Model solution

Advance Statistics Group Assignment

1) Problem - Cereal Data Factor Analysis (20 points)

The data file labeled Cereal has the following information

As part of a study of consumer consideration of ready-to-eat cereals sponsored by Kellogg Australia, Roberts and Lattin (1991) surveyed consumers regarding their perceptions of their favorite brands of cereals. Each respondent was asked to evaluate three preferred brands on each of 25 different attributes. Respondents used a five point likert scale to indicate the extent to which each brand possessed the given attribute.

For the purpose of this assignment, a subset of the data collected by Roberts and Lattin, reflecting the e
12 most frequently cited cereal brands in the sample (in the original study, a total of 40 different brands
by 121 respondents, but the majority of brands were rated by only a small number of consumers). The 2
12 brands are listed below:

Cereal Brand	Attributes 1-12	Attributes 13-25
All Bran	Filling	Family
Cerola Muesli	Natural	Calories
Just Right	Fibre	Plain
Kellogg's corn falkes	Sweet	Crisp
Komplete	Easy	Regular
Nutrigrain	Salt	Sugar
Purina Muesli	Satisfying	Fruit
Rice Bubbles	Energy	Process
Special K	Fun	Quality
Sustain	Kids	Treat
Vitabrit	Soggy	Boring
Weetbix	Economical	Nutritious
Health		

In total 116 respondents provided 235 observations of the 12 selected brands.

1) **How do you characterize the consideration behavior of the 12 selected brands?**
2) **Analyze and interpret your results using factor analysis.**

Advance Statistics Group Assignment

Solution

Steps in Factor Analysis

A) Test assumptions of Factor Analysis such as Factorability

B) Select the type of Analysis - Extraction & Rotation

C) Determine the number of Factors

D) Identify which item belong in which factor

E) Drop items as necessary and repeat steps c and d

F) Name and define factors.

G) Examine correlations among factors

H) Analyze internal reliability

Ref: Revelle, W.(2011),psych:Procedure for Personality and Psychological Research. *http://personality-project.org/r/psych.manual.pdf*)

A) Test assumptions of Factor Analysis such as Factorability

Re: *https://en.wikiversity.org/wiki/Exploratory_factor_analysis/Assumptions*

We choose to analyze this data using principal component analysis in R software.

Load required packages

```
#Load required libraries

#specify the packages of interest
packages <- c("psych","corpcor","GPArotation","ggplot2","MASS","MVN","psy")

#use this function to check if each package is on the local machine
#if a package is installed, it will be loaded
#if any are not, the missing package(s) will be installed and loaded

package.check <- lapply(packages, FUN = function(x) {
    if (!require(x, character.only = TRUE)) {
        install.packages(x, dependencies = TRUE,repos='http://cran.us.r-proje
ct.org')
        library(x, character.only = TRUE)
    }
})
```

```
## Loading required package: psych

## Loading required package: corpcor

## Loading required package: GPArotation

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.3

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

## Loading required package: MASS

## Loading required package: MVN

## Warning: package 'MVN' was built under R version 3.4.3

## sROC 0.1-2 loaded

##
## Attaching package: 'MVN'

## The following object is masked from 'package:psych':
##
##     mardia

## Loading required package: psy

##
## Attaching package: 'psy'

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

Load data and get basic descriptive statistics

```
cereals_data <- read.csv("D:/GL/AS/Cereals/data/cereal.csv", header=T)

str(cereals_data)

## 'data.frame':    235 obs. of  26 variables:
##  $ Cereals   : Factor w/ 12 levels "AllBran","CMuesli",..: 12 9 9 2 3 8 9
9 8 3 ...
##  $ Filling   : int  5 1 5 5 4 4 4 4 4 4 ...
```

```
##  $ Natural   : int  5 2 4 5 5 4 4 3 3 3 ...
##  $ Fibre     : int  5 2 5 5 3 4 3 3 3 3 ...
##  $ Sweet     : int  1 1 5 3 2 2 2 2 2 2 ...
##  $ Easy      : int  2 5 5 5 5 5 5 5 5 5 ...
##  $ Salt      : int  1 2 3 2 2 2 1 1 1 1 ...
##  $ Satisfying: int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Energy    : int  4 1 5 5 4 4 5 4 4 4 ...
##  $ Fun       : int  1 1 5 5 5 5 5 4 4 4 ...
##  $ Kids      : int  4 5 5 5 5 5 5 5 5 5 ...
##  $ Soggy     : int  5 3 3 3 1 1 1 1 1 1 ...
##  $ Economical: int  5 5 3 3 5 5 5 3 3 3 ...
##  $ Health    : int  5 2 5 5 5 4 5 4 4 4 ...
##  $ Family    : int  5 5 5 5 3 5 5 5 5 5 ...
##  $ Calories  : int  1 1 1 1 3 3 3 2 2 2 ...
##  $ Plain     : int  3 5 1 1 1 1 1 3 3 3 ...
##  $ Crisp     : int  1 5 5 1 5 5 5 4 4 4 ...
##  $ Regular   : int  4 1 4 4 3 3 3 4 4 4 ...
##  $ Sugar     : int  1 2 3 2 1 2 2 1 1 1 ...
##  $ Fruit     : int  1 1 1 5 1 1 1 1 1 1 ...
##  $ Process   : int  3 5 2 2 3 3 3 2 2 2 ...
##  $ Quality   : int  5 2 5 5 5 5 5 4 4 4 ...
##  $ Treat     : int  1 1 4 5 5 5 5 2 2 2 ...
##  $ Boring    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Nutritious: int  5 3 5 5 4 4 4 3 3 3 ...
```

```
describe(cereals_data)
```

```
##            vars   n mean   sd median trimmed  mad min max range  skew
## Cereals*      1 235 6.89 3.47      7    6.95 4.45   1  12    11 -0.08
## Filling       2 235 3.88 0.88      4    3.95 1.48   1   5     4 -0.54
## Natural       3 235 3.78 0.89      4    3.84 1.48   1   5     4 -0.63
## Fibre         4 235 3.53 1.00      4    3.58 1.48   1   5     4 -0.53
## Sweet         5 235 2.51 1.12      2    2.45 1.48   1   5     4  0.38
## Easy          6 235 4.53 0.77      5    4.69 0.00   1   6     5 -1.78
## Salt          7 235 1.99 0.83      2    1.94 1.48   1   4     3  0.41
## Satisfying    8 235 4.00 0.81      4    4.04 1.48   2   6     4 -0.29
## Energy        9 235 3.64 0.90      4    3.69 1.48   1   5     4 -0.41
## Fun          10 235 2.62 1.26      2    2.52 1.48   1   5     4  0.48
## Kids         11 235 3.84 1.20      4    3.99 1.48   1   6     5 -0.77
## Soggy        12 235 2.26 1.20      2    2.11 1.48   1   5     4  0.76
## Economical   13 235 3.22 1.12      3    3.25 1.48   1   5     4 -0.16
## Health       14 235 3.81 0.86      4    3.86 1.48   1   5     4 -0.50
## Family       15 235 3.88 1.11      4    3.99 1.48   1   6     5 -0.65
## Calories     16 235 2.70 0.99      3    2.71 1.48   1   5     4 -0.04
## Plain        17 235 2.27 1.09      2    2.18 1.48   1   5     4  0.44
## Crisp        18 235 3.20 1.21      3    3.24 1.48   1   6     5 -0.12
## Regular      19 235 3.07 1.15      3    3.09 1.48   1   5     4 -0.11
```

```
## Sugar         20 235 2.14 1.04      2    2.03 1.48   1   5     4  0.65
## Fruit         21 235 1.69 1.08      1    1.49 0.00   1   5     4  1.19
## Process       22 235 2.94 1.14      3    2.92 1.48   1   6     5  0.19
## Quality       23 235 3.69 0.91      4    3.74 1.48   1   5     4 -0.41
## Treat         24 235 2.63 1.26      3    2.53 1.48   1   6     5  0.38
## Boring        25 235 1.83 0.95      2    1.71 1.48   1   5     4  0.92
## Nutritious    26 235 3.66 0.89      4    3.72 1.48   1   5     4 -0.61
##             kurtosis   se
## Cereals*       -1.23 0.23
## Filling        -0.18 0.06
## Natural         0.42 0.06
## Fibre          -0.18 0.07
## Sweet          -0.68 0.07
## Easy            3.27 0.05
## Salt           -0.59 0.05
## Satisfying     -0.56 0.05
## Energy         -0.11 0.06
## Fun            -0.71 0.08
## Kids           -0.31 0.08
## Soggy          -0.26 0.08
## Economical     -0.59 0.07
## Health          0.06 0.06
## Family         -0.48 0.07
## Calories       -0.41 0.06
## Plain          -0.67 0.07
## Crisp          -0.86 0.08
## Regular        -0.84 0.08
## Sugar          -0.31 0.07
## Fruit          -0.03 0.07
## Process        -0.48 0.07
## Quality        -0.01 0.06
## Treat          -0.74 0.08
## Boring          0.08 0.06
## Nutritious      0.43 0.06
```

```
headTail(cereals_data)
```

```
##      Cereals Filling Natural Fibre Sweet Easy Salt Satisfying Energy Fun
## 1   Weetabix       5       5     5     1    2    1           5      4   1
## 2   SpecialK       1       2     2     1    5    2           5      1   1
## 3   SpecialK       5       4     5     5    5    3           5      5   5
## 4    CMuesli       5       5     5     3    5    2           5      5   5
## ...     <NA>     ...     ...   ...   ...  ...  ...         ...    ... ...
## 232  PMuesli       5       4     4     3    4    3           4      4   4
## 233 Weetabix       4       4     4     1    4    1           4      4   3
## 234 SpecialK       3       3     3     3    4    2           3      3   2
## 235 Weetabix       4       4     4     1    4    1           4      3   2
```

```
##     Kids Soggy Economical Health Family Calories Plain Crisp Regular Sugar
## 1      4     5          5      5      5        1     3     1       4     1
## 2      5     3          5      2      5        1     5     5       1     2
## 3      5     3          3      5      5        1     1     5       4     3
## 4      5     3          3      5      5        1     1     1       4     2
## ...  ...   ...        ...    ...    ...      ...   ...   ...     ...   ...
## 232    4     1          3      4      4        4     1     4       4     3
## 233    4     2          4      4      3        3     3     3       4     1
## 234    3     2          3      4      3        2     3     2       3     2
## 235    4     3          4      4      4        2     2     2       4     1
##     Fruit Process Quality Treat Boring Nutritious
## 1       1       3       5     1      1          5
## 2       1       5       2     1      1          3
## 3       1       2       5     4      1          5
## 4       5       2       5     5      1          5
## ...   ...     ...     ...   ...    ...        ...
## 232     4       2       4     4      1          4
## 233     1       2       3     3      2          4
## 234     1       3       3     2      2          3
## 235     1       2       4     2      3          4
```

Check Factorability

1. Factorability is the assumption that there are at least some correlations amongst the variables so that coherent factors can be identified. Basically, there should be some degree of collinearity among the variables but not an extreme degree or singularity among the variables.

Factorability can be examined via any of the following:

a. Inter-item correlations (correlation matrix) - are there at least several sizable correlations - e.g. > 0.5?

Correlation Matrix: To do the factor analysis we must have variables that correlate fairly well with each other. The correlation matrix is generated in R to check the pattern of relationship between variables.

create a correlation matrix

```
cereals_num_data <- subset(cereals_data, select = -c(1))
cerealsMatrix<-cor(cereals_num_data)
round(cerealsMatrix, 2)
```

```
##             Filling Natural Fibre Sweet  Easy  Salt Satisfying Energy   Fun
## Filling        1.00    0.54  0.55  0.19  0.24 -0.04       0.65   0.64  0.27
## Natural        0.54    1.00  0.65 -0.09  0.23 -0.22       0.46   0.49  0.08
## Fibre          0.55    0.65  1.00 -0.04  0.17 -0.17       0.41   0.50  0.06
## Sweet          0.19   -0.09 -0.04  1.00  0.13  0.44       0.18   0.18  0.33
## Easy           0.24    0.23  0.17  0.13  1.00  0.03       0.36   0.18  0.25
## Salt          -0.04   -0.22 -0.17  0.44  0.03  1.00       0.00  -0.07  0.03
## Satisfying     0.65    0.46  0.41  0.18  0.36  0.00       1.00   0.60  0.35
## Energy         0.64    0.49  0.50  0.18  0.18 -0.07       0.60   1.00  0.35
## Fun            0.27    0.08  0.06  0.33  0.25  0.03       0.35   0.35  1.00
## Kids           0.16    0.06 -0.09  0.12  0.25  0.03       0.31   0.13  0.35
## Soggy         -0.06    0.07 -0.04 -0.08 -0.01  0.02      -0.01  -0.05 -0.10
## Economical     0.05    0.10 -0.03 -0.24  0.09 -0.13       0.21   0.03  0.04
## Health         0.55    0.69  0.68 -0.12  0.20 -0.23       0.52   0.52  0.10
## Family         0.23    0.11 -0.01  0.04  0.24 -0.08       0.35   0.19  0.35
## Calories       0.05   -0.16 -0.19  0.47 -0.02  0.44       0.01   0.03  0.11
## Plain         -0.25   -0.14 -0.12 -0.29  0.02  0.02      -0.18  -0.26 -0.32
## Crisp          0.13    0.02  0.05  0.26  0.25  0.10       0.28   0.25  0.40
## Regular        0.42    0.42  0.65 -0.03  0.11 -0.16       0.33   0.39  0.14
## Sugar         -0.08   -0.32 -0.23  0.65 -0.01  0.59      -0.08  -0.09  0.17
## Fruit          0.26    0.30  0.29  0.35  0.04  0.03       0.25   0.27  0.25
## Process       -0.23   -0.30 -0.20  0.12 -0.05  0.30      -0.16  -0.10  0.00
## Quality        0.44    0.58  0.51 -0.08  0.17 -0.22       0.47   0.46  0.22
## Treat          0.34    0.17  0.14  0.38  0.20  0.13       0.38   0.32  0.59
## Boring        -0.18   -0.22 -0.10 -0.20 -0.17  0.11      -0.32  -0.22 -0.30
## Nutritious     0.53    0.65  0.71 -0.05  0.21 -0.16       0.50   0.54  0.16
##               Kids Soggy Economical Health Family Calories Plain Crisp
## Filling       0.16 -0.06       0.05   0.55   0.23     0.05 -0.25  0.13
## Natural       0.06  0.07       0.10   0.69   0.11    -0.16 -0.14  0.02
## Fibre        -0.09 -0.04      -0.03   0.68  -0.01    -0.19 -0.12  0.05
## Sweet         0.12 -0.08      -0.24  -0.12   0.04     0.47 -0.29  0.26
## Easy          0.25 -0.01       0.09   0.20   0.24    -0.02  0.02  0.25
## Salt          0.03  0.02      -0.13  -0.23  -0.08     0.44  0.02  0.10
## Satisfying    0.31 -0.01       0.21   0.52   0.35     0.01 -0.18  0.28
## Energy        0.13 -0.05       0.03   0.52   0.19     0.03 -0.26  0.25
## Fun           0.35 -0.10       0.04   0.10   0.35     0.11 -0.32  0.40
## Kids          1.00  0.09       0.30  -0.01   0.73     0.01  0.03  0.30
## Soggy         0.09  1.00       0.12   0.01   0.08    -0.08  0.35 -0.34
## Economical    0.30  0.12       1.00   0.19   0.23    -0.21  0.23  0.09
```

```
## Health      -0.01  0.01       0.19   1.00   0.08    -0.31 -0.10  0.08
## Family       0.73  0.08       0.23   0.08   1.00    -0.06 -0.03  0.29
## Calories     0.01 -0.08      -0.21  -0.31  -0.06     1.00 -0.08  0.15
## Plain        0.03  0.35       0.23  -0.10  -0.03    -0.08  1.00 -0.21
## Crisp        0.30 -0.34       0.09   0.08   0.29     0.15 -0.21  1.00
## Regular     -0.03 -0.14       0.08   0.54   0.04    -0.16 -0.08  0.13
## Sugar       -0.02 -0.09      -0.29  -0.38  -0.05     0.53 -0.15  0.17
## Fruit       -0.23 -0.14      -0.34   0.27  -0.12     0.13 -0.34  0.09
## Process      0.03  0.06      -0.13  -0.29  -0.01     0.27  0.11  0.02
## Quality      0.12 -0.03       0.22   0.69   0.24    -0.20 -0.23  0.13
## Treat        0.29 -0.25      -0.04   0.21   0.30     0.19 -0.43  0.47
## Boring      -0.19  0.23      -0.02  -0.23  -0.25    -0.03  0.33 -0.32
## Nutritious   0.03  0.03       0.13   0.76   0.09    -0.23 -0.14  0.10
##             Regular Sugar Fruit Process Quality Treat Boring Nutritious
## Filling        0.42 -0.08  0.26   -0.23    0.44  0.34  -0.18       0.53
## Natural        0.42 -0.32  0.30   -0.30    0.58  0.17  -0.22       0.65
## Fibre          0.65 -0.23  0.29   -0.20    0.51  0.14  -0.10       0.71
## Sweet         -0.03  0.65  0.35    0.12   -0.08  0.38  -0.20      -0.05
## Easy           0.11 -0.01  0.04   -0.05    0.17  0.20  -0.17       0.21
## Salt          -0.16  0.59  0.03    0.30   -0.22  0.13   0.11      -0.16
## Satisfying     0.33 -0.08  0.25   -0.16    0.47  0.38  -0.32       0.50
## Energy         0.39 -0.09  0.27   -0.10    0.46  0.32  -0.22       0.54
## Fun            0.14  0.17  0.25    0.00    0.22  0.59  -0.30       0.16
## Kids          -0.03 -0.02 -0.23    0.03    0.12  0.29  -0.19       0.03
## Soggy         -0.14 -0.09 -0.14    0.06   -0.03 -0.25   0.23       0.03
## Economical     0.08 -0.29 -0.34   -0.13    0.22 -0.04  -0.02       0.13
## Health         0.54 -0.38  0.27   -0.29    0.69  0.21  -0.23       0.76
## Family         0.04 -0.05 -0.12   -0.01    0.24  0.30  -0.25       0.09
## Calories      -0.16  0.53  0.13    0.27   -0.20  0.19  -0.03      -0.23
## Plain         -0.08 -0.15 -0.34    0.11   -0.23 -0.43   0.33      -0.14
## Crisp          0.13  0.17  0.09    0.02    0.13  0.47  -0.32       0.10
## Regular        1.00 -0.09  0.25   -0.15    0.44  0.17  -0.09       0.57
## Sugar         -0.09  1.00  0.15    0.37   -0.26  0.22   0.00      -0.27
## Fruit          0.25  0.15  1.00   -0.14    0.16  0.31  -0.26       0.31
## Process       -0.15  0.37 -0.14    1.00   -0.18  0.03   0.17      -0.28
## Quality        0.44 -0.26  0.16   -0.18    1.00  0.33  -0.28       0.66
## Treat          0.17  0.22  0.31    0.03    0.33  1.00  -0.36       0.24
## Boring        -0.09  0.00 -0.26    0.17   -0.28 -0.36   1.00      -0.17
## Nutritious     0.57 -0.27  0.31   -0.28    0.66  0.24  -0.17       1.00
```

```
#nth(sort(x,decreasing=T),2)
df <- as.data.frame(cerealsMatrix)
apply(df,2,function(x) nth(sort(x,decreasing=T),2))
```

```
##     Filling     Natural       Fibre       Sweet        Easy        Salt
##   0.6485072   0.6880977   0.7130650   0.6483827   0.3623098   0.5917709
##  Satisfying      Energy         Fun        Kids       Soggy  Economical
##   0.6485072   0.6367588   0.5864972   0.7270111   0.3461298   0.3049121
##      Health      Family    Calories       Plain       Crisp     Regular
##   0.7576148   0.7270111   0.5258262   0.3461298   0.4680856   0.6483757
##       Sugar       Fruit     Process     Quality       Treat      Boring
##   0.6483827   0.3465054   0.3692534   0.6863048   0.5864972   0.3305255
##Nutritious
## 0.7576148
```

Observation

From the above correlation matrix we can see that all the variables value of correlation coefficient greater than 0.3 with at least one other variable. 17 variables have correlation coefficient of at least 0.5 and only 8 variables have at least 0.3.

Hence we can assume that variables are fairly correlated with each other and we can run Factor Analysis on this data.

b. Anti-image correlation matrix diagonals - they should be > 0.5

```
X <- cor(cerealsMatrix)
iX <- ginv(X)
S2 <- diag(diag((iX^-1)))
AIS <- S2%*%iX%*%S2                      # anti-image covariance matrix
IS <- X+AIS-2*S2                         # image covariance matrix
Dai <- sqrt(diag(diag(AIS)))
IR <- ginv(Dai)%*%IS%*%ginv(Dai)         # image correlation matrix
AIR <- ginv(Dai)%*%AIS%*%ginv(Dai)       # anti-image correlation matrix

print(diag(AIR), row.names = FALSE)
```

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

Observation

We observe that the diagonals of the Anti - Image Correlation matrix to be 1.

Since Anti - Image Correlation matrix diagonals > 0.5, we can run Factor Analysis on this data.

c. Measure of Sampling Accuracy (MSA)

• Kaisar-Meyer-Olkin (KMO) should be > 0.5

Kaiser-Meyer-Olkin (KMO) Test :

```
KMO(r = cerealsMatrix)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cerealsMatrix)
## Overall MSA =  0.85
## MSA for each item =
##    Filling    Natural      Fibre      Sweet       Easy       Salt
##       0.89       0.90       0.88       0.78       0.83       0.82
## Satisfying     Energy        Fun       Kids      Soggy Economical
##       0.91       0.91       0.85       0.68       0.63       0.73
##     Health     Family   Calories      Plain      Crisp    Regular
##       0.92       0.73       0.86       0.82       0.83       0.87
##      Sugar      Fruit    Process    Quality      Treat     Boring
##       0.78       0.77       0.79       0.91       0.88       0.88
## Nutritious
##       0.92
```

Observation

We have applied the function, KMO() on the correlation matrix and it has returned the following:

a) Overall MSA to be 0.85 which yields a degree of common variance middling

b) The estimates of MSA for each item to be 0.89, 0.90, 0.88, 0.78, 0.83, 0.82, 0.91, 0.91, 0.85, 0.68, 0.63, 0.73,0.92, 0.73, 0.86, 0.82, 0.83, 0.87, 0.78, 0.77, 0.79, 0.91, 0.88, 0.88, 0.92

Since MSA > 0.5, we can run Factor Analysis on this data.

ref: *http://personality-project.org/r/html/KMO.html*

• Bartlett's test of sphericity (Should be significant)

```
cortest.bartlett(cerealsMatrix, n = 100)
```

```
## $chisq
## [1] 1153.931
##
## $p.value
## [1] 2.437858e-100
##
## $df
## [1] 300
```

Bartlettas test was conducted in R and it was found to be significant (P < 0.001) . The significance of this test tells us that the correlation matrix is not an identity matrix. Hence, we assume that there is some relationship between the variable.

2. Sample Size - The sample size should be large enough to yield reliable estimates of correlations among the variables:

a. Ideally, there should be a large ratio of N / k (Cases / items)
b. EFA can still be reasonably done with > ~ 5:1
c. In this case we have N = 236 and k = 26 and the ratio is ~ 9 : 1. Hence the sample size is large enough to yield reliable estimates of the correlations among the variables.

Hence the sample size is large enough to yield reliable estimates of the correlations among the variables.

(B) Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis (EFA) is generally used to discover structure of a measure and to examine its internal reliability. EFA is often recommended when researchers have no hypotheses about the nature of the underlying factor structure of their measure.

Exploratory factor analysis has three basic decision points:

(1) Decide the number of factors
(2) Choosing an extraction method
(3) Choosing a rotation method

C) Determine the number of Factors

The most common approach to deciding the number of factors is to generate a scree plot.

The scree plot is a two-dimensional graph with factors on the x-axis and *eigenvalues* on the y-axis.

Eigenvalues are produced by a process called **principal component analysis** (PCA) and represent the variance accounted for by each underlying factor. They are represented by scores that total to the number of items.

A 12-item scale will theoretically have 12 possible underlying factors, each factor will have an eigen value that indicates the amount of variation in the items accounted by each factor. If a, the first factor has an eigen value of 3.0, it accounts for 25% of the variance (3/12= .25). The total of all the eigen values will be 12 if there are 12 items, so some factors will have smaller eigenvalues. They are typically arranged in a scree plot in descending order.

(2) Choosing an extraction method and extraction

Once the number of factors are decided, you need to decide which mathematical solution to find the loadings. There are five basic extraction methods:

1. PCA - which assumes there is no measurement error and is considered not to be true exploratory factor analysis.
2. Maximum Likelihood (a.k.a canonical factoring)
3. Alpha Factoring
4. Image Factoring
5. Principal axis factoring with iterated communalities (a.k.a least squares)

Calculate initial factor loadings:

This can be done in a number of different ways: the two most common methods are described very briefly below:

- Principal Component Analysis (PCA) Method

As the name suggests, this method uses the method used to carry out a principal component analysis. However, the factors obtained will not actually be the principal components (although the loadings for the k^{th} factor will be proportional to the coefficients of the k^{th} principal component.

- Principal Axis Factoring

This is a method which tries to find the lowest number of factors which can account for the variability in the original variables, that is associated with these factors (this is in contrast to the

principal components method which looks for a set of factors which can account for the total variability in the original variables).

These two methods will tend to give similar results if the variables are quite highly correlated and / or the number of original variables is quite high. Whichever method is used, the resulting factors at this stage will be uncorrelated.

PCA

Let us perform an initial PCA and understand the eigen values and the variance explained by them.

```
#On raw data

pc1 <-  principal(cereals_num_data, nfactors = length(cereals_num_data), rota
te = "none")
pc1
```

```
## Principal Components Analysis
## Call: principal(r = cereals_num_data, nfactors = length(cereals_num_data),

##     rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
## Filling     0.75  0.09 -0.08  0.22 -0.12 -0.01 -0.24 -0.35 -0.02  0.04
## Natural     0.75 -0.26 -0.12  0.13 -0.14 -0.06  0.00  0.05 -0.18 -0.07
## Fibre       0.73 -0.25 -0.33  0.19  0.16  0.02  0.12 -0.11  0.02 -0.15
## Sweet       0.09  0.77 -0.20  0.18 -0.17 -0.12 -0.06  0.07  0.17 -0.10
## Easy        0.35  0.16  0.28  0.16  0.03 -0.63  0.45 -0.09 -0.15  0.19
## Salt       -0.22  0.55 -0.13  0.48  0.13 -0.11 -0.18  0.20  0.03 -0.01
## Satisfying  0.74  0.17  0.18  0.20 -0.10 -0.12 -0.16 -0.08 -0.08  0.11
## Energy      0.73  0.13 -0.08  0.17 -0.04  0.13 -0.13 -0.25 -0.10  0.18
## Fun         0.42  0.53  0.24 -0.16 -0.09  0.17  0.21  0.03  0.30  0.33
## Kids        0.22  0.27  0.78  0.09 -0.09  0.10  0.01 -0.12  0.05 -0.25
## Soggy      -0.11 -0.27  0.19  0.57 -0.51  0.16  0.18  0.26  0.10  0.06
## Economical  0.16 -0.27  0.59  0.10  0.23 -0.09 -0.41  0.33  0.15  0.20
## Health      0.81 -0.32 -0.12  0.09  0.08 -0.01 -0.01  0.17 -0.06  0.02
## Family      0.32  0.21  0.72  0.01 -0.15  0.20  0.09 -0.15  0.00 -0.31
## Calories   -0.17  0.63 -0.18  0.28 -0.02 -0.05 -0.27 -0.07 -0.22 -0.03
## Plain      -0.33 -0.39  0.27  0.49  0.15 -0.22  0.11  0.07  0.10 -0.15
## Crisp       0.31  0.50  0.26 -0.24  0.42 -0.17  0.03  0.02  0.06  0.00
## Regular     0.62 -0.15 -0.22  0.09  0.39  0.09  0.11 -0.05  0.33 -0.28
## Sugar      -0.25  0.75 -0.23  0.26  0.10  0.01  0.00  0.10  0.11 -0.15
## Fruit       0.39  0.27 -0.55 -0.14 -0.29 -0.05  0.19  0.15  0.16 -0.05
## Process    -0.33  0.32  0.02  0.35  0.35  0.41  0.32  0.05 -0.40  0.08
## Quality     0.75 -0.16  0.04 -0.01  0.09  0.21 -0.02  0.28 -0.17  0.04
```

```
## Treat       0.49  0.59  0.09 -0.19  0.06  0.18  0.08  0.13  0.05  0.17
## Boring     -0.41 -0.29 -0.11  0.44  0.17  0.16  0.04 -0.34  0.32  0.32
## Nutritious  0.81 -0.23 -0.15  0.15  0.07  0.03  0.04  0.14  0.06 -0.02
##              PC11  PC12  PC13  PC14  PC15  PC16  PC17  PC18  PC19  PC20
## Filling    -0.04 -0.08  0.06  0.08 -0.08  0.03  0.09 -0.19  0.14 -0.01
## Natural    -0.04  0.20  0.08 -0.06  0.16 -0.25  0.15  0.07  0.24 -0.01
## Fibre      -0.03  0.02 -0.01 -0.10 -0.02 -0.14  0.09 -0.07 -0.05 -0.18
## Sweet      -0.05 -0.17 -0.05 -0.14  0.10  0.18  0.28  0.10 -0.10 -0.06
## Easy       -0.17 -0.09  0.14 -0.06  0.04  0.02 -0.06  0.01 -0.03  0.10
## Salt       -0.29  0.12 -0.21  0.19 -0.16 -0.23 -0.14  0.03  0.00  0.06
## Satisfying  0.09 -0.16 -0.08  0.26  0.03  0.11 -0.16 -0.12 -0.04 -0.21
## Energy      0.23 -0.10 -0.26 -0.12 -0.06 -0.02  0.00  0.21 -0.03  0.27
## Fun         0.14 -0.04  0.12 -0.06 -0.21 -0.20 -0.04  0.14  0.08 -0.17
## Kids       -0.10  0.08  0.01  0.02  0.08 -0.12  0.10  0.09 -0.19 -0.03
## Soggy       0.09  0.01 -0.12 -0.24  0.04  0.01 -0.09 -0.24  0.01  0.04
## Economical  0.08 -0.18  0.15  0.02  0.22 -0.10  0.07  0.00  0.01  0.06
## Health     -0.03  0.03 -0.07  0.06 -0.05  0.08  0.08  0.01 -0.01  0.01
## Family     -0.07  0.05  0.01  0.10  0.06  0.09 -0.13  0.05  0.12  0.07
## Calories    0.26  0.25  0.37 -0.16  0.01  0.00 -0.14  0.00 -0.11 -0.01
## Plain       0.37  0.14  0.03  0.17 -0.25  0.13  0.16  0.10  0.08 -0.01
## Crisp       0.20  0.25 -0.34 -0.18  0.16  0.06 -0.06 -0.11  0.10 -0.07
## Regular     0.08 -0.19  0.18 -0.06 -0.02 -0.06 -0.15 -0.11 -0.03  0.14
## Sugar      -0.12 -0.16  0.03 -0.09  0.02  0.07 -0.01  0.04  0.22 -0.01
## Fruit       0.23  0.07  0.03  0.34  0.26 -0.01 -0.06  0.08 -0.01  0.07
## Process     0.12 -0.18 -0.01  0.12  0.13 -0.06  0.08 -0.03 -0.03 -0.04
## Quality    -0.17  0.04  0.12 -0.08 -0.05  0.27 -0.13  0.19  0.10 -0.04
## Treat      -0.10  0.23  0.12  0.12 -0.12  0.09  0.21 -0.26 -0.04  0.17
## Boring     -0.19  0.23  0.04  0.05  0.21  0.13 -0.01  0.08  0.02 -0.02
## Nutritious -0.09  0.11 -0.06 -0.05 -0.01  0.03 -0.04  0.08 -0.25 -0.05
##              PC21  PC22  PC23  PC24  PC25 h2       u2 com
## Filling     0.21  0.05 -0.19  0.10  0.01  1  7.8e-16 3.0
## Natural    -0.17  0.03 -0.02 -0.02 -0.13  1  3.3e-16 3.0
## Fibre       0.09 -0.13  0.24  0.08  0.14  1  1.4e-15 3.3
## Sweet       0.04  0.12  0.10  0.00 -0.13  1 -1.3e-15 2.7
## Easy        0.03  0.00  0.00  0.01  0.02  1  0.0e+00 4.5
## Salt        0.06  0.08  0.06  0.04 -0.03  1  1.2e-15 6.0
## Satisfying -0.24 -0.02  0.07 -0.03 -0.04  1  1.0e-15 3.1
## Energy     -0.07 -0.06  0.06  0.02  0.05  1  1.2e-15 3.3
## Fun         0.05  0.03 -0.02 -0.05 -0.03  1 -6.7e-16 7.2
## Kids       -0.11  0.12 -0.12  0.04  0.16  1  3.3e-16 2.6
## Soggy       0.01  0.03  0.00  0.02  0.03  1 -6.7e-16 5.1
## Economical  0.08 -0.09  0.04  0.03  0.02  1  7.8e-16 5.8
## Health      0.10  0.11 -0.02 -0.34  0.11  1  1.0e-15 2.2
## Family      0.16 -0.13  0.13 -0.08 -0.10  1  6.7e-16 3.4
## Calories    0.05  0.00  0.03 -0.07  0.01  1 -4.4e-16 5.0
## Plain      -0.02 -0.02 -0.01  0.06 -0.01  1  3.3e-16 8.0
```

```
## Crisp       0.04  0.05 -0.05  0.03  0.00  1  1.0e-15 7.6
## Regular    -0.09  0.15 -0.01 -0.01 -0.08  1  1.4e-15 5.1
## Sugar      -0.12 -0.21 -0.12 -0.06  0.14  1  8.9e-16 3.0
## Fruit       0.08  0.01 -0.02  0.07  0.09  1  3.3e-16 6.6
## Process     0.06  0.03 -0.04  0.01 -0.05  1  4.4e-16 8.4
## Quality     0.00  0.13  0.03  0.17  0.06  1  2.0e-15 2.9
## Treat      -0.11 -0.08  0.07  0.02  0.00  1  2.0e-15 5.2
## Boring     -0.04  0.03  0.03 -0.04  0.00  1 -1.8e-15 8.6
## Nutritious  0.04 -0.23 -0.21  0.01 -0.13  1  7.8e-16 2.3
##
##                         PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10
## SS loadings            6.50 3.82 2.50 1.68 1.09 0.93 0.85 0.79 0.73 0.70
## Proportion Var         0.26 0.15 0.10 0.07 0.04 0.04 0.03 0.03 0.03 0.03
## Cumulative Var         0.26 0.41 0.51 0.58 0.62 0.66 0.70 0.73 0.76 0.78
## Proportion Explained   0.26 0.15 0.10 0.07 0.04 0.04 0.03 0.03 0.03 0.03
## Cumulative Proportion  0.26 0.41 0.51 0.58 0.62 0.66 0.70 0.73 0.76 0.78
##                        PC11 PC12 PC13 PC14 PC15 PC16 PC17 PC18 PC19 PC20
## SS loadings            0.65 0.55 0.53 0.49 0.42 0.39 0.36 0.36 0.30 0.27
## Proportion Var         0.03 0.02 0.02 0.02 0.02 0.02 0.01 0.01 0.01 0.01
## Cumulative Var         0.81 0.83 0.85 0.87 0.89 0.90 0.92 0.93 0.95 0.96
## Proportion Explained   0.03 0.02 0.02 0.02 0.02 0.02 0.01 0.01 0.01 0.01
## Cumulative Proportion  0.81 0.83 0.85 0.87 0.89 0.90 0.92 0.93 0.95 0.96
##                        PC21 PC22 PC23 PC24 PC25
## SS loadings            0.26 0.24 0.22 0.20 0.16
## Proportion Var         0.01 0.01 0.01 0.01 0.01
## Cumulative Var         0.97 0.98 0.99 0.99 1.00
## Proportion Explained   0.01 0.01 0.01 0.01 0.01
## Cumulative Proportion  0.97 0.98 0.99 0.99 1.00
##
## Mean item complexity =  4.7
## Test of the hypothesis that 25 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

```
cat("\n\n Eigen values \n")
```

```
##
##
##  Eigen values
```

```
print(pc1$values)
```

```
##  [1] 6.5044682 3.8210452 2.5019953 1.6839941 1.0853540 0.9330069 0.8516405
##  [8] 0.7868366 0.7317169 0.6958507 0.6468294 0.5479585 0.5291569 0.4896236
## [15] 0.4177428 0.3870740 0.3624610 0.3588295 0.3047541 0.2741921 0.2624481
## [22] 0.2422794 0.2179754 0.1985309 0.1642362
```

Observation

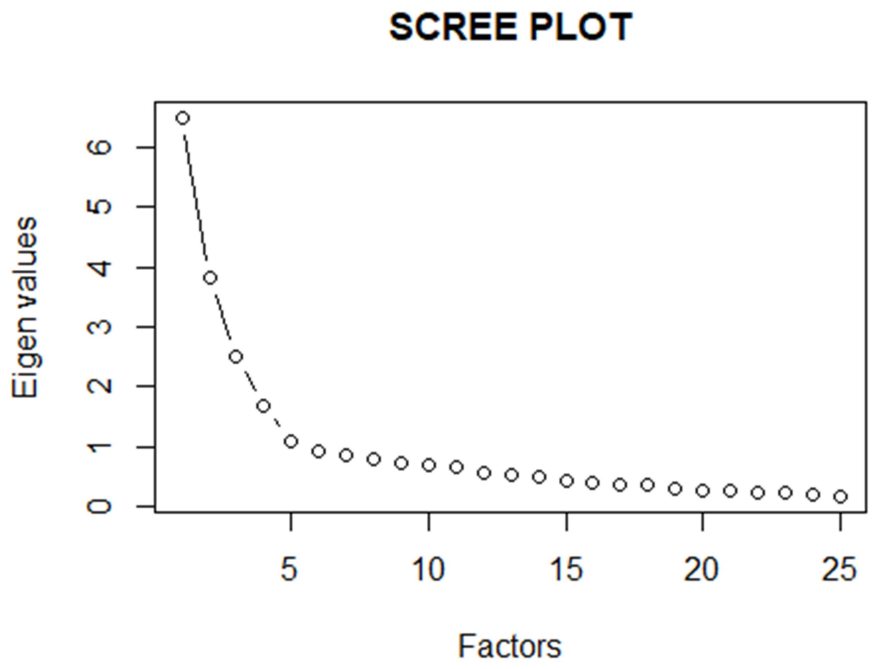
1) Output of this analysis show us that only 5 components have eigenvalues greater than 1, suggesting that we extract 5 components.

2) The above output also suggests that extracting 5 components explains 62% of the total variance.

We shall draw a scree plot to decide on the number of factors.

Scree Plot:

```
### Scree plot

plot(pc1$values, type = "b",xlab = "Factors", ylab = "Eigen values", main = "
SCREE PLOT")
```



Observation

From the **scree plot**, we notice a steep curve before the factor 6, that starts the flat line. We retain those components or factors in the steep curve before the first point that starts the flat line. We notice that 5 of those factors explain most of the variation - 62%.

So we shall use 5 as the number of factors for performing Factor Analysis.

Principal axis factoring

We shall use **Principal axis factoring**, (fm="pa") because we are most interested in identifying the underlying constructs in the data.

The extraction method will produce factor loadings for every item in every extracted factor.

Now, We will use **fa()** function from the *psych* package, which received the following primary arguments:

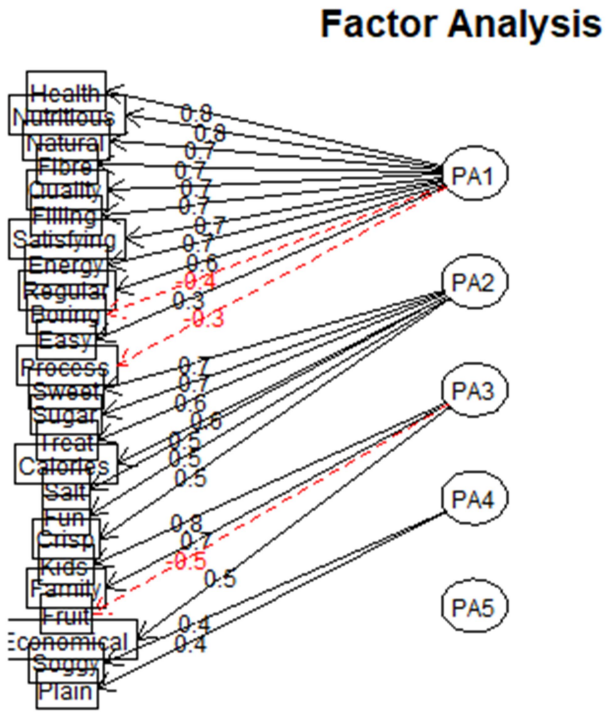
- r: the correlation matrix
- nfactors: number of factors to be extracted (default 1)
- rotate: one of several matrix rotation methods, such as "varimax" or "oblimin" or "none"
- fm: one of several factoring methodsm such as **pa** (principal axis) or **ml** (maximum likelihood)

```
solution <- fa(r=cerealsMatrix, nfactors = 5, rotate="none",fm="pa")

###

print(solution)
```

```
## Factor Analysis using method =  pa
## Call: fa(r = cerealsMatrix, nfactors = 5, rotate = "none", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              PA1   PA2   PA3   PA4   PA5   h2   u2 com
## Filling     0.72  0.10 -0.07  0.19 -0.14 0.59 0.41 1.3
## Natural     0.73 -0.24 -0.11  0.11 -0.14 0.64 0.36 1.4
## Fibre       0.73 -0.24 -0.31  0.16  0.15 0.73 0.27 1.8
## Sweet       0.09  0.74 -0.21  0.16 -0.13 0.65 0.35 1.3
## Easy        0.32  0.14  0.20  0.10  0.03 0.17 0.83 2.4
## Salt       -0.21  0.50 -0.14  0.41  0.11 0.50 0.50 2.6
## Satisfying  0.72  0.18  0.16  0.17 -0.10 0.62 0.38 1.4
## Energy      0.70  0.13 -0.06  0.12 -0.06 0.53 0.47 1.2
## Fun         0.39  0.49  0.21 -0.15 -0.04 0.46 0.54 2.5
## Kids        0.22  0.28  0.76  0.12 -0.06 0.72 0.28 1.5
## Soggy      -0.10 -0.24  0.15  0.45 -0.29 0.37 0.63 2.7
## Economical  0.15 -0.23  0.48  0.10  0.15 0.34 0.66 2.0
## Health      0.81 -0.31 -0.11  0.08  0.08 0.78 0.22 1.4
## Family      0.31  0.22  0.67  0.03 -0.09 0.60 0.40 1.7
## Calories   -0.16  0.56 -0.17  0.21 -0.03 0.42 0.58 1.7
## Plain      -0.31 -0.36  0.22  0.42  0.12 0.47 0.53 3.6
## Crisp       0.29  0.47  0.22 -0.21  0.33 0.51 0.49 3.6
## Regular     0.59 -0.13 -0.19  0.07  0.29 0.49 0.51 1.8
## Sugar      -0.25  0.74 -0.26  0.26  0.12 0.75 0.25 1.8
## Fruit       0.37  0.25 -0.48 -0.15 -0.22 0.51 0.49 3.2
```

```
## Process    -0.30  0.27  0.00  0.23  0.18 0.25 0.75 3.6
## Quality     0.73 -0.14  0.05 -0.03  0.07 0.55 0.45 1.1
## Treat       0.47  0.56  0.08 -0.20  0.07 0.59 0.41 2.3
## Boring     -0.38 -0.26 -0.10  0.33  0.09 0.34 0.66 3.1
## Nutritious  0.80 -0.22 -0.14  0.13  0.08 0.73 0.27 1.3
##
##                        PA1  PA2  PA3  PA4  PA5
## SS loadings           6.11 3.39 2.07 1.15 0.56
## Proportion Var        0.24 0.14 0.08 0.05 0.02
## Cumulative Var        0.24 0.38 0.46 0.51 0.53
## Proportion Explained  0.46 0.26 0.16 0.09 0.04
## Cumulative Proportion 0.46 0.72 0.87 0.96 1.00
##
## Mean item complexity =  2.1
## Test of the hypothesis that 5 factors are sufficient.
##
## The degrees of freedom for the null model are  300  and the objective func
tion was  12.85
## The degrees of freedom for the model are 185  and the objective function w
as  1.51
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1  PA2  PA3  PA4  PA5
## Correlation of (regression) scores with factors   0.97 0.95 0.92 0.84 0.74
## Multiple R square of scores with factors          0.94 0.90 0.85 0.70 0.55
## Minimum correlation of possible factor scores     0.89 0.79 0.69 0.40 0.09
```

```
fa.diagram(solution)
```



(3) Choosing a rotation method

We observe that the **components loadings** are not clear. Once an initial solution is obtained, the loadings are rotated. **Factor Rotation is used to increase interpretability**.

Rotation is a way of maximizing high loadings and minimizing low loadings so that the simplest possible structure is achieved.

There are two types of rotation method, **orthogonal** and **oblique** rotation. In orthogonal rotation the rotated factors will remain uncorrelated whereas in oblique rotation the resulting factors will be correlated.

There are a number of different methods of rotation of each type.

The most common orthogonal method is called **varimax** rotation; this is the method that many books recommend.

http://www.statstutor.ac.uk/resources/uploaded/factoranalysis.pdf

Orthogonal Rotation (varimax):

Assuming that there is no correlation between the extracted factors, we will carry out a varimax rotation.

Rotated component matrix obtained after **varimax** rotation is shown below:

```
# Orthogonal varimax rotation

solution1 <- fa(r=cerealsMatrix, nfactors = 5, rotate="varimax",fm="pa")
print(solution1)
```

```
## Factor Analysis using method =  pa
## Call: fa(r = cerealsMatrix, nfactors = 5, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              PA1   PA3   PA2   PA4   PA5   h2   u2 com
## Filling     0.68  0.25  0.06  0.25 -0.07 0.59 0.41 1.6
## Natural     0.73  0.09 -0.23  0.16 -0.13 0.64 0.36 1.4
## Fibre       0.84 -0.12 -0.10  0.02  0.05 0.73 0.27 1.1
## Sweet       0.04  0.15  0.67  0.40  0.06 0.65 0.35 1.8
## Easy        0.24  0.33  0.06  0.01  0.04 0.17 0.83 1.9
## Salt       -0.08  0.01  0.69 -0.07 -0.03 0.50 0.50 1.1
## Satisfying  0.60  0.47  0.05  0.19 -0.02 0.62 0.38 2.1
## Energy      0.64  0.24  0.06  0.24  0.03 0.53 0.47 1.6
## Fun         0.15  0.48  0.16  0.33  0.26 0.46 0.54 2.9
## Kids       -0.03  0.84  0.02 -0.10 -0.02 0.72 0.28 1.0
## Soggy       0.01  0.08  0.00 -0.21 -0.57 0.37 0.63 1.3
## Economical  0.09  0.36 -0.26 -0.36 -0.01 0.34 0.66 2.9
```

```
## Health      0.83  0.06 -0.28  0.03  0.04 0.78 0.22 1.2
## Family      0.05  0.77 -0.07 -0.01  0.00 0.60 0.40 1.0
## Calories   -0.13  0.02  0.61  0.16  0.01 0.42 0.58 1.2
## Plain      -0.12 -0.03 -0.03 -0.59 -0.32 0.47 0.53 1.7
## Crisp       0.10  0.40  0.17  0.09  0.55 0.51 0.49 2.2
## Regular     0.66 -0.05 -0.07 -0.06  0.23 0.49 0.51 1.3
## Sugar      -0.18 -0.03  0.83  0.12  0.14 0.75 0.25 1.2
## Fruit       0.34 -0.18  0.15  0.58  0.08 0.51 0.49 2.1
## Process    -0.21 -0.01  0.40 -0.20  0.03 0.25 0.75 2.0
## Quality     0.65  0.22 -0.24  0.08  0.13 0.55 0.45 1.7
## Treat       0.24  0.40  0.23  0.39  0.41 0.59 0.41 4.2
## Boring     -0.14 -0.29  0.09 -0.39 -0.27 0.34 0.66 3.1
## Nutritious  0.83  0.06 -0.17  0.05  0.04 0.73 0.27 1.1
##
##                        PA1  PA3  PA2  PA4  PA5
## SS loadings           5.10 2.66 2.62 1.73 1.16
## Proportion Var        0.20 0.11 0.10 0.07 0.05
## Cumulative Var        0.20 0.31 0.42 0.48 0.53
## Proportion Explained  0.38 0.20 0.20 0.13 0.09
## Cumulative Proportion 0.38 0.58 0.78 0.91 1.00
##
## Mean item complexity =  1.8
## Test of the hypothesis that 5 factors are sufficient.
##
## The degrees of freedom for the null model are  300  and the objective func
tion was  12.85
## The degrees of freedom for the model are 185  and the objective function w
as  1.51
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1  PA3  PA2  PA4  PA5
## Correlation of (regression) scores with factors   0.96 0.93 0.92 0.83 0.79
## Multiple R square of scores with factors          0.92 0.86 0.84 0.69 0.63
## Minimum correlation of possible factor scores     0.84 0.72 0.68 0.38 0.26
```

```
print(solution1$loadings, cutoff=0.4)
```

```
##
## Loadings:
##            PA1    PA3    PA2    PA4    PA5
## Filling     0.677
## Natural     0.730
## Fibre       0.836
## Sweet                      0.674  0.405
## Easy
## Salt                       0.695
## Satisfying  0.600  0.467
## Energy      0.638
## Fun                0.478
## Kids               0.840
## Soggy                                -0.566
## Economical
## Health      0.831
## Family             0.771
## Calories                   0.611
## Plain                            -0.590
## Crisp              0.401                0.545
## Regular     0.657
## Sugar                      0.829
## Fruit                             0.575
## Process                    0.404
## Quality     0.650
## Treat              0.404                0.408
## Boring
## Nutritious  0.834
##
##                  PA1   PA3   PA2   PA4   PA5
## SS loadings    5.102 2.661 2.623 1.732 1.157
## Proportion Var 0.204 0.106 0.105 0.069 0.046
## Cumulative Var 0.204 0.311 0.415 0.485 0.531
```

Observation:

From the above output we can see that on applying varimax rotation, the component loadings are very clear.

Observation

The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variable(s). The curved arrows are the correlations between the factors. If no curved arrow is present, then the correlation between the factors is not great.

We observe the cross loadings as shown below:

1) Sweet loading on PA2 (0.674) and PA4 (.405)

2) Satisfying loading on PA1 (0.60) and PA3 (0.467)

3) Crisp loading on PA3 (0.401) and PA5 (0.545)

4) Treat loading on PA3 (0.404) and PA5 (0.408)

Let us try another rotation method.

Oblique oblimin rotation

Oblique Rotation (oblimin): AWe will use oblique rotation, which recognizes that there is likely to be some correlation between pain relief factors in the real world.

Rotation matrix obtained after oblimin rotation is shown below:

```
solution2 <- fa(r=cerealsMatrix, nfactors = 5, rotate="oblimin",fm="pa")
print(solution2)
```

```
## Factor Analysis using method =  pa
## Call: fa(r = cerealsMatrix, nfactors = 5, rotate = "oblimin", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              PA1   PA2   PA3   PA4   PA5   h2   u2 com
## Filling     0.66  0.14  0.18 -0.09 -0.20 0.59 0.41 1.5
## Natural     0.71 -0.14  0.04 -0.15 -0.17 0.64 0.36 1.3
## Fibre       0.88  0.02 -0.20  0.01  0.07 0.73 0.27 1.1
## Sweet       0.02  0.67  0.07  0.04 -0.29 0.65 0.35 1.4
## Easy        0.23  0.09  0.30  0.04  0.03 0.17 0.83 2.2
## Salt        0.02  0.73  0.00 -0.09  0.17 0.50 0.50 1.1
## Satisfying  0.57  0.12  0.40 -0.03 -0.15 0.62 0.38 2.1
## Energy      0.62  0.13  0.15  0.01 -0.17 0.53 0.47 1.4
## Fun         0.06  0.13  0.38  0.30 -0.24 0.46 0.54 3.0
## Kids       -0.07  0.02  0.85  0.01  0.08 0.72 0.28 1.0
## Soggy       0.07  0.06  0.20 -0.61  0.05 0.37 0.63 1.3
## Economical  0.12 -0.22  0.40 -0.01  0.33 0.34 0.66 2.8
## Health      0.84 -0.17 -0.01  0.02  0.02 0.78 0.22 1.1
## Family     -0.01 -0.08  0.77  0.04 -0.02 0.60 0.40 1.0
## Calories   -0.10  0.61 -0.01 -0.02 -0.08 0.42 0.58 1.1
## Plain       0.01  0.04  0.10 -0.38  0.50 0.47 0.53 2.0
## Crisp       0.07  0.14  0.28  0.58  0.10 0.51 0.49 1.7
## Regular     0.70  0.03 -0.13  0.19  0.19 0.49 0.51 1.4
## Sugar      -0.11  0.82 -0.08  0.09  0.03 0.75 0.25 1.1
## Fruit       0.26  0.14 -0.29  0.08 -0.51 0.51 0.49 2.4
## Process    -0.13  0.41  0.01  0.00  0.26 0.25 0.75 1.9
## Quality     0.62 -0.17  0.14  0.13 -0.02 0.55 0.45 1.4
```

```
## Treat       0.16  0.20  0.26  0.44 -0.23 0.59 0.41 3.0
## Boring     -0.03  0.14 -0.20 -0.32  0.34 0.34 0.66 3.0
## Nutritious  0.85 -0.06 -0.01  0.01  0.02 0.73 0.27 1.0
##
##                         PA1  PA2  PA3  PA4  PA5
## SS loadings            5.17 2.62 2.44 1.57 1.48
## Proportion Var         0.21 0.10 0.10 0.06 0.06
## Cumulative Var         0.21 0.31 0.41 0.47 0.53
## Proportion Explained   0.39 0.20 0.18 0.12 0.11
## Cumulative Proportion 0.39 0.59 0.77 0.89 1.00
##
##  With factor correlations of
##       PA1   PA2   PA3   PA4   PA5
## PA1  1.00 -0.19  0.15  0.15 -0.29
## PA2 -0.19  1.00  0.05  0.19 -0.21
## PA3  0.15  0.05  1.00  0.16 -0.10
## PA4  0.15  0.19  0.16  1.00 -0.27
## PA5 -0.29 -0.21 -0.10 -0.27  1.00
##
## Mean item complexity =  1.7
## Test of the hypothesis that 5 factors are sufficient.
##
## The degrees of freedom for the null model are  300  and the objective func
tion was  12.85
## The degrees of freedom for the model are 185  and the objective function w
as  1.51
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1  PA2  PA3  PA4  PA5
## Correlation of (regression) scores with factors   0.97 0.93 0.93 0.85 0.85
## Multiple R square of scores with factors          0.94 0.87 0.86 0.72 0.72
## Minimum correlation of possible factor scores     0.88 0.74 0.72 0.45 0.44
```

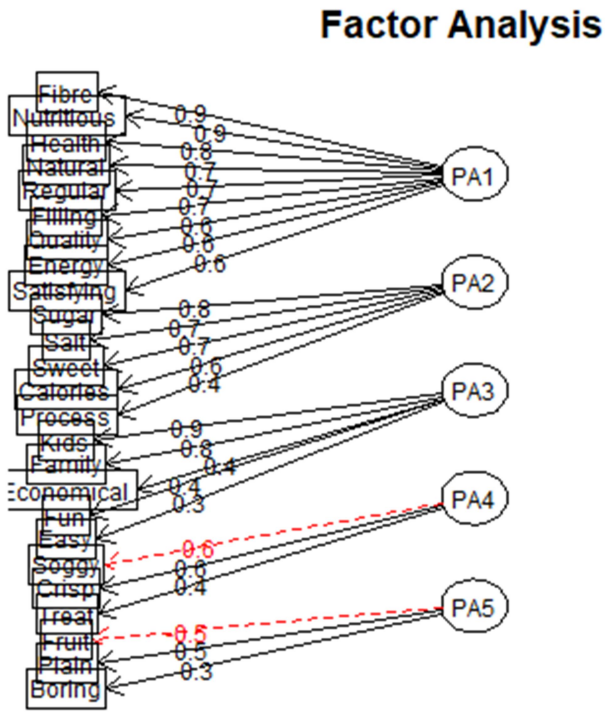
```
print(solution2$loadings, cutoff=0.4)
```

```
##
## Loadings:
##             PA1    PA2    PA3    PA4    PA5
## Filling      0.658
## Natural      0.708
## Fibre        0.881
## Sweet                0.666
## Easy
## Salt                 0.731
## Satisfying   0.570
## Energy       0.616
## Fun
## Kids                        0.855
## Soggy                              -0.611
## Economical                  0.402
## Health       0.837
## Family                      0.768
## Calories             0.606
## Plain                                     0.504
## Crisp                              0.578
## Regular      0.702
## Sugar                0.825
## Fruit                                    -0.511
## Process              0.414
## Quality      0.624
## Treat                              0.444
## Boring
## Nutritious   0.850
##
##                  PA1   PA2   PA3   PA4   PA5
## SS loadings    4.940 2.519 2.339 1.357 1.223
## Proportion Var 0.198 0.101 0.094 0.054 0.049
## Cumulative Var 0.198 0.298 0.392 0.446 0.495
```

Visualization

We can visualize the factors by calling the function fa.diagram.

```
fa.diagram(solution2)
```



From the above output we do not observe any cross loading of items.

(D) Identify which item belong in which factor

Criteria for Practical and Statistical Significance of Factor Loadings:

Factor loading can be classified based on their magnitude.

Value	Interpretation
> 30	Minimum consideration level
> 40	More important
> 50	Practically significant

Items belonging to factor - PA1:

1. Fibre
2. Nutritious
3. Health
4. Natural
5. Filling
6. Regular
7. Quality
8. Energy
9. Satisfying

Items belonging to factor - PA2:

1. Sugar
2. Salt
3. Sweet
4. Calories
5. Process

Items belonging to factor - PA3:

1. Kids
2. Family
3. Economical

Items belonging to factor - PA4:

1. Soggy
2. Crisp
3. Treat

Items belonging to factor - PA5:

1. Plain
2. Fruit

(E) Drop items as necessary and repeat steps (C) and (D)

Dropping items are not necessary.

(F) Name and define factors

The variables that load highly on factor - PA1:

1. Fibre
2. Nutritious
3. Health
4. Natural
5. Filling
6. Regular
7. Quality
8. Energy
9. Satisfying

All these items are related to the health aspects of the Cereal, so we can label this factor as **Healthy**.

The variables that load highly on factor - PA2:

1. Sugar
2. Salt
3. Sweet
4. Calories
5. Process

All these items are related to the taste aspects of Cereal, so we can label this factor as **Tasty**.

The variables that load highly on factor - PA3:

1. Kids
2. Family
3. Economical

All these items are related to the Family, so we can label this factor as **For Family**.

The variables that load highly on factor - PA4:

1. Soggy
2. Crisp
3. Treat

All these items are related to the Crunchy aspects of Cereal, so we can label this factor as **Crunchy**.

The variables that load highly on factor - PA5:

1. Plain
2. Fruit

All these items are related to the Fruits aspects of Cereal, so we can label this factor as **Fruits**.

(G) Examine correlations among factors

```
##  With factor correlations of
##       PA1   PA2   PA3   PA4   PA5
## PA1  1.00 -0.19  0.15  0.15 -0.29
## PA2 -0.19  1.00  0.05  0.19 -0.21
## PA3  0.15  0.05  1.00  0.16 -0.10
## PA4  0.15  0.19  0.16  1.00 -0.27
## PA5 -0.29 -0.21 -0.10 -0.27  1.00
```

Observation

We notice that our factors are least correlated at 5% and recall our choice of oblique rotation allowed for the recognition of good relationship; hence our assumption of correlation between components while running oblique rotation is violated.

(H) Analyze internal reliability

Crombach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. Cronbach's alpha is computed by correating the score for each scale item with the total score for each observation (usually individual survey responds or test takers), and then comparing that to the variance for all individual item scores. Cronbach's alpha is a function of the number of items in a test, the average covariance between pairs of items, and the variance of the total score.

```
## $sample.size
## [1] 235
##
## $number.of.items
## [1] 25
##
## $alpha
## [1] 0.7276847
```

Interpretation of Cronbach's alpha *http://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/*

The alpha coefficient of reliability ranges from 0 to 1 in providing this overall assessment of a measure's reliability. If all of the scale items are entirely independent from one another (i.e., are not correlated or share no covariance), the alpha = 0; and, if all of the items have high covariances, then alpha will approach 1 as more the items in the scale approaches infinity.

A **good** alpha coefficient depends on your theoretical knowledge of the scale in question. Many methodologies recommend a minimum alpha coefficient between 0.65 and 0.80; alpha coefficients less than 0.50 are usually unacceptable.

Conclusion

Overall assessment of this measure's reliability is **good** since the absolute value of the alpha coefficient is above 0.65.

2) Problem - Leslie Salt Data Set (10 points)

In 1968, the city of Mountain View, California, began the necessary legal proceedings to acquire a parcel of land owned by the Leslie Sal Company. The Leslie property contained 246.8 acres and was located right on the San Francisco Bay. The land had been used for salt evaporation and had an elevation of exactly sea level. However, the property was diked so that the waters from the bay park were kept out. The city of Mountain View intended to fill the property and use it for a city park.

Ultimately, it fell into the courts to determine a fair market value for the property. Appraisers were hired, but what made the processes difficult was that there were few sales of byland property and none of them corresponded exactly to the characteristics of the Leslie property. The experts involved decided to build a regression model to better understand the factors that might influence market valuation. They collected data on 31 byland properties that were sold during the previous 10 years. In addition to the transaction price for each property, they collected data oina large number of other factors, including size, time of sale, elevation, location, and access to sewers. A listing of these data, including only those variables deemed relevant for this exercise. A description of the variables is provided below:

Variable name	Description
Price	Sales price in $000 per acre
County	San Mateo=0, Santa Clara =1
Size	Size of the property in acres
Elevation	Average Elevation in foot above sea level
Sewer	Distance (in feet) to nearest sewer connection
Date	Date of sale counting backward from current time (in months)
Flood	Subject to flooding by tidal action =1; otherwise =0
Distance	Distance in miles from Leslie Property (in almost all cases, this is toward San Francisco

Solution

Discuss and Answer the following questions:

1. What is the nature of each of the variables? Which variable is dependent variable and what are the independent variables in the model?

```
library(corrgram)

## Warning: package 'corrgram' was built under R version 3.3.3

data <- read.csv('D:/GL/AS/Answers/data/Leslie_Salt.csv',header=T)
names(data)

## [1] "Price"     "County"    "Size"      "Elevation" "Sewer"     "Date"

## [7] "Flood"     "Distance"
```

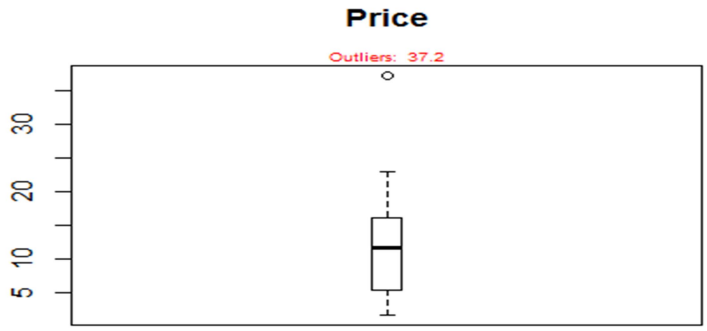
Answer

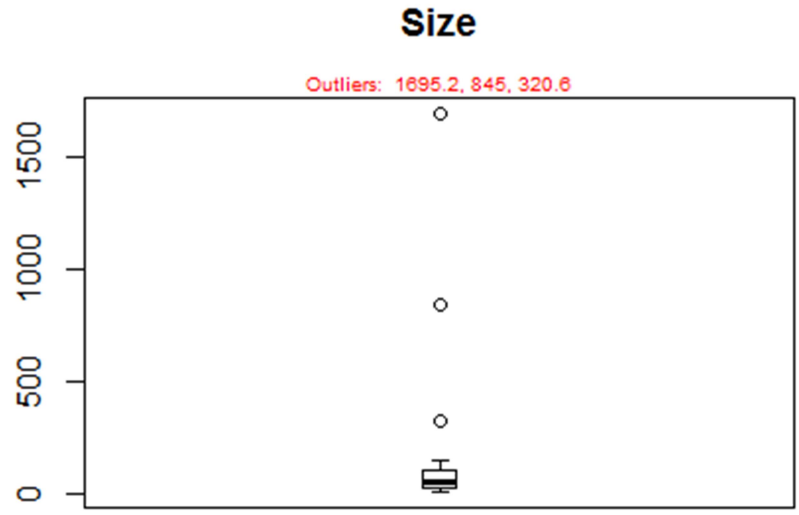
S.No	Variable	Nature of Variable	Dependent / Independent
1	Price	Continuous Numeric	Dependent
2	County	Categorical	Independent
3	Size	Continuous Numeric	Independent
4	Elevation	Continuous Numeric	Independent
5	Sewer	Continuous Numeric	Independent
6	Date	Continuous Numeric	Independent
7	Flood	Categorical	Independent
8	Distance	Continuous Numeric	Independent

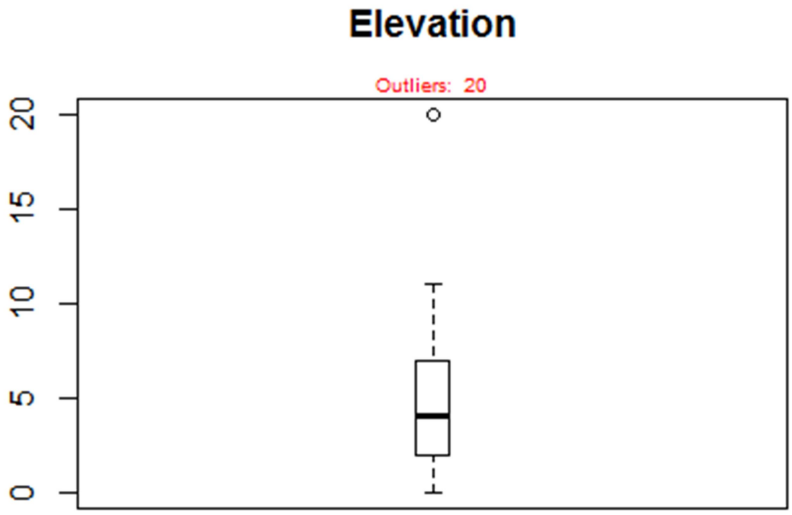
1a. Detect outiers

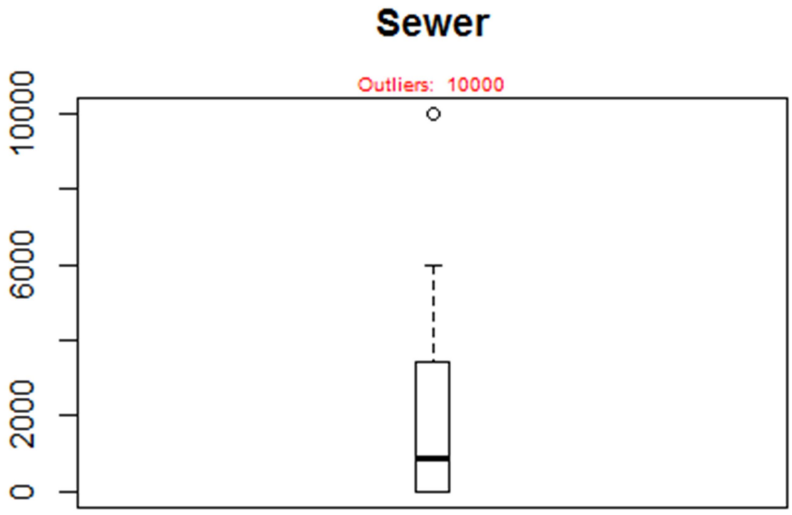
```
showoutlier_values <- function(x, mainlab) {

   outlier_values <- boxplot.stats(x)$out  # outlier values.
##
   boxplot(x, main=mainlab, boxwex=0.1)
   mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=0.6,
col="red")
}

is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 *
 IQR(x))
}

cols <- c("Price",  "Size", "Elevation", "Sewer", "Date", "Distance")
```

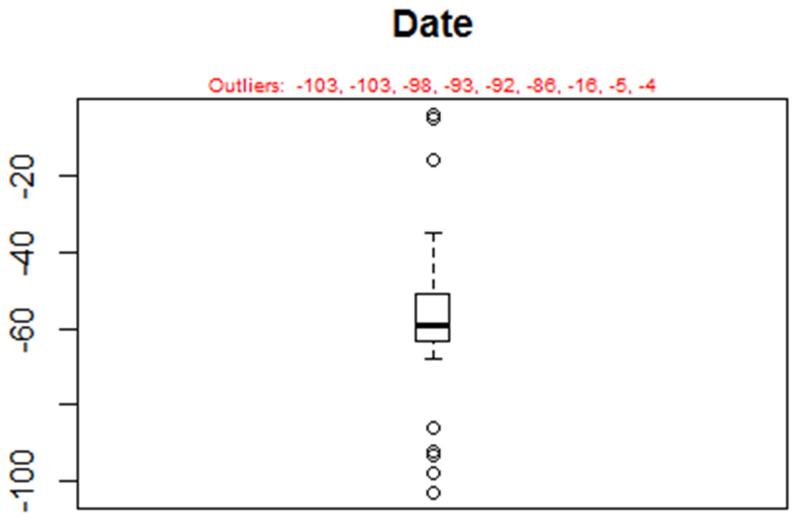
```
for (i in 1:length(cols)) {

  ### Box plot for each numeric variables

  showoutlier_values(data[,cols[i]],cols[i])
}
```

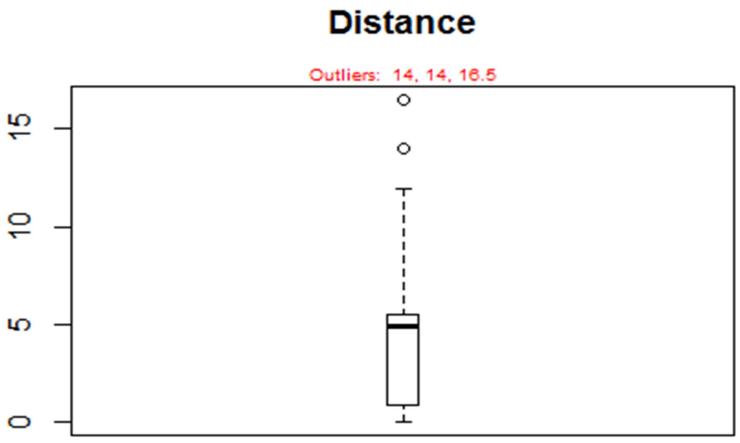





Elevation
Outliers: 20
0
5
10
15
20


Sewer
Outliers: 10000
0
2000
6000
10000





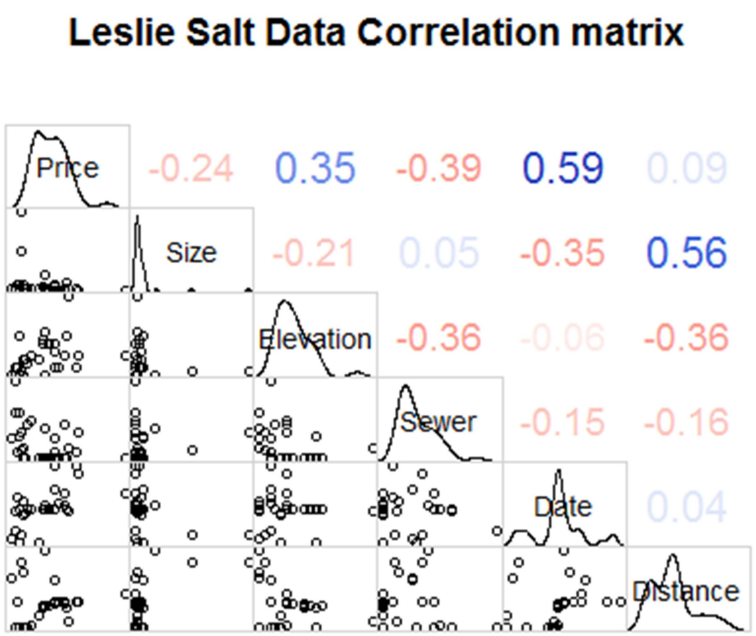
Observation: Outliers exist for all almost all predictor variables.

1 b) Correlation matrix

```
new_data <- data[,-c(2,7)]

library(corrgram)

round(cor(new_data, use="complete.obs", method="kendall") ,2)
```

```
##           Price  Size Elevation Sewer  Date Distance
## Price      1.00 -0.16      0.36 -0.29  0.51     0.26
## Size      -0.16  1.00     -0.02 -0.07 -0.13     0.14
## Elevation  0.36 -0.02      1.00 -0.40 -0.01    -0.30
## Sewer     -0.29 -0.07     -0.40  1.00 -0.07     0.01
## Date       0.51 -0.13     -0.01 -0.07  1.00     0.40
## Distance   0.26  0.14     -0.30  0.01  0.40     1.00
```

```
corrgram(new_data, main="Leslie Salt Data Correlation matrix",
         lower.panel=panel.pts, upper.panel=panel.cor,
         diag.panel=panel.density)
```



Observation

- The correlation between Price and Distance is very less 0.09.
- Correlation between Price and Size is negative -0.24.
- Correlation between Price and Sewer is negative -0.39.

1 c) Replace outliers with NA

```
new <- apply( new_data,2,function(x) ifelse(is_outlier(x),NA,x))
#print(new['Price'])

### Get row numbers of NAs

rn <- which(is.na(new['Price']) | is.na(new['Size']) |
            is.na(new['Elevation']) | is.na(new['Sewer']) |
            is.na(new['Date']) | is.na(new['Distance']))
print(rn)

##  [1]  1  2  3  4  5  6  9 21 26 29 30 31

print(new[-rn,])

##       Price  Size Elevation Sewer Date Distance
##  [1,]   5.7 105.9         4     0  -68      0.0
##  [2,]   6.2  56.6         4     0  -64      0.0
##  [3,]   3.2  22.1         0  6000  -62      0.0
##  [4,]   4.7  22.1         0  6000  -61      0.0
##  [5,]   6.9  27.7         3  4500  -60      0.0
##  [6,]   8.1  18.6         5  5000  -59      0.5
##  [7,]  11.6  69.9         8     0  -59      4.4
##  [8,]  19.3 145.7        10     0  -59      4.2
##  [9,]  11.7  77.2         9     0  -59      4.5
## [10,]  13.3  26.2         8     0  -59      4.7
## [11,]  15.1 102.3         6     0  -59      4.9
## [12,]  12.4  49.5        11     0  -59      4.6
## [13,]  15.3  12.2         8     0  -59      5.0
## [14,]  18.1   9.9         5     0  -54      5.2
## [15,]  16.8  15.3         2     0  -53      5.5
## [16,]   5.9  55.2         0  1320  -49     11.9
## [17,]   4.0 116.2         2   900  -45      5.5
## [18,]  18.2  23.4         5  4420  -39      5.5
## [19,]  15.1 132.8         2  2640  -35     10.2

new_data <- data[-rn,-c(2,7)]
```

Observation

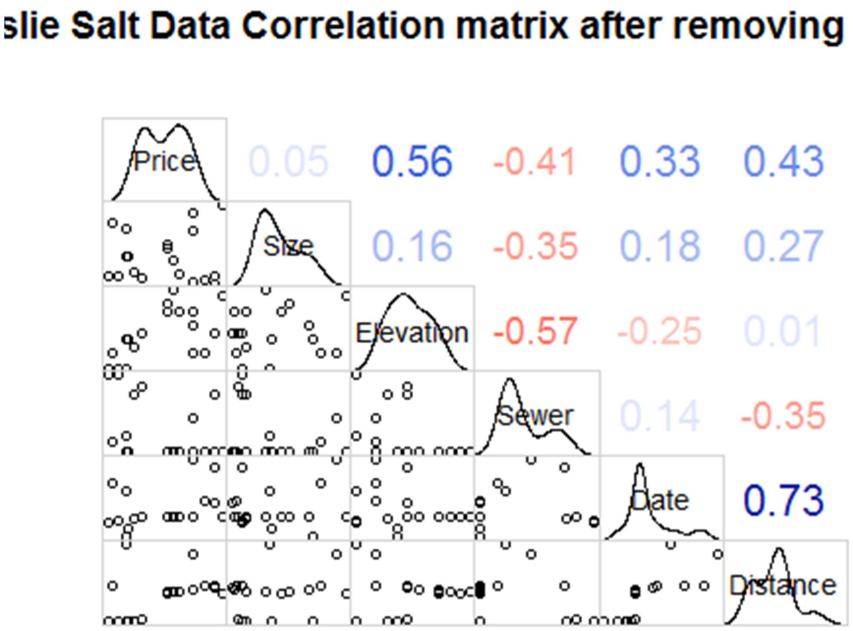
We have detected and removed outliers from the data frame

1 d) Correlation matrix after removing outliers

```
round(cor(new_data, use="complete.obs", method="kendall") ,2)
```

```
##           Price  Size Elevation Sewer  Date Distance
## Price      1.00 -0.09      0.42 -0.34  0.40     0.45
## Size      -0.09  1.00      0.09 -0.16  0.03     0.00
## Elevation  0.42  0.09      1.00 -0.53 -0.05    -0.08
## Sewer     -0.34 -0.16     -0.53  1.00  0.05    -0.07
## Date       0.40  0.03     -0.05  0.05  1.00     0.82
## Distance   0.45  0.00     -0.08 -0.07  0.82     1.00
```

```
corrgram(new_data, main="Leslie Salt Data Correlation matrix after removing o
utliers",
         lower.panel=panel.pts, upper.panel=panel.cor,
         diag.panel=panel.density)
```



Observation

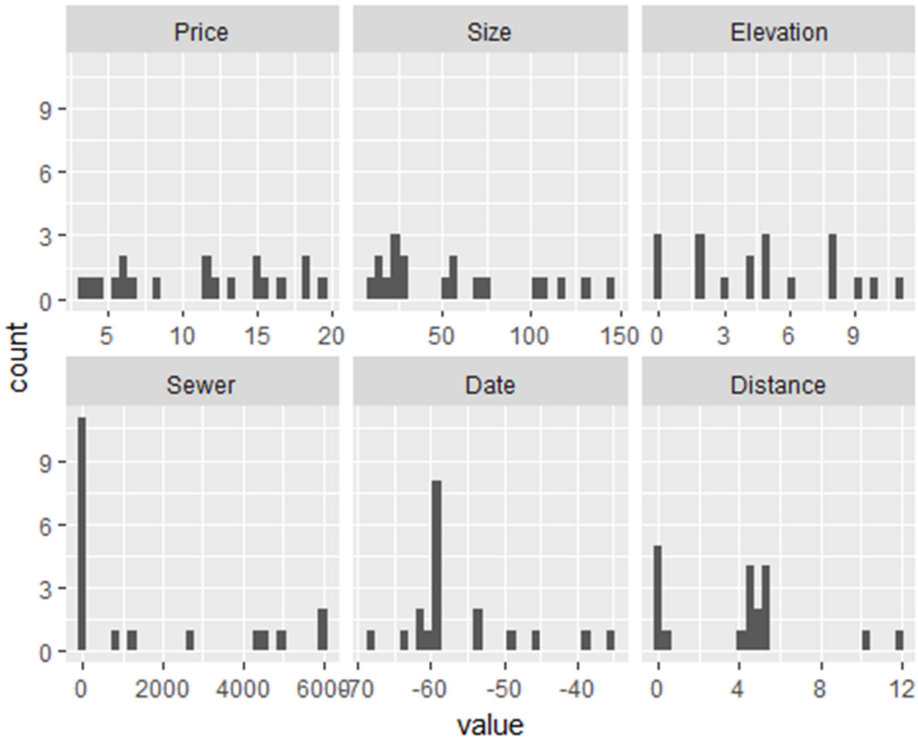
- Correlation between Price and Size is 0.05, which is very less.
- Correlation between Price and Sewer is negative -0.41.
- Correlation between Price and Distance is improved and it is 0.43.

2. Check whether the variables require any transformation individually

If a measurement variable does not fit a normal distribution or has greatly different standard deviations in different groups, you should try a data transformation.

Let us find how the variables are distributed by plotting histograms.

```
library(reshape2)
library(ggplot2)
d <- melt(new_data)

## No id variables; using all as measure variables

ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
### Check mean and median for these variables

apply(new_data,2, function(x) summary(x))
```

```
##         Price   Size Elevation Sewer   Date Distance
## Min.     3.20   9.90     0.000     0 -68.00    0.000
## 1st Qu.  6.05  22.10     2.000     0 -59.50    0.250
## Median  11.70  49.50     5.000     0 -59.00    4.600
## Mean    11.14  57.31     4.842  1620 -55.89    4.032
## 3rd Qu. 15.20  89.75     8.000  3530 -53.50    5.350
## Max.    19.30 145.70    11.000  6000 -35.00   11.900
```

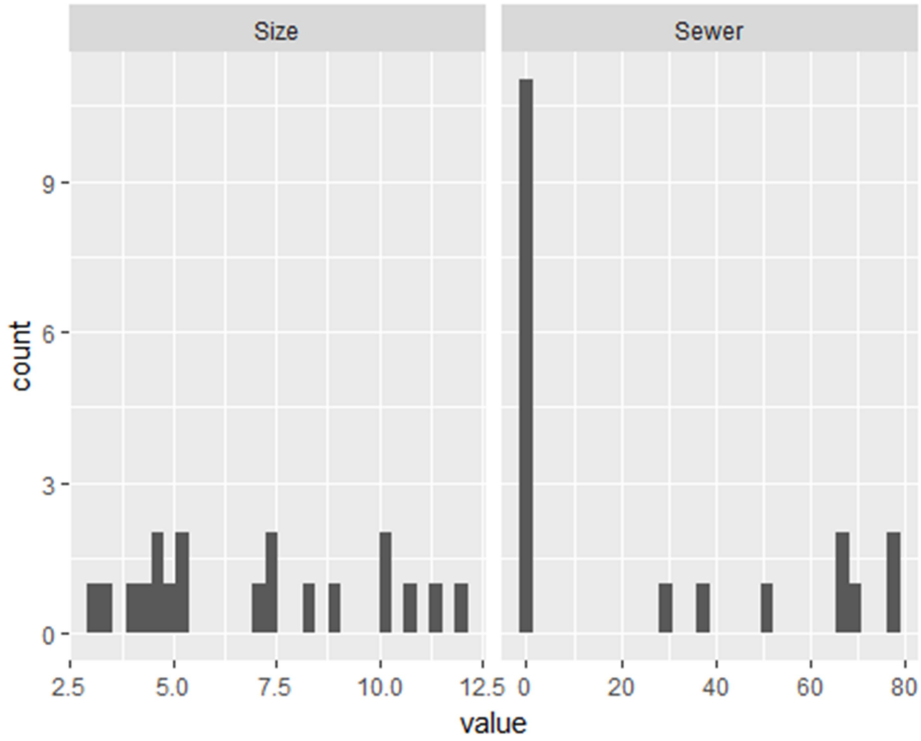
Observation

We will apply square root transformation for Size and Sewer and again check normalcy.

```
Size     <- sqrt(new_data[,2])
Sewer    <- sqrt(new_data[,4])
dt1      <- data.frame(Size,Sewer)

d <- melt(dt1)

## No id variables; using all as measure variables
```

```
ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
### Check mean and median for these variables

apply(dt1,2, function(x) summary(x))
```

```
##           Size Sewer
## Min.     3.146  0.00
## 1st Qu.  4.701  0.00
## Median   7.036  0.00
## Mean     7.021 25.10
## 3rd Qu.  9.450 58.93
## Max.    12.070 77.46
```

Observation

We find that after square root transformation, Sewer and Size is almost normally distributed.

3. Set up a regression equation, run the model and discuss your results

```
Price     <- new_data[,1]
County    <- data[-rn,2]
Size      <- dt1$Size
Elevation <- new_data[,3]
Sewer     <- dt1$Sewer
Date      <- new_data[,5]
Flood     <- data[-rn,7]
Distance  <- new_data[,6]

dt1 <- data.frame(Price, County, Size, Elevation, Sewer, Date, Flood, Distanc
e)

fit <- lm(dt1$Price ~ ., data = dt1)
fit_aov <- aov(fit)
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = dt1$Price ~ ., data = dt1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9467 -1.3437  0.1552  0.9017  5.3863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.67749   11.99458   3.058   0.0109 *
## County       6.31135    5.76808   1.094   0.2973
## Size        -0.17038    0.27190  -0.627   0.5437
## Elevation    0.40447    0.26818   1.508   0.1597
## Sewer       -0.04743    0.03407  -1.392   0.1914
## Date         0.55688    0.25095   2.219   0.0484 *
## Flood       -6.54737    3.88284  -1.686   0.1199
## Distance     0.42675    0.39626   1.077   0.3046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 11 degrees of freedom
## Multiple R-squared:  0.8459, Adjusted R-squared:  0.7478
## F-statistic: 8.626 on 7 and 11 DF,  p-value: 0.001015
```

```
summary(fit_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## County       1   0.57    0.57   0.079 0.783277
## Size         1   0.08    0.08   0.011 0.919152
## Elevation    1 204.49  204.49  28.261 0.000246 ***
## Sewer        1  21.40   21.40   2.957 0.113480
## Date         1 187.64  187.64  25.933 0.000348 ***
## Flood        1  14.36   14.36   1.984 0.186570
## Distance     1   8.39    8.39   1.160 0.304559
## Residuals   11  79.59    7.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observation

From the above regression, we find that the Multiple R squared is **84.59 %** and adjusted R^2 is **74.78%**.
All the independent variables except Date are not significant.
From the ANOVA table, we observe that the regression equation is

Price = 36.6775 + 6.3114 * County - 0.17038 * Size + 0.40447 * Elevation - 0.04743 * Sewer + 0.55688 * Date - 6.54737 * Flood + 0.42675 * Distance

Price = 36.6775 + 6.3114 * County - 0.17038 * Size + 0.40447 * Elevation - 0.04743 * Sewer + 0.55688 * Date - 6.54737 * Flood + 0.42675 * Distance

- ***County and Flood are categorical variables***

a. The equation shows that the coefficient for County is 6.3114.
b. So compared to County, San Mateo (base-lined at 0), the houses in Santa Clara, we would expect Price to increase by $6311.4, on average, assuming if it were the only variable in the model.
c. The equation shows that the coefficient for Flood is - 6.54737.
d. So compared to Flood - flooding by tidal action - No - (base-lined at 0), the houses in the comparison group (flooding by tidal action - 1 Yes), we would expect Price to **decrease** by $6547.37, on average, assuming if it were the only variable in the model.

- ***Size, Elevation, Sewer, Date, Distance are Continuous Numeric variables.***

a. The equation shows that the coefficient for Size, in acres is - 0.17038 acres .
b. The coefficient indicates that for every additional acre in Size you can expect Price to decrease by an average of $170.38, assuming if it were the only variable in the model.
c. The equation shows that the coefficient for Elevation in foot is 0.40447 ft.
d. The coefficient indicates that for every additional foot in Elevation you can expect Price to increase by an average of $404.47, assuming if it were the only variable in the model.
e. The equation shows that the coefficient for Sewer, in foot is - 0.04743 ft.
f. The coefficient indicates that for every foot increase in Sewer (Distance (in feet) to nearest sewer connection), you can expect Price to decrease by an average of $47.43, assuming if it were the only variable in the model.
g. The equation shows that the coefficient for Date, in months is 0.55688.
h. The coefficient indicates that for every increase in month in Date field (Date of sale counting backward from current time (in months)), you can expect Price to increase by an average of $556.88, assuming if it were the only variable in the model.
i. The equation shows that the coefficient for Distance, in miles is 0.42675.
j. The coefficient indicates that for every increase in miles in Distance field (Distance in miles from Leslie Property), you can expect Price to increase by an average of $426.75, assuming if it were the only variable in the model.

3) Problem 3 - All Greens Franchise (10 points)

Explain the importance of X2, X3, X4, X5, X6 on Annual Net Sales, X1.

The data (X1, X2, X3, X4, X5, X6) are for each franchise store.

1) X1 = annual net sales/$1000
2) X2 = number sq. ft./1000
3) X3 = inventory/$1000
4) X4 = amount spent on advertising/$1000
5) X5 = size of sales district/1000 families
6) X6 = number of competing stores in district

Solution

Since the regular coefficients use different scales, we cannot compare them directly to determine variable importance. We use standardized coefficients for all continuous predictors so that we can compare them.

Ref: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models

```
###   Data ----------------------------------------------------------------
X1 <- c(231,156,10,519,437,487,299,195,20,68,570,428,464,15,65,98,398,161,397
,497,528,99,0.5,347,341,507,400)
X2 <- c(3,2.2,0.5,5.5,4.4,4.8,3.1,2.5,1.2,0.6,5.4,4.2,4.7,0.6,1.2,1.6,4.3,2.6
,3.8,5.3,5.6,0.8,1.1,3.6,3.5,5.1,8.6)
X3 <- c(294,232,149,600,567,571,512,347,212,102,788,577,535,163,168,151,342,1
96,453,518,615,278,142,461,382,590,517)
X4 <- c(8.2,6.9,3,12,10.6,11.8,8.1,7.7,3.3,4.9,17.4,10.5,11.3,2.5,4.7,4.6,5.5
,7.2,10.4,11.5,12.3,2.8,3.1,9.6,9.8,12,7)
X5 <- c(8.2,4.1,4.3,16.1,14.1,12.7,10.1,8.4,2.1,4.7,12.3,14,15,2.5,3.3,2.7,16
,6.3,13.9,16.3,16,6.5,1.6,11.3,11.5,15.7,12)
X6 <- c(11,12,15,1,5,4,10,12,15,8,1,7,3,14,11,10,4,13,7,1,0,14,12,6,5,0,8)
###  ---------------------------------------------------------------------

SX1 <- scale(X1)
SX2 <- scale(X2)
SX3 <- scale(X3)
SX4 <- scale(X4)
SX5 <- scale(X5)
SX6 <- scale(X6)

### Regular fit - non-standardized
```

```
lm_fit1 <-  lm(X1 ~ X2 + X3 + X4 + X5 + X6)

summary(lm_fit1)

##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4 + X5 + X6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.338  -9.699  -4.496   4.040  41.139
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.85941   30.15023  -0.626 0.538372
## X2           16.20157    3.54444   4.571 0.000166 ***
## X3            0.17464    0.05761   3.032 0.006347 **
## X4           11.52627    2.53210   4.552 0.000174 ***
## X5           13.58031    1.77046   7.671 1.61e-07 ***
## X6           -5.31097    1.70543  -3.114 0.005249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.65 on 21 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9916
## F-statistic: 611.6 on 5 and 21 DF,  p-value: < 2.2e-16
```

```
### Standardized fit after scaling

lm_fit2 <-  lm(SX1 ~ SX2 + SX3 + SX4 + SX5 + SX6)

summary(lm_fit2)

##
## Call:
## lm(formula = SX1 ~ SX2 + SX3 + SX4 + SX5 + SX6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13713 -0.05050 -0.02341  0.02103  0.21420
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.499e-17  1.768e-02   0.000 1.000000
## SX2          1.696e-01  3.711e-02   4.571 0.000166 ***
## SX3          1.738e-01  5.734e-02   3.032 0.006347 **
## SX4          2.265e-01  4.976e-02   4.552 0.000174 ***
## SX5          3.634e-01  4.738e-02   7.671 1.61e-07 ***
## SX6         -1.354e-01  4.347e-02  -3.114 0.005249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09189 on 21 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9916
## F-statistic: 611.6 on 5 and 21 DF,  p-value: < 2.2e-16
```

Observations

1) The standardized coefficients are given below in the order of their importance on **Annual net sales/$1000 (X1)**:

- Size of sales district/1000 families (X5)
- Amount spent on advertizing/$1000 (X4)
- Number sq. ft./1000 (X2)
- Number of competing stores in district (X6)
- Inventory/$1000 (X3)

All these variables are significant in predicting **Annual net sales/$1000 (X1)** as shown by their respective p-values.

2) Using the non-standardized coefficients, we get different result!!