zadanie

```r
library(glmnet)
```

```
Warning: package 'glmnet' was built under R version 4.3.2
```

```
Loading required package: Matrix
```

```
Loaded glmnet 4.1-8
```

```r
library(readxl)
```

```
Warning: package 'readxl' was built under R version 4.3.2
```

```r
library(readr)
```

```
Warning: package 'readr' was built under R version 4.3.2
```

```r
library(tidymodels)
```

```
Warning: package 'tidymodels' was built under R version 4.3.2
```

```
-- Attaching packages -------------------------------------- tidymodels 1.1.1 --
```

```
v broom        1.0.5     v recipes      1.0.9
v dials        1.2.0     v rsample      1.2.0
v dplyr        1.1.4     v tibble       3.2.1
v ggplot2      3.4.4     v tidyr        1.3.0
v infer        1.0.5     v tune         1.1.2
v modeldata    1.2.0     v workflows    1.1.3
v parsnip      1.1.1     v workflowsets 1.0.1
v purrr        1.0.2     v yardstick    1.2.0

Warning: package 'broom' was built under R version 4.3.2

Warning: package 'dials' was built under R version 4.3.2

Warning: package 'scales' was built under R version 4.3.2

Warning: package 'dplyr' was built under R version 4.3.2

Warning: package 'ggplot2' was built under R version 4.3.2

Warning: package 'infer' was built under R version 4.3.2

Warning: package 'modeldata' was built under R version 4.3.2

Warning: package 'parsnip' was built under R version 4.3.2

Warning: package 'purrr' was built under R version 4.3.2

Warning: package 'recipes' was built under R version 4.3.2

Warning: package 'rsample' was built under R version 4.3.2

Warning: package 'tibble' was built under R version 4.3.2

Warning: package 'tidyr' was built under R version 4.3.2

Warning: package 'tune' was built under R version 4.3.2

Warning: package 'workflows' was built under R version 4.3.2
```

```
Warning: package 'workflowsets' was built under R version 4.3.2
```

```
Warning: package 'yardstick' was built under R version 4.3.2
```

```
-- Conflicts ----------------------------------------- tidymodels_conflicts() --
x purrr::discard()  masks scales::discard()
x tidyr::expand()   masks Matrix::expand()
x dplyr::filter()   masks stats::filter()
x dplyr::lag()      masks stats::lag()
x tidyr::pack()     masks Matrix::pack()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
x tidyr::unpack()   masks Matrix::unpack()
x recipes::update() masks Matrix::update(), stats::update()
* Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```r
library(rsample)
library(dplyr)
library(fastDummies)
```

```
Warning: package 'fastDummies' was built under R version 4.3.2
```

```
Thank you for using fastDummies!
```

```
To acknowledge our work, please cite the package:
```

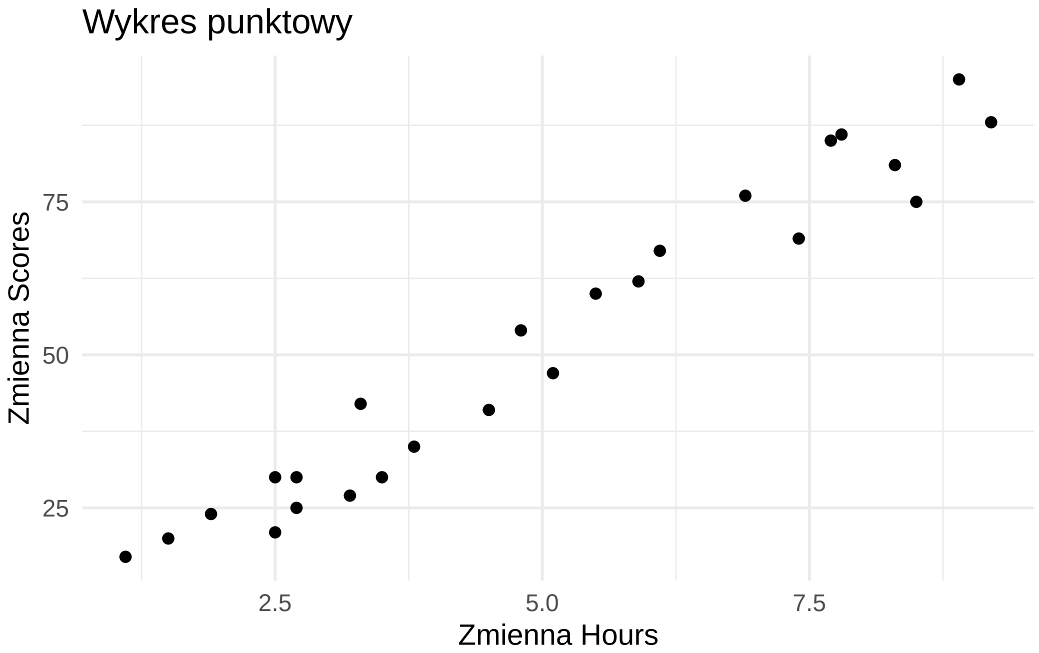
```
Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and R
```

Zadanie 1

SCORES

```r
data <- read.csv("SCORES.csv")

# Rysowanie wykresu punktowego
plot1 <- ggplot(data, aes(x = Hours, y = Scores)) +
  geom_point()+
  labs(title = "Wykres punktowy", x = "Zmienna Hours", y = "Zmienna Scores") +
  theme_minimal()
plot1
```



```
# Podział danych na zbiory X i Y
X <- data$Hours
Y <- data$Scores

# Podział na zbiór treningowy i testowy
set.seed(19)
split <- initial_split(data, prop = 0.7, strata = Scores)
```

```
Warning: The number of observations in each quantile is below the recommended threshold of 20
* Stratification will use 1 breaks instead.
```

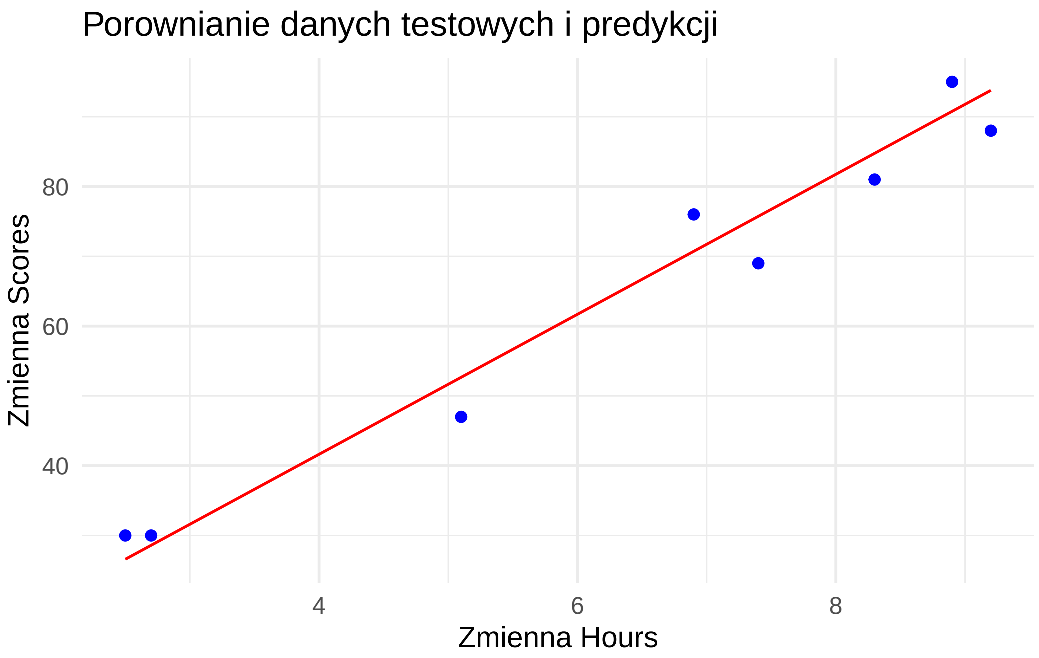
```
Warning: Too little data to stratify.
* Resampling will be unstratified.
```

```
train_data <- training(split)
test_data <- testing(split)

# Tworzenie modelu regresji liniowej
lm_model <- lm(Scores ~ Hours, data = train_data)

# Predykcja dla danych testowych
y_pred <- predict(lm_model, newdata = test_data)

# Wykres konfrontujący dane testowe i predykcje za pomocą ggplot2
plot2 <- ggplot() +
  geom_point(data = test_data, aes(x = Hours, y = Scores), color = "blue") +
  geom_line(data = data.frame(Hours = test_data$Hours, Scores = y_pred), aes(x = Hours, y
  labs(title = "Porownianie danych testowych i predykcji", x = "Zmienna Hours", y = "Zmien
  theme_minimal()
plot2
```



```
# Analiza dopasowania modelu
summary(lm_model)
```

```
Call:
lm(formula = Scores ~ Hours, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.759  -5.596   3.322   4.341   7.382 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.5280     3.2761   0.466    0.648    
Hours        10.0272     0.6701  14.964 2.01e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.054 on 15 degrees of freedom
Multiple R-squared:  0.9372,    Adjusted R-squared:  0.933 
F-statistic: 223.9 on 1 and 15 DF,  p-value: 2.008e-10
```

```r
# Walidacja predykcji
mae <- mean(abs(test_data$Scores - y_pred))
mse <- mean((test_data$Scores - y_pred) ^ 2)
rmse <- sqrt(mse)

cat("Sredni blad bezwzgledny (Mean Absolute Error):", mae, "\n")
```

```
Sredni blad bezwzgledny (Mean Absolute Error): 4.530642 
```

```r
cat("Blad sredniokwadratowy (Mean Squared Error:", mse, "\n")
```

```
Blad sredniokwadratowy (Mean Squared Error: 23.02989 
```

```r
cat("Pierwiastek bledu sredniokwadratowego (Root Mean Squared Error):", rmse, "\n")
```

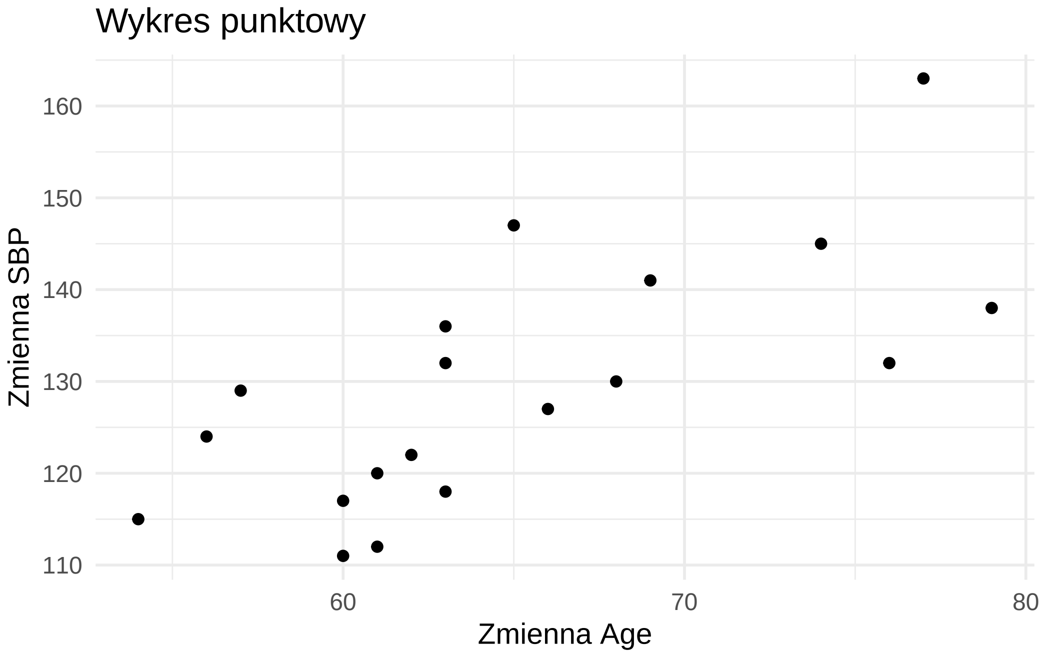
```
Pierwiastek bledu sredniokwadratowego (Root Mean Squared Error): 4.798947 
```

Wyniki sa calkiem zadowalajace, MAE wynosi prawie 4 co w naszym przypadku gdy dane maja wielkosci z zakresu 20-100 nie jest najgorszym wynikiem

SBP

```r
data <- read.csv("SBP.csv")

plot1 <- ggplot(data, aes(x = Age, y = SBP)) +
  geom_point() +
  labs(title = "Wykres punktowy", x = "Zmienna Age", y = "Zmienna SBP") +
  theme_minimal()
plot1
```



```r
X <- data$Age
Y <- data$SBP

set.seed(19 * 2)
split <- initial_split(data, prop = 0.7, strata = SBP)
```

```
Warning: The number of observations in each quantile is below the recommended threshold of 20
* Stratification will use 0 breaks instead.
```

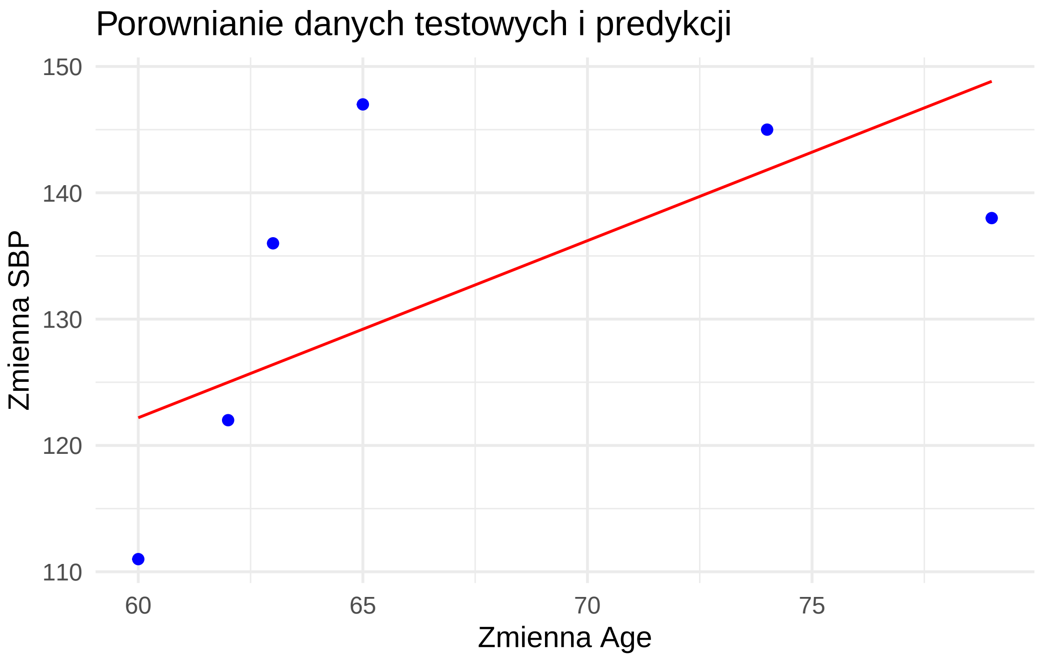
```
Warning: Too little data to stratify.
* Resampling will be unstratified.
```

```
train_data = training(split)
test_data = testing(split)

lm_model = lm(SBP ~ Age, data = train_data)

y_pred = predict(lm_model, newdata = test_data)

plot2 <- ggplot() +
  geom_point(data = test_data, aes(x = Age, y = SBP), color = "blue") +
  geom_line(data = data.frame(Age = test_data$Age, SBP = y_pred), aes(x = Age, y = SBP), c
  labs(title = "Porownianie danych testowych i predykcji", x = "Zmienna Age", y = "Zmienna
  theme_minimal()
plot2
```



```
summary(lm_model)
```

```
Call:
lm(formula = SBP ~ Age, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.620  -5.194  -3.407   6.192  16.979 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  38.0947    24.4337   1.559  0.14726   
Age           1.4016     0.3801   3.688  0.00358 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.363 on 11 degrees of freedom
Multiple R-squared:  0.5529,    Adjusted R-squared:  0.5122 
F-statistic:  13.6 on 1 and 11 DF,  p-value: 0.003577
```

```r
mae <- mean(abs(test_data$SBP - y_pred))
mse <- mean((test_data$SBP - y_pred) ^ 2)
rmse <- sqrt(mse)

cat("Sredni blad bezwzgledny (Mean Absolute Error):", mae, "\n")
```

```
Sredni blad bezwzgledny (Mean Absolute Error): 9.266392
```

```r
cat("Blad sredniokwadratowy (Mean Squared Error:", mse, "\n")
```

```
Blad sredniokwadratowy (Mean Squared Error: 111.7589
```

```r
cat("Pierwiastek bledu sredniokwadratowego (Root Mean Squared Error):", rmse, "\n")
```

```
Pierwiastek bledu sredniokwadratowego (Root Mean Squared Error): 10.57161
```

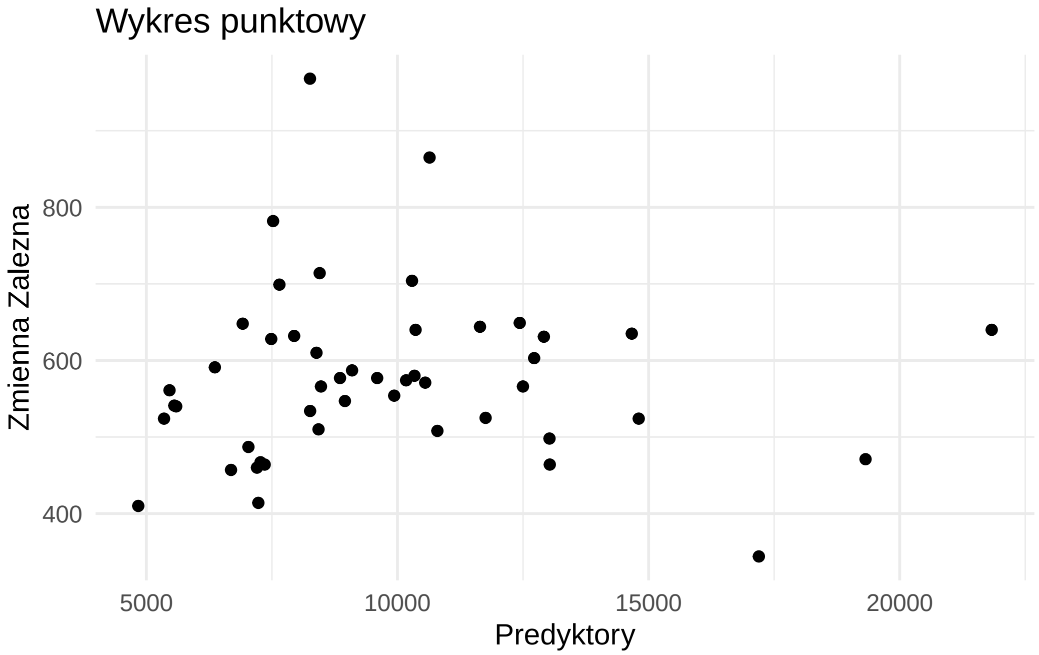
Wyniki sa lepsze niz w poprzednim przypadku.

Zadanie 2

PETROL

```r
data = read.csv("PETROL.csv")

plot1 <- ggplot(data, aes(x = Podatek_paliwowy + Sredni_przychod + Utwardzone_autostrady +
  geom_point() +
  labs(title = "Wykres punktowy", x = "Predyktory", y = "Zmienna Zalezna") +
  theme_minimal()
plot1
```



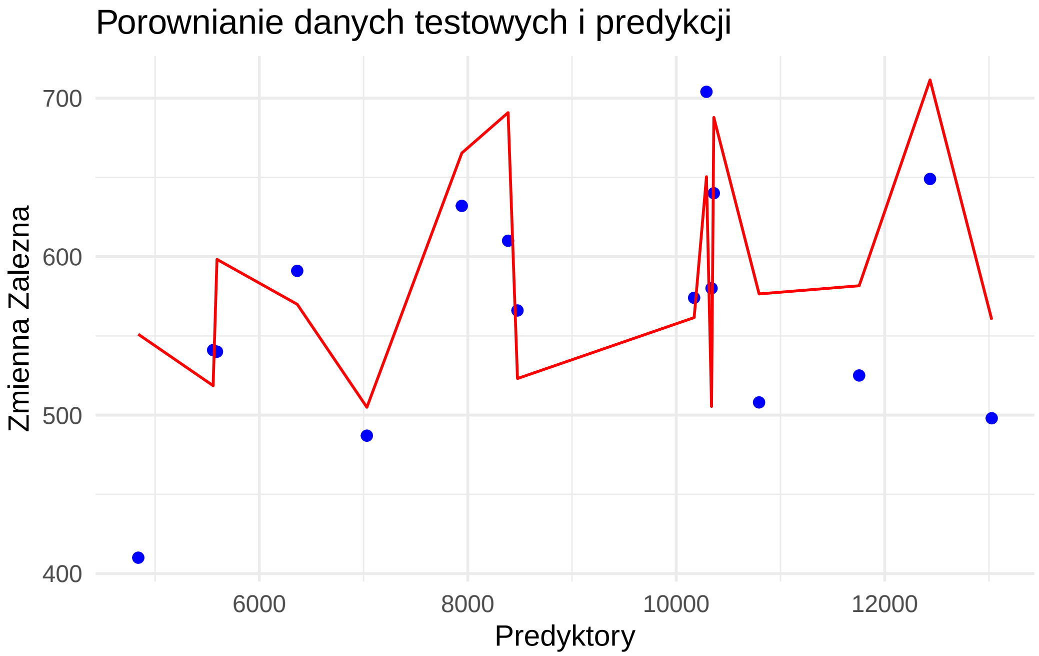
```r
set.seed(19 * 3)
split <- initial_split(data, prop = 0.7, strata = Zuzycie_paliwa)
```

```
Warning: The number of observations in each quantile is below the recommended threshold of 20
* Stratification will use 2 breaks instead.
```

```r
train_data = training(split)
test_data = testing(split)

lm_model = lm(Zuzycie_paliwa ~ Podatek_paliwowy + Sredni_przychod + Utwardzone_autostrady
y_pred = predict(lm_model, newdata = test_data)
```

```r
plot2 <- ggplot() +
  geom_point(data = test_data, aes(x = Podatek_paliwowy + Sredni_przychod + Utwardzone_aut
  geom_line(data = data.frame(Podatek_paliwowy = test_data$Podatek_paliwowy, Sredni_przych
  labs(title = "Porownianie danych testowych i predykcji", x = "Predyktory", y = "Zmienna
  theme_minimal()
plot2
```



```r
summary(lm_model)
```

```
Call:
lm(formula = Zuzycie_paliwa ~ Podatek_paliwowy + Sredni_przychod +
    Utwardzone_autostrady + Procent_ludnosci_z_prawem_jazdy,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-103.418  -51.265   -3.958   27.514  211.639

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                       3.926e+02  2.661e+02   1.475   0.1517
Podatek_paliwowy                 -4.359e+01  1.748e+01  -2.494   0.0191 *
Sredni_przychod                  -5.403e-02  2.039e-02  -2.649   0.0133 *
Utwardzone_autostrady            -4.861e-03  4.508e-03  -1.078   0.2904
Procent_ludnosci_z_prawem_jazdy  1.373e+03  2.652e+02   5.176  1.9e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.51 on 27 degrees of freedom
Multiple R-squared:  0.7268,    Adjusted R-squared:  0.6863
F-statistic: 17.96 on 4 and 27 DF,  p-value: 2.671e-07
```

```
mae <- mean(abs(test_data$Zuzycie_paliwa - y_pred))
mse <- mean((test_data$Zuzycie_paliwa - y_pred) ^ 2)
rmse <- sqrt(mse)

cat("Sredni blad bezwzgledny (Mean Absolute Error):", mae, "\n")
```

```
Sredni blad bezwzgledny (Mean Absolute Error): 53.50356
```

```
cat("Blad sredniokwadratowy (Mean Squared Error:", mse, "\n")
```

```
Blad sredniokwadratowy (Mean Squared Error: 3786.053
```

```
cat("Pierwiastek bledu sredniokwadratowego (Root Mean Squared Error):", rmse, "\n")
```

```
Pierwiastek bledu sredniokwadratowego (Root Mean Squared Error): 61.53091
```

Wyniki predykcji modelu sa dobre, wplyw na to maja dane do trenowania oraz testowe

Zadanie 3

HEART

```r
data = read.csv("HEART.csv")
data[data == "?"] <- NA
data <- data |> select(-c("slope", "ca", "thal"))
data <- na.omit(data)

data <- dummy_cols(data, select_columns = c("cp", "restecg"))

X <- data |> select(-"num")
y <- data$num

set.seed(19 * 4)
split <- initial_split(data, prop = 0.8, strata = y)
```

```
Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
i Please use `all_of()` or `any_of()` instead.
  # Was:
  data %>% select(y)

  # Now:
  data %>% select(all_of(y))

See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

```r
train_data <- training(split)
test_data <- testing(split)

#model <- logistic_reg(mixture = double(1), penalty = double(1)) |>
#   set_engine("glmnet") |>
#   set_mode("classification") |>
#   fit(num ~ ., data = train_data)
#
#
#pred_class <- predict(model,
#                       new_data = test_data,
#                       type = "class")
#
#pred_proba <- predict(model,
#                       new_data = test_data,
#                       type = "prob")
#
#results <- test_data |>
```

```
#           select(y) |>
#           bind_cols(pred_class, pred_proba)
#
#accuracy(results, truth = y, estimate = .pred_class)
```