M00361014\_Kaushal\_Guragain\_CIS663\_Section01\_Project\_Spring2024

Kaushal Guragain

2024-05-07

By: Kaushal Guragain To: Dr. Vlad Krotov

ID: M00361014 Course Instructor

Table of Contents

[Introduction 2](#_Toc166357054)

[Literature review 2](#_Toc166357055)

[Research Question 3](#_Toc166357056)

[Theory 4](#_Toc166357057)

[Data 4](#_Toc166357058)

[Data Analysis and Results 6](#_Toc166357059)

[Implications 8](#_Toc166357060)

[Conclusion 8](#_Toc166357061)

# Introduction

YouTube, with its vast user base and extensive reach, stands as a cornerstone of the social media landscape. Boasting over 2.49 billion monthly active users and more than 80 million paid subscribers, it presents an unparalleled platform for content creators and businesses alike to showcase their offerings to a global audience (Backlinko.com, [insert year]). However, amidst this abundance of opportunity lies a fundamental challenge: how to garner the attention and engagement necessary to drive views, interactions, and ultimately, revenue.

At the heart of this challenge lies the question: what factors influence video view counts on YouTube? In pursuit of an answer, this project delves into the rich tapestry of YouTube trending data, seeking to unravel the impact of tags – those concise descriptors appended to videos – on view counts. By analyzing a dataset encompassing trending YouTube videos, this study endeavors to illuminate the relationship between the content descriptors encapsulated in tags and the resulting viewer engagement, as measured by view counts.

# Literature review

I read the paper Elango, Dinesh, Social Media Video Creators Monetization and Business on YouTube by Dinesh Elango. In the paper he talks about how advertisers will select the video for ads marketing and how the earning is distributed to the video creators. He also talks about possible conversation rates of views into revenue. Another interesting paper I came across was, Modelling and statistical analysis of YouTube’s educational videos: A Channel’s owners’ perspective. Some on the things that the paper talked about was, using moving average to analyze the trend of view per day for the channel, the correlation between video uploading activity, the age of channel and its popularity, classifying videos based on metrices like number of comments, subscribers, shares and likes. Zipf distribution and pareto law, devices used to watch videos, and more. All of the analysis that have been done with above matrices for content-based analysis and finding user behavioral patters in response to the videos they have watched or the playlist they have viewed. After reading the paper the things that are known about the problem are: YouTube’s massive user base, importance of views for revenue, challenges in generating revenue, lack of concrete solution for predicting the views on the video. Some not known things are, an accurate statistical tool for predicting views, Effective strategies for video making, strategic sponsorship of channels based on category and location.A very interesting paper i came across was of “Hashtag recommendation for enhancing the popularity of social media posts”. Their proposed method utilizes the trending nature of hashtags by using post keywords along with the popularity of users and posts. They were able to devise a novel evaluation algorithm that is more robust and reliable than what is in the market for such analysis.

# Research Question

Does the frequency of tags or tags in general affect the overall viewcount of youtube trending videos.

The null hypothesis was that the tags did not affect the viewcount for the trending videos. However, we were able to devise that tags had effected the viewcount of videos by 16% according to the regression model. The initial model devised and presented gave a result of 18%, but removing weights and doing the regression generated the model r-squared and adjusted r-squared to “0.16”, and “0.15”.

## # A tibble: 16,404 × 5  
## video\_id title categoryId view\_count tags   
## <chr> <chr> <dbl> <dbl> <chr>  
## 1 J5N3JVCOipo Batman gets Riddled but it's Animated 1 338968 [Non…  
## 2 6ipksYvlaqA COLLEGE DECISIONS REACTIONS VLOG + a… 26 1212966 [Non…  
## 3 c9yc8X16w8c Worst Tik Tok Trend 24 2302190 [Non…  
## 4 AyexgXyBTl4 Making rolled ice cream for a regula… 24 2290002 [Non…  
## 5 KrLj6nc516A $1 Vs $100,000,000 Car! 24 154717067 [Non…  
## 6 tWYsfOSY9vY I Survived 7 Days In An Abandoned Ci… 24 111897842 [Non…  
## 7 oSBh6a0QJg8 Morgan Wallen - You Proof (Lyric Vid… 22 1065144 [Non…  
## 8 JZzl1UlVwFs World Record Packing Slimes Attempt 26 1141369 [Non…  
## 9 JbKhPDQH5hA Green Bay Packers vs. Philadelphia E… 17 1958366 [Non…  
## 10 SGQPEavFew0 EL AMOR DE TU VIDA (Video Oficial) -… 10 1374402 [Non…  
## # ℹ 16,394 more rows

## `summarise()` has grouped output by 'tags'. You can override using the  
## `.groups` argument.

##   
## Call:  
## lm(formula = view\_count ~ . - 1, data = train\_data[, c("view\_count",   
## independent\_variable\_after\_ANOVA)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1888759 -945676 85320 1056022 87106591   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## tagschallenge 939301 137101 6.851 7.7e-12 \*\*\*  
## tagscomedy 1498107 120642 12.418 < 2e-16 \*\*\*  
## tagsfortnite 1314729 144155 9.120 < 2e-16 \*\*\*  
## tagsfunny 1969393 53429 36.860 < 2e-16 \*\*\*  
## tagsnews 1325020 103281 12.829 < 2e-16 \*\*\*  
## tagsvlog 979419 118957 8.233 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3350000 on 11389 degrees of freedom  
## Multiple R-squared: 0.1477, Adjusted R-squared: 0.1473   
## F-statistic: 329.1 on 6 and 11389 DF, p-value: < 2.2e-16

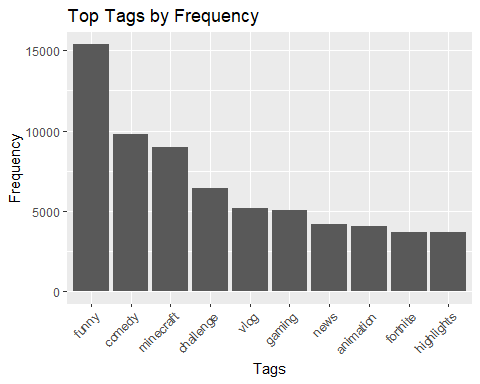
# Theory

My initial theory from literature review is that tags do affect the view count in some form. While predicting the view count of a particular YouTube video doesn’t have a simple solution. But, tags will in fact affect the video to some extent. Since, the data is for data of trending YouTube videos, even a small prediction as 16% is significant because we are found out what tags out of trending YouTube videos affected their view count. There are millions of videos which do not generate even a single view. Tags, which was able to predict as much as it could is definitely significant.

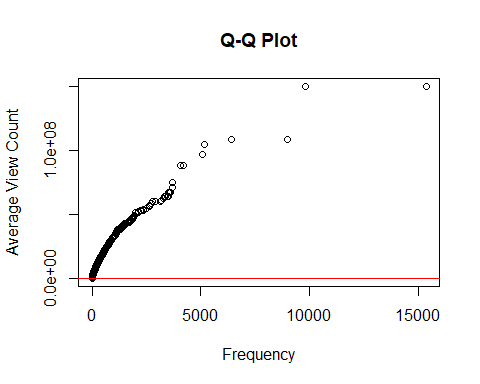
# Data

The data I require for my research project needed to encompass a high volume of data with numerous metrics involved, as discussed previously. These metrics included variables such as ‘tags’ for comparing different categoris of data, ‘channel title,’ ‘channel category,’ ‘view count,’ ‘likes,’ and others. I have identified relevant data sources for my project on Kaggle. One such dataset is available at <https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/> , which provides column variables like ‘published year’ for historical data, along with various other metrics.

The data I utilized was focused on tags and view count. Specially the top 10 tags out of all the data sets.The data was in excel (xlsx) format. The data was clean in various aspect and divided based on top 10 tags before we used a regreeion model to analyze it.



# Data Analysis and Results

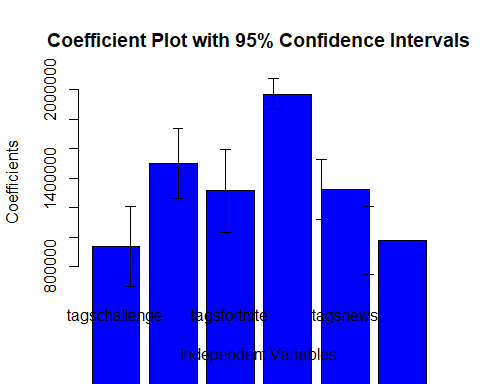
I analyzed my data in many ways. I used a QQ plot to see the distribution.It showed that the view count increases as the frequency of tags increases as well.  I also used the ANOVA test and Regression model to see the variance in the data and filtering independed variables for regression model.

## Analysis of Variance Table  
##   
## Response: view\_count  
## Df Sum Sq Mean Sq F value Pr(>F)   
## tagsanimation 1 1.7637e+13 1.7637e+13 1.9690 0.160570   
## tagschallenge 1 1.4705e+13 1.4705e+13 1.6417 0.200109   
## tagscomedy 1 7.5731e+13 7.5731e+13 8.4547 0.003646 \*\*   
## tagsfortnite 1 1.6753e+14 1.6753e+14 18.7030 1.536e-05 \*\*\*  
## tagsfunny 1 2.6276e+14 2.6276e+14 29.3351 6.175e-08 \*\*\*  
## tagsgaming 1 2.8430e+12 2.8430e+12 0.3174 0.573183   
## tagsminecraft 1 3.5852e+10 3.5852e+10 0.0040 0.949555   
## tagsnews 1 3.4119e+14 3.4119e+14 38.0908 6.913e-10 \*\*\*  
## tagsvlog 1 1.9255e+15 1.9255e+15 214.9654 < 2.2e-16 \*\*\*  
## Residuals 16269 1.4572e+17 8.9572e+12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

After looking at the results, i removed the independent variables with low F-value and high Pr(>0.5). Then the regression was done and the final result was as follows:

##   
## Call:  
## lm(formula = view\_count ~ . - 1, data = train\_data[, c("view\_count",   
## independent\_variable\_after\_ANOVA)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1888759 -945676 85320 1056022 87106591   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## tagschallenge 939301 137101 6.851 7.7e-12 \*\*\*  
## tagscomedy 1498107 120642 12.418 < 2e-16 \*\*\*  
## tagsfortnite 1314729 144155 9.120 < 2e-16 \*\*\*  
## tagsfunny 1969393 53429 36.860 < 2e-16 \*\*\*  
## tagsnews 1325020 103281 12.829 < 2e-16 \*\*\*  
## tagsvlog 979419 118957 8.233 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3350000 on 11389 degrees of freedom  
## Multiple R-squared: 0.1477, Adjusted R-squared: 0.1473   
## F-statistic: 329.1 on 6 and 11389 DF, p-value: < 2.2e-16

We can see that, the pr value is small for all the independent variables and good t values. The multiple R-squared and adjusted R-squared was 16% and 15% respectively. I also created a bar plot which will produce blue bars for value > 0 showing some degree of significance for prediction for view count by tags and red for not showing any. I got the output as blue bars which mean that the independed variables were significant in predicting the average view count.



# Implications

The results imply that, while tags are not the best way to estimate how the view count fluctuates, but it is still a significant factor. While 16% isn’t much, the percent is for the data of trending videos. So, even being able to predict 16% on such data holds some level of significane.

# Conclusion

The insights gained from this project hold significant implications. Equipped with a statistical understanding of the tags that drive higher view counts, creators, businesses, and marketers can optimize their video content strategy. They can discern the categories, types, and lengths of videos most likely to resonate with their target audience. This enables them to tailor their content to maximize engagement and achieve desired outcomes, whether it’s increased brand exposure, product promotion, or revenue generation.

In essence, this project aims to empower YouTube creators and businesses with actionable insights. It guides them towards creating customer-oriented videos that captivate audiences and drive meaningful outcomes. By embracing data-driven decision-making, stakeholders can unlock the full potential of YouTube as a marketing platform. This fosters greater connection, engagement, and success in the dynamic realm of online video content.