

Analiza korespondencji i Skalowanie Wielowymiarowe

Daria Plewa

Plan Prezentacji

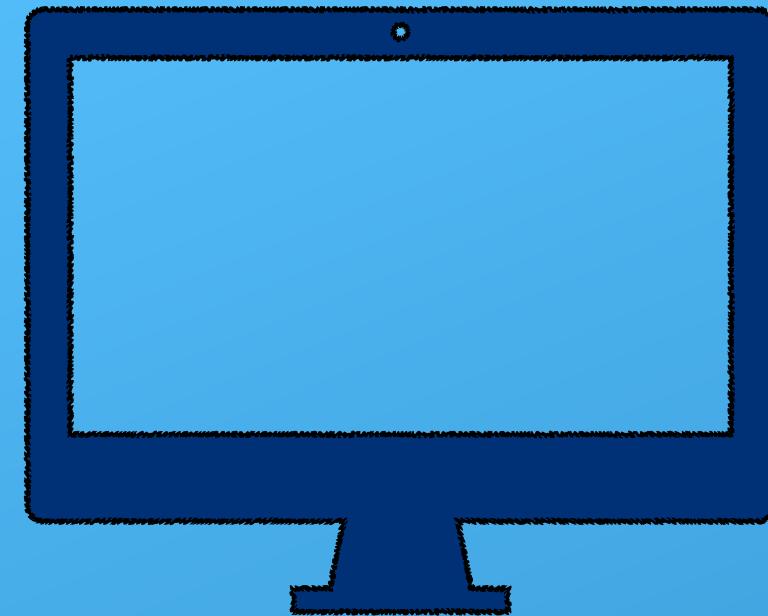


- 1. Opis Metod
- 2. Dane
- 3. Analiza Korespondencji
- 4. Analiza Korespondencji
Wieloczynnikowa
- 5. Skalowanie
Wielowymiarowe
- 6. Wnioski
- 7. Źródła

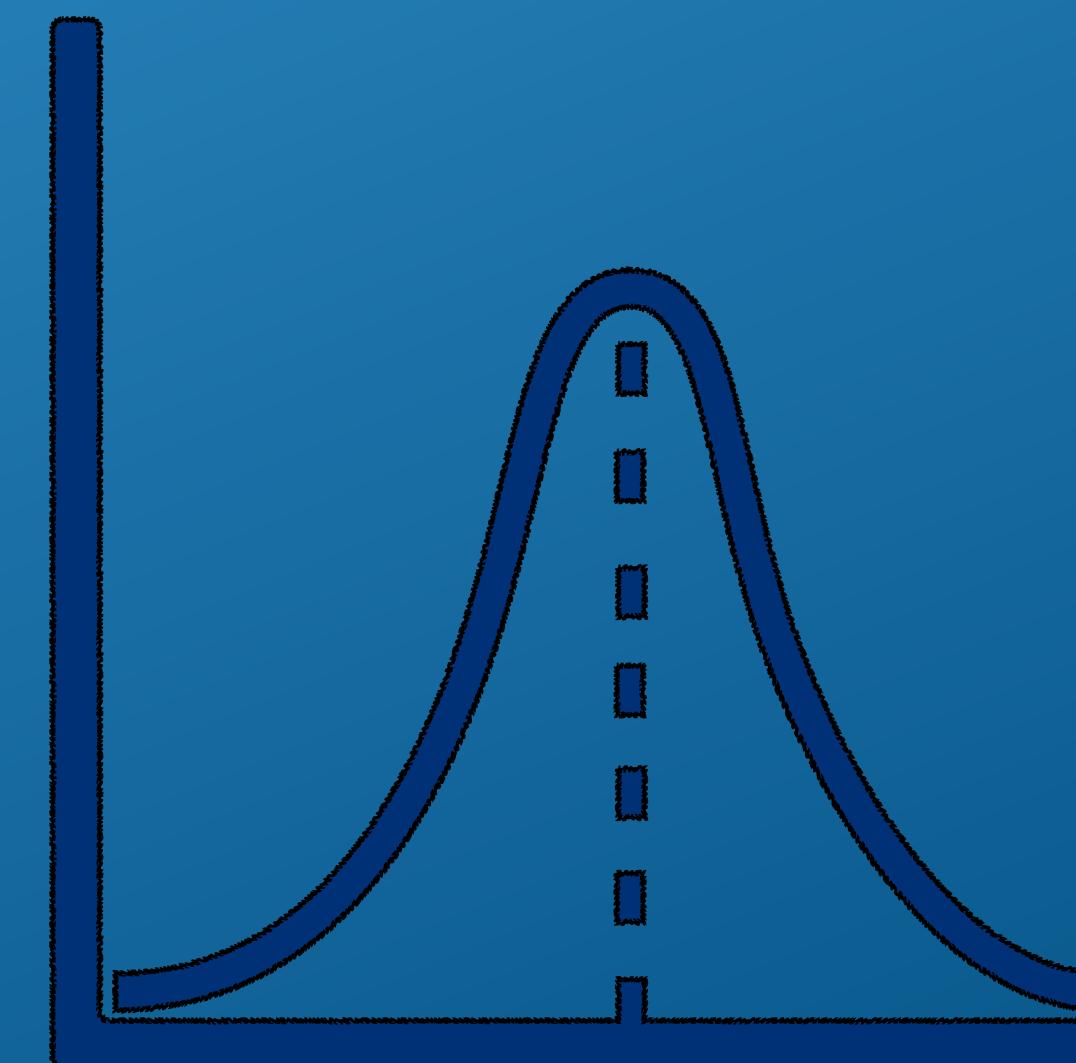


Opis metod

Analiza Korespondencji



Analiza korespondencji jest to opisowa i eksploracyjna technika analizy zmiennych jakościowych, która ma na celu m.in. prezentację w postaci wizualnej relacji z tabeli krzyżowej. Analiza korespondencji może być szczególnie przydatna w przypadku rozbudowanych tabel kontyngencji w celu zobrazowania relacji między kategoriami zmiennych. Jest często wykorzystywana w badaniach marketingowych, społecznych ale również ekonomicznych, w których to dominują zmienne jakościowe.



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

p_{ij} - prawdopodobieństwo zaobserwowania pierwszej zmiennej na poziomie i i jednocześnie drugiej na poziomie j

p_i - prawdopodobieństwo zaobserwowania zmiennej pierwszej na poziomie i

p_j - prawdopodobieństwo zaobserwowania zmiennej drugiej na poziomie j

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

Aby ocenić, które zmienne występują częściej lub rzadziej niż wynikało by to z niezależności, wyznaczymy standaryzowane reszty Pearsonowskie, zastąpimy też prawdopodobieństwa ich częstościowymi ocenami.

\hat{p} - liczba obserwowanych zdarzeń podzielona przez liczbę wszystkich obserwowanych zdarzeń

Duże dodatnie wartości \hat{e}_{ij} odpowiadają wysokiemu współwystępowaniu.

$$E_{k \times l} = U_{k \times k} \sum_{k \times k} V_{l \times l}^T$$

Następnie macierz $E = \hat{e}_{ij}$ przedstawię w postaci graficznej używając tzw. Bibliotu. Innymi słowy wyznaczamy dekompozycję SVD macierzy E.

Kolumny macierzy $U_{k \times k}$ to wektory własne macierzy $E^T E$, a kolumny macierzy V to wektory własne macierzy EE^T . Na przekątnej macierzy diagonalnej σ znajdują się tzw. wartości singularne równe pierwiastkom z wartości własnych macierzy $E^T E$ i EE^T .

Skalowanie Wielowymiarowe

Skalowanie wielowymiarowe służy do znajdowania struktury w zbiorze miar odległości między poszczególnymi obiektami lub obserwacjami. Jest to możliwe dzięki przypisywaniu obserwacji do poszczególnych miejsc w przestrzeni pojęciowej (zwykle dwu- lub trójwymiarowej) w taki sposób, że odległości między punktami w przestrzeni możliwie blisko odpowiadają danym miarom niepodobieństwa. W wielu przypadkach wymiary tej przestrzeni pojęciowej mogą być interpretowane i wykorzystywane do lepszego zrozumienia danych.

Skalowanie Wielowymiarowe Metryczne i Niemetryczne

Skalowanie metryczne wielowymiarowe (ang. metric multidimensional scaling, MDS) również opiera się na macierzy podobieństwa lub odległości między obiektyami. Jednak w przypadku MDS zakłada się, że odległości w przestrzeni pierwotnej są zgodne z odległościami w przestrzeni niewymiarowej. Celem MDS jest znalezienie takiej reprezentacji danych, w której odległości między punktami w przestrzeni niewymiarowej są jak najbardziej zgodne z odległościami między obiektyami w przestrzeni pierwotnej. Skalowanie metryczne wykorzystuje różne metody optymalizacyjne, takie jak metoda gradientowa lub analiza wartości własnych.

Skalowanie niemetryczne wielowymiarowe (ang. nonmetric multidimensional scaling, NMDS) opiera się na macierzy podobieństwa lub odległości między obiektyami. Celem NMDS jest znalezienie takiej reprezentacji danych, w której odległości między punktami w przestrzeni niewymiarowej są jak najbardziej zgodne z pierwotnymi odległościami między obiektyami. Skalowanie to jest "niemetryczne", ponieważ nie zakłada żadnego konkretnego związku między odległościami w przestrzeni pierwotnej a odległościami w przestrzeni niewymiarowej. NMDS jest oparte na metodzie gradientowej i może być stosowane zarówno do danych metrycznych, jak i niemetrycznych.

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

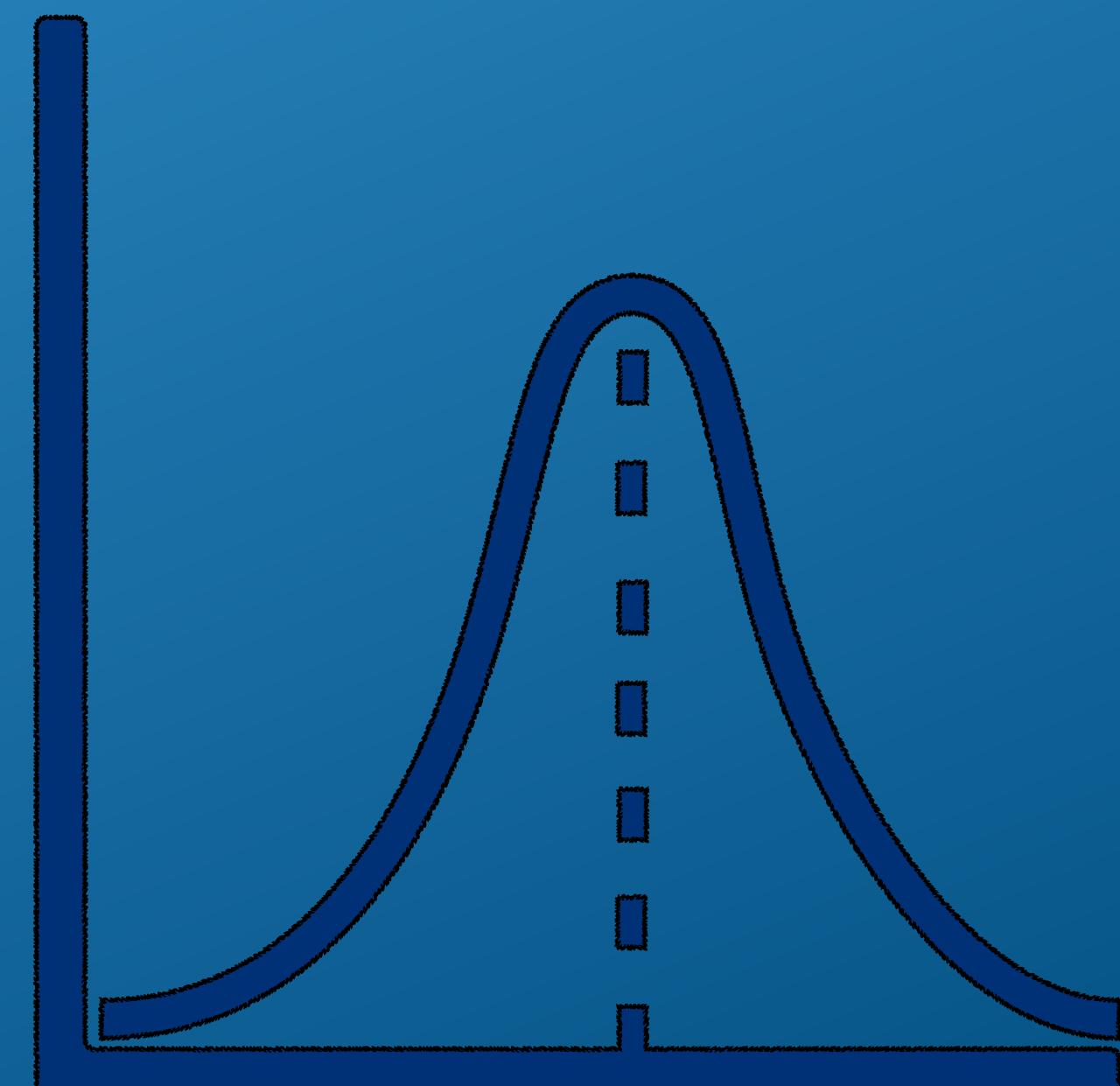
Efekty tych skalowań mogą różnić się w zależności od charakterystyki danych i użytych metryk. Obydwa skalowania mają na celu zachowanie relatywnych odległości między obiektami, ale różnią się pod względem interpretowalności wyników. Skalowanie niemetryczne może generować reprezentacje, które nie odzwierciedlają dokładnie odległości w oryginalnych danych, ale zachowują ich porządek względny. Z drugiej strony, skalowanie metryczne stara się odwzorować dokładne odległości, co może prowadzić do bardziej precyzyjnych reprezentacji, ale może być bardziej wrażliwe na błędy w danych

Po prawej znajduje się wzór stress'u (ang. Standardized Residual Sum of Squares). Minimalizuje on wartość standaryzowanej sumy kwadratów reszt.

\hat{d}_{ik} - odległość pomiędzy obiektami i i k w nowej x-miarowej przestrzeni

d_{ik} - to oryginalne odległości pomiędzy obiektami

Skalowanie metryczne wielowymiarowe głównie służy do analiz danych ilościowych gdy niemetryczne do danych jakościowych.



Opis danych

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, płeci ⓘ ⓘ ⓘ
Wymiary Typy szkół ⓘ; Grupy osób; Płeć; Lata
Ostatnia aktualizacja 04.09.2019

Tablica Wykres Mapa

Wybór jednostek terytorialnych Agregaty Kod Puste Export Objąść

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>

Dane dla większości Analiz Korespondencji zostały pobrane z GUS'u (Głównego Urzędu Statystycznego). Pobrałem stamtąd dane na temat liczby absolwentów uczelni wyższych, odstrzałów zwierząt, spisu powszechnego ludności względem różnych województw.

	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

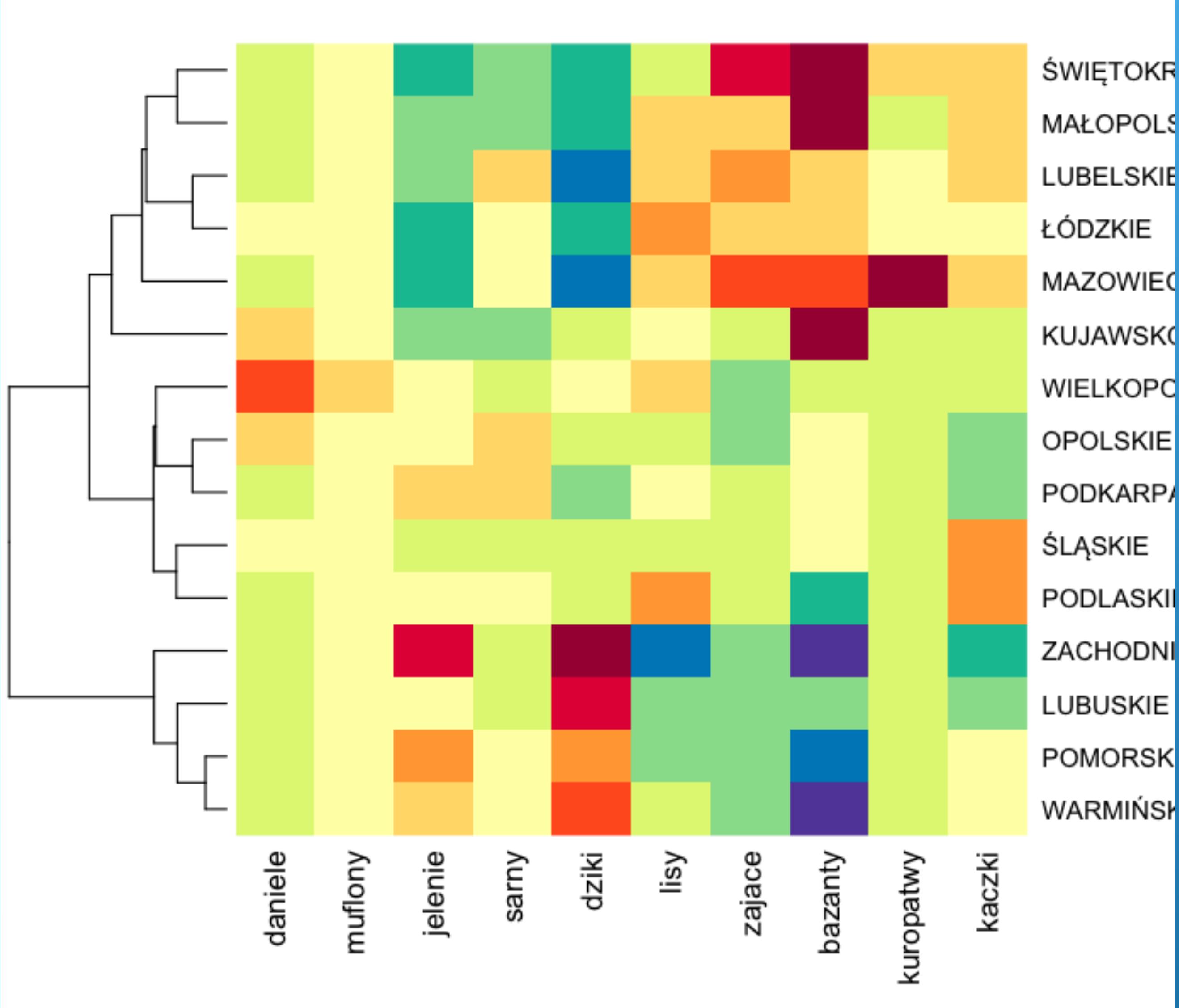
Dane tea

Dane tea zawierają 300 rekordów na temat preferencji picia herbaty. Wszystkie dane są wartościami jakościowymi. Zbiór zawiera 16 kolumn w których są opisane zwyczaje picia herbaty (obecność cukru, rodzaj herbaty, kiedy jest ona pita)

Na tych danych zostaną wykonane Analizy Korespondencji Wieloczynnikowej

Analiza Korespondencji

Liczba odstrzałów dzikich zwierząt względem województw



Pearson's Chi-squared test

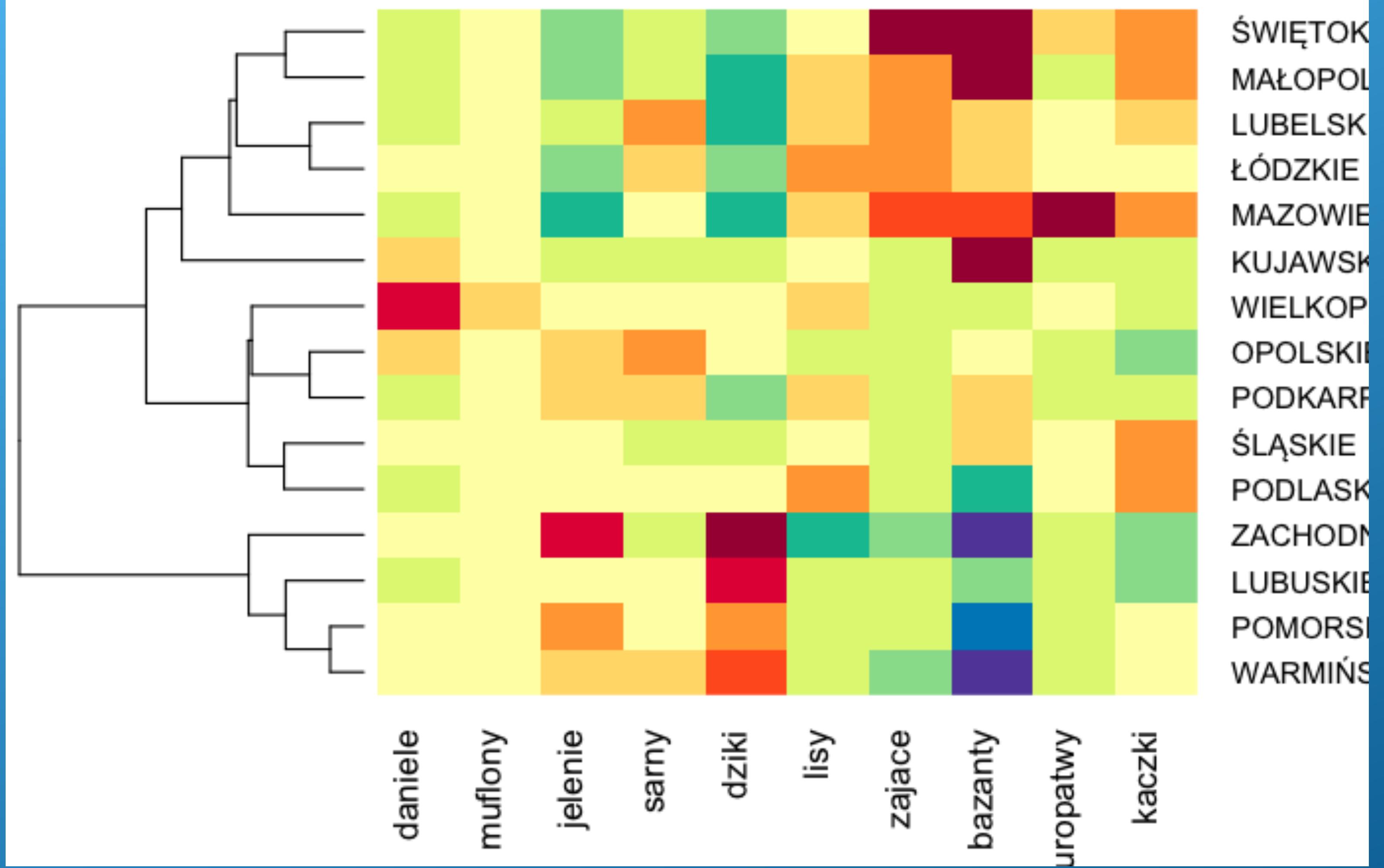
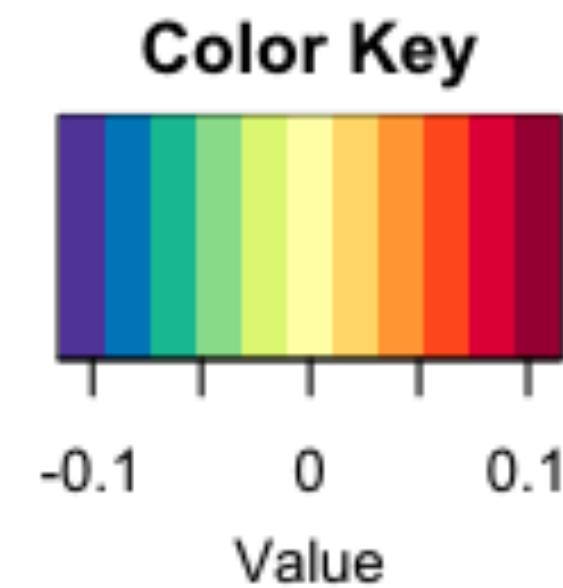
```
data: data2  
X-squared = 214748, df = 135, p-value < 2.2e-16
```

H0: Nie ma różnic pomiędzy województwami w liczebnościach odstrzałów zwierząt.

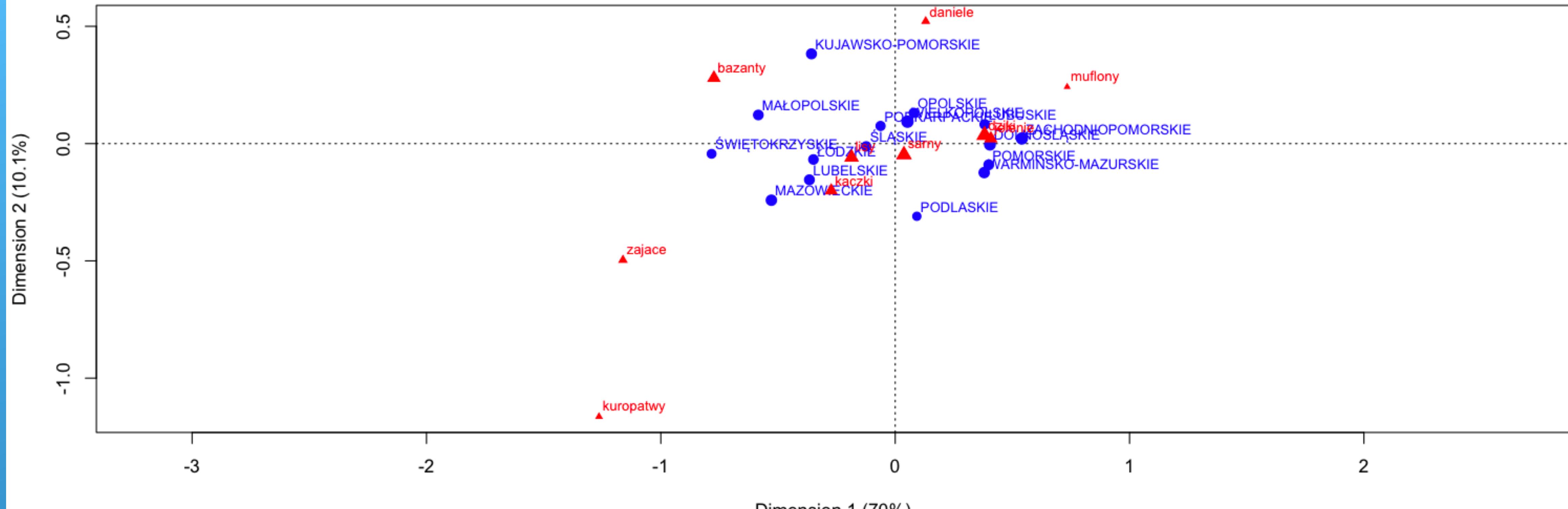
H1: Są różnice w liczebnościach odstrzałów zwierząt w województwach.

P-value jest małe (poniżej 0.05), mamy podstawy do odrzucenia H0 i przyjęcia H1 - są różnice w liczebnościach odstrzałów zwierząt w województwach.

Residua

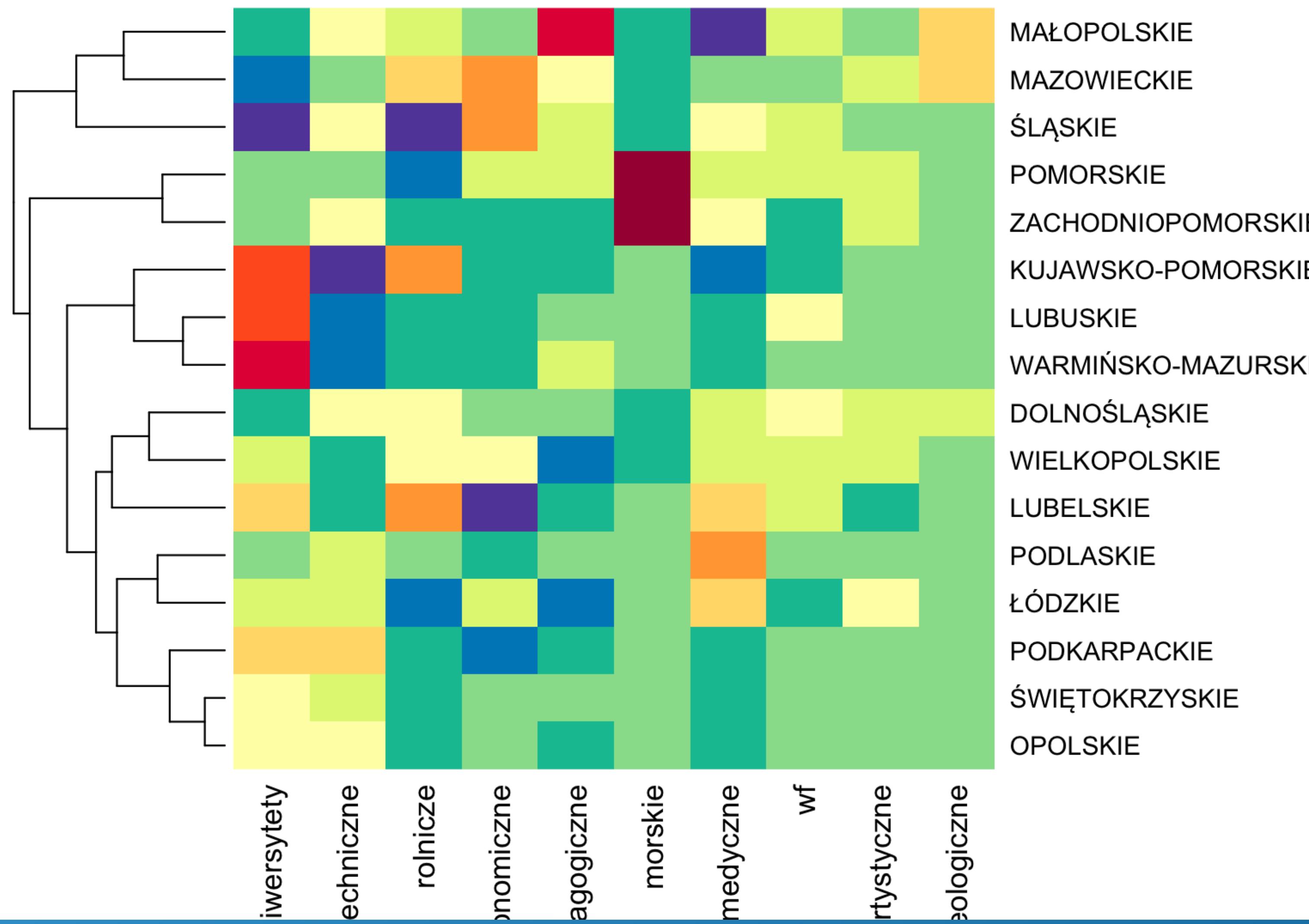


Najwięcej odstrzałów podlega bażantom aż w 3 województwach. Również popularne są łowy na kuropatwy w w. Mazowieckim, czy dzików w Zachodniopomorskim. Mniej popularne są odstrzały danieli w Wielkopolskim i jeleni w Zachodniopomorskim oraz dzików w Lubuskim i Warmińsko-Mazurskim. Za to jest bardzo mało połowów bądź żadnych dla bażantów w w. Zachodniopomorskim, Pomorskim i Warmińsko-Mazurskim.



We wszystkich województwach występują odstrzały saren, kaczek, dzików, lisów. Najwięcej danieli odstrzeliwuje się w Kujawsko-Pomorskim. Bażantów w Kujawsko-Pomorskim oraz Małopolskim. Muflony w Zachodnio-Pomorskim, Dolnośląskim i Wielkopolskim. Bażenty, danieli, muflony, zajęce oraz kuropatwy najbardziej różnicują województwa w ostrzałach, z czego najczęściej zajęcy i kuropatwy odstrzeliwuje się w Mazowieckim.

Liczba studentów różnych typów uczelni względem województw

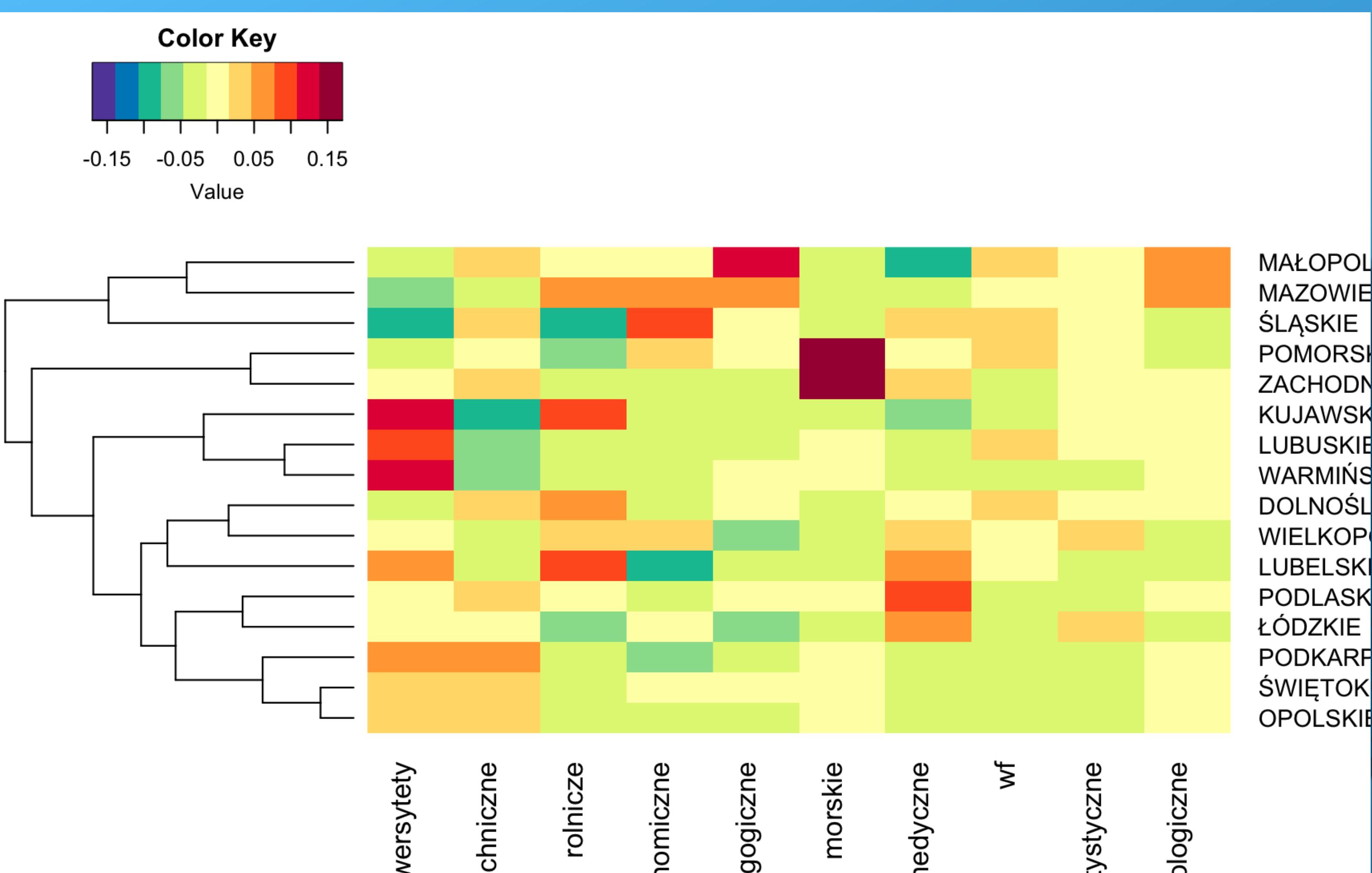


H0: Nie ma różnic pomiędzy województwami a liczbą studentów różnych typów uczelni.

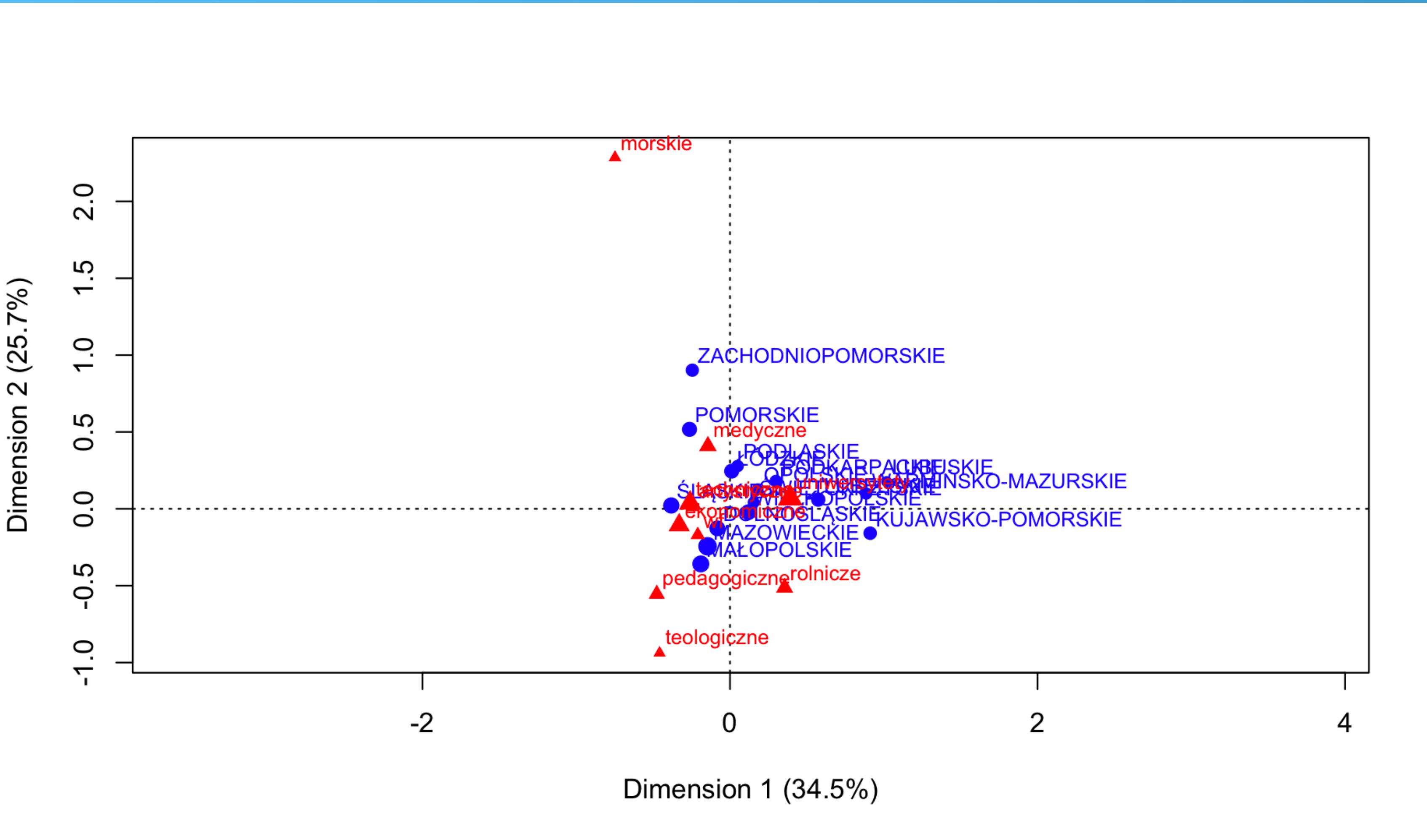
H1: Są różnice pomiędzy województwami a liczbą studentów na różnych typach uczelni.

P-value jest małe (poniżej 0.05), mamy podstawy do odrzucenia H0 i przyjęcia H1 - są różnice w liczebnościach studentów na różnych typach uczelni w różnych województwach.

Residua

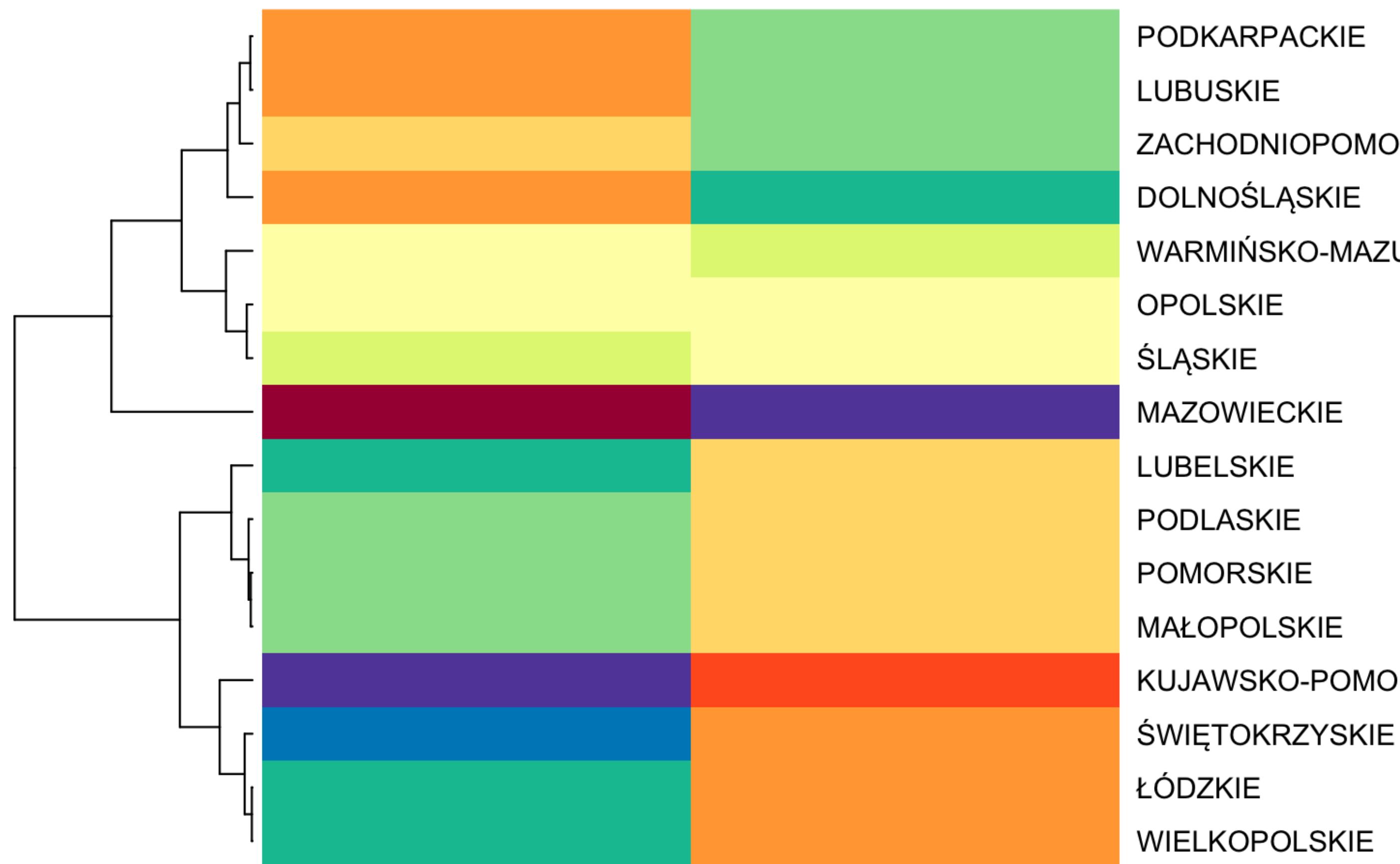


Najwięcej studentów uczelni morskich jest w województwie Pomorskim i Zachodniopomorskim. Najwięcej studentów dla uniwersytetów jest w województwie Kujawskim, Lubuskim oraz Warmińsko-mazurskim. Najwięcej studentów na uczelni pedagogicznej studiuje w województwie Małopolskim.



Najbardziej różnicującym uczelniami są uczelnie morskie, rolnicze, pedagogiczne oraz teologiczne. Na uczelniach morskich studiuje najwięcej studentów w Zachodniopomorskim oraz Pomorskim. Na innych rodzajach uczelni ciężko dostrzec znaczące różnice pomiędzy województwami.

Płeć względem województw wśród studentów



żni
ietym

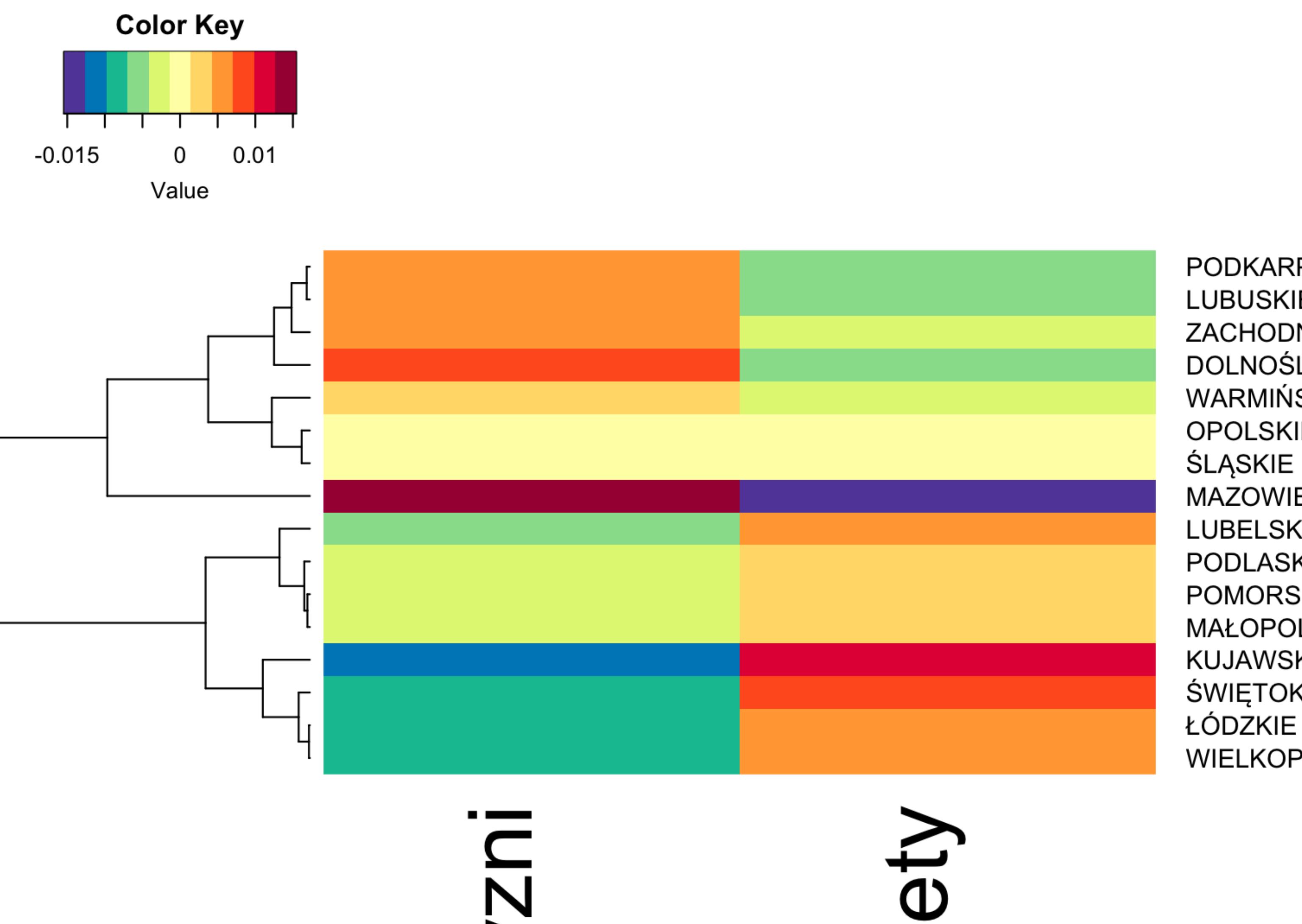
```
Pearson's Chi-squared test  
data: data2  
X-squared = 1691.2, df = 15, p-value < 2.2e-16
```

H0: Nie ma różnic pomiędzy województwami a płcią studentów.

H1: Są różnice pomiędzy województwami a płcią studentów.

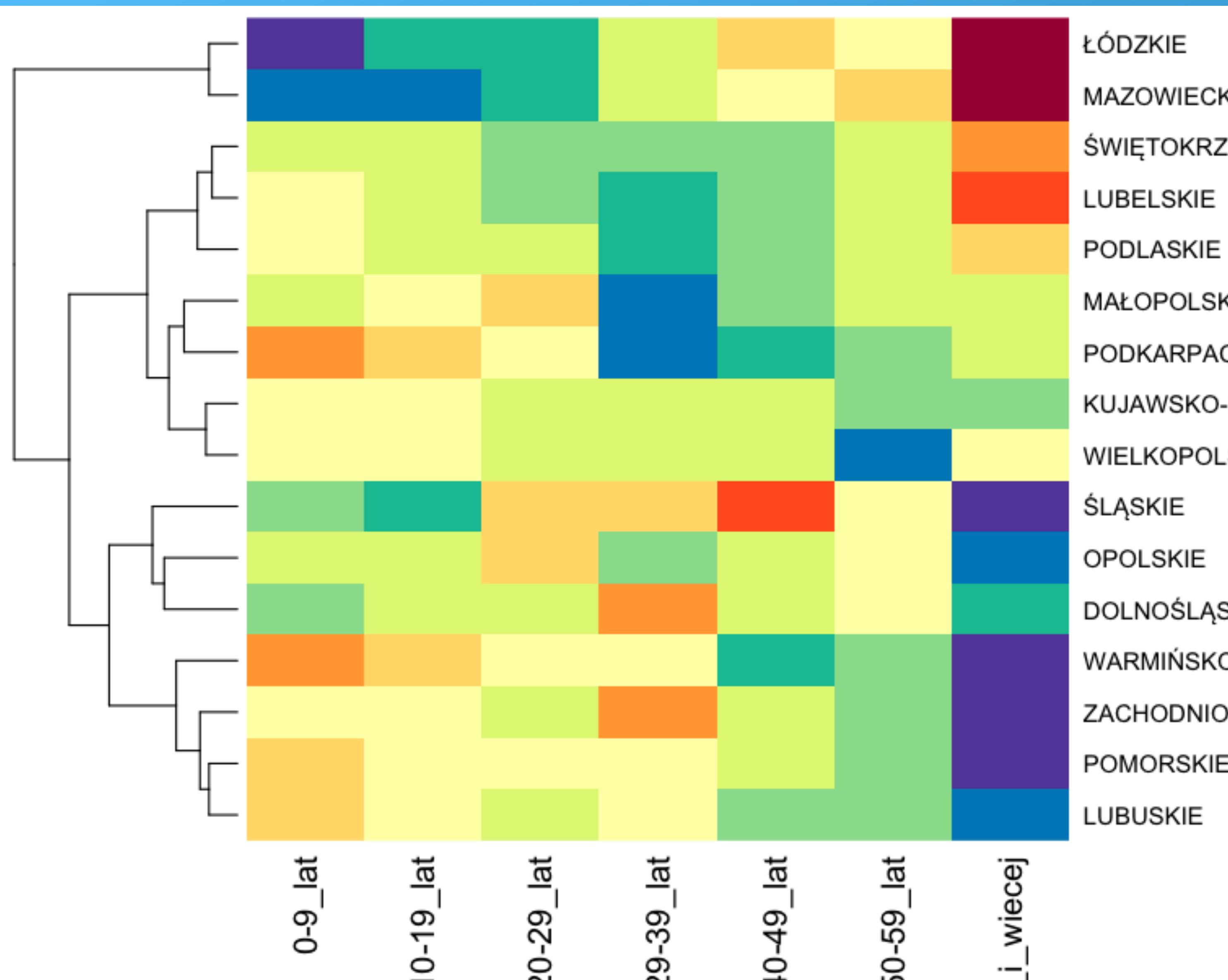
P-value jest małe (poniżej 0.05), mamy podstawy do odrzucenia H0 i przyjęcia H1 - są różnice w płciach studentów w różnych województwach. (Pomiędzy liczebnościami kobiet i mężczyzn)

Residua



Więcej mężczyzn niż kobiet studiuje w w. Podkarpackim, Lubuskim, Dolnośląskim, Mazowieckim. Więcej kobiet względem mężczyzn studiuje w w. Kujawskim, Świętokrzyskim, Łódzkim oraz Wielkopolskim i Lubelskim. Z czego największa dysproporcja w płciach występuje w w. Dolnośląskim, Kujawskim oraz Mazowieckim. A najbardziej zbliżone liczebności kobiet i mężczyzn występują w w. Opolskim i Śląskim.

Spis powszechny ludności względem województw



H₀: Nie ma różnic pomiędzy województwami a wiekiem ich mieszkańców.

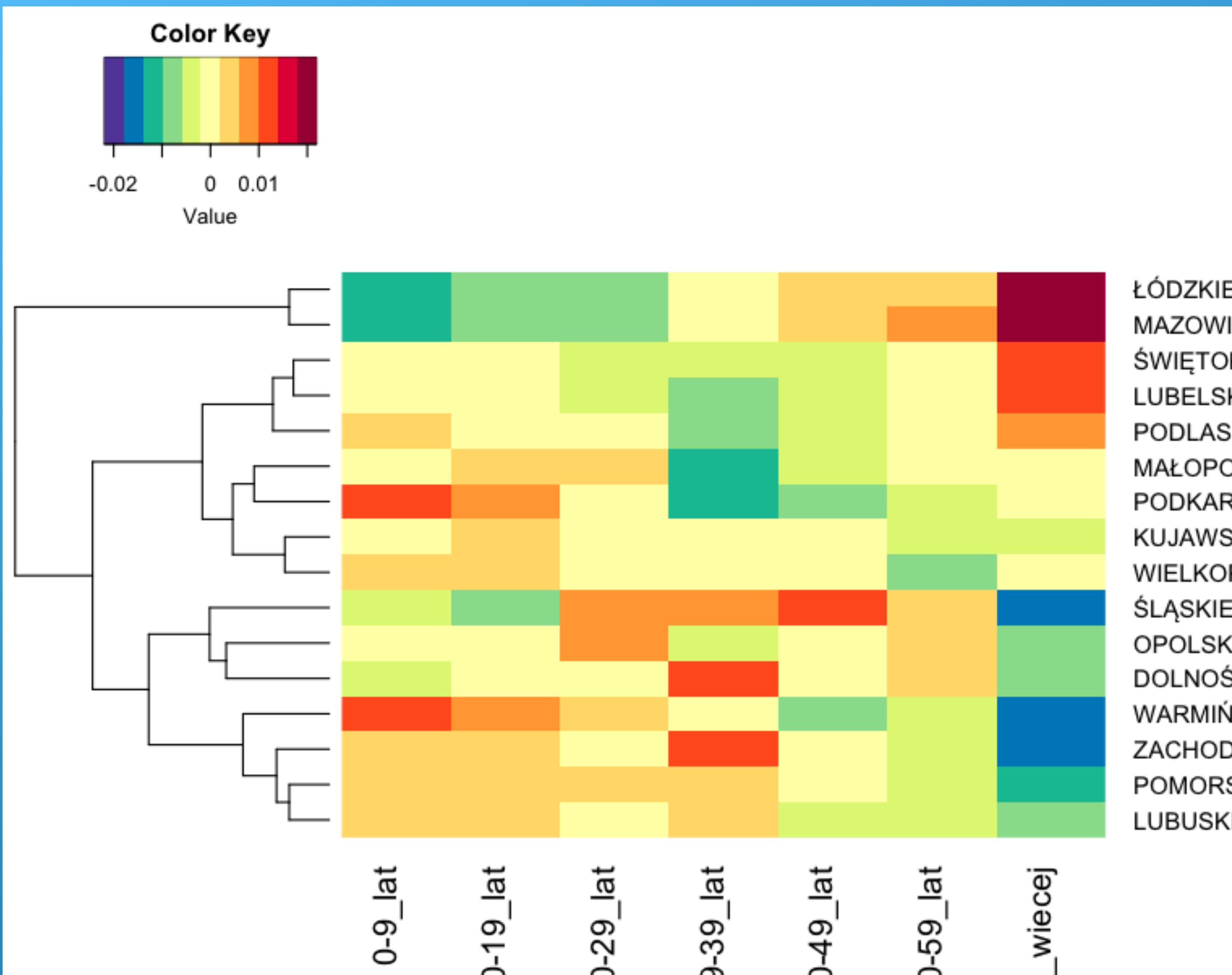
H₁: Są różnice pomiędzy województwami a wiekiem ich mieszkańców.

P-value jest małe (poniżej 0.05), mamy podstawy do odrzucenia H₀ i przyjęcia H₁ - są różnice pomiędzy województwami a wiekiem ich mieszkańców.

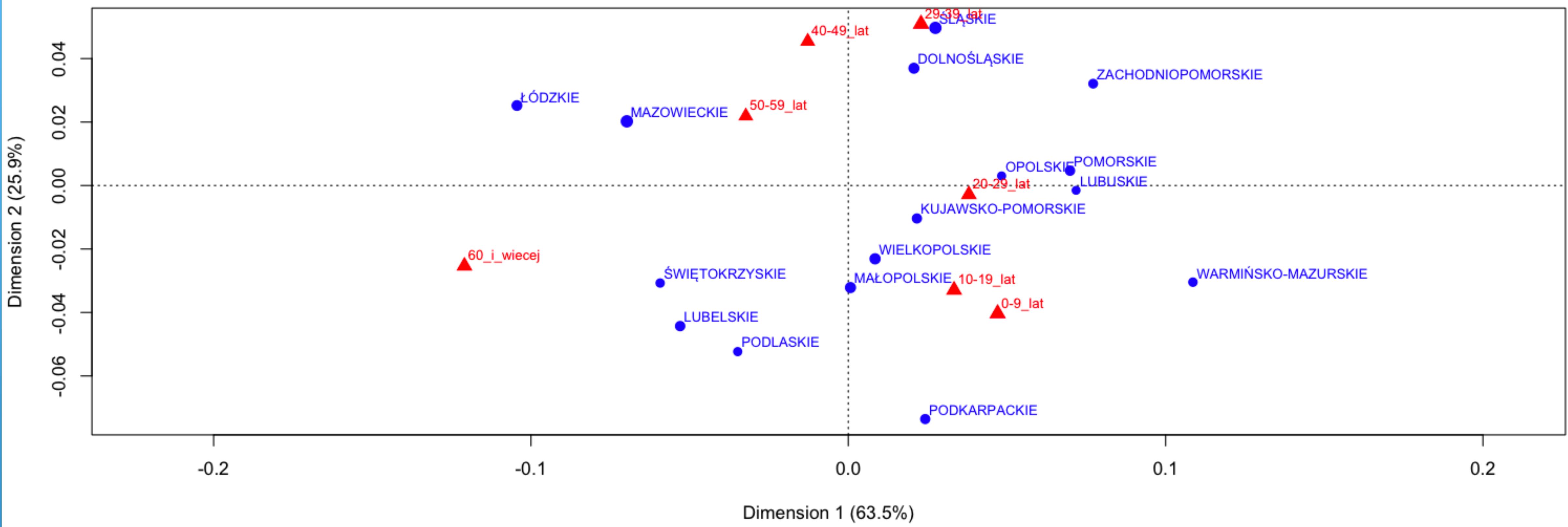
Pearson's Chi-squared test

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

Residua



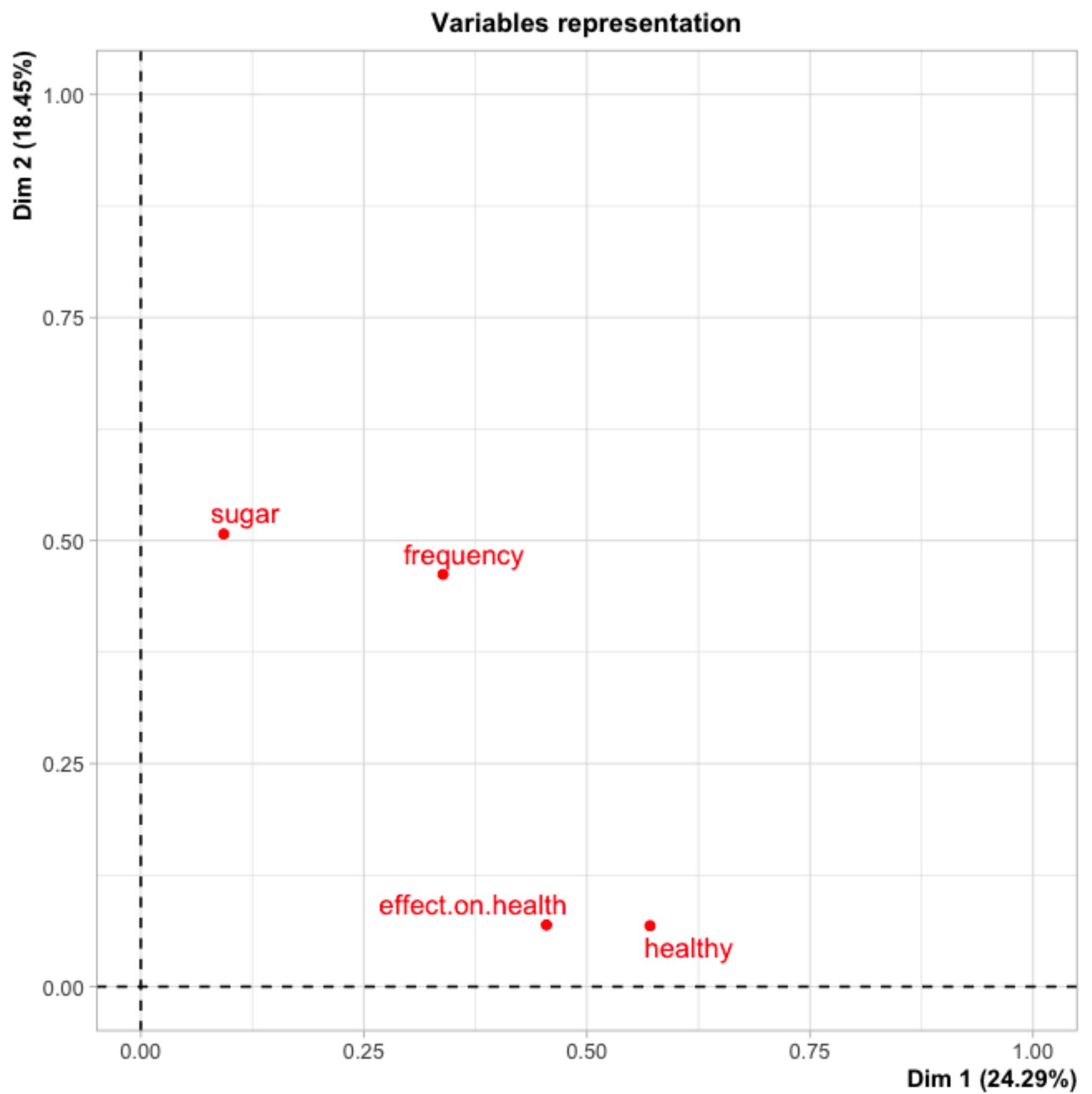
Najwięcej mieszkańców powyżej 60 roku życia mieszka w województwie Łódzkim i Mazowieckim, a najmniej w Śląskim, Warmińsko-Mazurskim oraz Zachodniopomorskim. Z takich widocznych różnic widać jeszcze dodatkową różnicę w przedziale 30-39 lat dla województw Dolnośląskiego oraz Zachodniopomorskiego. Po tych wynikach możemy też stwierdzić, że najwięcej dzieci rodziło się w ostatnich latach w województwie Podkarpackim oraz Warmińsko-Mazurskim.



Najstrasze osoby przebywają na terenie województwa Świętokrzyskiego, Lubelskiego, Podlaskiego, Łódzkiego oraz Mazowieckiego. W Podkarpackim, Małopolskim, Wielkopolskim oraz Warmińsko-Mazurskim znajduje się bardzo dużo osób w wieku 0-29 lat. Gdy w Dolnośląskim, Śląskim, Zachodniopomorskim znajduje się najwięcej osób w wieku 29-59 lat. W Opolskim, pomorskim oraz Lubuskim znajduje się najwięcej osób w przedziale wiekowym 20-29 lat. W zachodniopomorskim znajduje się najwięcej osób w przedziale 20 - 39 lat.

Analiza Korespondencji - wieloczynnikowa

Herbata



```
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
```

Do analiz wykorzystałam kolumny sugar, frequency, healthy, effect.on.health. Mówią one o tym czy osoba słodzi herbatę, jak często ją pije, czy jest zdrowa, czy herbata ma wpływ na zdrowie.

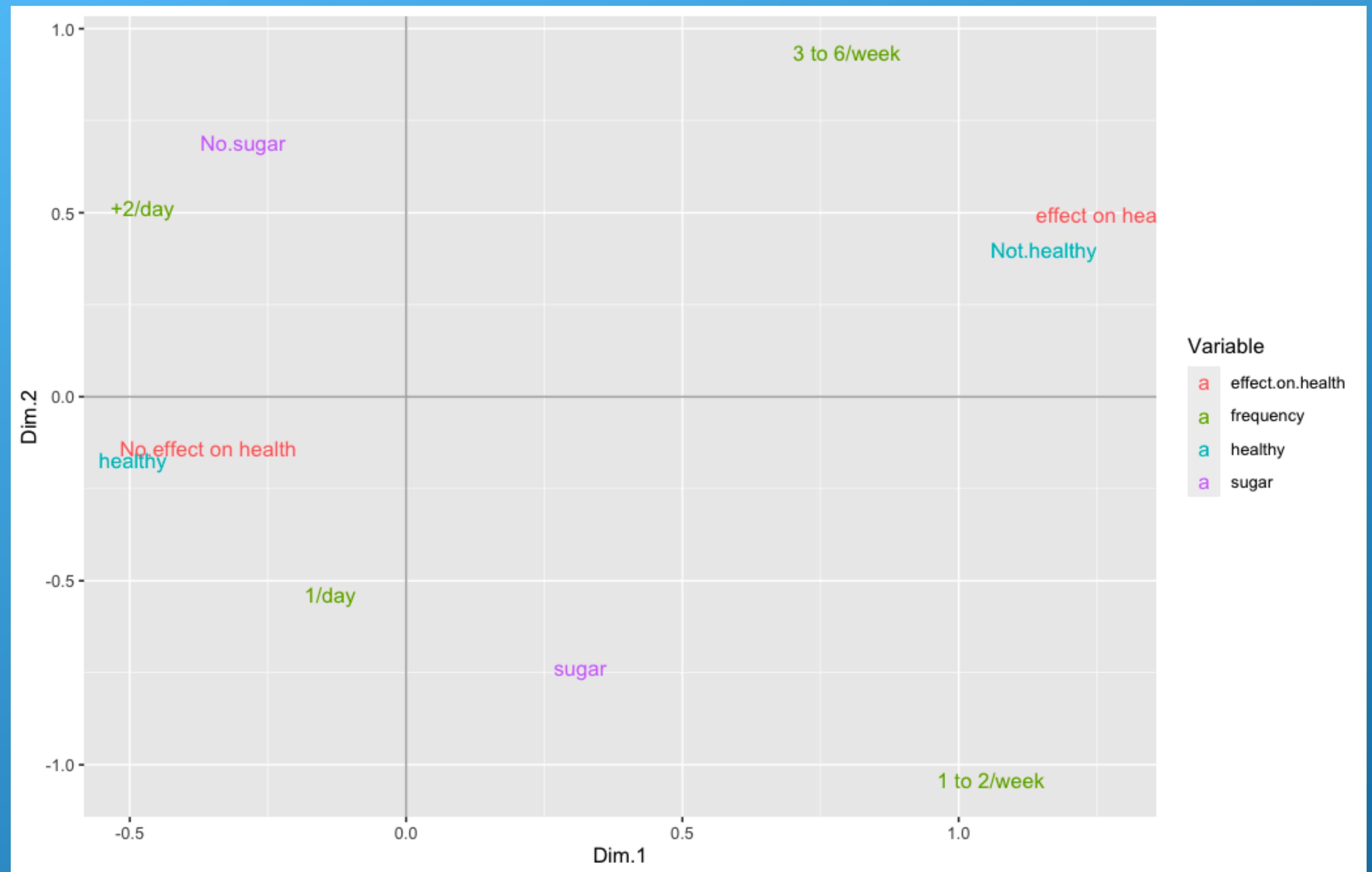
H0: Nie ma różnicy we frekwencjach wyników pomiędzy zmienną a zmienną b

H1: Są różnice we frekwencjach wyników dla zmiennej a i zmiennej b

Wykonany test Chi2 wskazuje na istotne różnice pomiędzy grupami zaznaczonymi na wynikach testu Chi2 na czerwono.

Wykres po lewej stronie pozwala na określenie które zmienne są najbardziej skorelowane z każdym wymiarem. Kwadraty korelacji między zmiennymi i wymiarami są wykorzystywane jako współrzędne.

Effect.on.health i healthy są najbardziej skorelowane z wymiarem 1 gdy sugar i frequency są najbardziej skorelowane z wymiarem 2.



Po otrzymanych wynikach możemy dostrzec, że osoby chore sięgają po herbatę w celu poprawy zdrowia i piją ją częściej niż osoby zdrowe, które nie oczekują po herbacie poprawy ich zdrowia. Dodatkowo wraz ze zwiększeniem częstotliwości picia herbaty, respondenci przestawali ją słodzić.

Skalowanie Wielowymiarowe

	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

Dane Swiss

Dane Swiss zawierają informacje na temat płodności i statusu socjo-ekonomicznego dla 47 prowincji w których głównym językiem jest język francuski w Szwajcarii.

Dane mtcars zawierają informacje na temat modelu samochodów oraz ich parametrów silnika.

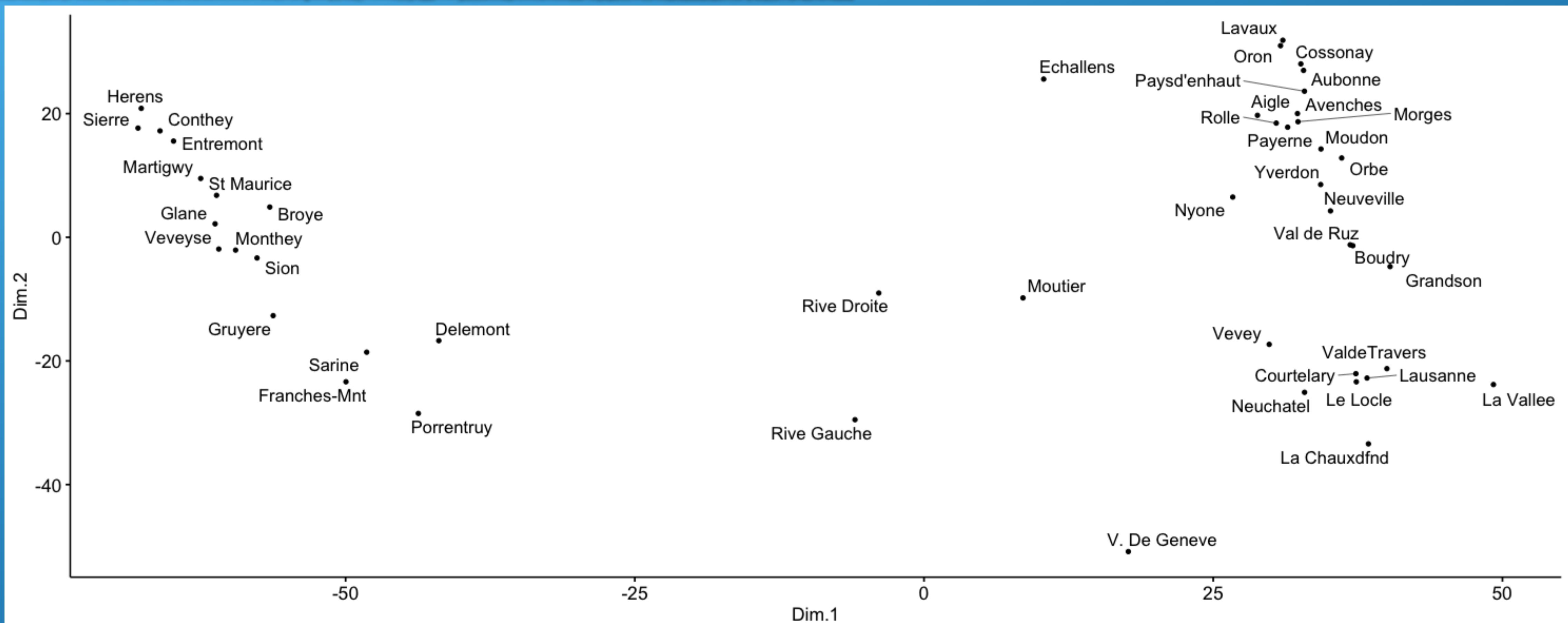
Na tych danych zostaną wykonane analizy Skalowania Wielowymiarowego.

	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

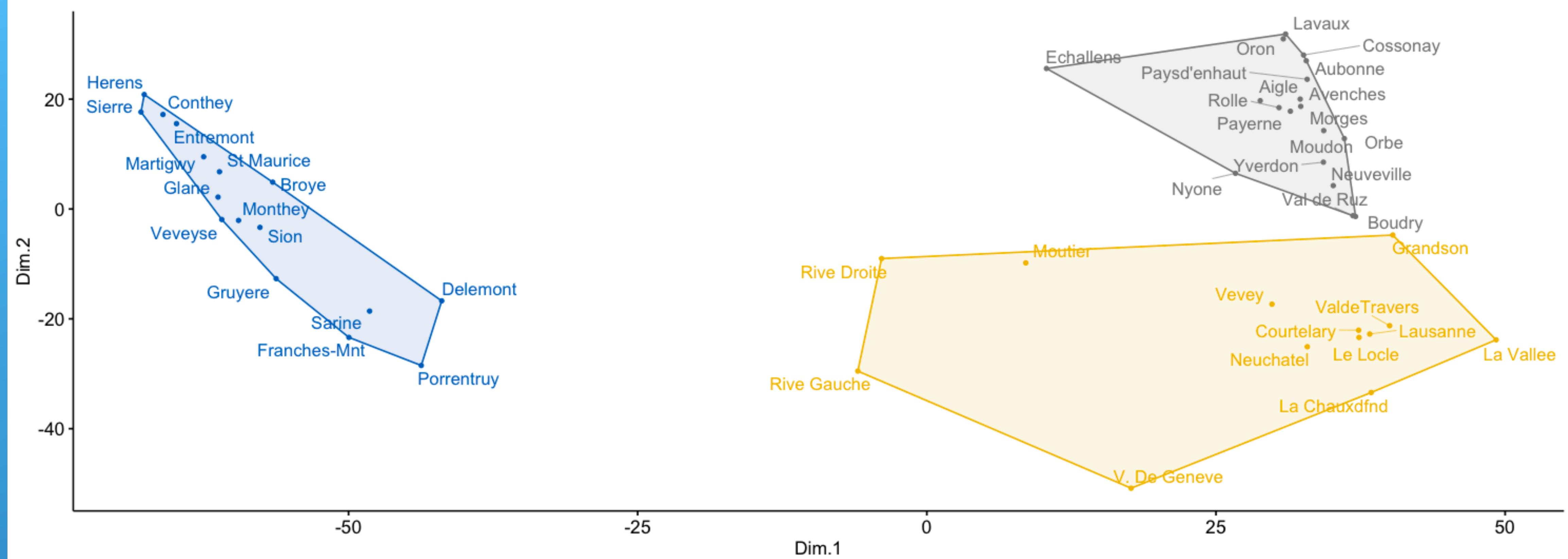
Dane mtcars

Takie wyniki otrzymaliśmy dla Skalowania Wielowymiarowego Metrycznego dla wybranych kolumn 1,2,5, i 6 (wszystkie dane ilościowe). Widać że wykres przedstawia 2-3 ścisłe grupy które otrzymują podobne wyniki z niewieloma outlierami. Na kolejnym slajdzie wykonam grupowanie k-średnich dla tych wyników.

Swiss - metryczny

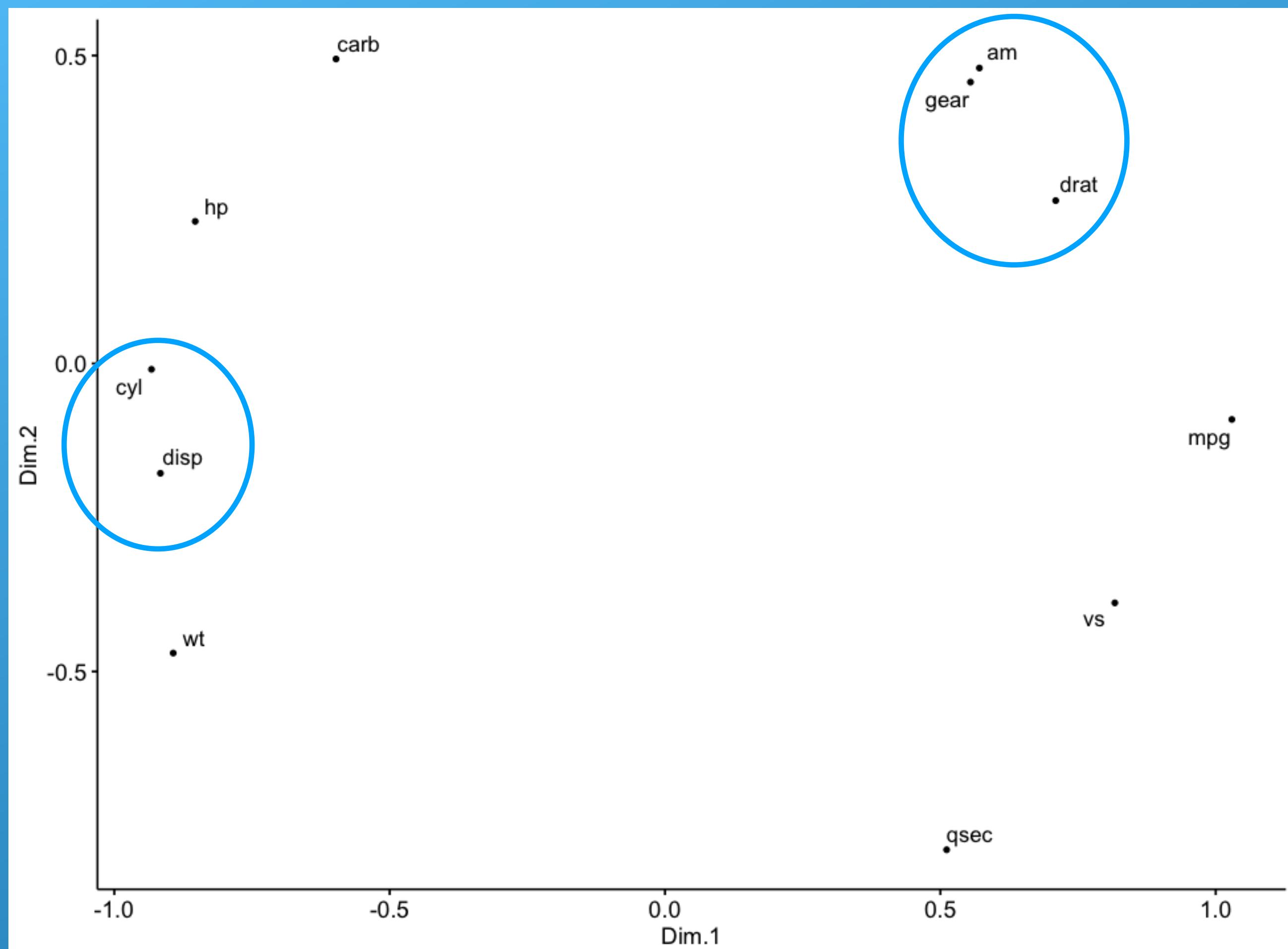


groups a 1 a 2 a 3



Za pomocą metody grupowania k-średnich (k-mean clustering), można było utworzyć 3 widoczne grupy, które oznaczone są kolorami. Grupy zostały utworzone ze względu na podobieństwo w otrzymanych wynikach w naszych danych. Jak widać grupa oznaczona kolorem niebieskim jest najbardziej spoista, czarna trochę mniej, żółta jest najbardziej rozległa i są tam największe różnice w wynikach. Być może gdybyśmy zmienili liczbę grup, jaką chcielibyśmy otrzymać to właśnie ta większa liczba grup wpłynęłaby na rozbicie tej grupy na pomniejsze podgrupy z bardziej zbliżonymi (podobnymi) wartościami.

mtcars- niemetryczny



Pozytywnie skorelowane (podobne) obiekty znajdują się blisko siebie po tej samej stronie wykresu. Am, gear oraz drat będą mocno skorelowane ze sobą. Cyl i disp również znajdują się blisko siebie. Obie grupy zakreśliłem na prezentacji. Reszta obiektów znajduje się dosyć daleko, żeby móc dostrzec teraz jakiś związek pomiędzy nimi.

Podsumowanie

Podsumowanie



Główne cele analiz

- Głównym celem MDS (Skalowania Wielowymiarowego) jest **redukcja wielowymiarowości danych** zachowując przy tym **najwięcej informacji o odległościach między obiektami**- czyli szuka reprezentacji obiektów, która najlepiej odzwierciedla strukturę odległości między obiektami.
- AC (Analiza korespondencji) ma na celu **analizę związku między dwoma zmiennymi kategorycznymi**. Próbuje znaleźć **wzorce i zależności między kategoriami** zmiennych w celu **wykrycia ukrytej struktury danych**.

Rodzaj danych

- MDS: Może być stosowane do danych **numerycznych lub miar odległości między obiektami**. Przykładami mogą być macierze odległości, podobieństwa, korzystanie z metryk euklidesowych itp.
- CA: Jest stosowana do danych **kategorycznych lub danych, które można przekształcić na formę kategoryczną**. Dane wejściowe dla CA to **tabela kontyngencji**, która zawiera liczbę obserwacji dla różnych kombinacji kategorii.

Wynik analiz

- MDS: Wynikiem MDS jest **przekształcona reprezentacja danych w niższej wymiarowości**, która najlepiej zachowuje odległości między obiektami. Wynikiem może być np. dwuwymiarowa mapa, która przedstawia relacje między obiektami.
- CA: Wynikiem CA są **dwie mapy korespondencji**, które **przedstawiają związki między kategoriami zmiennych**. Mapa korespondencji może pokazywać, jakie kategorie zmiennych są ze sobą powiązane i jakie kombinacje kategorii są najczęstsze.

Podsumowanie

- Podsumowując, **MDS** jest techniką **skalowania wielowymiarowego**, która **redukuje wymiarowość danych numerycznych**, podczas gdy **AC** jest techniką **analizy związków między zmiennymi kategorycznymi**. Oba podejścia są używane w różnych kontekstach i mają różne cele analizy danych.

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%2Fstmulsc.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>

Dziękuję za uwagę

Daria Plewa