Analiza Netflixovih filmova

Katarina Gaćina, Lara Grgurić, Ema Moškatelo, Ivan Plazibat

18.1.2024

Za početak je potrebno učitati potrebne biblioteke.

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(stats)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(nortest)
library(anytime)
```

Sljedeći korak je čitanje podataka iz datoteke NetflixOriginals.csv i spremanje u varijablu data.

```
data <- read.csv("NetflixOriginals.csv")
```

Atributi koji se pojavljuju u datoteci su: naslov, žanr, premijera (datum), trajanje (u minutama), IMDB ocjena i jezik.

Zadatak 1. Jesu li neki žanrovi popularniji u pojedinim područjima?

Naš prvi zadatak je ispitati jesu li neki žanrovi popularniji u pojedinim područjima. Pod "područja" misli se na jezična govorna područja stoga zapravo uspoređujemo atribute žanr i jezik.

Prilikom ručne inspekcije primjećujemo da nisu svi podatci u jednakom formatu. Npr. kod žanrova primjećujemo praznine viška zbog kojih može doći do krivoga grupiranja podataka. Stoga na početku moramo napraviti "čišćenje" podataka. Radimo funkciju combine_genres koja točno grupira svaki žanr bez obzira na različite formate zapisa.

```
combine_genres <- function(genre) {
  if (grepl("drama", genre, ignore.case = TRUE)) {
    return("Drama")
  } else if (grepl("comedy", genre, ignore.case = TRUE)) {
    return("Comedy")
  } else if (grepl("romantic", genre, ignore.case = TRUE)) {
    return("Romantic")
  } else if (grepl("science fiction", genre, ignore.case = TRUE)) {
    return("Science Fiction")
  } else if (grepl("horror", genre, ignore.case = TRUE)) {
    return("Horror")
  } else if (grepl("action", genre, ignore.case = TRUE)) {
    return("Action")
  } else if (grepl("christmas", genre, ignore.case = TRUE)) {
    return("Christmas")
  } else if (grepl("adventure", genre, ignore.case = TRUE)) {
    return("Adventure")
  } else if (grepl("family", genre, ignore.case = TRUE)) {
    return("Family")
  } else if (grepl("anime", genre, ignore.case = TRUE)) {
    return("Anime")
  } else if (grepl("heist", genre, ignore.case = TRUE)) {
    return("Heist")
  } else if (grepl("animation", genre, ignore.case = TRUE)) {
    return("Animation")
  } else {
    return("Other")
  }
}
```

Također, primjećujemo da neki filmovi spadaju u dvije kategorije žanrova i dvije kategorije jezika, stoga ćemo dogovorno, za potrebe ove vježbe, uzimati samo prvi žanr odnosno jezik.

```
data$FirstGenre <- sapply(strsplit(as.character(data$Genre), "/"), function(x) combine_genres(x[1]))
data$FirstLanguage <- sapply(strsplit(as.character(data$Language), "/"), function(x) x[1])
```

Sljedeći korak je pobrojavanje koliko je filmova iz kojeg žanra i koliko je filmova na kojem jeziku (frekvencije žanrova i jezika).

```
genre_counts <- data %>%
  group_by(FirstGenre) %>%
  summarize(count = n()) %>%
  filter(FirstGenre != "Other") %>%
  arrange(desc(count))

language_counts <- data %>%
  group_by(FirstLanguage) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
```

Primijetili smo da je puno filmova kategorizirano pod žanr “Other” što bi nam moglo kasnije stvarati probleme pri testiranju hipoteze, a i ne bi imalo smisla analizirati žanr “Other”, stoga smo ga izbacili iz naše analize. Sljedeći korak je izabrati top 5 žanrova i jezika po broju pojavljivanja.

```
top5_genres <- genre_counts$FirstGenre[1:5]
```

```r
top5_languages <- language_counts$FirstLanguage[1:5]
```

Sada ćemo napraviti novi dataset samo sa filtriranim podatcima (najpopularnijim žanrovima i jezicima).

```r
data_filtered <- data %>%
  filter(FirstGenre %in% top5_genres, FirstLanguage %in% top5_languages)
```

Dalje, radimo kontingencijsku tablicu kako bi mogli nad njom napraviti hi-kvadrat test.

```r
contingency_table <- table(data_filtered$FirstGenre, data_filtered$FirstLanguage)
```

Naša početna hipoteza H_0 je da su žanr i jezik nezavisne varijable. Kako bi testirali ovu hipotezu provesti ćemo hi-kvadrat test. Za kritičnu p-vrijednost uzeti ćemo klasičnih 5%.

```r
chi_square_test <- chisq.test(contingency_table)
```

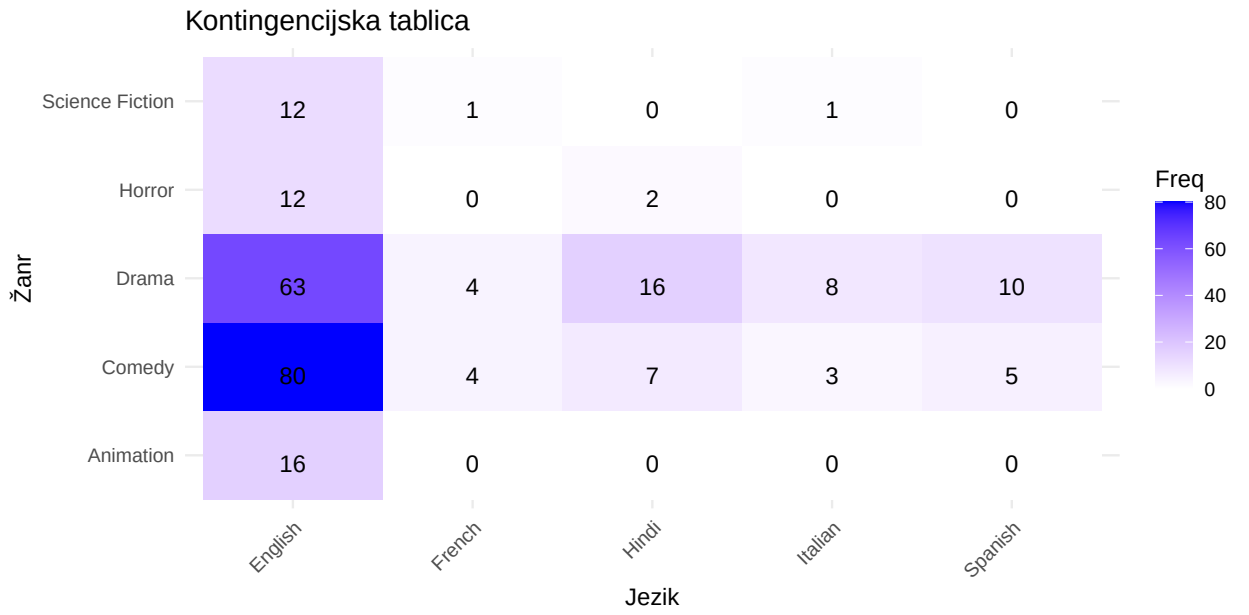
```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

```r
print(chi_square_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 22.722, df = 16, p-value = 0.1214
```

Po rezultatima hi-kvadrat testa zaključujemo da ne možemo odbaciti hipotezu H_0 pa ju prihvaćamo, odnosno zaključujemo da su jezik i žanr dvije nezavisne varijable. Za kraj ćemo još napraviti i vizualizaciju podataka pomoću heatmap-a.

```r
ggplot(data = as.data.frame(contingency_table), aes(x = Var2, y = Var1, fill = Freq)) +
  geom_tile() +
  geom_text(aes(label = Freq), vjust = 1) +
  labs(title = "Kontingencijska tablica",
       x = "Jezik",
       y = "Žanr") +
  theme_minimal() +
  scale_fill_gradient(low = "white", high = "blue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

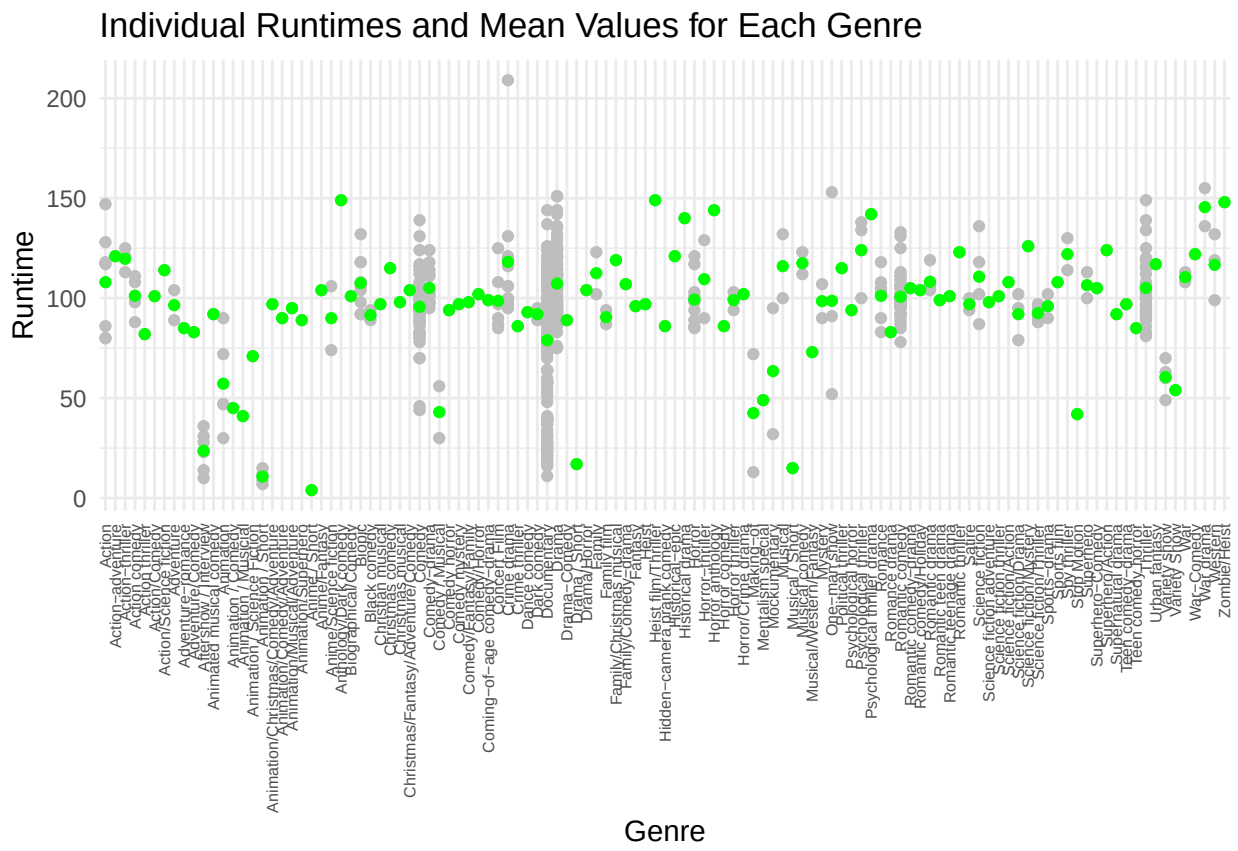


Zadatak 2. Traju li svi žanrovi jednako dugo?

```
netflix_data = read.csv("NetflixOriginals.csv")
netflix_df = data.frame(netflix_data)
```

Za motivaciju smo razmotrili graf koji predstavlja odnos između žanra filma i pojedinačnog trajanja filmova u minutama, kao i srednje vrijednosti trajanja filmova u minutama koji pripadaju tom žanru.

```
mean_runtimes <- aggregate(Runtime ~ Genre, data = netflix_df, mean)

ggplot(netflix_df, aes(x = Genre, y = Runtime)) +
  geom_point(color = "gray") +
  geom_point(data = mean_runtimes, aes(x = Genre, y = Runtime), color = "green") +
  labs(title = "Individual Runtimes and Mean Values for Each Genre", x = "Genre", y = "Runtime") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size = 6))
```



S obzirom na malen broj filmova koji pripadaju pojedinim žanrovima, odlučili smo testirati hipotezu na 5 najpopularnijih žanrova filma.

```
#unique_genres <- unique(netflix_df$Genre)

genre_counts <- netflix_df %>%
  group_by(Genre) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

top_genres <- genre_counts %>%
  head(5)

print(top_genres)
```

```
## # A tibble: 5 x 2
##   Genre           count
##   <chr>           <int>
## 1 Documentary       159
## 2 Drama              77
## 3 Comedy             49
## 4 Romantic comedy    39
## 5 Thriller           33
```

Želimo testirati je li pet najpopularnijih žanrova filma traju jednako dugo.

Nad pet najpopularnijih filmova prvo provodimo Liliforsova inačicu Kolmogorov-Smirnovljevog testa, kako bismo utvrdili dolaze li podaci, tj. duljine trajanja filmova koji pripadaju pojedinom žanru, iz normalne

distribucije. Dakle, nulta hipoteza za Lillie test kaže kako podaci dolaze iz normalne distribucije.

Za p-vrijednosti koje su manje od 0.05 odbacujemo nultu hipotezu. Analizom rezultata vidimo kako su p-vrijednosti za 3 od 5 najpopularnijih žanrova manje od 0.05, te za te podatke odbacujemo nultu hipotezu.

```
for (g in top_genres$Genre) {
  df_genre <- filter(netflix_df, Genre == g)

  cat("\n", g, "\n")
  #hist(df_genre$Runtime, breaks = 20, col = "lightblue", main = g, xlab = "Runtime")

  if (nrow(df_genre) > 4) {
    result <- lillie.test(df_genre$Runtime)
    print(result)

    p_value <- result$p.value
    if (p_value < 0.05) {
      cat("Data is not normally distributed for genre:", g, "\n")
    }

  } else {
    cat("Not enough data to perform Lillie test for genre:", g, "\n")
  }
}
```

```
##
##  Documentary
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_genre$Runtime
## D = 0.17347, p-value = 1.07e-12
##
## Data is not normally distributed for genre: Documentary
##
##  Drama
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_genre$Runtime
## D = 0.11753, p-value = 0.01033
##
## Data is not normally distributed for genre: Drama
##
##  Comedy
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_genre$Runtime
## D = 0.11489, p-value = 0.1089
##
##
##  Romantic comedy
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  df_genre$Runtime
## D = 0.1462, p-value = 0.03494
##
## Data is not normally distributed for genre: Romantic comedy
##
##  Thriller
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_genre$Runtime
## D = 0.12563, p-value = 0.2059
```

Razmatrali smo korištenje t-testa, ali kako radimo usporedbu između više kategorija (žanrova), odlučili smo se za ANOVA test, zato što bi se ponavljanjem t-testova vjerojatnost pogreške tipa I cijelog postupka znatno povećala.

S obzirom na to da smo prethodnom analizom zaključili kako nam žanrovi, tj. kategorije, nemaju normalnu distribuciju, nećemo provesti parametarski ANOVA test, nego neparametarsku alternativu, odnosno Kruskal-Wallisov test, kako bismo testirali nultu hipotezu: medijani distribucija svih uzoraka su jednaki.

```
filtered_df <- netflix_df[netflix_df$Genre %in% top_genres$Genre, ]

kruskal.test(Runtime ~ Genre, data = filtered_df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Runtime by Genre
## Kruskal-Wallis chi-squared = 77.262, df = 4, p-value = 6.62e-16
```

Zaključak: P-vrijednost je 6.62e-16 « 0.05, zbog čega odbacujemo nultu hipotezu. Pet najpopularnijih žanrova ne traju jednako dugo.

Zadatak 3. Kako su filmovi ocijenjeni s obzirom na datum premijere?

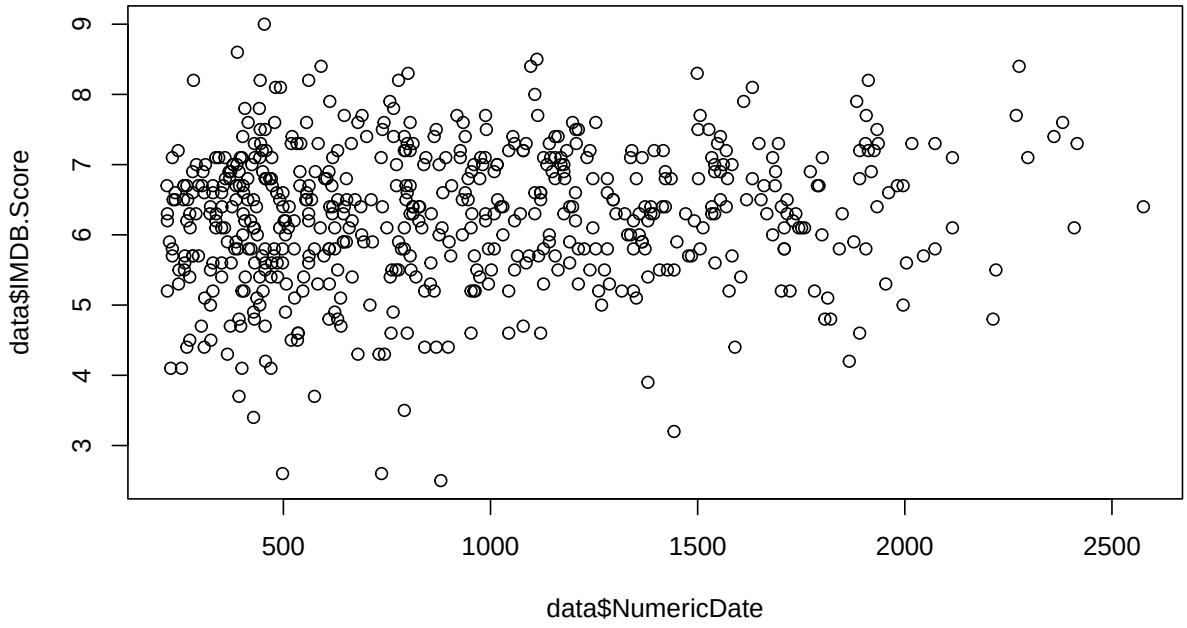
Pretvorba datuma u broj dana od referentnog datuma.

```
data$Premiere2 = anytime(data$Premiere)
data$NumericDate <- as.numeric(difftime(as.Date("2022-01-01"), data$Premiere2, units = "days"))
```

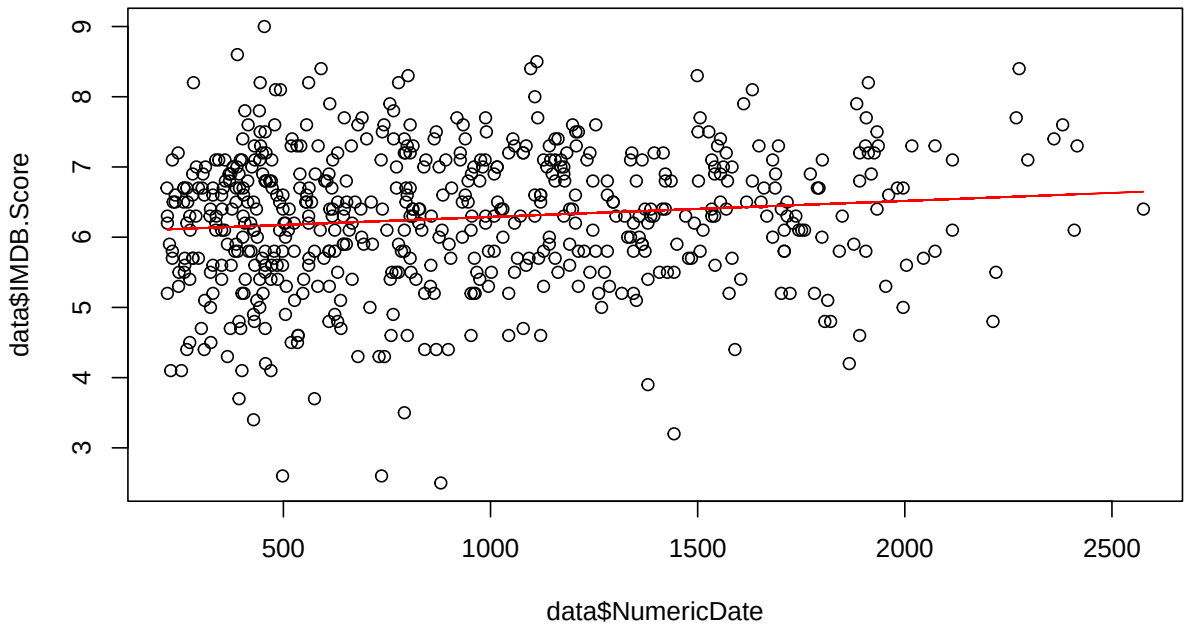
Kako su filmovi ocijenjeni s obzirom na datum premijere?

Moramo ispitati utjecaj datuma premijere na ocjenu filma. Crtanjem grafa koji prikazuje ovisnost ocjene filma o vremenu proteklom od datuma njegove premijere do 1.1.2022. možemo steći dojam o odnosu te dvije varijable. Iz grafa je vidljivo da je utjecaj vremena proteklog od premijere filma na ocjenu filma slab.

```
plot(data$NumericDate, data$IMDB.Score) #grafički prikaz podataka
```



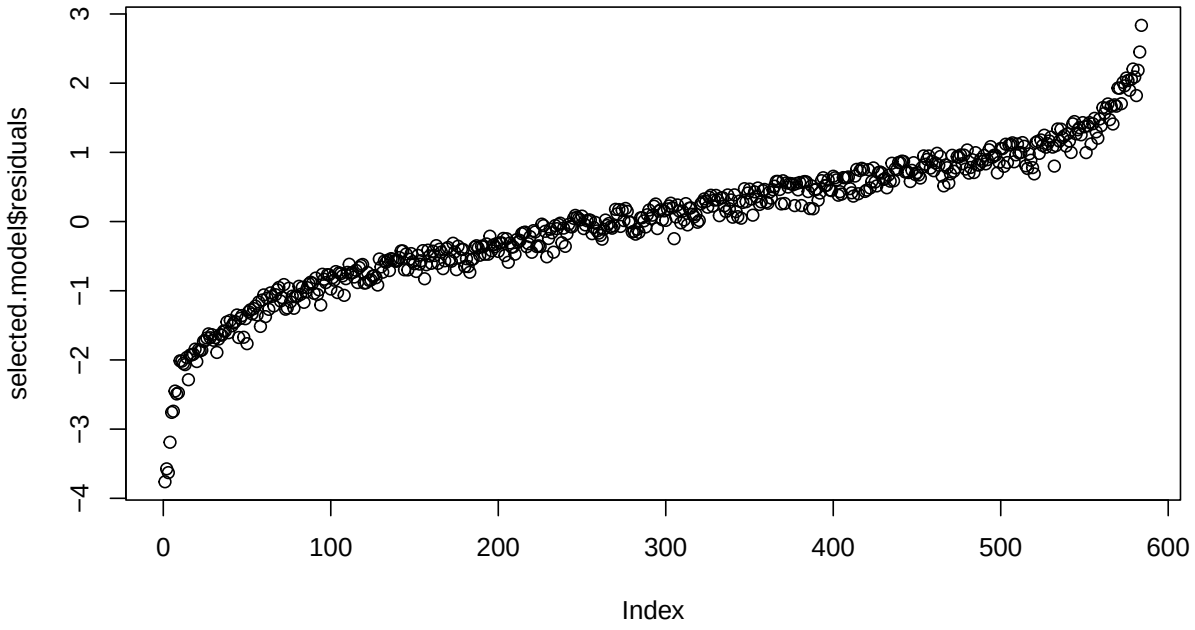
Modeliramo odnos dvije varijable linearnom regresijom. Regresijski model procjenjuje se funkcijom lm() koja kao parametre prima zavisne i nezavisne varijable, odnosno data.frame sa svim varijablama i definiciju varijabli u modelu. Grafički je prikazan pravac malog pozitivnog nagiba.



Potrebno je provjeriti da pretpostavke modela zadovoljene. Provjerimo pretpostavku o rezidualima (normalnost reziduala i homogenost varijance). Normalnost reziduala moguće je provjeriti grafički, pomoću kvantil-kvantil plota (usporedbom s linijom normalne razdiobe), te statistički pomoću Kolmogorov-Smirnovljevog testa.

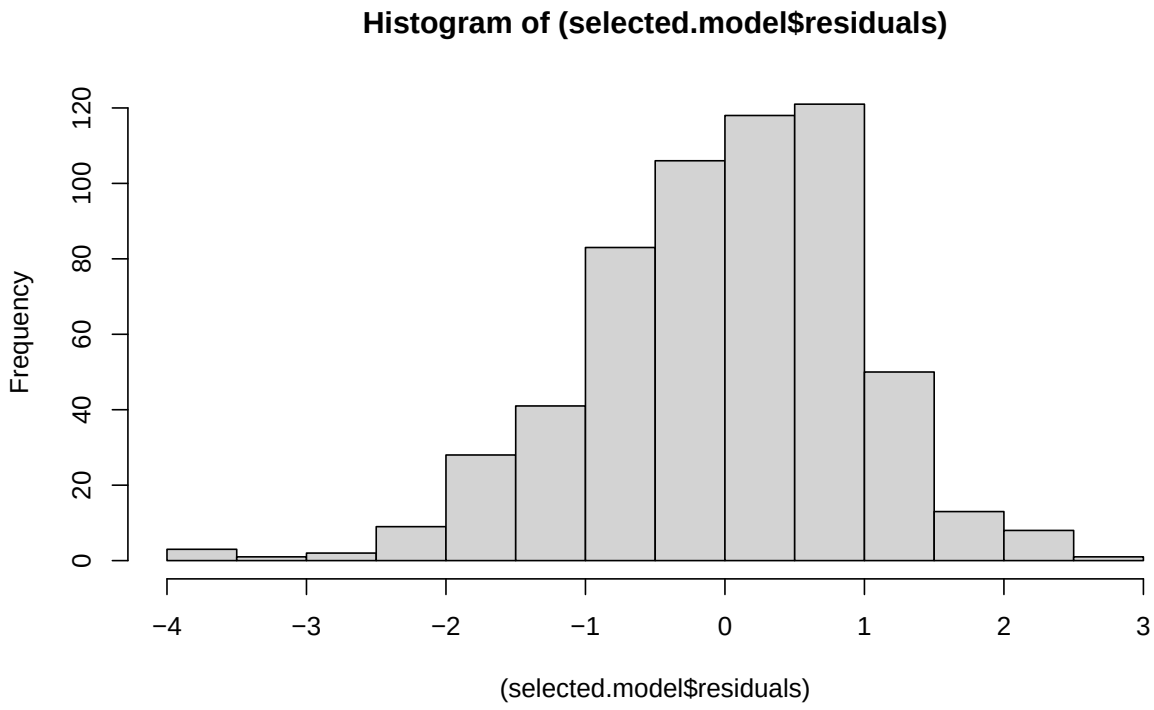
Reziduali rastu s indeksom što nam govori o linearnosti odnosa ocjene i starosti filma.

```
selected.model = fit.age
plot(selected.model$residuals)
```

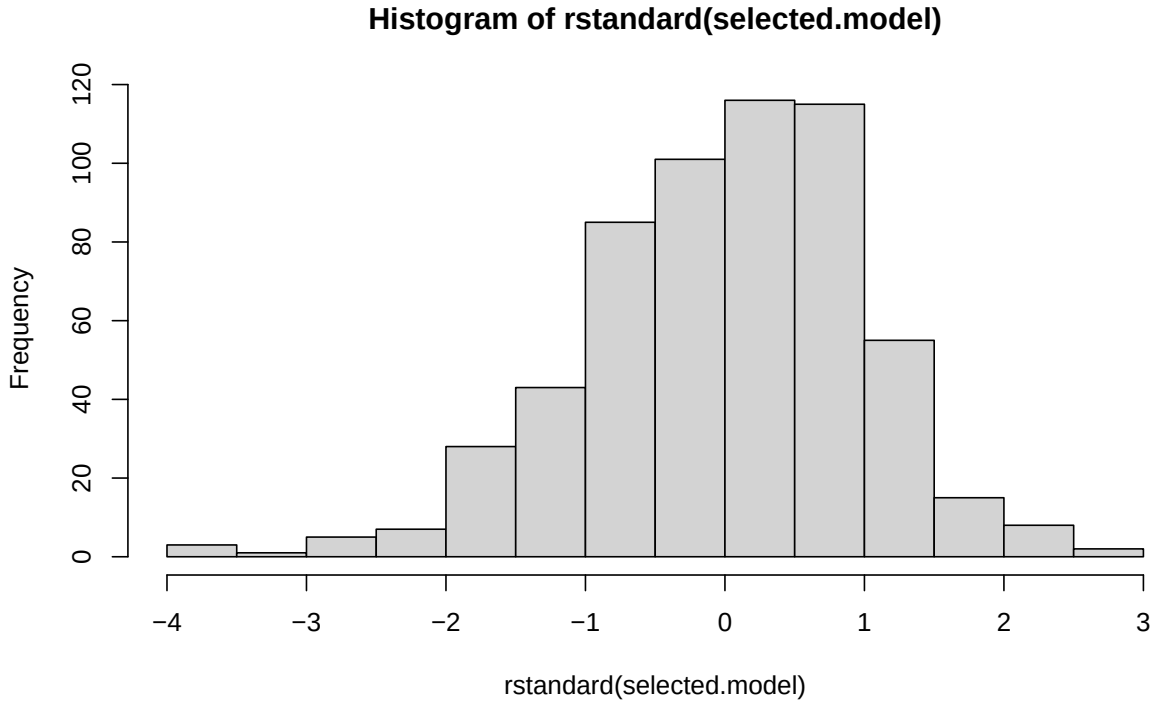


Iz histograma reziduala vidljiva je sličnost normalnoj razdiobi.

```
hist((selected.model$residuals))
```

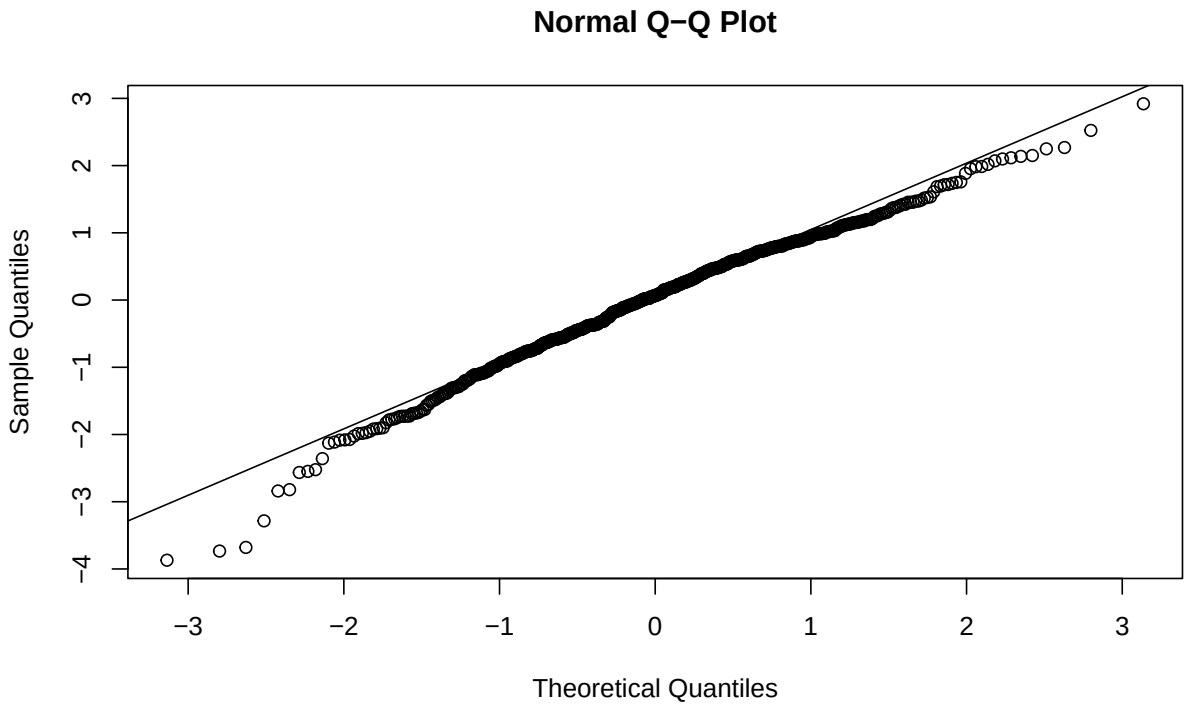



```
hist(rstandard(selected.model))
```

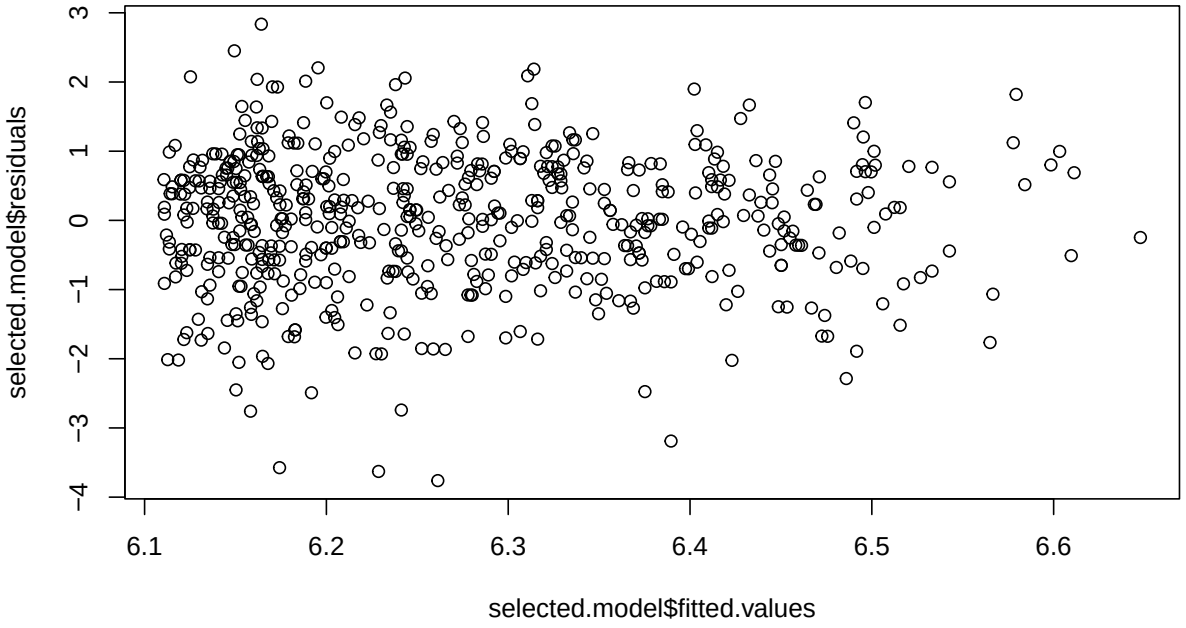


Kvantil-kvantil plot prikazuje blizinu liniji normalne distribucije.

```r
#q-q plot reziduala s linijom normalne distribucije
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```



Reziduali su na grafu ovisnosti o izlaznim vrijednostima modela položeni oko pravca $y = 0$ nasumičnim rasporedom što potvrđuje valjanost modela lienarne regresije.



Kolmogorov-Smirnovljev test je neparametarski test koji služi za provjeru dolaze li podaci iz neke točno određene distribucije. Pogodan je za kontinuirane razdiobe (ne treba podjelu u razrede). Test je aproksimatvan tj. vrijedi za velike uzorke. Lilliefors je njegova inačica pogodna za manje uzorke, no s obzirom na veličinu skupa podataka, uzimamo u obzir rezultat KS testa. S p-vrijednosti od 0.3571, što je dosta više od 0.05, zaključujemo da su reziduali normalno distribuirani.

```
## Warning in ks.test.default(rstandard(fit.age), "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.age)
## D = 0.03834, p-value = 0.3572
## alternative hypothesis: two-sided
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.age)
## D = 0.038295, p-value = 0.04056
```

Izračun Pearsonovog koeficijenta daje koeficijent 0.12217. Nul-hipoteza nalaže da je korelacija jednaka 0. S p-vrijednosti od 0.003104 ju odbacujemo.

```
cor.test(data$NumericDate,data$IMDB.Score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$NumericDate and data$IMDB.Score
## t = 2.9697, df = 582, p-value = 0.003104
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04145122 0.20131335
## sample estimates:
##       cor
## 0.1221746
```

Koeficijente modela ispitujemo t-testom. Njegova nul-hipoteza glasi da je koeficijent jednak 0. Zbog p-vrijednosti od 0.00314, odbacujem nul-hipotezu u korist alternativne hipoteze. Prema tome, koeficijent uz nezavisnu varijablu starosti filma je značajan. R^2 mjera kojom ćemo odrediti koji postotak varijance u izlaznoj varijabli Y je estimirani linearni model objasnio/opisao. Vrijednost R^2 govori nam da starost filma ima relativno malen udio u objašnjavanju ukupne varijance ocjene filma. F-statistikom ćemo ispitati signifikantnost modela. Pretpostavka F-statistike je normalnost populacija te jednakost njihovih standardnih devijacija, dok je hipoteza da su srednje vrijednosti populacija jednake. Ipak, f-testom koji daje p-vrijednost 0.003104 zaključujemo da postoji statistički značajan utjecaj vremena proteklog od premijere filma na njegovu ocjenu.

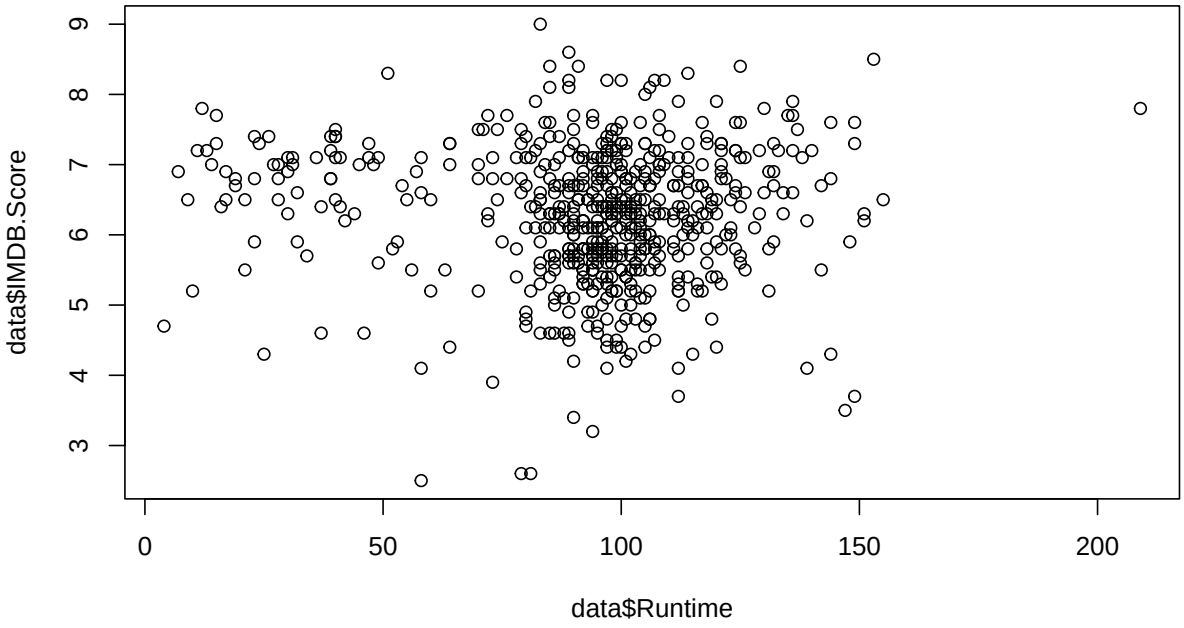
```
summary(fit.age)
```

```
##
## Call:
## lm(formula = IMDB.Score ~ NumericDate, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7613 -0.5885  0.0644  0.7046  2.8358
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.061e+00  8.166e-02   74.22   <2e-16 ***
## NumericDate 2.279e-04  7.673e-05    2.97   0.0031 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9728 on 582 degrees of freedom
## Multiple R-squared:  0.01493,    Adjusted R-squared:  0.01323
## F-statistic: 8.819 on 1 and 582 DF,  p-value: 0.003104
```

Zadatak 4. Jesu li dugi filmovi lošije ocijenjeni?

Proučavamo kako duljina filma (značajka Runtime) utječe na IMDB ocjenu (značajka IMDB.Score). Prikaz vrijednosti istaknutih značajki:

```
plot(data$Runtime,data$IMDB.Score) #runtime vs IMDB ocjena
```

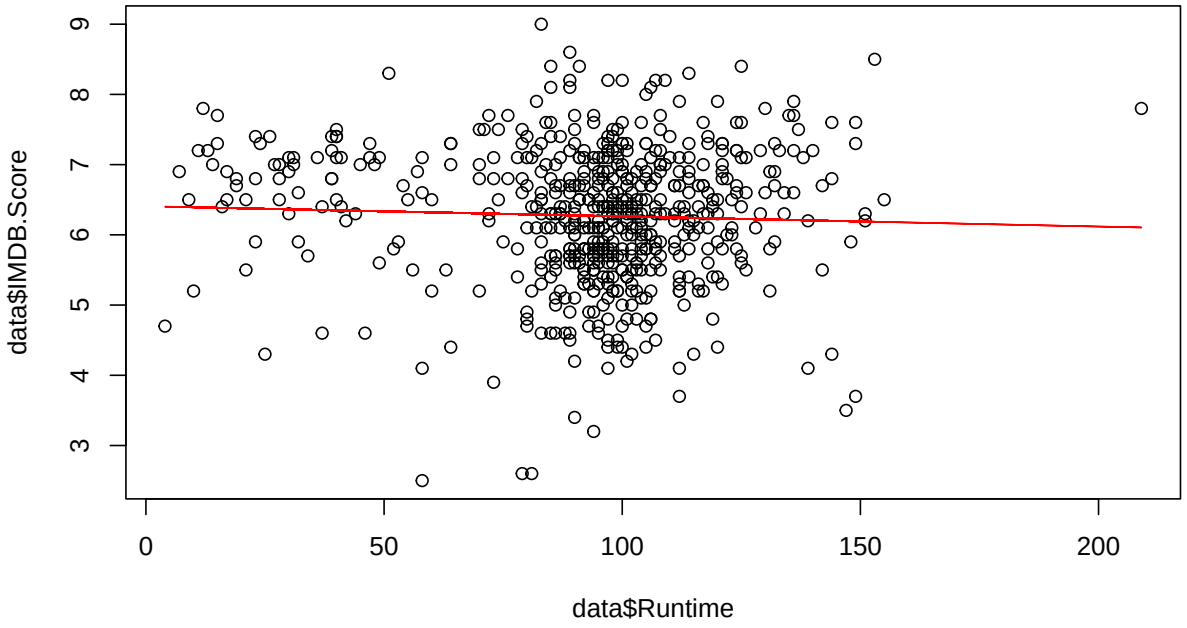


Iz grafičkog prikaza možemo vidjeti da duljina filma nema jak utjecaj na ocjenu.

Pošto se radi o numeričkim značajkama, jednostavnom regresijom možemo provjeriti utjecaj jedne na drugu značajku. Procijenit ćemo model jednostavne regresije s duljinom filma kao nezavisnom varijablom i IMDB ocjenom kao zavisnom varijablom.

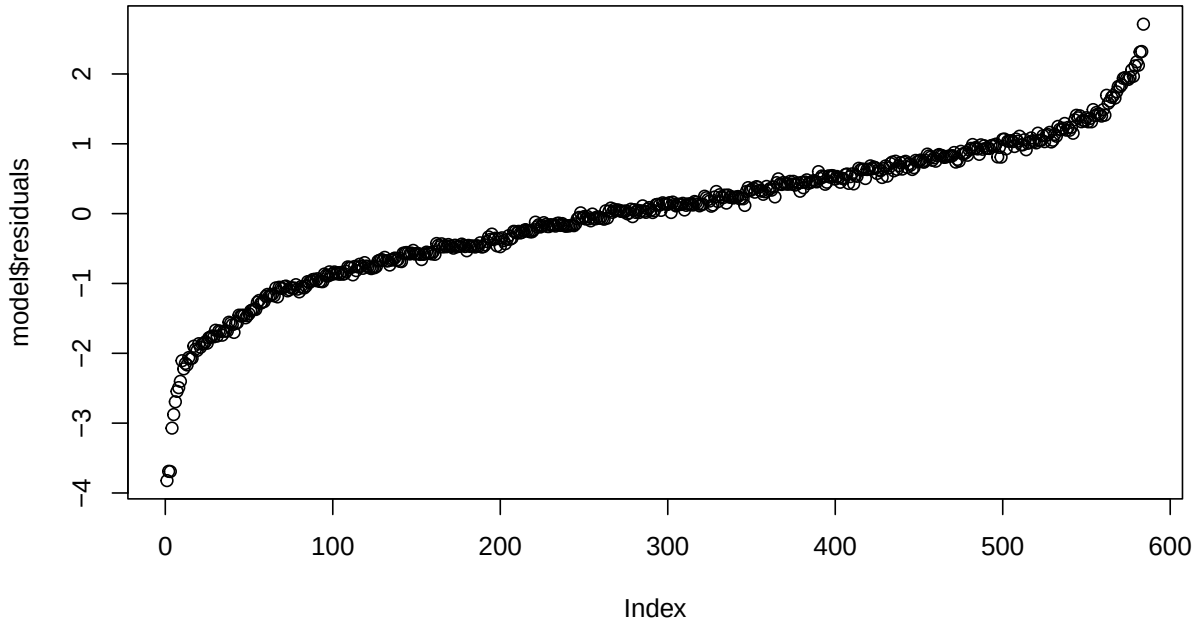
```
fit.runtime = lm(IMDB.Score~Runtime,data=data) #linearni model duljine trajanja filma (Runtime) i IMDB
```

```
plot(data$Runtime,data$IMDB.Score) #runtime vs IMDB ocjena
lines(data$Runtime,fit.runtime$fitted.values,col='red') #graficki prikaz procijenjenih vrijednosti iz m
```

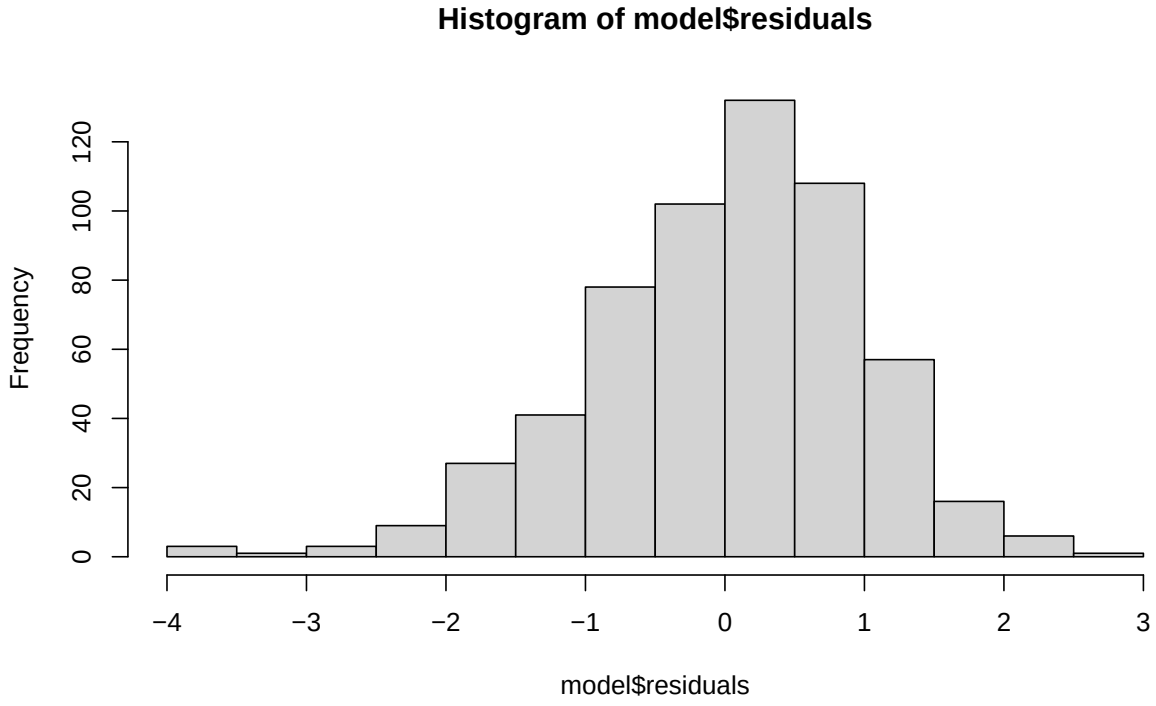


Prikazani graf sugerira da vrijednost značajke Runtime ne utječe na vrijednost značajke IMDB.Score. Normalnost reziduala provjerit ćemo grafički, pomoću kvantil-kvantil plota (usporedbom s linijom normalne razdiobe), te statistički pomoću Kolmogorov-Smirnovljevog testa.

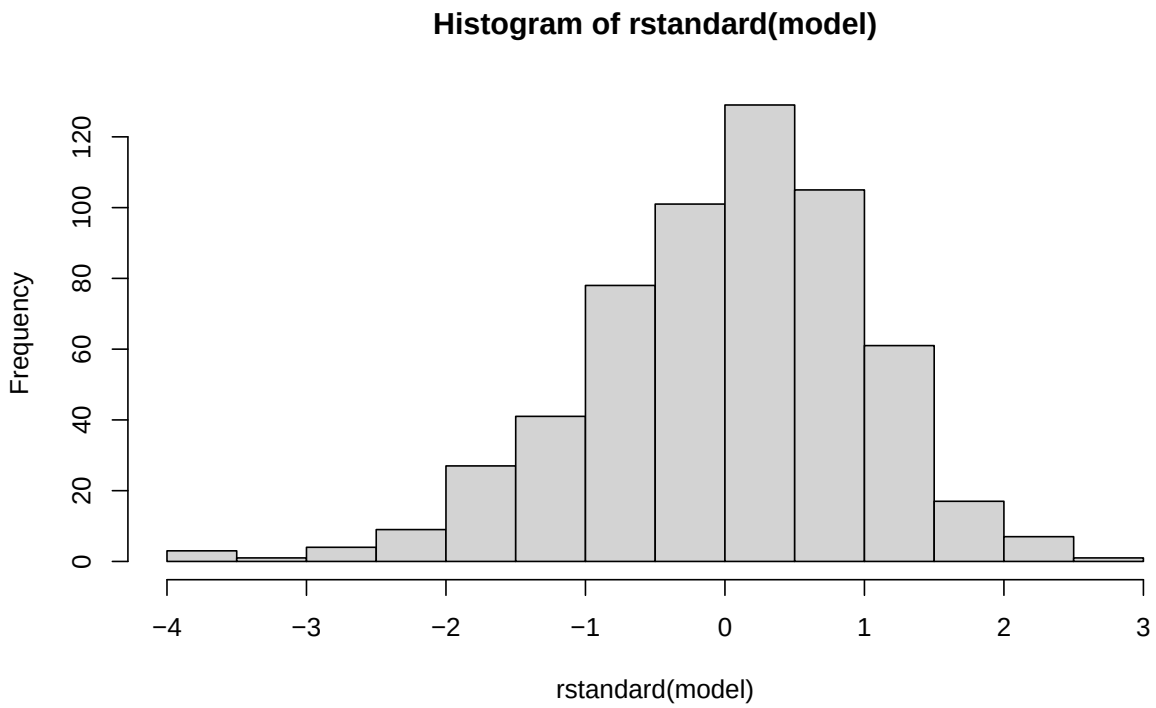
```
model = fit.runtime

plot(model$residuals)
```



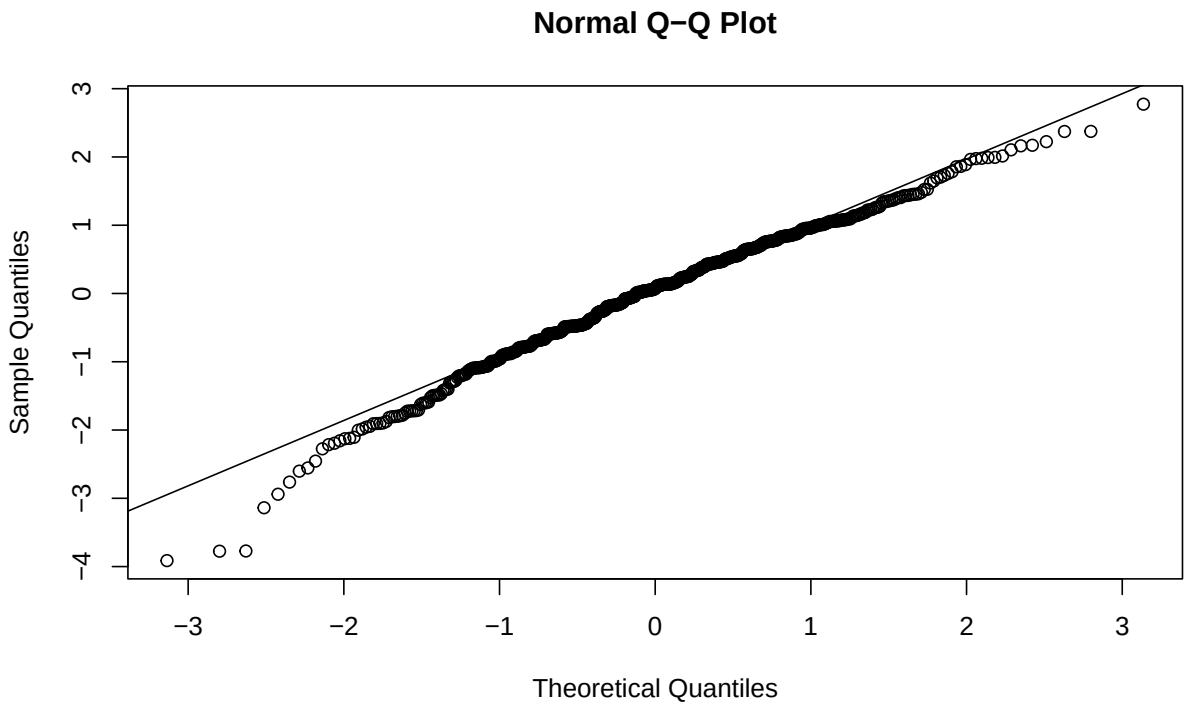

```
hist(model$residuals)
```



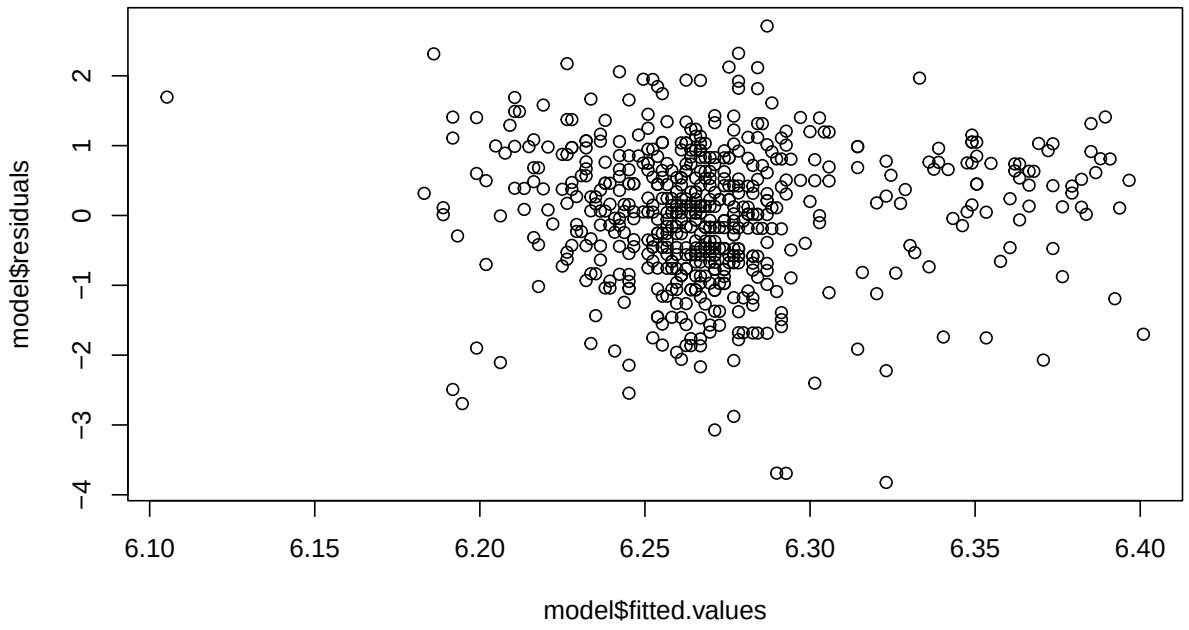
```
hist(rstandard(model))
```




```
qqnorm(rstandard(model))
qqline(rstandard(model))
```




```r
plot(model$fitted.values, model$residuals) #prikaz reziduala u ovisnosti o procjenama modela
```



```r
ks.test(rstandard(fit.runtime), 'pnorm')
```

```
## Warning in ks.test.default(rstandard(fit.runtime), "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.runtime)
## D = 0.052483, p-value = 0.08012
## alternative hypothesis: two-sided
```

```r
require(nortest)
lillie.test(rstandard(fit.runtime))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.runtime)
## D = 0.052389, p-value = 0.0006225
```

Distribucija reziduala, što je prikazano u histogramima, izgleda kao normalna distribucija. Iako se statistički testovi (Kolmogorov-Smirnovljev i Lillieforsov test) razlikuju u rezultatima, pošto se u rezultatima ne vidi veliko odstupanje od normalnosti, donijet ćemo statističke zaključke iz ovog modela.

```r
summary(fit.runtime)
```

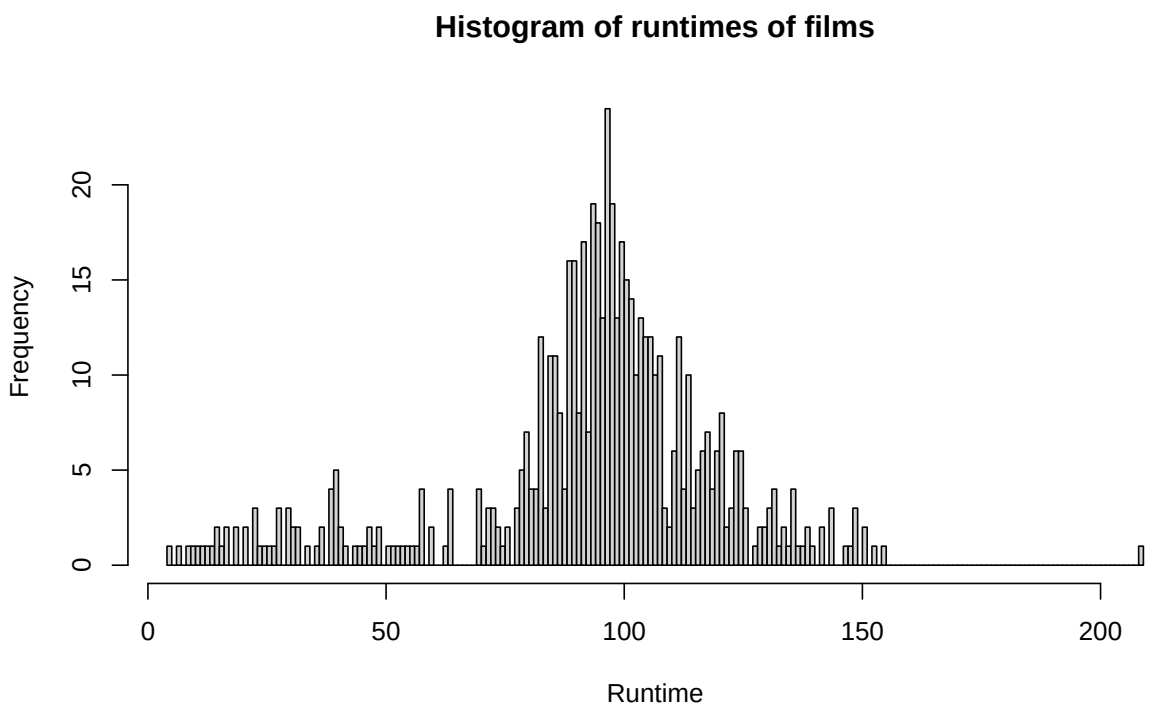
```
##
## Call:
## lm(formula = IMDB.Score ~ Runtime, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8231 -0.5783  0.0729  0.6842  2.7130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.406737   0.142588  44.932   <2e-16 ***
## Runtime     -0.001443   0.001461  -0.987    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9793 on 582 degrees of freedom
## Multiple R-squared:  0.001673,   Adjusted R-squared:  -4.283e-05
## F-statistic: 0.975 on 1 and 582 DF,  p-value: 0.3238
```

Rezultat t-testa upućuje na to da trajanje filma nije značajan koeficijent. Koeficijent deteminacije R^2 poprima vrlo malu vrijednost, što nam govori da je utjecaj značajke trajanja filma neznatan na IMDB ocjenu. F-test također upućuje na to da model nije značajan, to jest da značajka duljina trajanja filma ne utječe na IMDB ocjenu.

Iako ne postoji direktna veza između duljine trajanja filma i IMDB ocjene, možemo numeričku varijablu "runtime" pretvoriti u kategoričku varijablu s dvije kategorije: "kratki" i "dugi". Koristeći novostvorenu varijablu provjerit ćemo povezanost između duljine trajanja filma i IMDB ocjene. Prvo ćemo na histogramu procijeniti vrijednost Runtime varijable koja će odrediti granicu između kratkih i dugih filmova.

```r
hist(data$Runtime,
     breaks = seq(min(data$Runtime),
```

```
                max(data$Runtime), 1),
     main = 'Histogram of runtimes of films',
     xlab = 'Runtime')
```




```
summary(data$Runtime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   86.00   97.00   93.58  108.00  209.00
```

Iz histograma možemo procijeniti da bi vrijednost značajke Runtime od 75 minuta mogla biti dobra granica između kratkih i dugih filmova. Sad možemo podijeliti filmove na dva skupa:

```
shortfilms = data[data$Runtime <= 75,]
longfilms = data[data$Runtime > 75,]
```

Prikažimo prosječne vrijednosti ocjene za kratke i duge filmove:

```
cat('Prosjecna ocjena kratkih filmova iznosi ', mean(shortfilms$IMDB.Score), '\n')
```

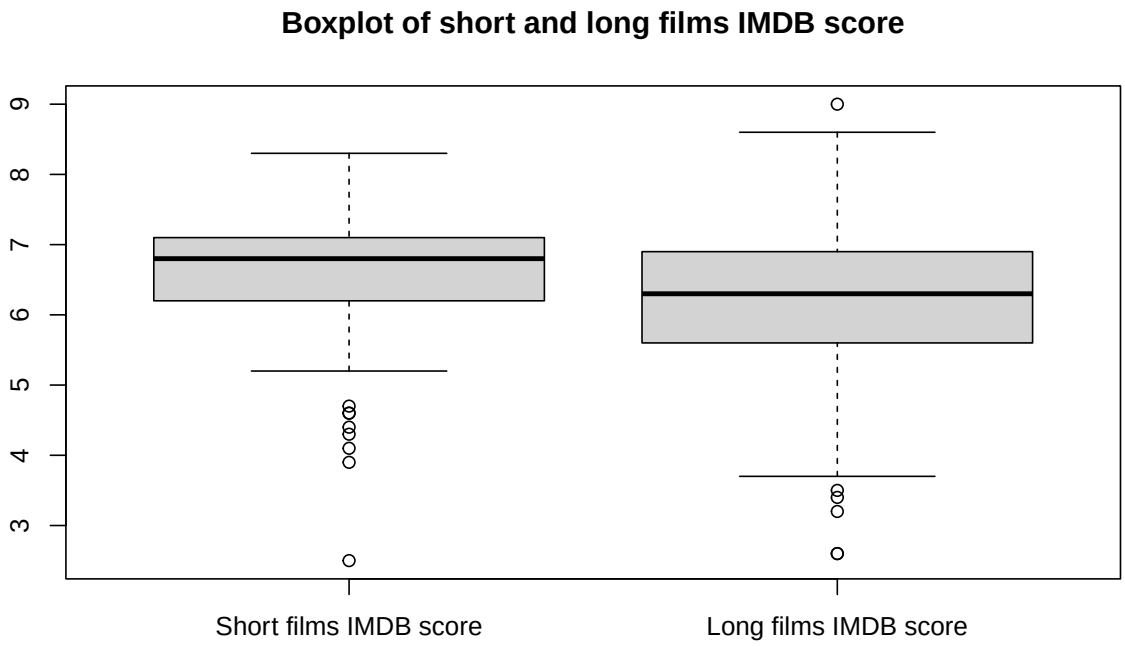
```
## Prosjecna ocjena kratkih filmova iznosi  6.518889
```

```
cat('Prosjecna ocjena dugih filmova iznosi ', mean(longfilms$IMDB.Score), '\n')
```

```
## Prosjecna ocjena dugih filmova iznosi  6.226721
```

Boxplotom možemo lakše primijetiti razliku između ove dvije skupine:

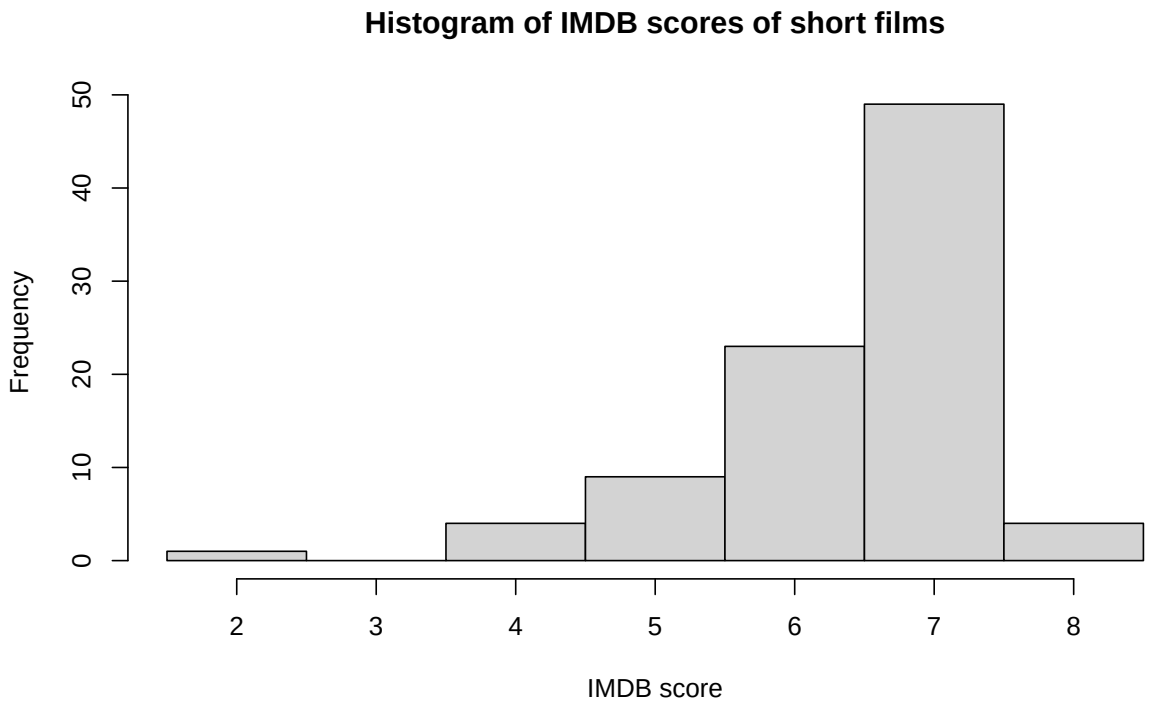
```
boxplot(shortfilms$IMDB.Score, longfilms$IMDB.Score,
        names = c('Short films IMDB score', 'Long films IMDB score'),
        main = 'Boxplot of short and long films IMDB score')
```



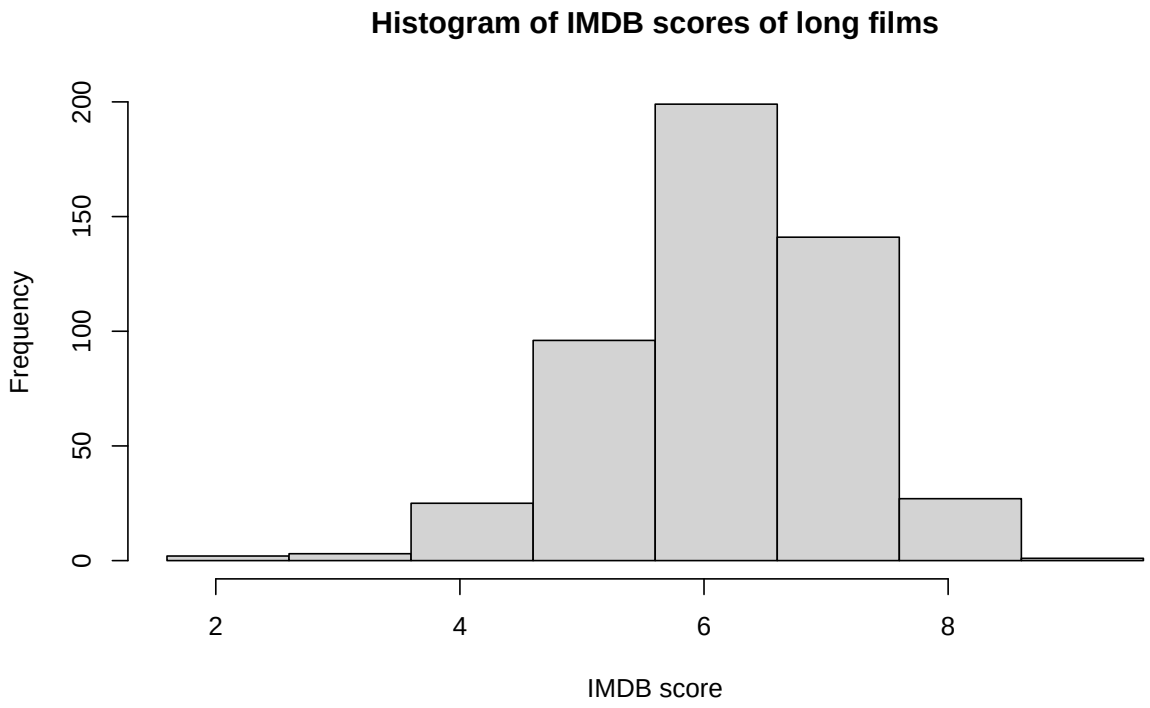
Na grafu možemo primijetiti da bi filmovi u trajanju do 75 minuta trebali biti bolje ocijenjeni nego filmovi u trajanju od više od 75 minuta.

Provjerimo prvo normalnost podataka. Prikažimo histograme obje populacije:

```
hist(shortfilms$IMDB.Score,
     breaks = seq(min(shortfilms$IMDB.Score)-1,
                  max(shortfilms$IMDB.Score)+1, 1),
     main = 'Histogram of IMDB scores of short films',
     xlab = 'IMDB score')
```

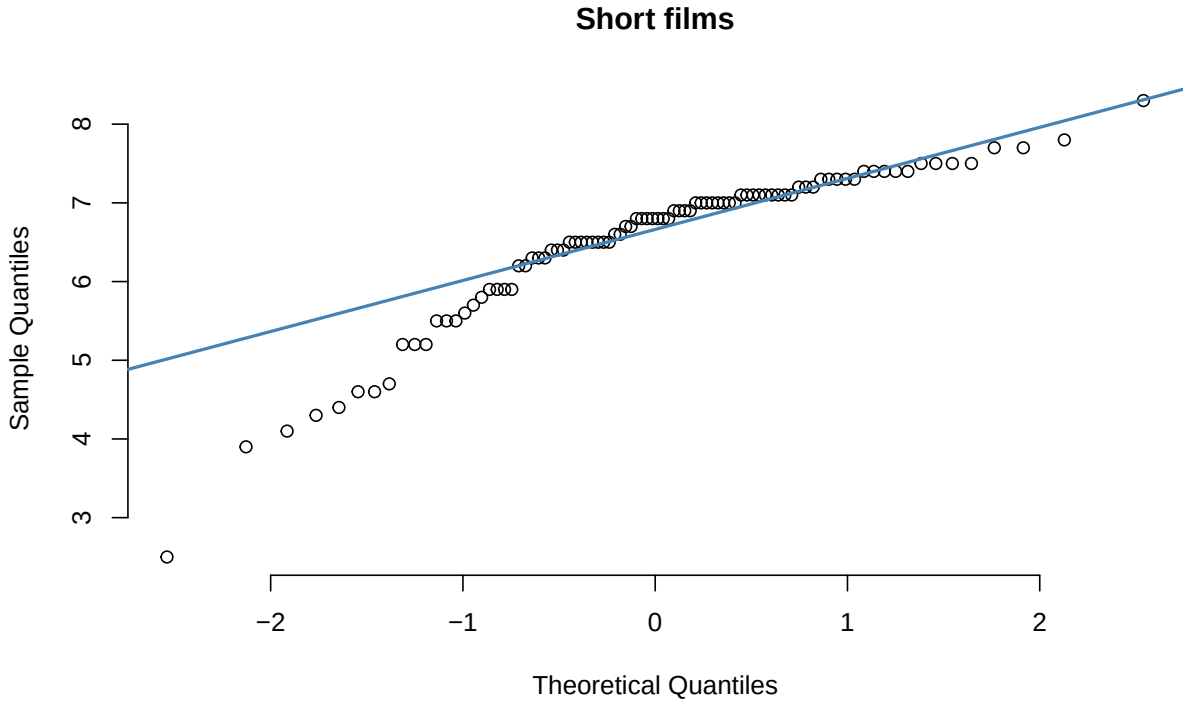



```
hist(longfilms$IMDB.Score,
     breaks = seq(min(longfilms$IMDB.Score)-1,
                  max(longfilms$IMDB.Score)+1, 1),
     main = 'Histogram of IMDB scores of long films',
     xlab = 'IMDB score')
```

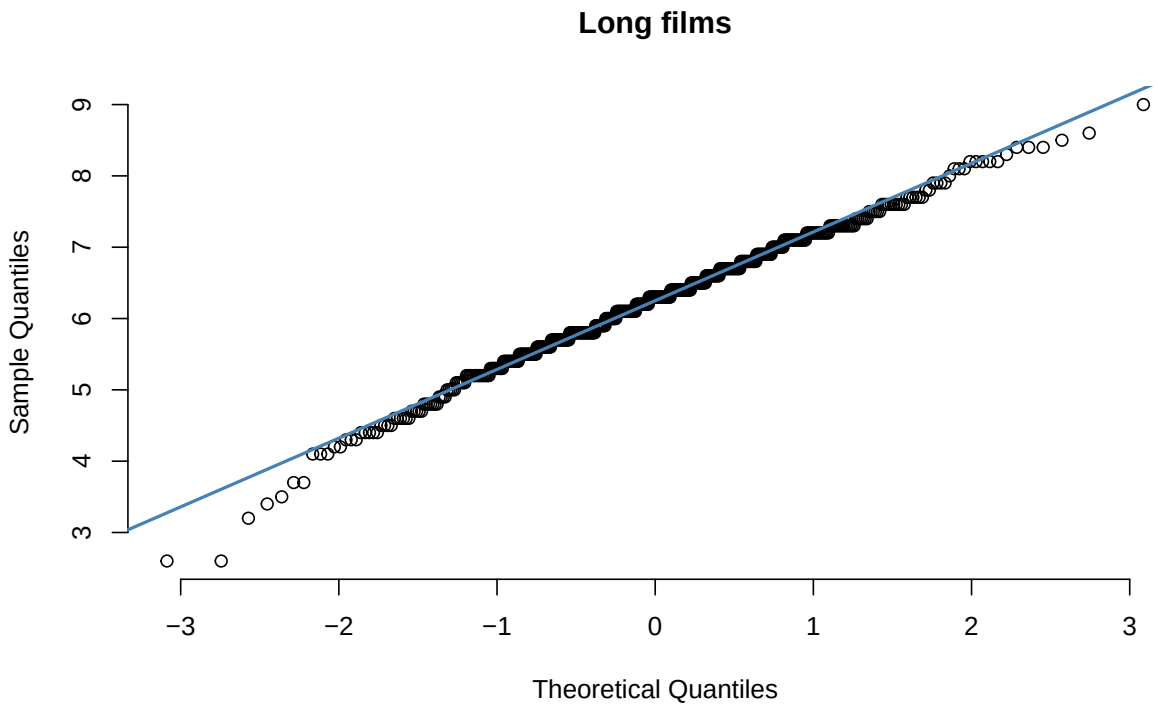


Prikažimo i q-q plotove obje populacije:

```
qqnorm(shortfilms$IMDB.Score, pch = 1, frame = FALSE,main='Short films')
qqline(shortfilms$IMDB.Score, col = "steelblue", lwd = 2)
```



```
qqnorm(longfilms$IMDB.Score, pch = 1, frame = FALSE,main='Long films')
qqline(longfilms$IMDB.Score, col = "steelblue", lwd = 2)
```



Izgled histograma i q-q plota za skup kratkih filmova upućuje na nenormalnost podataka. Testirajmo normalnost distribucije populacija Lillieforsovim testom.

```
h = lillie.test(shortfilms$IMDB.Score)
h
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  shortfilms$IMDB.Score
## D = 0.17016, p-value = 8.925e-07
```

```
h = lillie.test(longfilms$IMDB.Score)
h
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  longfilms$IMDB.Score
## D = 0.046253, p-value = 0.0134
```

Histogrami i qqplot-ovi, kao i Lillieforsov test upućuju na nenormalnu distribuciju podataka. Zbog toga ne možemo provesti test o jednakosti srednjih vrijednosti dvije populacije. Kako bismo proveli taj test naši podaci bi trebali podržavati pretpostavku normalnosti i nezavisnosti uzorka.

Alternativa t-testu su neparametarske metode. Metoda koju ćemo primijeniti na ovaj skup podataka bit će Wilcoxonov test predznačenih rangova. Prije primjene testa, stvorit ćemo novu varijablu length, koja poprima vrijednost 'short' ako je vrijednost varijable Runtime veća od 75 minuta ili 'long' u suprotnom. Wilcoxonov

test radimo nad varijablama IMDB.Score i length.

```
data$length <- with(data, ifelse(data$Runtime < 75, 'short', 'long'))

wilcox.test(data$IMDB.Score ~ data$length, data = data, exact = FALSE, alternative = 'less')
```

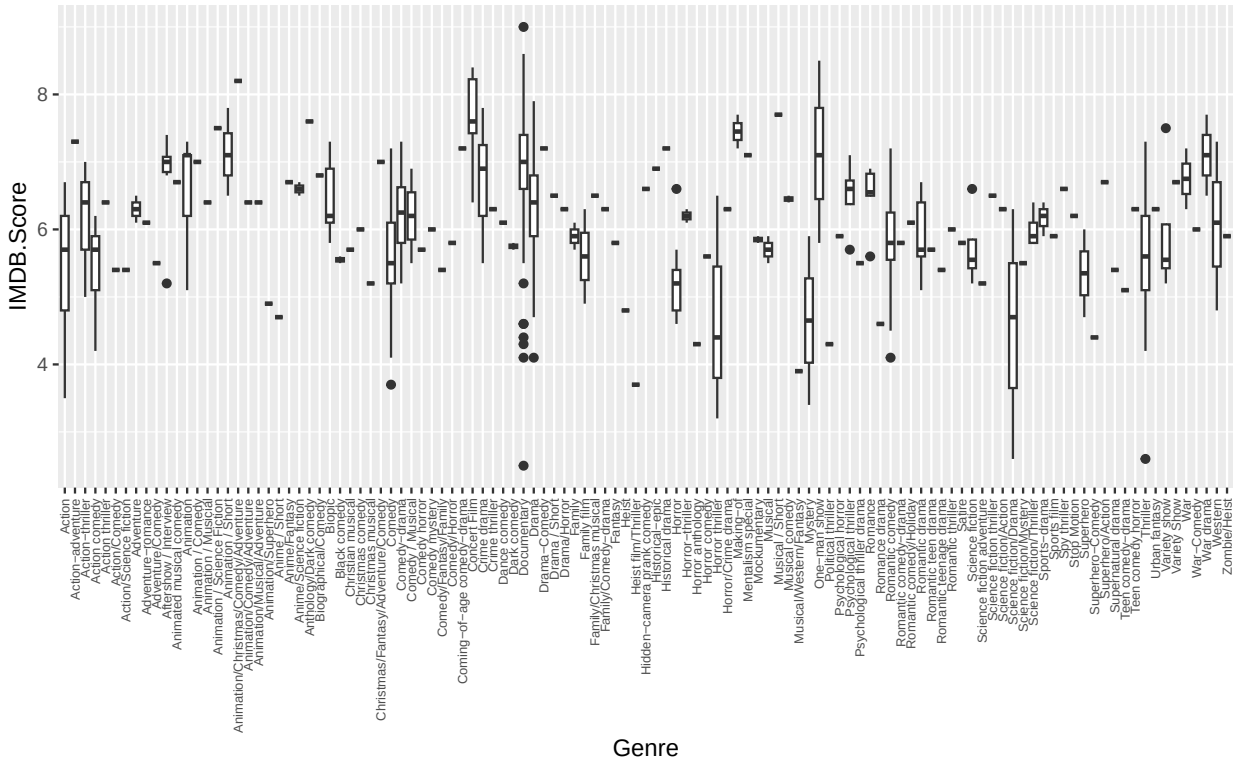
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data$IMDB.Score by data$length
## W = 16581, p-value = 0.0001002
## alternative hypothesis: true location shift is less than 0
```

Wilcoxonovim testom dobivena je p vrijednost iznosa 0.0001002, što je manje od 0.05. Stoga odbacujemo nultu hipotezu i zaključujemo kako je prosječna IMDB ocjena kratkih filmova veća od prosječne IMDB ocjene dugih filmova.

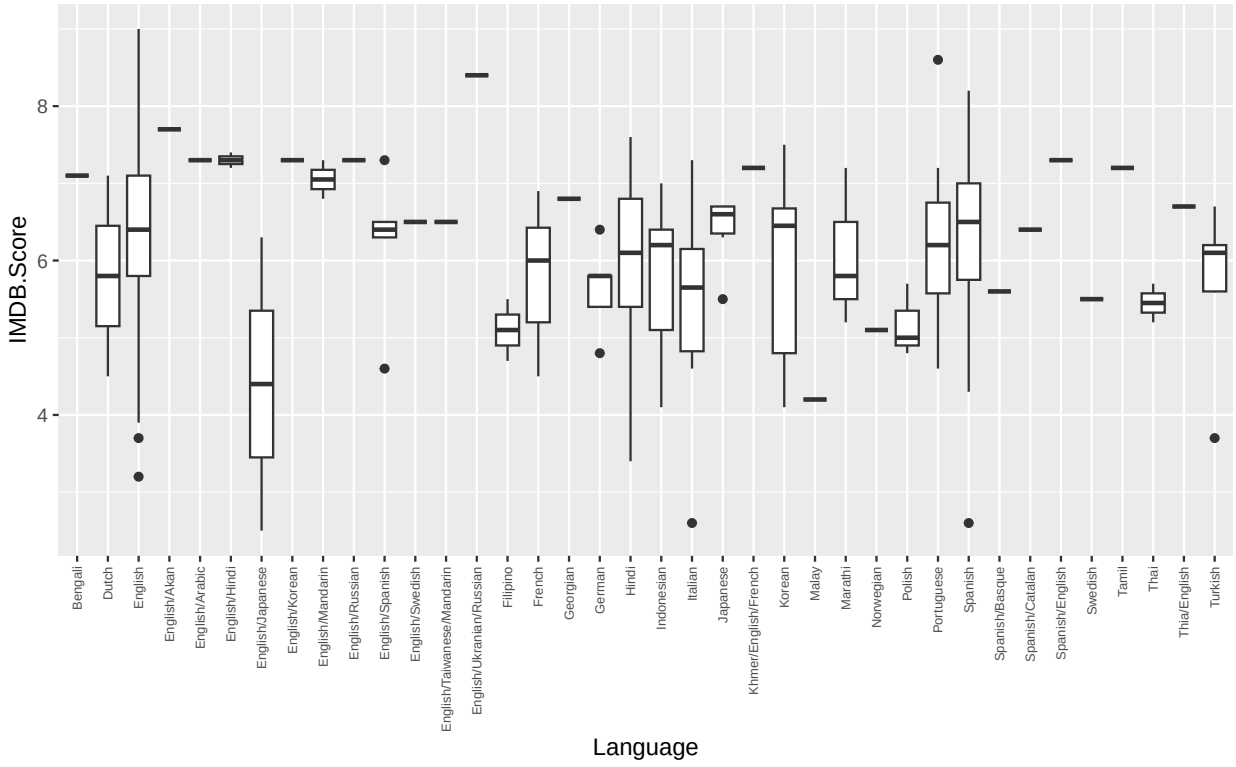
Zadatak 5. Možemo li naslutiti kako je film ocijenjen pomoću drugih značajki?

Za motivaciju ćemo prikazati postoji li neka povezanost između pojedine značajke i IMDB.Score-a. Na prva dva boxplota vidimo kako se medijani i kvartili razlikuju za različite žanrove i jezike. Zadnji graf upućuje na neovisnost između Runtime-a i IMDB.Score-a.

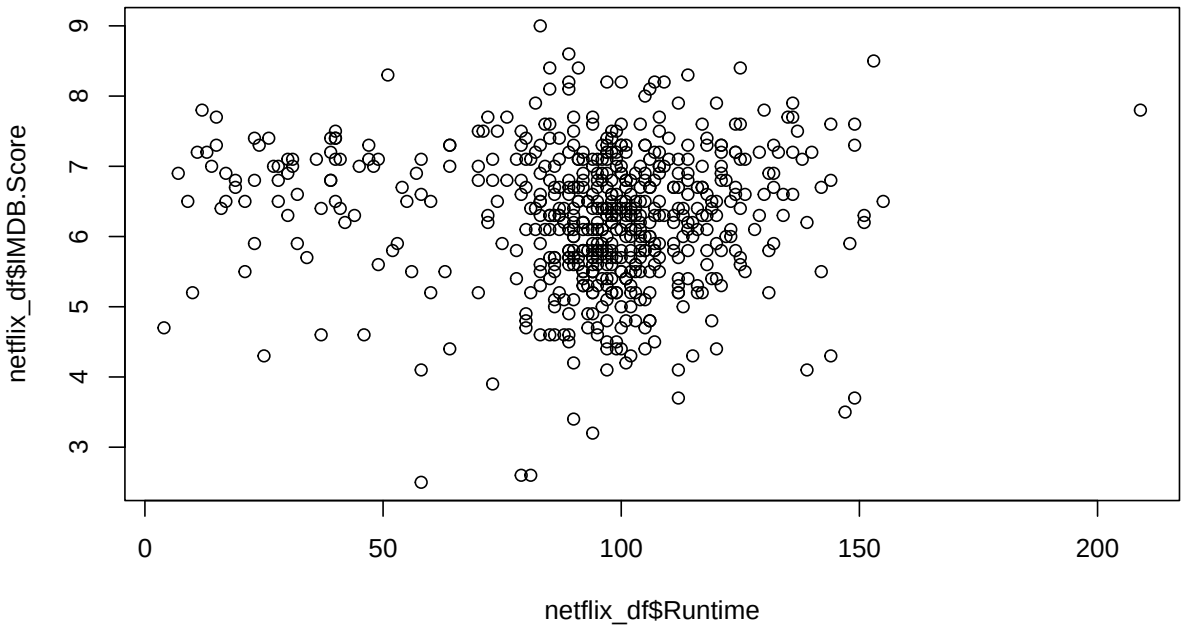
```
ggplot(netflix_df, aes(x = Genre, y = IMDB.Score)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size = 6))
```




```
ggplot(netflix_df, aes(x = Language, y = IMDB.Score)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size = 6))
```




```
plot(netflix_df$Runtime, netflix_df$IMDB.Score)
```



Pretvorit ćemo kategoričke varijable žanr i jezik u numeričke.

```
netflix_df$Genre <- factor(netflix_df$Genre)
netflix_df$Language <- factor(netflix_df$Language)
```

Računamo Pearsonov korelacijski koeficijent za sve kombinacije značajki: jezik, žanr i trajanje, što je prikazano u tablici. Vidimo kako značajke nisu međusobno linearno zavisne.

```
cor(cbind(netflix_df$Language,netflix_df$Genre,netflix_df$Runtime))
```

```
##            [,1]       [,2]      [,3]
## [1,] 1.00000000 0.05288443 0.1164943
## [2,] 0.05288443 1.00000000 0.1995784
## [3,] 0.11649428 0.19957844 1.0000000
```

Procijenit ćemo model višestruke regresije s nezavisnim varijablama: žanr, jezik, trajanje i zavisnom varijablom ocjena(IMDB Score).

```
fit.model = lm(IMDB.Score ~ Genre + Language + Runtime, netflix_df)
summary(fit.model)
```

```
##
## Call:
## lm(formula = IMDB.Score ~ Genre + Language + Runtime, data = netflix_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7564 -0.2782  0.0000  0.3447  1.9877
##
## Coefficients: (1 not defined because of singularities)
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              5.004460   0.816056   6.132 1.95e-09
## GenreAction-adventure                    1.088045   1.107468   0.982 0.326422
## GenreAction-thriller                     0.578112   0.530238   1.090 0.276194
## GenreAction comedy                       0.216936   0.475061   0.457 0.648152
## GenreAction thriller                     1.084038   0.807098   1.343 0.179935
## GenreAction/Comedy                      -0.105569   0.805948  -0.131 0.895846
## GenreAction/Science fiction             -0.235300   0.805933  -0.292 0.770458
## GenreAdventure                           0.839338   0.605989   1.385 0.166746
## GenreAdventure-romance                   0.754100   0.806828   0.935 0.350492
## GenreAdventure/Comedy                    0.174058   0.807004   0.216 0.829336
## GenreAftershow / Interview               2.016163   0.446820   4.512 8.28e-06
## GenreAnimated musical comedy             1.284245   0.806326   1.593 0.111956
## GenreAnimation                           1.491524   0.452728   3.295 0.001067
## GenreAnimation / Comedy                  2.053271   0.813155   2.525 0.011925
## GenreAnimation / Musicial                1.493188   0.814109   1.834 0.067321
## GenreAnimation / Science Fiction         2.293810   0.808374   2.838 0.004760
## GenreAnimation / Short                   2.520062   0.503734   5.003 8.24e-07
## GenreAnimation/Christmas/Comedy/Adventure 2.734348  0.806079   3.392 0.000757
## GenreAnimation/Comedy/Adventure          1.004203   0.806451   1.245 0.213728
## GenreAnimation/Musical/Adventure         0.954307   0.806167   1.184 0.237160
## GenreAnimation/Superhero                -0.485818   0.806519  -0.602 0.547248
## GenreAnime / Short                       0.162422   0.825631   0.197 0.844135
## GenreAnime/Fantasy                       1.309895   0.966414   1.355 0.175993
## GenreAnime/Science fiction               1.349605   0.807181   1.672 0.095249
## GenreAnthology/Dark comedy               1.808420   0.816711   2.214 0.027331
## GenreBiographical/Comedy                 1.294431   0.805948   1.606 0.108983
## GenreBiopic                              0.887963   0.384135   2.312 0.021270
```

```
## GenreBlack comedy                          0.064286   0.806216   0.080 0.936483
## GenreChristian musical                     0.234348   0.806079   0.291 0.771400
## GenreChristmas comedy                      0.354721   0.805958   0.440 0.660068
## GenreChristmas musical                    -0.275631   0.806041  -0.342 0.732550
## GenreChristmas/Fantasy/Adventure/Comedy    1.464493   0.805889   1.817 0.069873
## GenreComedy                                0.128553   0.309888   0.415 0.678467
## GenreComedy-drama                          0.776295   0.356116   2.180 0.029804
## GenreComedy / Musical                      1.273230   0.615964   2.067 0.039324
## GenreComedy horror                         0.264286   0.806216   0.328 0.743214
## GenreComedy mystery                        0.534348   0.806079   0.663 0.507749
## GenreComedy/Fantasy/Family                -0.075631   0.806041  -0.094 0.925287
## GenreComedy/Horror                         0.284452   0.805925   0.353 0.724298
## GenreComing-of-age comedy-drama            1.842932   0.818562   2.251 0.024860
## GenreConcert Film                          2.151049   0.422739   5.088 5.40e-07
## GenreCrime drama                           1.482958   0.373737   3.968 8.49e-05
## GenreCrime thriller                        0.944120   0.806745   1.170 0.242532
## GenreDance comedy                          0.674265   0.806269   0.836 0.403461
## GenreDark comedy                           0.334245   0.606291   0.551 0.581717
## GenreDocumentary                           1.686389   0.299037   5.639 3.09e-08
## GenreDrama                                 0.912615   0.302784   3.014 0.002729
## GenreDrama-Comedy                          1.814182   0.806519   2.249 0.024990
## GenreDrama / Short                         1.832691   0.821032   2.232 0.026115
## GenreDrama/Horror                          0.764493   0.805889   0.949 0.343338
## GenreFamily                                0.343940   0.609563   0.564 0.572883
## GenreFamily film                           0.199214   0.606413   0.329 0.742684
## GenreFamily/Christmas musical              0.814804   0.806094   1.011 0.312676
## GenreFamily/Comedy-drama                   0.734555   0.805863   0.912 0.362533
## GenreFantasy                               1.014122   0.973898   1.041 0.298318
## GenreHeist                                -0.665652   0.806079  -0.826 0.409379
## GenreHeist film/Thriller                  -2.284575   0.809006  -2.824 0.004963
## GenreHidden-camera prank comedy            1.244120   0.806745   1.542 0.123770
## GenreHistorical-epic                       1.194845   0.806185   1.482 0.139043
## GenreHistorical drama                      1.305239   0.807783   1.616 0.106862
## GenreHorror                               -0.032422   0.403167  -0.080 0.935941
## GenreHorror-thriller                       0.609607   0.605680   1.006 0.314747
## GenreHorror anthology                     -1.441684   0.816186  -1.766 0.078041
## GenreHorror comedy                         0.244120   0.806745   0.303 0.762341
## GenreHorror thriller                      -0.785610   0.522381  -1.504 0.133337
## GenreHorror/Crime drama                    0.784452   0.805925   0.973 0.330921
## GenreMaking-of                             2.592491   0.619116   4.187 3.42e-05
## GenreMentalism special                     2.113354   0.812259   2.602 0.009592
## GenreMockumentary                          0.718654   0.610505   1.177 0.239785
## GenreMusical                               0.044742   0.605839   0.074 0.941163
## GenreMusical / Short                       3.052650   0.821702   3.715 0.000230
## GenreMusical comedy                        0.844044   0.609796   1.384 0.167029
## GenreMusical/Western/Fantasy              -1.326149   0.808109  -1.641 0.101515
## GenreMystery                              -0.734123   0.607550  -1.208 0.227580
## GenreOne-man show                          1.651049   0.522398   3.161 0.001686
## GenrePolitical thriller                   -1.345279   0.805958  -1.669 0.095809
## GenrePsychological horror                  0.464286   0.806216   0.576 0.564994
## GenrePsychological thriller                0.764907   0.475750   1.608 0.108611
## GenrePsychological thriller drama         -0.269318   0.969250  -0.278 0.781251
## GenreRomance                               1.095719   0.426639   2.568 0.010556
## GenreRomance drama                        -0.258455   0.815368  -0.317 0.751412
```

```
## GenreRomantic comedy                  0.435514   0.312793   1.392 0.164536
## GenreRomantic comedy-drama            0.254514   0.805877   0.316 0.752290
## GenreRomantic comedy/Holiday          0.564493   0.805889   0.700 0.484017
## GenreRomantic drama                   0.592648   0.460166   1.288 0.198470
## GenreRomantic teen drama              0.744958   0.840181   0.887 0.375753
## GenreRomantic teenage drama           0.424999   0.840174   0.506 0.613222
## GenreRomantic thriller                0.274887   0.806290   0.341 0.733324
## GenreSatire                           0.492012   0.638292   0.771 0.441232
## GenreScience fiction                  0.224864   0.480983   0.468 0.640372
## GenreScience fiction adventure       -0.275631   0.806041  -0.342 0.732550
## GenreScience fiction thriller         1.461918   0.814612   1.795 0.073413
## GenreScience fiction/Action           0.724576   0.805862   0.899 0.369084
## GenreScience fiction/Drama           -0.705566   0.528225  -1.336 0.182341
## GenreScience fiction/Mystery         -0.255051   0.806476  -0.316 0.751962
## GenreScience fiction/Thriller         0.579255   0.475663   1.218 0.223971
## GenreSports-drama                     0.710994   0.522558   1.361 0.174348
## GenreSports film                      0.855144   0.840264   1.018 0.309386
## GenreSpy thriller                     0.884866   0.606170   1.460 0.145082
## GenreStop Motion                      1.283209   0.813865   1.577 0.115600
## GenreSuperhero                       -0.210455   0.605678  -0.347 0.728407
## GenreSuperhero-Comedy                -1.145486   0.805877  -1.421 0.155917
## GenreSuperhero/Action                 0.964907   0.806348   1.197 0.232104
## GenreSupernatural drama               0.299572   0.901992   0.332 0.739958
## GenreTeen comedy-drama               -0.365652   0.806079  -0.454 0.650332
## GenreTeen comedy horror               0.954100   0.806828   1.183 0.237645
## GenreThriller                         0.178933   0.318054   0.563 0.574008
## GenreUrban fantasy                    0.634762   0.806019   0.788 0.431405
## GenreVariety show                     0.848592   0.481926   1.761 0.078974
## GenreVariety Show                     1.663458   0.811221   2.051 0.040913
## GenreWar                              1.149628   0.605691   1.898 0.058357
## GenreWar-Comedy                       0.284866   0.806236   0.353 0.724015
## GenreWar drama                        0.455549   0.809986   0.562 0.574124
## GenreWestern                          0.340364   0.527540   0.645 0.519146
## GenreZombie/Heist                    -0.074596   0.808855  -0.092 0.926563
## LanguageDutch                        -1.176594   0.934520  -1.259 0.208697
## LanguageEnglish                      -0.506800   0.757903  -0.669 0.504052
## LanguageEnglish/Akan                  0.882807   1.304275   0.677 0.498859
## LanguageEnglish/Arabic               -0.528488   1.070627  -0.494 0.621822
## LanguageEnglish/Hindi                 0.284824   0.920835   0.309 0.757234
## LanguageEnglish/Japanese             -3.077290   0.930334  -3.308 0.001019
## LanguageEnglish/Korean                      NA         NA      NA       NA
## LanguageEnglish/Mandarin             -0.229627   0.921244  -0.249 0.803280
## LanguageEnglish/Russian              -0.288985   1.066527  -0.271 0.786552
## LanguageEnglish/Spanish              -0.862037   0.823520  -1.047 0.295790
## LanguageEnglish/Swedish              -0.590021   1.063154  -0.555 0.579201
## LanguageEnglish/Taiwanese/Mandarin   -0.325191   1.070450  -0.304 0.761435
## LanguageEnglish/Ukranian/Russian      0.801035   1.066665   0.751 0.453079
## LanguageFilipino                     -1.443174   0.929219  -1.553 0.121131
## LanguageFrench                       -0.974287   0.777699  -1.253 0.210963
## LanguageGeorgian                     -0.120373   1.063608  -0.113 0.909945
## LanguageGerman                       -0.822127   0.859939  -0.956 0.339592
## LanguageHindi                        -0.699795   0.774045  -0.904 0.366460
## LanguageIndonesian                   -0.916684   0.806445  -1.137 0.256295
## LanguageItalian                      -1.037368   0.794176  -1.306 0.192173
```

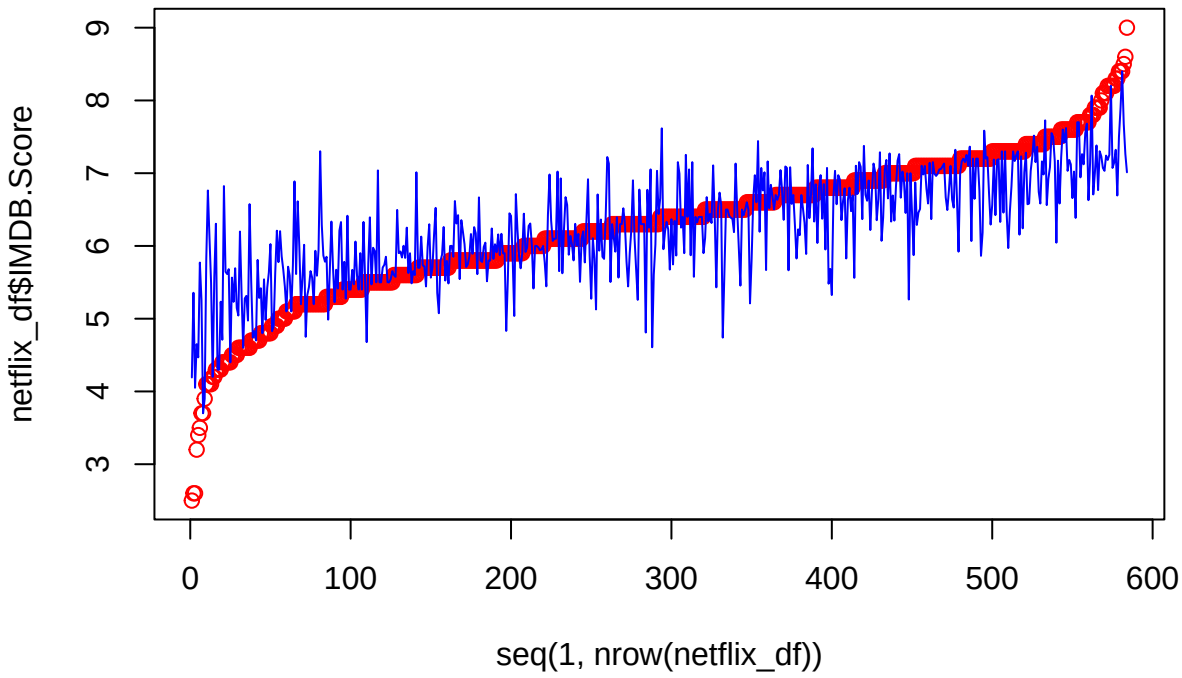
```
## LanguageJapanese                            -0.652202   0.924214  -0.706 0.480766
## LanguageKhmer/English/French                -0.074259   1.076537  -0.069 0.945038
## LanguageKorean                              -0.917723   0.823489  -1.114 0.265713
## LanguageMalay                               -2.029305   1.131719  -1.793 0.073654
## LanguageMarathi                             -1.019116   0.881651  -1.156 0.248354
## LanguageNorwegian                           -0.730257   1.102565  -0.662 0.508116
## LanguagePolish                              -1.016425   0.880392  -1.155 0.248927
## LanguagePortuguese                          -0.313626   0.789837  -0.397 0.691507
## LanguageSpanish                             -0.635342   0.770200  -0.825 0.409880
## LanguageSpanish/Basque                      -0.356903   1.305205  -0.273 0.784641
## LanguageSpanish/Catalan                     -1.448447   1.071040  -1.352 0.176963
## LanguageSpanish/English                     -0.348861   1.067402  -0.327 0.743953
## LanguageSwedish                             -0.541613   1.075001  -0.504 0.614641
## LanguageTamil                                0.275016   1.071317   0.257 0.797526
## LanguageThai                                -1.389352   0.936177  -1.484 0.138520
## LanguageThia/English                        -0.789192   1.065291  -0.741 0.459203
## LanguageTurkish                             -0.995963   0.833926  -1.194 0.233013
## Runtime                                      0.009979   0.001730   5.769 1.53e-08
##
## (Intercept)                                 ***
## GenreAction-adventure
## GenreAction-thriller
## GenreAction comedy
## GenreAction thriller
## GenreAction/Comedy
## GenreAction/Science fiction
## GenreAdventure
## GenreAdventure-romance
## GenreAdventure/Comedy
## GenreAftershow / Interview                  ***
## GenreAnimated musical comedy
## GenreAnimation                              **
## GenreAnimation / Comedy                     *
## GenreAnimation / Musicial                   .
## GenreAnimation / Science Fiction            **
## GenreAnimation / Short                      ***
## GenreAnimation/Christmas/Comedy/Adventure ***
## GenreAnimation/Comedy/Adventure
## GenreAnimation/Musical/Adventure
## GenreAnimation/Superhero
## GenreAnime / Short
## GenreAnime/Fantasy
## GenreAnime/Science fiction                  .
## GenreAnthology/Dark comedy                  *
## GenreBiographical/Comedy
## GenreBiopic                                 *
## GenreBlack comedy
## GenreChristian musical
## GenreChristmas comedy
## GenreChristmas musical
## GenreChristmas/Fantasy/Adventure/Comedy     .
## GenreComedy
## GenreComedy-drama                           *
## GenreComedy / Musical                       *
```

```
## GenreComedy horror
## GenreComedy mystery
## GenreComedy/Fantasy/Family
## GenreComedy/Horror
## GenreComing-of-age comedy-drama          *
## GenreConcert Film                        ***
## GenreCrime drama                         ***
## GenreCrime thriller
## GenreDance comedy
## GenreDark comedy
## GenreDocumentary                         ***
## GenreDrama                               **
## GenreDrama-Comedy                        *
## GenreDrama / Short                       *
## GenreDrama/Horror
## GenreFamily
## GenreFamily film
## GenreFamily/Christmas musical
## GenreFamily/Comedy-drama
## GenreFantasy
## GenreHeist
## GenreHeist film/Thriller                 **
## GenreHidden-camera prank comedy
## GenreHistorical-epic
## GenreHistorical drama
## GenreHorror
## GenreHorror-thriller
## GenreHorror anthology                    .
## GenreHorror comedy
## GenreHorror thriller
## GenreHorror/Crime drama
## GenreMaking-of                           ***
## GenreMentalism special                   **
## GenreMockumentary
## GenreMusical
## GenreMusical / Short                     ***
## GenreMusical comedy
## GenreMusical/Western/Fantasy
## GenreMystery
## GenreOne-man show                        **
## GenrePolitical thriller                  .
## GenrePsychological horror
## GenrePsychological thriller
## GenrePsychological thriller drama
## GenreRomance                             *
## GenreRomance drama
## GenreRomantic comedy
## GenreRomantic comedy-drama
## GenreRomantic comedy/Holiday
## GenreRomantic drama
## GenreRomantic teen drama
## GenreRomantic teenage drama
## GenreRomantic thriller
## GenreSatire
```

```
## GenreScience fiction
## GenreScience fiction adventure
## GenreScience fiction thriller             .
## GenreScience fiction/Action
## GenreScience fiction/Drama
## GenreScience fiction/Mystery
## GenreScience fiction/Thriller
## GenreSports-drama
## GenreSports film
## GenreSpy thriller
## GenreStop Motion
## GenreSuperhero
## GenreSuperhero-Comedy
## GenreSuperhero/Action
## GenreSupernatural drama
## GenreTeen comedy-drama
## GenreTeen comedy horror
## GenreThriller
## GenreUrban fantasy
## GenreVariety show                         .
## GenreVariety Show                         *
## GenreWar                                  .
## GenreWar-Comedy
## GenreWar drama
## GenreWestern
## GenreZombie/Heist
## LanguageDutch
## LanguageEnglish
## LanguageEnglish/Akan
## LanguageEnglish/Arabic
## LanguageEnglish/Hindi
## LanguageEnglish/Japanese                  **
## LanguageEnglish/Korean
## LanguageEnglish/Mandarin
## LanguageEnglish/Russian
## LanguageEnglish/Spanish
## LanguageEnglish/Swedish
## LanguageEnglish/Taiwanese/Mandarin
## LanguageEnglish/Ukranian/Russian
## LanguageFilipino
## LanguageFrench
## LanguageGeorgian
## LanguageGerman
## LanguageHindi
## LanguageIndonesian
## LanguageItalian
## LanguageJapanese
## LanguageKhmer/English/French
## LanguageKorean
## LanguageMalay                             .
## LanguageMarathi
## LanguageNorwegian
## LanguagePolish
## LanguagePortuguese
```

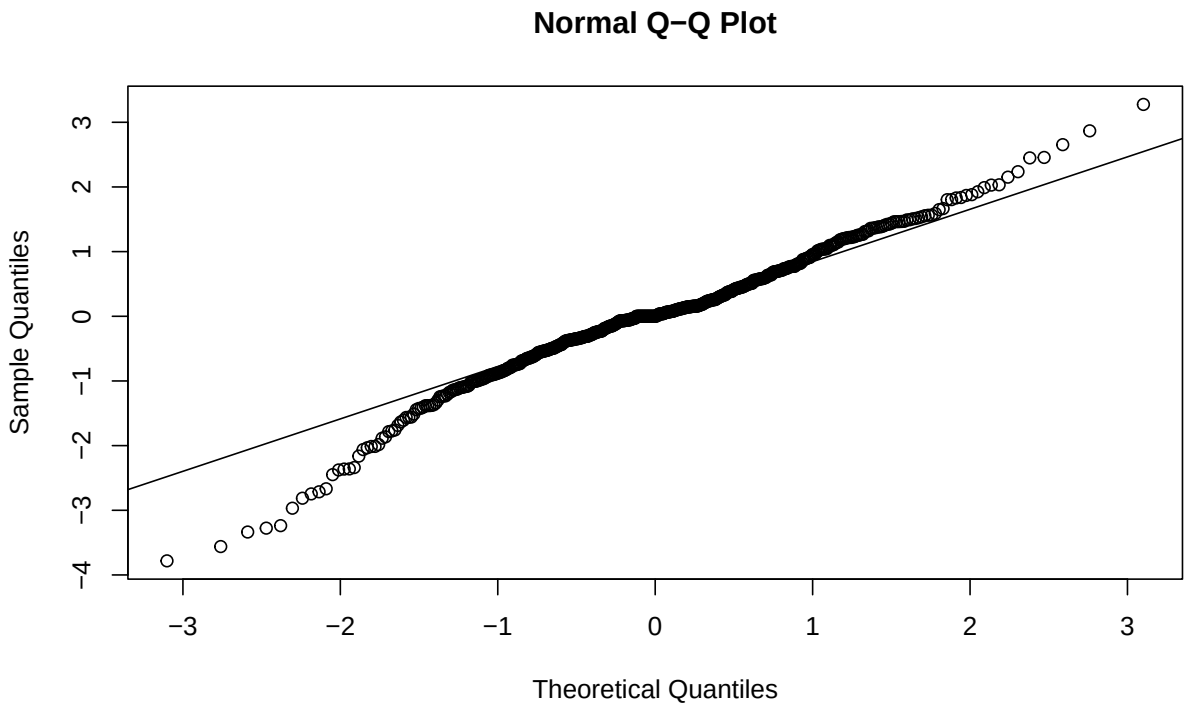
```
## LanguageSpanish
## LanguageSpanish/Basque
## LanguageSpanish/Catalan
## LanguageSpanish/English
## LanguageSwedish
## LanguageTamil
## LanguageThai
## LanguageThia/English
## LanguageTurkish
## Runtime                                    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7518 on 432 degrees of freedom
## Multiple R-squared:  0.5633, Adjusted R-squared:  0.4107
## F-statistic:  3.69 on 151 and 432 DF,  p-value: < 2.2e-16
```

```
plot(seq(1, nrow(netflix_df)), netflix_df$IMDB.Score, col="red")
lines(seq(1, nrow(netflix_df)), predict(fit.model), col = "blue")
```



S obzirom na to da je p-vrijednost F-statistike manje od 2.2e-16 « 0.05, odbacujemo njezinu nultu hipotezu. Iz toga zaključujemo da postoji utjecaj nezavisnih varijabli na vrijednost zavisne varijable, tj. model je značajan. Prema t-vrijednostima možemo vidjeti značajnost koeficijenata. Nulta hipoteza je $H_0 : \beta_i = 0$. Ovu hipotezu možemo odbaciti za npr. GenreAftershow / Interview, te je to onda značajni koeficijent modela. Multiple R-squared iznosi 0.5633, što zapravo znači kako 56,33% varijabilnosti zavisne varijable može biti objašnjeno navedenim regresijskim modelom.

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




```
ks.test(rstandard(fit.model),'pnorm')
```

```
## Warning in ks.test.default(rstandard(fit.model), "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.model)
## D = 0.067625, p-value = 0.01736
## alternative hypothesis: two-sided
```

```
require(nortest)
lillie.test(rstandard(fit.model))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.model)
## D = 0.067074, p-value = 8.077e-06
```

Pretpostavka testova je da su reziduali iz normalne distribucije. P-vrijednost Kolmogorov-Smirnov testa iznosi 0.01736, a prema Lilliefors (Kolmogorov-Smirnov) testu normalnosti 8.077e-06. S obzirom da su vrijednosti manje od 0.05, odbacujemo hipotezu normalnosti.

Iz svega navedenog zaključujemo kako ovaj model nije dobar za predviđanje ocjene filma.