Statystyczna analiza danych - projekt zaliczeniowy 1

Marta Solarz

1. Wczytanie i podsumowanie danych

```
df <- read.csv('/home/websensa/Studies/SAD/projekt_1/people.csv')

# nagłówek
head(df)
```

```
##   wiek waga wzrost plec stan_cywilny liczba_dzieci      budynek wydatki
## 1   25 61.7 121.12 <NA>            0             2         loft 1662.91
## 2   37 63.9 145.00    M            1             6 wielka_plyta 4041.86
## 3   41 50.2 145.03    K            1             2   apartament 3853.45
## 4   43 72.4 179.90    M            0             1 wielka_plyta 2398.88
## 5   26 78.4 163.91    M            0             1   apartament 2344.45
## 6   49 59.4 151.86    K            1             2         loft 1967.87
##   wydatki_zywnosc oszczednosci
## 1         1466.37        23.44
## 2         3347.84        96.84
## 3         3220.90       312.68
## 4         2036.12       447.43
## 5         1992.61       -78.23
## 6         1706.45      1241.98
```

```
# liczba wierszy i kolumn
dim(df)
```

```
## [1] 499  10
```

```
# informacje o każdej zmiennej
str(df)
```

```
## 'data.frame':    499 obs. of  10 variables:
##  $ wiek           : int  25 37 41 43 26 49 27 49 38 33 ...
##  $ waga           : num  61.7 63.9 50.2 72.4 78.4 59.4 67.5 82.3 64.1 77.4 ...
##  $ wzrost         : num  121 145 145 180 164 ...
##  $ plec           : Factor w/ 2 levels "K","M": NA 2 1 2 2 1 2 1 1 2 ...
##  $ stan_cywilny   : int  0 1 1 0 0 1 0 0 1 0 ...
##  $ liczba_dzieci  : int  2 6 2 1 1 2 1 0 5 2 ...
##  $ budynek        : Factor w/ 5 levels "apartament","jednorodzinny",..: 4 5 1 5 1 4 2 5 5 1 ...
##  $ wydatki        : num  1663 4042 3853 2399 2344 ...
##  $ wydatki_zywnosc: num  1466 3348 3221 2036 1993 ...
##  $ oszczednosci   : num  23.4 96.8 312.7 447.4 -78.2 ...
```

```
# zmienna stan cywilny mimo, że reprezentowana jest liczbami,
# w rzeczywistości jest rozróżnieniem jakościowym - zamieńmy ją zatem na factor
df$stan_cywilny <- as.factor(df$stan_cywilny)
# sprawdzenie w których kolumnach występują NA
```

```r
sapply(df, function(x) any(is.na(x)))
```

```
##            wiek            waga          wzrost            plec    stan_cywilny
##           FALSE           FALSE           FALSE            TRUE           FALSE
##   liczba_dzieci         budynek         wydatki wydatki_zywnosc    oszczednosci
##           FALSE           FALSE           FALSE           FALSE           FALSE
```

```r
# liczba wartości NA
sum(is.na(df))
```

```
## [1] 38
```

```r
# statystyki opisowe dla kolumn ilościowych
summary(df[c("wiek", "waga", "wzrost", "liczba_dzieci",
             "wydatki", "wydatki_zywnosc", "oszczednosci")])
```

```
##       wiek            waga            wzrost      liczba_dzieci
##  Min.   :17.00   Min.   : 45.20   Min.   :113.6   Min.   :0.000
##  1st Qu.:33.00   1st Qu.: 59.20   1st Qu.:155.8   1st Qu.:1.000
##  Median :39.00   Median : 67.50   Median :169.0   Median :1.000
##  Mean   :39.47   Mean   : 68.03   Mean   :168.2   Mean   :1.561
##  3rd Qu.:45.00   3rd Qu.: 75.60   3rd Qu.:180.2   3rd Qu.:2.000
##  Max.   :72.00   Max.   :107.20   Max.   :235.2   Max.   :6.000
##     wydatki       wydatki_zywnosc   oszczednosci
##  Min.   : 524.9   Min.   : 523.1   Min.   :-685.68
##  1st Qu.:1810.7   1st Qu.:1562.8   1st Qu.:  72.87
##  Median :2493.3   Median :2111.0   Median : 401.00
##  Mean   :2515.4   Mean   :2112.7   Mean   : 476.64
##  3rd Qu.:3086.6   3rd Qu.:2575.5   3rd Qu.: 802.15
##  Max.   :5574.6   Max.   :4531.6   Max.   :3503.90
```

```r
# tabele częstości dla zmiennych jakościowych
table(df$plec)
```

```
##
##   K   M
## 238 223
```

```r
table(df$budynek)
```

```
##
##    apartament jednorodzinny     kamienica          loft  wielka_plyta
##            54           187           105            53           100
```

```r
table(df$stan_cywilny)
```

```
##
##   0   1
## 326 173
```

W zbiorze mamy 499 obserwacji, zmiennych ilościowych jest 7 (“wiek”, “waga”, “wzrost”, “liczba_dzieci”, “wydatki”, “wydatki_zywnosc”, “oszczednosci”), a jakościowych 3 (“plec”, “budynek”, “stan_cywilny” - wyrażony liczbą, ale będący jakościowym rozróżnieniem). Występuje 38 braków danych, wszystkie w kolumnie “plec”.

Statystyki opisowe: mają sens dla zmiennych ilościowych (wygenerowane powyżej), dzięki funkcji summary można sprawdzić m.in. wartości mediany, średniej, wartości maksymalnej i minimalnej. Analizując opis danych, można zauważyć m.in., że wartość mediany wieku wynosi 39 lat, a wartość mediany wagi 67,5 kg. Średnia liczba dzieci wynosi 1,561, a wartość mediany wydatków to 2493,3. Oszczędności wynoszą

średnio 476,64, a 25% badanych ma oszczędności mniejsze niż 72,87. Zauważmy także, że wszystkie wartości maksymalne i minimalne są prawdopodobne, więc możemy założyć że nie ma błędów w danych.

Tabele częstości: mają sens dla rozróżnień jakościowych. W zbiorze danych mamy 238 kobiety i 223 mężczyzn. 54 osoby mieszkają w apartamencie, 187 w domu jednorodzinnym, 105 w kamienicy, 53 w loftach, a w budynkach wielkopłytowych 100 osób. Ponadto 326 osób jest stanu wolnego (kawaler/panna), a 173 w związkach małżeńskich.

Przed przystąpieniem do dalszych analiz sprawdzimy normalność rozkładu zmiennych ilościowych (potem ta informacja przyda się do wyboru odpowiedniego testu statystycznego).

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
choose_numeric_columns <- function(df) {
  return(select_if(df, is.numeric))
}

# Sprawdzenie, które zmienne mają rozkład normalny:
# H0: rozkład i-tej zmiennej jest normalny
# H1: rozkład i-tej zmiennej jest różny od normalnego
# poziom istotności alpha - 0.05
choose_gauss_variables <- function(df, alpha=0.05) {
  dane <- select_if(df, function(x) {
    if (length(unique(x)) == 1) {
      return(FALSE)
    } else {
      result <- shapiro.test(x)
      return(result$p.value > alpha)
    }
  })
  return(dane)
}

ilosciowe <- choose_numeric_columns(df)
names(choose_gauss_variables(ilosciowe, 0.05))
```

```
## [1] "wzrost"
```

Na podstawie przeprowadzonych testów normalności Shapiro-Wilka dla zmiennych ilościowych na poziomie istotności 0.05 możemy stwierdzić, że tylko w przypadku zmiennej "wzrost" nie ma przesłanek do odrzucenia H0. Dla pozostałych zmiennych przyjmujemy hipotezę H1 (rozkład jest różny od rozkładu normalnego).

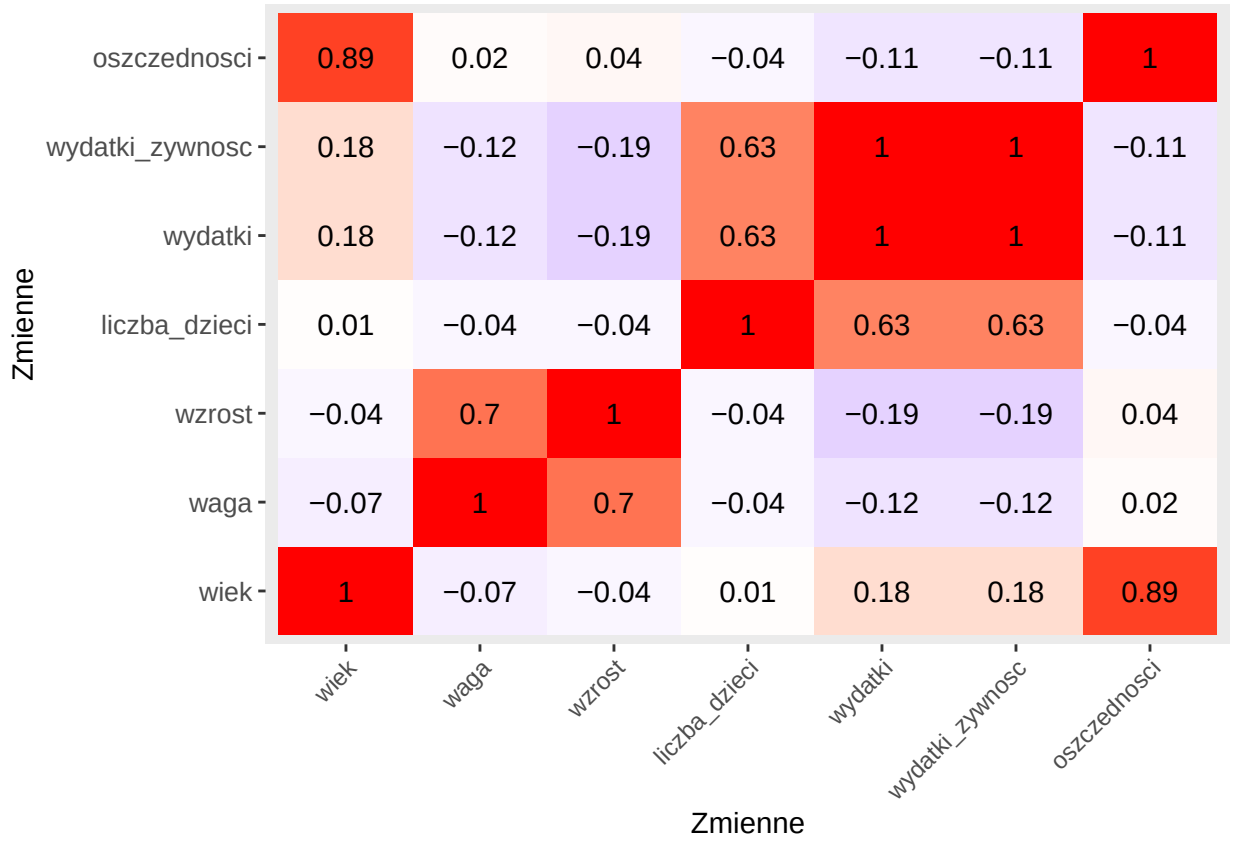
2. Sprawdzenie, czy występują pomiędzy zmiennymi zależności

I. Zmienne ilościowe

Kroki postępowania:

1. Liczymy korelację między danymi ilościowymi
2. Przygotowujemy macierz korelacji (dla wszystkich zmiennych ilościowych)
3. Sprawdzamy istotność zależności między zmiennymi
4. Przygotowujemy macierz korelacji tylko dla tych zmiennych, które są istotne statystycznie

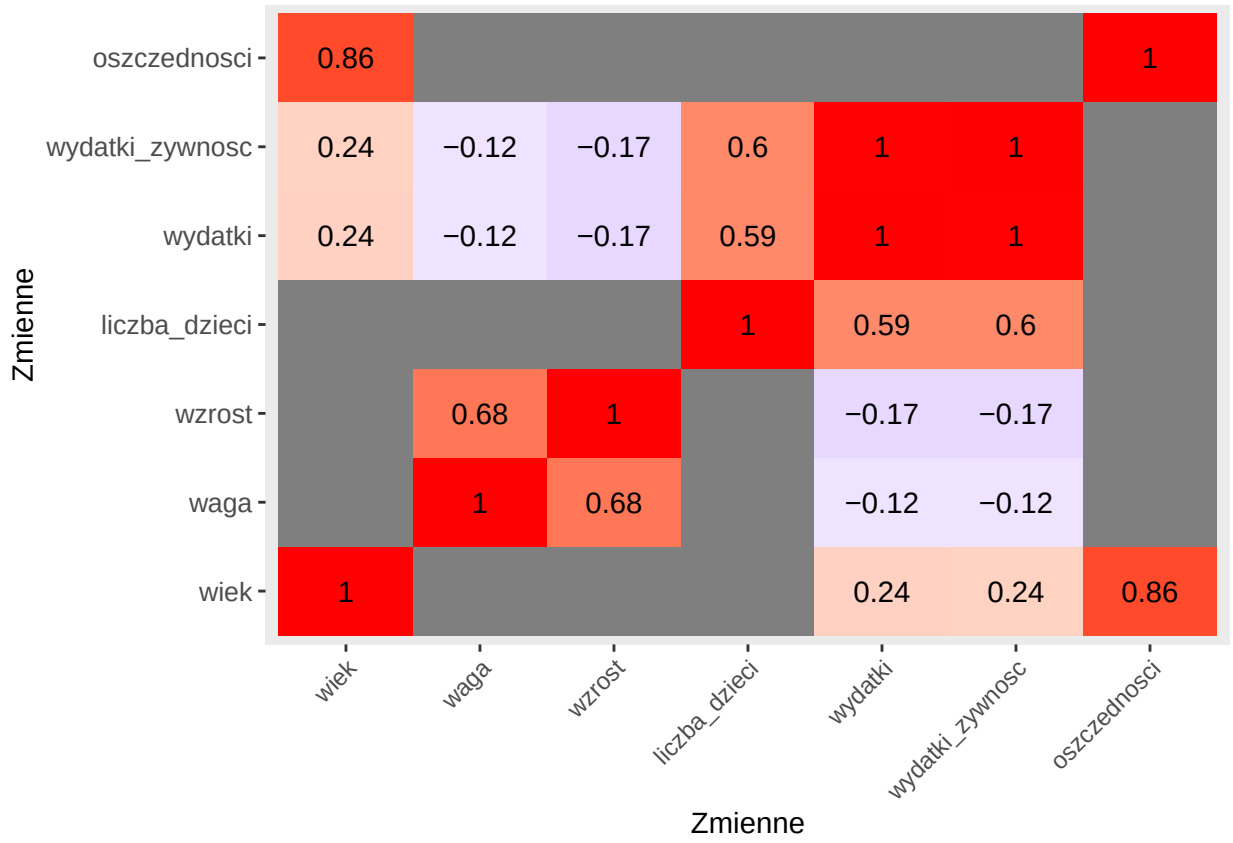
```r
library(ggplot2)
library(reshape2)

# 1.
cor_matrix <- round(cor(df[c("wiek", "waga",
                             "wzrost", "liczba_dzieci",
                             "wydatki", "wydatki_zywnosc",
                             "oszczednosci")]), 2)

# 2.
ggplot(data = reshape2::melt(cor_matrix)) +
  geom_tile(aes(x = Var1, y = Var2, fill = value)) +
  scale_fill_gradient2(low = "blue", high = "red") +
  geom_text(aes(x = Var1, y = Var2, label = value)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.text.y = element_text(size = 10),
        panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Zmienne", y = "Zmienne", fill = "Wsp. korelacji")
```




```r
# 3.
# H0: brak zależności między zmiennymi ilościowymi X i Y
```

```r
# H1: istnieje zależność między zmiennymi ilościowymi X i Y
# alpha = 0.05 - domyślny
# stosujemy test spearmana z powodu braku rozkładu normalnego
# dla conajmniej jednej zmiennej wśród badanych par
macierz_korelacji <- cor(ilosciowe, use = "pairwise.complete.obs", method = "spearman")
macierz_p <- matrix(nrow=ncol(ilosciowe), ncol=ncol(ilosciowe))

# Uzupełniamy macierz p-value
for (i in seq(ncol(ilosciowe))) {
  for (j in seq(ncol(ilosciowe))) {
    korelacja_wsk <- cor.test(ilosciowe[, i], ilosciowe[, j],
                              use="complete.obs", method = "spearman", exact = FALSE)
    p_val <- korelacja_wsk$p.value
    macierz_p[i,j] <- p_val
  }
}

istotne <- macierz_p
istotne[macierz_p < 0.05] <- 1
istotne[macierz_p >= 0.05] <- NA

macierz_ist_korelacji <- macierz_korelacji
macierz_ist_korelacji[is.na(istotne)] <- NA

# 4.
ggplot(data = reshape2::melt(macierz_ist_korelacji), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red") +
  geom_text(aes(label = round(value, 2))) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.text.y = element_text(size = 10),
        panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Zmienne", y = "Zmienne", fill = "Wsp. korelacji")
```

```
## Warning: Removed 20 rows containing missing values (`geom_text()`).
```



Zależność istotna statystycznie widoczna jest dla:

- wydatków i wydatków na żywność (korelacja wynosi 1)
- wiek i oszczędności (0.86)
- wzrostu i wagi (0.68)
- wydatków na żywność i liczby dzieci (0.6)
- wydatków i liczby dzieci (0.59)
- wydatki i wiek oraz wydatki na żywność i wiek (0.24)
- wydatki i waga oraz wydatki na żywność i waga (-0.12)
- wydatki i wzrost oraz wydatki na żywność i wzrost (-0.17)

A zatem otrzymaliśmy wyniki, które wydają się bardzo prawdopodobne - uzasadnione jest zakładanie, że wydatki są mocno powiązane z wydatkami na żywność, wiek z oszczędnościami, wzrost z wagą, a także wydatki i wydatki na żywność z liczbą dzieci. Natomiast między wydatkami, wydatkami na żywność a wiekiem, wagą i wzrostem nie ma intuicyjnie powodu, aby obserwowalna była silna zależność.

Pozostałe zależności (widoczne na pierwszym rysunku) nie są istotne statystycznie.

II. Zmienne jakościowe

```
# H0: brak zależności między zmiennymi jakościowymi X i Y
# H1: istnieje zależność między zmiennymi jakościowymi X i Y
# alpha = 0.05 - domyślny
# test chi2 - badamy zmienne jakościowe

# Dla stanu cywilnego i rodzaju budynku:
tab <- table(df$stan_cywilny, df$budynek)
```

```r
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 5.5743, df = 4, p-value = 0.2333
```

```r
# Dla płci i rodzaju budynku
tab <- table(df$plec, df$budynek)
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 1.2154, df = 4, p-value = 0.8756
```

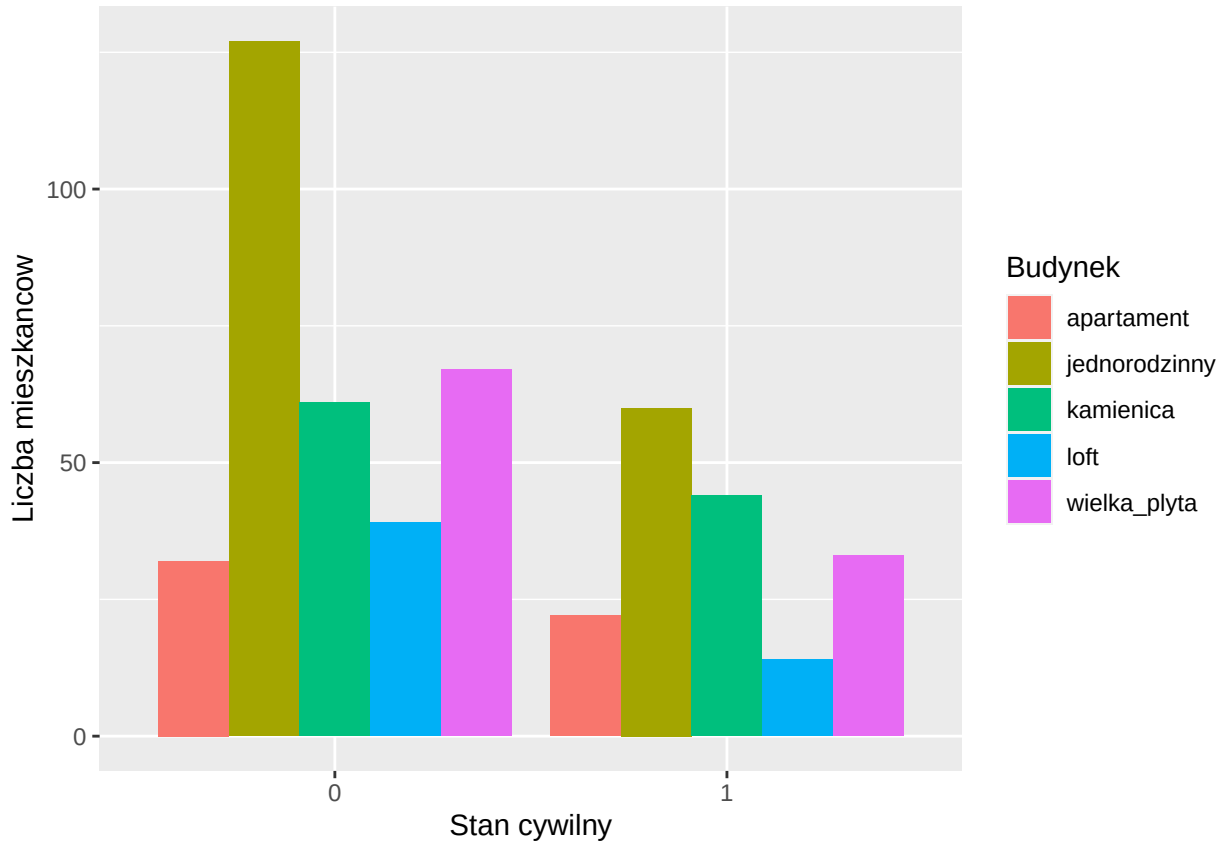
```r
# Dla płci i stanu cywilnego
tab <- table(df$plec, df$stan_cywilny)
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 2.3903, df = 1, p-value = 0.1221
```

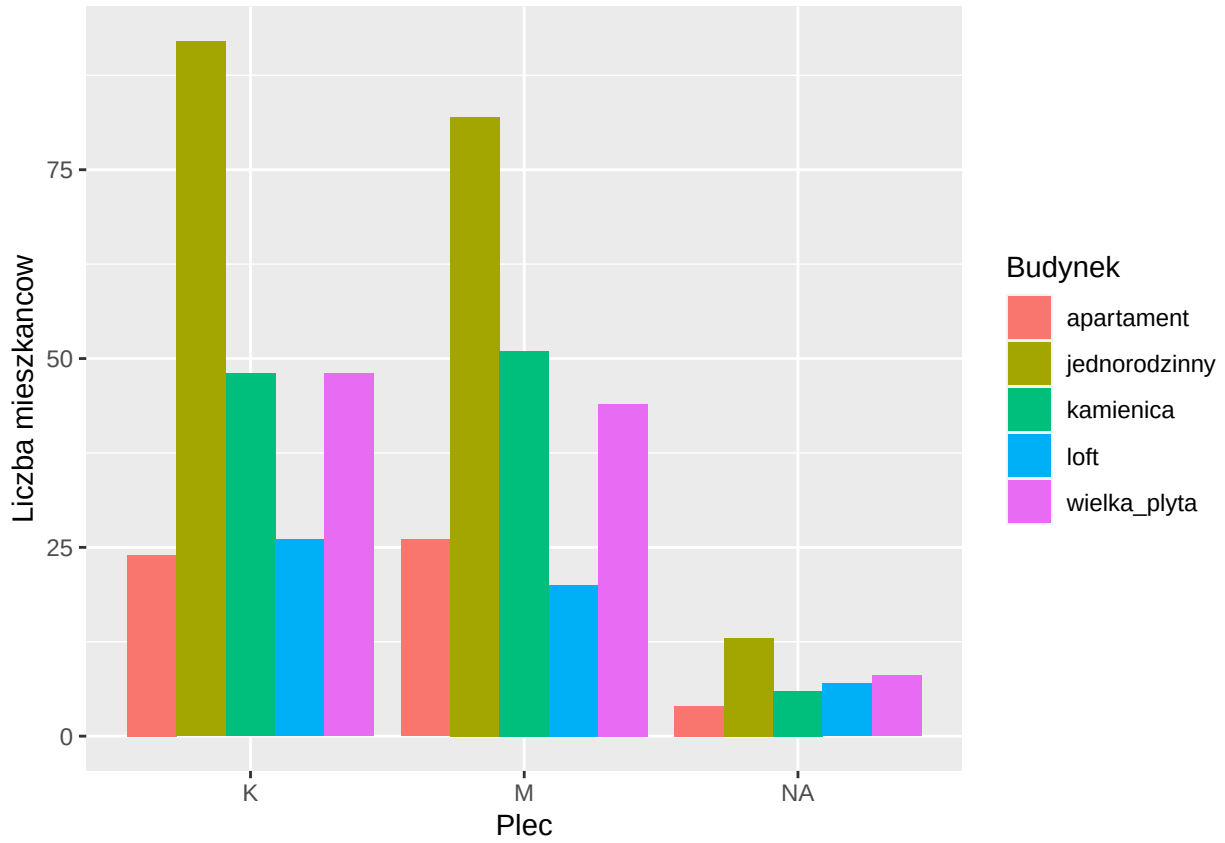
Dla wszystkich przypadków otrzymane wartości p są większe od alpha, stąd możemy wnioskować, że nie ma podstaw do odrzucenia H0 we wszystkich trzech powyższych przypadkach (nie istnieje istotna statystycznie zależność między tymi zmiennymi).

Poniżej znajdują się wykresy słupkowe ukazujące powiązania między zmiennymi jakościowymi.

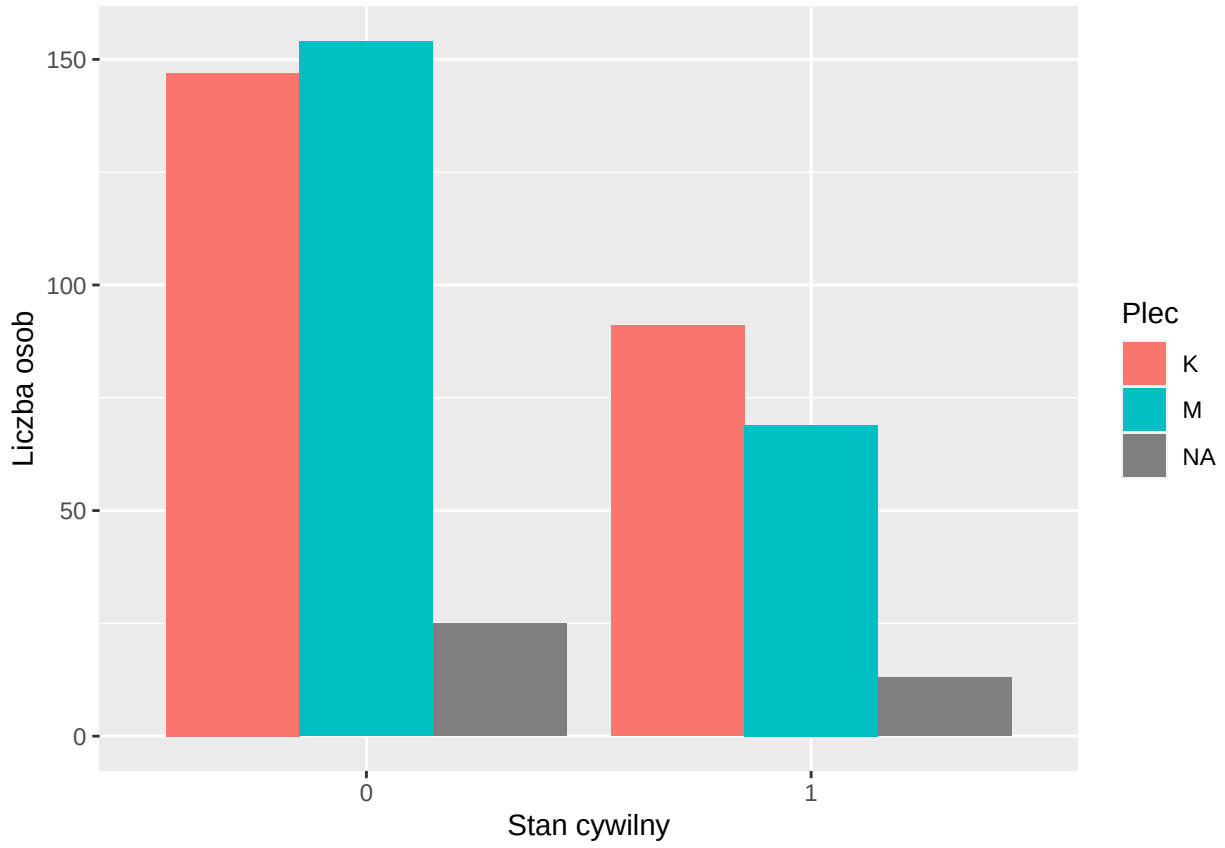
```r
ggplot(data = df, aes(x = stan_cywilny, fill = budynek)) +
  geom_bar(position = "dodge") +
  labs(x = "Stan cywilny", y = "Liczba mieszkancow", fill = "Budynek")
```




```
ggplot(data = df, aes(x = plec, fill = budynek)) +
  geom_bar(position = "dodge") +
  labs(x = "Plec", y = "Liczba mieszkancow", fill = "Budynek")
```



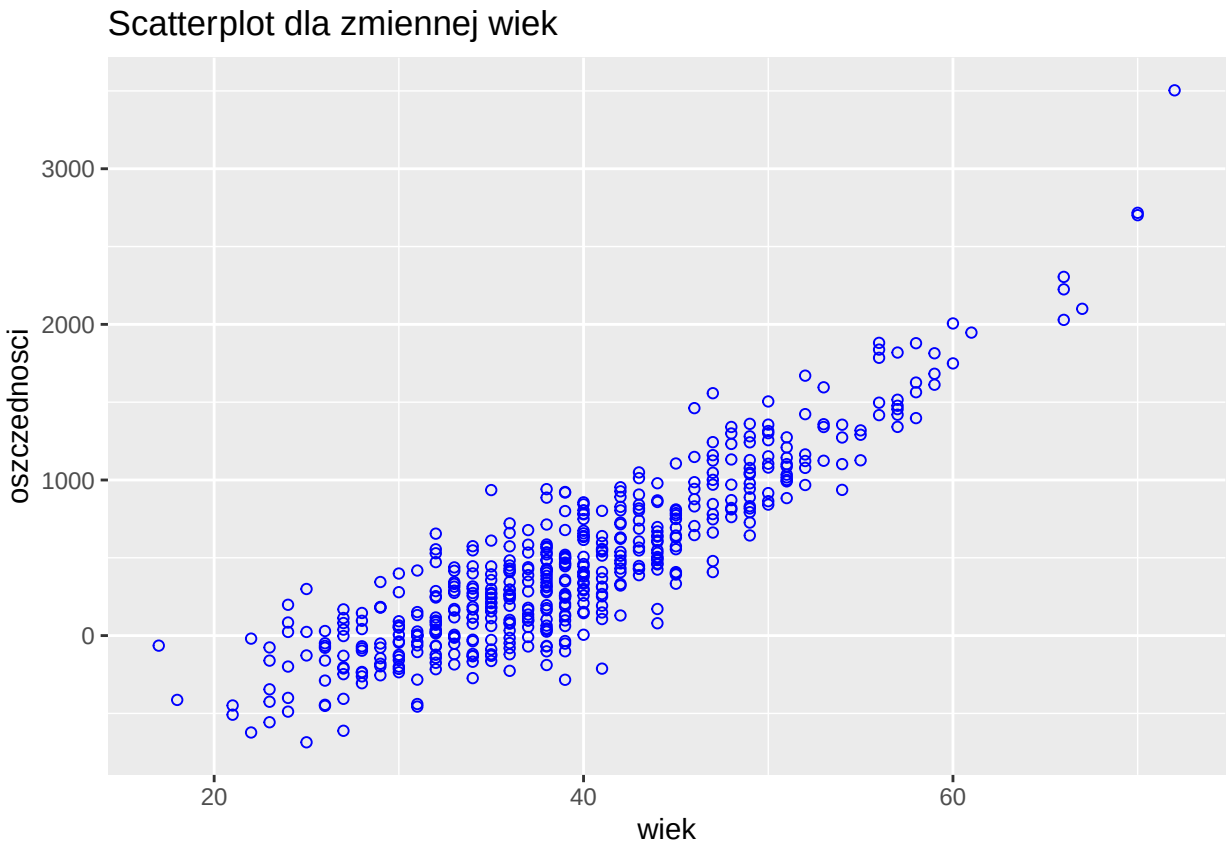

```
ggplot(data = df, aes(x = stan_cywilny, fill = plec)) +
  geom_bar(position = "dodge") +
  labs(x = "Stan cywilny", y = "Liczba osob", fill = "Plec")
```

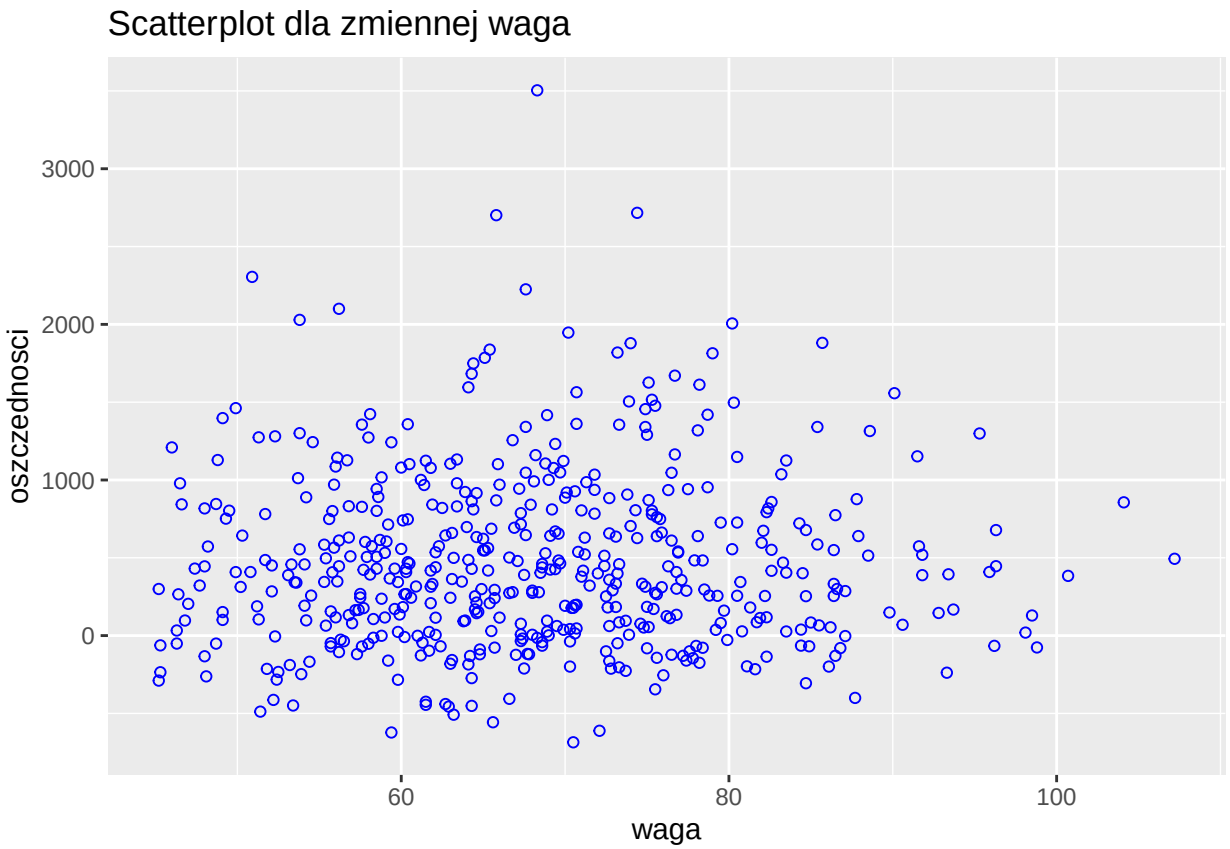


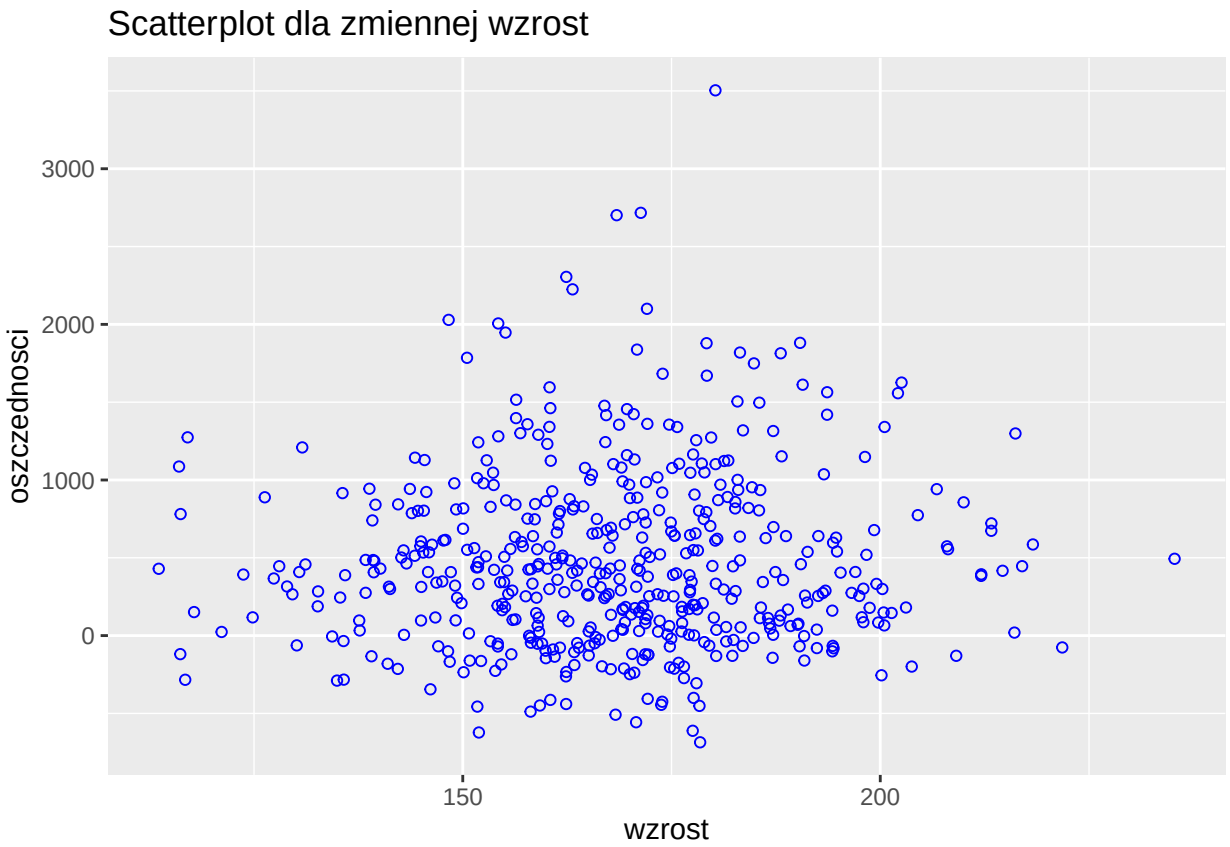
3. Podsumowanie danych przynajmniej trzema różnymi wykresami

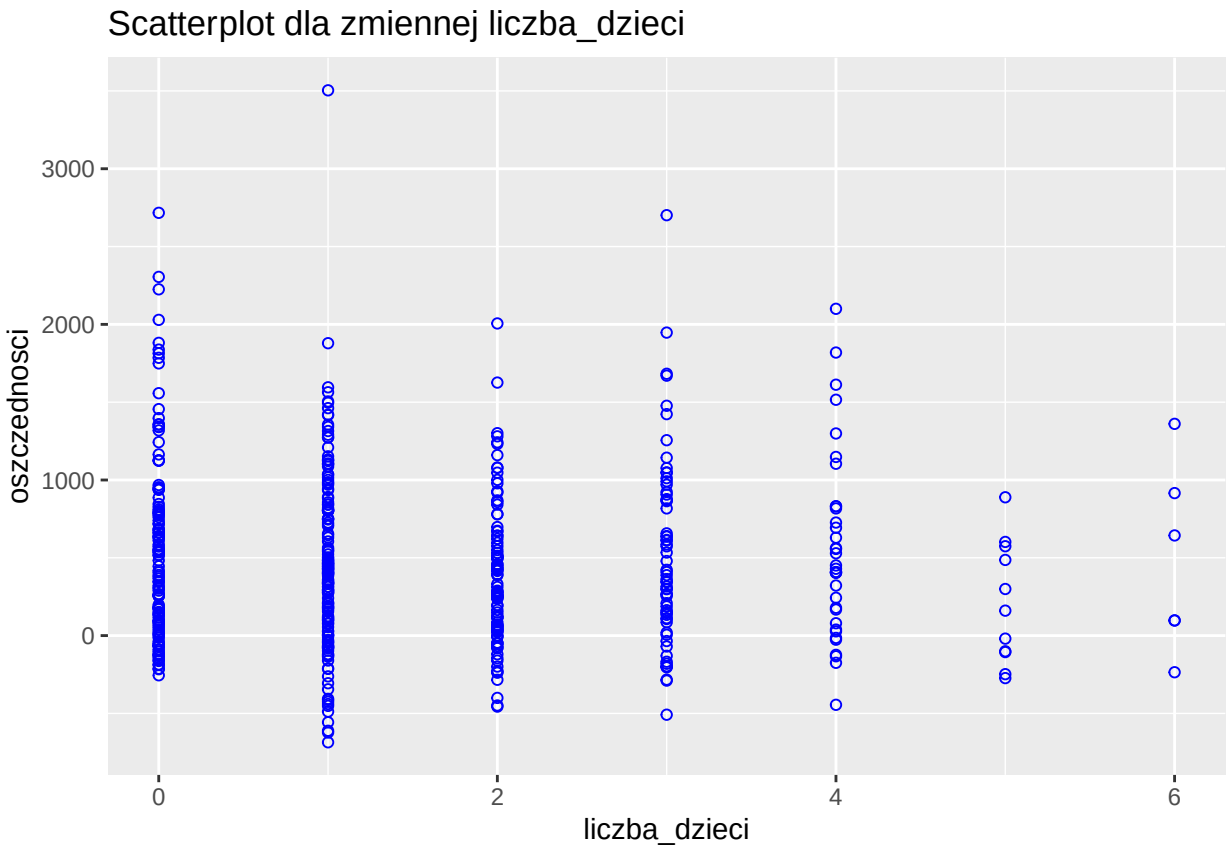
Część dodatkowych wykresów została wykonana powyżej. Teraz zajmę się wygenerowaniem wykresów obowiązkowych.

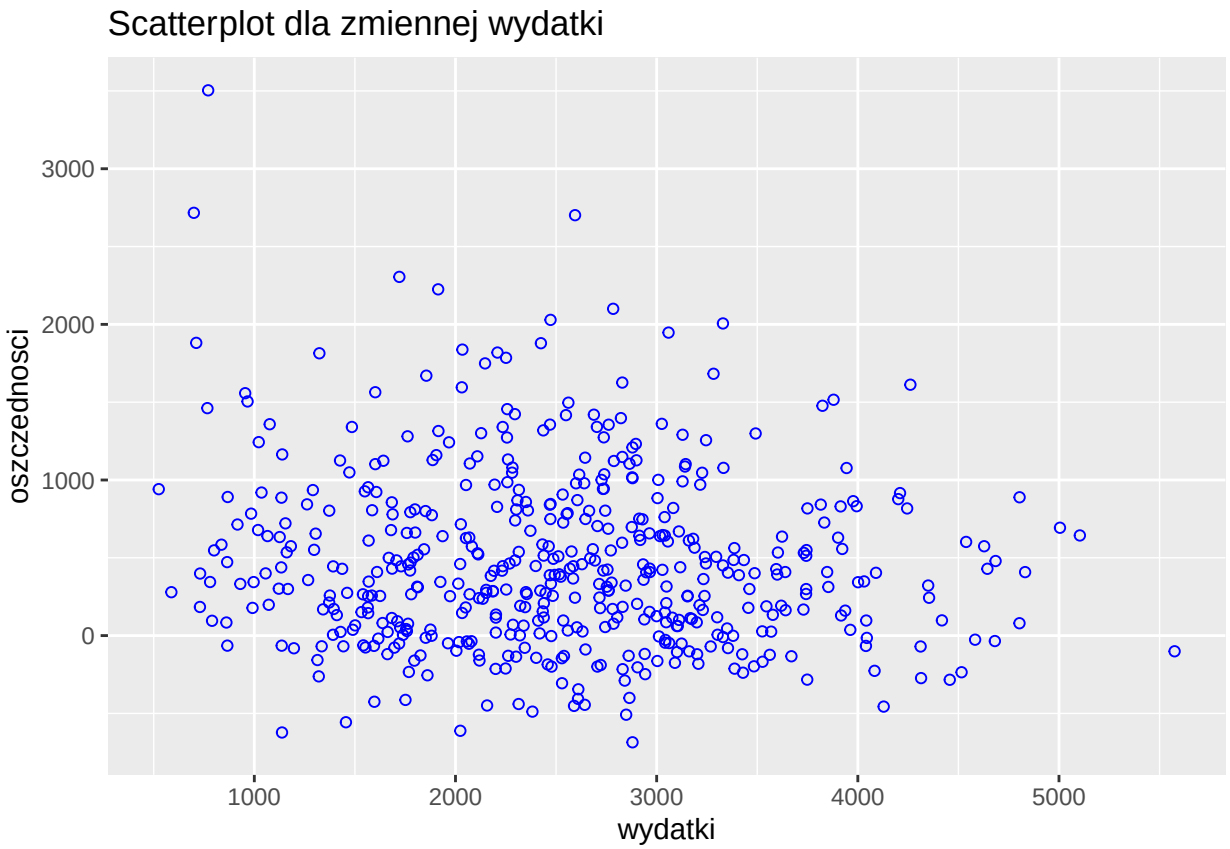
```
# scatter-plot
ilosciowe_vars <- names(ilosciowe)

for (i in seq(along = ilosciowe_vars)[-7]) {
  print(ggplot(data = ilosciowe, aes(x = !!sym(ilosciowe_vars[i]), y = oszczednosci)) +
          geom_point(shape = 1, color = "blue") +
          labs(title = paste("Scatterplot dla zmiennej", ilosciowe_vars[i])))
}
```

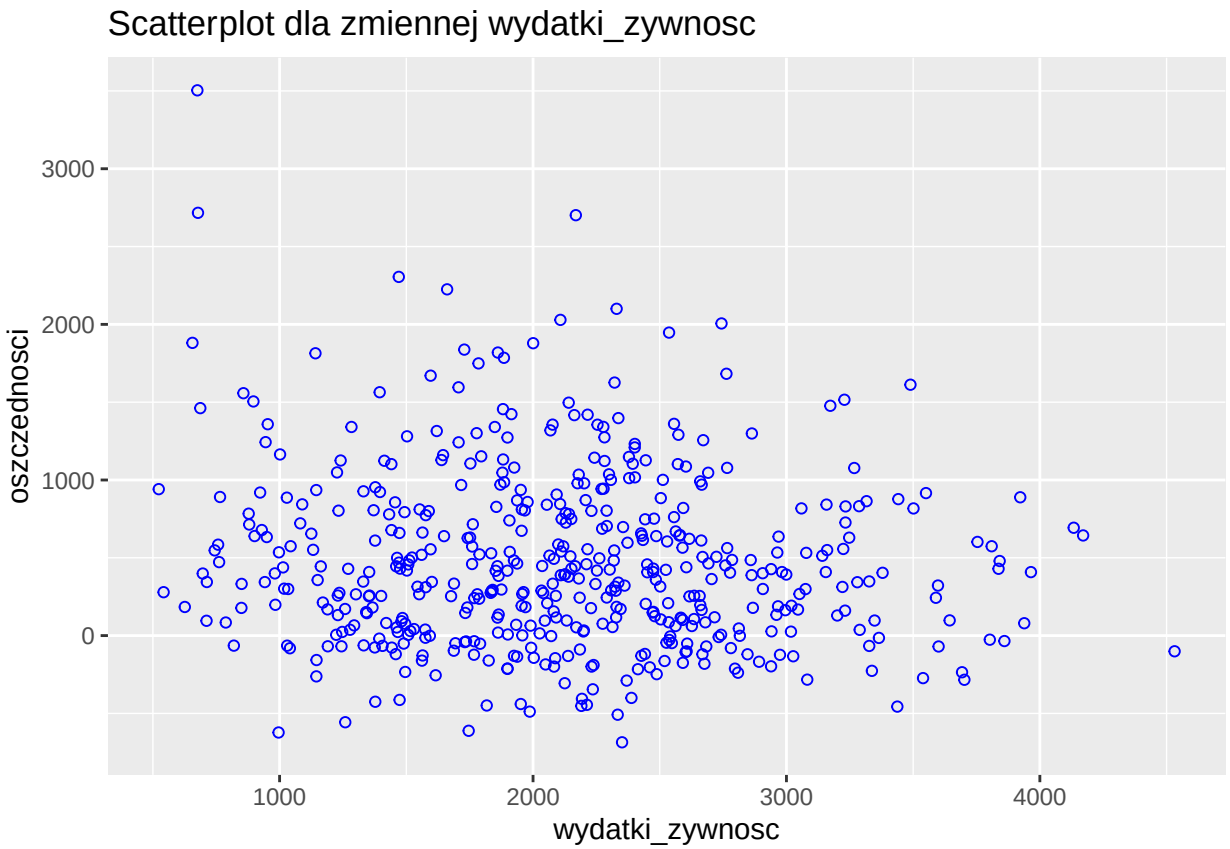

Scatterplot dla zmiennej wiek
oszczednosci
3000
2000
1000
0
20
40
60
wiek


Scatterplot dla zmiennej waga
oszczednosci
3000
2000
1000
0
60
80
100
waga

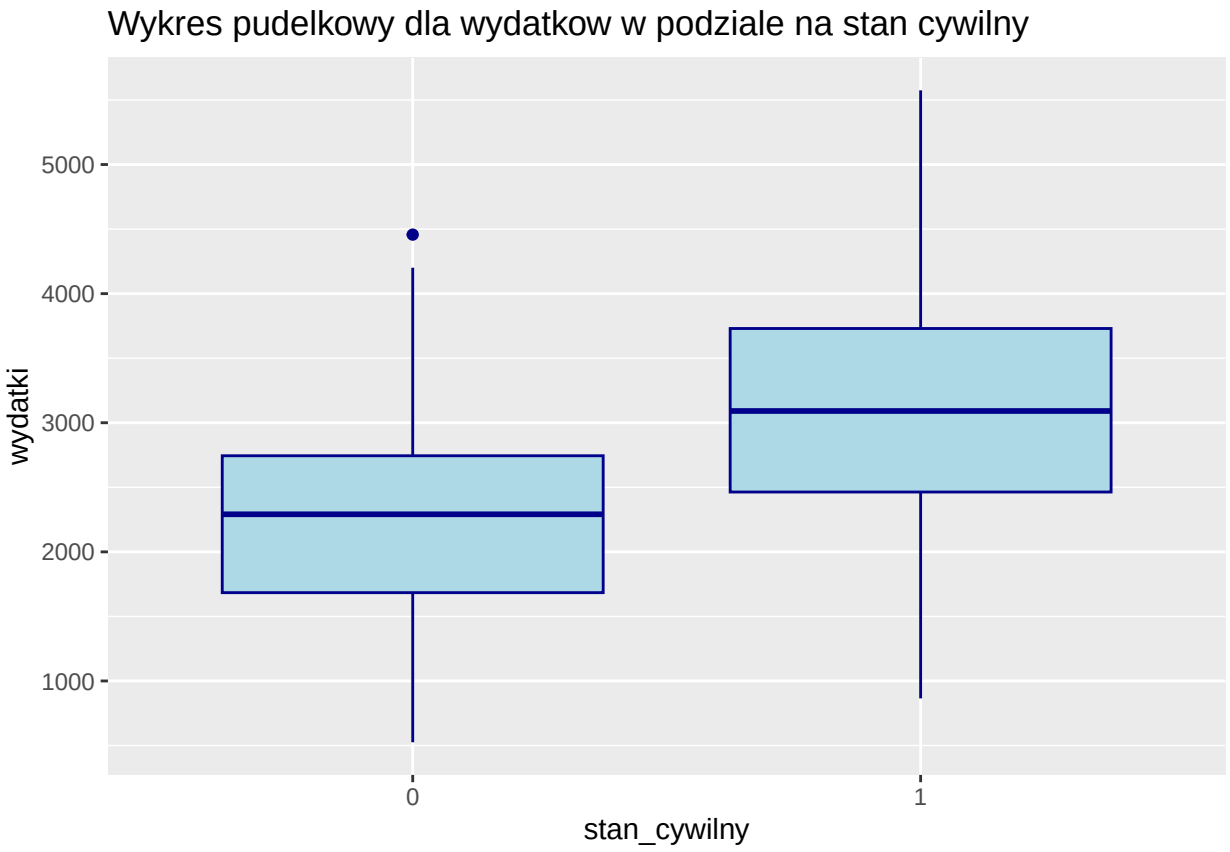

Scatterplot dla zmiennej wzrost
oszczednosci
3000
2000
1000
0
150
200
wzrost


Scatterplot dla zmiennej liczba_dzieci
oszczednosci
3000
2000
1000
0
0
2
4
6
liczba_dzieci

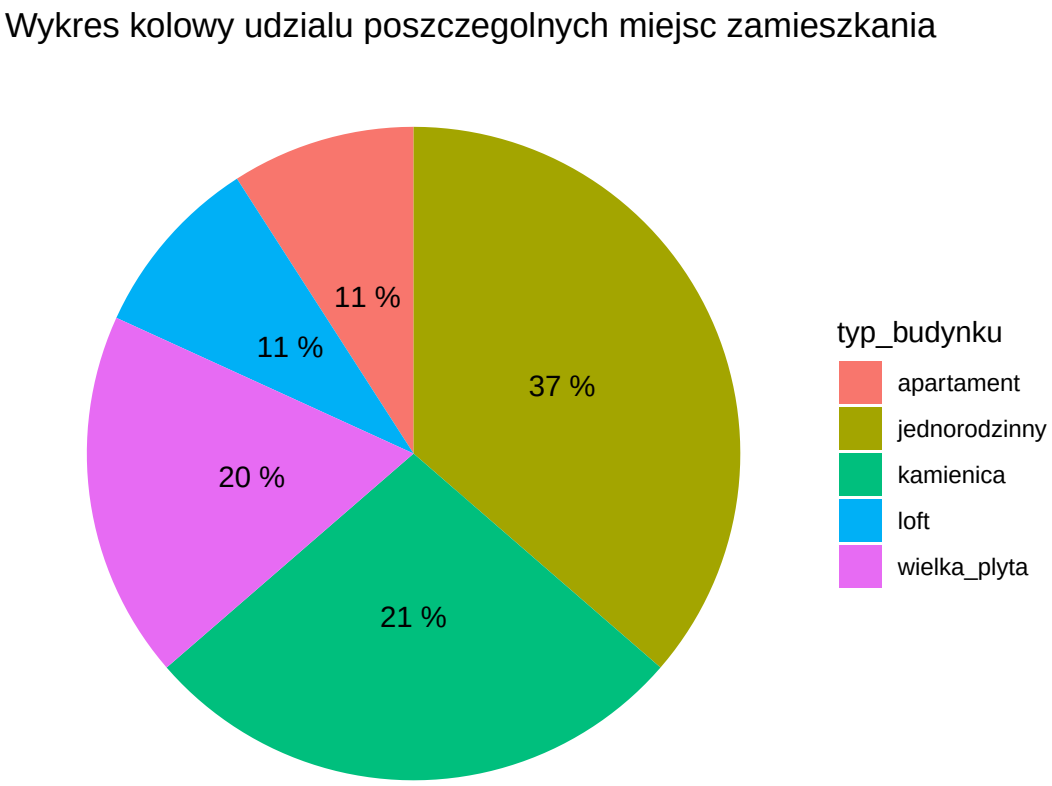

Scatterplot dla zmiennej wydatki
oszczednosci
3000
2000
1000
0
1000
2000
3000
4000
5000
wydatki




```
# boxplot - dla wydatków w podziale na stan cywilny
ggplot(data = df, aes(x = stan_cywilny, y = wydatki)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Wykres pudelkowy dla wydatkow w podziale na stan cywilny")
```




```
# kołowy - dla budynków
freq_table <- table(df$budynek)
freq_table_prop <- as.numeric(round(prop.table(freq_table) * 100))
prec <- paste(freq_table_prop, "%")
df_kolowy <- data.frame(freq_table, prec)
colnames(df_kolowy) <- c("typ_budynku", "liczebnosc", "procent")

ggplot(df_kolowy, aes(x = "", y = procent, fill = typ_budynku)) +
  geom_col() +
  geom_text(aes(label = procent), position = position_stack(vjust = 0.5), show.legend = FALSE) +
  coord_polar(theta = "y", clip = "off") + theme_void() +
        labs(title = "Wykres kolowy udzialu poszczegolnych miejsc zamieszkania")
```



4. Hipotezy o średniej i medianie wagi

Polecenie: Policzyć p-wartości dla hipotez o wartości średniej m=70kg i medianie me=65kg dla zmiennej waga, osobno na próbach dla kobiet i mężczyzn, przyjąć statystykę testową dla alternatywy lewostronnej.

```
women <- na.omit(df[df$plec == "K", ])
men <- na.omit(df[df$plec == "M", ])
# mamy próby > 30 każda
# sprawdzamy normalność tych danych:

# H0: rozkład X jest normalny
# H1: rozkład X jest różny od normalnego
# alpha = 0.01 --> chcemy odrzucać normalność danych tylko jeśli przesłanka jest bardzo silna
shapiro.test(women$waga)
```

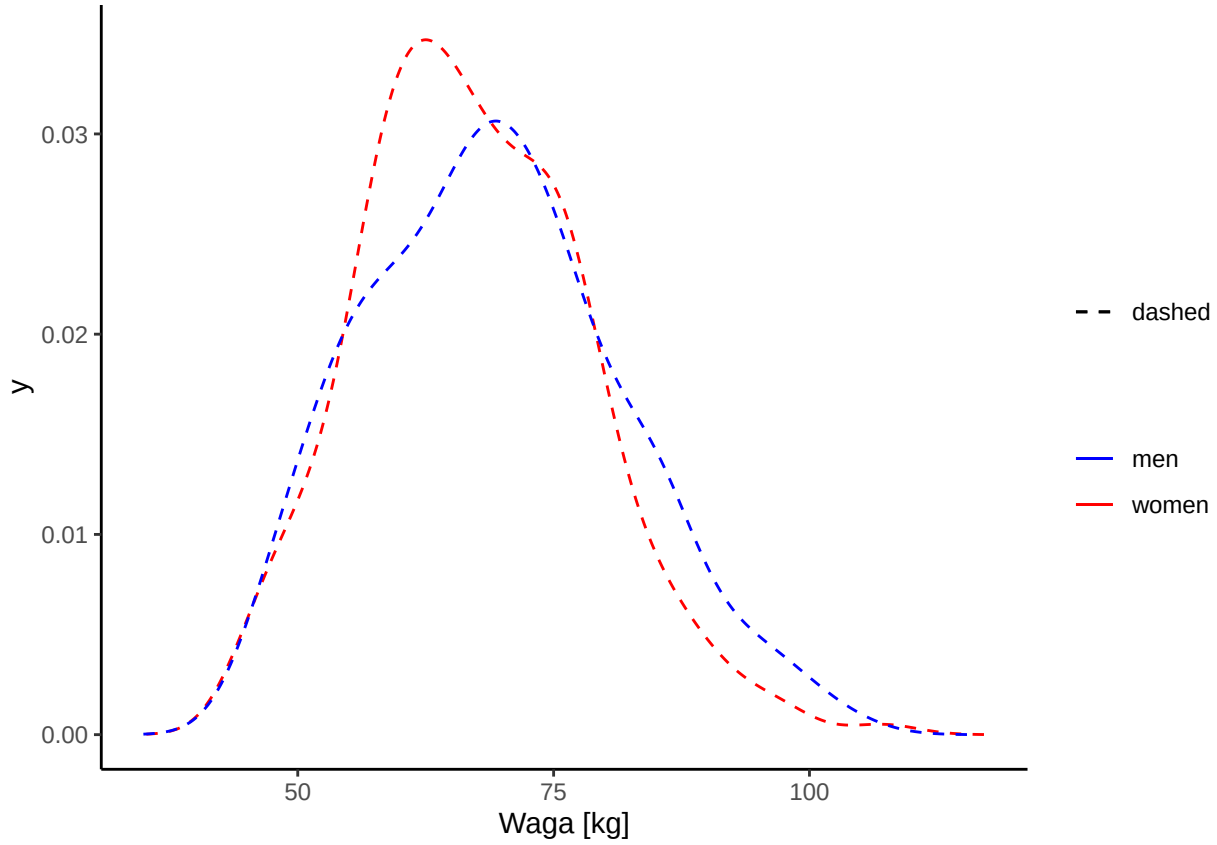
```
##
##  Shapiro-Wilk normality test
##
## data:  women$waga
## W = 0.98663, p-value = 0.02528
```

```
# p > alpha -> brak podstaw do odrzucenia H0

shapiro.test(men$waga)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  men$waga
## W = 0.98463, p-value = 0.01617
```

```r
# p > alpha -> brak podstaw do odrzucenia H0

# poniżej dodatkowo przygotowany wykres gęstości:
density_women <- density(women$waga, na.rm = TRUE)
density_men <- density(men$waga, na.rm = TRUE)
ggplot() +
  geom_line(data = data.frame(x = density_women$x, y = density_women$y, sex = "women"),
            aes(x, y, color = sex, linetype = "dashed")) +
  geom_line(data = data.frame(x = density_men$x, y = density_men$y, sex = "men"),
            aes(x, y, color = sex, linetype = "dashed")) +
  scale_color_manual(values = c("women" = "red", "men" = "blue")) +
  scale_linetype_manual(values = c("dashed", "dashed")) +
  labs(x = "Waga [kg]", color = "", linetype = "") +
  theme_classic()
```




```r
# Test dla wartości średniej
# mamy dane z rozkładu normalnego o nieznanych średniej i wariancji
# założenie jest uprawnione na mocy powyższego testu normalności
# H0: m = 70 kg
# H1: m < 70 kg
# alpha = 0.05 - uprawione założenie, standardowa wartość alpha
# a zatem stosujemy test t-studenta
```

```r
# Test średniej dla kobiet
t.test(women$waga, mu = 70, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  women$waga
## t = -3.8787, df = 237, p-value = 6.804e-05
## alternative hypothesis: true mean is less than 70
## 95 percent confidence interval:
##      -Inf 68.40365
## sample estimates:
## mean of x
##  67.22017
```

```r
# p < alpha --> przyjmujemy H1

# Test średniej dla mężczyzn
t.test(men$waga, mu = 70, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  men$waga
## t = -1.0265, df = 222, p-value = 0.1529
## alternative hypothesis: true mean is less than 70
## 95 percent confidence interval:
##      -Inf 70.52058
## sample estimates:
## mean of x
##  69.14529
```

```r
# p > alpha --> brak podstaw do odrzucenia H0

# Test dla mediany
# do testowania mediany korzystamy z testów nieparametrycznych
# (nie ma znaczenia rozkład danych)
# H0: me = 65 kg
# H1: me < 65 kg
# alpha = 0.05 - uprawione założenie, standardowa wartość alpha

# Test mediany dla kobiet
wilcox.test(women$waga, mu = 65, alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  women$waga
## V = 16834, p-value = 0.9952
## alternative hypothesis: true location is less than 65
```

```r
# p > alpha --> brak podstaw do odrzucenia H0

# Test mediany dla mężczyzn
wilcox.test(men$waga, mu = 65, alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  men$waga
## V = 16620, p-value = 1
## alternative hypothesis: true location is less than 65
```

```
# p > alpha --> brak podstaw do odrzucenia H0
```

5. Przedziały ufności

Polecenie: Policzyć dwustronne przedziały ufności na poziomie ufności 0.99 dla zmiennej wiek dla następujących parametrów rozkładu:

- średnia i odchylenie standardowe;
- kwantyle 1/4, 2/4 i 3/4.

```
# Średnia
mean_age <- mean(df$wiek)
sd_age <- sd(df$wiek)

# wartość kwantyla t-studenta dla danego poziomu ufności i liczności próby
n <- length(df$wiek)
t_critical <- qt(0.995, n-1) # 0.995 -> bo dwustronny (z obu stron po 0.005)

# granice przedziału ufności
margin_of_error <- t_critical * sd_age / sqrt(n)
lower_limit <- mean_age - margin_of_error
upper_limit <- mean_age + margin_of_error

cat("Dwustronny przedział ufności dla średniej wieku na poziomie ufności 0.99:\n",
    paste0("[", round(lower_limit, 2), ", ", round(upper_limit, 2), "]"))
```

```
## Dwustronny przedział ufności dla średniej wieku na poziomie ufności 0.99:
##  [38.43, 40.51]
```

```
# Odchylenie standardowe
lower_quantile <- qchisq(0.005, n-1)
upper_quantile <- qchisq(0.995, n-1)

margin_of_error_lower <- sqrt((n-1)*sd_age^2/lower_quantile)
margin_of_error_upper <- sqrt((n-1)*sd_age^2/upper_quantile)
lower_limit <- sd_age - margin_of_error_lower
upper_limit <- sd_age + margin_of_error_upper

cat("Dwustronny przedział ufności dla odchylenia standardowego wieku
na poziomie ufności 0.99:\n",
    paste0("[", round(lower_limit, 2), ", ", round(upper_limit, 2), "]"))
```

```
## Dwustronny przedział ufności dla odchylenia standardowego wieku
## na poziomie ufności 0.99:
##  [-0.79, 17.27]
```

```
# Kwantyle
library(EnvStats)
```

```
##
```

```
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm

## The following object is masked from 'package:base':
##
##     print.default
```

```
cat("Przedział ufności dla pierwszego kwartyla (0.25) na poziomie ufności 0.99 wynosi")
```

```
## Przedział ufności dla pierwszego kwartyla (0.25) na poziomie ufności 0.99 wynosi
```

```
as.numeric(EnvStats::eqnpar(x=df$wiek, p=0.25, ci=TRUE, ci.method="exact",
                            approx.conf.level=0.99)$interval$limits)
```

```
## [1] 32 35
```

```
cat("Przedział ufności dla drugiego kwartyla (0.5) na poziomie ufności 0.99 wynosi")
```

```
## Przedział ufności dla drugiego kwartyla (0.5) na poziomie ufności 0.99 wynosi
```

```
as.numeric(EnvStats::eqnpar(x=df$wiek, p=0.5, ci=TRUE, ci.method="exact",
                            approx.conf.level=0.99)$interval$limits)
```

```
## [1] 38 40
```

```
cat("Przedział ufności dla trzeciego kwartyla (0.75) na poziomie ufności 0.99 wynosi")
```

```
## Przedział ufności dla trzeciego kwartyla (0.75) na poziomie ufności 0.99 wynosi
```

```
as.numeric(EnvStats::eqnpar(x=df$wiek, p=0.75, ci=TRUE, ci.method="exact",
                            approx.conf.level=0.99)$interval$limits)
```

```
## [1] 43 47
```

Przyjęte założenia:

- Dane pochodzą z próby losowej i są niezależne od siebie - uzasadnione, brak podstaw aby zakładać ich zależność;
- rozkład zmiennej w próbie jest mniej więcej normalny lub przyjmujemy, że próba jest wystarczająco liczna, aby na mocy Centralnego Twierdzenia Granicznego zbiegała do rozkładu normalnego (w naszym przypadku zmienna wiek nie ma rozkładu normalnego, ale zakładam, że 499 obserwacji to wystarczająco dużo, aby powołać sie na CTG).

6. Pytania badawcze

1. Czy istnieją różnice w średnich wartościach wybranej zmiennej pomiędzy osobami zamężnymi/żonatymi a pannami/kawalerami w podpróbie osób w wieku poniżej 40 lat?

```
# H0: Średni wydatki osób poniżej 40 r.ż. w zależności od stanu cywilnego
# nie różnią się istotnie statystycznie.
# H1: Średnie wydatki osób poniżej 40 r.ż. w zależności od stanu cywilnego
# różnią się istotnie statystycznie.
# alpha = 0.01

ponizej_40 <- df[df$wiek < 40, ]
stan_wolny <- ponizej_40[ponizej_40$stan_cywilny == "0",]
w_malzenstwie <- ponizej_40[ponizej_40$stan_cywilny == "1",]
```

```r
# Najpierw sprawdzamy czy te dane mają rozkład normalny
# (to determinuje wybór odpowiedniego testu wartości średnich dla dwóch grup)
# H0: rozkład jest normalny
# H1: rozkład jest różny od normalnego
# alpha: 0.01
shapiro.test(stan_wolny$wydatki)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stan_wolny$wydatki
## W = 0.98746, p-value = 0.1082
```

```r
# p > alpha --> brak podstaw do odrzucenia H0

shapiro.test(w_malzenstwie$wydatki)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  w_malzenstwie$wydatki
## W = 0.9835, p-value = 0.2913
```

```r
# p > alpha --> brak podstaw do odrzucenia H0

# A zatem uzasadnione jest zakładanie, że mamy dwie grupy o rozkładzie normalnym,
# możemy korzystać z testu t-studenta dla grup niezależnych (zakładamy, że
# osoby, które są w związku małżeńskim oraz osoby, które są w stanie wolnym,
# nie są ze sobą powiązane w sposób umyślny).

t.test(stan_wolny$wydatki, w_malzenstwie$wydatki,
                alternative = "two.sided", paired = FALSE, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  stan_wolny$wydatki and w_malzenstwie$wydatki
## t = -6.7308, df = 157.79, p-value = 2.957e-10
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -1069.9414  -472.4523
## sample estimates:
## mean of x mean of y
##  2109.180  2880.377
```

p-wartość jest mniejsza niż alpha, stąd odrzucamy H0 i przyjmujemy H1. Zatem istnieje istotna statystycznie różnica w wydatkach między analizowanymi grupami.

2. Czy w podpróbie osób w wieku poniżej 25 lat średnie wydatki ogółem są równe średnim wydatkom na żywność?

```r
# H0: Średnie wydatki ogółem są równe średnim wydatkom na żywność
# w grupie wiekowej poniżej 25 r.ż.
# H1: Średnie wydatki ogółem nie są równe średnim wydatkom na żywność
# w grupie wiekowej poniżej 25 r.ż.
```

```r
# alpha = 0.01

ponizej_25 <- df[df$wiek < 25,]

# sprawdzamy normalność dla tych zmiennych:
# H0: rozkład jest normalny
# H1: rozkład jest różny od normalnego
# alpha: 0.01
shapiro.test(ponizej_25$wydatki)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ponizej_25$wydatki
## W = 0.9254, p-value = 0.1822
```

```r
# p > alpha --> brak podstaw do odrzucenia H0
shapiro.test(ponizej_25$wydatki_zywnosc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ponizej_25$wydatki_zywnosc
## W = 0.92538, p-value = 0.182
```

```r
# p > alpha --> brak podstaw do odrzucenia H0

# sprawdzamy zależność zmiennych --> mamy rozkład normalny,
# więc stosujemy test do sprawdzenia zależności metodą Pearsona
# H0: brak zależności między zmiennymi
# H1: zmienne są zależne
# alpha = 0.01
cor.test(ponizej_25$wydatki_zywnosc, ponizej_25$wydatki, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  ponizej_25$wydatki_zywnosc and ponizej_25$wydatki
## t = 88.433, df = 15, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9972722 0.9996640
## sample estimates:
##       cor
## 0.9990423
```

```r
# p < alpha --> przyjmujemy H1

# A zatem na mocy testów uzasadnione jest zakładanie,
# że mamy dwie grupy zmiennych zależnych o rozkładzie normalnym
# korzystamy zatem z testu t-studenta dla grup zależnych

t.test(ponizej_25$wydatki, ponizej_25$wydatki_zywnosc, paired = TRUE,
       alternative = "two.sided", conf.level = 0.95)
```

```
##
```

```
##  Paired t-test
##
## data:  ponizej_25$wydatki and ponizej_25$wydatki_zywnosc
## t = 7.3764, df = 16, p-value = 1.563e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  185.8780 335.8055
## sample estimates:
## mean of the differences
##                260.8418
```

p-wartość jest mniejsza niż alpha, stąd odrzucamy H0 i przyjmujemy H1. Istnieją zatem istotne statystycznie różnice między średnimi wydatkami ogółem a wydatkami na żywność w grupie osób poniżej 25 r.ż.

3. Czy niższy udział wydatków na żywność w wydatkach ogółem jest skorelowany z wyższymi oszczędnościami?

```
# H0: brak korelacji między oszczędnościami a wydatkami na żywność
# H1: istnieje korelacja (dodatnia bądź ujemna) między oszczędnościami a wydatkami na żywność
# alpha = 0.01

# zakładamy brak rozkładu normalnego zmiennych (na mocy testu na początku projektu)
# zatem korzystamy z testu korelacji rang Spearmana
cor.test(df$wydatki_zywnosc, df$oszczednosci, method = "spearman", conf.level = 0.99)
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$wydatki_zywnosc and df$oszczednosci
## S = 22309116, p-value = 0.08456
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.07729271
```

Wynik testu korelacji rangowej Spearmana wskazuje na słabą, ujemną korelację między wydatkami na żywność a oszczędnościami. Wartość współczynnika korelacji wynosi -0.077, co oznacza, że osoby, które wydają mniej na jedzenie, zwykle mają nieznacznie wyższe oszczędności. Jednakże, wynik testu nie jest istotny na poziomie istotności alpha=0.01 (p-value = $0.08456 > 0.01$), co oznacza, że nie ma wystarczających dowodów, aby odrzucić H0 o braku korelacji między zmiennymi.

4. Przetestuj hipotezę o zgodności z konkretnym rozkładem parametrycznym dla wybranej zmiennej (np. "zmienna A ma rozkład wykładniczy z parametrem 10")

```
# H0: rozkład zmiennej oszczędności jest rozkładem wykładniczym
# z parametrem lambda = 1/mean(oszczędności)
# H1: rozkład zmiennej oszczędności jest różny od wykładniczego
# z parametrem lambda = 1/mean(oszczędności)
# alpha: 0.01
# wybieramy test Kołmogorowa-Smirnova a nie Pearsona, bo mamy zmienną typu ciągłego

ks.test(df$oszczednosci, "pexp", rate = 1/mean(df$oszczednosci))
```

```
##
##  One-sample Kolmogorov-Smirnov test
```

```
##
## data:  df$oszczednosci
## D = 0.19639, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

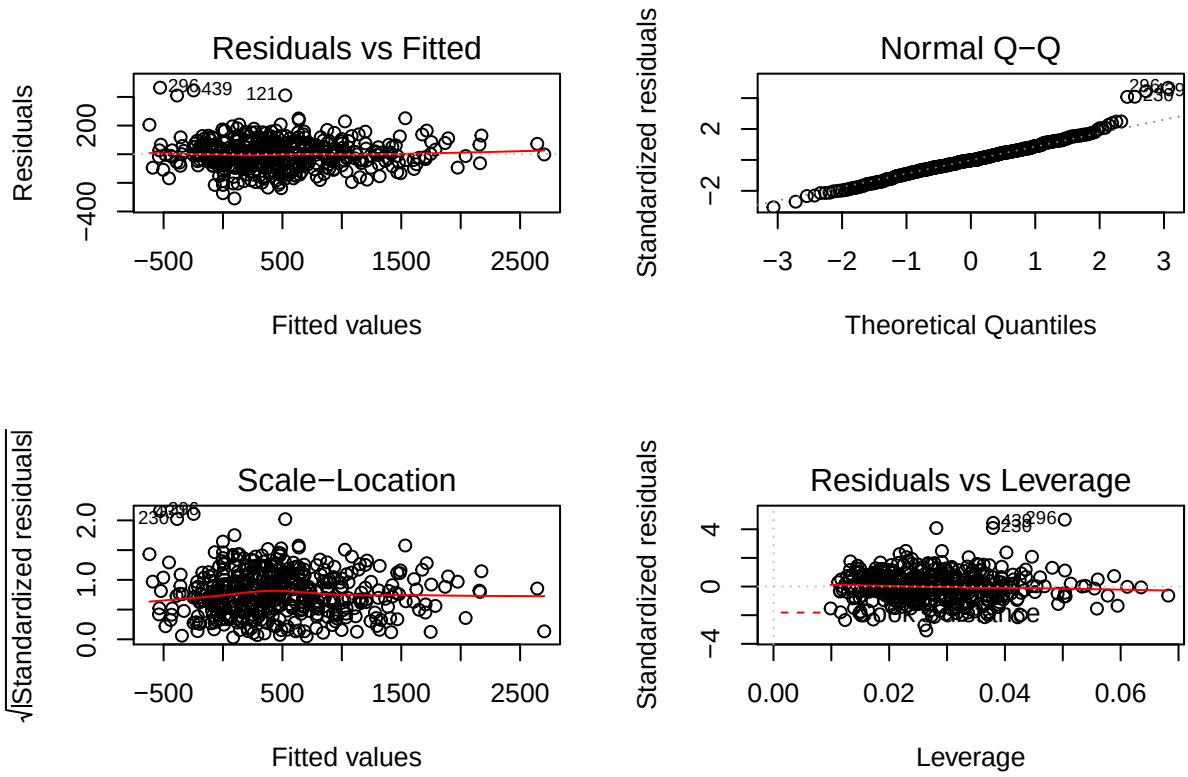
A zatem p wartość jest mniejsza od alpha, więc istnieją podstawy do odrzucenia H0 i przyjęcia H1. A zatem nie ma podstaw aby sądzić, że zmienna oszczędności ma rozkład wykładniczy z parametrem lambda = 1/mean(oszczędności).

7. Regresja liniowa

```r
library(stats)

# Oszacowanie pełnego modelu regresji liniowej
model <- lm(oszczednosci ~ ., data=df)
summary(model)
```

```
##
## Call:
## lm(formula = oszczednosci ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -309.00  -60.73   -1.89   57.94  464.80
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -859.7620    61.8744 -13.895  < 2e-16 ***
## wiek                    63.9407     0.5698 112.210  < 2e-16 ***
## waga                     3.9748     0.5715   6.955 1.25e-11 ***
## wzrost                  -2.4067     0.3538  -6.803 3.29e-11 ***
## plecM                    1.1701     9.6545   0.121   0.9036
## stan_cywilny1           -4.6099    12.9319  -0.356   0.7217
## liczba_dzieci          151.6980     6.1679  24.595  < 2e-16 ***
## budynekjednorodzinny -181.6680    16.4765 -11.026  < 2e-16 ***
## budynekkamienica     -305.3630    17.9242 -17.036  < 2e-16 ***
## budynekloft          -337.1887    25.2324 -13.363  < 2e-16 ***
## budynekwielka_plyta  -563.4955    20.6933 -27.231  < 2e-16 ***
## wydatki                 -0.2977     0.1497  -1.988   0.0474 *
## wydatki_zywnosc         -0.1229     0.1863  -0.660   0.5097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.1 on 448 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9662
## F-statistic:  1097 on 12 and 448 DF,  p-value: < 2.2e-16
```

```r
# wyświetlenie diagramów diagnostycznych w celu stwierdzenia, czy konieczne są transformacje
par(mfrow=c(2,2))
plot(model)
```



Na podstawie wykresów możemy uznać, że transformacje nie są konieczne:

- wykres 1 (residuals vs fitted): mamy mniej więcej równomierne rozłożenie punktów wokół linii, zatem warunek liniowości jest spełniony.
- wykres 2 (normal Q-Q): punkty w większości są dobrze dopasowane do teoretycznego rozkładu normalnego.
- wykres 3 (scale-location): homoskedadyczność jest zachowywana.
- wykres 4 (residuals vs leverage): brak widocznych dźwigni w danych.

```
# wykres regresji dla pełnego modelu
ggplot(na.omit(df), aes(x=oszczednosci, y=predict(model))) +
  geom_point(col='red', shape=1) +
  geom_smooth(method="lm") +
  xlab("Oszczednosci") +
  ylab("Przewidywane oszczednosci")
```

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

Wartości współczynnika determinacji Multiple R-squared i Adjusted R-squared opisują, jak wiele zmienności zmiennej zależnej jest wyjaśniane przez model. Multiple R-squared wynosi 0,9671, co oznacza, że 96,71% zmienności zmiennej oszczędności jest wyjaśnione przez zmienne objaśniające. Adjusted R-squared jest zbliżony do Multiple R-squared, co oznacza, że dodanie nowych zmiennych do modelu nie poprawia jego jakości.

F-statistic i p-wartość testu F opisują jakość dopasowania modelu. Wartość F-statistic jest wyższa od 1, co oznacza, że model ma lepsze dopasowanie niż model zerowy (bez zmiennych objaśniających). P-wartość testu F wynosi $<$ 2,2e-16, co oznacza, że istnieje istotna zależność między zmiennymi objaśniającymi a zmienną objaśnianą.

Residual standard error opisuje jak dobrze model dopasowuje się do danych, a wartość wynosząca 102.1 oznacza, że błąd resztowy ma przeciętną wartość 102.1.

Podsumowując:

- RSS wynosi 102.1.
- Multiple R-squared wynosi 0.9671, a Adjusted R-squared wynosi 0.9662.
- p-wartości dla poszczególnych współczynników można znaleźć w kolumnie "Pr(>|t|)" w podsumowaniu modelu.
- Oszacowania współczynników znajdują się w kolumnie "Estimate" w podsumowaniu modelu.

```
# sprawdzamy jak zmieni się model gdy usuniemy kolejno każdą ze zmiennych:
model_bez_wzrostu <- lm(oszczednosci ~ .-wzrost, data=df)
summary(model_bez_wzrostu)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - wzrost, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -335.85  -65.34   -2.72   67.22  506.22
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.122e+03  5.073e+01 -22.123   <2e-16 ***
## wiek                  6.361e+01  5.957e-01 106.780   <2e-16 ***
## waga                  1.331e+00  4.397e-01   3.028   0.0026 **
## plecM                 2.139e+00  1.013e+01   0.211   0.8329
## stan_cywilny1        -4.087e+00  1.357e+01  -0.301   0.7634
## liczba_dzieci         1.434e+02  6.344e+00  22.605   <2e-16 ***
## budynekjednorodzinny -1.847e+02  1.728e+01 -10.689   <2e-16 ***
## budynekkamienica     -3.065e+02  1.881e+01 -16.296   <2e-16 ***
## budynekloft          -3.115e+02  2.618e+01 -11.900   <2e-16 ***
## budynekwielka_plyta  -5.485e+02  2.159e+01 -25.407   <2e-16 ***
## wydatki              -3.260e-01  1.570e-01  -2.076   0.0385 *
## wydatki_zywnosc      -6.162e-02  1.952e-01  -0.316   0.7524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 107.2 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9637, Adjusted R-squared:  0.9628
## F-statistic:  1083 on 11 and 449 DF,  p-value: < 2.2e-16
```

```
model_bez_wagi <- lm(oszczednosci ~ .-waga, data=df)
summary(model_bez_wagi)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - waga, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -306.99  -60.55   -0.31   62.49  527.23
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -873.36025   65.02431 -13.431   <2e-16 ***
## wiek                   63.75211    0.59846 106.527   <2e-16 ***
## wzrost                 -0.73379    0.27275  -2.690   0.0074 **
## plecM                   5.20915   10.13263   0.514   0.6074
## stan_cywilny1          -1.59395   13.58939  -0.117   0.9067
## liczba_dzieci         149.26811    6.47469  23.054   <2e-16 ***
## budynekjednorodzinny -186.07221   17.31110 -10.749   <2e-16 ***
## budynekkamienica     -309.17808   18.83722 -16.413   <2e-16 ***
## budynekloft          -339.25116   26.52830 -12.788   <2e-16 ***
## budynekwielka_plyta  -566.67079   21.75235 -26.051   <2e-16 ***
## wydatki                -0.35243    0.15722  -2.242   0.0255 *
## wydatki_zywnosc        -0.05097    0.19554  -0.261   0.7945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 107.4 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9635, Adjusted R-squared:  0.9626
## F-statistic:  1079 on 11 and 449 DF,  p-value: < 2.2e-16
```

```
model_bez_wieku <- lm(oszczednosci ~.-wiek, data=df)
summary(model_bez_wieku)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - wiek, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1213.24  -399.58   -84.31   289.62  2293.11
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           454.5647   327.4054   1.388    0.166
## waga                    0.9238     3.0762   0.300    0.764
## wzrost                  0.9493     1.8996   0.500    0.617
## plecM                 -14.2992    52.0215  -0.275    0.784
## stan_cywilny1          22.0568    69.6769   0.317    0.752
## liczba_dzieci         -10.5574    32.3116  -0.327    0.744
## budynekjednorodzinny -124.5161    88.7474  -1.403    0.161
## budynekkamienica     -133.2656    96.2371  -1.385    0.167
## budynekloft           214.2572   133.3705   1.606    0.109
## budynekwielka_plyta  -126.8984   109.5251  -1.159    0.247
## wydatki                 0.9273     0.8047   1.152    0.250
## wydatki_zywnosc        -1.1565     1.0025  -1.154    0.249
##
## Residual standard error: 550.5 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.04207,    Adjusted R-squared:  0.0186
## F-statistic: 1.793 on 11 and 449 DF,  p-value: 0.05284
```

```
model_bez_wydatkow <- lm(oszczednosci ~.-wydatki, data=df)
summary(model_bez_wydatkow)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - wydatki, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -315.83  -58.70   -4.23   59.21  472.30
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -826.86082   59.81584 -13.823  < 2e-16 ***
## wiek                   63.85808    0.57018 111.996  < 2e-16 ***
## waga                    4.03458    0.57258   7.046 6.95e-12 ***
## wzrost                 -2.42627    0.35479  -6.839 2.62e-11 ***
## plecM                   1.15085    9.68613   0.119    0.905
## stan_cywilny1          -4.49720   12.97423  -0.347    0.729
```

```
## liczba_dzieci         151.18206    6.18263  24.453  < 2e-16 ***
## budynekjednorodzinny -180.48702   16.51980 -10.925  < 2e-16 ***
## budynekkamienica     -304.07323   17.97121 -16.920  < 2e-16 ***
## budynekloft          -333.08070   25.23015 -13.202  < 2e-16 ***
## budynekwielka_plyta  -558.94653   20.63394 -27.089  < 2e-16 ***
## wydatki_zywnosc        -0.49229    0.01324 -37.186  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.5 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.966
## F-statistic:  1189 on 11 and 449 DF,  p-value: < 2.2e-16
```

```
model_bez_wydatki_na_zywnosc <- lm(oszczednosci ~ .-wydatki_zywnosc, data=df)
summary(model_bez_wydatki_na_zywnosc)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - wydatki_zywnosc, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -307.30  -60.34   -1.89   58.66  462.15
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -872.2158    58.8886 -14.811  < 2e-16 ***
## wiek                   63.9593     0.5688 112.451  < 2e-16 ***
## waga                    3.9539     0.5703   6.934 1.44e-11 ***
## wzrost                 -2.3954     0.3531  -6.783 3.71e-11 ***
## plecM                   1.1997     9.6483   0.124    0.901
## stan_cywilny1          -4.6578    12.9236  -0.360    0.719
## liczba_dzieci         151.6669     6.1638  24.606  < 2e-16 ***
## budynekjednorodzinny -182.0788    16.4544 -11.066  < 2e-16 ***
## budynekkamienica     -305.6845    17.9063 -17.071  < 2e-16 ***
## budynekloft          -337.8974    25.1936 -13.412  < 2e-16 ***
## budynekwielka_plyta  -564.5422    20.6195 -27.379  < 2e-16 ***
## wydatki                -0.3962     0.0106 -37.378  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.1 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9662
## F-statistic:  1198 on 11 and 449 DF,  p-value: < 2.2e-16
```

```
model_bez_liczby_dzieci <- lm(oszczednosci ~.-liczba_dzieci, data=df)
summary(model_bez_liczby_dzieci)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - liczba_dzieci, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -598.22 -107.87    7.00   99.01  468.41
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.332e+03  9.007e+01 -14.788  < 2e-16 ***
## wiek                  6.066e+01  8.483e-01  71.503  < 2e-16 ***
## waga                  3.179e+00  8.738e-01   3.638 0.000307 ***
## wzrost               -6.856e-01  5.310e-01  -1.291 0.197305
## plecM                 2.548e+01  1.471e+01   1.733 0.083824 .
## stan_cywilny1         1.332e+02  1.785e+01   7.466 4.33e-13 ***
## budynekjednorodzinny -1.767e+02  2.523e+01  -7.005 9.06e-12 ***
## budynekkamienica     -2.563e+02  2.728e+01  -9.397  < 2e-16 ***
## budynekloft          -9.271e+01  3.551e+01  -2.610 0.009348 **
## budynekwielka_plyta  -3.647e+02  2.917e+01 -12.501  < 2e-16 ***
## wydatki              -1.427e-01  2.291e-01  -0.623 0.533628
## wydatki_zywnosc      -8.779e-02  2.852e-01  -0.308 0.758394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 156.4 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9226, Adjusted R-squared:  0.9208
## F-statistic: 486.9 on 11 and 449 DF,  p-value: < 2.2e-16
```

```
model_bez_plci <- lm(oszczednosci ~.-plec, data=df)
summary(model_bez_plci)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - plec, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -308.31  -60.39   -2.14   57.40  465.38
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -859.1414    61.5945 -13.948  < 2e-16 ***
## wiek                   63.9397     0.5691 112.343  < 2e-16 ***
## waga                    3.9790     0.5698   6.983 1.05e-11 ***
## wzrost                 -2.4073     0.3533  -6.813 3.08e-11 ***
## stan_cywilny1          -4.7973    12.8251  -0.374   0.7085
## liczba_dzieci         151.7746     6.1287  24.764  < 2e-16 ***
## budynekjednorodzinny -181.7290    16.4507 -11.047  < 2e-16 ***
## budynekkamienica     -305.3819    17.9038 -17.057  < 2e-16 ***
## budynekloft          -337.3746    25.1581 -13.410  < 2e-16 ***
## budynekwielka_plyta  -563.6224    20.6441 -27.302  < 2e-16 ***
## wydatki                -0.2977     0.1496  -1.990   0.0472 *
## wydatki_zywnosc        -0.1230     0.1861  -0.661   0.5089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9663
## F-statistic:  1199 on 11 and 449 DF,  p-value: < 2.2e-16
```

```
model_bez_stanu_cyw<- lm(oszczednosci ~.-stan_cywilny, data=df)
summary(model_bez_stanu_cyw)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - stan_cywilny, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -309.12  -59.47   -2.57   58.83  464.70
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -859.6811    61.8138 -13.908  < 2e-16 ***
## wiek                   63.9369     0.5692 112.331  < 2e-16 ***
## waga                    3.9680     0.5706   6.954 1.26e-11 ***
## wzrost                 -2.4060     0.3534  -6.808 3.19e-11 ***
## plecM                   1.5815     9.5759   0.165   0.8689
## liczba_dzieci         150.7451     5.5530  27.146  < 2e-16 ***
## budynekjednorodzinny -181.2668    16.4220 -11.038  < 2e-16 ***
## budynekkamienica     -305.3205    17.9063 -17.051  < 2e-16 ***
## budynekloft          -336.9569    25.1994 -13.372  < 2e-16 ***
## budynekwielka_plyta  -562.9595    20.6186 -27.304  < 2e-16 ***
## wydatki                -0.2974     0.1496  -1.988   0.0474 *
## wydatki_zywnosc        -0.1233     0.1861  -0.663   0.5080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102 on 449 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9663
## F-statistic:  1199 on 11 and 449 DF,  p-value: < 2.2e-16
```

```
model_bez_budynku <- lm(oszczednosci ~.-budynek, data=df)
summary(model_bez_budynku)
```

```
##
## Call:
## lm(formula = oszczednosci ~ . - budynek, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -465.53 -110.35   10.13  106.66  483.51
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1428.9892    93.4459 -15.292  < 2e-16 ***
## wiek               60.9481     0.9379  64.987  < 2e-16 ***
## waga                4.2395     0.9685   4.377 1.49e-05 ***
## wzrost             -1.4476     0.5867  -2.467    0.014 *
## plecM              10.4920    16.3297   0.643    0.521
## stan_cywilny1      16.1381    21.7917   0.741    0.459
```

```
## liczba_dzieci      83.1726     8.8362   9.413  < 2e-16 ***
## wydatki             0.1607     0.2521   0.638    0.524
## wydatki_zywnosc    -0.5169     0.3149  -1.642    0.101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173.3 on 452 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.9045, Adjusted R-squared:  0.9028
## F-statistic:   535 on 8 and 452 DF,  p-value: < 2.2e-16
```

Zatem porównując wyniki z pominięciem którejś ze zmiennych zauważyć można, że odrzucenie zmiennej wiek zdecydowanie obniża jakość modelu (gwałtowanie zmniejsza się wartość R2, a RSS rośnie). W przypadku zmiennych: wydatki na żywność, płeć i stan cywilny wartości R2 i RSS nie ulegają żadnej zmianie w stosunku do pełnego modelu - stąd można wnioskować, że te zmienne nie mają dużego znaczenia w modelu i są kandydatami do odrzucenia. Pozbawienie pełnego modelu jednej z pozostałych zmiennych powoduje jego nieznaczne pogorszenie (nieznaczny spadek R2 i wzrost RSS).

Analizując p-wartości w pełnym modelu, można zauważyć, że ich wartości potwierdzają powyższą konkluzję: wydatki na żywność, płeć i stan cywilny nie wpływają istotnie statystycznie na zmienną objaśnianą. Największa p-wartość otrzymana została dla zmiennej płeć, zatem ją w pierwszej kolejności należałoby usunąć.

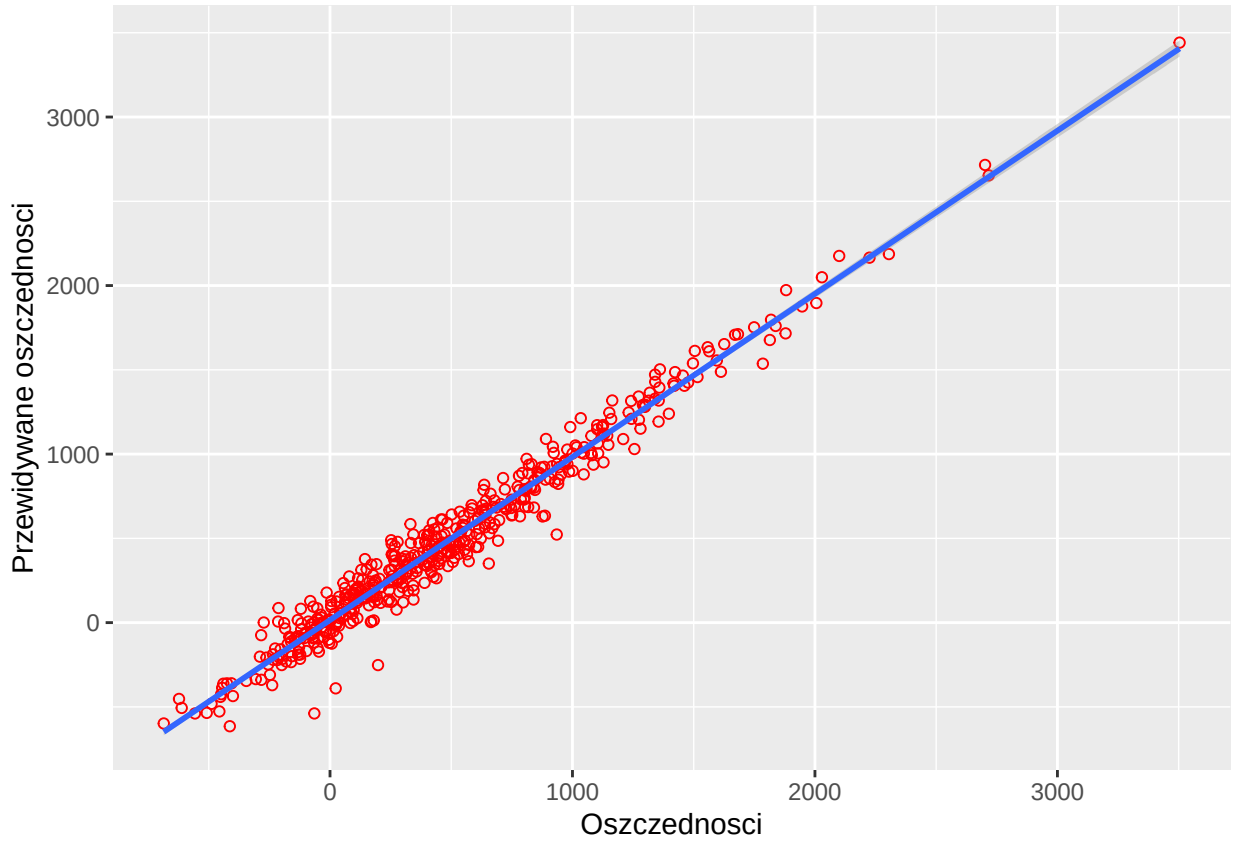
Nowy model bez zmiennej płeć:

```r
nowy_df <- df[, -which(names(df) == "plec")]
nowy_model <- lm(oszczednosci~., data=nowy_df)
summary(nowy_model)
```

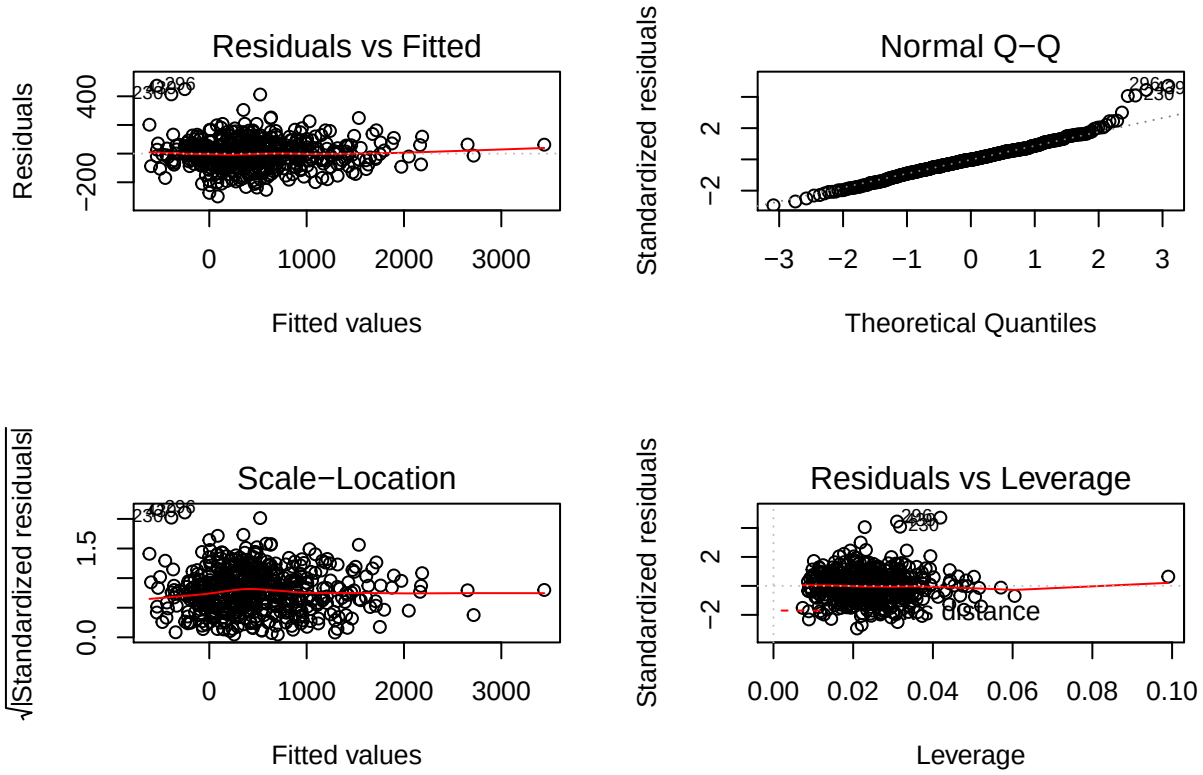
```
##
## Call:
## lm(formula = oszczednosci ~ ., data = nowy_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -298.95  -63.13   -1.98   58.47  474.22
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -839.15973   59.62891 -14.073  < 2e-16 ***
## wiek                   64.14059    0.54204 118.333  < 2e-16 ***
## waga                    3.65366    0.54849   6.661 7.35e-11 ***
## wzrost                 -2.37164    0.34376  -6.899 1.63e-11 ***
## stan_cywilny1          -4.88111   12.38053  -0.394   0.6936
## liczba_dzieci         154.93674    5.90900  26.220  < 2e-16 ***
## budynekjednorodzinny -185.22250   15.96468 -11.602  < 2e-16 ***
## budynekkamienica     -308.82051   17.35518 -17.794  < 2e-16 ***
## budynekloft          -348.49415   23.68482 -14.714  < 2e-16 ***
## budynekwielka_plyta  -572.08815   19.69696 -29.044  < 2e-16 ***
## wydatki                -0.33975    0.14582  -2.330   0.0202 *
## wydatki_zywnosc        -0.07799    0.18139  -0.430   0.6674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.8 on 487 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9671
## F-statistic:  1330 on 11 and 487 DF,  p-value: < 2.2e-16
```

```
# wykres regresji
ggplot(nowy_df, aes(x=oszczednosci, y=predict(nowy_model))) +
  geom_point(col='red', shape=1) +
  geom_smooth(method="lm") +
  xlab("Oszczednosci") +
  ylab("Przewidywane oszczednosci")
```

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




```
# wykresy diagnostyczne
par(mfrow=c(2,2))
plot(nowy_model)
```



Analiza szczegółowa wykresów diagnostycznych:

1. Wykres 1 (residuals vs fitted): pokazuje zależność między wartościami reszt a wartościami przewidywanymi przez model. Każdy punkt na wykresie odpowiada jednemu obserwowanemu punktowi danych, a jego położenie pokazuje, jak bardzo odpowiadająca mu wartość resztowa różni się od wartości przewidywanej przez model. W naszym przypadku mamy mniej więcej równomierne rozłożenie punktów wokół linii –> możemy przyjąć, że warunek liniowości jest spełniony.
2. Wykres 2 (normal Q-Q): służy do diagnozowania założenia normalności rozkładu reszt w modelu regresji liniowej. W naszym przypadku punkty w większości są dobrze dopasowane –> pojedyncze punkty leżą dalej od linii rozkładu normalnego.
3. Wykres 3 (scale-location): używany jest do sprawdzenia założenia homoskedastyczności (jednorodności wariancji) reszt. W naszym przypadku możemy uznać, że wariancja reszt jest stała (punkty rozkładają się mniej więcej równomiernie).
4. Wykres 4 (residuals vs leverage): służy do identyfikacji obserwacji, które mają duży wpływ na dopasowanie modelu (tzw. obserwacje odstające lub obserwacje z dużą dźwignią). Im bardziej odległa od środka jest wartość na wykresie dla danej obserwacji, tym większy wpływ ma ta obserwacja na dopasowanie modelu. Wartości zbyt odległe od środka (poza czerwoną linią) sugerują, że obserwacja ta może wpłynąć na wyniki modelu i warto ją zbadać dokładniej. W naszym przypadku wszystkie punkty znajdują się wewnątrz obszaru krytycznego, zatem możemy uznać, że nie ma wartości odstających, które w znaczący sposób wpływałyby na nasz model.

Na podstawie powyższej analizy możemy zatem uznać, że dla nowego modelu (bez zmiennej płeć) założenia modelu liniowego są spełnione.