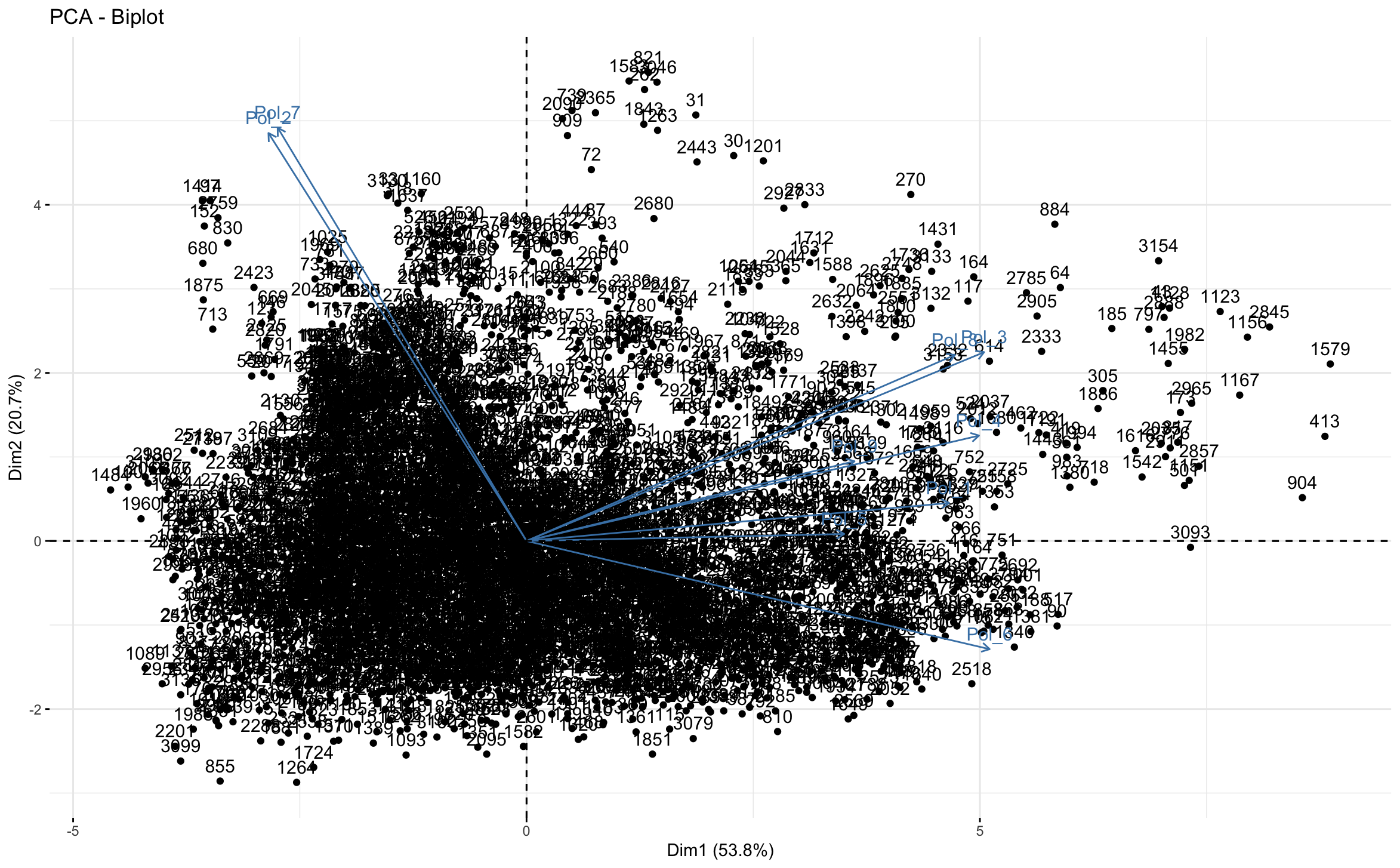
**Question 1**

The Excel data set Pollution.data.xlsx contains measurements on 9 pollutants made daily at a

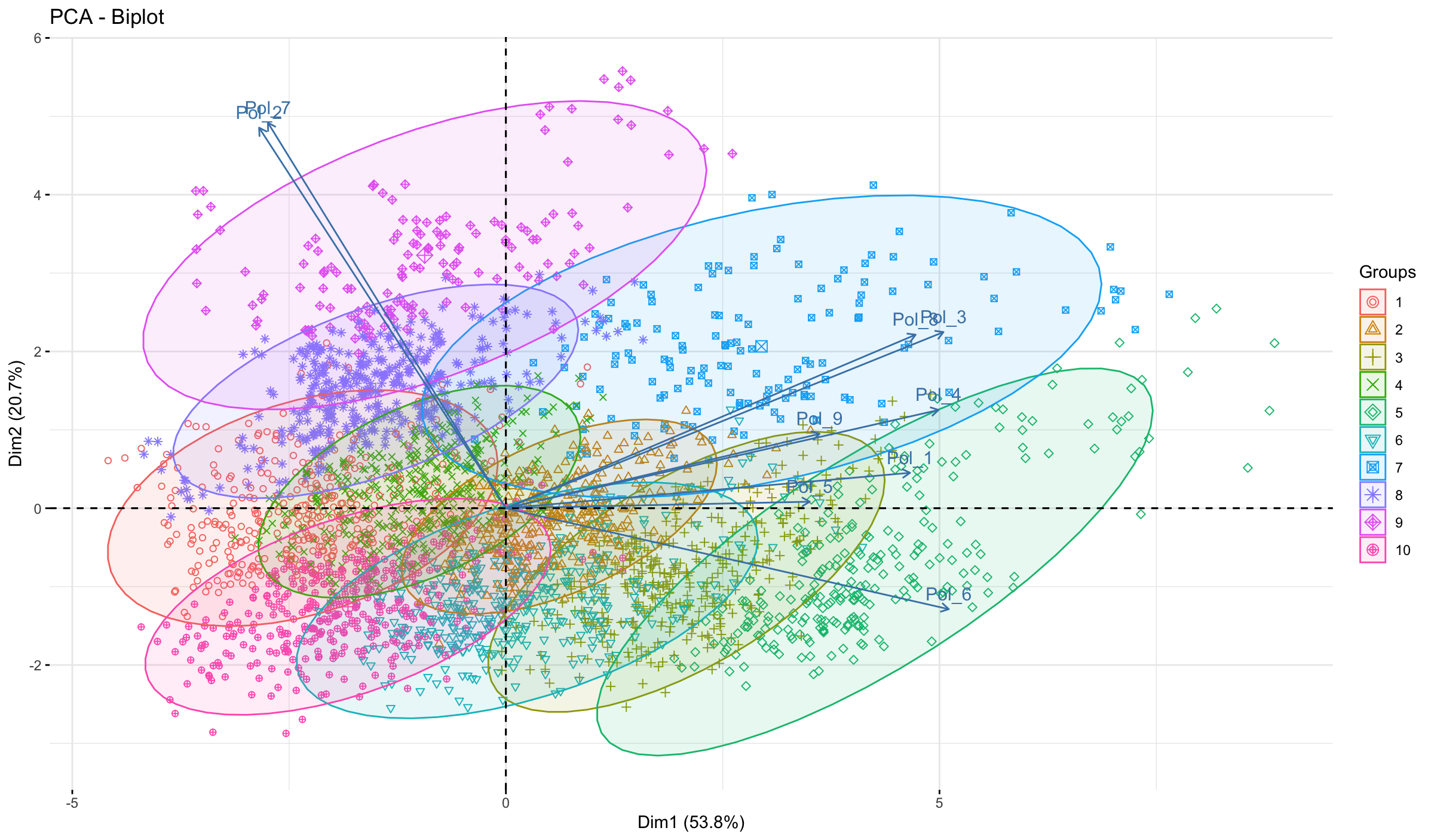
large airport. Different pre-defined groups in the data are included as the variable Cluster.

(a) Construct a PCA biplot of the Pollution data without showing alpha bags.



***Figure 1: PCA Biplot***

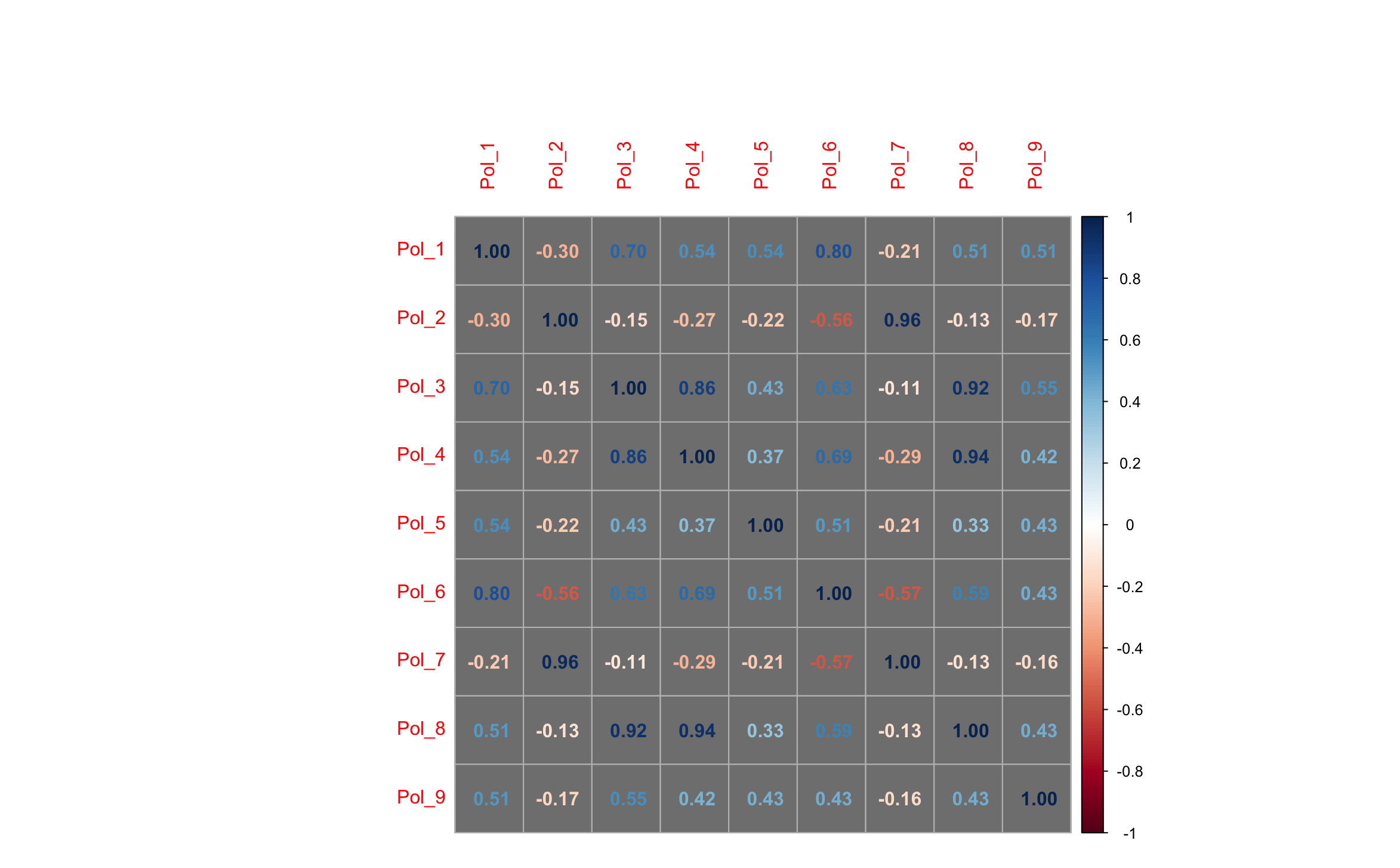
(b) Repeat (a) but instead of sample labels show the different groups (clusters) as 90% bags.



***Figure 2: PCA Cluster Biplot***

(c) Repeat (a) but give an optimal two-dimensional display of the correlations between the

variables.



***Figure 3: Correlation Matrix***

(d) Give a detailed interpretation of the plots constructed in (a), (b), and (c).

From the first bi-plot the table below shows the Principal components and the variables they are strongly correlated to. Strong correlation (>=0.5).

|  |  |
| --- | --- |
| **Principal Component** | **Strongly Correlated Variables** |
| 1 | - |
| 2 | Pol\_2, Pol\_7 |
| 3 | Pol\_5 |
| 4 | Pol\_9 |
| 5 | Pol\_1, Pol\_5 |
| 6 | Pol\_3, Pol\_6 |
| 7 | Pol\_4 |
| 8 | Pol3, Pol\_8 |
| 9 | Pol\_2, Pol\_7 |

***Table 1: PC vs Variables***

For the second bi-plot the table below shows the Groups and the variables they are strongly correlated with.

|  |  |
| --- | --- |
| **Group** | **Strong correlated with** |
| 1 | Pol\_2, Pol\_7 |
| 2 | Pol\_1, Pol\_5, Pol\_3, Pol\_4, Pol\_6, Pol\_8, Pol\_9 |
| 3 | Pol\_1, Pol\_5, Pol\_9 |
| 4 | Pol\_2, Pol\_7 |
| 5 | Pol\_1, Pol\_6 |
| 6 | Pol\_5 |
| 7 | Pol\_3, Pol\_8 |
| 8 | Pol\_2, Pol\_7 |
| 9 | Pol\_2, Pol\_7 |
| 10 | - |

***Table 2: Cluster vs Variables***

Based on the correlation matrix it shows there are strong relationships between

Pol\_1 and Pol\_6, Pol\_2 and Pol\_7, Pol\_3 and Pol\_8, Pol\_4 and Pol\_3 & Pol\_4 and Pol\_8,

(e) Construct a CVA biplot of the Pollution data with 90% bags added. Interpret and discuss the use of this biplot.



***Figure 4: CVA Biplot***

CVA is used to analyze group structure in multivariate data. It can be interpreted in terms of similarity or dissimilarity between groups. The bi-plot provides visual representation of the key distances among groups the different clusters. It shows that cluster 5, 9 and 10 are very unique.

**Question 2**

Consider again the Pollution data set introduced in Question 1

(a) Compute the following dissimilarity/distance matrices for Cluster 4:

i) Euclidean distances.

Used the dist function with method equal to Euclidean.

ii) Canberra distances (Hint: Study the help file of function dist() ).

Used the dist function with method equal to Canberra.

(b) Perform separate classical scalings on both of the dissimilarity/distance matrices

computed in (a).

Used the R function cmdscale on the Euclidean and Canberra distance matrices.

(c) Interpret and discuss the classical scalings in (b). Include in your answer a comparison with the biplots constructed in Question 1.

Metric Multidimensional Scaling (metric MDS) is a statistical method that converts data on distances between items into map-based visualization of those items. The generated mappings can be used for better understanding which items are close to each other, and which are different. It can also allow you to identify groups or clusters.

Both distance measure calculates the different between pair of objects but in different ways. Euclidean distance computes the root of square difference between co-ordinates of pair of objects. Canberra distance examines the sum of series of a fraction differences between coordinates of a pair of objects.

Since we are only looking at cluster 4, the results when plotted can show how far each item in cluster 4 is spaced out from each other. The outcomes is very similar to the bi-plot outlined in question 1 looking at cluster 4.

**Question 3**

The Excel data set ‘Brands.data.xlsx’ contains data of 10 features associated with 23 brands of a certain product.

(a) Obtain a Euclidean distance matrix of the Brands.

Used the dist function with method set to Euclidean.

(b) Give a brief description of the use of a stress function associated with metric least

squares scaling.

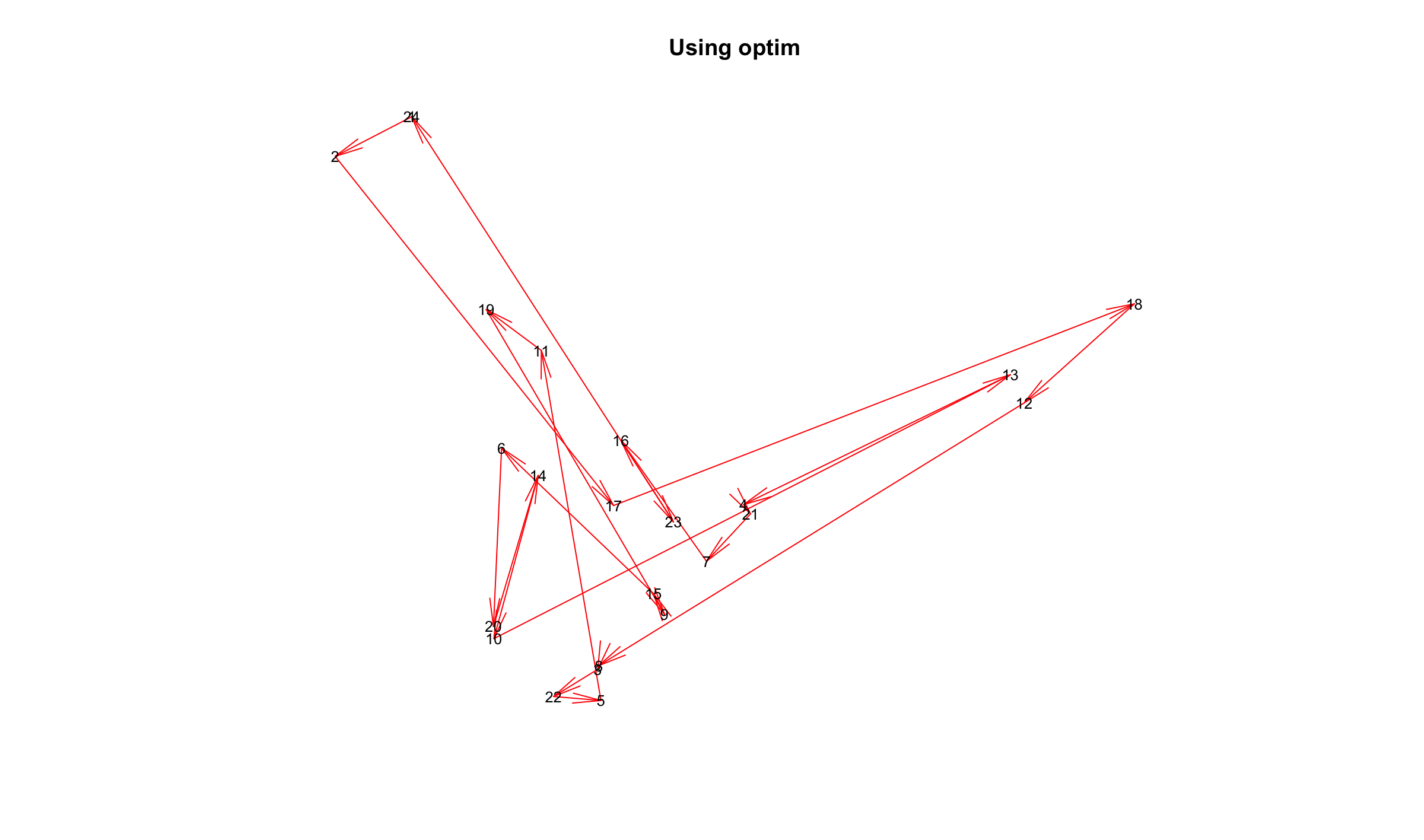
It is the measure of goodness of fit in multidimensional scaling. It measures the difference between the observed (dis)similarity matrix e.g. reaction time between semantic pairs and the estimated one using one or more estimated stimuli dimensions. The lower the stress the better the fit.

(c) Use the function optim() to perform a metric least squares scaling on the Brands data set.

Used the optim package and created a distance function. Set method to SANN

(d) Discuss the output (graphical and statistical) of the analysis in (c).

From the model, the squared value was 390.48. Graphically it is outlined in figure 5 which shows how the different brands interact.



***Figure 5: Using optim function***

(e) Transform the Brands.data into the R object Brands.data.ord where each feature is an ordered categorical variable.

Used base R apply function and then did a sort for all columns.

(f) Inspect the help file of function daisy() in the R package cluster. Use daisy() to

construct an ordinal dissimilarity matrix for the different brands considered in

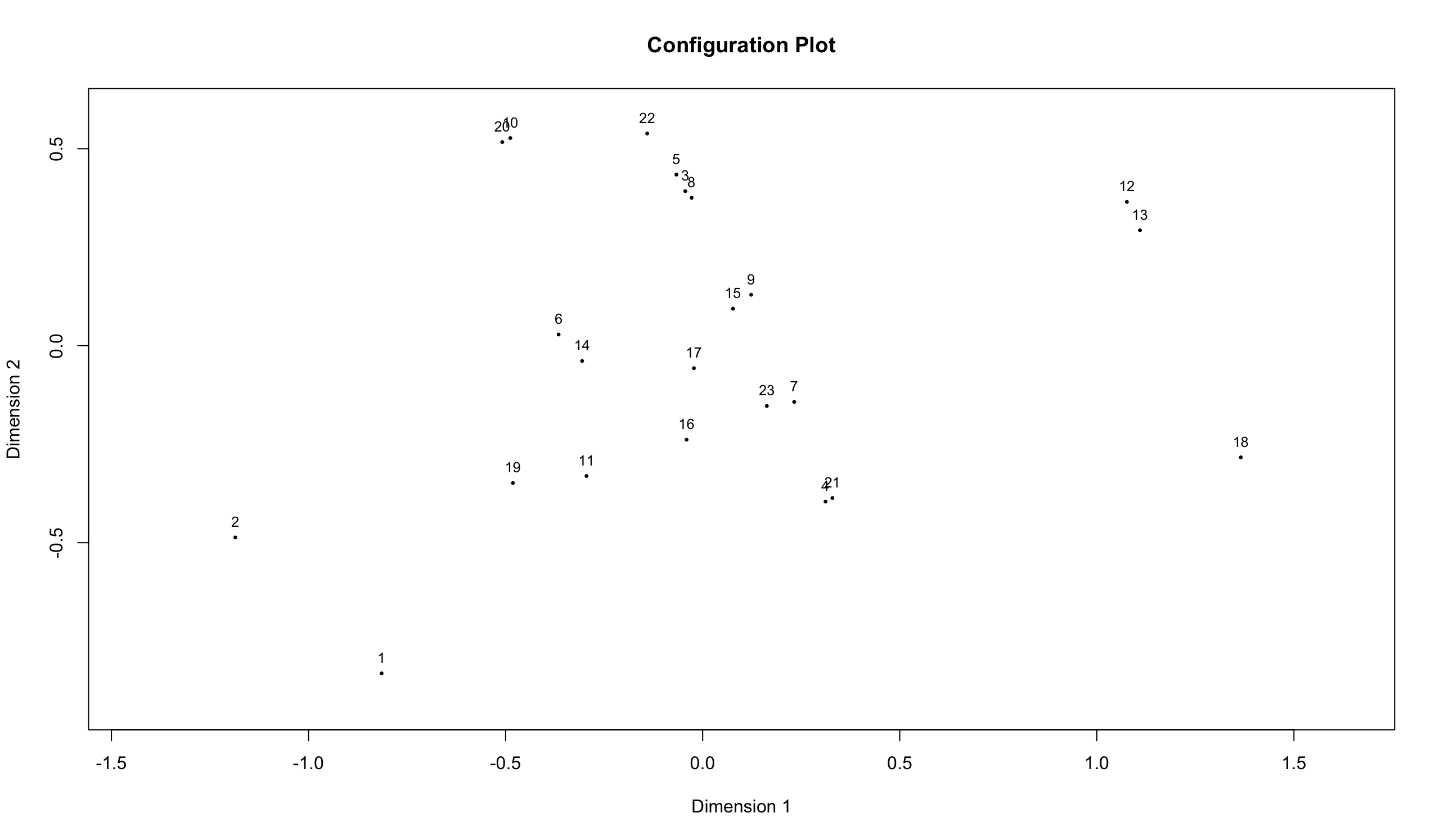
Brands.data.ord based on Gower's coefficient.

Leveraged the daisy package and set metric to gower.

(g) Perform a nonmetric MDS on the Brand dataset by uitilising the R package smacof.

Give a detailed interpretation of your nonmetric MDS. Refer also to your findings in (d).

Mds function in smacof was used with type set to ordinal to calculate the non-metric MDS. The output of the nonmetric MDS shows the configuration to minimize the stress, the square root of the ratio of the sum of squared differences between the input distances and those of the configuration to the sum of configuration distances squared. However, the input distances are allowed a monotonic transformation In this case. The values in Figure 5 follow a similar pattern to Figure 6 but without the directional arrows.

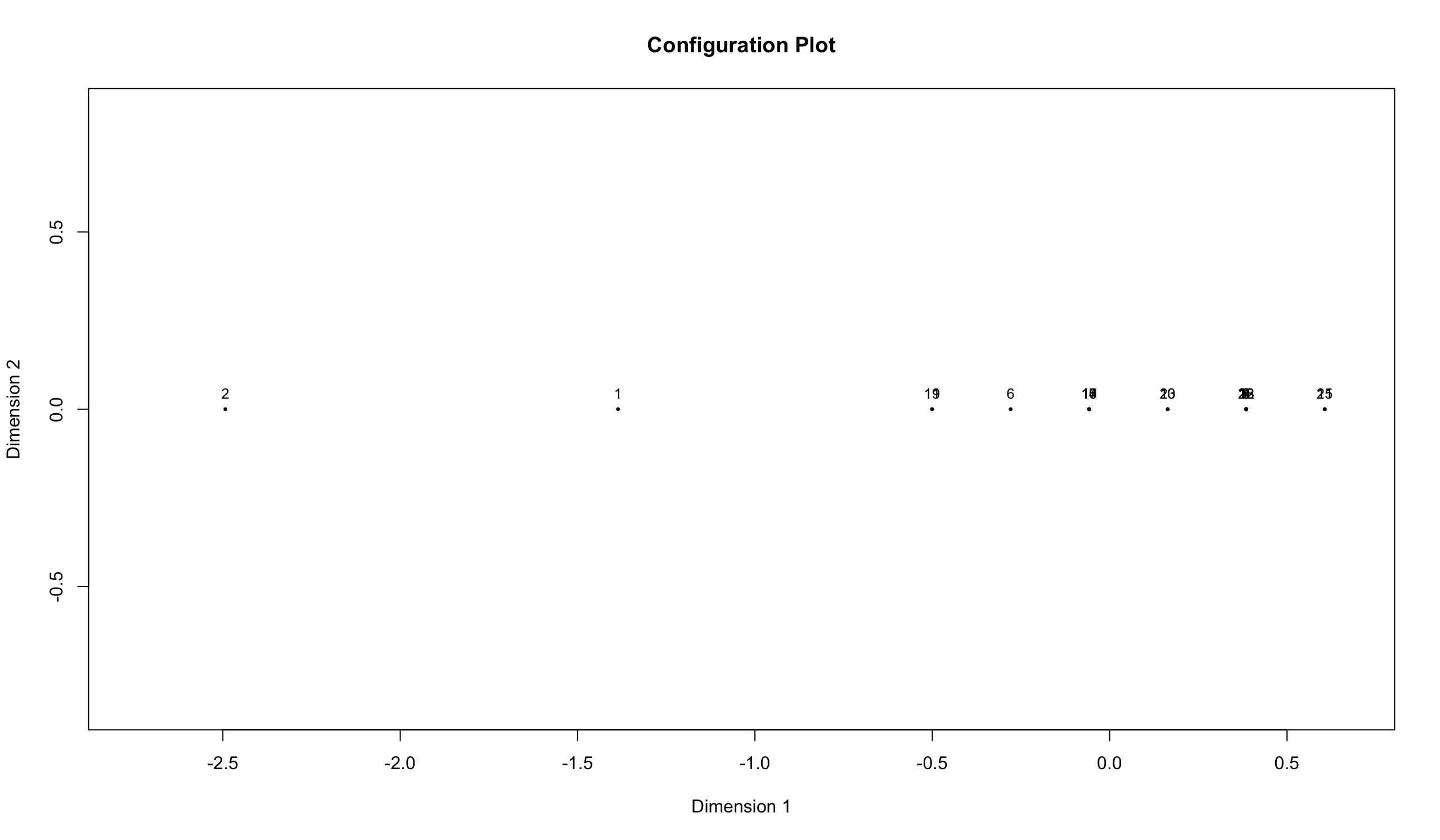


***Figure 6: Config. Plot***

(h) Repeat the nonmetric MDS graphical representation of the Brands in (g) but with the

values of Feature 5 replacing the different Brands. Interpret your graph.

The items in Feature 5 are all on the same line which shows they are all part of the same group.



***Figure 7: Config. plot***

**Question 4**

Perform a detailed Procrustes analysis on the metric and non-metric configurations obtained in Question 3 by using as target configuration the metric least squares MDS configuration.

Here is the output of the Procrustes Analysis

procrustes(X = brand\_metric\_mds1, Y = brand\_non\_metric\_mds$init, symmetric = FALSE)

Number of objects: 23 Number of dimensions: 2

Procrustes sum of squares:

0

Procrustes root mean squared error:

0

Quantiles of Procrustes errors:

Min 1Q Median 3Q Max

8.881784e-16 2.915641e-15 5.333694e-15 6.755087e-15 1.930128e-14

Rotation matrix:

[,1] [,2]

[1,] 1.000000e+00 -3.934575e-16

[2,] 3.934575e-16 1.000000e+00

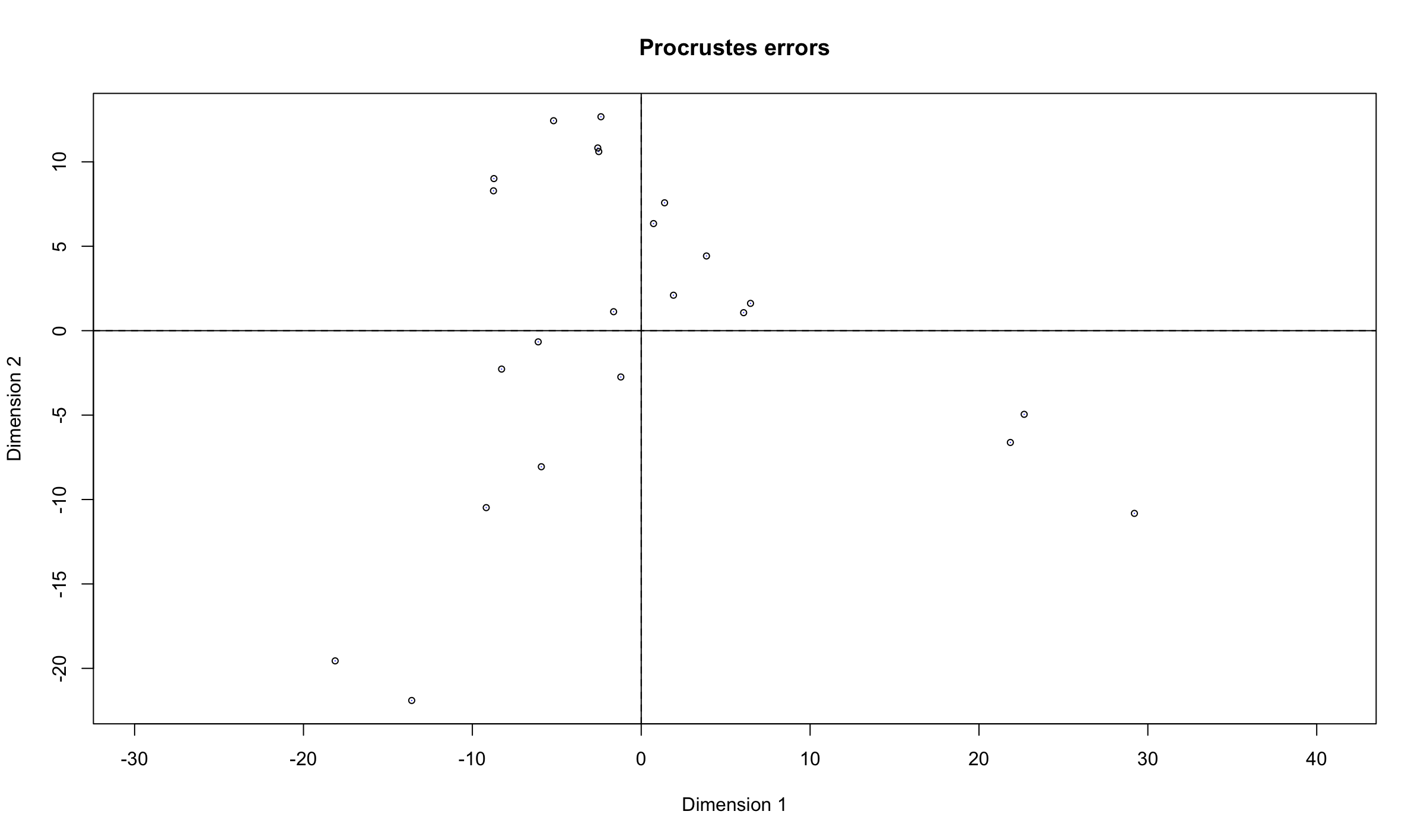
Translation of averages:

[,1] [,2]

[1,] -1.635825e-16 2.500085e-15

Scaling of target:

[1] 1



***Figure 8: Procrustes plot***

**Question 5**

Ninety-two subjects took part in an opinion survey consisting of five questions. The subjects

were recruited in three different districts. The data are given in the file

OpinionSurvey.data.xlsx.

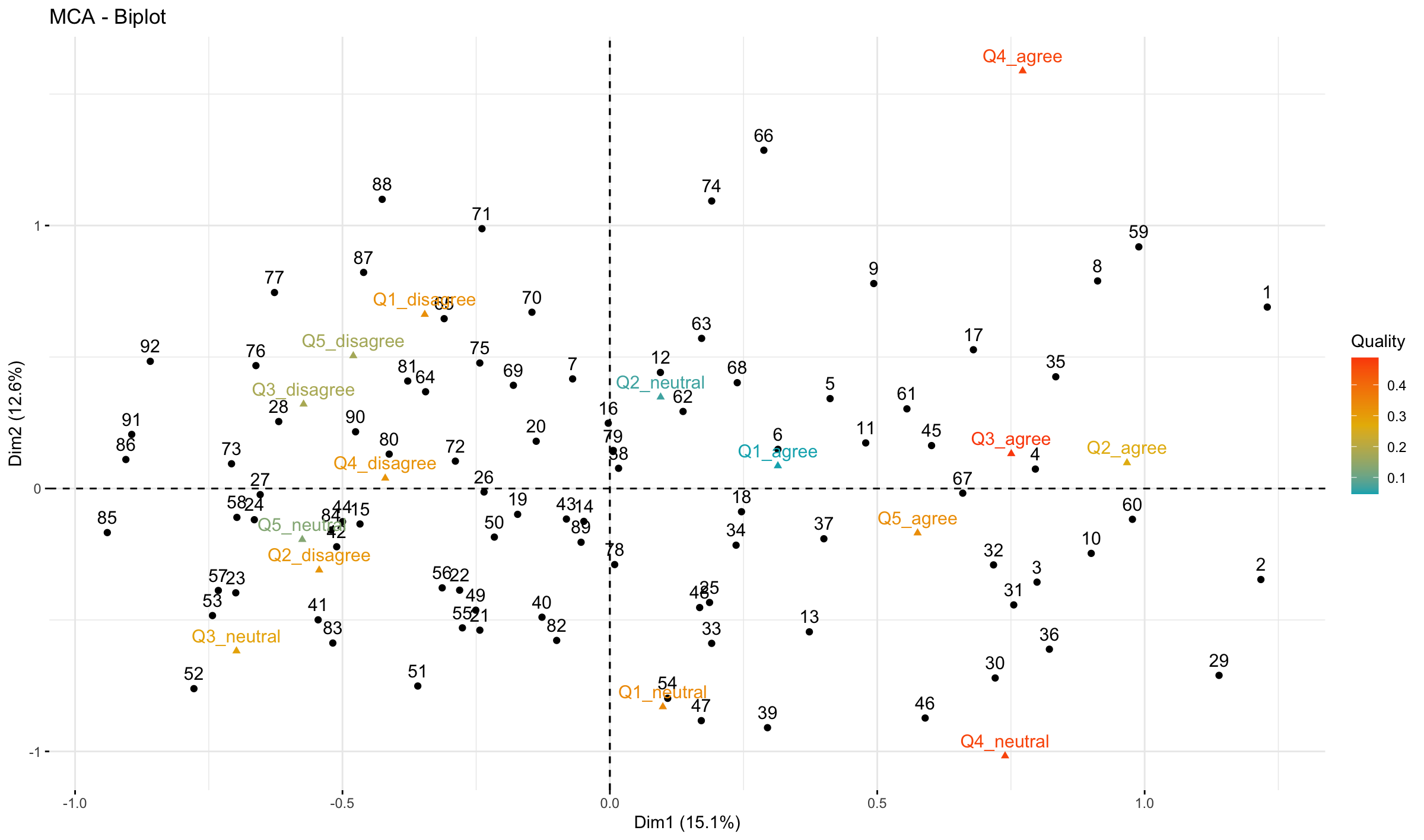
(a) Ensure that the answer to each question is a categorical variable.

Unique base R function was used to check the variables (Q1, Q2, Q3, Q4 and Q5).

(b) Construct an MCA biplot on the associated indicator matrix of the questions. Do not

colour the sample points but label them using their IDs as labels. Represent each

categorical variable in a different colour. Add a suitable legend to the MCA biplot.



***Figure 9: Question MCA Biplot***

(c) Repeat (b) but this time colour all CLPs in the same colour while distinguishing the

samples from the different districts using colour coding.



***Figure 10: Samples MCA Biplot***

(d) Discuss in detail the conclusions to be drawn from the MCA biplots in (b) and (c) and

the associated optimal scores. (20)

Figure 9 shows the quality of each individual while Figure 10 shows the contributions made by each group. The individual bi-plot shows there are clusters of information among the different questions asked. Also data point like 1,2, 59 had significant weight. Answers around Q4\_agree has very no answers associated with it compared to Q4\_disagree.

**Question 6**

Six features of products produced by three manufactures were scored as ordered categorical

variables and saved in Excel format as Manufacture.data.xlsx. Feature 6 was scored in

reversed order as the other features.

(a) Ensure that all features are ordered categorical variables in your imported data set.

Reviewed each feature (1 to 6) using the base R unique function.

(b) Use function CATPCAbipl as given in package UBbipl40 to carry out a Categorical

Principal Components Analysis on the Manufacture data. (Hint: due to many similar

response patterns argument jitter.bags should be set to TRUE when requesting alpha

bags)

Needs clarification - Can’t install the package due to my version of R not being compatible

(c) Give a brief discussion of the aim(s) of a Categorical Principal Components Analysis

followed by a detailed discussion/interpretation of the result of the analysis in (b). (20)

Categorical Principal Components Analysis simultaneously allows at reducing an original large set of variables into a smaller set of uncorrelated components. Categorical variables are optimally quantified in the reduced data dimensionality and nonlinear relationships between variables can be modelled.

Needs clarification - Can’t answer question because of b

**Question 7**

A research group aimed to construct a questionnaire for screening candidates for having an

underlying disease. They started with 50 binary questions (symptoms) given to a carefully

selected sample consisting of 5000 subjects (respondents or persons). The data are given in the file symptoms.data.xlsx where a 0 denotes no indication of the underlying disease and a 1 is an indication that the underlying disease is present.

(a) First, inspect the contents of R package CTT and then perform an Item Analysis on the Symptoms data.

Used the help function to look at the CTT package.

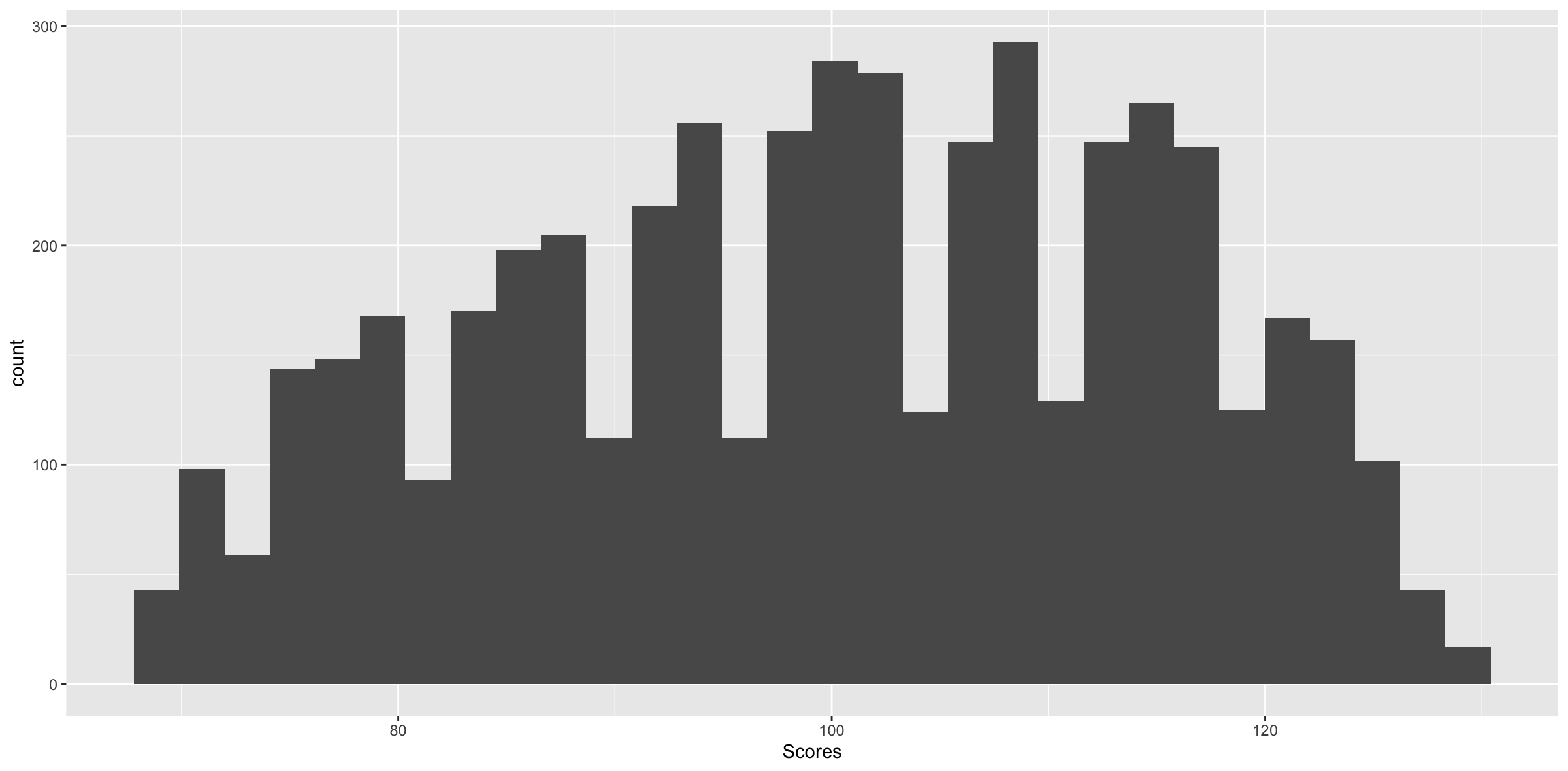
Analysis was done using the CTT package. The Function called was CTT::itemAnalysis. The analysis generated the following values. Alpha - 0.956, scaleMean - 26.37, scaleSD – 12.39.

(b) Obtain the ‘person’ scores and transform these scores to a scale having a mean of 100 and a standard deviation of 15.

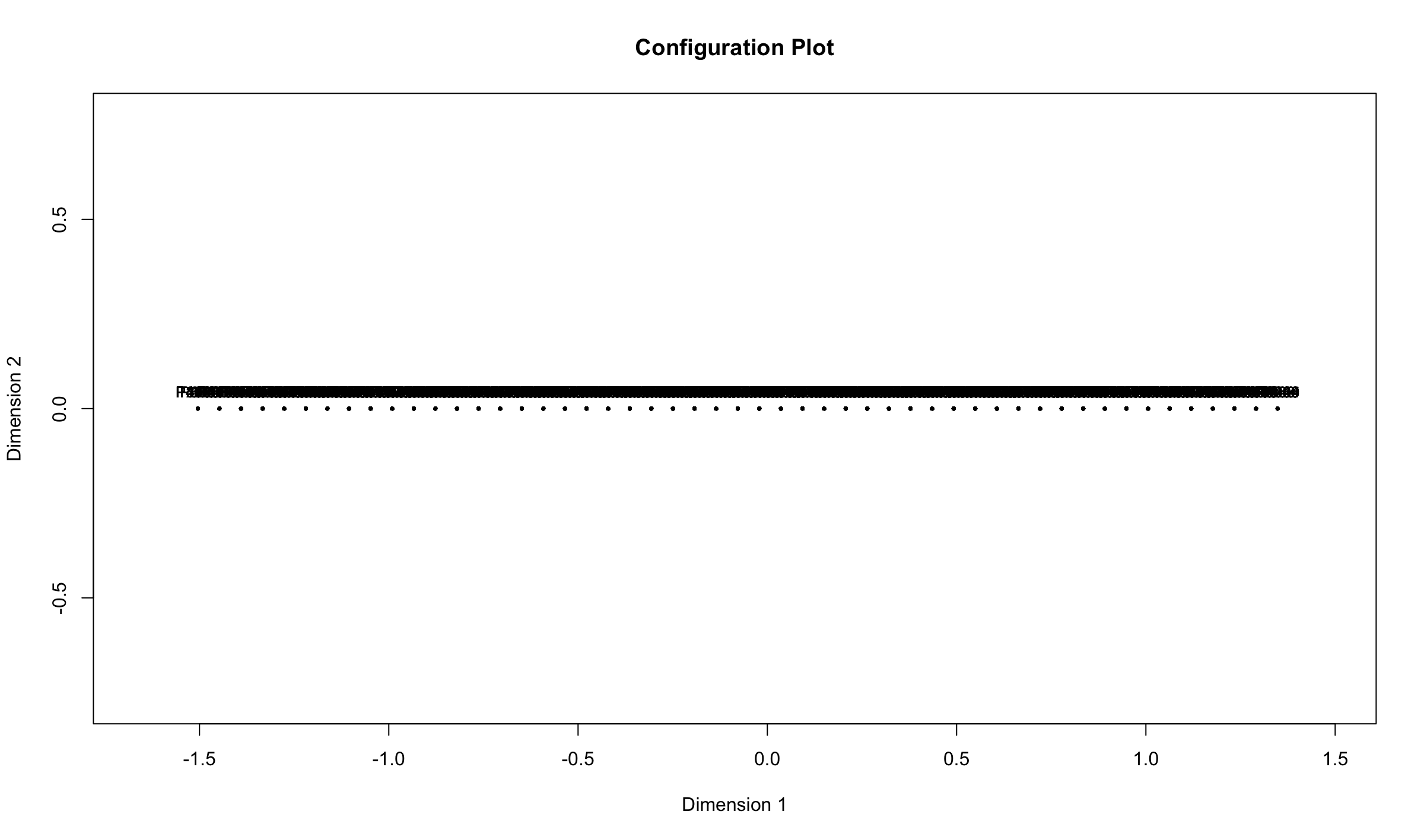
Used CTT::score and CTT::score.transform to generate the solutions.

(c) Represent the transformed person scores in the form of a unidimensional scaling graph. Explain how to use this graph in practice.

Figure 11 shows the transformed data as a histogram which shows how the scores are distributed and grouped. Figure 12 shows the score transformation. They are all in one line showing and different numbers clustered together.



***Figure 11: Histogram***



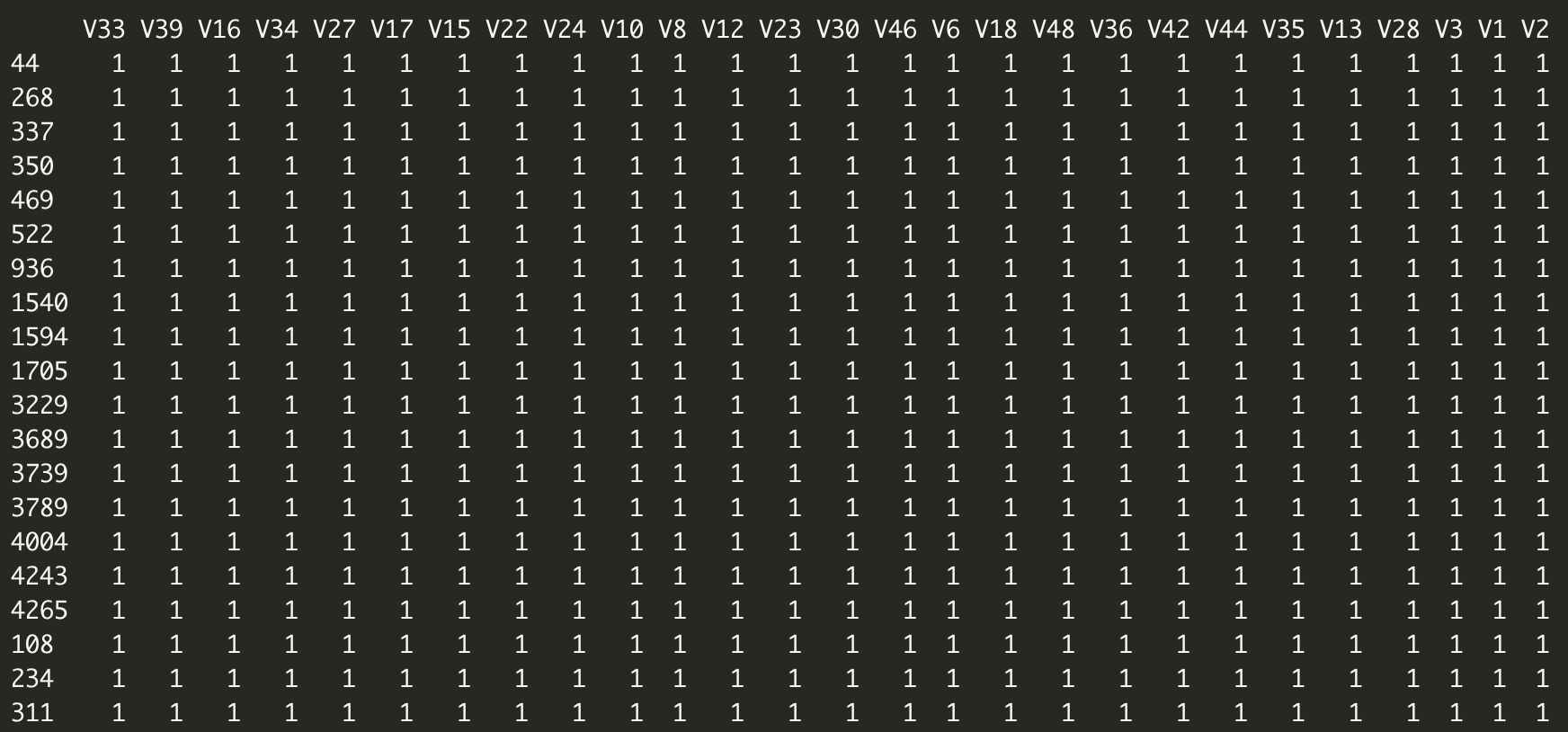
***Figure 12: Unidimensional graph***

(d) Construct a unidimensional scale (in table and graph form) for the items and explain

how to interpret the scale.

|  |
| --- |
| V39 |
| V16 |
| V34 |
| V27 |
| V17 |
| V15 |
| V22 |
| V24 |
| V10 |
| V8 |
| V12 |
| V23 |
| V30 |
| V46 |
| V6 |
| V18 |
| V48 |
| V36 |
| V42 |
| V44 |

***Table 3: Top 20 for Unidimensional scale table***



***Figure 13: Unidimensional scale***

(e) Motivate in detail which 20 items from the original 50 items would you recommend for a final screening test. (15)

Guttman scaling is also sometimes known as cumulative scaling or scalogram analysis. The purpose of Guttman scaling is to establish a one-dimensional continuum for a concept you wish to measure. What does that mean? Essentially, we would like a set of items or statements so that a respondent who agrees with any specific question in the list will also agree with all previous questions. It assumes that people’s willingness to participate in social relations with other people vary in degrees of intensity, and measures that intensity using a list of items arranged from “least intense” to “most intense”. Based on the results of the Guttman scaling the top 20 to select are highlighted in table 3. The guttman scaling had a coefficient of Reproducibility of 0.82516 which is good.

**Question 8**

Obtain a new data set (new.data) consisting of the 20 recommended items in Question 7(e) for a systematic sample of persons chosen by taking the first and thereafter every 50th person in the Symptoms data set.

(a) Before attempting (b) study Unit 5 Slides 15-24. If you have not already installed

package ltm then install it from (https://cran.r-project.org › web › packages › ltm ).

Next, work through the R script file: IRT.Examples.R.

Needs clarification – IRT.Examples.R not provided

(b) Use item response theory and the data set, new.data created above to

(i) fit a Rasch model to the data;

(ii) find disease scores and express them in unidimensional scaling format;

(iii) obtain item characteristic curves;

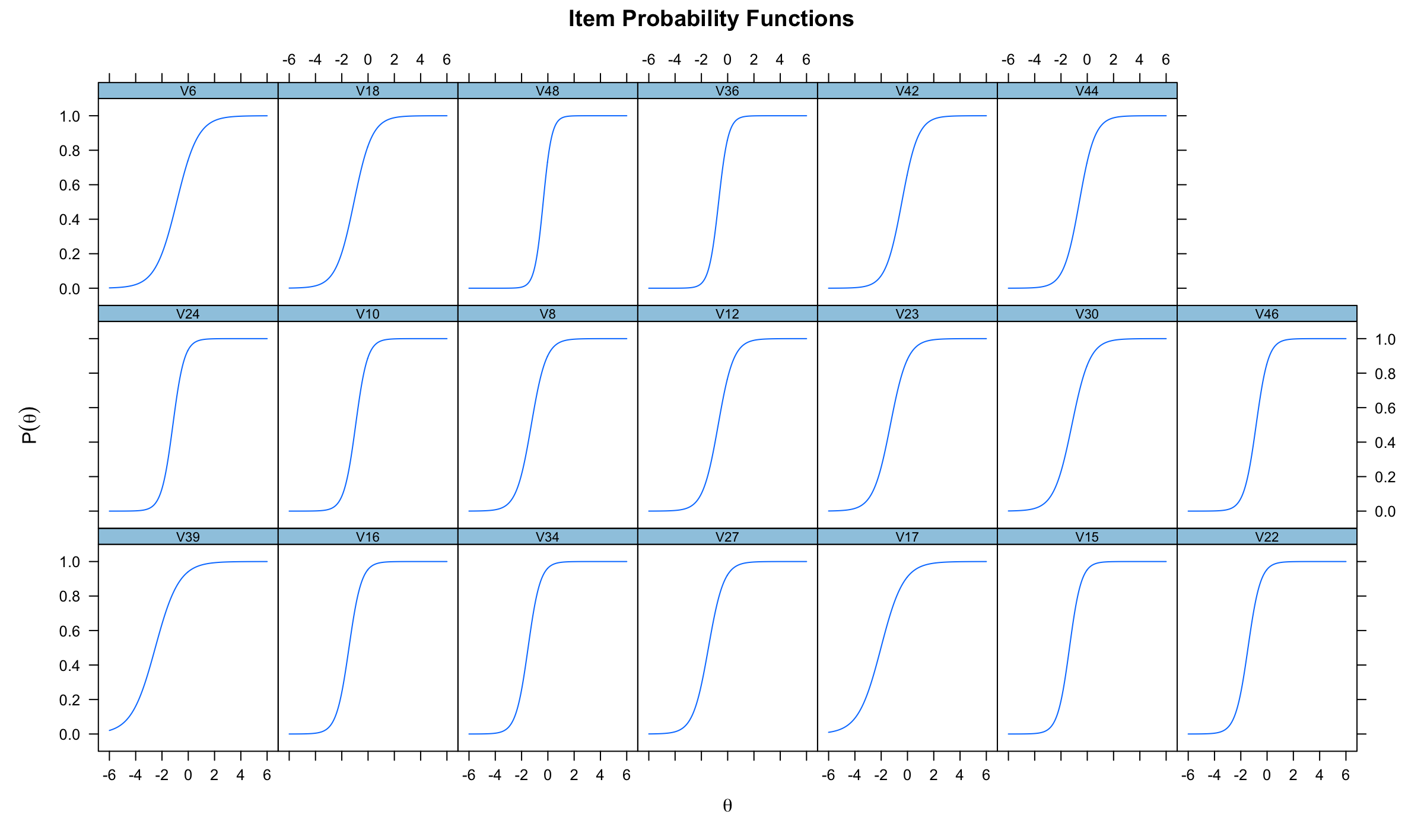
(iv) obtain item information curves and

(v) write a report on your findings

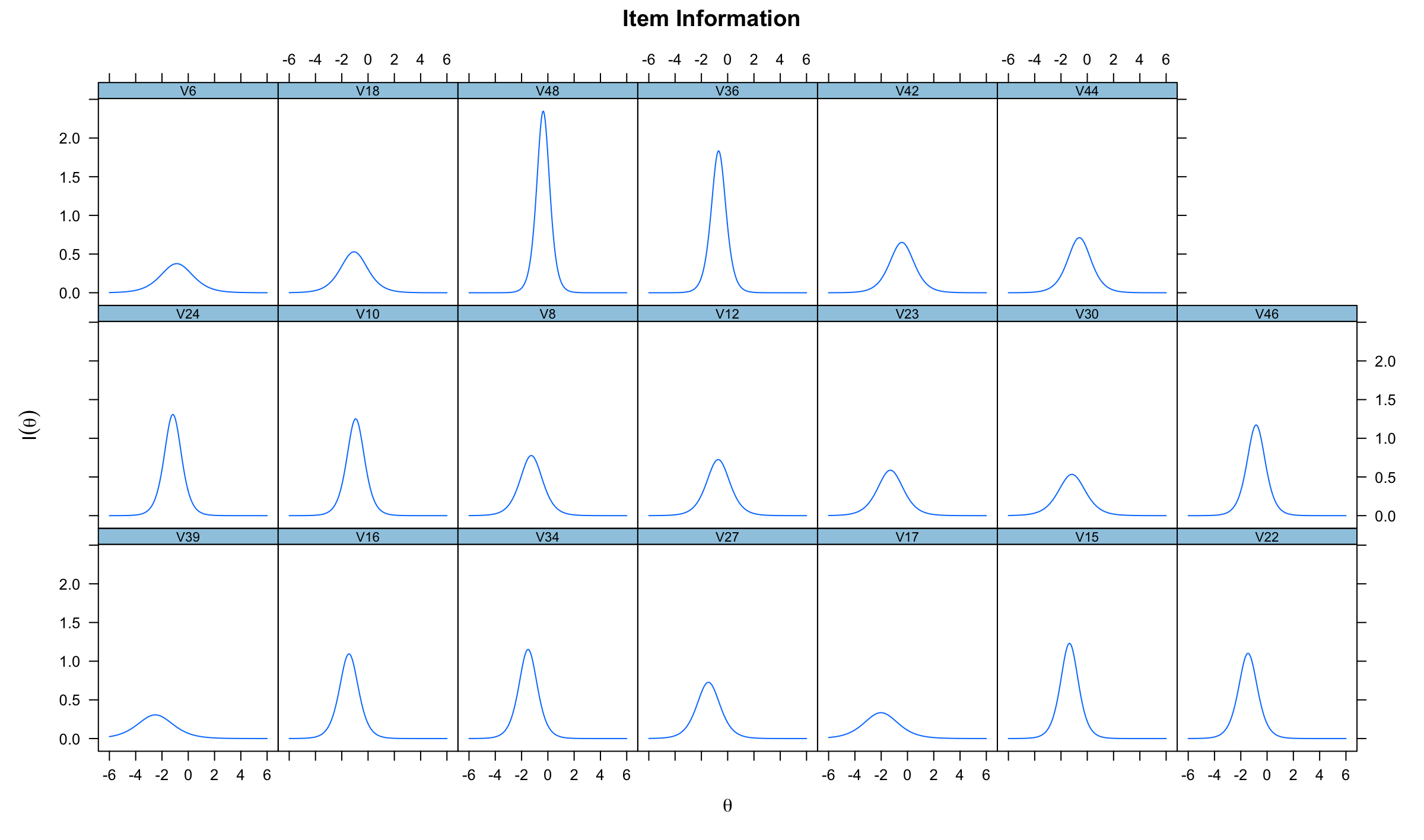
Based on analysis it shows that the top 20 items selected were pretty decent. Checking the itemfit function it shows that most of the RMSEA values are less than .06 indicating that the items had adequate fit with the model.

The characteristic has a functional relationship with theta; As theta increases, the probability of endorsing a category increase and then decreases as responses transition to the next higher category. It indicates that the response categories have a pretty consistent range. This is highlighted in figure 14.

Item level information clarifies how well each item contributes to score estimation precision with higher levels of information leading to more accurate score estimates. These curves show that item information is not a static quantity, rather, it is conditional on levels of theta. The relationship between slopes and information is illustrated here. Item V6 had the lowest slope and is, therefore, the least informative item. On the other hand, Item V48 had the highest slope and provides the highest amount of statistical information.



***Figure 14: Characteristic curves***



***Figure 15: Information curves***