Osnove statističkog programiranja

Ak. god. 2023./2024.

Neki naslov

Dokumentacija

Julijana Kolarec, Lucija Topolko

siječanj 2024., Zagreb

Nastavnik: *prof.dr.sc. Damir Pintar*

Sadržaj

1. Uvod

Cilj ovog projektnog zadatka detaljna je analiza odabranog skupa podataka. Skup podataka čine različiti atributi, a naš je zadatak doći do dubokog razumijevanja njihovih međusobnih odnosa i potencijalnih trendova. Sve to namjeravamo ostvariti kroz proces čišćenja podataka, statističke analize i vizualizacije. Za eksploratornu analizu odabran je podatkovni skup *IMDb movie dataset* koji sadrži informacije o filmovima, uključujući ocjene, godine premijere, glumačku postavu i druge relevantne podatke, prikupljene s popularnog filmskog portala IMDb. Očekujemo da ćemo kroz analizu ovog skupa podataka istražiti i shvatiti karakteristike filmova, odnose između različitih atributa skupa te steći dublji uvid u svijet filmova.

1.1 Pregled i čišćenje podataka

Originalni skup podata *IMDb movie dataset* sastoji se od ukupno 5043 zapisa s ukupno 28 atributa. Izvođenjem jednostavne naredbe

```
1 sum(duplicated(data))
```

utvrđeno je da dupli cirani zapisi čine 45 redaka izvornog skupa. Dupli cirani su redci izbačeni iz skupa i konačni se skup sastoji od 4998 zapisa. Prije eksportiranja uređenog skupa za daljnje korištenje tijekom analize, zbog lakšeg je snalaženja promijenjen i redoslijed stupaca. Redoslijed stupaca promijenjen je izvršavanjem sljedeće naredbe:

```
1 new_order <- c('movie_title', 'duration', 'director_name', 'director_
    facebook_likes', 'actor_1_name', 'actor_1_facebook_likes', 'actor_2_
    name', 'actor_2_facebook_likes', 'actor_3_name', 'actor_3_facebook_
    likes', 'num_user_for_reviews', 'num_critic_for_reviews', 'num_voted_
    users', 'cast_total_facebook_likes', 'movie_facebook_likes', 'plot_
    keywords', 'facenumber_in_poster', 'color', 'genres', 'title_year', '
    language', 'country', 'content_rating', 'aspect_ratio', 'movie_imdb_
    link', 'gross', 'budget', 'imdb_score')
2 data <- data[, new_order]
```

Imena varijabli i opisi značenja mogu se provjeriti u tablici na web stranici Kaggle[1].

[1] https://www.kaggle.com/code/harshadeepvattikunta/predicting-movie-success

Nakon navedenih izmjena, podatkovni je skup spreman za eksportiranje u `.csv` formatu. Sve daljnje analize provode se nad novodobivenom, *očišćenom*, verzijom skupa podataka.

1.2 Osnovne informacije o atributima podatkovnog skupa

U ovom ćemo dijelu, s ciljem boljeg upoznavanja sa skupom podatka, provesti jednostavne analize nad podacima svakog stupca zasebno.

Najprije primjenom funkcije `sapply` saznajemo broj vrijednosti koje nedostaju (*NA* vrijednosti) u svakom stupcu.

Stupac	**Broj**	**Stupac**	**Broj**
movie_title	0	num_user_for_reviews	21
duration	15	num_critic_for_reviews	49
director_name	103	num_voted_users	0
director_facebook_likes	103	cast_total_facebook_likes	0
actor_1_name	7	movie_facebook_likes	0
actor_1_facebook_likes	7	plot_keywords	152
actor_2_name	13	facenumber_in_poster	13
actor_2_facebook_likes	13	color	19
actor_3_name	23	genres	0
actor_3_facebook_likes	23	title_year	107
language	12	country	5
content_rating	301	aspect_ratio	327
movie_imdb_link	0	gross	874
budget	487	imdb_score	0

Tablica 1.1: Broj nepoznatih vrijednosti po stupcu

Podaci u stupcima s nazivom `actor_n_facebook_likes`, $n = 1, 2, 3$, sadrže podatke o broju *lajkova* na Facebook stranici glumca `actor_n_name`, $n = 1, 2, 3$. Izdvajanjem imena glumaca i njihovih odgovarajućih brojeva `lajkova` te uzimajući u obzir samo najviši broj *lajkova* za pojedinog glumca, dobivamo podatke o najpoznatijim glumcima (najpoznatiji u ovom kontekstu znači s najviše *lajkova*). Imena i broj *lajkova* pet najpoznatijih glumaca navedeni su u tablici 1.2.

Ime glumca	**Broj Facebook *lajkova***
Darcy Donavan	640,000
Matthew Ziff	260,000
Krista Allen	164,000
Andrew Fiscella	137,000
Jimmy Bennett	87,000

Tablica 1.2: Najpoznatiji glumci

Također, na temelju podataka iz stupaca `actor_n_name`, `n` = 1,2,3 te `director_name` izdvojili smo glumce i redatelje s najviše filmova. U tablici 1.3 prikazan je popis 5 glumaca s najviše uloga, dok su u tablici 1.4 prikazani redatelji koji su režirali najviše filmova.

Ime glumca	**Broj uloga**
Robert De Niro	54
Morgan Freeman	47
Bruce Willis	40
Johnny Depp	40
Matt Damon	38

Tablica 1.3: Glumci s najviše uloga

Ime redatelja	**Broj režija**
Steven Spielberg	26
Woody Allen	22
Clint Eastwood	20
Martin Scorsese	20
Ridley Scott	17

Tablica 1.4: Najčešći redatelji

Iz podataka sadržanih u stupcu nazvanom `cast_total_facebook_likes` moguće je identificirati filmove s najpoznatijom glumačkom postavom, a to su filmovi navedeni u tablici 1.5.

Naslov filma	Godina premijere	IMDB ocjena	Broj *lajkova* postave
Anchorman: The Legend of Ron Burgundy	2004	7.2	656,730
The Final Destination	2009	5.2	303,717
Treachery	2013	3.9	283,939
Hardflip	2012	5.6	263,584
Kickboxer: Vengeance	2016	9.1	261,818

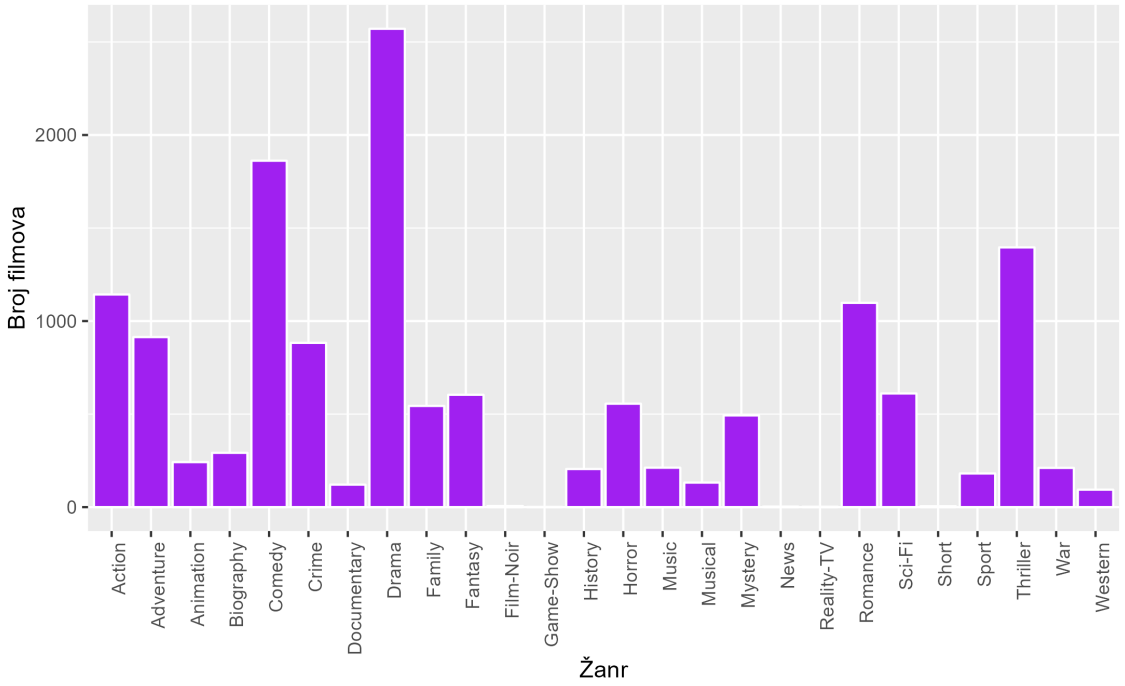
Tablica 1.5: Filmovi s najpoznatijom glumačkom postavom

Stupac `plot_keywords` sastoji se od ključnih riječi koje opisuju radnju filma odvojenih znakom '|'. Razdvajanjem sadržaja stupca po znaku '|' izdvajamo pojedinačne ključne riječi i saznajemo koje su najčešće te ih navodimo u tablici 1.6.

Ključna riječ	Broj filmova	Ključna riječ	Broj filmova
love	194	fbi	71
friend	165	revenge	70
murder	159	friendship	67
death	132	drugs	66
police	126	prison	62
new york city	91	money	61
high school	89	marriage	60
alien	82	female protagonist	57
school	73	island	57
boy	72	dog	56

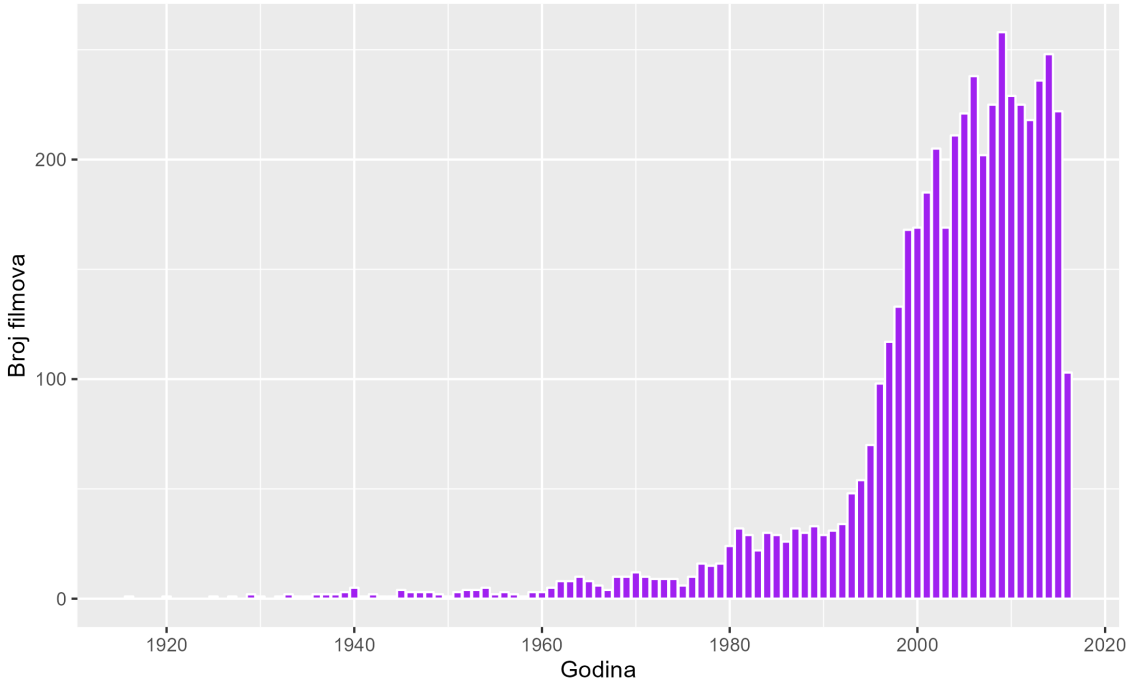
Tablica 1.6: Najčešće ključne riječi koje opisuju radnju filma

Sličan stupcu `plot_keywords` stupac je `genres` koji, odvojene znakom '|', sadrži informacije o žanrovima filmova. Izdvajamo žanrove za svaki film i prikazujemo broj filmova po svakom od žanrova u histogramu na slici 1.1.



Slika 1.1: Podjela filmova po žanru

Stupac `title_year` poprima vrijednosti od 1916 do 2016, a predstavlja godinu premijere filma. Koje je godine premijerno prikazano koliko filmova prikazano je na slici 1.2.



Slika 1.2: Broj filmova po godini premijere

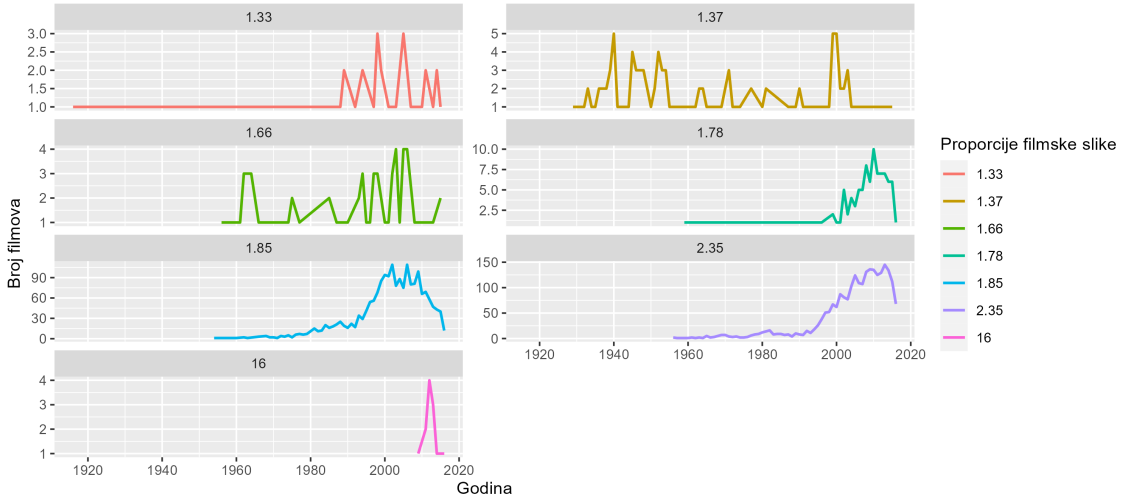
U stupcu `facenumber_in_poster` zapisan je broj glumaca koji se pojavljuju na plakatu filma. Za većinu filmova (njih ukupno 2136) taj je broj nula, a najveći broj glumaca na plakatu iznosi 43 (za film *500 Days of Summer*). Prosječna je vrijednost atributa `facenumber_in_poster` 1.37, a medijan 1.

Za koju je dobnu skupinu film namijenjen sadržano je u stupcu `content_rating`. Popis najčešćih starosnih ograničenja i njihova značenja dana su u tablici 1.7.

Ograničenje	Broj filmova	Značenje
R	2098	Za osobe starije od 17 godina
PG-13	1444	Za osobe starije od 13 godina
PG	688	Za osobe starije od 8 godina

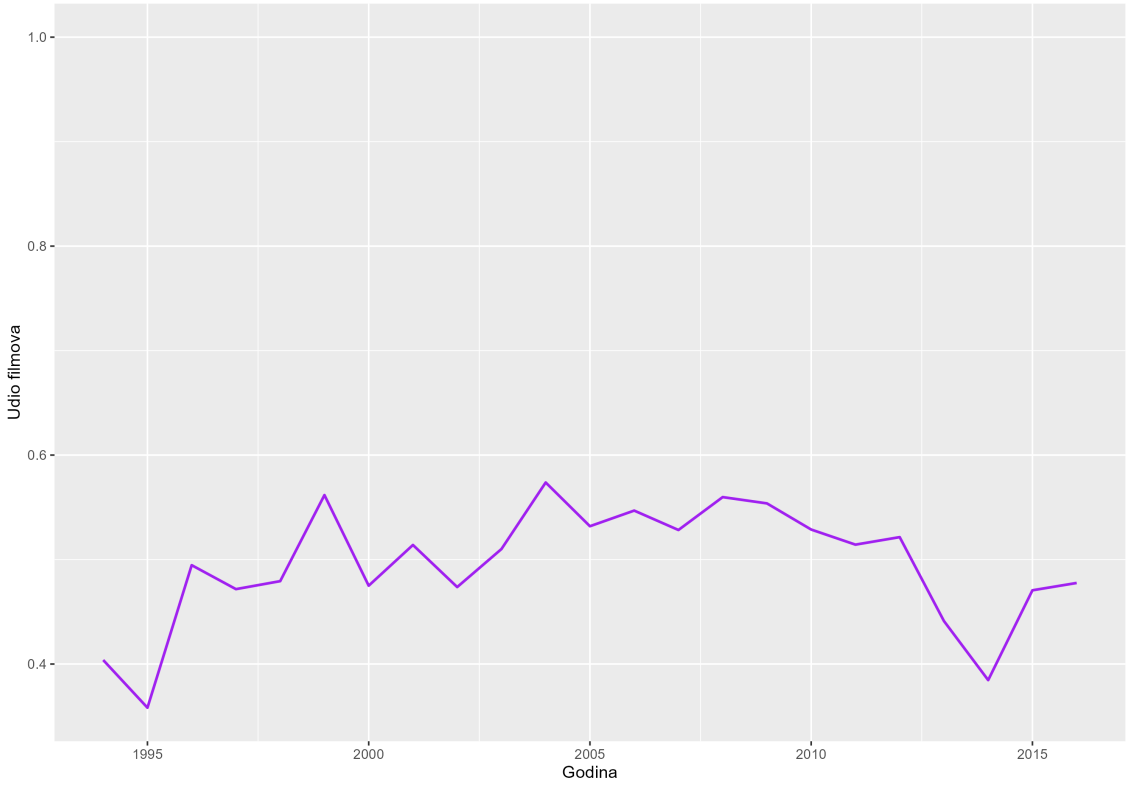
Tablica 1.7: Najčešća starosna ograničenja filmova

Omjer širine i visine (proporcije) filmske slike za pojedini film zapisan je u stupcu `aspect_ratio`. U skupu podataka pojavljuje se ukupno 23 različitih omjera, a za sedam najčešćih napravljen je pregled (slika 1.3) kretanja broja filmova s tim proporcijama po godinama.



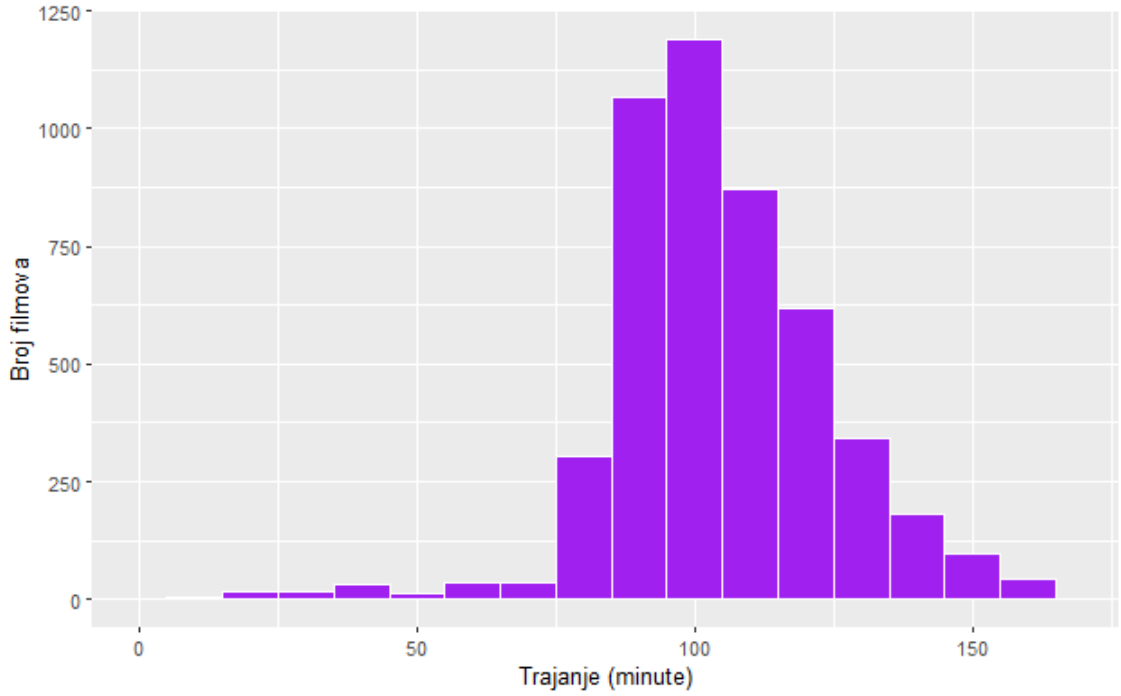
Slika 1.3: Broj filmova s najčešćim proporcijama filmske slike po godinama

U stupcima `budget` i `gross` nalaze se podaci o budžetu filma i ukupnoj bruto zaradi filma u američkim dolarima. Koliki je postotak filmova svake godine ostvario manju zaradu od iznosa budžeta prikazujemo na slici 1.4.

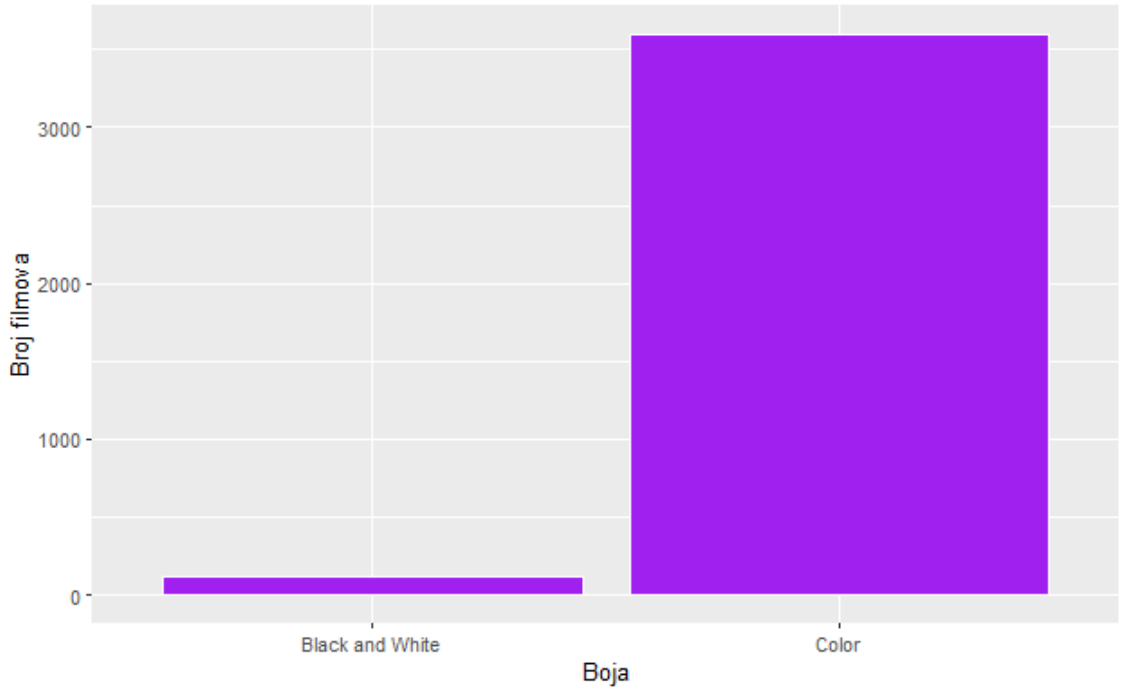


Slika 1.4: Postotak filmova koji su ostvarili manji prihoda od iznosa budžeta

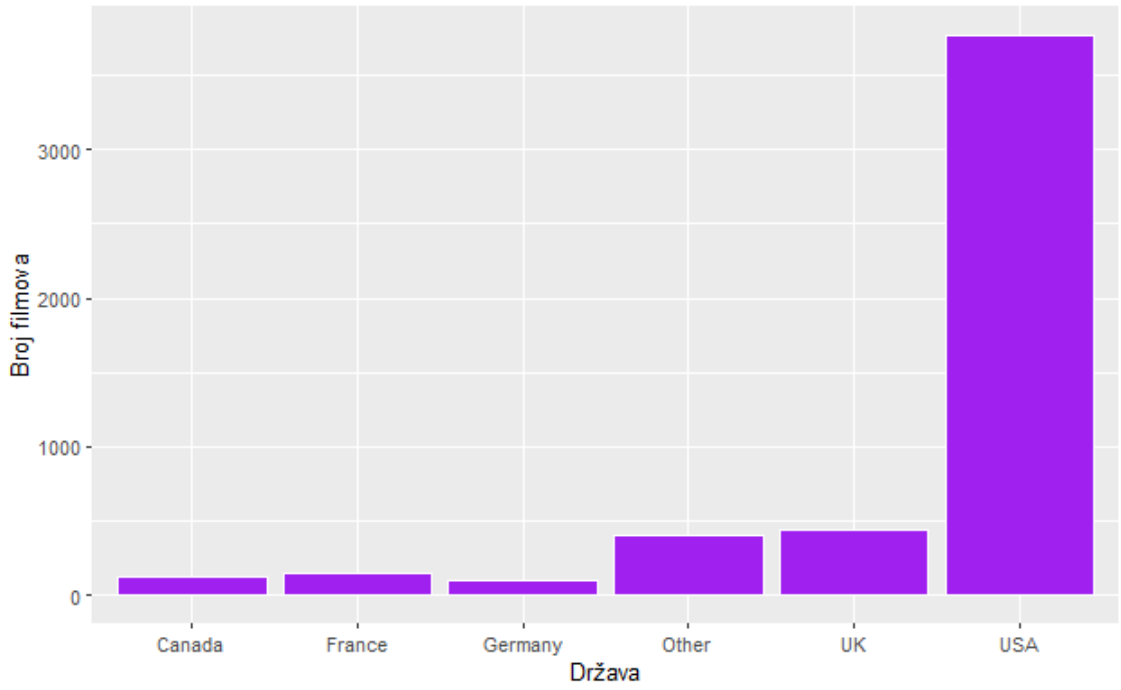
Stupac `duration` sadrži podatke o trajanju filmova. Najkraći film iz našeg skupa podataka traje 7, a najduži 511 minuta. Prosječno film traje 107 minuta. Histogram na slici 1.5 prikazuje broj filmova u ovisnosti o njihovom trajanju.



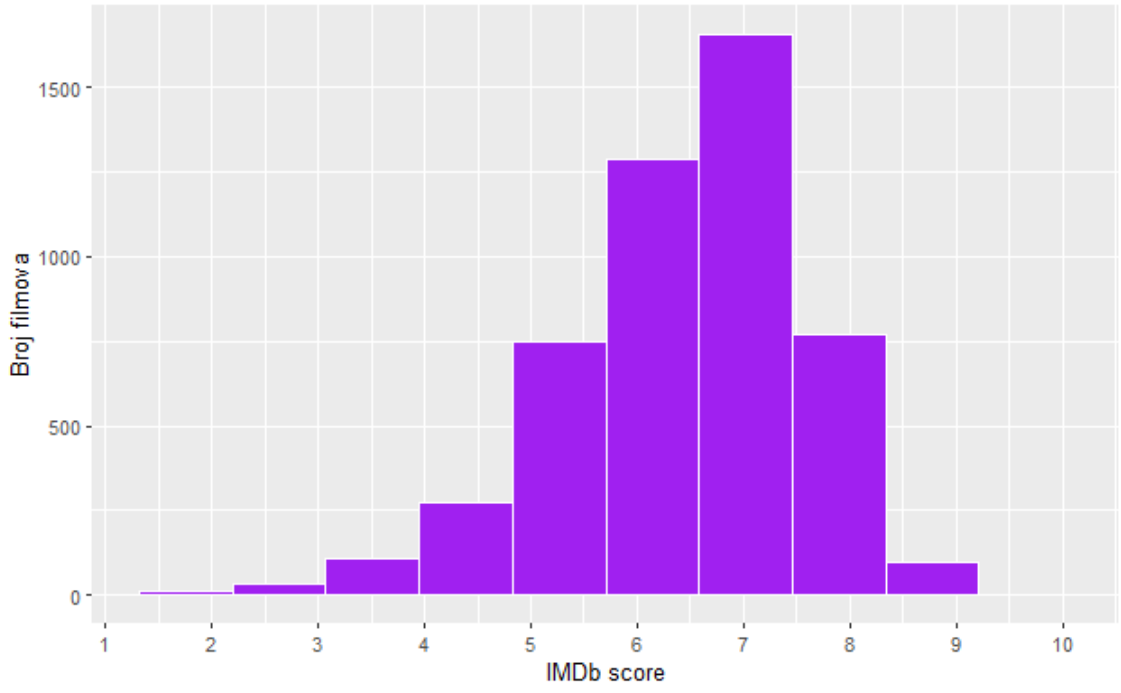
Slika 1.5: Broj filmova po trajanju



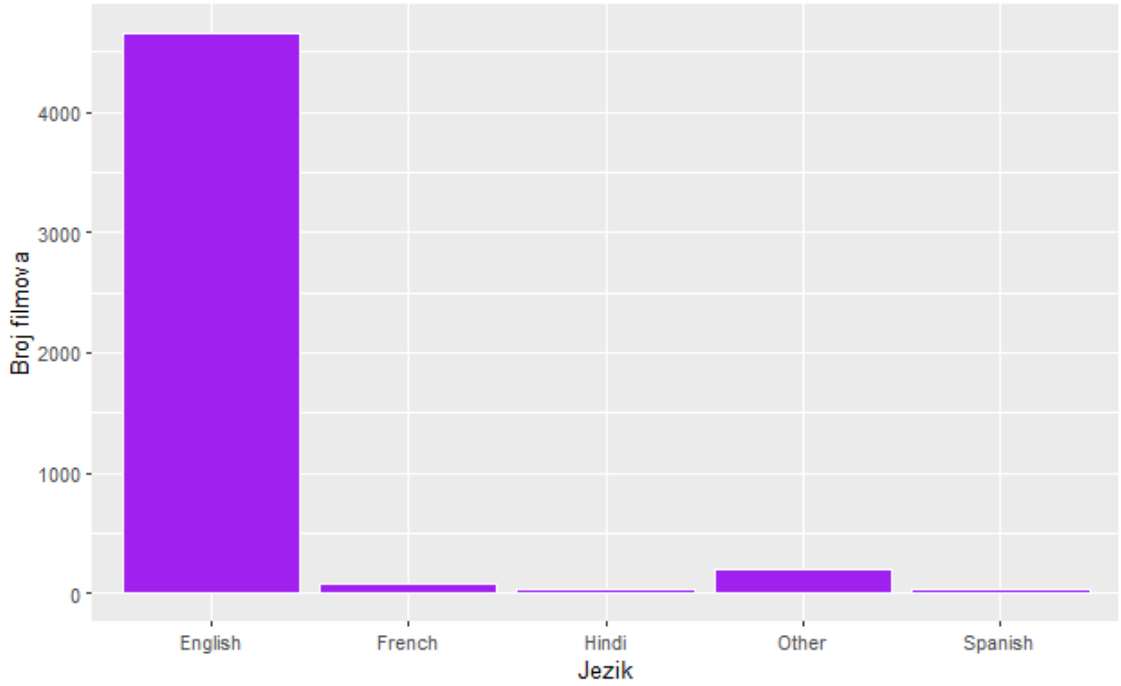
Slika 1.6: Podjela filmova po boji



Slika 1.7: Podjela filmova po državi nastanka



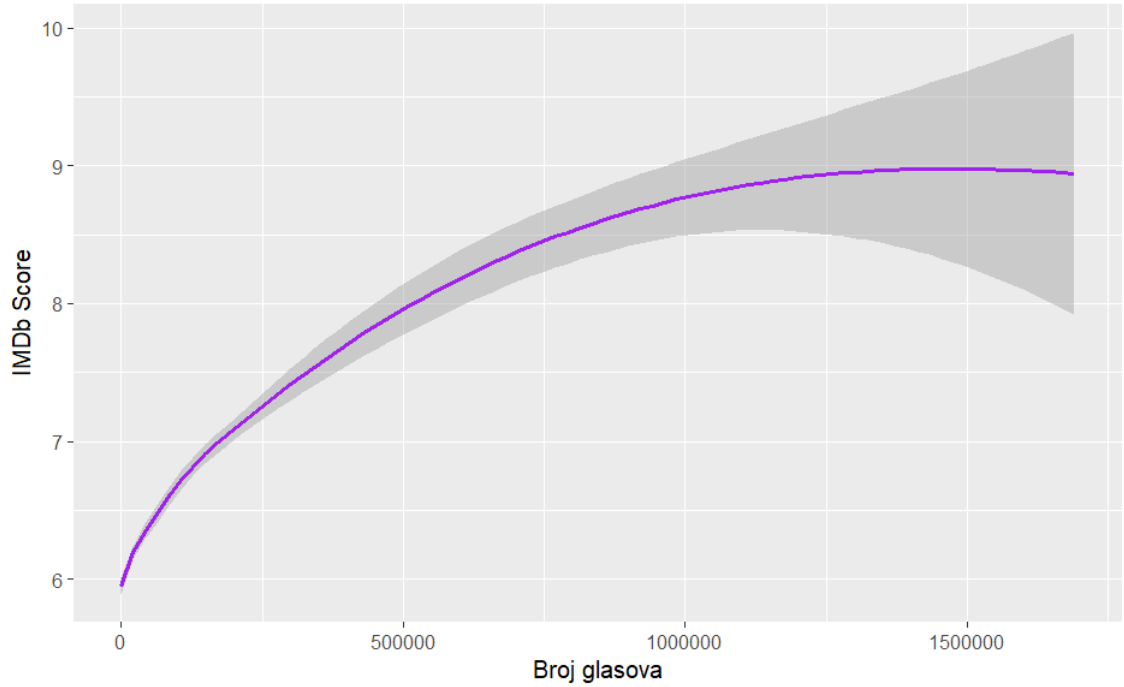
Slika 1.8: Podjela filmova po uspješnosti



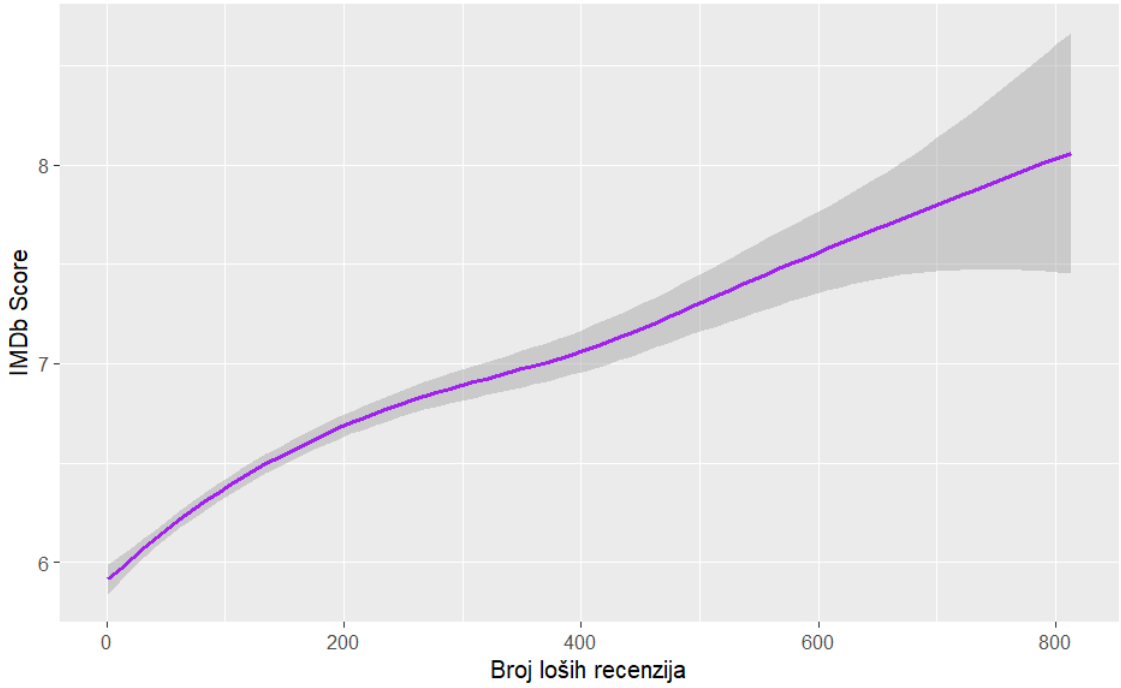
Slika 1.9: Broj filmova s obzirom na jezik

2. Naprednije analize podataka

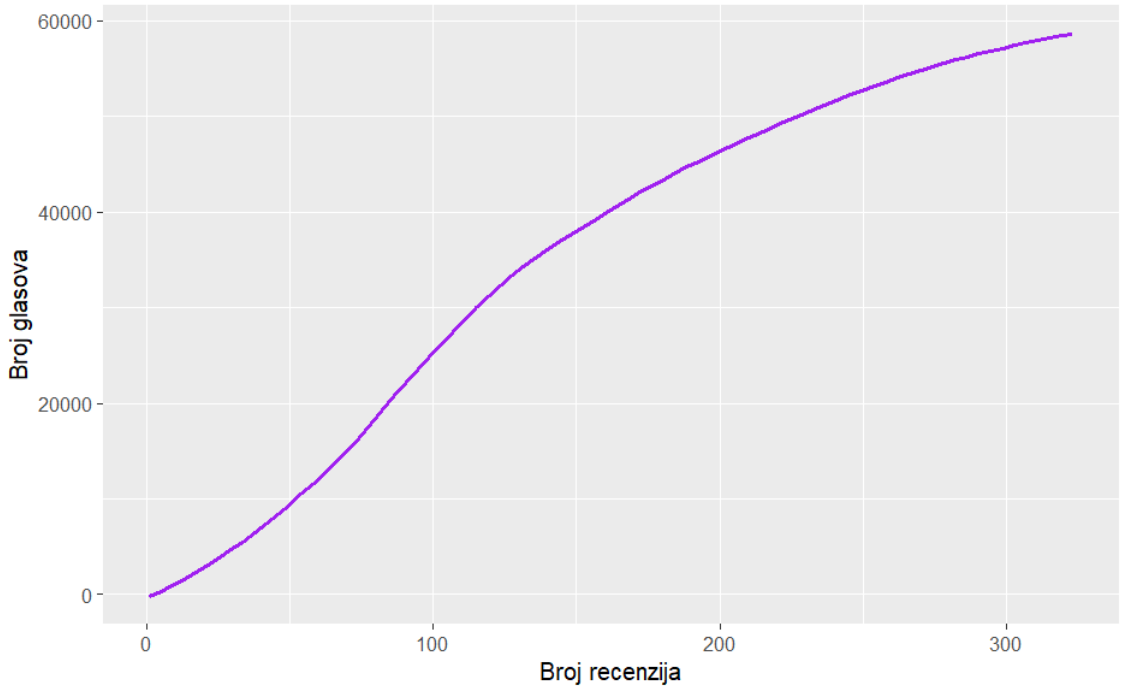
2.1 Proučavanje međusobne ovisnosti atributa



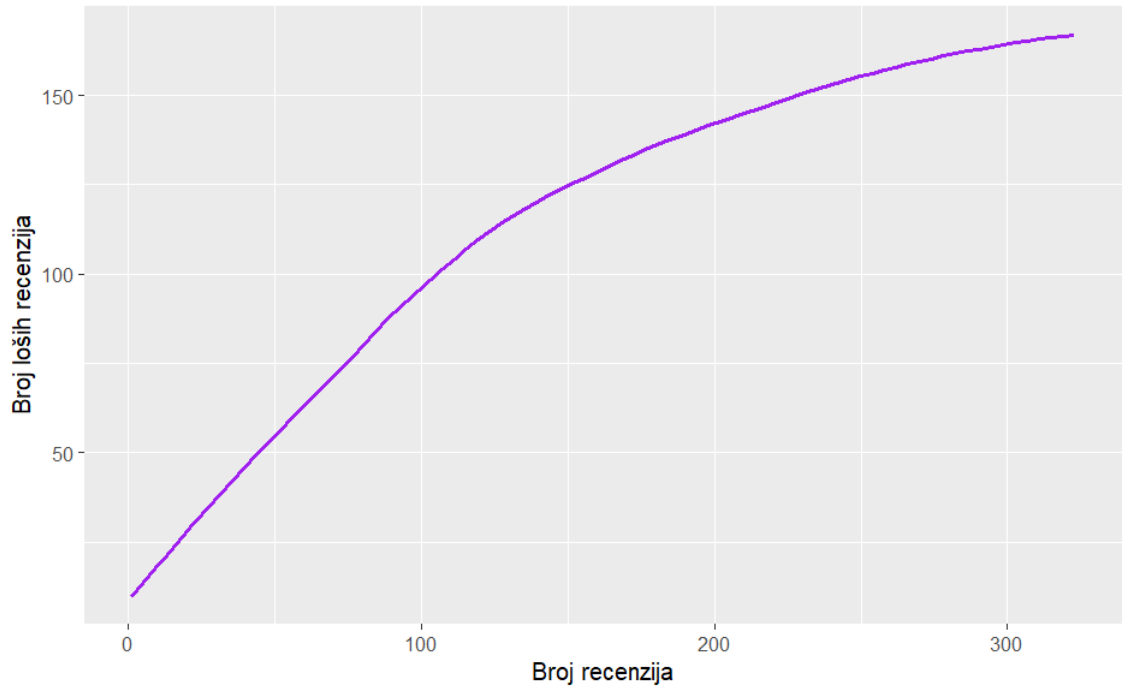
Slika 2.1: IMDb ocjena u ovisnosti o broju glasova



Slika 2.2: IMDb ocjena u ovisnosti o broju loših recenzija



Slika 2.3: Broj loših recenzija u ovisnosti o ukupnom broju recenzija



Slika 2.4:

2.2 Dodatne zanimljive vizualizacije

3. Prediktivni modeli primjenom strojnog učenja

Kako bismo bolje razumjeli što film čini uspješnim ili neuspješnim, provele smo analizu dobivenog skupa podataka primjenom strojnog učenja. Cilj nam je bio razviti model koji može čim točnije predviđati uspjeh filma na temelju njegovih karakteristika.

3.1 Priprema podataka

Iz dobivenih podataka izbacile smo retke kojima su nedostajali neki podaci. Takvih je redaka bilo 1261. Također, uklonile smo stupce koji su sadržavali jedinstvene ili skoro jedinstvene vrijednosti (*movie_title, movie_imdb_link, plot_keywords, genres*). Još smo izbacile tekstualne stupce koji su bili prekorelirani s nekim numeričkim stupcem. Na primjer, *actor_1_name* je prekoreliran s *actor_1_facebook_likes*.

```
columns <- c('duration', 'director_facebook_likes', 'actor_1_facebook_
    likes','actor_2_facebook_likes', 'actor_3_facebook_likes', 'num_user_
    for_reviews', 'num_critic_for_reviews', 'num_voted_users', 'cast_
    total_facebook_likes', 'movie_facebook_likes', 'facenumber_in_poster'
    , 'color', 'title_year', 'language', 'country', 'content_rating', '
    aspect_ratio', 'gross', 'budget', 'imdb_score')

mldata <- data[,columns]

mldata <- na.omit(mldata)
```

Preostale nenumeričke stupce pretvorile smo u tip integer.

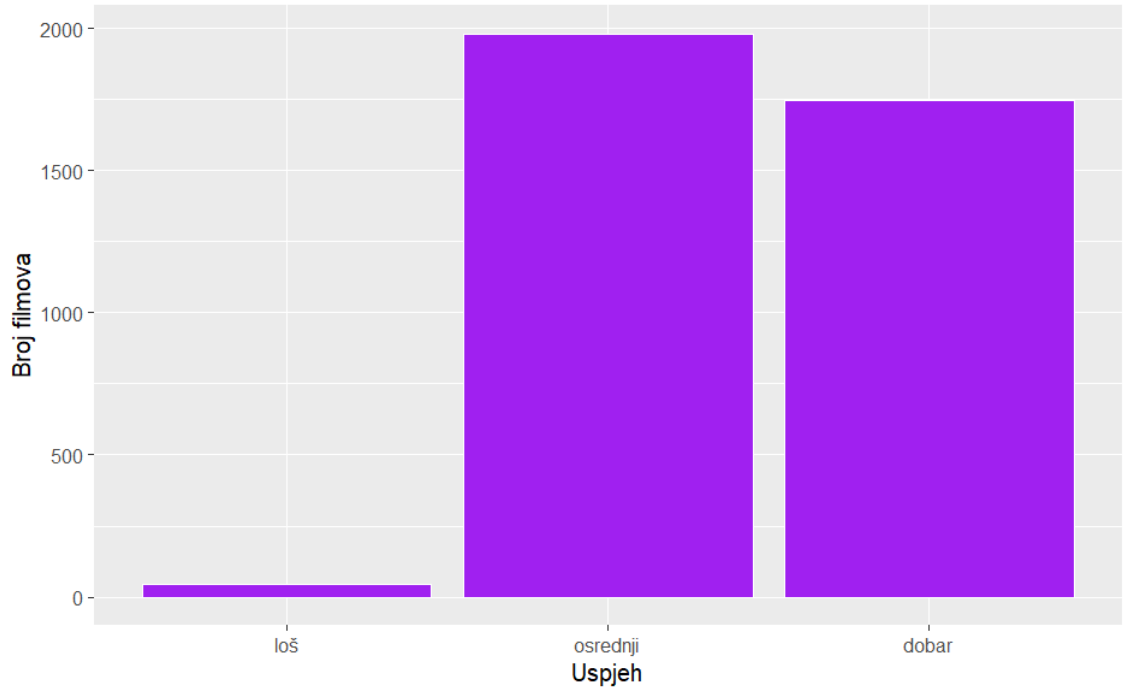
```
label_encode <- function(column) {
  as.integer(factor(column, levels = unique(column)))
}
```

Značajka koju predviđamo je *imdb_score*. To broj zaokružen na jednu decimalu, pa smo za bolje rezultate uspjeh filma podijelile u tri skupine: loš, osrednji i dobar,

a stupac imdb_score smo zbog prekoreliranosti uklonile.

```
mldata$score <- ifelse(mldata$imdb_score < 3.33, "los", ifelse(mldata$
    imdb_score < 6.66, "osrednji", "dobar"))
```

Graf 3.1 prikazuje omjer broja filmova po uspjehu. Filmova koji su ocijenjeni kao loši znatno je manje od ostalih. Točnije, loših je filmova 43, osrednjih 1981, a dobrih 1746.



Slika 3.1: Podjela filmova po uspjehu

3.2 Modeli s nebalansiranim skupom podataka

Kako bismo razvile model za predviđanje uspješnosti, skup podataka podijelile smo u omjeru 80:20. Na temelju 80% gradile smo model, a na 20% ga testirale.
Prvi model koji smo razvile koristi metodu potpornih vektora.

```
Model <- train(score ~ ., data = training_set,
               method = "svmPoly",
               na.action = na.omit,
               preProcess=c("scale","center"),
               trControl= trainControl(method="none"),
               tuneGrid = data.frame(degree=1,scale=1,C=1)
)
```

Model radi s uspješnošću od 72.8%.

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš        0        0     0
  osrednji   8      339   141
  dobar      0       54   205

Overall Statistics

               Accuracy : 0.7282
                 95% CI : (0.6948, 0.7599)
    No Information Rate : 0.5261
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4518

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity             0.00000          0.8626       0.5925
Specificity             1.00000          0.5791       0.8653
Pos Pred Value              NaN          0.6947       0.7915
Neg Pred Value          0.98929          0.7915       0.7111
Prevalence              0.01071          0.5261       0.4632
Detection Rate          0.00000          0.4538       0.2744
Detection Prevalence    0.00000          0.6533       0.3467
Balanced Accuracy       0.50000          0.7208       0.7289
```

Slika 3.2: Metoda potpornih vektora - rezultati

Drugi model koji smo razvile koristi metodu slučajne šume. Rezultati su nešto bolji, uspješnost je 78.3%.

```
Model_rf <- randomForest(score ~ ., data = training_set, ntree = 500,
    importance = TRUE)
```

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš        0        0     0
  osrednji   7      325    86
  dobar      1       68   260

Overall Statistics

               Accuracy : 0.7831
                 95% CI : (0.7518, 0.8122)
    No Information Rate : 0.5261
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5677

 Mcnemar's Test P-Value : 0.0177

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity             0.00000          0.8270       0.7514
Specificity             1.00000          0.7373       0.8279
Pos Pred Value              NaN          0.7775       0.7903
Neg Pred Value          0.98929          0.7933       0.7943
Prevalence              0.01071          0.5261       0.4632
Detection Rate          0.00000          0.4351       0.3481
Detection Prevalence    0.00000          0.5596       0.4404
Balanced Accuracy       0.50000          0.7821       0.7897
```

Slika 3.3: Metoda slučajne šume - rezultati

Iako su ovi rezultati na prvi pogled donekle zadovoljavajući, nijedan od ovih modela nije predvidio da će ijedan film biti loš. To je očekivani rezultat jer podaci nisu nimalo balansirani - loših filmova je znatno manje pa ih je i puno teže predvidjeti. Produkciji filma bilo bi najkorisnije imati model koji može predvidjeti neuspjeh filma, a ovi modeli to ne uspijevaju pa smo ih odbacile.

3.3 Modeli s balansiranim skupom podataka

S ciljem poboljšanja točnosti predviđanja loših filmova, podatke smo balansirale. Nastojale smo broj loših i dobrih filmova približiti broju osrednjih filmova (Graf 3.4).

```r
oversample <- ovun.sample(score~., data = over, method = "both", N =
    3932)$data
```

Slika 3.4: Podjela filmova po uspjehu - balansirani podaci

Nad novim smo podacima ponovo testirale naše modele. Ovaj je put metoda potpornih vektora postigla uspješnost od 75.6% (Slika 3.5), a metoda slučajne šume visokih 90.9% (Slika 3.6).

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš      374       69    34
  osrednji   7      256   109
  dobar      0       68   261

Overall Statistics

               Accuracy : 0.7564
                 95% CI : (0.7308, 0.7806)
    No Information Rate : 0.343
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6352

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity              0.9816          0.6514       0.6460
Specificity              0.8708          0.8522       0.9121
Pos Pred Value           0.7841          0.6882       0.7933
Neg Pred Value           0.9900          0.8300       0.8316
Prevalence               0.3234          0.3336       0.3430
Detection Rate           0.3175          0.2173       0.2216
Detection Prevalence     0.4049          0.3158       0.2793
Balanced Accuracy        0.9262          0.7518       0.7791
```

Slika 3.5: Metoda potpornih vektora - rezultati s balansiranim podacima

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš      381        0     0
  osrednji   0      330    44
  dobar      0       63   360

Overall Statistics

               Accuracy : 0.9092
                 95% CI : (0.8913, 0.925)
    No Information Rate : 0.343
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8637

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity              1.0000          0.8397       0.8911
Specificity              1.0000          0.9439       0.9186
Pos Pred Value           1.0000          0.8824       0.8511
Neg Pred Value           1.0000          0.9216       0.9417
Prevalence               0.3234          0.3336       0.3430
Detection Rate           0.3234          0.2801       0.3056
Detection Prevalence     0.3234          0.3175       0.3591
Balanced Accuracy        1.0000          0.8918       0.9048
```

Slika 3.6: Metoda slučajne šume - rezultati s balansiranim podacima

Ovim smo rezultatima zadovoljne jer oba modela s visokom točnošću predviđaju loše filmove.

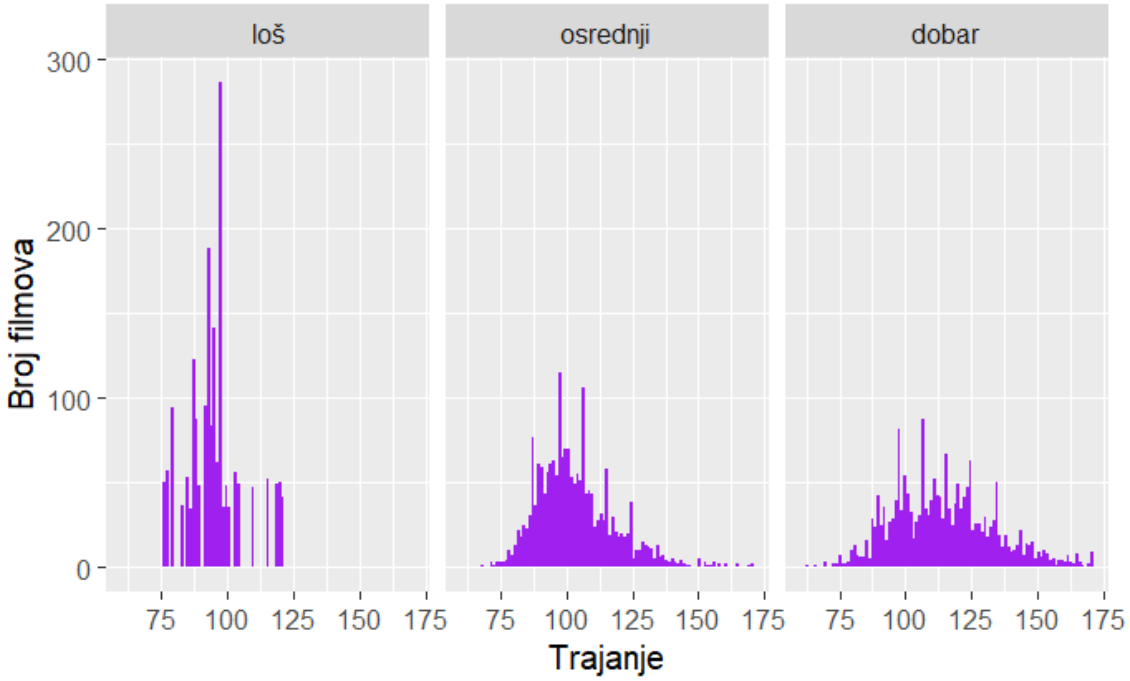
3.4 Atributi najznačajniji za predviđanje

Idući je korak u analizi bio otkriti koji atributi najviše koriste pri predviđanju uspješnosti filmova, posebice onih loših.

Analizu smo provele nad modelom koji koristi metodu slučajne šume i balansirani skup podataka jer upravo taj model daje najbolje rezultate.

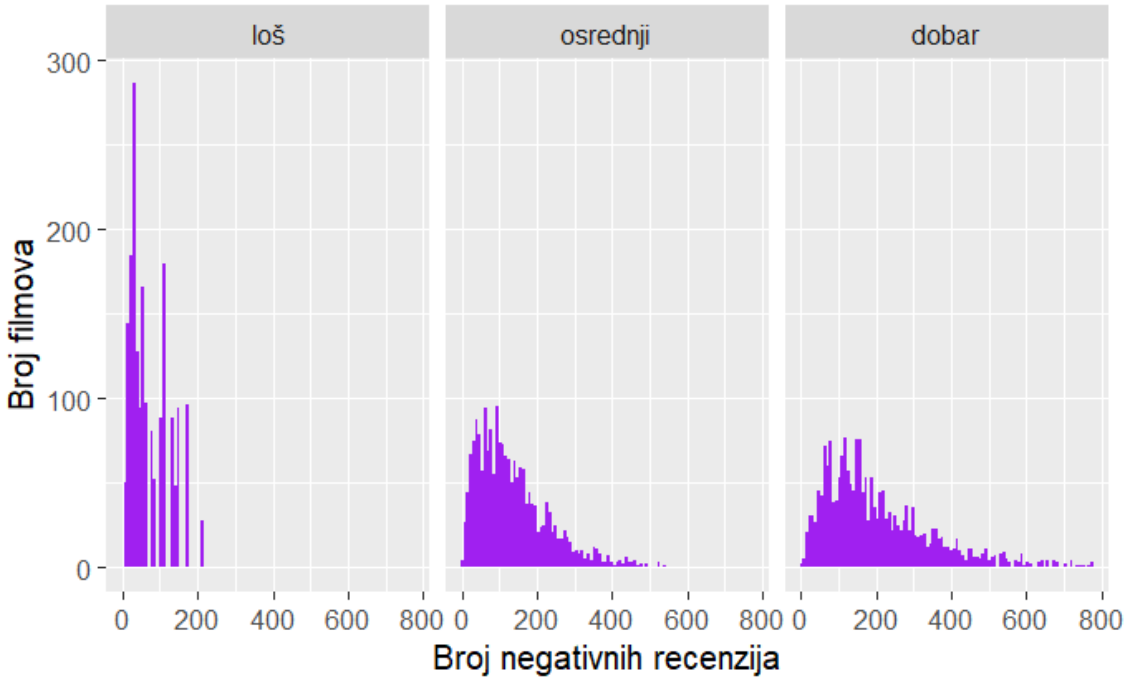
Atributi koji su u našem modelu u najvećoj korelaciji s uspješnosti su trajanje i broj negativnih recenzija.

Filmovi koji su dobili loše ocjene gledatelja najčešće traju između sat i dva sata, a većina ih traje do 100 minuta. Grafovi za ostale filmove također prikazuju da najviše filmova traje 100 ili više minuta.



Slika 3.7: Metoda slučajne šume - rezultati s balansiranim podacima

Za predviđanje uspješnih i neuspješnih filmova bio je važan i broj negativnih recenzija. Zanimljivo, filmovi koje smo klasificirale kao neuspješne imali su manji broj negativnih recenzija. Razlog tome je vjerojatno taj što se velik broj ljudi odlučio uopće ne pogledati film kad je vidio da je većina recenzija negativna. Uspješnije filmove pogleda puno više ljudi različitih mišljenja pa je očekivano da se nekima neće svidjeti.



Slika 3.8: Metoda slučajne šume - rezultati s balansiranim podacima

Atributi koji su najmanje korelirani s uspjehom filma su jezik i broj ljudi na plakatu.

Indeks slika i dijagrama