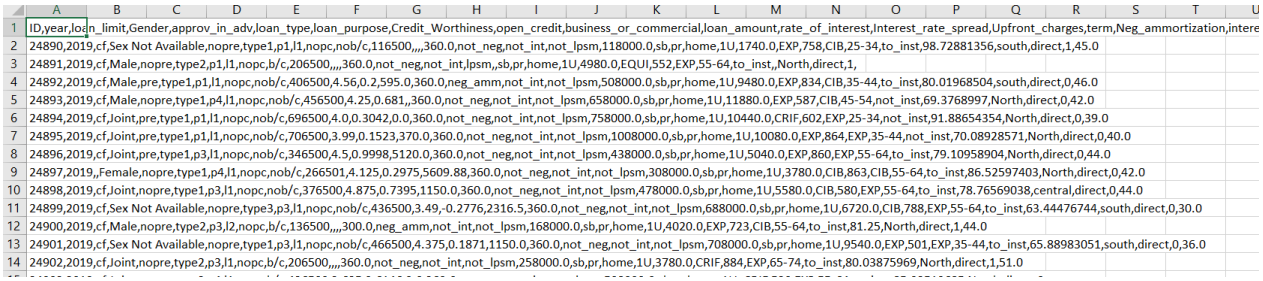
Statystyczna analiza danych - Projekt

Patrycja Milo 173182

11.06.2024

Dane zostały pobrane ze strony Kaggle: Loan Default Dataset. Dane zawarte w zbiorze dotyczą pożyczek udzielanych przez bank. Banki gromadzą różnorodne informacje na temat swoich klientów oraz pożyczek, aby podejmować świadome decyzje i skutecznie zarządzać ryzykiem. Cechy, które wybrałam do analizy, to kwota pożyczki, którą klient chce pożyczyć oraz dochód klienta.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	ID,year,loan_limit,Gender,approv_in_adv,loan_type,loan_purpose,Credit_Worthiness,open_credit,business_or_commercial,loan_amount,rate_of_interest,Interest_rate_spread,Upfront_charges,term,Neg_ammortization,intere																				
2	24890,2019,cf,Sex Not Available,nopre,type1,p1,l1,nopc,nob/c,116500,,,,360.0,not_neg,not_int,not_lpsm,118000.0,sb,pr,home,1U,1740.0,EXP,758,CIB,25-34,to_inst,98.72881356,south,direct,1,45.0																				
3	24891,2019,cf,Male,nopre,type2,p1,l1,nopc,b/c,206500,,,,360.0,not_neg,not_int,lpsm,,sb,pr,home,1U,4980.0,EQUI,552,EXP,55-64,to_inst,,North,direct,1,																				
4	24892,2019,cf,Male,pre,type1,p1,l1,nopc,nob/c,406500,4.56,0.2,595.0,360.0,neg_amm,not_int,not_lpsm,508000.0,sb,pr,home,1U,9480.0,EXP,834,CIB,35-44,to_inst,80.01968504,south,direct,0,46.0																				
5	24893,2019,cf,Male,nopre,type1,p4,l1,nopc,nob/c,456500,4.25,0.681,,360.0,not_neg,not_int,not_lpsm,658000.0,sb,pr,home,1U,11880.0,EXP,587,CIB,45-54,not_inst,69.3768997,North,direct,0,42.0																				
6	24894,2019,cf,Joint,pre,type1,p1,l1,nopc,nob/c,696500,4.0,0.3042,0.0,360.0,not_neg,not_int,not_lpsm,758000.0,sb,pr,home,1U,10440.0,CRIF,602,EXP,25-34,not_inst,91.88654354,North,direct,0,39.0																				
7	24895,2019,cf,Joint,pre,type1,p1,l1,nopc,nob/c,706500,3.99,0.1523,370.0,360.0,not_neg,not_int,not_lpsm,1008000.0,sb,pr,home,1U,10080.0,EXP,864,EXP,35-44,not_inst,70.08928571,North,direct,0,40.0																				
8	24896,2019,cf,Joint,pre,type1,p3,l1,nopc,nob/c,346500,4.5,0.9998,5120.0,360.0,not_neg,not_int,not_lpsm,438000.0,sb,pr,home,1U,5040.0,EXP,860,EXP,55-64,to_inst,79.10958904,North,direct,0,44.0																				
9	24897,2019,,Female,nopre,type1,p4,l1,nopc,nob/c,266501,4.125,0.2975,5609.88,360.0,not_neg,not_int,not_lpsm,308000.0,sb,pr,home,1U,3780.0,CIB,863,CIB,55-64,to_inst,86.52597403,North,direct,0,42.0																				
10	24898,2019,cf,Joint,nopre,type1,p3,l1,nopc,nob/c,376500,4.875,0.7395,1150.0,360.0,not_neg,not_int,not_lpsm,478000.0,sb,pr,home,1U,5580.0,CIB,580,EXP,55-64,to_inst,78.76569038,central,direct,0,44.0																				
11	24899,2019,cf,Sex Not Available,nopre,type3,p3,l1,nopc,nob/c,436500,3.49,-0.2776,2316.5,360.0,not_neg,not_int,not_lpsm,688000.0,sb,pr,home,1U,6720.0,CIB,788,EXP,55-64,to_inst,63.44476744,south,direct,0,30.0																				
12	24900,2019,cf,Male,nopre,type2,p3,l2,nopc,b/c,136500,,,,300.0,neg_amm,not_int,not_lpsm,168000.0,sb,pr,home,1U,4020.0,EXP,723,CIB,55-64,to_inst,81.25,North,direct,1,44.0																				
13	24901,2019,cf,Sex Not Available,nopre,type1,p3,l1,nopc,nob/c,466500,4.375,0.1871,1150.0,360.0,not_neg,not_int,not_lpsm,708000.0,sb,pr,home,1U,9540.0,EXP,501,EXP,35-44,to_inst,65.88983051,south,direct,0,36.0																				
14	24902,2019,cf,Joint,nopre,type2,p3,l1,nopc,b/c,206500,,,,360.0,not_neg,not_int,not_lpsm,258000.0,sb,pr,home,1U,3780.0,CRIF,884,EXP,65-74,to_inst,80.03875969,North,direct,1,51.0																				

Załadowanie potrzebnych bibliotek

```
library(dplyr)
library(DescTools)
library(moments)
library(openxlsx)
```

Przygotowanie danych

Ten kod wczytuje dane z pliku CSV, a następnie tworzy nowy zestaw danych, usuwając wiersze z brakującymi wartościami i wybierając tylko pierwsze 1000 wierszy oraz dwie kolumny: “kwota_pozyczki” (kwota pożyczki) i “dochod” (dochód).

```
dane_pozyczki <- read.csv("C:/Users/vanae/Desktop/loans_static_analysis_proje
ct/loans_data.csv")
przygotowane_dane <- na.omit(dane_pozyczki[1:1000, c("loan_amount", "income")
])
names(przygotowane_dane) <- c("kwota_pozyczki", "dochod")
```

Wyznaczenie podstawowych parametrów opisowych

Dla cechy “kwota pożyczki”

Średnia

```
srednia_kwota_pozyczki <- mean(przygotowane_dane$kwota_pozyczki)

## 332777.1
```

Odchylenie standardowe

```
odch_stand_kwota_pozyczki <- sd(przygotowane_dane$kwota_pozyczki)

## 184117.4
```

Mediana

```
mediana_kwota_pozyczki <- median(przygotowane_dane$kwota_pozyczki)

## 306500
```

Minimum

```
min_kwota_pozyczki <- min(przygotowane_dane$kwota_pozyczki)

## 26500
```

Maksimum

```
max_kwota_pozyczki <- max(przygotowane_dane$kwota_pozyczki)

## 1506500
```

Kwartyle

```
kwantyle_kwota_pozyczki <- quantile(przygotowane_dane$kwota_pozyczki)

##      0%     25%     50%     75%    100%
##   26500  196500  306500  436500 1506500
```

Współczynnik zmienności

```
wsp_zmien_kwota_pozyczki <- odch_stand_kwota_pozyczki/srednia_kwota_pozyczki*
100

## 55.32757
```

Dominanta

```
dominanta_kwota_pozyczki <- Mode(przygotowane_dane$kwota_pozyczki)

## 156500
```

Wariancja

```
war_kwota_pozyczki <- var(przygotowane_dane$kwota_pozyczki)

## 33899233956
```

Wskaźnik asymetrii

```
wskaz_asym_kwota_pozyczki <- skewness(przygotowane_dane$kwota_pozyczki)

## 1.50577
```

Trzeci moment centralny

```
moment_kwota_pozyczki <- moment(przygotowane_dane$kwota_pozyczki, 3)

## 8.004088e+16
```

Podsumowanie dla podstawowych parametrów kwoty pożyczki

Analiza kwoty pożyczki wykazała, że średnia wynosiła około $332,777.1, z odchyleniem standardowym w wysokości $184,117.4, co świadczy o znacznej zmienności wokół średniej. Mediana kwoty pożyczki wynosiła $306,500, co sugeruje, że połowa próby miała niższą kwotę pożyczki, a druga połowa wyższą. Dominanta, czyli najczęściej występująca wartość, wynosiła $156,500. Współczynnik zmienności wyniósł około 55.33%, co sugeruje umiarkowany poziom zmienności w stosunku do średniej wartości. Dodatni wskaźnik asymetrii (około 1.51) wskazuje na skośny rozkład danych w prawo, co oznacza tendencję do większych kwot pożyczek.

Dla cechy “dochód”

Średnia

```
srednia_dochod <- mean(przygotowane_dane$dochod)

## 6910.768
```

Odchylenie standardowe

```
odchyl_stand_dochod <- sd(przygotowane_dane$dochod)

## 5759.665
```

Mediana

```
mediana_dochod <- median(przygotowane_dane$dochod)

## 5700
```

Minimum

```
min_dochod <- min(przygotowane_dane$dochod)

## 0
```

Maksimum

```
max_dochod <- max(przygotowane_dane$dochod)

## 78120
```

Kwantyle

```
kwantyle_dochod <- quantile(przygotowane_dane$dochod)

##    0%   25%   50%   75%  100%
##     0  3780  5700  8400 78120
```

Współczynnik zmienności

```
wspol_zmien_dochod <- odchyl_stand_dochod/srednia_dochod*100

## 83.34333
```

Dominanta

```
dominanta_dochod <- Mode(przygotowane_dane$dochod)

## 4680
```

Wariancja

```
war_dochod <- var(przygotowane_dane$dochod)

## 33173736
```

Wskaźnik asymetrii

```
wskaz_asym_dochod <- skewness(przygotowane_dane$dochod)

## 5.428348
```

Trzeci moment centralny

```
moment_dochod <- moment(przygotowane_dane$dochod, 3)

## 2.052582e+12
```

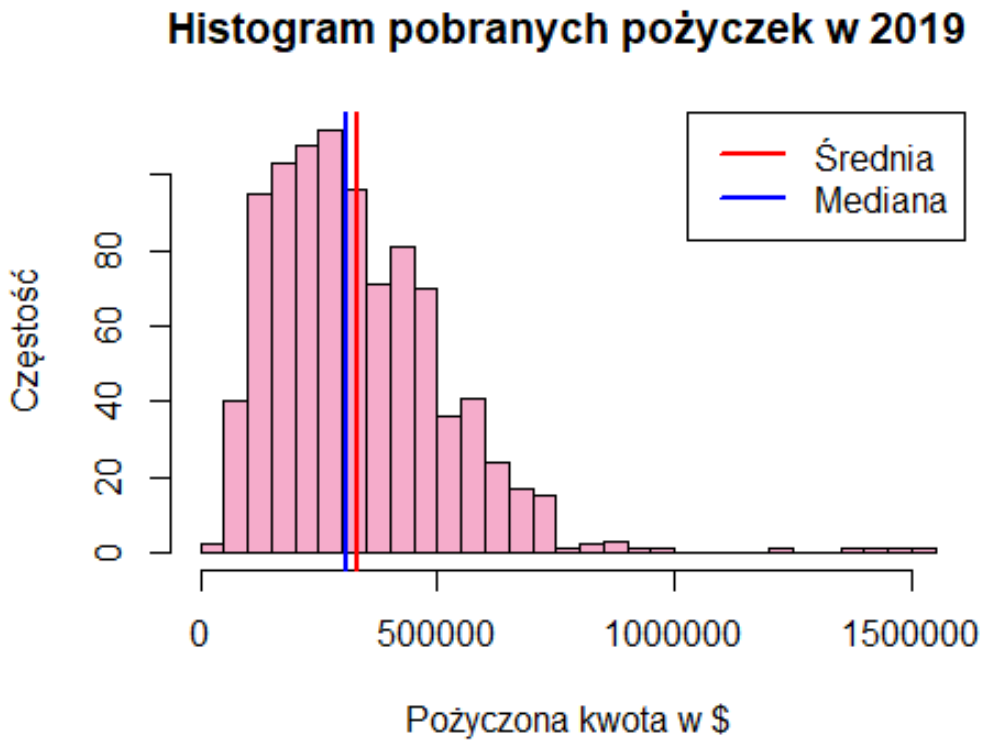
Podsumowanie dla podstawowych parametrów dochodu

Analiza dochodów klientów wykazała, że średni dochód wynosił około $6910.768, z dużym odchyleniem standardowym ($5759.665), co świadczy o znacznej zmienności w dochodach. Mediana dochodu była niższa i wynosiła $5700. Maksymalny dochód wynosił $78120, a minimalny $0, co wskazuje na różnorodność dochodów w badanej grupie. Współczynnik zmienności wynoszący około 83.34% oraz dodatni wskaźnik asymetrii (około 5.43) sugerują dużą zmienność i skośność rozkładu dochodów w prawo.

Graficzna prezentacja danych

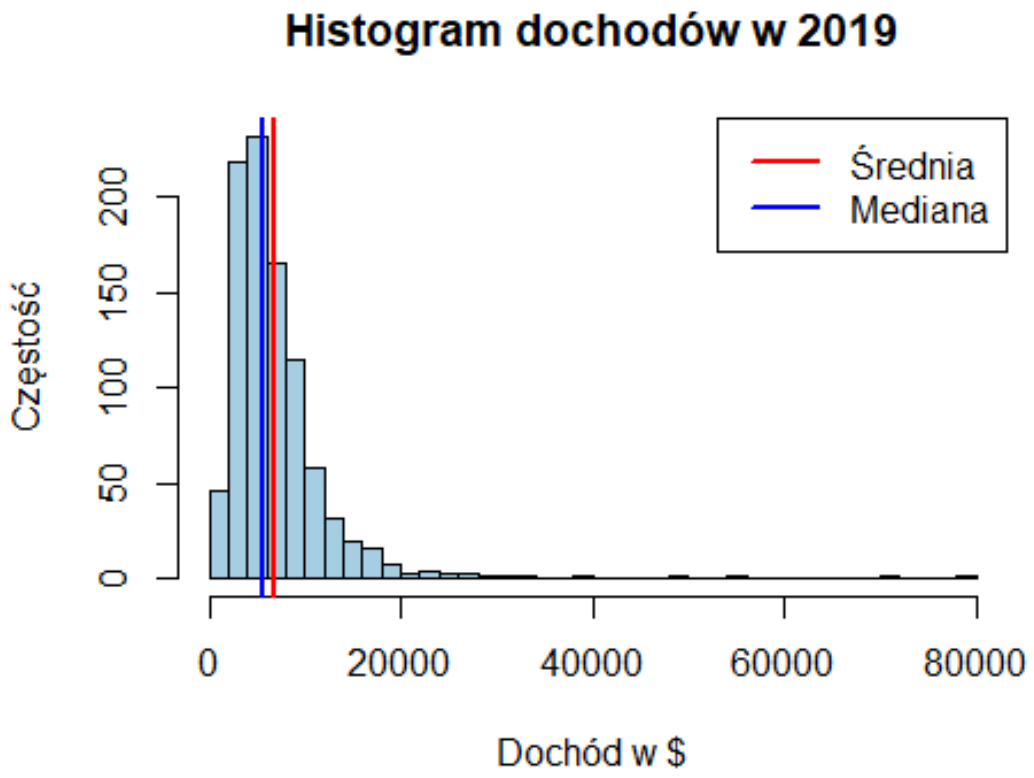
Histogram

```
histogram_kwota_pozyczki <- hist(przygotowane_dane$kwota_pozyczki,
                                 breaks = 50,
                                 main = "Histogram pobranych pożyczek w 2019"
,
                                 xlab = "Pożyczona kwota w $",
                                 ylab = "Częstość",
                                 col = "#F5ACCB")

abline(v = mean(przygotowane_dane$kwota_pozyczki), col = "red", lwd = 2)
abline(v = median(przygotowane_dane$kwota_pozyczki), col = "blue", lwd = 2)
legend("topright", legend = c("Średnia", "Mediana"), col = c("red", "blue"),
lwd = 2)
```



W 2019 roku większość pożyczek była zaciągana na kwoty głównie między 100 000 a 400 000 dolarów. Rozkład jest skośny w prawo, co oznacza, że występuje kilka bardzo wysokich kwot pożyczek. Średnia (czerwona linia) jest wyższa niż mediana (niebieska linia), co wskazuje na wpływ kilku bardzo wysokich pożyczek na średnią wartość. Najwięcej pożyczek zaciągnięto na kwoty około 300 000 dolarów.

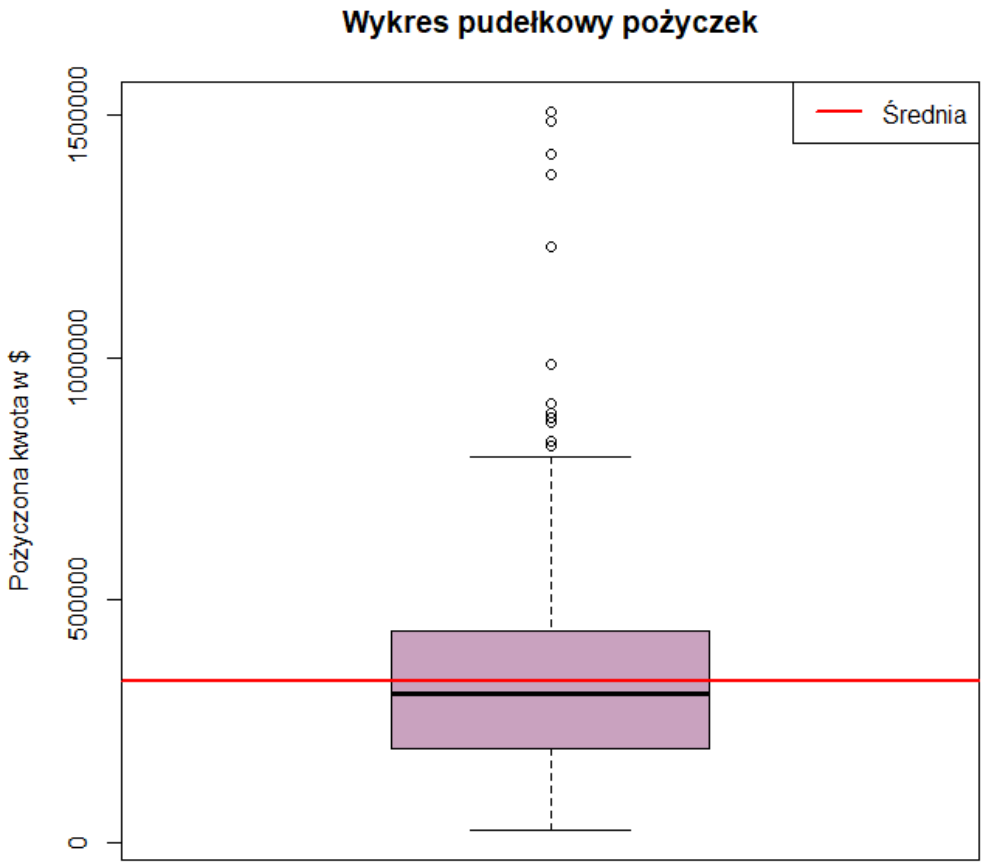
```r
histogram_dochod <- hist(przygotowane_dane$dochod,
                         breaks = 50,
                         main = "Histogram dochodów w 2019",
                         xlab = "Dochód w $",
                         ylab = "Częstość",
                         col = "#A6CEE3")

abline(v = mean(przygotowane_dane$dochod), col = "red", lwd = 2)
abline(v = median(przygotowane_dane$dochod), col = "blue", lwd = 2)
legend("topright", legend = c("Średnia", "Mediana"), col = c("red", "blue"),
lwd = 2)
```



Histogram dochodów w 2019 roku pokazuje, że większość dochodów skoncentrowana jest poniżej 20 000 dolarów. Rozkład jest prawostronnie asymetryczny, z niewielką liczbą wysokich dochodów powyżej 20 000 dolarów. Średnia dochodów (czerwona linia) jest wyższa od mediany (niebieska linia), co oznacza, że wysokie dochody kilku osób podnoszą średnią.

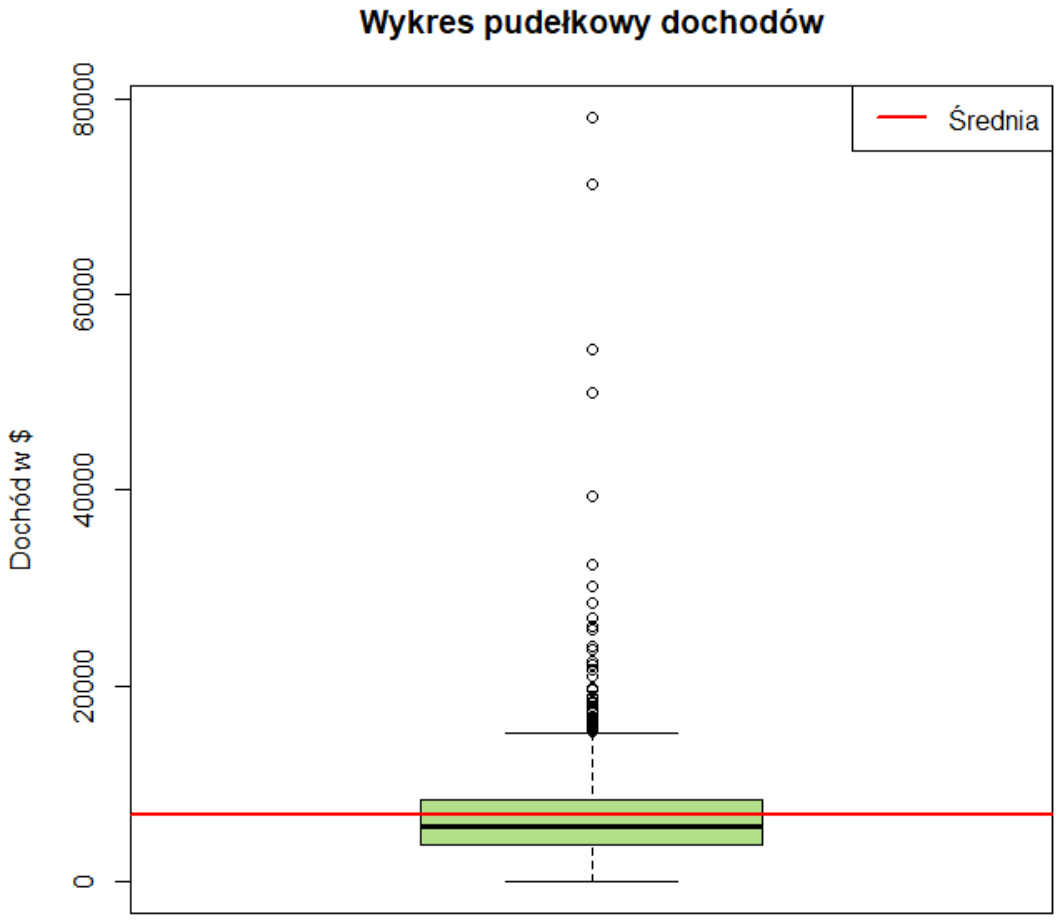
Wykres pudełkowy

```
pudełkowy_kwota_pozyczki <- boxplot(przygotowane_dane$kwota_pozyczki,
                                    main = "Wykres pudełkowy pożyczek",
                                    ylab = "Pożyczona kwota w $",
                                    col = "#C9A2BF",
                                    outline = TRUE)  # Wyświetlanie wartości
odstających

average_value <- mean(przygotowane_dane$kwota_pozyczki)
abline(h = average_value, col = "red", lwd = 2)
legend("topright", legend = c("Średnia"), col = c("red"), lwd = 2)
```



Wykres pudełkowy przedstawia rozkład kwot pożyczek w 2019 roku. Mediana pożyczek jest niższa niż średnia, co wskazuje na prawoskośny rozkład danych. Większość pożyczek mieści się w przedziale od 0 do 500,000 dolarów, ale kilka bardzo dużych pożyczek podnosi średnią kwotę. Widoczne wartości odstające świadczą o obecności wyjątkowo wysokich pożyczek.

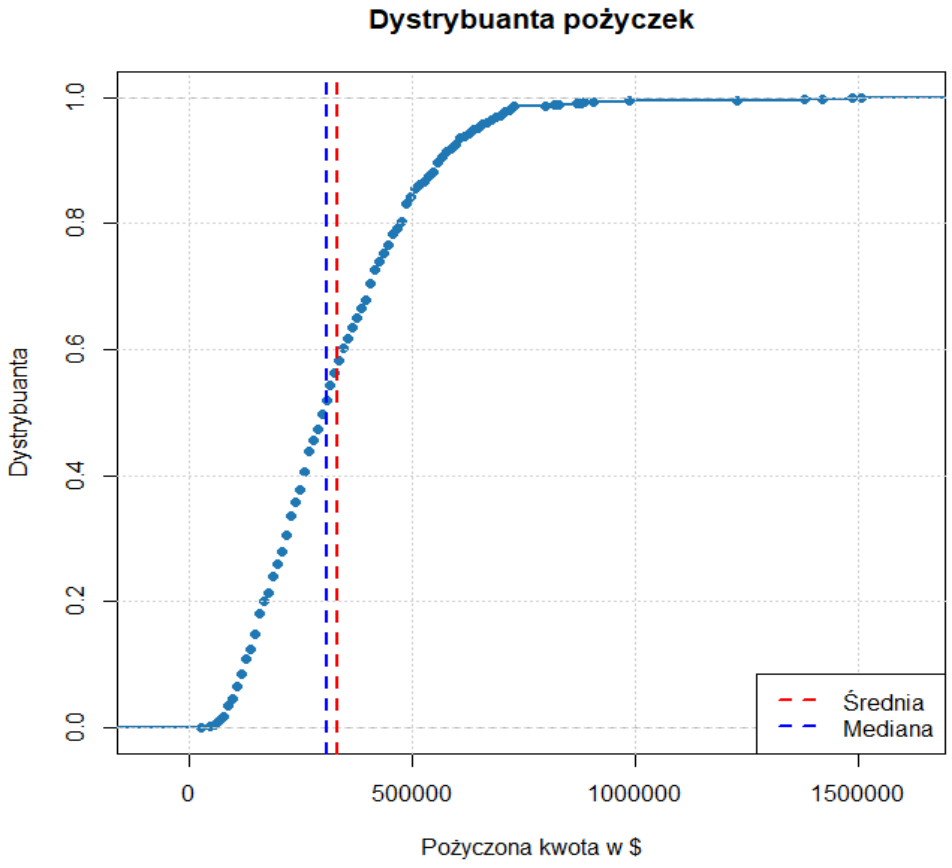
```r
pudełkowy_dochod <- boxplot(przygotowane_dane$dochod,
                            main = "Wykres pudełkowy dochodów",
                            ylab = "Dochód w $",
                            col = "#B2DF8A",
                            outline = TRUE)

average_dochod <- mean(przygotowane_dane$dochod)
abline(h = average_dochod, col = "red", lwd = 2)
legend("topright", legend = c("Średnia"), col = c("red"), lwd = 2)
```



Wykres pudełkowy dochodów pokazuje, że większość dochodów mieści się między około 5 000 a 15 000 dolarów. Mediana dochodów (środkowa linia w pudełku) jest poniżej średniej (czerwona linia), co sugeruje prawostronną asymetrię rozkładu. Widoczna jest również duża liczba wartości odstających powyżej 20 000 dolarów, co wskazuje na obecność kilku bardzo wysokich dochodów.

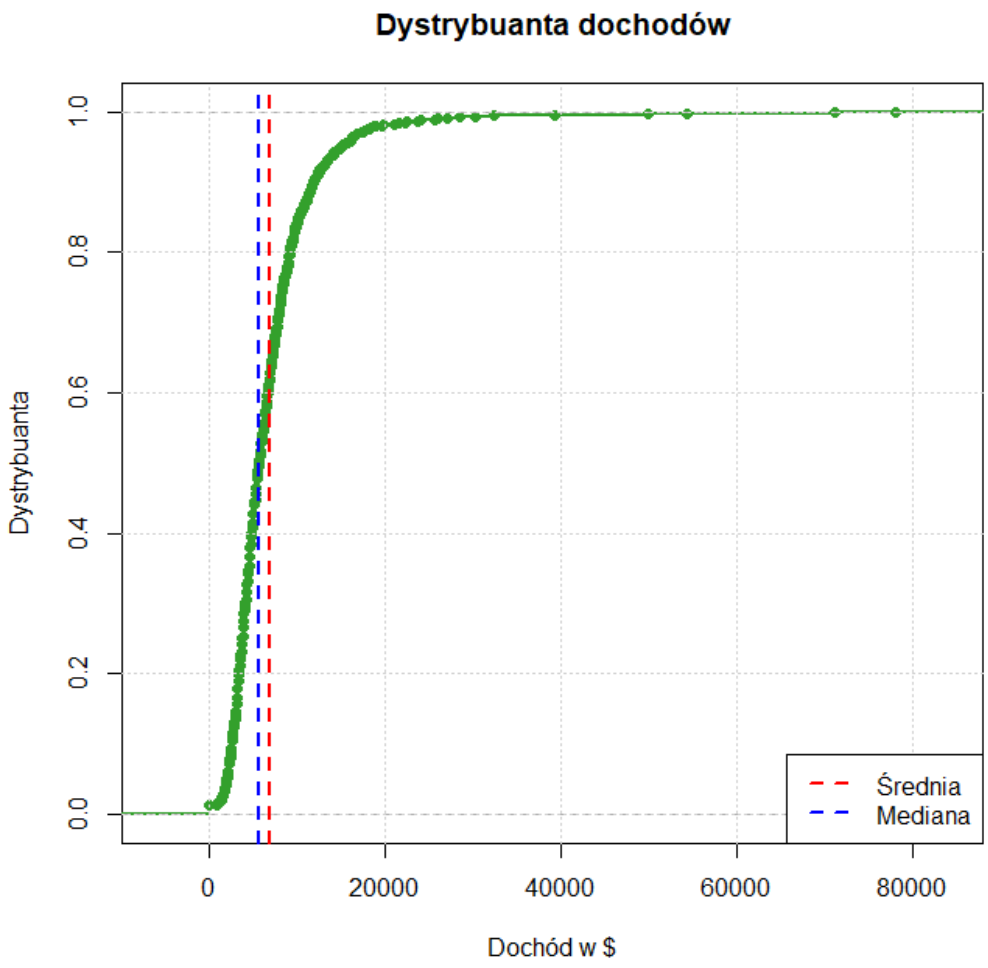
Dystrybuanta

```
dystrybuanta_kwota_pozyczki <- plot(ecdf(przygotowane_dane$kwota_pozyczki),
                                    main = "Dystrybuanta pożyczek",
                                    xlab = "Pożyczona kwota w $",
                                    ylab = "Dystrybuanta",
                                    col = "#1F78B4",
                                    lwd = 2)

grid()
abline(v = mean(przygotowane_dane$kwota_pozyczki), col = "red", lwd = 2, lty
= 2)
abline(v = median(przygotowane_dane$kwota_pozyczki), col = "blue", lwd = 2, l
ty = 2)
legend("bottomright", legend = c("Średnia", "Mediana"), col = c("red", "blue"
), lwd = 2, lty = 2)
```



Wykres przedstawia dystrybuantę kwot pożyczek. Krzywa dystrybuanty pokazuje, że większość pożyczek ma wartości poniżej około 500 tysięcy dolarów. Wartości powyżej tej kwoty są rzadsze, co widać po szybkim wzroście krzywej w początkowym zakresie i jej wypłaszczeniu w dalszej części. Pożyczki mają wyraźnie prawoskośny rozkład, co oznacza, że większość pożyczek jest na stosunkowo niższe kwoty, ale istnieje niewielka liczba pożyczek o bardzo wysokich wartościach, które wpływają na średnią.

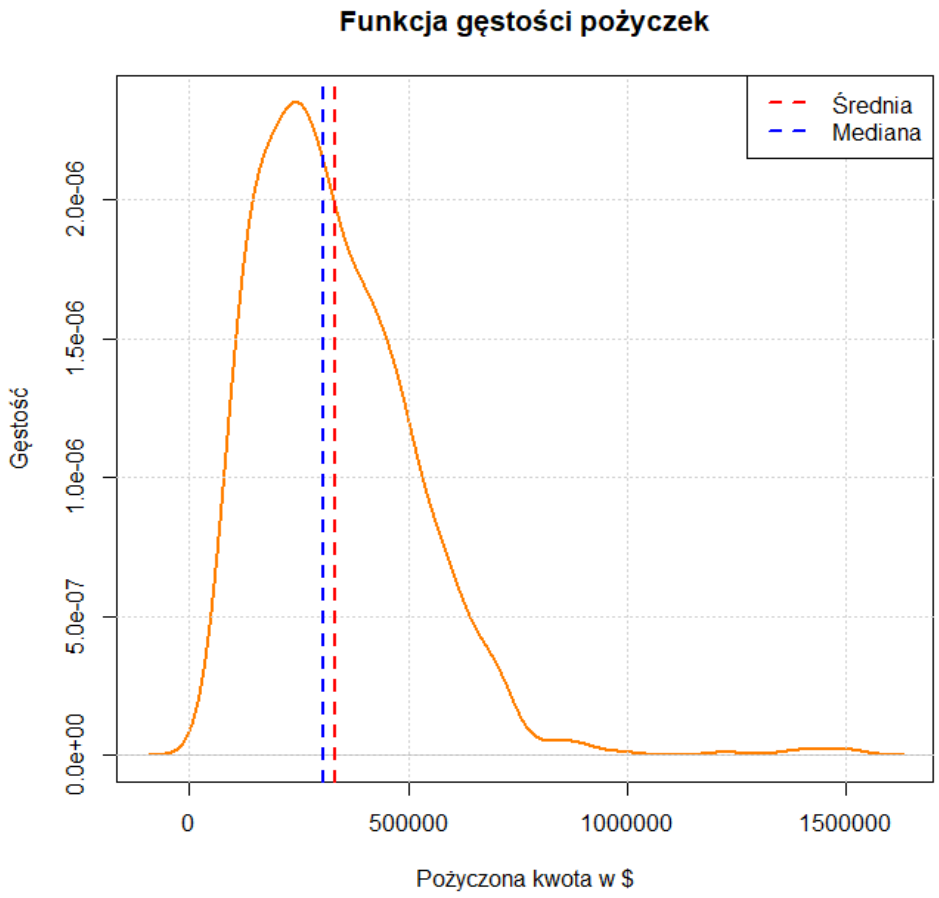
```r
dystrybuanta_dochod <- plot(ecdf(przygotowane_dane$dochod),
                            main = "Dystrybuanta dochodów",
                            xlab = "Dochód w $",
                            ylab = "Dystrybuanta",
                            col = "#33A02C",
                            lwd = 2)

grid()
abline(v = mean(przygotowane_dane$dochod), col = "red", lwd = 2, lty = 2)
abline(v = median(przygotowane_dane$dochod), col = "blue", lwd = 2, lty = 2)
legend("bottomright", legend = c("Średnia", "Mediana"), col = c("red", "blue"
), lwd = 2, lty = 2)
```



Wykres przedstawia dystrybuantę dochodów. Krzywa dystrybuanty pokazuje, że większość dochodów jest skoncentrowana w przedziale do około 20 tysięcy dolarów. Po tym punkcie krzywa zaczyna się wypłaszczać, co sugeruje, że wyższe dochody są rzadsze.
Podsumowując, wykres ukazuje, że typowy dochód (mediana) wynosi około 6 tysięcy dolarów, podczas gdy średni dochód jest nieco wyższy, co wskazuje na obecność kilku wyższych dochodów wpływających na średnią wartość.

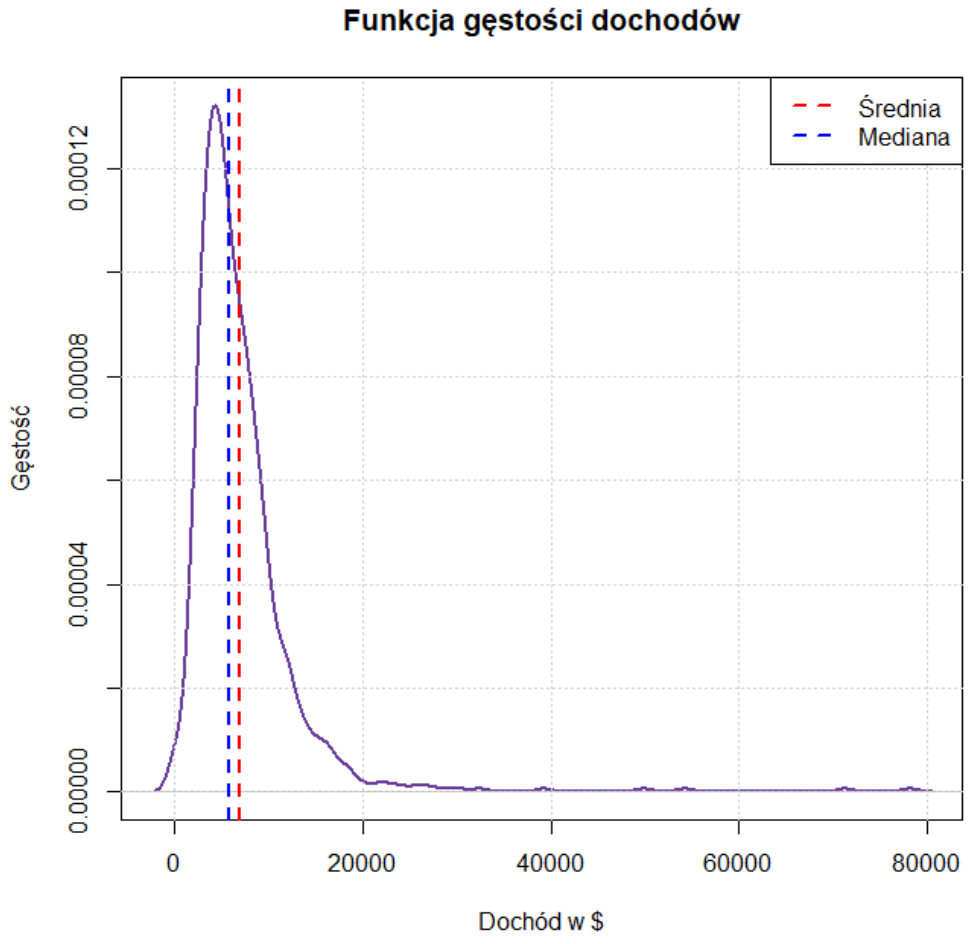
Gęstość

```
gęstość_kwota_pozyczki <- plot(density(przygotowane_dane$kwota_pozyczki),
                               main = "Funkcja gęstości pożyczek",
                               xlab = "Pożyczona kwota w $",
                               ylab = "Gęstość",
                               col = "#FF7F00",
                               lwd = 2)

grid()
average_kwota_pozyczki <- mean(przygotowane_dane$kwota_pozyczki)
median_kwota_pozyczki <- median(przygotowane_dane$kwota_pozyczki)
abline(v = average_kwota_pozyczki, col = "red", lwd = 2, lty = 2)
abline(v = median_kwota_pozyczki, col = "blue", lwd = 2, lty = 2)
legend("topright", legend = c("Średnia", "Mediana"), col = c("red", "blue"),
lwd = 2, lty = 2)
```



Wykres przedstawia funkcję gęstości rozkładu kwot pożyczek. Funkcja gęstości pokazuje, że większość pożyczek jest skoncentrowana w zakresie poniżej 500 tysięcy dolarów, z wyraźnym maksimum około 300 tysięcy dolarów. Rozkład jest silnie prawoskośny, co oznacza, że większa część pożyczek ma niższe wartości, a tylko nieliczne mają bardzo wysokie kwoty.

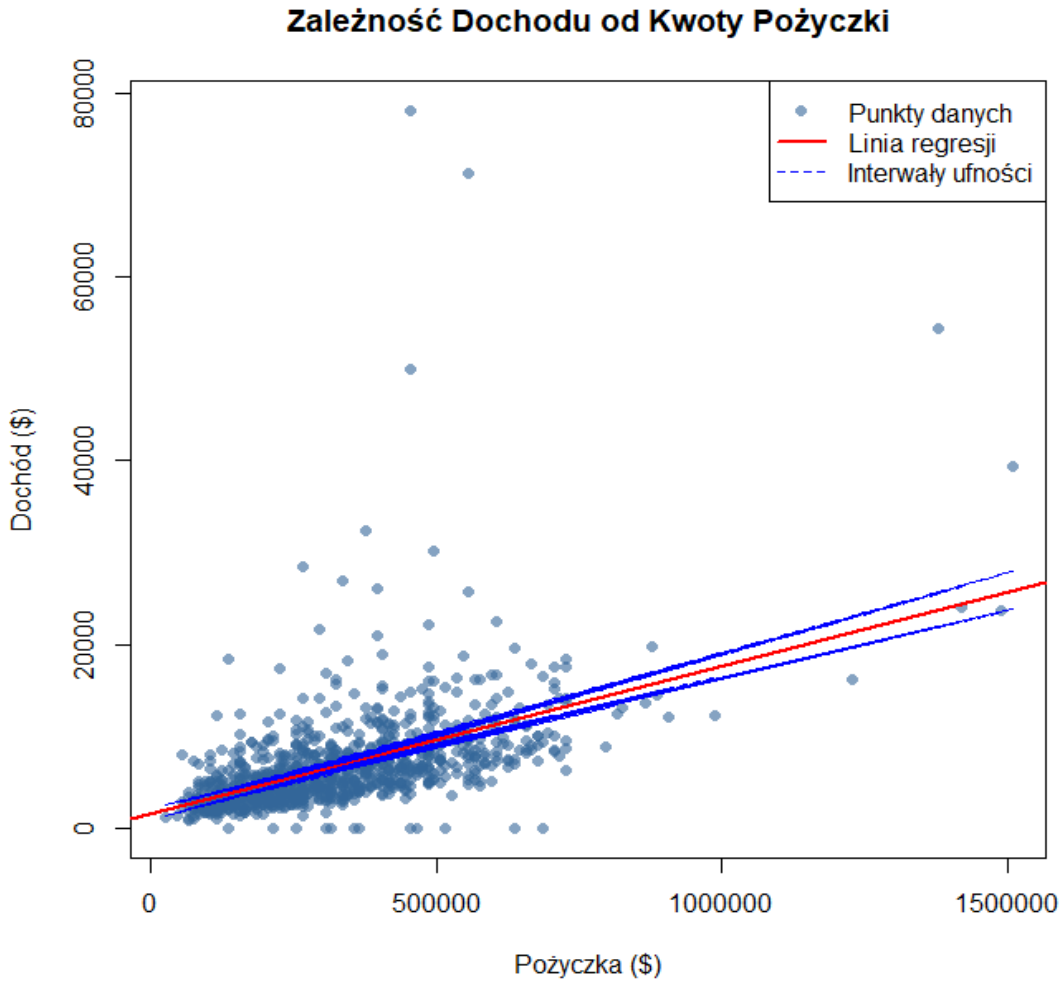
```r
gęstość_dochod <- plot(density(przygotowane_dane$dochod),
                       main = "Funkcja gęstości dochodów",
                       xlab = "Dochód w $",
                       ylab = "Gęstość",
                       col = "#6A3D9A",
                       lwd = 2)

grid()
average_dochod <- mean(przygotowane_dane$dochod)
median_dochod <- median(przygotowane_dane$dochod)
abline(v = average_dochod, col = "red", lwd = 2, lty = 2)
abline(v = median_dochod, col = "blue", lwd = 2, lty = 2)
legend("topright", legend = c("Średnia", "Mediana"), col = c("red", "blue"),
lwd = 2, lty = 2)
```



Wykres przedstawia rozkład dochodów, który jest asymetryczny i prawoskośny. Większość dochodów koncentruje się w niższych przedziałach, z długim ogonem w kierunku wyższych wartości. Średnia dochodów (czerwona linia) jest wyższa od mediany (niebieska linia), co wskazuje na wpływ wysokich dochodów na średnią. Większość dochodów znajduje się poniżej 20 000 dolarów.

Wykres punktowy

```
model <- lm(dochod ~ kwota_pozyczki, data = przygotowane_dane)

plot(przygotowane_dane$kwota_pozyczki, przygotowane_dane$dochod,
     xlab = "Pożyczka ($)", ylab = "Dochód ($)",
     main = "Zależność Dochodu od Kwoty Pożyczki",
     col = rgb(0.2, 0.4, 0.6, 0.6), pch = 16)

abline(model, col = "red", lwd = 2)
predicted_values <- predict(model, interval = "confidence")
lines(przygotowane_dane$kwota_pozyczki, predicted_values[, "lwr"], col = "blu
e", lwd = 1, lty = 2)
lines(przygotowane_dane$kwota_pozyczki, predicted_values[, "upr"], col = "blu
e", lwd = 1, lty = 2)
legend("topright", legend = c("Punkty danych", "Linia regresji", "Interwały u
fności"),
       col = c(rgb(0.2, 0.4, 0.6, 0.6), "red", "blue"), pch = c(16, NA, NA),
lty = c(NA, 1, 2), lwd = c(NA, 2, 1))
```



Wykres przedstawia zależność dochodu od kwoty pożyczki. Widoczna jest dodatnia korelacja między tymi zmiennymi, co oznacza, że wyższy dochód wiąże się z wyższą kwotą pożyczki. Linia regresji (czerwona) i wąskie przedziały ufności (niebieskie linie przerywane) wskazują na stosunkowo dobrą przewidywalność dochodu na podstawie kwoty pożyczki, choć istnieją pewne odchylenia, zwłaszcza dla wyższych wartości.

Podsumowanie dla graficznej prezentacji danych

Analiza wykresów dla dochodów

Analiza wykresów dochodów sugeruje, że większość obserwacji koncentruje się w niższych przedziałach, głównie poniżej 20 000 dolarów, co wskazuje na istnienie pewnej grupy osób zarabiających znacznie mniej niż reszta. Rozkład jest prawostronnie asymetryczny, co oznacza, że istnieje kilka osób z wyższymi dochodami, które znacząco podnoszą ogólną średnią wartość dochodów. Wartość mediany jest niższa od średniej, co sugeruje, że rozkład ten może być mocno wpływany przez kilka osób z bardzo wysokimi dochodami, co oznacza, że jednostki te mogą stanowić mniejszy procent populacji, ale mają znaczący wpływ na ogólny rozkład dochodów.

Analiza wykresów dla kwoty pożyczki

Analiza wykresów kwot pożyczek ujawnia rozpiętość tych kwot, z większością pożyczek mieszczących się w przedziale między około 250 000 a 750 000 dolarów. Rozkład ten również jest prawostronnie asymetryczny, co sugeruje, że istnieje kilka bardzo wysokich pożyczek, które znacznie podnoszą ogólną średnią kwotę pożyczek. Mediana jest niższa od średniej, co wskazuje na obecność wartości odstających w prawo, co może być efektem kilku bardzo wysokich pożyczek. Jednakże, nawet pomimo obecności tych wartości odstających, większość pożyczek skupia się w niższych przedziałach, co może być interpretowane jako istnienie grupy klientów preferujących niższe kwoty pożyczek.

Hipoteza 1: Średnia kwota pożyczki jest równa medianie kwoty pożyczki

Hipoteza zerowa (H0): Średnia kwota pożyczki jest równa medianie pożyczki

Hipoteza alternatywna (H1): Średnia kwota pożyczki nie jest równa medianie pożyczki

```
t.test(przygotowane_dane$kwota_pozyczki, mu = mediana_kwota_pozyczki, conf.le
vel = 0.95)
```

```
##
##  One Sample t-test
##
## data:  przygotowane_dane$kwota_pozyczki
## t = 4.3383, df = 923, p-value = 1.594e-05
## alternative hypothesis: true mean is not equal to 306500
## 95 percent confidence interval:
##  320889.9 344664.2
## sample estimates:
## mean of x
##  332777.1
```

Analiza wyniku testu t-Studenta dla danych dotyczących kwot pożyczek wykazała istotne statystycznie różnice między średnią kwotą pożyczki a wartością referencyjną 306500. Wynik testu ($t = 4.3383$, $df = 923$, $p < 0.05$) sugeruje, że prawdziwa średnia kwoty pożyczki nie jest równa 306500. Przedział ufności dla średniej kwoty pożyczki wynosi od 320889.9 do 344664.2 z 95% pewnością. Średnia kwota pożyczki wynosi 332777.1. Ostatecznie, istnieje istotna różnica między średnią kwotą pożyczki a wartością referencyjną, co potwierdza wynik testu na poziomie istotności alfa = 0.05.

Hipoteza 2: Hipoteza dotycząca współczynnika korelacji

Hipoteza zerowa (H0): Nie ma korelacji między kwotą pożyczki a dochodem klientów.

Hipoteza alternatywna (H1): Istnieje korelacja między kwotą pożyczki a dochodem klientów.

```
cor.test(przygotowane_dane$kwota_pozyczki, przygotowane_dane$dochod, conf.lev
el = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  przygotowane_dane$kwota_pozyczki and przygotowane_dane$dochod
## t = 18.334, df = 922, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4679969 0.5626268
## sample estimates:
##       cor
## 0.5168891
```

Analiza korelacji Pearsona pomiędzy kwotą pożyczki a dochodem klientów wykazała istotne statystycznie dodatnie powiązanie między tymi zmiennymi. Wynik testu (t = 18.334, df = 922, p < 2.2e-16) sugeruje odrzucenie hipotezy zerowej, co potwierdza istnienie korelacji. Wartość współczynnika korelacji wynosi 0.517, co wskazuje na umiarkowaną dodatnią zależność między kwotą pożyczki a dochodem klientów. Przedział ufności dla współczynnika korelacji (0.468 do 0.563) potwierdza istotność tego wyniku z 95% pewnością. Ostatecznie, analiza sugeruje, że osoby z wyższym dochodem mają tendencję do zaciągania wyższych pożyczek, co jest istotne statystycznie.

Hipoteza 3: Średni dochód klientów z wyższą kwotą pożyczki (> 500,000) jest większy niż średni dochód klientów z niższą kwotą pożyczki (<= 500,000)

Hipoteza zerowa (H0): Średni dochód klientów z wyższą kwotą pożyczki jest mniejszy lub równy średniemu dochodowi klientów z niższą kwotą pożyczki.

Hipoteza alternatywna (H1): Średni dochód klientów z wyższą kwotą pożyczki jest większy niż średni dochód klientów z niższą kwotą pożyczki.

```
wysokie_dochody_pozyczka <- przygotowane_dane$dochod[przygotowane_dane$kwota_
pozyczki > 500000]
niskie_dochody_pozyczka <- przygotowane_dane$dochod[przygotowane_dane$kwota_p
ozyczki <= 500000]
t.test(wysokie_dochody_pozyczka, niskie_dochody_pozyczka, alternative = "grea
ter", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  wysokie_dochody_pozyczka and niskie_dochody_pozyczka
## t = 7.7144, df = 165.73, p-value = 5.382e-13
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4114.099      Inf
## sample estimates:
## mean of x mean of y
## 11320.274  6083.278
```

Analiza testu t-Studenta Welch'a na danych dotyczących dochodów klientów z różnymi kwotami pożyczek wykazała istotne statystycznie różnice między grupami. Średni dochód klientów z wyższą kwotą pożyczki wynosił 11320.274, co jest istotnie wyższe od średniego dochodu klientów z niższą kwotą pożyczki (6083.278). Wartość p-value wynosi 5.382e-13, co jest znacznie niższe niż ustalony poziom istotności alfa = 0.05, potwierdzając istotność wyników. Przedział ufności dla różnicy między średnimi dochodami wskazuje, że różnica jest istotna i sięga od 4114.099 do nieskończoności dla wyższych dochodów z 95% pewnością. Ostatecznie można stwierdzić, że istnieje statystycznie istotna różnica w średnich dochodach klientów z różnymi kwotami pożyczek, przy czym średni dochód klientów z wyższą kwotą pożyczki jest istotnie wyższy niż średni dochód klientów z niższą kwotą pożyczki.

Opis użytych funkcji środowiska R:

Funkcja read.csv() służy do odczytywania danych z plików CSV do środowiska R. W powyższym kodzie jest używana do wczytania danych z pliku CSV do zmiennej dane_pozyczki.

Funkcja na.omit() usuwa wiersze zawierające brakujące wartości z danego obiektu danych. Jest stosowana do przetwarzania danych w zmiennej przygotowane_dane, usuwając wiersze zawierające brakujące wartości.

Funkcja mean() oblicza średnią wartość z wektora liczb. Jest wykorzystywana do obliczenia średniej kwoty pożyczki oraz średniego dochodu.

Funkcja sd() oblicza odchylenie standardowe próbki. W kodzie jest używana do obliczenia odchylenia standardowego kwoty pożyczki oraz dochodu.

Funkcja median() oblicza medianę wektora liczb. Jest wykorzystywana do obliczenia mediany kwoty pożyczki oraz mediany dochodu.

Funkcja min() zwraca najmniejszą wartość wektora liczb. Jest używana do znalezienia minimalnej wartości kwoty pożyczki oraz dochodu.

Funkcja max() zwraca największą wartość wektora liczb. Jest wykorzystywana do znalezienia maksymalnej wartości kwoty pożyczki oraz dochodu.

Funkcja quantile() oblicza kwantyle danego wektora danych. Jest stosowana do obliczenia kwantyli kwoty pożyczki.

Funkcja Mode() zwraca dominującą (najczęściej występującą) wartość w wektorze danych. W kodzie jest używana do znalezienia dominującej wartości kwoty pożyczki.

Funkcja var() oblicza wariancję próbki. Jest stosowana do obliczenia wariancji kwoty pożyczki oraz dochodu.

Funkcja skewness() oblicza wskaźnik asymetrii próbki. Jest wykorzystywana do obliczenia wskaźnika asymetrii kwoty pożyczki oraz dochodu.

Funkcja moment() oblicza moment danego rzędu dla próbki. W kodzie jest używana do obliczenia trzeciego momentu centralnego kwoty pożyczki oraz dochodu.

Funkcja hist(): Generuje histogram dla danych numerycznych. Parametry breaks, main, xlab, ylab i col kontrolują odpowiednio liczbę przedziałów, tytuł główny, etykiety osi x i y oraz kolor słupków histogramu.

Funkcja abline(): Dodaje linie na wykresie. W tym przypadku używana jest do dodania linii o wartościach średniej i mediany do histogramu. Parametr v określa poziomą pozycję linii, col ustawia kolor linii, a lwd kontroluje grubość.

Funkcja legend(): Dodaje legendę do wykresu. Parametr legend zawiera etykiety, col określa kolory, a lwd kontroluje grubość linii.

Funkcja boxplot(): Generuje wykres pudełkowy dla danych numerycznych. Parametry main, ylab, col i outline kontrolują odpowiednio tytuł główny, etykiety osi y, kolor pudełek i wyświetlanie wartości odstających.

Funkcja ecdf(): Oblicza dystrybuantę empiryczną dla danych numerycznych.

Funkcja plot(): Rysuje wykres punktowy lub liniowy. W tym kodzie używana jest do rysowania dystrybuanty oraz wykresu punktowego. Parametry main, xlab, ylab, col, lwd i pch kontrolują odpowiednio tytuł główny, etykiety osi x i y, kolor, grubość linii oraz typ punktów.

Funkcja density(): Oblicza funkcję gęstości dla danych numerycznych.

Funkcja lm(): Tworzy model liniowy na podstawie danych. Jest używana do obliczenia regresji liniowej.