Assignment 01 - Factor Analysis of a real-world problem

S17403

1. Introduction

The statistical tool used in this study, Factor Analysis. It uses the correlation structure amongst observed variables to model a smaller number of unobserved, latent variables known as factors. Factor analysis simplifies a complex dataset by taking a larger number of observed variables and reducing them to a smaller set of unobserved factors. It is widely employed in various fields, including psychology, sociology, marketing, and finance, to explore complex data structures and understand the underlying dimensions that influence them.

The aim of the present study was to use exploratory and confirmatory factor analysis (CFA) to investigate the factorial structure of the 14 variables which help to predict the Quality of water. This study opens the doors to exploiting Factor Analysis as a tool for use in the field of analyzing the quality of water.

2. Methodology

Dataset Description: - This is a set of data on water quality in an urban environment. Dataset Link: - https://www.kaggle.com/datasets/mssmartypants/water-quality

All attributes are numeric variables and they are listed below:

- aluminum - dangerous if greater than 2.8
- ammonia - dangerous if greater than 32.5
- arsenic - dangerous if greater than 0.01
- barium - dangerous if greater than 2
- cadmium - dangerous if greater than 0.005
- chloramine - dangerous if greater than 4
- chromium - dangerous if greater than 0.1
- copper - dangerous if greater than 1.3
- fluoride - dangerous if greater than 1.5
- bacteria - dangerous if greater than 0
- viruses - dangerous if greater than 0
- lead - dangerous if greater than 0.015
- nitrates - dangerous if greater than 10
- nitrites - dangerous if greater than 1
- mercury - dangerous if greater than 0.002
- perchlorate - dangerous if greater than 56
- radium - dangerous if greater than 5
- selenium - dangerous if greater than 0.5
- silver - dangerous if greater than 0.1
- uranium - dangerous if greater than 0.3
- is_safe - class attribute {0 - not safe, 1 - safe}

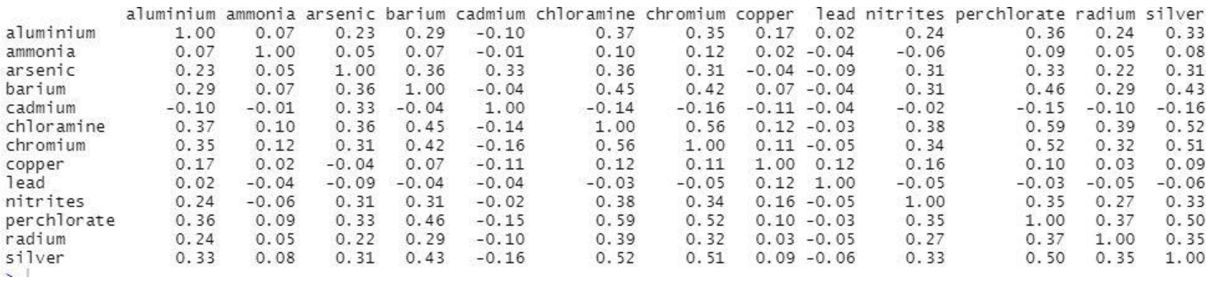
key steps

- Data Preparation: Clean and preprocess data to ensure it is suitable for factor analysis. This involves handling missing values, checking for outliers, and appropriately transforming variables if necessary.
- Determine the Factor Analysis Technique: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). EFA is used when to identify latent factors without pre-specified hypotheses, while CFA is employed to test a pre-established factor structure.
- Factor Extraction: Apply the chosen factor analysis technique to extract the underlying factors. During this step, the analysis identifies the number of factors that best explain the relationships among the observed variables. Methods can be used,
    - principal component analysis (PCA) or maximum likelihood estimation (MLE)
    - General methods used in determining the number of factors
        - Cumulative proportion of at least 0.80
        - Eigen Values of at least one
        - Based on Scree Plot
- Factor Rotation: Once the factors are extracted, need to rotate them to enhance interpretability. Rotation aims to achieve a simpler and more understandable factor structure by minimizing the number of variables that load heavily on each factor.
    - rotation methods include varimax, oblique
- Interpretation of Results: Interpret the factor analysis results to understand the meaning and characteristics of each factor.

3. Results and discussion

EXPLORATORY FACTOR ANALYSIS

Correlation matrix for the data: -



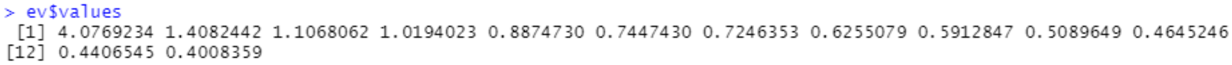
```
            aluminium ammonia arsenic barium cadmium chloramine chromium copper  lead nitrites perchlorate radium silver
aluminium        1.00    0.07    0.23   0.29   -0.10       0.37     0.35   0.17  0.02     0.24        0.36   0.24   0.33
ammonia          0.07    1.00    0.05   0.07   -0.01       0.10     0.12   0.02 -0.04    -0.06        0.09   0.05   0.08
arsenic          0.23    0.05    1.00   0.36    0.33       0.36     0.31  -0.04 -0.09     0.31        0.33   0.22   0.31
barium           0.29    0.07    0.36   1.00   -0.04       0.45     0.42   0.07 -0.04     0.31        0.46   0.29   0.43
cadmium         -0.10   -0.01    0.33  -0.04    1.00      -0.14    -0.16  -0.11 -0.04    -0.02       -0.15  -0.10  -0.16
chloramine       0.37    0.10    0.36   0.45   -0.14       1.00     0.56   0.12 -0.03     0.38        0.59   0.39   0.52
chromium         0.35    0.12    0.31   0.42   -0.16       0.56     1.00   0.11 -0.05     0.34        0.52   0.32   0.51
copper           0.17    0.02   -0.04   0.07   -0.11       0.12     0.11   1.00  0.12     0.16        0.10   0.03   0.09
lead             0.02   -0.04   -0.09  -0.04   -0.04      -0.03    -0.05   0.12  1.00    -0.05       -0.03  -0.05  -0.06
nitrites         0.24   -0.06    0.31   0.31   -0.02       0.38     0.34   0.16 -0.05     1.00        0.35   0.27   0.33
perchlorate      0.36    0.09    0.33   0.46   -0.15       0.59     0.52   0.10 -0.03     0.35        1.00   0.37   0.50
radium           0.24    0.05    0.22   0.29   -0.10       0.39     0.32   0.03 -0.05     0.27        0.37   1.00   0.35
silver           0.33    0.08    0.31   0.43   -0.16       0.52     0.51   0.09 -0.06     0.33        0.50   0.35   1.00
```

Figure 1:Corrrelation Matrix

The highest correlation shows between perchlorate and chloramine variables.

Determining the number of factors:-

- After standardizing the variables, then calculated the Eigenvalues



```
> ev$values
 [1] 4.0769234 1.4082442 1.1068062 1.0194023 0.8874730 0.7447430 0.7246353 0.6255079 0.5912847 0.5089649 0.4645246
[12] 0.4406545 0.4008359
```

- And also use the scree plot

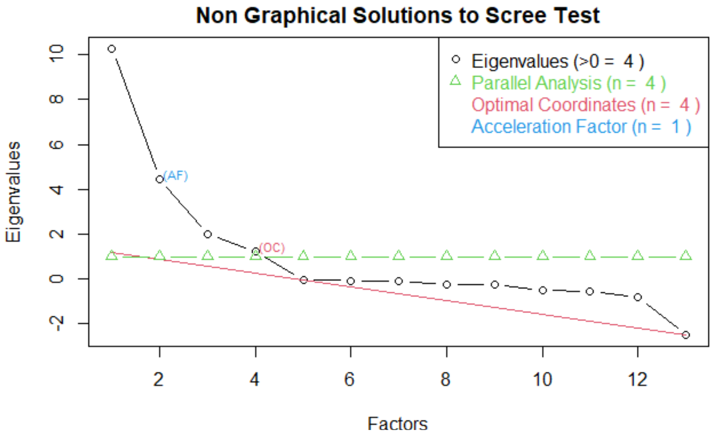
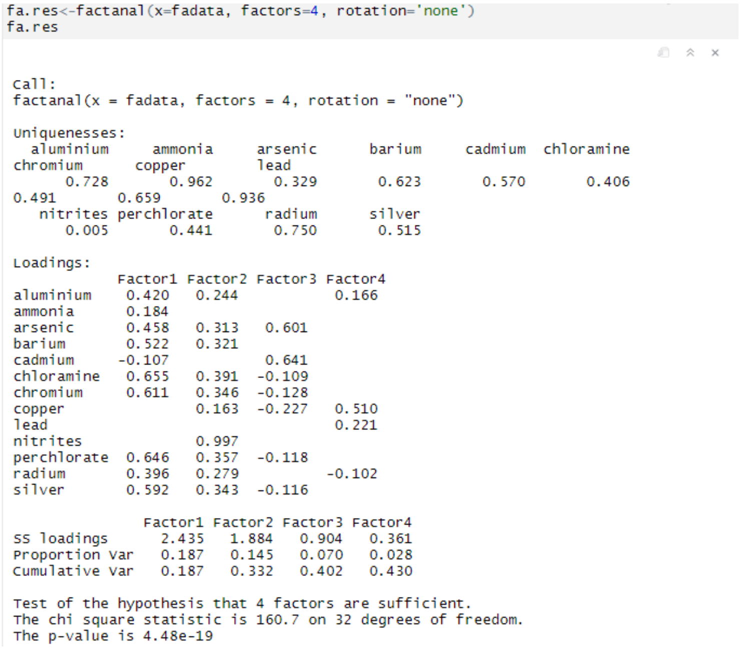


Figure 2:Scree plot

Once the number of factors (as four) is decided, conduct exploratory factor analysis using the R function factanal().First do the factor analysis with no rotation.Its hard to interpret using these loadings. Therefore, again used the varimax rotaion to get an interpretable output.

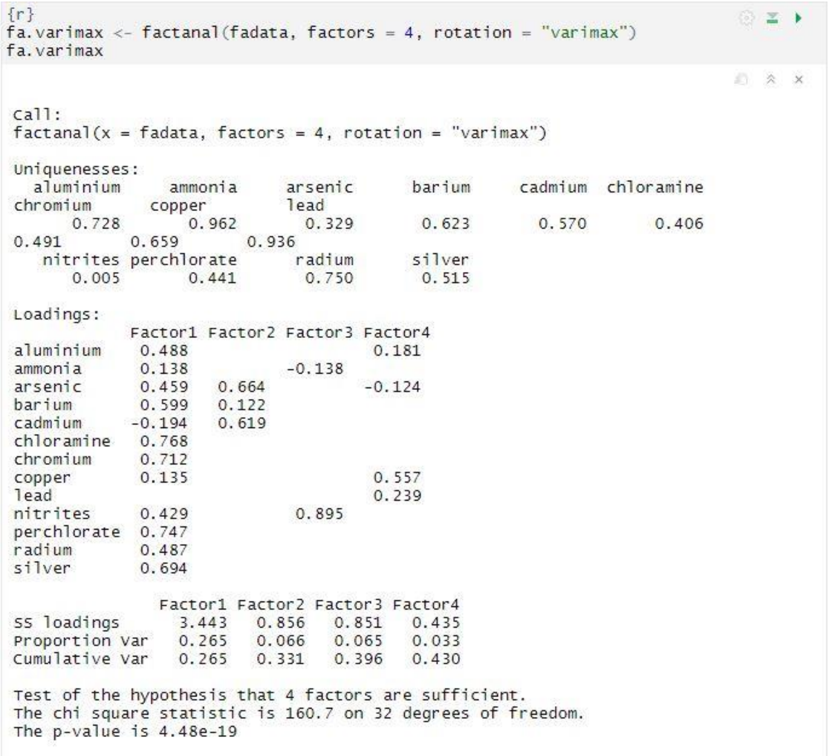


```
fa.res<-factanal(x=fadata, factors=4, rotation='none')
fa.res
```

```
Call:
factanal(x = fadata, factors = 4, rotation = "none")

Uniquenesses:
  aluminium     ammonia     arsenic      barium     cadmium  chloramine 
chromium      copper        lead 
      0.728       0.962       0.329       0.623       0.570       0.406 
0.491       0.659       0.936 
   nitrites perchlorate      radium      silver 
      0.005       0.441       0.750       0.515 

Loadings:
            Factor1 Factor2 Factor3 Factor4
aluminium    0.420   0.244           0.166 
ammonia      0.184                         
arsenic      0.458   0.313   0.601         
barium       0.522   0.321                 
cadmium     -0.107           0.641         
chloramine   0.655   0.391  -0.109         
chromium     0.611   0.346  -0.128         
copper               0.163  -0.227   0.510 
lead                                 0.221 
nitrites             0.997                 
perchlorate  0.646   0.357  -0.118         
radium       0.396   0.279          -0.102 
silver       0.592   0.343  -0.116         

               Factor1 Factor2 Factor3 Factor4
SS loadings      2.435   1.884   0.904   0.361
Proportion Var   0.187   0.145   0.070   0.028
Cumulative Var   0.187   0.332   0.402   0.430

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 160.7 on 32 degrees of freedom.
The p-value is 4.48e-19 
```

Figure 3:Factoring with no rotation

Varimax Rotation

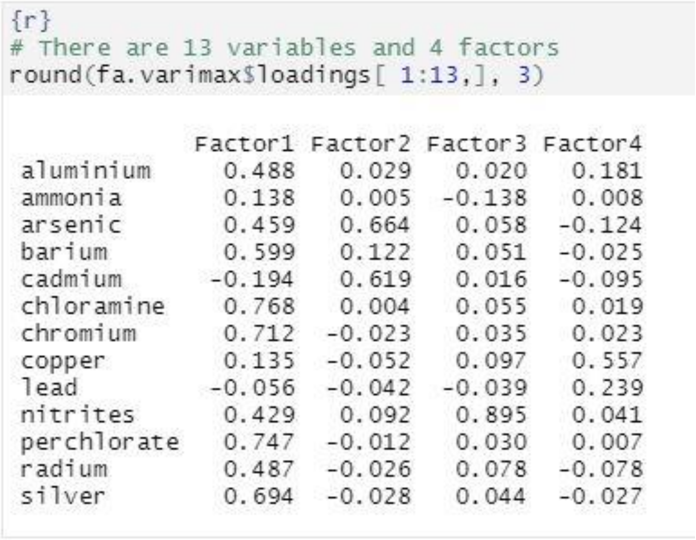


```
{r}
fa.varimax <- factanal(fadata, factors = 4, rotation = "varimax")
fa.varimax
```

```
Call:
factanal(x = fadata, factors = 4, rotation = "varimax")

Uniquenesses:
  aluminium     ammonia     arsenic      barium     cadmium  chloramine 
chromium      copper        lead 
      0.728       0.962       0.329       0.623       0.570       0.406 
0.491       0.659       0.936 
   nitrites perchlorate      radium      silver 
      0.005       0.441       0.750       0.515 

Loadings:
            Factor1 Factor2 Factor3 Factor4
aluminium    0.488                   0.181 
ammonia      0.138          -0.138         
arsenic      0.459   0.664          -0.124 
barium       0.599   0.122                 
cadmium     -0.194   0.619                 
chloramine   0.768                         
chromium     0.712                         
copper       0.135                   0.557 
lead                                 0.239 
nitrites     0.429           0.895         
perchlorate  0.747                         
radium       0.487                         
silver       0.694                         

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.443   0.856   0.851   0.435
Proportion Var   0.265   0.066   0.065   0.033
Cumulative Var   0.265   0.331   0.396   0.430

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 160.7 on 32 degrees of freedom.
The p-value is 4.48e-19 
```

Figure 4:Factoring with varimax rotation

Factor loading of varimax rotation:-



```
{r}
# There are 13 variables and 4 factors
round(fa.varimax$loadings[ 1:13,], 3)
```

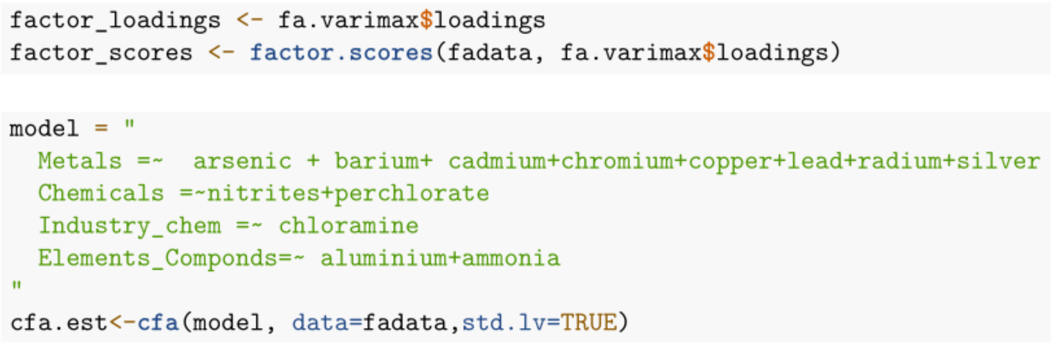
```
            Factor1 Factor2 Factor3 Factor4
aluminium     0.488   0.029   0.020   0.181
ammonia       0.138   0.005  -0.138   0.008
arsenic       0.459   0.664   0.058  -0.124
barium        0.599   0.122   0.051  -0.025
cadmium      -0.194   0.619   0.016  -0.095
chloramine    0.768   0.004   0.055   0.019
chromium      0.712  -0.023   0.035   0.023
copper        0.135  -0.052   0.097   0.557
lead         -0.056  -0.042  -0.039   0.239
nitrites      0.429   0.092   0.895   0.041
perchlorate   0.747  -0.012   0.030   0.007
radium        0.487  -0.026   0.078  -0.078
silver        0.694  -0.028   0.044  -0.027
```

Figure 5:Factor loading of Varimax rotation

- using varimax factor rotations we can explain factor one as the **metal factor**.Because heavy metals get the low factor loadings(cadmium,Led) compare to other factor loadings.
- we can explain the second factor as the **health factor**. Because values of arsenic, cadmium, and barium are high compared to other factor loadings. These elements and compounds are naturally occurring with health risks.
- In third-factor loadings, Nitrites got the highest value compared to others. Other all-factor loadings are lower than 0.1. nitrites are used in food preservatives. Most of the other variables are used in industries. Therefore, factor three can name as **Industry Factor**.
- Factor loadings of arsenic, barium, cadmium, radium, and silver are lower than other factor loadings. These lower factors have many negative environmental implications. Therefore, this factor can consider as **environmental factor**.

CONFIRMATORY FACTOR ANALYSIS

CFA and EFA are both methods of factor analysis. It is said that EFA extracts a factor structure from the data whereas CFA is used to test if a factor structure fits the data (or in other words to test a hypothesis). The cfa() function in lavaan can be used to estimate a factor model. To use the function, we need to first specify the factor model



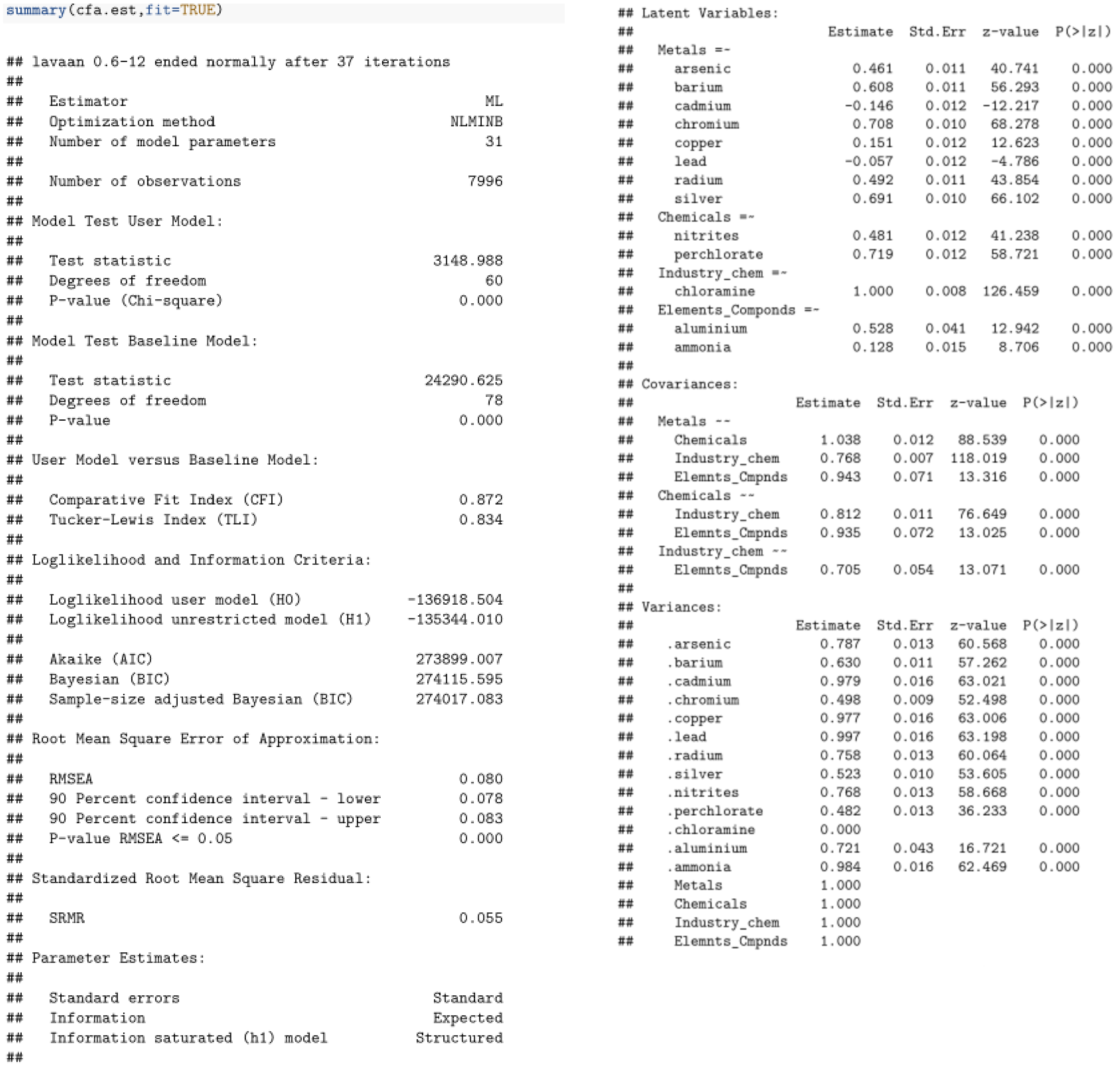
```
factor_loadings <- fa.varimax$loadings
factor_scores <- factor.scores(fadata, fa.varimax$loadings)
```

```
model = "
  Metals =~  arsenic + barium+ cadmium+chromium+copper+lead+radium+silver
  Chemicals =~nitrites+perchlorate
  Industry_chem =~ chloramine
  Elements_Componds=~ aluminium+ammonia
"
cfa.est<-cfa(model, data=fadata,std.lv=TRUE)
```

Figure 6:Confirmatory Factor Analysis

Using these criteria, we can evaluate whether the confirmatory factor model identified

- The chi-square statistic is 3148 with a degree of freedom of 60 and a p-value close to 0. Therefore, one would reject the hypothesis that the model fits the data simply based on it.
- Comparative Fit Index (CFI) is 0.872, which is smaller than the cut-off value of 0.95. It also suggests a bad fit.
- The RMSEA = 0.085, which lies in the range of a reasonable fit model.



```
summary(cfa.est,fit=TRUE)
```

```
## lavaan 0.6-12 ended normally after 37 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        31
##
##   Number of observations                          7996
##
## Model Test User Model:
##
##   Test statistic                              3148.988
##   Degrees of freedom                                60
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                             24290.625
##   Degrees of freedom                                78
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.872
##   Tucker-Lewis Index (TLI)                       0.834
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)            -136918.504
##   Loglikelihood unrestricted model (H1)    -135344.010
##
##   Akaike (AIC)                              273899.007
##   Bayesian (BIC)                            274115.595
##   Sample-size adjusted Bayesian (BIC)       274017.083
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.080
##   90 Percent confidence interval - lower         0.078
##   90 Percent confidence interval - upper         0.083
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.055
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
```

```
## Latent Variables:
##                        Estimate  Std.Err  z-value  P(>|z|)
##   Metals =~
##     arsenic               0.461    0.011   40.741    0.000
##     barium                0.608    0.011   56.293    0.000
##     cadmium              -0.146    0.012  -12.217    0.000
##     chromium              0.708    0.010   68.278    0.000
##     copper                0.151    0.012   12.623    0.000
##     lead                 -0.057    0.012   -4.786    0.000
##     radium                0.492    0.011   43.854    0.000
##     silver                0.691    0.010   66.102    0.000
##   Chemicals =~
##     nitrites              0.481    0.012   41.238    0.000
##     perchlorate           0.719    0.012   58.721    0.000
##   Industry_chem =~
##     chloramine            1.000    0.008  126.459    0.000
##   Elements_Componds =~
##     aluminium             0.528    0.041   12.942    0.000
##     ammonia               0.128    0.015    8.706    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Metals ~~
##     Chemicals         1.038    0.012   88.539    0.000
##     Industry_chem     0.768    0.007  118.019    0.000
##     Elemnts_Cmpnds    0.943    0.071   13.316    0.000
##   Chemicals ~~
##     Industry_chem     0.812    0.011   76.649    0.000
##     Elemnts_Cmpnds    0.935    0.072   13.025    0.000
##   Industry_chem ~~
##     Elemnts_Cmpnds    0.705    0.054   13.071    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .arsenic           0.787    0.013   60.568    0.000
##    .barium            0.630    0.011   57.262    0.000
##    .cadmium           0.979    0.016   63.021    0.000
##    .chromium          0.498    0.009   52.498    0.000
##    .copper            0.977    0.016   63.006    0.000
##    .lead              0.997    0.016   63.198    0.000
##    .radium            0.758    0.013   60.064    0.000
##    .silver            0.523    0.010   53.605    0.000
##    .nitrites          0.768    0.013   58.668    0.000
##    .perchlorate       0.482    0.013   36.233    0.000
##    .chloramine        0.000
##    .aluminium         0.721    0.043   16.721    0.000
##    .ammonia           0.984    0.016   62.469    0.000
##     Metals            1.000
##     Chemicals         1.000
##     Industry_chem     1.000
##     Elemnts_Cmpnds    1.000
```

Figure 7:Confirmatory factor analysis summary

4. Conclusion and recommendation

The null hypothesis is that a 4-factor model is sufficient. For this model, the chi-square statistic is 160.7 with degrees of freedom 32. The p-value for the chi-square test is 4.48-e19 which is lower than .05. Therefore, we reject the null hypothesis that the factor model needs more factors to fit the data. Therefore, need to try this model with a higher number of factor models. The results of this study could serve as a starting point for future studies into checking the water quality with its compounds and the use of Factor analysis for dimension reduction in other areas.

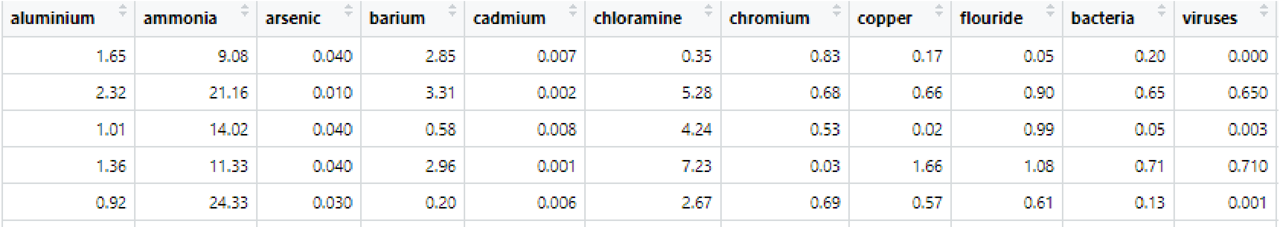
5. References

1. Zimmer, C. (2019). Learn to Perform Confirmatory Factor Analysis in Stata With Data From the General Social Survey (2016). In *SAGE Publications Ltd eBooks*. https://doi.org/10.4135/9781529700091
2. *Confirmatory factor analysis -- Advanced Statistics using R*. (n.d.). https://advstats.psychstat.org/book/factor/cfa.php
3. *Intro - Basic Exploratory Factor Analysis | QuantDev Methodology*. (n.d.). https://quantdev.ssri.psu.edu/tutorials/intro-basic-exploratory-factor-analysis
4. Kim, J., Ahtola, O., Spector, P. E., Kim, J., Mueller, C. W., & Wales, G. S. O. N. S. (1978). *Introduction to Factor Analysis: What It Is and How To Do It*. SAGE.
5. *RPubs - Exploratory Factor Analysis in R*. (n.d.). https://rpubs.com/pjmurphy/758265

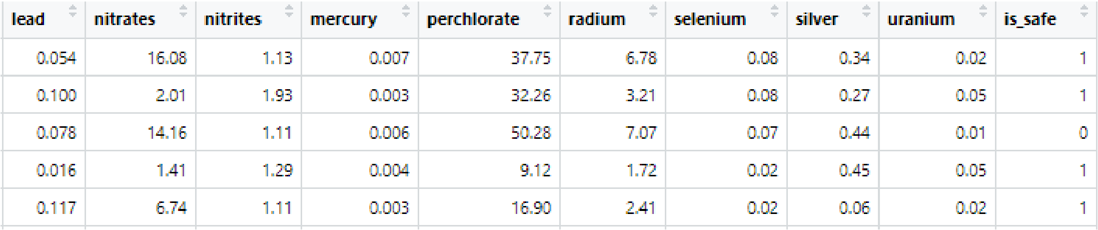
6. Appendices

- Part of the Dataset

Table 1:Water Quality Dataset



aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride	bacteria	viruses
1.65	9.08	0.040	2.85	0.007	0.35	0.83	0.17	0.05	0.20	0.000
2.32	21.16	0.010	3.31	0.002	5.28	0.68	0.66	0.90	0.65	0.650
1.01	14.02	0.040	0.58	0.008	4.24	0.53	0.02	0.99	0.05	0.003
1.36	11.33	0.040	2.96	0.001	7.23	0.03	1.66	1.08	0.71	0.710
0.92	24.33	0.030	0.20	0.006	2.67	0.69	0.57	0.61	0.13	0.001



lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	is_safe
0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1
0.100	2.01	1.93	0.003	32.26	3.21	0.08	0.27	0.05	1
0.078	14.16	1.11	0.006	50.28	7.07	0.07	0.44	0.01	0
0.016	1.41	1.29	0.004	9.12	1.72	0.02	0.45	0.05	1
0.117	6.74	1.11	0.003	16.90	2.41	0.02	0.06	0.02	1

- R codes, written in markdown.

Mini Project

S17403

2023-05-26

Factor Analysis

Explanatory Factor Analysis

Load dataset

```
#https://www.kaggle.com/datasets/mssmartypants/water-quality
library(readr)
waterQuality1 <- read_csv("waterQuality1.csv")
```

```
## Warning: One or more parsing issues, see ‘problems()‘ for details
```

```
## Rows: 7999 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (21): aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
#View(waterQuality1)
str(waterQuality1)
```

```
## spec_tbl_df [7,999 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ aluminium  : num [1:7999] 1.65 2.32 1.01 1.36 0.92 0.94 2.36 3.93 0.6 0.22 ...
##  $ ammonia    : num [1:7999] 9.08 21.16 14.02 11.33 24.33 ...
##  $ arsenic    : num [1:7999] 0.04 0.01 0.04 0.04 0.03 0.03 0.01 0.04 0.01 0.02 ...
##  $ barium     : num [1:7999] 2.85 3.31 0.58 2.96 0.2 2.88 1.35 0.66 0.71 1.37 ...
##  $ cadmium    : num [1:7999] 0.007 0.002 0.008 0.001 0.006 0.003 0.004 0.001 0.005 0.007 ...
##  $ chloramine : num [1:7999] 0.35 5.28 4.24 7.23 2.67 0.8 1.28 6.22 3.14 6.4 ...
##  $ chromium   : num [1:7999] 0.83 0.68 0.53 0.03 0.69 0.43 0.62 0.1 0.77 0.49 ...
##  $ copper     : num [1:7999] 0.17 0.66 0.02 1.66 0.57 1.38 1.88 1.86 1.45 0.82 ...
##  $ flouride   : num [1:7999] 0.05 0.9 0.99 1.08 0.61 0.11 0.33 0.86 0.98 1.24 ...
##  $ bacteria   : num [1:7999] 0.2 0.65 0.05 0.71 0.13 0.67 0.13 0.16 0.35 0.83 ...
##  $ viruses    : num [1:7999] 0 0.65 0.003 0.71 0.001 0.67 0.007 0.005 0.002 0.83 ...
##  $ lead       : num [1:7999] 0.054 0.1 0.078 0.016 0.117 0.135 0.021 0.197 0.167 0.109 ...
##  $ nitrates   : num [1:7999] 16.08 2.01 14.16 1.41 6.74 ...
##  $ nitrites   : num [1:7999] 1.13 1.93 1.11 1.29 1.11 1.89 1.78 1.81 1.84 1.46 ...
##  $ mercury    : num [1:7999] 0.007 0.003 0.006 0.004 0.003 0.006 0.007 0.001 0.004 0.01 ...
##  $ perchlorate: num [1:7999] 37.75 32.26 50.28 9.12 16.9 ...
##  $ radium     : num [1:7999] 6.78 3.21 7.07 1.72 2.41 5.42 2.84 7.24 4.99 0.08 ...
##  $ selenium   : num [1:7999] 0.08 0.08 0.07 0.02 0.02 0.08 0.1 0.08 0.08 0.03 ...
##  $ silver     : num [1:7999] 0.34 0.27 0.44 0.45 0.06 0.19 0.24 0.08 0.25 0.31 ...
##  $ uranium    : num [1:7999] 0.02 0.05 0.01 0.05 0.02 0.02 0.08 0.07 0.08 0.01 ...
##  $ is_safe    : num [1:7999] 1 1 0 1 1 1 0 0 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   aluminium = col_double(),
##   ..   ammonia = col_double(),
##   ..   arsenic = col_double(),
##   ..   barium = col_double(),
##   ..   cadmium = col_double(),
##   ..   chloramine = col_double(),
##   ..   chromium = col_double(),
##   ..   copper = col_double(),
##   ..   flouride = col_double(),
##   ..   bacteria = col_double(),
##   ..   viruses = col_double(),
##   ..   lead = col_double(),
##   ..   nitrates = col_double(),
##   ..   nitrites = col_double(),
##   ..   mercury = col_double(),
##   ..   perchlorate = col_double(),
##   ..   radium = col_double(),
##   ..   selenium = col_double(),
##   ..   silver = col_double(),
##   ..   uranium = col_double(),
##   ..   is_safe = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
fadata<-waterQuality1[,-21]

#remove null values
fadata <- na.omit(fadata) # Remove NA
colSums(is.na(fadata))
```

```
##   aluminium     ammonia     arsenic      barium     cadmium  chloramine
##           0           0           0           0           0           0
##    chromium      copper    flouride    bacteria     viruses        lead
##           0           0           0           0           0           0
##    nitrates    nitrites     mercury perchlorate      radium    selenium
##           0           0           0           0           0           0
##      silver     uranium
##           0           0
```

In factor analysis, the Kaiser-Meyer-Olkin (KMO) measure is used to assess the sampling adequacy for factor analysis.The KMO values range between 0 and 1. A higher KMO value (close to 1) indicates a better suitability of the dataset for factor analysis. Generally, a KMO value above 0.7 is considered acceptable. Additionally, we can examine the KMO values per variable to identify variables with low individual KMO values. Variables with KMO values below 0.5 may indicate poor sampling adequacy and may need to be excluded from the factor analysis.

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.2.3
```

```
df_corr <- cor(fadata) # Create a correlation matrix
KMO(df_corr) # Kaiser-Meyer-Olkin factor adequacy
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = df_corr)
## Overall MSA =  0.8
## MSA for each item = 
##   aluminium     ammonia     arsenic      barium     cadmium  chloramine 
##        0.86        0.70        0.75        0.93        0.50        0.90 
##    chromium      copper    flouride    bacteria     viruses        lead 
##        0.91        0.65        0.42        0.48        0.43        0.58 
##    nitrates    nitrites     mercury perchlorate      radium    selenium 
##        0.53        0.77        0.42        0.91        0.93        0.42 
##      silver     uranium 
##        0.92        0.64
```

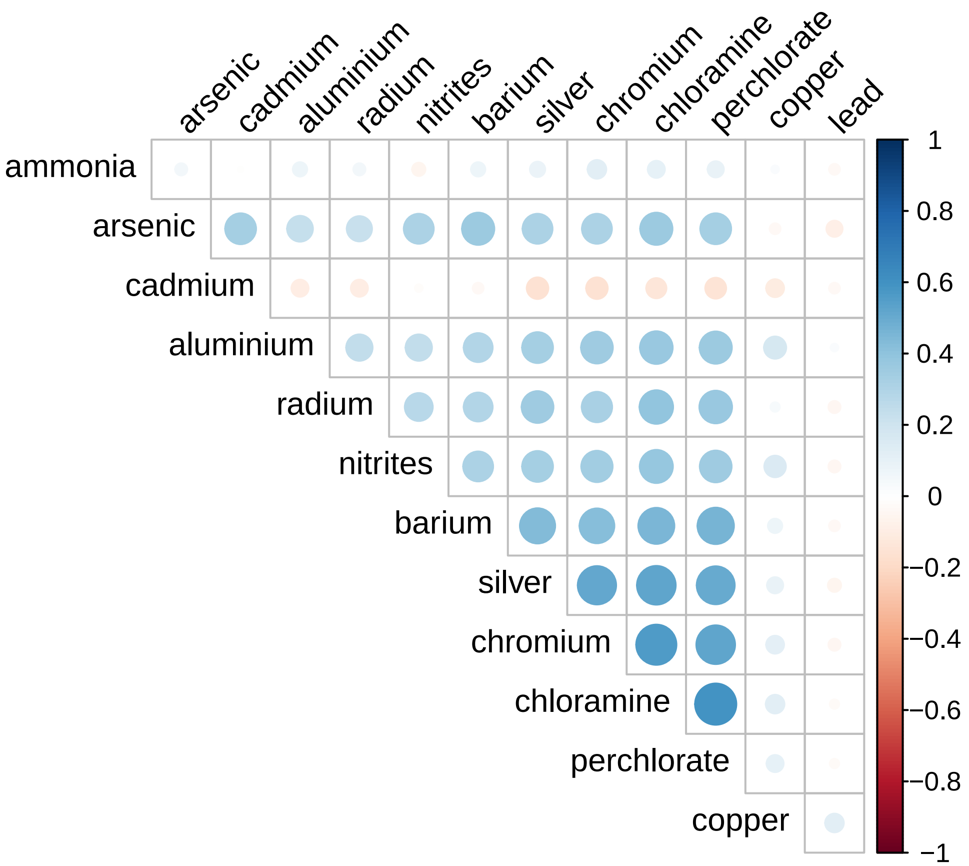
```
fa.var<-c('aluminium','ammonia' ,'arsenic','barium','cadmium','chloramine','chromium','copper','lead','
fadata<-fadata[,fa.var]
```

Correlation Matrix for the data

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
df1_corr <- cor(fadata)
corrplot(round(df1_corr, 2),
         type="upper", order="hclust",
         tl.col="black", tl.srt=45, #Text label color and rotation
         diag=FALSE) # hide correlation coefficient on the principal diagonal
```



```
round(df1_corr, 2)
```

```
##               aluminium ammonia arsenic barium cadmium chloramine chromium copper
## aluminium          1.00    0.07    0.23   0.29   -0.10       0.37     0.35   0.17
## ammonia            0.07    1.00    0.05   0.07   -0.01       0.10     0.12   0.02
## arsenic            0.23    0.05    1.00   0.36    0.33       0.36     0.31  -0.04
## barium             0.29    0.07    0.36   1.00   -0.04       0.45     0.42   0.07
## cadmium           -0.10   -0.01    0.33  -0.04    1.00      -0.14    -0.16  -0.11
## chloramine         0.37    0.10    0.36   0.45   -0.14       1.00     0.56   0.12
## chromium           0.35    0.12    0.31   0.42   -0.16       0.56     1.00   0.11
## copper             0.17    0.02   -0.04   0.07   -0.11       0.12     0.11   1.00
## lead               0.02   -0.04   -0.09  -0.04   -0.04      -0.03    -0.05   0.12
## nitrites           0.24   -0.06    0.31   0.31   -0.02       0.38     0.34   0.16
## perchlorate        0.36    0.09    0.33   0.46   -0.15       0.59     0.52   0.10
## radium             0.24    0.05    0.22   0.29   -0.10       0.39     0.32   0.03
## silver             0.33    0.08    0.31   0.43   -0.16       0.52     0.51   0.09
##                lead nitrites perchlorate radium silver
## aluminium      0.02     0.24        0.36   0.24   0.33
## ammonia       -0.04    -0.06        0.09   0.05   0.08
## arsenic       -0.09     0.31        0.33   0.22   0.31
## barium        -0.04     0.31        0.46   0.29   0.43
## cadmium       -0.04    -0.02       -0.15  -0.10  -0.16
## chloramine    -0.03     0.38        0.59   0.39   0.52
## chromium      -0.05     0.34        0.52   0.32   0.51
## copper         0.12     0.16        0.10   0.03   0.09
## lead         1.00    -0.05       -0.03  -0.05  -0.06
## nitrites    -0.05     1.00        0.35   0.27   0.33
## perchlorate -0.03     0.35        1.00   0.37   0.50
## radium      -0.05     0.27        0.37   1.00   0.35
## silver      -0.06     0.33        0.50   0.35   1.00
```

Standardizing each variable

```
fadata <- apply(fadata, 2, scale)
head(fadata)
```

```
##      aluminium     ammonia    arsenic     barium    cadmium chloramine
## [1,] 0.7773543 -0.58545472 -0.4808447  1.0541387 -0.9931786 -0.7118970
## [2,] 1.3068633  0.77506955 -0.5995943  1.4323575 -1.1318775  1.2084760
## [3,] 0.2715546 -0.02908139 -0.4808447 -0.8122887 -0.9654388  0.8033669
## [4,] 0.5481638 -0.33204581 -0.4808447  1.1445823 -1.1596173  1.9680556
## [5,] 0.2004265  1.13209454 -0.5204279 -1.1247303 -1.0209184  0.1918079
## [6,] 0.2162328  0.02160039 -0.5204279  1.0788051 -1.1041378 -0.5366094
##        chromium     copper         lead    nitrites perchlorate     radium
## [1,]  2.1528583 -0.9729890 -0.781021271 -0.34860691  1.20328692  1.6617381
## [2,]  1.5986646 -0.2232889  0.009784669  1.04689407  0.89292150  0.1248033
## [3,]  1.0444708 -1.2024891 -0.368426868 -0.38349444  1.91164368  1.7865871
## [4,] -0.8028416  1.3067112 -1.434295744 -0.06950672 -0.41524893 -0.5166625
## [5,]  1.6356108 -0.3609890  0.302039038 -0.38349444  0.02457674 -0.2196079
## [6,]  0.6750083  0.8783112  0.611484841  0.97711902  0.60516923  1.0762391
##          silver
## [1,]  1.3386525
## [2,]  0.8510810
## [3,]  2.0351832
## [4,]  2.1048363
## [5,] -0.6116334
## [6,]  0.2938565
```

General methods used in determining the number of factors

- Cumulative proportion of at least 0.80
- Eigen Values of at least one
- Based on Scree Plot

Calculate Eigen Values

```
#Evaluate the correlation matrix
fa.cor<-cor(fadata)
# get eigenvalues
ev <- eigen(round(fa.cor,3))
ev$values
```

```
##  [1] 4.0769234 1.4082442 1.1068062 1.0194023 0.8874730 0.7447430 0.7246353
##  [8] 0.6255079 0.5912847 0.5089649 0.4645246 0.4406545 0.4008359
```

```
round(ev$values,5)
```

```
##  [1] 4.07692 1.40824 1.10681 1.01940 0.88747 0.74474 0.72464 0.62551 0.59128
## [10] 0.50896 0.46452 0.44065 0.40084
```

```
sum(ev$values)
```

```
## [1] 13
```

```
cumsum(ev$values)
```

```
##  [1]  4.076923  5.485168  6.591974  7.611376  8.498849  9.243592  9.968227
##  [8] 10.593735 11.185020 11.693985 12.158510 12.599164 13.000000
```

```
cumsum(ev$values)/13
```

```
##  [1] 0.3136095 0.4219360 0.5070749 0.5854905 0.6537576 0.7110455 0.7667867
##  [8] 0.8149027 0.8603862 0.8995373 0.9352700 0.9691665 1.0000000
```

```
ev$values
```

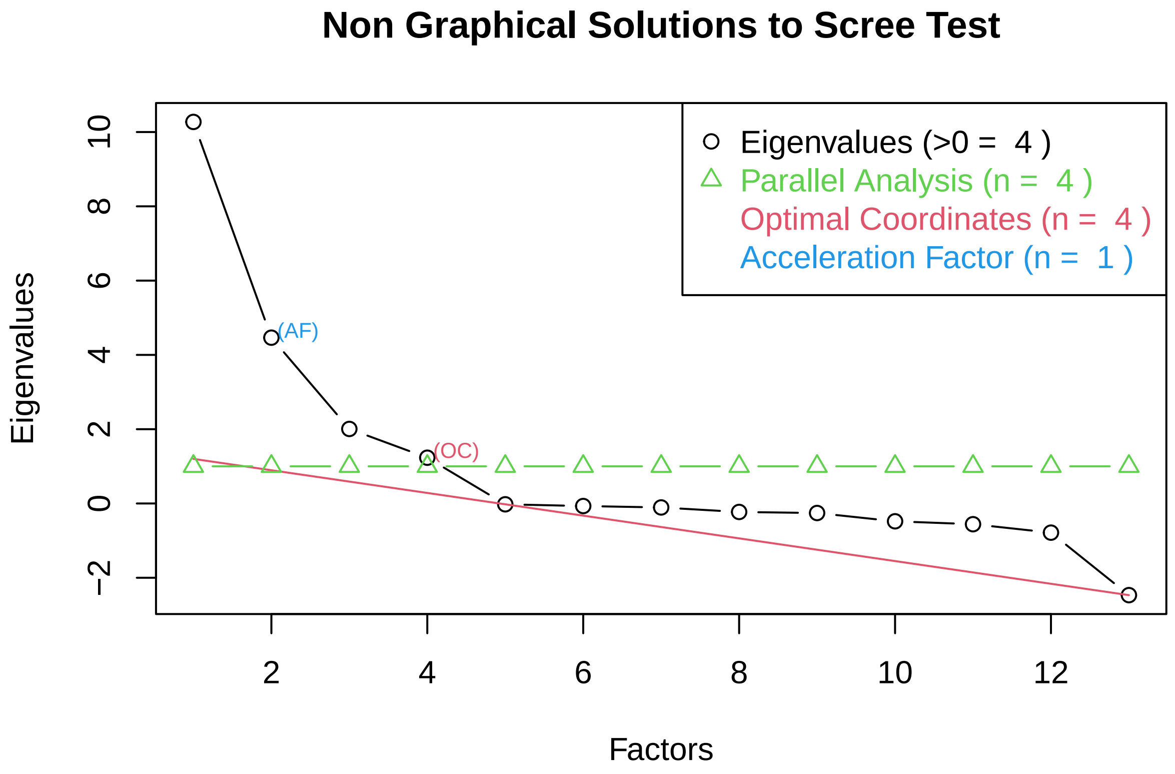
```
##  [1] 4.0769234 1.4082442 1.1068062 1.0194023 0.8874730 0.7447430 0.7246353
##  [8] 0.6255079 0.5912847 0.5089649 0.4645246 0.4406545 0.4008359
```

Scree Plot

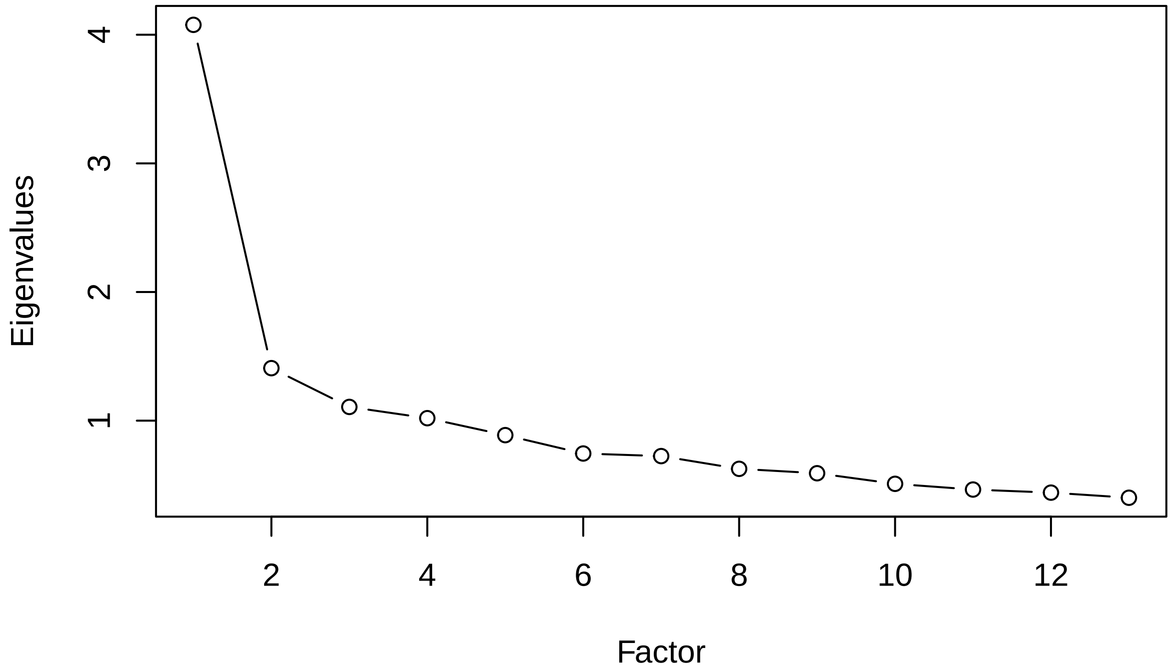
```
library(nFactors)
```

```
## Warning: package 'nFactors' was built under R version 4.2.3

## Loading required package: lattice

##
## Attaching package: 'nFactors'

## The following object is masked from 'package:lattice':
##
##     parallel
```

```
plot(nScree(x=fa.cor,model="factors"))
```



```r
plot(ev$values, type='b', ylab='Eigenvalues', xlab='Factor')
```



Factor Analysis with No Rotation

```
fa.res<-factanal(x=fadata, factors=4, rotation='none')
fa.res
```

```
##
## Call:
## factanal(x = fadata, factors = 4, rotation = "none")
##
## Uniquenesses:
##   aluminium     ammonia     arsenic      barium     cadmium  chloramine
##       0.728       0.962       0.329       0.623       0.570       0.406
##    chromium      copper        lead    nitrites perchlorate      radium
##       0.491       0.659       0.936       0.005       0.441       0.750
##      silver
##       0.515
##
## Loadings:
##             Factor1 Factor2 Factor3 Factor4
## aluminium    0.420   0.244           0.166
## ammonia      0.184
## arsenic      0.458   0.313   0.601
## barium       0.522   0.321
## cadmium     -0.107           0.641
## chloramine   0.655   0.391  -0.109
## chromium     0.611   0.346  -0.128
## copper               0.163  -0.227   0.510
## lead                                 0.221
## nitrites             0.997
## perchlorate  0.646   0.357  -0.118
## radium       0.396   0.279          -0.102
## silver       0.592   0.343  -0.116
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      2.435   1.884   0.904   0.361
## Proportion Var   0.187   0.145   0.070   0.028
## Cumulative Var   0.187   0.332   0.402   0.430
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 160.7 on 32 degrees of freedom.
## The p-value is 4.48e-19
```

```r
# There are 13 variables and 4 factors
round(fa.res$loadings[ 1:13,], 3)
```

```
##             Factor1 Factor2 Factor3 Factor4
## aluminium     0.420   0.244  -0.092   0.166
## ammonia       0.184  -0.061  -0.014   0.013
## arsenic       0.458   0.313   0.601   0.049
## barium        0.522   0.321   0.039  -0.011
## cadmium      -0.107  -0.017   0.641   0.087
## chloramine    0.655   0.391  -0.109  -0.007
## chromium      0.611   0.346  -0.128  -0.008
## copper        0.052   0.163  -0.227   0.510
## lead         -0.044  -0.053  -0.099   0.221
## nitrites     -0.015   0.997   0.000  -0.001
## perchlorate   0.646   0.357  -0.118  -0.021
## radium        0.396   0.279  -0.072  -0.102
## silver        0.592   0.343  -0.116  -0.058
```

Communalities

```r
#communality

#fa.res$uniquenesses

apply(fa.res$loadings^2,1,sum) # communality
```

```
##   aluminium     ammonia     arsenic      barium     cadmium  chloramine
##  0.27219086  0.03794645  0.67086494  0.37703347  0.43034224  0.59390013
##    chromium      copper        lead    nitrites perchlorate      radium
##  0.50947341  0.34089639  0.06356752  0.99500002  0.55865640  0.25031351
##      silver
##  0.48459249
```

```r
sum(apply(fa.res$loadings^2,1,sum))/13
```

```
## [1] 0.4295983
```

Residual Matrix

```r
#residuals
Lambda <- fa.res$loadings
Psi <- diag(fa.res$uniquenesses)
S <- fa.res$correlation
Sigma <- Lambda %*% t(Lambda) + Psi


# residual matrix
round(S - Sigma, 5)
```

```
##             aluminium  ammonia  arsenic   barium  cadmium chloramine chromium
## aluminium     0.00000  0.00134  0.00380  0.00166 -0.00612   -0.01044  0.00143
## ammonia       0.00134  0.00001 -0.01066 -0.00571  0.02017    0.00645  0.03161
## arsenic       0.00380 -0.01066  0.00000  0.00066 -0.00013    0.00031  0.00167
## barium        0.00166 -0.00571  0.00066  0.00000 -0.00044   -0.01629 -0.00910
## cadmium      -0.00612  0.02017 -0.00013 -0.00044  0.00001    0.00336 -0.00347
## chloramine   -0.01044  0.00645  0.00031 -0.01629  0.00336    0.00000  0.00631
## chromium      0.00143  0.03161  0.00167 -0.00910 -0.00347    0.00631  0.00000
## copper        0.00148  0.00626 -0.00034 -0.00016  0.00042   -0.00010 -0.00013
## lead          0.00694 -0.03649 -0.00238  0.00368  0.00297    0.01053 -0.01542
## nitrites     -0.00001 -0.00001 -0.00001  0.00000  0.00002    0.00001  0.00004
## perchlorate  -0.00297 -0.00765 -0.00349  0.01499  0.00331    0.01314 -0.00884
## radium        0.01897 -0.00514 -0.00204 -0.00768  0.00335    0.01199 -0.03312
## silver        0.00130 -0.01348  0.00240  0.01660 -0.00638   -0.01242  0.01514
##               copper     lead nitrites perchlorate   radium   silver
## aluminium    0.00148  0.00694   -1e-05    -0.00297  0.01897  0.00130
## ammonia      0.00626 -0.03649   -1e-05    -0.00765 -0.00514 -0.01348
## arsenic     -0.00034 -0.00238   -1e-05    -0.00349 -0.00204  0.00240
## barium      -0.00016  0.00368    0e+00     0.01499 -0.00768  0.01660
## cadmium      0.00042  0.00297    2e-05     0.00331  0.00335 -0.00638
## chloramine  -0.00010  0.01053    1e-05     0.01314  0.01199 -0.01242
## chromium    -0.00013 -0.01542    4e-05    -0.00884 -0.03312  0.01514
## copper       0.00000 -0.00248    1e-05    -0.00347 -0.00426  0.00538
## lead        -0.00248  0.00000   -3e-05     0.01334 -0.00071 -0.01136
## nitrites     0.00001 -0.00003    0e+00    -0.00001  0.00000 -0.00002
## perchlorate -0.00347  0.01334   -1e-05     0.00000  0.00730 -0.01486
## radium      -0.00426 -0.00071    0e+00     0.00730  0.00000  0.01000
## silver       0.00538 -0.01136   -2e-05    -0.01486  0.01000  0.00000
```

Numbers close to 0 indicate that our factor model is a good representation of the underlying concept.

Factor Rotations Factor rotations in factor analysis are used to achieve a more interpretable and meaningful solution. The primary goal of factor rotation is to simplify and clarify the factor structure by creating more distinct and easily interpretable factors.

```r
fa.varimax <- factanal(fadata, factors = 4, rotation = "varimax")
fa.varimax
```

```
##
## Call:
## factanal(x = fadata, factors = 4, rotation = "varimax")
```

```
##
## Uniquenesses:
##   aluminium     ammonia     arsenic      barium     cadmium  chloramine
##       0.728       0.962       0.329       0.623       0.570       0.406
##    chromium      copper        lead    nitrites perchlorate      radium
##       0.491       0.659       0.936       0.005       0.441       0.750
##      silver
##       0.515
##
## Loadings:
##             Factor1 Factor2 Factor3 Factor4
## aluminium    0.488                   0.181
## ammonia      0.138          -0.138
## arsenic      0.459   0.664          -0.124
## barium       0.599   0.122
## cadmium     -0.194   0.619
## chloramine   0.768
## chromium     0.712
## copper       0.135                   0.557
## lead                                 0.239
## nitrites     0.429           0.895
## perchlorate  0.747
## radium       0.487
## silver       0.694
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      3.443   0.856   0.851   0.435
## Proportion Var   0.265   0.066   0.065   0.033
## Cumulative Var   0.265   0.331   0.396   0.430
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 160.7 on 32 degrees of freedom.
## The p-value is 4.48e-19
```

```
# There are 13 variables and 4 factors
round(fa.varimax$loadings[ 1:13,], 3)
```

```
##             Factor1 Factor2 Factor3 Factor4
## aluminium     0.488   0.029   0.020   0.181
## ammonia       0.138   0.005  -0.138   0.008
## arsenic       0.459   0.664   0.058  -0.124
## barium        0.599   0.122   0.051  -0.025
## cadmium      -0.194   0.619   0.016  -0.095
## chloramine    0.768   0.004   0.055   0.019
## chromium      0.712  -0.023   0.035   0.023
## copper        0.135  -0.052   0.097   0.557
## lead         -0.056  -0.042  -0.039   0.239
## nitrites      0.429   0.092   0.895   0.041
## perchlorate   0.747  -0.012   0.030   0.007
## radium        0.487  -0.026   0.078  -0.078
## silver        0.694  -0.028   0.044  -0.027
```

- using varimax factor rotations we can explain factor one as the **metal factor**.Because heavy metals get the low factor loadings(cadmium,Led) compare to other factor loadings.

- we can explain second factor as the **health factor**.Because values of arsenic,cadmium and bariums are high compared to other factor loadings.These elements and compounds are naturally occuring with health risks.

- In third factor loadings ,Nitrites got the highest value compare to others.Other all factor loading are lower than 0.1.nitrites are used in food preservatives.Most of the other variables are used in industries.Therefore factor three can name as **Industry Factor**.

- Factor loadings of arsenic,barium,cadmium, radium and silver are lower than other factor loadings.These lower factors have many negative environmental implications.Therefore this factor can consider as **environmental factor**.

  Estimation of Factor Scores

```
factor_scores <- factor.scores(fadata, fa.varimax$loadings)
#factor_scores
```

Confirmatory Factor Analysis

difference between EFA and CFA?

CFA and EFA are both methods of factor analysis. It is said that EFA extracts a factor structure from the data whereas CFA is used to test if a factor structure fits the data (or in other words to test a hypothesis)

```
library(lavaan)
```

```
## This is lavaan 0.6-12
## lavaan is FREE software! Please report any bugs.
```

```
##
## Attaching package: 'lavaan'
```

```
## The following object is masked from 'package:psych':
##
##     cor2cov
```

```
factor_loadings <- fa.varimax$loadings
factor_scores <- factor.scores(fadata, fa.varimax$loadings)
```

```
model = "
  Metals =~  arsenic + barium+ cadmium+chromium+copper+lead+radium+silver
  Chemicals =~nitrites+perchlorate
  Industry_chem =~ chloramine
  Elements_Componds=~ aluminium+ammonia
"
cfa.est<-cfa(model, data=fadata,std.lv=TRUE)
```

```
## Warning in lav_object_post_check(object): lavaan WARNING: covariance matrix of latent variables
##                   is not positive definite;
##                   use lavInspect(fit, "cov.lv") to investigate.
```

```r
summary(cfa.est,fit=TRUE)
```

```
## lavaan 0.6-12 ended normally after 37 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        31
##
##   Number of observations                          7996
##
## Model Test User Model:
##
##   Test statistic                              3148.988
##   Degrees of freedom                                60
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                             24290.625
##   Degrees of freedom                                78
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.872
##   Tucker-Lewis Index (TLI)                       0.834
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)            -136918.504
##   Loglikelihood unrestricted model (H1)    -135344.010
##
##   Akaike (AIC)                              273899.007
##   Bayesian (BIC)                            274115.595
##   Sample-size adjusted Bayesian (BIC)       274017.083
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.080
##   90 Percent confidence interval - lower         0.078
##   90 Percent confidence interval - upper         0.083
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.055
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
```

```
## Latent Variables:
##                        Estimate  Std.Err  z-value  P(>|z|)
##   Metals =~
##     arsenic              0.461    0.011   40.741    0.000
##     barium               0.608    0.011   56.293    0.000
##     cadmium             -0.146    0.012  -12.217    0.000
##     chromium             0.708    0.010   68.278    0.000
##     copper               0.151    0.012   12.623    0.000
##     lead                -0.057    0.012   -4.786    0.000
##     radium               0.492    0.011   43.854    0.000
##     silver               0.691    0.010   66.102    0.000
##   Chemicals =~
##     nitrites             0.481    0.012   41.238    0.000
##     perchlorate          0.719    0.012   58.721    0.000
##   Industry_chem =~
##     chloramine           1.000    0.008  126.459    0.000
##   Elements_Componds =~
##     aluminium            0.528    0.041   12.942    0.000
##     ammonia              0.128    0.015    8.706    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Metals ~~
##     Chemicals          1.038    0.012   88.539    0.000
##     Industry_chem      0.768    0.007  118.019    0.000
##     Elemnts_Cmpnds    0.943    0.071   13.316    0.000
##   Chemicals ~~
##     Industry_chem      0.812    0.011   76.649    0.000
##     Elemnts_Cmpnds    0.935    0.072   13.025    0.000
##   Industry_chem ~~
##     Elemnts_Cmpnds    0.705    0.054   13.071    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .arsenic            0.787    0.013   60.568    0.000
##    .barium             0.630    0.011   57.262    0.000
##    .cadmium            0.979    0.016   63.021    0.000
##    .chromium           0.498    0.009   52.498    0.000
##    .copper             0.977    0.016   63.006    0.000
##    .lead               0.997    0.016   63.198    0.000
##    .radium             0.758    0.013   60.064    0.000
##    .silver             0.523    0.010   53.605    0.000
##    .nitrites           0.768    0.013   58.668    0.000
##    .perchlorate        0.482    0.013   36.233    0.000
##    .chloramine         0.000
##    .aluminium          0.721    0.043   16.721    0.000
##    .ammonia            0.984    0.016   62.469    0.000
##     Metals             1.000
##     Chemicals          1.000
##     Industry_chem      1.000
##     Elemnts_Cmpnds    1.000
```

Using these criteria, we can evaluate whether the confirmatory factor model identified

- The chi-square statistic is 3148 with the degrees of freedom 60 and a p-value close to 0. Therefore, one would reject the hypothesis that the model fits the data simply based on it.

- Comparative Fit Index (CFI) is 0.872, which is smaller than the cut-off value 0.95. It also suggests a bad fit.
- The RMSEA = 0.085, which lies the range of a reasonable fit model.