YouTube Video Titles & Their Influence Over Views

Analyzing How Title Characteristics Affect the Number of Views

Sai Bharath Bobba, Jagadeesh Chandrabose Gurram, Dillon Price, Rohit Raju, Ankit Rai Sharma & Punitha Vancha

18 December 2023

YouTube creators are prolific and deploy many tactics in an attempt to gain traction and virality on their content. This paper looks at the interplay between title characteristics, like length and categorization, and the number of views a video receives. The analyses suggest that video titles play a massive role in influencing the number of views.

.

1 Introduction

Since 2005, YouTube has become the cornerstone of content creation, and emerged as one of the most advanced and important platforms in this dynamic landscape. The increasing number of content creators joining the platform to effect change in society and share their beliefs highlights the growing significance of optimizing videos to reach the intended audience and beyond.

YouTube provides content creators with the freedom to tailor their titles, thumbnails, tags, and descriptions to craft an enticing preview of their videos before uploading them. While content creators might recognize the importance of an intriguing title, the value a well-engineered title brings is overlooked. Video titles must be descriptive but not prolonged to be recommended to an audience, and this click-through decision is often influenced by crucial factors such as the title, thumbnail, and duration of videos—features visible to the audience. Among these features, titles often carry a greater weight, as a strong title can significantly impact a video's click-through rate.

While research efforts aim to decipher the YouTube algorithm to increase impressions, it's essential to note that impressions alone do not guarantee an improvement in click-through rates. YouTube titles vary widely, ranging from short to long, cased to uncased, belonging to different video categories, and featuring different use of special characters. Upon closer inspection, it becomes apparent that certain videos within a YouTube channel outperform others, primarily because some video titles are more attractive and concise.

This study focuses on uncovering the influence of YouTube titles on views. Through rigorous hypothesis testing on parameters such as title length and its impact on views, as well as the title category's influence on views. The aim is to derive a comprehensive understanding of the importance of constructing a strong title for videos to enhance their visibility and engagement.

2 Data

2.1 Data Management

This paper uses the Python 3 programming language as well as the packages SciPy (Virtanen, Paul & et. al. 2020), NumPy (Harris, C.R. & et. al. 2020) and Pandas (McKinney, W. & et. al. 2010). The figures in this paper were created using Plotly (Plotly Technologies Inc. 2015), Seaborn (Waskom, M. L., 2021) and Matplotlib (Hunter, J.D. 2007).

2.2 Source

For this paper, data was extracted via YouTube’s public API. In order to use the public API, one must specify a specific YouTube channel ID to obtain information associated with the channel’s videos. To achieve a diverse collection of videos across many subject matters, YouTube channels were chosen whose self-described focus aligned with one of the following subject matters: News & Current Events, Gaming, Comedy, Film, Entertainment & Music, Education, or Family / Lifestyle.

2.3 Sampling

For any given channel on YouTube, there can be a large variance of how long a channel has been publishing videos and how many people are subscribed to said channel. In order to control these large variances, the data was limited to videos that were published more than 3 months and less than 2 years ago. These constraints made certain that A) a video was likely to have entered YouTube's recommender system and B) views were not over inflated by videos living on service for an extended period of time.

2.4 Key Features

The publicly available data from YouTube API scrapes encapsulates the number of views a video received, the algorithmically tagged category associated with the video, the video's title, publish date, the number of "likes" and "shares", and video and channel IDs. For the purposes of this research, the key features of the data set are the publish date, number of views, video category and title. The video's title length was derived by measuring the number of characters used which maxes out at 140 characters (a flaw that is addressed in Section 3.1).

3 Methods

3.1 Exploratory Analyses

Since the relationship between a video's categorization and a channel's self described focus are independent of each other, it was important to understand the distribution of video categories in the data set. Figure 1 examines the volume of videos within each algorithmically tagged category with the most common ones in the data set being "Education", "Gaming" and "How-to & Style".

Figure 1

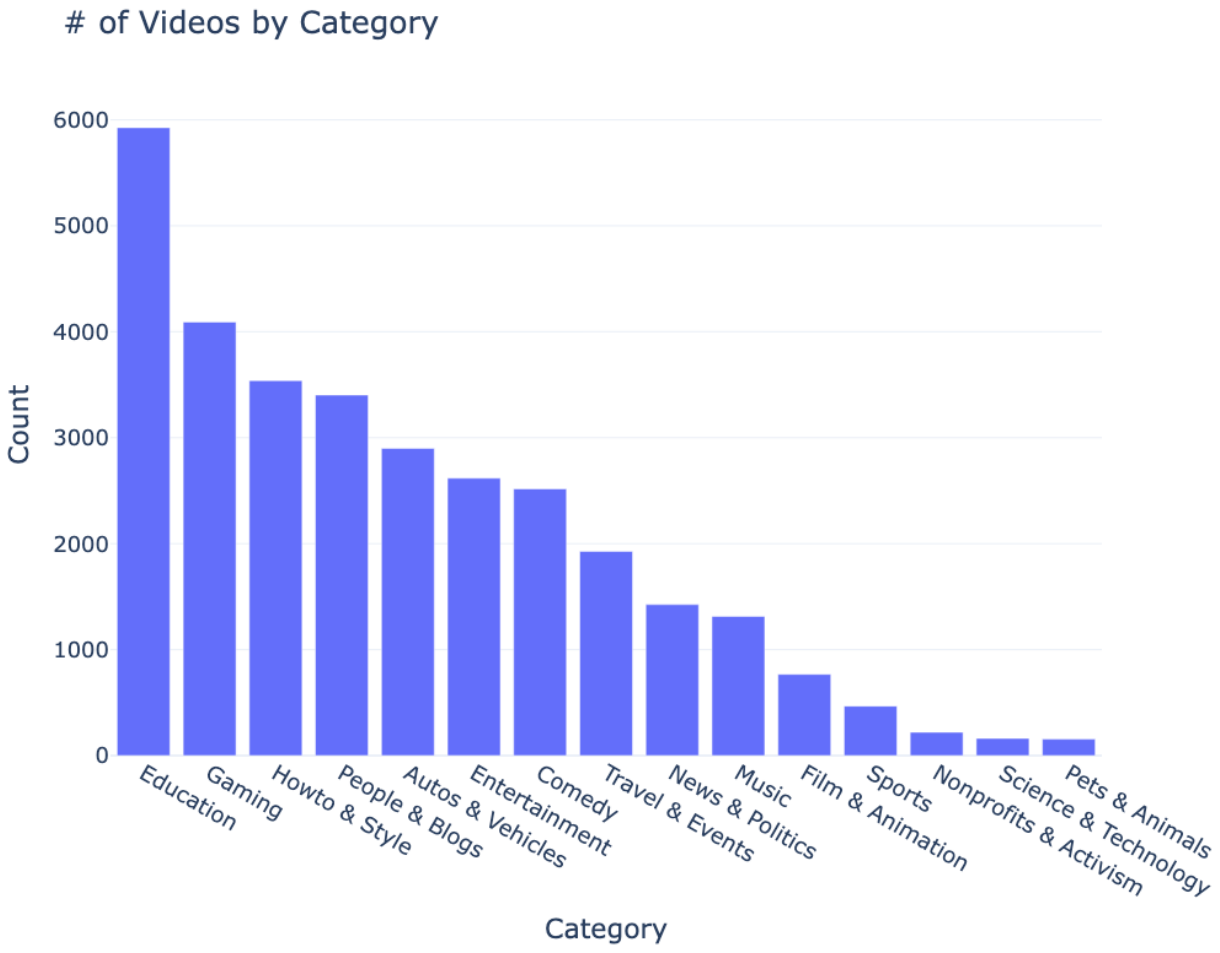
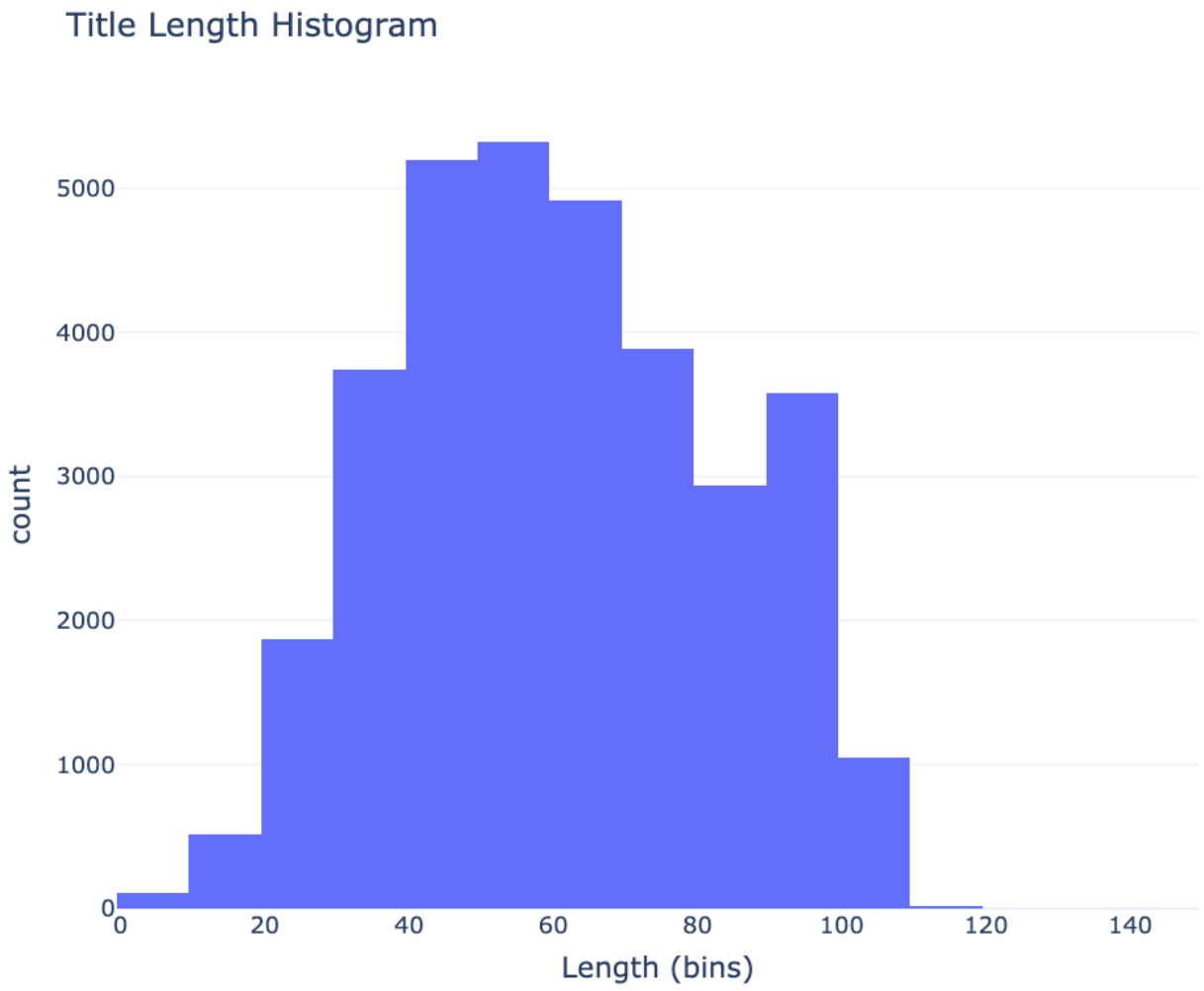


Figure 2 showcases the distribution of title lengths. YouTube sets a limit of 100 characters for a video's title. However, due to the way special characters, like quotation marks, ampersands, etc. are encoded and subsequently scraped from the channels, a handful of records (277 of the ~33,000) had title lengths that exceeded YouTube's limit. Those records were dropped from the data set for posterity and alignment to the content platform.

Figure 2



Additionally, there was interest to understand the distributions of the two title characteristics – title length and category – both in video volume and view counts. The following three figures were used in conjunction to determine which video categories should be tested against each other, and how to best approach testing title lengths.

Figure 3 looks at the total number of views in a category against the total number of videos in the data set. From this view, it is clear that videos categorized as Education are a clear outlier both in view count and the number of videos.

Figure 3

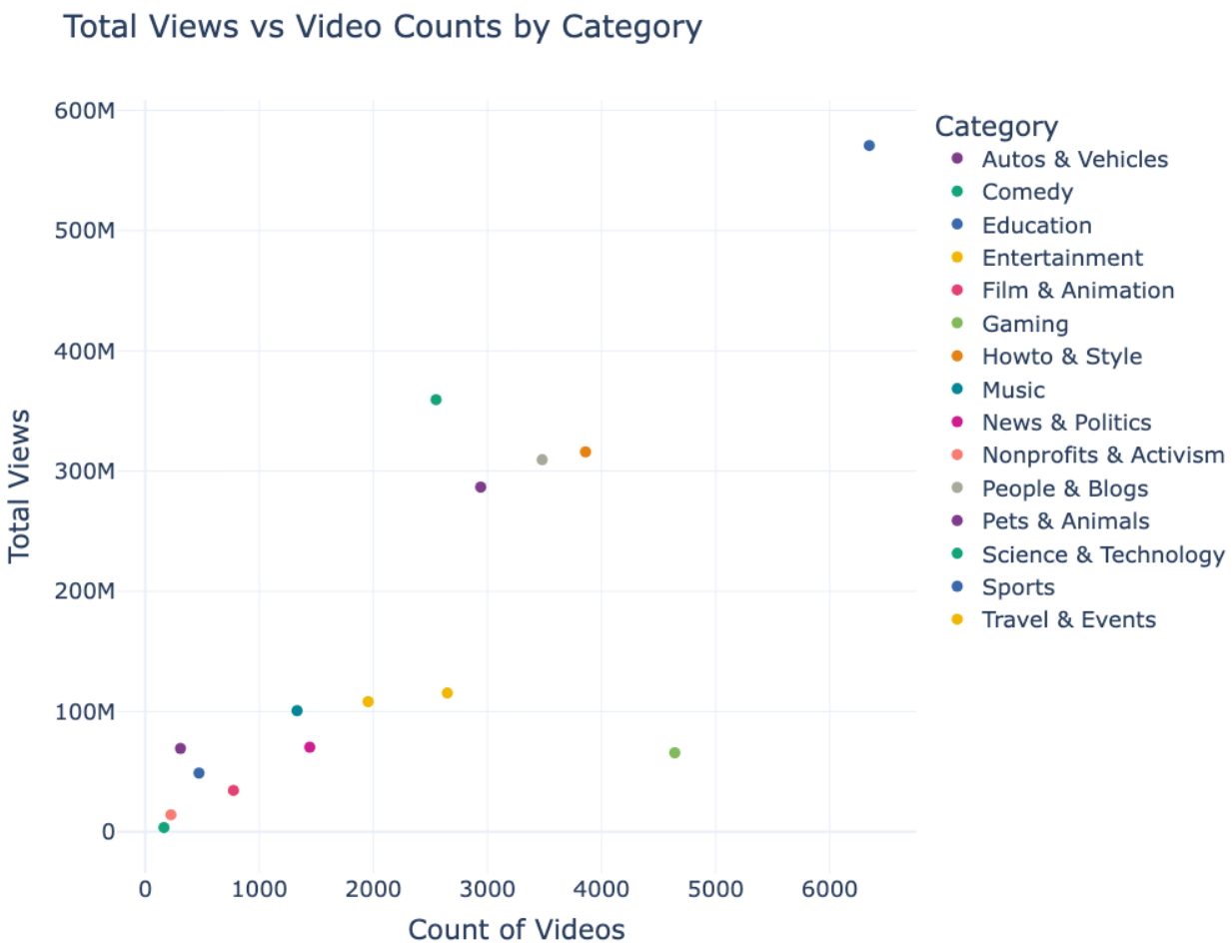
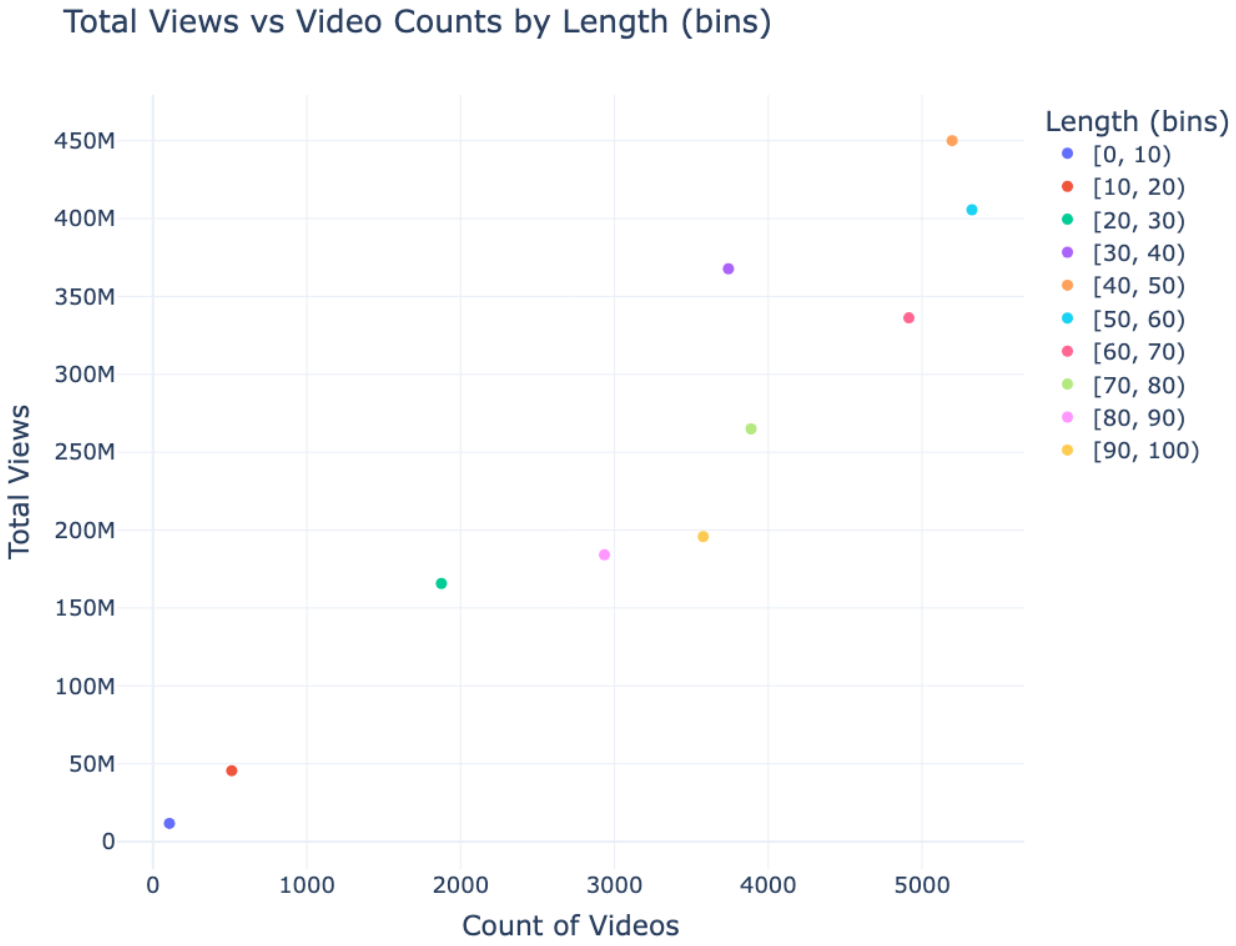


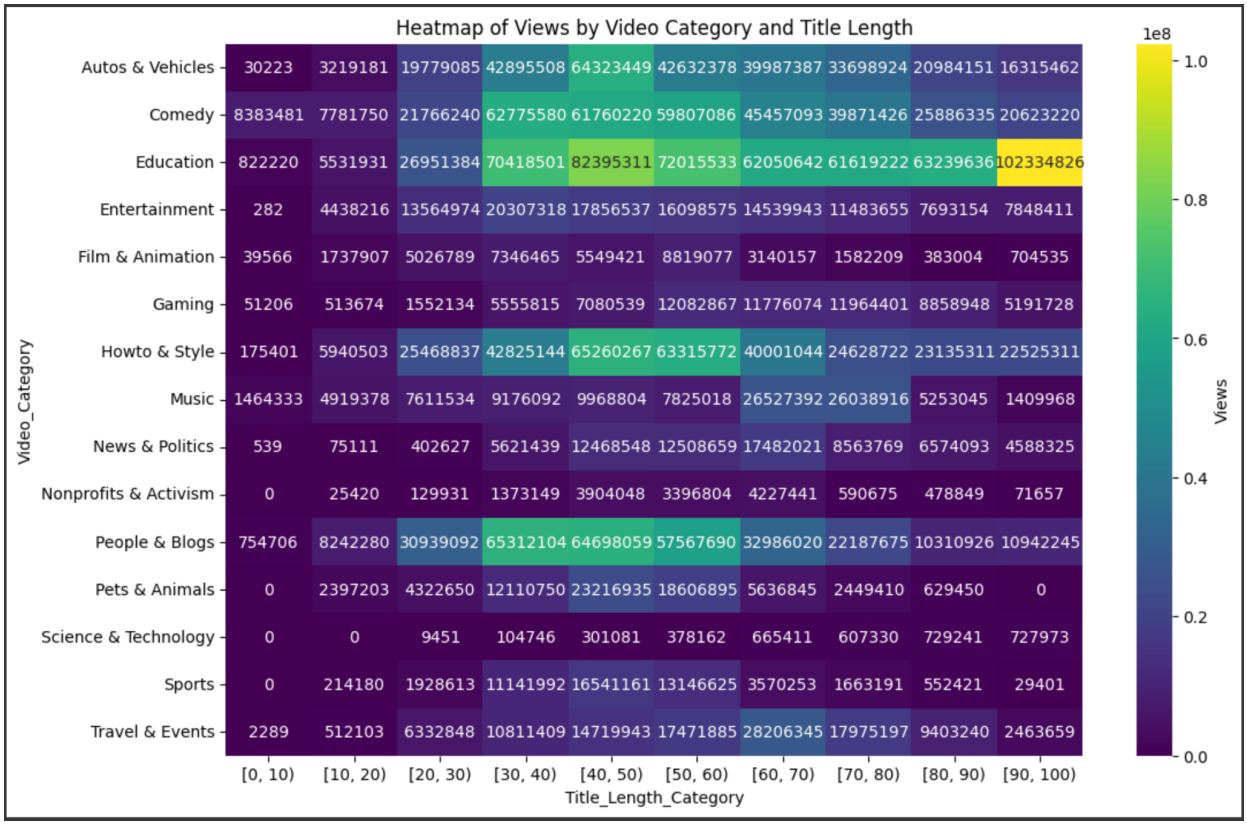
Figure 4 takes the same approach as Figure 3 but instead of category it is broken out by title length bins. In this view, it is observed that title lengths between 0 to 20 characters have the lowest number of views as well as number of videos.

Figure 4



Lastly, Figure 5 shows the total number of views by title length and video category. In this view, there is a concentration of views around half the character limit imposed by YouTube (a similar observation is made in Figure 2 but with the emphasis that this tendency exists across multiple categories). This threshold of 50 characters was used for the title length hypothesis test.

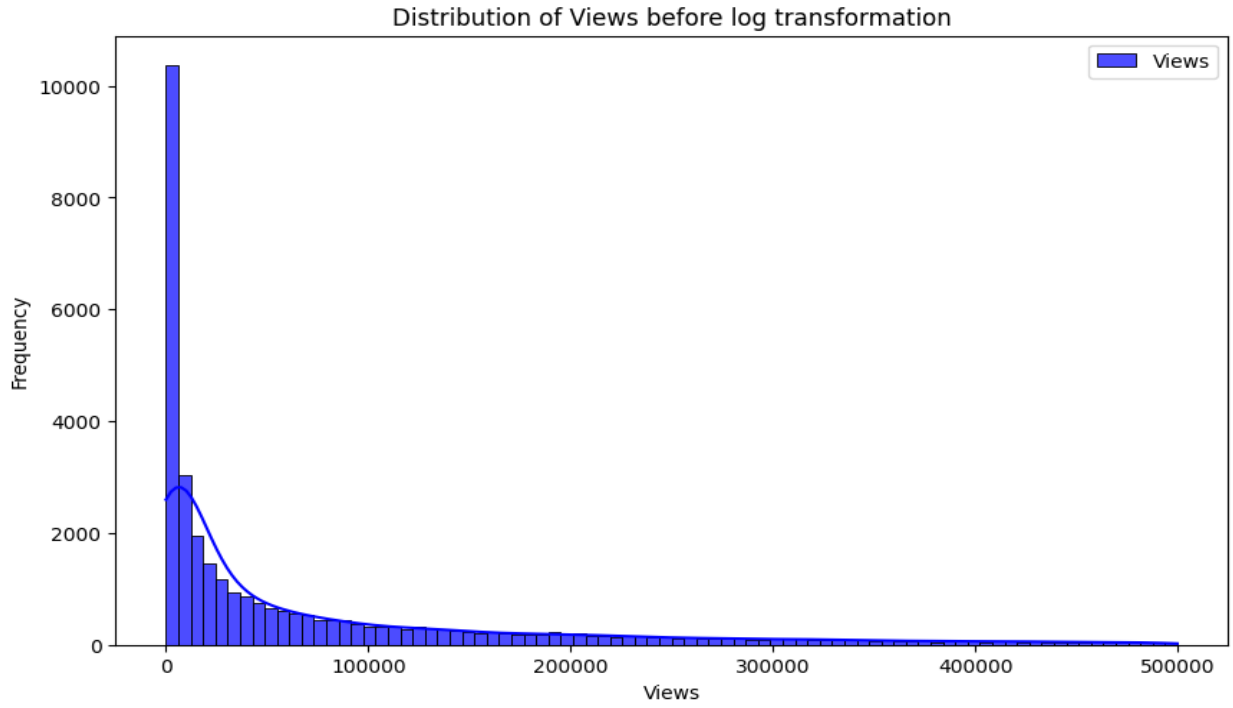
Figure 5



3.2 Data preprocessing for hypothesis testing

The key prerequisite to accurately carry out the hypothesis testing is to ensure that the data satisfies the modeling assumptions, most importantly the normality of the dataset with constant variance. Figure 6 displays the dataset exhibits a strong right skewed distribution with a long tail, which would result in a biased hypothesis.

Figure 6



To prepare the raw dataset for hypothesis testing meeting model assumptions, a natural log transformation has been applied. As seen in Figure 7, the log transformation has removed the strong right skew and closely resembles normal distribution. The R^2 value of 0.9520 in Figure 8 signifies that the variance has covered about 95% of the data, which is considered optimal representation of population.

Figure 7

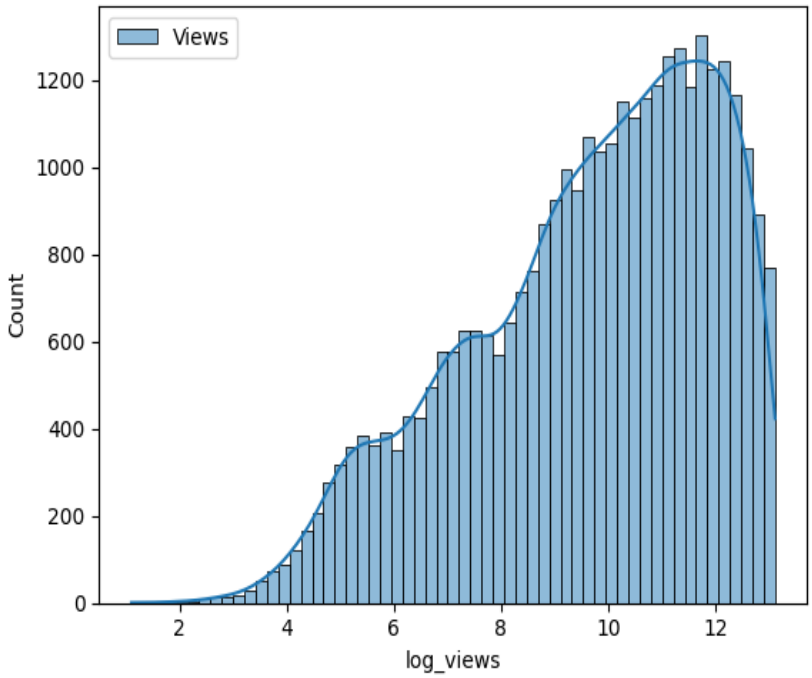
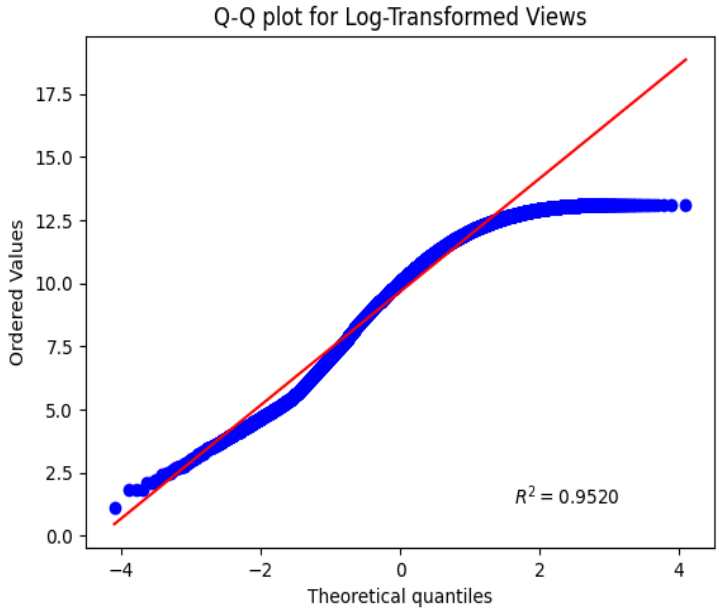


Figure 8



3.3 Examining Views in Relation to Title Character Length & Video Categorization

This section and the following section states the hypotheses to evaluate the research question: Does video title significantly influence the number of views? To analyze the relation between video title and views, three hypotheses are stated as:

Hypothesis 1:

- Null Hypothesis (H_0): The mean log-transformed views of videos with a title length 50 characters or less are equal to the mean log-transformed views of videos with title length greater than 50 characters.
- Alternative Hypothesis (H_1): The mean log-transformed views of videos with a title length 50 characters or less are not equal to the mean log-transformed views of videos with title length greater than 50 characters.

Hypothesis 2:

- Null Hypothesis (H_0): The mean log-transformed views of videos under the "Gaming" category are equal to the mean log-transformed views of videos under the "Entertainment" category.
- Alternative Hypothesis (H_1): The mean log-transformed views of videos under the "Gaming" category are not equal to the mean log-transformed views of videos under the "Entertainment" category.

Hypothesis 3:

- Null Hypothesis (H_0): The mean log-transformed views of videos under the "How-to & Style" category are equal to the mean log-transformed views of videos under the "People & Blogs" category.
- Alternative Hypothesis (H_1): The mean log-transformed views of videos under the "How-to & Style" category are not equal to the mean log-transformed views of videos under the "People & Blogs" category.

Since views in each category used for comparison are transformed into a logarithmic space, the results obtained from the log-transformed space can be leveraged to gain insights into views in the non-log-transformed space. Consequently, working in a log-transformed space will not impact the results in the non-log-transformed space.

Establishing significance level:

The threshold for the statistical significance in this analysis is p-value = 0.05. Any p-values calculated below 0.05 will indicate sufficient statistical evidence to reject the null hypothesis.

Analysis Approach:

Inorder to test this hypothesis, A two-tailed t-test has been performed. If the t-test value is positive it provides evidence that former is greater than later.

4 Results

The results of the hypothesis test are as states in Table 1:

Table 1

Hypothesis	t-value	p-value
Hypothesis 1	18.457	1.445×10^{-75}
Hypothesis 2	13.816	7.062×10^{-43}
Hypothesis 3	10.499	1.329×10^{-25}

Hypothesis 1 Analysis:

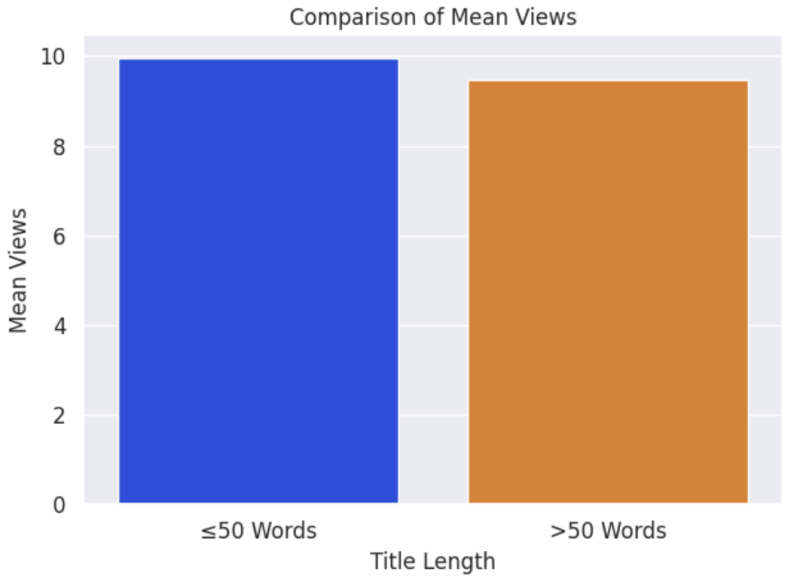
Test Statistic:

The t-test resulted in a test statistic value of ~18.475. This substantial absolute value suggests a significant difference between the mean log-transformed views for shorter and longer video titles. Since views in both categories, shorter and longer video titles, are transformed into a logarithmic space, it can be inferred that a significant difference exists between the mean views for shorter and longer video titles.

p-value

The p-value from the statistical test is 1.445×10^{-75}. Being substantially smaller than the 0.05 significance level, this provides strong evidence to reject the null hypothesis. There is very statistically significant evidence that the shorter and longer title length groups do not have equal mean log-transformed views.

Figure 9



As displayed in Figure 9, the group of videos with ≤50 character titles has a higher mean log-transformed views compared to a lower mean view for the >50 character title group.

The substantially higher viewership for more concise video titles is evident in this summarized comparison of central tendency. This aligns cleanly with the interpretation of the hypothesis test results rejecting the null hypothesis of no difference between mean log-transformed views. Together, the quantitative tests and visual data exploration provide consistent evidence that more concise video titles relate to higher viewership based on this dataset.

Hypothesis 2 Analysis:

Test Statistic:
The t-test yielded a test statistic value of ~13.82. This value indicates that videos in the population under the "Gaming" category have a higher mean of log-transformed views compared to that of the "Entertainment" category.

p-value:
The p-value from the statistical test is 7.062×10^{-43}. Being substantially smaller than the 0.05 significance level, it provides enough evidence to reject the null hypothesis. Consequently, we reject the null hypothesis, suggesting that videos under the "Gaming" category have a mean log-transformed value of views not equal to that of the "Entertainment" category.

Figure 10

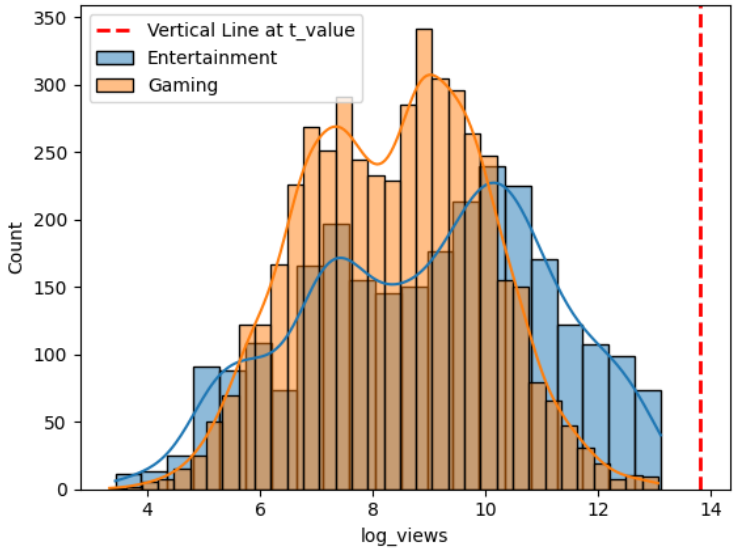


Figure 10 visualizes the log-transformed views plotted against the density plot of the categories, with the red line representing the test statistic value. The test statistic value, being distant from the distribution, provides strong evidence to reject the null hypothesis. From the test statistic, it is evident that videos under the "Gaming" category obtained a higher mean log-transformed value of views compared to that of "Entertainment". This implies that videos under the "Gaming" category also obtained a higher mean view compared to that of "Entertainment".

Hypothesis 3 analysis

Test Statistic:
The t-test yielded a test statistic value of ~10.50. This value indicates that videos in the population under the "How-to & Style" category have a higher mean of log-transformed views compared to that of the "People & Blogs" category.

p-value:
The p-value from the statistical test is 1.329×10^{-25}. Being substantially smaller than the 0.05 significance level, it provides enough evidence to reject the null hypothesis. Consequently, we reject the null hypothesis, suggesting that videos under the "How-to & Style" category have a mean log-transformed value of views not equal to that of the "People & Blogs" category.

Figure 11

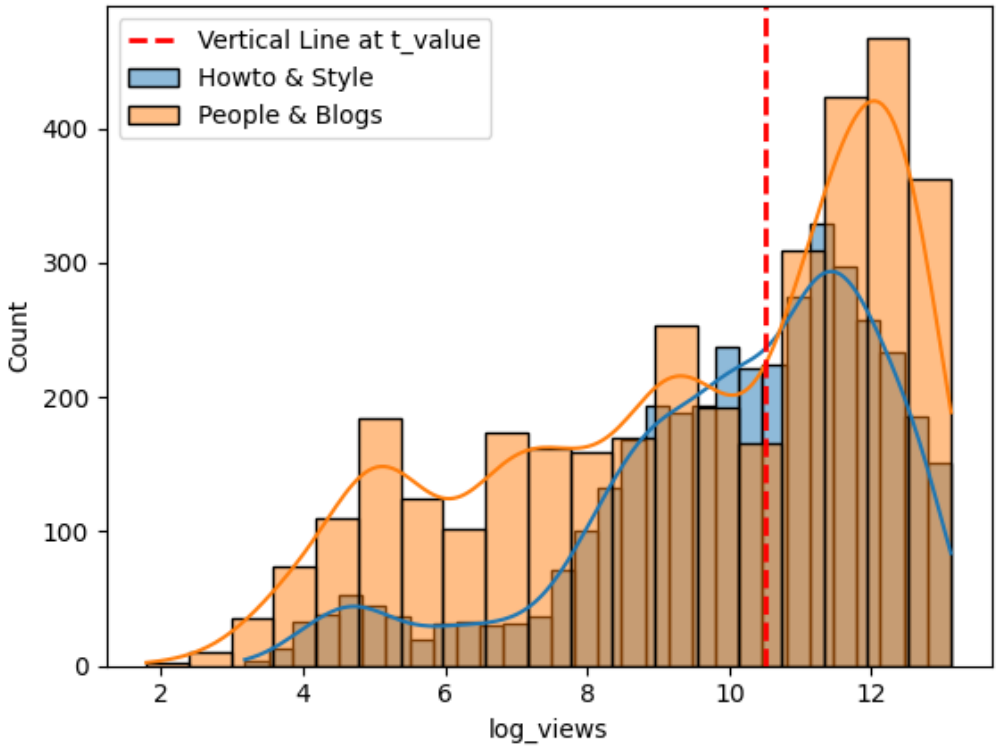


Figure 11 visualizes the distribution of both categories, accompanied by a red line representing the t-value of the t-test. The test statistic indicates that videos in the "How-to & Style" category achieved a higher mean log-transformed value of views compared to those in "People & Blogs". Similarly, videos in the "How-to & Style" category also obtained a higher mean view count compared to those in the "People & Blogs" category, as log transformation of views was applied to both categories.

5 Conclusions

For content creators on YouTube, choosing the right title makes all the difference when it comes to gaining views. While crafting the title for a video, creators should be as concise as possible with their word choice and ensure they are using keywords that fit the intention of the video without unintentionally categorizing their video with lower visibility tags.

One possible reason for wordy titles receiving less views is that the title gets cut off in the user interface and potential viewers cannot fully grasp why the video might be one they want to engage. Because this paper did not test for optimal length, this takeaway is merely directional and further research would need to be conducted to understand the ideal character length. Additionally, the research highlighted that certain algorithmic tags almost guarantee more views. For example, if the content of the video borders between two similar tags like “People & Blogs” and “How-to & Style”, it would be crucial for the video to be tagged as “How-to & Style” to garner more engagement. However, predicting what tag could be applied to a video was beyond the scope of this paper, and would require significant research and testing.

With all that said, YouTube creators should strategically consider both title length and category to enhance the visibility and engagement of their content.

6 References

Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. https://plot.ly.

Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**. *Nature Methods*, 17(3), 261-272.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).