Student ID: 112077423

```
library(ggplot2)
library(compstatslib)
library(data.table)
library(tidyr)
library(lsa)
```

```
## Warning:     'lsa'           R     4.3.3
```

```
library(readxl)
```

```
## Warning:     'readxl'           R     4.3.3
```

```
library(tidyverse)
```

```
## Warning:     'readr'           R     4.3.2
```

```
library(psych)
```

```
## Warning:     'psych'           R     4.3.3
```

```
df <- read_excel('security_questions.xlsx', sheet = 2)
head(df)
```

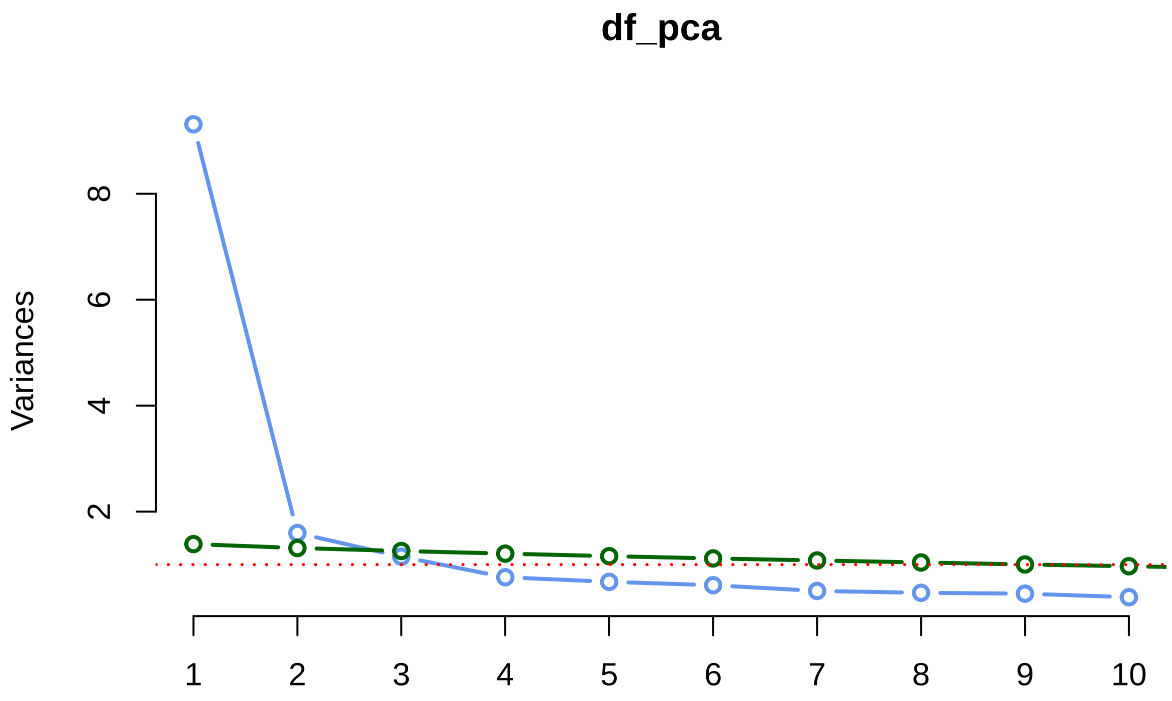
```
## # A tibble: 6 x 18
##      Q1    Q2    Q3    Q4    Q5    Q6    Q7    Q8    Q9   Q10   Q11   Q12   Q13
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     7     5     5     7     7     4     4     7     5     7     5     7     5
## 2     5     5     6     6     6     5     5     7     5     6     6     6     6
## 3     6     6     6     6     7     6     6     6     5     7     6     6     5
## 4     5     5     5     5     5     5     5     5     5     5     5     5     4
## 5     7     7     7     7     7     4     5     7     6     7     6     7     6
## 6     6     5     4     5     4     4     4     5     6     2     5     5     5
## # i 5 more variables: Q14 <dbl>, Q15 <dbl>, Q16 <dbl>, Q17 <dbl>, Q18 <dbl>
```

Question 1(a)

Show a single visualization with scree plot of data, scree plot of simulated noise (use average eigenvalues of 100 noise samples), and a horizontal line showing the eigenvalue = 1 cutoff.

```
set.seed(64528409)

df_pca <- prcomp(df, scale.=TRUE)
# function to get eigenvalues from noise data
sim_noise_ev <- function(n, p) {
  noise <- data.frame(replicate(p, rnorm(n)))
  return(eigen(cor(noise))$values)
}
# generate noise data
evalues_noise <- replicate(100, sim_noise_ev(nrow(df), ncol(df)))
# get mean of each row
evalues_mean <- apply(evalues_noise, 1, mean)
# plot
screeplot(df_pca, type="lines", col='cornflowerblue', lwd=2)
lines(evalues_mean, type="b", col='darkgreen', lwd=2)
abline(h=1, col="red", lty='dotted', lwd=1.5)
```



Question 1(b)

How many dimensions would you retain if we used Parallel Analysis?

In the Parallel Analysis we reatain PC when its ev of original data > ev of noise data. In this case, I'd retain only two dimensions.

Question 2(a)

Looking at the loadings of the first 3 principal components, to which components does each item seem to best belong?

```
df_principal <- principal(df, nfactor=18, rotate="none", scores=TRUE)
df_principal$loadings[,1:3]
```

```
##           PC1         PC2          PC3
## Q1  0.8169846 -0.13941235 -0.002115927
## Q2  0.6726084 -0.01375526  0.089174403
## Q3  0.7655215 -0.03269651  0.089686106
## Q4  0.6233733  0.64307826  0.108031860
## Q5  0.6900841 -0.03126466 -0.542354570
## Q6  0.6828029 -0.10462094  0.207232000
## Q7  0.6566249 -0.31763196  0.324176779
## Q8  0.7861054  0.04235983 -0.343212951
## Q9  0.7230295 -0.23164618  0.203556038
## Q10 0.6861529 -0.09868038 -0.532678749
## Q11 0.7529735 -0.26100673  0.172516196
## Q12 0.6303505  0.63753124  0.121522834
## Q13 0.7119085 -0.06463837  0.084335919
## Q14 0.8114677 -0.09970016  0.156787046
## Q15 0.7040428  0.01057936 -0.332546876
## Q16 0.7575616 -0.20281591  0.183170175
## Q17 0.6175336  0.66426051  0.110061160
## Q18 0.8067284 -0.11360432 -0.065189145
```

Q4, Q12, and Q17 best belong to PC2 whereas the rest best belong to PC1.

Question 2(b)

How much of the total variance of the security dataset do the first 3 PCs capture?

```
df_principal$Structure
```

```
##
## Loadings:
##      PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9    PC10
## Q1   0.817 -0.139         0.110         0.143 -0.337        -0.107
## Q2   0.673                0.225         0.624        -0.254
## Q3   0.766               -0.349         0.105  0.211               -0.391
## Q4   0.623  0.643  0.108
## Q5   0.690        -0.542        -0.159  0.117  0.137  0.129  0.147
## Q6   0.683 -0.105  0.207         0.502                0.368  0.223
## Q7   0.657 -0.318  0.324  0.286                0.322  0.157 -0.159  0.195
## Q8   0.786        -0.343         0.172 -0.157        -0.140 -0.156
## Q9   0.723 -0.232  0.204 -0.109        -0.211        -0.309  0.401  0.161
## Q10  0.686        -0.533        -0.205         0.111  0.171
## Q11  0.753 -0.261  0.173  0.231 -0.173 -0.151         0.117 -0.195
## Q12  0.630  0.638  0.122                                            0.104
## Q13  0.712               -0.526                             -0.189  0.305
## Q14  0.811         0.157 -0.317                             -0.151
## Q15  0.704        -0.333         0.422 -0.201  0.112 -0.209 -0.169 -0.119
## Q16  0.758 -0.203  0.183  0.178 -0.282 -0.171        -0.127        -0.132
## Q17  0.618  0.664  0.110        -0.129
## Q18  0.807 -0.114                             -0.414          0.124
##     PC11   PC12   PC13   PC14   PC15   PC16   PC17   PC18  
## Q1  -0.156 -0.201        -0.110        -0.128         0.223
## Q2                                                         
## Q3  -0.128               -0.196                            
## Q4  -0.109        -0.173  0.275        -0.126  0.178       
## Q5                -0.223         0.203        -0.121       
## Q6   0.137                                                 
## Q7  -0.263                                                 
## Q8  -0.130 -0.169               -0.251        -0.145 -0.145
## Q9                                                   0.101
## Q10                0.294        -0.133         0.114       
## Q11  0.236  0.227 -0.120        -0.149 -0.136              
## Q12                0.213 -0.238        -0.143        -0.171
## Q13  0.182        -0.108                       0.122       
## Q14         0.127  0.159  0.196  0.156        -0.231       
## Q15  0.106                       0.163         0.101       
## Q16  0.229 -0.264                                    -0.119
## Q17                                     0.246 -0.179  0.191
## Q18 -0.136  0.210 -0.106                0.203        -0.138
## 
##                  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
## SS loadings    9.311 1.596 1.150 0.762 0.675 0.612 0.503 0.468 0.452 0.385
## Proportion Var 0.517 0.089 0.064 0.042 0.038 0.034 0.028 0.026 0.025 0.021
## Cumulative Var 0.517 0.606 0.670 0.712 0.750 0.784 0.812 0.838 0.863 0.884
##                 PC11  PC12  PC13  PC14  PC15  PC16  PC17  PC18
## SS loadings    0.355 0.301 0.292 0.262 0.235 0.230 0.209 0.202
## Proportion Var 0.020 0.017 0.016 0.015 0.013 0.013 0.012 0.011
## Cumulative Var 0.904 0.921 0.937 0.951 0.964 0.977 0.989 1.000
```

The cumulative variance of the first three principal components is 0.67.

Question 2(c)

Looking at commonality and uniqueness, which items are less than adequately explained by the first 3 principal components?

```r
principal(df, nfactor=3, rotate="none", scores=TRUE)
```

```
## Principal Components Analysis
## Call: principal(r = df, nfactors = 3, rotate = "none", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PC1   PC2   PC3   h2   u2 com
## Q1  0.82 -0.14  0.00 0.69 0.31 1.1
## Q2  0.67 -0.01  0.09 0.46 0.54 1.0
## Q3  0.77 -0.03  0.09 0.60 0.40 1.0
## Q4  0.62  0.64  0.11 0.81 0.19 2.1
## Q5  0.69 -0.03 -0.54 0.77 0.23 1.9
## Q6  0.68 -0.10  0.21 0.52 0.48 1.2
## Q7  0.66 -0.32  0.32 0.64 0.36 2.0
## Q8  0.79  0.04 -0.34 0.74 0.26 1.4
## Q9  0.72 -0.23  0.20 0.62 0.38 1.4
## Q10 0.69 -0.10 -0.53 0.76 0.24 1.9
## Q11 0.75 -0.26  0.17 0.66 0.34 1.4
## Q12 0.63  0.64  0.12 0.82 0.18 2.1
## Q13 0.71 -0.06  0.08 0.52 0.48 1.0
## Q14 0.81 -0.10  0.16 0.69 0.31 1.1
## Q15 0.70  0.01 -0.33 0.61 0.39 1.4
## Q16 0.76 -0.20  0.18 0.65 0.35 1.3
## Q17 0.62  0.66  0.11 0.83 0.17 2.0
## Q18 0.81 -0.11 -0.07 0.67 0.33 1.1
##
##                        PC1  PC2  PC3
## SS loadings           9.31 1.60 1.15
## Proportion Var        0.52 0.09 0.06
## Cumulative Var        0.52 0.61 0.67
## Proportion Explained  0.77 0.13 0.10
## Cumulative Proportion 0.77 0.90 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.05
##  with the empirical chi square  258.65  with prob <  1.4e-15
##
## Fit based upon off diagonal values = 0.99
```

Items that are less than adequately explained by the first 3 principal components: Q1, Q2, Q3, Q6, Q7, Q9, Q11, Q13, Q14, Q15, Q16, Q18. Communality is less than 0.7

Question 2(d)

How many measurement items share similar loadings between 2 or more components?

```
loadings <- round(df_principal$loadings[, 1:18], 3)
num <- 0
lst <- list()
for (i in 1:18) {
  for (j in 1:18) {
    diff <- abs(abs(loadings[i,]) - abs(loadings[i, j]))
    diff[j] <- 5
    lst <- append(lst, names(diff)[diff == 0])
  }
  lst <- unlist(lst, recursive = FALSE)
  if(length(unique(lst)) >= 2) num <- num + 1
  lst <- list()
}
print(paste(num, 'measurement items share similar loadings between 2 or more components'))
```

```
## [1] "9 measurement items share similar loadings between 2 or more components"
```

Question 2(e)

Can you interpret a 'meaning' behind the first principal component from the items that load best upon it?

```
tmp <- round(df_principal$loadings[,1], 2)
tmp[tmp > 0.8]
```

```
##   Q1  Q14  Q18
## 0.82 0.81 0.81
```

Q1 and Q4 are more related to confidentiality whereas Q14 is more related to the accuracy of the information.

Question 3(a)

Individually, does each rotated component (RC) explain the same, or different, amount of variance than the corresponding principal components (PCs)?

```
df_pca_rot <- principal(df, nfactor=3, rotate="varimax", scores=TRUE)
df_pca_rot
```

```
## Principal Components Analysis
## Call: principal(r = df, nfactors = 3, rotate = "varimax", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1  RC3  RC2   h2   u2 com
## Q1  0.66 0.45 0.22 0.69 0.31 2.0
## Q2  0.54 0.29 0.29 0.46 0.54 2.1
## Q3  0.62 0.34 0.31 0.60 0.40 2.1
## Q4  0.22 0.19 0.85 0.81 0.19 1.2
## Q5  0.24 0.83 0.16 0.77 0.23 1.3
## Q6  0.65 0.20 0.23 0.52 0.48 1.5
## Q7  0.79 0.10 0.06 0.64 0.36 1.0
## Q8  0.38 0.71 0.30 0.74 0.26 2.0
## Q9  0.74 0.23 0.14 0.62 0.38 1.3
## Q10 0.28 0.82 0.10 0.76 0.24 1.3
## Q11 0.76 0.28 0.12 0.66 0.34 1.3
## Q12 0.23 0.19 0.85 0.82 0.18 1.2
## Q13 0.59 0.32 0.26 0.52 0.48 1.9
## Q14 0.72 0.31 0.28 0.69 0.31 1.7
## Q15 0.34 0.66 0.24 0.61 0.39 1.8
## Q16 0.74 0.27 0.17 0.65 0.35 1.4
## Q17 0.21 0.19 0.87 0.83 0.17 1.2
## Q18 0.61 0.50 0.23 0.67 0.33 2.2
##
##                        RC1  RC3  RC2
## SS loadings           5.61 3.49 2.95
## Proportion Var        0.31 0.19 0.16
## Cumulative Var        0.31 0.51 0.67
## Proportion Explained  0.47 0.29 0.24
## Cumulative Proportion 0.47 0.76 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  0.05
##  with the empirical chi square  258.65  with prob <  1.4e-15
```

```
##
## Fit based upon off diagonal values = 0.99
```

Each rotated component (RC) explain **different** amount of variance than the corresponding principal component.

Question 3(b)

Together, do the three rotated components explain the same, more, or less cumulative variance as the three principal components combined?

Three rotated components explain the **same** cumulative variance as the three principal components combined.

Question 3(c)

Looking back at the items that shared similar loadings with multiple principal components (2d), do those items have more clearly differentiated loadings among rotated components?

Rotated components are not principal Components. Therefore, we have different loadings.

Question 3(d)

Can you now more easily interpret the "meaning" of the 3 rotated components from the items that load best upon each of them?

```r
tmp <- round(df_pca_rot$loadings[], 2)

for (i in 1:nrow(tmp)) {
  for (j in 1:ncol(tmp)) {
    if (tmp[i, j] > 0.7) {
      #cat("\033[31m", tmp[i, j], "\033[0m", "\t", sep='')
      cat(tmp[i, j], 'x\t')
    } else {
      cat(tmp[i, j], "\t")
    }
  }
  cat("\n")
}
```

```
## 0.66     0.45    0.22
## 0.54     0.29    0.29
## 0.62     0.34    0.31
## 0.22     0.19    0.85 x
## 0.24     0.83 x  0.16
## 0.65     0.2     0.23
## 0.79 x   0.1     0.06
## 0.38     0.71 x  0.3
## 0.74 x   0.23    0.14
## 0.28     0.82 x  0.1
## 0.76 x   0.28    0.12
## 0.23     0.19    0.85 x
## 0.59     0.32    0.26
## 0.72 x   0.31    0.28
## 0.34     0.66    0.24
## 0.74 x   0.27    0.17
## 0.21     0.19    0.87 x
## 0.61     0.5     0.23
```

RC1 is more about personal information-related things. RC2 is about data transmission. RC3 is about providing transaction-related evidence.

Question 3(e)

If we reduced the number of extracted and rotated components to 2, does the meaning of our rotated components change?

```
df_pca_rot <- principal(df, nfactor=2, rotate="varimax", scores=TRUE)


tmp <- round(df_pca_rot$loadings[], 2)

for (i in 1:nrow(tmp)) {
  for (j in 1:ncol(tmp)) {
    if (tmp[i, j] > 0.7) {
      #cat("\033[31m", tmp[i, j], "\033[0m\t", sep='')
      cat(tmp[i, j], 'x\t')
    } else {
      cat(tmp[i, j], "\t")
    }
  }
  cat("\n")
}
```

```
## 0.78 x   0.27
## 0.6  0.31
## 0.69     0.34
## 0.24     0.86 x
## 0.62     0.31
## 0.65     0.24
## 0.73 x   0.04
## 0.67     0.42
## 0.75 x   0.15
## 0.65     0.24
## 0.79 x   0.13
## 0.25     0.86 x
## 0.65     0.29
## 0.76 x   0.3
## 0.61     0.35
## 0.76 x   0.19
## 0.22     0.88 x
## 0.76 x   0.29
```

I think the meaning does change to an extent.

Additional Question

Looking back at all our results and analyses of this dataset (from this week and previous), how many components (1-3) do you believe we should extract and analyze to understand the security dataset? Feel free to suggest different answers for different purposes.

I'd still retain only one dimension. I don't think the second component has a great value even if it passed the Parallel Analysis.