Eksploratorna analiza - IMDB_movie_dataset

Kristo Palic

2023-01-21

Učitavanje podataka

Pozicioniranje u radni repozitorij

```
root_dir <- setProjectWorkingDirectory()
```

Učitavanje obavljamo pomoću read.csv funkcije

```
data_file <- file.path(root_dir, "data", "IMDB_movie_dataset.txt")
data <- read.csv(data_file)
```

Podatke spremamo u globalni spremnik kako bi joj svi ostali dijelovi projekta mogli pristupiti.

```
save(data, file = "data.RData")
```

Prilagodba podataka

Podatke ucitavamo iz data.RData globalnog spremnika.

```
load("data.RData")
```

5043 filmova, svaki sa 28 atributa 5043 x 28 data.frame Ispisani medijan, srednja vrijednost, kvartali, minimum i maximum za numeričke varijable kao i broj NA atributa u određenom stupcu.

```
pander(summary(data))
```

Table 1: Table continues below

color	director_name	num_critic_for_reviews	duration
Length:5043	Length:5043	Min. : 1.0	Min. : 7.0
Class :character	Class :character	1st Qu.: 50.0	1st Qu.: 93.0
Mode :character	Mode :character	Median :110.0	Median :103.0
NA	NA	Mean :140.2	Mean :107.2
NA	NA	3rd Qu.:195.0	3rd Qu.:118.0
NA	NA	Max. :813.0	Max. :511.0
NA	NA	NA's :50	NA's :15

Table 2: Table continues below

director_facebook_likes	actor_3_facebook_likes	actor_2_name
Min. : 0.0	Min. : 0.0	Length:5043
1st Qu.: 7.0	1st Qu.: 133.0	Class :character
Median : 49.0	Median : 371.5	Mode :character
Mean : 686.5	Mean : 645.0	NA
3rd Qu.: 194.5	3rd Qu.: 636.0	NA
Max. :23000.0	Max. :23000.0	NA
NA's :104	NA's :23	NA

Table 3: Table continues below

actor_1_facebook_likes	gross	genres
Min. : 0	Min. : 162	Length:5043
1st Qu.: 614	1st Qu.: 5340988	Class :character
Median : 988	Median : 25517500	Mode :character
Mean : 6560	Mean : 48468408	NA
3rd Qu.: 11000	3rd Qu.: 62309438	NA
Max. :640000	Max. :760505847	NA
NA's :7	NA's :884	NA

Table 4: Table continues below

actor_1_name	movie_title	num_voted_users
Length:5043	Length:5043	Min. : 5
Class :character	Class :character	1st Qu.: 8594
Mode :character	Mode :character	Median : 34359
NA	NA	Mean : 83668
NA	NA	3rd Qu.: 96309
NA	NA	Max. :1689764
NA	NA	NA

Table 5: Table continues below

cast_total_facebook_likes	actor_3_name	facenumber_in_poster
Min. : 0	Length:5043	Min. : 0.000
1st Qu.: 1411	Class :character	1st Qu.: 0.000
Median : 3090	Mode :character	Median : 1.000
Mean : 9699	NA	Mean : 1.371
3rd Qu.: 13756	NA	3rd Qu.: 2.000
Max. :656730	NA	Max. :43.000
NA	NA	NA's :13

Table 6: Table continues below

plot_keywords	movie_imdb_link	num_user_for_reviews	language
Length:5043	Length:5043	Min. : 1.0	Length:5043
Class :character	Class :character	1st Qu.: 65.0	Class :character
Mode :character	Mode :character	Median : 156.0	Mode :character
NA	NA	Mean : 272.8	NA
NA	NA	3rd Qu.: 326.0	NA
NA	NA	Max. :5060.0	NA
NA	NA	NA's :21	NA

Table 7: Table continues below

country	content_rating	budget	title_year
Length:5043	Length:5043	Min. :2.180e+02	Min. :1916
Class :character	Class :character	1st Qu.:6.000e+06	1st Qu.:1999
Mode :character	Mode :character	Median :2.000e+07	Median :2005
NA	NA	Mean :3.975e+07	Mean :2002
NA	NA	3rd Qu.:4.500e+07	3rd Qu.:2011
NA	NA	Max. :1.222e+10	Max. :2016
NA	NA	NA's :492	NA's :108

actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes
Min. : 0	Min. :1.600	Min. : 1.18	Min. : 0
1st Qu.: 281	1st Qu.:5.800	1st Qu.: 1.85	1st Qu.: 0
Median : 595	Median :6.600	Median : 2.35	Median : 166
Mean : 1652	Mean :6.442	Mean : 2.22	Mean : 7526
3rd Qu.: 918	3rd Qu.:7.200	3rd Qu.: 2.35	3rd Qu.: 3000
Max. :137000	Max. :9.500	Max. :16.00	Max. :349000
NA's :13	NA	NA's :329	NA

Detaljna struktura varijabli unutar podatkovnog skupa.

```
str(data, width = 85, strict.width = "cut")
```

```
## ’data.frame’:    5043 obs. of  28 variables:
##  $ color                    : chr  "Color" "Color" "Color" "Color" ...
##  $ director_name            : chr  "James Cameron" "Gore Verbinski" "Sam Mendes" ""..
##  $ num_critic_for_reviews   : int  723 302 602 813 NA 462 392 324 635 375 ...
##  $ duration                 : int  178 169 148 164 NA 132 156 100 141 153 ...
##  $ director_facebook_likes  : int  0 563 0 22000 131 475 0 15 0 282 ...
##  $ actor_3_facebook_likes   : int  855 1000 161 23000 NA 530 4000 284 19000 10000 ...
##  $ actor_2_name             : chr  "Joel David Moore" "Orlando Bloom" "Rory Kinnea"..
##  $ actor_1_facebook_likes   : int  1000 40000 11000 27000 131 640 24000 799 26000 2..
##  $ gross                    : int  760505847 309404152 200074175 448130642 NA 73058..
##  $ genres                   : chr  "Action|Adventure|Fantasy|Sci-Fi" "Action|Adven"..
##  $ actor_1_name             : chr  "CCH Pounder" "Johnny Depp" "Christoph Waltz" ""..
##  $ movie_title              : chr  "Avatar " "Pirates of the Caribbean: At World’s"..
##  $ num_voted_users          : int  886204 471220 275868 1144337 8 212204 383056 294..
##  $ cast_total_facebook_likes: int  4834 48350 11700 106759 143 1873 46055 2036 9200..
##  $ actor_3_name             : chr  "Wes Studi" "Jack Davenport" "Stephanie Sigman""..
##  $ facenumber_in_poster     : int  0 0 1 0 0 1 0 1 4 3 ...
##  $ plot_keywords            : chr  "avatar|future|marine|native|paraplegic" "godde"..
##  $ movie_imdb_link          : chr  "http://www.imdb.com/title/tt0499549/?ref_=fn_t"..
##  $ num_user_for_reviews     : int  3054 1238 994 2701 NA 738 1902 387 1117 973 ...
##  $ language                 : chr  "English" "English" "English" "English" ...
##  $ country                  : chr  "USA" "USA" "UK" "USA" ...
##  $ content_rating           : chr  "PG-13" "PG-13" "PG-13" "PG-13" ...
##  $ budget                   : num  2.37e+08 3.00e+08 2.45e+08 2.50e+08 NA ...
##  $ title_year               : int  2009 2007 2015 2012 NA 2012 2007 2010 2015 2009 ..
##  $ actor_2_facebook_likes   : int  936 5000 393 23000 12 632 11000 553 21000 11000 ..
##  $ imdb_score               : num  7.9 7.1 6.8 8.5 7.1 6.6 6.2 7.8 7.5 7.5 ...
##  $ aspect_ratio             : num  1.78 2.35 2.35 2.35 NA 2.35 2.35 1.85 2.35 2.35 ..
##  $ movie_facebook_likes     : int  33000 0 85000 164000 0 24000 0 29000 118000 1000..
```

```
missing_values <- sapply(data, function(x) sum(is.na(x)))
print(missing_values, width = 100)
```

```
##                     color             director_name    num_critic_for_reviews
##                         0                         0                        50
##                  duration   director_facebook_likes    actor_3_facebook_likes
##                        15                       104                        23
##              actor_2_name    actor_1_facebook_likes                     gross
##                         0                         7                       884
##                    genres              actor_1_name               movie_title
##                         0                         0                         0
##           num_voted_users cast_total_facebook_likes              actor_3_name
##                         0                         0                         0
##      facenumber_in_poster             plot_keywords           movie_imdb_link
##                        13                         0                         0
##      num_user_for_reviews                  language                   country
##                        21                         0                         0
##            content_rating                    budget                title_year
##                         0                       492                       108
##    actor_2_facebook_likes                imdb_score              aspect_ratio
##                        13                         0                       329
##      movie_facebook_likes
##                         0
```

sapply() funkcija primjenjuje is.na() funkciju na svaki stupac data.framea, a funkcija sum() prebrojava NA vrijednosti svakog stupca. Rezultat funkcije je vektor s brojem NA vrijednosti za svaki stupac.

Izbacujemo duplikate

```
# Identificiramo duplikate na temelju imena filma
duplicate_rows <- duplicated(data, by = "movie_title")

# Izbacujemo duplikate iz originalnog seta podataka
data <- data[!duplicate_rows, ]
save(data, file = "data.RData")
```

Cistimo podatke potrebne za odgovaranje na prvo pitanje

```
## 1. Pitanje
modifiedDataForFirst <- data %>%
    mutate(genres = strsplit(genres, "\\|")) %>%
    tidyr::unnest(genres) %>%
    filter(imdb_score != 0) %>%
    filter(!is.na(imdb_score)) %>%
    filter(genres == "Action" |
           genres == "Comedy" |
           genres == "Drama" |
           genres == "Romance" |
           genres == "Horror" |
           genres == "Thriller" |
           genres == "Animation")

# Na ovaj način rastavili smo filmove koji pripadaju u više od jedne kategorije
# i izbrisali retke koji nemaju ocjenu.

save(modifiedDataForFirst, file = "data.RData")
```

Kreiranje grafova

```
# Potrebno je koristiti drugačiju funkciju za spremanje jer ggsave ima
# neobjašnjivi problem s histogramima

hist(action$imdb_score,
     breaks=30,
     main="Histogram of imdb_score",
     xlab="Scores")
dev.copy(png, file = "../figures/report/actionHistogram.png")
dev.off()

hist(as.double(drama$imdb_score),
     breaks=50,
     main='Histogram of imdb scores of Drama movies',
     xlab='Scores')
dev.copy(png, file = "../figures/report/dramaHistogram.png")
dev.off()

hist(as.double(romance$imdb_score),
  breaks=50,
  main='Histogram of imdb scores of Romance movies',
  xlab='Scores')
dev.copy(png, file = "../figures/report/romanceHistogram.png")
dev.off()

hist(as.double(comedy$imdb_score),
  breaks=50,
  main='Histogram of imdb scores of Romance movies',
  xlab='Scores')
dev.copy(png, file = "../figures/report/comedyHistogram.png")
dev.off()

hist(as.double(thriller$imdb_score),
  breaks=50,
  main='Histogram of imdb scores of Thriller movies',
  xlab='Scores')
dev.copy(png, file = "../figures/report/thrillerHistogram.png")
dev.off()

hist(as.double(horror$imdb_score),
  breaks=50,
  main='Histogram of imdb scores of Horror movies',
  xlab='Scores')
dev.copy(png, file = "../figures/report/horrorHistogram.png")
dev.off()

hist(as.double(animation$imdb_score),
  breaks=50,
  main='Histogram of imdb scores of Animation movies',
  xlab='Scores')
dev.copy(png, file = "../figures/report/animationHistogram.png")
dev.off()
```

Kreiranje QQ-plota

```
qqnorm(action$imdb_score, xlab = "Scores",
       main = "QQ plot of imdb scores of ACTION movies")
dev.copy(png, file = "../figures/report/actionQQplot.png")
dev.off()

qqnorm(drama$imdb_score, xlab = "Scores",
       main = "QQ plot of imdb scores of DRAMA movies")
dev.copy(png, file = "../figures/report/dramaQQplot.png")
dev.off()

qqnorm(romance$imdb_score, xlab = "Scores",
       main = "QQ plot of imdb scores of ROMANCE movies")
dev.copy(png, file = "../figures/report/romanceQQplot.png")
dev.off()

qqnorm(comedy$imdb_score, xlab = "Scores",
       main = "QQ plot of imdb scores of COMEDY movies")
dev.copy(png, file = "../figures/report/comedyQQplot.png")
dev.off()

qqnorm(thriller$imdb_score, xlab = "Scores",
       main = "QQ plot of imdb scores of THRILLER movies")
dev.copy(png, file = "../figures/report/thrillerQQplot.png")
dev.off()

qqnorm(horror$imdb_score, xlab = "Scores",
       main = "QQ plot of imdb scores of HORROR movies")
dev.copy(png, file = "../figures/report/horrorQQplot.png")
dev.off()

qqnorm(animation$imdb_score, xlab = "Scores",
       main = "QQ plot of imdb scores of ANIMATION movies")
dev.copy(png, file = "../figures/report/animationQQplot.png")
dev.off()
```

Kreiranje Box-plota

```
boxplot(data$imdb_score, xlab = "imdb scores")
ggsave(path = "../figures/expl/", filename = "imdbScoresBoxPlot.png", device = "png")

boxplot(data$gross, xlab = "gross income")
ggsave(path = "../figures/expl/", filename = "imdbGrossBoxPlot.png", device = "png")

boxplot(data$cast_total_facebook_likes, xlab = "total fb likes")
ggsave(path = "../figures/expl/", filename = "imdbFBLikes.png", device = "png")
```

Po uzoru na projekt iz SAP-a odabrati ću par varijabli i nad njima napraviti statističku analizu i iznijeti zaključke

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

Promatrat ćemo sljedeće žanrove:

Action, Comedy, Drama, Romance, Thriller, Horror, Western, Animation, History i Documentary

```r
genresSplit = unlist(strsplit(data$genres, "\\|"))
print(table(genresSplit), width = 80)
```

```
## genresSplit
##      Action   Adventure   Animation   Biography      Comedy       Crime
##        1143         914         242         292        1862         883
## Documentary       Drama      Family     Fantasy   Film-Noir   Game-Show
##         121        2571         544         604           6           1
##     History      Horror       Music     Musical     Mystery        News
##         205         556         212         132         493           3
##  Reality-TV     Romance      Sci-Fi       Short       Sport    Thriller
##           2        1098         611           5         181        1396
##         War     Western
##         211          94
```

Dijelimo žanrove pojedinih filmova svaki u svoj redak radi lakšeg upravljanja podatcima. Podatke imamo spremljene u varijabli modifiedDataForFirst

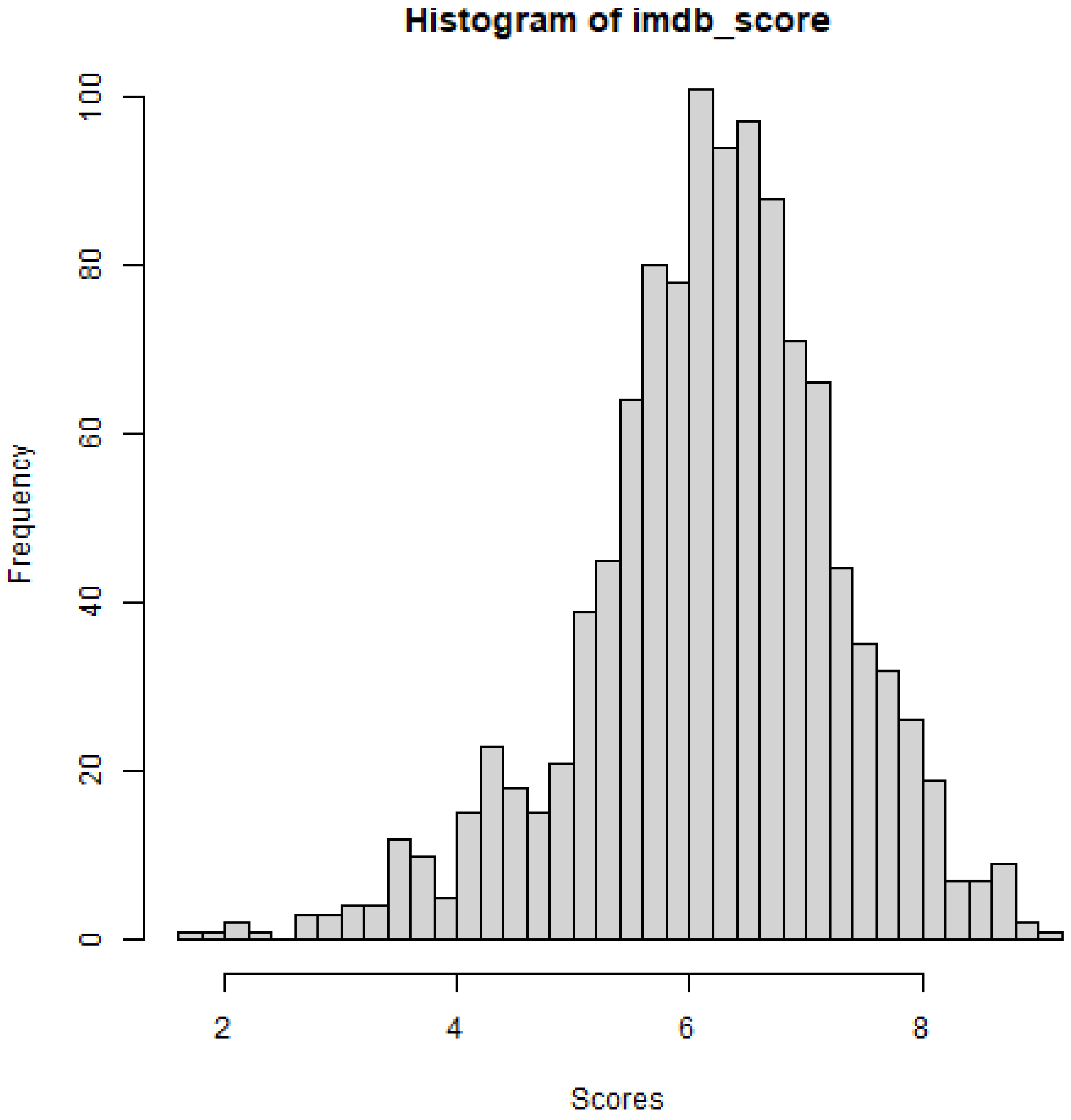
```r
action <- subset(modifiedDataForFirst, genres == "Action")
save(action, file = "data.RData")

comedy <- subset(modifiedDataForFirst, genres == "Comedy")
save(comedy, file = "data.RData")

drama <- subset(modifiedDataForFirst, genres == "Drama")
save(drama, file = "data.RData")

romance <- subset(modifiedDataForFirst, genres == "Romance")
save(romance, file = "data.RData")

thriller <- subset(modifiedDataForFirst, genres == "Thriller")
save(thriller, file = "data.RData")

horror <- subset(modifiedDataForFirst, genres == "Horror")
save(horror, file = "data.RData")

animation <- subset(modifiedDataForFirst, genres == "Animation")
save(animation, file = "data.RData")
```

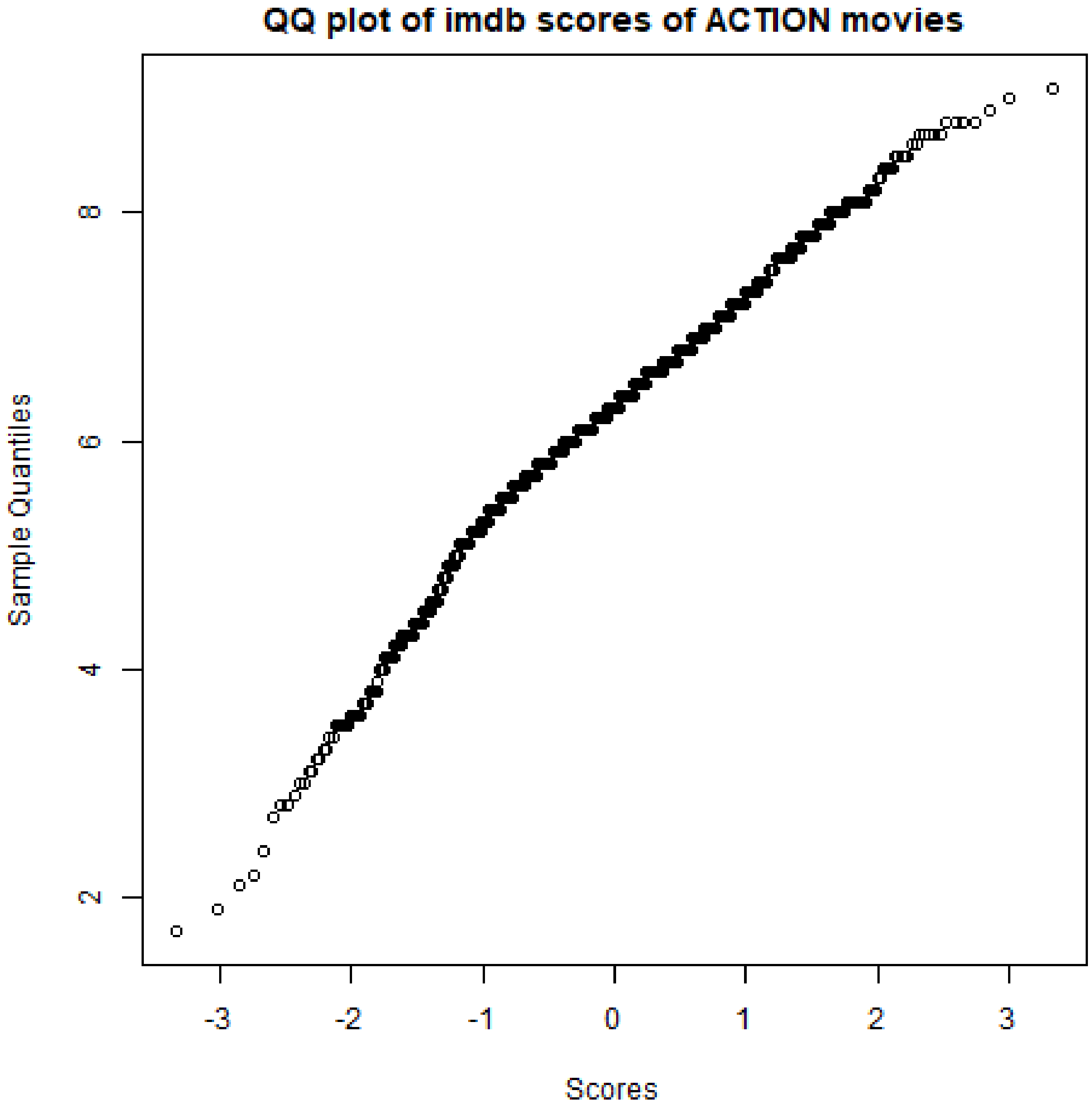
ANOVA

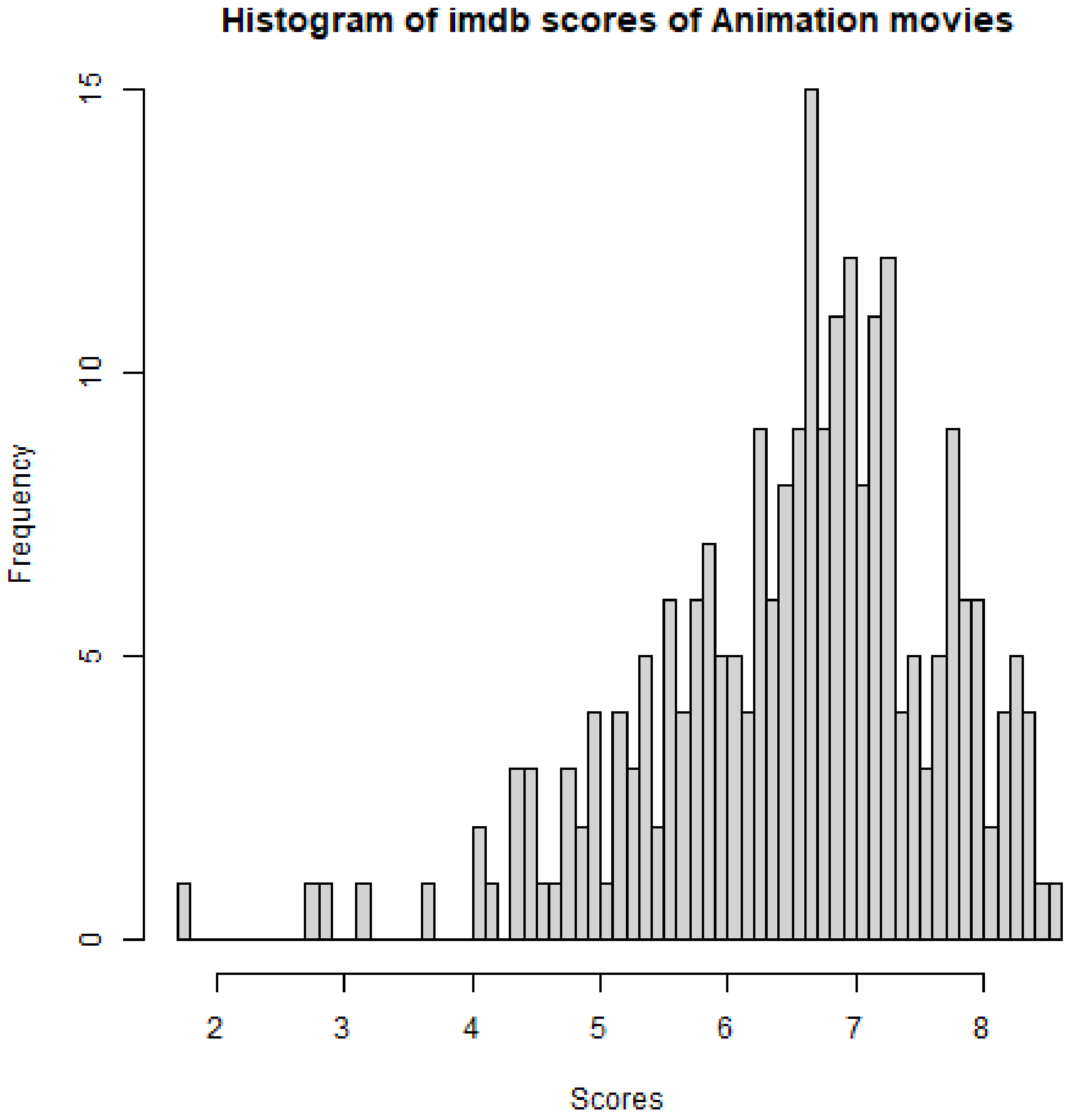
ANOVA (ANalysis Of VAriance) je metoda kojom testiramo sredine više populacija. U analizi varijance pretpostavlja se da je ukupna varijabilnost u podatcima posljedica varijabilnosti podataka unutar svakog pojedine grupe (populacije) i varijabilnosti između različitih grupa. Varijabilnost unutar pojedinog uzorka je rezultat slučajnosti, a ukoliko postoje razlike u srednimana populacija, one će biti odražene u varijabilnosti među grupama. Jedan od glavnih ciljeva analize varijance je ustanoviti jesu li upravo te razlike između grupa samo posljedica slučajnosti ili je statistički značajna.

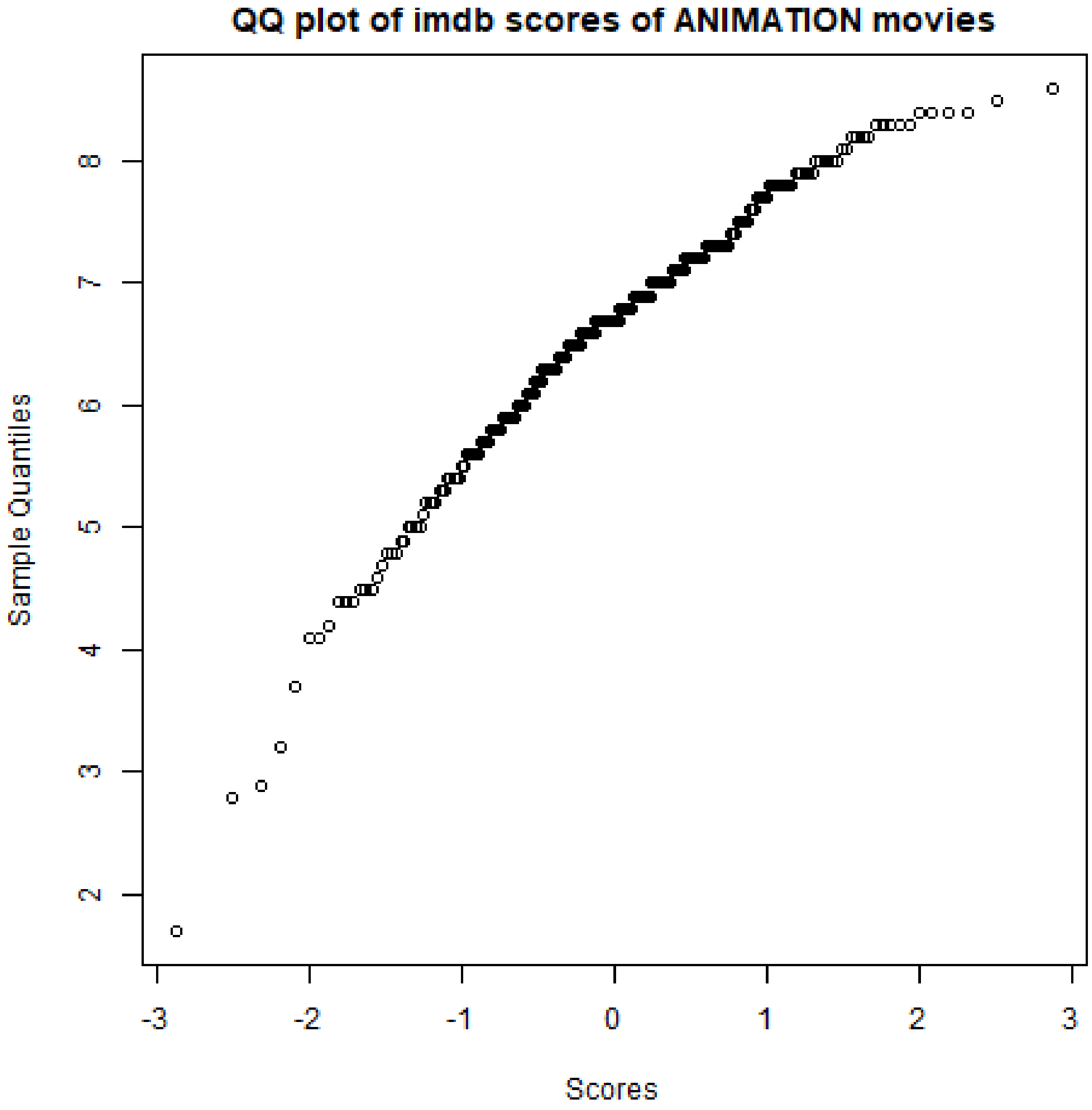
Pretpostavke ANOVA-e su:
1.) nezavisnost pojedinih podataka u uzorcima
2.) normalna razdioba podataka
3.) homogenost varijanci među populacijama.

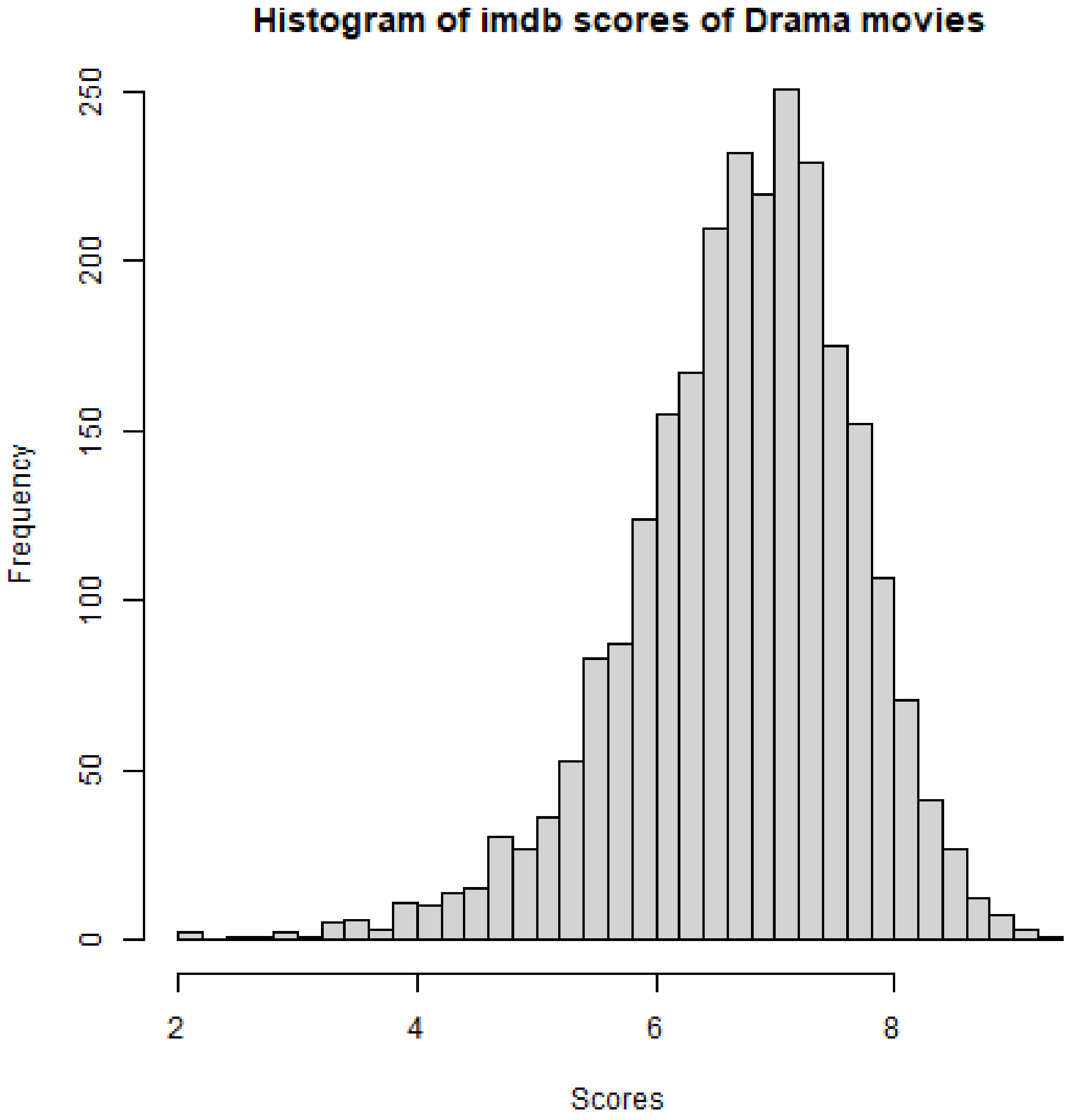
Provjeru normalnosti podataka radit cemo preko histograma i qqplota.

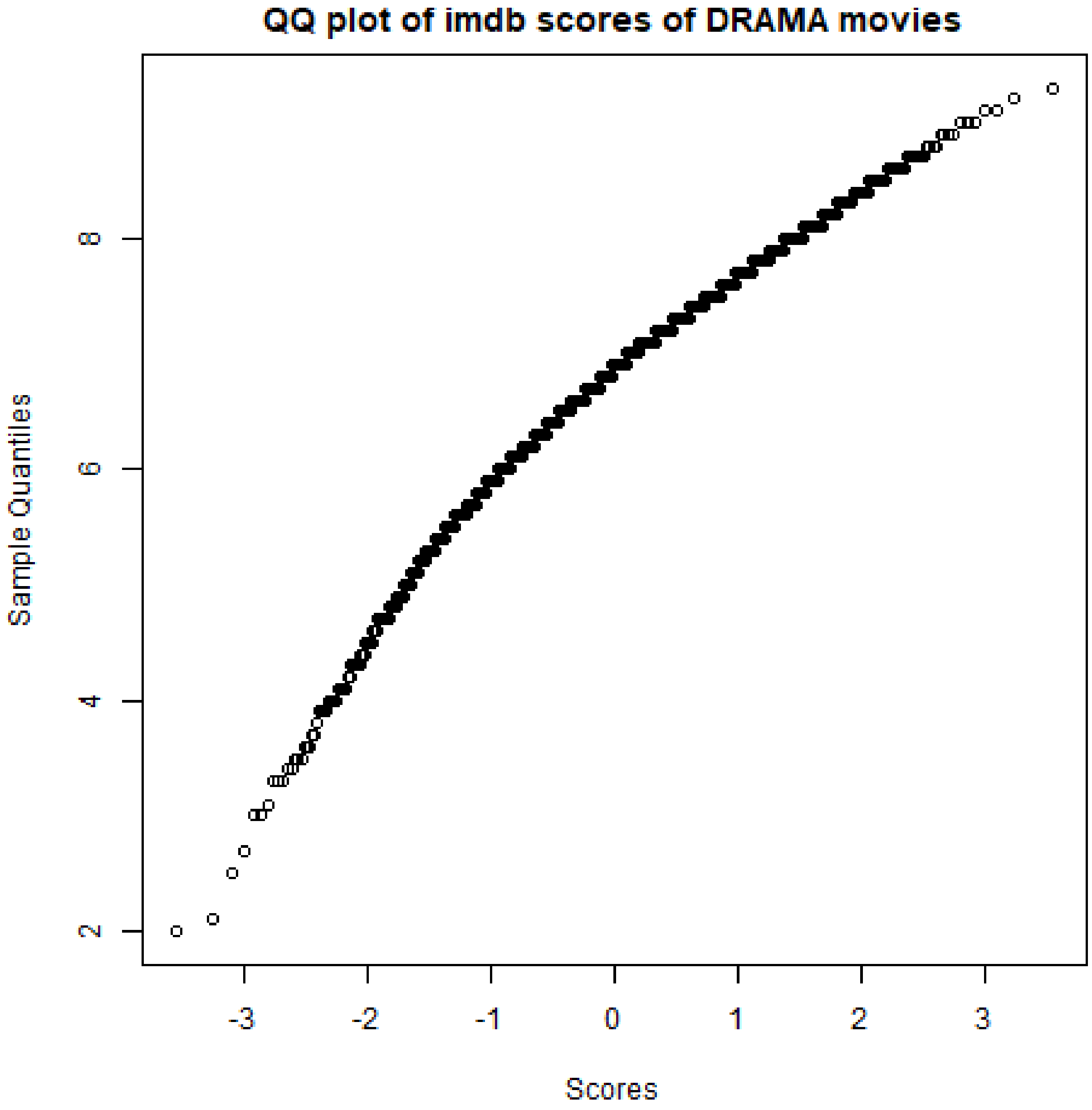

Histogram of imdb_score
Frequency
0
20
40
60
80
100
2
4
6
8
Scores

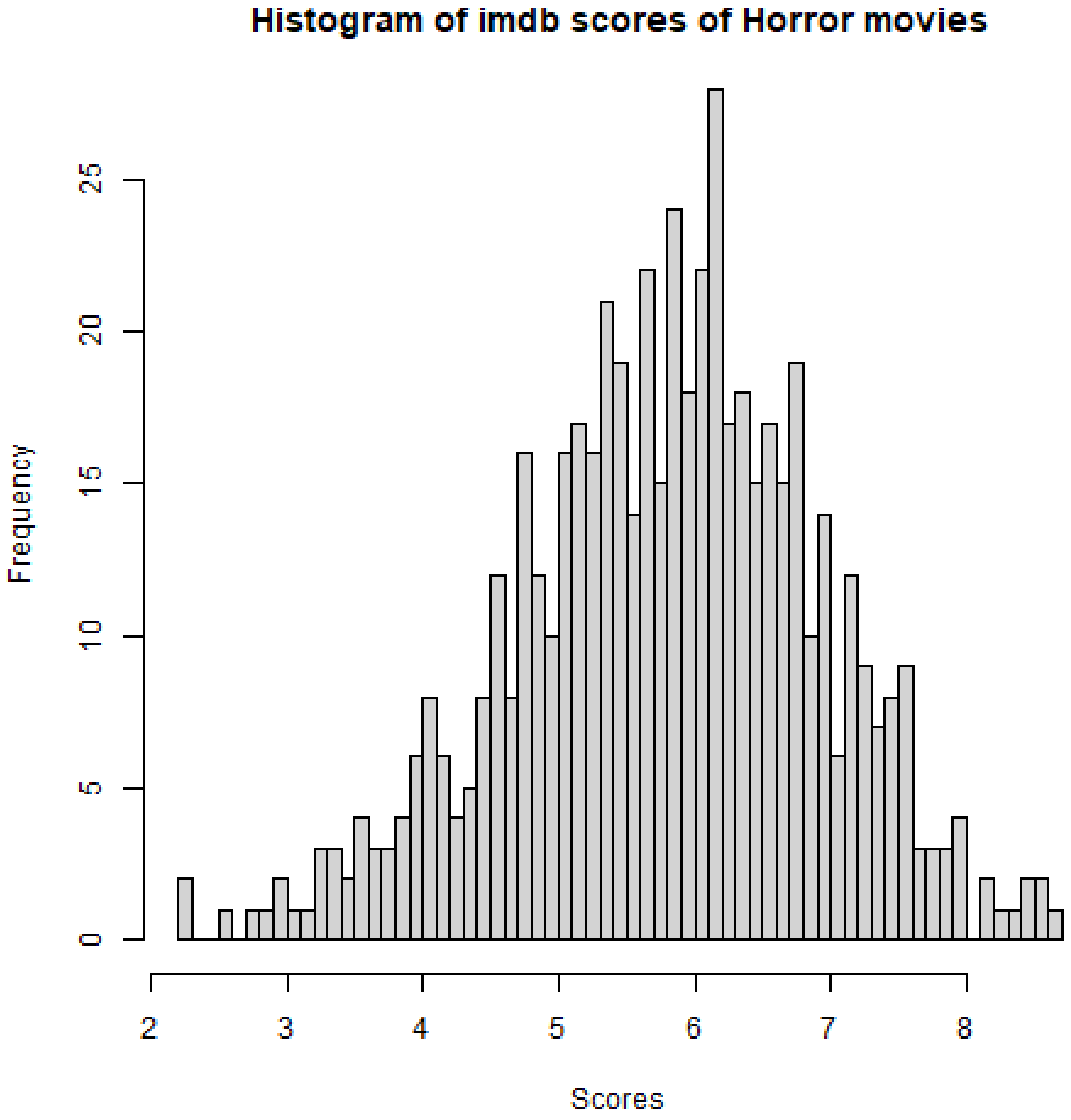

QQ plot of imdb scores of ACTION movies
Sample Quantiles
2
4
6
8
-3
-2
-1
0
1
2
3
Scores

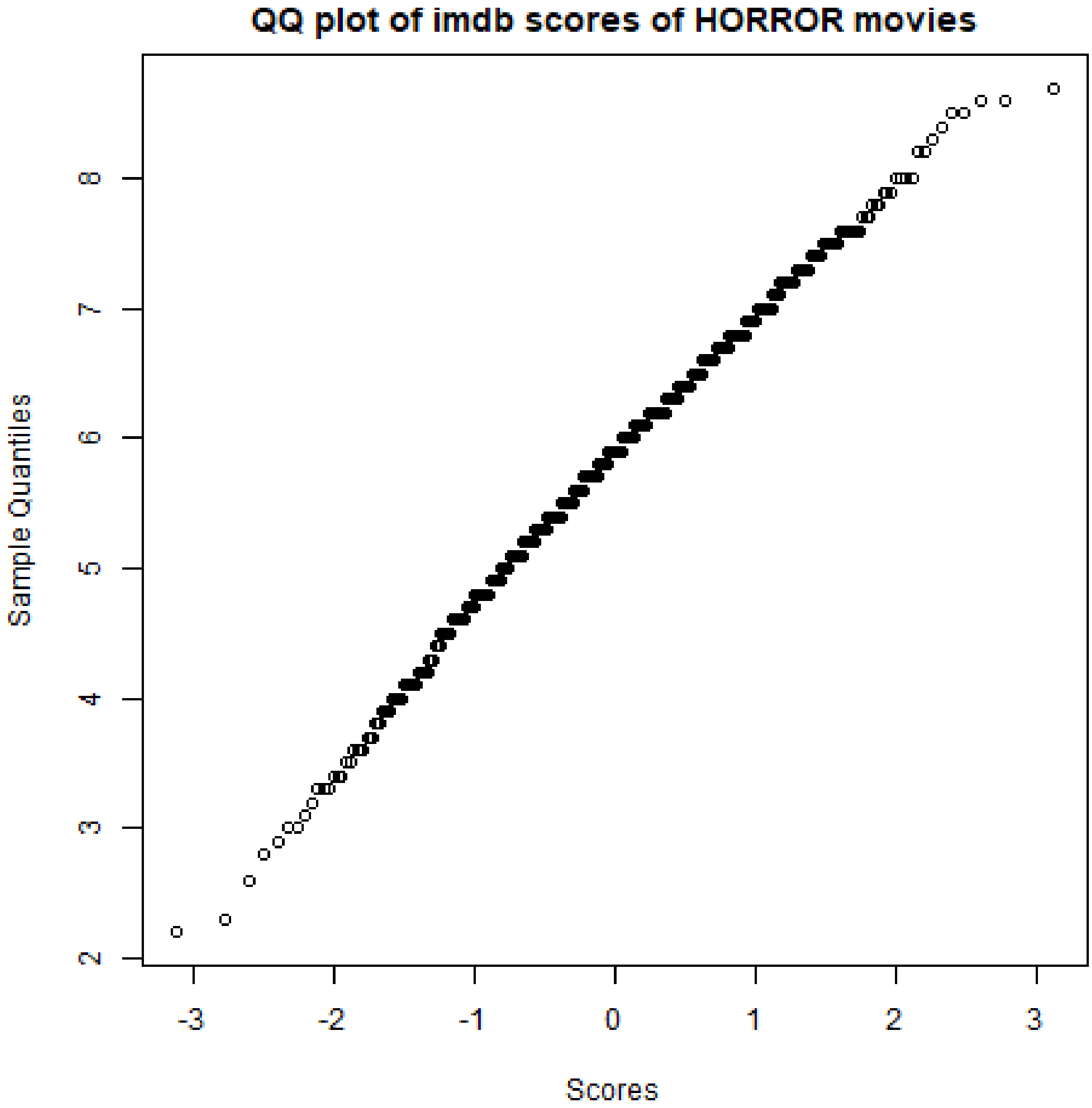

Histogram of imdb scores of Animation movies
Frequency
0
5
10
15
Scores
2
3
4
5
6
7
8

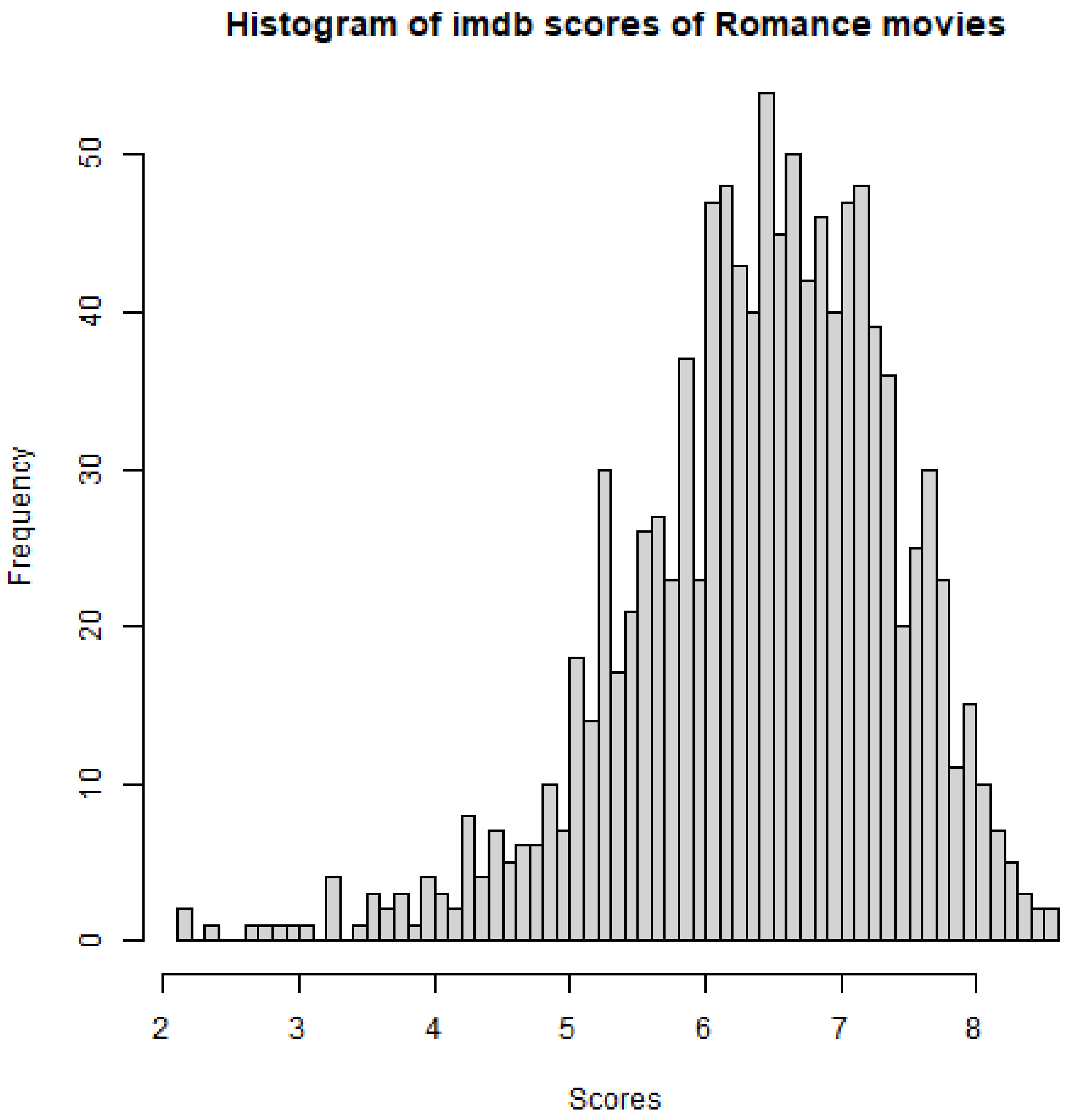

QQ plot of imdb scores of ANIMATION movies
Sample Quantiles
2
3
4
5
6
7
8
Scores
-3
-2
-1
0
1
2
3

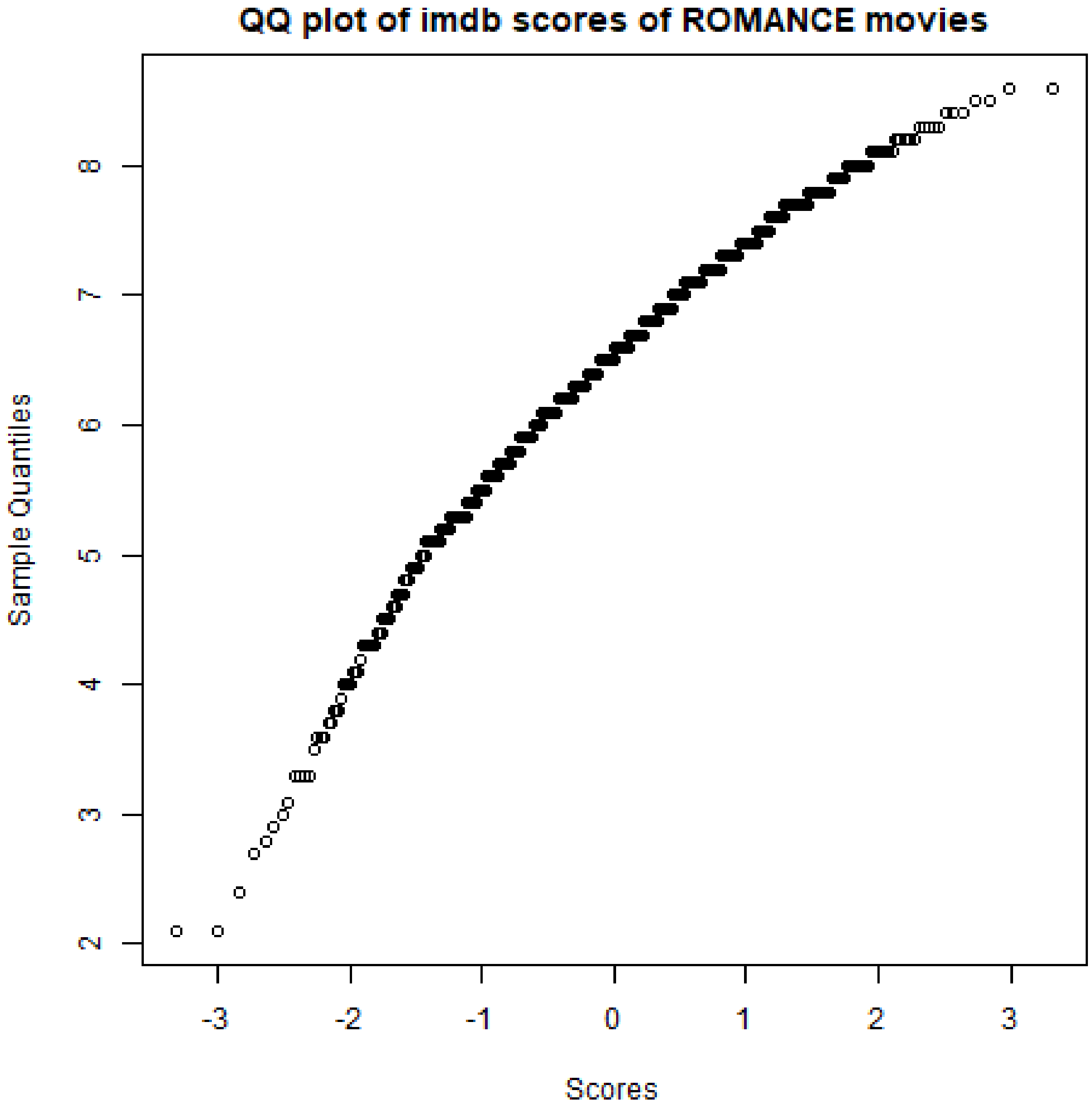

Histogram of imdb scores of Drama movies
Frequency
0
50
100
150
200
250
2
4
6
8
Scores

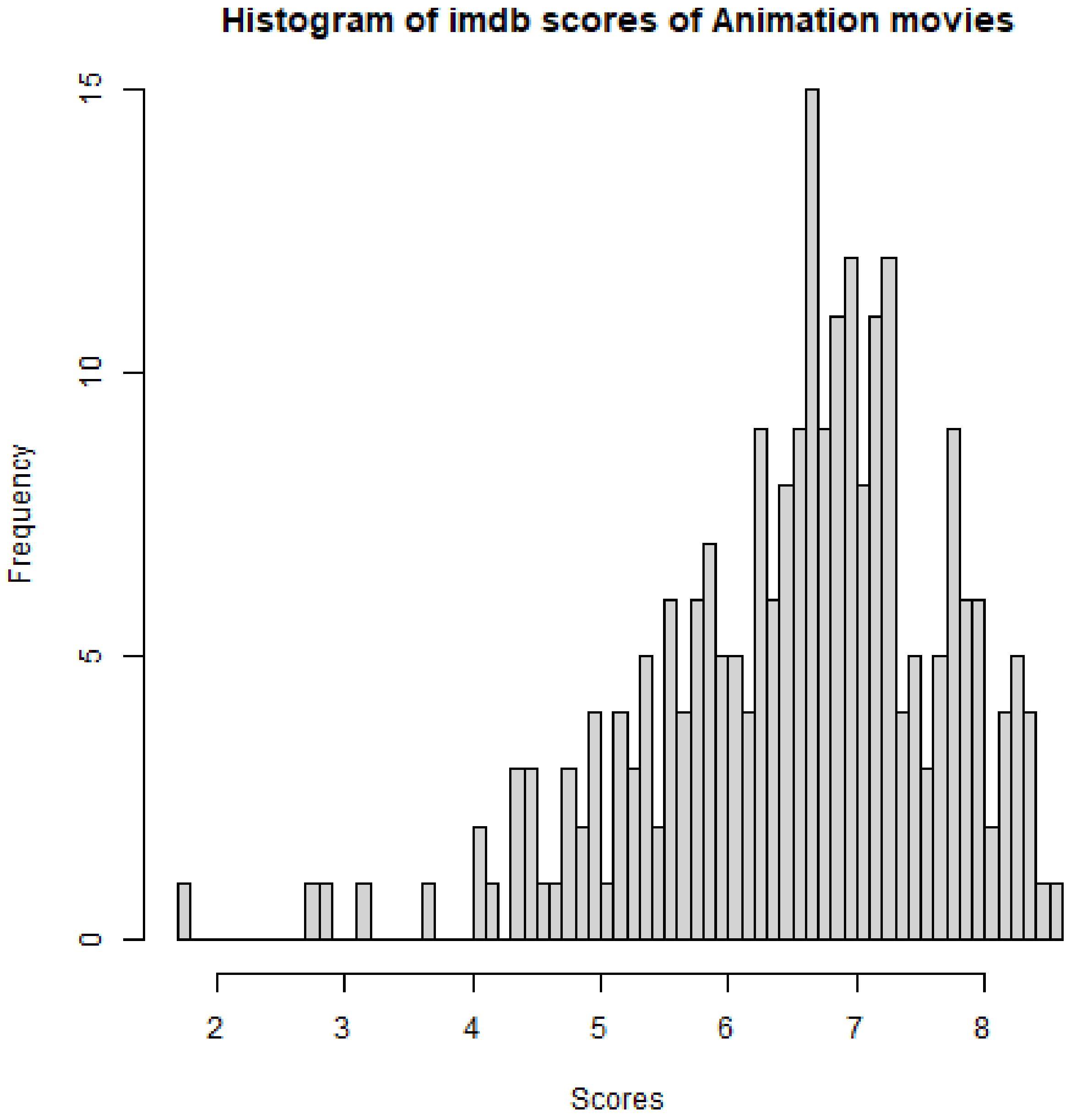

QQ plot of imdb scores of DRAMA movies
Sample Quantiles
2
4
6
8
-3
-2
-1
0
1
2
3
Scores

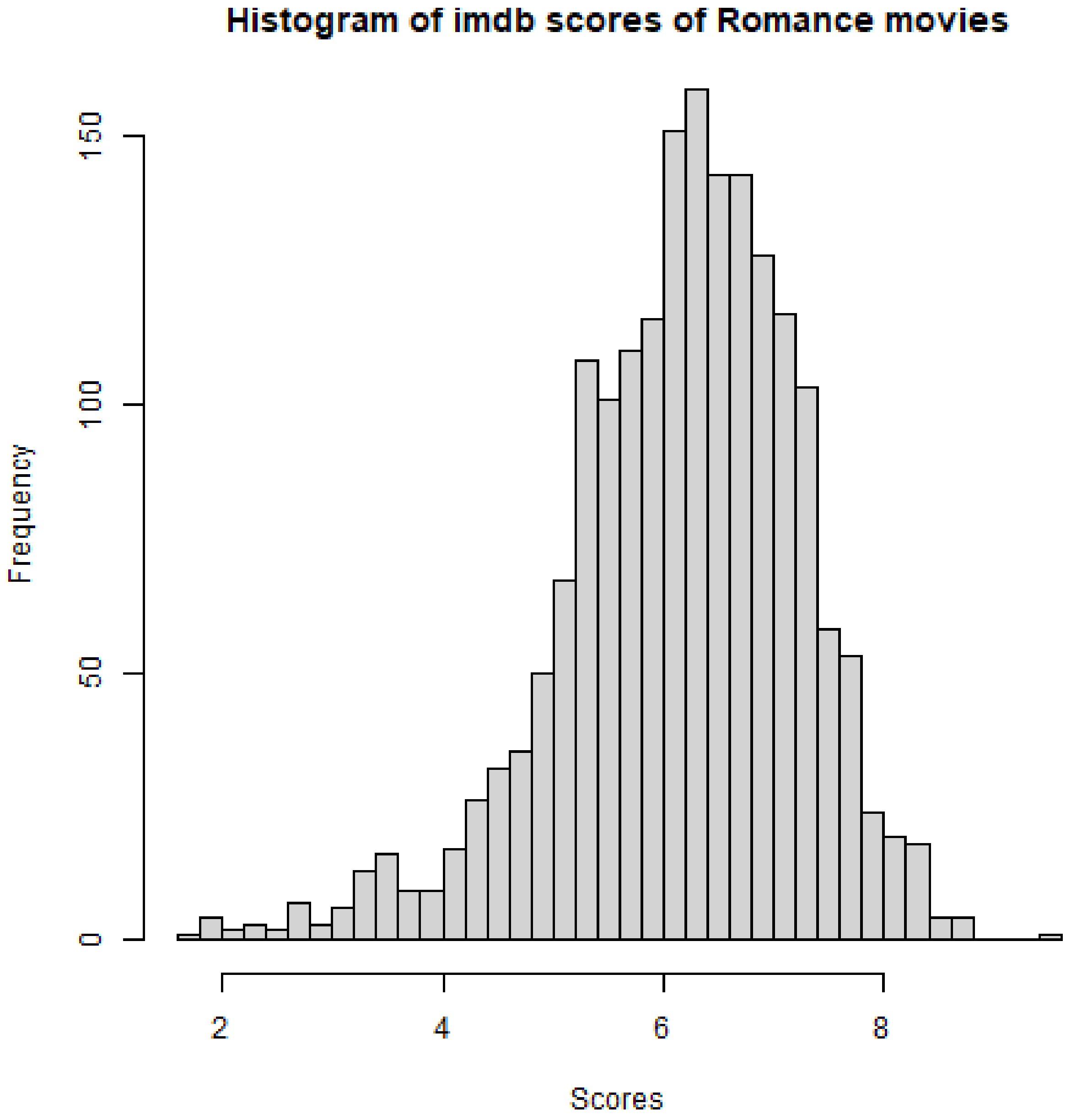

Histogram of imdb scores of Horror movies
Frequency
0
5
10
15
20
25
2
3
4
5
6
7
8
Scores


QQ plot of imdb scores of HORROR movies
Sample Quantiles
2
3
4
5
6
7
8
-3
-2
-1
0
1
2
3
Scores


Histogram of imdb scores of Romance movies
Frequency
0
10
20
30
40
50
2
3
4
5
6
7
8
Scores


QQ plot of imdb scores of ROMANCE movies
Sample Quantiles
2
3
4
5
6
7
8
Scores
-3
-2
-1
0
1
2
3


Histogram of imdb scores of Animation movies
Frequency
0
5
10
15
2
3
4
5
6
7
8
Scores


Histogram of imdb scores of Romance movies
Frequency
0
50
100
150
2
4
6
8
Scores

Testiranje homogenosti varijance uzoraka radili bi Bartletovim testom kad bi imali uzorke jednakih veličina. Umjesto njega, koristit ćemo Levenov test koji ne pretpostavlja jednaku veličinu uzoraka.

```
leveneTest(y = c(action$imdb_score, comedy$imdb_score,
                 drama$imdb_score, romance$imdb_score,
                 thriller$imdb_score, horror$imdb_score,
                 animation$imdb_score),
           group = factor(c(rep("action", length(action$imdb_score)),
                            rep("drama", length(drama$imdb_score)),
                            rep("comedy", length(comedy$imdb_score)),
                            rep("thriller", length(thriller$imdb_score)),
                            rep("animation", length(animation$imdb_score)),
                            rep("romance", length(romance$imdb_score)),
                            rep("horror", length(horror$imdb_score)))),
           center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##         Df F value    Pr(>F)
## group    6  12.993 1.101e-14 ***
##       8861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ne možemo koristiti ANOVU jer pretpostavka jednakosti varijanci nije zadovoljena. Normalnost distribucija bi mogli provjeriti neparametarskim testovima poput Kolmogorljev-Smirnovljevog testa i LillieForce inačice. Međutim zbog nejednakosti varijanci moramo koristiti neparametarski Kruskal-Wallis H test pa nam je normalnost nebitna i nećemo je dalje testirati. Kruskal-Wallis H test pretpostavlja da distribucije dolaze iz jednakih distribucija što vidimo iz qq-plotova

```
kruskal.test(modifiedDataForFirst$imdb_score ~ modifiedDataForFirst$genres)
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  modifiedDataForFirst$imdb_score by modifiedDataForFirst$genres
## Kruskal-Wallis chi-squared = 606.56, df = 6, p-value < 2.2e-16
```

Užasno mala p-vrijednost sugerira da postoje značajne razlike u medijanima između imdb_score-ova različitih žanrova, stoga odbacujemo H0 u korist H1.

Postoji značajna razlika u ocjenama filmova koji dolaze iz različitih žanrova.