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

```r
    "own"
  )
)
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"
  )
)
data$Default <- factor(
  data$Default,
  levels = c(0,1),
  labels = c(FALSE, TRUE)
)
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
```

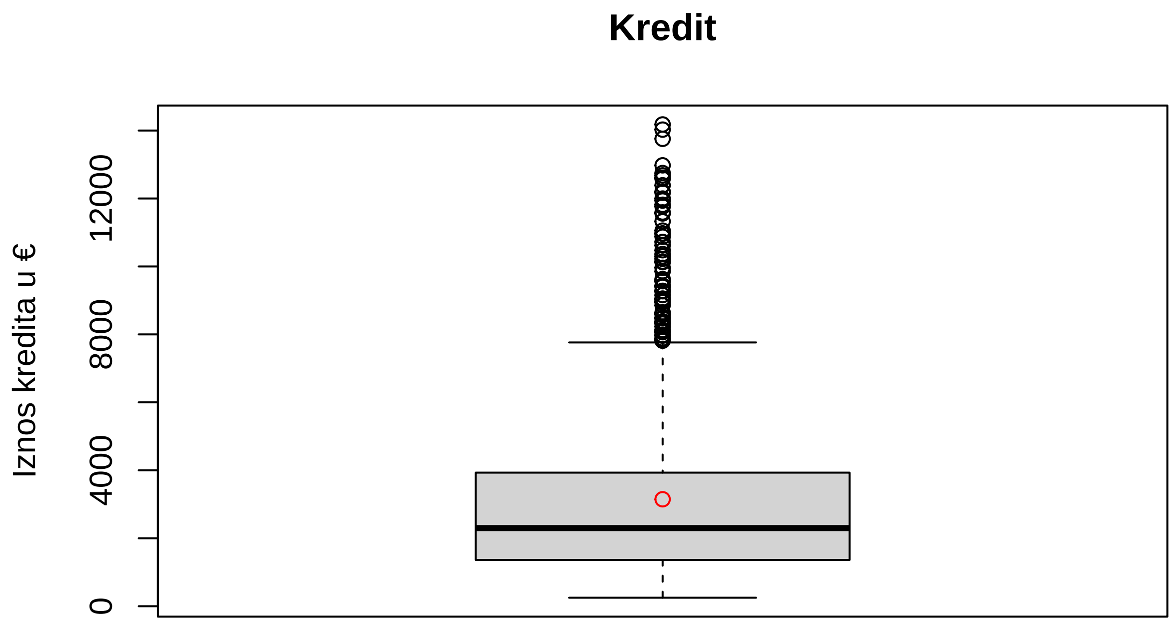
```
##  critical account/ other credits existing (not at this bank):293
##  no credits taken/ all credits paid back duly              : 40
##  existing credits paid back duly till now                  :530
##  all credits at this bank paid back duly                   : 49
##
##
##                Purpose     CreditAmount                          Account
##  radio/television   :280   Min.   :  250   unknown/ no savings account:183
##  car (new)          :234   1st Qu.: 1366   ... < 100                  :603
##  furniture/equipment:181   Median : 2320   100 <= ... < 500           :103
##  car (used)         :103   Mean   : 3271   500 <= ... < 1000          : 63
##  business           : 97   3rd Qu.: 3972   ... >= 1000                : 48
##  education          : 50   Max.   :18424
##  (Other)            : 55
##            EmploymentSince         PercentOfIncome
##  unemployed        : 62    ... < 20%       :476
##  ... < 1 year      :172    20% <= ... < 25%:157
##  1 <= ... < 4 years:339    25% <= ... < 35%:231
##  4 <= ... < 7 years:174    ... >= 35%      :136
##  ... >= 7 years    :253
##
##
##                             PersonalStatus       OtherDebtors
##  female - divorced/separated/married:310    none        :907
##  male - divorced/separated          : 50    guarantor   : 52
##  male - married/widowed             : 92    co-applicant: 41
##  male - single                      :548
##
##
##
##            ResidenceSince
##  ... < 1 year      :130
##  1 <= ... < 4 years:308
##  4 <= ... < 7 years:149
##  .. >= 7 years     :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       FALSE:700
##  yes:963       TRUE :300
##
##
##
##
##
```

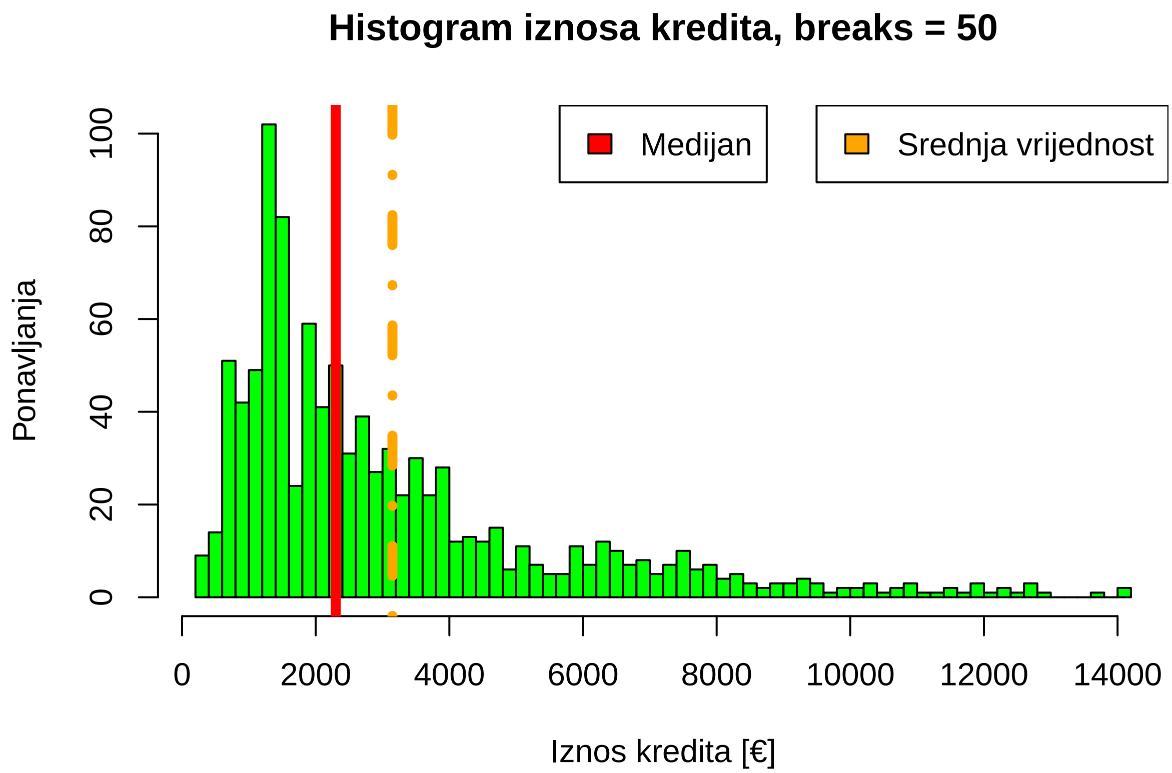
Vidimo da je skup poprilično čist (nema nedostajućih vrijednosti). Iako bi neki stupci moguće bili korisniji da su numerički prije nego kategorički.

Uvodni grafovi

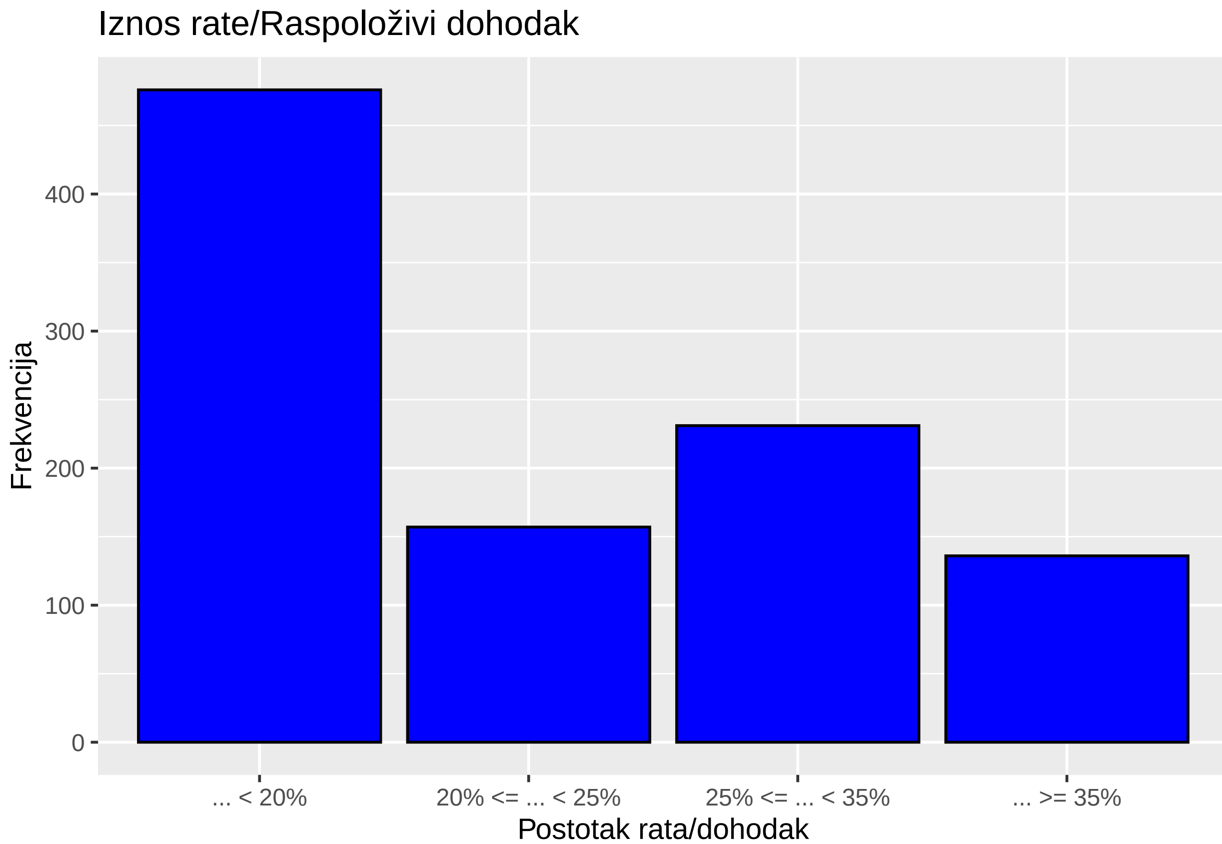
```r
threshold <- quantile(data$CreditAmount, 0.99)  # Set threshold at 99th percentile
# Exclude data points above the threshold
filtered_data <- subset(data, CreditAmount <= threshold)
boxplot(filtered_data$CreditAmount, main="Kredit", ylab="Iznos kredita u €")
points(mean(filtered_data$CreditAmount),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




```r
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)
```




```r
# 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?

Primjeren način za odgovoriti na ovo pitanje je razavoj dobrog modela logističke regresije. Kako bi dobili nekakvu okvirnu sliku o međuovisnostima naših regresora valjalo bi dobiti korelacijsku matricu.

```
cor_matrix <- cor(sapply(data, as.numeric))
# cor_matrix
```

Budući je izlaz u R-u nepregledan predočit ćemo koeficijente varijable čija je apsolutna vrijednost veća od 0.3. Varijable kod kojih dolazi do takvih korelacija su sljedeće:

Duration i CreditAmount

```
cor_matrix["Duration", "CreditAmount"]
```

```
## [1] 0.6249842
```

CreditHistory i NumExistingCredits

```
cor_matrix["CreditHistory", "NumExistingCredits"]
```

```
## [1] -0.5340804
```

ResidenceSince i Housing

```r
cor_matrix["ResidenceSince","Housing"]
```

```
## [1] -0.3040227
```

Property i Housing

```r
cor_matrix["Property", "Housing"]
```

```
## [1] 0.4932404
```

```r
logreg.mdl <- glm(Default ~ as.numeric(AccountStatus) + Duration + as.numeric(CreditHistory) + Purpose +
summary(logreg.mdl)
```

```
##
## Call:
## glm(formula = Default ~ as.numeric(AccountStatus) + Duration +
##     as.numeric(CreditHistory) + Purpose + CreditAmount + as.numeric(Account) +
##     as.numeric(EmploymentSince) + as.numeric(PercentOfIncome) +
##     PersonalStatus + OtherDebtors + as.numeric(ResidenceSince) +
##     as.numeric(Property) + Age + OtherInstallPlans + Housing +
##     as.numeric(NumExistingCredits) + as.numeric(Job) + as.numeric(NumberOfDependents) +
##     Telephone + ForeignWorker, family = binomial(), data = data)
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                             -2.255e+00  1.137e+00  -1.983
## as.numeric(AccountStatus)                4.344e-01  8.469e-02   5.129
## Duration                                 2.893e-02  8.664e-03   3.339
## as.numeric(CreditHistory)                3.118e-01  8.241e-02   3.783
## Purposecar (new)                         6.094e-01  3.005e-01   2.028
## Purposecar (used)                       -1.088e+00  3.958e-01  -2.748
## Purposedomestic appliances               2.172e-01  7.487e-01   0.290
## Purposeeducation                         5.209e-01  4.194e-01   1.242
## Purposefurniture/equipment               8.276e-03  3.143e-01   0.026
## Purposeothers                           -6.721e-01  7.882e-01  -0.853
## Purposeradio/television                 -3.478e-01  3.020e-01  -1.152
## Purposerepairs                           4.395e-01  5.513e-01   0.797
## Purposeretraining                       -1.138e+00  1.130e+00  -1.007
## CreditAmount                             1.187e-04  4.056e-05   2.926
## as.numeric(Account)                     -7.717e-02  8.643e-02  -0.893
## as.numeric(EmploymentSince)             -1.691e-01  7.210e-02  -2.345
## as.numeric(PercentOfIncome)             -3.289e-01  8.245e-02  -3.989
## PersonalStatusmale - divorced/separated  4.561e-01  3.606e-01   1.265
## PersonalStatusmale - married/widowed    -4.457e-02  2.972e-01  -0.150
## PersonalStatusmale - single             -4.727e-01  1.949e-01  -2.425
## OtherDebtorsguarantor                   -6.305e-01  3.996e-01  -1.578
## OtherDebtorsco-applicant                 4.966e-01  3.907e-01   1.271
## as.numeric(ResidenceSince)               3.754e-02  7.973e-02   0.471
## as.numeric(Property)                    -2.116e-01  9.805e-02  -2.158
## Age                                     -1.469e-02  8.318e-03  -1.766
## OtherInstallPlansstores                  6.988e-01  3.402e-01   2.054
## OtherInstallPlansbank                    6.476e-01  2.196e-01   2.949
## Housingrent                              4.526e-01  3.544e-01   1.277
## Housingown                              -8.931e-02  3.154e-01  -0.283
## as.numeric(NumExistingCredits)           2.629e-01  1.671e-01   1.573
## as.numeric(Job)                          4.976e-02  9.494e-02   0.524
```

```
## as.numeric(NumberOfDependents)                        -3.504e-01  2.345e-01  -1.494
## Telephoneyes, registered under the customers name -3.952e-01  1.754e-01  -2.253
## ForeignWorkeryes                                       1.192e+00  5.784e-01   2.061
##                                                      Pr(>|z|)
## (Intercept)                                          0.047333 *
## as.numeric(AccountStatus)                            2.91e-07 ***
## Duration                                             0.000842 ***
## as.numeric(CreditHistory)                            0.000155 ***
## Purposecar (new)                                     0.042576 *
## Purposecar (used)                                    0.005991 **
## Purposedomestic appliances                           0.771720
## Purposeeducation                                     0.214277
## Purposefurniture/equipment                           0.978993
## Purposeothers                                        0.393830
## Purposeradio/television                              0.249361
## Purposerepairs                                       0.425262
## Purposeretraining                                    0.313776
## CreditAmount                                         0.003439 **
## as.numeric(Account)                                  0.371932
## as.numeric(EmploymentSince)                          0.019012 *
## as.numeric(PercentOfIncome)                          6.63e-05 ***
## PersonalStatusmale - divorced/separated              0.205872
## PersonalStatusmale - married/widowed                 0.880794
## PersonalStatusmale - single                          0.015296 *
## OtherDebtorsguarantor                                0.114602
## OtherDebtorsco-applicant                             0.203659
## as.numeric(ResidenceSince)                           0.637767
## as.numeric(Property)                                 0.030938 *
## Age                                                  0.077465 .
## OtherInstallPlansstores                              0.039999 *
## OtherInstallPlansbank                                0.003193 **
## Housingrent                                          0.201608
## Housingown                                           0.777026
## as.numeric(NumExistingCredits)                       0.115772
## as.numeric(Job)                                      0.600193
## as.numeric(NumberOfDependents)                       0.135192
## Telephoneyes, registered under the customers name 0.024233 *
## ForeignWorkeryes                                     0.039349 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.73  on 999  degrees of freedom
## Residual deviance:  998.43  on 966  degrees of freedom
## AIC: 1066.4
##
## Number of Fisher Scoring iterations: 5
```

Sad ćemo razmotriti neke od mjera kvalitete modela.

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

```
## [1] 0.1827695
```

```r
Ypredicted <- logreg.mdl$fitted.values >= 0.5
tab <- table(data$Default, Ypredicted)

tab
```

```
##        Ypredicted
##         FALSE TRUE
##   FALSE   628   72
##   TRUE    187  113
```

```r
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.741
```

```r
precision
```

```
## [1] 0.6108108
```

```r
recall
```

```
## [1] 0.3766667
```

```r
specificity
```

```
## [1] 0.7705521
```

```r
logreg.mdl2 <- glm(Default ~ as.numeric(AccountStatus) + Duration + as.numeric(CreditHistory) + Purpose
summary(logreg.mdl2)
```

```
##
## Call:
## glm(formula = Default ~ as.numeric(AccountStatus) + Duration +
##     as.numeric(CreditHistory) + Purpose + CreditAmount + as.numeric(EmploymentSince) +
##     as.numeric(PercentOfIncome) + as.numeric(Property) + Age +
##     OtherInstallPlans + Telephone, family = binomial(), data = data)
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                -1.179e+00  6.347e-01  -1.858
## as.numeric(AccountStatus)                   4.129e-01  8.173e-02   5.052
## Duration                                    2.969e-02  8.368e-03   3.548
## as.numeric(CreditHistory)                   2.612e-01  6.907e-02   3.783
## Purposecar (new)                            4.928e-01  2.888e-01   1.707
## Purposecar (used)                          -1.057e+00  3.769e-01  -2.805
## Purposedomestic appliances                  1.119e-01  7.325e-01   0.153
## Purposeeducation                            5.410e-01  4.068e-01   1.330
## Purposefurniture/equipment                  1.532e-02  3.012e-01   0.051
## Purposeothers                              -6.995e-01  7.449e-01  -0.939
## Purposeradio/television                    -4.635e-01  2.894e-01  -1.601
## Purposerepairs                              4.032e-01  5.409e-01   0.745
## Purposeretraining                          -1.042e+00  1.108e+00  -0.940
## CreditAmount                                1.027e-04  3.919e-05   2.621
## as.numeric(EmploymentSince)                -1.760e-01  6.734e-02  -2.614
```

```
## as.numeric(PercentOfIncome)                         -2.818e-01  7.881e-02  -3.576
## as.numeric(Property)                                -2.103e-01  8.130e-02  -2.587
## Age                                                 -1.671e-02  7.615e-03  -2.194
## OtherInstallPlansstores                              5.229e-01  3.326e-01   1.572
## OtherInstallPlansbank                                5.764e-01  2.131e-01   2.705
## Telephoneyes, registered under the customers name -3.348e-01  1.696e-01  -1.974
##                                                     Pr(>|z|)
## (Intercept)                                         0.063231 .
## as.numeric(AccountStatus)                           4.38e-07 ***
## Duration                                            0.000388 ***
## as.numeric(CreditHistory)                           0.000155 ***
## Purposecar (new)                                    0.087909 .
## Purposecar (used)                                   0.005024 **
## Purposedomestic appliances                          0.878541
## Purposeeducation                                    0.183609
## Purposefurniture/equipment                          0.959433
## Purposeothers                                       0.347756
## Purposeradio/television                             0.109276
## Purposerepairs                                      0.456067
## Purposeretraining                                   0.347175
## CreditAmount                                        0.008769 **
## as.numeric(EmploymentSince)                         0.008959 **
## as.numeric(PercentOfIncome)                         0.000349 ***
## as.numeric(Property)                                0.009692 **
## Age                                                 0.028250 *
## OtherInstallPlansstores                             0.115913
## OtherInstallPlansbank                               0.006832 **
## Telephoneyes, registered under the customers name 0.048413 *
## ---
## 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: 1033.6  on 979  degrees of freedom
## AIC: 1075.6
##
## Number of Fisher Scoring iterations: 4
```

```r
Rsq = 1 - logreg.mdl2$deviance/logreg.mdl$null.deviance
Rsq
```

```
## [1] 0.153956
```

```r
Ypredicted <- logreg.mdl2$fitted.values >= 0.5
tab <- table(data$Default, Ypredicted)

tab
```

```
##        Ypredicted
##         FALSE TRUE
##   FALSE   627   73
##   TRUE    194  106
```

```r
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
```

```r
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.733
```

```r
precision
```

```
## [1] 0.5921788
```

```r
recall
```

```
## [1] 0.3533333
```

```r
specificity
```

```
## [1] 0.7637028
```

```r
anova(logreg.mdl, logreg.mdl2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Default ~ as.numeric(AccountStatus) + Duration + as.numeric(CreditHistory) +
##     Purpose + CreditAmount + as.numeric(Account) + as.numeric(EmploymentSince) +
##     as.numeric(PercentOfIncome) + PersonalStatus + OtherDebtors +
##     as.numeric(ResidenceSince) + as.numeric(Property) + Age +
##     OtherInstallPlans + Housing + as.numeric(NumExistingCredits) +
##     as.numeric(Job) + as.numeric(NumberOfDependents) + Telephone +
##     ForeignWorker
## Model 2: Default ~ as.numeric(AccountStatus) + Duration + as.numeric(CreditHistory) +
##     Purpose + CreditAmount + as.numeric(EmploymentSince) + as.numeric(PercentOfIncome) +
##     as.numeric(Property) + Age + OtherInstallPlans + Telephone
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1       966     998.43
## 2       979    1033.64 -13  -35.202 0.0007882 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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
num_female_clients <- data %>% filter(str_detect(PersonalStatus, "female")) %>% count() %>% as.numeric()
num_male_clients <- data %>% filter(!str_detect(PersonalStatus, "female")) %>% count() %>% as.numeric()
num_female_default <- data %>% filter(str_detect(PersonalStatus, "female") & Default == T) %>% count() %
num_male_default <- data %>% filter(!str_detect(PersonalStatus, "female") & Default == T) %>% count() %
```

```r
proportion_matrix <- matrix(c(num_male_clients-num_male_default,
                            num_male_default,
                         num_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).

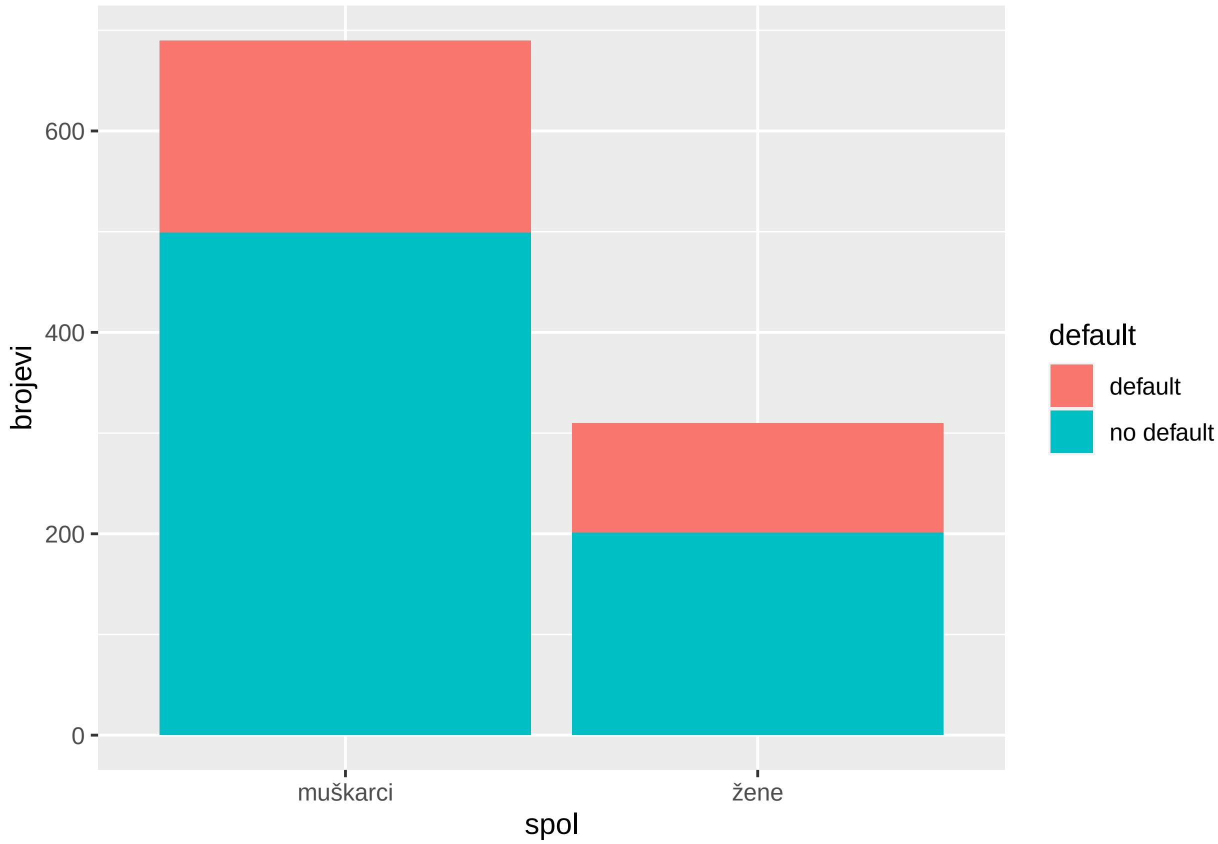
Provodimo Z-test o dvije proporcije s očekivanjem da će nam dati vrlo slične rezultate kao i χ^2-test.

```r
n1 <- num_male_clients
n2 <- num_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))
```

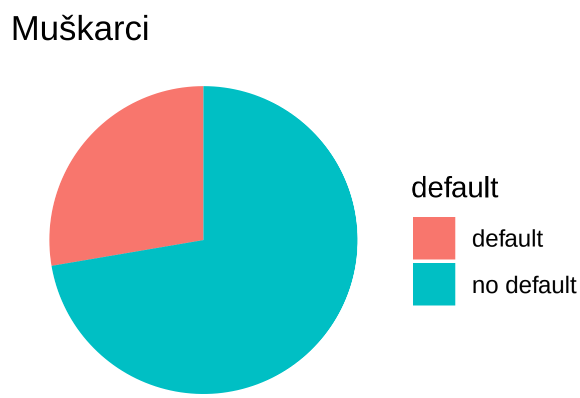
```
## The p-value of the Z statistic is:  0.9915134
```

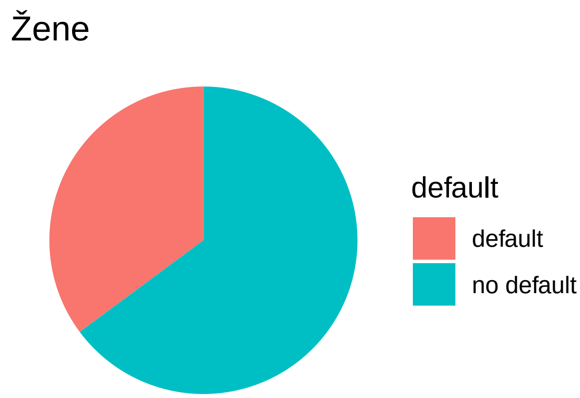
Kao što možemo uočiti Z-test nam daje isti zaključak i vrlo sličnu p-vrijednost kao i χ^2-test pa ćemo nadalje koristiti χ^2 jer je on implementiran u R-u.

```r
df <- data.frame(spol=c("žene", "muškarci", "žene", "muškarci"),
                 brojevi=c(num_female_default, num_male_default, k2, k1),
                 default=c("default", "default", "no default", "no default"))

ggplot(df, aes(x=spol, y=brojevi, fill=default)) + geom_bar(stat="identity")
```




```
g1 <- df %>% filter(spol=="žene") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Žene")
g2 <- df %>% filter(spol=="muškarci") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Muškarci")
grid.arrange(g2, g1)
```





Statistika nad slobodnim muškarcima i ženama u skupu podataka

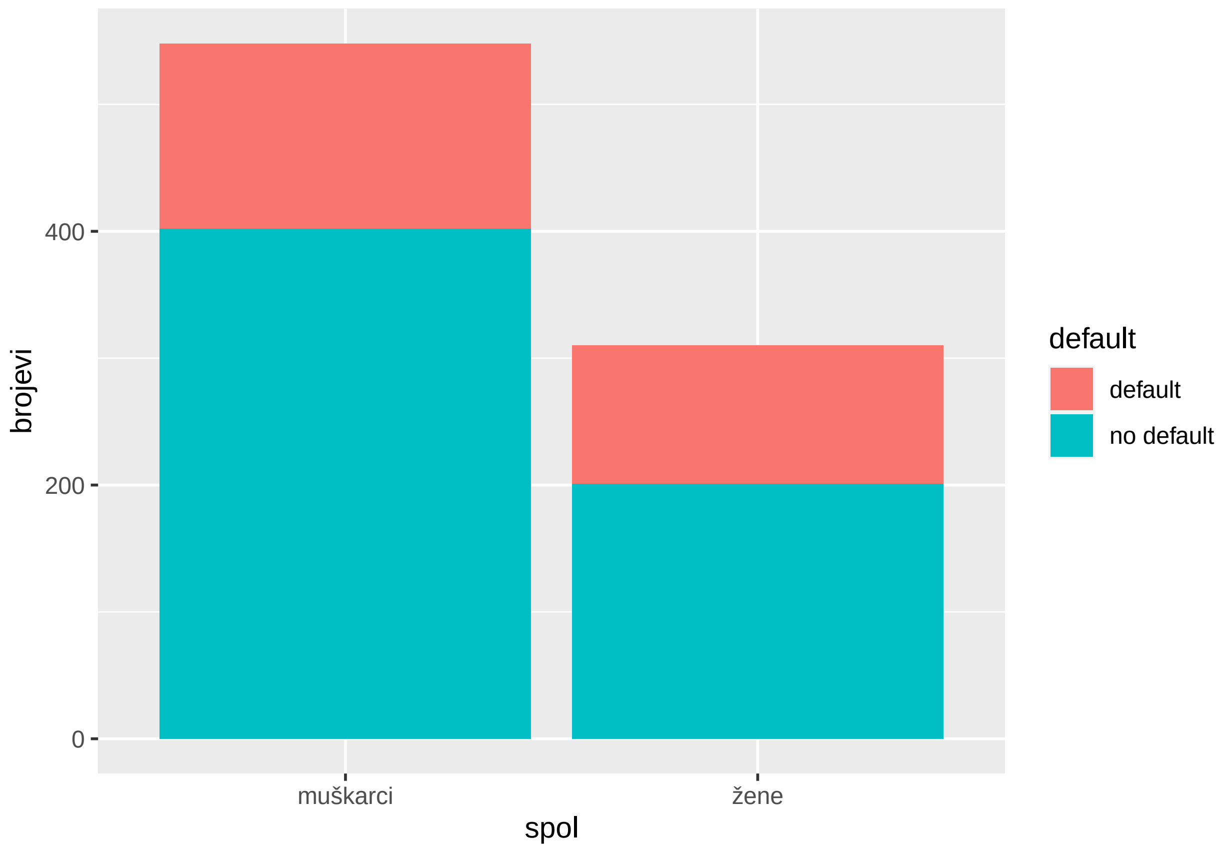
H0: Proporcija onih koji nisu ispunili obveza naspram onih koji su ispunili obaveze jednaka je kod slobodnih muškaraca i žena (ili je manja kod slobodnih muškaraca).
H1: Proporcija osoba koje nisu ispunile obaveze naspram onih koji su ispunili obaveze veća je kod slobodnih muškaraca.

```
num_male_single <- data %>% filter(str_detect(PersonalStatus, "single")) %>% count() %>% as.numeric()
num_male_single_default <- data %>% filter(str_detect(PersonalStatus, "single") & Default == T) %>% cou

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

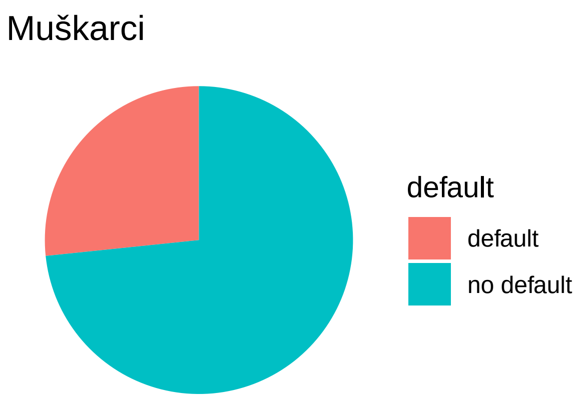
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  proportion_matrix
## X-squared = 6.4775, df = 1, p-value = 0.9945
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.1420715
## sample estimates:
##    prop 1    prop 2
## 0.7335766 0.6483871
```

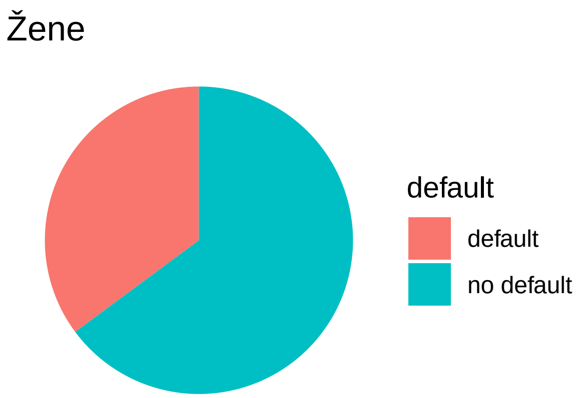
Iz ovoga zaključujemo, na razini značajnosti 0.05, da slobodni muškarci ispunjavaju kreditne obveze razmjerno ženama (tj. ne možemo reći da su skloniji neispunjavanju obveza).

```
df <- data.frame(spol=c("žene", "muškarci", "žene", "muškarci"),
                 brojevi=c(num_female_default, num_male_single_default, k2,
                           num_male_single-num_male_single_default),
                 default=c("default", "default", "no default", "no default"))

ggplot(df, aes(x=spol, y=brojevi, fill=default)) + geom_bar(stat="identity")
```




```
g1 <- df %>% filter(spol=="žene") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Žene")
g2 <- df %>% filter(spol=="muškarci") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Muškarci")
grid.arrange(g2, g1)
```





Statistika nad rastavljenim muškarcima i ženama u skupu podataka

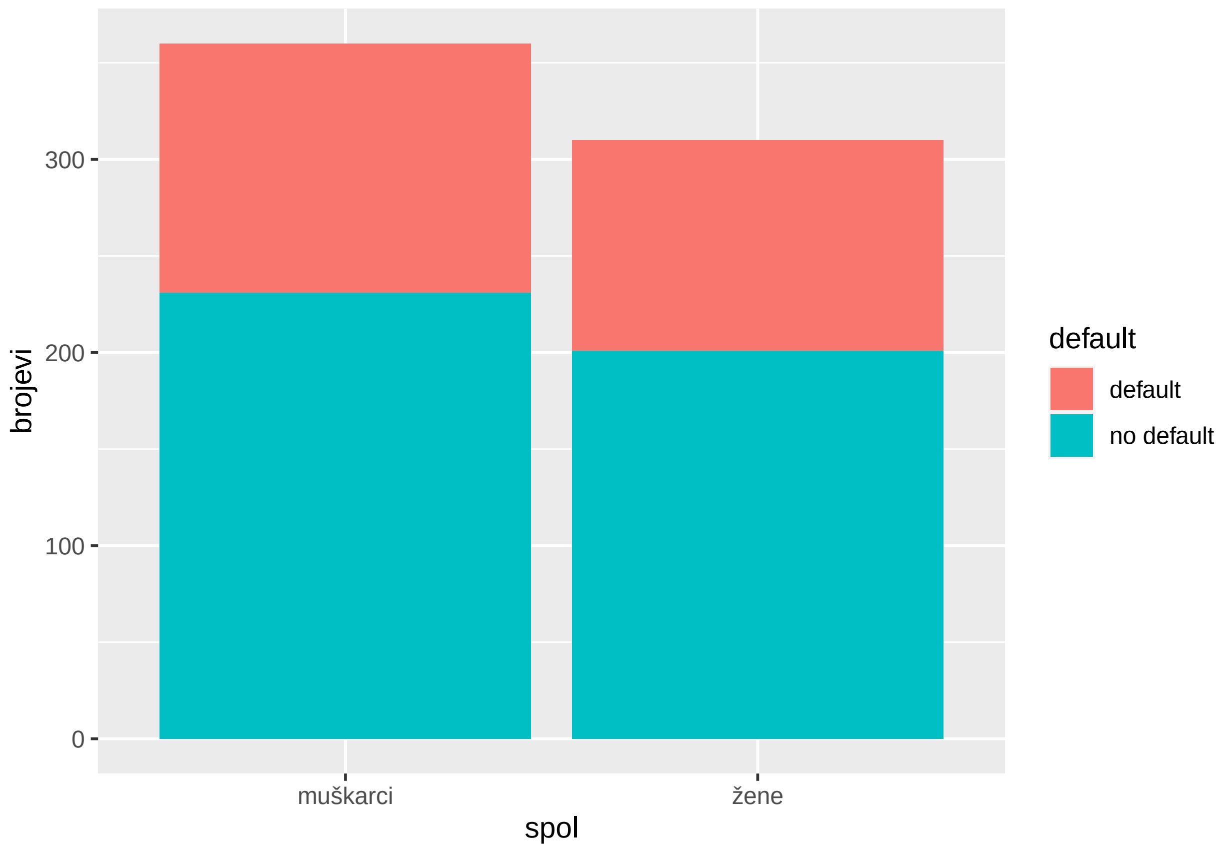
H0: Proporcija onih koji nisu ispunili obveza naspram onih koji su ispunili obaveze jednaka je kod rastavljenih muškaraca i žena (ili je manja kod rastavljenih muškaraca).
H1: Proporcija osoba koje nisu ispunile obaveze naspram onih koji su ispunili obaveze veća je kod rastavljenih muškaraca.

```
num_male_divor <- data %>% filter(str_detect(PersonalStatus, "divorced")) %>% count() %>% as.numeric()
num_male_divor_default <- data %>% filter(str_detect(PersonalStatus, "divorced") & Default == T) %>% co

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

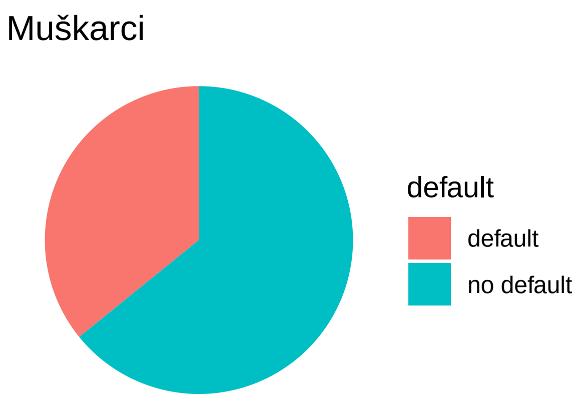
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  proportion_matrix
## X-squared = 0.010056, df = 1, p-value = 0.4601
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000  0.05725463
## sample estimates:
##    prop 1    prop 2
## 0.6416667 0.6483871
```

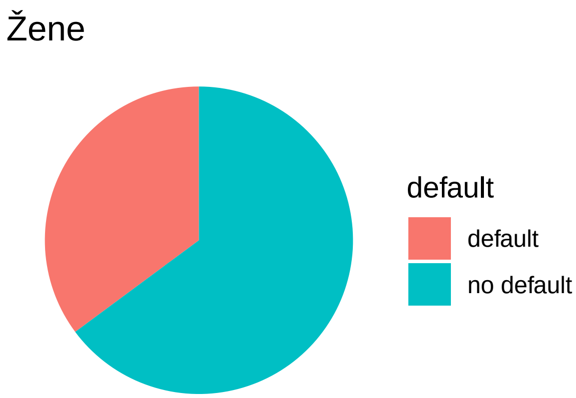
Iz ovoga zaključujemo, na razini značajnosti 0.05, da rastavljeni muškarci ispunjavaju kreditne obveze razmjerno ženama (tj. ne možemo reći da su skloniji neispunjavanju obveza).

```
df <- data.frame(spol=c("žene", "muškarci", "žene", "muškarci"),
                 brojevi=c(num_female_default, num_male_divor_default, k2,
                           num_male_divor-num_male_divor_default),
                 default=c("default", "default", "no default", "no default"))

ggplot(df, aes(x=spol, y=brojevi, fill=default)) + geom_bar(stat="identity")
```




```
g1 <- df %>% filter(spol=="žene") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Žene")
g2 <- df %>% filter(spol=="muškarci") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Muškarci")
grid.arrange(g2, g1)
```





Statistika nad oženjenim muškarcima i ženama u skupu podataka

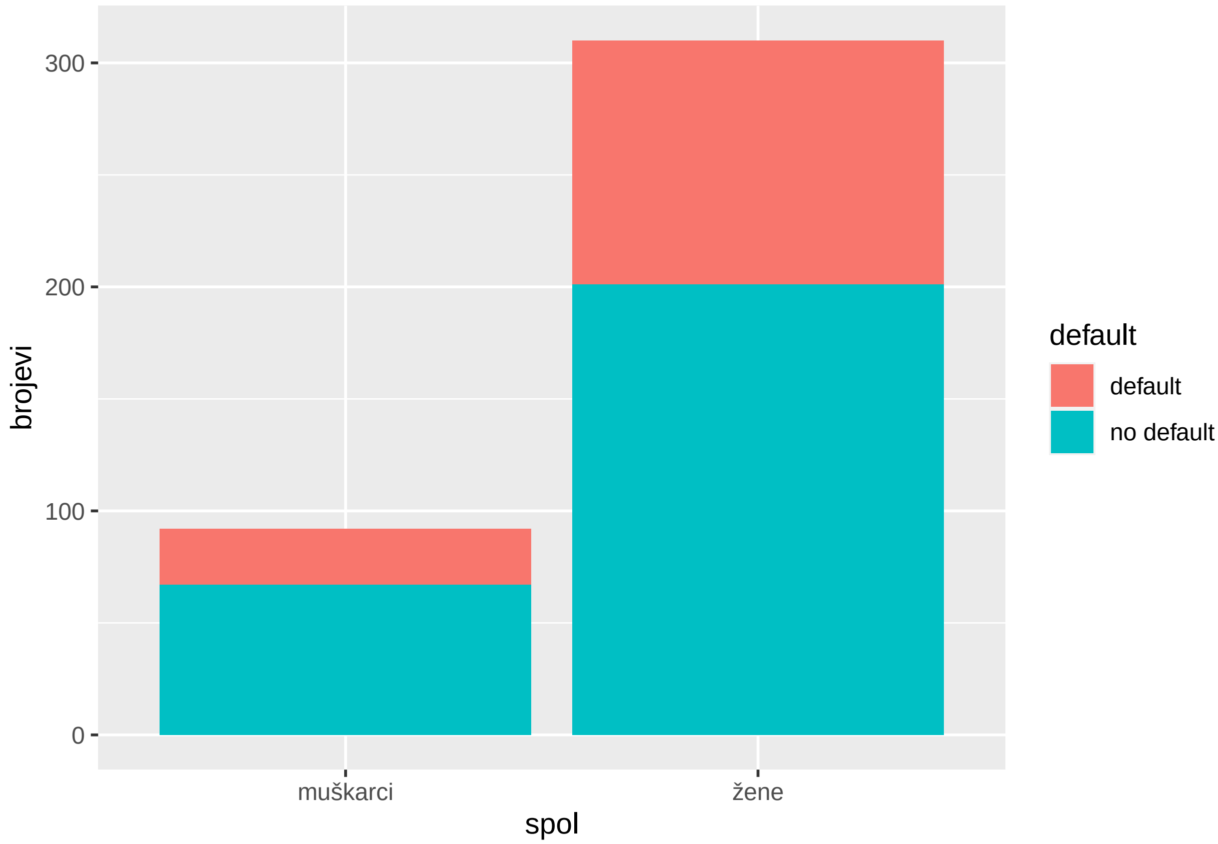
H0: Proporcija onih koji nisu ispunili obveza naspram onih koji su ispunili obaveze jednaka je kod oženjenih muškaraca i žena (ili je manja kod oženjenih muškaraca).
H1: Proporcija osoba koje nisu ispunile obaveze naspram onih koji su ispunili obaveze veća je kod oženjenih muškaraca.

```
num_male_married <- data %>% filter(str_detect(PersonalStatus, "widowed")) %>% count() %>% as.numeric()
num_male_married_default <- data %>% filter(str_detect(PersonalStatus, "widowed") & Default == T) %>% co

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

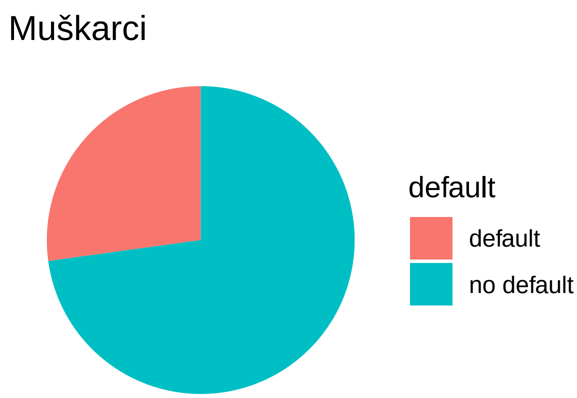
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  proportion_matrix
## X-squared = 1.6932, df = 1, p-value = 0.9034
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.1752928
## sample estimates:
##    prop 1    prop 2
## 0.7282609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

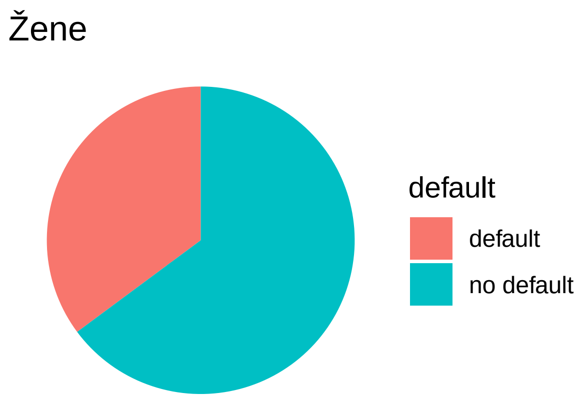
Iz ovoga zaključujemo, na razini značajnosti 0.05, da oženjeni muškarci ispunjavaju kreditne obveze razmjerno ženama (tj. ne možemo reći da su skloniji neispunjavanju obveza).

```
df <- data.frame(spol=c("žene", "muškarci", "žene", "muškarci"),
                 brojevi=c(num_female_default, num_male_married_default, k2,
                           num_male_married-num_male_married_default),
                 default=c("default", "default", "no default", "no default"))

ggplot(df, aes(x=spol, y=brojevi, fill=default)) + geom_bar(stat="identity")
```




```
g1 <- df %>% filter(spol=="žene") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Žene")
g2 <- df %>% filter(spol=="muškarci") %>% ggplot(aes(x="", y=brojevi, fill=default)) +
  geom_bar(stat="identity") + coord_polar("y", start=0) + theme_void() + ggtitle("Muškarci")
grid.arrange(g2, g1)
```





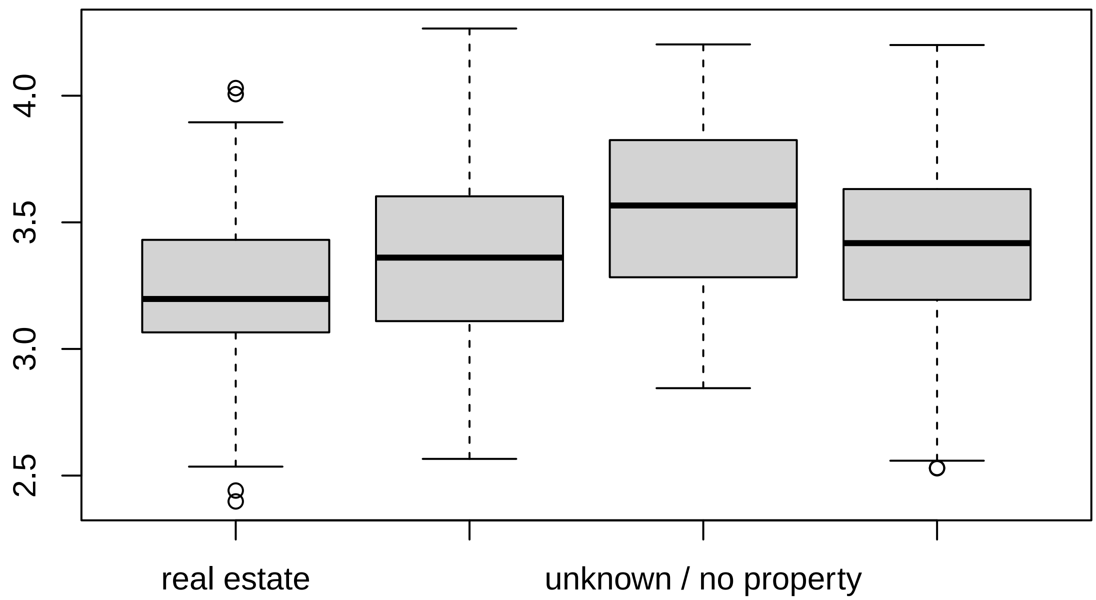
3. pitanje: Postoje li razlike u traženom iznosu kredita prema imovini klijenta?

```
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")
    log10(numbers)
  }) -> Prop_category
```

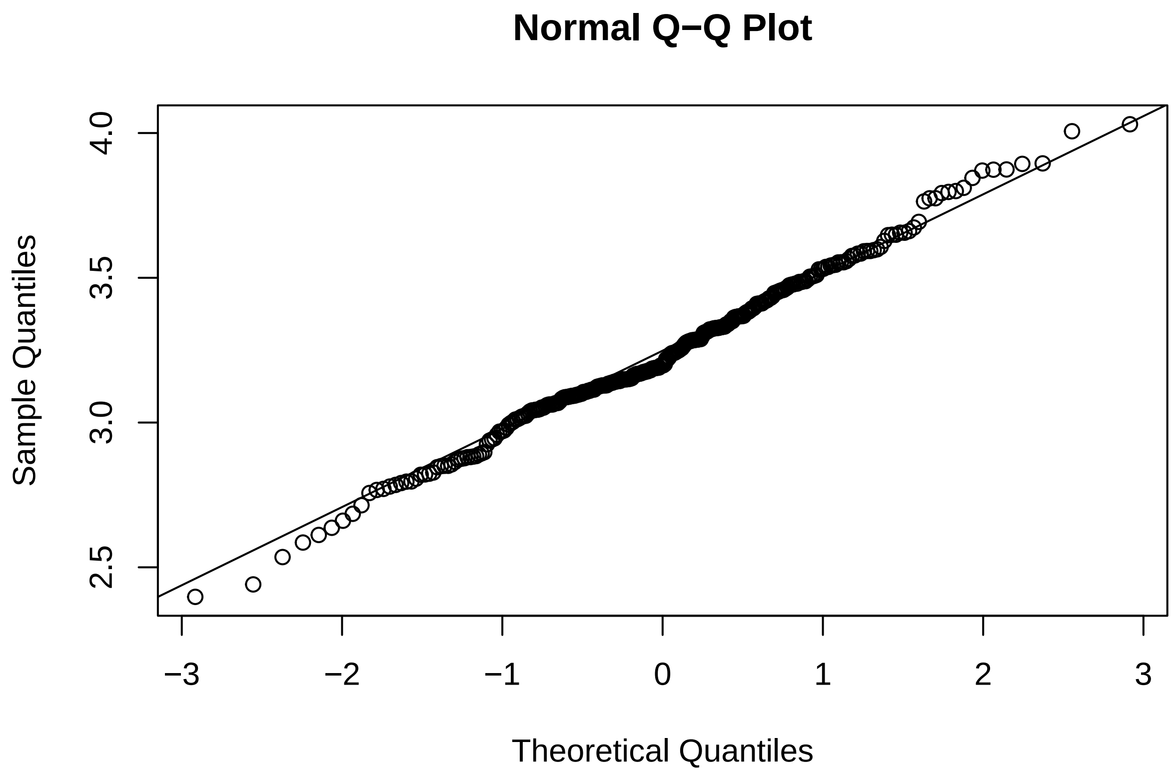
```
## real estate n: 282
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     250    1164    1576    2153    2694   10722
## real estate standard deviation: 1606.27879330167
## -------------------------------------------
## building society savings agreement/ life insurance n: 232
##    Min. 1st 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
## ------------------------------------------
```

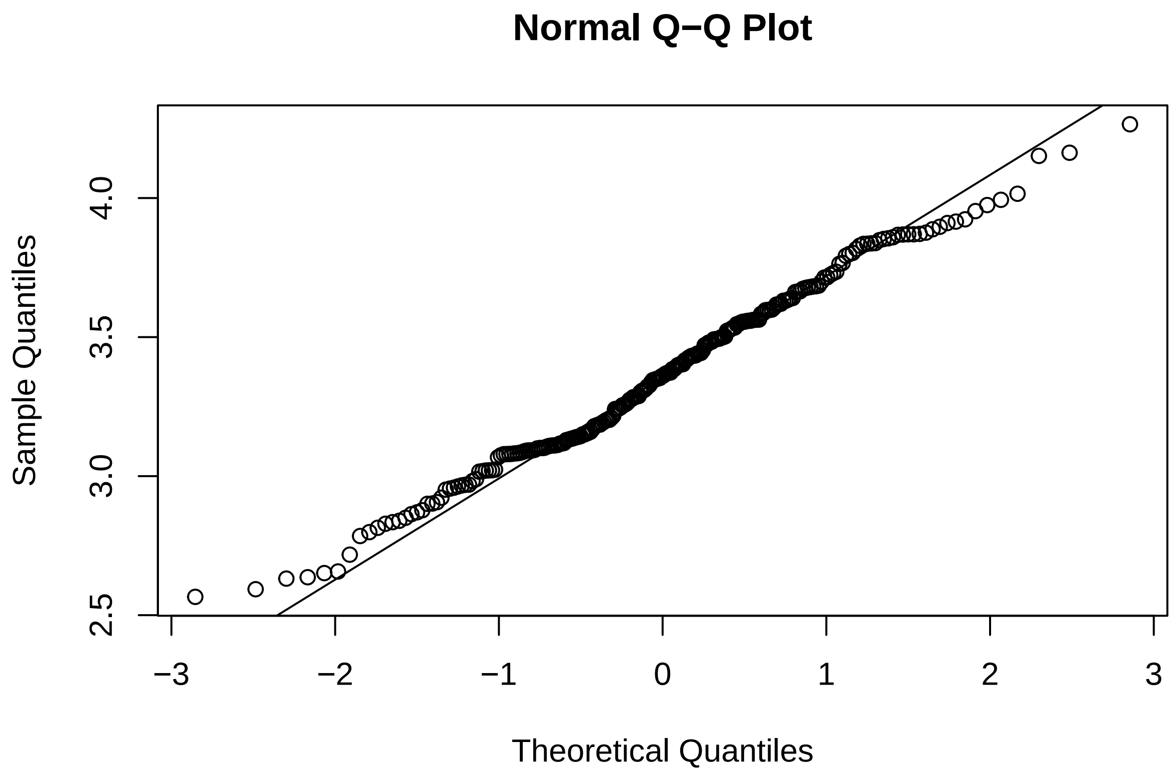
```r
boxplot(Prop_category)
```



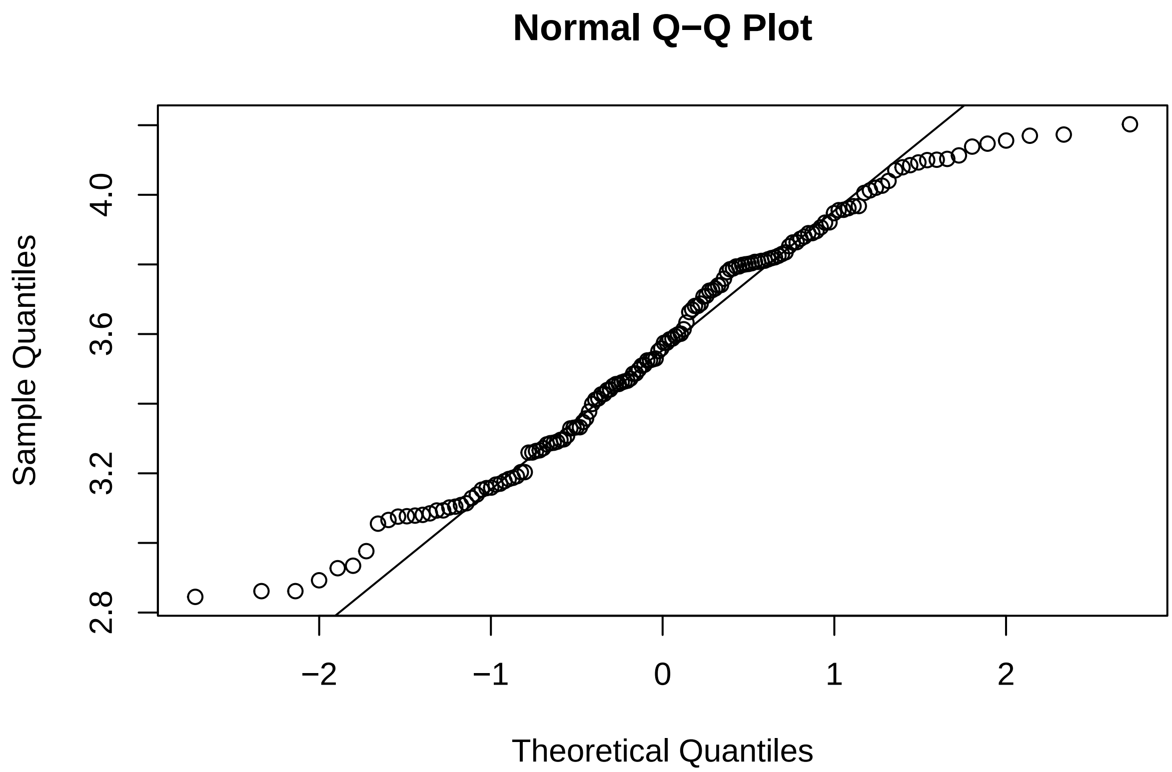

```r
qqnorm(Prop_category$`real estate`)
qqline(Prop_category$`real estate`)
```



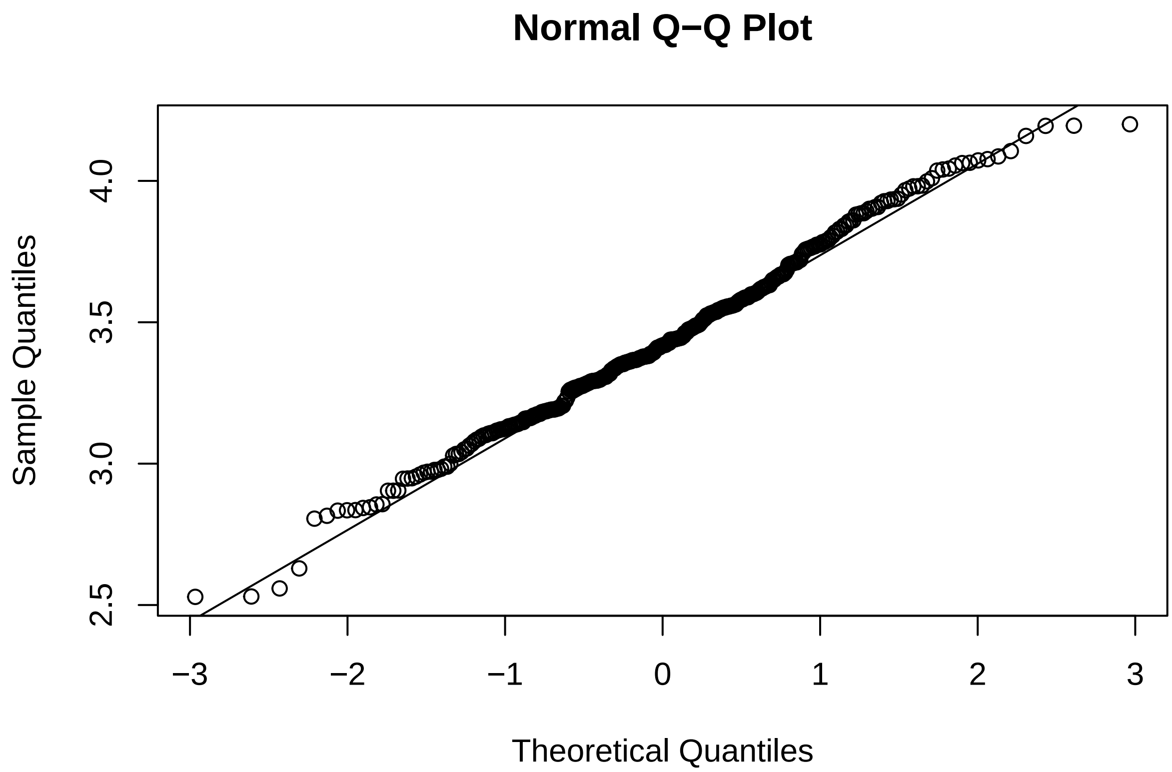

```r
qqnorm(Prop_category$`building society savings agreement/ life insurance`)
qqline(Prop_category$`building society savings agreement/ life insurance`)
```




```r
qqnorm(Prop_category$`unknown / no property`)
qqline(Prop_category$`unknown / no property`)
```




```r
qqnorm(Prop_category$`car or other, not in attribute Account`)
qqline(Prop_category$`car or other, not in attribute Account`)
```




```r
a = aov(data$CreditAmount ~ data$Property)
summary(a)
```

```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## data$Property    3 8.067e+08 268884018   37.44 <2e-16 ***
## Residuals      996 7.153e+09   7181951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```