# OSP Projekt: Eksploratorna analiza dataseta *IMDB_movie_dataset*

Jakov Krčadinac

2023-01-20

```
library(tidyverse)

library(gridExtra)

library(GGally)

library(ggplot2)
library(dplyr)
knitr::opts_chunk$set(results = 'hold')

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

Za izradu projekta odabrao sam dataset *IMDB_movie_dataset* jer sam oduvijek bio vrlo strastven oko filmova. Tijekom ranog djetinjstva sam naravno najviše gledao animirane filmove, a nakon što sam u osnovnoj školi značajno unaprijedio svoje poznavanje engleskog jezika, počeo sam puno više gledati prezežito američke dugometražne igrane filmove. U srednjoj sam školi skoro svaki tjedan gledao bar jedan film koji slovi kao "klasik", a za donošenje konačne odluke je li film za koji sam čuo vrijedan gledanja uglavnom bih se služio stranicama https://www.rottentomatoes.com/ ili https://www.imdb.com/

Cilj eksploratorne analize mi je proučiti najprofitabilnije i najbolje ocjenjene filmove, budući da su to dva kriterija koja me najviše zanimaju kada se informiram o filmu.

Za početak provjerimo koje sve stupce sadrži *IMDB_movie_dataset*.

```
colnames(data)
```

```
##  [1] "color"                     "director_name"
##  [3] "num_critic_for_reviews"    "duration"
##  [5] "director_facebook_likes"   "actor_3_facebook_likes"
##  [7] "actor_2_name"              "actor_1_facebook_likes"
##  [9] "gross"                     "genres"
## [11] "actor_1_name"              "movie_title"
## [13] "num_voted_users"           "cast_total_facebook_likes"
## [15] "actor_3_name"              "facenumber_in_poster"
## [17] "plot_keywords"             "movie_imdb_link"
## [19] "num_user_for_reviews"      "language"
## [21] "country"                   "content_rating"
## [23] "budget"                    "title_year"
## [25] "actor_2_facebook_likes"    "imdb_score"
## [27] "aspect_ratio"              "movie_facebook_likes"
```

Vidimo da dataset sadrži čak 28 stupaca. Odlučio sam izbaciti stupce *movie_imdb_link*, *facenumber_in_poster* i sve stupce koji sadeže informacije o broju Facebook lajkova filma ili ljudi koji su radili na filmu jer te podatke ne planiram koristiti u eksploratornoj analizi i smatram da su prilično irelevantni, pogotovo stupci o broju lajkova. Smatram da analiza broja lajkova daje jako loš uvid u bilo što jer je taj broj vrlo nepouzdan, budući da je Facebook zatrpan plaćenim oglasima. Također ne volim Facebook zbog svih ostalih manipulativnih praksi u njihovim aplikacijama, ali to je manje bitno.

```
#izbacivanje irelevantnih stupaca iz dataseta
df <- data %>% select(-c("movie_imdb_link", "facenumber_in_poster",
                         "director_facebook_likes", "actor_3_facebook_likes",
"actor_1_facebook_likes",
                         "cast_total_facebook_likes",
"actor_2_facebook_likes", "movie_facebook_likes"))
#df
```

Sljedeće što bih volio zamijeniti u datasetu je redoslijed stupaca. Smatram da na prvim mjestima trebaju biti stupci koji su najbitniji za analizu, a neki manje bitni detalji o filmovima neka budu među zadnjim stupcima u data frameu.

```
col_order <- c("movie_title", "title_year", "director_name", "duration",
               "budget", "gross", "imdb_score", "country", "language",
               "actor_1_name", "actor_2_name", "actor_3_name", "genres",
               "color", "content_rating", "aspect_ratio", "plot_keywords",
               "num_critic_for_reviews","num_user_for_reviews",
"num_voted_users")
df <- df[, col_order]
#df
```

Budući da sam u datasetu ostavio samo stupce koje smatram relevantnima i bitnima za analizu, razmišljao sam o izbacivanju svih redova koji sadrže NA vrijednosti. Međutim time bih izbacio 1228 redaka, što je više od 20% dataseta. Umjesto toga ću radije na početku svakog pitanja izbaciti retke koji sadrže NA vrijednosti u stupcima koji su relevantni za to pitanje, kako bih sačuvao što više podataka te dobio potpuniji i ispravniji odgovor na pitanje.

```
nrow(df)  # broj redaka cijelog dataseta
nrow(na.omit(df)) # broj redaka koji ne sadrže niti jednu NA vrijednost
nrow(df %>% drop_na(movie_title, title_year, duration)) # broj redaka u
kojima je poznat naziv filma, godina kada je film izašao i trajanje filma
# Za odgovor na neko pitanje su nam ovi podaci sasvim dovoljni, stoga je
pogrešno izbaciti sve retke koji sadrže NA vrijednosti na bilo kojem mjestu.
```

```
## [1] 5043
## [1] 3815
## [1] 4923
```

## Analiza filmova koji su zaradili najveću svotu novaca

### 1. Pitanje: Iz koje države dolaze filmovi koji zarađuju najviše?

Za početak bismo mogli probjeriti iz kojih država dolazi najviše filmova iz dataseta.

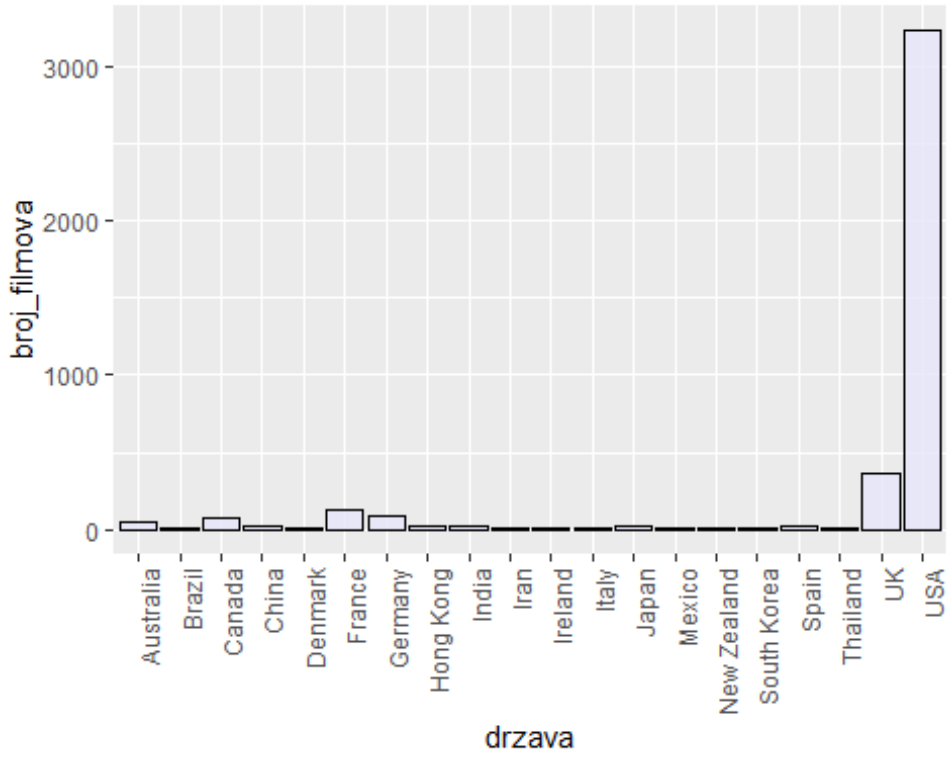
```
df_1 <- df %>% drop_na(movie_title, country, gross) #koristimo samo retke u
kojima imamo informacije o državi i zaradi filma
#nrow(df_1) koliko je redaka ostalo u dataframeu

broj_filmova_iz_drzave <- table(df_1$country)
top_20_drzava_po_broju_filmova <- sort(broj_filmova_iz_drzave, decreasing =
TRUE)[1:20] #top 20 država po broju filmova

top_20_drzava_po_broju_filmova["USA"]/nrow(df_1)  # postotak filmova koji
dolaze iz SAD-a
```

```
##       USA
## 0.7778312
```

Vidimo da daleko najviše (čak 77.8%) filmova iz dataseta dolazi iz SAD-a, što je za očekivati s obzirom na Hollywood.

```
top_20_drzava_po_broju_filmova_df <- data.frame(
  drzava = names(top_20_drzava_po_broju_filmova),
  broj_filmova = as.numeric(top_20_drzava_po_broju_filmova)
)
#top_20_drzava_po_broju_filmova_df
```

```
ggplot(top_20_drzava_po_broju_filmova_df) + geom_bar(aes(x=drzava,
y=broj_filmova), stat="identity", color="black", fill="lavender", alpha=0.8)
+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




S obzirom na veliku razliku između broja filmova koji dolaze iz SAD-a i ostalih država, za očekivati je da će i većina filmova koji su među top 20 filmova s najvećom zaradom također biti iz SAD-a.

```
# sortiram dataframe po stupcu gross silazno, dodao sam unique() jer sam
primijetio da postoje dupli redovi u tablici
highest_grossing <- df_1[order(df_1$gross, decreasing = T), ]  %>% unique()

top_20_highest_grossing <- highest_grossing[1:20, ]
top_20_highest_grossing <- top_20_highest_grossing %>% select(movie_title,
gross, country)
top_20_highest_grossing
```

```
##                                       movie_title     gross country
## 1                                          Avatar  760505847     USA
## 26                                        Titanic  658672302     USA
## 29                                 Jurassic World  652177271     USA
## 17                                   The Avengers  623279547     USA
## 66                                The Dark Knight  533316061     USA
## 234   Star Wars: Episode I - The Phantom Menace  474544677     USA
## 2813         Star Wars: Episode IV - A New Hope  460935665     USA
## 8                         Avengers: Age of Ultron  458991599     USA
## 4                           The Dark Knight Rises  448130642     USA
## 563                                       Shrek 2  436471036     USA
## 2860                 E.T. the Extra-Terrestrial  434949459     USA
## 183             The Hunger Games: Catching Fire  424645577     USA
## 13   Pirates of the Caribbean: Dead Man's Chest  423032628     USA
## 492                                 The Lion King  422783777     USA
## 43                                    Toy Story 3  414984497     USA
## 32                                     Iron Man 3  408992272     USA
## 426                               The Hunger Games  407999255     USA
## 27                    Captain America: Civil War  407197282     USA
## 159                                     Spider-Man  403706375     USA
## 36          Transformers: Revenge of the Fallen  402076689     USA
```

Vidimo da svih top 20 filmova s najvećom zaradom dolaze iz SAD-a, što je sukladno očekivanju.

## 2. Pitanje: Koji su glumci najzastupljeniji u filmovima koji su najviše zaradili?

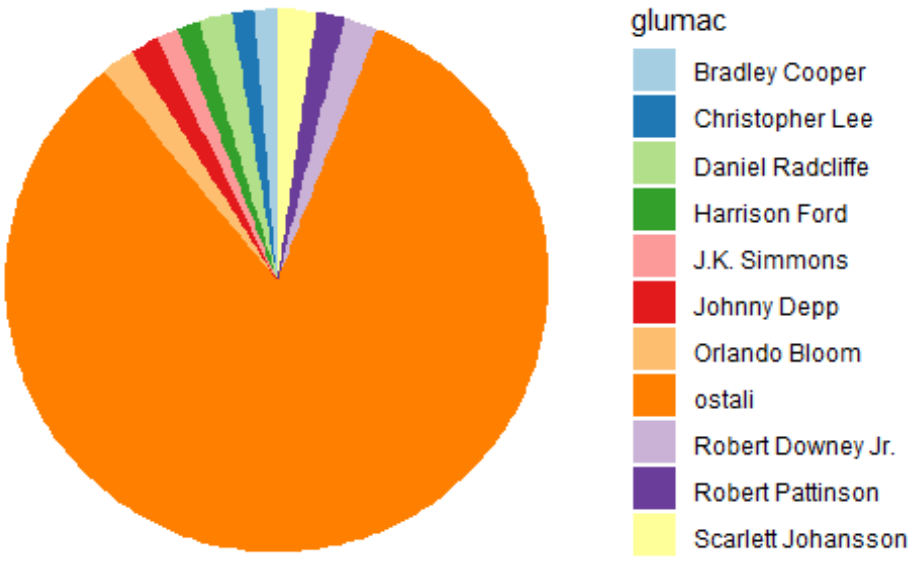
Znamo da se veliki filmovi često oslanjaju na slavu i prepoznatljivost glumaca koji su u njima. U datasetu za svaki film pišu 3 najpoznatija glumca koja glume u njemu. Napravit ćemo data frame sa top 100 filmova koji su najvišezaradili te ćemo iz tih podataka izvući koji su najzastupljeniji glumci. Za očekivati je da će se među njima nalaziti upravo neki od najpoznatijih hollywoodskih glumaca današnjice.

```
top_100_highest_grossing <- highest_grossing[1:100, ]
#top_100_highest_grossing

svi_glumci_u_top_100 <- c(top_100_highest_grossing$actor_1_name,
top_100_highest_grossing$actor_2_name, top_100_highest_grossing$actor_3_name)

tablica_glumaca <- sort(table(svi_glumci_u_top_100), decreasing = TRUE)
#tablica_glumaca

top_10_glumaca <- tablica_glumaca[1:10] #top 10 glumaca koji su u najviše
filmova s najvećom zaradom


ostali_glumci <-tail(tablica_glumaca, length(tablica_glumaca)-10) #glumci
koji nisu u top 10 glumaca po prisutnosti u filmovima s najvećom zaradom, ali
su prisutni u 100 filmova s najvećom zaradom.
```

```
top_10_glumaca
#ostali_glumci

## svi_glumci_u_top_100
## Scarlett Johansson   Daniel Radcliffe      Orlando Bloom  Robert Downey
Jr.
##                  7                   6                   6
6
##        Johnny Depp   Robert Pattinson     Bradley Cooper    Christopher
Lee
##                  5                   5                   4
4
##      Harrison Ford       J.K. Simmons
##                  4                   4
```

Od svih glumaca je upravo Scarlett Johansson bila u najviše filmova s najvećom zaradom. Osim Bradleya Coopera, Christophera Leea, Harrisona Forda i J.K. Simmonsa postoji još 7 glumaca koji su bili u 4 filma s najvećom zaradom, ali nisu upali u top 10 zbog abecednog poretka imena u tablici *tablica_glumaca*.

```
# kolika je zastupljenost svih ostalih glumaca zajedno?
broj_ostalih_glumaca <- sum(as.numeric(ostali_glumci))  #suma brojeva filmova
u kojima su bili glumci koji nisu u top 10 najzastupljenijih
#broj_ostalih_glumaca
# sum(as.numeric(top_10_glumaca))

# radim dataframe kako bih ga mogao koristiti u pie chartu
df_glumci <- data.frame(
  glumac = c(names(top_10_glumaca), "ostali"),
  broj_filmova = c(as.numeric(top_10_glumaca), broj_ostalih_glumaca)
)
#df_glumci

# pie chart zastupljenosti glumaca u top 10 filmova s najvećom zaradom
ggplot(df_glumci, aes(x="", y=broj_filmova, fill=glumac)) +
  geom_bar(stat="identity", width=1) +
  labs(name=NULL) +
  coord_polar("y", start=0) + theme_void() + scale_fill_brewer(palette =
"Paired")
```




Vidimo da se među najzastupljenijim glumcima u top 100 filmova s najvećom zaradom upravo neki od najpoznatijih glumaca u Hollywoodu trenutno, što je sukladno očekivanjima.

## 3. Pitanje: top 10 najprofitabilnijih filmova

U filmskoj je industriji nažalost jedan od najbitnijih faktora koji odlučuje hoće li se neki film napraviti ili ne upravo profitabilnost filma, odnosno želi se maksimizirati količina zarađenog novca. Zbog toga su u Hollywoodu sve zastupljeniji nastavci filmova u nekoj franšizi ukoliko je 1. film bio jako profitabilan, što znatno smanjuje raznolikost novih blockbustera.

Pogledajmo top 20 najprofitabilnijih filmova. Očekujem da će među njima biti većina filmova koji su i u top 20 filmova s najvećom zaradom.

```
df_3 <- df %>% drop_na(movie_title, budget, gross) #koristimo samo retke u
kojima imamo informacije o budžetu i zaradi filma
#df_3

#dodajem u dataframe df_3 stupac "profit" koji predstavlja čisti profit koji
je film ostvatio. Naravno svi filmovi ne moraju biti profitabilni, pa će tako
neki imati negativan profit, odnosno filmski studio zaslužan za izradu filma
je izgubio novac na njima.
df_3$profit <- df_3$gross - df_3$budget

most_lucrative <- df_3[order(df_3$profit, decreasing = T), ]  %>% unique()
```

```r
most_lucrative <- most_lucrative %>% select(movie_title, gross, budget,
profit)
top_20_most_lucrative <- most_lucrative[1:20, ]
top_20_most_lucrative
```

```
##                                         movie_title     gross   budget
## 1                                            Avatar  760505847 2.37e+08
## 29                                   Jurassic World  652177271 1.50e+08
## 26                                          Titanic  658672302 2.00e+08
## 2727           Star Wars: Episode IV - A New Hope  460935665 1.10e+07
## 2771                   E.T. the Extra-Terrestrial  434949459 1.05e+07
## 17                                   The Avengers  623279547 2.20e+08
## 488                                  The Lion King  422783777 4.50e+07
## 233     Star Wars: Episode I - The Phantom Menace  474544677 1.15e+08
## 65                                  The Dark Knight  533316061 1.85e+08
## 425                                The Hunger Games  407999255 7.80e+07
## 776                                        Deadpool  363024263 5.80e+07
## 182               The Hunger Games: Catching Fire  424645577 1.30e+08
## 667                                  Jurassic Park  356784000 6.30e+07
## 500                                Despicable Me 2  368049635 7.60e+07
## 778                                American Sniper  350123553 5.88e+07
## 328                                   Finding Nemo  380838870 9.40e+07
## 558                                        Shrek 2  436471036 1.50e+08
## 329  The Lord of the Rings: The Return of the King  377019252 9.40e+07
## 1447    Star Wars: Episode VI - Return of the Jedi  309125409 3.25e+07
## 796                                   Forrest Gump  329691196 5.50e+07
##         profit
## 1    523505847
## 29   502177271
## 26   458672302
## 2727 449935665
## 2771 424449459
## 17   403279547
## 488  377783777
## 233  359544677
## 65   348316061
## 425  329999255
## 776  305024263
## 182  294645577
## 667  293784000
## 500  292049635
## 778  291323553
## 328  286838870
## 558  286471036
## 329  283019252
## 1447 276625409
## 796  274691196
```

Vidimo da se otprilike pola svih filmova koji su bili u top 20 filmova po ukupnoj zaradi nalazi i u top 20 najprofitabilnijih filmova. Međutim nešto sam zanimljivo opazio promatrajući top 20 najlošije profitabilnih filmova.

```
top_10_least_lucrative <- tail(most_lucrative, 10) #zadnjih 10 redova nisu
poredani tako da je najmanje profitabilan film na 1. mjestu

#želimo obrnuti poredak redova
top_10_least_lucrative <- apply(top_10_least_lucrative, 2, rev)

# vraćanje rezultata u dataframe
top_10_least_lucrative <- as.data.frame(top_10_least_lucrative)

top_10_least_lucrative
```

```
##                   movie_title    gross      budget       profit
## 2701                The Host   2201412 12215500000 -12213298588
## 3310         Lady Vengeance    211667  4200000000  -4199788333
## 2716                Fateless    195888  2500000000  -2499804112
## 2150      Princess Mononoke   2298191  2400000000  -2397701809
## 2161                Steamboy    410388  2127519898  -2127109510
## 3015                   Akira    439162  1100000000  -1099560838
## 3662          Godzilla 2000  10037390  1000000000   -989962610
## 3305                   Tango   1687311   700000000   -698312689
## 2766 Kabhi Alvida Naa Kehna   3275443   700000000   -696724557
## 2914                   Kites   1602466   600000000   -598397534
```

Među filmovima se nalazi dugometražni animirani film Akira, jedan od najpoznatijih japanskih anime filmova i prvi anime koji je postao poznat izvan granica Japana. Po dobivenim podacima Akira je bio iznimno neprofitabilan film, međutim brzim istraživanjem na internetu jasno je vidljivo da je taj film itekako ostvario profit, međutim većina je kino ulaznica logično prodana u Japanu. Gross tog filma koji piše u tablici otprilike odgovara zaradi koju je ostvario u SAD-u, iz čega možemo zaključiti da cijeli stupac *gross* predstavlja zaradu koju je film ostvario u SAD-u.

Razmotrimo drugi način za mjerenje profitabilnosti. U filmskoj su industriji također poznate priče o malim indie filmovima nepoznatih redatelja oko kojih se stvori veliki hype na filmskim festivalima, te u konačnici ti filmovi postanu senzacije diljem svijeta i kasnije klasici. Jedan od takvih filmova je kultni Napoleon Dynamite iz 2004., film koji je zbog svoje originalnosti i jedinstvenog humora postao svjetski hit te je s budžetom od svega 400 000$ zaradio preko 44 milijuna dolara od prodaje ulaznica.

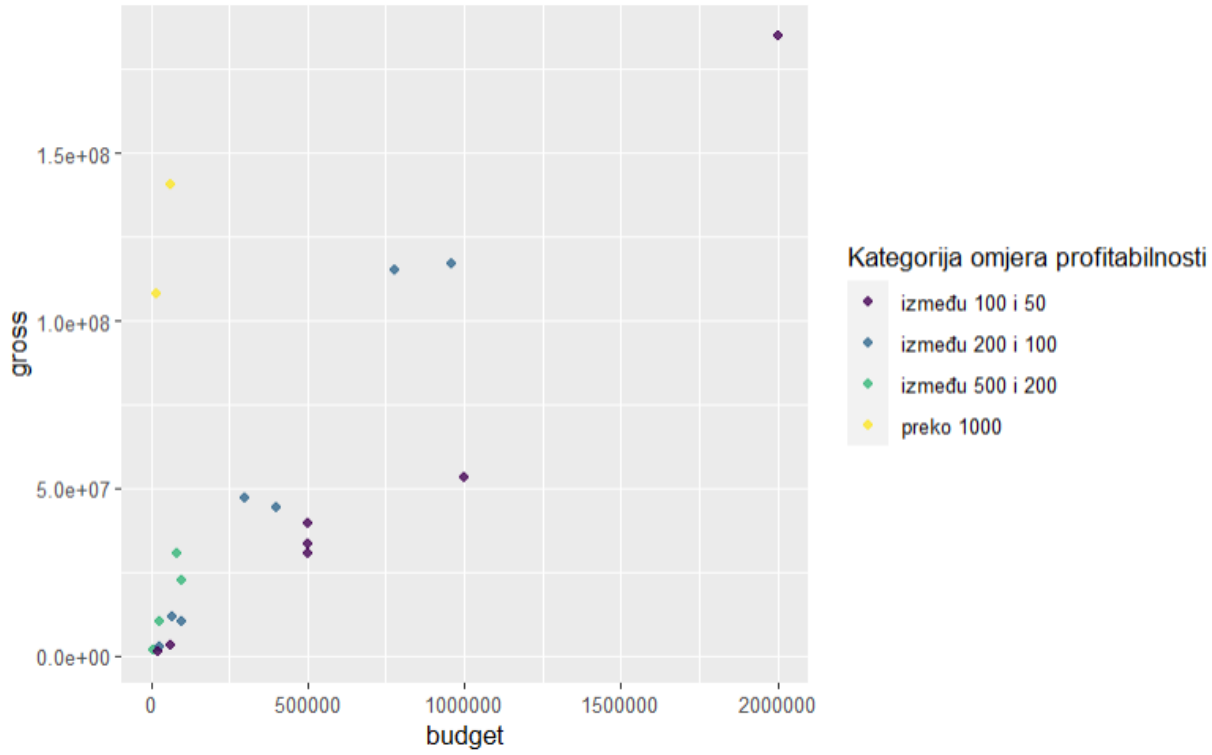
U nastavku dakle dodajemo novi stupac *omjer_profitabilnosti* pomoću kojeg ćemo poredati filmove silazno po omjeru zarađene svote i budžeta filma. Kao jedan od uvjeta također ćemo staviti da je film morao zaraditi minimalno milijun dolara.

```
df_3$omjer_profitabilnosti <- df_3$gross / df_3$budget
df_omjer <- df_3 %>% filter(gross>1000000)


df_omjer <- df_omjer[order(df_omjer$omjer_profitabilnosti, decreasing = T), ]
%>% select(movie_title, gross, budget, profit, omjer_profitabilnosti) %>%
unique()


top_20_df_omjer <- df_omjer[1:20, ] # top 20 filmova po omjeru
profitabilnosti

#top_20_df_omjer

# dodavanje novog stupca koji kategorizira svaki omjer profitabilnosti radi
lakšeg prikazivanja na grafu

kategoriziraj <- function(omjer){

  if(omjer>1000)
    kategorija <- "preko 1000"
  else if(omjer <= 1000 && omjer > 500)
    kategorija <- "između 1000 i 500"
  else if(omjer <= 500 && omjer > 200)
    kategorija <- "između 500 i 200"
  else if(omjer <= 200 && omjer > 100)
    kategorija <- "između 200 i 100"
  else if(omjer <= 100 && omjer > 50)
    kategorija <- "između 100 i 50"
  return(kategorija)
}

top_20_df_omjer$kategorija_omjera <-
sapply(top_20_df_omjer$omjer_profitabilnosti, FUN=kategoriziraj)


top_20_df_omjer
```

```
##                            movie_title     gross  budget    profit
## 3386              Paranormal Activity  107917283   15000 107902283
## 3366          The Blair Witch Project  140530114   60000 140470114
## 3426            The Brothers McMullen   10246600   25000  10221600
## 2788    The Texas Chain Saw Massacre   30859000   83532  30775468
## 3431                      El Mariachi    2040920    7000   2033920
## 3420                      The Gallows   22757819  100000  22657819
## 3425                    Super Size Me   11529368   65000  11464368
## 2222                       Halloween   47000000  300000  46700000
## 3355               American Graffiti  115000000  777000 114223000
## 3316                           Rocky  117235247  960000 116275247
## 3428            In the Company of Men    2856622   25000   2831622
## 3385                Napoleon Dynamite   44540956  400000  44140956
## 3419                Facing the Giants   10174663  100000  10074663
## 3295 Snow White and the Seven Dwarfs  184925485 2000000 182925485
## 3369                           Benji   39552600  500000  39052600
## 3368                       Fireproof   33451479  500000  32951479
## 3370                      Open Water   30500882  500000  30000882
## 3424                              Pi    3216970   60000   3156970
## 3430                         Slacker    1227508   23000   1204508
## 3328                 The Devil Inside   53245055 1000000  52245055
##      omjer_profitabilnosti kategorija_omjera
## 3386            7194.48553        preko 1000
## 3366            2342.16857        preko 1000
## 3426             409.86400  između 500 i 200
## 2788             369.42729  između 500 i 200
## 3431             291.56000  između 500 i 200
## 3420             227.57819  između 500 i 200
## 3425             177.37489  između 200 i 100
## 2222             156.66667  između 200 i 100
## 3355             148.00515  između 200 i 100
## 3316             122.12005  između 200 i 100
## 3428             114.26488  između 200 i 100
## 3385             111.35239  između 200 i 100
## 3419             101.74663  između 200 i 100
## 3295              92.46274   između 100 i 50
## 3369              79.10520   između 100 i 50
## 3368              66.90296   između 100 i 50
## 3370              61.00176   između 100 i 50
## 3424              53.61617   između 100 i 50
## 3430              53.36991   između 100 i 50
## 3328              53.24506   između 100 i 50
```

Vidimo da je većina ima budžet znatno ispod milijun dolara. Pogledajmo kako podaci izgledaju na grafu.

```
ggplot(top_20_df_omjer, aes(budget, gross, col=kategorija_omjera)) +
  geom_point(alpha = 0.8) +
  scale_color_ordinal(name="Kategorija omjera profitabilnosti")
```

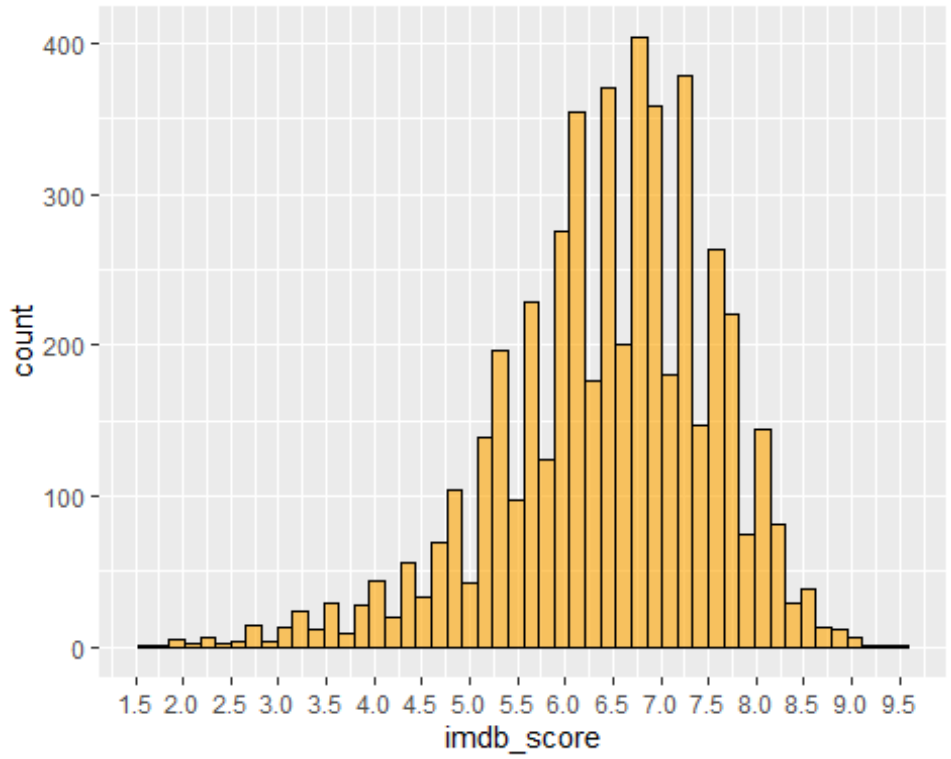



Zanimljivo je vidjeti da među filmovima kojima je omjer profitabilnosti između 100 i 50 ima filmova kojima je budžet čak 2 milijuna dolara, a opet postoje u potpuno drugom kraju grafa filmovi s budžetom znatno manjim od 100 000$.

## Analiza filmova koji su najbolje ocijenjeni

Za početak promotrimo razdiobu ocjena, odnosno *imdb_scorea* u datasetu.

```
df_ocjene <- df %>% drop_na(movie_title, imdb_score)
ggplot(df_ocjene, aes(x=imdb_score)) + geom_histogram(bins=50, color="black",
fill="orange", alpha=0.6) +
  scale_x_continuous(breaks = seq(0, 10, by = 0.5))
```




Vidimo da je razdioba ocjena filmova približno normalna i većina filmova ima ocjenu između 5 i 8.

## 4. Pitanje: postoji li razlika u medijanima ocjena filmova ovisno o uzrastu za koji je film namijenjen (*content_rating*)?

Pogledajmo prvo koliko filmova ima u svakoj kategoriji preporučenog uzrasta.

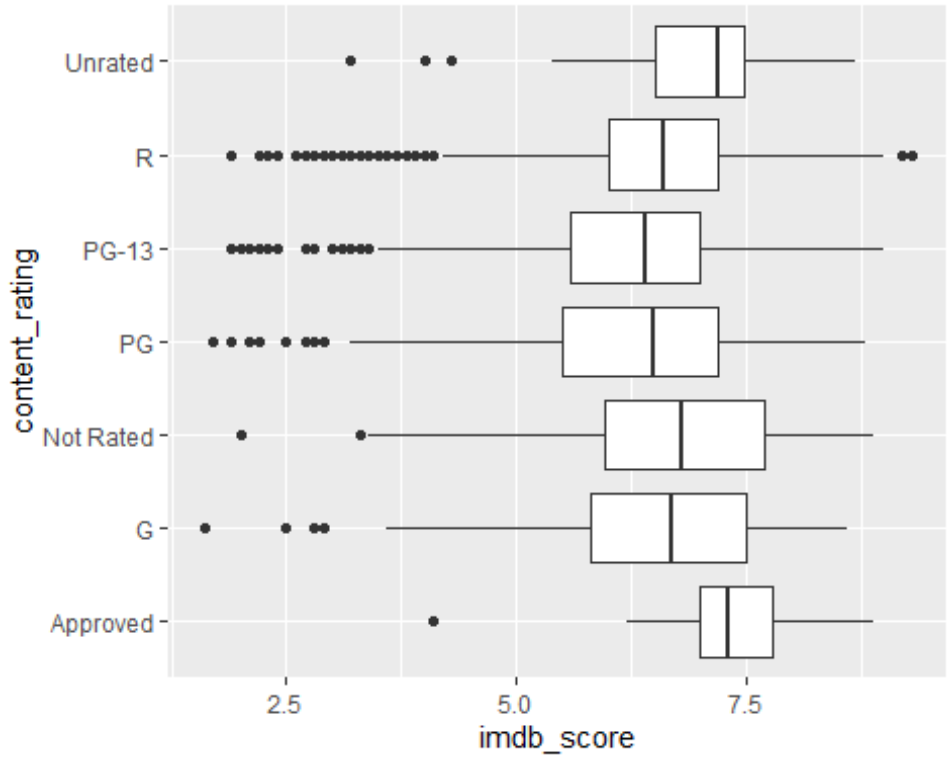
```
df_ocjene_pg_rating <- df_ocjene %>% drop_na(movie_title, imdb_score,
content_rating) %>% filter(content_rating != "")

frekvencije_kategorija <- sort(table(df_ocjene_pg_rating$content_rating),
decreasing = TRUE)
frekvencije_kategorija
```

```
##
##         R     PG-13        PG Not Rated         G   Unrated  Approved
TV-14
##      2118      1461       701       116       112        62        55
30
##     TV-MA     TV-PG         X      TV-G    Passed     NC-17        GP
M
##        20        13        13        10         9         7         6
```

```
5
##      TV-Y     TV-Y7
##         1         1
```

Vidimo da za neke kategorije imamo jako malo filmova, stoga ćemo izbaciti sve kategorije koje sadrže manje od 50 filmova.

```
prihvatljive_frekvencije_kategorija <-
names(frekvencije_kategorija[frekvencije_kategorija>=50])
prihvatljive_frekvencije_kategorija

# u data frameu ostavljamo filmove koji su isključivo u dobnim kategorijama
koje sadrže bar 50 filmova iz originalnog dataseta.
df_ocjene_pg_rating_analiza <- df_ocjene_pg_rating %>% filter(content_rating
%in% prihvatljive_frekvencije_kategorija)

ggplot(df_ocjene_pg_rating_analiza, aes(x=imdb_score, y=content_rating)) +
  geom_boxplot()
```




```
## [1] "R"         "PG-13"     "PG"        "Not Rated" "G"         "Unrated"
## [7] "Approved"
```

Iznad su navedene sve prihvatljive preporučene dobne kategorije (sadrže više od 50 filmova).
Vidimo da su medijani većine kategorija podjednaki, a kategorija *Approved* ima najveći medijan te 1. i 3. kvartil.

## 5. Pitanje: Koji redatelj ima najviše vrlo dobro ocijenjenih filmova?

Vrlo dobro ocijenjenim filmovima smatramo filmove koji su imali *imdb_score* veći ili jednak 7.5

```
df_5 <- df %>% drop_na(movie_title, imdb_score, director_name) %>%
filter(director_name != "")


vrlo_dobre_ocj <- df_5 %>% filter(imdb_score>=7.5)


najbolje_ocjenjeni_redatelji <- sort(table(vrlo_dobre_ocj$director_name),
decreasing = TRUE)[1:10]
najbolje_ocjenjeni_redatelji


as.numeric(najbolje_ocjenjeni_redatelji[1]) / nrow(df_5 %>%
filter(director_name == "Steven Spielberg"))  #postotak Spilbergovih filmova
koji su vrlo dobro ocijenjeni

##
##      Steven Spielberg       Martin Scorsese        Clint Eastwood
##                    14                    12                     8
##         David Fincher     Quentin Tarantino     Christopher Nolan
##                     8                     8                     7
## Alejandro G. Iñárritu  Francis Ford Coppola         James Cameron
##                     6                     6                     6
##         Peter Jackson
##                     6
## [1] 0.5384615
```

Riječ je dakako o Stevenu Spilbergu, redatelju koji je osvojio 3 prestižne nagrade Oscar za najboljeg redatelja. Čak je 54% filmova koje je režirao, vrlo dobro ocijenjeno.

## Zaključak

Tijekom analize sam se imao prilike vrlo dobro upoznati s data setom i bilo mi je drago među filmovima vidjeti neke od mojih omiljenih.
Iako sam koristio nekoliko različitih grafičkih prikaza podataka, pred kraj izrade projekta sam primijetio da se dosta pitanja na koje sam htio naći odgovor u datasetu, svodilo na rangiranje data frameova i traženje "naj" nečega. Zbog toga bi se analiza mogla smatrati donekle ograničenom ili repetitivnom, ali iskreno mogu reći da sam uživao u traženju odgovora na pitanja koja me stvarno zanimaju i korištenju stečenog znanja da dođem do odgovora.
Također smatram da sam pri odgovaranju na ta pitanja dobio dobar pregled data seta i izvukao nekoliko bitnih zaključaka o njemu, što je koliko sam shvatio poanta eksploratorne analize.