

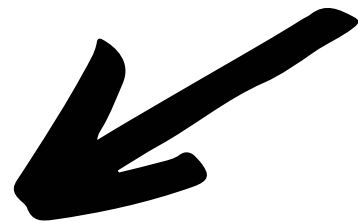
ANALIZA ANKIETY WŚRÓD STUDENTÓW

*WPŁYW CZYNNIKÓW SPOŁECZNO-EKONOMICZNYCH, SYTUACJI RODZINNEJ ORAZ
INDYWIDUALNYCH PREDYSPOZYCJI NA WYNIKI AKADEMICKIE*

CELE I MOTYWACJA

Celem niniejszego projektu jest zbadanie zależności pomiędzy różnymi czynnikami z życia studentów, a ich wynikami w bieżących semestrach.

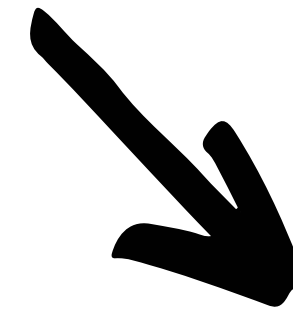
Przeprowadzenie tej analizy może przynieść istotne korzyści w kilku kontekstach:



*Dla instytucji edukacyjnych,
szczególnie niepublicznych*



Dla rodziców




Dla młodzieży


WPROWADZENIE DO ZBIORU

Zbiór jest reprezentacją wyników ankiety przeprowadzonej wśród studentów Uniwersytetu w Malezji (został dodany 07.10.2024r.). Zawiera on odpowiedzi na pytania dotyczące zarówno edukacji, jak i życia poza szkolnego. Zawiera niemalże 500 obserwacji oraz 16 cech przedstawionych obok.

 **DEPARTMENT**

 **GENDER**

 **HSC** Wynik uzyskany w wykształceniu średnim wyższego szczebla.

 **SSC** Wynik uzyskany w wykształceniu średnim niższego szczebla.

 **INCOME** Miesięczny przychód rodziców

 **HOMETOWN**

 **COMPUTER** Poziom zaawansowania w obsłudze komputera

 **PREPARATION** Czas spędzony na samodzielnej nauce

 **GAMING**

 **ATTENDANCE**

 **JOB**

 **ENGLISH** Zdolność do komunikacji w języku angielskim

 **EXTRA** Uczestnictwo w zajęciach dodatkowych

 **SEMESTER**

 **LAST** Średnia z ostatniego semestru

 **OVERALL** Średnia ogółem

PROBLEMY BADAWCZE

I

HIPOTEZY

PROBLEM BADAWCZY

Jak czynniki demograficzne
oraz społeczno-
ekonomiczne wpływają na
osiągnięcia edukacyjne?

HIPOTEZA

Studenci z wyższym
statusem społeczno-
ekonomicznym osiągają
lepsze wyniki akademickie
niż studenci z niższym
statusem.

PROBLEM BADAWCZY

Czy aktywności
pozalekcyjne mają
pozytywny wpływ na wyniki
akademickie studentów?

HIPOTEZA

Studenci angażujący się w
aktywności pozalekcyjne
osiągają lepsze wyniki
akademickie niż ci, którzy
nie biorą udziału w takich
zajęciach.

PROBLEM BADAWCZY

Jak regularna obecność na
zajęciach przekłada się na
wyniki akademickie?

HIPOTEZA

Większa liczba godzin
poświęconych na naukę i
obecność na zajęciach jest
skorelowana z lepszymi
wynikami w testach.

PRZYKŁADOWE REKORDY

Department	Gender	HSC	SSC	Income	Hometown	Computer	Preparation	Gaming	Attendance	Job	English	Extra	Semester	Last	Overall
Business Administration	Male	4.17	4.84	Low (Below 15,000)	Village	3	More than 3 Hours	0-1 Hour	80%-100%	No	3	Yes	6th	3.220	3.350
Business Administration	Female	4.92	5.00	Upper middle (30,000-50,000)	City	3	0-1 Hour	0-1 Hour	80%-100%	No	3	Yes	7th	3.467	3.467
Business Administration	Male	5.00	4.83	Lower middle (15,000-30,000)	Village	3	0-1 Hour	More than 3 Hours	80%-100%	No	4	Yes	3rd	4.000	3.720
Business Administration	Male	4.00	4.50	High (Above 50,000)	City	5	More than 3 Hours	More than 3 Hours	80%-100%	No	5	Yes	4th	3.800	3.750
Business Administration	Female	2.19	3.17	Lower middle (15,000-30,000)	Village	3	0-1 Hour	2-3 Hours	80%-100%	No	3	Yes	4th	3.940	3.940
Computer Science and Engineering	Male	4.75	4.05	Lower middle (15,000-30,000)	Village	3	0-1 Hour	More than 3 Hours	Below 40%	No	4	No	2nd	1.000	1.000
Computer Science and Engineering	Male	4.42	5.00	High (Above 50,000)	Village	4	0-1 Hour	More than 3 Hours	60%-79%	No	2	No	2nd	1.060	1.060
Computer Science and Engineering	Male	4.50	4.81	Upper middle (30,000-50,000)	City	3	2-3 Hours	More than 3 Hours	80%-100%	No	4	Yes	11th	2.950	1.250
Computer Science and Engineering	Male	3.32	4.50	Low (Below 15,000)	City	4	0-1 Hour	More than 3 Hours	Below 40%	No	3	Yes	5th	1.420	1.440
Computer Science and Engineering	Female	3.33	4.95	Lower middle (15,000-30,000)	City	3	0-1 Hour	More than 3 Hours	Below 40%	No	4	No	2nd	1.500	1.500

PRZYGOTOWANIE DANYCH

0 IMPLEMENTACJA BIBLIOTEK

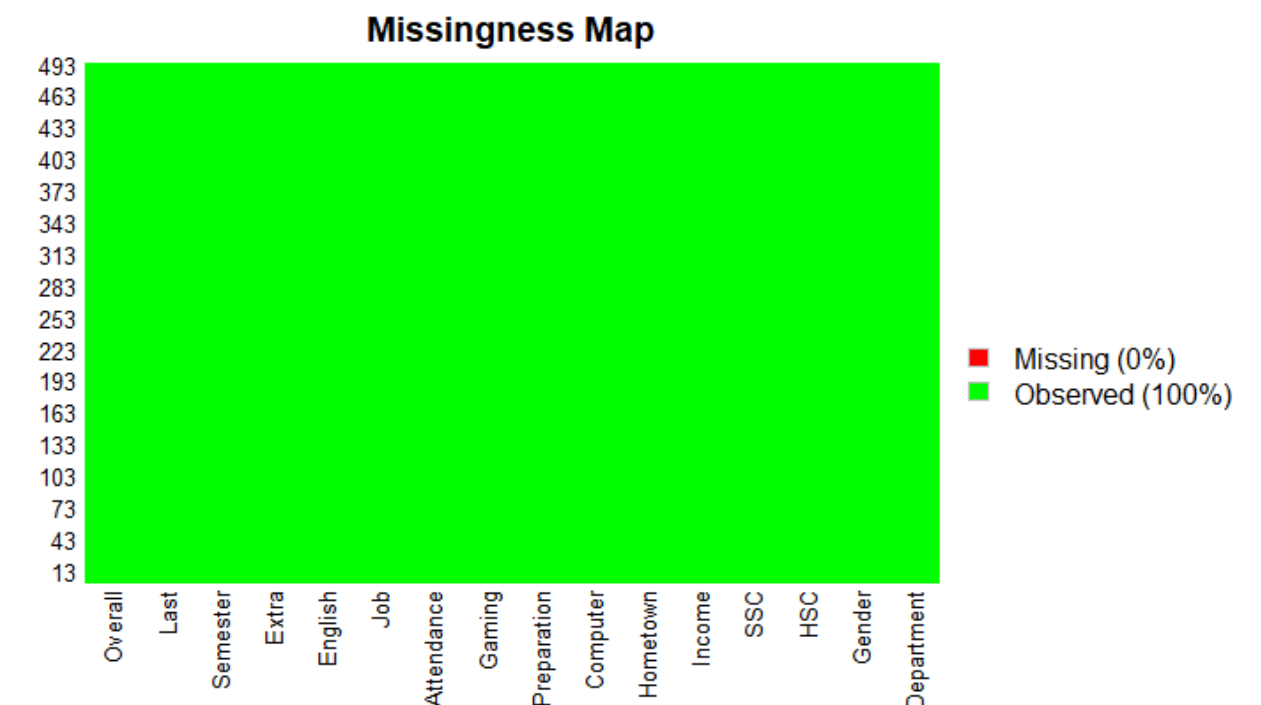
```
library(rio)
library(tidyverse)
library(gridExtra)
library(RColorBrewer)
library(patchwork)
library(ggrepel)
library(ggthemes)
library(ggcorrplot)
library(corrplot)
library(vcd)
library(viridis)
library(GGally)
library(Amelia)
library(cluster)
library(clustMixType)
library(caret)
library(networkD3)
library(dplyr)
library(htmltools)
```

1 IMPORT, SPRAWDZENIE BRAKÓW DANYCH ORAZ LICZBY REKORDÓW

```
data <- import("ResearchInformation3.csv")

any(is.na(data))
missmap(data, col = c("red", "green"), margins = c(5,2))
nrow(data)
```

```
[1] FALSE
[1] 493
```



2 DANE PRZED ZAMIANĄ ICH TYPÓW

Department	Gender	HSC	SSC	Income	Hometown
Length:493	Length:493	Min. :2.170	Min. :3.000	Length:493	Length:493
Class :character	Class :character	1st Qu.:3.830	1st Qu.:4.680	Class :character	Class :character
Mode :character	Mode :character	Median :4.170	Median :4.940	Mode :character	Mode :character
		Mean :4.157	Mean :4.768		
		3rd Qu.:4.500	3rd Qu.:5.000		
		Max. :5.000	Max. :5.000		
Computer	Preparation	Gaming	Attendance	Job	English
Min. :1.000	Length:493	Length:493	Length:493	Length:493	Min. :1.00
1st Qu.:3.000	Class :character	Class :character	Class :character	Class :character	1st Qu.:3.00
Median :3.000	Mode :character	Mode :character	Mode :character	Mode :character	Median :4.00
Mean :3.339					Mean :3.57
3rd Qu.:4.000					3rd Qu.:4.00
Max. :5.000					Max. :5.00
Extra	Semester	Last	Overall		
Length:493	Length:493	Min. :1.000	Min. :1.000		
Class :character	Class :character	1st Qu.:2.810	1st Qu.:2.880		
Mode :character	Mode :character	Median :3.250	Median :3.270		
		Mean :3.164	Mean :3.188		
		3rd Qu.:3.670	3rd Qu.:3.680		
		Max. :4.000	Max. :4.000		

3 ZMIANA TYPÓW DANYCH

```
```{r}
data$Income[data$Income == "Low (Below 15,000) "] <- "Low (Below 15,000)"
data$Income[data$Income == "Lower middle (15,000-30,000) " |
 data$Income == "Lower middle (15,000-30,000) "] <- "Lower middle (15,000-30,000)"
data$Income[data$Income == "Upper middle (30,000-50,000) "] <- "Upper middle (30,000-50,000)"
data$Income[data$Income == "High (Above 50,000) " | data$Income == "High (Above 50,000) "] <- "High (Above 50,000)"
```
```

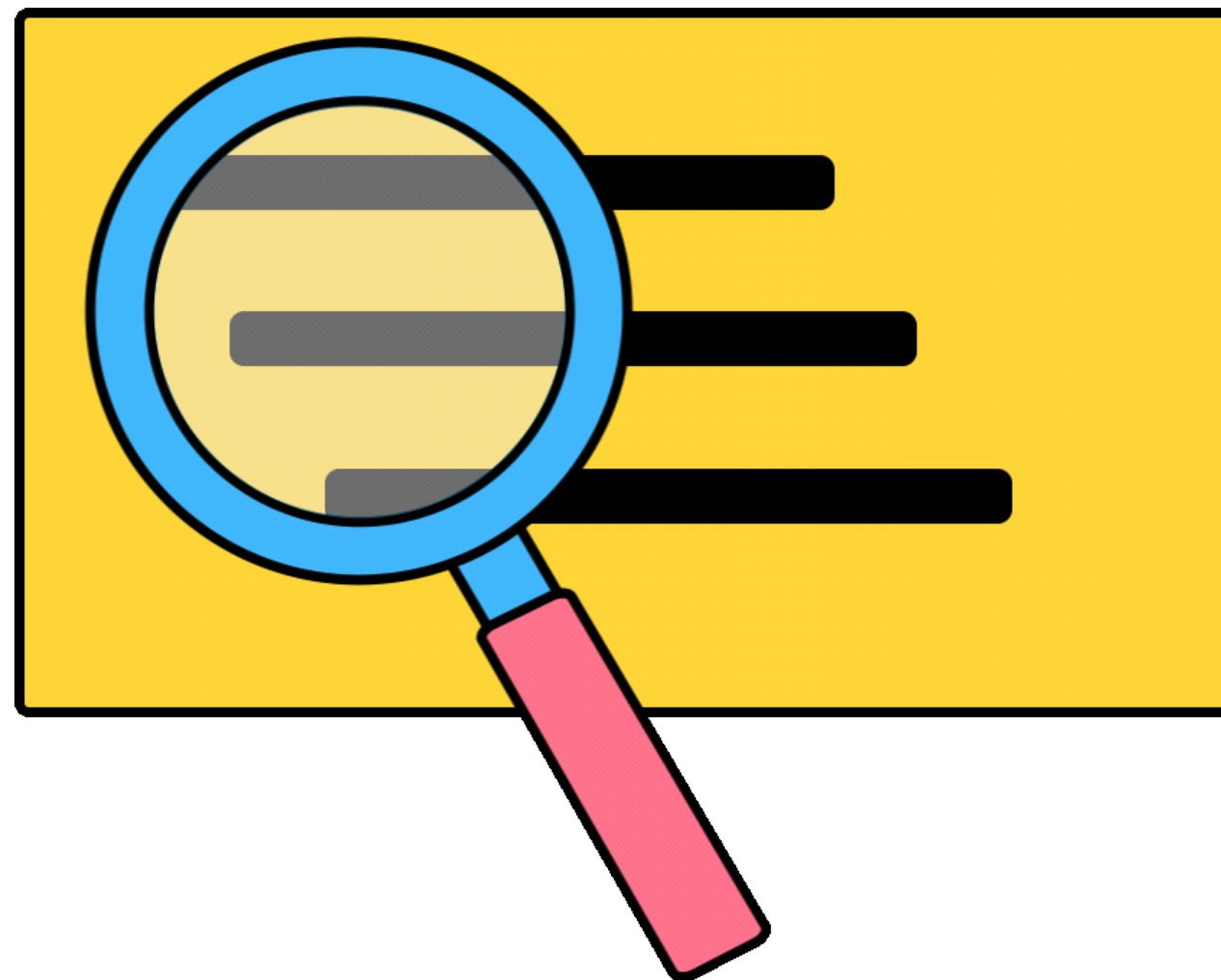
```
```{r}
data$Department <- as.factor(data$Department)
data$Gender <- as.factor(data$Gender)
data$Income <- ordered(data$Income, levels=c("Low (Below 15,000)", "Lower middle (15,000-30,000)",
 "Upper middle (30,000-50,000)", "High (Above 50,000)"))

data$Hometown <- as.factor(data$Hometown)
data$Computer <- ordered(data$Computer)
data$Preparation <- ordered(data$Preparation)
data$Gaming <- ordered(data$Gaming)
data$Attendance <- ordered(data$Attendance, levels=c("Below 40%", "40%-59%", "60%-79%", "80%-100%"))
data$Job <- as.factor(data$Job)
data$English <- ordered(data$English)
data$Extra <- as.factor(data$Extra)
data$Semester <- ordered(data$Semester, levels=c("2nd", "3rd", "4th", "5th", "6th", "7th", "8th", "9th", "10th", "11th", "12th"))
```
```

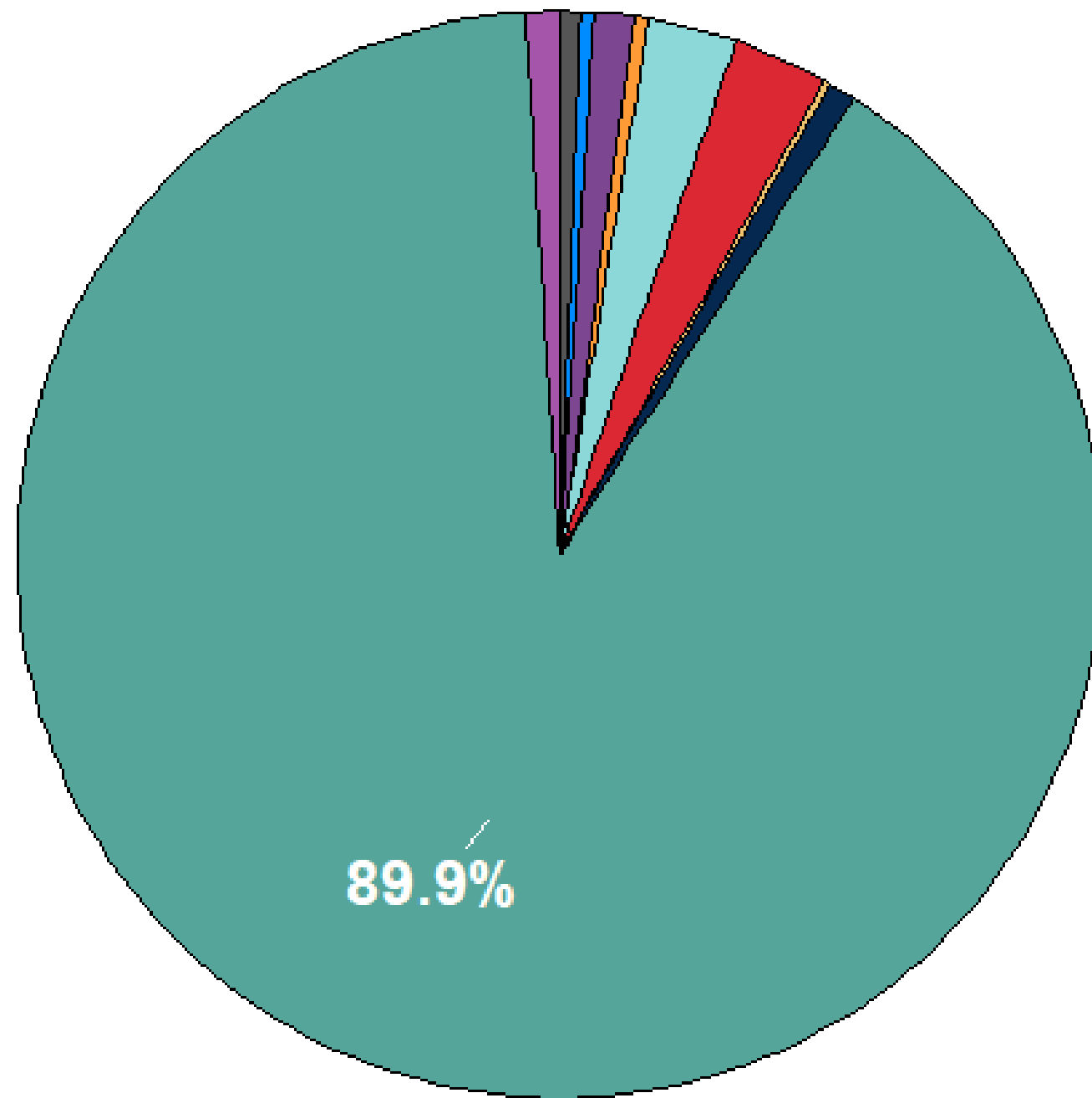

4 ZMIANA TYPÓW DANYCH

| | | Department | Gender | | HSC | | SSC | | |
|--|---------------|------------|-------------|----------|--------------------|-------------|--------------------|--------|---------------|
| Computer Science and Engineering | | :443 | Female: | 165 | Min. : | 2.170 | Min. : | 3.000 | |
| English | | : 14 | Male : | 328 | 1st Qu.: | 3.830 | 1st Qu.: | 4.680 | |
| Journalism, Communication and Media Studies: | | 13 | | | Median : | 4.170 | Median : | 4.940 | |
| Political Science | | : 6 | | | Mean : | 4.157 | Mean : | 4.768 | |
| Business Administration | | : 5 | | | 3rd Qu.: | 4.500 | 3rd Qu.: | 5.000 | |
| Economics | | : 4 | | | Max. : | 5.000 | Max. : | 5.000 | |
| (Other) | | : 8 | | | | | | | |
| | | Income | Hometown | Computer | | Preparation | | Gaming | Attendance |
| Low (Below 15,000) | | : 74 | City :213 | 1: 58 | 0-1 Hour | :218 | 0-1 Hour | : 50 | Below 40%: 11 |
| Lower middle (15,000-30,000): | | 180 | Village:280 | 2: 42 | 2-3 Hours | :228 | 2-3 Hours | :141 | 40%-59% : 68 |
| Upper middle (30,000-50,000): | | 110 | | 3:183 | More than 3 Hours: | 47 | More than 3 Hours: | 302 | 60%-79% :217 |
| High (Above 50,000) | | :129 | | 4: 95 | | | | | 80%-100% :197 |
| | | | | 5:115 | | | | | |
| | | | | | | | | | |
| Job | English Extra | | Semester | | Last | | Overall | | |
| No :459 | 1: 8 | No :288 | 2nd | :183 | Min. : | 1.000 | Min. : | 1.000 | |
| Yes: 34 | 2: 38 | Yes:205 | 10th | : 57 | 1st Qu.: | 2.810 | 1st Qu.: | 2.880 | |
| | 3:176 | | 8th | : 54 | Median : | 3.250 | Median : | 3.270 | |
| | 4:207 | | 3rd | : 35 | Mean : | 3.164 | Mean : | 3.188 | |
| | 5: 64 | | 5th | : 33 | 3rd Qu.: | 3.670 | 3rd Qu.: | 3.680 | |
| | | | 9th | : 30 | Max. : | 4.000 | Max. : | 4.000 | |
| | | | (Other): | 101 | | | | | |

PRZEKROJOWY PRZEGLĄD DANYCH

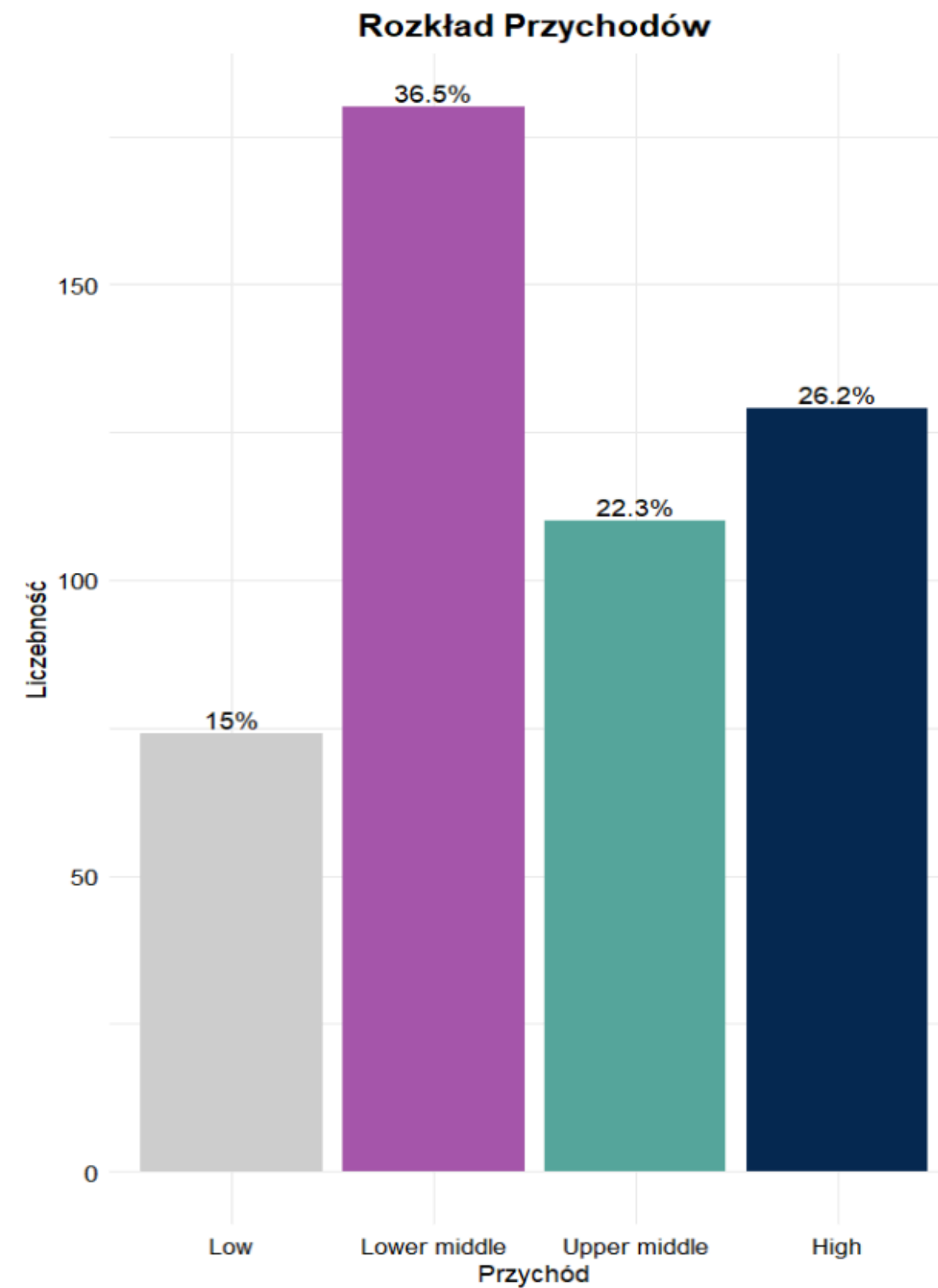
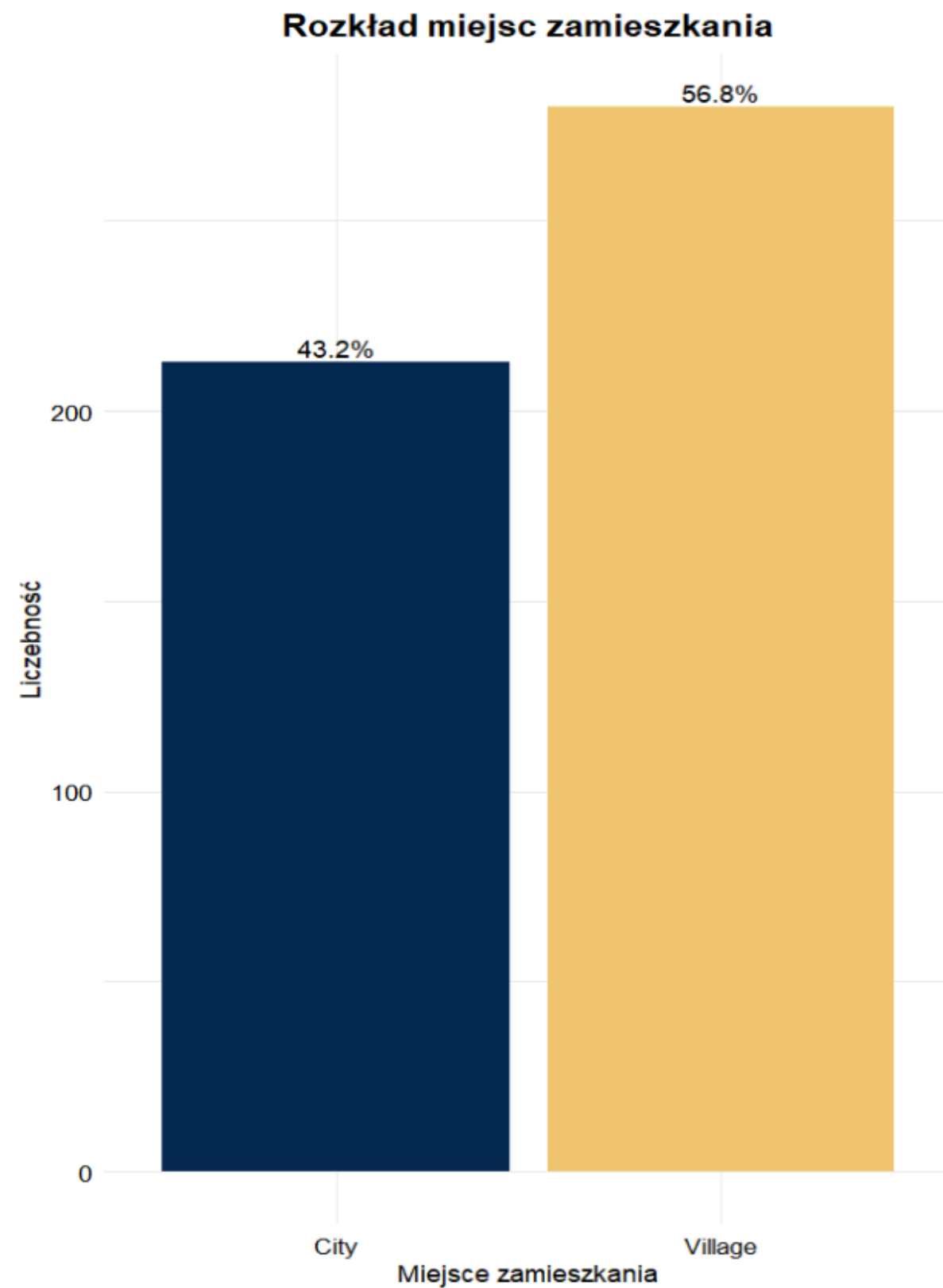
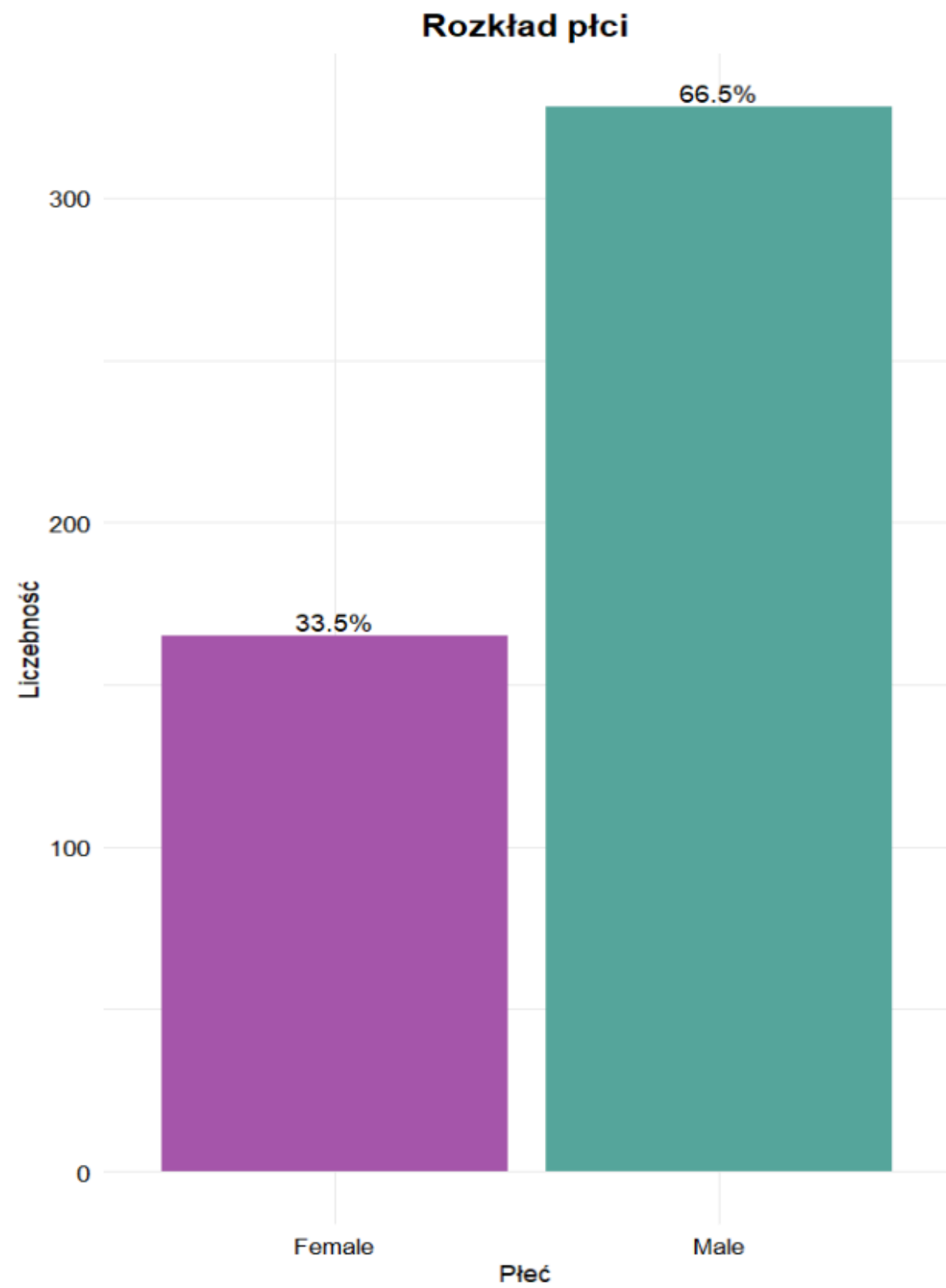


Kierunek studiów

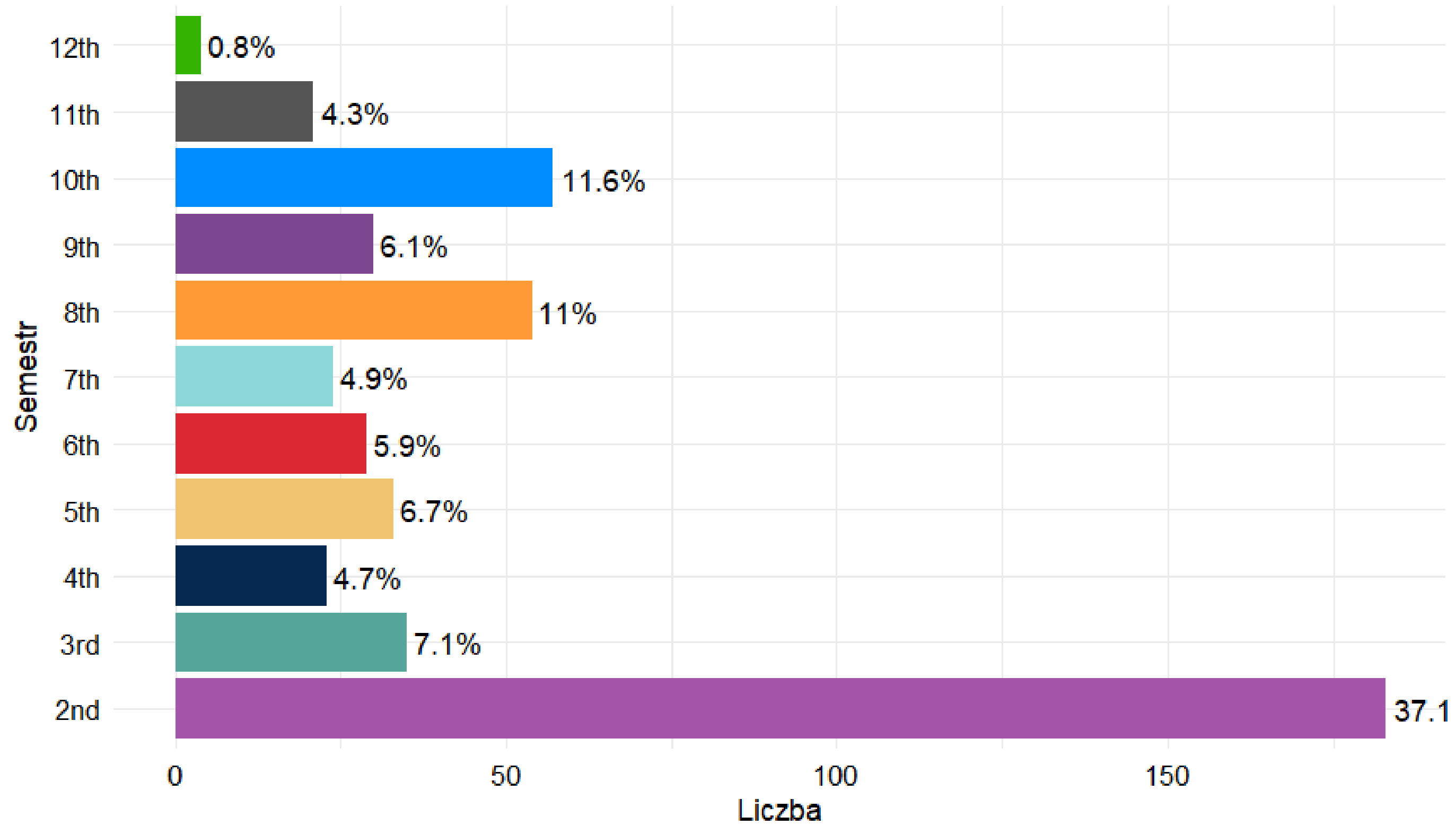


Department

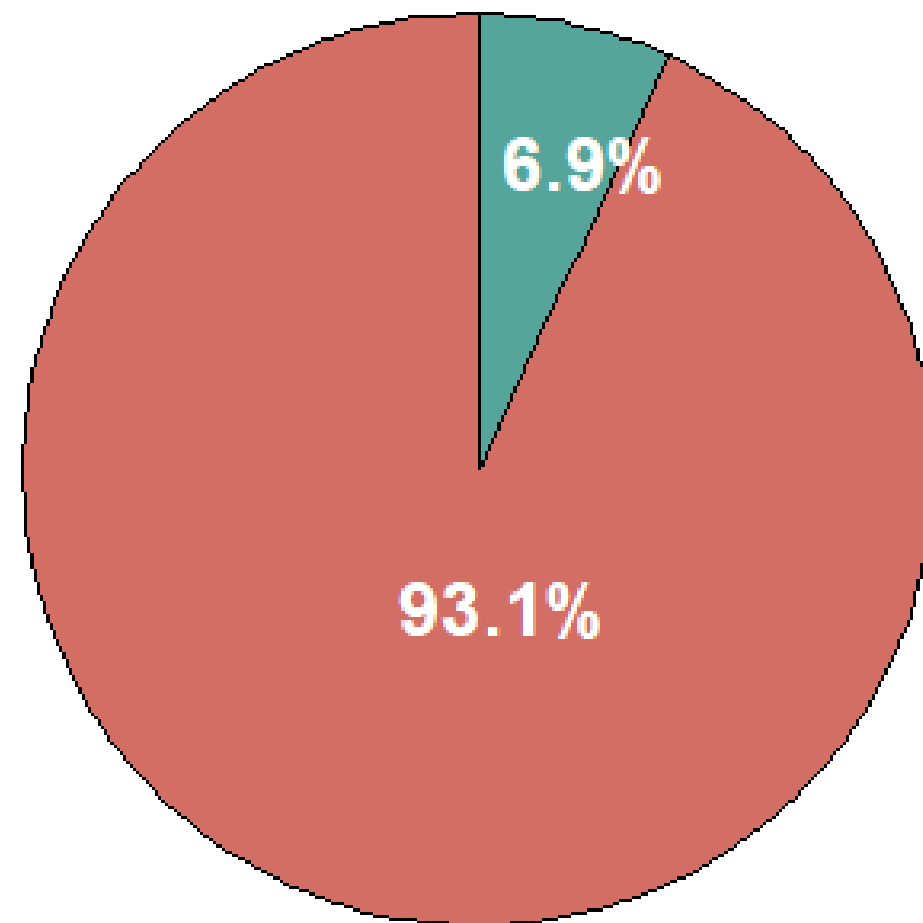
- Business Administration
- Computer Science and Engineering
- Economics
- Electrical and Electronic Engineering
- English
- Journalism, Communication and Media Studies
- Law and Human Rights
- Political Science
- Public Health
- Sociology



Rozkład według semestru studiów



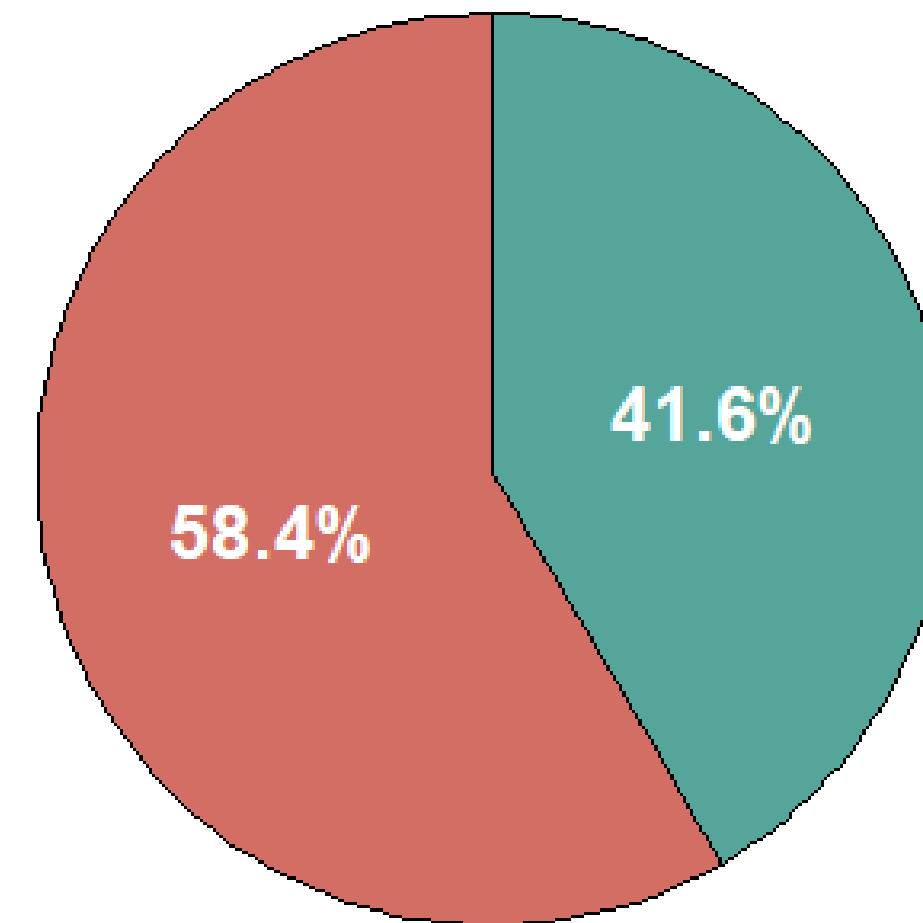
Praca



Job

| |
|-----|
| No |
| Yes |

Zajęcia dodatkowe



Extra

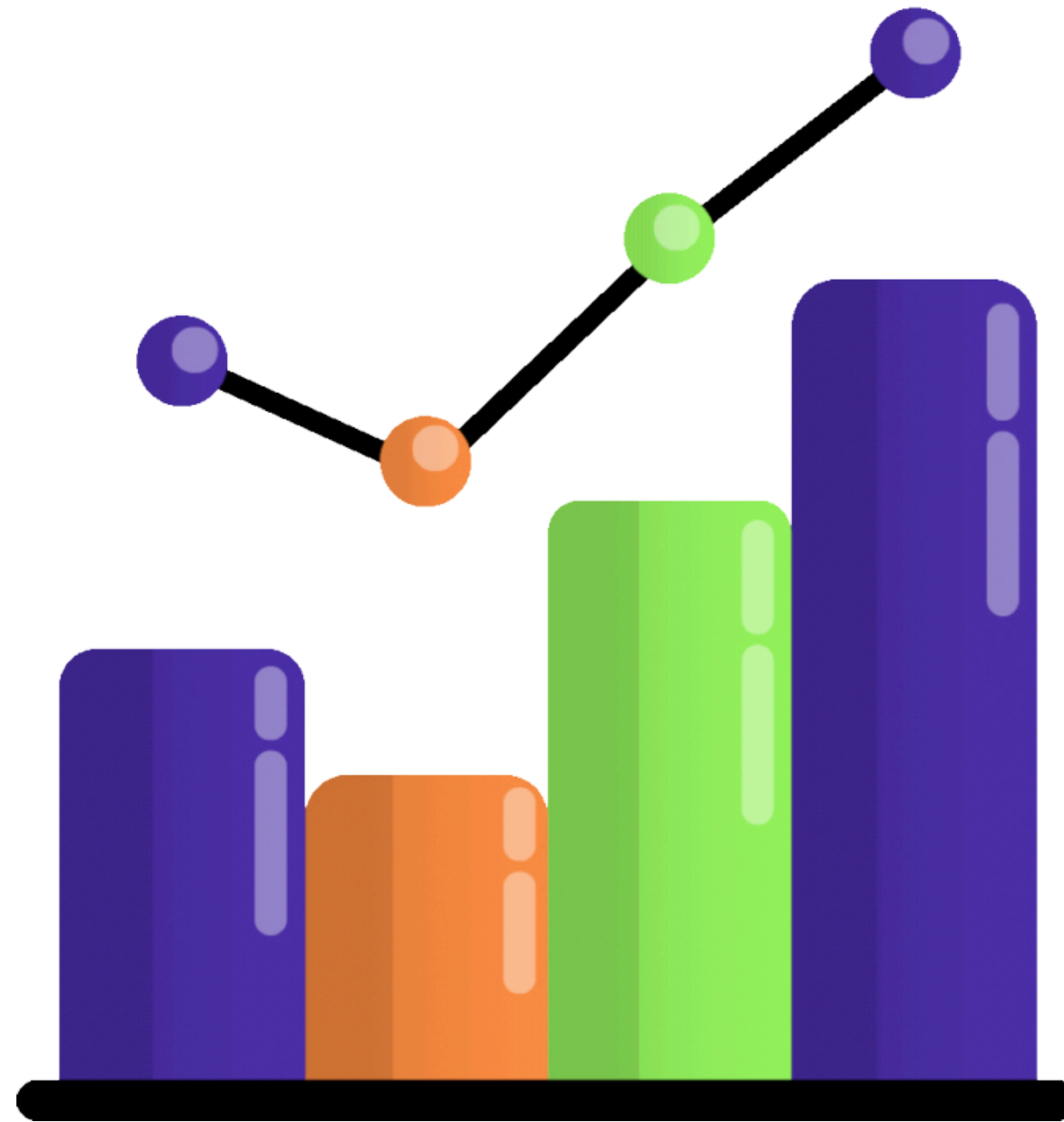
| |
|-----|
| No |
| Yes |

SYLWETKA PRZECIĘTNEGO ANKIETOWANEGO

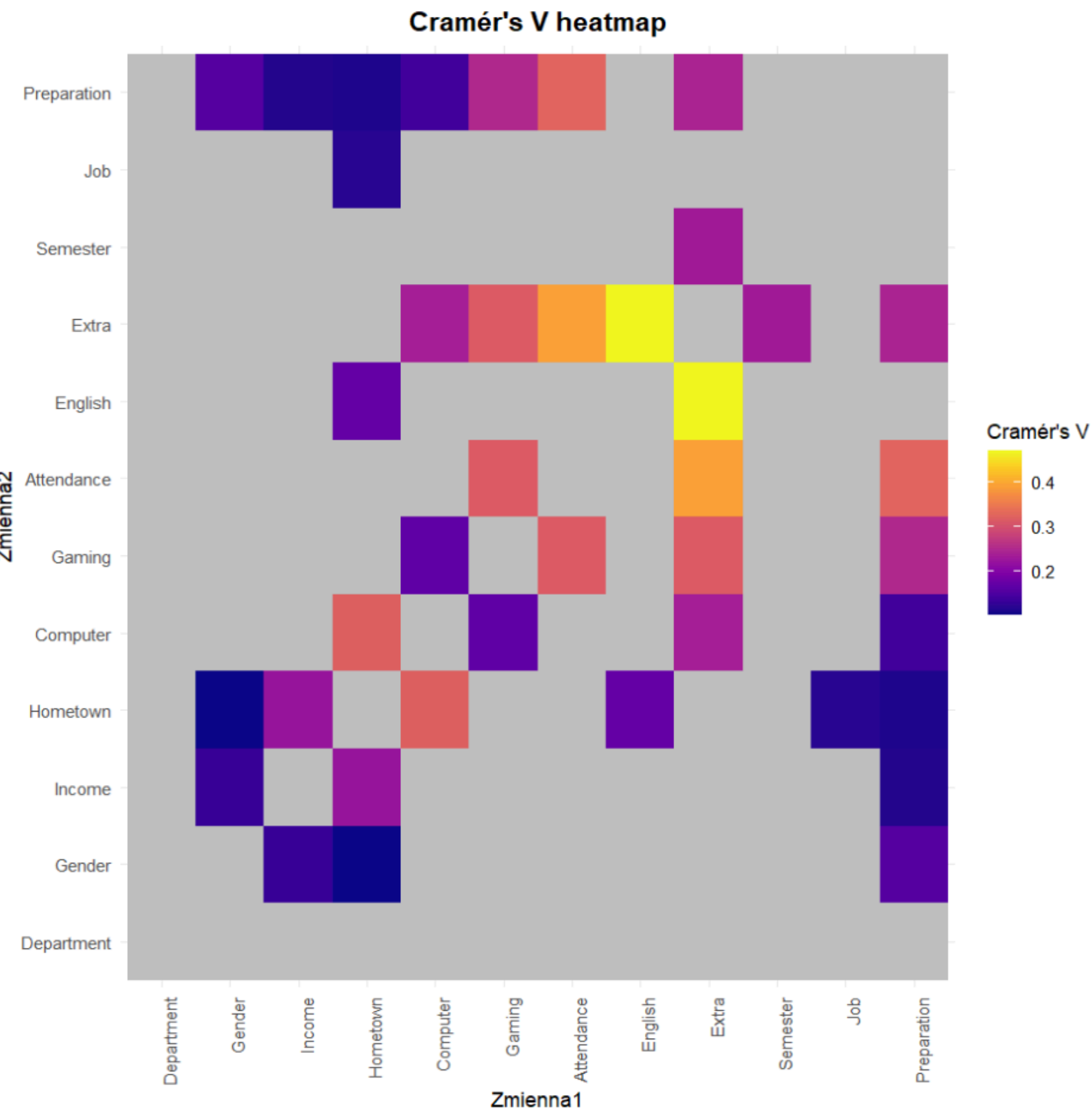
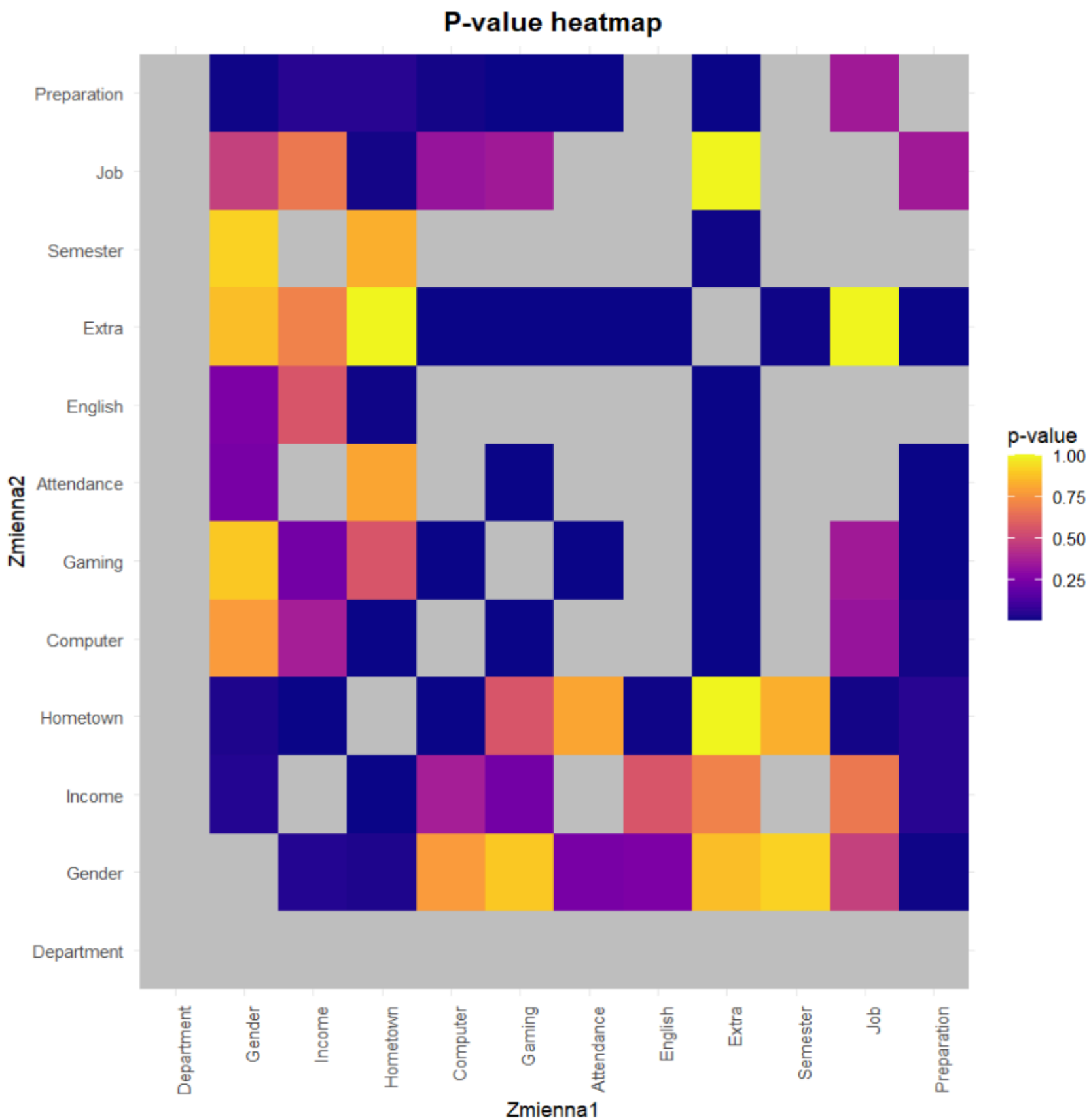


- Płeć: **mężczyzna**
- Miejsce zamieszkania: **wieś**
- Kierunek: **computer science and engineering (umysł ścisły)**
- Semestr: **drugi**
- Zamożność: **klasa średnia**
- Praca: **brak**
- Zajęcia dodatkowe: **nie**

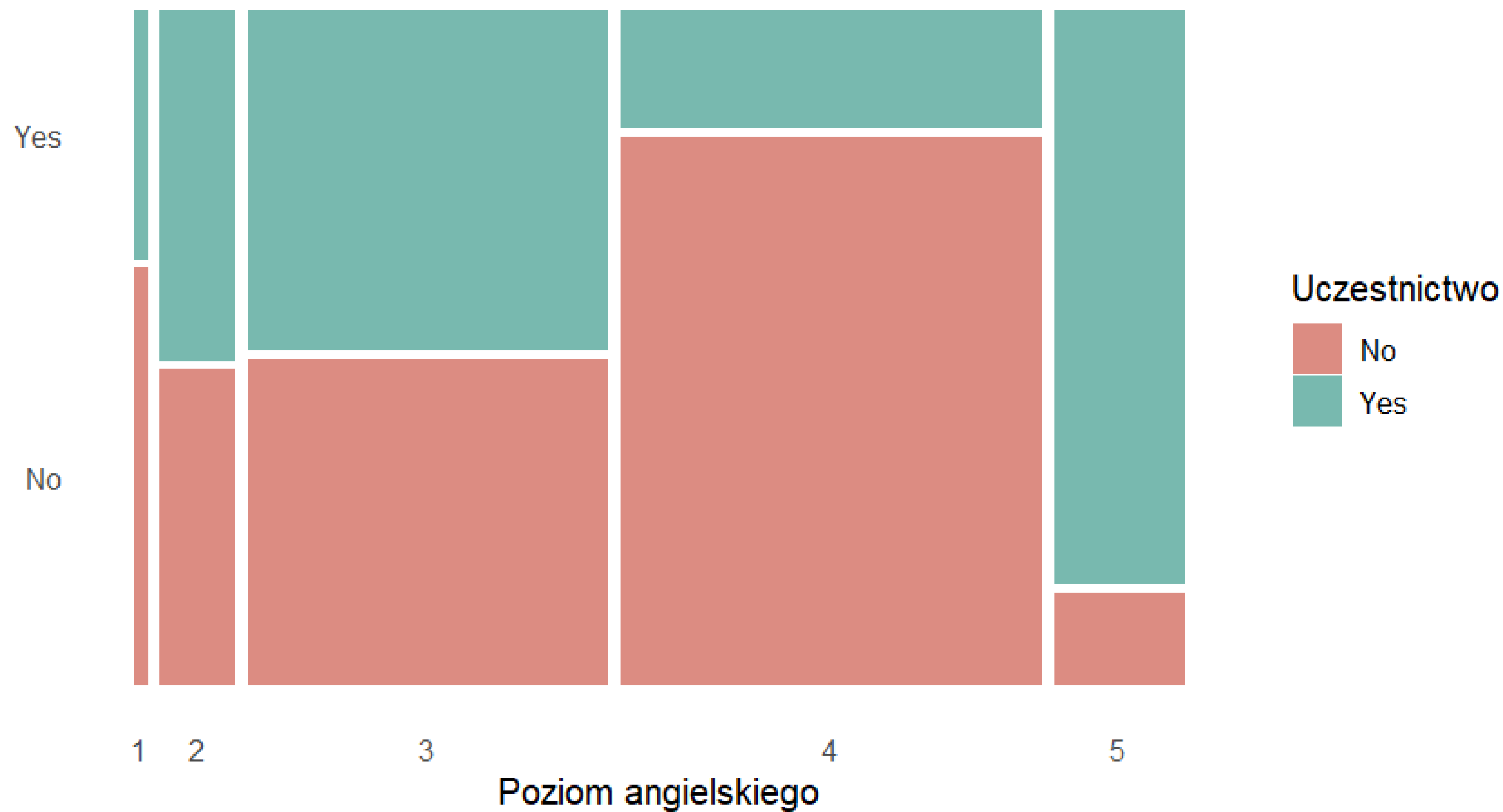
ZALEŻNOŚCI POMIĘDZY ZMIENNYMI



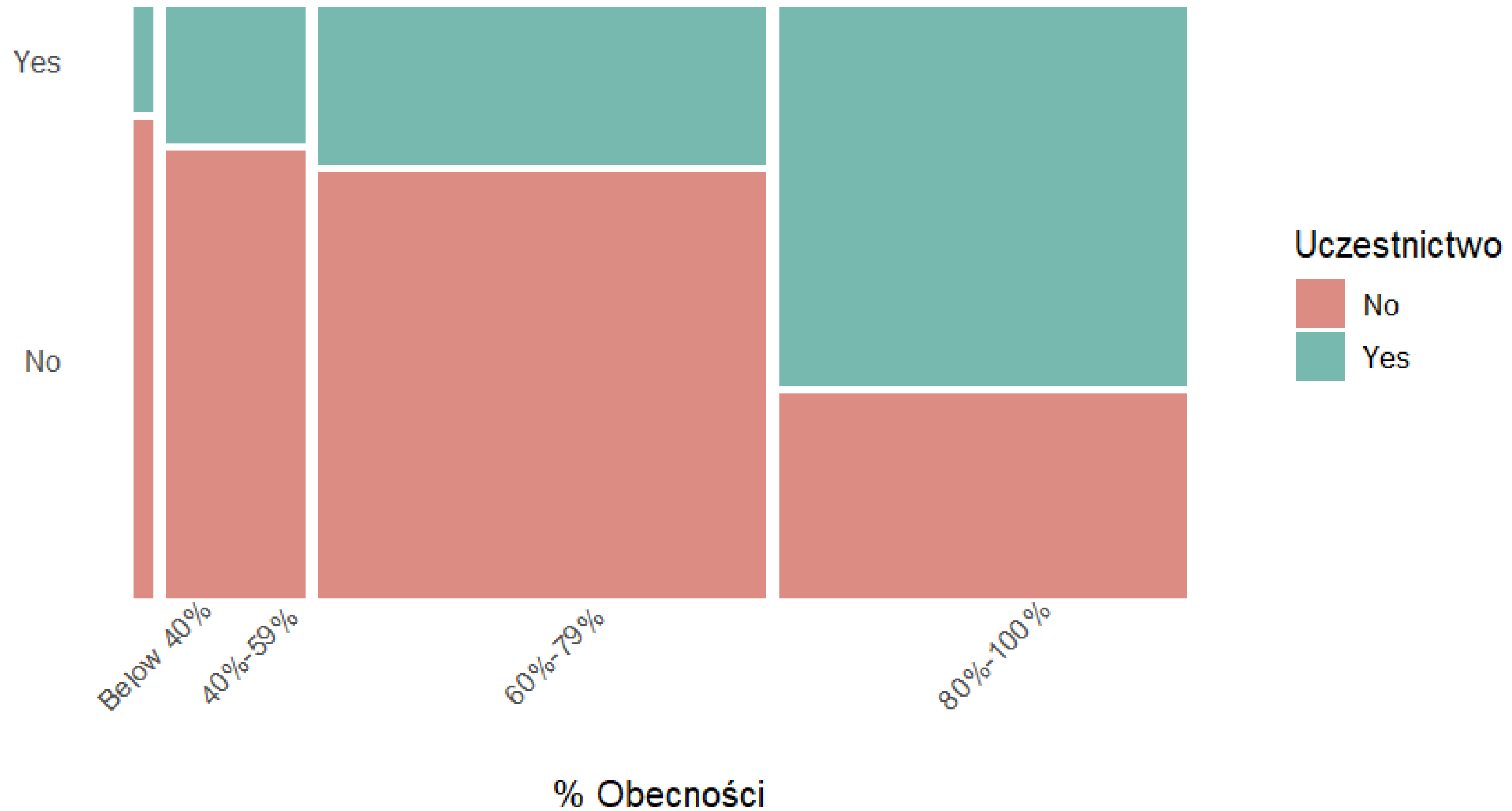
ZMIENNE KATEGORYCZNE



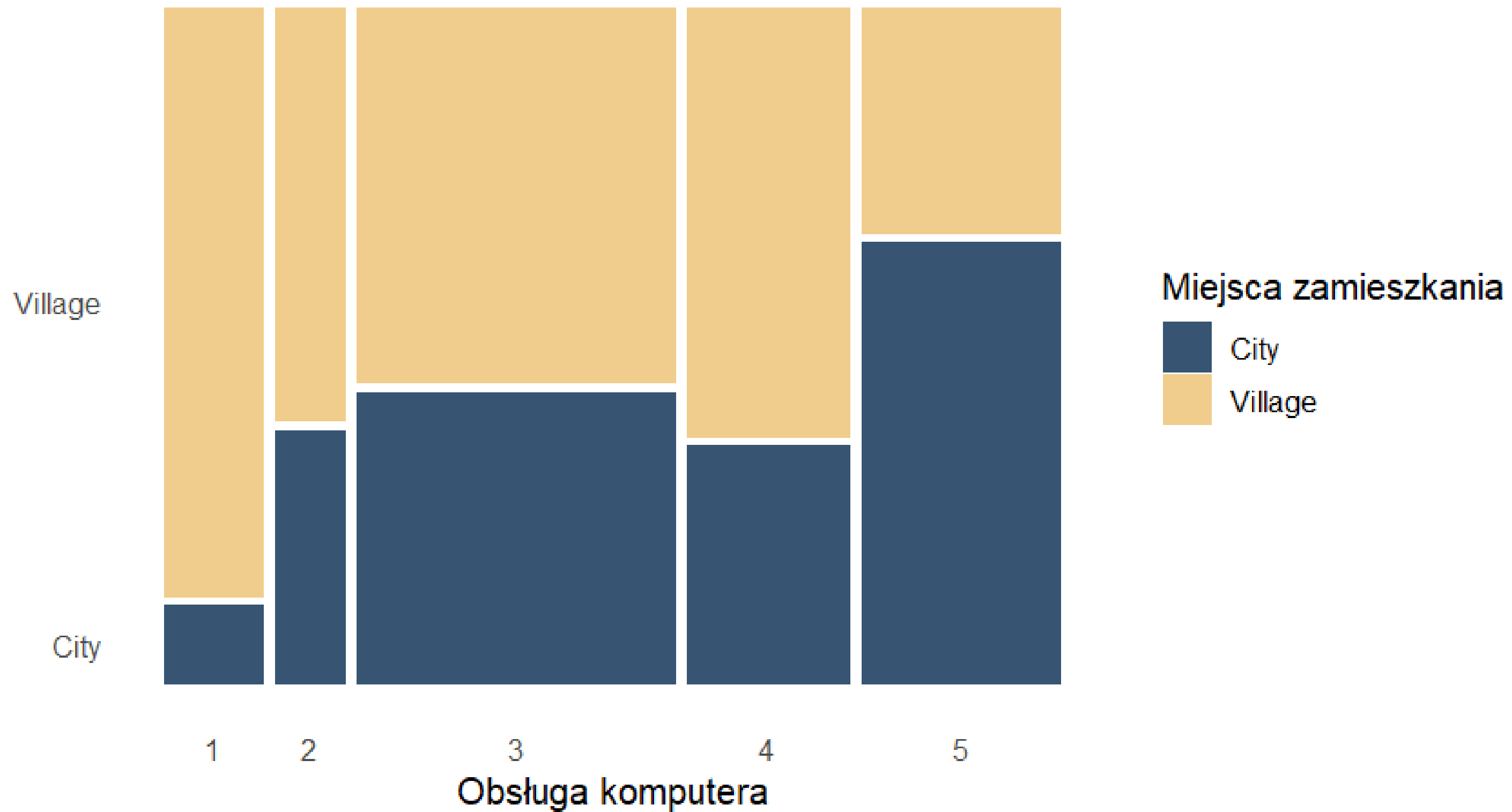
Związek między poziomem angielskiego, a uczestnictwem w zajęciach dodatkowych



Związek między obecnością, a uczestnictwem w zajęciach pozalekcyjnych



Związek między umiejętnościami obsługi komputera, a miejscem zamieszkania

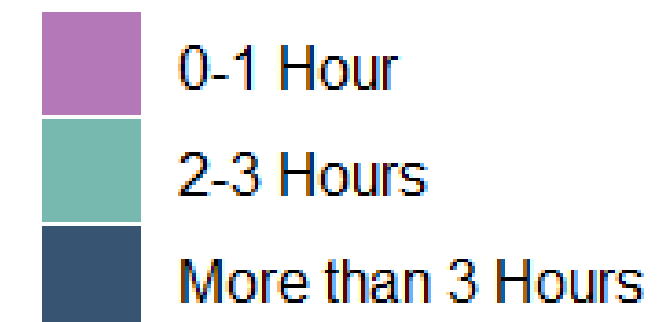


Związek między Obecnością, a czasem spędzonym na naukę

More than 3 Hours

0-1 Hour

Czas poświęcony na naukę



Below 40%
40%-59%

60%-79%

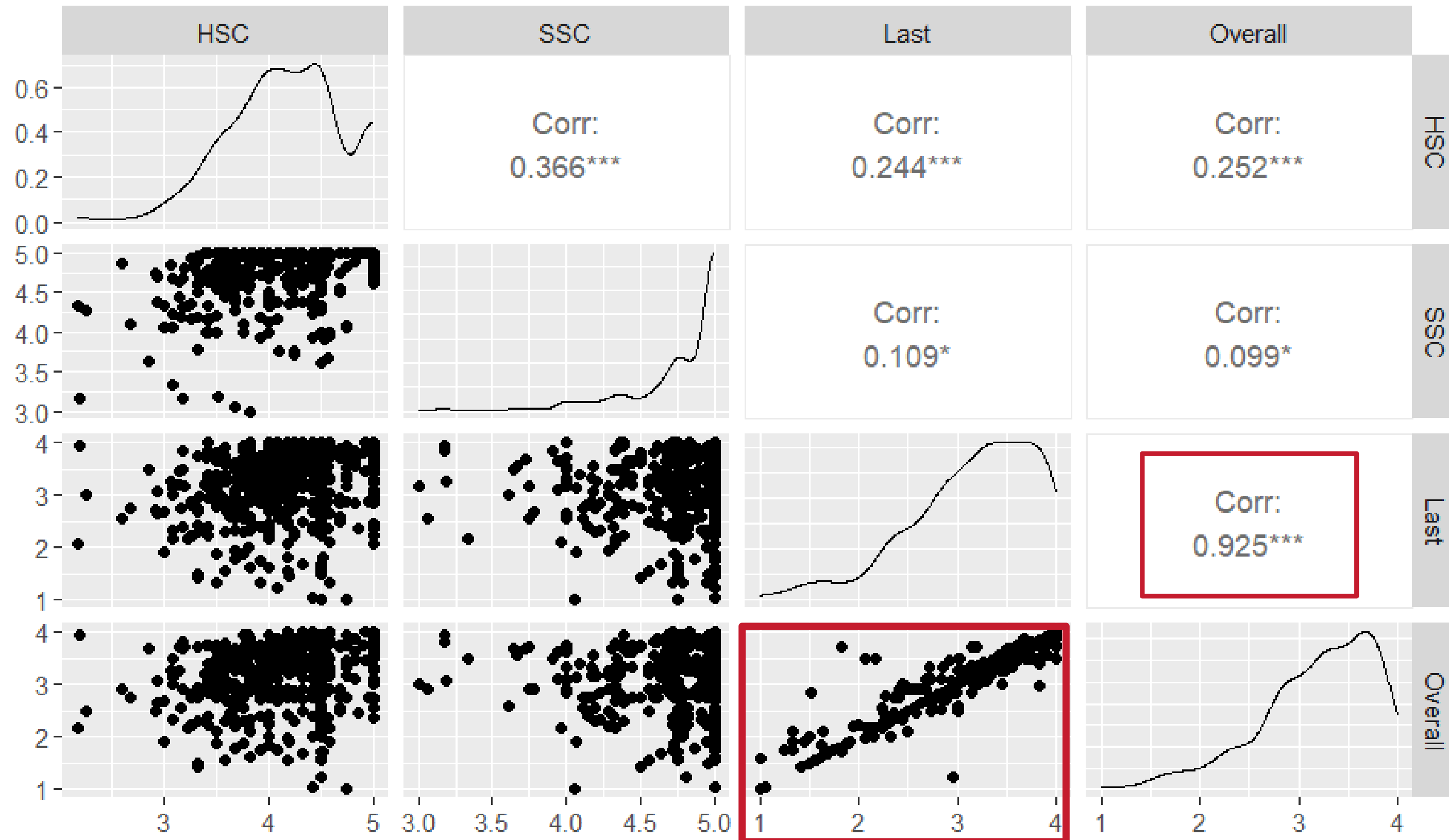
80%-100%

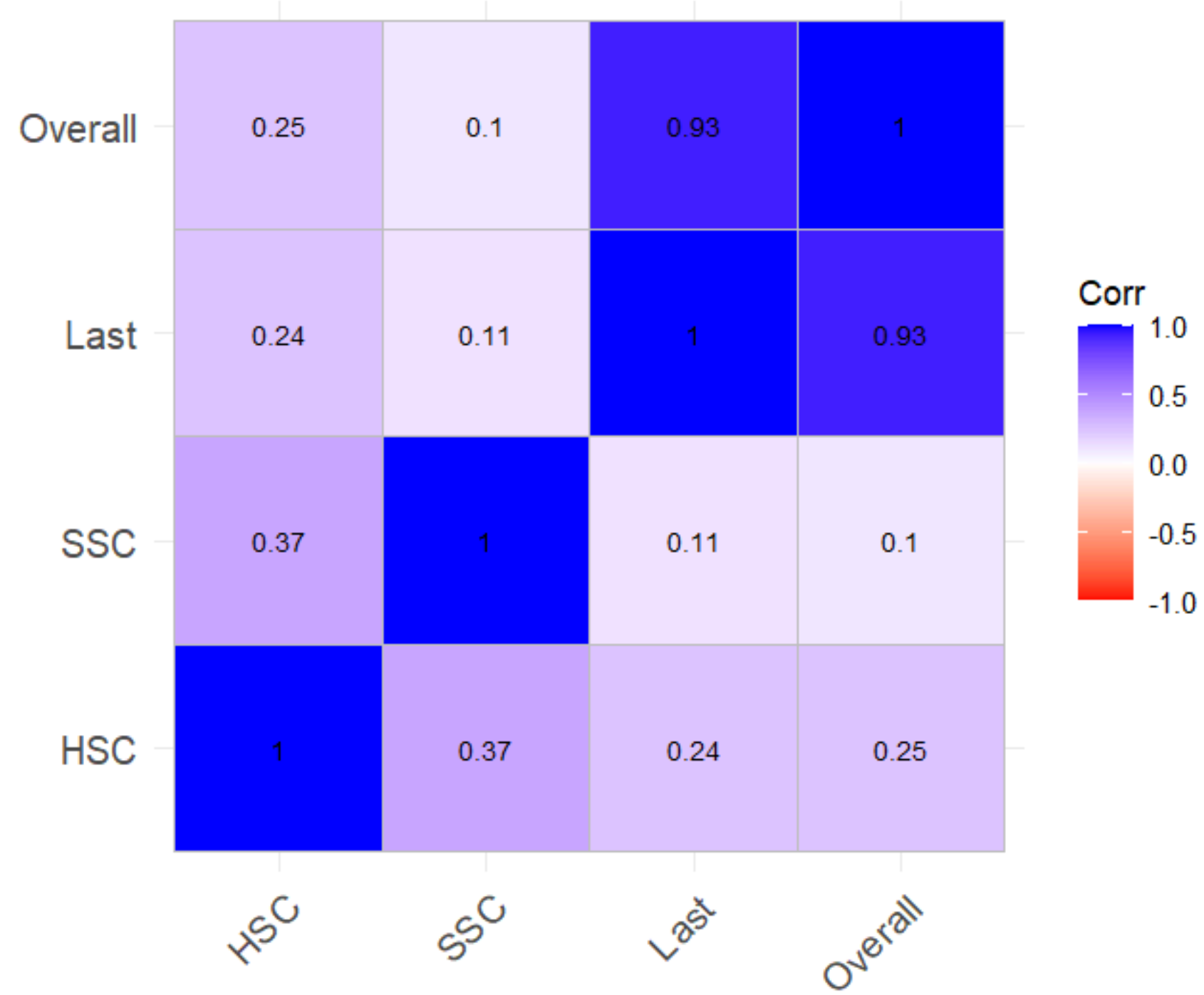
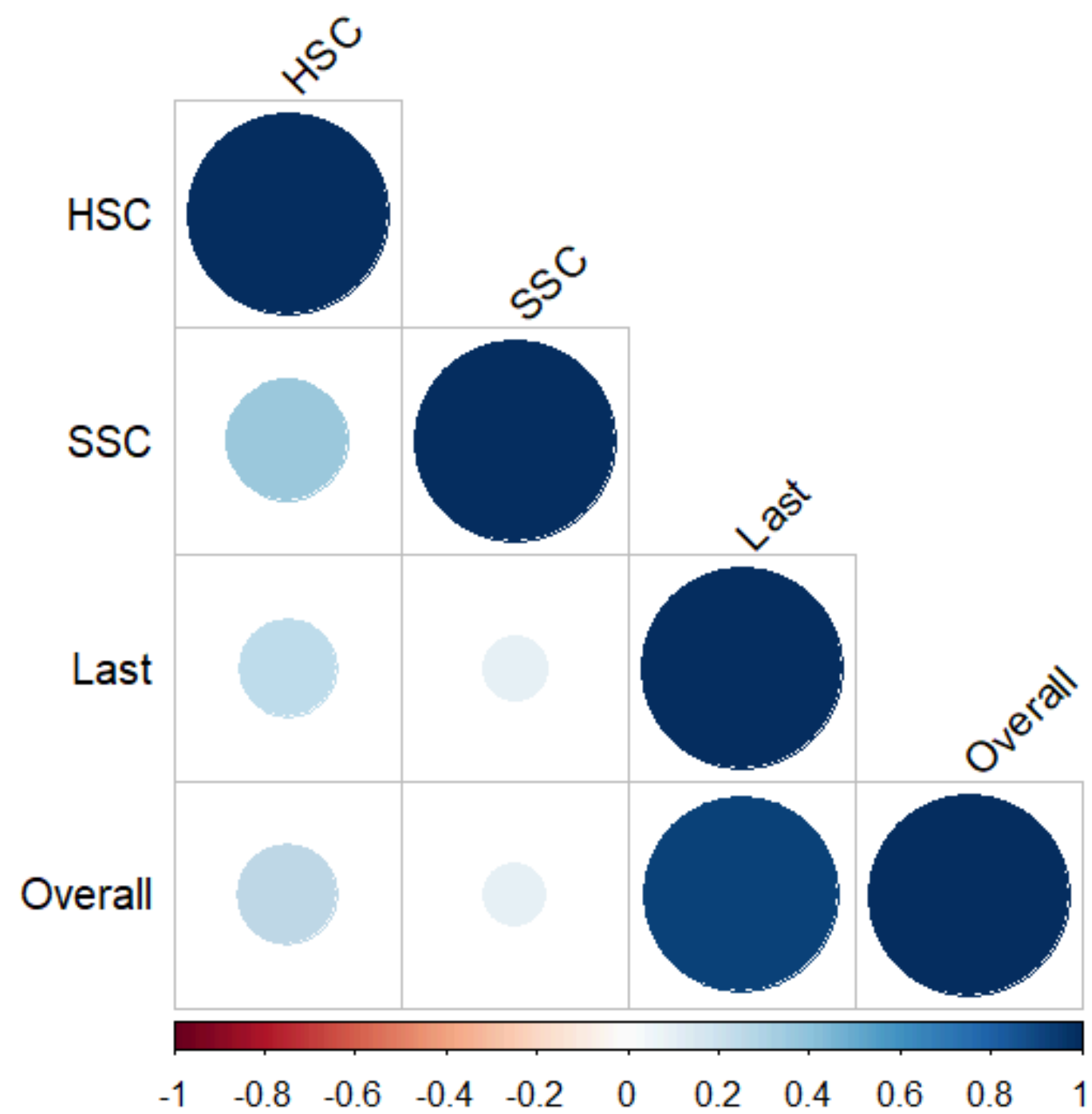
% Obecności



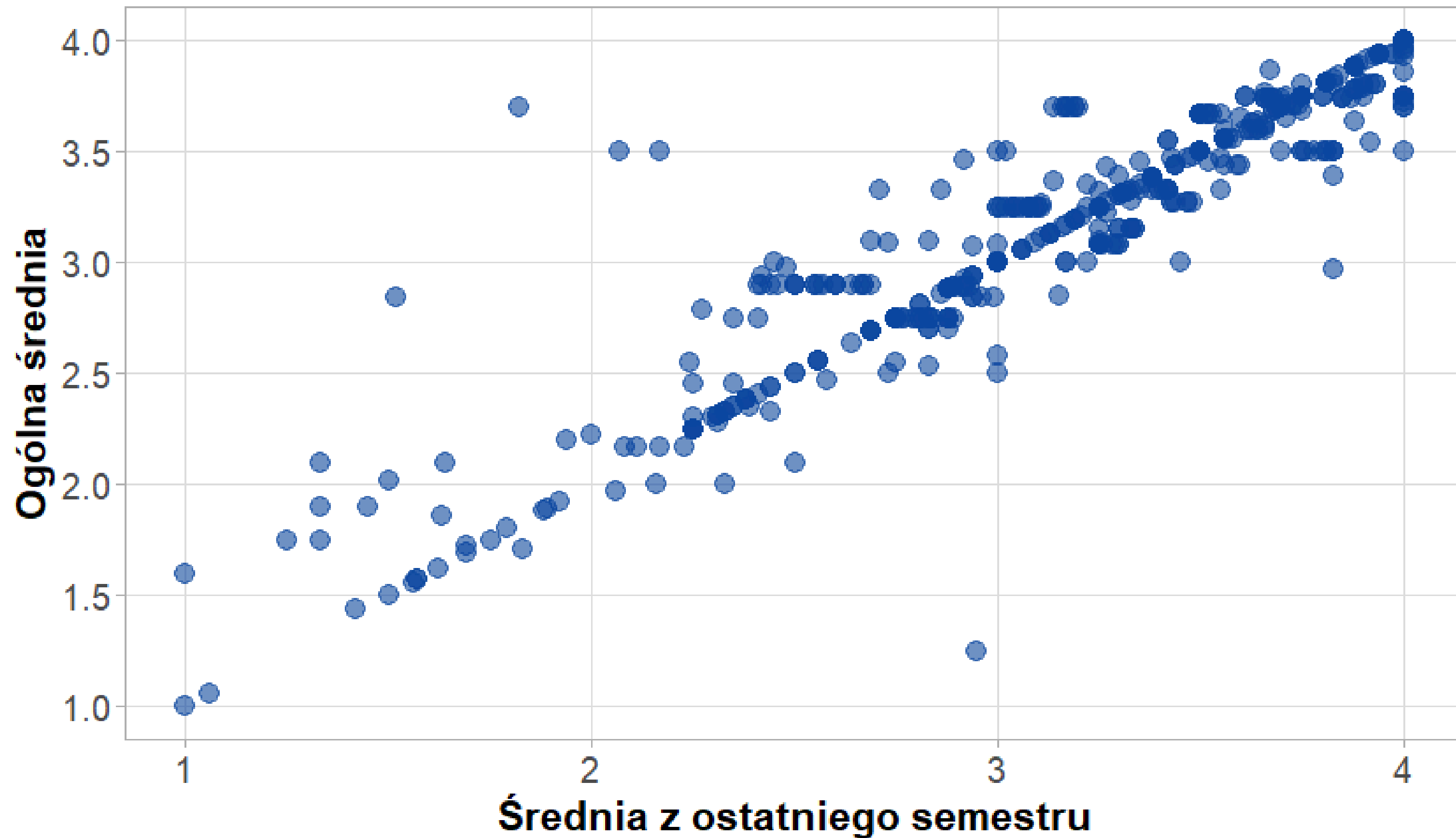
ZMIENNE NUMERYCZNE

Macierz wykresów rozrzutu





Relacja między średnią z ostatniego semestru a ogólną średnią



RÓŻNE TYPY ZMIENNYCH



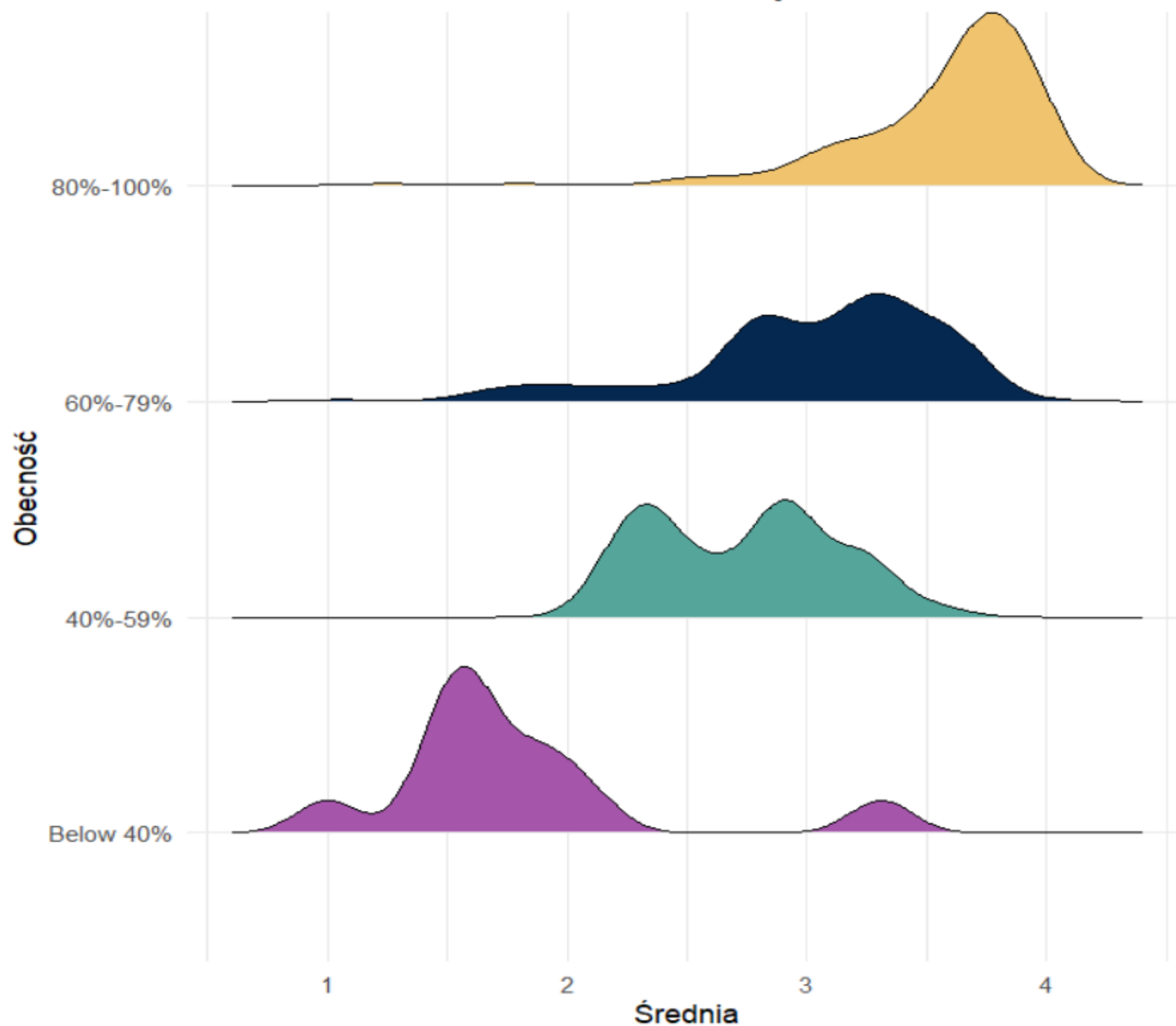
HIPOTEZA

Większa liczba godzin poświęconych na naukę i obecność na zajęciach jest skorelowana z lepszymi wynikami w testach.

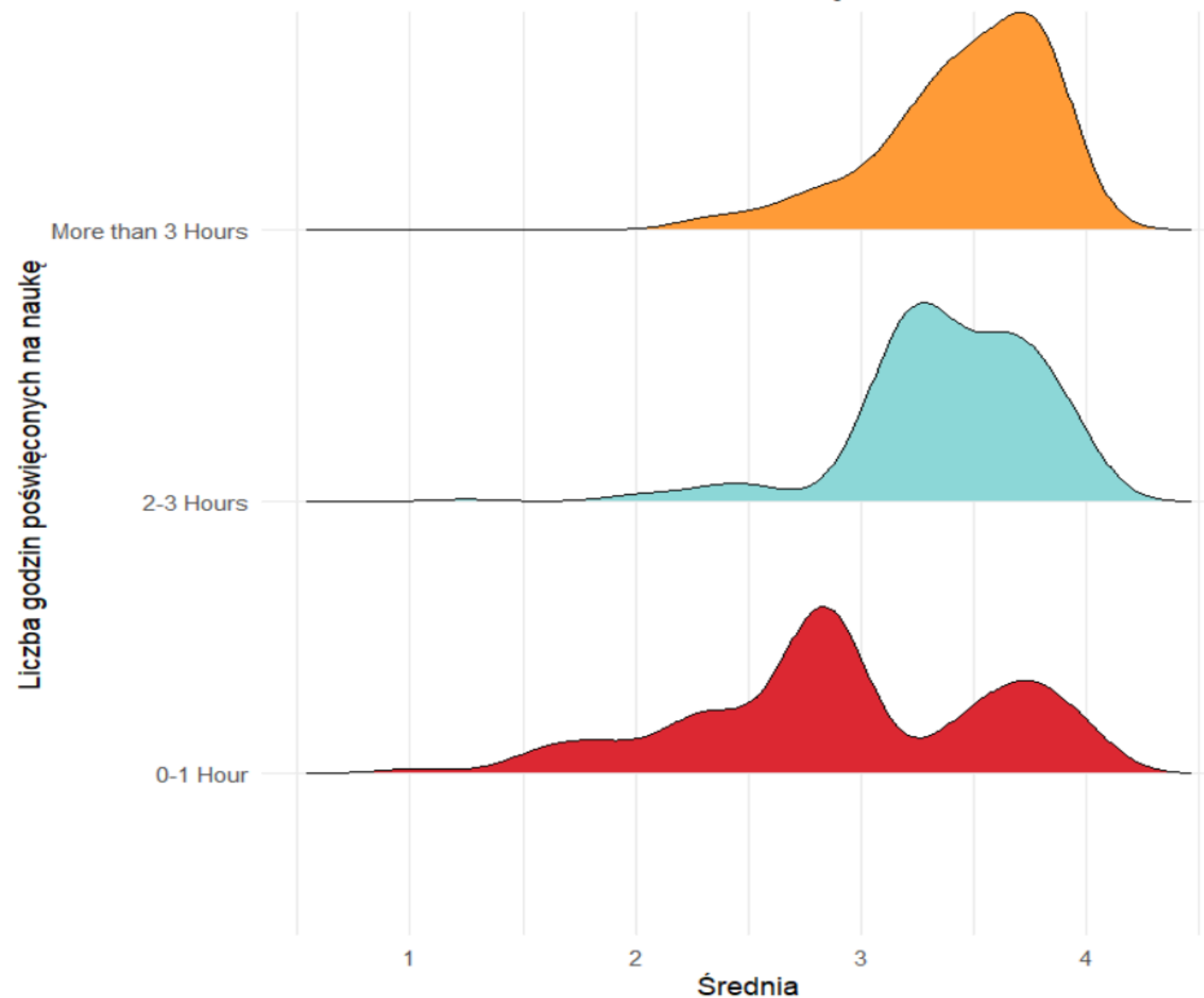
PROBLEM BADAWCZY

Jak regularna obecność na zajęciach przekłada się na wyniki akademickie?

Wpływ obecności na średnią



Wpływ liczby godzin nauki na średnią



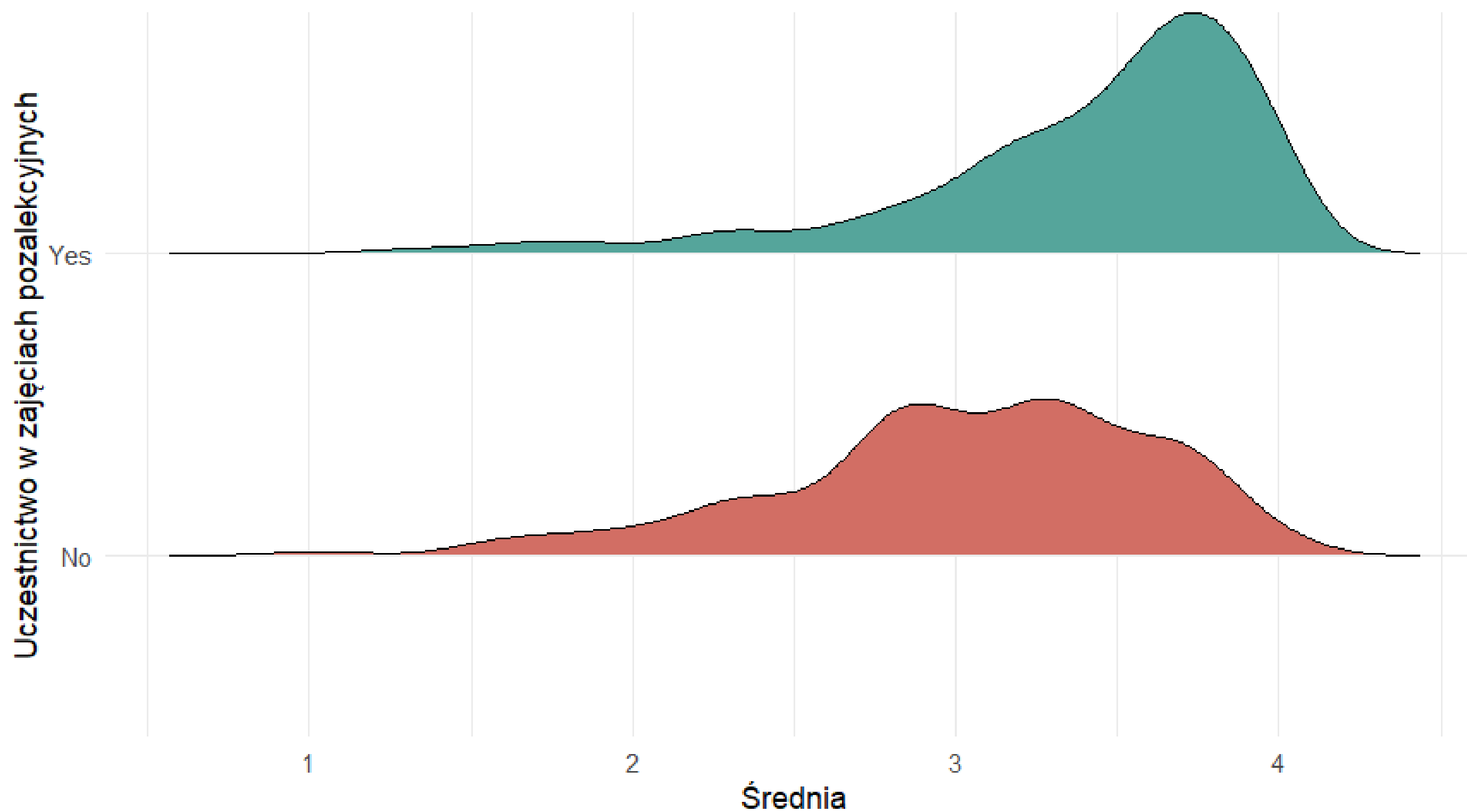
HIPOTEZA

Studenci angażujący się w aktywności pozalekcyjne osiągają lepsze wyniki akademickie niż ci, którzy nie biorą udziału w takich zajęciach.

PROBLEM BADAWCZY

Czy aktywności pozalekcyjne mają pozytywny wpływ na wyniki akademickie studentów?

Wpływ zajęć pozalekcyjnych na średnią



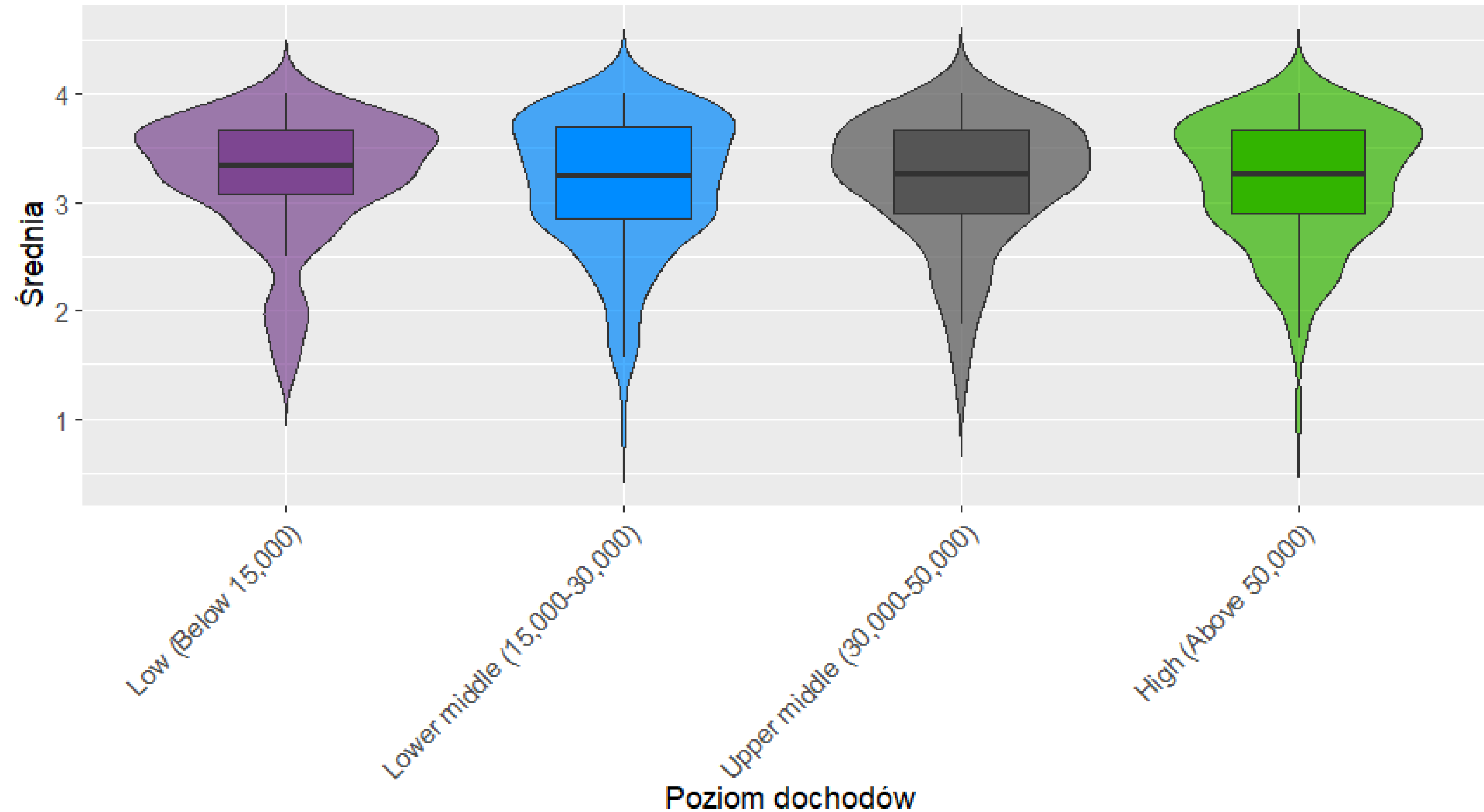
HIPOTEZA

Studenci z wyższym statusie społeczno-ekonomicznym osiągają lepsze wyniki akademickie niż studenci z niższym statusem.

PROBLEM BADAWCZY

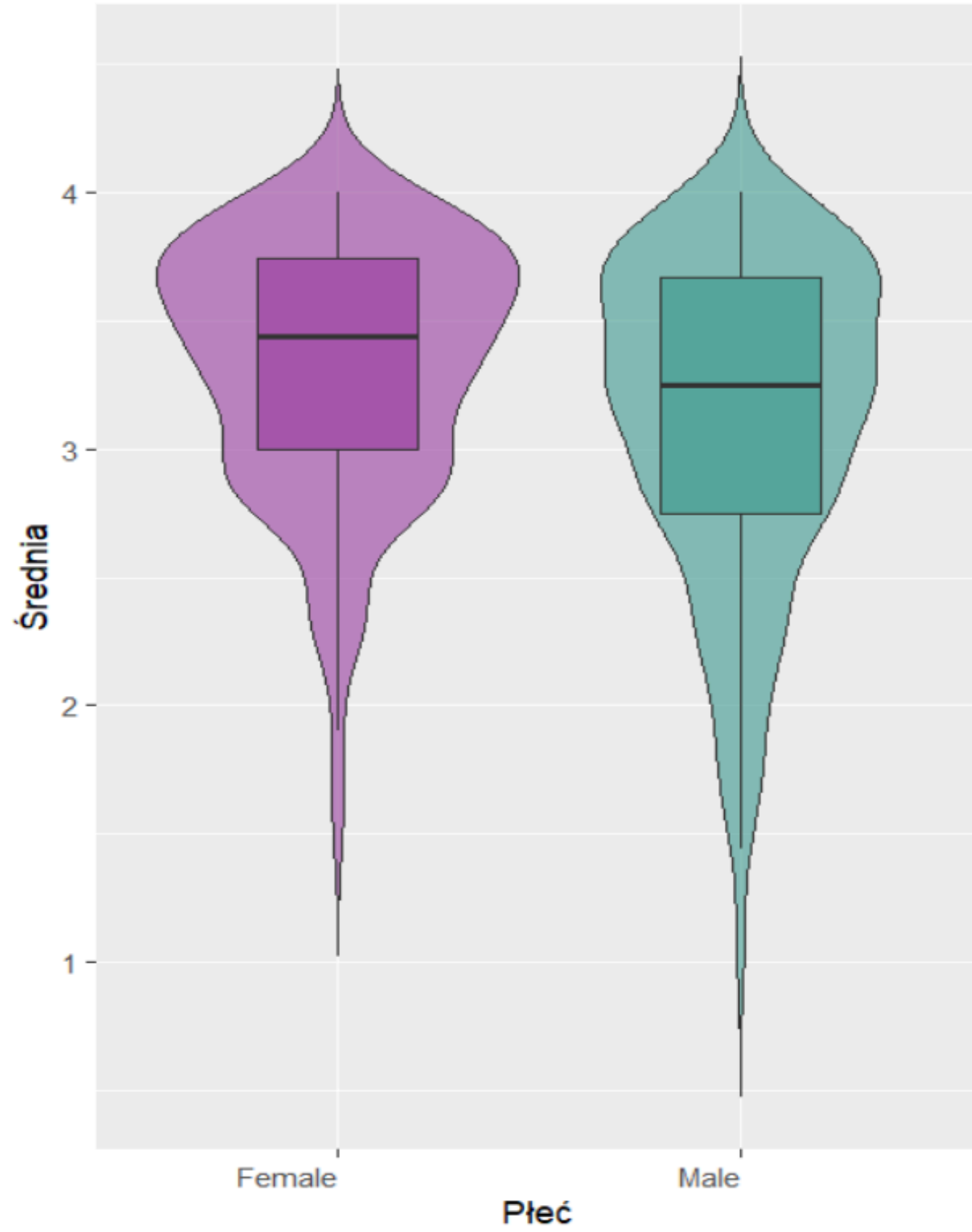
Jak czynniki demograficzne oraz społeczno-ekonomiczne wpływają na osiągnięcia edukacyjne?

Wpływ zamożności studenta a na jego wyniki w nauce

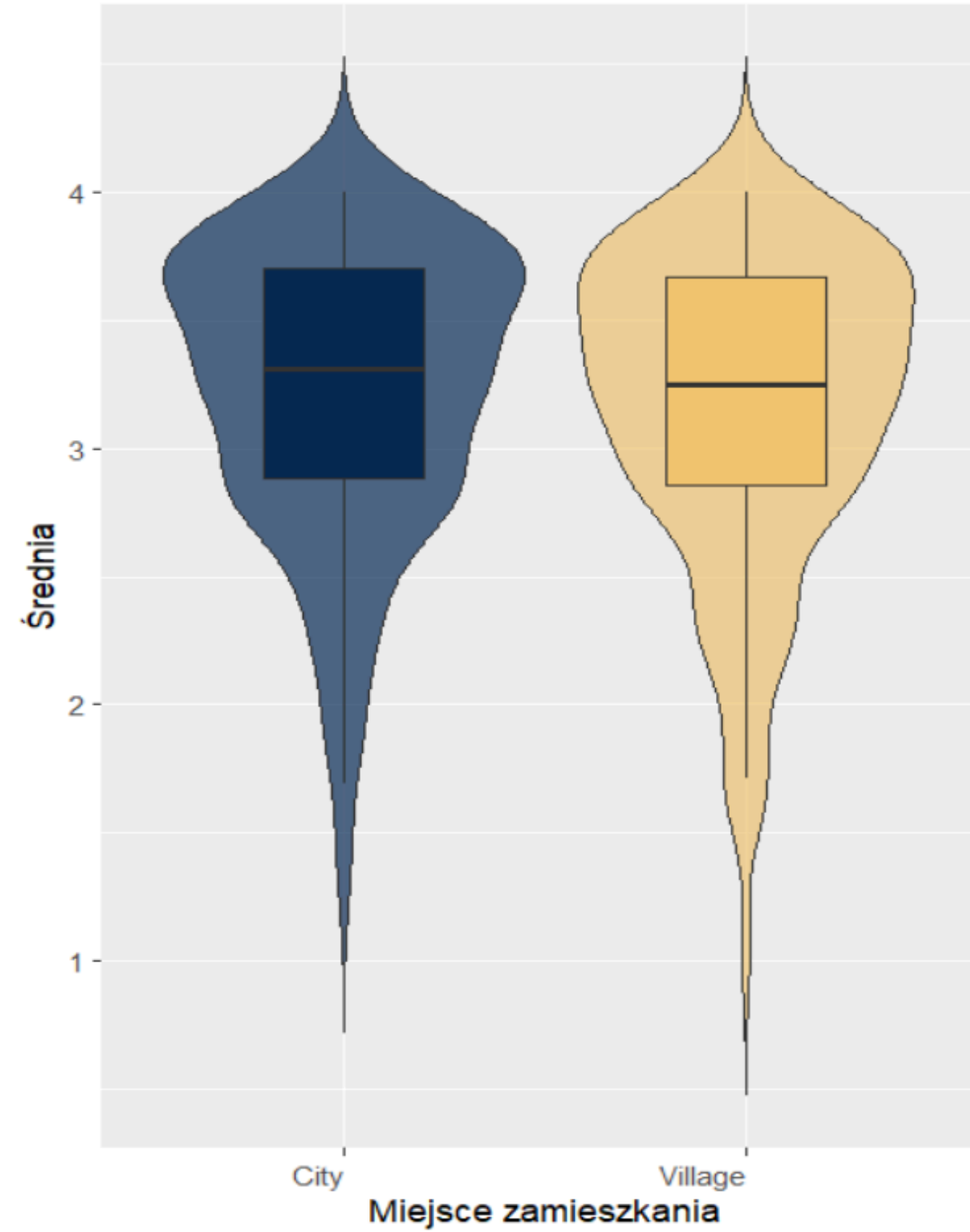


PROBLEM BADAWCZY

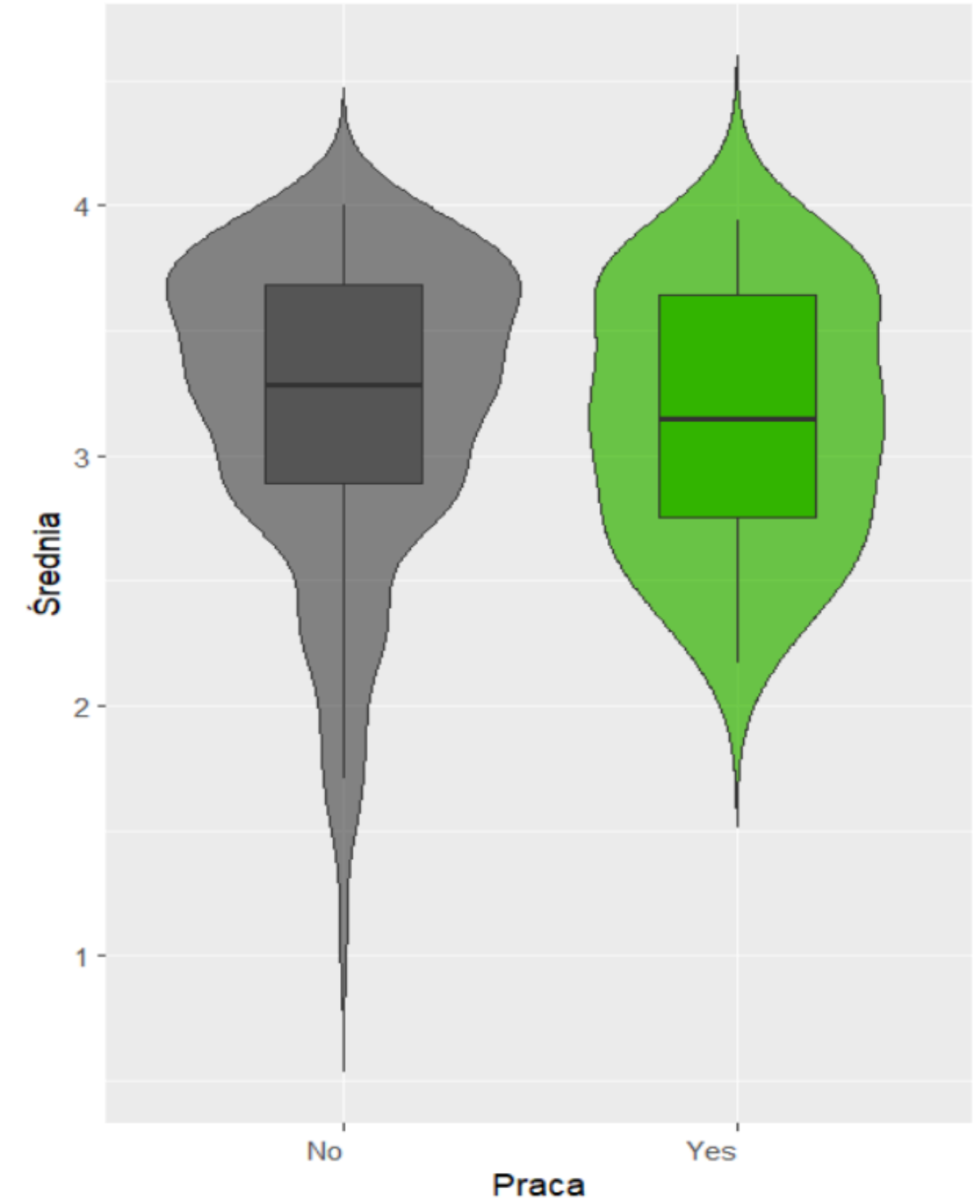
Wpływ płci
a na wyniki w nauce



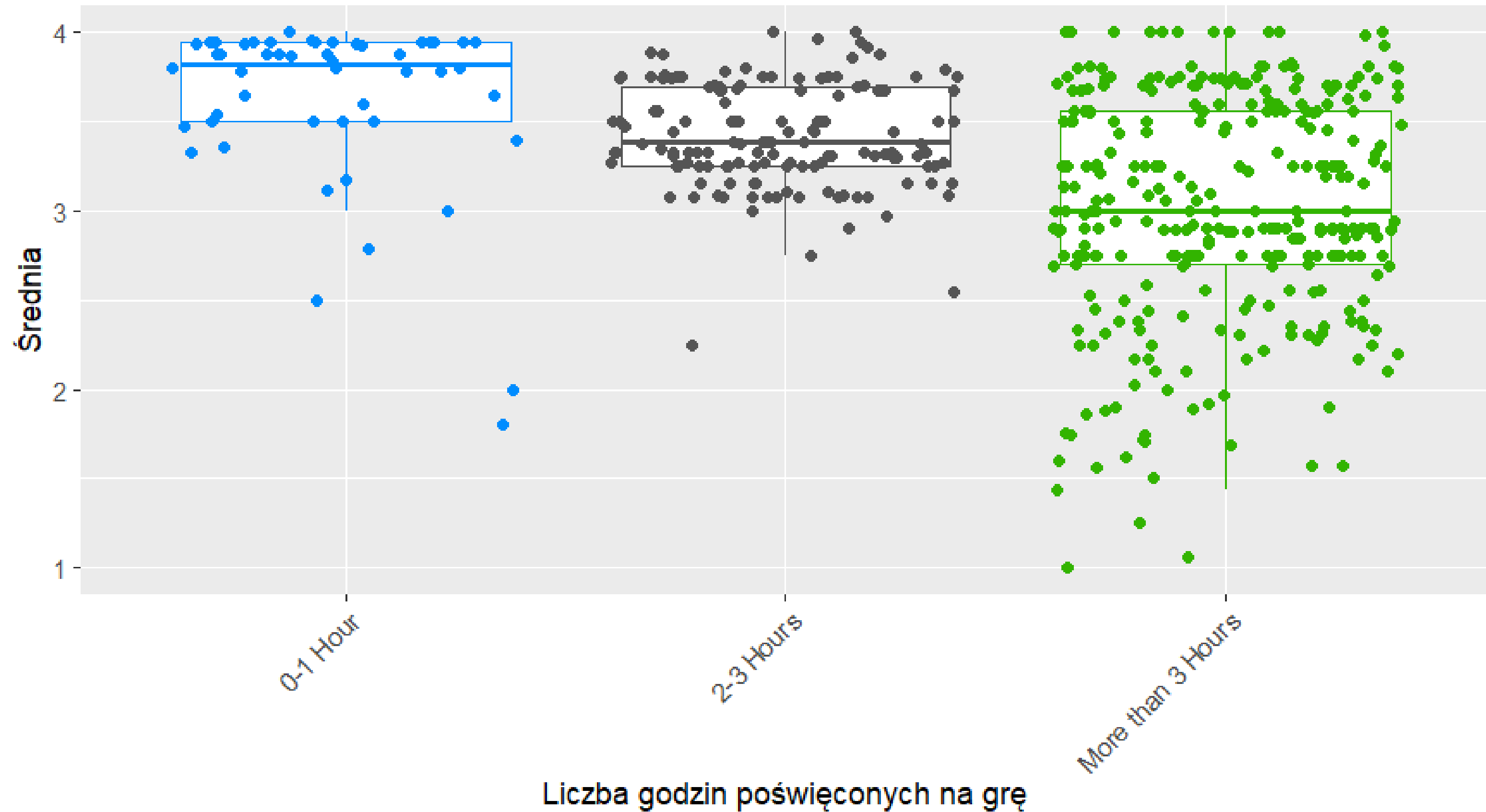
Wpływ miejsca zamieszkania
a na wyniki w nauce



Wpływ posiadania pracy
a na wyniki w nauce



Wpływ liczby godzin grania a na wyniki w nauce



SYLWETKA IDEALNEGO KANDYDATA?

*wskazówki dla placówek niepublicznych na
jakie czynniki zwracać uwagę chcąc kształcić
jednostki z największym potencjałem
naukowym*



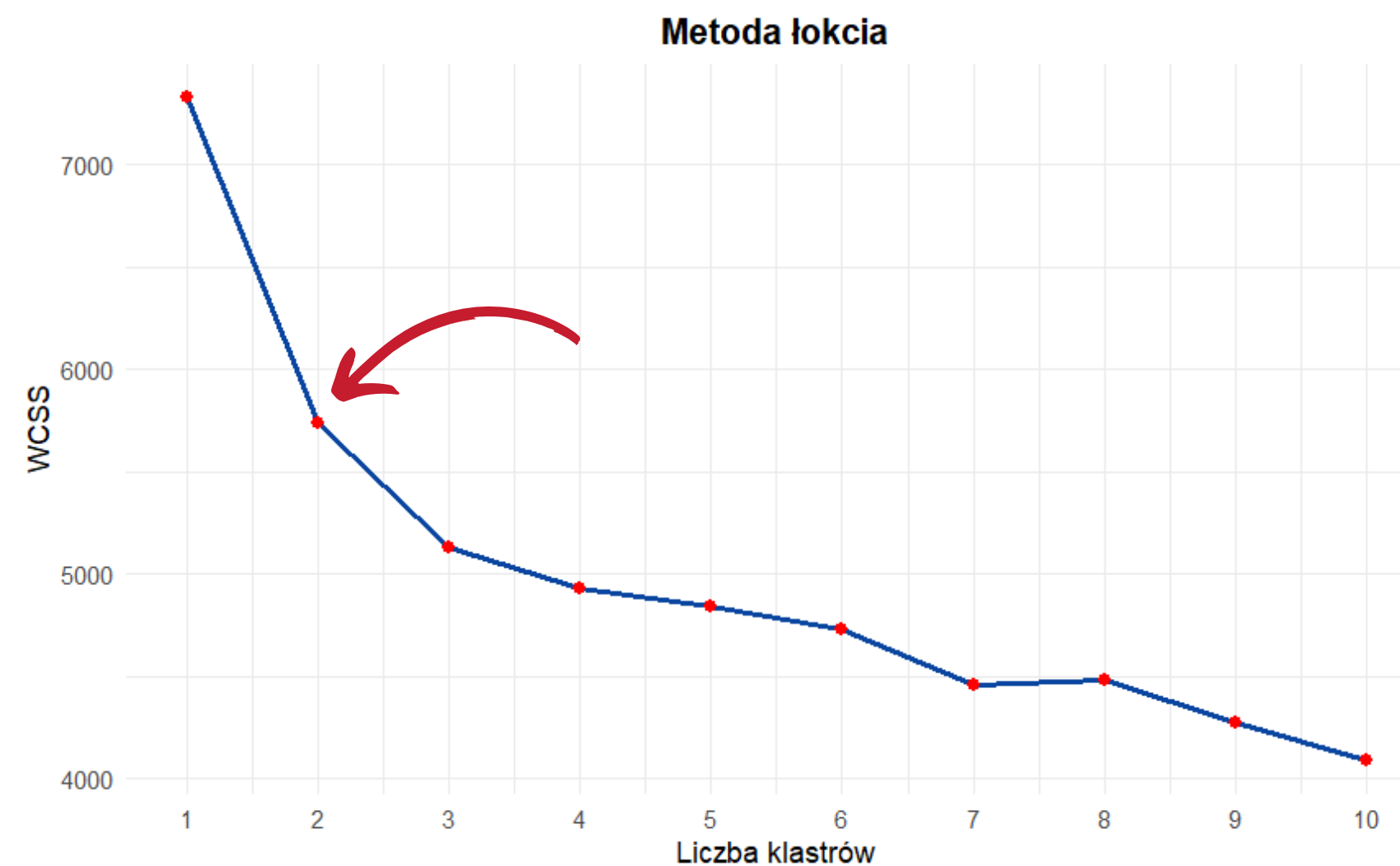
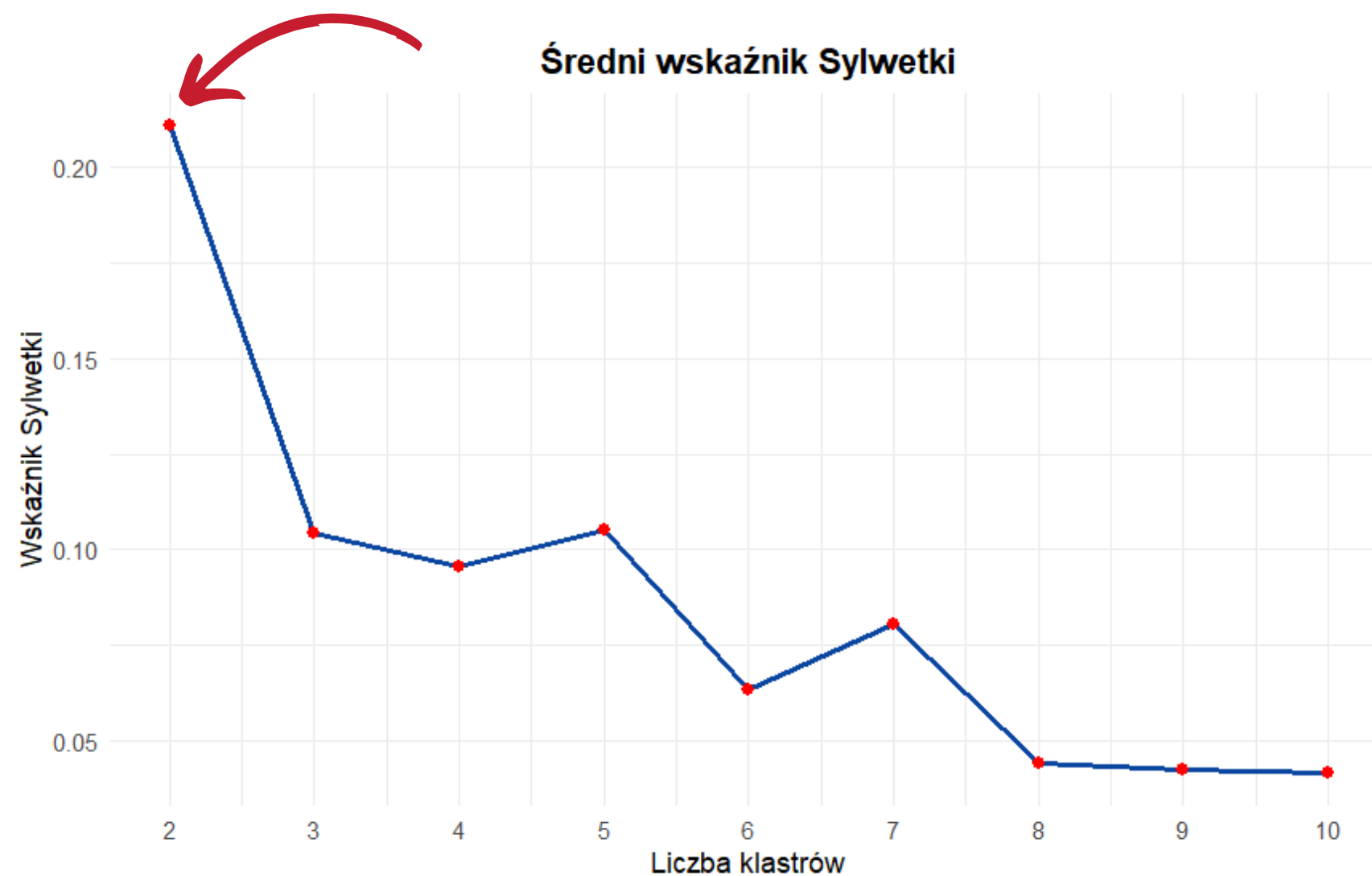
- Płeć: **kobieta**
- Miejsce zamieszkania: **miasto**
- Praca: **brak**
- Zajęcia dodatkowe: **tak**
- Gaming: **0-1 godzina**

```
best2 <- data %>%  
  filter(Gender == "Female",  
         Hometown == "City",  
         Job == "No",  
         Extra == "Yes", #wpływa na obecność, a ta na oceny  
         Gaming == "0-1 Hour") %>%  
  as.data.frame()  
  
mean(best2$Overall)
```

```
[1] 3.7597  
  
best2 | 10 obs.
```

PODZIAŁ DANYCH NA KLASTRY

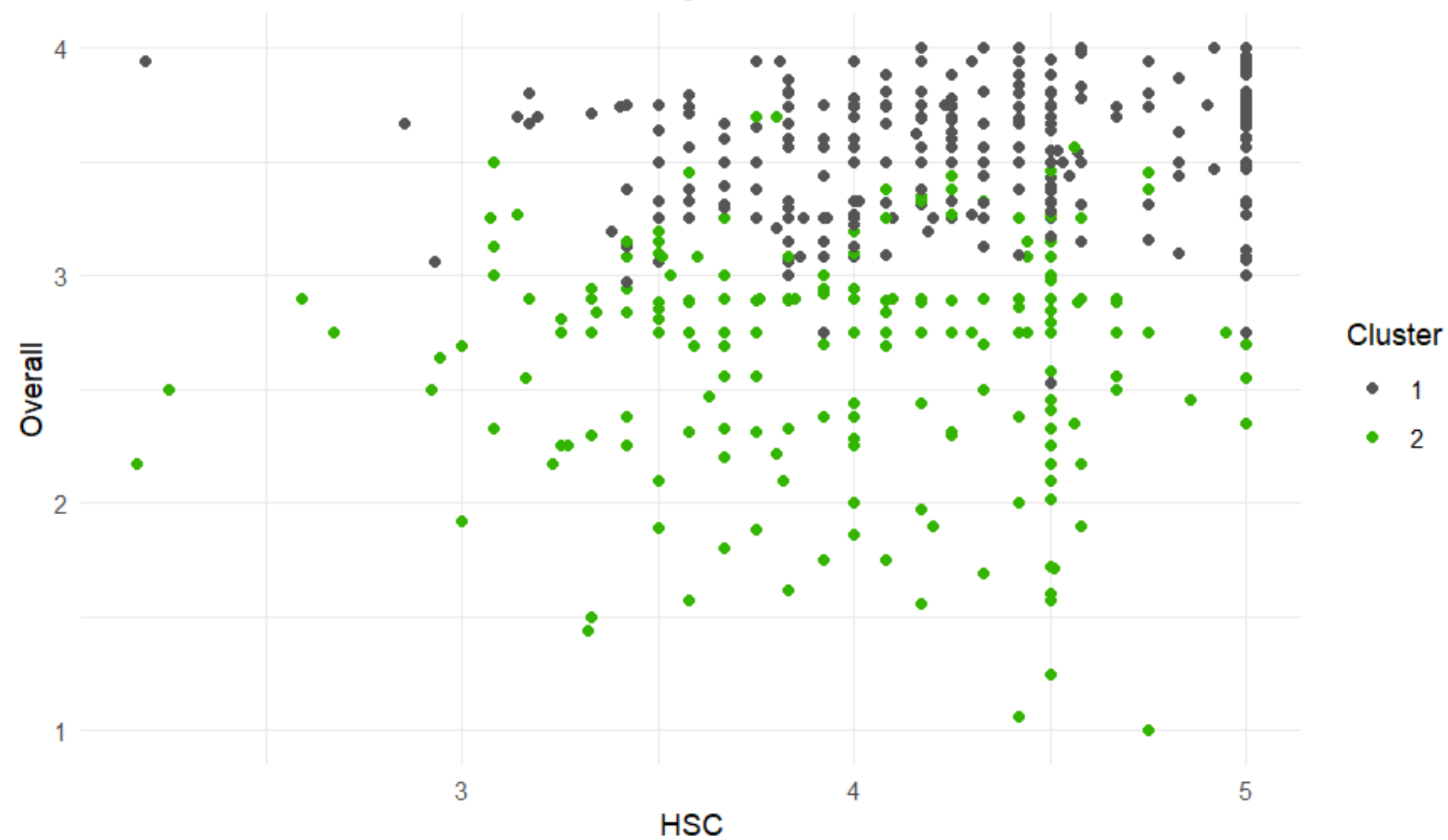
1 ZNALEZIENIE OPTYMALNEJ LICZBY KLASTRÓW



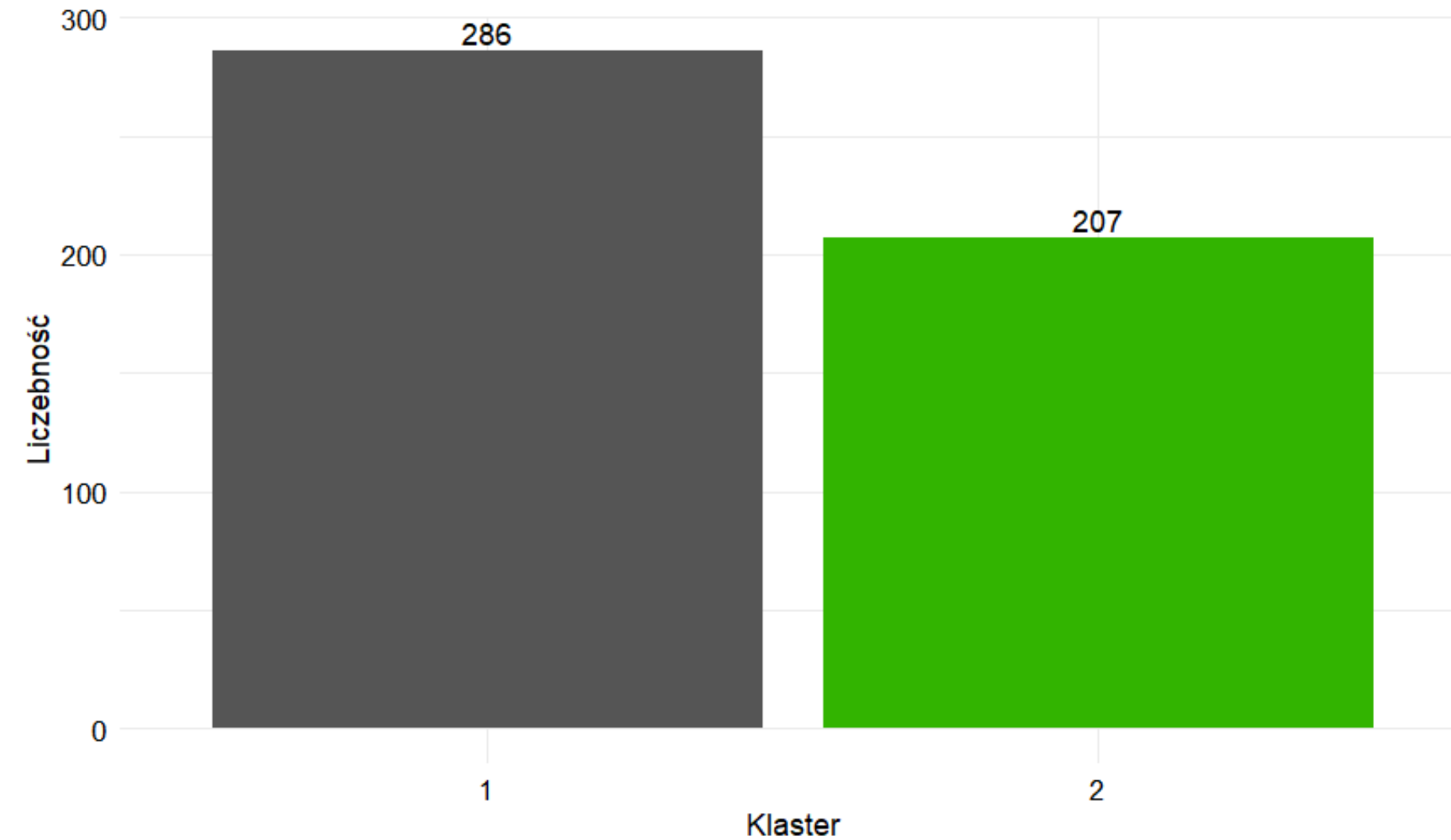
K=2

2 ALGORYTM K-PROTOTYPES

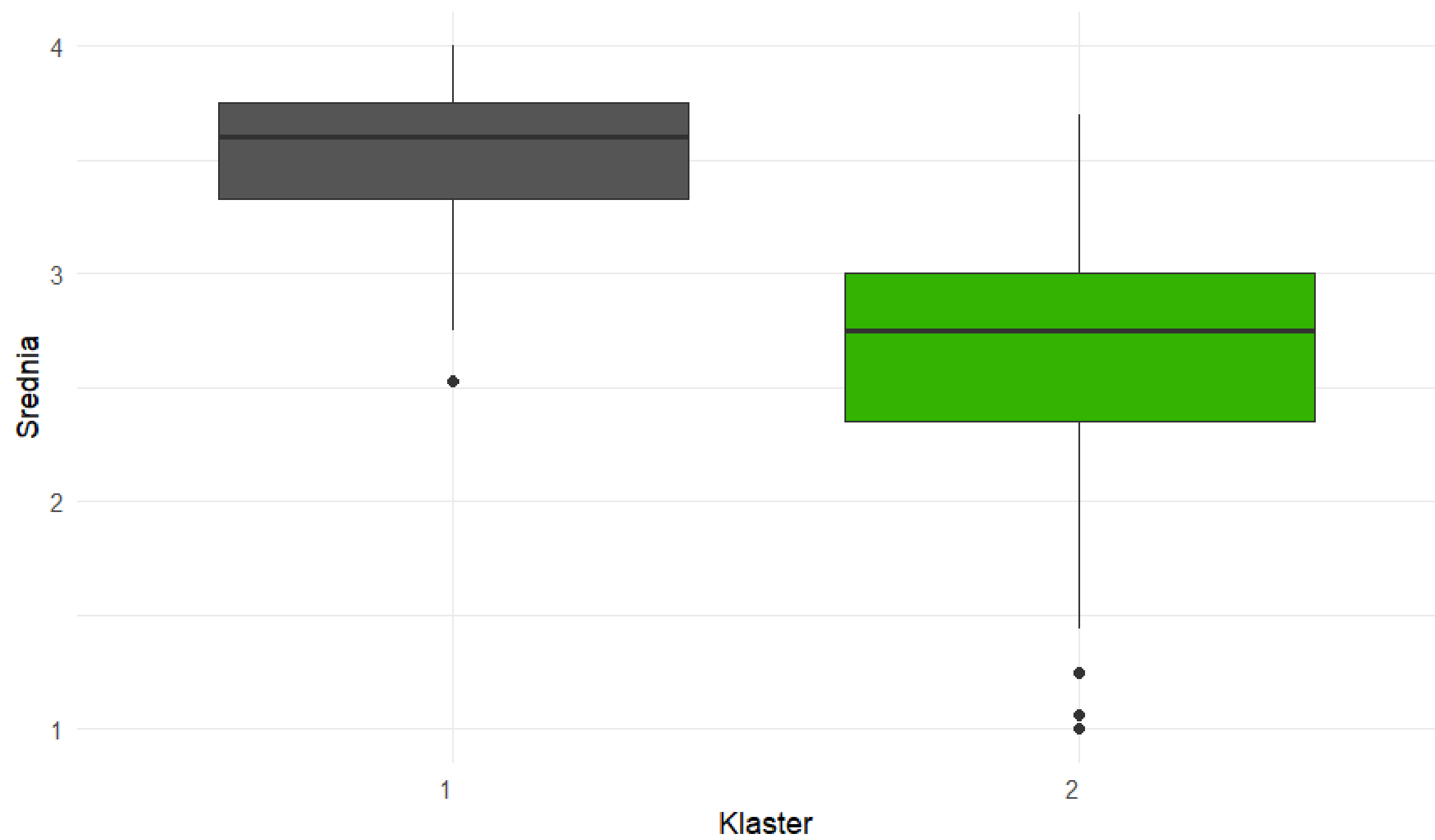
Wizualizacja klastrów



Liczebność klastrów

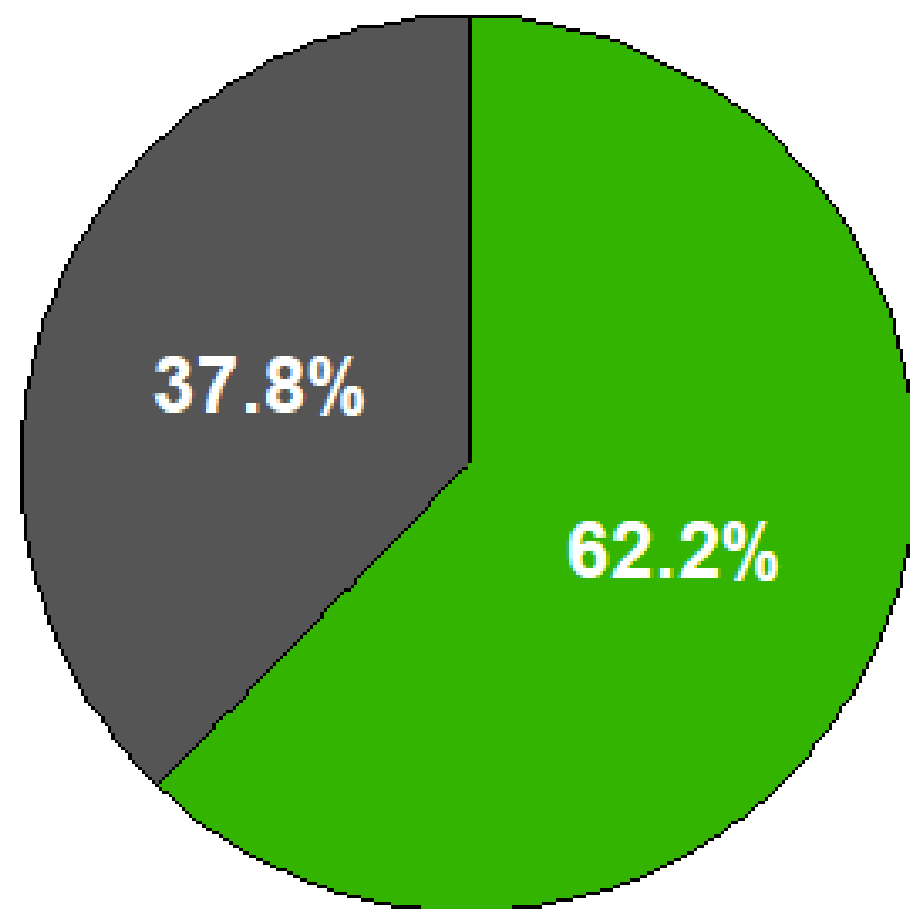


Rozkład średniej w poszczególnych klastrach



ANALIZA KLASTRÓW

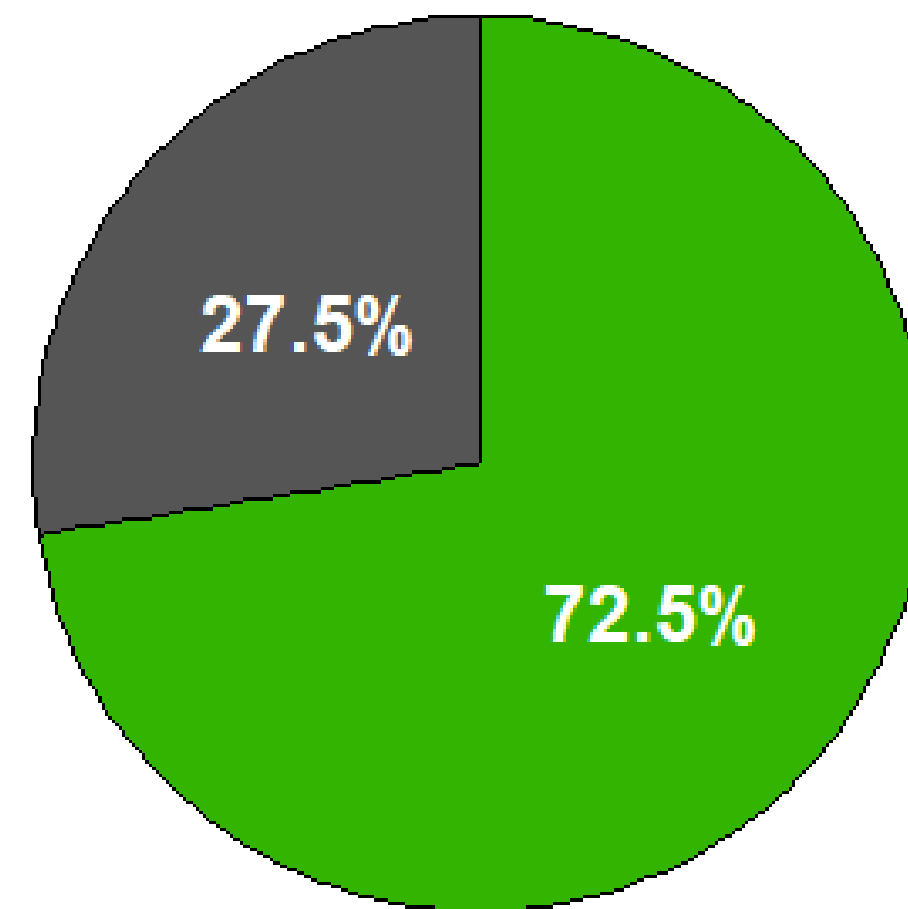
Rozkład płci w klastrze 1



Gender

- Female
- Male

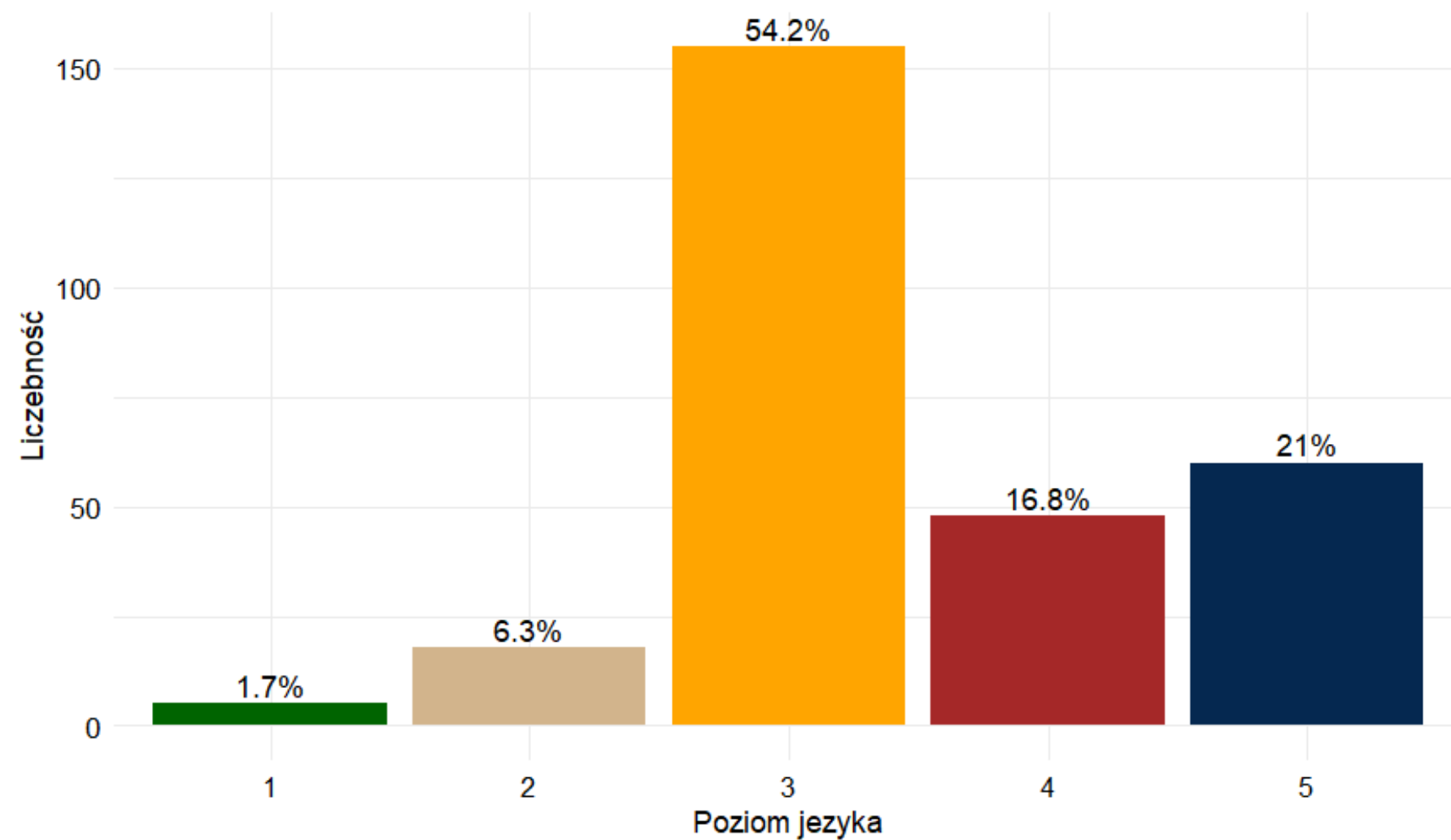
Rozkład płci w klastrze 2



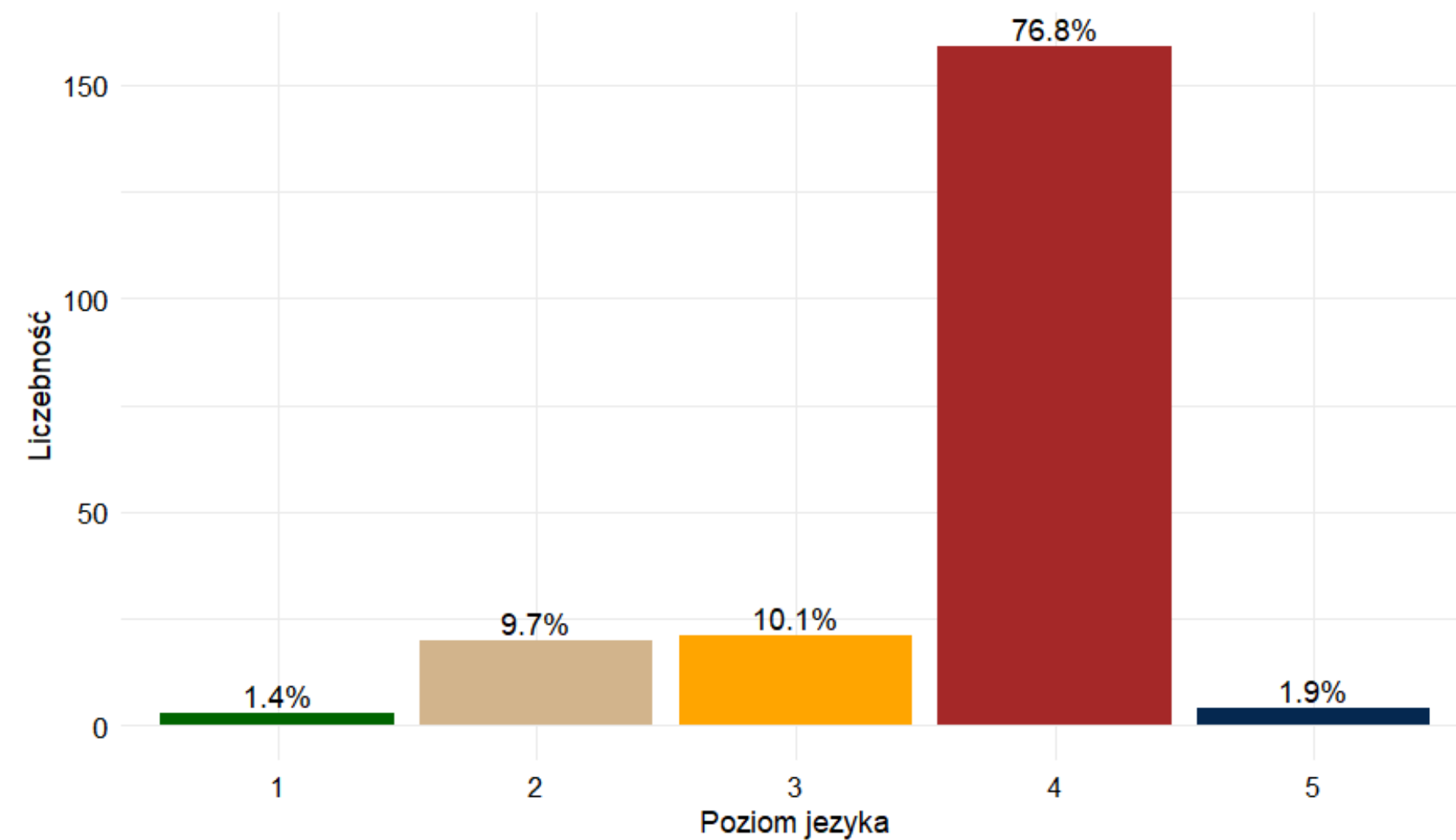
Gender

- Female
- Male

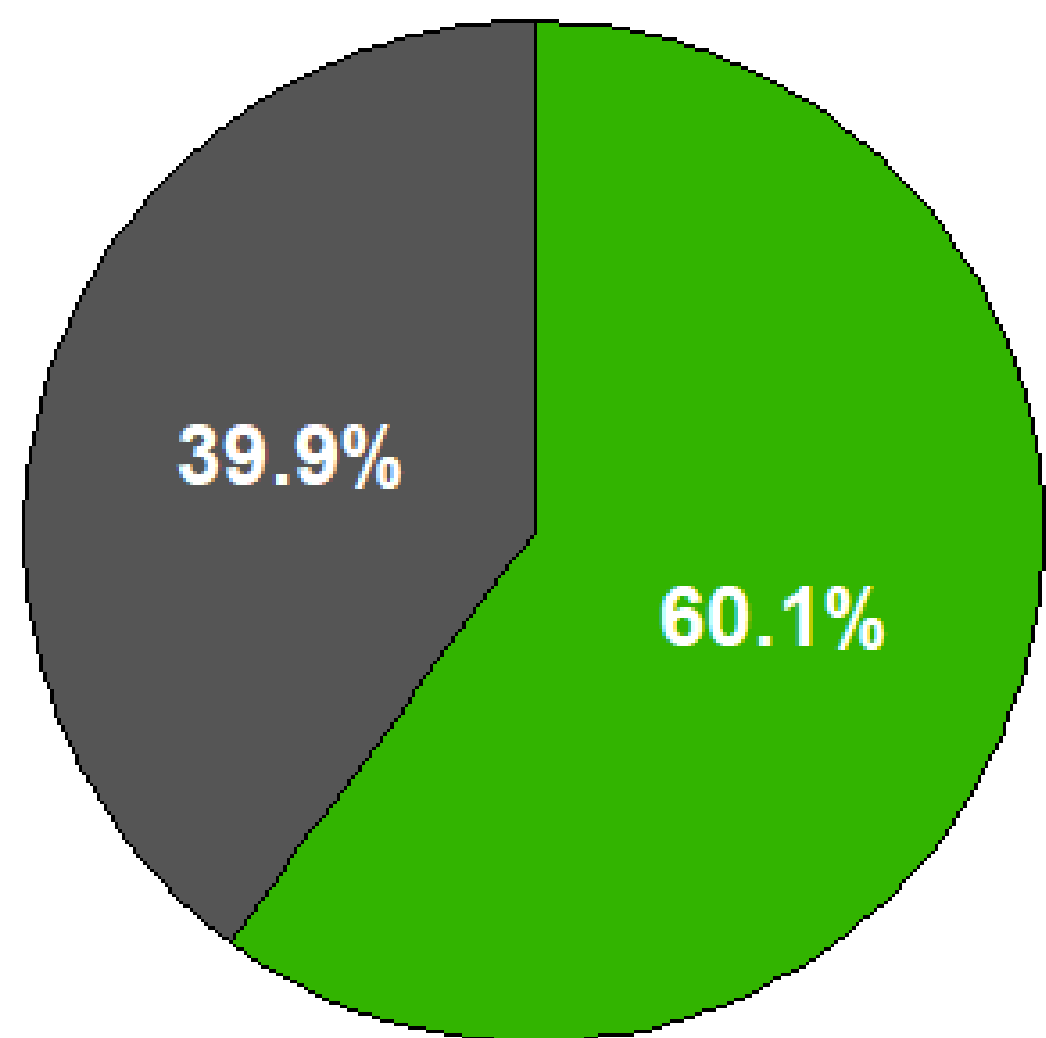
Znajomość języka angielskiego (klaster 1)



Znajomość języka angielskiego (klaster 2)

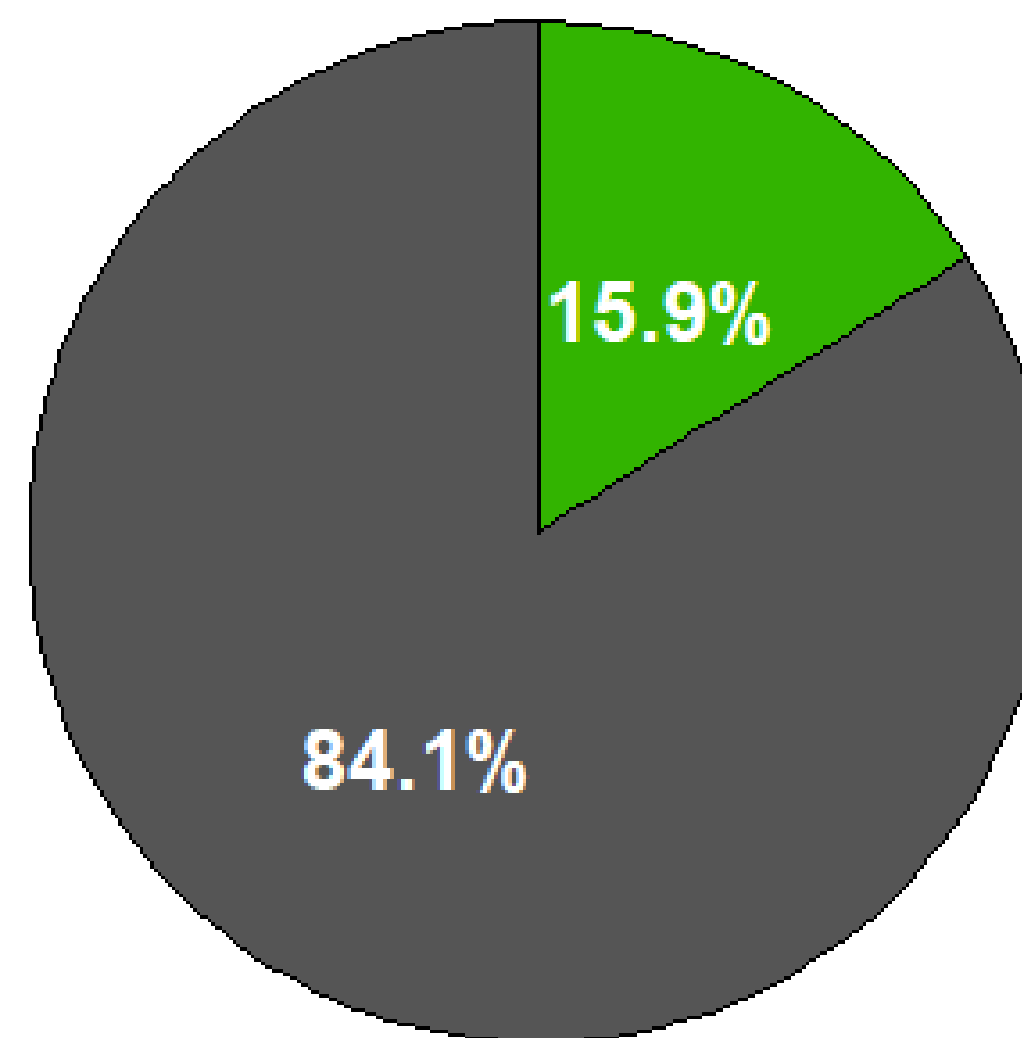


Zajęcia dodatkowe w klastrze 1



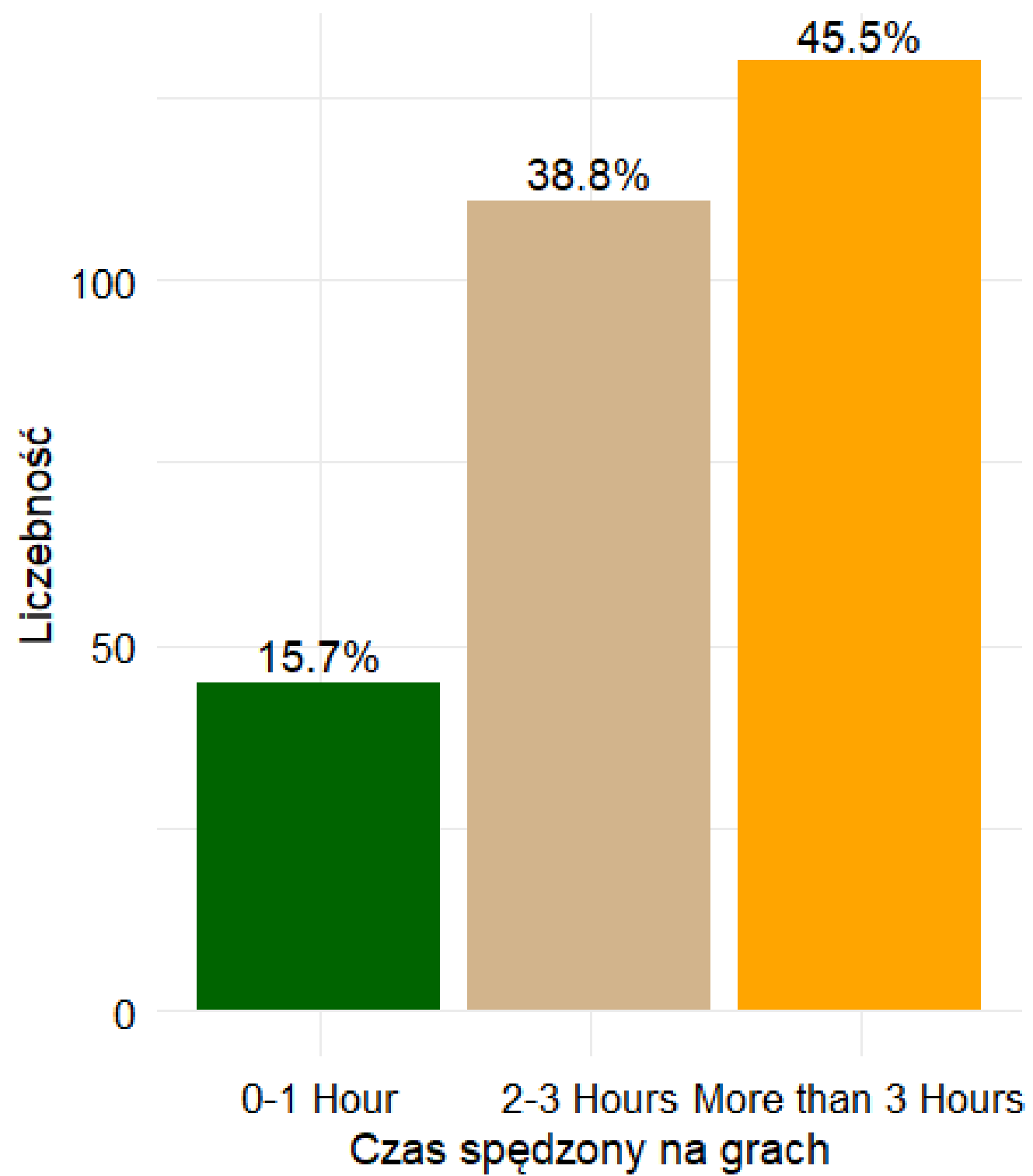
Extra
No
Yes

Zajęcia dodatkowe w klastrze 2

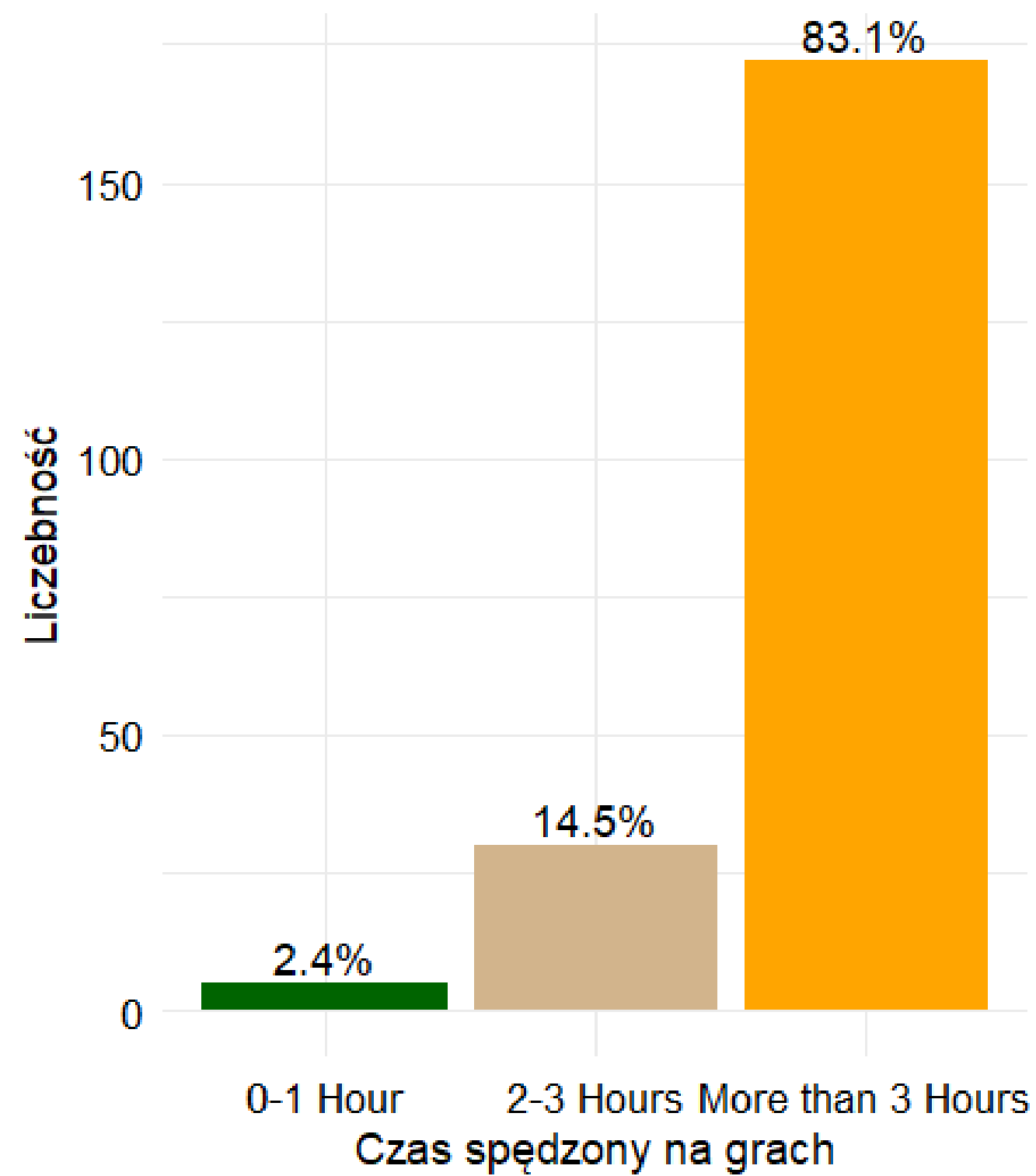


Extra
No
Yes

Czas spędzony na grach (klaster 1)



Czas spędzony na grach (klaster 2)



SYLWETKA IDEALNEGO KANDYDATA

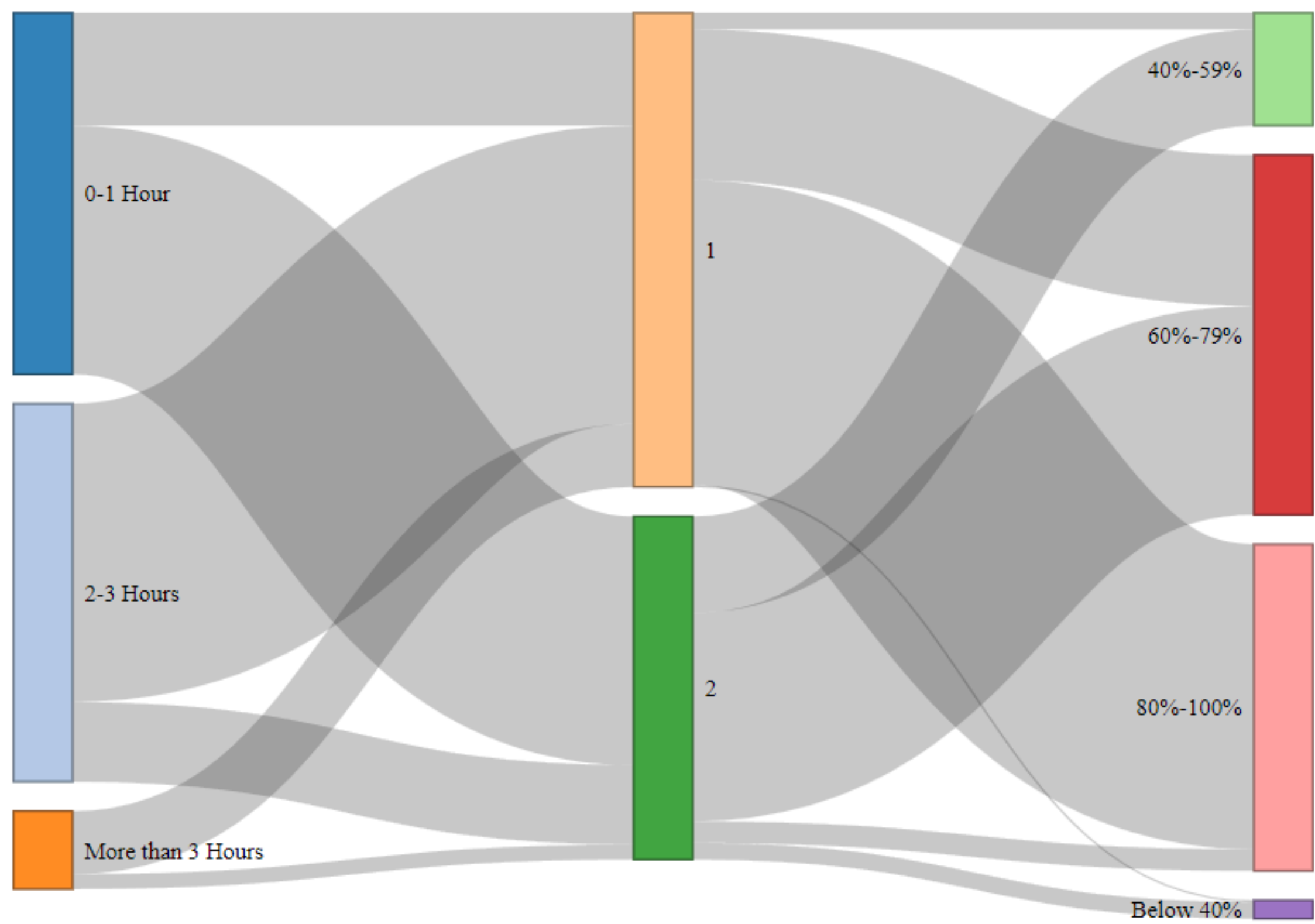
*wskazówki dla placówek niepublicznych na
jakie czynniki zwracać uwagę chcąc kształcić
jednostki z największym potencjałem
naukowym*

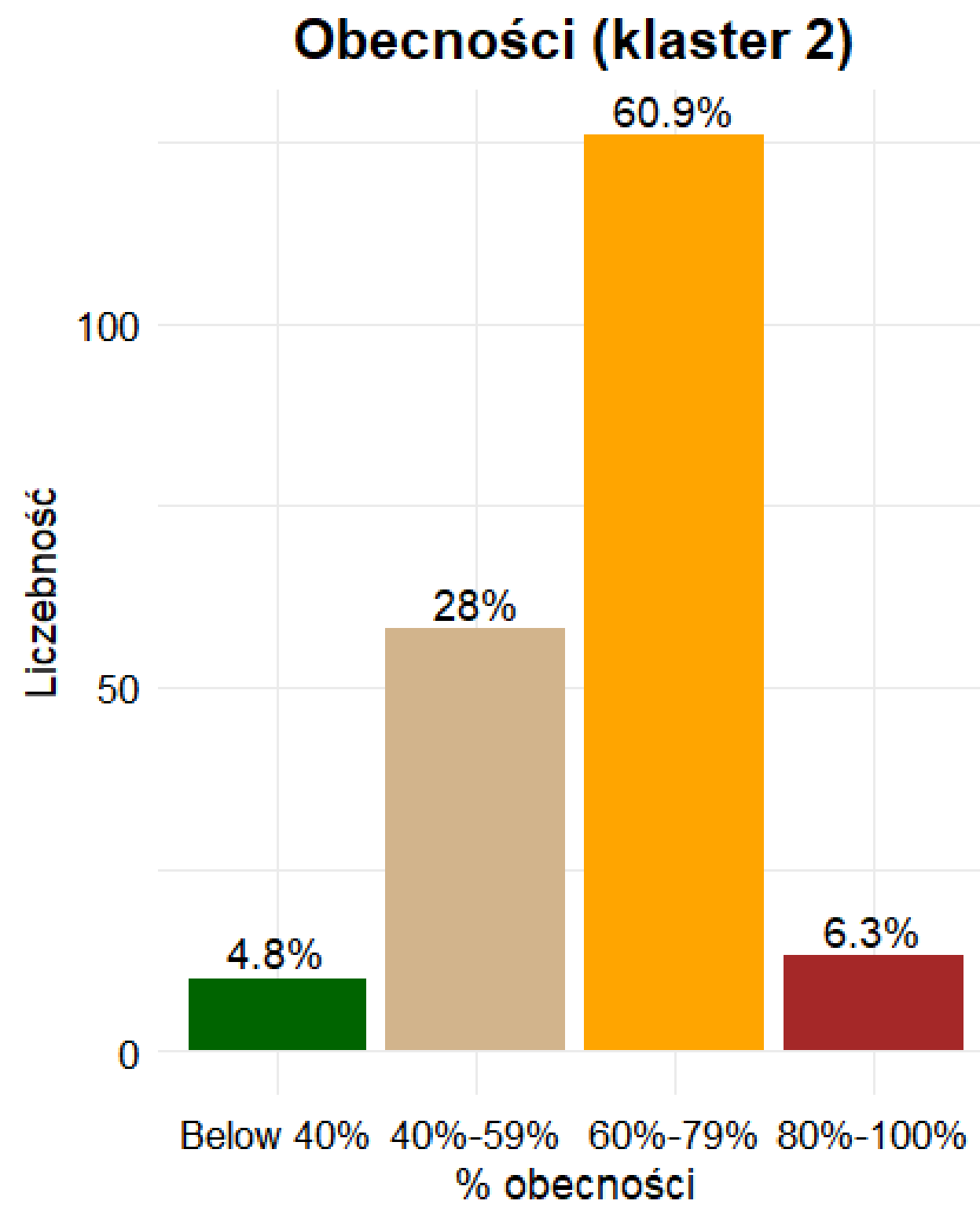
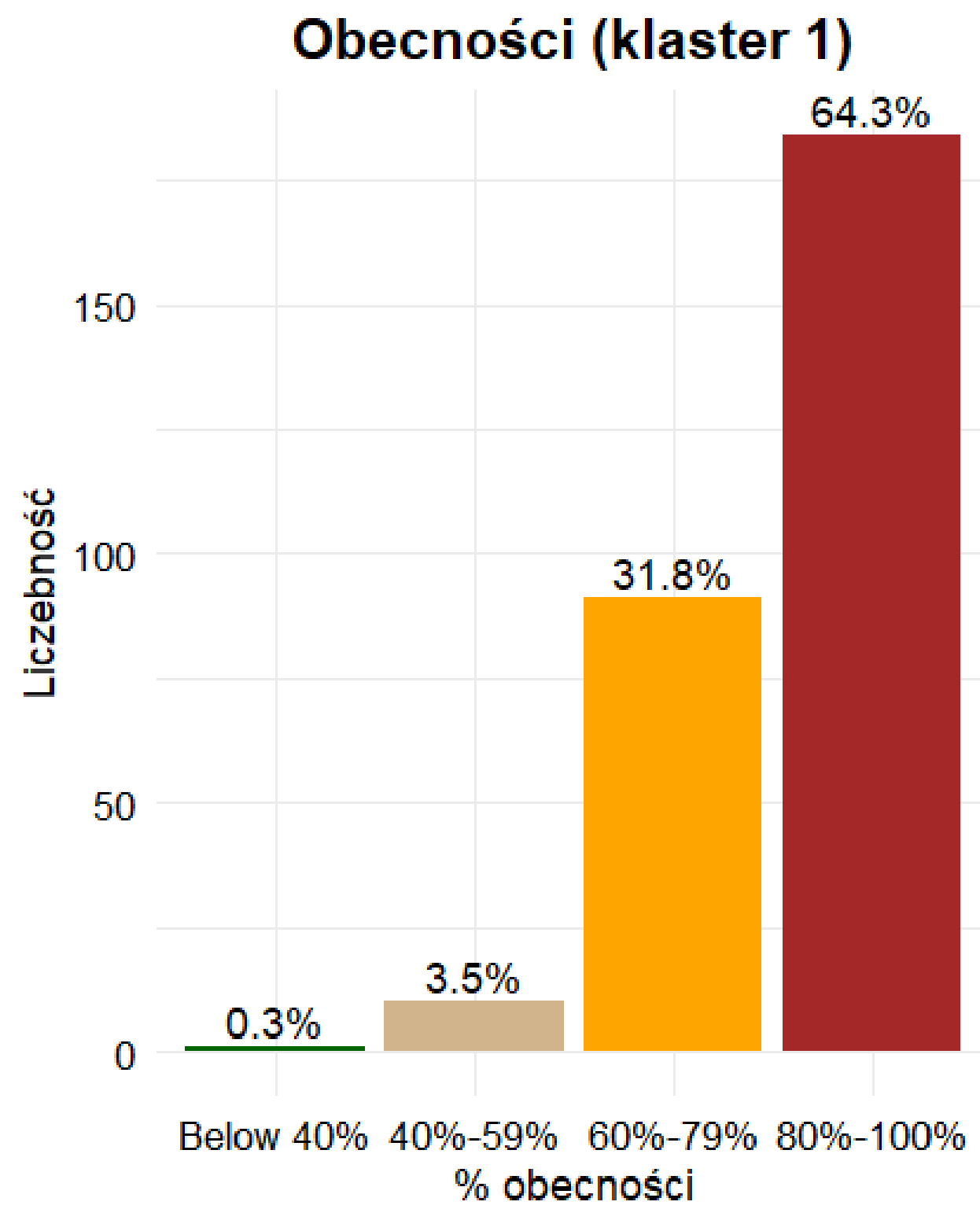


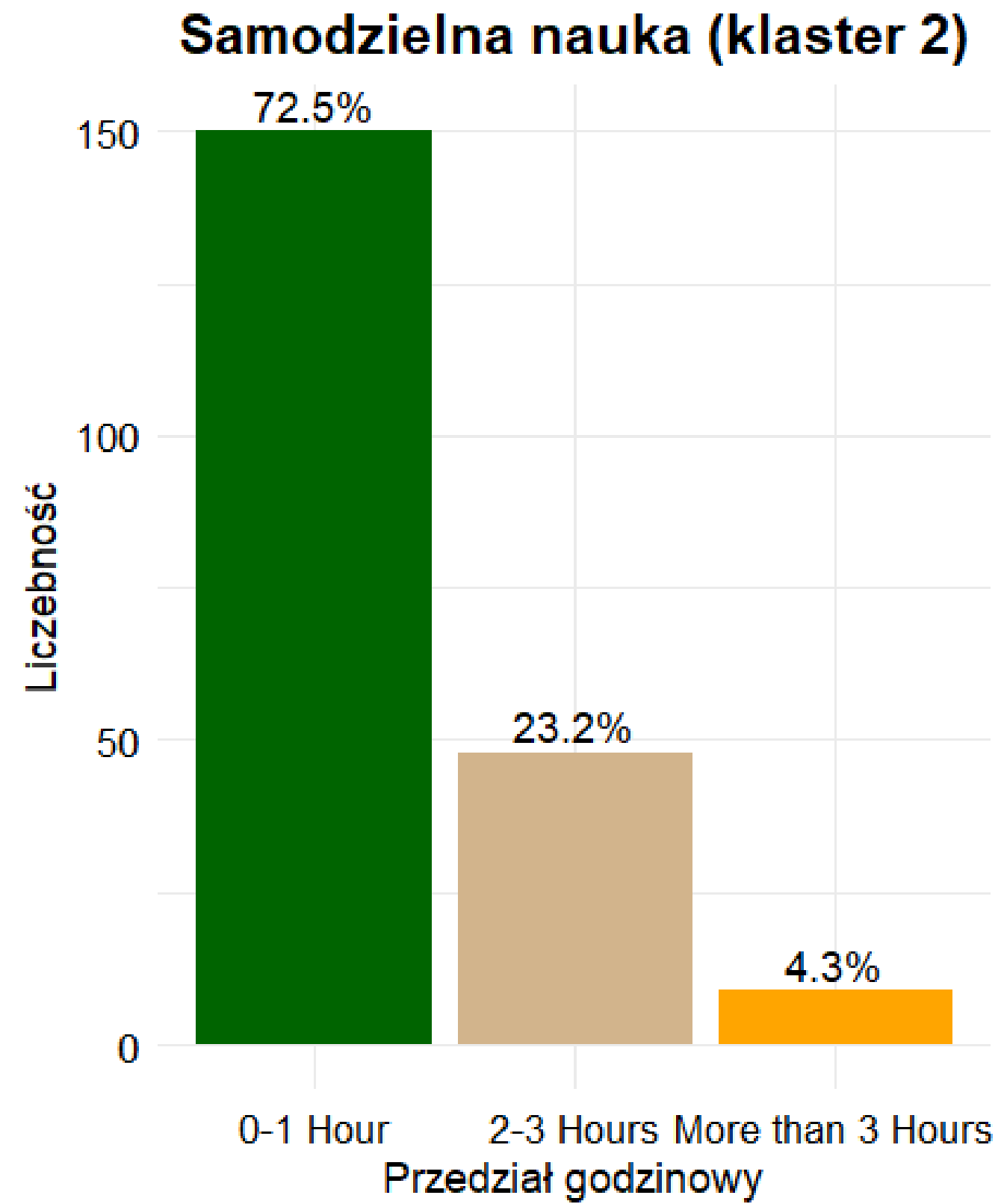
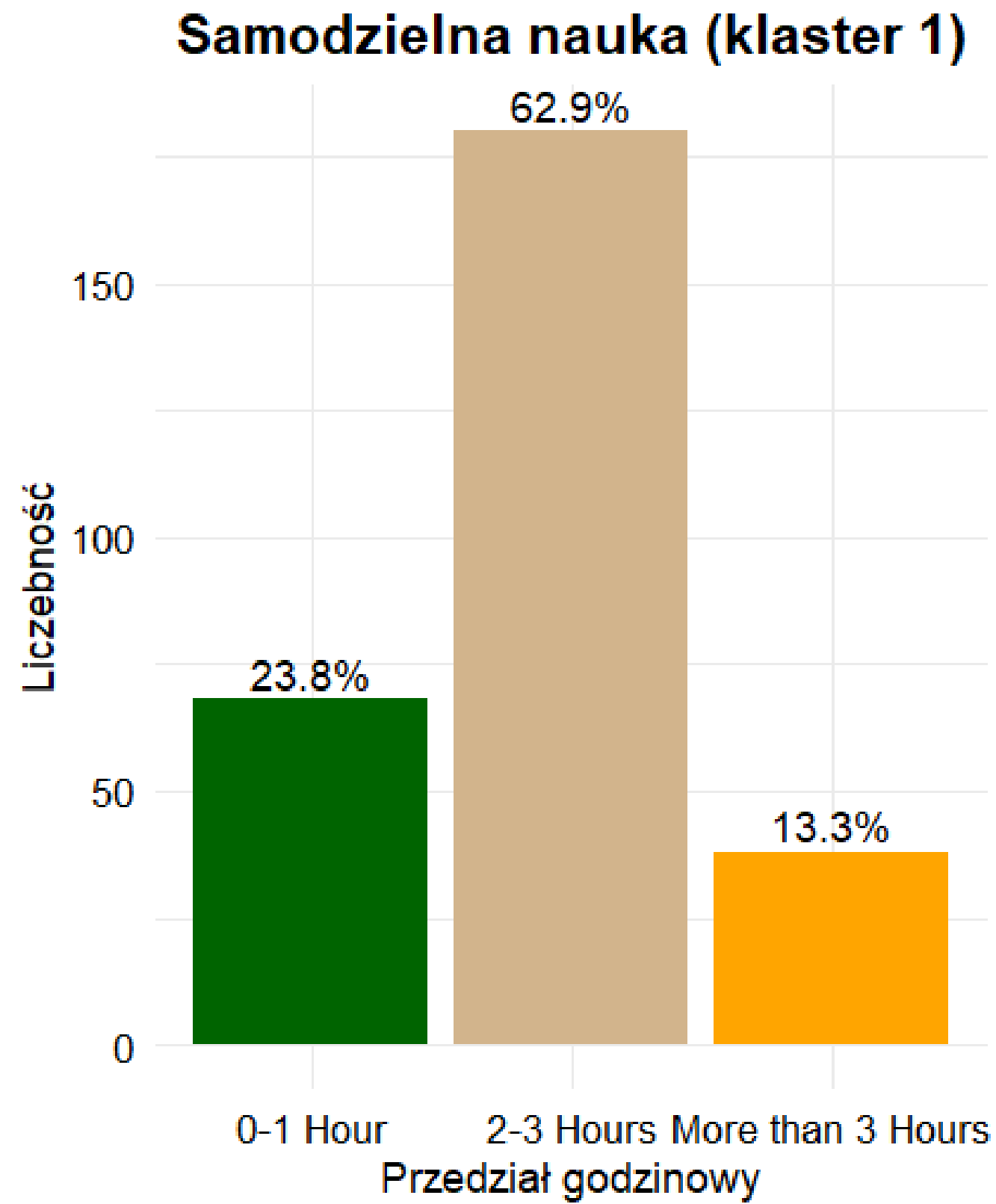
- Płeć: **kobieta (niewielki wpływ)**
- Miejsce zamieszkania: **brak istotnego wpływu**
- Praca: **brak**
- Zajęcia dodatkowe: **tak**
- Gaming: **<3 godziny**
- Angielski: **Wysoki poziom nie jest wymagany**

[1] 3.562752

Przygotowanie - Klaster - Obecność

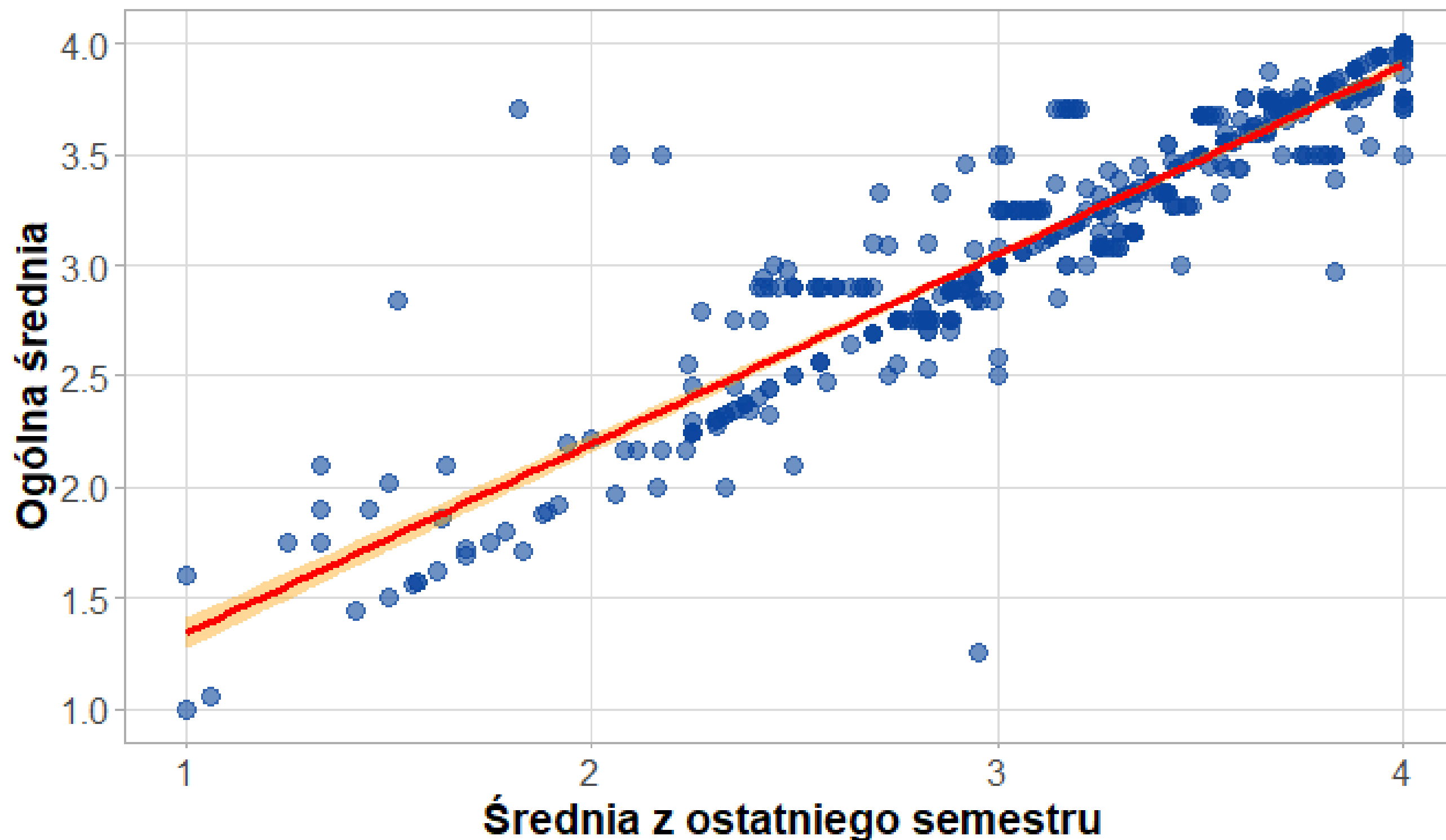






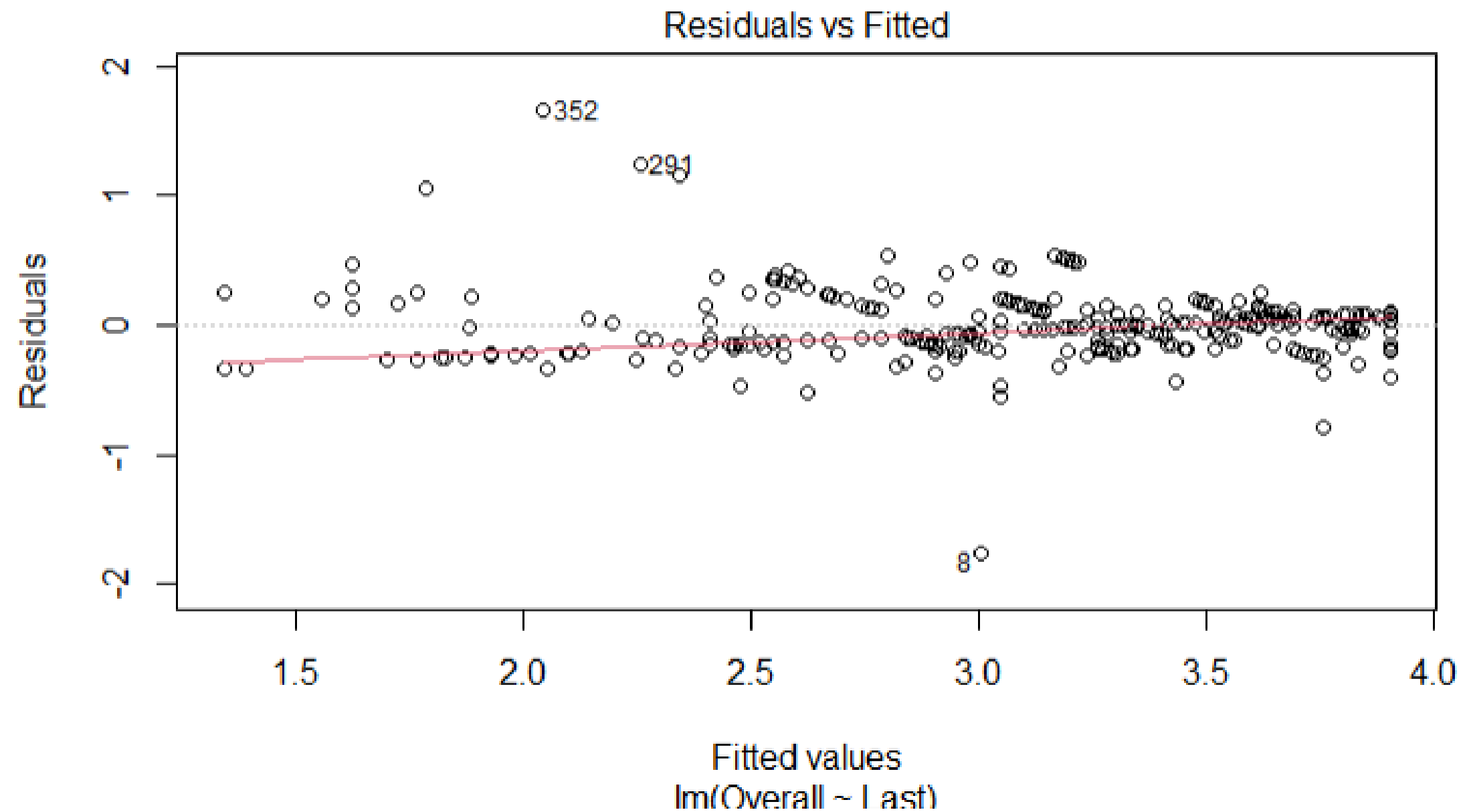
ANALIZA REGRESJI, SLR MODEL

Relacja między średnią z ostatniego semestru a ogólną średnią



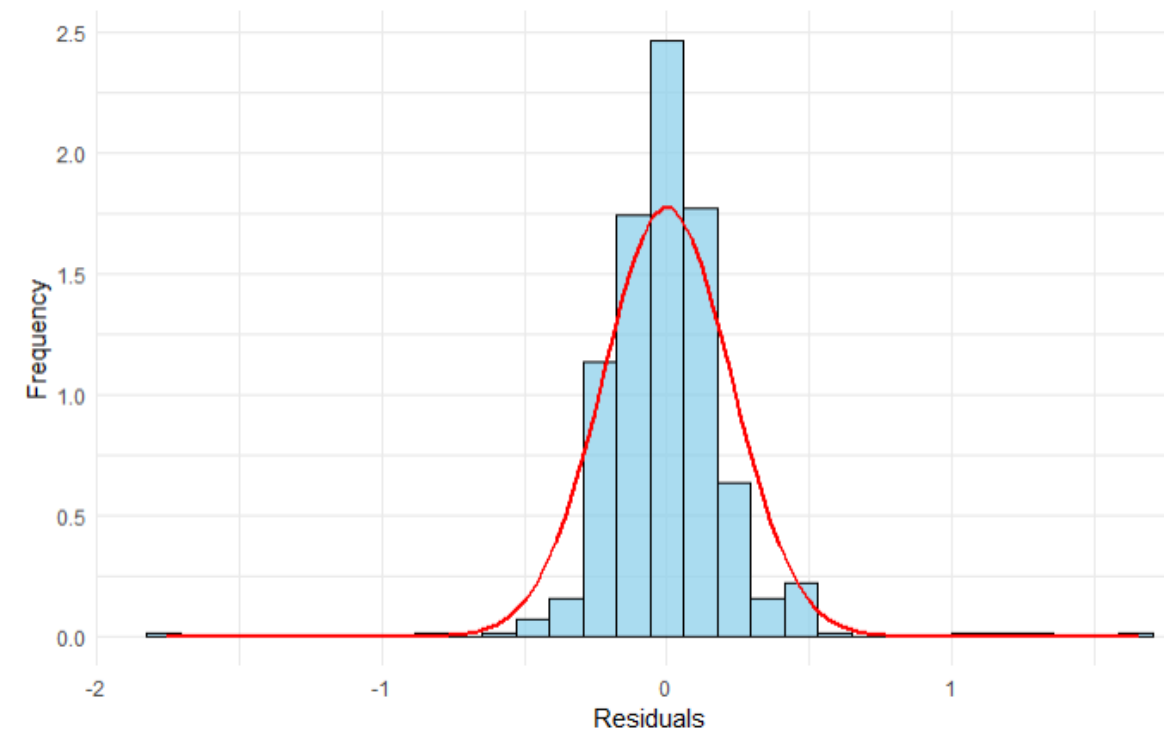
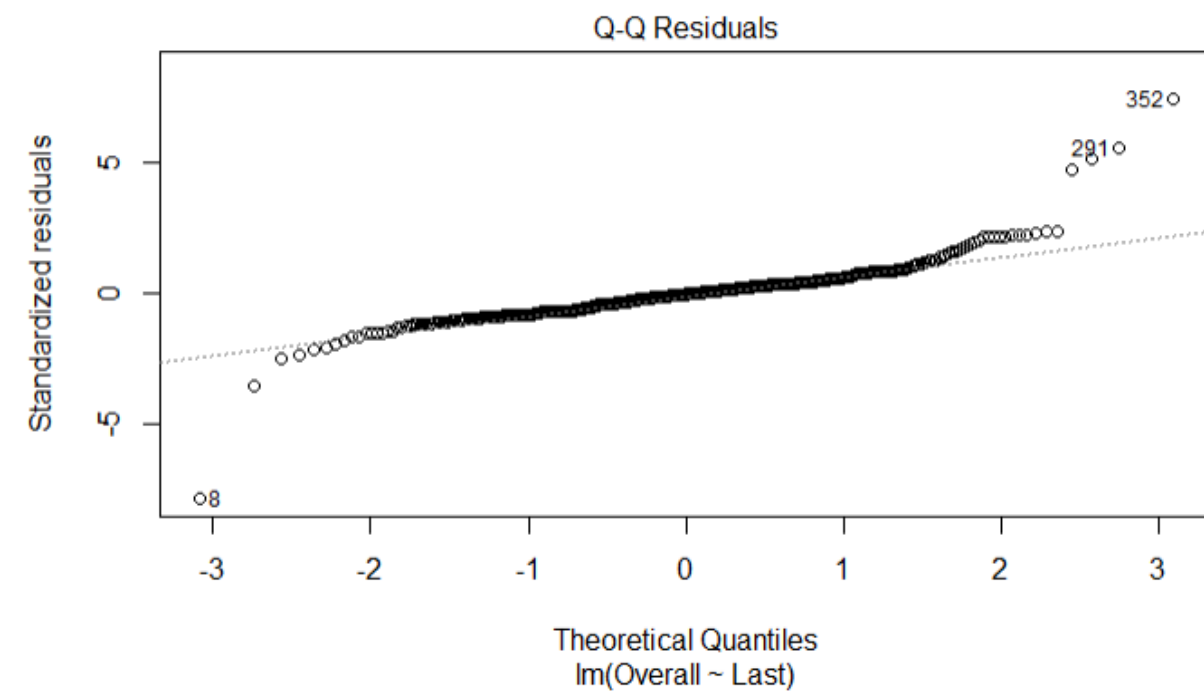
ZAŁOŻENIA

1 (RELACJA LINIOWA)-CZY W CAŁYM OBSZARZE ZMIENNOŚCI Y RESZTY ROZKŁADAJĄ SIĘ RÓWNOMIERNIE

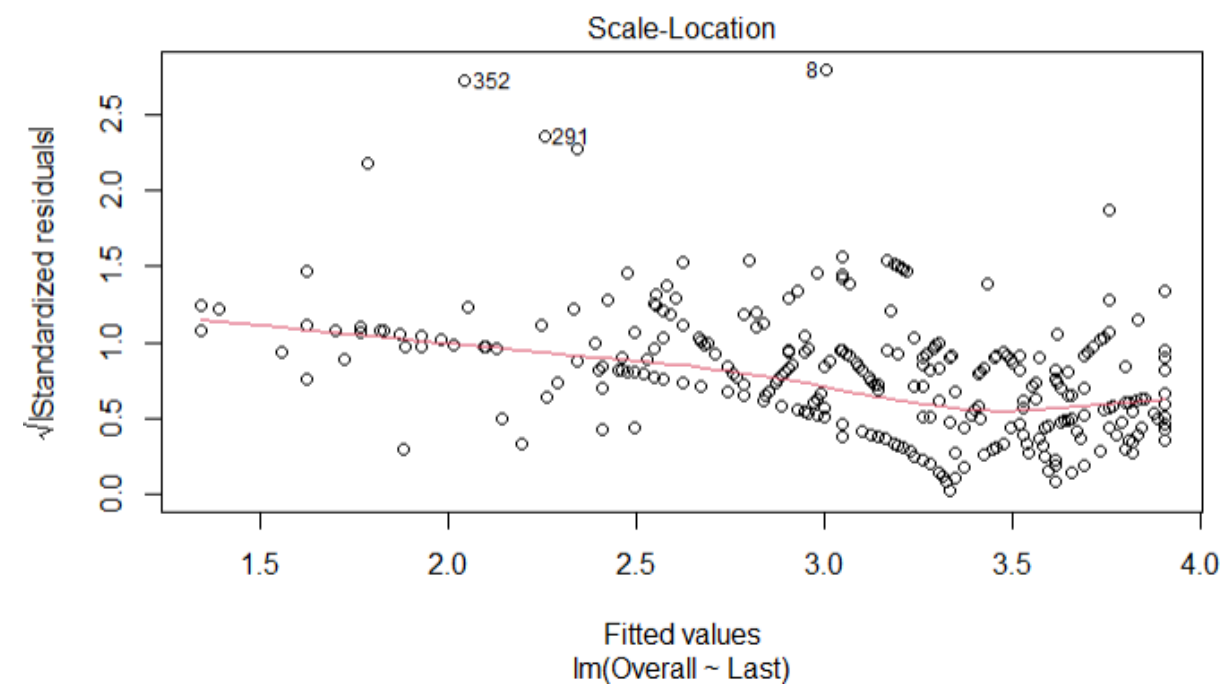


LINIA WZGLĘDNIĄ POZIOMA - WARUNEK SPEŁNIONY

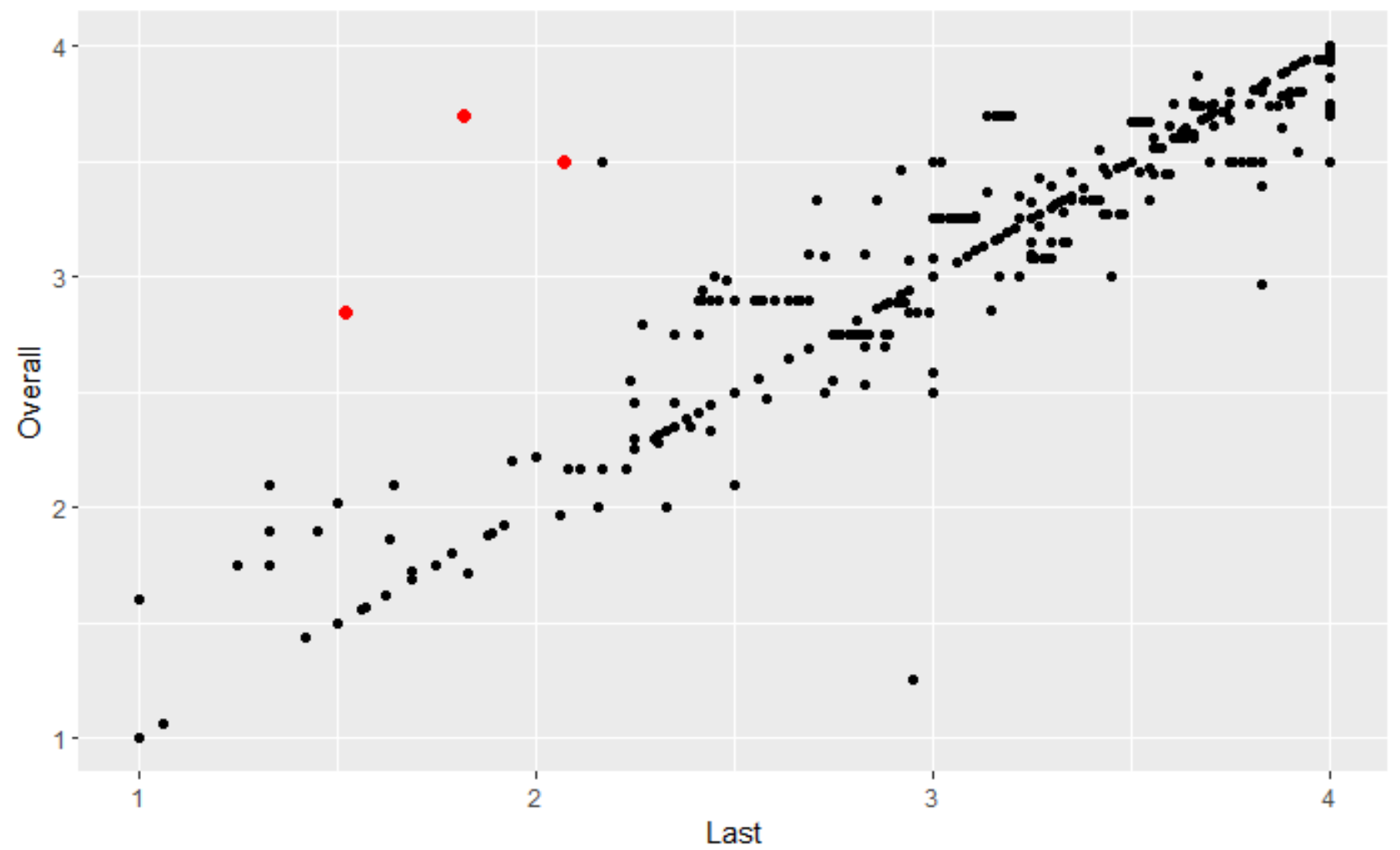
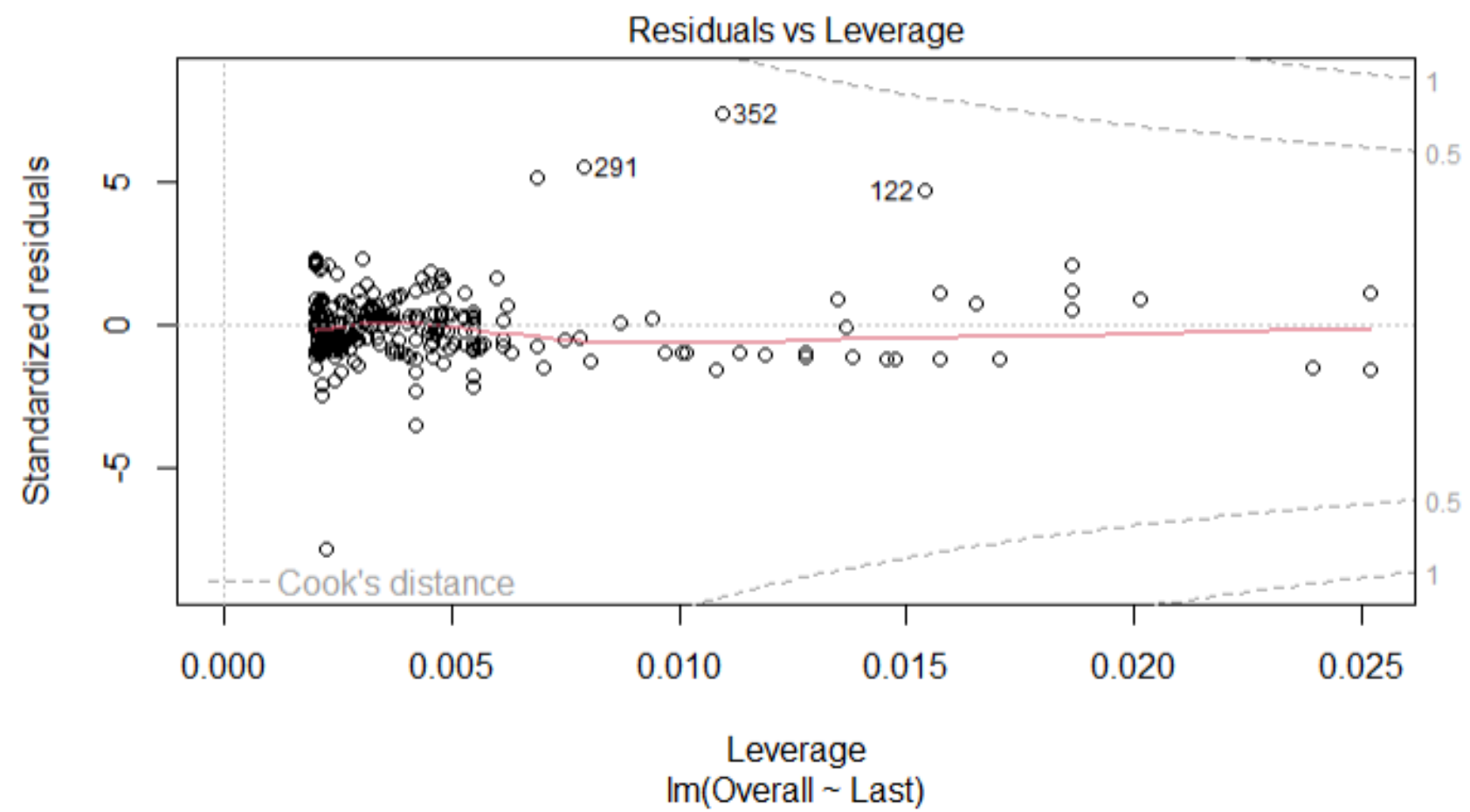
2 RESZTY MAJĄ ROZKŁAD NORMALNY



3 HOMOSCEDASTYCZNOŚĆ - RESZTY MAJĄ PORÓWNYWALNĄ WARIANCJĘ W CAŁYM OBSZARZE ZMIENNOŚCI ZMIENNEJ X



4 ZNACZENIE WARTOSCI EKSTREMALNYCH



model liniowy

Błędy/reszty

$$y = 0.85344 * \text{Last} + 0.48819$$

przeciętnie popełnimy
błąd na poziomie ~0.2

call:

```
lm(formula = overall ~ Last, data = data)
```

Residuals:

| Min | 1q | Median | 3q | Max |
|----------|-----------------|-----------------|----------------|---------|
| -1.75584 | <u>-0.13636</u> | <u>-0.00446</u> | <u>0.08925</u> | 1.65855 |

Odrzucenie hipotezy zerowej -
Last w istotny sposób pomaga
przewidzieć Overall

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------------|------------|---------|----------------------|
| (Intercept) | <u>0.48819</u> | 0.05104 | 9.564 | <2e-16 *** |
| Last | <u>0.85344</u> | 0.01581 | 53.974 | <u><2e-16 ***</u> |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.225 on 491 degrees of freedom

Multiple R-squared: 0.8558, Adjusted R-squared: 0.8555

F-statistic: 2913 on 1 and 491 DF, p-value: < 2.2e-16

(0-1) ~85% zmiennej zależnej jest
wyjaśnione przez zmienną niezależną

DZIĘKUJĘ ZA UWAGĘ!



Wszelkie pliki (dane, zarys, kod źródłowy) znajdują się na GitHubie :)

DAWID PAŁKA