STAA 574: Homework 5

Spring 2020

Jon Dollard

Use R-Markdown to organize your work.

You will need to carry out a factor analysis on two data sets. You should type up approximately a 1-2 page summary. At a minimum, this will include:

- Give the correlation matrix used for the analysis.
- Run the factor analysis with and without using varimax rotation and compare the results.
- Describe the criteria you used for how many factors you decide to keep.
- Make a table that includes the factor loadings and communalities for the variables. Also, note the eigenvalues and the proportion of variance accounted for by each factor. This table will be similar ot the one constructed on page 116 of the notes.

The data sets that are to be analyzed are on Canvas and are:

1. Flea Beatle Data
2. Engineer Data

Each dataset has two groups. Run a factor analysis on each group individually for each dataset. Compare your results from the two groups. Note that measurements are not give on the engineer dataset, so you won't be able to interpret the results as thoroughly.

Use at least two factors in each model. Also, note the that psych package will sometimes give errors if you try to use more than one factor. If this is the case, use a different package or do the factor analysis manually as shown in the notes.

We will begin our factor analysis by reading the data into R for analysis. Our approach for this factor analysis will be to calculate the correlation matrix for the flea beetle data and look for the higher correlations that will indicate to us variables that form groups (variables that are highly correlated). From the correlation matrix we should be able to make a decision about a reasonable number of groups that exist and number of factors to use. One of the goals of factor analysis, similar to principal component analysis, is dimension reduction. Dimension reduction is reasonable if we see a high degree of correlation in the data set variables. Once we have decided if factor analysis seems reasonable and how many factors make sense based on correlation groupings we can complete the factor analysis calculations using R. From the calculations we will attempt to interpret the factors in the context of the data set and provide an interpretability conclusion.

Please note that for this assignment I collaborated with fellow student Dilyara Murtazina. Specifically, I have used code that she shared with me to help me be able to construct professional looking tables for the factors, eigenvalues, and cumulative proportion of variance. I want to acknowledge her collaboration and contribution to this HW assignment and report.

Factor Analysis for Haltica Oleracea Flea Beetle Data:

Begin by reading the data for the H. oleracea into R and manipulating it into a numeric matrix format. It is good practice to look at the first few lines of the dataset to ensure everything looks correct. We complete this using the head() function in R.

```r
#read in the flea beetle data set and manipulate it to a usable form
Flea_Beetle_Data <- read_excel("fleabeetledata.xlsx")
```

```
New names:
* `` -> ...1
* x1 -> x1...2
* x2 -> x2...3
* x3 -> x3...4
* x4 -> x4...5
* ... and 4 more problems
```

```r
Haltica_Oleracea_Data <- Flea_Beetle_Data[1:19,2:5]
Haltica_Oleracea_Data <- data.frame(Haltica_Oleracea_Data)
colnames(Haltica_Oleracea_Data) <- c("TG", "Elytra", "Second Antenna","Third Antenna")
head(Haltica_Oleracea_Data)
```

```
   TG Elytra Second Antenna Third Antenna
1 189    245            137           163
2 192    260            132           217
3 217    276            141           192
4 221    299            142           213
5 171    239            128           158
6 192    262            147           173
```

Next we calculate the correlation matrix from the data set. It is shown below.

```r
#calculate the correlation matrix for the Haltica oleracea data set
HO_corr <- cor(Haltica_Oleracea_Data)
round(HO_corr,3)
```

```
##                   TG Elytra Second Antenna Third Antenna
## TG             1.000  0.695          0.434         0.535
## Elytra         0.695  1.000          0.502         0.413
## Second Antenna 0.434  0.502          1.000         0.129
## Third Antenna  0.535  0.413          0.129         1.000
```

$$\mathbf{R}_{H.oleracea} = \begin{bmatrix} 1.000 & 0.695 & 0.434 & 0.535 \\ 0.695 & 1.000 & 0.502 & 0.413 \\ 0.434 & 0.502 & 1.000 & 0.129 \\ 0.535 & 0.413 & 0.129 & 1.000 \end{bmatrix}$$

If we look at the correlation matrix above we can see that TG and Elytra show the largest correlation in this data set and could be viewed as a grouping. A second grouping seems plausible between Elytra and the Second Antenna and Third Antenna variables. While the correlations for those variables are noticeably less than TG and Elytra they still seem significant enough to consider them as a second group. Therefore, for the factor analysis of the Haltica oleracea data set I chose 2 factors.

There are 3 factor analysis estimation methods we could consider for this data. Principal component (PC) and principal factor (PF) methods are similar and an extension of the principal components analysis (PCA) methodology. The maximum likelihood estimation (MLE) method is the third potential method and assumes the data is distributed multivariate normal (MVN). For the H. oleracea data I chose to use the PC method for estimation. The "psych" library in R contains a very useful function called principal() that I will use with 2 factors to complete the FA calculations.

Using the principal function with no rotation of the the PCs we obtain the following.

```
#use the principal components method to perform factor analysis of the Haltica oleracea data set
#use the function principal found in the "psych" library
#no rotation

FA_HO <- principal(Haltica_Oleracea_Data, nfactors = 2, rotate = "none")
FA_HO
```

```
## Principal Components Analysis
## Call: principal(r = Haltica_Oleracea_Data, nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 PC1   PC2   h2   u2 com
## TG             0.89 -0.11 0.80 0.20   1
## Elytra         0.87  0.11 0.77 0.23   1
## Second Antenna 0.65  0.66 0.86 0.14   2
## Third Antenna  0.66 -0.65 0.86 0.14   2
##
##                        PC1  PC2
## SS loadings           2.40 0.89
## Proportion Var        0.60 0.22
## Cumulative Var        0.60 0.82
## Proportion Explained  0.73 0.27
## Cumulative Proportion 0.73 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  2.53  with prob <  NA
##
## Fit based upon off diagonal values = 0.95
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$		Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2		
TG	0.89	-0.11	0.80	0.20
Elytra	0.87	0.11	0.77	0.23
Second Antenna	0.65	0.66	0.86	0.14
Third Antenna	0.66	-0.65	0.86	0.14
Eigenvalues	2.40	0.89		
Cumulative proportion of total (standardized) sample variance	0.60	0.82		

Table 1: Factor Analysis Summary for H. oleracea using the Principal Components Method, No Rotation.

Looking at factor 1 (F1) we see the significant loadings on the TG and Elytra variables. The second and third antenna loadings are similar but smaller. We might interpret this as the overall size of the beetle with measurements for the TG and Elytra being the dominant variables for describing size. Looking at factor 2 (F2) we see a clear contrast between the second and third antenna measurements. The loadings for TG and Elytra appear negligible. So we might think of F2 as the antenna size factor. Using 2 factors we can account for about 82% of the variance in the data. It seems reasonable for this data set to use 2 factors.

Let's now investigate the FA using a varimax rotation of the factors to determine if this improves or changes our interpretability.

```r
#use the principal components method to perform factor analysis of the Haltica oleracea data set
#use the function principal found in the "psych" library
#use the varimax rotation

FA_HO_r <- principal(Haltica_Oleracea_Data, nfactors = 2, rotate = "varimax")
FA_HO_r
```

```
## Principal Components Analysis
## Call: principal(r = Haltica_Oleracea_Data, nfactors = 2, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 RC1  RC2   h2   u2 com
## TG             0.70 0.55 0.80 0.20 1.9
## Elytra         0.54 0.69 0.77 0.23 1.9
## Second Antenna 0.00 0.93 0.86 0.14 1.0
## Third Antenna  0.93 0.00 0.86 0.14 1.0
##
##                        RC1  RC2
## SS loadings           1.65 1.63
## Proportion Var        0.41 0.41
## Cumulative Var        0.41 0.82
## Proportion Explained  0.50 0.50
## Cumulative Proportion 0.50 1.00
##
```

```
## Mean item complexity =  1.4
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  2.53  with prob <  NA
##
## Fit based upon off diagonal values = 0.95
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$		Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2		
TG	0.70	0.55	0.80	0.20
Elytra	0.54	0.69	0.77	0.23
Second Antenna	0.00	0.93	0.86	0.14
Third Antenna	0.93	0.00	0.86	0.14
Eigenvalues	1.65	1.63		
Cumulative proportion of total (standardized) sample variance	0.41	0.82		

Table 2: Factor Analysis Summary for H. oleracea using the Principal Components Method, Factor Rotation.

Observing the factors after rotation we see that for F1 the third antenna measurement has the highest loading. TG and Elytra also have signficant loadings but not nearly as large as the third antenna measurement. In my opinion the rotated factors became harder to interpret. Looking at both F1 and F2 together I would say that both describe the beetle's size in different ways with different dominant factors. The contrast between the second and third antenna is no longer apparent in these factors. If we look at actual data from the data set there is a contrast between the second and third antenna. It is interesting that factor rotation is typically done to improve interpretability, but in this case, it appears to make the interpretablity more difficult.

Factor Analysis for Haltica Corduourum Flea Beetle Data:

Begin by reading the data for the H. corduourum into R and manipulating it into a matrix format. We'll take a look at the first few lines of the data using the head() function

```
Flea_Beetle_Data <- read_excel("fleabeetledata.xlsx")
```

```
## New names:
## * `` -> ...1
## * x1 -> x1...2
## * x2 -> x2...3
## * x3 -> x3...4
## * x4 -> x4...5
## * ... and 4 more problems
```

```
Haltica_carduourum_Data <- Flea_Beetle_Data[1:20,6:9]
Haltica_carduourum_Data <- data.frame(Haltica_carduourum_Data)
colnames(Haltica_carduourum_Data) <- c("TG", "Elytra", "Second Antenna","Third Antenna")
head(Haltica_carduourum_Data)
```

```
##    TG Elytra Second Antenna Third Antenna
## 1 181    305            184           209
## 2 158    237            133           188
## 3 184    300            166           231
## 4 171    273            162           213
## 5 181    297            163           224
## 6 181    308            160           223
```

Next we calculate the correlation matrix from the data set. It is shown below.

```
#calculate the correlation matrix for the Haltica carduourum data set
Hc_corr <- cor(Haltica_carduourum_Data)
round(Hc_corr, 3)
```

```
##                   TG Elytra Second Antenna Third Antenna
## TG             1.000  0.643          0.283         0.242
## Elytra         0.643  1.000          0.648         0.359
## Second Antenna 0.283  0.648          1.000         0.385
## Third Antenna  0.242  0.359          0.385         1.000
```

$$\mathbf{R}_{H.corduourum} = \begin{bmatrix} 1.000 & 0.643 & 0.283 & 0.242 \\ 0.643 & 1.000 & 0.648 & 0.359 \\ 0.283 & 0.648 & 1.000 & 0.385 \\ 0.242 & 0.359 & 0.385 & 1.000 \end{bmatrix}$$

Observing the correlation matrix above for the Haltica carduourum data set we see that, similar to the H. oleracea data, a large correlation exists for the TG and Elytra. Once again we could look at this as a single group that could be combined into one factor. A second group appears for the Elytra and Second Antenna due to the correlation between those variables. The other correlations that exist are to a much lesser degree so it appears that 2 factors would be reasonable for this data set.

First, let's perform the FA for 2 factors without rotation on the H. carduourum data set. Then we will investigate the FA with 2 factors using a varimax factor rotation.

```
#use the principal components method to perform factor analysis of the Haltica carduourum data set
#use the function principal found in the "psych" library
#no rotation
```

```
FA_Hc <- principal(Haltica_carduourum_Data, nfactors = 2, rotate = "none")
FA_Hc
```

```
## Principal Components Analysis
## Call: principal(r = Haltica_carduourum_Data, nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                  PC1   PC2   h2   u2 com
## TG              0.72 -0.56 0.83 0.17 1.9
## Elytra          0.90 -0.20 0.86 0.14 1.1
## Second Antenna 0.78  0.26 0.67 0.33 1.2
## Third Antenna  0.61  0.64 0.78 0.22 2.0
##
##                         PC1  PC2
## SS loadings            2.31 0.83
## Proportion Var         0.58 0.21
## Cumulative Var         0.58 0.79
## Proportion Explained  0.74 0.26
## Cumulative Proportion 0.74 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.15
##  with the empirical chi square  5.06  with prob <  NA
##
## Fit based upon off diagonal values = 0.9
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$		Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2		
TG	0.72	-0.56	0.83	0.17
Elytra	0.90	-0.20	0.86	0.14
Second Antenna	0.78	0.26	0.67	0.33
Third Antenna	0.61	0.64	0.78	0.22
Eigenvalues	2.31	0.83		
Cumulative proportion of total (standardized) sample variance	0.58	0.79		

Table 3: Factor Analysis Summary for H. corduourum using the Principal Components Method, No Rotation.

Looking at factor 1 (F1) we see the most significant loading is on the Elytra measurement. The TG and second antenna loadings are similar size but noticably smaller than the Elytra. We might interpret this as the overall size of the beetle with measurement for the Elytra being the dominant variable for describing size. Looking at factor 2 (F2) we see a definite contrast between the TG and third antenna measurements. The loadings for Elytra and second antenna are quite a bit smaller and possibly negligible. We could view F2 as a Elytra and Second Antenna contrast factor. Using 2 factors we can account for about 79% of the variance in the data. It seems reasonable for this data set to use 2 factors.

The interpretation was fairly straightforward for the FA without rotation. Factor rotation is a technique that is used to help improve interpretability by performing a rotation of the factors to maximize the variance of squares of scaled loadings for the jth factor. Next, let's investigate the H. corduourum data set using FA with two factors and utilizing a varimax rotation.

```
#use the principal components method to perform factor analysis of the Haltica carduourum data set
#use the function principal found in the "psych" library
#use the varimax rotation
FA_Hc_r <- principal(Haltica_carduourum_Data, nfactors = 2, rotate = "varimax")
FA_Hc_r
```

```
## Principal Components Analysis
## Call: principal(r = Haltica_carduourum_Data, nfactors = 2, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 RC1  RC2   h2   u2 com
## TG             0.91 0.04 0.83 0.17 1.0
## Elytra         0.82 0.43 0.86 0.14 1.5
## Second Antenna 0.43 0.70 0.67 0.33 1.6
## Third Antenna  0.05 0.88 0.78 0.22 1.0
##
##                        RC1  RC2
## SS loadings           1.69 1.46
## Proportion Var        0.42 0.36
## Cumulative Var        0.42 0.79
## Proportion Explained  0.54 0.46
## Cumulative Proportion 0.54 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.15
##  with the empirical chi square  5.06  with prob <  NA
##
## Fit based upon off diagonal values = 0.9
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$		Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2		
TG	0.91	0.04	0.83	0.17
Elytra	0.82	0.43	0.86	0.14
Second Antenna	0.43	0.70	0.67	0.33
Third Antenna	0.05	0.88	0.78	0.22
Eigenvalues	1.69	1.46		
Cumulative proportion of total (standardized) sample variance	0.42	0.79		

Table 4: Factor Analysis Summary for H. corduourum using the Principal Components Method, Factor Rotation.

After rotation we get different loadings for the variables for F1 and F2. Interestingly, the TG measurement has the highest loading after rotation and the Elytra variable also has a significant loading. After rotation we could interpret F1 the same way as we did for the unrotated FA and call F1 a “size” factor with the TG and Elytra measurements being the most dominant factors for describing H. carduourum size. Unlike the rotated F2 of H. oleracea the rotated F2 for the H. carduourum data set seems to have a clearer interpretation. Both the second and third antenna measurements load highly after rotation and we could call F2 a “antenna size” factor, where the values for these variables would describe the relative sizes of the antennas for the data set.

Conclusion and Comparison of Flea Beetle Data Sets.

If we compare the data sets we find that for the unrotated factor analysis for both beetle species resulted in an F1 that seemed to describe the overall size of the beetle. The loadings for the variables were different, but that would not necessarily be surprising since the beetles are different species and likely have different shapes such that different measurements would describe their size differently. For F2 we saw contrasts for both species although they were contrasts of different variables. Once again, I don’t find this too surprising for variables that describe different species of beetle.

Observing the rotated factors I thought that the H. oleracea variables became more difficult to interpret and the contrast for the second factor was no longer apparent. Comparing that with the H. carduourum factors after rotation I thought the interpretability was straightforward, but with a different interpretation for F2 than the unrotated F2. I found this to be interesting and a bit confusing. Which interpretation is correct for F2? I think that further analysis would have to be performed in order to understand this better and answer that question.

Factor Analysis for Engineer Apprentice Data:

We'll start the FA by reading in the engineer apprentice data.

```
#read in the engineer data for apprentices
Eng_App_Data <- read_excel("engineers_pilots.xls")
```

```
## New names:
## * `` -> ...1
## * X1 -> X1...2
## * X2 -> X2...3
## * X3 -> X3...4
## * X4 -> X4...5
## * ... and 9 more problems
```

```
Eng_App_Data <- Eng_App_Data[1:20,2:7]
Eng_App_Data <- data.frame(Eng_App_Data)
colnames(Eng_App_Data) <- c("x1", "x2", "x3","x4","x5","x6")
head (Eng_App_Data)
```

```
##    x1 x2 x3  x4 x5  x6
## 1 121 22 74 223 54 254
## 2 108 30 80 175 40 300
## 3 122 49 87 266 41 223
## 4  77 37 66 178 80 209
## 5 140 35 71 175 38 261
## 6 108 37 57 241 59 245
```

As we have done previously with the flea beetle data we will start with the correlation matrix to look for potential groupings to determine how many factors to use.

```
#calculate the correlation matrix for the engineer apprentice data set
Ea_corr <- cor(Eng_App_Data)
round(Ea_corr, 3)
```

```
##        x1     x2     x3     x4     x5     x6
## x1  1.000  0.128  0.377  0.043 -0.045  0.195
## x2  0.128  1.000  0.105  0.103 -0.432  0.037
## x3  0.377  0.105  1.000 -0.109  0.000 -0.017
## x4  0.043  0.103 -0.109  1.000 -0.043 -0.013
## x5 -0.045 -0.432  0.000 -0.043  1.000 -0.340
## x6  0.195  0.037 -0.017 -0.013 -0.340  1.000
```

$$\mathbf{R}_{apprentice} = \begin{bmatrix} 1.000 & 0.128 & 0.377 & 0.043 & -0.045 & 0.195 \\ 0.128 & 1.000 & 0.105 & 0.103 & -0.432 & 0.037 \\ 0.377 & 0.105 & 1.000 & -0.109 & 0.000 & -0.017 \\ 0.043 & 0.103 & -0.109 & 1.000 & -0.043 & -0.013 \\ -0.045 & -0.432 & 0.000 & -0.043 & 1.000 & -0.340 \\ 0.195 & 0.037 & -0.017 & -0.013 & -0.340 & 1.000 \end{bmatrix}$$

If we take a close look at the correlation matrix we do see some correlation between variables. Clearly a dimensional reduction method such as factor analysis would be reasonable for this data. Looking more closely we see a correlation between X1 and X3 that could represent a grouping. X2 and X5 have a negative correlation and could be a second grouping. X4 doesn't appear to be correlated with any of the other variables. X5 and X6 show a negative correlation and we could reasonable consider those 2 variables as a third grouping. Given our investigation of the correlation matrix I am going to select 3 factors for FA of the engineer apprentice data.

We will use the principal components method with no rotatation first.

```
#use the principal components method to perform factor analysis of the engineer apprentice data set
#use the function principal found in the "psych" library
#no rotation
FA_Ea <- principal(Eng_App_Data, nfactors = 3, rotate = "none")
FA_Ea
```

```
## Principal Components Analysis
## Call: principal(r = Eng_App_Data, nfactors = 3, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PC1   PC2   PC3   h2   u2 com
## x1  0.53  0.61  0.09 0.66 0.34 2.0
## x2  0.66 -0.23  0.38 0.63 0.37 1.9
## x3  0.36  0.76  0.13 0.72 0.28 1.5
## x4  0.11 -0.31  0.71 0.61 0.39 1.4
## x5 -0.73  0.45  0.13 0.75 0.25 1.7
## x6  0.54 -0.16 -0.60 0.67 0.33 2.1
##
##                        PC1  PC2  PC3
## SS loadings           1.68 1.32 1.04
## Proportion Var        0.28 0.22 0.17
## Cumulative Var        0.28 0.50 0.67
## Proportion Explained  0.42 0.33 0.26
## Cumulative Proportion 0.42 0.74 1.00
##
## Mean item complexity =  1.8
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.16
##  with the empirical chi square  15.41  with prob <  NA
##
## Fit based upon off diagonal values = 0.29
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$			Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2	F3		
X1	0.53	0.61	0.09	0.66	0.34
X2	0.66	-0.23	0.38	0.63	0.37
X3	0.36	0.76	0.13	0.72	0.28
X4	0.11	-0.31	0.71	0.61	0.39
X5	-0.73	0.45	0.13	0.75	0.25
X6	0.54	-0.16	-0.60	0.67	0.33
Eigenvalues	1.68	1.32	1.04		
Cumulative proportion of total (standardized) sample variance	0.28	0.50	0.67		

Table 5: Factor Analysis Summary for the Engineer Apprentice Data set using the Principal Components Method, No Rotation.

Looking first at F1, the interpretability is a bit difficult without the context of the variables and their measurements. The most dominant feature we see is a contrast between X2 and X5 so we might consider this the "X2 and X5 contrast" factor. F2 appears to have the highest loadings on X1 and X3 so we could consider this the "X1 and X3 magnitude" factor. Finally, for F3 we see a contrast between X4 and X6. We could call this the "X4 and X6" contrast factor. Using 3 factors provided a nice dimensional reduction with each factor showing somewhat clear grouping such that each variable grouped with another one. Any further interpretation is difficult without knowing what the variables represent and what they measure. We see with 3 factors that we account for about 67% of the variability in the data set, which is not bad, given that we reduced the dimension of the data by half.

Let's take a look now at the same data set using the same method for FA, three factors, and factor rotation using varimax.

```r
#use the principal components method to perform factor analysis of the engineer apprentice data set
#use the function principal found in the "psych" library
#use the varimax rotation

FA_Ea_r <- principal(Eng_App_Data, nfactors = 3, rotate = "varimax")
FA_Ea_r
```

```
## Principal Components Analysis
## Call: principal(r = Eng_App_Data, nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1   RC2   RC3   h2   u2 com
## x1  0.13  0.80  0.04 0.66 0.34 1.1
## x2  0.41  0.20  0.65 0.63 0.37 1.9
## x3 -0.08  0.85 -0.04 0.72 0.28 1.0
## x4 -0.15 -0.10  0.76 0.61 0.39 1.1
## x5 -0.80  0.02 -0.31 0.75 0.25 1.3
## x6  0.78  0.06 -0.26 0.67 0.33 1.2
```

```
##
##                         RC1  RC2  RC3
## SS loadings           1.47 1.41 1.17
## Proportion Var        0.24 0.23 0.19
## Cumulative Var        0.24 0.48 0.67
## Proportion Explained  0.36 0.35 0.29
## Cumulative Proportion 0.36 0.71 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.16
##  with the empirical chi square  15.41  with prob <  NA
##
## Fit based upon off diagonal values = 0.29
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$			Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2	F3		
X1	0.13	0.80	0.04	0.66	0.34
X2	0.41	0.20	0.65	0.63	0.37
X3	-0.08	0.85	-0.04	0.72	0.28
X4	-0.15	-0.10	0.76	0.61	0.39
X5	-0.80	0.02	-0.31	0.75	0.25
X6	0.78	0.06	-0.26	0.67	0.33
Eigenvalues	1.47	1.41	1.17		
Cumulative proportion of total (standardized) sample variance	0.24	0.48	0.67		

Table 6: Factor Analysis Summary for the Engineer Apprentice Data set using the Principal Components Method, Factor Rotation.

Observing the rotated factors we see for F1 that a clear contrast emerges between X5 and X6. This is different from our interpretation of the unrotated F1. For F2 we see the same interpretation as we did for unrotated F2. This factor shows the highest loadings for X1 and X3 so we can still look at this as the "X1 and X3 magnitude" factor. Rotated F3 also has different loadings and interpretation from the unrotated F3. For F3 we see high loadings between X2 and X4 and could consider this as an "X2 and X4 magnitude" factor. In this case the rotated factors were a bit easier to interpret because 2 variables in each factor emerged as dominant in describing the data for that factor.

As we noted above it would be necessary to understand what the variables and their measurements represent in order to more fully interpret the data set. That said, since we have already completed the FA we could add that information later to provide contextual interpretation. In some sense this might be viewed as a way to perform FA while minimizing bias since the investigator has no knowledge of the variables, their measurements or contextual meaning. Since the data is general and anonymous we minimize the bias that could be introduced if we are seeking particular meaning from the data.

Factor Analysis for Engineer Pilot Data:

We'll start the FA by reading in the engineer apprentice data.

```r
#read in the engineer data for pilots
Eng_Pilot_Data <- read_excel("engineers_pilots.xls")
```

```
## New names:
## * `` -> ...1
## * X1 -> X1...2
## * X2 -> X2...3
## * X3 -> X3...4
## * X4 -> X4...5
## * ... and 9 more problems
```

```r
Eng_Pilot_Data <- Eng_Pilot_Data[1:20,9:14]
Eng_Pilot_Data <- data.frame(Eng_Pilot_Data)
colnames(Eng_Pilot_Data) <- c("x1", "x2", "x3","x4","x5","x6")
head(Eng_Pilot_Data)
```

```
##    x1 x2 x3  x4 x5  x6
## 1 132 17 77 232 50 249
## 2 123 32 79 192 64 315
## 3 129 31 96 250 55 319
## 4 131 23 67 291 48 310
## 5 110 24 96 239 42 268
## 6  47 22 87 231 40 217
```

Next we will use R to calculate the correlation matrix so we can investigate the potential for factor analysis and variable groupings.

```r
#calculate the correlation matrix for the engineer pilot data set
Ep_corr <- cor(Eng_Pilot_Data)
round(Ep_corr, 3)
```

```
##        x1     x2     x3     x4     x5     x6
## x1  1.000  0.368 -0.189  0.104 -0.004  0.473
## x2  0.368  1.000  0.044 -0.232 -0.373  0.222
## x3 -0.189  0.044  1.000  0.148 -0.057 -0.300
## x4  0.104 -0.232  0.148  1.000  0.022  0.130
## x5 -0.004 -0.373 -0.057  0.022  1.000  0.249
## x6  0.473  0.222 -0.300  0.130  0.249  1.000
```

$$\mathbf{R}_{pilot} = \begin{bmatrix} 1.000 & 0.368 & -0.189 & 0.104 & -0.004 & 0.473 \\ 0.368 & 1.000 & 0.044 & -0.232 & -0.373 & 0.222 \\ -0.189 & 0.044 & 1.000 & 0.148 & -0.057 & -0.300 \\ 0.104 & -0.232 & 0.148 & 1.000 & 0.022 & 0.130 \\ -0.004 & -0.373 & -0.057 & 0.022 & 1.000 & 0.249 \\ 0.473 & 0.222 & -0.300 & 0.130 & 0.249 & 1.000 \end{bmatrix}$$

Looking carefully at the correlation matrix for the engineer pilot data we can see that some correlation between variables does exist. Factor analysis does seem reasonable for this data set. X1 and X6, and X1 and X2 appear to be correlated and could be grouped. X2 and X5 are negatively correlated and could also potentially be grouped together. Variables X3 and X6 also show some negative correlation. Given the multicollinearity of the variables in this data set I think that 3 factors seems to be reasonable for factor analysis.

For the engineer pilot data set we will use the principal components method with 3 factors and no rotation to begin.

```
#use the principal components method to perform factor analysis of the engineer pilot data set
#use the function principal found in the "psych" library
#no rotation
FA_Ep <- principal(Eng_Pilot_Data, nfactors = 3, rotate = "none")
FA_Ep
```

```
## Principal Components Analysis
## Call: principal(r = Eng_Pilot_Data, nfactors = 3, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PC1   PC2   PC3   h2   u2 com
## x1  0.82 -0.02  0.24 0.72 0.28 1.2
## x2  0.55 -0.70  0.06 0.80 0.20 1.9
## x3 -0.45 -0.25  0.59 0.62 0.38 2.3
## x4 -0.01  0.40  0.82 0.84 0.16 1.4
## x5  0.02  0.79 -0.19 0.67 0.33 1.1
## x6  0.80  0.35  0.06 0.76 0.24 1.4
##
##                        PC1  PC2  PC3
## SS loadings           1.82 1.47 1.12
## Proportion Var        0.30 0.24 0.19
## Cumulative Var        0.30 0.55 0.73
## Proportion Explained  0.41 0.33 0.25
## Cumulative Proportion 0.41 0.75 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.13
##  with the empirical chi square  10.4  with prob <  NA
##
## Fit based upon off diagonal values = 0.69
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$			Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2	F3		
X1	0.82	-0.02	0.24	0.72	0.28
X2	0.55	-0.70	0.06	0.80	0.20
X3	-0.45	-0.25	0.59	0.62	0.38
X4	-0.01	0.40	0.82	0.84	0.16
X5	0.02	0.79	-0.19	0.67	0.33
X6	0.80	0.35	0.06	0.76	0.24
Eigenvalues	1.81	1.47	1.12		
Cumulative proportion of total (standardized) sample variance	0.30	0.55	0.73		

Table 7: Factor Analysis Summary for the Engineer Pilot Data set using the Principal Components Method, No Rotation.

Observing factor one (F1) first we can note that variables X1 and X6 show the highest loadings. We could call F1 the "X1 and X6 magnitude" factor since the first factor loads the heaviest on those variables. The second factor (F2) appears to be a contrast between X2 and X5 so we could call this the "X2 and X5 contrast" factor. The third factor (F3) loads X4 the highest but X3 also appears to be significant. So let's call this the "X3 and X4 magnitude" factor. For this FA we ended up with a dimension reduction of half the variables and also we are able to explain about 73% of the variability in this data set. It was nice that we were able to account for all of the variables in the the three factor model. This way, even though we only explain 73% of the variability, we have representation of each variable in the interpretation of the factors.

We can use the varimax rotation of the factors to see if we can further improve interpretability.

```r
#use the principal components method to perform factor analysis of the engineer pilot data set
#use the function principal found in the "psych" library
#no rotation
FA_Ep_r <- principal(Eng_Pilot_Data, nfactors = 3, rotate = "varimax")
FA_Ep_r
```

```
## Principal Components Analysis
## Call: principal(r = Eng_Pilot_Data, nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1   RC2   RC3   h2   u2 com
## x1  0.81  0.24  0.06 0.72 0.28 1.2
## x2  0.37  0.80 -0.16 0.80 0.20 1.5
## x3 -0.41  0.26  0.62 0.62 0.38 2.1
## x4  0.22 -0.22  0.86 0.84 0.16 1.3
## x5  0.18 -0.79 -0.06 0.67 0.33 1.1
## x6  0.86 -0.15 -0.04 0.76 0.24 1.1
##
##                        RC1  RC2  RC3
## SS loadings           1.78 1.47 1.16
## Proportion Var        0.30 0.24 0.19
## Cumulative Var        0.30 0.54 0.73
## Proportion Explained  0.40 0.33 0.26
## Cumulative Proportion 0.40 0.74 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.13
##  with the empirical chi square  10.4  with prob <  NA
##
## Fit based upon off diagonal values = 0.69
```

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e_{ij}}$			Communalities $\tilde{h_i}^2$	Specific variances $\tilde{\psi} = 1 - \tilde{h_i}^2$
	F1	F2	F3		
X1	0.81	0.24	0.06	0.72	0.28
X2	0.37	0.80	-0.16	0.80	0.20
X3	-0.41	0.26	0.62	0.62	0.38
X4	0.22	-0.22	0.86	0.84	0.16
X5	0.18	-0.79	-0.06	0.67	0.33
X6	0.86	-0.15	-0.04	0.76	0.24
Eigenvalues	1.78	1.47	1.16		
Cumulative proportion of total (standardized) sample variance	0.30	0.54	0.73		

Table 8: Factor Analysis Summary for the Engineer Pilot Data set using the Principal Components Method, Factor Rotation.

After rotation we find that F1 again loads the highest on X1 and X6. After rotation the interpretation of F1 remains the same as the unrotated F1. F1 is our “X1 and X6 magnitude” factor. After rotation F2 retains the same interpretation as we clearly see a contrast between X2 and X5 as we did with the unrotated F2. F2 is the “X2 and X5 contrast” factor. Similary, F3, after rotation has a similar interpretation to the unrotated F3. F3 loads the highest on X4 and X3 so we consider this the “X3 and X4 magnitiude” factor. The rotated factors account for the same cumulative proportion of variance as the unrotated factors.

Conclusion and Comparison of the Engineer Data Sets.

If we compare the unrotated and rotated factors loadings and interpretations between the apprentice and pilot data sets we see that they are different. We are not able to interpret this in the context of the variables since we don’t know what the variables mean or what they measure. We can draw a general conclusion that whatever they mean we can clearly see they describe the apprentices and pilots differently. This is not unreasonable as we might expect that two distinct groups of engineers could be described differently if the variables were measuring things like skill, experience, or ability. We might expect that the pilots would score or be measured differently then apprentices. It would be very interesting to now see what the variables are defined as and their respective measurements in order to fill in the missing context. It may also help us understand if a 3 factor model is descriptive enough for our goals or if a 4 factor model should be investigated. Overall, in the context of the available information for this problem, the 3 factor model does a good job of dimension reduction, variable description and interpretability.