Advanced Statistics: Group Assignment

Factor Analysis





**1.Factor Analysis ( Cereal Data Factor Analysis):**

**Reading the data**

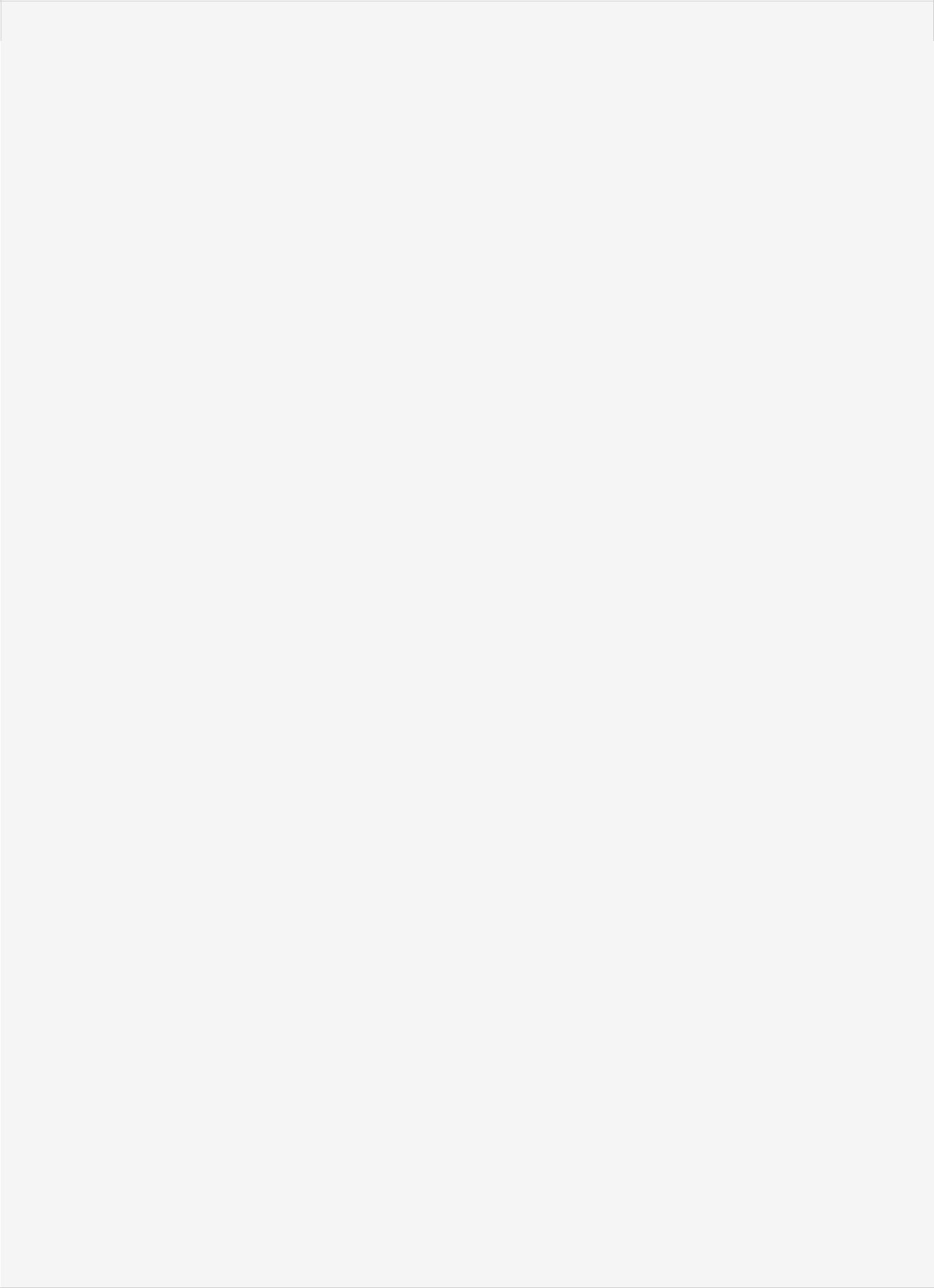
cereal <- read.csv("D:/great lakes work/AS./data/cereal.csv")

**Exploring the data**

dim(cereal)

## [1] 235 26

Summarize the data to see if the data is as per the expectations.



summary(cereal)

## Cereals Filling Natural Fibre

* CornFlakes :27 Min. :1.000 Min. :1.000 Min. :1.000
* Weetabix :27 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000
* Vitabrit :25 Median :4.000 Median :4.000 Median :4.000

## NutriGrain :24 Mean :3.881 Mean :3.783 Mean :3.528

* SpecialK :23 3rd Qu.:4.500 3rd Qu.:4.000 3rd Qu.:4.000
* RiceBubbles:21 Max. :5.000 Max. :5.000 Max. :5.000
* (Other) :88

## Sweet Easy Salt Satisfying

* Min. :1.000 Min. :1.000 Min. :1.000 Min. :2.000
* 1st Qu.:2.000 1st Qu.:4.000 1st Qu.:1.000 1st Qu.:3.000
* Median :2.000 Median :5.000 Median :2.000 Median :4.000
* Mean :2.506 Mean :4.532 Mean :1.991 Mean :4.004
* 3rd Qu.:3.000 3rd Qu.:5.000 3rd Qu.:3.000 3rd Qu.:5.000
* Max. :5.000 Max. :6.000 Max. :4.000 Max. :6.000

##

## Energy Fun Kids Soggy

* Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000
* 1st Qu.:3.000 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:1.000
* Median :4.000 Median :2.000 Median :4.000 Median :2.000
* Mean :3.643 Mean :2.617 Mean :3.843 Mean :2.255
* 3rd Qu.:4.000 3rd Qu.:3.000 3rd Qu.:5.000 3rd Qu.:3.000
* Max. :5.000 Max. :5.000 Max. :6.000 Max. :5.000

##

## Economical Health Family Calories

* Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000
* 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:2.000
* Median :3.000 Median :4.000 Median :4.000 Median :3.000
* Mean :3.217 Mean :3.809 Mean :3.877 Mean :2.702
* 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:5.000 3rd Qu.:3.000
* Max. :5.000 Max. :5.000 Max. :6.000 Max. :5.000

##

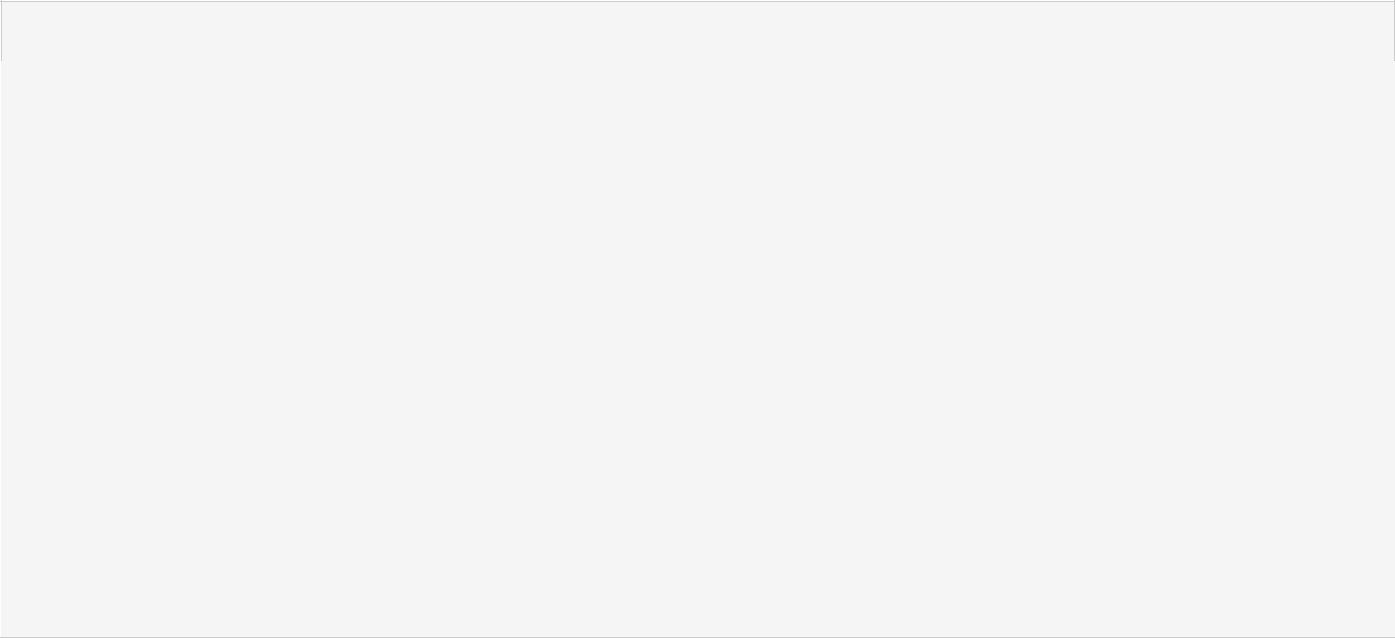
## Plain Crisp Regular Sugar

* Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000
* 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:1.000
* Median :2.000 Median :3.000 Median :3.000 Median :2.000
* Mean :2.268 Mean :3.204 Mean :3.072 Mean :2.145
* 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:3.000
* Max. :5.000 Max. :6.000 Max. :5.000 Max. :5.000

##

## Fruit Process Quality Treat

## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00



* 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:2.00
* Median :1.000 Median :3.000 Median :4.000 Median :3.00
* Mean :1.694 Mean :2.936 Mean :3.694 Mean :2.63
* 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:3.00
* Max. :5.000 Max. :6.000 Max. :5.000 Max. :6.00

##

## Boring Nutritious

* Min. :1.00 Min. :1.000
* 1st Qu.:1.00 1st Qu.:3.000
* Median :2.00 Median :4.000
* Mean :1.83 Mean :3.664
* 3rd Qu.:2.00 3rd Qu.:4.000
* Max. :5.00 Max. :5.000

We can see that there are values of 6 which is not expected; the max. of the scale is 5. Let’s repla



cereal[cereal==6] <- 5

Seven 6s replaced by 5.

Recode the scores on negative variables like Soggy, Boring etc.

1.converting the negative score variables into reverse for example

example: positive scores:1 2 3 4 5

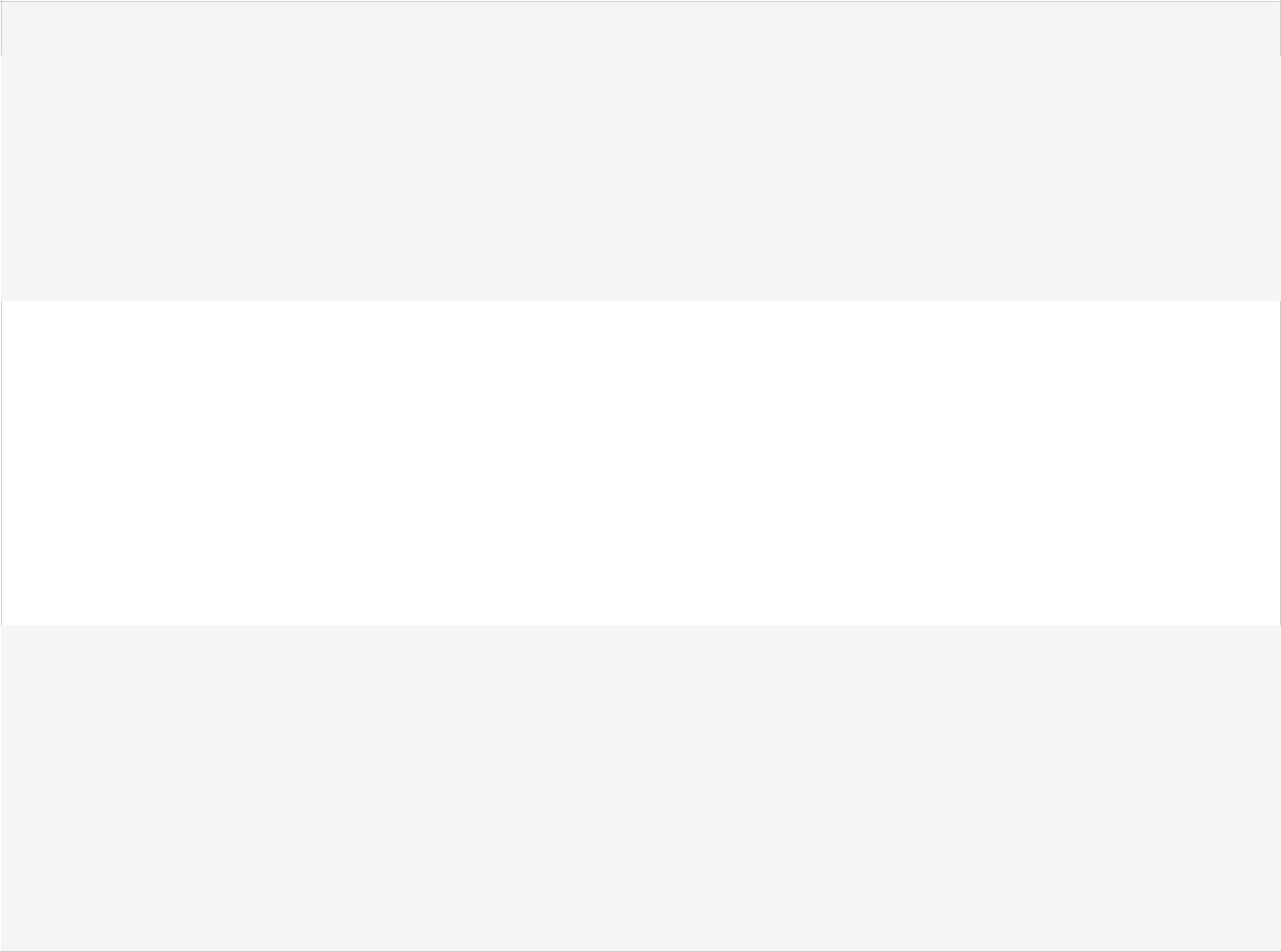
Negative scores:5 4 3 2 1



cereal[,c(12,25)] <- 6 - cereal[,c(12,25)]

Before constructing the factor analysis we need to determine whether we have a large enough sample to perform Factor Analysis or Principal Component Analysis and if the dimensionality reduction is a possibility at all(correlation adequacy). to check this below tests are performed

* KMO test of sampling adequacy.(large enough sample?)
* Perform the Bartlett Test of Sphericity (dimensionality reduction possible?)



library(psych)

cerealKMO <- KMO(cereal[,-1])

cerealKMO

* Kaiser-Meyer-Olkin factor adequacy
* Call: KMO(r = cereal[, -1])
* 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 | 0.67 | 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 | 0.80 | 0.91 | 0.88 | 0.87 |
|  |  |  |  |  |  |  |

* Nutritious
* 0.92

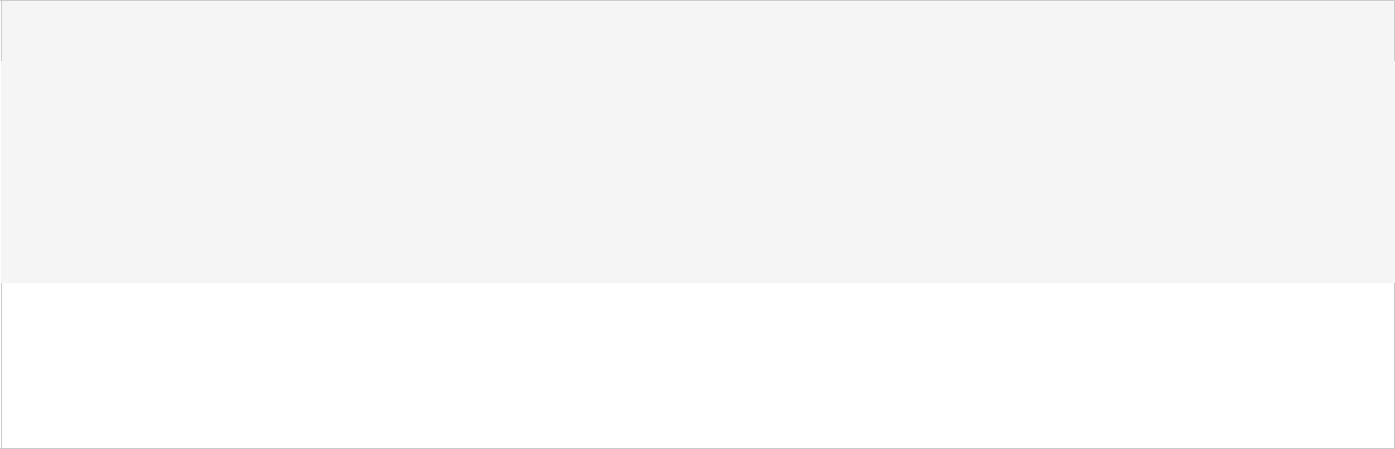
cerealMatrix <- cor(cereal[,-1])

cerealMatrix <- round(cerealMatrix, 2)

cerealBartlett <- cortest.bartlett(cerealMatrix, n = nrow(cereal))

cerealBartlett

* $chisq
* [1] 2878.65



##

* $p.value
* [1] 0

##

* $df
* [1] 300

**Result Interpretation:**

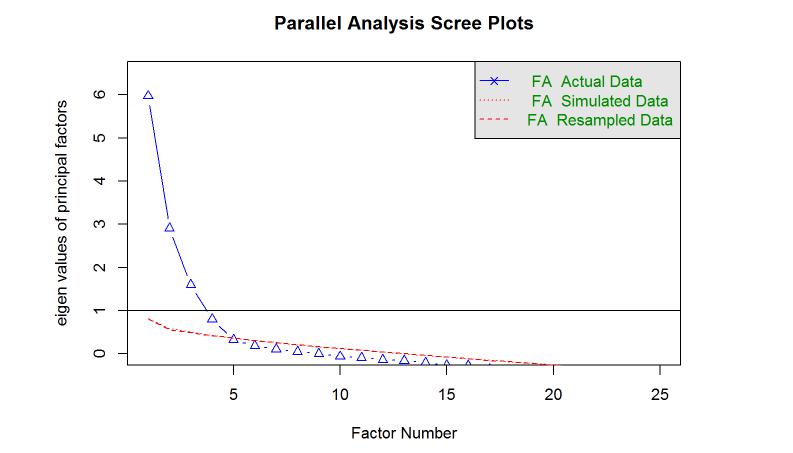
KMO Test - The KMO test yields a degree of common variance meritorious. Thus, our sample is large enough for factor analysis or principal component analysis.

Bartlett Test of Sphericity - The p-value is <.001, thus the null hypothesis is rejected (The null hypothesis is that the correlation matrix is an identity matrix i.e. there is no scope for dimensionality reduction.). Thus, the dimensionality reduction is a possibility using PCA/FA.

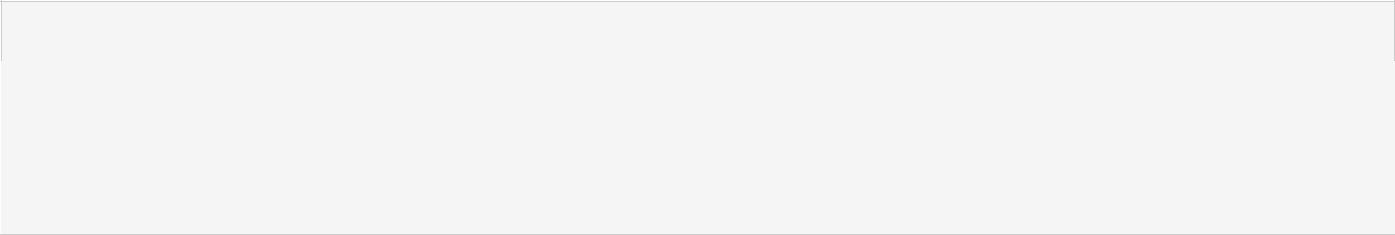
**To decide no: of factors:**



numFactors <- fa.parallel(cereal[,-1], fm="ml", fa="fa")



**result interpretation:**



## Parallel analysis suggests that the number of factors = 4 and the number of components = NA

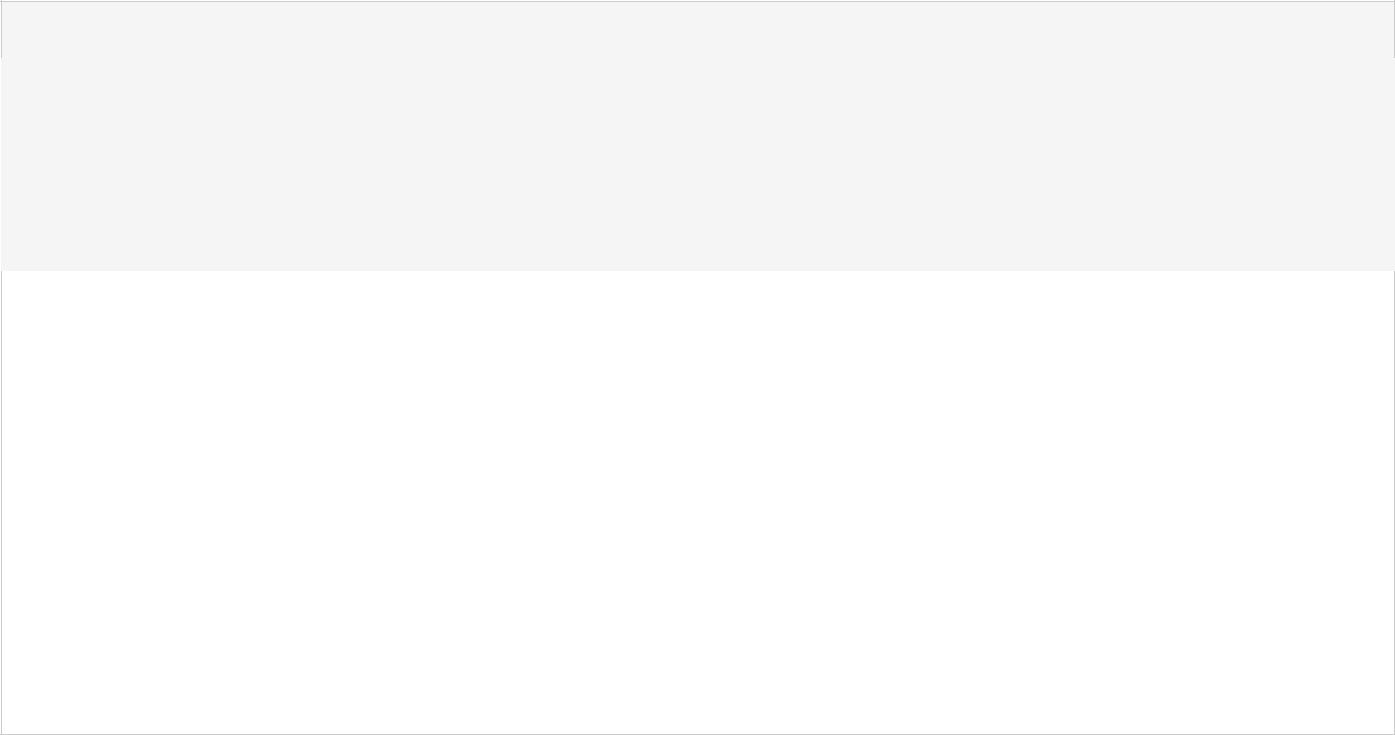
sum(numFactors$fa.values>1.0) ##old Kaiser criterion

## [1] 3

sum(numFactors$fa.values>0.7) ##new Kaiser criterion

## [1] 4

Parallel analysis helps you to decide how many factors to retain. scree plot suggests that 4 factors should be retained. Next step is to create four factor model



fit <- fa(cereal[,-1], nfactors=4, fm="ml", rotate="oblimin")

* Loading required namespace: GPArotation fit
* Factor Analysis using method = ml
* Call: fa(r = cereal[, -1], nfactors = 4, rotate = "oblimin", fm = "ml")
* Standardized loadings (pattern matrix) based upon correlation matrix

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ## | ML1 | ML2 | ML3 | ML4 | h2 | u2 | com |
| ## Filling | 0.70 | 0.17 | 0.14 | 0.06 | 0.56 | 0.44 | 1.2 |
| ## Natural | 0.76 | -0.11 | 0.00 | -0.03 | 0.61 0.39 1.0 | | |
| ## Fibre | 0.86 | 0.02 | -0.16 -0.08 | | 0.69 0.31 1.1 | | |
| ## Sweet | 0.07 | 0.71 | 0.03 | 0.22 | 0.65 | 0.35 | 1.2 |
| ## Easy | 0.25 | 0.09 | 0.27 | 0.01 | 0.15 | 0.85 | 2.2 |
| ## Salt | 0.02 | 0.73 | 0.01 | -0.22 | 0.49 0.51 1.2 | | |
| ## Satisfying | 0.61 | 0.12 | 0.32 | 0.10 | 0.57 | 0.43 | 1.7 |
| ## Energy | 0.64 | 0.13 | 0.10 | 0.14 | 0.51 | 0.49 | 1.2 |
| ## Fun | 0.04 | 0.11 | 0.32 | 0.50 | 0.47 | 0.53 | 1.8 |
| ## Kids | -0.04 | 0.03 | 0.88 | -0.02 | 0.77 0.23 1.0 | | |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | Soggy |  | -0.15 | -0.09 | -0.17 0.52 0.23 | | 0.77 | 1.5 |  |  |  |  |
| ## | Economical | | 0.10 | -0.24 | 0.41 -0.21 0.28 | | 0.72 | 2.4 |  |  |  |  |
| ## | Health | | 0.84 | -0.17 | -0.04 -0.01 0.78 | | 0.22 | 1.1 |  |  |  |  |
| ## | Family | | 0.02 | -0.07 | 0.79 0.08 0.65 | | 0.35 | 1.0 |  |  |  |  |
| ## | Calories | | -0.09 | 0.61 | -0.02 0.03 0.41 | | 0.59 | 1.1 |  |  |  |  |
| ## | Plain |  | 0.00 | 0.01 | 0.15 -0.69 0.45 | | 0.55 | 1.1 |  |  |  |  |
| ## | Crisp |  | -0.01 | 0.10 | 0.27 0.42 0.32 | | 0.68 | 1.9 |  |  |  |  |
| ## | Regular | | 0.65 | 0.02 | -0.09 -0.01 0.41 | | 0.59 | 1.0 |  |  |  |  |
| ## | Sugar |  | -0.14 | 0.82 | -0.07 0.03 0.74 | | 0.26 | 1.1 |  |  |  |  |
| ## | Fruit |  | 0.31 | 0.18 | -0.35 0.41 0.44 | | 0.56 | 3.3 |  |  |  |  |
| ## | Process | | -0.18 | 0.37 | 0.04 -0.18 0.21 | | 0.79 | 2.0 |  |  |  |  |
| ## | Quality | | 0.63 | -0.18 | 0.11 0.15 0.56 | | 0.44 | 1.3 |  |  |  |  |
| ## | Treat |  | 0.13 | 0.16 | 0.21 0.59 0.58 | | 0.42 | 1.5 |  |  |  |  |
|  |  | |  |  |  | |  |  |  |  |  |  |
| ## | Boring | | 0.04 | -0.12 | 0.15 0.53 0.33 | | 0.67 | 1.3 |  |  |  |  |
| ## Nutritious | | | 0.85 | -0.05 | -0.02 -0.02 0.73 | | 0.27 | 1.0 |  |  |  |  |
| ## |  |  |  |  |  |  |  |  |  |  |  |  |
| ## |  |  |  |  | ML1 ML2 ML3 | ML4 | |  |  |  |  |  |
| ## | SS loadings | |  |  | 5.27 2.61 2.30 2.40 | | |  |  |  |  |  |
| ## | Proportion Var | | |  | 0.21 0.10 0.09 0.10 | | |  |  |  |  |  |
| ## | Cumulative Var | | |  | 0.21 0.32 0.41 0.50 | | |  |  |  |  |  |
| ## | Proportion Explained | | | | 0.42 0.21 0.18 0.19 | | |  |  |  |  |  |
| ## Cumulative Proportion 0.42 0.63 0.81 1.00 | | | | | | | |  |  |  |  |  |
| ## |  |  |  |  |  |  |  |  |  |  |  |  |
| ## With factor correlations of | | | | | |  |  |  |  |  |  |  |
| ## |  | ML1 | ML2 | ML3 | ML4 |  |  |  |  |  |  |  |
| ## | ML1 | 1.00 -0.18 0.11 0.31 | | | |  |  |  |  |  |  |  |
| ## | ML2 -0.18 | | 1.00 0.03 0.28 | | |  |  |  |  |  |  |  |
| ## | ML3 | 0.11 | 0.03 1.00 0.18 | | |  |  |  |  |  |  |  |
| ## ML4 | | 0.31 | 0.28 0.18 1.00 | | |  |  |  |  |  |  |  |
| ## |  |  |  |  |  |  |  |  |  |  |  |  |
| ## Mean item complexity = | | | | | 1.4 |  |  |  |  |  |  |  |
| ## Test of the hypothesis that 4 factors are sufficient. | | | | | | | | |  |  |  |  |
| ## |  |  |  |  |  |  |  |  |  |  |  |  |
| ## The degrees of freedom for the null model are | | | | | | | | 300 and the objective function was | | | | 12.8 |
| with Chi Square of | | | | 2877.74 | |  |  |  |  |  |  |  |
| ## The degrees of freedom for the model are 206 | | | | | | | | and the objective function was 1.79 | | | |  |
| ## |  |  |  |  |  |  |  |  |  |  |  |  |
| ## The root mean square of the residuals (RMSR) is 0.04 | | | | | | | | |  |  |  |  |
| ## The df corrected root mean square of the residuals is | | | | | | | | | 0.05 |  |  |  |
| ## |  |  |  |  |  |  |  |  |  |  |  |  |
| ## The harmonic number of observations is | | | | | |  | 235 with the empirical chi square | | | | 220.91 | with |
| prob < | | 0.23 |  |  |  |  |  |  |  |  |  |  |
| ## The total number of observations was | | | | | | 235 with Likelihood Chi Square = | | | | | 398.19 with prob | |
| < | 2.3e-14 | |  |  |  |  |  |  |  |  |  |  |
| ## |  |  |  |  |  |  |  |  |  |  |  |  |
| ## Tucker Lewis Index of factoring reliability = | | | | | | | | 0.89 |  |  |  |  |
| ## RMSEA index = 0.004 | | | | | and the 90 % confidence intervals are | | | | | 0.004 0.072 |  |  |
| ## BIC = | | -726.48 | |  |  |  |  |  |  |  |  |  |
| ## Fit based upon off diagonal values = 0.98 | | | | | | | |  |  |  |  |  |
| ## Measures of factor score adequacy | | | | | |  |  |  |  |  |  |  |
| ## |  |  |  |  |  |  |  | ML1 | ML2 ML3 | ML4 |  |  |
| ## | Correlation of scores with factors | | | | |  |  | 0.97 0.93 0.93 0.90 | | |  |  |
| ## | Multiple R square of scores with factors | | | | | |  | 0.94 0.87 0.87 0.81 | | |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |



## Minimum correlation of possible factor scores 0.87 0.74 0.74 0.62

we have to check how adequate is the model? We will look at the goodness of fit and residual statistics. We are looking for large values for the former and small ones for the latter. Most of the metrics are given in the output of the fa() function.

**Goodness of fit:**

1. Tucker Lewis Index of factoring reliability = 0.906. >0.90 is acceptable. >0.95 is excellent.
2. Comparative Fit Index(CFI) - calculated manually later.

**Residual Fit Statistics:**

1. RMSEA index = 0.004. <0.06 is excellent
2. The root mean square of the residuals (RMSR) is 0.04. <0.06 is excellent. We need to calculate comparative fit index (a goodness of fit metric) manually.



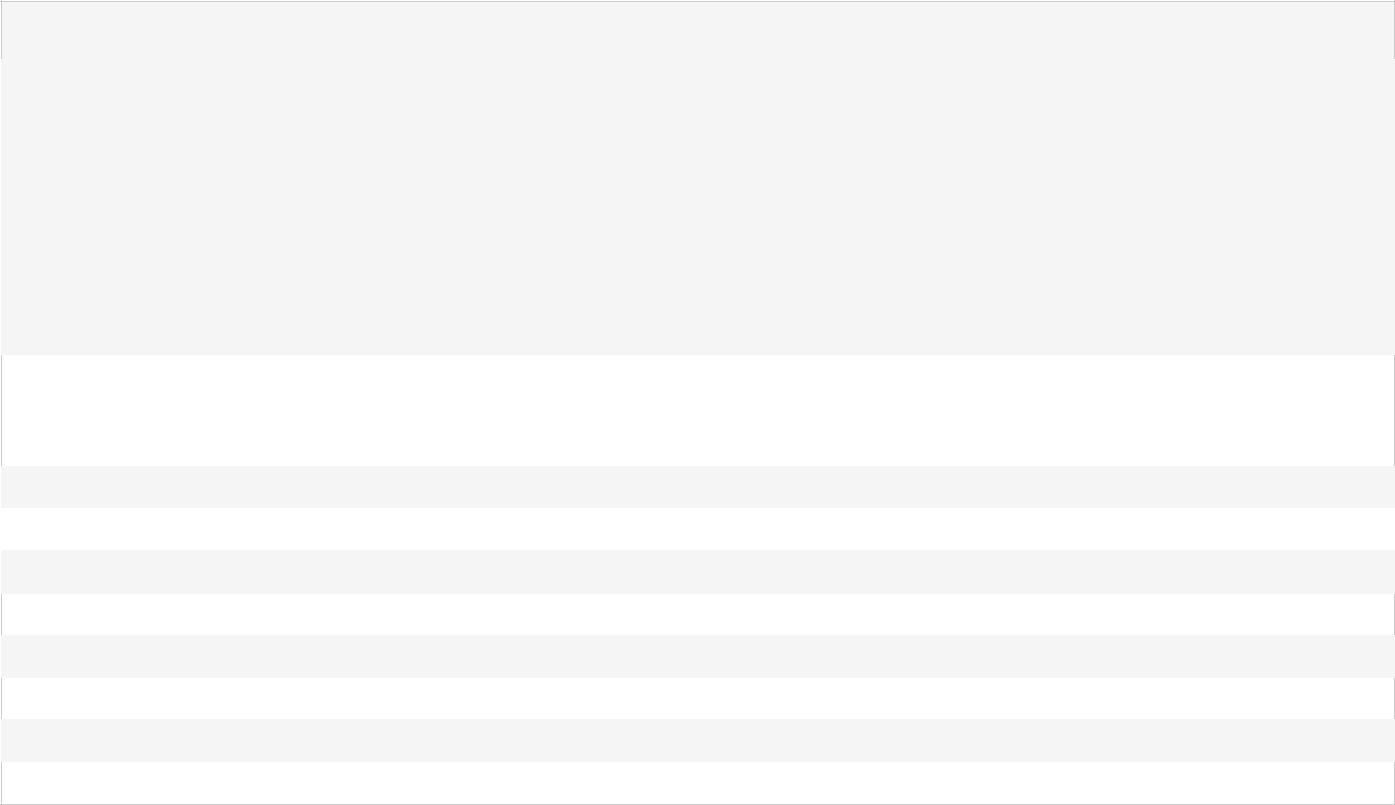
1-((fit$STATISTIC-fit$dof)/fit$null.chisq-fit$null.dof))

## [1] 0.9254409

A value of >0.90 for CFI is acceptable while >0.95 is excellent.

Looking at the metrics, we can conclude that we have an acceptable model.

Next step is to check the factors reliability



factor1 <- c(2,3,4,8,9,14,19,23,26)

factor2 <- c(5,7,16,20,22)

factor3 <- c(11,13,15)

factor4 <- c(12,17,18,24,25)

factor1alpha <- psych::alpha(cereal[,factor1], check.keys = TRUE)

factor2alpha <- psych::alpha(cereal[,factor2], check.keys = TRUE)

factor3alpha <- psych::alpha(cereal[,factor3], check.keys = TRUE)

factor4alpha <- psych::alpha(cereal[,factor4], check.keys = TRUE)

* Warning in psych::alpha(cereal[, factor4], check.keys = TRUE): Some items were negatively correlated with total scale and were automatically reversed.
* This is indicated by a negative sign for the variable name.

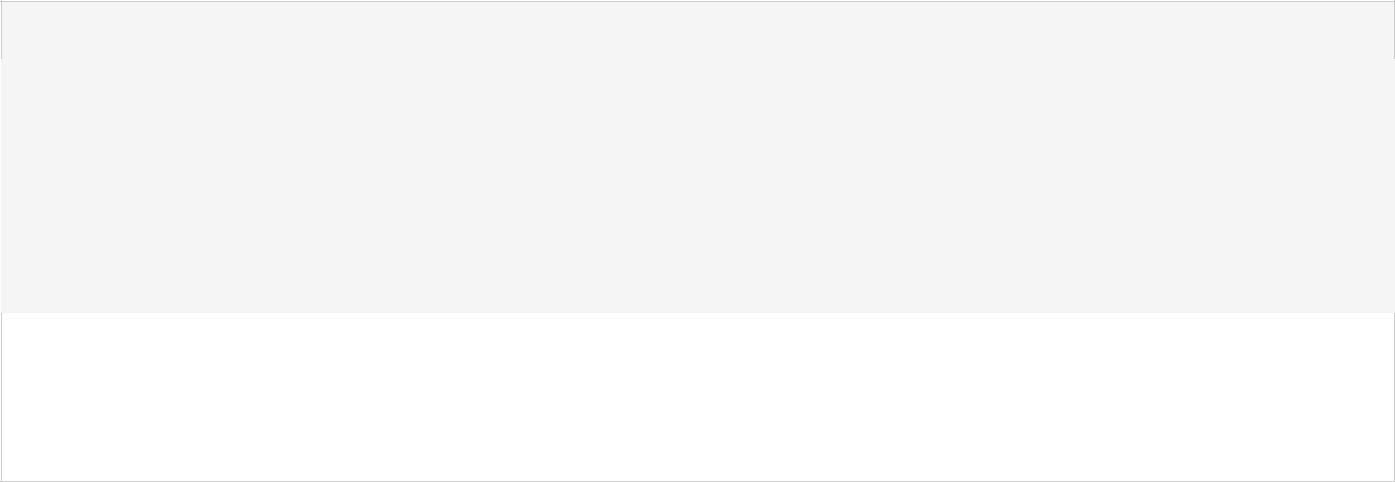
factor1alpha$total$raw\_alpha

* [1] 0.9131105 factor2alpha$total$raw\_alpha
* [1] 0.7713322 factor3alpha$total$raw\_alpha
* [1] 0.6867598 factor4alpha$total$raw\_alpha
* [1] 0.708883

As the alpha values are >0.7, the factors are reliable. Next step is too decide the what our factor mean

|  |  |  |
| --- | --- | --- |
| Factor1 | Health/Strength | Filling, Natural, Fibre, Satisfying, Energy, Health, Regular, Quality, Nutritious |
| Factor2 | Taste/wish | Sweet, Salt, Calories, Sugar, Process |
| Factor3 | Family | Kids, Economical, Family |
| Factor4 | Texture/Excitement | Fun, Soggy, Plain, Crisp, Treat, Boring |

Score and check our factors



cereal$factor1Score <- apply(cereal[,factor1],1,mean)

cereal$factor2Score <- apply(cereal[,factor2],1,mean)

cereal$factor3Score <- apply(cereal[,factor3],1,mean)

cereal$factor4Score <- apply(cereal[,factor4],1,mean)

colnames(cereal)[27:30] <-c("Strength", "Taste/wish", "Family", "Texture/Excitement")

aggregateCereal<-aggregate(cereal[,27:30], list(cereal[,1]), mean)

format(aggregateCereal, digits = 2)

* Group.1 Health Taste Family Texture/Excitement

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## 1 | | AllBran | 3.9 | 2.2 | 3.0 | 2.9 |
| ## | 2 | CMuesli | 4.0 | 2.8 | 3.5 | 3.3 |
| ## | 3 | CornFlakes | 3.3 | 2.7 | 4.1 | 3.3 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## 4 | JustRight | 3.6 | 2.7 | 3.2 | 3.2 |
| ## 5 | Komplete | 4.0 | 2.6 | 2.6 | 3.2 |
| ## 6 | NutriGrain | 3.4 | 3.1 | 4.0 | 3.5 |
| ## 7 | PMuesli | 4.1 | 2.9 | 3.2 | 3.5 |
| ## 8 | RiceBubbles | 2.9 | 2.2 | 4.2 | 3.3 |
| ## 9 | SpecialK | 3.5 | 2.3 | 3.7 | 3.4 |
| ## 10 | Sustain | 4.2 | 2.2 | 3.3 | 3.4 |
| ## 11 | Vitabrit | 3.9 | 1.9 | 3.9 | 2.9 |
| ## 12 | Weetabix | 3.9 | 2.1 | 3.8 | 2.7 |
|  |  |  |  |  |  |