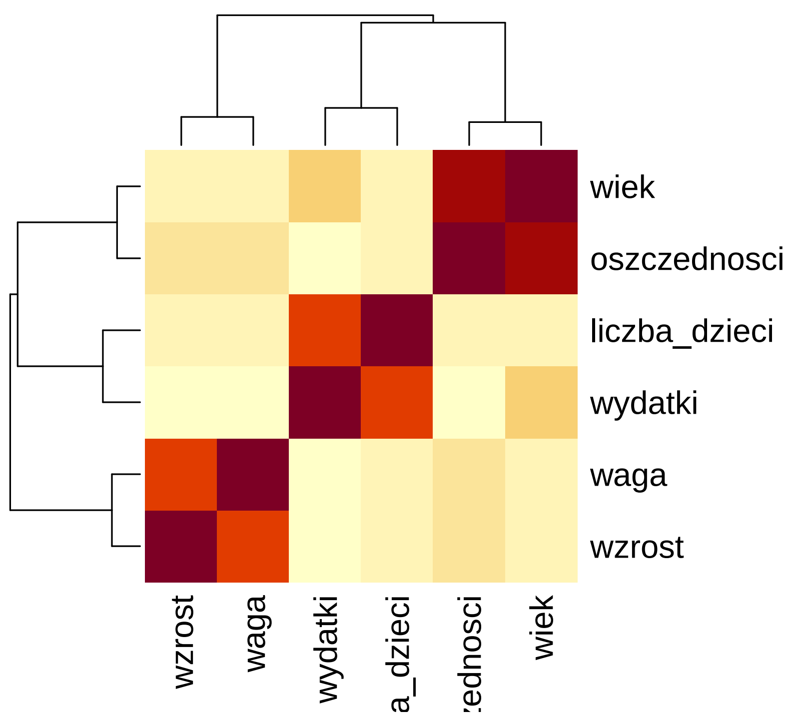
projekt1

Julia Gołębiowska

1

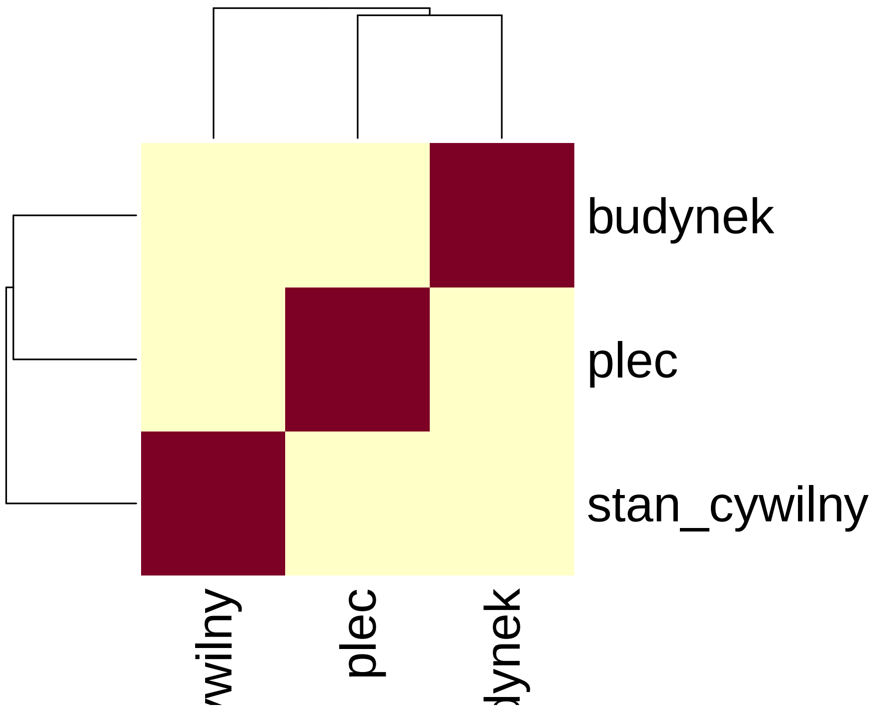
Wczytaj dane, obejrzyj je i podsumuj w dwóch-trzech zdaniach. Pytania pomocnicze: ile jest obserwacji, ile zmiennych ilościowych, a ile jakościowych? Czy są zależności w zmiennych objaśniających (policz i zaprezentuj na wykresach korelacje pomiędzy zmiennymi ilościowymi, a także zbadaj zależność zmiennych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

```
setwd("~/Dokumenty/VIrok/SAD")
#read data
data = read.delim("people.tab.csv", sep="\t")
#Lets check corelation between variables (ilościowe)
p.corr <- cor(data[, c(1,2,3,6,8,9)])
heatmap(p.corr, scale = "none")
```




```
#change to factor
data[, 4] <- as.factor(data[, 4])
data[, 5] <- as.factor(data[, 5])
data[, 7] <- as.factor(data[, 7])
#spearman correlation for categorical variables
#eliminate rows with NA values
```

```r
not_NA_values <- !is.na(data$plec)
plec <- as.numeric(as.factor(data[not_NA_values, 4]))
stan_cywilny <- as.numeric(as.factor(data[not_NA_values, 5]))
budynek <- as.numeric(as.factor(data[not_NA_values, 7]))
#new dataframe
for_corr_data <- data.frame(plec=plec, stan_cywilny=stan_cywilny, budynek=budynek)
s.corr <- cor(for_corr_data, method = c("spearman"))
heatmap(s.corr, scale = "none")
```



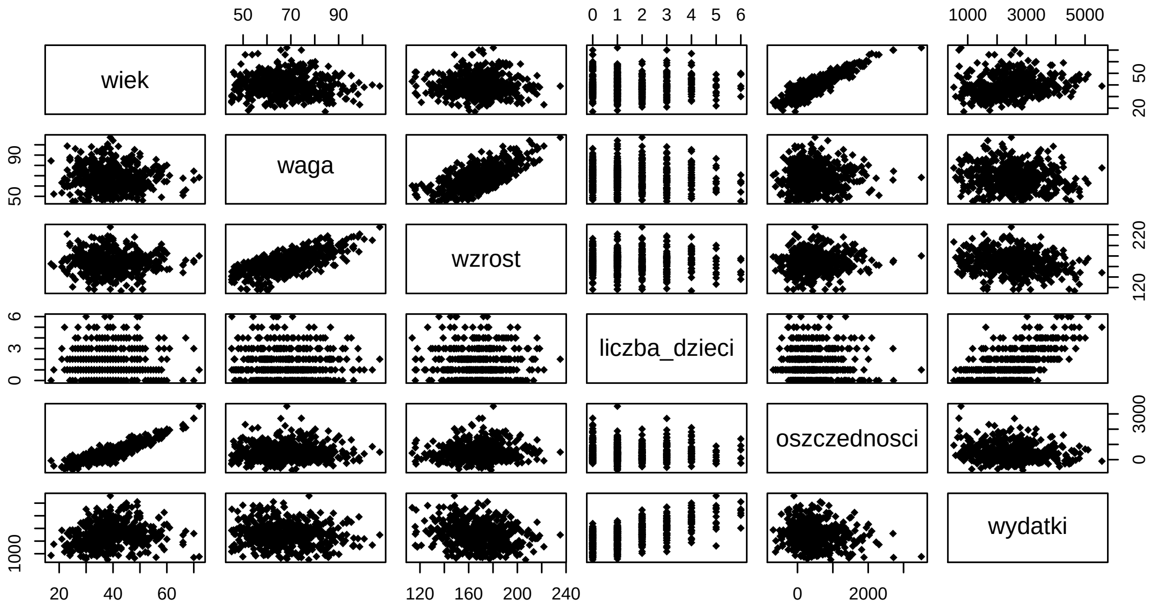
W zbiorze jest 500 obserwacji. Jest 6 zmiennych ilościowych i 3 zmienne jakościowe. Najwyższe korelacje występują pomiędzy wiekiem i oszczędnościami, liczbą dzieci i wydatkami oraz wagą i wzrostem. W danych występuje 38 brakujących obserwacji w kolumnie płeć. Między zmiennymi jakościowych korelacje wynoszą ok zero, czyli nie ma między nimi korelacji.

2

2. Podsumuj dane przynajmniej trzema różnymi wykresami. Należy przygotować:

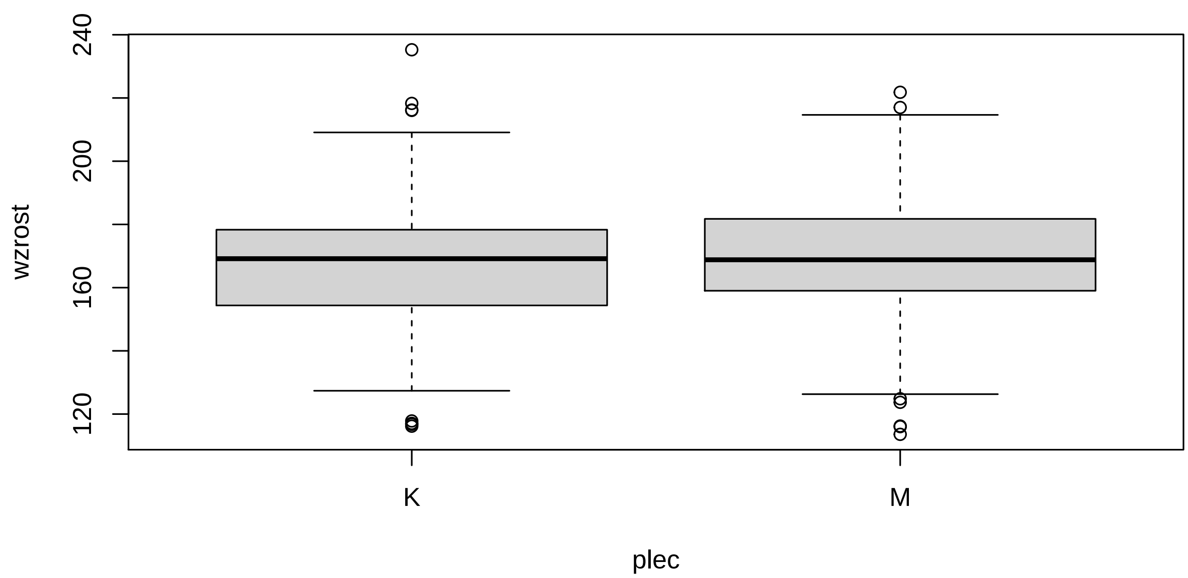
a) wykres typu scatter-plot (taki jak na wykładzie 7, slajd 3) dla wszystkich zmiennych objaśniających ilościowych i zmiennej objaśnianej.
b) Wykresy typu pudełkowy (boxplot) dla jednej wybranej zmiennej ilościowej w podziale na płeć respondentów.
c) Wykres kołowy (pie chart) dla jednej wybranej zmiennej jakościowej (wykres ma zawierać etykiety z procentami wystąpień danych kategorii).

```r
#a
pairs(data[,c("wiek", "waga", "wzrost", "liczba_dzieci", "oszczednosci", "wydatki")], pch=18)
```



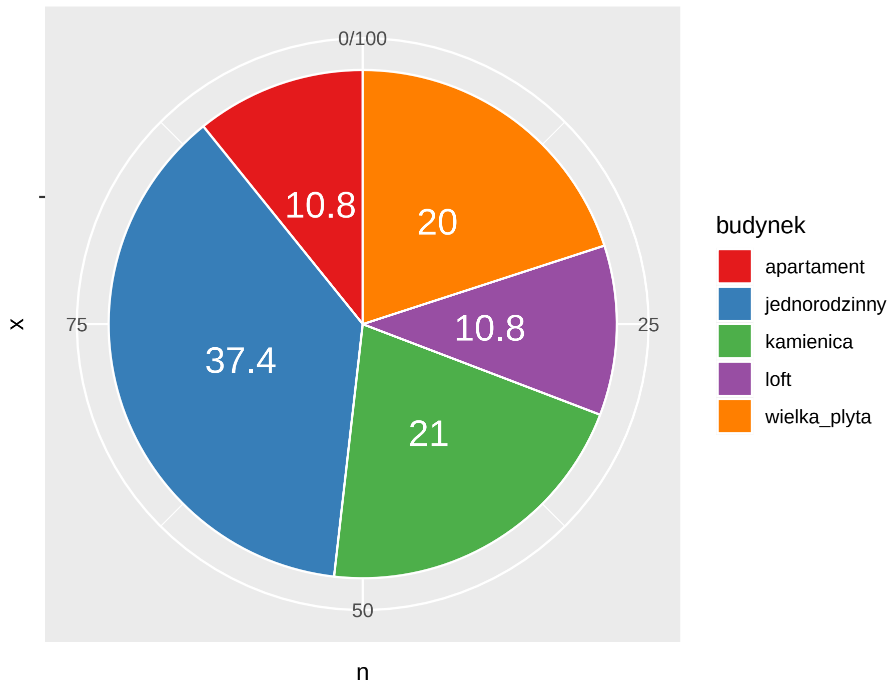
Między zmiennymi ilościowymi występują głównie zależności linniowe lub zbliżone do linniowych. Niektóre jak np. między wiekiem i wagą mają bardzo dużą wariancje i nie widać żadnego trendu.

```
#b
boxplot(wzrost ~ plec, data = data)
```



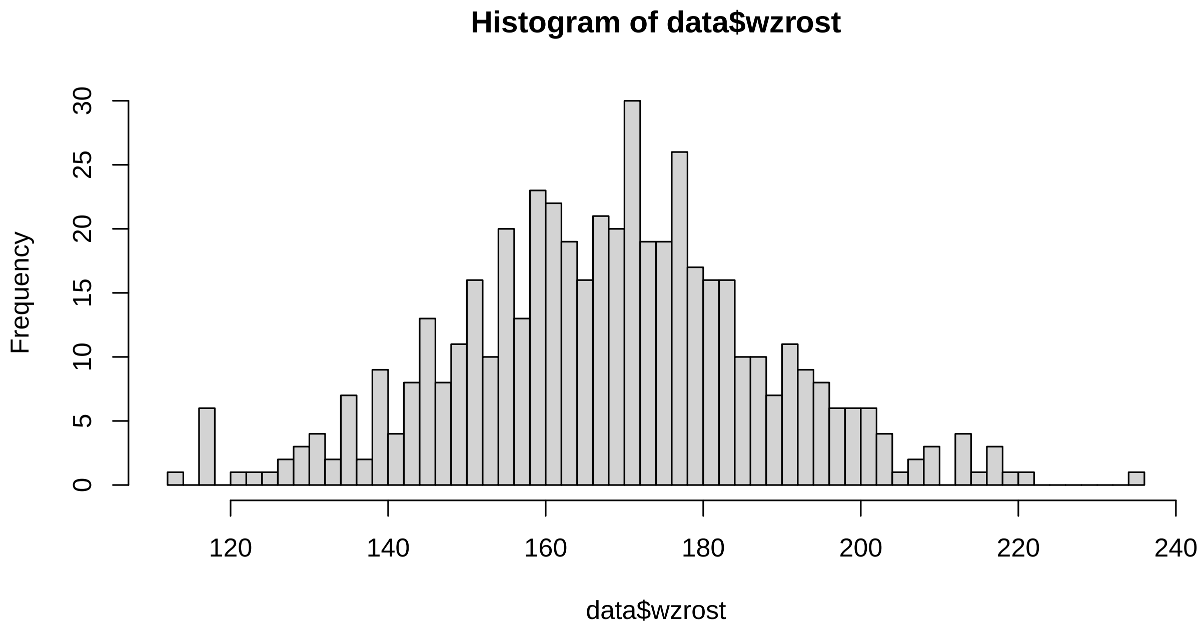
Rozkłady wzrostu u różnych płci są zbliżone. Mają one podobną medianę, ale rozkład wzrostu dla mężczyzn jest przesunięty w kierunku wyższych wartości.

```
#c
#calculate procentage of building type
c_budynek <- data %>% count(budynek)
s_budynek <- c_budynek$n %>% sum
c_budynek$n <-c_budynek$n/s_budynek *100
p1 = ggplot(c_budynek, aes(x="", y=n, fill=budynek))  + geom_bar(stat="identity", width=1, color="white
p1 +  scale_fill_brewer(palette="Set1")
```

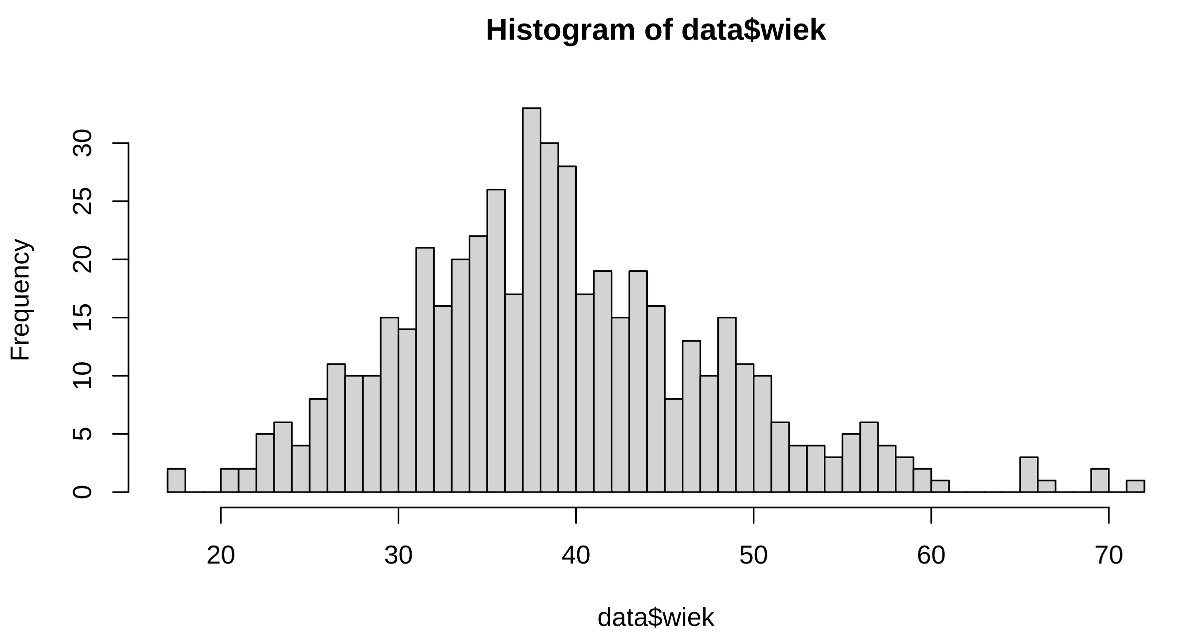


Najwięcej ludzi zmaieszkuje domu jednorodzinne, a najmniej apartamenty i lofty.

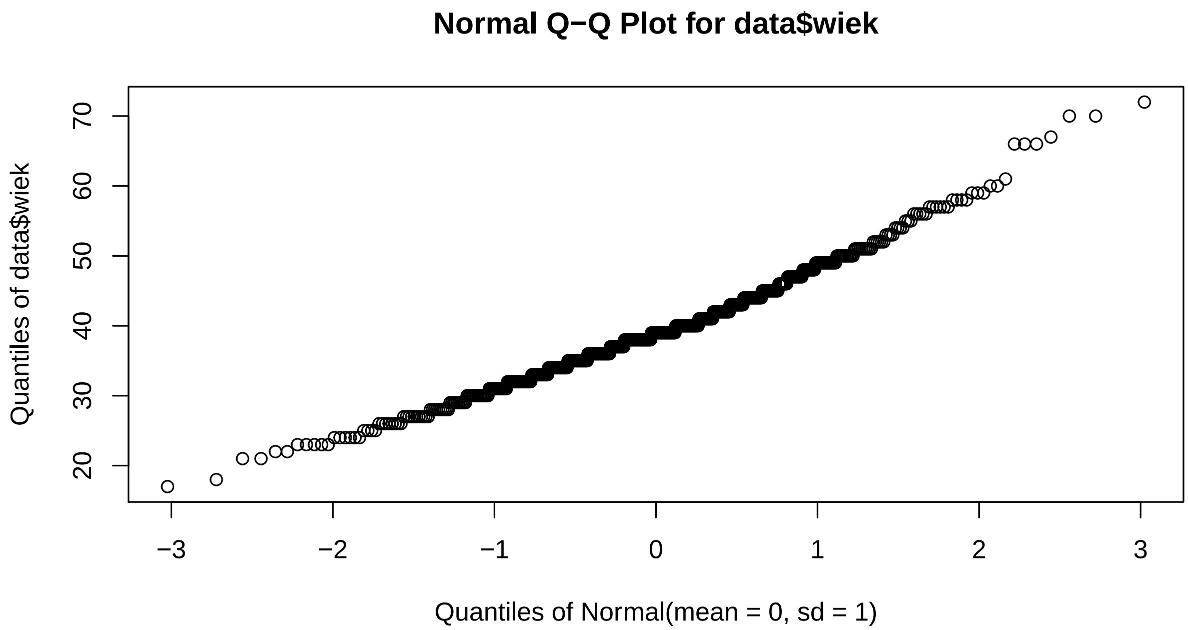
```
#added
hist(data$wzrost, breaks = 50)
```




```
#added
hist(data$wiek, breaks = 50)
```



```r
qqPlot(data$wiek)
```



Rozkłady wzorstu i wieku są zbliżone do normalnego.

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. (2 pkt)

```r
shapiro.test.wzrost.res <- shapiro.test(data$wzrost)
t.test.res<-t.test(data$wzrost, mu=170, alternative = c("less"))
wilcox.test.res <-wilcox.test(data$wzrost, md = 165, alternative = "less")
wilcox.test.res
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data$wzrost
## V = 125250, p-value = 1
## alternative hypothesis: true location is less than 0
```

P-wartość dla średniej wzrostu w powyższym teście 0.019487. Test t-studenta zakłada rozkład normalny - można przyjąć to założenie biorąc pod uwagę histogram dla zmiennej wzrost oraz wysoka p-wartość dla testu Shapiro-Wilka (0.3777816). P-wartość dla mediany wzrostu równej 165 wynosi aż 1. Dla tego testu nie potrzebujemy szczególnych założeń.

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 ¾. Podaj założenia, z jakich korzystałeś i skomentuj czy wydają Ci się uprawnione (2 pkt).

```
#1
# dla średniej
shapiro.test.res <- shapiro.test(data$wiek)
mu1 <- mean(data$wiek)
sigma <- var(data$wiek)
sd <- sd(data$wiek)
lower_bound_mean  <- mu1+qt(c(0.005), df = nrow(data)-1)*sd/sqrt(nrow(data)-1)
upper_bound_mean <- mu1+qt(c(0.995), df = nrow(data)-1)*sd/sqrt(nrow(data)-1)
res_var_test <- varTest(data$wiek, alternative = "two.sided", conf.level = 0.99, sigma.squared = sigma,
lower_bound_var <- res_var_test$conf.int["LCL"]
upper_bound_var <- res_var_test$conf.int["UCL"]
#dla sd
```

Założyłam rozkład normalny. Mimo niskiej p-wartość testu Shapiro-Wilka (6.5890806×10^{-6}), to biorąc pod uwagę histogram oraz QQ-plot dla tej zmiennej w części 2, uważam to założenie za uprawnione. Przedział ufności dla średniej wynosi (38.4449637, 40.5230363), a dla wariancji (68.832637, 95.4160336).

```
cl_025 <- quantileCI(x=data$wiek, prob=c(0.25), method="asymptotic",conf.level=0.99)
lower_bound_025<- cl_025$conf.int[1]
upper_bound_025<- cl_025$conf.int[2]
cl_05 <- quantileCI(x=data$wiek, prob=c(0.5), method="asymptotic",conf.level=0.99)
lower_bound_05<- cl_05$conf.int[1]
upper_bound_05<- cl_05$conf.int[2]
cl_075 <- quantileCI(x=data$wiek, prob=c(0.75), method="asymptotic",conf.level=0.99)
lower_bound_075<- cl_075$conf.int[1]
upper_bound_075<- cl_075$conf.int[2]
```

Przedział ufności dla kwantylu 0.25 wynosi (32, 35), dla kwantylu 0.5 (38, 40), a dla kwantylu 0.75 (43, 47).

5

Przetestuj na poziomie istotności 0.01 trzy hipotezy: 1. średnie wartości wybranej zmiennej pomiędzy osobami zamężnymi/żonatymi a pannami/kawalerami są równe; 2. dwie wybrane zmienne ilościowe są niezależne; 3. dwie wybrane zmienne jakościowe są niezależne. 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. Każda hipoteza po 1 punkcie (w sumie 4). Punktowane jest sformułowanie hipotezy zerowej, wybranie właściwego testu, przeprowadzenie testu i podjęcie decyzji czy odrzucamy hipotezę zerową

```
#1 dla zmiennnej wiek
married <- data$wiek[data$stan_cywilny == TRUE]
not_married  <- data$wiek[data$stan_cywilny == FALSE]
t.test.res <- t.test(married, not_married, conf.level = 0.99)
```

Z powodu dużej p-wartości 0.5575688 » 0.01 nie mogę odrzucić hipotezy zerowej o braku różnic dla średnich wieku u grupy ludzi zamężnych/żonatych i grupy kawalerów/panien. Założyłam rozkład normalny zmiennej wiek. Biorąc pod uwagę histogram tej zmiennej oraz niską p-wartość dla testu Shapiro-Wilka (6.5890806×10^{-6}) uważam to założenie za uprawnione.

```
#2 dla zmiennych wiek i wydatki
corr.res <- cor.test(data$wiek, data$wydatki)
corr.res
##
##  Pearson's product-moment correlation
##
## data:  data$wiek and data$wydatki
## t = 4.0577, df = 498, p-value = 5.753e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0926583 0.2624684
## sample estimates:
##       cor
## 0.1788953
```

Bardzo niska p-wartość (5.7533512×10^{-5} « 0.01) pozwala nam odrzucić hipotezę zerową o niezależności zmiennych na poziomie istotności 0.01. Założyłam, że zmienne są ciągłe - co wynikało bezpośrednio z zadania.

```
#3 dla zmiennych płeć i budynek
#lets not take into account rows with NA values
ct<-data.frame(data[not_NA_values,c("plec", "budynek")])
vec_count <- ct %>% count(plec, budynek)
contingency_table <- matrix(vec_count$n, ncol=2)
contingency_table
##      [,1] [,2]
## [1,]   24   26
## [2,]   92   82
## [3,]   48   51
## [4,]   27   20
## [5,]   48   44
chisq.test.res <- chisq.test(contingency_table, correct=F)
```

Z powodu dużej p-wartości 0.8425133» 0.01 nie mogę odrzucić hipotezy zerowej o braku zależności między płcią o rodzajem zamieszkałego budynku. Założyłam, że zmienne są ketegoryczne co wynika z zadania; obserwacje są niezależne - tak, są to osobne obserwacje dla każdej osoby; jedna obserwacja może się znajdować tylko w jednej komórce tabeli kontugencji - tak, ponieważ warunki są rozłączne - mamy jeden budynek przyporządkowany do osoby i zawartość komórek powinna być większa niż 5 w 80% komórek - tak, ponieważ najmniejsza wartość to 24.

```
#4 sprawdźmy czy zmienna oszczednosci ma rozklad normalny
y = rnorm(500)
ks.test.res <- ks.test(data$oszczednosci,y)
```

P-wartość wynosi 0, zatem na poziomie istotności 0.01 możemy odrzucić hipotezę zerową o rozkładzie normalnym o średniej 0 i odchyleniu standardowym równym 1.

6

Oszacuj model regresji liniowej, przyjmując za zmienną zależną (y) bilans dochodów na koniec miesiąca (oszczednosci) a jako zmienne niezależne (x) przyjmując pozostałe zmienne. Rozważ, czy konieczne są transformacje zmiennych (objaśniających lub 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 oszczednosci). Aby dokonać wyboru takiej zmiennej, dla każdej ze zmiennych objaśniających sprawdź: - Jaką ma p-wartość w pełnym modelu? - O ile zmniejsza się R2 , 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. Przedstaw (i skomentuj) wykresy diagnostyczne: wykres zależności reszt od zmiennej objaśnianej, wykres reszt studentyzowanych i dźwigni

```r
y_full <- lm(data = data[not_NA_values,], oszczednosci ~ .) #because we want to compere modeles we elim
y_full_summary <- summary(y_full)
#check R^2
r_squared_full <- y_full_summary$r.squared
#save p.values
p.values <-y_full_summary$coefficients[,4]
#eliminate NA values - is it correct? -better replace by 0? is it the plec ?
y_pred_full = predict(y_full,  data[not_NA_values,], interval='prediction')[,"fit"]
y_true = data$oszczednosci[not_NA_values]
#count RSS for foull model
RSS_full = sum((y_true - y_pred_full)^2)
#prepare models
#prepare formulas
Vars <- as.list(colnames(data[,-length(data[1,])])) # I dont need value oszczednosci
allModelsList <- lapply(paste("oszczednosci ~ . - ", Vars), as.formula)
allModelsResults <- lapply(allModelsList, function(x) lm(x, data = data[not_NA_values,]))
allModelssummary <- lapply(allModelsResults, summary)

fstats <- function(list){
  list$r.squared
}
allModelsstatistics <- lapply(allModelssummary, fstats)

prediction <- function(model){
   predict(model,  data[not_NA_values,], interval='prediction')[,"fit"]
}

allModelspredicted <- lapply(as.vector(allModelsResults), prediction)
allModelspredicted <- as.data.frame(allModelspredicted)
#count RSS
colnames(allModelspredicted) <- Vars
RSS <-function(vec1, vec2){
  sum((vec1-vec2)^2)
}

RSSres <- apply(allModelspredicted, 2, RSS, y_true)
#put all together
allModelsstatistics <- unlist(allModelsstatistics)
stats <- as.data.frame(cbind(RSSres, R_square=allModelsstatistics))
#choose from p-values eliminate various buildings
p.values <- t(as.data.frame(p.values))
no_buildings <- p.values[,c(-1,-8,-9,-10,-11)]
#condence p.values of all buildings

no_buildings["budynek"] <- paste(as.vector(p.values[,c(8, 9, 10, 11)]), sep=" ", collapse=" ")
no_buildings <- as.data.frame(t(no_buildings))
```

```r
#reorder columns
no_buildings <-no_buildings[, c(1:6,8,7)]
#build table with statistics
stats <- cbind(stats, p.value = unlist(no_buildings))
stats$RSSreschange <- stats$RSSres - rep(c(RSS_full), times=length(stats$RSSres))
stats$R_squarechange <- stats$R_square - rep(c(r_squared_full), times=length(stats$R_square))
stats
##                  RSSres   R_square
## wiek          136929242 0.04241364
## waga            5180546 0.96377092
## wzrost          5158668 0.96392392
## plec            4681548 0.96726056
## stan_cywilny    4682661 0.96725277
## liczba_dzieci  10988814 0.92315200
## budynek        13661076 0.90446409
## wydatki        19275685 0.86519948
##                                                                                       p.value
## wiek                                                                                        0
## waga                                                                     1.49011162168742e-11
## wzrost                                                                   3.93728564678018e-11
## plec                                                                        0.886080559289635
## stan_cywilny                                                                0.721086110052031
## liczba_dzieci                                                            2.06138736792693e-85
## budynek       2.25639059750172e-25 8.45959310256749e-51 5.9003727481203e-35 5.81625115689573e-98
## wydatki                                                                 2.1935604428315e-140
##               RSSreschange R_squarechange
## wiek          1.322479e+08  -9.248484e-01
## waga          4.992122e+05  -3.491137e-03
## wzrost        4.773341e+05  -3.338138e-03
## plec          2.137629e+02  -1.494907e-06
## stan_cywilny  1.327566e+03  -9.284059e-06
## liczba_dzieci 6.307480e+06  -4.411006e-02
## budynek       8.979742e+06  -6.279797e-02
## wydatki       1.459435e+07  -1.020626e-01
```

Biorąc pod uwagę wykresy punktowe dla zmiennych ilościowych z części 2, nie ma potrzeby przekształcania danych - mają one trend liniowy.
Dla pełnego modelu RSS wynosi 4.6813339×10^6, a R^2 0.9672621.

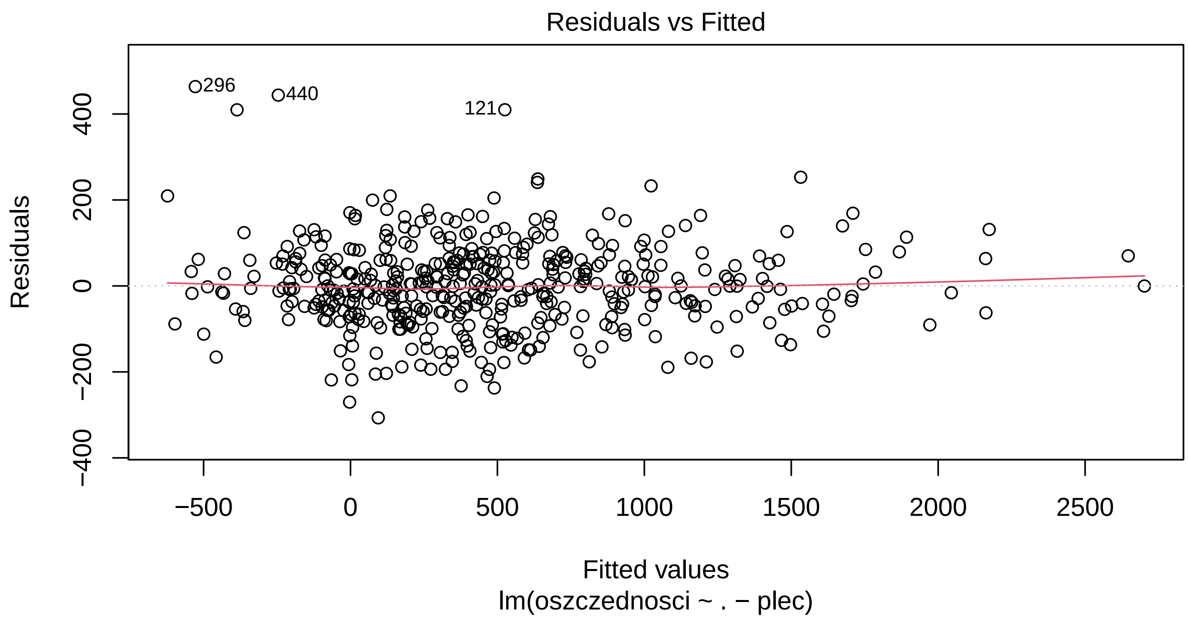
Tutaj podane są współczynniki (Estmate) i p-wartości (Pr(>|t|)) dla wszytkich zmiennych tego modelu.

```r
y_full_summary$coefficients
##                         Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept)          -873.5910639 58.76097840 -14.8668570  5.903953e-41
## wiek                   63.9425778  0.56711911 112.7498195  0.000000e+00
## waga                    3.9440904  0.56935460   6.9273005  1.490112e-11
## wzrost                 -2.3846402  0.35203850  -6.7738050  3.937286e-11
## plecM                   1.3806850  9.63178947   0.1433467  8.860806e-01
## stan_cywilnyTRUE       -4.6125227 12.91186549  -0.3572313  7.210861e-01
## liczba_dzieci         151.6035490  6.15686993  24.6234776  2.061387e-85
## budynekjednorodzinny -182.0703090 16.43991452 -11.0748939  2.256391e-25
## budynekkamienica     -305.6314445 17.89019676 -17.0837386  8.459593e-51
## budynekloft          -338.4700061 25.14077670 -13.4629892  5.900373e-35
## budynekwielka_plyta  -564.2601455 20.59225436 -27.4015723  5.816251e-98
```

```
## wydatki                    -0.3959261  0.01057062 -37.4553491 2.193560e-140
```

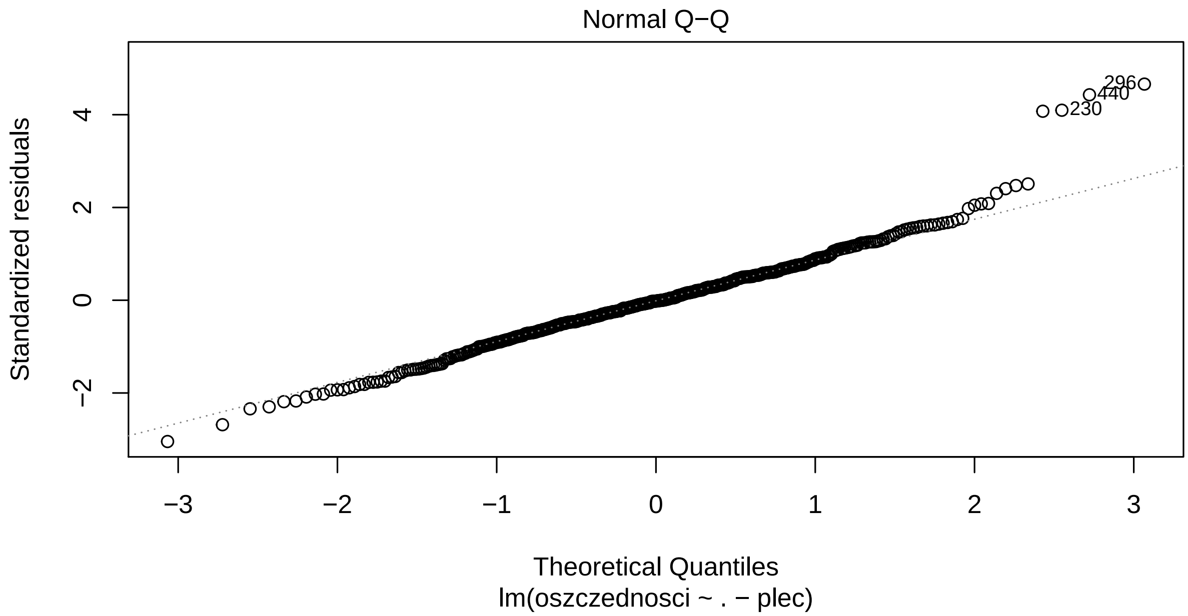
Tutaj podsumowanie zmian RSS, R^2 dla modeli bez poszczególnych zmiennych oraz p-value dla tych zmiennych w pełnym modelu. Najmniej zwiększył się RSS dla zmiennej płeć (o 213.762914), dla tej zmiennej również $ Rˆ2$ zmniejszył się najmniej (o 1.4949068×10^{-6}). Zmienna ta nie jest istotna statystycznie. Jej p-wartość wynosi aż 0.89. Ponieważ usunięcie zmiennej płeć nie powoduje znaczącego wzrostu RSS (nie zwiększa znacząco błędu modelu) oraz nie zmniejsza znacząco R^2 - czyli nie wplywa znacząco na wyjaśnienie wariancji w modelu oraz nie jest istotna statystycznie to typuję ją do usunięcia z modelu.

```
model <-  lm(data = data, oszczednosci ~ . -plec) #we can use all data as only in plec colum were NA va
y_pred_model = predict(model,  data[not_NA_values,], interval='prediction')[,"fit"]
plot(model, which=1)
```



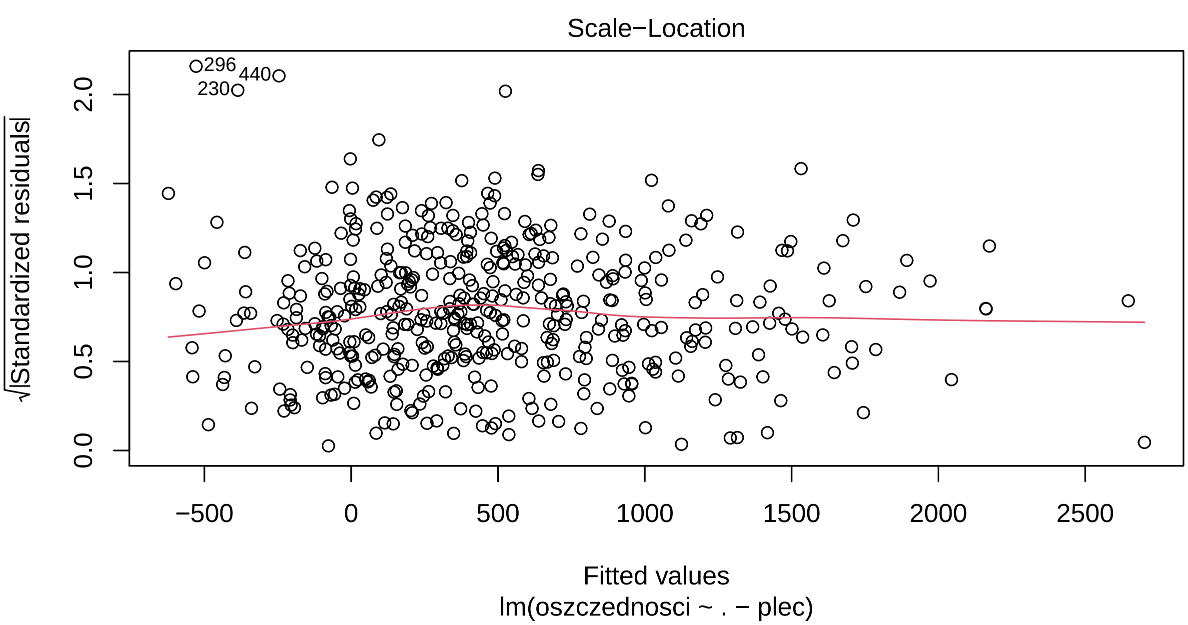
Wykres jest zbliżony do linni równoległej do osi OX. Zatem zależność w danych jest linniowa. Możemy pobieżnie stwierdzić również losowość błędów - różnice między sąsiednimi residuami nie są mniejsze niż różnice dla dalszych residuów.

```
plot(model, which=2)
```



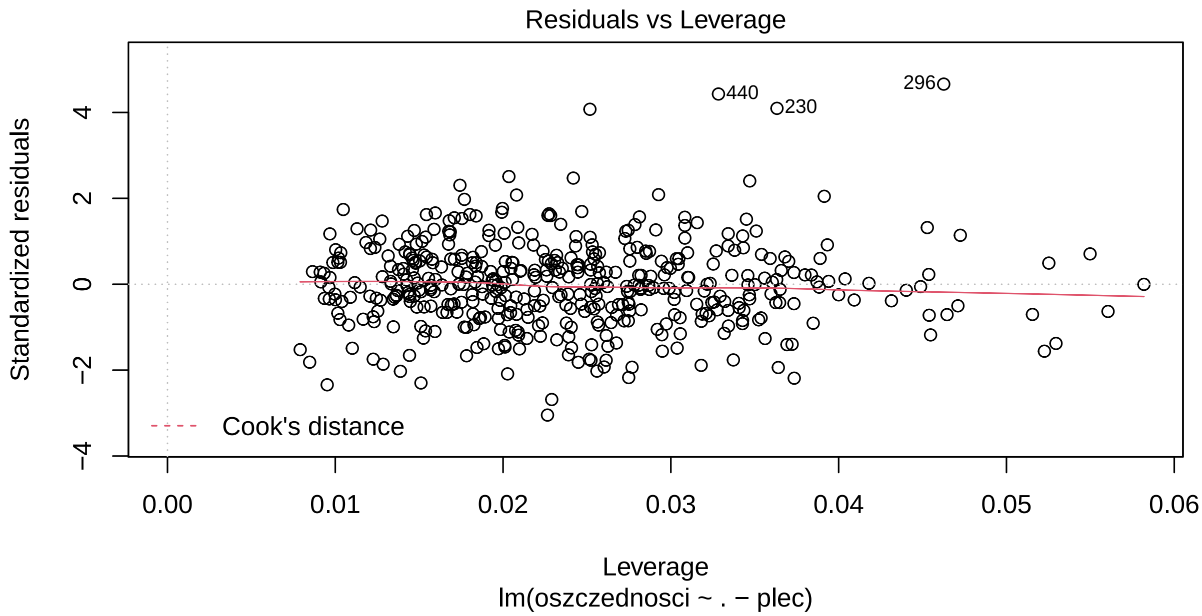
Na powyższym wykresie możemy przeanalizować rozkład residuów. Większoć z nich tworzy linie - choć residua po prawej stronie wykresu odbiegają od tego trendu. Mimo to mozemy przyjąć, że rozkład błędów jest zbliżony do normalnego.

```
plot(model, which=3)
```



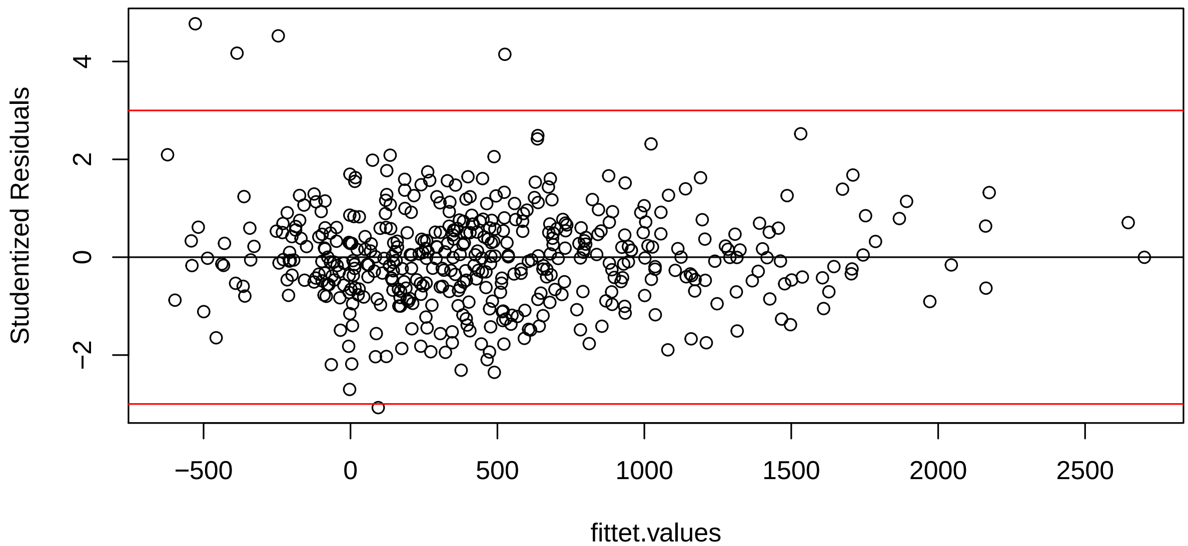
Na podstawie powyższego wykresu możemy ocenić homoskedastyczność modelu. Linia trendu jest zbliżona do linni równoległej do osi OX, co świadczy o homoskedastyczności modelu - czyli, że wariancja nie zależy od kolejnych zmiennych objaśnianych. Możemy zauważyć, że dla weiększych wartości mamy mniej obserwacji, co sprawia, że ocena homoskedastyczności dla tych zmiennych jest mniej pewna.

```
plot(model, which=5)
```



Ostatni wykres umożliwia znalezinie obserwacji ostających (z ang. outliers). Czerwona przerywana linia, która miała oddzielać obserwacje o wysokiej dźwigni nie jest widoczna na powyższym wykresie, zatem możemy uznać, że nie ma obserwacji o wysokiej dźwigni w danych.

```
#wykres reszt studentyzowanych
stud_resids <- studres(model)
plot(y_pred_model, stud_resids,  ylab='Studentized Residuals', xlab='fittet.values')
#add horizontal line at 0
abline(0, 0)
abline(3, 0, col = "red")
abline(-3, 0, col = "red")
```



Dla wykresów reszt studentyzowanych przyjmuje się, że reszty odchylone o więcej niż 3 od zera są resztami obserwacji odstających. Zgodnie z tą zasadą możemy uznać pięć obserwacji za odstające.