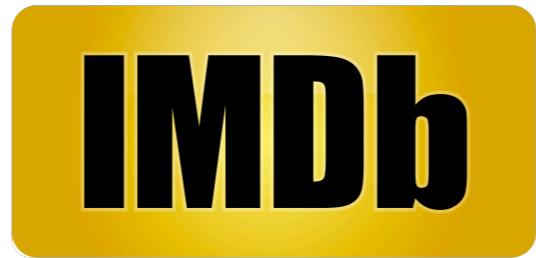




Анализ отзывов на IMDb

Анастасия Костяницина



- ★ IMDb - крупнейшая в мире база данных и веб-сайт о кинематографе
- ★ Особенностью отзывов на IMDb является пометка "**Warning: Spoilers**", которая предупреждает читателя, о наличии в тексте раскрытия сюжета
- ★ Пометку о наличии спойлера пользователь устанавливает сам

 1/10

Disneys absolute worst in many years

mtgospiller 11 January 2014

Warning: Spoilers

253 out of 443 found this helpful. Was this review helpful?

Yes

No

[Report this](#) | [Permalink](#)



Оценка пользователя



Холодное сердце (I) (2013)

★ 1/10

Disneys absolute worst in many years

11 January 2014

Warning: Spoilers

Дата

Заголовок

Пометка

Текст

One of the worst Disney animated movies I can recall seeing. 75% inane songs, way to obvious plot even for Disney. Very, very little humour. Bland characters and just plain below par compared to what Disney and other animation studios have been able to produce in the last 10 years. The usual villains and sidekicks appear but with nothing new to offer.

I could not stop myself from skipping at least a minute from most of the songs. The voice acting was, well, bland is the best word. Absolutely mediocre. I was thoroughly disappointed

Do not by any means pay to watch this and be sure to be in the company of easily amused small children.

Оценившие

253 out of 443 found this helpful. Was this review helpful?

Yes

No

[Report this](#) | [Permalink](#)

Просмотры

Исследование

Несмотря на то, что спойлеры, особенно в сети, считаются нежелательными, люди все равно оставляют такие отзывы. Интересно исследовать на "причины" их написания.

Будем проверять следующие гипотезы:

- ★ Больше просмотров будет у отзывов с низкой оценкой
- ★ Отзывы со спойлерами будут иметь меньше просмотров
- ★ Спойлеры скорее будут писаться к фильмам/сериалам, которые не понравились пользователю (низкая оценка фильма)
- ★ Спойлеры скорее будут писаться к сериалам
- ★ Спойлеры скорее будут писаться к фильмам с низким рейтингом
- ★ Чем старше фильм, тем вероятнее наличие спойлеров

Данные

.....

Основой послужили следующие чарты на IMDb

- ★ Top Rated Movies
- ★ Top Rated TV Shows
- ★ Lowest Rated Movies
- ★ Lowest Rated TV Shows

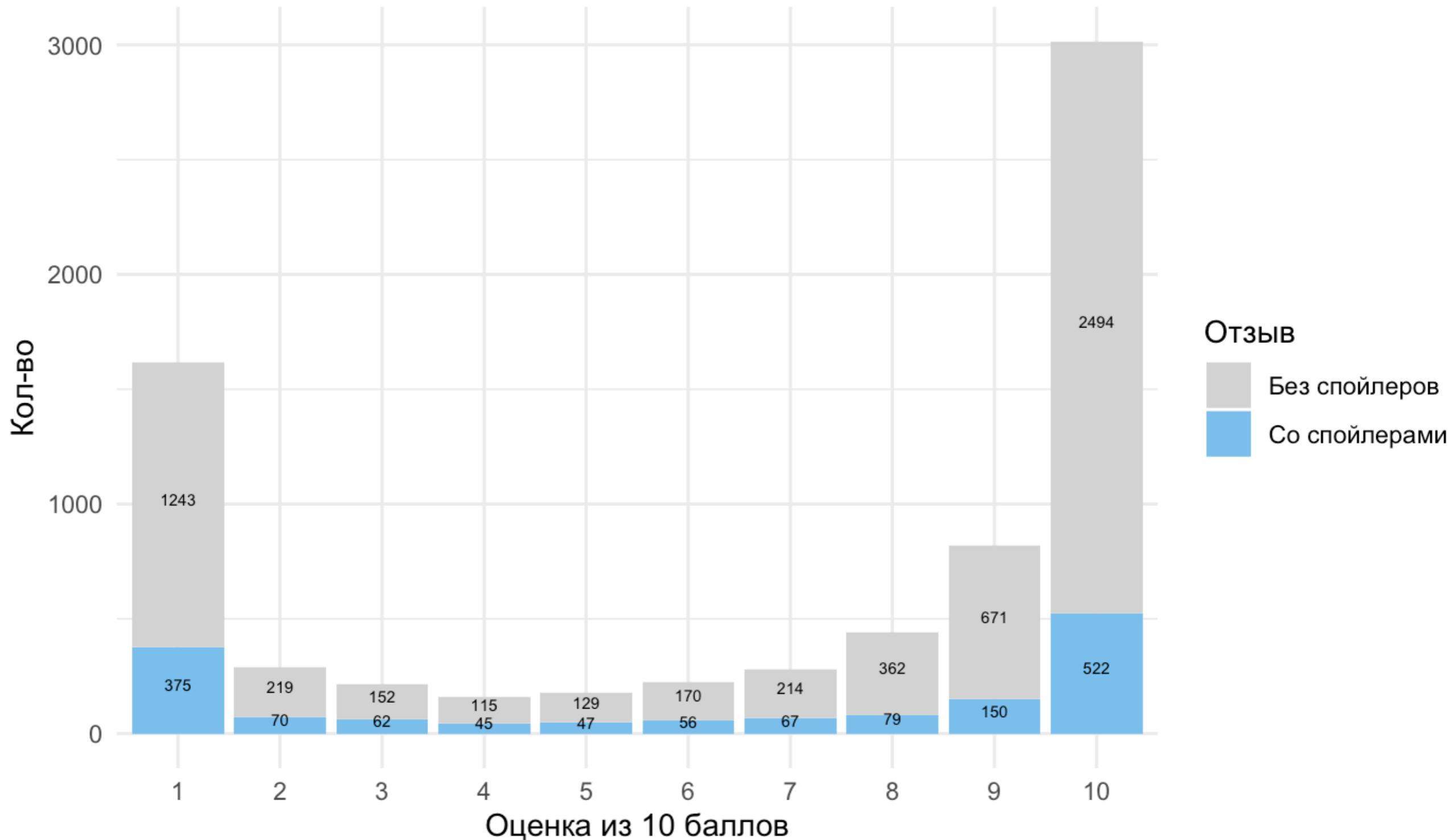
Данные

• •

- ★ Из каждого списка 100 случайных названий
- ★ Для каждого фильма и сериала максимум 25 отзывов и соответствующие метаданные
- ★ В итоге получилось 8382 отзыва
- ★ Спойлерских отзывов - 20%

movie_id	rate	spoiler	date	user	title	link	text	help_plus	help_all	type	top	year
tt0120179	3	0	7 April 2003	ur1980092	12: Snooze	rw0432965	o a complete	134	160	tv	low	2014
tt0120179	2	0	2 March 2007	ur1293485	I down this c	rw1610253	interrupted w	52	62	tv	low	2014
tt0120179	NA	0	September 20	ur1219578	ower, poorer.	rw0432978	ock almost st	79	102	tv	low	2014
tt0120179	1	0	6 June 2004	ur3515639	g boat full of	rw0432989	e 4 times bef	82	108	tv	low	2014
tt0120179	NA	0	1 January 200	ur1002035	vledge of the	rw0432983	ng also is dat	53	68	tv	low	2014

Кол-во оценок пользователей, написавших отзыв, по группам



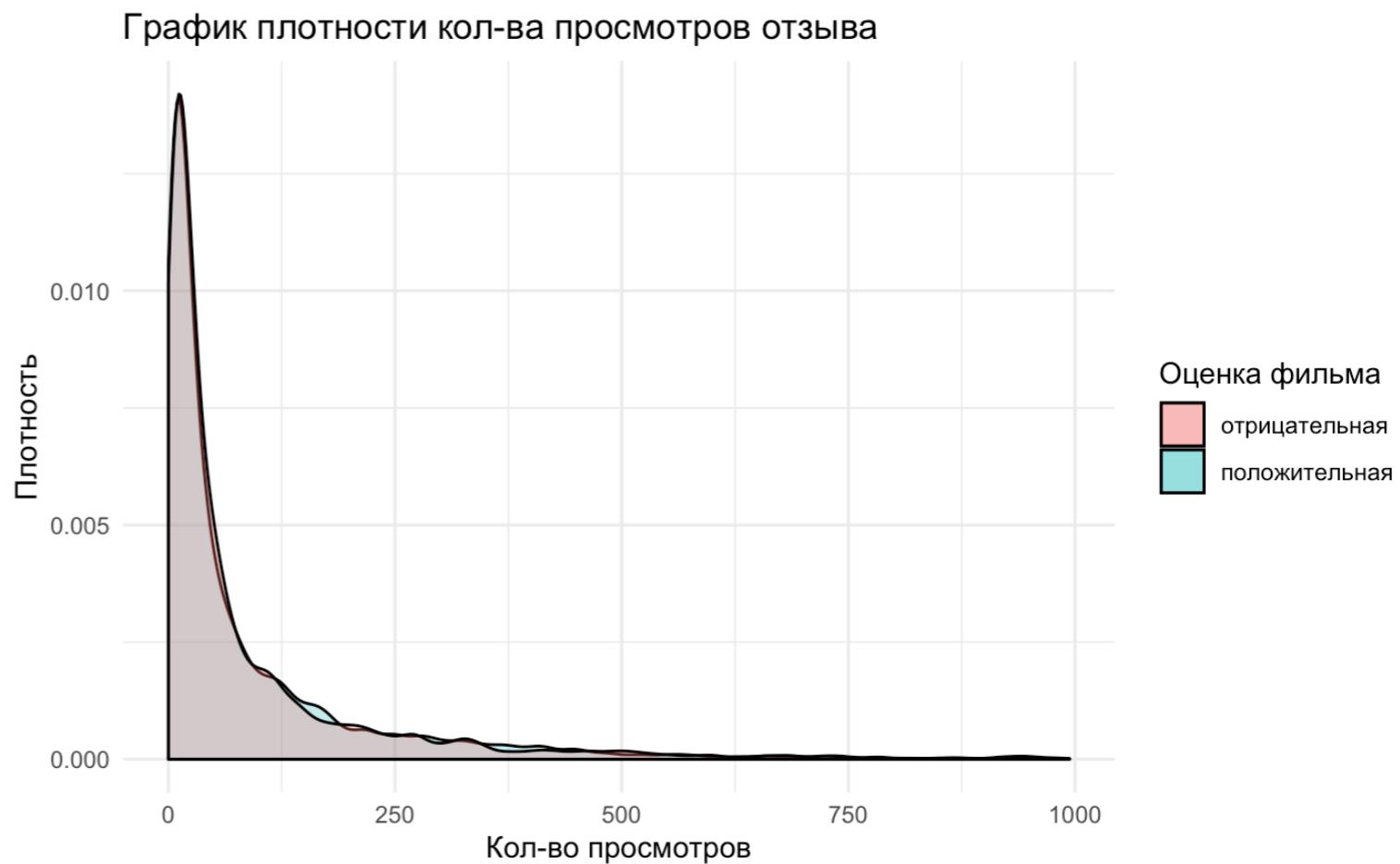
Анализ

- ★ Критерий Манна-Уитни для порядковых и количественных данных
- ★ Критерий Хи-квадрат для качественных номинальных данных

Зависимость кол-ва просмотров комментария от оценки фильма рецензентом



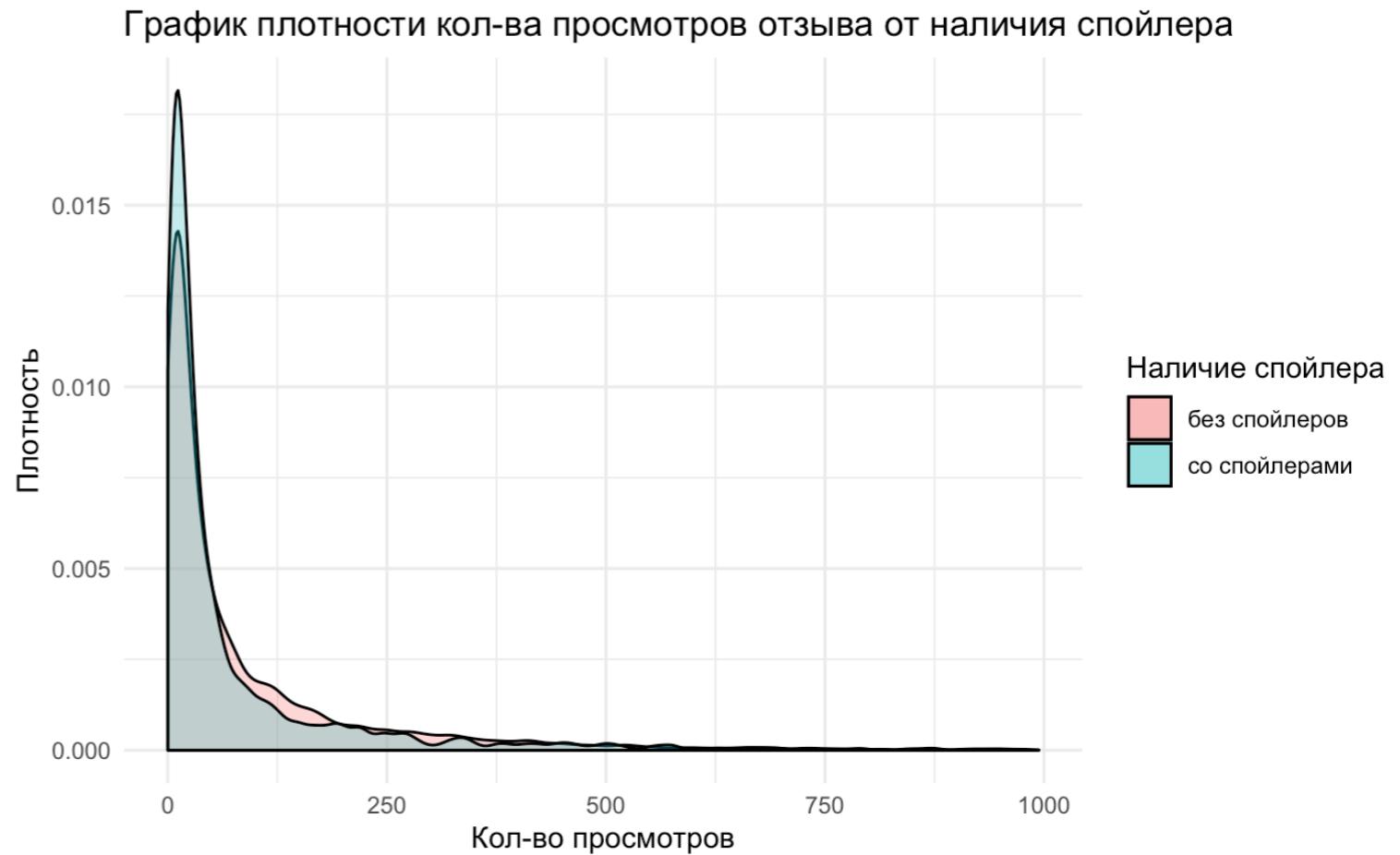
- ★ Распределены не нормально
- ★ **Тип групп:** Независимые
- ★ **Тест:** Критерий Манна-Уитни
- ★ Удаление выбросов
- ★ **p-value = 0.2755438**
- ★ Принимаем нулевую гипотезу



Зависимость кол-ва просмотров комментария от наличия спойлеров



- ★ Распределены не нормально
- ★ Тип групп: Независимые
- ★ Тест: Критерий Манна-Уитни
- ★ $p\text{-value} = 2.488404\text{e-}05$
- ★ Отвергаем нулевую гипотезу



Зависимость наличия спойлеров с оценкой фильма пользователем

• • • • • • • • • • • • • • • •

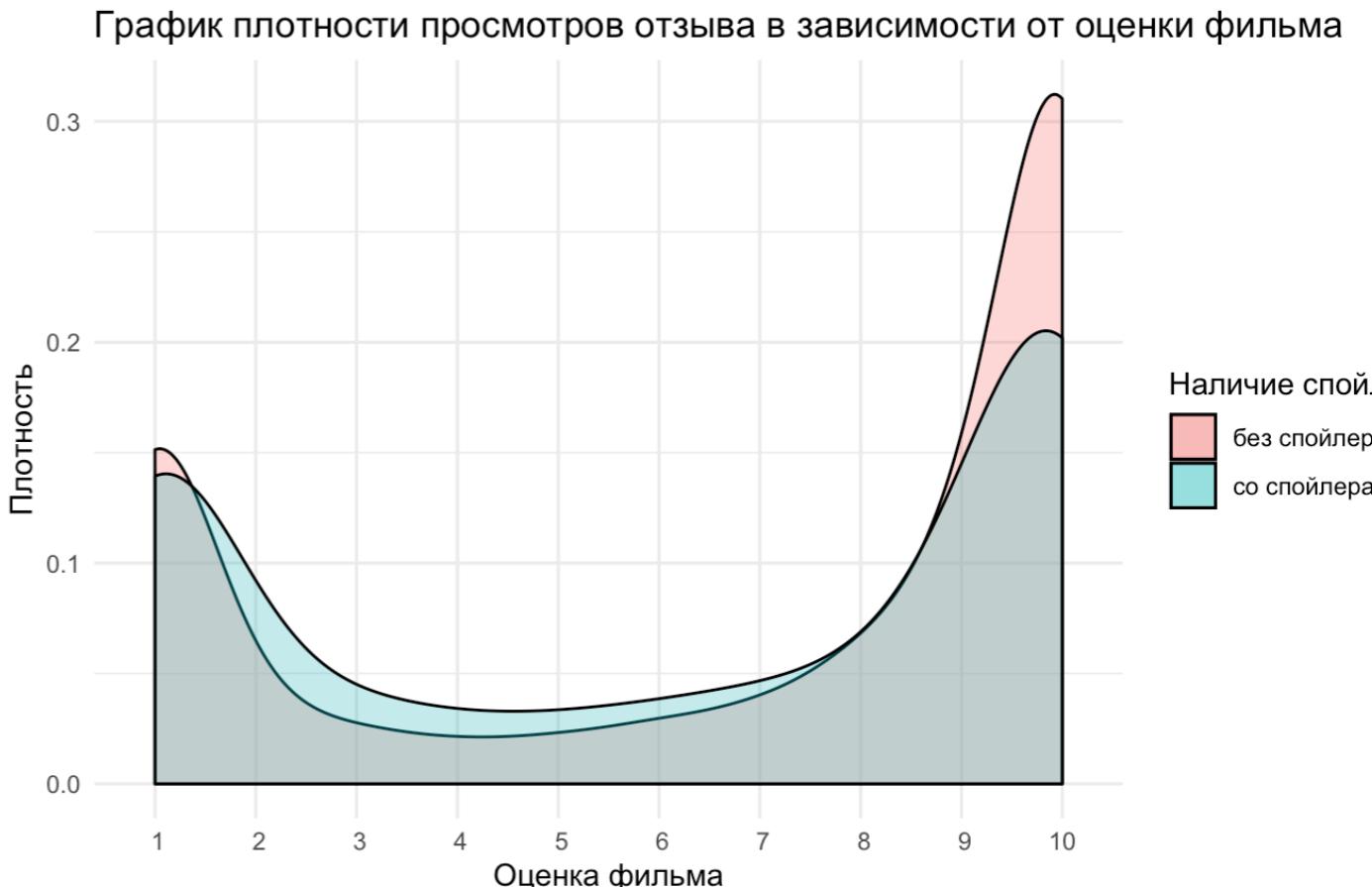
★ Распределены не нормально

★ Тип групп: Независимые

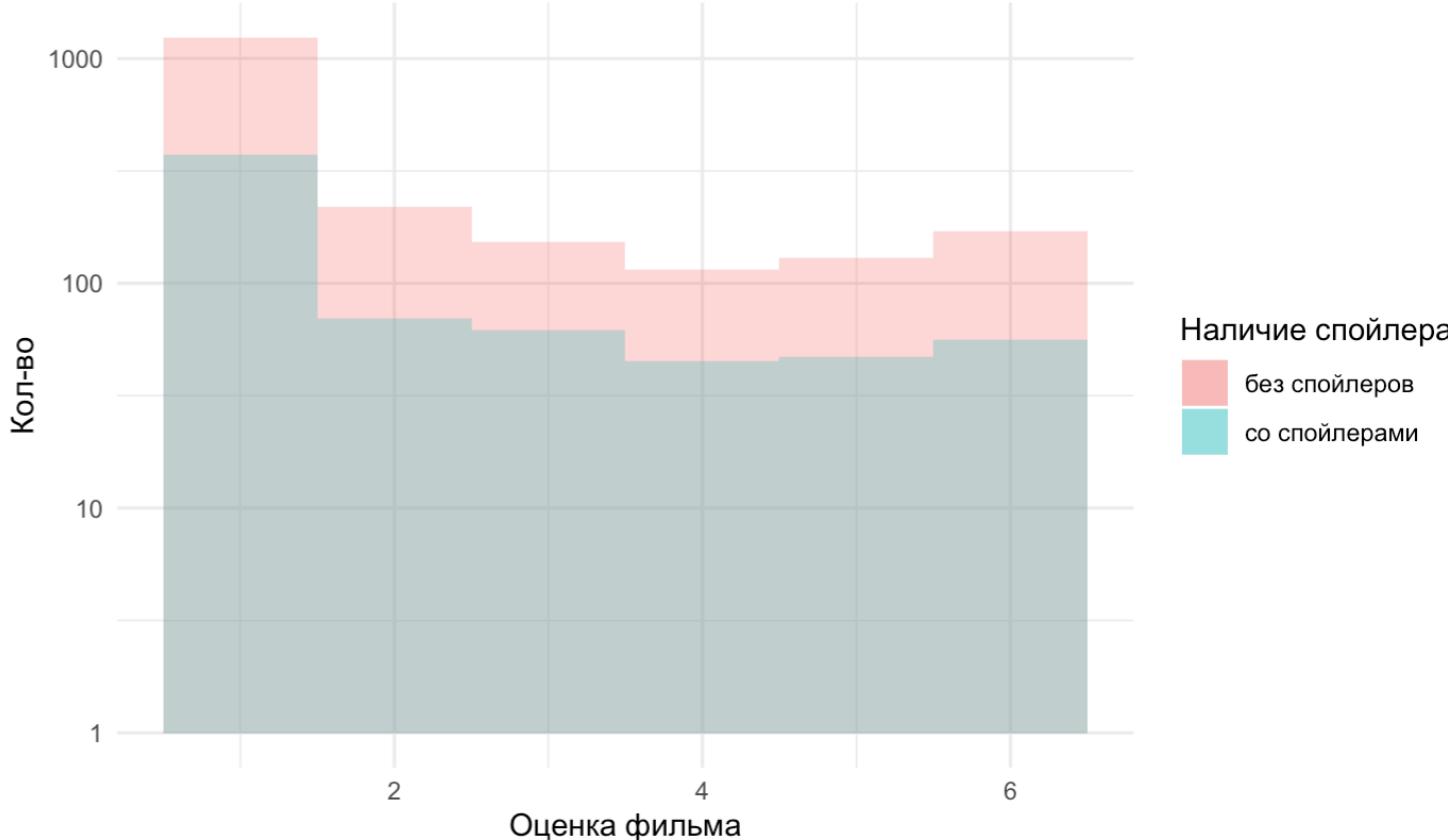
★ Тест: Критерий Манна-Уитни

★ $p\text{-value} = 7.515512\text{e-}10$

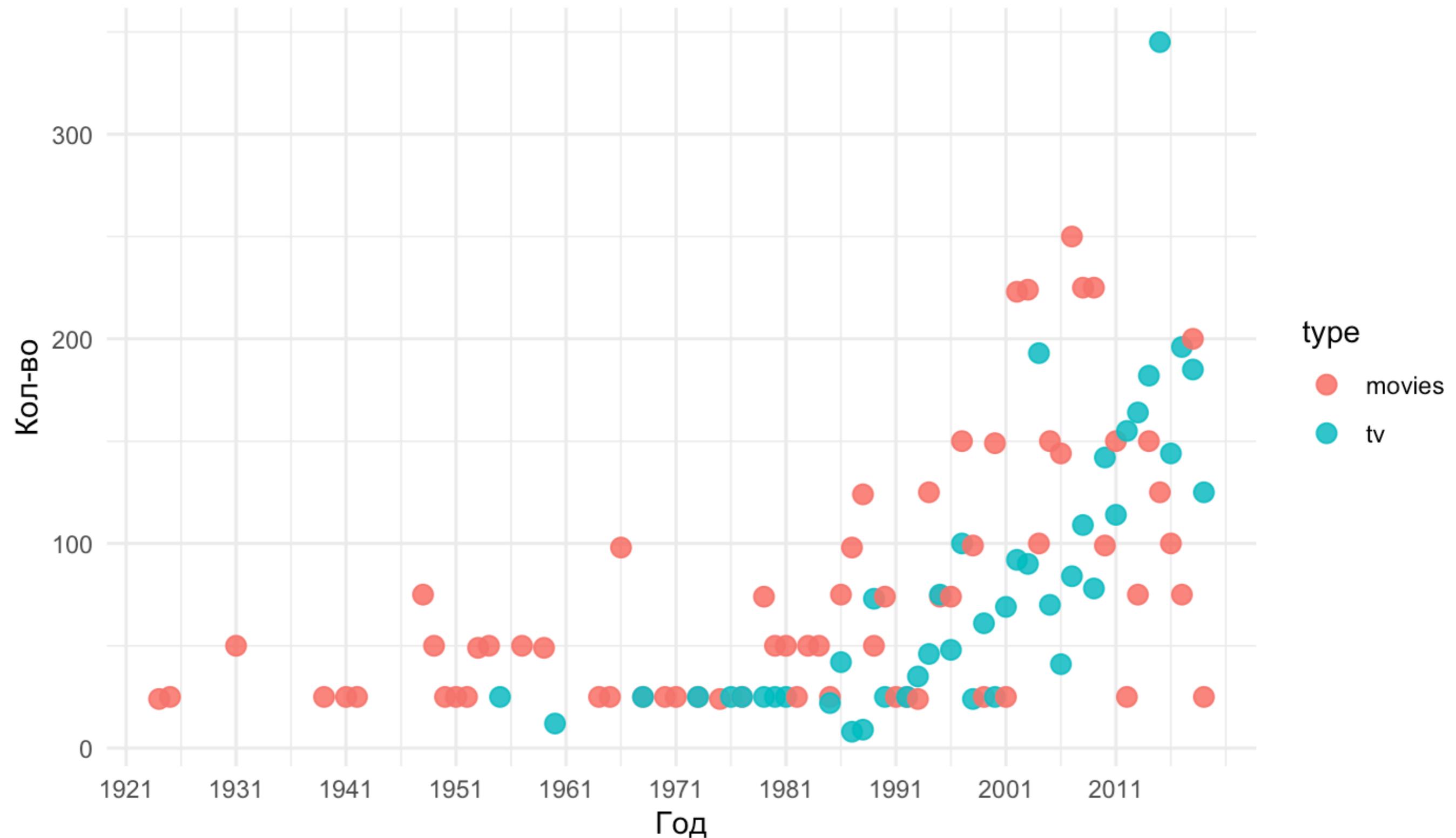
★ Отвергаем нулевую гипотезу



Кол-во просмотров фильма для отзывов с оценкой ≤ 6

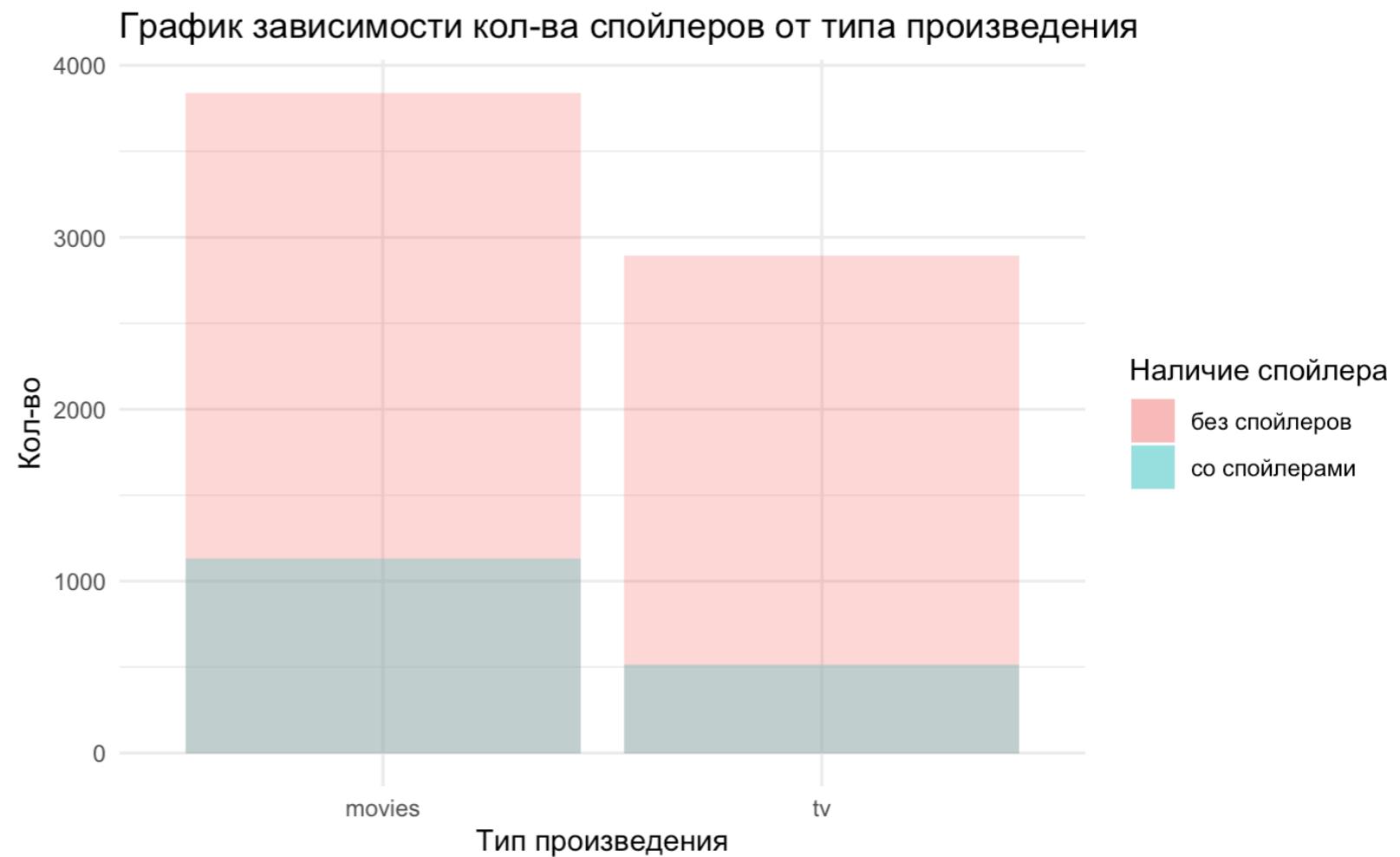


Количество фильмов и сериалов по годам выхода

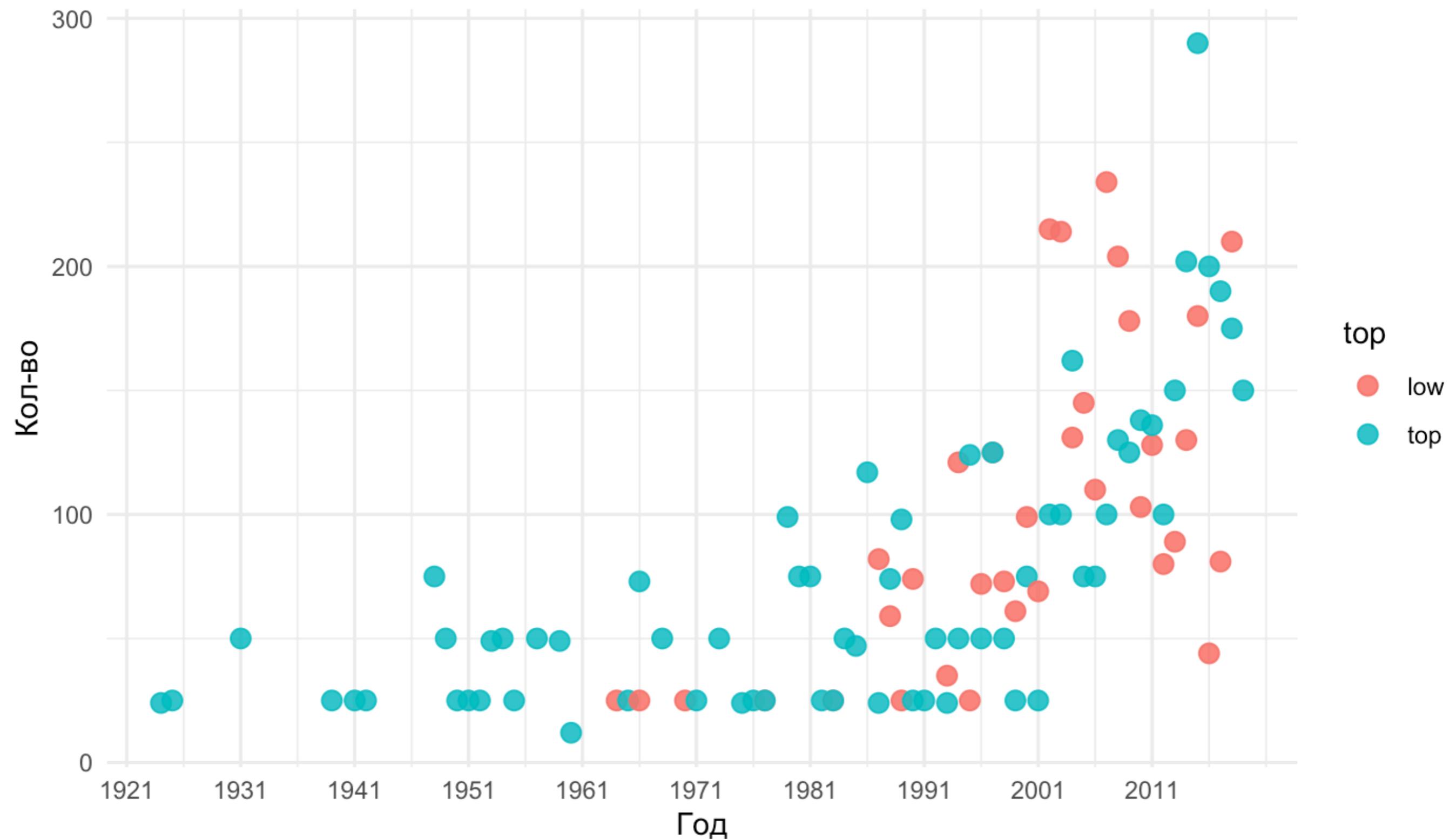


Зависимость спойлера от типа: тв шоу vs фильм

- ★ **Тип групп:** Независимые
 - ★ **Тест:** Критерий Хи-квадрат
 - ★ **p-value = 1.341e-05**
 - ★ Отвергаем нулевую гипотезу



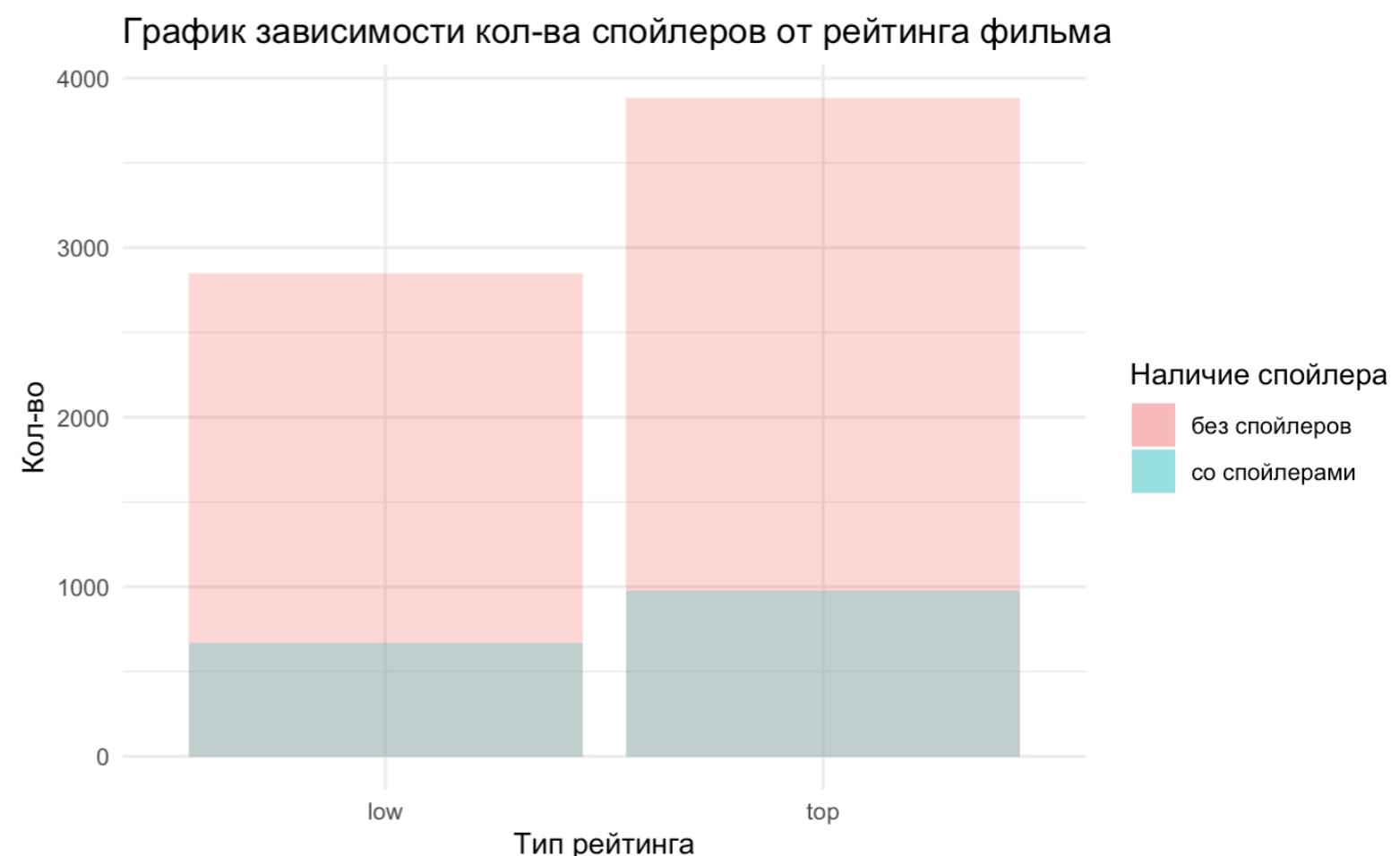
Количество low rated и high rated произведений по годам выхода



Зависимость спойлера от рейтинга фильма: high vs low

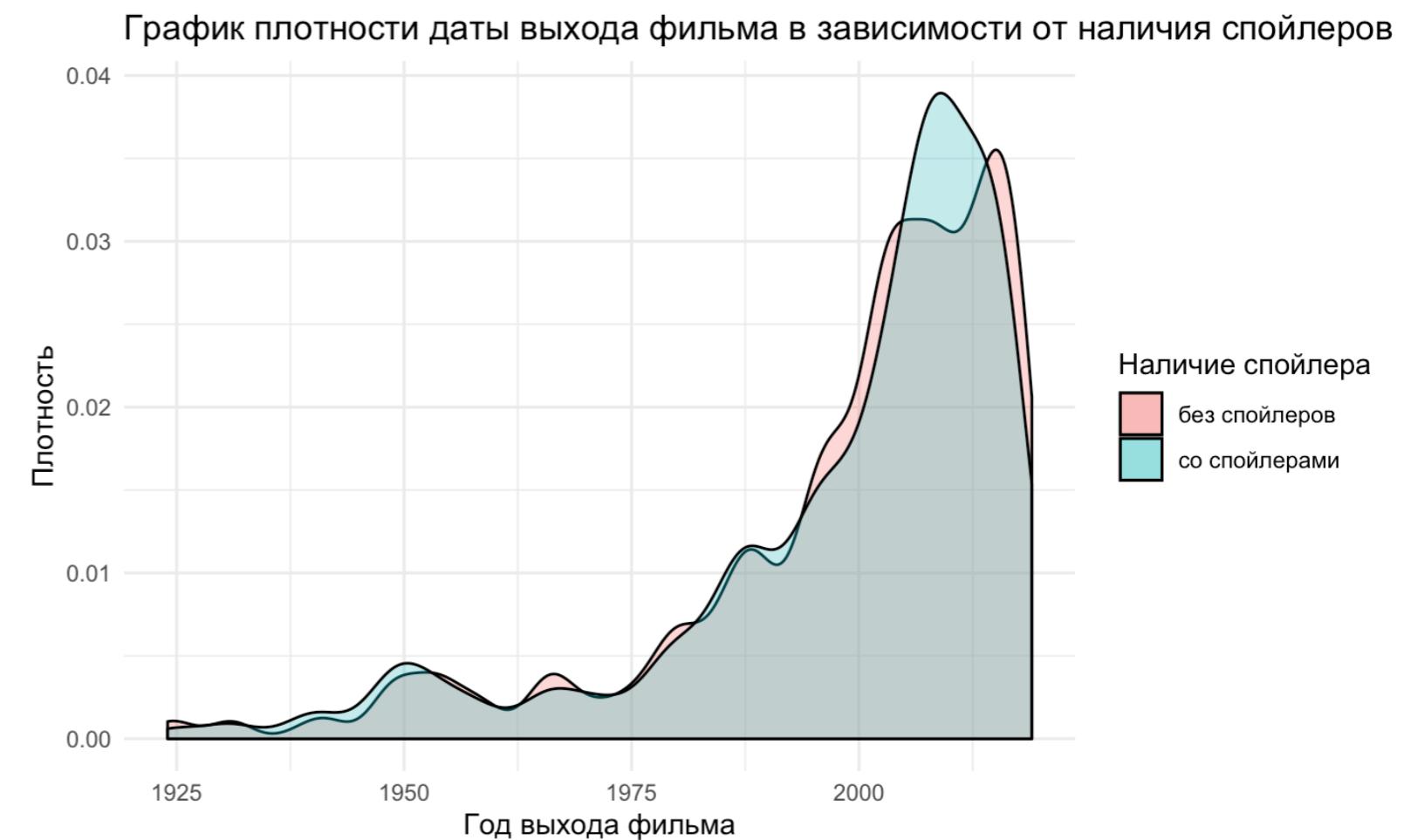


- ★ **Тип групп:** Независимые
 - ★ **Тест:** Критерий Хи-квадрат
 - ★ **p-value = 0.2675**
 - ★ Принимаем нулевую гипотезу



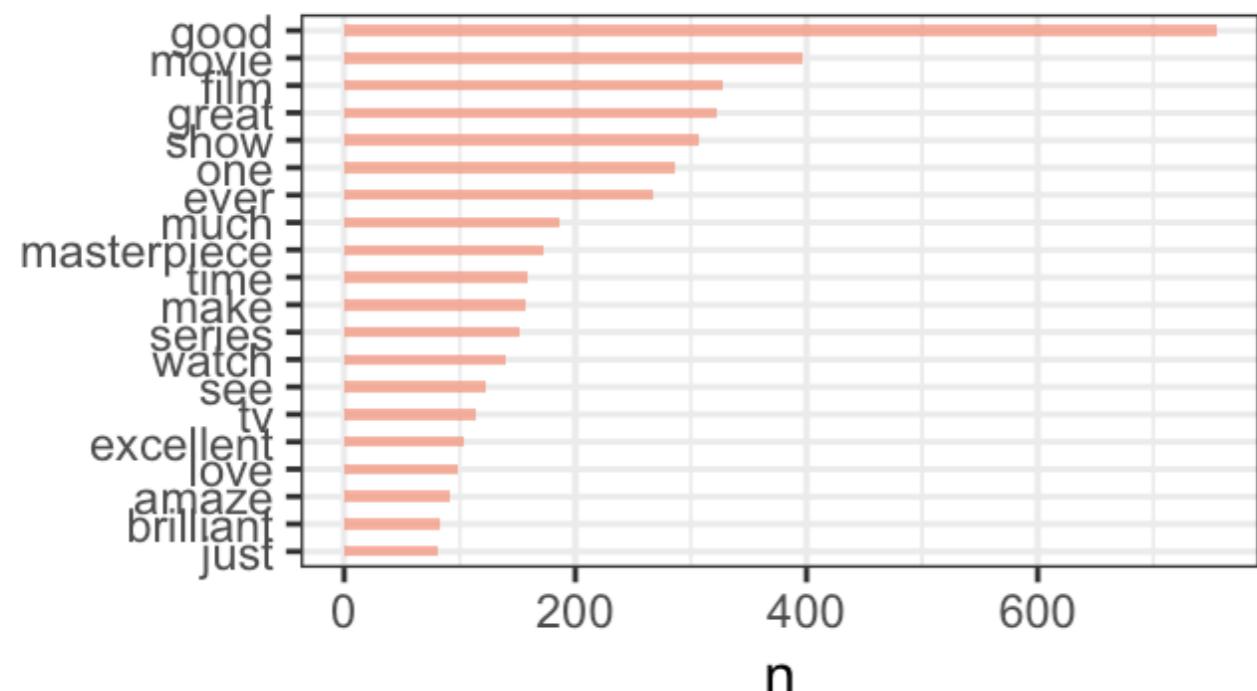
Зависимость спойлера от времени выхода фильма

- ★ Распределены не нормально
- ★ Тип групп: Независимые
- ★ Тест: Критерий Манна-Уитни
- ★ $p\text{-value} = 0.3926894$
- ★ Принимаем нулевую гипотезу

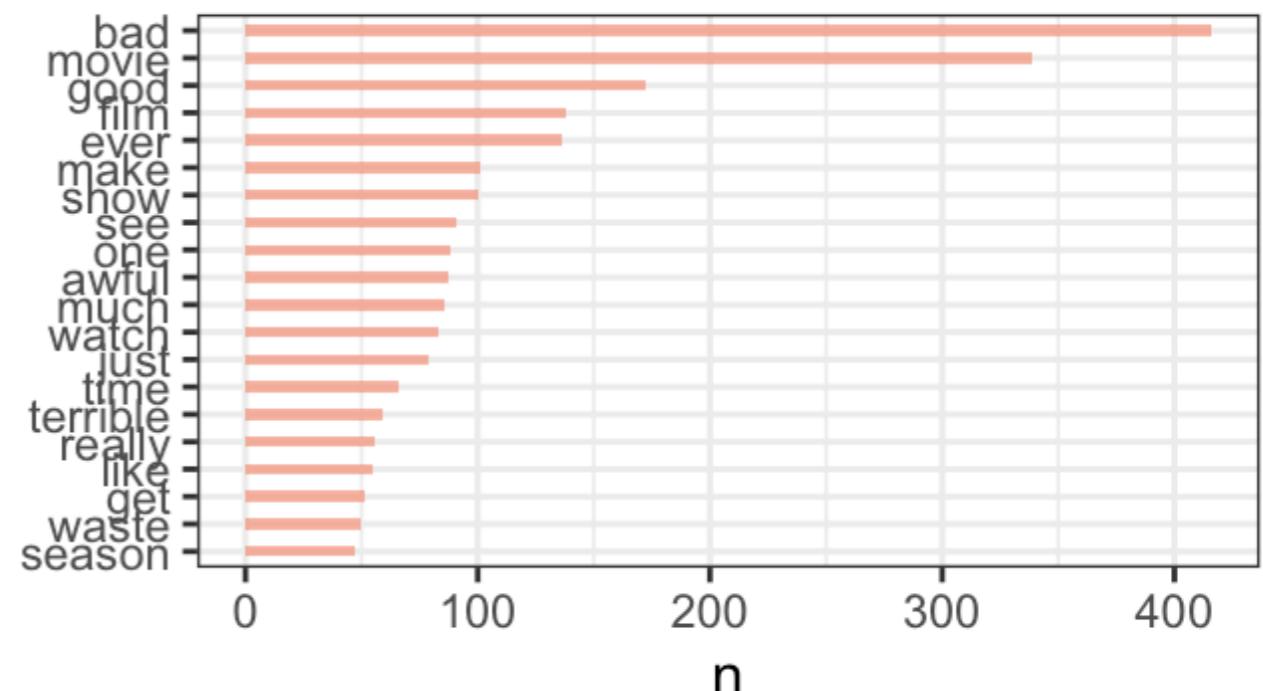


Частотные слова в заголовках отзывов

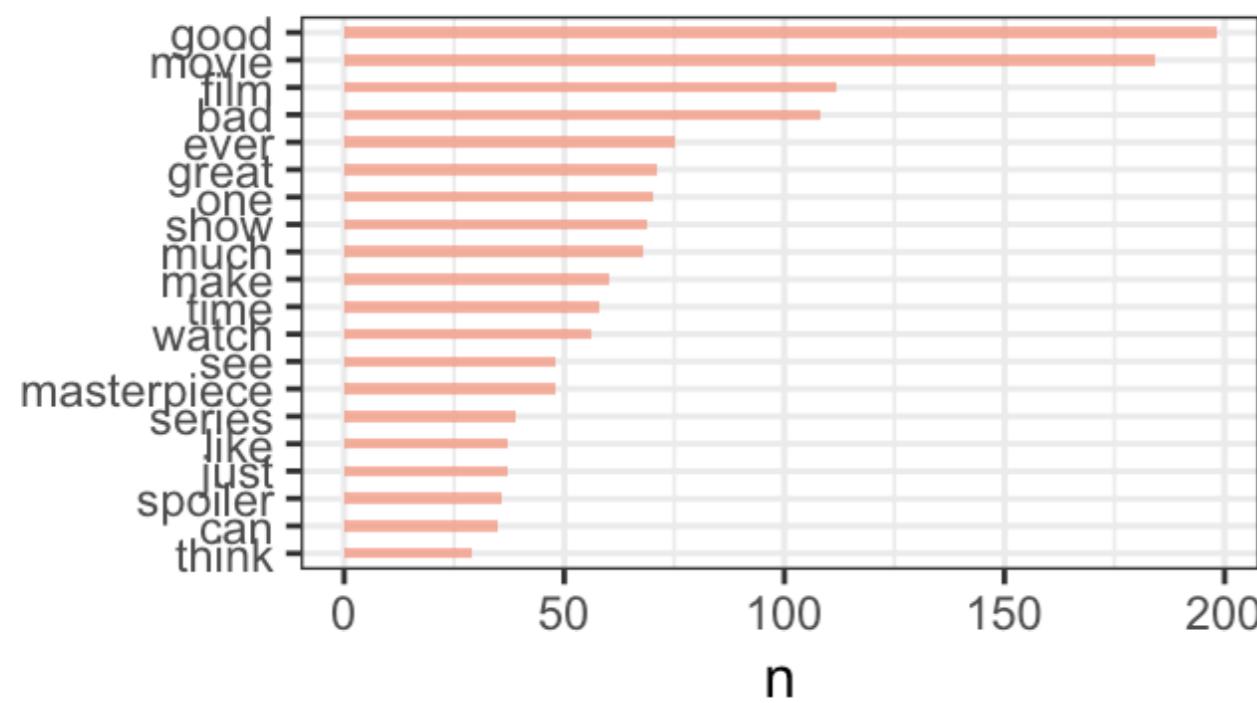
Положительная оценка



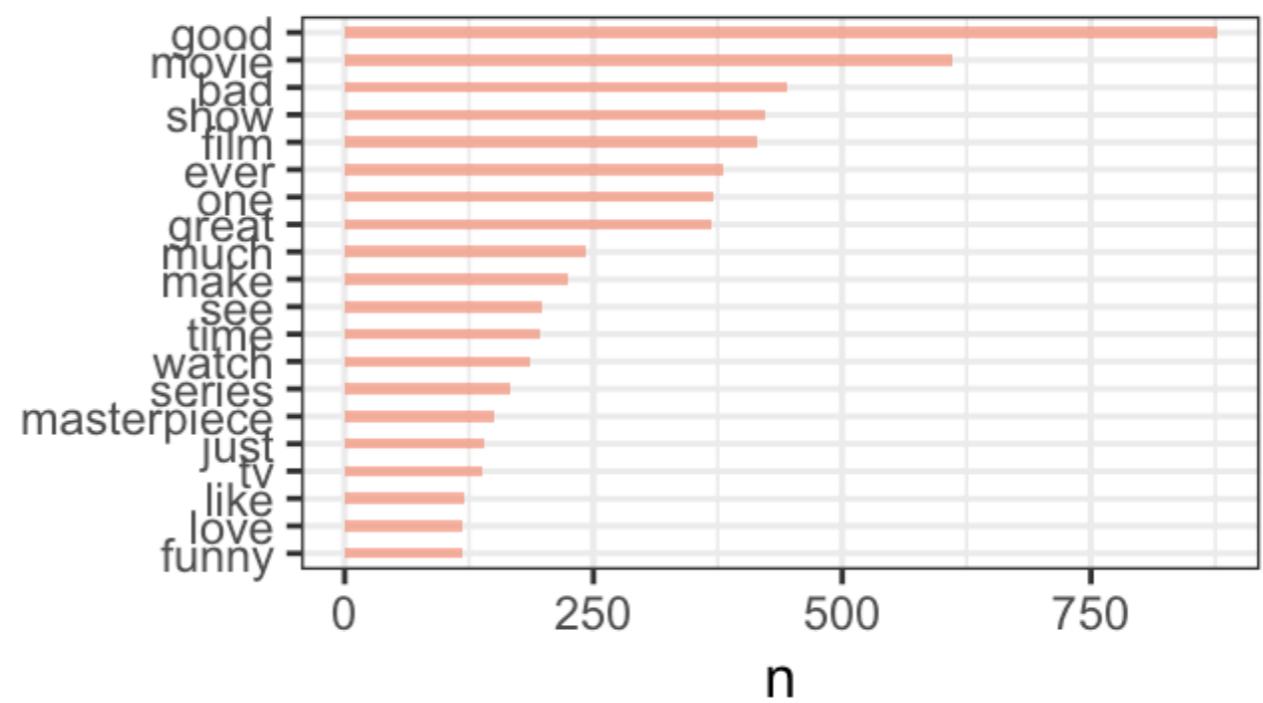
Отрицательная оценка



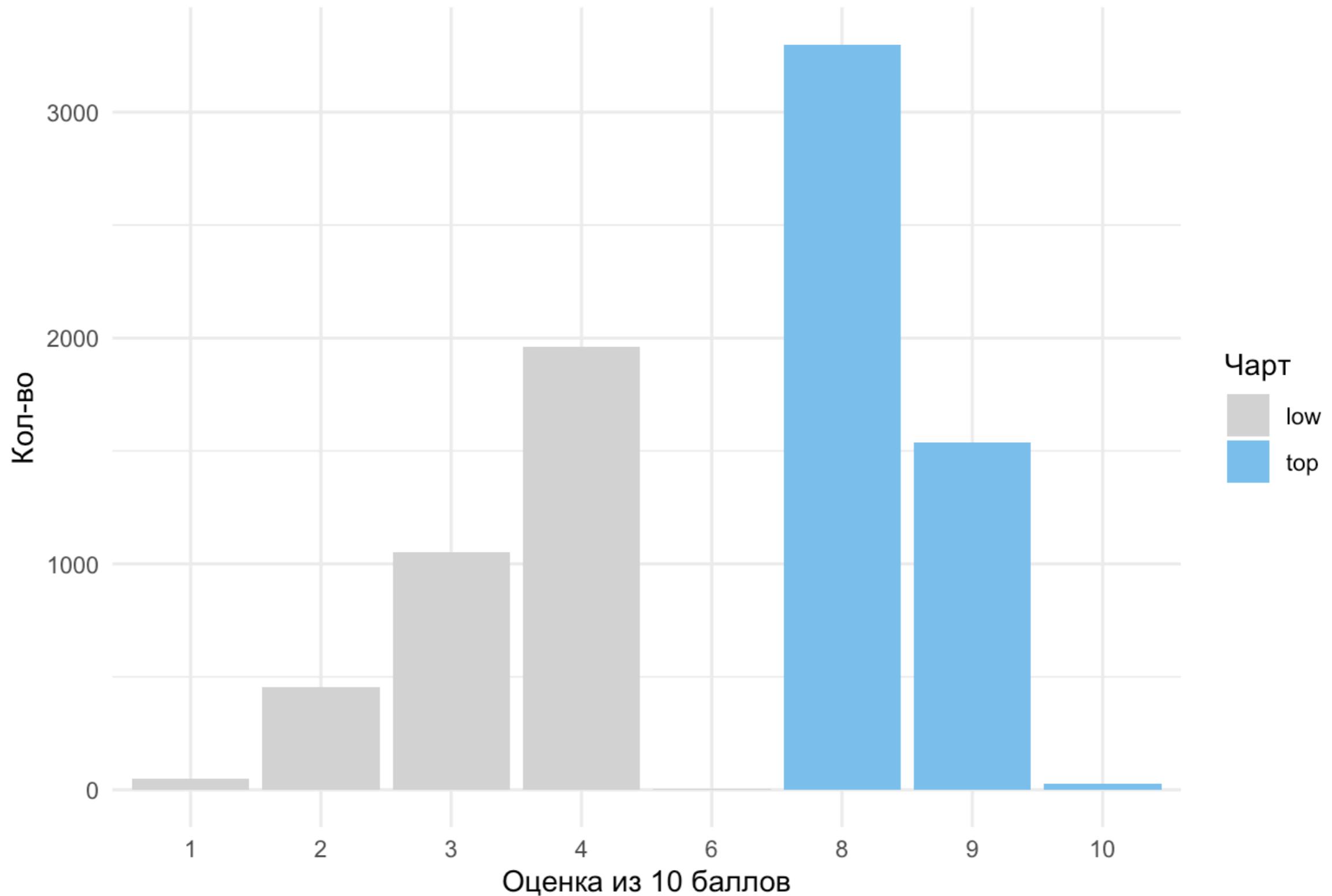
Отзыв со спойлерами



Отзыв без спойлеров



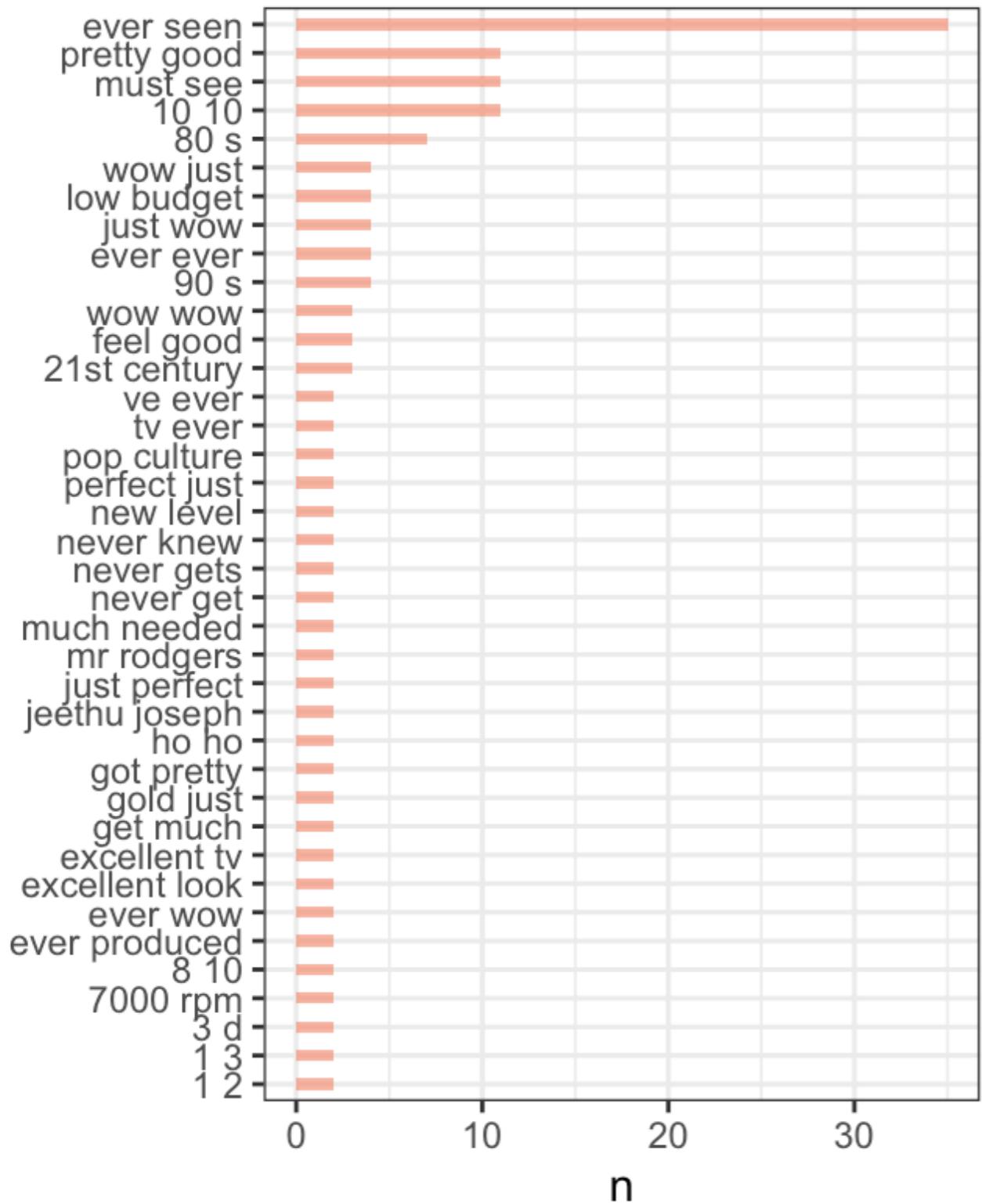
Оценки фильмов в зависимости от типа чарта



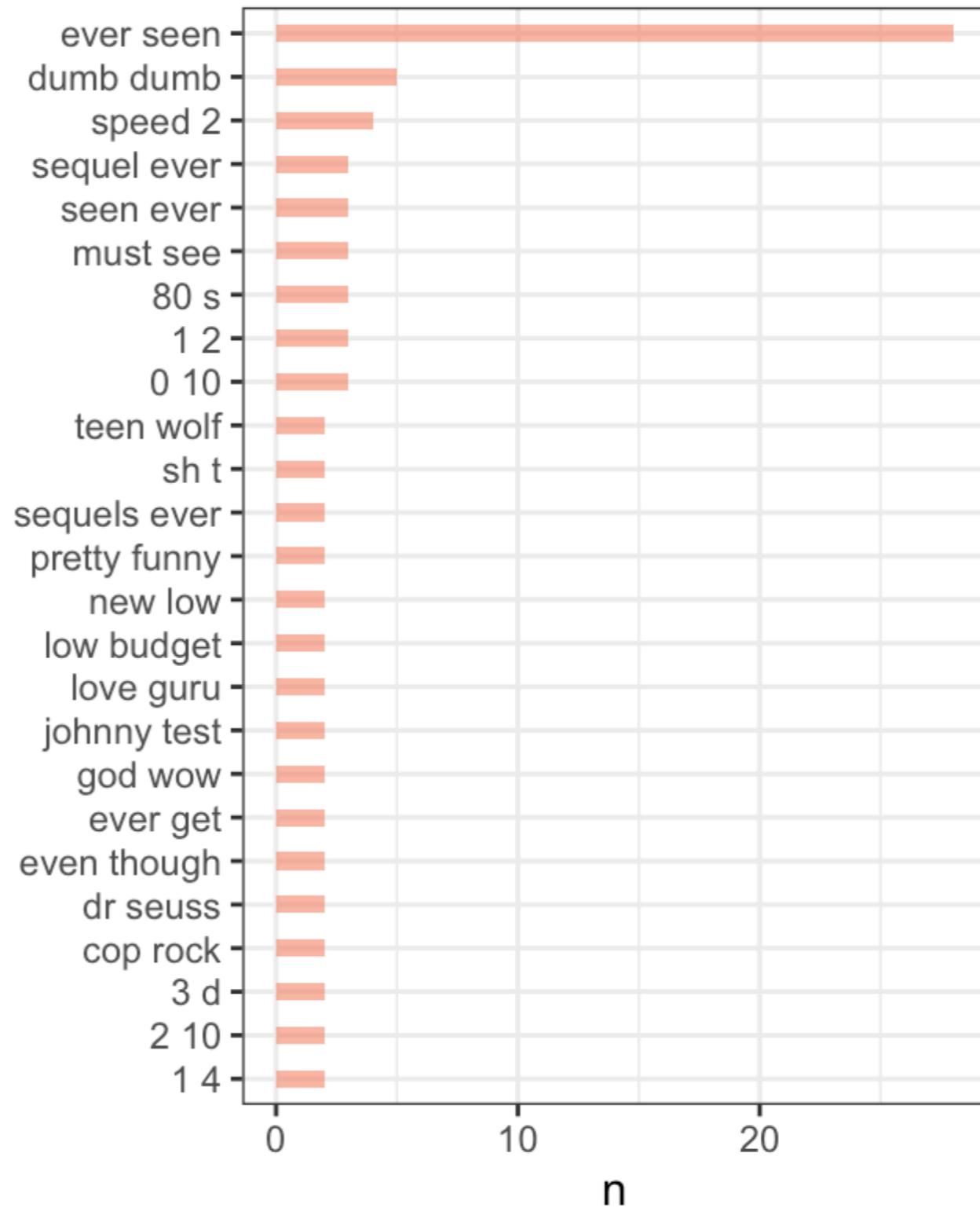
Частотные bigrams в заголовках отзывов

.....

Положительная оценка



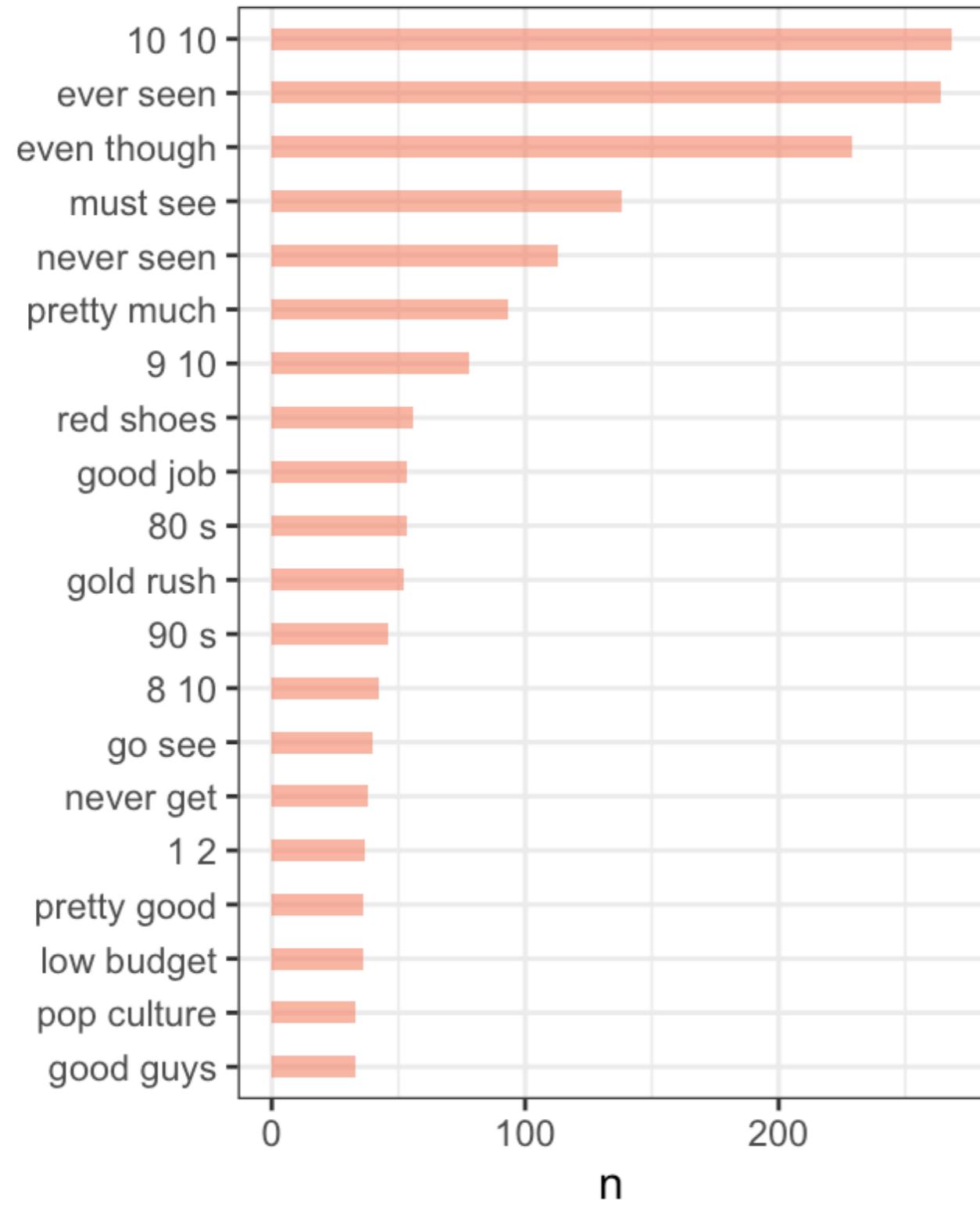
Отрицательная оценка



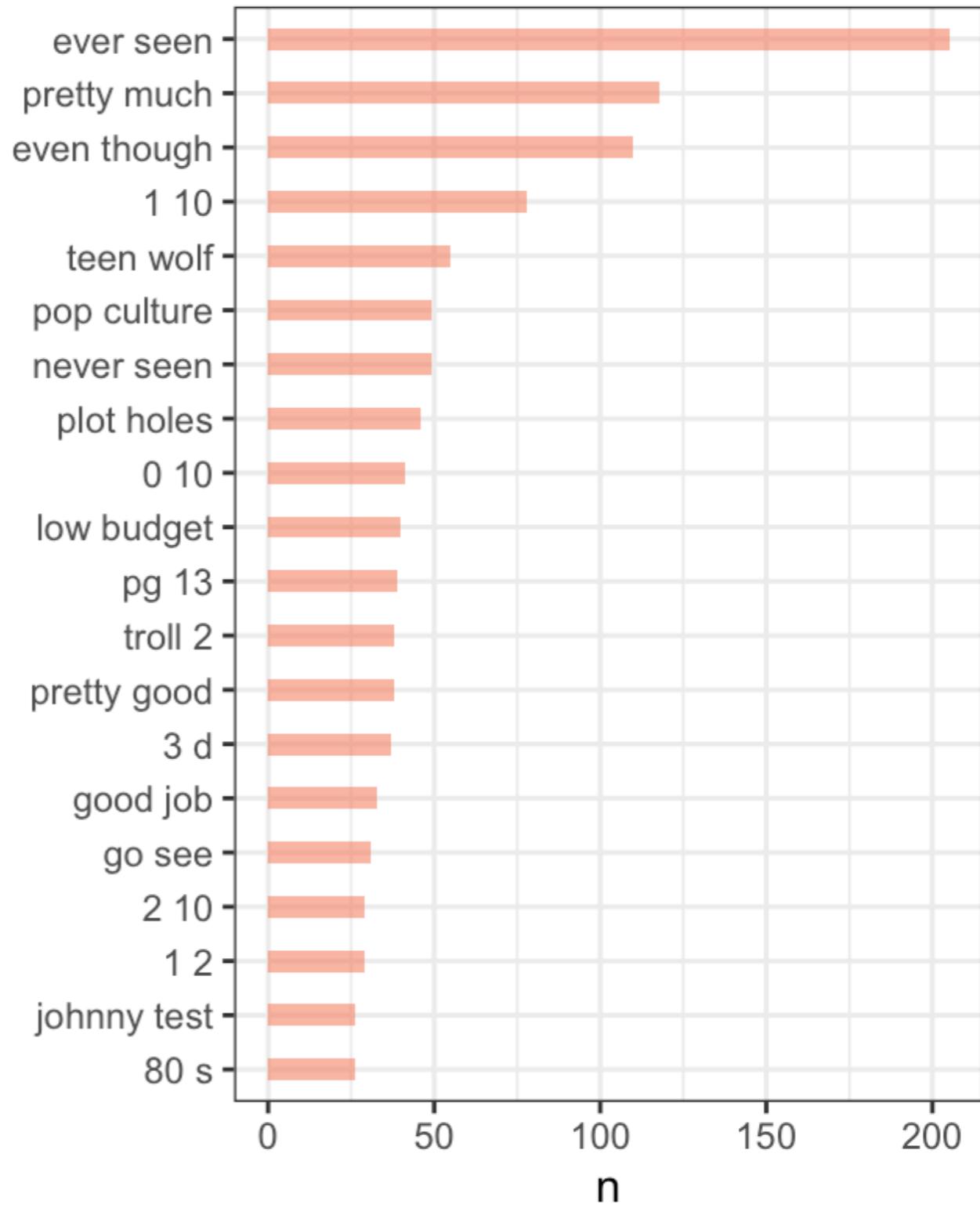
Частотные слова в полных текстах отзывов

.....

Положительная оценка



Отрицательная оценка



Оценки в текстах отзывов

rate	text_rate	out	old	new
Min. : 1.000	Min. : -10000	Min. : 5.0	Min. : 0.1000	Min. : -1000.0
1st Qu.: 1.000	1st Qu.: 1	1st Qu.: 10.0	1st Qu.: 0.1000	1st Qu.: 0.1
Median : 7.000	Median : 6	Median : 10.0	Median : 0.7000	Median : 0.8
Mean : 5.966	Mean : 49665	Mean : 11.6	Mean : 0.5966	Mean : 4966.4
3rd Qu.: 10.000	3rd Qu.: 9	3rd Qu.: 10.0	3rd Qu.: 1.0000	3rd Qu.: 1.0
Max. : 10.000	Max. : 10000000	Max. : 100.0	Max. : 1.0000	Max. : 1000000.0

График разницы между выставленной оценкой и оценкой в тексте отзыва

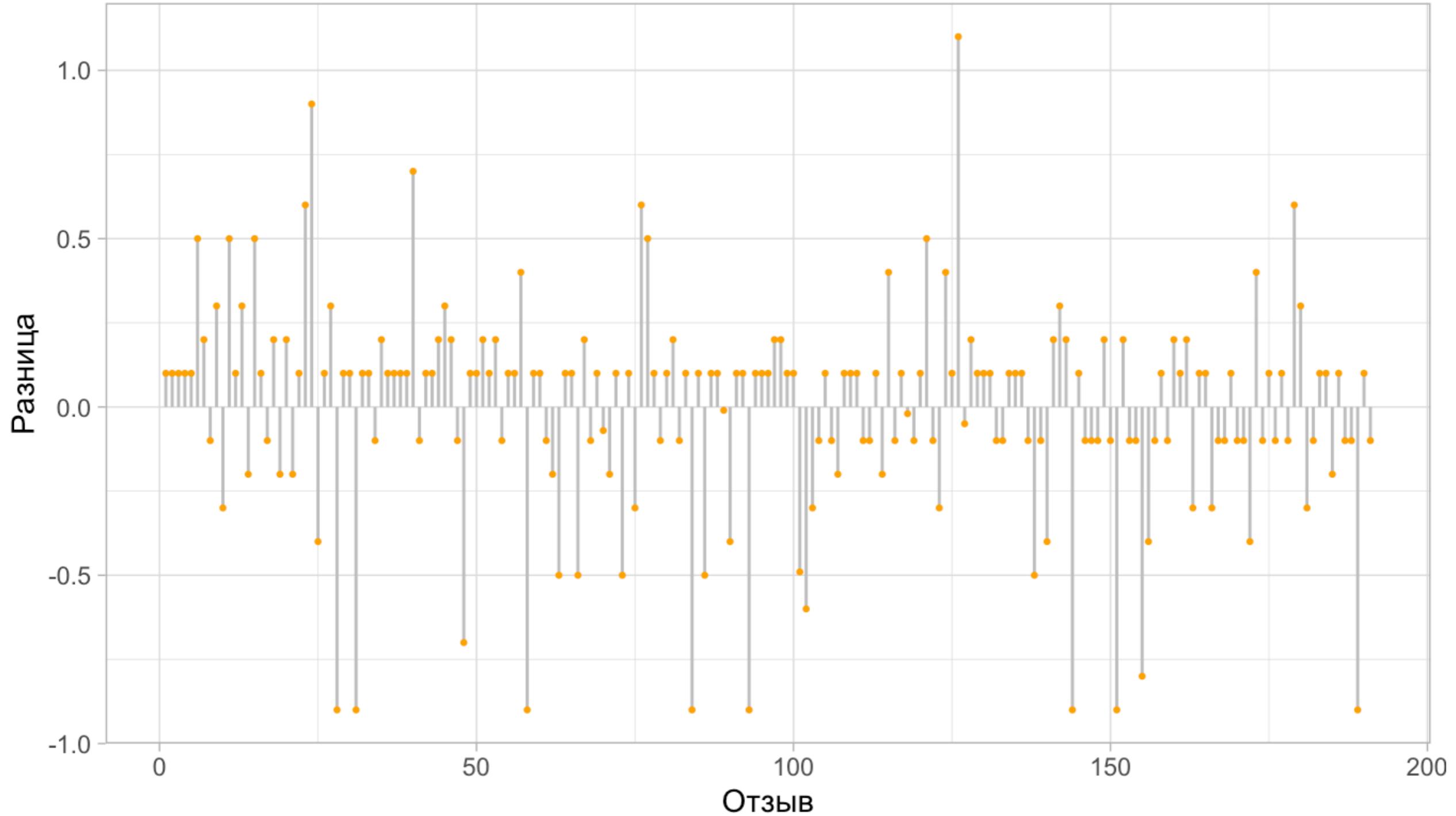
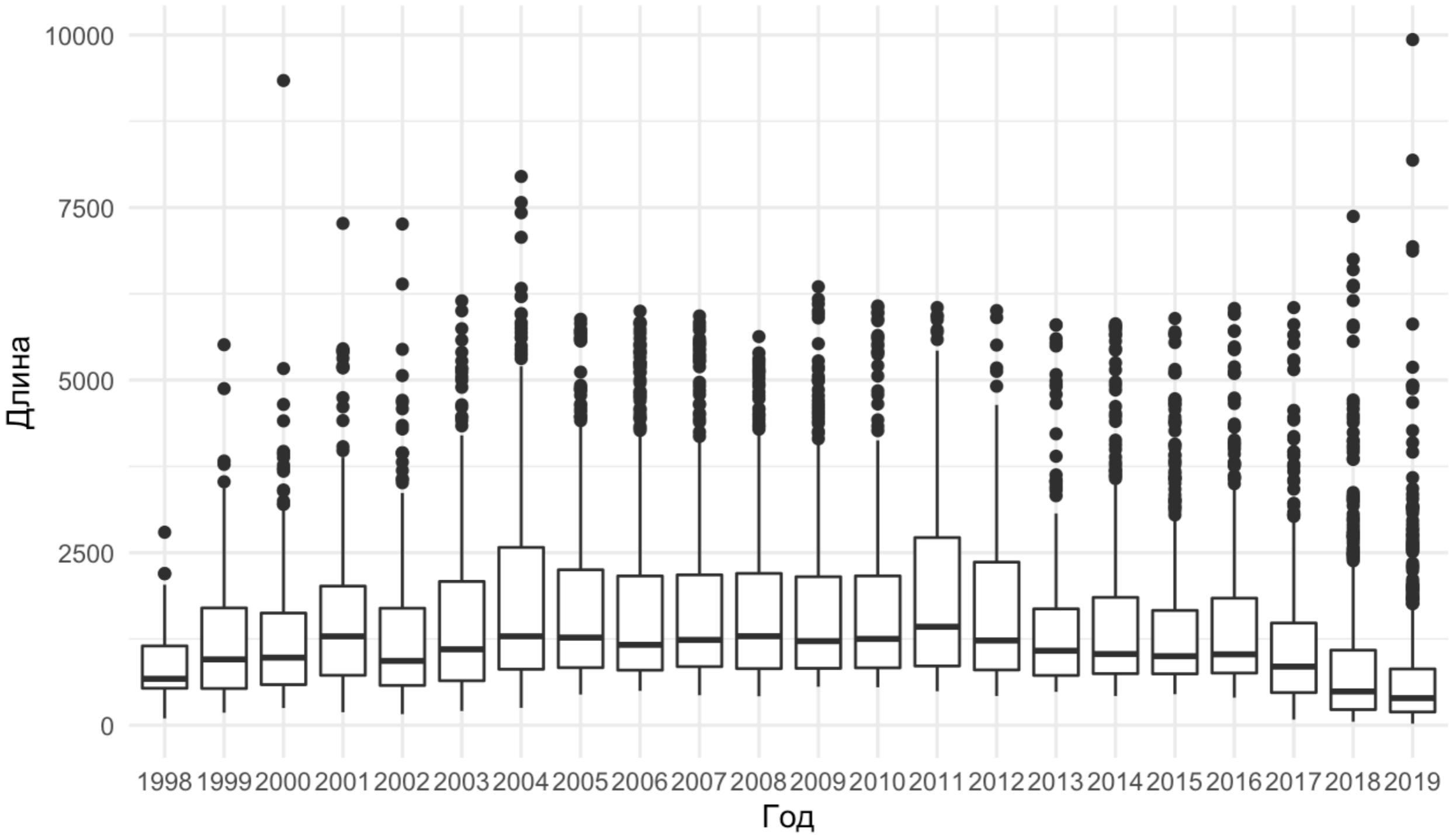
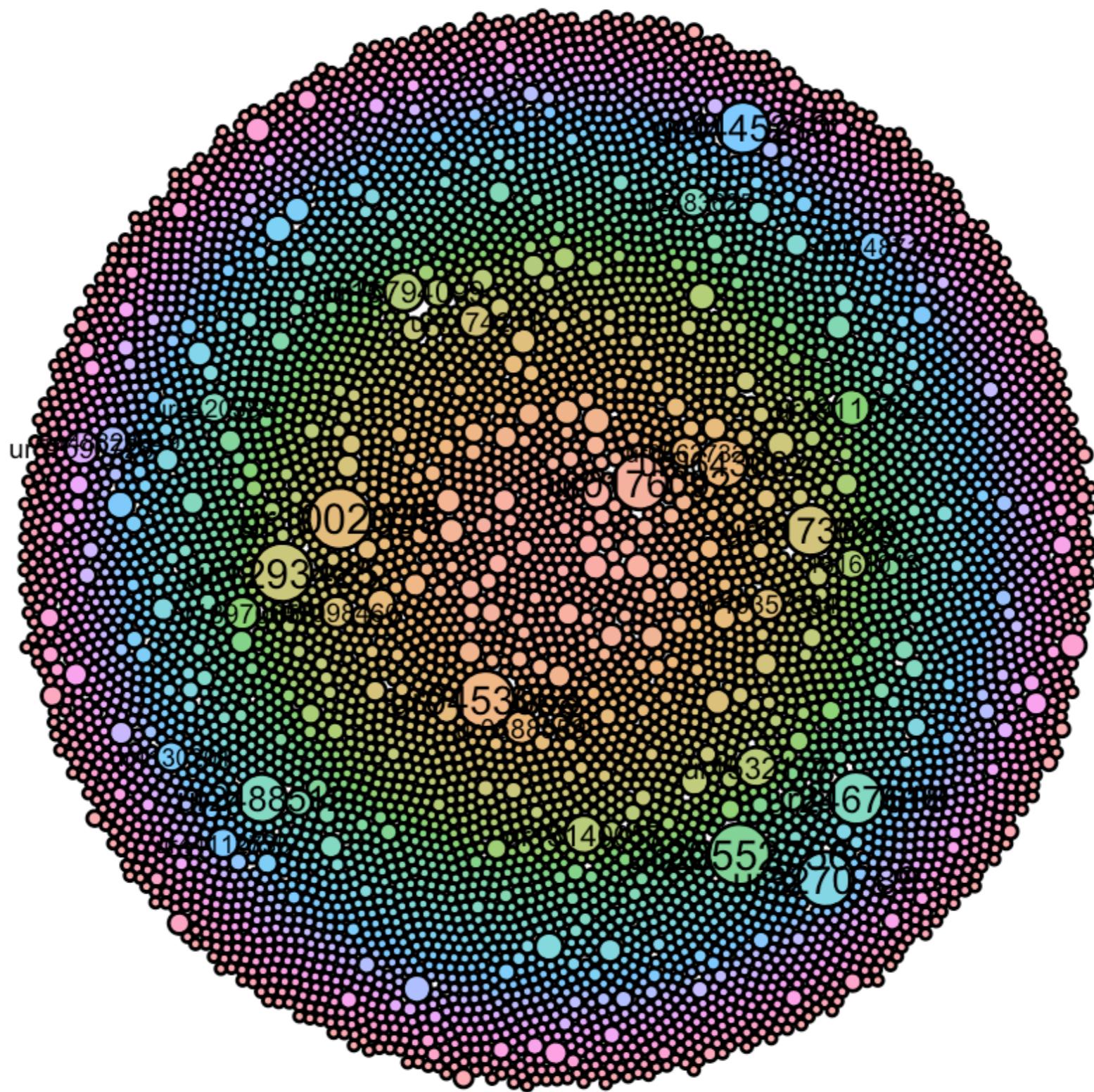


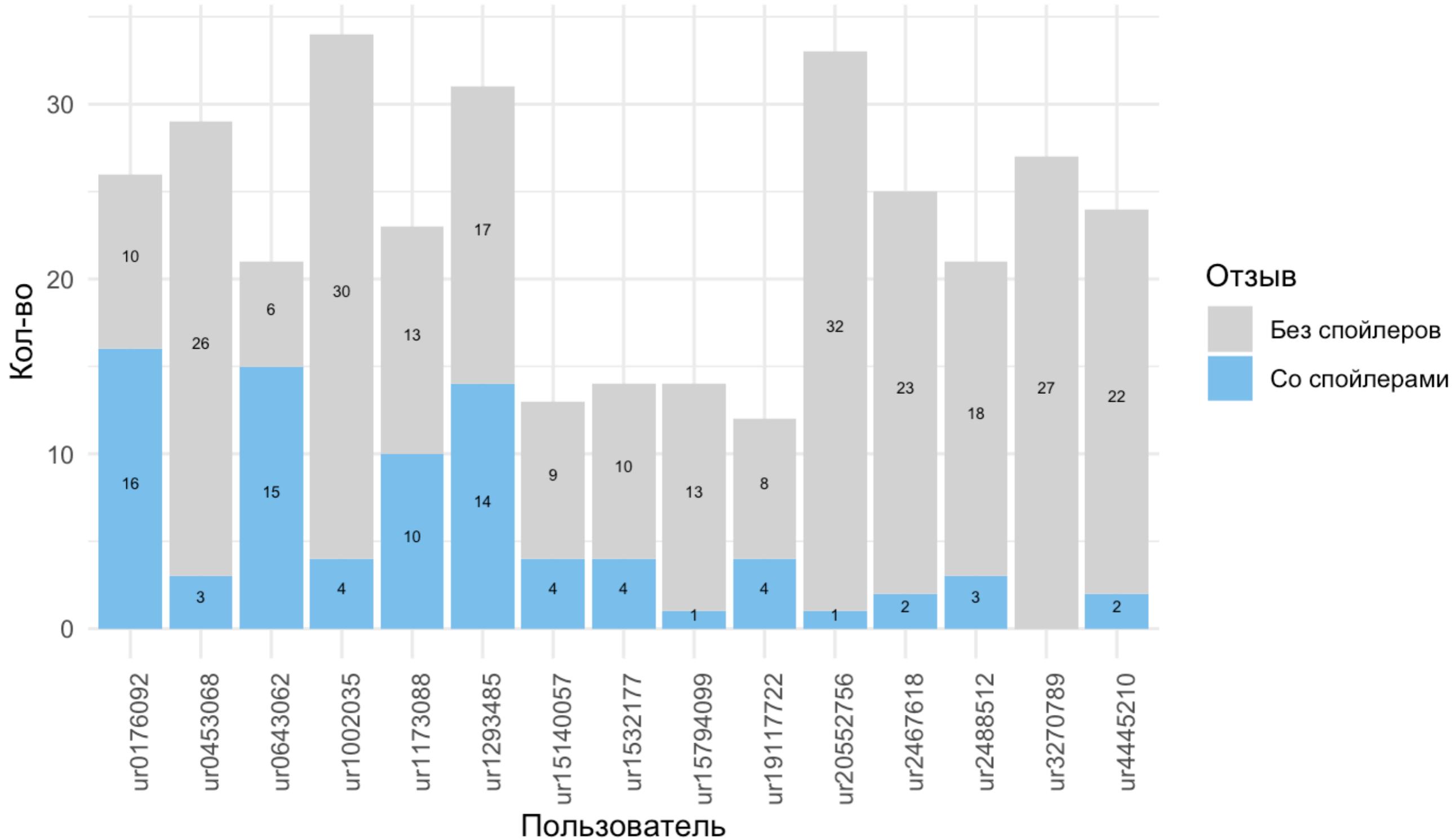
График длин отзывов в каждый год



Активность пользователей



Кол-во отзывов пользователей в зависимости от наличия в нем спойлеров

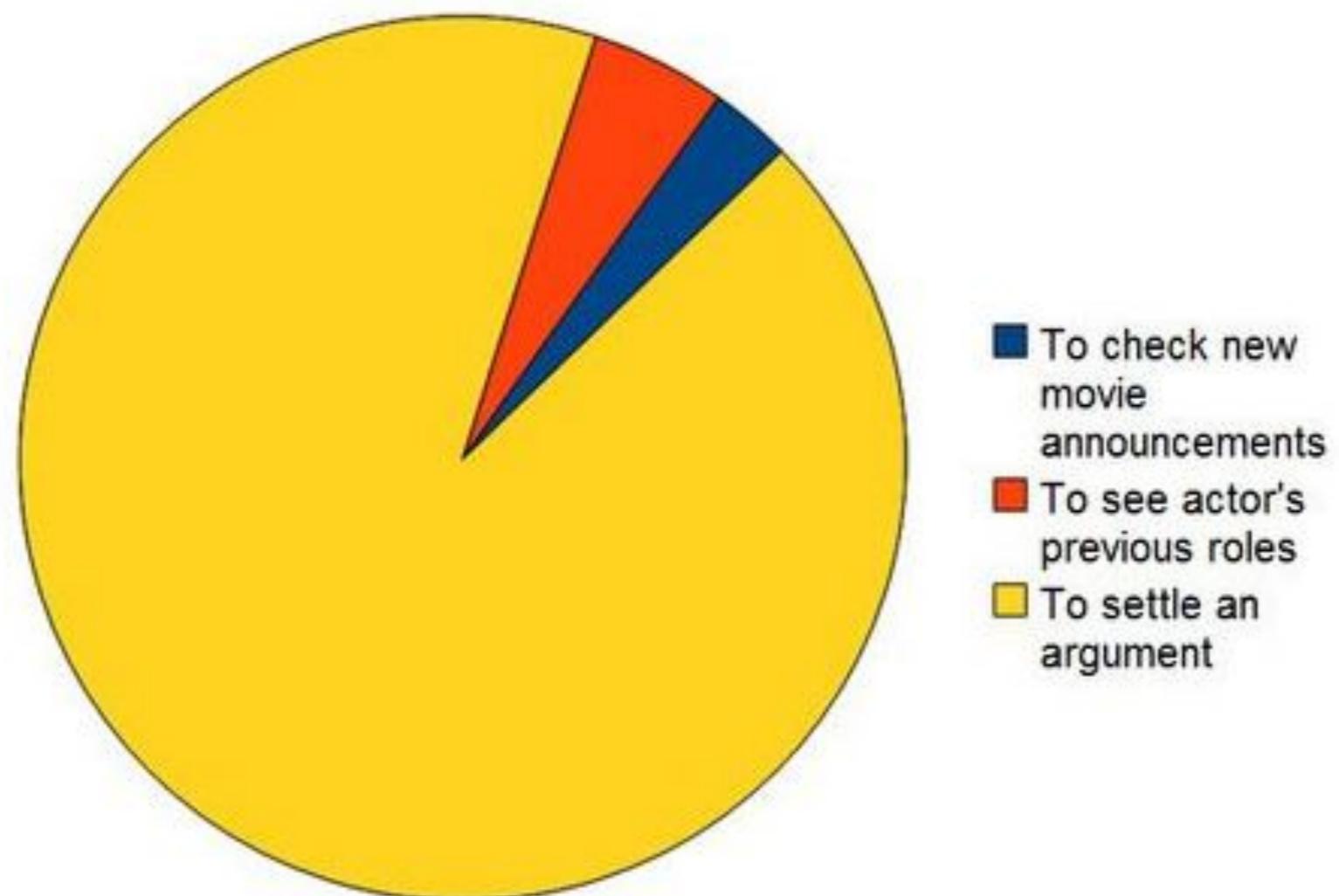


Итог

• •

Отзывы со спойлерами будут иметь меньше просмотров	+
Спойлеры скорее будут писаться к фильмам/сериалам, которые не понравились пользователю	+/-
Спойлеры скорее будут писаться к сериалам (обратная зависимость)	+/-
Больше просмотров будет у отзывов с низкой оценкой	-
Спойлеры скорее будут писаться к фильмам с низким рейтингом	-
Чем старше фильм, тем вероятнее наличие спойлеров	-

Why I use IMDB



Спасибо за внимание!