Введение в анализ данных

Отчёт по задаче
Youtube

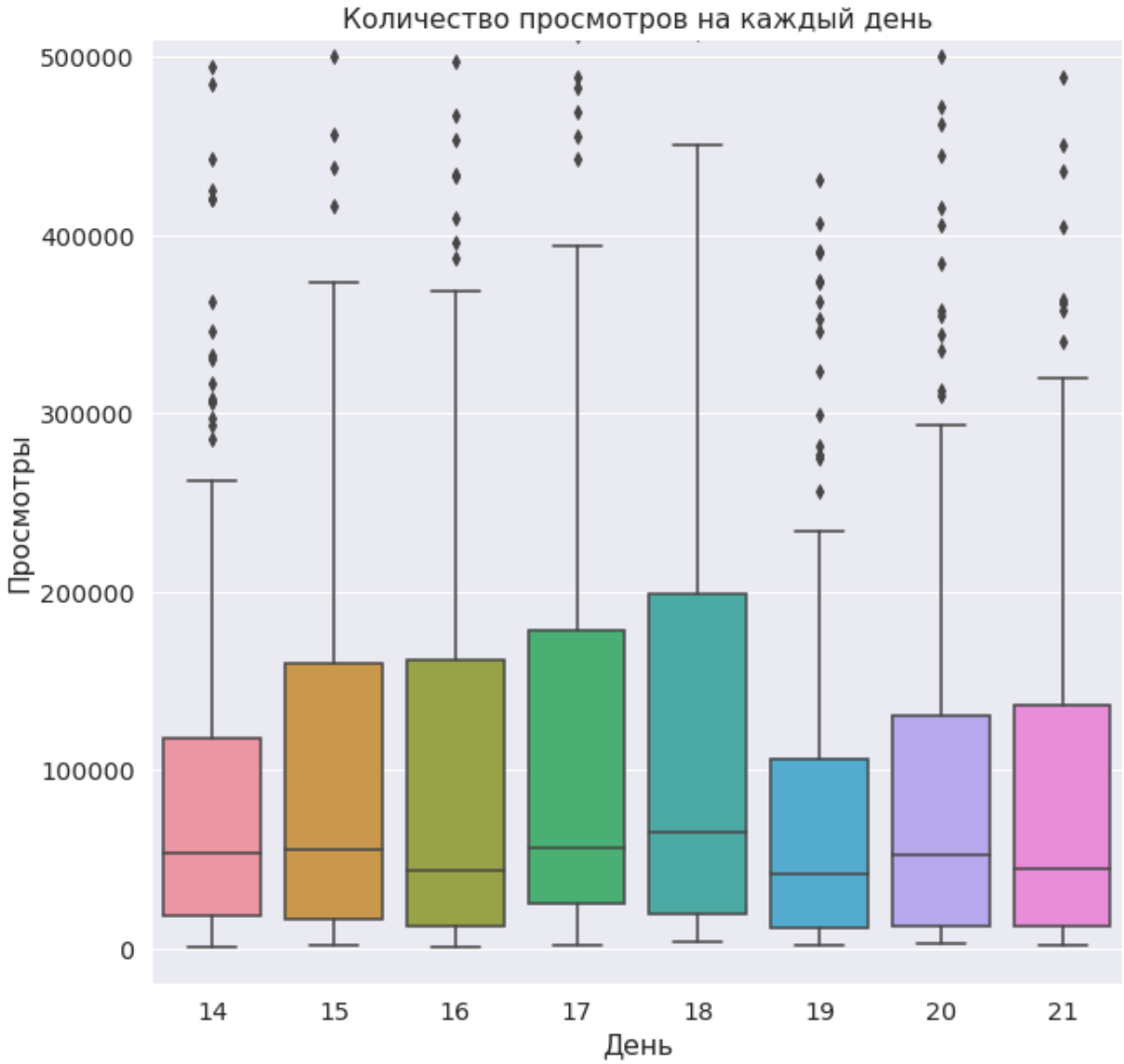
Артём Юков
Б05-102

Постановка задачи

У нас есть данные о видео с русскоязычного YouTube'а за период с 14 по 21 ноября 2017 года. Мы хотим исследовать корреляции между данными, построив графики по различным параметрам.

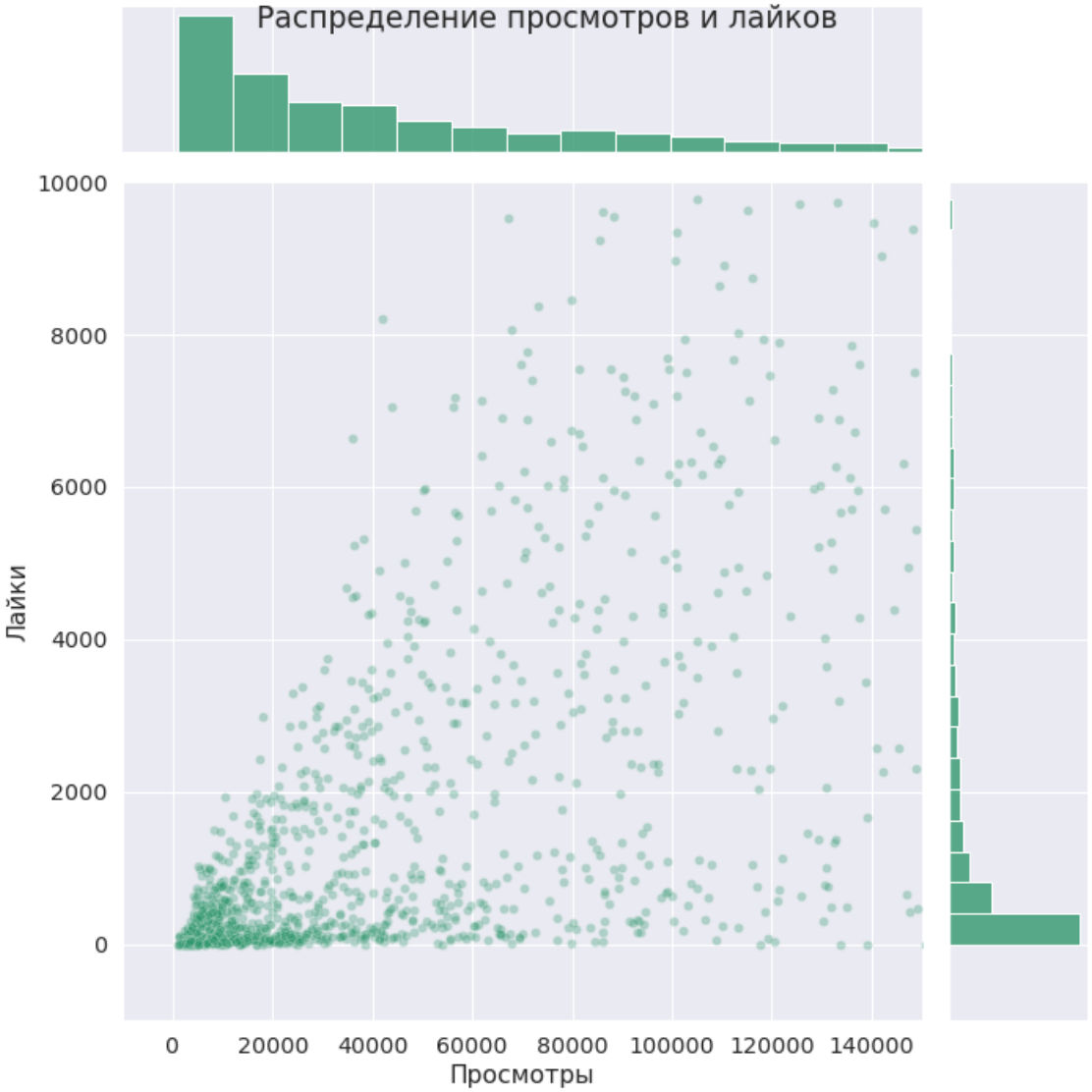
Просмотры по дням

Сперва, посмотрим на распределение просмотров по каждому дню. Для этого построим boxplot. Так как мы имеем большое количество выбросов, что мешает читаемости графика – ограничим количество просмотров полумиллионом, тогда "усы"ящиков будет видно чётче и масштаб станет удобным для анализа.



Лайки и просмотры

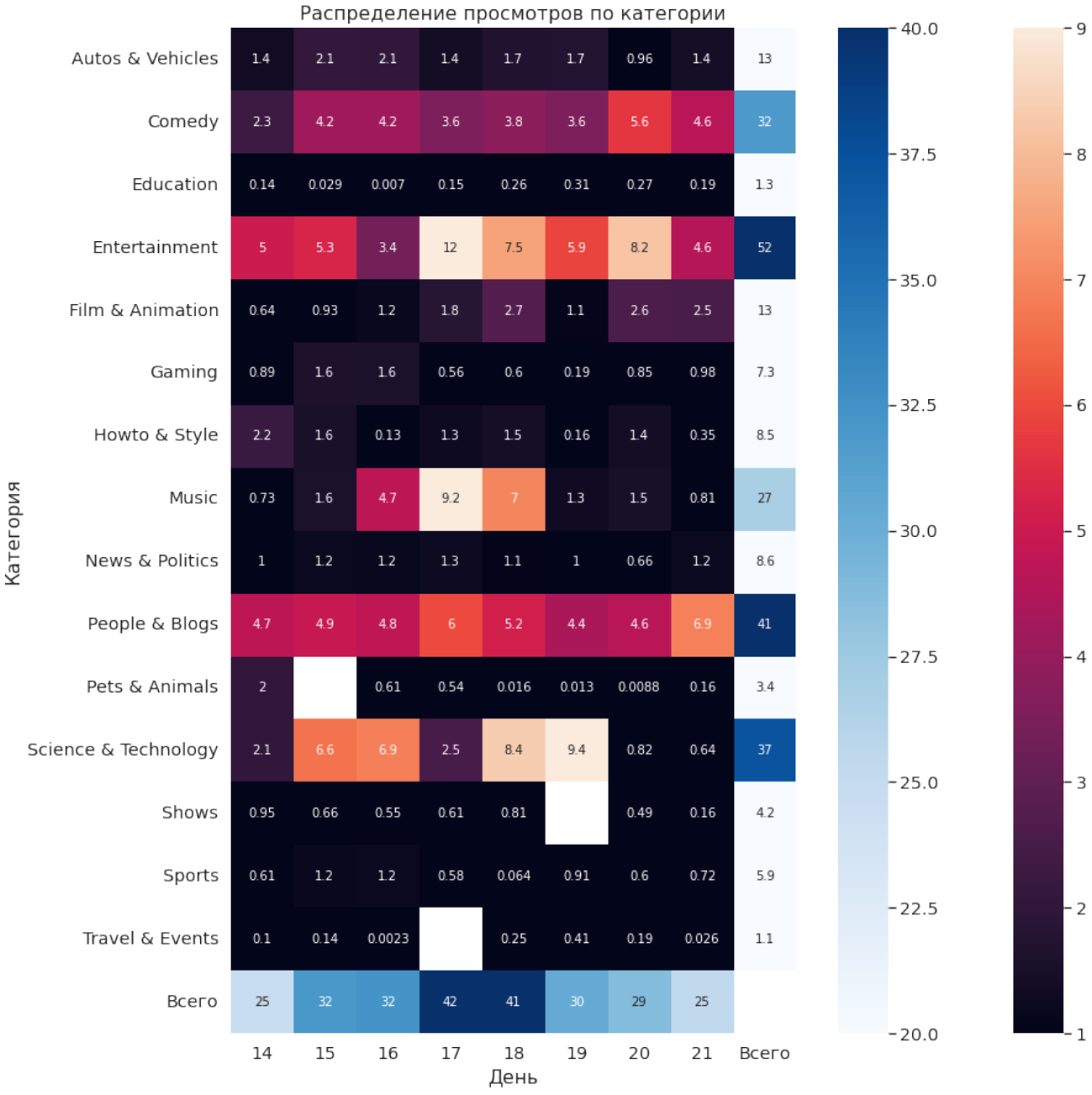
Нам также интересно пронаблюдать за распределением по просмотрам и лайкам одновременно. Для этого построим jointplot по этим двум параметрам. Для большей информативности увеличим прозрачность, чтобы сгущения точек выделялись сильнее.



Категории

По расширенной таблице с категориями видео построим heatmap зависимости числа просмотров (в миллионах) от дня и тематики контента. Пропуски в данных практически не мешают, поэтому оставим их как есть.

Для большей информативности введём две шкалы: для дневных просмотров и просмотров за неделю, грамотно подобрав границы, а также введём аннотацию для каждой ячейки.



Выводы

После анализа этих графиков напрашивается несколько выводов:

- число просмотров 17 и 18 числа было наибольшим. Довольно логично – в пятницу и субботу человек хочет отдохнуть от прошедшей рабочей недели и выделяет свободное время на просмотр YouTube.

- тогда встаёт вопрос, почему просмотров в воскресенье меньше всего. Скорее всего, после двух дней релаксации новые видео заманивают человека сильно меньше.

  Отмечу, что в таблице на каждый день записано ровно по 200 видео, поэтому выводы являются справедливыми.

- люди редко ставят лайки. Причём с ростом числа просмотров среднее количество лайков растёт примерно так же.

- самые популярные категории – 'Развлечение' и 'Блог', а наименее популярные – 'Образование' и 'Путешествия'.

Замечание

И проведём небольшое исследование по последнему графику: вышеупомянутые тематики более–менее стабильны, но, например, категории 'Наука и Технологии' и 'Музыка' имеют некие выбросы. Я решил разобраться в этом, и оказалось, что в первом случае на это повлиял анонс двух новых роботов Boston Dynamics, чьи видео занимают три первые позиции по просмотрам за рассматриваемый период, а во втором – выход 'Розового вина' и новой песни группы BTS,

- значит, в целом эти категории не так интересны, но, например, если у артиста выходит новый альбом или удаётся получить фотографию чёрной дыры, то люди могут вывести категорию в топ.

Так же к этим категориям можно отнести 'Политика и новости', 'Спорт' и 'Игры'. Просто нам не повезло увидеть какого-либо связанного инфоповода за рассматриваемую неделю.