

# 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

1. Sadržaj
2. Osnovna prilagodba podataka
3. Možemo li temeljem drugih dostupnih varijabli predvidjeti hoće li nastupiti “default” za određenog klijenta? Koje varijable povežavaju tu vjerojatnost?
4. Jesu li muškarci skloniji nesipunjavanja obaveza po kreditu od žena?
5. Postoje li razlike u traženom iznosu klijenta prema imovini klijenta?
6. Zaključak

### 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", "...")
)

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?

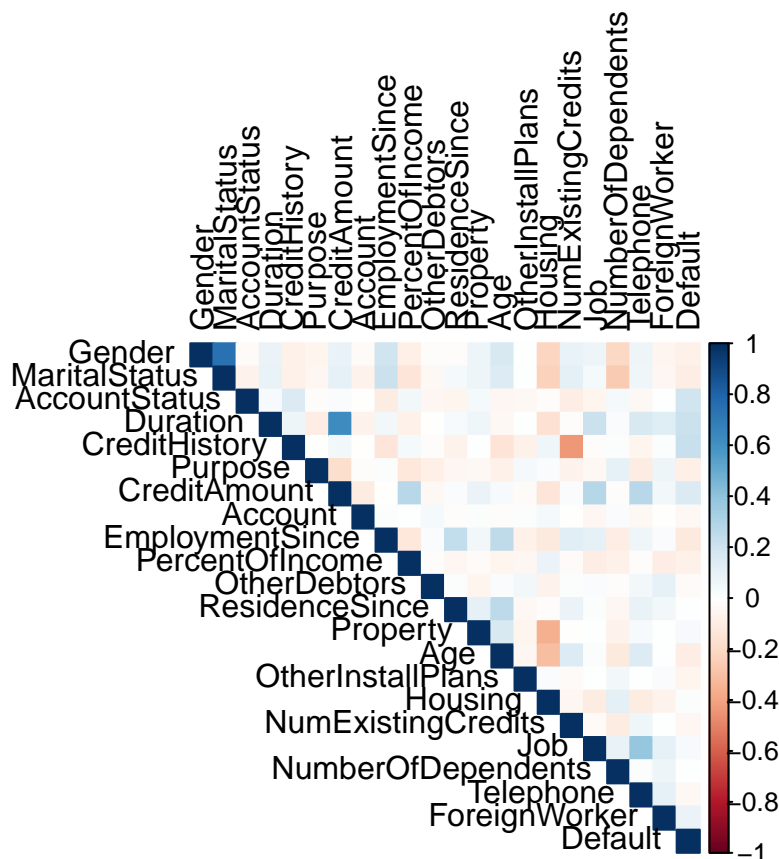
Računamo i prikazujemo matricu korelacije. Želimo vidjeti kako se pojedinačne varijable posebno koreliraju s varijablom “default”, stoga ćemo zasebno nacrtati taj grafikon. Cilj je pronaći i bolje istražiti varijable koje imaju veći utjecaj na konačni ishod varijable “default”:

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

Ispisujemo matricu korelacije:

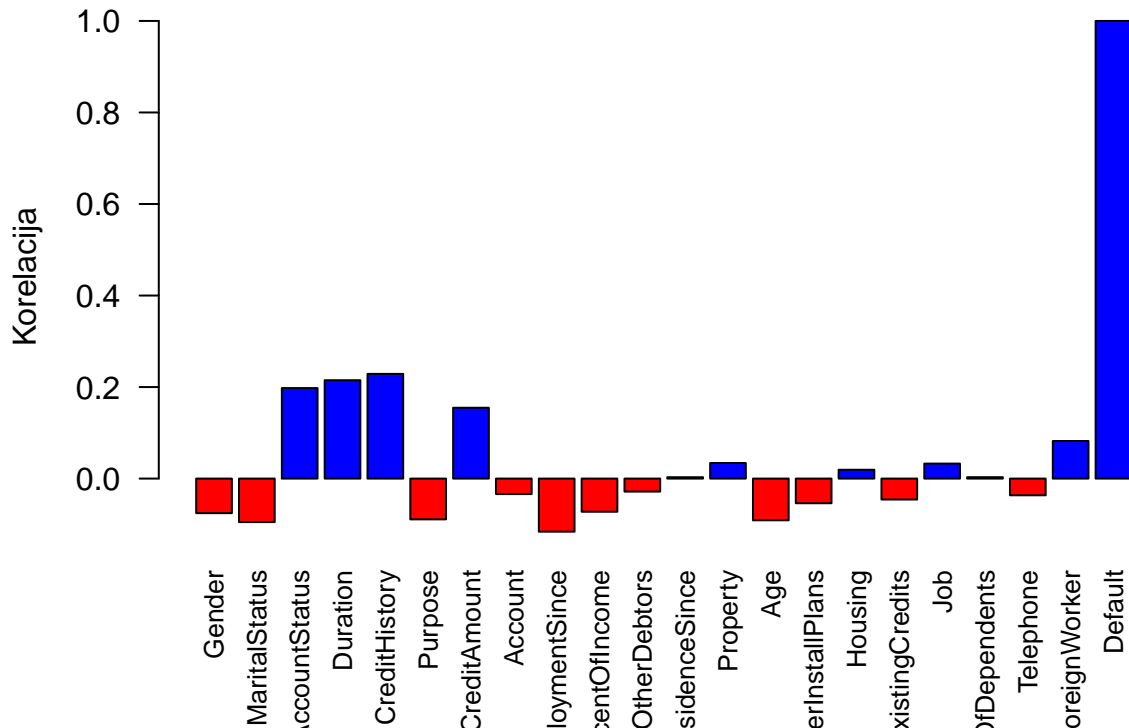
```
corr_matrix <- cor(data.frame(lapply(data, function(x) as.numeric(x))))
corrplot(corr_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 90)
```



```
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")
```

## Korelacije s varijablom "Default"



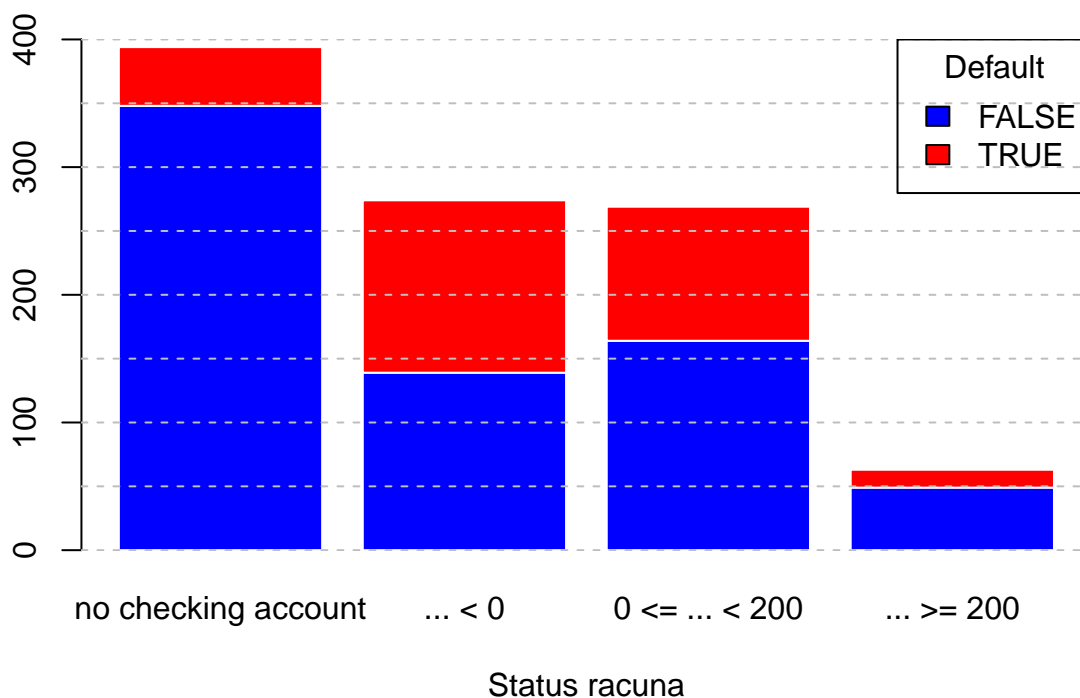
```
filt_cor_with_target <- cor_with_target[abs(cor_with_target) >= 0.1 & cor_with_target != 1]
print(names(filt_cor_with_target))
```

```
## [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")
```

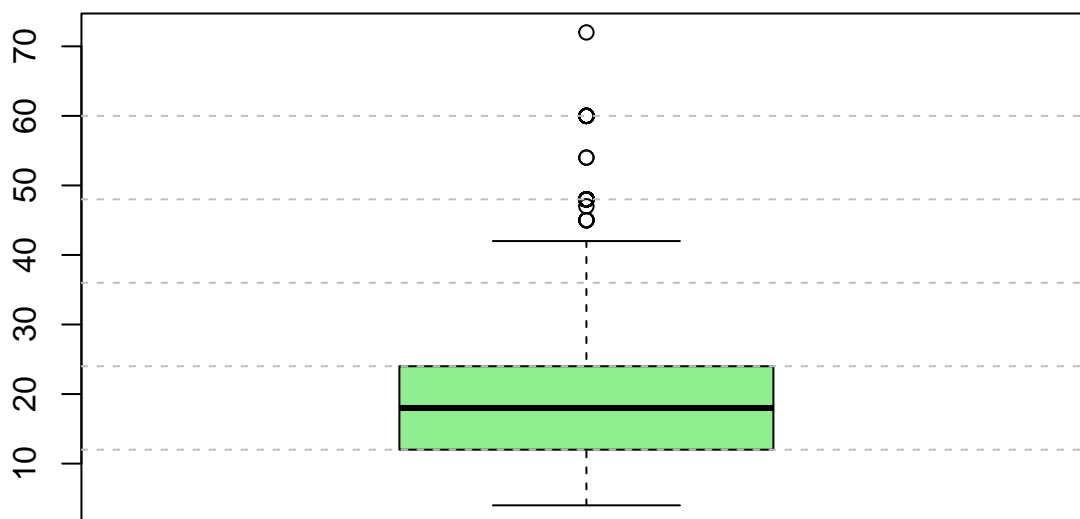
## Broj "defaultova" prema statusu racuna



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)
```

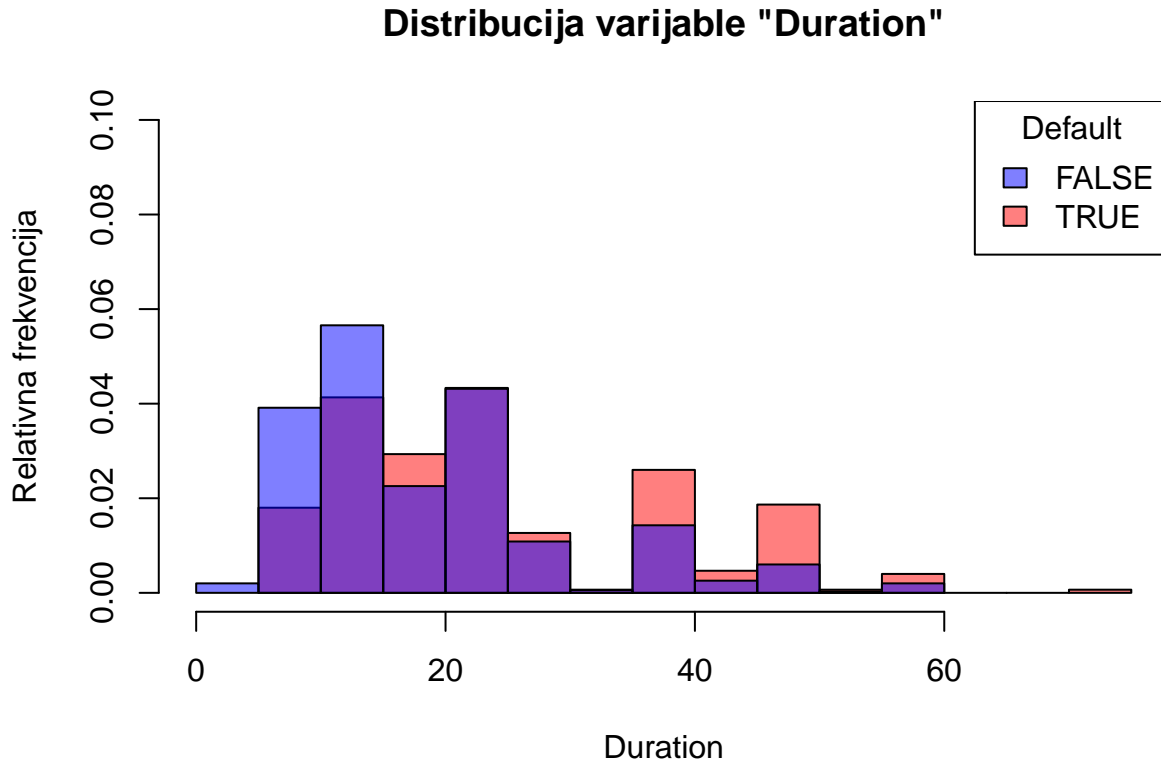
## Boxplot varijable "Duration"



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



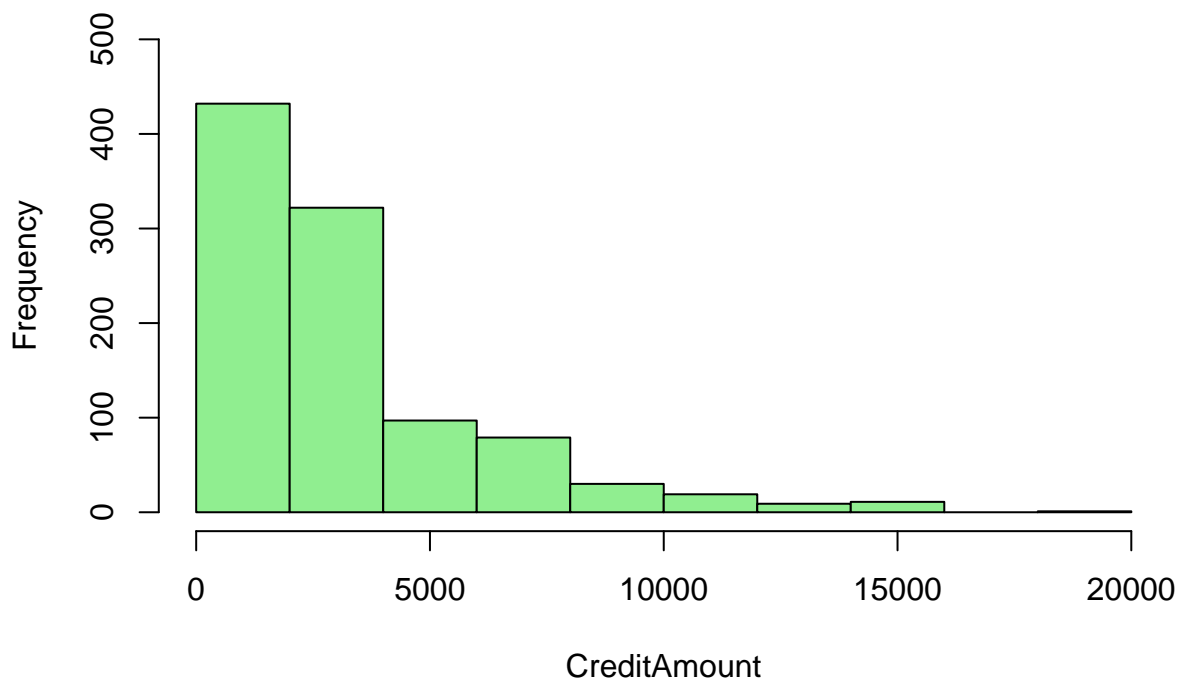
```
hist(defaulted$Duration, breaks = 10, xlim = c(0, 75), ylim = c(0, 0.1), freq = FALSE, col = rgb(1, 0, 0),
      ylab = "Relativna frekvencija", main = 'Distribucija varijable "Duration"')
hist(not_defaulted$Duration, breaks = 10, freq = FALSE, xlim = c(0, 75), col = rgb(0, 0, 1, 0.5), add =
   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



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 upotrijebili tu varijablu, koristit ćemo “CreditAmount”:

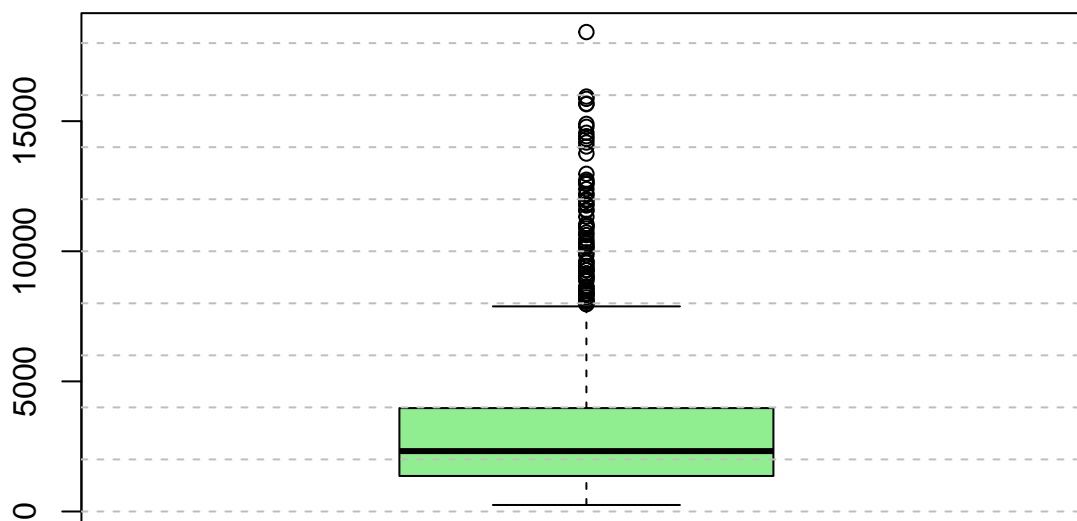
```
hist(CreditAmount, col = 'lightgreen', ylim = c(0, 500), main = 'Histogram varijable "CreditAmount"')
```

## Histogram varijable "CreditAmount"



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

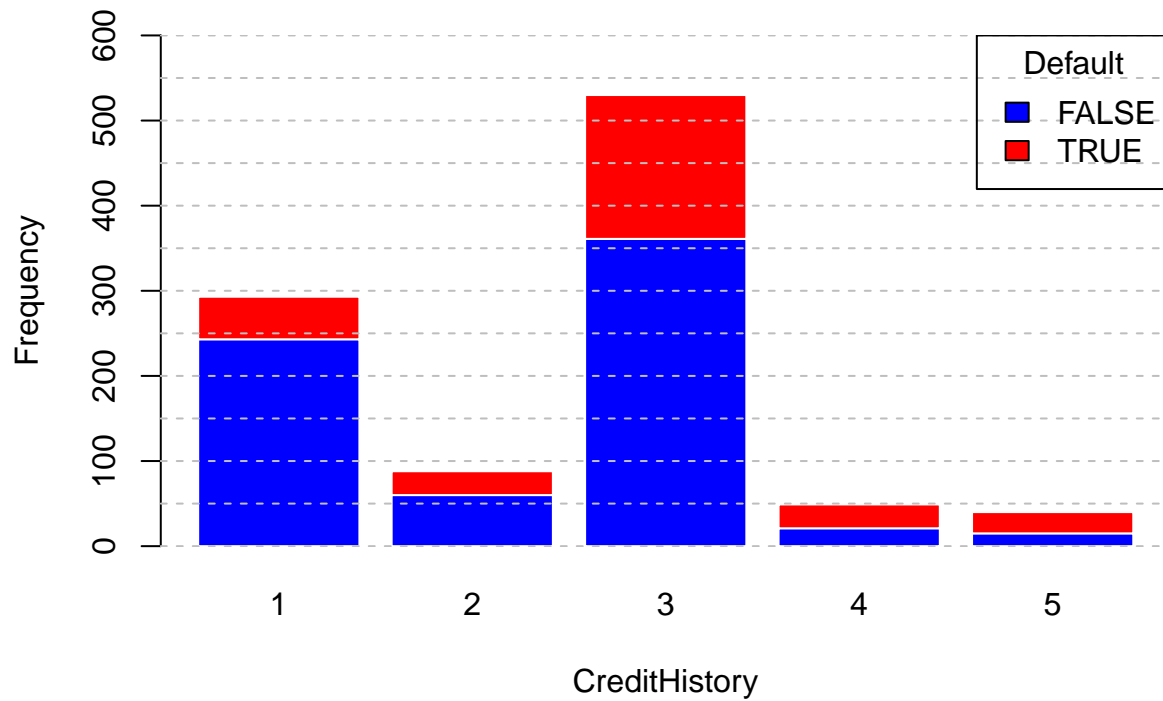
## Boxplot varijable "CreditAmount"



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")
```

## Broj "defaultova" prema varijabli "CreditHistory"

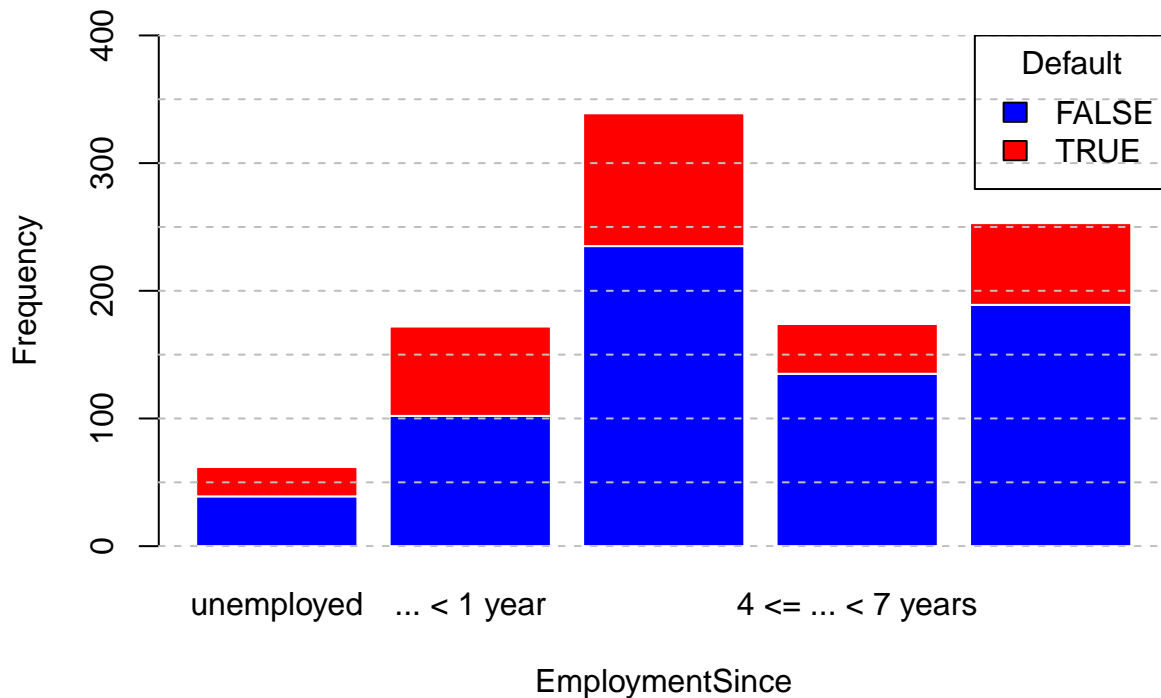


Opis stupaca grafa:

- 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, 600))
abline(h = seq(0, 400, by = 50), col = "gray", lty = 2)
legend("topright", legend = levels(Default), fill = c('blue', 'red'), title = "Default")
```

## Broj "defaultova" prema godinama radnog staza



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 određenog 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 -> Default, data = data, family = binomial())
summary(logreg.mdl)
```

```
##
```

```
## Call:
```

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

```
##
```

```
## Deviance Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```

## -2.2809 -0.6999 -0.3695 0.6869 2.7790
##
## Coefficients: (1 not defined because of singularities)
##
## (Intercept) -4.520e+00
## AccountStatus... < 0 1.780e+00
## AccountStatus0 <= ... < 200 1.397e+00
## AccountStatus... >= 200 8.061e-01
## Duration 2.801e-02
## CreditHistorydelay in paying off in the past 5.461e-01
## CreditHistoryexisting credits paid back duly till now 9.285e-01
## CreditHistoryall credits at this bank paid back duly 1.665e+00
## CreditHistoryno credits taken/ all credits paid back duly 1.496e+00
## Purposecar (new) 6.888e-01
## Purposecar (used) -9.717e-01
## Purposedomestic appliances 1.779e-01
## Purposeeducation 8.018e-01
## Purposefurniture/equipment -5.930e-02
## Purposeothers -7.966e-01
## Purposeradio/television -1.855e-01
## Purposerepairs 5.285e-01
## Purposeretraining -1.242e+00
## CreditAmount 1.233e-04
## Account... < 100 9.732e-01
## Account100 <= ... < 500 6.094e-01
## Account500 <= ... < 1000 6.068e-01
## Account... >= 1000 -4.872e-01
## EmploymentSince... < 1 year 6.662e-02
## EmploymentSince1 <= ... < 4 years -2.293e-01
## EmploymentSince4 <= ... < 7 years -7.634e-01
## EmploymentSince... >= 7 years -2.213e-01
## PercentOfIncome20% <= ... < 25% -3.109e-01
## PercentOfIncome25% <= ... < 35% -6.727e-01
## PercentOfIncome... >= 35% -9.369e-01
## Gendermale 2.616e-01
## MaritalStatusdivorced/separated/married NA
## MaritalStatusmarried/widowed -3.764e-01
## MaritalStatussingle -8.427e-01
## OtherDebtorsguarantor -1.416e+00
## OtherDebtorsnone -4.329e-01
## ResidenceSince1 <= ... < 4 years 7.613e-01
## ResidenceSince4 <= ... < 7 years 5.246e-01
## ResidenceSince... >= 7 years 3.885e-01
## Propertycar or other, not in attribute Account -1.091e-01
## Propertyreal estate -2.698e-01
## Propertyunknown / no property 4.669e-01
## Age -1.279e-02
## OtherInstallPlansnone -6.475e-01
## OtherInstallPlansstores -8.884e-02
## Housingown 1.729e-01
## Housingrent 6.303e-01
## NumExistingCredits2 || 3 4.050e-01
## NumExistingCredits4 || 5 2.741e-01
## NumExistingCredits... >= 6 4.550e-01

```

## Jobunskilled - resident	4.416e-01
## Jobskilled employee / official	4.694e-01
## Jobmanagement/ self-employed/highly qualified employee/ officer	3.691e-01
## NumberOfDependents... >= 3	-2.628e-01
## Telephoneyes, registered under the customers name	-2.848e-01
## ForeignWorkeryes	1.461e+00
##	Std. Error
## (Intercept)	1.321e+00
## AccountStatus... < 0	2.358e-01
## AccountStatus0 <= ... < 200	2.358e-01
## AccountStatus... >= 200	3.852e-01
## Duration	9.448e-03
## CreditHistorydelay in paying off in the past	3.392e-01
## CreditHistoryexisting credits paid back duly till now	2.724e-01
## CreditHistoryall credits at this bank paid back duly	4.506e-01
## CreditHistoryno credits taken/ all credits paid back duly	4.452e-01
## Purposecar (new)	3.377e-01
## Purposecar (used)	4.456e-01
## Purposedomestic appliances	8.182e-01
## Purposeeducation	4.677e-01
## Purposefurniture/equipment	3.570e-01
## Purposeothers	8.137e-01
## Purposeradio/television	3.402e-01
## Purposerepairs	5.945e-01
## Purposeretraining	1.202e+00
## CreditAmount	4.502e-05
## Account... < 100	2.661e-01
## Account100 <= ... < 500	3.557e-01
## Account500 <= ... < 1000	4.520e-01
## Account... >= 1000	5.811e-01
## EmploymentSince... < 1 year	4.396e-01
## EmploymentSince1 <= ... < 4 years	4.212e-01
## EmploymentSince4 <= ... < 7 years	4.596e-01
## EmploymentSince... >= 7 years	4.236e-01
## PercentOfIncome20% <= ... < 25%	2.551e-01
## PercentOfIncome25% <= ... < 35%	2.341e-01
## PercentOfIncome... >= 35%	3.047e-01
## Gendermale	3.885e-01
## MaritalStatusdivorced/separated/married	NA
## MaritalStatusmarried/widowed	4.576e-01
## MaritalStatussingle	3.813e-01
## OtherDebtorsguarantor	5.719e-01
## OtherDebtorsnone	4.127e-01
## ResidenceSince1 <= ... < 4 years	2.994e-01
## ResidenceSince4 <= ... < 7 years	3.359e-01
## ResidenceSince... >= 7 years	3.029e-01
## Propertycar or other, not in attribute Account	2.336e-01
## Propertyreal estate	2.551e-01
## Propertyunknown / no property	4.180e-01
## Age	9.317e-03
## OtherInstallPlansnone	2.403e-01
## OtherInstallPlansstores	4.166e-01
## Housingown	4.607e-01
## Housingrent	4.854e-01

## NumExistingCredits2    3	2.456e-01
## NumExistingCredits4    5	6.087e-01
## NumExistingCredits... >= 6	1.072e+00
## Jobunskilled - resident	6.867e-01
## Jobskilled employee / official	6.625e-01
## Jobmanagement/ self-employed/highly qualified employee/ officer	6.708e-01
## NumberOfDependents... >= 3	2.518e-01
## Telephoneyes, registered under the customers name	2.031e-01
## ForeignWorkeryes	6.265e-01
##	z value
## (Intercept)	-3.422
## AccountStatus... < 0	7.547
## AccountStatus0 <= ... < 200	5.923
## AccountStatus... >= 200	2.093
## Duration	2.965
## CreditHistorydelay in paying off in the past	1.610
## CreditHistoryexisting credits paid back duly till now	3.409
## CreditHistoryall credits at this bank paid back duly	3.695
## CreditHistoryno credits taken/ all credits paid back duly	3.360
## Purposecar (new)	2.040
## Purposecar (used)	-2.181
## Purposedomestic appliances	0.217
## Purposeeducation	1.714
## Purposefurniture/equipment	-0.166
## Purposeothers	-0.979
## Purposeradio/television	-0.545
## Purposerepairs	0.889
## Purposeretraining	-1.034
## CreditAmount	2.740
## Account... < 100	3.657
## Account100 <= ... < 500	1.713
## Account500 <= ... < 1000	1.343
## Account... >= 1000	-0.838
## EmploymentSince... < 1 year	0.152
## EmploymentSince1 <= ... < 4 years	-0.544
## EmploymentSince4 <= ... < 7 years	-1.661
## EmploymentSince... >= 7 years	-0.523
## PercentOfIncome20% <= ... < 25%	-1.219
## PercentOfIncome25% <= ... < 35%	-2.874
## PercentOfIncome... >= 35%	-3.075
## Gendermale	0.673
## MaritalStatusdivorced/separated/married	NA
## MaritalStatusmarried/widowed	-0.823
## MaritalStatussingle	-2.210
## OtherDebtorsguarantor	-2.476
## OtherDebtorsnone	-1.049
## ResidenceSince1 <= ... < 4 years	2.543
## ResidenceSince4 <= ... < 7 years	1.562
## ResidenceSince... >= 7 years	1.282
## Propertycar or other, not in attribute Account	-0.467
## Propertyreal estate	-1.058
## Propertyunknown / no property	1.117
## Age	-1.373
## OtherInstallPlansnone	-2.694

## OtherInstallPlansstores	-0.213
## Housingown	0.375
## Housingrent	1.299
## NumExistingCredits2    3	1.649
## NumExistingCredits4    5	0.450
## NumExistingCredits... >= 6	0.424
## Jobunskilled - resident	0.643
## Jobskilled employee / official	0.709
## Jobmanagement/ self-employed/highly qualified employee/ officer	0.550
## NumberOfDependents... >= 3	-1.044
## Telephoneyes, registered under the customers name	-1.402
## ForeignWorkeryes	2.333
##	Pr(> z )
## (Intercept)	0.000622 ***
## AccountStatus... < 0	4.45e-14 ***
## AccountStatus0 <= ... < 200	3.15e-09 ***
## AccountStatus... >= 200	0.036373 *
## Duration	0.003028 **
## CreditHistorydelay in paying off in the past	0.107362
## CreditHistoryexisting credits paid back duly till now	0.000653 ***
## CreditHistoryall credits at this bank paid back duly	0.000220 ***
## CreditHistoryno credits taken/ all credits paid back duly	0.000780 ***
## Purposecar (new)	0.041386 *
## Purposecar (used)	0.029211 *
## Purposedomestic appliances	0.827872
## Purposeeducation	0.086459 .
## Purposefurniture/equipment	0.868070
## Purposeothers	0.327552
## Purposeradio/television	0.585437
## Purposerepairs	0.374052
## Purposeretraining	0.301348
## CreditAmount	0.006153 **
## Account... < 100	0.000255 ***
## Account100 <= ... < 500	0.086644 .
## Account500 <= ... < 1000	0.179389
## Account... >= 1000	0.401801
## EmploymentSince... < 1 year	0.879539
## EmploymentSince1 <= ... < 4 years	0.586261
## EmploymentSince4 <= ... < 7 years	0.096714 .
## EmploymentSince... >= 7 years	0.601303
## PercentOfIncome20% <= ... < 25%	0.222947
## PercentOfIncome25% <= ... < 35%	0.004050 **
## PercentOfIncome... >= 35%	0.002106 **
## Gendermale	0.500728
## MaritalStatusdivorced/separated/married	NA
## MaritalStatusmarried/widowed	0.410748
## MaritalStatussingle	0.027106 *
## OtherDebtorsguarantor	0.013303 *
## OtherDebtorsnone	0.294219
## ResidenceSince1 <= ... < 4 years	0.010985 *
## ResidenceSince4 <= ... < 7 years	0.118342
## ResidenceSince... >= 7 years	0.199687
## Propertycar or other, not in attribute Account	0.640384
## Propertyreal estate	0.290201



```
## Propertyunknown / no property 0.264061
## Age 0.169876
## OtherInstallPlansnone 0.007056 **
## OtherInstallPlansstores 0.831134
## Housingown 0.707338
## Housingrent 0.194111
## NumExistingCredits2 || 3 0.099170 .
## NumExistingCredits4 || 5 0.652475
## NumExistingCredits... >= 6 0.671370
## Jobunskilled - resident 0.520167
## Jobskilled employee / official 0.478594
## Jobmanagement/ self-employed/highly qualified employee/ officer 0.582115
## NumberOfDependents... >= 3 0.296625
## Telephoneyes, registered under the customers name 0.160870
## ForeignWorkeryes 0.019658 *
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
```

```
## Null deviance: 1221.7 on 999 degrees of freedom
```

```
## Residual deviance: 887.4 on 945 degrees of freedom
```

```
## AIC: 997.4
```

```
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
odds_ratios <- exp(coef(logreg.mdl))
print(odds_ratios)
```

```
## (Intercept)
## 0.01088555
## AccountStatus... < 0
## 5.92985675
## AccountStatus0 <= ... < 200
## 4.04124311
## AccountStatus... >= 200
## 2.23926854
## Duration
## 1.02840788
## CreditHistorydelay in paying off in the past
## 1.72652476
## CreditHistoryexisting credits paid back duly till now
## 2.53074040
## CreditHistoryall credits at this bank paid back duly
## 5.28416202
## CreditHistoryno credits taken/ all credits paid back duly
## 4.46254976
## Purposecar (new)
## 1.99128577
## Purposecar (used)
## 0.37844358
## Purposedomestic appliances
## 1.19471579
## Purposeeducation
## 2.22958351
```

```

##          Purposefurniture/equipment
##          0.94242187
##          Purposeothers
##          0.45084611
##          Purposeradio/television
##          0.83065669
##          Purposerepairs
##          1.69632708
##          Purposeretraining
##          0.28877108
##          CreditAmount
##          1.00012333
##          Account... < 100
##          2.64639538
##          Account100 <= ... < 500
##          1.83928964
##          Account500 <= ... < 1000
##          1.83463962
##          Account... >= 1000
##          0.61432760
##          EmploymentSince... < 1 year
##          1.06888562
##          EmploymentSince1 <= ... < 4 years
##          0.79512525
##          EmploymentSince4 <= ... < 7 years
##          0.46608807
##          EmploymentSince... >= 7 years
##          0.80146202
##          PercentOfIncome20% <= ... < 25%
##          0.73282168
##          PercentOfIncome25% <= ... < 35%
##          0.51030914
##          PercentOfIncome... >= 35%
##          0.39185003
##          Gendermale
##          1.29903604
##          MaritalStatusdivorced/separated/married
##          NA
##          MaritalStatusmarried/widowed
##          0.68634086
##          MaritalStatussingle
##          0.43053671
##          OtherDebtorsguarantor
##          0.24275492
##          OtherDebtorsnone
##          0.64862056
##          ResidenceSince1 <= ... < 4 years
##          2.14112517
##          ResidenceSince4 <= ... < 7 years
##          1.68980799
##          ResidenceSince... >= 7 years
##          1.47471405
##          Propertycar or other, not in attribute Account
##          0.89660877

```

```
## Propertyreal estate
## 0.76351409
## Propertyunknown / no property
## 1.59497702
## Age
## 0.98729303
## OtherInstallPlansnone
## 0.52332692
## OtherInstallPlansstores
## 0.91499292
## Housingown
## 1.18880442
## Housingrent
## 1.87813665
## NumExistingCredits2 || 3
## 1.49923311
## NumExistingCredits4 || 5
## 1.31533987
## NumExistingCredits... >= 6
## 1.57614275
## Jobunskilled - resident
## 1.55516689
## Jobskilled employee / official
## 1.59903715
## Jobmanagement/ self-employed/highly qualified employee/ officer
## 1.44647276
## NumberOfDependents... >= 3
## 0.76889433
## Telephoneyes, registered under the customers name
## 0.75217185
## ForeignWorkeryes
## 4.31206829
```

```
Rsq = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
Rsq
```

```
## [1] 0.2736555
```

```
coef_table <- data.frame(
  Variable = names(coef(logreg.mdl)),
  Coefficient = coef(logreg.mdl),
  OddsRatio = exp(coef(logreg.mdl))
)

significant_vars <- coef_table[coef_table$`Pr(>|z|)` < 0.05, ]

print(significant_vars)
```

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

```
set.seed(123)
```

```
formula <- Default ~ AccountStatus + Duration + CreditHistory + Purpose + CreditAmount + Account + Empl

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

```

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

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

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

```

#### 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 prema banci:

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

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

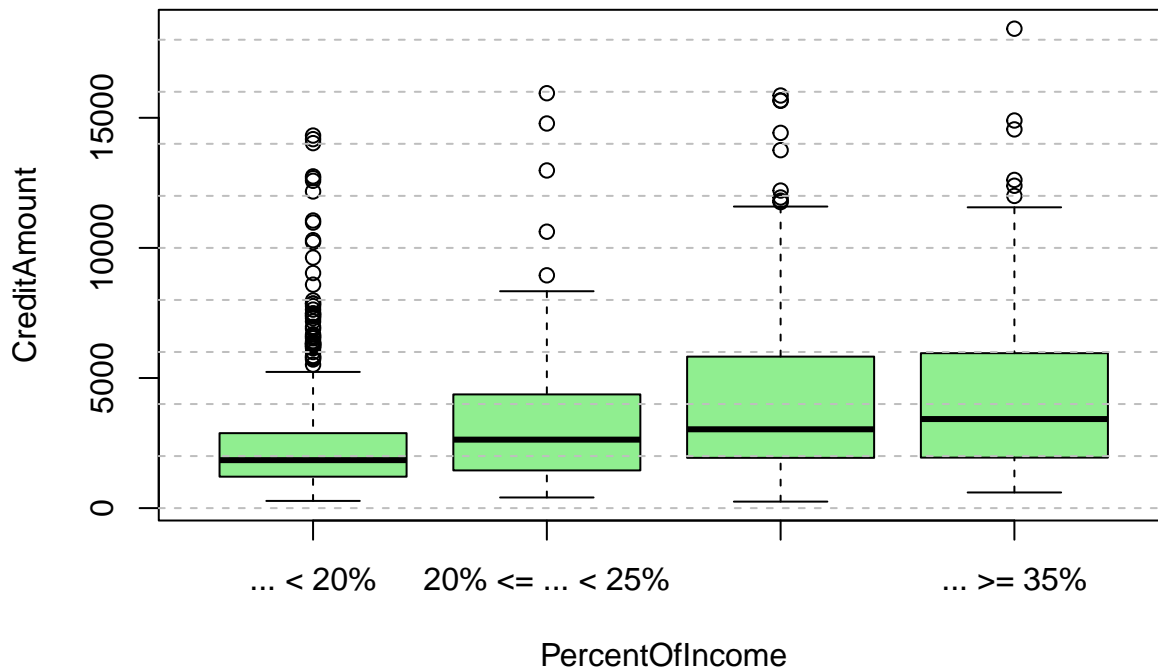
Postotak žena koje nisu redovito ispunjavali obaveze prema banci:

```
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:

```
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.

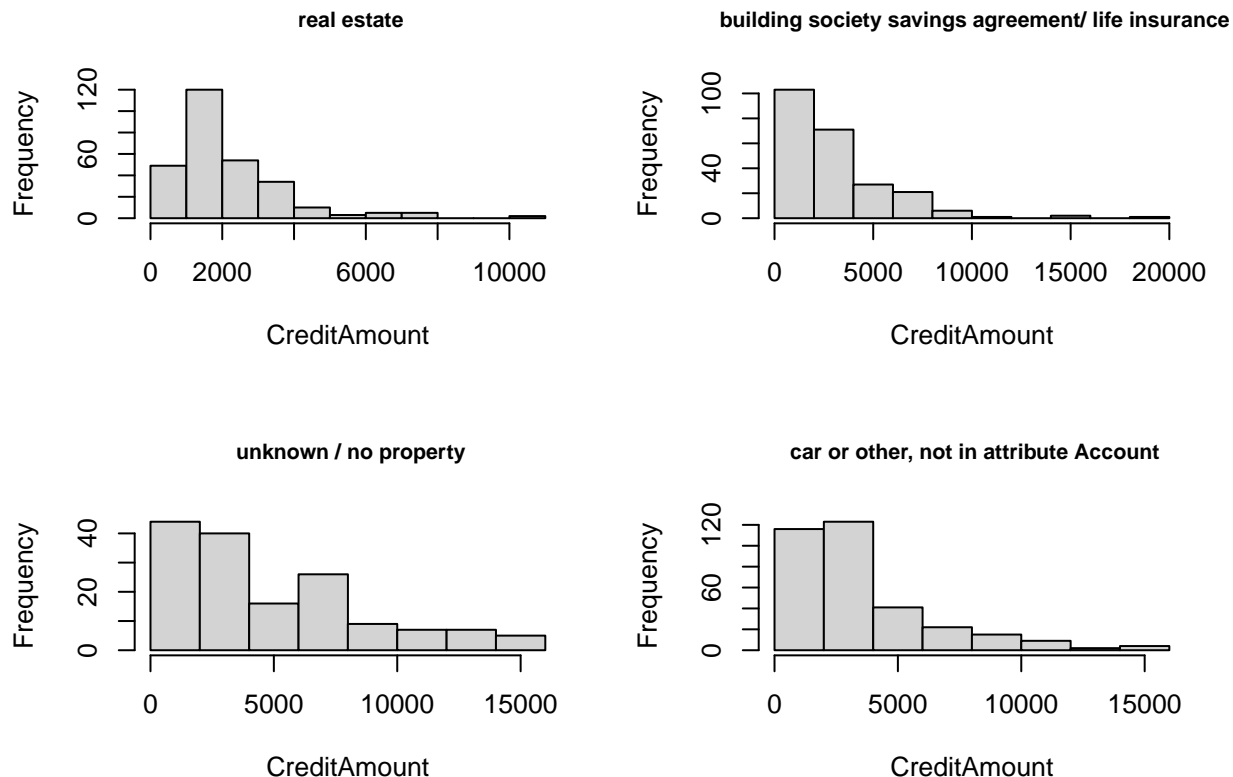
## 5. Postoje li razlike u traženom iznosu klijenta prema imovini klijenta?

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:

```
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", ylab = "Frequency")
}
```



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

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

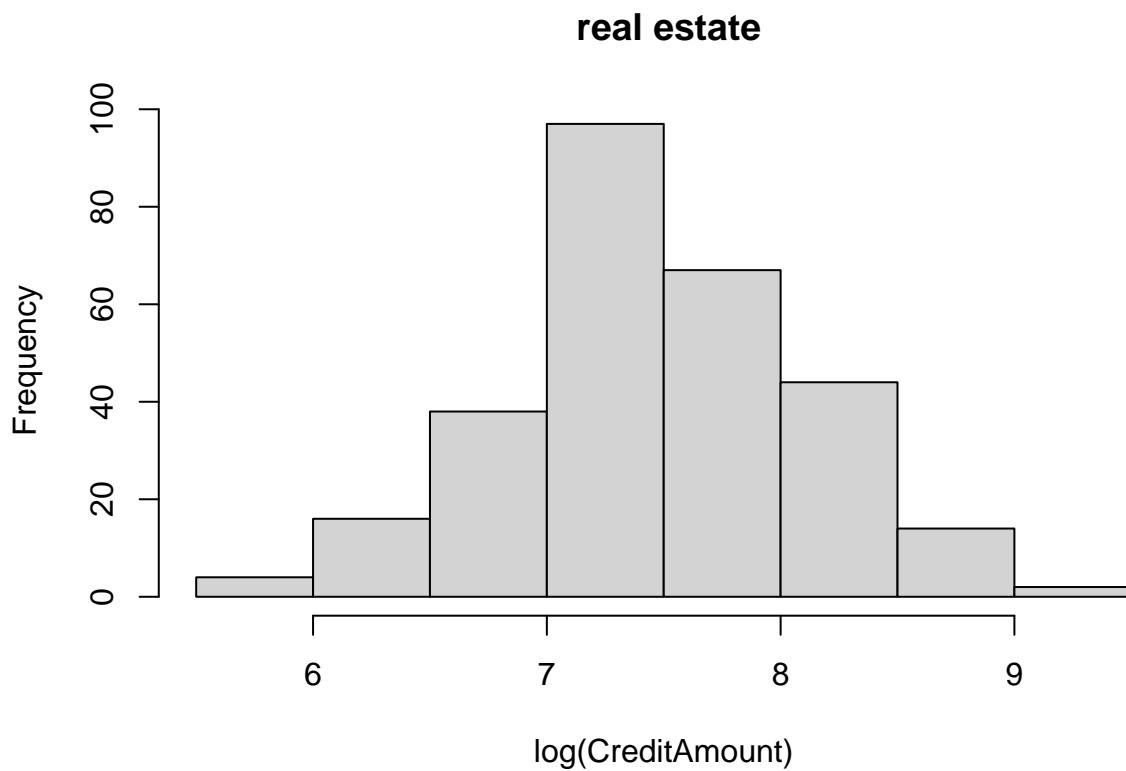
Histogrami logaritmiranog iznosa kredita obzirom na vrstu imovine:

```
require(nortest)

## Loading required package: nortest
data$LogCreditAmount <- log(data$CreditAmount)

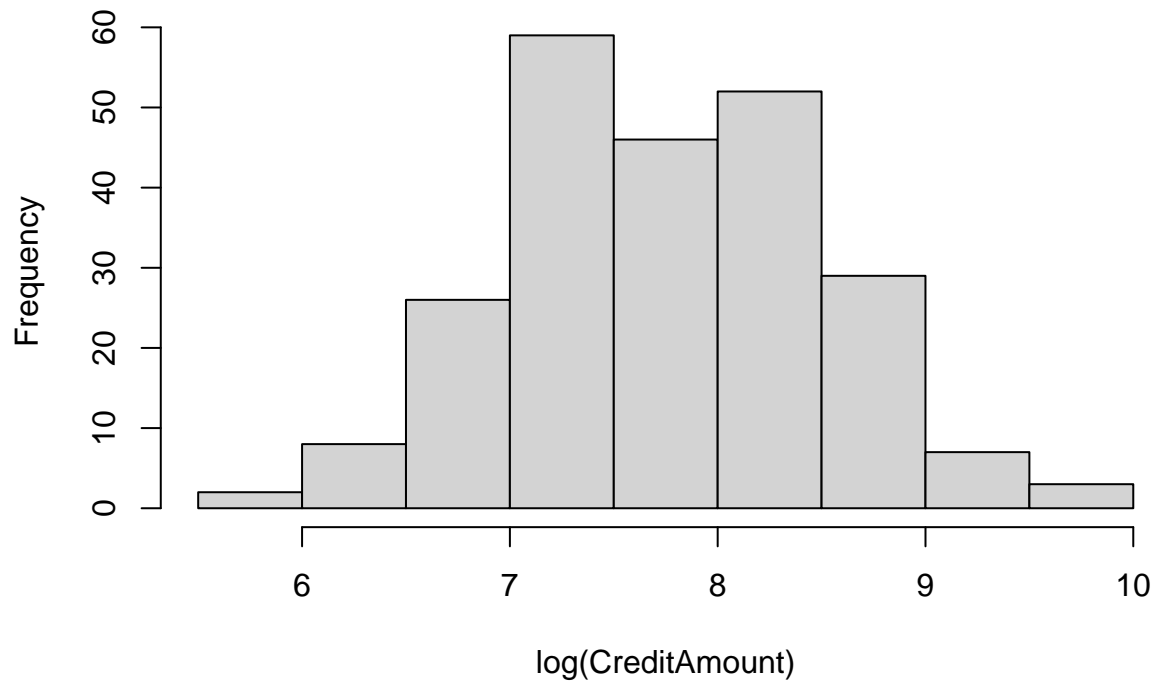
for (item in unique(data$Property)) {
  print(lillie.test(data$LogCreditAmount[data$Property == item]))
  hist(data$LogCreditAmount[data$Property == item], main = paste("", item), xlab = 'log(CreditAmount)')
}

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



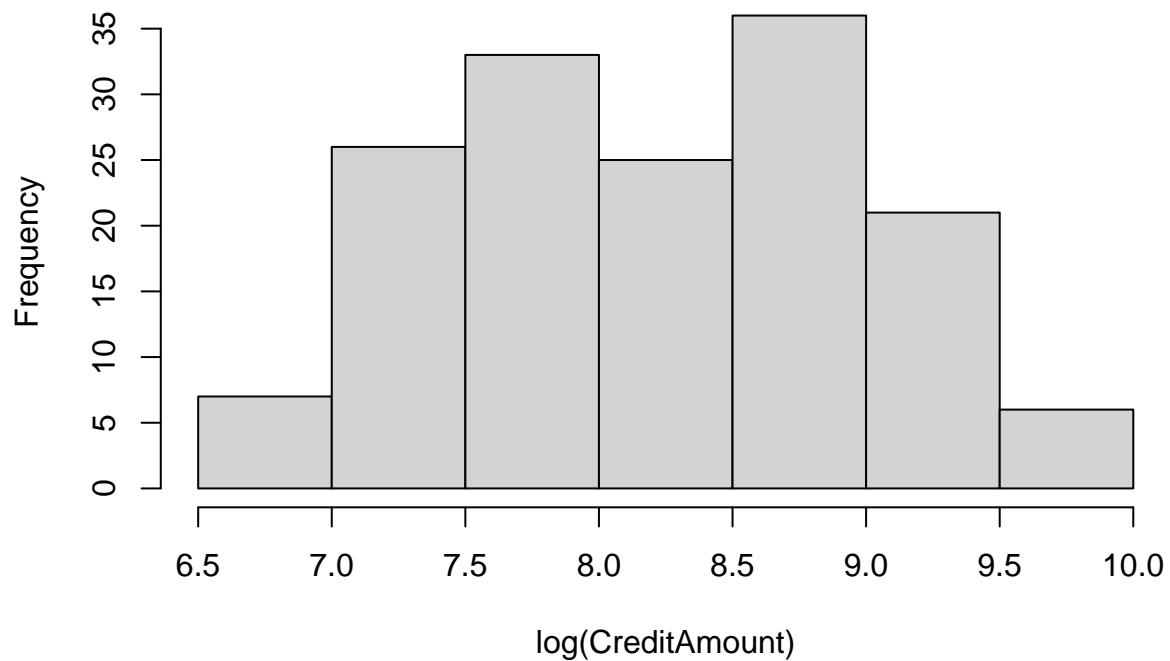
```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: data$LogCreditAmount[data$Property == item]
## D = 0.061582, p-value = 0.03277
```

## building society savings agreement/ life insurance



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

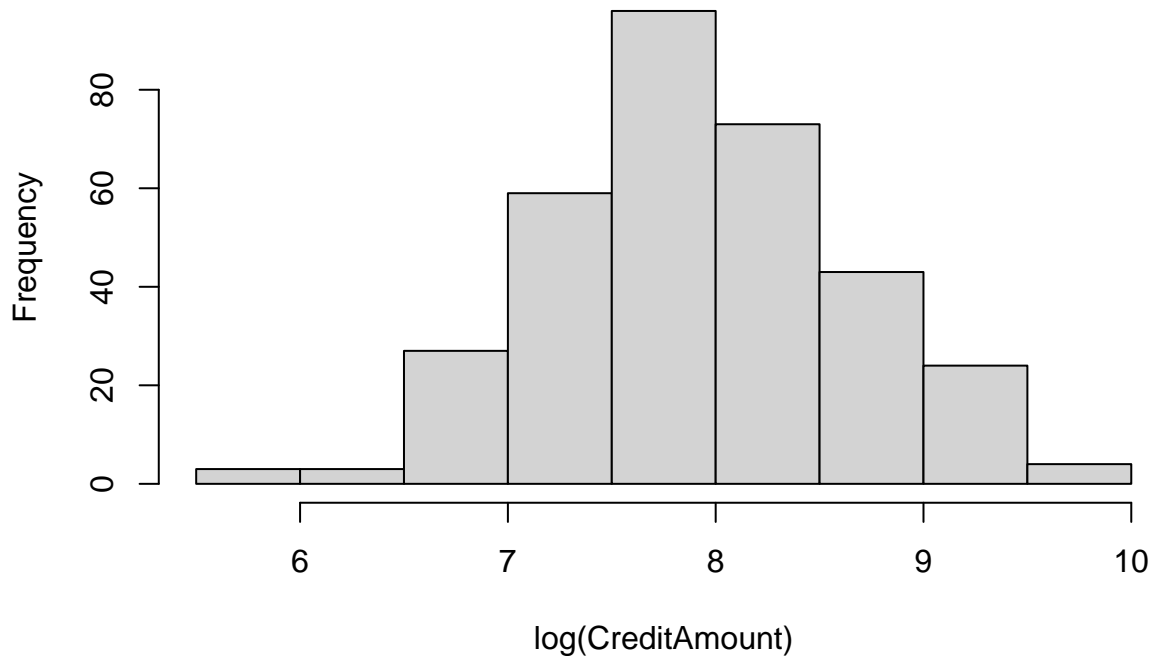
## unknown / no property





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

### car or other, not in attribute Account



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

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1 : \text{barem dvije varijance nisu iste.}$$

```
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:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \text{barem dvije sredine nisu iste.}$$

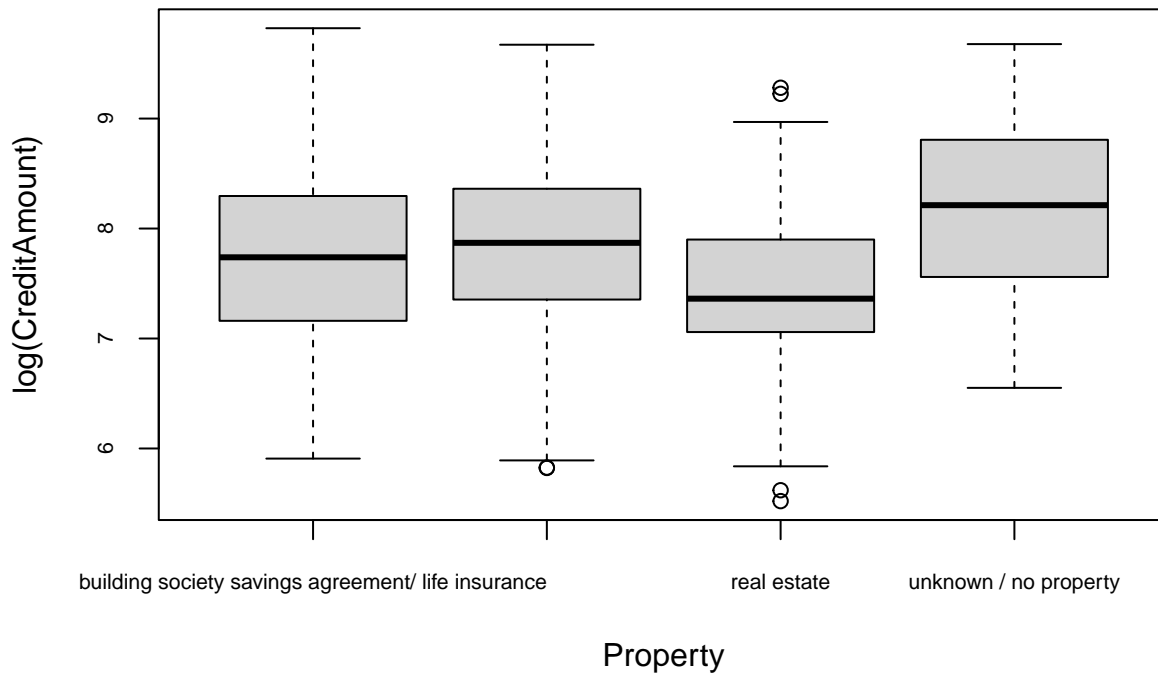
```
a = aov(data$LogCreditAmount ~ data$Property)
summary(a)
```

```
##              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

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

## Boxplot logaritmiranog iznosa kredita prema vrsti imovine



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