Topic Final Project – SMDA

Project title: How Video Length Influences the Number of Viewers and their Engagement

Name and email address: Hanna Müller, [hanna.3.mueller@uni-konstanz.de](mailto:hanna.3.mueller@uni-konstanz.de)

Research questions:

My goal is to investigate the relationship between the length of a YouTube video and the engagement of viewers, measured by their likes and comments as well as the relationship with the number of views. Here I differentiate between different video categories to see if there is a difference in the relationships. Then I will also apply this question to movies and see if the relationship can be replicated there.

The following questions will guide my research:

1. How is the length of a YouTube video related to the engagement of viewers?

* Here is hypothesize that the engagement for very short videos is low, then increases with a higher length and decreases again for long videos.

1. How is the length of a YouTube video related to the number of views?

* Here I hypothesize that the number of views decreases with the video length.

1. Do the relationships differ for videos out of different video categories?

* Here I hypothesize that there are category-specific effects, e.g. in the category entertainment there is more of a negative linear effect while in the category education videos with a higher length are more engaging.

1. What is the optimal length of a video in each category, maximizing the engagement of viewers and the number of views?
2. Can the relationship found for YouTube videos also be replicated for movies when looking at the relationship between the duration of a movie and its success as well as its rating?

* Here I hypothesize that the quadratic relationship is even more pronounced so that longer movies are more successful than shorter ones and only for very long movies the success decreases again.

Planned data retrieval and preparation:

1. From the website viviq.com, I will scrape the channel names of the YouTubers with the most followers out of the categories “overall", "autos", "education", "food", "entertainment" and "gaming" which are listed there as top categories of YouTube videos.
2. Out of these channels, I will create a sample of 30 channels for each category and get the information on those channels as well as their latest 50 videos using the YouTube API. I will focus on the metrics number of views, likes and comments and the duration of the videos.
3. I will exclude videos that were published only one week ago and videos that were published before 2023 to capture the current trend. Videos that are shorter than 1 minute are also excluded because those are usually YouTube #shorts which I do not want to include in my analysis.
4. For every video, I will calculate an engagement score as the number of comments and likes per user by summing the number of comments and likes and dividing it by the number of views.
5. For the movies, I use a dataset provided on Kaggle, that contains daily updated movie data from TMDb. In this dataset, I will focus on the duration of movies as well as their rating, popularity and the revenue they generated.
6. For the movies, I will only keep those released after 2023 to capture the same time span as with the videos.

Planned data analysis

1. First, I will summarize the YouTube data to understand the basic properties such as average video length, average engagement score, and average number of views across categories.
2. I will then correlate the video length with the number of views and the engagement score. I will calculate a confidence interval around the coefficient by bootstrapping. The coefficients and their confidence intervals are additionally compared between the different categories to see if the relationship differs between videos of different categories.
3. Because I expect a curvilinear relationship between the video length and engagement, I will compare a linear regression with a quadratic regression that would capture this nonlinearity using goodness-of-fit metrics like .
4. To be able to give recommendations regarding the video length for each category, I will group the videos by their length, average their engagement and view count and compute the maximum engagement score as well as the maximum number of views over all groups.
5. Finally, I will compare the found relationships to the relationship between the movie length and their revenue as well as their rating and popularity score. I will follow the same steps as before with the videos: summarizing the data, calculating correlations, bootstrapping and comparing linear with quadratic regressions by goodness-of-fit measures.

Weitere Ideen:

Contentanalyse der Videos, die trotz hoher Länge hohes Engagement haben mithilfe der Description

* videos die länger als durchschnitt sind und engagement im oberen perzentil
* description von denen analysieren möglicherweise mit large language model
* vielleicht sowas wie Wordcloud