Zadanie zaliczeniowe Statystyka 24/25L

Maksymilian Kulicki

2024-12-27

Contents

1. Opis Danych

Dane do zadania zostały pobrane ze strony https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval?select=Loan.csv. Zostały one wygenerowane sztucznie przez skrypt napisany przez autora.Są to dane finansowe, używane do celów edukacyjnych w Data Sciencie. Dane zawierają podstawowe informacje dotyczące osób ubiegających się o kredyt. W ostatnich dwóch kolumnach podana jest informacja czy dana osoba uzyskała kredyt oraz wskaźnik ryzyka związany z tą osobą. Popatrzmy na dane :

```
dataSet <- read.csv("C:/Users/Maksym/Downloads/Loan.csv")

dataSet[1:7,1:11]
```

```
##   ApplicationDate Age AnnualIncome CreditScore EmploymentStatus EducationLevel
## 1      2018-01-01  45        39948         617         Employed         Master
## 2      2018-01-02  38        39709         628         Employed      Associate
## 3      2018-01-03  47        40724         570         Employed       Bachelor
## 4      2018-01-04  58        69084         545         Employed    High School
## 5      2018-01-05  37       103264         594         Employed      Associate
## 6      2018-01-06  37       178310         626    Self-Employed         Master
## 7      2018-01-07  58        51250         564         Employed    High School
##   Experience LoanAmount LoanDuration MaritalStatus NumberOfDependents
## 1         22      13152           48       Married                 2
## 2         15      26045           48        Single                 1
## 3         26      17627           36       Married                 2
## 4         34      37898           96        Single                 1
## 5         17       9184           36       Married                 1
## 6         16      15433           72       Married                 0
## 7         39      12741           48       Married                 0
```

```
dim(dataSet)
```

```
## [1] 20000    36
```

```
any(is.na(dataSet))
```

```
## [1] FALSE
```

Dane są dość spore. Zawierają 20000 obserwacji oraz 36 cech i żadnych brakujących wartości. W raporcie zajmiemy się tylko niektórymi z nich. Przyjrzyjmy się cechom, które będziemy badać w dalszej części raportu. Są to kolumny : “Age”, “Annual Income”, “CreditScore”, “EducationLevel”, “BaseInterestRate”, “InterestRate,”LoanAmount”. Wybrałem te zmienne ze względu na najciekawsze właściwości statystyczne.

Informacje o zmiennych ilościowych

Popatrzmy na podstawowe statystyki danych ilościowych.

```
columns <- c("Age", "AnnualIncome", "CreditScore",
"BaseInterestRate", "InterestRate", "LoanAmount")


for (i in columns){
  cat("\nSummary for:", i, "\n")
  print(summary(dataSet[[i]]))
  cat("Standard Deviation:",sd(dataSet[[i]]), "\n")
}
```

```
##
## Summary for: Age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   32.00   40.00   39.75   48.00   80.00
## Standard Deviation: 11.62271
##
## Summary for: AnnualIncome
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15000   31679   48566   59162   74391  485341
## Standard Deviation: 40350.85
##
## Summary for: CreditScore
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   343.0   540.0   578.0   571.6   609.0   712.0
## Standard Deviation: 50.99736
##
## Summary for: BaseInterestRate
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1301  0.2139  0.2362  0.2391  0.2615  0.4050
## Standard Deviation: 0.03550949
##
## Summary for: InterestRate
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1133  0.2091  0.2354  0.2391  0.2655  0.4468
## Standard Deviation: 0.04220519
##
## Summary for: LoanAmount
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3674   15575   21915   24883   30835  184732
## Standard Deviation: 13427.42
```

Powyżej mamy informacje o maksimach, minimach, średnich, medianach, kwantylach oraz o odchyleniu standardowym poszczególnych zmiennych ilościowych.

Informacje o zmiennych jakościowych

Popatrzmy również na liczności poszczególnych grup zmiennych jakościowych.

```
table(dataSet$EducationLevel)
```

```
##
##   Associate    Bachelor   Doctorate High School      Master
##        4034        6054         954        5908        3050
```

```
table(dataSet$EmploymentStatus)
```

```
##
##      Employed Self-Employed    Unemployed
##         17036          1573          1391
```

Pierwsza zmienna jakościowa pokazuje poziom edukacji poszczególnych osób. Grupy co do rzędu są równoliczne, poza grupą osób z Doktoratem.

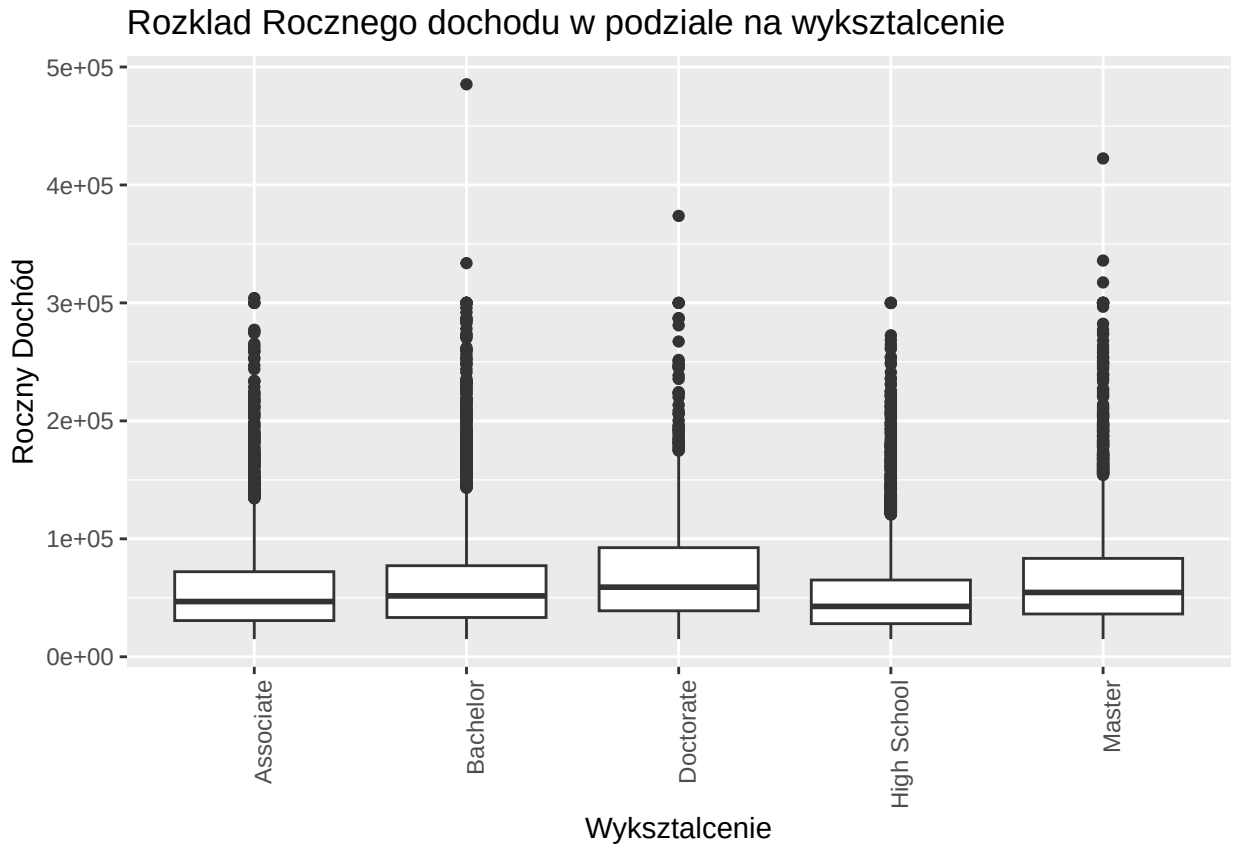
Druga zmienna jakościowa dzieli osoby ze względu na status zatrudnienia. Najliczenijsza grupa to osoby zatrudnione.

2. Analiza zmiennej Annual Income w podziale na zmienną Age

Popatrzmy jak wygląda wykres zmiennej AnnualIncome w podziale na zmienną Age na wykresie pudełkowym.

Wykres pudełkowy rozkładu w podziale na wykształcenie nr 1

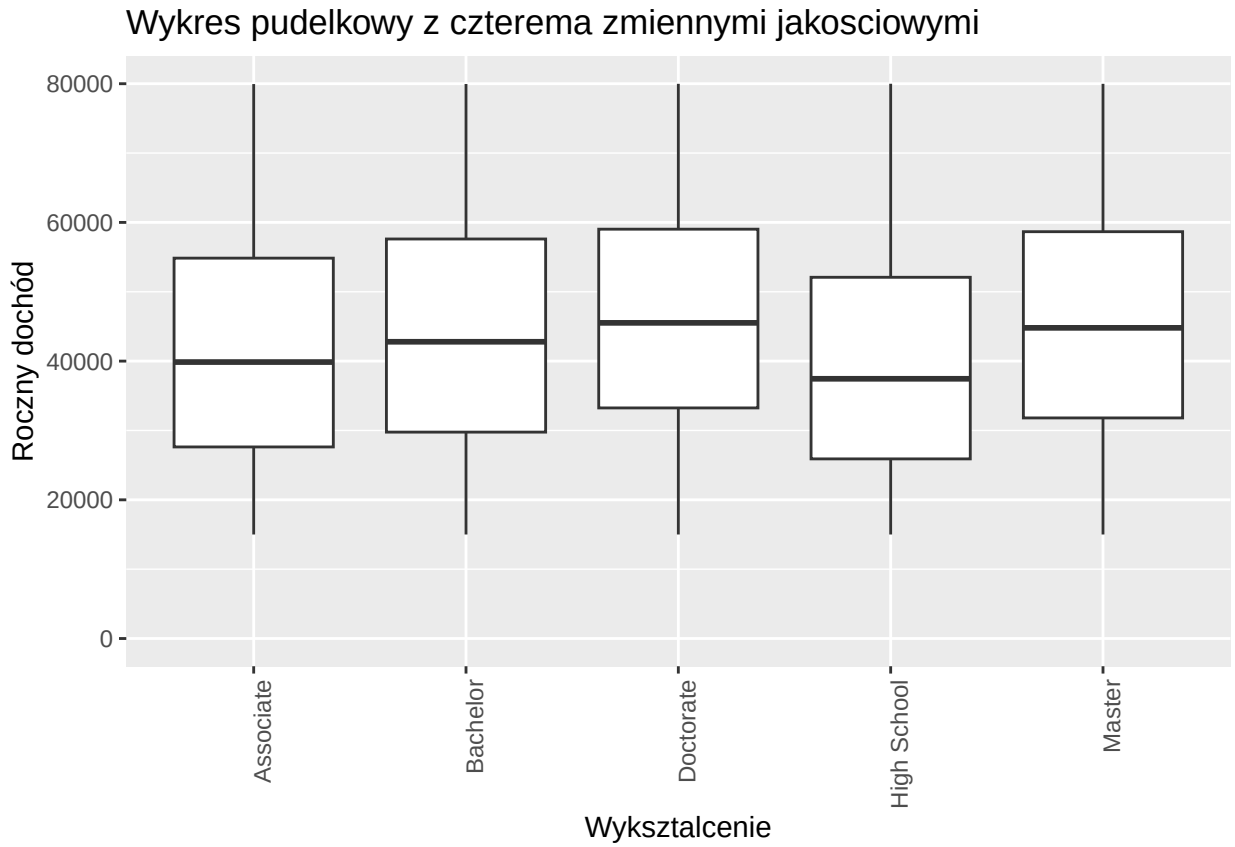
```
ggplot(dataSet, aes(x = EducationLevel, y = AnnualIncome)) +
  geom_boxplot() +
  labs(title = "Rozkład Rocznego dochodu w podziale na wykształcenie",
       x = "Wykształcenie", y = "Roczny Dochód") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Spróbujmy zmiejszyć zakres zmiennej y.

Wykres pudełkowy rozkładu w podziale na wykształcenie nr 2

```
ggplot(dataSet, aes(x = EducationLevel, y = AnnualIncome)) +
geom_boxplot() +
labs(title = "Wykres pudełkowy z czterema zmiennymi jakościowymi",
     x = "Wykształcenie", y = "Roczny dochód") +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
ylim(0,80000)
```



Można by podejrzewać, że roczny dochód będzie większy razem z lepszym wykształceniem. Rzeczywiście widać nieznaczną różnicę pomiędzy ludźmi, którzy skończyli liceum, a ludźmi po magisterium i doktoracie. Różnica ta wynosi średnio około 8000.

Pomimo tej różnicy rozkłady są zbliżone. Jest to zapewne spowodowane wygenerowaniem danych sztucznie.

Tabela statystyk w podziale na wykształcenie

Popatrzmy na tabelkę z statystykami w podziale na wykształcenie:

```
tabela_stat <- dataSet %>%
  group_by(EducationLevel) %>%
  summarise(
    srednia = mean(AnnualIncome),
    mediana = median(AnnualIncome),
    minimum = min(AnnualIncome),
    maksimum = max(AnnualIncome),
    odchylenie_std = sd(AnnualIncome)
  )
kable(tabela_stat, caption = "Statystyki opisowe w podziale na wykształcenie")
```

Table 1: Statystyki opisowe w podziale na wykształcenie

EducationLevel	srednia	mediana	minimum	maksimum	odchylenie_std
Associate	56973.19	46834.0	15000	304122	38343.81
Bachelor	61678.47	51626.0	15000	485341	41136.20
Doctorate	72682.68	59025.0	15000	373724	48478.26
High School	52103.01	42721.5	15000	300000	35418.93
Master	66503.06	54543.5	15000	422480	44541.42

Z powyższej tabelki możemy wywnioskować nieznaczną ale istniejącą korelacje pomiędzy lepszym wykształceniem, a większym rocznym dochodem. Poza tym warto zauważyć istnienie swojego rodzaje wynagrodzenia minimalnego wynoszącego 15000.

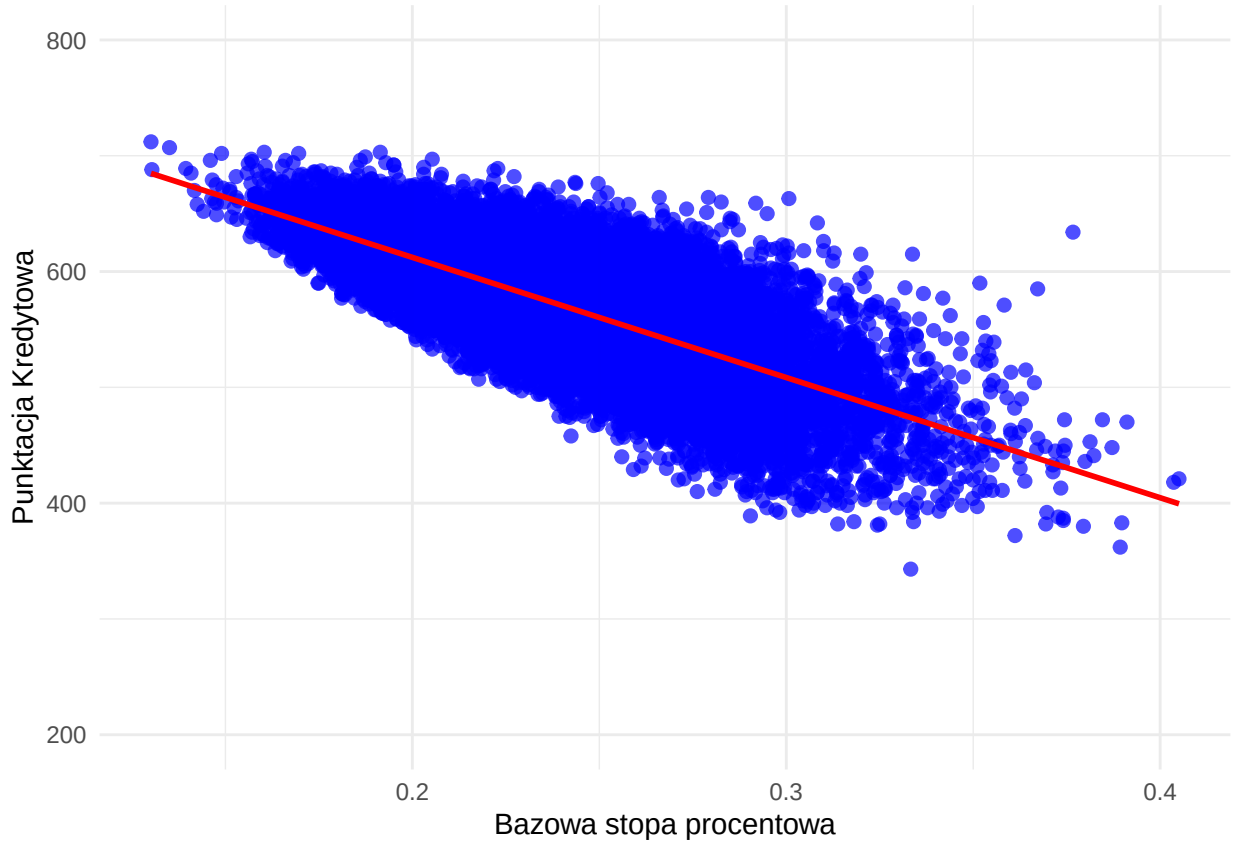
3. Zależność liniowa pomiędzy zmienną BaseInterestRate, a CreditScore

Popatrzmy na wykres zmiennej BaseInterestRate, a zmienną CreditScore:

Wykres

```
ggplot(dataSet,aes(x = BaseInterestRate,y = CreditScore)) +
  geom_point(color = "blue", size = 2, alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = FALSE, linewidth = 1) +
  theme_minimal() +
  ylim(200, 800) +
  labs(
    y = "Punktacja Kredytowa",
    x = "Bazowa stopa procentowa"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Widzimy, że odsetki, które kredytobiorca będzie płacił zależą od punktów kredytowych, czego można się było spodziewać. Nie jest to zależność perfekcyjna, ponieważ dane są rozstrzelone, ale widać korelacje

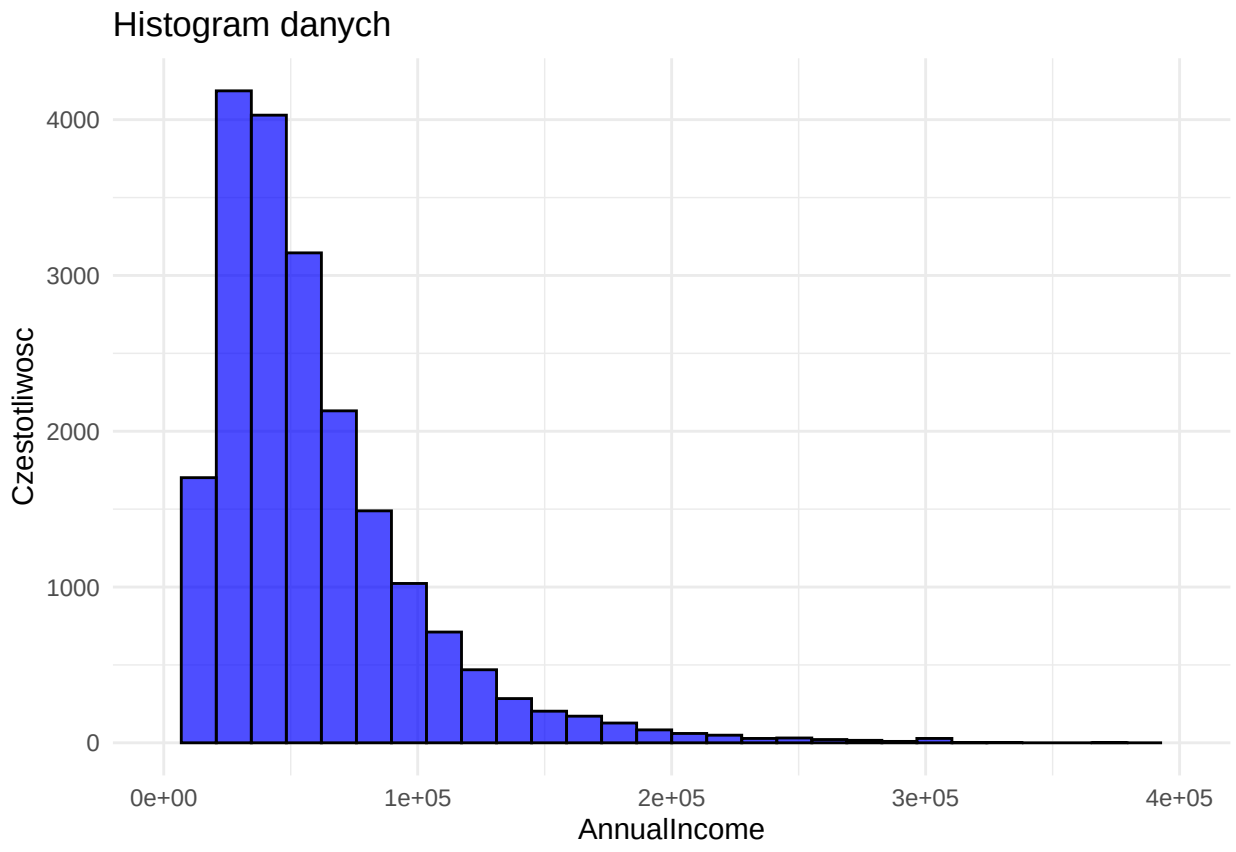
ujemną między zmiennymi.Napewno jest to zależność liniowa i nie potrzebujemy dopasowywać wielomianów wyższego stopnia.

4. Dopasowanie rozkładu dla zmiennej AnnualIncome

Wizualizacja rozkładu danych empirycznych

Przyjrzyjmy się histogramowi naszej zmiennej.

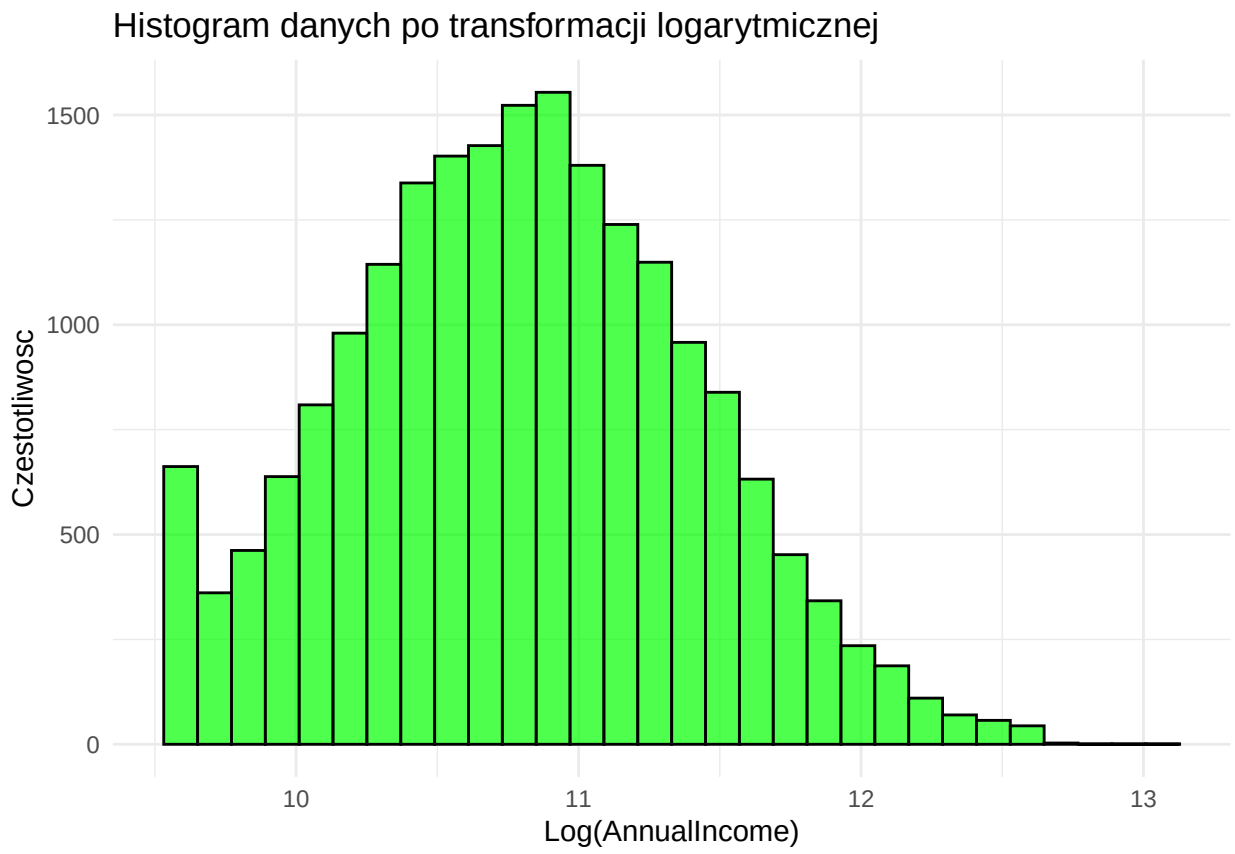
```
ggplot(dataSet, aes(x = dataSet$AnnualIncome)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7, color = "black") +
  labs(title = "Histogram danych", x = "AnnualIncome", y = "Częstotliwość") +
  theme_minimal() +
  xlim(0, 400000)
```



Spróbujmy użyć transformacji naszego rozkładu, logarytmu z rozkładu.

```
log_AnnualIncome <- log(dataSet$AnnualIncome)


ggplot(dataSet, aes(x = log_AnnualIncome)) +
  geom_histogram(bins = 30, fill = "green", alpha = 0.7, color = "black") +
  labs(title = "Histogram danych po transformacji logarytmicznej", x = "Log(AnnualIncome)", y = "Częst
  theme_minimal()
```



Na pierwszy rzut oka logarytm z naszego rozkładu wygląda na rozkład normalny. Przeprowadźmy dalszą analizę statystyczną, aby to potwierdzić lub zaprzeczyć.

Analiza rozkładu - wykresy

Na początku dopasujmy parametry. Niezależnie czy wybierzemy metodę momentów, czy metodę największej wiarygodności otrzymamy ten sam wynik dla rozkładu normalnego :

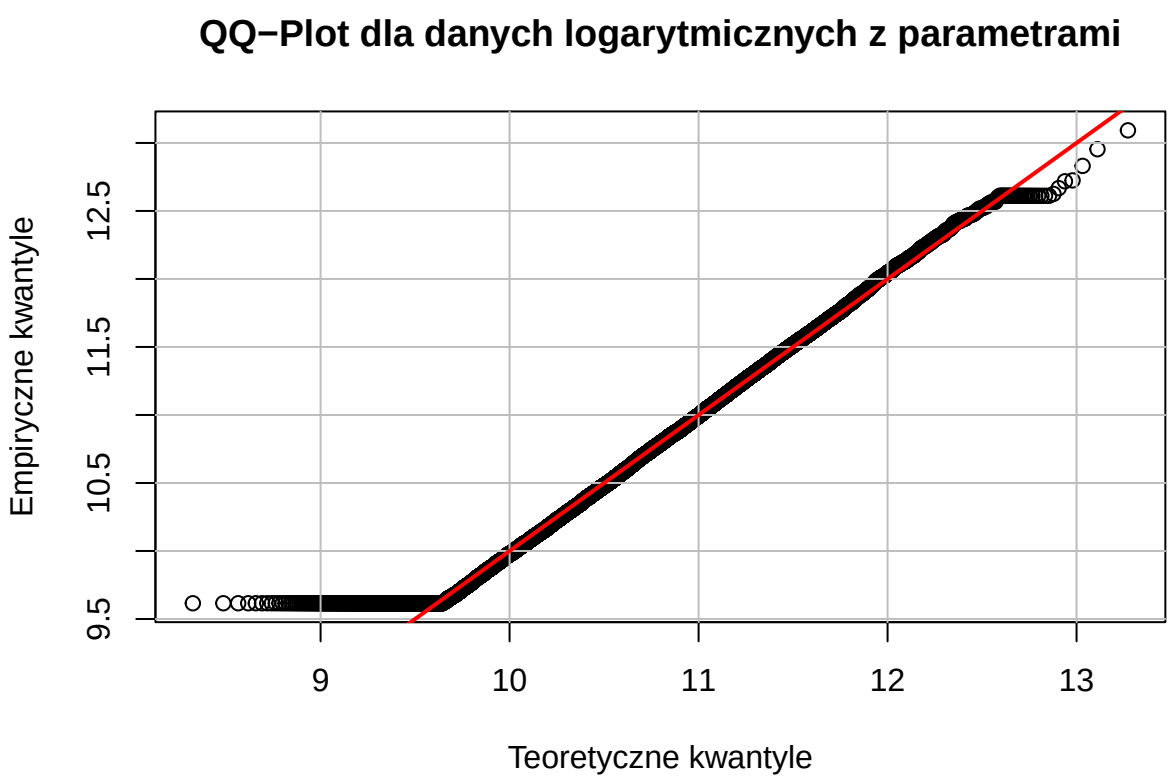
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

, gdzie X_i to pojedyncza próbka, mu to wartość oczekiwana, a sigma^2 to wariancja.

Teraz narysujmy wykres kwantylowy dla dopasowania

```
qqplot(qnorm(ppoints(length(log_AnnualIncome)), mean = 10.798, sd = 0.610),
       log_AnnualIncome,
       main = "QQ-Plot dla danych logarytmicznych z parametrami",
       xlab = "Teoretyczne kwantyle", ylab = "Empiryczne kwantyle")
grid(nx = NULL, ny = NULL, col = "gray", lty = "solid", lwd = 1)
abline(0, 1, col = "red", lwd = 2)
```

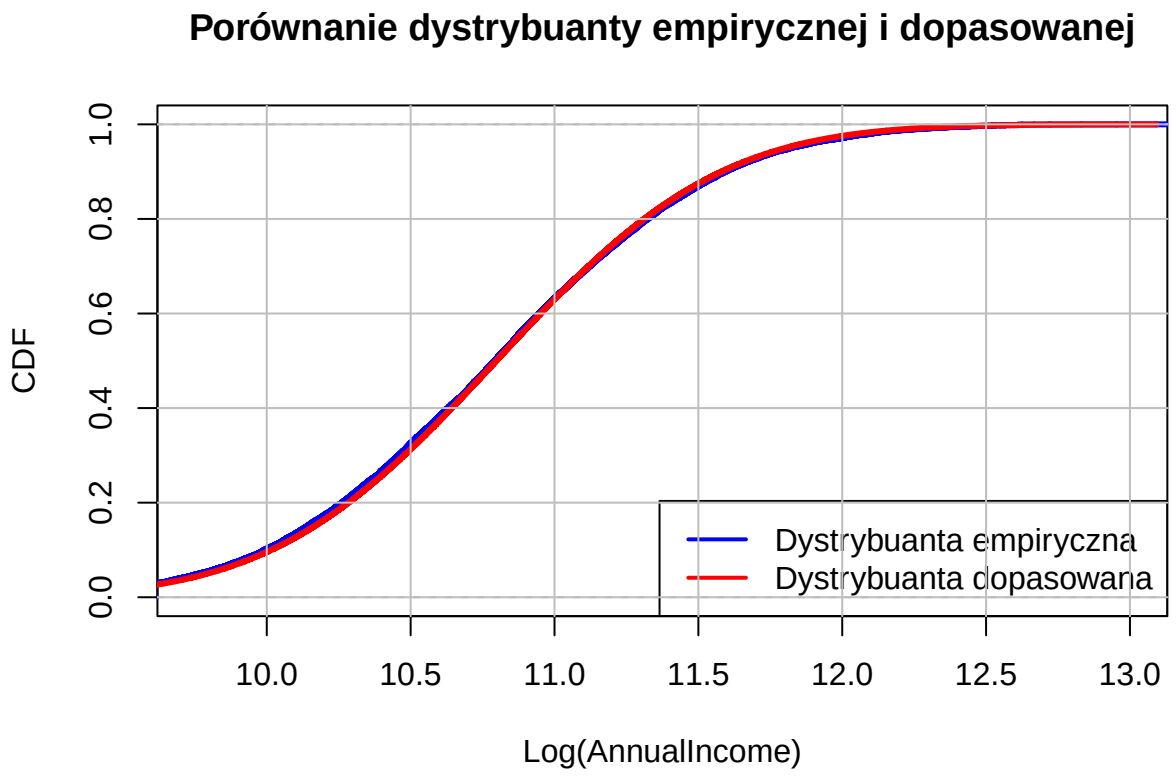


Na wykresie widzimy całkiem niezłe dopasowanie, problem stanowią największe i najmniejsze wartości. Rzeczywiście na histogramie mogliśmy zauważyć, że pierwszy słupek był nienaturalnie wyższy.

Popatrzmy na dopasowanie dystrybuanty.

```
empirical_cdf <- ecdf(log_AnnualIncome)
fitted_cdf <- pnorm(sort(log_AnnualIncome), mean = mean(log_AnnualIncome), sd = sd(log_AnnualIncome))

plot(empirical_cdf, main = "Porównanie dystrybuanty empirycznej i dopasowanej",
     xlab = "Log(AnnualIncome)", ylab = "CDF", col = "blue", lwd = 3, xlim = c(9.75, 13))
lines(sort(log_AnnualIncome), fitted_cdf, col = "red", lwd = 3)
legend("bottomright", legend = c("Dystrybuanta empiryczna", "Dystrybuanta dopasowana"),
       col = c("blue", "red"), lwd = 2)
grid(nx = NULL, ny = NULL, col = "gray", lty = "solid", lwd = 1)
```



Tutaj również dopasowanie jest na pierwszy rzut oka idealne. Żeby potwierdzić domysły przeprowadźmy test Kołmogorowa - Smirnowa.

Analiza rozkładu - test Kołmogorowa-Smirnowa

```
ks_test <- ks.test(log_AnnualIncome, "pnorm", mean = mean(log_AnnualIncome), sd = sd(log_AnnualIncome))
ks_test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  log_AnnualIncome
## D = 0.026369, p-value = 1.666e-12
## alternative hypothesis: two-sided
```

Ponieważ p-wartość jest niezwykle mała, odrzucamy hipotezę zerową, że dane po transformacji logarytmicznej pochodzą z dopasowanego rozkładu normalnego. Chociaż wizualne oceny sugerowały rozsądne dopasowanie, test statystyczny wskazuje, że rozkład normalny może nie w pełni oddawać charakterystykę danych po transformacji logarytmicznej.

5. Prawdopodobieństo wystąpienia wartości zmiennych jakościowych w całej populacji

Dla zmiennej jakościowej EducationLevel poniższy kod estymuje prawdopodobieństwo wystąpienia każdej wartości w całej populacji na podstawie obserwacji oraz oblicza przedziały ufności Wilsona na poziomie 99%.

```
value_counts <- dataSet %>%
  group_by(EducationLevel) %>%
  summarise(Count = n())


total <- sum(value_counts$Count)

value_counts <- value_counts %>%
  mutate(
    Proportion = Count / total,
    CI_Lower = binom.confint(Count, total, conf.level = 0.99, methods = "wilson")$lower,
    CI_Upper = binom.confint(Count, total, conf.level = 0.99, methods = "wilson")$upper
  )
print(value_counts)
```

```
## # A tibble: 5 x 5
##   EducationLevel Count Proportion CI_Lower CI_Upper
##   <chr>          <int>      <dbl>    <dbl>    <dbl>
## 1 Associate       4034     0.202    0.194    0.209
## 2 Bachelor        6054     0.303    0.294    0.311
## 3 Doctorate        954     0.0477   0.0440   0.0517
## 4 High School     5908     0.295    0.287    0.304
## 5 Master          3050     0.152    0.146    0.159
```

W powyższej tabelce mamy poszczególne wartości zmiennej EducationLevel, ilość ich wystąpień w populacji, szacowane prawdopodobieństwo wystąpienia w całej populacji na podstawie próby. W ostatnich dwóch kolumnach mamy dolną i górną granicę przedziału ufności na poziomie 99%. Zauważmy, że nasze wyestymowane wartości prawdopodobieństwa leżą w przedziałach ufności.

Zróbmy to samo dla drugiej zmiennej jakościowej - EmploymentStatus.

```
value_counts2 <- dataSet %>%
  group_by(EmploymentStatus) %>%
  summarise(Count = n())


total2 <- sum(value_counts2$Count)

value_counts2 <- value_counts2 %>%
  mutate(
    Proportion = Count / total2,
    CI_Lower = binom.confint(Count, total, conf.level = 0.99, methods = "wilson")$lower,
    CI_Upper = binom.confint(Count, total, conf.level = 0.99, methods = "wilson")$upper
  )
print(value_counts2)
```

```
## # A tibble: 3 x 5
##   EmploymentStatus Count Proportion CI_Lower CI_Upper
##   <chr>            <int>      <dbl>    <dbl>    <dbl>
## 1 Employed         17036     0.852    0.845    0.858
## 2 Self-Employed     1573     0.0786   0.0739   0.0837
## 3 Unemployed        1391     0.0696   0.0651   0.0743
```

---

6. Regresja liniowa zmiennej InterestRate względem zmiennych Age i AnnualIncome

---

Model

```
model <- lm(InterestRate ~ Age + AnnualIncome, data = dataSet)

summary_model <- summary(model)
coefficients <- summary_model$coefficients
mse <- mean(summary_model$residuals^2)


print("Współczynniki : ")
## [1] "Współczynniki : "
print(coefficients)
##                   Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)   2.698140e-01 1.079545e-03 249.93306  0.000000e+00
## Age          -7.175461e-04 2.539743e-05 -28.25271 3.151472e-172
## AnnualIncome -3.684273e-08 7.315509e-09  -5.03625  4.788705e-07
print("Błąd średniokwadratowy : ")
## [1] "Błąd średniokwadratowy : "
print(mse)
## [1] 0.00170584
```

W powyższym modelu zmiennymi objaśniającymi są zmienne ciągłe Age i AnnualIncome, a zmienną objaśnialną jest zmienna InterestRate. Wyestymowane współczynniki to: wyraz wolny = 2.698e-01, współczynnik przy zmiennej Age = -7.175e-04 oraz przy zmiennej AnnualIncome = -3.684e-08. Tak małe wartości są między innymi związane z jednostką w jakiej podana jest zmienna InterestRate. Ponadto możemy zauważyć, że współczynniki są ujemne, więc razem ze wzrostem wieku i rocznej pensji maleje oprocentowanie kredytu, czego można się było spodziewać. Błąd średniokwadratowy modelu wynosi : 1.7e-04.

Wszystkie współczynniki nie są statycznie istotne. Współczynnik przy zmiennej AnnualIncome jest pomijalnie mały. Model nie byłby wcale o wiele gorszy jeśli jedyną zmienną objaśniającą byłaby zmienna Age.

7. Testowanie hipotez

Test chi-kwadrat na niezależność zmiennej Age i zmiennej LoanApproved

Przeprowadzimy teraz test chi-kwadrat.

Hipoteza zerowa : Wiek nie ma wpływu na akceptacje wniosku o pożyczkę.

Hipoteza alternatywna : Wiek ma wpływ na akceptacje wniosku o pożyczkę.

Przed wykonaniem testu podziele wiek na grupy wiekowe co ułatwi interpretowalność testu i zmiejszy liczbę unikalnych wartości.

```
dataSet <- dataSet %>%
  mutate(AgeGroup = cut(Age, breaks = c(18, 30, 40, 50, 60, 70),
                        labels = c("18-30", "31-40", "41-50", "51-60", "61-70")))
```

```
table <- table(dataSet$AgeGroup, dataSet$LoanApproved)

chi_test <- chisq.test(table)



cat("Hipoteza 1: Wiek a akceptacja wniosku o pożyczkę\n")
```

```
## Hipoteza 1: Wiek a akceptacja wniosku o pożyczkę
```

```
cat("Statystyka Chi2: ", chi_test$statistic, "\n")
```

```
## Statystyka Chi2:  323.4294
```

```
cat("p-wartość: ", chi_test$p.value, "\n")
```

```
## p-wartość:  9.541913e-69
```

p-wartość jest bardzo mała, więc hipotezę zerową należy odrzucić. Więc zmienna Age ma wpływ na zmienną LoanApproved.

Test normalności Shapiro-Wilka na zmiennej CreditScore

Hipoteza zerowa : Punkty kredytowe mają rozkład normalny.

Hipoteza alternatywna : Punkty kredytowe nie mają rozkładu normalnego.

```
sample <- sample(dataSet$CreditScore, size = 5000, replace = FALSE)
shapiro_test <- shapiro.test(sample)



cat("Test Shapiro-Wilka dla zmiennej AnnualIncome:\n")
```

```
## Test Shapiro-Wilka dla zmiennej AnnualIncome:
```

```
cat("Statystyka Shapiro-Wilka: ", shapiro_test$statistic, "\n")
```

```
## Statystyka Shapiro-Wilka:  0.9727161
```

```
cat("p-wartość: ", shapiro_test$p.value, "\n")
```

```
## p-wartość:  7.295412e-30
```

W powyższym teście pobrałem próbkę z danych, ponieważ dla tak dużej próbki test Shapiro-Wilka działa gorzej.

p-wartość jest bardzo mała, więc hipotezę zerową należy odrzucić. Zmienna nie ma rozkładu normalnego.