OSP - Analiza podatkovnog skupa IMDb movie dataset

0035235027 Adam Vuković & 0036542276 Ivan Zeba

2024-01-26

Uvod

U sklopu predmeta Osnove statističkog programiranja provodimo projekt u kojem analiziramo podatkovni skup "IMDb movie dataset". U ovom projektu obrađujemo i analiziramo podatkovni skup, u kojem se nalaze informacije o filmovima sa portala IMDB, sa podatcima od 1916. do 2016. godine. Neke od dostupnih varijabli su: movie_title, title_year, director_name, duration, gross, budget itd. Pokušat ćemo prikazati dostupne podatke na što zanimljiviji i korisniji način.

Učitavanje i prilagodba podataka

Učitavanje podataka

Podatkovni skup "IMDb movie dataset" je pohranjen u .csv datoteci tako da smo za otvaranje datoteke koristili funkciju read_csv iz paketa {readr}. Neobrađeni skup imao je 5043 redaka i 28 stupaca. Prilikom istraživanja skupa istražili smo postotke nedostajućih vrijednosti za svaku varijablu.

```
## Postotak nedostajućih vrijednosti za varijablu  color :  0.3767599 %
## Postotak nedostajućih vrijednosti za varijablu  director_name :  2.062265 %
## Postotak nedostajućih vrijednosti za varijablu  num_critic_for_reviews :  0.9914733 %
## Postotak nedostajućih vrijednosti za varijablu  duration :  0.297442 %
## Postotak nedostajućih vrijednosti za varijablu  director_facebook_likes :  2.062265 %
## Postotak nedostajućih vrijednosti za varijablu  actor_3_facebook_likes :  0.4560777 %
## Postotak nedostajućih vrijednosti za varijablu  actor_2_name :  0.2577831 %
## Postotak nedostajućih vrijednosti za varijablu  actor_1_facebook_likes :  0.1388063 %
## Postotak nedostajućih vrijednosti za varijablu  gross :  17.52925 %
## Postotak nedostajućih vrijednosti za varijablu  actor_1_name :  0.1388063 %
## Postotak nedostajućih vrijednosti za varijablu  actor_3_name :  0.4560777 %
## Postotak nedostajućih vrijednosti za varijablu  facenumber_in_poster :  0.2577831 %
## Postotak nedostajućih vrijednosti za varijablu  plot_keywords :  3.033908 %
## Postotak nedostajućih vrijednosti za varijablu  num_user_for_reviews :  0.4164188 %
## Postotak nedostajućih vrijednosti za varijablu  language :  0.2379536 %
## Postotak nedostajućih vrijednosti za varijablu  country :  0.09914733 %
## Postotak nedostajućih vrijednosti za varijablu  content_rating :  6.008328 %
## Postotak nedostajućih vrijednosti za varijablu  budget :  9.756098 %
## Postotak nedostajućih vrijednosti za varijablu  title_year :  2.141582 %
## Postotak nedostajućih vrijednosti za varijablu  actor_2_facebook_likes :  0.2577831 %
## Postotak nedostajućih vrijednosti za varijablu  aspect_ratio :  6.523895 %
```

Čišćenje podataka

Kod čišćenja podataka smo obrisali duplikate filmova, obrisali smo filmove koji nemaju title_year jer smo zaključili da su to zapravo serije. Također smo izbrisali stupce koje nismo koristili ili stupce s velikim postotkom nedostajućih vrijednosti (movie_imdb_link, facenumber_in_poster, actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes i aspect_ratio)

Prilagođavanje podataka

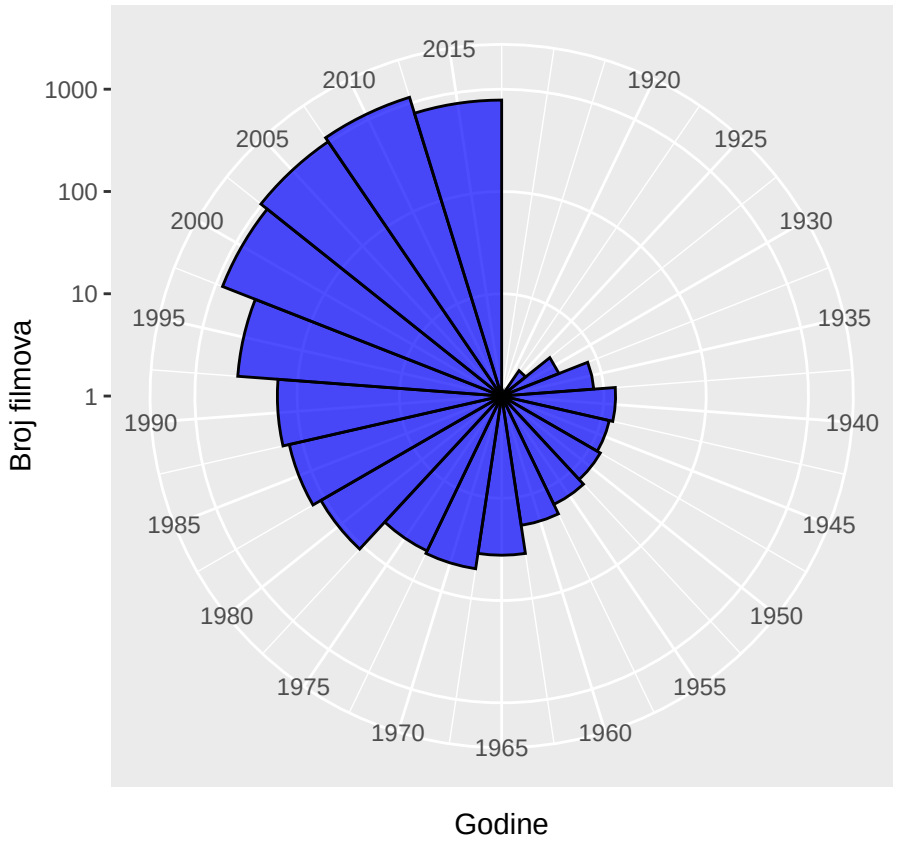
Kako bismo što bolje prikazali podatke, dodali smo našem podatkovnom skupu još jedan stupac "decade", npr. 2000.-2009. -> 2000s (iako desetljeće počinje prve godine).

```
data$decade <- ifelse(!is.na(data$title_year), paste0(data$title_year %/% 10 * 10, "s"), NA)
```

Deskriptivna analiza

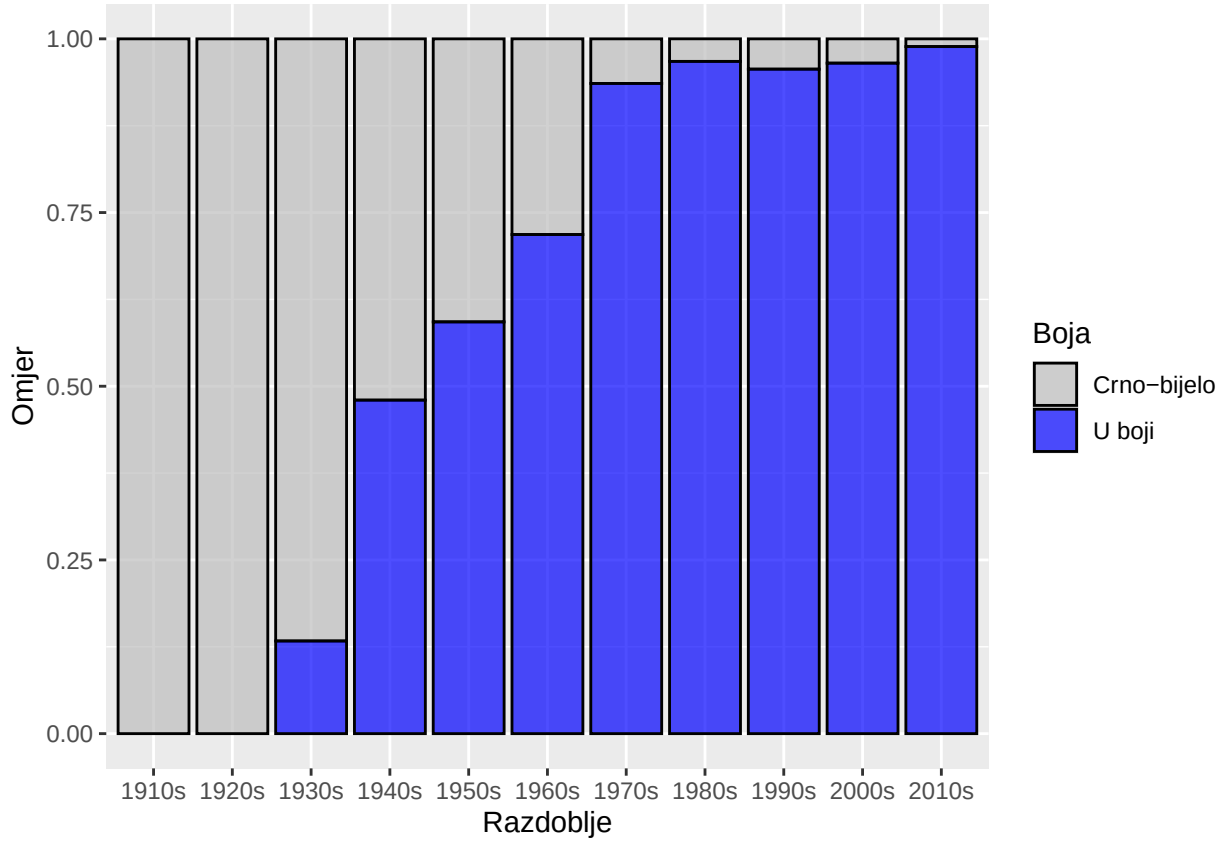
Filmovi kroz godine

Broj filmova kroz godine



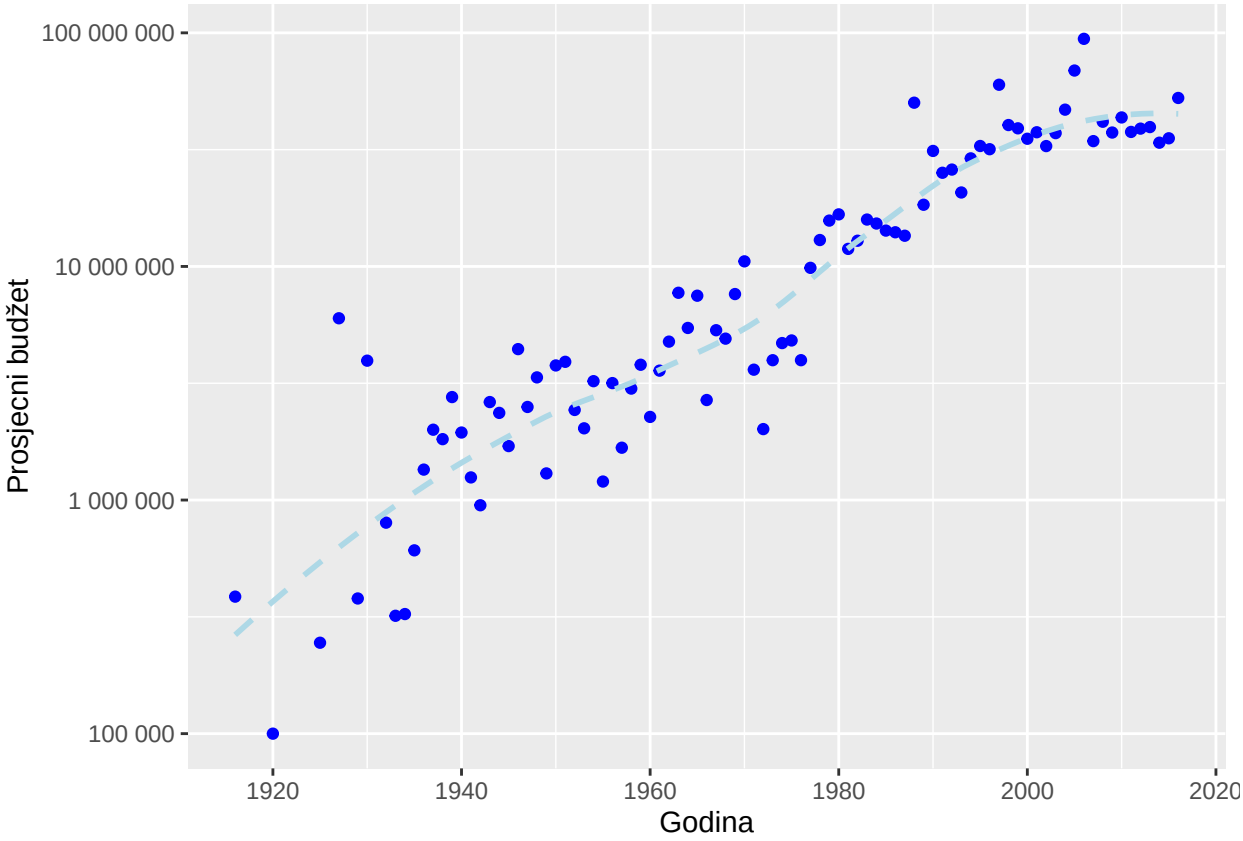
Graf prikazuje trend porasta broja filmova po godinama smještenih u ladice od veličine 5 godina. Zadnja ladica je anomalija, ali zamo zato što ta ladica ne obuhvaća svih 5 godina.

Omjer filmova u boji i crno-bijelih filmova kroz desetljeća



Graf nam pokazuje omjer filmova u boji i crno-bijelih filmova kroz desetljeća do danas. Prema dostupnim podatcima prvi film u boji je nastao prije najstarijeg filma u našem podatkovnom skupu. Tako da na grafu možemo vidjeti kako je 40-ih godina došlo do nagle ekspanzije filmova u boji te je do 70-ih godina film u boji postao sasvim uobičajena pojava. U zadnjih 10 godina crno-bijeli film je rijetka pojava.

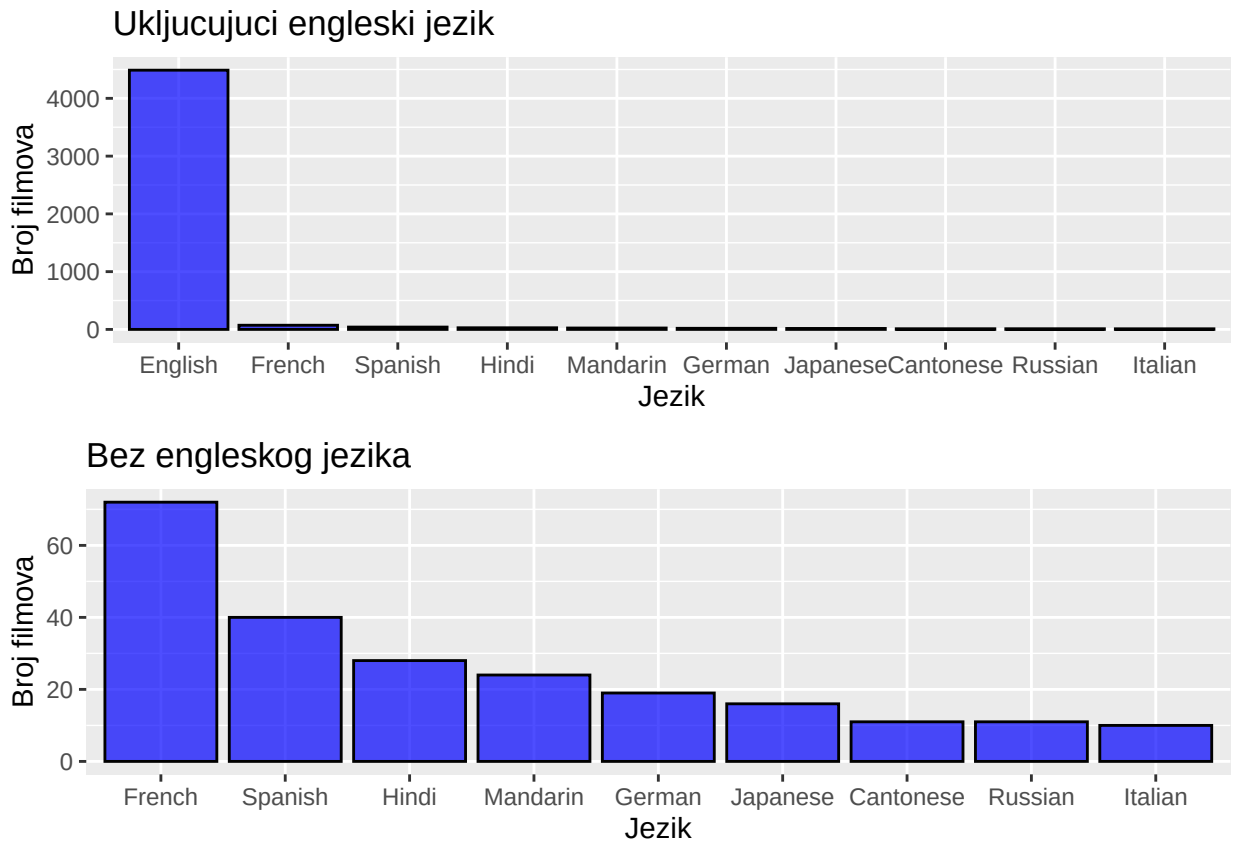
Prosječni budžet filmova kroz godine



Na grafu možemo primjeniti veliki porast prosječnog budžeta filmova kroz godine. Kako bismo bolje pokazali o kakvom se rastu radi, y os smo morali logaritamski skalirati.

Jezici

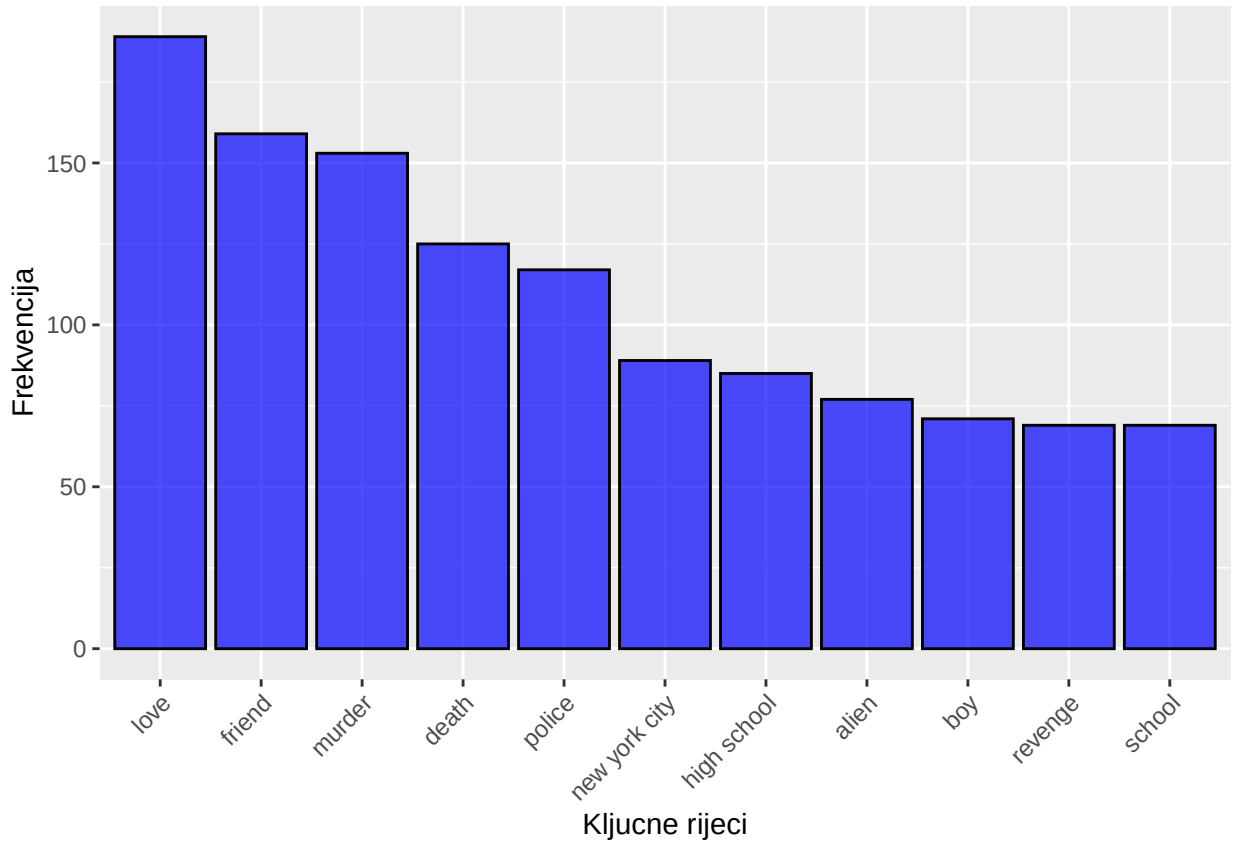
Razdioba govornih jezika u filmovima



U prvome grafu možemo vidjeti kako u našem podatkovnom skupu ima daleko najviše filmova na engleskom jeziku. Kako je broj filmova na engleskom jeziku toliko veći od ostalih, dodan je još jedan graf koji prikazuje razdiobu filmova koji nisu na engleskom jeziku. Može se zaključiti da je razdioba relativno sukladna s brojem ljudi u svijetu koji priča taj jezik.

Ključne riječi

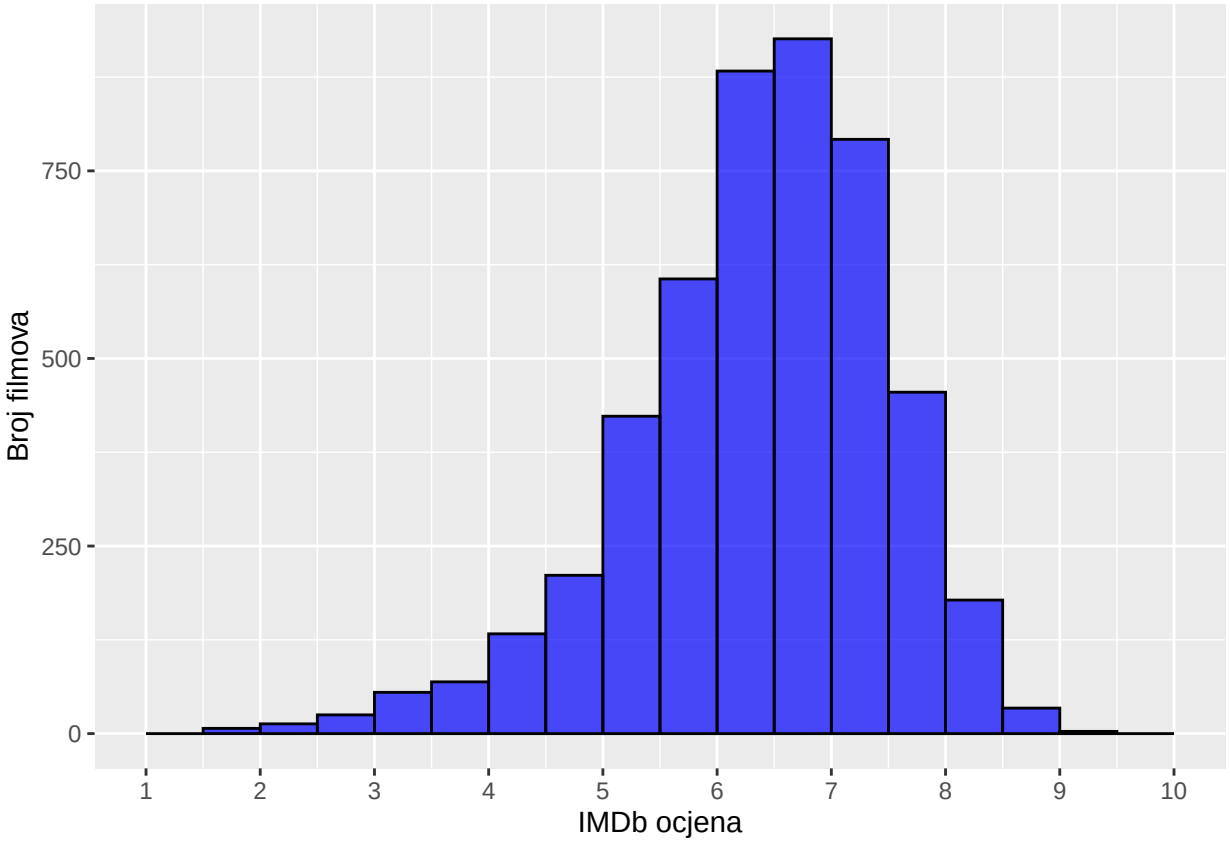
Top 10 najčešćih ključnih riječi



Iz ovog grafa možemo iščitati glavne motive filmova i neke norme industrije. Vidimo da studiji i scenaristi najviše vole raditi filmove o svakakvim temama, bilo to o ljubavi ili o ubojstvu (ili oboje). Također možemo vidjeti da je policija česti lik u filmovima te da radnja često bude smještena u New Yorku, što je zanimljivo s obzirom da je Hollywood blizu Los Angelesa koji se ne nalazi na listi.

IMDb ocjene filmova

Razdioba IMDb ocjena

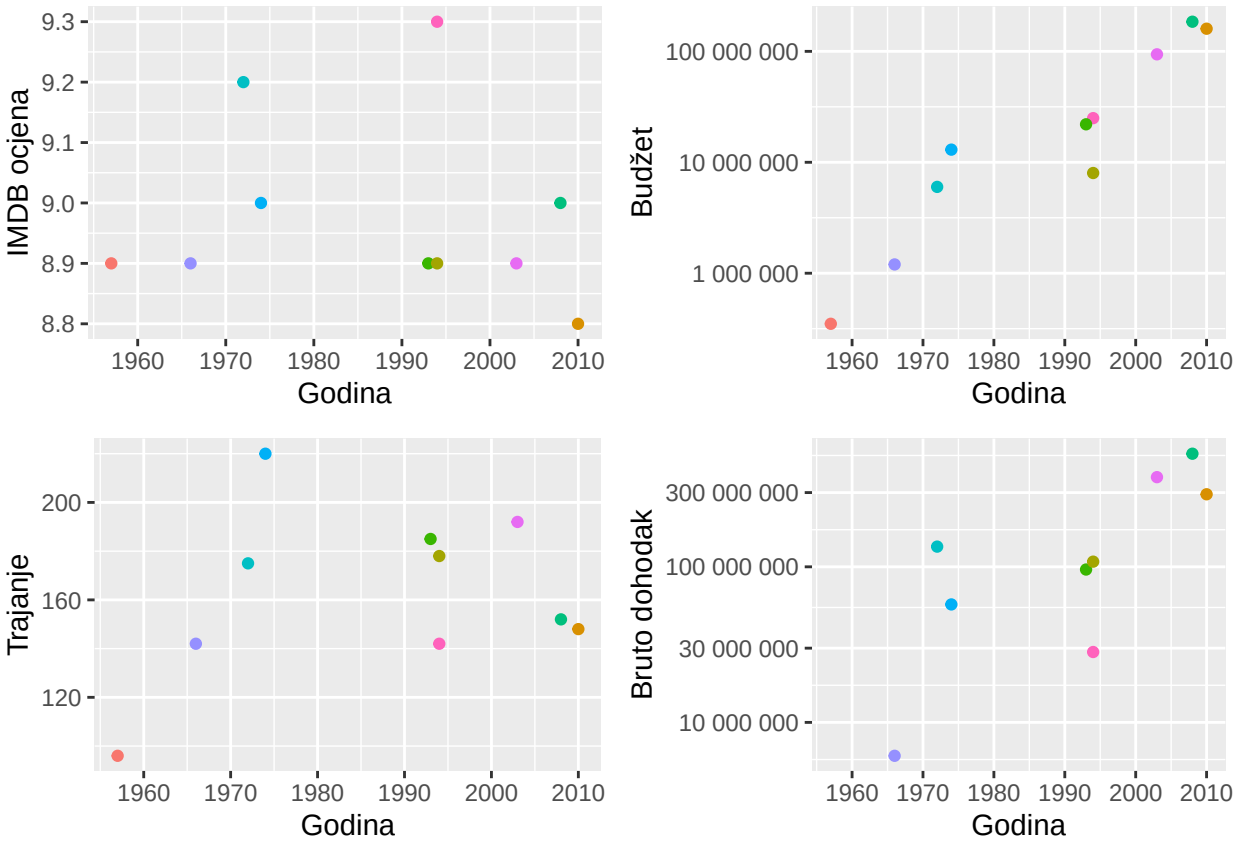


Prvo, pogledajmo razdiobu ocjena filmova. Vidimo da graf prati normalnu razdiobu, ali pomaknutu u desno. Očekivano bi bilo da ima najviše ocjena oko 5.5 (pošto je najniža ocjena 1, a ne 0), ali vidimo da su IMDb korisnici općenito blaži u ocjenjivanju filmova tako da je najviše ocjena između 6 i 7.

Top 10 filmova - ocjene i ostalo

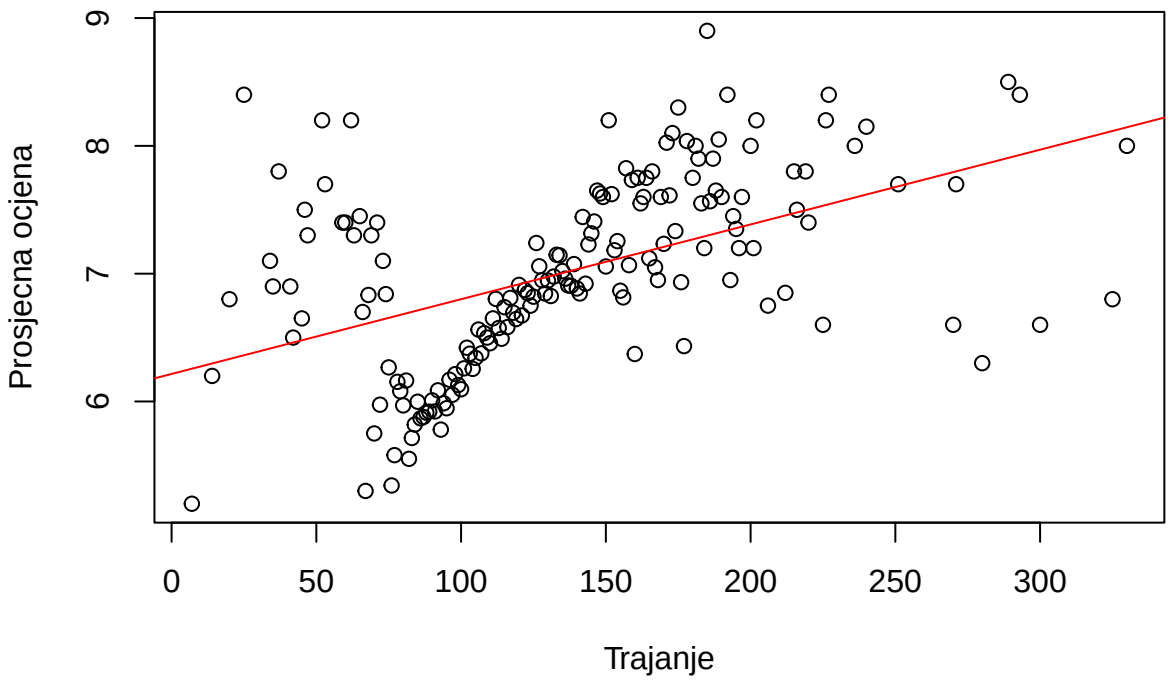
Table 1: Top 10 filmova

Naslov	Godina	Ocjena	Budzet	Trajanje	Bruto_dohodak
The Shawshank Redemption	1994	9.3	$25,000,000	142	$28,341,469
The Godfather	1972	9.2	$6,000,000	175	$134,821,952
The Dark Knight	2008	9.0	$185,000,000	152	$533,316,061
The Godfather: Part II	1974	9.0	$13,000,000	220	$57,300,000
The Lord of the Rings: The Return of the King	2003	8.9	$94,000,000	192	$377,019,252
Schindler's List	1993	8.9	$22,000,000	185	$96,067,179
Pulp Fiction	1994	8.9	$8,000,000	178	$107,930,000
The Good, the Bad and the Ugly	1966	8.9	$1,200,000	142	$6,100,000
12 Angry Men	1957	8.9	$350,000	96	NA
Inception	2010	8.8	$160,000,000	148	$292,568,851



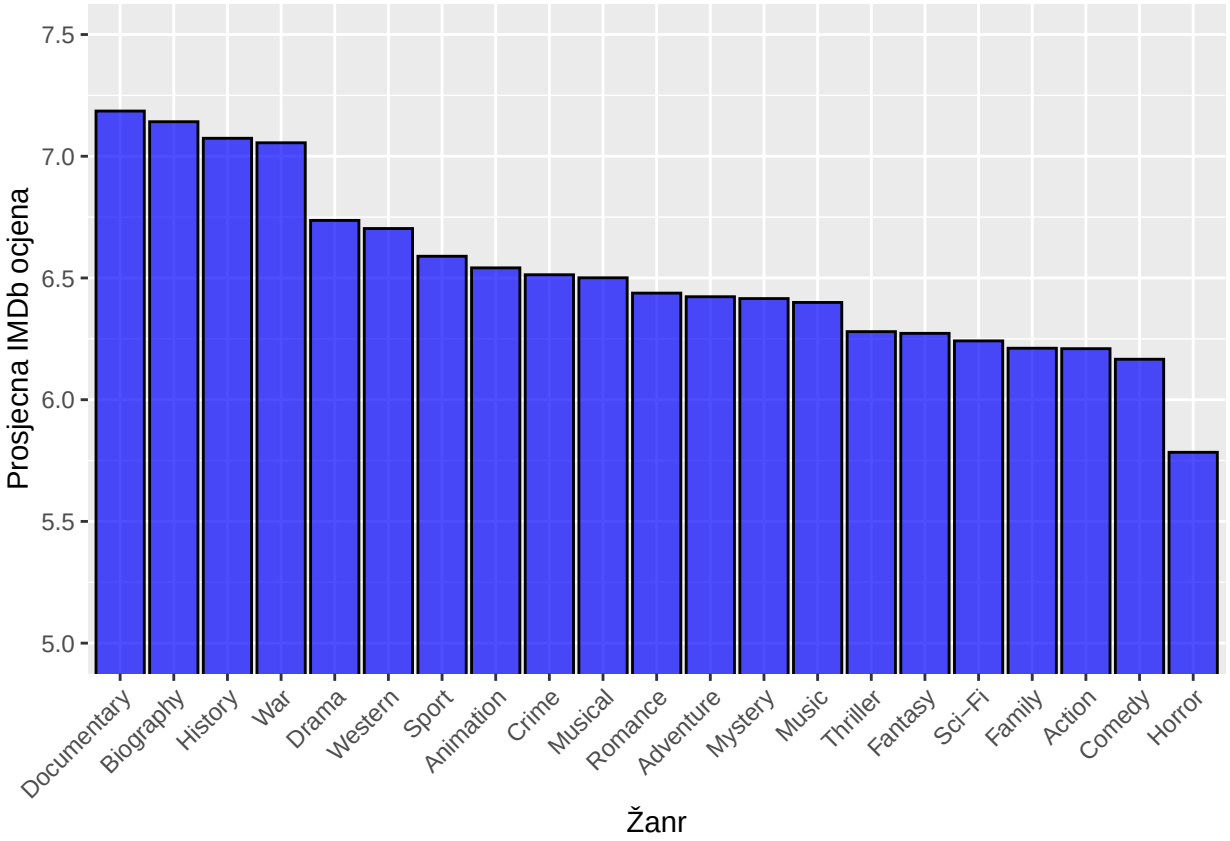
Tablicom smo prikazali 10 najbolje ocjenjenih filmova te njihove atribute: godinu izlaska, IMDb ocjenu, budžet, trajanje i bruto dobit. Nakon toga smo napravili četiri manja grafa gdje smo prikazali vrijednosti njihovih atirbuta(IMDb ocjena, budžet, trajanje i bruto dobit) u odnosu na godinu izlaska.

Prosječna ocjena filmova po trajanju



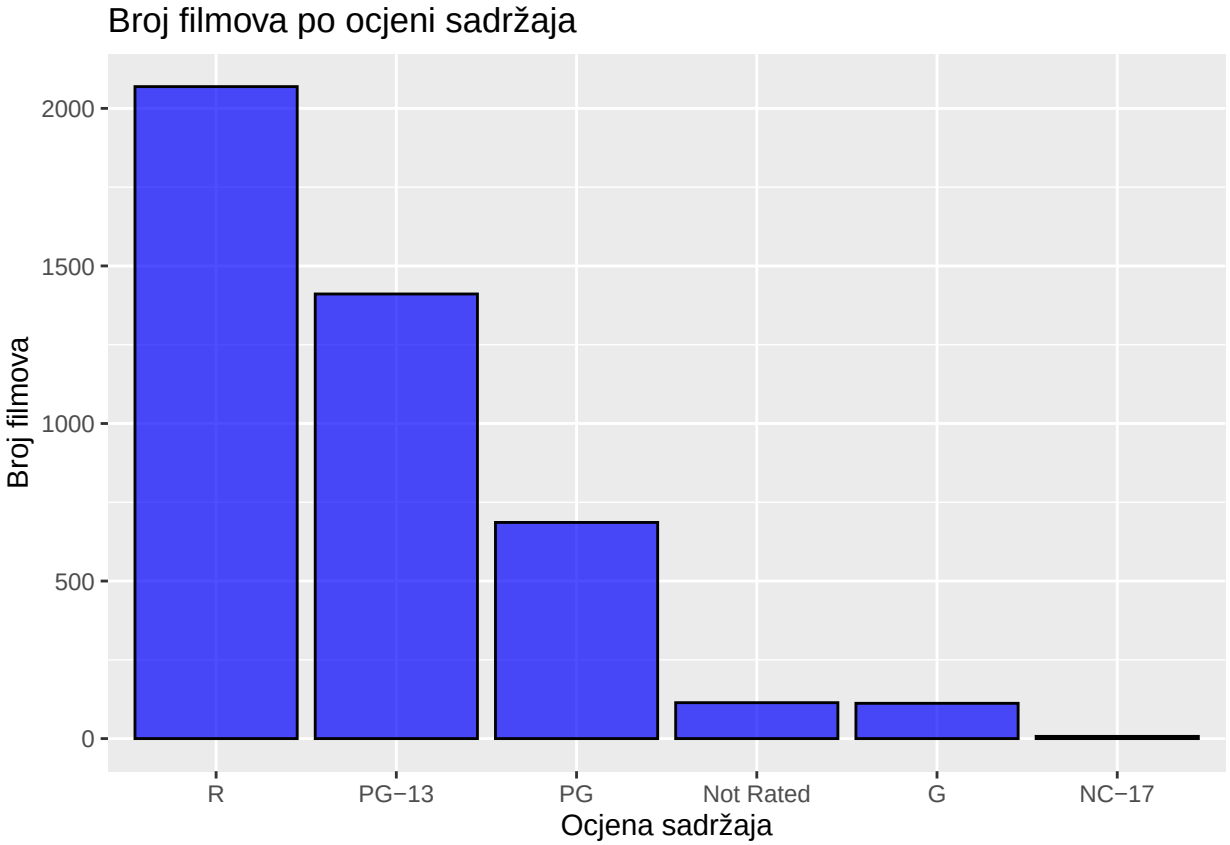
Ovaj graf nam pokazuje prosječnu ocjenu u odnosu na trajanje filma. Grupirali smo podatke po minutama trajanja filma te dobili prosječne ocjene za svaku minutu. Pravac crvene boje predstavlja funkciju koja predviđa prosječnu ocjenu za pojedino trajanje filmova. Jednadžba pravca je oblika: (Prosječna ocjena) = 6.2155021 + 0.0058487 * (Trajanje).

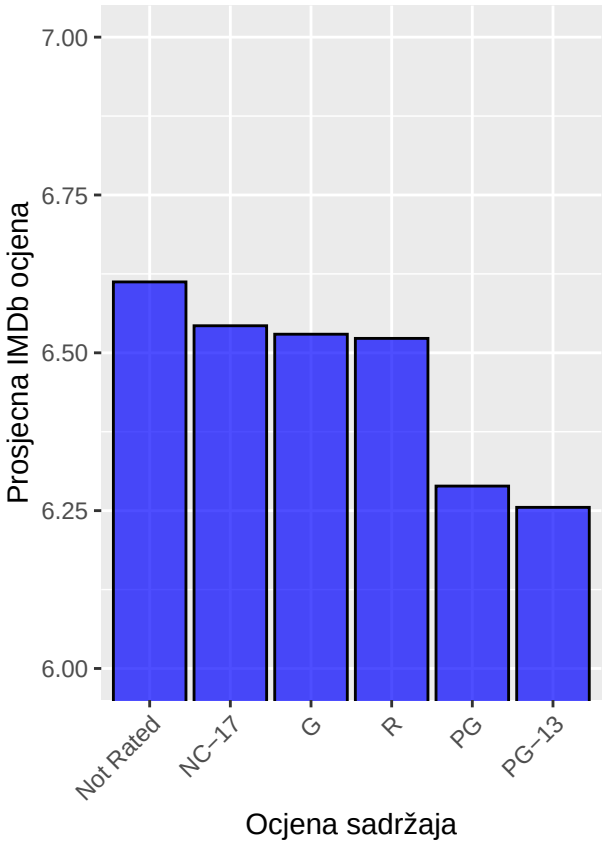
Prosječna IMDb ocjena po žanru

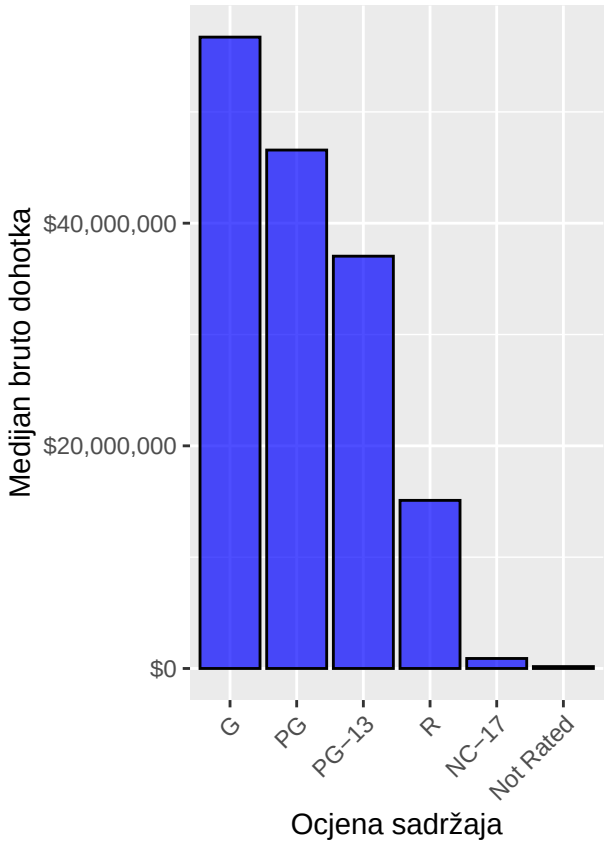


Pogledajmo koje žanrove ljudi najviše cijene. Izbacili smo žanrove koji imaju manje od 10 filmova u podatkovnom skupu (poput vijesti i film-noira). Možemo zaključiti da ljudi daju najviše ocjene "ozbiljnim" filmovima (žanrovi poput dokumentarnih filmova, biografija, povijesnih filmova, itd.). Još jedna stvar koju bismo mogli zaključiti je da ljudi bolje ocjenjuju filmove čija radnja se dogodila ili se može dogoditi, tu bih kao primjer opet naveli iste žanrove kao i prije. S druge strane, lošije su ocijenjeni žanrovi koji su "neozbiljni" poput komedija i obiteljskih filmova ili žanrovi čija je radnja često nerealna poput horora, akcije i znanstvene fantastike.

Usporedbe ocjena sadržaja (engl. content rating)







Ova 3 grafa istražuju utjecaj ocjene sadržaja na uspješnost filma. U analizi smo ostavili samo ocjene sadržaja koje se koriste od 1996. do danas u SAD-u jer tih filmova ima najviše te radi preglednosti. Prema MPA (Motion Picture Association), to su redom: - Rated G: General audiences – All ages admitted. - Rated PG: Parental guidance suggested – Some material may not be suitable for children. - Rated PG-13: Parents strongly cautioned – Some material may be inappropriate for children under 13. - Rated R: Restricted – Under 17 requires accompanying parent or adult guardian. - Rated NC-17: Adults Only – No one 17 and under admitted. Od ostalih ocjena sadržaja ostavili smo i Not Rated koji se koristi kad filmu iz nekog razloga još uvijek nije dana ocjena sadržaja.

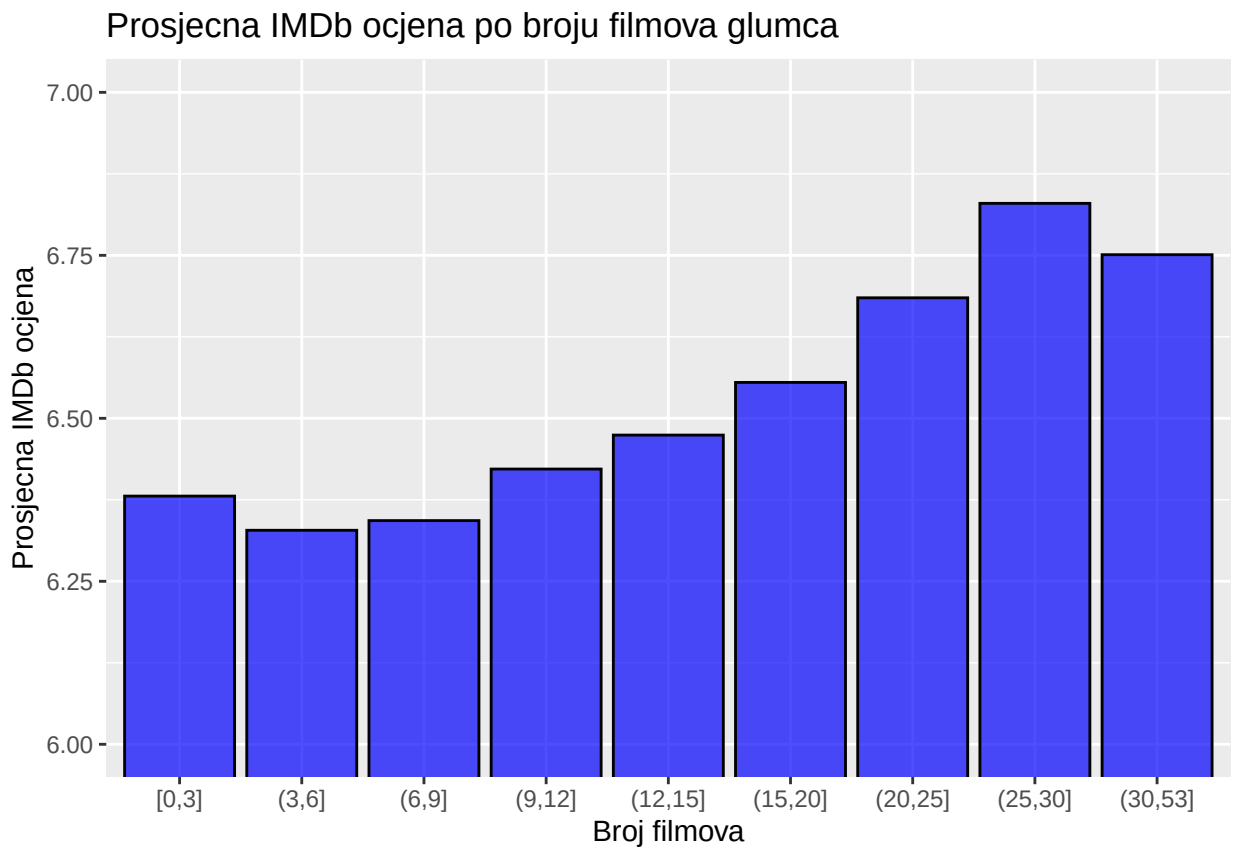
Prvi graf prikazuje razdiobu broja filmova gdje vidimo da broj filmova pada što je film primjereniji većoj publici. Jedina anomalija je to što ima najmanje filmova koji su 18+, izgleda da ta ocjena u globalu studijima predstavlja problem. S druge strane, drugi graf nam govori kako su ipak Not Rated i 18+ filmovi najbolje ocijenjeni, iako je razlika IMDb ocjena između svih kategorija vrlo mala. Zadnji graf potvrđuje opću pretpostavku da primjereniji filmovi zarađuju najviše novaca. To je jedini od 3 grafa koji nema anomalija nego su ocjene sadržaja sortirane po primjerenosti počevši od najprimjerenije.

Glumci

Najboljih 10 glumaca koji su napravili barem 10 filmova

Ime_glumca	Prosjecna_ocjena	Broj_filmova
John Ratzenberger	7.89	10
Leonardo DiCaprio	7.52	21
Orlando Bloom	7.47	11
Tom Hardy	7.47	12
Tom Hanks	7.40	28
Clint Eastwood	7.34	16
Christian Bale	7.30	27
Eddie Marsan	7.29	11
Joan Allen	7.26	10
Benedict Cumberbatch	7.19	10

Tablicom je prikazano top 10 glumaca koji su napravili 10 ili više filmova, broj filmova u kojima su sudjelovali te prosječna ocjena tih filmova. U tablici stvarno prepoznajemo najpoznatija imena u Hollywoodu. Na vrhu se nalazi John Ratzenberger koji je najpoznatiji po svojoj ulozi u seriji Cheers, a razlog njegove visoke ocjene je to što je posuđivao glas sporednim likovima u mnoštvu uspješnih Pixarovih filmova.



Na grafu vidimo kako je broj filmova u kojima je glumac glumio direknto koreliran s boljom prosječnom ocjenom filmova na kojima je radio. To je, naravno, očekivano jer najuspješniji glumci generalno glume u najcjenjenijim filmovima i stalno dobivaju nove uloge. "Ladice" grafa podijeljene su tako da prvo idu 3 po 3, zatim 5 po 5, a na kraju je prikazana ladica s glumcima koji su odglumili više od 30 filmova. "Ladice" su sve veće i veće iz razloga što je mali broj glumaca glumio u puno filmova.

Prvi i zadnji film glumaca

```
## Medijan ocjena prvog filma svih glumaca: 7
```

```
## Medijan ocjena zadnjeg filma svih glumaca: 6.5
```

```
## Medijan budžeta prvog filma svih glumaca: $16,000,000
```

```
## Medijan budžeta zadnjeg filma svih glumaca: $30,000,000
```

```
## Medijan bruto dohotka prvog filma svih glumaca: $33,000,000
```

```
## Medijan bruto dohotka zadnjeg filma svih glumaca: $52,200,504
```

U ovom odjeljku promatramo prve i zadnje filmove glumaca. Rezultati nam govore kako je unatoč manjem budžetu i bruto dohotku prvog filma u odnosu na zadnji film ipak ocjena prvog filma bolja za dosta velikih 0.5. Dolazimo do zaključka da glumci na kraju svoje karijere ipak dobivaju uloge u lošijim filmovima, vjerojatno radi gubitka glumačkih sposobnosti u starijoj dobi.

Redatelji

Top 10 najboljih redatelja

Table 3: Top 10 redatelja s jednim filmom

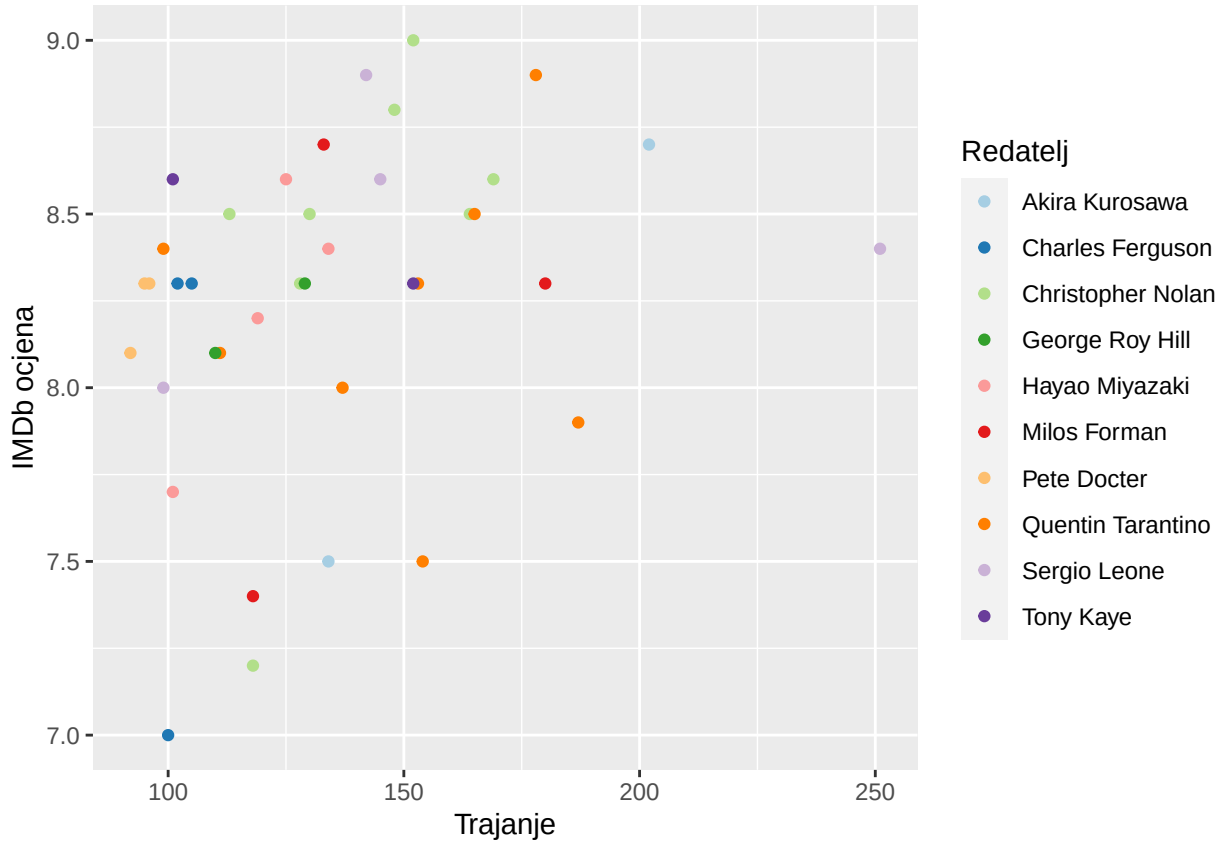
Redatelj	Prosjecna_ocjena
Sadyk Sher-Niyaz	8.7
Charles Chaplin	8.6
Damien Chazelle	8.5
Majid Majidi	8.5
Raja Menon	8.5
Ron Fricke	8.5
Asghar Farhadi	8.4
Bill Melendez	8.4
Catherine Owens	8.4
Jay Oliva	8.4

Table 4: Top 10 redatelja s barem 2 filma

Redatelj	Prosjecna_ocjena	Broj_filmova
Sergio Leone	8.47	4
Tony Kaye	8.45	2
Christopher Nolan	8.43	8
Charles Ferguson	8.30	2
Pete Docter	8.23	3
Hayao Miyazaki	8.22	4
George Roy Hill	8.20	2
Quentin Tarantino	8.20	8
Milos Forman	8.13	3
Akira Kurosawa	8.10	2

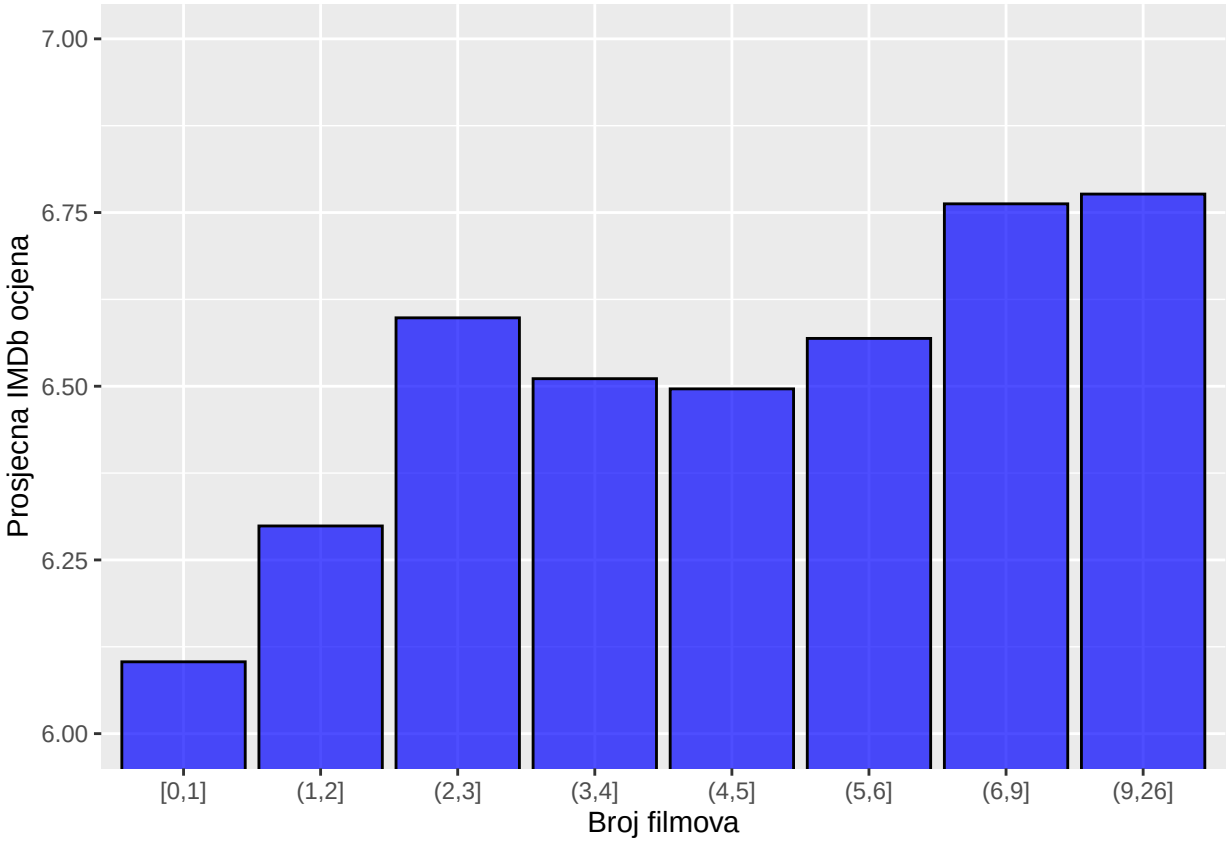
Najbolje redatelje po prosječnoj IMDb ocjeni smo odlučili prikazati u obliku talica. Prethodne dvije tablice prikazuju 10 najboljih redatelja s 1 filmom i 10 najboljih redatelja s barem 2 filma. Primjetili smo da pri odabiru 10 najboljih redatelja po prosječnoj IMDb ocjeni, većina ih ima samo jedan film, jedini film tog redatelja koji je poznat i dobro ocijenjen. Zbog toga smo odlučili podijeliti redatelje u 2 skupine, kako bi do izražaja došli i redatelji s većim brojem filmova.

Ocjene i trajanje filmova najboljih 10 redatelja (s barem 2 filma)



Prikaz svih filmova najboljih 10 redatelja koji imaju barem 2 filma. Na x osi je prikazano trajanje filma, na y osi IMDb ocjena, te se u legendi nalazi objašnjenje kojem redatelju film pripada.

Prosječna IMDb ocjena po broju filmova redatelja



Na ovome grafu, za razliku od sukladnog grafa glumaca, možemo vidjeti kako se kvaliteta filmova poboljšava samo na početku. Možemo zaključiti da je broj napravljenih filmova dosta zanemariv već nakon drugog filma premda su filmovi iskusniji redatelja još uvijek u prosjeku malo bolji nego onih manje iskusnih.

Prvi i zadnji film redatelja

```
## Medijan ocjena prvog filma svih redatelja: 7.1
```

```
## Medijan ocjena zadnjeg filma svih redatelja: 6.5
```

```
## Medijan budžeta prvog filma svih redatelja: $9,500,000
```

```
## Medijan budžeta zadnjeg filma svih redatelja: $40,000,000
```

```
## Medijan bruto dohotka prvog filma svih redatelja: $39,219,762
```

```
## Medijan bruto dohotka zadnjeg filma svih redatelja: $47,170,911
```

U ovom odjeljku promatramo prve i zadnje filmove redatelja. Ovi rezultati se ne razlikuju pretjerano od rezultata za prvi i zadnji film glumaca. Razlike vidimo u većoj razlici budžeta i manjoj razlici bruto dohotka.

Zaključak

Analizirajući ovaj podatkovni skup došli smo do nekoliko zaključaka, što o samome skupu, što o industriji i ljudima koji se njom bave. Što se tiče skupa, iznenadilo nas su se u njemu našle serije koje smo uspjeli identificirati uz pomoć stupca title_year koji je imao NA vrijednosti samo za serije. Još jedna opaska u skupu je ta da je skup vjerojatno iz 2016. godine i da filmovi iz te godine znaju imati drastično drugačiju ocjenu danas (2024. godine) nego tada, vjerojatno zbog malog broja glasova u to vrijeme. Što se tiče stranice IMDb, saznali smo kako korisnici ipak ocjenjuju filmove blaže nego što bi trebali. Što se tiče industrije, zanimljivo je to da redatelji s najboljom prosječnom ocjenom svojih filmova većinom imaju samo jedan film kojim su se proslavili. Ostale opaske bile su u skladu s očekivanjima te su objašnjene neposredno ispod iznešenih podataka.