Statystyka matematyczna i ekonometria

Projekt zaliczeniowy

Natalia Rusin (252725) i Tomasz Szandała (6169 / 184712)

12.06.2023

Opis bazy danych

Wykorzystywany w projekcie zbiór danych został opubliowany przez Kaggle i dotyczy czynników, wpływających na udzielenie kredytu hipotecznego. Baza danych jest dostępna na licencji Creative Commons 0. Zbiór zawiera trzynaście zmiennych:

- ID kredytu (Loan_ID) – wartości unikatowe
- Płeć (Gender) – zmienna jakościowa binarna
- Czy w związku małżeńskim (Married) – zmienna jakościowa binarna
- Liczba osób na utrzymaniu (Dependents) – zmienna ilościowa skokowa
- Wykształcenie (Education) – zmienna jakościowa binarna
- Samozatrudnienie (Self_Employed) – zmienna jakościowa binarna
- Dochód Aplikanta (ApplicantIncome) – zmienna ilościowa skokowa
- Dochód Współaplikanta (CoapplicantIncome) – zmienna ilościowa skokowa
- Wartość kredytu (tys) (LoanAmount) - zmienna ilościowa skokowa
- Długość kredytu (miesiące) (Loan_Amount_Term) – zmienna ilościowa skokowa
- Historia kredytu (Credit_History) – zmienna jakościowa binarna
- Lokalizacja nieruchomości (Property_Area) – zmienna jakościowa wielodzielna
- Status kredytu (Loan_Status) – zmienna jakościowa binarna

Pracę nad projektem rozpoczęto od wczytania bazy danych.

```
WHERE <- "/home/szandala/pwr/statystyka/"
source(paste(WHERE, "funkcje.r", sep=""))
source(paste(WHERE, "drawings.r", sep=""))
source(paste(WHERE, "outliers.r", sep=""))

loans_data <- read.csv(paste(WHERE, "loan_sanction_train.csv", sep=""))
```

Następnie, przy użyciu funkcji `as.factor` zamieniono część danych na zmienne kategoryczne:

```
loans_data$Gender = as.factor(loans_data$Gender)
loans_data$Married = as.factor(loans_data$Married)
loans_data$Education = as.factor(loans_data$Education)
loans_data$Self_Employed = as.factor(loans_data$Self_Employed)
loans_data$Credit_History = as.factor(loans_data$Credit_History)
loans_data$Property_Area = as.factor(loans_data$Property_Area)
loans_data$Loan_Status = as.factor(loans_data$Loan_Status)
```

Po wykonaniu powyższych czynności możliwe było przystąpienie do dalszej pracy z danymi.

Wyliczenie podstawowych statystyk

Podstawowe statystyki można zaprezentować funkcją summary lub wyliczyć ręcznie.:

```
summary(loans_data)
   Loan_ID             Gender    Married   Dependents        Education
 Length:614              : 13      :  3     : 15     Graduate    :480
 Class :character   Female:112   No :213   0 :345     Not Graduate:134
 Mode  :character   Male  :489   Yes:398   1 :102
                                          2 :101
                                          3+: 51


 Self_Employed ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term
    : 32       Min.   :  150   Min.   :    0     Min.   :  9.0   Min.   : 12
 No :500       1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.0   1st Qu.:360
 Yes: 82       Median : 3812   Median : 1188     Median :128.0   Median :360
               Mean   : 5403   Mean   : 1621     Mean   :146.4   Mean   :342
               3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:168.0   3rd Qu.:360
               Max.   :81000   Max.   :41667     Max.   :700.0   Max.   :480
                                                 NA's   :22      NA's   :14
 Credit_History   Property_Area Loan_Status
 0   : 89       Rural    :179   N:194
 1   :475       Semiurban:233   Y:420
 NA's: 50       Urban    :202
```

W celu wyliczenia podstawowych statystyk stworzono funkcję count_statistics, która jako parametr przyjmuję wybraną kolumnę. Wewnątrz funkcja korzysta z wbudowanych w R funkcji takich jak min(), variance(), median() itp. Otrzymane wyniki są prezentowane w postaci ramki danych.

```
count_statistics <- function(column) {
  c_mean <- mean(column)
  c_median <- median(column)
  c_min <- min(column)
  c_max <- max(column)
  c_quantiles <- quantile(column,  probs = seq(0, 1, 0.25)) # default:
probs = seq(0, 1, 0.25)
  #print(c_quantiles)
  c_mode <- get_domination(column)
  c_variance <- var(column) # populacji, czy próby
  c_sd <- sqrt(c_variance)

  df <- data.frame(Name = c("Średnia_Arytmetyczna",
                            "Wartość_Minimalna","Kwantyl_Dolny",
                            "Mediana",
                            "Kwantyl_Górny", "Wartość_Maxymalna",
                            "Dominanta", "Wariancja",
"Odchylenie_Standardowe"),
                   Value = c(c_mean,
                             c_min, c_quantiles[2],
                             c_median,
                             c_quantiles[4], c_max,
                             c_mode, c_variance, c_sd)
  )
  return(df)
}

get_domination <- function(v) {
  uniqv <- unique(v)
```

```
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

Aby zbadać zróżnicowane danych zdecydowano się na wyliczenie dziewięciu podstawowych statystyk takich jak średnia arytmetyczna, wartość minimalna, kwantyl dolny, mediana, kwantyl górny, wartość maksymalna, dominanta, wariancja i odchylenie standardowe. Jako kolumny wybrano dochód aplikanta oraz dochód współaplikanta. W wyniku działania funkcji otrzymano następujące rezultaty:

```
count_statistics(loans_data$ApplicantIncome)
```

	Name	Value
1	Średnia_Arytmetyczna	5403.459
2	Wartość_Minimalna	150.000
3	Kwantyl_Dolny	2877.500
4	Mediana	3812.500
5	Kwantyl_Górny	5795.000
6	Wartość_Maxymalna	81000.000
7	Dominanta	2500.000
8	Wariancja	37320390.167
9	Odchylenie_Standardowe	6109.042

```
count_statistics(loans_data$CoapplicantIncome
```

	Name	Value
1	Średnia_Arytmetyczna	1621.244
2	Wartość_Minimalna	0.000
3	Kwantyl_Dolny	0.000
4	Mediana	1188.500
5	Kwantyl_Górny	2297.250
6	Wartość_Maxymalna	41667.000
7	Dominanta	0.000
8	Wariancja	8562931.806
9	Odchylenie_Standardowe	2926.249

Otrzymane wyniki sugerują, że osoby o wyższych dochodach stają się głównym aplikantem o kredyt hipoteczny co potwierdza zarówno wyższa średnia arytmetyczna jak i mediana. Ponadto, każdy aplikant posiada jakikolwiek dochód, nawet jeśli jest on niewielki, natomiast w przypadku współaplikantów występuje sytuacja gdzie osoba może nie posiadać żadnego dochodu.

Ponadto zdecydowano się także na zbadanie korelacji pomiędzy wymienionymi wyżej zmiennymi, aby zbadać czy istnieje jakiś związek wartościowy między nimi. W tym celu również posłużono się funkcją wbudowaną w R – `cor()`

```
correlation <- cor(loans_data$ApplicantIncome,
loans_data$CoapplicantIncome)
print(correlation)
```

Otrzymany wynik to -0.1166043

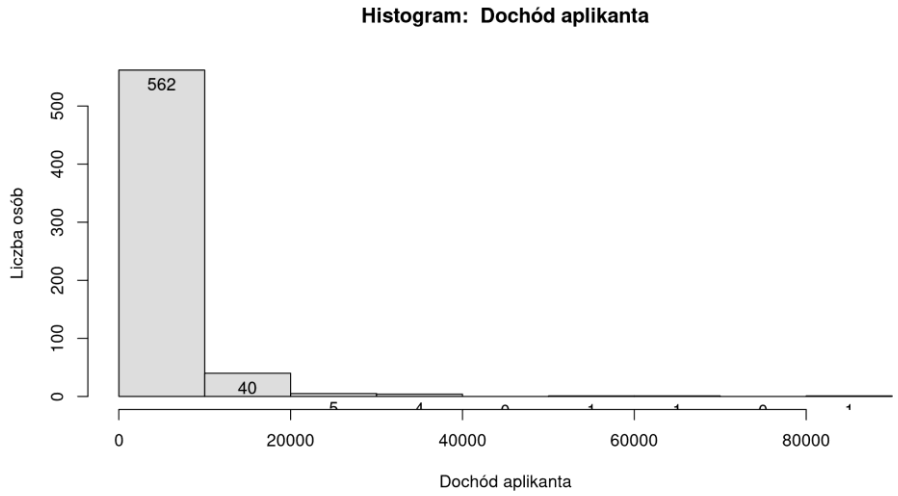
Korelacja o wartości -0.1166 między dwiema zmiennymi (ApplicantIncome i CoapplicantIncome) wskazuje na słabe, lecz ujemne powiązanie między nimi. Ujemna korelacja sugeruje, że w miarę wzrostu jednej zmiennej, druga zmienna ma tendencję malejącą, i vice versa. Jednak w przypadku korelacji o wartości -0.1166, powiązanie to jest stosunkowo słabe. A zatem dla badanych danych, jeżeli dochód głównego aplikanta rośnie to dochód współaplikanta może nieznacząco maleć. Z drugiej strony jeśli dochód współaplikanta wzrasta, to ten głównego aplikanta może być mniejszy.

Wykresy

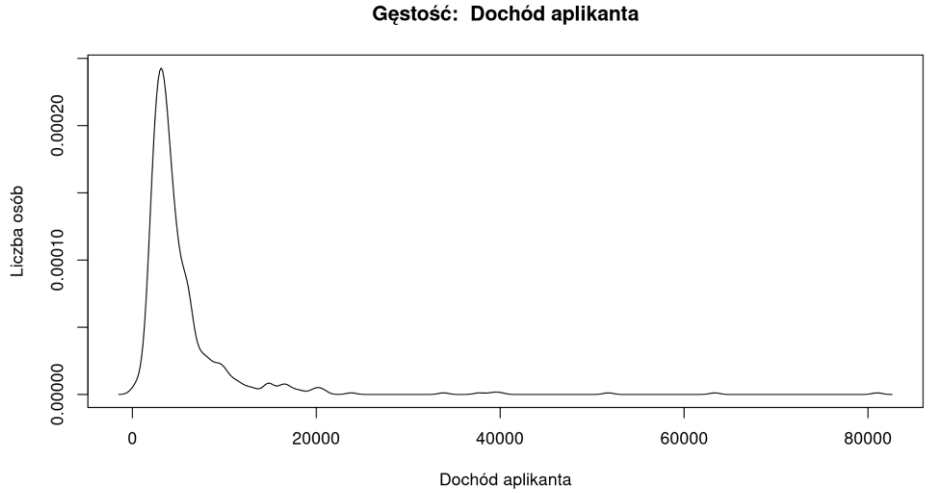
W celu wizualizacji badanych danych wygenerowano takie wykresy jak: histogram, wykres słupkowy, wykres pudełko-wąsy, gęstości. Zrezygnowano z rysowania wykresu liniowego ze względu na rodzaj posiadanych danych. Zamiast tego wyrysowano wykres kołowy. Wszystkie wykresy zostały wyrysowane w oparciu o funkcje dostępne w R.

```
draw_histogram <- function(column, title = "", xlabel="", ylabel = "Liczba
osób") {
  if (title == "")
    title = xlabel
  histogram <- hist(column, main = paste("Histogram: ", title), xlab =
xlabel, ylab = ylabel)

  for (i in 1:length(histogram$counts)) {
    text(histogram$mids[i], histogram$counts[i], labels =
histogram$counts[i], pos = 1)
  }
}

draw_box <- function(column, title = "", xlabel="", ylabel = "Liczba osób")
{
  if (title == "")
    title = xlabel
  boxplot(column, main = paste("Wykres pudełkowy: ", title),
          horizontal = TRUE,
          xlab = xlabel, ylab = ylabel, outline = FALSE)

  stats <- boxplot.stats(column)
  Q1 <- stats$stats[2]
  median <- stats$stats[3]
  Q3 <- stats$stats[4]

  # Create legend box
  legend("topright",
         legend = c(paste("Q1 =", Q1),paste("Median =", median), paste("Q3
=", Q3)),
         border = "black",
         text.col = "black", bty = "l")
}

draw_box2 <- function(column, column2, title = "", xlabel="", ylabel =
"Liczba osób") {
  if (title == "")
    title = xlabel

  combined_data <- c(column, column2)

  # Tworzenie wektorów oznaczających grupy
  group1 <- rep("Dochód Aplikanta", length(column))
  group2 <- rep("Dochód co-Aplikanta", length(column2))
  groups <- c(group1, group2)

  boxplot(combined_data ~ groups, main = paste("Wykres pudełkowy: ",
title),
          horizontal = TRUE,
          xlab = xlabel, ylab = ylabel, outline = FALSE)
}

draw_density <- function(column, title = "", xlabel="", ylabel = "Liczba
osób") {
  if (title == "")
    title = xlabel
  data <- density(column)
  plot(data, main = paste("Gęstość: ", title), xlab = xlabel, ylab =
ylabel)
}
```

```
draw_bar <- function (columnX, columnY, title = "", xlabel="", ylabel =
"Liczba osób") {
  if (title == "")
    title = xlabel

  barplot(height=columnY, names.arg = columnX,
          main = paste("Wykres: ", title), xlab = xlabel, ylab = ylabel)
}

### kod wykresów 2x2
par(mfrow = c(2, 2))
par(mar = c(2, 2, 1, 1))
par(oma = c(0.5, 0.5, 0.5, 0.5))

pie(table(loans_data$Property_Area),  main = "Lokalizacje inwestycji")
pie(table(loans_data$Gender),  main = "Płeć aplikantów")
pie(table(loans_data$Married),  main = "W związku")
pie(table(loans_data$Dependents),  main = "Osoby na utrzymaniu")
```



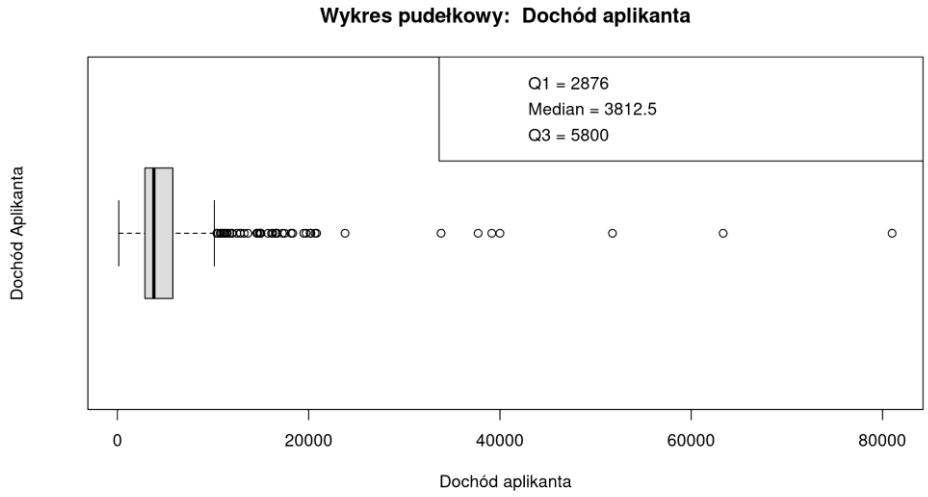
Dominującym przedziałem dochodów aplikantów jest [0,10 000], co jest zgodne z wyliczoną wcześniej wartością dominanty (2500). Z uwagi na fakt, iż większość obserwacji znajduje się we wspomnianym przedziale można stwierdzić iż, zarobki aplikantów nie są zróżnicowane.



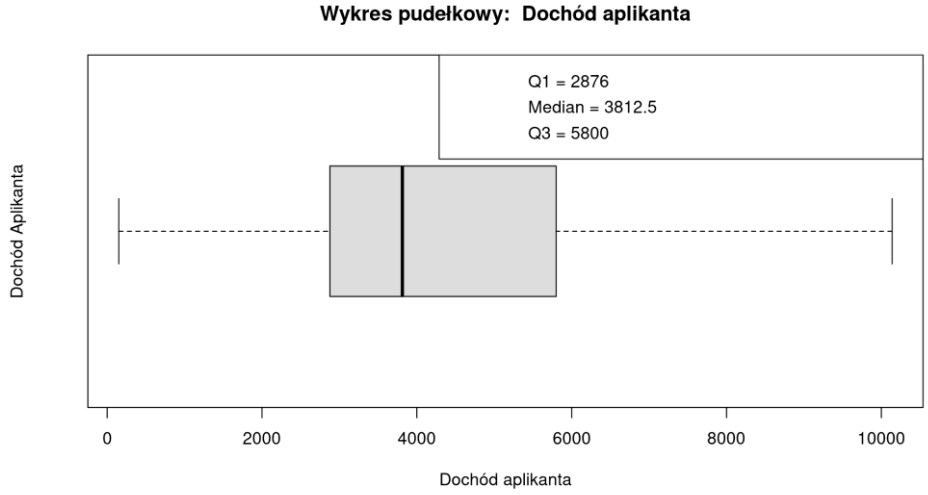
Podobnie jak w przypadku histogramu, można zaobserwować skupienie danych na przedziale [0;10 000]. Ponadto długi ogon wykresu po prawej stronie wskazuje na prawostronną skośność i skupienie większości wartości poniżej średniej. Zatem dane o dochodzie aplikanta odbiegają od rozkładu normalnego. Aby zweryfikować w jakim stopniu dane są asymetryczne wyliczono współczynnik skośności. Skorzystano z dodatkowej biblioteki *moments* dostępnej w R.

```
library(moments)
skew <- skewness(loans_data$ApplicantIncome)
print(skew)
```

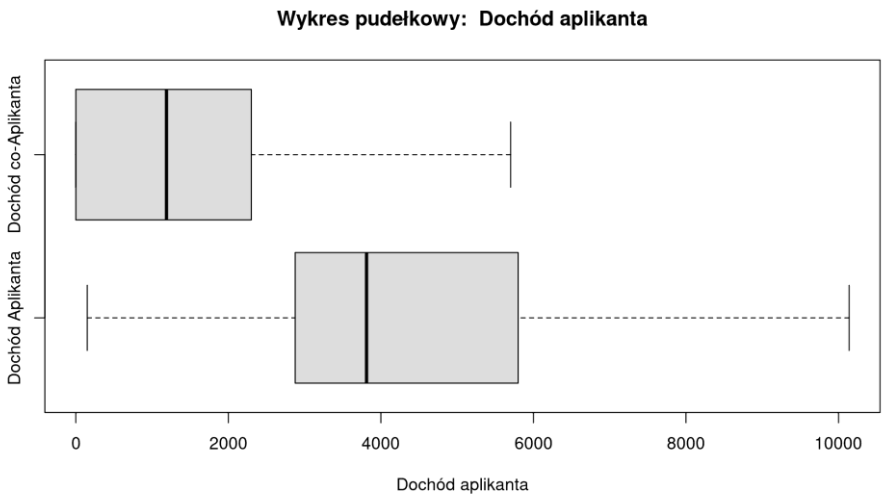
Otrzymana wartość wynosi 6.523526, co świadczy o występowaniu wartości odstających – zarówno bardzo niewielkich wartości dochodu jak i tych wysokich, co znajduje także potwierdzenie w wyliczonych wcześniej wartościach minimalnych i maksymalnych.



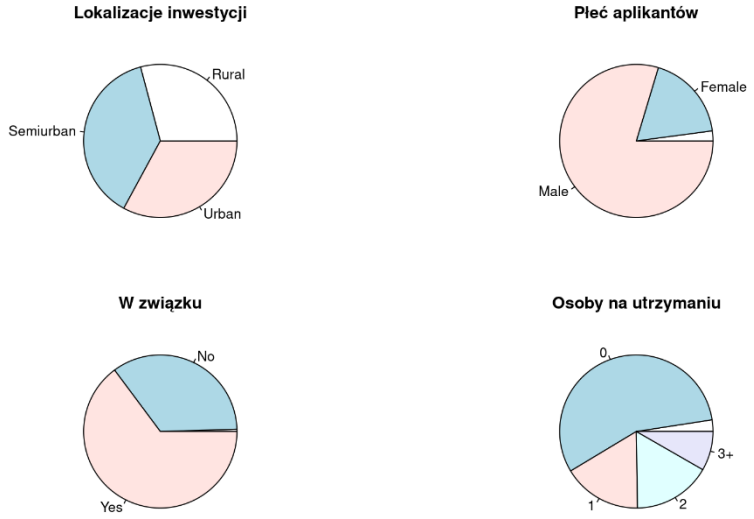
Przedstawiony powyżej wykres pudełkowy bardzo wyraźnie wskazuje na występowanie danych odstających, co potwierdza występowanie wyliczonej wcześniej skośności. W celu poprawy możliwości interpretacji wykresu usunięto wartości odstające, dzięki czemu uzyskano poniższy wykres.



Tak wygenerowany wykres potwierdza prawostronną asymetrię danych, ponieważ oznaczenie mediany jest przesunięte w lewą stronę. Ponadto warto dodać, iż nawet po usunięciu wartości odstających, te pozostawione są rozproszone o czym świadczą długie wąsy po obu stronach pudełka.



Postanowiono także porównać dochód aplikanta i współaplikanta. Dochód współaplikanta jest symetryczny, w związku z czym można stwierdzić, że jest on mniej zróżnicowany niż dochód aplikanta, gdzie można zaobserwować zarówno asymetrię jak i różne długości wąsów. Warto zaznaczyć, że w przypadku tej analizy nie badano danych odstających.



Dla wybranych zmiennych jakościowych wyrysowano wykresy kołowe. Na ich podstawie można wywnioskować, iż w większości o kredyt hipoteczny ubiegają się mężczyźni. Wśród aplikantów więcej osób jest w związku, a także aplikanci deklarują, iż nie posiadają nikogo na swoim utrzymaniu. Wygenerowano także wykres kołowy dla lokalizacji inwestycji, który pokazuje równomierny podział pomiędzy możliwe lokalizacje.

Obserwacje odstające

Przeanalizowano zmienną Dochód aplikanta pod względem występowania obserwacji odstających. Na podstawie przeprowadzonych wcześniej obliczeń wiadomo, że wykorzystywany zbiór danych posiada wartości odstające, natomiast wykorzystane metody, pozwalają stwierdzić, które dokładnie wartości można uznać za odstające. Wykorzystano 4 metody:

- Standard Deviation Method (Metoda odchylenia standardowego) – punkty, które znajdują się powyżej lub poniżej pewnej liczby standardowych odchyleń od średniej, są uznawane za odstające.
- IQR Method (Metoda IQR - Interquartile Range - punkty, które wychodzą poza pewien zakres wyznaczony przez mnożnik IQR (rozstęp międzykwantylowy), są uznawane za odstające.
- Z-Score Method (Metoda wyników standaryzowanych) - wartości, które mają z-score przekraczający pewien próg, są uznawane za odstające, gdzie z-score określa, jak daleko od średniej znajduje się wartość, wyrażając ją w jednostkach odchylenia standardowego
- Tukey's Fences Method (Metoda ogrodzeń Tukeya) - punkty, które znajdują się poza ogrodzeniami, są uznawane za odstające. Ogrodzenie to zakres wyznaczony, na podstawie wartości skrajnych w danych

```
outlierSD <- function(x, nmads = 3) {
  z_scores <- (x - mean(x)) / sd(x)
  outliers <- x[abs(z_scores) > nmads]
  return(outliers)
}
get_outliers <- function(column) {
  sd_outliers <- outlierSD(column, nmads = 2)  # Adjust the number of
standard deviations as desired

  # Interquartile Range (IQR) Method
  boxplot_stats <- boxplot.stats(column)
  Q1 <- boxplot_stats$stats[2]
  Q3 <- boxplot_stats$stats[4]
  IQR <- Q3 - Q1
  k <- 1.5  # Adjust this value as desired
  iqr_outliers <- column[column < Q1 - k * IQR | column > Q3 + k * IQR]

  # Z-Score Method
  threshold <- 2  # Adjust the z-score threshold as desired
  z_scores <- scale(column)
  z_score_outliers <- column[abs(z_scores) > threshold]

  # Tukey's Fences Method
  tukey_outliers <- boxplot_stats$out

  cat("Standard Deviation Method - Outliers:\n", sort(sd_outliers), "\n")
  cat("IQR Method - Outliers:\n", sort(iqr_outliers), "\n")
  cat("Z-Score Method - Outliers:\n", sort(z_score_outliers), "\n")
  cat("Tukey's Fences Method - Outliers:\n", sort(tukey_outliers), "\n")
  }
```

W wyniku działania funkcji otrzymano następujące wartości:

```
get_outliers(loans_data$ApplicantIncome)
```

Standard Deviation Method - Outliers:

18165 18333 19484 19730 20166 20233 20667 20833 23803 33846 37719 39147 39999 51763 63337 81000

IQR Method - Outliers:

10408 10416 10513 10750 10833 11000 11146 11250 11417 11500 11757 12000 12000 12500 12841 12876 13262 13650 14583 14583 14683 14866 14880 14999 15000 15759 16120 16250 16525 16666

16667 16692 17263 17500 18165 18333 19484 19730 20166 20233 20667 20833 23803 33846 37719 39147 39999 51763 63337 81000

Z-Score Method - Outliers:
18165 18333 19484 19730 20166 20233 20667 20833 23803 33846 37719 39147 39999 51763 63337 81000

Tukey's Fences Method - Outliers:
10408 10416 10513 10750 10833 11000 11146 11250 11417 11500 11757 12000 12000 12500 12841 12876 13262 13650 14583 14583 14683 14866 14880 14999 15000 15759 16120 16250 16525 16666 16667 16692 17263 17500 18165 18333 19484 19730 20166 20233 20667 20833 23803 33846 37719 39147 39999 51763 63337 81000

Metody Standard Deviation i Z-Score zwróciły takie same wartości, natomiast IQR oraz Tukey's Fences zwróciły więcej wartości odstających, które również są takie same. Różnica pomiędzy tymi metodami wynika z kryteriów jakie przyjmowane są do wyznaczania wartości odstających.

Budowanie macierzy

Do zbudowania macierzy wykorzystano następujące zmienne: dochód aplikanta, dochód współaplikanta oraz kwotę kredytu. Następnie wyliczono dla macierzy odpowiednie statystyki.

```
describe_matrix <- function(aaa, bbb, ccc, loans_data) {

  library(pastecs)  # Do obliczania skośności i kurtozy
  selected_vars <- c(aaa, bbb, ccc)
  data_matrix <- as.matrix(loans_data[, selected_vars])

  stats <- data.frame(
    Minimum = numeric(3),
    Maximum = numeric(3),
    Median = numeric(3),
    SD = numeric(3),
    Variance = numeric(3),
    Cumulative_Sum = numeric(3),
    Quantile_25 = numeric(3),
    Quantile_75 = numeric(3),
    Skewness = numeric(3),
    Kurtosis = numeric(3),
    Unique_Values = numeric(3),
    Zero_Percentage = numeric(3)
  )
  for (i in 1:3) {
    print(dim(data_matrix))
    var_values <- data_matrix[, i]
    #print(var_values)
    non_na_values <- var_values[!is.na(var_values)]
    #print(non_na_values)
    stats[i, "Minimum"] <- min(non_na_values)
    stats[i, "Maximum"] <- max(non_na_values)
    stats[i, "Median"] <- median(non_na_values)
    stats[i, "SD"] <- sd(non_na_values)
    stats[i, "Variance"] <- var(non_na_values)
```

```
    stats[i, "Cumulative_Sum"] <- sum(non_na_values)
    stats[i, "Quantile_25"] <- quantile(non_na_values, probs = 0.25)
    stats[i, "Quantile_75"] <- quantile(non_na_values, probs = 0.75)
    stats[i, "Skewness"] <- skewness(non_na_values)
    stats[i, "Kurtosis"] <- kurtosis(non_na_values)
    stats[i, "Unique_Values"] <- length(unique(non_na_values))
    stats[i, "Zero_Percentage"] <- sum(var_values == 0, na.rm = TRUE) /
sum(!is.na(var_values)) * 100
  }
  print(stats)
}
describe_matrix("ApplicantIncome", "CoapplicantIncome", "LoanAmount",
loans_data)
```

Otrzymano następujące wyniki:

```
Minimum Maximum Median          SD    Variance Cumulative_Sum Quantile_25
1     150   81000 3812.5 6109.04167 37320390.17        3317724      2877.5
2       0   41667 1188.5 2926.24876  8562931.81         995444         0.0
3       9     700  128.0   85.58733     7325.19          86676       100.0
  Quantile_75 Skewness Kurtosis Unique_Values Zero_Percentage
1     5795.00 6.523526 63.03904           505         0.00000
2     2297.25 7.473215 87.25635           287        44.46254
3      168.00 2.670763 13.30377           203         0.00000
```

Macierz zawiera wszystkie dane, również te odstające. Uwzględniając wszystkie dane okazuje się, że dochód współaplikanta jest zmienną bardziej asymetryczną niż dochód aplikanta. W przypadku przeprowadzonej wcześniej analizy wykresu pudełko-wąsy sytuacja była odwrotna, natomiast nie uwzględniano wtedy danych odstających. Ponadto można stwierdzić, iż wartość kredytu nie jest tak rozproszona jak dochody aplikantów.

Wyliczanie przedziałów ufności

Przedział ufności pozwala zbadać zakres, w którym można oczekiwać, że prawdziwa wartość parametru populacyjnego występuje z określonym prawdopodobieństwem. Dla zmiennej numerycznej Dochód Aplikanta wybrano poziom ufności wynoszący 90%, natomiast dla zmiennej jakościowej wybrano decyzje kredytową, gdzie przyznanie kredytu jest sukcesem. Ten problem jest opisywalny rozkładem dwumianowym. Do przeprowadzenia testu przedziałowego dla rozkładu dwumianowego wykorzystano funkcję `binom.test()`. W tym przypadku przyjęto poziom ufności wynoszący 99%.

```
# numeryczny przedział
dane_numeryczne <- loans_data$ApplicantIncome
wynik_testu <- t.test(dane_numeryczne, na.rm = TRUE,  conf.level = 0.90)
przedzial_ufnosci <- wynik_testu$conf.int
print(przedzial_ufnosci)

# jakościowy przedział
dane_jakosciowe <- loans_data$Loan_Status
sukcesy <- sum(dane_jakosciowe == "Y")  # Liczba sukcesów
proba <- length(dane_jakosciowe)  # Liczba próbek
wynik_testu <- binom.test(sukcesy, proba, conf.level = 0.99)
przedzial_ufnosci <- wynik_testu$conf.int
print(przedzial_ufnosci)
```

Dla zmiennej numerycznej otrzymany przedział ufności to `[4997.322, 5809.597]` a zatem istnieje 90% prawdopodobieństwo, że prawdziwa średnia wartość zmiennej Dochód aplikanta mieści się między 4997 a 5809.

Dla zmiennej jakościowej wynik testu binom.test zwrócił przedział ufności [0.6335889, 0.7315829]. W związku z otrzymanym wynikiem można z 99% pewnością stwierdzić, że prawdopodobieństwo uzyskania kredytu w populacji mieści się między 0.63 a 0.73.

Testowanie hipotez

W obu przypadkach przyjęto poziom ufności wynoszący 95%.

Testy parametryczne

H0: dochód osoby z wykształceniem wyższym jest równy dochodowi osoby bez takiego wykształcenia.

```
# Welch Two Sample t-test

graduate_income <- loans_data[loans_data$Education == "Graduate", "ApplicantIncome"]

non_graduate_income <- loans_data[loans_data$Education == "Not Graduate", "ApplicantIncome"]


t.test(graduate_income, non_graduate_income)
```

Otrzymane wyniki:

```
t = 5.7258, df = 596.88, p-value = 0.00000001632
```

Wniosek: p-value < 0.05 a zatem hipoteza jest odrzucona.

H0: średnio dochód aplikanta jest zbliżony do średnio dochodu koaplikantów

Wykonano 3 testy.

```
# Pearson's product-moment correlation

cor.test(income, coapplicant_income, method = "pearson")
```

Otrzymane wyniki:

```
t = -2.9044, df = 612, p-value = 0.003812
```

```
#Spearman's rank correlation rho
cor.test(income, coapplicant_income, method = "spearman")
```

Otrzymane wyniki:

```
S = 50926632, p-value = 0.0000000000000004318
```

```
#Kendall's rank correlation tau
cor.test(income, coapplicant_income, method = "kendall")
```

Otrzymane wyniki:

```
z = -7.9993, p-value = 0.000000000000001252
```

Wniosek: Dla wszystkich przeprowadzonych testów p-value < 0.05 a zatem hipoteza jest odrzucona.

Testy nieparametryczne

H0: Istnieje zależność między historią kredytową a udzieleniem pożyczki.

```
contingency_table <- table(loans_data$Credit_History, loans_data$Loan_Statu
s)
# Test chi-kwadrat
chisq.test(contingency_table)
```

Otrzymano następujące wyniki:

```
X-squared = 171.56, df = 1, p-value < 0.00000000000000022
```

Wniosek: p-value < 0.05 a zatem hipoteza jest odrzucona.

H0: Istnieje zależność między płcią a udzieleniem pożyczki.

```
# bo są inne płcie
subset_data <- subset(loans_data, Gender %in% c("Male", "Female"))

contingency_table <- table(subset_data$Gender, subset_data$Loan_Status)
contingency_table <- contingency_table[rowSums(contingency_table) != 0, ]


# Test chi-kwadrat
chisq.test(contingency_table)
```

Otrzymane wyniki:

```
X-squared = 0.54719, df = 1, p-value = 0.4595
```

```
# Test Fishera
fisher.test(contingency_table)
```

Otrzymane wyniki:

```
p-value = 0.4299
```

Wniosek: zarówno dla testu chi-kwadrat jaki i testu Fishera wartość p-value jest większa niż 0.05 a zatem nie ma podstaw do odrzucenia hipotezy.

Ponadto warto zwrócić uwagę na fakt, iż uzyskane wartości p-value są różne, a zatem może wystąpić sytuacja gdy w wyniku zastosowania danego testu hipotezę należałoby odrzucić a w wyniku zastosowania innego testu nie byłoby przesłanek do jej odrzucenia.

Regresja liniowa

Dla badanego zbioru danych nie da się obliczyć regresji liniowej.

Analiza głównych składowych (PCA)

Analiza głównych składowych ma na celu wskazanie, które zmienne po kolei prowadzą do zmniejszenia entropii.

```
# Usunięcie kolumny z identyfikatorem
data <- loans_data[, -1]
# Usunięcie wierszy z brakującymi danymi
data <- na.omit(data)
# Wyodrębnienie zmiennych numerycznych do macierzy
numeric_data <- as.matrix(data[, sapply(data, is.numeric)])
# Standaryzacja zmiennych
scaled_data <- scale(numeric_data)
# Obliczenie macierzy kowariancji
cov_matrix <- cov(scaled_data)
# Obliczenie składowych głównych (PCA)
pca <- prcomp(scaled_data)
# Wyświetlenie wyników
print(pca)
```

Wyniki analizy składowych głównych dla podanych danych są następujące:

```
                          PC1         PC2          PC3          PC4
ApplicantIncome   -0.70152950 -0.2408230  0.04502801  0.66920337
CoapplicantIncome -0.04691879  0.9050825 -0.30091834  0.29677000
LoanAmount        -0.70945958  0.1979208  0.04002080 -0.67519835
Loan_Amount_Term   0.04818821  0.2892358  0.95174522  0.09056271
```

PC1: Składowa główna PC1 jest najbardziej wyjaśniającą zmienną, która ma duże ujemne obciążenie dla ApplicantIncome, LoanAmount i Loan_Amount_Term. Oznacza to, że te zmienne są silnie skorelowane ze sobą w kontekście PC1.

PC2: Składowa główna PC2 ma duże dodatnie obciążenie dla CoapplicantIncome. To sugeruje, że ta zmienna ma silny wpływ na PC2. Można interpretować PC2 jako miarę zależności między dochodem współaplikanta a resztą danych.

PC3: Składowa główna PC3 ma duże dodatnie obciążenie dla Loan_Amount_Term. Oznacza to, że PC3 odzwierciedla zmienność w liczbie miesięcy trwania kredytu. Im większa wartość tej składowej, tym większe wahania w długości kredytu.

PC4: Składowa główna PC4 ma ujemne obciążenie dla LoanAmount. Oznacza to, że ta zmienna jest negatywnie skorelowana z PC4. Można interpretować PC4 jako miarę wielkości pożyczki.