Pierwszy projekt zaliczeniowy

Statystyczna analiza danych 2020/2021

Joanna Kęczkowska

26.04.2021

Celem zadania jest statystyczna analiza danych znajdujących się w pliku people.tab. Dane: Są to dane symulowane; opisują wiek (zmienna age), wagę (weight), wzrost (height), płeć (gender), stan cywilny (married), liczbę dzieci (number_of_kids), posiadane zwierzę domowe (pet) oraz miesięczne wydatki (expenses) pewnych osób. We wszystkich zadaniach poniżej zmienna expenses jest zmienną objaśnianą (zależną), a pozostałe zmienne są zmiennymi objaśniającymi (niezależnymi).

1. Wczytaj dane, obejrzyj je i podsumuj w dwóch-trzech zdaniach. Pytania pomocnicze: ile jest obserwacji, ile zmiennych ilościowych, a ile jakościowych? Czy są zależności w zmiennych objaśniających (policz i zaprezentuj na wykresach korelacje pomiędzy zmiennymi ilościowymi, a także zbadaj zależność zmiennych jakościowych). Skomentuj wyniki. Czy występują jakieś braki danych?

```
df <- read.delim("peopletab.txt", header = TRUE, sep='\t')
sprintf("Dane zawierają %d obserwacji i %d cech", dim(df)[1], dim(df)[2])
```

```
## [1] "Dane zawierają 500 obserwacji i 8 cech"
```

```
summary(df)
```

```
##       age            weight           height         gender
##  Min.   :17.00   Min.   : 19.40   Min.   :113.6   Length:500
##  1st Qu.:33.00   1st Qu.: 57.60   1st Qu.:155.6   Class :character
##  Median :39.00   Median : 66.60   Median :169.0   Mode  :character
##  Mean   :39.48   Mean   : 66.39   Mean   :168.2
##  3rd Qu.:45.00   3rd Qu.: 75.30   3rd Qu.:180.1
##  Max.   :72.00   Max.   :107.20   Max.   :235.2
##   married        number_of_kids      pet              expenses
##  Mode :logical   Min.   :0.000   Length:500         Min.   :-685.68
##  FALSE:327       1st Qu.:0.750   Class :character   1st Qu.:  74.51
##  TRUE :173       Median :1.000   Mode  :character   Median : 402.22
##                  Mean   :1.558                      Mean   : 478.60
##                  3rd Qu.:2.000                      3rd Qu.: 802.72
##                  Max.   :6.000                      Max.   :3503.90
```

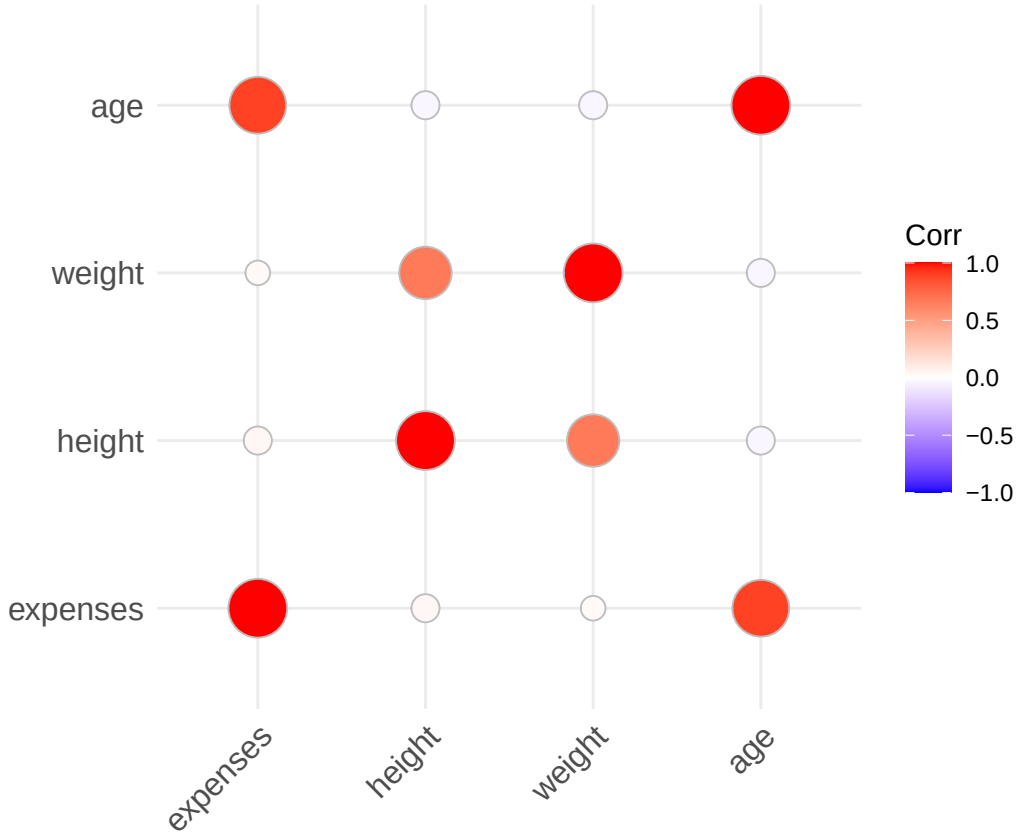
Dane zawierają **500 obserwacji**.
Zmienne ilościowe: 'age', 'weight', 'height', 'expenses'.
Zmienne jakościowe: 'gender', 'married', 'pet', 'number_of_kids'. Niepokojące są ujemne wartości w cesze 'expanses', jak również factor 'other' w cesze 'gender'. Wartość 'none' w cesze 'pet' interpretuję jako nieposiadanie zwierzęcia. W zmiennej 'gender' potraktuję factor 'other' jako brak informacji.

```
df$gender <- factor(df$gender)
df$pet <- factor(df$pet)
df$married <- factor(df$married)
df$number_of_kids <- factor(df$number_of_kids)
summary(df)
```

```
##       age            weight           height        gender     married
##  Min.   :17.00   Min.   : 19.40   Min.   :113.6   man  :223   FALSE:327
##  1st Qu.:33.00   1st Qu.: 57.60   1st Qu.:155.6   other: 38   TRUE :173
##  Median :39.00   Median : 66.60   Median :169.0   woman:239
##  Mean   :39.48   Mean   : 66.39   Mean   :168.2
##  3rd Qu.:45.00   3rd Qu.: 75.30   3rd Qu.:180.1
##  Max.   :72.00   Max.   :107.20   Max.   :235.2
##
##  number_of_kids       pet         expenses
##  0:125          cat     :105   Min.   :-685.68
##  1:161          dog     :100   1st Qu.:  74.51
##  2: 99          ferret  : 54   Median : 402.22
##  3: 63          hedgehog: 54   Mean   : 478.60
##  4: 35          none    :187   3rd Qu.: 802.72
##  5: 11                         Max.   :3503.90
##  6:  6
```

Współczynnik korelacji r jest liczbą pomiędzy –1 i 1, która określa, w jakim stopniu dwie zmienne są współzależne. Wartość r = 0 oznacza, że nie ma żadnego powiązania, a wartość 1 lub –1 oznacza idealne powiązanie. Znak współczynnika korelacji wskazuje, czy zmienne są skorelowane dodatnio (większe wartości w jednej zmiennej pokrywają się z większymi wartościami w drugiej), czy też ujemnie (większe wartości w jednej zmiennej pokrywają się z mniejszymi wartościami w drugiej).

```
library(ggcorrplot)
numerical <- df[c("expenses", "height", "weight", "age")]
categorial <- df[c("married", "gender", "pet", "number_of_kids")]

corr <- round(cor(numerical), 2)

ggcorrplot(corr, method = "circle")
```



Zgodnie z intuicją **wiek jest dodatnio skorelowany z zarobkami** i **wzrost jest dodatnio skorelowany z wagą**.

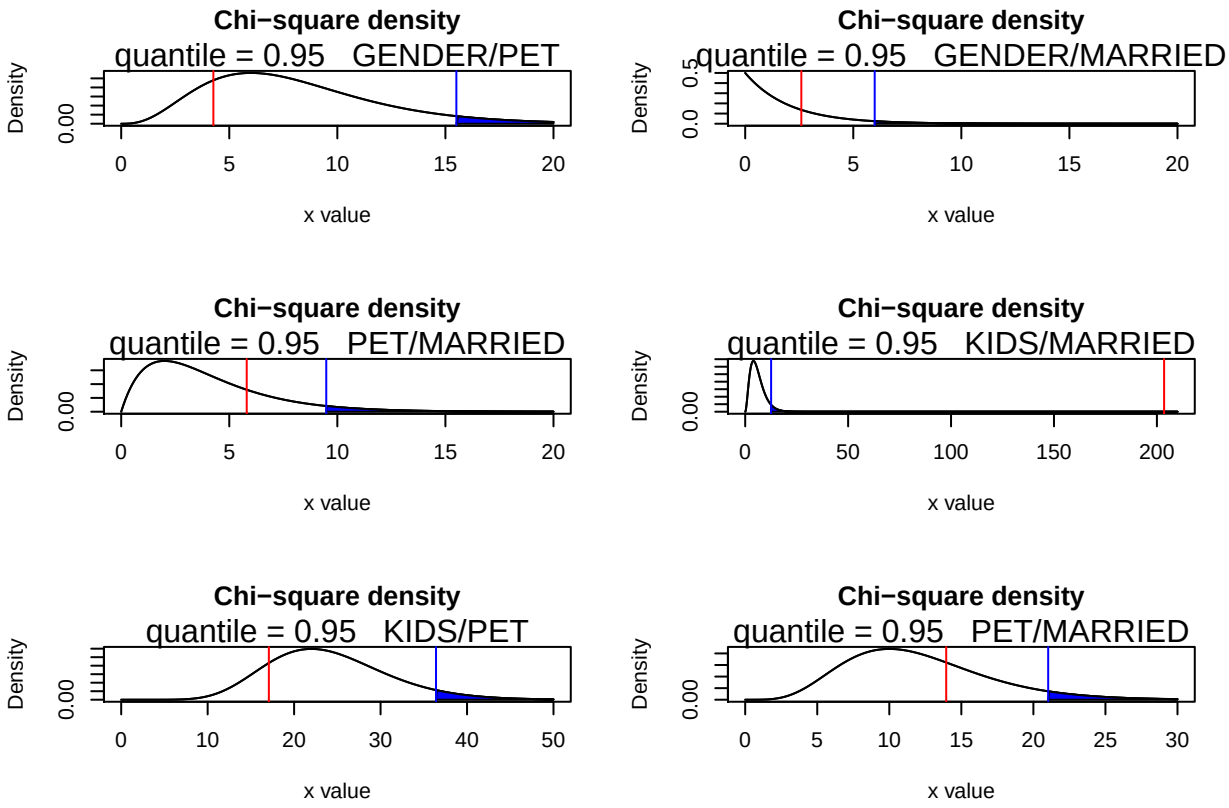
W przypadku zmiennych jakościowych nie możemy zbadać korelacji tak jak dla zmiennych ilościowych - przypisane do nich wartości liczbowe są jedynie symboliczne.
Dla tego typu zmiennych posłużymy się testem zgodności χ^2

dla danej komórki wartość oczekiwana: $e = \frac{row.sum * col.sum}{grand.total}$
Chi-square statistic: $\chi^2 = \sum \frac{(o-e)^2}{e}$, gdzie o - obserwacja, e - wartość oczekiwana

Hipoteza zerowa H_0: Zmienne są **niezależne**.
Hipoteza alternatywna H_1: Zmienne są **zależne**.

```
#funkcja do testowania korelacji zmiennych jakościowych
#przyjmuje dwie kolumny zmiennych kategorycznych, które zamienia na tablicę wielodzielczą

testchi <- function(feature1, feature2, sq = 20, t) {
  alpha <- 0.05 #5% level of significance
  TAB <- table(feature1, feature2)
  total <- sum(TAB)

  n <- nlevels(feature1)
  m <- nlevels(feature2)

  sumRows <- margin.table(TAB, 1) #rows
  sumCols <- margin.table(TAB,2) #columns
```

```
  sumRows <- as.vector(sumRows)
  sumCols <- as.vector(sumCols)

  exp <- matrix(rep(0, n*m), nrow=n, ncol=m)
  exp[] <- 0L
  for(i in 1:n) {
    exp[i, ] <- sumRows[i]*sumCols/total
  }

  Tab <- data.frame(TAB)
  obs <- matrix(Tab[["Freq"]], nrow = n, ncol = m)

  chi_sq <- sum((obs-exp)^2/exp) #test statistic
  df <- (nrow(obs)-1)*(ncol(obs)-1) #deg of freedom
  pval <- pchisq(chi_sq, df, lower.tail=FALSE)  #right-tailed

  quantile <- qchisq(alpha, df, lower.tail = FALSE) #quantile of chi-square distribution

  x <- seq(0, sq, by = 0.1)
  chi_dense <- dchisq(x, df)

  plot(x, chi_dense,type='l', xlab="x value",
  ylab="Density", main="Chi-square density")

  i <- x >= quantile
  lines(x, chi_dense)
  polygon(c(quantile,x[i],sq), c(0,chi_dense[i],0), col="blue")

  area <- pchisq(quantile, df, lower.tail = TRUE)
  result <- paste("quantile =", signif(area, digits=3), " ", t)
  mtext(result,3)
  abline(v=chi_sq, col="red")
  abline(v=quantile, col="blue")

  c <- list(chi_sq, pval, quantile)
  return (c)

}
par(mfrow = c(3, 2))
gp <- testchi(df$gender, df$pet, t="GENDER/PET")
gm <- testchi(df$gender, df$married, t="GENDER/MARRIED")
pm <- testchi(df$pet, df$married, t="PET/MARRIED")
nm <- testchi(df$number_of_kids, df$married, sq=210, t="KIDS/MARRIED")
np <- testchi(df$number_of_kids, df$pet, t="KIDS/PET", sq=50)
gn <- testchi(df$gender, df$number_of_kids, t="PET/MARRIED", sq=30)
```




```
sprintf("GENDER/PET, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gp[[1]]
```

```
## [1] "GENDER/PET, test statistic = 4.264540 , p-value = 0.832502, confidece interval = [-infinity, 15
```

```
sprintf("GENDER/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gm[
```

```
## [1] "GENDER/MARRIED, test statistic = 2.597089 , p-value = 0.272929, confidece interval = [-infinity
```

```
sprintf("PET/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", pm[[1]]
```

```
## [1] "PET/MARRIED, test statistic = 5.806971 , p-value = 0.214035, confidece interval = [-infinity, 9
```

```
sprintf("KIDS/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", nm[[1]
```

```
## [1] "KIDS/MARRIED, test statistic = 203.501380 , p-value = 0.000000, confidece interval = [-infinity
```

```
sprintf("KIDS/PET, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", np[[1]], n
```

```
## [1] "KIDS/PET, test statistic = 17.076893 , p-value = 0.845364, confidece interval = [-infinity, 36.4
```

```
sprintf("PET/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gn[[1]]
```

```
## [1] "PET/MARRIED, test statistic = 13.950817 , p-value = 0.303860, confidece interval = [-infinity, 1
```

Jedyne dwie skorelowane zmienne jakościowe to 'number_of_kids' i 'married' - statystyka testowa wpada do obszaru krytycznego. W przypadku pozostałych par zmiennych nie mamy podstawy do odrzucenia hipotezy zerowej. Żadne dwie inne zmienne nie wydają się być skorelowane.

Jeszcze tylko szybkie sprawdzenie:

```
#sprawdźmy
chisq.test(table(df$gender, df$pet))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$gender, df$pet)
## X-squared = 4.2645, df = 8, p-value = 0.8325
```

```
chisq.test(table(df$gender, df$married))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$gender, df$married)
## X-squared = 2.5971, df = 2, p-value = 0.2729
```

```
chisq.test(table(df$pet, df$married))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$pet, df$married)
## X-squared = 5.807, df = 4, p-value = 0.214
```

```
chisq.test(table(df$number_of_kids, df$married))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$number_of_kids, df$married)
## X-squared = 203.5, df = 6, p-value < 2.2e-16
```

```
chisq.test(table(df$number_of_kids, df$pet))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$number_of_kids, df$pet)
## X-squared = 17.077, df = 24, p-value = 0.8454
```

```
chisq.test(table(df$gender, df$number_of_kids))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$gender, df$number_of_kids)
## X-squared = 13.951, df = 12, p-value = 0.3039
```

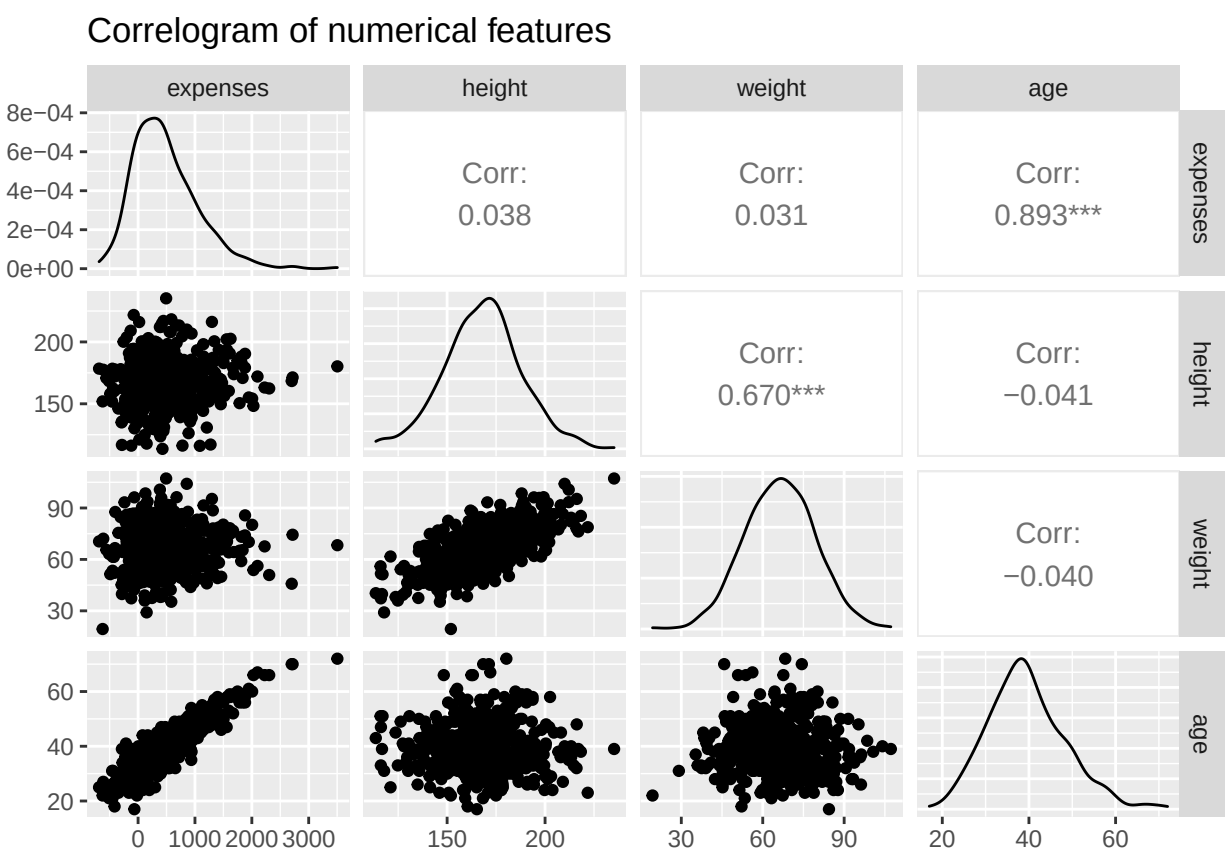
2. Podsumuj dane przynajmniej trzema różnymi wykresami. Należy przygotować: **a)** wykres typu scatter-plot (taki jak na wykładzie 6, slajd 3) dla wszystkich zmiennych objaśniających ilościowych i zmiennej objaśnianej. **b)** Wykresy typu pudełkowy (boxplot) dla jednej wybranej zmiennej ilościowej. **c)** Wykres typu słupkowy (barplot) dla jednej wybranej zmiennej jakościowej. Dodatkowe wykresy wg własnej inwencji (np. histogram, punktowy, liniowy, mapa ciepła. . .).

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

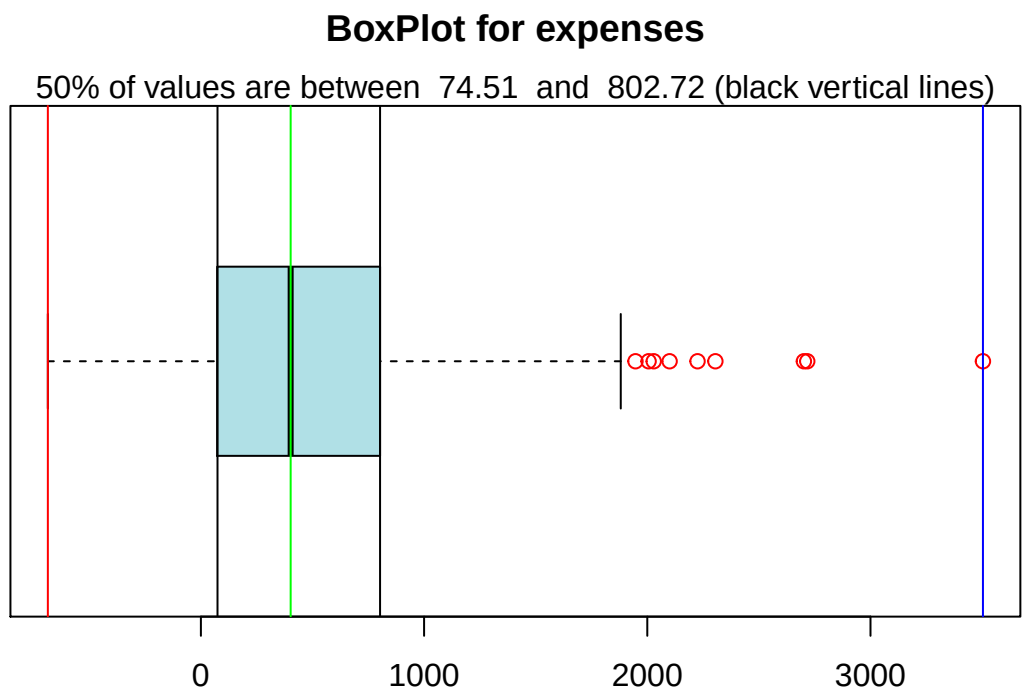
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(numerical, title="Correlogram of numerical features")
```

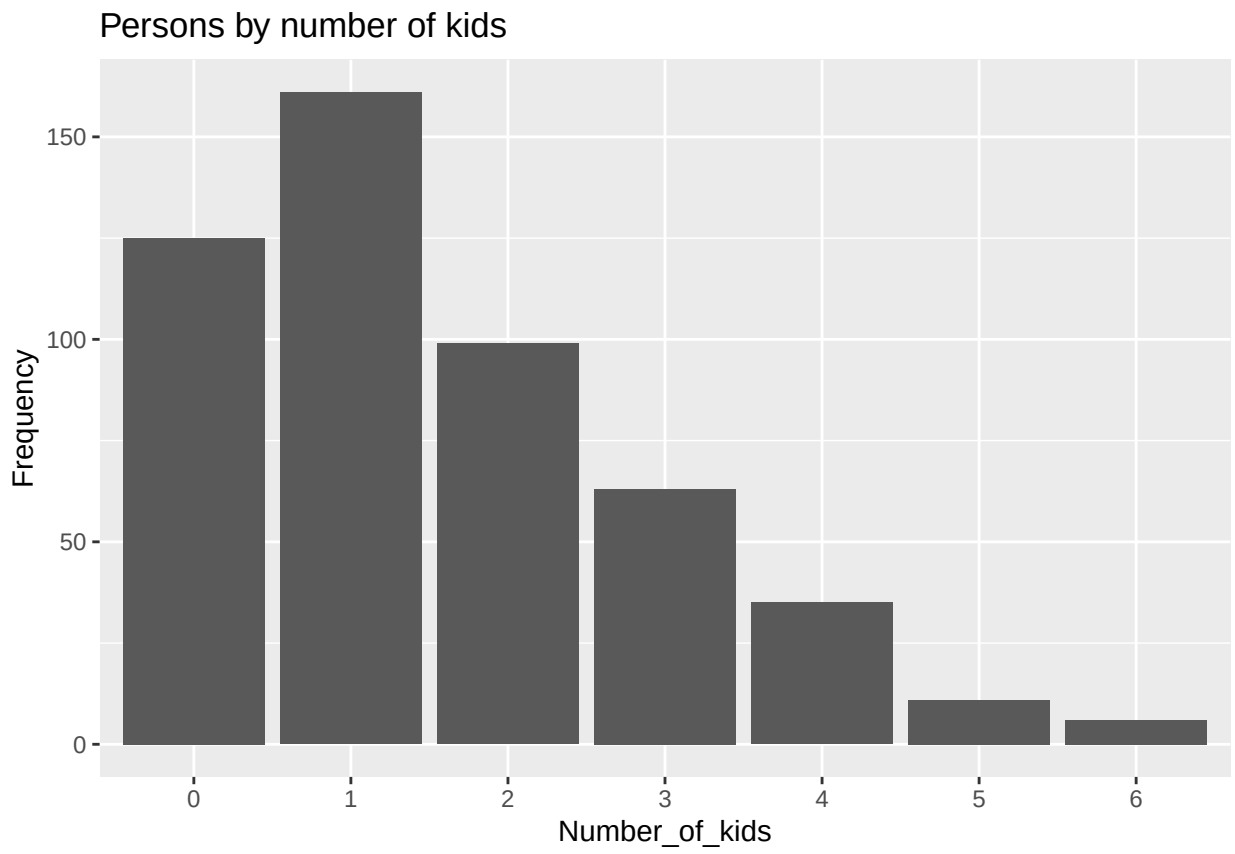


Trochę inny wykres od 'ggcorrplot(corr, method = "circle")' ale prowadzący do tych samych wniosków: dodatnia korelacja wzrostu z wagą i dodatnia korelacja wieku z zarobkami.

```
expenses <- df$expenses
quantiles <- unname(quantile(expenses))
boxplot(expenses, horizontal = TRUE, col="powderblue", outcol="red", main="BoxPlot for expenses")

#linie pomocnicze
abline(v = max(expenses), col="blue")
abline(v = min(expenses), col="red")
abline(v = median(expenses), col="green")
result <- paste("50% of values are between ", round(quantiles[2], 2), " and ",round(quantiles[4], 2), "
mtext(result,3)
abline(v=quantiles[4], col="black") #quantile 75%
abline(v=quantiles[2], col="black") #quantile 25%
```

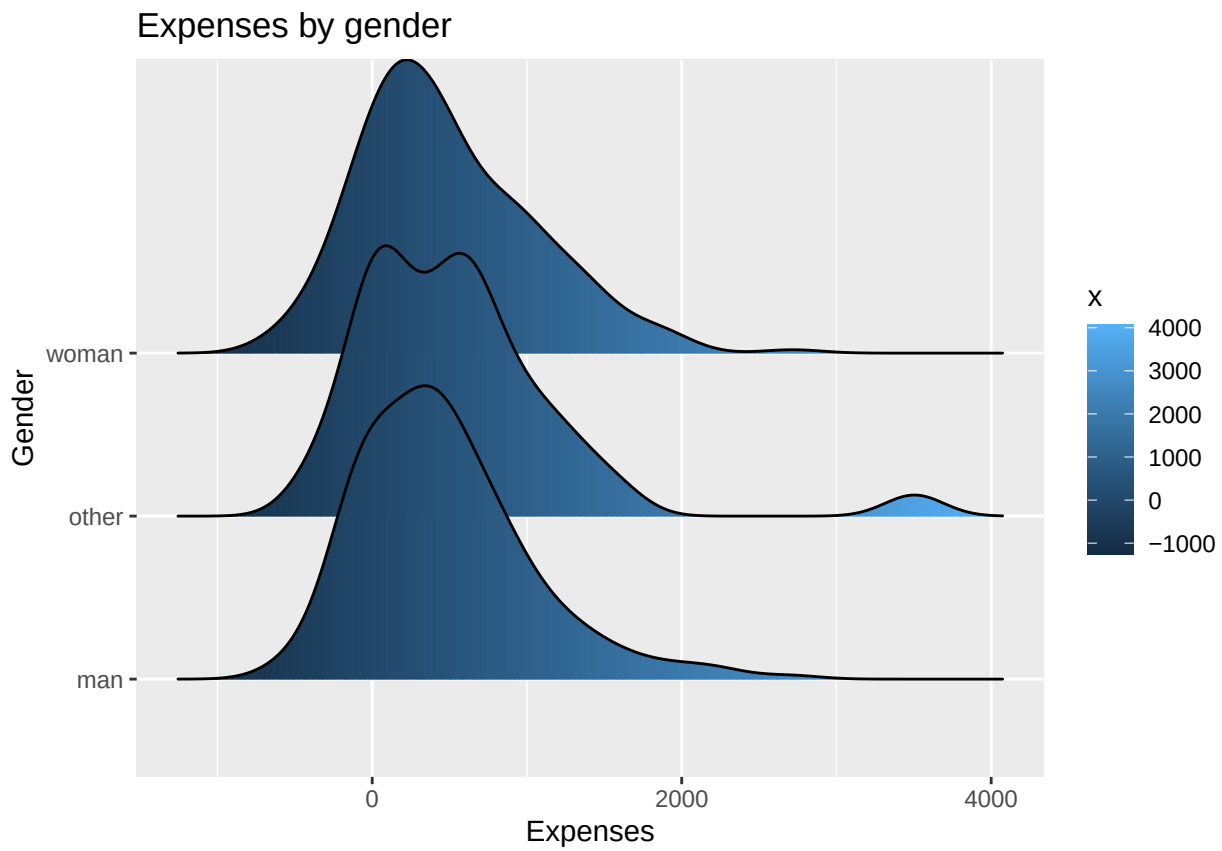



```
library(ggplot2)

ggplot(df, aes(x=number_of_kids))+ geom_bar() + labs(x = "Number_of_kids",
       y = "Frequency",
       title  = "Persons by number of kids")
```




```
ggplot(df, aes(x = expenses, y = gender, fill=stat(x))) +  geom_density_ridges_gradient()+
labs(x = "Expenses",
       y = "Gender",
       title  = "Expenses by gender")
```

```
## Picking joint bandwidth of 190
```



3. Policz p-wartości dla hipotez o wartości średniej m = 170 i medianie me = 165 (cm) dla zmiennej wzrost. Wybierz statystykę testową dla alternatywy lewostronnej, podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione.

Test dla wariancji: $t = \frac{\overline{X}-\mu}{\sigma}:\sqrt{n}$, gdzie $\overline{X}$ - średnia próby, n - liczba obserwacji, σ - odchylenie standardowe

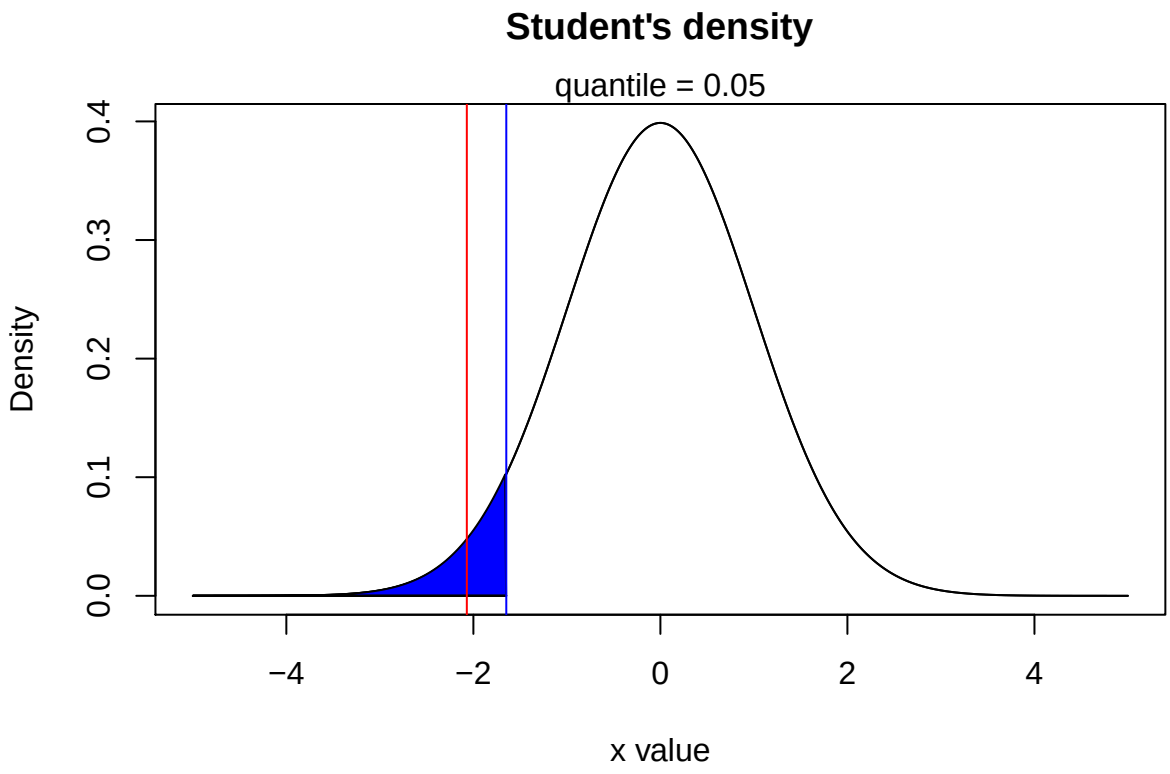
Hipoteza zerowa H0: $\mu = 170$ Hipoteza alternatywna H1: $\mu < 170$

```
alpha <- 0.05
m <- 170
me <- 165
height <- df$height
n <- length(height)

test <- (mean(height)-m)/sd(height)*sqrt(n)
def <- n-1

quantile <- qt(alpha, def) #left-tailed
pval <- pt(test, def)

x<- seq(-5, 5, by=0.01)
t_dense <- dt(x, n-1)

plot(x, t_dense,type='l', xlab="x value",
  ylab="Density", main="Student's density")

i<- x<=quantile
lines(x, t_dense)
```

```
polygon(c(-5,quantile,x[i]), c(0,t_dense[i],0),col="blue")
sprintf("test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", test, pval, quantile
```

```
## [1] "test statistic = -2.069917 , p-value = 0.019487, confidece interval = [-infinity, -1.647913]"
```

```
area <- pt(quantile, def)
result <- paste("quantile =", signif(area, digits=3))
mtext(result,3)
abline(v=test, col="red")
abline(v=quantile, col="blue")
```




```
if(alpha > pval) {
  print("H0 rejected.")
}else {
  print("There is not enough evidence to reject H_0")
}
```

```
## [1] "H0 rejected."
```

```
#sprawdźmy
t.test(height, mu=m)
```

```
##
```

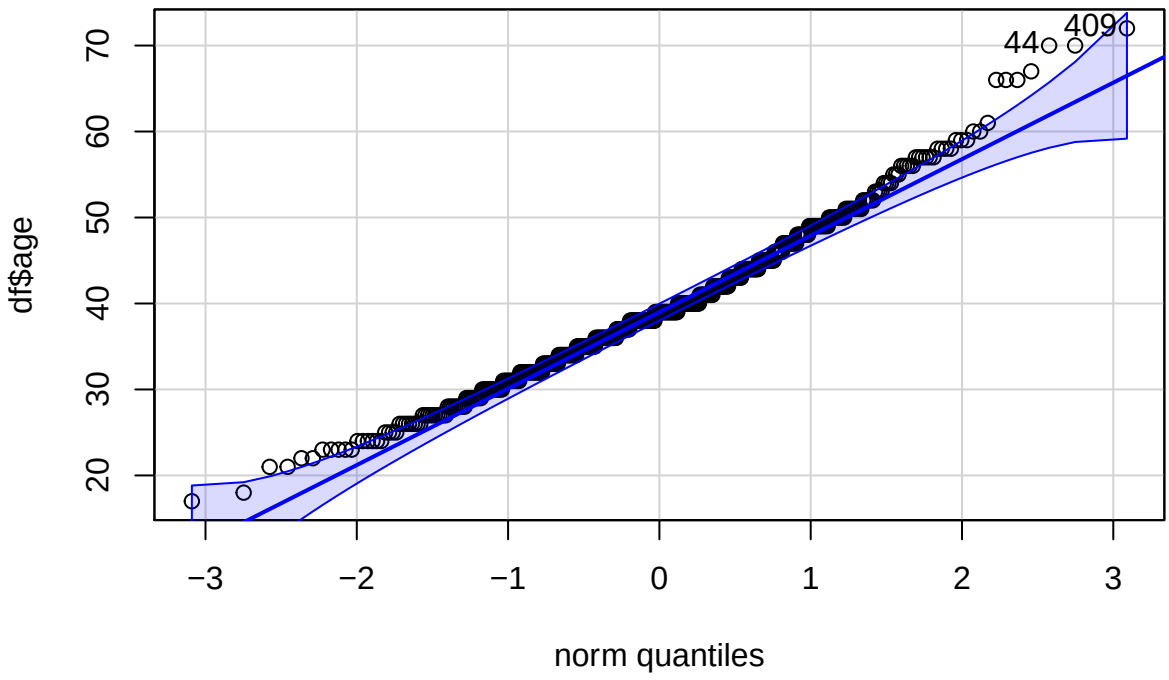
```
##  One Sample t-test
##
## data:  height
## t = -2.0699, df = 499, p-value = 0.03897
## alternative hypothesis: true mean is not equal to 170
## 95 percent confidence interval:
##  166.4532 169.9075
## sample estimates:
## mean of x
##  168.1804
```

```
#tu wpisać test dla mediany
```

4. Policz dwustronne przedziały ufności na poziomie 0.99 dla zmiennej wiek dla następujących parametrów rozkładu: 1. średnia i odchylenie standardowe; 2. kwantyle 1/4, 2/4 i 3/4. Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione.

Studentyzowany przedział ufności:
$\left(\bar{X} - \frac{t(1-\alpha/2,n-1)}{\sqrt{n}}\hat{S}, \bar{X} + \frac{t(1-\alpha/2,n-1)}{\sqrt{n}}\hat{S}\right)$ gdzie $\bar{X}$ to średnia, $\hat{S}$ to pierwiastek z *nieobciążonego* estymatora wariancji, a $t(1-\alpha/2, n-1)$ to kwantyl na poziomie $1-\alpha/2$ dla rozkładu t Studenta o $n-1$ stopniach swobody.
Asymptotyczny przedział ufności:
$\left(\bar{X} - \frac{q(1-\alpha/2)}{\sqrt{n}}\hat{S}, \bar{X} + \frac{q(1-\alpha/2)}{\sqrt{n}}\hat{S}\right)$ gdzie $q(1-\alpha/2)$ jest kwantylem na poziomie $1-\alpha/2$ ze standardowego rozkładu normalnego.

```
alpha <- 0.01
#ocena czy zmienna age ma rozkład normalny
age <- df$age
qqPlot(df$age)
```




```
## [1] 409  44
```

```
n<- length(age)
```

```
rightstud <-  mean(age) + 1/sqrt(n-1)*sd(age)*qt(1-alpha/2, (n-1))
leftstud <-  mean(age) - 1/sqrt(n-1)*sd(age)*qt(1-alpha/2, (n-1))

rightasympt <- mean(age) + (qnorm(1-alpha/2))/sqrt(n-1)*sd(age)
leftasympt <- mean(age) - (qnorm(1-alpha/2))/sqrt(n-1)*sd(age)

sprintf("(%f, %f)", leftstud, rightstud)
```

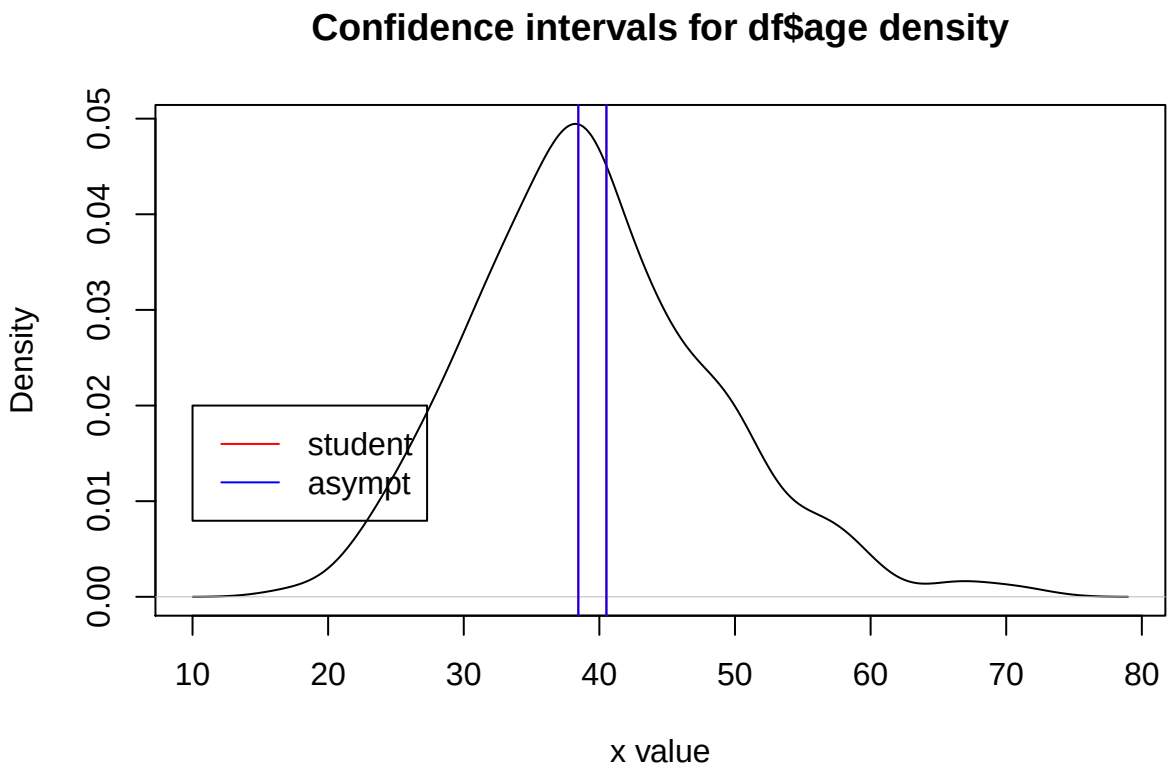
```
## [1] "(38.444964, 40.523036)"
```

```
sprintf("(%f, %f)",leftasympt, rightasympt)
```

```
## [1] "(38.448937, 40.519063)"
```

```
plot(density(age),type='l', xlab="x value", ylab="Density", main="Confidence intervals for df$age densi

abline(v=leftstud, col="red")
abline(v=rightstud, col="red")
abline(v=leftasympt, col="blue")
```

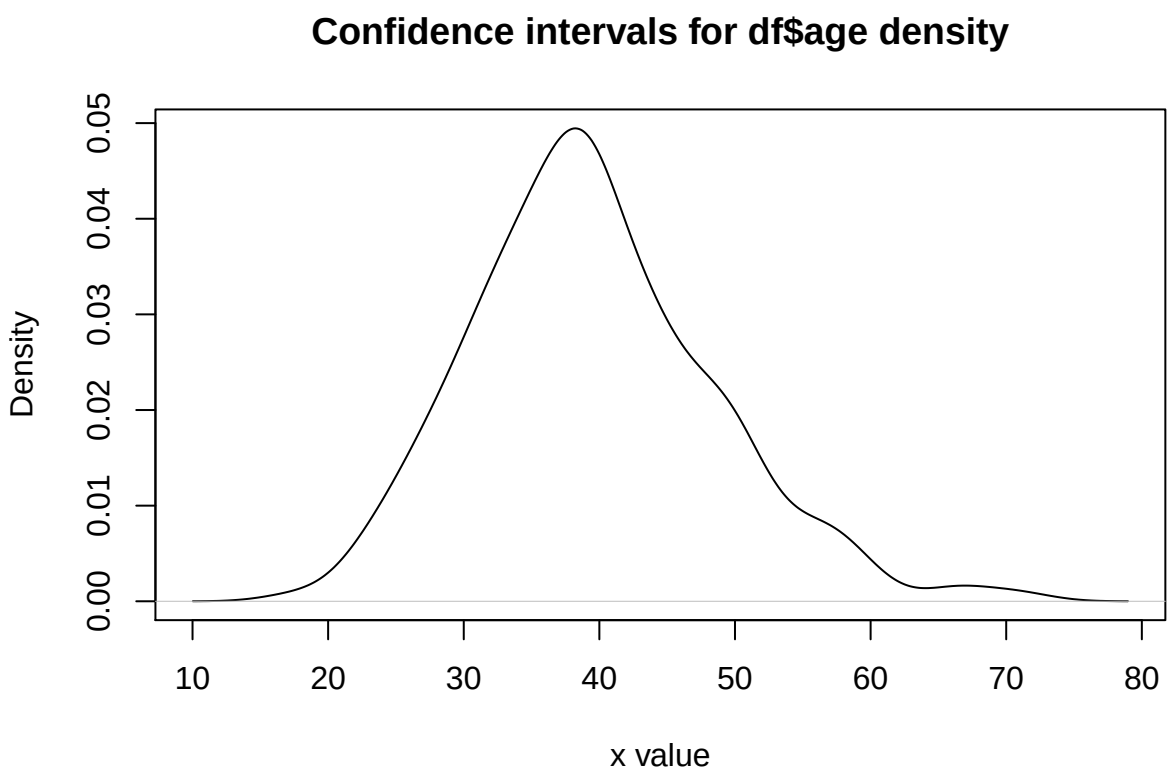
```
abline(v=rightasympt, col="blue")

legend(10,0.02,c("student","asympt"), lty = c(1,1), col=c('red','blue'),ncol=1)
```




```
lv <- levels(quantcut(age, probs=seq(0, 1, 1/4)))
lv
```

```
## [1] "[17,33]" "(33,39]" "(39,45]" "(45,72]"
```

```
plot(density(age),type='l', xlab="x value", ylab="Density", main="Confidence intervals for df$age densi
```




```
#jeszcze cos tu bedzie
```

5. Przetestuj na poziomie istotności 0.01 trzy hipotezy istotności:
1. różnicy między średnią wartością wybranej zmiennej dla kobiet i dla mężczyzn;
2. zależności między dwiema zmiennymi ilościowymi;
3. zależności między dwiema zmiennymi jakościowymi.

Ponadto,
4. przetestuj hipotezę o zgodności z konkretnym rozkładem parametrycznym dla wybranej zmiennej (np. "zmienna A ma rozkład wykładniczy z parametrem 10"). Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione.

3. była robiona wyżej dla wszystkich zmiennych jakościowych.

6. Oszacuj model regresji liniowej, przyjmując za zmienną zależną (y) wydatki domowe (expenses) a jako zmienne niezależne (x) przyjmując pozostałe zmienne.
Rozważ, czy konieczne są transformacje zmiennych lub zmiennej objaśnianej. Podaj RSS, R^2, p-wartości i oszacowania współczynników w pełnym modelu (w modelu zawierającym wszystkie zmienne). Następnie wybierz jedną zmienną objaśniającą, którą można by z pełnego modelu odrzucić (która najgorzej tłumaczy expenses). Aby dokonać wyboru takiej zmiennej, dla każdej ze zmiennych objaśniających sprawdź:
- Jaką ma p-wartość w pełnym modelu?
- O ile zmniejsza się R^2, gdy ją usuniemy z pełnego modelu?
- O ile zwiększa się RSS, gdy ją usuniemy z pełnego modelu?
Opisz wnioski.
Oszacuj model ze zbiorem zmiennych objaśniających pomniejszonym o wybraną zmienną. Sprawdź czy w otrzymanym przez Ciebie modelu spełnione są założenia modelu liniowego i przedstaw na wykresach diag-

nostycznych: wykresie zależności reszt od zmiennej objaśnianej, na wykresie reszt studentyzowanych i na wykresie dźwigni i przedyskutuj, czy są spełnione./

```
linear_regression <- lm(expenses ~ age + weight + height + gender + married + number_of_kids + pet, df)
summary(linear_regression)
```

```
##
## Call:
## lm(formula = expenses ~ age + weight + height + gender + married +
##     number_of_kids + pet, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -758.69 -119.55    3.06  128.17  885.29
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2299.7769   101.8435 -22.581  < 2e-16 ***
## age                57.5620     1.0753  53.529  < 2e-16 ***
## weight              1.3272     0.9963   1.332  0.18345
## height              2.0825     0.6572   3.169  0.00163 **
## genderother        48.1590    37.7395   1.276  0.20254
## genderwoman       -17.6926    20.1808  -0.877  0.38108
## marriedTRUE       -17.7722    26.3260  -0.675  0.49995
## number_of_kids1    14.9060    25.9026   0.575  0.56525
## number_of_kids2   -56.9082    30.1096  -1.890  0.05935 .
## number_of_kids3    25.9653    38.2434   0.679  0.49750
## number_of_kids4   -52.4638    46.6479  -1.125  0.26129
## number_of_kids5   -31.8876    72.4424  -0.440  0.66000
## number_of_kids6  -115.3335    93.2361  -1.237  0.21669
## petdog             34.8071    30.1051   1.156  0.24818
## petferret         413.5364    36.2367  11.412  < 2e-16 ***
## pethedgehog       244.8257    35.8244   6.834  2.5e-11 ***
## petnone            24.2985    26.1476   0.929  0.35321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 212.9 on 483 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8592
## F-statistic: 191.3 on 16 and 483 DF,  p-value: < 2.2e-16
```

~TBA