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 №1

«Разведочный анализ 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

Изучение различных методов визуализация данных.

2. Описание задание

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь https://github.com/ugapanyuk/courses_current/wiki/DSLIST.
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть здесь - https://github.com/ugapanyuk/courses_current/blob/main/notebooks/ds/sklearn_datasets.ipynb.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на github.

Средства и способы визуализации данных можно посмотреть здесь - https://github.com/ugapanyuk/courses_current/wiki/VISUAL.

В качестве опорного примера для выполнения лабораторной работы можно использовать пример - https://github.com/ugapanyuk/courses_current/blob/main/notebooks/eda/eda_visualization.ipynb.

3. 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:

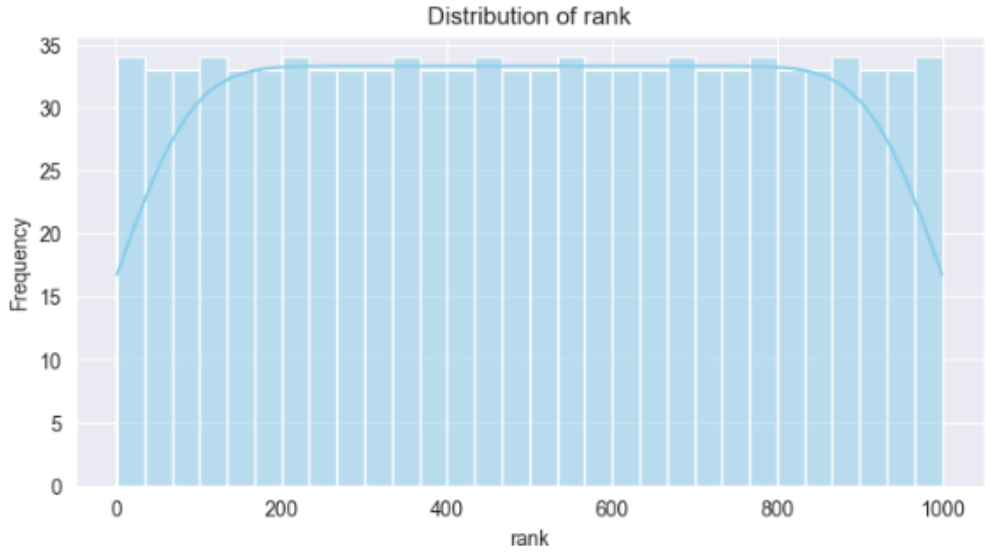
- **rank: Ранг видео в списке (от 1 до 1000).**
- **Video: Название видео.**
- **Video views: Количество просмотров видео.**
- **Likes: Количество лайков видео.**
- **Dislikes: Количество дизлайков видео.**
- **Category: Категория, к которой относится видео (например, Music, Entertainment, Sports).**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

4. Визуальное исследование датасета

4.1.Гистограммы для числовых колонок

Basic statistics:

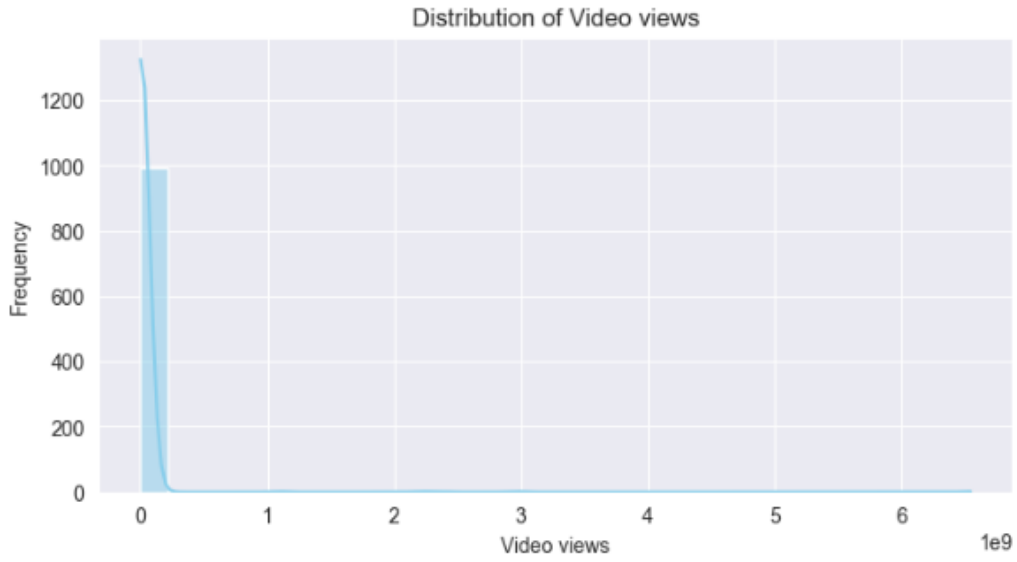
	rank	Video views	Likes	Dislikes	published
count	1000.000000	1.000000e+03	1.000000e+03	527.000000	1000.000000
mean	500.500000	2.453435e+07	3.685451e+05	2322.324478	2019.100000
std	288.819436	2.512570e+08	1.629418e+06	9653.170360	5.384328
min	1.000000	4.493900e+04	4.330000e+02	0.000000	2005.000000
25%	250.750000	9.815690e+05	9.427250e+03	200.000000	2017.000000
50%	500.500000	2.341652e+06	3.026200e+04	477.000000	2021.000000
75%	750.250000	1.162638e+07	1.649858e+05	1469.000000	2024.000000
max	1000.000000	6.547981e+09	4.442854e+07	178042.000000	2025.000000



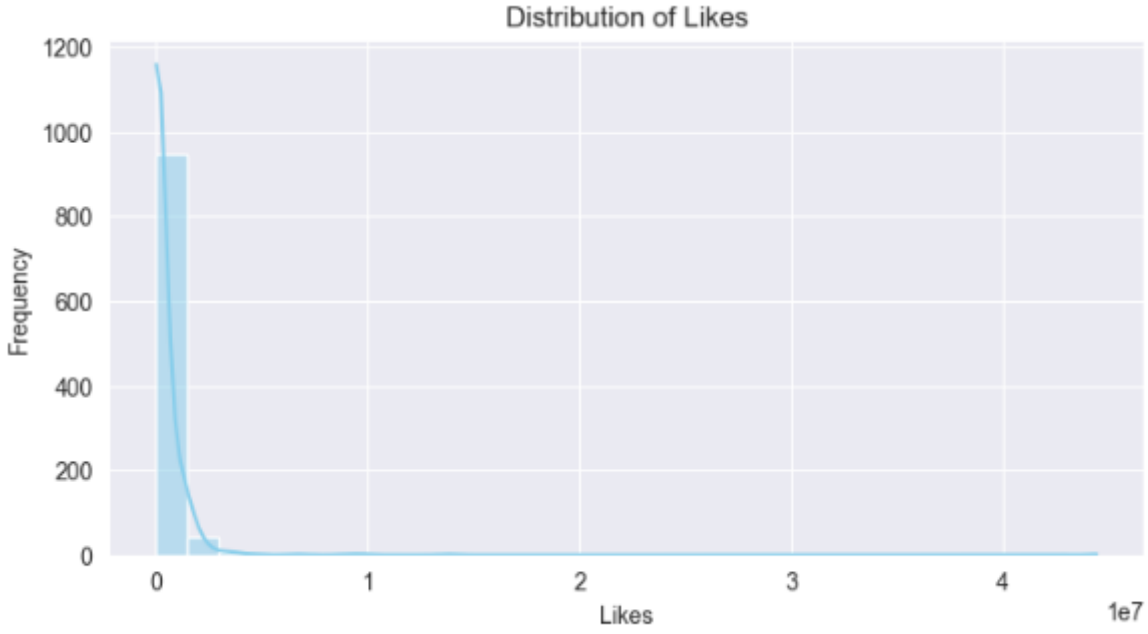
Ось X: Значения переменной (например, количество просмотров, год публикации).

Ось Y: Частота (количество) наблюдений, попадающих в каждый интервал (бин).

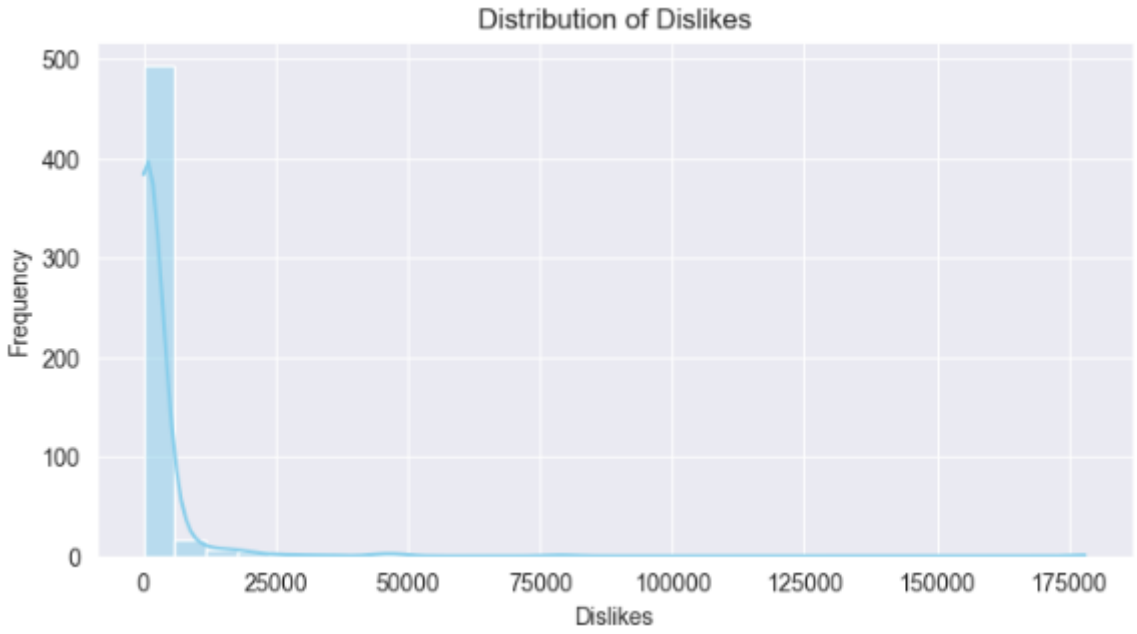
rank: Распределение рангов близкое к равномерному, так как есть 1000 видео, и каждое имеет свой уникальный ранг от 1 до 1000.



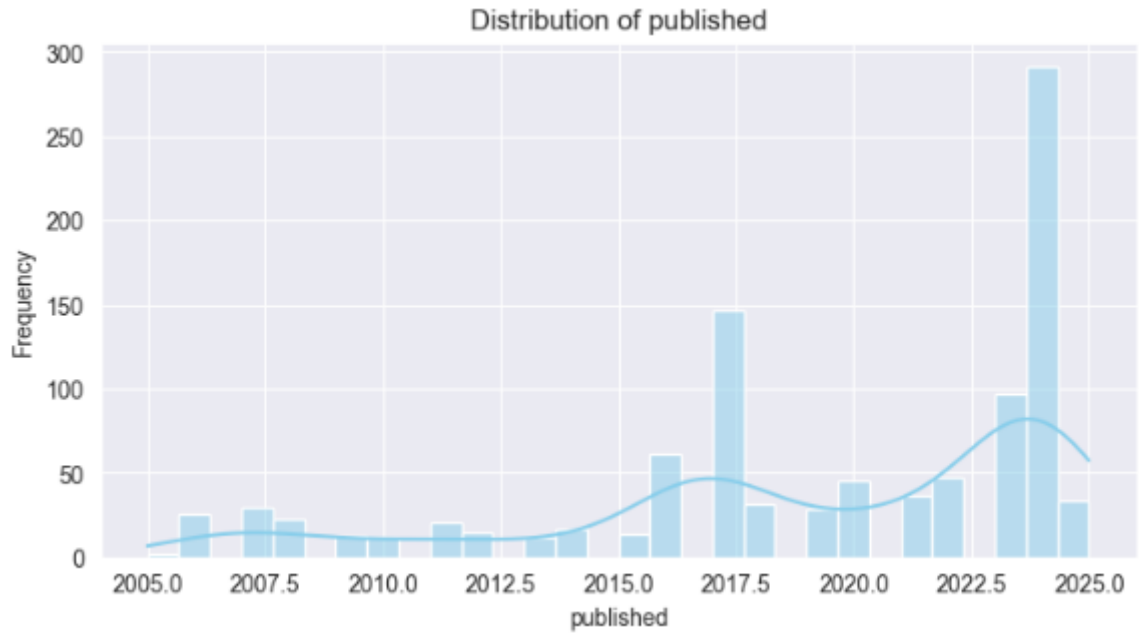
Video views (Количество просмотров): Имеет сильно скошенное вправо распределение, т.к. большинство видео имеют относительно небольшое количество просмотров, а лишь небольшое количество видео имеют очень большое количество просмотров. Это типично для YouTube.



Likes (Количество лайков): Как и с просмотрами, имеют скошенное вправо распределение. Лайки часто коррелируют с просмотрами, поэтому логично ожидать похожую форму.

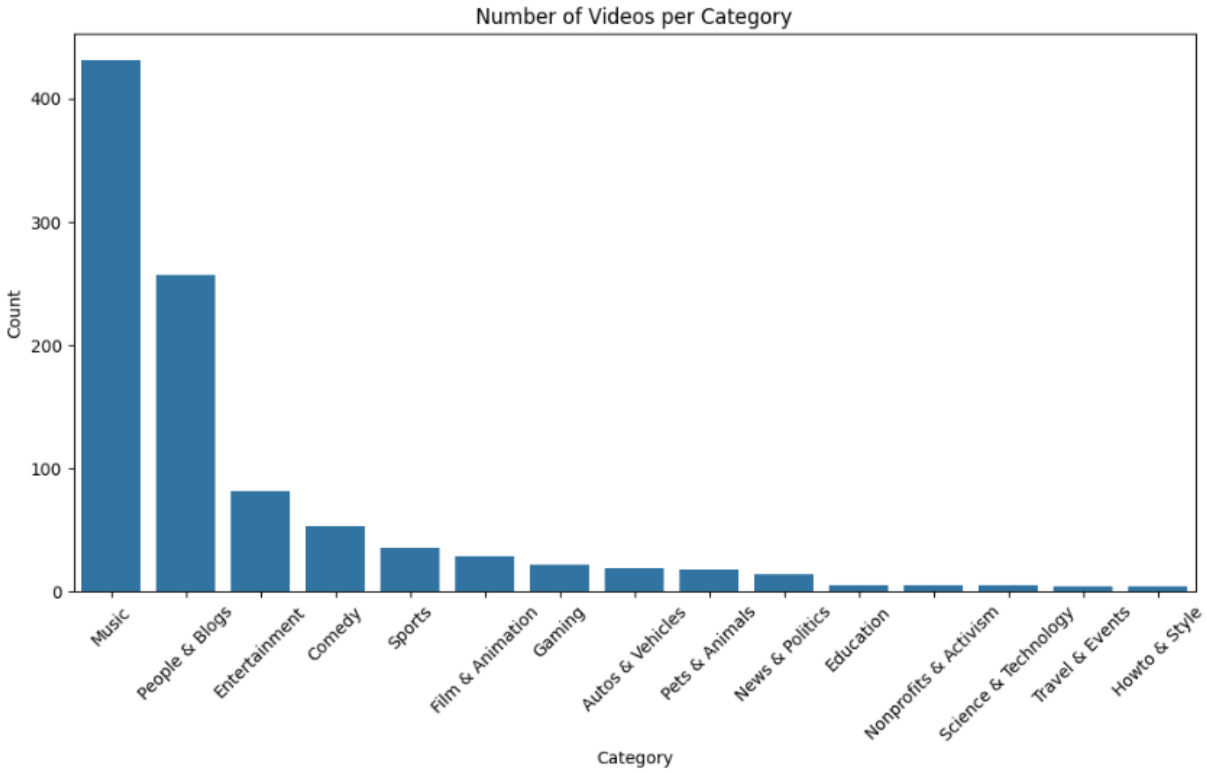


Dislikes (Количество дизлайков): Тоже, вероятно, скошенное вправо распределение, но, возможно, с большим количеством нулевых или отсутствующих значений (помните пропуски в данных). Важно помнить, что количество дизлайков может быть подвержено изменениям политики YouTube, видимости дизлайков и т.д.



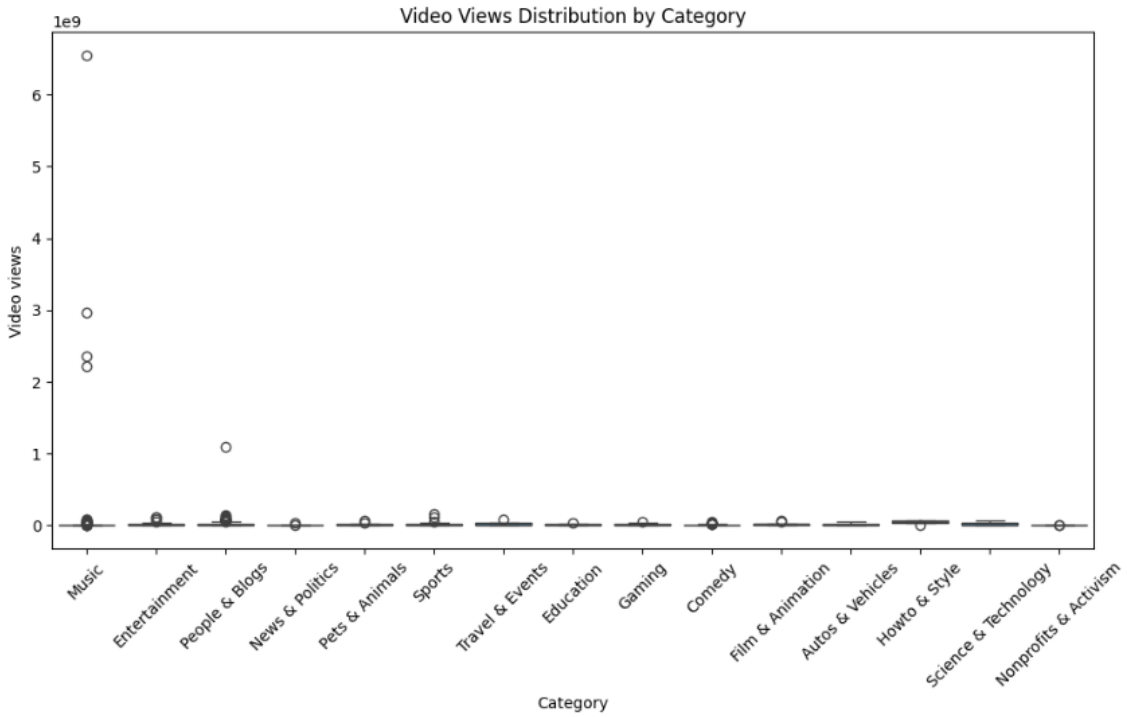
published (Год публикации): показывает тренд увеличения количества популярных видео с течением времени (если YouTube растет). Видны пики в определенные годы, что может указывать на изменения в алгоритмах YouTube или трендах в контенте.

4.2. Число видео в категориях



Данный график показывает, какие категории наиболее популярны, мы видим, что это - "музыка", "блоги", и т.д.

4.3. Распределение количества просмотров видео для каждой категории



Можем сравнить медианы количества просмотров между разными категориями. Все медианы равны и указывают на то, что в среднем видео во всех категориях имеют одинаковое кол-во просмотров (т.к. это топ-видео). Видим, что категория “Music” имеет самый высокий “ящик” и много выбросов сверху. Это может означать следующее:

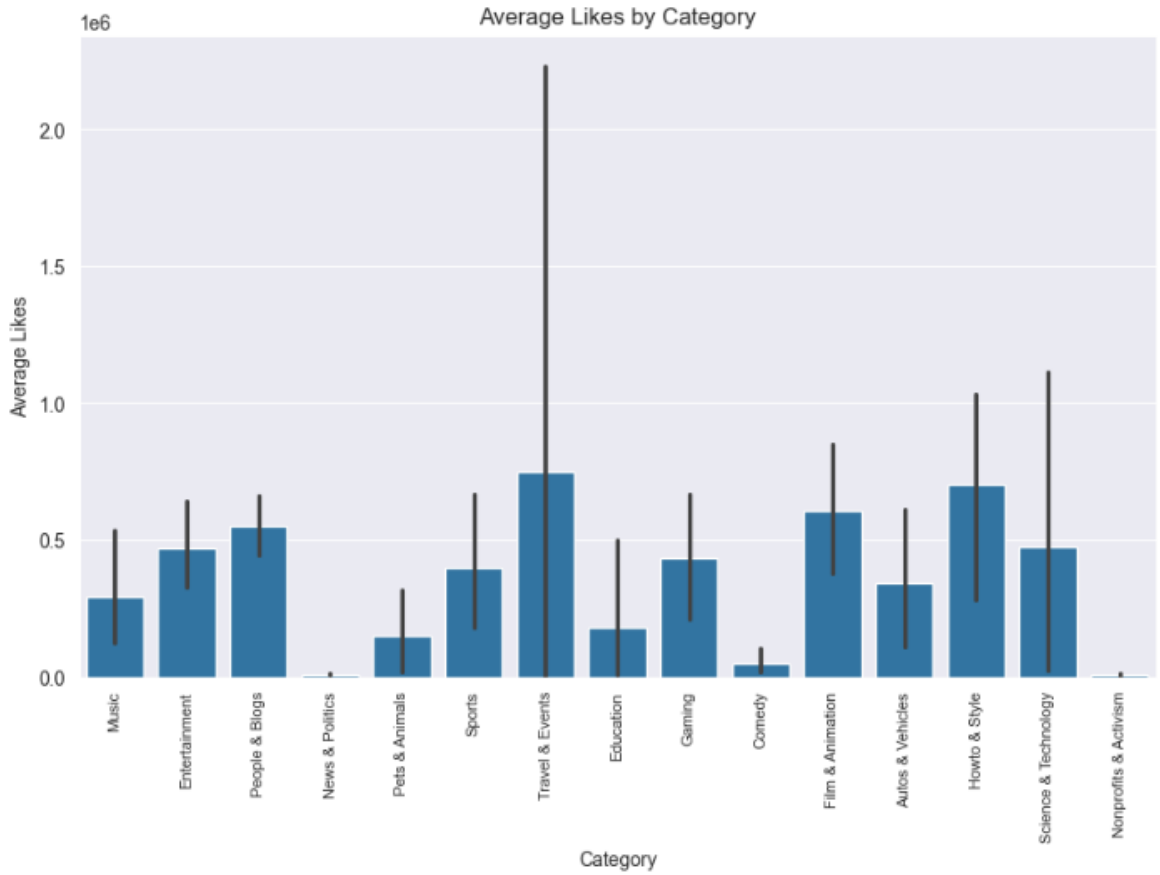
- В среднем музыкальные видео имеют больше просмотров, чем видео в других категориях.

- Разброс количества просмотров среди музыкальных видео очень велик (некоторые музыкальные видео становятся невероятно популярными, а другие - нет).

- Есть несколько музыкальных видео с аномально большим количеством просмотров (выбросы).

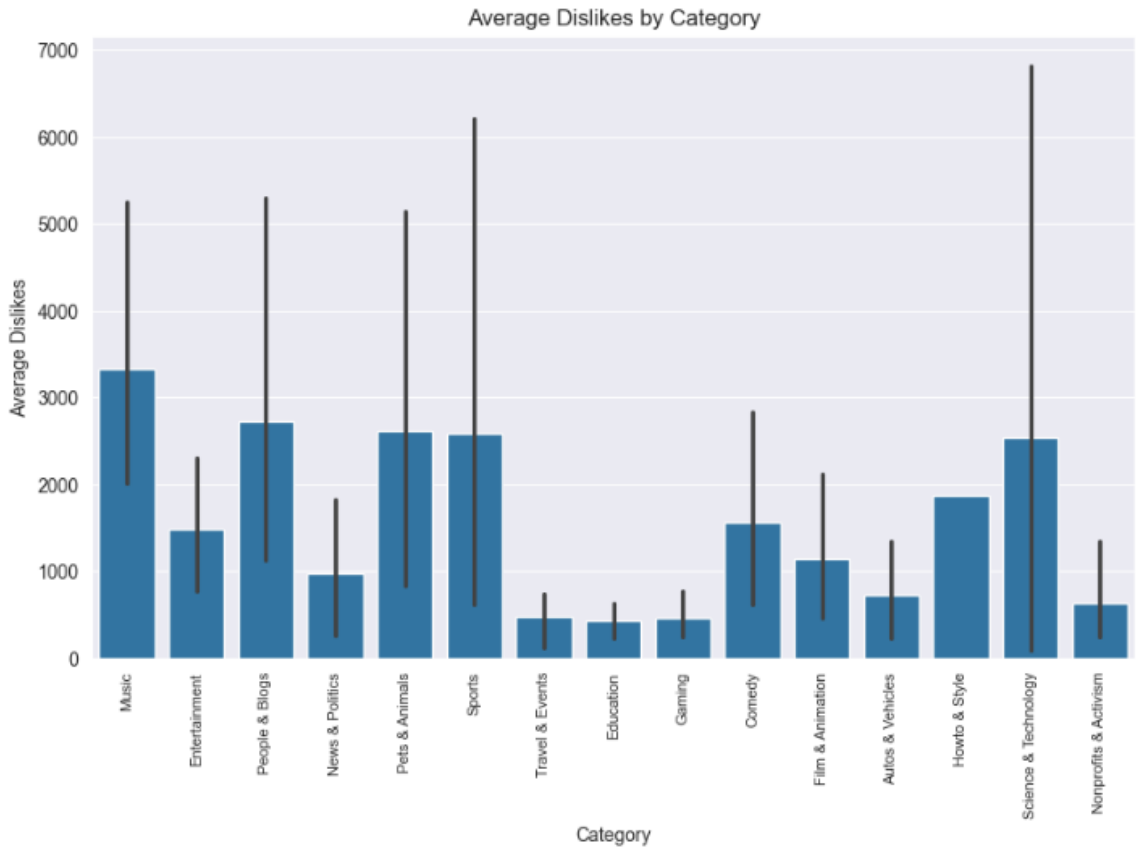
То же самое можем сказать про категорию "блоги".

4.4. Диаграмма распределения лайков для категорий



Столбчатая диаграмма показывает среднее количество лайков для видео в каждой категории. Видим, что лидируют "путешествия".

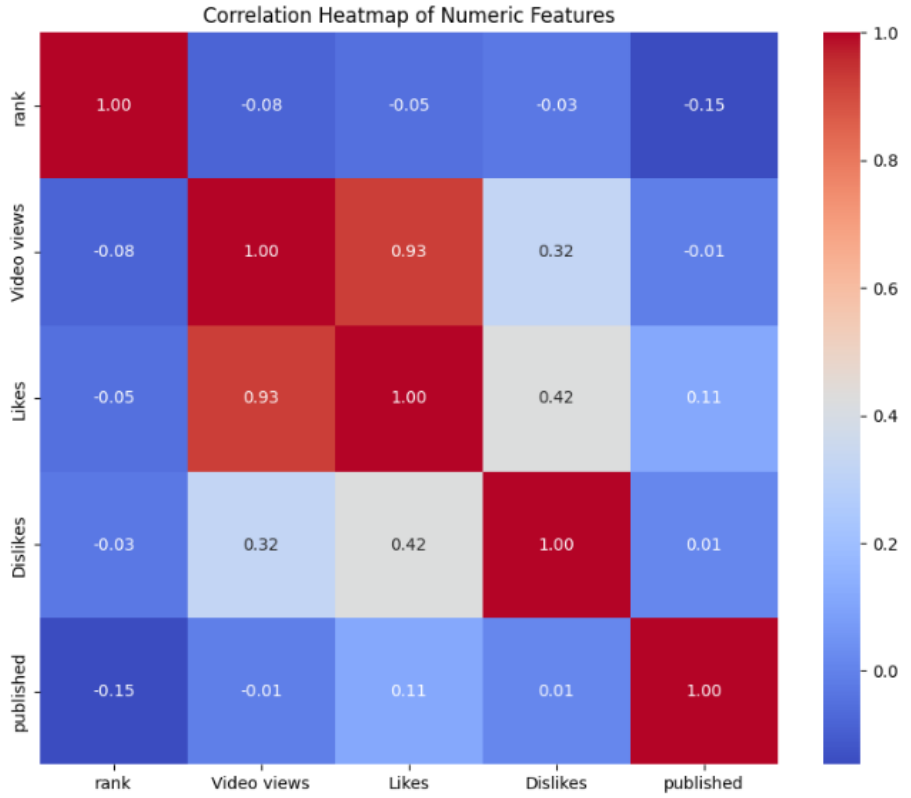
4.5. Диаграмма распределения дизлайков для категорий



Столбчатая диаграмма показывает среднее количество дизлайков для видео в каждой категории. Видим, что больше всего дизлайки ставят спортивным видео, музыкальным и "науке и технологиям".
Интересно, что спортивные видео вызывают больше всего эмоций как по лайкам, так по дизлайкам.

5. Информация о корреляции признаков

Для анализа взаимосвязей между числовыми признаками была построена корреляционная матрица:



- **Сильная положительная корреляция между просмотрами и лайками (0.93):** Это ожидаемый результат, который подтверждает гипотезу о том, что чем больше просмотров набирает видео, тем больше вероятность, что пользователи поставят лайк. Эта взаимосвязь является одной из основных метрик вовлеченности аудитории на YouTube. Такой высокий коэффициент корреляции говорит о том, что эти две переменные движутся практически синхронно.
- **Умеренная положительная корреляция между лайками и дизлайками (0.42):** Этот результат говорит о том, что видео, которые вызывают больше положительных реакций (лайки), также имеют тенденцию вызывать и больше отрицательных реакций (дизлайки). Это может указывать на то, что контент, вызывающий сильные эмоции (не обязательно только положительные), привлекает больше внимания.

Также это может говорить о том, что вокруг видео разворачиваются дискуссии и споры.

- **Слабая положительная корреляция между просмотрами и дизлайками (0.32):** Как и в случае с лайками, большее количество просмотров может приводить к большему количеству дизлайков, хотя эта связь значительно слабее, чем связь между просмотрами и лайками. Вероятно, просмотры генерируются не только довольными зрителями, но и теми, кто просто любопытствует или критикует контент.
- **Слабая отрицательная корреляция между годом публикации (published) и рангом (rank) (-0.15):** Эта небольшая отрицательная корреляция предполагает слабую тенденцию к тому, что более новые видео (более поздний год публикации) имеют более низкий ранг (то есть занимают более высокие позиции в топе). Однако, связь довольно слабая, что указывает на то, что год публикации не является определяющим фактором для ранга видео. Возможно, более свежие видео получают больше продвижения от YouTube, но это не является гарантией высокого рейтинга.
- **Отсутствие значимой корреляции с рангом (rank):** В целом, ранг видео слабо коррелирует с другими переменными. Это говорит о том, что на ранг видео влияет множество факторов, не все из которых представлены в этом наборе данных (например, алгоритмы YouTube, продвижение видео, актуальность темы, и т.д.).

6. Итог

6.1. Анализ данных

- Сильная положительная корреляция между просмотрами и лайками (0.93): Это ожидаемый результат, который подтверждает гипотезу о том, что чем больше просмотров набирает видео, тем больше вероятность, что пользователи поставят лайк. Эта взаимосвязь является одной из основных метрик вовлеченности аудитории на YouTube. Такой высокий коэффициент корреляции говорит о том, что эти две переменные движутся практически синхронно.
- Умеренная положительная корреляция между лайками и дизлайками (0.42): Этот результат говорит о том, что видео, которые вызывают больше положительных реакций (лайки), также имеют тенденцию вызывать и больше отрицательных реакций (дизлайки). Это может указывать на то, что контент, вызывающий сильные эмоции (не обязательно только положительные), привлекает больше внимания. Также это может говорить о том, что вокруг видео разворачиваются дискуссии и споры.
- Слабая положительная корреляция между просмотрами и дизлайками (0.32): Как и в случае с лайками, большее количество просмотров может приводить к большему количеству дизлайков, хотя эта связь значительно слабее, чем связь между просмотрами и лайками. Вероятно, просмотры генерируются не только довольными зрителями, но и теми, кто просто любопытствует или критикует контент.
- Слабая отрицательная корреляция между годом публикации (published) и рангом (rank) (-0.15): Эта небольшая отрицательная корреляция предполагает слабую тенденцию к тому, что более новые видео (более поздний год публикации) имеют более низкий ранг (то есть занимают более высокие позиции в топе). Однако, связь довольно слабая, что указывает на то, что год публикации не является определяющим

фактором для ранга видео. Возможно, более свежие видео получают больше продвижения от YouTube, но это не является гарантией высокого рейтинга.

- Отсутствие значимой корреляции с рангом (rank): В целом, ранг видео слабо коррелирует с другими переменными. Это говорит о том, что на ранг видео влияет множество факторов, не все из которых представлены в этом наборе данных (например, алгоритмы YouTube, продвижение видео, актуальность темы, и т.д.).