# Домашнее задание 2. Библиотеки pandas и matplotlib

Куринова Ксения, Б05-011

24 февраля 2022 г.

## YouTube

### Постановка задачи

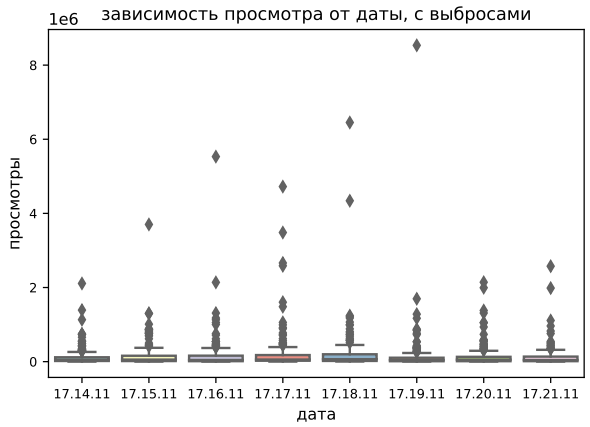
Задача заключается в работе с данными о трендах на YouTube. В работе использована библиотека seaborn, которая была рассмотрена на одной из последних лекций.

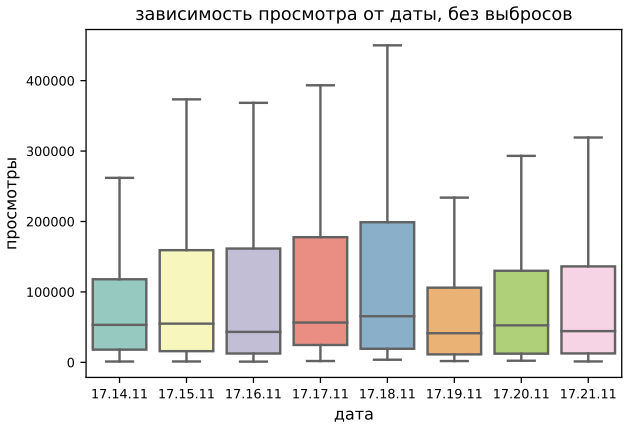
### Подготовка данных

В ходе выполнения работы были считаны данные и изменена исходная таблица:оставлены только дни и изменен исходный формат поля даты.

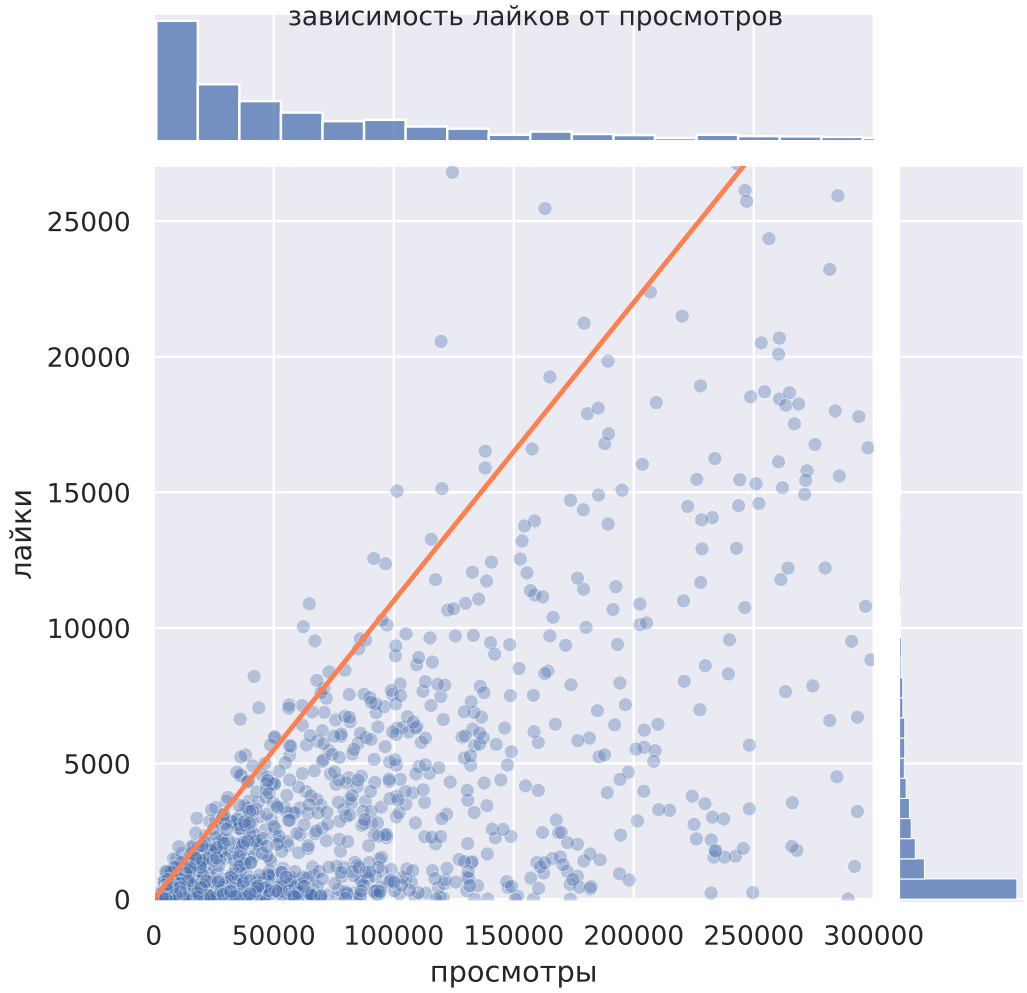
### Некоторая визуализация

Построим ящики с усами по количеству просмотров. Заметим, что выбросы сильно снижают информативность графика, приведём исправленную версию. И изначальную для примера:




Таким образом, просмотры видеороликов в среднем не превышают $150-200$ тысяч в сутки. Далее построим объединённый график по количеству просмотров и лайков.




### Выводы

В результате проделанной работы мы увидели, что бОльшая часть точек на графике числа просмотров лежит ниже некоторой прямой, что позволяет нам оценить сверху число лайков, зная просмотры. Чем меньше будет число просмотров, тем точнее будет наша оценка – большее число точек будет лежать в нашей выборке. В среднем, отношение лайков к просмотрам равно $0.8 - 0.12$, что говорит о том, что gпримерно каждый десятый зритель лайкает видео.

Имеем, что график просмотра от даты даёт нам понять: люди смотрели видео всё чаще с понедельника к субботе, но в воскресенье просмотры упали. Суточные просмотры не превышают двухста тысяч в сутки. Возможно, динамика просмотра привязана к рабочей неделе и к ожиданию выходных.

## YouTube2

### Постановка задачи

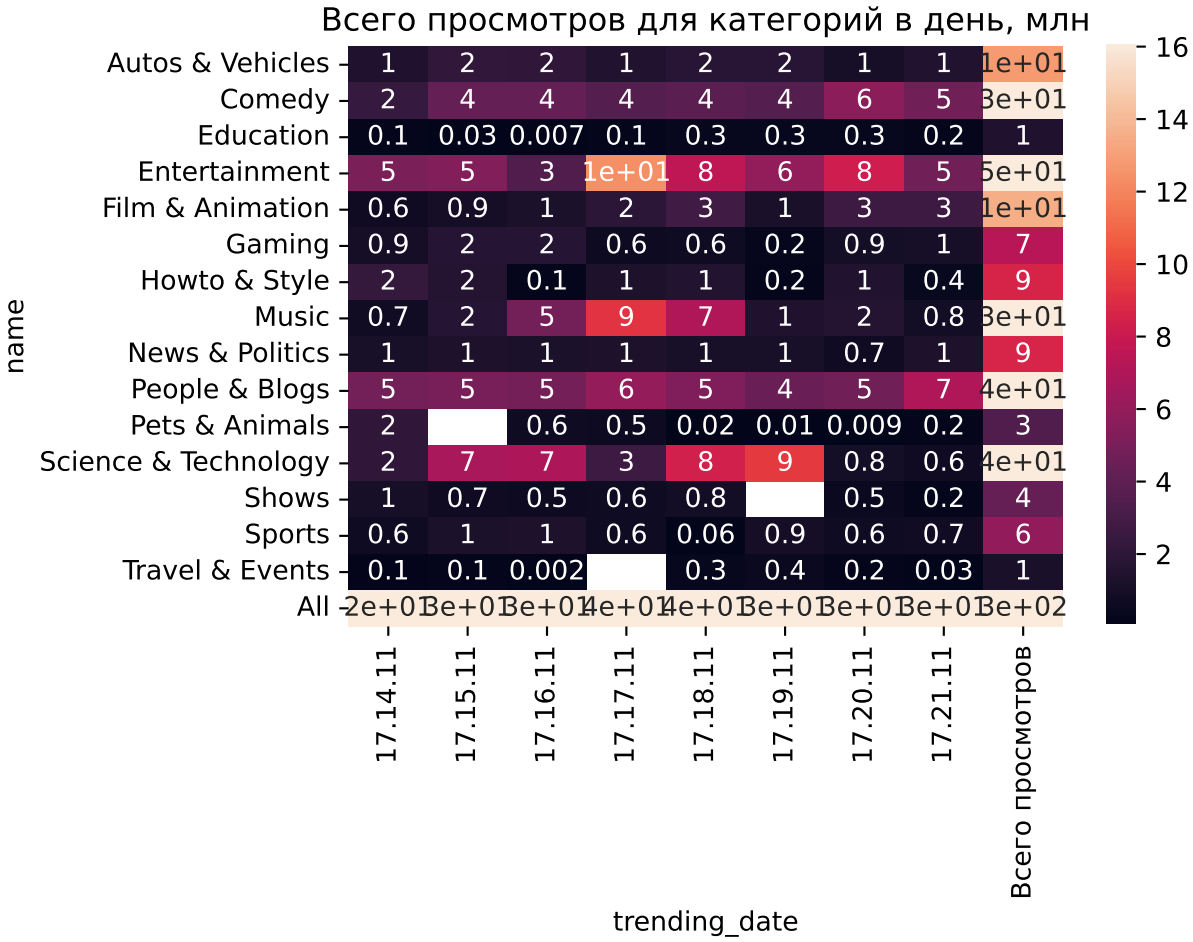
В данной задаче нам нужно будет продолжить анализ данных о видео на YouTube.

### Обработка данных

Объединим две таблицы на основе индекса категории, составим свободную таблицу о количестве просмотров. Для информативности поделим всё на $10^6$.

## Визуализация

Рассмотрим визуализацию данных, полученную при выполнении задания:




Заметим, что регулировка количества цветов позволила графику обрести более лаконичную визуальную форму. Динамика просмотров стала более наглядной.

## Выводы

Рассмотрим различные категории контента:

1. Развлекательный. У данного вида контента максимальное число просмотров в неделю и за день. Он популярен преимущественно в выходные дни.

2. Science & Technology. Заметим, что второй по популярности контент имеет самые высокие просмотры в будние дни. Возможно это связано с тем, что процесс обучения проходит именно по будним дням.

3. Заметим, что видео, связанные с игровой индустрией, приобретают популярность в выходной, что может подтверждать гипотезу, изложеннукю в предыдущем пункте. Когда люди перестают учиться – они играют и отдыхают :)

4. Следующая по популярности категория это блогерский контент. Как и игровая индустрия, наиболее популярен по выходным. Как мне кажется, причины те же.