|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по лабораторной работе №2

**«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б   
Падалко К.Р.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2025 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc194515078)

[2. Описание задания 3](#_Toc194515079)

[3. Основные характеристики датасета 3](#_Toc194515080)

[4. Изучение данных 5](#_Toc194515081)

[5. Описательная статистика 6](#_Toc194515082)

[6. Предобработка данных 7](#_Toc194515083)

[7. Итог 11](#_Toc194515084)

[8. Вывод 11](#_Toc194515085)

# Цель лабораторной работы

Изучение способов предварительной обработки данных для дальнейшего формирования моделей.

# Описание задания

1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
   * + обработку пропусков в данных;
     + кодирование категориальных признаков;
     + масштабирование данных.

# Основные характеристики датасета

Название датасета: **Most Popular YouTube 1000 videos** (Самые популярные видео с YouTube 1000)

Ссылка: https://www.kaggle.com/datasets/samithsachidanandan/most-popular-1000-youtube-videos

**О датасетах**

Этот набор данных содержит информацию о 1000 самых популярных видео на YouTube. Он включает в себя различные параметры, такие как количество просмотров, лайков, дизлайков, категорию видео и год публикации. Этот датасет предоставляет возможность исследовать факторы, которые влияют на популярность видео на YouTube, анализировать тенденции в контенте и изучать взаимосвязи между различными характеристиками видео.

Структура данных:

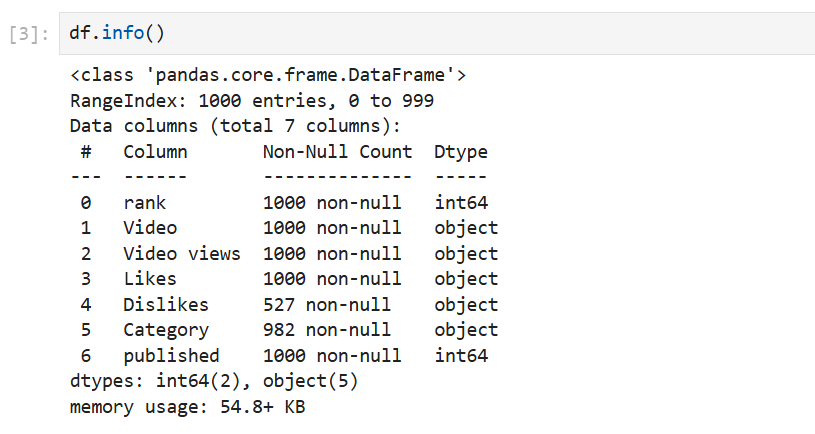
* **rank: Ранг видео в списке (от 1 до 1000).**
* **Video: Название видео.**
* **Video views: Количество просмотров видео.**
* **Likes: Количество лайков видео.**
* **Dislikes: Количество дизлайков видео.**
* **Category: Категория, к которой относится видео (например, Music, Entertainment, Sports).**
* **published: Год публикации видео.**

# Изучение данных

Подключаем необходимые библиотеки.



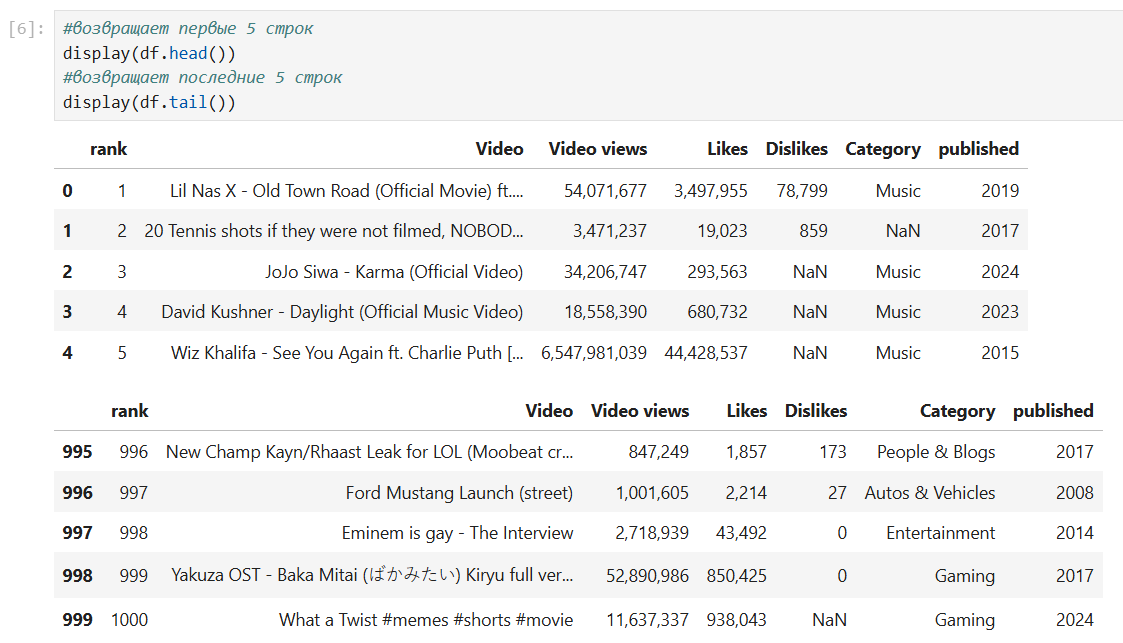
Выводим информацию.



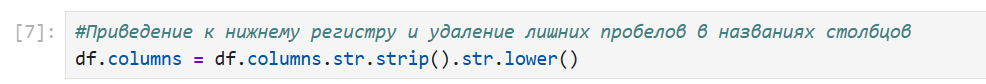
Выводим названия столбцов датасета.



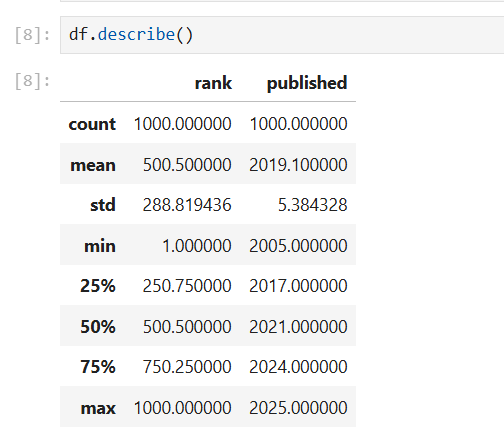
Выводим первые и последние 5 строк датасета.



Устраним и приводим к нижнему регистру.



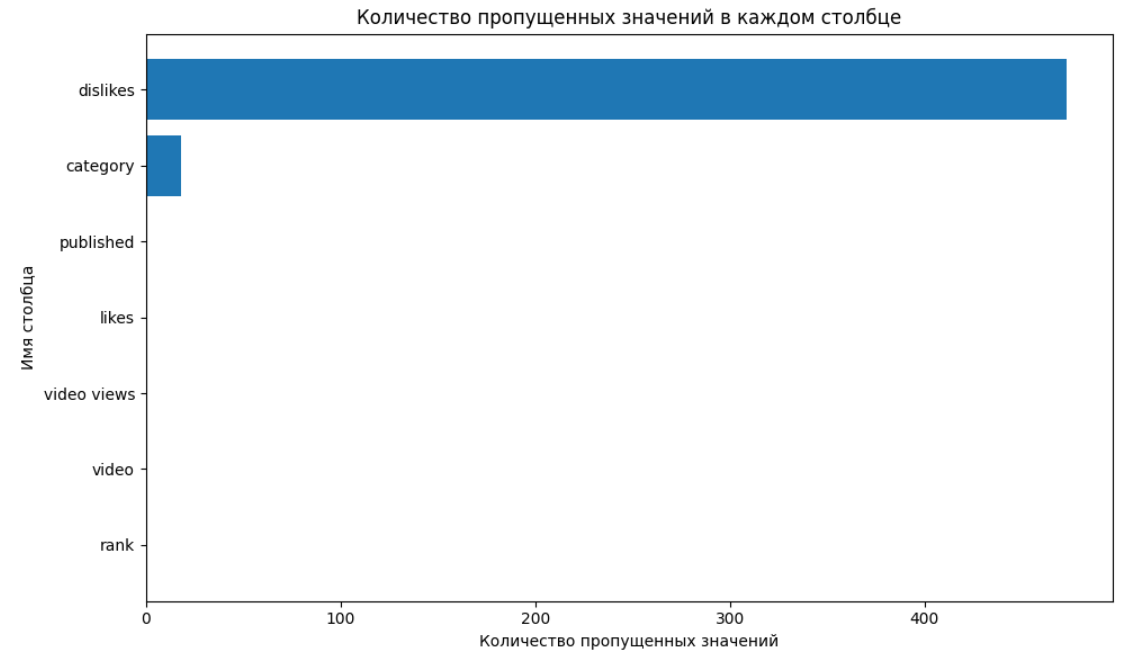
# Описательная статистика

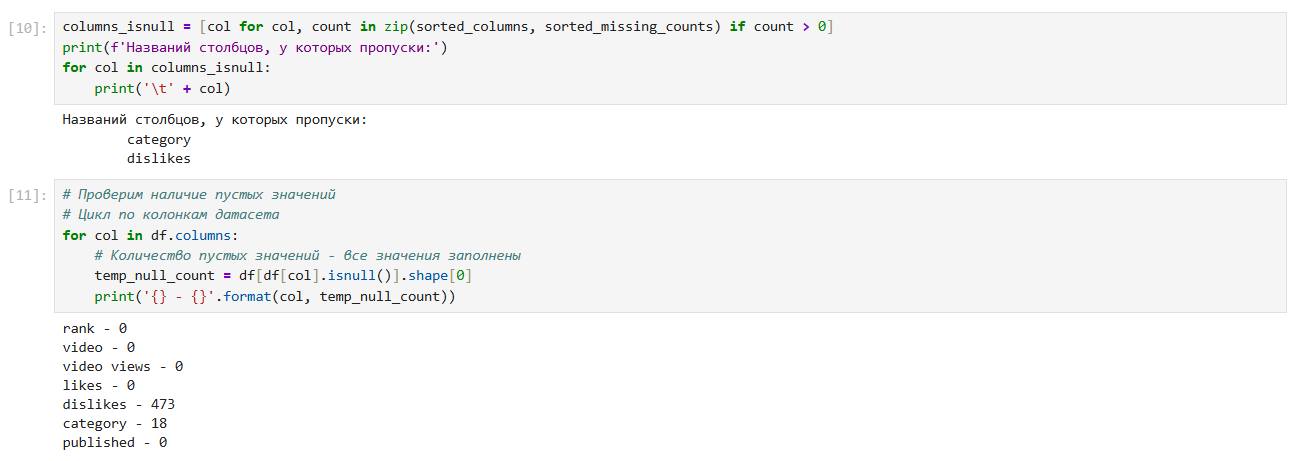


# Предобработка данных

**6.1. Пропуск значения**





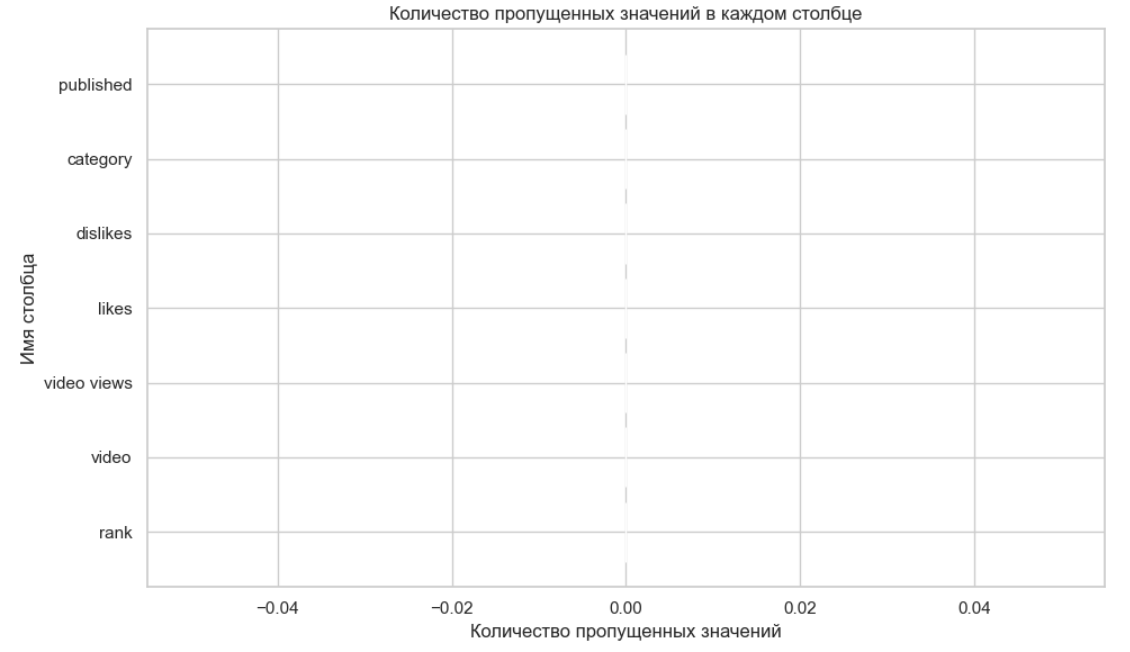


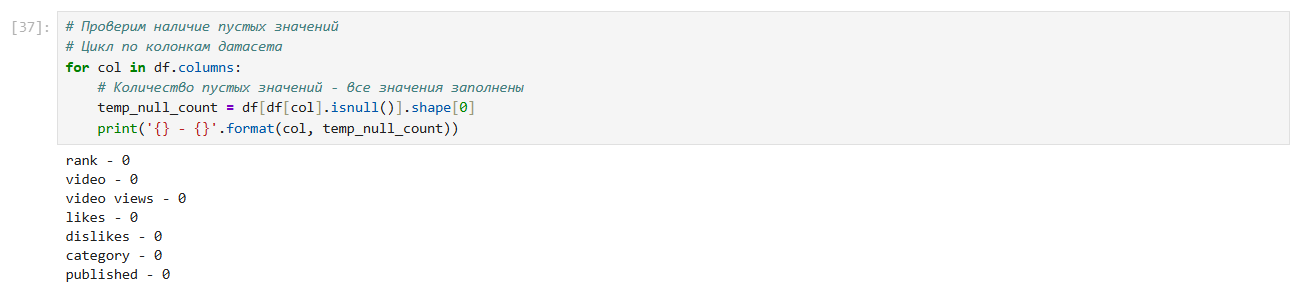
Для заполнения пропусков dislikes и category выполним команды:



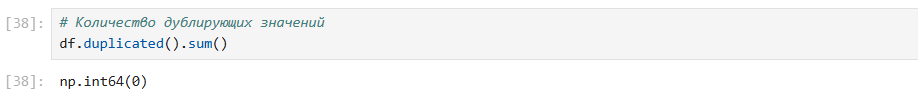
Еще раз проверим наличие пропусков.





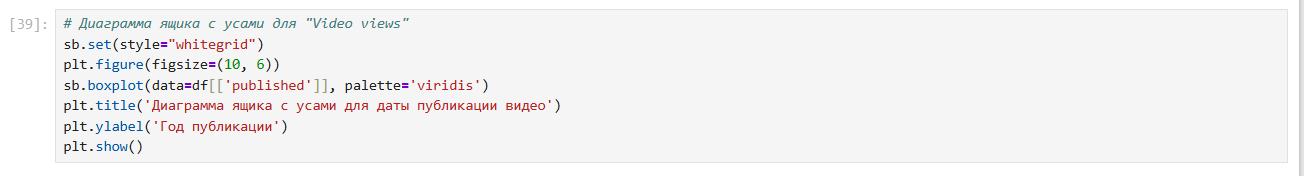


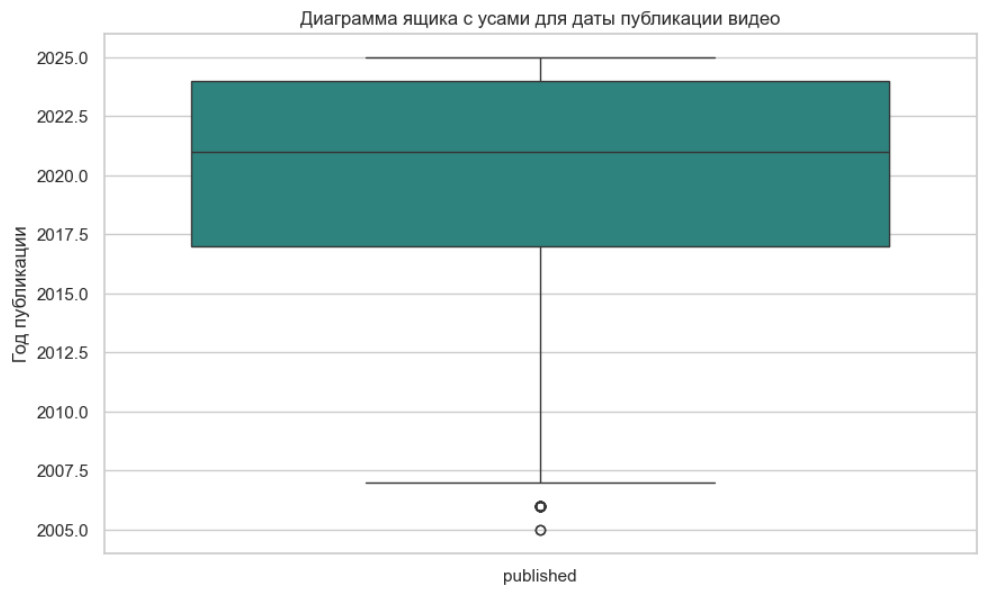
**6.2. Дубликаты**



Дубликатов нет.

**6.3. Выбросы (Ящик с усами)**

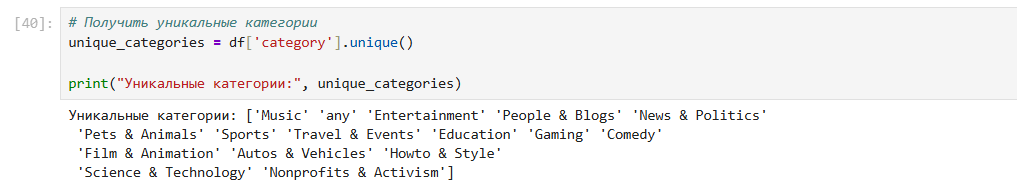




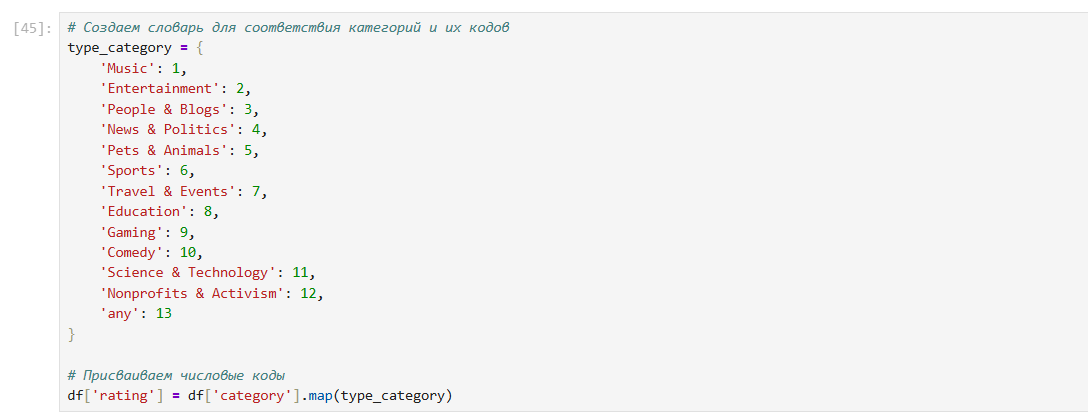
Часто видео публиковали с 2017 по 2024.

**6.4. Преобразование категорий в числа**

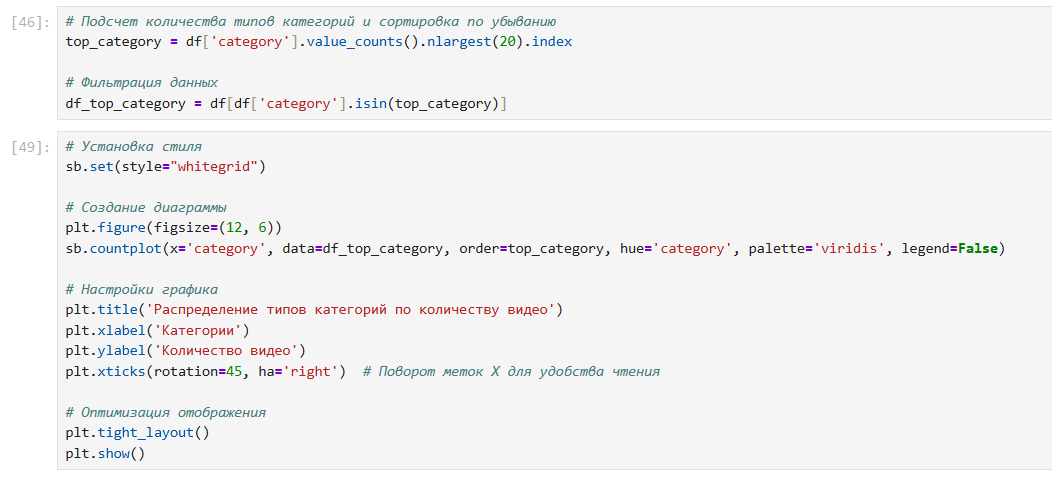
Получаем какие есть категории.

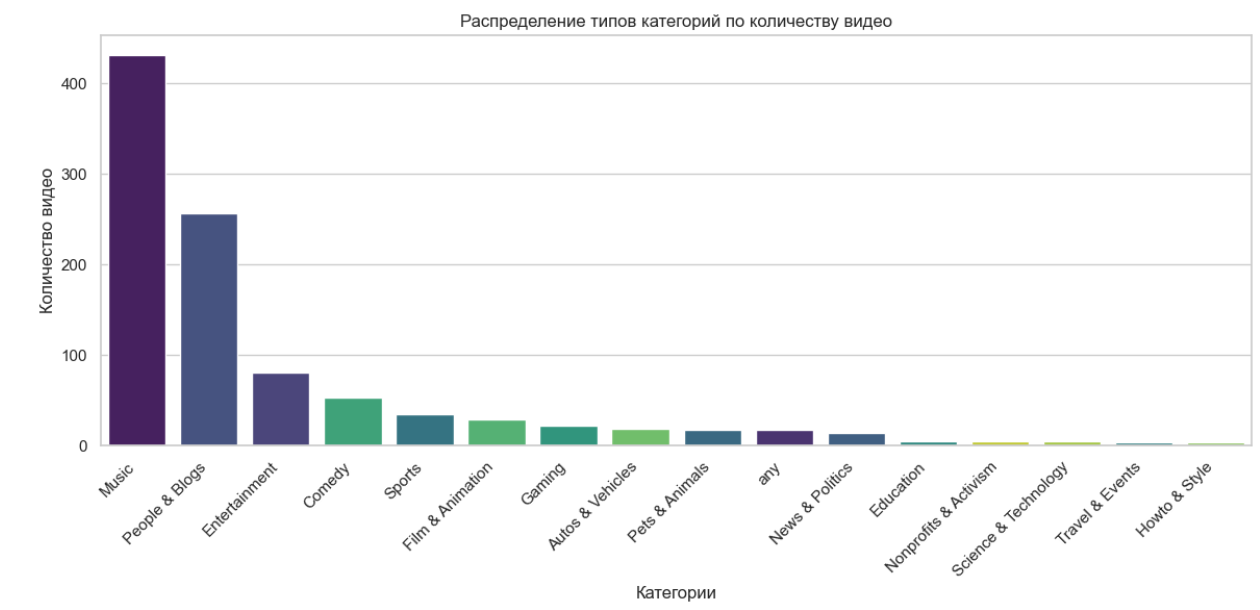


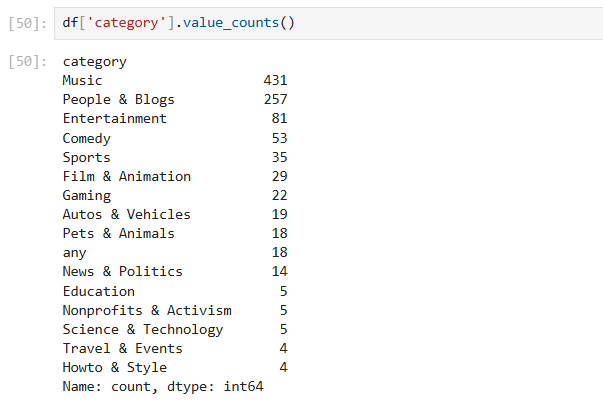
Преобразуем категории в числа.



Проведем исследование на распределение категорий по количеству видео.







Результат: Категория «Music» имеет самое большое количество - более 400 видео. Эта категория является самой популярной категорией в YouTube среди 1000 видео.

# Итог

* Пропусков нет
* Дубликатов нет
* Количество выбросов отсутствует
* В ходе предобработки данных было выявлено, что есть колонка, которая дает информацию типов категорий, но, однако информация носит HTML формата, что на парсинг уходит много времени и необходимо сопровождать кода. Для тщательного исследования будет полезно.

# Вывод

В ходе выполнения лабораторной работы изучили способы предварительной обработки данных для дальнейшего формирования моделей.