Contents

Assignment 1: PCA

Task 1 Get the multivariate data

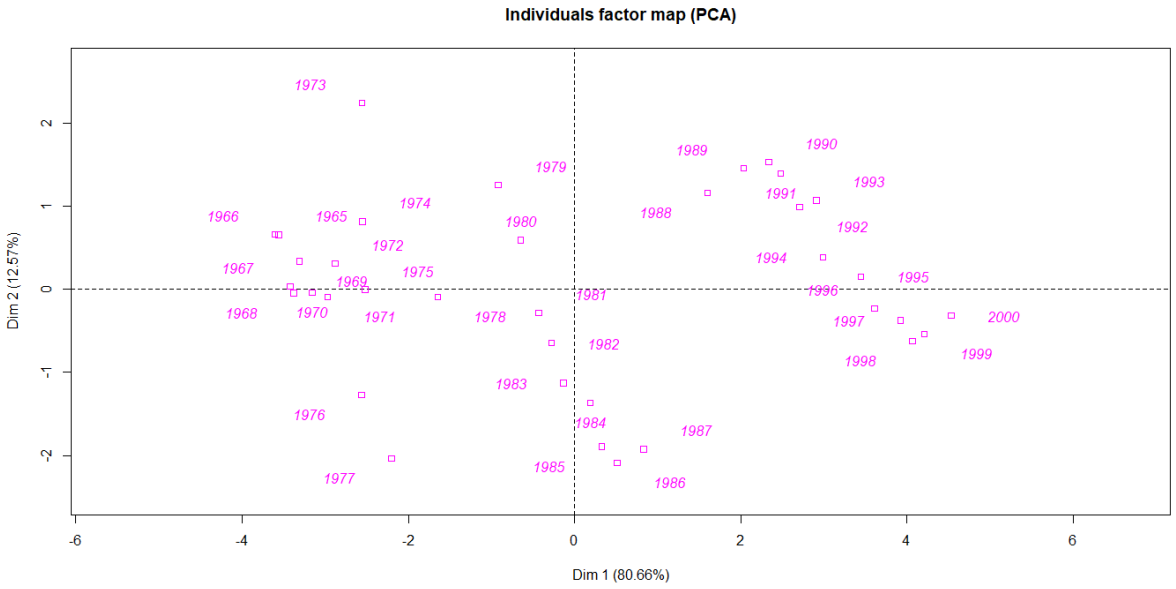
```
ls()

beef<-read.table("BeefDemand.txt", header = TRUE)

## The dataset consists of a few variables that may influence the demand for Beef

## in the United States.  It provides an example of the influence of inflation in

## monetary time series data as well as providing some interesting statistical

## features in building demand models in regression.
```

Task 2

a) Plot the individuals in the plane corresponding to the first two principal components (PCs), see [1], p.31. Comment on the resulting cloud.

We can draw conclusions based on the distance between objects, we focus on Dim1 since it explains almost 81% variance. Thus we base our comparison based on the 1st dimension. For example, let's compare observations in 1973 and 2000:

All the values of the variables are different except the BeefConsump column, hence the large distance between two objects in the graph is confirmed.

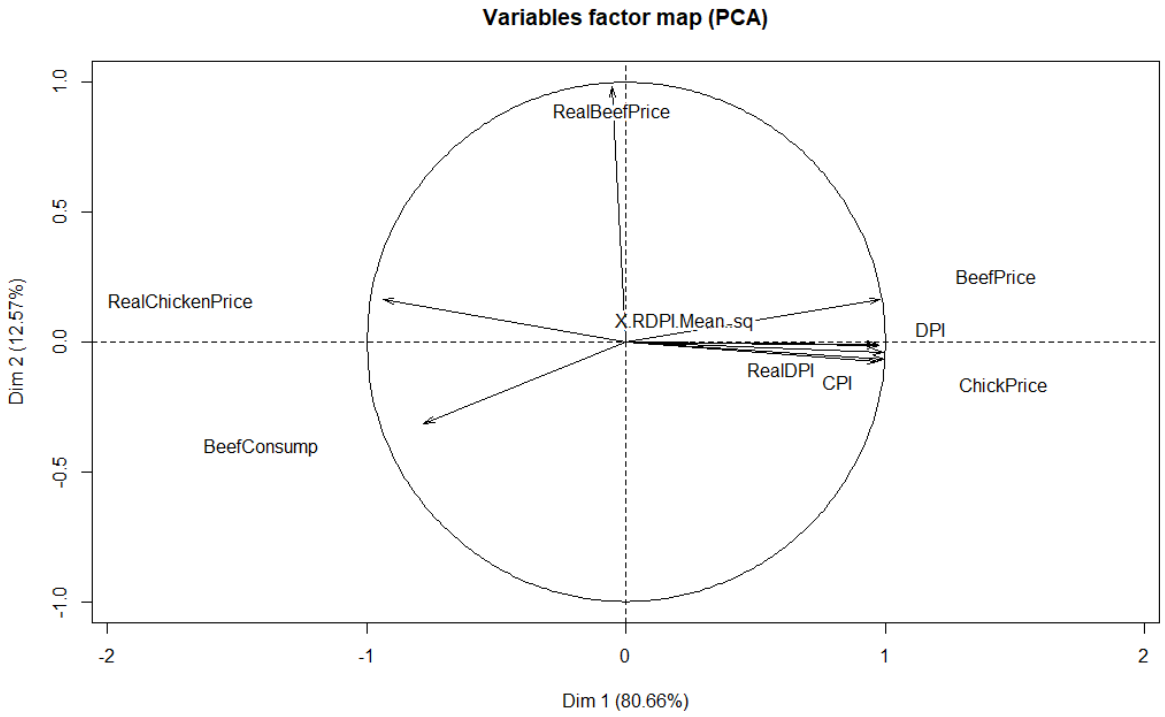


Individuals factor map

Let's check now similar objects in 1998 and 1999:

Here the values for variables are close to each other, hence the proximity of the two objects in the graph is confirmed too.

From the variables factor map we can conclude that prices for beef and chicken are positively correlated. Whereas, beef consumption and beef price are negatively correlated, which is logical. The vectors / variables that are close to unity on the variables factor map are represented well in 2D artificial space.



Variables factor map

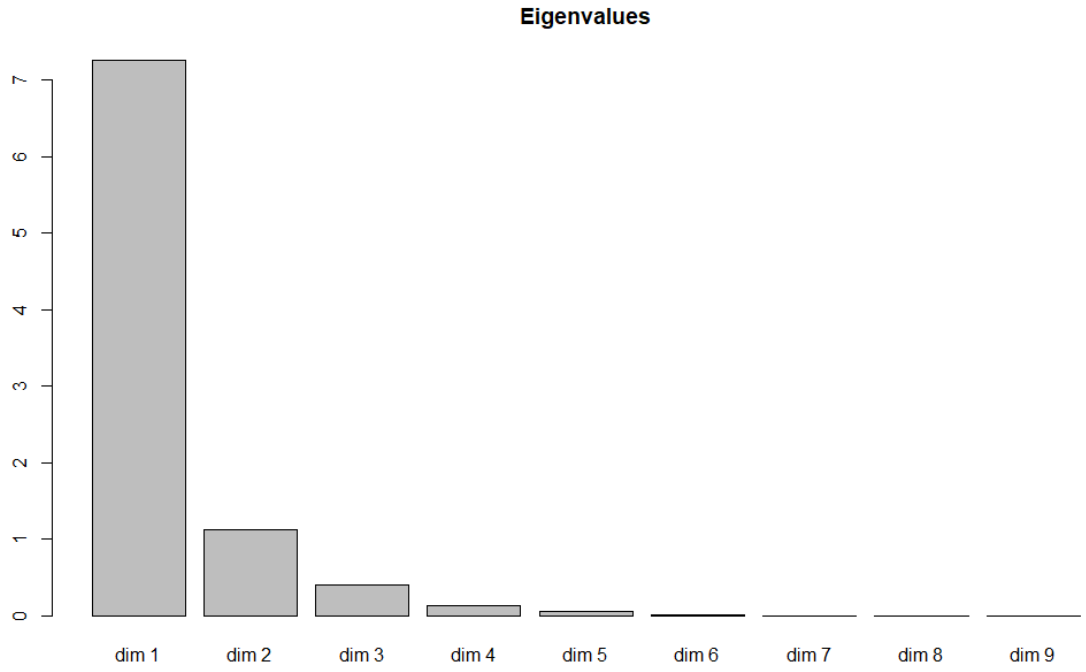
b) Justify the choice of the PCs by plotting the eigenvalues, [1],p.32. Calculate how much of the total variability is explained by the first two PCs.

The first two dimensions explain variability very well: overall explanation of data variability is 93.23%

comp 1 explains 80.66% of variance

comp 2 explains 12.57%

since we have 9 dimensions there are 9 components, but we need only first component since it explains majority of variance.



Eigenvalues barplot

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	7.259098e+00	8.065664e+01	80.65664
comp 2	1.131182e+00	1.256869e+01	93.22533
comp 3	4.098767e-01	4.554186e+00	97.77952
comp 4	1.353133e-01	1.503481e+00	99.28300
comp 5	5.201060e-02	5.778955e-01	99.86090
comp 6	9.806389e-03	1.089599e-01	99.96986
comp 7	1.777911e-03	1.975456e-02	99.98961
comp 8	9.164319e-04	1.018258e-02	99.99979
comp 9	1.865596e-05	2.072885e-04	100.00000

Eigenvalues

c) Discuss the quality of the PCA representation: provide cos2 and the contributions for each individual, [1], p.34.

We have 36 individuals, we should concentrate on Dim.1 since it explains the most variability. In general majority of the individuals are well represented by the first dimension. However, objects 18-23 are poorly presented by the dim 1. They are better represented by the second dimension

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	0.80	0.03	0.14	0.03	0.00
2	0.88	0.03	0.05	0.03	0.00
3	0.93	0.00	0.06	0.01	0.00
4	0.97	0.00	0.02	0.00	0.00
5	0.98	0.01	0.00	0.00	0.00
6	0.98	0.00	0.00	0.00	0.02
7	0.97	0.00	0.00	0.00	0.02
8	0.93	0.01	0.02	0.00	0.03
9	0.52	0.40	0.04	0.03	0.01
10	0.83	0.08	0.09	0.00	0.00
11	0.85	0.00	0.14	0.00	0.01
12	0.61	0.15	0.24	0.00	0.00
13	0.48	0.41	0.12	0.00	0.00
14	0.65	0.00	0.34	0.00	0.01
15	0.33	0.61	0.02	0.01	0.01
16	0.39	0.32	0.08	0.09	0.11
17	0.24	0.10	0.29	0.24	0.13
18	0.05	0.30	0.37	0.26	0.02
19	0.01	0.68	0.16	0.15	0.00
20	0.02	0.84	0.06	0.01	0.07
21	0.03	0.91	0.05	0.02	0.00
22	0.06	0.92	0.01	0.00	0.02
23	0.15	0.77	0.07	0.00	0.00
24	0.58	0.31	0.07	0.04	0.00
25	0.64	0.33	0.01	0.01	0.01
26	0.69	0.29	0.00	0.02	0.00
27	0.74	0.23	0.00	0.03	0.00
28	0.86	0.11	0.01	0.01	0.00
29	0.85	0.11	0.02	0.01	0.00
30	0.97	0.02	0.01	0.00	0.00
31	0.99	0.00	0.00	0.00	0.01
32	0.99	0.00	0.00	0.01	0.00
33	0.97	0.01	0.00	0.02	0.00
34	0.95	0.02	0.00	0.03	0.00
35	0.96	0.02	0.01	0.02	0.00
36	0.97	0.00	0.01	0.02	0.00

cos^2 values for each individual

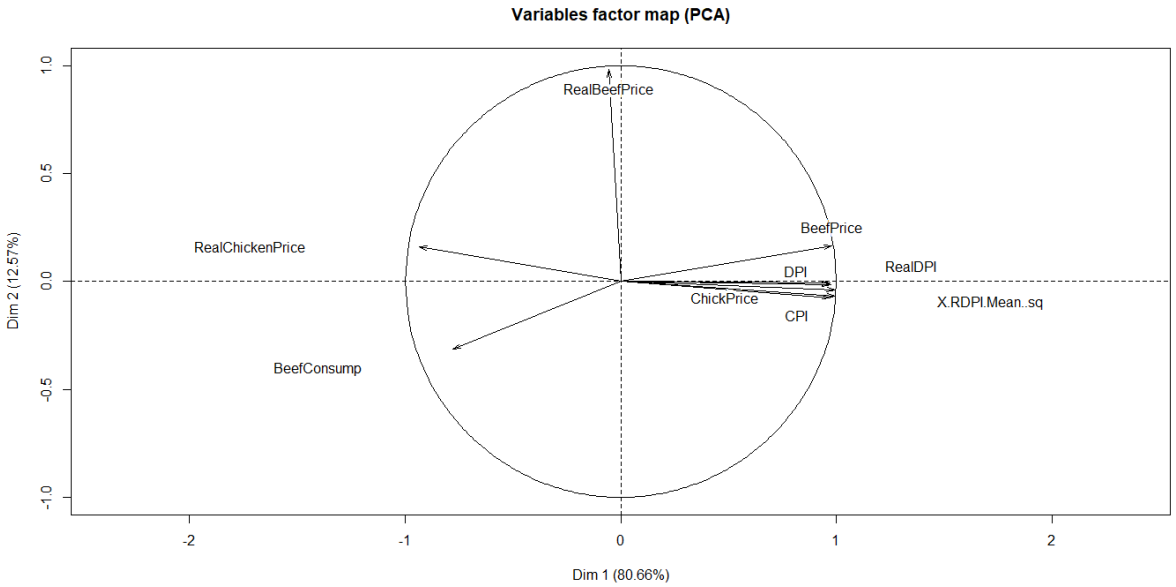
d) If there are categorical variables, paint the individuals with different colors according to the categories. Draw the confidence ellipses and interpret them, [1], p. 36.

There were no categorical variables available for accomplishment of this task. In general, if the confidence ellipses overlap, the category doesn't make sense in terms of quantitative variables.

Task 3

a) Using the graphical output of pca command, discuss correlation between the variables including presence of groups of variables that are closely related.

In general, from the graph we can conclude that all variables are presented very well, because the arrows of each feature is close to the border of the circle. For example, highly positively correlated variables include BeefPrice, ChickPrice, CPI, DPI, RealDpi. Whereas, real chicken price and BeefConsump are negatively correlated with the rest of features mentioned before.



Variables factor map

b) Discuss the quality of the PCA representation: provide cos2 and the contributions for variables.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
ChickPrice	0.94	0.01	0.01	0.00	0.04
BeefPrice	0.96	0.03	0.00	0.01	0.00
BeefConsump	0.61	0.10	0.28	0.01	0.00
CPI	0.99	0.00	0.00	0.00	0.00
DPI	0.99	0.00	0.00	0.01	0.00
RealChickenPrice	0.88	0.03	0.02	0.07	0.01
RealBeefPrice	0.00	0.97	0.03	0.00	0.00
RealDPI	0.95	0.00	0.04	0.01	0.00
X.RDPI.Mean..sq	0.94	0.00	0.03	0.02	0.00

Cos^2 values for variables

This table represents the quality values of variables: cos2 according to dimensions. For instance, all of the variables except RealBeefPrice are perfectly represented by the first dimension (all values are close to one). However, RealBeefPrice is ideally represented by the second dimension. This is the reason why the second dimension represents 12% of the total data variance (100% / 9 variables = approx 11%).

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	4.85	1.05	15.17	8.68	1.68
2	4.97	1.06	5.43	8.80	2.33
3	4.48	0.00	5.25	1.68	2.01
4	4.37	0.01	1.57	1.18	2.54
5	4.19	0.28	0.30	0.42	1.50
6	3.81	0.00	0.03	0.73	10.62
7	3.38	0.02	0.04	0.19	10.46
8	3.17	0.23	1.26	0.88	14.11
9	2.51	12.31	3.72	7.92	4.60
10	2.49	1.62	4.83	0.03	0.05
11	2.43	0.00	6.94	0.04	3.26
12	2.51	3.98	17.33	0.62	0.10
13	1.85	10.18	7.97	0.00	0.12
14	1.03	0.02	9.59	0.00	1.22
15	0.32	3.87	0.34	0.47	1.47
16	0.16	0.86	0.56	2.07	6.15
17	0.07	0.20	1.51	3.85	5.39
18	0.03	1.03	3.58	7.60	1.23
19	0.01	3.14	2.03	5.70	0.41
20	0.01	4.58	0.86	0.55	8.27
21	0.04	8.79	1.29	1.49	0.00
22	0.10	10.68	0.18	0.05	5.02
23	0.27	9.08	2.31	0.02	0.33
24	0.98	3.31	1.96	3.94	0.06
25	1.59	5.20	0.36	0.89	4.80
26	2.09	5.74	0.00	2.46	0.34
27	2.35	4.76	0.24	4.39	0.05
28	2.81	2.40	0.61	2.14	1.89
29	3.25	2.80	1.55	1.70	0.49
30	3.43	0.36	0.38	0.69	1.52
31	4.54	0.06	0.00	0.27	6.45
32	4.99	0.13	0.01	1.65	0.92
33	5.89	0.35	0.03	5.78	0.38
34	6.33	0.95	0.14	9.04	0.02
35	6.79	0.72	0.74	7.35	0.00
36	7.87	0.25	1.88	6.73	0.21

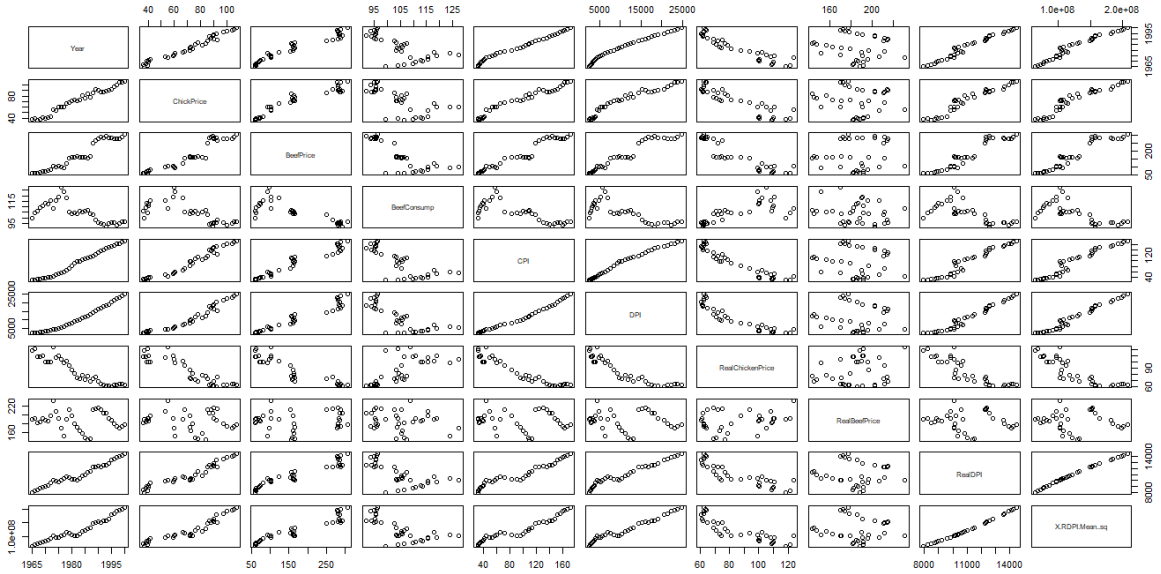
cos^2 values for individuals vs dimensions

The distribution of the contributions of individuals is scattered for each dimension.Ideally the values should be equally distributed along the individuals. For example, the individuals from 15 to 23 contribute insignificantly to the first dimension in comparison to the rest of the observations. However, objects 9, 13, and 22 contribute to the second dimension immensely.

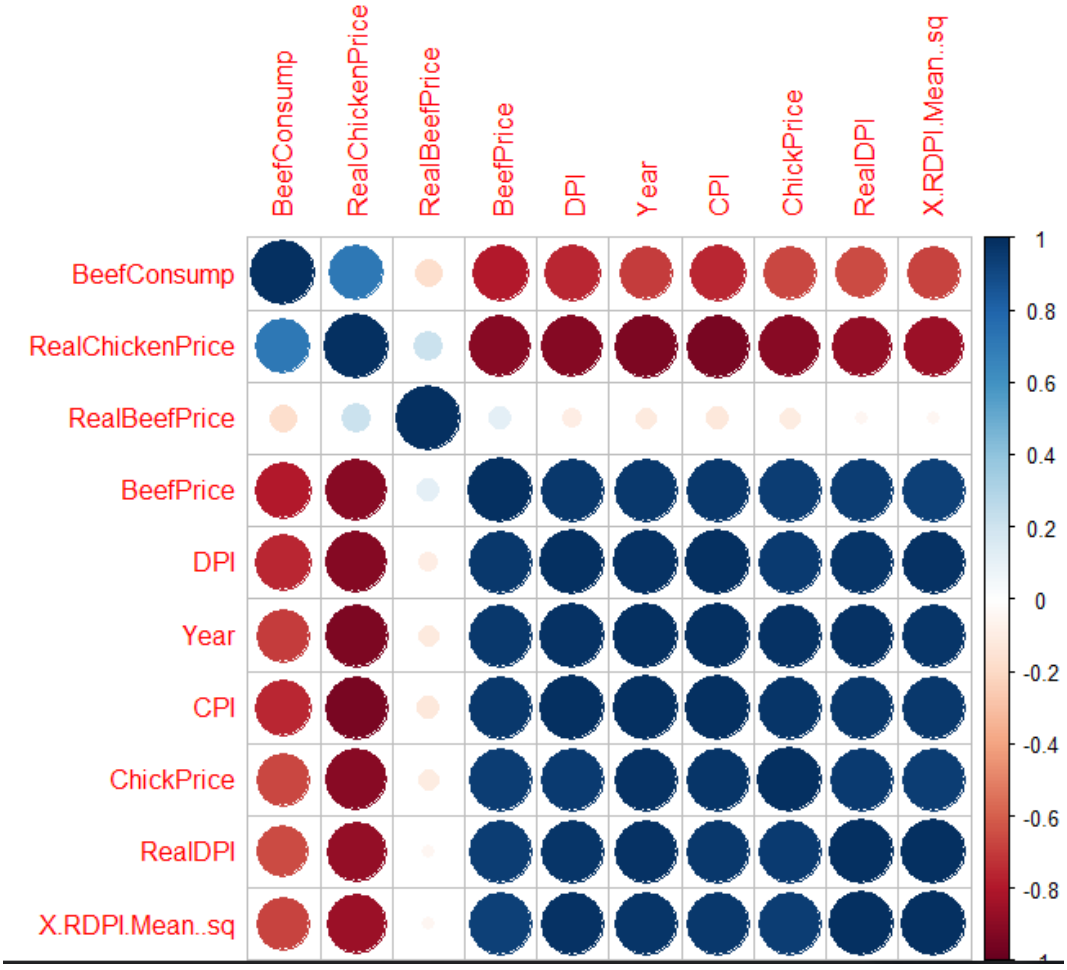
c) Use dimdesc function to summarize the variables. Comment on the p-values.

p-values are very small for all of the variables: hypothesis that respective cor coeffiecient is 0 where H0=cor(CPI, Dim1)=0 vs H1=cor(CPI, Dim1) not = 0. Therefore, we should reject the H0 hypothesis in favor of H1. For example, BeefPrice is highly correlated with principal component Dim1.

d) Plot the correlations between variables using pairs function. Compare the result with that of 3a.



Pairs scatter plot



Correlation plot, order="hclust"

From the correlation plot we can see that the variables BeefPrice, DPI, CPI, ChickPrice, RelaDPI are positively correlated (dark blue circles). However, BeefConsump and RealChickenPrice variables are negatively correlated with the rest of the variables. This correlation pattern is in a line with to the one mentioned in a) part of the current task.

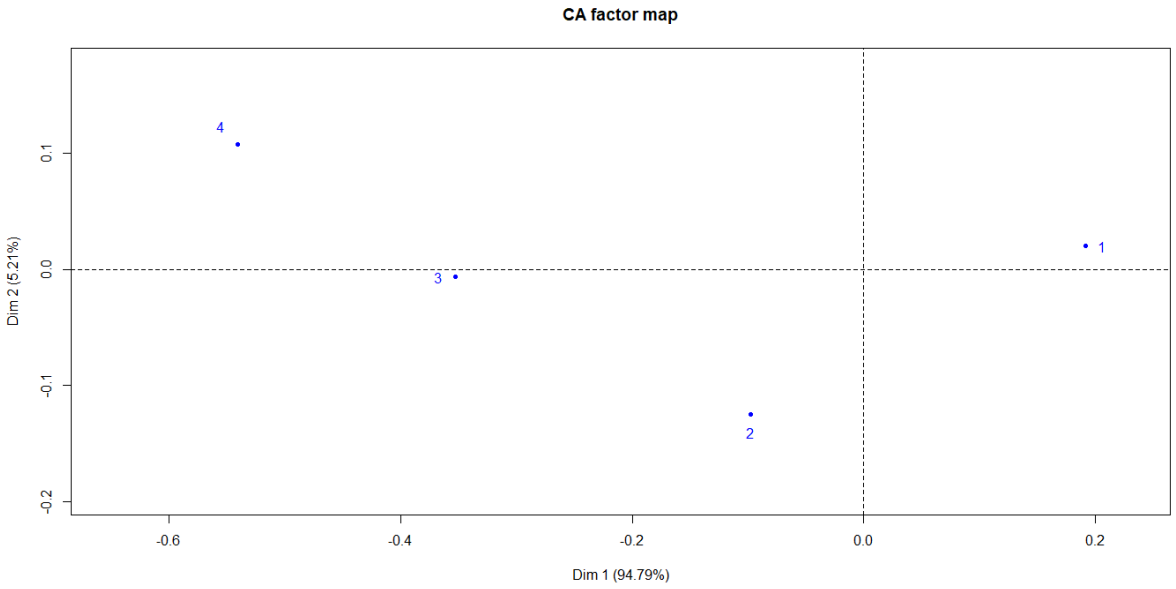
Assigment 2

Task 2

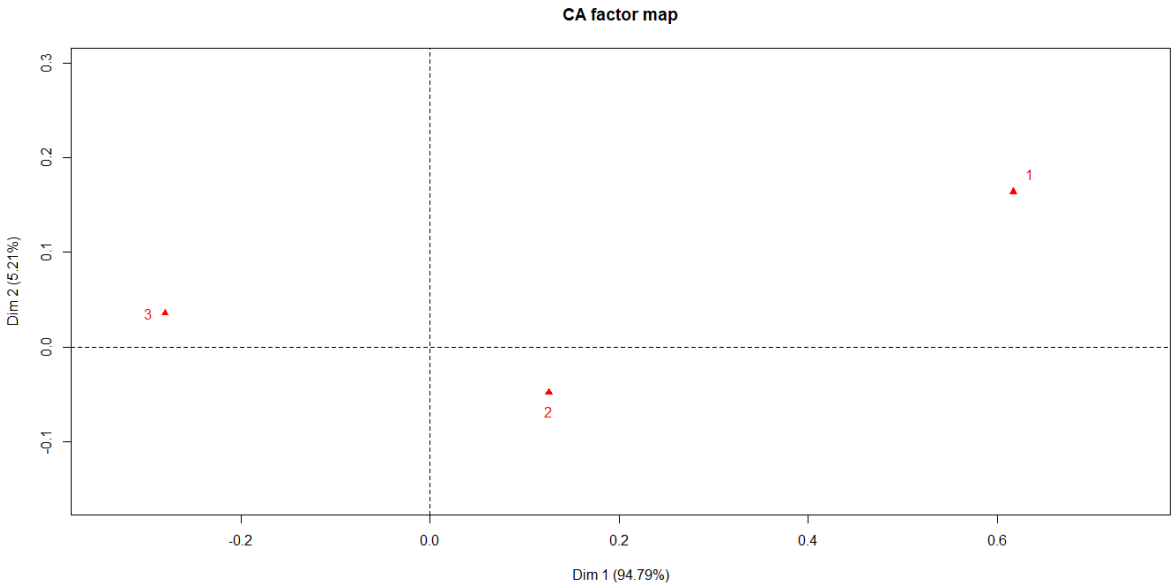
a) Do the X^2 test for independence and interpret it, see Section 2.2.2 of [1].

Chi - sqr is 43.38 with p-value: 9.810217e-08. p is very small: H0 (that both variables are independent) should be rejected in favor of H1. Conclusion: there is a relationship between the two categories - marichuana consumption and partying frequency.

b) Perform the CA, get the 2D representation of row and column profiles separately in the same graph.



CA factor map for rows



CA factor map for columns

Gives depiction of the rows, artificial 2D plane. Along Dim1: row 1 and row 4 are most distinct: 22-11-1 vs 5-9-2. The relative distribution is the completely different.

c) Analyze the patterns obtained in item 2b. Focus on the total variability, similarities/dissimilarities and the conclusions that can be made from the simultaneous representation of rows and columns. See examples, [1], pp. 92-125.

Based on the CA factor map for rows, we can conclude that the rows 1 and 4 are totally different. Whereas, rows 4 and 3 are more likely to be similar.

1 2 3

40 213 118

1 2 3

0 17 32

Indeed the values for each row are completely different. Let's check the same for columns: columns 1 and 3 should be different

1 2 3 4

40 3 1 0

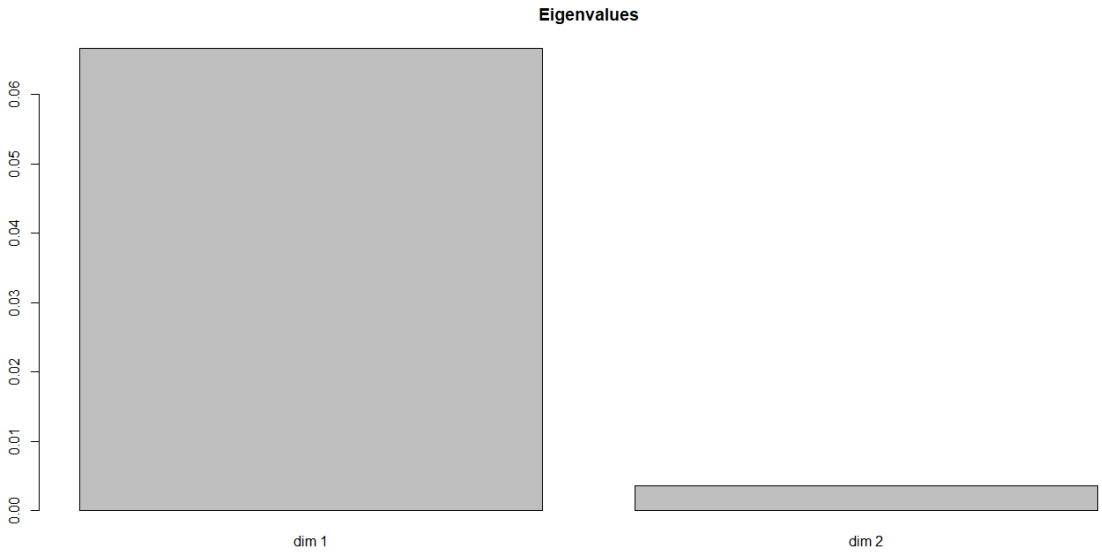
1 2 3 4

118 40 54 32

Here we also observe a big difference in values, although not so extreme as in previous example.

d) Provide the table and graph of eigenvalues, justify the choice of principal components.

barplot(daCA$eig[,1], main="Eigenvalues", names.arg=paste("dim", 1:nrow(daCA$eig))). As we can see from the barplot, the first dimension is mostly responsible for the variation in data. Second principal component contributes little to the data variance.



Eigenvalues barplot

	Dim 1	Dim 2
1	0.9887123	0.0112877399
2	0.3777105	0.6222894630
3	0.9996903	0.0003096752
4	0.9617638	0.0382361729

cos^2 values for rows

	Dim 1	Dim 2
1	0.9341272	0.06587283
2	0.8708900	0.12910998
3	0.9840762	0.01592376

cos^2 values for columns

e) Discuss the quality of the CA representation based on cos2 for rows and columns daCArowcos2.

From the table above we can conclude that the majority of the rows are well represented by the first dimension. The second row, however, is better represented by the second dimension. In general, all of the columns are well represented by the first dimension.

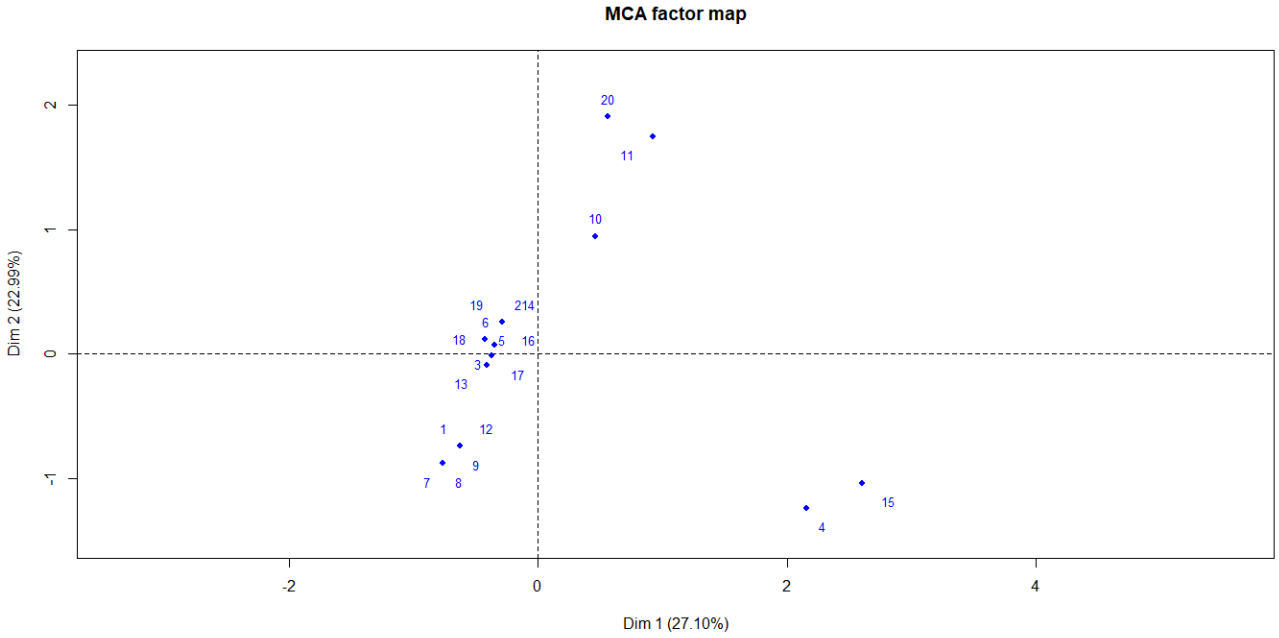
Assignment 3: Multiple correspondence analysis

When we have more than 2 qualitative variables!

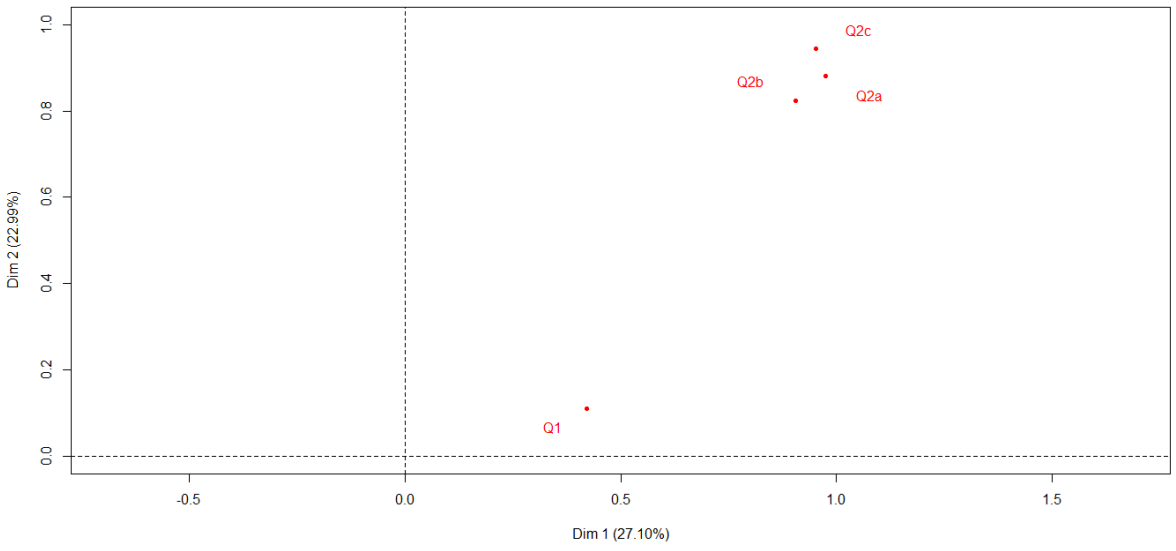
Task 1: Get multivariate data set

Task 2

a) Conduct the MCA. Visualize individuals and categories



Individuals MCA factor map



Variables factor map

b) Provide a detailed interpretation of the obtained patterns. Focus on variability of individuals and categories, comment on the extreme cases.

Here we get 27% and 23% variation explained by Dim1 and Dim2 respectively. Again we can compare the objects on the graph for confirmation of difference/similarity of objects. Close to origin we have majority of the similar objects. However, objects that are far away from each other should be different. For example, the answers to all of the questions # 7 and 20 are different:

Q1 Q2a Q2b Q2c

7 5 5 5 5

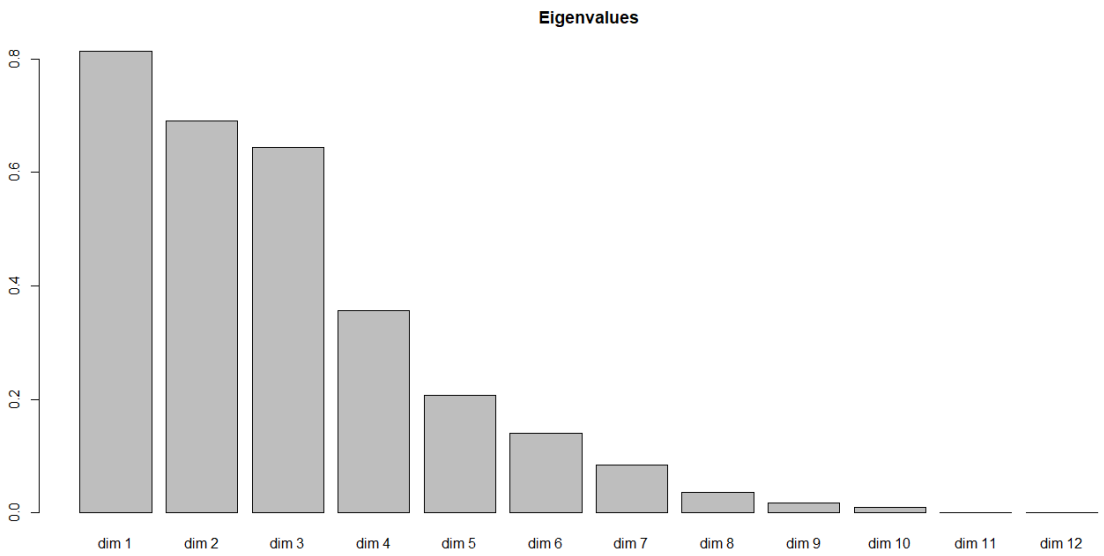
Q1 Q2a Q2b Q2c

20 4 3 2 3

Similarly, we can compare the questions themselves. For instance, Q2a, Q2b, and Q2c were answered similarly. Whereas, Q1 should have different answers that the previous ones.

c) Provide a table of eigenvalues, comment on the values of the largest ones and justify the choice of principal components. Do you need to look at the PCs other than the firrst two ones?

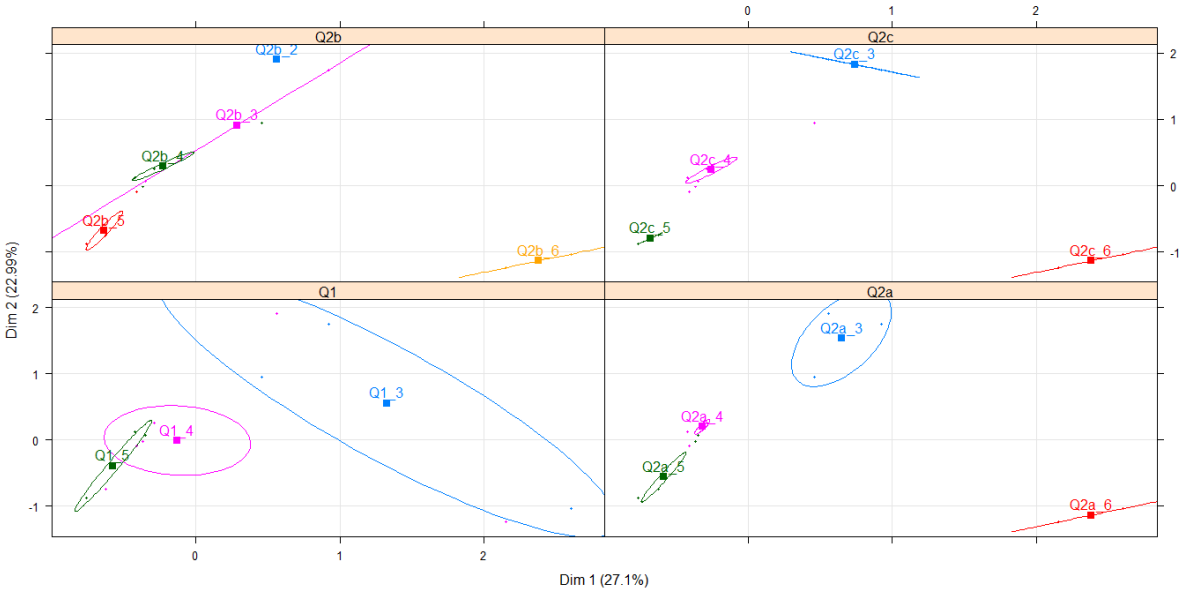
The significant eigenvalues are for dimensions 1, 2, 3 and 4 that comprise together 83% of variation. Therefore, we cannot justify 2d space only based on the first 2 dimensions. Because other dimensions are also substantial. Thus we can make mistakes in interpreting the graphs based on only two dimensions.



Eigenvalues for MCA dimensions

d) Draw the confidence ellipses around the categories and interpret the results

In general, the ellipses answer the question of whether an answer to one question is affected by another answer. Hence all of the answers to questions except to Q1 are independent. However, the answer to the question Q1 as category 4 might be influenced by category 5 or vice versa.

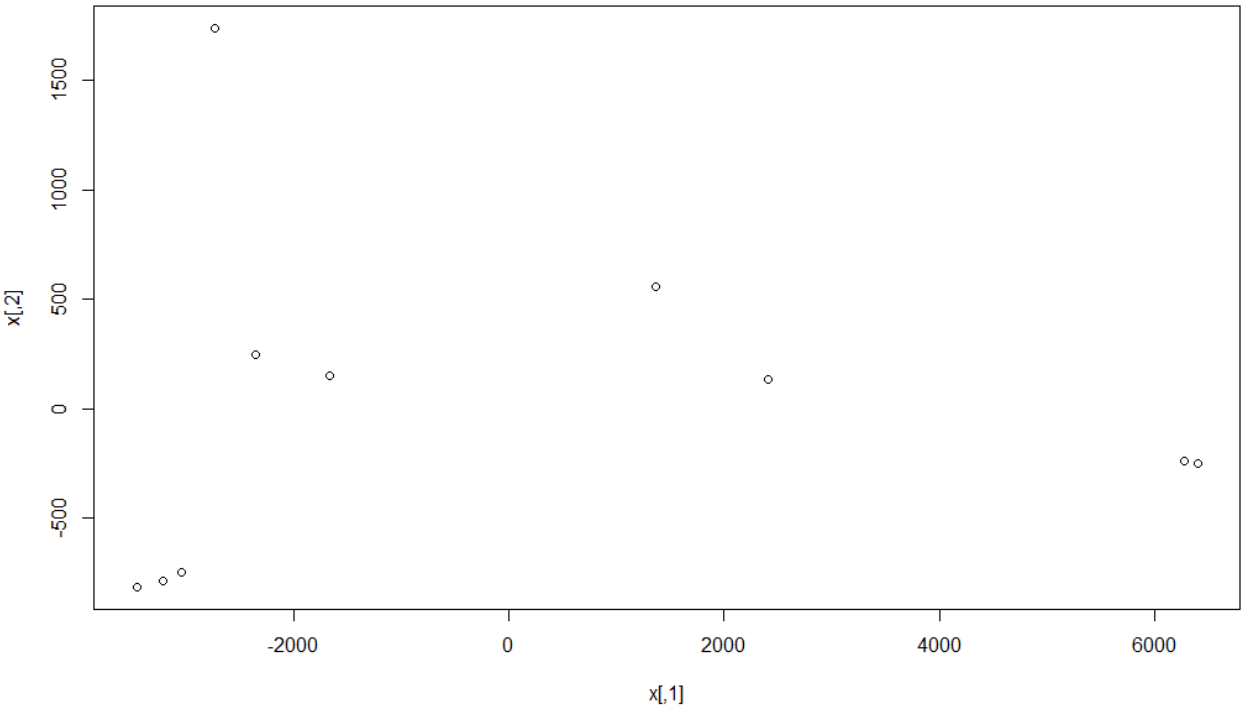


Confidence ellipses around categories

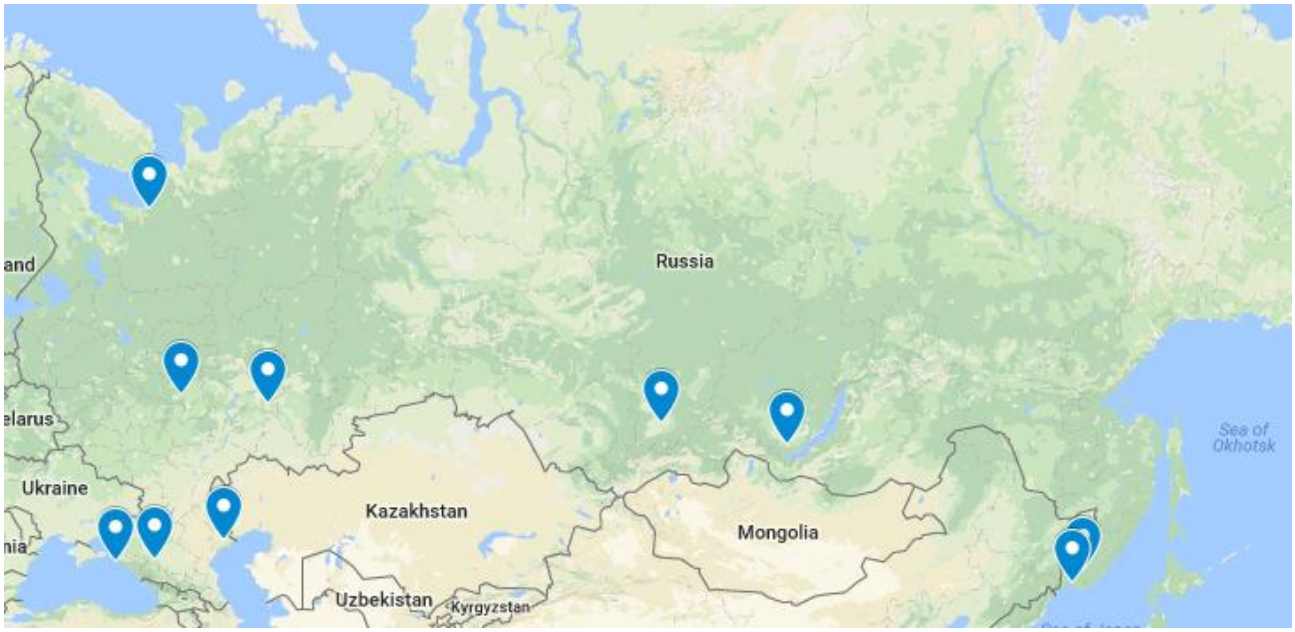
Assignment 4: Multidimensional Scaling

Task 1: Do the classical multidimensional scaling using command cmdscale from MASS package

a) Plot a two-dimensional MDS configuration representing the cities. Compare the result with the actual geographical location of the cities across the country.



MDS Plot of 10 Russian cities



Cities' location according to google maps

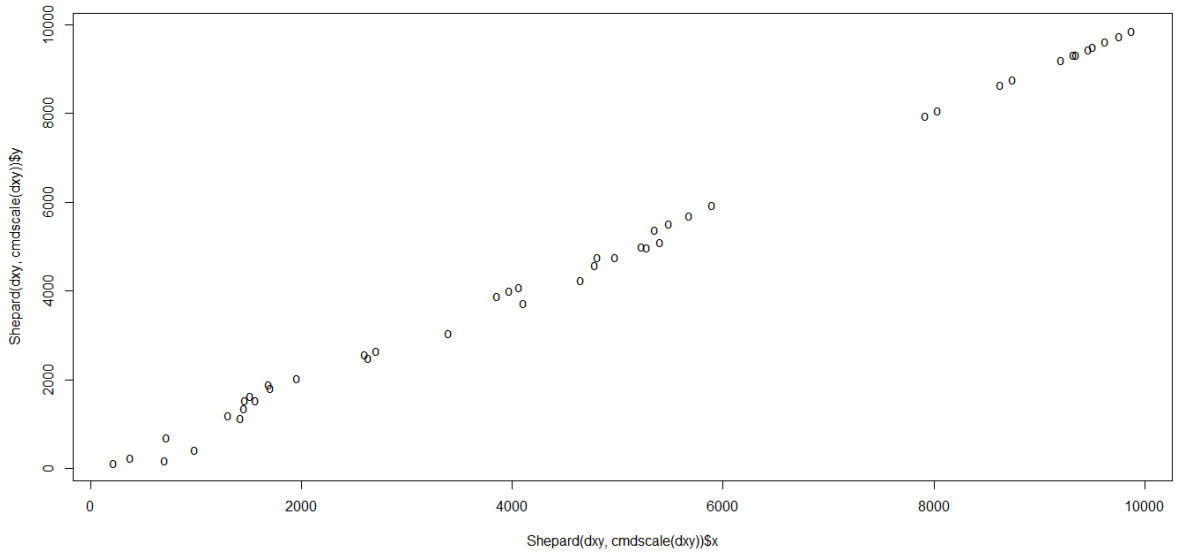
By comparing the MDS plot and locations of the cities on google map we can see the similarity between both.

b) Based on the computed eigenvalues, discuss the quality of representation in the 2D space.

Based on the generated 2D plane representing the locations of the cities and actual places on the map we can conclude that the quality of the MDS is high.

c) Plot the Shepard diagram and discuss it.

Shepard diagram represents the quality of multidimensional scaling. Ideally it should be a bissecting line. In our case with the distances of the cities we can observe slight deviations. But in general the points lie on the bissecting line. Thus, the quality of the multidimensional scaling that we derived is high.



Shepard diagram

d) Check whether the MDS configuration you obtained does restore the original distances in a sufficiently high dimensional space.

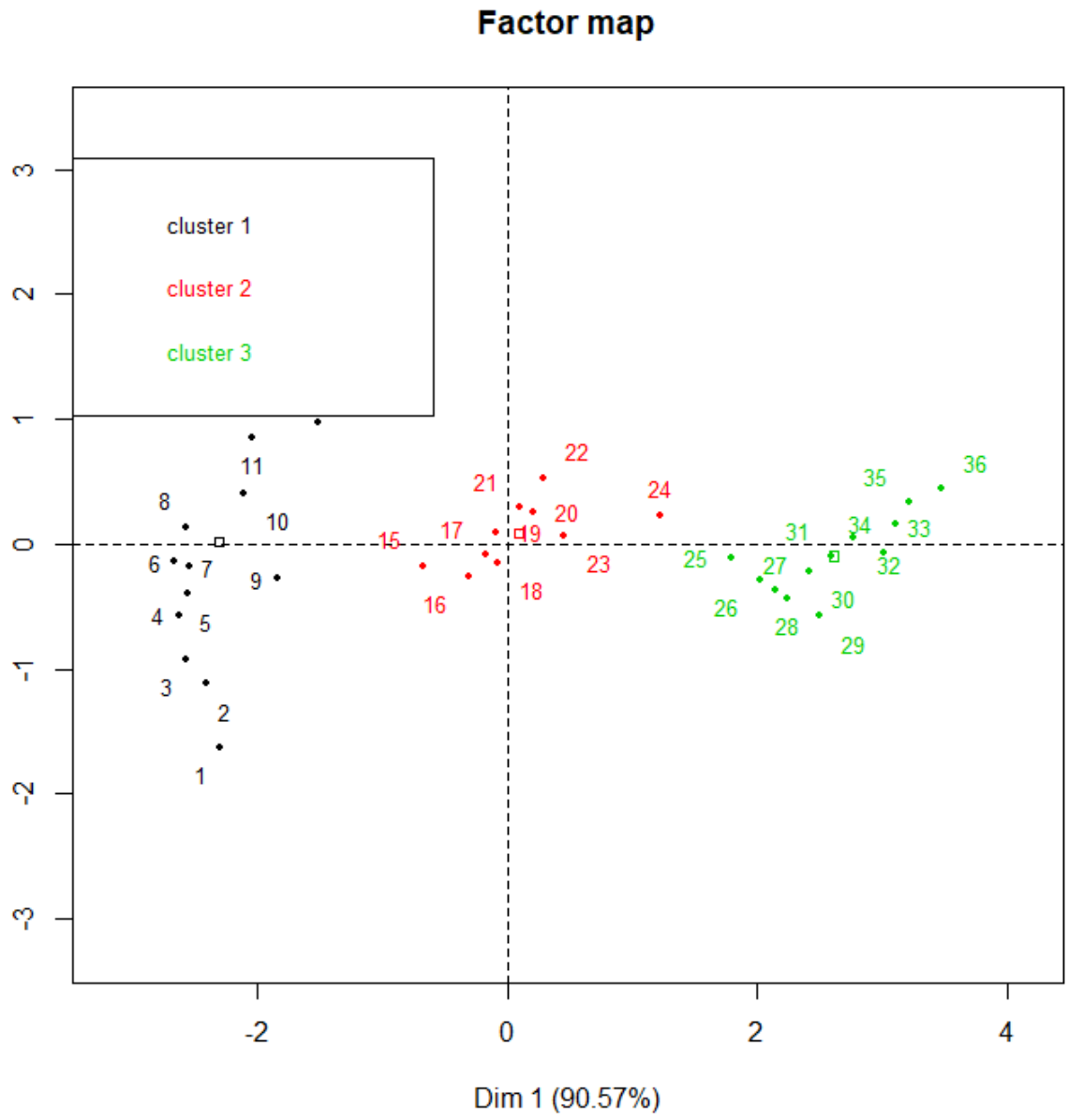
In contrast to the previous results, the max difference is relatively high (14570.85) stating that the observed distance matrix is not recovered by the seven-dimensional classical scaling.

Assignment 5: k-means

Task 1: Do the hierarchical clustering (preceded by the PCA) using command HCPC from FactoMineR package

a) Clearly name the recommended (by HCPC) clusters

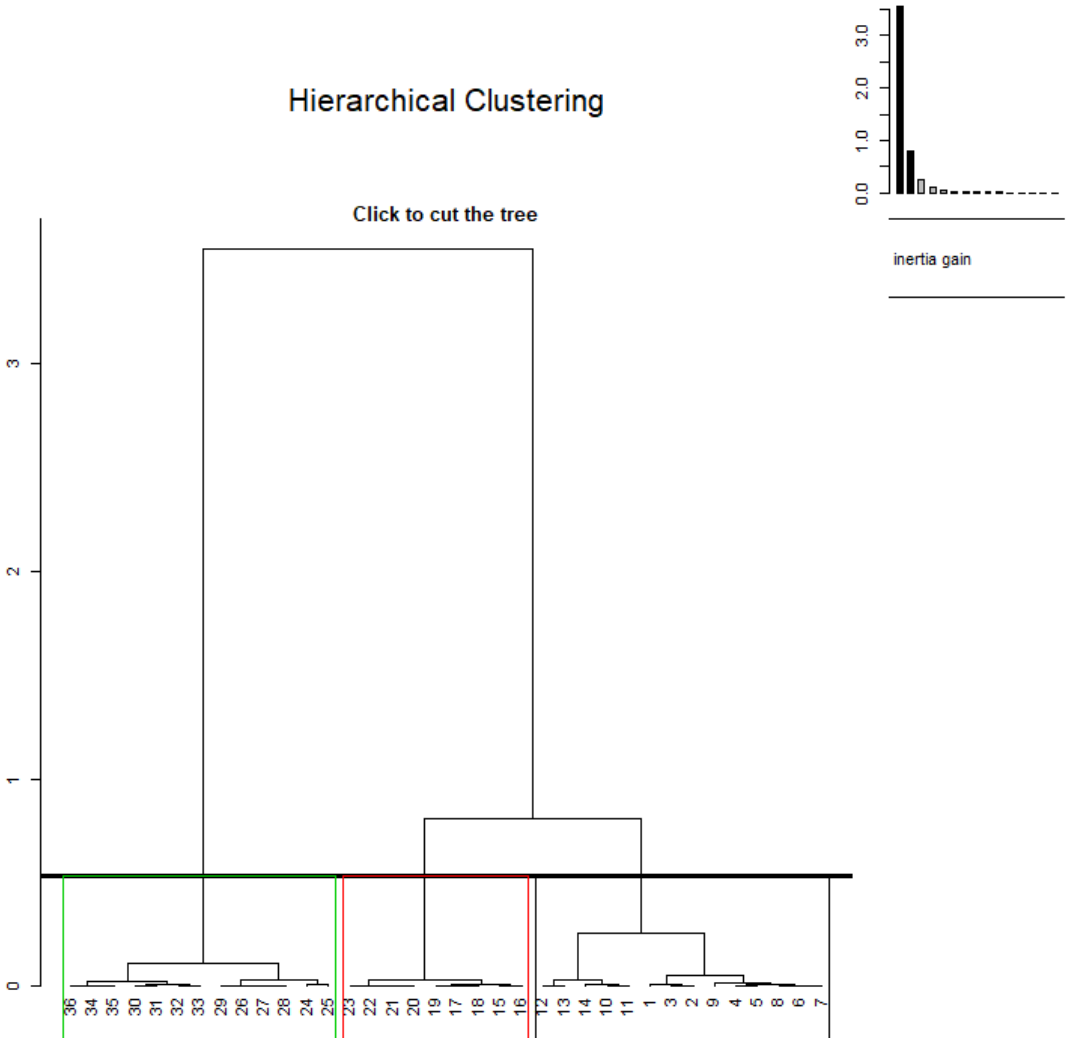
HCPC command recommended to divide the dataset into 3 distinct clusters mainly by considering the first principal component: cluster 1, cluster 2, and cluster 3.



Recommended clusters by HCPC

b) Explain the meaning of the barplot in the upper-right corner of the output.

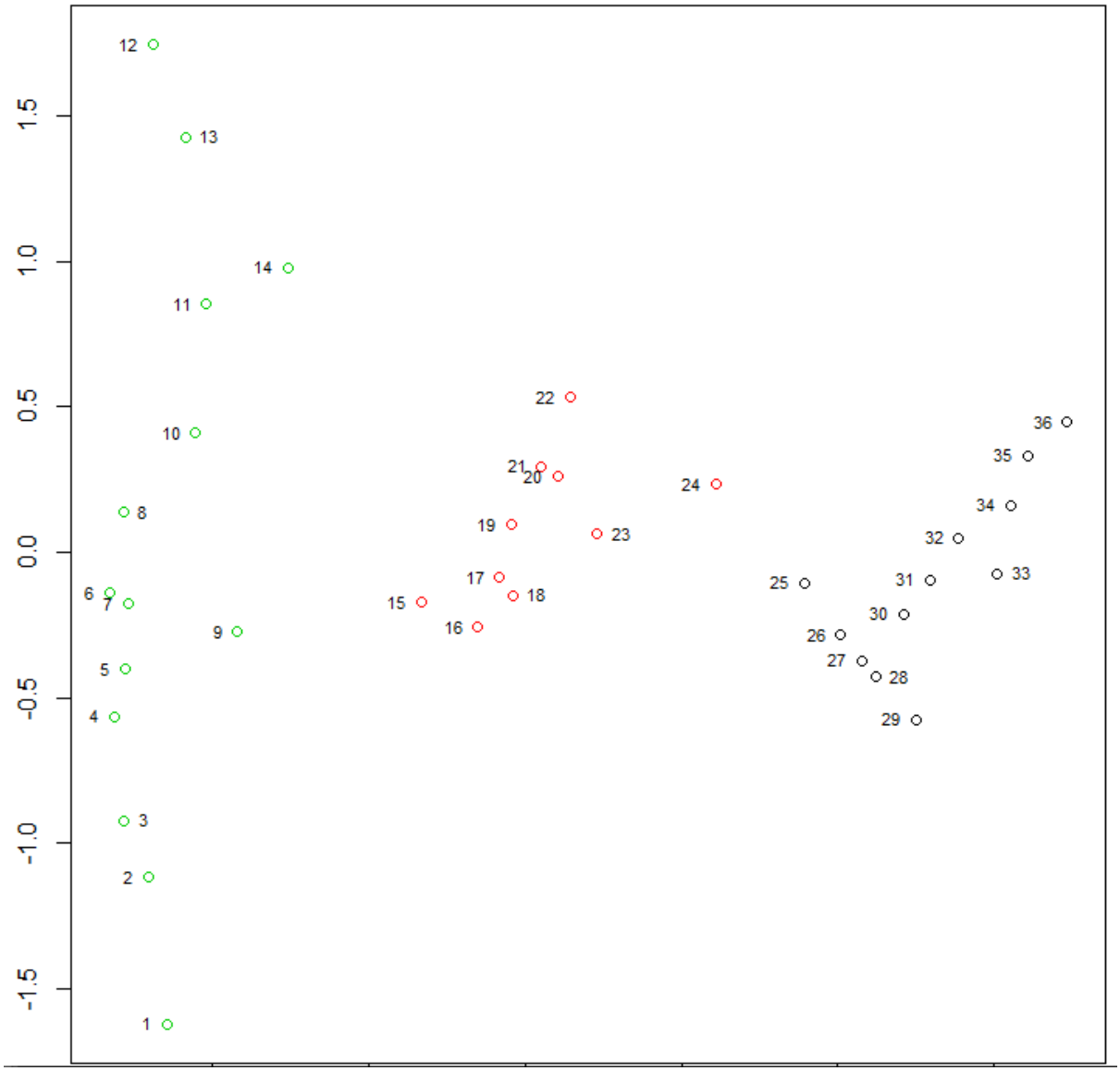
The bar plot in hierarchical clustering graph represents how to choose to how many clusters to divide the dataset. The bars illustrate the level of inertia / explained variance explained by each hierarchical level. For example, if we change from 1st hierarchy to the 2 one, we will gain reliable amount of variance. If between cluster inertia change is small then it does not make much sense to increase number of clusters. This barplot helps us to make a compromise between k and between cluster variance. Therefore, algorithm recommends to cut cluster hierarchy on bold line: 3 clusters overall, 2nd hierarchy level.



Dendrogram

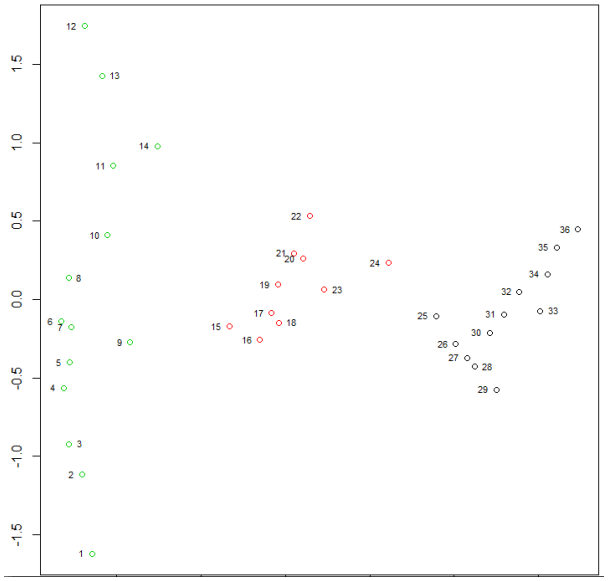
Task 2: Perform the K-means clustering, choosing K according to the results of hierarchical clustering.

a) Plot the results

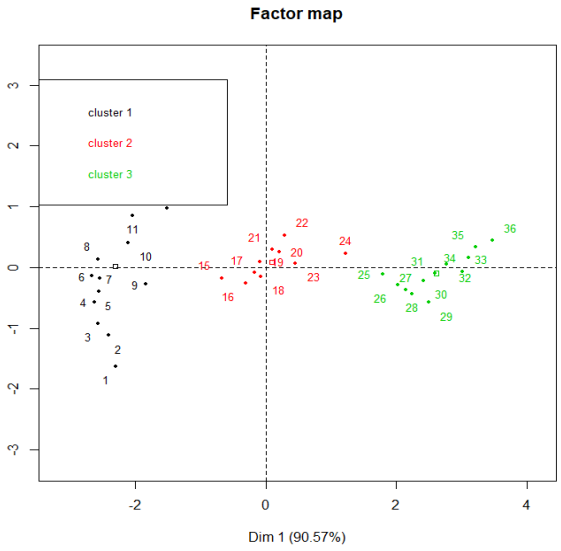


k-means clustering result

b) Compare distribution of points over clusters with that of hierarchical approach



k-means



HCPC

Now we can compare the result with Hclustering. After comparing both hierarchical and k-means clustering results we conclude that there is no difference in distribution of points over clusters.