SAP - projekt

Procjena kreditnog rizika

Vedran Knežević, Andre Medvedić, Jan Celin, Ante Čavar

20.10.2021.

Uvod

U našem projektu obrađujemo veliki skup podataka kreditnog stanja korisnika neke banke. Naš je zadatak procijentiti koji čimbenici utječu na sposobnost otplate kredita u zadanome roku.

Skup podataka - statistika

```
data <- read.csv("procjena_kreditnog_rizika.csv")


cat("number of missing values: ", sum(is.na(data[,])), "\n")
```

```
## number of missing values:  0
```

```
data$AccountStatus <- factor(
  data$AccountStatus, levels = c(
    "no checking account",
    "... < 0",
    "0 <= ... < 200",
    "... >= 200")
)
data$CreditHistory <- factor(
  data$CreditHistory, levels = c(
    "delay in paying off in the past",
    "critical account/ other credits existing (not at this bank)",
    "no credits taken/ all credits paid back duly",
    "existing credits paid back duly till now",
    "all credits at this bank paid back duly"
  )
)
data$Purpose <- factor(data$Purpose)
data$Account <- factor(
  data$Purpose, 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$PercentOfIncome, levels = c(
    "... < 20%",
    "20% <= ... < 25%",
    "25% <= ... < 35%",
    "... >= 35%"
  )
)
data$OtherDebtors <- factor(
  data$OtherDebtors, levels = c(
    "none",
    "guarantor",
    "co-applicant"
  )
)
data$ResidenceSince <- factor(
  data$ResidenceSince, levels = c(
    "... < 1 year",
    "1 <= ... < 4 years",
    "4 <= ... < 7 years",
    "... >= 7 years"
  )
)
data$Property <- factor(
  data$Property, levels = c(
    "unknown / no property",
    "building society savings agreement/ life insurance",
    "car or other, not in attribute Account",
    "real estate"
  )
)
data$OtherInstallPlans <- factor(
  data$OtherInstallPlans, levels = c(
    "none",
    "stores",
    "bank"
  )
)
data$Housing <- factor(
  data$Housing, levels = c(
    "for free",
    "rent",
    "own"
  )
)
```

```r
data$NumExistingCredits <- factor(
  data$NumExistingCredits, levels = c(
    "1",
    "2 or 3",
    "4 or 5",
    "above 6"
  )
)
data$Job <- factor(
  data$Job, levels = c(
    "unemployed/ unskilled - non-resident",
    "unskilled - resident",
    "management/ self-employed/highly qualified employee/ officer",
    "skilled employee / official"
  )
)
data$NumberOfDependents <- factor(
  data$NumberOfDependents, levels = c(
    "less than 3",
    "3 or more"
  )
)
data$Telephone <- factor(
  data$Telephone, levels = c(
    "none",
    "yes, registered under the customers name"
  )
)
data$ForeignWorker <- factor(
  data$ForeignWorker, levels = c(
    "no",
    "yes"
  )
)
summary(data)
```

```
##              AccountStatus    Duration
##  no checking account:394   Min.   : 4.0
##  ... < 0            :274   1st Qu.:12.0
##  0 <= ... < 200     :269   Median :18.0
##  ... >= 200         : 63   Mean   :20.9
##                            3rd Qu.:24.0
##                            Max.   :72.0
##
##                                                        CreditHistory
##  delay in paying off in the past                               : 88
##  critical account/ other credits existing (not at this bank):293
##  no credits taken/ all credits paid back duly                  : 40
##  existing credits paid back duly till now                      :530
##  all credits at this bank paid back duly                       : 49
##
##
##                 Purpose     CreditAmount                            Account
##  radio/television   :280   Min.   :  250   Unknown / no savings account:   0
```

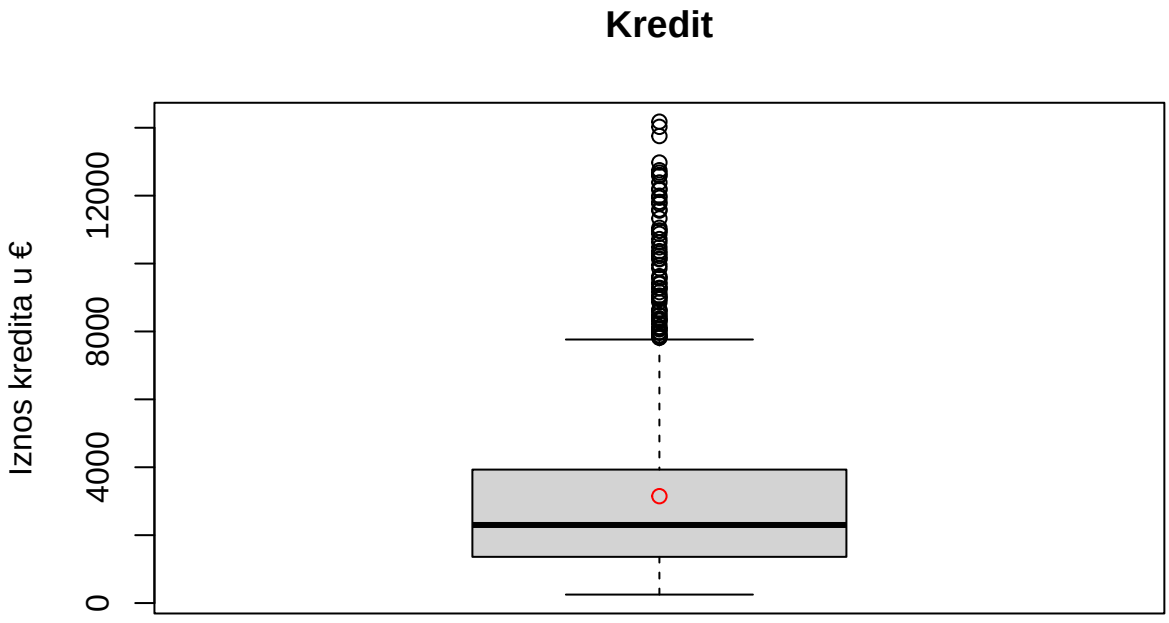
```
##  car (new)          :234   1st Qu.: 1366   ... < 100                  :   0
##  furniture/equipment:181   Median : 2320   100 <= ... < 500           :   0
##  car (used)         :103   Mean   : 3271   500 <= ... < 1000          :   0
##  business           : 97   3rd Qu.: 3972   ... >= 1000                :   0
##  education          : 50   Max.   :18424   NA's                       :1000
##  (Other)            : 55
##           EmploymentSince         PercentOfIncome PersonalStatus
##  unemployed        : 62    ... < 20%       :476    Length:1000
##  ... < 1 year      :172    20% <= ... < 25%:157    Class :character
##  1 <= ... < 4 years:339    25% <= ... < 35%:231    Mode  :character
##  4 <= ... < 7 years:174    ... >= 35%      :136
##  ... >= 7 years    :253
##
##
##        OtherDebtors            ResidenceSince
##  none        :907   ... < 1 year      :130
##  guarantor   : 52   1 <= ... < 4 years:308
##  co-applicant: 41   4 <= ... < 7 years:149
##                     ... >= 7 years    :  0
##                     NA's              :413
##
##
##                                                 Property        Age
##  unknown / no property                             :154   Min.   :19.00
##  building society savings agreement/ life insurance:232   1st Qu.:27.00
##  car or other, not in attribute Account            :332   Median :33.00
##  real estate                                       :282   Mean   :35.55
##                                                           3rd Qu.:42.00
##                                                           Max.   :75.00
##
##  OtherInstallPlans     Housing    NumExistingCredits
##  none  :814        for free:108   1      :633
##  stores: 47        rent    :179   2 or 3 :333
##  bank  :139        own     :713   4 or 5 : 28
##                                   above 6:  6
##
##
##
##                                                        Job
##  unemployed/ unskilled - non-resident                    : 22
##  unskilled - resident                                    :200
##  management/ self-employed/highly qualified employee/ officer:148
##  skilled employee / official                             :630
##
##
##
##    NumberOfDependents                                 Telephone
##  less than 3:155      none                                :596
##  3 or more  :845      yes, registered under the customers name:404
##
##
##
##
##
```

```
##  ForeignWorker    Default
##  no : 37       Min.   :0.0
##  yes:963       1st Qu.:0.0
##                Median :0.0
##                Mean   :0.3
##                3rd Qu.:1.0
##                Max.   :1.0
##
```

Vidimo da je skup poprilično čist (nema nedostajućih vrijednosti). Iako bi neki stupci moguće bili korisniji da su numerički prije nego kategorički.

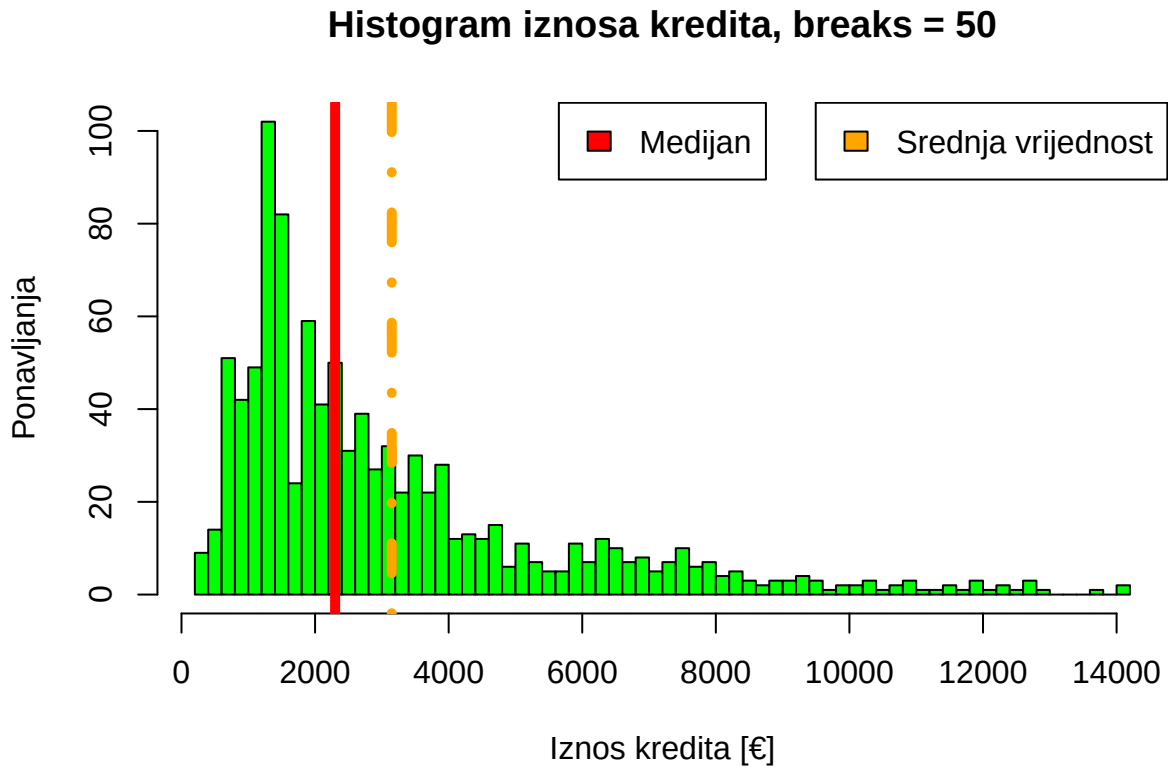
Uvodni grafovi

```
threshold <- quantile(data$CreditAmount, 0.99)  # Set threshold at 99th percentile
# Exclude data points above the threshold
filtered_data <- subset(data, CreditAmount <= threshold)
boxplot(filtered_data$CreditAmount, main="Kredit", ylab="Iznos kredita u €")
points(mean(filtered_data$CreditAmount), col = "red")
```

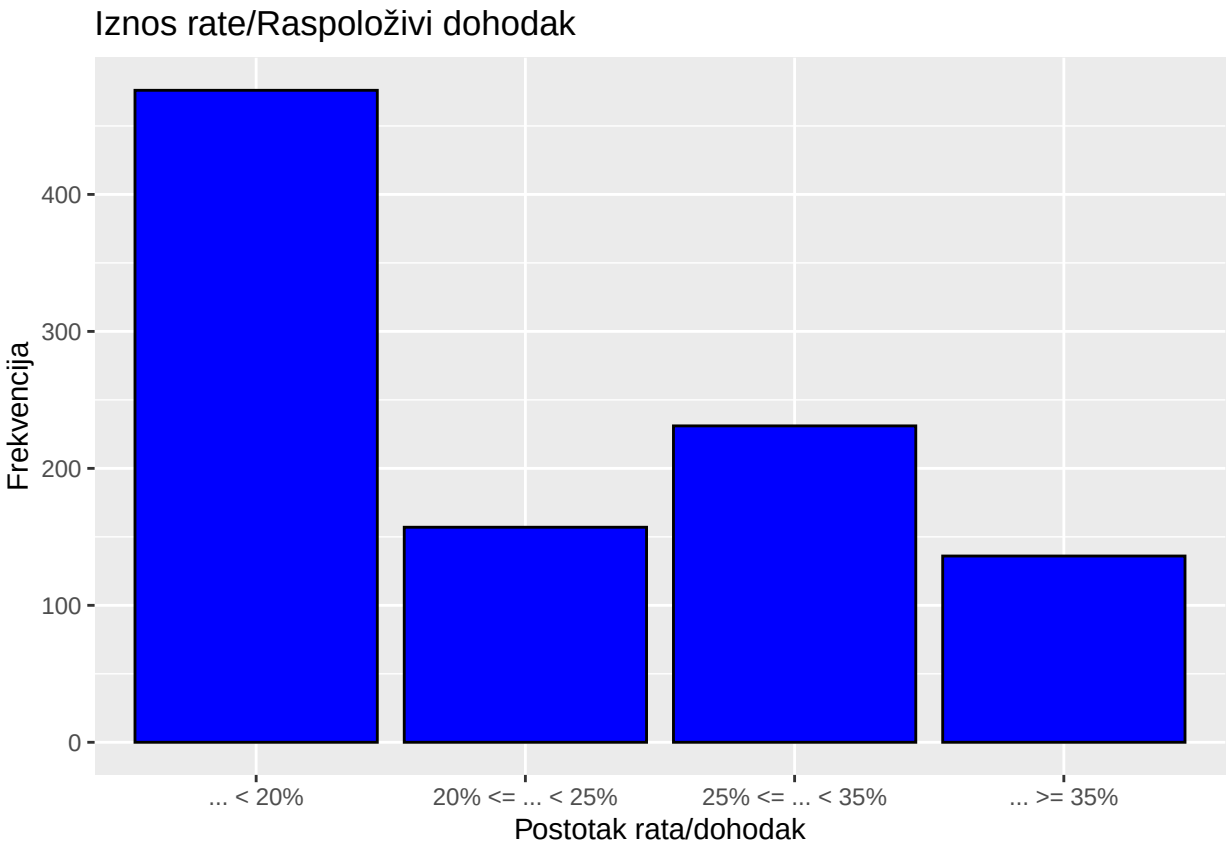



```
threshold <- quantile(data$CreditAmount, 0.99)  # Set threshold at 99th percentile
# Exclude data points above the threshold
filtered_data <- subset(data, CreditAmount <= threshold)
h = hist(filtered_data$CreditAmount,
         breaks = 50,
         main="Histogram iznosa kredita, breaks = 50",
         xlab="Iznos kredita [€]",
```

```
         ylab='Ponavljanja',
         col="green"
         )
legend("topright", legend = "Srednja vrijednost", fill = "orange")
legend("top", legend = "Medijan", fill = "red")
abline(v = mean(filtered_data$CreditAmount), col = "orange", lwd=5, lty=10)
abline(v = median(filtered_data$CreditAmount), col= "red", lwd=5)
```




```
# Create a bar plot
bar_plot <- ggplot(data, aes(x = PercentOfIncome)) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Iznos rate/Raspoloživi dohodak",
       x = "Postotak rata/dohodak",
       y = "Frekvencija")

# Print the plot
print(bar_plot)
```



TESTOVI

1.pitanje: Možemo li temeljem drugih dostupnih varijabli predvidjeti hoće li nastupiti *default* za određenog klijenta? Koje varijable povećavaju tu vjerojatnost?

U sljedećem odsječku uspoređujemo odnos između dviju kategorijskih varijabli: CreditHistory i Default. Cilj je testirati postoji li veza između nečije kreditne povijesti i toga ispunjavaju li kreditne obveze ili ne. H0: Varijable Default i CreditHistory su nezavisne.
H1: Varijable Default i CreditHistory su zavisne.

```
credit_history_default <- data.frame(category = data$CreditHistory %>% unique)

credit_history_default$no_default <- sapply(credit_history_default$category, function(x){
  nrow(data[data$CreditHistory == x & data$Default==0,])
})

credit_history_default$default <- sapply(credit_history_default$category, function(x){
  nrow(data[data$CreditHistory == x & data$Default==1,])
})

credit_history_matirx <- matrix(
  c(credit_history_default$no_default, credit_history_default$default), nrow = 2, byrow = T
)
rownames(credit_history_matirx) <- c("no_default", "default")
colnames(credit_history_matirx) <- credit_history_default$category
```

```r
chisq.test(credit_history_matirx, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  credit_history_matirx
## X-squared = 61.691, df = 4, p-value = 1.279e-12
```

Na temelju ovog testa na razini značajnosti $\alpha = 0.05$ odbijamo nultu hipotezu u korist alternativne hipoteze. Drugim riječima, zaključujemo da postoji zavisnost između varijabli Default i CreditHistory.

2.pitanje: Jesu li muškarci skloniji neispunjavanju obveza po kreditu od žena?

U ovom odsječku uspoređujemo odnos između dviju kategorijskih varijabli (spol, izvršavanje svojih novčanih obveza). Uspoređivat ćemo je li kod muškaraca i žena jednaka proporcija onih koji nisu izvršili svoje novčane obaveze (default).

Sve statistike provjeravamo na razina značajnosti $\alpha = 0.05$. Ispitujemo jednostranu alternativu (neispunjavanje obveza je češće kod muškaraca).

Statistika nad svim muškarcima i ženama u skupu podataka

H0: Proporcija onih koji nisu ispunili obveza naspram onih koji su ispunili obaveze jednaka je kod muškaraca i žena (ili je manja kod muškaraca).
H1: Proporcija osoba koje nisu ispunile obaveze naspram onih koji su ispunili obaveze veća je kod muškaraca.

```r
female_clients <- data[str_detect(data$PersonalStatus, "female"),]
male_clients <- data[!str_detect(data$PersonalStatus, "female"),]
num_female_default <- nrow(female_clients[female_clients$Default == 1,])
num_male_default <- nrow(male_clients[male_clients$Default == 1,])

proportion_matrix <- matrix(c(nrow(male_clients)-num_male_default,
                              num_male_default,
                            nrow(female_clients)-num_female_default,
                             num_female_default), nrow=2, byrow = T)
colnames(proportion_matrix) <- c("no_default", "default")
rownames(proportion_matrix) <- c("male", "female")
# proportion_matrix
prop.test(proportion_matrix, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  proportion_matrix
## X-squared = 5.3485, df = 1, p-value = 0.9896
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.000000  0.129814
## sample estimates:
##    prop 1    prop 2
## 0.7231884 0.6483871
```

Iz ovoga zaključujemo, na razini značajnosti 0.05, da muškarci ispunjavaju kreditne obveze razmjerno ženama (tj. ne možemo reći da su skloniji neispunjavanju obveza).

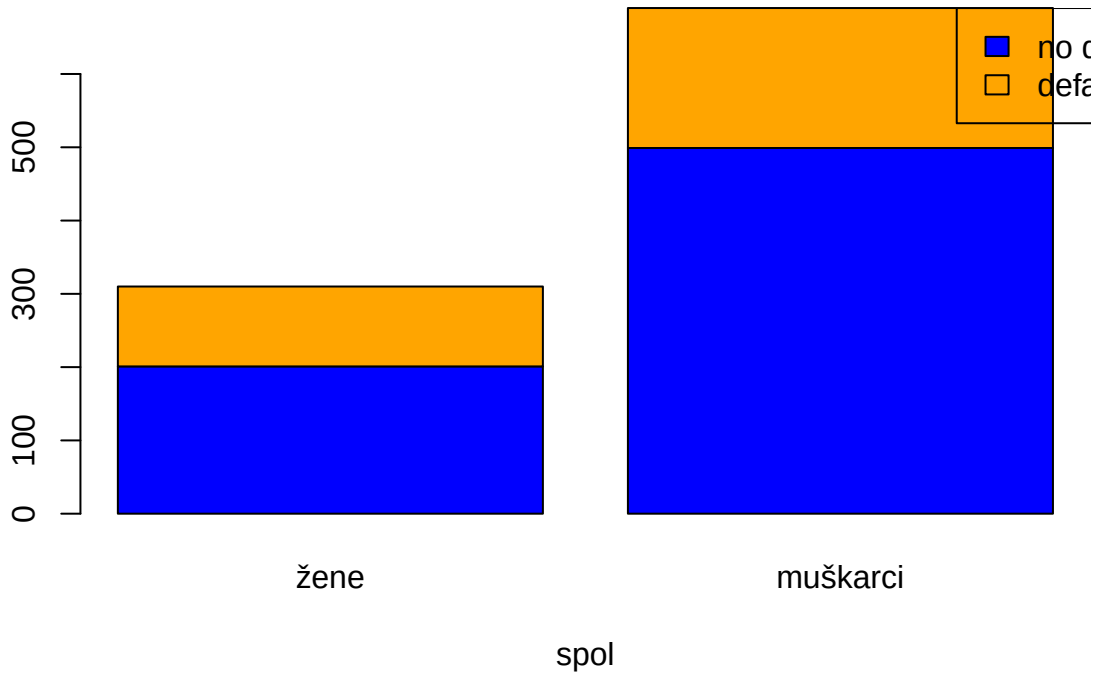
Provodimo Z-test o dvije proporcije s očekivanjem da će nam dati vrlo slične rezultate kao i χ^2-test.

```r
n1 <- nrow(male_clients)
n2 <- nrow(female_clients)
k1 <- n1 - num_male_default
k2 <- n2 - num_female_default

Z_stat <- (k1/n1-k2/n2)/sqrt(((k1+k2)/(n1+n2))*(1-(k1+k2)/(n1+n2))*(1/n1+1/n2))
cat("The p-value of the Z statistic is: ", pnorm(Z_stat))
```

```
## The p-value of the Z statistic is:  0.9915134
```

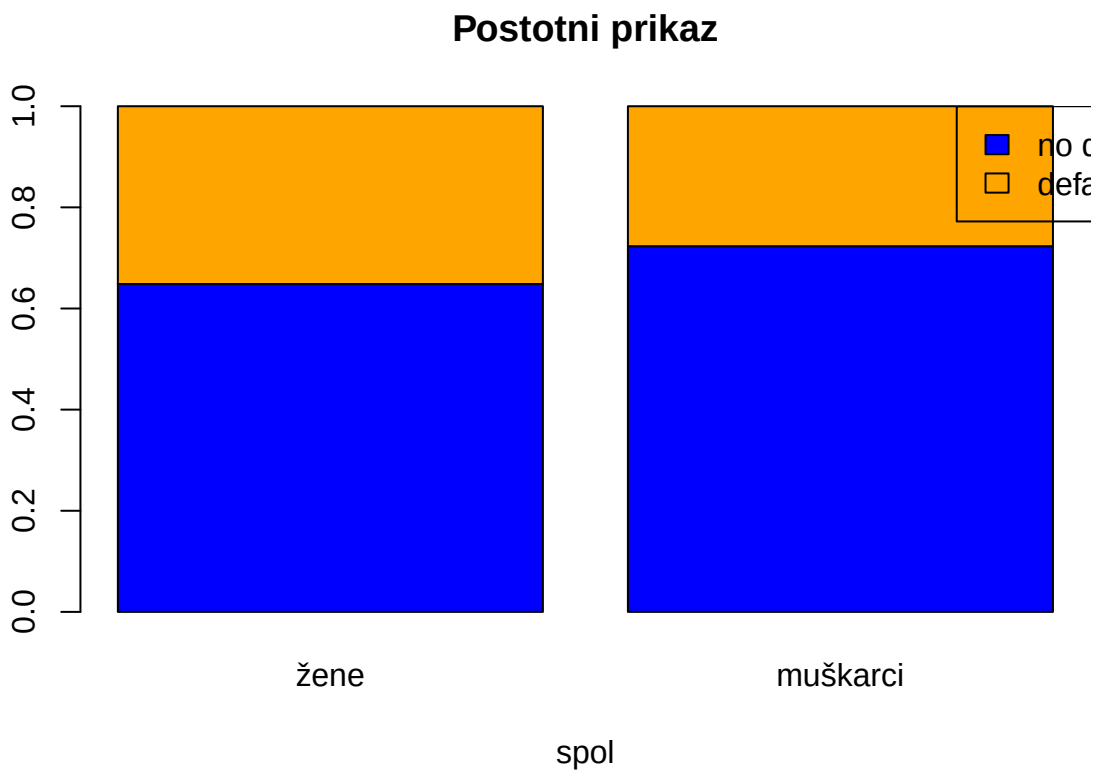
Kao što možemo uočiti Z-test nam daje isti zaključak i vrlo sličnu p-vrijednost kao i χ^2-test pa ćemo nadalje koristiti χ^2 jer je on implementiran u R-u.

```r
Values <- matrix(c((nrow(female_clients)-num_female_default),
                   (nrow(male_clients)-num_male_default),
                   num_female_default,
                   num_male_default
                     ), nrow=2, ncol=2, byrow = T)

barplot(Values, main="Kvantitativni prikaz", names.arg=c("žene", "muškarci"), xlab="spol", col = c("blu

legend("topright", inset = c(-0.1, 0), c("no default","default"), fill = c("blue", "orange"))
```




```r
Values <- matrix(c((nrow(female_clients)-num_female_default)/nrow(female_clients),
                   (nrow(male_clients)-num_male_default)/nrow(male_clients),
                   num_female_default/nrow(female_clients),
                   num_male_default/nrow(male_clients)
```

```
                    ), nrow=2, ncol=2, byrow = T)

barplot(Values, main="Postotni prikaz", names.arg=c("žene", "muškarci"), xlab="spol", col = c("blue", "

legend("topright", inset = c(-0.1, 0), c("no default","default"), fill = c("blue", "orange"))
```



Statistika nad slobodnim muškarcima i ženama u skupu podataka

H0: Proporcija onih koji nisu ispunili obveza naspram onih koji su ispunili obaveze jednaka je kod slobodnih muškaraca i žena (ili je manja kod slobodnih muškaraca).
H1: Proporcija osoba koje nisu ispunile obaveze naspram onih koji su ispunili obaveze veća je kod slobodnih muškaraca.

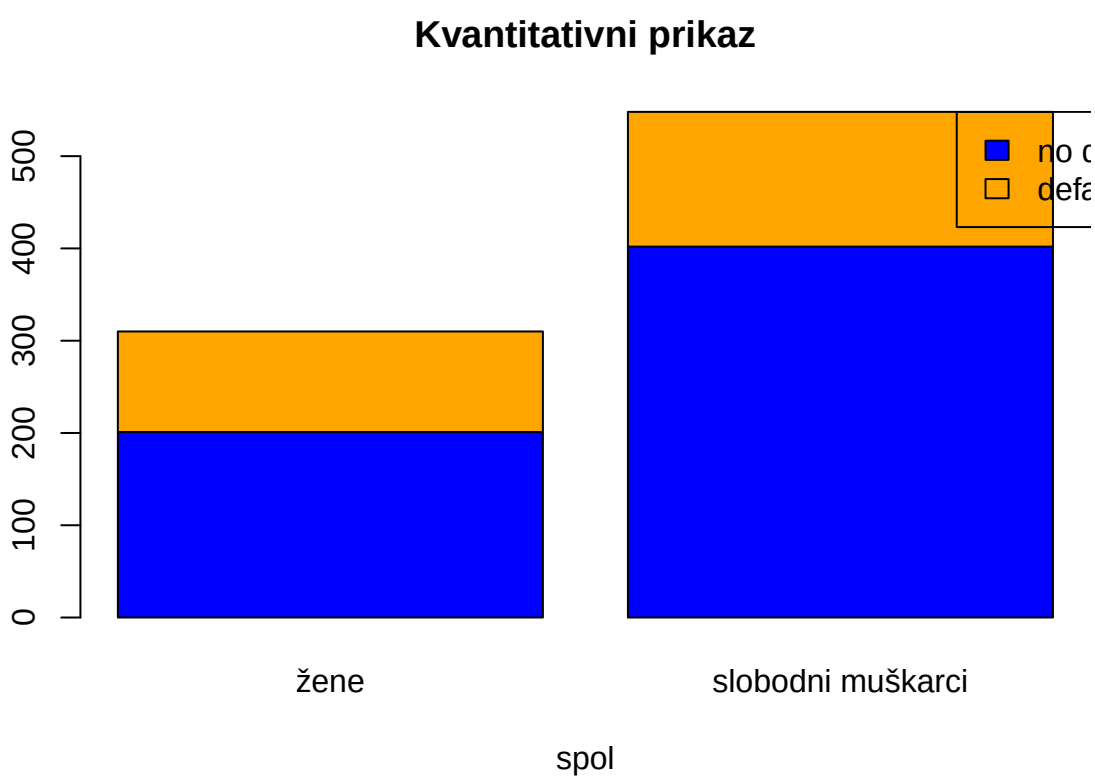
```
male_single_clients <- male_clients[str_detect(male_clients$PersonalStatus, "single"),]
num_male_single_default <- sum(male_single_clients$Default == 1)

proportion_matrix[1,] <- c(nrow(male_single_clients) - num_male_single_default,
                           num_male_single_default)
# proportion_matrix
prop.test(proportion_matrix, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  proportion_matrix
```

```
## X-squared = 6.4775, df = 1, p-value = 0.9945
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.1420715
## sample estimates:
##    prop 1    prop 2
## 0.7335766 0.6483871
```

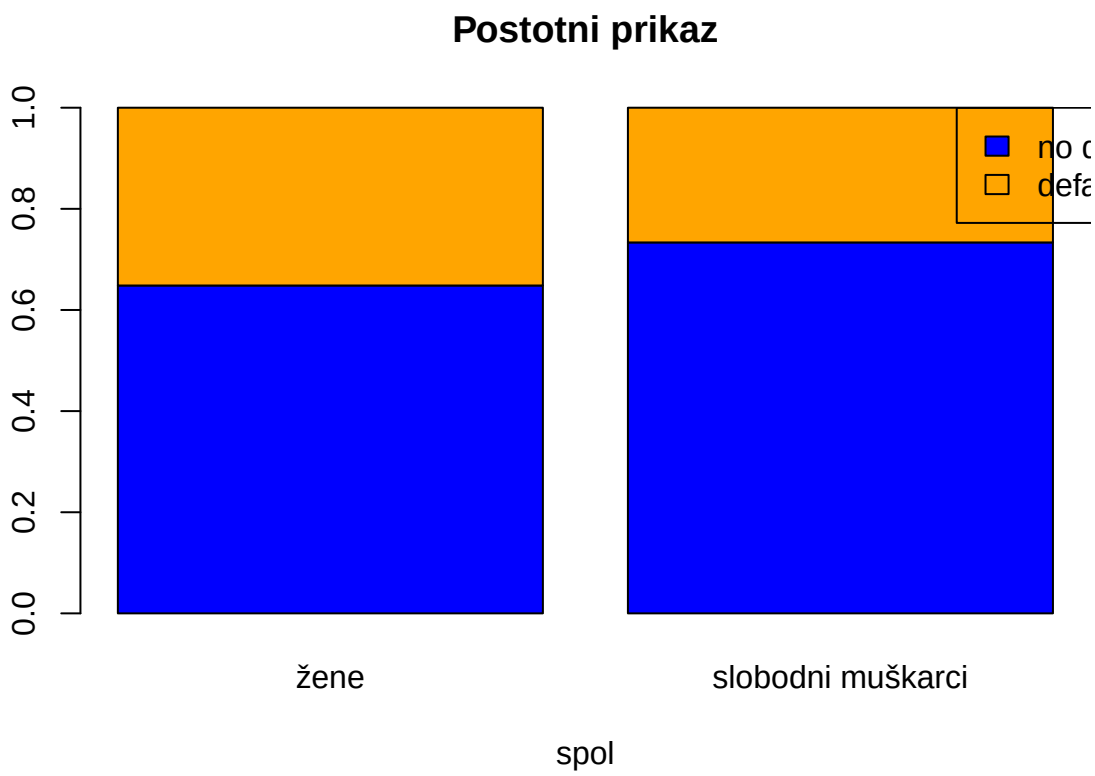
Iz ovoga zaključujemo, na razini značajnosti 0.05, da slobodni muškarci ispunjavaju kreditne obveze razmjerno ženama (tj. ne možemo reći da su skloniji neispunjavanju obveza).

```
Values <- matrix(c((nrow(female_clients)-num_female_default),
                   (nrow(male_single_clients)-num_male_single_default),
                   num_female_default,
                   num_male_single_default
                     ), nrow=2, ncol=2, byrow = T)

barplot(Values, main="Kvantitativni prikaz", names.arg=c("žene", "slobodni muškarci"), xlab="spol", col

legend("topright", inset = c(-0.1, 0), c("no default","default"), fill = c("blue", "orange"))
```




```
Values <- matrix(c((nrow(female_clients)-num_female_default)/nrow(female_clients),
                   (nrow(male_single_clients)-num_male_single_default)/nrow(male_single_clients),
                   num_female_default/nrow(female_clients),
                   num_male_single_default/nrow(male_single_clients)
                     ), nrow=2, ncol=2, byrow = T)
```

```r
barplot(Values, main="Postotni prikaz", names.arg=c("žene", "slobodni muškarci"), xlab="spol", col = c(

legend("topright", inset = c(-0.1, 0), c("no default", "default"), fill = c("blue", "orange"))
```



Statistika nad rastavljenim muškarcima i ženama u skupu podataka

H0: Proporcija onih koji nisu ispunili obveza naspram onih koji su ispunili obaveze jednaka je kod rastavljenih muškaraca i žena (ili je manja kod rastavljenih muškaraca).
H1: Proporcija osoba koje nisu ispunile obaveze naspram onih koji su ispunili obaveze veća je kod rastavljenih muškaraca.

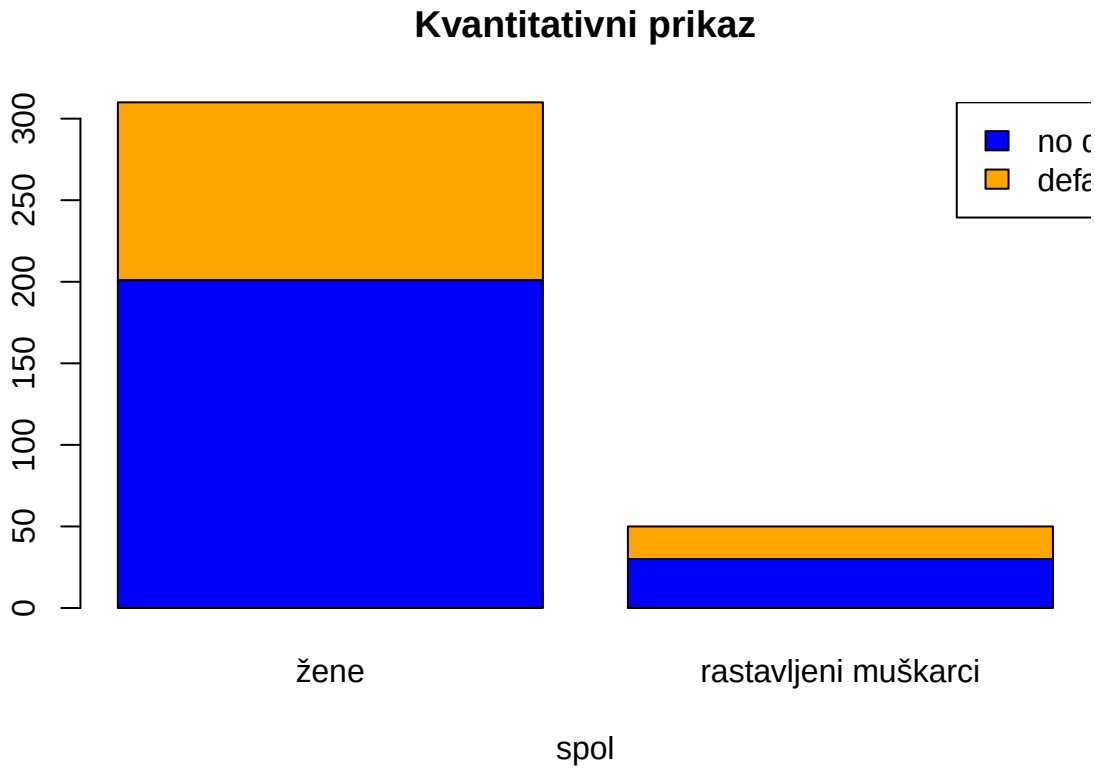
```r
male_divor_clients <- male_clients[str_detect(male_clients$PersonalStatus, "divorced"),]
num_male_divor_default <- sum(male_divor_clients$Default == 1)

proportion_matrix[1,] <- c(nrow(male_divor_clients) - num_male_divor_default,
                           num_male_divor_default)
# proportion_matrix
prop.test(proportion_matrix, alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  proportion_matrix
## X-squared = 0.25323, df = 1, p-value = 0.3074
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000  0.08560362
```

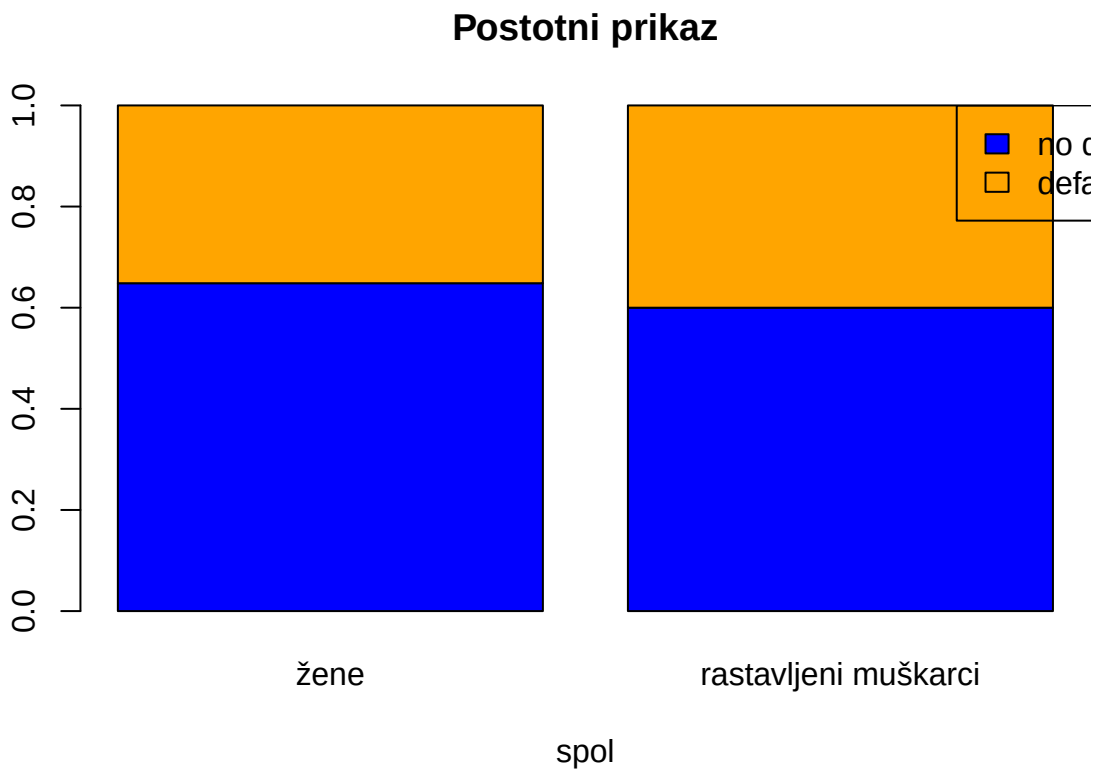
```
## sample estimates:
##    prop 1    prop 2
## 0.6000000 0.6483871
```

Iz ovoga zaključujemo, na razini značajnosti 0.05, da rastavljeni muškarci ispunjavaju kreditne obveze razmjerno ženama (tj. ne možemo reći da su skloniji neispunjavanju obveza).

```
Values <- matrix(c((nrow(female_clients)-num_female_default),
                   (nrow(male_divor_clients)-num_male_divor_default),
                   num_female_default,
                   num_male_divor_default
                     ), nrow=2, ncol=2, byrow = T)

barplot(Values, main="Kvantitativni prikaz", names.arg=c("žene", "rastavljeni muškarci"), xlab="spol", c

legend("topright", inset = c(-0.1, 0), c("no default", "default"), fill = c("blue", "orange"))
```




```
Values <- matrix(c((nrow(female_clients)-num_female_default)/nrow(female_clients),
                   (nrow(male_divor_clients)-num_male_divor_default)/nrow(male_divor_clients),
                   num_female_default/nrow(female_clients),
                   num_male_divor_default/nrow(male_divor_clients)
                     ), nrow=2, ncol=2, byrow = T)

barplot(Values, main="Postotni prikaz", names.arg=c("žene", "rastavljeni muškarci"), xlab="spol", col =

legend("topright", inset = c(-0.1, 0), c("no default", "default"), fill = c("blue", "orange"))
```



Statistika nad oženjenim muškarcima i ženama u skupu podataka

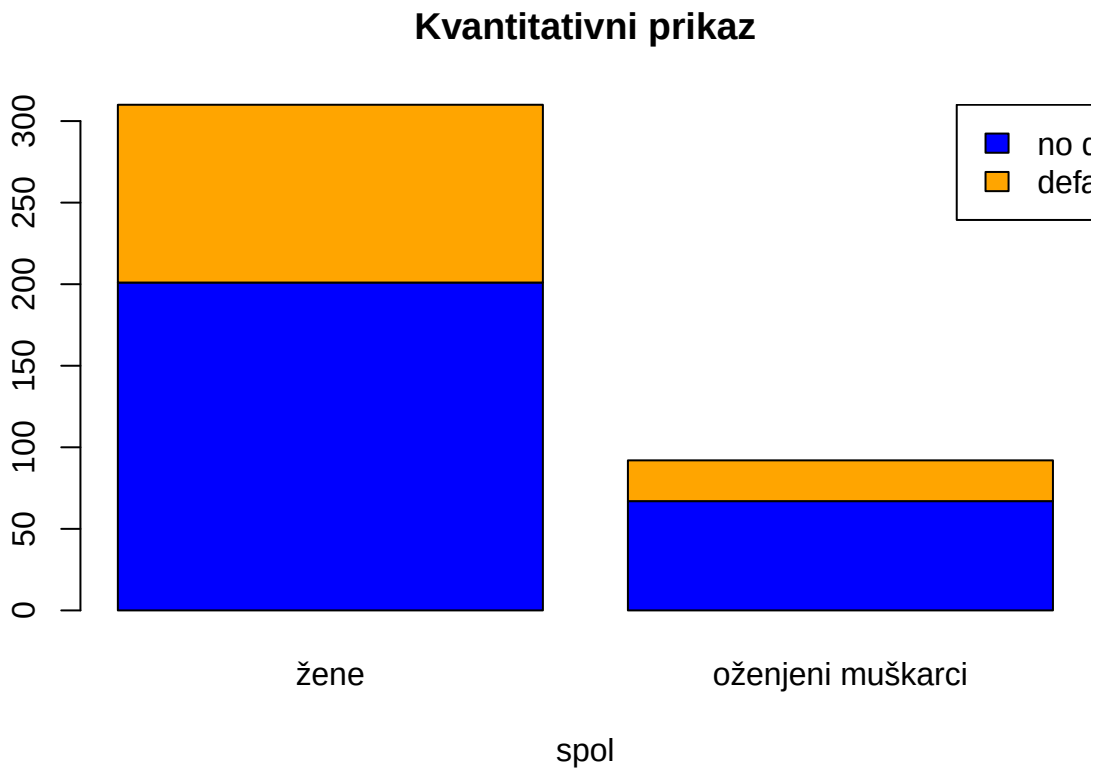
H0: Proporcija onih koji nisu ispunili obveza naspram onih koji su ispunili obaveze jednaka je kod oženjenih muškaraca i žena (ili je manja kod oženjenih muškaraca).
H1: Proporcija osoba koje nisu ispunile obaveze naspram onih koji su ispunili obaveze veća je kod oženjenih muškaraca.

```
male_married_clients <- male_clients[str_detect(male_clients$PersonalStatus, "married"),]
num_male_married_default <- sum(male_married_clients$Default == 1)

proportion_matrix[1,] <- c(nrow(male_married_clients) - num_male_married_default,
                           num_male_married_default)
# proportion_matrix
prop.test(proportion_matrix, alternative = "less")
```

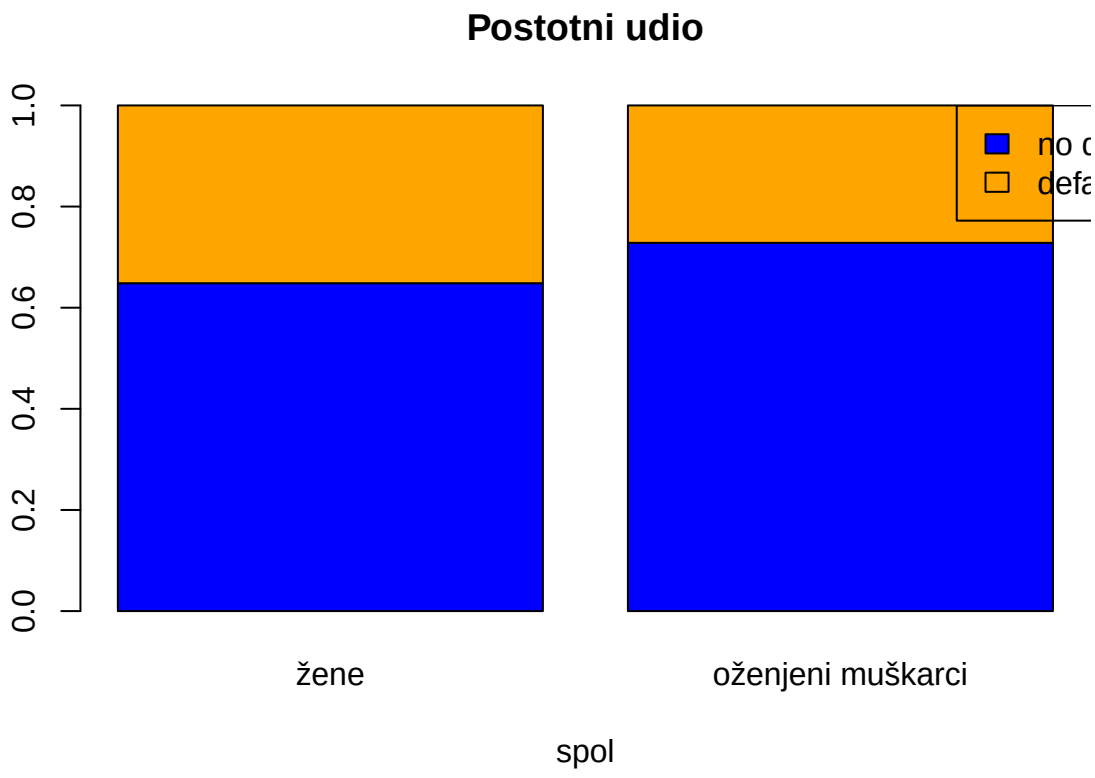
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  proportion_matrix
## X-squared = 1.6932, df = 1, p-value = 0.9034
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.1752928
## sample estimates:
##    prop 1    prop 2
## 0.7282609 0.6483871
```

Iz ovoga zaključujemo, na razini značajnosti 0.05, da oženjeni muškarci ispunjavaju kreditne obveze razmjerno ženama (tj. ne možemo reći da su skloniji neispunjavanju obveza).

```
Values <- matrix(c((nrow(female_clients)-num_female_default),
                   (nrow(male_married_clients)-num_male_married_default),
                   num_female_default,
                   num_male_married_default
                     ), nrow=2, ncol=2, byrow = T)

barplot(Values, main="Kvantitativni prikaz", names.arg=c("žene", "oženjeni muškarci"), xlab="spol", col

legend("topright", inset = c(-0.1, 0), c("no default", "default"), fill = c("blue", "orange"))
```




```
Values <- matrix(c((nrow(female_clients)-num_female_default)/nrow(female_clients),
                   (nrow(male_married_clients)-num_male_married_default)/nrow(male_married_clients),
                   num_female_default/nrow(female_clients),
                   num_male_married_default/nrow(male_married_clients)
                     ), nrow=2, ncol=2, byrow = T)

barplot(Values, main="Postotni udio", names.arg=c("žene", "oženjeni muškarci"), xlab="spol", col = c("bl

legend("topright", inset = c(-0.1, 0), c("no default", "default"), fill = c("blue", "orange"))
```



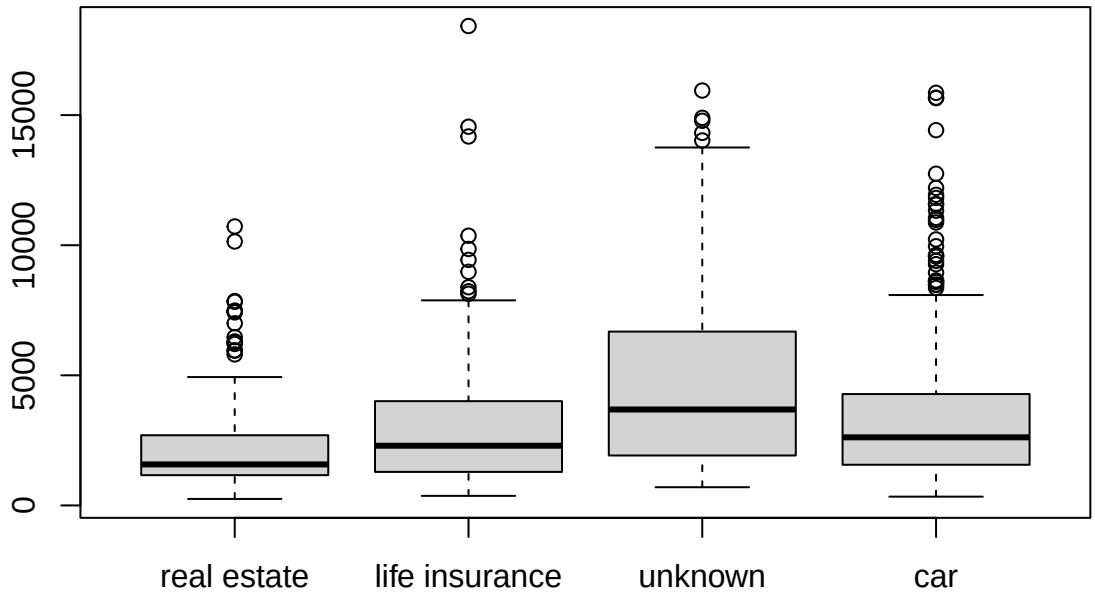
3. pitanje: Postoje li razlike u traženom iznosu kredita prema imovini klijenta?

```
c("real estate", "building society savings agreement/ life insurance",
  "unknown / no property", "car or other, not in attribute Account") %>%
  sapply(function(x) {
    filter(data, Property==x) %>% pull(CreditAmount) -> numbers
    str_c(x, " n: ", length(numbers), "\n") %>% cat()
    print(summary(numbers))
    str_c(x, " standard deviation: ", sd(numbers), "\n") %>% cat()
    cat("--------------------------------------------\n")
    numbers
  }) -> Prop_category
```

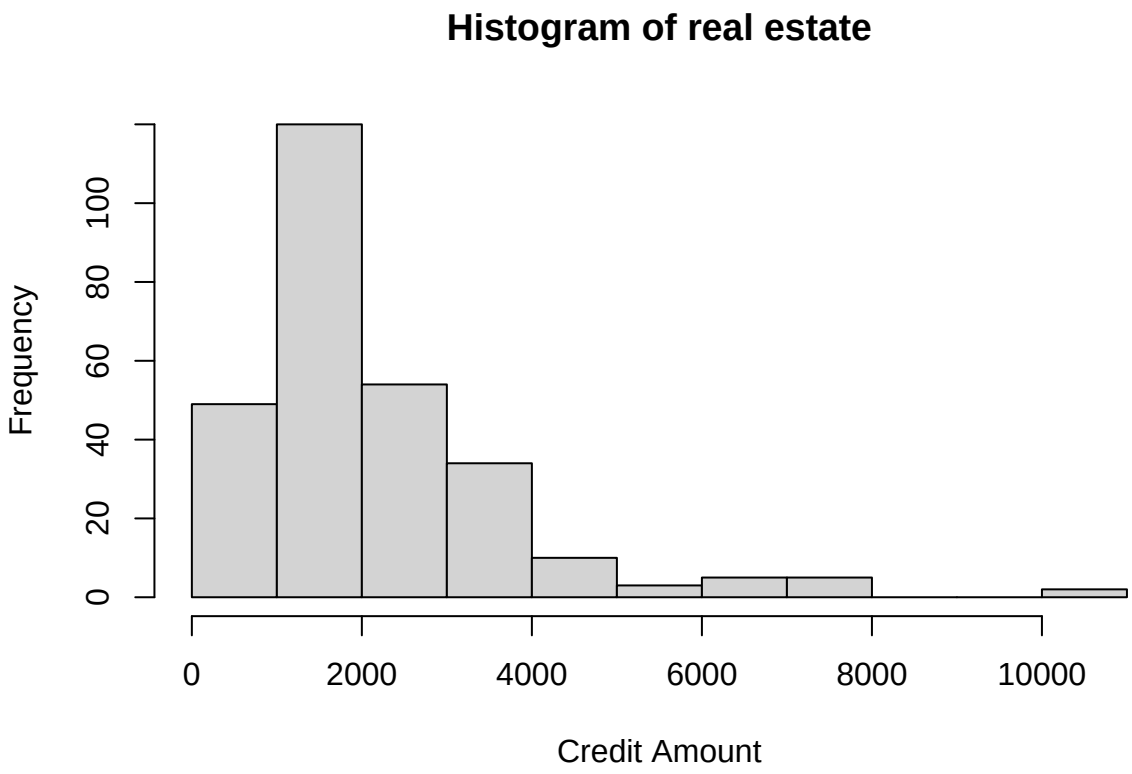
```
## real estate n: 282
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     250    1164    1576    2153    2694   10722
## real estate standard deviation: 1606.27879330167
## --------------------------------------------
## building society savings agreement/ life insurance n: 232
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     368    1288    2294    3104    3990   18424
## building society savings agreement/ life insurance standard deviation: 2602.53168475544
## --------------------------------------------
## unknown / no property n: 154
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     700    1923    3687    4917    6664   15945
```

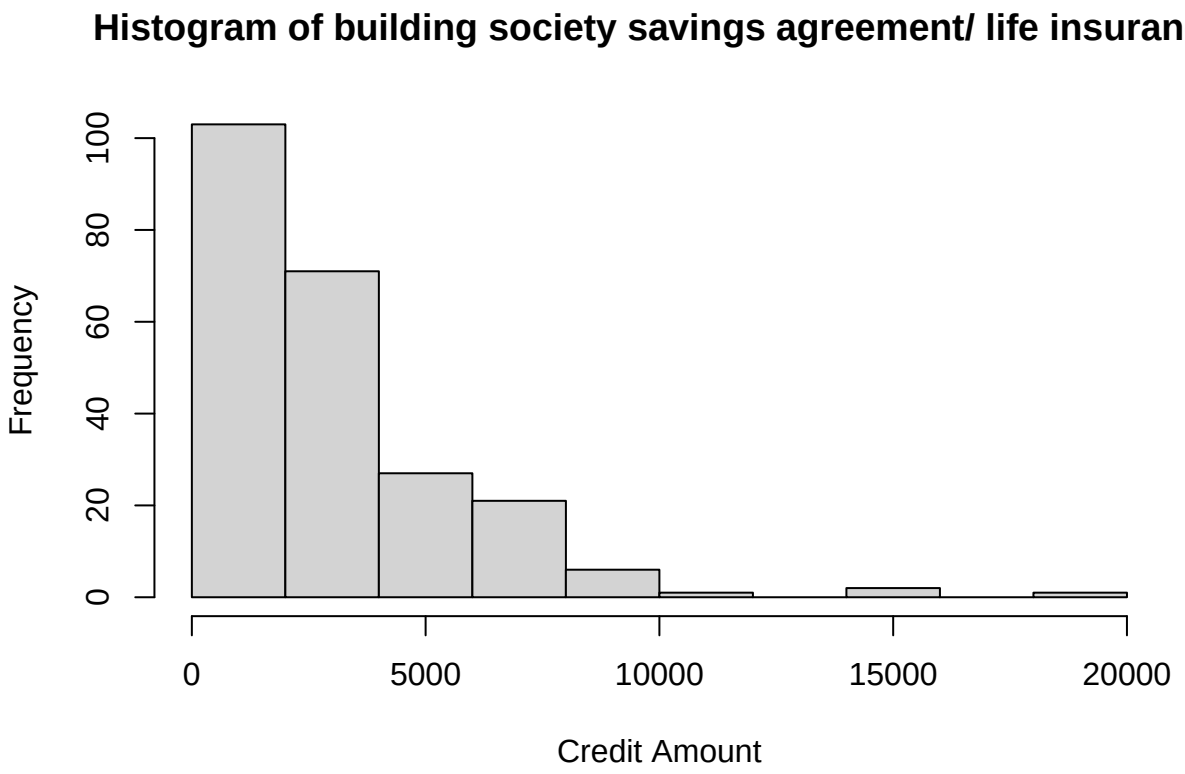
```
## unknown / no property standard deviation: 3725.2304734243
## ------------------------------------------
## car or other, not in attribute Account n: 332
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     338    1565    2618    3574    4280   15857
## car or other, not in attribute Account standard deviation: 2877.33655331269
## ------------------------------------------
```

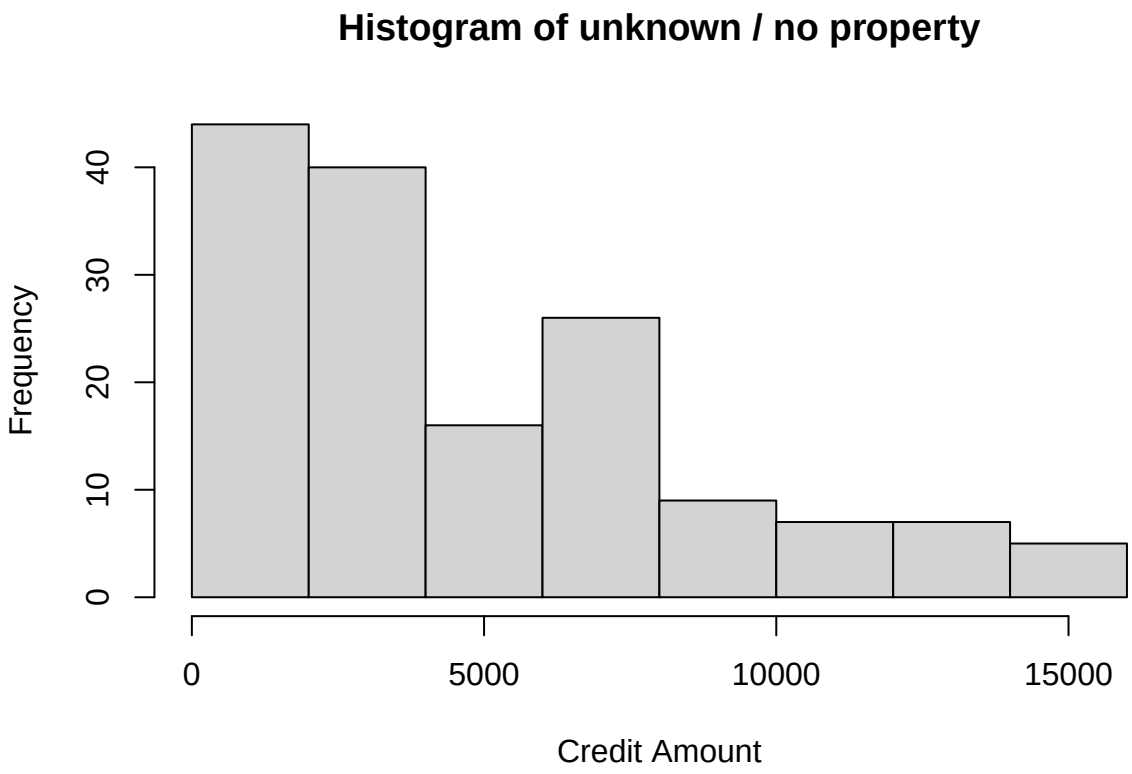
```
boxplot(Prop_category, names=c("real estate", "life insurance", "unknown", "car"))
```

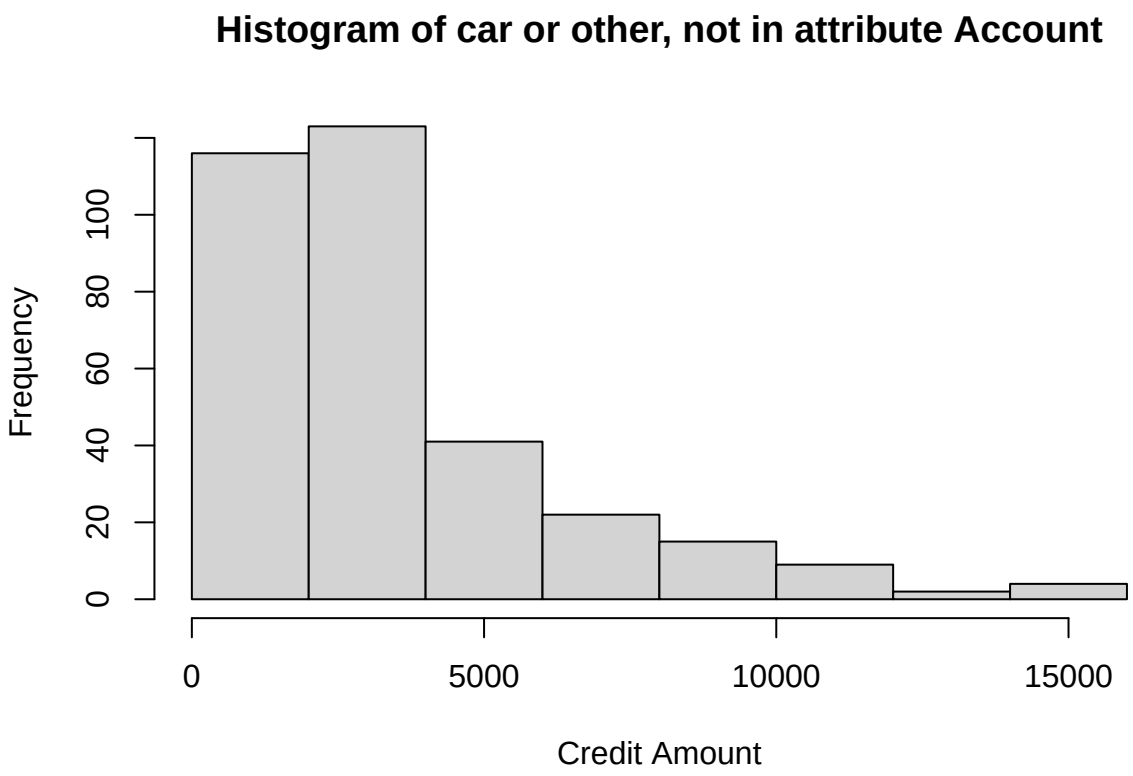



```
for(x in 1:length(Prop_category)) {
  hist(Prop_category[[x]], main = str_c("Histogram of ", names(Prop_category)[x]), xlab="Credit Amount"
}
```


Histogram of real estate
Frequency
0
20
40
60
80
100
0
2000
4000
6000
8000
10000
Credit Amount


Histogram of building society savings agreement/ life insurance
Frequency
0
20
40
60
80
100
0
5000
10000
15000
20000
Credit Amount


Histogram of unknown / no property
Frequency
0
10
20
30
40
0
5000
10000
15000
Credit Amount


Histogram of car or other, not in attribute Account
Frequency
0
20
40
60
80
100
0
5000
10000
15000
Credit Amount