IS507 - Data, Statistical Models and Information - Assignment 4

Problem 1: Journal Article Review

A) How are they applying Factoring Analysis?

Solution:

The experiment was performed during and after the Covid 19 period. Colleges were closed and so the nursing institutions wanted to check how the students would react to the E Learning method. The study makes use of the Exploratory Factor Analysis to check if the attitudes of the Filipino nursing students towards the E learning holds true and is reliable or not.

B) What kind of factor rotation do they use?

Solution:

The type of rotation used is 'varimax' rotation when performing the Exploratory Factor Analysis using the Principal Component Analysis.

C) How many factors do they concentrate on in their analysis? How did they arrive at these number of factors?

Solution:

They have provided a 5-point Likert scale for each of the original 11 variables in the dataset. They have used close to 10 observations for each variable and hence 111 rows.
The Exploratory Factor Analysis using PCA was performed on the data. They have set a threshold of 0.6 based on the previous study/analysis done by experts for the factor loadings and any values below them are rejected. For 2 variables 6 and 11 the values were less than 0.6 and hence they finally arrived at the 9 variables used.

D) Explain the breakdown of the factors and the significance of their names.

Solution:

The 9 factors that they arrive at are as follows:

1. I am interested in studying courses that utilize e-learning
2. I think that e-learning promotes my learning experiences
3. Presenting courses on the internet makes learning more efficient
4. I intend to use e-learning tools during the semester if available
5. I am positive about e-learning.
6. I would prefer to have courses on the internet rather than in the classroom or face-to-face.
7. Online learning is more comfortable and enjoying to me.
8. E-learning is a favorable alternative to the pen-paper based system
9. E-learning is not an efficient learning method

They have not provided us with the original dataset as to what were the initial names and whether the above 9 items are the actual names of the variables.

We will assume them as the names of the variables.
The variable names seem justifiable because these are the metrics which are definitely used in the questionnaire for asking the students to rate their experiences to the e-learning systems, how comfortable they are, whether they look forward to adopting the technology for studying, whether it is better than the conventional pen-paper based method and if it is efficient or not. Other variables that also are critical include whether they would prefer online or face-to-face classes and whether the e-learning does provide any positive experience on the learning of the students. Comfort of the students is also important whether they prefer to use the e-learning tools frequently or not.

E) How do they evaluate the stability of the components (i.e. factorability)?

Solution:

The stability of the PCA and FA has been verified by the following metrics:

1. The KMO test is used to check the adequacy of the samples is terms of correlation. The value is 0.9 which is very good and it ensures that the sample is adequately correlated.
2. The Bartlett's test of sphericity has a value of 644.38 which suggests that the data is suitable for performing PCA and that the correlation matrix of the variables is not an identity matrix.
3. The item mean, standard deviation and item total correlation are calculated.
4. The item mean ranges from 2.28 to 3.07 and the item-total correlation ranges from 0.409 to 0.854 which is more than the recommended 0.3 as discussed in paper in previous studies.
5. The Cronbach's alpha coefficient has an overall value of 0.917 which is greater than the recommended value of 0.7 based on the results of previous studies as discussed in the paper. This suggests that the internal consistency of the set of items inside the group is high and it indicate that it is good.

F) Do they use these factors in later analysis, such as regression? If so, what do they discover?

Solution:

The factors arrived at in the paper are not used in any further analysis. They have just calculated the mean and standard deviations of the 9 items obtained along with Cronbach's alpha if each variable is removed.

G) What overall conclusions does Factor Analysis allow them to draw?

Solution:

Using the metrics discussed above they conclude that the **study done was successful** in understanding the validity and reliability of the of the attitudes of the Filipino students towards the E-learning scale. The 9-item scale is suitable for use among the Filipino students to understand their attitudes towards the e-learning technology.
The instrument and the construct validity were proven using the dataset and the metrics of Cronbach's reliability, KMO, total item correlation and the Bartlett's test.

Problem 2: Principal Component Analysis

Dataset Cleaning:

```r
# Load library readr to read the dataset
library(readr)
library(plyr)
library(dplyr)
library(tidyr)
library(stringr)

# Set working directory
setwd("~/Desktop/IS507 - Data, Statistical Models and Information/Assignments/Assignment_5")

# Import the dataset using read_csv()
raw_dataset <- read.table("16personality.csv", sep="\t", header=TRUE)
dataset <- raw_dataset

# Shape of dataset
dim(dataset)

# Count of missing values
sum(is.na(dataset))

# Delete rows with missing values
dataset <- na.omit(dataset)

# Shape of new dataset after listwise deletion
dim(dataset)

new_dataset <- raw_dataset[, c(1:162)]
new_dataset
```

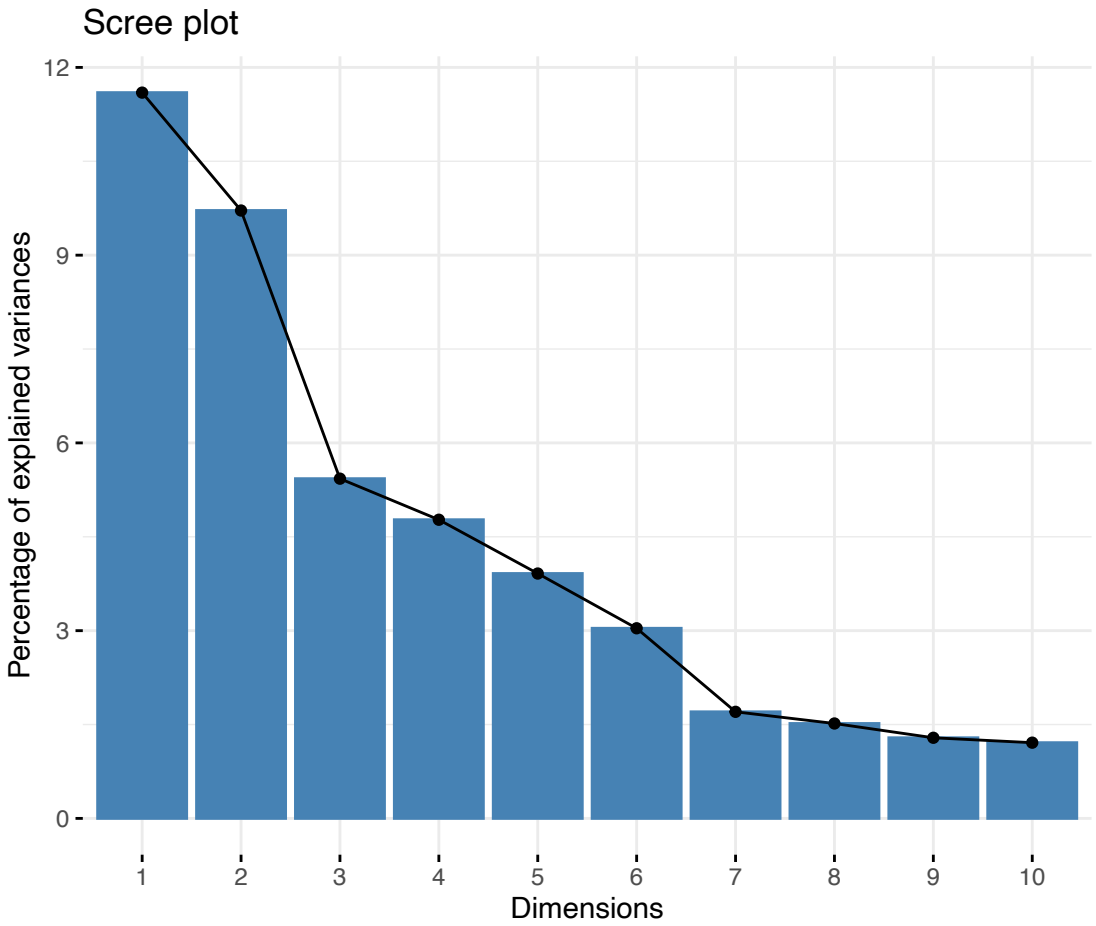
```
Console  Terminal  Background Jobs
R 4.2.1 · ~/Desktop/IS507 - Data, Statistical Models and Information/Assignments/Assignment_5/
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:plyr':

    arrange, count, desc, failwith, id, mutate, rename, summarise, summarize

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> library(tidyr)
> library(stringr)
>
>
> # Set working directory
> setwd("~/Desktop/IS507 - Data, Statistical Models and Information/Assignments/Assignment_5")
>
> # Import the dataset using read_csv()
> raw_dataset <- read.table("16personality.csv", sep="\t", header=TRUE)
> dataset <- raw_dataset
>
> # Shape of dataset
> dim(dataset)
[1] 49159   169
>
> # Count of missing values
> sum(is.na(dataset))
[1] 7
>
> # Delete rows with missing values
> dataset <- na.omit(dataset)
>
> # Shape of new dataset after listwise deletion
> dim(dataset)
[1] 49152   169
>
> library(gmodels)
>
> new_dataset <- raw_dataset[, c(1:162)]
> new_dataset
```

A) How many components are determined from the scree plot using the knee method?

Solution:

We run the initial PCA model on the cleaned dataset using the prcomp() function and provide scale = TRUE.
When we plot the scree plot of the resultant PCA model with the knee method we find the first critical bend or knee at the dimension number 7 because from this point the percentage of variance explained by the components is almost equal and the graph flattens out.
So, the number of components determined from the scree plot using the knee method is 7.

```
43
44  # A) Number of components using Scree Plot and Knee Method
45
46  # You need to run the dataset in PCA function once before you decide on the number of components
47  pca = prcomp(new_dataset, scale = TRUE)
48  fviz_eig(pca)
```



B) What number of components would you using the eigenvalue method?

Solution:

We run the initial PCA model on the cleaned dataset using the prcomp() function and provide scale = TRUE.
When we plot the PCA model and use the abline(1,0) function to find out which values are greater than 1, we get only the first 10 components. But there may be more values and hence we use the

get_eig() function we get all the 162 components with their Eigen values. Since, we are interested in those whose value is more than 1 and which contribute to the maximum proportion of variance. When we find them we get 23 such components.
So, the number of components determined from the Eigen values is 23.

```
50  # B) Number of components using Eigen values
51  plot(pca)
52  abline(1,0)
53  sum(get_eig(pca)$eigenvalue>1)
```

```
> plot(pca)
> abline(1,0)
> sum(get_eig(pca)$eigenvalue>1)
[1] 23
```

C) Based upon your answers from parts A and B, what number of components would you wish to start with for the model? Run the PCA model.

Solution:

Based on the answers from parts A and B, I have chosen to go with the number of components obtained from the A using the knee plot = 7. We do this because, there is a possibility that the number of components determined by the Eigen value method may be very large and those many variables may tend to overfit the data.
We run the new PCA model on the cleaned dataset using the principal() function from 'psych' library with nfactors=7, rotation = 'varimax' and scores = TRUE.

First PCA:

```
70  options(max.print = 10000)
71
72  pca_with_7_components = principal(new_dataset, rotate="varimax", nfactors = 7, scores = TRUE)
73
74  comps = print(pca_with_7_components$loadings, cutoff=0.4, sort=T)
75
76  # New components with cutoff 0.45 trying to remove cross loadings
77  comps = print(pca_with_7_components$loadings, cutoff=0.45, sort=T)
```

```
> # New components with cutoff 0.45 trying to remove cross loadings
> comps = print(pca_with_7_components$loadings, cutoff=0.45, sort=T)

Loadings:
    RC6    RC1    RC2    RC3    RC4    RC5    RC7
C5  -0.568
C6   0.644
C7   0.625
C8   0.579
C9   0.586
C10  0.647
L1   0.590
L2   0.647
L3   0.681
L4   0.663
L5   0.579
L7   0.601
L9  -0.540
P1   0.664
P2   0.594
P5   0.517
P8  -0.505
E1         -0.626
E2         -0.596
E4         -0.572
E8          0.624
G1         -0.504
G2         -0.574
G4         -0.509
G5         -0.503
G6          0.585
G7          0.540
G9          0.590
G10         0.631
K3          0.664
N1          0.551
N4          0.659
N6          0.525
B1                 0.513
B2                 0.513
B3                 0.569
B4                 0.589
B5                 0.560
D1                 0.579
D2                 0.522
D4                 0.505
```

```
D4                 0.505
D5                 0.563
D6                 0.555
M4                 0.556
O4                 0.514
O5                 0.564
A1                        0.524
A2                        0.593
A3                        0.568
A4                        0.555
A5                        0.526
A6                        0.539
A7                        0.544
F4                        0.530
I7                        0.584
I8                        0.581
I9                        0.640
I10                       0.524
N8                        0.506
N9                        0.504
P9                        0.516
E6                               0.506
F6                               0.620
F7                               0.675
F9                               0.631
F10                              0.522
J7                               0.521
O7                               0.535
O8                               0.544
B12                                     0.543
H9                                      0.549
J9                                      0.508
J10                                     0.522
M6                                      0.609
M7                                      0.561
M8                                      0.597
M9                                      0.594
M10                                     0.583
K1                                             0.541
K6                                            -0.566
K7                                            -0.614
K9                                            -0.576
A8
A9
A10
B6                 0.468
```

```
A8
A9
A10
B6                 0.468
B7
B8
B9
B10  0.484
B11
B13                                     0.456
C1
C2
C3                 0.463
C4  -0.451
D3                 0.454
D7
D8
D9
D10
E3
E5
E7                                      0.452
E9
E10
F1
F2                        0.500
F3
F5
F8
G3         -0.455
G8          0.489
H1
H2
H3
H4
H5
H6
H7
H8                                      0.499
H10
I1
I2
I3
I4
I5
I6
```

```
I2
I3
I4
I5
I6
J1                               0.474
J2
J3                               0.483
J4
J5
J6                               0.491
J8
K2          0.465
K4                                             0.469
K5          0.476
K8
K10
L6
L8
L10 -0.486
M1
M2                 0.462
M3
M5
N2          0.458
N3          0.499
N5
N7
N10
O1
O2
O3                 0.476
O6                               0.454
O9                               0.468
O10
P3
P4
P6
P7

                   RC6    RC1    RC2    RC3   RC4   RC5   RC7
SS loadings     12.067 11.382 10.938 10.902 8.437 7.536 3.806
Proportion Var   0.074  0.070  0.068  0.067 0.052 0.047 0.023
Cumulative Var   0.074  0.145  0.212  0.280 0.332 0.378 0.402
>
```

Final 6th PCA:

```
# PCA 6
# Removing variables with 0 intercorrelations
removed_vars_5 = removed_vars_4 %>% select(-c('O6'))

pca5_with_7_components = principal(removed_vars_5, rotate="varimax", nfactors = 7, scores = TRUE)

comps = print(pca5_with_7_components$loadings, cutoff=0.4, sort=T)

# New components with cutoff 0.448 trying to remove cross loadings
comps = print(pca5_with_7_components$loadings, cutoff=0.48, sort=T)
```

```
> # New components with cutoff 0.448 trying to remove cross loadings
> comps = print(pca5_with_7_components$loadings, cutoff=0.448, sort=T)

Loadings:
    RC2    RC1    RC4    RC3    RC5    RC6    RC7
B10  0.557
C5  -0.531
C6   0.678
C7   0.662
C8   0.594
C9   0.622
C10  0.681
L1   0.637
L2   0.672
L3   0.724
L4   0.686
L5   0.605
L7   0.632
P1   0.667
P2   0.595
P5   0.557
E1          0.700
E2          0.704
E4          0.670
E8         -0.646
G1          0.580
G2          0.683
G3          0.538
G4          0.593
G5          0.584
G6         -0.583
G7         -0.537
G9         -0.576
G10        -0.609
K3         -0.624
N4         -0.558
N8          0.524  0.458
A1                 0.526
A2                 0.565
A3                 0.551
A4                 0.547
A6                 0.535
A7                 0.517
I7                 0.617
I8                 0.611
I9                 0.673
I9                 0.673
I10                0.545
P8                 0.515
P9                 0.551
B1                        0.541
B2                        0.571
B3                        0.613
B4                        0.629
B5                        0.598
B6                        0.501
D1                        0.587
D2                        0.536
D4                        0.511
D5                        0.582
D6                        0.558
M4                        0.564
O4                        0.542
O5                        0.597
B12                              0.521
H8                               0.505
H9                               0.547
J9                               0.555
J10                              0.572
M6                               0.659
M7                               0.592
M8                               0.647
M9                               0.629
M10                              0.607
F2                                     -0.515
F4                                     -0.542
F6                                      0.741
F7                                      0.715
F9                                      0.746
F10                                     0.545
J6                                      0.511
K1                                             0.671
K4                                             0.573
K5                                             0.584
K6                                            -0.611
K7                                            -0.660
K9                                            -0.629
A5                 0.500
L9  -0.494
N9                 0.492
O7                                      0.495
O8
```

```
                   RC2   RC1   RC4   RC3   RC5   RC6   RC7
SS loadings      8.370 7.711 7.268 6.031 4.438 4.213 3.322
Proportion Var   0.097 0.090 0.085 0.070 0.052 0.049 0.039
Cumulative Var   0.097 0.187 0.271 0.342 0.393 0.442 0.481
>
```

D) For the number of components in part C, give the formula for the first component.

Solution:

The formula for the first component based on part C is as follows:

The first component is RC2:

RC2 = (0.557) * B10
+ (-0.531) * C5
+ (0.678) * C6
+ (0.662) * C7
+ (0.594) * C8
+ (0.622) * C9
+ (0.681) * C10
+ (0.637) * L1
+ (0.672) * L2
+ (0.724) * L3
+ (0.686) * L4
+ (0.605) * L5
+ (0.632) * L7
+ (0.667) * P1
+ (0.595) * P2
+ (0.557) * P5
+ (-0.494) * L9

E) Give a brief interpretation of the components after rotation. What do these components mean? What names might you give for each of the components?

Solution:

Since, we have 7 components we need to suggest appropriate names for these components as they are critical and may be used for further analysis. So, giving meaningful names to the components is very important.

Principal Component 1: RC2 —> Irascible - The positive variables of B10, C6, C7, C8, C9, C10, L1, L2, L3, L4, L5, L7, P1, P2, P5 indicate a person who can be angered easily, is always sad, and dislikes themselves. The negative variables of C5 and L9 contradict and show a person who cannot be flustered easily. So, this indicates a Irascible personality.

Principal Component 2: RC1 —> Gregarious - The positive variables like E1, E2, E4, G1, G2, G3, G4, G5, N8 define the characteristics of a person who is very social and likes to mingle with others. This person does not shy away from being people's favourite. The negative variables like E8, G6, G7, G9, G10, K3, N4 indicate person who is shy and does not talk to people. So. this indicates an Extrovert nature.

Principal Component 3: RC4 —> Generous - The positive variables from A1, A2, A3, A4, A6, A7, I7, I8, I9, I10, P8, P9, A5, A9 indicate a person who tries to soothe others and takes time to understand them. It also shows a person who does good to others and believes people around are kind. So, this indicates Generosity.

Principal Component 4: RC3 —> Tenacious - The positive variables of B1, B2, B3, B4, B5, B6, D1, D2, D4, D5, D6, M4, O4, O5 suggest a person who is smart, can grasp things quickly, has a take-charge attitude and one who has knack for doing work correctly. So, this indicates a person who is Tenacious.

Principal Component 5: RC5 —> Veritable - The positive variables of B12, H8, H9, J9, J10, M6, M7, M8, M9, M10 is a type of person who believes in real facts over fiction and does not like vague or philosophical discussions. So, this is a person who just likes to work with existing entities and is Veritable.

Principal Component 6: RC6 —> Recalcitrant - The positive variables of F6, F7, F9, F10, J6 indicate a person who is stubborn and does not like rules and cannot resist a superior person. The negative variables describe a person who follows rules and stays within means. So, this indicates a person who is Recalcitrant and opposes authority.

Principal Component 7: RC7 —> Introvert - The positive variables of K1, K2, K5 is a type of person who prefers to keep everything within themselves and do no reveal or speak about their feelings to others. The negative variables of K6, K7, K9 on the other hand tell us a person likes speaking about themselves and their feelings to others. So, this is an Introvert person.

F) What are the highest and lowest scores for each principal component conducted in Part C?

Solution:

In this problem, we have used the last (6th PCA) for the analysis and calculated the maximum and the minimum scores for the same for each principal component.
Here, we can see that that scores are in negative and positive which indicates that the maximum scores are that many units on the positive side or greater than the mean and vice versa for the negative values.

- For RC2 (Irascible) more people tend to be less angered as compared to others, because the negative scale is high and more people are calm.
- For RC1 (Gregarious) more people tend to be less social as compared to others, because the negative scale is high and more people prefer to be alone.
- For RC4 (Generous) more people tend to be not good to others or do not help others, because the negative scale is high and more people are selfish.
- For RC3 (Tenacious) more people are less tenacious and not courageous or tend to follow others rather than taking charge, because the negative scale is high and more people are followers.
- For RC5 (Veritable) a high number of people believe in reality and will base their decisions on facts. The negative scale is small which means there are a few people who still believe in fiction.
- For RC6 (Recalcitrant) majority of the people follow rules and obey the superior authorities, because the positive scale is low. But few people do break rules and disobey.
- For RC7 (Introvert) the positive value is less than the negative and indicates that the majority of the population does speak about their feelings to others.

```r
# F) Max and Min scores for each component

scores <- pca5_with_7_components$scores

max_scores = apply(scores, 2, max)
min_scores = apply(scores, 2, min)
print(max_scores)
print(min_scores)
```

```
> max_scores = apply(scores, 2, max)
> min_scores = apply(scores, 2, min)
> print(max_scores)
     RC2      RC1      RC4      RC3      RC5      RC6      RC7
3.173754 3.535169 3.379689 3.258672 4.568548 3.816834 3.822719
> print(min_scores)
      RC2       RC1       RC4       RC3       RC5       RC6       RC7
-3.438674 -3.897457 -5.450705 -5.538684 -3.447616 -4.870815 -4.156276
```

G) Finally, run a common factor analysis on the same data. Is there a difference between the Principal Component Analysis and the factor analysis? Does the factor analysis change your ability to interpret the results practically?

Solution:

There is a difference between PCA and FA.
PCA takes into account all the different types of variances - shared, unique and error variances. So, it chooses correlations of variables higher as the variances is higher in PCA due to the presence of unique and error variances.
FA just takes into account only the shared variances.

Here we come to know that the factor RC2 (Irascible) has variables which are same in the PCA and FA. But for the remaining components there are a few variables which are there in FA but not in PCA because we have refined the PCA a lot but not FA.

Also there a few variables like L9, O7, O8, H8, J6 which do not make it in the FA as it just makes use of the shared variances.

When we check the results of the FA and PCA we see that the RC2, RC1 component and Factor1 and Factor2 are same and so we will conclude the same. Also when we look at the outputs for them it may looked jumbled but the remaining components in PCA and factors in FA convey the same meaning for both FA and PCA except 1 or 2 variables which may not be common between them. Overall, the analysis does not change with the results of FA and PCA.

```r
# G) Factor analysis


factor_analysis = factanal(removed_vars_5,7)
print(factor_analysis$loadings, cutoff = 0.4, sort=TRUE)

print(factor_analysis$loadings, cutoff = 0.47, sort=TRUE)

```

```
> print(factor_analysis$loadings, cutoff = 0.47, sort=TRUE)

Loadings:
    Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
B10  0.541
C5  -0.506
C6   0.659
C7   0.645
C8   0.577
C9   0.599
C10  0.663
L1   0.609
L2   0.649
L3   0.700
L4   0.652
L5   0.578
L7   0.595
P1   0.656
P2   0.581
P5   0.535
E1           0.667
E2           0.660
E4           0.615
E8          -0.599
G1           0.560
G2           0.658
G3           0.506
G4           0.576
G5           0.573
G6          -0.581
G7          -0.524
G9          -0.574
G10         -0.600
K3          -0.621
N4          -0.531
A1                   0.508
A2                   0.557
A3                   0.515
A4                   0.516
A6                   0.515
I7                   0.555
I8                   0.554
I9                   0.614
P8                   0.523
```

```
B1                      0.520
B2                      0.539
B3                      0.569
B4                      0.602
B5                      0.571
D1                      0.557
D5                      0.545
D6                      0.532
M4                      0.543
O5                      0.535
M6                              0.624
M7                              0.545
M8                              0.611
M9                              0.574
M10                             0.570
F2                                     -0.513
F4                                     -0.557
F6                                      0.749
F7                                      0.668
F9                                      0.755
K1                                              0.615
K6                                             -0.614
K7                                             -0.668
K9                                             -0.590
A5                  0.480
A7                  0.480
B6                      0.480
B12                             0.483
D2                      0.500
D4                      0.481
F10                                    0.487
H8
H9                              0.475
I10                 0.489
J6
J9                              0.478
J10                             0.498
K4                                              0.497
K5                                              0.499
L9
N8          0.478
N9                  0.484
O4                      0.472
O7
O8
P9                  0.496
```

```
                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
SS loadings       7.866   7.183   6.781   5.493   3.845   3.612   2.693
Proportion Var    0.091   0.084   0.079   0.064   0.045   0.042   0.031
Cumulative Var    0.091   0.175   0.254   0.318   0.362   0.404   0.436
>
```