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))

Chart, histogram

Description automatically generated

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")

Chart, pie chart

Description automatically generated

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()   
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")

Chart, scatter chart

Description automatically generated

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))

Chart, histogram

Description automatically generated

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*)?

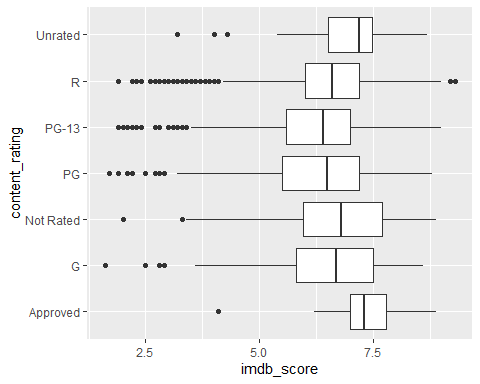
Pogledajmo prvo koliko filmova ima u svakoj kategoriji preporučenog uzrasta.

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.