Rutracker: Немного визуализации

Ульяна Сенцова

29 июня 2016 г

Загрузим данные. Для этого создадим сначала вектор классов колонок. Объединим данные по русским фильмам и по иностранным.

```
col.classes <- c("character", "character", "integer", "factor", "numeric", "numeric",
 "numeric", "numeric")
awards <- vector(mode="character", length=6)
awards[1:6] <- "integer"
col.classes <- c(col.classes, awards)
other <- awards <- vector(mode="character", length=29)
other[1:29] <- "factor"
col.classes <- c(col.classes, other)
col.classes.rus <- c(col.classes, "factor", "factor")

films <- read.csv("films_data.tsv", header = T, sep = "\t", colClasses = col.classes)
russian_films <- read.csv("russian_films.tsv", header=F, sep="\t", colClasses = col.c
lasses.rus)
films$melodrama <- as.factor(c("0"))
films$russia <- as.factor(c("0"))
colnames(russian_films) <- names(films)
film.data <- rbind(films, russian_films)
```

Теперь посмотрим на структуру наших данных:

```
str(film.data)
```

```
## 'data.frame':    10146 obs. of  45 variables:
##  $ id                      : chr  "tt1219289" "tt1284575" "tt0318403" "tt0307758"
 ...
##  $ title                   : chr  "Limitless" "Bad Teacher" "The Lion King 1 1/2"
 "Hodejegerne" ...
##  $ downloads               : int  342204 285162 254935 244397 244036 226320 215949
 214396 213984 197567 ...
##  $ year                    : Factor w/ 274 levels "1896","1909",..: 219 219 162 14
5 228 178 212 225 225 219 ...
##  $ idbm                    : num  7.4 5.7 6.6 4.5 7.1 8.4 7.4 6.8 6.4 6.8 ...
##  $ tomato_rating           : num  6.4 5.3 6.4 NA 7.3 NA 5.9 5.7 4.8 6.9 ...
##  $ tomato_user_rating      : num  3.7 2.9 3.2 NA 3.6 NA 3.5 3.6 3.1 3.5 ...
##  $ runtime                 : num  105 92 77 NA 106 22 133 115 114 117 ...
##  $ wins                    : int  2 6 6 0 0 25 0 1 0 2 ...
##  $ nominations             : int  7 3 10 0 4 87 4 7 6 7 ...
##  $ oscar_wins              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ oscar_nominations       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ golden_globe_wins       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ golden_globe_nominations: int  0 0 0 0 0 2 0 0 0 0 ...
##  $ usa                     : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 2 2 1 2 ...
##  $ canada                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ france                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 2 1 ...
##  $ uk.                     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ germany                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ european_country        : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ china                   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ asia                    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
##  $ other_country           : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1 1 ...
##  $ thriller                : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 1 1 1 1 ...
##  $ music                   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ drama                   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 1 2 2 ...
##  $ documentary             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ crime                   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 2 1 1 ...
##  $ history                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ animation               : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ fantasy                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ sci_fi                  : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
##  $ biography               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ romance                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 2 1 ...
##  $ war                     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ comedy                  : Factor w/ 2 levels "0","1": 1 2 2 1 1 2 1 1 1 1 ...
##  $ mystery                 : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 2 1 1 ...
##  $ adventure               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ western                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ action                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 2 ...
##  $ horror                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ family                  : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ short                   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ melodrama               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ russia                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Преобразование данных

Немного преобразуем наши данные. Во-первых, разберемся с годами. Как мы видим, год выпуска фильма встречается в двух форматах: в формате “год” и в формате “год начала-год окончания”. Для упрощения, создадим колонку start_year, в которую поместим год выпуска фильма либо год выпуска

первой части/серии фильма.

```
# Демонстрация различных форматов представления года выпуска:
head(film.data$year, 10)
```

```
##  [1] 2011      2011      2004      2002      2013      2005—2014 2010
##  [8] 2012      2012      2011
## 274 Levels: 1896 1909 1915 1917 1921 1924 1926 1927 1929 1931 1933 ... 2013–1080
```

```
N <- length(film.data$year)
start_year <- vector("integer", N)
for (i in 1:N) {
        if (film.data$year[i] == 4) {
                start_year[i] <- film.data$year[i]
                end_year[i] <- film.data$year[i]
        } else {
                x <- strsplit(as.character(film.data$year[i]), "[—–]")
                x <- unlist(x)
                start_year[i] <- x[1]
                }
        }
film.data$start_year <- as.factor(start_year)
```

Теперь, когда мы разобрались с годами, можно посмотреть на рейтинги. Рейтинг imdb и оба рейтинга сайта “Гнилые помидоры” представлены оценками от 0 до 10. Количество скачиваний на рутрекере - тоже своеобразный рейтинг, который нужно преобразовать. Сначала посмотрим на основные показатели колонки downloads.

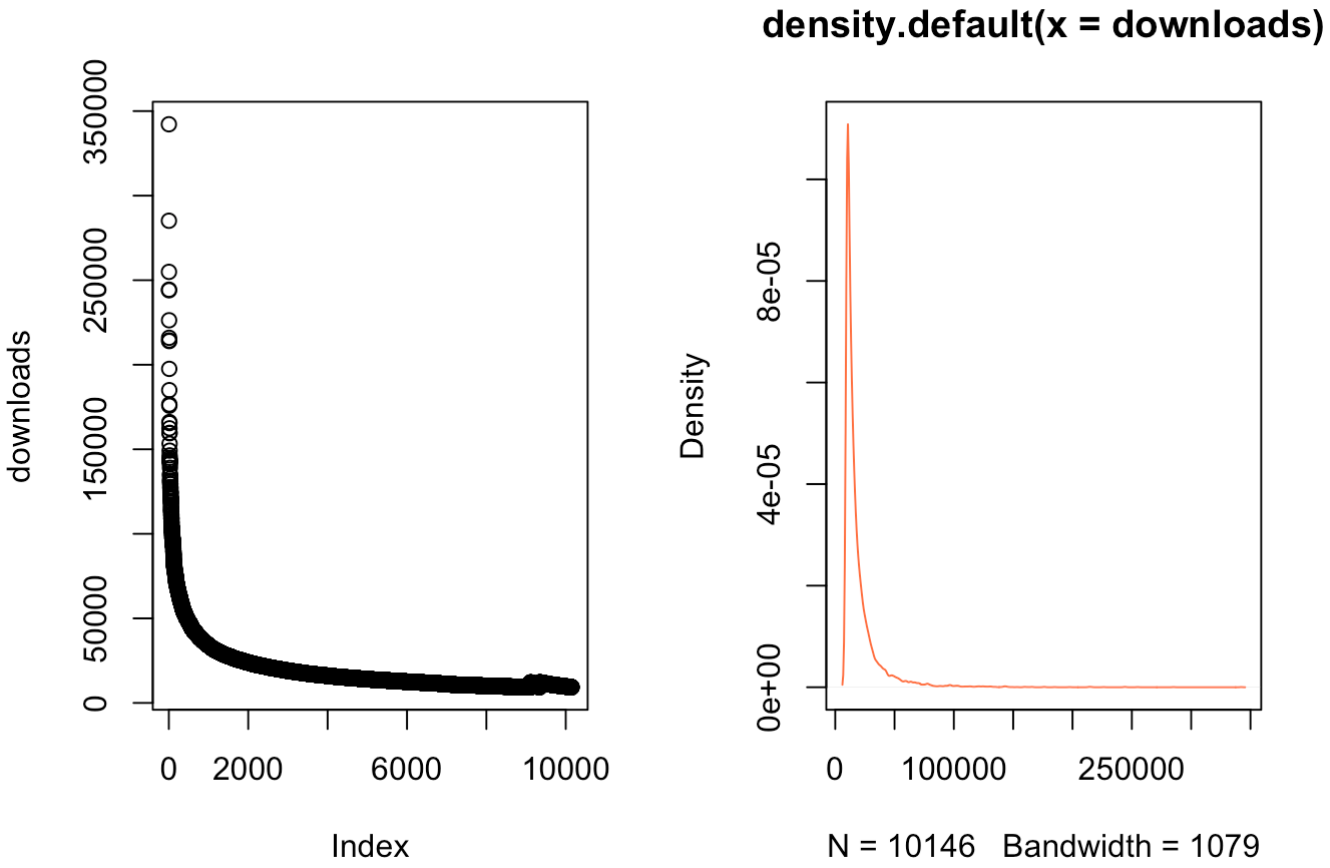
```
attach(film.data)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     start_year
```

```
summary(downloads)
```

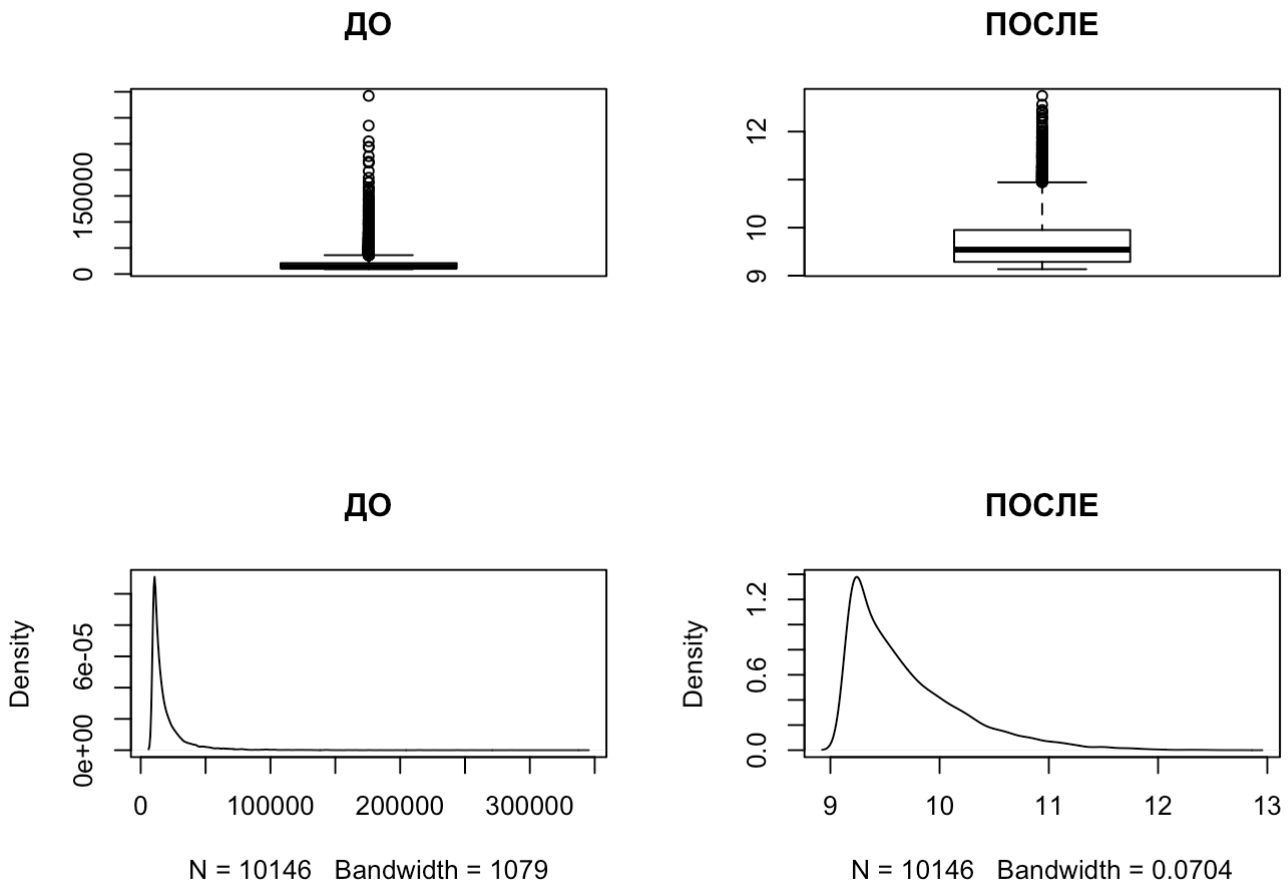
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9276   10800   13920   19450   20960  342200
```

```
par(mfrow = c(1,2))
plot(downloads)
plot(density(downloads), col="coral")
```



Как мы видим из обоих графиков, данные значительно skewed. Попробуем выполнить логарифмическое преобразование:

```
par(mfrow=c(2,2))
boxplot(downloads, main="ДО")
boxplot(log(downloads), main="ПОСЛЕ")
plot(density(downloads), main="ДО")
plot(density(log(downloads)), main="ПОСЛЕ")
```




```
downloads <- log(downloads)
summary(downloads)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.135   9.287   9.541   9.696   9.951  12.740
```

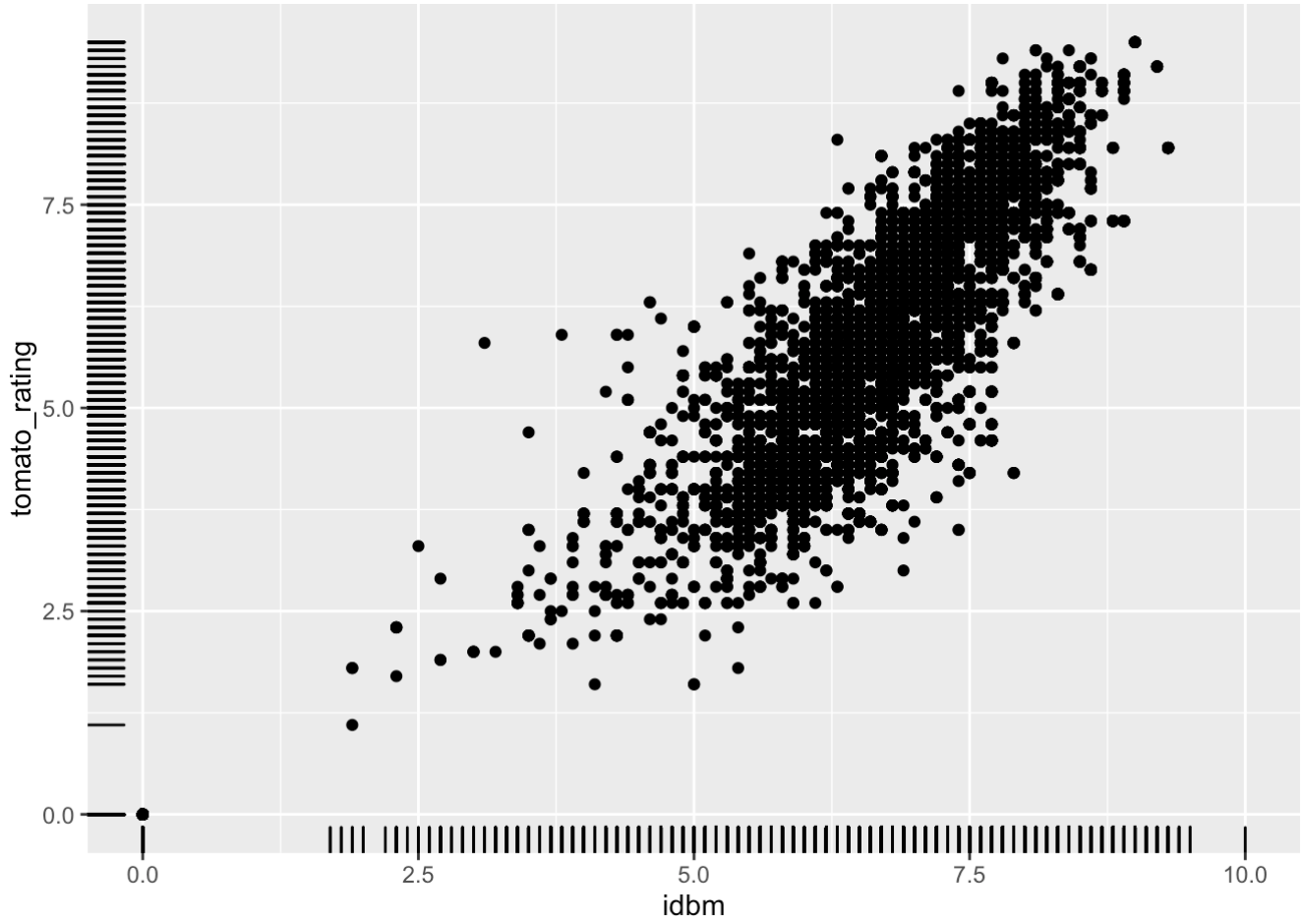
```
film.data$rutracker_rating <- log(downloads)
detach(film.data)
```

Зависимость между разными рейтингами

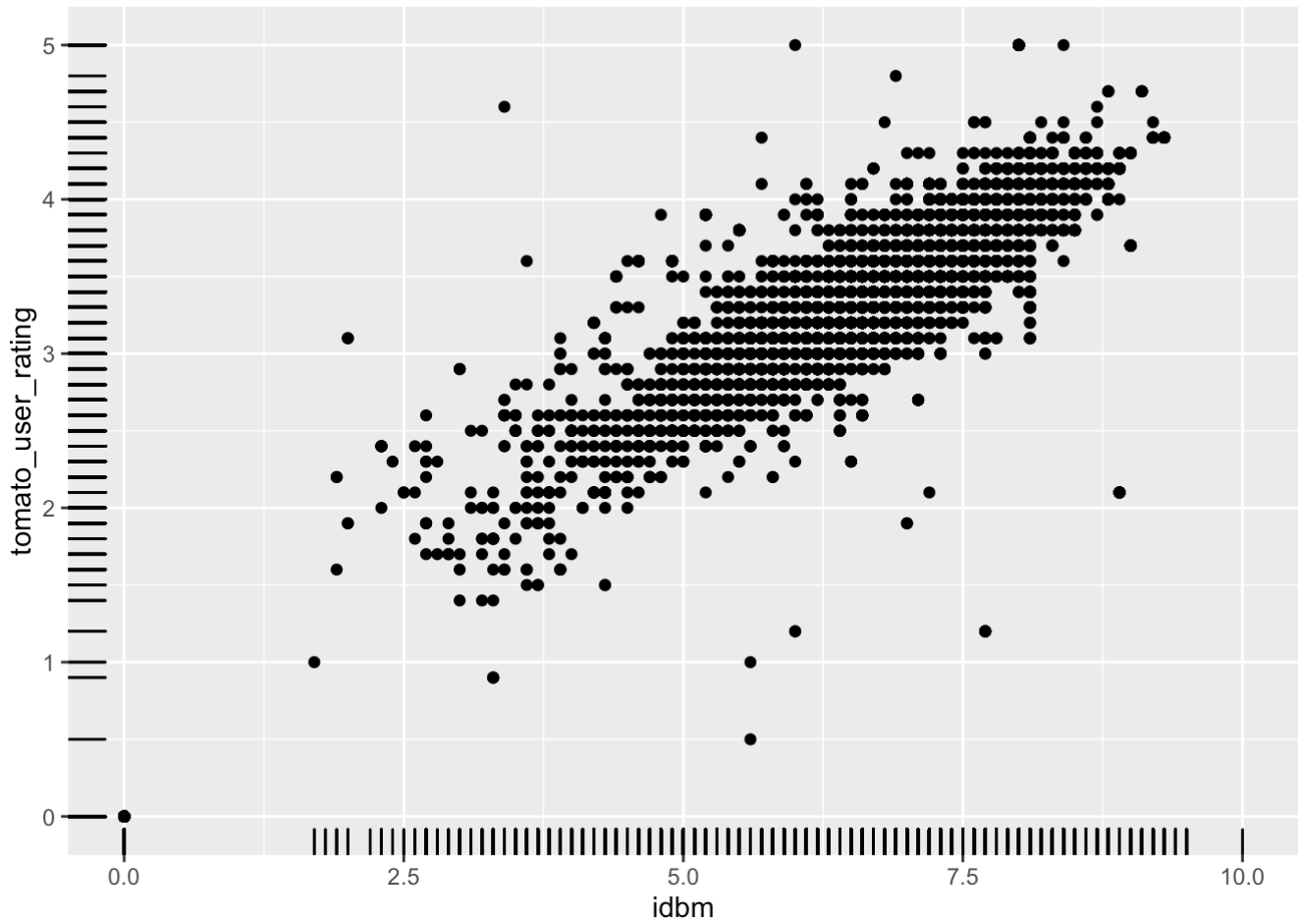
Ура! Теперь всё выглядит намного лучше. Интересно, есть ли зависимость между различными международными рейтингами (idbm и rotten tomatoes рейтинг)?

```
## The following objects are masked _by_ .GlobalEnv:
##
##     downloads, start_year
```

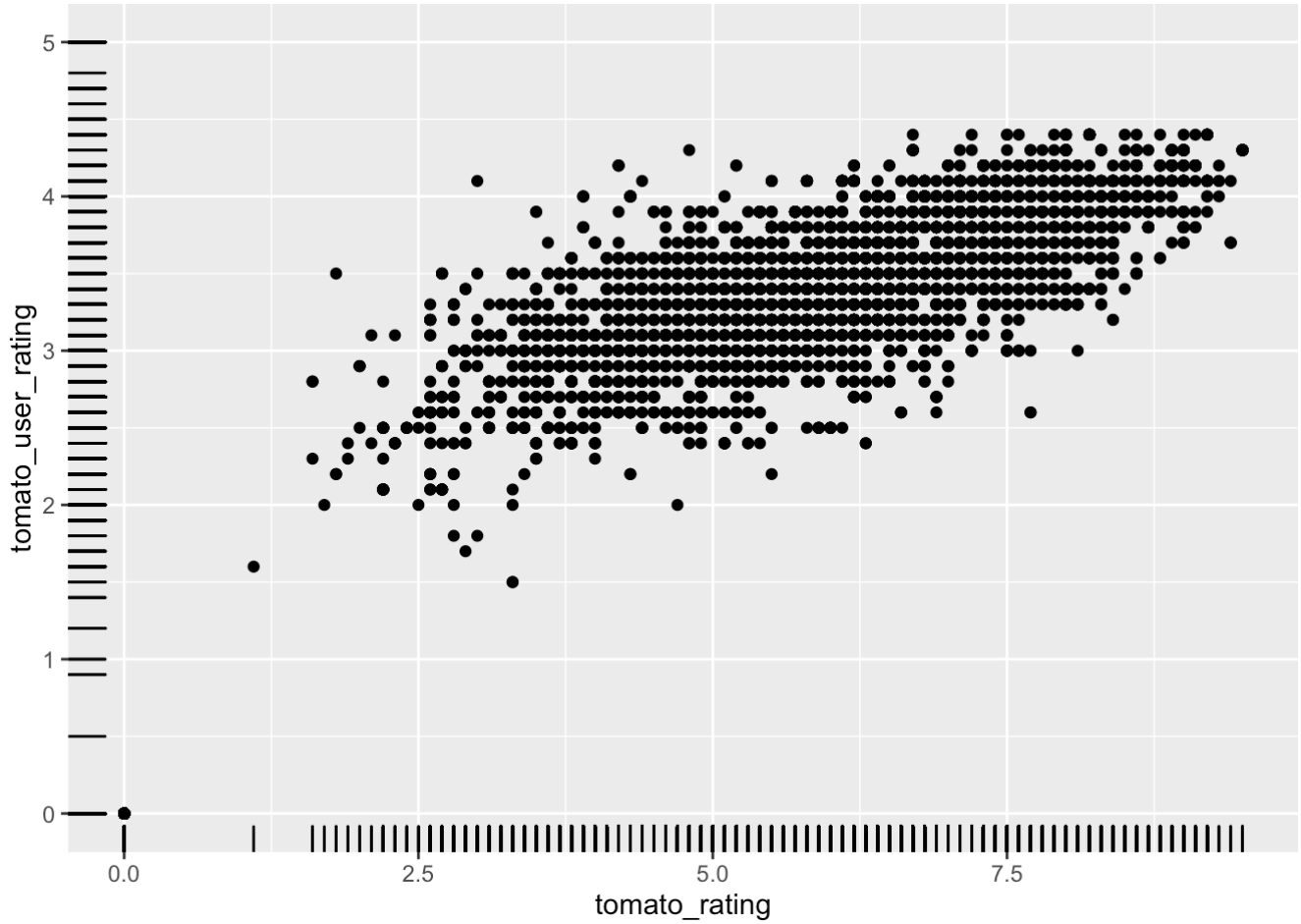
```
## Warning: Removed 3427 rows containing missing values (geom_point).
```




```
## Warning: Removed 2501 rows containing missing values (geom_point).
```



```
## Warning: Removed 3430 rows containing missing values (geom_point).
```



Да, действительно, похоже, что рейтинги неплохо работают. Прослеживается линейная зависимость.

Новинки или старые фильмы?

Вернемся к годам. Например, можно посмотреть, есть ли зависимость от новизны фильма и желанием рутрекерщиков скачивать этот фильм. Для этого преобразуем года в группы по пять лет.

```
attach(film.data)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     downloads, start_year
```

```
## The following objects are masked from film.data (pos = 3):
##
##     action, adventure, animation, asia, biography, canada, china,
##     comedy, crime, documentary, downloads, drama,
##     european_country, family, fantasy, france, germany,
##     golden_globe_nominations, golden_globe_wins, history, horror,
##     id, idbm, melodrama, music, mystery, nominations,
##     oscar_nominations, oscar_wins, other_country, romance,
##     runtime, russia, rutracker_rating, sci_fi, short, start_year,
##     thriller, title, tomato_rating, tomato_user_rating, uk., usa,
##     war, western, wins, year
```

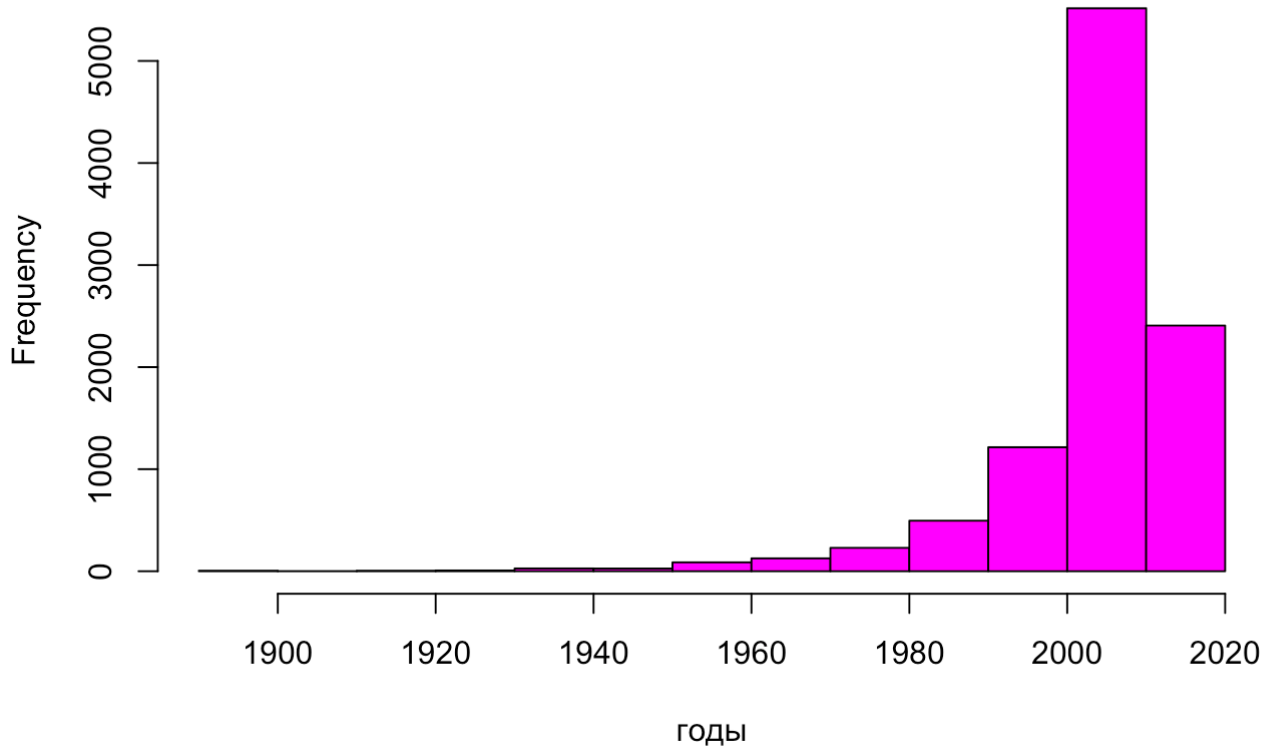
```
# Посмотрим на данные по количеству фильмов в разные годы:
table(start_year)
```

```
## start_year
## 1896 1909 1915 1917 1921 1924 1926 1927 1929 1931 1933 1934 1935 1936 1937
##    5    1    3    2    1    1    2    2    1    1    1    1    2    1    6
## 1938 1939 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953
##    3   12    5    4    5    1    1    3    1    2    1    3    1    9    7
## 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968
##    6    4    8   14   10   19    9   15    5    5   17   13    8   15   18
## 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983
##   17   13   21   16   25   20   30   20   27   12   37   21   44   40   22
## 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998
##   69   42   40   47   46   71   74   55   69   64  132  103   65  167  159
## 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
##  212  189  242  240  294  401  481  530  620  725  849 1134 1198  760  355
## 2014 2015 2016 2018
##   51   37    5    1
```

```
# у нас есть явный выброс: фильм, датируемый 2018-м годом. Оказывается, этот фильм на самом д
еле снят в 1978 году. Заменяем.
film.data[film.data$start_year == 2018,]$start_year <- 1978
```

Из гистограммы по годам уже видно, что фильмы конца 90-х и нулевых - самые популярные. Разбиваем фильмы по пятилеткам. При этом уберем фильмы до 1940 года: их слишком мало.

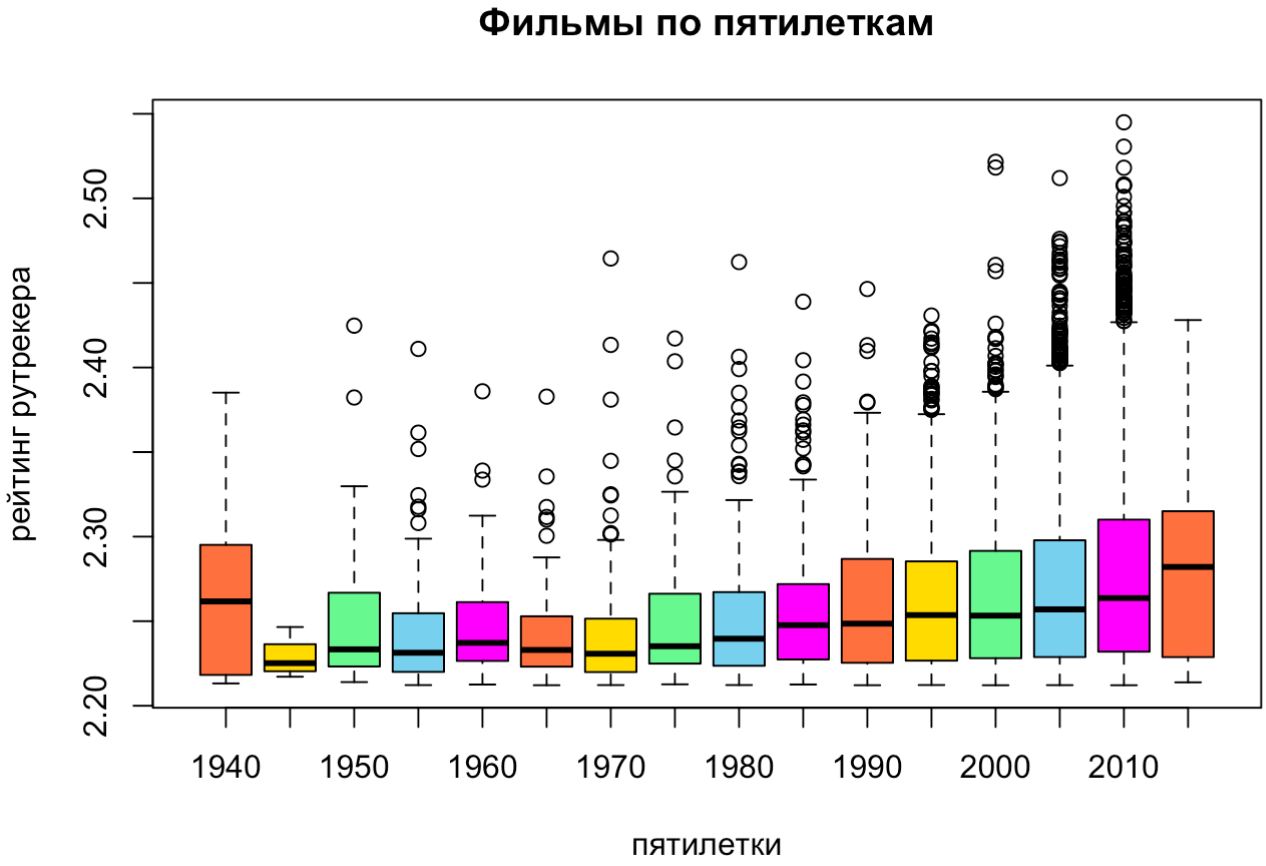
```
start_year <- as.integer(start_year)
hist(start_year, main="Популярность фильмов по годам", xlab="годы", col="magenta")
```




```
# Теперь разобьем все фильмы по "пятилеткам".
film.data$start_year <- as.integer(start_year)
films.after.1935 <- subset(film.data, film.data$start_year >= 1940)
year.group <- floor(films.after.1935$start_year/5)*5
year.group <- as.factor(year.group)
table(year.group)
```

```
## year.group
## 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010
##   15    8   26   55   51   71   95  126  196  246  394  706 1366 3205 3498
## 2015
##   43
```

```
plot(films.after.1935$rutracker_rating ~ year.group, col=c("coral", "gold", "lightgre
en", "skyblue", "magenta"), main="Фильмы по пятилеткам", xlab = "пятилетки",ylab="рейтинг р
утрекера")
```



Некоторая зависимость есть: в среднем, начиная с 1970 года, фильмы становятся все более “скачиваемыми”. Много скачиваний также у фильмов из промежутка от 1940 до 1945 года. Интересно посмотреть, что это за фильмы? Сравним их с фильмами, которые очень мало скачиваются на рутрекере, а именно с группой фильмов, снятых с 1945 по 1950.

```
film.data[film.data$start_year >= 1940 & film.data$start_year < 1945, c("title", "sta
rt_year", "idbm", "tomato_rating", "downloads")]
```

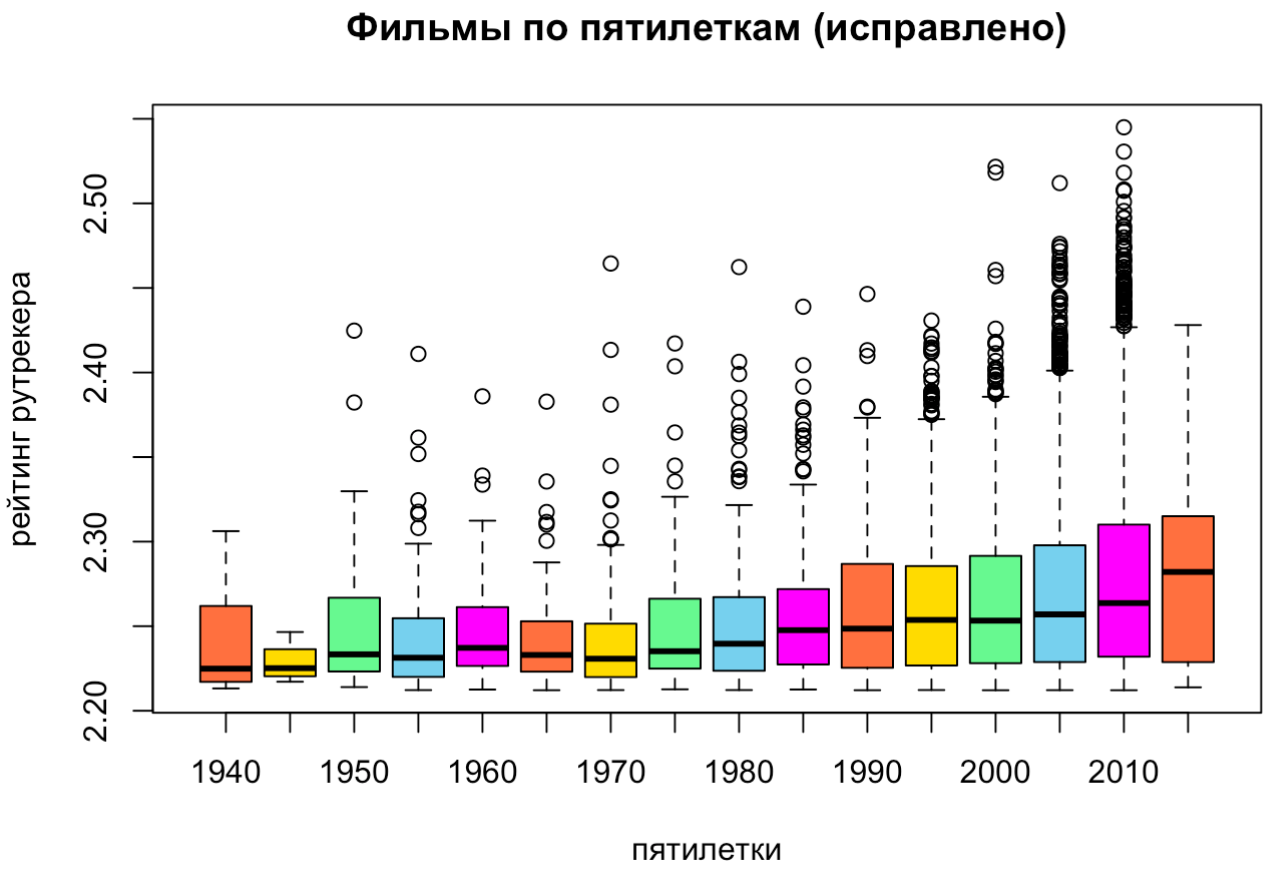
```
##                   title start_year idbm tomato_rating downloads
## 413      La vita è bella       1943  7.9            NA     52080
## 997      La vita è bella       1943  7.9            NA     34110
## 1483     La vita è bella       1943  7.9            NA     27748
## 2162          Casablanca       1942  8.6           9.3     22838
## 3192     La vita è bella       1943  7.9            NA     18333
## 3981       Meet John Doe       1941  7.7           7.5     16116
## 4569          Casablanca       1942  8.6           9.3     14829
## 4611               Bambi       1942  7.4           8.3     14754
## 6416        Citizen Kane       1941  8.4           9.4     12003
## 7656          Gran Hotel       1944  7.3            NA     10429
## 8304     Spooks Run Wild       1941  6.1            NA      9854
## 8421               Bambi       1942  7.4           8.3      9763
## 8574               Dumbo       1941  7.3           8.3      9650
## 8962         Dive Bomber       1941  6.6            NA      9367
## 9326 The Ox-Bow Incident       1943  8.1           8.1      9443
```

```
film.data[film.data$start_year >= 1945 & film.data$start_year < 1950, c("title", "sta
rt_year","idbm", "tomato_rating", "downloads")]
```

```
##                                      title start_year idbm tomato_rating
## 5847                       Mildred Pierce       1945    8            7.7
## 6583                            Notorious       1946    8            8.9
## 7192                     Racketeer Rabbit       1946    8             NA
## 9510                    Снеговик-почтовик       1948   NA             NA
## 9878  Сказки русских писателей. Выпуск III       1948   NA             NA
## 9892                  Волк и семеро козлят       1946   NA             NA
## 9920                                Весна        1947   NA             NA
## 10005          Приключения Домовёнка Кузи       1949   NA             NA
##       downloads
## 5847      12772
## 6583      11803
## 7192      10840
## 9510      11416
## 9878      10092
## 9892      10045
## 9920       9968
## 10005      9714
```

А вот и неправильные данные. Действительно, существует фильм “La vita è bella”, снятый в 1943 году, но он далеко не такой популярный, как фильм с тем же названием 1997 года, получивший множество наград. Если поменять год у этого фильма, то всё встанет на свои места.

```
film.data$start_year <- as.integer(start_year)
film.data[film.data$title == "La vita è bella",]$start_year <- 1997
films.after.1935 <- subset(film.data, film.data$start_year >= 1940)
year.group <- floor(films.after.1935$start_year/5)*5
year.group <- as.factor(year.group)
plot(films.after.1935$rutracker_rating ~ year.group, col=c("coral", "gold", "lightgre
en", "skyblue", "magenta"), main="Фильмы по пятилеткам (исправлено)", xlab = "пятилетки",yl
ab="рейтинг рутрекера")
```




```
detach(film.data)
```

Ну вот, отлично. Теперь нет резких скачков популярности фильмов в те или иные годы, и рейтинг рутрекера плавно увеличивается во времени. Это логично, так как рейтинг рутрекера - всего лишь нормализованное количество загрузок фильмов. В этом плане интересно посмотреть на выборы над верхним “усом” боксплотов. Эти фильмы были скачаны наибольшее количество раз, и тут очевидна любовь посетителей рутрекера к фильмам, выпущенными в последнее пятнадцатилетие.

Награды фильмов и их рейтинг

Теперь посмотрим по наградам. Зависит ли желание скачивать фильмы от количества наград, полученных этим фильмом? Посмотрим на фильмы, у которых наград в среднем больше, чем у остальных фильмов. Больше ли среднее количество скачиваний в этой группе, чем в группе у остальных фильмов?

```
attach(film.data)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     downloads, start_year
```

```
## The following objects are masked from film.data (pos = 3):
##
##     action, adventure, animation, asia, biography, canada, china,
##     comedy, crime, documentary, downloads, drama,
##     european_country, family, fantasy, france, germany,
##     golden_globe_nominations, golden_globe_wins, history, horror,
##     id, idbm, melodrama, music, mystery, nominations,
##     oscar_nominations, oscar_wins, other_country, romance,
##     runtime, russia, rutracker_rating, sci_fi, short, start_year,
##     thriller, title, tomato_rating, tomato_user_rating, uk., usa,
##     war, western, wins, year
```

```
film.data$sum.awards <- wins + nominations + oscar_nominations + oscar_wins + golden_
globe_nominations + golden_globe_wins

#среднее количество наград:
mean(film.data$sum.awards,na.rm = T)
```

```
## [1] 35.55749
```

```
many_awards <- film.data[is.na(film.data$sum.awards)==F & film.data$sum.awards > 35,
c("title", "start_year", "downloads", "rutracker_rating")]
few_awards <- film.data[is.na(film.data$sum.awards)==F & film.data$sum.awards <= 35,
c("title", "start_year", "downloads", "rutracker_rating")]

mean(many_awards$downloads) - mean(few_awards$downloads) < 1*sd(downloads)
```

```
## [1] FALSE
```

```
mean(log(many_awards$downloads)) - mean(log(few_awards$downloads)) < 1*sd(log(downloa
ds))
```

```
## [1] FALSE
```

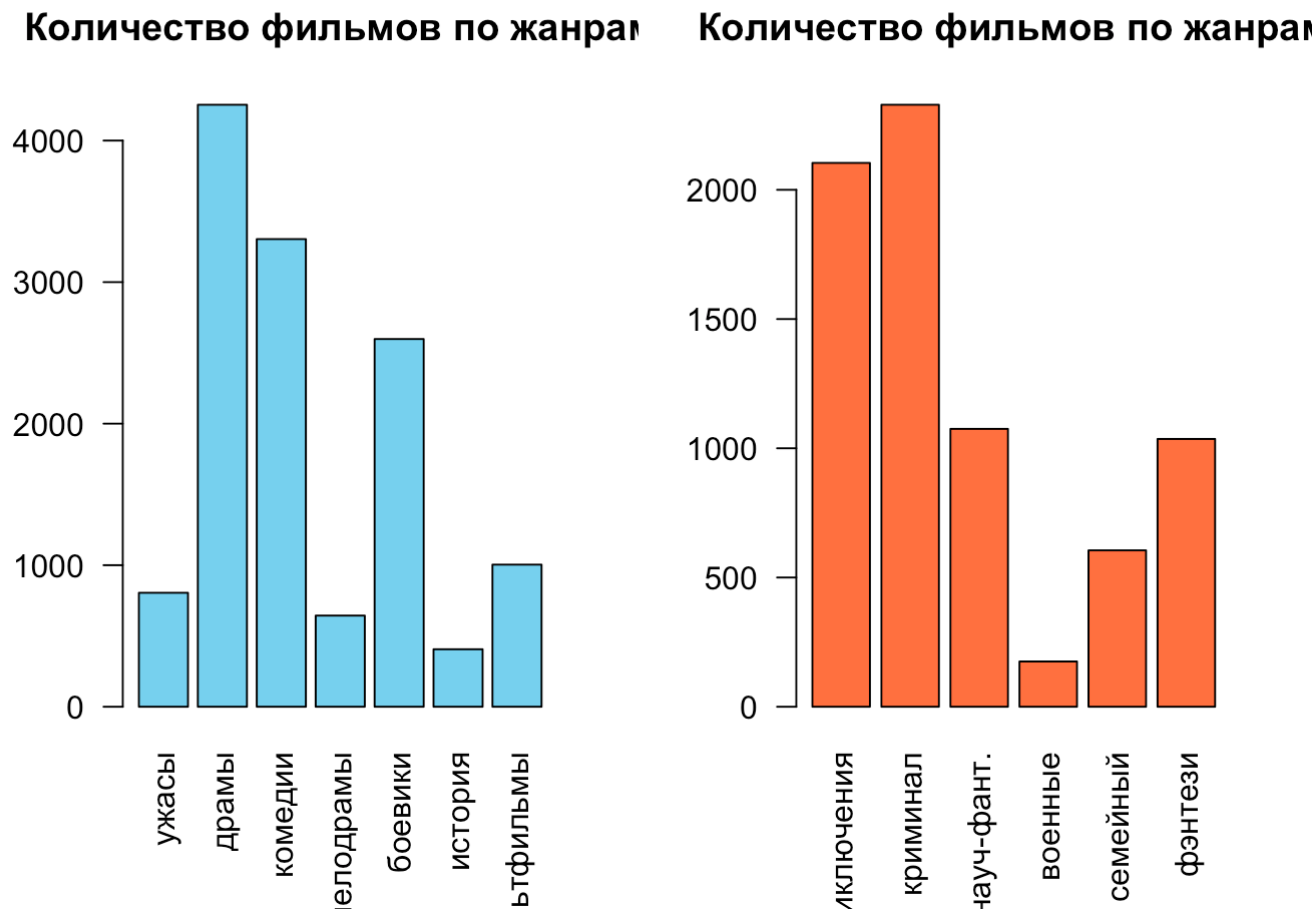
В среднем, их фильмы без наград действительно скачивают немного меньше. Однако это незначительное изменение: все в пределах одного стандартного отклонения.

Жанры рутрекера

Теперь посмотрим на жанры. Какие жанры самые чаще всего встречаются в подборке рутрекера?

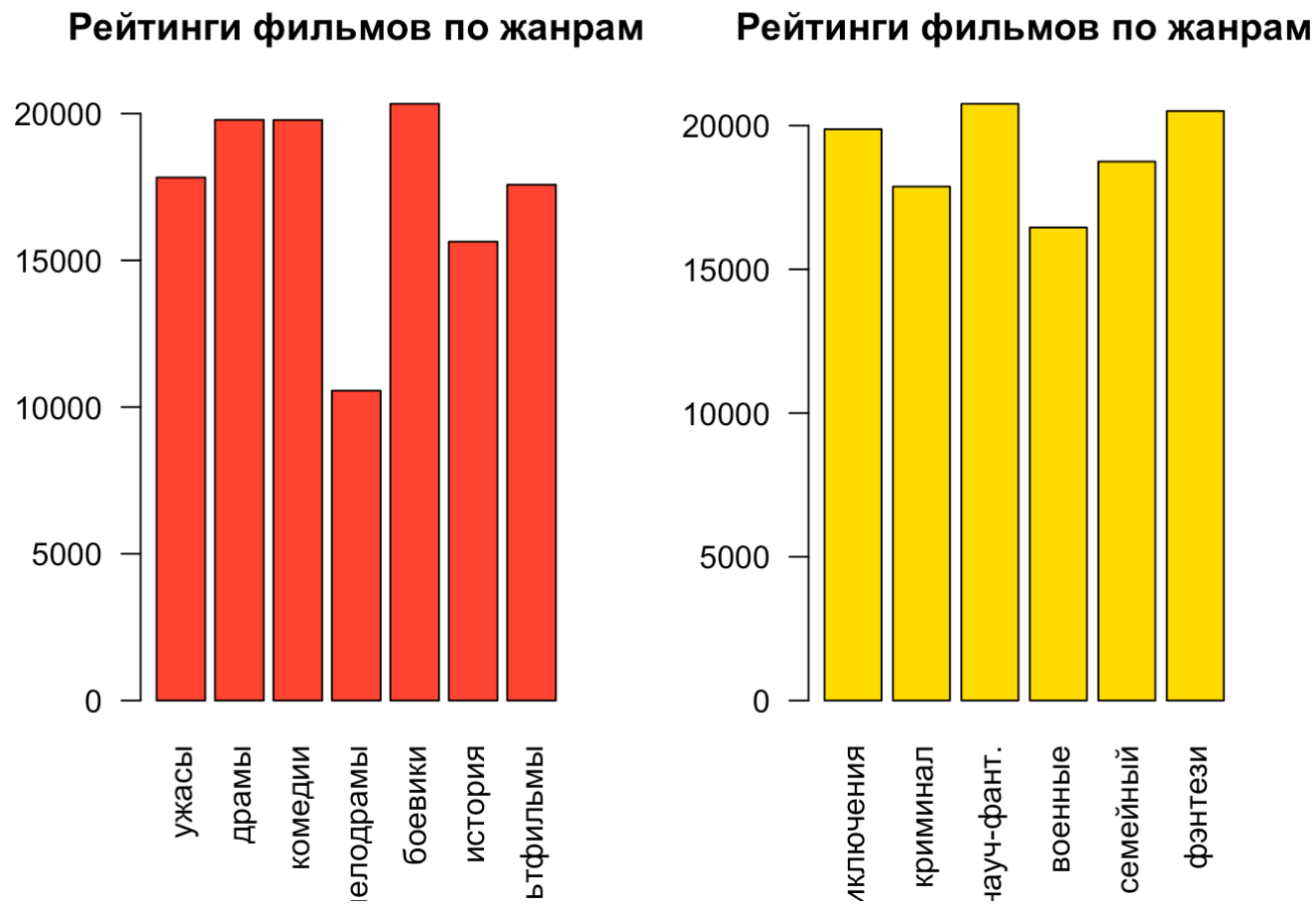
```
## [1] 14347572
```

```
## [1] 805
```



Теперь посмотрим не на количество фильмов разных жанров, а на их рутрекеровский рейтинг. Какие жанры самые популярные на рутрекере?

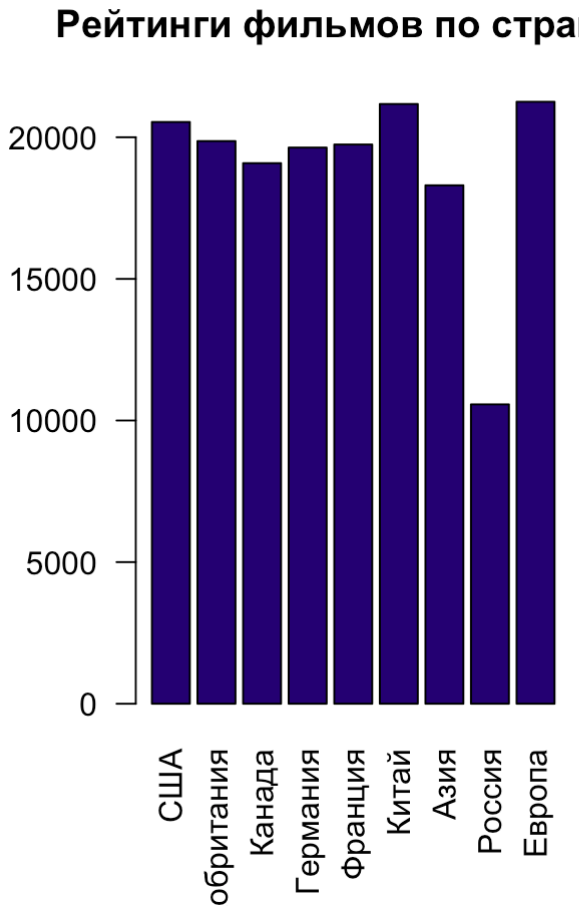
```
##  [1] "id"                    "title"
##  [3] "downloads"             "year"
##  [5] "idbm"                  "tomato_rating"
##  [7] "tomato_user_rating"    "runtime"
##  [9] "wins"                  "nominations"
## [11] "oscar_wins"            "oscar_nominations"
## [13] "golden_globe_wins"     "golden_globe_nominations"
## [15] "usa"                   "canada"
## [17] "france"                "uk."
## [19] "germany"               "european_country"
## [21] "china"                 "asia"
## [23] "other_country"         "thriller"
## [25] "music"                 "drama"
## [27] "documentary"           "crime"
## [29] "history"               "animation"
## [31] "fantasy"               "sci_fi"
## [33] "biography"             "romance"
## [35] "war"                   "comedy"
## [37] "mystery"               "adventure"
## [39] "western"               "action"
## [41] "horror"                "family"
## [43] "short"                 "melodrama"
## [45] "russia"                "start_year"
## [47] "rutracker_rating"      "sum.awards"
```

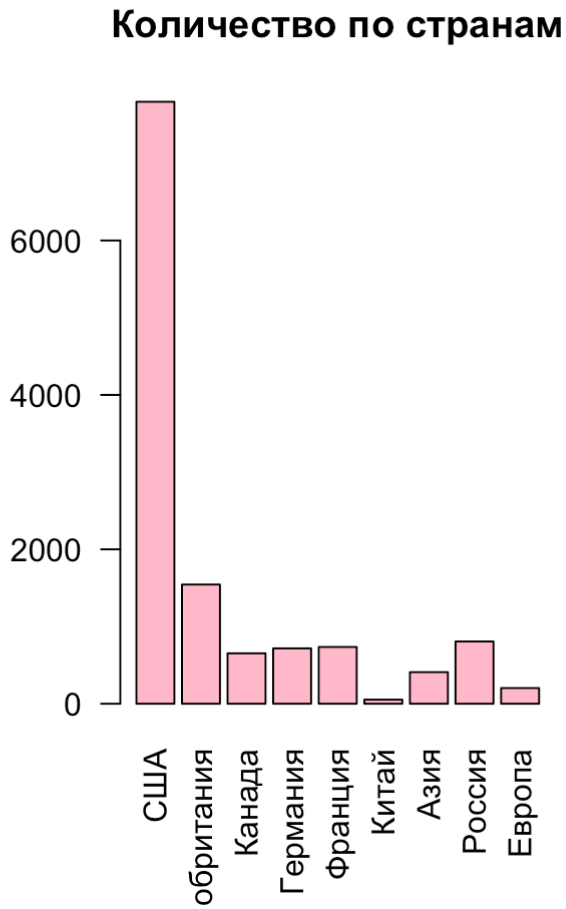


Рутрекер и фильмы по странам

А теперь подборка по странам. Посмотрим и на рейтинг, и на количество фильмов в подборке рутрекера.

```
##  [1] "id"                    "title"
##  [3] "downloads"             "year"
##  [5] "idbm"                  "tomato_rating"
##  [7] "tomato_user_rating"    "runtime"
##  [9] "wins"                  "nominations"
## [11] "oscar_wins"            "oscar_nominations"
## [13] "golden_globe_wins"     "golden_globe_nominations"
## [15] "usa"                   "canada"
## [17] "france"                "uk."
## [19] "germany"               "european_country"
## [21] "china"                 "asia"
## [23] "other_country"         "thriller"
## [25] "music"                 "drama"
## [27] "documentary"           "crime"
## [29] "history"               "animation"
## [31] "fantasy"               "sci_fi"
## [33] "biography"             "romance"
## [35] "war"                   "comedy"
## [37] "mystery"               "adventure"
## [39] "western"               "action"
## [41] "horror"                "family"
## [43] "short"                 "melodrama"
## [45] "russia"                "start_year"
## [47] "rutracker_rating"      "sum.awards"
```


Рейтинги фильмов по странам
20000
15000
10000
5000
0
США
обритания
Канада
Германия
Франция
Китай
Азия
Россия
Европа


Количество по странам
6000
4000
2000
0
США
обритания
Канада
Германия
Франция
Китай
Азия
Россия
Европа