

Metody przetwarzania i analizy danych w R

Łukasz Wawrowski

Contents

Wprowadzenie	5
1 Wprowadzenie	7
1.1 Narzędzie	7
1.2 Cele analiz	7
2 Testowanie hipotez	9
2.1 Hipoteza statystyczna	9
2.2 Poziom istotności i wartość p	9
2.3 Testy parametryczne i nieparametryczne	9
3 Regresja	11
3.1 Regresja prosta	11
3.2 Regresja wieloraka	14
4 Grupowanie	31
4.1 Metoda k-średnich	31
4.2 Metoda hierarchiczna	38
5 Klasyfikacja	43
5.1 Drzewa klasyfikacyjne	43
5.2 KNN	43
6 Materiały z zajęć	45
6.1 28.10.2018	45
6.2 18.11.2018	45
6.3 16.12.2018	45
6.4 26.01.2019	45

Wprowadzenie

Literatura podstawowa:

- Przemysław Biecek - *Przewodnik po pakiecie R*
- Marek Gągolewski - *Programowanie w języku R. Analiza danych, obliczenia, symulacje.*
- Garret Golemund, Hadley Wickham - *R for Data Science* (polska wersja)

Literatura dodatkowa:

- inne pozycje po polsku
- inne pozycje po angielsku

Internet:

- R-bloggers
- rweekly

Chapter 1

Wprowadzenie

1.1 Narzędzie

- darmowe
- wszechstronne
- wsparcie społeczności
- wersja desktopowa i serwerowa

czyli **R** - środowisko do obliczeń statystycznych i wizualizacji wyników

- strona projektu: r-project.org
- świetne IDE: RStudio
- wersja przeglądarkowa: rstudio.cloud

R + Python

1.2 Cele analiz

Podstawowe:

- wnioskowanie statystyczne - porównywanie grup
- regresja - poszukiwanie związków
- klasyfikacja - przyporządkowanie do grup
- grupowanie - poszukiwanie grup
- prognozowanie - patrzenie w przyszłość

Inne:

- analiza języka naturalnego
- rozpoznawanie obrazów
- analiza koszykowa
- ...

1.2.1 Eksporacja danych

Pakiet `tidyverse`

```
library(tidyverse)
```

- analiza częstości dla zmiennych jakościowych
- analiza struktury dla zmiennych ilościowych

Case study: Wybory 2018

Chapter 2

Testowanie hipotez

2.1 Hipoteza statystyczna

Przypuszczenie dotyczące własności analizowanej cechy, np. średnia w populacji jest równa 10, rozkład cechy jest normalny.

Formuluje się zawsze dwie hipotezy: hipotezę zerową (H_0) i hipotezę alternatywną (H_1). Hipoteza zerowa jest hipotezą mówiącą o równości:

$$H_0 : \bar{x} = 10$$

Z kolei hipoteza alternatywna zakłada coś przeciwnego:

$$H_1 : \bar{x} \neq 10$$

Zamiast znaku nierówności (\neq) może się także pojawić znak mniejszości ($<$) lub większości ($>$).

2.2 Poziom istotności i wartość p

Hipotezy statystyczne weryfikuje się przy określonym poziomie istotności α , który wskazuje maksymalny poziom akceptowalnego błędu (najczęściej $\alpha = 0,05$).

Większość programów statystycznych podaje w wynikach testu wartość p. Jest to prawdopodobieństwo uzyskania obserwowanych wyników przy założeniu prawdziwości hipotezy zerowej.

Generalnie jeśli $p < \alpha$ - odrzucamy hipotezę zerową.

Krytyka wartości p

2.3 Testy parametryczne i nieparametryczne

Testy statystyczne dzielą się na dwie grupy:

- parametryczne, które wymagają spełnienia założeń, ale są dokładniejsze,
- nieparametryczne, które nie wymagają tylu założeń, ale są mniej dokładne.

Chapter 3

Regresja

Metody regresji pozwalają na analizowanie zależności przyczynowo-skutkowych.

Na początku pracy wczytujemy biblioteki `tidyverse`, `readxl` oraz wyłączamy notację naukową.

```
library(tidyverse)
library(readxl)

options(scipen = 100)
```

3.1 Regresja prosta

Na podstawie danych dotyczących informacji o doświadczeniu i wynagrodzeniu pracowników zbuduj model określający ‘widełki’ dla potencjalnych pracowników o doświadczeniu równym 8, 10 i 11 lat.

Wczytujemy dane i sprawdzamy czy nie występują zera bądź braki danych z użyciem funkcji `summary()`.

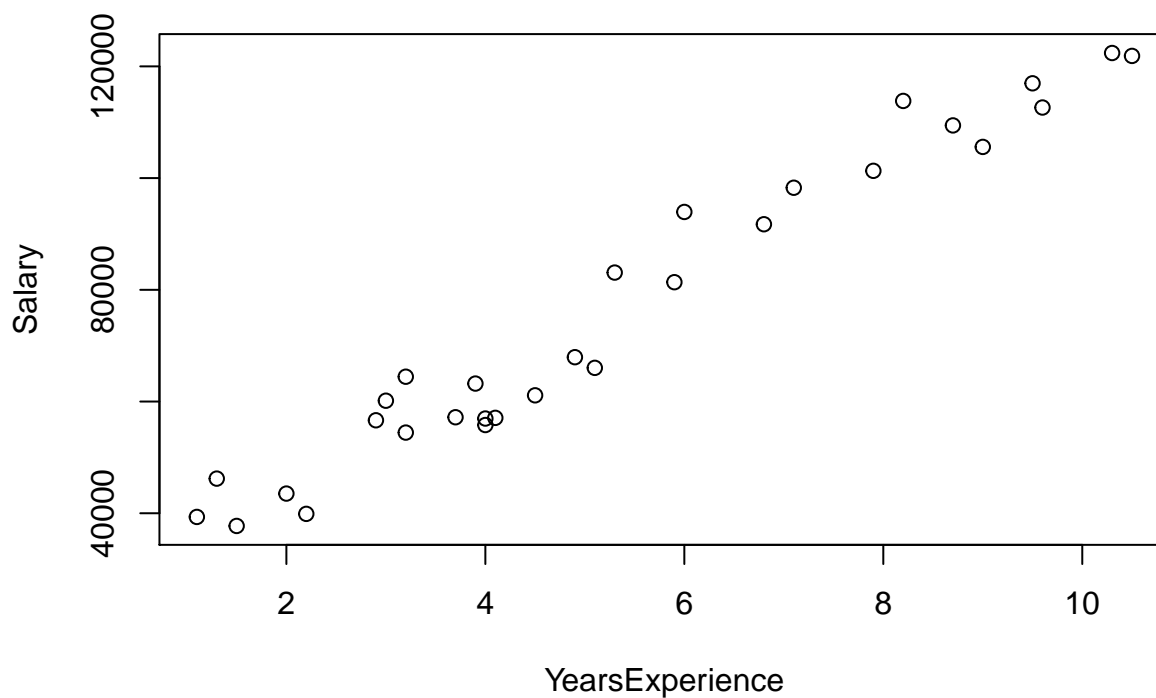
```
dane <- read.csv("data/Salary_Data.csv")

summary(dane)
```

```
##  YearsExperience      Salary
##  Min.   : 1.100   Min.    : 37731
##  1st Qu.: 3.200   1st Qu.: 56721
##  Median : 4.700   Median : 65237
##  Mean   : 5.313   Mean    : 76003
##  3rd Qu.: 7.700   3rd Qu.:100545
##  Max.   :10.500   Max.    :122391
```

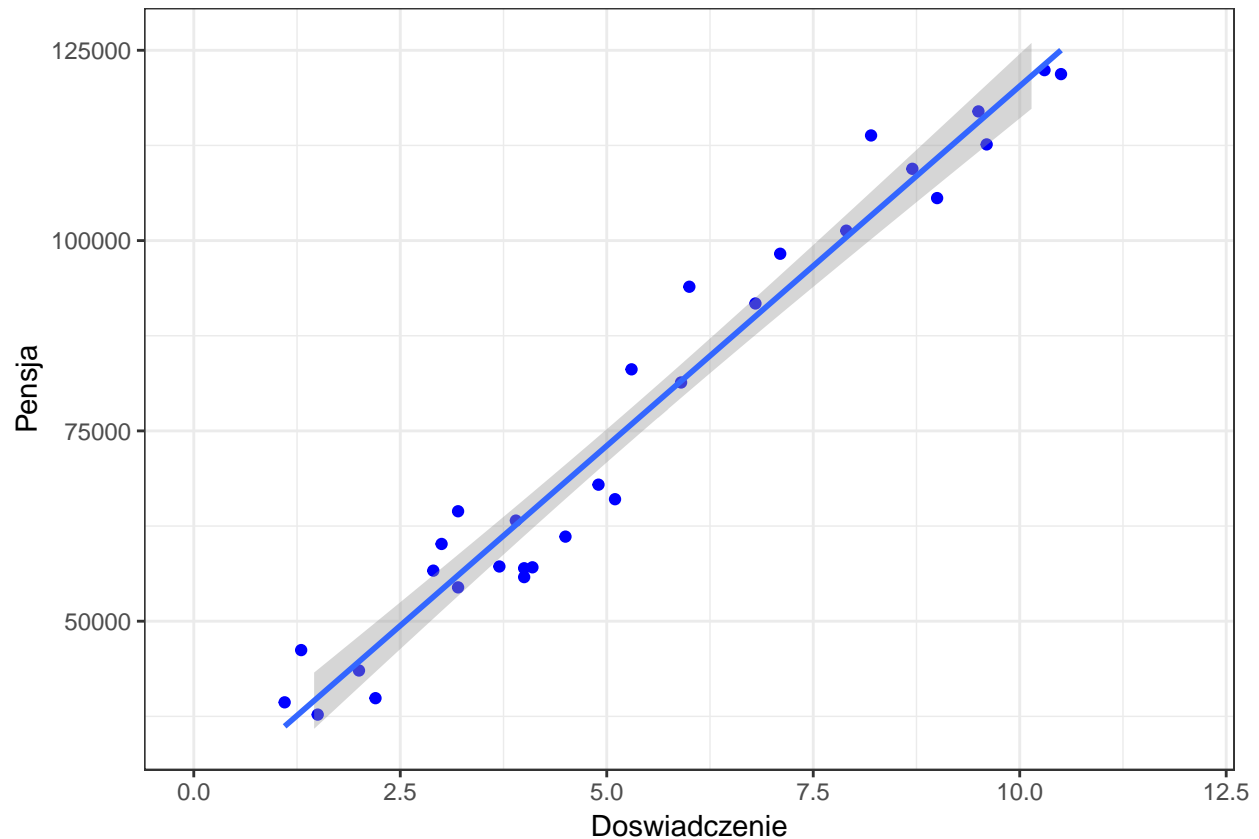
Następnie stworzymy wykres.

```
plot(dane)
```



Najprostszym sposobem wizualizacji jest wykorzystanie funkcji `plot()`, niemniej taki wykres nie jest najpiękniejszy i trudno się go formatuje. Dużo lepiej skorzystać z pakietu `ggplot2`.

```
ggplot(dane, aes(x=YearsExperience, y=Salary)) +  
  geom_point(colour = "blue") +  
  geom_smooth(method = "lm") +  
  xlab("Doświadczenie") +  
  ylab("Pensja") +  
  xlim(0,12) +  
  ylim(35000, 126000) +  
  theme_bw()
```



W argumentach funkcji `ggplot()` podajemy co wizualizujemy, natomiast sposób prezentacji ustalany jest przez funkcje `geom`. Funkcje `xlab()` i `ylab()` określają etykiety osi, a `xlim()` i `ylim()` wartości graniczne. Funkcje rozpoczynające się od `theme_` określają wygląd wykresu.

Modelowanie rozpoczynamy od określenia zmiennej zależnej i niezależnej.

- zmienna zależna/objaśniana: pensja
- zmienna niezależna/objaśniająca: doświadczenie

```
model <- lm(Salary ~ YearsExperience, data = dane)
summary(model)
```

```
##
## Call:
## lm(formula = Salary ~ YearsExperience, data = dane)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7958.0 -4088.5  -459.9   3372.6 11448.0
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   25792.2    2273.1    11.35 0.00000000000551 ***
## YearsExperience  9450.0     378.8    24.95 < 0.000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5788 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.957, Adjusted R-squared:  0.9554
## F-statistic: 622.5 on 1 and 28 DF,  p-value: < 0.00000000000000022
```

- $b_1 = 9450$ - wzrost doświadczenia o rok powoduje przeciętny wzrost pensji o 9450 zł
- $b_0 = 25792,2$ - pracownik o doświadczeniu 0 lat uzyska pensję w wysokości 25792,2 zł

Resztowy błąd standardowy S_u wynosi 5788, co oznacza wartości z modelu różnią się od rzeczywistości średnio o ± 5788 zł

$R^2 = 0,957$ - model wyjaśnia 96% zmienności pensji

Dokonujemy wyznaczenia wartości teoretycznych dla kilku wybranych lat.

```
nowiPracownicy <- data.frame(YearsExperience=c(8,10,11))

predict(model, nowiPracownicy)
```

```
##          1          2          3
## 101391.9 120291.8 129741.8
```

Tym sposobem uzyskujemy następujące widełki:

- pracownik o stażu 8 lat - proponowana pensja 101391,9 zł \pm 5788 zł
- pracownik o stażu 10 lat - proponowana pensja 120291,8 zł \pm 5788 zł
- pracownik o stażu 11 lat - proponowana pensja 129741,8 zł \pm 5788 zł

3.1.1 Zadanie

Dla danych dotyczących sklepu nr 77 opracuj model zależności sprzedaży od liczby klientów. Ile wynosi teoretyczna sprzedaż w dniach, w których liczba klientów będzie wynosiła 560, 740, 811 oraz 999 osób?

3.2 Regresja wieloraka

Na podstawie danych dotyczących zatrudnienia opracuj model, w którym zmienną zależną jest bieżące wynagrodzenie. Jaka cecha ma największy wpływ na tę wartość?

Opis zbioru:

- id - kod pracownika
- plec - płeć pracownika (0 - mężczyzna, 1 - kobieta)
- data_urodz - data urodzenia
- edukacja - wykształcenie (w latach nauki)
- kat_pracownika - grupa pracownicza (1 - ochroniarz, 2 - urzędnik, 3 - menedżer)
- bwynagrodzenie - bieżące wynagrodzenie
- pwynagrodzenie - początkowe wynagrodzenie
- staz - staż pracy (w miesiącach)
- doswiadczenie - poprzednie zatrudnienie (w miesiącach)
- związki - przynależność do związków zawodowych (0 - nie, 1 - tak)
- wiek - wiek (w latach)

Wczytanie danych.

```
pracownicy <- read_xlsx("data/pracownicy.xlsx")

pracownicy2 <- pracownicy %>%
  filter(!is.na(wiek)) %>%
  select(-id, -data_urodz) %>%
```

```
mutate(plec=as.factor(plec),
       kat_pracownika=as.factor(kat_pracownika),
       zwiazki=as.factor(zwiazki))

summary(pracownicy2)
```

##	plec	edukacja	kat_pracownika	bwynagrodzenie	pwynagrodzenie
##	0:257	Min. : 8.00	1:362	Min. : 15750	Min. : 9000
##	1:216	1st Qu.:12.00	2: 27	1st Qu.: 24000	1st Qu.:12450
##		Median :12.00	3: 84	Median : 28800	Median :15000
##		Mean :13.49		Mean : 34418	Mean :17009
##		3rd Qu.:15.00		3rd Qu.: 37050	3rd Qu.:17490
##		Max. :21.00		Max. :135000	Max. :79980
##	staz	doswiadczenie	zwiazki	wiek	
##	Min. :63.00	Min. : 0.00	0:369	Min. :24.00	
##	1st Qu.:72.00	1st Qu.: 19.00	1:104	1st Qu.:30.00	
##	Median :81.00	Median : 55.00		Median :33.00	
##	Mean :81.14	Mean : 95.95		Mean :38.67	
##	3rd Qu.:90.00	3rd Qu.:139.00		3rd Qu.:47.00	
##	Max. :98.00	Max. :476.00		Max. :66.00	

W zmiennej wiek występował brak danych, który został usunięty. Usunięto także kolumny, które nie przydadzą się w modelowaniu. Ponadto dokonujemy przekształcenia typu cech, które są jakościowe (pleć, kat_pracownika, zwiazki) z typu liczbowego na czynnik/faktor, który będzie poprawnie interpretowany przez model.

W modelu zmienna zależna to `bwynagrodzenie`, natomiast jako zmienne niezależne bierzemy pod uwagę wszystkie pozostałe cechy.

```
model <- lm(bwynagrodzenie ~ ., pracownicy2)
summary(model)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ ., data = pracownicy2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23185  -3041   -705    2591   46295
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -4764.87418   3590.49652  -1.327    0.18514
## plec1        -1702.43743    796.51779  -2.137    0.03309 *
## edukacja      482.43603    160.83977   2.999    0.00285 **
## kat_pracownika2 6643.17910   1638.06138   4.056 0.00005869172850407 ***
## kat_pracownika3 11169.64519   1372.73990   8.137 0.000000000000000377 ***
## pwynagrodzenie   1.34021     0.07317  18.315 < 0.0000000000000002 ***
## staz          154.50876     31.65933   4.880 0.00000145958620443 ***
## doswiadczenie  -15.77375     5.78369  -2.727    0.00663 **
## zwiazki1       -1011.55276    787.80884  -1.284    0.19978
## wiek          -64.78787     47.88015  -1.353    0.17668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6809 on 463 degrees of freedom
## Multiple R-squared:  0.8444, Adjusted R-squared:  0.8414
## F-statistic: 279.1 on 9 and 463 DF,  p-value: < 0.00000000000000022
```

Tak zbudowany model wyjaśnia 84% zmienności bieżącego wynagrodzenia, ale nie wszystkie zmienne są w tym modelu istotne.

Parametry regresji mają następujące interpretacje:

- plec1 - kobiety zarabiają przeciętnie o 1702,44 zł mniej niż mężczyźni,
- edukacja - wzrost liczby lat nauki o rok powoduje średni wzrost bieżącego wynagrodzenia o 482,44 zł
- kat_pracownika2 - pracownicy o kodzie 2 (urzędnik) zarabiają średnio o 6643,18 zł więcej niż pracownik o kodzie 1 (ochroniarz)
- kat_pracownika2 - pracownicy o kodzie 3 (menedżer) zarabiają średnio o 11169,65 zł więcej niż pracownik o kodzie 1 (ochroniarz)
- pwynagrodzenie - wzrost początkowego wynagrodzenia o 1 zł powoduje przeciętny wzrost bieżącego wynagrodzenia o 1,34 zł
- staz - wzrost stażu pracy o miesiąc skutkuje przeciętnym wzrostem bieżącego wynagrodzenia o 154,51 zł
- doswiadczenie - wzrost doświadczenia o miesiąc powoduje średni spadek bieżącego wynagrodzenia o 15,77 zł
- zwiazkil - pracownicy należący do związków zawodowych zarabiają średnio o 1011,55 zł mniej aniżeli pracownicy, którzy do związków nie należą
- wiek - wzrost wieku pracownika o 1 rok to przeciętnym spadek bieżącego wynagrodzenia o 64,79 zł

Wszystkie te zależności obowiązują przy założeniu *ceteris paribus* - przy pozostałych warunkach niezmiennych.

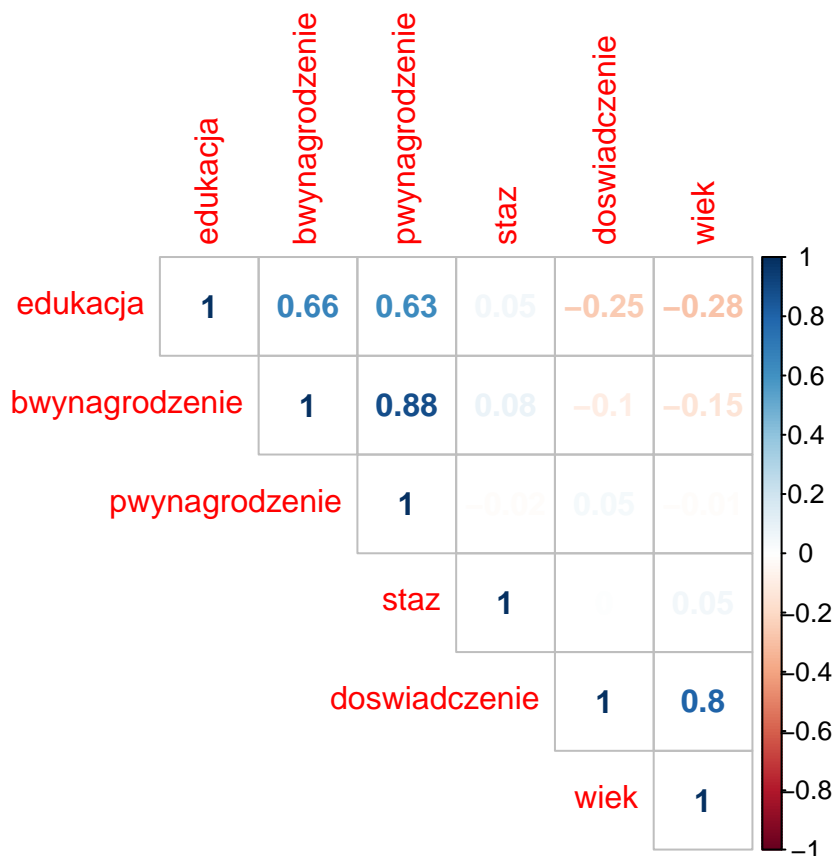
Ten model wymaga oczywiście ulepszenie do czego wykorzystamy m.in. pakiet `olsrr`.

Pierwszą kwestią, którą się zajmiemy jest współliniowość zmiennych. W regresji zmienne objaśniające powinny być jak najbardziej skorelowane ze zmienną objaśnianą, a możliwie nieskorelowane ze sobą. W związku z tym wybieramy ze zbioru wyłącznie cechy ilościowe, dla którym wyznaczymy współczynnik korelacji liniowej Pearsona.

```
library(corrplot)
library(olsrr)

korelacje <- pracownicy2 %>%
  select(-c(plec, kat_pracownika, zwiazki)) %>%
  cor()

corrplot(korelacje, method = "number", type = "upper")
```

Możemy zauważyć, że wartości bieżącego wynagrodzenia są najsilniej skorelowane z wartościami wynagrodzenia początkowego. Także doświadczenie i wiek są silnie ze sobą związane, co może sugerować, że obie zmienne wnoszą do modelu podobną informację.

W związku z tym powinniśmy wyeliminować niektóre zmienne z modelu pozostawiając te najważniejsze. Wyróżnia się trzy podejścia do tego zagadnienia:

- ekspercki dobór cech,
- budowa wszystkich możliwych modeli i wybór najlepszego według określonego kryterium,
- regresja krokowa.

W przypadku budowy wszystkich możliwych modeli należy pamiętać o rosnącej wykładniczo liczbie modeli - $2^p - 1$, gdzie p oznacza liczbę zmiennych objaśniających. w rozważanym przypadku liczba modeli wynosi 255.

```
wszystkie_modely <- ols_step_all_possible(model)
```

W uzyskanym zbiorze danych są informacje o numerze modelu, liczbie użytych zmiennych, nazwie tych zmiennych oraz wiele miar jakości. Te, które warto wziąć pod uwagę to przede wszystkim:

- **rsquare** - współczynnik R-kwadrat,
- **aic** - kryterium informacyjne Akaike,
- **msep** - błąd średniokwadratowy predykcji.

Najwyższa wartość współczynnika R^2 związana jest z modelem zawierającym wszystkie dostępne zmienne objaśniające. Jest to pewna niedoskonałość tej miary, która rośnie wraz z liczbą zmiennych w modelu, nawet jeśli te zmienne nie są istotne.

W przypadku kryteriów informacyjnych oraz błędu średniokwadratowego interesują nas jak najmniejsze wartości. Wówczas jako najlepszy należy wskazać model nr 219 zawierający 6 zmiennych objaśniających.

Metodą, która także pozwoli uzyskać optymalny model, ale przy mniejszym obciążeniu obliczeniowym jest regresja krokowa polegająca na krokowym budowaniu modelu.

```
ols_step_both_aic(model)
```

```
## Stepwise Selection Method
## -----
##
## Candidate Terms:
##
## 1 . plec
## 2 . edukacja
## 3 . kat_pracownika
## 4 . pwynagrodzenie
## 5 . staz
## 6 . doswiadczenie
## 7 . zwiazki
## 8 . wiek
##
##
## Variables Entered/Removed:
##
## - pwynagrodzenie added
## - kat_pracownika added
## - doswiadczenie added
## - staz added
## - edukacja added
## - plec added
##
## No more variables to be added or removed.
##
##
##                                     Stepwise Summary
## -----
```

## Variable	## Method	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
## pwynagrodzenie	addition	9862.260	31053506813.535	106862706669.340	0.77484	0.7743
## kat_pracownika	addition	9786.152	26215474648.689	111700738834.186	0.80992	0.8087
## doswiadczenie	addition	9743.487	23853248017.651	114062965465.224	0.82705	0.8255
## staz	addition	9719.469	22576592070.620	115339621412.255	0.83630	0.8345
## edukacja	addition	9707.338	21912088629.084	116004124853.791	0.84112	0.8390
## plec	addition	9703.188	21629051655.016	116287161827.859	0.84317	0.8408

```
## -----
```

Otrzymany w ten sposób model jest tożsamy z modelem charakteryzującym się najlepszymi miarami jakości spośród zbioru wszystkich możliwych modeli:

```
wybrany_model <- lm(bwynagrodzenie ~ pwynagrodzenie + kat_pracownika + doswiadczenie + staz + plec + edukacja, data = pracownicy2)
summary(wybrany_model)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
##     doswiadczenie + staz + plec + edukacja, data = pracownicy2)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22922  -3300   -673    2537   46524
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   -6547.147   3402.860   -1.924    0.05496 .
## pwynagrodzenie     1.342     0.073   18.382 < 0.0000000000000002 ***
## kat_pracownika2   6734.992   1631.122    4.129  0.0000431843301918 ***
## kat_pracownika3  11226.635   1368.413    8.204  0.0000000000000023 ***
## doswiadczenie    -22.302     3.571   -6.246  0.0000000009514655 ***
## staz             147.865     31.461    4.700  0.0000034337703087 ***
## plec1            -1878.949    761.703   -2.467    0.01399 *
## edukacja         501.391     160.270    3.128    0.00187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6820 on 465 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8408
## F-statistic: 357.1 on 7 and 465 DF,  p-value: < 0.00000000000000022
```

Uzyskany model charakteryzuje się mniejszym błędem standardowym od modelu ze wszystkimi zmiennymi i tylko jedną nieistotną zmienną. Wyraz wolny (Intercept) nie musi być istotny w modelu.

Wróćmy jeszcze na chwilę do tematu współliniowości zmiennych objaśniających:

```
ols_vif_tol(wybrany_model)
```

```
## # A tibble: 7 x 3
##   Variables      Tolerance    VIF
##   <chr>         <dbl> <dbl>
## 1 pwynagrodzenie 0.298 3.36
## 2 kat_pracownika2 0.687 1.46
## 3 kat_pracownika3 0.360 2.78
## 4 doswiadczenie 0.705 1.42
## 5 staz          0.986 1.01
## 6 plec1         0.683 1.46
## 7 edukacja      0.461 2.17
```

Współczynnik tolerancji wskazuje na procent niewyjaśnionej zmienności danej zmiennej przez pozostałe zmienne objaśniające. Przykładowo współczynnik tolerancji dla początkowego wynagrodzenia wynosi 0,2980, co oznacza, że 30% zmienności początkowego wynagrodzenia nie jest wyjaśnione przez pozostałe zmienne w modelu. Z kolei współczynnik VIF jest obliczany na podstawie wartości współczynnika tolerancji i wskazuje o ile wariancja szacowanego współczynnika regresji jest podwyższona z powodu współliniowości danej zmiennej objaśniającej z pozostałymi zmiennymi objaśniającymi. Wartość współczynnika VIF powyżej 4 należy uznać za wskazującą na współliniowość. W analizowanym przypadku takie zmienne nie występują.

Ocena siły wpływu poszczególnych zmiennych objaśniających na zmienną objaśnianą w oryginalnej postaci modelu nie jest możliwa. Należy wyznaczyć standaryzowane współczynniki beta, które wyliczane są na danych standaryzowanych, czyli takich, które są pozbawione jednostek i cechują się średnią równą 0, a odchyleniem standardowym równym 1. Standaryzacja ma sens tylko dla cech numerycznych, w związku z czym korzystamy z funkcji `mutate_if()`, która jako pierwszy argument przyjmuje warunek, który ma być spełniony, aby była zastosowana przekształcenie podawane jako drugi argument.

```
pracownicy2_std <- pracownicy2 %>%
  mutate_if(is.numeric, scale)
```

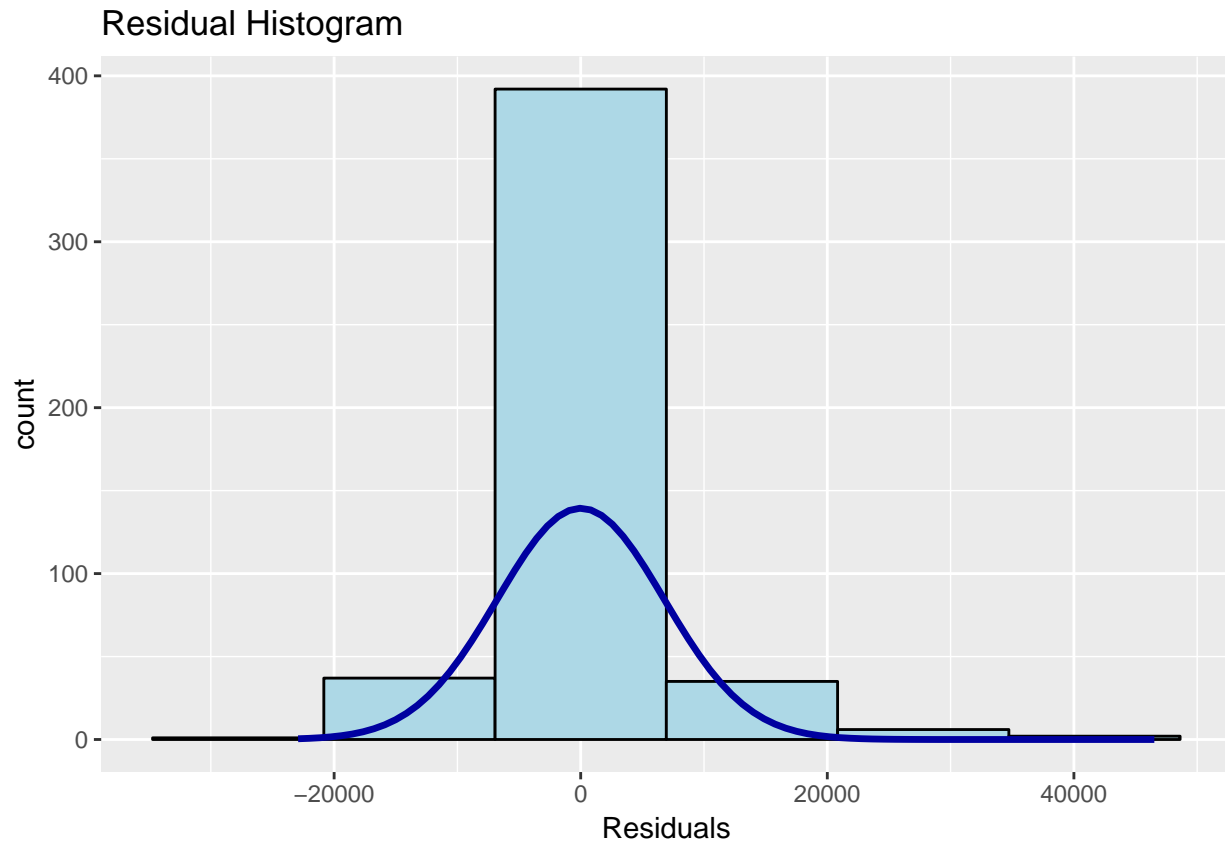
```
wybrany_model_std <- lm(bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
                        doswiadczenie + staz + plec + edukacja, data = pracownicy2_std)
summary(wybrany_model_std)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
##     doswiadczenie + staz + plec + edukacja, data = pracownicy2_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34098 -0.19306 -0.03939  0.14841  2.72171
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -0.08893    0.03144  -2.828    0.00488 **
## pwynagrodzenie  0.61842    0.03364  18.382 < 0.0000000000000002 ***
## kat_pracownika2 0.39400    0.09542   4.129  0.0000431843301918 ***
## kat_pracownika3 0.65677    0.08005   8.204  0.0000000000000023 ***
## doswiadczenie  -0.13657    0.02187  -6.246  0.00000000009514655 ***
## staz           0.08691    0.01849   4.700  0.0000034337703087 ***
## plec1          -0.10992    0.04456  -2.467    0.01399 *
## edukacja       0.08464    0.02706   3.128    0.00187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.399 on 465 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8408
## F-statistic: 357.1 on 7 and 465 DF,  p-value: < 0.00000000000000022
```

Spośród cech ilościowych największy wpływ na zmienną objaśnianą mają wartości wynagrodzenia początkowego, staż, edukacja i na końcu doświadczenie.

Reszty czyli różnice pomiędzy obserwowanymi wartościami zmiennej objaśnianej, a wartościami wynikającymi z modelu w klasycznej metodzie najmniejszych kwadratów powinny być zbliżone do rozkładu normalnego. Oznacza to, że najwięcej reszt powinno skupiać się wokół zerowych różnic, natomiast jak najmniej powinno być wartości modelowych znacznie różniących się od tych rzeczywistych.

```
ols_plot_resid_hist(wybrany_model)
```



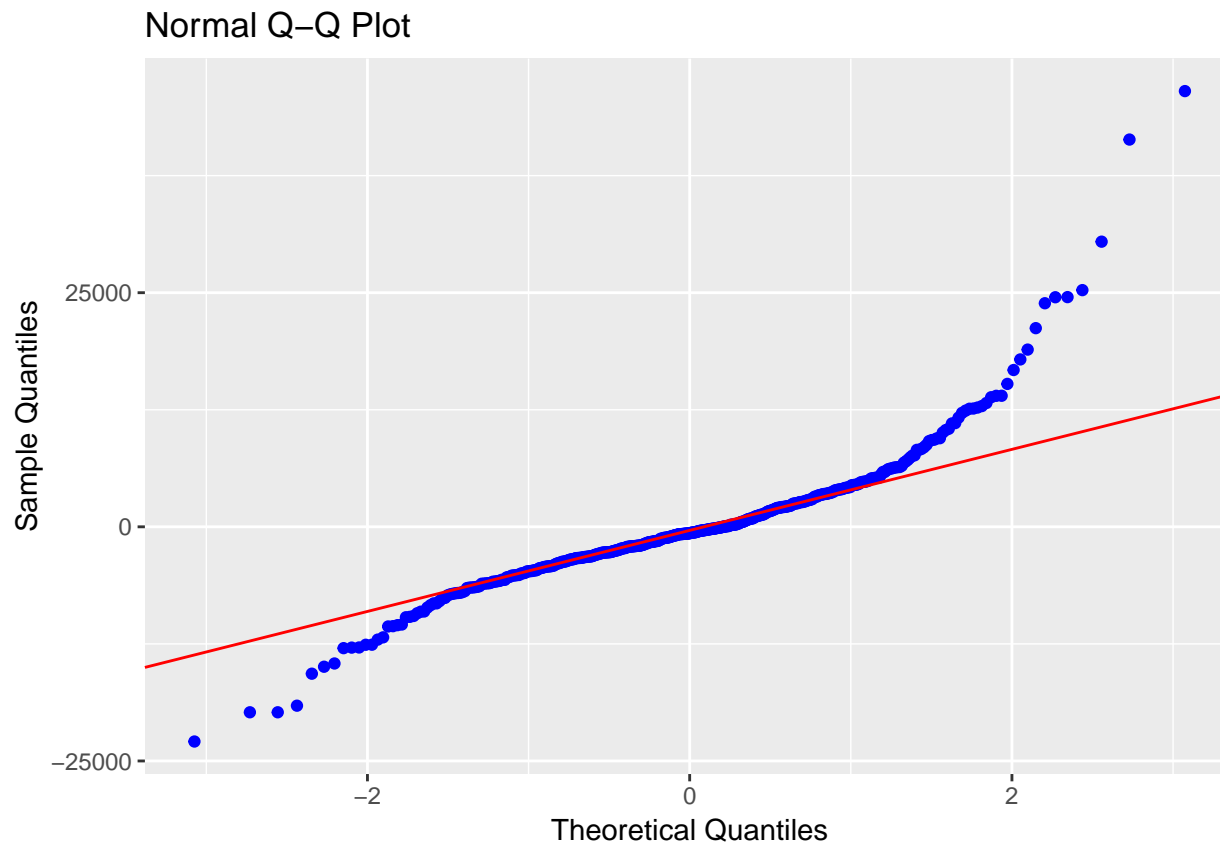
Reszty w naszym modelu wydają się być zbliżone do rozkładu normalnego. Jednoznaczną odpowiedź da jednak odpowiedni test.

```
ols_test_normality(wybrany_model)
```

```
## -----
##      Test           Statistic      pvalue
## -----
## Shapiro-Wilk         0.868         0.0000
## Kolmogorov-Smirnov    0.1092         0.0000
## Cramer-von Mises      42.5504         0.0001
## Anderson-Darling      13.0233         0.0000
## -----
```

Hipoteza zerowa w tych testach mówi o zgodności rozkładu reszt z rozkładem normalnym. Na podstawie wartości p, które są mniejsze od $\alpha = 0,05$ stwierdzamy, że są podstawy do odrzucenia tej hipotezy czyli reszty z naszego modelu nie mają rozkładu normalnego. W diagnostyce przyczyn takiego stanu rzeczy pomoże nam wykres kwantyl-kwantyl:

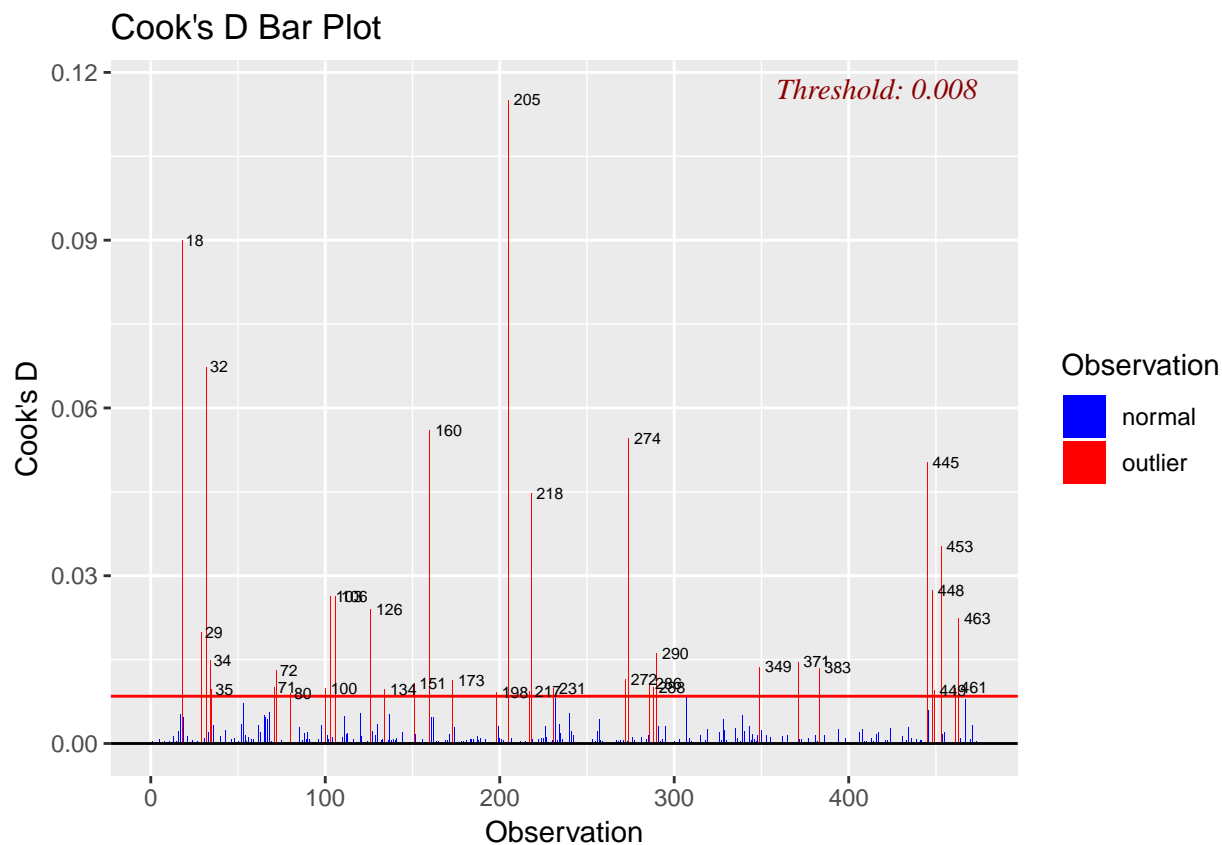
```
ols_plot_resid_qq(wybrany_model)
```



Gdyby wszystkie punkty leżały na prostej to oznaczałoby to normalność rozkładu reszt. Tymczasem po lewej i prawej stronie tego wykresu znajdują się potencjalne wartości odstające, które znacznie wpływają na rozkład reszt modelu.

Wartości odstające można ustalić na podstawie wielu kryteriów. Do jednych z najbardziej popularnych należy odległość Cooka:

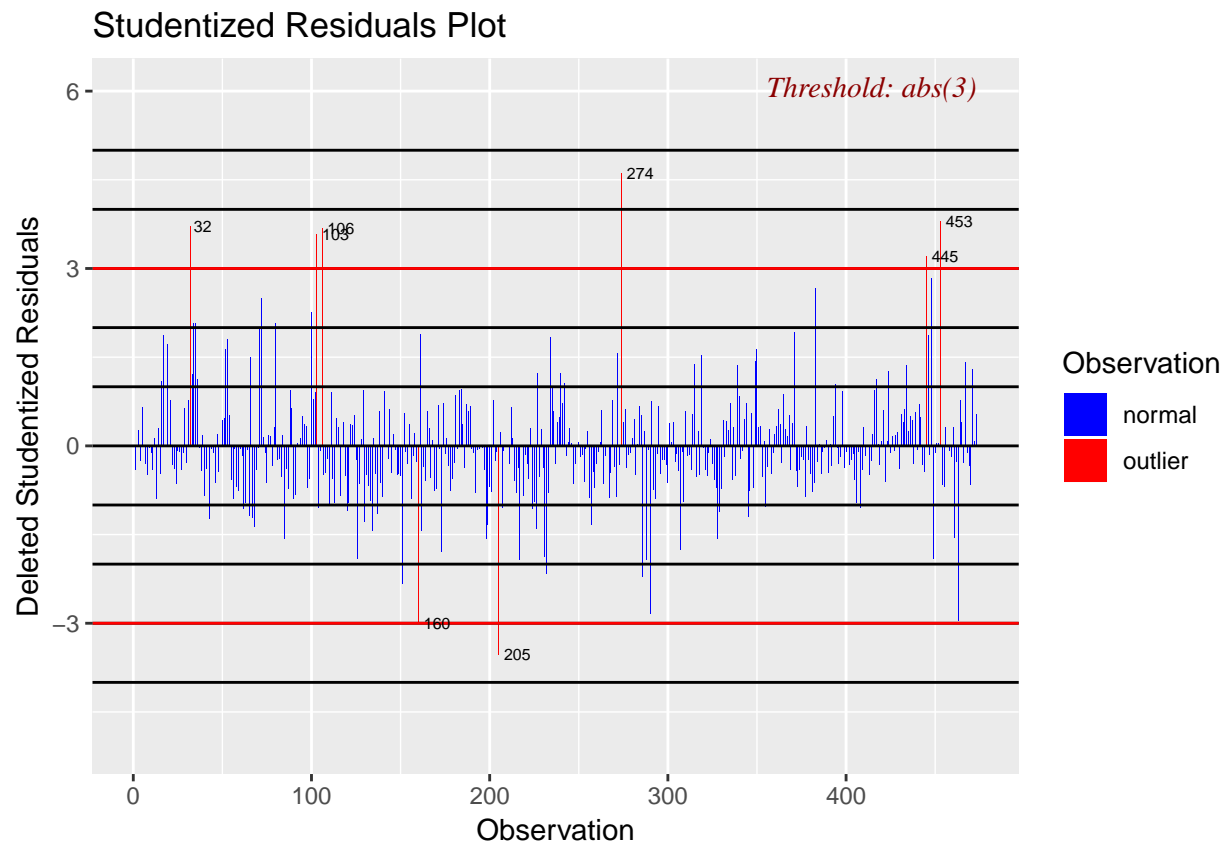
```
cook <- ols_plot_cooksd_bar(wybrany_model)
```



Przypisanie tej funkcji do obiektu zwraca nam tabelę z numerami zidentyfikowanych obserwacji wpływowych. W przypadku odległości Cooka jest to 35 obserwacji.

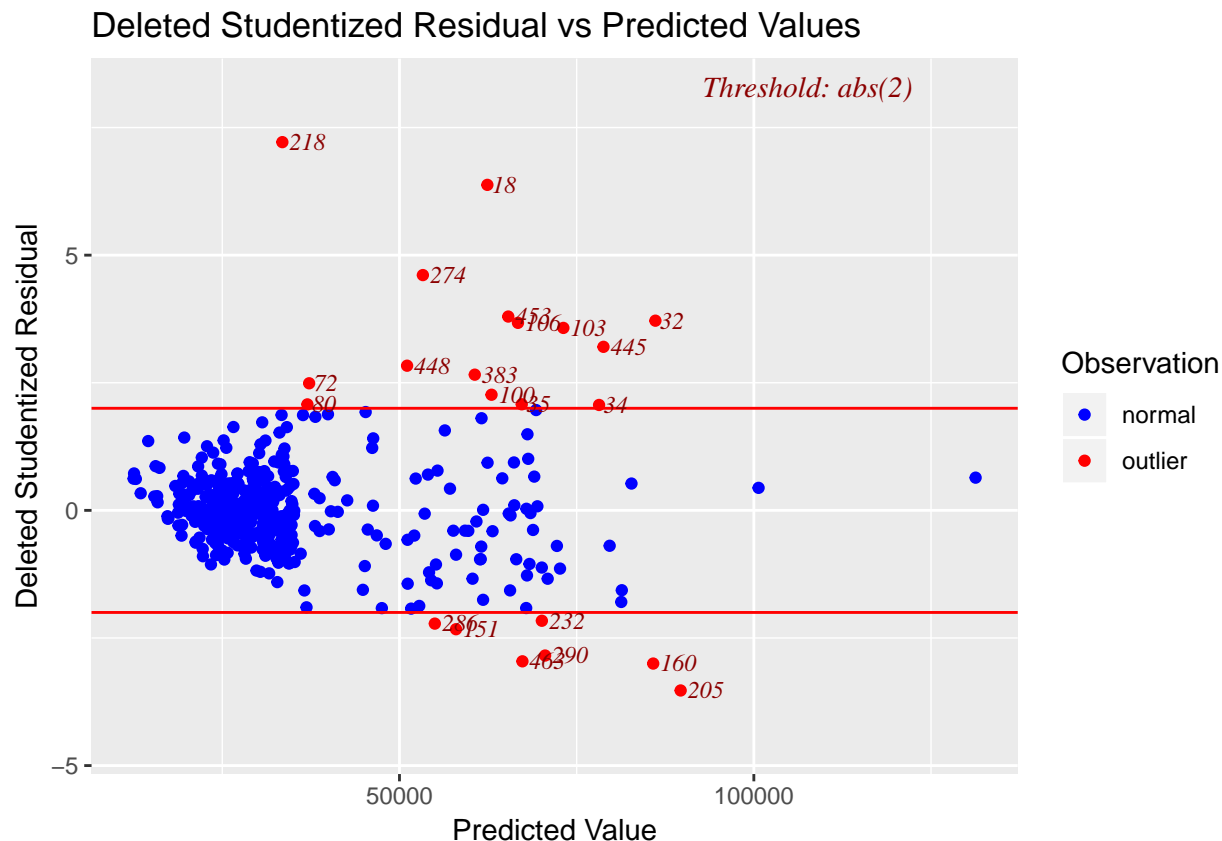
Inną miarą są reszty studentyzowane.

```
stud3 <- ols_plot_resid_stud(wybrany_model)
```



Wyżej wykorzystana funkcja jako kryterium odstawania przyjmuje wartość 3 identyfikując 10 obserwacji wpływowych. Z kolei dodanie do powyższej funkcji przyrostka *fit* powoduje przyjęcie jako granicy wartości równej 2.

```
obs_wplyw <- ols_plot_resid_stud_fit(wybrany_model)
```

W ten sposób zostało zidentyfikowanych 22 obserwacji odstających. Korzystając z odległości Cooka wyeliminujemy obserwacje odstające ze zbioru:

```
nr_obs_wplyw <- cook$outliers$observation

bez_obs_wplyw <- pracownicy2[-nr_obs_wplyw,]

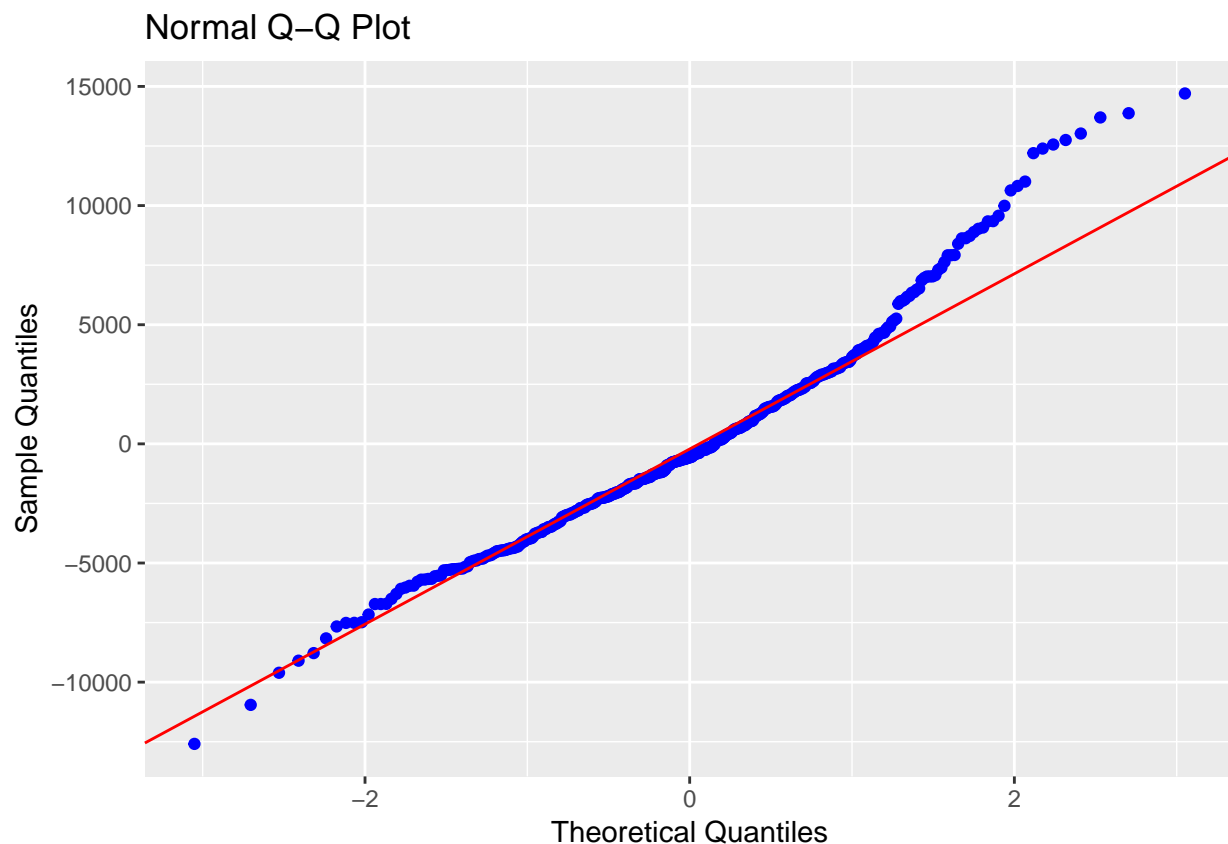
wybrany_model_out <- lm(bwynagrodzenie ~ pwynagrodzenie + kat_pracownika + doswiadczenie + staz + plec,
                        data = bez_obs_wplyw)
summary(wybrany_model_out)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
##     doswiadczenie + staz + plec + edukacja, data = bez_obs_wplyw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12591.2  -2696.4   -564.8   2263.0  14704.1
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   -3756.8288   2381.5361  -1.577    0.115420
## pwynagrodzenie    1.3696     0.0676  20.260 < 0.0000000000000002 ***
## kat_pracownika2  6480.6971   1049.9539   6.172  0.0000000015569534 ***
## kat_pracownika3  9791.5000   1059.4980   9.242 < 0.0000000000000002 ***
## doswiadczenie   -18.5808     2.3229  -7.999  0.0000000000000117 ***
```

```
## staz          116.3809    20.9704    5.550    0.0000000501086186 ***
## plec1        -1616.6030   505.1391   -3.200          0.001475 **
## edukacja     397.7652    105.8146    3.759          0.000194 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4271 on 430 degrees of freedom
## Multiple R-squared:  0.894, Adjusted R-squared:  0.8923
## F-statistic: 518.3 on 7 and 430 DF,  p-value: < 0.00000000000000022
```

Model dopasowany na takim zbiorze charakteryzuje się dużo mniejszym błędem standardowym oraz wyższym współczynnikiem R^2 . Sprawdźmy w takim razie normalność reszt.

```
ols_plot_resid_qq(wybrany_model_out)
```



Wykres kwantyl-kwantyl wygląda już dużo lepiej, ale dla pewności przeprowadzimy testy statystyczne.

```
ols_test_normality(wybrany_model_out)
```

```
## -----
##      Test           Statistic      pvalue
## -----
## Shapiro-Wilk        0.9696        0.0000
## Kolmogorov-Smirnov   0.0667        0.0405
## Cramer-von Mises     38.1644        0.0001
## Anderson-Darling     3.5421        0.0000
## -----
```

Tylko jeden test wskazał zgodność rozkładu reszt z rozkładem normalnym.

3.2.1 Zadanie

Na podstawie zbioru dotyczącego 50 startupów określ jakie czynniki wpływają na przychód startupów. Na zajęciach 27.04.2019 r. przedyskutujemy wyniki.

```
startupy <- read.csv("data/50_Startups.csv")
```

```
summary(startupy)
```

```
##      R.D.Spend      Administration      Marketing.Spend      State
## Min.       :    0      Min.       : 51283      Min.       :    0      California:17
## 1st Qu.: 39936      1st Qu.:103731      1st Qu.:129300      Florida   :16
## Median : 73051      Median :122700      Median :212716      New York  :17
## Mean   : 73722      Mean   :121345      Mean   :211025
## 3rd Qu.:101603      3rd Qu.:144842      3rd Qu.:299469
## Max.    :165349      Max.    :182646      Max.    :471784
##      Profit
## Min.       : 14681
## 1st Qu.: 90139
## Median :107978
## Mean   :112013
## 3rd Qu.:139766
## Max.    :192262
```

Podsumowanie danych pokazuje występowanie zer w wartościach zmiennych R.D.Spend i Marketing.Spend - należy je wyeliminować.

```
startupy <- startupy %>%
  filter(R.D.Spend != 0, Marketing.Spend != 0)
```

Na tak przygotowanym zbiorze możemy dopasować model ze wszystkimi zmiennymi.

```
model <- lm(Profit ~ ., data=startupy)
summary(model)
```

```
##
## Call:
## lm(formula = Profit ~ ., data = startupy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15464   -4887   -1424    5550   13560
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  59773.09290   6126.19072    9.757 0.0000000000000391 ***
## R.D.Spend      0.77507     0.03909   19.828 < 0.0000000000000002 ***
## Administration -0.05759     0.04515   -1.275    0.210
## Marketing.Spend  0.02102     0.01528    1.376    0.176
## StateFlorida -1752.43437   2744.96836   -0.638    0.527
## StateNew York -1597.22937   2764.85338   -0.578    0.567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7503 on 40 degrees of freedom
## Multiple R-squared:  0.9613, Adjusted R-squared:  0.9565
## F-statistic: 198.7 on 5 and 40 DF,  p-value: < 0.00000000000000022
```

Jedyną istotną zmienną objaśniającą jest cecha R.D.Spend. W związku z tym używamy funkcji krokowej metody doboru zmiennych w celu znalezienia istotnych czynników wpływających na przychód startupu.

```
ols_step_both_p(model)
```

```
## Stepwise Selection Method
## -----
##
## Candidate Terms:
##
## 1. R.D.Spend
## 2. Administration
## 3. Marketing.Spend
## 4. State
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - R.D.Spend added
## - Administration added
##
## No more variables to be added/removed.
##
## Final Model Output
## -----
##
##                               Model Summary
## -----
## R                0.979          RMSE                7427.091
## R-Squared         0.959          Coef. Var            6.344
## Adj. R-Squared    0.957          MSE                55161684.725
## Pred R-Squared    0.953          MAE                6057.542
## -----
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
## -----
##                               Sum of
##                               Squares      DF      Mean Square      F      Sig.
## -----
## Regression    55818532409.384          2      27909266204.692      505.954      0.0000
## Residual      2371952443.176          43           55161684.725
## Total         58190484852.560          45
## -----
##
##                               Parameter Estimates
## -----
##                               model      Beta      Std. Error      Std. Beta      t      Sig.      lower      upper
## -----
## (Intercept)    62996.044      5252.182              11.994      0.000      52404.009      73588.078
## R.D.Spend       0.814        0.026              0.989      31.507      0.000           0.762           0.866
```

```
## Administration      -0.079      0.042      -0.059      -1.875      0.068      -0.164      0.006
## -----
##
##                               Stepwise Selection Summary
## -----
##      Step      Variable      Added/      R-Square      Adj.      C(p)      AIC      RMSE
##      Step      Variable      Removed      R-Square      R-Square      C(p)      AIC      RMSE
## -----
##      1      R.D.Spend      addition      0.956      0.955      3.5750      957.0424      7636.5800
##      2      Administration      addition      0.959      0.957      2.1290      955.4259      7427.0913
## -----
```

Algorytm proponuje tylko dwie cechy w modelu R.D.Spend i Administration.

```
model_2zm <- lm(Profit ~ R.D.Spend + Administration, data=startupy)
summary(model_2zm)
```

```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + Administration, data = startupy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16632  -4867  -2488   6280  14410
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  62996.04359  5252.18190  11.994 0.000000000000000262 ***
## R.D.Spend      0.81412    0.02584  31.507 < 0.00000000000000002 ***
## Administration -0.07897    0.04211  -1.875    0.0675 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7427 on 43 degrees of freedom
## Multiple R-squared:  0.9592, Adjusted R-squared:  0.9573
## F-statistic: 506 on 2 and 43 DF, p-value: < 0.00000000000000022
```

W związku z czym stwierdzamy, że tylko te dwie cechy wpływają istotnie na przychód - wydatki na R&D wpływają na niego pozytywnie, a na administrację ujemnie.

Chapter 4

Grupowanie

Grupowanie polega na przypisaniu obiektów do określonych grup/klastrów/skupień/segmentów, w których znajdują się jednostki najbardziej do siebie podobne, a powstałe grupy będą się między sobą różnić. Całe utrudnienie polega na tym, że nie wiemy ile tych grup ma powstać.

4.1 Metoda k-średnich

Najpopularniejszą metodą grupowania jest metoda k-średnich. Do jej zalet należy zaliczyć to, że dobrze działa zarówno na małych, jak i dużych zbiorach i jest bardzo efektywny - zwykle osiąga zbieżność w mniej niż 10 iteracjach. Z wad należy wskazać losowy wybór początkowych centrów skupień, co może skutkować nieprawidłowym przypisaniem obiektów do grup.

Algorytm postępowania jest następujący:

1. Wskaż liczbę grup k .
2. Wybierz dowolne k punktów jako centra grup.
3. Przypisz każdą z obserwacji do najbliższego centroidu.
4. Oblicz nowe centrum grupy.
5. Przypisz każdą z obserwacji do nowych centroidów. Jeśli któraś obserwacja zmieniła grupę - przejdź do kroku nr 4, a w przeciwnym przypadku zakończ algorytm.

Wykorzystamy informacje ze zbioru zawierającego informacje o klientach sklepu i dokonamy grupowania tych klientów.

Opis zbioru:

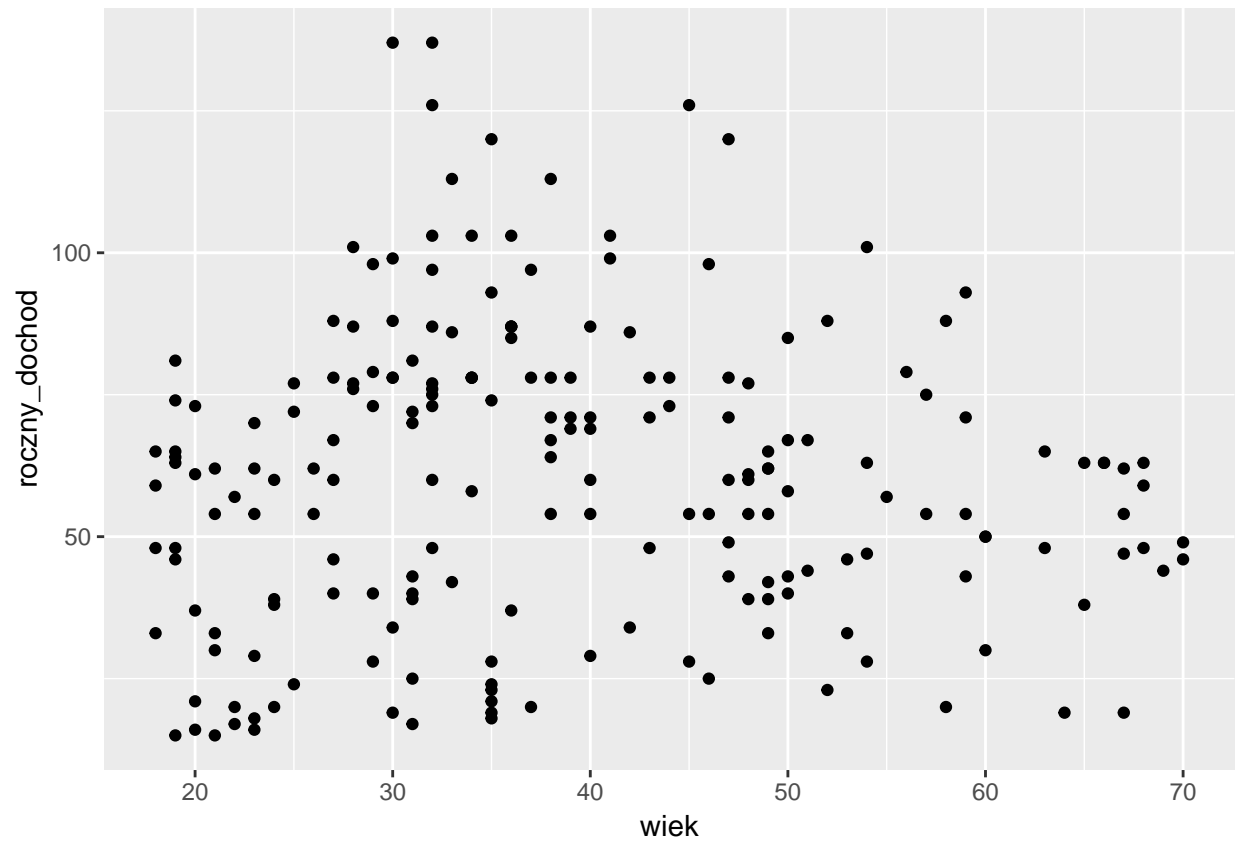
- klientID - identyfikator klienta
- plec - płeć
- wiek - wiek
- roczny_dochod - roczny dochód wyrażony w tys. dolarów
- wskaznik_wydatkow - klasyfikacja sklepu od 1 do 100

Wczytujemy zbiór danych i sprawdzamy czy pomiędzy zmiennymi są widoczne jakieś zależności.

```
library(tidyverse)

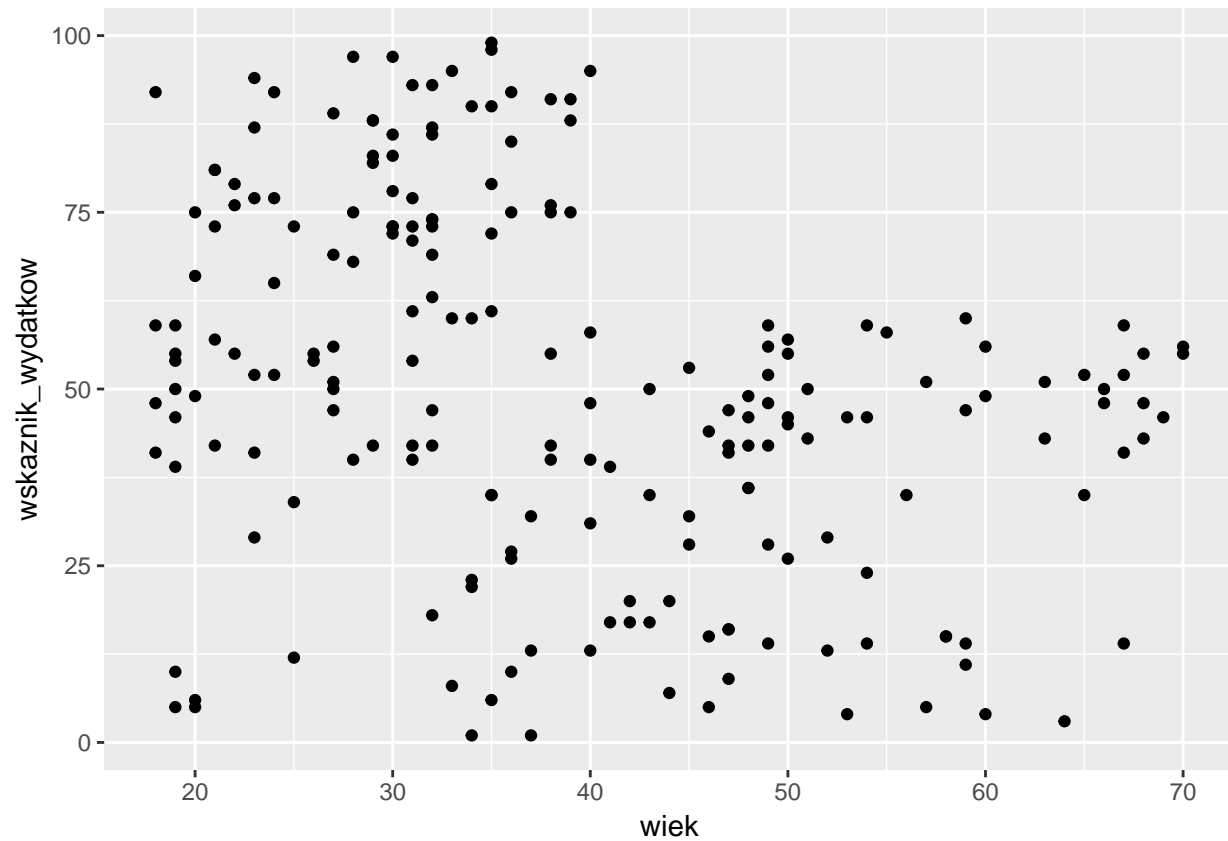
klienci <- read.csv("data/klienci.csv")

ggplot(klienci, aes(x=wiek, y=roczny_dochod)) +
  geom_point()
```



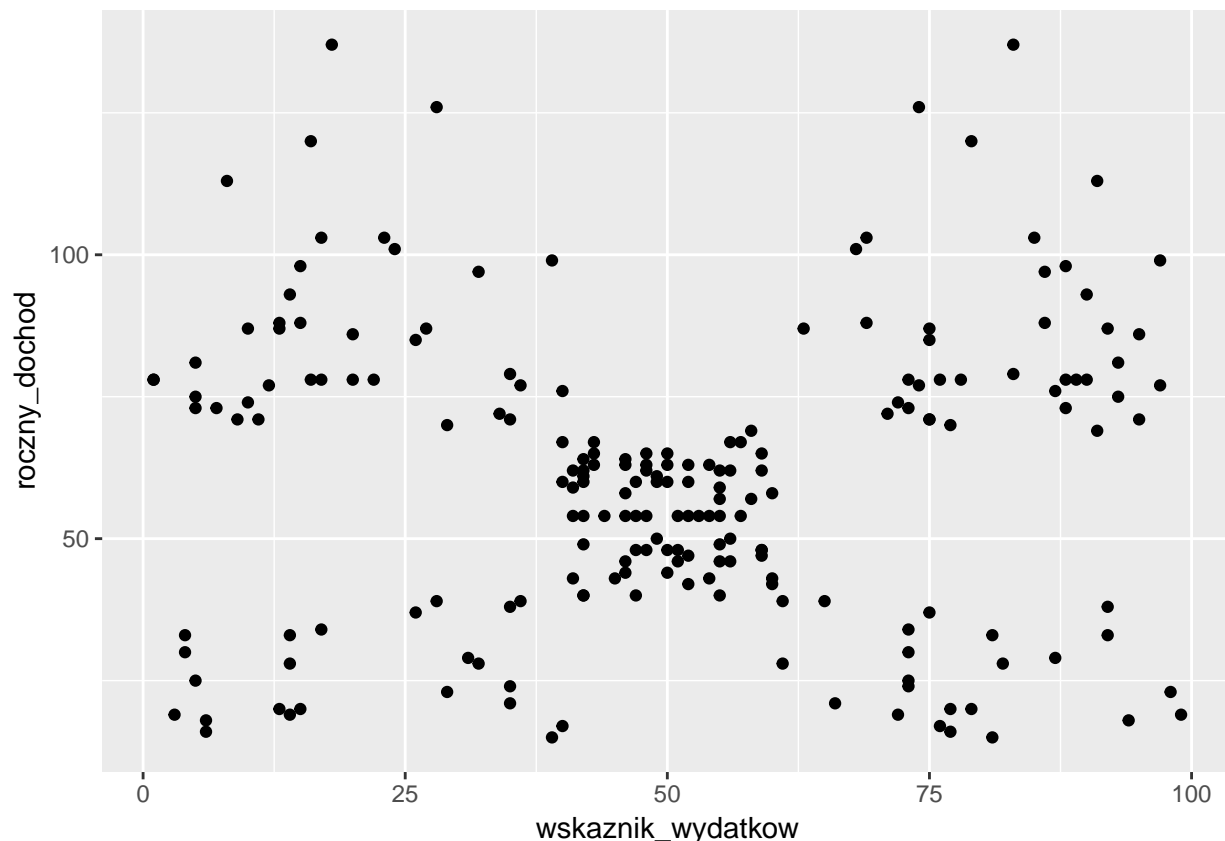
Pomiędzy wiekiem a rocznym dochodem nie widać zależności.

```
ggplot(klienci, aes(x=wiek, y=wskaznik_wydatkow)) +  
  geom_point()
```

W przypadku wieku i wskaźnika wydatków moglibyśmy się pokusić o podział zbioru na dwie grupy za pomocą przekątnej.

```
ggplot(klienci, aes(x=wskaznik_wydatkow, y=roczny_dochod)) +  
  geom_point()
```



Po zestawieniu rocznego dochodu i wskaźnika wydatków wyłania się 5 potencjalnych grup, zatem te dwie cechy wykorzystamy do grupowania. Jednak przed zastosowaniem algorytmu musimy te dane przygotować normalizując zakres obu cech - w tym przypadku za pomocą standaryzacji.

```
klienci_z <- klienci %>%
  select(roczny_dochod, wskaznik_wydatkow) %>%
  scale()
head(klienci_z)
```

```
##      roczny_dochod wskaznik_wydatkow
## [1,]      -1.734646        -0.4337131
## [2,]      -1.734646         1.1927111
## [3,]      -1.696572        -1.7116178
## [4,]      -1.696572         1.0378135
## [5,]      -1.658498        -0.3949887
## [6,]      -1.658498         0.9990891
```

W przypadku, gdy podział na grupy nie jest tak oczywisty lub bierzemy pod uwagę więcej niż dwa kryteria to wówczas w wyznaczeniu optymalnej liczby skupień może pomóc wykres osypiska (ang. elbow method). Polega to na przeprowadzeniu grupowania (z wykorzystaniem funkcji `kmeans()`) dla różnej liczby grup i porównanie wariancji wewnątrz-grupowej. Dane do stworzenia wykresu osypiska możemy obliczyć w pętli:

```
zm_w_gr <- numeric(15)

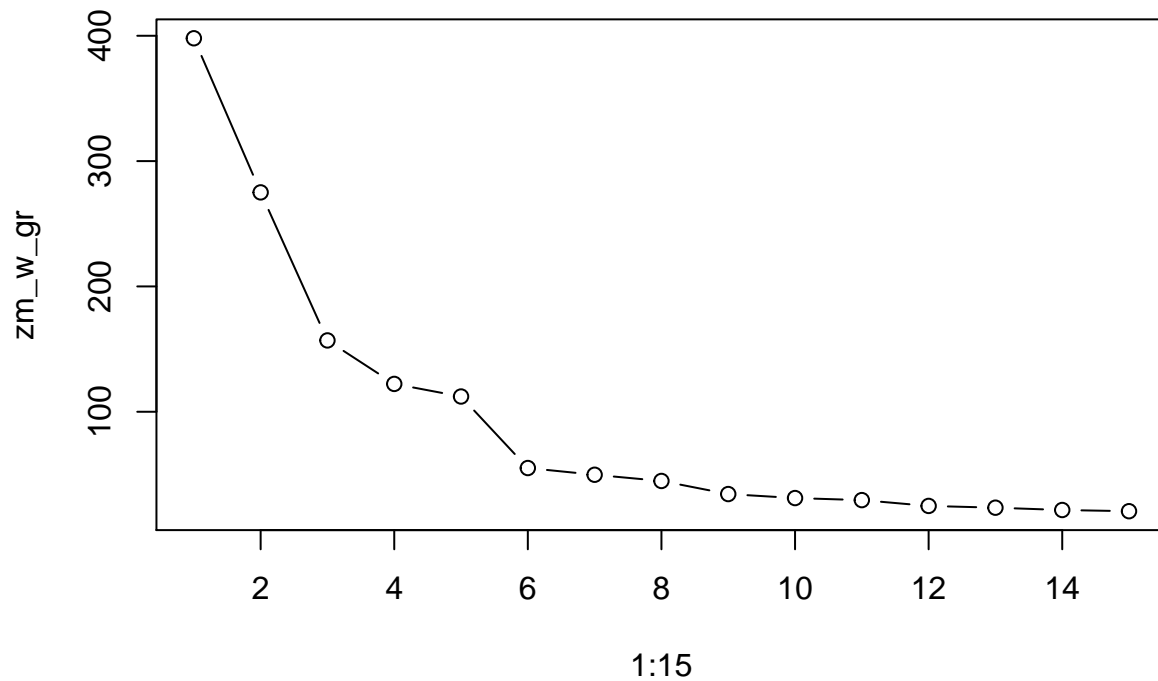
# wprowadzenie pętli
for(i in 1:length(zm_w_gr)) {
```

```

set.seed(14)
gr <- kmeans(klienci_z, centers = i)
zm_w_gr[i] <- gr$tot.withinss
}

plot(1:15, zm_w_gr, type="b")

```



Wybieramy liczbę skupień po której nie następuje już gwałtowny spadek wartości wariancji wewnątrzgrupowej. Według tego kryterium powinniśmy wybrać wartość 6 zamiast 5. Sprawdźmy zatem jakie otrzymamy przyporządkowanie do grup. Następnie informację o tym przypisaniu umieszczamy w oryginalnym zbiorze danych i przedstawiamy na wykresie. W celu uzyskania powtarzalnych wyników zastosujemy stałe ziarno generatora.

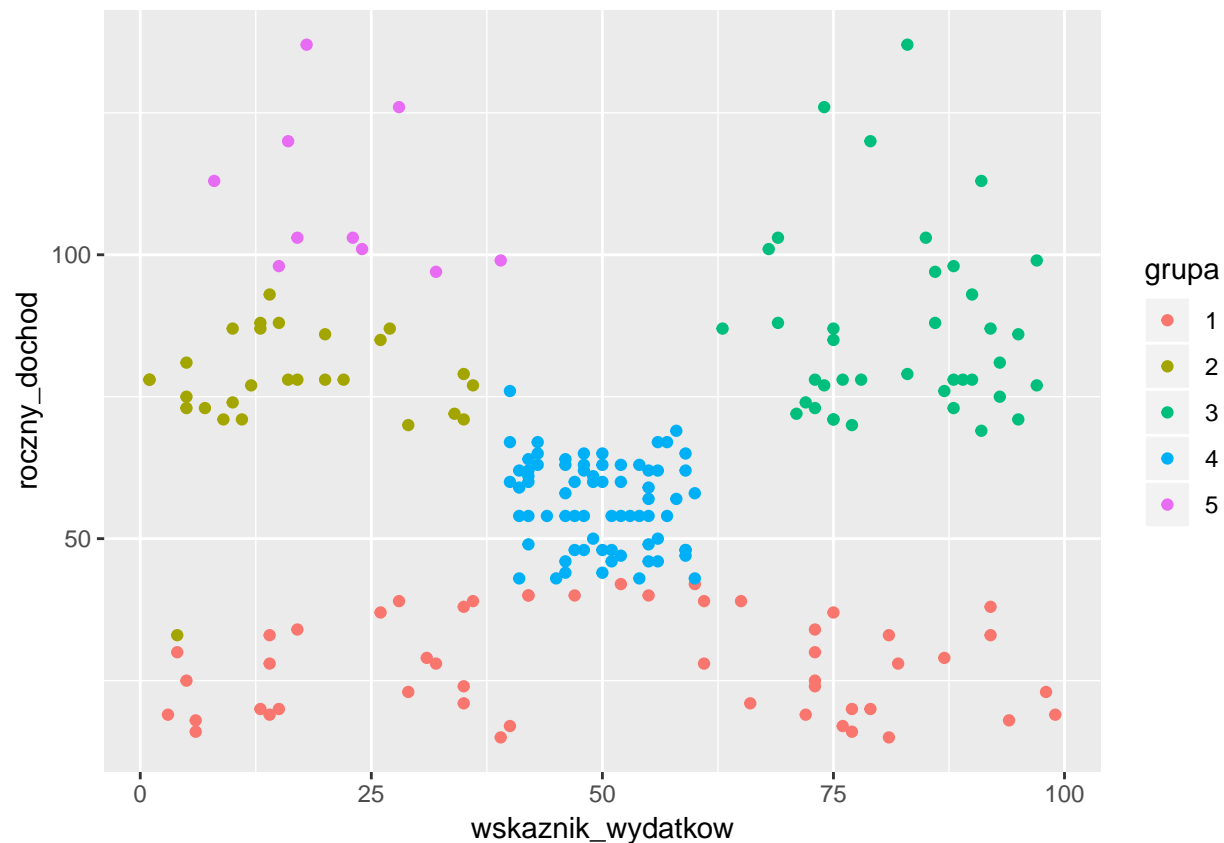
```

set.seed(14)
grupy <- kmeans(x = klienci_z, centers = 5)

klienci$grupa <- as.factor(grupy$cluster)

ggplot(klienci, aes(x=wskaznik_wydatkow,
                    y=roczny_dochod,
                    color=grupa)) +
  geom_point()

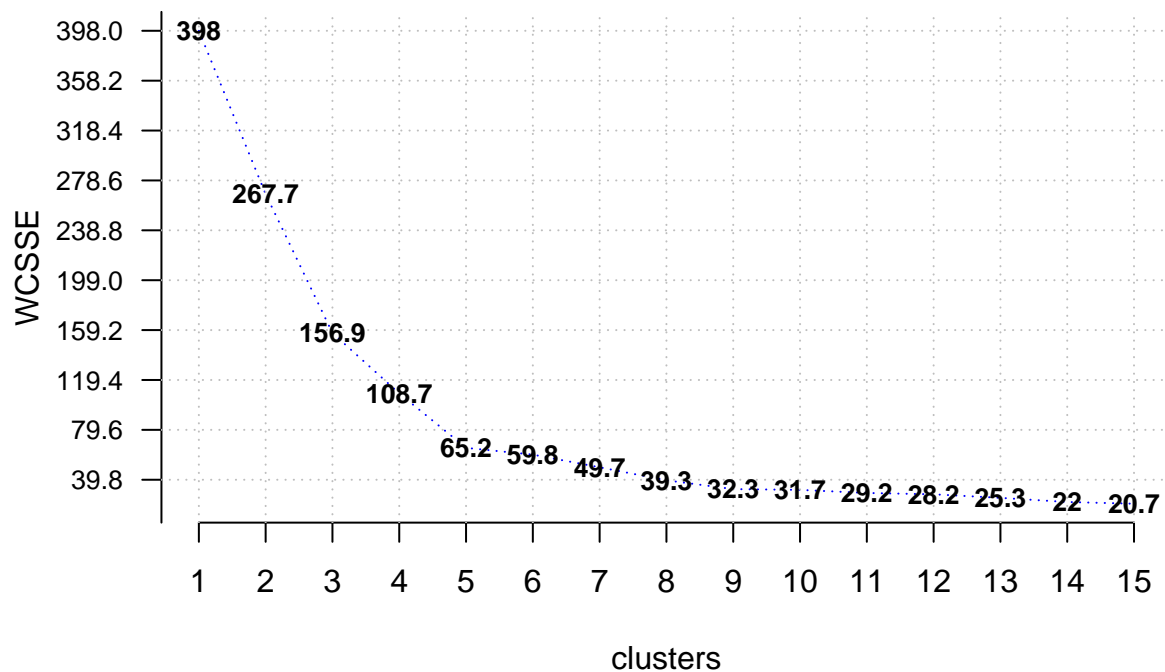
```



Jak zauważamy ten podział nie jest właściwy. Ze względu na losowy przydział centrów skupień na początku algorytmu istnieje spora szansa, że rozwiązanie nie będzie optymalne. Rozwiązaniem tego problemu jest użycie algorytmu `kmeans++` do początkowego ustalenia centrów. Ta metoda jest zaimplementowana w pakiecie `ClusterR`. Ponadto jest tam także funkcja do ustalenia optymalnej liczby skupień na podstawie wykresu osypiska.

```
library(ClusterR)
```

```
Optimal_Clusters_KMeans(data = klienci_z, max_clusters = 15, criterion = "WCSSE")
```



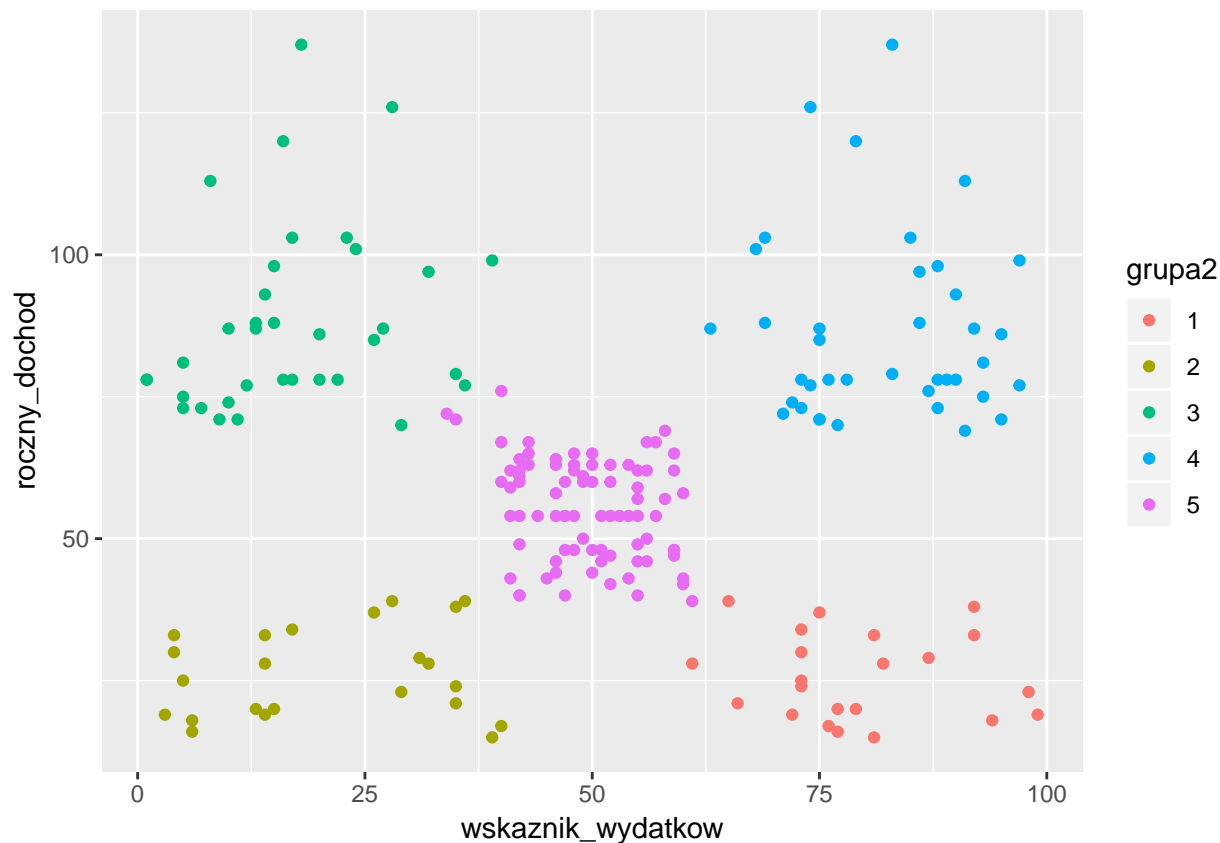
```
## [1] 398.00000 267.67171 156.91549 108.68209 65.24057 59.83221 49.72816
## [8] 39.31585 32.33261 31.70877 29.24003 28.18491 25.34096 22.01873
## [15] 20.67234
## attr(,"class")
## [1] "k-means clustering"
```

Wybieramy liczbę skupień po której nie następuje już gwałtowny spadek wartości wariancji wewnątrz-grupowej. W analizowanym przypadku będzie to 5 grup. Następnie korzystamy z funkcji `KMeans_rcpp` do wyznaczenia przynależności do grup. Ta funkcja domyślnie korzysta z algorytmu `kmeans++`, zatem nie ma niebezpieczeństwa, że uzyskamy niewłaściwe przyporządkowanie.

```
grupy2 <- KMeans_rcpp(data = klienci_z, clusters = 5)

klienci$grupa2 <- as.factor(grupy2$clusters)

ggplot(klienci, aes(x=wskaznik_wydatkow,
                    y=roczny_dochod,
                    color=grupa2)) +
  geom_point()
```



Ostatnim etapem analizy jest odpowiednia charakterystyka uzyskanych klastrow - najczęściej wyznacza się średnie wartości cech w ramach każdej grupy.

```
klienci %>%
  select(-klientID, -plec, -grupa) %>%
  group_by(grupa2) %>%
  summarise_all(.funs = "mean")
```

```
## # A tibble: 5 x 4
##   grupa2  wiek roczny_dochod wskaznik_wydatkow
##   <fct>  <dbl>      <dbl>          <dbl>
## 1 1      25.3      25.7           79.4
## 2 2      45.2      26.3           20.9
## 3 3      41.1      88.2           17.1
## 4 4      32.7      86.5           82.1
## 5 5      42.7      55.3           49.5
```

W grupie pierwszej znalazły się osoby z niskimi dochodami i wysokim wskaźnikiem wydatków. Grupa druga to klienci o niskich dochodach i wydatkach - ich przeciwieństwem jest grupa 4. W grupie 3 są osoby z wysokimi dochodami, ale niskimi wydatkami. Grupa 5 to z kolei średniacy - klienci o średnich dochodach i wydatkach.

4.2 Metoda hierarchiczna

Alternatywną metodą grupowania jest metoda hierarchiczna. Do jej zalet zaliczymy prosty sposób ustalenia liczby grup oraz praktyczny sposób wizualizacji. Niestety nie jest to metoda odpowiednia dla dużych zbiorów

danych.

Algorytm postępowania:

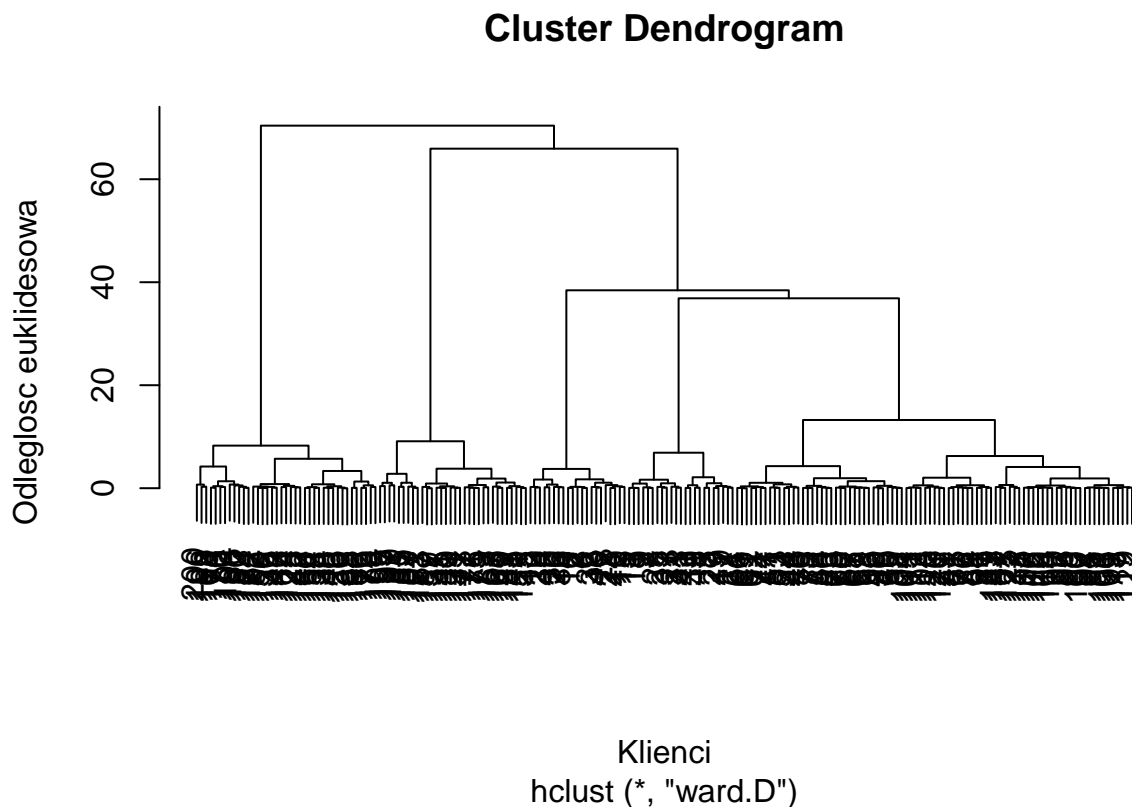
1. Każda obserwacji stanowi jedną z N pojedynczych grup.
2. Na podstawie macierzy odległości połącz dwie najbliższe leżące obserwacje w jedną grupę ($N - 1$ grup).
3. Połącz dwa najbliższe sobie leżące grupy w jedną ($N - 2$ grup).
4. Powtórz krok nr 3, aż do uzyskania jednej grupy.

Dla tych samych danych przeprowadzimy grupowanie, ale tym razem metodą hierarchiczną. W metodzie hierarchicznej bazuje się na macierzy odległości pomiędzy obserwacjami. Można zastosować wiele miar odległości, ale najczęściej wykorzystywana jest odległość euklidesowa. Drugą zmienną, na którą mamy wpływ to metoda łączenia skupień - w tym przypadku najlepsze rezultaty daje metoda Warda. Z kolei wyniki grupowania metodą hierarchiczną są prezentowane na dendrogramie.

```
macierz_odl <- dist(klienci_z)

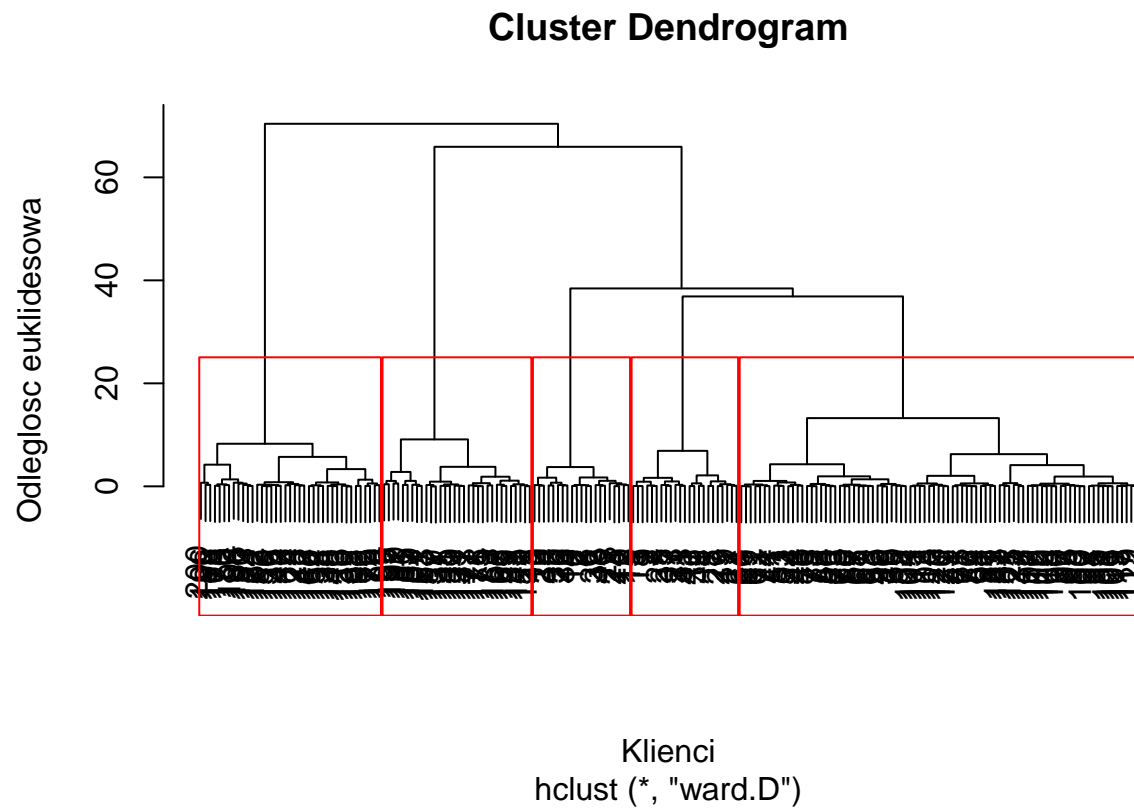
dendrogram <- hclust(macierz_odl, method = "ward.D")

plot(dendrogram, xlab="Klienci", ylab="Odległość euklidesowa")
```



Na podstawie dendrogramu identyfikujemy największe różnice odległości opisane na osi Y. Także w tym przypadku identyfikujemy 5 grup. Istnieje także wiele kryteriów, które mają na celu wyznaczyć optymalną liczbę grup - link.

```
plot(dendrogram, xlab="Klienci", ylab="Odległość euklidesowa")
rect.hclust(dendrogram, k=5, border="red")
```

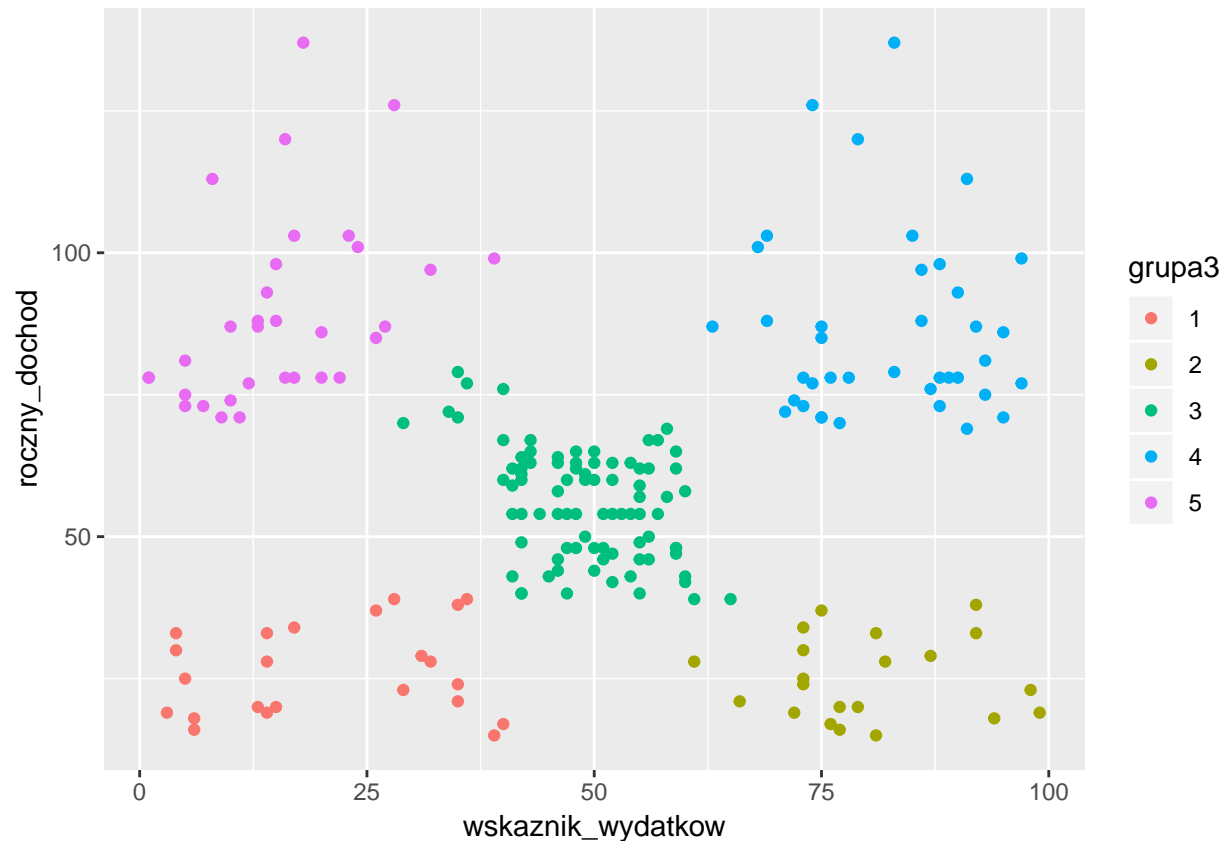


Następnie dopisujemy do oryginalnego zbioru danych etykiety utworzonych grup.

```
grupy_dendro <- cutree(dendrogram, 5)

klienci$grupa3 <- as.factor(grupy_dendro)

ggplot(klienci, aes(x=wskaznik_wydatkow,
                    y=roczny_dochod,
                    color=grupa3)) +
  geom_point()
```

Uzyskane wyniki są bardzo zbliżone do tych otrzymanych za pomocą algorytmu k-średnich.

```
klienci %>%
  select(-klientID, -plec, -grupa, -grupa2) %>%
  group_by(grupa3) %>%
  summarise_all(.funs = "mean")
```

```
## # A tibble: 5 x 4
##   grupa3  wiek roczny_dochod wskaznik_wydatkow
##   <fct>  <dbl>      <dbl>          <dbl>
## 1 1      45.2      26.3           20.9
## 2 2      25.3      25.1           80.0
## 3 3      42.5      55.8           49.1
## 4 4      32.7      86.5           82.1
## 5 5       41      89.4           15.6
```

Metoda hierarchiczna zastosowała inną numerację grup. Liczebności tych grup nieznacznie się różnią, ale charakterystyki wewnątrz grupowe są bardzo podobne do tych określonych na podstawie metody k-średnich.

Tworząc tabelę krzyżową możemy zobaczyć, że tylko 4 obserwacje zmieniły przypisanie do grup.

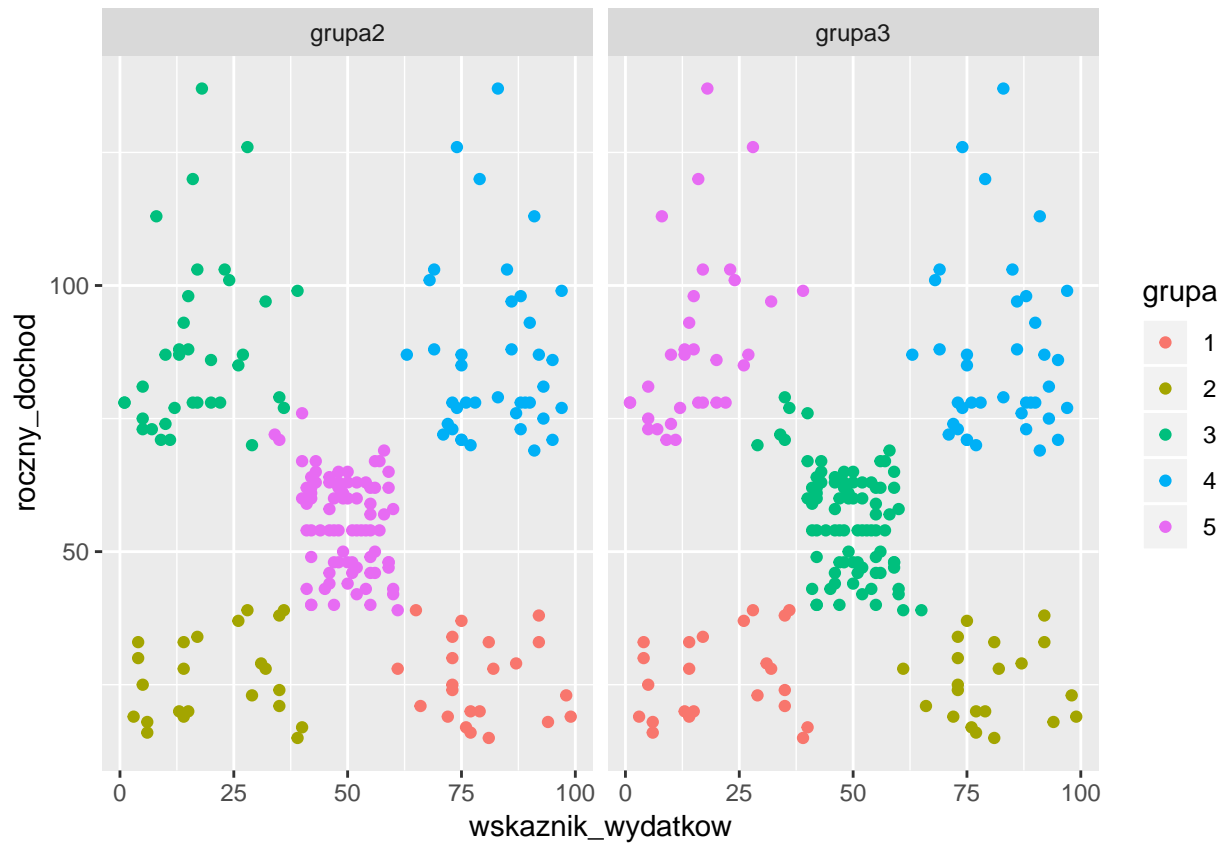
```
table(klienci$grupa2, klienci$grupa3)
```

```
##
##      1  2  3  4  5
## 1  0 21  1  0  0
## 2 23  0  0  0  0
## 3  0  0  3  0 32
## 4  0  0  0 39  0
```

```
## 5 0 0 81 0 0
```

Porównajmy jeszcze wyniki działania tych dwóch metod na wykresach:

```
klienci %>%
  select(rocny_dochod, wskaznik_wydatkow, grupa2, grupa3) %>%
  gather(metoda, grupa, -rocny_dochod, -wskaznik_wydatkow) %>%
  ggplot(aes(x=wskaznik_wydatkow, y=rocny_dochod)) +
  geom_point(aes(color=grupa)) +
  facet_wrap(~ metoda)
```



Problematic observations came from the group of clients with average incomes and expenditures.

4.2.1 Zadania

1. Dokonaj grupowania danych dotyczących 32 samochodów według następujących zmiennych: pojemność, przebieg, lata oraz cena.
2. Rozpoznawanie czynności na podstawie danych z przyspieszeniomierza w telefonie: User Identification From Walking Activity Data Set

Chapter 5

Klasyfikacja

A visual introduction to machine learning

5.1 Drzewa klasyfikacyjne

Zalety:

- łatwa interpretacja
- nie trzeba normalizować cech
- rozwiązuje problemy liniowe i nieliniowe

Wady:

- mała efektywność przy małych zbiorach danych
- łatwo można przeuczyć

5.2 KNN

Algorytm:

1. Określ liczbę sąsiadów - K
2. Wyznacz K sąsiadów dla nowego punktu na podstawie wybranej odległości
3. Oblicz liczbę sąsiadów, w każdej z grup
4. Przypisz nową obserwację do grupy, w której ma więcej najbliższych sąsiadów

Zalety:

- łatwa interpretacja
- szybki i efektywny

Wady:

- trzeba określić liczbę sąsiadów

5.2.1 Zadanie

Zbuduj model klasyfikacyjny dla zbioru danych dotyczących cech internautów oraz informacji czy zamówili reklamowany produkt czy nie.

Przeprowadź imputację braków danych dla zbioru pracowników.

Chapter 6

Materiały z zajęć

6.1 28.10.2018

Wprowadzenie do R

Analiza sejmików

6.2 18.11.2018

Analiza struktury

Rossmann

Analiza struktury w R

6.3 16.12.2018

Prezentacja

Pracownicy

Korelacja w R

Regresja w R

6.4 26.01.2019

Pensja i doświadczenie

Pracownicy

Opis zbioru:

- id - kod pracownika
- plec - płeć pracownika (0 - mężczyzna, 1 - kobieta)
- data_urodz - data urodzenia
- edukacja - wykształcenie (w latach nauki)
- kat_pracownika - grupa pracownicza (1 - ochroniarz, 2 - urzędnik, 3 - menedżer)

- bwynagrodzenie - bieżące wynagrodzenie
- pwynagrodzenie - początkowe wynagrodzenie
- staz - staż pracy (w miesiącach)
- doswiadczenie - poprzednie zatrudnienie (w miesiącach)
- zwiazki - przynależność do związków zawodowych (0 - nie, 1 - tak)
- wiek - wiek (w latach)

Regresja w R