Correspondence analysis and Multidimensional scaling

Daria Plewa

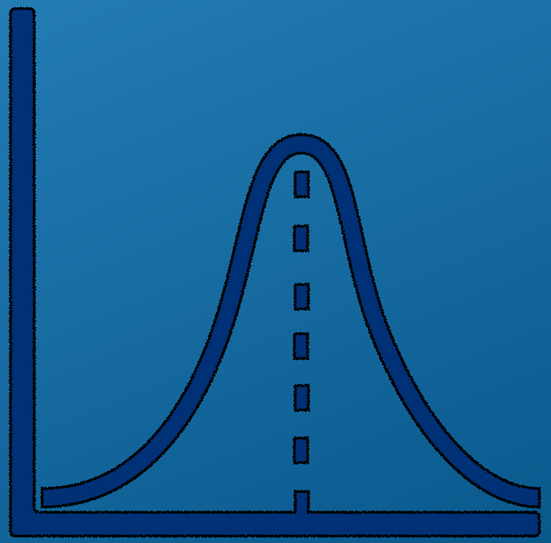
Presentation Plan

1. Description of the Methods
2. Data
3. Correspondence analysis
4. Multivariate Correspondence Analysis
5. Multidimensional scaling
6. Conclusions
7. Sources

Description of methods

Correspondence analysis

Correspondence analysis is a descriptive and exploratory technique for analysing qualitative variables that aims, among other things, to present cross-tabulation relationships in visual form. Correspondence analysis can be particularly useful for extensive contingency tables to illustrate relationships between categories of variables. It is often used in marketing, social but also economic research, where qualitative variables predominate.



$$p_{ij} = p_{i\cdot}p_{\cdot j}, \quad i \in \{1 \dots k\}, \quad j \in \{1 \dots l\}.$$

p_{ij} - the probability of observing the first variable at i and at the same time the second variable at j

p_i - probability of observing variable one at i

p_j - the probability of observing variable two at j

$$\hat{e}_{ij} = \frac{\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

In order to assess which variables are more or less frequent than independence would imply, we will determine the standardised Pearson residuals, and replace the probabilities with their frequency scores.

$\hat{p}$ - number of observed events divided by the number of all observed events

Large positive values $\hat{e}_{ij}$ correspond to high co-occurrence.

$$E_{k\times l} = U_{k\times k} \sum_{k\times k} V^T_{l\times l}$$

Then the matrix $E = \hat{e}_{ij}$ I will present it in graphical form using the so-called Biblot. In other words, we determine the SVD decomposition of the matrix E.

Columns of the matrix U_{kxk}*are the eigenvectors of the matrix* $E^T E$, and the columns of the matrix V are the eigenvectors of the matrix EE^T. On the diagonal of the diagonal matrix σ there are so-called singular values equal to the elements of the eigenvalues of the matrix $E^T E$ i EE^T.

Multidimensional scaling

Multidimensional scaling is used to find structure in a set of distance measures between individual objects or observations. This is done by assigning observations to particular locations in a conceptual space (usually two- or three-dimensional) in such a way that the distances between points in the space as closely as possible correspond to given measures of dissimilarity. In many cases, the dimensions of this conceptual space can be interpreted and used to better understand the data.

Multidimensional Metric and Non-metric Scaling

Metric multidimensional scaling (MDS) is also based on a similarity or distance matrix between objects. However, MDS assumes that the distances in the original space are consistent with the distances in the non-dimensional space. The aim of MDS is to find a representation of the data in which the distances between points in the non-dimensional space are as consistent as possible with the distances between objects in the original space. Metric scaling uses various optimisation methods, such as the gradient method or eigenvalue analysis.

Non-metric multidimensional scaling (NMDS) is based on a similarity or distance matrix between objects. The objective of NMDS is to find a data representation in which the distances between points in a non-dimensional space are as close as possible to the original distances between objects. This scaling is 'non-metric' in that it does not assume any specific relationship between distances in the original space and distances in the non-dimensional space. NMDS is based on the gradient method and can be applied to both metric and non-metric data.

The effects of these scaling may vary depending on the characteristics of the data and the metrics used. Both scaling aim to preserve the relative distances between objects, but differ in the interpretability of the results. Non-metric scaling can generate representations that do not accurately reflect the distances in the original data, but preserve their relative order. Metric scaling, on the other hand, attempts to represent exact distances, which can lead to more precise representations, but may be more sensitive to errors in the data

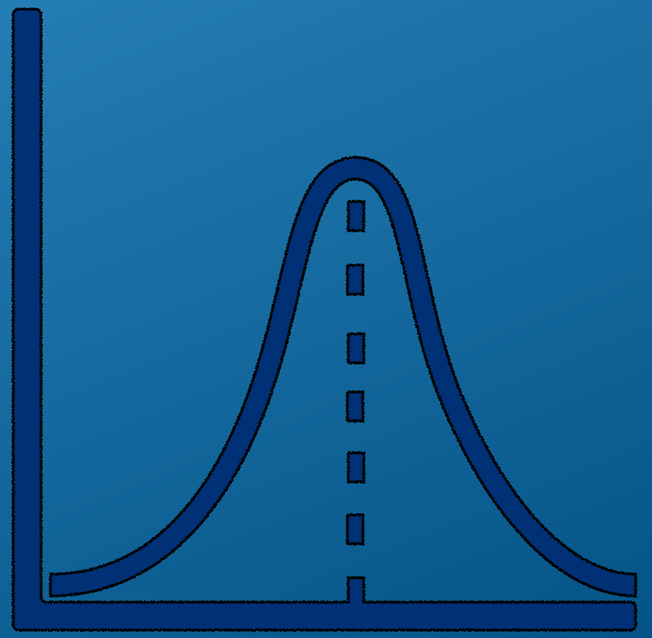
On the right is the stress formula (Standarised Sum of Squares). It minimises the value of the standardised sum of squares of the residuals.

$\hat{d}_{ik}$ - the distance between objects i and k in the new x-dimensional space

d_{ik} - are the original distances between objects

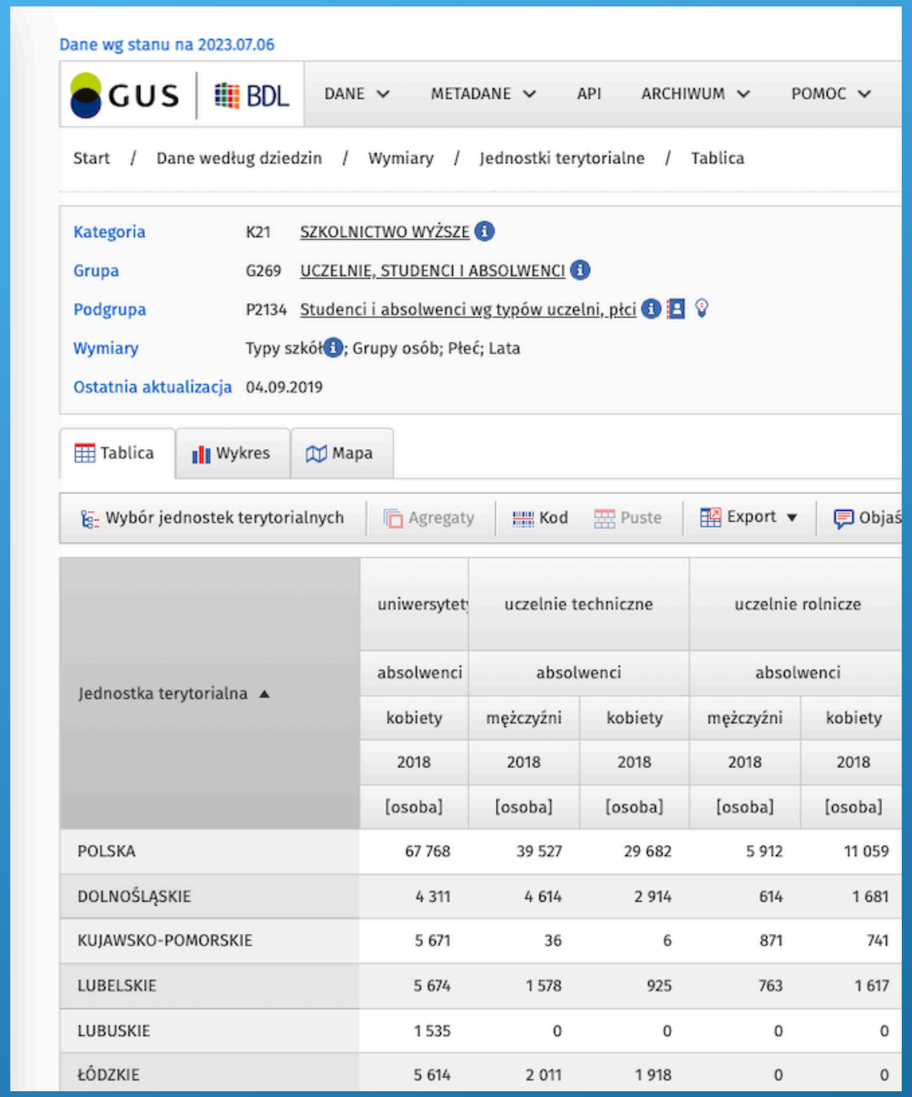
Metric multidimensional scaling is mainly used for analyses of quantitative data when non-metric scaling is used for qualitative data.

$$stress = \frac{\sum_{i,k}(d_{ik} - \tilde{d}_{i,k})^2}{\sum_{i,k} d_{ij}^2}$$



Data description

GUS



Dane wg stanu na 2023.07.06

GUS | BDL DANE METADANE API ARCHIWUM POMOC

Start / Dane według dziedzin / Wymiary / Jednostki terytorialne / Tablica

Kategoria	K21	SZKOLNICTWO WYŻSZE
Grupa	G269	UCZELNIE, STUDENCI I ABSOLWENCI
Podgrupa	P2134	Studenci i absolwenci wg typów uczelni, płci
Wymiary	Typy szkół; Grupy osób; Płeć; Lata	
Ostatnia aktualizacja	04.09.2019	

Tablica | Wykres | Mapa

Wybór jednostek terytorialnych | Agregaty | Kod | Puste | Export | Objaś

Jednostka terytorialna ▲	uniwersytet	uczelnie techniczne		uczelnie rolnicze	
	absolwenci	absolwenci		absolwenci	
	kobiety	mężczyźni	kobiety	mężczyźni	kobiety
	2018	2018	2018	2018	2018
	[osoba]	[osoba]	[osoba]	[osoba]	[osoba]
POLSKA	67 768	39 527	29 682	5 912	11 059
DOLNOŚLĄSKIE	4 311	4 614	2 914	614	1 681
KUJAWSKO-POMORSKIE	5 671	36	6	871	741
LUBELSKIE	5 674	1 578	925	763	1 617
LUBUSKIE	1 535	0	0	0	0
ŁÓDZKIE	5 614	2 011	1 918	0	0

https://bdl.stat.gov.pl/bdl/start

The data for most of the Correspondence Analysis was taken from the CSO (Central Statistical Office). From there, I downloaded data on the number of university graduates, animal culls, the population census in relation to different provinces.

	breakfast	tea.time	evening	lunch	dinner	always	home
1	breakfast	Not.tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
2	breakfast	Not.tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
3	Not.breakfast	tea time	evening	Not.lunch	dinner	Not.always	home
4	Not.breakfast	Not.tea time	Not.evening	Not.lunch	dinner	Not.always	home
5	breakfast	Not.tea time	evening	Not.lunch	Not.dinner	always	home
6	Not.breakfast	Not.tea time	Not.evening	Not.lunch	dinner	Not.always	home
7	breakfast	tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
8	Not.breakfast	tea time	evening	Not.lunch	Not.dinner	Not.always	home
9	breakfast	tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
10	breakfast	Not.tea time	evening	Not.lunch	Not.dinner	Not.always	home
11	Not.breakfast	Not.tea time	evening	Not.lunch	Not.dinner	Not.always	home
12	breakfast	Not.tea time	evening	Not.lunch	Not.dinner	Not.always	home
13	breakfast	tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
14	breakfast	Not.tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
15	breakfast	Not.tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
16	breakfast	tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
17	breakfast	tea time	evening	Not.lunch	Not.dinner	Not.always	home
18	breakfast	tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home
19	breakfast	tea time	evening	lunch	Not.dinner	Not.always	home
20	Not.breakfast	tea time	Not.evening	lunch	Not.dinner	Not.always	home
21	Not.breakfast	tea time	Not.evening	Not.lunch	Not.dinner	Not.always	home

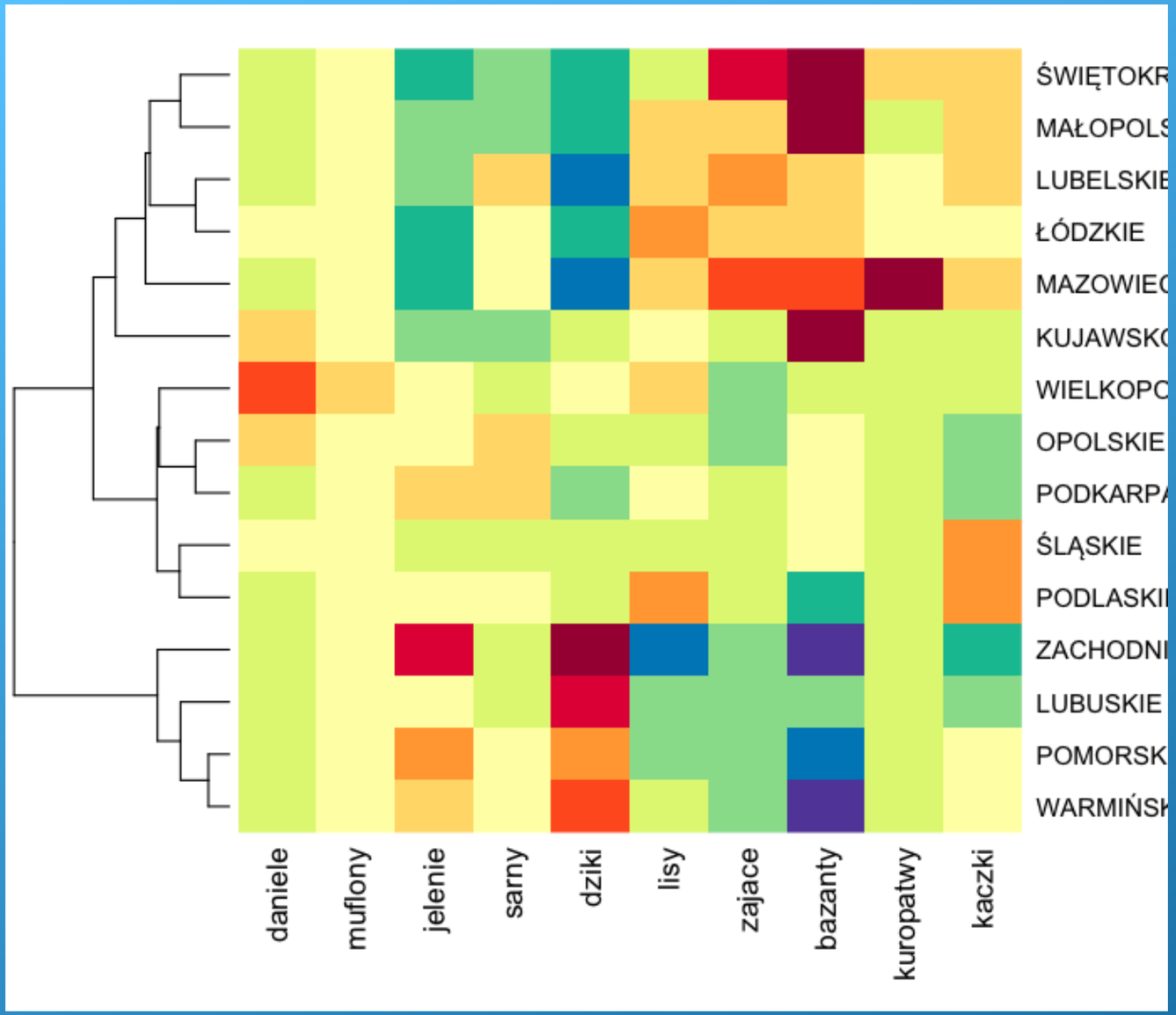
Dane tea

The tea data contains 300 records on tea drinking preferences. All data are qualitative values. The dataset contains 16 columns in which tea drinking habits (presence of sugar, type of tea, when it is drunk) are described.

Multivariate Correspondence Analysis will be performed on these data

Correspondence analysis

Number of wildlife culls by province



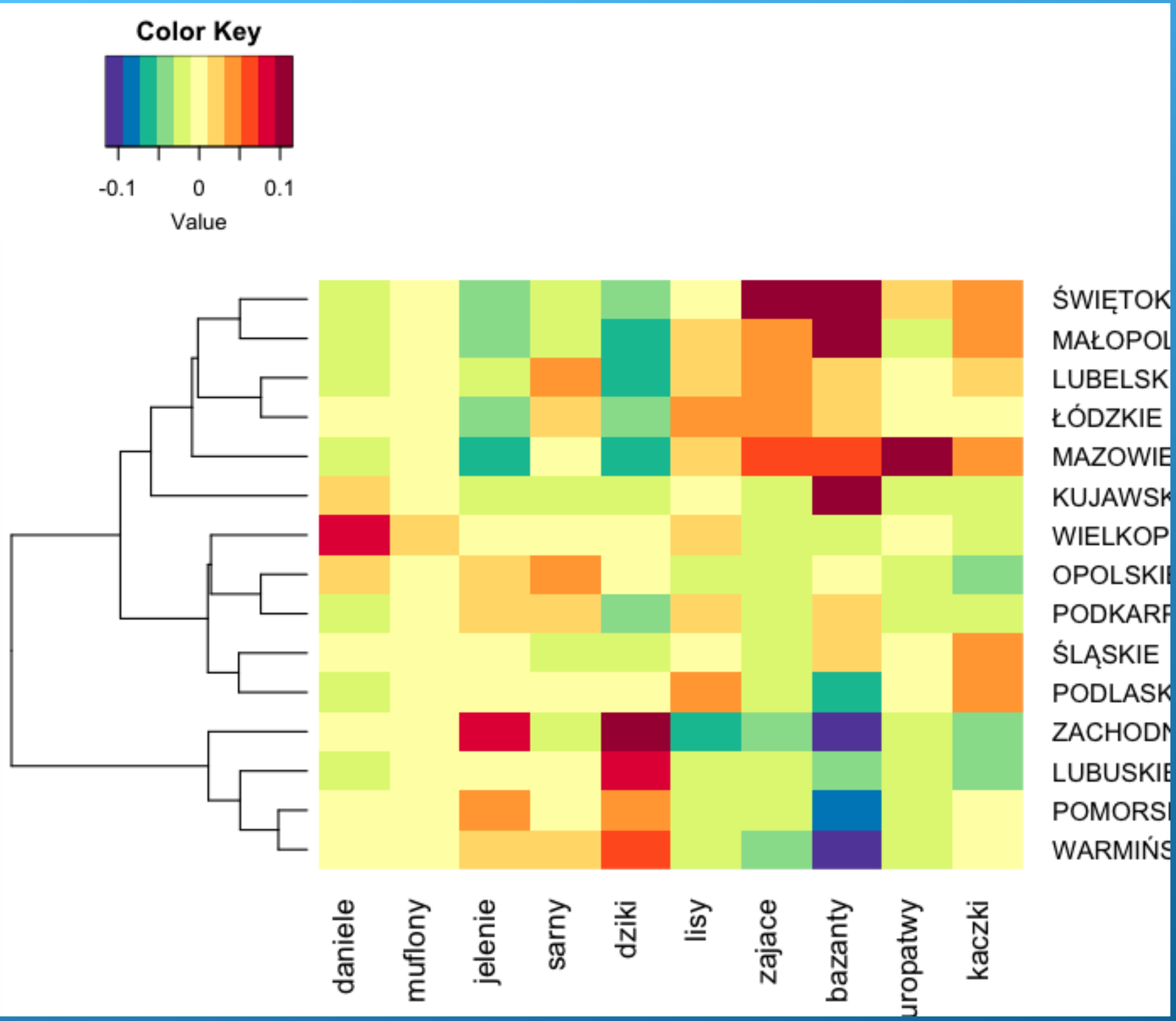

```
        Pearson's Chi-squared test

data:  data2
X-squared = 214748, df = 135, p-value < 2.2e-16
```

H0: There are no differences between provinces in animal shooting counts.

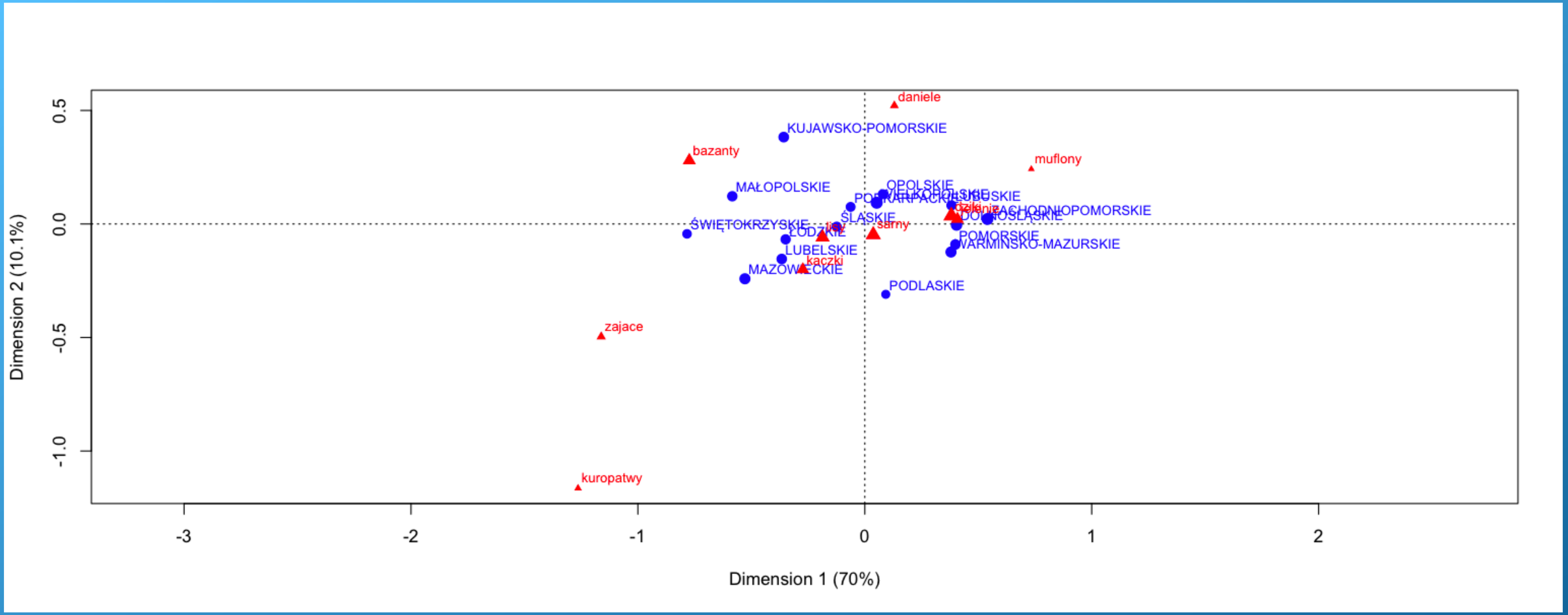
H1: There are differences in animal shooting abundances across provinces.

P-value is small (less than 0.05), we have grounds to reject H0 and accept H1 - there are differences in animal shot counts across provinces.

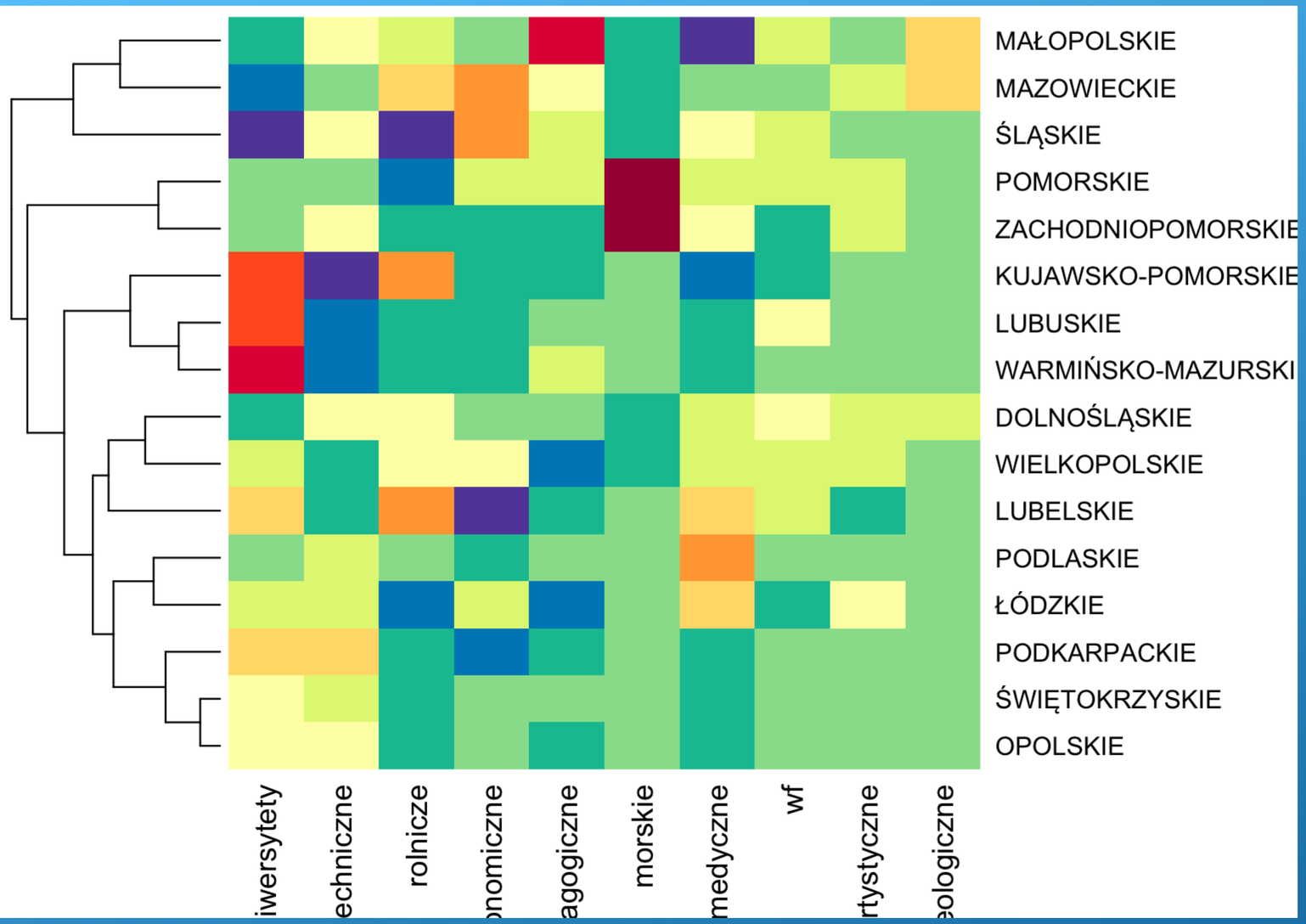
Residua



Pheasants are hunted the most in three provinces. Partridge hunting in the Mazowieckie Voivodeship and wild boar hunting in the Zachodniopomorskie Voivodeship are also popular. Less popular are fallow deer hunting in Wielkopolskie Voivodeship and deer hunting in Zachodniopomorskie Voivodeship, and wild boar hunting in Lubuskie and Warmińsko-Mazurskie Voivodeships. On the other hand, there is very little or none for pheasants in Zachodniopomorskie, Pomorskie and Warmińsko-Mazurskie.



Shooting of roe deer, ducks, wild boars and foxes occurs in all provinces. Most fallow deer are shot in Kujawsko-Pomorskie. Pheasants in Kujawsko-Pomorskie and Małopolskie. Mouflons in Zachodnio-Pomorskie, Dolnośląskie and Wielkopolskie. Pheasants, fallow deer, moufflons, hares and partridges differentiate the most between the provinces in shots, with the greatest number of hares and partridges shot in Mazovia.




```
        Pearson's Chi-squared test

data:  data2
X-squared = 333135, df = 135, p-value < 2.2e-16
```

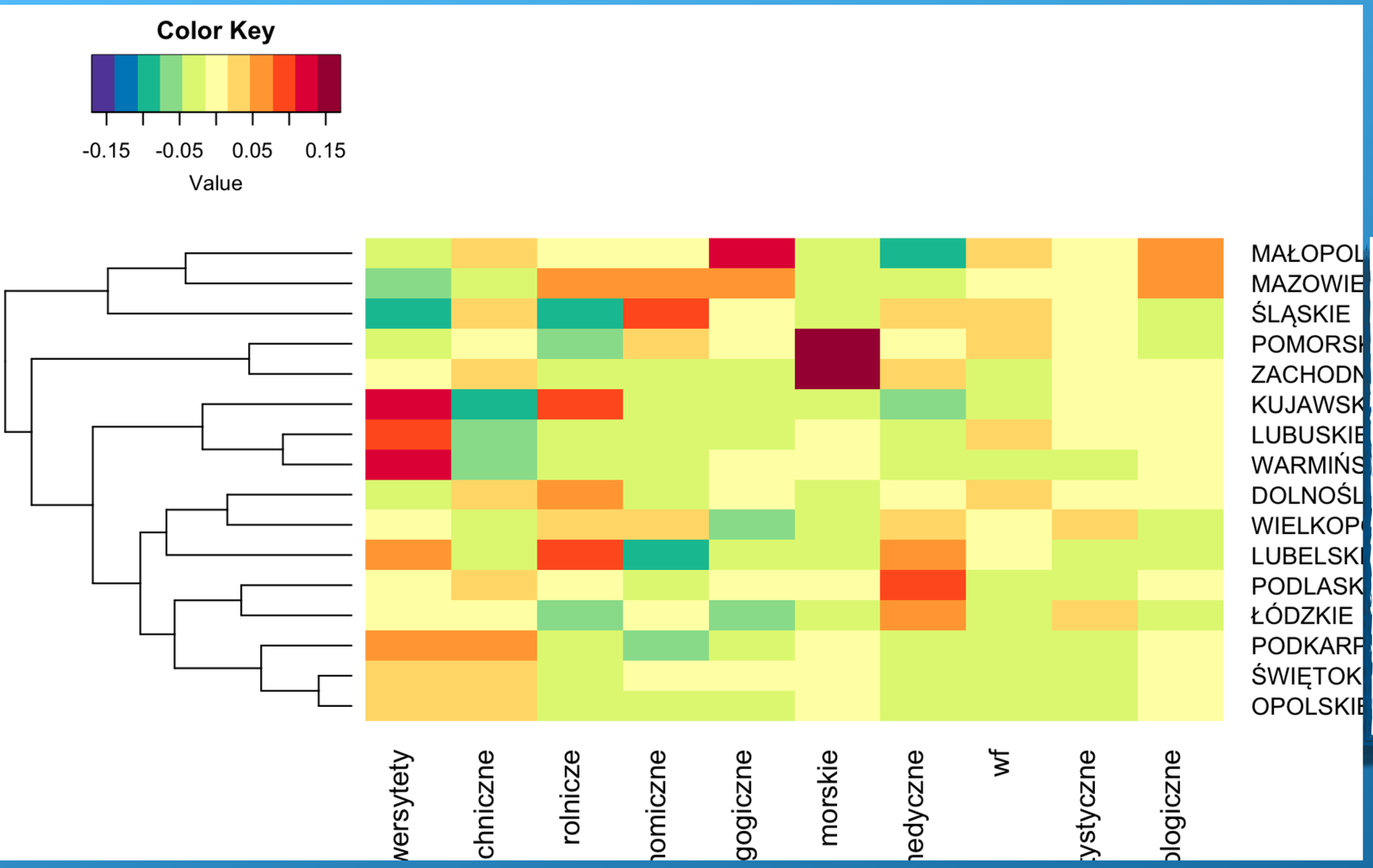
Number of students of different types of HEIs by province

H0: There are no differences between provinces and the number of students at different types of universities.

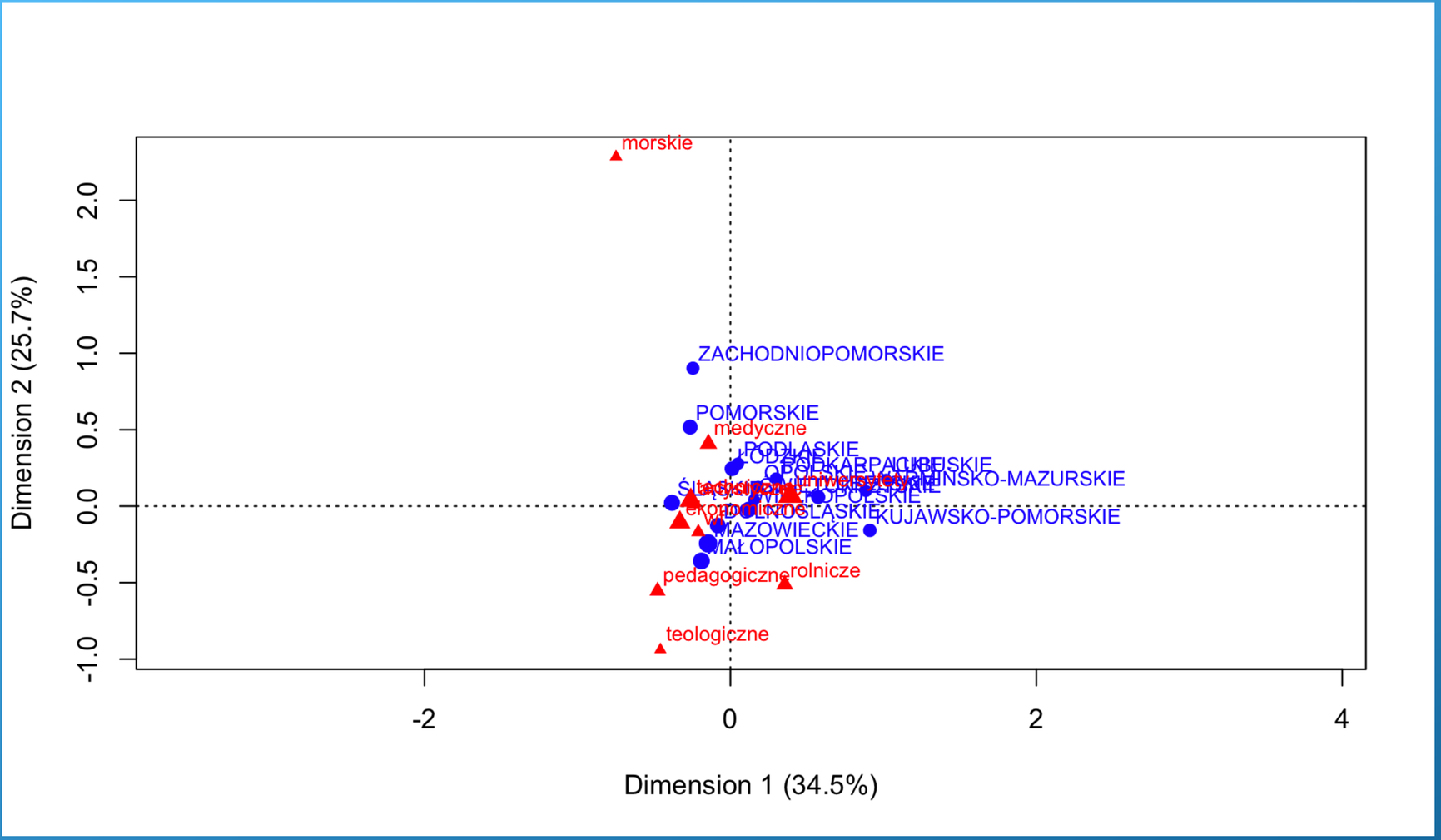
H1: There are differences between provinces and the number of students at different types of universities.

P-value is small (less than 0.05), we have grounds to reject H0 and accept H1 - there are differences in the numbers of students at different types of HEIs in different provinces.

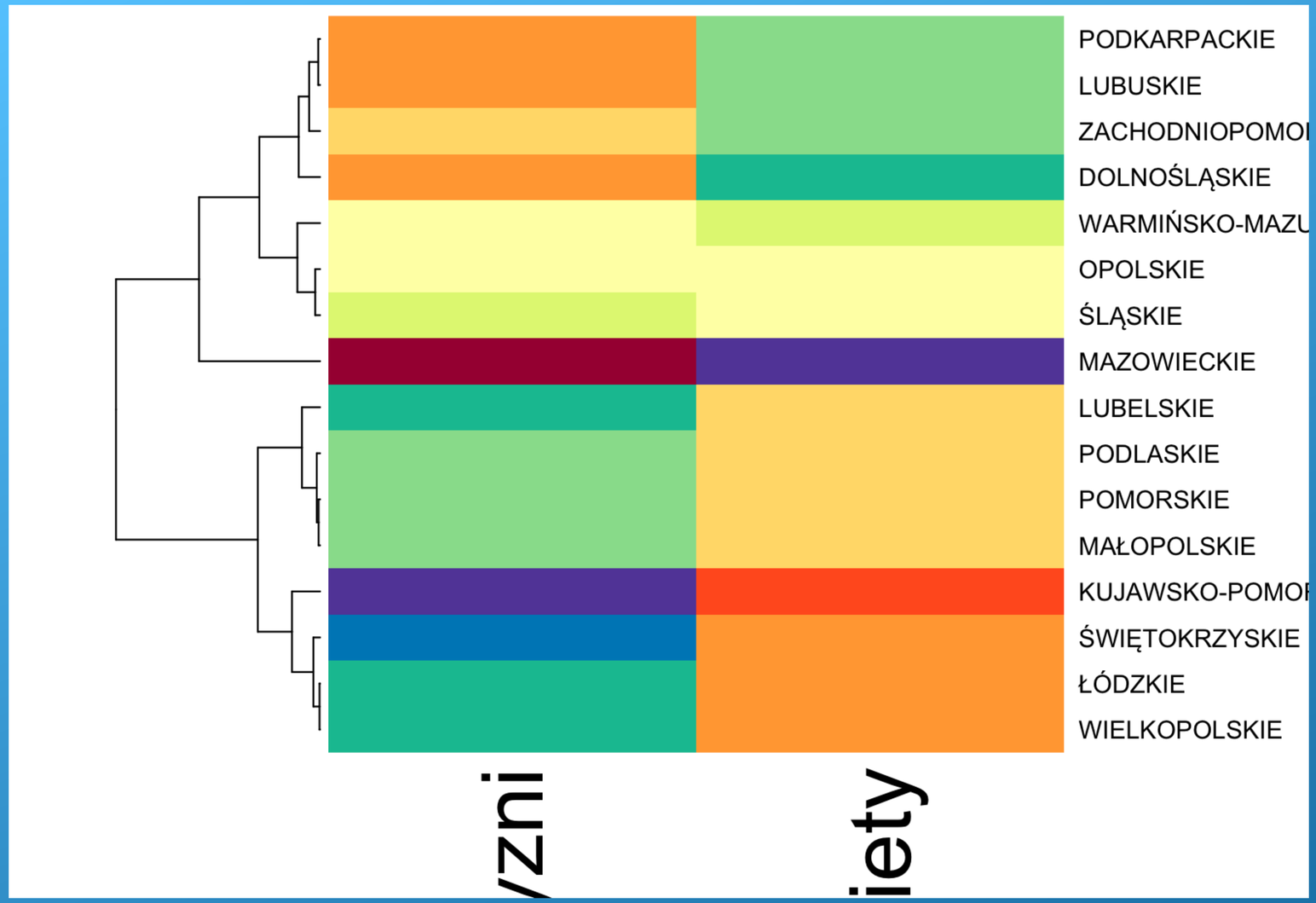
Residua



The largest number of students for maritime universities is in w. Pomorskie and Zachodniopomorskie. The largest number of students for universities is in w. Kujawskie, Lubuskie and Warmińsko-Mazurskie. The largest number of students for pedagogical universities is in w. Małopolskie.



The most diverse universities are maritime universities, agricultural universities, pedagogical universities and theological universities. Most students study at maritime universities in Zachodniopomorskie and Pomorskie. At other types of universities it is difficult to see significant differences between the voivodeships.



Gender by province among students

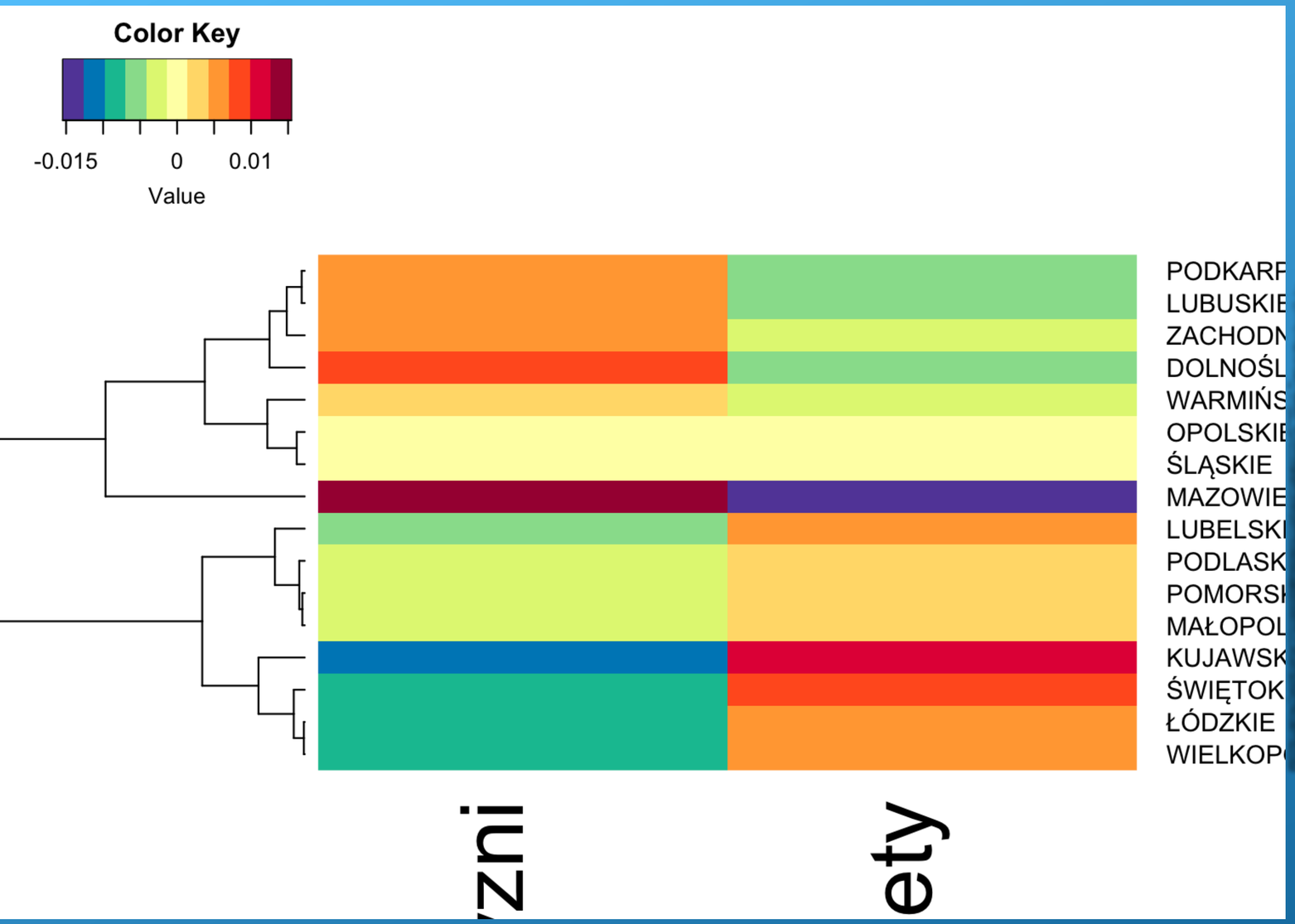
H0: There are no differences between provinces and student gender.

H1: There are differences between provinces and gender of students.

P-value is small (less than 0.05), we have grounds for rejecting H0 and accepting H1 - there are differences in student gender across provinces. (Between male and female numbers)

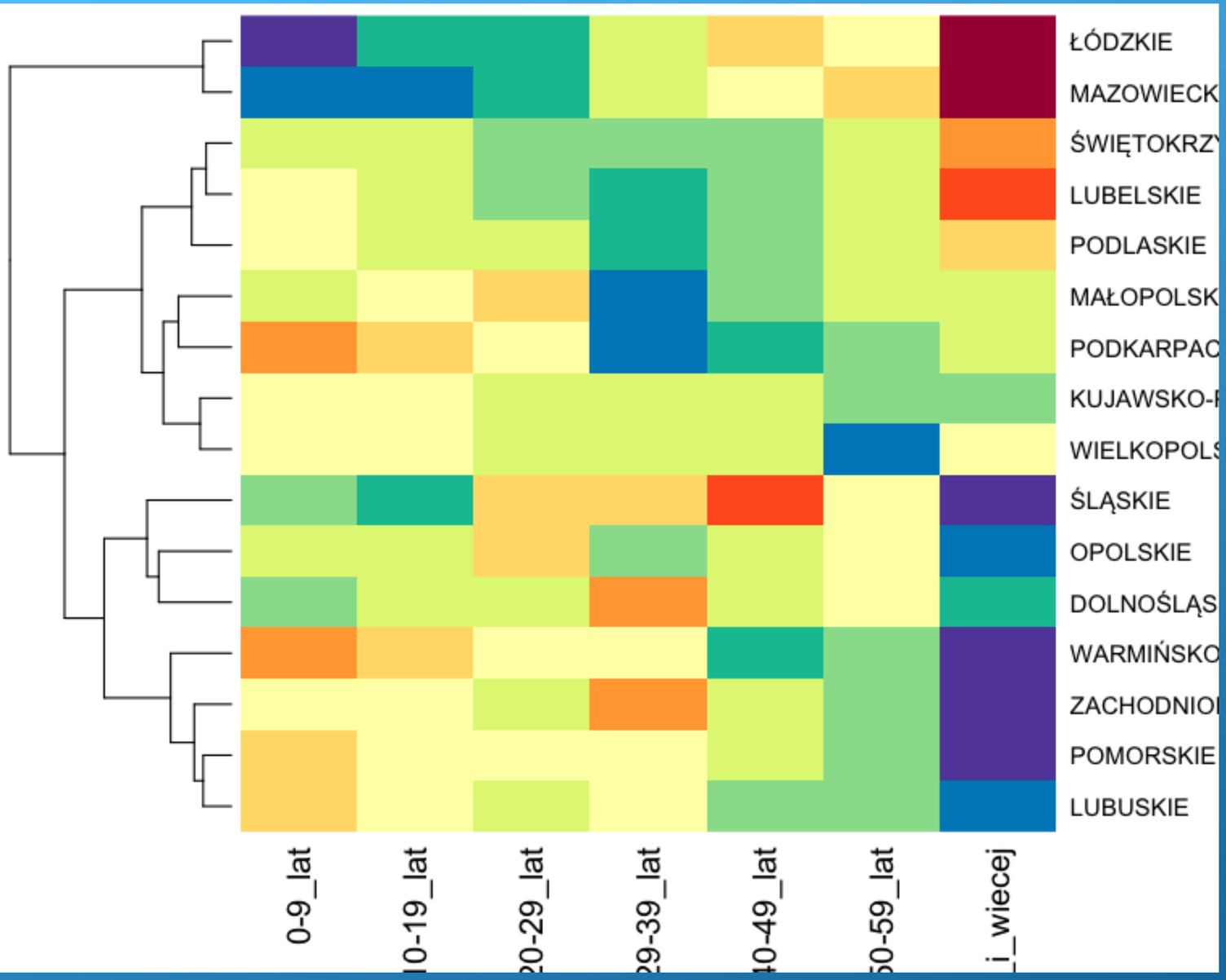
```
        Pearson's Chi-squared test

data:  data2
X-squared = 1691.2, df = 15, p-value < 2.2e-16
```

Residua



More men than women study in the Podkarpackie, Lubuskie, Dolnośląskie and Mazowieckie voivodeships. More women than men study in the Kujawski, Świętokrzyskie, Łódzkie, Wielkopolskie and Lubelskie voivodeships. The largest gender disproportion is found in Dolnośląskie, Kujawskie and Mazowieckie. And the most similar numbers of women and men are in Opolskie and Śląskie.

Population census by province



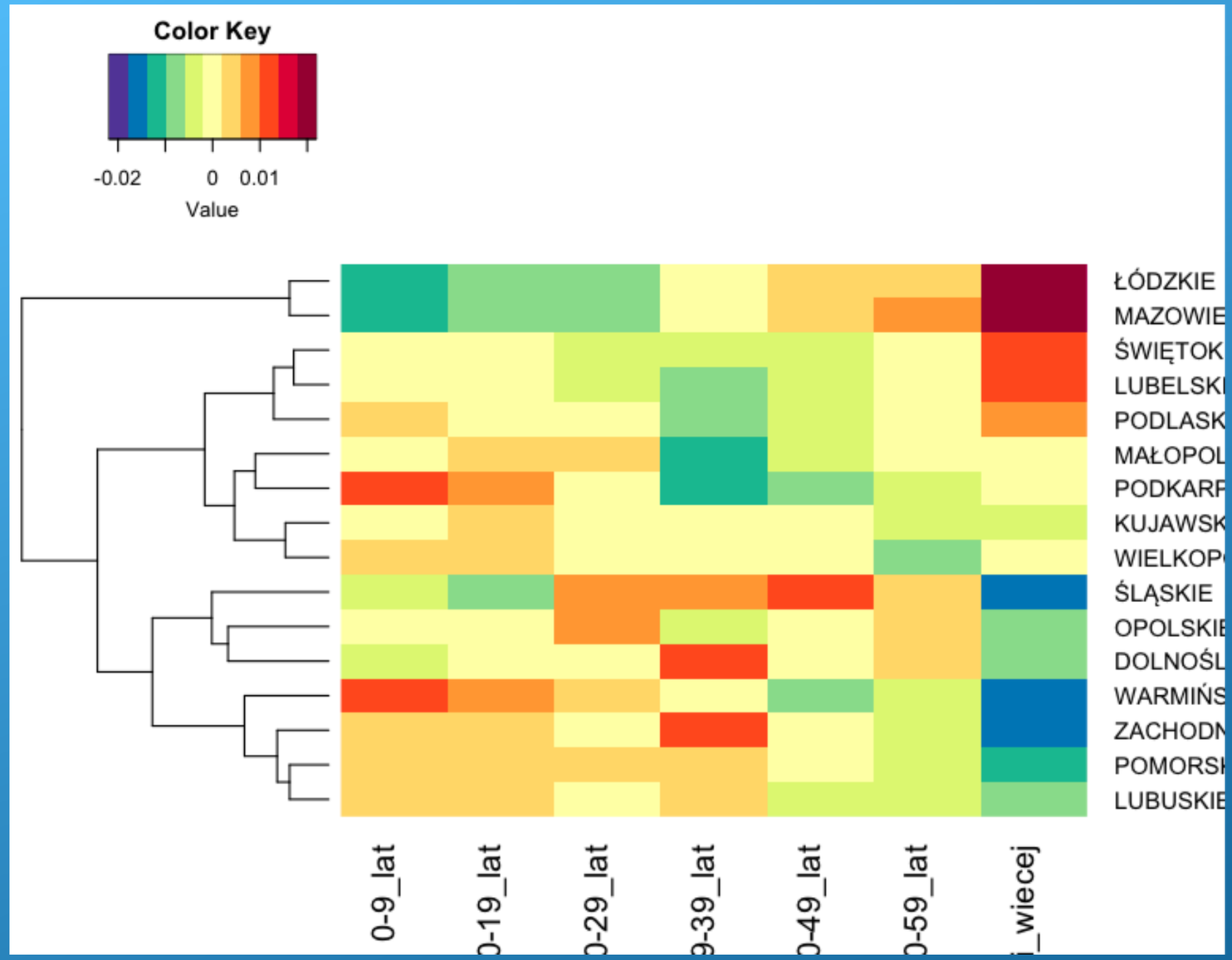
H0: There are no differences between provinces and the age of their inhabitants.

H1: There are differences between provinces and the age of their inhabitants.

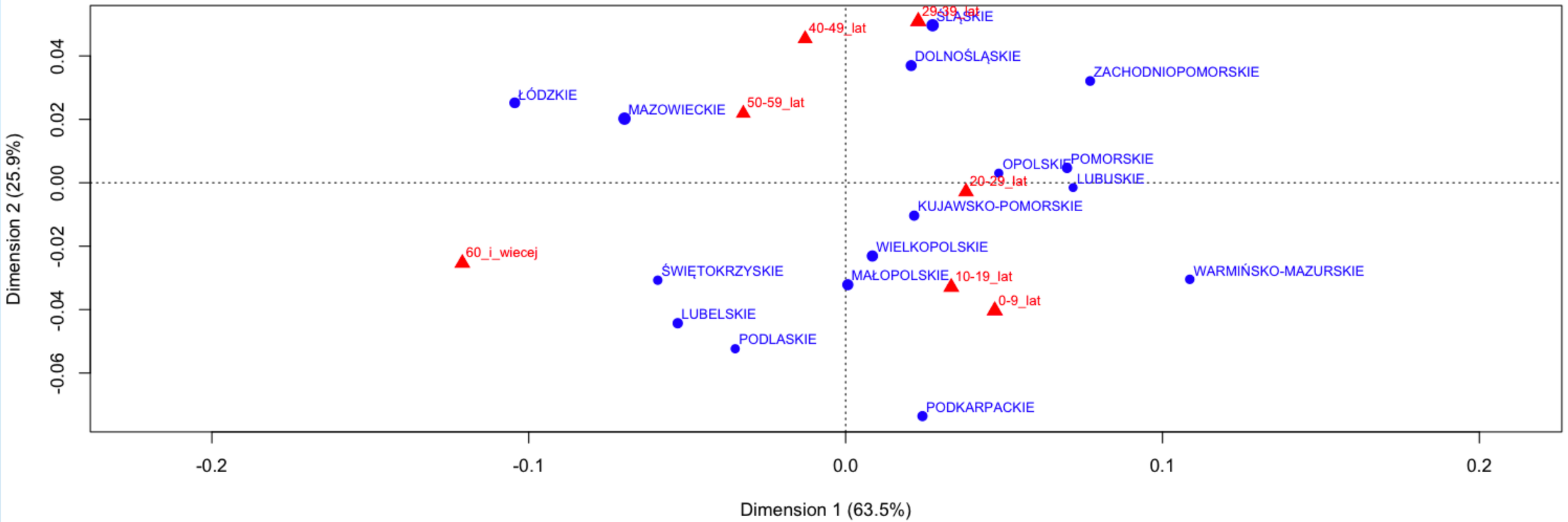
P-value is small (less than 0.05), we have grounds to reject H0 and accept H1 - there are differences between provinces and the age of their inhabitants.

```
        Pearson's Chi-squared test

data:  data2
X-squared = 184973, df = 90, p-value < 2.2e-16
```

Residua

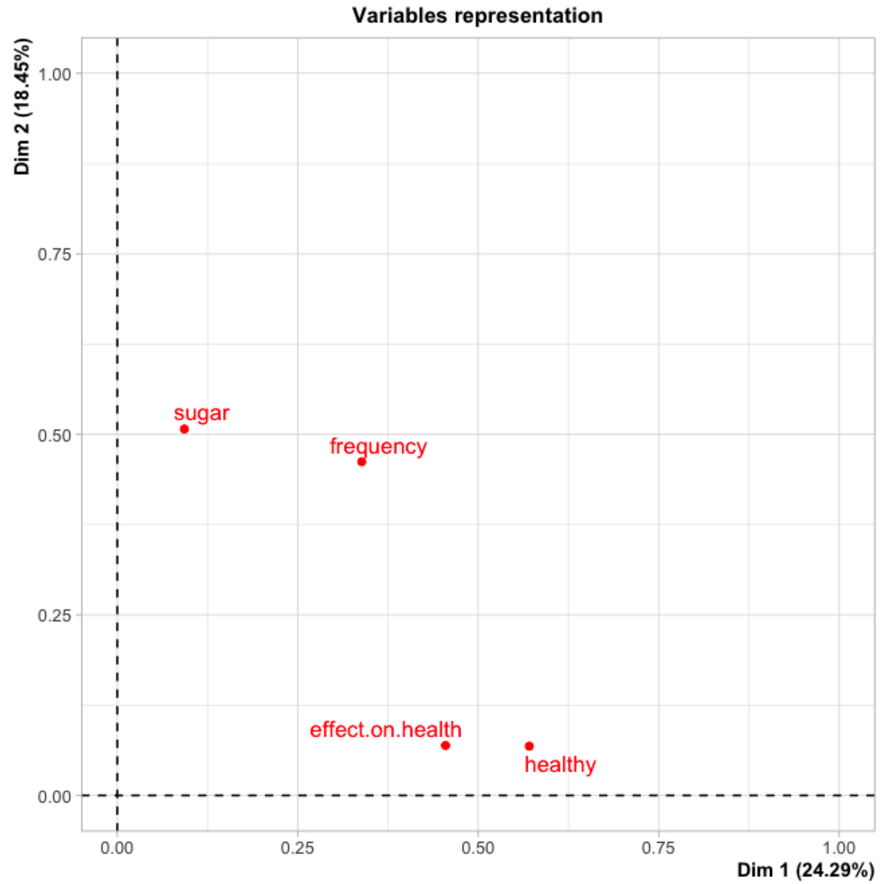


The largest number of residents over 60 live in the Łódzkie and Mazowieckie Voivodeships, and the smallest in the Śląskie, Warmińsko-Mazurskie and Zachodnio-Pomorskie Voivodeships. Of these visible differences, there is also a positive difference in the 30-39 age bracket for the Lower Silesian and West Pomeranian Voivodeships. From these results, we can also conclude that the greatest number of children were born in recent years in the voivodeships of Podkarpackie and Warmińsko-Mazurskie.



The most fearful people are in the w. Świętokrzyskie, Lubelskie, Podlaskie, Łódzkie and Mazowieckie. The Podkarpackie, Małopolskie, Wielkopolskie and Wamińsko-Mazurskie voivodeships have a very high number of people aged 0-29. While Dolnośląskie, Śląskie and Zachodniopomorskie have the highest number of people aged 29-59. The Opolskie, Pomorskie and Lubuskie Voivodeships have the largest number of people in the 20-29 age bracket. The West Pomeranian Voivodeship has the largest number of people in the 20-39 age bracket.

Correspondence analysis - multivariate




```
p-Value for sugar and effect.on.health 0.911171551769645
p-Value for sugar and healthy 0.207459453036099
p-Value for sugar and frequency 0.0284640048014936
p-Value for sugar and sugar 2.43957289266247e-66
p-Value for frequency and effect.on.health 0.143846456569731
p-Value for frequency and healthy 0.00782431696311168
p-Value for frequency and frequency 6.18680103239438e-188
p-Value for frequency and sugar 0.0284640048014936
p-Value for healthy and effect.on.health 7.31538284074612e-08
p-Value for healthy and healthy 3.55746141205215e-66
p-Value for healthy and frequency 0.00782431696311168
p-Value for healthy and sugar 0.207459453036099
p-Value for effect.on.health and effect.on.health 6.04315296663721e-66
p-Value for effect.on.health and healthy 7.31538284074612e-08
p-Value for effect.on.health and frequency 0.143846456569731
p-Value for effect.on.health and sugar 0.911171551769645
```

Tea

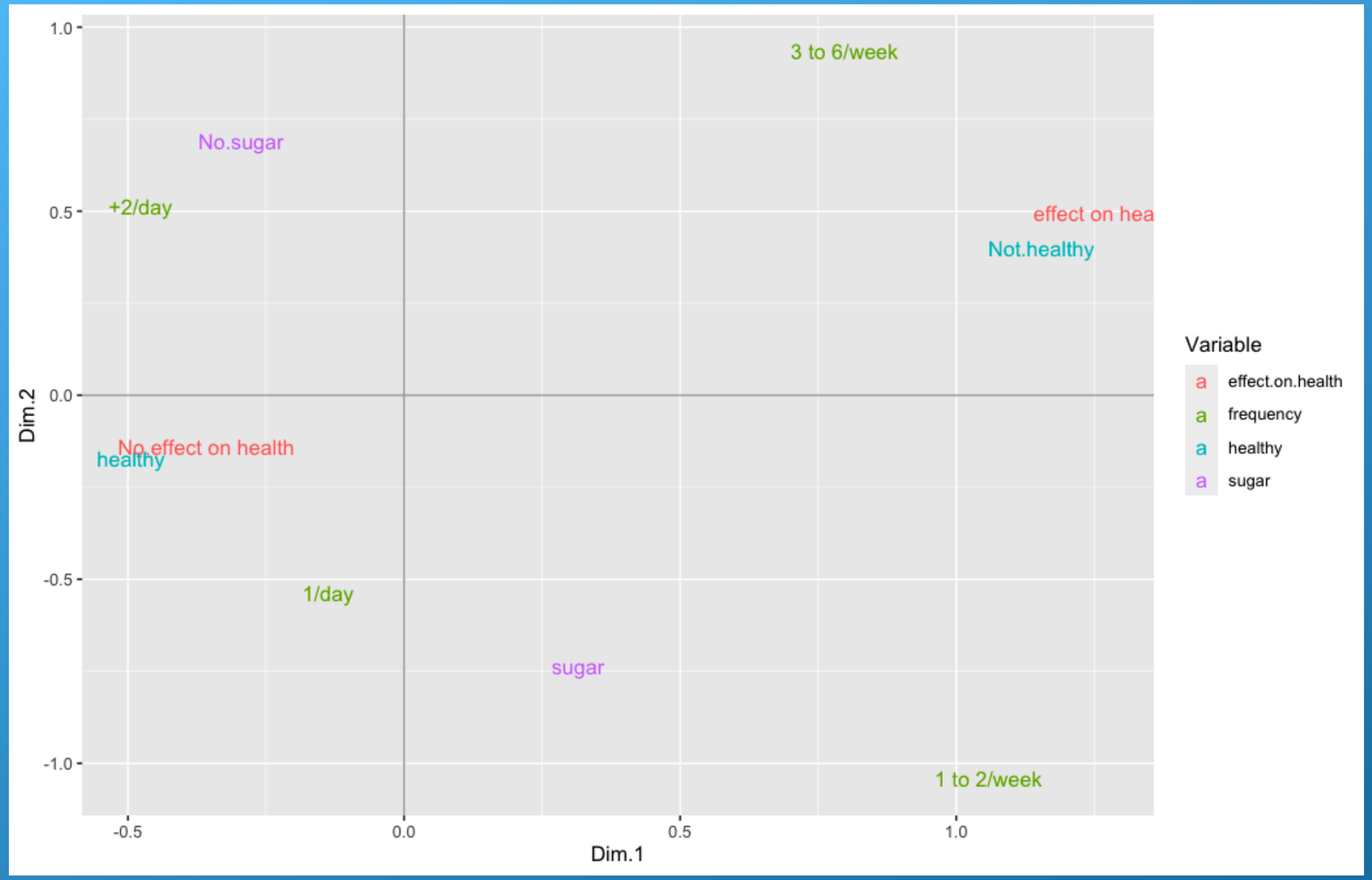
For the analyses I used the columns sugar, frequency, healthy, effect.on.health. These tell whether a person sweetens tea, how often they drink it, whether they are healthy, and whether tea has an effect on health.

H0: There is no difference in the frequencies of the results between variable a and variable b

H1: There are differences in the frequencies of the results for variable a and variable b

The Chi2 test performed indicates significant differences between the groups marked in red on the Chi2 test results.

The graph on the left allows us to identify which variables are most correlated with each dimension. The squares of the correlation between variables and dimensions are used as coordinates. Effect.on.health and healthy are most correlated with dimension 1 when sugar and frequency are most correlated with dimension 2.



From the results obtained, we can see that sick people reach for tea to improve their health and drink it more often than healthy people who do not expect tea to improve their health. In addition, as the frequency of tea drinking increased, the respondents stopped sweetening the tea.

Multidimensional scaling

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2
Grandson	71.7	34.0	17	8	3.30	20.0
Lausanne	55.7	19.4	26	28	12.11	20.2
La Vallee	54.3	15.2	31	20	2.15	10.8
Lavaux	65.1	73.0	19	9	2.84	20.0
Morges	65.5	59.8	22	10	5.23	18.0
Moudon	65.0	55.1	14	3	4.52	22.4

Data Swiss

Swiss data contain information on fertility and socio-economic status for the 47 provinces where French is the main language in Switzerland.

The mtcars data contain information on the model of cars and their engine parameters.

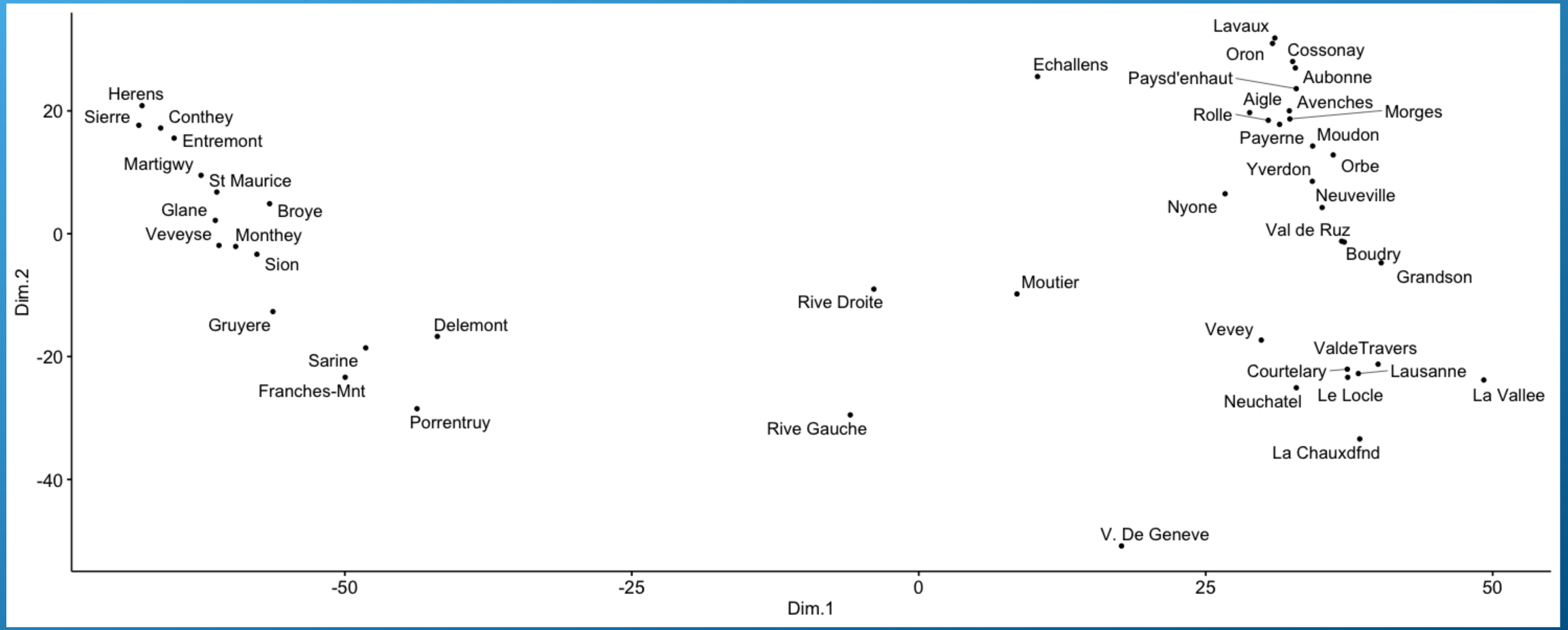
Multidimensional scaling analyses will be performed on these data.

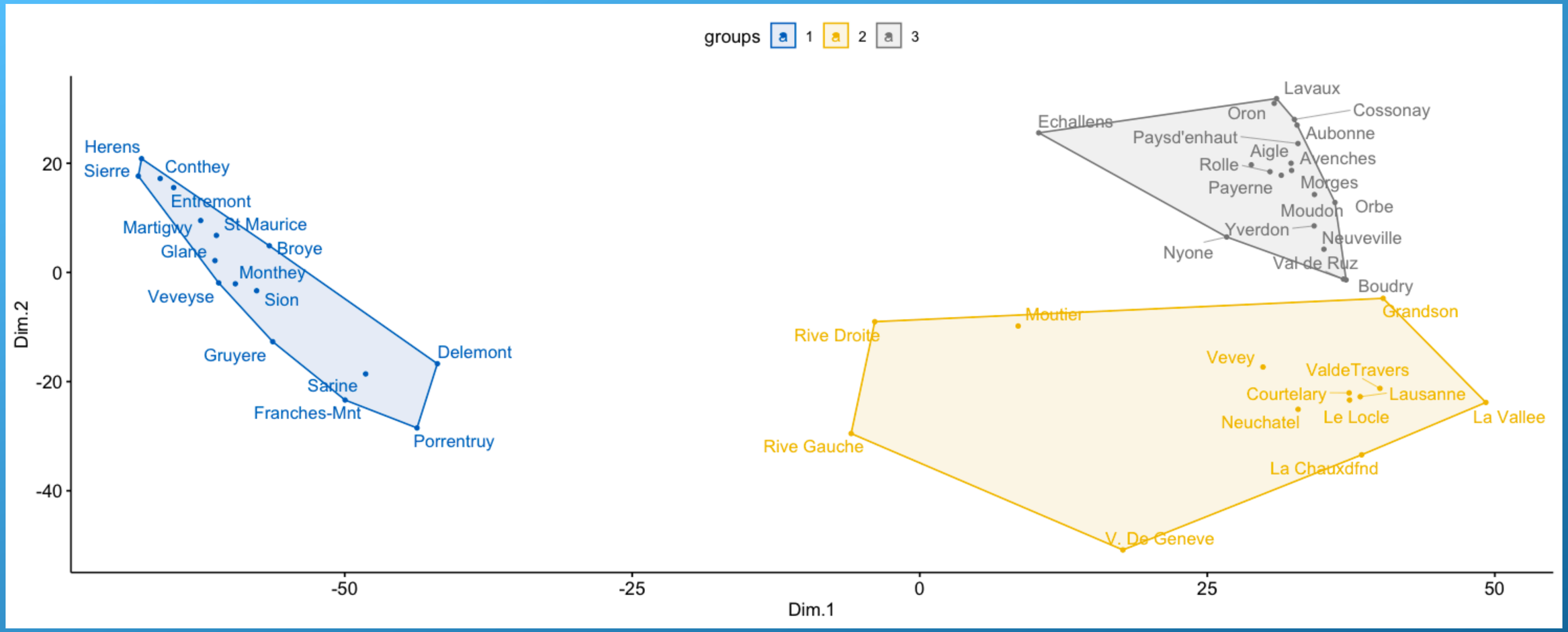
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2

Data mtcars

These are the results we obtained for Multidimensional Metric Scaling for selected columns 1,2,5, and 6 (all quantitative data). You can see that the graph shows 2-3 close groups that get similar results with few outliers. On the next slide I will perform k-means clustering for these results.

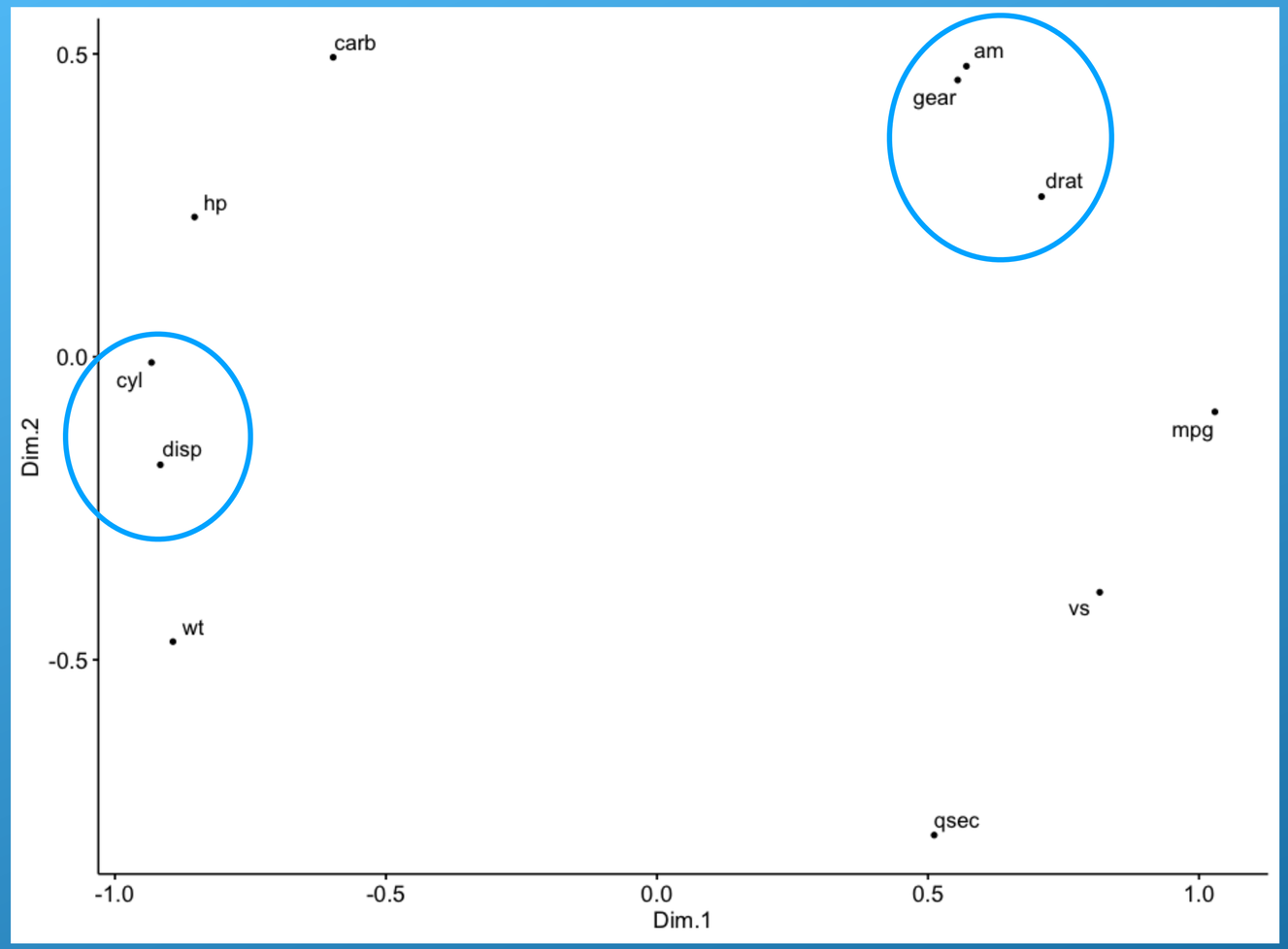
Swiss - metryczny





Using the k-mean clustering method, it was possible to create 3 visible groups, which are colour-coded. The groups were created due to the similarity in the results obtained in our data. As you can see, the group marked in blue is the most cohesive, black a little less so, yellow is the most extensive and there are the greatest differences in results. Perhaps if we changed the number of groups we would like to obtain it would be this larger number of groups that would influence the breakdown of this group into smaller subgroups with more similar (similar) values.

mtcars- non-metric



Positively correlated (similar) objects are close together on the same side of the graph. Am, gear and drat will be strongly correlated with each other. Cyl and disp are also close to each other. I have circled both groups on the presentation. The rest of the objects are quite far away to be able to see any relationship between them now.

Summary

Summary

Main objectives of the analyses

- The main objective of MDS (Multidimensional Scaling) is to **reduce the multidimensionality** of the data while **retaining the most information about distances between objects** - that is, it seeks a representation of objects that best reflects the structure of distances between objects.
- AC (Correspondence Analysis) aims to **analyse the relationship between two categorical variables**. It attempts to **find patterns and relationships between categorical variables** in order to detect the **hidden structure of the data**.

Type of data

- MDS: Can be applied to **numeric data or measures of distance between objects**. Examples include distance matrices, similarity matrices, use of Euclidean metrics, etc.
- CA: It is used for **categorical data or data that can be transformed into a categorical form**. The input for CA is a **contingency table** that contains the **number of observations for different combinations of categories**.

Result of the analyses

- MDS: The result of MDS is a **transformed representation of the data in a lower dimensionality** that best preserves the distances between objects. The result can be, for example, a two-dimensional map that represents the relationships between objects.
- CA: The result of CA is **two correspondence maps** that **represent relationships between categories of variables**. A correspondence map can show which categories of variables are related to each other and which combinations of categories are most frequent.

Summary

- In summary, **MDS** is a **multidimensional scaling technique that reduces the dimensionality of numerical data**, while **AC** is a **technique for analysing relationships between categorical variables**. Both approaches are used in different contexts and have different data analysis objectives.

Źródła

- https://cran.r-project.org/doc/contrib/Biecek-R-basics.pdf
- http://pbiecek.github.io/Przewodnik/Analiza/beznadzoru/mds_metric.html
- https://rpubs.com/gaston/MCA
- http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/
- https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstcoran.html
- https://www.ibm.com/docs/pl/spss-statistics/27.0.0?topic=categories-correspondence-analysis
- https://predictivesolutions.pl/tagi,analiza-korespondencji
- https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstmulsca.html
- https://www.lukaszderylo.pl/blog/skalowanie-wielowymiarowe.html
- https://www.ibm.com/docs/pl/spss-statistics/29.0.0?topic=features-multidimensional-scaling
- http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/122-multidimensional-scaling-essentials-algorithms-and-r-code/
- https://pbiecek.github.io/NaPrzelajDataMiningR/part-6.html
- http://www.biecek.pl/NaPrzelajPrzezDataMining/NaPrzelajPrzezDataMining.pdf

Thank you for your attention

Daria Plewa