PCA Reducing Dimentionality- vocation

Joseph Shu

2/28/2020

Golding and Seidman (1974) studied the vocational interests of 231 undergraduate males. Each respondent rated the strength of his interests in 22 vocational areas, listed below:

- X1 public speaking
- X2 law and politics
- X3 business management
- X4 sales
- X5 merchandising
- X6 office practice
- X7 military activities
- X8 technical supervision
- X9 mathematics
- X10 science
- X11 mechanical
- X12 nature
- X13 agriculture
- X14 adventure
- X15 recreational leadership
- X16 medical service
- X17 social service
- X18 religious activities
- X19 teaching
- X20 music
- X21 art
- X22 writing

Does there appear to be more than one dimension describing vocational interests among undergraduate males? How would you describe the under- lying dimension(s)? Which vocational interests seem to go together? Which seem most different?

Read in data

```
vocation <- read.csv("vocations.csv", header = FALSE)
```

```
vocation[upper.tri(vocation)] = t(vocation)[upper.tri(vocation)]
vocation
```

```
##        V1     V2     V3     V4     V5     V6     V7     V8     V9    V10    V11    V12
## 1   1.00  0.77  0.53  0.54  0.54  0.30  0.16  0.36 -0.11 -0.10 -0.02  0.14
## 2   0.77  1.00  0.50  0.44  0.48  0.28  0.20  0.34 -0.05 -0.09 -0.07 -0.02
## 3   0.53  0.50  1.00  0.74  0.91  0.72  0.28  0.79  0.08 -0.03  0.22  0.04
## 4   0.54  0.44  0.74  1.00  0.82  0.63  0.19  0.56  0.02 -0.07  0.23  0.05
## 5   0.54  0.48  0.91  0.82  1.00  0.75  0.26  0.70  0.05 -0.08  0.21  0.07
## 6   0.30  0.28  0.72  0.63  0.75  1.00  0.31  0.63  0.20  0.02  0.27 -0.03
## 7   0.16  0.20  0.28  0.19  0.26  0.31  1.00  0.38  0.03  0.15  0.29  0.23
## 8   0.36  0.34  0.79  0.56  0.70  0.63  0.38  1.00  0.14  0.05  0.37  0.11
## 9  -0.11 -0.05  0.08  0.02  0.05  0.20  0.03  0.14  1.00  0.50  0.44 -0.04
## 10 -0.10 -0.09 -0.03 -0.07 -0.08  0.02  0.15  0.05  0.50  1.00  0.62  0.37
## 11 -0.02 -0.07  0.22  0.23  0.21  0.27  0.29  0.37  0.44  0.62  1.00  0.31
## 12  0.14 -0.02  0.04  0.05  0.07 -0.03  0.23  0.11 -0.04  0.37  0.31  1.00
## 13  0.09 -0.01  0.06  0.10  0.09 -0.03  0.24  0.11 -0.10  0.08  0.21  0.73
## 14  0.21  0.18  0.15  0.15  0.14 -0.01  0.16  0.13  0.13  0.11  0.28  0.12
## 15  0.16  0.21  0.22  0.22  0.22  0.23  0.29  0.18  0.03 -0.07  0.09  0.10
## 16  0.23  0.24  0.09  0.12  0.12  0.05  0.19  0.08  0.08  0.41  0.24  0.33
## 17  0.38  0.36  0.13  0.21  0.14  0.10  0.07  0.00 -0.19 -0.04 -0.07  0.23
## 18  0.32  0.17  0.18  0.22  0.17  0.27  0.17  0.13 -0.01  0.12  0.14  0.33
## 19  0.37  0.23  0.29  0.35  0.28  0.30  0.15  0.20 -0.03  0.18  0.16  0.36
## 20  0.22  0.04 -0.01  0.05  0.06 -0.05 -0.22 -0.06  0.01  0.22  0.11  0.31
## 21  0.19 -0.01 -0.06  0.04  0.05 -0.13 -0.15 -0.10  0.02  0.22  0.12  0.49
## 22  0.49  0.26  0.04  0.16  0.10 -0.08 -0.10 -0.06 -0.23 -0.04 -0.12  0.28
##      V13   V14   V15  V16   V17   V18   V19   V20   V21   V22
## 1   0.09  0.21  0.16 0.23  0.38  0.32  0.37  0.22  0.19  0.49
## 2  -0.01  0.18  0.21 0.24  0.36  0.17  0.23  0.04 -0.01  0.26
## 3   0.06  0.15  0.22 0.09  0.13  0.18  0.29 -0.01 -0.06  0.04
## 4   0.10  0.15  0.22 0.12  0.21  0.22  0.35  0.05  0.04  0.16
## 5   0.09  0.14  0.22 0.12  0.14  0.17  0.28  0.06  0.05  0.10
## 6  -0.03 -0.01  0.23 0.05  0.10  0.27  0.30 -0.05 -0.13 -0.08
## 7   0.24  0.16  0.29 0.19  0.07  0.17  0.15 -0.22 -0.15 -0.10
## 8   0.11  0.13  0.18 0.08  0.00  0.13  0.20 -0.06 -0.10 -0.06
## 9  -0.10  0.13  0.03 0.08 -0.19 -0.01 -0.03  0.01  0.02 -0.23
## 10  0.08  0.11 -0.07 0.41 -0.04  0.12  0.18  0.22  0.22 -0.04
## 11  0.21  0.28  0.09 0.24 -0.07  0.14  0.16  0.11  0.12 -0.12
## 12  0.73  0.12  0.10 0.33  0.23  0.33  0.36  0.31  0.49  0.28
## 13  1.00  0.31  0.32 0.05  0.09  0.19  0.12  0.00  0.17  0.09
## 14  0.31  1.00  0.41 0.12 -0.01  0.00 -0.02 -0.05  0.02  0.08
## 15  0.32  0.41  1.00 0.10  0.18  0.19  0.12 -0.28 -0.22 -0.02
## 16  0.05  0.12  0.10 1.00  0.29  0.20  0.22  0.26  0.23  0.15
## 17  0.09 -0.01  0.18 0.29  1.00  0.47  0.51  0.27  0.26  0.42
## 18  0.19  0.00  0.19 0.20  0.47  1.00  0.41  0.37  0.25  0.31
## 19  0.12 -0.02  0.12 0.22  0.51  0.41  1.00  0.42  0.34  0.42
## 20  0.00 -0.05 -0.28 0.26  0.27  0.37  0.42  1.00  0.73  0.57
## 21  0.17  0.02 -0.22 0.23  0.26  0.25  0.34  0.73  1.00  0.62
## 22  0.09  0.08 -0.02 0.15  0.42  0.31  0.42  0.57  0.62  1.00
```

Fill in the NAs and make it symmetrical.

Rename the columns

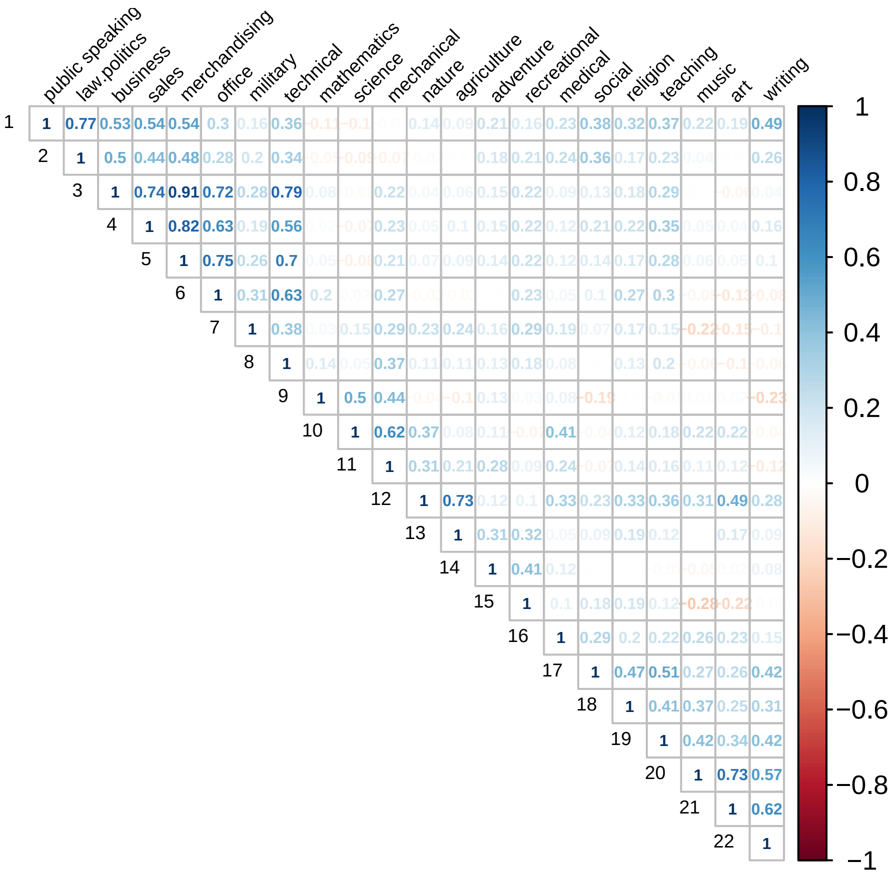
```r
names(vocation)[1] <- "public speaking"
names(vocation)[2] <- "law.politics"
names(vocation)[3] <- "business"
names(vocation)[4] <-"sales"
names(vocation)[5] <- "merchandising"
names(vocation)[6] <- "office"
names(vocation)[7] <- "military"
names(vocation)[8] <- "technical"
names(vocation)[9] <- "mathematics"
names(vocation)[10] <- "science"
names(vocation)[11] <- "mechanical"
names(vocation)[12] <- "nature"
names(vocation)[13] <- "agriculture"
names(vocation)[14] <- "adventure"
names(vocation)[15] <- "recreational"
names(vocation)[16] <- "medical"
names(vocation)[17] <- "social"
names(vocation)[18] <- "religion"
names(vocation)[19] <- "teaching"
names(vocation)[20] <- "music"
names(vocation)[21] <- "art"
names(vocation)[22] <- "writing"
```

Visualize the matrix for a better interpretation

```
library(corrplot)
```

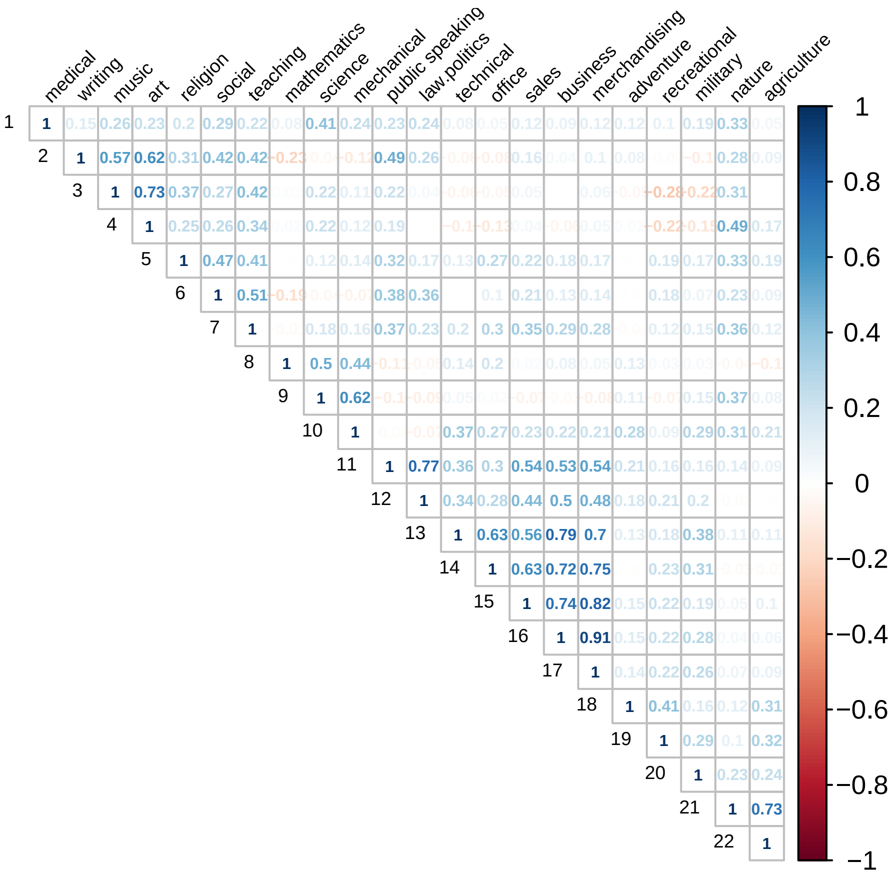
```
## corrplot 0.84 loaded
```

```
voc.matrix <- as.matrix(vocation)
corrplot(voc.matrix,
         method = "number", # correlation coefficients for numeric variables
         type = "upper",    # to display only the lower part of the symmetric correlation matrix
         tl.cex = 0.6,        # font size of the variable labels
         tl.col = "black",    # color of the variable labels
         tl.srt = 45,         # rotation angle for the variable labels
         number.cex = 0.5      # font size of the coefficients
)
```



A little peek at the correlation matrix, we can see that business is highly correlated with sales, merchandising, office operations and technical supervision, which makes a lot of sense.

```
corrplot(voc.matrix,
         method = "number",
         type = "upper",
         order = "hclust", # reorder by the size of the correlation coefficients
         tl.cex = 0.6, # font size of the variable labels
         tl.col = "black", # color of the variable labels
         tl.srt = 45, # rotation angle for the variable labels
         number.cex = 0.5 # font size of the coefficients
)
```



By rearranging the corrplot, we can already see some patterns in variables that are correlated strongly with another. For example, the business-related interests and artistic interests.

Unrotated PCA solution

```
library(psych)
unrotatedPCA <- principal(r = voc.matrix,
                          nfactors = 4,
                          rotate="none",
                          scores = FALSE)
print(unrotatedPCA, cut = 0, digits = 3)
```

```
## Principal Components Analysis
## Call: principal(r = voc.matrix, nfactors = 4, rotate = "none", scores = FALSE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##       PC1    PC2    PC3    PC4    h2    u2   com
## 1   0.723  0.087 -0.389  0.065 0.686 0.314 1.59
## 2   0.615 -0.102 -0.366  0.089 0.530 0.470 1.75
## 3   0.810 -0.418 -0.081 -0.159 0.863 0.137 1.61
## 4   0.778 -0.266 -0.143 -0.120 0.711 0.289 1.36
## 5   0.825 -0.357 -0.117 -0.164 0.848 0.152 1.50
## 6   0.675 -0.435  0.038 -0.252 0.710 0.290 2.03
## 7   0.403 -0.218  0.322  0.326 0.420 0.580 3.48
## 8   0.688 -0.426  0.140 -0.133 0.693 0.307 1.86
## 9  0.073 -0.131  0.585 -0.411 0.533 0.467 1.95
## 10 0.138  0.287  0.749 -0.281 0.741 0.259 1.68
## 11 0.358  0.004  0.758 -0.182 0.736 0.264 1.56
## 12 0.362  0.560  0.409  0.331 0.722 0.278 3.34
## 13 0.268  0.227  0.343  0.648 0.661 0.339 2.21
## 14 0.261 -0.048  0.281  0.423 0.328 0.672 2.51
## 15 0.341 -0.223  0.121  0.615 0.559 0.441 1.96
## 16 0.351  0.329  0.252 -0.019 0.295 0.705 2.82
## 17 0.435  0.423 -0.312  0.174 0.496 0.504 3.15
## 18 0.480  0.373 -0.015  0.071 0.375 0.625 1.94
## 19 0.578  0.400 -0.068 -0.082 0.506 0.494 1.86
## 20 0.256  0.736 -0.077 -0.395 0.768 0.232 1.84
## 21 0.235  0.776  0.011 -0.221 0.706 0.294 1.36
## 22 0.350  0.650 -0.382  0.006 0.691 0.309 2.22
##
##                         PC1   PC2   PC3   PC4
## SS loadings           5.596 3.479 2.624 1.881
## Proportion Var        0.254 0.158 0.119 0.086
## Cumulative Var        0.254 0.412 0.532 0.617
## Proportion Explained  0.412 0.256 0.193 0.139
## Cumulative Proportion 0.412 0.668 0.861 1.000
##
## Mean item complexity =  2.1
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.073
##
## Fit based upon off diagonal values = 0.936
```

```
unrotatedPCA$communality
```

```
## public speaking    law.politics        business           sales   merchandising
##       0.6864729       0.5302822       0.8631899       0.7113038       0.8484447
##          office        military       technical     mathematics         science
##       0.7099934       0.4199745       0.6925865       0.5331349       0.7414440
##      mechanical          nature     agriculture       adventure    recreational
##       0.7363333       0.7215051       0.6611104       0.3284178       0.5586898
##         medical          social        religion        teaching           music
##       0.2951819       0.4957999       0.3752616       0.5055699       0.7681034
##             art         writing
##       0.7061800       0.6912150
```

Eigenvalues

```
EV = eigen(voc.matrix)$values
EV
```

```
##  [1] 5.59637727 3.47853135 2.62393808 1.88134833 1.28794966 1.21677289
##  [7] 0.94491507 0.66186961 0.61826691 0.58644127 0.49916135 0.41020471
## [13] 0.39015558 0.35692400 0.28509856 0.26947987 0.23531220 0.20017131
## [19] 0.15734287 0.13834357 0.10983372 0.05156184
```

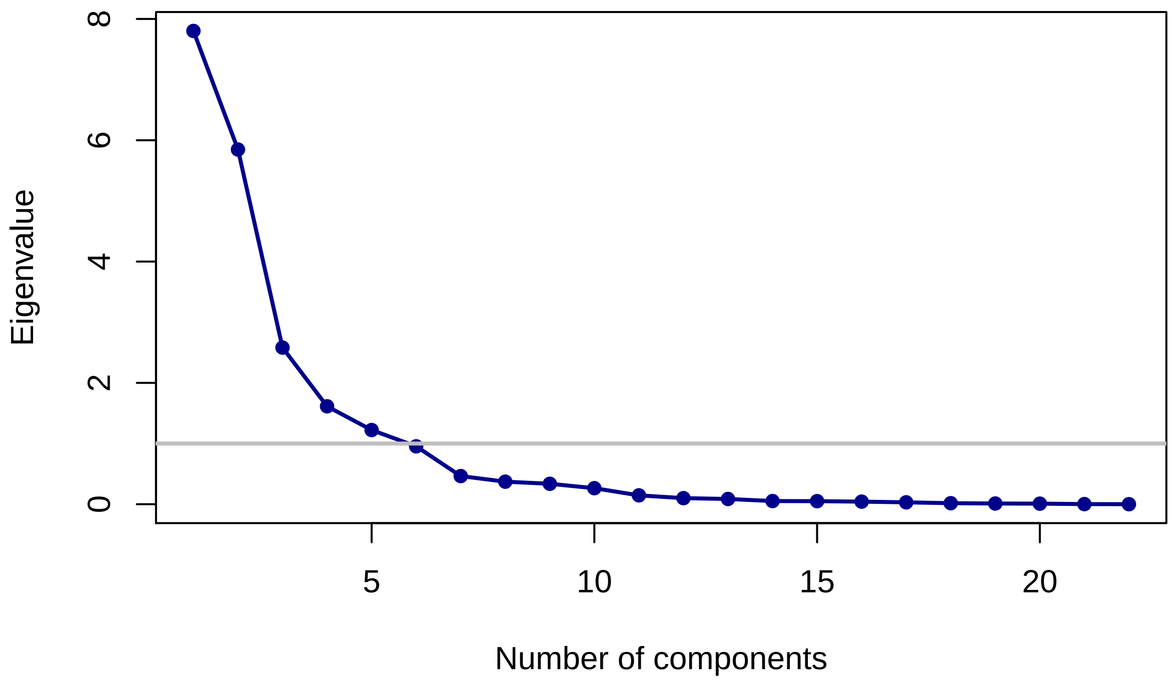
```
EV/length(EV)
```

```
##  [1] 0.254380785 0.158115061 0.119269913 0.085515833 0.058543166 0.055307859
##  [7] 0.042950685 0.030084982 0.028103042 0.026656421 0.022689152 0.018645669
## [13] 0.017734345 0.016223818 0.012959025 0.012249085 0.010696009 0.009098696
## [19] 0.007151948 0.006288344 0.004992442 0.002343720
```

The individual percentage of variance explained by a variable.

Scree plot

```
plot(eigen(cor(voc.matrix))$values,
     type = "o", # type of points
     col = "darkblue",
     pch = 16, # symbol type (here, filled circle)
     cex = 1, # size of plot symbols
     xlab = "Number of components",
     ylab = "Eigenvalue",
     lwd = 2) # line width
abline(h = 1, lwd = 2, col = "grey") # horizontal line at 1
```



The scree plot shows a slight elbow between the forth and the fifth components; from the fifth component on, the decline in eigenvalues is roughly linear and flatter than the first part of the graph. This suggests that four dimensions may be sufficient to capture to a reasonable extent the variability in the data. For simplicity's sake, we proceed with four dimentions.

```
unrotatedPCA$loadings
```

```
## 
## Loadings:
##        PC1    PC2    PC3    PC4   
##  [1,]  0.723        -0.389       
##  [2,]  0.615 -0.102 -0.366       
##  [3,]  0.810 -0.418        -0.159
##  [4,]  0.778 -0.266 -0.143 -0.120
##  [5,]  0.825 -0.357 -0.117 -0.164
##  [6,]  0.675 -0.435        -0.252
##  [7,]  0.403 -0.218  0.322  0.326
##  [8,]  0.688 -0.426  0.140 -0.133
##  [9,]        -0.131  0.585 -0.411
## [10,]  0.138  0.287  0.749 -0.281
## [11,]  0.358         0.758 -0.182
## [12,]  0.362  0.560  0.409  0.331
## [13,]  0.268  0.227  0.343  0.648
## [14,]  0.261         0.281  0.423
## [15,]  0.341 -0.223  0.121  0.615
## [16,]  0.351  0.329  0.252       
## [17,]  0.435  0.423 -0.312  0.174
## [18,]  0.480  0.373              
## [19,]  0.578  0.400              
## [20,]  0.256  0.736        -0.395
## [21,]  0.235  0.776        -0.221
## [22,]  0.350  0.650 -0.382       
## 
##                  PC1   PC2   PC3   PC4
## SS loadings    5.596 3.479 2.624 1.881
## Proportion Var 0.254 0.158 0.119 0.086
## Cumulative Var 0.254 0.412 0.532 0.617
```

PC1 explains 25.4% of variance in the data. For factors together explain about 62% of the total variability. I decided to proceed with this scenario since with more factors it'd become even more difficult to interpret.

Rotated PCA solution

```
rotatedPCA <- principal(r = voc.matrix,
                       nfactors = 4,
                       rotate="varimax",
                       scores = TRUE)
print(rotatedPCA,
      digits = 3 ,# round numbers to the 3d digits
      cut = 0.5, # to show only values > 0.5
      sort = TRUE # sort rows by factor
      )
```

```
## Principal Components Analysis
## Call: principal(r = voc.matrix, nfactors = 4, rotate = "varimax", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##    item    RC1    RC2    RC3    RC4    h2    u2  com
## 3     3  0.922                      0.863 0.137 1.03
## 5     5  0.911                      0.848 0.152 1.04
## 4     4  0.823                      0.711 0.289 1.10
## 6     6  0.819                      0.710 0.290 1.12
## 8     8  0.780                      0.693 0.307 1.28
## 1     1  0.606                      0.686 0.314 2.48
## 2     2  0.600                      0.530 0.470 1.97
## 20   20         0.801               0.768 0.232 1.40
## 21   21         0.799               0.706 0.294 1.22
## 22   22         0.785               0.691 0.309 1.24
## 19   19         0.634               0.506 0.494 1.50
## 17   17         0.603               0.496 0.504 1.75
## 12   12         0.567         0.553 0.722 0.278 2.56
## 18   18         0.544               0.375 0.625 1.55
## 16   16                             0.295 0.705 2.36
## 10   10                0.820        0.741 0.259 1.21
## 11   11                0.789        0.736 0.264 1.38
## 9     9                0.703        0.533 0.467 1.16
## 13   13                       0.776 0.661 0.339 1.20
## 15   15                       0.687 0.559 0.441 1.38
## 14   14                       0.560 0.328 0.672 1.10
## 7     7                       0.541 0.420 0.580 1.85
##
##                          RC1   RC2   RC3   RC4
## SS loadings            4.800 3.910 2.504 2.366
## Proportion Var         0.218 0.178 0.114 0.108
## Cumulative Var         0.218 0.396 0.510 0.617
## Proportion Explained   0.353 0.288 0.184 0.174
## Cumulative Proportion  0.353 0.641 0.826 1.000
##
## Mean item complexity =  1.5
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.073
##
## Fit based upon off diagonal values = 0.936
```

```
rotatedPCA$loadings
```

```
##
## Loadings:
##       RC1    RC2    RC3    RC4
##  [1,]  0.606  0.467 -0.295  0.122
##  [2,]  0.600  0.247 -0.307  0.121
##  [3,]  0.922
##  [4,]  0.823  0.164
##  [5,]  0.911  0.106
##  [6,]  0.819         0.191
##  [7,]  0.308         0.164  0.541
##  [8,]  0.780         0.230  0.165
##  [9,]  0.113 -0.135  0.703       
## [10,]         0.223  0.820  0.101
## [11,]  0.198         0.789  0.264
## [12,] -0.126  0.567  0.280  0.553
## [13,]         0.219         0.776
## [14,]                       0.560
## [15,]  0.231 -0.102 -0.151  0.687
## [16,]         0.413  0.286  0.190
## [17,]  0.165  0.603 -0.287  0.151
## [18,]  0.197  0.544         0.198
## [19,]  0.308  0.634              
## [20,]         0.801  0.180 -0.303
## [21,] -0.142  0.799  0.179 -0.125
## [22,]         0.785 -0.267       
## 
##                  RC1   RC2   RC3   RC4
## SS loadings    4.800 3.910 2.504 2.366
## Proportion Var 0.218 0.178 0.114 0.108
## Cumulative Var 0.218 0.396 0.510 0.617
```

We look at the attributes that are loading highly in the first factor, [1],[2],[3],[4],[5],[6],[8] which are “public speaking”,“law and politics”,“business management”,“sales”,“merchandising”,“office practice”,“technical supervision”. Attributes that are loading highly in the second factor, [17],[18],[19],[20],[21],[22] which are “social service”,“religious activities”,“teaching”," music“,”art“,”writing“. In the third factor, attributes like [9],[10],[11] are loading highly. They are denoted for”mathematics“,”science“,”mechanical“. In the last factor, [7],[12],[13],[14],[15] are the atrributes that are loading most highly, which are”military activities“,”nature“,”agriculture“,”adventure“,”recreational leadership“. We can see that in some dimentions, some of the attributes are cleary highly correlated, however, some might not seem to perfectly fit in. This is a trade-off between parsimony and interpretability that we made, with four dimentions, but can only explain 62% of the variability. Nonetheless, we can still label the four factors that would make the most sense. For the first factor, the attributes are at most closely related to business activities, thus we label it as **business profession**. Noteably,”public speaking“,”law and politics" seem to not belong to this factor. But as we mentioned, with only four factors, it’s very likely that some of the attributes may not fit in. They might belong to a fifth or even sixth factor. For the second factor, the attributes seem to be related to societal activities. We label it as **social work**. Thirdly, the factor that contains three attributes can be easily labeld as **scientific profession** Last, we can label the factor as **non-scientific professions**, since the attributes demonstrate not much scientific activities.