# Анализ некоторых данных с YouTube

Второе домашнее задание по курсу "Введение в анализ данных"

Якушев Георгий Б05-111

## Введение и задачи

В данной работе были исследованы данные с платформы YouTube за определенный промежуток времени. Несмотря на очень специфичные данные, удалось установить некоторые качественные зависимости количества просмотров в зависимости от некоторых параметров.

## Исходные данные

Исходными данными были данные о видео в российском сегменте YouTube с 14 ноября по 21 ноября 2017 года. Полная версия данных лежит на платформе *Kaggle*.

## Методология

Для анализа этих данных использовались базовые методы визуализации данных. Для обработки и исследования были использованы инструменты библиотек *python*: *matplotlib*, *pandas*, *seaborn*, *numpy* и др.

### Количества просмотров по дням

Для начала были изучены дни недели для предоставленных данных. Для ноября 2017 года.

| Дата | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|
| День | Вт | Ср | Чт | Пт | Сб | Вс | Пн | Вт |

Сгруппировав данные о просмотрах по дням, были построены коробки с усами на каждый из этих дней.

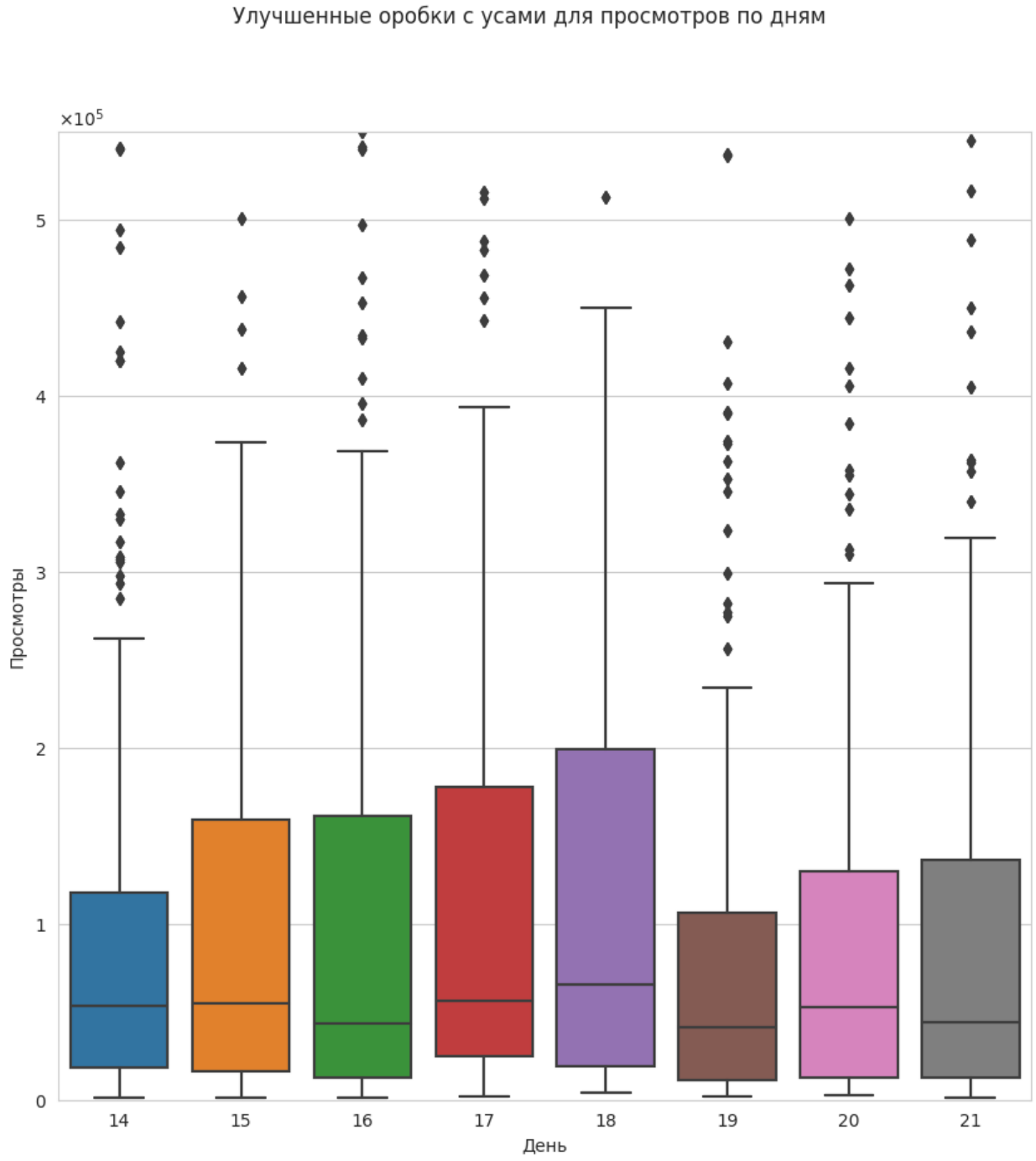


*Рис. 1 Коробки с усами для просмотров видео по дням ноября 2017 г.*

По графику видно, что на протяжении этих восьми дней количество просмотров возрастало. 12 ноября 2017 года - вторник, а значит на 18 число приходится суббота. Локальный пик по просмотрам приходится на субботу, а локальный минимум на воскресенье (18 и 19 числа ноября соответственно). Также по отношению медианы и бокса можно понять, что видео распределены неравномерно по просмотрам.

## Зависимость лайков и просмотров

Для исследования этой зависимости был построен присоединенный график лайков и просмотров.

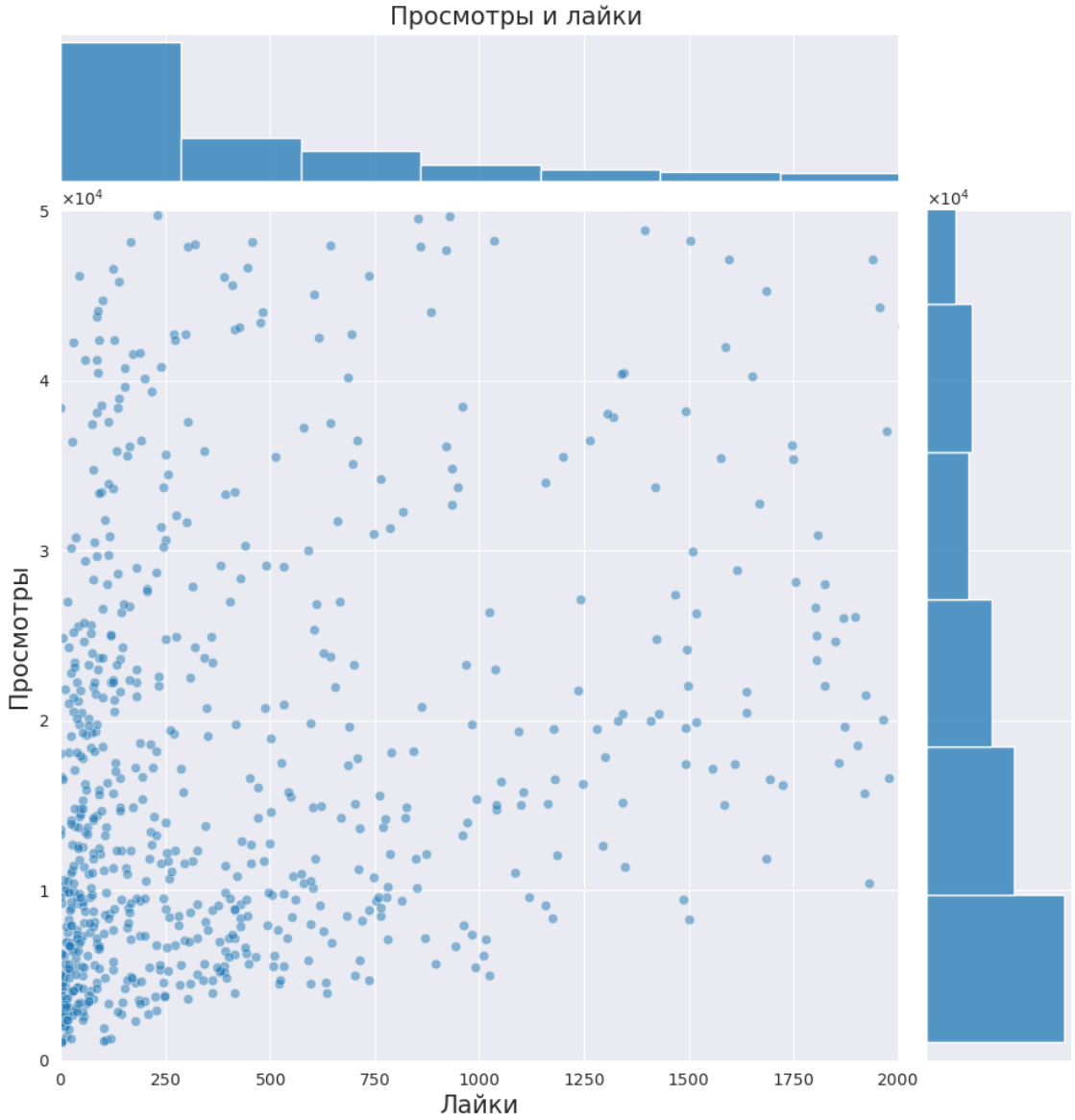


*Рис. 2 Присоединенный график Лайков и Просмотров видео с YouTube.*

Исходя из полученных данных можно сказать, что количество видео от количества лайков или просмотров падает экспоненциально, о чем свидетельствуют гистограммы по бокам. При этом большинство видео сгруппировано возле области с малым числом лайков и возможно большим количеством просмотров.

## Распределения просмотров по категориям

Построенные тепловые карты хорошо показывают особенности этих данных. Здесь одна карта - карта по просмотрам, другая - нормированная по суммарным просмотрам в день карта с добавленными колонкой суммарных просмотров.

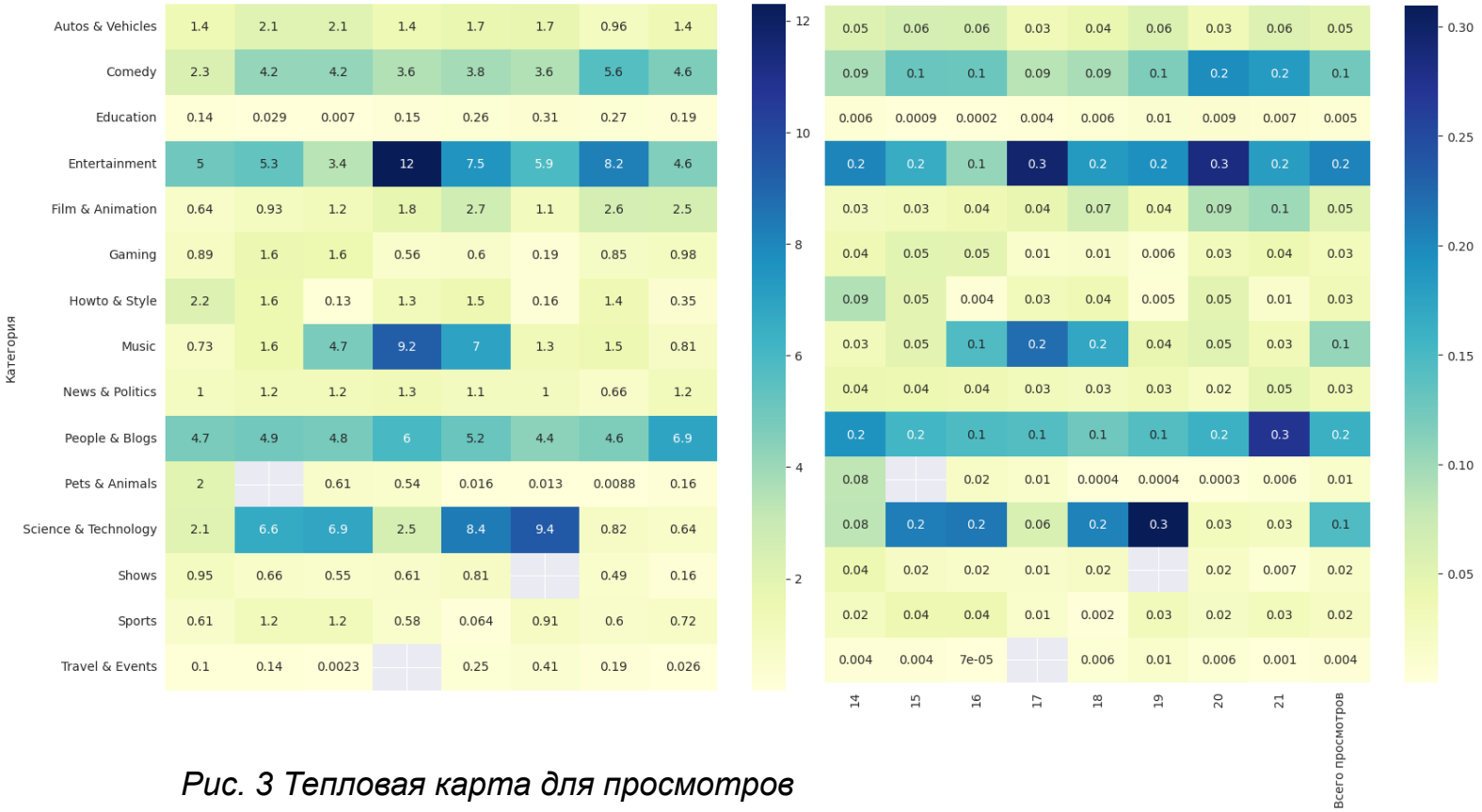


*Рис. 3 Тепловая карта для просмотров по категориям.*

*Рис. 3 Нормированная тепловая карта для просмотров по категориям.*

На этих тепловых картах видно, что среди категорий выделяются 4 самые популярные: *Entertainment*, *People & Blogs*, *Comedy*, *Science & Technology*.

Популярность многих категорий в среднем увеличивается ближе к выходным: *Entertainment*, *Science & Technology*. Другие категории распределены по дням равномерно, за исключением тех, о которых отсутствуют данные.

# Выводы

Стоит отметить, что изначально данные ограничены русским сегментом, некоторыми популярными видео, а также всего 8 днями. Несмотря на это получено некоторое качественное представление о поведении просмотров на видео на YouTube в зависимости от некоторых параметров: дней, категорий, лайков:

- В среднем число просмотров растет ближе к субботе и падает в воскресенье
- Число видео с определенным числом лайков или просмотров падает экспоненциально
- Самыми популярными категориями являются: *Entertainment*, *People & Blogs*, *Comedy*, *Science & Technology*
- А популярность следующих категорий растет вблизи выходных: *Entertainment*, *Science & Technology*.

В ходе этой работы были изучены только общие зависимости на качественном уровне. Соответственно не были изучены выбросы и их поведение. А также сильная специфика данных не позволяет делать количественных выводов.