

Projekt iz kolegija Statistička analiza podataka: Analiza filmova na IMDb-u

Varijacija

13 01 2022

Uvod

U ovom projektu analiziramo skup podataka o filmovima sa stranice IMDb. Podatke, koji se nalaze u CSV datoteci IMDB.csv, učitali smo u varijablu imdb. Pogledajmo kako izgleda naš set podataka - koje su mu dimenzije, kako se zovu njegove varijable i kojeg su tipa podataka.

```
# Dimenzije tablice podataka

cat("No. of rows:", nrow(imdb), '\n')

## No. of rows: 5043
cat("No. of columns:", ncol(imdb), '\n')

## No. of columns: 28
# Imena i tipovi varijabli

cat("\nVariables and their types:\n\n")

##
## Variables and their types:
sapply(imdb, class)

##          color      director_name num_critic_for_reviews
## "character"      "character"           "integer"
##          duration director_facebook_likes actor_3_facebook_likes
## "integer"          "integer"           "integer"
##         actor_2_name   actor_1_facebook_likes        gross
## "character"                  "integer"           "integer"
##          genres       actor_1_name      movie_title
## "character"                  "character"           "character"
##         num_voted_users cast_total_facebook_likes   actor_3_name
## "integer"                      "integer"           "character"
## facenumber_in_poster      plot_keywords      movie_imdb_link
## "integer"                  "character"           "character"
##         num_user_for_reviews        language      country
## "integer"                  "character"           "character"
##          content_rating        budget      title_year
## "character"                 "numeric"           "integer"
##     actor_2_facebook_likes      imdb_score aspect_ratio
## "integer"                  "numeric"           "numeric"
```

```
##      movie_facebook_likes
##              "integer"
```

Naš set podataka ima 5043 redaka (unosa) i 28 stupaca (varijabli).

Varijable su sljedeće:

- movie_title - naslov filma
- director_name - ime redatelja filma
- title_year - godina premijere filma
- actor_1_name - ime glumca br. 1
- actor_2_name - ime glumca br. 2
- actor_3_name - ime glumca br. 3
- genres - žanrovi kojima film pripada odvojeni znakom ‘|’
- content_rating - MPAA oznaka filma
- plot_keywords - ključne riječi koje opisuju radnju filma odvojene znakom ‘|’
- duration - trajanje filma u minutama
- language - jezik filma
- country - država u kojoj je produciran film
- budget - budžet filma u dolarima
- gross - zarada filma u dolarima
- color - informacija je li film crno-bijeli ili u boji
- aspect_ratio - format filma
- facenumber_in_poster - broj glumaca koji su prikazani na posteru filma
- movie_imdb_link - poveznica na IMDB stranicu filma
- imdb_score - IMDB-ova ocjena filma
- num_critic_for_reviews - broj filmskih kritičara koji su recenzirali film
- num_voted_users - broj korisnika IMDB-a koji su glasali za film
- num_user_for_reviews - broj korisnika IMDB-a koji su recenzirali film
- movie_facebook_likes - ukupan broj Facebook lajkova koje je film dobio
- director_facebook_likes - broj Facebook lajkova na Facebook stranici direktora
- actor_1_facebook_likes - broj Facebook lajkova na Facebook stranici glumca br. 1
- actor_2_facebook_likes - broj Facebook lajkova na Facebook stranici glumca br. 2
- actor_3_facebook_likes - broj Facebook lajkova na Facebook stranici glumca br. 3
- cast_total_facebook_likes - ukupan broj Facebook lajkova sa Facebook stranica svih glumaca iz postave filma

Uočavamo da metrički podaci čine većinu našeg skupa podataka, npr. varijable duration, budget, gross i sl. Neke od kategoriskih varijabli su movie_title, director_name, genres itd. Varijabla movie_imdb_link za nas nema značaj pa ćemo je maknuti iz našeg seta podataka.

```
# Izbacivanje "movie_imdb_link"
```

```
imdb <- imdb[-c(18)]
dim(imdb)
```

```
## [1] 5043   27
```

Pogledajmo koliki postotak nepostojećih vrijednosti imaju preostale varijable.

```
# NA vrijednosti
```

```
for (variable in names(imdb)) {
  nas = sum(is.na(imdb[variable]))
  p = round(nas / nrow(imdb) * 100, 2)

  cat(variable, ":", nas, '(', p, '%)\n')
}
```

```

## color : 0 ( 0 %)
## director_name : 0 ( 0 %)
## num_critic_for_reviews : 50 ( 0.99 %)
## duration : 15 ( 0.3 %)
## director_facebook_likes : 104 ( 2.06 %)
## actor_3_facebook_likes : 23 ( 0.46 %)
## actor_2_name : 0 ( 0 %)
## actor_1_facebook_likes : 7 ( 0.14 %)
## gross : 884 ( 17.53 %)
## genres : 0 ( 0 %)
## actor_1_name : 0 ( 0 %)
## movie_title : 0 ( 0 %)
## num_voted_users : 0 ( 0 %)
## cast_total_facebook_likes : 0 ( 0 %)
## actor_3_name : 0 ( 0 %)
## facenumber_in_poster : 13 ( 0.26 %)
## plot_keywords : 0 ( 0 %)
## num_user_for_reviews : 21 ( 0.42 %)
## language : 0 ( 0 %)
## country : 0 ( 0 %)
## content_rating : 0 ( 0 %)
## budget : 492 ( 9.76 %)
## title_year : 108 ( 2.14 %)
## actor_2_facebook_likes : 13 ( 0.26 %)
## imdb_score : 0 ( 0 %)
## aspect_ratio : 329 ( 6.52 %)
## movie_facebook_likes : 0 ( 0 %)

```

Većina stupaca sadrži samo malo nepostojećih vrijednosti osim stupca sa zaradom filmova koji sadrži 884 nepostojeće vrijednosti što je gotovo 20% ukupnog broja filmova te zaključci koje bismo donijeli na temelju tih podataka ne bi najbolje prikazivali stvarnu sliku.

Deskriptivna statistika

Kategoriskske varijable

Varijabla genres

Za varijablu genres već smo ranije utvrdili da nema nedostajućih vrijednosti, stoga ne moramo paziti na te vrijednosti pri radu s podacima.

Prvo moramo svaki redak razdvojiti na više njih ovisno o njegovom broju žanrova.

```
# Separacija žanrova

imdb_sep_bygenres <- separate_rows(imdb, genres, sep = "\\|")
```

Najprije pogledajmo koliko je različitih žanrova te koji je najzastupljeniji, a koji je najmanje zastupljen.

```
occurences = imdb_sep_bygenres$genres
occurences_df = as.data.frame(table(occurences))
colnames(occurences_df) = c('genre', 'occurences')
occurences_df = occurences_df[order(occurences_df$occurences, decreasing = TRUE), ]

cat("No. of Genres:", nrow(occurences_df))
```

```
## No. of Genres: 26
```

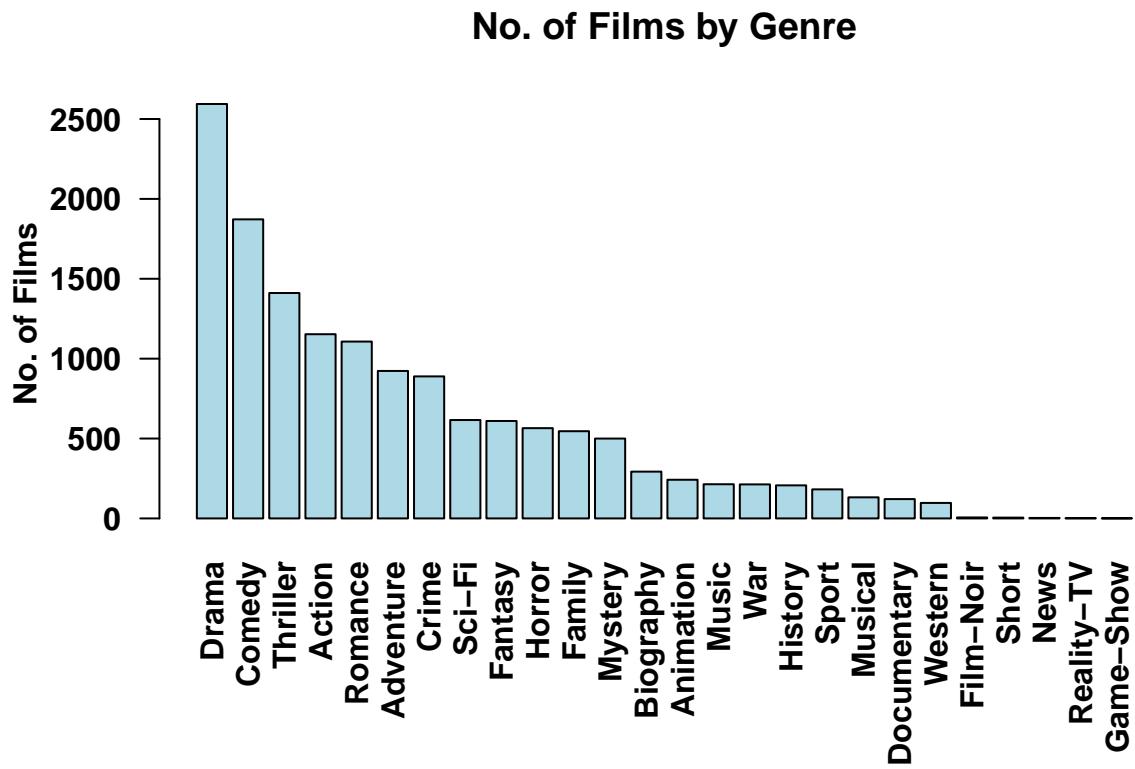
```

par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)

# Stupčasti dijagram

barplot(height = occurences_df$occurrences,
        names.arg = occurences_df$genre,
        main = "No. of Films by Genre",
        ylab = "No. of Films",
        las = 2, col = "Lightblue")

```



U našoj tablici podataka pojavljuje se ukupno 26 različitih žanrova. Žanr s najviše snimljenih filmova jest "Drama". Također, primjećujemo da žanrovi "Film-Noir", "Short", "News", "Reality-TV" i "Game-Show" jako mali broj snimljenih filmova u odnosu na ostatak žanrova te zato odbacujemo ove žanrove iz daljnje analize zbog premalog broja podataka.

```

# Izbacivanje vrijednosti "Film-Noir", "Short", "News", "Reality-TV" i "Game-Show"

tbd <- c('Film-Noir', 'Short', 'News', 'Reality-TV', 'Game-Show')
genres_edited <- imdb_sep_bygenres[!(imdb_sep_bygenres$genres %in% tbd), ]

```

Varijabla plot_keywords

Za varijablu plot_keywords već smo ranije utvrdili da ima 153 nedostajuće vrijednosti, stoga moramo pripaziti da te vrijednosti isključimo iz analize ključnih riječi.

Prvo moramo svaki redak razdvojiti na više njih ovisno o njegovom broju ključnih riječi.

```

# Separacija ključnih riječi

imdb_sep_bykeywords <- separate_rows(imdb[imdb$plot_keywords != " ", ], plot_keywords, sep = "\\|")

Najprije pogledajmo koliko je različitih ključnih riječi te koji su najzastupljenije.

occurrences = imdb_sep_bykeywords$plot_keywords
occurrences_df = as.data.frame(table(occurrences))
colnames(occurrences_df) = c('keyword', 'occurrences')
occurrences_df = occurrences_df[order(occurrences_df$occurrences, decreasing = TRUE), ]

cat("No. of rows:", nrow(occurrences_df))

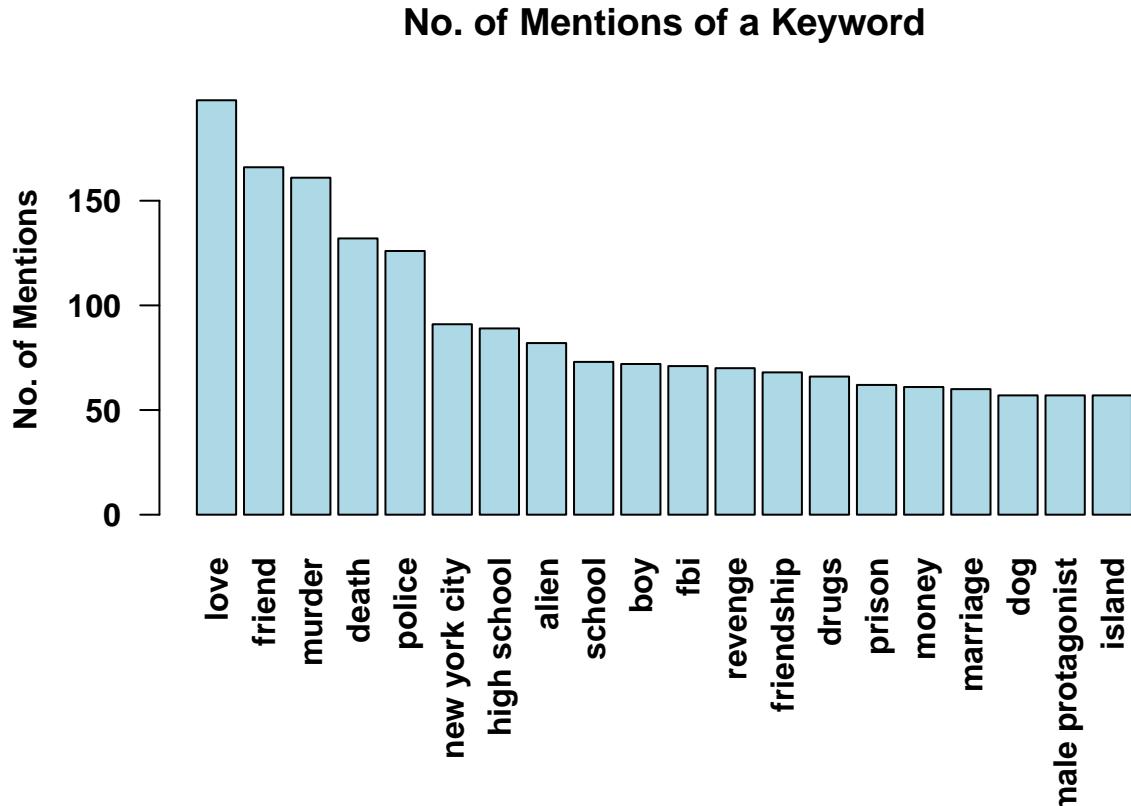
## No. of rows: 8086

par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)

# Stupčasti dijagram

barplot(height = occurrences_df[1:20, "occurrences"],
       names.arg = occurrences_df[1:20, "keyword"],
       main = "No. of Mentions of a Keyword",
       ylab = "No. of Mentions",
       las = 2, col = "Lightblue")

```



U našoj tablici podataka pojavljuje se ukupno 8086 različitih ključnih riječi. Tri najzastupljenije ključne riječi su "love", "friend" i "murder", što i ima smisla ako se prisjetimo da je najzastupljeniji žanr filmova drama.

Varijabla country

Pogledajmo koliko je različitih država te u kojoj je državi producirano najviše filmova, pritom zanemarujemo nedostajuće vrijednosti.

```
occurences = imdb[!is.na(imdb$country), "country"]
occurences_df = as.data.frame(table(occurences))
colnames(occurences_df) = c('country', 'occurences')
occurences_df = occurences_df[order(occurences_df$occurences, decreasing = TRUE), ]

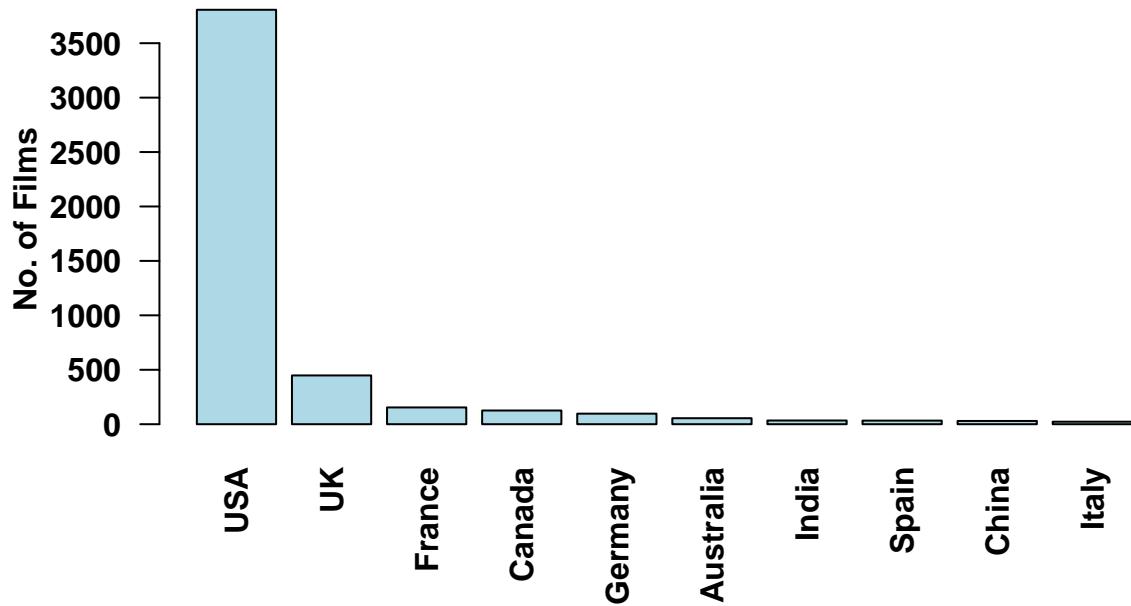
cat("No. of rows:", nrow(occurences_df))

## No. of rows: 66
par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)

# Stupčasti dijagram

barplot(height = occurences_df[1:10, "occurences"],
       names.arg = occurences_df[1:10, "country"],
       main = "No. of Films per Country",
       ylab = "No. of Films",
       las = 2, col = "Lightblue")
```

No. of Films per Country



Ukupno je 66 različitih država, pri čemu je SAD daleko najzastupljeniji, što je razumljivo s obzirom na razvijenost filmske industrije u SAD-u.

Pogledajmo koliko je točno filmova snimljeno u SAD-u.

```

n = occurences_df[occurences_df$country == 'USA', 'occurences']
p = n / nrow(imdb) * 100

cat("No. of films produced in USA:", n, "(", p, "%)")

## No. of films produced in USA: 3807 ( 75.49078 %)

```

Kako vrijednost "USA" sačinjava i više od 75% svih podataka, ova nam varijabla ne daje toliku raznolikost podataka. Zato ćemo u daljnjoj analizi filmove razdvajati na američke i neameričke.

Metričke varijable

Varijabla title_year

Pogledajmo kako se ponaša varijabla title_year.

```

summary(imdb$title_year)

##      Min.   1st Qu.   Median     Mean   3rd Qu.   Max.   NA's
##      1916    1999    2005    2002    2011    2016    108

```

Najstariji zabilježeni film je iz 1916. godine, a "najmlađi" je iz 2016. "Mlada" polovica filmova imala je premijeru u periodu između 2005. i 2016. godine, dok je interval za drugu polovicu filmova dug gotovo 90 godina. To zapažanje ima smisla s obzirom na razvoj filmske industrije od samih početaka pa sve do danas.

Također, možemo opaziti da je medijan veći od aritmetičke sredine iz čega možemo pretpostaviti da je distribucija lijevo (negativno) zakriviljena. Tu tvrdnju možemo potvrditi i histogramom.

```
par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)
```

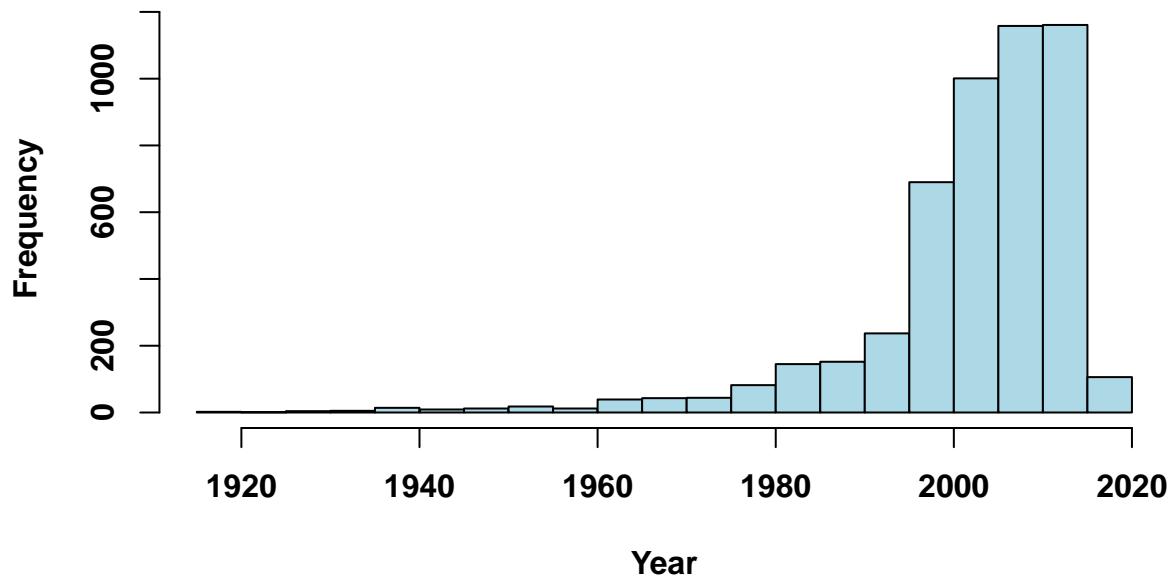
Histogram

```

hist(imdb$title_year,
      main="Title Year Histogram",
      xlab="Year",
      ylab='Frequency',
      breaks = 20, col="lightblue")

```

Title Year Histogram



Varijabla duration

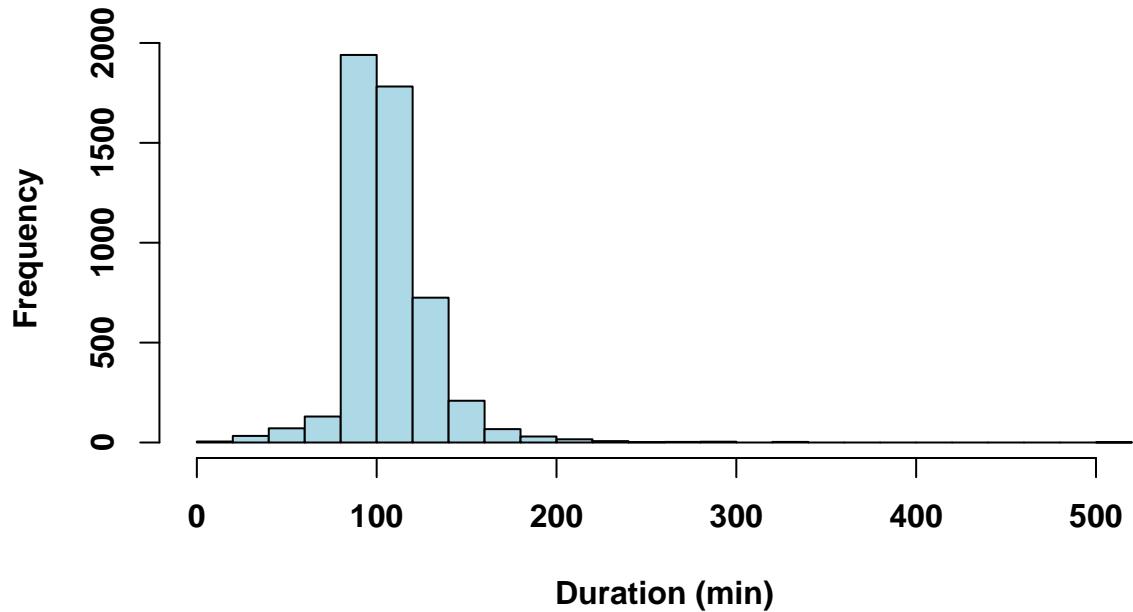
Prvo ćemo pogledati histogram i pravokutni dijagram.

```
par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)

# Histogram

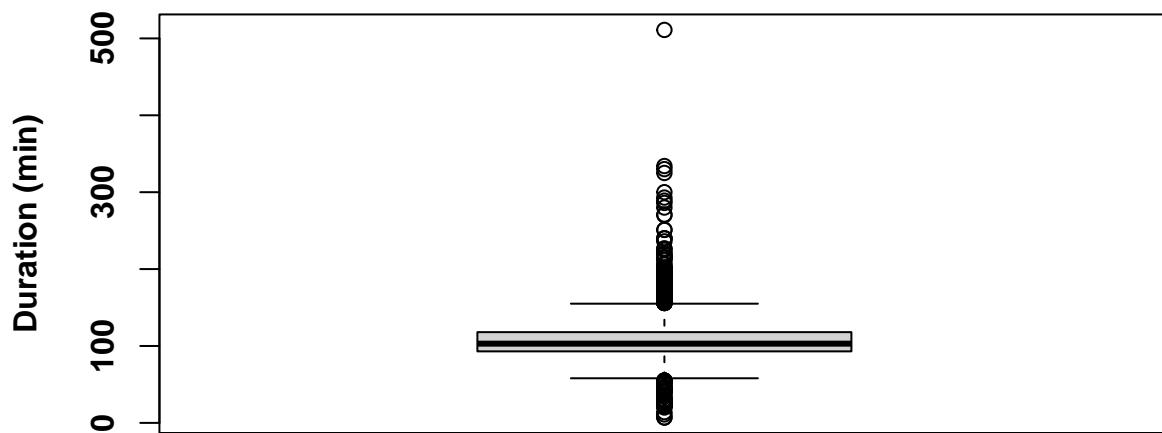
hist(imdb$duration,
      breaks=20,
      main="Film's duration in minutes",
      xlab="Duration (min)",
      ylab='Frequency',
      col = 'lightblue')
```

Film's duration in minutes



```
# Pravokutni dijagram  
boxplot(imdb$duration,  
        main="Box and whisker plot for duration",  
        ylab="Duration (min)")
```

Box and whisker plot for duration



Možemo vidjeti kako ovaj skup podataka i nije baš najzahvalniji za prikaz pomoću stupčastog dijagrama jer je rang ovog skupa podataka dosta velik te imamo jako puno stršećih vrijednosti. Međutim možemo pretpostaviti kako ova distribucija nalikuje na multimodalnu.

```
# Rang
cat("Rang:", max(imdb[!is.na(imdb$duration)], 'duration')) - min(imdb[!is.na(imdb$duration)], 'duration')

## Rang: 504

Pogledajmo sada mjere centra.

# Medijan
cat("Median:", median(imdb[!is.na(imdb$duration)], 'duration'), "\n")

## Median: 103

# Podrezana aritmetička sredina s uklanjanjem po 10% najmanjih i najvećih podataka
cat("10% trimmed mean:", mean(imdb[!is.na(imdb$duration)], 'duration'], trim=0.1), "\n")

## 10% trimmed mean: 105.2393

# Aritmeticka sredina
cat("Mean:", mean(imdb[!is.na(imdb$duration)], 'duration']), "\n")

## Mean: 107.2011

# Mod
require(modeest)
cat("Mod:", mfv(imdb[!is.na(imdb$duration)], 'duration')), "\n")
```

```

## Mod: 90

Zbog velikog broja stršećih vrijednosti najbolji pokazatelj centra je mod koji je robustan i neosjetljiv na ekstreme kojih u ovom skupu podataka ima mnogo.

Nadalje pogledajmo mjere rasipanja.

# IQR
cat("IQR:", IQR(imdb[(!is.na(imdb$duration)), 'duration']), "\n")

## IQR: 25

# Varijanca
cat("Variance:", var(imdb[(!is.na(imdb$duration)), 'duration']), "\n")

## Variance: 634.911

# Standardna devijacija
cat("Standard deviation:", sd(imdb[(!is.na(imdb$duration)), 'duration']), "\n")

## Standard deviation: 25.19744

# Koeficijent varijacije
cat("Variation coefficient:", sd(imdb[(!is.na(imdb$duration)), 'duration'])/mean(imdb[(!is.na(imdb$duration)), 'duration']), "\n")

## Variation coefficient: 0.2350484

Ovdje je zanimljivo pogledati interkvartilni rang koji je izuzetno mali u odnosu na rang cijelog skupa podataka. Iznosi tek 25 što je i očekivano s obzirom da većina filmova traje od otprilike 90 do 120 minuta. Upravo odavdje dolazi velik broj stršećih vrijednosti.

Pronadimo još najkraći i najdulji film te njihova trajanja.

# Uređivanje naslova filma

for (i in 1:nrow(imdb)) {
  row = imdb[i, 'movie_title']
  imdb[i, 'movie_title'] <- str_split(row, 'Ã')[[1]][1]
}

# Najkraći i najdulji film

lenmin = min(imdb[(!is.na(imdb$duration)), 'duration'])
lenmax = max(imdb[(!is.na(imdb$duration)), 'duration'])

moviemin = imdb[(!is.na(imdb$duration)) & (imdb$duration == lenmin), 'movie_title']
moviemax = imdb[(!is.na(imdb$duration)) & (imdb$duration == lenmax), 'movie_title']

cat('The shortest movie in dataset:\n', moviemin, '\nlasts:', lenmin, 'minutes\n')

## The shortest movie in dataset:
## Shaun the Sheep The Touch
## lasts: 7 minutes

cat('\nThe longest movie in dataset:\n', moviemax, '\nlasts:', lenmax, 'minutes\n')

##
## The longest movie in dataset:
## Trapped
## lasts: 511 minutes

```

Primjećujemo da dva filma dijele titulu najkraćeg filma: "Shaun the Sheep" i "The Touch", a najdulji film je "Trapped" koji traje više od 8 sati (vjerojatno je riječ o mini-seriji).

Varijabla budget

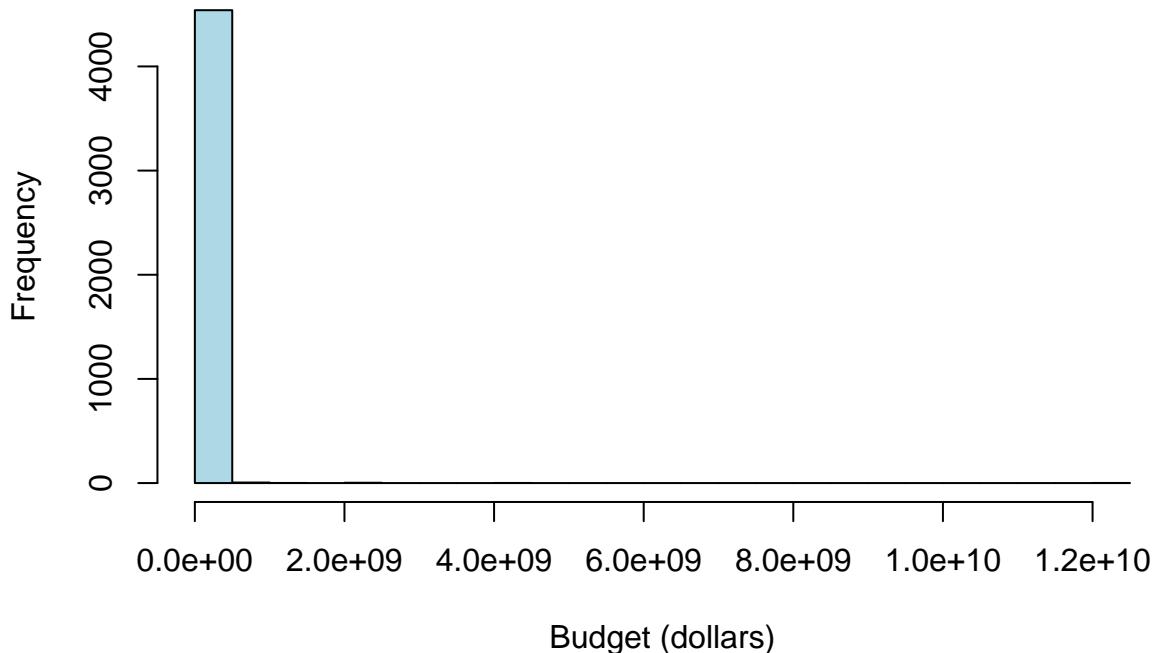
Varijabla budget ima 492 nedostajuće vrijednosti pa moramo pripraziti da ih ne uključimo u daljnju analizu.

Prvo ćemo pogledati histogram i pravokutni dijagram.

```
# Histogram
```

```
hist(imdb$budget,
      breaks=20,
      main="Film's budget in dollars",
      xlab="Budget (dollars)",
      ylab='Frequency',
      col = 'lightblue')
```

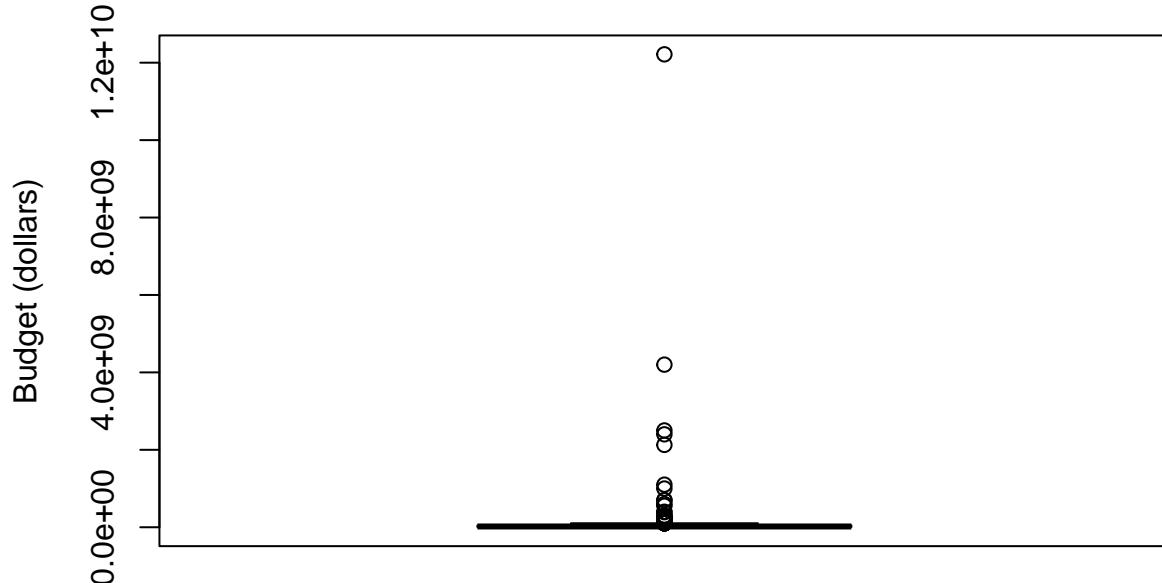
Film's budget in dollars



```
# Pravokutni dijagram
```

```
boxplot(imdb$budget,
        main="Box and whisker plot for budget",
        ylab="Budget (dollars)")
```

Box and whisker plot for budget



```
# Rang
cat("Rang:", max(imdb[(!is.na(imdb$budget)), 'budget']) - min(imdb[(!is.na(imdb$budget)), 'budget']))
```

Rang: 12215499782

Rang distribucije budžeta filmova je jako velik te imamo povelik broj stršećih vrijednosti, zbog čega ne možemo puno zaključiti iz histograma.

Pogledajmo mjere centra.

```
# Medijan
cat("Median:", median(imdb[(!is.na(imdb$budget)), 'budget']), "\n")
```

Median: 2e+07

Podrezana aritmetička sredina s uklanjanjem po 10% najmanjih i najvećih podataka
cat("10% trimmed mean:", mean(imdb[(!is.na(imdb\$budget)), 'budget'], trim=0.1), "\n")

```
## 10% trimmed mean: 25098686
```

Aritmeticka sredina
cat("Mean:", mean(imdb[(!is.na(imdb\$budget)), 'budget']), "\n")

Mean: 39752620

```
# Mod
require(modeest)
cat("Mod:", mfv(imdb[(!is.na(imdb$budget)), 'budget']), "\n")
```

Mod: 2e+07

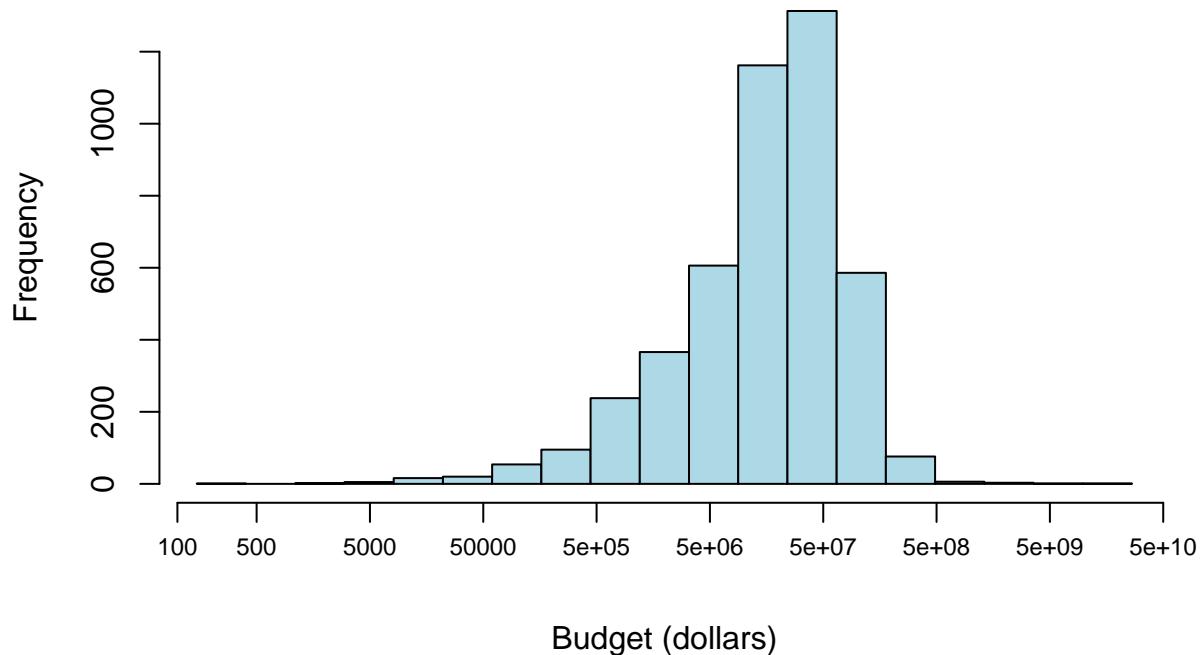
Uočavamo da su medijan i mod jednaki i "blizu" vrijednosti podrezane aritmetičke sredine. Dakle, distribucija bi mogla nalikovati normalnoj distribuciji.

Skalirajmo podatke prirodnim logaritmom.

```
# Histogram (log)
```

```
h = hist(log(imdb$budget),
          axes = FALSE,
          main="Film's budget in dollars (log)",
          xlab="Budget (dollars)",
          ylab='Frequency',
          breaks = 20, col="lightblue")
axis(side = 1,
     at = log(c(100, 500, 5000, 50000, 500000, 5000000, 50000000, 500000000, 5000000000, 50000000000, 500000000000)),
     labels = paste(c(100, 500, 5000, 50000, 500000, 5000000, 50000000, 500000000, 5000000000, 50000000000, 500000000000)),
     cex.axis = 0.75,
     padj = -1,
     hadj = 0.5,
     las = 1)
axis(2)
```

Film's budget in dollars (log)



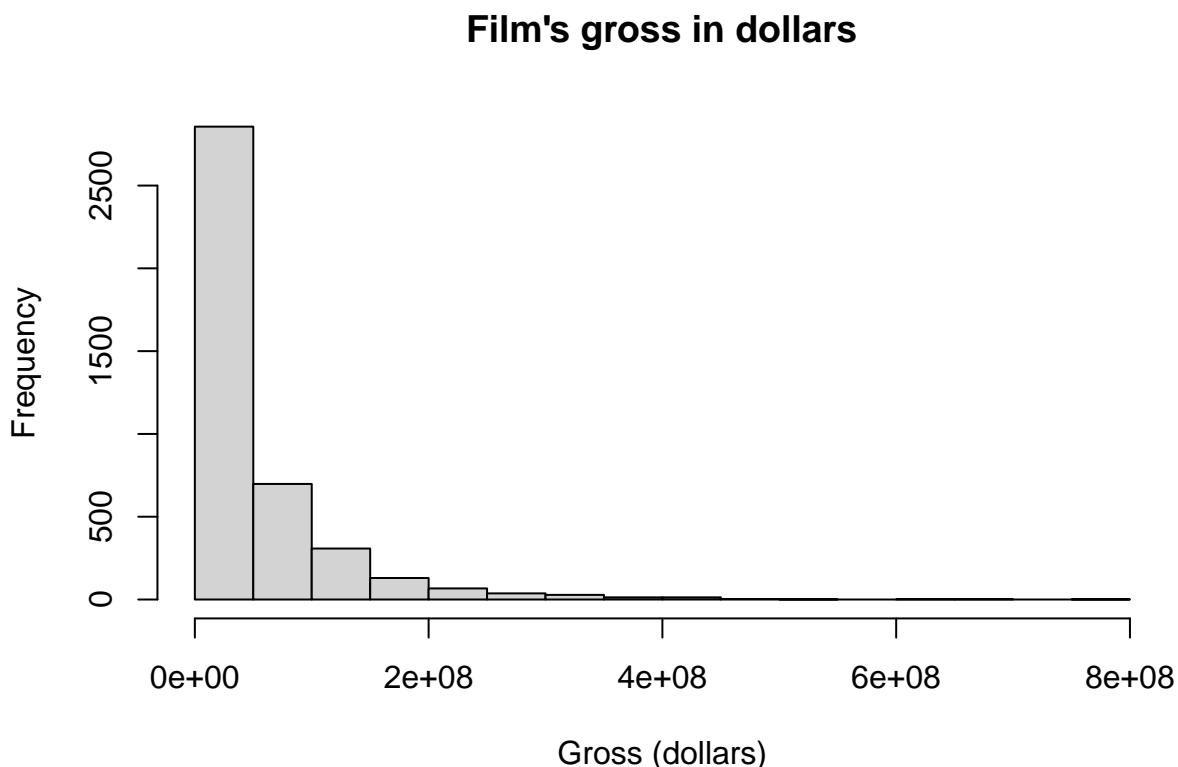
Sada puno bolje možemo vidjeti distribuciju budžeta. Distribucija nalikuje normalnoj, no malo je zakriviljena u lijevo.

Varijabla gross

Prvo ćemo pogledati histogram i pravokutni dijagram.

```
# Histogram

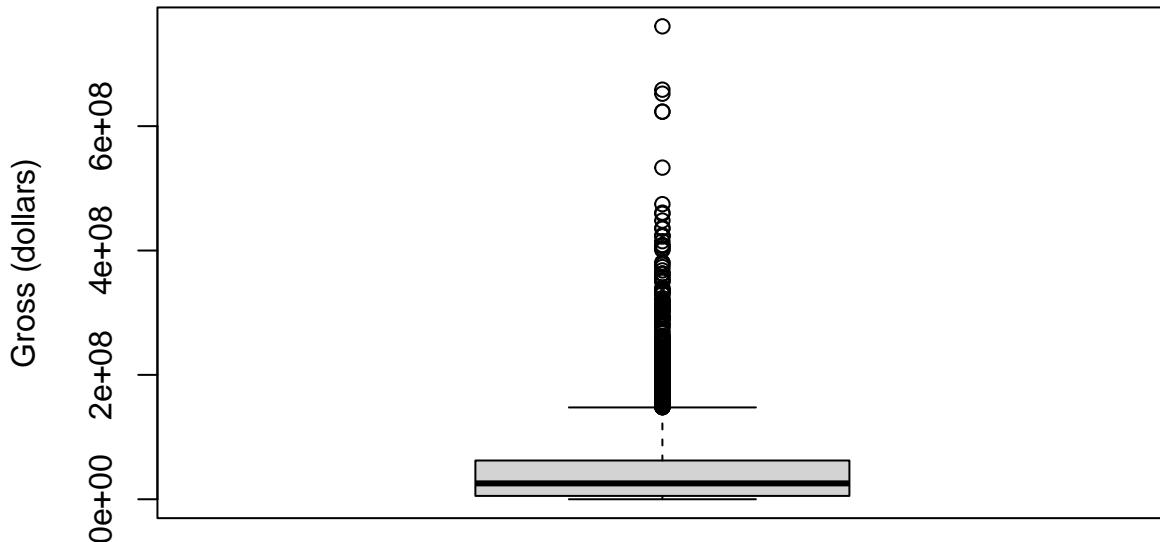
hist(imdb$gross,
      breaks=20,
      main="Film's gross in dollars",
      xlab="Gross (dollars)",
      ylab='Frequency' )
```



```
# Pravokutni dijagram

boxplot(imdb$gross,
        main="Box and whisker plot for gross",
        ylab="Gross (dollars)")
```

Box and whisker plot for gross



```
# Rang
cat("Rang:", max(imdb[(!is.na(imdb$gross)), 'gross']) - min(imdb[(!is.na(imdb$gross)), 'gross'])))

## Rang: 760505685
Slično kao i kod budžeta, imamo veliki rang i pozamašan broj stršećih vrijednosti.

Pogledajmo mjere centra.

# Medijan
cat("Median:", median(imdb[(!is.na(imdb$gross)), 'gross']), "\n")

## Median: 25517500
# Podrezana aritmetička sredina s uklanjanjem po 10% najmanjih i najvećih podataka
cat("10% trimmed mean:", mean(imdb[(!is.na(imdb$gross)), 'gross'], trim=0.1), "\n")

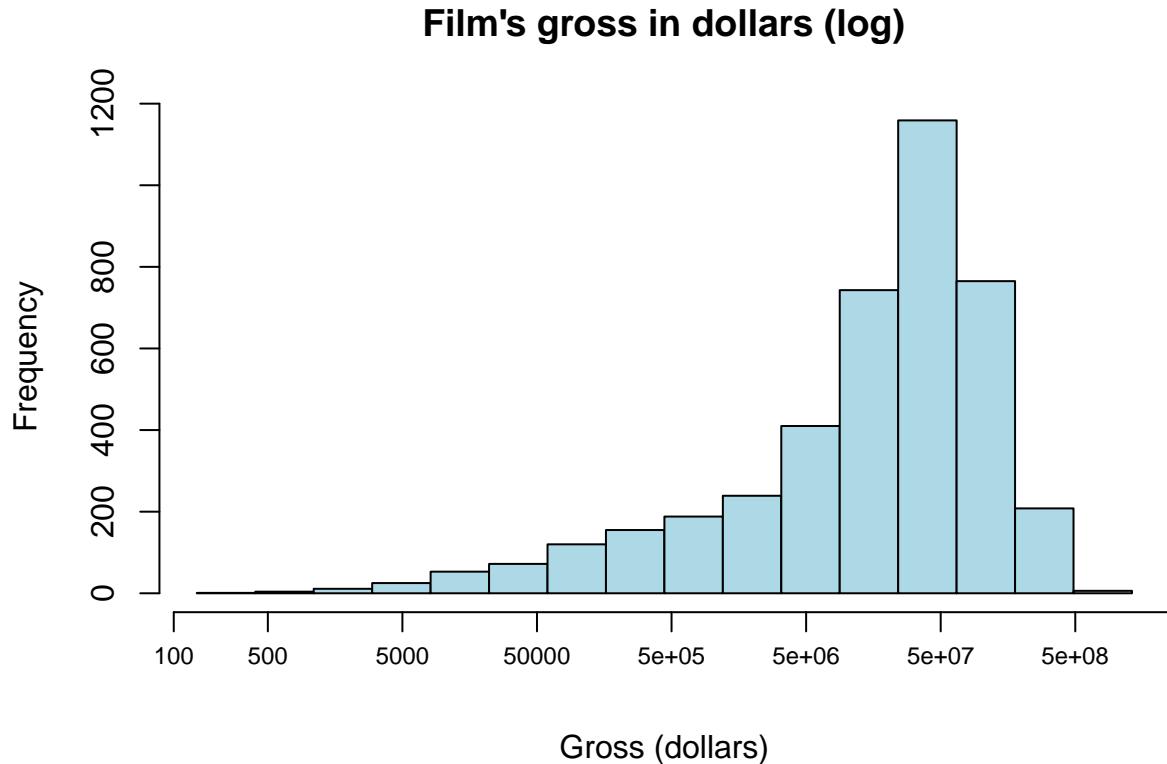
## 10% trimmed mean: 34169552
# Aritmeticka sredina
cat("Mean:", mean(imdb[(!is.na(imdb$gross)), 'gross']), "\n")

## Mean: 48468408
# Mod
require(modeest)
cat("Mod:", mfv(imdb[(!is.na(imdb$gross)), 'gross']), "\n")

## Mod: 3000000 5773519 8000000 34964818 47000000 144512310 177343675 218051260
```

Medijan i podrezana aritmetička sredina su također naoko "blizu" pa bismo mogli i ovdje skalirati podatke prirodnim logaritmom kako bismo bolje vizualizirali podatke.

```
h = hist(log(imdb$gross),
          axes = FALSE,
          main="Film's gross in dollars (log)",
          xlab="Gross (dollars)",
          ylab='Frequency',
          breaks = 20, col="lightblue")
axis(side = 1,
     at = log(c(100, 500, 5000, 50000, 500000, 5000000, 50000000, 500000000, 5000000000, 50000000000, 500000000000)),
     labels = paste(c(100, 500, 5000, 50000, 500000, 5000000, 50000000, 500000000, 5000000000, 50000000000, 500000000000)),
     cex.axis = 0.75,
     padj = -1,
     hadj = 0.5,
     las = 1)
axis(2)
```



Sada svakako bolje vidimo kako su distribuirani podaci o zaradi filmova. Vidimo da je riječ o lijevo (negativno) zakriviljenoj distribuciji.

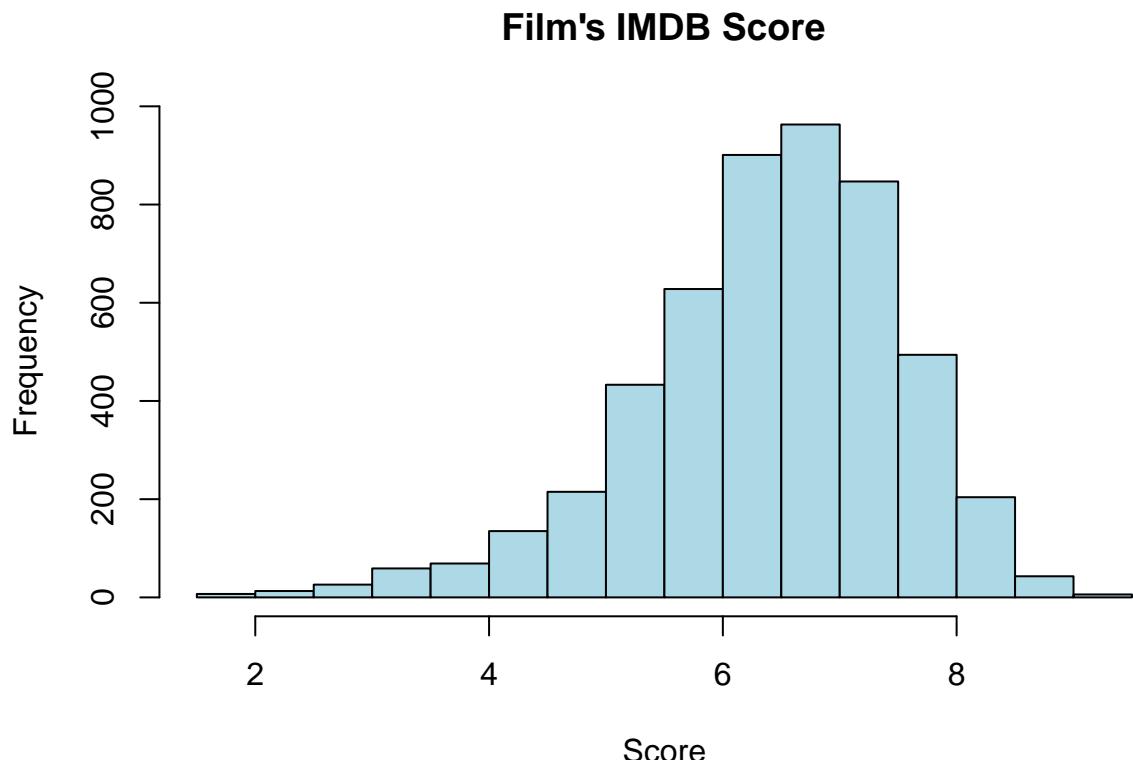
Varijabla `imdb_score`

Prehodno smo uvidjeli da varijabla `imdb_score` nema nedostajućih vrijednosti.

Pogledajmo prvo histogram i pravokutni dijagram.

```
# Histogram

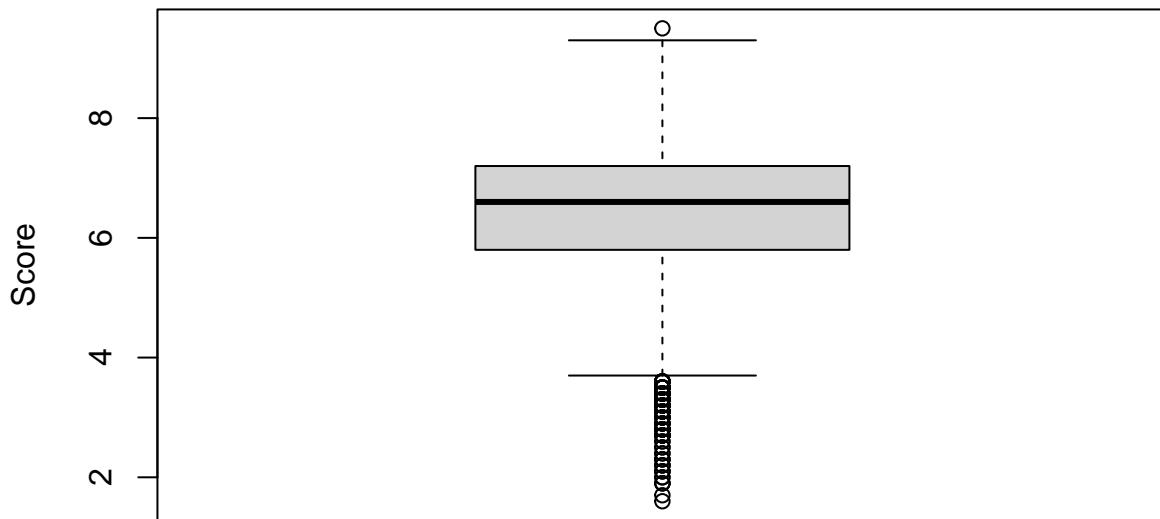
hist(imdb$imdb_score,
      breaks=20,
      main="Film's IMDB Score",
      xlab="Score",
      ylab='Frequency',
      col = 'lightblue')
```



```
# Pravokutni dijagram

boxplot(imdb$imdb_score,
        main="Box and whisker plot for IMDB Score",
        ylab="Score")
```

Box and whisker plot for IMDB Score



Pogledajmo prosječne ocjene uz pojedini žanr. Za mjeru centra koristili smo aritmetičku sredinu.

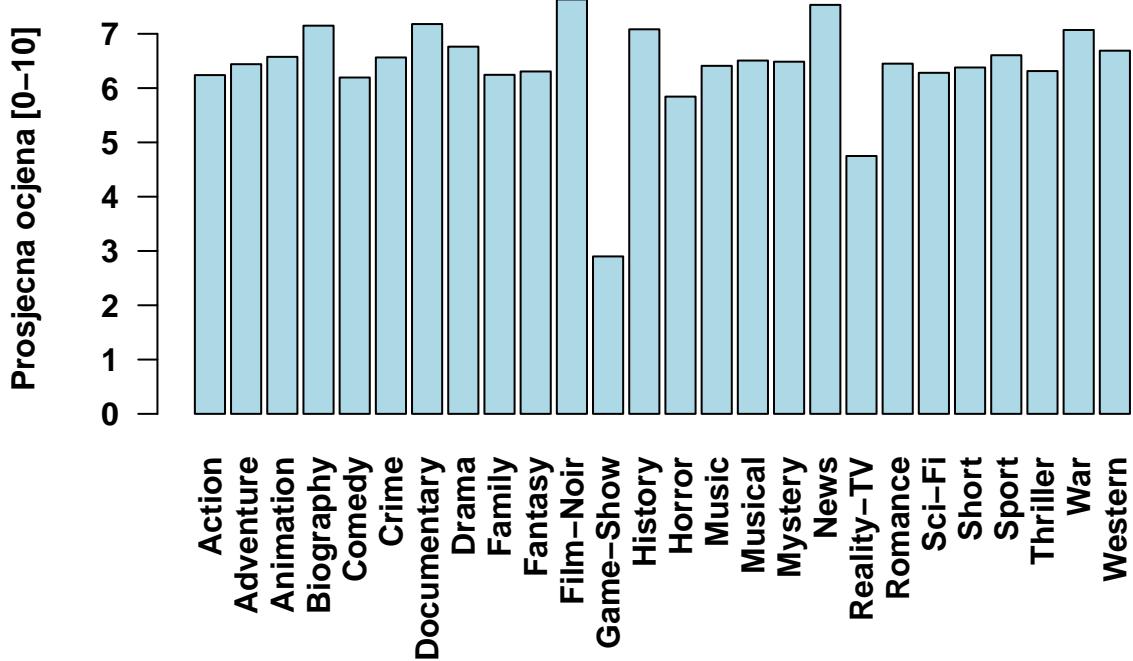
```
imdb_sep_bygenres %>% group_by(genres) %>% summarise(AVG_imdb_rating = mean(imdb_score)) -> summary.result

##postavljanje podataka na barplot
data_bar <- summary.result.avg$AVG_imdb_rating
names(data_bar) <- summary.result.avg$genres

##uredjivanje margeina
par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)

barplot(data_bar,
        main="Prosječna ocjena filmova za pojedine žanrove",
        ylab = "Prosječna ocjena [0-10]",
        las = 2, col="lightblue")
```

Prosjecna ocjena filmova za pojedine žanrove



Gore dobiveni dijagram jako podsjeća na uniformnu distribuciju. Pogledajmo sada standardne devijacije za pojedini žanr.

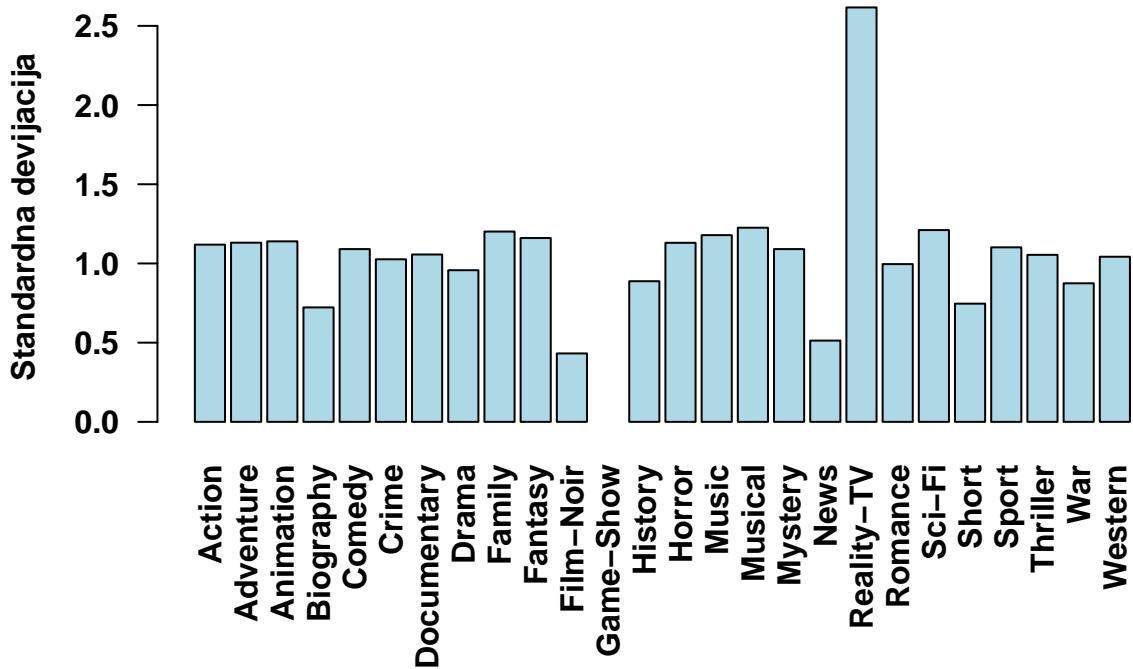
```
imdb_sep_bygenres %>% group_by(genres) %>% summarise(AVG_imdb_rating = sd(imdb_score)) -> summary.result

data_bar <- summary.result$AVG_imdb_rating
names(data_bar) <- summary.result$genres

par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)

barplot(data_bar,
        main="Standardna devijacija ocjena filmova za pojedine žanrove",
        ylab = "Standardna devijacija",
        las = 2, col="lightblue")
```

Standardna devijacija ocjena filmova za pojedine žanrove



Vidimo kako samo žanr "Biography" značajnije odstupa od drugih standardnih varijacija, no i dalje možemo prepostaviti kako se radi o uniformnoj distribuciji.

Variable movie_facebook_likes, director_facebook_likes, actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes, cast_total_facebook_likes

Prethodno smo uočili da je broj lajkova koje je dobio film zabilježen za svaki unos (redak) u tablici podataka. Isto vrijedi i za broj lajkova koje je dobila cijelokupna glumačka postava. Broj nezabilježenih unosa lajkova glavnih glumaca redom iznose: 7 (~0.14%) za prvog glumca, 13 (~0.26%) za drugog glumca te 23 (0.46%) za trećeg glumca. Najveći broj nedostajućih vrijednosti ima stupac lajkova koji je dobio redatelj - 104 (~2.12%).

Riječ je o jako malim udjelima pa možemo maknuti sve retke u kojima barem jedan od tih stupaca ima nedostajuću vrijednost.

```
likesdata <- imdb[!(is.na(imdb$movie_facebook_likes)
  | is.na(imdb$director_facebook_likes)
  | is.na(imdb$actor_1_facebook_likes)
  | is.na(imdb$actor_2_facebook_likes)
  | is.na(imdb$actor_3_facebook_likes)
  | is.na(imdb$cast_total_facebook_likes)), ]  
  
cat('No. of entries:', nrow(likesdata), '\n')  
  
## No. of entries: 4919
```

Nova tablica bez nedostajućih vrijednosti ima 4797 redaka, što je oko 97% od ukupnog broja redaka stare tablice podataka. Uklonili smo gotovo nezamjetan broj redaka.

Sada kada smo uklonili nedostajuće vrijednosti, pogledajmo kako se ponašaju te varijable.

```
summary(likesdata[c('movie_facebook_likes', 'director_facebook_likes', 'actor_1_facebook_likes', 'actor_2_facebook_likes', 'actor_3_facebook_likes', 'cast_total_facebook_likes')])

## movie_facebook_likes director_facebook_likes actor_1_facebook_likes
## Min. : 0      Min. : 0      Min. : 0
## 1st Qu.: 0      1st Qu.: 7      1st Qu.: 622
## Median : 166     Median : 49     Median : 997
## Mean   : 7614     Mean   : 689     Mean   : 6683
## 3rd Qu.: 3000     3rd Qu.: 197     3rd Qu.: 11000
## Max.  :349000     Max.  :23000     Max.  :640000
## actor_2_facebook_likes actor_3_facebook_likes cast_total_facebook_likes
## Min. : 0      Min. : 0.0      Min. : 0
## 1st Qu.: 284     1st Qu.: 133.0    1st Qu.: 1439
## Median : 600     Median : 372.0    Median : 3141
## Mean   : 1679     Mean   : 651.2    Mean   : 9877
## 3rd Qu.: 922     3rd Qu.: 637.0    3rd Qu.: 14094
## Max.  :137000     Max.  :23000.0   Max.  :656730

cat('\nIQR(Movie Facebook Likes) =', IQR(likesdata$movie_facebook_likes), '\n')

##
## IQR(Movie Facebook Likes) = 3000
cat('sd(Movie Facebook Likes) =', sd(likesdata$movie_facebook_likes), '\n')

## sd(Movie Facebook Likes) = 19481
cat('\nIQR(Director Facebook Likes) =', IQR(likesdata$director_facebook_likes), '\n')

##
## IQR(Director Facebook Likes) = 190
cat('sd(Director Facebook Likes) =', sd(likesdata$director_facebook_likes), '\n')

## sd(Director Facebook Likes) = 2818.742
cat('\nIQR(Actor1 Facebook Likes) =', IQR(likesdata$actor_1_facebook_likes), '\n')

##
## IQR(Actor1 Facebook Likes) = 10378
cat('sd(Actor1 Facebook Likes) =', sd(likesdata$actor_1_facebook_likes), '\n')

## sd(Actor1 Facebook Likes) = 15166.77
cat('\nIQR(Actor2 Facebook Likes) =', IQR(likesdata$actor_2_facebook_likes), '\n')

##
## IQR(Actor2 Facebook Likes) = 638
cat('sd(Actor2 Facebook Likes) =', sd(likesdata$actor_2_facebook_likes), '\n')

## sd(Actor2 Facebook Likes) = 4083.279
cat('\nIQR(Actor3 Facebook Likes) =', IQR(likesdata$actor_3_facebook_likes), '\n')

##
## IQR(Actor3 Facebook Likes) = 504
cat('sd(Actor3 Facebook Likes) =', sd(likesdata$actor_3_facebook_likes), '\n')
```

```

## sd(Actor3 Facebook Likes) = 1681.086
cat('\nIQR(Cast Total Facebook Likes) =', IQR(likesdata$cast_total_facebook_likes), '\n')

##
## IQR(Cast Total Facebook Likes) = 12654.5
cat('sd(Cast Total Facebook Likes) =', sd(likesdata$cast_total_facebook_likes), '\n')

## sd(Cast Total Facebook Likes) = 18344.11

```

Uočavamo da je minimum svih varijabli 0 lajkova, što je samo po sebi logično jer broj lajkova ne može biti negativan. Maksimume varijabli brojimo u desetinama i stotinama tisuća.

Kod varijable ‘movie_facebook_likes’ uočavamo da je barem 25% podataka jednako nuli. Prebrojimo koliko ih je točno jednako nuli.

```
nrow(likesdata[likesdata$movie_facebook_likes == 0, ])
```

```
## [1] 2132
```

Dakle, 2083 (~43%) redaka u stupcu ‘movie_facebook_likes’ jednako je nuli. To je zaista velik udio podataka pa je za pretpostaviti da će distribucija biti desno zakrivljena, što nam dodatno potvrđuje i činjenica da je aritmetička sredina višestruko veća od medijana, vrijednosti od koje je točno polovica podataka manja. Štoviše, sredina je zamjetno veća i od trećeg kvartila, što bi značilo da su u gornjoj četvrtini podaci dovoljno veliki da bi ‘nadvladali’ ostatak. Drugim riječima, postoji nekolicina filmova sa zabilježenim iznimno velikim brojem lajkova s obzirom na većinu filmova. Moguće je da je riječ o grešci ili je neki film doživio ekstremnu popularnost među korisnicima Facebooka. Slična zapažanja mogu se primijeniti i na ostale varijable.

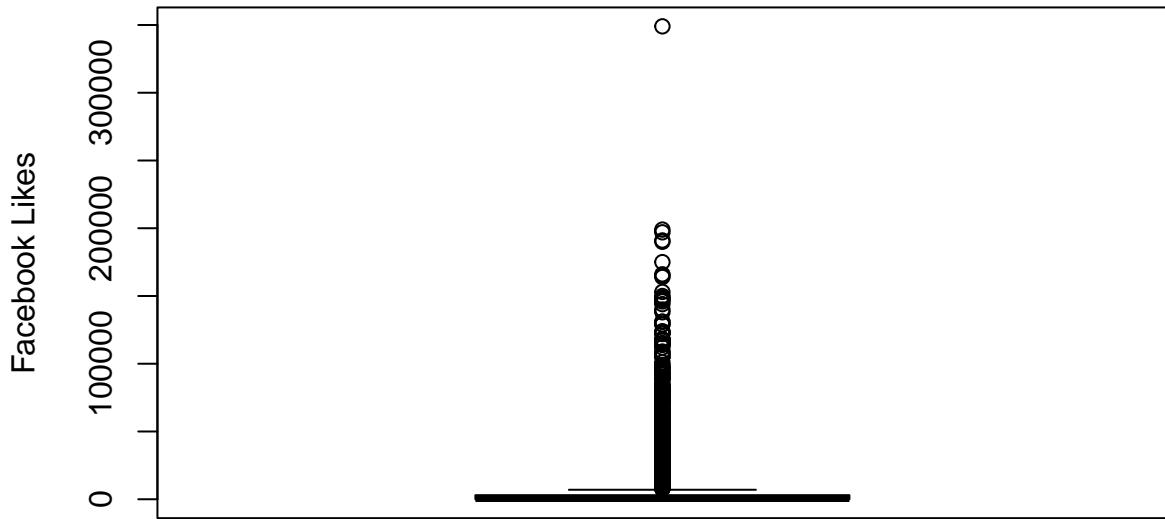
Interkvartilni rang i standardna devijacija su mjere raspršenosti podataka, no interkvartilni rang je za razliku od standardne devijacije neosjetljiv na ekstreme. Od svih promatranih varijabli, najviše odskače upravo varijabla ‘movie_facebook_likes’, koja u usporedbi s drugim varijablama ima osrednji rang, no zato ima najveći standardnu devijaciju, po čemu bismo mogli pretpostaviti da ima najviše stršećih vrijednosti.

Vizualizirajmo ponašanje promatranih varijabli korištenjem pravokutnog dijagrama.

```
# Pravokutni dijagrami
```

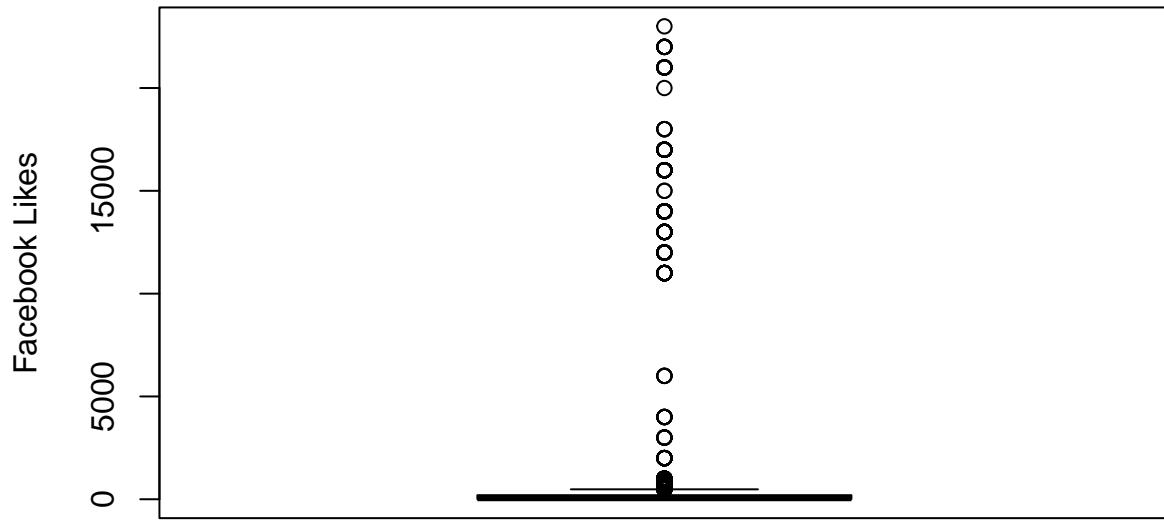
```
boxplot(likesdata$movie_facebook_likes,
        main = 'Movie Facebook Likes Boxplot',
        ylab = 'Facebook Likes')
```

Movie Facebook Likes Boxplot



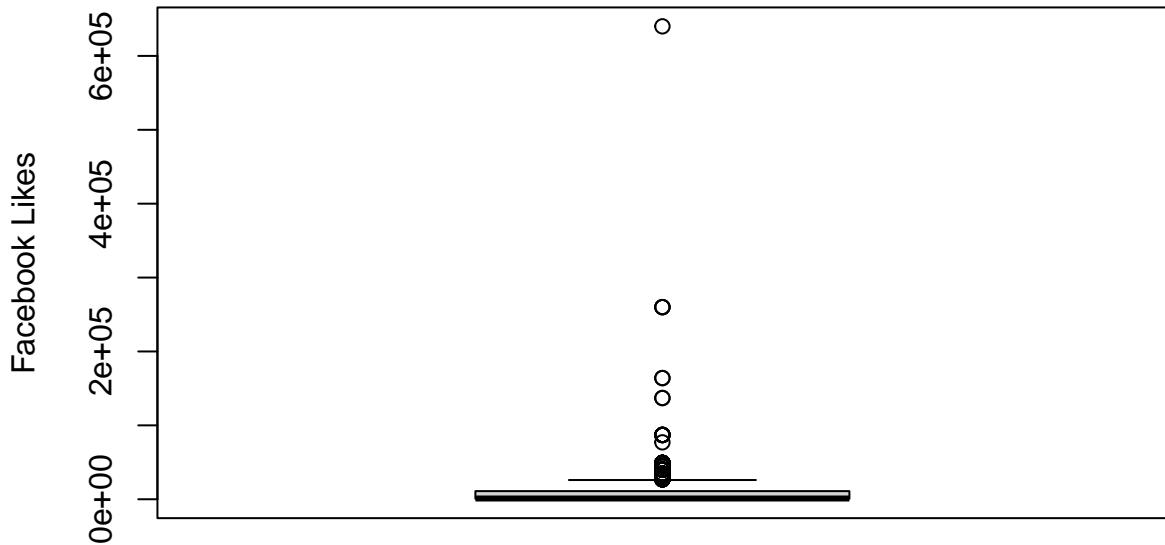
```
boxplot(likesdata$director_facebook_likes,  
       main = 'Director Facebook Likes Boxplot',  
       ylab = 'Facebook Likes')
```

Director Facebook Likes Boxplot



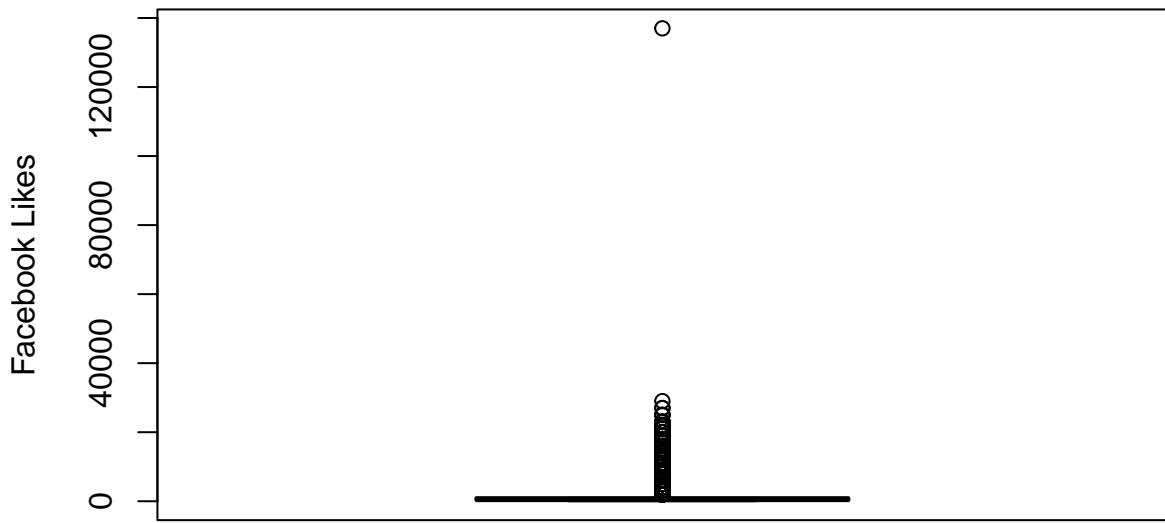
```
boxplot(likesdata$actor_1_facebook_likes,  
        main = 'Actor1 Facebook Likes Boxplot',  
        ylab = 'Facebook Likes')
```

Actor1 Facebook Likes Boxplot



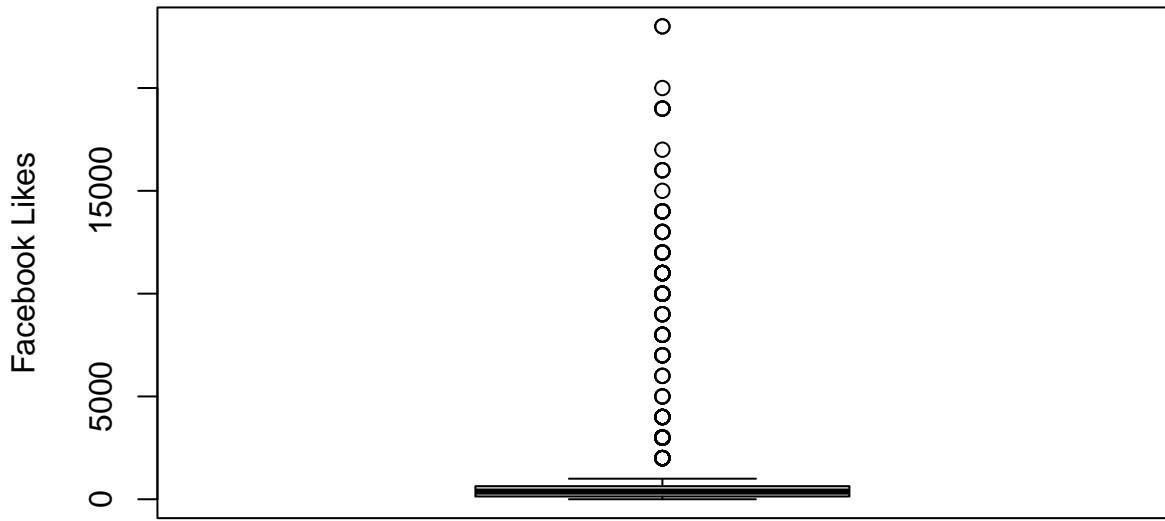
```
boxplot(likesdata$actor_2_facebook_likes,  
       main = 'Actor2 Facebook Likes Boxplot',  
       ylab = 'Facebook Likes')
```

Actor2 Facebook Likes Boxplot



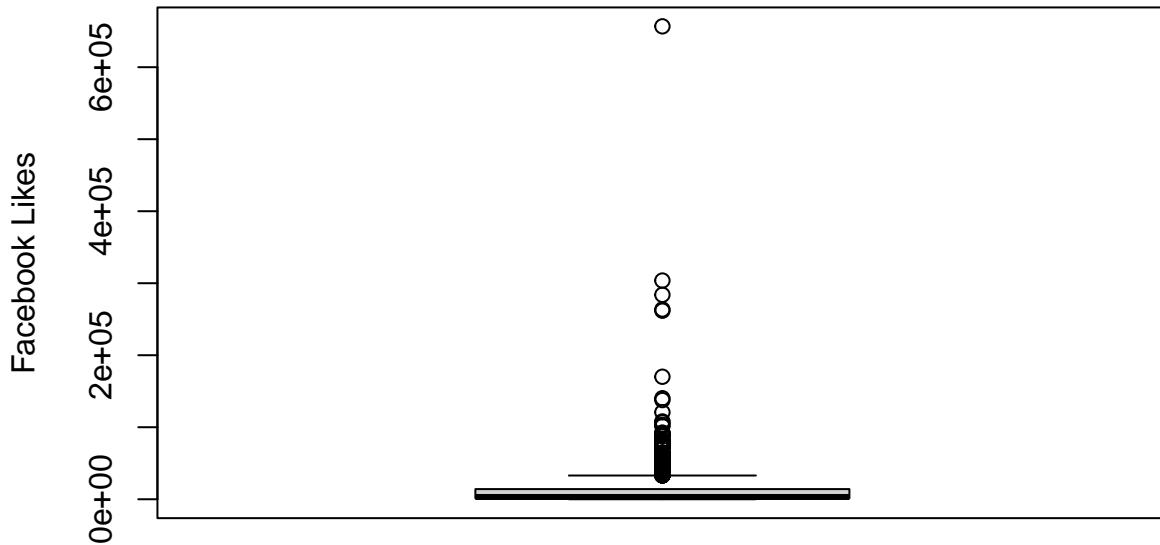
```
boxplot(likesdata$actor_3_facebook_likes,  
        main = 'Actor3 Facebook Likes Boxplot',  
        ylab = 'Facebook Likes')
```

Actor3 Facebook Likes Boxplot



```
boxplot(likesdata$cast_total_facebook_likes,  
       main = 'Cast Total Facebook Likes Boxplot',  
       ylab = 'Facebook Likes')
```

Cast Total Facebook Likes Boxplot



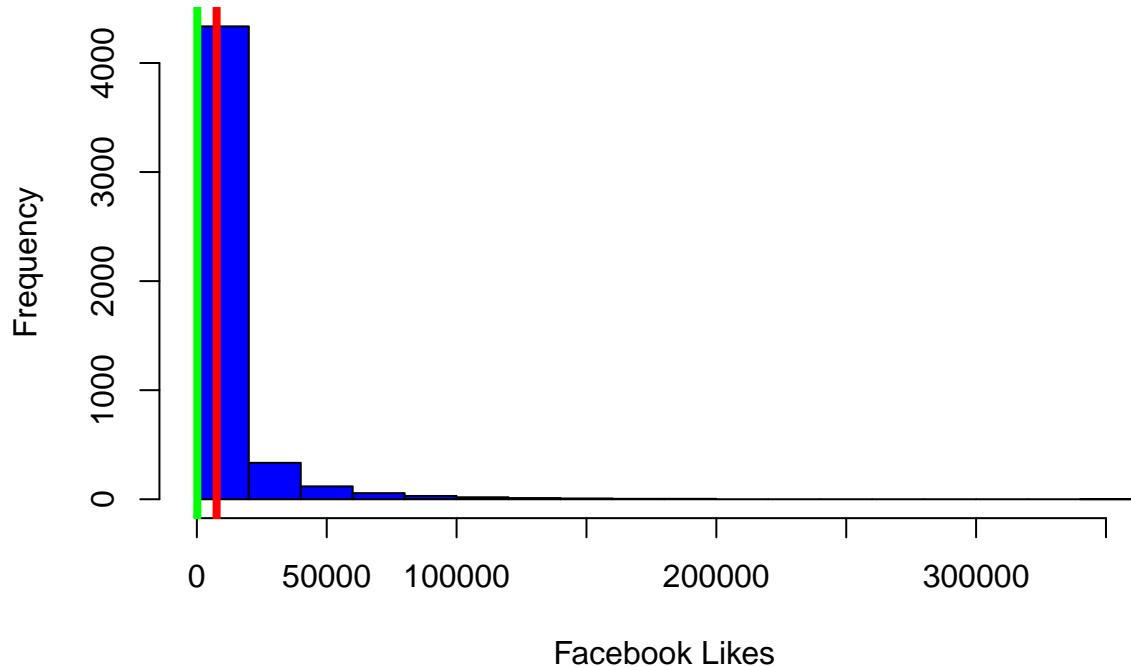
Grafovi nam ne daju toliko informacija o ponašanju podataka jer je raspon stršećih vrijednosti prevelik da bi se dobro vidi ostatak distribucije, no grafom smo uspjeli potvrditi da varijabla 'movie_facebook_likes' doista ima najviše stršećih vrijednosti.

Nadalje, pogledajmo distribucije promatranih varijabli:

```
# Histogrami

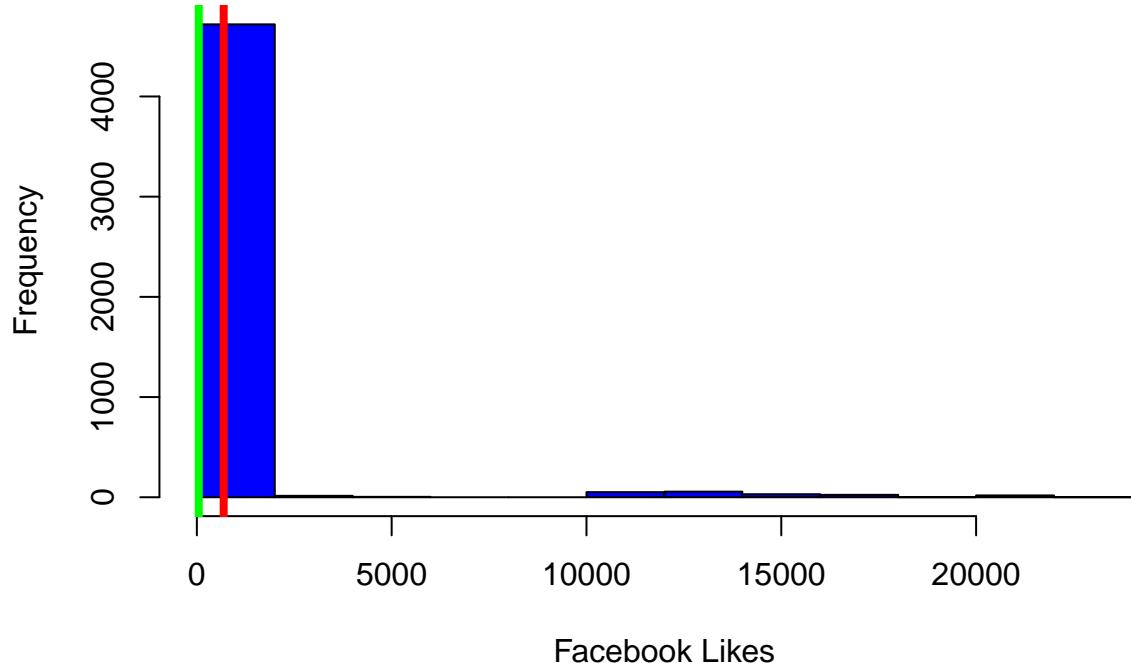
h = hist(likesdata$movie_facebook_likes,
         main="Movie Facebook Likes",
         xlab="Facebook Likes",
         ylab='Frequency',
         col="blue"
         )
abline(v = mean(likesdata$movie_facebook_likes, na.rm = TRUE), col = "red", lwd = 4)
abline(v = median(likesdata$movie_facebook_likes, na.rm = TRUE), col = "green", lwd = 4)
```

Movie Facebook Likes



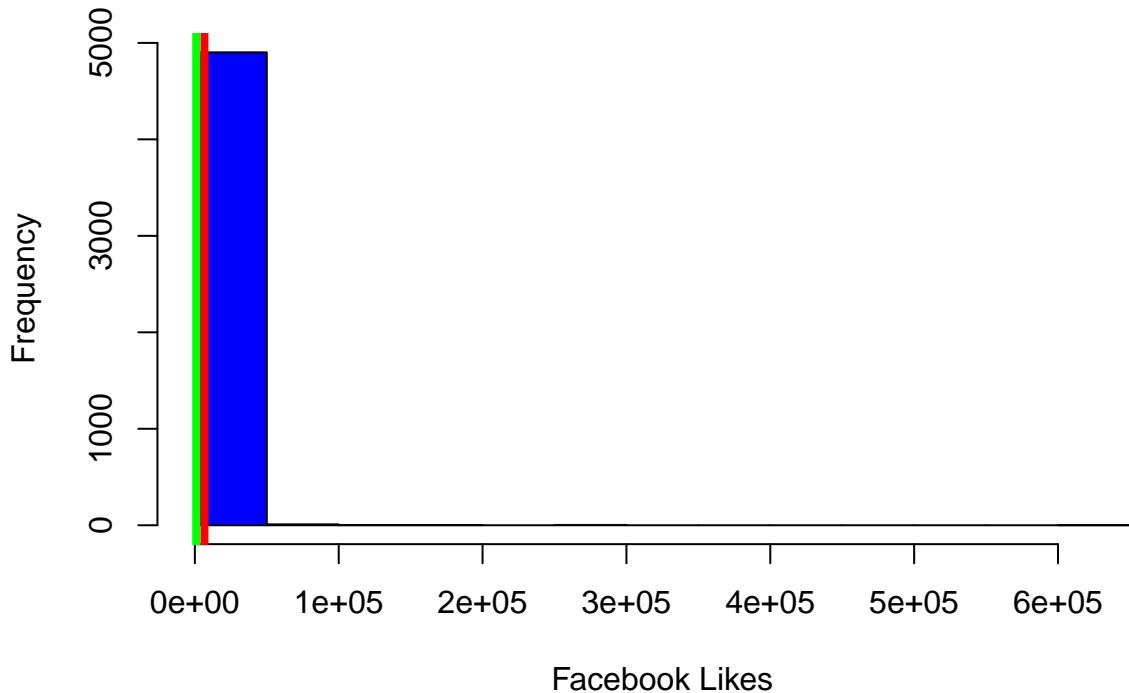
```
h = hist(likesdata$director_facebook_likes,
         main="Director Facebook Likes",
         xlab="Facebook Likes",
         ylab='Frequency',
         col="blue"
         )
abline(v = mean(likesdata$director_facebook_likes, na.rm = TRUE), col = "red", lwd = 4)
abline(v = median(likesdata$director_facebook_likes, na.rm = TRUE), col = "green", lwd = 4)
```

Director Facebook Likes



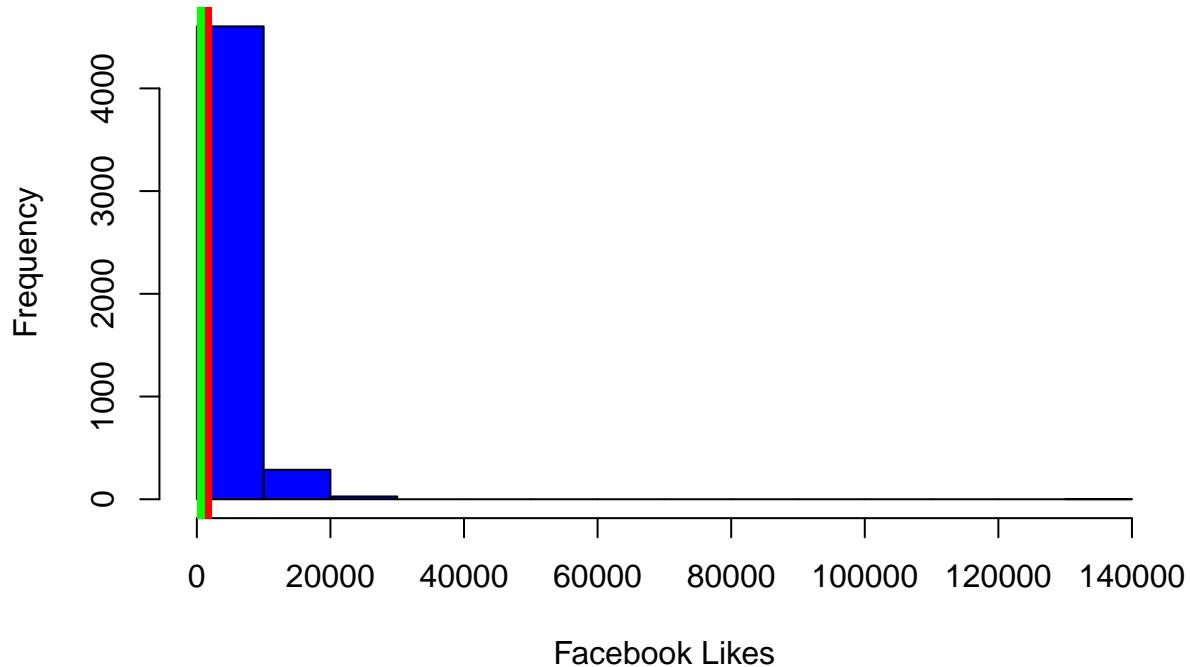
```
h = hist(likesdata$actor_1_facebook_likes,
         main="Actor1 Facebook Likes",
         xlab="Facebook Likes",
         ylab='Frequency',
         col="blue"
         )
abline(v = mean(likesdata$actor_1_facebook_likes, na.rm = TRUE), col = "red", lwd = 4)
abline(v = median(likesdata$actor_1_facebook_likes, na.rm = TRUE), col = "green", lwd = 4)
```

Actor1 Facebook Likes



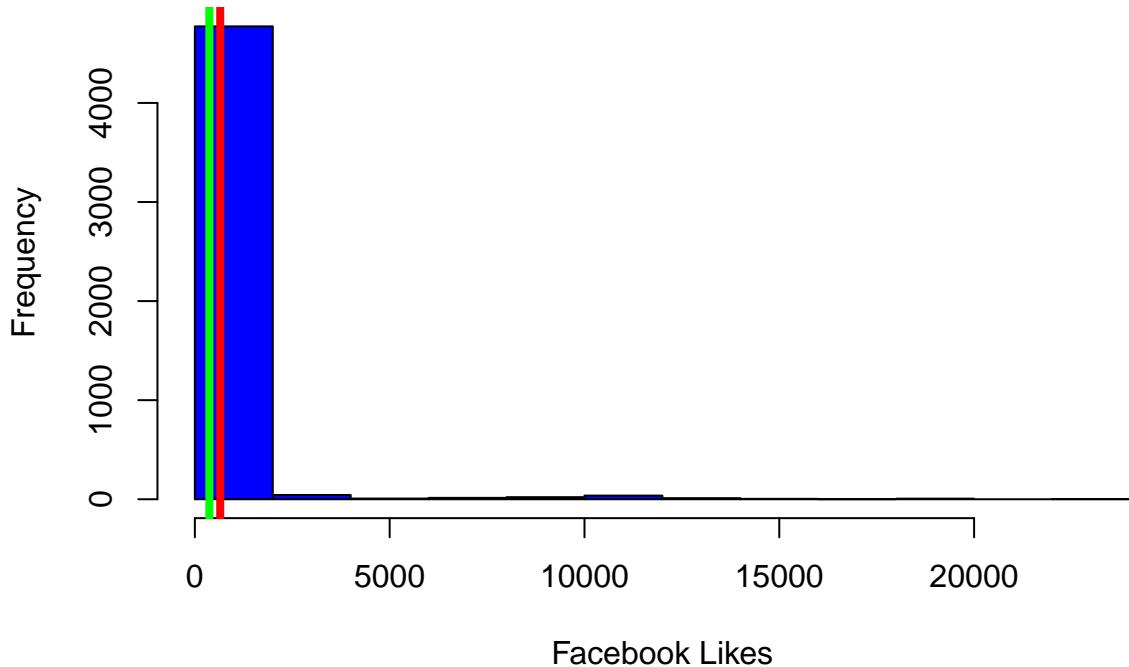
```
h = hist(likesdata$actor_2_facebook_likes,
         main="Actor2 Facebook Likes",
         xlab="Facebook Likes",
         ylab='Frequency',
         col="blue"
         )
abline(v = mean(likesdata$actor_2_facebook_likes, na.rm = TRUE), col = "red", lwd = 4)
abline(v = median(likesdata$actor_2_facebook_likes, na.rm = TRUE), col = "green", lwd = 4)
```

Actor2 Facebook Likes



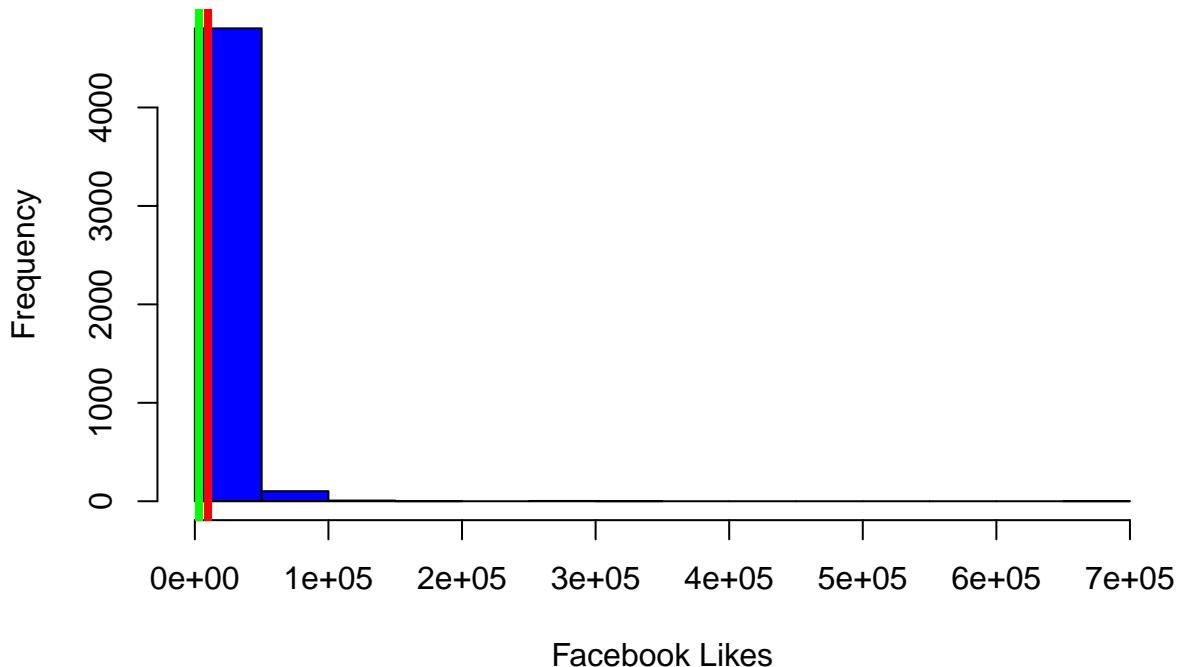
```
h = hist(likesdata$actor_3_facebook_likes,
         main="Actor3 Facebook Likes",
         xlab="Facebook Likes",
         ylab='Frequency',
         col="blue"
         )
abline(v = mean(likesdata$actor_3_facebook_likes, na.rm = TRUE), col = "red", lwd = 4)
abline(v = median(likesdata$actor_3_facebook_likes, na.rm = TRUE), col = "green", lwd = 4)
```

Actor3 Facebook Likes



```
h = hist(likesdata$cast_total_facebook_likes,
         main="Cast Total Facebook Likes",
         xlab="Facebook Likes",
         ylab='Frequency',
         col="blue"
         )
abline(v = mean(likesdata$cast_total_facebook_likes, na.rm = TRUE), col = "red", lwd = 4)
abline(v = median(likesdata$cast_total_facebook_likes, na.rm = TRUE), col = "green", lwd = 4)
```

Cast Total Facebook Likes



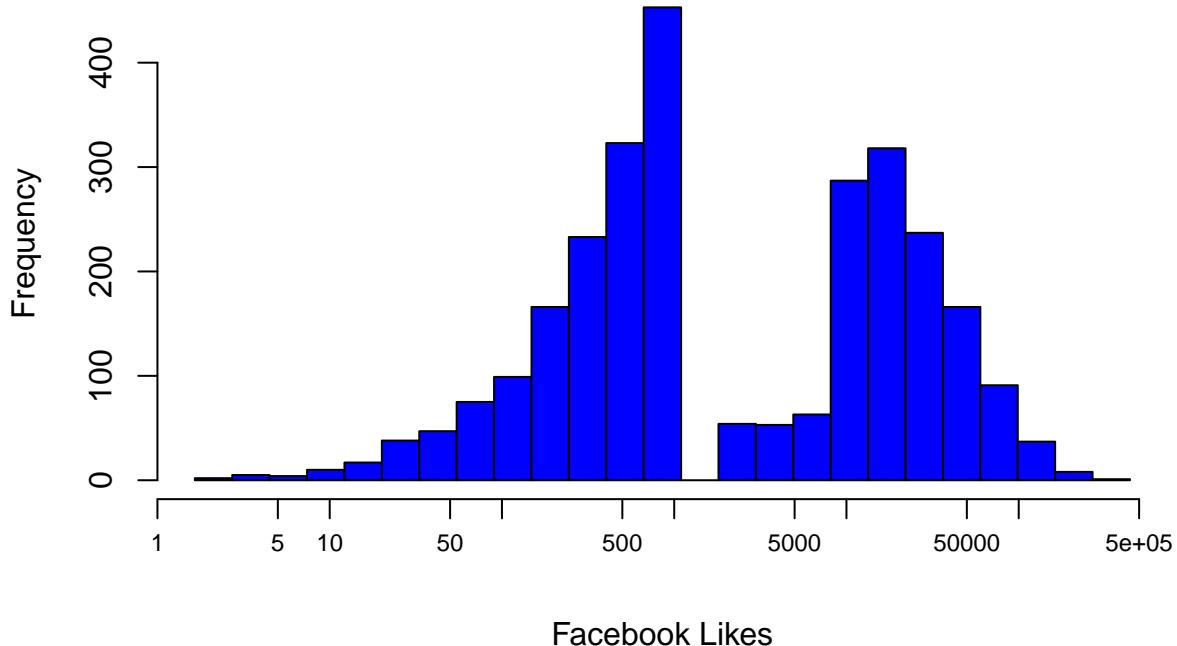
Baš kako smo i prepostavili, distribucije varijabli su desno zakriviljene.

Sami histogrami nisu nam toliko korisni u ovom obliku jer se većina vrijednosti mjeri u stotinama, stoga skalirajmo podatke prirodnim logaritmom.

```
# Histogrami (log)

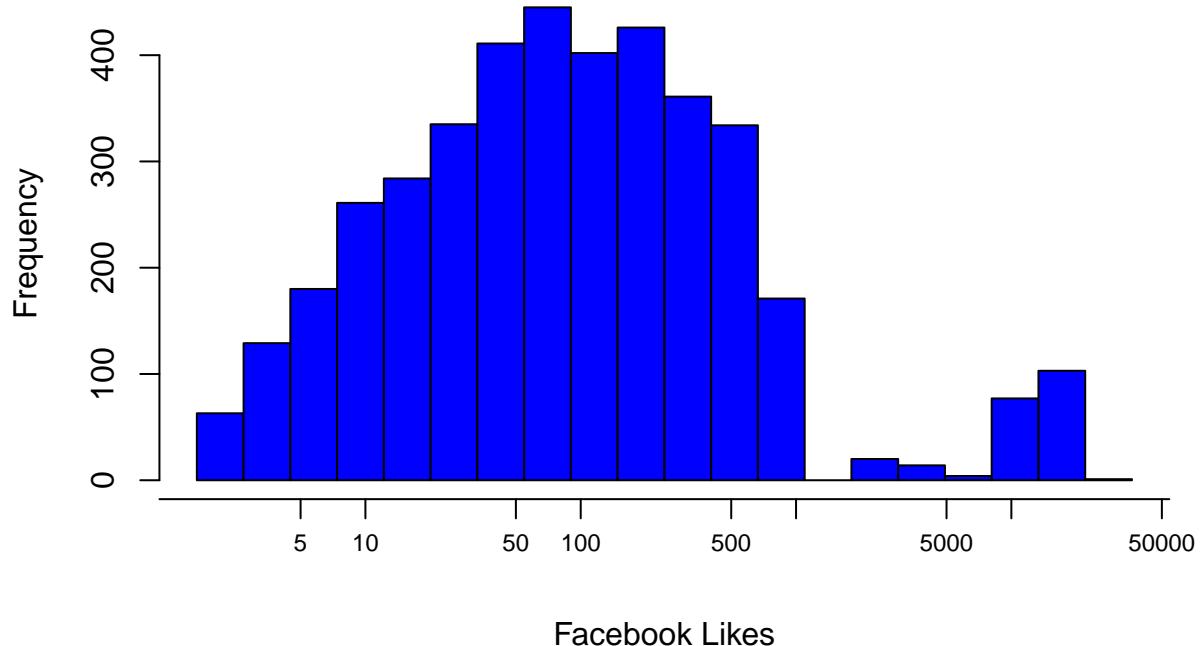
h = hist(log(likesdata$movie_facebook_likes),
          axes = FALSE,
          main="Movie Facebook Likes (log)",
          xlab="Facebook Likes",
          ylab='Frequency',
          breaks = 20, col="blue"
        )
axis(side = 1,
     at = log(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     labels = paste(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     cex.axis = 0.75,
     padj = -1,
     hadj = 0.5,
     las = 1)
axis(2)
```

Movie Facebook Likes (log)



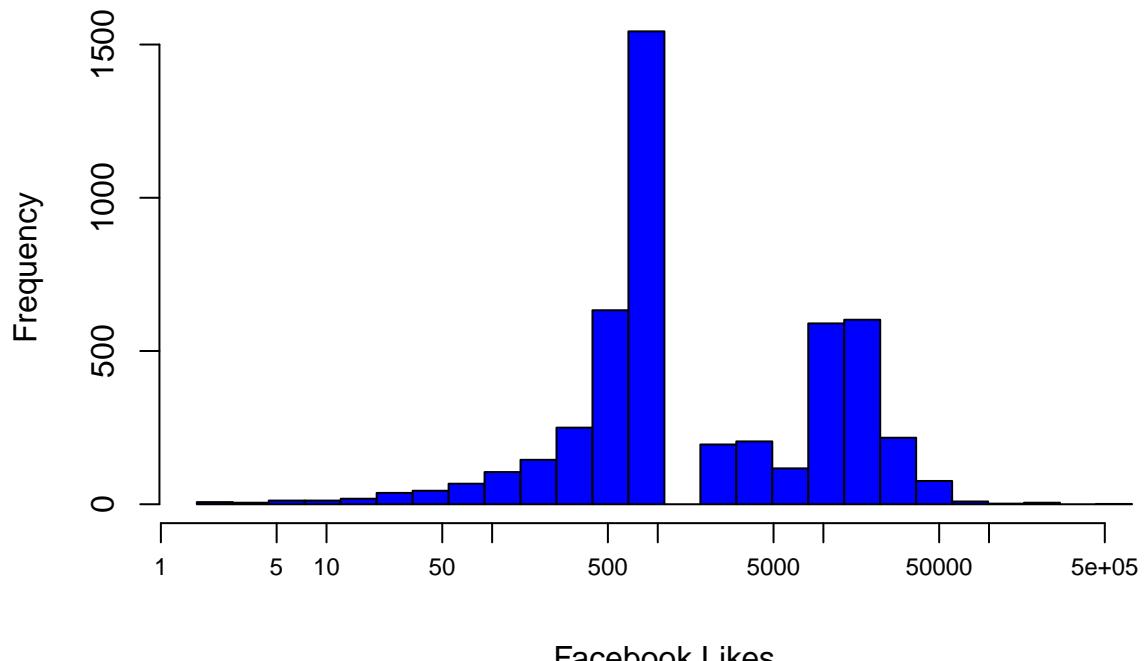
```
h = hist(log(likesdata$director_facebook_likes),
          axes = FALSE,
          main="Director Facebook Likes (log)",
          xlab="Facebook Likes",
          ylab='Frequency',
          breaks = 20, col="blue"
        )
axis(side = 1,
     at = log(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     labels = paste(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     cex.axis = 0.75,
     padj = -1,
     hadj = 0.5,
     las = 1)
axis(2)
```

Director Facebook Likes (log)



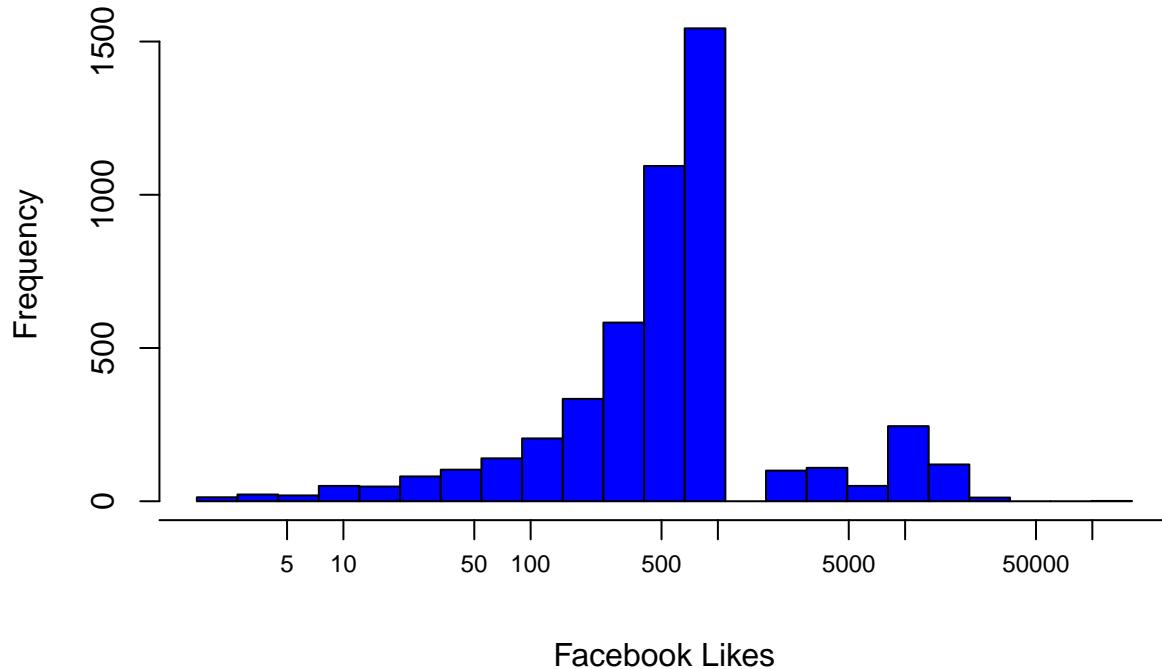
```
h = hist(log(likesdata$actor_1_facebook_likes),
          axes = FALSE,
          main="Actor1 Facebook Likes (log)",
          xlab="Facebook Likes",
          ylab='Frequency',
          breaks = 20, col="blue"
        )
axis(side = 1,
      at = log(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
      labels = paste(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
      cex.axis = 0.75,
      padj = -1,
      hadj = 0.5,
      las = 1)
axis(2)
```

Actor1 Facebook Likes (log)



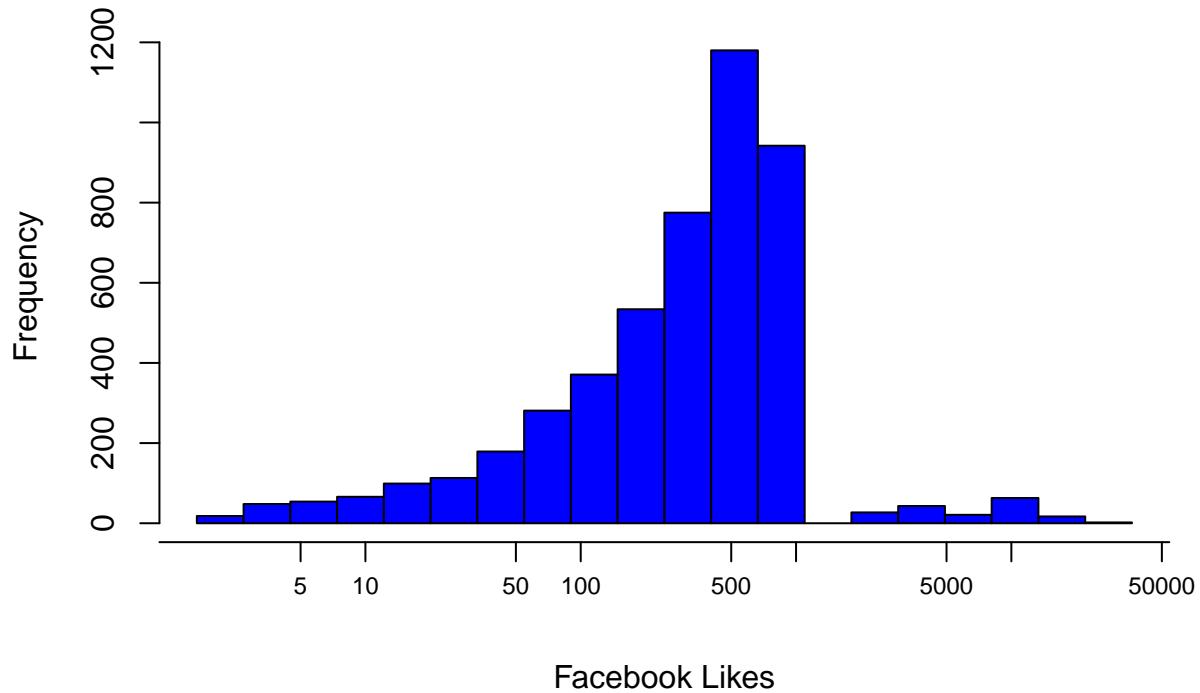
```
h = hist(log(likesdata$actor_2_facebook_likes),
          axes = FALSE,
          main="Actor2 Facebook Likes (log)",
          xlab="Facebook Likes",
          ylab='Frequency',
          breaks = 20, col="blue"
        )
axis(side = 1,
      at = log(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
      labels = paste(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
      cex.axis = 0.75,
      padj = -1,
      hadj = 0.5,
      las = 1)
axis(2)
```

Actor2 Facebook Likes (log)



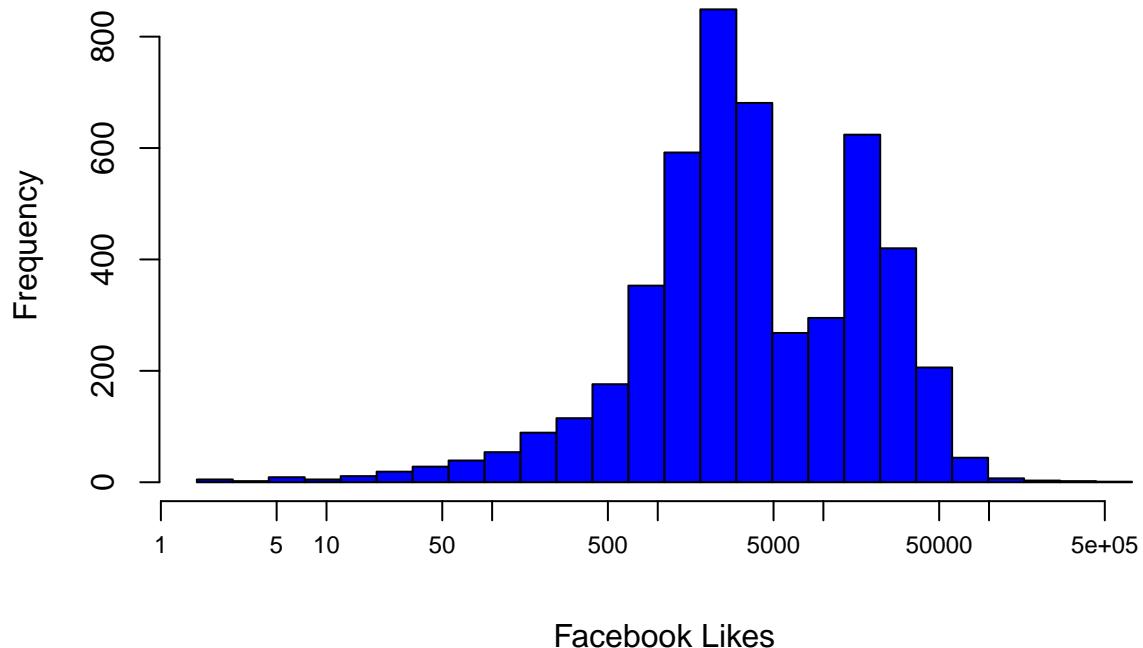
```
h = hist(log(likesdata$actor_3_facebook_likes),
          axes = FALSE,
          main="Actor3 Facebook Likes (log)",
          xlab="Facebook Likes",
          ylab='Frequency',
          breaks = 20, col="blue"
        )
axis(side = 1,
     at = log(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     labels = paste(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     cex.axis = 0.75,
     padj = -1,
     hadj = 0.5,
     las = 1)
axis(2)
```

Actor3 Facebook Likes (log)



```
h = hist(log(likesdata$cast_total_facebook_likes),
          axes = FALSE,
          main="Cast Total Facebook Likes (log)",
          xlab="Facebook Likes",
          ylab='Frequency',
          breaks = 20, col="blue"
        )
axis(side = 1,
     at = log(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     labels = paste(c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000, 50000, 100000, 500000)),
     cex.axis = 0.75,
     padj = -1,
     hadj = 0.5,
     las = 1)
axis(2)
```

Cast Total Facebook Likes (log)



Ovi nam histogrami puno jasnije prikazuju distribucije promatranih varijabli. Distribucije su višemodalne. Većina ih ima najviše pojave u intervalu između 500 i 1000 lajkova te zatim oko 10 000 lajkova. Distribucija varijable 'director_facebook_likes' bila bi najbliža normalnoj distribuciji da nema stršećih vrijednosti koje se nakupljaju između vrijednosti 10 000 i 50 000 lajkova.

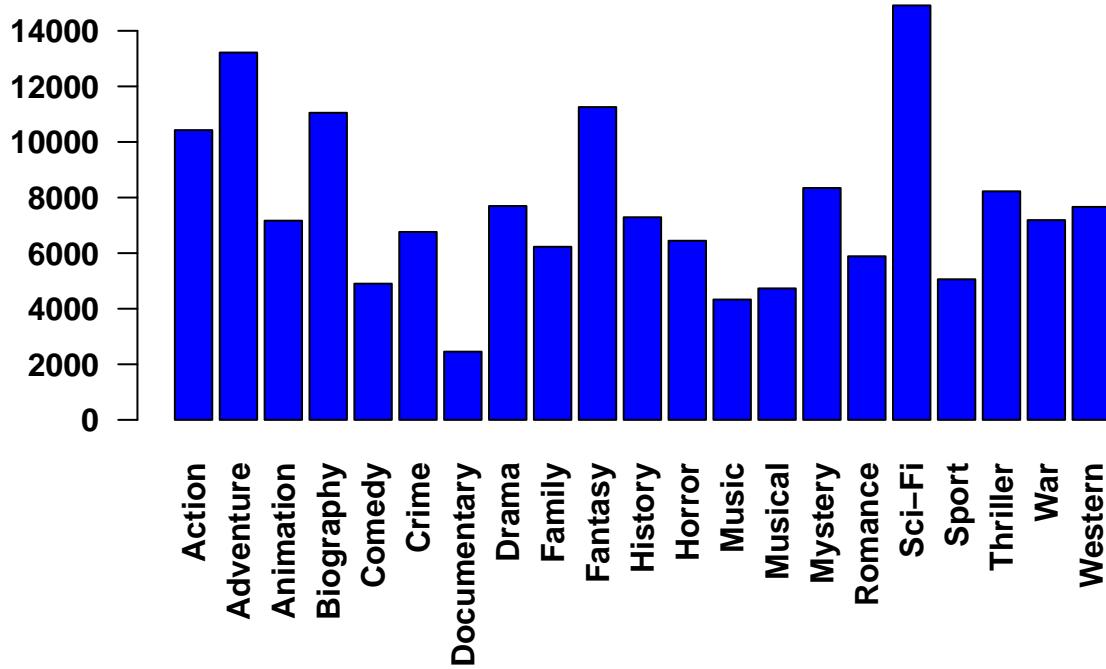
Prosječni broj Facebook lajkova po žanru

Zanima nas kakva je prosječna raspodjela lajkova po žanrovima. Ideja je grupirati filmove po žanrovima i prema lajkovima filmova dobiti prosječan broj lajkova žanra. Kako jedan film može imati više žanrova, prvo ćemo ih morati razdvojiti.

```
genres_edited %>% group_by(genres) %>% summarise(AVG_facebook_likes = round(mean(movie_facebook_likes))

par(mar=c(7.5, 4.1, 4.1, 2.1), font.axis = 2, font.lab = 2)
barplot(genre_likes$AVG_facebook_likes,
       names.arg = genre_likes$genres,
       main = 'Average No. of Facebook Likes by Genre',
       las = 2, col = 'blue')
```

Average No. of Facebook Likes by Genre



Vidimo da je žanr s najvećim prosječnim brojem lajkova SF, dok dokumentarci u prosjeku dobivaju najmanje lajkova. Pogledajmo koji su točni iznosi:

```
cat('Sci-Fi:', max(genre_likes$AVG_facebook_likes), '\n')
## Sci-Fi: 14914
cat('Documentary:', min(genre_likes$AVG_facebook_likes), '\n')
## Documentary: 2454
```

Jesu li povezani lajkovi lajkovi filma i lajkovi redatelja i glavnog glumca?

Pogledajmo postoji li povezanost između broja Facebook lajkova koje je dobio film i broja Facebook lajkova redatelja i glavnog glumca. Prvo ćemo izbaciti stršeće vrijednosti.

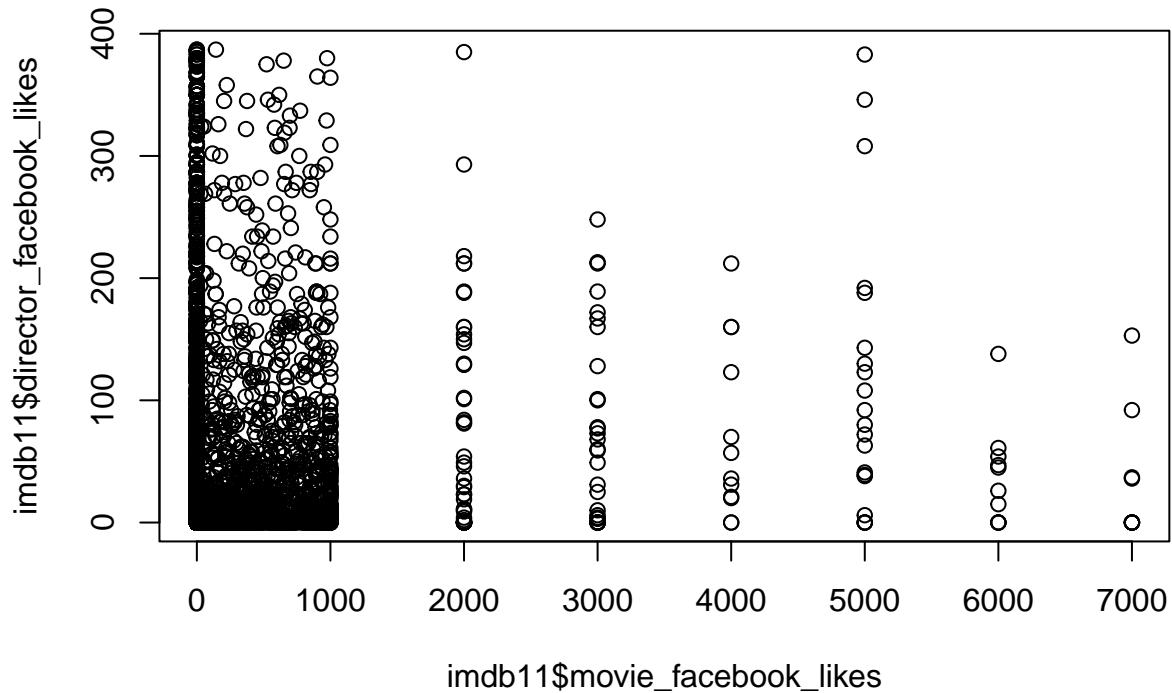
```
# Outliers
outliersgross <- boxplot(likesdata$movie_facebook_likes, plot=FALSE)$out
imdb1 <- likesdata[-which(likesdata$movie_facebook_likes %in% outliersgross),]

outliersgross <- boxplot(likesdata$gross, plot=FALSE)$out
imdb2 <- likesdata[-which(likesdata$gross %in% outliersgross),]

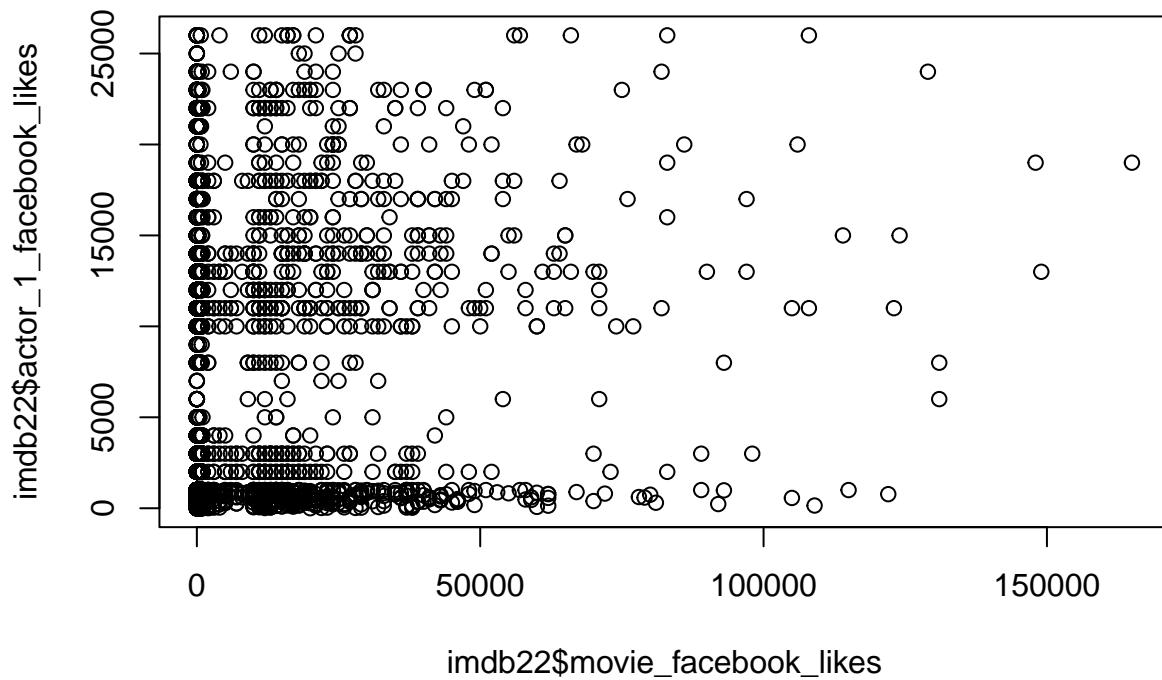
outliersbudget <- boxplot(imdb1$director_facebook_likes, plot=FALSE)$out
imdb11 <- imdb1[-which(imdb1$director_facebook_likes %in% outliersbudget),]

outliersdur <- boxplot(imdb2$actor_1_facebook_likes, plot=FALSE)$out
```

```
imdb22<- imdb2[-which(imdb2$actor_1_facebook_likes %in% outliersdur),]  
plot(imdb11$movie_facebook_likes,imdb11$director_facebook_likes)
```



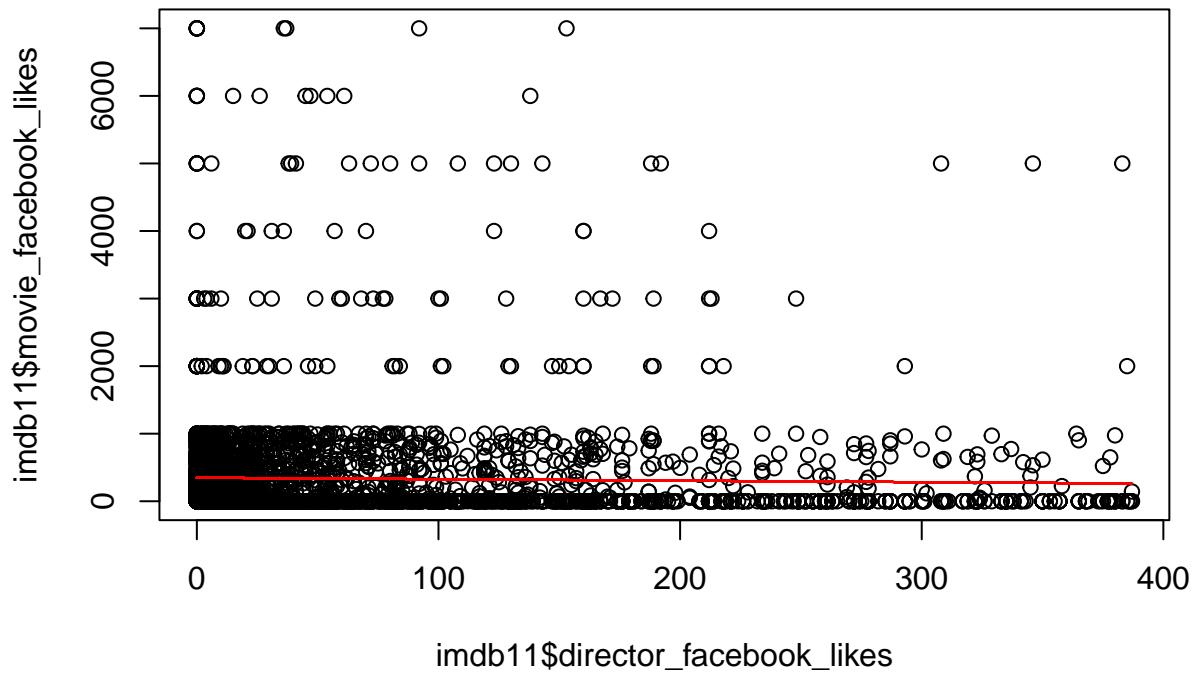
```
plot(imdb22$movie_facebook_likes,imdb22$actor_1_facebook_likes)
```



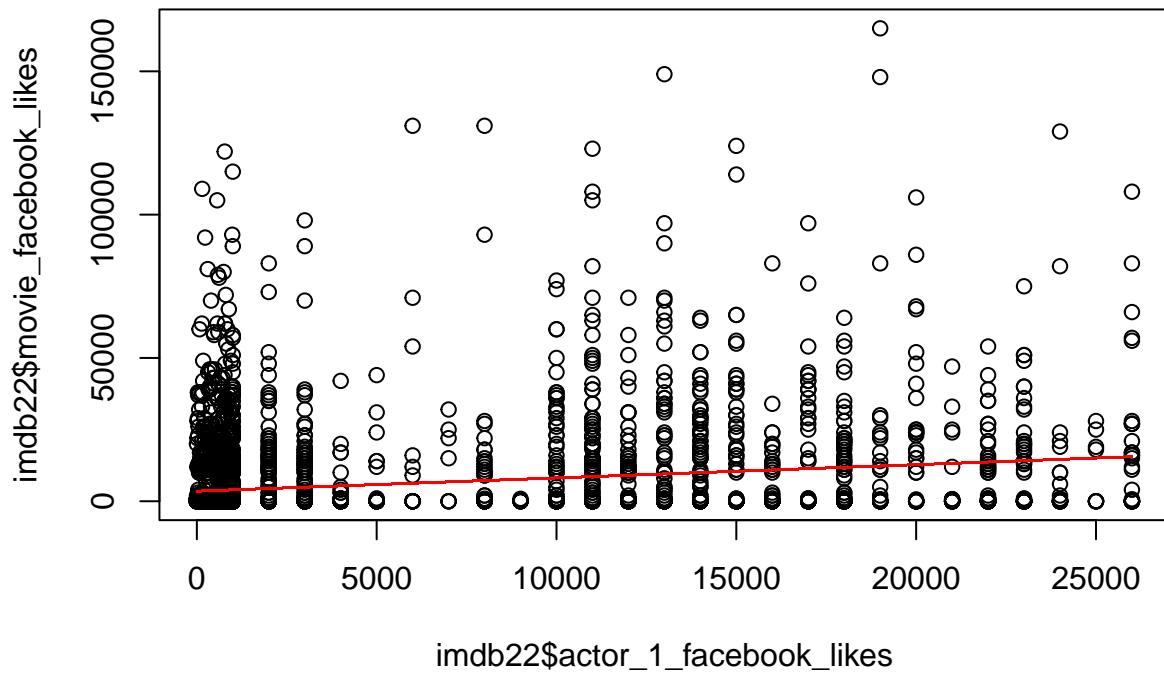
```
fit.1 = lm(movie_facebook_likes~director_facebook_likes,data=imdb11)

fit.2 = lm(movie_facebook_likes~actor_1_facebook_likes,data=imdb22)

plot(imdb11$director_facebook_likes,imdb11$movie_facebook_likes)
lines(imdb11$director_facebook_likes,fit.1$fitted.values,col='red')
```



```
plot(imdb22$actor_1_facebook_likes,imdb22$movie_facebook_likes)
lines(imdb22$actor_1_facebook_likes,fit.2$fitted.values,col='red')
```

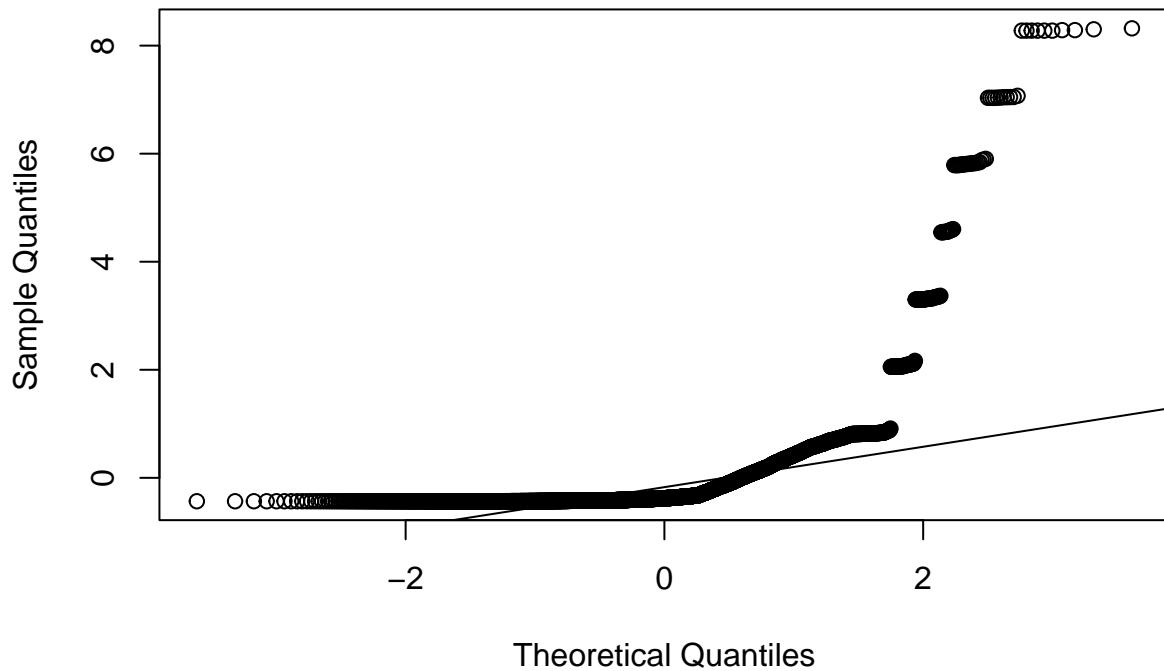


Iz scatter plotova vidimo da povezanosti gotovo i nema. Napravit ćemo i kvantil-kvantil plotove kako bismo vidjeli razdiobu reziduala.

```
selected.model = fit.1

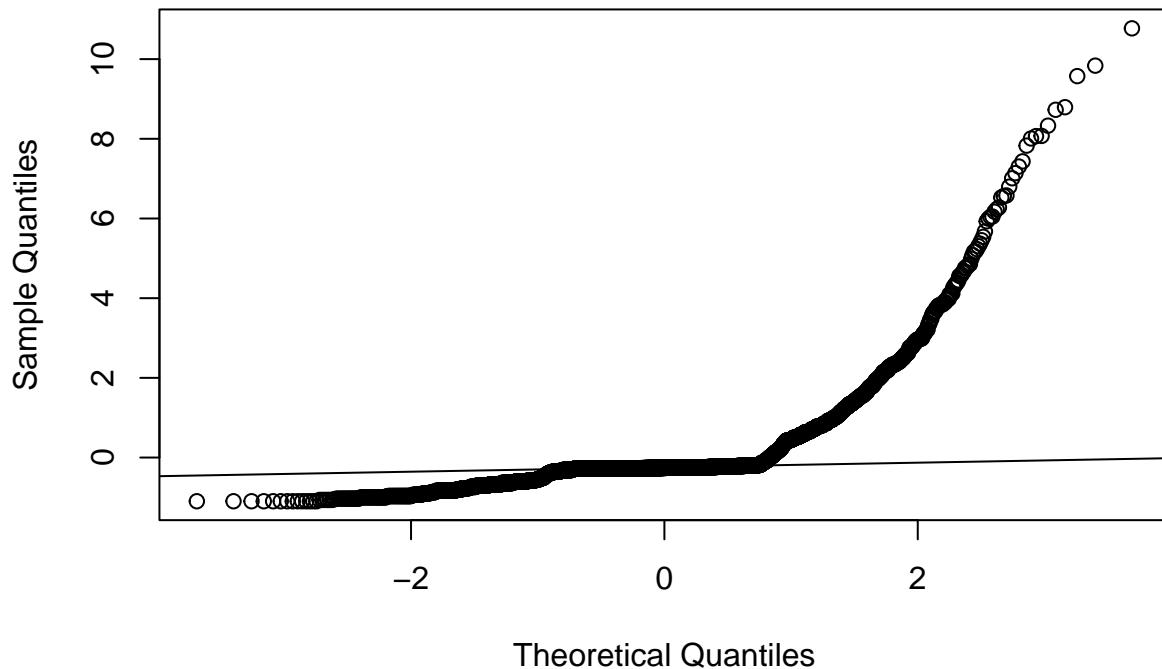
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```

Normal Q-Q Plot



```
selected.model = fit.2  
qqnorm(rstandard(selected.model))  
qqline(rstandard(selected.model))
```

Normal Q-Q Plot



Primjećujemo da razdiobe nisu niti blizu normalnoj.

Pogledajmo još mjere kvalitete prilagodbe modela podatcima.

```
summary(fit.1)
```

```
##  
## Call:  
## lm(formula = movie_facebook_likes ~ director_facebook_likes,  
##      data = imbd11)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -348.3 -341.4 -304.9   63.8 6685.4  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)             348.2518    17.4733 19.930 <2e-16 ***  
## director_facebook_likes -0.2202     0.1559 -1.413   0.158  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 803.8 on 3308 degrees of freedom  
## Multiple R-squared:  0.0006029, Adjusted R-squared:  0.0003008  
## F-statistic: 1.996 on 1 and 3308 DF,  p-value: 0.1579  
summary(fit.2)
```

```

## 
## Call:
## lm(formula = movie_facebook_likes ~ actor_1_facebook_likes, data = imdb22)
## 
## Residuals:
##      Min     1Q Median     3Q    Max 
## -15537  -3975  -3671  -2894 152700 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            3512.7947   259.6366   13.53   <2e-16 ***
## actor_1_facebook_likes  0.4625     0.0310   14.92   <2e-16 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14180 on 4471 degrees of freedom
## Multiple R-squared:  0.04743, Adjusted R-squared:  0.04722 
## F-statistic: 222.6 on 1 and 4471 DF,  p-value: < 2.2e-16

```

Obje vrijednosti R^2 su jako male što nam dodatno pokazuje koliko je mala povezanost između varijabli.

Imaju li neki žanrovi značajno različite ocjene na IMDB-u?

Kod ove hipoteze zanima nas postoji li neka značajna razlika kod ocjena različitih žanrova na IMDB-u. U nastavku ćemo testirati našu hipotezu. Promtoriti ćemo prosječne ocjene kod filmova koji pripadaju žanrovima "Biography" i "Horror".

```

biography <- genres_edited[genres_edited$genres=="Biography",]
horror <- genres_edited[genres_edited$genres=="Horror",]

#Izdvajamo samo stupce sa žanrovima, imenima filmova i ocjenama
biography <- biography[c(10,12,25)]
horror <- horror[c(10,12,25)]

##Provjera postoje li filmovi koji su ujedno i Biography i Horror
biography.movie_titles <- biography["movie_title"]
horror.movie_titles <- horror["movie_title"]
remove.list <- Reduce(intersect, list(biography.movie_titles, horror.movie_titles))

cat("Broj filmova koji pripadaju žanrovima \"Horror\" i \"Biography\": ", nrow(remove.list))

```

Broj filmova koji pripadaju žanrovima "Horror" i "Biography": 1

Vidimo da samo 1 film pripada u oba žanra te zbog prepostavke nezavisnosti koju ćemo koristiti u našem testu odbaciti ćemo ovaj podatak.

```

biography <- biography[!(biography$movie_title %in% remove.list), ]
horror <- horror[!(horror$movie_title %in% remove.list), ]

```

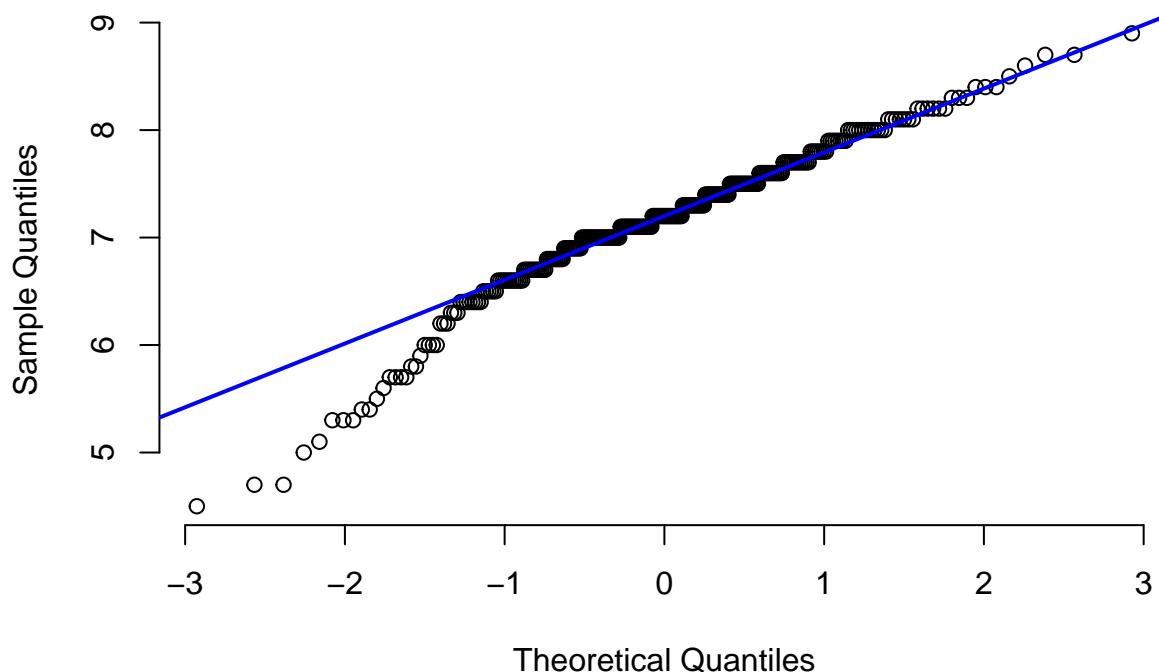
Sada kada smo dobili nezavisne podatke potrebno je provjeriti dolaze li podaci iz normalne razdiobe. Ovo ćemo napraviti uz pomoć QQ-plota.

```

# QQ plot za ocjene žanra biography
qqnorm(biography$imdb_score, pch = 1, frame = FALSE, main='Ocjene žanra biography')
qqline(biography$imdb_score, col = "blue", lwd = 2)

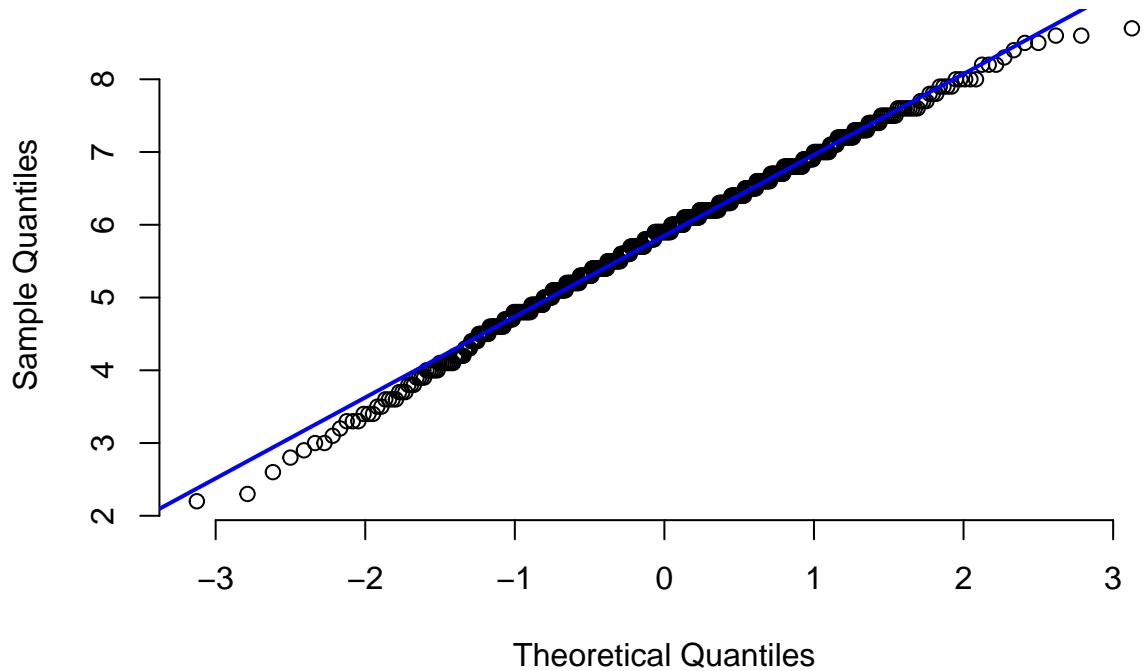
```

Ocjene zanra biography



```
# QQ plot za ocjene zanra horror
qqnorm(horror$imdb_score, pch = 1, frame = FALSE, main='Ocjene zanra horror')
qqline(horror$imdb_score, col = "blue", lwd = 2)
```

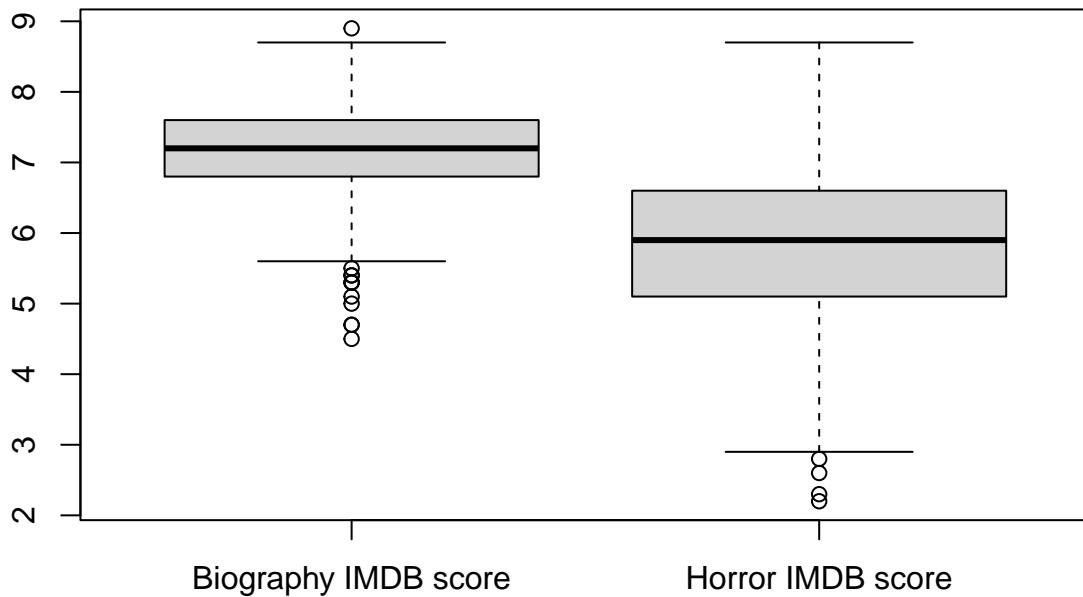
Ocjene zanra horror



Iz QQ-plota možemo vidjeti kako ocjene za žanr "Horror" dolaze iz normalne razdiobe, no isto ne možemo reći i za žanr "Biography". Pogledajmo imamo li veći broj stršećih vrijednosti koje utječu na ovaj rezultat

```
boxplot(biography$imdb_score, horror$imdb_score,  
        names=c("Biography IMDB score", "Horror IMDB score"),  
        main = "Boxplot of IMDB scores")
```

Boxplot of IMDB scores



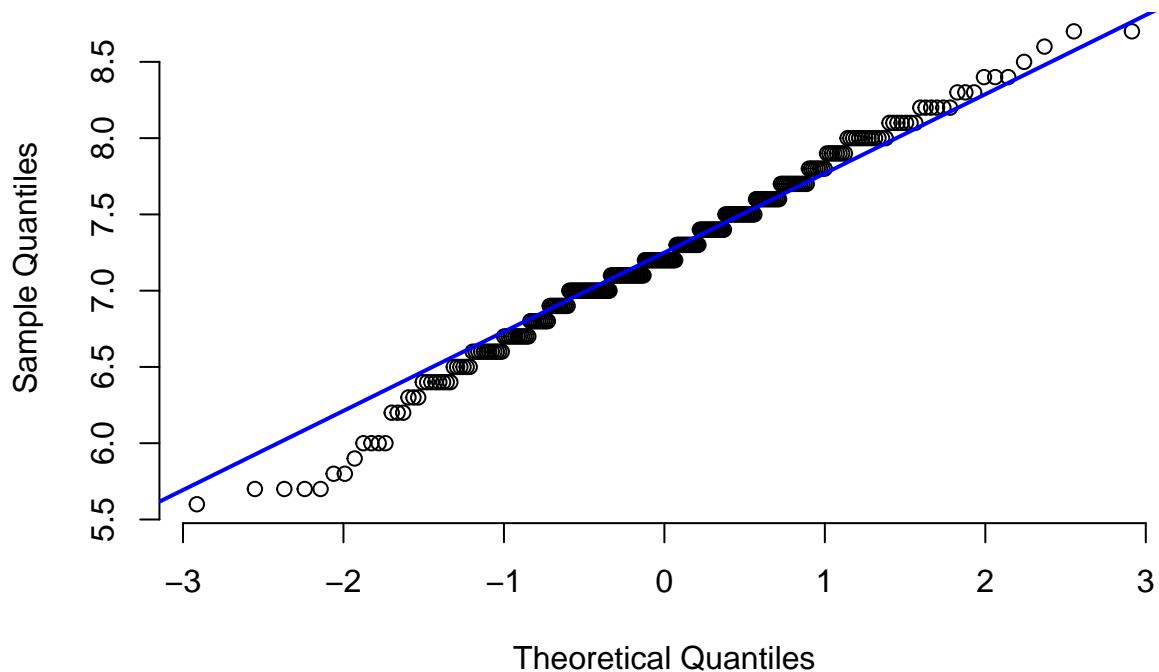
Možemo vidjeti kako imamo dosta stršećih vrijednosti kod žanra "Biography" te ih zato uklanjamo kako bi dobili bolju pretpostavku normalnosti.

```
biography.edited <- biography[!biography$imdb_score %in% boxplot.stats(biography$imdb_score)$out,]
```

Pogledajmo sada podatke na QQ-plotu.

```
qqnorm(biography.edited$imdb_score, pch = 1, frame = FALSE, main='Ocjene zanra biography')
qqline(biography.edited$imdb_score, col = "blue", lwd = 2)
```

Ocjene žanra biography



Sada kada smo postigli normalnost i nezavisnost ocjena za žanrove “Biography” i “Horror” možemo krenuti provoditi test.

1) $H_0 \dots$ Prosječna ocjena žanra *Biography* je jednaka ocjeni žanra *Horror*

$$\mu_1 = \mu_2$$

2) $H_1 \dots$ Prosječna ocjena žanra *Biography* nije jednaka ocjeni žanra *Horror*

$$\mu_1 \neq \mu_2$$

$$3) \alpha = 0.05$$

Prije nego krenemo s provođenjem testa moramo provjeriti jesu li varijance naših podataka jednake kako bi znali provest odgovarajući t-test.

1) $H_0 \dots$ Varijance distribucija se ne razlikuju

$$\sigma_1 = \sigma_2$$

2) $H_1 \dots$ Varijance distribucija se razlikuju

$$\sigma_1 \neq \sigma_2$$

$$3) \alpha = 0.05$$

```
var.test(biography$imdb_score, horror$imdb_score)
```

```

## 
## F test to compare two variances
##
## data: biography$imdb_score and horror$imdb_score
## F = 0.40351, num df = 291, denom df = 563, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.3312934 0.4946667
## sample estimates:
## ratio of variances
## 0.4035108

```

Zbog jako male p-vrijednosti zaključujemo da na razini značajnosti od 5% varijance nisu jednake. Nakon toga provodimo t-test za različite varijance.

```
t.test(biography.edited$imdb_score, horror$imdb_score, alt = "two.sided", var.equal = FALSE)
```

```

## 
## Welch Two Sample t-test
##
## data: biography.edited$imdb_score and horror$imdb_score
## t = 23.315, df = 839.47, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 1.268854 1.502129
## sample estimates:
## mean of x mean of y
## 7.229286 5.843794

```

Možemo vidjeti kako je p-vrijednost izrazito mala te iz tog razloga odbacujemo hipotezu H_0 u korist H_1 te zaključujemo da na razini značajnosti od 5% sredine ocjena za žanrove “Biography” i “Horror” nisu jednake.

Pogledajmo sada usporedbu 5 žanrova (Action, Comedy, Drama, Thriller, Romance) i njihovih prosječnih ocjena korištenjem ANOVE. Izabrali smo upravo ove žanrove zbog veličine njihvog uzorka ($n > 1000$) što naravno utječe na ispravnost naših zaključaka i smanjuje šanse za pogreške prve i druge vrste. Imamo problem jer podaci iz uzorka nisu nezavisni te je zato potrebno gledati samo filmove koji ne pripadaju nijednom od ostalih 4 žanrova.

Kao što smo vidjeli u deskriptivnoj analizi postoji velik broj stršećih vrijednosti što može znatno utjecati na pretpostavku normalnosti podataka koja nam je potrebna. Izbacimo prvo stršeće vrijednosti iz skupa podataka za pojedine žanrove

```

#Svi filmovi pojedinog zanra
action <- genres_edited[genres_edited$genres=="Action",]
comedy <- genres_edited[genres_edited$genres=="Comedy",]
drama <- genres_edited[genres_edited$genres=="Drama",]
thriller <- genres_edited[genres_edited$genres=="Thriller",]
romance <- genres_edited[genres_edited$genres=="Romance",]

#Izbacivanje strsecih vrijednosti
action <- action[!action$imdb_score %in% boxplot.stats(action$imdb_score)$out,]
comedy <- comedy[!comedy$imdb_score %in% boxplot.stats(comedy$imdb_score)$out,]
drama <- drama[!drama$imdb_score %in% boxplot.stats(drama$imdb_score)$out,]
thriller <- thriller[!thriller$imdb_score %in% boxplot.stats(thriller$imdb_score)$out,]
romance <- romance[!romance$imdb_score %in% boxplot.stats(romance$imdb_score)$out,]

```

Izbacimo nakon toga filmove koji pripadaju u više žanrova iz našeg skupa.

```

#Imena filmova pojedinog zanra (da lakse maknemo filmove koji pripadaju vise zanrova)
action.movie_title <- action[ "movie_title" ]
comedy.movie_title <- comedy[ "movie_title" ]
drama.movie_title <- drama[ "movie_title" ]
thriller.movie_title <- thriller[ "movie_title" ]
romance.movie_title <- romance[ "movie_title" ]

#Liste filmova drugih zanrova (ako je ime filma u toj listi onda pripada vise zanrova)
action.remove_list <- Reduce(union, list(comedy.movie_title, drama.movie_title, thriller.movie_title, rom))
comedy.remove_list <- Reduce(union, list(action.movie_title, drama.movie_title, thriller.movie_title, rom))
drama.remove_list <- Reduce(union, list(action.movie_title, comedy.movie_title, thriller.movie_title, rom))
thriller.remove_list <- Reduce(union, list(action.movie_title, drama.movie_title, comedy.movie_title, rom))
romance.remove_list <- Reduce(union, list(action.movie_title, drama.movie_title, thriller.movie_title, rom))

#Micanje svih filmova iz drugih zanrova iz skupa podataka

action.nez <- action[!(action$movie_title %in% action.remove_list$movie_title), ]
comedy.nez <- comedy[!(comedy$movie_title %in% comedy.remove_list$movie_title), ]
drama.nez <- drama[!(drama$movie_title %in% drama.remove_list$movie_title), ]
thriller.nez <- thriller[!(thriller$movie_title %in% thriller.remove_list$movie_title), ]
romance.nez <- romance[!(romance$movie_title %in% romance.remove_list$movie_title), ]

```

Nakon transformacija dobili smo filmove koji pripadaju samo jednom žanru, a ne nijednom od ostalih četiri. Pogledajmo sada koliko je filmova ostalo koji ne pripadaju ni jednom drugom žanru.

```
nrow(action.nez)
```

```
## [1] 205
```

```
nrow(comedy.nez)
```

```
## [1] 633
```

```
nrow(drama.nez)
```

```
## [1] 760
```

```
nrow(thriller.nez)
```

```
## [1] 270
```

```
nrow(romance.nez)
```

```
## [1] 23
```

Vidimo kako je broj filmova žanra “Romance” dosta manji od ostalih, međutim takvo ponašanje je očekivano s obzirom da većina filmova ovog žanra spada ili u žanr “Comedy” ili žanr “Drama”. Sada kada smo postigli nezavisnost podataka moramo provjeriti normalnost. To ćemo napraviti uz pomoć Lillieforsove inačice KS testa.

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(action.nez$imdb_score)
```

```
##
```

```
## Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
```

```
## data: action.nez$imdb_score
```

```

## D = 0.047553, p-value = 0.3109
lillie.test(comedy.nez$imdb_score)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: comedy.nez$imdb_score
## D = 0.041281, p-value = 0.01223
lillie.test(drama.nez$imdb_score)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: drama.nez$imdb_score
## D = 0.073463, p-value = 1.932e-10
lillie.test(thriller.nez$imdb_score)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: thriller.nez$imdb_score
## D = 0.054947, p-value = 0.04724
lillie.test(romance.nez$imdb_score)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: romance.nez$imdb_score
## D = 0.15081, p-value = 0.1912

```

Možemo vidjeti kako žanr "Drama" ima jako malu p-vrijednost pa za njega ne možemo pretpostaviti normalnost. Za ostale žanrove na razini značajnosti od 1% vrijedi pretpostavka normalnosti. U daljnoj provedbi testa izbacujemo žanr "Drama" iz analize zbog nezadovoljenih pretpostavki testa. Sljedeća pretpostavka koja nam je potrebna za analizu je homogenost varijanci svih uzoraka. To ćemo provjeriti uz pomoć Bartlettovog testa koji će zbog prijašnje pretpostavke normalnosti još bolje prikazati rezultate.

```

# Pojedinacne varijance svakog zanra
var(action.nez$imdb_score)

## [1] 1.291764
var(comedy.nez$imdb_score)

## [1] 1.033606
var(thriller.nez$imdb_score)

## [1] 1.20487
var(romance.nez$imdb_score)

## [1] 1.646798
# Dodajemo listu ocjena u varijable
a <- action.nez$imdb_score
b <- comedy.nez$imdb_score
c <- thriller.nez$imdb_score

```

```

d <- romance.nez$imdb_score

bartlett.test(list(a,b,c,d))

##
##  Bartlett test of homogeneity of variances
##
## data: list(a, b, c, d)
## Bartlett's K-squared = 6.646, df = 3, p-value = 0.08408

```

Vidimo kako Bartlettov test daje p-vrijednost od nešto više od 8% što nam, na razini značajnosti od 1%, dozvoljava da prepostavimo da su varijance ova 4 žanra jednake. Nakon što smo zadovoljili prepostavke normalnosti, nezavisnosti i homogenosti varijanci uzorka možemo provesti ANOVU sa sljedećim hipotezama:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \neg H_0.$$

```

# Sve filmove dodajemo u jednu listu
allNames <- c(action.nez$movie_title, comedy.nez$movie_title)
allNames <- c(allNames, romance.nez$movie_title)
allNames <- c(allNames, thriller.nez$movie_title)
allGenres <- c("Action", "Comedy", "Thriller", "Romance")
#Editamo pocetni skup
edited <- genres_edited[genres_edited$movie_title %in% allNames,]
edited <- edited[edited$genres %in% allGenres,]
#Provjedba ANOVE
test <- aov(imdb_score ~ genres, data=edited)
summary(test)

##                                Df Sum Sq Mean Sq F value    Pr(>F)
## genres             3   23.3   7.752   6.746 0.000164 ***
## Residuals      1132 1300.7   1.149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Vidimo da je p-vrijednost jako mala te na razini značajnosti od 1% odbacujemo nultu hipotezu i zaključujemo da srednje vrijednosti ocjena za barem dva žanra od žanrova “Action”, “Comedy”, “Thriller”, “Romance” nisu jednake. Tako se naša inicijalna pretpostavka o uniformnoj distribuciji srednjih vrijednosti ocjena pokazala netočnom te zaključujemo da postoji značajna razlika u prosječnim ocjenama pojedinih žanrova.

Dobivaju li američki filmovi veća financiranja od filmova drugih zemalja

Prebrojimo koliko Američkih filmova za svoj budget ima NA.

```

americkiFilmovi <- imdb[imdb$country == "USA",]$budget/1000000

cat("Broj Americkih filmova : ", length(americkiFilmovi), "\n")

## Broj Americkih filmova : 3807
cat("Broj NA vrijednosti: ", sum(is.na(americkiFilmovi)), "\n")

## Broj NA vrijednosti: 298

```

Vidimo da je manje od 20% te cemo te redke sada maknuti.

```

americkiFilmovi <- na.omit(americkiFilmovi)
cat("Broj Americkih filmova bez NA redaka: ", length(americkiFilmovi), "\n")

## Broj Americkih filmova bez NA redaka: 3509
neAmerickiFilmovi <- imdb[imdb$country != "USA",]$budget/1000000

cat("Broj ne Americkih filmova : ", length(neAmerickiFilmovi), "\n")

## Broj ne Americkih filmova : 1236
cat("Broj NA vrijednosti: ", sum(is.na(neAmerickiFilmovi)), "\n")

```

Broj NA vrijednosti: 194

Ovdje je također tih redaka manje od 20% te ćemo te redke također sada maknuti.

```

neAmerickiFilmovi <- na.omit(neAmerickiFilmovi)
cat("Broj Ne Americkih filmova bez NA redaka: ", length(neAmerickiFilmovi), "\n")

```

Broj Ne Americkih filmova bez NA redaka: 1042

Maknimo sada outliere iz vektora.

```

americkiFilmoviOutliers <- boxplot(americkiFilmovi, plot=FALSE)$out
americkiFilmovi <- americkiFilmovi[-which(americkiFilmovi %in% americkiFilmoviOutliers)]


neAmerickiFilmoviOutliers <- boxplot(neAmerickiFilmovi, plot=FALSE)$out
neAmerickiFilmovi <- neAmerickiFilmovi[-which(neAmerickiFilmovi %in% neAmerickiFilmoviOutliers)]

```

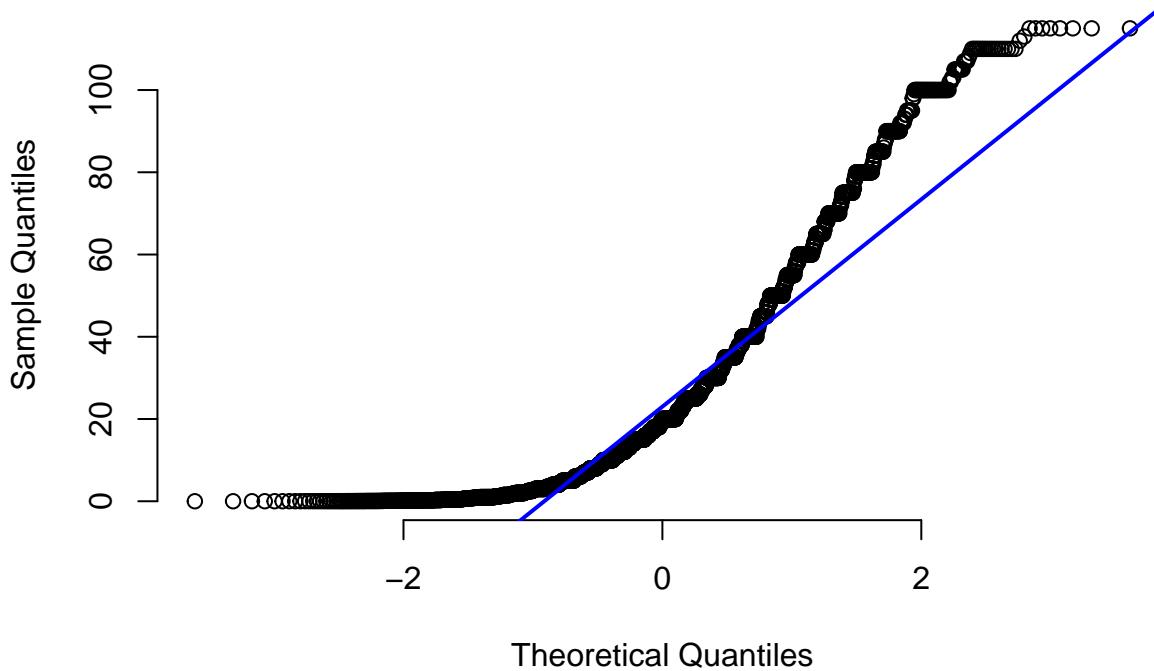
Sada ćemo vizualizirati te podatke da vidimo dolaze li iz normalne razdiobe.

```

qqnorm(americkiFilmovi, pch = 1, frame = FALSE, main='Financiranje Americkih filmova ')
qqline(americkiFilmovi, col = "blue", lwd = 2)

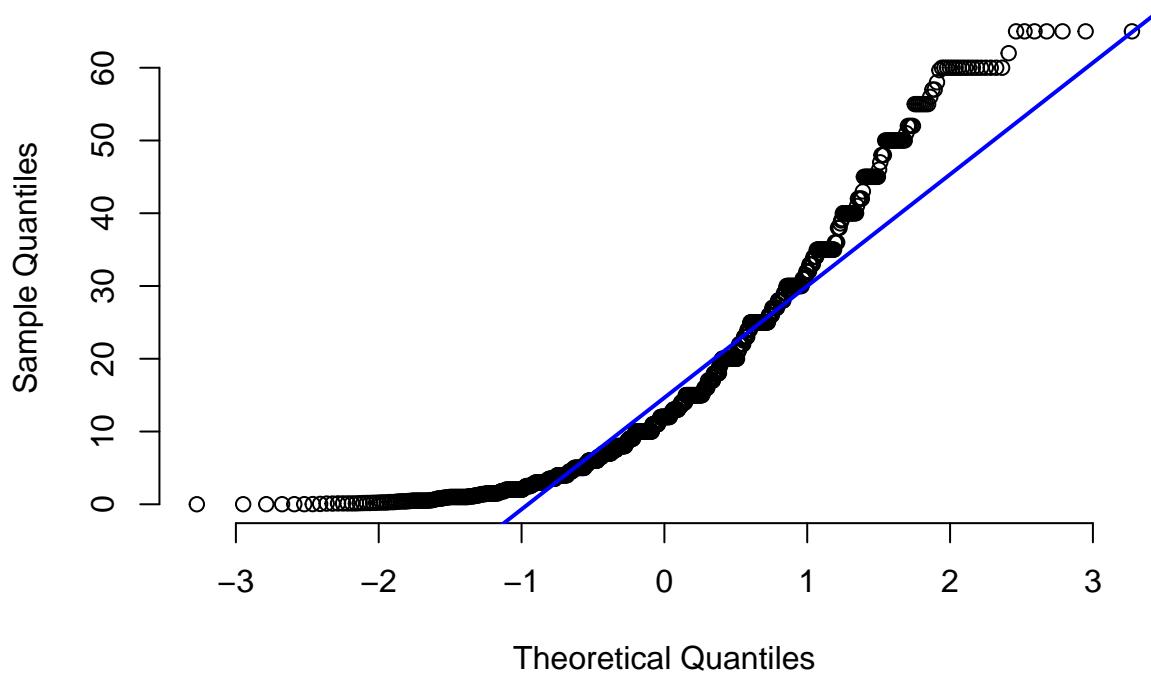
```

Financiranje Americkih filmova



```
qqnorm(neAmerickiFilmovi, pch = 1, frame = FALSE,main='Financiranje Ne Americkih filmova ')
qqline(neAmerickiFilmovi, col = "blue", lwd = 2)
```

Financiranje Ne Americkih filmova

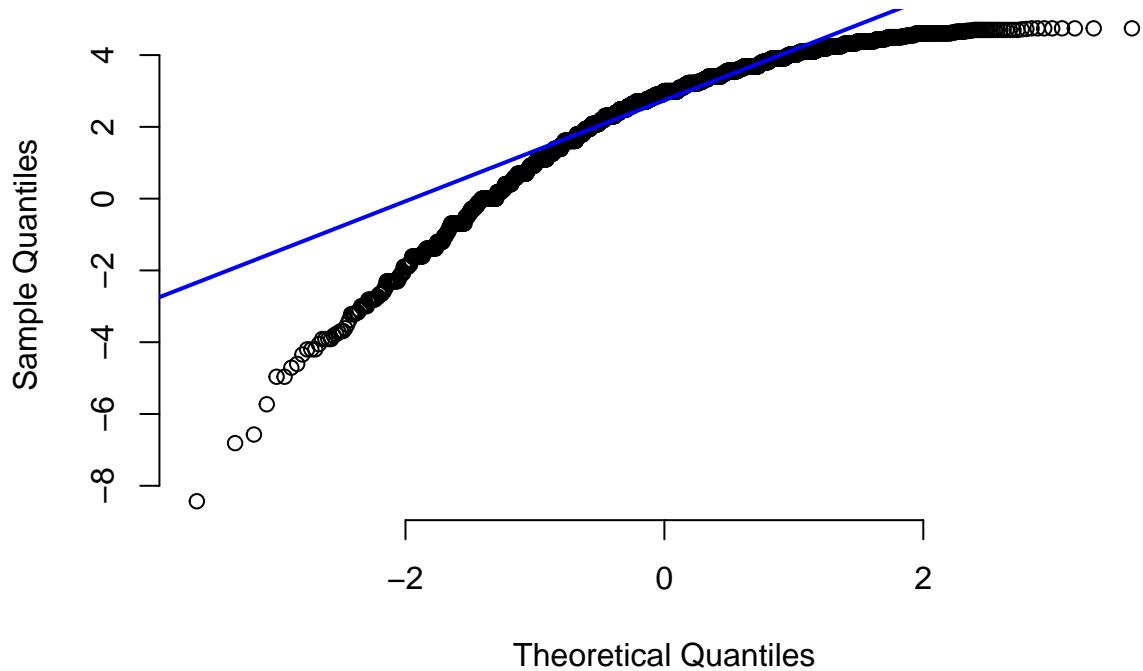


Vidimo da ove grafovi ne nalikuju na Normalnu razdiobu, no mozda će nalikovati ako na podatke primjenimo logaritamsku funkciju.

```
logAmericki <- log(americkiFilmovi)

qqnorm(logAmericki, pch = 1, frame = FALSE, main='Financiranje Americkih filmova sa log funkcijom')
qqline(logAmericki, col = "blue", lwd = 2)
```

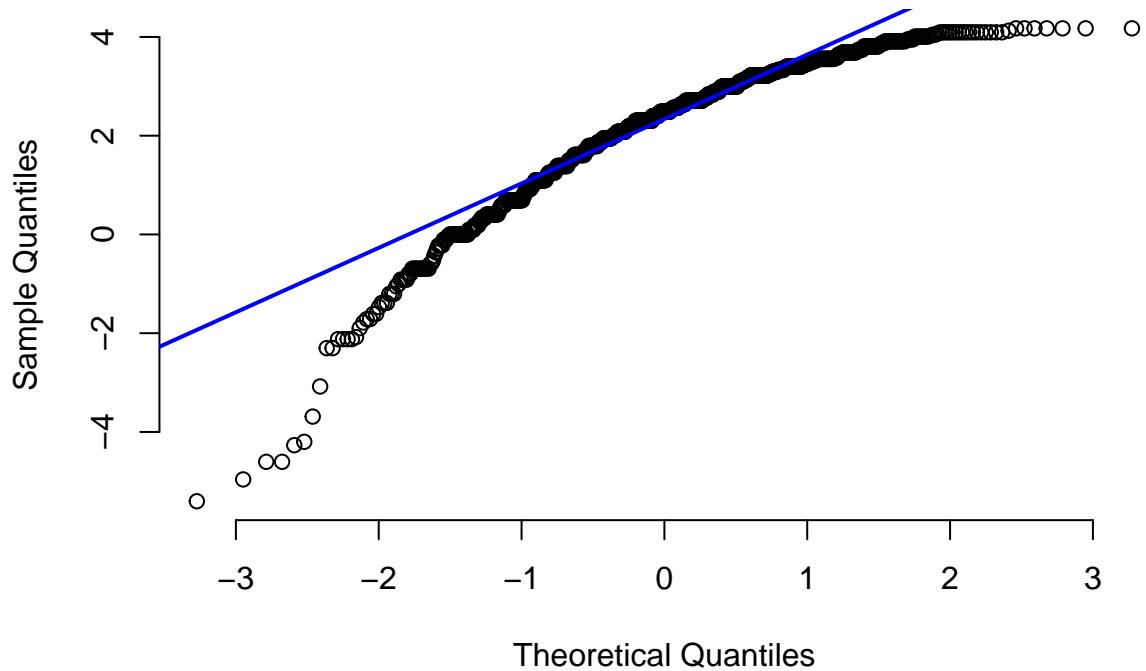
Financiranje Americkih filmova sa log funkcijom



```
logNeAmericki <- log(neAmerickiFilmovi)
```

```
qqnorm(logNeAmericki, pch = 1, frame = FALSE, main='Financiranje Ne Americkih filmova sa log funkcijom')  
qqline(logNeAmericki, col = "blue", lwd = 2)
```

Financiranje Ne Americkih filmova sa log funkcijom

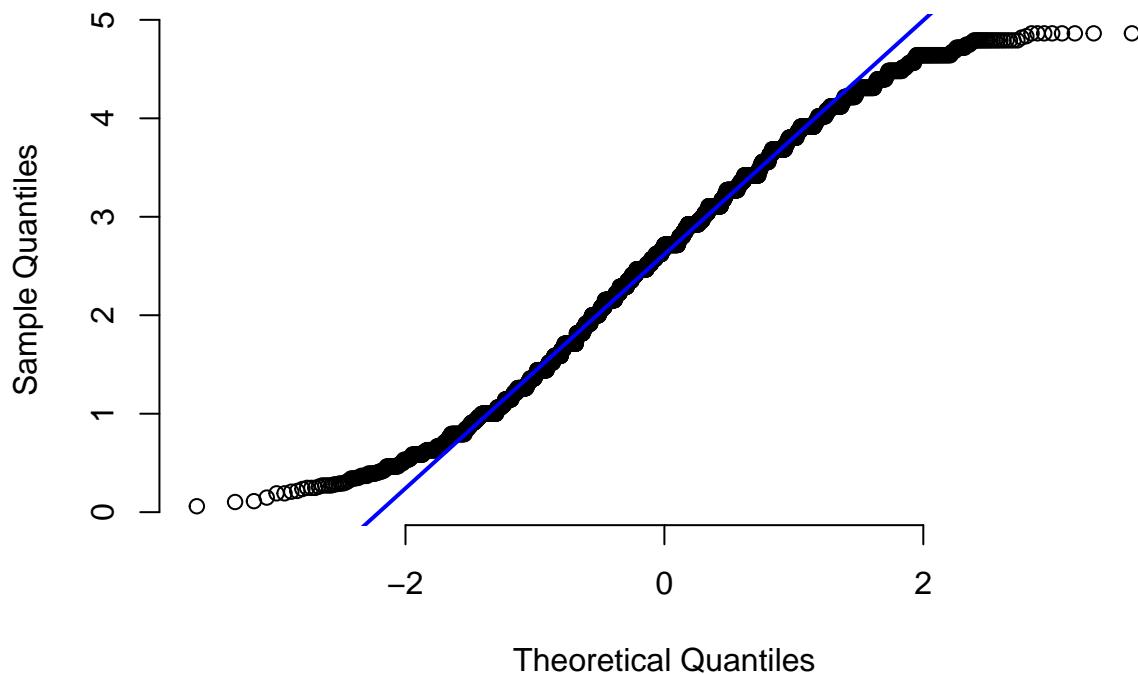


Vidimo da ni nakon toga ne nalikuju na normalnu, ali ako bolje proučimo, graf nalikuje na x^3 , te ako primjenimo treci korijen, možda će onda podatci biti normalni.

```
treciKorijenAmericki <- americkiFilmovi ^ (1/3)
```

```
qqnorm(treciKorijenAmericki, pch = 1, frame = FALSE, main='Financiranje Americkih filmova sa treci korijenom')
qqline(treciKorijenAmericki, col = "blue", lwd = 2)
```

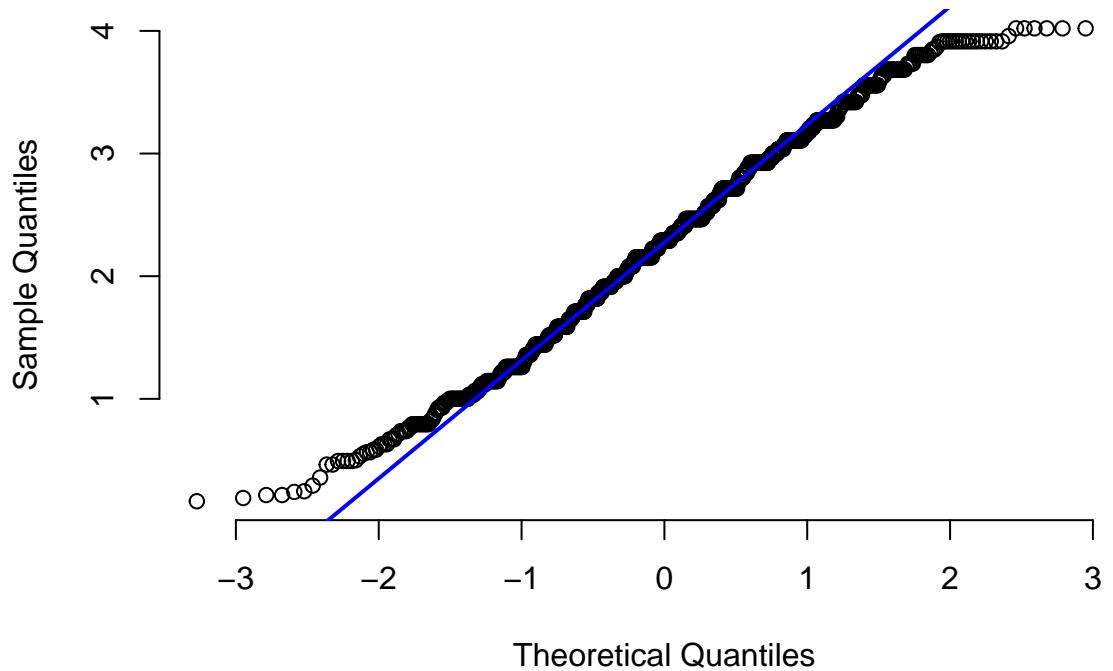
Financiranje Americkih filmova sa treci korijen funkcijom



```
treciKorijenNeAmericki <- neAmerickiFilmovi ^ (1/3)
```

```
qqnorm(treciKorijenNeAmericki, pch = 1, frame = FALSE, main='Financiranje Ne Americkih filmova sa treci')
```

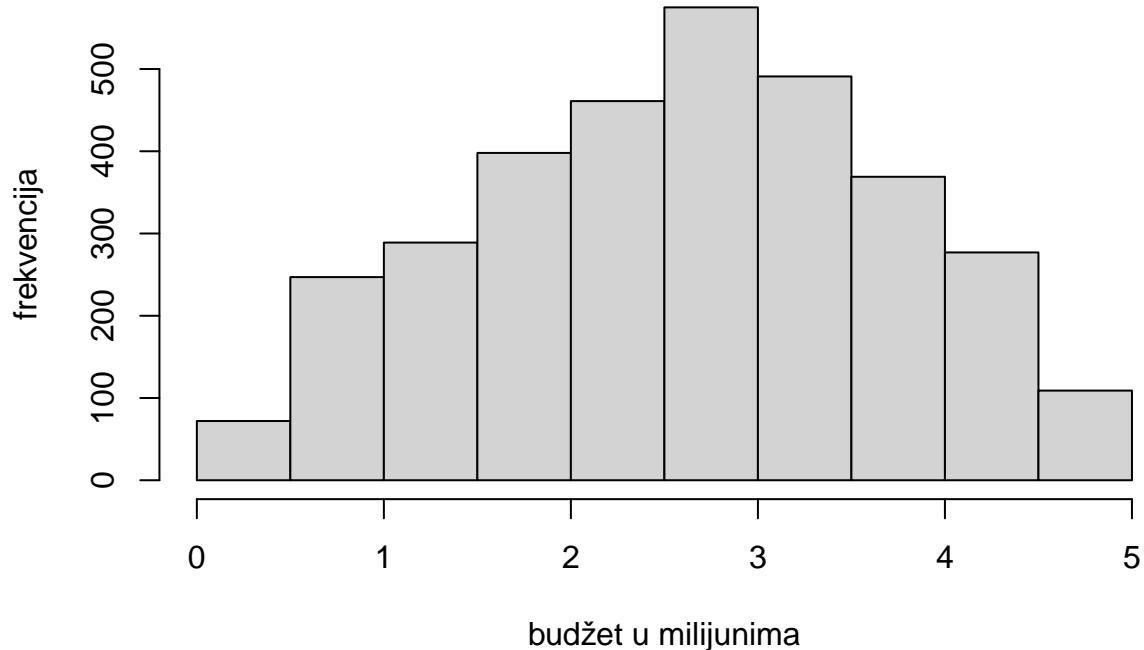
Financiranje Ne Americkih filmova sa treci korijen funkcijom



Pogledajmo sada i histograme za te novoskalirane podatke.

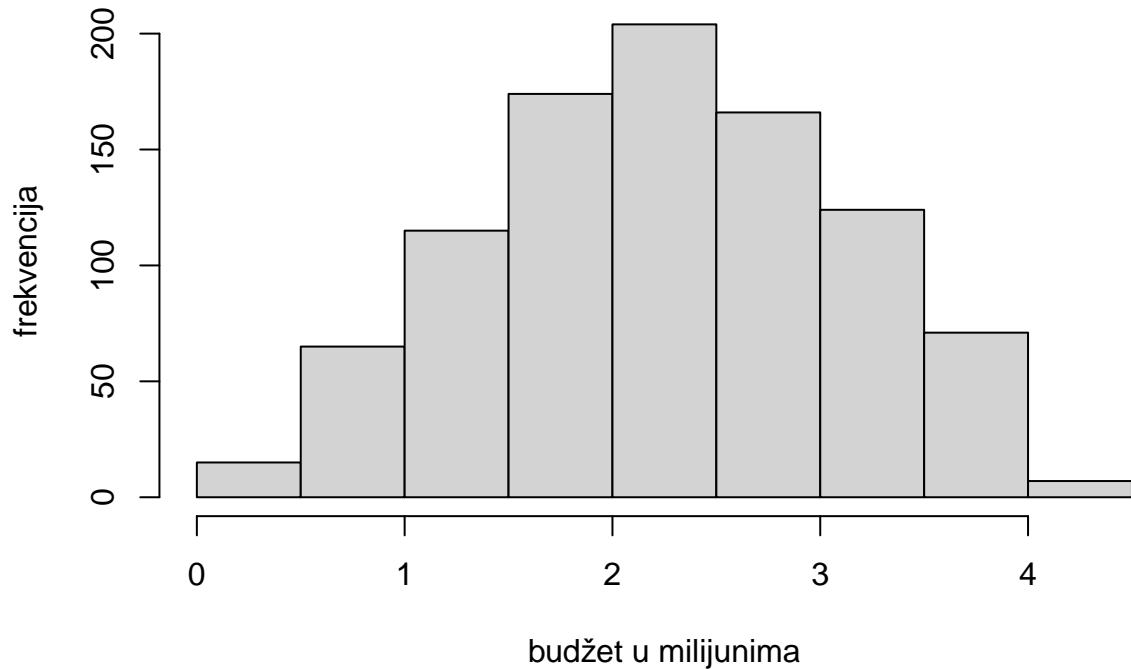
```
hist(treciKorijenAmericki,  
      main="Histogram Američkih filmova skaliranih funkcijom 3. korijen",  
      xlab="budžet u milijunima",  
      ylab="frekvencija"  
)
```

Histogram Americkih filmova skaliranih funkcijom 3. korijen



```
hist(treciKorijenNeAmericki,
      main="Histogram Američkih filmova filmova skaliranih funkcijom 3. korijen",
      xlab="budžet u milijunima",
      ylab="frekvencija"
    )
```

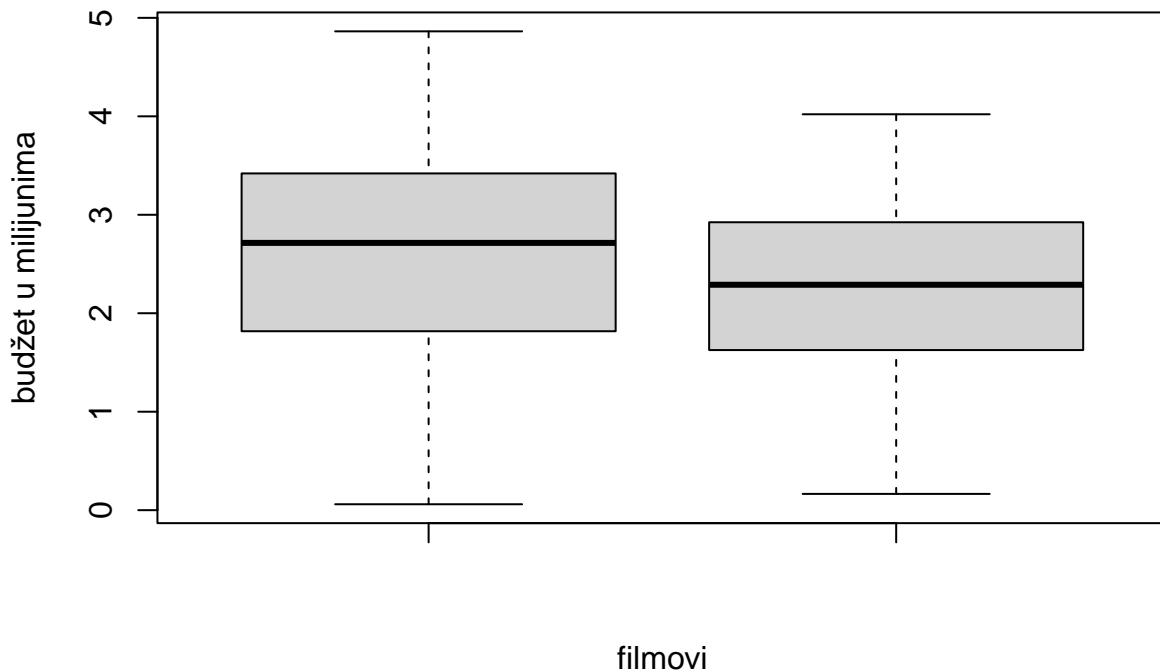
Histogram Americkih filmova filmova skaliranih funkcijom 3. korijen



Iz grafova vidimo da su distribucije sada poprilično normalne.

```
boxplot(treciKorijenAmericki, treciKorijenNeAmericki,  
        main="Box plot Američkih i ne Američkih filmova skaliranih funkcijom 3. korijen",  
        xlab="filmovi",  
        ylab="budžet u milijunima"  
)
```

Box plot Američih i ne Američkih filmova skaliranih funkcijom 3. kori



```
cat("Srednja vrijednost Američkih Filmova: ", mean(treciKorijenAmericki), "\n")
## Srednja vrijednost Američkih Filmova:  2.617406
cat("Srednja vrijednost ne Američkih Filmova: ", mean(treciKorijenNeAmericki), "\n")
## Srednja vrijednost ne Američkih Filmova:  2.268837
cat("Standardna devijacija Američkih Filmova: ", sd(treciKorijenAmericki), "\n")
## Standardna devijacija Američkih Filmova:  1.097046
cat("Standardna devijacija ne Američkih Filmova: ", sd(treciKorijenNeAmericki), "\n")
## Standardna devijacija ne Američkih Filmova:  0.8609953
Sada cemo provesti i F test za varijance.
var.test(treciKorijenAmericki,treciKorijenNeAmericki)

##
##  F test to compare two variances
##
## data: treciKorijenAmericki and treciKorijenNeAmericki
## F = 1.6235, num df = 3287, denom df = 940, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.463051 1.796185
## sample estimates:
## ratio of variances
```

```
##          1.623484
```

Iz njega vidimo da je p value skoro pa i 0 te iz toga zaključujemo da se varijance uistinu razlikuju

Sada postavimo hipoteze, H0... Američki filmovi imaju jednak budžet kao i ne Američki -> $m_1 = m_2$, $d_0 = 0$ H1... Američki filmovi imaju veći budžet -> $m_1 > m_2$, $d_0 > 0$ alfa = 5%

```
t.test(treciKorijenAmericki, treciKorijenNeAmericki, alternative = "greater", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: treciKorijenAmericki and treciKorijenNeAmericki
## t = 10.262, df = 1899.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.292669      Inf
## sample estimates:
## mean of x mean of y
## 2.617406  2.268837
```

Hipoteza H0 se odbija jer je p vrijednost manja od alfe.

Iz ovog možemo zaključiti da Američki filmovi imaju veći budžet od ne Američkih filmova, no ovo je bila provjera svih ne Američkih zajedno.

Idemo sada ponoviti isti postupak ali za neku zemlju zasebno, poput UK, izbacimo outliere, te prazne vrijednosti, te skaliramo sa funkcijom 3. korijen.

```
UKfilmovi <- imdb[imdb$country == "UK", ]$budget/1000000

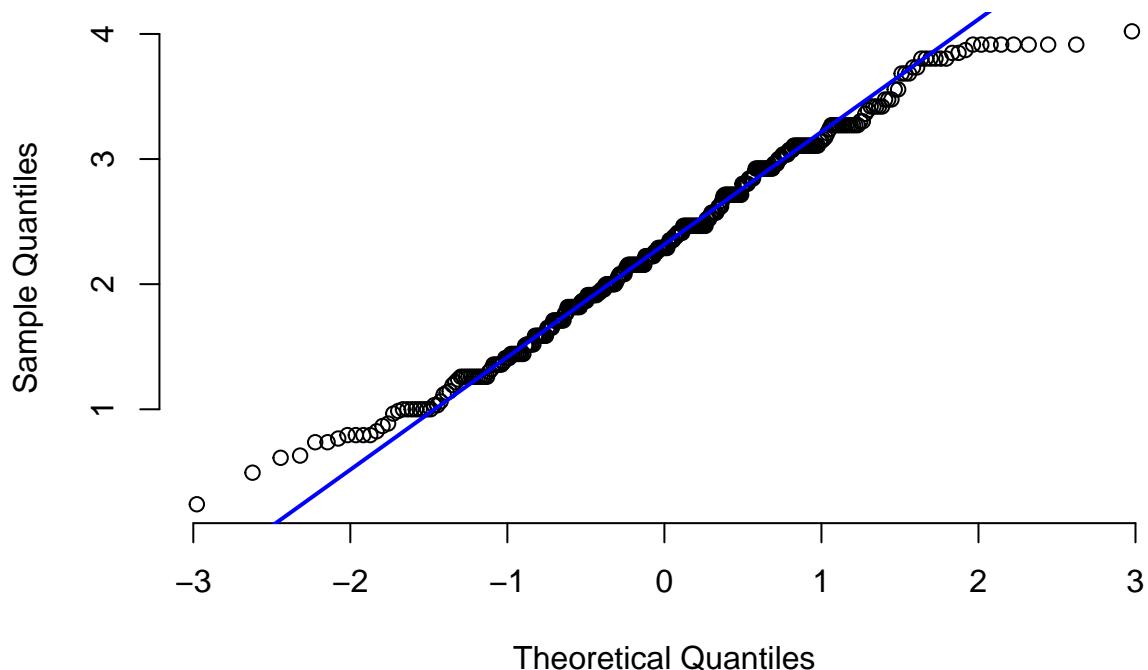
UKfilmovi <- na.omit(UKfilmovi)

UKfilmoviOutliers <- boxplot(UKfilmovi, plot=FALSE)$out
UKfilmovi <- UKfilmovi[-which(UKfilmovi %in% UKfilmoviOutliers)]


UKfilmovi <- UKfilmovi ^ (1/3)

qqnorm(UKfilmovi, pch = 1, frame = FALSE, main='Financiranje UK filmova sa treci korijen funkcijom')
qqline(UKfilmovi, col = "blue", lwd = 2)
```

Financiranje UK filmova sa treci korijen funkcijom



Iz grafa mozemo zaključiti normalnost.

```
cat("Srednja vrijednost UK Filmova: ", mean(UKfilmovi), "\n")
```

```
## Srednja vrijednost UK Filmova: 2.311431
```

```
cat("Standardna devijacija UK Filmova: ", sd(UKfilmovi), "\n")
```

```
## Standardna devijacija UK Filmova: 0.8210842
```

One sliče na vrijednosti koje smo dobili za sve ne američke filmove zajedno, no idemo provesti f test još jednom za svaki slučaj.

```
var.test(treciKorijenAmericki,UKfilmovi)
```

```
##
## F test to compare two variances
##
## data: treciKorijenAmericki and UKfilmovi
## F = 1.7851, num df = 3287, denom df = 343, p-value = 2.713e-11
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.517277 2.079029
## sample estimates:
## ratio of variances
## 1.785148
```

Kao što vidimo, opet možemo zaključiti da varijance nisu jednake jer je p vrijednost opet skoro pa jednaka nuli. Sada postavimo hipoteze, $H_0 \dots$ Američki filmovi imaju jednak budžet kao i UK filmovi $\rightarrow m_1 = m_2$, $d_0 = 0$ $H_1 \dots$ Američki filmovi imaju veći budžet $\rightarrow m_1 > m_2$, $d_0 > 0$ alfa = 5%

```
t.test(treciKorijenAmericki, UKfilmovi, alternative = "greater", var.equal = FALSE)

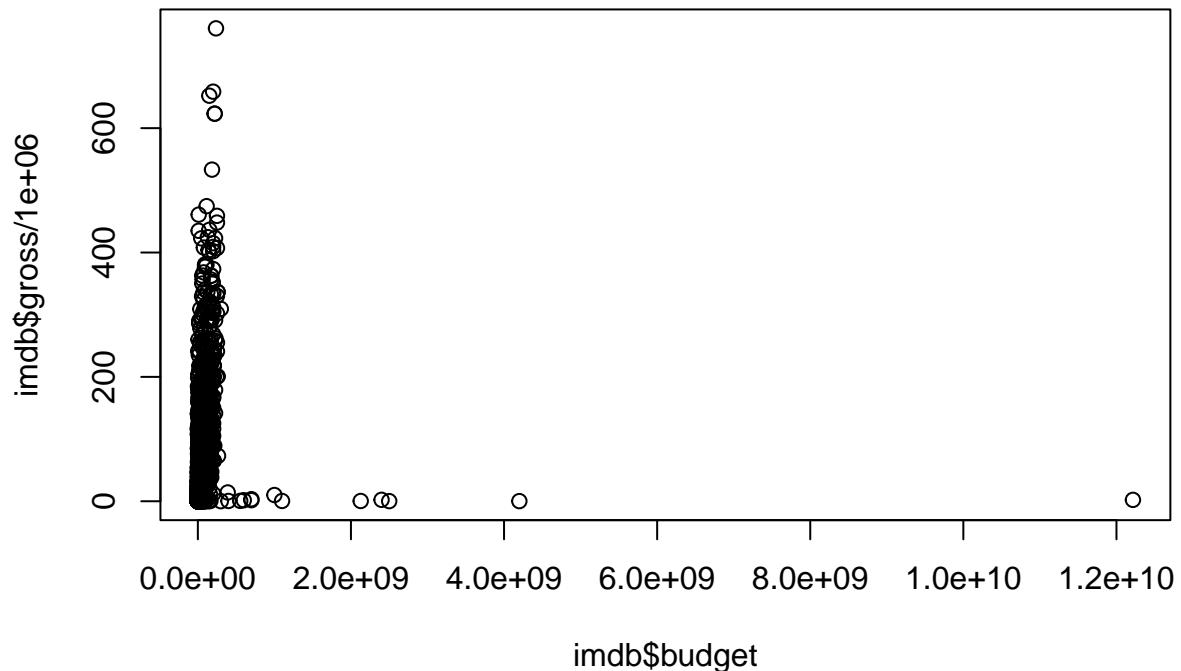
##
##  Welch Two Sample t-test
##
## data: treciKorijenAmericki and UKfilmovi
## t = 6.3444, df = 481.33, p-value = 2.578e-10
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2264948      Inf
## sample estimates:
## mean of x mean of y
##  2.617406   2.311431
```

p vrijednost je i u ovom testu značajno manja od alfe te zaključujemo da Američki filmovi dobivaju veći budžet od UK filmova

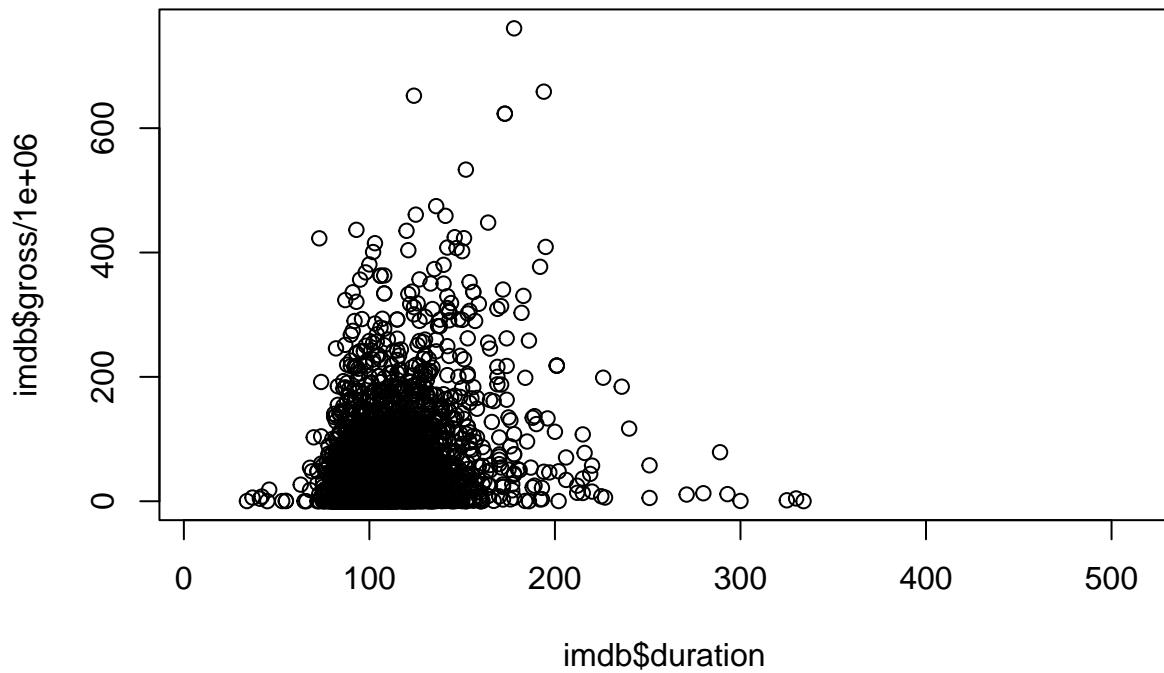
Možemo li temeljem danih varijabli predvidjeti zaradu filmova?

Trebamo provjeriti postoji li veza između ulaznih varijabli (regresora) i izlazne varijable (reakcije) koja je u našem slučaju zarada(gross) filma. Za to ćemo koristiti linearnu regresiju. Želimo saznati koliko su te veze jake, koje ulazne varijable najviše utječu na zaradu, te vidjeti možemo li predvidjeti zaradu na temelju nekih ulaznih varijabli i s kojom točnošću. Tražimo sve varijable koje bi mogle utjecati na zaradu filma, prvo bez uklanjanja NA vrijednosti i stršecih vrijednosti. Zarada je prikazana u milijunima dolara. Moramom uzeti u obzir da nisu sve varijable poznate pri izlasku novog filma. Prvo ćemo ispitati utjecaj sljedećih varijabli: - budžet (budget) - trajanje filma (duration) - godina izlaska filma (title_year)

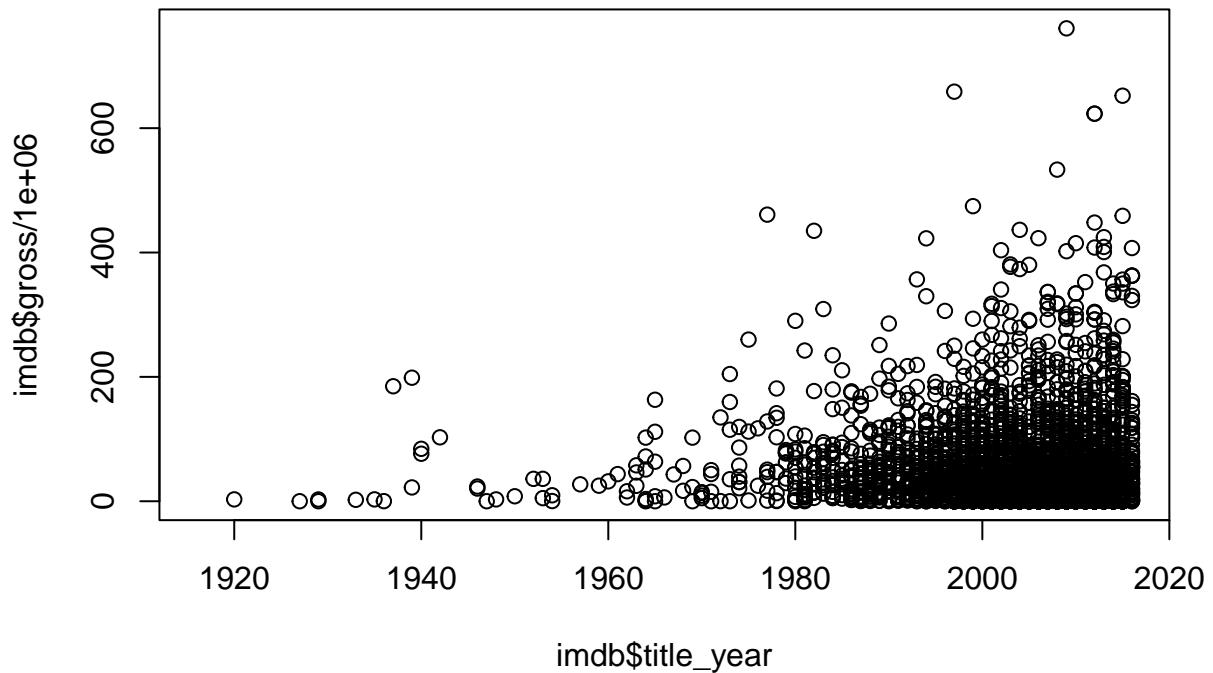
```
plot(imdb$budget, imdb$gross/1000000)
```



```
plot(imdb$duration,imdb$gross/1000000)
```



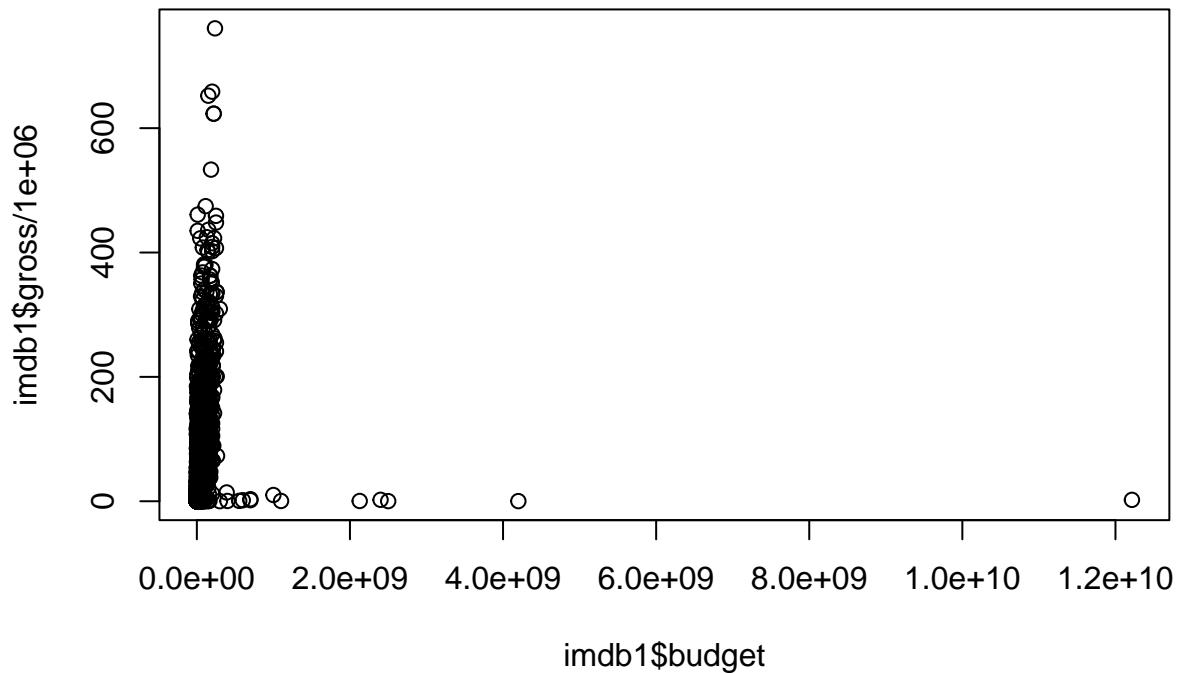
```
plot(imdb$title_year,imdb$gross/1000000)
```



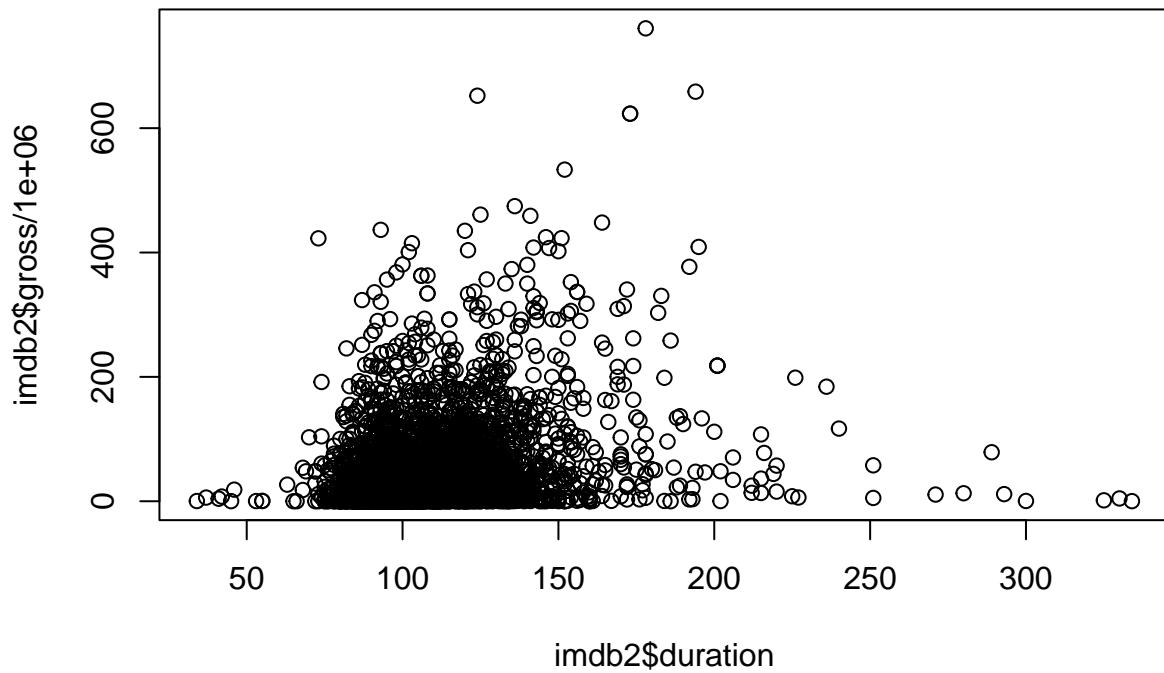
Utjecaj svake od varijabli na zarađu prikazali smo pomoću scatter plot-a. Vidimo da u dosta slučajeva postoje stršeće vrijednosti i NA vrijednosti zbog kojih je teže jasno vidjeti odnos varijabli. Radi preglednosti ukloniti ćemo NA vrijednosti i ponoviti postupak.

```
imdb0 <- imdb[!is.na(imdb$gross),]
imdb1 <- imdb0[!is.na(imdb0$budget),]
imdb2 <- imdb0[!is.na(imdb0$duration),]
imdb3 <- imdb0[!is.na(imdb0$title_year),]

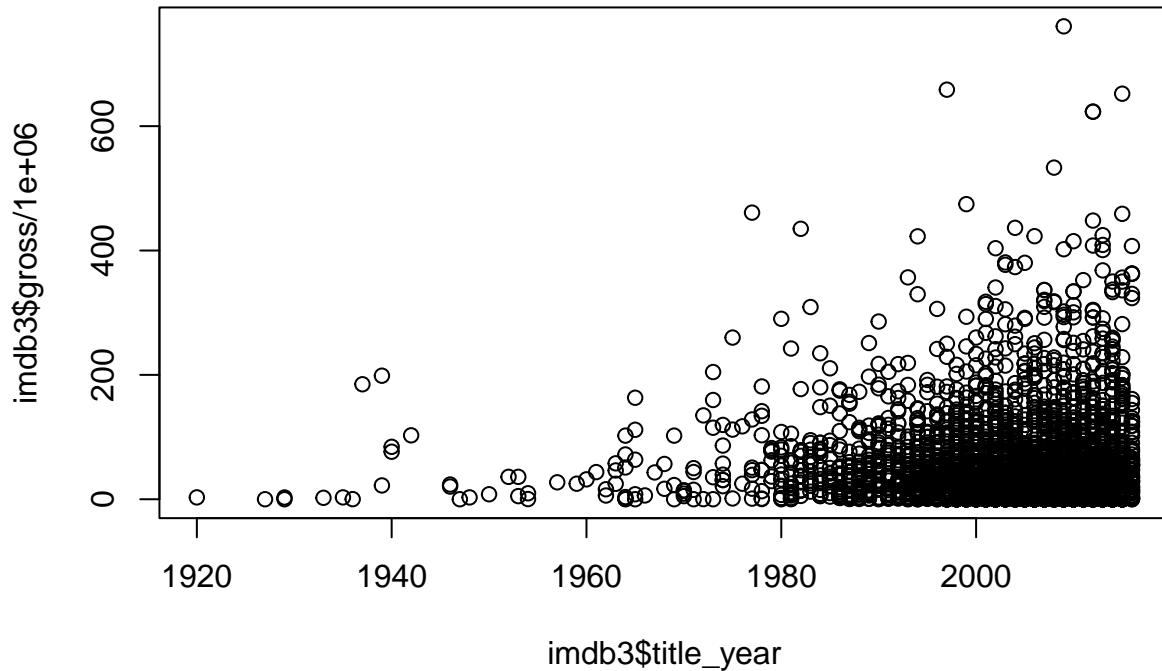
plot(imdb1$budget, imdb1$gross/1000000)
```



```
plot(imdb2$duration,imdb2$gross/1000000)
```



```
plot(imdb3$title_year,imdb3$gross/1000000)
```



Pošto ne vidimo preveliku promjenu, ukloniti ćemo i stršeće vrijednosti.

```

outliersgross <- boxplot(imdb1$gross, plot=FALSE)$out
imdbh1 <- imdb1[-which(imdb1$gross %in% outliersgross),]

outliersgross <- boxplot(imdb2$gross, plot=FALSE)$out
imdbh2 <- imdb2[-which(imdb2$gross %in% outliersgross),]

outliersgross <- boxplot(imdb3$gross, plot=FALSE)$out
imdbh3 <- imdb3[-which(imdb3$gross %in% outliersgross),]

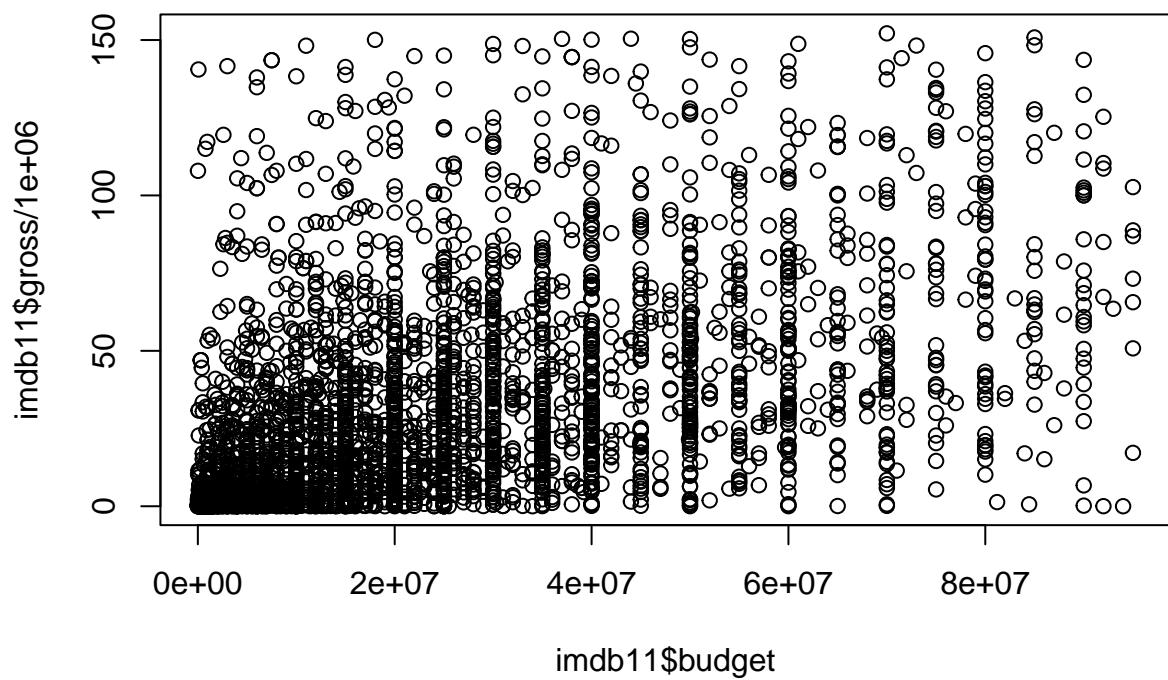
outliersbudget <- boxplot(imdbh1$budget, plot=FALSE)$out
imdb11 <- imdbh1[-which(imdbh1$budget %in% outliersbudget),]

outliersdur <- boxplot(imdbh2$duration, plot=FALSE)$out
imdb22 <- imdbh2[-which(imdbh2$duration %in% outliersdur),]

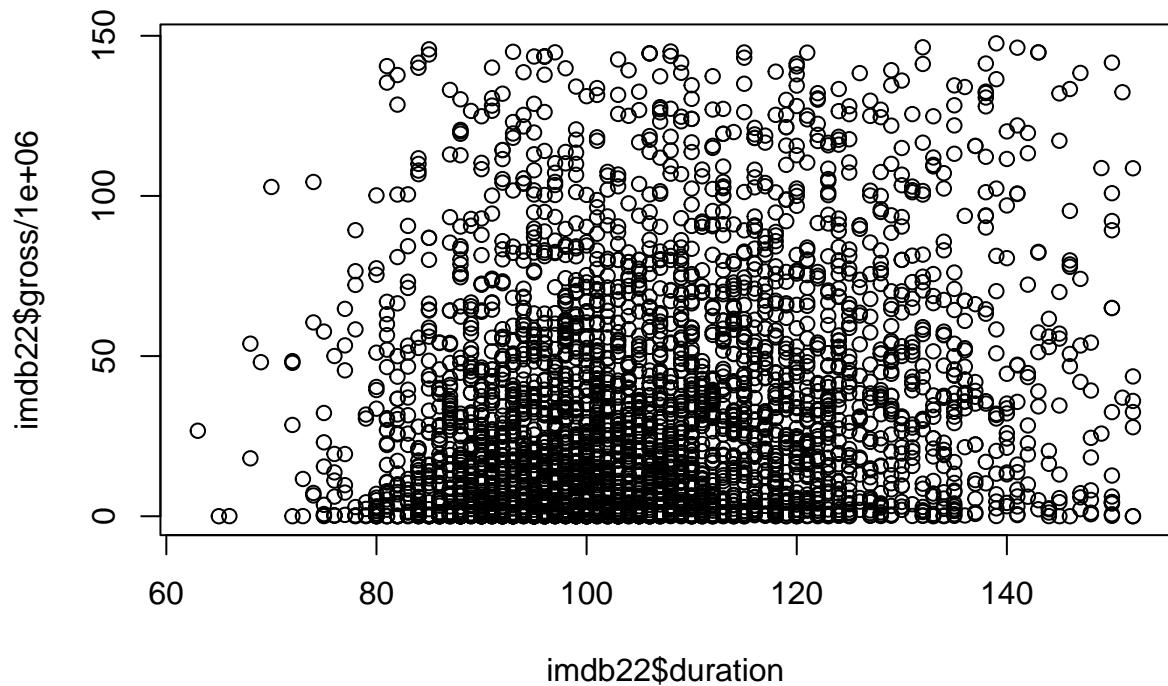
outliersty <- boxplot(imdbh3$title_year, plot=FALSE)$out
imdb33 <- imdbh3[-which(imdbh3$title_year %in% outliersty),]

plot(imdb11$budget,imdb11$gross/1000000)

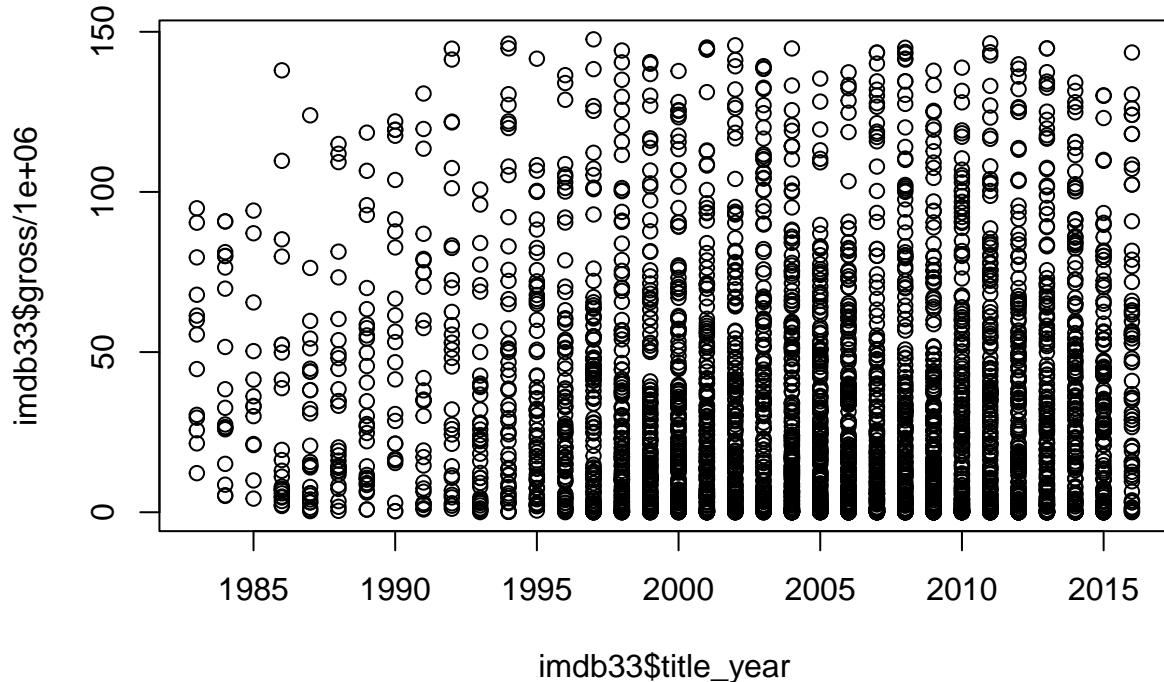
```



```
plot(imdb22$duration,imdb22$gross/1000000)
```



```
plot(imdb33$title_year,imdb33$gross/1000000)
```



Primjećujemo da niti jedna od ulaznih varijabli nema jako veliku povezanost, ali npr. budget pokazuje određenu povezanost.

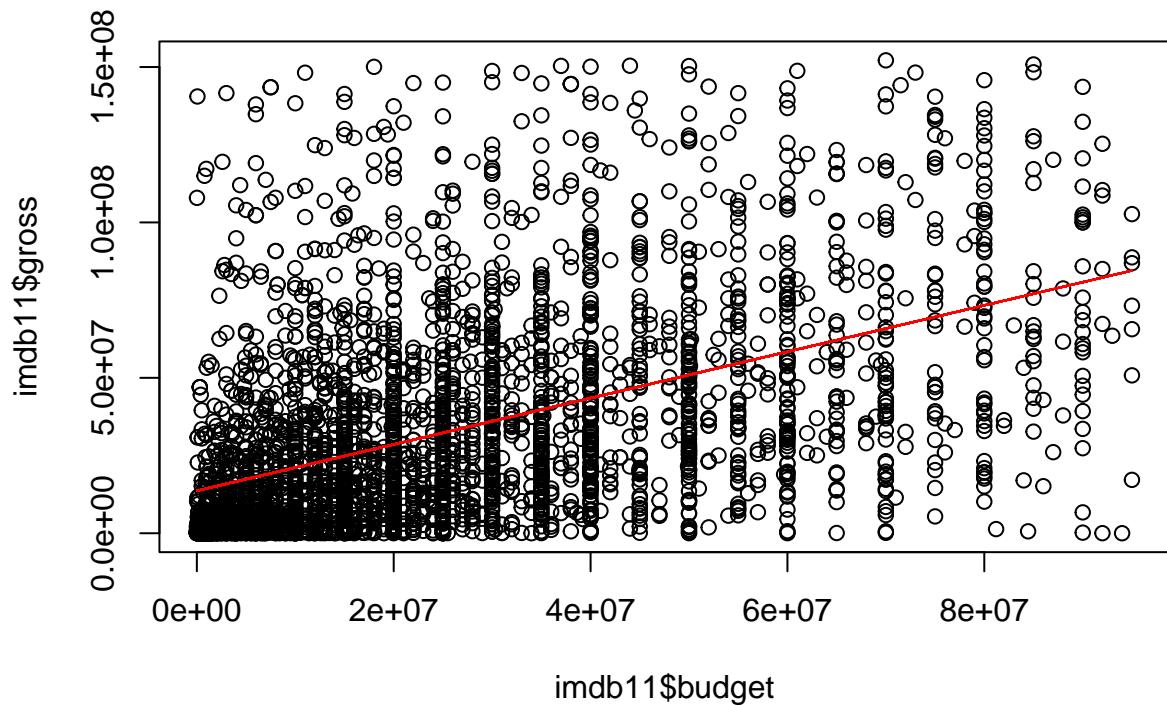
Sada ćemo na dijagrame dodati pravce linearne regresije kako bismo bolje vidjeli efekte varijabli na zaradu.

```
fit.budget = lm(gross~budget,data=imdb11)
```

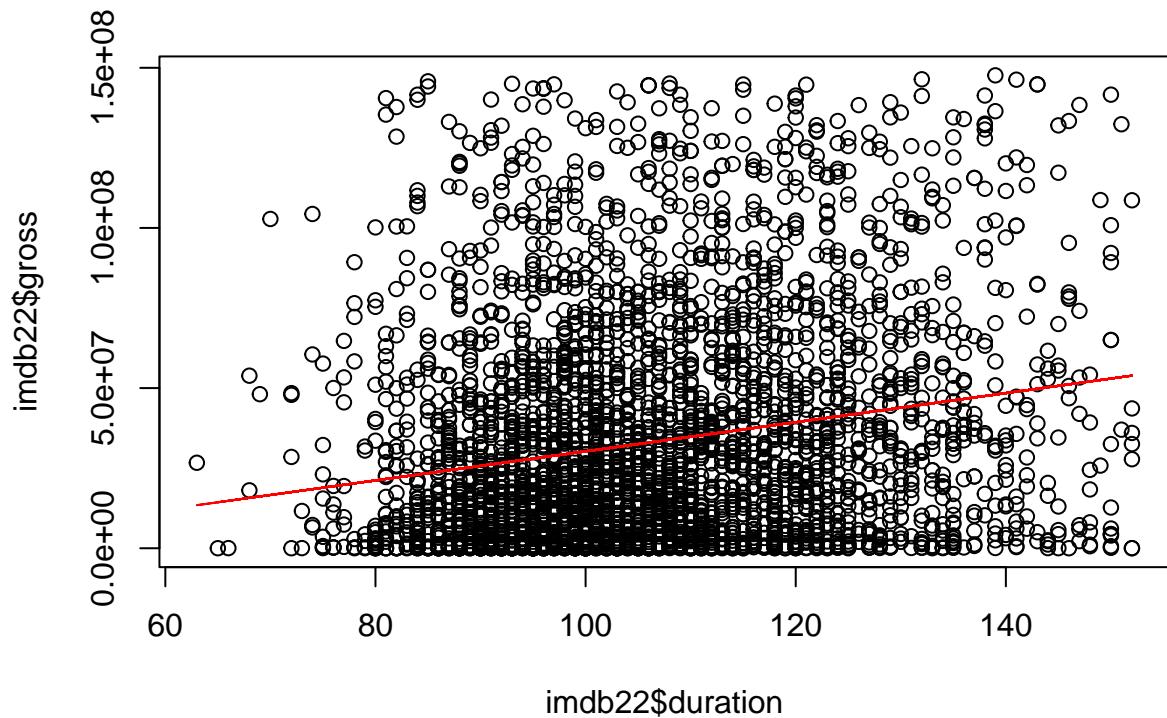
```
fit.dur = lm(gross~duration,data=imdb22)
```

```
fit.ty = lm(gross~title_year,data=imdb33)
```

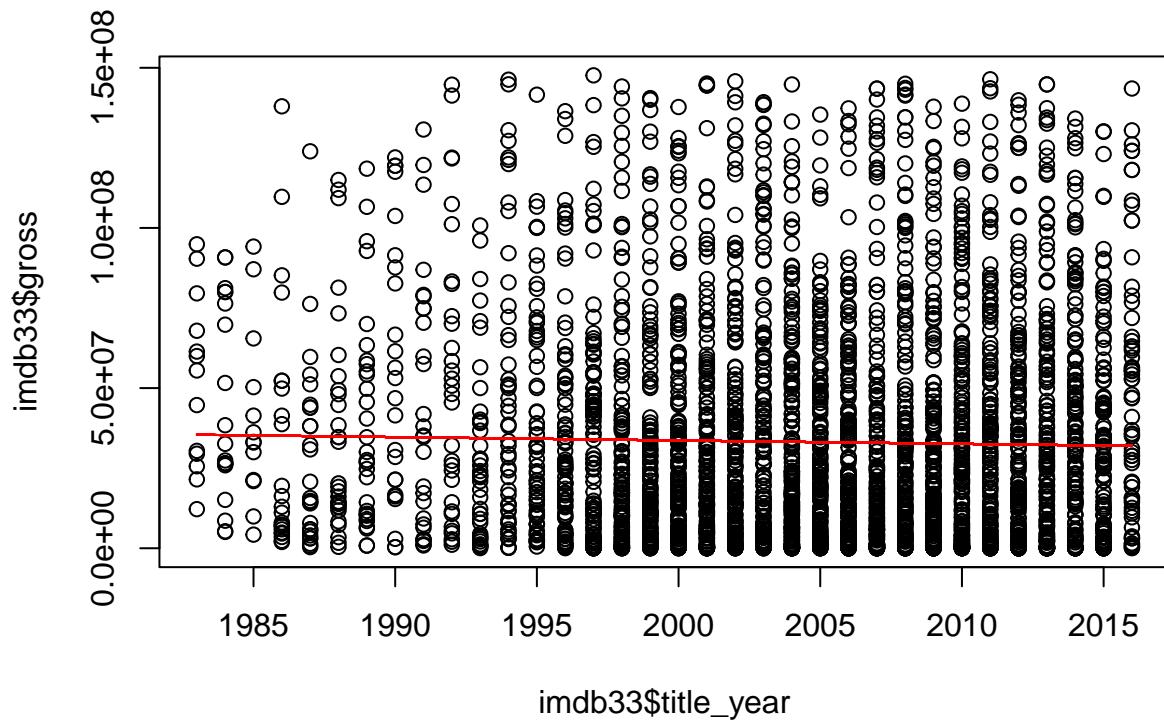
```
plot(imdb11$budget,imdb11$gross)
lines(imdb11$budget,fit.budget$fitted.values,col='red')
```



```
plot(imdb22$duration,imdb22$gross)
lines(imdb22$duration,fit.dur$fitted.values,col='red')
```



```
plot(imdb33$title_year,imdb33$gross)
lines(imdb33$title_year,fit.ty$fitted.values,col='red')
```



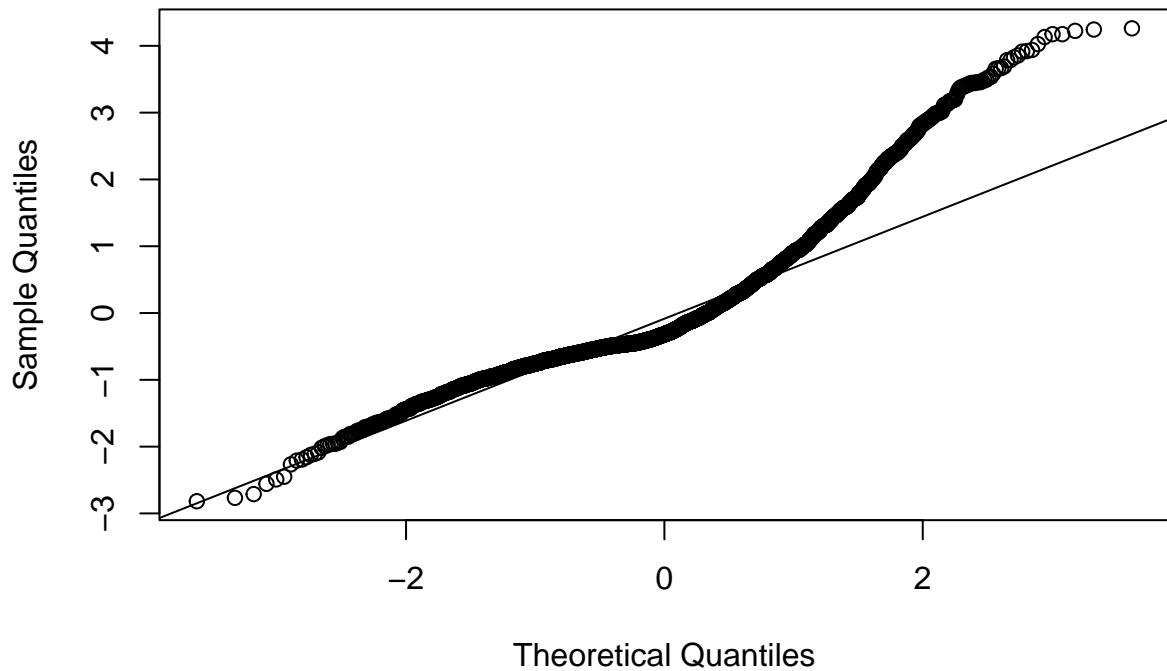
Sada bolje vidimo utjecaje na zaradu, budget ima najveći utjecaj što vidimo po nagibu pravca linearne regresije.

Sada ćemo provjeriti normalnost reziduala za svaki model pomoću kvantil-kvantil plota.

```
selected.model = fit.budget
```

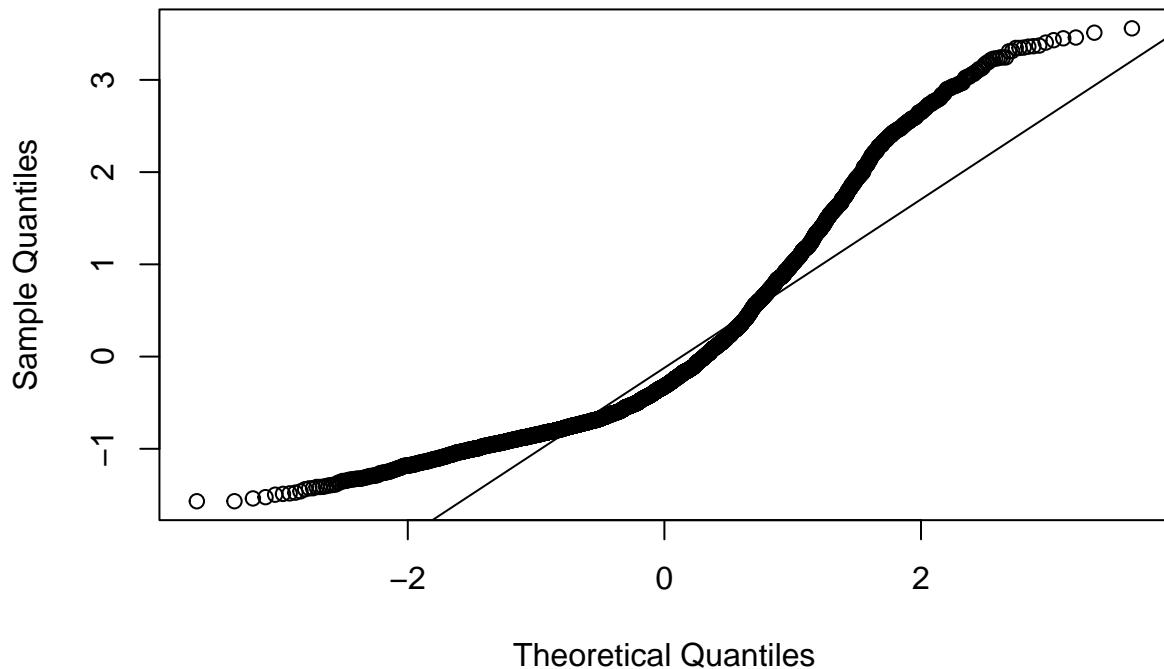
```
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```

Normal Q-Q Plot

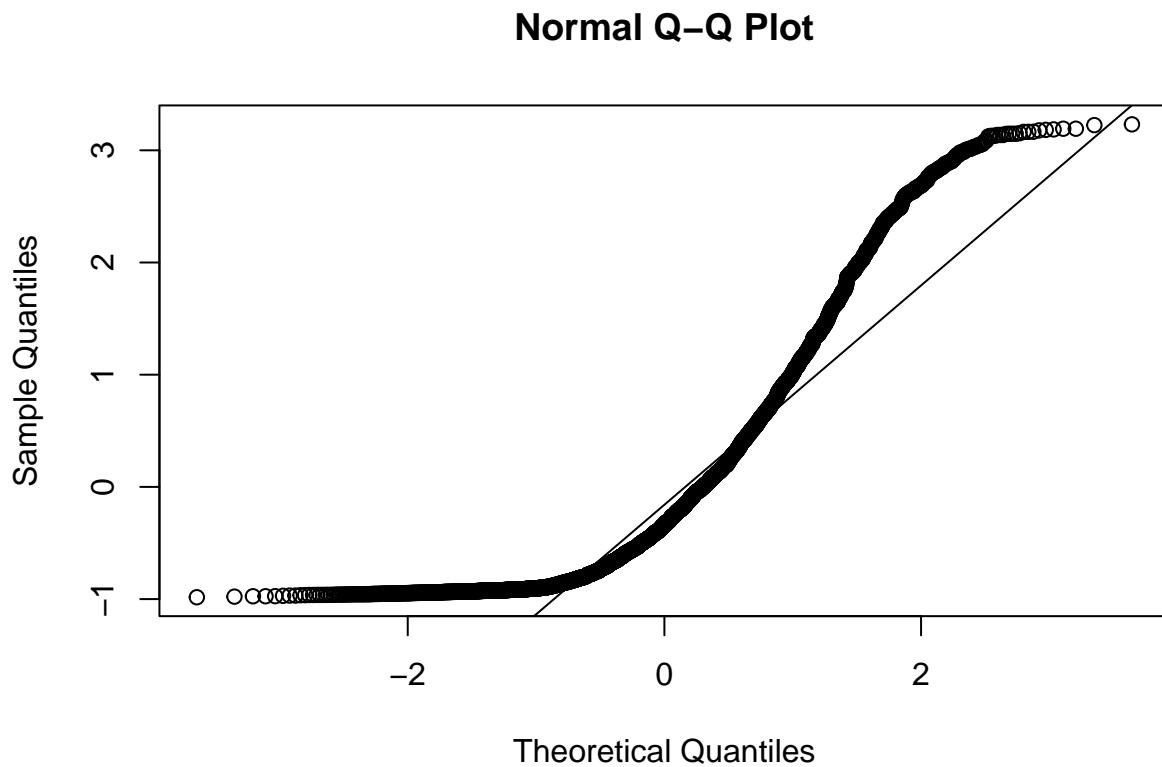


```
selected.model = fit.dur  
qqnorm(rstandard(selected.model))  
qqline(rstandard(selected.model))
```

Normal Q-Q Plot



```
selected.model = fit.ty  
qqnorm(rstandard(selected.model))  
qqline(rstandard(selected.model))
```



Distribucije donekle nalikuju na normalnu razdiobu, ali sve bi ih trebalo još više približiti normalnoj razdiobi ako želimo pretpostaviti normalnost reziduala. Pokušat ćemo to riješiti kasnije doadvanjem drugih varijabli u model te dodavanjem interakcijskih ili polinomijalnih članova.

Pogledat ćemo mjere kvalitete prilagodbe modela podatcima.

```
summary(fit.budget)

##
## Call:
## lm(formula = gross ~ budget, data = imbd11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -83750518 -17858044  -9519916  12785331 126825136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.366e+07 7.808e+05 17.50 <2e-16 ***
## budget      7.457e-01 2.290e-02 32.56 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29770000 on 3387 degrees of freedom
## Multiple R-squared:  0.2384, Adjusted R-squared:  0.2382
## F-statistic: 1060 on 1 and 3387 DF,  p-value: < 2.2e-16
```

```

summary(fit.dur)

##
## Call:
## lm(formula = gross ~ duration, data = imdb22)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53868950 -25370027 -11104499  16990537 122330959
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15181473    3870369  -3.922 8.92e-05 ***
## duration     454377     36081   12.593 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34390000 on 3722 degrees of freedom
## Multiple R-squared:  0.04087, Adjusted R-squared:  0.04061
## F-statistic: 158.6 on 1 and 3722 DF, p-value: < 2.2e-16
summary(fit.ty)

```

```

##
## Call:
## lm(formula = gross ~ title_year, data = imdb33)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34653105 -28845952 -11872125  17627663 113924713
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 244234867  162053641   1.507   0.132
## title_year   -105298     80849  -1.302   0.193
## 
## Residual standard error: 35260000 on 3711 degrees of freedom
## Multiple R-squared:  0.0004569, Adjusted R-squared:  0.0001875
## F-statistic: 1.696 on 1 and 3711 DF, p-value: 0.1929

```

Za svaki od modela možemo vidjeti intercept i slope, pomoću koji možemo procijeniti zaradu za bilo koju vrijednost ulazne varijable.

Primjećujemo da nisu svi modeli jednako kvalitetni. Vidimo da najveći utjecaj ima budget kao što smo i pretpostavili prije, to se očituje najvećim vrijednostima R^2 . Vidimo da duration nema utjecaj kao budget ali ima određen značaj, dok za title_year vidimo da nema gotovo nikakav utjecaj (R^2 je skoro 0).

Pogledajmo i Pearsonov koeficijent korelacije za sljedeće 3 varijable:

```

cor(imdb11$budget,imdb11$gross)

## [1] 0.4882675

cor(imdb22$duration,imdb22$gross)

## [1] 0.2021594

```

```

cor(imdb33$title_year,imdb33$gross)

## [1] -0.02137475

On dodatno potvrđuje veći utjecaj nekih varijabli.

Sada ćemo napraviti procjenu modela višestruke regresije. Prije procjene modela višestruke regresije moramo provjeriti da nam varijable nisu međusobno previše korelirane.

imdbc  $\leftarrow$  imdb[!is.na(imdb$gross),]
imdbc  $\leftarrow$  imdbc[!is.na(imdbc$budget),]
imdbc  $\leftarrow$  imdbc[!is.na(imdbc$duration),]
imdbc  $\leftarrow$  imdbc[!is.na(imdbc$title_year),]

cor(cbind(imdbc$budget,imdbc$duration,imdbc$title_year))

##           [,1]      [,2]      [,3]
## [1,] 1.00000000 0.0695740 0.04498474
## [2,] 0.06957400 1.0000000 -0.12819764
## [3,] 0.04498474 -0.1281976 1.00000000

```

Vidimo da nam varijable nisu previše korelirane i možemo procijeniti model.

```

fit.multi = lm(gross ~ budget + duration + title_year, imdb11)
summary(fit.multi)

##
## Call:
## lm(formula = gross ~ budget + duration + title_year, data = imdb11)
##
## Residuals:
##       Min        1Q        Median        3Q        Max
## -83815739 -17712733 -8992370  12470409 127527240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.939e+08 1.048e+08 6.624 4.06e-11 ***
## budget      7.556e-01 2.368e-02 31.913 < 2e-16 ***
## duration    5.566e+04 2.523e+04  2.206  0.0274 *
## title_year -3.428e+05 5.206e+04 -6.584 5.29e-11 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29510000 on 3384 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared: 0.2518, Adjusted R-squared: 0.2512
## F-statistic: 379.7 on 3 and 3384 DF, p-value: < 2.2e-16

```

Vidimo da nam ovaj model nije mnogo bolji nego onaj koji sadrži samo buget. Zato ćemo radi jednostavnosti nastaviti s modelom koji sadrži samo budget.

Sada ćemo probati poboljšati model dodavajući kategoriskske varijable. Kategoriskske varijable koje ćemo uzeti u obzir su: - color - language - country - content_rating - aspect_ratio

```

fit.multi.1 = lm(gross ~ budget + country , imdb11)
summary(fit.multi.1)

##

```

```

## Call:
## lm(formula = gross ~ budget + country, data = imdb11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -72445597 -18292812  -7067974  11990331 126946565
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           1.094e+06  2.919e+07  0.037   0.970    
## budget                7.224e-01  2.285e-02 31.616 <2e-16 ***
## countryArgentina     4.957e+06  3.370e+07  0.147   0.883    
## countryAruba          -1.630e+07 4.128e+07 -0.395   0.693    
## countryAustralia      1.842e+06  2.963e+07  0.062   0.950    
## countryBelgium        -1.811e+07 3.575e+07 -0.507   0.612    
## countryBrazil         -1.011e+06 3.197e+07 -0.032   0.975    
## countryCanada         4.854e+06  2.942e+07  0.165   0.869    
## countryChile          -7.687e+06 4.128e+07 -0.186   0.852    
## countryChina          -2.184e+07 3.039e+07 -0.719   0.472    
## countryColombia       3.256e+06  4.128e+07  0.079   0.937    
## countryCzech Republic -3.506e+07 3.372e+07 -1.040   0.299    
## countryDenmark        -9.733e+06 3.077e+07 -0.316   0.752    
## countryFinland        -3.264e+06 4.128e+07 -0.079   0.937    
## countryFrance         -1.099e+06 2.934e+07 -0.037   0.970    
## countryGeorgia        -1.552e+07 4.128e+07 -0.376   0.707    
## countryGermany        3.892e+06  2.938e+07  0.132   0.895    
## countryGreece         -1.039e+07 4.128e+07 -0.252   0.801    
## countryHong Kong      -7.735e+06 3.030e+07 -0.255   0.798    
## countryHungary        7.638e+06  4.128e+07  0.185   0.853    
## countryIceland        -6.063e+06 3.575e+07 -0.170   0.865    
## countryIndia          -6.177e+06 3.049e+07 -0.203   0.839    
## countryIndonesia      2.216e+06  4.128e+07  0.054   0.957    
## countryIran           -1.323e+06 3.263e+07 -0.041   0.968    
## countryIreland        -1.561e+06 3.120e+07 -0.050   0.960    
## countryIsrael          -1.100e+06 3.370e+07 -0.033   0.974    
## countryItaly           -6.876e+06 3.049e+07 -0.226   0.822    
## countryJapan           3.249e+06  3.038e+07  0.107   0.915    
## countryMexico          2.517e+06  3.049e+07  0.083   0.934    
## countryNetherlands     -6.313e+06 3.370e+07 -0.187   0.851    
## countryNew Line        -5.939e+07 4.133e+07 -1.437   0.151    
## countryNew Zealand     7.008e+05  3.153e+07  0.022   0.982    
## countryNorway          -1.265e+07 3.263e+07 -0.388   0.698    
## countryOfficial site   8.289e+06  4.128e+07  0.201   0.841    
## countryPeru            2.376e+07 4.129e+07  0.575   0.565    
## countryPhilippines     -1.029e+06 4.128e+07 -0.025   0.980    
## countryPoland          -1.849e+06 4.128e+07 -0.045   0.964    
## countryRomania         3.460e+06  3.575e+07  0.097   0.923    
## countryRussia          -6.918e+06 3.370e+07 -0.205   0.837    
## countrySouth Africa    3.876e+07 3.371e+07  1.150   0.250    
## countrySouth Korea     -1.137e+07 3.121e+07 -0.364   0.716    
## countrySpain           -4.300e+06 2.988e+07 -0.144   0.886    
## countrySweden          -1.914e+07 4.128e+07 -0.464   0.643    
## countryTaiwan          5.241e+07  3.575e+07  1.466   0.143    
## countryThailand        -7.595e+06 4.128e+07 -0.184   0.854

```

```

## countryUK           6.529e+06  2.924e+07   0.223    0.823
## countryUSA          1.625e+07  2.920e+07   0.556    0.578
## countryWest Germany 2.259e+05  4.128e+07   0.005    0.996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29190000 on 3341 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2677
## F-statistic: 27.35 on 47 and 3341 DF,  p-value: < 2.2e-16

```

Dodavanjem varijable country model nam se poboljšava, ali u stupcu country se pojavljuje jako puno država, što bi nam zakompliciralo model, a i unos podataka pri procjeni zarade za novi film. Pogledajmo koje sve vrijednosti može poprimiti stupac country.

```
table(imdb11['country'])
```

```

## 
##      Afghanistan      Argentina       Aruba      Australia      Belgium
##                      1                  3                  1                 33                  2
##      Brazil          Canada        Chile       China      Colombia
##                      5                  62                  1                 12                  1
## Czech Republic Denmark     Finland     France      Georgia
##                      3                  9                  1                100                  1
##      Germany        Greece Hong Kong     Hungary     Iceland
##                      77                  1                 13                  1                  2
##      India          Indonesia     Iran      Ireland      Israel
##                      11                  1                  4                  7                  3
##      Italy            Japan      Mexico  Netherlands New Line
##                      11                 12                 11                  3                  1
## New Zealand      Norway Official site      Peru Philippines
##                      6                  4                  1                  1                  1
##      Poland          Romania      Russia South Africa South Korea
##                      1                  2                  3                  3                  7
##      Spain            Sweden      Taiwan Thailand      UK
##                      21                  1                  2                  1                298
##      USA      West Germany
##                      2643                  1

```

Vidimo da ima uvjerljivo najviše američkih filmova, zato ćemo zbog jednostavnosti zamijeniti sve ostale države s OtherCountry.

```

imdbcou = imdb11
imdbcou$country[imdbcou$country!="USA"]<- "OtherCountry"
table(imdbcou['country'])

```

```

## 
##      OtherCountry      USA
##                      746      2643

```

Nakon zamjene isprobajmo novi model:

```

fit.multi.2 = lm(gross ~ budget + country, imdbcou)
summary(fit.multi.2)

```

```

## 
## Call:
## lm(formula = gross ~ budget + country, data = imdbcou)
## 
```

```

## Residuals:
##      Min       1Q    Median       3Q      Max
## -70795554 -18364305  -6907517   12226237 130188457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.477e+06  1.170e+06   2.972  0.00298 **
## budget      7.163e-01  2.262e-02  31.670 < 2e-16 ***
## countryUSA 1.403e+07  1.219e+06  11.513 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29210000 on 3386 degrees of freedom
## Multiple R-squared:  0.2671, Adjusted R-squared:  0.2667
## F-statistic:   617 on 2 and 3386 DF,  p-value: < 2.2e-16

```

Model je gotovo jednako dobar što možemo vidjeti po ako usporedimo vrijednosti Adjusted R-squared (prije 0.2677, sada 0.2667).

Isprobajmo sada dodati još neke kategoriske varijable u naš model.

```

fit.multi.3 = lm(gross ~ budget + country + color, imdbcou)
summary(fit.multi.3)

```

```

##
## Call:
## lm(formula = gross ~ budget + country + color, data = imdbcou)
##
## Residuals:
##      Min       1Q    Median       3Q      Max
## -70870456 -18423950  -6928318   12224538 130011883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.260e+07  2.066e+07   1.094  0.274
## budget      7.151e-01  2.264e-02  31.589 <2e-16 ***
## countryUSA 1.399e+07  1.220e+06  11.467 <2e-16 ***
## color Black and White -2.257e+07  2.082e+07  -1.084  0.279
## colorColor  -1.894e+07  2.066e+07  -0.917  0.359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29200000 on 3384 degrees of freedom
## Multiple R-squared:  0.2677, Adjusted R-squared:  0.2668
## F-statistic: 309.2 on 4 and 3384 DF,  p-value: < 2.2e-16

```

Dodavanjem varijable color nismo znatno poboljšali model, zato ćemo odbaciti color.

Prije dodavanja varijable language u model, pogledajmo koje sve vrijednosti postoje u tom stupcu.

```
table(imdbcou['language'])
```

	Aboriginal	Arabic	Aramaic	Bosnian	Cantonese	Czech
3	2	1	1	1	8	1
Danish	Dari	Dutch	Dzongkha	English	Filipino	French
3	2	3	1	3222	1	35

```

##      German      Hebrew      Hindi      Icelandic      Indonesian      Italian      Japanese
##      13          3          8          1          2          7          8
##      Kazakh      Korean      Mandarin      Maya      Mongolian      None      Norwegian
##      1          3          13          1          1          1          1          4
##      Persian      Portuguese      Romanian      Russian      Spanish      Swedish      Telugu
##      3          5          1          1          25          1          1
## Vietnamese      Zulu
##      1          1

```

Engleski je očekivano najčešći jezik, zato ćemo sve ostale zamijeniti s OtherLanguage.

```

imdblan = imdbcou
imdblan$language[imdblan$language!="English"] <- "OtherLanguage"
table(imdblan['language'])

```

```

##
##      English      OtherLanguage
##      3222          167
fit.multi.lan = lm(gross ~ budget + country + language, imdblan)
summary(fit.multi.lan)

##
## Call:
## lm(formula = gross ~ budget + country + language, data = imdblan)
##
## Residuals:
##      Min      1Q      Median      3Q      Max
## -71000893 -18514245  -7155874   12055940  127826996
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.888e+06  1.282e+06   4.591 4.57e-06 ***
## budget                  7.080e-01  2.262e-02  31.296 < 2e-16 ***
## countryUSA              1.192e+07  1.301e+06   9.163 < 2e-16 ***
## languageOtherLanguage -1.128e+07  2.492e+06  -4.528 6.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29120000 on 3385 degrees of freedom
## Multiple R-squared:  0.2715, Adjusted R-squared:  0.2709
## F-statistic: 420.5 on 3 and 3385 DF,  p-value: < 2.2e-16

```

Ponovno ne vidimo bitno poboljšanje u modelu, ponovit ćemo postupak za ostale kategoriske varijable, kako bismo pronašli najbolji model.

```
table(imdbcou['content_rating'])
```

```

##
##      Approved      G      GP      M      NC-17      Not Rated      Passed
##      48          16          68          1          2          6          42          3
##      PG      PG-13          R      Unrated          X
##      441         1061        1668         23         10

```

Mijenjamo sve vrijednosti osim PG, PG-13 i R u OtherRating.

```

imdbcrr = imdblan
imdbcrr$content_rating[imdbcrr$content_rating!="R" & imdbcrr$content_rating!="PG-13" & imdbcrr$content_rating!="NC-17"] <- "OtherRating"
table(imdbcrr['content_rating'])

```

```

table(imdbcrr['content_rating'])

##
## OtherRating          PG        PG-13         R
##      219           441       1061      1668
fit.multi.cr = lm(gross ~ budget + country + content_rating, imdbcrr)
summary(fit.multi.cr)

##
## Call:
## lm(formula = gross ~ budget + country + content_rating, data = imdbcrr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -73269138 -18109887 -6543939  11852489 131781444
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.630e+06  2.104e+06   1.250 0.211398
## budget                  6.899e-01  2.324e-02  29.680 < 2e-16 ***
## countryUSA              1.377e+07  1.221e+06   11.277 < 2e-16 ***
## content_ratingPG        8.738e+06  2.450e+06   3.567 0.000366 ***
## content_ratingPG-13    2.810e+06  2.214e+06   1.269 0.204412
## content_ratingR         -5.876e+05  2.108e+06  -0.279 0.780448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29060000 on 3383 degrees of freedom
## Multiple R-squared:  0.275, Adjusted R-squared:  0.274
## F-statistic: 256.7 on 5 and 3383 DF,  p-value: < 2.2e-16

```

Niti dodavanje content_rating-a nam značajno ne poboljšava model.

```

table(imdbcrr['aspect_ratio'])

##
## 1.33 1.37 1.66 1.75 1.77 1.78 1.85      2  2.2 2.35 2.39  2.4 2.55 2.76  16
## 17   46   38    2    1   35 1473      1   10 1674    10    3    1    3    1

```

Zamijenimo sve vrijednosti aspect_ratio osim 1.85 i 2.35 s OtherRatio.

```

imdbar = imdbcrr
imdbar$aspect_ratio[imdbar$aspect_ratio!="1.85" & imdbar$aspect_ratio!="2.35"]<- "OtherRatio"
table(imdbar['aspect_ratio'])

```

```

##
##      1.85      2.35 OtherRatio
##      1473     1674      168
fit.multi.ar = lm(gross ~ budget + country + aspect_ratio, imdbar)
summary(fit.multi.ar)

```

```

##
## Call:
## lm(formula = gross ~ budget + country + aspect_ratio, data = imdbar)
##
## Residuals:
##
```

```

##      Min       1Q     Median      3Q      Max
## -70292532 -19140613  -6962620   12364040 128636247
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.013e+06  1.317e+06   3.806 0.000144 ***
## budget                7.190e-01  2.396e-02  30.005 < 2e-16 ***
## countryUSA            1.392e+07  1.249e+06  11.146 < 2e-16 ***
## aspect_ratio2.35     -2.297e+06  1.088e+06  -2.110 0.034920 *
## aspect_ratioOtherRatio -2.098e+05  2.396e+06  -0.088 0.930233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29370000 on 3310 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.2617, Adjusted R-squared:  0.2608
## F-statistic: 293.3 on 4 and 3310 DF,  p-value: < 2.2e-16

```

Niti ovaj model nije bolji od modela koji samo sadrži budget i country.

Sada ćemo još pokušati transformirati neke varijable i približiti distribuciju reziduala normalnoj distribuciji.

```
library("SciViews")
```

```

## Warning: package 'SciViews' was built under R version 4.1.2
fit.multi.exp = lm(gross ~ budget + country + I(budget^(1/3)), imdbcou)
summary(fit.multi.exp)

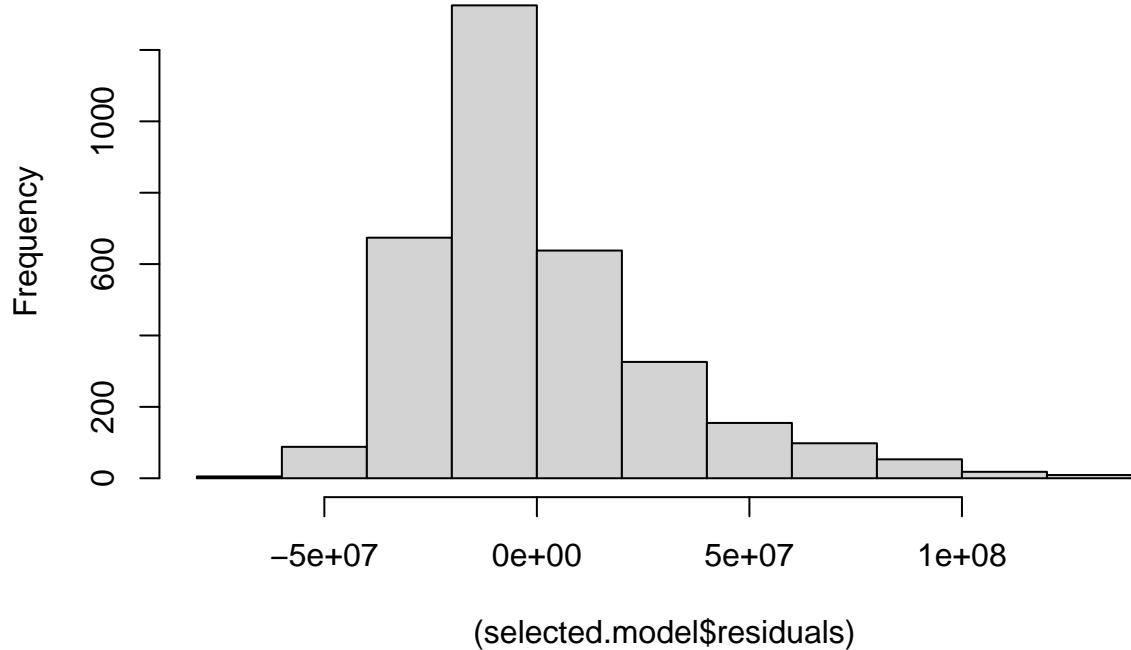
##
## Call:
## lm(formula = gross ~ budget + country + I(budget^(1/3)), data = imdbcou)
##
## Residuals:
##      Min       1Q     Median      3Q      Max
## -63017274 -18580455  -6743889   11286630 134120535
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.131e+07  2.548e+06  -4.439 9.34e-06 ***
## budget                 3.495e-01  6.056e-02   5.771 8.59e-09 ***
## countryUSA            1.412e+07  1.211e+06  11.660 < 2e-16 ***
## I(budget^(1/3))      9.124e+04  1.399e+04   6.523 7.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29030000 on 3385 degrees of freedom
## Multiple R-squared:  0.2762, Adjusted R-squared:  0.2756
## F-statistic: 430.6 on 3 and 3385 DF,  p-value: < 2.2e-16

```

Primijenili smo transformaciju $I(\text{budget}^{(1/3)})$ i time smo malo poboljšali model. Pogledajmo sada distribuciju reziduala.

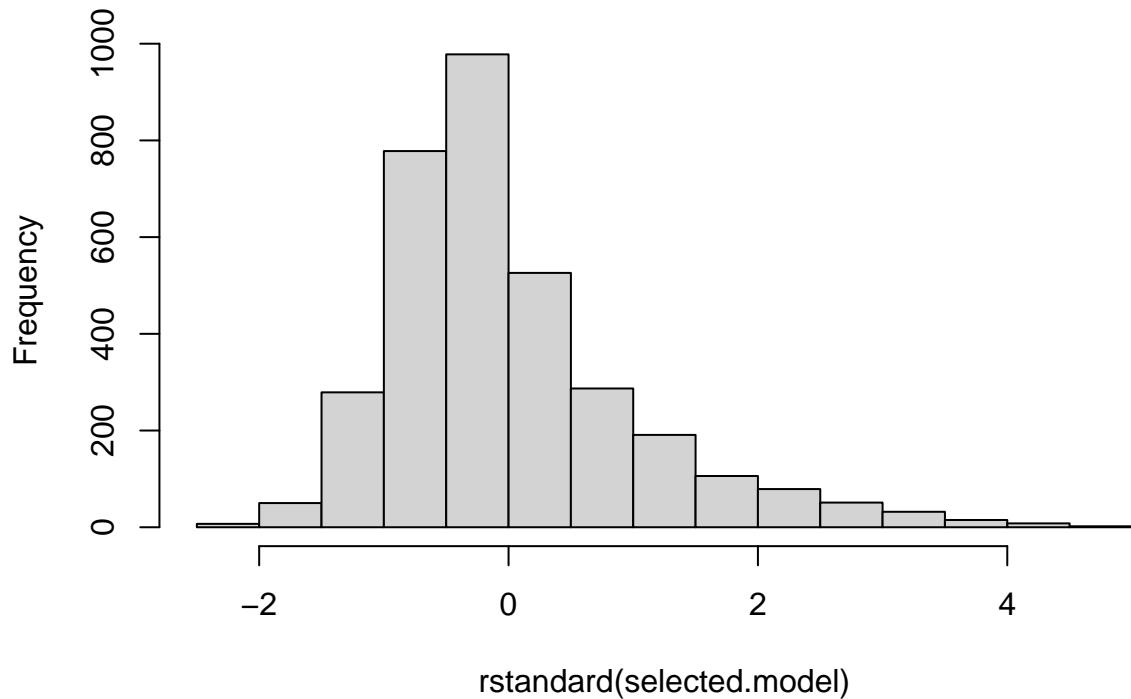
```
selected.model = fit.multi.exp
hist(selected.model$residuals)
```

Histogram of (selected.model\$residuals)



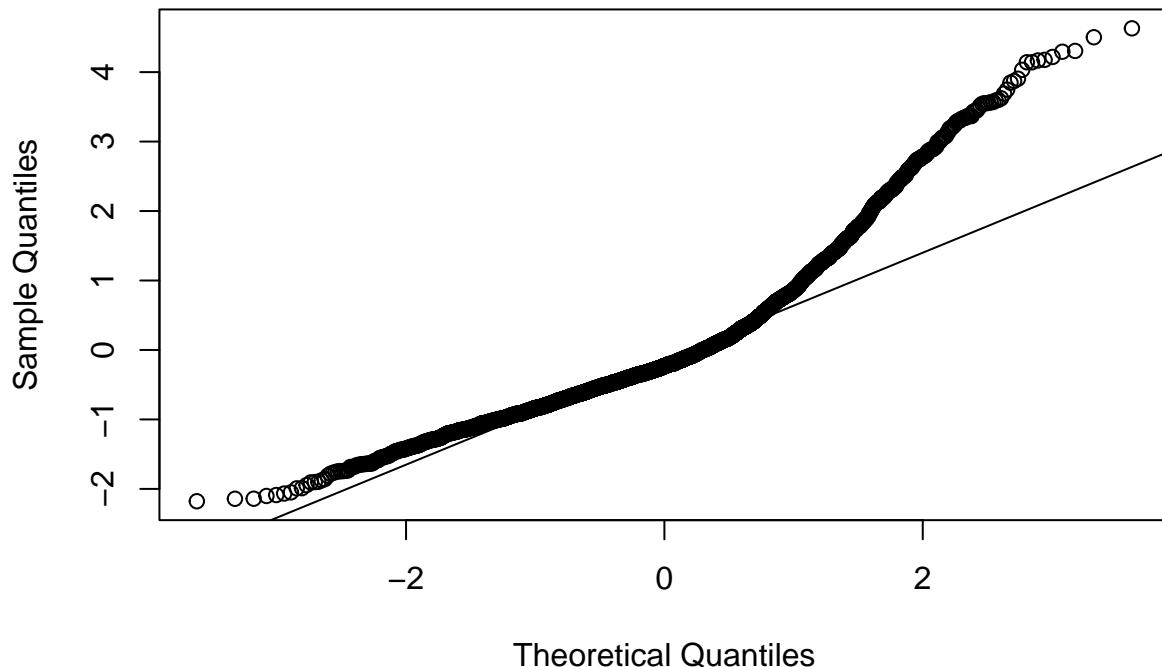
```
hist(rstandard(selected.model))
```

Histogram of rstandard(selected.model)



```
qqnorm(rstandard(selected.model))  
qqline(rstandard(selected.model))
```

Normal Q-Q Plot



Uklonit ćemo još stršećih vrijednosti kako bismo distribuciju reziduala još više približili normalnoj-nrow(imdbar)

```
## [1] 3389
outliers00 <- boxplot(imdbar$gross, plot=FALSE)$out
imdbtry <- imdbar[-which(imdbar$gross %in% outliers00),]

outliers01 <- boxplot(imdbtry$budget, plot=FALSE)$out
imdbtry <- imdbtry[-which(imdbtry$budget %in% outliers01),]

outliers02 <- boxplot(imdbtry$gross, plot=FALSE)$out
imdbtry <- imdbtry[-which(imdbtry$gross %in% outliers02),]

outliers03 <- boxplot(imdbtry$budget, plot=FALSE)$out
imdbtry <- imdbtry[-which(imdbtry$budget %in% outliers03),]

outliers04 <- boxplot(imdbtry$gross, plot=FALSE)$out
imdbtry <- imdbtry[-which(imdbtry$gross %in% outliers04),]

outliers05 <- boxplot(imdbtry$gross, plot=FALSE)$out
imdbtry <- imdbtry[-which(imdbtry$gross %in% outliers05),]

outliers06 <- boxplot(imdbtry$gross, plot=FALSE)$out
imdbtry <- imdbtry[-which(imdbtry$gross %in% outliers06),]

outliers07 <- boxplot(imdbtry$gross, plot=FALSE)$out
```

```

imdbtry <- imdbtry[-which(imdbtry$gross %in% outliers07),]

nrow(imdbtry)

## [1] 3045

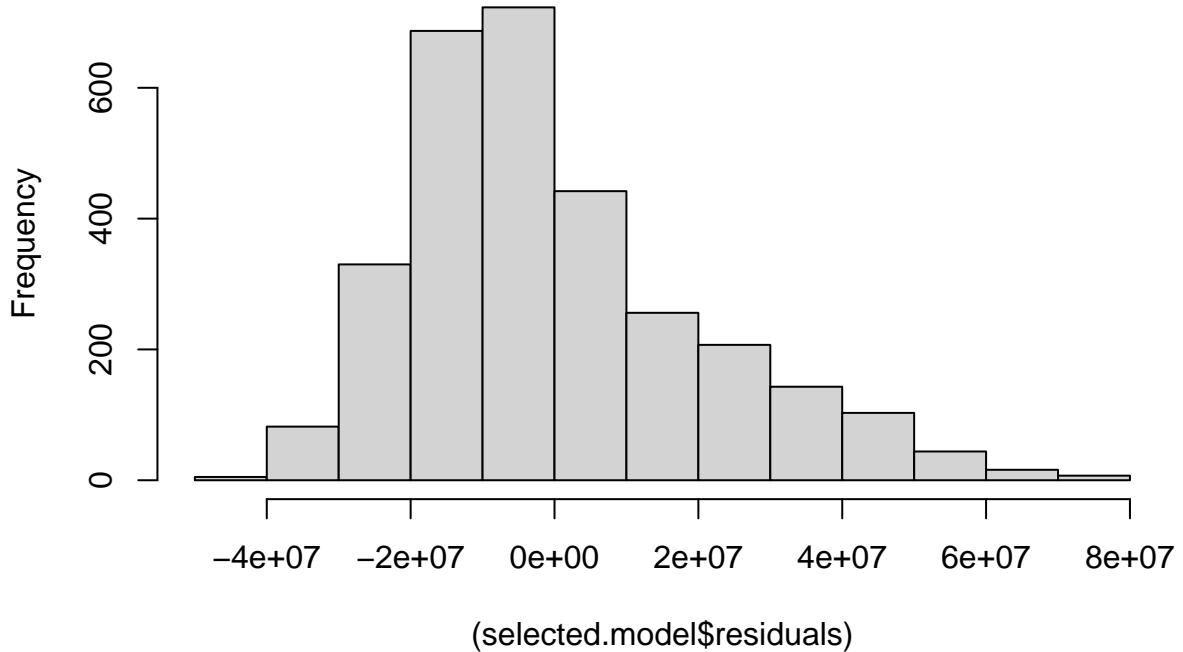
fit.multi.x = lm(gross ~ budget + country + I(budget^(1/3)), imdbtry)
summary(fit.multi.x)

## 
## Call:
## lm(formula = gross ~ budget + country + I(budget^(1/3)), data = imdbtry)
## 
## Residuals:
##       Min     1Q Median     3Q    Max 
## -45942041 -14499390 -4633555 10479796 75746481 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.132e+07  1.930e+06 -5.864 5.01e-09 ***
## budget       1.056e-01  5.477e-02   1.928  0.0539 .  
## countryUSA   9.991e+06  8.827e+05 11.318 < 2e-16 ***
## I(budget^(1/3)) 1.046e+05  1.126e+04   9.289 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 20560000 on 3041 degrees of freedom
## Multiple R-squared:  0.2687, Adjusted R-squared:  0.268 
## F-statistic: 372.5 on 3 and 3041 DF,  p-value: < 2.2e-16

selected.model = fit.multi.x
hist((selected.model$residuals))

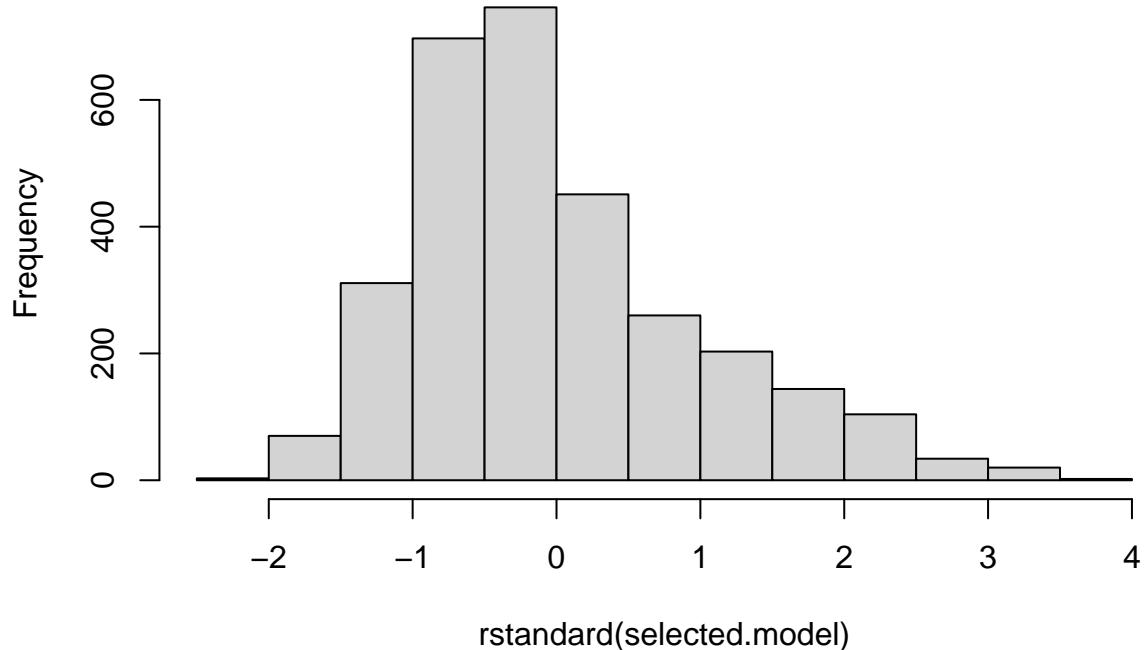
```

Histogram of (selected.model\$residuals)



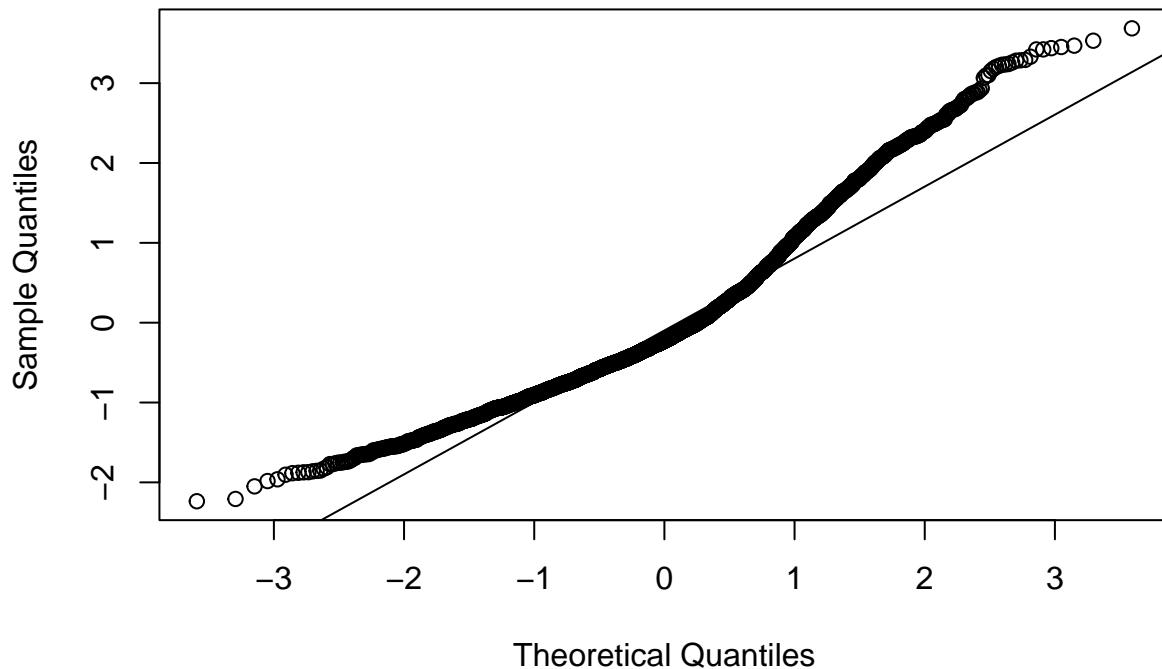
```
hist(rstandard(selected.model))
```

Histogram of rstandard(selected.model)



```
qqnorm(rstandard(selected.model))  
qqline(rstandard(selected.model))
```

Normal Q-Q Plot



Sada je distribucija nešto bliža normalnoj.

Pogledajmo sada naš model.

```
fit.multi.x = lm(gross ~ budget + country + I(budget^(1/3)), imdbtry)
summary(fit.multi.x)
```

```
##
## Call:
## lm(formula = gross ~ budget + country + I(budget^(1/3)), data = imdbtry)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45942041 -14499390  -4633555   10479796  75746481
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.132e+07  1.930e+06  -5.864 5.01e-09 ***
## budget       1.056e-01  5.477e-02   1.928  0.0539 .
## countryUSA  9.991e+06  8.827e+05  11.318 < 2e-16 ***
## I(budget^(1/3)) 1.046e+05  1.126e+04   9.289 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20560000 on 3041 degrees of freedom
## Multiple R-squared:  0.2687, Adjusted R-squared:  0.268
## F-statistic: 372.5 on 3 and 3041 DF,  p-value: < 2.2e-16
```

Na temelju ovog modela pokušat ćemo predvidjeti zaradu novog filma. Naš model sadržava relevantne varijable koje objašnjavaju oko 26% varijance zarade filma. Predviđanjem zarade novog filma vidjet je li tih 26% "dovoljna" vrijednost R².

Pošto je ovo stariji dataset, nekih filmova koji su u međuvremenu izašli nema (zadnja godina je 2016), pa na njima možemo ocijeniti model, ali isprobati ćemo i za neke filmove koji nisu ni sada izašli.

1) No time to die

```
input_var <- data.frame(budget = 300000000, country = "USA")
linear_model = fit.multi.x
predict(linear_model, newdata = input_var, interval = "confidence")

##           fit      lwr      upr
## 1 100388105 78869681 121906530
```

Rezultat koji smo dobili je interval pouzdanosti (95%) i fit predviđena vrijednost. Film No time to die je zaradio oko 500 milijuna dolara, što niti blizu ne pada u naš interval pouzdanosti.

Isprobajmo sad s filmom koji ima manji budget.

2) Nobody

```
input_var <- data.frame(budget = 16000000, country = "USA")
linear_model = fit.multi.x
predict(linear_model, newdata = input_var, interval = "confidence")

##           fit      lwr      upr
## 1 26724329 25651067 27797591
```

Nobody je zapravo zaradio oko 40 milijuna, što opet nije blizu našoj procjeni.

Pokušat ćemo sada s filmom koji nije američki.

3) The Battle at Lake Changjin

Ovo je kineski film s dosta velikim budžetom.

```
input_var <- data.frame(budget = 200000000, country = "OtherCountry")
linear_model = fit.multi.x
predict(linear_model, newdata = input_var, interval = "confidence")

##           fit      lwr      upr
## 1 70982818 58255569 83710067
```

Prava zarada filma je oko 700-800 milijuna, naš model je opet puno pogriješio.

Sada ćemo probati još probati predvidjeti zaradu nekih filmova koji još nisu izašli.

4) The Batman

```
input_var <- data.frame(budget = 100000000, country = "USA")
linear_model = fit.multi.x
predict(linear_model, newdata = input_var, interval = "confidence")

##           fit      lwr      upr
## 1 57791252 53391152 62191351
```

5) Spiderman: Across the Spiderverse

```
input_var <- data.frame(budget = 90000000, country = "USA")
linear_model = fit.multi.x
predict(linear_model, newdata = input_var, interval = "confidence")
```

```
##      fit      lwr      upr
## 1 55059427 51376146 58742708
```

Zaključujemo da naš model nije dobar za kvalitetnu procjenu zarade novog filma, trebao bi nam model s većim R^2 . Takav bolji model bismo dobili dodavanjem varijabli koje više utječu na zaradu, a te varijable nam često nisu dostupne prije samog izlaska filma ili su prekomplikirane za napraviti model s njima (npr glumci u filmu sigurno imaju veliki utjecaj na zaradu ali pošto ima jako puno glumaca, teže je napraviti model). Osim toga treba uzeti u obzir da na zaradu utječu i druge varijable osim ovih koje su prikazane u ovom datasetu.