SAD2021 - projekt 1

Jakub Skrajny

1

Wczytujemy dane

```
data <- read.table("data.csv", sep="\t", header=TRUE)
head(data)
```

```
##   age weight height gender married number_of_kids      pet   expenses
## 1  25   61.7 121.12  other   FALSE              2   ferret   23.44299
## 2  37   63.9 145.00    man    TRUE              6      dog   96.83683
## 3  41   50.2 145.03  woman    TRUE              2 hedgehog  312.67693
## 4  43   72.4 179.90    man   FALSE              1      dog  447.42838
## 5  26   78.4 163.91    man   FALSE              1 hedgehog  -78.22799
## 6  49   59.4 151.86  woman    TRUE              2   ferret 1241.98263
```

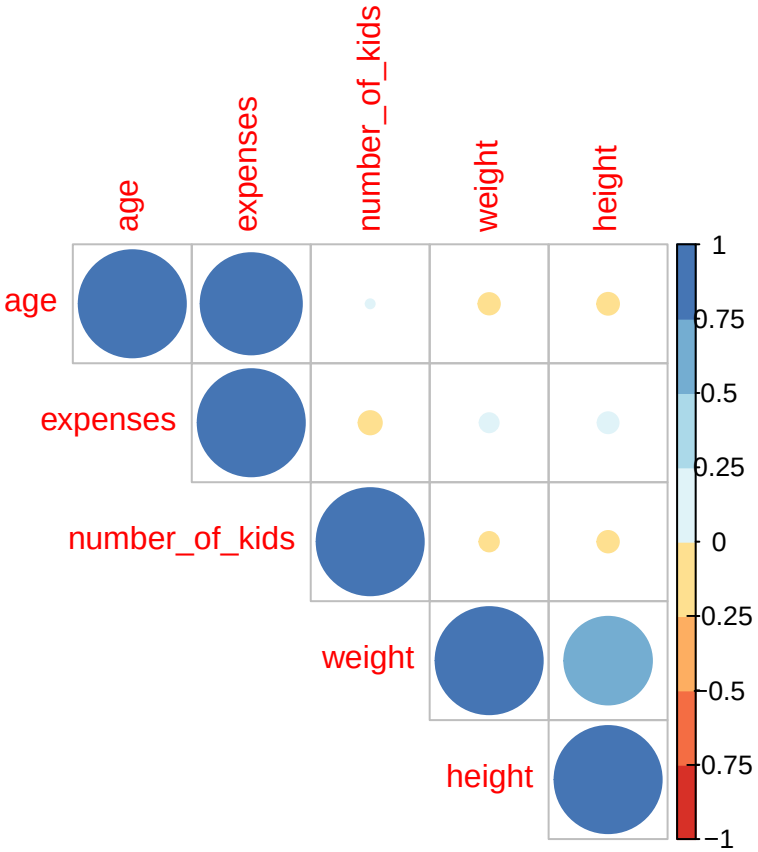
```
which(is.na(data))
```

```
## integer(0)
```

Mamy 500 obserwacji. Pięć z ośmiorga zmiennych to zmienne ilościowe. Dane są kompletne.

Zależności między zmiennymi ilościowymi

```
data_num <- Filter(is.numeric, data)
cor_num <- cor(data_num)
corrplot(cor_num, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```



Na podstawie powyższej grafiki widzimy, że jedyne dobrze skorelowane pary zmiennych, to 'age' z 'expenses' oraz 'height' z 'weight'.

Zależności między zmiennymi jakościowymi

```
chisq.test(data$pet, data$married)
```

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

```
chisq.test(data$married, data$gender)
```

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

```
chisq.test(data$gender, data$pet)
```

```
## Warning in chisq.test(data$gender, data$pet): Chi-squared approximation may be
## incorrect
```

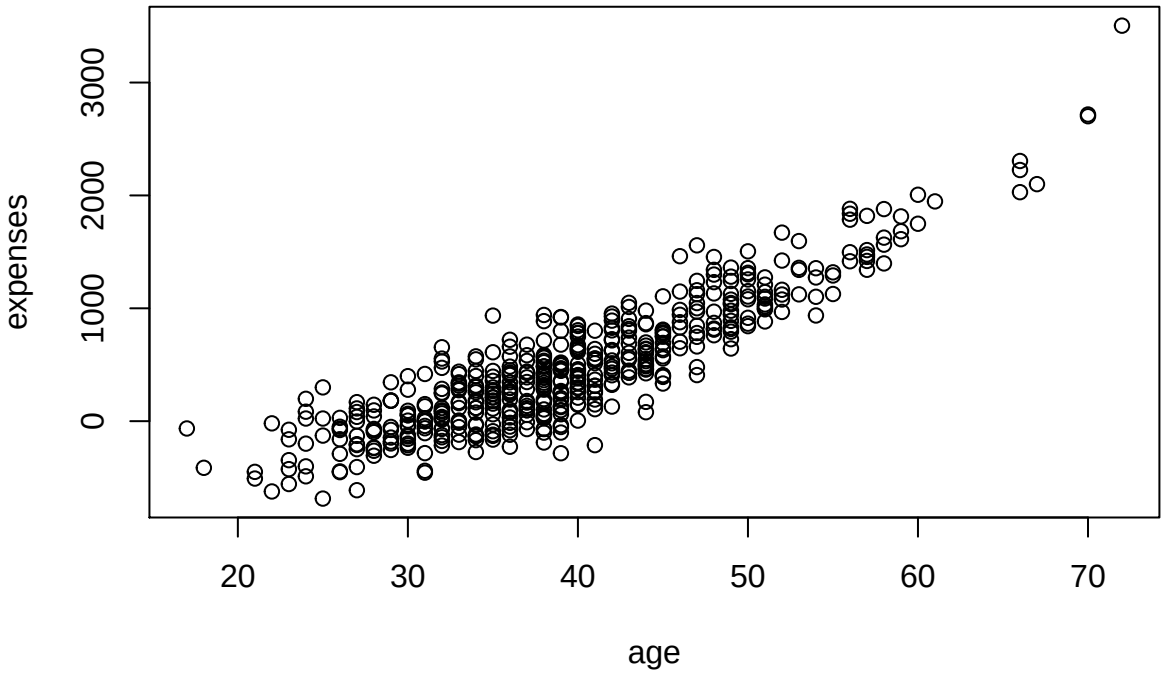
```
##
##  Pearson's Chi-squared test
##
## data:  data$gender and data$pet
## X-squared = 4.2645, df = 8, p-value = 0.8325
```

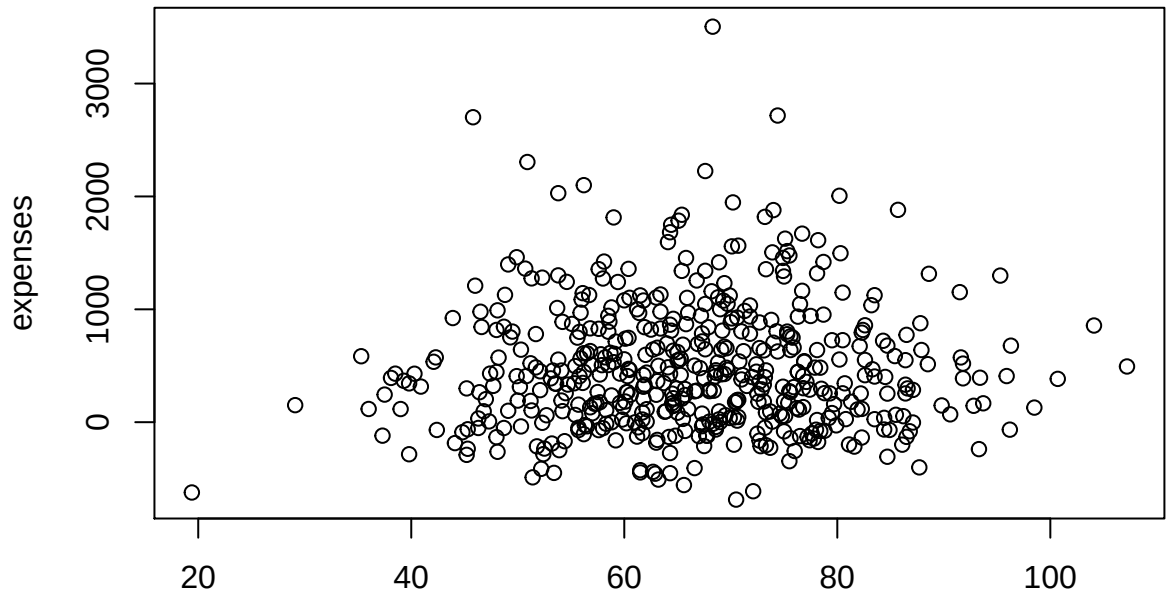
Nie ma podstaw aby odrzucić hipotezę o niezależności zmiennych jakościowych.

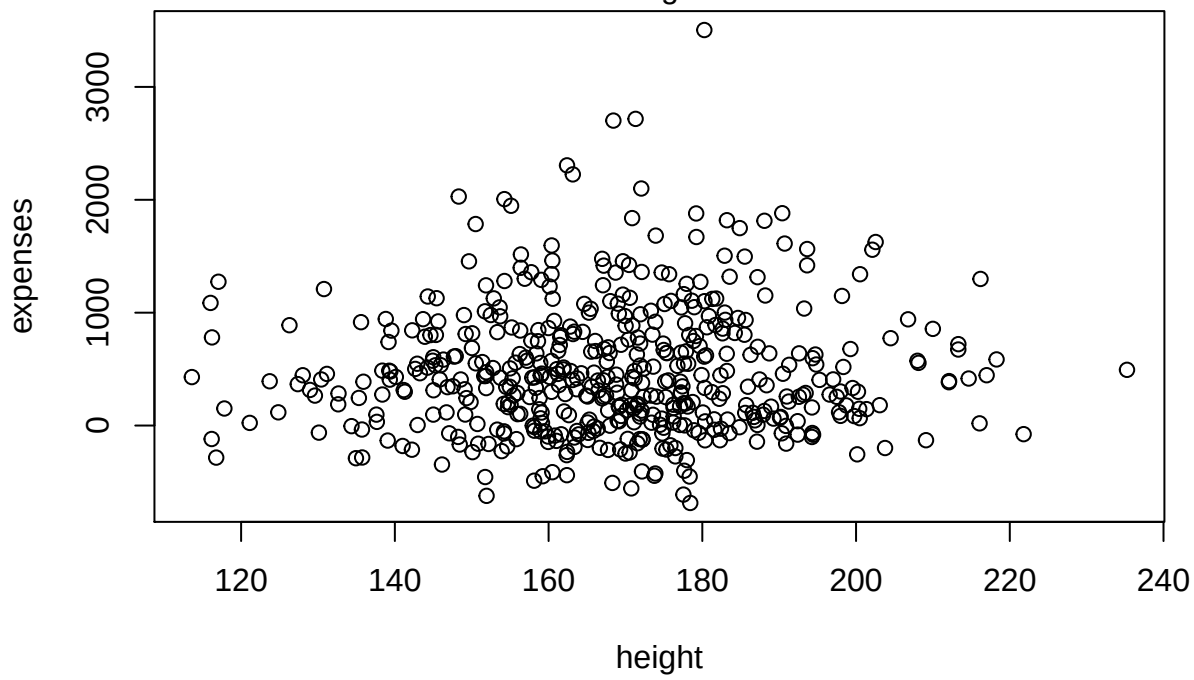
2

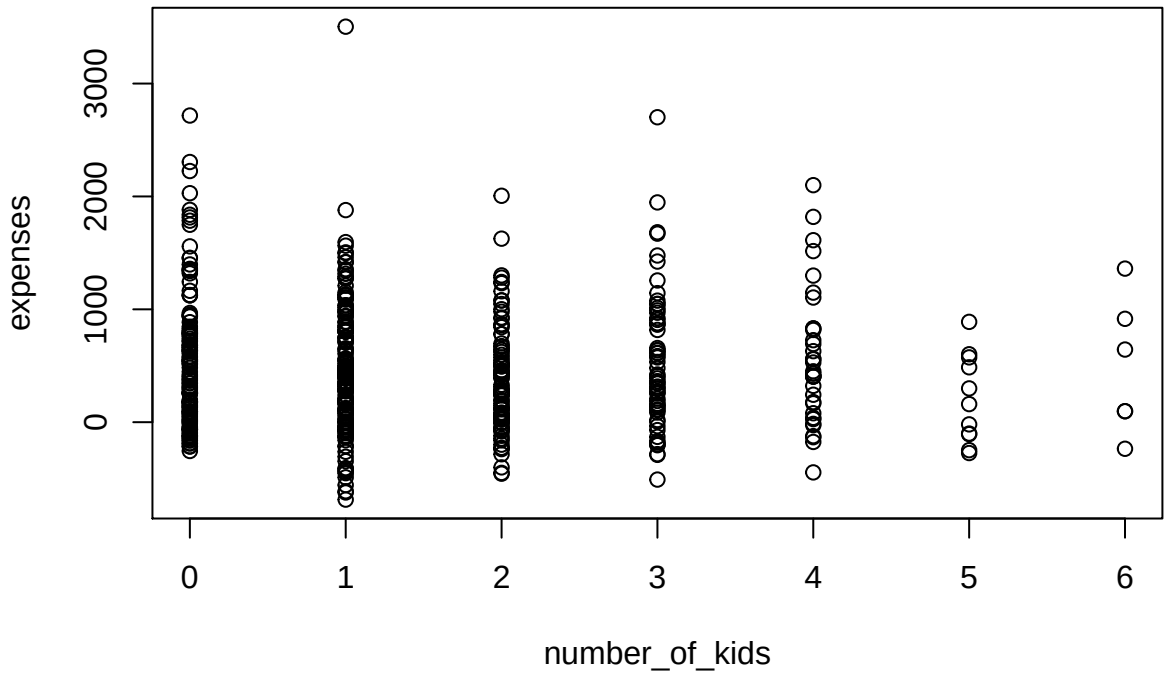
Scatter-plot ze zmiennymi ilościowymi

```
for (i in colnames(data_num)){
  if (i != "expenses") {
      plot(data_num[[i]], data_num$expenses, xlab=i, ylab="expenses")
  }
}
```



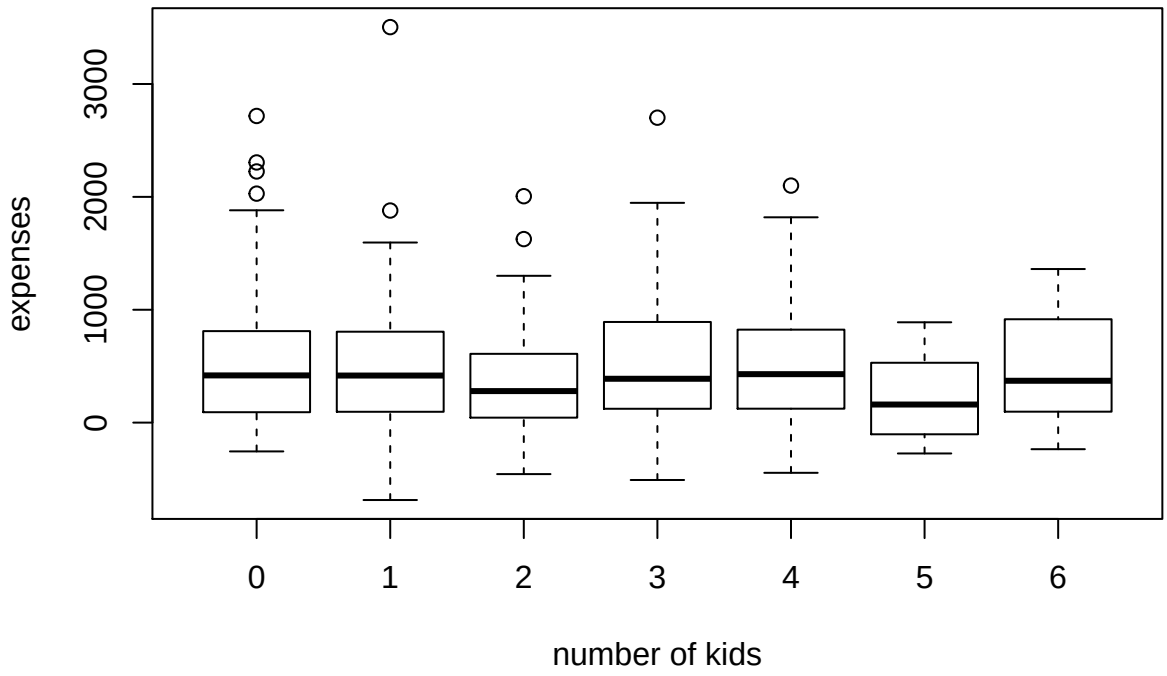

expenses
0
1000
2000
3000
20
40
60
80
100
weight


expenses
0
1000
2000
3000
120
140
160
180
200
220
240
height



Boxplot ze zmienną ilościową

```
boxplot(expenses~number_of_kids, data=data, xlab="number of kids", ylab="expenses")
```

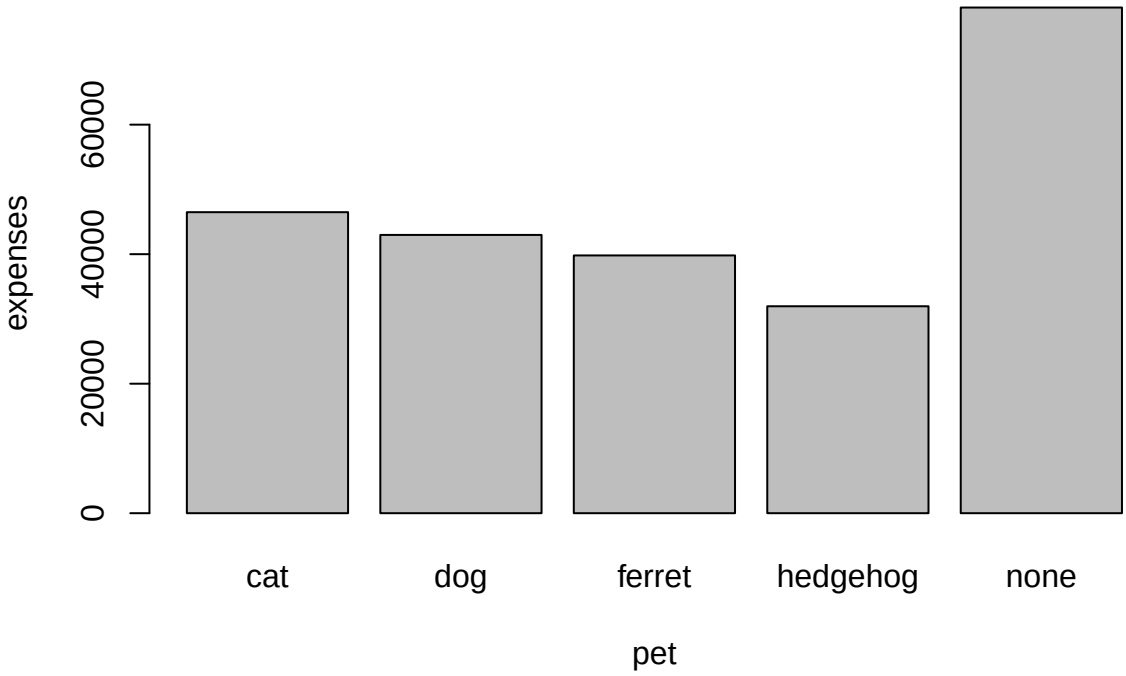


Barplot ze zmienną jakościową

```
data_bar <- data %>% group_by(pet) %>% summarise(across(expenses, sum))
data_bar
```

```
## # A tibble: 5 x 2
##   pet      expenses
##   <fct>       <dbl>
## 1 cat        46484.
## 2 dog        42970.
## 3 ferret     39805.
## 4 hedgehog   31954.
## 5 none       78086.
```

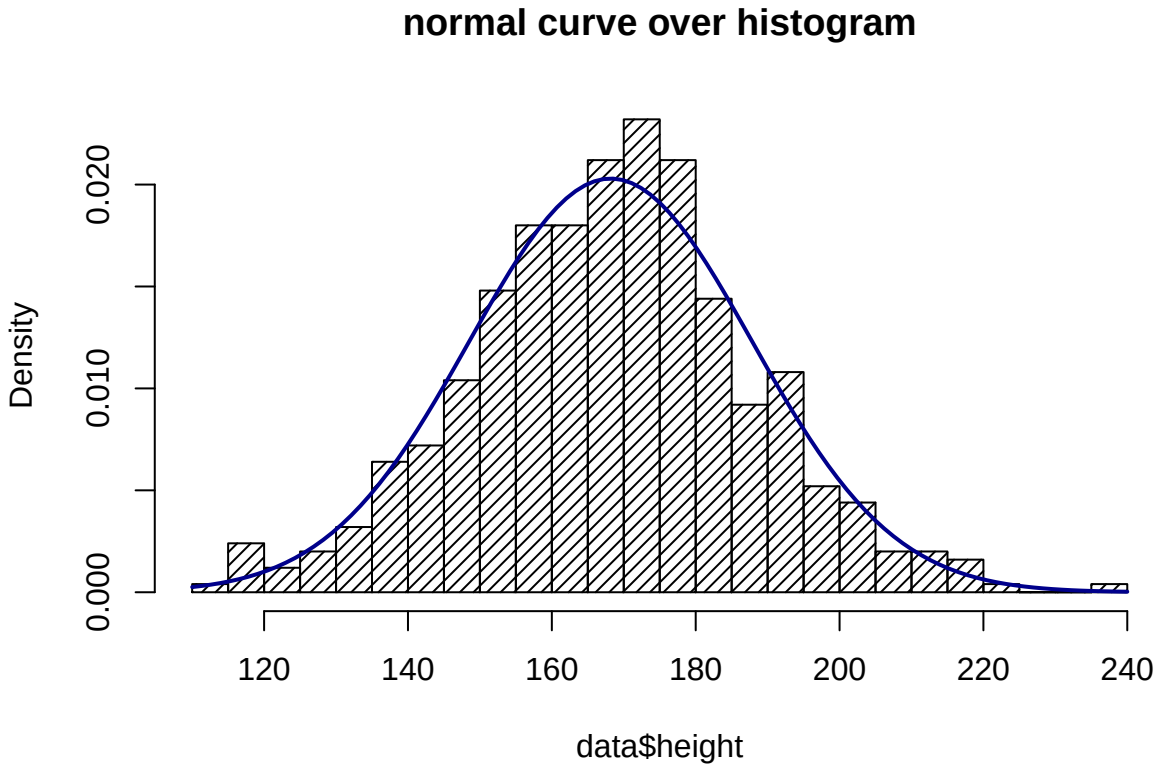
```
barplot(height=data_bar$expenses, names.arg=data_bar$pet, xlab="pet", ylab="expenses")
```



3

P value dla średniej

```
m<-mean(data$height)
std<-sd(data$height)
hist(data$height, density=20, breaks=20, prob=TRUE, main="normal curve over histogram")
curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```




```
t.test(data$height, mu=170)
```

```
##
##  One Sample t-test
##
## data:  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
```

Test odrzuca hipotezę zerową. Zakładam, że zmienna 'height' pochodzi z rozkładu normalnego. To założenie wydaje się być uprawnione. Pokazuje to wykres rozkładu normalnego na tle histogramu.

P value dla mediany

```
mediantest(x=data$height, y=c(165,165))
```

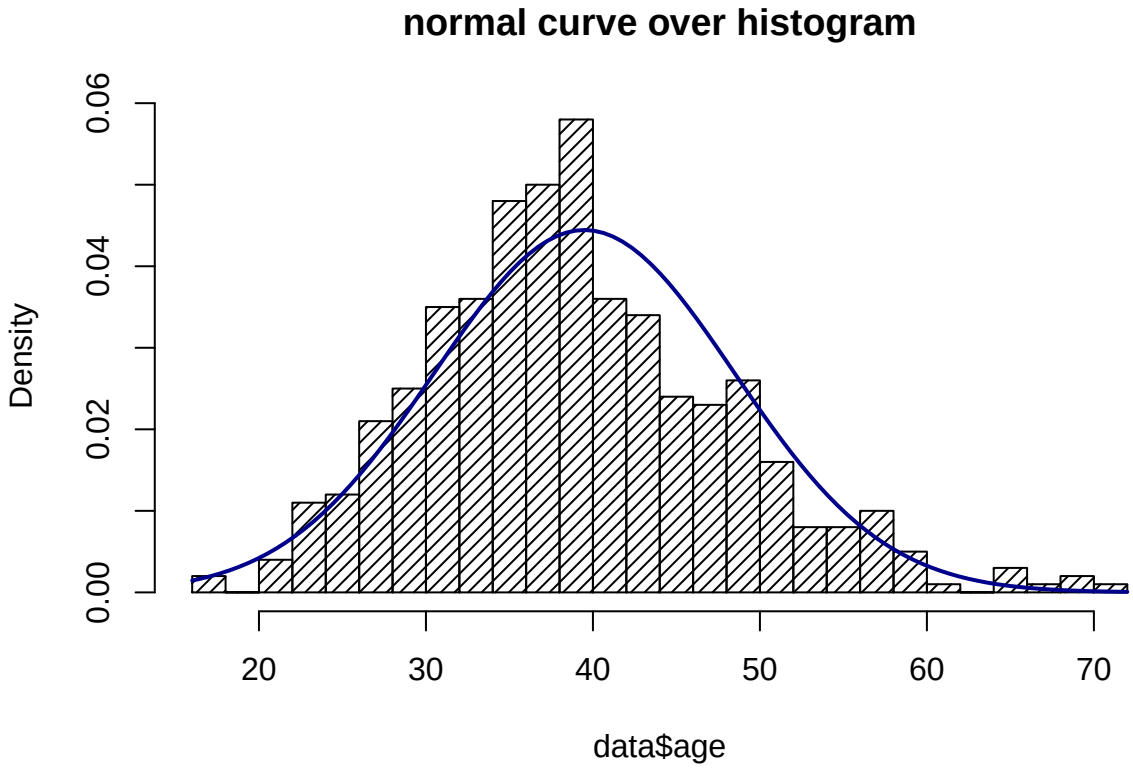
```
##
##  Exact Median Test
##
##  H0: The 2 population medians are equal.
##  HA: The 2 population medians are not equal.
```

```
##
##
##
##  Significance Level = 0.05
##  The p-value is  0.499001996007984
##  There is not enough evidence to conclude that the population medians are different at a significanc
##
```

Nie ma podstaw do odrzucenia hipotezy zerowej.

4

```
m<-mean(data$age)
std<-sd(data$age)
hist(data$age, density=20, breaks=20, prob=TRUE, main="normal curve over histogram")
curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```




```
# przedział ufności dla średniej
Mboot = boot(data$age, function(x,i) mean(x[i]), R=5000)
boot.ci(Mboot, conf = 0.99, type = c("norm", "basic" ,"perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
```

```
## CALL :
## boot.ci(boot.out = Mboot, conf = 0.99, type = c("norm", "basic",
##     "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 99%   (38.46, 40.50 )   (38.49, 40.49 )   (38.48, 40.48 )
## Calculations and Intervals on Original Scale
```

```
# przedział ufności dla odchylenia standardowego
Mboot = boot(data$age, function(x,i) sd(x[i]), R=5000)
boot.ci(Mboot, conf = 0.99, type = c("norm", "basic" ,"perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = Mboot, conf = 0.99, type = c("norm", "basic",
##     "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 99%   ( 8.167,  9.817 )   ( 8.168,  9.820 )   ( 8.132,  9.785 )
## Calculations and Intervals on Original Scale
```

```
# przedział ufności dla mediany
Mboot = boot(data$age, function(x,i) median(x[i]), R=5000)
boot.ci(Mboot, conf = 0.99, type = c("norm", "basic" ,"perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = Mboot, conf = 0.99, type = c("norm", "basic",
##     "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 99%   (38.04, 40.54 )   (38.00, 40.00 )   (38.00, 40.00 )
## Calculations and Intervals on Original Scale
```

```
# przedział ufności dla kwantyla 0.25
Mboot = boot(data$age, function(x,i) quantile(x[i], c(0.25)), R=5000)
boot.ci(Mboot, conf = 0.99, type = c("norm", "basic" ,"perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = Mboot, conf = 0.99, type = c("norm", "basic",
##     "perc"))
##
```

```
## Intervals :
## Level      Normal              Basic              Percentile
## 99%   (30.95, 34.14 )   (31.00, 34.00 )   (32.00, 35.00 )
## Calculations and Intervals on Original Scale
```

```
# przedział ufności dla kwantyla 0.75
Mboot = boot(data$age, function(x,i) quantile(x[i], c(0.75)), R=5000)
boot.ci(Mboot, conf = 0.99, type = c("norm", "basic" ,"perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = Mboot, conf = 0.99, type = c("norm", "basic",
##     "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 99%   (43.61, 47.02 )   (43.00, 47.00 )   (43.00, 47.00 )
## Calculations and Intervals on Original Scale
```

Założenia zależą od przyjętej metody. Używając metody "norm" należy założyć, że rozkład zmiennej jest bliski do normalnego. Używając "basic" należy się pogodzić z pewnymi niedokładnościamy, w szczególności, gdy rozkład jest "dziwny". Uzywając metody "perc" należy założyć, że próba zmiennej X, którą dysponujemy, ma bardzo podobny rozkład do X. Założenie, że dane pochodzą z rozkładu normalnego wydaje mi się lekko naciągane (gyby przeprowadzić test, to w zależności od przyjętego poziomu istotności otrzymamy różne wyniki). Natomiast uważam, że dana próbka danych jest reprezentatywna.

5

```
data_man <- subset(data, gender=="man")
data_woman <- subset(data, gender=="woman")
t.test(data_man$number_of_kids, data_woman$number_of_kids, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_man$number_of_kids and data_woman$number_of_kids
## t = 0.49592, df = 459.98, p-value = 0.6202
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -0.2716219  0.4004847
## sample estimates:
## mean of x mean of y
##  1.587444  1.523013
```

Test wskazuje na to, że średnia ilości dzieci mężczyzn nie jest równa średniej ilości dzieci kobiet(dla poziomu istotności 0.01).

```r
cor.test(data$age, data$expenses, alternative = c("two.sided"), method = c("pearson"), conf.level = 0.9
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$age and data$expenses
## t = 44.227, df = 498, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  0.8667704 0.9139551
## sample estimates:
##       cor
## 0.8927856
```

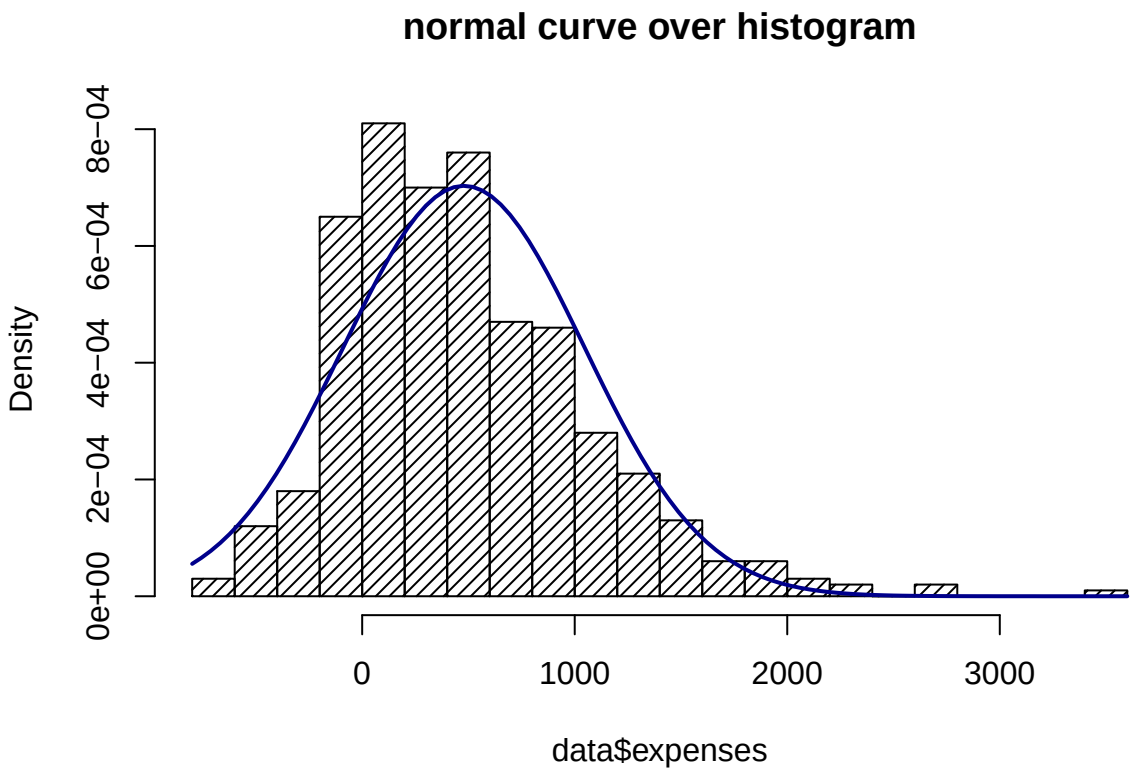
Test wskazuje na to, że zmienne są ze sobą skorelowane.

```r
chisq.test(data$married, data$gender)
```

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

Nie ma podstaw do odrzucenia hipotezy zerowej o niezależności danych.

```r
m<-mean(data$expenses)
std<-sd(data$expenses)
hist(data$expenses, density=20, breaks=20, prob=TRUE, main="normal curve over histogram")
curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```




```
shapiro_test(data$expenses)
```

```
## # A tibble: 1 x 3
##   variable      statistic  p.value
##   <chr>             <dbl>    <dbl>
## 1 data$expenses     0.946 1.75e-12
```

Test wskazuje, że dane nie pochodzą z rozkładu normalnego. Natomiast na wykresie widać pewne podobieństwo (podobnie jak w przy zmiennej "age").

6

```
dmy <- dummyVars(" ~ .", data = data, fullRank = T)
data_transformed <- data.frame(predict(dmy, newdata = data))
glimpse(data_transformed)
```

```
## Rows: 500
## Columns: 12
## $ age            <dbl> 25, 37, 41, 43, 26, 49, 27, 49, 38, 33, 44, 27, 46, 45,~
## $ weight         <dbl> 61.7, 63.9, 50.2, 72.4, 78.4, 59.4, 67.5, 82.3, 64.1, 7~
## $ height         <dbl> 121.12, 145.00, 145.03, 179.90, 163.91, 151.86, 169.31,~
## $ gender.other   <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ gender.woman   <dbl> 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0~
## $ marriedTRUE    <dbl> 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0~
## $ number_of_kids <dbl> 2, 6, 2, 1, 1, 2, 1, 0, 5, 2, 3, 4, 4, 0, 1, 1, 1, 1, 0~
```

```
## $ pet.dog        <dbl> 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0~
## $ pet.ferret     <dbl> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ pet.hedgehog   <dbl> 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ pet.none       <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0~
## $ expenses       <dbl> 23.44299, 96.83683, 312.67693, 447.42838, -78.22799, 12~
```

```
model <- lm(expenses ~ ., data = data_transformed)
summary(model)
```

```
##
## Call:
## lm(formula = expenses ~ ., data = data_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -768.28 -127.24   -2.79  130.54  905.98
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2276.7455   100.1447 -22.735  < 2e-16 ***
## age               57.5889     1.0712  53.759  < 2e-16 ***
## weight             1.2078     0.9984   1.210  0.22698
## height             2.0637     0.6580   3.136  0.00182 **
## gender.other      44.1656    37.7199   1.171  0.24222
## gender.woman     -21.7164    20.1406  -1.078  0.28146
## marriedTRUE       -9.8079    25.9461  -0.378  0.70559
## number_of_kids   -12.4393     8.8614  -1.404  0.16103
## pet.dog           29.2695    30.0801   0.973  0.33101
## pet.ferret       406.6324    36.3026  11.201  < 2e-16 ***
## pet.hedgehog     242.0460    35.9101   6.740 4.47e-11 ***
## pet.none          21.7454    26.2294   0.829  0.40748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213.9 on 488 degrees of freedom
## Multiple R-squared:  0.861,  Adjusted R-squared:  0.8579
## F-statistic: 274.9 on 11 and 488 DF,  p-value: < 2.2e-16
```

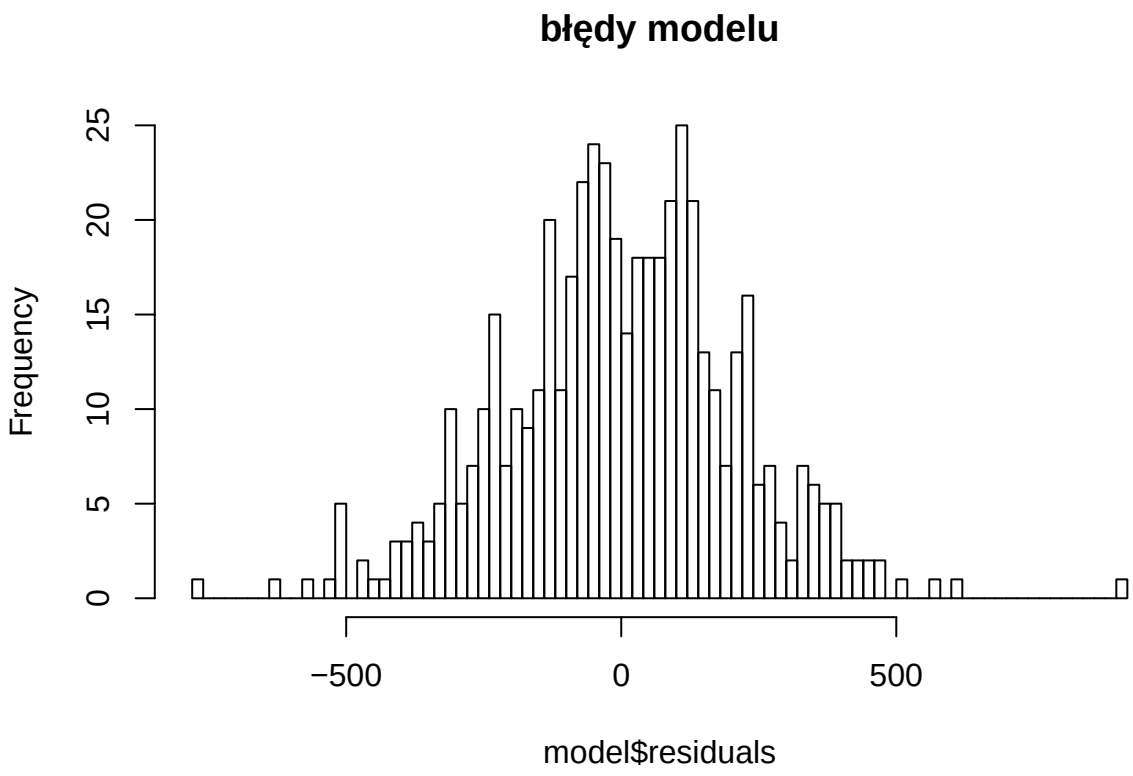
```
#RSS and sqrt(RSS / n)
sum(model$residuals^2)
```

```
## [1] 22328238
```

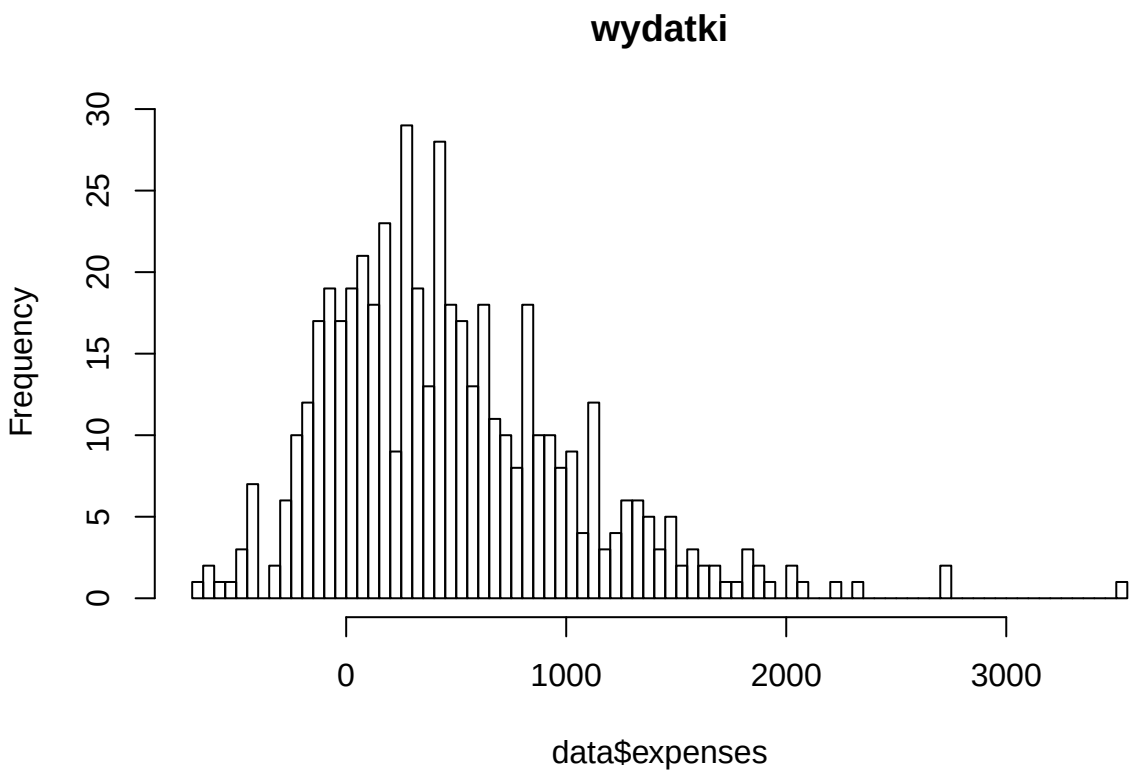
```
sqrt(sum(model$residuals^2) / length(data$expenses))
```

```
## [1] 211.3208
```

```
hist(model$residuals, breaks=80, main="błędy modelu")
```




```
hist(data$expenses, breaks=60, main="wydatki")
```



Model nie jest zbyt dokładny, ale daje szanse na przybliżenie wydatków. \ Na podstawie p-wartości opisujących zmienne używane przez model odrzucam marriedTrue.

```r
model <- lm(expenses ~ . - marriedTRUE, data = data_transformed)
summary(model)
```

```
##
## Call:
## lm(formula = expenses ~ . - marriedTRUE, data = data_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -768.63 -130.81   -1.07  131.16  909.18
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2276.5344   100.0553 -22.753  < 2e-16 ***
## age               57.5796     1.0700  53.811  < 2e-16 ***
## weight             1.1893     0.9963   1.194  0.23318
## height             2.0707     0.6572   3.151  0.00173 **
## gender.other      43.7879    37.6736   1.162  0.24568
## gender.woman     -22.5991    19.9873  -1.131  0.25875
## number_of_kids   -14.5238     6.9304  -2.096  0.03663 *
## pet.dog           30.2990    29.9303   1.012  0.31189
## pet.ferret       407.5562    36.1885  11.262  < 2e-16 ***
## pet.hedgehog     241.7696    35.8711   6.740 4.47e-11 ***
## pet.none          22.3861    26.1516   0.856  0.39241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213.7 on 489 degrees of freedom
## Multiple R-squared:  0.861,  Adjusted R-squared:  0.8582
## F-statistic: 302.9 on 10 and 489 DF,  p-value: < 2.2e-16
```

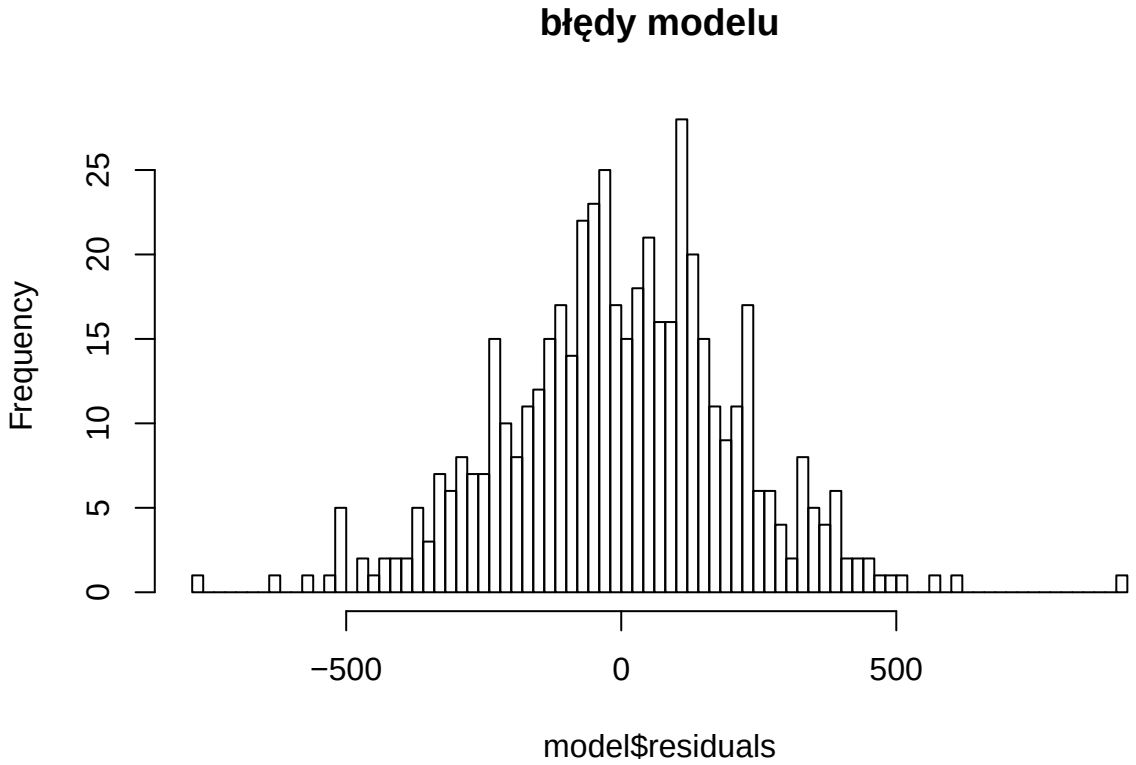
```r
#RSS and sqrt(RSS / n)
sum(model$residuals^2)
```

```
## [1] 22334776
```

```r
sqrt(sum(model$residuals^2) / length(data$expenses))
```

```
## [1] 211.3517
```

```r
hist(model$residuals, breaks=80, main="błędy modelu")
```



Widzimy, że model nieuwzględniający stanu cywilnego jest nieznacznie gorszy od modely uwzgędniającego wszystkie dane którymi dysponujemy.