МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Введение в анализ данных

YouTube, YouTube 2

Автор:
Дандаева Баина
Б06-106

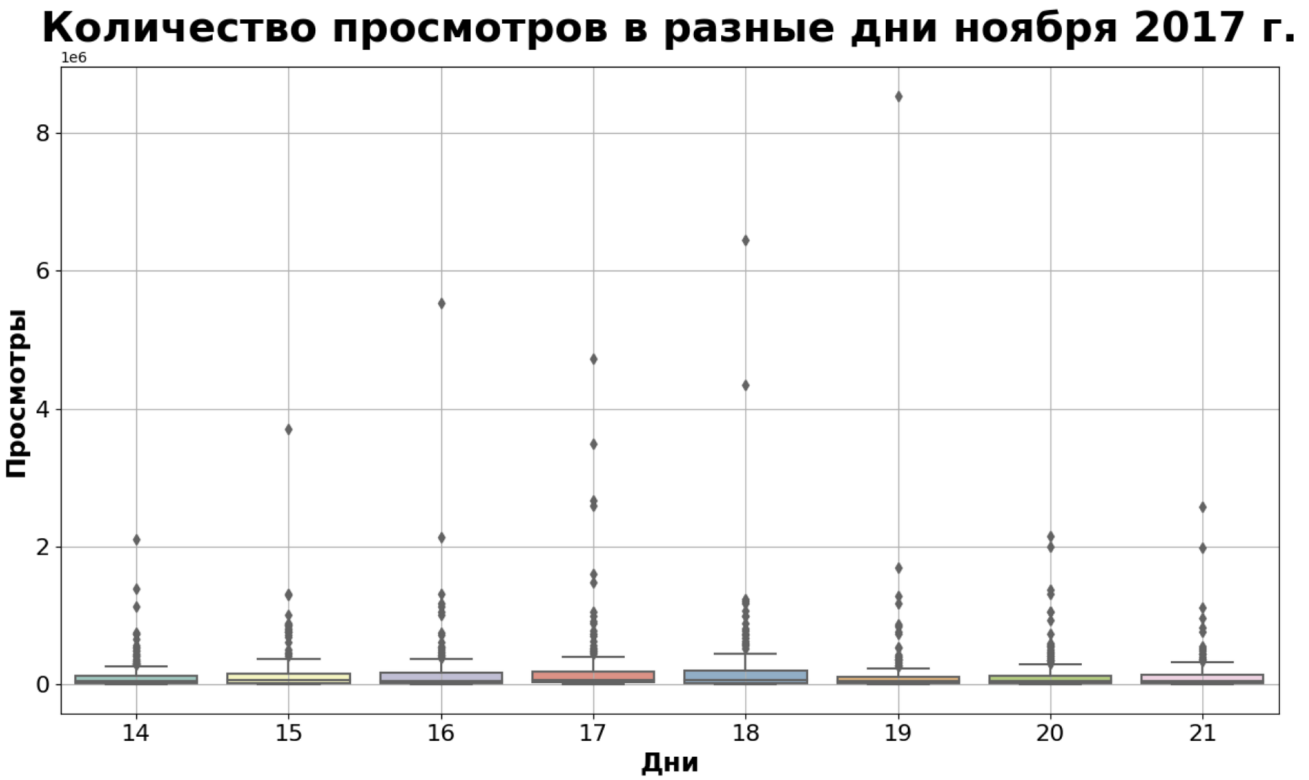
Долгопрудный 2023

Цель работы: Изучив и проанализировав данные видеороликов российского сегмента YouTbe, визуализировать полученную и обработанную информацию, используя разные графические представления библиотеки seaborn.

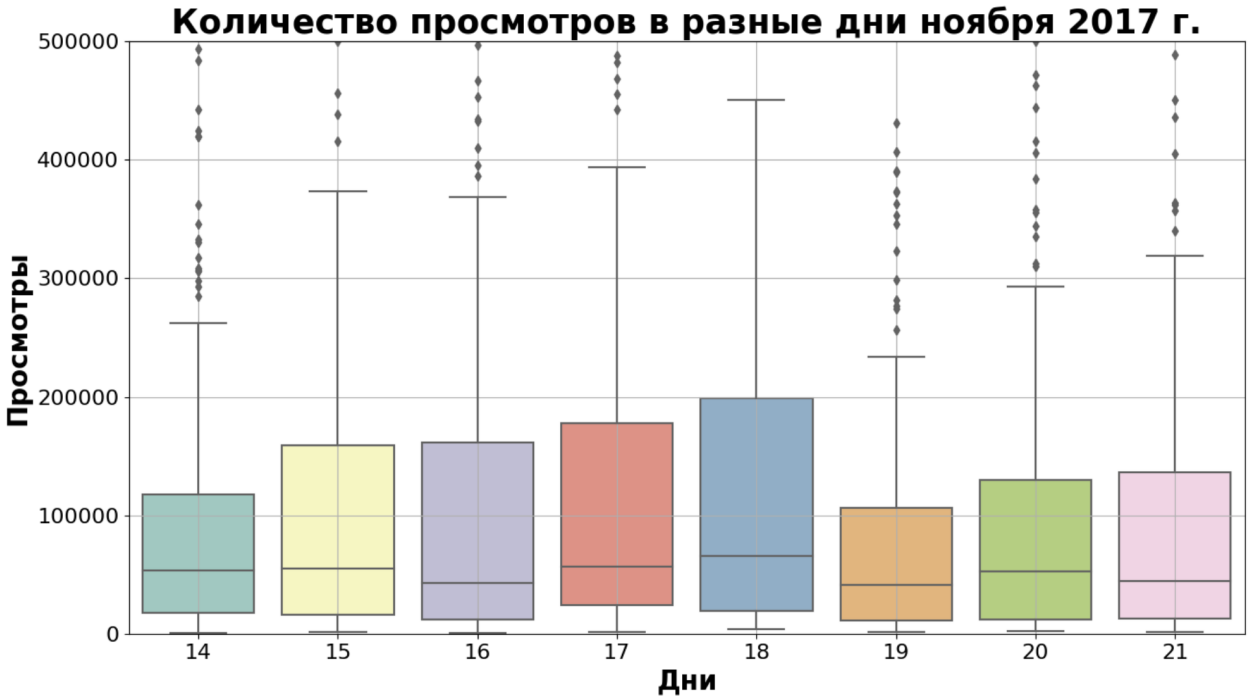
1 Youtube

1.1 Boxplot

Рассмотрим количество просмотров по дням, зафиксированных в ноябре 2017 года. Визуализируем данные с помощью Boxplot:



При таком посстроении графика мы получаем неинформативный формат, так как м не установили предел. Выбросы сильно влияют на общую картину, тем самым снижается информативность графика.



Используя метод plt.ylim() задали диапазон значений по оси y, что улучшило визуальное представление данных. Можно заметить, что мало видеороликов набирают больше 400 000, то есть на данном графике выбросов больше всего в диапазоне от 400 000 до 500000. Больше всего просмотров собиарется в 17 и 18 ноября 2017 года(пятница и суббота), их медианы и верхний квартиль выше других. Также в эти дни даже самые непопулярные видео набирают больше просмотров(по уровню минимальных значений).Менее популярные дни для просмотров видеородликов являются вторник и воскресенье (14 и 19 ноября 2017г)

1.2 Joinplot

Рассмотрим графическое представление данных с помощью Joinplot. Проверим есть ли зависимость между количеством просмотров видеоролика и количеством лайков:

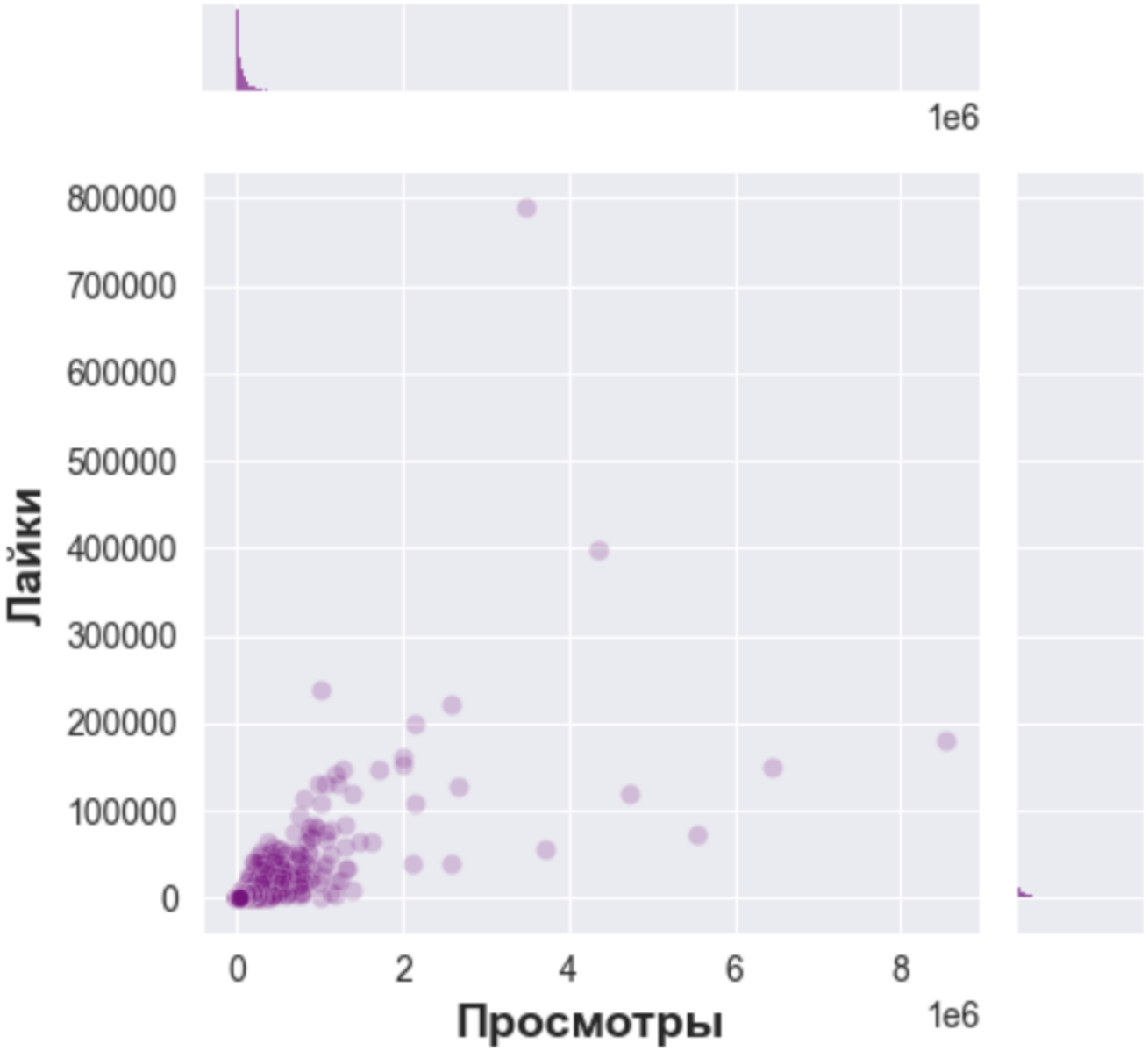
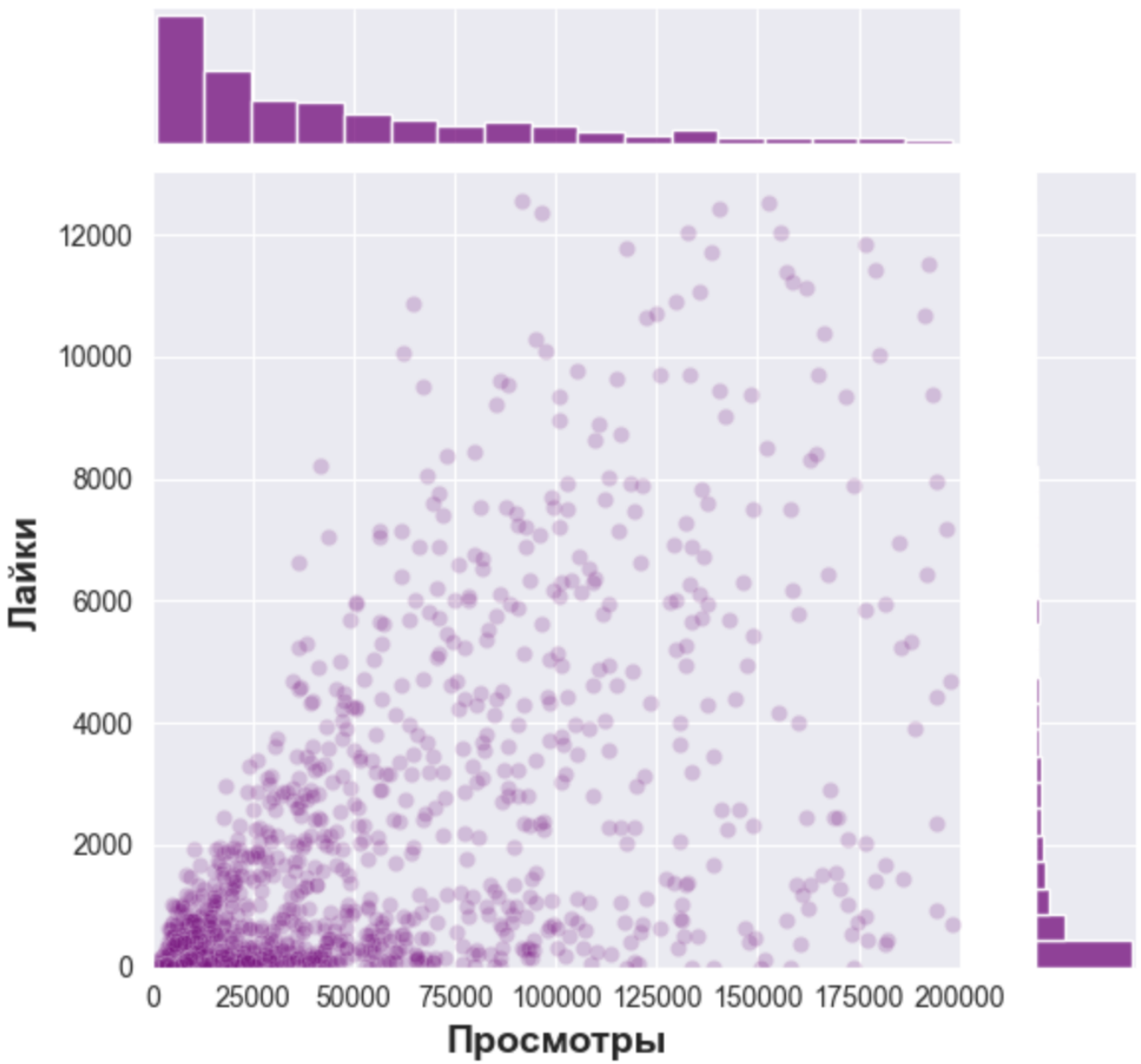


График без установки пределов снова неинформативен, так как большая плотность в значениях до 13 000 лайков и 200 000 просмотров. В данных значениях установить зависимость сложно. График с использованием plt.ylim():



Наибольшая плотность сосредоточена до 15 000 лаков и 30 000 просмотров, поэтому в данных значениях сложно сказать какая наблюдается зависимость между двумя параметрами. При рассмотрении от 20 000 лайков и 50 000 просмотров в большинстве случаев можно уже сказать, что число лайков прямо пропорционально количеству просмотров. То есть пользователи будут больше ставить лайков, если видео более популярное на просторе YouTube.

2 Вывод

С помощью метода boxplot выявили:

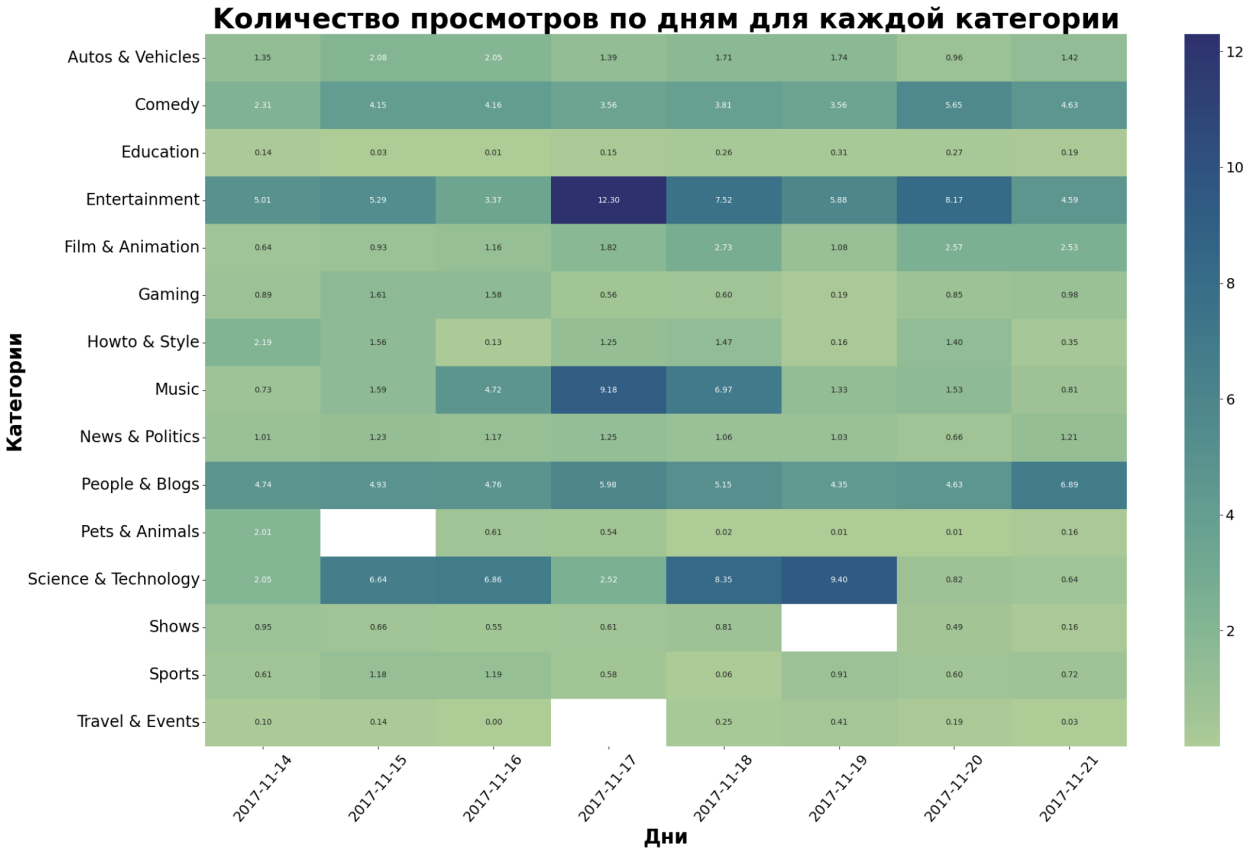
1. Больше всего просмотров видео набирают по пятницам и субботам
2. Меньше всего просмотров видео набирают по вторникам и воскресениям

С помощью метода joinplot выявили, чем более популярное видео на YouTube, тем больше лайков будут ставить пользователи.

3 Youtube 2

3.1 Heatmap task 3

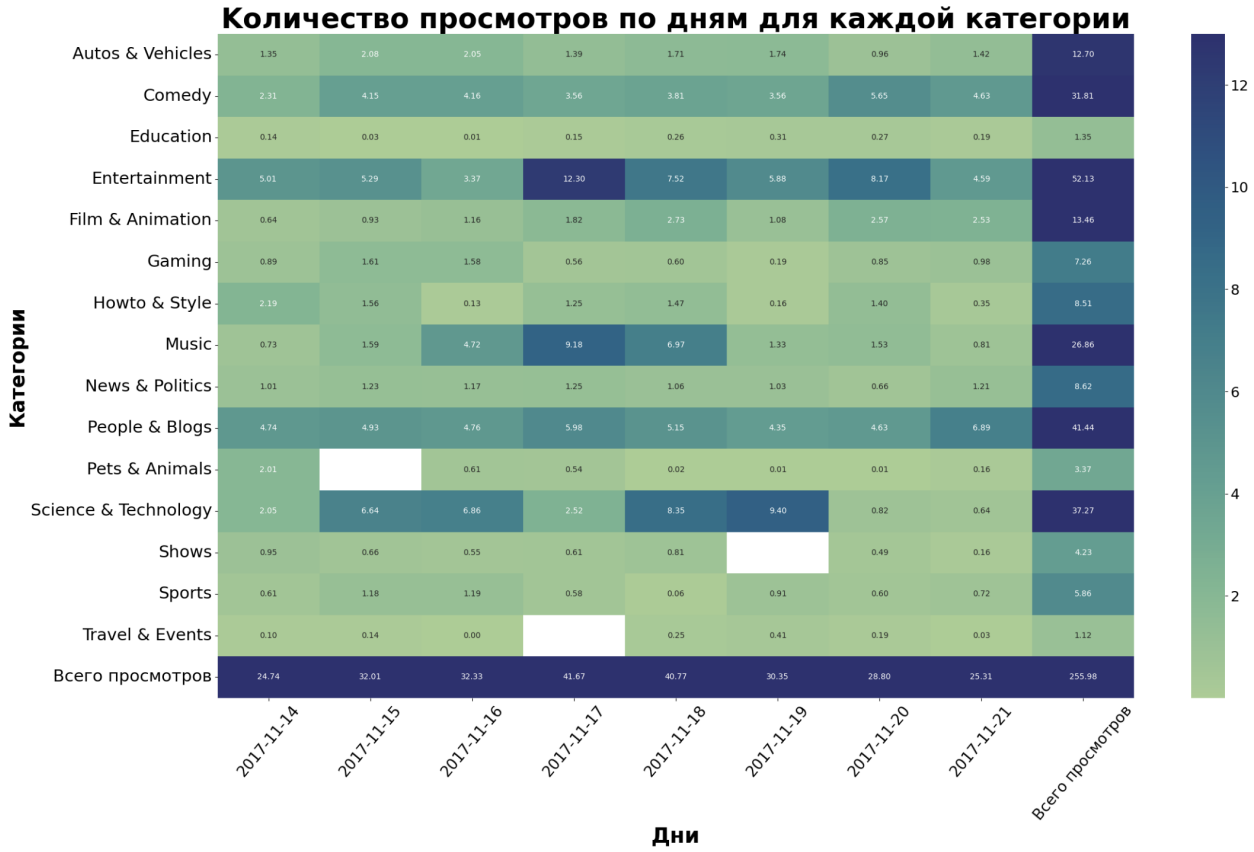
Здесь видеоролики мы разбили по категориям, чтобы выяснить какая же категория пользуется спросом среди пользователей



Больше всего просмотров набирают категория 'Entertainment', 'People&Blogs' и 'Science & Technology'. Меньше всего просмотров 'Education', 'Travel&Evants', 'Pets&Animals'. Максимум просмотров 12.3 миллионов по категории 'Entertainment.

3.2 Heatmap task 5

В данном графике добавили суммы по категориям и по дням. Так же, как и в прошлом разделе, если не использовать дополнительные условия, то график не информативен, так как все цвета сольются в один . Поэтому здесь нужно добавить параметр vmax, тогда будут видны различия между цветами ячеек.



Топ общему количеству просмотров являются 'Entertainment', 'People&Blogs', 'Science & Technology'.
И все тот же топ по меньшему количеству: 'Education', 'Travel&Evants', 'Pets&Animals'(подтверждение прошлого вывода)
Также можем заметить подтверждение вывода о том, что самыми популярными днями для просмотра - это пятница и суббота, непопулярными- вторник и воскресенье.

4 Вывод

С помошью метода heatmap выявили:

1. Больше всего просмотров набирают категории:'Entertainment', 'People&Blogs' и 'Science & Technology'.

2. Меньше всего просмотров набирают категории: 'Education', 'Travel&Evants', 'Pets&Animals'.

Также подтвердили вывод прошлого метода о популярности дней недели по просмотрам видео.