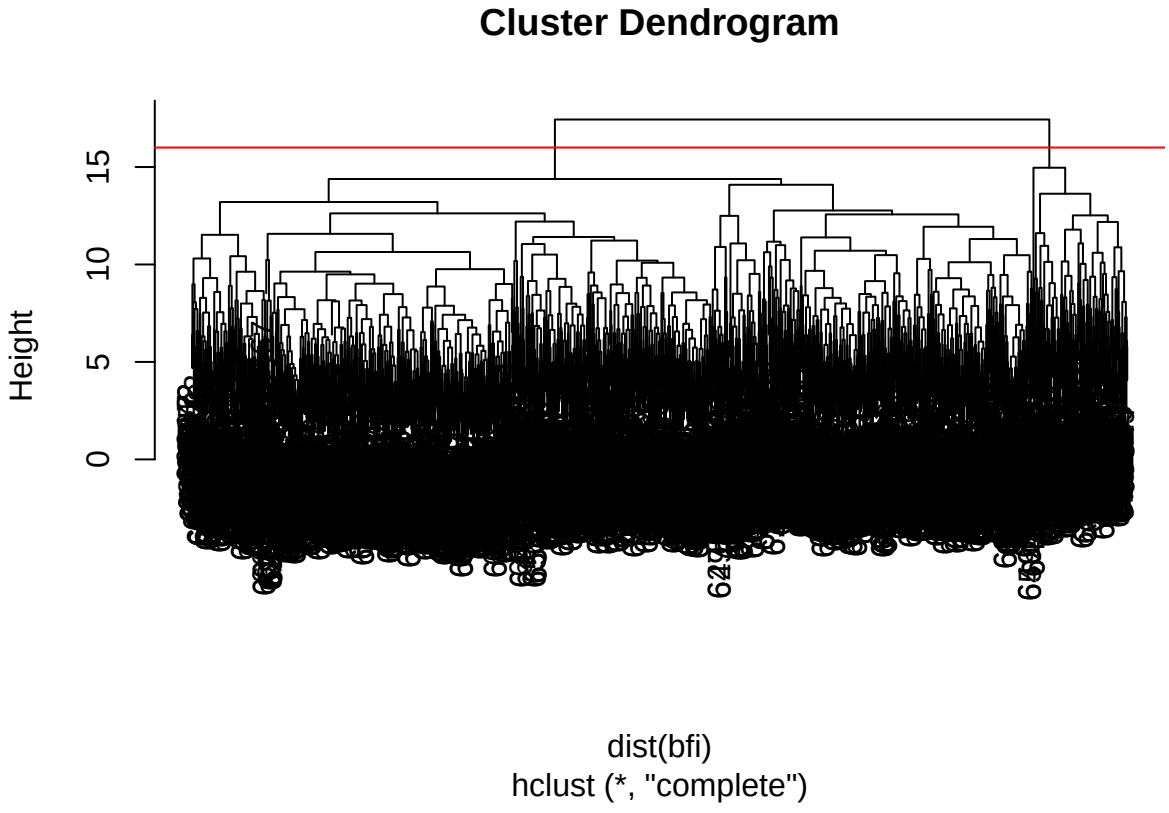
```
sort(pca$rotation[, 1])
```

```
##          E2          N4          C5          C4          E1          N1
## -0.28568502 -0.24234436 -0.22914087 -0.21755803 -0.19933450 -0.18661712
##          N2          N3          N5          A1          O5          O2
## -0.18053429 -0.17839176 -0.16249313 -0.10986654 -0.09890417 -0.09554183
##          O4   education      gender         age          C2          O1
## -0.03975386  0.03177401  0.05394314  0.07686110  0.15360272  0.15698527
##          C3          C1          O3          A4          A2          A3
##  0.15759463  0.15954124  0.18566786  0.19659188  0.21633627  0.24419609
##          E5          E3          A5          E4
##  0.24891167  0.25260715  0.26649341  0.27781216
```

The 1st cluster captures variables that are negative on the 1st principal component. The 2nd cluster captures variables that are postive on the 1st principal component. It is often the case that first two clusters will comprise the points on either end of the first factor.

Problem 4

```
hout <- hclust(dist(bfi), method="complete")
plot(hout)

abline(h=16, col="red")
```




```
cut <- cutree(hout, k=2)

clust_means <- aggregate(bfi, by=list(cut), FUN=mean)

tail(unlist(sort(clust_means[1, names(clust_means)!="Group.1"])))
```

```
##        E3        E5        A5        A2        A3        E4
## 0.1149700 0.1161775 0.1312102 0.1322361 0.1332015 0.1450269
```

```
tail(topvars_centroid2)
```

```
##        A2        E5        A3        E3        A5        E4
## 0.3526705 0.3746481 0.4078066 0.4181671 0.4501740 0.4568777
```

```
tail(unlist(sort(clust_means[2, names(clust_means)!="Group.1"])))
```

```
##        A1        C5        C4        N4        E1        E2
## 0.4304819 0.4569451 0.5025960 0.5993251 1.0223622 1.0901949
```

```
tail(topvars_centroid1)
```

```
##        N1        C4        E1        C5        N4        E2
## 0.4176598 0.4296434 0.4339739 0.4544198 0.5178946 0.6184365
```

The output above is grouped with results most similar from 3. We can see that similar variables are found as high scoring using kmeans and the clustered output from cutree.

Problem 5

There appears to be a few underlying personality types that explain much of the variation in the data. Traditionally, modern personality analysis group these variables into the "Big 5" clusters. However, as you can see from these data, it is much more of a judgment call whether we choose to go with two, five, or seven factors or clusters. In this case it even depends on whether you include the demographic variables (age, gender, education) or not. When you look under the hood and see how these things are done, you can learn a lot about how even sophisticated scientists will sweep the messy details under the carpet in order to produce a tidy model that is easily explained to others.