Я взял датасет из открытого источника: <https://www.kaggle.com/datasnaek/youtube-new>

Это статистика трендовых видео на YouTube. Я взял файлы RUvideos.csv и RU\_category\_id.json и загрузил эти данные в Excel как таблицы на листы trending и category соответственно. Таблица trending содержит информацию по всем трендовым видео России ежедневно за период 14.11.17 по 14.06.18. Таблица category это просто справочник, в котором каждому коду категории соответствует описание.

Тренды на YouTube – раздел, куда попадают видеоролики, набравшие наибольшее количество просмотров за десять-шестнадцать часов. Трендовые видео находятся здесь: <https://www.youtube.com/feed/trending>

Всего на лист trending загружено 40739 строк не включая заголовок. В таблице 16 столбцов.

Описание столбцов (переменных) ниже:

|  |  |  |
| --- | --- | --- |
| **Название колонки** | **Описание колонок** | **Тип данных** |
| video\_id | Код видео | String |
| trending\_date | Дата попадания в тренд | String (дата указана в формате ГГ.ДД.ММ) |
| title | Заголовок | String |
| channel\_title | Видео канал | String |
| category\_id | Код категории видео | ID внешний ключ |
| publish\_time | Дата публикации | Date |
| tags | Теги | String |
| views | Кол-во просмотров | Integer |
| likes | Кол-во лайков | Integer |
| dislikes | Кол-во дислайков | Integer |
| comment\_count | Количество комментариев | Integer |
| thumbnail\_link | Ссылка на изображение | URL |
| comments\_disabled | Признак отключенных комментариев | Boolean |
| ratings\_disabled | Признак отключенных рейтингов | Boolean |
| video\_error\_or\_removed | Видео имеет ошибку или удалено | Boolean |
| description | Описание видео | String |

По смыслу у данной таблицы составной ключ состоит из 2-х колонов video\_id и trending\_date, но по факту это не так. Присутствуют строки у которых все значения в столбцах идентичные или незначительно расходятся данные по кол-ву просмотров, лайков, дислайков. Фактически эта неточность получена из-за проблем в алгоритме извлечения данных.

Чтобы не перегружать файл вставлены первые 100 строк листа trending:



Все количественные данные в таблице можно считать дискретными, т.к. они принимают фиксированное количество различных значений. К тому же они конечны, лайки ограничены количеством зарегистрированных пользователей, просмотры тоже ограничены: с каждого IP адреса можно «накрутить» не более определенного количества просмотров

Данные на листе trending:



**Возможные варианты использования этого набора данных:**

* Анализ предпочтений пользователей.
* Классификация видео YouTube на основе их комментариев и статистики.
* Обучающие алгоритмы ML
* Предсказание популярности нового видео на YouTube.
* Статистический анализ с течением времени.