Prirodno-matematički fakultet, Univerzitet u Kragujevcu

Seminarski rad na temu
“Movie Industry”

Student
Katarina Mošić 90/2018

Predmetni profesor
Branko Arsić

Sadržaj

Uvod

Tema: Movie Industry

Naziv: movies.csv

Movie Industry je skup podataka koji daje informacije o filmovima napravljenim u periodu 1980-2020. Cilj analize podataka je da se otkrije da li filmska industrija propada. Kolona za predvidjanje gross. Podaci su izvučeni sa IMDb-a.

Link ka skupu podataka na kaggle sajtu:
https://www.kaggle.com/datasets/danielgrijalvas/movies

Priprema podataka

Učitavanje podataka

Učitavanje potrebnih biblioteka

```
library(tidyverse)
```

```
## Warning: package 'dplyr' was built under R version 4.4.1
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.4.1
```

```
## Loading required package: Rcpp
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.2, built: 2024-04-10)
## ## Copyright (C) 2005-2024 James Honaker, Gary King and Matthew Blackwell
```

```
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##

library(nortest)
library(ggplot2)
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.4.1

## corrplot 0.94 loaded

library(rpart)
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 4.4.1

library(randomForest)

## Warning: package 'randomForest' was built under R version 4.4.1

## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin

library(caret)

## Warning: package 'caret' was built under R version 4.4.1

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

Učitavanje podataka: movies.csv.

```
movies = read.csv("C:/Users/Nikola/Documents/data/movies.csv", sep=",")
View(movies)
```

Funkcija head nam pomaže da se upoznamo sa podacima prikazujući prvih 6 redova u skupu podataka.

```
head(movies)
```

```
##                                                name rating     genre year
## 1                                       The Shining      R     Drama 1980
## 2                                   The Blue Lagoon      R Adventure 1980
## 3 Star Wars: Episode V - The Empire Strikes Back     PG    Action 1980
## 4                                         Airplane!     PG    Comedy 1980
## 5                                        Caddyshack      R    Comedy 1980
## 6                                   Friday the 13th      R    Horror 1980
##                          released score   votes           director
## 1 June 13, 1980 (United States)   8.4  927000    Stanley Kubrick
## 2  July 2, 1980 (United States)   5.8   65000     Randal Kleiser
## 3 June 20, 1980 (United States)   8.7 1200000     Irvin Kershner
## 4  July 2, 1980 (United States)   7.7  221000       Jim Abrahams
## 5 July 25, 1980 (United States)   7.3  108000       Harold Ramis
## 6   May 9, 1980 (United States)   6.4  123000 Sean S. Cunningham
##                    writer           star        country  budget     gross
## 1           Stephen King Jack Nicholson United Kingdom 1.9e+07  46998772
## 2 Henry De Vere Stacpoole Brooke Shields  United States 4.5e+06  58853106
## 3          Leigh Brackett    Mark Hamill  United States 1.8e+07 538375067
## 4            Jim Abrahams    Robert Hays  United States 3.5e+06  83453539
## 5      Brian Doyle-Murray    Chevy Chase  United States 6.0e+06  39846344
## 6           Victor Miller   Betsy Palmer  United States 5.5e+05  39754601
##              company runtime
## 1       Warner Bros.     146
## 2  Columbia Pictures     104
## 3          Lucasfilm     124
## 4 Paramount Pictures      88
## 5     Orion Pictures      98
## 6 Paramount Pictures      95
```

Funkcije ncol i nrow daju informacije o dimenzija skupa podataka. Set podataka movies se sastoji od 15 kolona i 7668 redova.

```
ncol(movies)
```

```
## [1] 15
```

```
nrow(movies)
```

```
## [1] 7668
```

Funkcija str(movies) nam daje uvid u tipove podataka. U skupu podataka se nalazi 9 kategorijskih (chr) i 6 numeričkih promenljivih (int, num).

```
str(movies)
```

```
## 'data.frame':    7668 obs. of  15 variables:
##  $ name    : chr  "The Shining" "The Blue Lagoon" "Star Wars: Episode V -
The Empire Strikes Back" "Airplane!" ...
##  $ rating  : chr  "R" "R" "PG" "PG" ...
##  $ genre   : chr  "Drama" "Adventure" "Action" "Comedy" ...
##  $ year    : int  1980 1980 1980 1980 1980 1980 1980 1980 1980 1980 ...
##  $ released: chr  "June 13, 1980 (United States)" "July 2, 1980 (United
States)" "June 20, 1980 (United States)" "July 2, 1980 (United States)" ...
##  $ score   : num  8.4 5.8 8.7 7.7 7.3 6.4 7.9 8.2 6.8 7 ...
##  $ votes   : num  927000 65000 1200000 221000 108000 123000 188000 330000
101000 10000 ...
##  $ director: chr  "Stanley Kubrick" "Randal Kleiser" "Irvin Kershner" "Jim
Abrahams" ...
##  $ writer  : chr  "Stephen King" "Henry De Vere Stacpoole" "Leigh
Brackett" "Jim Abrahams" ...
##  $ star    : chr  "Jack Nicholson" "Brooke Shields" "Mark Hamill" "Robert
Hays" ...
##  $ country : chr  "United Kingdom" "United States" "United States" "United
States" ...
##  $ budget  : num  1.9e+07 4.5e+06 1.8e+07 3.5e+06 6.0e+06 5.5e+05 2.7e+07
1.8e+07 5.4e+07 1.0e+07 ...
##  $ gross   : num  4.70e+07 5.89e+07 5.38e+08 8.35e+07 3.98e+07 ...
##  $ company : chr  "Warner Bros." "Columbia Pictures" "Lucasfilm"
"Paramount Pictures" ...
##  $ runtime : num  146 104 124 88 98 95 133 129 127 100 ...
```

Opis podataka

Skup podataka sadrži sledeće kolone:

1. name - kategorijska promenljiva koja označava naziv filma
2. rating - kategorijska promenljiva koja označava kategoriju filma (R, PG,..)
3. genre - kategorijska promenljiva koja označava žanr filma
4. year - numerička promenljiva koja označava godinu objavljivanja filma
5. released - kategorijska promenljiva koja označava datum objavljivanja u formatu (YYYY-MM-DD)
6. score - numerička promenljiva koja označava IMDb ocenu korisnika
7. votes - numerička promenljiva koja označava broj ljudi koji su glasali
8. director - kategorijska promenljiva koja označava direktora filma
9. writer - kategorijska promenljiva koja označava pisca
10. star - kategorijska promenljiva koja označava glumca/glumicu koji je zvezda filma

11. country - kategorijska promenljiva koja označava zemlju porekla filma
12. budget - numerička promenljiva koja označava budžet filma.
13. gross - numerička promenljiva koja oznacava prihod filma
14. company - kategorijska promenljiva koja označava produkcijsku kuću
15. runtime - numerička promenljiva koja označava trajanje filma

Nedostajuće vrednosti

Rukovanje nedostajućim vrednostima obezbedjuje da podaci budu u odgovarajućem obliku za dalju analizu.

Funkcija summary nam daje statistički prikaz podataka.

```
summary(movies)
```

```
##      name              rating             genre               year
##  Length:7668        Length:7668        Length:7668        Min.   :1980
##  Class :character   Class :character   Class :character   1st Qu.:1991
##  Mode  :character   Mode  :character   Mode  :character   Median :2000
##                                                           Mean   :2000
##                                                           3rd Qu.:2010
##                                                           Max.   :2020
##
##    released            score           votes            director
##  Length:7668        Min.   :1.90   Min.   :      7   Length:7668
##  Class :character   1st Qu.:5.80   1st Qu.:   9100   Class :character
##  Mode  :character   Median :6.50   Median :  33000   Mode  :character
##                     Mean   :6.39   Mean   :  88109
##                     3rd Qu.:7.10   3rd Qu.:  93000
##                     Max.   :9.30   Max.   :2400000
##                     NA's   :3      NA's   :3
##     writer              star             country             budget
##  Length:7668        Length:7668        Length:7668        Min.   :
3000
##  Class :character   Class :character   Class :character   1st Qu.:
10000000
##  Mode  :character   Mode  :character   Mode  :character   Median :
20500000
##                                                           Mean   :
35589876
##                                                           3rd Qu.:
45000000
##                                                           Max.
:356000000
##                                                           NA's   :2171
```

```
##      gross              company             runtime
##  Min.   :3.090e+02   Length:7668        Min.   : 55.0
##  1st Qu.:4.532e+06   Class :character   1st Qu.: 95.0
##  Median :2.021e+07   Mode  :character   Median :104.0
##  Mean   :7.850e+07                      Mean   :107.3
##  3rd Qu.:7.602e+07                      3rd Qu.:116.0
##  Max.   :2.847e+09                      Max.   :366.0
##  NA's   :189                            NA's   :4
```

Ovaj skup podataka ima nedostajuće vrednosti u 5 numeričkih kolona i to su: score, votes, budget, gross i runtime. Ove podatke je potrebno pripremiti za dalju analizu.

Ispod je procentualni prikaz nedostajućih vrednosti.

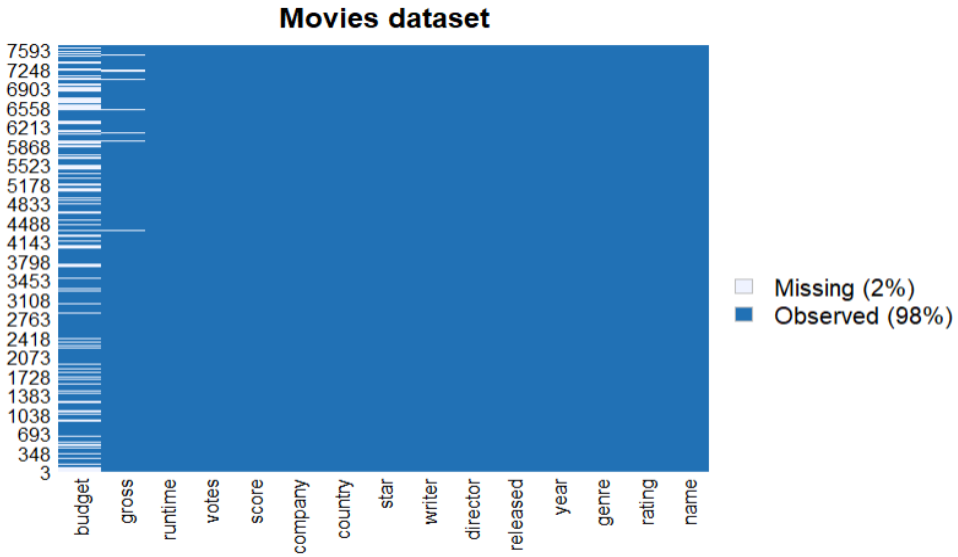
```
colMeans(is.na(movies))*100
```

```
##        name      rating       genre        year    released       score
##  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.03912363
##       votes    director      writer        star     country      budget
##  0.03912363  0.00000000  0.00000000  0.00000000  0.00000000 28.31246740
##       gross     company     runtime
##  2.46478873  0.00000000  0.05216484
```

Možemo da vidimo da nijedna kolona nema više od 60% observacija za koje podaci nedostaju, zato što bi jedino takva kolona mogla da se obriše, naravno u slučaju da je beznačajna. Kao što smo mogli i iz summary funkcije da vidimo, najviše podataka nedostaje u koloni budget, oko 28%.

Grafički prikaz kolona sa nedostajućim vrednostima:

```
par(mfrow=c(1,1))
missmap(obj = movies, main = "Movies dataset")
```



Numeričke kolone

Ovde možemo da vidimo da ukupan procenat numeričkih podataka koji nedostaju u odnosu na celi skup iznosi 2%.

Prva numerička kolona sa nedostajućim vrednostima je score.

```
sum(is.na(movies$score))
```

```
## [1] 3
```

U ovoj koloni imamo 3 podatka koji nedostaju. Pošto je podatke uvek bolje sačuvati nego obrisati, a u ovom slučaju imamo mali broj podataka koji nedostaje, možemo upotrebiti metodu umetanja vrednosti kao npr. mean ili median. Izbor metode zavisi od distribucije podataka, zato moram da proverim kakva je distribucija podataka u ovoj koloni. Shapiro-Wilk test koji smo ranije koristili je ograničen na 3-5000 podataka, što znači da ne može da se upotrebi na ovom koji ima 7668 podataka. Zbog toga koristm Anderson-Darling test za proveru normalnosti.

```
ad.test(movies$score)
```

```
##
##  Anderson-Darling normality test
##
## data:  movies$score
## A = 25.739, p-value < 2.2e-16
```

Velika vrednost statistike A ukazuje na veće odstupanje od normalne distribucije. p-vrednost < 2.2e-16 - manje od 0.05 Odbacujemo pretpostavku da su podaci u ovoj koloni normalno distribuirani. Kada varijabla nije normalno distribuirana, medijana je bolji izbor od proseka za popunjavanje nedostajućih vrednosti, zato što je manje podložna ekstremnim vrednostima (outlierima).

```
median.score <- median(x=movies$score, na.rm = T)
median.score
```

```
## [1] 6.5
```

```
movies$score[is.na(movies$score)] <- median.score

sum(is.na(movies$score)) #Provera da li jos uvek ima vrednosti koje nedostaju
```

```
## [1] 0
```

Druga numerička kolona sa nedostajućim vrednostima je votes.

```
sum(is.na(movies$votes))
```

```
## [1] 3
```

U ovoj koloni takodje imamo 3 podatka koji nedostaju. Ponovo koristim Anderson-Darling test za proveru normalnosti kako bih odredila šta ce biti vrednost umetanja.

```
ad.test(movies$votes)
```

```
##
##  Anderson-Darling normality test
##
## data:  movies$votes
## A = 1049.6, p-value < 2.2e-16
```

Velika vrednost statistike A i mala p-vrednost < 2.2e-16 ukazuju na to da podaci ni u ovoj koloni nisu normalno distribuirani - zato ponovo koristim medijanu.

```
median.votes <- median(x = movies$votes, na.rm=T)
median.votes
```

```
## [1] 33000
```

```
movies$votes[is.na(movies$votes)] <- median.votes

sum(is.na(movies$votes)) #Provera da li jos uvek ima vrednosti koje nedostaju
```

```
## [1] 0
```

Treća numerička kolona sa nedostajućim vrednostima je runtime.

```
sum(is.na(movies$runtime))
```

```
## [1] 4
```

Isto kao i do sad, zadržaću ove podatke, samo prvo proveravam distribuciju podataka.

```
ad.test(movies$runtime)
```

```
##
##  Anderson-Darling normality test
##
## data:  movies$runtime
## A = 146.08, p-value < 2.2e-16
```

I ovde vidimo da podaci nisu normalno distribuirani. Ponovo koristim medijanu umesto proseka za ovu varijablu.

```
median.runtime <- median(x = movies$runtime, na.rm=T)
median.runtime
```

```
## [1] 104
```

```
movies$runtime[is.na(movies$runtime)] <- median.runtime

sum(is.na(movies$runtime)) #Provera da li jos uvek ima vrednosti koje
nedostaju
```

```
## [1] 0
```

Naredne dve numerčke kolone sa nedostajućim vrednostima su budget i gross. Prvo proveravam da li postoje redovi koji nemaju ni gross ni budget.

```
cat("Broj nedostajucih u koloni budget:", sum(is.na(movies$budget)), "\n")

## Broj nedostajucih u koloni budget: 2171

cat("Broj nedostajucih u koloni gross:", sum(is.na(movies$gross)), "\n")

## Broj nedostajucih u koloni gross: 189

missing = which(is.na(movies$budget) & is.na(movies$gross))
movies_miss = movies[missing, ]

cat("Broj redova koji nemaju ni gross ni budget:",
sum(is.na(movies_miss$budget)), "\n")

## Broj redova koji nemaju ni gross ni budget: 128

cat("Procenat redova koji nemaju ni gross ni budget:",
(sum(is.na(movies_miss$gross))/ 7668) * 100, "%") #procenat zajednicikih
nedostajucih u odnosu na ceo skup

## Procenat redova koji nemaju ni gross ni budget: 1.669275 %
```

Postoje redovi koji nedostaju u obe kolone i ima ih ukupno 128 što predstavlja ~1.7% u odnosu na ceo skup podataka. Ove redove je možda najbolje obrisati zato što je budget bitan prediktor i zato što ih nema puno, a popunjavanje ovih kolona može da dovede do lažnih zaključaka što može negativno uticati na tačnost modela. Zato ću kombinovati pristupe. Ove redove ću obrisati, a ostale ću analizirati u nastavku zato što je previše da obrišem sve podatke iz obe kolone koji nedostaju.

```
movies <- movies[-missing, ]
```

Sada tražim pristup za popunjavanje kolone budget.

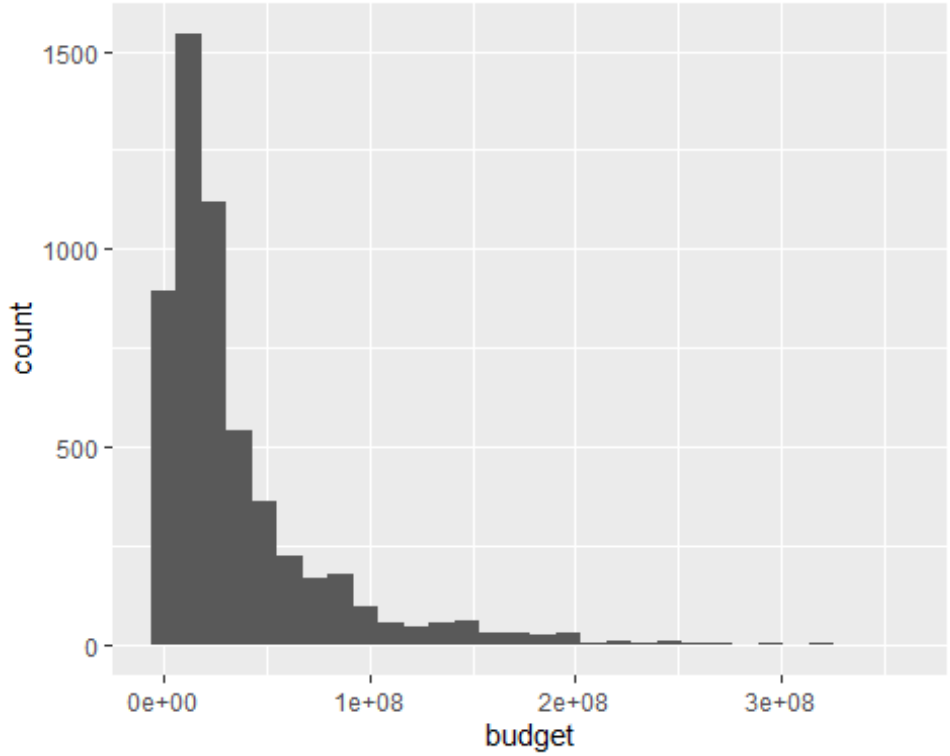
```
sum(is.na(movies$budget))

## [1] 2043
```

Sada nedostaje 2043 podataka za ovu kolonu.

```
na_budget <- movies[is.na(movies$budget), ]
View(na_budget)
```

Deluje da nema specifičnog obrasca medju podacima koji nedostaju. Nemamo podatke odredjenih kombinacija npr. žanrova ili godina koji nedostaju. Podaci koji nedostaju su dosta raznoliki. Brisanje podataka ili popunjavanje sa median/mean mi ne deluje kao dobra opcija zato što je veliki procenat podataka koji nedostaju. Bilo bi previše izbrisati oko 2000 redova podataka iz skupa koji ima ~7600. Isto tako koristiti vrednost mean/median za umetanje može dosta da umanji varijansu u podacima. Zbog toga sada gledam raspodelu kolone budget i da li postoji mogućnost da na osnovu nje kreiram novu kolonu.

```
ggplot(data = movies, mapping = aes(x=budget)) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2043 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

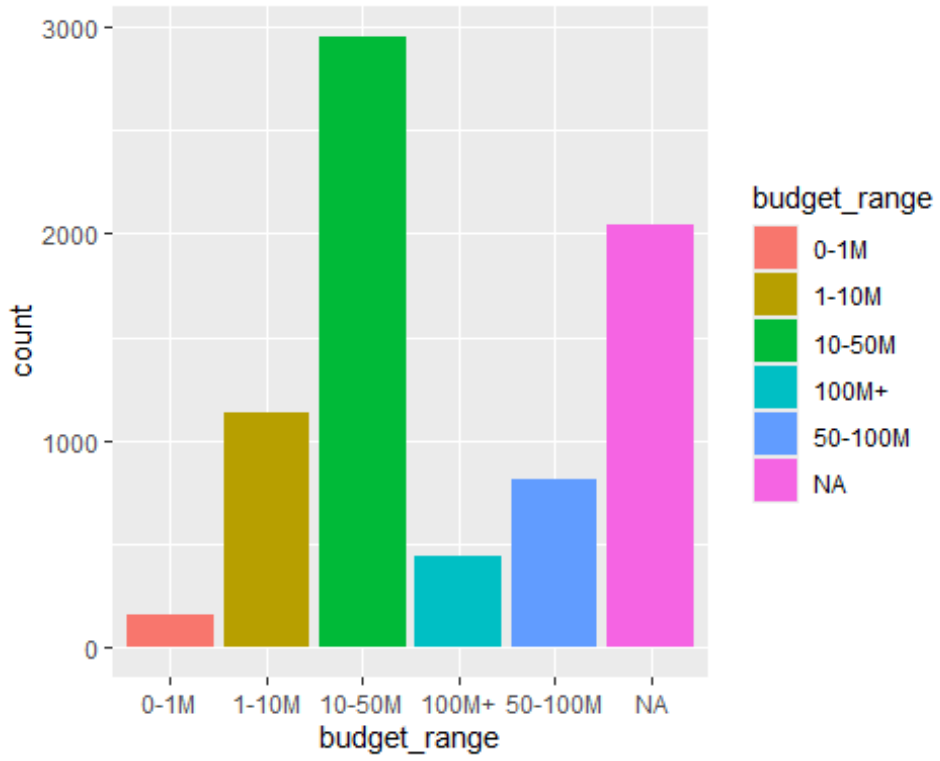


Vidimo da je jako mali broj filmova koji imaju budžet veći od 100 miliona (1e+08). Zbog ovakve raspodele, upotrebiću FE pristup i kreiraću kategorijsku promenljivu budget_range. Ova kolona će imati 6 kategorija i to: 1. Nedostajuće vrednosti predstavljaju kategoriju za sebe 2. 0 - 1M 3. 1 - 10M 4. 10 - 50M 5. 50 - 100M 6. 100M+

```
movies <- movies %>%
  mutate(budget_range = case_when(
    is.na(budget) ~ 'NA',
    budget >= 0 & budget < 1e6 ~ '0-1M',
    budget >= 1e6 & budget < 1e7 ~ '1-10M',
    budget >= 1e7 & budget < 5e7 ~ '10-50M',
    budget >= 5e7 & budget < 1e8 ~ '50-100M',
    budget >= 1e8 ~ '100M+'
  ))
```

Sada imam kolonu budget_range koju ću dalje da koristim kao zamenu za budget. Ovako sada izgleda raspodela po kategorijama:

```
ggplot(data=movies) +
  geom_bar(mapping = aes(x=budget_range, fill=budget_range))
```



Peta numerička kolona sa nedostajućim vrednostima je gross

```
sum(is.na(movies$gross))
```

```
## [1] 61
```

Isto kao i do sad, zadržaću ove podatke, samo prvo proveravam distribuciju podataka.

```
ad.test(movies$runtime)
```

```
##
##  Anderson-Darling normality test
##
## data:  movies$runtime
## A = 140.74, p-value < 2.2e-16
```

I ovde vidimo da podaci nisu normalno distribuirani. Ponovo koristim medijanu umesto proseka za ovu varijablu.

```
median.gross <- median(x = movies$gross, na.rm=T)
median.gross
```

```
## [1] 20205757
```

```
movies$gross[is.na(movies$gross)] <- median.gross

sum(is.na(movies$gross)) #Provera da li jos uvek ima vrednosti koje nedostaju
```

```
## [1] 0
```

Kategorijske kolone

Pomoću summary funkcije smo saznali da postoje nedostajuće vrednosti ali samo u numeričkim kolonama. Za kategorijske moram da proverim da li u kolonama postoje polja koja nemaju vrednost.

sapply funkciju ću da iskoristim kako bih odjednom proverila koje sve kolone imaju prazna polja.

```
sapply(movies, function(x) {
  if (is.character(x)) {
    sum(x == "")
  } else {
    NA
  }})
```

```
##          name        rating         genre          year      released
score
##             0            62             0            NA             0
NA
##         votes      director        writer          star       country
budget
##            NA             0             3             1             1
NA
##         gross       company       runtime budget_range
##            NA            14            NA             0
```

Ovde vidimo da imamo još vrednosti koje nedostaju i to u kolonama: rating, writer, star, country i company.

Promenljiva rating ima 66 praznih polja. Znači da 66 filmova ima “” vrednost za kategoriju filma. Ovakve vrednosti ne moraju uvek da znače da vrednosti nedostaju već mogu biti povezane sa vrednostima drugih promenljivih. U ovom slučaju mislim da “” kod rating nije posledica ni jedne druge promenljive već nedostajuća vrednost.

Funkcija unique nam daje uvid u jedinstvene vrednosti kategorijske promenljive rating.

```
unique(movies$rating)
```

```
##  [1] "R"         "PG"        "G"         ""          "Not Rated" "NC-17"
##  [7] "Approved"  "PG-13"     "Unrated"   "X"         "TV-PG"     "TV-MA"
## [13] "TV-14"
```

Ovde možemo da vidimo da postoji nepravilnost, a to je da su napravljene dve kategorije za neocenjene filmove (Not Rated i Unrated). To možemo da rešimo tako što ćemo ove dve kategorije da objedinimo zato što imaju isto značenje, sve će postati “Unrated”. Isto radim i sa podacima koji nedostaju u ovoj koloni. Najviše logike ima da pripadaju kategoriji neocenjenih. Zbog toga sada nije toliko važno koliko kojih vrednosti ima kada je već nedostajuća napravljena kao posebna kategorija.

```
movies$rating[which(movies$rating == "Not Rated")] <- "Unrated"

movies$rating[which(movies$rating == "")] <- "Unrated"
```

Provera nedostajućih za kolonu writer

```
length(which(movies$writer == ""))

## [1] 3

length(xtabs(~writer, data=movies)) #broj jedinstvenih vrednosti

## [1] 4444

(length(which(movies$writer == "")) / 7668) * 100 #procenat nedostajucih u
odnosu na ceo skup

## [1] 0.03912363
```

Ova kolona takođe ima veliki broj jedinstvenih vrednosti i mali broj onih koji nedostaju u odnosu na celokupan skup, pa ću ih obrisati.

```
movies = movies[movies$writer != "", ]
length(which(movies$writer == "")) #Provera da li ih jos uvek ima, tj. da li
su stvarno obrisane

## [1] 0
```

Sledeća je kolona star. Pošto je došlo do brisanja nekih redova, u preostalim kolonama moram da proverim da li još uvek ima nedostajućih.

```
length(which(movies$star == ""))

## [1] 1

length(xtabs(~star, data=movies))

## [1] 2733
```

Kolona star ima 1 polje bez vrednosti.Takođe ima veliki broj jedinstvenih vrednosti. Zbog toga se opet odlučujem da obrišem taj koji nedostaje.

```
movies = movies[movies$star != "", ]
length(which(movies$star == "")) #Provera da li ih jos uvek ima, tj. da li su
stvarno obrisane

## [1] 0
```

Ponovo proveravam broj za kolonu country.

```
length(which(movies$country == ""))

## [1] 1
```

```
length(xtabs(~country, data=movies))
```

```
## [1] 60
```

Promenljiva country IMA 1 polje bez vrednosti. Ovde postoji 60 jedinstvenih vrednosti što i nije toliko mnogo kao kod prethodnih kolona, pa ću videti da li postoji kategorija/vrednost koja se značajno izdvaja od ostalih.

Većina filmova potiče iz United States, pa ću ovo da koristim kao vrednost umetanja za podatak koji nedostaje.

```
movies$country[which(movies$country == "")] <- 'United States'
length(which(movies$country == "")) #Proveravam da li kolona company ima
nedostajuce vrednosti
```

```
## [1] 0
```

Ostala je još jedna kolona - company.

```
length(which(movies$company == ""))
```

```
## [1] 14
```

```
length(xtabs(~company, data=movies))
```

```
## [1] 2312
```

Kolona company IMA 14 polja bez vrednosti i 2312 jedinstvenih vrednosti.

```
(length(which(movies$company == "")) / 7668) * 100
```

```
## [1] 0.1825769
```

Vrednosti koje nedostaju čine 0.18% u odnosu na ceo skup podataka što nije puno. Takođe Ima veliki broj jedinstvenih vrednosti pa ću zbog toga obrisati one koji nedostaju.

```
movies = movies[movies$company != "", ]
length(which(movies$company == ""))
```

```
## [1] 0
```

Raspodela podataka po kolonama

U ovom delu ću prikazati raspodelu podataka po kolonama i na taj način saznati nešto više o podacima. Omogućava nam da vidimo kako su podaci raspodeljeni za svaku promenljivu, da li postoje izuzeci, itd,..

Kolona Name

Tabela učestalosti za kolonu name

```
movies %>% group_by(name) %>% summarise(count = n()) %>% arrange(desc(count))
%>% filter(count > 1)
```

```
## # A tibble: 139 × 2
##    name                       count
##    <chr>                      <int>
##  1 Anna                           3
##  2 Fever Pitch                    3
##  3 Hamlet                         3
##  4 Hercules                       3
##  5 Nobody's Fool                  3
##  6 Pulse                          3
##  7 Venom                          3
##  8 A Nightmare on Elm Street      2
##  9 After the Wedding              2
## 10 Aladdin                        2
## # ℹ 129 more rows
```

Filtritala sam sve > 1 kako bih videla koji su to filmovi koji se pojavljuju više puta. Da li se ponavljaju isti filmovi tj. duplikati i da li je to rezultat neke greške ili su to različiti filmovi.

```
movies %>% filter(name == "Anna" | name == "Fever Pitch")
```

```
##          name rating  genre year                         released score
votes
## 1 Fever Pitch      R  Drama 1985 November 22, 1985 (United States)   4.1
243
## 2        Anna  PG-13  Drama 1987 November 28, 1987 (United States)   6.5
639
## 3 Fever Pitch      R Comedy 1997    April 4, 1997 (United Kingdom)   6.7
10000
## 4 Fever Pitch  PG-13 Comedy 2005     April 8, 2005 (United States)   6.2
43000
## 5        Anna      R  Drama 2013          January 24, 2014 (Spain)   6.5
22000
## 6        Anna      R Action 2019     June 21, 2019 (United States)   6.6
69000
##           director           writer           star        country budget
## 1   Richard Brooks   Richard Brooks    Ryan O'Neal  United States  7e+06
## 2 Yurek Bogayevicz Yurek Bogayevicz Sally Kirkland  United States     NA
## 3      David Evans      Nick Hornby    Colin Firth United Kingdom     NA
```

```
## 4   Bobby Farrelly      Lowell Ganz Drew Barrymore  United States  3e+07
## 5     Jorge Dorado       Guy Holmes    Mark Strong          Spain  7e+06
## 6       Luc Besson       Luc Besson     Sasha Luss         France     NA
##      gross                   company runtime budget_range
## 1   618847 Metro-Goldwyn-Mayer (MGM)      96        1-10M
## 2  1236848              Magnus Films     100           NA
## 3     3736        Channel Four Films     102           NA
## 4 50605163         Fox 2000 Pictures     104       10-50M
## 5  1257142        The Safran Company      99        1-10M
## 6 31626978      Summit Entertainment     118           NA
```

Na osnovu ovih podataka deluje da ovu kolonu nema smisla analizirati. Svaki film ima drugačiji naziv, a ukoliko postoji više filmova sa istim nazivom radi se o različitim filmovima. Možemo videti na primeru filmova "Anna" i "Fever Pitch" da su različiti na osnovu drugih kolona, poput žanra, godine, pisca...

Kolona Rating

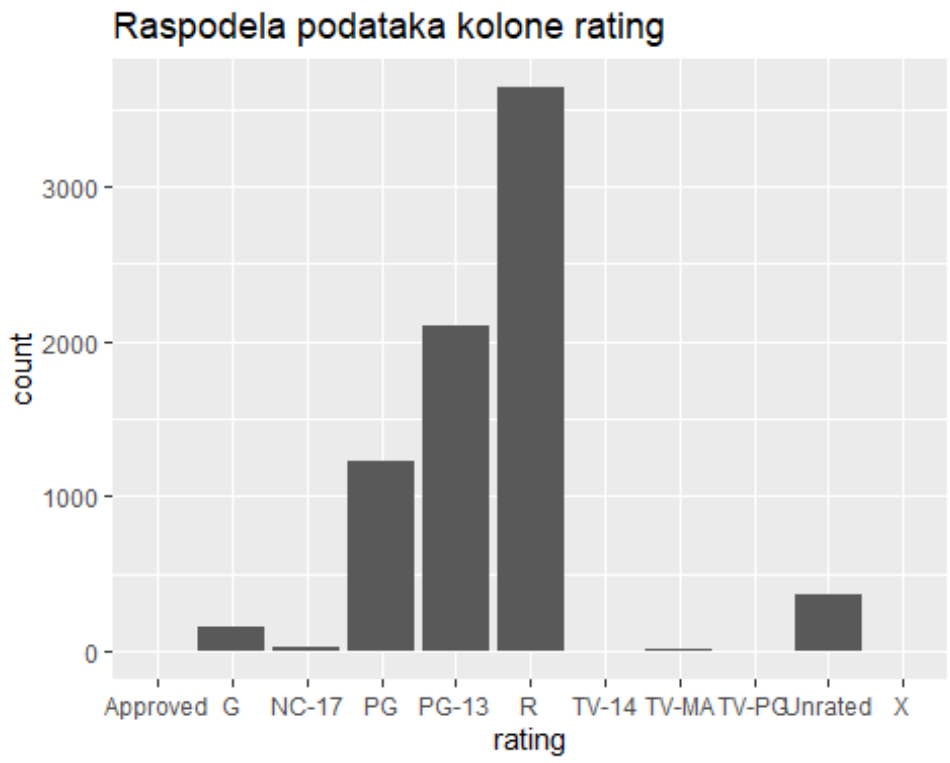
Tabela učestalosti za rating kako bismo videli koliko kategorija ima.

```
xtabs(~rating, movies)
```

```
## rating
## Approved        G    NC-17       PG    PG-13        R    TV-14    TV-MA
##        1      152       23     1229     2097     3640        1        9
##    TV-PG  Unrated        X
##        3      364        3
```

Grafik raspodele filmova prema kategoriji (rating)

```
  ggplot(data = movies, mapping = aes(x = rating)) +
  geom_bar() +
  labs(title = "Raspodela podataka kolone rating", x = "rating", y = "count")
```



Sa grafika možemo da vidimo da najveći broj filmova, oko 3500 ima ocenu “R” koja označava neko ograničenje (“Restricted”), a zatim “PG-13” koju ima nešto više od 2000 filmova. Mnogo manje filmova ima ocenu “G”.

U celom skupu imamo 1 film koji ima kategoriju Approved i 1 TV-14. Njihov uticaj na model je minimalan zato što ne daju dovoljno informacija za pouzdan zaključak.

Na osnovu analize značenja kategorija došla sam do zaključka da neke od kategorija mogu da spojim jer su bespotrebno odvojene.

Approved kategorija je zastarela i ranije se koristila kako bi označila filmove koji imaju dozvolu za javno prikazivanje. Tako da ovu kategoriju mogu da spojim sa G (General Audiences) koja se odnosi na filmove prikladne za sve uzraste. PG je oznaka da je film prikladan za većinu publike ali da roditelji treba da budu opreziji jer moze sadržati scene nasilja ili nepristojne reči. TV-PG je oznaka da film zahteva roditeljski nadzor. Approved, G, PG i TV-PG su kategorije koje predstavljaju sadržaj za sve uzraste tako da svi mogu da budu pod jednom kategorijom npr. PG - zato što svakako svi mogu da gledaju te filmove i sa decom uz opreznost.

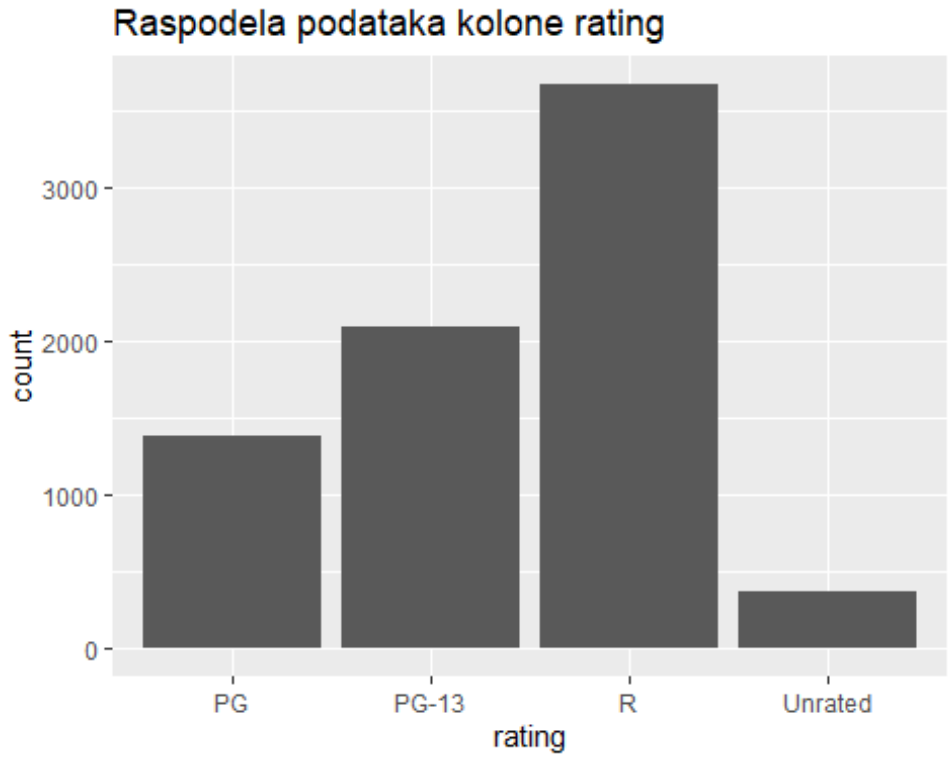
PG-13, TV-14 mogu da se spoje u jednu - PG-13. Bilo bi logičnije da spojim u TV-14 jer svakako i deca od 13 godina spadaju u tu kategoriju ali je PG-13 poznata filmska kategorija, a u TV-14 imam samo jedan film tako da mislim da nije toliko bitno.

R (Restricted) može da se spoji sa NC-17 koji je zabranjen osobama mladjim od 17 godina. Isto tako TV-MA (sadržaj namenjen odrasloj publici) i x (zastarela oznaka koja se kasnije transformisala u NC-17).

```
movies$rating[which(movies$rating == "Approved" | movies$rating == "G" |
movies$rating == "TV-PG")] <- "PG"
movies$rating[which(movies$rating == "TV-14")] <- "PG-13"
movies$rating[which(movies$rating == "NC-17" | movies$rating == "TV-MA" |
movies$rating == "X")] <- "R"
movies$rating[which(movies$rating == "Not Rated")] <- "Unrated"
```

Sada grafik izgleda ovako:

```
  ggplot(data = movies, mapping = aes(x = rating)) +
  geom_bar() +
  labs(title = "Raspodela podataka kolone rating", x = "rating", y = "count")
```



Kolona genre

Prikaz učestalosti za genre, kako bismo videli koliko kojih žanrova ima.

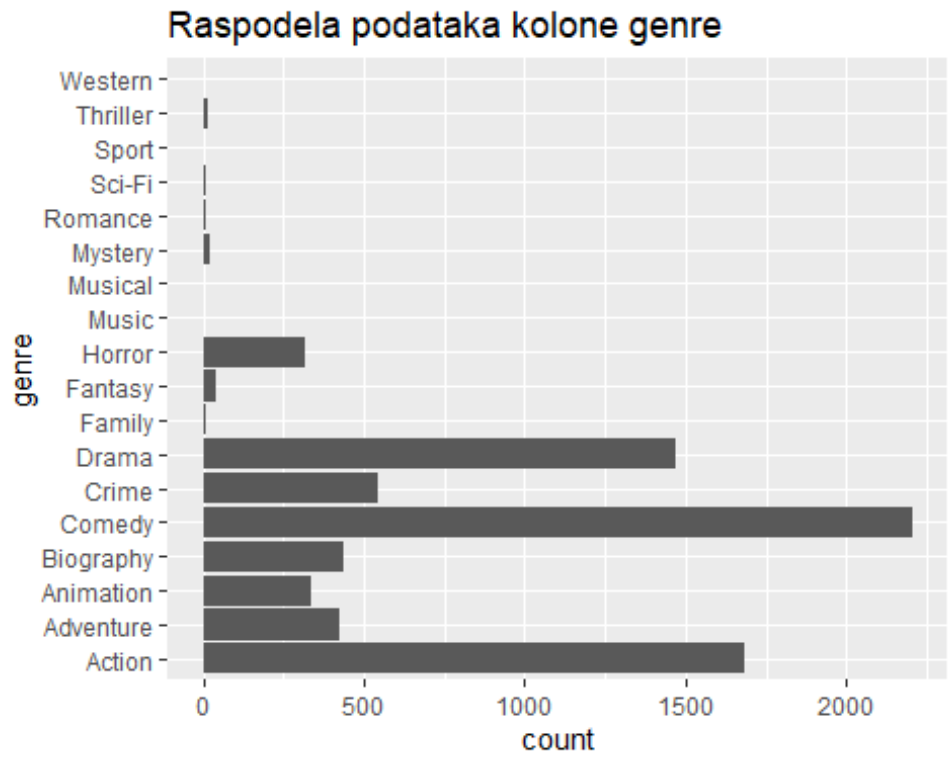
```
xtabs(~genre, data=movies)
```

```
## genre
##    Action Adventure Animation Biography    Comedy     Crime     Drama
Family
##      1680       422       335       435      2203       547      1470
11
##   Fantasy    Horror     Music   Musical   Mystery   Romance    Sci-Fi
Sport
##        43       317         1         2        20         8        10
1
```

```
##  Thriller   Western
##        14         3
```

Grafik raspodele filmova prema žanru (genre)

```
ggplot(data = movies, mapping=aes(y=genre)) +
  geom_bar() +
  labs(title = "Raspodela podataka kolone genre", y="genre", x="count")
```



Najveći broj filmova je iz kategorija Komedija, Akcija i Drama. To su najbrojnije kategorije sa više od 1500 podataka po kategoriji. Možemo da zaključimo da ljudi mnogo manje gledaju filmove koje pripadaju drugim kategorijama (npr. Biografija, Horor, Animacioni). Imamo po 1 film za žanrove Music i Sport, 2 za Musical, 3 za Western. Ovde je takodje njihov uticaj na model minimalan zato što ne daju dovoljno informacija za pouzdan zaključak. Mogu da razmotrim da neke od njih spojim u jednu kategoriju kao npr. Music i Musical iako je to i dalje malo podataka.

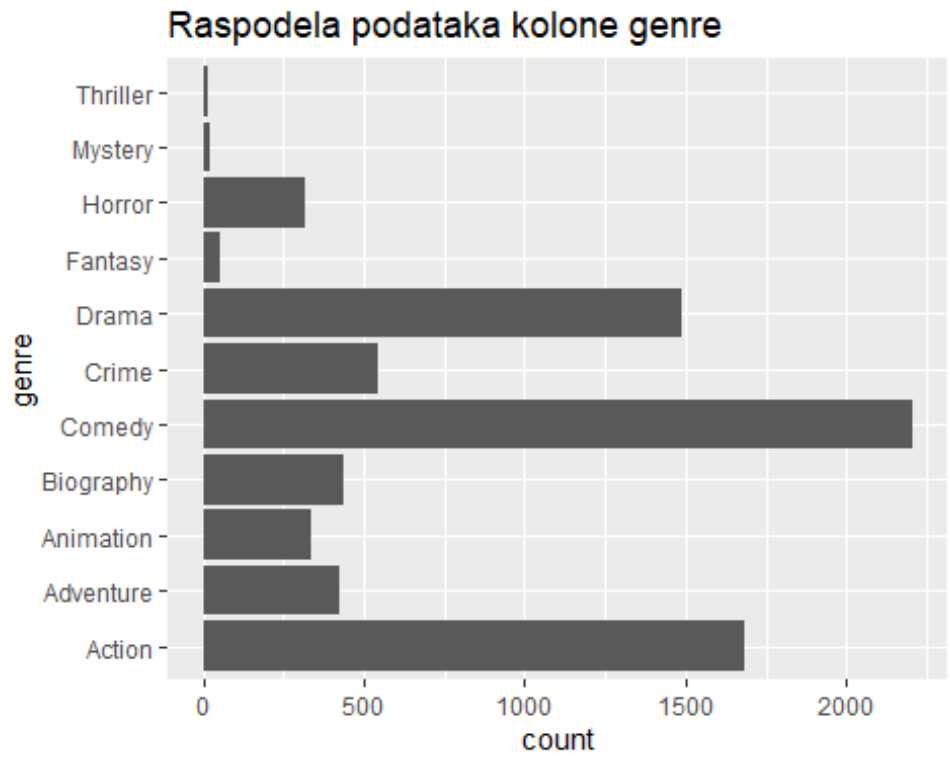
Ovde bih žanrove Music i Musical mogla da spojim u jedan jer su oba usmerena na muziku, ali ih je mnogo malo i zato ću da ih obrišem kao i Sport i Western. Žanr Fantasy ću pridružiti Sci-Fi, i žanr Romance - Drami, takođe Family i Drama često idu zajedno (porodični često imaju elemente drame koji mogu uključivati međuljudske odnose, odrastanje i slično).

```
movies$genre[which(movies$genre == "Sci-Fi")] <- "Fantasy"
movies$genre[which(movies$genre == "Romance")] <- "Drama"
movies$genre[which(movies$genre == "Family")] <- "Drama"
movies = movies[movies$genre != "Music", ]
```

```
movies = movies[movies$genre != "Musical", ]
movies = movies[movies$genre != "Western", ]
movies = movies[movies$genre != "Sport", ]
```

Sada raspodela izgleda ovako:

```
ggplot(data = movies, mapping=aes(y=genre)) +
  geom_bar() +
  labs(title = "Raspodela podataka kolone genre", y="genre", x="count")
```



Kolona year

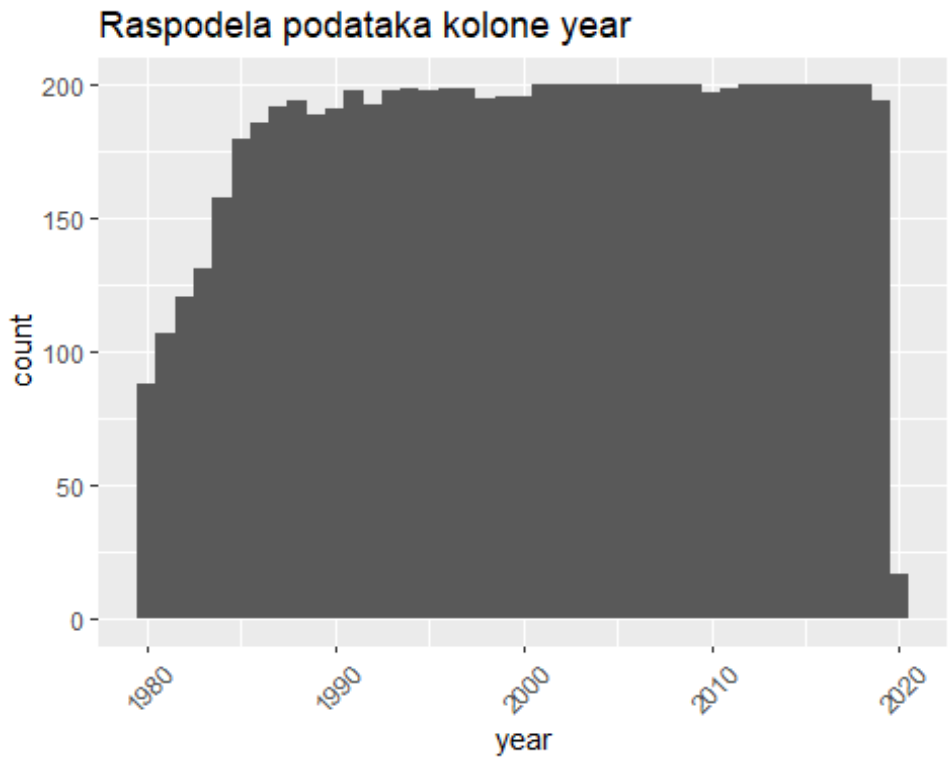
Tabela učestalosti za year

```
xtabs(~year, movies)
```

```
## year
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
1995
##   88  107  121  131  158  180  186  192  194  189  191  198  193  198  199
198
## 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
2011
##  199  199  195  196  196  200  200  200  200  200  200  200  200  200  197
199
## 2012 2013 2014 2015 2016 2017 2018 2019 2020
##  200  200  200  200  200  200  200  194   17
```

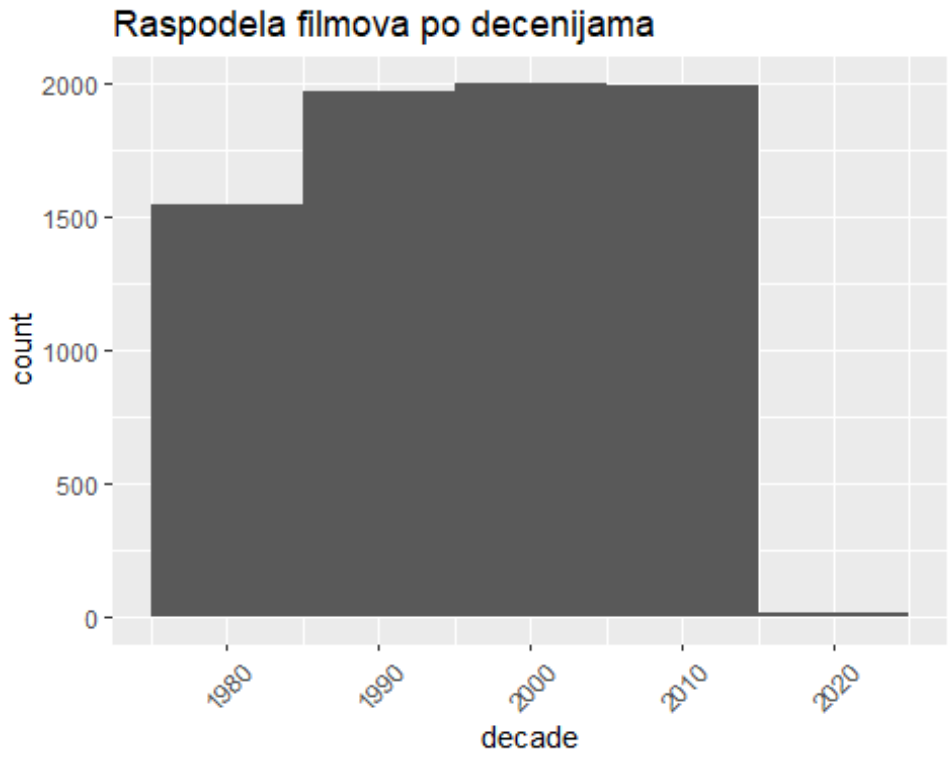
Grafik raspodele filmova prema godini (year)

```
ggplot(data = movies, mapping=aes(x=year)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Raspodela podataka kolone year") +
  theme(axis.text.x = element_text(angle=45, vjust=0.5))
```



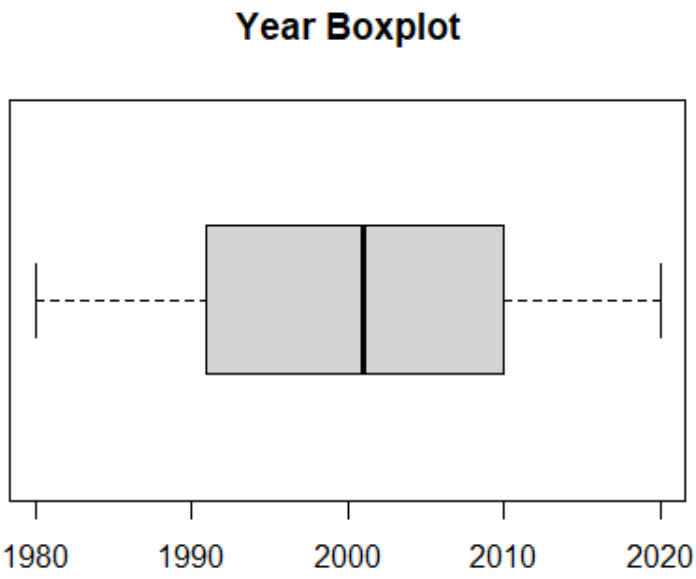
Sa grafika mozemo da vidimo da je pre 1985. godine broj filmova rastao iz godine u godinu, što ukazuje na povećanu produkciju tokom tog perioda. Izmedju 1985 - 2020. godine je snimljen približno isti broj filmova (200 po godini). Nakon 2020. se moze uočiti pad u broju filmova, može da bude rezultat smanjenja produkcije filmova zbog pandemije COVID-19 koji je značajno uticao i na ovu industriju.

```
movies$decade <- floor(movies$year / 10) * 10
ggplot(data = movies, mapping = aes(x = decade)) +
  geom_histogram(binwidth = 10) +
  labs(title = "Raspodela filmova po decenijama") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```



Proveriću outliere za ovu numeričku kolonu

```
boxplot(movies$year, main = "Year Boxplot", horizontal = TRUE)
```



Sa grafika jasno možemo da vidimo da nema odstupajućih vrednosti u koloni year.

Kolona released

Ova kolona sadrži informacije o datumu(YYY-MM-DD) kada je objavljen film. Pregledanjem podataka uočila sam da se godina u koloni year(release year) ne poklapa sa godinom u koloni released(release date). Pošto na sajtu sa kog sam uzela dataset nije pisalo ništa o tome, potražila sam informaciju zašto je to tako, pošto nemam dovoljno domenskog znanja da sama zaključim. Zapravo kolona released označava datum kada je film prvi put prikazan u nekoj specifičnoj zemlji, dok kolona year označava godinu kada je film prvi put prikazan, nebitno da li je to festival ili neka posebna projekcija.

Kod promenljive released imamo ogroman broj jedinstvenih vrednosti, tačnije 3415 što je skoro polovina svih podataka. Pošto je u pitanju datum objavljivanja, normalno je da je većina filmova objavljena drugog datuma pogotovo što je period od 30 godina u pitanju.

Kolona score

Statistika za score

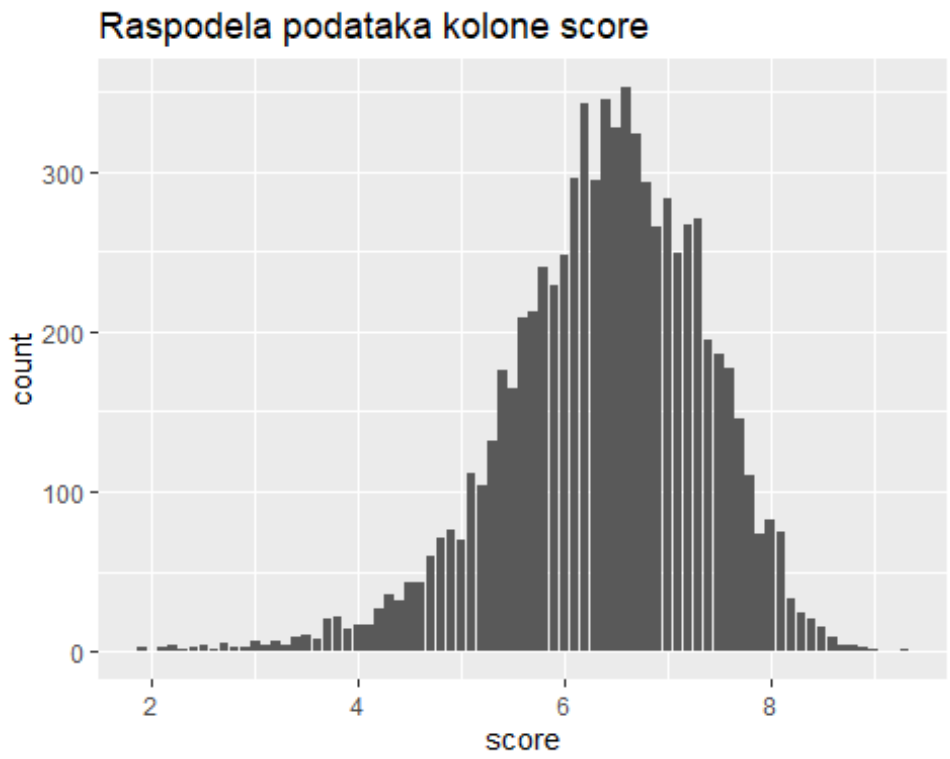
```
summary(movies$score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.900   5.800   6.500   6.395   7.100   9.300
```

Raspon vrednosti u ovoj koloni je od 1.9 do 9.3.

Grafik raspodele filmova prema oceni na IMDb-u (score)

```
ggplot(data = movies, mapping=aes(x=score)) +
  geom_bar() +
  labs(title = "Raspodela podataka kolone score")
```

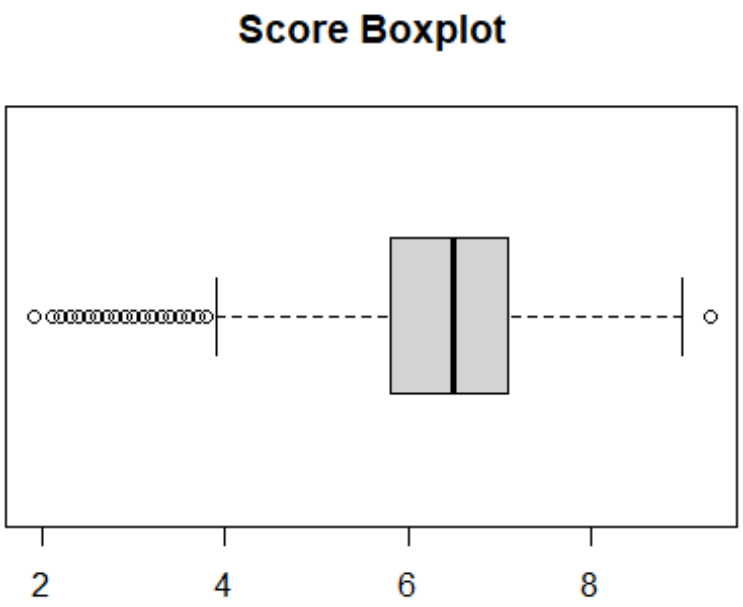


Možemo da vidimo da mali broj filmova ima najviše i najniže ocene. Najveći je broj onih filmova koji ima ocenu izmedju 6 i 7.

Zbog malog broj filmova u rasponu 0-5 ovo bi moglo da predstavlja kategoriju za sebe. Isto i za filmove za ocenom preko 8. Ovde bi mogao da se iskoristi Feature Engineering (FE) pristup za kreiranje nove promenljive na osnovu ovih ocena ali sačekaću da vidim u nastavku analize kakav odnos score ima sa gross.

Proveriću outliere za ovu numeričku kolonu

```
boxplot(movies$score, main = "Score Boxplot", horizontal = TRUE)
```



Filmovi sa ocenama nižim od 4 i višim od 9 izlaze iz okvira uobičajenih vrednosti. Ovi filmovi se retko pojavljuju pa su prepoznati kao izuzeci. Ali to ne znači da su ove ocene nemoguće u stvarnosti. Za konačno kreiranje modela mislim da je značajno imati raznolike podatke koji u ovom slučaju ne moraju nužno da ukazuju na niže prihode zbog niskih ocena. Isto tako skroz je realno da ima filmova koji se ljudima manje ili vise sviđaju.

Kolona votes

Statistika za votes

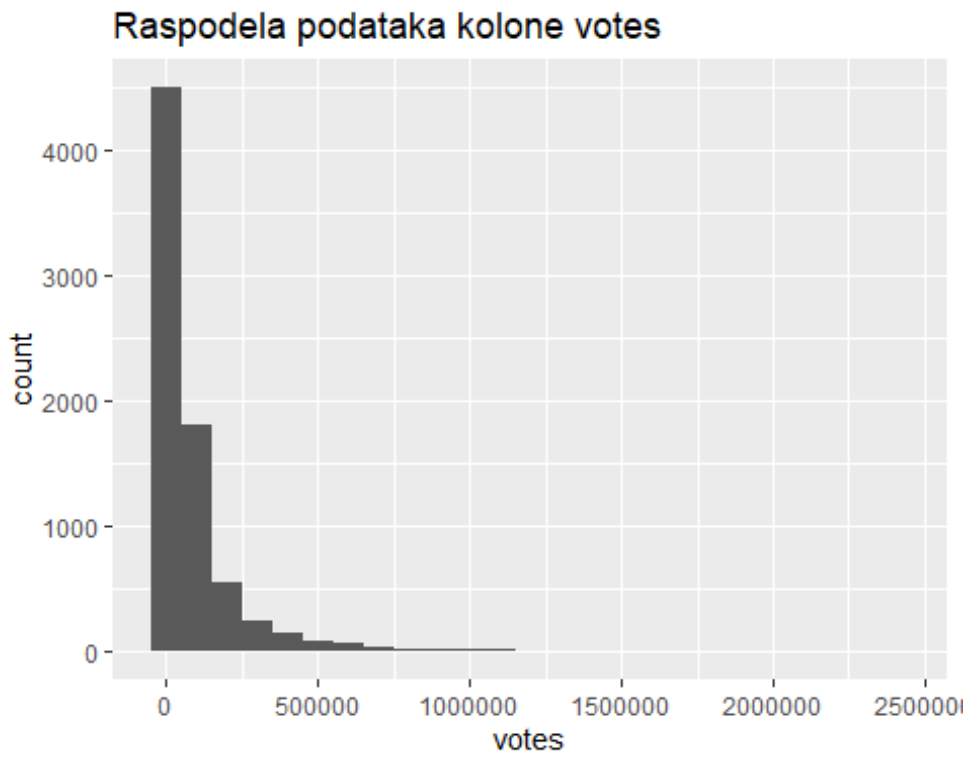
```
summary(movies$votes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      29    9800   34000   89770   95000 2400000
```

Raspon vrednosti u ovoj koloni je od 7 do 2400000.

Grafik raspodele filmova prema broju glasova (votes)

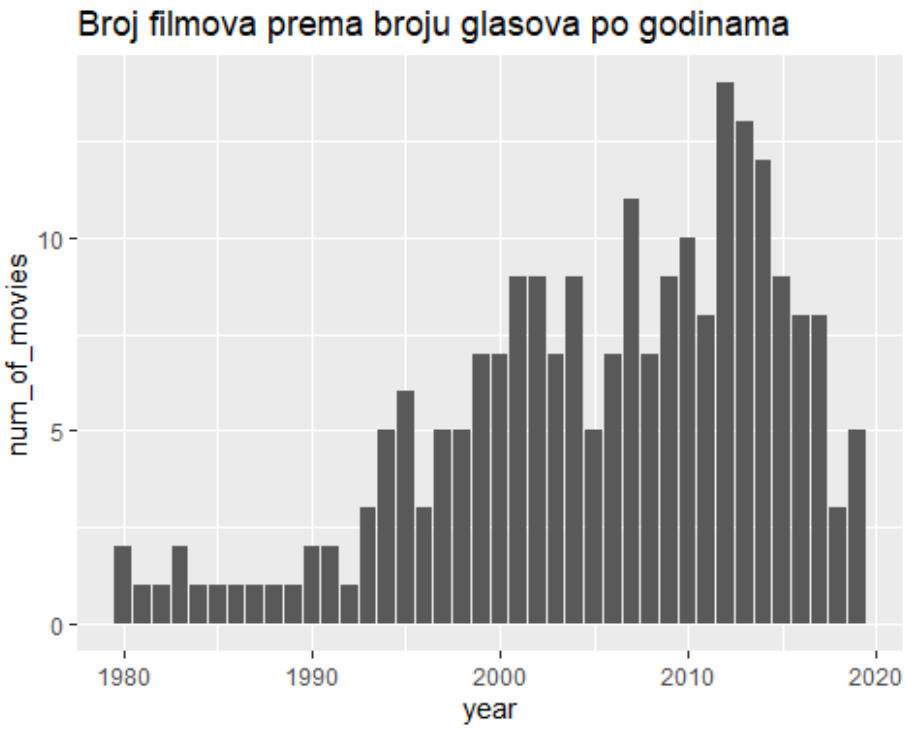
```
ggplot(data=movies, mapping=aes(x=votes)) + geom_histogram(binwidth = 100000)
+ labs(title = "Raspodela podataka kolone votes")
```



Najveći broj filmova ima manje od 500 000 glasova. Distribucija je asimetrična što u ovom slučaju ukazuje na to da postoji veći broj filmova sa manjim brojem glasova i mali broj filmova koji imaju baš veliki broj glasova. Na osnovu ovog grafika se vidi da postoje izuzeci - veliki broj glasova.

Proveravam iz kojih godina su ovi filmovi i da li je starost filma povezana sa brojem glasova.

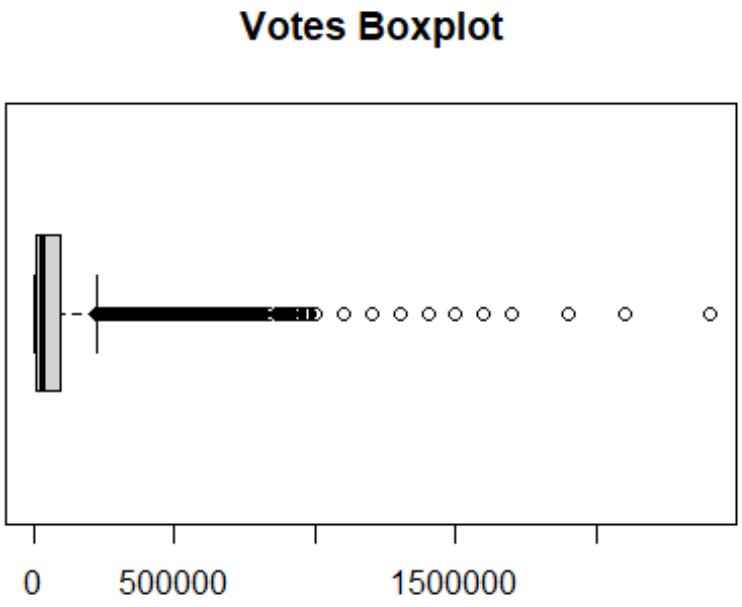
```
movies_out = movies %>% filter(votes > 500000)
movies1 <- movies_out %>% group_by(year) %>% summarise(num_of_movies=n())

ggplot(data=movies1, mapping=aes(x=year, y=num_of_movies)) +
  geom_bar(stat="identity") +
  labs(title = "Broj filmova prema broju glasova po godinama")
```



Sa grafika možemo da vidimo da to što filmovi dugo postoje nema veze sa njihovim brojem glasova, zapravo stariji filmovi imaju manji broj glasova u odnosu na novije.

Proveriću outliere za ovu numeričku kolonu

```
boxplot(movies$votes, main = "Votes Boxplot", horizontal = TRUE)
```



```
movies %>% transmute(name, votes, z_score = (votes-mean(votes))/sd(votes))
%>% filter(abs(z_score) > 3.0)
```

```
##                                                      name   votes
z_score
## 1                                             The Shining  927000
5.089351
## 2           Star Wars: Episode V - The Empire Strikes Back 1200000
6.748861
## 3            Indiana Jones and the Raiders of the Lost Ark  905000
4.955617
## 4                                           Blade Runner  710000
3.770252
## 5                                               Scarface  766000
4.110665
## 6              Star Wars: Episode VI - Return of the Jedi  973000
5.368975
## 7                                          The Terminator  812000
4.390289
## 8                                      Back to the Future 1100000
6.140982
## 9                                                 Aliens  668000
3.514943
## 10                                      Full Metal Jacket  691000
3.654755
## 11                                              Die Hard  810000
```

Iako se u ovom skupu podataka nalazi 175 filmova sa ekstremnim vrednostima broja glasova, te podatke ne treba zanemariti. Popularnost filmova može da varira, pa veći broj glasova može da ukazuje na veću popularnost koja može doprineti visokom prihodu.

Kolona director

```
length(xtabs(~director, movies))
```

```
## [1] 2868
```

Tabela učestalosti za director

Veliki je broj jedinstvenih vrednosti za direktora pa sam filtrirala sve one veće od 15 kako bi bilo moguće videti nešto sa grafika.

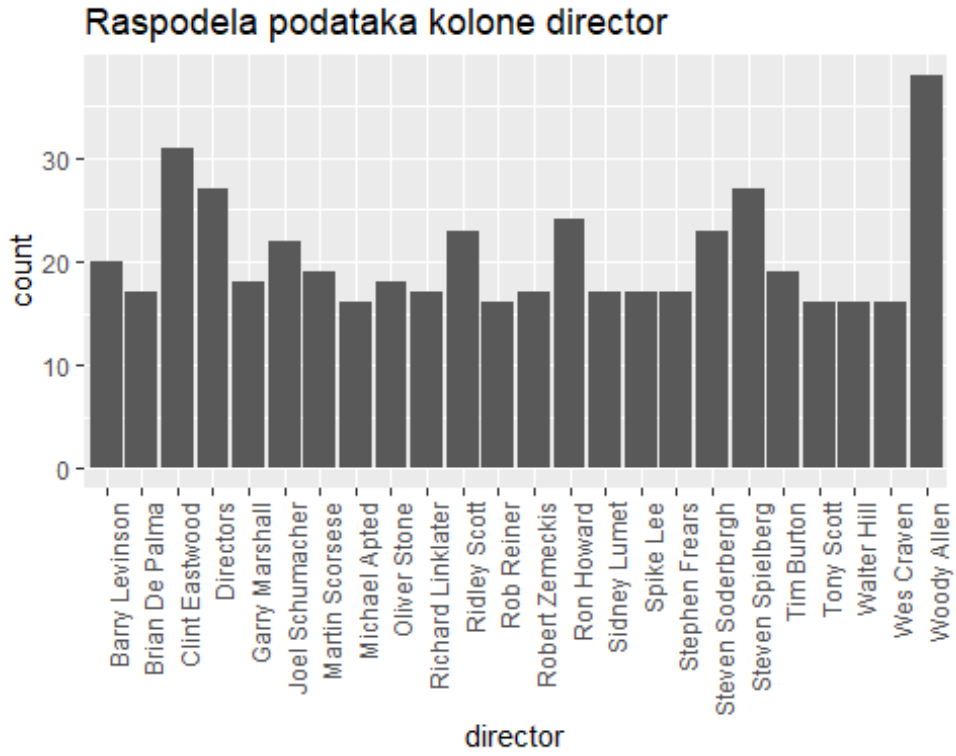
```
movies_by_director = movies %>% group_by(director) %>% summarise(count = n())
%>% arrange(desc(count)) %>% filter(count > 15)
movies_by_director
```

```
## # A tibble: 24 × 2
##    director          count
##    <chr>             <int>
##  1 Woody Allen          38
```

```
##  2 Clint Eastwood       31
##  3 Directors            27
##  4 Steven Spielberg     27
##  5 Ron Howard           24
##  6 Ridley Scott         23
##  7 Steven Soderbergh    23
##  8 Joel Schumacher      22
##  9 Barry Levinson       20
## 10 Martin Scorsese      19
## # ℹ 14 more rows
```

Grafik raspodele filmova prema direktoru filma (director)

```
  ggplot(data = movies_by_director, mapping=aes(x=director, y=count)) +
  geom_col()+
  theme(axis.text.x = element_text(angle= 90, hjust=1)) +
  labs(title = "Raspodela podataka kolone director", x = "director", y =
"count")
```



Ovde na filtriranom grafiku možemo da vidimo direktore sa najvećim brojem filmova. Medju njima se izdvaja Woody Allen sa 38 filmova. Ove informacije mi ne znače puno, baš zbog velikog broja različitih direktora. Ne deluje kao da bi nesto moglo da se zaključi sa grafika korelacije gross~director zbog velikog broja različitih podataka.

Kolona writer

```
length(xtabs(~writer, movies))
```

```
## [1] 4426
```

Tabela učestalosti za writer

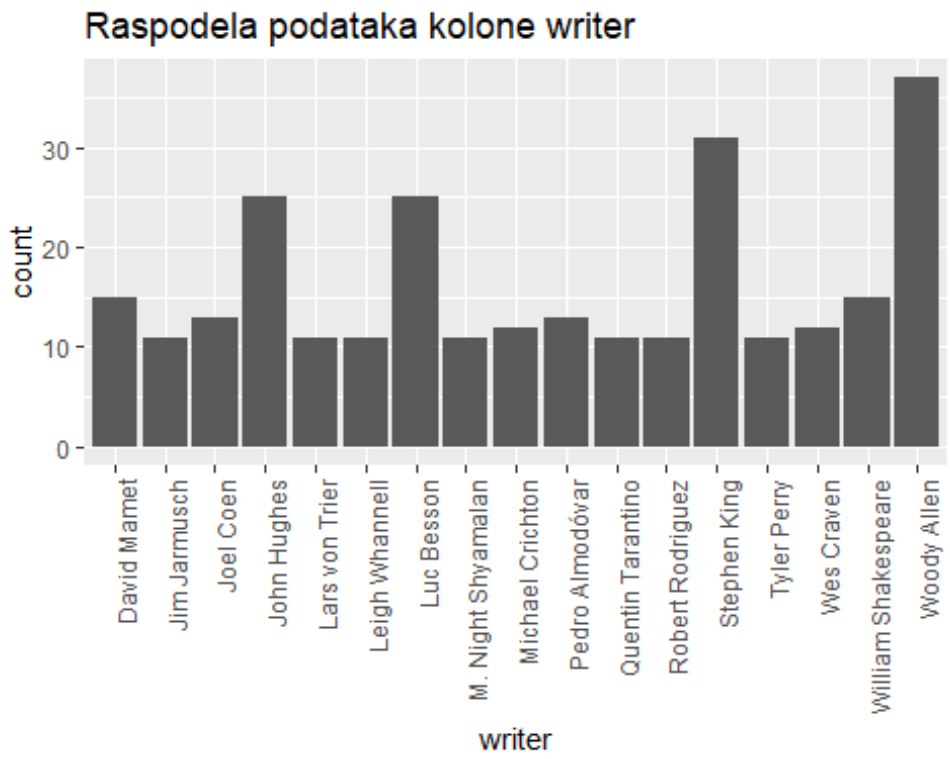
```
movies_by_writer = movies %>% group_by(writer) %>% summarise(count = n()) %>%
arrange(desc(count)) %>% filter(count > 10)
movies_by_writer
```

```
## # A tibble: 17 × 2
##    writer              count
##    <chr>               <int>
##  1 Woody Allen            37
##  2 Stephen King           31
##  3 John Hughes            25
##  4 Luc Besson             25
##  5 David Mamet            15
##  6 William Shakespeare    15
##  7 Joel Coen              13
##  8 Pedro Almodóvar        13
##  9 Michael Crichton       12
## 10 Wes Craven             12
## 11 Jim Jarmusch           11
## 12 Lars von Trier         11
## 13 Leigh Whannell         11
## 14 M. Night Shyamalan     11
## 15 Quentin Tarantino      11
## 16 Robert Rodriguez       11
## 17 Tyler Perry            11
```

Veliki je broj pisaca i filtrirala sam sve one koji su napisali vise od 10 filmova kako bi bilo moguće videti nešto sa grafika.

Grafik raspodele filmova prema piscu (writer)

```
  ggplot(data = movies_by_writer, mapping=aes(x=writer, y=count)) +
  geom_col()+
  theme(axis.text.x = element_text(angle= 90, hjust=1)) +
  labs(title = "Raspodela podataka kolone writer", x = "writer", y = "count")
```



Ovde takođe možemo videti da Woody Allen dominira po broju filmova. Sve što sam zaključila za director kolonu važi i ovde. Previše je različitih vrednosti da bi moglo nešto da se vidi za kasnije.

Kolona star

```
length(xtabs(~star, movies))
```

```
## [1] 2721
```

Tabela učestalosti za star

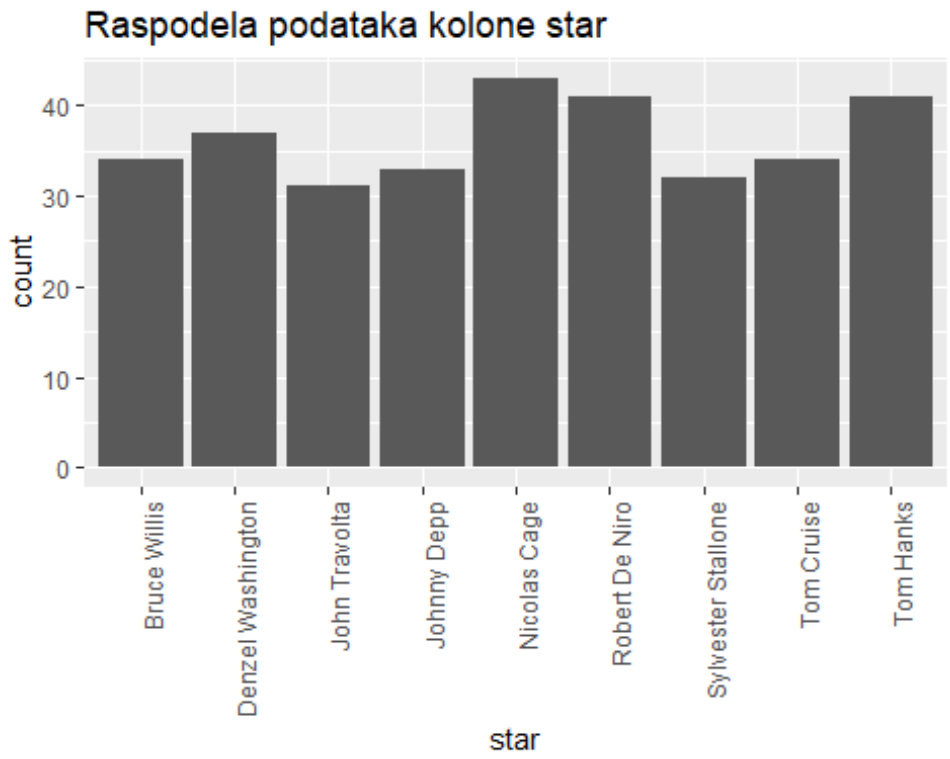
```
movies_by_star = movies %>% group_by(star) %>% summarise(count = n()) %>%
arrange(desc(count)) %>% filter(count > 30)
movies_by_star
```

```
## # A tibble: 9 × 2
##   star               count
##   <chr>              <int>
## 1 Nicolas Cage          43
## 2 Robert De Niro        41
## 3 Tom Hanks             41
## 4 Denzel Washington     37
## 5 Bruce Willis          34
## 6 Tom Cruise            34
## 7 Johnny Depp           33
## 8 Sylvester Stallone    32
## 9 John Travolta         31
```

Veliki je broj zvezda i filtrirala sam sve one koji su glumili u više od 30 filmova kako bi bilo moguće videti nešto sa grafika.

Grafik raspodele filmova prema zvezdi filma (star)

```
  ggplot(data = movies_by_star, mapping=aes(x=star, y=count)) +
  geom_col()+
  theme(axis.text.x = element_text(angle= 90, hjust=1)) +
  labs(title = "Raspodela podataka kolone star", x = "star", y = "count")
```



Ono što možemo da vidimo da se najveće svetske zvezde nalaze u ovih top 10 glumaca sa najvećim brojem filmova. Ovde isto ima veliki broj jedinstvenih vrednosti, slično kao kod prethodne sve kolone: director i writer.

Kolona country

Tabela učestalosti za country

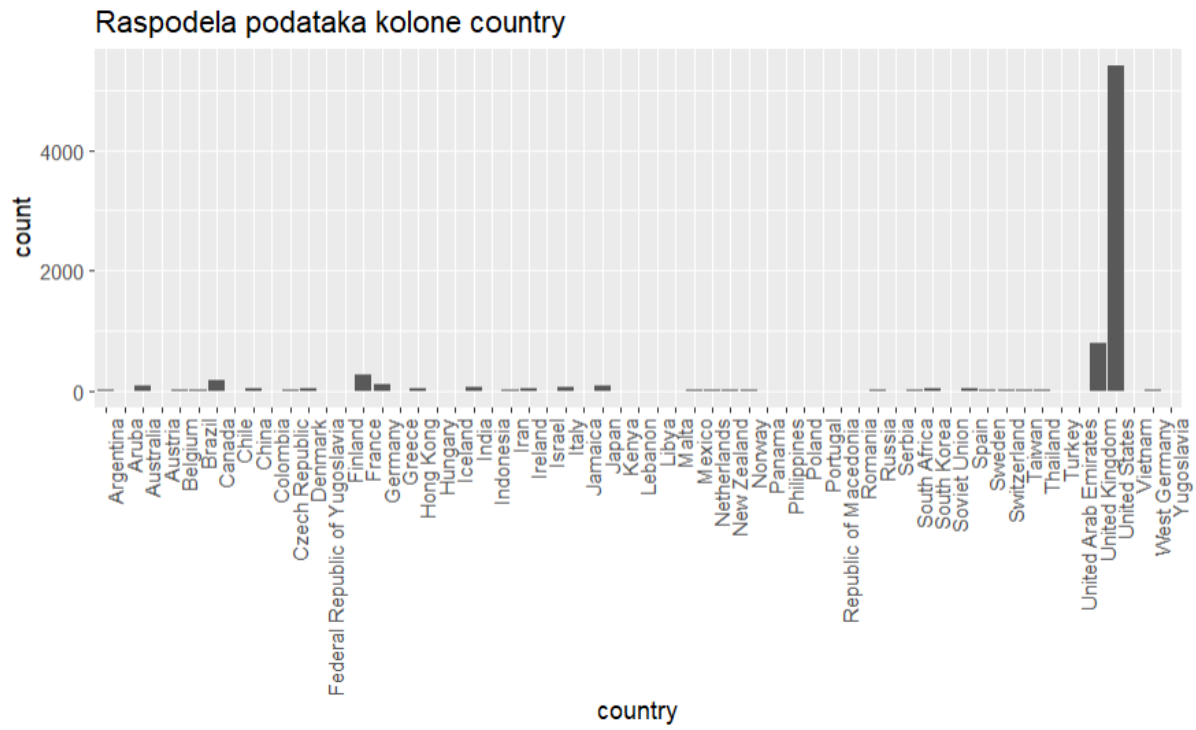
```
by_country = movies %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  mutate(percentage = round((count/nrow(movies)*100),2)) %>%
  arrange(desc(count))

by_country
```

```
## # A tibble: 59 × 3
##    country        count percentage
```

```
##    <chr>          <int>      <dbl>
##  1 United States   5415      72.1
##  2 United Kingdom   794      10.6
##  3 France           261       3.47
##  4 Canada           184       2.45
##  5 Germany          116       1.54
##  6 Australia         86       1.14
##  7 Japan             77       1.02
##  8 India             60       0.8
##  9 Italy             56       0.75
## 10 Spain             44       0.59
## # ℹ 49 more rows
```

Grafik raspodele filmova prema zemlji iz kojih potiču (country)

```
  ggplot(data = by_country, mapping=aes(x=country, y = count)) +
  geom_col()+
  theme(axis.text.x = element_text(angle= 90, hjust=1)) +
  labs(title = "Raspodela podataka kolone country", x = "country", y =
"count")
```



Najveći broj snimljenih filmova potiče iz Amerike, što ukazuje na dominantnost američke filmske industrije, verovatno zbog Hollywood-a. Nakon toga Ujedinjeno Kraljevstvo i to sa mnogo manjim brojem filmova u odnosu na Ameriku. Zemlje poput Francuske, Kanade i Nemačke su takođe prisutne ali sa znatno manjim udelom doprinose filmskoj industriji u okviru ovog skupa podataka. Ovde imamo 12 država iz kojih dolazi samo jedan film, 9 iz

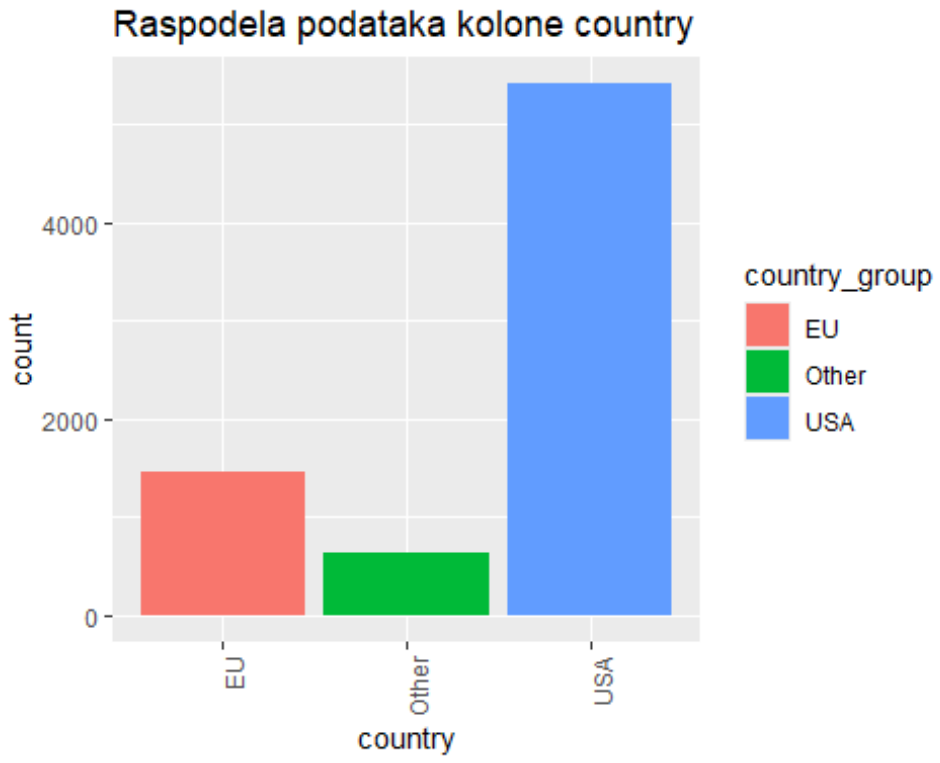
kojih dolaze samo dva filma, ovakvi podaci nisu od neke koristi pri kreiranju modela, pogotovo što je ovde u više kolona slična situacija.

Napraviću novu kolonu country_group i grupisaću države prema USA, EU i Ostale. Na taj način će ova kolona možda biti korisnija za model i rešiću problem velikog broja kategorija koji ima po jednu ili dve vrednosti.

```
movies$country_group = ifelse(movies$country %in% c("United States"),
"USA",ifelse(movies$country %in% c("United Kingdom", "France", "Germany",
"Italy", "Spain", "Ireland", "Denmark", "Sweden", "Norway", "Netherlands",
"West Germany", "Switzerland", "Belgium", "Czech Republic", "Russia",
"Austria", "Hungary", "Poland", "Finland", "Yugoslavia", "Federal Republic of
Yugoslavia", "Portugal", "Greece", "Malta", "Republic of Macedonia",
"Romania", "Serbia", "Soviet Union"), "EU", "Other"))
View(movies)
```

Sada grafik izgleda ovako:

```
  ggplot(data = movies, mapping=aes(x=country_group, fill=country_group)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle= 90, hjust=1)) +
  labs(title = "Raspodela podataka kolone country", x = "country", y =
"count")
```



Kolona company

```
length(xtabs(~company, movies))
```

```
## [1] 2306
```

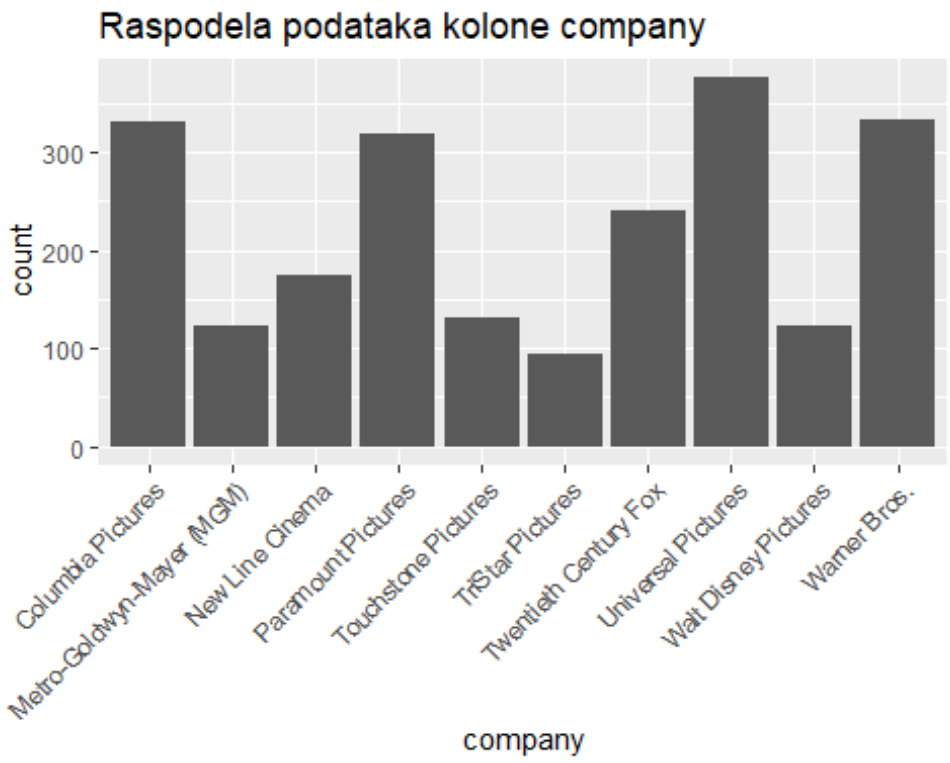
Tabela učestalosti za company

```
movies_by_company <- movies %>% group_by(company) %>% summarise(count = n())
%>% arrange(desc(count)) %>% filter(count > 90)
movies_by_company
```

```
## # A tibble: 10 × 2
##    company                   count
##    <chr>                     <int>
##  1 Universal Pictures          376
##  2 Warner Bros.                333
##  3 Columbia Pictures           332
##  4 Paramount Pictures          319
##  5 Twentieth Century Fox       240
##  6 New Line Cinema             174
##  7 Touchstone Pictures         132
##  8 Metro-Goldwyn-Mayer (MGM)   124
##  9 Walt Disney Pictures        123
## 10 TriStar Pictures             94
```

Grafik raspodele filmova prema produkcijskoj kući(company). Filtrirala sam one sa najvećim brojem filmova kako bih videla nešto na grafiku.

```
  ggplot(data = movies_by_company, mapping=aes(x=company, y=count)) +
  geom_col()+
  theme(axis.text.x = element_text(angle= 45, hjust=1)) +
  labs(title = "Raspodela podataka kolone company", x = "company", y =
"count")
```



Najproduktivnije kompanije su: “Universal Pictures”, “Warner Bros.”,“Columbia Pictures”, “Paramount Pictures”.

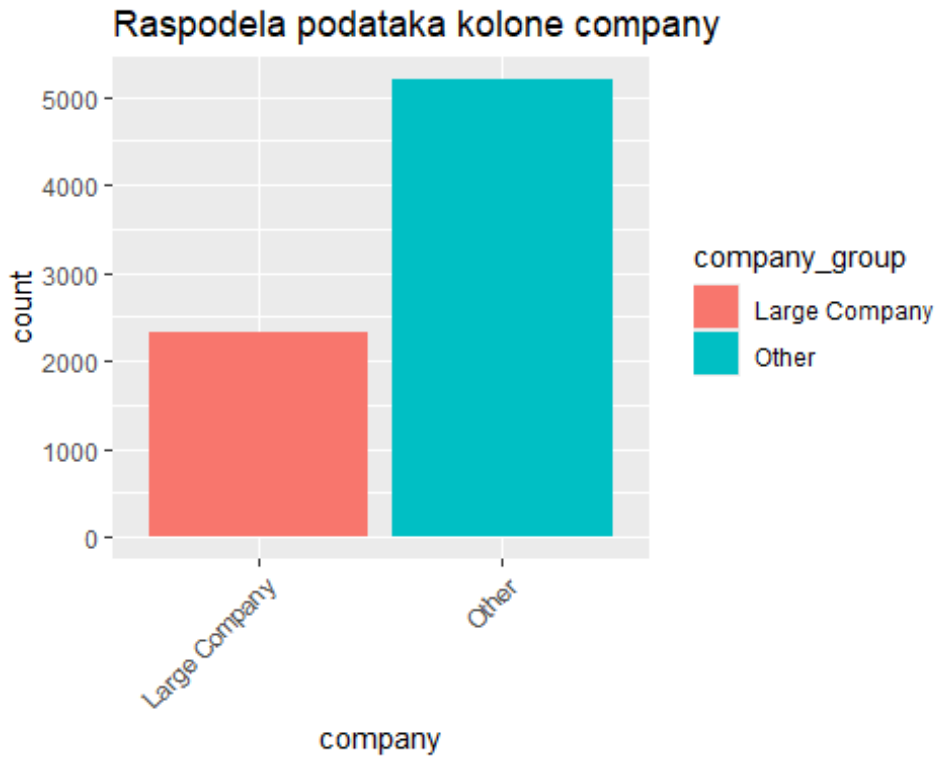
Na osnovu domenskog znanja je poznato da filmovi koji pripadaju velikim produkcijskim kućama često donose veće prihode. Pošto se u ovom skupu podataka nalazi veliki broj različitih produkcijskih kuća (2381) teško je doneti neke zaključke. Pošto ovo deluje kao jedan od važnih prediktora napraviću novu kolonu na osnovu kolone company koja će da ima kategorije vezane za to da li film pripada nekoj poznatoj produkcijskoj kući ili ne. Potražila sam informaciju o najpoznatijim produkcijskim kućama na svetu i to su: “Warner Bros.”, “Walt Disney”, “Pixar Animation”, “Universal Pictures”, “Marvel Studios”, “MGM Studios”, “Lionsgate”, “Sony Pictures”, “Paramount Pictures Studios”, “DreamWork Studios”, “20th Century Fox”, “Weinstein”, “Columbia Pictures”.

```
movies$company_group = ifelse(grepl("Warner Bros\\.|Walt Disney|New Line
Cinema|Pixar Animation Studios|Universal|Marvel|MGM|Lionsgate|Sony
Pictures|Paramount|Dreamworks|Twentieth Century Fox|The Weinstein
Company|Columbia Pictures", movies$company, ignore.case = TRUE), "Large
Company", "Other")
View(movies)
```

Sada grafički mozemo da vidimo raspodelu onih koji pripadaju navedenim velikim produkcijskim kućama i ostalih

```
  ggplot(data = movies, mapping=aes(x=company_group, fill=company_group)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle= 45, hjust=1)) +
```

```
  labs(title = "Raspodela podataka kolone company", x = "company", y =
"count")
```



Kategorija 'Other' sadrži mnogo veći broj filmova od kategorije 'Large'. Ali zato najveći broj filmova potiče upravo iz velikih produkcijskih kuća (mnogo je veliki broj onih kuća iz skupa ostalih koji imaju po 1, 2, 3 filma).

Kolona runtime

Statistika za runtime

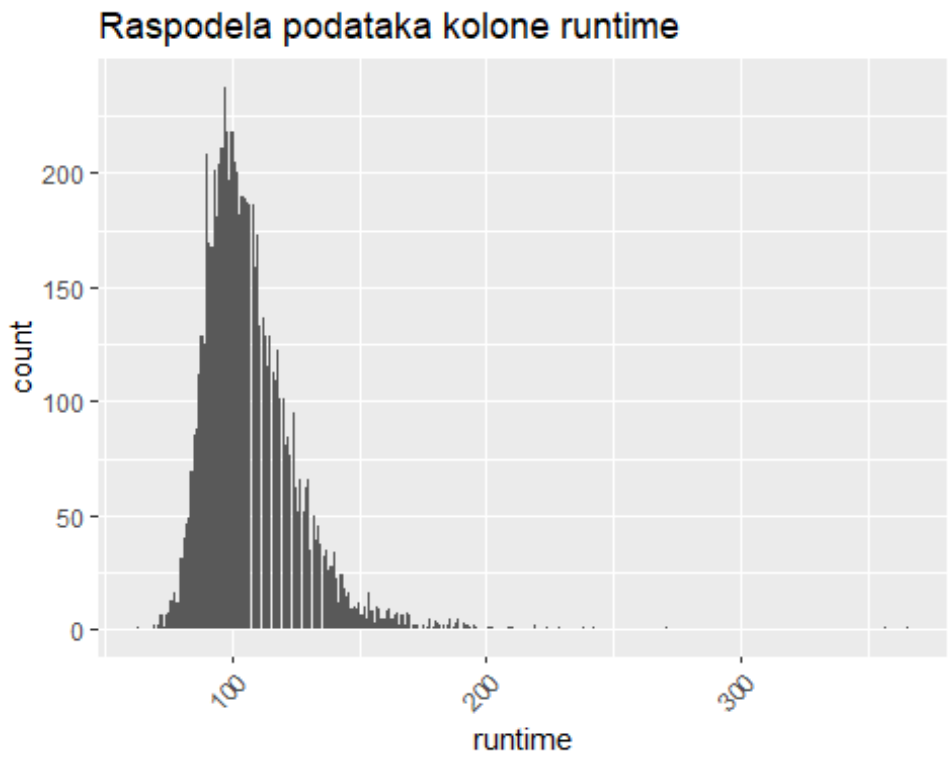
```
summary(movies$runtime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    63.0    95.0   104.0   107.4   116.0   366.0
```

Vreme trajanja filma iznosi između 63 i 366 minuta.

Grafik raspodele filmova prema dužini trajanja filma (runtime)

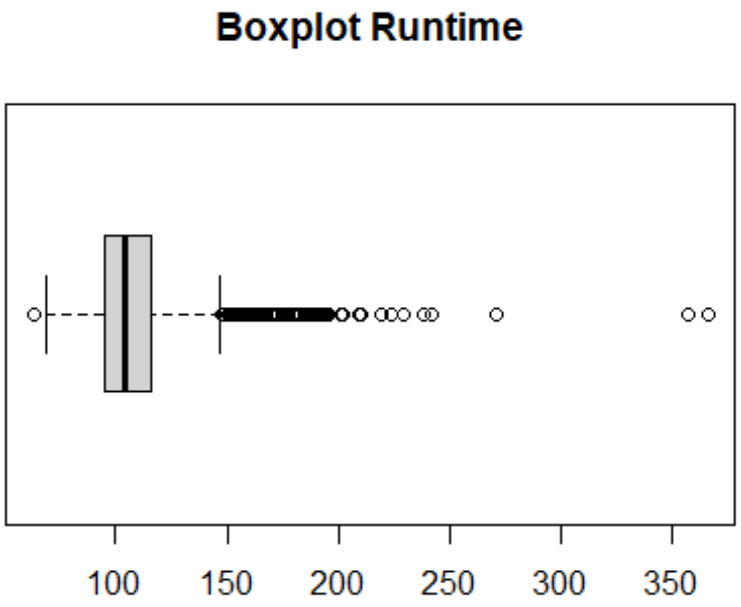
```
  ggplot(data = movies, mapping=aes(x=runtime)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle= 45, hjust=1)) +
  labs(title = "Raspodela podataka kolone runtime", x = "runtime", y =
"count")
```



Najčešća dužina trajanja filma je oko 100 minuta. Većina filmova traje između 95 i 116 minuta i to je neka standardna dužina trajanja filmova koja bi zadržala pažnju publike. Distribucija je asimetrična što u ovom slučaju ukazuje na to da postoji veći broj filmova kraćih trajanja i mali broj filmova koji dugo traju.

Proveriću outliere za ovu numeričku kolonu

```
boxplot(movies$runtime, main = "Boxplot Runtime", horizontal = TRUE)
```



```
movies %>% transmute(name, runtime, z_score = (runtime-
mean(runtime))/sd(runtime)) %>% filter(abs(z_score) > 3.0)
```

```
##                                                 name runtime   z_score
## 1                                       Heaven's Gate     219  6.023807
## 2                                   Lion of the Desert     173  3.541015
## 3                                                Reds     195  4.728437
## 4                                   Prince of the City     167  3.217172
## 5                                               Gandhi     191  4.512542
## 6                                  Fanny and Alexander     188  4.350621
## 7                                            Scarface     170  3.379093
## 8                                      The Right Stuff     193  4.620489
## 9                           Once Upon a Time in America     229  6.563544
## 10                                   A Passage to India     164  3.055251
## 11                                          Betty Blue     185  4.188700
## 12                                    The Last Emperor     163  3.001277
## 13                                       Little Dorrit     357 13.472182
## 14                          The Last Temptation of Christ     164  3.055251
## 15                                        The Big Blue     168  3.271146
## 16                     The Unbearable Lightness of Being     171  3.433067
## 17                                    Camille Claudel     175  3.648962
## 18                                  Dances with Wolves     181  3.972805
## 19                                                 JFK     189  4.404594
## 20                           The Beautiful Troublemaker     238  7.049308
## 21                     At Play in the Fields of the Lord     189  4.404594
## 22                                          Malcolm X     202  5.106253
```

U ovom skupu podataka filmovi sa dužinom većom od 164 minuta smatraju se ekstremnim vrednostima. Standardna dužina trajanja filma kreće se između 90 i 150 minuta ali trajanje značajno može da varira od žanra. Neki filmovi poput istorijskih ili dokumentaraca mogu da traju veoma dugo, čak i do 600 minuta. U ovom skupu podataka postoji mali broj filmova koji pripadaju ovim žanrovima zbog čega su ove vrednosti verovatno prepoznate kao izuzeci. Duže trajanje filma je karakteristika nekih žanrova.

Raspodela podataka izlazne kolone ~ gross

Statistika kolone gross

```
summary(movies$gross)
```

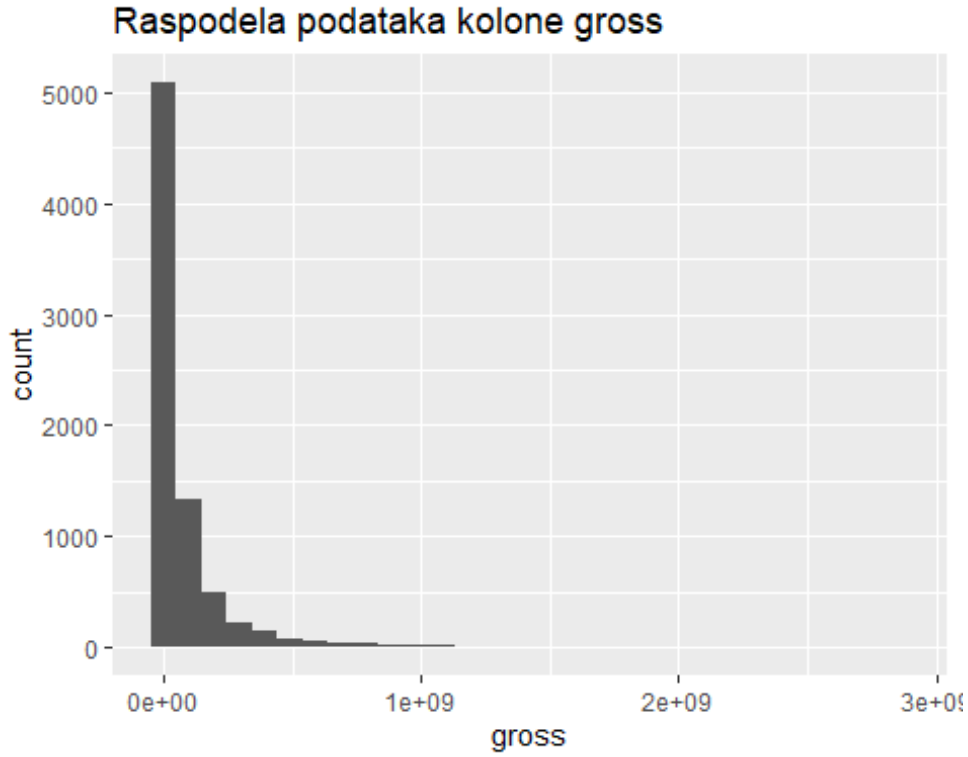
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 3.090e+02 4.617e+06 2.021e+07 7.823e+07 7.546e+07 2.847e+09
```

U skupu podataka prihodi se kreću od 309 do 2 847 000 000.

Grafički prikaz raspodele

```
ggplot(data=movies, mapping=aes(x=gross)) +
  geom_histogram() +
    labs(title = "Raspodela podataka kolone gross", x = "gross", y = "count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



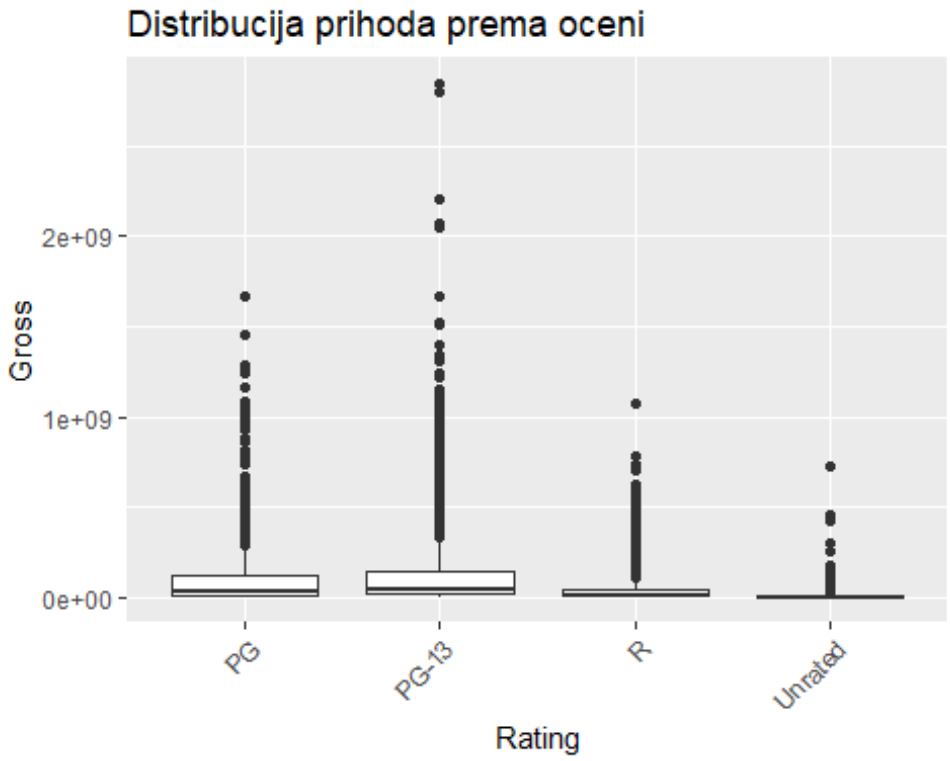
Analiza podataka

Analiza izmedju prediktora i odgovora

Kolona Rating

Grafikom ispod je prikazan odnos između kategorije filma i prihoda.

```
ggplot(data = movies, mapping = aes(x = rating, y = gross)) +
  geom_boxplot() +
  labs(title = "Distribucija prihoda prema oceni", x = "Rating", y = "Gross")
+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



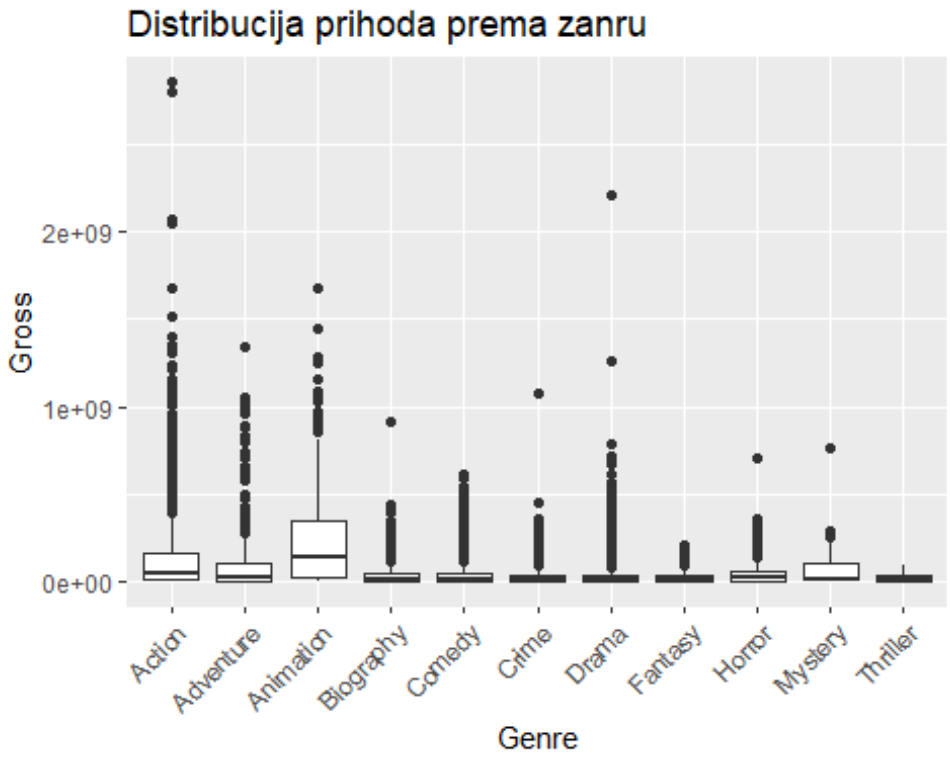
Filmove sa kategorijom PG-13 karakteriše najširi raspon prihoda. Filmovi sa kategorijom PG-13 i PG imaju medianu koja je viša u poređenju sa ostalim kategorijama, što znači da ti filmovi u proseku ostvaruju veće prihode. Medijana kod R je nešto niža nego kod PG-13, što može značiti da, iako neki filmovi sa kategorijom R ostvaruju veliki prihod, u proseku su filmovi sa kategorijom PG-13 uspešniji. Filmove iz kategorije PG-13 mogu da gledaju deca, pa se ovakvi filmovi puštaju u svako doba dana i dostupniji su široj populaciji. Za razliku od njih R filmovi se obično puštaju uveče pa imaju i manju gledanost.

Zaključak je da kolona rating moze biti značajan faktor u modelu predikcije prihoda.

Kolona genre

Grafikom ispod je prikazan odnos između žanra filma i prihoda.

```
ggplot(data = movies, mapping = aes(x = genre, y = gross)) +
  geom_boxplot() +
  labs(title = "Distribucija prihoda prema zanru", x = "Genre", y = "Gross")
+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



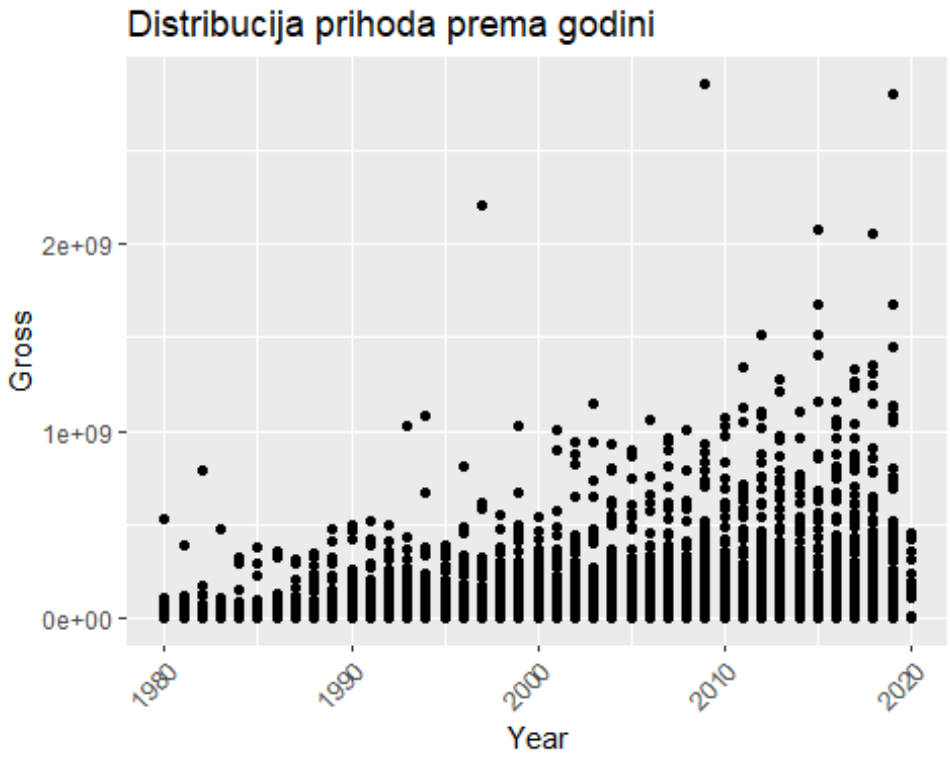
Filmovi sa žanrom Action imaju najširi raspon prihoda što sugeriše da mogu imati veoma visoke ali i niske prihode. Animation ima medijanu koja je viša u poređenju sa ostalim žanrovima, što znači da ovi filmovi u proseku ostvaruju veće prihode od drugih. Disney filmovi su Animacioni i logično je da takvi filmovi zarađuju više zbog popularnosti. Nasuprot tome, žanrovi poput Drama i Horor imaju niže medijane prihoda i manji raspon što ukazuje da filmovi iz ovih žanrova ostvaruju skromnije prihode.

Zaključak je da kolona genre takođe može biti značajan faktor u modelu predikcije prihoda.

Kolona year

Grafički prikaz odnosa između prihoda filmova prema godinama (Year)

```
ggplot(data = movies, mapping = aes(x = year, y = gross)) +
  geom_point() +
  labs(title = "Distribucija prihoda prema godini", x = "Year", y = "Gross")
+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

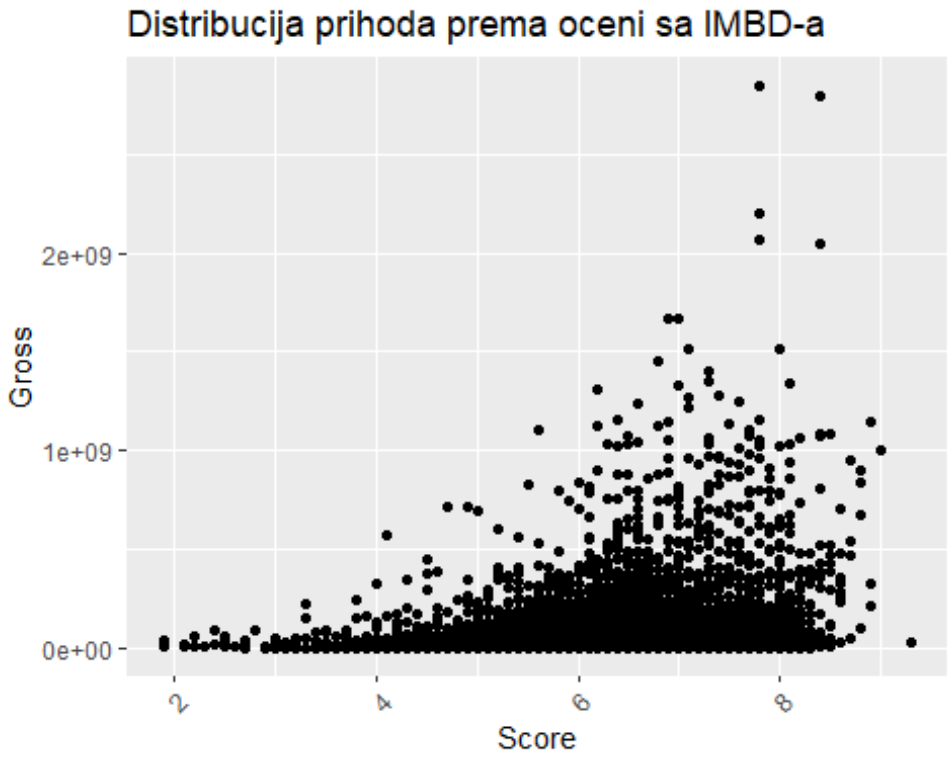


Kako se približavamo novijim godinama vidi se porast u prihodima filmova. Pre 2000. godine većina filmova ima niže prihode, dok su posle 2000. godine filmovi dostizali veće prihode. Ovo ukazuje na postojanje korelacije između godine proizvodnje i prihoda filma. Sa grafika možemo da vidimo i da se nakon 2000. pojavljuju filmovi sa izuzetno velikim prihodima, primer mogu da budu “blockbuster” filmovi sa velikim budžetima. Vidi se da nakon 2010. godine postoji značajan porast u prihodima, na šta su možda uticali neki dodatni faktori.

Kolona score

Grafikom ispod je prikazan odnos između ocene sa IMDb-a i prihoda.

```
ggplot(data = movies, mapping = aes(x = score, y = gross)) +
  geom_point() +
  labs(title = "Distribucija prihoda prema oceni sa IMDb-a", x = "Score", y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

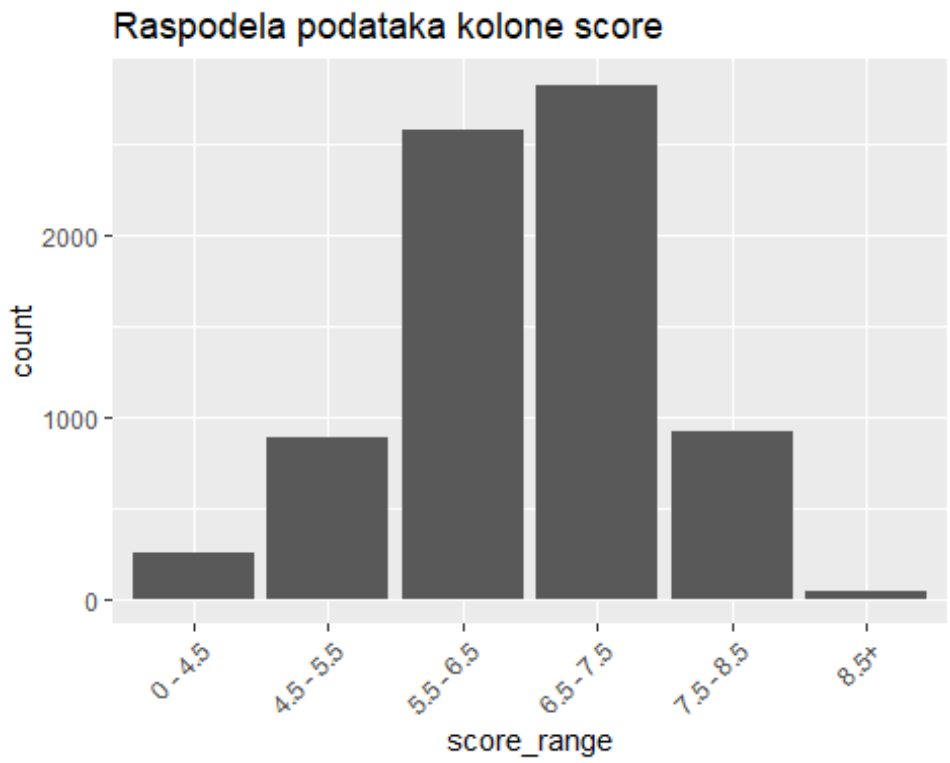


Ovde možemo da zaključimo da filmovi sa većim ocenama imaju i veće prihode. Isto tako i filmovi sa većim ocenama imaju veće skokove u prihodima (ekstremne vrednosti).

Iskoristicu Feature Engineering (FE) pristup da kreiram novu kolonu zasnovanu na postojećoj koloni score. Ovo može potencijalno da doprinese poboljšanju performansi modela jer omogućava transformacije koje bolje opisuju relacije u podacima.

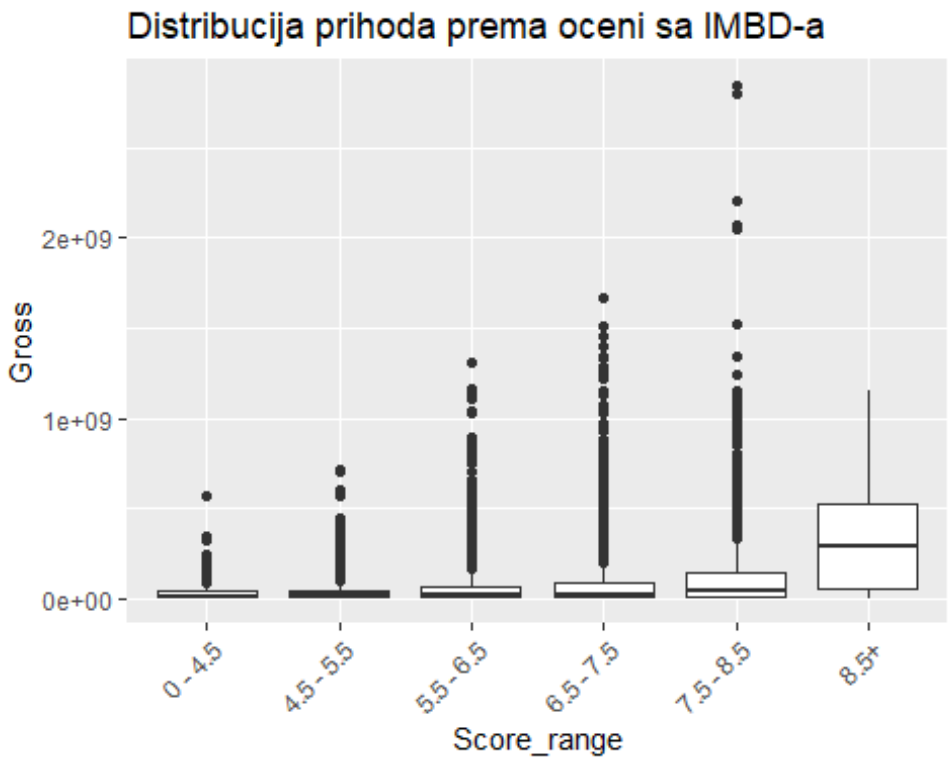
Na osnovu stanja sa grafika se vidi da se sa povećanjem score-a povećava i prihod. Nije bitno da li je score 1, 2, 3 ili 4 zato što se svakako očekuje mala zarada. To nije slučaj od 4 do 8 zato što se tu razlikuje zarada.

```
movies <- movies %>%
  mutate(score_range = case_when(
    score >= 0 & score < 4.5 ~ '0 - 4.5',
    score >= 4.5 & score < 5.5 ~ '4.5 - 5.5',
    score >= 5.5 & score < 6.5 ~ '5.5 - 6.5',
    score >= 6.5 & score < 7.5 ~ '6.5 - 7.5',
    score >= 7.5 & score < 8.5 ~ '7.5 - 8.5',
    score >=8.5 ~ '8.5+',
  ))
View(movies)

  ggplot(data = movies, mapping=aes(x=score_range)) +
  geom_bar() +
  labs(title = "Raspodela podataka kolone score", x="score_range") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Nakon transformacije kolone, grafik izgleda ovako:

```
ggplot(data = movies, mapping = aes(x = score_range, y = gross)) +
  geom_boxplot() +
  labs(title = "Distribucija prihoda prema oceni sa IMDb-a", x =
"Score_range", y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

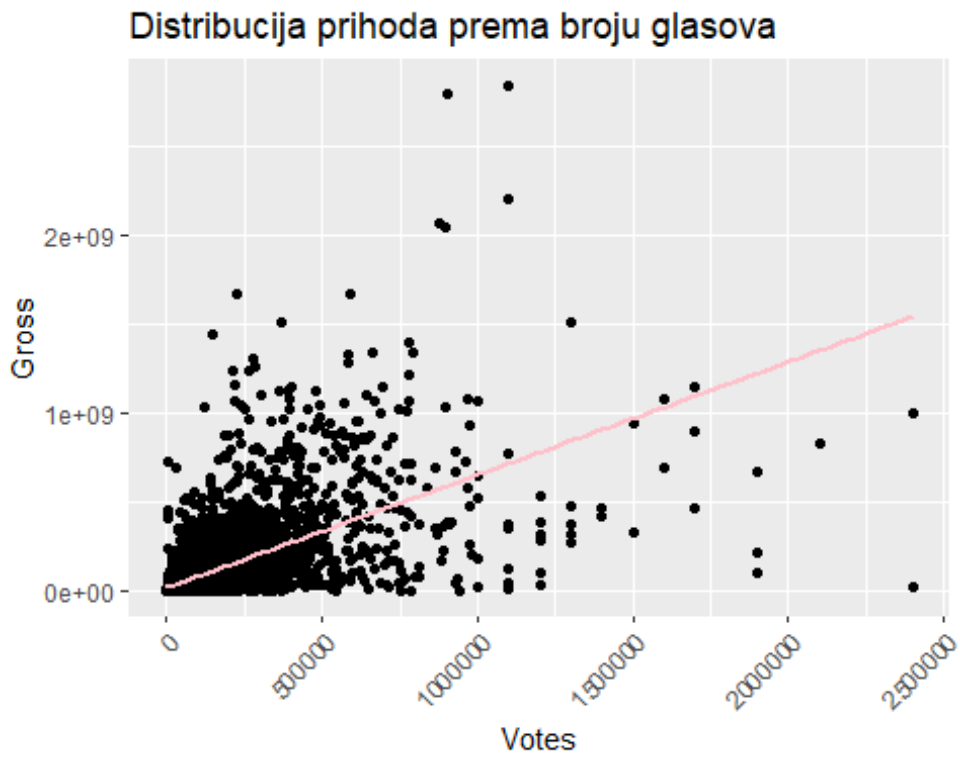


Ovde sad možemo da zaključimo da najveći raspon prihoda imaju filmovi sa ocenom između 7.5-8.5, dok zapravo filmovi sa ocenom većom od 8.5 u proseku imaju najveći prihod.

Kolona votes

Grafikom ispod je prikazan odnos između broja glasova filma i prihoda.

```
ggplot(data = movies, mapping = aes(x = votes, y = gross)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col="pink", formula="y~x") +
  labs(title = "Distribucija prihoda prema broju glasova", x = "Votes", y =
"Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Roze linija na grafiku koja predstavlja linearnu regresiju pokazuje da veći broj glasova može biti povezan sa većim prihodima. Ali rasipanje podataka je značajno i većina podataka se nalazi u donjem delu grafika. Na osnovu domenskog znanja mogu da kažem da je očekivano da su filmovi sa više glasova popularniji i obično imaju i veće prihode. Isto tako neki filmovi sa malim brojem glasova mogu ostvariti velike prihode.

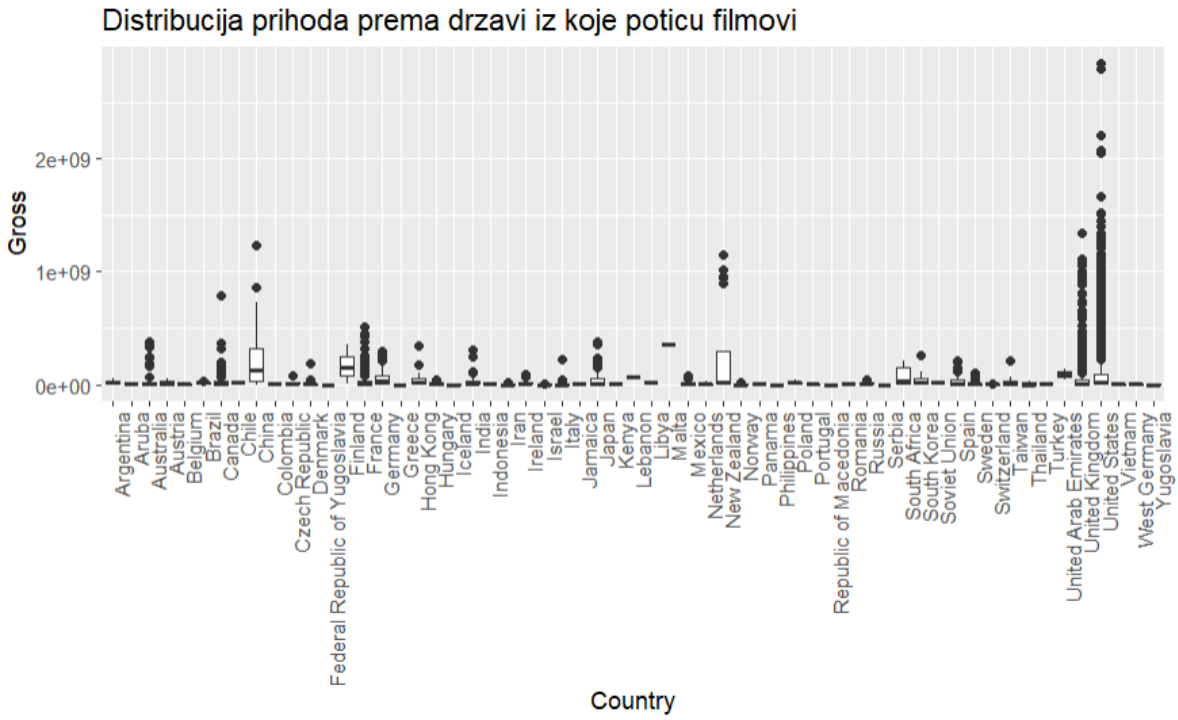
Zaključak je da može biti jedan od faktora koji će učestvovati u kreiranju modela.

Kolone director, writer, star i releised imaju ogroman broj jedinstvenih kategorija i na grafiku odnosa sa prihodom ne može ništa da se zaključi. Na osnovu domenskog znanja, kolona star bi mogla da bude značajna zato što filmovi u kojima glume najpoznatije svetske zvezde su sigurno i među najpopularnijim što dovodi do velkih prihoda.

Kolona country

Grafikom ispod je prikazan odnos između zemlje porekla filma i prihoda.

```
ggplot(data = movies, mapping = aes(x = country, y = gross)) +
  geom_boxplot() +
  labs(title = "Distribucija prihoda prema drzavi iz koje poticu filmovi", x
= "Country", y = "Gross") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

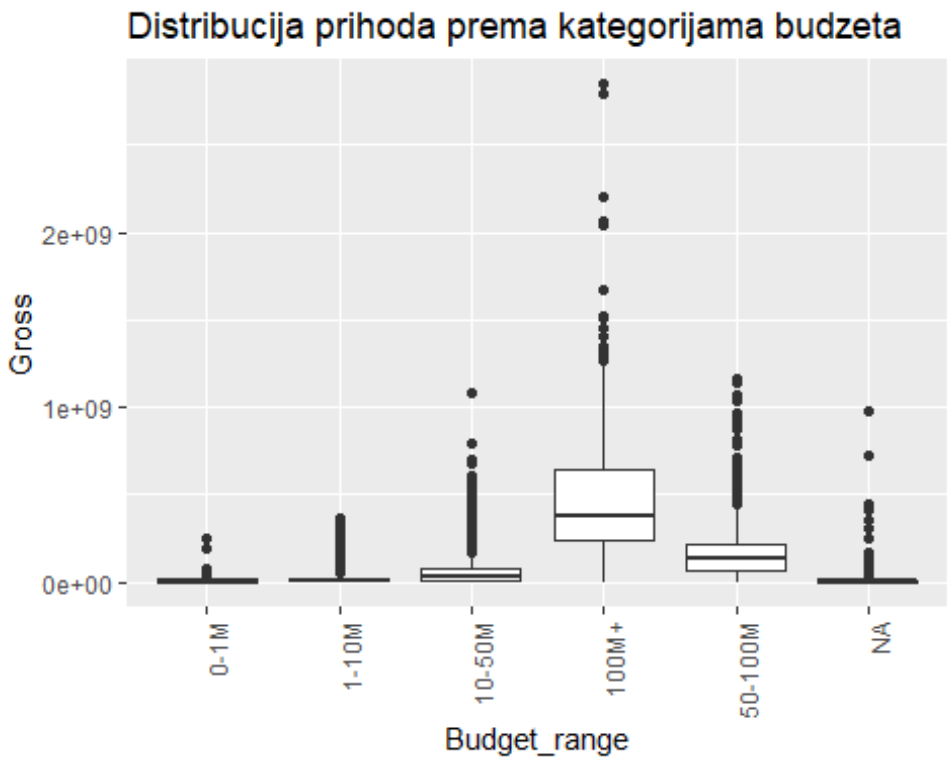


Najveće prihode očekivano ostvaruje Amerika. Razlog tome može da bude to sto najveće produkcijske kuće dolaze upravo iz Amerike. Takođe Amerika je centar filmske industrije, tu je i Holivud. Očigledno da velika ulaganja koja produkcijske kuće mogu da obezbede donose i velike prihode. Nakon Amerike tu je i Engleska. Neki od najpopularnijih filmova dolaze upravo iz ove zemlje kao npr. Hari Poter.

Kolona budget_range

Grafikom ispod je prikazan odnos izmedju budžeta filma i prihoda.

```
ggplot(data = movies, mapping = aes(x = budget_range, y = gross)) +
  geom_boxplot() +
  labs(title = "Distribucija prihoda prema kategorijama budzeta", x =
"Budget_range", y = "Gross") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

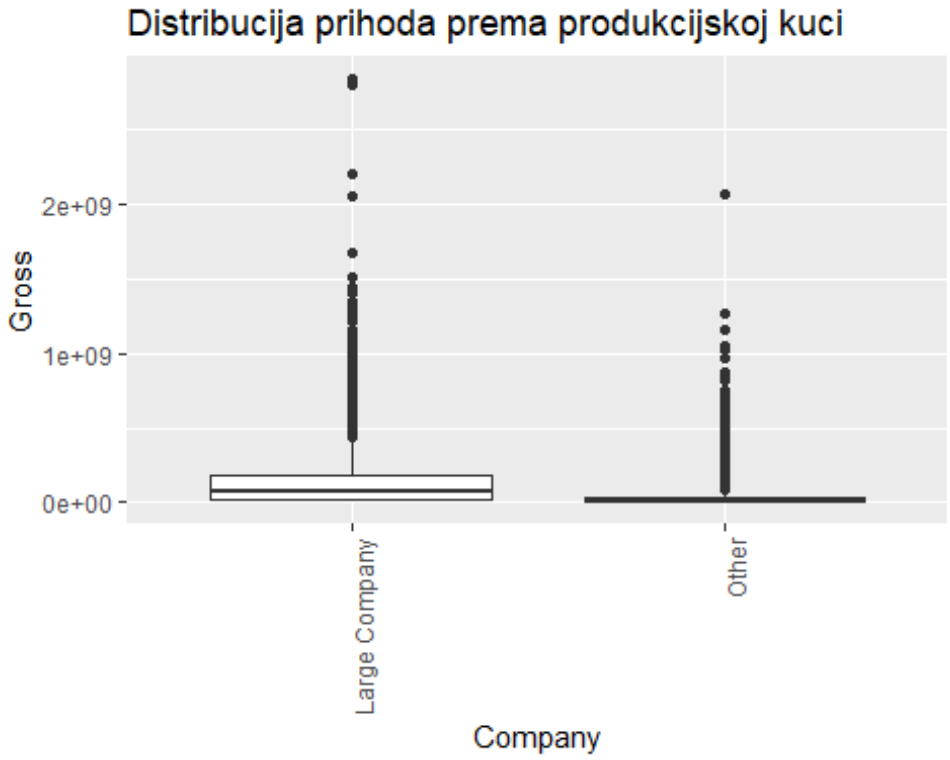


Ovde jasno mozemo da vidimo da filmovi sa budžetima preko 100M+ imaju najveće prihode. Prihodi kategorije budžeta 100M+ se dosta razlikuju od ostalih kao i medijana koja je najveća. U proseku ovi filmovi donose najveće prihode. Tu se nalaze i filmovi sa najvećim prihodima (ekstremne vrednosti) i to može ukazivati na blockbuster-e recimo, ne mora da znači da su izuzeci.

Ovaj prediktor može da bude značajan pri kreiranju modela pored još nekih faktora.

Kolona company

Grafikom ispod je prikazan odnos između produkcijske kuće i prihoda.

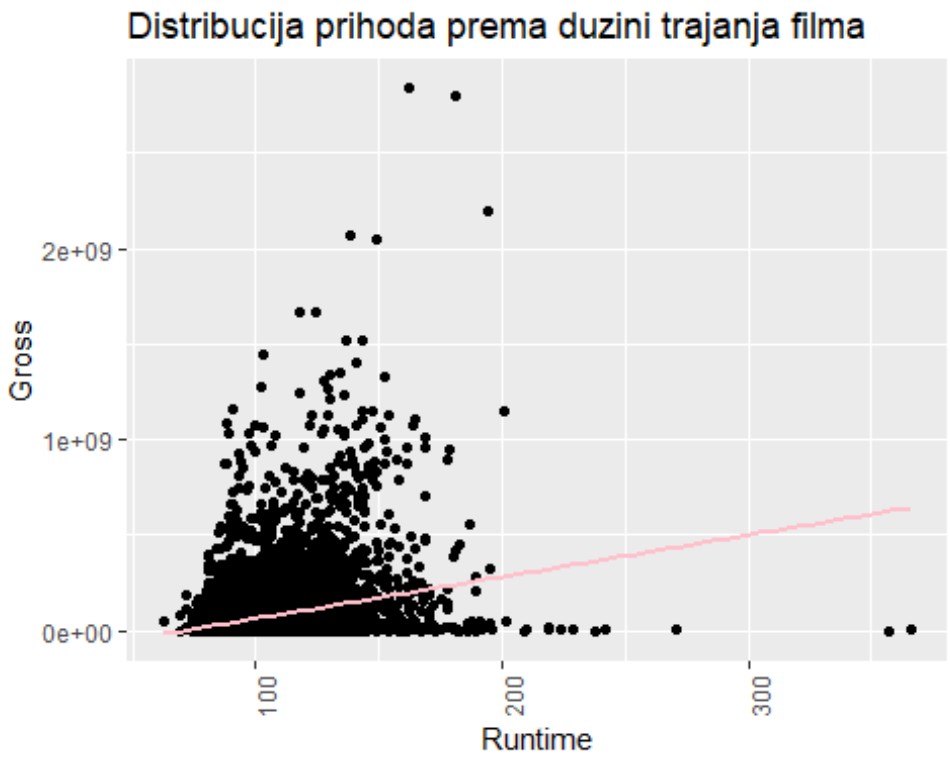


Sa grafika mozemo da vidimo da veće kompanije zapravo ostvaruju veće prihode iako je broj takvih filmova manji. Kod većih produkcijskih kuća je veća medijana što znači da je prosečna zarada ovih filmova veća u odnosu na "ostale" filmove. Isto tako kod većih kompanija imamo više vrednosti koje "odskaču" iz skupa i može ukazivati na blockbuster-e koje su karakteristične za ovakve produkcije (npr. filmovi kao sto su Harry Poter, Deadpool, Hobbit, Diznijevi filmovi...) .

Kolona runtime

Grafikom ispod je prikazan odnos između dužine trajanja filma i prihoda.

```
ggplot(data = movies, mapping = aes(x = runtime, y = gross)) +
  geom_point() +
  geom_smooth(method = "lm", formula="y~x", se=FALSE, col="pink") +
  labs(title = "Distribucija prihoda prema duzini trajanja filma", x =
"Runtime", y = "Gross") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



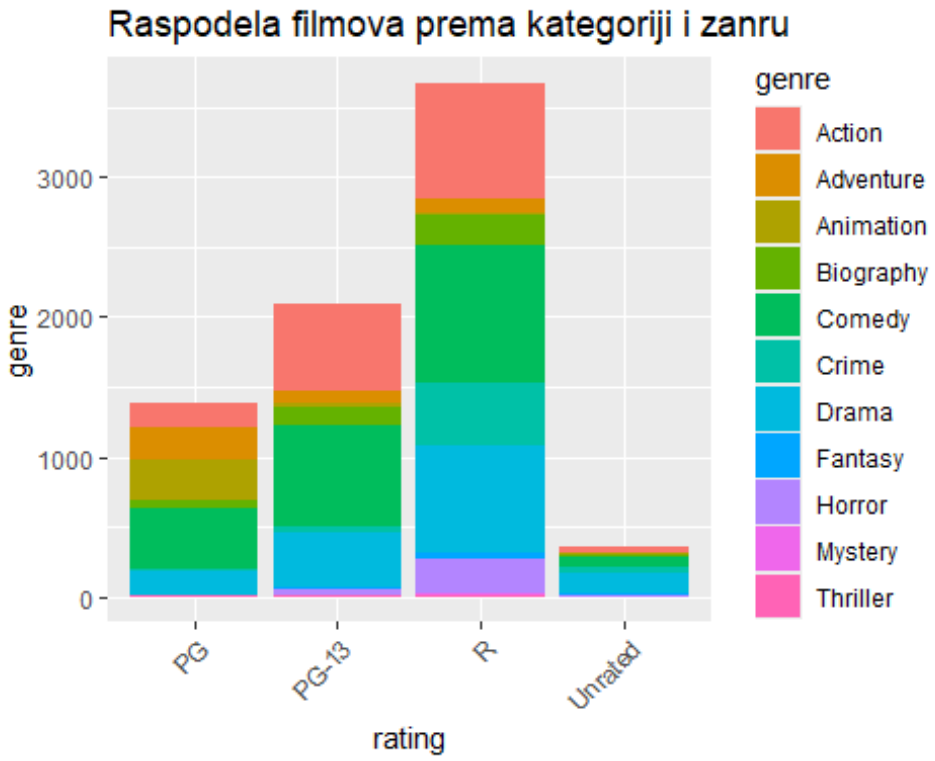
Na ovom grafiku izgleda kao da nema jasne korelacije između dužine trajanja filma i prihoda.

Multivarijantna analiza

Ovde ću pokušati da pronađem povezanost između više kolona na osnovu dosadašnje analize.

Povezanost između rating i genre

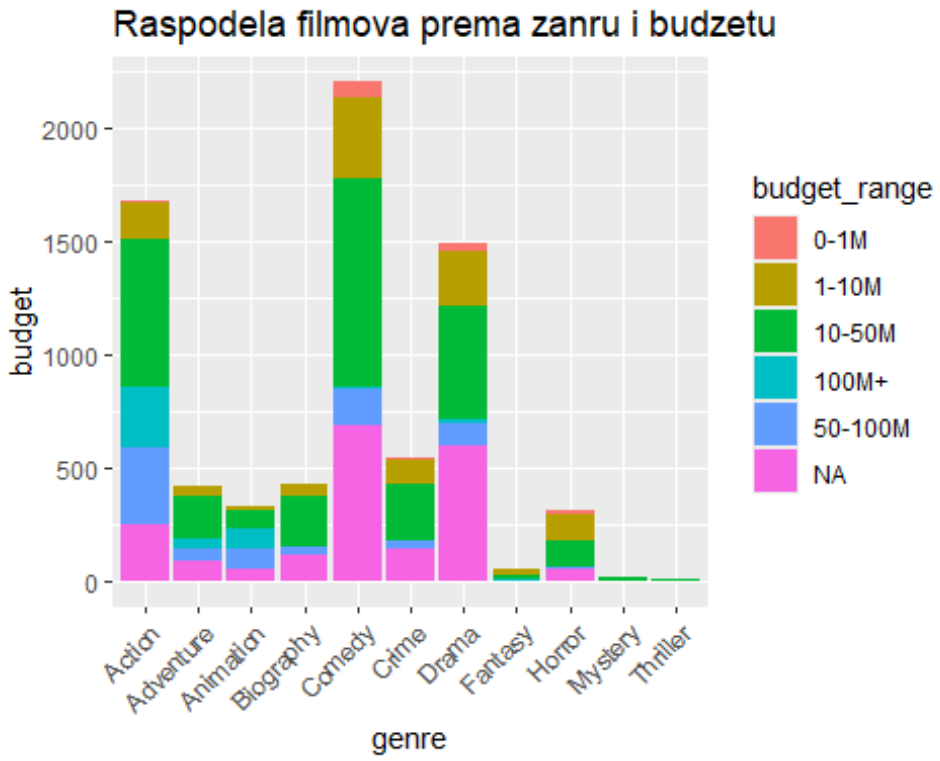
```
ggplot(data = movies, mapping=aes(x=rating, fill=genre)) +
  geom_bar(position = "stack") +
  labs(title = "Raspodela filmova prema kategoriji i zanru", x = "rating", y
= "genre") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Generalno se vidi raznolikost žanrova u svim kategorijama. Unutar R kategorije se nalazi žanr koga nema kod drugih npr.Crime. Takođe kod PG kategorije je manja zastupljenost akcionih filmova, kao i Horora i Misterije.

Povezanost između genre i budget

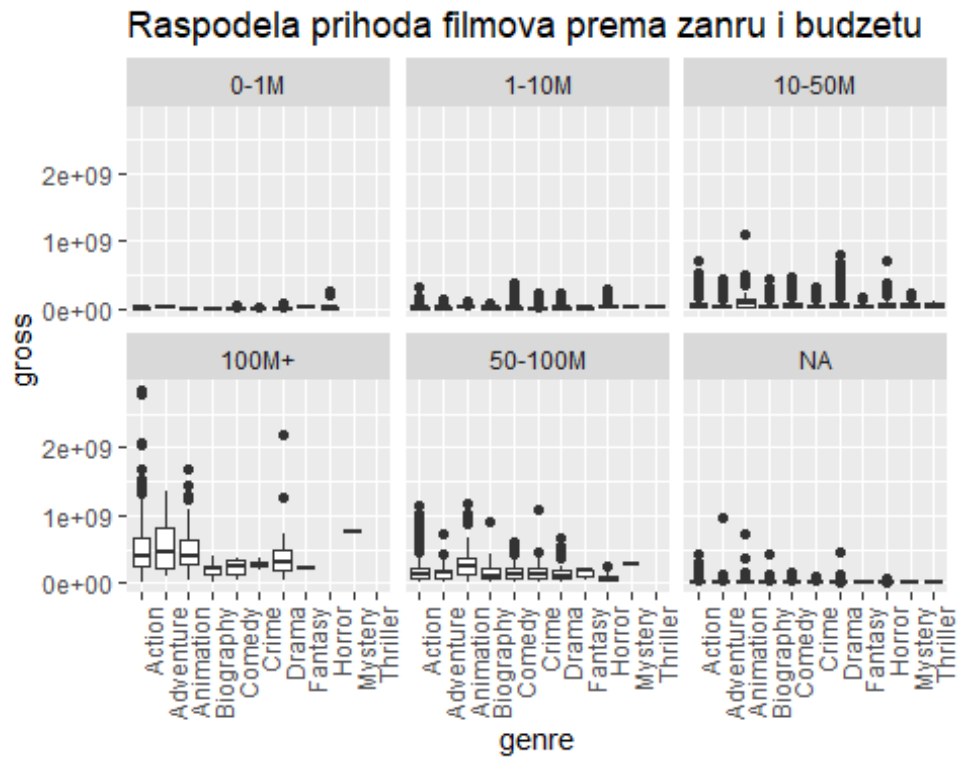
```
ggplot(data = movies, mapping=aes(x=genre, fill=budget_range)) +
  geom_bar(position = "stack") +
  labs(title = "Raspodela filmova prema zanru i budzetu", x = "genre", y =
"budget") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Budžet od 100+ miliona je zastupljen samo kod određenih žanrova: akcije, avanture i animacije. Ostalo je raznoliko. Uglavnom je najzastupljeniji srednji budžet: 10-50 miliona. Veza između ova dva prediktora može da bude značajna.

Povezanost između žanra, budžeta i prihoda

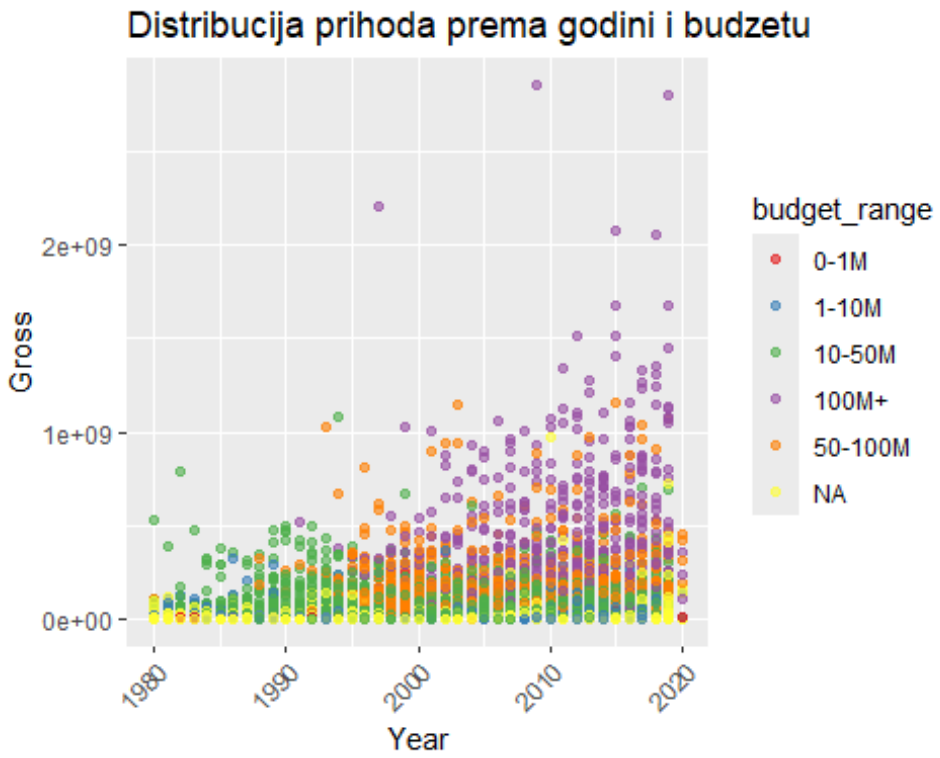
```
ggplot(data = movies, mapping=aes(x=genre, y=gross)) +
  geom_boxplot() +
  labs(title = "Raspodela prihoda filmova prema zanru i budzetu", x =
"genre", y = "gross") +
  facet_wrap(~budget_range) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Najmanje prihode imaju filmovi svih žanrova sa budžetom do 10 miliona, dok najveće prihode imaju filmovi: akcija, avantura, animacija sa budžetom preko 100 miliona. Ova veza, između ova tri prediktora je značajna.

Povezanost između godine, budžeta i prihoda

```
ggplot(data = movies, mapping = aes(x = year, y = gross, color=budget_range))
+
  geom_point(alpha=0.6) +
  labs(title = "Distribucija prihoda prema godini i budzetu", x = "Year", y =
"Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_colour_brewer(palette = "Set1")
```

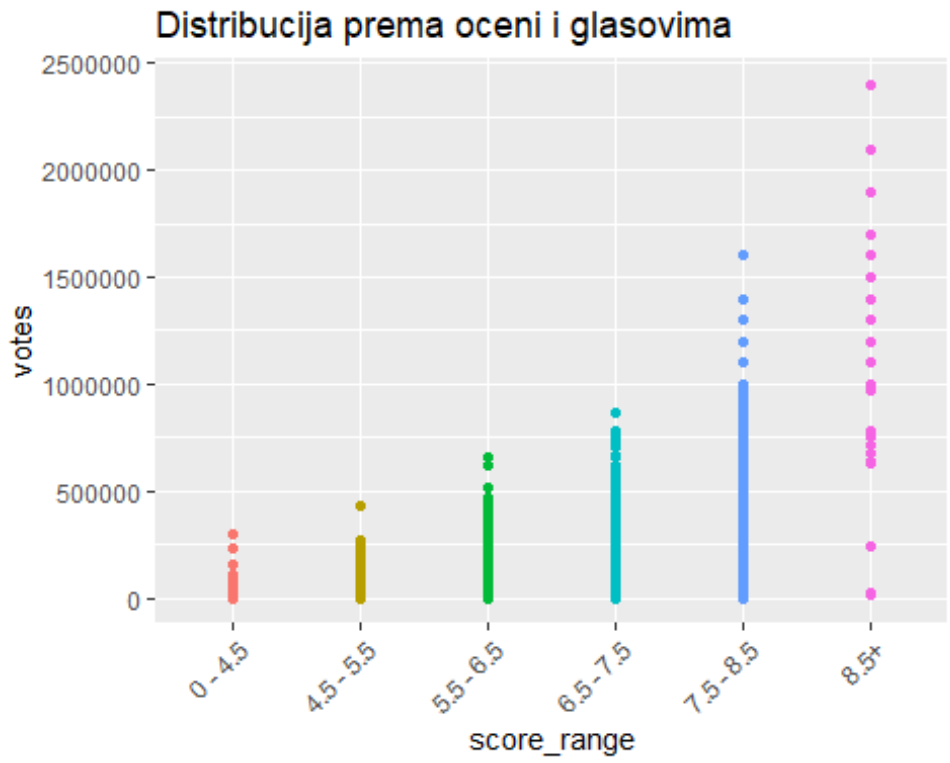


Ovde možemo da zaključimo da filmovi sa budžetom preko 100 miliona počinju da postaju prisutni nakon 2000. godine i ti filmovi uglavnom ostvaruju veće prihode. Ovo može da ukazuje na vezu većih budžeta sa potencijalno većim prihodima. Broj filmova sa visokim prihodima se povećava nakon 2000. godine, dok pre toga možemo da vidimo malo ujednačeniju raspodelu. Moguće je da su neki faktori poput marketinga ili boljih tehnologija doprinele tome. Filmovi sa budžetom ispod 10 miliona (plave tačke) obično ostvaruju niže prihode i ograničeni su na mogućnost ostvarivanja velikih prihoda.

Zaključak: Očigledno je da budžet igra značajnu ulogu u predviđanju prihoda filma. Budget i year zajedno mogu da budu značajni za model.

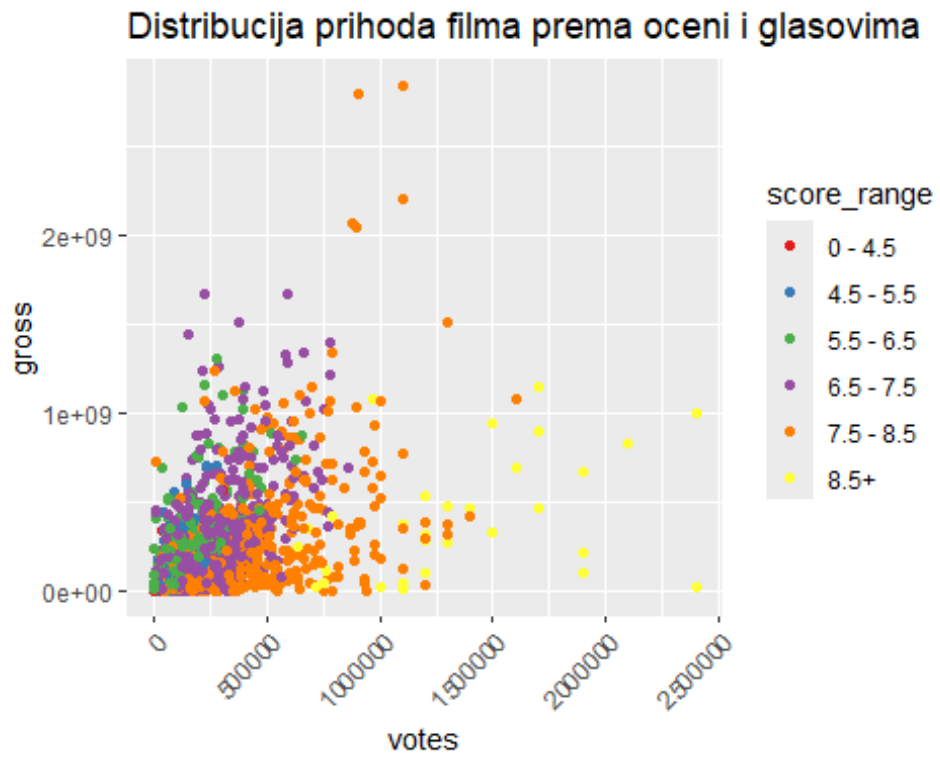
Povezanost između score i votes

```
ggplot(data = movies, mapping = aes(x = score_range, y=votes,
col=score_range)) +
  geom_point() +
  labs(title = "Distribucija prema oceni i glasovima", x = "score_range", y =
"votes") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position =
"none")
```



U grafik iznad sam dodatno ubacila i prihod (gross).

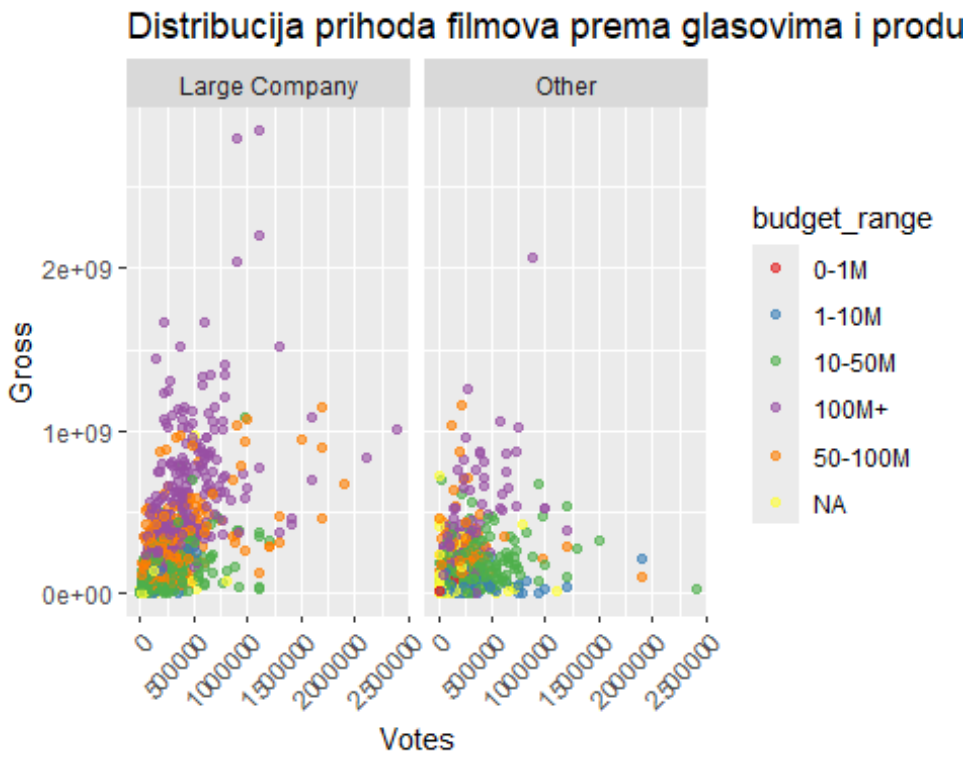
```
ggplot(data = movies, mapping = aes(x = votes, y=gross, col=score_range)) +
  geom_point() +
  labs(title = "Distribucija prihoda filma prema oceni i glasovima", x =
"votes", y = "gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_colour_brewer(palette = "Set1")
```



Najpopularniji filmovi, sa najvećim ocenama imaju i najveći broj glasova. Ova dva grafika su takođe značajna, pokazuju nam da postoji povezanost između broja glasova i ocene.

Povezanost izmedju budget , votes i gross

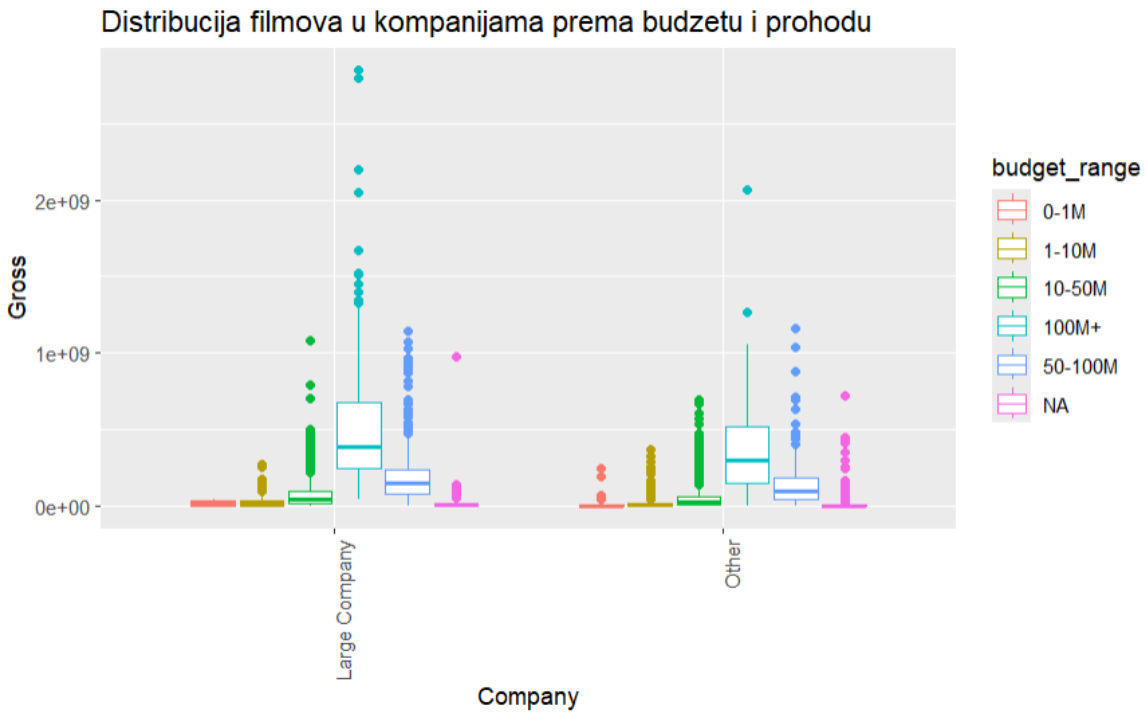
```
ggplot(data = movies, mapping = aes(x = votes, y = gross, col=budget_range))
+
  geom_point(alpha=0.6) +
  labs(title = "Distribucija prihoda filmova prema glasovima i produkcijskoj
kuci i budzetu", x = "Votes", y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_colour_brewer(palette = "Set1") +
  facet_wrap(~company_group)
```



Filmovi sa većim budžetom (preko 100 miliona) uglavnom ostvaruju veće prihode i imaju više glasova posebno kod velikih kompanija.

Povezanost izmedju budget , company i gross

```
ggplot(data=movies, mapping = aes(x=company_group, y=gross,
col=budget_range)) +
  geom_boxplot() +
  labs(title = "Distribucija filmova u kompanijama prema budzetu i prohodu",
x = "Company", y = "Gross") +
  theme(axis.text.x = element_text(angle= 90, hjust=1))
```



Ovi rezultati potvrđuju dominaciju velikih produkcionih kuca u filmskoj industriji, dok manje kompanije verovatno imaju manji tržišni udeo.

Zaključak na osnovu celokupne analize

Na osnovu svih dosadašnjih analiza došla sam do zaključka da kolona prihod zavisi od nekoliko bitnih kolona kao što su budget, votes, genre, year i company. Kolone budget i votes deluju veoma bitno za predviđanje prihoda, pogotovo u korelaciji sa drugim kolonama, npr. budget~year, budget~genre, votes~score_range. Primetila sam trend da su neki žanrovi poput akcije, drame i komedije popularniji od drugih, da isto tako imaju veće budžete i veće prihode. Isto tako u filmsku industriju je krenulo više da se ulaže nakon 2000. godine kada može da se primeti značajniji rast u prihodima filma, posebno nakon 2010. godine. Razlozi tome su mnogobrojni poput razvoja tehnologija, marktetinga, itd... Važan faktor je i produkcijska kuća. Velike produkcijske kuće imaju više novca da ulažu u budžet filma, pa su oni najpopularniji sa najvećim prihodom zapravo oni koji potiču iz velikih kompanija. Manje produkcijske kuće imaju manji udeo na tržištu i manju zaradu pa tako i manje budžete. Još jedna bitna činjenica je da u ovom skupu podataka najviše filmova potiče iz Amerike što je očekivano jer je ona centar filmske industrije. Ovo su neki od bitnijih prediktora koje sam dobila kombinacijom domenskom znanja i analize grafika. U nastavku ću videti koliko dobar model može da se napravi.

Korelaciona matrica

```
cor_matrix <- cor(select_if(movies, is.numeric))
cor_matrix

##               year      score     votes budget     gross   runtime
decade
## year    1.00000000 0.09421531 0.2176638     NA 0.2578376 0.1141885
0.96653493
## score   0.09421531 1.00000000 0.4129376     NA 0.1865923 0.3990243
0.09541778
## votes   0.21766380 0.41293757 1.0000000     NA 0.6310780 0.3099364
0.22248646
## budget          NA         NA        NA      1        NA        NA
NA
## gross   0.25783758 0.18659229 0.6310780     NA 1.0000000 0.2452095
0.25140729
## runtime 0.11418854 0.39902435 0.3099364     NA 0.2452095 1.0000000
0.11341379
## decade  0.96653493 0.09541778 0.2224865     NA 0.2514073 0.1134138
1.00000000
```

Korelaciona matrica omogućava prikaz korelacija između numeričkih kolona. Na osnovu korelacione matrice može da se uoči korelacija između votes i gross ~63%, dok sa ostalim kolonama gross ima korelaciju ispod 30% što nije dobra korelacija.

Kreiranje modela

Prvo pripremam podatke tako što kategorijske promenljive pretvaram u factor.

```
movies$rating = factor(movies$rating)
movies$genre = factor(movies$genre)
movies$budget_range = factor(movies$budget_range)
movies$score_range = factor(movies$score_range)
movies$company_group = factor(movies$company_group)
movies$country_group = factor(movies$country_group)

movies$director = factor(movies$director)
movies$writer = factor(movies$writer)
movies$star = factor(movies$star)
```

Podela podataka na train:test skup u odnosu 80:20. Stratifikacija na osnovu kolone budget_range.

```
set.seed(123)
trainIndex <- createDataPartition(movies$budget_range, p = 0.8, list = FALSE)

train <- movies[trainIndex, ]
test <- movies[-trainIndex, ]
```

Linearna regresija

Za početak ću koristiti samo kolone budget_range i votes jer na osnovu dosadašnje analize deluju najbitnije.

```
lm.fit1 = lm(gross ~ budget_range + votes, data = train)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = gross ~ budget_range + votes, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -954774508  -22162613   -3785857    9775160 2086915642
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.956e+06  9.624e+06  -0.203  0.83898
## budget_range1-10M   -1.340e+06  1.023e+07  -0.131  0.89581
## budget_range10-50M   2.592e+07  9.872e+06   2.625  0.00868 **
## budget_range100M+    3.515e+08  1.145e+07  30.691  < 2e-16 ***
## budget_range50-100M  1.069e+08  1.054e+07  10.143  < 2e-16 ***
## budget_rangeNA       5.034e+06  9.966e+06   0.505  0.61347
## votes                3.998e+02  9.049e+00  44.185  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104900000 on 6007 degrees of freedom
## Multiple R-squared:  0.5938, Adjusted R-squared:  0.5934
## F-statistic:  1464 on 6 and 6007 DF,  p-value: < 2.2e-16
```

Tačnost modela iznosi ~59% što i nije tako dobar rezultat, opisuje nešto više od polovine varijanse u podacima. Model greši za 101 700 000 jedinica što ukazuje na to da treba uključiti i druge promenljive kako bi bolje objašnjavao podatke.

Ostale numeričke kolone pokazuju slabu korelaciju sa gross i to smo videli pomoću korelacione matrice. Zato ću dodati neku kategorijsku promenljivu da vidim kako će se model ponašati. Na osnovu prethodne analize kolone genre, score_range i company mogu imati uticaj na prihod.

```
lm.fit2 = lm(gross ~ votes + budget_range + score_range, data = train)
summary(lm.fit2)
```

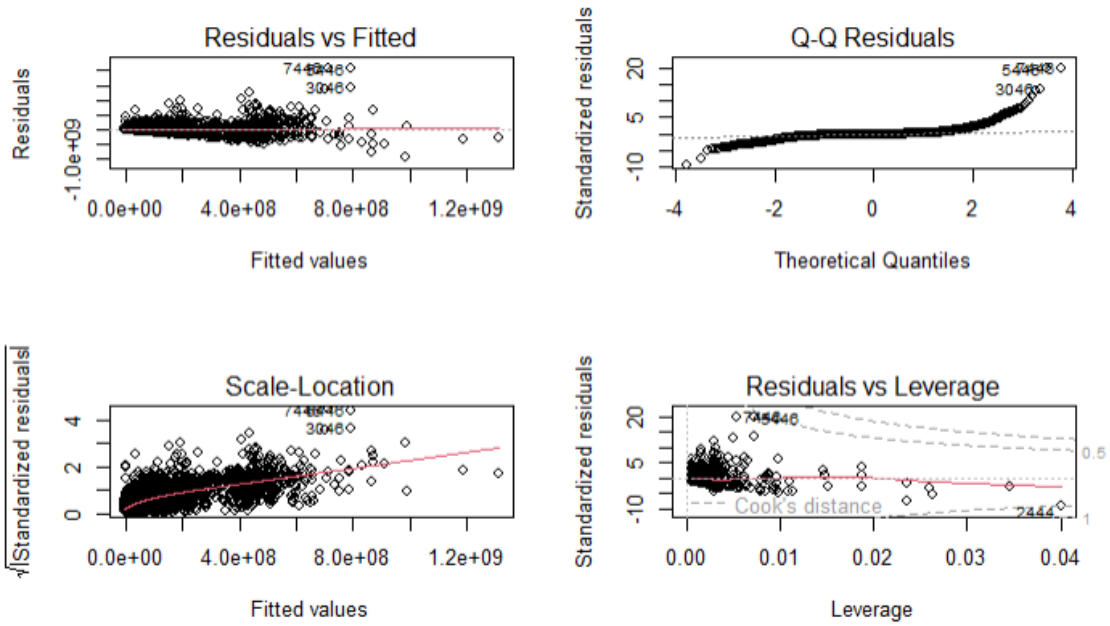
```
##
## Call:
## lm(formula = gross ~ votes + budget_range + score_range, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -896083833  -20402131   -5134779   11170234 2050193975
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.935e+06  1.153e+07  -0.688   0.4915
## votes                 4.996e+02  1.212e+01  41.220   <2e-16 ***
## budget_range1-10M    -1.896e+06  1.010e+07  -0.188   0.8511
## budget_range10-50M    2.168e+07  9.763e+06   2.221   0.0264 *
## budget_range100M+     3.255e+08  1.150e+07  28.307   <2e-16 ***
## budget_range50-100M   9.601e+07  1.047e+07   9.173   <2e-16 ***
## budget_rangeNA        7.654e+06  9.846e+06   0.777   0.4370
## score_range4.5 - 5.5  6.390e+06  8.089e+06   0.790   0.4296
## score_range5.5 - 6.5  6.503e+06  7.447e+06   0.873   0.3825
## score_range6.5 - 7.5  4.954e+06  7.442e+06   0.666   0.5057
## score_range7.5 - 8.5 -2.135e+07  8.441e+06  -2.529   0.0114 *
## score_range8.5+      -2.878e+08  2.375e+07 -12.118   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103500000 on 6002 degrees of freedom
## Multiple R-squared:  0.6055, Adjusted R-squared:  0.6048
## F-statistic: 837.5 on 11 and 6002 DF,  p-value: < 2.2e-16
```

Dodavanjem prediktora score_range model se minimalno poboljšao ~61%. Ono što primećujem je da Adjusted R-squared sve vreme prati Multiple R-squared što je bitno kako ne bi došlo do overfitting-a. F-statistika > 1, p-vrednost < 0.5 Znači da postoji povezanost između prediktora i odgovora. RSE: 98620000 pokazuje da je greška predviđanja modela velika u odnosu na raspon prihoda

Dodavanjem ostalih promenljivih u model (tj. bez writer, director, star, released, jer imaju ogroman broj kategorija) dobije se minimalno poboljšanje modela, skoro da ga nema. Zbog toga ću pokušati da upotrebim neke druge modele.

```
par(mfrow=c(2,2))

plot(lm.fit1)
```



Analiza dobijenih grafika

Residuals vs Fitted - služi za proveru pretpostavki linearnog odnosa, horizontalna linija bez jasnih obrazaca je dobar pokazatelj. Ovde možemo da vidimo da su podaci uglavnom raspoređeni oko horizontalne linije ali su vrlo zbijeni.

Normal Q-Q - služi za ispitivanje da li se reziduali normalno distribuiraju. Dobro je ukoliko prate isprekidanu liniju. Na grafiku ovog modela se vide odstupanja, pa raspodela očigledno nije normalna. Razlog tome može biti veliki broj ekstremnih vrednosti.

Scale-Location - služi za proveru homogenosti varijanse reziduala. Horizontalna linija sa jednako raspršenim tačkama je pokazatelj dobrog reziduala. Na grafiku se vidi da su skupljene tačke između 0.0e+00 i 6.0e+08, i da linija nije horizontalna što znači da varijansa reziduala nije konstantna.

Residuals vs Leverage - ovaj grafik služi za identifikaciju ekstremnih vrednosti koje mogu da utiču na rezultate. Na ovom grafiku se nalaze tačke sa ekstremnim vrednostima: 7446, 2444, 5446

Decision tree

Stablo odlučivanja može efikasno da opiše nelinarne odnose pa ću u nastavku napraviti nov model.

Kategorijski podaci su vec pretvoreni u factor promenljive pre linearne regresije. Kolone koje imaju veliki broj kategorija: name, director, writer, star i released ne uključujem u model.

```
lm.fit3 <- rpart(gross ~ rating + genre + year + votes + runtime +
budget_range + score_range + company_group + country_group,  method =
"anova", data = train)
lm.fit3
```

```
## n= 6014
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 6014 1.628424e+20   77935960
##    2) budget_range=0-1M,1-10M,10-50M,50-100M,NA 5662 5.208556e+19
52742280
##      4) votes< 115500 4672 1.012814e+19   29138820
##        8) budget_range=0-1M,1-10M,10-50M,NA 4322 5.728160e+18   23350290 *
##        9) budget_range=50-100M 350 2.466870e+18  100618900 *
##      5) votes>=115500 990 2.707105e+19  164131500
##       10) budget_range=0-1M,1-10M,10-50M,NA 691 9.416010e+18  120692500 *
##       11) budget_range=50-100M 299 1.333785e+19  264520600
##         22) votes< 495500 261 7.668847e+18  226636300
##           44)
genre=Action,Adventure,Biography,Comedy,Crime,Drama,Fantasy,Horror,Mystery
240 3.403254e+18  198658000 *
##           45) genre=Animation 21 1.930649e+18  546388800 *
##         23) votes>=495500 38 2.721555e+18  524725800 *
##    3) budget_range=100M+ 352 4.935602e+19  483182200
##      6) votes< 349500 226 1.530793e+19  344381100
##       12) votes< 137000 78 1.236947e+18  204393600 *
##       13) votes>=137000 148 1.173687e+19  418158400
##         26) year< 2015.5 110 3.059945e+18  337869500 *
##         27) year>=2015.5 38 5.915198e+18  650573500
##           54) rating=PG-13,R 27 2.211649e+18  491772000 *
##           55) rating=PG 11 1.351404e+18 1040359000 *
##      7) votes>=349500 126 2.188439e+19  732142800
##       14) votes< 786500 108 6.945836e+18  667774300 *
##       15) votes>=786500 18 1.180622e+19 1118354000 *
```

```
summary(lm.fit3)
```

```
## Call:
## rpart(formula = gross ~ rating + genre + year + votes + runtime +
##     budget_range + score_range + company_group + country_group,
```

```
##     data = train, method = "anova")
##   n= 6014
##
##            CP nsplit rel error    xerror       xstd
## 1  0.37705671      0 1.0000000 1.0001522 0.09203379
## 2  0.09141581      1 0.6229433 0.6239565 0.06182897
## 3  0.07469620      2 0.5315275 0.5409628 0.06069775
## 4  0.02651147      3 0.4568313 0.4788496 0.05135118
## 5  0.01923537      4 0.4303198 0.4759044 0.05128951
## 6  0.01810000      5 0.4110844 0.4545653 0.04528531
## 7  0.01564653      6 0.3929844 0.4459164 0.04527387
## 8  0.01444430      8 0.3616914 0.4315924 0.04281878
## 9  0.01433867      9 0.3472471 0.4191478 0.03879555
## 10 0.01187104     10 0.3329084 0.4056511 0.03609983
## 11 0.01000000     11 0.3210374 0.3909739 0.03534936
##
## Variable importance
## budget_range        votes  score_range      runtime       rating
year
##           54           29            5            4            3
2
##        genre
##            2
##
## Node number 1: 6014 observations,    complexity param=0.3770567
##   mean=7.793596e+07, MSE=2.707722e+16
##   left son=2 (5662 obs) right son=3 (352 obs)
##   Primary splits:
##       budget_range  splits as  LLLRLL,      improve=0.37705670, (0
missing)
##       votes         < 196500  to the left,  improve=0.32558650, (0
missing)
##       company_group splits as  RL,          improve=0.10335540, (0
missing)
##       genre         splits as  RRRLLLLLLRL, improve=0.09355917, (0
missing)
##       runtime       < 126.5   to the left,  improve=0.06475830, (0
missing)
##   Surrogate splits:
##       votes < 2e+06   to the left,  agree=0.942, adj=0.003, (0 split)
##
## Node number 2: 5662 observations,    complexity param=0.09141581
##   mean=5.274228e+07, MSE=9.199145e+15
##   left son=4 (4672 obs) right son=5 (990 obs)
##   Primary splits:
##       votes         < 115500  to the left,  improve=0.28580610, (0
missing)
##       budget_range  splits as  LLL-RL,      improve=0.21426070, (0
missing)
##       company_group splits as  RL,          improve=0.08314070, (0
```

```
missing)
##       genre         splits as  LLRLLLLLLL, improve=0.04228667, (0
missing)
##       score_range   splits as  LLLLLR,     improve=0.04155680, (0
missing)
##   Surrogate splits:
##       score_range splits as  LLLLLR, agree=0.83, adj=0.026, (0 split)
##
## Node number 3: 352 observations,    complexity param=0.0746962
##   mean=4.831822e+08, MSE=1.40216e+17
##   left son=6 (226 obs) right son=7 (126 obs)
##   Primary splits:
##       votes       < 349500  to the left,  improve=0.24644830, (0 missing)
##       score_range splits as  LLLLRR,     improve=0.12699640, (0 missing)
##       runtime     < 135.5   to the left,  improve=0.10218680, (0 missing)
##       year        < 2016.5  to the left,  improve=0.04786288, (0 missing)
##       rating      splits as  RRL-,       improve=0.04164181, (0 missing)
##   Surrogate splits:
##       score_range splits as  LLLLRR,     agree=0.770, adj=0.357, (0
split)
##       runtime     < 138.5   to the left,  agree=0.710, adj=0.190, (0
split)
##       genre       splits as  LLLLLRLL-R-, agree=0.651, adj=0.024, (0
split)
##
## Node number 4: 4672 observations,    complexity param=0.01187104
##   mean=2.913882e+07, MSE=2.167838e+15
##   left son=8 (4322 obs) right son=9 (350 obs)
##   Primary splits:
##       budget_range  splits as  LLL-RL,     improve=0.19086510, (0
missing)
##       votes         < 44500   to the left,  improve=0.18312930, (0
missing)
##       company_group splits as  RL,         improve=0.08140397, (0
missing)
##       rating        splits as  RRLL,       improve=0.05779456, (0
missing)
##       year          < 1997.5  to the left,  improve=0.03526413, (0
missing)
##
## Node number 5: 990 observations,    complexity param=0.02651147
##   mean=1.641315e+08, MSE=2.734449e+16
##   left son=10 (691 obs) right son=11 (299 obs)
##   Primary splits:
##       budget_range  splits as  LLL-RL,     improve=0.15947630, (0
missing)
##       genre         splits as  LLRLLLLLL-, improve=0.10585840, (0
missing)
##       rating        splits as  RRLL,       improve=0.10154550, (0
missing)
```

```
##       votes         < 342500  to the left,  improve=0.09879704, (0
missing)
##       company_group splits as  RL,          improve=0.06143157, (0
missing)
##   Surrogate splits:
##       votes   < 1250000 to the left,  agree=0.702, adj=0.013, (0 split)
##       genre   splits as  LLRLLLLLLL-, agree=0.701, adj=0.010, (0 split)
##       runtime < 168.5   to the left,  agree=0.700, adj=0.007, (0 split)
##
## Node number 6: 226 observations,    complexity param=0.01564653
##   mean=3.443811e+08, MSE=6.773418e+16
##   left son=12 (78 obs) right son=13 (148 obs)
##   Primary splits:
##       votes       < 137000  to the left,  improve=0.15247690, (0 missing)
##       year        < 2016.5  to the left,  improve=0.14434690, (0 missing)
##       genre       splits as  LLRLL-LL---, improve=0.06003296, (0 missing)
##       rating      splits as  RRL-,        improve=0.04380500, (0 missing)
##       score_range splits as  LLLRR-,      improve=0.03850135, (0 missing)
##   Surrogate splits:
##       genre         splits as  RLRRL-RR---, agree=0.686, adj=0.090, (0
split)
##       runtime       < 86.5    to the left,  agree=0.686, adj=0.090, (0
split)
##       score_range   splits as  RLRRR-,      agree=0.681, adj=0.077, (0
split)
##       country_group splits as  RLR,         agree=0.659, adj=0.013, (0
split)
##
## Node number 7: 126 observations,    complexity param=0.01923537
##   mean=7.321428e+08, MSE=1.736856e+17
##   left son=14 (108 obs) right son=15 (18 obs)
##   Primary splits:
##       votes   < 786500  to the left,  improve=0.14313100, (0 missing)
##       runtime < 132.5   to the left,  improve=0.11241200, (0 missing)
##       rating  splits as  RRL-,        improve=0.08054138, (0 missing)
##       year    < 2017.5  to the left,  improve=0.06698369, (0 missing)
##       genre   splits as  RRLL-LR--R-, improve=0.02220370, (0 missing)
##   Surrogate splits:
##       score_range splits as  --LLLR,      agree=0.881, adj=0.167, (0
split)
##       year        < 1997.5  to the right, agree=0.873, adj=0.111, (0
split)
##       runtime     < 174.5   to the left,  agree=0.873, adj=0.111, (0
split)
##
## Node number 8: 4322 observations
##   mean=2.335029e+07, MSE=1.325349e+15
##
## Node number 9: 350 observations
##   mean=1.006189e+08, MSE=7.048199e+15
```

```
##
## Node number 10: 691 observations
##   mean=1.206925e+08, MSE=1.362664e+16
##
## Node number 11: 299 observations,    complexity param=0.0181
##   mean=2.645206e+08, MSE=4.460819e+16
##   left son=22 (261 obs) right son=23 (38 obs)
##   Primary splits:
##       votes       < 495500  to the left,  improve=0.22098370, (0 missing)
##       genre       splits as  LLRLLLLLLL-, improve=0.18710470, (0 missing)
##       rating      splits as  RLL-,        improve=0.13335870, (0 missing)
##       score_range splits as  -LLLLR,      improve=0.08068626, (0 missing)
##       runtime     < 91.5    to the right, improve=0.04501369, (0 missing)
##   Surrogate splits:
##       score_range splits as  -LLLLR,      agree=0.90, adj=0.211, (0 split)
##       runtime     < 159.5   to the left,  agree=0.88, adj=0.053, (0 split)
##
## Node number 12: 78 observations
##   mean=2.043936e+08, MSE=1.585829e+16
##
## Node number 13: 148 observations,    complexity param=0.01564653
##   mean=4.181584e+08, MSE=7.93032e+16
##   left son=26 (110 obs) right son=27 (38 obs)
##   Primary splits:
##       year        < 2015.5  to the left,  improve=0.23530380, (0 missing)
##       genre       splits as  LRRLL-RL---, improve=0.15720580, (0 missing)
##       rating      splits as  RLL-,        improve=0.12416130, (0 missing)
##       votes       < 261000  to the left,  improve=0.04615347, (0 missing)
##       score_range splits as  LLLRR-,      improve=0.03095949, (0 missing)
##   Surrogate splits:
##       votes         < 141500  to the right, agree=0.764, adj=0.079, (0
split)
##       score_range   splits as  LLLLR-,      agree=0.757, adj=0.053, (0
split)
##       country_group splits as  LRL,         agree=0.757, adj=0.053, (0
split)
##
## Node number 14: 108 observations
##   mean=6.677743e+08, MSE=6.43133e+16
##
## Node number 15: 18 observations
##   mean=1.118354e+09, MSE=6.55901e+17
##
## Node number 22: 261 observations,    complexity param=0.01433867
##   mean=2.266363e+08, MSE=2.938256e+16
##   left son=44 (240 obs) right son=45 (21 obs)
##   Primary splits:
##       genre         splits as  LLRLLLLLLL-, improve=0.30447130, (0
missing)
##       rating        splits as  RLL-,        improve=0.23519240, (0
```

```
missing)
##       runtime       < 91.5    to the right, improve=0.10936030, (0
missing)
##       votes         < 180500  to the left,  improve=0.06543695, (0
missing)
##       country_group splits as  LLR,         improve=0.02874193, (0
missing)
##   Surrogate splits:
##       runtime < 91.5    to the right, agree=0.954, adj=0.429, (0 split)
##       rating  splits as  RLL-,        agree=0.950, adj=0.381, (0 split)
##
## Node number 23: 38 observations
##   mean=5.247258e+08, MSE=7.161986e+16
##
## Node number 26: 110 observations
##   mean=3.378695e+08, MSE=2.781768e+16
##
## Node number 27: 38 observations,    complexity param=0.0144443
##   mean=6.505735e+08, MSE=1.556631e+17
##   left son=54 (27 obs) right son=55 (11 obs)
##   Primary splits:
##       rating  splits as  RLL-,        improve=0.39764430, (0 missing)
##       genre   splits as  LRRL--R----, improve=0.31858780, (0 missing)
##       votes   < 213500  to the left,  improve=0.16714720, (0 missing)
##       runtime < 108.5   to the right, improve=0.07187364, (0 missing)
##       year    < 2017.5  to the left,  improve=0.04668886, (0 missing)
##   Surrogate splits:
##       runtime     < 106.5   to the right, agree=0.842, adj=0.455, (0
split)
##       votes       < 144000  to the right, agree=0.737, adj=0.091, (0
split)
##       score_range splits as  -LLLR-,      agree=0.737, adj=0.091, (0
split)
##
## Node number 44: 240 observations
##   mean=1.98658e+08, MSE=1.418023e+16
##
## Node number 45: 21 observations
##   mean=5.463888e+08, MSE=9.193565e+16
##
## Node number 54: 27 observations
##   mean=4.91772e+08, MSE=8.191293e+16
##
## Node number 55: 11 observations
##   mean=1.040359e+09, MSE=1.228549e+17
```

Variable importance budget_range: 54, votes: 29, score_range: 5, runtime: 4, rating: 3, year: 2, genre: 2

Na osnovu ovoga možemo da vidimo koje su kolone bitne u modelu. Veći broj znači da je kolona bitnija za model - budget_range je najvažnija, dok je country_group najmanje važna kolona. Na osnovu grafika mogu da vidim da je stablo u kreiranje modela uključilo kolone: budget_range, votes, genre i year.

Sada koristim metrike MAE, MSE i RMSE za procenu performansi modela.

```
predictions <- predict(lm.fit3, test)

mae <- mean(abs(predictions-test$gross))
mse <- mean((predictions-test$gross)^2)
rmse <-sqrt(mse)
r2 <- 1- (sum((test$gross - predictions)^2) / sum((test$gross-
mean(test$gross))^2))

cat("MAE:", mae, "\n")

## MAE: 50476249

cat("MSE:", mse, "\n")

## MSE: 1.097494e+16

cat("RMSE:", rmse, "\n")

## RMSE: 104761339

cat("R2:", r2, "\n")

## R2: 0.6132627
```

MAE(Mean Absolute Error) meri prosečnu grešku između stvarnih i predviđenih vrednosti. Greška između stvarnih i predviđenih vrednosti iznosi oko 50 miliona. MSE(Mean Squared Error) meri prosečnu kvadratnu grešku između stvarnih i predviđenih vrednosti. U ovom slučaju ona iznosi 1.097494e+16 što ukazuje na velike greške u modelu. RMSE(Root Mean Squared Error) koristi se za procenu veličine greške u istim jedinicama kao originalni podaci. Prosečna greška modela je oko 104 miliona. R2 iznosi ~61%.

Feature selection preko stabla odlučivanja

```
print(lm.fit3$variable.importance)

##  budget_range         votes   score_range       runtime        rating
##  6.765112e+19  3.612802e+19  6.416440e+18  5.128258e+18  3.241647e+18
##          year         genre country_group
##  3.109768e+18  2.877343e+18  1.752787e+17
```

Ovde potvrđujemo da su budget_range i votes najbitniji prediktori. Sada ću samo njih ubaciti u model.

```
lm.fit3.1 <- rpart(gross ~ votes + budget_range,  method = "anova", data =
train)
lm.fit3.1

## n= 6014
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 6014 1.628424e+20   77935960
##    2) budget_range=0-1M,1-10M,10-50M,50-100M,NA 5662 5.208556e+19
52742280
##      4) votes< 115500 4672 1.012814e+19   29138820
##        8) budget_range=0-1M,1-10M,10-50M,NA 4322 5.728160e+18   23350290 *
##        9) budget_range=50-100M 350 2.466870e+18  100618900 *
##      5) votes>=115500 990 2.707105e+19  164131500
##       10) budget_range=0-1M,1-10M,10-50M,NA 691 9.416010e+18  120692500 *
##       11) budget_range=50-100M 299 1.333785e+19  264520600
##         22) votes< 495500 261 7.668847e+18  226636300 *
##         23) votes>=495500 38 2.721555e+18  524725800 *
##    3) budget_range=100M+ 352 4.935602e+19  483182200
##      6) votes< 349500 226 1.530793e+19  344381100
##       12) votes< 137000 78 1.236947e+18  204393600 *
##       13) votes>=137000 148 1.173687e+19  418158400 *
##      7) votes>=349500 126 2.188439e+19  732142800
##       14) votes< 786500 108 6.945836e+18  667774300 *
##       15) votes>=786500 18 1.180622e+19 1118354000 *

predictions <- predict(lm.fit3.1, test)

mae <- mean(abs(predictions-test$gross))
mse <- mean((predictions-test$gross)^2)
rmse <-sqrt(mse)
r2 <- 1- (sum((test$gross - predictions)^2) / sum((test$gross-
mean(test$gross))^2))

cat("MAE:", mae, "\n")

## MAE: 49773938

cat("MSE:", mse, "\n")

## MSE: 1.033862e+16

cat("RMSE:", rmse, "\n")

## RMSE: 101678994

cat("R2:", r2, "\n")
```

```
## R2: 0.6356855
```

Sa selekcijom prediktora se dobija malo poboljšanje R2 - malopre je bilo ~61%, sada je ~64%. Isto tako MAE, MSE i RMSE imaju nešto niže vrednosti.

Random Forest

```
lm.fit4 = randomForest(gross ~ rating + genre + year + votes + runtime +
budget_range + score_range + company_group + country_group, data = train)
lm.fit4

##
## Call:
##  randomForest(formula = gross ~ rating + genre + year + votes +
runtime + budget_range + score_range + company_group + country_group,
data = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 7.220407e+15
##                     % Var explained: 73.33
```

Sada koristim metrike MAE, MSE i RMSE za procenu performansi modela.

```
predictions <- predict(lm.fit4, newdata=test)

mae <- mean(abs(predictions-test$gross))
mse <- mean((predictions-test$gross)^2)
rmse <-sqrt(mse)

cat("MAE:", mae, "\n")

## MAE: 36098917

cat("MSE:", mse, "\n")

## MSE: 7.310756e+15

cat("RMSE:", rmse, "\n")

## RMSE: 85502961
```

Procenat objašnjene varijanse iznosi ~74% što je dosta bolje u odnosu na model linearne regresije i stabla odlučivanja. MAE iznosi oko 36 miliona, MSE iznosi 7.354478e+15, RMSE je oko 85 miliona. Na osnovu ovih metrika ovaj model je najbolji do sada.

Zaključak

Cilj kreiranja dataset-a je bio da se na osnovu analize otkrije da li filmska industrija propada. Na osnovu svih grafika i zaključaka do sad mogu da zaključim da to nije tačno, pogotovo sa većim porastom budžeta i prihoda nakon 2010. godine. Podaci su prikupljeni zaključno sa 2020. godinom tako da nemamo najnovije analize. Što se tiče kreiranja modela i predviđanja kolone prihod, najbolji model koji sam dobila je pomoću Random Forest-a da je procenat objašnjene varijanse ~74%.

Literatura

1. Uvod u nauku o podacima – Predavanja i vezbe
2. https://ggplot2.tidyverse.org/index.html
3. https://www.geeksforgeeks.org/how-to-conduct-an-anderson-darling-test-in-r/
4. https://www.geeksforgeeks.org/decision-tree-for-regression-in-r-programming/
5. https://www.geeksforgeeks.org/random-forest-approach-in-r-programming/?ref=lbp