YouTube Videos and Channels ANALYSIS

Analyze the statistical relation between videos

This project was analyzed with Python, with the help of Pandas, and the following libraries were imported into it:

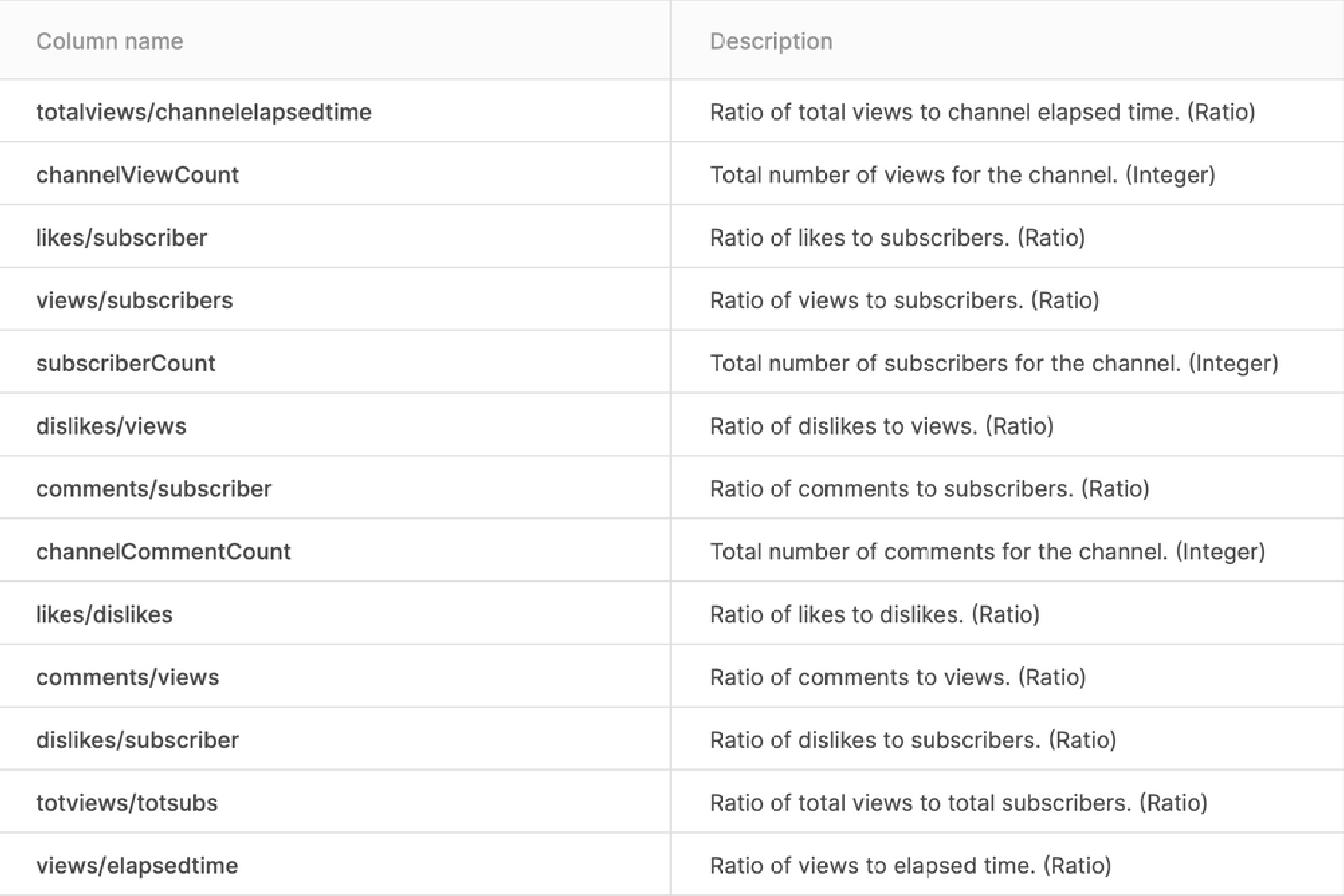
- ***Numpy***
- ***Matplotlib.Pyplot***
- ***Sklearn***

By: Yakir Attias

Data Overview

This data contains YouTube video and channel data.

Columns-



Column name	Description
totalviews/channelelapsedtime	Ratio of total views to channel elapsed time. (Ratio)
channelViewCount	Total number of views for the channel. (Integer)
likes/subscriber	Ratio of likes to subscribers. (Ratio)
views/subscribers	Ratio of views to subscribers. (Ratio)
subscriberCount	Total number of subscribers for the channel. (Integer)
dislikes/views	Ratio of dislikes to views. (Ratio)
comments/subscriber	Ratio of comments to subscribers. (Ratio)
channelCommentCount	Total number of comments for the channel. (Integer)
likes/dislikes	Ratio of likes to dislikes. (Ratio)
comments/views	Ratio of comments to views. (Ratio)
dislikes/subscriber	Ratio of dislikes to subscribers. (Ratio)
totviews/totsubs	Ratio of total views to total subscribers. (Ratio)
views/elapsedtime	Ratio of views to elapsed time. (Ratio)

Data Overview

- The data can provide information on youtube videos/youtube channels/youtube categories/youtube as a platform, and more.
- The data is divided into 27 columns and 575610 rows.
- Most of the data is numeric, we should change the date to DATE type, and we can add a "year" column.

Data Overview

- Only some of the columns are detailed; for example, the 'elapsed_time' column - doesn't say in which time unit it is measured.
- Some of the ratio columns are negative, which is impossible.
- channel_id: 449980 unique values
- video_id: 555627 unique values
- video_category: 18 unique value

Data Cleaning and preparation

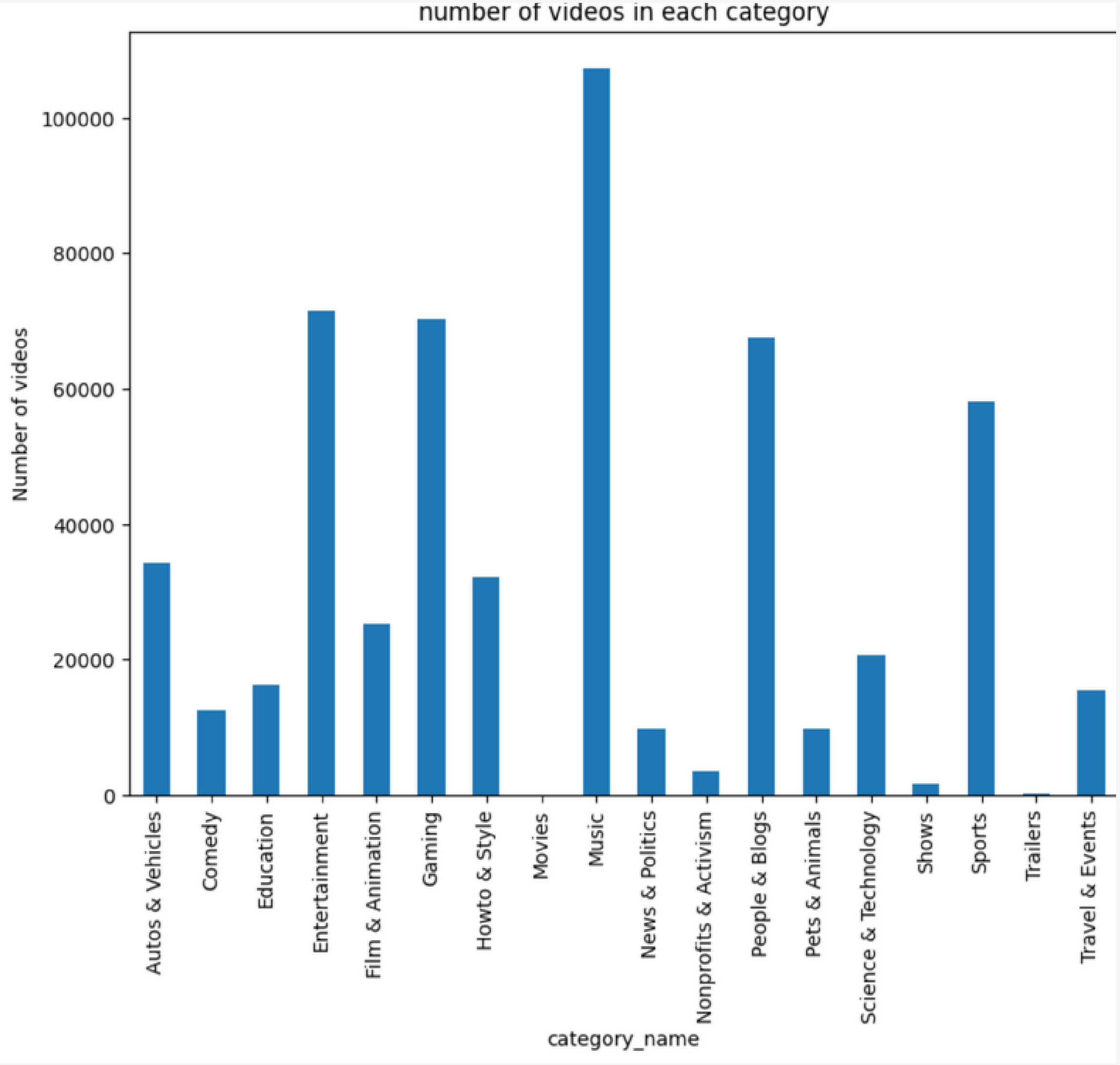
Actions that I've been to clean the data in order to analyze it:

- **We had 2 index columns, so I've deleted one:**
- **I've found 19186 duplicated rows and deleted them:**
- **I've set the date column as a Date type.**
- **Added a "Years" row. It can be very useful.**
- **Categories were presented in the dataset as numbers. We added a column for the category names.**

data visualizations

Publish a video in the Leading Category promises a large number of interactions?

We will check whether the leading category (in terms of the number of videos) is the category that brings the creator the highest amount of interactions.



Explain-

As we can see here, **the leading category is "Music"** (in terms of the number of videos), with more than 100,000 videos in our time frame.

Continue in the next slide.

data visualizations

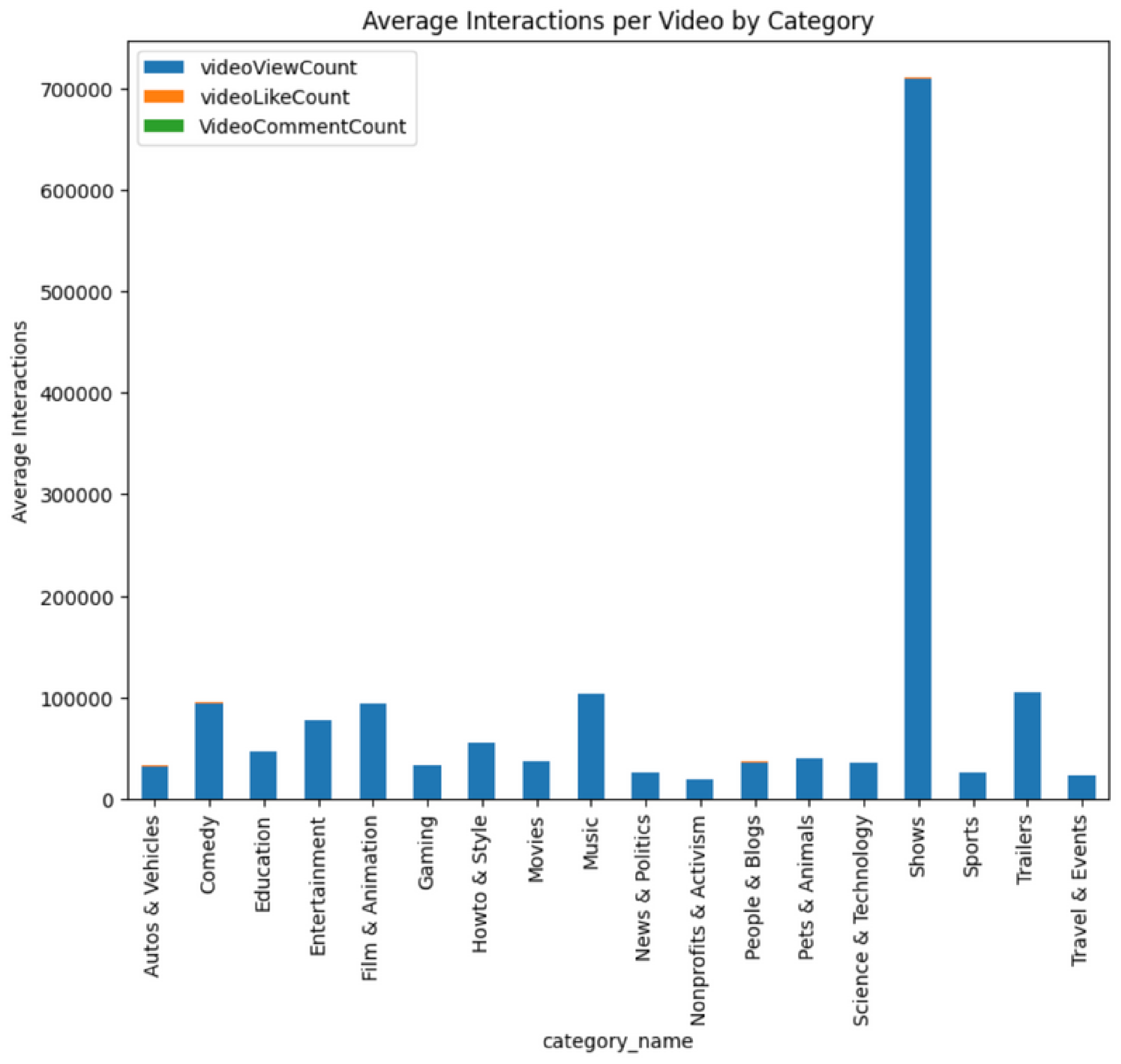
Publish a video in the Leading Category promises a large number of interactions?

Explain-

As we showed previously, the category that is published the most is number 10 (music), but it is far from being the category that collects the most interactions, which is the **"performance" category.**

We can think of a few reasons for it:

1. The "music" category may have a larger number of active channels that regularly upload new content.
2. The "shows" category may include a smaller number of highly popular shows that drive a disproportionate amount of interactions. This could result in a higher average number of interactions per video in the "shows" category, even if there are fewer videos overall.



data visualizations

The Number of Views by The Number of Subscribers

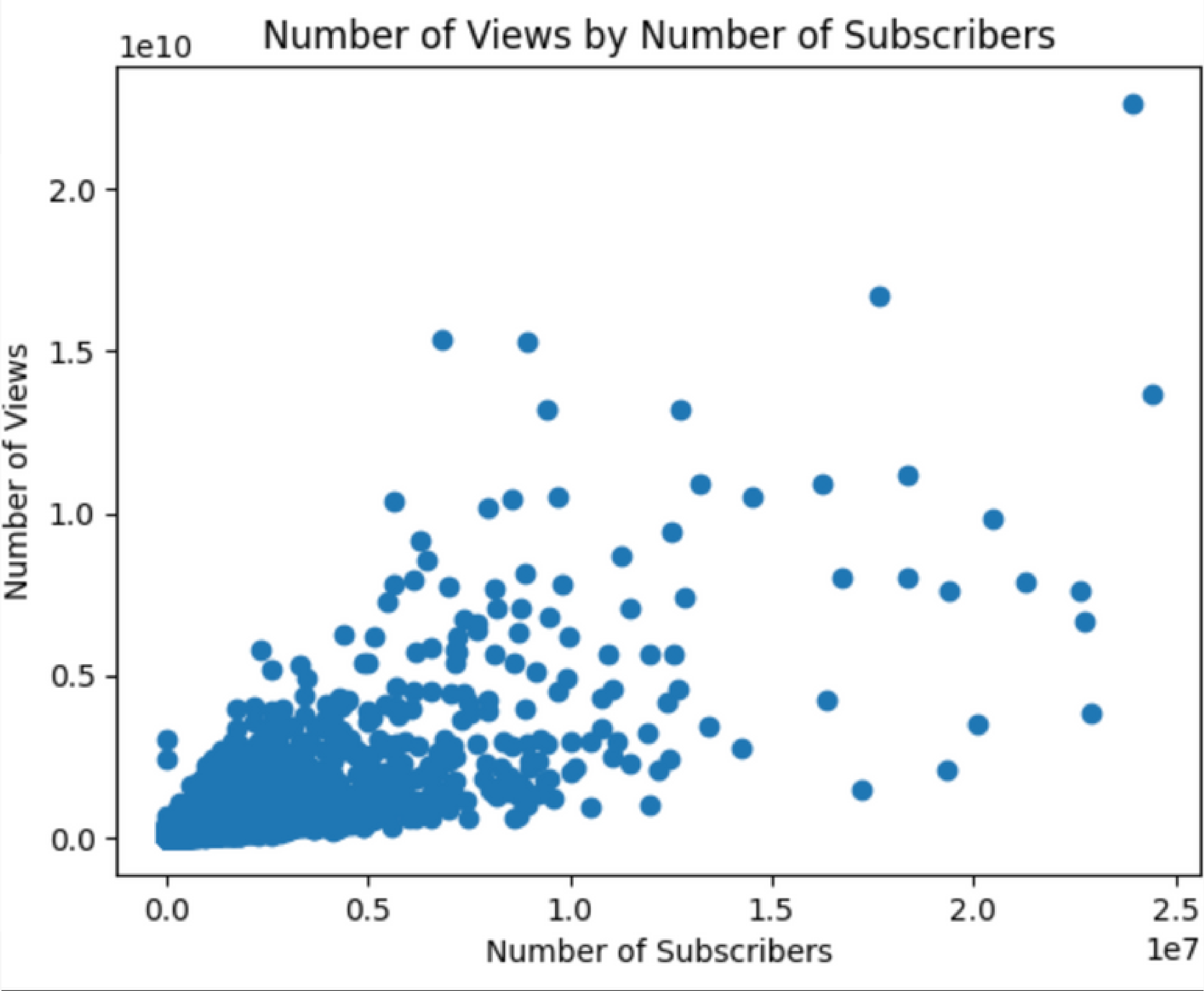
We will check if there is a connection between the popularity of the videos on the channel (total number of views on the channel) and the number of subscribers to the channel.

Explain-

I chose a scatter plot in order to show how much one variable is affected by another

According to the resulting graph, it is difficult to determine exactly how much the two variables influence each other, but it can be seen that there is indeed a strong relationship between them.

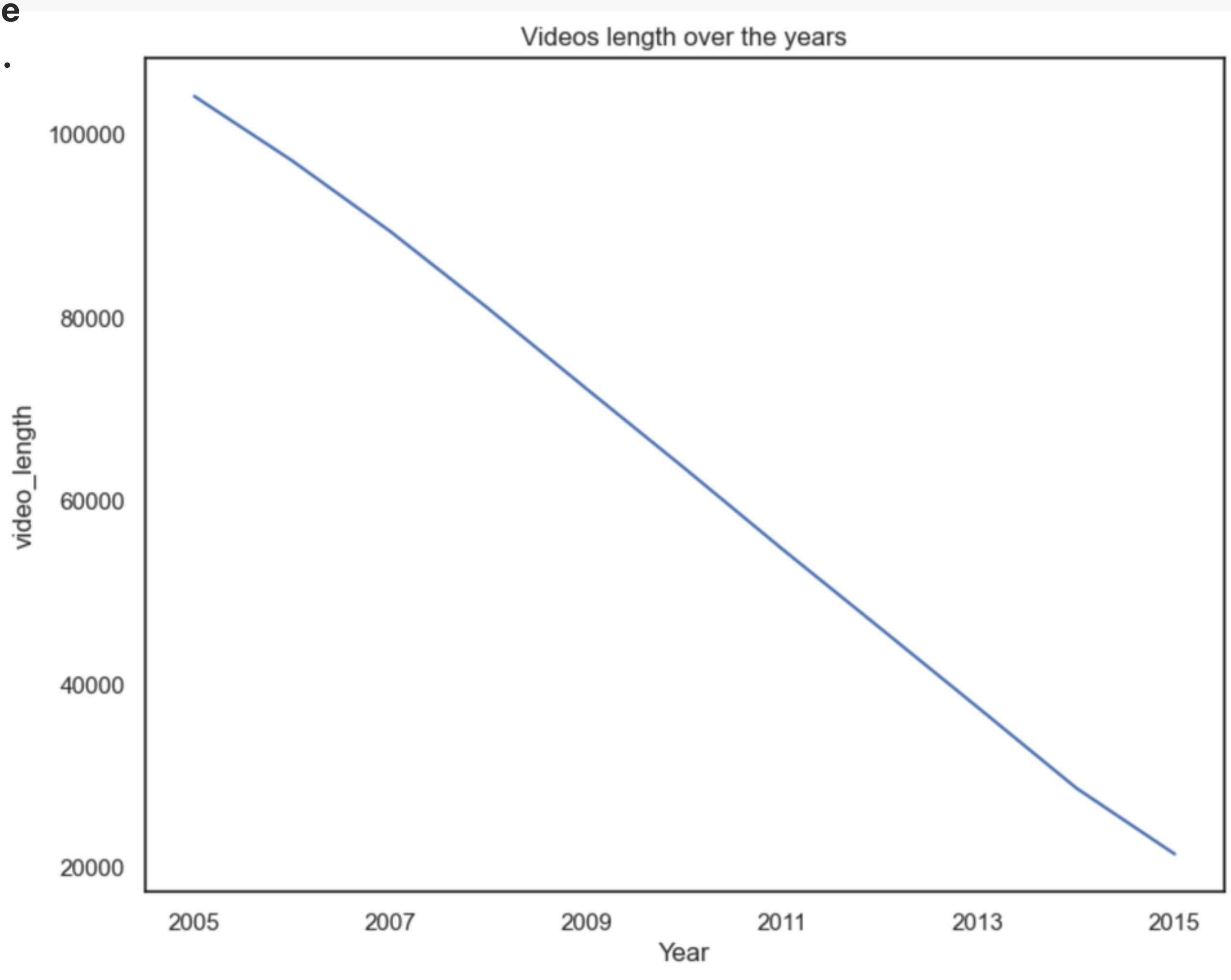
That's why we will try to check also with the help of correlation.
The correlation is: 0.7888077028423459
which is a strong correlation, so we can certainly see that there is a connection between these two variables.



data visualizations

Do videos get shorter as the years go by?

We will check whether the videos become shorter as the years go by (which makes sense since people prefer to receive information in as short a time as possible).



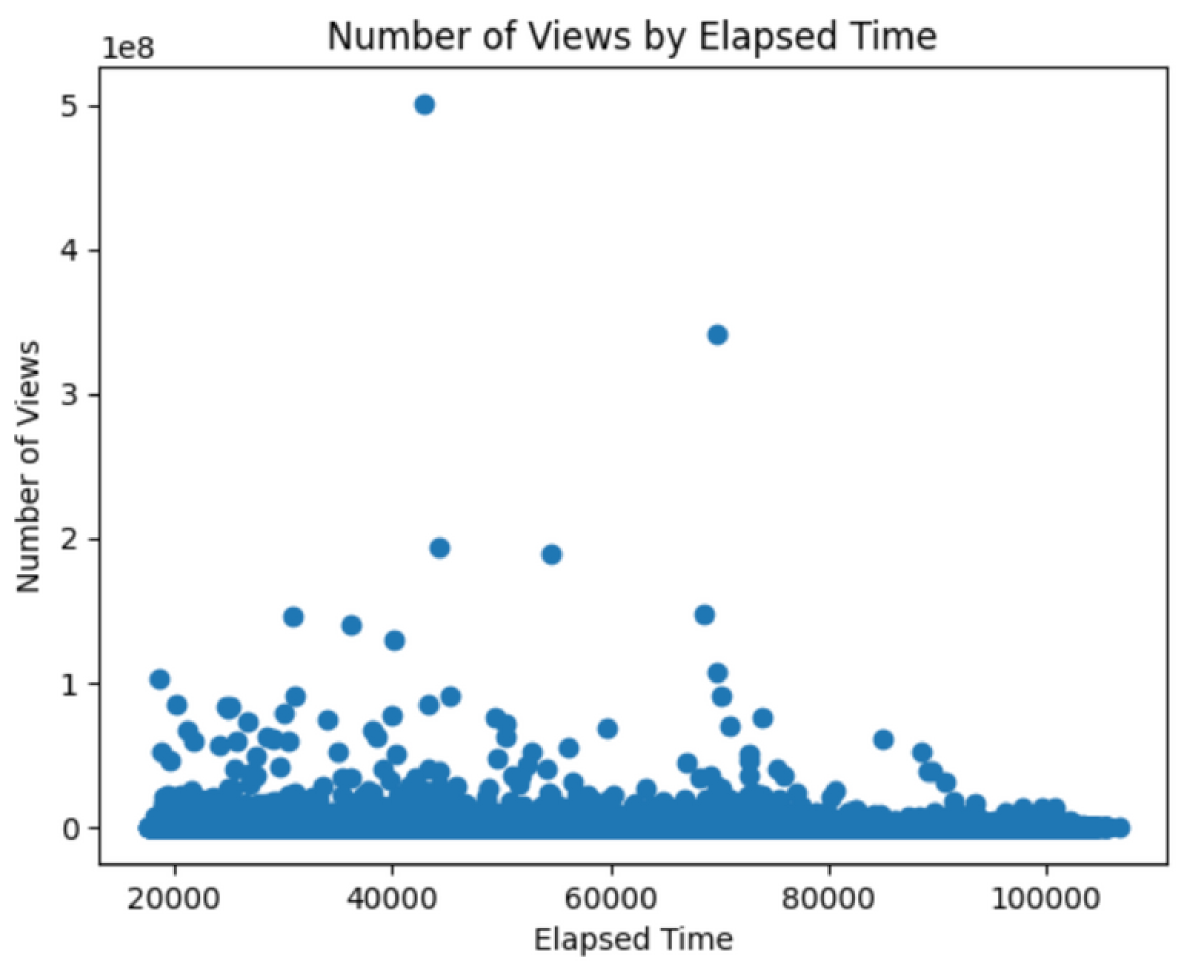
Explain-

We checked the average length of the videos in all the categories by year, we can see that there is indeed a constant decrease in the length of the videos every year, in a remarkably gradual way.

data visualizations

Short Videos = More Views?

Following on from the previous slide, we will now check whether, as the videos became shorter, so did the number of views on each video increase.



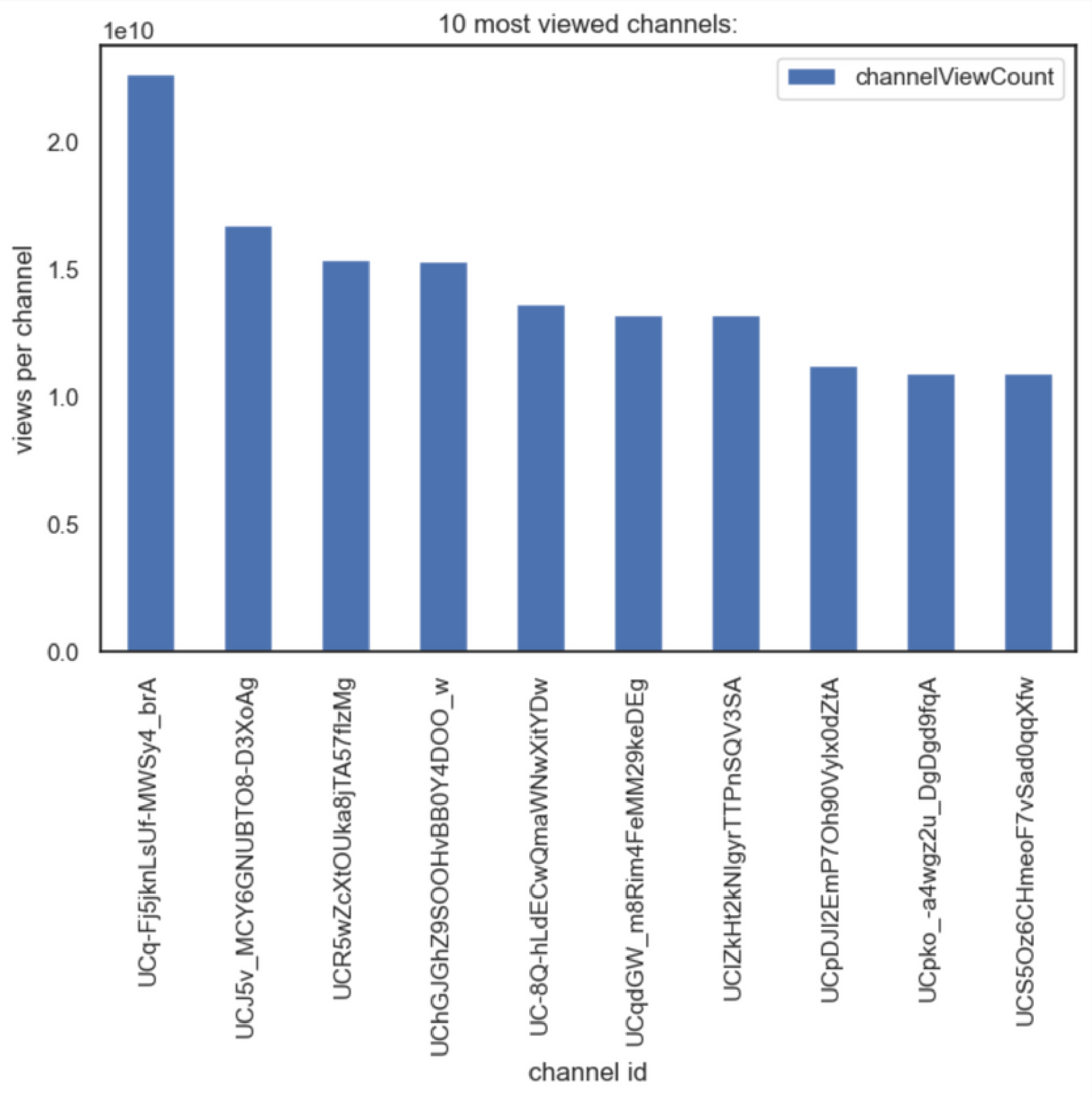
Explain-

We can see from the graph that the correlation between these 2 variables is not particularly strong. there is an increase in the number of views as the videos became shorter, but on a tiny scale.

data visualizations

10 most viewed channels:

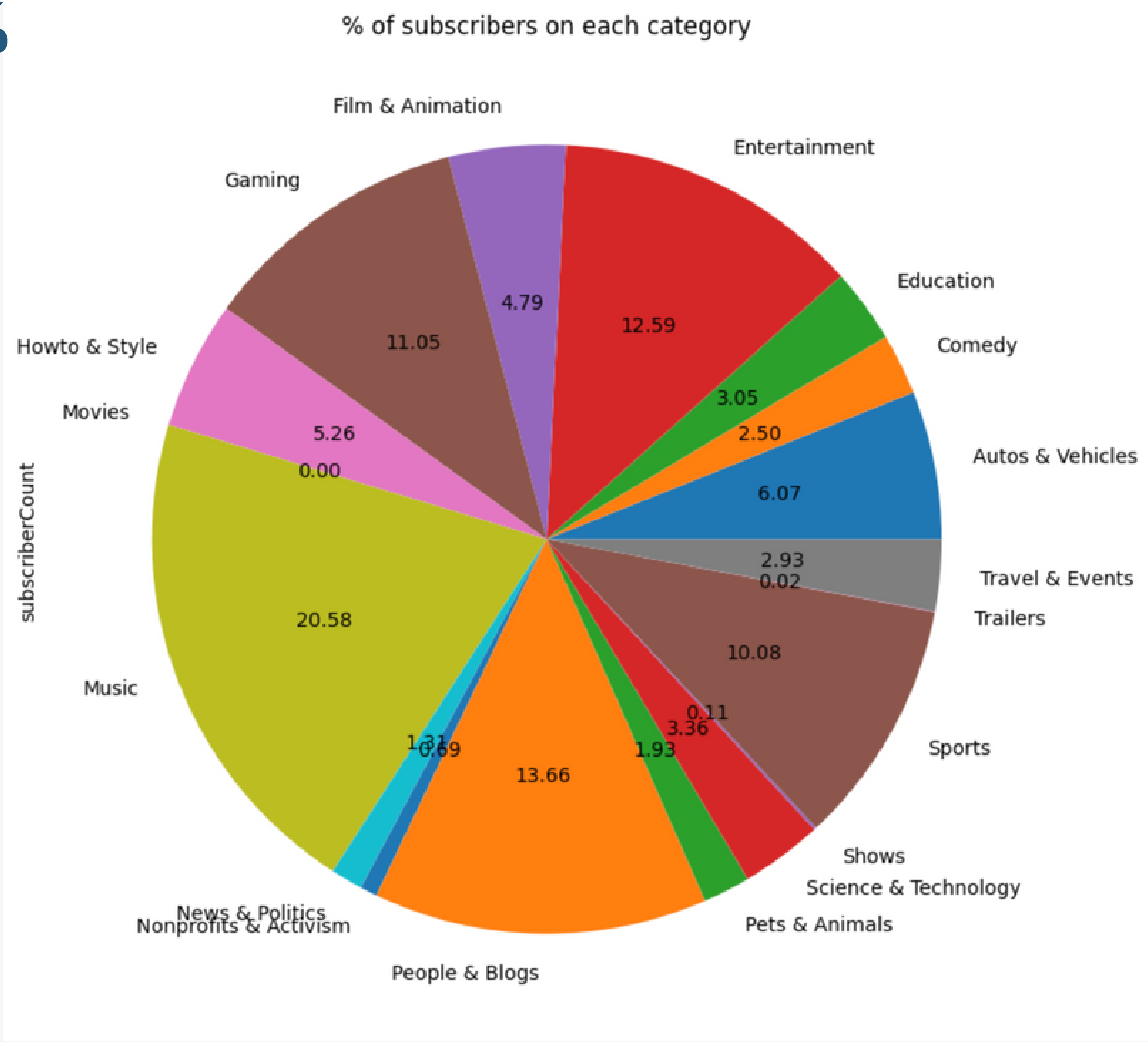
Let's see which channels are in the top 10 views amount:



data visualizations

Amount of subscribers in each category in %

Let's see the number of subscribers in each category in percentage:



Explain-

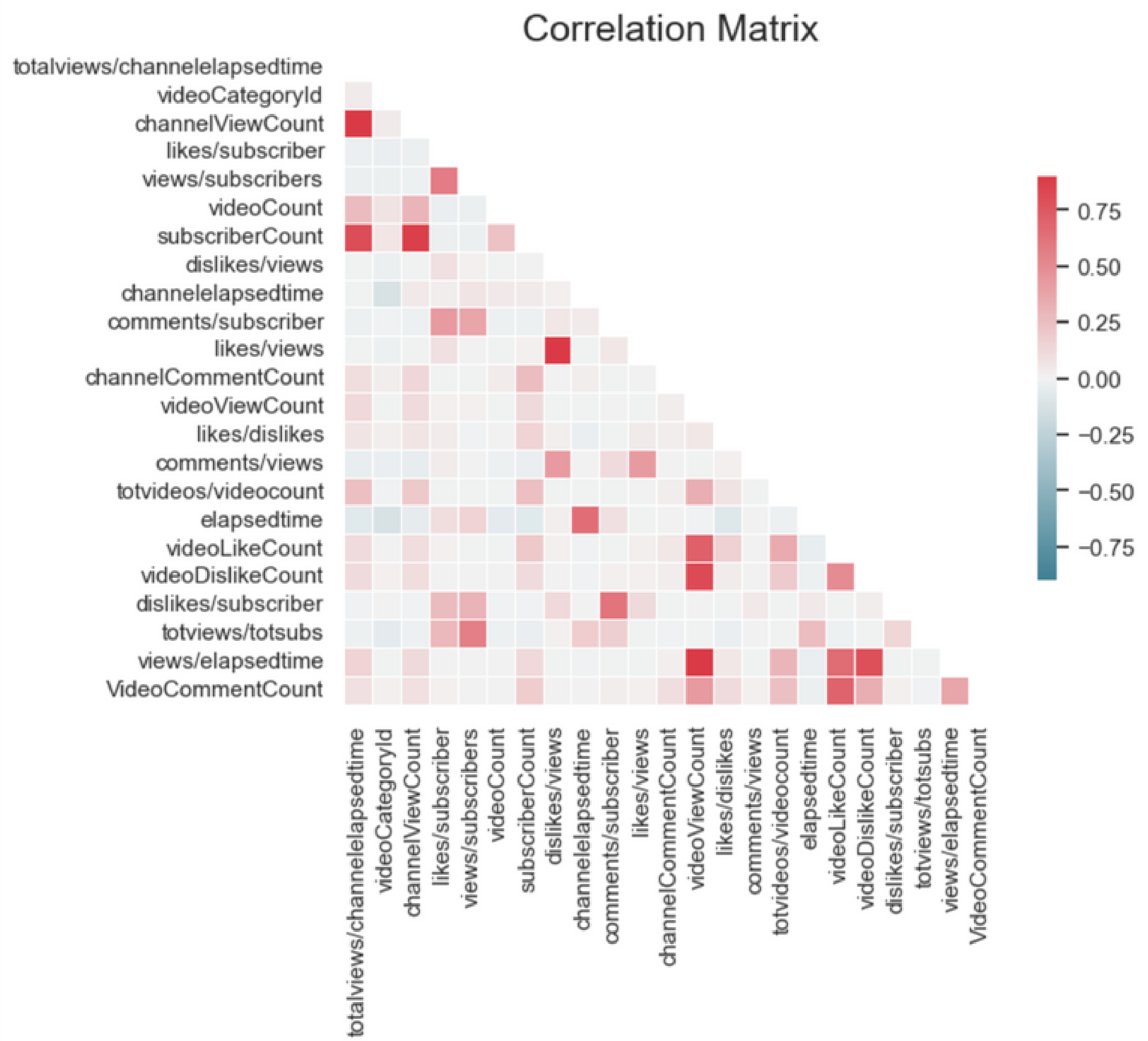
We can see that the leading category is music, with 20.58% of all category subscribers, immediately followed by the "people and blogs" category with 13.66 percent.

On the other hand, in last place is the movie category (I estimate due to the fact that there may not be channels that aim to publish movies because it is illegal).

data visualizations

Correlation Matrix

Let's see the Correlation Matrix and spot few analysis:



Explain-

Very strong positive correlation

- *channelViewCount - totalviews/channelElapsedTime*
- *views/elapsed_time - videoViewCount*

Strong positive linear correlation

- *video_dislike_count - video_view_count*
- *video_comment_count - video_like_count*
- *views/elapsed_time - video_dislike_count*

(And few more)

data visualizations

Number of Followers by Number of Videos:

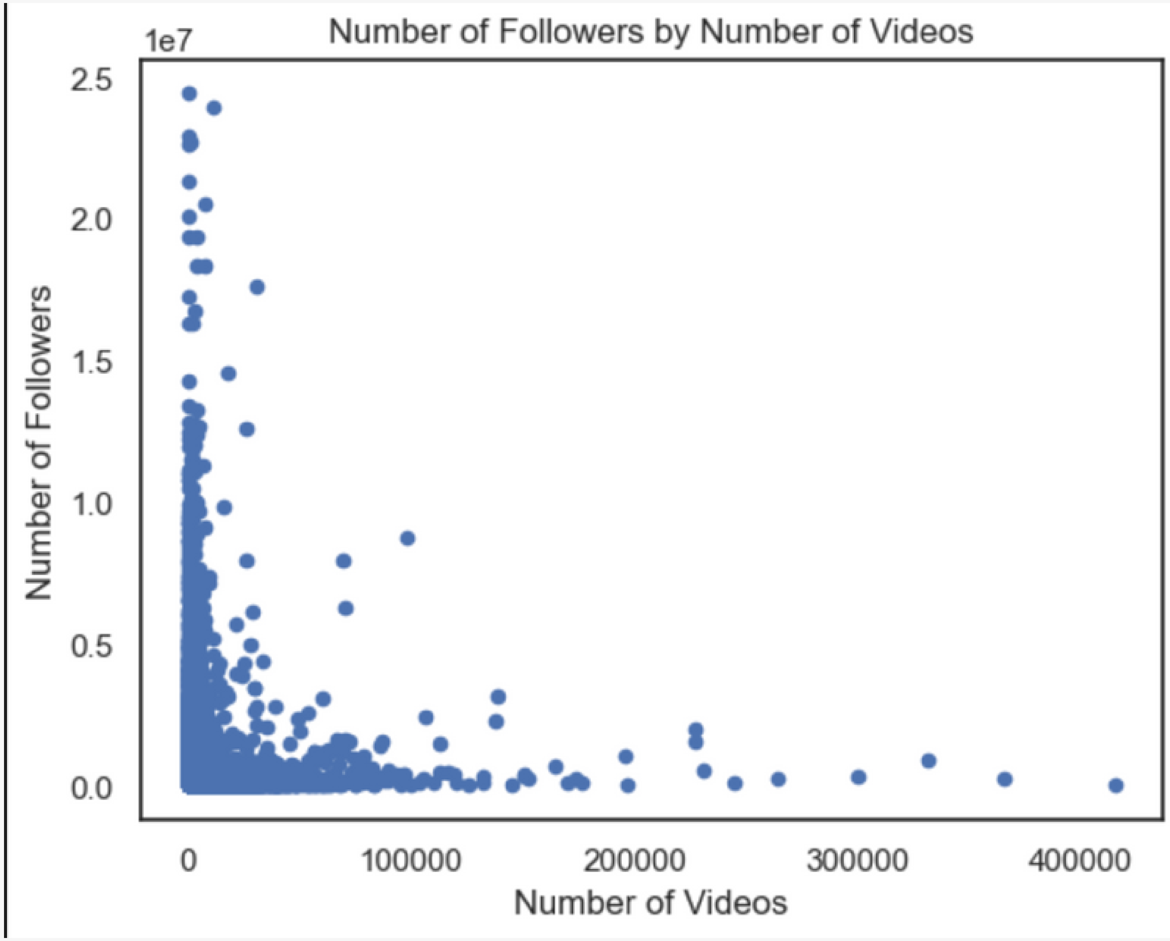
Let's see if- The more videos there are on each channel, the more subscribers there are.

Explain-

We can see that our claim is wrong, the correlation between the 2 variables is **very weak.**

Probably despite everything, the quality definitely exceeds the number of contents uploaded to the channel.

When checking the **correlation** between the 2 variables, the result was **0.09671830654836816** - a very weak correlation.


Number of Followers by Number of Videos
1e7
Number of Followers
Number of Videos