IMDB_movie_dataset

Mario Slatinac, Alen Murtić

5 veljače 2017

IMDB movie dataset

```
data <- read.csv("../../data/IMDB_movie_dataset.csv", encoding="UTF-8")
```

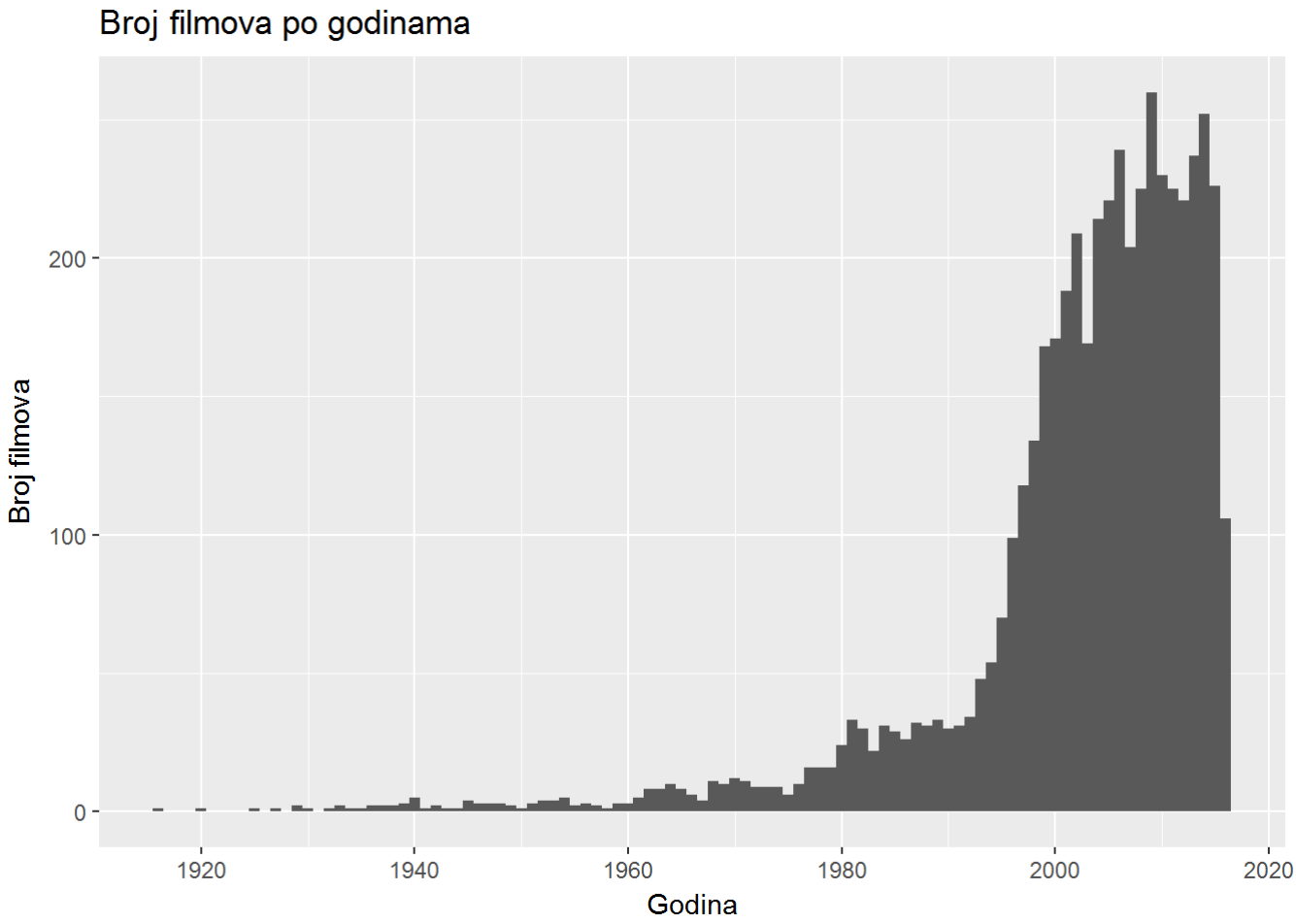
Pregled strukture podatkovnog skupa.

```
str(data)
```

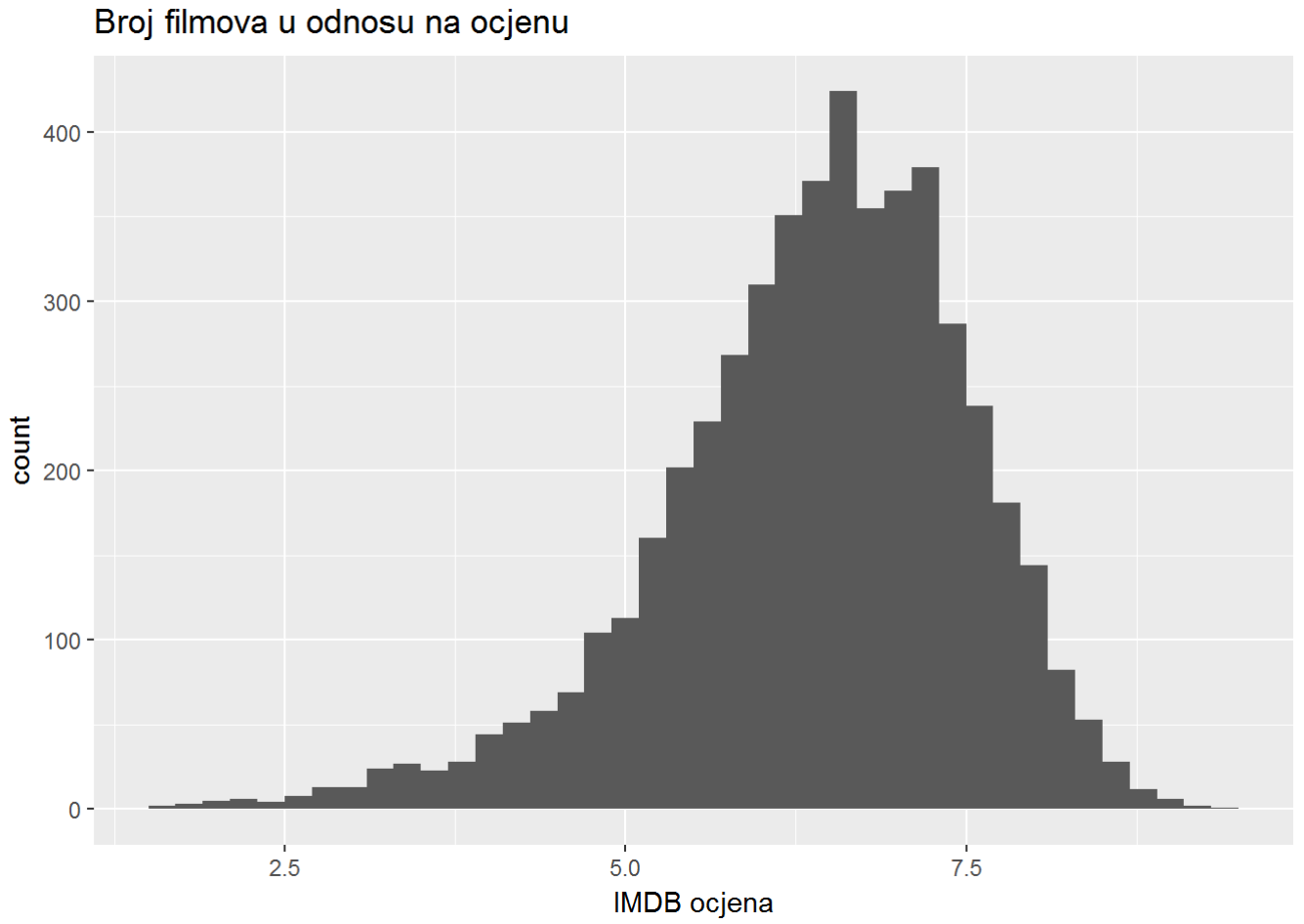
```
## 'data.frame':    5043 obs. of  28 variables:
##  $ color                   : Factor w/ 3 levels "","," Black and White",..: 3 3 3 3 1 3 3 3
 3 3 ...
##  $ director_name           : Factor w/ 2399 levels "","A. Raven Cruz",..: 927 801 2027 37
7 603 106 2030 1652 1228 551 ...
##  $ num_critic_for_reviews  : 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            : Factor w/ 3033 levels "","50 Cent","A. Michael Baldwin",..:
 1407 2218 2488 534 2432 2549 1227 801 2439 653 ...
##  $ actor_1_facebook_likes  : int  1000 40000 11000 27000 131 640 24000 799 26000 25000
 ...
##  $ gross                   : int  760505847 309404152 200074175 448130642 NA 73058679 336
530303 200807262 458991599 301956980 ...
##  $ genres                  : Factor w/ 914 levels "Action","Action|Adventure",..: 107 101
 128 288 754 126 120 308 126 447 ...
##  $ actor_1_name            : Factor w/ 2098 levels "","50 Cent","A.J. Buckley",..: 302 97
9 353 1968 526 440 785 221 336 32 ...
##  $ movie_title             : Factor w/ 4917 levels "#Horror ","[Rec] 2 ",..: 398 2731 327
9 3707 3332 1961 3289 3459 399 1631 ...
##  $ num_voted_users         : int  886204 471220 275868 1144337 8 212204 383056 294810 462
669 321795 ...
##  $ cast_total_facebook_likes: int  4834 48350 11700 106759 143 1873 46055 2036 92000 58753
 ...
##  $ actor_3_name            : Factor w/ 3522 levels "","50 Cent","A.J. Buckley",..: 3442 1
392 3134 1769 1 2714 1969 2162 3018 2941 ...
##  $ facenumber_in_poster    : int  0 0 1 0 0 1 0 1 4 3 ...
##  $ plot_keywords           : Factor w/ 4761 levels "","10 year old|dog|florida|girl|super
market",..: 1320 4283 2076 3484 1 651 4745 29 1142 2005 ...
##  $ movie_imdb_link         : Factor w/ 4919 levels "http://www.imdb.com/title/tt0006864/?
ref_=fn_tt_tt_1",..: 2965 2721 4533 3756 4918 2476 2526 2458 4546 2551 ...
##  $ num_user_for_reviews    : int  3054 1238 994 2701 NA 738 1902 387 1117 973 ...
##  $ language                : Factor w/ 48 levels "","Aboriginal",..: 13 13 13 13 1 13 13
 13 13 13 ...
##  $ country                 : Factor w/ 66 levels "","Afghanistan",..: 65 65 63 65 1 65 65
 65 65 63 ...
##  $ content_rating          : Factor w/ 19 levels "","Approved",..: 10 10 10 10 1 10 10 9
 10 9 ...
##  $ 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 10000 ...
```

```
ggplot(data, aes(title_year)) + geom_histogram(binwidth = 1) + labs(x="Godina", y="Broj filmo
va", title="Broj filmova po godinama")
```

```
## Warning: Removed 108 rows containing non-finite values (stat_bin).
```

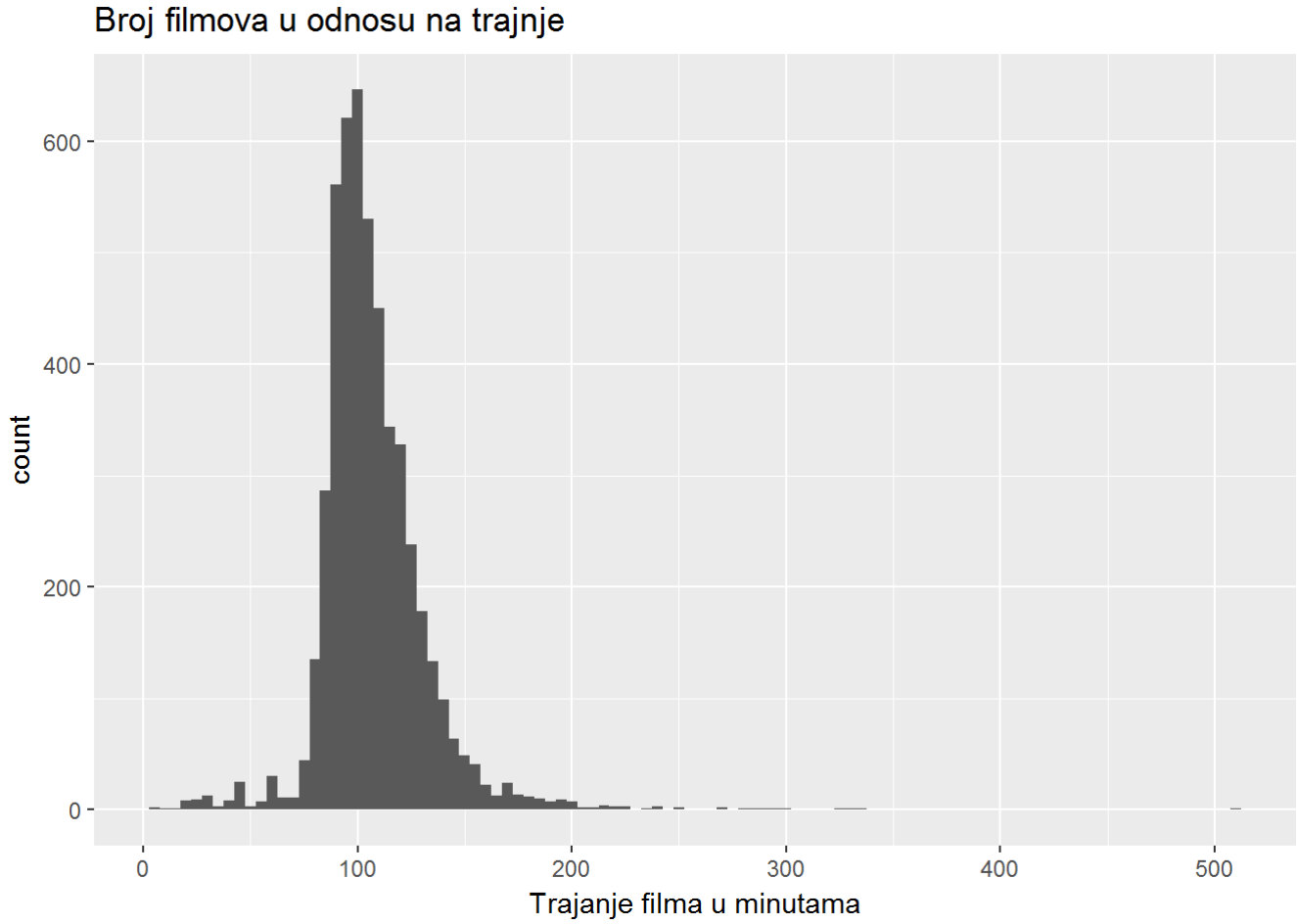



```
ggplot(data, aes(imdb_score)) + geom_histogram(binwidth = 0.2) + labs(x="IMDB ocjena", title
= "Broj filmova u odnosu na ocjenu")
```

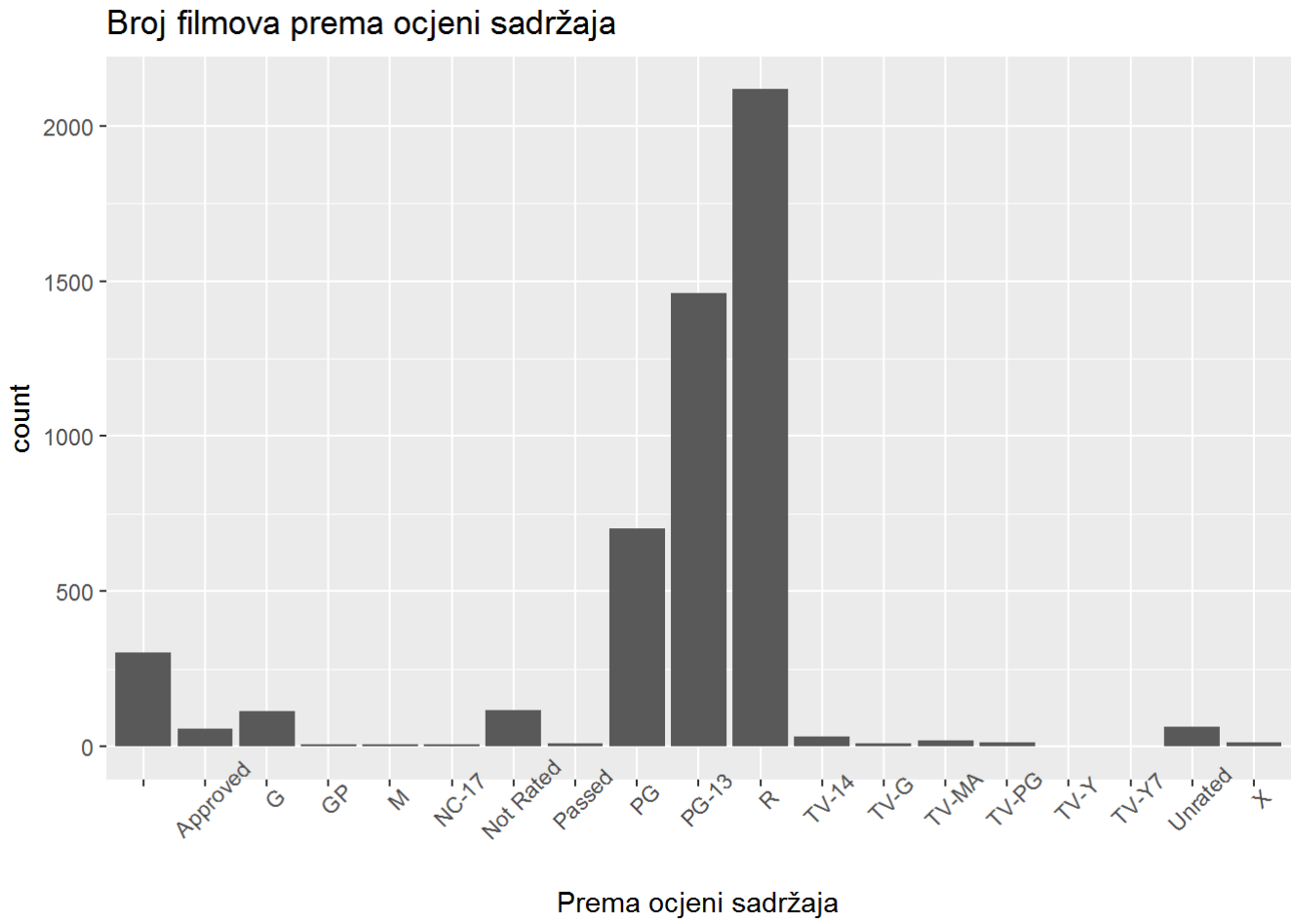


```
ggplot(data, aes(duration)) + geom_histogram(binwidth = 5) + labs(x="Trajanje filma u minutam
a", title="Broj filmova u odnosu na trajnje")
```

```
## Warning: Removed 15 rows containing non-finite values (stat_bin).
```



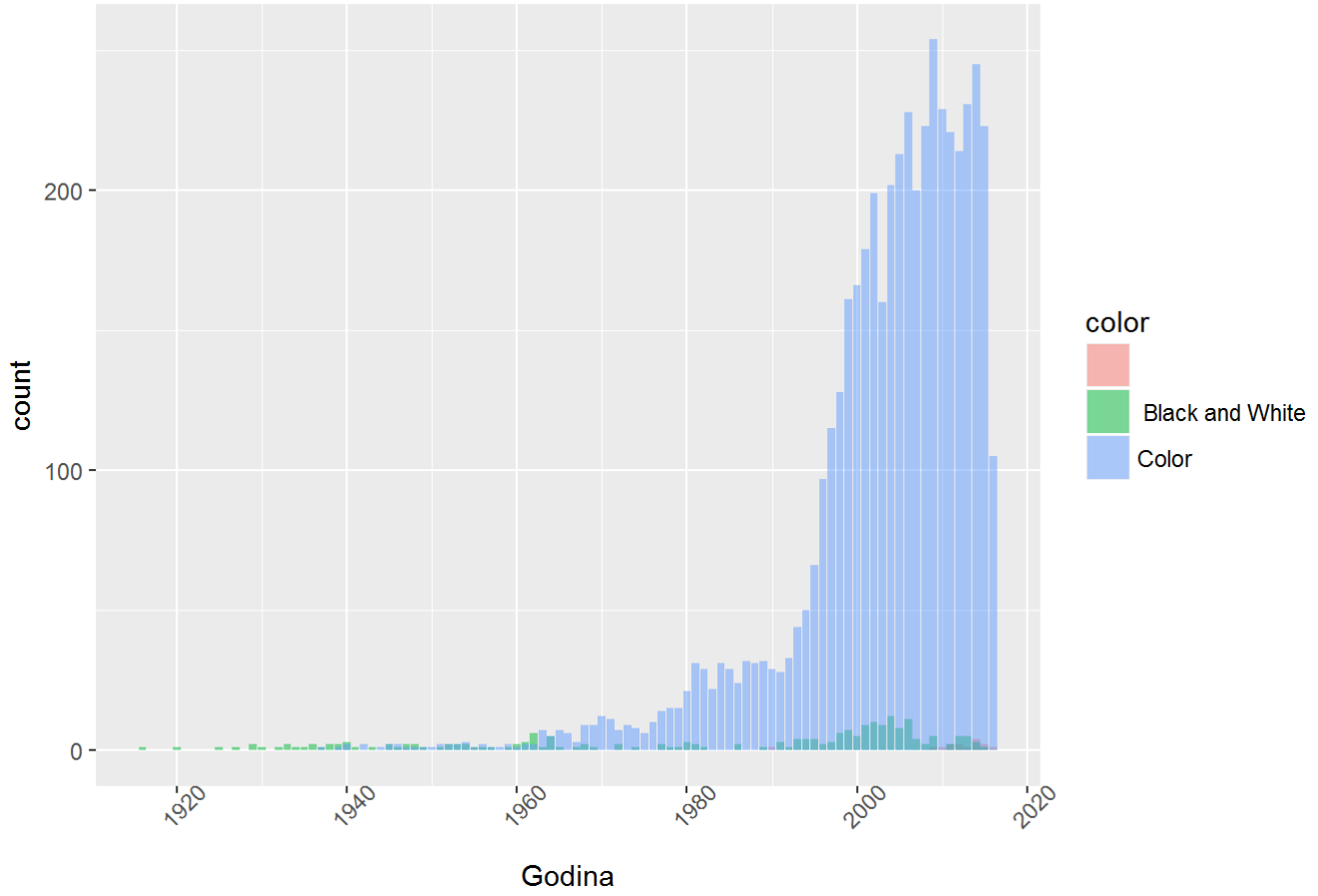

```
ggplot(data,aes(content_rating)) + geom_bar() + theme(axis.text.x = element_text(angle = 45))
 + labs(x="Prema ocjeni sadržaja", title="Broj filmova prema ocjeni sadržaja")
```




```
ggplot(data,aes(title_year, fill=color)) + geom_bar(position = "identity", alpha = 0.5) + the
me(axis.text.x = element_text(angle = 45)) + labs(x="Godina", title="Broj filmova prema boji
 u odnosu na godine.")
```

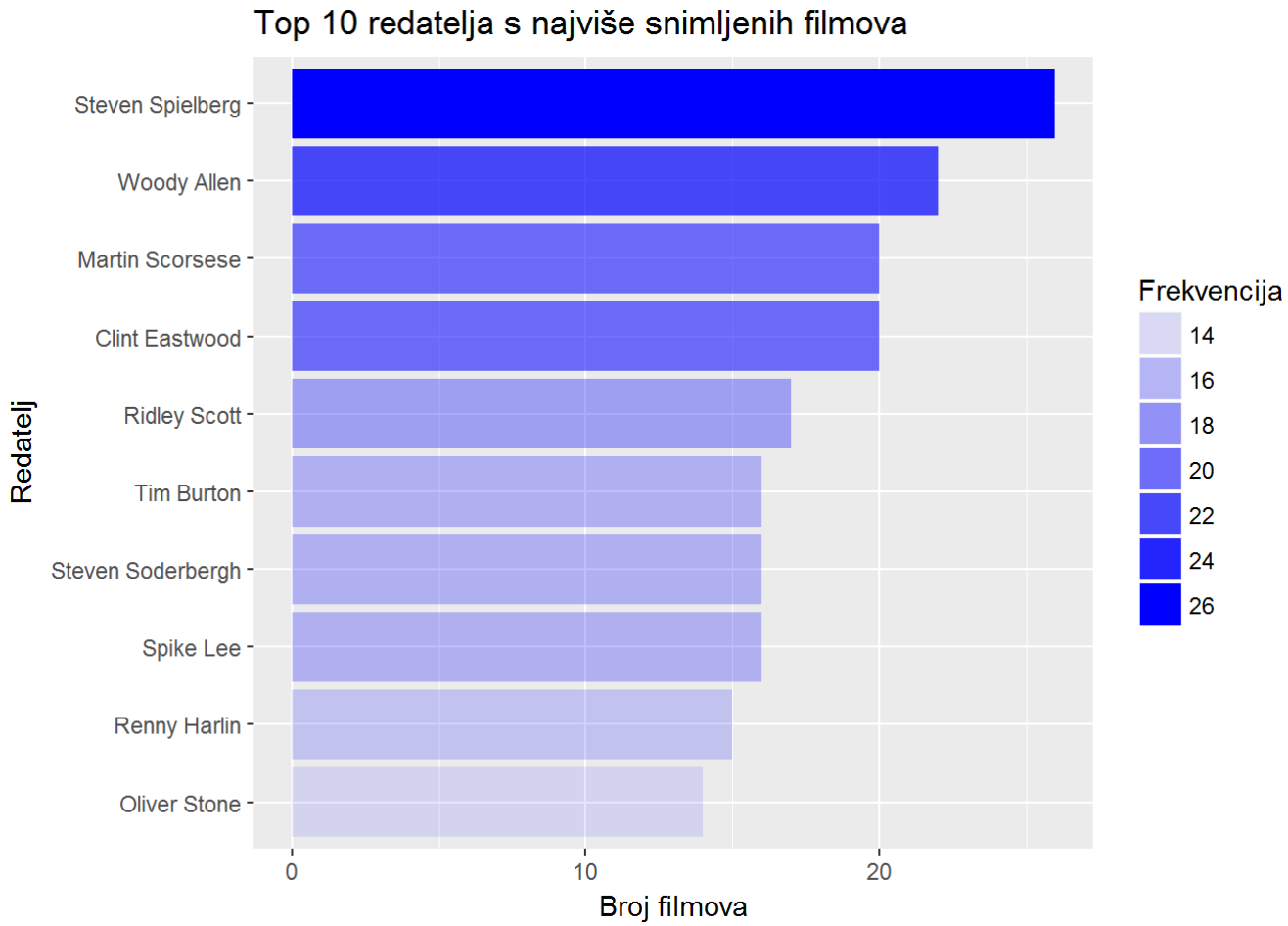
```
## Warning: Removed 108 rows containing non-finite values (stat_count).
```

Broj filmova prema boji u odnosu na godine.



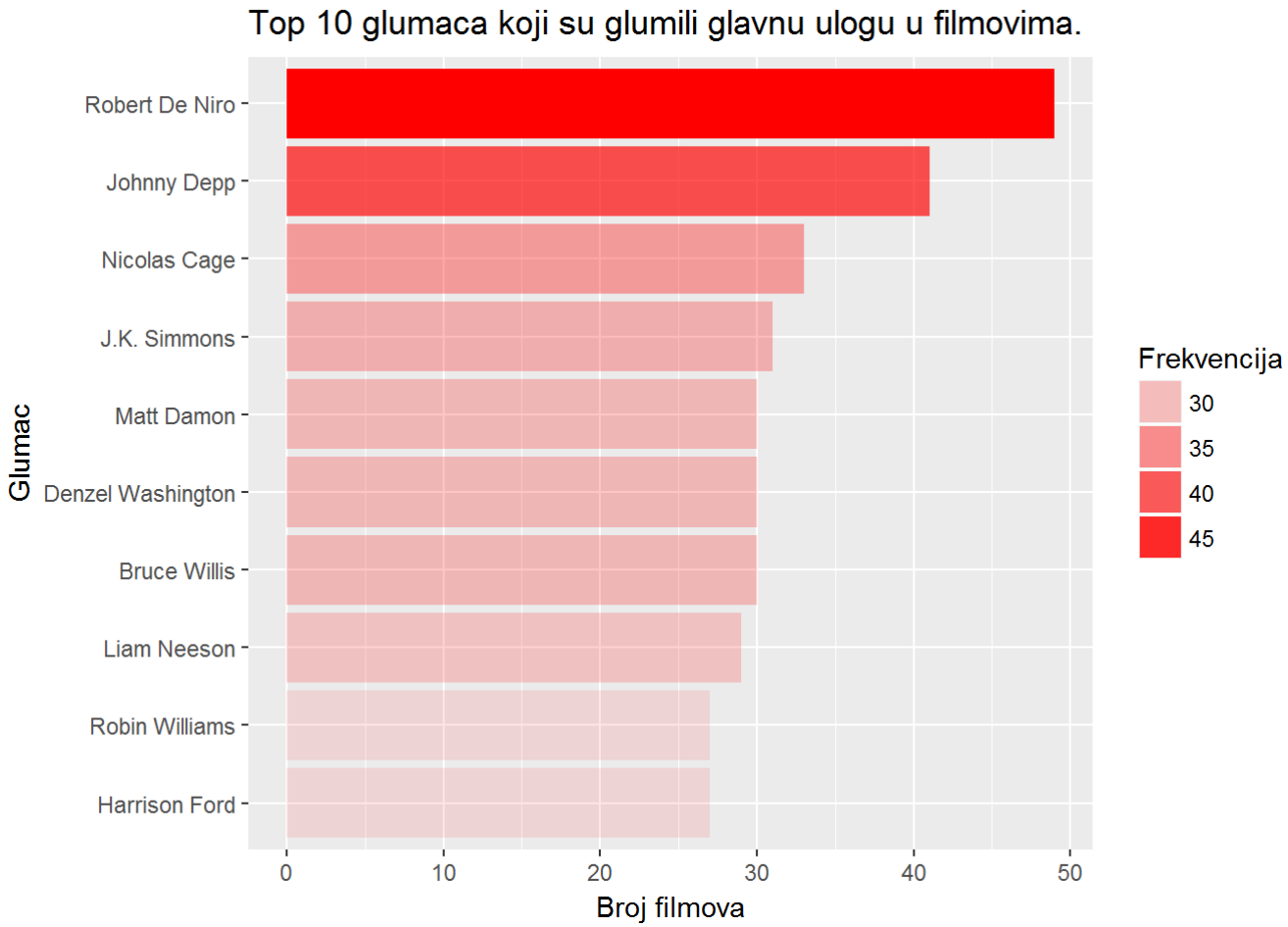

```
  director <- data.frame(table(data$director_name))
  director <- director[-c(1),]
  director <- director[order(director$Freq, decreasing = TRUE),]

  names(director)[1] <- "director_name"
```

```
  ggplot(director[1:10,], aes(x=reorder(factor(director_name), Freq), y=Freq, alpha=Freq)) +
geom_bar(stat="identity",fill="blue") + coord_flip() + labs(x="Redatelj", y = "Broj filmova",
 title="Top 10 redatelja s najviše snimljenih filmova" , alpha="Frekvencija")
```



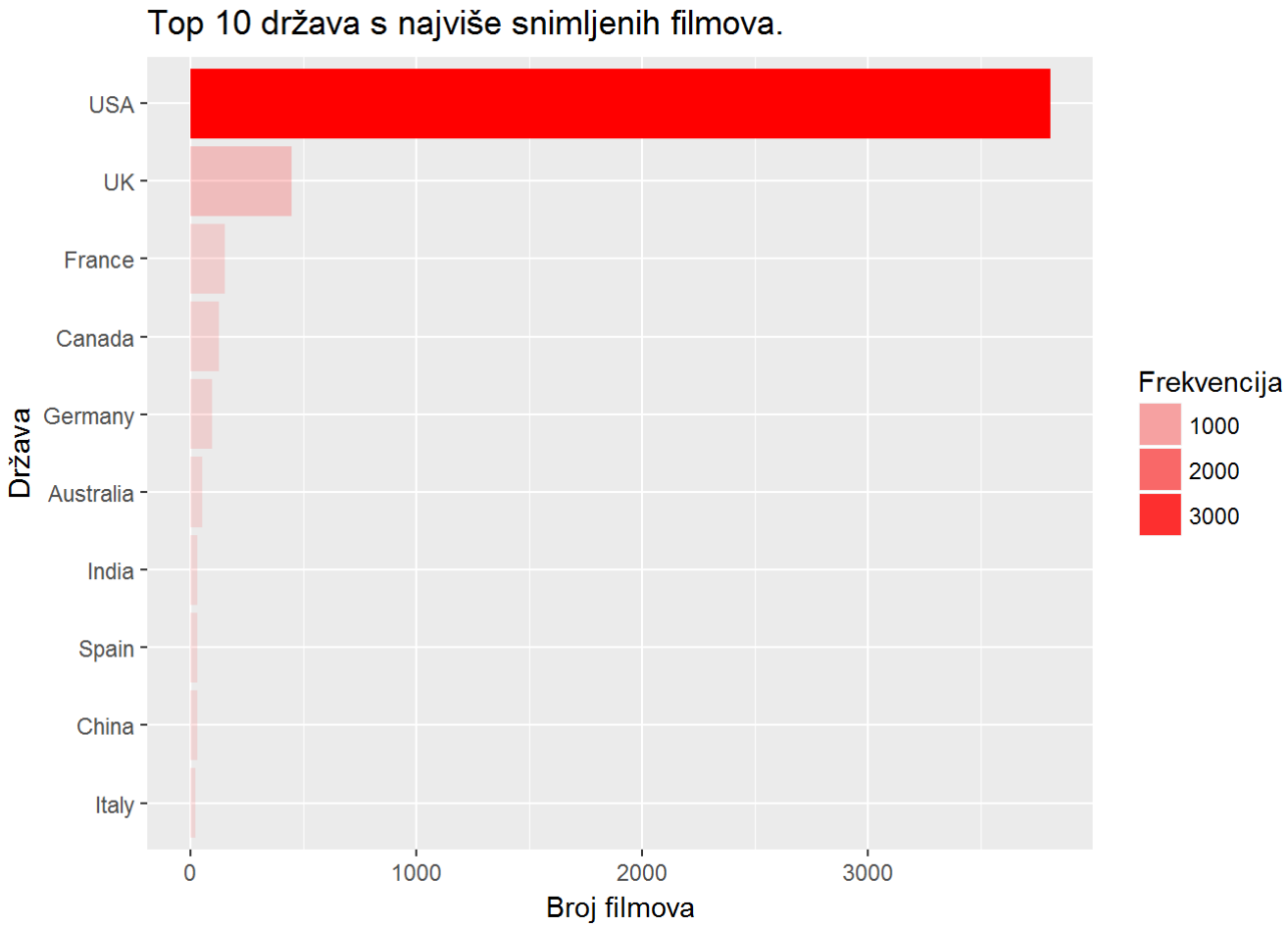

```
  actor_1 <- data.frame(table(data$actor_1_name))
  actor_1 <- actor_1[order(actor_1$Freq, decreasing = TRUE),]
  names(actor_1)[1] <- "actor_name"
```

```
  ggplot(actor_1[1:10,], aes(reorder(factor(actor_name),Freq), Freq, alpha = Freq)) + geom_ba
r(stat="identity",fill="red") + coord_flip() + labs(x="Glumac", y = "Broj filmova", title="To
p 10 glumaca koji su glumili glavnu ulogu u filmovima.", alpha="Frekvencija")
```



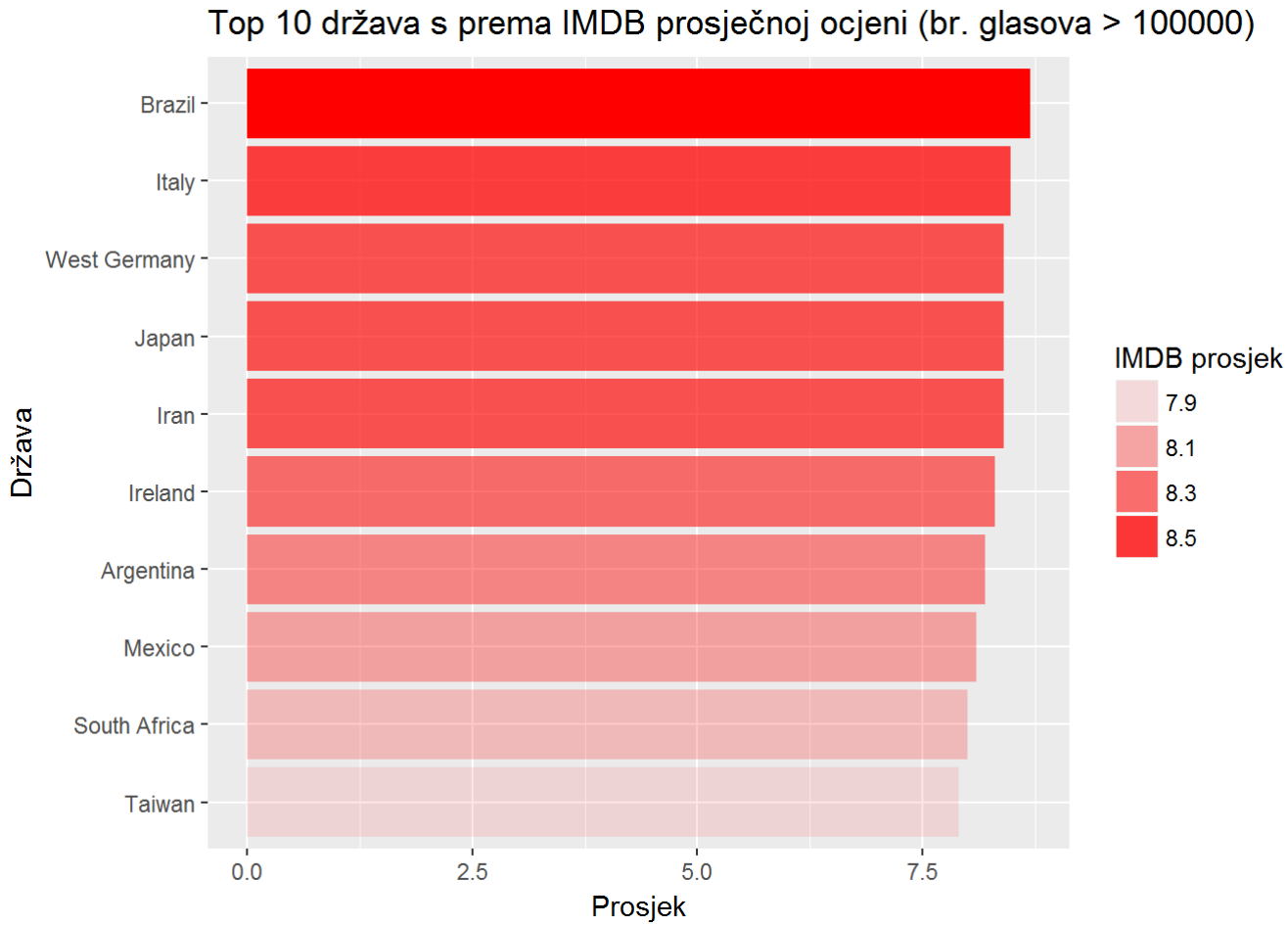

```
  country <- data.frame(table(data$country))
  country <- country[order(country$Freq, decreasing = TRUE),]
  names(country)[1] <- "country"
```

```
  ggplot(country[1:10,], aes(reorder(factor(country),Freq), Freq, alpha=Freq)) + geom_bar(sta
t="identity",fill="red") + coord_flip() + labs(x="Država", y="Broj filmova", title="Top 10 dr
žava s najviše snimljenih filmova.", alpha="Frekvencija")
```



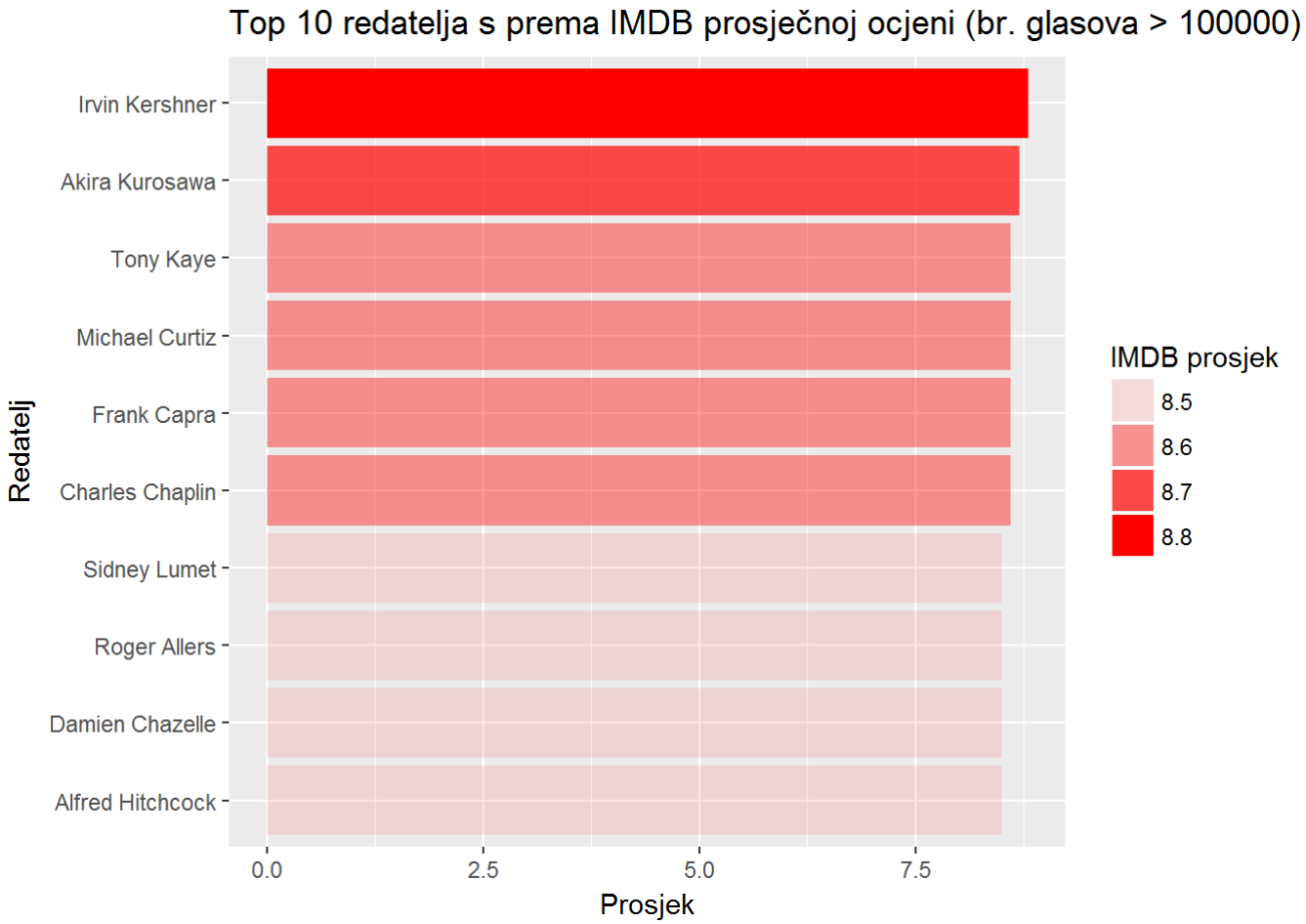

```
imdb_scores_country = as.data.table(subset(data, data$country != ''  & data$num_voted_users >
 100000))
  imdb_scores_country = imdb_scores_country[, mean(imdb_score), by=country]
  names(imdb_scores_country) = c("country", "average_score")

  imdb_scores_country = imdb_scores_country[order(imdb_scores_country$average_score, decreasi
ng = TRUE),]
```

```
  ggplot(imdb_scores_country[1:10,], aes(reorder(factor(country),average_score), average_scor
e, alpha=average_score)) + geom_bar(stat="identity",fill="red") + coord_flip() + labs(x="Drža
va", y="Prosjek", title="Top 10 država s prema IMDB prosječnoj ocjeni (br. glasova >
100000)", alpha="IMDB prosjek")
```



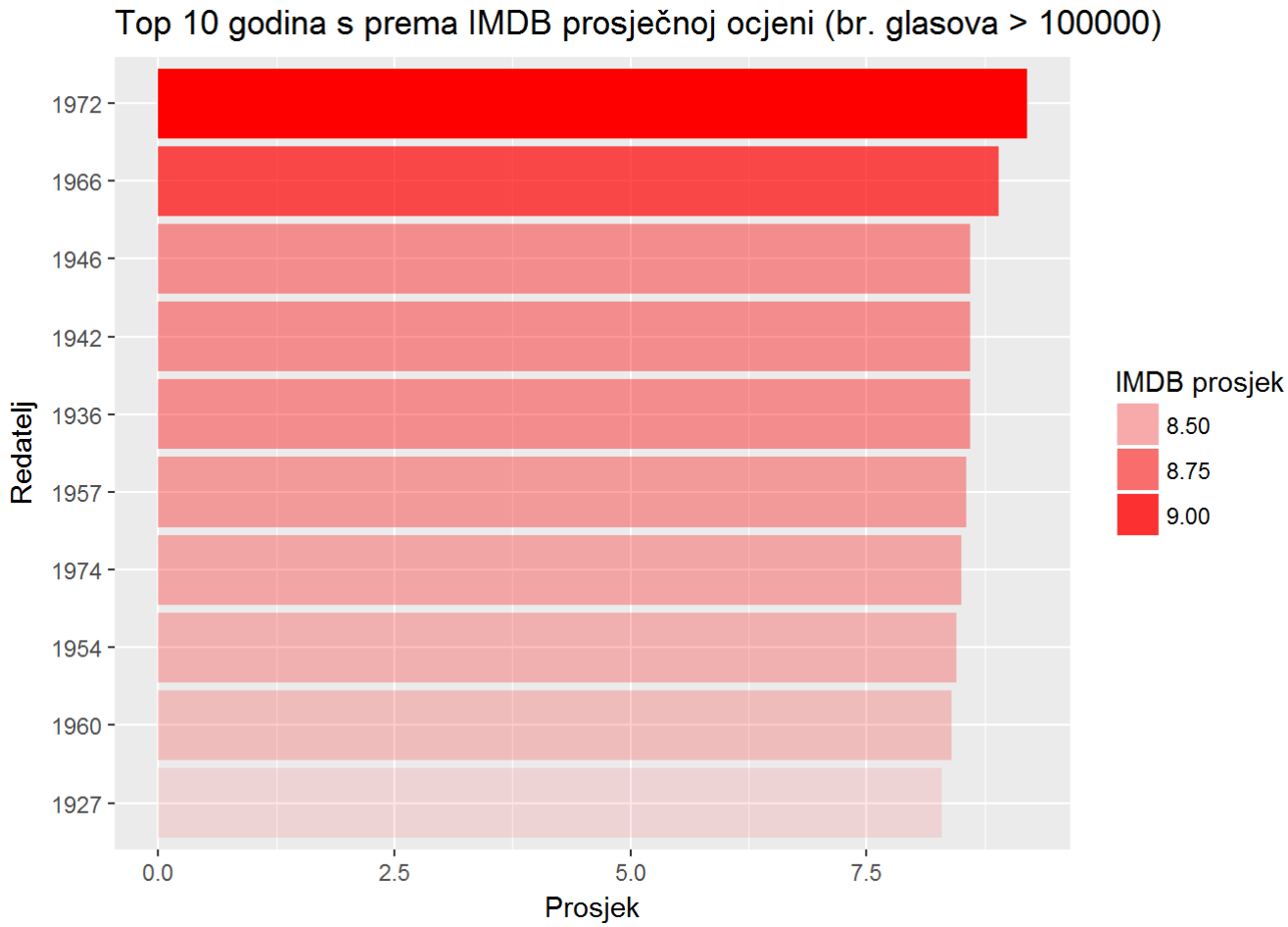

```
  imdb_scores_director = as.data.table(subset(data, data$director_name != ''  & data$num_vote
d_users > 100000))
  imdb_scores_director = imdb_scores_director[, mean(imdb_score), by=director_name]
  names(imdb_scores_director) = c("director_name", "average_score")

  imdb_scores_director = imdb_scores_director[order(imdb_scores_director$average_score, decre
asing = TRUE),]
```

```
  ggplot(imdb_scores_director[1:10,], aes(reorder(factor(director_name),average_score), avera
ge_score, alpha=average_score)) + geom_bar(stat="identity",fill="red") + coord_flip() +
labs(x="Redatelj", y="Prosjek", title="Top 10 redatelja s prema IMDB prosječnoj ocjeni (br. g
lasova > 100000) ", alpha="IMDB prosjek")
```




```
  imdb_scores_year = as.data.table(subset(data, data$title_year != '' & data$num_voted_users
> 100000))
  imdb_scores_year = imdb_scores_year[, mean(imdb_score), by=title_year]
  names(imdb_scores_year) = c("year", "average_score")

  imdb_scores_year = imdb_scores_year[order(imdb_scores_year$average_score, decreasing =
TRUE),]
```

```
  ggplot(imdb_scores_year[1:10,], aes(reorder(factor(year),average_score), average_score, alp
ha=average_score)) + geom_bar(stat="identity",fill="red") + coord_flip() + labs(x="Redatelj",
 y="Prosjek", title="Top 10 godina s prema IMDB prosječnoj ocjeni (br. glasova > 100000) ", a
lpha="IMDB prosjek")
```


Top 10 godina s prema IMDB prosječnoj ocjeni (br. glasova > 100000)
Redatelj
1972
1966
1946
1942
1936
1957
1974
1954
1960
1927
0.0
2.5
5.0
7.5
Prosjek
IMDB prosjek
8.50
8.75
9.00