Statistička analiza podataka - projekt

AMMP

2024-01-21

Procjena kreditnog rizika

Studenti: Marita Radić, Antun Tišljar, Petar Knežević i Mislav Rendulić Asistent: Andro Merćep

Cilj ovoga projekta je uzeti dane podatke i iz njih probati izvući zaključke i faktore koji mogu utjecati na veću vjerojatnost neispunjavanja obaveza prema banci te odlazak u status "default". Ključna stvar je korištenje ispravnih testova te dobivanje validnih rezultata.

1. Sadržaj

2. Osnovna prilagodba podataka

Najprije učitamo podatke. Od iznimne je važnosti ispravno proučavanje istih, kako ne bismo donijeli neispravne zaključke. Tek nakon iscrpne analize možemo započeti sa testiranjem naših hipoteza.

```
data = read.csv("procjena_kreditnog_rizika.csv")
```

Radi preglednosti čistimo i uljepšavamo podatke:

```
data$Default <- as.logical(data$Default)

data$ResidenceSince <- ifelse(data$ResidenceSince ==
    ".. >= 7 years", "... >= 7 years", data$ResidenceSince)

data$NumExistingCredits <- gsub("above", "... >=",
    data$NumExistingCredits)
data$NumExistingCredits <- gsub("or", "||", data$NumExistingCredits)

data$NumberOfDependents <- gsub("less than", "... <",
    data$NumberOfDependents)
data$NumberOfDependents <- gsub("3 or more", "... >= 3",
    data$NumberOfDependents)
```

Sažetak očišćenih podataka:

```
summary(data)
```

```
##  AccountStatus         Duration    CreditHistory        Purpose
##  Length:1000        Min.   : 4.0   Length:1000        Length:1000
##  Class :character   1st Qu.:12.0   Class :character   Class :character
##  Mode  :character   Median :18.0   Mode  :character   Mode  :character
##                     Mean   :20.9
##                     3rd Qu.:24.0
##                     Max.   :72.0
##   CreditAmount     Account           EmploymentSince    PercentOfIncome
##  Min.   :  250   Length:1000        Length:1000        Length:1000
##  1st Qu.: 1366   Class :character   Class :character   Class :character
##  Median : 2320   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 3271
##  3rd Qu.: 3972
##  Max.   :18424
##  PersonalStatus     OtherDebtors       ResidenceSince       Property
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       Age        OtherInstallPlans    Housing          NumExistingCredits
##  Min.   :19.00   Length:1000        Length:1000        Length:1000
##  1st Qu.:27.00   Class :character   Class :character   Class :character
##  Median :33.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :35.55
##  3rd Qu.:42.00
##  Max.   :75.00
##      Job            NumberOfDependents  Telephone         ForeignWorker
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Default
##  Mode :logical
##  FALSE:700
##  TRUE :300
##
##
##
```

Prvih nekoliko redova očišćenih podataka:

```
head(data)
```

```
##         AccountStatus Duration
## 1             ... < 0        6
## 2      0 <= ... < 200       48
## 3 no checking account       12
## 4             ... < 0       42
## 5             ... < 0       24
## 6 no checking account       36
##                                          CreditHistory
```

```
## 1 critical account/ other credits existing (not at this bank)
## 2                   existing credits paid back duly till now
## 3 critical account/ other credits existing (not at this bank)
## 4                   existing credits paid back duly till now
## 5                           delay in paying off in the past
## 6                   existing credits paid back duly till now
##              Purpose CreditAmount                    Account
## 1    radio/television         1169 unknown/ no savings account
## 2    radio/television         5951                  ... < 100
## 3           education         2096                  ... < 100
## 4 furniture/equipment         7882                  ... < 100
## 5           car (new)         4870                  ... < 100
## 6           education         9055 unknown/ no savings account
##      EmploymentSince  PercentOfIncome                      PersonalStatus
## 1     ... >= 7 years        ... < 20%                       male - single
## 2 1 <= ... < 4 years 25% <= ... < 35% female - divorced/separated/married
## 3 4 <= ... < 7 years 25% <= ... < 35%                       male - single
## 4 4 <= ... < 7 years 25% <= ... < 35%                       male - single
## 5 1 <= ... < 4 years 20% <= ... < 25%                       male - single
## 6 1 <= ... < 4 years 25% <= ... < 35%                       male - single
##   OtherDebtors     ResidenceSince
## 1         none     ... >= 7 years
## 2         none 1 <= ... < 4 years
## 3         none 4 <= ... < 7 years
## 4    guarantor     ... >= 7 years
## 5         none     ... >= 7 years
## 6         none     ... >= 7 years
##                                          Property Age OtherInstallPlans
## 1                                     real estate  67              none
## 2                                     real estate  22              none
## 3                                     real estate  49              none
## 4 building society savings agreement/ life insurance  45              none
## 5                           unknown / no property  53              none
## 6                           unknown / no property  35              none
##    Housing NumExistingCredits                        Job NumberOfDependents
## 1      own             2 || 3 skilled employee / official           ... >= 3
## 2      own                  1 skilled employee / official           ... >= 3
## 3      own                  1        unskilled - resident            ... < 3
## 4 for free                  1 skilled employee / official            ... < 3
## 5 for free             2 || 3 skilled employee / official            ... < 3
## 6 for free                  1        unskilled - resident            ... < 3
##                               Telephone ForeignWorker Default
## 1 yes, registered under the customers name          yes   FALSE
## 2                                   none          yes    TRUE
## 3                                   none          yes   FALSE
## 4                                   none          yes   FALSE
## 5                                   none          yes    TRUE
## 6 yes, registered under the customers name          yes   FALSE
```

Poredajmo varijable i pretvarimo podatkovni tip u faktor kako bismo ih kasnije mogli jednostavnije analizirati:

```
data$AccountStatus <- factor(data$AccountStatus,
    levels = c("no checking account", "... < 0",
        "0 <= ... < 200", "... >= 200"))
```

```
data$CreditHistory <- factor(data$CreditHistory,
    levels = c("critical account/ other credits existing (not at this bank)",
        "delay in paying off in the past", "existing credits paid back duly till now",
        "all credits at this bank paid back duly",
        "no credits taken/ all credits paid back duly"))

data$Purpose <- factor(data$Purpose)

data$Account <- factor(data$Account, levels = c("unknown/ no savings account",
    "... < 100", "100 <= ... < 500", "500 <= ... < 1000",
    "... >= 1000"))

data$EmploymentSince <- factor(data$EmploymentSince,
    levels = c("unemployed", "... < 1 year", "1 <= ... < 4 years",
        "4 <= ... < 7 years", "... >= 7 years"))

data$PercentOfIncome <- factor(data$PercentOfIncome,
    levels = c("... < 20%", "20% <= ... < 25%",
        "25% <= ... < 35%", "... >= 35%"))

split_parts <- strsplit(as.character(data$PersonalStatus),
    " - ")
data$Gender <- sapply(split_parts, function(x) x[1])
data$MaritalStatus <- sapply(split_parts, function(x) x[2])
data <- data[, !(names(data) %in% c("PersonalStatus"))]
data <- data %>%
    select(Gender, MaritalStatus, everything())
data$Gender <- factor(data$Gender)
data$MaritalStatus <- factor(data$MaritalStatus)
data$OtherDebtors <- factor(data$OtherDebtors)

data$ResidenceSince <- factor(data$ResidenceSince,
    levels = c("... < 1 year", "1 <= ... < 4 years",
        "4 <= ... < 7 years", "... >= 7 years"))

data$Property <- factor(data$Property)
data$OtherInstallPlans <- factor(data$OtherInstallPlans)
data$Housing <- factor(data$Housing)

data$NumExistingCredits <- factor(data$NumExistingCredits,
    levels = c("1", "2 || 3", "4 || 5", "... >= 6"))

data$Job <- factor(data$Job, levels = c("unemployed/ unskilled - non-resident",
    "unskilled - resident", "skilled employee / official",
    "management/ self-employed/highly qualified employee/ officer"))

data$NumberOfDependents <- factor(data$NumberOfDependents,
    levels = c("... < 3", "... >= 3"))
data$Telephone <- factor(data$Telephone)
data$ForeignWorker <- factor(data$ForeignWorker)
data$Default <- factor(data$Default)

attach(data)
```

```
head(data)
```

```
##   Gender              MaritalStatus       AccountStatus Duration
## 1   male                     single             ... < 0        6
## 2 female divorced/separated/married      0 <= ... < 200       48
## 3   male                     single no checking account       12
## 4   male                     single             ... < 0       42
## 5   male                     single             ... < 0       24
## 6   male                     single no checking account       36
##                                           CreditHistory
## 1 critical account/ other credits existing (not at this bank)
## 2                   existing credits paid back duly till now
## 3 critical account/ other credits existing (not at this bank)
## 4                   existing credits paid back duly till now
## 5                            delay in paying off in the past
## 6                   existing credits paid back duly till now
##              Purpose CreditAmount                     Account
## 1    radio/television         1169 unknown/ no savings account
## 2    radio/television         5951                  ... < 100
## 3           education         2096                  ... < 100
## 4 furniture/equipment         7882                  ... < 100
## 5           car (new)         4870                  ... < 100
## 6           education         9055 unknown/ no savings account
##      EmploymentSince  PercentOfIncome OtherDebtors     ResidenceSince
## 1     ... >= 7 years        ... < 20%         none     ... >= 7 years
## 2 1 <= ... < 4 years 25% <= ... < 35%         none 1 <= ... < 4 years
## 3 4 <= ... < 7 years 25% <= ... < 35%         none 4 <= ... < 7 years
## 4 4 <= ... < 7 years 25% <= ... < 35%    guarantor     ... >= 7 years
## 5 1 <= ... < 4 years 20% <= ... < 25%         none     ... >= 7 years
## 6 1 <= ... < 4 years 25% <= ... < 35%         none     ... >= 7 years
##                                          Property Age OtherInstallPlans
## 1                                       real estate  67              none
## 2                                       real estate  22              none
## 3                                       real estate  49              none
## 4 building society savings agreement/ life insurance  45              none
## 5                              unknown / no property  53              none
## 6                              unknown / no property  35              none
##    Housing NumExistingCredits                          Job NumberOfDependents
## 1      own             2 || 3 skilled employee / official           ... >= 3
## 2      own                  1 skilled employee / official           ... >= 3
## 3      own                  1        unskilled - resident            ... < 3
## 4 for free                  1 skilled employee / official            ... < 3
## 5 for free             2 || 3 skilled employee / official            ... < 3
## 6 for free                  1        unskilled - resident            ... < 3
##                                  Telephone ForeignWorker Default
## 1 yes, registered under the customers name           yes   FALSE
## 2                                     none           yes    TRUE
## 3                                     none           yes   FALSE
## 4                                     none           yes   FALSE
## 5                                     none           yes    TRUE
## 6 yes, registered under the customers name           yes   FALSE
```

3. Možemo li temeljem drugih dostupnih varijabli predvidjeti hoće li nastupiti *default* za određenog klijenta? Koje varijable povežavaju tu vjerojatnost?

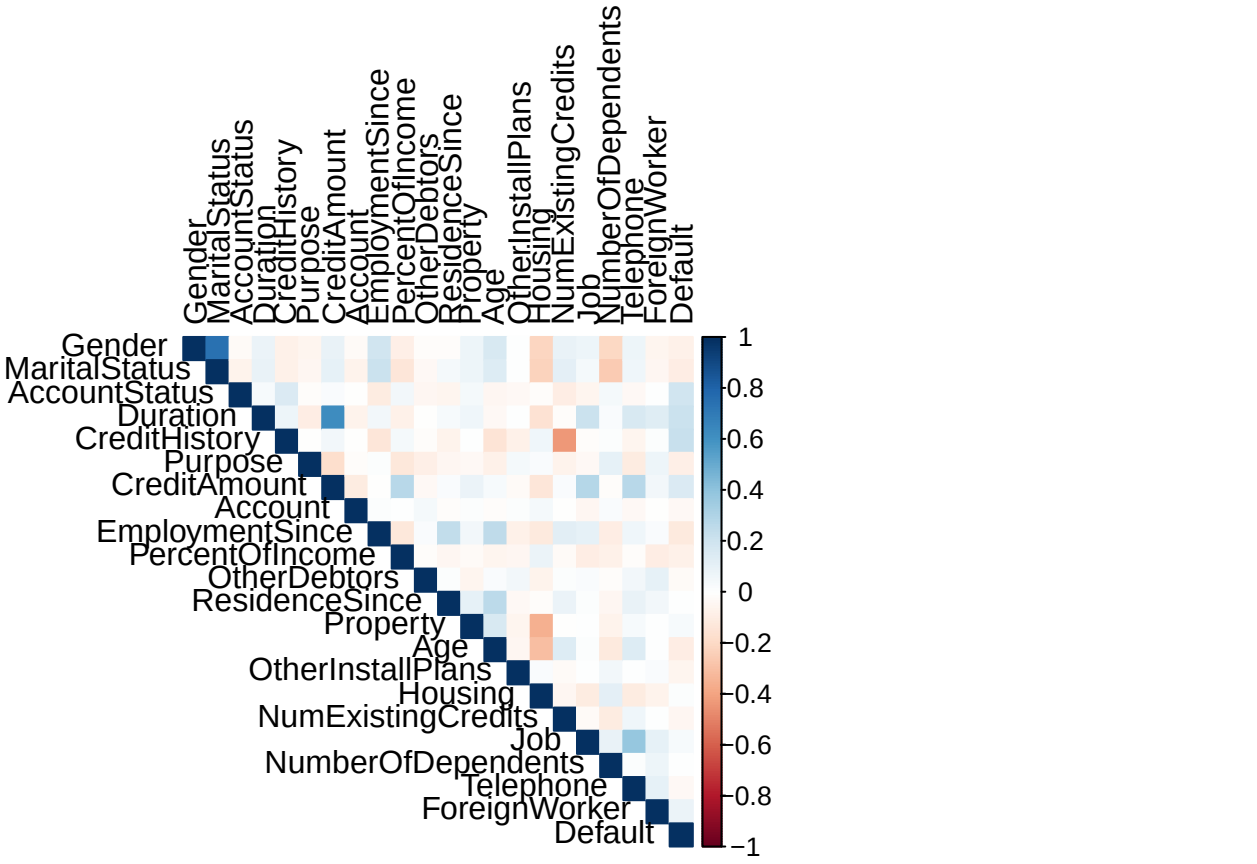
Računamo i prikazujemo matricu korelacije. Želimo vidjeti kako se pojedinačne varijable posebno koreliraju s varijablom "default", stoga ćemo zasebno nacrtati taj grafikon. Cilj je pronaći i bolje istražiti varijable koje imaju veći utjecaj na konačni ishod varijable "default":

```
library(corrplot)
```

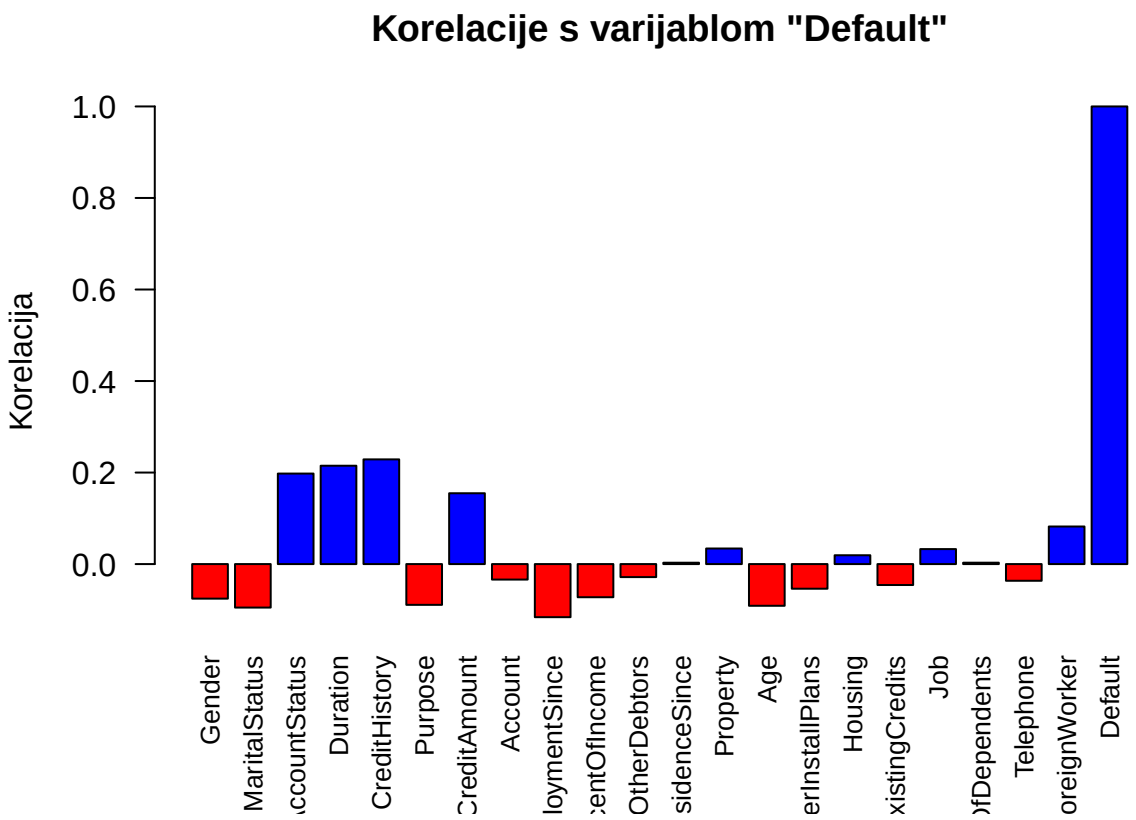
```
## corrplot 0.92 loaded
```

Ispisujemo matricu korelacije:

```
corr_matrix <- cor(data.frame(lapply(data, function(x) as.numeric(x))))
corrplot(corr_matrix, method = "color", type = "upper",
    tl.col = "black", tl.srt = 90)
```




```
cor_with_target <- corr_matrix["Default", ]
barplot(cor_with_target, names.arg = names(cor_with_target),
    las = 2, cex.names = 0.8, col = ifelse(cor_with_target >
        0, "blue", "red"), main = paste("Korelacije s varijablom \"Default\""),
    ylab = "Korelacija")
```

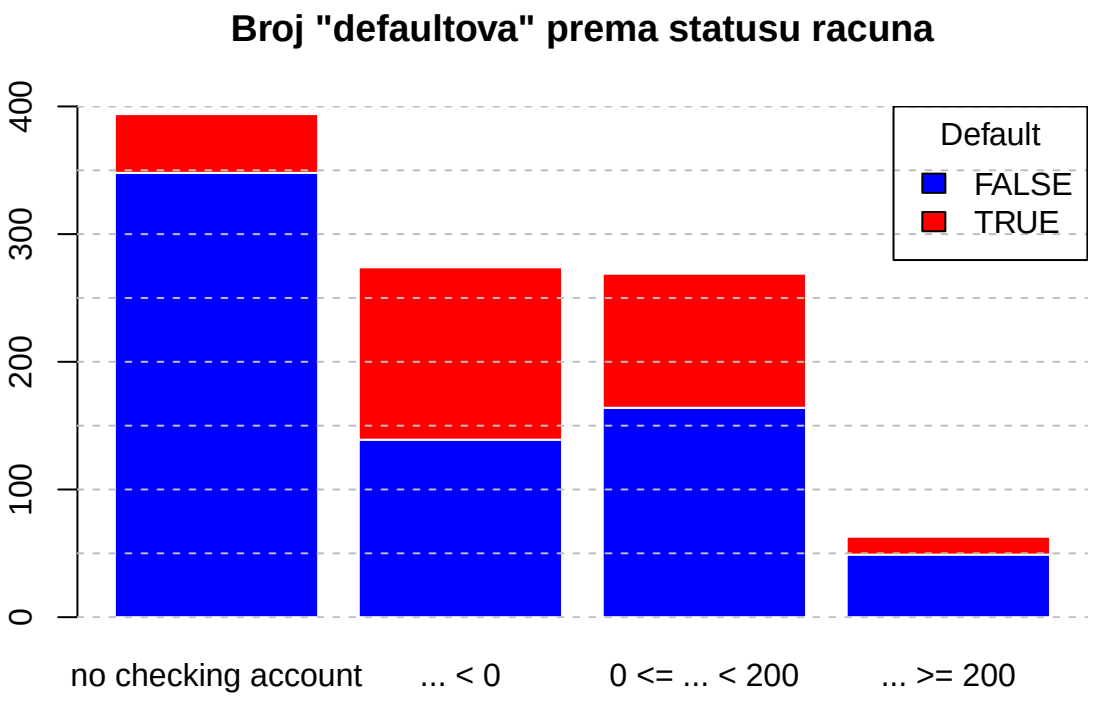



```
filt_cor_with_target <- cor_with_target[abs(cor_with_target) >=
    0.1 & cor_with_target != 1]
print(names(filt_cor_with_target))
```

```
## [1] "AccountStatus"   "Duration"        "CreditHistory"   "CreditAmount"
## [5] "EmploymentSince"
```

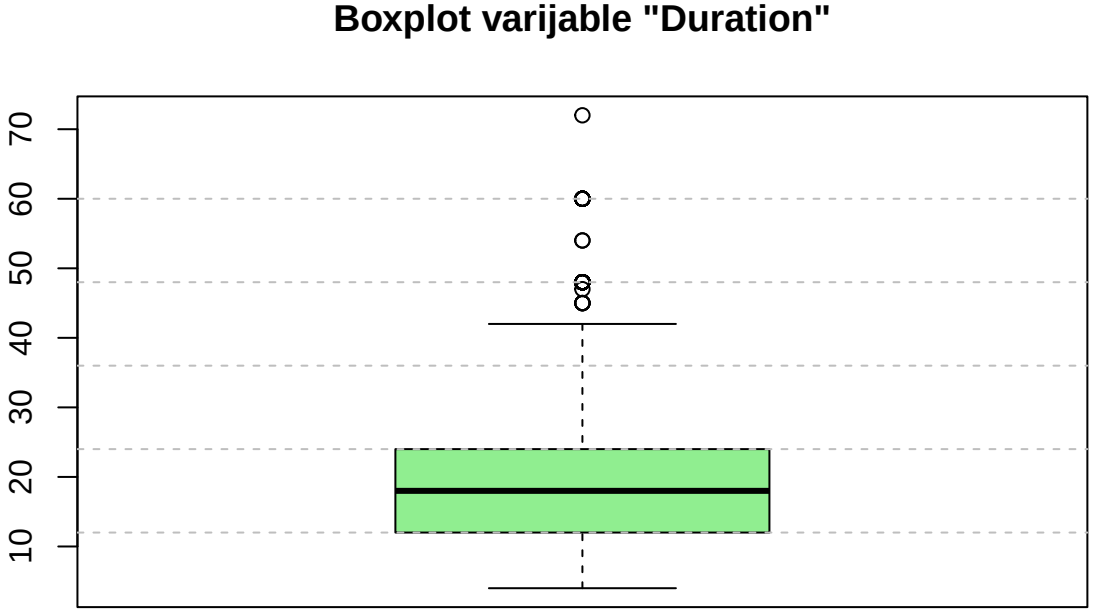
Postoje 5 varijabli koje imaju apsolutnu korelaciju veću ili jednaku 0,1. Te varijable mogu imati veću prediktivnu moć od ostalih koje slabo koreliraju s varijablom "Default". Proučit ćemo ih detaljnije. Počet ćemo s "AccountStatus", koji nam govori o trenutnom stanju računa osobe, ako ga uopće ima:

```
barplot(table(Default, AccountStatus), main = "Broj \"defaultova\" prema statusu računa",
    border = "white", col = c("blue", "red"),
    xlab = "Status računa", ylim = c(0, 400))
abline(h = seq(0, 500, by = 50), col = "gray",
    lty = 2)
legend("topright", legend = levels(Default), fill = c("blue",
    "red"), title = "Default")
```



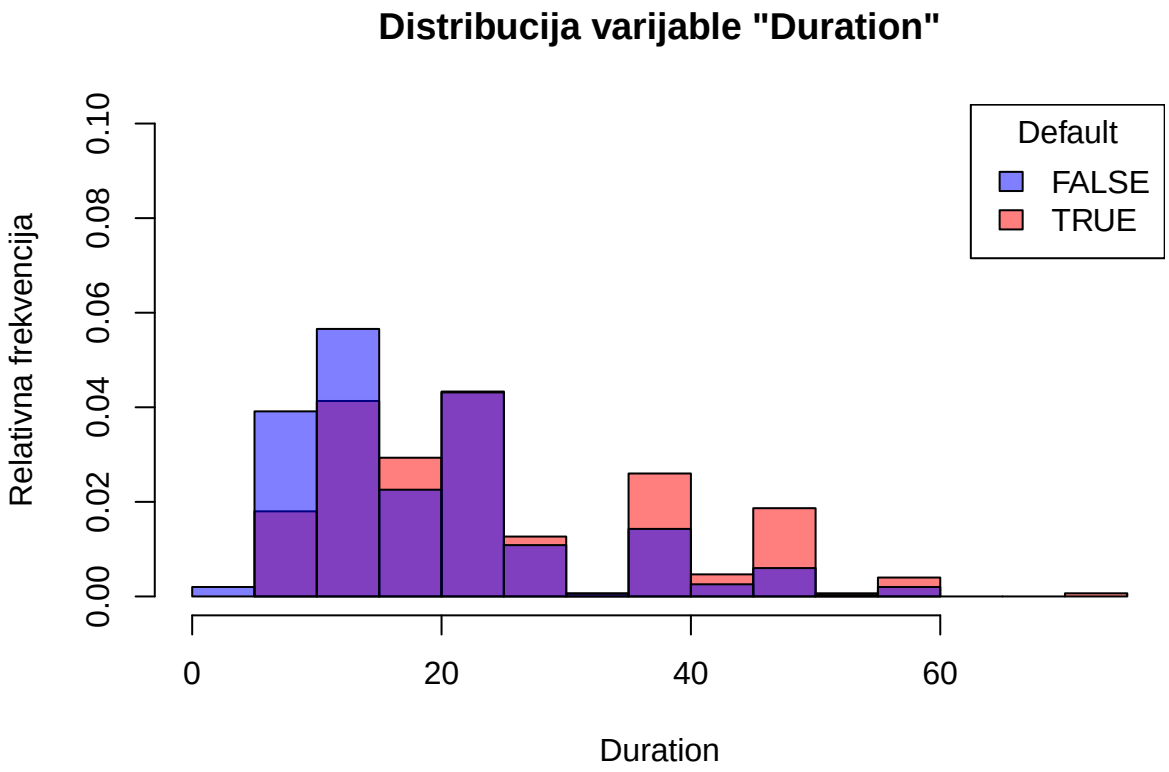
Ono što vidimo ovdje jest da je rezultat korelacije snažno utjecan vrijednošću varijable "no checking account", što nam ne pruža puno informacija. Ako bismo isključili tu vrijednost (što ne možemo jer bismo izgubili gotovo 40% podataka), vidjeli bismo negativnu korelaciju. To bi više odgovaralo našem očekivanju da što netko ima više novca na računu, to je manja vjerojatnost da će doći do neizvršenja plaćanja. Nastavljamo s varijablom "Duration", koja je numerička varijabla i mogla bi nam pružiti više informacija:

```
boxplot(Duration, col = "lightgreen", main = "Boxplot varijable \"Duration\"")
abline(h = seq(0, 70, by = 12), col = "gray",
    lty = 2)
```



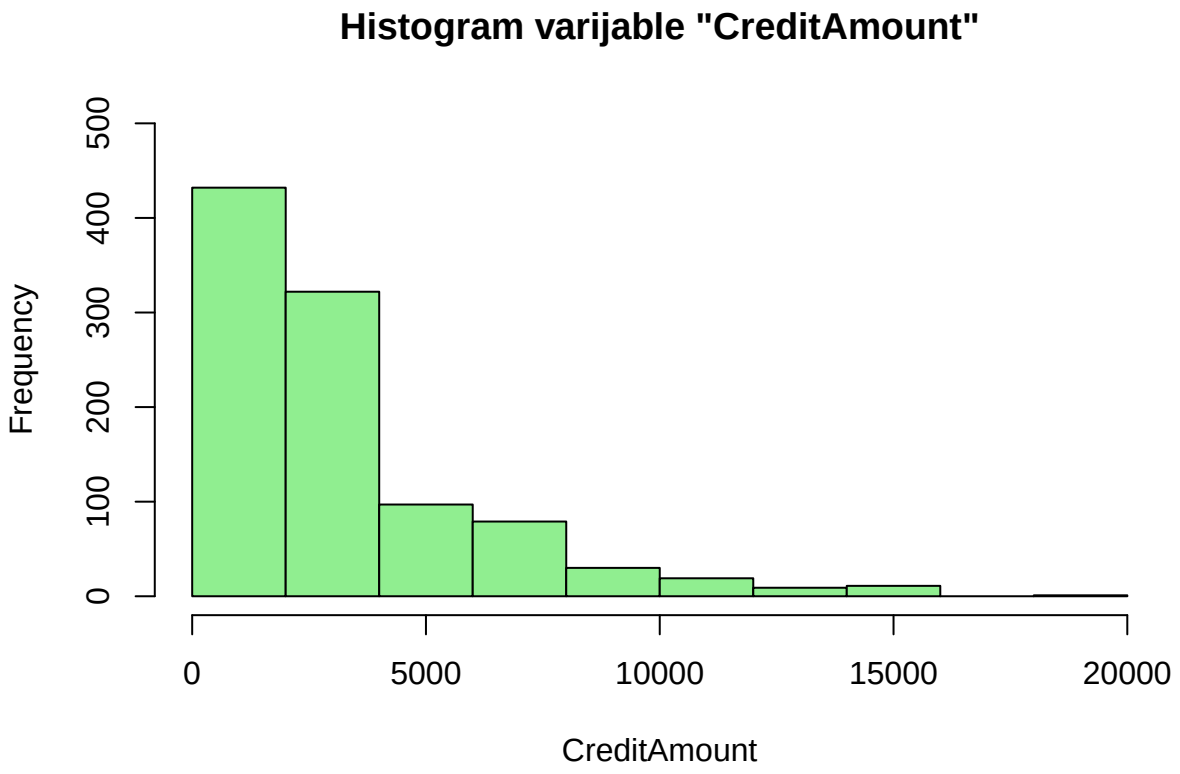
```
defaulted <- data[Default == "TRUE", ]
not_defaulted <- data[Default == "FALSE", ]

hist(defaulted$Duration, breaks = 10, xlim = c(0,
    75), ylim = c(0, 0.1), freq = FALSE, col = rgb(1,
    0, 0, 0.5), xlab = "Duration", ylab = "Relativna frekvencija",
    main = "Distribucija varijable \"Duration\"")
hist(not_defaulted$Duration, breaks = 10, freq = FALSE,
    xlim = c(0, 75), col = rgb(0, 0, 1, 0.5),
    add = TRUE)
legend("topright", legend = levels(Default), fill = c(rgb(0,
    0, 1, 0.5), rgb(1, 0, 0, 0.5)), title = "Default")
```

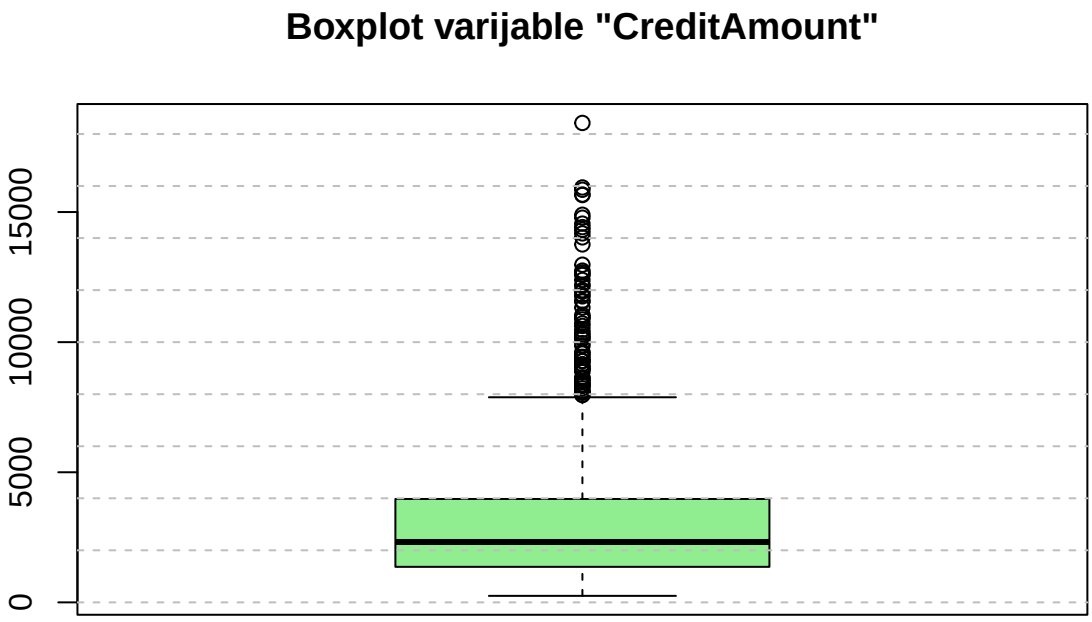


Gledajući distribuciju varijable "Duration", možemo zaključiti da je većina kredita kratkoročna. Također pretpostavljamo da postoje neki izvanredni podaci, krediti s trajanjem od 50 ili više mjeseci. Boxplot nam to potvrđuje i također nam govori da je barem 50% podataka između 12 i 24 mjeseca. Sljedeće što nas zanima je koji od tih kredita su završili s plaćanjem. Pozitivna korelacija sugerira da što je duže trajanje kredita, veća je vjerojatnost neizvršenja plaćanja. To je i ono što vidimo u odvojenim histogramima za kredite s neizvršenjem plaćanja i one bez neizvršenja plaćanja. Trajanje kredita samo po sebi neće nam pružiti potpunu perspektivu. Kako bismo upotpunili tu varijablu, koristit ćemo "CreditAmount":

```
hist(CreditAmount, col = "lightgreen", ylim = c(0,
    500), main = "Histogram varijable \"CreditAmount\"")
```



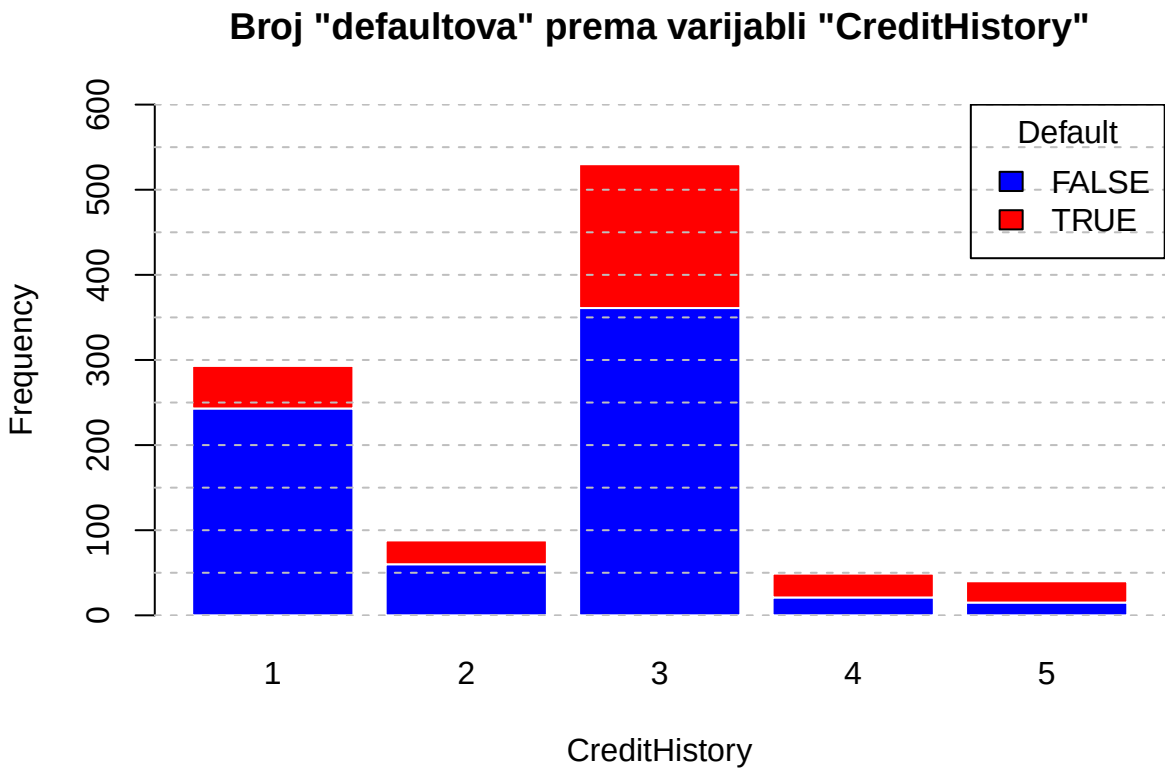

```
boxplot(CreditAmount, col = "lightgreen", main = "Boxplot varijable \"CreditAmount\"")
abline(h = seq(0, 20000, by = 2000), col = "gray",
    lty = 2)
```



Nijedna od ovih varijabli ne pruža nam potpune informacije. Kredit može imati dugotrajnost i mali iznos, i obrnuto. Kako bismo zaključili naš popis koreliranih varijabli, provjerimo posljednje dvije, "CreditHistory" i "EmploymentSince":

```
barplot(table(Default, as.numeric(CreditHistory)),
    main = "Broj \"defaultova\" prema varijabli \"CreditHistory\"",
    border = "white", col = c("blue", "red"),
```

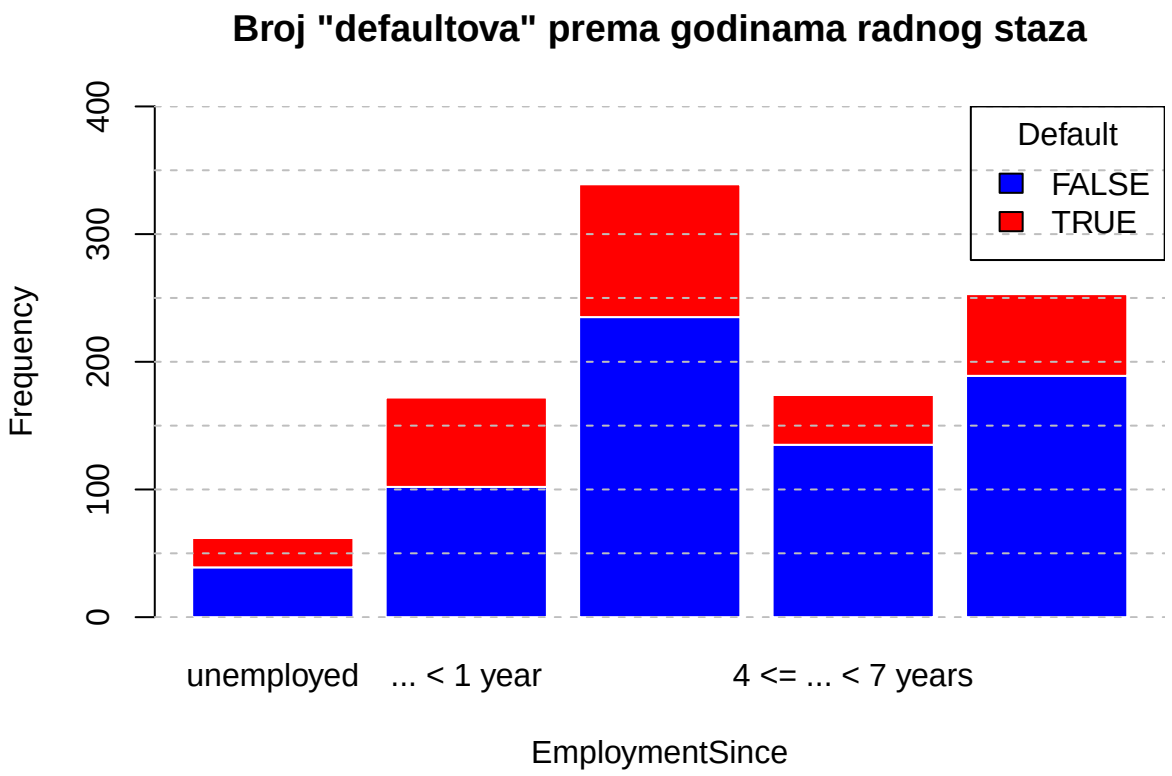
```
    xlab = "CreditHistory", ylab = "Frequency",
    ylim = c(0, 600))
abline(h = seq(0, 600, by = 50), col = "gray",
    lty = 2)
legend("topright", legend = levels(Default), fill = c("blue",
    "red"), title = "Default")
```



Opis stupaca grafa:

1 - Kritični račun / ostali postojeći krediti (ne na ovoj banci)
2 - U prošlosti kašnjenje u otplati
3 - Postojeći krediti do sada uredno vraćeni
4 - Svi krediti u ovoj banci uredno vraćeni
5 - Nema kredita / svi krediti uredno vraćeni

```
barplot(table(Default, EmploymentSince), main = "Broj \"defaultova\" prema godinama radnog staža",
    border = "white", col = c("blue", "red"),
    xlab = "EmploymentSince", ylab = "Frequency",
    ylim = c(0, 400))
abline(h = seq(0, 400, by = 50), col = "gray",
    lty = 2)
legend("topright", legend = levels(Default), fill = c("blue",
    "red"), title = "Default")
```



Suprotno onome što bi smo prvotno zaključili, analiza govori da dobra kreditna povijest ne znači da će kredit uredno biti vraćan.

Nakon obavljene analize testirajmo sada statističkim testom možemo li temeljem drugih dostupnih varijabli predvidjeti hoće li nastupiti default za odredenog klijenta i koje varijable povećavaju tu vjerojatnost.

```
require(caret)
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
logreg.mdl <- glm(Default ~ AccountStatus + Duration +
    CreditHistory + Purpose + CreditAmount + Account +
    EmploymentSince + PercentOfIncome + Gender +
    MaritalStatus + OtherDebtors + ResidenceSince +
    Property + Age + OtherInstallPlans + Housing +
    NumExistingCredits + Job + NumberOfDependents +
    Telephone + ForeignWorker, data = data, family = binomial())

Rsq <- 1 - logreg.mdl$deviance/logreg.mdl$null.deviance

coef_summary <- summary(logreg.mdl)$coefficients

coef_table <- data.frame(Variable = rownames(coef_summary),
    Coefficient = coef_summary[, "Estimate"],
```

```
    OddsRatio = exp(coef_summary[, "Estimate"]),
    `Pr(>|z|)` = coef_summary[, "Pr(>|z|)"], stringsAsFactors = FALSE)
significant_vars <- coef_table[coef_table$`Pr(>|z|)` <
    0.05, ]

cat("R-squared:", Rsq, "\n")
```

```
## R-squared: 0.2736555
```

```
cat("Significant variables:", paste(significant_vars$Variable,
    collapse = ", "), "\n")
```

```
## Significant variables:
```

```
print(significant_vars)
```

```
## [1] Variable    Coefficient OddsRatio   Pr...z..
## <0 rows> (or 0-length row.names)
```

Nema značajnih varijabli za p-vrijednost 0.05 što ukazuje da nijedna od varijabli nije statistički relevantna za predviđanje ciljne varijable. Logistička regeresija se u ovom slučaju nije pokazala kao dobar model. Također, obzirom na nisku vrijednost R-kvadrata ukazano nam je da model nije dovoljno dobar u objašnjavanju varijance za varijablu "Default".

```
set.seed(123)

formula <- Default ~ AccountStatus + Duration +
    CreditHistory + Purpose + CreditAmount + Account +
    EmploymentSince + PercentOfIncome + Gender +
    MaritalStatus + OtherDebtors + ResidenceSince +
    Property + Age + OtherInstallPlans + Housing +
    NumExistingCredits + Job + NumberOfDependents +
    Telephone + ForeignWorker

ctrl <- trainControl(method = "cv", number = 10)

cv_model <- suppressWarnings(train(formula, data = data,
    method = "glm", family = binomial(), trControl = ctrl))

print(cv_model)
```

```
## Generalized Linear Model
##
## 1000 samples
##   21 predictor
##    2 classes: 'FALSE', 'TRUE'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
## Resampling results:
##
##   Accuracy  Kappa
##   0.754     0.379993
```

Cross-validated generaliziranog linearnog modela na 1000 uzoraka točno je generalizirao 75.4% uzoraka što ukazuje na (samo) vrlo dobre performanse modela. Stupanj suglasnosti (kappa) između stvarnih i predviđenih

klasa jest 0.38. Želimo da je ta vrijednost bliže 1 te zaključujemo da rezultat nije idealan.

Ovaj model, kako je opisan, pokazuje određenu prediktivnu sposobnost, ali nedostatak značajnih varijabli i umjereni kappa vrijednost sugeriraju da ima prostora za poboljšanja. Trebali bi smo uvažiti mogućnosti daljnjeg istraživanja, selekciju varijabli, podešavanje modela ili istraživanje različitih algoritama.

4. Jesu li muškarci skloniji nesipunjavanja obaveza po kreditu od žena?

```
male <- data[Gender == "male", ]
female <- data[Gender == "female", ]
```

Postotak mušakraca koji nisu redovito ispunjavali obaveze oprema banci:

```
print(as.numeric(1 - table(male$Default)/count(male)))
```

```
## [1] 0.2768116
```

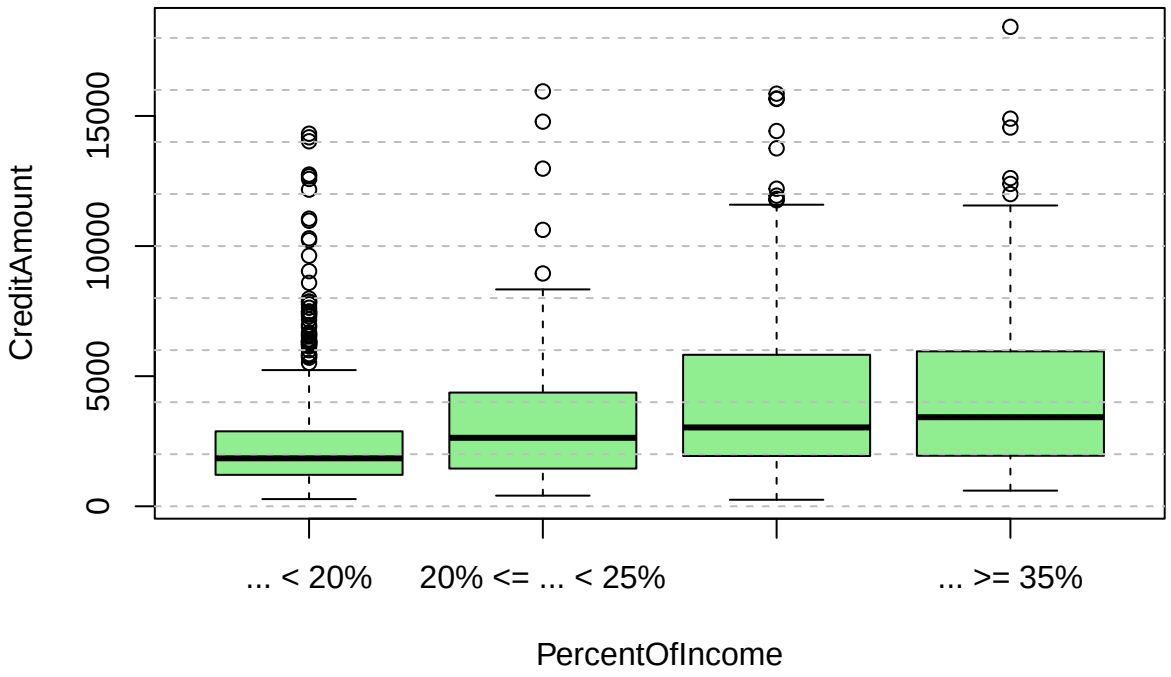
Postotak žena koje nisu redovito ispunjavali obaveze oprema banci:

```
print(as.numeric(1 - table(female$Default)/count(female)))
```

```
## [1] 0.3516129
```

Vidimo da je samo oko 27% muškaraca u našem skupu podataka imalo neizvršenje plaćanja, u usporedbi s otprilike 35% žena. Naši će statistički testovi reći je li ta razlika značajna:

```
boxplot(CreditAmount ~ PercentOfIncome, col = "lightgreen",
    ylab = "CreditAmount", xlab = "PercentOfIncome")
abline(h = seq(0, 20000, by = 2000), col = "gray",
    lty = 2)
```



Na temelju proučavanja podataka ima smisla testirati hipotezu da su žene sklonije neispunjavanju kreditnih obaveza od muškaraca. Kako bismo to testirali koristimo test proporcija. Kao nultu hipotezu pretpostavljamo jednakost proporcija dok za alternativnu hipotezu stavimo da je manji udio muškaraca nego žena koji ne ispunjavaju kreditne obaveze:

```r
male_default_count = sum(male$Default == TRUE)
female_default_count = sum(female$Default == TRUE)
x = c(male_default_count, female_default_count)
n = c(count(male)$n, count(female)$n)

prop.test(x, n, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 5.3485, df = 1, p-value = 0.01037
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.01978865
## sample estimates:
##    prop 1    prop 2
## 0.2768116 0.3516129
```

Na razini značajnosti od 5% možemo zaključiti da su žene sklonije neispunjavanju kreditnih obaveza od muškaraca na temelju ovih podataka.

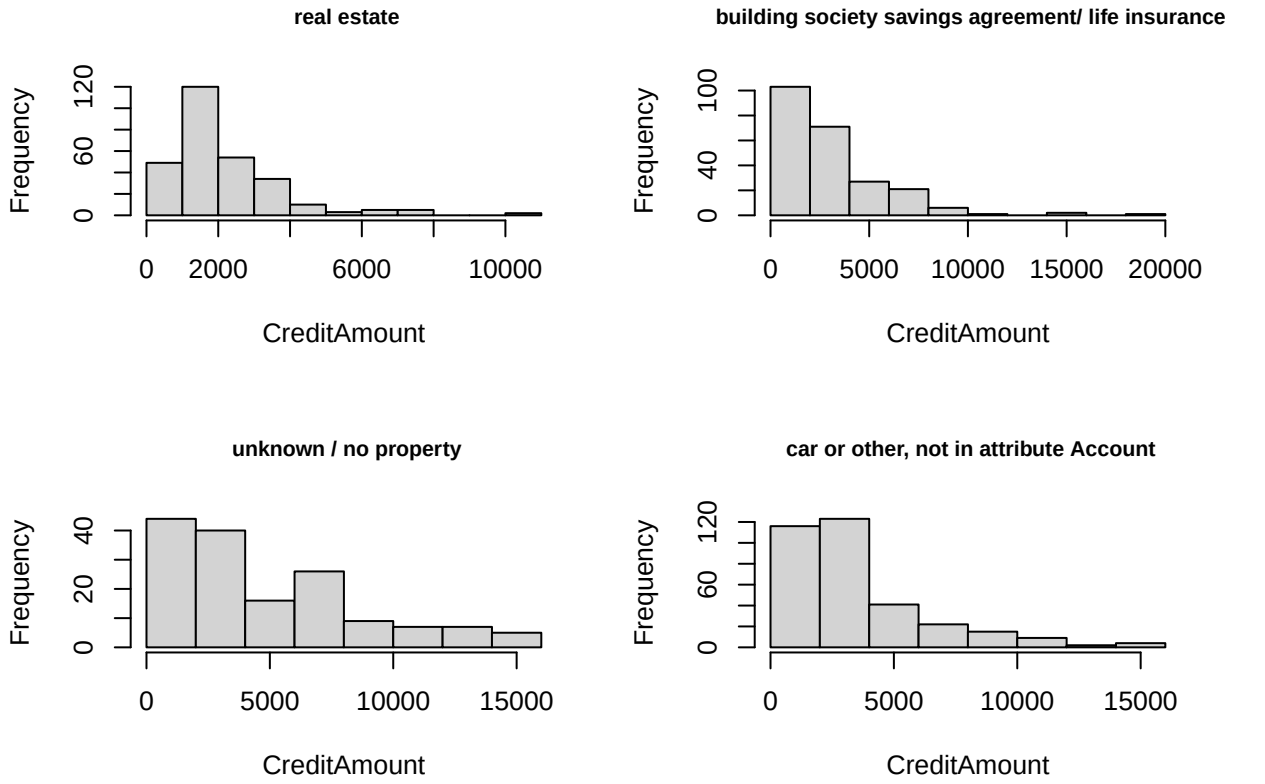
5. Postoje li razlike u traženom iznosu klijenta prema imovini klijenta?

Testirajmo sada postoji li razlika u traženom iznosu kredita prema imovini klijenta. Da bismo to testirali koristimo ANOVA test. On ima određene pretpostavke u čiju se zadovoljenost moramo uvjeriti prije nego krenemo na testiranje. Prva je pretpostavka pojedinih podataka u uzorcima, druge je pretpostavka normalne razdiobe podataka, a treća je pretpostavka homogenosti varijanci među populacijama. Naše populacije se razlikuju s obzirom na imovinu koju osoba posjeduje, a proučavamo iznos kredita. Pogledajmo prvo kako izgledaju histogrami da vidimo ima li pretpostavka o normalnosti smisla.

Histogrami iznosa kredita obzirom na vrstu imovine:

```r
par(mfrow = c(2, 2))

for (item in unique(data$Property)) {
    hist(data$CreditAmount[data$Property == item],
        main = paste("", item), cex.main = 0.8,
        xlab = "CreditAmount")
}
```




```
par(mfrow = c(1, 1))
```

Vidimo kako pretpostavka o normalnosti nema smisla. No, probajmo sada logaritmirati podatke pa provesti Lillieforsovu inačicu KS testa.

Histogrami logaritmiranog iznosa kredita obzirom na vrstu imovine:

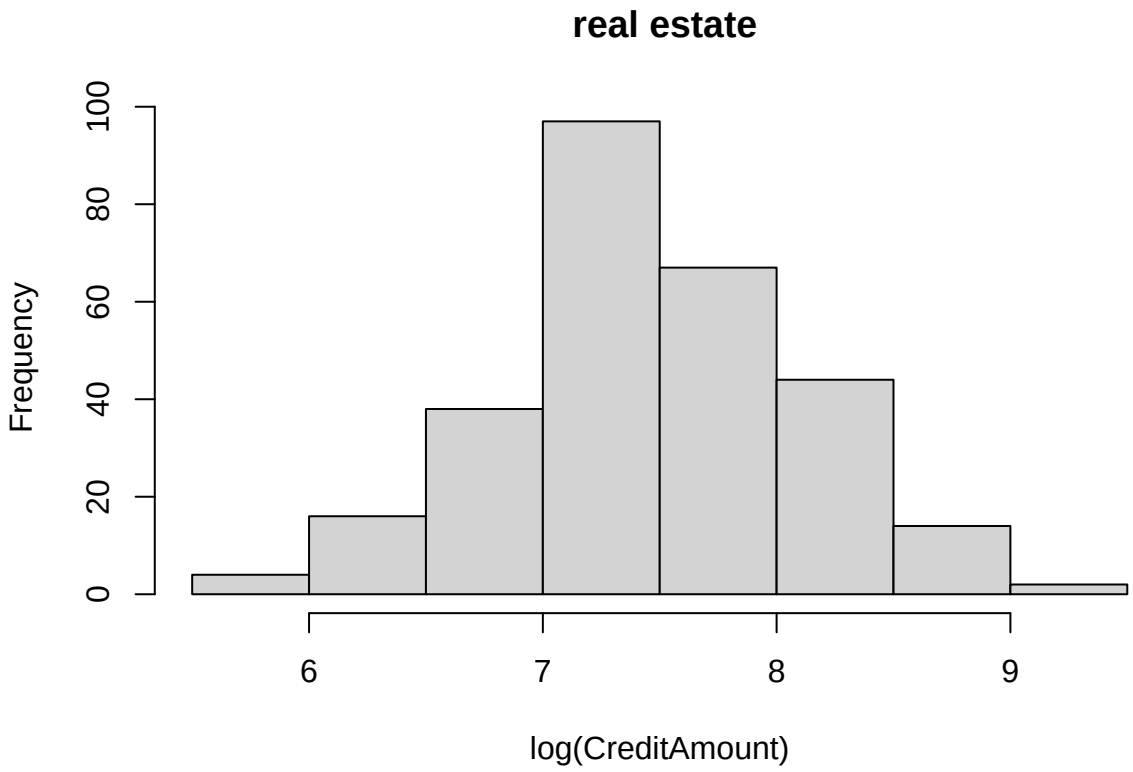
```
require(nortest)
```

```
## Loading required package: nortest
```

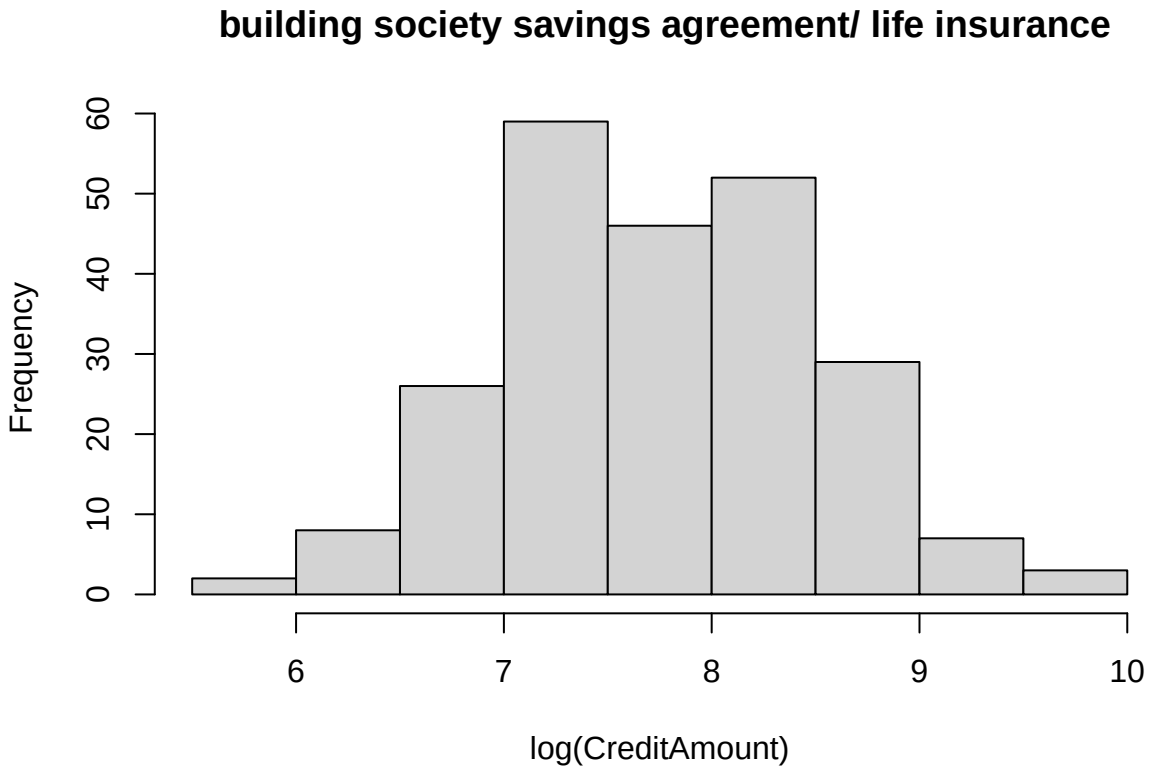
```
data$LogCreditAmount <- log(data$CreditAmount)

for (item in unique(data$Property)) {
    print(lillie.test(data$LogCreditAmount[data$Property ==
        item]))
    hist(data$LogCreditAmount[data$Property ==
        item], main = paste("", item), xlab = "log(CreditAmount)")
}
```

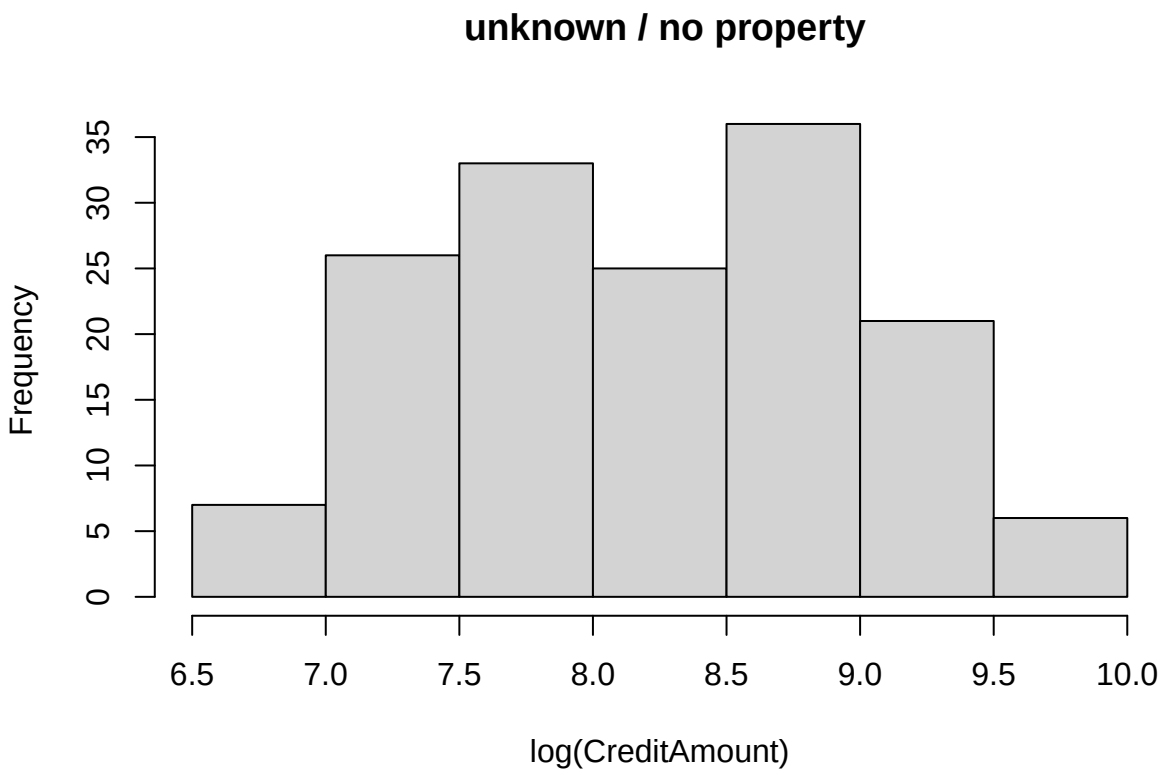
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$LogCreditAmount[data$Property == item]
## D = 0.058694, p-value = 0.02017
```



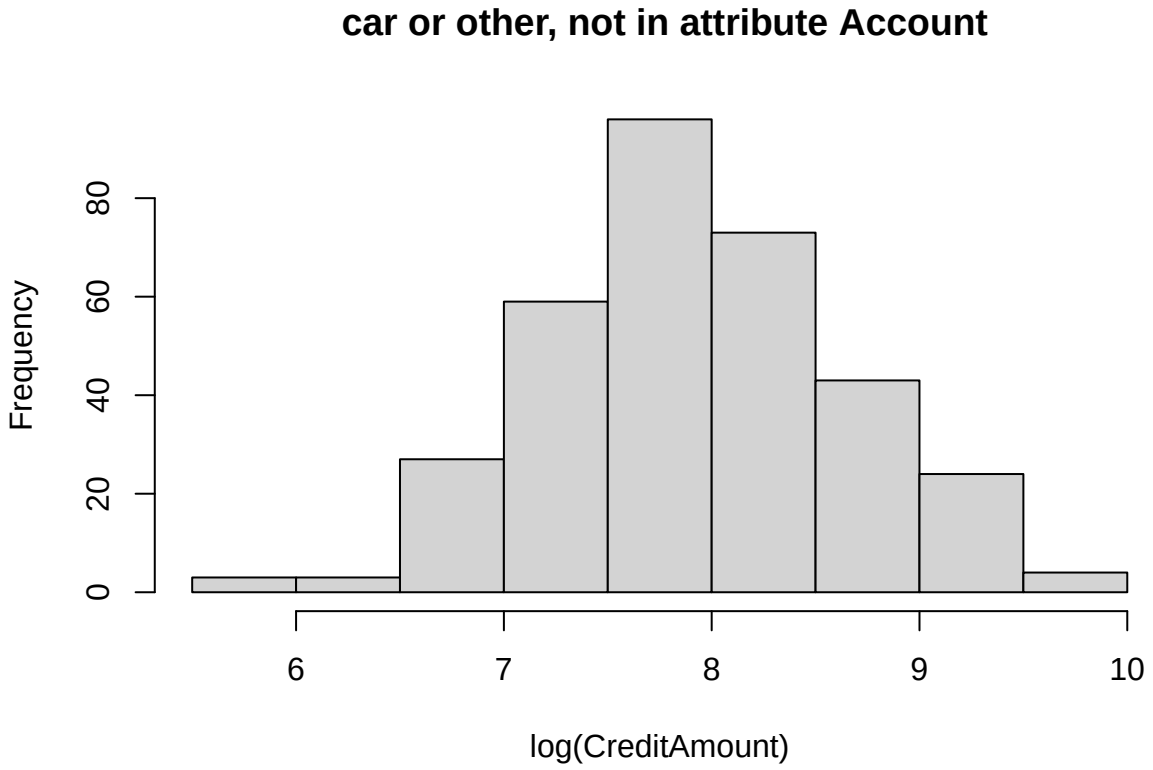

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$LogCreditAmount[data$Property == item]
## D = 0.061582, p-value = 0.03277
```



```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$LogCreditAmount[data$Property == item]
## D = 0.09026, p-value = 0.003752
```




```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$LogCreditAmount[data$Property == item]
## D = 0.035506, p-value = 0.3915
```



Bartlettovim testom testiramo homogenost varijanci kod razlicitih populacija. Postavljamo hipoteze:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
$$H_1 : \text{barem dvije varijance nisu iste.}$$

```
bartlett_result <- bartlett.test(data$LogCreditAmount ~
    data$Property)
print(bartlett_result)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data$LogCreditAmount by data$Property
## Bartlett's K-squared = 9.7812, df = 3, p-value = 0.02052
```

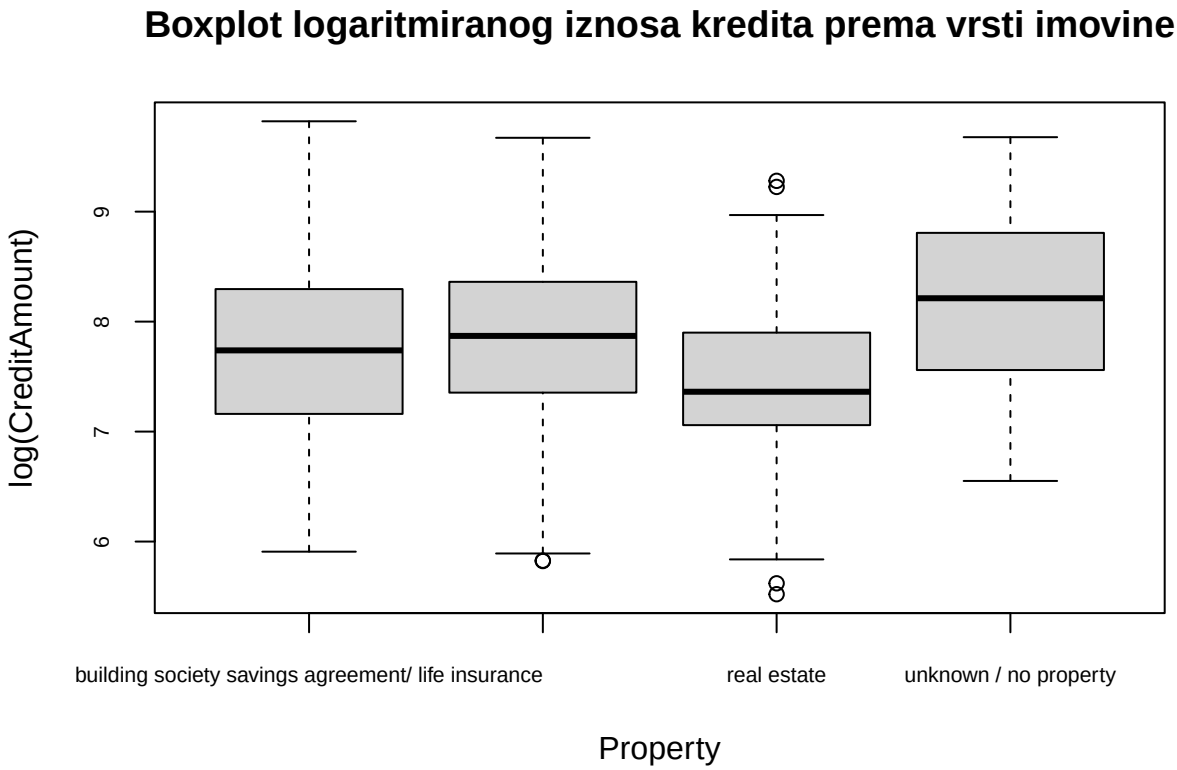
Vidimo da su pretpostavke o normalnosti i pretpostavka o homogenosti valjane ako pogledamo p-vrijednosti testova. Provedimo sada ANOVA test kako bi testirali našu osnovnu pretpostavku. Postavljamo hipoteze:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \text{barem dvije sredine nisu iste.}$$

```
a = aov(data$LogCreditAmount ~ data$Property)
summary(a)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## data$Property   3   63.1  21.024   38.83 <2e-16 ***
## Residuals     996  539.2   0.541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
boxplot(data$LogCreditAmount ~ data$Property,
    main = "Boxplot logaritmiranog iznosa kredita prema vrsti imovine",
    xlab = "Property", ylab = "log(CreditAmount)",
    cex.axis = 0.7, )
```



Na temelju p-vrijednosti zaključujemo kako na razini značajnosti od 5% možemo odbaciti nultu hipotezu, odnosno vidimo da postoje razlike u traženom iznosu kredita s obzirom na imovinu klijenta.

6. Zaključak

Ukupni zaključak cijelog rada je podijeljen. Puno smo vremena proveli u razmatranju pravog pristupa zadacima. Problem nam je predstavljala "realnost" podataka jer smo naviknuti na prilagođene podatke, podatke koji ne odstupaju, prema kojima lakše pretpostavljamo homogenost, normalnost... Ovdje to nije jednostavno, podatke treba filtrirati iprilagošavati za zadovoljavanje uvjeta testova. Iz tog su razloga testovi i njihovi zaključci ograničeni i treba ih uzeti sa zadrškom. Mnogo smo naučili iz ovog projekta i to će nam iskustvo zasigurno biti od pomoći ako se i kada budemo susretali s problemima ove vrste.