Goodreads – analiza popularnosti knjiga

Transupstancija

Mario Hladek, Mihael Kožul, Matija Sever, Mirta Vučinić

2022-12-16

Motivacija i opis projekta

Goodreads, kao društvena mreža za obožavatelje čitanja, svojim korisnicima omogućuje pretraživanje i ocjenjivanje velikog kataloga knjiga. Zahvaljujući tome, nastala je iscrpna Goodreads baza podataka koja sadrži atribute poput naslova knjige, formata knjige, imena autora, ocjene i komentara korisnika i dr. Skup podataka koji je korišten unutar projekta odgovara knjigama na popisu Goodreads Best Book Ever te sadrži čak 52,478 knjige.

Cilj projekta

Cilj projekta pod nazivom "Goodreads – analiza popularnosti knjiga" jest na temelju dostupnog skupa podataka odgovoriti na naredna pitanja:

- Postoje li razlike u ocjenama knjiga s obzirom na žanr
- Jesu li knjige s manje stranica jeftinije
- Možete li odrediti popularnost knjige na temelju dostupnih varijabli
- Postoje li razlike u popularnosti knjiga s obzirom na njihovu starost
- Možete li na temelju dostupnih varijabli odrediti je li knjiga bila nagrađivana,

te pritom saznati i naučiti nešto novo.

Skup podataka

```
#Učitavanje podataka
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(MLmetrics)
```

```
## Warning: package 'MLmetrics' was built under R version 4.2.2
```

```
##
## Attaching package: 'MLmetrics'
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

```
data <- read.csv("Goodreads-dataset.csv", sep = ";", header = TRUE)
```

```
#Pregled podataka
head(data)
```

```
##   X             title            series                              author
## 1 0 Attracted to Fire                       DiAnn Mills (Goodreads Author)
## 2 1         Elemental Soul Guardians #2 Kim Richardson (Goodreads Author)
## 3 2      Unbelievable      Port Fare #2  Sherry Gammon (Goodreads Author)
## 4 3         Fractured        Fateful #2  Cheri Schmidt (Goodreads Author)
## 5 4           Anasazi Sense of Truth #2                      Emma Michaels
## 6 5            Marked Soul Guardians #1 Kim Richardson (Goodreads Author)
##   rating language
## 1   4.14  English
## 2   4.07  English
## 3   4.16  English
## 4   4.00  English
## 5   4.19  English
## 6   3.70  English
##
## 1 ['Christian Fiction', 'Christian', 'Suspense', 'Romance', 'Mystery', 'Romantic Suspense', 'Fiction
## 2                   ['Fantasy', 'Young Adult', 'Angels', 'Romance', 'Paranormal', 'Demons', 'Fiction
## 3                ['Romance', 'Young Adult', 'Contemporary', 'Contemporary Romance', 'Suspense', 'Abus
## 4                    ['Vampires', 'Paranormal', 'Young Adult', 'Romance', 'Fantasy', 'Paranormal Roma
## 5
## 6                     ['Fantasy', 'Young Adult', 'Paranormal', 'Angels', 'Romance', 'Demons', 'Superr
##       bookFormat pages               publisher
## 1      Paperback   416 Tyndale House Publishers
## 2 Kindle Edition   151           Kim Richardson
## 3      Paperback   360  Wordpaintings Unlimited
## 4           Nook     0            Cheri Schmidt
## 5      Paperback   190       Bokheim Publishing
## 6      Paperback   280              CreateSpace
##                                                                                 awards
## 1 ['HOLT Medallion by Virginia Romance Writers Nominee for Long Inspirational (2012)']
## 2                                                                                   []
```

```
## 3                                                                                []
## 4                                                                                []
## 5                                                                                []
## 6                                                  ["Readers' Favorite Book Award (2011)"]
##   numRatings                          ratingsByStars likedPercent price  genre1
## 1       2143      ['945', '716', '365', '78', '39']           95  5.55 Fiction
## 2       1947      ['801', '636', '391', '84', '35']           94       Fiction
## 3       1028      ['442', '384', '142', '48', '12']           94 19.18   Other
## 4        871      ['311', '310', '197', '42', '11']           94       Fiction
## 5         37            ['16', '14', '5', '2', '0']           95         Other
## 6       6674 ['2109', '1868', '1660', '647', '390']           84  7.37 Fiction
##        genre2
## 1     Romance
## 2     Fantasy
## 3     Romance
## 4 Young Adult
## 5 Young Adult
## 6     Fantasy
```

```
summary(data)
```

```
##        X            title              series             author
##  Min.   :    0   Length:52478       Length:52478       Length:52478
##  1st Qu.:13119   Class :character   Class :character   Class :character
##  Median :26239   Mode  :character   Mode  :character   Mode  :character
##  Mean   :26239
##  3rd Qu.:39358
##  Max.   :52477
##
##      rating        language           genres           bookFormat
##  Min.   :0.000   Length:52478       Length:52478       Length:52478
##  1st Qu.:3.820   Class :character   Class :character   Class :character
##  Median :4.030   Mode  :character   Mode  :character   Mode  :character
##  Mean   :4.022
##  3rd Qu.:4.230
##  Max.   :5.000
##
##     pages            publisher           awards            numRatings
##  Length:52478       Length:52478       Length:52478       Min.   :      0
##  Class :character   Class :character   Class :character   1st Qu.:    341
##  Mode  :character   Mode  :character   Mode  :character   Median :   2307
##                                                           Mean   :  17879
##                                                           3rd Qu.:   9380
##                                                           Max.   :7048471
##
##  ratingsByStars      likedPercent       price              genre1
##  Length:52478       Min.   :  0.00   Length:52478       Length:52478
##  Class :character   1st Qu.: 90.00   Class :character   Class :character
##  Mode  :character   Median : 94.00   Mode  :character   Mode  :character
##                     Mean   : 92.23
##                     3rd Qu.: 96.00
##                     Max.   :100.00
##                     NA's   :622
##     genre2
```

```
##  Length:52478
##  Class :character
##  Mode  :character
##
##
##
##
```

Opis dataset-a:

"title": Naslov knjige.

"series": Serija kojoj knjiga pripada, ako postoji.

"author": Autor ili autori knjige.

"rating": Prosječna ocjena koju je knjiga dobila, određena prema recenzijama ili čitateljima.

"language": Jezik na kojem je knjiga napisana.

"genres": Popis žanrova u kojima se knjiga nalazi.

"bookForm": Format knjige, kao što su tvrdi uvez, meki uvez ili elektronska knjiga.

"pages": Broj stranica u knjizi.

"publisher": Izdavač knjige.

"awards": Nagrade koje je knjiga dobila.

"numRatings": Broj ocjena koje je knjiga dobila.

"ratingByStars": Razdioba ocjena koje je knjiga dobila, razvrstana po broju zvjezdica (npr. ocjene s 5 zvjezdica, ocjene s 4 zvjezdice itd.).

"likedPercent": Postotak čitatelja koji su voljeli knjigu.

"price": Cijena knjige.

"genre1": Prvi žanr knjige.

"genre2": Drugi žanr knjige.

```
# Dimenzije dataseta:
dim(data)  # broj redaka, broj stupaca (broj primjera, broj varijabli)
```

```
## [1] 52478    17
```

```
# Tip podataka unutar dataseta
str(data)
```

```
## 'data.frame':    52478 obs. of  17 variables:
##  $ X             : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ title         : chr  "Attracted to Fire" "Elemental" "Unbelievable" "Fractured" ...
##  $ series        : chr  "" "Soul Guardians #2" "Port Fare #2" "Fateful #2" ...
##  $ author        : chr  "DiAnn Mills (Goodreads Author)" "Kim Richardson (Goodreads Author)" "Sherry
##  $ rating        : num  4.14 4.07 4.16 4 4.19 3.7 3.85 4.02 4.09 4.67 ...
##  $ language      : chr  "English" "English" "English" "English" ...
##  $ genres        : chr  "['Christian Fiction', 'Christian', 'Suspense', 'Romance', 'Mystery', 'Roman
##  $ bookFormat    : chr  "Paperback" "Kindle Edition" "Paperback" "Nook" ...
```

```
##  $ pages         : chr  "416" "151" "360" "0" ...
##  $ publisher     : chr  "Tyndale House Publishers" "Kim Richardson" "Wordpaintings Unlimited" "Cheri
##  $ awards        : chr  "['HOLT Medallion by Virginia Romance Writers Nominee for Long Inspirational
##  $ numRatings    : int  2143 1947 1028 871 37 6674 238 246 6196 3 ...
##  $ ratingsByStars: chr  "['945', '716', '365', '78', '39']" "['801', '636', '391', '84', '35']" "['4
##  $ likedPercent  : num  95 94 94 94 95 84 90 90 94 NA ...
##  $ price         : chr  "5.55" "" "19.18" "" ...
##  $ genre1        : chr  "Fiction" "Fiction" "Other" "Fiction" ...
##  $ genre2        : chr  "Romance" "Fantasy" "Romance" "Young Adult" ...
```

Mjere centralne tendencije

Mjere centralne tendencije (ili središnje mjere) opisuju skup podataka jednom vrijednošću oko koje se podatci grupiraju. Najčešće korištene mjere centralne tendencije su: aritmetička sredina, medijan, mod i podrezana aritmetička sredina.

```
sprintf("Aritmetička sredina (mean)= %f",mean(data$rating))
```

```
## [1] "Aritmetička sredina (mean)= 4.021878"
```

```
sprintf("Podrezana aritmetička sredina s uklanjanjem po 20%% najmanjih i najvećih podataka = %f",mean(da
```

```
## [1] "Podrezana aritmetička sredina s uklanjanjem po 20% najmanjih i najvećih podataka = 4.027296"
```

```
sprintf("Medijan - robusna mjera centralne tendencije(točno 50%% podataka je manje i 50%% podataka veće
```

```
## [1] "Medijan - robusna mjera centralne tendencije(točno 50% podataka je manje i 50% podataka veće od
```

```
print("1., 2. i 3. kvartil")
```

```
## [1] "1., 2. i 3. kvartil"
```

```
quantile(data$rating, probs = c(0.25,0.5,0.75))
```

```
##  25%  50%  75%
## 3.82 4.03 4.23
```

1.Postoje li razlike u ocjenama knjiga s obzirom na žanr?

Imamo jedan numerički stupac rating odnosno ocijenu knjige i dva kategorička stupca genre1 i genre2. Kada imamo kombinaciju numeričkih i kategoričkih varijabli najbolje je koristiti ANOVA-u.

ANOVA je statistički test za procjenu kako se kvantitativna zavisna varijabla mijenja prema razinama jedne ili više kategoričkih nezavisnih varijabli.Jedan od glavnih ciljeva analize varijance je ustanoviti jesu li upravo te razlike između grupa samo posljedica slučajnosti ili je statistički značajna.

```r
#Spajanje genre1 genre2 i ratings u dataframe
data_subset <- cbind(data$genre1, data$genre2)
data_subset <- cbind(data_subset, data$rating)

colnames(data_subset) <- c("genre1", "genre2", "rating")

data_subset <- as.data.frame(data_subset)

#Aritmetička sredina ratings-a zavisna o genre1 i genre2
mean_ratings <- data_subset %>%
  group_by(genre1, genre2) %>%
  summarise(mean_rating = mean(as.numeric(rating))) %>%
  ungroup() %>%
  arrange(desc(mean_rating))
```

```
## `summarise()` has grouped output by 'genre1'. You can override using the
## `.groups` argument.
```

```r
mean_ratings
```

```
## # A tibble: 57 x 3
##    genre1 genre2    mean_rating
##    <chr>  <chr>           <dbl>
##  1 Poetry Drama            4.38
##  2 Poetry Religion         4.24
##  3 Other  Religion         4.23
##  4 Poetry Other            4.22
##  5 Other  Adventure        4.20
##  6 Other  Other            4.19
##  7 Poetry War              4.18
##  8 Other  Memoir           4.18
##  9 Poetry Fantasy          4.17
## 10 Poetry Memoir           4.17
## # ... with 47 more rows
```

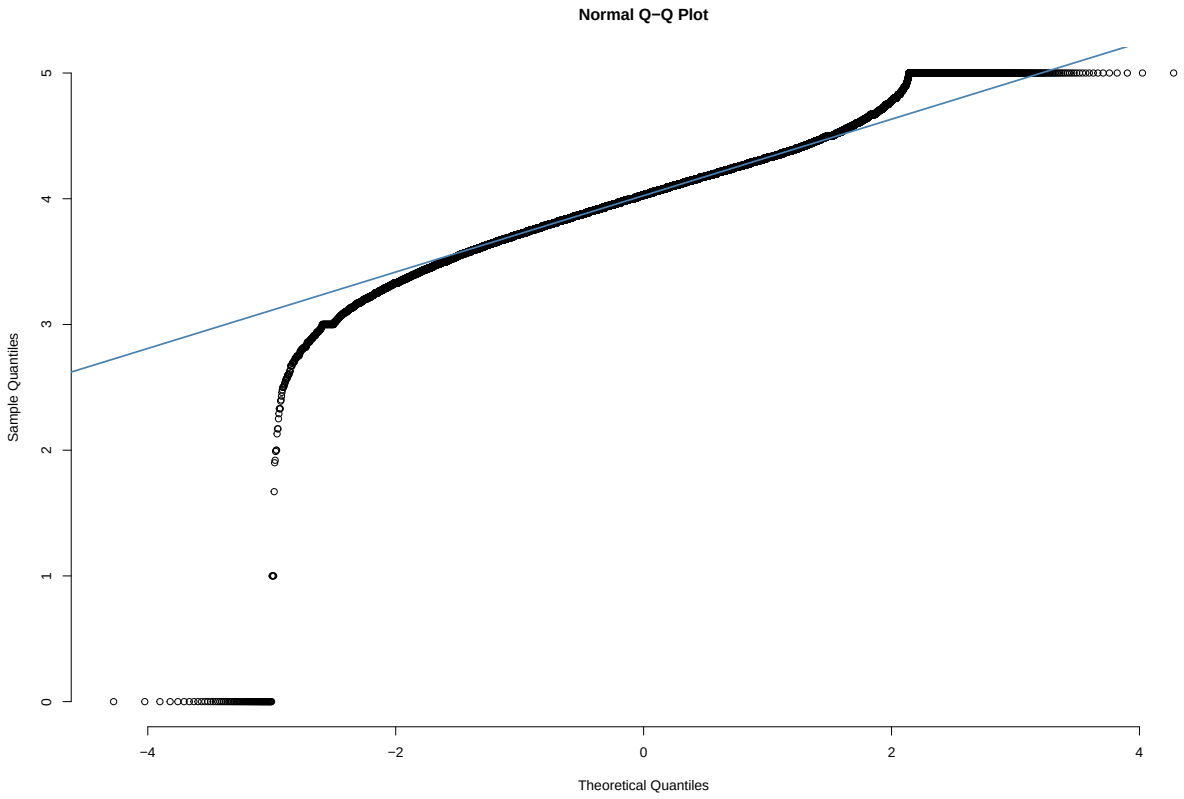
```r
#dvofaktorska anova - graficka provjera, provjera normalnosti i homogenosti varijanci

require(nortest)
```

```
## Loading required package: nortest
```

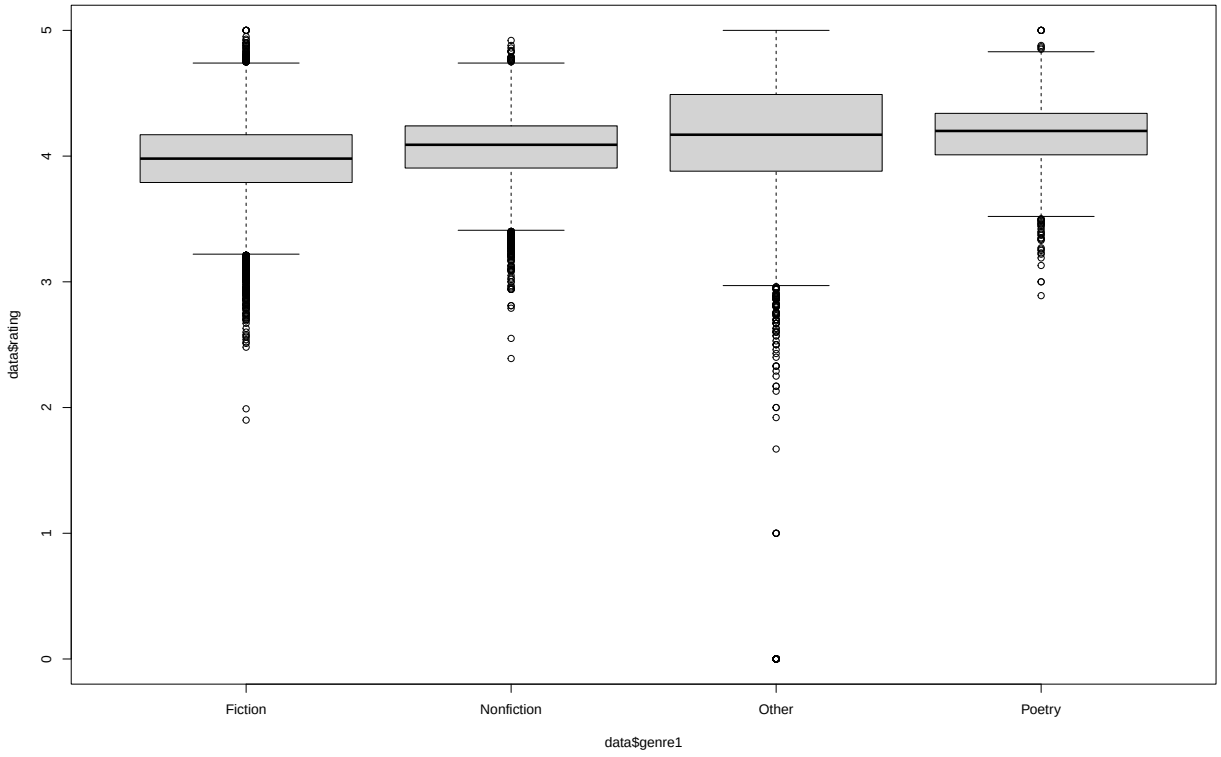
```r
lillie.test(data$rating)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$rating
## D = 0.051635, p-value < 2.2e-16
```

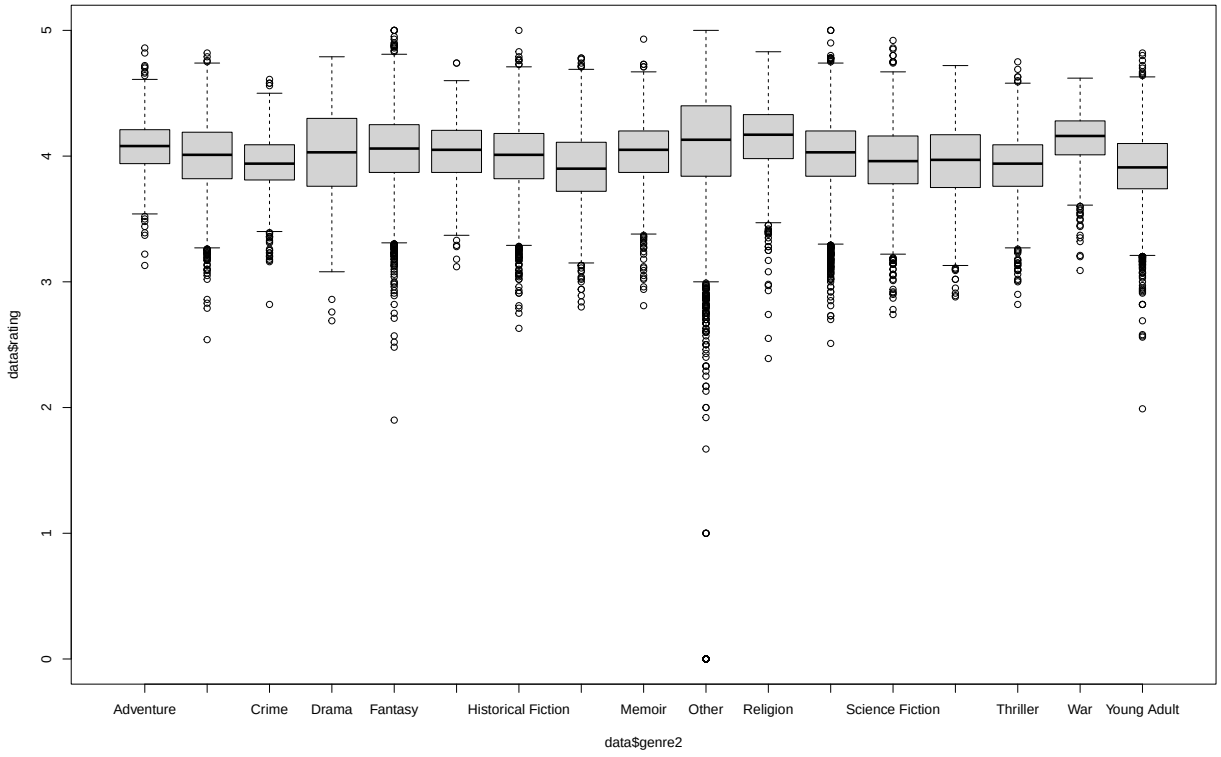
```
qqnorm(data$rating, pch = 1, frame = FALSE)
qqline(data$rating, col = "steelblue", lwd = 2)
```



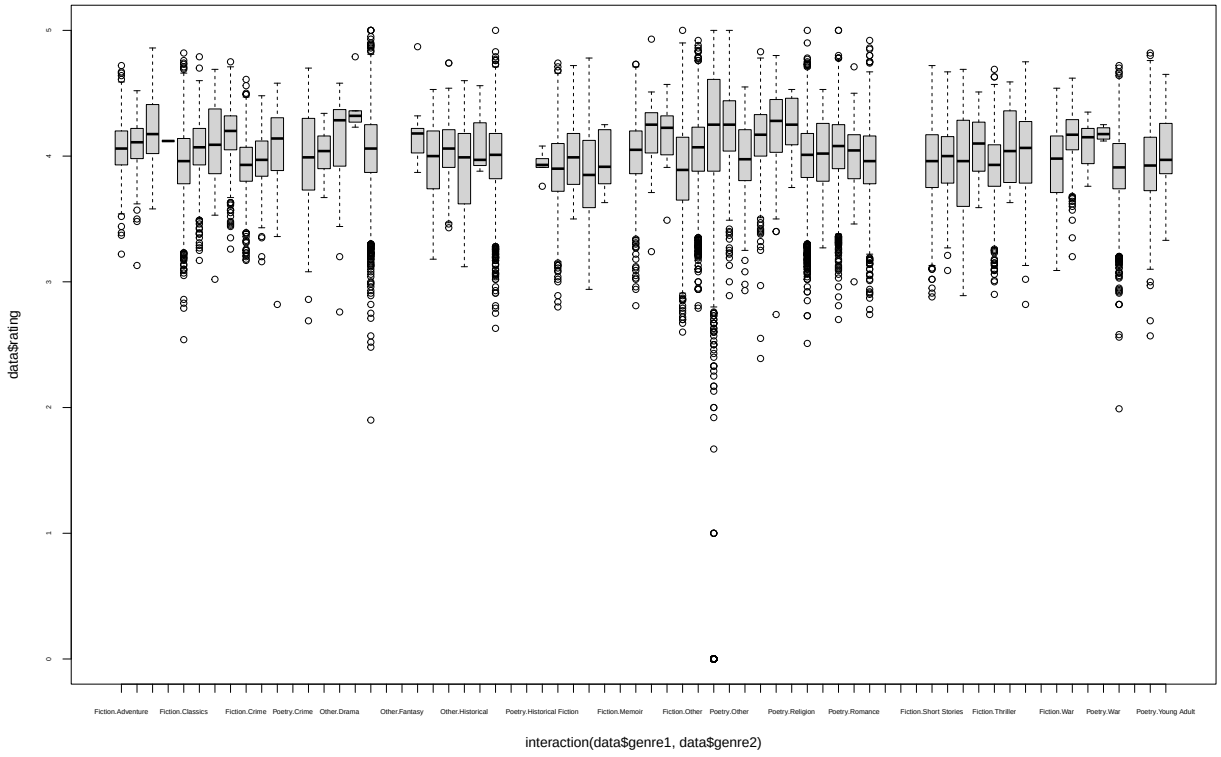

```
# Graficki prikaz podataka
boxplot(data$rating~data$genre1)
```




```
boxplot(data$rating ~ data$genre2)
```




```
boxplot(data$rating ~ interaction(data$genre1,data$genre2),cex.axis=0.5)
```




```
#Levene-ov test za jednakost varijanci između pojedinih grupa
require(car)
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.2.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.2.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
leveneTest(data$rating~interaction(data$genre1,data$genre2),data=data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group    56  110.22 < 2.2e-16 ***
##       52421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Iako je dosta nepregledno, jer imamo puno kategorija žanrova, grafički prikaz sugerira da postoji jasna razlika između žanrova.

Vidimo kako podaci nisu normalno distrubuirano zbog male p vrijednosti i također zbog istog razloga vidimo da ni varijance nisu homogene. Kako nisu zadovoljeni uvjeti za testiranje podataka ANOVA-om, koristit ćemo Kruskal- Wallis test.

```
#ANOVA
#a = aov(rating ~ genre1 * genre2, data = data)
#summary(a)
```

Kruskal-Wallisov test po rangovima je neparametarska metoda za testiranje potječu li uzorci iz istih distribucija Koristi se za usporedbu dva ili više neovisnih uzoraka jednake ili različite veličine uzorka. Proširuje Mann–Whitneyjev U test koji se koristi za usporedbu samo dvije skupine.

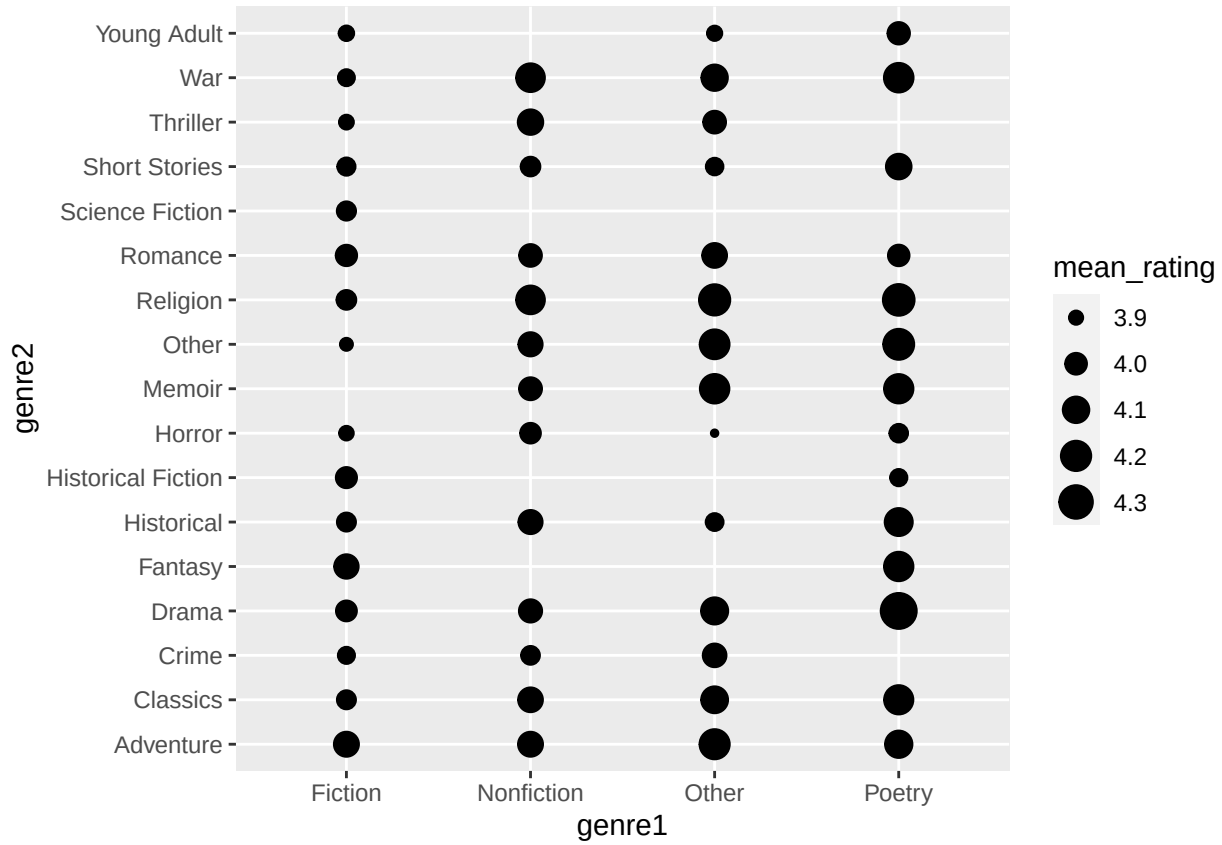
U ovom zadataku uspoređujemo kako genre1 i genre2 utječu na rating knjige.

```
#Kruskal- Wallis test
kruskal.test(rating~interaction(genre1,genre2),data=data)
```

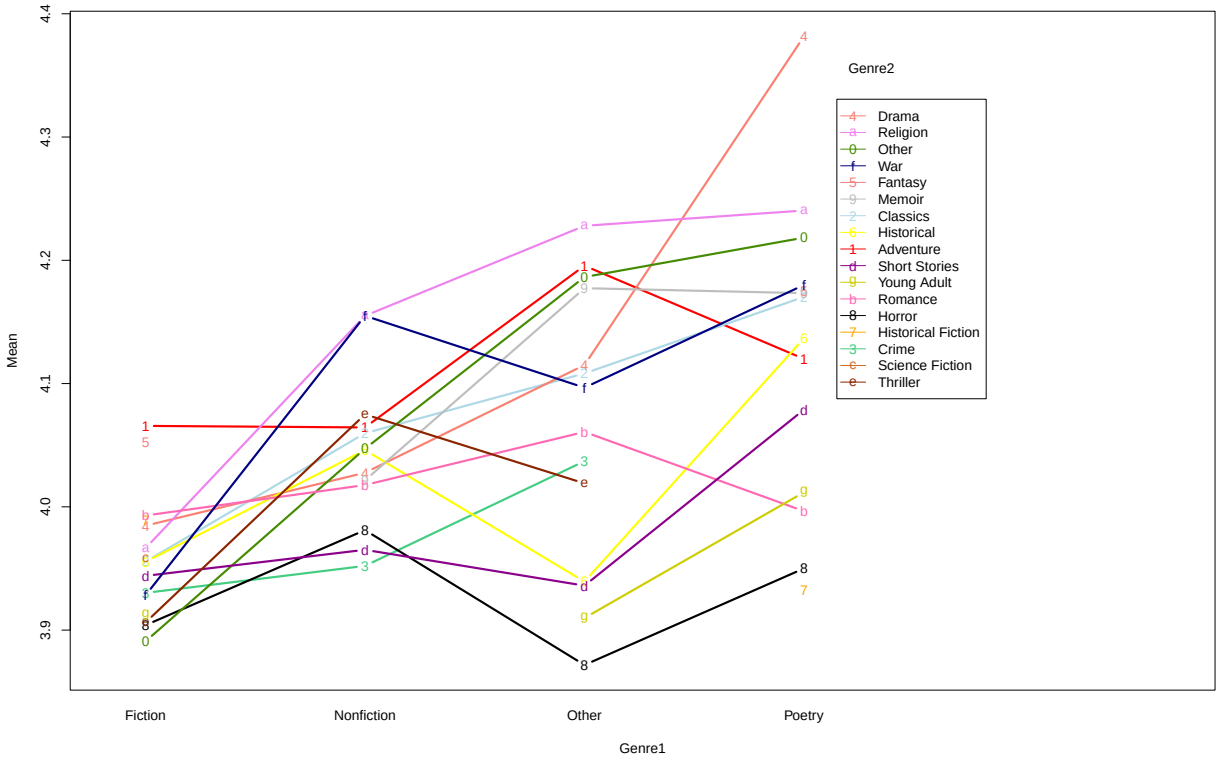
```
##
##  Kruskal-Wallis rank sum test
##
## data:  rating by interaction(genre1, genre2)
## Kruskal-Wallis chi-squared = 4165.6, df = 56, p-value < 2.2e-16
```

Analizom ispisanih podataka vidimo kako p-value iznosi <2.2e-16.Kako je p vrijednost izrazito mala možemo zaključiti da kategorije, genre1 i genre2 imaju značajan utjecaj na numeričku varijablu, odnosno možemo zaključiti da žanr knjige utječe na ocijenu.

```
ggplot(mean_ratings, aes(x = genre1, y = genre2, size = mean_rating)) +
  geom_point()
```




```
interaction.plot(x.factor = data$genre1,
                 trace.factor =data$genre2,
                 response = data$rating,
                 fun = mean,
                 type = "b",
                 ylab = "Mean",
                 xlab = "Genre1",
                 col = c("red", "lightblue","seagreen3","salmon","lightcoral","yellow","orange","black"
                 lty = 1,
                 lwd=2.5,
                 trace.label = "Genre2",
                 xpd=FALSE,
                 leg.bty = "o",
                  )
```



Zbog velike količine kategorija u genre2 stupcu introduction.plot se na nekim mjestima teško interpretira. Zbog toga imamo i ggplot iznad gdje za neke podatke nečitljive iz introduction.plota možemo viditi njihove vrijednosti.

2.Jesu li knjige s manje stranica jeftinije?

```
# Odvajanje 'pages' i 'price' u zaseban skup, pretvaranje vrijednosti u int i double.

data_drugi <- cbind(data$pages, data$price)
colnames(data_drugi) <- c("pages", "price")

data_drugi <- as.data.frame(data_drugi)

data_drugi$pages = as.integer(data_drugi$pages)
```

```
## Warning: NAs introduced by coercion
```

```
data_drugi$price = as.double(data_drugi$price)
```

```
## Warning: NAs introduced by coercion
```

```
data_drugi_clean <- na.omit(data_drugi)
p11 = cor(data_drugi_clean$price, data_drugi_clean$pages)
cat("\n\nJako nizak koeficijent koeficijent korelacije dviju zadanih znacajki:", p11)
```

```
##
##
## Jako nizak koeficijent koeficijent korelacije dviju zadanih znacajki: 0.1066827
```

```
# Prikaz podataka i njihovih osnovnih informacija (jako puno nedostajucih vrijednosti)

summary(data_drugi)
```

```
##      pages             price
##  Min.   :    0.0   Min.   :  0.840
##  1st Qu.:  210.0   1st Qu.:  3.240
##  Median :  304.0   Median :  5.200
##  Mean   :  328.7   Mean   :  9.657
##  3rd Qu.:  392.0   3rd Qu.:  8.860
##  Max.   :14777.0   Max.   :898.640
##  NA's   :2370      NA's   :14377
```

```
head(data_drugi)
```

```
##   pages price
## 1   416  5.55
## 2   151    NA
## 3   360 19.18
## 4     0    NA
## 5   190    NA
## 6   280  7.37
```

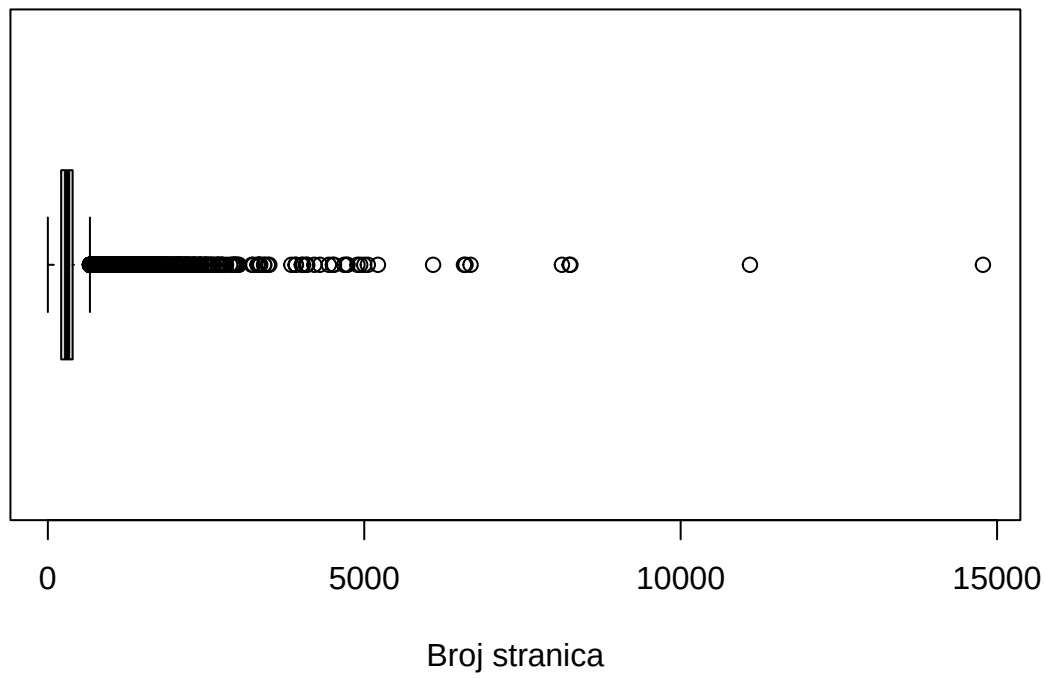
```
data_drugi_prazni <- data_drugi
```

Prikazi boxplot-a i histograma za značajku 'pages' i 'price'.

```
summary(data_drugi$pages)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   210.0   304.0   328.7   392.0 14777.0    2370
```

```
boxplot(
  data_drugi$pages,
  main = "Boxplot za znacajku 'pages'",
  xlab = "Broj stranica",
  horizontal = TRUE
)
```

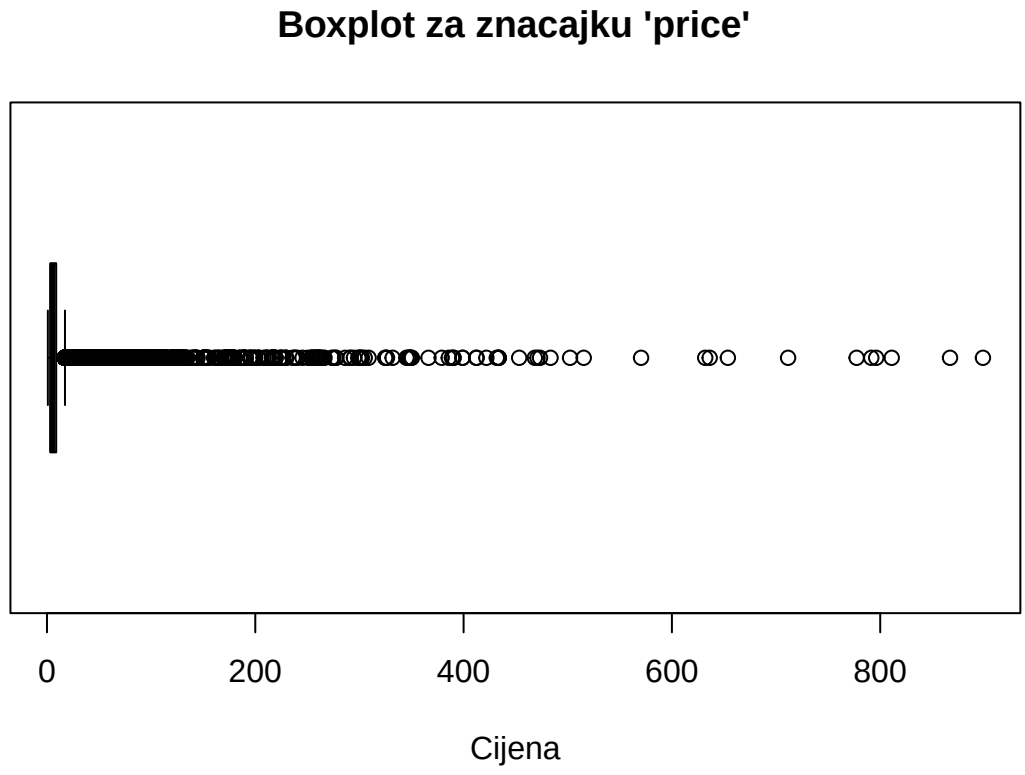



```
hist(
  data_drugi$pages,
  main = "Histogram za znacajku 'pages'",
  xlab = "Broj stranica",
  )
```


```
summary(data_drugi$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.840   3.240   5.200   9.657   8.860 898.640   14377
```

```
boxplot(
  data_drugi$price,
  main = "Boxplot za znacajku 'price'",
  xlab = "Cijena",
  horizontal = TRUE
)
```




```
hist(
  data_drugi$price,
  main = "Histogram za znacajku 'price'",
  xlab = "Cijena",
  )
```

Vidimo da grafovi ne izgledaju baš dobro, jer su jako ukošeni u jednu stranu. Nedostajuće vrijednosti možemo postaviti da su jednake medijanu. Medutim prvo ćemo provesti pokus izgleda grafova ako izbacimo samo 1% podataka sa gornje i donje strane.

```
missing_values <- which(is.na(data_drugi$pages))
extrem <- data_drugi$pages[-missing_values]
missing_values <- which(is.na(data_drugi$price))
extrem2 <- data_drugi$price[-missing_values]

extrem <- sort(extrem)
extrem2 <- sort(extrem2)
extrem <- extrem[(length(extrem) * 0.01) : (length(extrem)* 0.99)]
extrem2 <- extrem2[(length(extrem2) * 0.01) : (length(extrem2)* 0.99)]

length(data_drugi$pages)
```

```
## [1] 52478
```

```
length(extrem)
```

```
## [1] 49106
```

```
summary(extrem)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    22.0   213.0   304.0   316.6   389.0  1088.0
```

```
length(data_drugi$price)
```

```
## [1] 52478
```

```
length(extrem2)
```

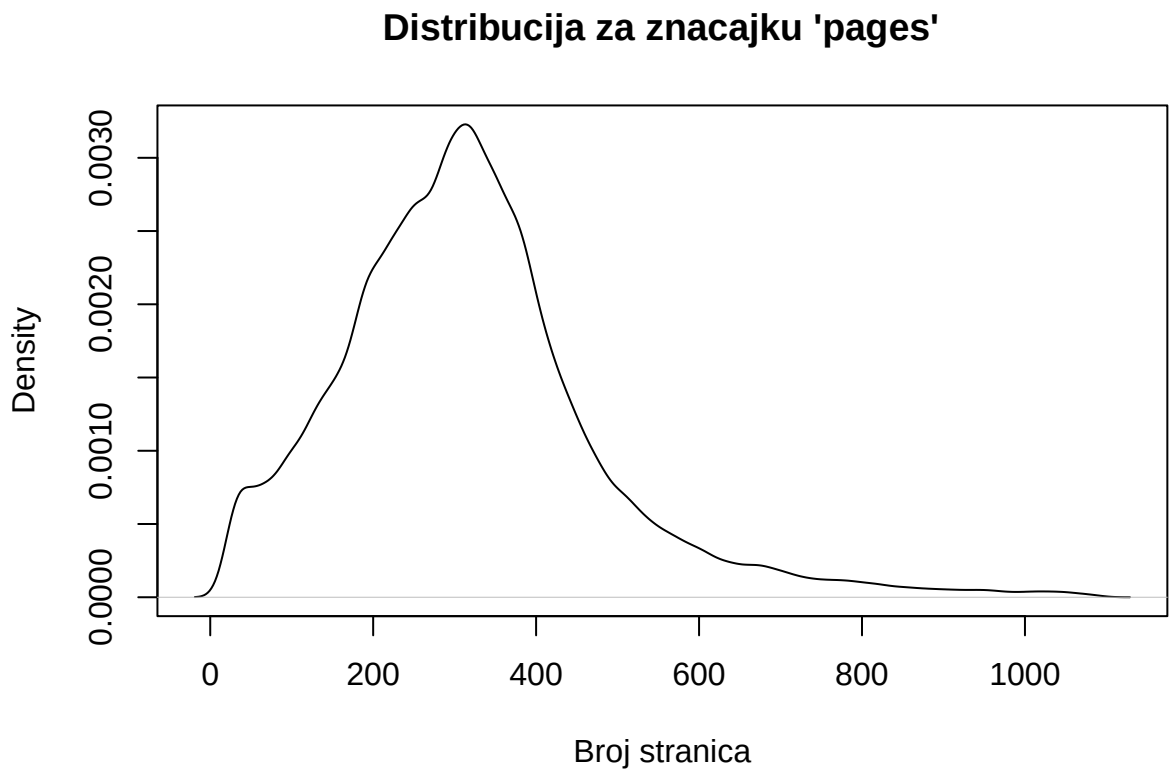
```
## [1] 37339
```

```
summary(extrem2)
```

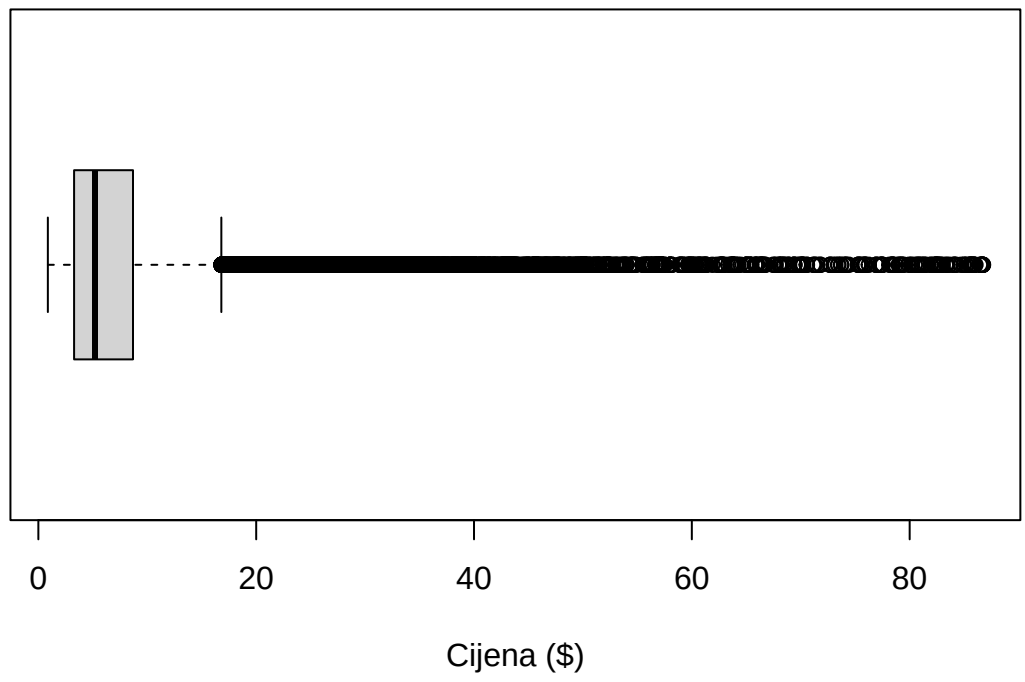
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.870   3.280   5.200   7.873   8.690  86.730
```

```
boxplot(extrem,
  main = "Boxplot za znacajku 'pages'",
  xlab = "Broj stranica",horizontal = TRUE)
```

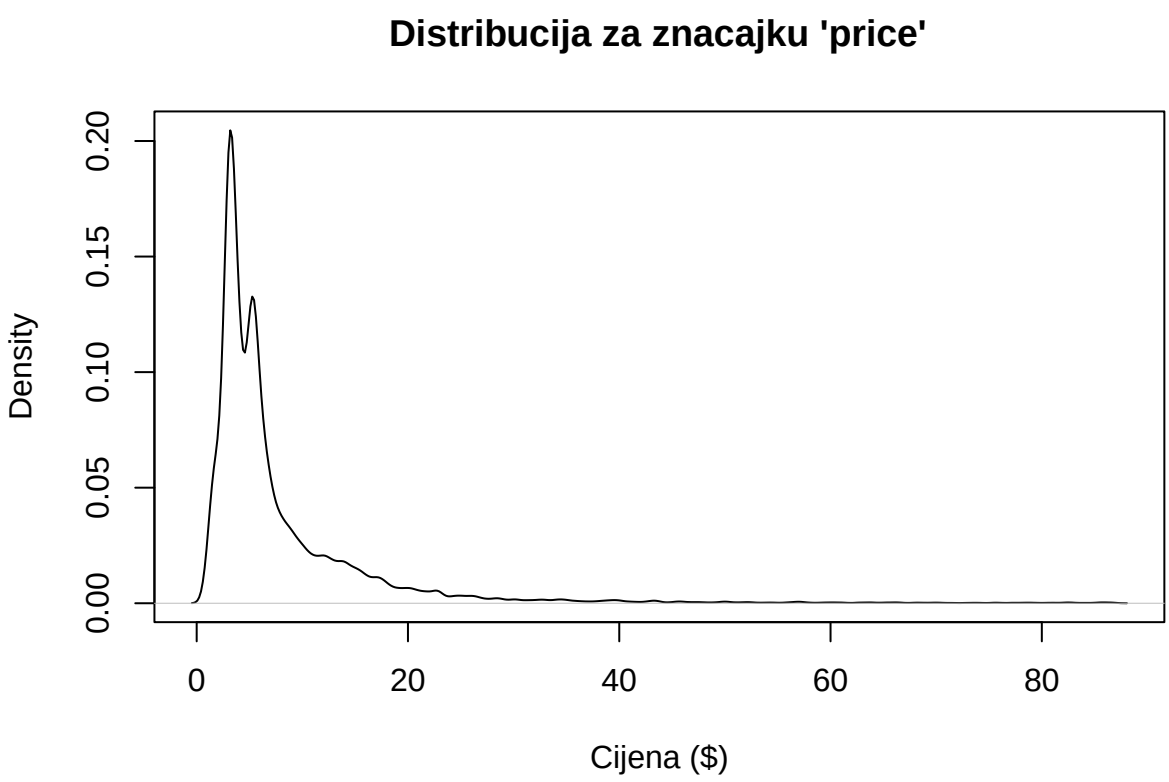

```
d <- density(extrem)
plot(d,xlab = "Broj stranica",
  main = "Distribucija za znacajku 'pages'")
```




```
boxplot(extrem2,
  main = "Boxplot za znacajku 'price'",
  xlab = "Cijena ($)",
  horizontal = TRUE)
```




```
d <- density(extrem2)
plot(d, xlab = "Cijena ($)",
  main = "Distribucija za znacajku 'price'")
```



Iako su sada grafovi puno pregledniji, medijan za 'price' i 'pages' ostao je gotovo isti (čak se ni srednja vrijednost nije puno promijenila). Odlučili smo se da ćemo nedostajuće vrijednosti postaviti na medijan.

```
data_drugi_prosireni <- data_drugi

x_median <- median(data_drugi$pages[!is.na(data_drugi$pages)])
data_drugi_prosireni$pages[is.na(data_drugi$pages)] <- x_median

x_median <- median(data_drugi$price[!is.na(data_drugi$price)])
data_drugi_prosireni$price[is.na(data_drugi$price)] <- x_median

summary(data_drugi_prosireni$pages)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   217.0   304.0   327.6   385.0 14777.0
```

```
summary(data_drugi_prosireni$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.840   3.720   5.200   8.436   6.570 898.640
```

Možemo započeti test. –> Jesu li knjige s manje stranica jeftinije? <–

Ovdje možemo iskoristiti Hi-kvadrat test koji ispituje postoji li zavisnost između broja stranica i cijene, time odgovoriti na zadano pitanje.

Za početak trebamo napraviti kontigencijsku tablicu. Značajke cemo grupirati u 3×3 razreda. Malo stranica, srednje, puno stranica. Mala cijena, srednja, velika cijena.

```
subset_data <- subset(data_drugi_prosireni, pages<=90 & price<=20)
nrow(subset_data)
```

```
## [1] 3026
```

```
data_drugi_prosireni$pages_cat <- cut(data_drugi_prosireni$pages, breaks = c(-Inf, 90, 199, Inf), labels
data_drugi_prosireni$price_cat <- cut(data_drugi_prosireni$price, breaks = c(-Inf, 20, 40, Inf), labels

table <- table(data_drugi_prosireni$price_cat, data_drugi_prosireni$pages_cat)

added_margins_table = addmargins(table)
added_margins_table
```

```
##
##              less_pages semi_pages many_pages   Sum
##   low_price        3026       7472      39212 49710
##   semi_price         95        307       1389  1791
##   high_price         71        172        734   977
##   Sum              3192       7951      41335 52478
```

```
#chisq.test(table)
```

Nakon sto grupiramo razrede, za svaki provjeravamo je li očekivana frekvencija >= 5. To je pretpostavka hi-kvadrat testa.

```
for (col_names in colnames(added_margins_table)){
  for (row_names in rownames(added_margins_table)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije :',col_names,'-',row_names,':',(added_margins_table[row_names,'Sum'] *
    }
  }
}
```

```
## Očekivane frekvencije : less_pages - low_price : 3023.635
## Očekivane frekvencije : less_pages - semi_price : 108.9385
## Očekivane frekvencije : less_pages - high_price : 59.4265
## Očekivane frekvencije : semi_pages - low_price : 7531.617
## Očekivane frekvencije : semi_pages - semi_price : 271.3564
## Očekivane frekvencije : semi_pages - high_price : 148.0264
## Očekivane frekvencije : many_pages - low_price : 39154.75
## Očekivane frekvencije : many_pages - semi_price : 1410.705
## Očekivane frekvencije : many_pages - high_price : 769.5471
```

Vidimo da frekvencije zadovoljavaju uvjet, možemo nastaviti s testom.

Testom ispitujemo postoji li veza između cijene i stranica knjige. H0 pretpostavka govori da su ove varijable nezavisne. Ukoliko p-vrijednost ispadne manja od 0.05 odbacujemo H0 pretpostavku.

```
chisq.test(added_margins_table, correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  added_margins_table
## X-squared = 15.135, df = 9, p-value = 0.08728
```

p-vrijednost rezultira s 0.087 sto znači da su zadane varijable nezavisne jedna od druge. Time smo odgovorili na početno pitanje "Jesu li knjige s manje stranica jeftinije?" –> varijable su nezavisne, ako knjiga ima manje stranica ne mora značiti da ce biti jeftinija.

3. Možete li odrediti popularnost knjige (po vašoj definiciji, npr. broj glasača, prosječna ocjena...) na temelju dostupnih varijabli?

U svrhu pronalaženja rješenja na prethodno pitanje izabrana je logistička regresija.

Logistička regresija

```
library(zoo)
```

```
## Warning: package 'zoo' was built under R version 4.2.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
data_2 <- data
data_2 <- select(data_2, -genres)
# zamjena praznih vrijednosti interpolacijom
data_2 <- data_2 %>%
        mutate(likedPercent = round(na.approx(likedPercent)))
data_2["price"][data_2["price"] == ''] <- NA
data_2 <- data_2 %>%
        mutate(price = na.approx(price))
```

```
## Warning in xy.coords(x, y, setLab = FALSE): NAs introduced by coercion
```

```
data_2["pages"][data_2["pages"] == ''] <- NA
data_2 <- data_2 %>%
        mutate(pages = na.approx(pages))
```

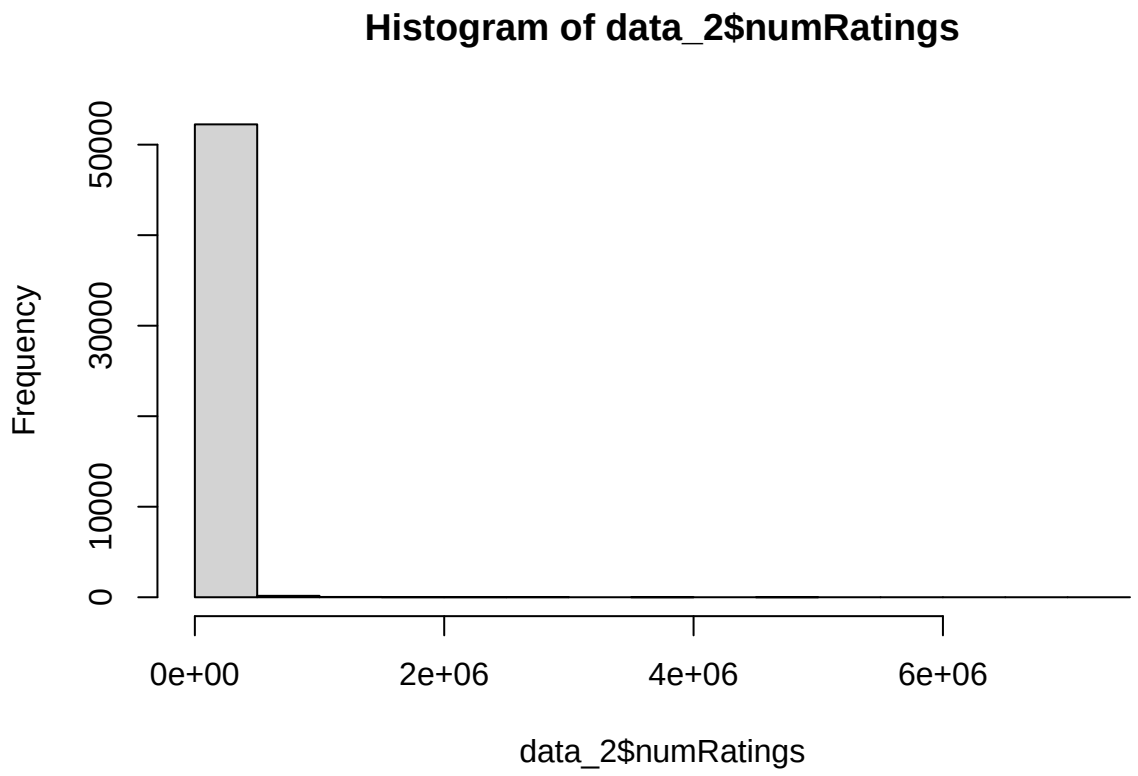
```
## Warning in xy.coords(x, y, setLab = FALSE): NAs introduced by coercion
```

```r
# Vrsta podataka
cat("\n\n")
```

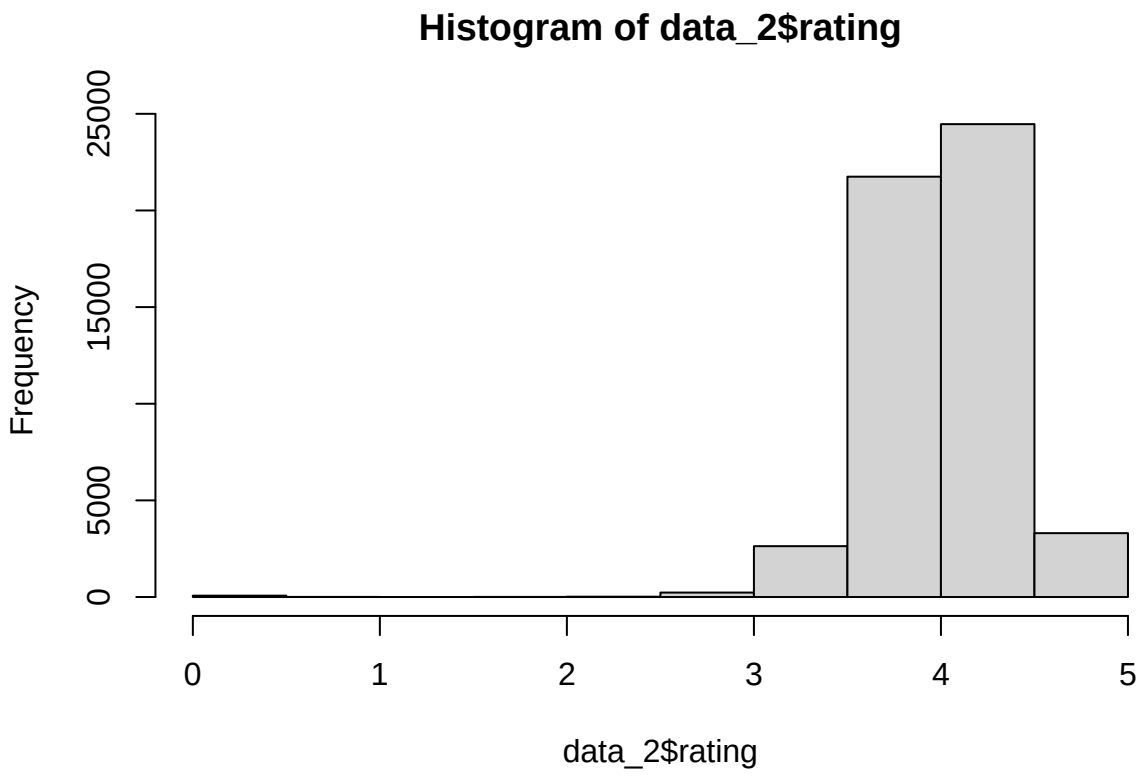
```r
str(data_2)
```

```
## 'data.frame':    52478 obs. of  16 variables:
##  $ X             : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ title         : chr  "Attracted to Fire" "Elemental" "Unbelievable" "Fractured" ...
##  $ series        : chr  "" "Soul Guardians #2" "Port Fare #2" "Fateful #2" ...
##  $ author        : chr  "DiAnn Mills (Goodreads Author)" "Kim Richardson (Goodreads Author)" "Sherry
##  $ rating        : num  4.14 4.07 4.16 4 4.19 3.7 3.85 4.02 4.09 4.67 ...
##  $ language      : chr  "English" "English" "English" "English" ...
##  $ bookFormat    : chr  "Paperback" "Kindle Edition" "Paperback" "Nook" ...
##  $ pages         : num  416 151 360 0 190 280 507 201 518 350 ...
##  $ publisher     : chr  "Tyndale House Publishers" "Kim Richardson" "Wordpaintings Unlimited" "Cheri
##  $ awards        : chr  "['HOLT Medallion by Virginia Romance Writers Nominee for Long Inspirational
##  $ numRatings    : int  2143 1947 1028 871 37 6674 238 246 6196 3 ...
##  $ ratingsByStars: chr  "['945', '716', '365', '78', '39']" "['801', '636', '391', '84', '35']" "['4
##  $ likedPercent  : num  95 94 94 94 95 84 90 90 94 90 ...
##  $ price         : num  5.55 12.36 19.18 15.24 11.31 ...
##  $ genre1        : chr  "Fiction" "Fiction" "Other" "Fiction" ...
##  $ genre2        : chr  "Romance" "Fantasy" "Romance" "Young Adult" ...
```

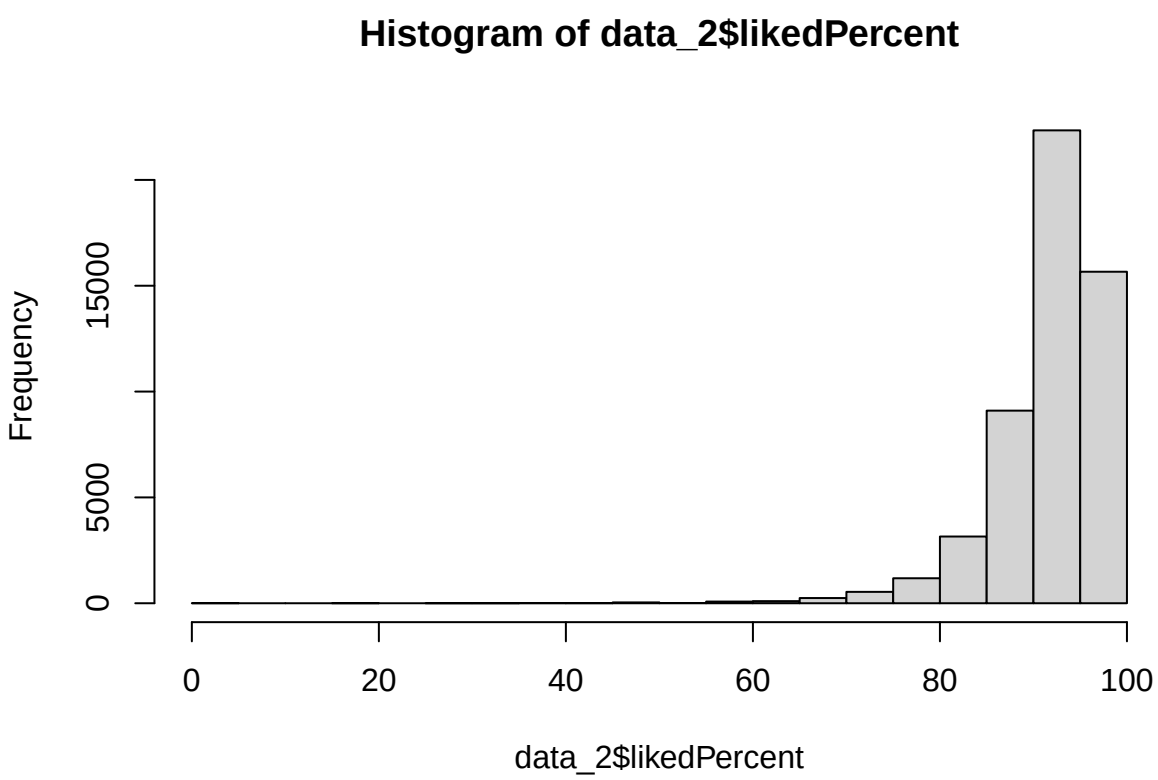
```r
# Histogrami potencijalnih atributa popularnosti
hist(data_2$numRatings)
```



```
hist(data_2$rating)
```




```
hist(data_2$likedPercent)
```




```
max(data_2$numRatings)
```

```
## [1] 7048471
```

```
summary(data_2$numRatings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     341    2307   17879    9380 7048471
```

```
# Popularne su one knjige koje imaju vise od 9380 ocjena (top 25%)
data_2$popularity <- ifelse(data_2$numRatings > 9380, 1, 0)
str(data_2)
```

```
## 'data.frame':    52478 obs. of  17 variables:
##  $ X             : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ title         : chr  "Attracted to Fire" "Elemental" "Unbelievable" "Fractured" ...
##  $ series        : chr  "" "Soul Guardians #2" "Port Fare #2" "Fateful #2" ...
##  $ author        : chr  "DiAnn Mills (Goodreads Author)" "Kim Richardson (Goodreads Author)" "Sherry
##  $ rating        : num  4.14 4.07 4.16 4 4.19 3.7 3.85 4.02 4.09 4.67 ...
##  $ language      : chr  "English" "English" "English" "English" ...
##  $ bookFormat    : chr  "Paperback" "Kindle Edition" "Paperback" "Nook" ...
##  $ pages         : num  416 151 360 0 190 280 507 201 518 350 ...
##  $ publisher     : chr  "Tyndale House Publishers" "Kim Richardson" "Wordpaintings Unlimited" "Cheri
##  $ awards        : chr  "['HOLT Medallion by Virginia Romance Writers Nominee for Long Inspirational
```

```
##  $ numRatings    : int  2143 1947 1028 871 37 6674 238 246 6196 3 ...
##  $ ratingsByStars: chr  "['945', '716', '365', '78', '39']" "['801', '636', '391', '84', '35']" "['4
##  $ likedPercent  : num  95 94 94 94 95 84 90 90 94 90 ...
##  $ price         : num  5.55 12.36 19.18 15.24 11.31 ...
##  $ genre1        : chr  "Fiction" "Fiction" "Other" "Fiction" ...
##  $ genre2        : chr  "Romance" "Fantasy" "Romance" "Young Adult" ...
##  $ popularity    : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
# Priprema podataka
data_2$price <- as.numeric(data_2$price)
data_2$pages <- as.numeric(data_2$pages)
data_2$author_enc <- as.numeric(factor(data_2$author))
data_2$title_enc <- as.numeric(factor(data_2$title))
data_2$series_enc <- as.numeric(factor(data_2$series))
data_2$language_enc <- as.numeric(factor(data_2$language))
data_2$bookFormat_enc <- as.numeric(factor(data_2$bookFormat))
data_2$publisher_enc <- as.numeric(factor(data_2$publisher))
data_2$genre1_enc <- as.numeric(factor(data_2$genre1))
data_2$genre2_enc <- as.numeric(factor(data_2$genre2))
data_2$awards_enc <- as.numeric(factor(data_2$awards))
data_2$ratingByStars_enc <- as.numeric(factor(data_2$ratingsByStars))
data_2$popularity <- factor(data_2$popularity, levels = c(0,1), labels = c(FALSE,TRUE))
summary(data_2)
```

```
##        X            title              series             author
##  Min.   :    0   Length:52478       Length:52478       Length:52478
##  1st Qu.:13119   Class :character   Class :character   Class :character
##  Median :26239   Mode  :character   Mode  :character   Mode  :character
##  Mean   :26239
##  3rd Qu.:39358
##  Max.   :52477
##      rating        language          bookFormat            pages
##  Min.   :0.000   Length:52478       Length:52478       Min.   :    0
##  1st Qu.:3.820   Class :character   Class :character   1st Qu.:  212
##  Median :4.030   Mode  :character   Mode  :character   Median :  304
##  Mean   :4.022                                         Mean   :  328
##  3rd Qu.:4.230                                         3rd Qu.:  390
##  Max.   :5.000                                         Max.   :14777
##   publisher            awards            numRatings       ratingsByStars
##  Length:52478       Length:52478       Min.   :      0   Length:52478
##  Class :character   Class :character   1st Qu.:    341   Class :character
##  Mode  :character   Mode  :character   Median :   2307   Mode  :character
##                                        Mean   :  17879
##                                        3rd Qu.:   9380
##                                        Max.   :7048471
##   likedPercent        price             genre1             genre2
##  Min.   :  0.00   Min.   :  0.840   Length:52478       Length:52478
##  1st Qu.: 90.00   1st Qu.:  3.423   Class :character   Class :character
##  Median : 94.00   Median :  5.350   Mode  :character   Mode  :character
##  Mean   : 92.23   Mean   : 10.004
##  3rd Qu.: 96.00   3rd Qu.:  9.570
##  Max.   :100.00   Max.   :898.640
##  popularity      author_enc      title_enc       series_enc     language_enc
##  FALSE:39358   Min.   :    1   Min.   :    1   Min.   :    1   Min.   : 1.0
```

```
##  TRUE :13120   1st Qu.: 7038   1st Qu.:12423   1st Qu.:    1   1st Qu.:25.0
##                Median :14190   Median :24957   Median :    1   Median :25.0
##                Mean   :14056   Mean   :24909   Mean   : 5102   Mean   :25.2
##                3rd Qu.:20986   3rd Qu.:37339   3rd Qu.:10073   3rd Qu.:25.0
##                Max.   :28227   Max.   :49925   Max.   :22803   Max.   :82.0
##  bookFormat_enc  publisher_enc     genre1_enc      genre2_enc
##  Min.   :  1.0   Min.   :    1   Min.   :1.000   Min.   : 1.000
##  1st Qu.: 56.0   1st Qu.: 2124   1st Qu.:1.000   1st Qu.: 5.000
##  Median : 94.0   Median : 4655   Median :1.000   Median :10.000
##  Mean   : 75.2   Mean   : 5020   Mean   :1.568   Mean   : 9.376
##  3rd Qu.: 94.0   3rd Qu.: 8046   3rd Qu.:2.000   3rd Qu.:12.000
##  Max.   :137.0   Max.   :11111   Max.   :4.000   Max.   :17.000
##    awards_enc   ratingByStars_enc
##  Min.   :   1   Min.   :    1
##  1st Qu.:9215   1st Qu.:12929
##  Median :9215   Median :25799
##  Mean   :8266   Mean   :25707
##  3rd Qu.:9215   3rd Qu.:38579
##  Max.   :9215   Max.   :49908
```

Model logističke regresije prima sve atribute skupa podataka s time da su kategoričke varijable enkodirane kako bi model mogao konvergirati. Pomoću Rsq koristi se kako bi se vidjelo koliko je procijenjeni model blizu, odnosno daleko od null modela, dakle Rsq prikazuje koliko je naučeni model dobar.

```
# sample_data <- data_2 %>% sample_frac(0.8)

# Model logisticke regresije
logreg.mdl = glm(popularity ~ rating + author_enc + title_enc + series_enc + language_enc + bookFormat_
summary(logreg.mdl)
```

```
##
## Call:
## glm(formula = popularity ~ rating + author_enc + title_enc +
##     series_enc + language_enc + bookFormat_enc + genre1_enc +
##     genre2_enc + awards_enc + ratingByStars_enc + pages + price +
##     likedPercent, family = binomial(), data = data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0099  -0.7706  -0.5312   0.3961   5.3532
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.014e+00  2.287e-01  -8.804  < 2e-16 ***
## rating            -1.692e-01  5.055e-02  -3.348 0.000815 ***
## author_enc        -2.473e-06  1.375e-06  -1.799 0.072056 .
## title_enc         -1.571e-07  7.719e-07  -0.204 0.838686
## series_enc         1.756e-05  1.525e-06  11.518  < 2e-16 ***
## language_enc      -4.989e-03  1.062e-03  -4.699 2.61e-06 ***
## bookFormat_enc    -2.297e-03  4.958e-04  -4.633 3.60e-06 ***
## genre1_enc        -7.052e-01  1.840e-02 -38.336  < 2e-16 ***
## genre2_enc        -5.373e-03  2.442e-03  -2.200 0.027782 *
## awards_enc        -2.297e-04  4.405e-06 -52.148  < 2e-16 ***
```

```
## ratingByStars_enc  1.595e-05  7.593e-07  21.012  < 2e-16 ***
## pages              3.882e-04  4.509e-05   8.611  < 2e-16 ***
## price             -2.196e-02  1.361e-03 -16.132  < 2e-16 ***
## likedPercent       4.709e-02  3.384e-03  13.913  < 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: 59022  on 52477  degrees of freedom
## Residual deviance: 51372  on 52464  degrees of freedom
## AIC: 51400
##
## Number of Fisher Scoring iterations: 6
```

```
# Pseudo-R2
Rsq = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
Rsq
```

```
## [1] 0.1296122
```

Matrica zabune jedan je od pokazatelja kvalitete modela te je baza za daljnji izračun metrika performansi modela. Ona je zapravo kontingencijska matriac oznaka iz podataka i modela.

```
# Matrica zabune
yHat <- logreg.mdl$fitted.values >= 0.5
tab <- table(data_2$popularity, yHat)
tab
```

```
##        yHat
##         FALSE  TRUE
##   FALSE 37497  1861
##   TRUE  10272  2848
```

```
# Metrike performansi - tocnost, preciznost, odziv, specificnost
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] 0.7687984
```

```
precision
```

```
## [1] 0.6047993
```

```
recall
```

```
## [1] 0.2170732
```

```
specificity
```

```
## [1] 0.7849651
```

```
# Novi atribut
data_2$RL <- data_2$rating * data_2$likedPercent
```

```
# Model 2
logreg.mdl.2 = glm(popularity ~ rating + author_enc + title_enc + series_enc + language_enc + bookForma
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logreg.mdl.2)
```

```
##
## Call:
## glm(formula = popularity ~ rating + author_enc + title_enc +
##     series_enc + language_enc + bookFormat_enc + genre1_enc +
##     genre2_enc + awards_enc + pages + price + likedPercent +
##     RL, family = binomial(), data = data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0718  -0.7896  -0.5425   0.4721   5.2148
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.530e+01  2.437e+00 -14.485  < 2e-16 ***
## rating          8.983e+00  6.525e-01  13.768  < 2e-16 ***
## author_enc     -2.483e-06  1.371e-06  -1.812   0.0700 .
## title_enc      -3.512e-07  7.695e-07  -0.456   0.6481
## series_enc      1.706e-05  1.517e-06  11.251  < 2e-16 ***
## language_enc   -5.343e-03  1.066e-03  -5.011 5.41e-07 ***
## bookFormat_enc -2.471e-03  4.954e-04  -4.988 6.10e-07 ***
## genre1_enc     -6.696e-01  1.838e-02 -36.442  < 2e-16 ***
## genre2_enc     -5.719e-03  2.429e-03  -2.355   0.0185 *
## awards_enc     -2.245e-04  4.378e-06 -51.284  < 2e-16 ***
## pages           3.885e-04  4.510e-05   8.614  < 2e-16 ***
## price          -2.089e-02  1.346e-03 -15.521  < 2e-16 ***
## likedPercent    4.009e-01  2.585e-02  15.506  < 2e-16 ***
## RL             -9.622e-02  6.853e-03 -14.040  < 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: 59022  on 52477  degrees of freedom
## Residual deviance: 51596  on 52464  degrees of freedom
## AIC: 51624
##
## Number of Fisher Scoring iterations: 6
```

```r
# Pseudo-R2
Rsq.2 = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
Rsq.2
```

```
## [1] 0.1296122
```

Uz pomoć testa omjera izglednosti u nastavku se uspoređuju rezultati dvaju modela - originalnog te modela s dodatnim atributom nastalim kao kombinacija dva postojeća atributa.

```r
# Test omjera izglednosti
anova(logreg.mdl, logreg.mdl.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: popularity ~ rating + author_enc + title_enc + series_enc + language_enc +
##     bookFormat_enc + genre1_enc + genre2_enc + awards_enc + ratingByStars_enc +
##     pages + price + likedPercent
## Model 2: popularity ~ rating + author_enc + title_enc + series_enc + language_enc +
##     bookFormat_enc + genre1_enc + genre2_enc + awards_enc + pages +
##     price + likedPercent + RL
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     52464      51372
## 2     52464      51596  0  -224.34
```

Iz modela 3 izbačena je enkodirana varijabla naslova knjige s obzirom da je ona nesignifikantni regresor.

```r
# Model 3
logreg.mdl.3 = glm(popularity ~ rating + author_enc + series_enc + language_enc + bookFormat_enc + genr
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(logreg.mdl.3)
```

```
##
## Call:
## glm(formula = popularity ~ rating + author_enc + series_enc +
##     language_enc + bookFormat_enc + genre1_enc + genre2_enc +
##     awards_enc + pages + price + likedPercent + RL, family = binomial(),
##     data = data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0717  -0.7896  -0.5427   0.4715   5.2161
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.531e+01  2.437e+00 -14.489  < 2e-16 ***
## rating          8.983e+00  6.525e-01  13.767  < 2e-16 ***
## author_enc     -2.512e-06  1.369e-06  -1.835   0.0665 .
## series_enc      1.699e-05  1.509e-06  11.264  < 2e-16 ***
## language_enc   -5.311e-03  1.064e-03  -4.993 5.95e-07 ***
```

```
## bookFormat_enc -2.470e-03  4.954e-04  -4.987 6.13e-07 ***
## genre1_enc     -6.698e-01  1.837e-02 -36.462  < 2e-16 ***
## genre2_enc     -5.674e-03  2.427e-03  -2.338   0.0194 *
## awards_enc     -2.245e-04  4.378e-06 -51.284  < 2e-16 ***
## pages           3.881e-04  4.509e-05   8.607  < 2e-16 ***
## price          -2.089e-02  1.346e-03 -15.522  < 2e-16 ***
## likedPercent    4.009e-01  2.586e-02  15.505  < 2e-16 ***
## RL             -9.621e-02  6.854e-03 -14.039  < 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: 59022  on 52477  degrees of freedom
## Residual deviance: 51596  on 52465  degrees of freedom
## AIC: 51622
##
## Number of Fisher Scoring iterations: 6
```

```
# Pseudo-R2
Rsq.3 = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
Rsq.3
```

```
## [1] 0.1296122
```

```
# Test omjera izglednosti
anova(logreg.mdl, logreg.mdl.3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: popularity ~ rating + author_enc + title_enc + series_enc + language_enc +
##     bookFormat_enc + genre1_enc + genre2_enc + awards_enc + ratingByStars_enc +
##     pages + price + likedPercent
## Model 2: popularity ~ rating + author_enc + series_enc + language_enc +
##     bookFormat_enc + genre1_enc + genre2_enc + awards_enc + pages +
##     price + likedPercent + RL
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     52464      51372
## 2     52465      51596 -1  -224.55 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Matrica zabune
yHat <- logreg.mdl.3$fitted.values >= 0.5
tab <- table(data_2$popularity, yHat)
tab
```

```
##        yHat
##         FALSE  TRUE
##   FALSE 37478  1880
##   TRUE  10263  2857
```

```r
# Metrike performansi - tocnost, preciznost, odziv, specificnost
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] 0.7686078
```

```r
precision
```

```
## [1] 0.6031243
```

```r
recall
```

```
## [1] 0.2177591
```

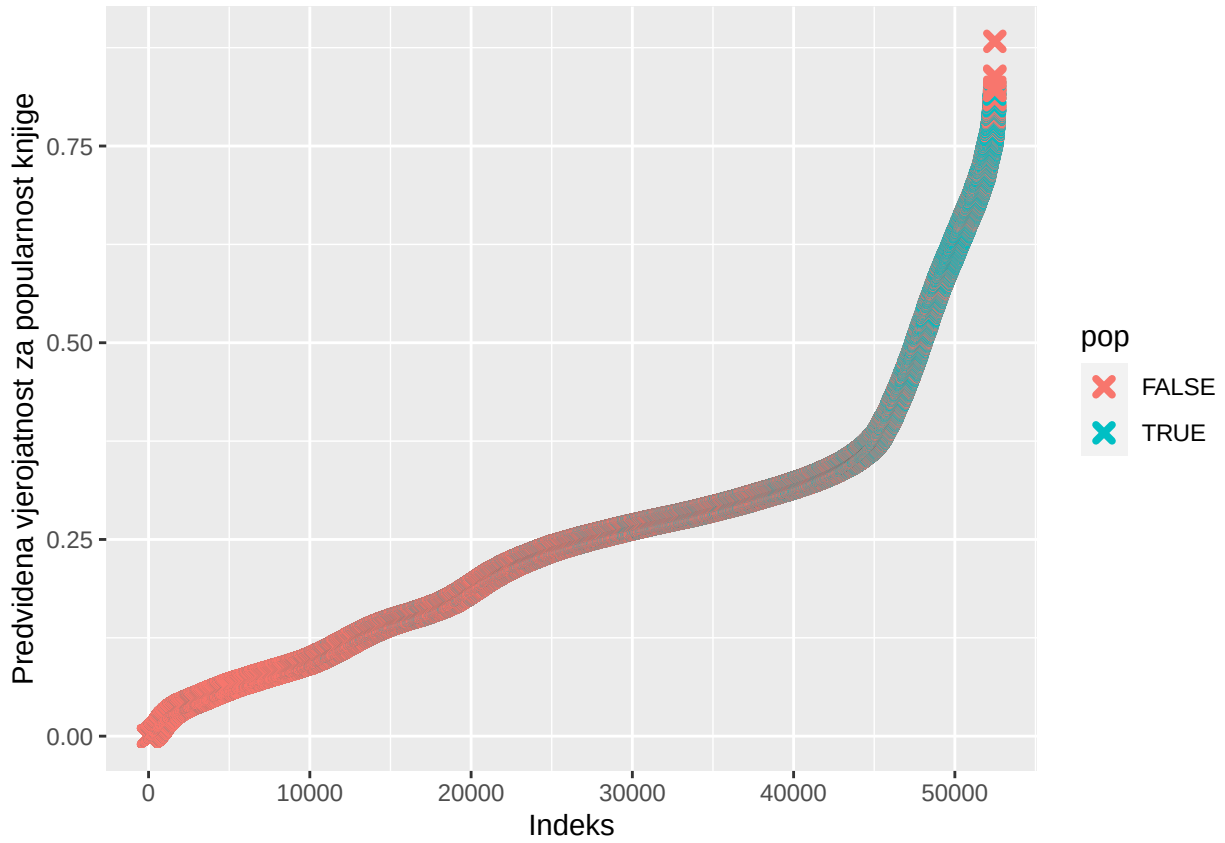
```r
specificity
```

```
## [1] 0.7850275
```

```r
# Originalni model
# [1] 0.7682648
# [1] 0.6023916
# [1] 0.2150152
# [1] 0.7845172

# Model 3 bez RL
# [1] 0.7683029
# [1] 0.602649
# [1] 0.2150152
# [1] 0.7845262
```

Najboljim se pokazao treći model iako su generalno razlike između modela minimalne.

```r
# Graf predvidanja
predicted.data <- data.frame(
  probability.of.popularity=logreg.mdl.3$fitted.values,
  pop=data_2$popularity)

predicted.data <- predicted.data[
  order(predicted.data$probability.of.popularity, decreasing=FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)

ggplot(data=predicted.data, aes(x=rank, y=probability.of.popularity)) +
  geom_point(aes(color=pop), alpha=1, shape=4, stroke=2) +
  xlab("Indeks") +
  ylab("Predvidena vjerojatnost za popularnost knjige")
```



Dobiveni graf prikazuje uspješnost predviđanja modela na način da bi sve popularne knjige trebale biti iznad granice od 0.5 jer 0 predstavlja nepopularnu knjigu, a 1 popularnu. Vidljivo je kako je većina vrijednosti stvarno iznad te granice te da su dobro klasificirane. Iako postoje greške i model ne ostvaruje odlične rezultate (koji bi bili mogući uz dodatna poboljšanja), može se donijeti zaključak kako je moguće odrediti popularnost knjige na temelju varijabli unutar zadanog skupa podataka. Točnost konačnog modela iznosi 76.86%, preciznost 60.31%, odziv 21.77%, a specifičnost 78.5%.

4.Možete li na temelju dostupnih varijabli odrediti je li knjiga bila nagrađivana?

Pretvorba podataka za učenje modela:

```
y <- data$awards
y <- ifelse(y == "[]", 0, 1)

X <- data %>% select(-language,-publisher, -bookFormat,-author, -series, -X, -title, -genres, -ratingsB

X$genre1 <- factor(X$genre1)
X$genre2 <- factor(X$genre2)


X$pages <- as.numeric(as.character(X$pages))
```

```
## Warning: NAs introduced by coercion
```

```
X$pages[!is.numeric(X$pages)] <- NA
average_pages <- mean(X$pages, na.rm = TRUE)
X$pages[is.na(X$pages)] <- average_pages

X$price <- as.numeric(as.character(X$price))
```

```
## Warning: NAs introduced by coercion
```

```
X$price[!is.numeric(X$price)] <- NA
average_price <- mean(X$price, na.rm = TRUE)
X$price[is.na(X$price)] <- average_price

average_likedPercent <- mean(X$likedPercent, na.rm = TRUE)
X$likedPercent[is.na(X$likedPercent)] <- average_likedPercent
```

Učenje modela na svim podacima:

```
logreg.mdl.3 = glm(y ~ ., data = X, family = binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logreg.mdl.3)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(), data = X)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.3840  -0.7273  -0.5508  -0.2092   2.8771
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.528e+00  2.446e-01  -6.247 4.17e-10 ***
## rating                   -5.535e-01  5.337e-02 -10.372  < 2e-16 ***
## pages                     2.183e-04  4.540e-05   4.809 1.51e-06 ***
## numRatings                9.782e-06  3.097e-07  31.585  < 2e-16 ***
## likedPercent              2.689e-02  3.388e-03   7.937 2.08e-15 ***
## price                    -4.306e-03  9.207e-04  -4.677 2.90e-06 ***
## genre1Nonfiction         -4.990e-01  5.544e-02  -9.000  < 2e-16 ***
## genre1Other              -2.025e+00  7.310e-02 -27.704  < 2e-16 ***
## genre1Poetry             -3.765e-02  8.852e-02  -0.425 0.670608
## genre2Classics           -7.284e-01  1.240e-01  -5.876 4.19e-09 ***
## genre2Crime               2.858e-01  1.311e-01   2.180 0.029292 *
## genre2Drama              -1.750e-01  1.642e-01  -1.066 0.286378
## genre2Fantasy            -9.574e-02  1.176e-01  -0.814 0.415710
## genre2Historical          7.005e-01  1.515e-01   4.623 3.79e-06 ***
## genre2Historical Fiction  4.581e-01  1.193e-01   3.839 0.000124 ***
## genre2Horror              5.932e-03  1.332e-01   0.045 0.964473
## genre2Memoir              5.448e-01  1.329e-01   4.099 4.15e-05 ***
## genre2Other              -2.631e-01  1.210e-01  -2.174 0.029713 *
```

```
## genre2Religion           -6.485e-01  1.488e-01  -4.359 1.31e-05 ***
## genre2Romance            -5.132e-01  1.197e-01  -4.286 1.82e-05 ***
## genre2Science Fiction     2.083e-01  1.243e-01   1.676 0.093775 .
## genre2Short Stories      -2.416e-02  1.381e-01  -0.175 0.861089
## genre2Thriller           -2.180e-01  1.291e-01  -1.689 0.091282 .
## genre2War                 9.725e-01  1.521e-01   6.394 1.62e-10 ***
## genre2Young Adult         2.169e-01  1.192e-01   1.820 0.068725 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 52847  on 52477  degrees of freedom
## Residual deviance: 46581  on 52453  degrees of freedom
## AIC: 46631
##
## Number of Fisher Scoring iterations: 6
```

```
yHat <- logreg.mdl.3$fitted.values > 0.5
tab <- table(y, yHat)

tab
```

```
##    yHat
## y   FALSE  TRUE
##   0 41548   316
##   1  9753   861
```

```
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])

print('accuracy:')
```

```
## [1] "accuracy:"
```

```
accuracy
```

```
## [1] 0.8081291
```

```
print('precision')
```

```
## [1] "precision"
```

```
precision
```

```
## [1] 0.7315208
```

```
print('recall')
```

```
## [1] "recall"
```

```
recall
```

```
## [1] 0.08111928
```

```
print('specificity')
```

```
## [1] "specificity"
```

```
specificity
```

```
## [1] 0.8098867
```

Učenjem modela logističkom regresijom dobivamo dobar accuracy, ali loš recall što ukazuje na ne balans između klasa.

Upsampaling data:

```
data_prediction <- cbind(X, y)
data <- cbind(X, y)

majority <- max(table(data$y))
balanced_data <- data %>% group_by(y) %>% sample_n(majority, replace = TRUE)
```

Učenje modela na ujednačenom skupu podataka:

```
y <- balanced_data$y
X <- balanced_data %>% select(-y)
```

```
## Adding missing grouping variables: `y`
```

```
logreg.mdl.3 = glm(y ~ ., data = X, family = binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logreg.mdl.3)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(), data = X)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -1.0846  -0.0543   1.0366   2.3406
##
## Coefficients:
```

```
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                7.495e-02  1.594e-01   0.470 0.638215
## rating                    -6.301e-01  3.945e-02 -15.972  < 2e-16 ***
## pages                      2.361e-04  3.628e-05   6.506 7.71e-11 ***
## numRatings                 1.489e-05  3.148e-07  47.290  < 2e-16 ***
## likedPercent               2.651e-02  2.362e-03  11.226  < 2e-16 ***
## price                     -3.561e-03  5.094e-04  -6.990 2.75e-12 ***
## genre1Nonfiction          -4.834e-01  3.439e-02 -14.057  < 2e-16 ***
## genre1Other               -1.870e+00  3.903e-02 -47.897  < 2e-16 ***
## genre1Poetry              -6.000e-03  5.515e-02  -0.109 0.913371
## genre2Classics            -6.642e-01  8.097e-02  -8.203 2.35e-16 ***
## genre2Crime                3.073e-01  8.817e-02   3.485 0.000491 ***
## genre2Drama               -2.843e-01  1.082e-01  -2.627 0.008625 **
## genre2Fantasy             -1.009e-01  7.775e-02  -1.297 0.194467
## genre2Historical           6.803e-01  1.017e-01   6.692 2.20e-11 ***
## genre2Historical Fiction   4.626e-01  7.950e-02   5.820 5.90e-09 ***
## genre2Horror              -2.453e-02  8.856e-02  -0.277 0.781814
## genre2Memoir               5.128e-01  8.717e-02   5.883 4.04e-09 ***
## genre2Other               -2.775e-01  7.889e-02  -3.517 0.000436 ***
## genre2Religion            -6.328e-01  9.344e-02  -6.772 1.27e-11 ***
## genre2Romance             -5.000e-01  7.864e-02  -6.358 2.04e-10 ***
## genre2Science Fiction      1.701e-01  8.266e-02   2.057 0.039655 *
## genre2Short Stories        4.778e-02  9.084e-02   0.526 0.598907
## genre2Thriller            -3.244e-01  8.550e-02  -3.794 0.000148 ***
## genre2War                  1.020e+00  1.046e-01   9.750  < 2e-16 ***
## genre2Young Adult          1.959e-01  7.892e-02   2.483 0.013043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116072  on 83727  degrees of freedom
## Residual deviance:  99616  on 83703  degrees of freedom
## AIC: 99666
##
## Number of Fisher Scoring iterations: 9
```

```
yHat <- logreg.mdl.3$fitted.values > 0.5
tab <- table(y, yHat)

tab
```

```
##    yHat
## y   FALSE  TRUE
##   0 25680 16184
##   1 10928 30936
```

```
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])

print('accuracy')
```

```
## [1] "accuracy"
```

```
accuracy
```

```
## [1] 0.6761896
```

```
print('precision')
```

```
## [1] "precision"
```

```
precision
```

```
## [1] 0.6565365
```

```
print('recall')
```

```
## [1] "recall"
```

```
recall
```

```
## [1] 0.7389643
```

```
print('specificity')
```

```
## [1] "specificity"
```

```
specificity
```

```
## [1] 0.701486
```

Nakon popravljenog balansa klasa u podacima accuracy se smanjila ali je sada recall puno bolji.

Priprema podataka:

```
balanced_data$z <- abs(scale(balanced_data$pages))
data_clean <- balanced_data %>% filter(z < 3)

balanced_data$z <- abs(scale(balanced_data$numRatings))
data_clean <- balanced_data %>% filter(z < 3)

balanced_data$z <- abs(scale(balanced_data$price))
data_clean <- balanced_data %>% filter(z < 3)

balanced_data$z <- abs(scale(balanced_data$rating))
data_clean <- balanced_data %>% filter(z < 3)

balanced_data$z <- abs(scale(balanced_data$likedPercent))
data_clean <- balanced_data %>% filter(z < 3)

data_clean <- data_clean %>% select(-z)
```

Učenje modela na ujednačenom skupu podataka s izbacenim strsecim vrijednostima:

```r
y <- data_clean$y
X <- data_clean %>% select(-y)
```

```
## Adding missing grouping variables: `y`
```

```r
logreg.mdl.3 = glm(y ~ ., data = X, family = binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(logreg.mdl.3)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(), data = X)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -1.0909   0.0025   1.0353   2.3199
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               5.709e-01  1.861e-01   3.068 0.002158 **
## rating                   -5.637e-01  4.027e-02 -14.000  < 2e-16 ***
## pages                     2.249e-04  3.616e-05   6.219 4.99e-10 ***
## numRatings                1.483e-05  3.147e-07  47.132  < 2e-16 ***
## likedPercent              1.845e-02  2.733e-03   6.748 1.50e-11 ***
## price                    -3.509e-03  5.081e-04  -6.906 4.99e-12 ***
## genre1Nonfiction         -4.682e-01  3.467e-02 -13.502  < 2e-16 ***
## genre1Other              -1.880e+00  3.994e-02 -47.071  < 2e-16 ***
## genre1Poetry              1.528e-03  5.549e-02   0.028 0.978036
## genre2Classics           -6.791e-01  8.116e-02  -8.368  < 2e-16 ***
## genre2Crime               3.109e-01  8.842e-02   3.516 0.000439 ***
## genre2Drama              -3.049e-01  1.088e-01  -2.802 0.005086 **
## genre2Fantasy            -1.004e-01  7.793e-02  -1.289 0.197566
## genre2Historical          6.651e-01  1.019e-01   6.527 6.70e-11 ***
## genre2Historical Fiction  4.532e-01  7.967e-02   5.688 1.28e-08 ***
## genre2Horror             -5.331e-02  8.911e-02  -0.598 0.549681
## genre2Memoir              5.002e-01  8.741e-02   5.723 1.05e-08 ***
## genre2Other              -2.935e-01  7.913e-02  -3.709 0.000208 ***
## genre2Religion           -6.517e-01  9.366e-02  -6.958 3.45e-12 ***
## genre2Romance            -5.025e-01  7.883e-02  -6.375 1.84e-10 ***
## genre2Science Fiction     1.641e-01  8.286e-02   1.980 0.047650 *
## genre2Short Stories       5.210e-02  9.116e-02   0.572 0.567642
## genre2Thriller           -3.186e-01  8.574e-02  -3.716 0.000202 ***
## genre2War                 1.008e+00  1.048e-01   9.616  < 2e-16 ***
## genre2Young Adult         1.731e-01  7.915e-02   2.187 0.028722 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 114284  on 82441  degrees of freedom
## Residual deviance:  98329  on 82417  degrees of freedom
## AIC: 98379
##
## Number of Fisher Scoring iterations: 8
```

```
yHat <- logreg.mdl.3$fitted.values > 0.5
tab <- table(y, yHat)

tab
```

```
##    yHat
## y   FALSE  TRUE
##   0 24578 16328
##   1 10625 30911
```

```
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])

print('accuracy')
```

```
## [1] "accuracy"
```

```
accuracy
```

```
## [1] 0.6730671
```

```
print('precision')
```

```
## [1] "precision"
```

```
precision
```

```
## [1] 0.6543534
```

```
print('recall')
```

```
## [1] "recall"
```

```
recall
```

```
## [1] 0.7441978
```

```
print('specificity')
```

```
## [1] "specificity"
```

```
specificity
```

```
## [1] 0.6981791
```

Kada smo maknuli stršeće vrijednosti model je malo izgubio na accuracy-u, ali model sada bolje generalizira.

Zaključno, da, možemo s nekom malom pristranošću odrediti je li je knjiga bila nagrađivana, ali podatke moramo prije balansirati. Balansiranje skupa podataka je važno jer algoritmi učenja često daju prednost većem broju primjera jedne klase u odnosu na druge. To može dovesti do "preučenosti" modela na većini primjera jedne klase i lošijeg generaliziranja na primjere druge klase. Balansiranjem skupa podataka, osiguravamo da algoritmi učenja imaju sličan broj primjera svake klase, što povećava njihovu sposobnost generaliziranja na nove primjere.