

**ASEEM MALHOTRA**  
**101557400**

**STA 30005**  
**Multivariate Analysis**

**Assignment 2**

### Question 1

We start by entering the covariance matrix in R. We can weather that the matrix is symmetric to avoid any mistakes. There are 61 observations and 6 variables. That is more than the minimum 5 observation to each variable ratio ideal for PCA.

(a)

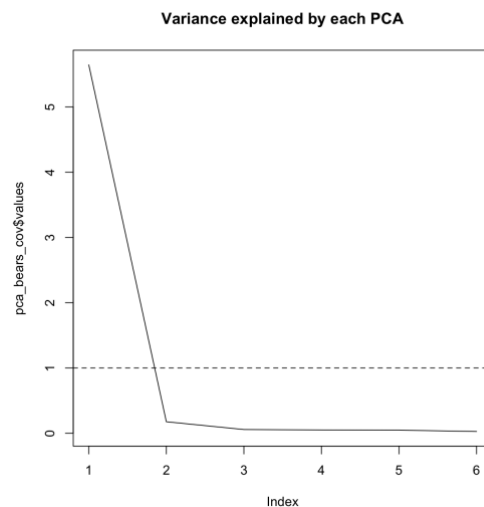
We can perform PCA on the covariance matrix with 6 factors and no rotation and we get the following loadings shown in *Table 1*.

*Table 1: Table of PCA & Parallel analysis results values*

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
SS Loadings	5.64	0.18	0.06	0.05	0.05	0.03
Proportionate Variance	0.94	0.03	0.01	0.01	0.01	0.00
Cumulative Variance	0.94	0.97	0.98	0.99	1.00	1.00
95 <sup>th</sup> percentile of PAR	1.64	1.36	1.14	1.01	0.86	0.72

We can obtain the PCA loadings values and variance and PC 1 explains variance of 5.64 which is 94%  $[(5.64 \cdot 100)/6]$  of total variance. The next best PC 2 only accounts for 3% of variance. *Figure 1* shows that it's the only principle component with variance above one which is the usual guideline to select the number of PCs.

When we compare the variance of each PC with 95<sup>th</sup> percentile values with 6 variables and 61 observations we only get one PC with variance greater than 95<sup>th</sup> percentile value, i.e. PC 1.



*Figure 1: PCA loadings with bar equal 1*

Hence the data can be easily summarized with just one PC dimension.

(b)

Principle component analysis can be calculated via Covariance matrix and correlation matrix. We can calculate the correlation between two variables by the formula.

$$COR_{x_1x_2} = \frac{COV_{x_1x_2}}{\sqrt{V_{x_1}}\sqrt{V_{x_2}}}$$

$COR_{x_1x_2}$  = Correlation between variable  $x_1$  and  $x_2$

$COV_{x_1x_2}$  = Covariance between variable  $x_1$  and  $x_2$

$V_{x_1}, V_{x_2}$  = Variance of variable  $x_1$  and  $x_2$  (the diagonal of covariance matrix)

We will use R (cov2cor()) function to calculate the correlation from covariance matrix but we can also calculate correlation between our 6 variables using the above formula.

E.g. Correlation between X2: body length and X6 head width will be

$$\text{COR}_{X2X6} = 117.73 / \sqrt{721.91 \times 21.26} = 0.9503$$

The PCA with correlation matrix produces the identical output shown in *Table 1*. Even *figure 2* shows that just one PC (PC 1) is enough as it accounts for 94% of total variance.

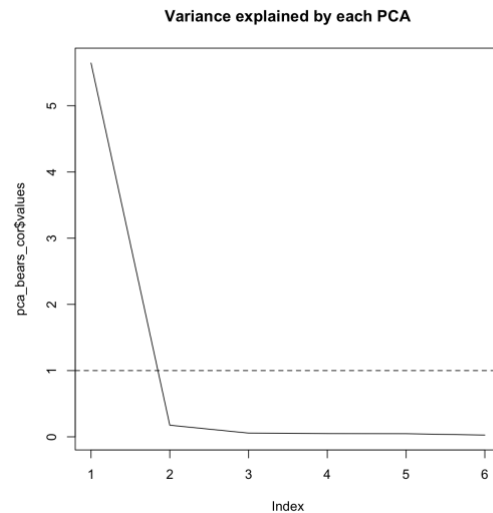


Figure 2: PCA loadings with bar at 1

(c)

Both covariance matrix and correlation matrix can be used to perform Principal component analysis. Usually correlation matrix is preferred if the data is not measured in same units and scale. Otherwise covariance matrix of scaled data is taken.

In this case, the covariance matrix and correlation matrix produces the same principle component analysis result. One reason could be that the data was scaled before covariance matrix was calculated hence we see no difference in the two Principal component analysis. Another reason could be as 5 of the 6 variables are measured in cm the data is already uniformed. But even though such similarity is highly likely if the data was scaled.

## APPENDIX 1

(a)

# entering values in covariance matrix

```
bears_cov <- matrix(c(3266.46, 1343.97, 731.54, 1175.50, 162.68, 238.37,
  1343.97, 721.91, 324.25, 537.35, 80.17, 117.73,
  731.54, 324.25, 179.28, 281.17, 39.15, 56.80,
  1175.50, 537.35, 281.17, 474.98, 63.73, 94.85,
  162.68, 80.17, 39.15, 63.73, 9.95, 13.88,
  238.37, 117.73, 56.80, 94.85, 13.88, 21.26), ncol = 6, byrow = T)
```

bears\_cov

# checking matrix is symmetric

isSymmetric(bears\_cov)

# performing PCA

```
pca_bears_cov <- principal(bears_cov, nfactors = 6, rotate = "none")
pca_bears_cov
```

# PCA Output

```
> pca_bears_cov
Principal Components Analysis
Call: principal(r = bears_cov, nfactors = 6, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	PC1	PC2	PC3	PC4	PC5	PC6	h2	u2	com
1	0.96	-0.23	0.07	0.06	0.13	0.02	1	0.0e+00	1.2
2	0.96	0.22	-0.05	0.16	-0.01	0.00	1	1.1e-16	1.2
3	0.97	-0.16	0.01	0.01	-0.12	-0.10	1	4.4e-16	1.1
4	0.98	-0.09	-0.14	-0.05	-0.05	0.09	1	0.0e+00	1.1
5	0.97	0.13	0.17	-0.06	-0.05	0.05	1	4.4e-16	1.1
6	0.97	0.13	-0.06	-0.11	0.10	-0.07	1	4.4e-16	1.1

```

          PC1  PC2  PC3  PC4  PC5  PC6
SS loadings      5.64 0.18 0.06 0.05 0.05 0.03
Proportion Var   0.94 0.03 0.01 0.01 0.01 0.00
Cumulative Var   0.94 0.97 0.98 0.99 1.00 1.00
Proportion Explained 0.94 0.03 0.01 0.01 0.01 0.00
Cumulative Proportion 0.94 0.97 0.98 0.99 1.00 1.00

Mean item complexity = 1.1
Test of the hypothesis that 6 components are sufficient.

The root mean square of the residuals (RMSR) is 0

Fit based upon off diagonal values = 1
```

# Plotting loadings chart with bar of 1

```
plot(pca_bears_cov$values, type = "l", main = "Variance explained by each PCA")
abline(h=1, lty=2)
```

## # Parallel Analysis Results

hornpa(k = 6, size = 61, reps = 500, seed = 100)

## Parallel Analysis Results

Method: pca

Number of variables: 6

Sample size: 61

Number of correlation matrices: 500

Seed: 100

Percentile: 0.95

Compare your observed eigenvalues from the table below generated using random percentile indicated (not the mean), you

Component	Mean	0.95
1	1.443	1.642
2	1.223	1.359
3	1.053	1.144
4	0.913	1.006
5	0.763	0.860
6	0.605	0.722

(b)

## # converting covariance matrix into correlation matrix

bears\_cor &lt;- cov2cor(bears\_cov)

bears\_cor

```
> bears_cor
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	1.0000000	0.8752042	0.9559463	0.9437268	0.9023686	0.9045473
[2,]	0.8752042	1.0000000	0.9013066	0.9176521	0.9459296	0.9503071
[3,]	0.9559463	0.9013066	1.0000000	0.9635303	0.9269457	0.9200274
[4,]	0.9437268	0.9176521	0.9635303	1.0000000	0.9270320	0.9438827
[5,]	0.9023686	0.9459296	0.9269457	0.9270320	1.0000000	0.9543250
[6,]	0.9045473	0.9503071	0.9200274	0.9438827	0.9543250	1.0000000

## # PCA on correlation matrix

```
pca_bears_cor <- principal(bears_cor, nfactors = 6, rotate = "none")
pca_bears_cor
```

```
Principal Components Analysis
Call: principal(r = bears_cor, nfactors = 6, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	PC1	PC2	PC3	PC4	PC5	PC6	h2	u2	com
1	0.96	-0.23	0.07	0.06	0.13	0.02	1	4.4e-16	1.2
2	0.96	0.22	-0.05	0.16	-0.01	0.00	1	1.1e-15	1.2
3	0.97	-0.16	0.01	0.01	-0.12	-0.10	1	3.3e-16	1.1
4	0.98	-0.09	-0.14	-0.05	-0.05	0.09	1	0.0e+00	1.1
5	0.97	0.13	0.17	-0.06	-0.05	0.05	1	3.3e-16	1.1
6	0.97	0.13	-0.06	-0.11	0.10	-0.07	1	4.4e-16	1.1

	PC1	PC2	PC3	PC4	PC5	PC6
SS loadings	5.64	0.18	0.06	0.05	0.05	0.03
Proportion Var	0.94	0.03	0.01	0.01	0.01	0.00
Cumulative Var	0.94	0.97	0.98	0.99	1.00	1.00
Proportion Explained	0.94	0.03	0.01	0.01	0.01	0.00
Cumulative Proportion	0.94	0.97	0.98	0.99	1.00	1.00

```
Mean item complexity = 1.1
Test of the hypothesis that 6 components are sufficient.

The root mean square of the residuals (RMSR) is 0

Fit based upon off diagonal values = 1
```

## # Plotting loadings chart with bar of 1

```
plot(pca_bears_cor$values, type = "l", main = "Variance explained by each PCA")
abline(h=1, lty=2)
```

## Question 2

(a)

We start by looking into the dataset for missing values. There were none. When we examine the structure of each variable the first variable which is the cultivator is characterized as number. There are only 3 cultivators so it's more useful to change cultivator to factors. The summary of the wine dataset can now be studied for each cultivator separately. There are 59 samples for first cultivator, 71 for the second and 48 for the third.

Principle component analysis is highly sensitive to outliers. For a dataset with 13 variables it's hard to detect outliers that have impact on PCA so can look for outliers across different methods and remove the observations that are common. A list of outliers by different method is given in *Table 2*.

### Methods used for identifying outliers:

**Mahalanobis Distance.** We calculate Mahalanobis distance and the one furthest from mean are considered outliers.

**Boxplots.** We can look at boxplots for each chemical to look for observations that are outliers.

**K-means clustering.** Clustering can be used to identify outliers. If we make lots of clusters( $k=20$ ) the small clusters can be considered as outliers. When we make 20 clusters we get 3 clusters with 1, 2 and 4 observations so we will consider them as outliers.

*Table 2: Outliers of different methods*

Method	Outliers
Mahalanobis	14, 60, 70, 72, 74, 96, 97, 111, 116, 122, 159, 160
Boxplot chem2	124, 138, 174
Boxplot chem3	26, 60, 122
Boxplot chem4	60, 74, 122
Boxplot chem5	70, 74, 79, 96
Boxplot chem9	96, 111
Boxplot chem10	152, 159, 160
Boxplot chem11	116
k-means cluster 1	60
k-means cluster 2	159, 160
k-means cluster 3	70, 74, 96, 122

The observations that are often outliers are 60, 74, 96, 122. So we will remove them from further analysis.

(b)

Principal component analysis can be done via correlation and covariance matrix. The wine dataset is not measured on same scale so correlation matrix will be more suitable of the two.

The *Figure 3* below shows correlation matrix between 13 chemicals. It's a symmetric matrix and the diagonal represents correlation of a chemical with itself, which is always 1.

	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	chem9	chem10	chem11	chem12	chem13
chem1	1	0.08	0.25	-0.32	0.33	0.31	0.28	-0.15	0.15	0.55	-0.07	0.09	0.65
chem2	0.08	1	0.16	0.31	-0.03	-0.34	-0.43	0.29	-0.23	0.24	-0.56	-0.38	-0.2
chem3	0.25	0.16	1	0.38	0.29	0.09	0.03	0.18	-0.04	0.24	-0.07	-0.06	0.24
chem4	-0.32	0.31	0.38	1	-0.19	-0.41	-0.49	0.39	-0.27	0.01	-0.32	-0.37	-0.49
chem5	0.33	-0.03	0.29	-0.19	1	0.19	0.17	-0.26	0.16	0.25	0.01	0.03	0.41
chem6	0.31	-0.34	0.09	-0.41	0.19	1	0.87	-0.46	0.62	-0.05	0.43	0.69	0.51
chem7	0.28	-0.43	0.03	-0.49	0.17	0.87	1	-0.58	0.67	-0.19	0.57	0.78	0.52
chem8	-0.15	0.29	0.18	0.39	-0.26	-0.46	-0.58	1	-0.38	0.13	-0.25	-0.52	-0.31
chem9	0.15	-0.23	-0.04	-0.27	0.16	0.62	0.67	-0.38	1	-0.02	0.29	0.52	0.33
chem10	0.55	0.24	0.24	0.01	0.25	-0.05	-0.19	0.13	-0.02	1	-0.52	-0.45	0.32
chem11	-0.07	-0.56	-0.07	-0.32	0.01	0.43	0.57	-0.25	0.29	-0.52	1	0.57	0.23
chem12	0.09	-0.38	-0.06	-0.37	0.03	0.69	0.78	-0.52	0.52	-0.45	0.57	1	0.32
chem13	0.65	-0.2	0.24	-0.49	0.41	0.51	0.52	-0.31	0.33	0.32	0.23	0.32	1

Figure 3: Correlation matrix of 13 Chemicals

(c)

Since we have removed the 4 outliers the mean and standard deviation of 13 chemicals out of 174 observations is given below in table

Table 3: Mean and Standard Deviations of each Chemical

Chemicals	Mean	Standard Deviation
chem 1	13.02	0.81
chem 2	2.35	1.12
chem 3	2.37	0.26
chem 4	19.44	3.14
chem 5	99.11	13.19
chem 6	2.28	0.62
chem 7	2.01	0.97
chem 8	0.36	0.12
chem 9	1.58	0.56
chem 10	5.09	2.32
chem 11	0.95	0.23
chem 12	2.60	0.71
chem 13	747.36	316.49

By looking at the mean and standard deviation in *Table 3* we can clearly see the contrast in the scale of each variable. Chemical 8 is measured in the lowest scale (mean = 0.36, sd = 0.12) and chemical 13 the highest (mean = 747.36, sd = 316.49).



So, a PCA of unscaled data will produce a highly-skewed result. So, either a correlation matrix or covariance matrix of scaled data should be used.

(d)

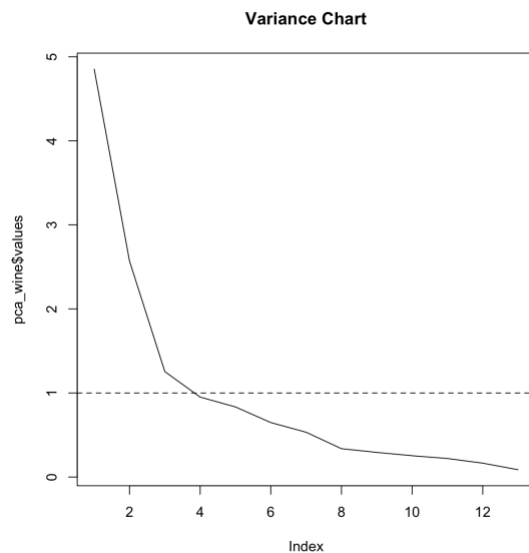
We use the correlation matrix to perform PCA and the results are shown in *Table 4*.

*Table 4: Principal component analysis results*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
SS Loadings	4.85	2.57	1.25	0.95	0.83	0.65	0.53	0.34	0.29	0.25	0.22	0.17	0.09
Proportionate Variance	0.37	0.20	0.10	0.07	0.06	0.05	0.04	0.03	0.02	0.02	0.02	0.01	0.01
Cumulative Variance	0.37	0.57	0.67	0.74	0.80	0.85	0.90	0.92	0.94	0.96	0.98	0.99	1.0
95 <sup>th</sup> percentile of PAR													

As per the PCA results from *Table 4* we will retain 3 PCAs as PC 1, PC 2 and PC 3 have SS Loading above 1 and they all explain 37%, 20% and 10% of variance respectively. So, all 3 together explain 67% of the total variance. *Figure 4* shows a variance chart which shows the same result.

Sometimes we have a limit of minimum variance requirement. E.g. if we need minimum of 80% variance then we will choose 5 PCAs.



*Figure 4: Variance chart*

(e)

Parallel analysis can be used to choose how many PCA's to retain. If the Variance explained by a PC is more than the 95<sup>th</sup> percentile with equal variables and size than that PC should be retained. For this sample, we have 13 variables and 174 size (after reducing 4 outliers). The values are compared in *Table 5* below and the screen plot of output can be viewed in APPENDIX 2 (e).

*Table 5: PCS loadings compared with Parallel analysis*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
SS Loadings	4.85	2.57	1.25	0.95	0.83	0.65	0.53	0.34	0.29	0.25	0.22	0.17	0.09
95 <sup>th</sup> percentile	1.6	1.44	1.33	1.24	1.17	1.1	1.03	0.96	0.91	0.85	0.79	0.73	0.66

According to the parallel analysis only 2 PC's qualify. PC 1 and PC 2 are the only one with SS Loadings higher than 95<sup>th</sup> percentile of parallel analysis. So, we select only PC 1 and PC 2.

(f)

**Linear combination for PC 1**

$$\Rightarrow 0.336*(\text{chem1}) - 0.531*(\text{chem2}) - 0.044*(\text{chem3}) - 0.624*(\text{chem4}) + 0.279*(\text{chem5}) + 0.854*(\text{chem6}) + 0.936*(\text{chem7}) - 0.654*(\text{chem8}) + 0.681*(\text{chem9}) - 0.181*(\text{chem10}) + 0.635*(\text{chem11}) + 0.814*(\text{chem12}) + 0.628*(\text{chem13})$$

**Score of PC 1 of standardizes value with observation 1 = 7.497**

**Linear combination for PC 2**

$$\Rightarrow 0.772*(\text{chem1}) + 0.335*(\text{chem2}) + 0.495*(\text{chem3}) - 0.071*(\text{chem4}) + 0.537*(\text{chem5}) + 0.089*(\text{chem6}) - 0.03*(\text{chem7}) + 0.031*(\text{chem8}) + 0.026*(\text{chem9}) + 0.835*(\text{chem10}) - 0.439*(\text{chem11}) - 0.287*(\text{chem12}) + 0.582*(\text{chem13})$$

**Score of PC 1 of standardizes value with observation 2 = 2.43**

After we multiply the scaled observation with PC 1 and PC 2 we get a correlation of 0 among the new data. The correlation matrix is given in *Table 6*.

*Table 6: Correlation between PCAs*

	PC 1	PC 2
PC 1	1	0
PC 2	0	1

## APPENDIX 2

(a)

# Read file and look for missing values

wine &lt;- read.csv(file.choose())

sum(is.na(wine))

[1] 0

# Cultivar can be better used if changed to factor as there are only cultivators

str(wine)

wine\$cultivar &lt;- as.factor(wine\$cultivar)

summary(wine)

cultivar	chem1	chem2	chem3	chem4	chem5
1:59	Min. :11.03	Min. :0.740	Min. :1.360	Min. :10.60	Min. : 70.00
2:71	1st Qu.:12.36	1st Qu.:1.603	1st Qu.:2.210	1st Qu.:17.20	1st Qu.: 88.00
3:48	Median :13.05	Median :1.865	Median :2.360	Median :19.50	Median : 98.00
	Mean :13.00	Mean :2.336	Mean :2.367	Mean :19.49	Mean : 99.74
	3rd Qu.:13.68	3rd Qu.:3.083	3rd Qu.:2.558	3rd Qu.:21.50	3rd Qu.:107.00
	Max. :14.83	Max. :5.800	Max. :3.230	Max. :30.00	Max. :162.00
	chem6	chem7	chem8	chem9	chem10
	Min. :0.980	Min. :0.340	Min. :0.1300	Min. :0.410	Min. : 1.280
	1st Qu.:1.742	1st Qu.:1.205	1st Qu.:0.2700	1st Qu.:1.250	1st Qu.: 3.220
	Median :2.355	Median :2.135	Median :0.3400	Median :1.555	Median : 4.690
	Mean :2.295	Mean :2.029	Mean :0.3619	Mean :1.591	Mean : 5.058
	3rd Qu.:2.800	3rd Qu.:2.875	3rd Qu.:0.4375	3rd Qu.:1.950	3rd Qu.: 6.200
	Max. :3.880	Max. :5.080	Max. :0.6600	Max. :3.580	Max. :13.000
	chem11	chem12	chem13		
	Min. :0.4800	Min. :1.270	Min. : 278.0		
	1st Qu.:0.7825	1st Qu.:1.938	1st Qu.: 500.5		
	Median :0.9650	Median :2.780	Median : 673.5		
	Mean :0.9574	Mean :2.612	Mean : 746.9		
	3rd Qu.:1.1200	3rd Qu.:3.170	3rd Qu.: 985.0		
	Max. :1.7100	Max. :4.000	Max. :1680.0		

# Mahalanobis outliers

md &lt;- mahalanobis(wine[,2:14], colMeans(wine[,2:14]), cov(wine[,2:14]))

match(boxplot(md)\$out, md)

14 60 70 72 74 96 97 111 116 122 159 160

#Boxplots

match(boxplot(wine[,3])\$out, wine[,3])

[1] 124 138 174

match(boxplot(wine[,4])\$out, wine[,4])

[1] 26 60 122

match(boxplot(wine[,5])\$out, wine[,5])

[1] 60 74 122

```
match(boxplot(wine[,6])$out, wine[,6])
[1] 70 74 79 96
```

```
match(boxplot(wine[,10])$out, wine[,10])
[1] 96 111
```

```
match(boxplot(wine[,11])$out, wine[,11])
[1] 152 159 160
```

```
match(boxplot(wine[,12])$out, wine[,12])
[1] 116
```

### # k-means clustering

```
wine_scale <- dist(scale(wine[, 2:14]))
wine_kmean <- kmeans(wine_scale, 20, 50)
wine_kmean$size
[1] 14 5 7 2 7 5 16 12 11 8 16 10 8 10 1 14 10 4 11 7
```

```
wine_kmean$cluster[wine_kmean$cluster == 4]
159 160
4 4
wine_kmean$cluster[wine_kmean$cluster == 15]
60
15
wine_kmean$cluster[wine_kmean$cluster == 18]
70 74 96 122
18 18 18 18
```

### # removing observations

```
wine <- wine[-c(60, 74, 96, 122), ]
dim(wine)
[1] 174 14
```

### (b)

#### # Correlation Matrix for all chemicals

```
cor(wine[, 2:14])
write.csv(round(cor(wine[, 2:14]), 2), "cor_martix.csv")
```

(c)

# Mean and Standard deviation

```
> sapply(wine[,2:14], mean)
  chem1      chem2      chem3      chem4      chem5      chem6      chem7
13.0156322  2.3545402  2.3669540 19.4367816 99.1149425  2.2848851  2.0137931
  chem8      chem9     chem10     chem11     chem12     chem13
0.3628161  1.5841954  5.0944828  0.9538851  2.6048276 747.3563218
> sapply(wine[, 2:14], sd)
  chem1      chem2      chem3      chem4      chem5      chem6      chem7
0.8112649  1.1217732  0.2576111  3.1371581 13.1930313  0.6240681  0.9745884
  chem8      chem9     chem10     chem11     chem12     chem13
0.1248842  0.5562999  2.3204792  0.2289775  0.7076630 316.4898880
```

(d)

# PCA of correlated variables and chart for loading above 1

```
pca_wine <- pca(cor(wine[, 2:14]), nfactors = 13, rotate = "none")
pca_wine
plot(pca_wine$values, type = "l", main = "Variance Chart")
abline(h=1, lty =2)
```

```
Principal Components Analysis
Call: principal(r = r, nfactors = nfactors, residuals = residuals,
  rotate = rotate, n.obs = n.obs, covar = covar, scores = scores,
  missing = missing, impute = impute, oblique.scores = oblique.scores,
  method = method)
Standardized loadings (pattern matrix) based upon correlation matrix
   PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9  PC10  PC11  PC12  PC13  h2      u2 com
chem1  0.34  0.77 -0.14 -0.01  0.25  0.18 -0.12  0.22  0.29 -0.08 -0.10 -0.11 -0.02  1 -4.4e-16 2.6
chem2 -0.53  0.34  0.06  0.51 -0.08  0.49  0.24  0.07 -0.06  0.14  0.04  0.05  0.01  1  0.0e+00 4.5
chem3 -0.04  0.49  0.79 -0.16  0.02  0.07 -0.10 -0.11 -0.14  0.03 -0.23  0.01 -0.02  1  1.1e-16 2.2
chem4 -0.62 -0.07  0.62  0.15 -0.08 -0.16 -0.21  0.21  0.11 -0.04  0.24 -0.03  0.02  1  0.0e+00 3.2
chem5  0.28  0.54  0.08 -0.32 -0.64 -0.02  0.29 -0.04  0.15 -0.04  0.06  0.01  0.02  1  1.1e-15 3.6
chem6  0.85  0.09  0.18  0.21  0.12 -0.05  0.00 -0.27  0.14  0.17  0.15 -0.08 -0.14  1  1.0e-15 1.8
chem7  0.94 -0.03  0.13  0.14  0.07  0.00 -0.03 -0.08  0.02  0.09  0.00 -0.07  0.25  1  1.4e-15 1.3
chem8 -0.65  0.03  0.21 -0.16  0.51 -0.07  0.44 -0.13  0.11 -0.11  0.05  0.01  0.04  1  1.1e-15 3.4
chem9  0.68  0.03  0.12  0.45 -0.03 -0.39  0.31  0.21 -0.10 -0.08 -0.12 -0.03 -0.04  1  0.0e+00 3.5
chem10 -0.18  0.84 -0.19  0.08  0.13 -0.34 -0.11 -0.01  0.02  0.14  0.03  0.25  0.03  1  1.7e-15 2.0
chem11  0.63 -0.44  0.18 -0.41  0.15  0.09  0.14  0.29  0.03  0.25  0.01  0.11 -0.02  1  7.8e-16 4.2
chem12  0.81 -0.29  0.18  0.16  0.01  0.22 -0.06 -0.05  0.10 -0.26  0.02  0.25 -0.01  1  8.9e-16 2.2
chem13  0.63  0.58 -0.05 -0.21  0.17  0.10  0.00  0.07 -0.33 -0.12  0.23 -0.04 -0.01  1  2.7e-15 3.5

   PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9  PC10  PC11  PC12  PC13
SS loadings      4.85 2.57 1.25 0.95 0.83 0.65 0.53 0.34 0.29 0.25 0.22 0.17 0.09
Proportion Var    0.37 0.20 0.10 0.07 0.06 0.05 0.04 0.03 0.02 0.02 0.02 0.01 0.01
Cumulative Var    0.37 0.57 0.67 0.74 0.80 0.85 0.90 0.92 0.94 0.96 0.98 0.99 1.00
Proportion Explained 0.37 0.20 0.10 0.07 0.06 0.05 0.04 0.03 0.02 0.02 0.02 0.01 0.01
Cumulative Proportion 0.37 0.57 0.67 0.74 0.80 0.85 0.90 0.92 0.94 0.96 0.98 0.99 1.00

Mean item complexity = 2.9
Test of the hypothesis that 13 components are sufficient.

The root mean square of the residuals (RMSR) is 0

Fit based upon off diagonal values = 1
```

(e)

## # Parallel Analyses results

dim(wine[,2:14])

hornpa(k = 13, size = 174, reps = 500, seed = 100)

```

Parallel Analysis Results

Method: pca
Number of variables: 13
Sample size: 174
Number of correlation matrices: 500
Seed: 100
Percentile: 0.95

Compare your observed eigenvalues from your original dataset to the 95 percentile in the table below
generated using random data. If your eigenvalue is greater than the percentile indicated (not the me
an), you have support to retain that factor/component.

Component Mean 0.95
 1 1.478 1.597
 2 1.357 1.441
 3 1.264 1.331
 4 1.183 1.239
 5 1.112 1.165
 6 1.047 1.096
 7 0.981 1.030
 8 0.917 0.962
 9 0.858 0.907
10 0.797 0.847
11 0.739 0.791
12 0.673 0.728
13 0.593 0.659

```

(f)

## # PCA coefficients of PC 1

```

> pca_wine$loadings[,1]
      chem1      chem2      chem3      chem4      chem5      chem6      chem7
0.33604581 -0.53082033 -0.04445462 -0.62383046 0.27882890 0.85433613 0.93562150
      chem8      chem9      chem10      chem11      chem12      chem13
-0.65393603 0.68098125 -0.18112686 0.63465362 0.81410934 0.62759519

```

## # PCA coefficients of PC 2

```

> pca_wine$loadings[,2]
      chem1      chem2      chem3      chem4      chem5      chem6      chem7
0.77210813 0.33549510 0.49481869 -0.07129168 0.53733621 0.08896987 -0.03097594
      chem8      chem9      chem10      chem11      chem12      chem13
0.03105881 0.02576939 0.83537745 -0.43874950 -0.28672965 0.58198061

```

## # Creating matrix of PC 1 &amp; PC 2 score

```
a1 <- as.matrix(c(0.336, -0.531, -0.044, -0.624, 0.279, 0.854, 0.936, -0.654, 0.681, -0.181,
0.635, 0.814, 0.628))
```

```
a2 <- as.matrix(c(.772, .355, .495, -0.071, .537, 0.089, -0.03, 0.031, 0.026, .835, -.439, -.287,
.582))
```

**# scaling observations**

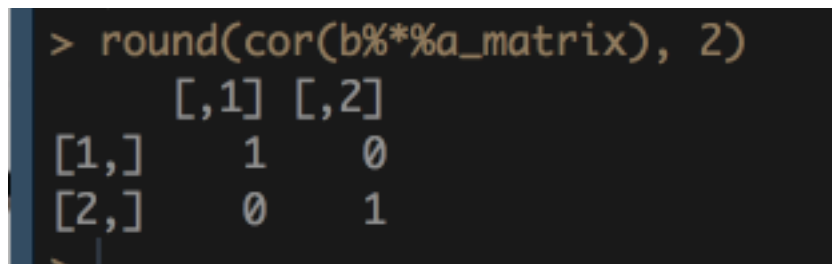
```
b <- as.matrix(scale(wine[,2:14]))  
dim(b)
```

**# Calculating score for PC 1 & PC 2 for the first observation**

```
b[1,]%*%a1  
b[1,]%*%a2
```

**# checking correlation of observation after cross product with PCs**

```
round(cor(b%*%a_matrix), 2)
```



```
> round(cor(b%*%a_matrix), 2)  
      [,1] [,2]  
[1,]    1    0  
[2,]    0    1
```

## Question 3

(a)

When we have an initial look at dataset it has 30 columns for questions and 1 each for ID and gender. There are 1001 observations. Each of the 30 questions is rated on the scale of 1 to 4.

**Missing values:** When looking for missing values we found 136 missing values but they were all for gender which is a categorical variable and more importantly not being used in our analysis. So, then we will remove the first 2 columns (ID and gender) from our dataset.

**Outliers:** we did find few Mahalanobis outliers, but they are very few out of 1001 observations and the range of values is in-between 1 and 4 we will keep them.

(b)

To perform factor analysis, we need strong correlation among variables. We look for a minimum correlation of 0.30. We can see from *Figure 5* below of the correlation Matrix that question 3, 4, 6, 8, 14, 15, 29 are not correlated to any other variable positively to negatively at 0.30. So, we proceed by removing those variables.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30
Q1	1	0.4	-0.1	-0.1	0.4	0.2	0.3	0.3	0.4	-0.3	0.2	0.3	0	0.2	-0.2	0.1	0.3	0.2	0.2	-0.3	0.2	0.3	0.3	0.1	0.1	0.4	0.2	0.3	-0.3	0.1
Q2	0.4	1	-0.1	0.1	0.3	0.1	0.4	0.1	0.4	-0.2	0.3	0.4	0.2	0.3	-0.2	0.3	0.2	0.1	-0.2	0.2	0.3	0.6	0.2	0.2	0.2	0.4	0.2	-0.1	0.2	0
Q3	-0.1	-0.1	1	0.1	-0.1	-0.1	0	-0.1	0.1	-0.1	0.1	-0.1	-0.1	-0.1	0	0	-0.1	-0.1	0	-0	0	-0	-0	-0	-0	-0.1	-0.1	-0.1	0.1	-0
Q4	-0.1	0.1	0.1	1	-0.1	-0	-0	-0	-0	0.1	0	0.1	0.1	0.2	0.1	0.1	-0.1	0.1	0.1	0.1	0.2	-0	0	0	0.1	0	0.1	-0.1	0.1	0.1
Q5	0.4	0.3	-0.1	-0.1	1	0.2	0.3	0.2	0.4	-0.3	0.2	0.2	0.1	0.1	-0.2	0.1	0.4	0.2	0.1	-0.4	0.1	0.3	0.2	0.1	0.1	0.5	0.1	0.3	-0.2	0.1
Q6	0.2	0.1	-0	-0	0.2	1	0.3	0.1	0.2	-0.1	0.3	0.2	0.1	0.2	-0.1	0.2	0.2	0.1	0.3	-0.1	0.3	0.3	0.2	0.2	0.3	0.2	0.1	0.1	-0.1	0.2
Q7	0.3	0.4	-0.1	-0	0.3	0.3	1	0.1	0.4	-0.2	0.4	0.4	0.1	0.3	-0.1	0.2	0.3	0.1	0.1	-0.3	0.2	0.4	0.5	0.2	0.2	0.3	0.4	0.4	-0.2	0.3
Q8	0.3	0.1	0.1	-0	0.2	0.1	0.1	1	0.2	-0.1	0.1	0.1	-0	0.1	-0.1	-0	0.2	0.1	0.1	-0.1	0.1	0.2	0.1	0	0.1	0.2	0.1	0.2	-0.1	0.1
Q9	0.4	0.4	-0.1	-0	0.4	0.2	0.4	0.2	1	-0.4	0.3	0.4	0.2	0.2	-0.2	0.1	0.4	0.1	0.1	-0.4	0.1	0.3	0.4	0.2	0.1	0.4	0.2	0.4	-0.3	0.1
Q10	-0.3	-0.2	0.1	0.1	-0.3	-0.1	-0.2	-0.1	-0.4	1	-0.2	-0.2	0	-0.1	0.2	-0	-0.5	-0.1	-0.1	0.4	-0.1	-0.1	-0.2	-0.1	-0.1	-0.4	-0.1	-0.3	0.3	-0.1
Q11	0.2	0.3	-0.1	0	0.2	0.3	0.4	0.1	0.3	-0.2	1	0.4	0.2	0.3	-0.1	0.3	0.3	0.2	0.2	-0.2	0.3	0.3	0.4	0.2	0.3	0.2	0.3	0.3	-0.2	0.2
Q12	0.3	0.4	-0.1	0.1	0.2	0.2	0.4	0.1	0.4	-0.2	0.4	1	0.2	0.3	0	0.2	0.3	0.2	0.2	-0.2	0.2	0.3	0.4	0.2	0.2	0.2	0.4	0.3	-0.1	0.2
Q13	0	0.2	-0.1	0.1	0.1	0.1	0.1	-0	0.2	0	0.2	0.2	1	0.2	0.1	0.2	0.1	0	0	-0	0.1	0.2	0.2	0.4	0.1	0	0.2	0	0	0.1
Q14	0.2	0.3	0	0.2	0.1	0.2	0.3	0.1	0.2	-0.1	0.3	0.3	0.2	1	-0	0.2	0.2	0.3	0.2	-0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.2	-0.1	0.2
Q15	-0.2	-0	0	0.1	-0.2	-0.1	-0.1	-0.1	-0.2	0.2	-0.1	0	0.1	-0	1	0.1	-0.2	-0	-0.1	0.2	0	-0.1	-0	-0.1	-0	-0.2	0	-0.2	0.2	-0.1
Q16	0.1	0.2	-0.1	0.1	0.1	0.2	0.2	-0	0.1	-0	0.3	0.2	0.2	0.2	0.1	1	0.2	0.2	0.1	-0.1	0.2	0.2	0.2	0.1	0.2	0.1	0.2	0.1	-0.1	0.1
Q17	0.3	0.3	-0.1	-0.1	0.4	0.2	0.3	0.2	0.4	-0.5	0.3	0.3	0.1	0.2	-0.2	0.2	1	0.2	0.1	-0.4	0.2	0.3	0.3	0.2	0.1	0.4	0.2	0.4	-0.2	0.1
Q18	0.2	0.2	0	0.1	0.2	0.1	0.1	0.1	0.1	-0.1	0.2	0.2	0	0.3	-0	0.2	0.2	1	0.2	-0.1	0.2	0.1	0.2	0.1	0.2	0.2	0.2	0.2	-0.1	0.3
Q19	0.2	0.1	-0	0.1	0.1	0.3	0.1	0.1	0.1	-0.1	0.2	0.2	0	0.2	-0.1	0.1	0.1	0.2	1	-0.1	0.5	0.2	0.1	0.1	0.5	0.1	0.1	0.1	-0.1	0.2
Q20	-0.3	-0.2	0	0.1	-0.4	-0.1	-0.3	-0.1	-0.4	0.4	-0.2	-0.2	-0	-0.2	0.2	-0.1	-0.4	-0.1	-0.1	1	-0.1	-0.2	-0.2	-0.1	-0.1	-0.4	-0.1	-0.4	0.3	-0
Q21	0.2	0.2	-0	0.2	0.1	0.3	0.2	0.1	0.1	-0.1	0.3	0.2	0.1	0.3	0	0.2	0.2	0.2	0.5	-0.1	1	0.2	0.2	0.1	0.8	0.1	0.2	0.1	-0.1	0.2
Q22	0.3	0.3	-0	-0	0.3	0.3	0.4	0.2	0.3	-0.1	0.3	0.3	0.2	0.2	-0.1	0.2	0.3	0.1	0.2	-0.2	0.2	1	0.5	0.2	0.2	0.2	0.3	0.3	-0.2	0.2
Q23	0.3	0.6	-0.1	0	0.2	0.2	0.5	0.1	0.4	-0.2	0.4	0.4	0.2	0.3	-0	0.2	0.3	0.2	0.1	-0.2	0.2	0.5	1	0.3	0.2	0.2	0.4	0.3	-0.1	0.3
Q24	0.1	0.2	-0.1	0	0.1	0.2	0.2	0	0.2	-0.1	0.2	0.2	0.4	0.2	-0.1	0.1	0.2	0.1	0.1	-0.1	0.1	0.2	0.3	1	0.2	0.2	0.3	0.2	-0.1	0.2
Q25	0.1	0.2	-0	0.1	0.1	0.3	0.2	0.1	0.1	-0.1	0.3	0.2	0.1	0.3	-0	0.2	0.1	0.2	0.5	-0.1	0.8	0.2	0.2	0.2	1	0.1	0.2	0.1	-0.1	0.2
Q26	0.4	0.2	-0.1	0	0.5	0.2	0.3	0.2	0.4	-0.4	0.2	0.2	0	0.2	-0.2	0.1	0.4	0.2	0.1	-0.4	0.1	0.2	0.2	0.2	0.1	1	0.2	0.4	-0.3	0.1
Q27	0.2	0.4	-0.1	0.1	0.1	0.1	0.4	0.1	0.2	-0.1	0.3	0.4	0.2	0.3	0	0.2	0.2	0.1	-0.1	0.2	0.3	0.4	0.3	0.2	0.2	1	0.3	-0	0.2	0
Q28	0.3	0.2	-0.1	-0.1	0.3	0.1	0.4	0.2	0.4	-0.3	0.3	0.3	0	0.2	-0.2	0.1	0.4	0.2	0.1	-0.4	0.1	0.3	0.3	0.2	0.1	0.4	0.3	1	-0.2	0.2
Q29	-0.3	-0.1	0.1	0.1	-0.2	-0.1	-0.2	-0.1	-0.3	0.3	-0.2	-0.1	0	-0.1	0.2	-0.1	-0.2	-0.1	-0.1	0.3	-0.1	-0.2	-0.1	-0.1	-0.1	-0.3	-0	-0.2	1	-0
Q30	0.1	0.2	-0	0.1	0.1	0.2	0.3	0.1	0.1	-0.1	0.2	0.2	0.1	0.2	-0.1	0.1	0.1	0.3	0.2	-0	0.2	0.2	0.3	0.2	0.2	0.1	0.2	0.2	-0	0

Figure 5: Correlation matrix of all 30 questions



After removing the selected questions, in *Figure 6* we can see that all variables are correlated with at least one other variable with absolute value of greater than equal to 0.3.

	Q1	Q2	Q5	Q7	Q9	Q10	Q11	Q12	Q13	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q30
Q1	1	0.35	0.43	0.3	0.42	-0.3	0.23	0.27	0	0.1	0.32	0.17	0.15	-0.31	0.15	0.25	0.3	0.13	0.13	0.39	0.16	0.32	0.13
Q2	0.35	1	0.25	0.43	0.43	-0.2	0.31	0.39	0.19	0.21	0.26	0.15	0.12	-0.2	0.2	0.34	0.57	0.16	0.16	0.23	0.35	0.24	0.19
Q5	0.43	0.25	1	0.26	0.41	-0.34	0.21	0.24	0.06	0.06	0.41	0.16	0.06	-0.4	0.06	0.26	0.23	0.13	0.09	0.47	0.12	0.3	0.14
Q7	0.3	0.43	0.26	1	0.42	-0.2	0.37	0.42	0.13	0.15	0.34	0.14	0.13	-0.3	0.17	0.39	0.5	0.21	0.16	0.31	0.39	0.38	0.25
Q9	0.42	0.43	0.41	0.42	1	-0.4	0.28	0.35	0.15	0.11	0.4	0.1	0.06	-0.42	0.14	0.34	0.41	0.17	0.14	0.42	0.24	0.42	0.08
Q10	-0.3	-0.2	-0.34	-0.2	-0.4	1	-0.2	-0.2	0	-0	-0.47	-0.1	-0.1	0.41	-0.1	-0.1	-0.2	-0.1	-0.1	-0.36	-0.1	-0.3	-0.1
Q11	0.23	0.31	0.21	0.37	0.28	-0.2	1	0.37	0.19	0.31	0.33	0.18	0.18	-0.2	0.27	0.31	0.35	0.24	0.25	0.21	0.32	0.28	0.22
Q12	0.27	0.39	0.24	0.42	0.35	-0.2	0.37	1	0.19	0.2	0.29	0.17	0.15	-0.2	0.21	0.3	0.39	0.21	0.22	0.23	0.36	0.25	0.17
Q13	0	0.19	0.06	0.13	0.15	0	0.19	0.19	1	0.15	0.09	0.04	0.01	-0	0.1	0.16	0.18	0.35	0.09	0.04	0.21	0.03	0.08
Q16	0.1	0.21	0.06	0.15	0.11	-0	0.31	0.2	0.15	1	0.2	0.15	0.13	-0.1	0.21	0.15	0.17	0.13	0.2	0.09	0.2	0.13	0.14
Q17	0.32	0.26	0.41	0.34	0.4	-0.47	0.33	0.29	0.09	0.2	1	0.16	0.12	-0.36	0.16	0.25	0.27	0.21	0.12	0.37	0.18	0.38	0.14
Q18	0.17	0.15	0.16	0.14	0.1	-0.1	0.18	0.17	0.04	0.15	0.16	1	0.18	-0.1	0.19	0.13	0.17	0.13	0.17	0.2	0.21	0.15	0.31
Q19	0.15	0.12	0.06	0.13	0.06	-0.1	0.18	0.15	0.01	0.13	0.12	0.18	1	-0.1	0.49	0.17	0.13	0.05	0.46	0.09	0.11	0.11	0.19
Q20	-0.31	-0.2	-0.4	-0.3	-0.42	0.41	-0.2	-0.2	-0	-0.1	-0.36	-0.1	-0.1	1	-0.1	-0.2	-0.2	-0.1	-0.1	-0.43	-0.1	-0.35	-0
Q21	0.15	0.2	0.06	0.17	0.14	-0.1	0.27	0.21	0.1	0.21	0.16	0.19	0.49	-0.1	1	0.19	0.18	0.11	0.75	0.13	0.19	0.12	0.16
Q22	0.25	0.34	0.26	0.39	0.34	-0.1	0.31	0.3	0.16	0.15	0.25	0.13	0.17	-0.2	0.19	1	0.47	0.23	0.21	0.2	0.28	0.29	0.2
Q23	0.3	0.57	0.23	0.5	0.41	-0.2	0.35	0.39	0.18	0.17	0.27	0.17	0.13	-0.2	0.18	0.47	1	0.26	0.2	0.19	0.37	0.28	0.25
Q24	0.13	0.16	0.13	0.21	0.17	-0.1	0.24	0.21	0.35	0.13	0.21	0.13	0.05	-0.1	0.11	0.23	0.26	1	0.18	0.15	0.25	0.16	0.16
Q25	0.13	0.16	0.09	0.16	0.14	-0.1	0.25	0.22	0.09	0.2	0.12	0.17	0.46	-0.1	0.75	0.21	0.2	0.18	1	0.13	0.2	0.09	0.18
Q26	0.39	0.23	0.47	0.31	0.42	-0.36	0.21	0.23	0.04	0.09	0.37	0.2	0.09	-0.43	0.13	0.2	0.19	0.15	0.13	1	0.16	0.4	0.12
Q27	0.16	0.35	0.12	0.39	0.24	-0.1	0.32	0.36	0.21	0.2	0.18	0.21	0.11	-0.1	0.19	0.28	0.37	0.25	0.2	0.16	1	0.29	0.22
Q28	0.32	0.24	0.3	0.38	0.42	-0.3	0.28	0.25	0.03	0.13	0.38	0.15	0.11	-0.35	0.12	0.29	0.28	0.16	0.09	0.4	0.29	1	0.15
Q30	0.13	0.19	0.14	0.25	0.08	-0.1	0.22	0.17	0.08	0.14	0.14	0.31	0.19	-0	0.16	0.2	0.25	0.16	0.18	0.12	0.22	0.15	1

Figure 6: Correlation matrix after removing variables

(c)

**Kaiser-Meyer-Olkin Test:** The KMO test tells us weather the data is suitable for factor analysis or not. It generates a measure of sampling adequacy (MSA) score out of one. The higher the MSA the better is the data suitable for factor analysis. For the household data we are working on, we get an MSA of **0.89** which is **meritorious**.

**Bartlett test:** a  $p\text{-value} < .01^{***}$  in the Bartlett test tells us that there is significant correlation in the data.

Both the test produces results that tells us that data is ready for a factor analysis.

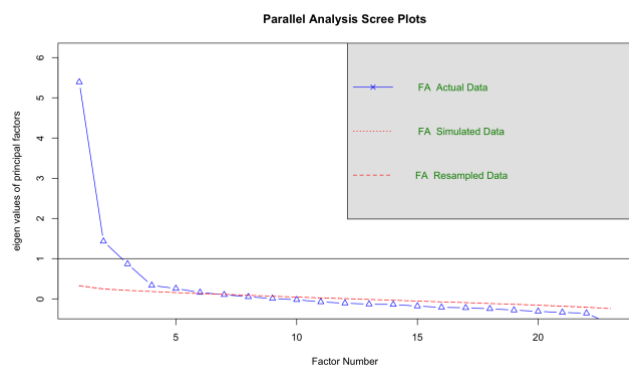
(d)

We run a parallel analysis on the household dataset with 23 *nfactors* and get 23 loadings values. The results are shown in *Table 7* below. As per the initial selectin as per the Kaiser criteria we select the factors with eigenvalues greater then 1. That gives us 3 factors, MR1, MR2 and MR3. Combine these three factors accounts for 38.64% ( $8.887/23 = 0.386$ ) of variance in data. Which is not enough.

*Table 7: Results of Factor Analysis*

	SS Loadings	Proportionate Variance	Cumulative Variance
MR 1	5.692	0.247	0.247
MR 2	1.895	0.082	0.330
MR 3	1.3	0.057	0.386
MR 4	0.669	0.029	0.416
MR 5	0.614	0.027	0.442
MR 6	0.483	0.021	0.463
MR 7	0.397	0.017	0.480
MR 8	0.343	0.015	0.495
MR 9	0.311	0.014	0.509
MR 10	0.226	0.010	0.519
MR 11	0.2	0.009	0.527
MR 12	0.157	0.007	0.534
MR 13	0.148	0.006	0.541
MR 14	0.139	0.006	0.547
MR 15	0.13	0.006	0.552
MR 16	0.114	0.005	0.557
MR 17	0.083	0.004	0.561
MR 18	0.069	0.003	0.564
MR 19	0.042	0.002	0.566
MR 20	0.026	0.001	0.567
MR 21	0.005	0.0	0.567
MR 22	0.001	0.0	0.567
MR 23	0.0	0.0	0.567

Kaiser criteria is not the best way for choosing factors. So, we run a parallel analysis and *Figure 7* shows as that 6 factors are ideal. These 6 factors account for 46.3% of variance in data. So, we go with 6 factors for further analysis.



*Figure 7: Parallel Analysis plot*

(e)

We rerun the factor analysis with 6 factors and the results are shown below in *Table 8*. The 6<sup>th</sup> factor does not produce any results significant results and only contributes 1% of total variance. And overall the 6 factors account for 43% of total variance. *Figure 8* below shows the correlation coefficient of each variable for each factor.

*Table 8: Results of Factor analysis with 6 factors*

	MR 1	MR 2	MR 3	MR 4	MR 5	MR 6
<b>SS Loadings</b>	5.55	1.78	1.18	0.54	0.49	0.33
<b>Propo Variance</b>	0.24	0.08	0.05	0.02	0.02	0.01
<b>Cumulative Variance</b>	0.24	0.32	0.37	0.39	0.41	0.43

(f)

As we can see in *Figure 8* we have failed to achieve simple structure. There are 20 values of more than 0.3 in factor 1 and none in factor 6. Factor 4 and 5 have two values each which is the minimum requirement. Also, numerous rows have more than 1 value above 0.3 but the difference between them is less than 0.2

So, a reduction in factors and rotation is necessary for further factor analysis.

	MR1	MR2	MR3	MR4	MR5	MR6
Q1	0.53	-0.18	-0.14	-0.13	0.03	0.17
Q2	0.6	0.01	0.25	-0.22	-0.11	0.03
Q5	0.53	-0.33	-0.2	0.06	0.06	0.23
Q7	0.64	-0.05	0.19	-0.14	0.01	-0.1
Q9	0.65	-0.27	-0.04	-0.06	-0.21	0.04
Q10	-0.43	0.31	0.31	-0.1	0.05	0.17
Q11	0.55	0.13	0.11	0.12	0.05	-0.22
Q12	0.56	0.06	0.17	-0.01	-0.03	-0.06
Q13	0.24	0.14	0.33	0.4	-0.19	0.12
Q16	0.3	0.19	0.08	0.12	0.07	-0.18
Q17	0.58	-0.22	-0.16	0.19	0.03	-0.18
Q18	0.32	0.13	-0.04	0.05	0.38	0.07
Q19	0.3	0.41	-0.26	-0.08	0.07	-0.01
Q20	-0.5	0.31	0.22	-0.02	0.07	0.04
Q21	0.42	0.67	-0.38	-0.02	-0.14	-0.01
Q22	0.53	0.06	0.18	-0.1	-0.04	0.07
Q23	0.64	0.05	0.35	-0.24	-0.07	0.05
Q24	0.35	0.11	0.23	0.36	-0.03	0.15
Q25	0.41	0.65	-0.33	0.01	-0.13	0.08
Q26	0.54	-0.29	-0.26	0.08	0.08	0.13
Q27	0.48	0.15	0.28	0.01	0.07	-0.07
Q28	0.54	-0.19	-0.06	-0.01	0.07	-0.09
Q30	0.33	0.19	0.11	-0.01	0.41	0.07

*Figure 8: Correlation coefficient of each factor*

(g)

**Varimax rotation:** After applying varimax rotation the results are better. The first factor has 9 values with correlation more than 0.3 and the last one has 1 which is still below the minimum requirement of 2. The results are shown below in *Figure 9*.

	MR3	MR1	MR2	MR4	MR5	MR6
Q1	0.5	0.29	0.1	-0.01	0.14	-0.14
Q2	0.22	0.64	0.09	0.09	0.05	0
Q5	0.64	0.14	0.01	0.1	0.16	-0.15
Q7	0.31	0.57	0.06	0.07	0.14	0.17
Q9	0.58	0.44	0.05	0.11	-0.09	0
Q10	-0.61	-0.04	-0.05	0.03	0.05	-0.19
Q11	0.22	0.35	0.18	0.2	0.17	0.35
Q12	0.23	0.46	0.13	0.18	0.11	0.16
Q13	-0.02	0.17	0.02	0.6	-0.02	0.06
Q16	0.05	0.17	0.17	0.15	0.15	0.28
Q17	0.59	0.16	0.07	0.12	0.1	0.29
Q18	0.14	0.09	0.15	0.05	0.47	0.05
Q19	0.06	0.09	0.53	-0.05	0.18	0.06
Q20	-0.6	-0.17	-0.04	0	0.03	-0.07
Q21	0.07	0.13	0.87	0.07	0.06	0.09
Q22	0.19	0.49	0.13	0.15	0.13	0
Q23	0.16	0.73	0.08	0.13	0.12	0.01
Q24	0.1	0.17	0.06	0.53	0.14	0.05
Q25	0.05	0.12	0.83	0.13	0.09	0.02
Q26	0.65	0.11	0.06	0.06	0.16	-0.05
Q27	0.07	0.45	0.1	0.22	0.21	0.19
Q28	0.47	0.29	0.03	0.02	0.14	0.15
Q30	0.03	0.21	0.11	0.06	0.52	0.06

Figure 9: Correlation coefficient for Varimax rotation

**Oblimin rotation:** Oblimin rotation improves the results even more the results are shown in *Figure 10*. Every factor has the minimum of 2 values above 0.3, but it still fails to reach a simple structure. Question 9 and Question 10 has 2 correlation values of above 0.3 (0.41 and 0.37 for question 9 and -.33 and -.43 for question 10) and the difference between them is less than 0.2. Question 16 and Question 28 has no values above 0.3.

	MR1	MR2	MR3	MR6	MR4	MR5
Q1	0.25	0.06	0.49	-0.05	-0.07	0.05
Q2	0.69	0.03	0.04	-0.03	0	-0.04
Q5	0.02	-0.02	0.67	0	0.07	0.06
Q7	0.58	-0.02	0.02	0.19	-0.02	0.08
Q9	0.41	0.02	0.37	0.13	0.05	-0.21
Q10	0.05	-0.04	-0.33	-0.43	0.06	0.1
Q11	0.25	0.11	-0.12	0.38	0.15	0.13
Q12	0.41	0.07	0	0.15	0.11	0.05
Q13	0.01	0	-0.03	-0.02	0.63	-0.07
Q16	0.09	0.12	-0.17	0.27	0.11	0.14
Q17	0.02	0.02	0.26	0.48	0.08	0.04
Q18	-0.03	0.08	0.18	0.03	0.01	0.46
Q19	0.01	0.53	0.02	0.03	-0.12	0.14
Q20	-0.11	-0.02	-0.4	-0.26	0.04	0.1
Q21	0	0.9	-0.04	0.04	-0.02	-0.02
Q22	0.47	0.07	0.09	-0.04	0.08	0.05
Q23	0.78	0	-0.01	-0.07	0.03	0.03
Q24	-0.01	0.02	0.11	0	0.55	0.08
Q25	-0.01	0.85	0.03	-0.06	0.05	0
Q26	-0.02	0.03	0.6	0.13	0.03	0.08
Q27	0.4	0.02	-0.13	0.12	0.16	0.17
Q28	0.23	-0.03	0.24	0.27	-0.04	0.08
Q30	0.12	0.02	0.07	-0.02	0.01	0.51

Figure 10: Correlation coefficient with Oblimin rotation

Both the rotation produces better results than no rotation but Oblimin rotation is better of the two. It's still needs more work. Reducing the number of factors and removing some variables will be required to produce simple structure.

(h)

Figure 11 shows us the residual values from factor analysis with 6 factors and Oblimin rotation. Ideally, we want residual values of greater than 0.05 to be less than 10%. In this case, the percentage of residual value above 0.05 is 1.19% ( $6/506 * 100$ ). So even though we are not able to achieve a simple structure we can still get residual percentage within limits.

	Q1	Q2	Q5	Q7	Q9	Q10	Q11	Q12	Q13	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q30
Q1	0.62	0.04	0.03	-0.03	0.01	0.03	0.02	0.02	-0.02	0.03	0	0	0.02	0.03	0	-0.02	-0.02	0.01	-0.02	0	-0.01	0	-0.02
Q2	0.04	0.52	0	-0.03	0.01	-0.02	-0.01	0.01	0.03	0.05	0	0.02	0	0	0.02	-0.05	0.04	-0.04	-0.03	0	0	-0.05	0
Q5	0.03	0	0.52	-0.02	-0.02	0.01	0.02	0.02	0.01	0.01	0.03	-0.02	-0.02	-0.01	-0.01	0.03	-0.01	-0.02	0.01	0	-0.02	-0.04	0.01
Q7	-0.03	-0.03	-0.02	0.52	0	0	-0.01	0.02	-0.01	-0.05	-0.01	-0.04	0	0	0	0.01	-0.01	0.01	0.01	0.03	0.03	0.03	0.03
Q9	0.01	0.01	-0.02	0	0.45	-0.03	-0.02	0	0.02	0	-0.02	0.01	-0.02	0.01	0	0.01	-0.02	-0.01	0.01	0	0	0.03	0.01
Q10	0.03	-0.02	0.01	0	-0.03	0.58	0.01	0.03	-0.01	0.04	-0.06	-0.02	-0.02	0.02	0.01	0.05	-0.04	0	-0.01	0.02	0.01	0.04	-0.03
Q11	0.02	-0.01	0.02	-0.01	-0.02	0.01	0.6	0.02	-0.01	0.06	-0.01	-0.02	-0.01	-0.01	0	0.02	-0.01	0	0	-0.01	-0.02	-0.01	0
Q12	0.02	0.01	0.02	0.02	0	0.03	0.02	0.65	0	0	0	0.01	0	0	-0.01	-0.03	-0.03	-0.02	0.01	0	0.03	-0.03	-0.03
Q13	-0.02	0.03	0.01	-0.01	0.02	-0.01	-0.01	0	0.61	0.01	-0.01	0.01	0.02	-0.01	0.02	-0.01	-0.01	0	-0.03	0	0	-0.03	0.01
Q16	0.03	0.05	0.01	-0.05	0	0.04	0.06	0	0.01	0.82	0.03	0.02	-0.01	0	0	-0.01	-0.01	-0.03	0	0.01	-0.01	-0.01	-0.02
Q17	0	0	0.03	-0.01	-0.02	-0.06	-0.01	0	-0.01	0.03	0.52	-0.01	0.01	0.04	0.01	0.02	0.03	0.03	-0.01	-0.04	-0.04	0	0.01
Q18	0	0.02	-0.02	-0.04	0.01	-0.02	-0.02	0.01	0.01	0.02	-0.01	0.73	0	-0.02	0.01	-0.03	0.01	-0.01	-0.01	0.01	0.02	-0.02	0.02
Q19	0.02	0	-0.02	0	-0.02	-0.02	-0.01	0	0.02	-0.01	0.01	0	0.67	-0.01	0	0.03	0	0	0	-0.02	-0.02	0.01	0.01
Q20	0.03	0	-0.01	0	0.01	0.02	-0.01	0	-0.01	0	0.04	-0.02	-0.01	0.6	0	0	-0.01	0.01	0.01	-0.02	0	-0.01	0.02
Q21	0	0.02	-0.01	0	0	0.01	0	-0.01	0.02	0	0.01	0.01	0	0	0.21	-0.01	-0.01	-0.02	0	0.01	0	0.01	-0.01
Q22	-0.02	-0.05	0.03	0.01	0.01	0.05	0.02	-0.03	-0.01	-0.01	0.02	-0.03	0.03	0	-0.01	0.67	0.04	0.01	0	-0.02	-0.02	0.03	0
Q23	-0.02	0.04	-0.01	-0.01	-0.02	-0.04	-0.01	-0.03	-0.01	-0.01	0.03	0.01	0	-0.01	-0.01	0.04	0.4	0.03	0.01	-0.03	-0.03	-0.02	0.02
Q24	0.01	-0.04	-0.02	0.01	-0.01	0	0	-0.02	0	-0.03	0.03	-0.01	0	0.01	-0.02	0.01	0.03	0.65	0.03	0	0.01	0.02	0
Q25	-0.02	-0.03	0.01	0.01	0.01	-0.01	0	0.01	-0.03	0	-0.01	-0.01	0	0.01	0	0	0.01	0.03	0.28	0.01	0.01	-0.01	0
Q26	0	0	0	0.03	0	0.02	-0.01	0	0	0.01	-0.04	0.01	-0.02	-0.02	0.01	-0.02	-0.03	0	0.01	0.53	0.02	0.05	-0.02
Q27	-0.01	0	-0.02	0.03	0	0.01	-0.02	0.03	0	-0.01	-0.04	0.02	-0.02	0	0	-0.02	-0.03	0.01	0.01	0.02	0.66	0.07	-0.02
Q28	0	-0.05	-0.04	0.03	0.03	0.04	-0.01	-0.03	-0.03	-0.01	0	-0.02	0.01	-0.01	0.01	0.03	-0.02	0.02	-0.01	0.05	0.07	0.66	-0.01
Q30	-0.02	0	0.01	0.03	0.01	-0.03	0	-0.03	0.01	-0.02	0.01	0.02	0.01	0.02	-0.01	0	0.02	0	0	-0.02	-0.02	-0.01	0.67

Figure 11: residual values

(i)

To achieve simple structure, we start by running a factor analysis with 5 factors and Oblimin rotation. The results are below in *Figure 12*. Question 9 have 2 values above 0.3 correlation and the difference between them is less than 0.2. So, we remove Question 9 and run the analysis again.

	MR3	MR1	MR2	MR4	MR5	
Q1	0.43	0.24	0.04	-0.13	0.05	
Q2	0.02	0.68	0.03	-0.01	-0.04	
Q5	0.63	0.02	-0.05	-0.01	0.06	
Q7	0.16	0.54	-0.01	0.03	0.09	
Q9	0.5	0.36	0.01	0.04	-0.19	
Q10	-0.66	0.11	-0.03	0.01	0.06	
Q11	0.16	0.21	0.13	0.24	0.15	
Q12	0.11	0.39	0.08	0.16	0.06	
Q13	-0.05	0.02	0	0.61	-0.07	
Q16	0.02	0.08	0.14	0.19	0.15	
Q17	0.61	-0.03	0.02	0.15	0.07	
Q18	0.13	-0.02	0.07	0	0.46	
Q19	0.02	0.02	0.53	-0.11	0.14	
Q20	-0.61	-0.06	-0.01	0.03	0.08	
Q21	0	0	0.91	-0.01	-0.03	
Q22	0.05	0.47	0.07	0.07	0.05	
Q23	-0.06	0.79	0	0.02	0.03	
Q24	0.08	0.01	0.01	0.51	0.08	
Q25	-0.01	0	0.83	0.04	0	
Q26	0.68	-0.05	0.01	-0.01	0.09	
Q27	-0.05	0.39	0.03	0.21	0.18	
Q28	0.44	0.18	-0.03	0	0.1	
Q30	-0.02	0.15	0.02	0	0.51	

Figure 12: factor analysis with 5 factors

After removing Question 9 we notice (Figure 13) that Question 11 and 16 has no values above 0.3 so we remove those 2 variables.

	MR3	MR1	MR2	MR4	MR5
Q1	0.42	0.24	0.04	-0.13	0.05
Q2	0.03	0.69	0.03	-0.02	-0.04
Q5	0.63	0.03	-0.05	-0.01	0.05
Q7	0.16	0.56	-0.02	0.02	0.06
Q10	-0.65	0.08	-0.03	0.01	0.07
Q11	0.16	0.23	0.13	0.23	0.11
Q12	0.11	0.4	0.07	0.15	0.04
Q13	-0.04	0.04	0	0.59	-0.06
Q16	0.01	0.08	0.14	0.19	0.13
Q17	0.61	0	0.03	0.15	0.03
Q18	0.09	-0.07	0.06	0	0.51
Q19	0.02	0.01	0.53	-0.11	0.14
Q20	-0.62	-0.09	-0.01	0.03	0.11
Q21	0	0	0.9	-0.01	-0.03
Q22	0.05	0.47	0.06	0.06	0.04
Q23	-0.06	0.8	-0.01	0	0.01
Q24	0.08	0.01	0.01	0.51	0.08
Q25	-0.01	0	0.83	0.04	0
Q26	0.66	-0.04	0	-0.01	0.09
Q27	-0.06	0.38	0.02	0.21	0.17
Q28	0.43	0.19	-0.03	0	0.09
Q30	-0.06	0.09	0.01	0	0.55

Figure 13: Factor analysis results after removing Question 9



After removing question 11 and 16 we finally get simple structure (*Figure 14*). Each column has at least 2 coefficients above 0.3 and no row has more than 1. The root mean square of the residuals (RMSR) is 0.02, (it should be below 0.06). TLI is 0.927 (should be between 0.9 and 1). RMSEA is 0.046 (should be below 0.08).

	MR3	MR1	MR2	MR4	MR5
Q1	0.43	0.22	0.04	-0.1	0.06
Q2	0.03	0.7	0.02	-0.03	-0.05
Q5	0.63	0.02	-0.05	0.02	0.06
Q7	0.16	0.56	-0.02	0.01	0.07
Q10	-0.64	0.07	-0.04	0.02	0.07
Q12	0.11	0.42	0.08	0.12	0.02
Q13	-0.06	0.08	0.01	0.56	-0.07
Q17	0.6	0.03	0.03	0.11	0.01
Q18	0.11	-0.05	0.07	0.01	0.48
Q19	0.02	0.01	0.53	-0.11	0.14
Q20	-0.62	-0.09	-0.01	0.04	0.11
Q21	0	0.01	0.9	-0.02	-0.03
Q22	0.06	0.47	0.06	0.07	0.04
Q23	-0.05	0.79	-0.01	0.01	0.02
Q24	0.08	-0.01	0.02	0.6	0.07
Q25	-0.01	-0.01	0.84	0.05	0
Q26	0.67	-0.05	0.01	0.02	0.09
Q27	-0.06	0.41	0.03	0.18	0.15
Q28	0.43	0.19	-0.03	0	0.09
Q30	-0.05	0.09	0.01	0.01	0.56

Figure 14: Simple structure

Residual value (*Figure 15*) is at 3.16% ( $12/380 * 100$ ). Its, below the 10% maximum limit so the result is acceptable.

	Q1	Q2	Q5	Q7	Q10	Q12	Q13	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q30
Q1	0.66	0.05	0.08	-0.04	0.05	0.01	-0.02	-0.03	0	0.01	0.03	0	0	0	0.03	0	0.02	-0.03	0	-0.02
Q2	0.05	0.53	0.01	-0.03	-0.02	0.01	0.04	-0.01	0.02	-0.01	0	0.02	-0.04	0.04	-0.02	-0.02	0.01	0	-0.05	0
Q5	0.08	0.01	0.57	-0.04	0.03	0	0.02	0	-0.01	-0.02	0	-0.01	0.05	0	-0.01	0.02	0.03	-0.04	-0.06	0.02
Q7	-0.04	-0.03	-0.04	0.53	-0.02	0.03	-0.02	0.01	-0.04	0	-0.01	0	0.01	-0.02	-0.01	0	0.01	0.04	0.04	0.02
Q10	0.05	-0.02	0.03	-0.02	0.63	0.02	-0.01	-0.1	0	-0.01	0.03	0.01	0.06	-0.02	0.02	0	0.03	0	0.02	-0.02
Q12	0.01	0.01	0	0.03	0.02	0.67	0.01	0.02	0.02	0	0	-0.01	-0.03	-0.04	-0.02	0.01	0	0.04	-0.02	-0.02
Q13	-0.02	0.04	0.02	-0.02	-0.01	0.01	0.67	-0.01	0	0.01	-0.01	0.02	-0.01	-0.01	0	-0.03	0	0	-0.04	0.01
Q17	-0.03	-0.01	0	0.01	-0.1	0.02	-0.01	0.57	0	0.01	0.04	0.01	0.01	0.01	0.01	-0.02	-0.06	-0.02	0.02	0.01
Q18	0	0.02	-0.01	-0.04	0	0.02	0	0	0.73	0	-0.02	0.01	-0.03	0.01	-0.01	-0.02	0.01	0.03	-0.03	0.01
Q19	0.01	-0.01	-0.02	0	-0.01	0	0.01	0.01	0	0.67	-0.01	0	0.03	-0.01	0	0	-0.03	-0.02	0	0.01
Q20	0.03	0	0	-0.01	0.03	0	-0.01	0.04	-0.02	-0.01	0.59	0	0	-0.01	0.01	0	-0.01	0	-0.02	0.01
Q21	0	0.02	-0.01	0	0.01	-0.01	0.02	0.01	0.01	0	0	0.21	-0.01	-0.01	-0.02	0	0	0	0.01	0
Q22	0	-0.04	0.05	0.01	0.06	-0.03	-0.01	0.01	-0.03	0.03	0	-0.01	0.68	0.04	0.01	0.01	-0.01	-0.03	0.03	0
Q23	0	0.04	0	-0.02	-0.02	-0.04	-0.01	0.01	0.01	-0.01	-0.01	-0.01	0.04	0.4	0.03	0.01	-0.02	-0.04	-0.03	0.02
Q24	0.03	-0.02	-0.01	-0.01	0.02	-0.02	0	0.01	-0.01	0	0.01	-0.02	0.01	0.03	0.59	0.02	0	-0.01	0.01	0
Q25	0	-0.02	0.02	0	0	0.01	-0.03	-0.02	-0.02	0	0	0	0.01	0.01	0.02	0.28	0.02	0.01	-0.01	0
Q26	0.02	0.01	0.03	0.01	0.03	0	0	-0.06	0.01	-0.03	-0.01	0	-0.01	-0.02	0	0.02	0.54	0	0.04	-0.02
Q27	-0.03	0	-0.04	0.04	0	0.04	0	-0.02	0.03	-0.02	0	0	-0.03	-0.04	-0.01	0.01	0	0.67	0.08	-0.02
Q28	0	-0.05	-0.06	0.04	0.02	-0.02	-0.04	0.02	-0.03	0	-0.02	0.01	0.03	-0.03	0.01	-0.01	0.04	0.08	0.67	-0.02
Q30	-0.02	0	0.02	0.02	-0.02	-0.02	0.01	0.01	0.01	0.01	0.01	0	0	0.02	0	0	-0.02	-0.02	-0.02	0.64

Figure 15: Residual values

We can check the reliability of our factors by looking at alpha values of all variables in each factor. An alpha score of 0.7 is considered acceptable. *Table 9* shows that only 2 factors has value above 0.7.

Table 9: Alpha values of each factors.

Factor	alpha
MR 3	0.158
MR 1	0.796
MR 2	0.800
MR 4	0.519
MR 5	0.461

## APPENDIX 3

(a)

# loading data in R and looking at dimensions and head

```
household <- read.csv(file.choose())
head(household)
dim(household)
```

# looking for missing values in dataset and row

```
> sum(is.na(household))
[1] 136
> View(household)
> sum(is.na(household[,1]))
[1] 0
> sum(is.na(household[,2]))
[1] 136
```

# extracting the 30 questions

```
household <- household[,3:32]
```

# new Dimensions

```
> dim(household)
[1] 1001 30
```

# missing values

```
> sum(is.na(household))
[1] 0
head(household)
```

# Mahalanobis outliers

```
md <- mahalanobis(household, colMeans(household), cov(household))
match(boxplot(md)$out, md)
[1] 91 93 175 186 319 334 373 386 388 404 468 489 619 724 774 785
[17] 823 878 893 909 932 945 962 981 992 994 999
```

(b)

# Correlation Matrix, saving it to observe further

```
household_cor <- round(cor(household), 2)
write.csv(household_cor, "household_cor.csv")
```

# We remove variables with no correlation and check again

```
household_2 <- household[, -c(3,4,6,8,14,15,29)]
household_2_cor <- round(cor(household_2), 2)
write.csv(household_2_cor, "household_2_cor.csv")
household <- household_2
dim(household)
```

```
[1] 1001 23
```

(c)

# Kaiser-Meyer-Olkin (KMO) sample adequacy test

```
KMO(household)
```

```
> KMO(household) #sample adequacy
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = household)
Overall MSA = 0.89
MSA for each item =
  Q1  Q2  Q5  Q7  Q9  Q10  Q11  Q12  Q13  Q16  Q17  Q18  Q19  Q20
0.93 0.90 0.90 0.94 0.93 0.86 0.94 0.95 0.75 0.88 0.91 0.87 0.88 0.93
  Q21  Q22  Q23  Q24  Q25  Q26  Q27  Q28  Q30
0.72 0.93 0.89 0.83 0.72 0.91 0.92 0.93 0.86
```

# Bartlett test

```
bartlett.test(household)
```

```
round(bartlett.test(household)$p.value, 10)
```

```
> bartlett.test(household) # p < .05

      Bartlett test of homogeneity of variances

data:  household
Bartlett's K-squared = 232.99, df = 22, p-value < 2.2e-16

> round(bartlett.test(household)$p.value, 10)
[1] 0
```

(d)

# Factor analysis with 23 factors, minres method and no rotation

dim(household)

household\_fa23 &lt;- fa(household, nfactors = 23, fm = "minres", rotate = "none")

household\_fa23\$loadings

Loadings:

	MR1	MR2	MR3	MR4	MR5	MR6	MR7	MR8	MR9	MR10	MR11	MR12	MR13	MR14	MR15	MR16	MR17	MR18	MR19	MR20	MR21	MR22	MR23
Q1	0.540	-0.195	-0.136	-0.160		0.214		0.168			0.208												
Q2	0.605		0.269	-0.275	-0.108		-0.250		-0.116	0.114					-0.102								
Q5	0.534	-0.351	-0.203			0.307	-0.106		0.146	-0.139			0.111			-0.117							
Q7	0.647		0.206	-0.137		-0.124	0.143		-0.159					0.124									
Q9	0.657	-0.274			-0.218										0.237								
Q10	-0.443	0.354	0.344	-0.126		0.239	0.176	0.227	0.154							0.103							
Q11	0.557	0.138	0.102	0.145		-0.201		0.116	0.147				-0.125	0.111									
Q12	0.567		0.179					0.162	-0.281														
Q13	0.241	0.148	0.318	0.380	-0.238	0.149				-0.107	0.147												
Q16	0.306	0.207		0.161		-0.212	-0.118	0.264	0.132	0.142													
Q17	0.593	-0.236	-0.170	0.239		-0.201	-0.106		0.189		0.178		0.114										
Q18	0.320	0.142			0.425		-0.106	-0.159						-0.166									
Q19	0.299	0.414	-0.294								0.233						0.125						
Q20	-0.499	0.317	0.208								0.200		0.134										
Q21	0.420	0.650	-0.405		-0.136												-0.169						
Q22	0.538		0.198	-0.122		0.102	0.111	-0.140	0.277		-0.124			-0.109									
Q23	0.641		0.354	-0.251			-0.107	-0.185				-0.114											
Q24	0.360	0.126	0.240	0.412		0.219		-0.124			0.140		-0.138										
Q25	0.413	0.665	-0.383		-0.154							-0.174											
Q26	0.539	-0.299	-0.251			0.149	0.118	0.102	-0.111							0.106			0.101				
Q27	0.487	0.175	0.296				0.154	-0.177								-0.192							
Q28	0.556	-0.210			0.101	-0.109	0.379			0.175								-0.114					
Q30	0.334	0.199	0.101		0.452		-0.176							0.148									

	MR1	MR2	MR3	MR4	MR5	MR6	MR7	MR8	MR9	MR10	MR11	MR12
SS loadings	5.692	1.895	1.300	0.669	0.614	0.483	0.397	0.343	0.311	0.226	0.200	0.157
Proportion Var	0.247	0.082	0.057	0.029	0.027	0.021	0.017	0.015	0.014	0.010	0.009	0.007
Cumulative Var	0.247	0.330	0.386	0.416	0.442	0.463	0.480	0.495	0.509	0.519	0.527	0.534
	MR13	MR14	MR15	MR16	MR17	MR18	MR19	MR20	MR21	MR22	MR23	
SS loadings	0.148	0.139	0.130	0.114	0.083	0.069	0.042	0.026	0.005	0.001	0.000	
Proportion Var	0.006	0.006	0.006	0.005	0.004	0.003	0.002	0.001	0.000	0.000	0.000	
Cumulative Var	0.541	0.547	0.552	0.557	0.561	0.564	0.566	0.567	0.567	0.567	0.567	

# Parallel analysis

fa.parallel(household, fa = "fa")

Parallel analysis suggests that the number of factors = 6 and the number of components = NA

(e)

# Factor analysis with 6 factors, minres method and no rotation

household\_fa6 &lt;- fa(household, nfactors = 6, fm = "minres", rotate = "none")

temp &lt;- household\_fa6\$loadings

write.csv(round(temp, 2), "q3\_d.csv")

```

Factor Analysis using method = minres
Call: fa(r = household, nfactors = 6, rotate = "none", fm = "minres")
Standardized loadings (pattern matrix) based upon correlation matrix
      MR1  MR2  MR3  MR4  MR5  MR6  h2  u2  com
Q1   0.53 -0.18 -0.14 -0.13  0.03  0.17 0.38 0.62 1.8
Q2   0.60  0.01  0.25 -0.22 -0.11  0.03 0.48 0.52 1.7
Q5   0.53 -0.33 -0.20  0.06  0.06  0.23 0.48 0.52 2.5
Q7   0.64 -0.05  0.19 -0.14  0.01 -0.10 0.48 0.52 1.4
Q9   0.65 -0.27 -0.04 -0.06 -0.21  0.04 0.55 0.45 1.6
Q10 -0.43  0.31  0.31 -0.10  0.05  0.17 0.42 0.58 3.2
Q11  0.55  0.13  0.11  0.12  0.05 -0.22 0.40 0.60 1.6
Q12  0.56  0.06  0.17 -0.01 -0.03 -0.06 0.35 0.65 1.2
Q13  0.24  0.14  0.33  0.40 -0.19  0.12 0.39 0.61 3.7
Q16  0.30  0.19  0.08  0.12  0.07 -0.18 0.18 0.82 3.1
Q17  0.58 -0.22 -0.16  0.19  0.03 -0.18 0.48 0.52 1.9
Q18  0.32  0.13 -0.04  0.05  0.38  0.07 0.27 0.73 2.3
Q19  0.30  0.41 -0.26 -0.08  0.07 -0.01 0.33 0.67 2.8
Q20 -0.50  0.31  0.22 -0.02  0.07  0.04 0.40 0.60 2.2
Q21  0.42  0.67 -0.38 -0.02 -0.14 -0.01 0.79 0.21 2.5
Q22  0.53  0.06  0.18 -0.10 -0.04  0.07 0.33 0.67 1.4
Q23  0.64  0.05  0.35 -0.24 -0.07  0.05 0.60 0.40 1.9
Q24  0.35  0.11  0.23  0.36 -0.03  0.15 0.35 0.65 3.3
Q25  0.41  0.65 -0.33  0.01 -0.13  0.08 0.72 0.28 2.4
Q26  0.54 -0.29 -0.26  0.08  0.08  0.13 0.47 0.53 2.3
Q27  0.48  0.15  0.28  0.01  0.07 -0.07 0.34 0.66 2.0
Q28  0.54 -0.19 -0.06 -0.01  0.07 -0.09 0.34 0.66 1.4
Q30  0.33  0.19  0.11 -0.01  0.41  0.07 0.33 0.67 2.6

```

	MR1	MR2	MR3	MR4	MR5	MR6
SS loadings	5.55	1.78	1.18	0.54	0.49	0.33
Proportion Var	0.24	0.08	0.05	0.02	0.02	0.01
Cumulative Var	0.24	0.32	0.37	0.39	0.41	0.43
Proportion Explained	0.56	0.18	0.12	0.05	0.05	0.03
Cumulative Proportion	0.56	0.74	0.86	0.92	0.97	1.00

**(g)**

**# Factor analysis with 6 factors, minres method and varimax rotation**

```
household_fa6_v <- fa(household, nfactors = 6, fm = "minres", rotate = "varimax")
temp <- household_fa6_v$loadings
write.csv(round(temp, 2), "q3_f.csv")
```

**# Factor analysis with 6 factors, minres method and oblimin rotation**

```
household_fa6_o <- fa(household, nfactors = 6, fm = "minres", rotate = "oblimin")
temp <- household_fa6_o$loadings
write.csv(round(temp, 2), "q3_f2.csv")
```

**(h)**

**# getting the residuals figures from factor analysis with**

```
temp <- round(household_fa6_o$residual, 2)
write.csv(temp, "q3_h_residual.csv")
```

(i)

# Factor analysis with 5 factors, minres method and oblimin rotation

household\_fa5\_o &lt;- fa(household, nfactors = 5, fm = "minres", rotate = "oblimin")

household\_fa5\_o\$loadings

temp &lt;- household\_fa5\_o\$loadings

write.csv(round(temp, 2), "q3\_i1.csv")

```

Loadings:
      MR3  MR1  MR2  MR4  MR5
Q1   0.432  0.239      -0.131
Q2      0.684
Q5   0.633
Q7   0.165  0.544
Q9   0.500  0.363      -0.188
Q10 -0.663  0.108
Q11  0.163  0.212  0.133  0.237  0.148
Q12  0.112  0.391      0.155
Q13      0.607
Q16      0.139  0.191  0.148
Q17  0.610      0.151
Q18  0.133      0.457
Q19      0.530 -0.111  0.145
Q20 -0.614
Q21      0.905
Q22      0.468
Q23      0.786
Q24      0.513
Q25      0.833
Q26  0.678
Q27      0.389      0.212  0.181
Q28  0.441  0.183      0.105
Q30      0.147      0.511

```

```

      MR3  MR1  MR2  MR4  MR5
SS loadings  2.784 2.218 1.859 0.856 0.673
Proportion Var 0.121 0.096 0.081 0.037 0.029
Cumulative Var 0.121 0.217 0.298 0.335 0.365

```



## # removing question 9

```
household_2 <- household[,-5]
head(household_2)
```

## # Factor analysis with 5 factors, minres method and oblimin rotation

```
household_2_fa5_o <- fa(household_2, nfactors = 5, fm = "minres", rotate = "oblimin")
temp <- household_2_fa5_o$loadings
write.csv(round(temp, 2), "q3_i2.csv")
```

Loadings:					
	MR3	MR1	MR2	MR4	MR5
Q1	0.421	0.239		-0.132	
Q2		0.686			
Q5	0.625				
Q7	0.163	0.557			
Q10	-0.651				
Q11	0.162	0.229	0.135	0.232	0.108
Q12	0.113	0.402		0.148	
Q13				0.591	
Q16			0.140	0.189	0.128
Q17	0.607			0.151	
Q18					0.511
Q19			0.527	-0.110	0.140
Q20	-0.619				0.113
Q21			0.905		
Q22		0.474			
Q23		0.800			
Q24				0.512	
Q25			0.830		
Q26	0.663				
Q27		0.381		0.206	0.175
Q28	0.426	0.189			
Q30					0.550

### # removing question 11 and 16

```
household_3 <- household_2[, -c(6,9)]
head(household_3)
```

### # Factor analysis with 5 factors, minres method and oblimin rotation

```
household_3_fa5 <- fa(household_3, nfactors = 5, fm = "minres", rotate = "oblimin")
temp <- household_3_fa5$loadings
write.csv(round(temp, 2), "q3_i3.csv")
household_3_fa5
```

```
Mean item complexity = 1.2
Test of the hypothesis that 5 factors are sufficient.

The degrees of freedom for the null model are 190 and the objective function was
5.7 with Chi Square of 5655.08
The degrees of freedom for the model are 100 and the objective function was 0.31

The root mean square of the residuals (RMSR) is 0.02
The df corrected root mean square of the residuals is 0.03

The harmonic number of observations is 1001 with the empirical chi square 217.07
with prob < 1.2e-10
The total number of observations was 1001 with Likelihood Chi Square = 310.6 w
ith prob < 2e-23

Tucker Lewis Index of factoring reliability = 0.927
RMSEA index = 0.046 and the 90 % confidence intervals are 0.04 0.052
BIC = -380.28
Fit based upon off diagonal values = 0.99
Measures of factor score adequacy
```

### # residuals

```
temp<-round(household_3_fa5$residual,2)
write.csv(temp, "q3_i_residual.csv")
```

**# alpha factor 1**

```
factor1 <- household_3[, c(1, 3, 5, 8, 11, 17, 19)]
```

```
alpha(factor1)$total
```

```
> factor1 <- household_3[, c(1, 3, 5, 8, 11, 17, 19)]
> alpha(factor1)$total
Some items ( Q10 Q20 ) were negatively correlated with the total scale and
probably should be reversed.
To do this, run the function again with the 'check.keys=TRUE' option
raw_alpha std.alpha  G6(smc) average_r      S/N      ase
0.1584961 0.1814674 0.4214794 0.03069893 0.2216984 0.03779129
      mean      sd
2.183103 0.3644553
Warning message:
In alpha(factor1) :
  Some items were negatively correlated with the total scale and probably
should be reversed.
To do this, run the function again with the 'check.keys=TRUE' option
```

**# alpha factor 2**

```
factor2 <- household_3[, c(2, 4, 6, 13, 14, 18)]
```

```
alpha(factor2)$total
```

```
> factor2 <- household_3[, c(2, 4, 6, 13, 14, 18)]
> alpha(factor2)$total
raw_alpha std.alpha  G6(smc) average_r      S/N      ase
0.796286 0.7968496 0.7741605 0.3953114 3.922463 0.009916603
      mean      sd
2.066101 0.6448848
```

**# alpha factor 3**

```
factor3 <- household_3[, c(10, 12, 16)]
```

```
alpha(factor3)$total
```

```
> factor3 <- household_3[, c(10, 12, 16)]
> alpha(factor3)$total
raw_alpha std.alpha  G6(smc) average_r      S/N      ase
0.8000294 0.7979915 0.7548363 0.5683631 3.950286 0.01094236
      mean      sd
1.811855 0.80181
```

## # alpha factor 4

factor4 &lt;- household\_3[, c(7, 15)]

alpha(factor4)\$total

```

> factor4 <- household_3[, c(7, 15)]
> alpha(factor4)$total
raw_alpha std.alpha  G6(smc) average_r      S/N      ase
0.5190737 0.5192826 0.3506966 0.3506966 1.080224 0.03038242
      mean      sd
2.275225 0.737174

```

## # alpha factor 5

factor5 &lt;- household\_3[, c(9, 20)]

alpha(factor5)\$total

```

> factor5 <- household_3[, c(9, 20)]
> alpha(factor5)$total
raw_alpha std.alpha  G6(smc) average_r      S/N      ase
0.4606219 0.4692853 0.3065792 0.3065792 0.8842517 0.03319692
      mean      sd
1.954545 0.7798601

```