Analiza uspješnosti marketinške kampanje

Andrija Petrusić, Matija Luka Kukić, Dominik Gračner, Ivan Džanija

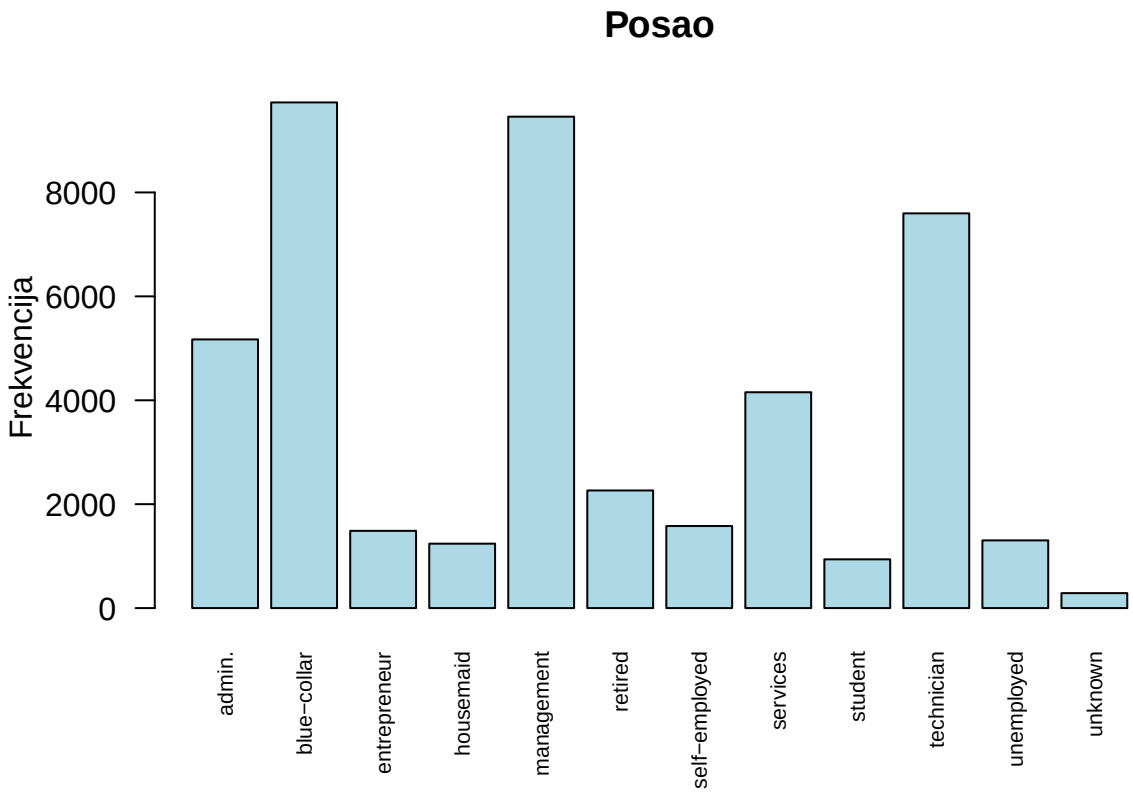
2025-01-15

Deskriptivna statistika

Generalni pregled podataka i vizualizacija.

```
## Uplaćen depozit(uspješna kampanja) - Binarna varijabla
## 39922 5289
##
## Uspješnost prethodne kampanje
## 4901 1840 1511 36959
##
## Bračni status
## 5207 27214 12790
##
## Razina edukacije
## 6851 23202 13301 1857
##
## Ima li stambeni kredit?
## 20081 25130
##
## Ima li osobni zajam?
## 37967 7244
```

```r
barplot(table(marketingData$job), main = "Posao",
        col = "lightblue",
        ylab = "Frekvencija",
        las = 2,
        cex.names = 0.7)
```



Postoji li zavisnost između zanimanja i bračnog statusa klijenta?

Prvo ćemo provjeriti imamo li nedostajućih vrijednosti i uzeti samo stupce koji su nam bitni za ovo testiranje. Također iz stupca "job" ćemo makuti vrijednosti "unknown" jer ne nose nikakvu informaciju za ovo testiranje.

```
rel = marketingData[names(marketingData) %in% c('job','marital_status')]
rel = rel[rel$job != 'unknown', ]
status <- c('job', 'marital_status')
for (colName in status){
  if (sum(is.na(marketingData[,colName])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu ',colName, ': ', sum(is.na(marketingData[,colName]
  }
  else {
    cat('Nema nedostajućih vrijednosti za ', colName, '\n')
  }
}
```

```
## Nema nedostajućih vrijednosti za  job
## Nema nedostajućih vrijednosti za  marital_status
```

Za testiranje zavisnosti zanimanja i bračnog statusa razmatramo χ^2 test nezavinosti.

Pretpostavke:

- kategorički podatci - zadovoljeno
- očekivane frekvencije svake ćelije tablice mora biti minimalno 5

Provjera očekivanih vrijednosti

Izrađujemo kontigencijsku tablicu i provjeravamo kolika je očekivana vrijednost za svaku ćeliju.

```r
tab = addmargins(table(rel$job,rel$marital_status))
cat("\t\tKontigencijska tablica\n")
```

```
##      Kontigencijska tablica
```

```r
print(tab)
```

```
##
##                 divorced married single   Sum
##   admin.             750    2693   1728  5171
##   blue-collar        750    6968   2014  9732
##   entrepreneur       179    1070    238  1487
##   housemaid          184     912    144  1240
##   management        1111    5400   2947  9458
##   retired            425    1731    108  2264
##   self-employed      140     993    446  1579
##   services           549    2407   1198  4154
##   student              6      54    878   938
##   technician         925    4052   2620  7597
##   unemployed         171     731    401  1303
##   Sum               5190   27011  12722 44923
```

```r
cat("H0: Kategorijski podatci su nezavisni\n")
```

```
## H0: Kategorijski podatci su nezavisni
```

```r
cat("H1: Kategorijski podatci nisu nezavisni\n")
```

```
## H1: Kategorijski podatci nisu nezavisni
```

```r
cat("Alpha value = 0.05\n")
```

```
## Alpha value = 0.05
```

```r
chi_squared_result <- chisq.test(tab)
expected_values <- chi_squared_result$expected
for (val in expected_values)
  if (val < 5){
    cat("Očekivana vrijednost manja od 5!")
  }
print(chi_squared_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 3819.6, df = 33, p-value < 2.2e-16
```

Zaključak

Prvo vidimo kako niti jedna očekivana vrijednost nije manja od 5 te zaključujemo da možemo provesti zamišljeni test.

Na temelju testa odbacujemo H0(Kategorijski podatci su nezavisni) u korist H1(Kategorijski podatci nisu nezavisni) te zaključemo da postoji statistički značajna zavisnost između zanimanja i bračnost statusa klijenta na razini značajnosti $\alpha = 5\%$.

Imaju li klijenti s otvorenim kreditom više novca na računu od ostalih klijenata?

Za provjeru zavisnosti financijskog stanja klijenta i trenutno otvorenog kredita razmatramo T-test za dva uzorka.

Pretpostavke:

- Numerički podatci - zadovoljeno(razdvajamo na dvije skupine numeričkih podataka)
- Normalna distribucija podataka

Provjera normalnosti podataka

Uzimamo stupce koji su nam bitni - kredit i stanje računa te dodajmo stupac koji sadrži "yes" ako klijent ima neki od dva kredita, a inače "no".

```
stripped = select(marketingData, c("balance", "housing_loan", "personal_loan"))
stripped$open_any_loan <- ifelse(stripped$housing_loan == "yes" | stripped$personal_loan == "yes", "yes"
summary(stripped)
```

```
##     balance       housing_loan       personal_loan      open_any_loan
##  Min.   : -8019   Length:45211       Length:45211       Length:45211
##  1st Qu.:    72   Class :character   Class :character   Class :character
##  Median :   448   Mode  :character   Mode  :character   Mode  :character
##  Mean   :  1362
##  3rd Qu.:  1428
##  Max.   :102127
```

Provjerimo vrijednosti kategoričkih podataka i nalazimo li na nedostajuće vrijednosti.

```
## [1] "Moguće vrijednosti za stambeni kredit: "
```

```
## [1] "yes" "no"
```

```
## [1] "Moguće vrijednosti za osobni zajam: "
```

```
## [1] "no"  "yes"
```

```
## Broj negativnih stanja računa:  3766
```

```
##
## Dimenzije podataka:  45211 4
```

Vidimo kako nema nedostajućih vrijednosti.

Vizualiziramo podatke i provodimo moguće testove na normalnost podataka.

```
hloan <- table(stripped$housing_loan)
cat("\nIma li stambeni kredit?")
```

```
##
## Ima li stambeni kredit?
```

```
print(hloan)
```

```
##
##    no   yes
## 20081 25130
```

```
ploan <- table(stripped$personal_loan)
cat("\nIma li osobni zajam?")
```

```
##
## Ima li osobni zajam?
```

```
print(ploan)
```

```
##
##    no   yes
## 37967  7244
```

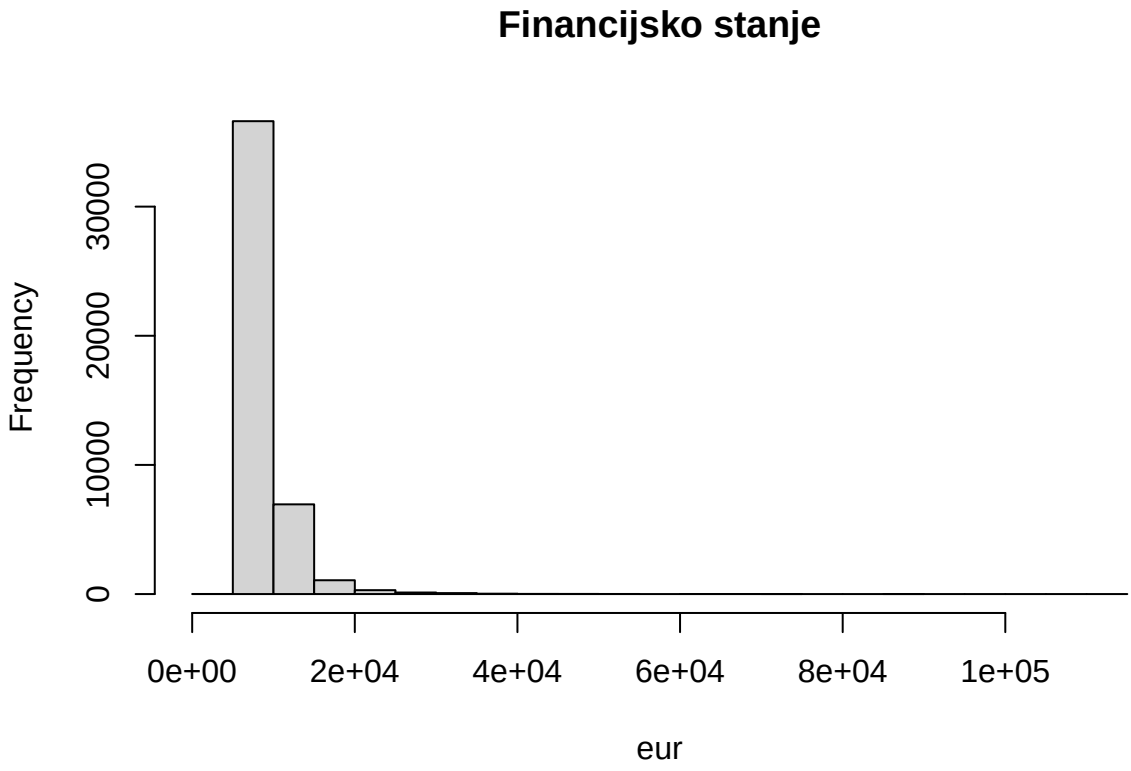
```
aloan <- table(stripped$open_any_loan)
cat("\nIma li osobni zajam?")
```

```
##
## Ima li osobni zajam?
```

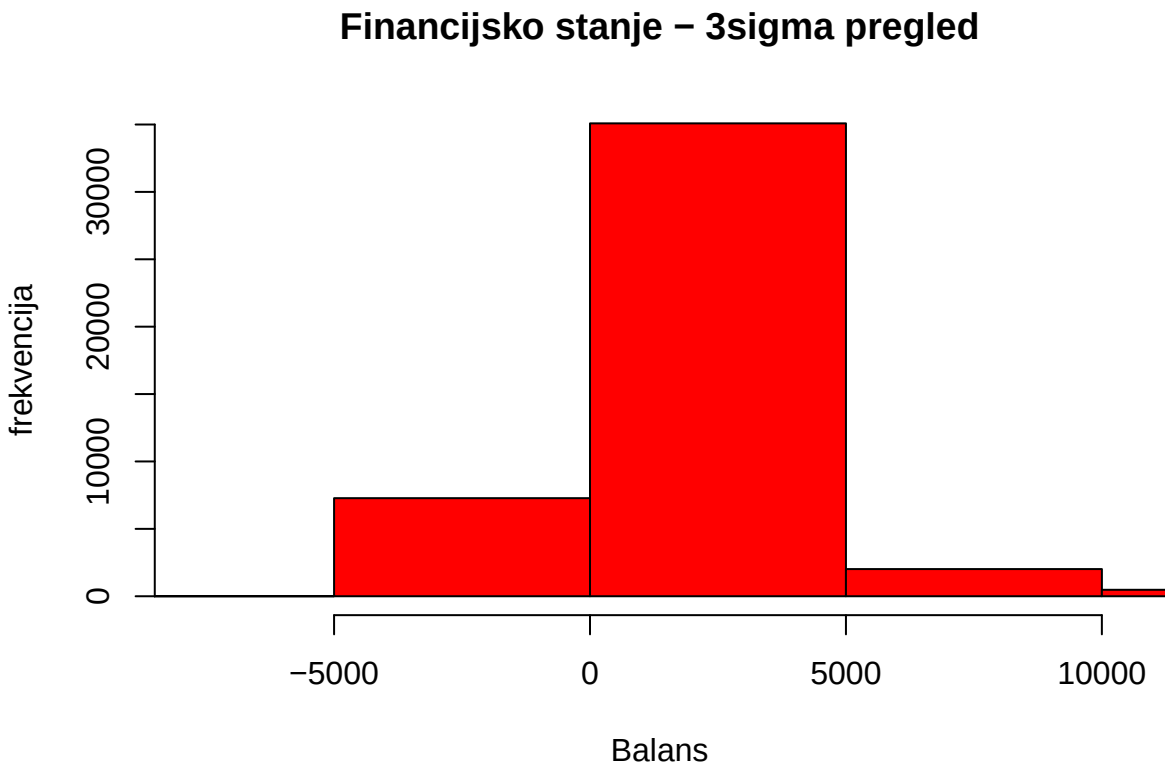
```
print(aloan)
```

```
##
##    no   yes
## 17204 28007
```

```
hist(stripped$balance - min(stripped$balance)+1,main='Financijsko stanje', xlab='eur', ylab='Frequency')
```

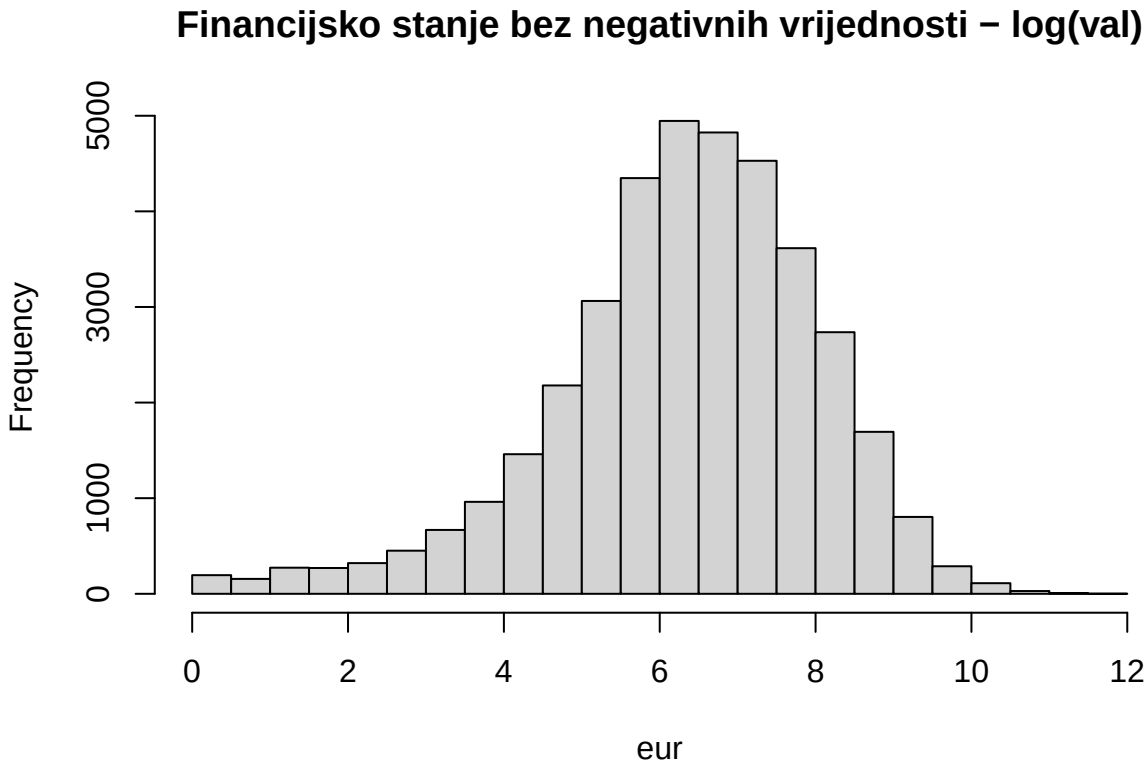



```
balance_mean <- mean(stripped$balance)
balance_sd <- sd(stripped$balance)
h = hist(stripped$balance,
  main="Financijsko stanje - 3sigma pregled",
  xlab="Balans",
  ylab="frekvencija",
  xlim = c(balance_mean - 3 * balance_sd, balance_mean + 3 * balance_sd),
  col="red"
)
```




```
hist(log(stripped$balance),main='Financijsko stanje bez negativnih vrijednosti - log(val)', xlab='eur',
```

```
## Warning in log(stripped$balance): NaNs produced
```



Primjećujemo postojanje velikih outliera, analiziramo njihovu frekvenciju te ih uklanjamo ukoliko nije značajna.

```
stripped$z <- scale(stripped$balance)
summary(stripped$z)
```

```
##        V1
##  Min.   :-3.08111
##  1st Qu.:-0.42377
##  Median :-0.30028
##  Mean   : 0.00000
##  3rd Qu.: 0.02159
##  Max.   :33.09441
```

```
cat('\nbroj vrijednosti sa z-vrijednošću većom od 3.29: ',sum(stripped$z > 3))
```

```
##
## broj vrijednosti sa z-vrijednošću većom od 3.29:  744
```

```
cat('\nbroj vrijednosti sa z-vrijednošću manjom od -3.29: ',sum(stripped$z < -3))
```

```
##
## broj vrijednosti sa z-vrijednošću manjom od -3.29:  1
```

```
cat('\nukupan broj vrijednosti prvog seta: ', sum(stripped$balance))
```

```
##
## ukupan broj vrijednosti prvog seta:  61589682
```

```
final <- data.frame(stripped)
final <- subset(final, balance >= quantile(balance, 0.01) & balance <= quantile(balance, 0.99))
```

Vidimo kako su stršeće vrijednosti stvarno samo manjina podataka te ćemo i maknuti kako bi mogli lakše dalje vizualizirati i provoditi testove bez da pretjerano utječju stršeće vrijednosti. Izbacujemo samo 2% podataka.

```
otvoren = final[final$open_any_loan == 'yes',]
neotvoren = final[final$open_any_loan == 'no',]

cat('Prosječno stanje računa klijenata s otvorenim kreditom: ', mean(otvoren$balance))
```

```
## Prosječno stanje računa klijenata s otvorenim kreditom:  1031.766
```

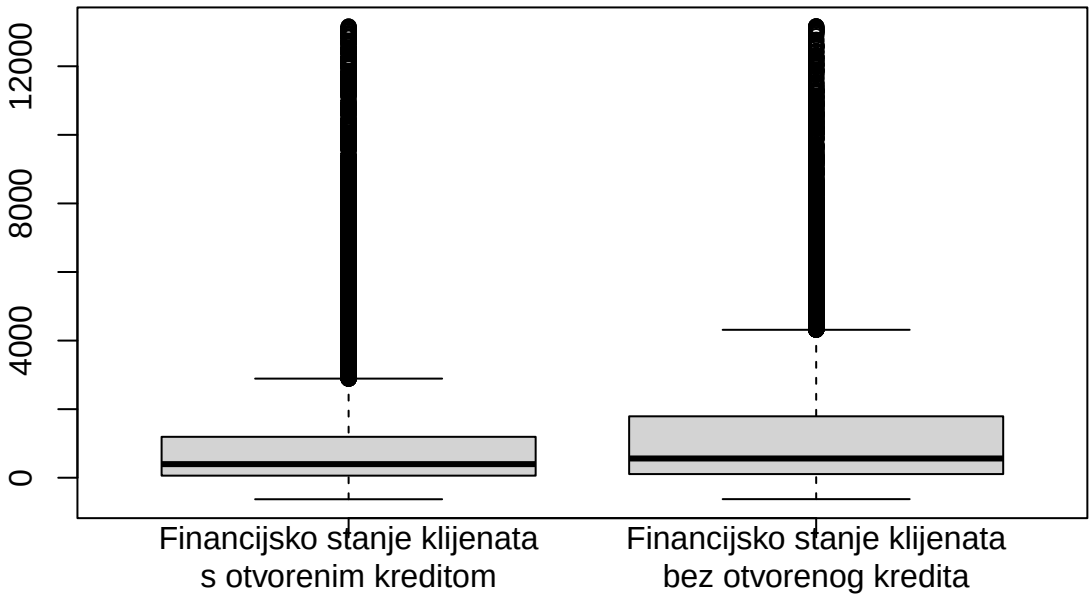
```
cat('\nProsječno stanje računa klijenata bez otvorenog kredita: ', mean(neotvoren$balance))
```

```
##
## Prosječno stanje računa klijenata bez otvorenog kredita:  1409.694
```

```
boxplot(otvoren$balance, neotvoren$balance,
        names = c("Financijsko stanje klijenata\ns otvorenim kreditom",
                  "Financijsko stanje klijenata\nbez otvorenog kredita"),
        main='Usporedba stanja računa')
```

```
## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'Usporedba stanja računa' in 'mbcsToSbcs': dot
## substituted for <c4>
```

```
## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'Usporedba stanja računa' in 'mbcsToSbcs': dot
## substituted for <8d>
```



Sada ćemo vizualizirati histogramom i qq-plotom izgled distribucija te ćemo također provesti Kolmogorov-Smirnov test. Za Kolmogorov-Smirnov moramo testirati specifičnu distribuciju što znači da moramo prosljediti i parametre distribucije prema kojoj hoćemo testirati.

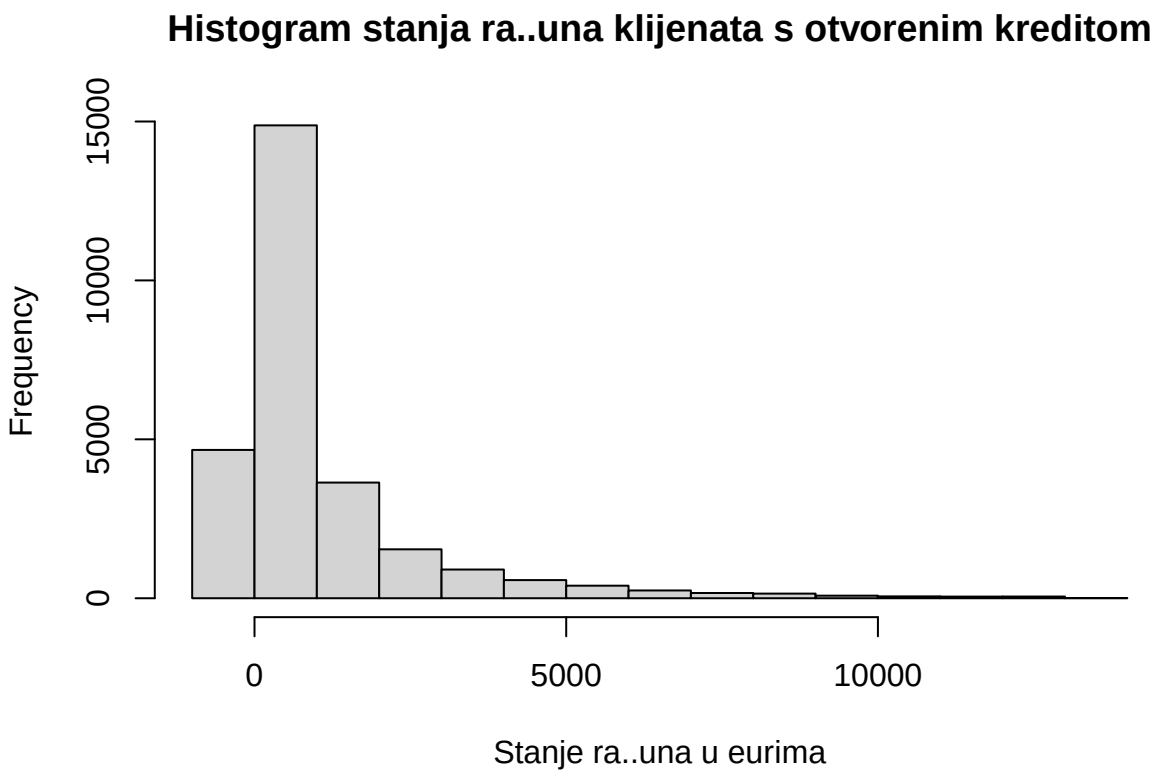
```
hist(otvoren$balance,
     main='Histogram stanja računa klijenata s otvorenim kreditom',
     xlab='Stanje računa u eurima')
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Histogram stanja računa klijenata s otvorenim kreditom'
## in 'mbcsToSbcs': dot substituted for <c4>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Histogram stanja računa klijenata s otvorenim kreditom'
## in 'mbcsToSbcs': dot substituted for <8d>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Stanje računa u eurima' in 'mbcsToSbcs': dot substituted
## for <c4>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Stanje računa u eurima' in 'mbcsToSbcs': dot substituted
## for <8d>
```



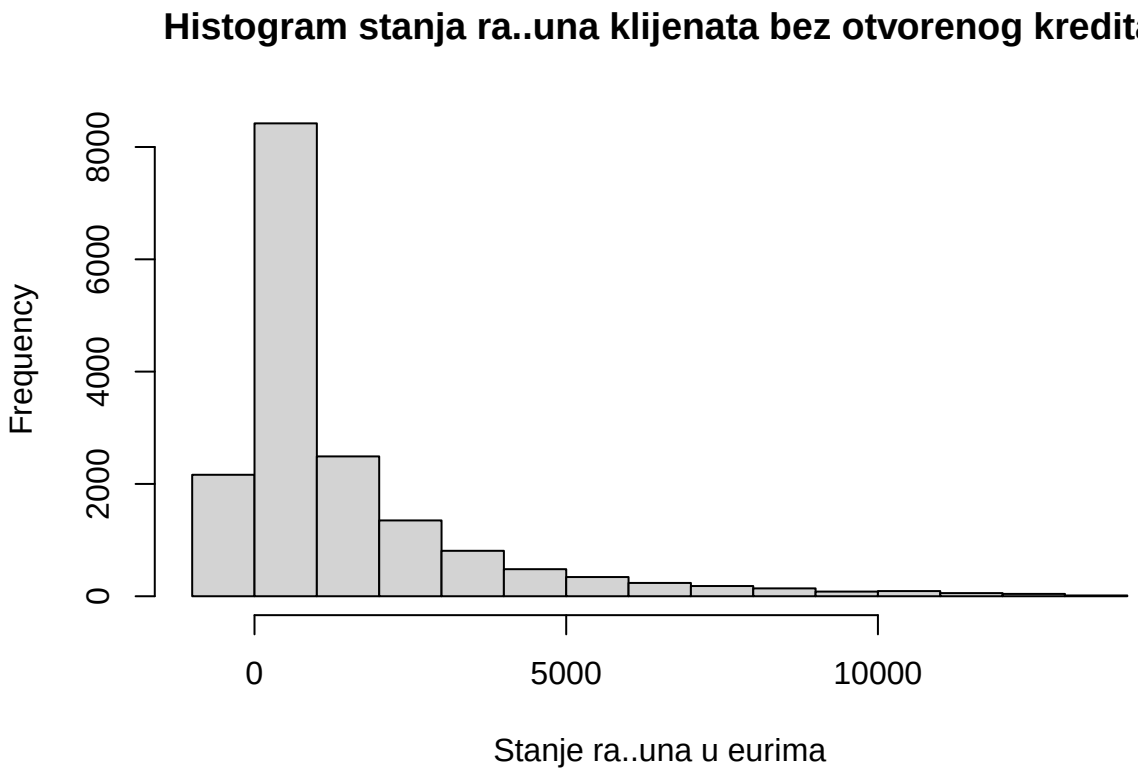

```
hist(neotvoren$balance,
     main='Histogram stanja računa klijenata bez otvorenog kredita',
     xlab='Stanje računa u eurima')
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Histogram stanja računa klijenata bez otvorenog kredita'
## in 'mbcsToSbcs': dot substituted for <c4>
```

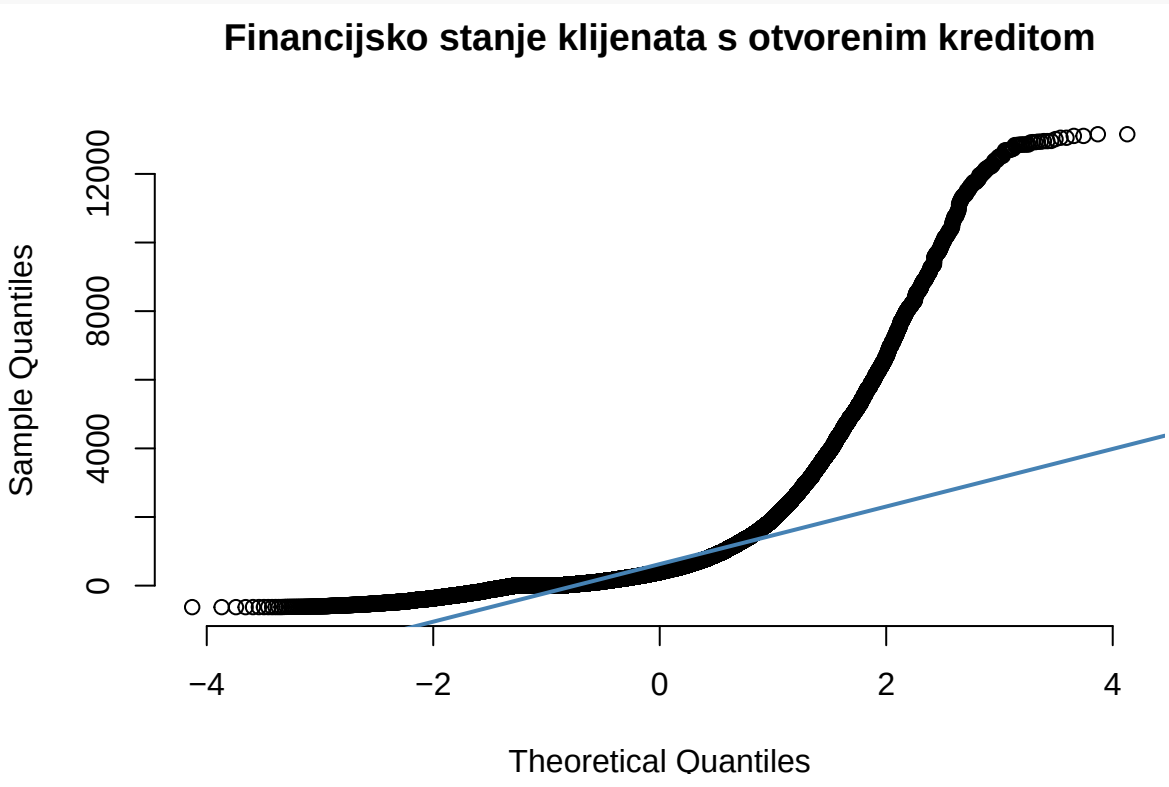
```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Histogram stanja računa klijenata bez otvorenog kredita'
## in 'mbcsToSbcs': dot substituted for <8d>
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Stanje računa u eurima' in 'mbcsToSbcs': dot substituted
## for <c4>
```

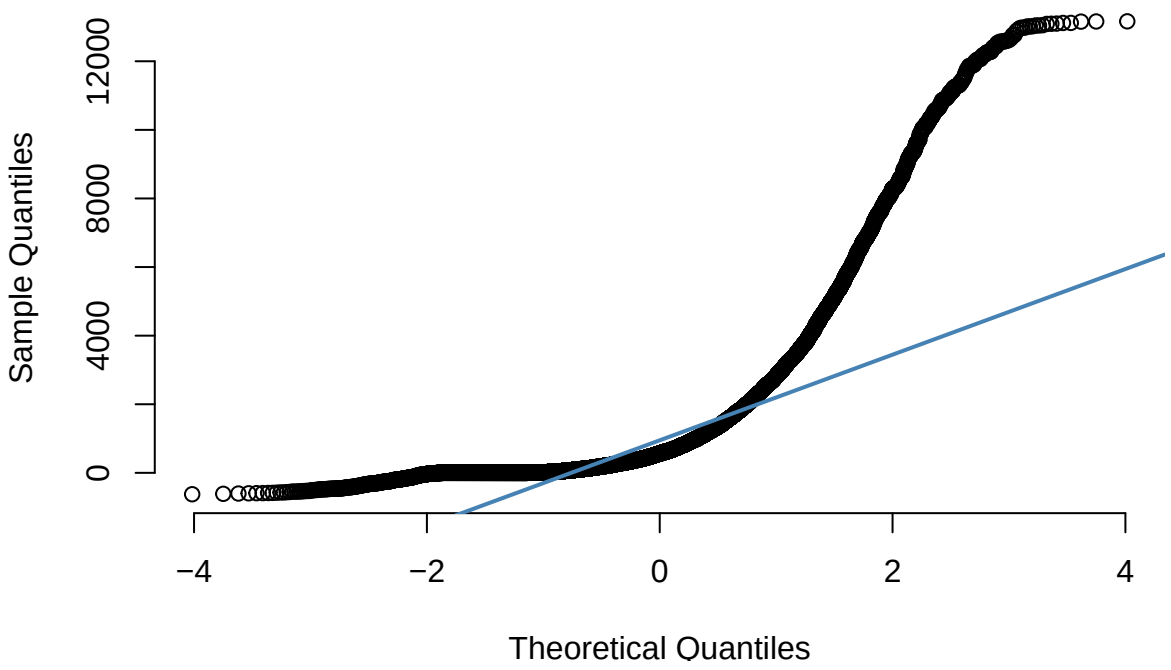
```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## conversion failure on 'Stanje računa u eurima' in 'mbcsToSbcs': dot substituted
## for <8d>
```




```
qqnorm(otvoren$balance, pch = 1, frame = FALSE,main='Financijsko stanje klijenata s otvorenim kreditom'
qqline(otvoren$balance, col = "steelblue", lwd = 2)
```




```
qqnorm(neotvoren$balance, pch = 1, frame = FALSE,main='Financijsko stanje klijenata bez otvorenog kredi
qqline(neotvoren$balance, col = "steelblue", lwd = 2)
```




```
cat("Alpha value = 0.05\n")
```

```
## Alpha value = 0.05
```

```
ks.test(otvoren$balance, "pnorm", mean = mean(otvoren$balance), sd = sd(otvoren$balance))
```

```
## Warning in ks.test.default(otvoren$balance, "pnorm", mean =
## mean(otvoren$balance), : ties should not be present for the Kolmogorov-Smirnov
## test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  otvoren$balance
## D = 0.22155, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(neotvoren$balance, "pnorm", mean = mean(neotvoren$balance), sd = sd(neotvoren$balance))
```

```
## Warning in ks.test.default(neotvoren$balance, "pnorm", mean =
## mean(neotvoren$balance), : ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  neotvoren$balance
## D = 0.22471, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Zaključak: Odbacujemo H0(normalnost distribucije) u koristi H1(nemamo normalnost distribucije) za oba uzorka. Kao što smo mogli i pretpostaviti financijsko stanje klijenata nije normalno distribuirano. Znači ne možemo koristiti T-test za provjeru.

Neparametski test

Pošto nemamo pretpostavku normalnosti ne možemo koristiti T-test te provodimo neparametarski test. Test koji provodimo je Mann-Whitney-Wilcoxonov test/Mann–Whitney U test/Wilcoxon rank-sum test

```
cat("H0: Medijani su jednaki\n")
```

```
## H0: Medijani su jednaki
```

```
cat("H1: Medijani su različiti\n")
```

```
## H1: Medijani su različiti
```

```
cat("Alpha value = 0.05\n")
```

```
## Alpha value = 0.05
```

```
wilcox.test(otvoren$balance, neotvoren$balance, paired = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  otvoren$balance and neotvoren$balance
## W = 204515578, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Zaključak

Nismo mogli provesti T-test jer nismo imali zadovoljenu pretpostavku normalnosti te smo odlučili provesti neparametarski MWW/MWU test za 2 nezavisna uzorka. Na temelju testa odbacujemo H0(medijani su jednaki) u koristi H1(medijani su različiti) na razini značajnosti $\alpha = 5\%$.