The Study of YouTube Video Popularity

Group 7

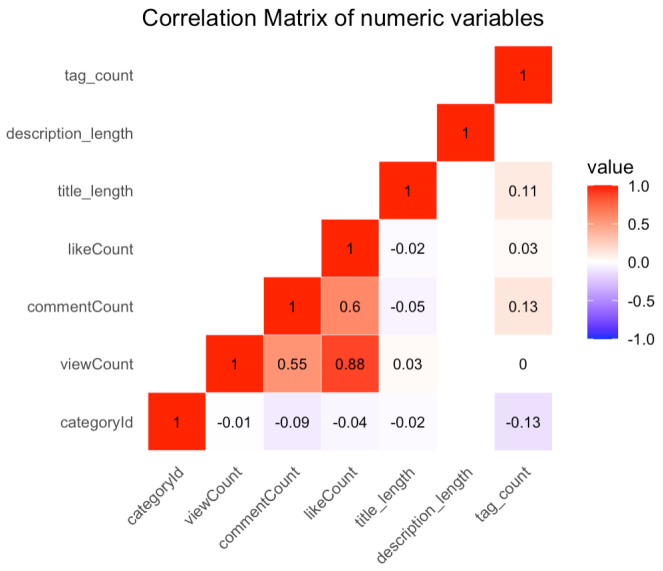
Xinyi Liang, Jinxin Li, Tianyuan Deng,
Yuhao Liu, Zahradeen Ibrahim(Deeno)

1. Statement of problem

Multi Channel Networks (MCN), also known as Internet Celebrity Incubator, is a business model derived from the video site - Youtube. MCN companies sign up promising channels and nurture them to become outlets for popular content and cash in through advertising, bandwagoning, etc. The challenge for MCNs is determining which channels have the prospects to become popular and how to guide them to create highly sought-after content. Our team stands in the shoes of these MCNs and tries to shed light on how to solve this problem. Specifically, based on Youtube, the world's largest video platform, this report aims to examine what published details related to the video affects its popularity.

2. Data and Suitability

In order to analyze this problem, we used the Youtube API tool to extract detailed information from the most popular YouTube videos. We have generated 1180 unique observations from YouTube daily trending videos over a period of 6 weeks (Feb 13, 2022 to Mar 30, 2022). Unused or unavailable variables were dropped and missing values were imputed based on other numerical variables. The dataset contains 13 numerical and textual columns, and columns of title length, description length and tag count were added for more quantitative inspiration in the subsequent analysis. The independent variables are the published details, including publishedAt, title, description, channelTitle, tags, categoryId, category_name, title_length, description_length and tag_count. The dependent variables are popularity-related data, which are defined based on view count, like count, and comment count.
(Description of Variables see Appendix A.)



The selection of variables is based on factors affecting video popularity and selected comment, view and like counts to represent the popularity based on our study of YouTube algorithm. To attract views, content creators are advised to consider factors such as the number of comments, likes, and keyword-rich video

tags in order to rank higher when users search for videos (Dean, 2017). Except for unavailable data such as the efficiency of a video generating views, our existing variables are able to largely imitate the profile of popular videos on YouTube. However, we still need to comprehensively define popularity using view, like and comment counts .

Our act of crawling the latest data is also desirable for MCNs to generate video popularity patterns on Youtube. This is because although there have been several studies on YouTube video trends, the studies indicate that trends are constantly changing. Therefore, iIt still makes sense to crawl the latest data for analysis. For instance, Music was the most viewed category in 2011-2012, and entertainment was most popular from 2013 onwards (Bärtl, 2018). Thus, for decision making of MCN companies or Youtuber investment, it is necessary to understand current trends on video category and other metadata that would potentially impact popularity of contents.

Nevertheless, for this study, our dataset has two shortcomings. The first is the small volume of data. For trend analysis, the features that can be generated with a data volume of thousand levels are not necessarily accurate. In particular, this dataset only covers a one-and-a-half-month period which makes it impossible to discharge the impact of certain events on video buzz during this specific time. The second concern is that our dimensionality is small, and the data volume of 14 columns may not cover all features included in YouTube's algorithm for trending videos. The shortcoming is mainly limited by the availability of YouTube data.

3. Suitability of analytical techniques

After inspecting and exploring the YouTube trending video dataset, several analytical techniques were being considered suitable for analyzing characteristics of popular videos. Analytical techniques we have considered and attempted include text analysis, time series analysis, correlation analysis, and prediction models.

Text analysis, using NLP to extract meaningful patterns from unstructured text in title and description, would identify important keywords and underlying themes in popular videos. In addition, we considered certain words such as giveaway or promo code included in the description could prompt more interaction behavior (more likes and comments). This valuable information allows MCN companies and content creators to design and deliver their videos within the popular themes and include most common words extracted from text columns to potentially optimize video popularity. The common words extracted would also enhance our prediction model as input columns. Sentiment analysis of text columns can help understand the relationship between positive and negative words used and the corresponding view counts.

Time series analysis can help us effectively observe the periodicity of the target variable. For video popularity, we believe that there is a time-based regularity. Since our data is not consistent (dates of popular video releases are uncontinued), it is difficult for us to use time as a variable to predict our three popularity metrics. But effective visualization can help us observe the occurrence of abnormal events, and as the data expands, we can also see the fluctuation trend of popularity and future trends

Another important analysis for this dataset is correlation analysis for each video category, which investigates the relationship of YouTube metrics between each other. These metrics include independent variables that content creators can manage (title length, description length, tag count) and dependent variables of view count, like count, and comment count. This analysis shows the differences of relationships between these variables within different categories. This could be valuable to content creators within certain categories to adjust their videos' user input information and optimize audience interaction. For instance, in one category, longer description length would have a highly negative correlation with view/like/comment counts, while it might be the opposite result in another category.

In our dataset, we have also constructed a positive action index from the viewers' interaction columns. Since the algorithm that YouTube uses to determine which popular videos will be on the trending list is unknown, we theorized two scenarios. On one hand, popularity is evaluated by the average view count of any video category as YouTube officially stated that how quickly the video accumulates views since being published is an important factor in the algorithm. High popularity is a goal for videos and can be self-determined as a KPI for MCN companies. Positive action index, on the other hand, determines the percentage of interactions per view, and is calculated from **(number of comments + number of likes)*100 / view counts**. High positive action index indicates high viewer stickiness and would also be meaningful targets for MCN companies.

After all the variables are constructed, we are able to perform predictions on popularity index and positive action index. For the independent variable, we choose category name, title length, description length, number of tags and number of keywords in description and title. There will be two sets of keywords: the first set focuses on video metrics, such as 'subscriptions', 'twitter', 'reply' and the second set focuses on contents, such as 'games', 'music'. The two sets of keywords will be examined in separate models. By comparison, we can determine how the content and video settings affect the user's decision on view and interaction using linear model, decision tree and logistic regression. The goal for the linear model is to provide a basic line for all the parameters.. From the results of the linear model, we could obtain p-value for each variable which will be helpful to conclude what is more important when videos are being published. To understand people's decision about whether to watch the video and interact with the video, a new variable was created which selects the top 50% scores as true and bottom 50% scores as false. The classification decision tree and logistic model are the best tool to predict categorical variables. Compared with Random Forest and Neural Network, decision trees have the best interpretability. For our models, high accuracy is ideal, but it is not the primary goal. It is more important for us to interoperate user's behaviors. The logistic model helps us to assess the accuracy of the logistic model and avoid extreme results.

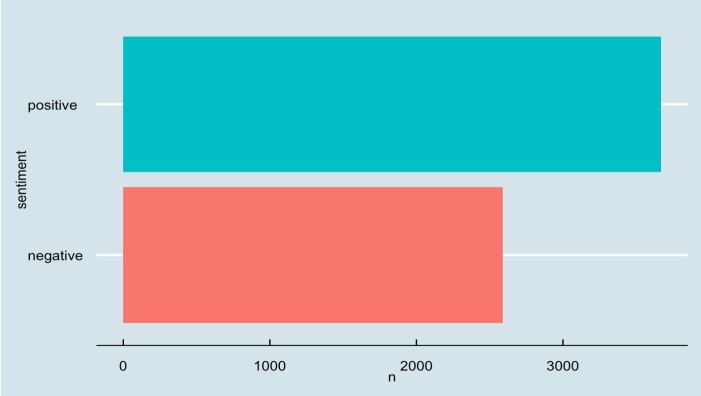
4. Results from analysis

Text Analysis:

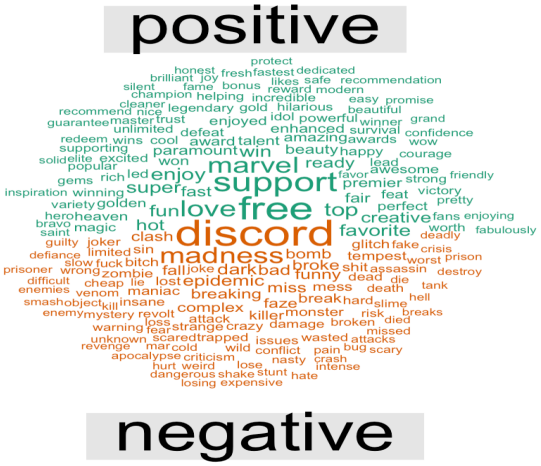
Our sentimental analysis explores the relationship between three variables namely, Description, Tags and Title with the Number of views.

First, we explored the Description variable which has an average of 945 characters, 108 words and 14 sentences per video. While the majority of the descriptions are in English, some are in other languages.

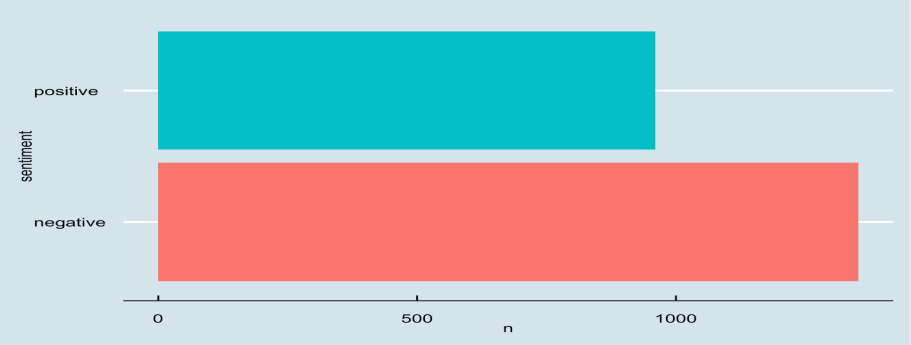
No strong correlations between length of characters, words and sentences, capitalizations, and exclamations with the respective view count. The description also exhibits 3,668 positive sentiment words, representing 58.62% of total sentiments compared to 2,589 or 41.38% negative sentiment words.



Sentiment of words in the description variable do not have influence on the category of the videos as there are both positive and negative words in each Category ID. (See Appendix B. for distribution of sentiments in video description). Furthermore, correlation between positive words in the description column and the viewCount (-0.032) does not suggest that reviews with a lot of positive words are viewed more. Below is a comparison cloud for the sentiments.



Secondly, we explored the Tags variable. This variable has an average of 250 characters and 30 words per video. Tags are usually a list of words preceded by the # sign and do not make up full sentences and the variable contains 295 NULL entries. No strong correlations between length of characters, words, capitalizations, or exclamations with the respective view count. The Tags variable also exhibits 1,352 negative sentiment words, representing 58.48% of total sentiments compared to 960 or 41.52% positive sentiment words.

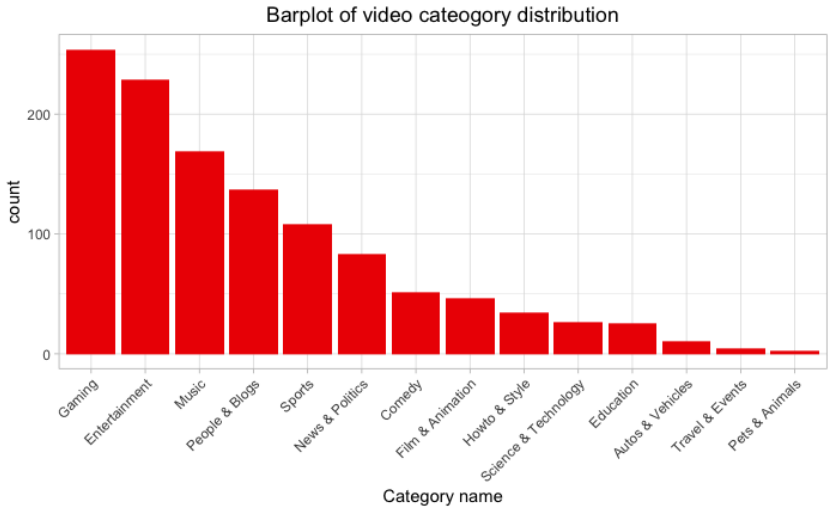


Sentiments of words in the Tags variable have no influence on the category of the videos as there are both positive and negative words in most Category ID (See Appendix C). Additionally, the correlation between positive words in the Tags column and the view count (-0.028) does not suggest that reviews with a lot of positive words are viewed more (See Appendix D. for word cloud of the Tags variable).

Finally, we analyzed the Title variable which has an average of 55 characters, 9 words and 0.43 sentences per video. No strong correlation exists between length of characters, words and sentences, capitalizations, and exclamations with the respective viewCount. The Title variable exhibits 341 negative sentiment words (56.36%) and 264 (43.64%) positive sentiment words (See Appendix E. for sentiment distribution in title variable). CategoryId is not dependent on sentiment of words in the Title of the videos, as there are both positive and negative words in each Catergory ID (See Appendix F. for title sentiment distribution between categories) . The correlation between positive words in the Tags variable and the viewCount (-0.032) does not suggest that reviews with a lot of positive words are viewed more (See Appendix G. for comparison cloud on tag variable).

Category analysis:

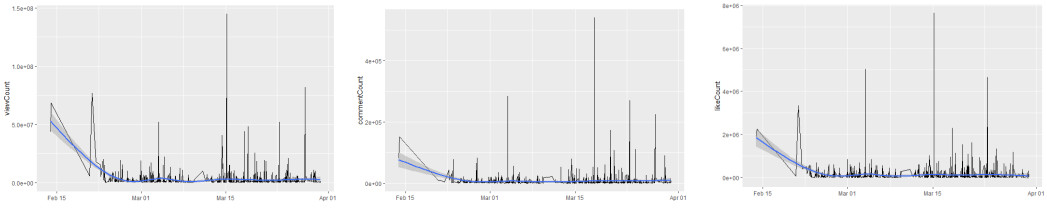
Based on the category distribution of the dataset, Gaming category has the highest number of videos, followed by entertainment and music. The trending video dataset has the least number of Pets & Animals videos among all. This indicates a change in popular video categories from time, as Music was the most viewed category in 2011-2012, and entertainment was most popular from 2013 onwards. Considering the effect of Covid19 Pandemic Lockdown, gaming videos could have gained huge popularity while travel videos might have lost many viewers. The distribution of video categories on YouTube's popular list is critically important to show MCN companies and content creators the ever changing trends and size of viewer groups.



Within different categories, education, music, and People & Blog videos have the highest average view counts and videos in Howto & Style, Travel & Events, Autos & Vehicles have the lowest average view counts (See Appendix H. for average view count in each category). However, the distribution of like and comment counts is slightly different. News & Politics category, which is the lowest in average like count, has the second highest comment count among all categories. Considering recent turmoil in politics, it could be possible people are commenting more in negative news videos instead of liking the videos (See Appendix I. for average like & comment count for each category) .

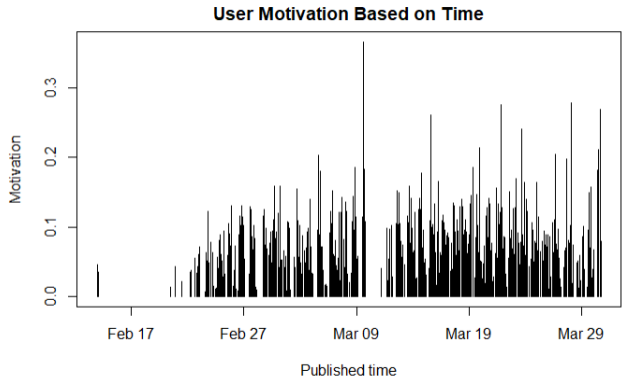
Time Analysis:

The original data format was character, and we had to convert it to POSIXct format. Since our data uses popular videos at release time, time is discontinuous and cannot be predicted by time, since popular videos are not a periodic event either. Therefore, we decided to observe the characteristics of our popular indicators during the month according to time. The following picture shows the relationship between the total value of views, comments, and likes of popular videos and time from February 15th to April 1st.

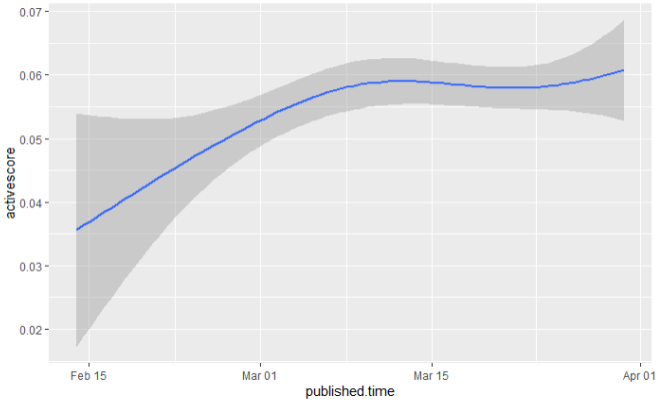


The charts show that starting from March 8, there was a significant increase in views, comments and likes, signaling some exceptionally popular videos that stand out on this date. Below is a time series

based on our definition of popularity ((view+comment)/like).



From this figure, we can find that on March 19th and 20th, user activity increased by leaps and bounds. The reason could be the result of something big happening, setting off a trend of public opinion on the Internet. The last graph is about activity trends. According to the figure below, we can see that the user activity is gradually increasing, and then fluctuates slightly.



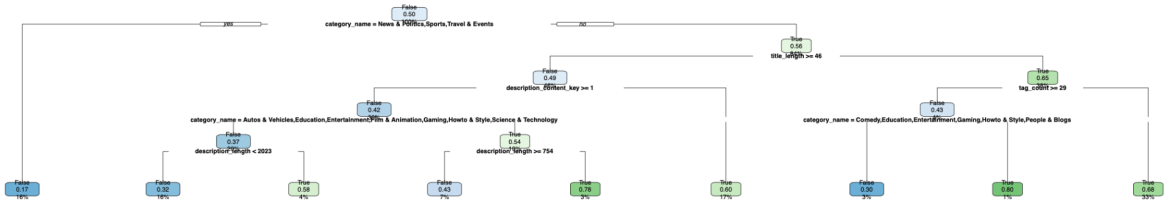
Correlation Analysis:

We conducted correlation analysis between video metrics among the 14 different video categories included in this dataset. Interestingly, for the category of Pets & Animals, Howto & Style and Autos & Vehicles, title length has significantly high negative correlation with comment, like and view counts. This means that videos in these categories would appear less attractive to viewers with longer titles. Moreover, most categories have high positive correlation between comment, like and view counts, which means videos in these two categories would usually have more likes and comments with more views. However, comedy and entertainment videos have a much weaker positive correlation between comment count and like count. (Correlation plots see Appendix J.)

Modeling and prediction:

We used some modeling techniques to examine the word choice, title length and choice of the video. Using the linear regression model, none of the independent variables were significant when predicting views, but a few variables were significant when predicting action scores, which were description length(p-value= 0.02655), description_content_key(p-value = 0.05280), title_content_key(p-value =

0.01282) and title_metrics_key(p-value = 0.20680). It appears that the content and video metric choices will not affect the views but it will impact on how users interact with the video. The interaction rate may vary in each category, and the relationship is worth investigating.
The prediction from logistic regression and decision tree provides a similar conclusion. The accuracy scores in view predicting were 55.37%(logistic regression) and 58.19%(decision tree); the accuracy scores in action predicting were 64.41%(logistic regression) and 59.32(decision tree).



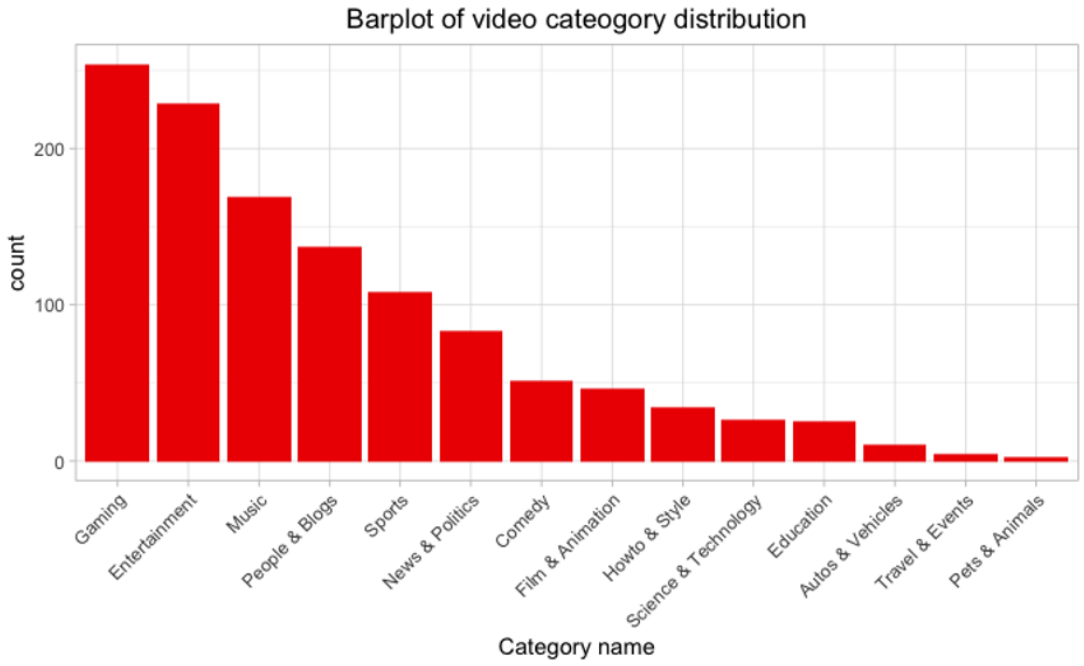
From the graph, we can see that category has the highest impact on action scores. Categories including News & Politics, Sports and Travel & Events have the lowest action score. If we are looking for high user engagement in our videos, the above topics should be avoided. Longer title length (more than 46 characters) and higher number of tags (more than 29) could lead to more interactions.

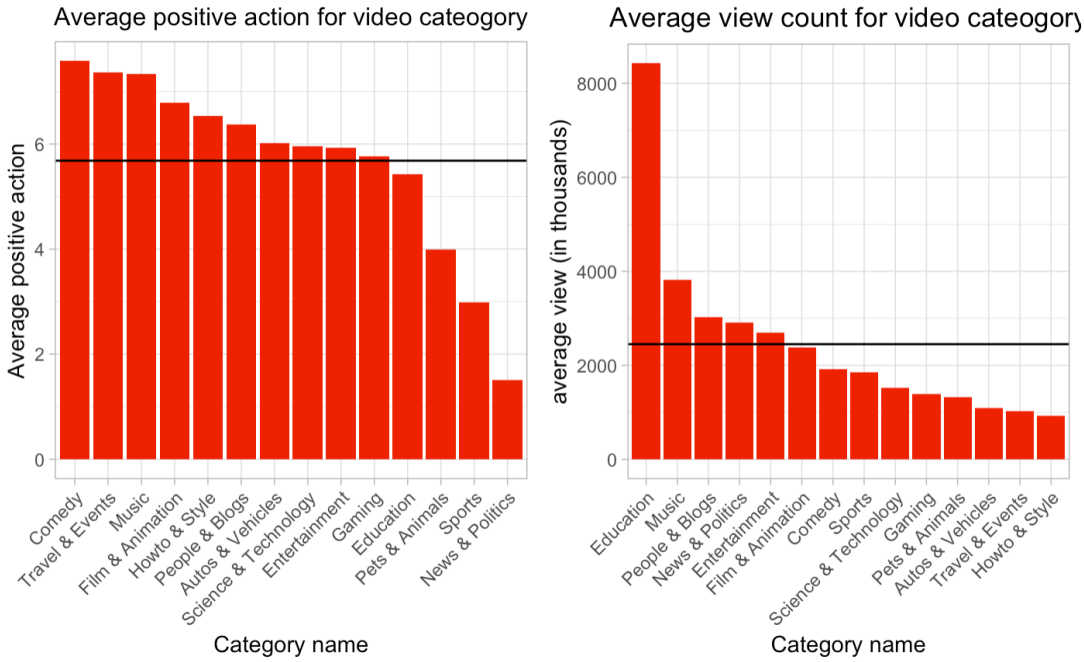
5. Conclusions and recommendations

Consequent upon our analysis of the data related to Youtube videos, we provide the following insights that would enable MCNs determine the appropriate content that will generate maximum views, likes and comments from viewers:

First, in terms of choosing the channels with potential, we suggest MCNs give primary consideration to the Music category. For one thing, Music has remained in the top three “hottest” videos after becoming YouTube's most popular category in 2011 to 2012. This indicates that audiences’ love for Music is enduring, which is important for MCNs who may need to invest in nurturing accounts over time to start reaping the rewards. On top of that, even though Gaming and Entertainment are also featured in popular videos with high frequency, Music is the only category that ranks in the top three in terms of frequency of popular video appearances, average views and positive action metrics (including likes and comments). This indicates that videos in the Music category tend to be more popular, with more active user interaction, and that this distribution of this popularity is more evenly distributed across all music videos. This even distribution is important for MCNs because the risk of investing in a particular channel in a category that is not popular is lower for MCNs if the videos in that category are more popular overall. A clear example of this comparison is the Gaming category, which accounts for the majority of popular videos, well ahead of the second and third places. But it ranks in the bottom few in both average views and positive interactions. This suggests that audiences’ attention to the category is focused on a small number of videos and channels, and that investing in accounts in this category is as high-risk as it is high-reward. However, a viable strategy for risk-averse MCNs is to look specifically at which specific gaming videos and accounts are popular and see if there are possibilities for collaboration.

Secondly, the Education category is worth paying attention to because educational videos on YouTube enjoy a very high average view count, which is twice that of second place - Music, but appear less frequently in popular videos. This shows that the audience likes educational videos but just not usually to make them breakout videos. Therefore this category is also suitable for MCNs who seek a more balanced risk and reward to consider.





In terms of optimizing video publish details to improve channel operations, we found that audience attention and interaction are not evenly distributed in time, especially when there are big events and popular news broadcasts, audience attention rises sharply making them more likely to interact with the video. Therefore, MCN companies need to detect news and events from all walks of life in a timely manner, and the "password" to get YouTube traffic is to follow the “hot spot” and produce related content in good time. In addition, we noticed that there is no clear positive connection between audiences’ attention and the text part of the video (description, title, etc.), including the length of the text, sentiment, keywords, etc. However, according to the decision tree prediction, longer videos are more likely to be viewed by users. However, according to the decision tree prediction, longer texts, more tags and keywords tend to cause more positive user interactions. Therefore, for MCN companies, text optimization still makes sense for improving video publishing.

Finally, a small tip for MCNs from the relevance analysis is that videos in the Pets & Animals, Howto & Style, and Autos & Vehicles categories should not be titled too long, as they may face low user attention and interaction.

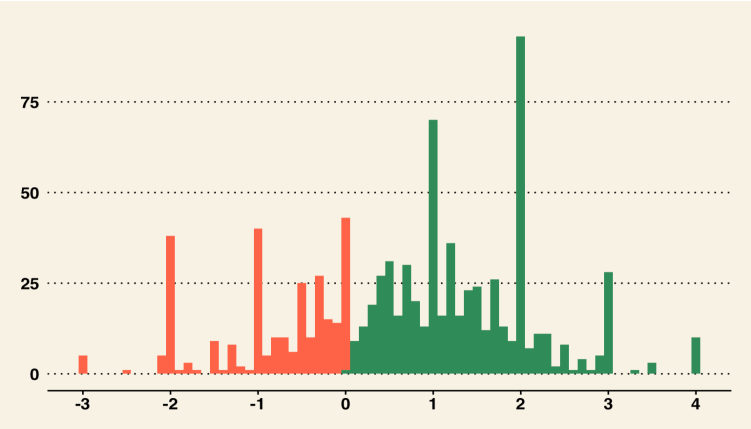
Overall this study provides an attempt for MCN companies to understand popular trends through data and optimize account operations, and gives some corresponding suggestions. However, due to the limited data - limited amount, narrow dimension, and time inconsistency - it does not give more comprehensive and accurate guidance. Nevertheless, MCNs can continue to track popular video data along the lines of this attempt, and use more comprehensive data sets to dig more helpful and inspiring insights.

6. Appendix

A. Description of Variables

Variable Name	Variable Explanation
publishedAt	The publish time of the video
title	The title of the video
description	The detailed description of the video
channelTitle	channel name of the video published
tags	Tags provided to creators on YouTube to highlight the video content.
categoryId	Video type numbers, each number corresponds to a video category name
category_name	Video category corresponding to category number
title_length	The length of the video title in words
description_length	The length of the video description in words
tag_count	The number of tags used by a video
viewCount	How many times a video was viewed
commentCount	How many times a video was commented
likeCount	How many times a video was liked

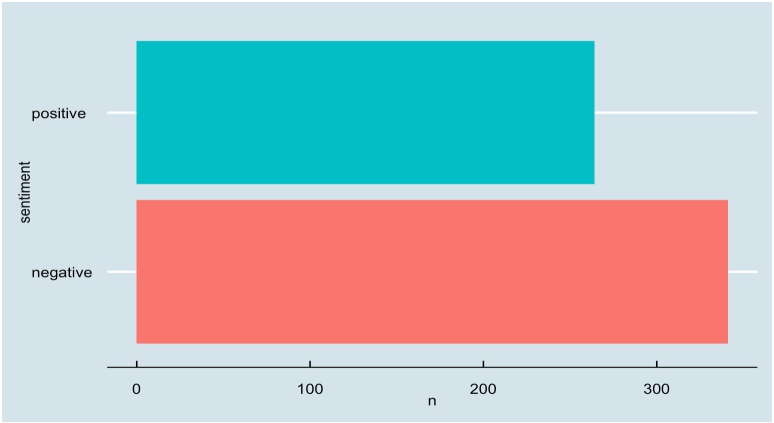
B. Distribution of sentiments in description variable



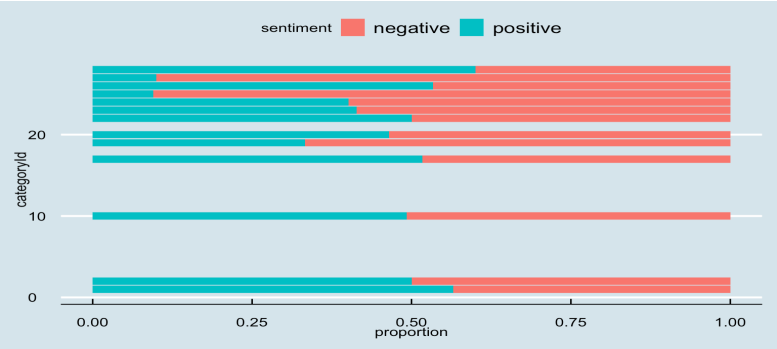
C. Sentiment distribution between categories in tag variable

D. Word cloud of the Tags variable

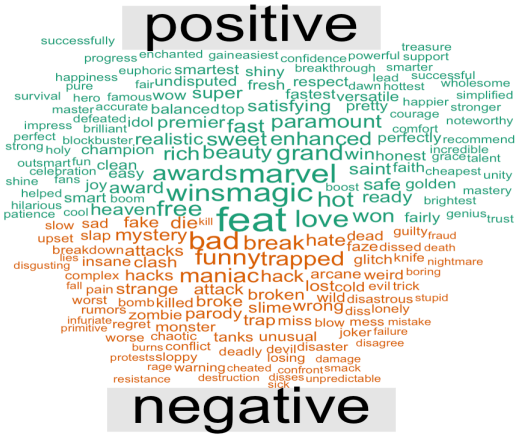
E. Sentiment distribution in title variable



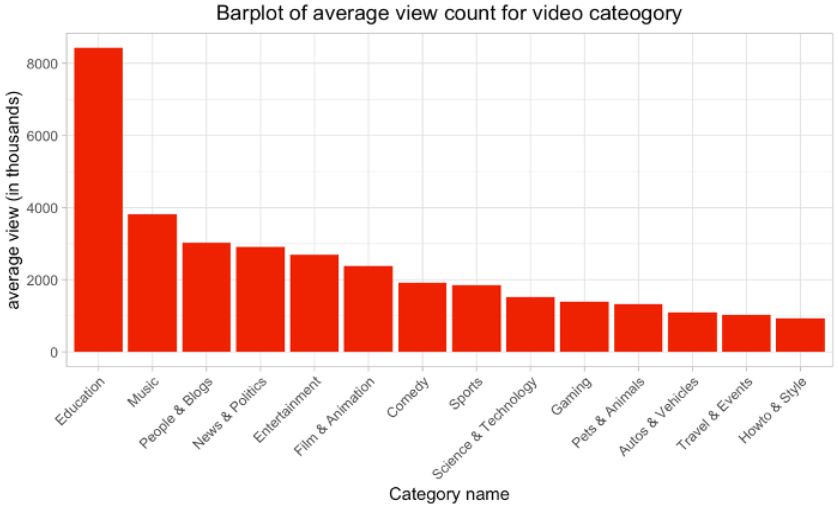
F. Title sentiment distribution between categories



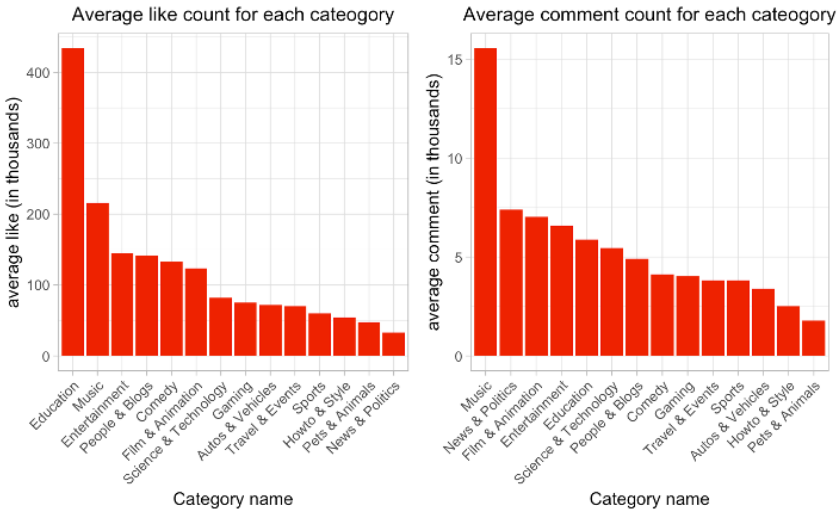
G. Comparison cloud on tag variable



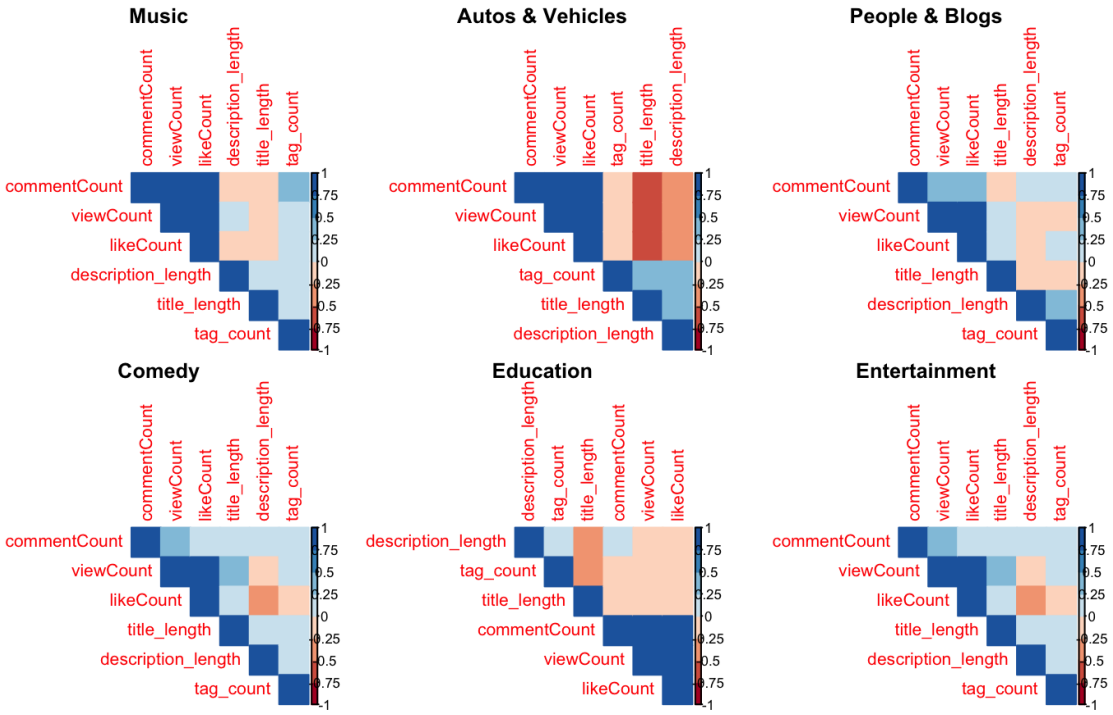
H. Bar Plot of average view count in each video category



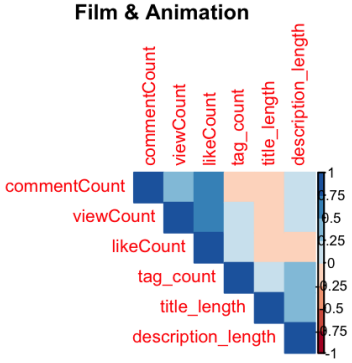
I. Average like count & comment count for each video category



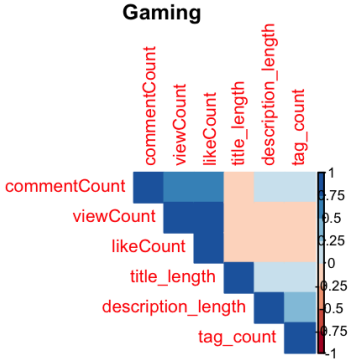
J. Correlation plot between categories for YouTube metrics



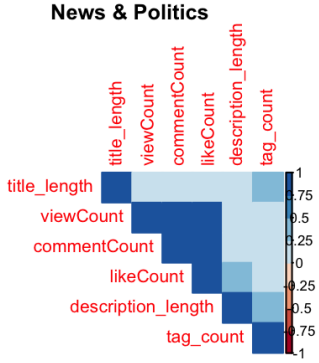
Film & Animation



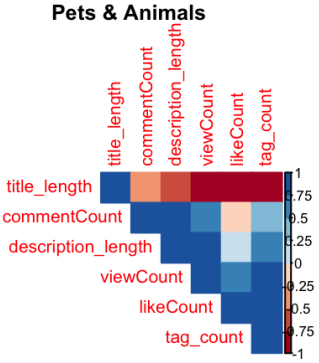
Gaming



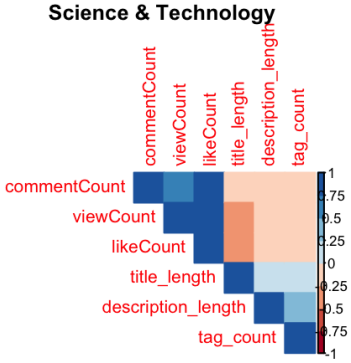
News & Politics



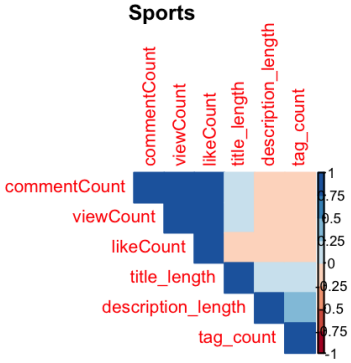
Pets & Animals



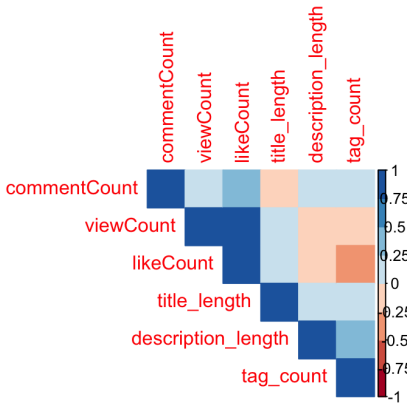
Science & Technology



Sports



Howto & Style



Travel & Events