Eksploratorna analiza IMDb movie skupa podataka

36557473 Milković Borna

2026-01-28

Opis IMDb movies skupa podataka

Za početak ćemo učitati skup podataka i proučiti ga. U nastavku su navedena objašnjenja pojedinih stupaca:

- `movie_title` - naslov filma
- `title_year` - godina izlaska filma
- `duration` - trajanje filma (u minutama)
- `color` - govori nam je li film u boji ili je crno bijeli ("Color", "Black and White")
- `genres` - žanrovi u kojima film pripada
- `director_name` - ime i prezime redatelja
- `country` - država produkcije filma
- `language` - jezik kojim se primarno priča u filmu
- `budget` - novac uložen u kreiranje filma
- `gross` - novac zarađen na prodaji ulaznica za film
- `imdb_score` - ocjena filma na IMDb stranici (0-10)
- `content_rating` - oznaka za primjerenu dob gledatelja: `PG-13` - neprimjereno za djecu mlađu od 13 godina, `R` - ograničeno gledanje za osobe mlađe od 17 godina, `NC-17` - samo za odrasle, ... (za više informacija: https://www.filmratings.com/)
- `num_user_for_reviews` - broj pisanih komentara gledatelja
- `num_critic_for_reviews` - broj profesionalnih recenzija
- `num_voted_users` - broj ljudi koji su dali ocjenu
- `movie_facebook_likes` - broj oznaka sviđanja na filmovoj facebook stranici
- `facenumber_in_poster` - broj ljudi vidljiv na plakatu filma
- `cast_total_facebook_likes` - oznake sviđanja svih glumaca u filmu
- `actor_1_name`, `actor_2_name`, `actor_3_name` - imena i prezimena glavna 3 glumca filma
- `director_facebook_likes` - oznake sviđanja za redatelja na facebooku

Učitavanje i proučavanje skupa

```
movies <- read.csv("../data/IMDB_movie_dataset.csv")
#head(movies)
glimpse(movies)
```

```
## Rows: 5,043
## Columns: 28
## $ color                    <chr> "Color", "Color", "Color", "Color", "", "Col~
## $ director_name            <chr> "James Cameron", "Gore Verbinski", "Sam Mend~
## $ num_critic_for_reviews   <int> 723, 302, 602, 813, NA, 462, 392, 324, 635, ~
## $ duration                 <int> 178, 169, 148, 164, NA, 132, 156, 100, 141, ~
```

```
## $ director_facebook_likes   <int> 0, 563, 0, 22000, 131, 475, 0, 15, 0, 282, 0~
## $ actor_3_facebook_likes    <int> 855, 1000, 161, 23000, NA, 530, 4000, 284, 1~
## $ actor_2_name              <chr> "Joel David Moore", "Orlando Bloom", "Rory K~
## $ actor_1_facebook_likes    <int> 1000, 40000, 11000, 27000, 131, 640, 24000, ~
## $ gross                     <int> 760505847, 309404152, 200074175, 448130642, ~
## $ genres                    <chr> "Action|Adventure|Fantasy|Sci-Fi", "Action|A~
## $ actor_1_name              <chr> "CCH Pounder", "Johnny Depp", "Christoph Wal~
## $ movie_title               <chr> "Avatar ", "Pirates of the Caribbean: At Wor~
## $ num_voted_users           <int> 886204, 471220, 275868, 1144337, 8, 212204, ~
## $ cast_total_facebook_likes <int> 4834, 48350, 11700, 106759, 143, 1873, 46055~
## $ actor_3_name              <chr> "Wes Studi", "Jack Davenport", "Stephanie Si~
## $ facenumber_in_poster      <int> 0, 0, 1, 0, 0, 1, 0, 1, 4, 3, 0, 0, 1, 2, 1,~
## $ plot_keywords             <chr> "avatar|future|marine|native|paraplegic", "g~
## $ movie_imdb_link           <chr> "http://www.imdb.com/title/tt0499549/?ref_=f~
## $ num_user_for_reviews      <int> 3054, 1238, 994, 2701, NA, 738, 1902, 387, 1~
## $ language                  <chr> "English", "English", "English", "English", ~
## $ country                   <chr> "USA", "USA", "UK", "USA", "", "USA", "USA",~
## $ content_rating            <chr> "PG-13", "PG-13", "PG-13", "PG-13", "", "PG-~
## $ budget                    <dbl> 237000000, 300000000, 245000000, 250000000, ~
## $ title_year                <int> 2009, 2007, 2015, 2012, NA, 2012, 2007, 2010~
## $ actor_2_facebook_likes    <int> 936, 5000, 393, 23000, 12, 632, 11000, 553, ~
## $ imdb_score                <dbl> 7.9, 7.1, 6.8, 8.5, 7.1, 6.6, 6.2, 7.8, 7.5,~
## $ aspect_ratio              <dbl> 1.78, 2.35, 2.35, 2.35, NA, 2.35, 2.35, 1.85~
## $ movie_facebook_likes      <int> 33000, 0, 85000, 164000, 0, 24000, 0, 29000,~
```

```
#summary(movies)
```

Skup podataka sastoji se od 5043 podatka opisanih s 28 stupaca. Skup podataka opisuje informacije o filmovima koji se nalaze na stranici IMDb (https://www.imdb.com), a najnoviji podaci su iz 2016. godine.

Nedostajuće vrijednosti

U prvih par redaka možemo vidjeti neke nedostajuće vrijednosti (`NA`), ali također i prazne znakovne nizove (`""`) unutar pojedinih znakovnih stupaca. Prvo ćemo pretvoriti prazne znakovne nizove u nedostajuće vrijednosti.

```
broj_nedostajucih_prije <- movies %>% is.na() %>% sum()

movies$color <- na_if(movies$color, "")
movies$director_name <- na_if(movies$director_name, "")
movies$actor_2_name <- na_if(movies$actor_2_name, "")
movies$genres <- na_if(movies$genres, "")
movies$actor_1_name <- na_if(movies$actor_1_name, "")
movies$movie_title <- na_if(movies$movie_title, "")
movies$actor_3_name <- na_if(movies$actor_3_name, "")
movies$plot_keywords <- na_if(movies$plot_keywords, "")
movies$movie_imdb_link <- na_if(movies$movie_imdb_link, "")
movies$language <- na_if(movies$language, "")
movies$country <- na_if(movies$country, "")
movies$content_rating <- na_if(movies$content_rating, "")

broj_nedostajucih_poslije <- movies %>% is.na() %>% sum()
```

```
cat ("Broj novih NA vrijednosti: " , broj_nedostajucih_poslije - broj_nedostajucih_prije)
```

```
## Broj novih NA vrijednosti:  639
```

Sada ćemo proučiti gdje se nalazi najviše nedostajućih vrijednosti i odlučiti kako ćemo postupiti s njima.

```
na_counts <- sapply(movies, function(x) sum(is.na(x)))
sort(na_counts, decreasing = TRUE)
```

```
##                     gross                    budget              aspect_ratio
##                       884                       492                       329
##            content_rating             plot_keywords                title_year
##                       303                       153                       108
##             director_name   director_facebook_likes    num_critic_for_reviews
##                       104                       104                        50
##    actor_3_facebook_likes              actor_3_name      num_user_for_reviews
##                        23                        23                        21
##                     color                  duration              actor_2_name
##                        19                        15                        13
##      facenumber_in_poster    actor_2_facebook_likes                  language
##                        13                        13                        12
##    actor_1_facebook_likes              actor_1_name                   country
##                         7                         7                         5
##                    genres               movie_title           num_voted_users
##                         0                         0                         0
## cast_total_facebook_likes          movie_imdb_link                imdb_score
##                         0                         0                         0
##      movie_facebook_likes
##                         0
```

Stupci s najviše nedostajućih vrijednosti su povezani s potrošnjom i zaradom na filmu. Za sada nećemo uklanjati nedostajuće vrijednosti jer čine dosta velik udio u uzorku podataka (883 od 5043), ali ćemo ih ukloniti prije analiza koje koriste navedene stupce.

Uklanjanje duplikata

Pregledom skupa, uočena su ista imena filmova, ovaj problem riješiti ćemo uklanjanjem duplikata.

```
movies <- distinct(.data = movies, movie_title, .keep_all = TRUE)
```

Kategorijske varijable i uređivanje skupa

Za početak možemo uočiti neke varijable znakovnog tipa poput imena države i jezika filma možemo kategorizirati pretvaranjem u tip `factor`. Također, stupac `movie_imdb_link` nam neće biti potreban za daljnju analizu pa ćemo ga ukloniti.

```
movies$language <- factor(movies$language)
movies$country <- factor(movies$country)
movies$color <- factor(movies$color)
movies$content_rating <- factor(movies$content_rating)
movies$movie_imdb_link <- NULL
```

Trenutačno nam je stupac o žanrovima filma strukturiran tako da su svi žanrovi navedeni u jednom stupcu odvojeni okomitom crtom. Takav način mogao bi biti problematičan kod linearne regresije i treniranja modela, stoga ćemo stvoriti još jedan skup koji ćemo proširiti s stupcem za svaki postojeći žanr i vrijednostima 0 i 1 u zavisnosti je li film tog žanra ili ne.

```
movies %>% separate_rows(genres, sep = "\\|") %>% count(genres, sort = TRUE)
```

```
## # A tibble: 26 x 2
##    genres        n
##    <chr>     <int>
##  1 Drama      2533
##  2 Comedy     1847
##  3 Thriller   1364
##  4 Action     1113
##  5 Romance    1084
##  6 Adventure   888
##  7 Crime       868
##  8 Sci-Fi      594
##  9 Fantasy     583
## 10 Horror      539
## # i 16 more rows
```

```
movies_reg <- movies %>%
  separate_rows(genres, sep = "\\|") %>%
  mutate(value = 1) %>%
  pivot_wider(names_from  = genres, values_from = value, values_fill = 0)

#glimpse(movies_reg)[28:52]
```

Zanimljivosti iz skupa podataka

Za početak, koristeći jednostavne upite koji podsjećaju na one u jeziku SQL, odgovorimo na nekoliko zanimljivih pitanja o filmovima korištenjem danog podatkovnog okvira.

5 najbolje ocijenjenih filmova na IMDb-u

```
top5 <- movies %>% slice_max(order_by = imdb_score, n = 5) %>%
  dplyr::select(movie_title, imdb_score)
top5
```

```
##                     movie_title imdb_score
## 1 Towering Inferno                      9.5
## 2     The Shawshank Redemption          9.3
## 3                The Godfather          9.2
## 4           Dekalog                     9.1
## 5         Kickboxer: Vengeance          9.1
```

5 filmova s najviše ocjena korisnika

```
top5_num_votes <- movies %>% slice_max(order_by = num_voted_users, n = 5) %>%
  dplyr::select(movie_title, num_voted_users, imdb_score)
top5_num_votes
```

```
##                 movie_title num_voted_users imdb_score
## 1 The Shawshank Redemption          1689764        9.3
## 2          The Dark Knight          1676169        9.0
## 3                Inception          1468200        8.8
## 4              Fight Club          1347461        8.8
## 5             Pulp Fiction          1324680        8.9
```

5 filmova s najviše ocjena kritičara

```
top5_crit_votes <- movies %>% slice_max(order_by = num_critic_for_reviews, n = 5) %>%
  dplyr::select(movie_title, num_critic_for_reviews, imdb_score)
top5_crit_votes
```

```
##              movie_title num_critic_for_reviews imdb_score
## 1 The Dark Knight Rises                    813        8.5
## 2            Prometheus                    775        7.0
## 3     Django Unchained                     765        8.5
## 4              Skyfall                     750        7.8
## 5    Mad Max: Fury Road                    739        8.1
```

5 najskupljih filmova

```
top5_budget <- movies %>% slice_max(order_by = budget, n = 5) %>%
  dplyr::select(movie_title, budget, imdb_score)
top5_budget
```

```
##          movie_title     budget imdb_score
## 1    Lady Vengeance  4200000000        7.7
## 2          Fateless  2500000000        7.1
## 3 Princess Mononoke  2400000000        8.4
## 4          Steamboy  2127519898        6.9
## 5             Akira  1100000000        8.1
```

Ovdje je zanimljivo primijetiti da su čak 3 animirana filma (Princess Mononoke, Steamboy, Akira) u top 5 najskupljih filmova.

5 filmova s najvećom zaradom od prodaje ulaznica

```
top5_gross <- movies %>% slice_max(order_by = gross, n = 5) %>%
  dplyr::select(movie_title, gross, imdb_score)
top5_gross
```

```
##        movie_title     gross imdb_score
## 1           Avatar 760505847        7.9
## 2          Titanic 658672302        7.7
## 3  Jurassic World 652177271        7.0
## 4    The Avengers 623279547        8.1
## 5 The Dark Knight 533316061        9.0
```

Kao što vidimo to su neki od najpoznatijih filmova današnjice.

5 crno-bijelih filmova s najboljim IMDb ocjenama

```
top5_bw <- movies %>% filter(color != "Color") %>%
  slice_max(order_by = imdb_score, n = 5) %>%
  dplyr::select(movie_title, color, imdb_score)

top5_bw
```

```
##                     movie_title           color imdb_score
## 1             Schindler's List   Black and White        8.9
## 2                 12 Angry Men   Black and White        8.9
## 3                 Forrest Gump   Black and White        8.8
## 4 The Honeymooners               Black and White        8.7
## 5                Seven Samurai   Black and White        8.7
```

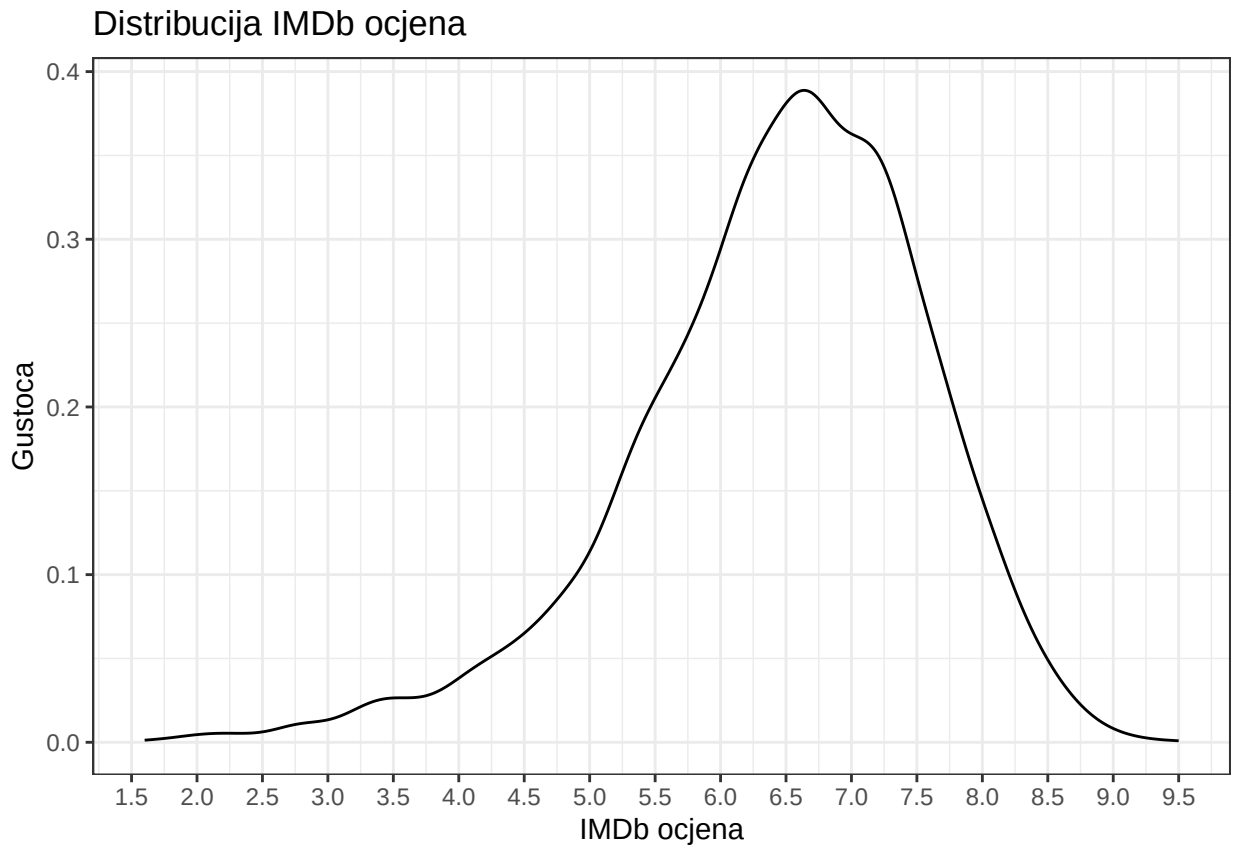
Vizualizacija podataka

U ovom dijelu analize, fokusirat ćemo se na grafičke prikaze varijabli podatkovnog skupa te odnose, odnosno zavisnosti parova varijabli.

Distribucija IMDb ocjena

Pogledajmo kako su distribuirane ocjene za filmove u podatkovnom okviru:

```
ggplot(movies, aes(imdb_score)) + geom_density() +
  scale_x_continuous(breaks = seq(0, 10, by = 0.5)) +
  labs(title = "Distribucija IMDb ocjena", x = "IMDb ocjena", y = "Gustoca") + theme_bw()
```




```
summary(movies$imdb_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.600   5.800   6.600   6.438   7.200   9.500
```

IMDb ocjena može biti bilo koji broj iz intervala [1, 10], a ocjene za filmove unutar promatranih podataka nalaze se unutar intervala [1.6, 9.5] s medijanom 6.6. Graf koji smo dobili nazivamo negativno zakrivljenom Gaussovom krivuljom. U nastavku analize pitat ćemo se imaju li neke druge dane varijable utjecaj na ovu ocjenu.

Jezik filma

Prvo ćemo proučiti jezik filma. Očekivano je pretpostaviti da je najviše filmova na engleskom jeziku. S takvim filmovima susrećemo se svakodnevno kada upalimo televizor. Pogledajmo sada broj filmova po pojedinom jeziku u našem skupu podataka.

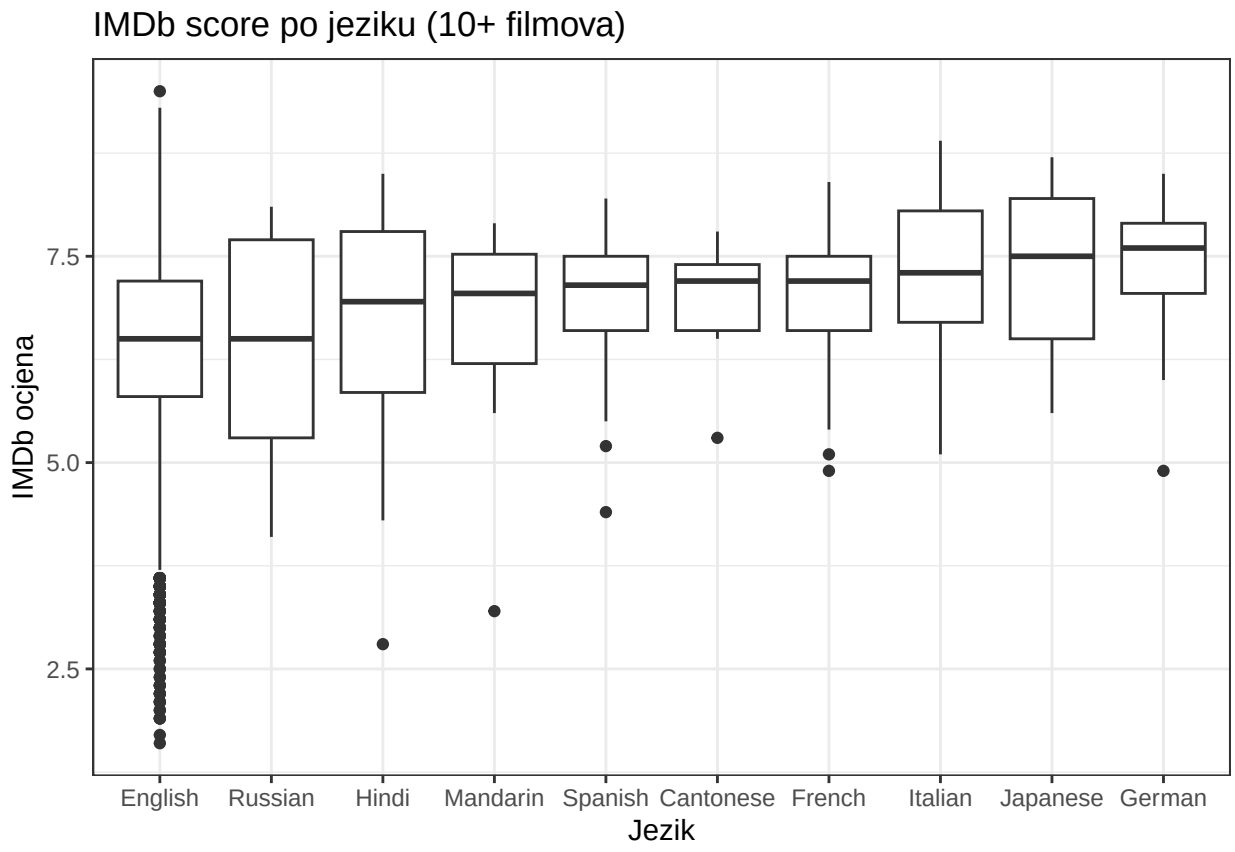
```
sort(table(movies[!is.na(movies$language), ]$language), decreasing = TRUE)
```

```
##
##    English     French    Spanish      Hindi   Mandarin     German   Japanese
##       4583         73         40         28         24         19         17
##  Cantonese    Italian    Russian Portuguese     Korean     Arabic     Danish
##         11         11         11          8          7          5          5
##     Hebrew    Swedish      Dutch  Norwegian    Persian    Chinese     Polish
```

```
##           5           5           4           4           4           3           3
##        Thai Aboriginal        Dari  Icelandic Indonesian        None   Romanian
##           3           2           2           2           2           2           2
##       Zulu     Aramaic     Bosnian      Czech   Dzongkha   Filipino      Greek
##           2           1           1           1           1           1           1
##  Hungarian     Kannada      Kazakh       Maya  Mongolian    Panjabi  Slovenian
##           1           1           1           1           1           1           1
##    Swahili       Tamil      Telugu       Urdu Vietnamese
##           1           1           1           1           1
```

Vidimo da su naše pretpostavke istinite, filmova na engleskom ima znatno više nego ostalih. Uz filmove na engleskom u skupu su najviše zastupljeni filmovi na francuskom i španjolskom. Pogledajmo sada kako izgleda IMDb ocjena za jezika s više od po 10 filmova.

```
movies_lan <- movies[!is.na(movies$language) , ]
movies_lan %>%
  count(language) %>%
  filter(n >= 10) %>%
  inner_join(movies, by = "language") %>%
  ggplot(aes(x = reorder(language, imdb_score, FUN = median),
             y = imdb_score)) +
  geom_boxplot() + scale_y_continuous() +
  labs(title = "IMDb score po jeziku (10+ filmova)",
       x = "Jezik",
       y = "IMDb ocjena") + theme_bw()
```



Na pravokutnim dijagramima možemo vidjeti raspone ocjena po najčešćim jezicima filmova, kao i medijan i ostale kvartile. Najveći raspon imaju filmovi na engleskom jeziku upravo zbog velike brojnosti istih. Najveći medijan ocjena imaju filmovi na njemačkom jeziku.

Pogledajmo još jednu zanimljivost vezanu za filmove koji nisu na engleskom jeziku.

```
top5_not_english <- movies %>% filter(language != "English") %>%
  slice_max(order_by = imdb_score, n = 5) %>%
  dplyr::select(movie_title, language, imdb_score)

top5_not_english
```

5 najbolje ocijenjenih filmova koji nisu na engleskom

```
##                       movie_title   language imdb_score
## 1            Dekalog                  Polish        9.1
## 2 The Good, the Bad and the Ugly     Italian        8.9
## 3           Gomorrah                  Italian        8.7
## 4                   City of God  Portuguese        8.7
## 5                 Seven Samurai    Japanese        8.7
```

Država produkcije filma

Osim jezika imamo podatke i o državi u kojoj je film napravljen. Za očekivati je da su američki filmovi najbrojniji zbog snažnog utjecaja Hollywooda na američku filmsku industriju. Ovaj utjecaj najbolje prikazuje brojnost američkih filmova naspram ostalih:

```
number_usa <- movies %>% filter(country == "USA") %>% count()
number_no_usa <- movies %>% filter(country != "USA") %>% count()

number_usa
number_no_usa
```

```
##      n
## 1 3711
##      n
## 1 1201
```

S otprilike trostruko većim brojem filmova, američka filmska industrija ima puno veći uzorak iz kojeg filmovi mogu biti bolje ili lošije ocijenjeni, ali pogledajmo sada tablice najbolje ocijenjenih američkih filmova i najbolje ocijenjenih filmova koji ne potječu iz Amerike:

```
top5_usa <- movies %>% filter(country == "USA") %>%
  slice_max(order_by = imdb_score, n = 5) %>%
  dplyr::select(movie_title, country, imdb_score)
cat("5 najbolje ocijenjenih američkih filmova:\n")
top5_usa
```

```
## 5 najbolje ocijenjenih američkih filmova:
##                 movie_title country imdb_score
## 1 The Shawshank Redemption      USA        9.3
## 2            The Godfather      USA        9.2
## 3     Kickboxer: Vengeance      USA        9.1
## 4          The Dark Knight      USA        9.0
## 5   The Godfather: Part II      USA        9.0
## 6        Fargo                 USA        9.0
```

```
top5_no_usa <- movies %>% filter(country != "USA") %>%
  slice_max(order_by = imdb_score, n = 5) %>%
  dplyr::select(movie_title, country, imdb_score)
cat("5 najbolje ocijenjenih filmova koji nisu američki:\n")
top5_no_usa
```

```
## 5 najbolje ocijenjenih filmova koji nisu američki:
##                                         movie_title     country imdb_score
## 1                      Towering Inferno                  Canada        9.5
## 2                               Dekalog                  Poland        9.1
## 3                   The Good, the Bad and the Ugly         Italy        8.9
## 4 The Lord of the Rings: The Fellowship of the Ring  New Zealand        8.8
## 5                             Gomorrah                    Italy        8.7
## 6                                      City of God       Brazil        8.7
## 7                          Queen of the Mountains   Kyrgyzstan        8.7
## 8                                  Seven Samurai        Japan        8.7
```

Zanimljivo je uočiti da je unatoč malom broju filmova, ostatak svijeta proizveo jako dobro ocijenjene filmove.

Filmski budžet i zarada od filma

Sljedeće varijable koje ćemo promatrati su budžet koji je utrošen za snimanje i produkciju filma te zarada od filma. Obje vrijednosti su zapisane u američkim dolarima, a kako su iznosi prilično visoki, u većini slučajeva ćemo koristiti logaritamsku transformaciju koja će nam urednije predstaviti podatke i ovisnosti o drugim varijablama. Kao česta praksa kod logaritamske transformacije dodajemo svim podacima vrijednost jedan iz predostrožnosti jer log(0) nije definirano.

```
movies_budget <- movies[!is.na(movies$budget),]
movies_gross <- movies[!is.na(movies$gross),]

summary(movies_budget$budget)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 2.180e+02 6.000e+06 1.980e+07 3.654e+07 4.300e+07 4.200e+09
```

```
summary(movies_gross$gross)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##       162   5019656  25043962  47644515  61108413 760505847
```

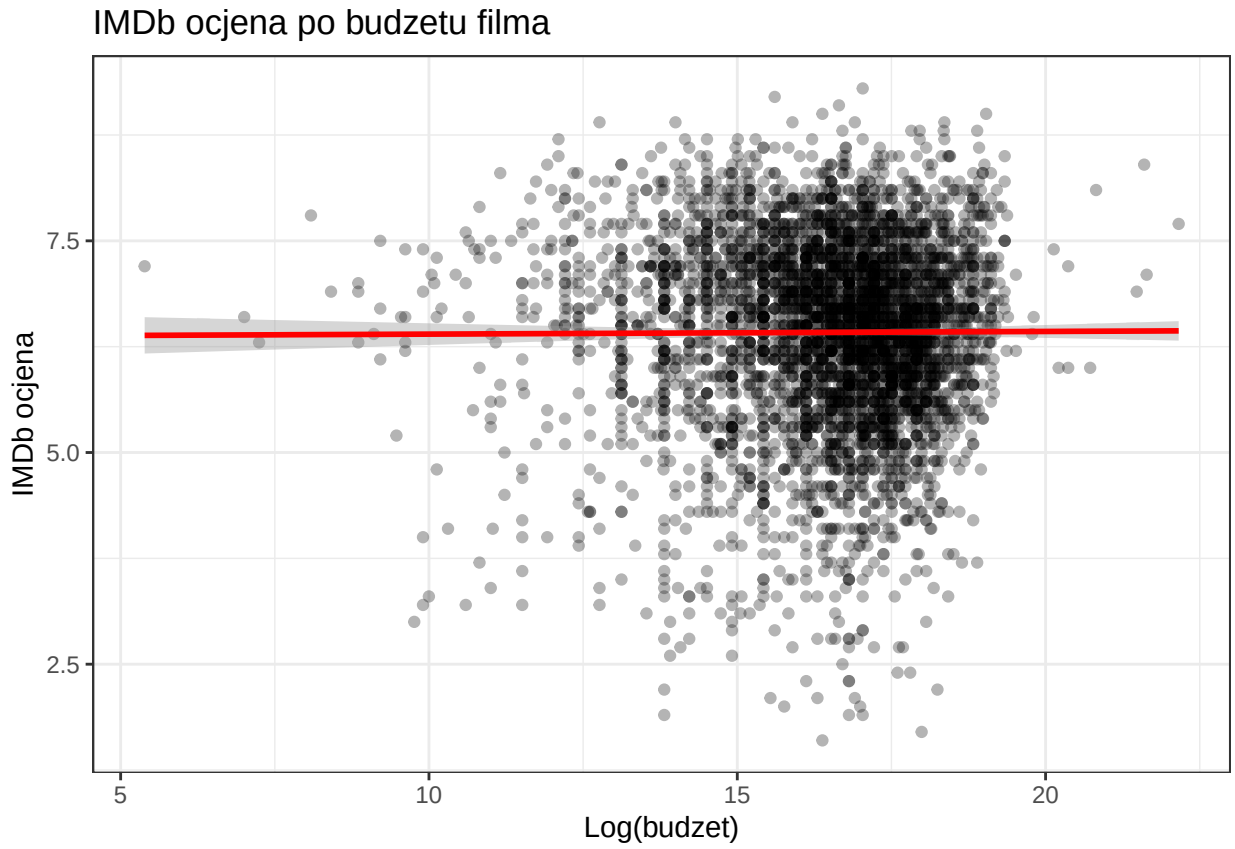
Vidimo da su vrijednosti zarade raspršenije od vrijednosti budžeta. Minimalna vrijednost zarade od filma iznosi 162, što se na prvi pogled čini kao krivo unesen podatak, daljnjim proučavanjem saznajemo da IMDb ocjena tog filma iznosi 5.7, što može donekle objasniti podatak o zaradi kao filmski neuspjeh.

Omjer budžeta i ocjene

```
cor(movies_budget$budget, movies_budget$imdb_score)
```

```
## [1] 0.05094991
```

```
ggplot(movies_budget, aes(x = log(budget + 1), y = imdb_score)) +
  geom_point(alpha = 0.3) +
  geom_smooth(formula = y ~ x, method = "lm", color = "red") +
  labs(title = "IMDb ocjena po budzetu filma",
       x = "Log(budzet)",
       y = "IMDb ocjena") + theme_bw()
```



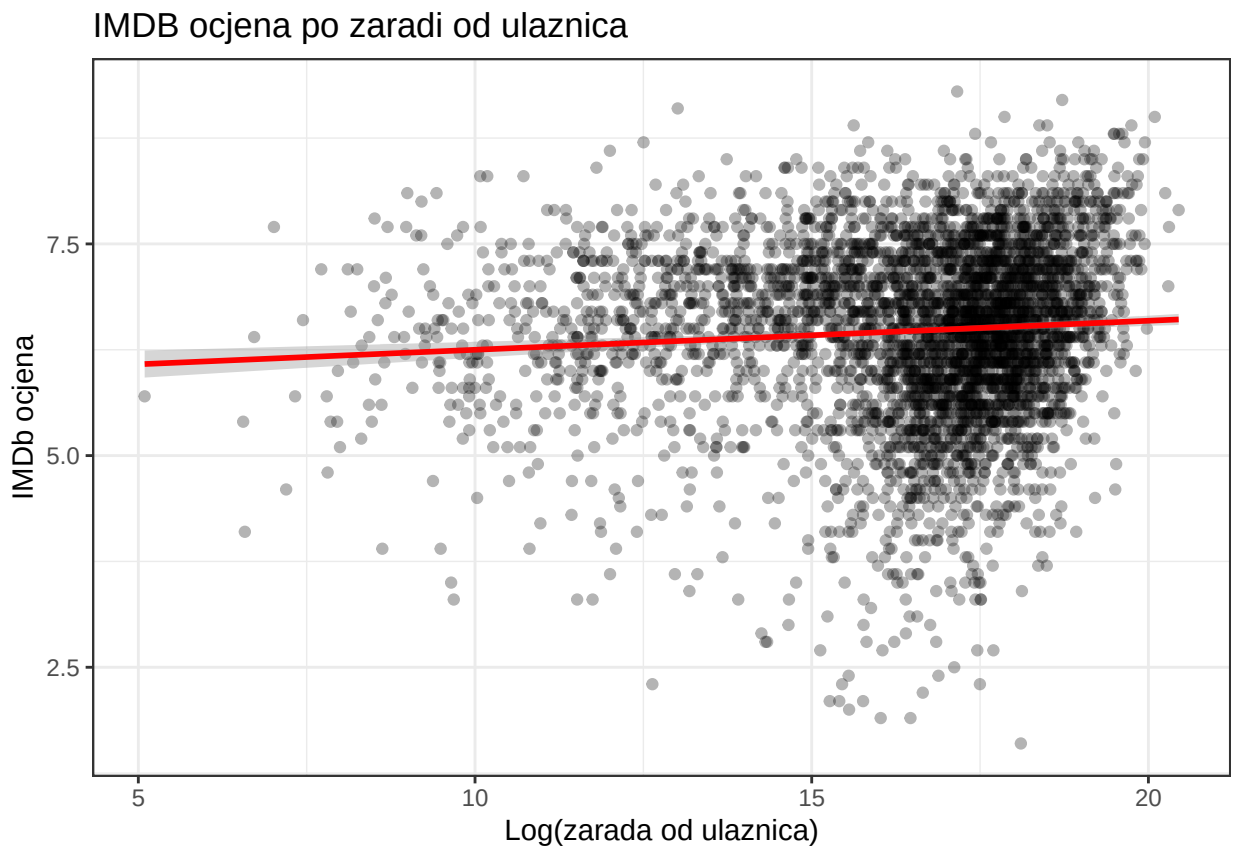
Prema grafu i malom koeficijentu korelacije možemo zaključiti da nema linearne ovisnosti između ocjene i budžeta.

Omjer zarade i ocjene

```
cor(movies_gross$gross, movies_gross$imdb_score)
```

```
## [1] 0.1999136
```

```
ggplot(movies_gross, aes(x = log(gross + 1), y = imdb_score)) +
  geom_point(alpha = 0.3) +
  geom_smooth(formula = y ~ x, method = "lm", color = "red") +
  labs(title = "IMDB ocjena po zaradi od ulaznica",
       x = "Log(zarada od ulaznica)",
       y = "IMDb ocjena") + theme_bw()
```



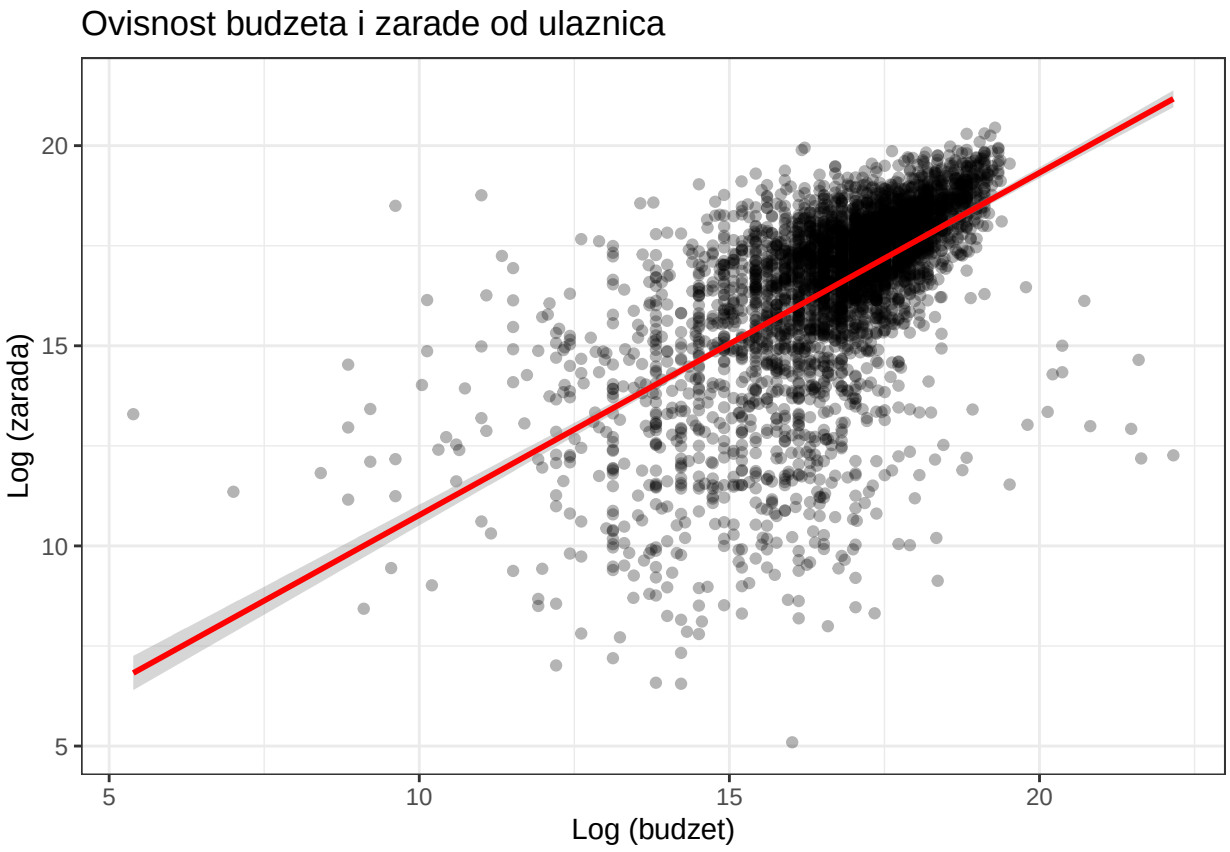
Iako je u ovom slučaju koeficijent korelacije veći nego kod budžeta, i dalje ne možemo zaključiti postojanje linearne ovisnosti.

Ovisnost budžeta i zarade od ulaznica

Pogledajmo koeficijent korelacije te graf ovisnosti logaritamskih vrijednosti budžeta filma i zarade od prodaje ulaznica.

```
movies_money <- movies_gross[!is.na(movies_gross$budget),]

ggplot(movies_money, aes(x = log(budget + 1), y = log(gross + 1))) +
  geom_point(alpha = 0.3) +
  geom_smooth(formula = y ~ x, method = "lm", color = "red") +
  labs(title = "Ovisnost budzeta i zarade od ulaznica",
       x = "Log (budzet)",
       y = "Log (zarada)") + theme_bw()
```




```
cor(log(movies_money$budget + 1),log(movies_money$gross + 1) )
```

```
## [1] 0.5887431
```

Ono što slutimo po koeficijentu korelacije, a graf nam dodatno potvrđuje je da opravdano sumnjamo na postojanje linearna ovisnosti između dvije varijable. Vidimo da su točke na grafu dosta raspršene za male vrijednosti, ali za visoke iznose budžeta i zarade ovisnost se više naslućuje s određenim odstupanjima. Pokušajmo potvrditi ovu ovisnost linearnim modelom.

Linearni model zarade i budžeta

Statistički istražimo postoji li formula koja opisuje zaradu od ulaznica preko budžeta: `zarada` = koeficijent * `budžet`.

```
linMod <- lm(log(gross + 1) ~ log(budget + 1), data = movies_money)

summary(linMod)
```

```
##
## Call:
## lm(formula = log(gross + 1) ~ log(budget + 1), data = movies_money)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

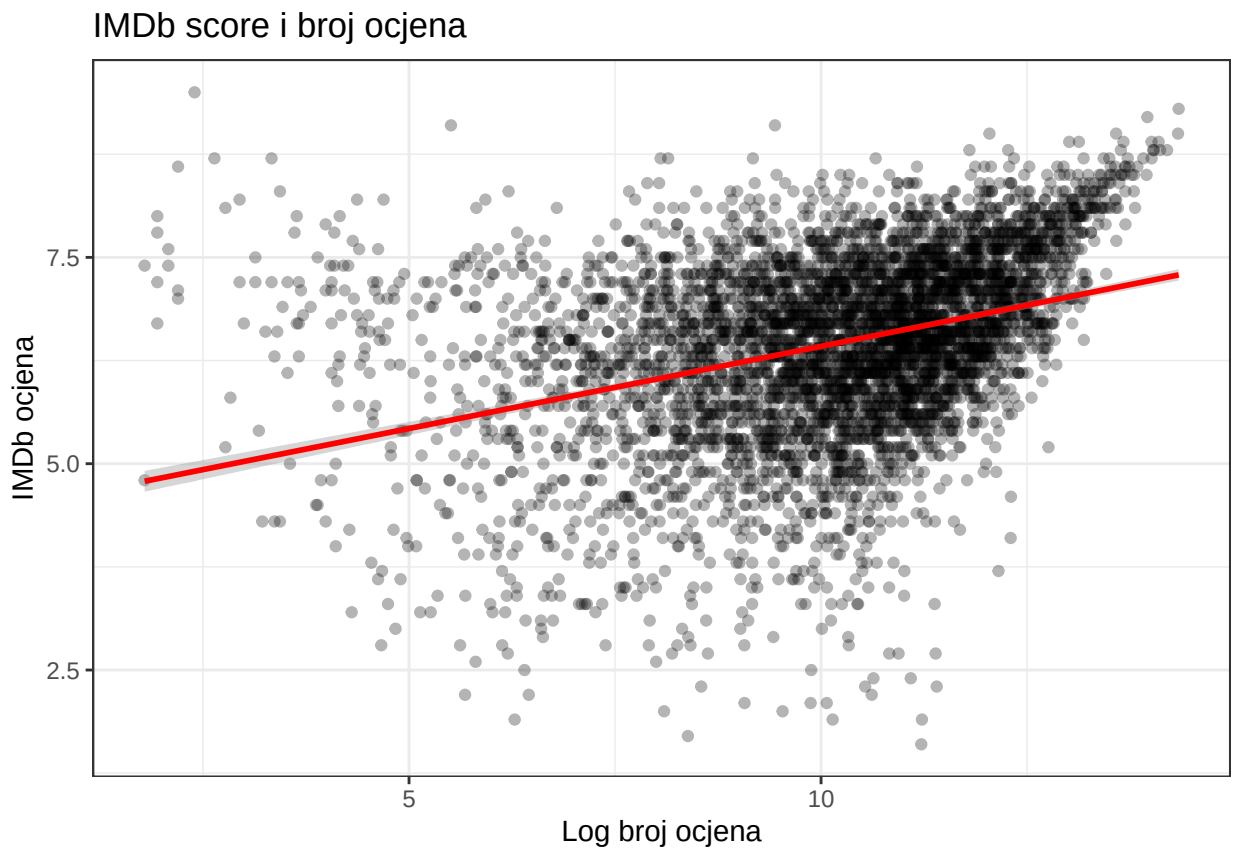
```
## -10.8192  -0.5577   0.3506   1.0453   8.0560
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.21500    0.32025   6.917 5.41e-12 ***
## log(budget + 1)  0.85544    0.01909  44.822  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.811 on 3787 degrees of freedom
## Multiple R-squared:  0.3466, Adjusted R-squared:  0.3464
## F-statistic:  2009 on 1 and 3787 DF,  p-value: < 2.2e-16
```

Možemo vidjeti da postoji statistički značajna ovisnost između dvije logaritmirane varijable, ali R-squared = 0.35 pokazuje da model objašnjava samo trećinu varijance, što upućuje na samo umjerenu prediktivnu moć modela zbog prevelike raspršenosti podataka.

Broj ocjena korisnika

Sljedeća varijabla koju ćemo istražiti je `num_voted_users` koja nam govori koliko korisnika IMDb stranice je dalo svoju ocjenu za film. Ovaj podatak nam može reći koliki doseg je film imao, ali moramo imati na umu da većina ljudi samo pogleda film a da ga ocijeni na internetu.

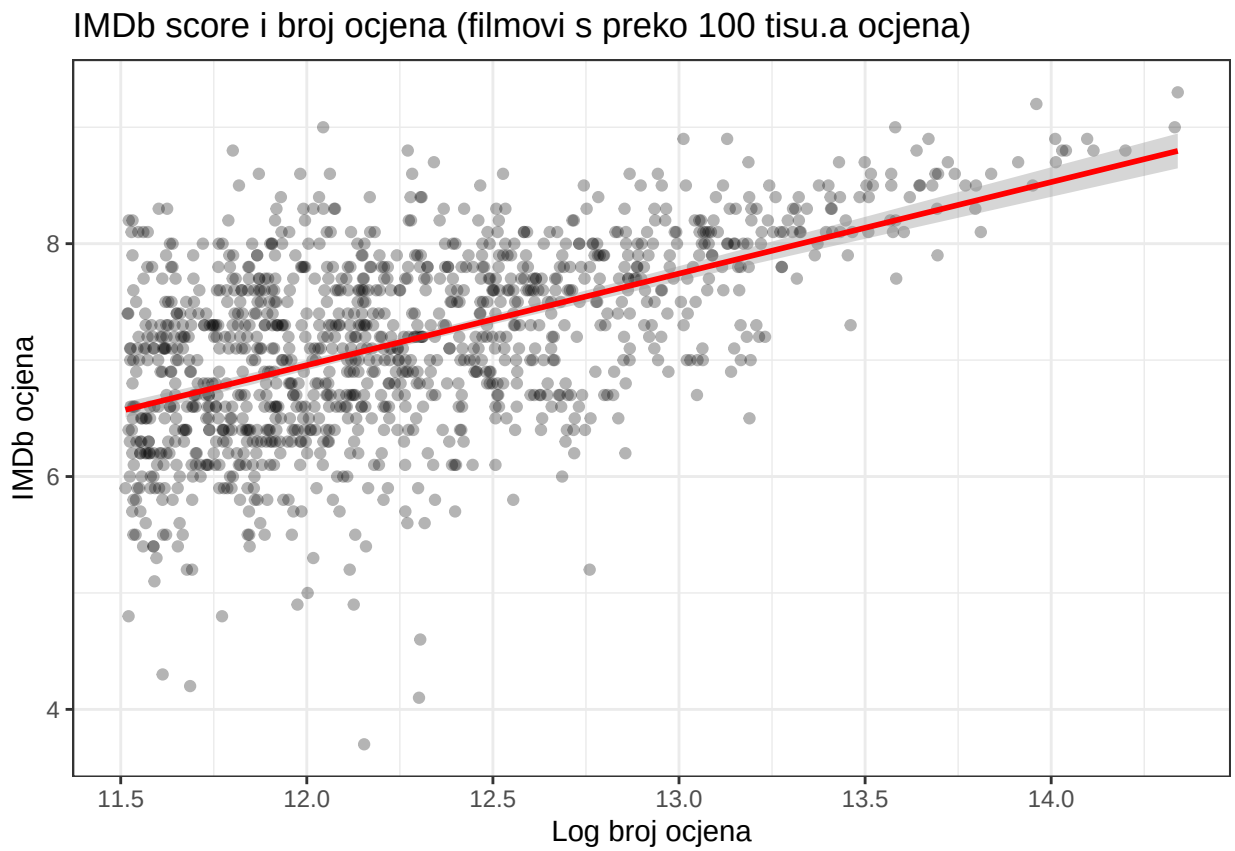
```r
movies_glasovi <- movies[!is.na(movies$num_voted_users),]
ggplot(movies_glasovi, aes(x = log(num_voted_users + 1), y = imdb_score)) +
  geom_point(alpha = 0.3) +
  geom_smooth(formula = y ~ x, method = "lm", color = "red") +
  labs(title = "IMDb score i broj ocjena",
       x = "Log broj ocjena",
       y = "IMDb ocjena") + theme_bw()
```




```
koef_korelacija_prije <- cor(movies_glasovi$num_voted_users, movies_glasovi$imdb_score)
```

Na ovom grafu vidimo veliku raspršenost ocjena za malen broj ocjenjivača. Zanimljivo je pogledati kako se podaci "nakupljaju", odnosno grupiraju za sve veći broj ocjena. Iz ovog razloga pogledat ćemo još jedan ovakav graf, ali samo za filmove koje je više od 100 tisuća ljudi ocijenilo. Ovime ćemo probati pokazati postoji li tendencija da što je film "bolji" to će više ljudi ostaviti svoju ocjenu za njega. Naravno treba biti oprezan na to da je dobra ocjena jedan od mogućih razloga zašto neki ljudi uopće saznaju za neke filmove i odluče ih pogledati, stoga veća gledanost može rezultirati i većim brojem glasova, ali nažalost podatak o gledanosti nemamo u našem podatkovnom okviru tako da ne možemo taj utjecaj potvrditi u ovom trenutku.

```
movies_glasovi <- movies_glasovi[movies_glasovi$num_voted_users > 100000,]
ggplot(movies_glasovi, aes(x = log(num_voted_users + 1), y = imdb_score)) +
  geom_point(alpha = 0.3) +
  geom_smooth(formula = y ~ x, method = "lm", color = "red") +
  labs(title = "IMDb score i broj ocjena (filmovi s preko 100 tisuća ocjena)",
       x = "Log broj ocjena",
       y = "IMDb ocjena") + theme_bw()
```




```
koef_korelacija_poslije <- cor(movies_glasovi$num_voted_users, movies_glasovi$imdb_score)
```

Na ovom grafu vidimo jaču linearnu vezu, provjerimo s koeficijentom korelacije koliko su varijable zaista povezanije nakon odabiranja češće ocjenjivanih filmova.

```
koef_korelacija_prije
```

```
## [1] 0.4123669
```

```
koef_korelacija_poslije
```

```
## [1] 0.5305801
```

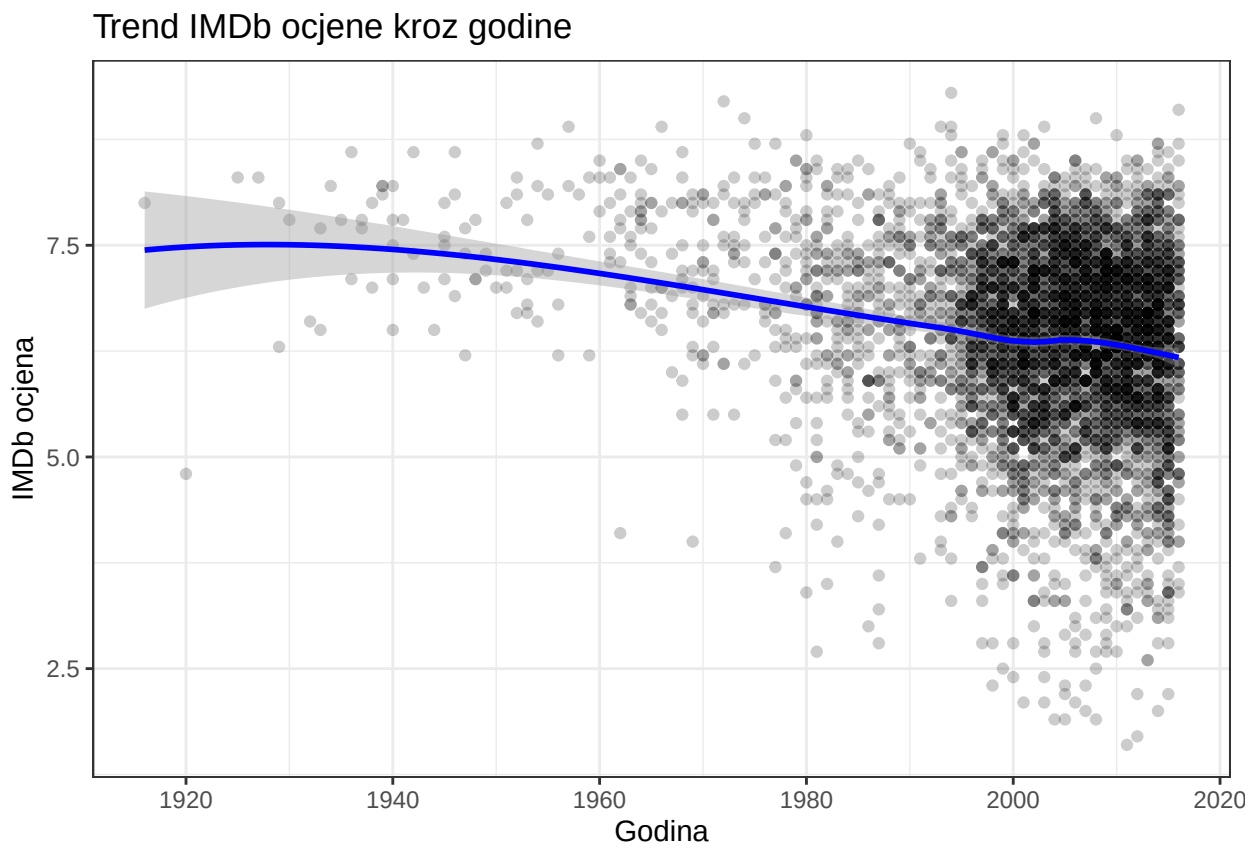
Kao što smo i pretpostavili, koeficijent korelacije se povećao.

Godina izlaska filma

Uz pomoć varijable `title_year` promotrimo trend IMDb ocjena po godinama izlaska filma.

```
movies_trend <- movies[!is.na(movies$title_year), ]
ggplot(movies_trend, aes(x = title_year, y = imdb_score)) +
  geom_point(alpha = 0.2) +
  geom_smooth(formula = y ~ x, method = "loess", color = "blue") +
```

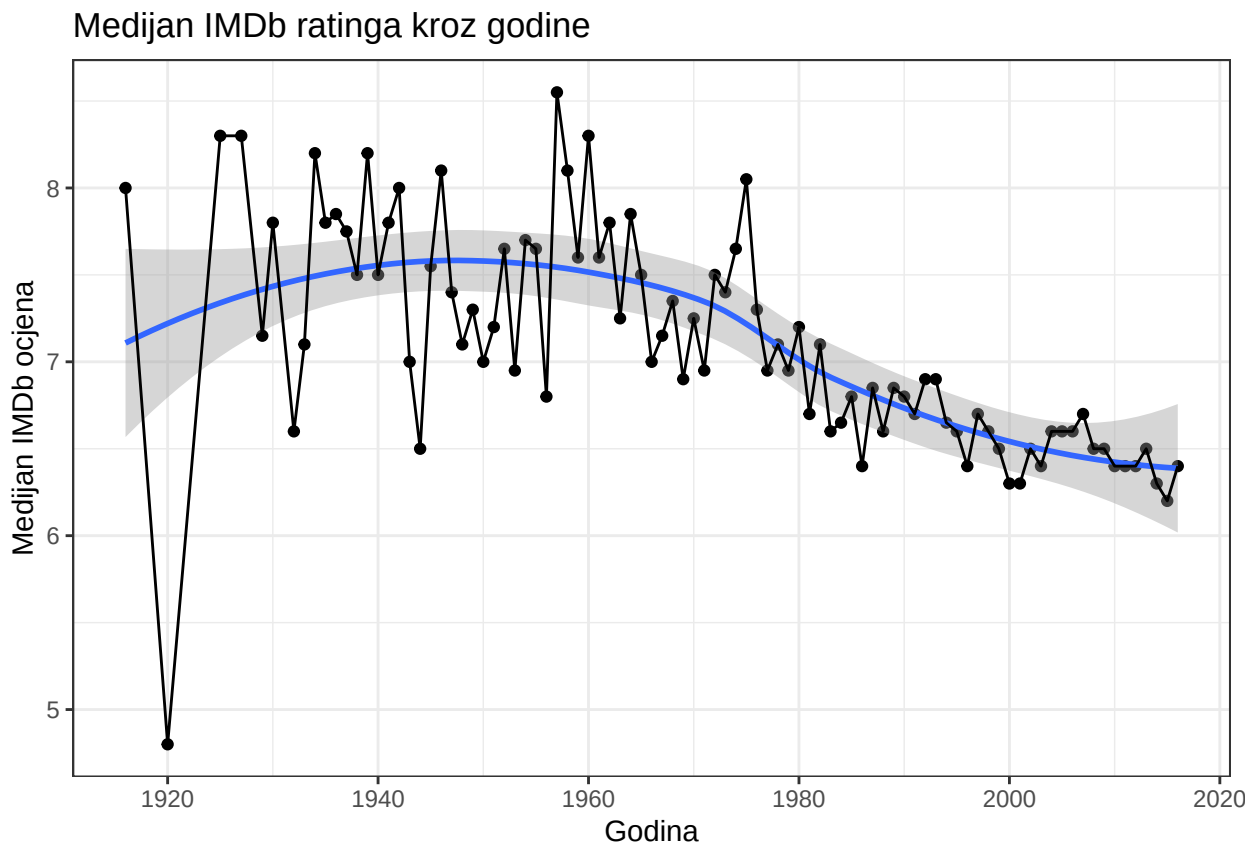
```
  labs(title = "Trend IMDb ocjene kroz godine",
       x = "Godina",
       y = "IMDb ocjena") + theme_bw()
```



Ovaj graf nam pokazuje IMDb ocijene za filmove prema godinama izlaska. Možemo vidjeti kako je broj ocijenjenih filmova znatno veći za razdoblje nakon 1990. godine, a razlog tomu je podatak da je IMDb stranica nastala upravo 1990. godine, a kupljena od strane Amazona 1998. godine što joj je pridonijelo još većoj popularnosti. Promotrimo sada medijan ocjena za svaku zabilježenu godinu filma. Za ovaj graf uzimamo medijan jer je robusniji na stršeće vrijednosti.

```
movies_trend <- movies_trend %>% group_by(title_year) %>%
  summarise(median_rating = median(imdb_score))

ggplot(movies_trend, aes(x = title_year, y = median_rating)) + geom_point() +
  geom_smooth(formula = y ~ x, method = "loess") + geom_line(group = 1) +
  labs(title = "Medijan IMDb ratinga kroz godine",
       x = "Godina",
       y = "Medijan IMDb ocjena") + theme_bw()
```

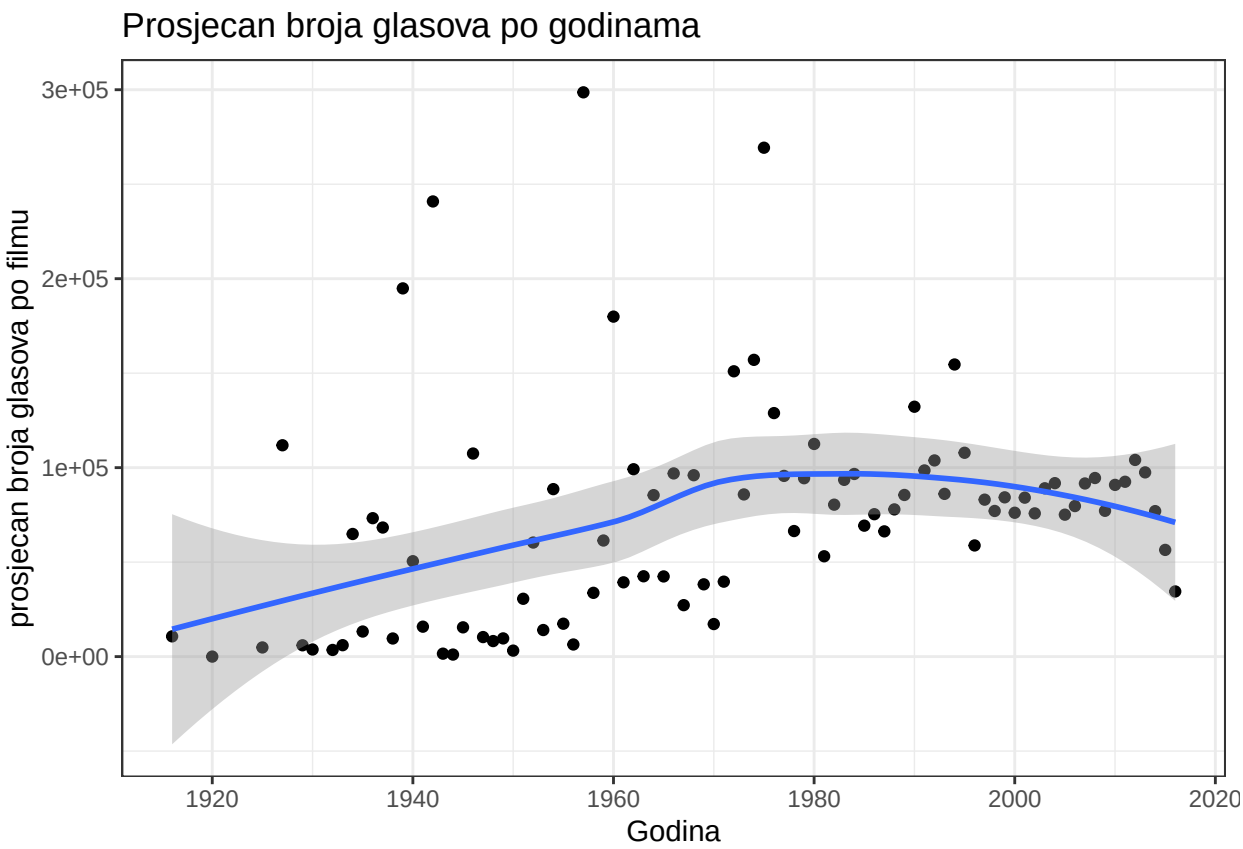


Ovaj graf daje nam par zanimljivih informacija. 2 najlošija medijana u 20.stoljeću dogodila su se u poslijeratnom periodu oba svjetska rata. Trend nam pokazuje da su ocjene novijih filmova lošije od ocjena starijih. Jedan od mogućih objašnjenja ovog trenda je da korisnici starije filmove češće gledaju jer su već čuli da je film jako dobar ili su vidjeli da ima jako dobru ocjenu pa su ga odlučili pogledat. Vjerojatno se radi o filmskim klasicima kojima se ljudi rado vraćaju i ponovo gledaju. Nakon osnutka IMDb-a korisnici su počeli ocjenjivati filmove sve češće, odnosno ocjenjivali bi filmove nakon što ih prvi puta pogledaju. Sa sve više filmova, došlo je i sve više lošije ocijenjenih filmova zbog kojih je median niži u 21.stoljeću.

Istražimo sada kako se prosječan broj glasova mijenjao po godinama.

```
movies_trend2 <- movies[!is.na(movies$num_voted_users)
                        & !is.na(movies$title_year), ] %>%
  group_by(title_year) %>% summarise(avg_votes = (mean(num_voted_users)))

ggplot(movies_trend2, aes(x = title_year, y = (avg_votes))) +
  geom_point() + geom_smooth(formula = y ~ x, method = "loess")  +
  labs(title = "Prosjecan broja glasova po godinama",
       x = "Godina",
       y = "prosjecan broja glasova po filmu") + theme_bw()
```



Uz trend koji upućuje na povećanje prosječnog broja glasova, na ovom grafu primjećujemo i par točaka koje jako odstupaju od dugih. 3 godine s najvećim prosjekom su 1942., 1957., 1975. Pogledajmo koji su najbolje ocijenjeni filmovi iz tih godina.

```
movies_trend3 <- movies[!is.na(movies$movie_title),] %>% group_by(title_year) %>% arrange(desc(imdb_sco
movies_1942 <- movies_trend3[movies_trend3$title_year == 1942, ] %>% slice_head(n = 3)
movies_1957 <- movies_trend3[movies_trend3$title_year == 1957, ] %>% slice_head(n = 3)
movies_1975 <- movies_trend3[movies_trend3$title_year == 1975, ] %>% slice_head(n = 3)

cat("Najbolje ocjenjeni filmovi iz 1942. godine: ",
paste(movies_1942$movie_title[!is.na(movies_1942$movie_title)], collapse = ", "), "\n")
cat("Najbolje ocjenjeni filmovi iz 1957. godine: ",
    paste(movies_1957$movie_title[!is.na(movies_1957$movie_title)], collapse = ", "),
     "\n")
cat("Najbolje ocjenjeni filmovi iz 1975. godine: ",
    paste(movies_1975$movie_title[!is.na(movies_1975$movie_title)], collapse = ", "),
     "\n")
```

```
## Najbolje ocjenjeni filmovi iz 1942. godine:  Casablanca , Bambi
## Najbolje ocjenjeni filmovi iz 1957. godine:  12 Angry Men , The Bridge on the River Kwai
## Najbolje ocjenjeni filmovi iz 1975. godine:  One Flew Over the Cuckoo's Nest , Monty Python and the
```

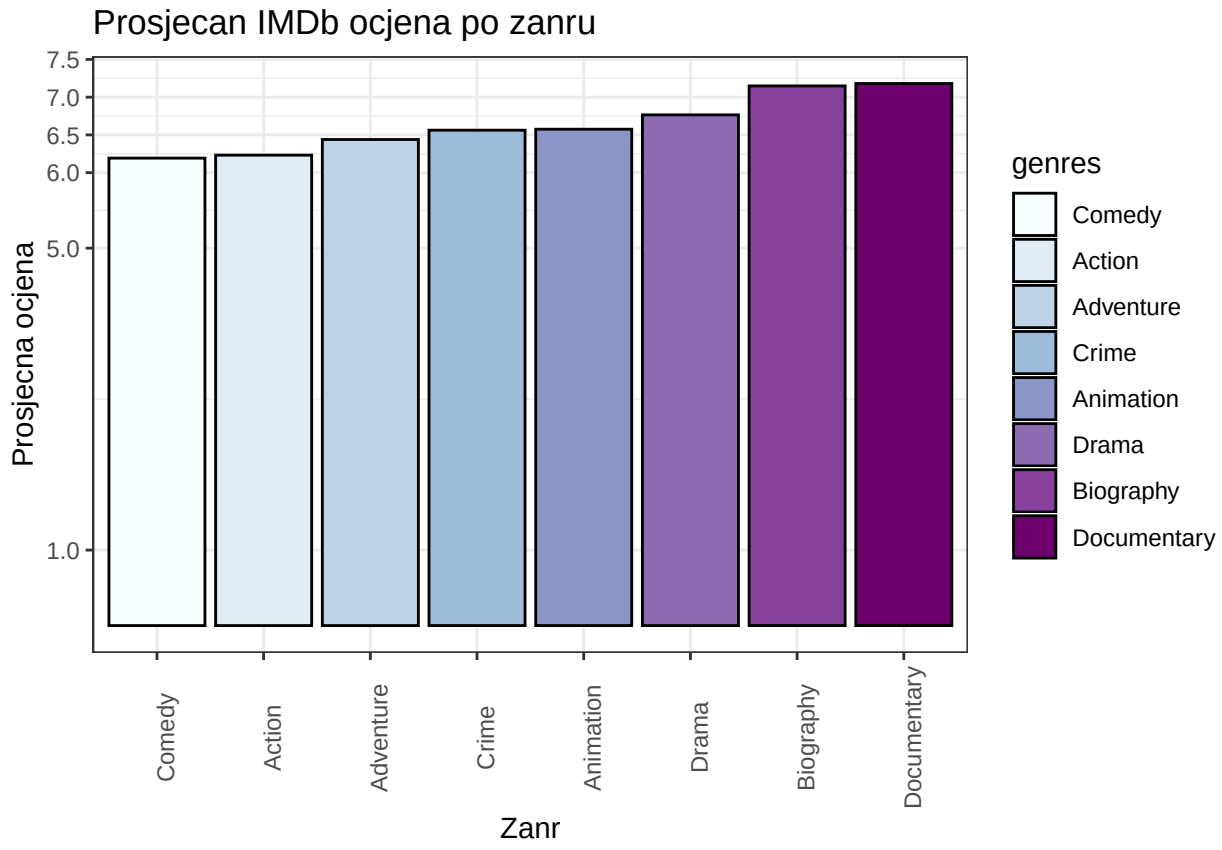
Radi se o vrlo poznatim i uspješnim filmovima koje ljudi i dan danas gledaju.

Žanr filma

Promotrimo sada varijablu `genres`. U uvodnom dijelu smo upoznali kako ovaj stupac izgleda kada smo ga pripremali za linearnu regresiju. Sada ćemo raditi nad originalnim stupcem i pogledati prosječnu ocjenu po žanru. Promatrat ćemo samo žanrove sa 100 ili više filmova.

```
genre_stats <- movies %>%
  separate_rows(genres, sep = "\\|") %>%
  group_by(genres) %>%
  summarise(avg_rating = mean(imdb_score, na.rm = TRUE), n = n()) %>%
  filter(n >= 100) %>%
  slice_head( n = 8) %>%
  mutate(genres = fct_reorder(genres, avg_rating))


ggplot(genre_stats,
       aes(x = reorder(genres, avg_rating), y = avg_rating, fill = genres)) +
  geom_col(color = "black") + scale_fill_brewer(palette = "BuPu") +
  scale_y_continuous(name ="Prosjecna ocjena",  breaks = c(1, 5, 6, 6.5, 7, 7.5, 8)) +
  labs(title = "Prosjecan IMDb ocjena po zanru",
       x = "Zanr") + theme_bw() + theme(axis.text.x = element_text(angle = 90))
```



Graf nam pokazuje kako je prosjek ocjena najveći za dokumentarce, biografije, povijesne te ratne filmove. Filmovi tog žanra znaju biti često nagrađivani prestižnim filmskim nagradama zbog tematike kojoj se bave.

Linearni model ratinga

Pokušajmo sada odrediti linearnu funkciju koja najbolje opisuje IMDb rating. Za početak moramo ukloniti sve stupce koji nam ne trebaju, a to su svi znakovni stupci koji nisu kategorijski. Također, logaritmirati ćemo veće brojke poput zarade, budžeta i broja lajkova.

```
movies_reg <- movies_reg %>% dplyr::select(-movie_title, -director_name, -actor_1_name, -actor_2_name,

movies_reg <- movies_reg %>%
  mutate(
    log_budget = log(budget + 1),
    log_gross = log(gross + 1),
    log_votes = log(num_voted_users + 1),
    log_user_reviews = log(num_user_for_reviews + 1),
    log_cast_likes = log(cast_total_facebook_likes + 1),
    log_director_likes = log(director_facebook_likes + 1),
    log_movies_fb_likes = log(movie_facebook_likes + 1)
  ) %>%
  dplyr::select(-budget, -gross, -num_voted_users,
         -num_user_for_reviews,
         -cast_total_facebook_likes, -director_facebook_likes, -movie_facebook_likes)
```

Kako u skupu imamo puno kategorija za varijablu države i jezika filma, zbog brojnosti pojednostavit ćemo stupce tako da ćemo za `engLanguage` upisati 1 ako je film na engleskom jeziku, a 0 ako nije te za `USA_country` upisati 1 ako je američki film, a 0 ako nije. Za kraj ćemo obrisati sve nedostajuće vrijednosti bi svi modeli radili nad istim podacima.

```
movies_reg <- movies_reg %>%
  mutate(
    engLanguage = if_else(language == "English", 1L, 0L),
    USA_country = if_else(country == "USA", 1L, 0L)
  ) %>%
  dplyr::select(-language, -country)

movies_reg <- movies_reg %>% drop_na()
```

Iterativna (stepwise) izgradnja prediktivnog modela

Prvo ćemo stvoriti jedan linearni model koji sadrži sve varijable iz tablice `movies_reg` te jedan potpuno prazni linearni model. Uz pomoć funkcije `stepAIC` stvorit ćemo nova 2 modela. Jedan od njih (`lm1`) će nastati iterativnom selekcijom prediktora od punog modela, dok će drugi (`lm2`) od potpuno praznog.

```
lm_sve <- lm(imdb_score ~ ., data = movies_reg)
lm_prazan <- lm(imdb_score ~ 1, data = movies_reg)

lm1 <- stepAIC(lm_sve, direction = "backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan), direction = "forward", trace
```

Vrijednost koja nam govori o tome koliko je model prediktivan i koliko varijance objašnjava naziva se prilagođeni koeficijent determinacije (`Adjusted R-squared`). Usporedimo vrijednosti početnog modela i 2 modela dobivena stepwise funkcijom.

```r
cat("Adjusted R-squared vrijednosti od punog modela: ", summary(lm_sve)$adj.r.squared, '\n')
cat("Adjusted R-squared vrijednosti od lm1 modela: ", summary(lm1)$adj.r.squared,'\n')
cat("Adjusted R-squared vrijednosti od lm2 modela: ", summary(lm2)$adj.r.squared,'\n')
```

```
## Adjusted R-squared vrijednosti od punog modela:  0.5680386
## Adjusted R-squared vrijednosti od lm1 modela:  0.5688987
## Adjusted R-squared vrijednosti od lm2 modela:  0.5685917
```

Usporedba prilagođenih koeficijenata determinacije pokazuje da iterativna selekcija prediktora nije dovela do značajnog poboljšanja objašnjene varijance u odnosu na puni model, što sugerira da većina varijabli doprinosi modelu ili da su njihovi učinci relativno mali.

Pogledajmo detaljnije `lm1` model.

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = imdb_score ~ color + num_critic_for_reviews + duration +
##     actor_1_facebook_likes + facenumber_in_poster + content_rating +
##     title_year + aspect_ratio + Action + Fantasy + `Sci-Fi` +
##     Thriller + Documentary + Romance + Animation + Comedy + Family +
##     Drama + Horror + Biography + Music + log_budget + log_gross +
##     log_votes + log_user_reviews + log_cast_likes + log_director_likes +
##     engLanguage + USA_country, data = movies_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2944 -0.3397  0.0464  0.4361  2.6314
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.346e+01  3.483e+00  15.349  < 2e-16 ***
## colorColor               -1.589e-01  6.652e-02  -2.388  0.01697 *
## num_critic_for_reviews    1.917e-03  1.656e-04  11.574  < 2e-16 ***
## duration                  8.033e-03  6.630e-04  12.117  < 2e-16 ***
## actor_1_facebook_likes    1.837e-06  8.992e-07   2.042  0.04118 *
## facenumber_in_poster     -1.375e-02  5.901e-03  -2.330  0.01986 *
## content_ratingG           2.778e-01  1.973e-01   1.408  0.15927
## content_ratingGP         -2.377e-01  7.168e-01  -0.332  0.74016
## content_ratingM          -4.951e-02  5.213e-01  -0.095  0.92435
## content_ratingNC-17      -3.363e-01  3.371e-01  -0.998  0.31856
## content_ratingNot Rated   2.341e-01  2.174e-01   1.076  0.28181
## content_ratingPassed     -1.490e-01  4.396e-01  -0.339  0.73477
## content_ratingPG          3.408e-01  1.854e-01   1.838  0.06608 .
## content_ratingPG-13       2.730e-01  1.879e-01   1.453  0.14623
## content_ratingR           4.027e-01  1.856e-01   2.170  0.03011 *
## content_ratingUnrated     4.345e-01  2.310e-01   1.881  0.06009 .
## content_ratingX           4.497e-01  2.926e-01   1.537  0.12439
## title_year               -2.423e-02  1.769e-03 -13.699  < 2e-16 ***
## aspect_ratio              1.041e-01  3.519e-02   2.960  0.00310 **
## Action                   -1.819e-01  3.268e-02  -5.567 2.77e-08 ***
## Fantasy                  -7.059e-02  3.778e-02  -1.868  0.06180 .
```

```
## `Sci-Fi`                -1.028e-01  3.812e-02  -2.697  0.00703 **
## Thriller                -1.642e-01  3.082e-02  -5.328 1.05e-07 ***
## Documentary              7.839e-01  1.118e-01   7.013 2.76e-12 ***
## Romance                 -9.184e-02  2.986e-02  -3.076  0.00211 **
## Animation                8.005e-01  6.647e-02  12.043  < 2e-16 ***
## Comedy                  -1.422e-01  3.144e-02  -4.522 6.33e-06 ***
## Family                  -1.031e-01  6.132e-02  -1.682  0.09264 .
## Drama                    4.270e-01  2.992e-02  14.275  < 2e-16 ***
## Horror                  -4.753e-01  4.533e-02 -10.484  < 2e-16 ***
## Biography                1.281e-01  5.038e-02   2.542  0.01105 *
## Music                   -1.296e-01  5.932e-02  -2.185  0.02897 *
## log_budget              -1.630e-01  1.195e-02 -13.642  < 2e-16 ***
## log_gross               -4.820e-02  9.066e-03  -5.317 1.12e-07 ***
## log_votes                5.008e-01  1.935e-02  25.888  < 2e-16 ***
## log_user_reviews        -2.457e-01  2.520e-02  -9.752  < 2e-16 ***
## log_cast_likes          -1.786e-02  1.101e-02  -1.622  0.10482
## log_director_likes       1.039e-02  4.904e-03   2.119  0.03413 *
## engLanguage             -3.992e-01  6.565e-02  -6.080 1.32e-09 ***
## USA_country             -1.440e-01  3.210e-02  -4.486 7.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6953 on 3629 degrees of freedom
## Multiple R-squared:  0.5735, Adjusted R-squared:  0.5689
## F-statistic: 125.1 on 39 and 3629 DF,  p-value: < 2.2e-16
```

Iz ispisa možemo vidjeti koje su varijable statistički značajno (***) povezane s IMDb ocjenom. Neke od njih koje najviše pridonose boljoj ocjeni su: `num_critic_for_reviews`, `duration` i `Animation`. A neke koje najviše negativno utječu su: `title_year`, `log_gross` i `Horror`.

Valja napomenuti da statistička značajnost pojedinih varijabli ne implicira nužno uzročnu povezanost, već isključivo njihovu povezanost s IMDb ocjenom unutar promatranog modela. Linearni model opisuje 57 % varijance, što je statistički nedovoljno da proglasimo naš model jako dobrim prediktorom budućih IMDb ocjena. Ovdje ostavljamo mjesta za napredak budućim analizama da uz pomoć drugih transformacija i metoda prediktivnih modela poboljšaju prediktivni model za izračun ocjene.

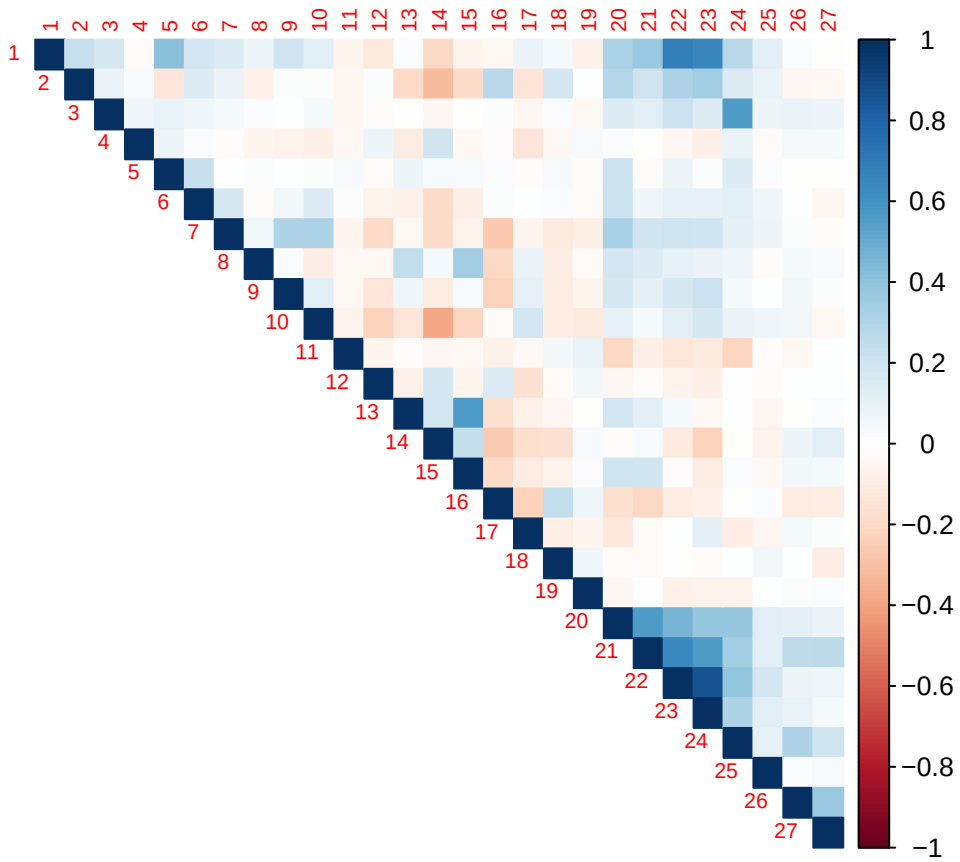
Kolinearnost ulaznih varijabli

Sada ćemo provjeriti u kojoj mjeri su ulazni podaci međusobno korelirani.

```
moviesNumInputs <- movies_reg %>% dplyr::select(color, num_critic_for_reviews, duration,
    actor_1_facebook_likes ,facenumber_in_poster , content_rating ,
    title_year , aspect_ratio , Action , Fantasy , `Sci-Fi` ,
    Thriller , Documentary , Romance , Animation , Comedy , Family ,
    Drama , Horror , Biography , Music , log_budget , log_gross ,
    log_votes , log_user_reviews , log_cast_likes , log_director_likes ,
    engLanguage , USA_country) %>% select_if(is.numeric)

corMatrix <- moviesNumInputs %>% cor () %>% round(digits=2)
dimnames(corMatrix) <- NULL


corrplot(corMatrix, method = "color", type = "upper", tl.cex = 0.7)
```



Većina parova varijabli u korelacijskoj matrici ne pokazuju izrazito visoke korelacijske vrijednosti (vrijednosti su približne 0), što upućuje na nisku razinu linearne povezanosti među prediktorima.

Multikolinearnost

Kako bi provjerili odsutnost multikolinearnosti, koja je jedna od ključnih pretpostavki višestruke linearne regresije koristit ćemo faktor inflacije varijance (VIF).

```
vif(lm1)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## color                  1.070928  1        1.034857
## num_critic_for_reviews 3.110206  1        1.763577
## duration               1.716062  1        1.309986
## actor_1_facebook_likes 1.490590  1        1.220897
## facenumber_in_poster   1.115359  1        1.056106
## content_rating         5.185578 11        1.077682
## title_year             2.316379  1        1.521966
## aspect_ratio           1.185572  1        1.088840
## Action                 1.528086  1        1.236158
## Fantasy                1.247393  1        1.116867
## 'Sci-Fi'               1.249956  1        1.118014
## Thriller               1.494641  1        1.222555
## Documentary            1.199000  1        1.094989
## Romance                1.194442  1        1.092905
```

```
## Animation              1.679216  1        1.295846
## Comedy                 1.790180  1        1.337976
## Family                 2.964742  1        1.721843
## Drama                  1.697898  1        1.303034
## Horror                 1.424620  1        1.193575
## Biography              1.173306  1        1.083193
## Music                  1.047086  1        1.023272
## log_budget             2.342406  1        1.530492
## log_gross              2.750567  1        1.658483
## log_votes              5.911645  1        2.431388
## log_user_reviews       4.949228  1        2.224686
## log_cast_likes         2.059655  1        1.435150
## log_director_likes     1.059592  1        1.029365
## engLanguage            1.339914  1        1.157547
## USA_country            1.276863  1        1.129984
```

Analiza multikolinearnosti pokazala je da su sve VIF vrijednosti manje od 2.5, čime je potvrđeno da multikolinearnost ne predstavlja problem u promatranom regresijskom modelu.

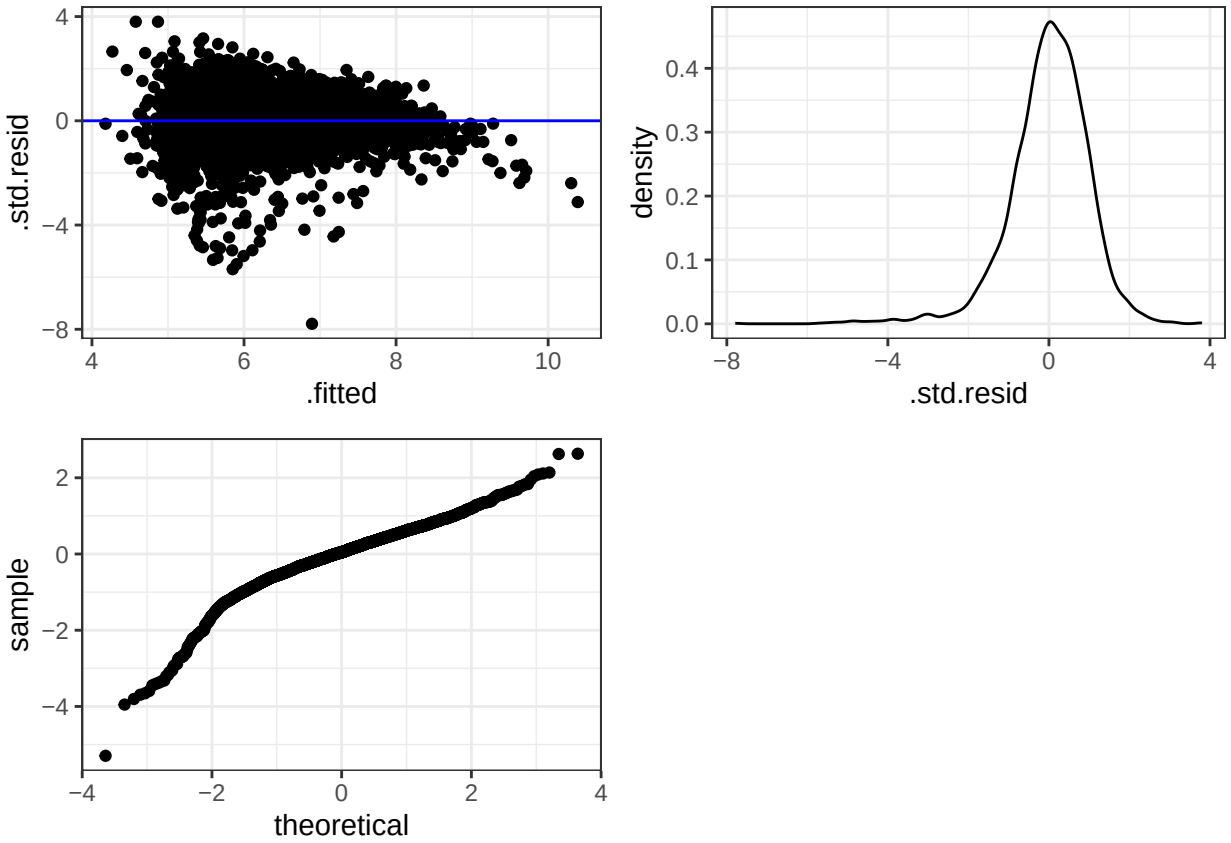
Provjera normalnosti reziduala

Za kraj ćemo provjeriti pretpostavke višestrukog regresijskog modela vezane uz reziduale. Prikazat ćemo odnos standardiziranih reziduala i predviđenih vrijednosti, gustoću razdiobe standardiziranih reziduala te kvantil-kvantil (Q–Q) graf reziduala.

```
predikcije <- augment(lm1)
g1 <- ggplot(predikcije, aes(x = .fitted, y = .std.resid)) + geom_point()+ geom_hline(yintercept = 0, c

g2 <- ggplot(predikcije, aes(x = .std.resid)) + geom_density() + theme_bw()

g3 <- ggplot(predikcije, aes(sample = .resid)) + geom_qq() + theme_bw()
grid.arrange(g1, g2, g3, ncol= 2)
```



Vizualni pregledi reziduala ukazuju na to da su pretpostavke normalnosti reziduala u velikoj mjeri zadovoljene. Gustoća razdiobe standardiziranih reziduala približno je simetrična, dok Q–Q graf pokazuje blaga odstupanja u repovima, što je očekivano kod većih uzoraka te ne predstavlja ozbiljno kršenje pretpostavke normalnosti.

Zaključak

U ovoj analizi smo se pokušali, prije svega, što bolje upoznati sa skupom podataka i iz njega izvući neke zanimljive činjenice o filmovima objavljenima do 2016. godine. Također, istraživali smo kakav utjecaj pojedine varijable imaju na IMDb ocjenu filma uz pomoć koeficijenata korelacije, vizualizacija i prediktivnog linearnog modela koji objašnjava približno 57 % varijance uzorka. Kao mjesto za napredak i budući rad smatram učitavanje novijih filmova u podatkovni skup te treniranje naprednijih prediktivnih modela kao i igranje sa samim parametrima modela.