

# Analiza uspješnosti marketinške kampanje

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

2025-01-15

```
marketingData <- read.csv("data/data.csv")
# head(marketingData)
```

## Deskriptivna statistika

Generalni pregled podataka i vizualizacija.

```
# summary(marketingData)

deposit_count <- table(marketingData$term_deposit_accepted)
cat("Uplaćen depozit(uspješna kampanja) - Binarna varijabla\n")

## Uplaćen depozit(uspješna kampanja) - Binarna varijabla
print(deposit_count)

##
##      no      yes
## 39922   5289

previous_deposit_count <- table(marketingData$previous_campaign_outcome)
cat("\nUspješnost prethodne kampanje\n")

##
## Uspješnost prethodne kampanje
print(previous_deposit_count)

##
## failure    other success unknown
##     4901     1840     1511   36959

marital_status_count <- table(marketingData$marital_status)
cat("\nBračni status\n")

##
## Bračni status
print(marital_status_count)

##
## divorced   married    single
##     5207     27214     12790

education_count <- table(marketingData$education)
cat("\nRazina edukacije\n")
```

```

##  

## Razina edukacije  

print(education_count)

##  

##   primary secondary tertiary unknown  

##      6851       23202      13301      1857

housing_loan_count <- table(marketingData$housing_loan)
cat("\nIma li stambeni kredit?\n")

##  

## Ima li stambeni kredit?  

print(housing_loan_count)

##  

##   no     yes  

## 20081  25130

personal_loan_count <- table(marketingData$personal_loan)
cat("\nIma li osobni zajam?\n")

##  

## Ima li osobni zajam?  

print(personal_loan_count)

##  

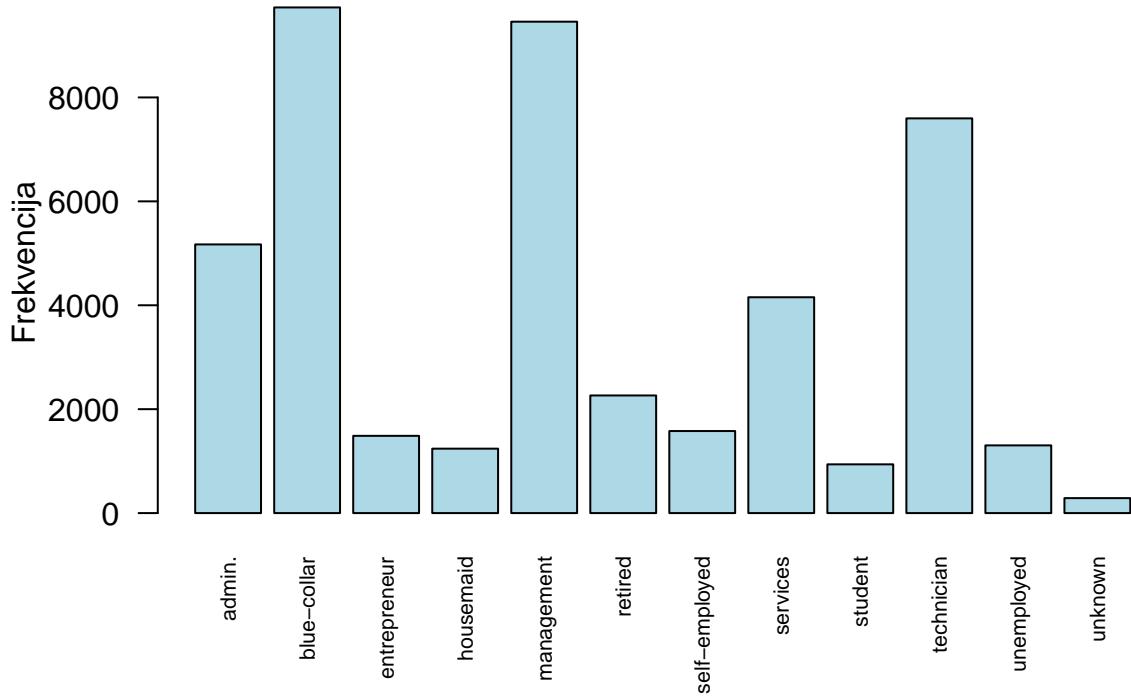
##   no     yes  

## 37967  7244

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

```

## Posao



## *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  $\chi^2$  test nezaviniosti.

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.

```

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

## Kontigencijska tablica
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

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

## H0: Kategorijski podatci su nezavisni
cat("H1: Kategorijski podatci nisu nezavisni\n")

## H1: Kategorijski podatci nisu nezavisni
cat("Alpha value = 0.05\n")

## Alpha value = 0.05
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  $H_0$ (Kategorijski podatci su nezavisni) u korist  $H_1$ (Kategorijski podatci nisu nezavisni) te zaključemo da postoji statistički značajna zavisnost između zanimanja i bračnosti 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.

```
'Moguće vrijednosti za stambeni kredit: '
```

```
## [1] "Moguće vrijednosti za stambeni kredit: "
unique(stripped$housing_loan)
```

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

```
'Moguće vrijednosti za osobni zajam: '
```

```
## [1] "Moguće vrijednosti za osobni zajam: "
unique(stripped$personal_loan)
```

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

```
for (col_name in names(stripped)){
  if (sum(is.na(stripped[,col_name])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu ', col_name, ': ',
        sum(is.na(stripped[,col_name])), '\n')
  }
}
```

```
count = 0
```

```
for(vrijednost in stripped$balance){
  if(vrijednost < 0){
    count = count + 1
  }
}
cat('Broj negativnih stanja računa: ', count)
```

```

## Broj negativnih stanja računa: 3766
cat('\nDimenziije podataka: ',dim(stripped))

##
## Dimenziije 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?")

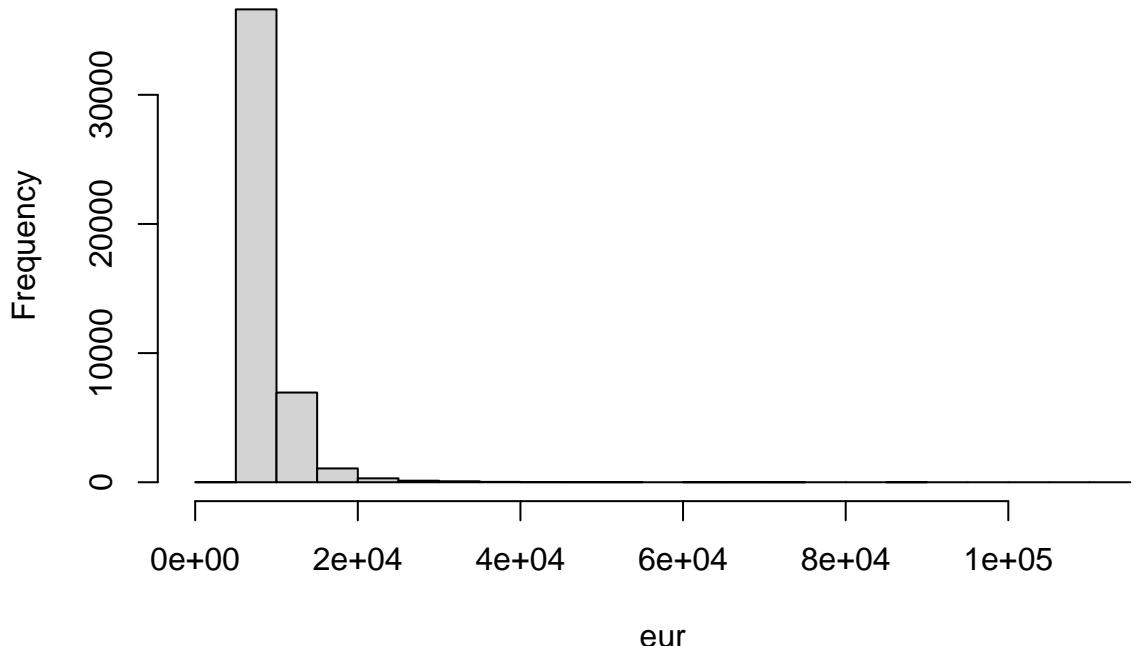
##
## Ima li osobni zajam?
print(aloan)

##
##      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)

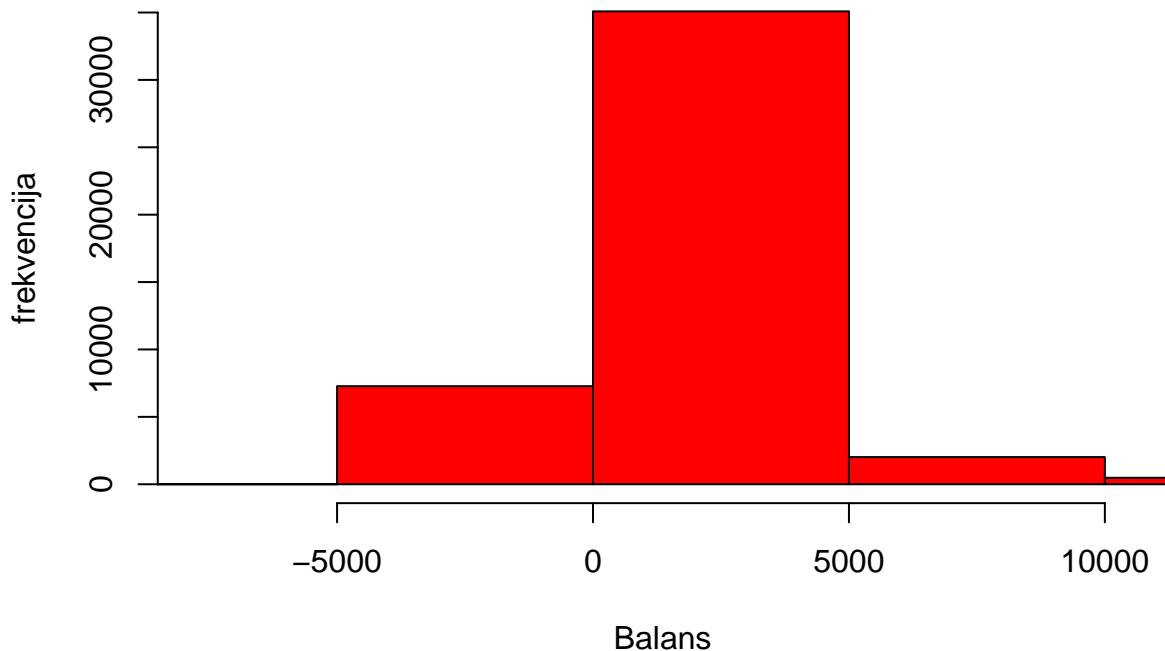
```

## Financijsko stanje



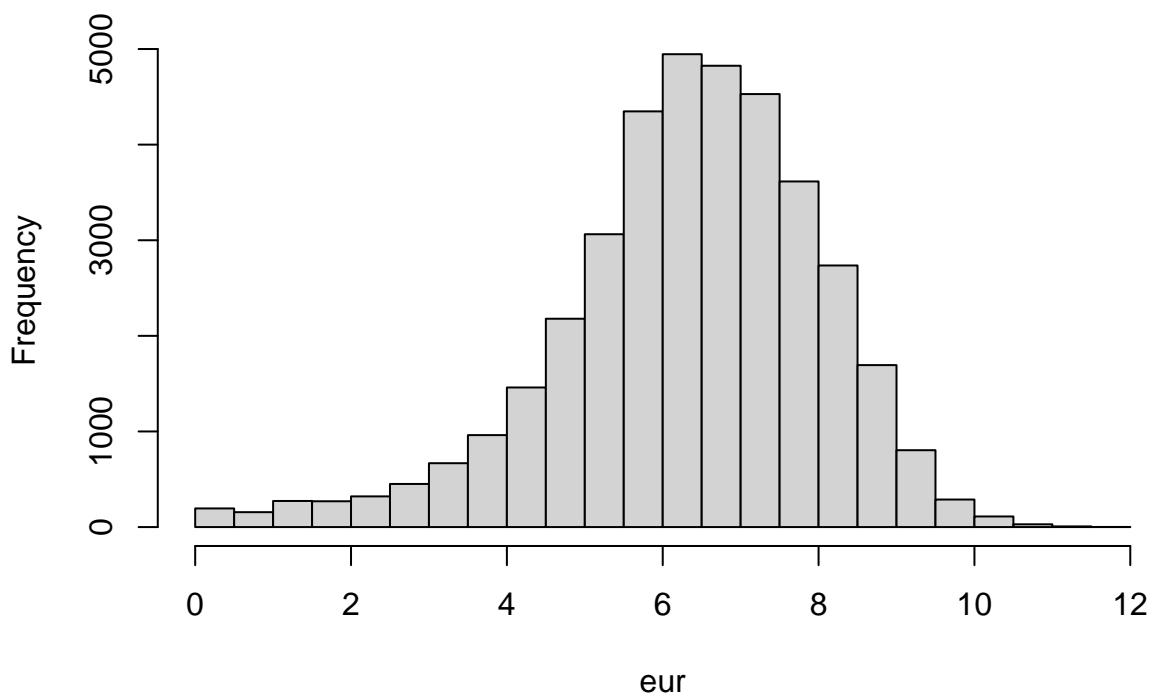
```
h = hist(striped$balance,
  main="Financijsko stanje - 3sigma pregled",
  xlab="Balans",
  ylab="frekvencija",
  xlim = c(balance_mean - 3 * balance_sd, balance_mean + 3 * balance_sd),
  col="red"
)
```

## Financijsko stanje – 3sigma pregled



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

## Financijsko stanje bez negativnih vrijednosti – log(val)



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 utječu stršeće vrijednosti. Izbacujemo samo 2% podataka.

```

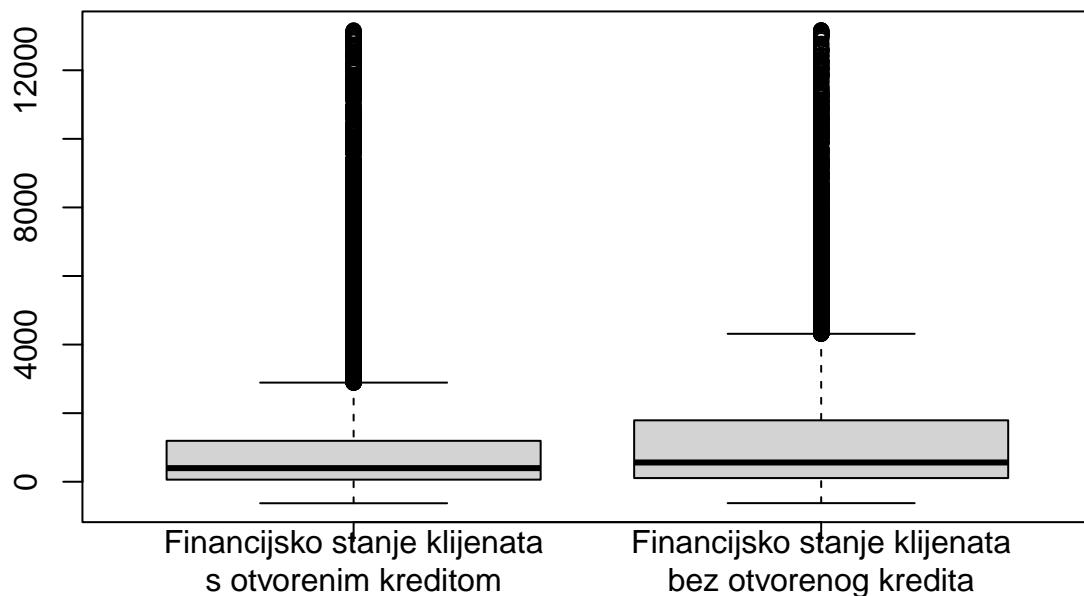
otvoren = final[final$open_any_loan == 'yes',]
neotvoren = final[final$open_any_loan == 'no',]
cat('Prosječno stanje računa klijenata s otvorenim kreditom: ', mean(otvoren$balance))

## 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')

```

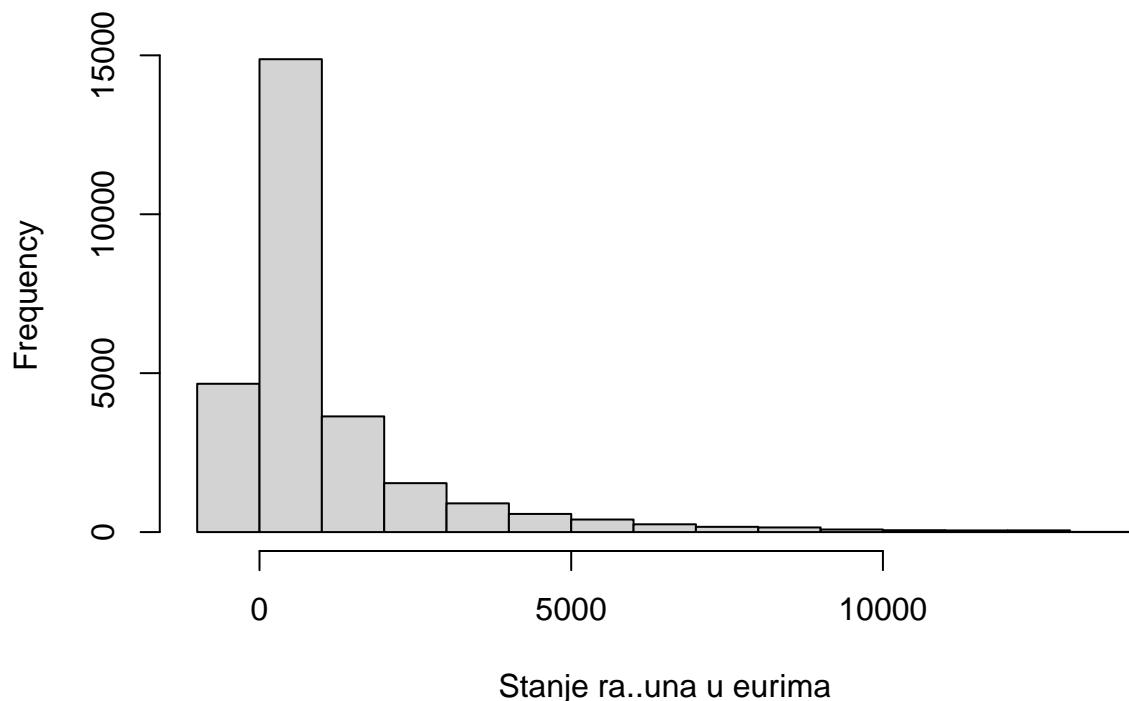
## Usporedba stanja računa



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.

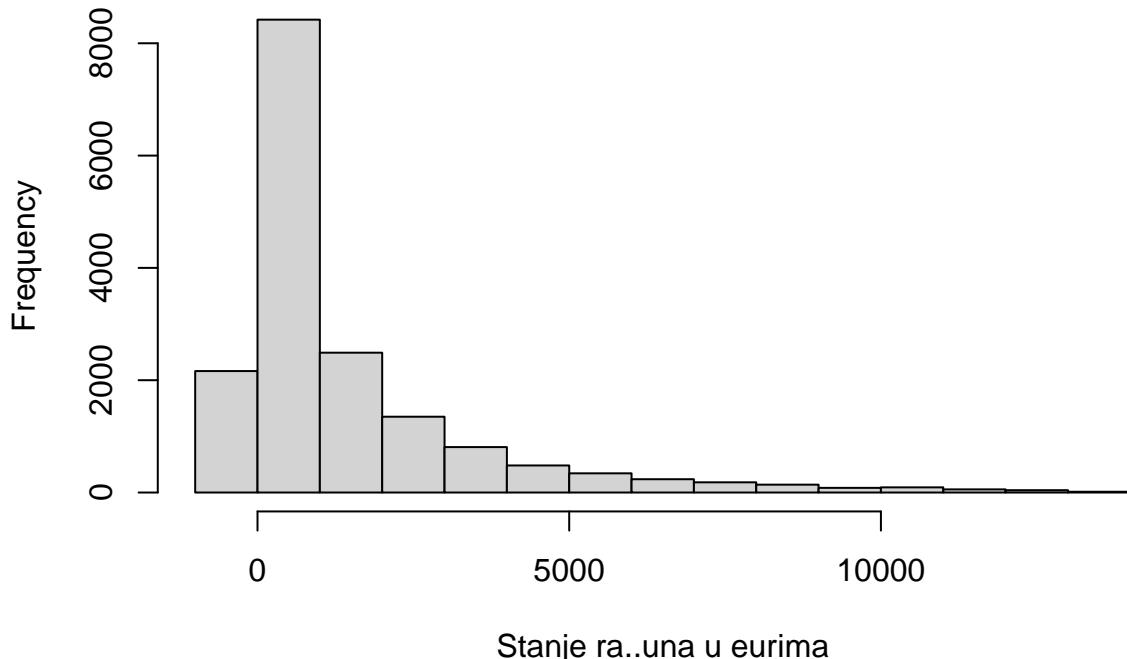
```
hist(otvoren$balance,  
     main='Histogram stanja računa klijenata s otvorenim kreditom',  
     xlab='Stanje računa u eurima')
```

## Histogram stanja računa klijenata s otvorenim kreditom



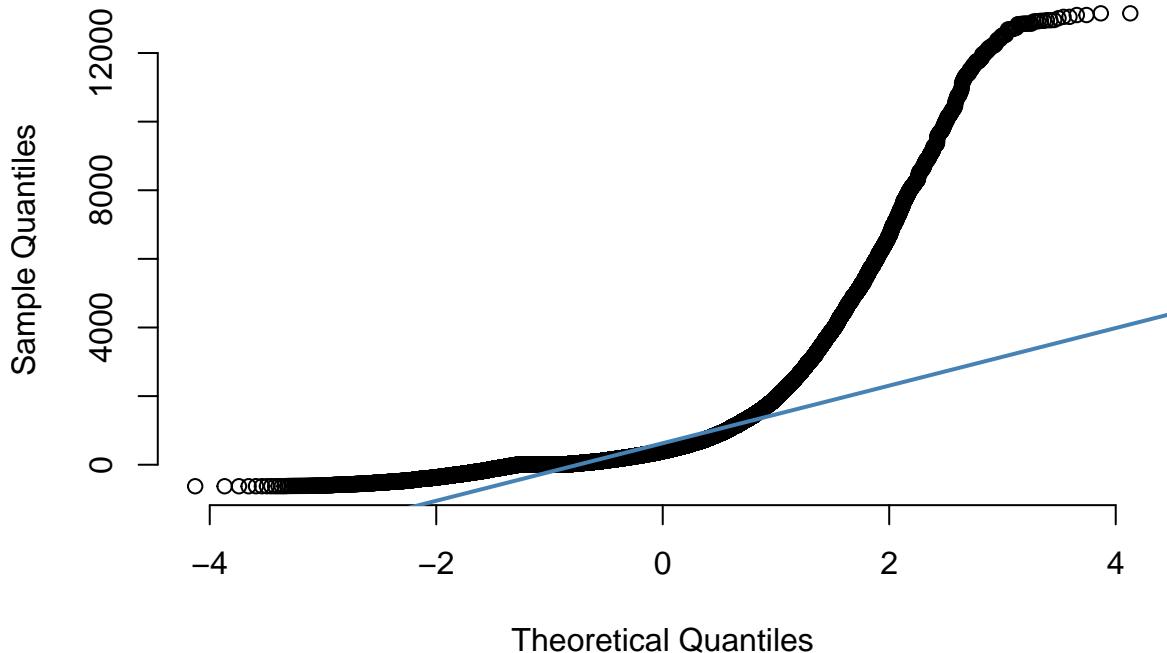
```
hist(neotvoren$balance,  
     main='Histogram stanja računa klijenata bez otvorenog kredita',  
     xlab='Stanje računa u eurima')
```

### Histogram stanja računa klijenata bez otvorenog kredita



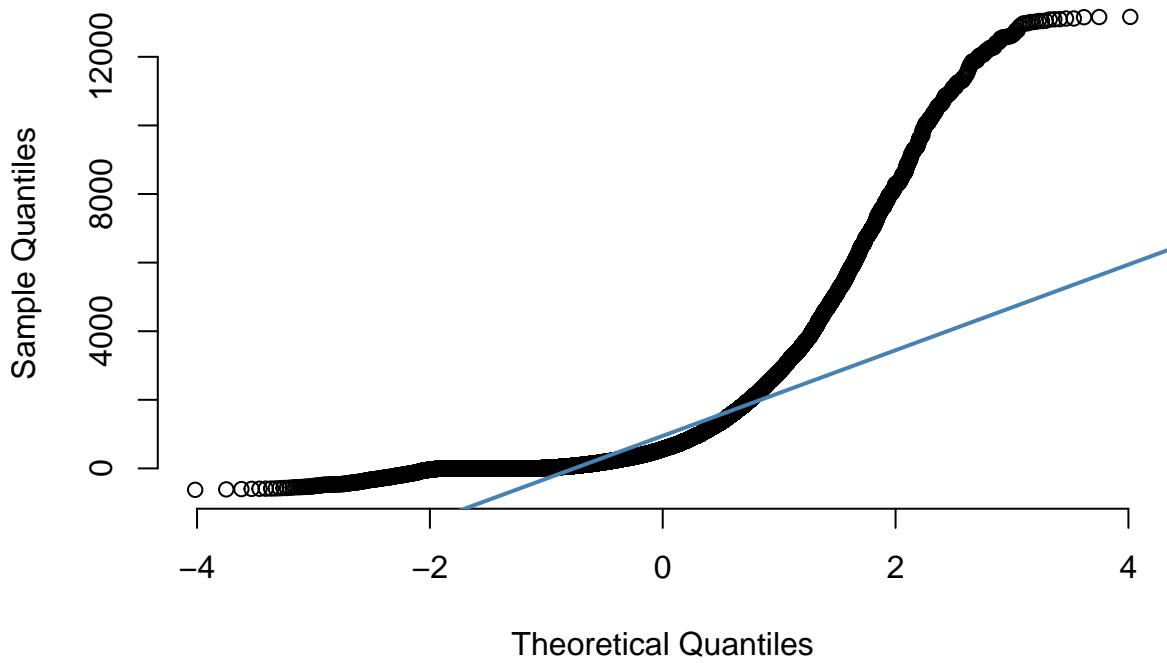
```
qqnorm(otvoren$balance, pch = 1, frame = FALSE,main='Financijsko stanje klijenata s otvorenim kreditom')  
qqline(otvoren$balance, col = "steelblue", lwd = 2)
```

## Financijsko stanje klijenata s otvorenim kreditom



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

## Financijsko stanje klijenata bez otvorenog kreditom



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

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

```

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

##
##  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))

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

```

Zaključak: Odbacujemo  $H_0$ (normalnost distribucije) u koristi  $H_1$ (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  $H_0$ (medijani su jednaki) u koristi  $H_1$ (medijani su različiti) na razini značajnosti  $\alpha = 5\%$ .

## *Postoji li razlika trajanja poziva marketinške kampanje među klijentima različitog stupnja obrazovanja?*

Prvo uzmimo samo podatke koji su nam potrebni za provedbu ovog testa te provjeravamo postoje li nedostajuće vrijednosti. Također ćemo maknuti podatke za "education" s vrijednostima "unknown" jer nam ne nose informaciju u ovom testiranju.

```
stripped = select(marketingData, c("last_contact_duration", "education"))
```

```
for (col_name in names(stripped)){
  if (sum(is.na(stripped[,col_name])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu ', col_name, ': ', sum(is.na(stripped[,col_name])))
  }
}
final <- subset(stripped, education != "unknown")
```

Razdvajamo podatke na 3 različite skupine: Primary, Secondary i Tertiary koje predstavljaju stupnjeve obrazovanja klijenata

```
primary = final[final$education == 'primary',]
secondary = final[final$education == 'secondary',]
tertiary = final[final$education == 'tertiary',]

cat('Prosječno trajanje razgovora - primary: ', median(primary$last_contact_duration))

## Prosječno trajanje razgovora - primary: 178
cat('\nProsječno trajanje razgovora - secondary: ', median(secondary$last_contact_duration))

##
## Prosječno trajanje razgovora - secondary: 183
cat('\nProsječno trajanje razgovora - tertiary: ', median(tertiary$last_contact_duration))

##
```

Za provođenje ovog istraživačko pitanja razmatramo ANOVA test.

Pretpostavke:

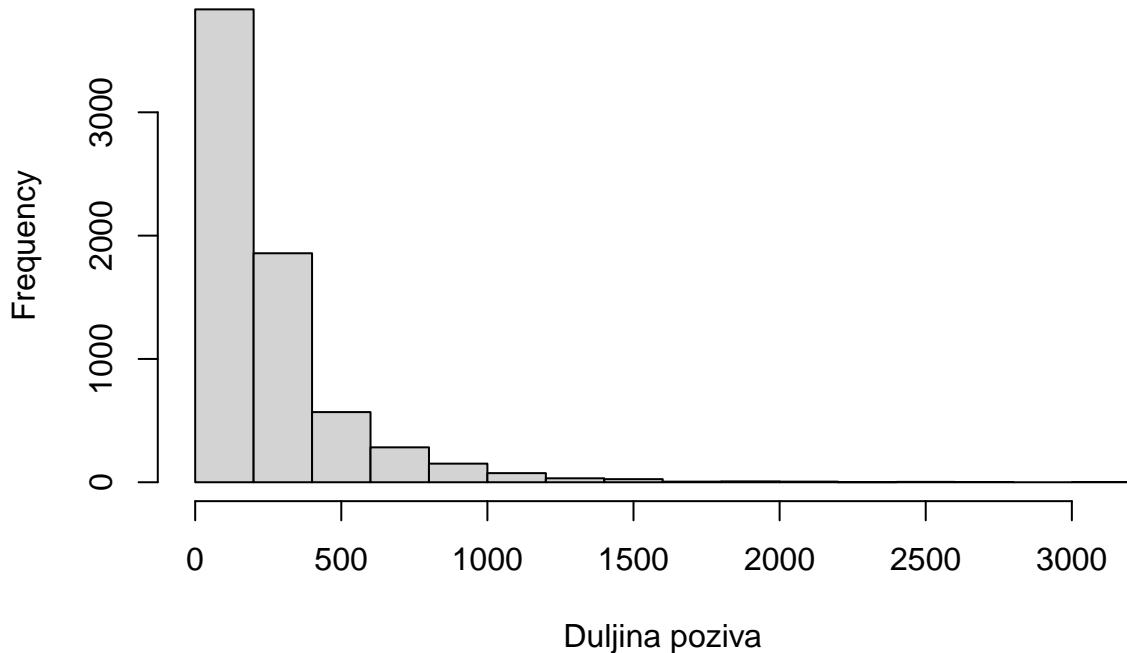
- Normalna distribucija podataka unutar pojedine skupine
- Homogenost varijance između skupina(homoskedastičnost)
- Nezavisnost podataka - zadovoljeno

### **Provjera normalnosti podataka**

Distribuciju podataka unutar svake skupine prikazati ćemo histogramom i qq-plotom. Također testiramo normalnost Lilliforsovom inačicom Kolmogorov-Smirnovljeva testa.

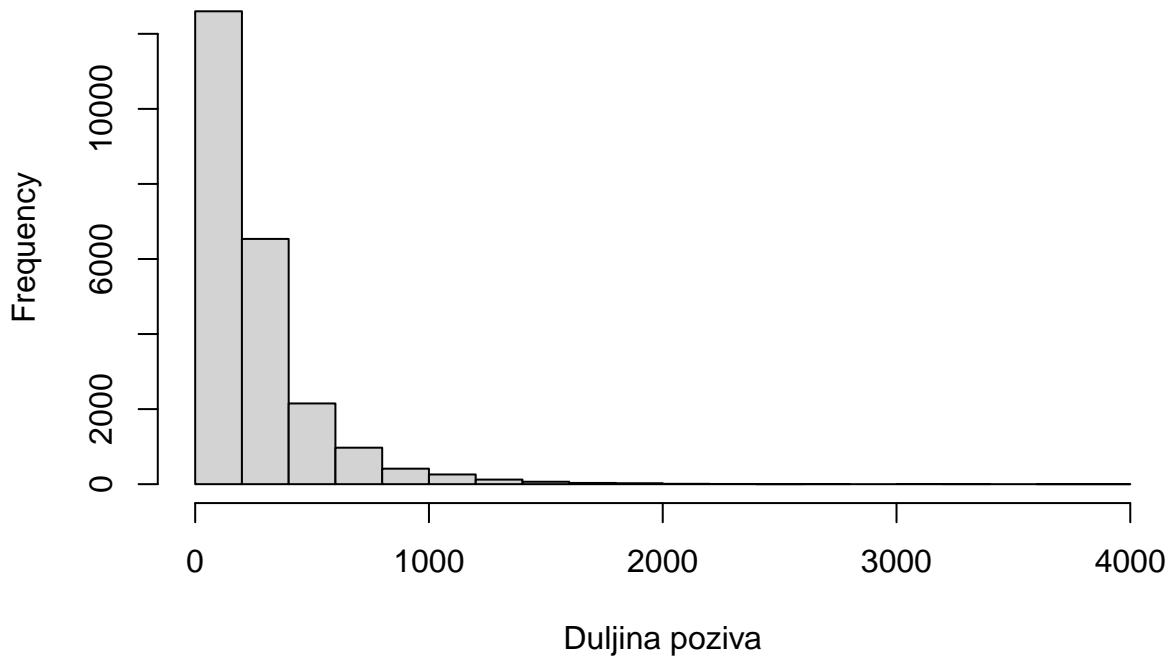
```
library(nortest)
hist(primary$last_contact_duration,
  main='Histogram primary',
  xlab='Duljina poziva')
```

### Histogram primary



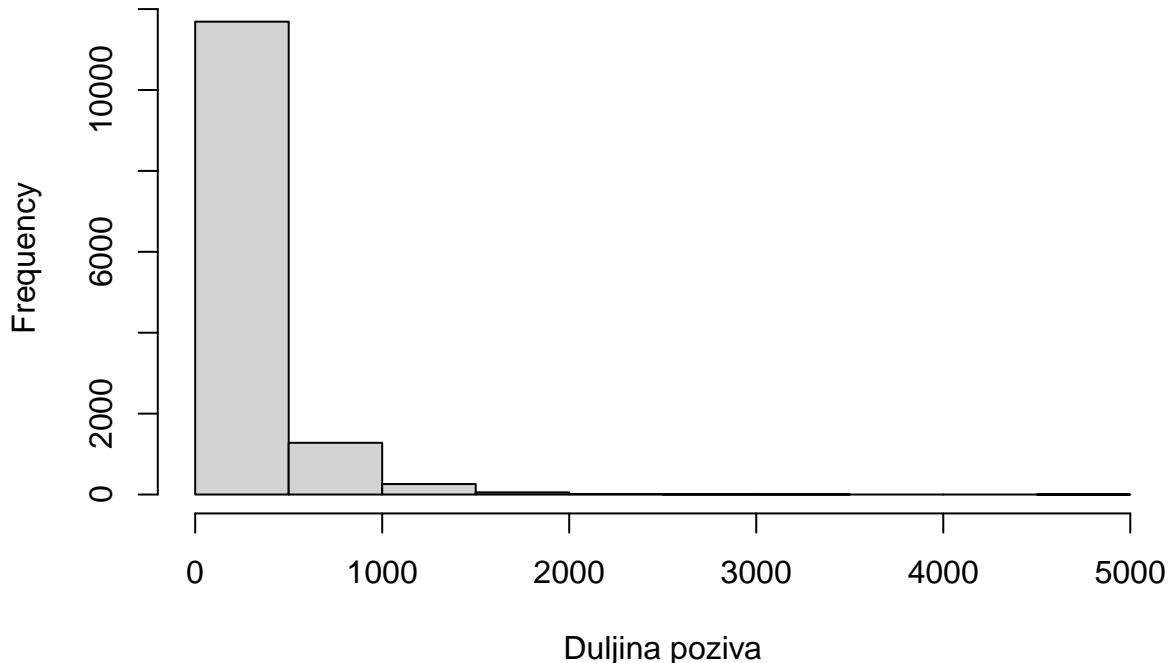
```
hist(secondary$last_contact_duration,  
     main='Histogram secondary',  
     xlab='Duljina poziva')
```

### Histogram secondary



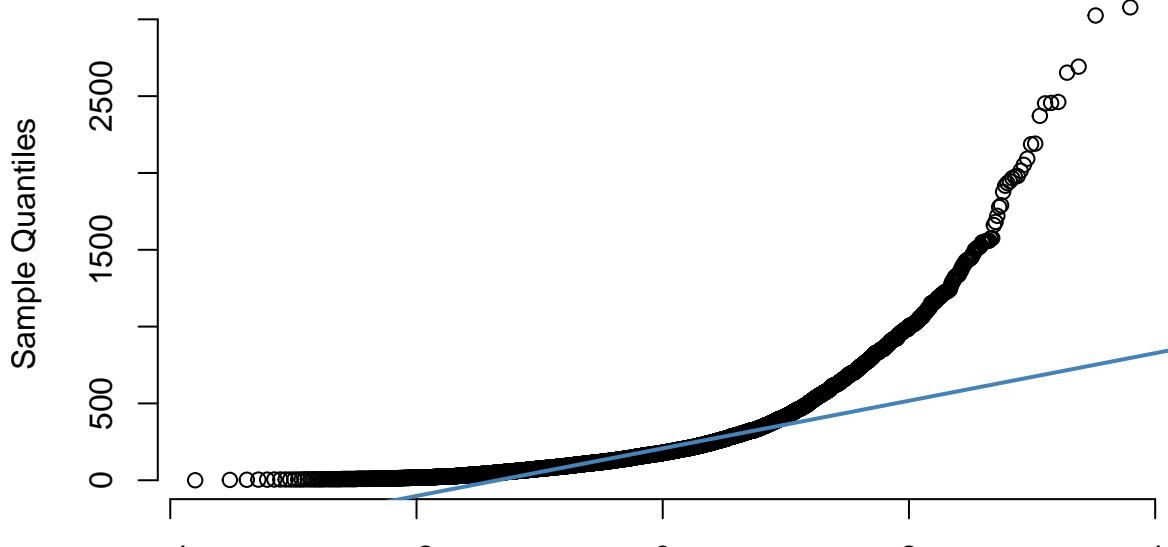
```
hist(tertiary$last_contact_duration,  
     main='Histogram tertiary',  
     xlab='Duljina poziva')
```

Histogram tertiary



```
qqnorm(primary$last_contact_duration, pch = 1, frame = FALSE,main='primary')  
qqline(primary$last_contact_duration, col = "steelblue", lwd = 2)
```

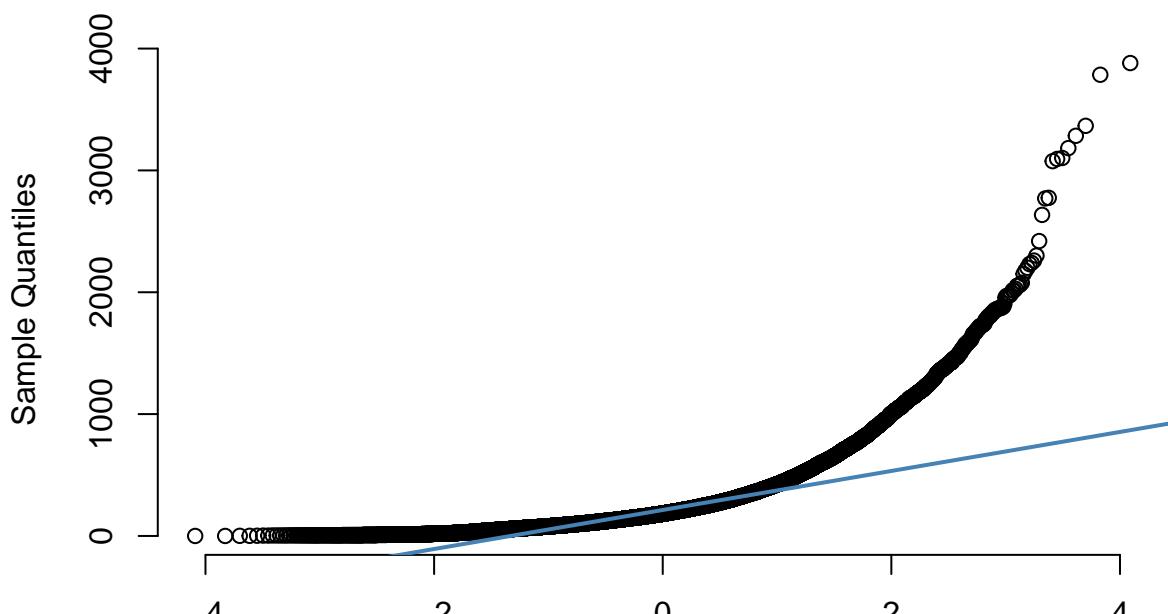
### primary



Theoretical Quantiles

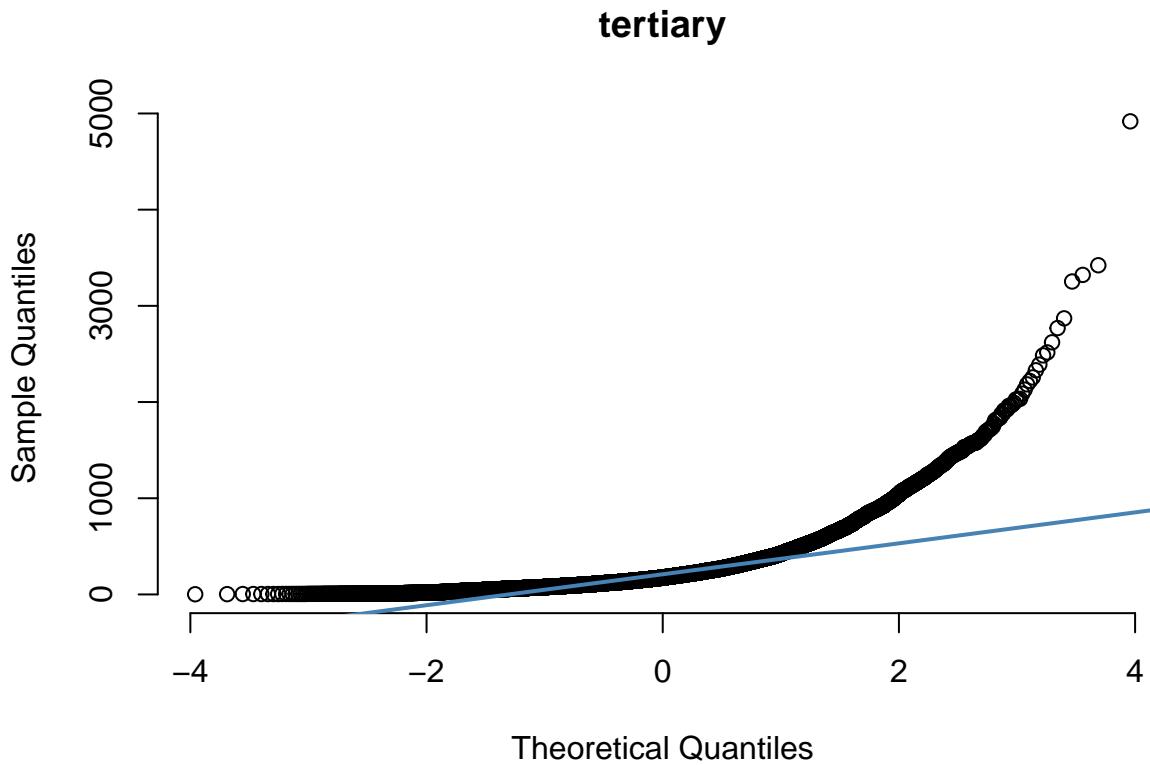
```
qqnorm(secondary$last_contact_duration, pch = 1, frame = FALSE, main='secondary')
qqline(secondary$last_contact_duration, col = "steelblue", lwd = 2)
```

### secondary



Theoretical Quantiles

```
qqnorm(tertiary$last_contact_duration, pch = 1, frame = FALSE, main='tertiary')
qqline(tertiary$last_contact_duration, col = "steelblue", lwd = 2)
```



```

cat("Alpha value = 0.05\n")

## Alpha value = 0.05
lillie.test(primary$last_contact_duration)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: primary$last_contact_duration
## D = 0.17433, p-value < 2.2e-16
lillie.test(secondary$last_contact_duration)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: secondary$last_contact_duration
## D = 0.16607, p-value < 2.2e-16
lillie.test(tertiary$last_contact_duration)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: tertiary$last_contact_duration
## D = 0.175, p-value < 2.2e-16

```

Zaključak: Odbacujemo H<sub>0</sub>(normalnost distribucije) u koristi H<sub>1</sub>(nemamo normalnost distribucije) za sve uzorke. Znači ne možemo koristiti ANOVA test te onda nemamo ni potrebe dalje testirati homoskedastičnost uzoraka.

## Neparametski test

Pošto nemamo pretpostavku normalnosti ne možemo koristiti ANOVA test te provodimo neparametarski test. Test koji provodimo je Kruskal–Wallis test. Jedini uvjet za primjenjivost Kruskal–Wallis testa je: veličina svakog uzorka barem 5 što je zadovoljeno.

```
data <- marketingData
data <- data.frame(
  value = c(primary$last_contact_duration, secondary$last_contact_duration, tertiary$last_contact_duration),
  group = rep(c("primary", "secondary", "tertiary"), times = c(length(primary$last_contact_duration), length(secondary$last_contact_duration), length(tertiary$last_contact_duration)))
)
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
kruskal_result <- kruskal.test(value ~ group, data = data)
print(kruskal_result)

##
## Kruskal-Wallis rank sum test
##
## data: value by group
## Kruskal-Wallis chi-squared = 11.465, df = 2, p-value = 0.003238
```

## Zaključak

Nismo mogli provesti ANOVA test jer nismo imali zadovoljenu pretpostavku normalnosti te smo odlučili provesti neparametarski Kruskal–Wallis test. Na temelju testa odbacujemo  $H_0$ (medijani su jednaki) u koristi  $H_1$ (medijani su različiti) na razini značajnosti  $\alpha = 5\%$ .

## *Mogu li dostupne varijable predvidjeti uspješnost marketinške kampanje?*

Prvo ćemo pogledati korelacije između podataka.

```
cor_matrix <- cor(marketingData[, c("age", "balance", "previous_contacts_count", "campaign_contacts_count", "last_contact_duration")])
cor_matrix

##                                     age      balance previous_contacts_count
## age                         1.000000000  0.09778274          0.001288319
## balance                      0.097782739  1.00000000          0.016673637
## previous_contacts_count     0.001288319  0.01667364          1.000000000
## campaign_contacts_count    0.004760312 -0.01457828         -0.032855290
## last_contact_duration      -0.004648428  0.02156038          0.001203057
##                                         campaign_contacts_count last_contact_duration
## age                               0.004760312           -0.004648428
## balance                          -0.014578279            0.021560380
## previous_contacts_count        -0.032855290            0.001203057
## campaign_contacts_count       1.000000000           -0.084569503
```

```
## last_contact_duration           -0.084569503           1.0000000000
```

Ne vidimo veliku korelaciju između ovih podataka.

### Logistička regresija

```
marketingData$job <- as.factor(marketingData$job)
marketingData$job <- relevel(marketingData$job, ref = "unknown")
marketingData$marital_status <- as.factor(marketingData$marital_status)
marketingData$education <- as.factor(marketingData$education)
marketingData$education <- relevel(marketingData$education, ref = "unknown")
marketingData$previous_campaign_outcome <- as.factor(marketingData$previous_campaign_outcome)
marketingData$previous_campaign_outcome <- relevel(marketingData$previous_campaign_outcome, ref = "unkn")
marketingData$housing_loan <- as.factor(marketingData$housing_loan)
marketingData$personal_loan <- as.factor(marketingData$personal_loan)
marketingData$term_deposit_accepted <- ifelse(marketingData$term_deposit_accepted == "yes", 1, 0)
marketingData$housing_loan <- ifelse(marketingData$housing_loan == "yes", 1, 0)
marketingData$personal_loan <- ifelse(marketingData$personal_loan == "yes", 1, 0)

marketingData %>%
  count(term_deposit_accepted)

##   term_deposit_accepted      n
## 1                      0 39922
## 2                      1  5289

weights <- ifelse(marketingData$term_deposit_accepted == 1, 9, 1)
model <- glm(term_deposit_accepted ~ age + job + marital_status + education + balance + default + housin
summary(model)

##
## Call:
## glm(formula = term_deposit_accepted ~ age + job + marital_status +
##       education + balance + default + housing_loan + personal_loan +
##       last_contact_duration + previous_contacts_count + campaign_contacts_count +
##       previous_campaign_outcome, family = binomial(), data = marketingData,
##       weights = weights)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.690e+00  1.298e-01 -13.024  < 2e-16 ***
## age                     2.257e-03  1.067e-03   2.115  0.034425 *
## jobadmin.                6.076e-01  1.151e-01   5.281  1.29e-07 ***
## jobblue-collar            2.271e-02  1.144e-01   0.199  0.842647
## jobentrepreneur           3.976e-02  1.241e-01   0.320  0.748626
## jobhousemaid              8.604e-02  1.254e-01   0.686  0.492501
## jobmanagement             3.436e-01  1.142e-01   3.009  0.002621 **
## jobretired                 1.023e+00  1.177e-01   8.687  < 2e-16 ***
## jobself-employed           7.532e-02  1.220e-01   0.617  0.537075
## jobservices                1.567e-01  1.167e-01   1.343  0.179380
## jobstudent                  1.204e+00  1.231e-01   9.786  < 2e-16 ***
## jobtechnician               3.246e-01  1.141e-01   2.845  0.004445 **
## jobunemployed                3.878e-01  1.227e-01   3.161  0.001574 **
## marital_statusmarried      -1.132e-01  2.919e-02  -3.879  0.000105 ***
## marital_statussingle        2.468e-01  3.364e-02   7.336  2.20e-13 ***
```

```

## educationprimary           -3.584e-01  5.128e-02 -6.989 2.77e-12 ***
## educationsecondary        -7.777e-02  4.557e-02 -1.707 0.087880 .
## educationtertiary         2.644e-01  4.801e-02  5.507 3.66e-08 ***
## balance                   3.087e-05  3.036e-06 10.165 < 2e-16 ***
## defaultyes                -3.337e-01  7.568e-02 -4.409 1.04e-05 ***
## housing_loan               -1.094e+00  1.964e-02 -55.704 < 2e-16 ***
## personal_loan              -6.615e-01  2.861e-02 -23.119 < 2e-16 ***
## last_contact_duration     5.506e-03  4.477e-05 122.985 < 2e-16 ***
## previous_contacts_count   2.707e-02  5.367e-03  5.044 4.56e-07 ***
## campaign_contacts_count   -1.253e-01  4.525e-03 -27.699 < 2e-16 ***
## previous_campaign_outcomefailure 6.355e-01  3.147e-02 20.193 < 2e-16 ***
## previous_campaign_outcomeother 8.606e-01  4.532e-02 18.990 < 2e-16 ***
## previous_campaign_outcomesuccess 2.987e+00  5.075e-02 58.861 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 120658  on 45210  degrees of freedom
## Residual deviance:  76390  on 45183  degrees of freedom
## AIC: 76446
##
## Number of Fisher Scoring iterations: 6

```

Vidimo kako imamo nekoliko parametara koji nisu značajni, ali ti parametri su specifični unutar kategorija te ih nismo uspjeli kategoriski razdvojiti.

Prije izrade novog modela ćemo pogledati neke mjere kvalitete modela.

```
yHat <- model$fitted.values > 0.5
tab <- table(marketingData$term_deposit_accepted,yHat)
cat("\n")
```

```
tab
```

```
##      yHat
##      FALSE  TRUE
##      0 32308 7614
##      1  993 4296
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])
cat("\naccuracy:", accuracy)
```

```
##
## accuracy: 0.809626
cat("\nprecision:", precision)
```

```
##
## precision: 0.3607053
cat("\nrecall:", recall)
```

```
##
## recall: 0.8122518
```

```

cat("\nspecificity:", specificity)

##
## specificity: 0.9701811
F1 = 2 * ((precision*recall)/(precision+recall))
cat("\nF1:", F1)

##
## F1: 0.4995639

Rsq = 1 - model$deviance/model>null.deviance
cat("RSQ", Rsq)

## RSQ 0.3668921

```

Vidimo da imamo točnost preko 80% dok nam je preciznost tek nešto iznad 36%. U trenutnom slučaju smatramo F1 kao najbolju opisnu mjeru ovog modela jer balansira točnost i preciznost. To nam je bitno jer znamo da imamo veliku neuravnoteženost između 0 i 1 u podatcima te bi i "najgluplji" model mogao u ovakvom skupu podataka imati popriličnost visoku točnost.

Razmatramo maknuti varijablu "previouse\_contact\_count" te provodimo Kruskall-Wallis test kako bi provjerili postojanje korelacije.

```

stripped <- select(marketingData, c("previous_contacts_count", "previous_campaign_outcome"))
final <- subset(stripped, previous_campaign_outcome != "unknown")
# Micanje stršećih vrijednosti
final <- subset(final, previous_contacts_count <=50)

prev_success = final[final$previous_campaign_outcome=="success",]
prev_failure = final[final$previous_campaign_outcome=="failure",]
prev_other = final[final$previous_campaign_outcome=="other",]
cat("Srednje vrijednosti:\n")

## Srednje vrijednosti:
cat( mean(prev_success$previous_contacts_count), '\n')
## 3.075447
cat( mean(prev_failure$previous_contacts_count), '\n')
## 2.876097
cat( mean(prev_other$previous_contacts_count), '\n')

## 3.832427

data <- data.frame(
  value = c(prev_success$previous_contacts_count, prev_failure$previous_contacts_count, prev_other$previous_contacts_count),
  group = rep(c("success", "failure", "other"), times = c(length(prev_success$previous_contacts_count),
) )
kruskal_result <- kruskal.test(value ~ group, data = data)
print(kruskal_result)

##
##  Kruskal-Wallis rank sum test
##
## data:  value by group
## Kruskal-Wallis chi-squared = 57.067, df = 2, p-value = 4.055e-13

```

Primjenom Kruskal-Wallis odbacujemo hipotezu H0(medijani su isti) u koristi H1(medijani su različiti) i time zapravo vidimo kako postoji korelacija između broja komunikacija i ishoda prošle kampanje te možemo maknuti jedan od tih regresora. Prepostavili smo da nema razloga da neki razgovori traju dulje ili kraće u ovisnosti u ishodu kampanje ukoliko nisu korelirani.

Ponovo izrađujemo model logističke regresije.

```
reduced_model <- glm(term_deposit_accepted ~ age + job + last_contact_duration + default + housing_lo
# summary(reduced_model)

yHat <- reduced_model$fitted.values > 0.5
tab <- table(marketingData$term_deposit_accepted,yHat)
tab

##      yHat
##      FALSE  TRUE
## 0 32273 7649
## 1 1022 4267

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])

cat("\naccuracy:", accuracy)

##
## accuracy: 0.8082104
cat("\nprecision:", precision)

##
## precision: 0.35809
cat("\nrecall:", recall)

##
## recall: 0.8067688
cat("\nspecificity:", specificity)

##
## specificity: 0.9693047
F1 = 2 * (precision*recall)/(precision+recall)
cat("\nF1:", F1)

##
## F1: 0.4960186
Rsq = 1 - reduced_model$deviance/reduced_model>null.deviance
cat("\nRSQ: ", Rsq)

##
## RSQ: 0.3606931
```

## Zaključak

Prvo što smo uspjeli postići je smanjenje broja regresora, a da pri tome dobijemo jako bliske rezultate kao i s potpunim modelom. To nam je naravno bio i cilj jer time lakše možemo donositi daljnje zaključke i lakše

možemo objasniti neke pojave. Također zaključujemo da je izrazito teško odrediti što specifično utječe na uspješnost kampanje jer i dalje imamo dosta regresora. I dalje ne možemo reći da model može predvidjeti uspješnost kampanje iako ima dobru točnost zato što mu je preciznost poprilično loša. Uzrok tome bi mogli biti razni faktori, ali mislimo da je najveći faktor neuravnoteženost između distribucije nula i jedinica u samom uzorku. To je dakako teško promjeniti jer uvijek očekujemo da će izrazito mali broj ljudi uzeti kredit uz neku kampanju.

Možda bi nam uspješnost kampanje bila bolje interpretirana da napravimo usporedbu između dvije vrste kampanje ili nečega sličnog kao što je i navedeno u opisu i motivaciji problema.