Mini raport analityczny przygotowany na zaliczenie kursu ZintegrUJ 'Język R wsparciem warsztatu badacza'

Michał Bakalarz*

30 grudzień, 2022

Zmienna nominalna "gndr" i zmienna porządkowa "stflife"

Opis zmiennej "gndr"

```
str(esspl$gndr)
```

```
##  dbl+lbl [1:1500] 2, 1, 2, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, 2, 2, 2, 1, ...
##  @ label        : chr "Gender"
##  @ format.spss  : chr "F1.0"
##  @ display_width: int 6
##  @ labels       : Named num [1:3] 1 2 9
##   ..- attr(*, "names")= chr [1:3] "Male" "Female" "No answer"
```

```
describe(esspl$gndr)
```

```
##    vars    n mean  sd median trimmed mad min max range  skew kurtosis   se
## X1    1 1500 1.53 0.5      2    1.53   0   1   2     1 -0.11    -1.99 0.01
```

Opis zmiennej "stflife"

```
str(esspl$stflife)
```

```
##  dbl+lbl [1:1500]  7,  8,  5,  5,  3,  7,  7,  8,  6,  1,  8,  8,  7,  8,  ...
##  @ label        : chr "How satisfied with life as a whole"
##  @ format.spss  : chr "F2.0"
##  @ display_width: int 9
##  @ labels       : Named num [1:14] 0 1 2 3 4 5 6 7 8 9 ...
##   ..- attr(*, "names")= chr [1:14] "Extremely dissatisfied" "1" "2" "3" ...
```

```
describe(esspl$stflife)
```

```
##    vars    n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 1480 7.04 2.06      7    7.04 1.48   0  10    10 -0.76     0.55 0.05
```

```
summary(czad1)
```

```
##    esspl.gndr    esspl.stflife
##  Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.000   1st Qu.: 6.000
##  Median :2.000   Median : 7.000
##  Mean   :1.524   Mean   : 7.043
##  3rd Qu.:2.000   3rd Qu.: 8.000
##  Max.   :2.000   Max.   :10.000
```

*michal.bakalarz@student.uj.edu.pl

Tabela krzyżowa pommiędzy zmienną nominalna “gndr” i zmienna porządkowa “stflife”

```
tbl_cross(
  czad1,
  row = esspl.stflife,
  col = esspl.gndr,
  label = list(esspl.gndr ~ "Gender",
               esspl.stflife ~ "How satisfied with life as a whole"),
  statistic = "{n} ({p}%)",
  digits = c(0, 1),
  percent = c("column"),
  margin = c("column", "row"),
  missing = c("ifany"),
  missing_text = "Unknown",
  margin_text = "Total"
)
```

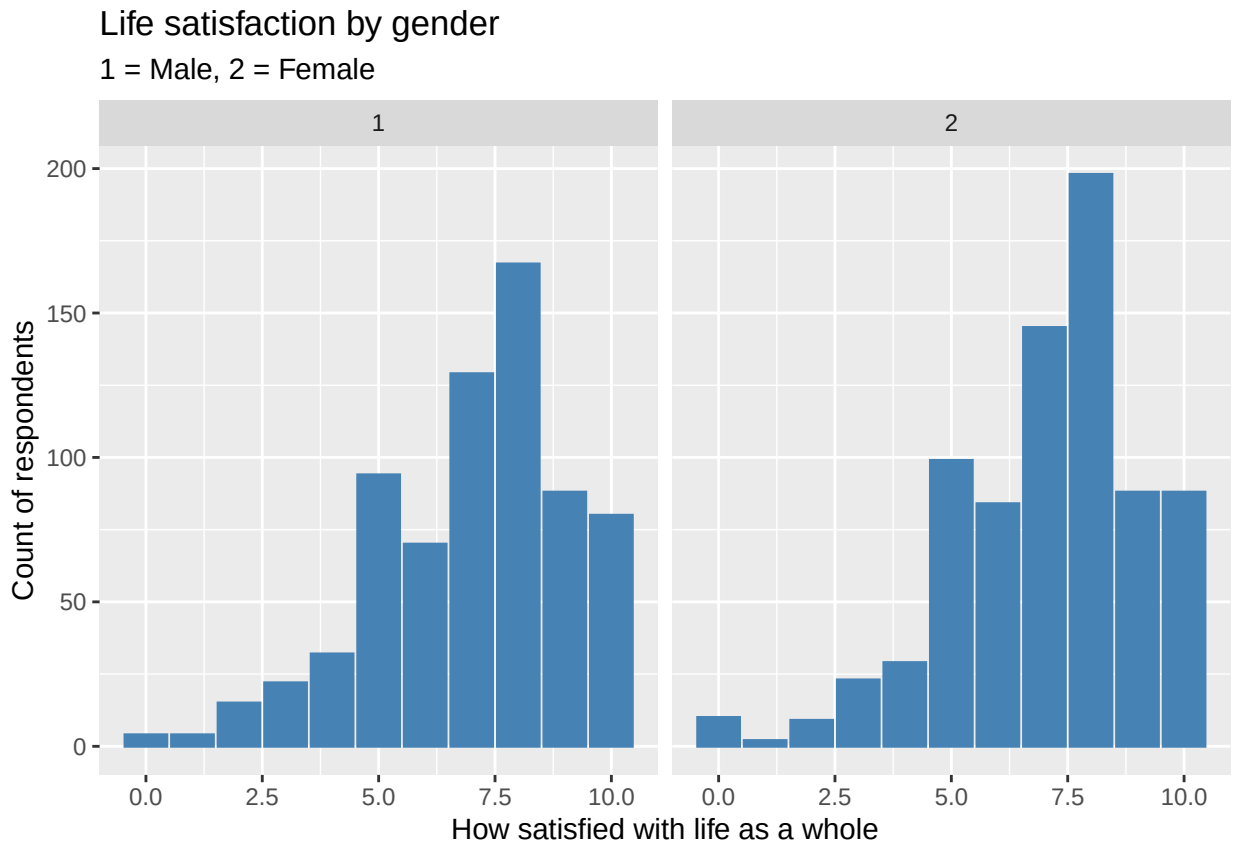
```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

	1	2	Total
How satisfied with life as a whole			
0	4 (0.6%)	10 (1.3%)	14 (0.9%)
1	4 (0.6%)	2 (0.3%)	6 (0.4%)
2	15 (2.1%)	9 (1.2%)	24 (1.6%)
3	22 (3.1%)	23 (3.0%)	45 (3.0%)
4	32 (4.5%)	29 (3.7%)	61 (4.1%)
5	94 (13.3%)	99 (12.8%)	193 (13.0%)
6	70 (9.9%)	84 (10.8%)	154 (10.4%)
7	129 (18.3%)	145 (18.7%)	274 (18.5%)
8	167 (23.7%)	198 (25.5%)	365 (24.7%)
9	88 (12.5%)	88 (11.4%)	176 (11.9%)
10	80 (11.3%)	88 (11.4%)	168 (11.4%)
Total	705 (100.0%)	775 (100.0%)	1,480 (100.0%)

Wykres słupkowy z podziałem na grupy wg kategorii zmiennej “gndr”

```
ggplot(czad1, aes(esspl.stflife)) +
  geom_bar(color = "steelblue", fill = "steelblue") +
  facet_wrap(vars(esspl.gndr)) +
  labs(title = "Life satisfaction by gender",
       subtitle = "1 = Male, 2 = Female ",
       y = "Count of respondents", x = "How satisfied with life as a whole")
```

```
## Don't know how to automatically pick scale for object of type
## <haven_labelled/vctrs_vctr/double>. Defaulting to continuous.
```



Testy niezależności

```
tblzad1 = table(esspl$gndr, esspl$stflife)

# chi2 test
chisq.test(tblzad1) # brak zależności między płcią a satysfakcją z życia
```

```
## Warning in chisq.test(tblzad1): Aproksymacja chi-kwadrat może być niepoprawna
```

```
##
##  Pearson's Chi-squared test
##
## data:  tblzad1
## X-squared = 6.963, df = 10, p-value = 0.7289
```

```
# Fisher test

fisher = fisher.test(tblzad1, simulate.p.value=TRUE)
fisher # test fishera potwierdza brak zależności pomiędzy płcią a satysfakcją z życia
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  tblzad1
## p-value = 0.7371
## alternative hypothesis: two.sided
```

H0 = Nie ma liniowego związku między satysfakcją z życia w Polsce, a płcią.

H1 = Istnieje liniowy związek między satysfakcją z życia w Polsce, a płcią.

Na podstawie testu Chi^2 przyjmujemy hipotezę zerową mówiącą, iż nie ma liniowego związku między analizowaną zmienną zależną, a daną zmienną niezależną. Zmienne w regresji są nieistotne statystycznie, ponieważ p>0,05.

H0 = Nie istnieje zależność pomiędzy satysfakcją z życia w Polsce, a płcią.

H1 = Istnieje zależność pomiędzy satysfakcją z życia w Polsce, a płcią.

Na podstawie dokładnego testu Fishera przyjmujemy hipotezę zerową mówiącą, iż nie istnieje zależność między analizowaną zmienną zależną, a daną zmienną niezależną. Zmienne w regresji są nieistotne statystycznie, ponieważ p>0,05.

Korelacja dwuseryjna (test siły związku)

```
b = biserial(esspl$stflife, esspl$gndr)
```

```
b # korelacja dwuseryjna
```

```
##            [,1]
## [1,] 0.01357968
```

Zmienna nominalna "gndr" i zmienna ilościowa "lrscale"

Opis zmiennej "gndr"

```
str(esspl$gndr)
```

```
##  dbl+lbl [1:1500] 2, 1, 2, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, 2, 2, 2, 1, ...
##  @ label        : chr "Gender"
##  @ format.spss  : chr "F1.0"
##  @ display_width: int 6
##  @ labels       : Named num [1:3] 1 2 9
##   ..- attr(*, "names")= chr [1:3] "Male" "Female" "No answer"
```

```
describe(esspl$gndr)
```

```
##    vars    n mean  sd median trimmed mad min max range  skew kurtosis   se
## X1    1 1500 1.53 0.5      2    1.53   0   1   2     1 -0.11    -1.99 0.01
```

```
IQR(esspl$gndr)
```

```
## [1] 1
```

Opis zmiennej "lrscale"

```
str(esspl$lrscale)
```

```
##  dbl+lbl [1:1500]  8,  2,  3,  5,  5, NA,  2, NA,  5, NA, NA,  5,  6,  7,  ...
##  @ label        : chr "Placement on left right scale"
##  @ format.spss  : chr "F2.0"
##  @ display_width: int 9
##  @ labels       : Named num [1:14] 0 1 2 3 4 5 6 7 8 9 ...
##   ..- attr(*, "names")= chr [1:14] "Left" "1" "2" "3" ...
```

```
describe(esspl$lrscale)
```

```
##    vars    n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 1236  5.8 2.48      5     5.8 2.97   0  10    10 -0.12    -0.36 0.07
```

```
IQR(na.omit(esspl$lrscale))
```

```
## [1] 3
```

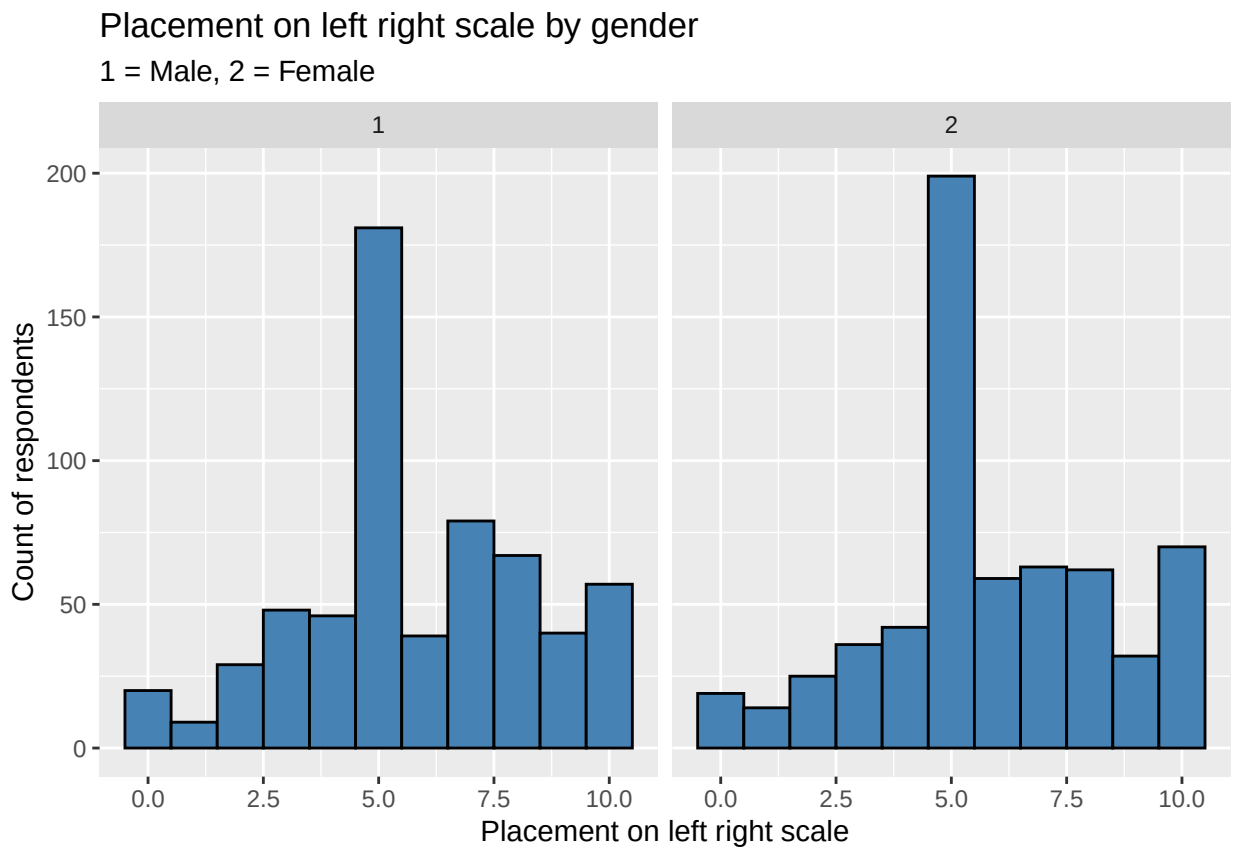
```
summary(czad2)
```

```
##    esspl.gndr    esspl.lrscale
##  Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.000   1st Qu.: 5.000
##  Median :2.000   Median : 5.000
##  Mean   :1.502   Mean   : 5.799
##  3rd Qu.:2.000   3rd Qu.: 8.000
##  Max.   :2.000   Max.   :10.000
```

Histogram w podziale na grupy wg kategorii zmiennej “gndr”

```
ggplot(czad2, aes(esspl.lrscale)) +
  geom_histogram(binwidth = 1, color = "black", fill = "steelblue") +
  facet_wrap(vars(esspl.gndr))+
  labs(title = "Placement on left right scale by gender",
       subtitle = "1 = Male, 2 = Female ",
       y = "Count of respondents", x = "Placement on left right scale")
```

```
## Don't know how to automatically pick scale for object of type
## <haven_labelled/vctrs_vctr/double>. Defaulting to continuous.
```




```
zależna2 = as.numeric(esspl$lrscale)
niezależna2 = esspl$gndr
```

```r
describeBy(zależna2, group = niezależna2)
```

```
##
##  Descriptive statistics by group
## group: 1
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis  se
## X1    1 615 5.78 2.49      5    5.82 2.97   0  10    10 -0.13    -0.42 0.1
## ------------------------------------------------------------
## group: 2
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis  se
## X1    1 621 5.82 2.48      5    5.87 2.97   0  10    10 -0.1     -0.3 0.1
```

```r
t.test(zależna2 ~ niezależna2) # t-test, średnie w grupach nie różnią się przyjmuje h0
```

```
##
##  Welch Two Sample t-test
##
## data:  zależna2 by niezależna2
## t = -0.30003, df = 1233.9, p-value = 0.7642
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.3197249  0.2349050
## sample estimates:
## mean in group 1 mean in group 2
##        5.777236        5.819646
```

```r
res_aov = aov(zależna2 ~ niezależna2) # ANOVA
summary(res_aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## niezależna2    1      1   0.556    0.09  0.764
## Residuals   1234   7618   6.174
## 264 obserwacje zostały skasowane z uwagi na braki w nich zawarte
```

Analiza równości średnich (parametryczne): t-test i ANOVA

```r
zależna2 = as.numeric(esspl$lrscale)
niezależna2 = esspl$gndr

describeBy(zależna2, group = niezależna2)
```

```
##
##  Descriptive statistics by group
## group: 1
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis  se
## X1    1 615 5.78 2.49      5    5.82 2.97   0  10    10 -0.13    -0.42 0.1
## ------------------------------------------------------------
## group: 2
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis  se
## X1    1 621 5.82 2.48      5    5.87 2.97   0  10    10 -0.1     -0.3 0.1
```

```r
t.test(zależna2 ~ niezależna2) # t-test, średnie w grupach nie różnią się przyjmuje h0
```

```
##
##  Welch Two Sample t-test
##
## data:  zależna2 by niezależna2
```

```
## t = -0.30003, df = 1233.9, p-value = 0.7642
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.3197249  0.2349050
## sample estimates:
## mean in group 1 mean in group 2
##        5.777236        5.819646
```

```
res_aov = aov(zależna2 ~ niezależna2) # ANOVA
summary(res_aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## niezależna2    1      1   0.556    0.09  0.764
## Residuals   1234   7618   6.174
## 264 obserwacje zostały skasowane z uwagi na braki w nich zawarte
```

H0 = Średnie w grupach kobiet i mężczyzn nie różnią się.

H1 = Średnie w grupach kobiet i mężczyzn różnią się.

Na podstawie testu t przyjmujemy hipotezę zerową mówiącą, iż średnie w grupach nie różnią się. Zmienne w regresji są nieistotne statystycznie, ponieważ p>0,05.

H0 = Nie ma liniowego związku pomiędzy uplasowaniem się na skali politycznej lewicy i prawicy w Polsce, a płcią.

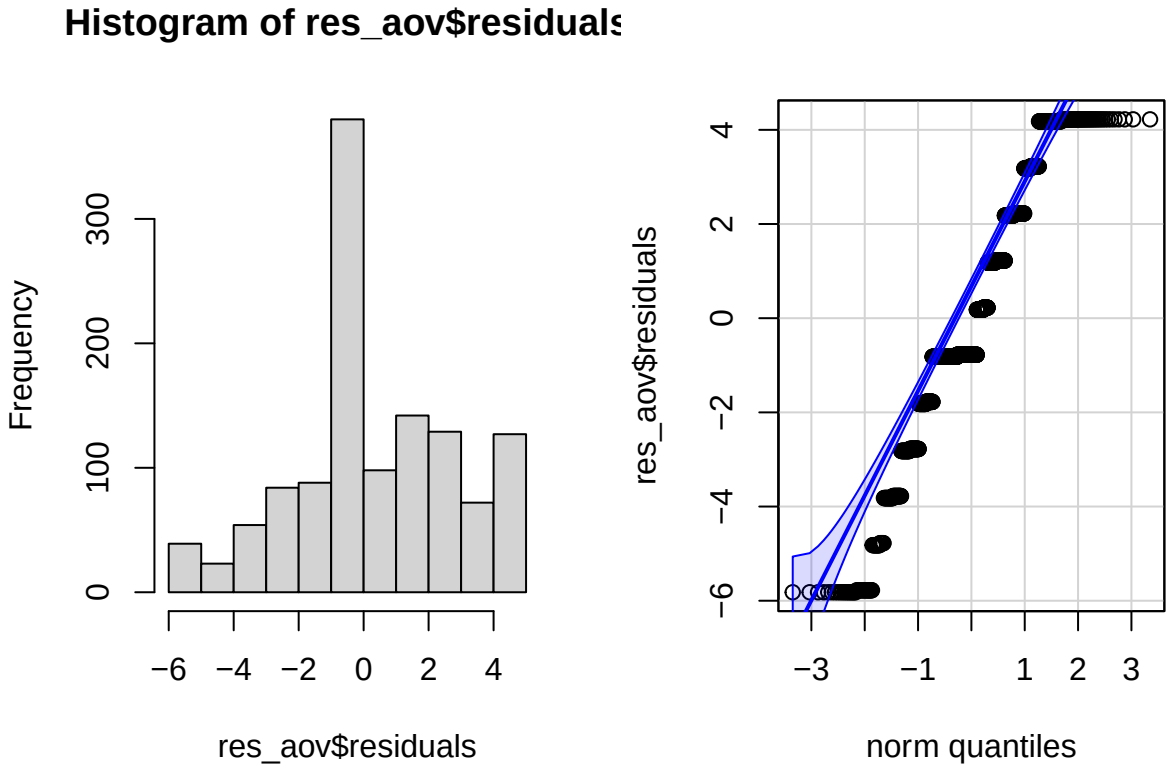
H1 = Istnieje liniowy związek pomiędzy uplasowaniem się na skali politycznej lewicy i prawicy w Polsce, a płcią.

Na podstawie testu F przyjmujemy hipotezę zerową mówiącą, iż nie ma liniowego związku między analizowanymi zmiennymi. Zmienne w regresji są nieistotne statystycznie, ponieważ p>0,05.

Sprawdzanie normalności rozkładu

```
par(mfrow = c(1, 2))
hist(res_aov$residuals) # histogram

qqPlot(res_aov$residuals,
       id = FALSE
) # QQ-plot
```



Korelacja dwuseryjna (test siły związku)

```
b2 = biserial(esspl$lrscale, esspl$gndr)
```

```
b2 # korelacja dwuseryjna
```

```
##             [,1]
## [1,] 0.01070002
```

Zmienna ilościowa “netustm” i zmienna ilościowa “stfeco”

Opis zmiennej “netustm”

```
str(esspl$netustm)
```

```
##  dbl+lbl [1:1500]  60,  60,  NA, 180,  30,  60, 180,  NA, 180,  NA,  NA,  6...
##  @ label        : chr "Internet use, how much time on typical day, in minutes"
##  @ format.spss  : chr "F4.0"
##  @ display_width: int 9
##  @ labels       : Named num [1:4] 6666 7777 8888 9999
##   ..- attr(*, "names")= chr [1:4] "Not applicable" "Refusal" "Don't know" "No answer"
```

```
describe(esspl$netustm)
```

```
##    vars    n   mean     sd median trimmed   mad min max range skew kurtosis   se
## X1    1 853 182.27 144.41    120  182.27 88.96   0 900   900 1.81     4.09 4.94
```

```
IQR(na.omit(esspl$netustm))
```

```
## [1] 150
```

Opis zmiennej “stfeco”

```
str(esspl$stfeco)
```

```
##  dbl+lbl [1:1500]  8,  4,  6,  3,  5,  7,  6, NA,  4,  5,  6,  7,  8,  6,  ...
##  @ label      : chr "How satisfied with present state of economy in country"
##  @ format.spss: chr "F2.0"
##  @ labels     : Named num [1:14] 0 1 2 3 4 5 6 7 8 9 ...
##   ..- attr(*, "names")= chr [1:14] "Extremely dissatisfied" "1" "2" "3" ...
```

```
describe(esspl$stfeco)
```

```
##    vars    n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 1427 5.79 2.03      6    5.79 1.48   0  10    10 -0.54     0.43 0.05
```

```
IQR(na.omit(esspl$stfeco))
```

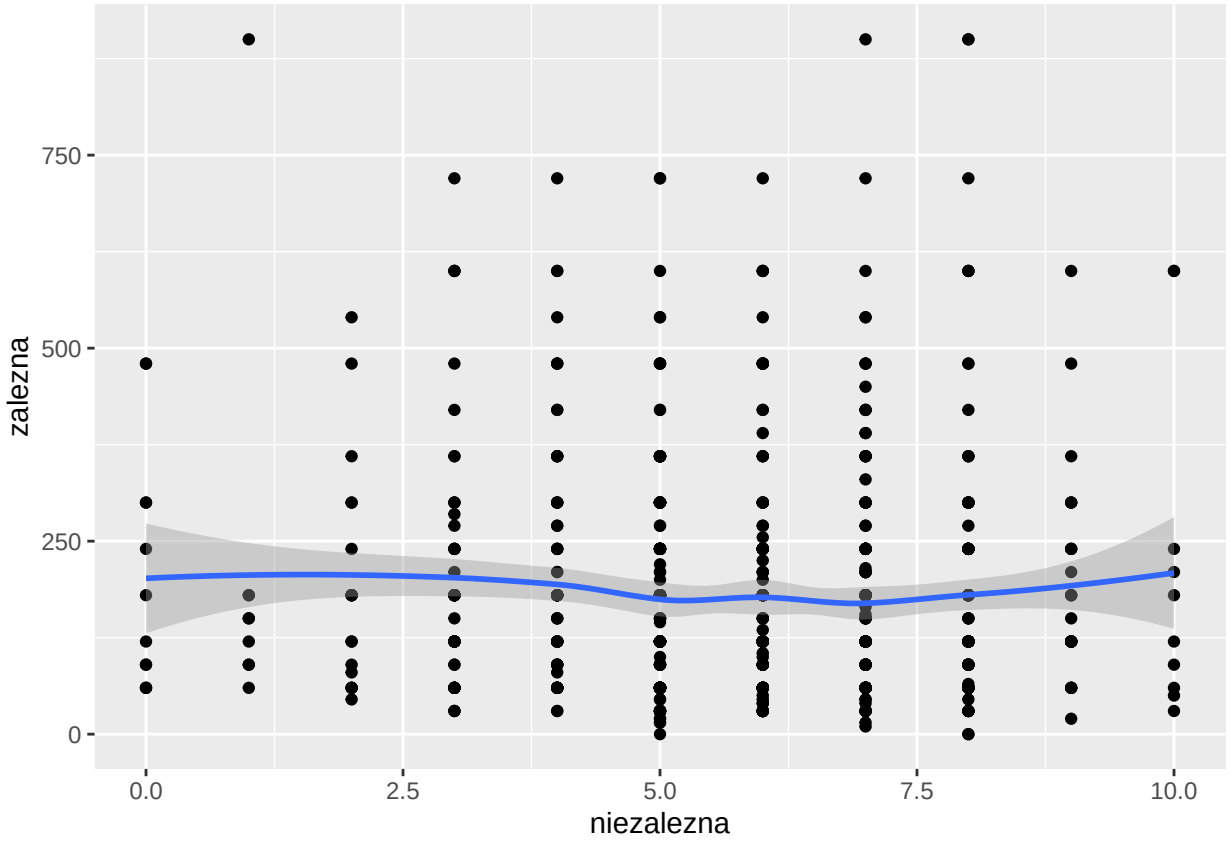
```
## [1] 2
```

```
summary(czad3) # podsumowanie df
```

```
##  esspl.netustm     esspl.stfeco
##  Min.   :  0.00   Min.   : 0.000
##  1st Qu.: 61.25   1st Qu.: 5.000
##  Median :120.00   Median : 6.000
##  Mean   :182.30   Mean   : 5.852
##  3rd Qu.:240.00   3rd Qu.: 7.000
##  Max.   :900.00   Max.   :10.000
```

```
ggplot(reg_df, aes(x = niezależna, y = zależna)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
summary(model) # model Regresji
```

```
##
## Call:
## lm(formula = zależna ~ niezależna, data = reg_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -184.91 -112.16  -52.64   58.16  724.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  200.252     16.589  12.071   <2e-16 ***
## niezależna    -3.068      2.698  -1.137    0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 145.8 on 816 degrees of freedom
## Multiple R-squared:  0.001583,   Adjusted R-squared:  0.0003591
## F-statistic: 1.294 on 1 and 816 DF,  p-value: 0.2557
```

```
cor(reg_df$niezależna, reg_df$zależna, method = c("pearson", "kendall", "spearman"))
```

```
## [1] -0.03978288
```

```
cor(na.omit(reg_df))
```

```
##             niezależna     zależna
```

```
## niezależna  1.00000000 -0.03978288
## zależna    -0.03978288  1.00000000
```

```
cor(reg_df, use = "pairwise.complete.obs")
```

```
##             niezależna     zależna
## niezależna  1.00000000 -0.03978288
## zależna    -0.03978288  1.00000000
```

H0 = Nie ma liniowego związku pomiędzy przeciętną długością korzystania z internetu w ciągu dnia przedstawioną w minutach, a satysfakcją z aktualnej sytuacji ekonomicznej w Polsce.

H1 = Istnieje liniowy związek pomiędzy przeciętną długością korzystania z internetu w ciągu dnia przedstawioną w minutach, a satysfakcją z aktualnej sytuacji ekonomicznej w Polsce.

Wartość p statystyki F wynosi 0.2537, oznacza to, iż nie ma istotnego związku, pomiędzy przeciętną długością korzystania z internetu w ciągu dnia przedstawioną w minutach, a satysfakcją z aktualnej sytuacji ekonomicznej w Polsce.

Ujemna korelacja r Pearsona wskazuję na śladowy związek, pomiędzy przeciętną długością korzystania z internetu w ciągu dnia przedstawioną w minutach, a satysfakcją z aktualnej sytuacji ekonomicznej w Polsce.