МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Введение в анализ данных

YouTube

Автор: Дубенков Родион

Б05-112

Физтех-школа: ФПМИ
Направление: КТ

Долгопрудный 2022

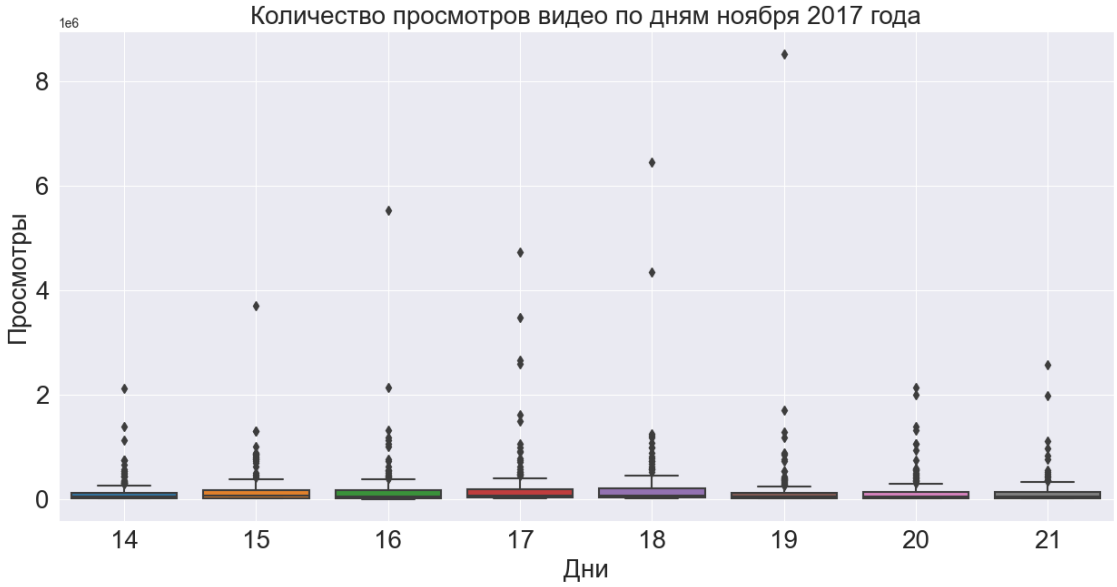
1 Аннотация

Цель работы: Проанализировать и обработать исходный набор данных, представленный в виде таблицы с информацией по каждому опубликованному видео на платформе YouTube российского сигмента за определенный период времени. Визуализировать данные с помощью библиотеки *seaborn* и выявить в них закономерности.

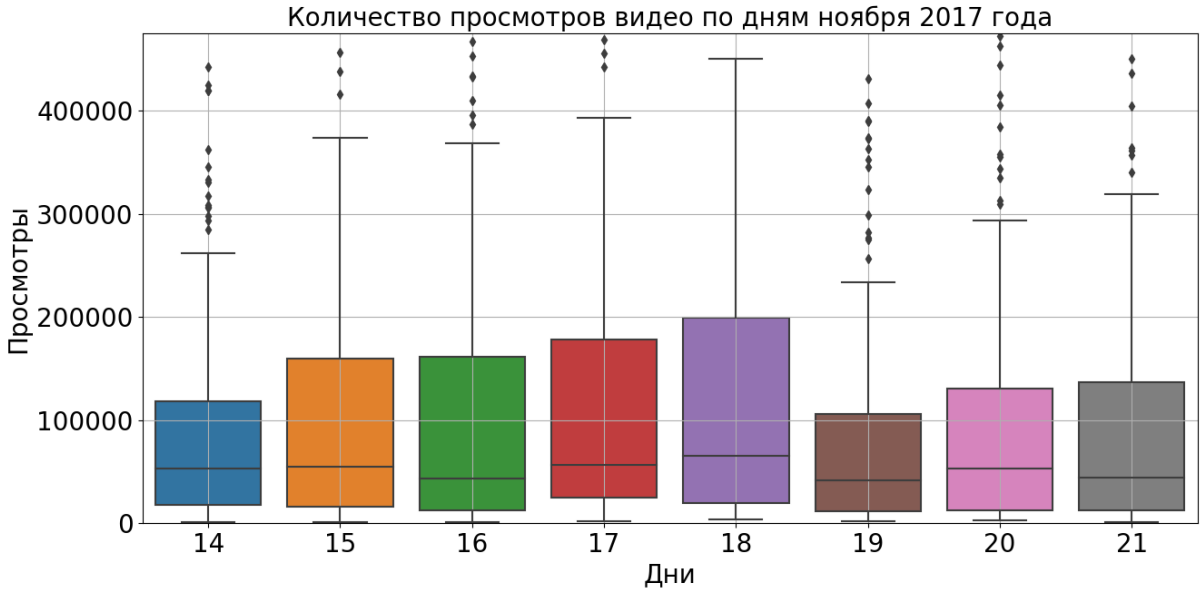
2 YouTube 1

2.1 Boxplot

Рассмотрим диаграмму, описывающую количество просмотров по дням. По ней мы сможем определить активность пользователей в зависимости от дня недели. С учётом того, что 14 ноября 2017 года является вторником.



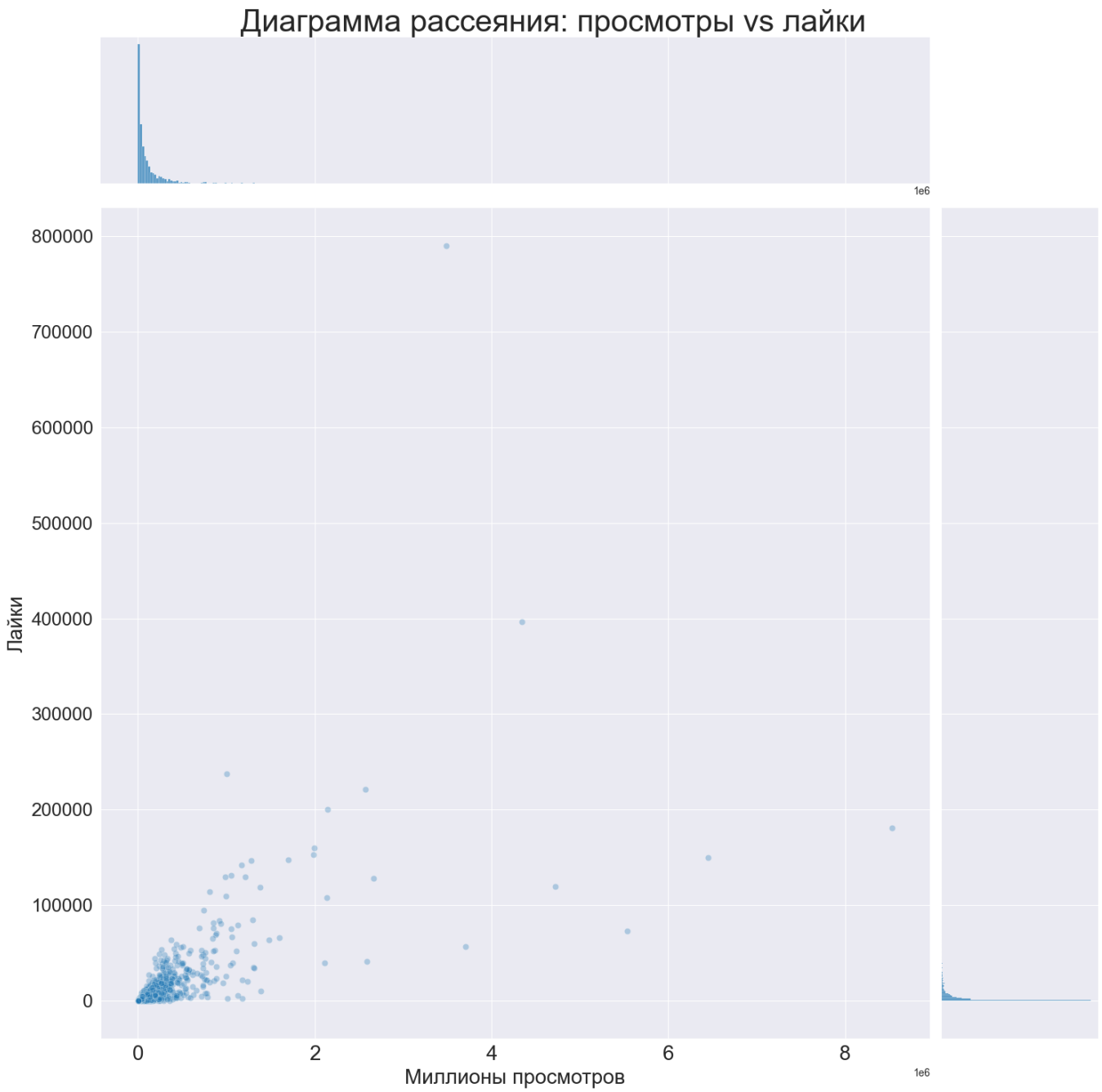
По вышепоказанной диаграмме достаточно проблематично отпределить в какой день недели активность пользователей выше. Нам мешают выбросы, которые уменьшили масштаб картинки. Построим более информативную диаграмму, задав пределы вертикальной оси $[0; 475000]$. Важно заметить, что мы не выкидываем все выбросы, так как они тоже несут в себе важную информацию.



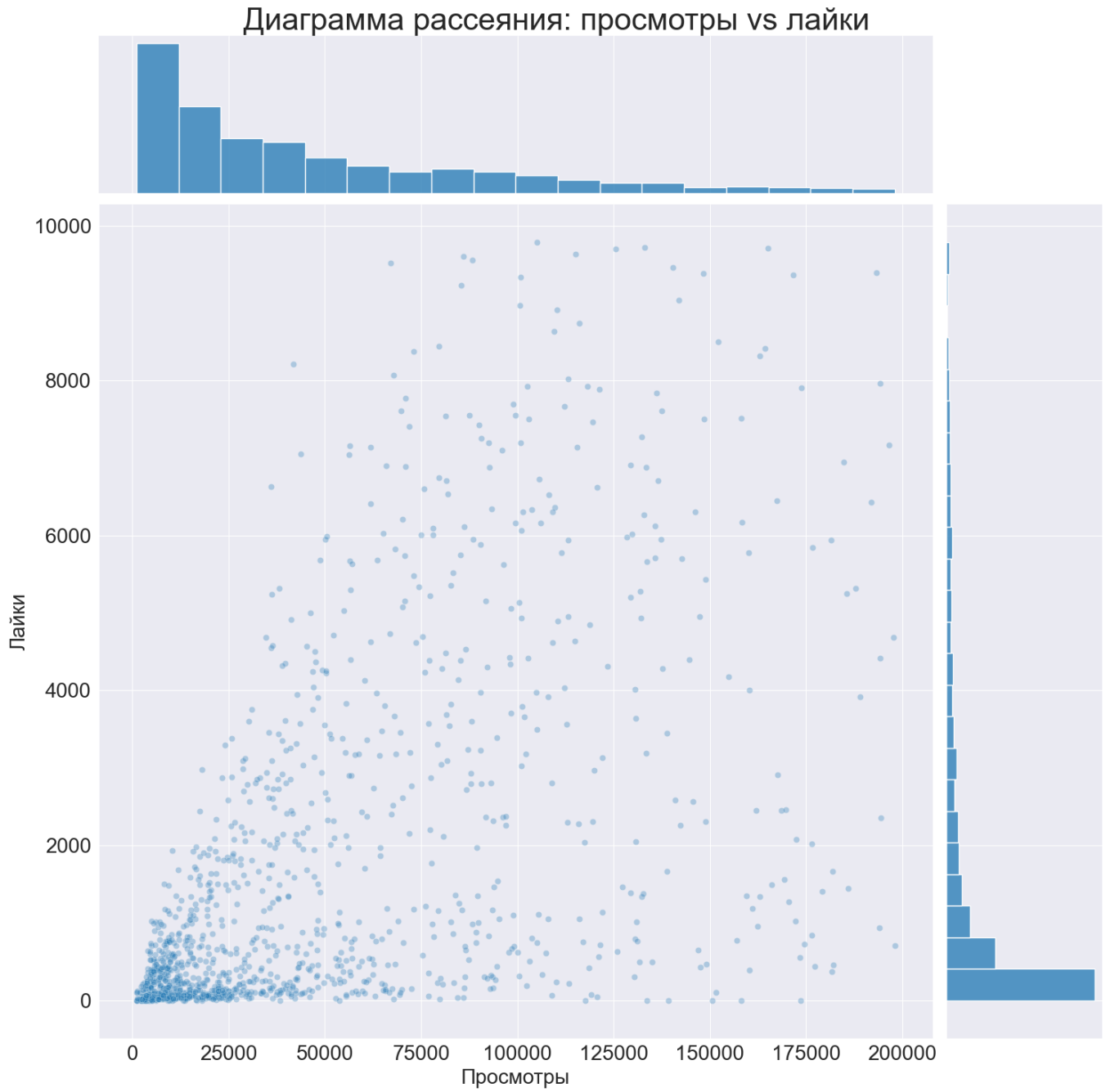
По такой диаграмме гораздо проще сделать вывод: в воскресенье и вторник активность пользователей сервиса YouTube - наименьшая, в пятницу и субботу - наибольшая.

2.2 Joinplot

Теперь посмотрим на диаграмму рассеяния, описывающую зависимость количество просмотров от количества лайков на видео.



Опять же из-за выбросов масштаб диаграммы уменьшился. Давайте посмотрим на зависимость в таких пределах: количество лайков - $[0; 10000]$, а количество просмотров - $[0; 200000]$.



По картинке несложно понять, что спектр разброса точек довольно большой. Есть видео с большим количеством просмотров, но которые набрали мало лайков, и наоборот. Но при этом прослеживается чёткая зависимость: чем больше просмотров, тем больше лайков.

2.3 Вывод

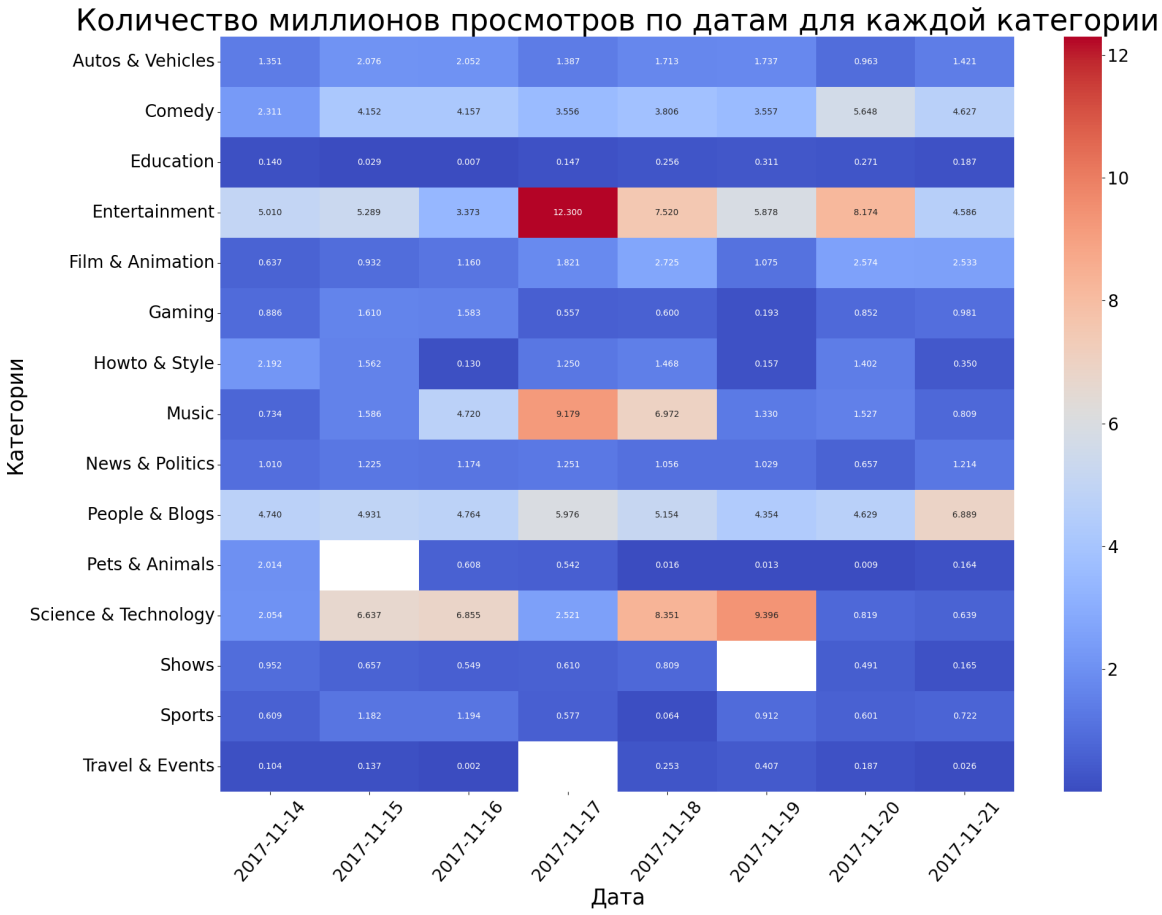
1. Активность пользователей YouTube выше по пятницам и субботам, и ниже по воскресеньям и вторникам.

2. Чем больше просмотров, тем больше лайков на видео.

3 YouTube 2

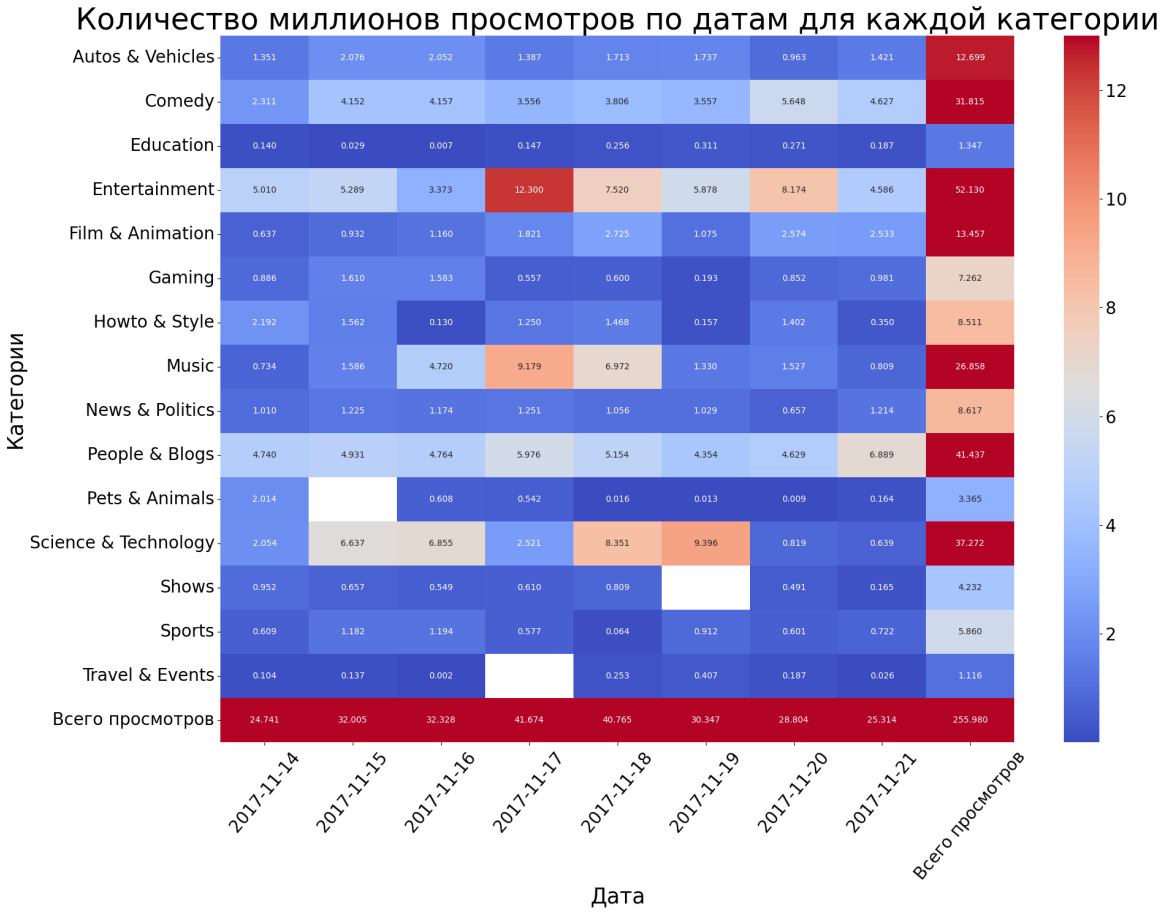
3.1 Heatmap

Далее посмотрим на диаграмму ниже, и оценим, какие категории люди больше предпочитают смотреть.



По данной картинке можно сделать вывод, что люди предпочитают смотреть больше категорию видео: Entertainment и People & Blogs. Меньше всего предпочитают - Education и Travel & Events. Однако нам необходимо знать информацию «на дистанции», то есть сумммарно по дням.

И чтобы тепловая карта осталась информативной, мы добавим параметр $vmax = 13$. То есть мы вручную изменили диапазон количества промотров. Иначе картинка автоматически подстроится под максимальное значение категории "Всего просмотров"и будет сложно визуально отличить значения ячеек.



В данном случае, информация по количеству просмотров в сумме по дням подтвердила предыдущий вывод.

3.2 Вывод

1. Самые популярные категории: Entertainment и People & Blogs
2. Самые непопулярные: Education и Travel & Events.