Introduction

Youtube creators are always looking to maximize views and increase their engagement. However, the relationship between tags, engagement and views remains blurry. Therefore, in this data analysis, we will be looking at variables such as 'views' , 'likes' , 'engagment' etc, to figure out, what exactly drives the top performing videos.

Data exploratory plan

The data analysis will be divided into four stages:

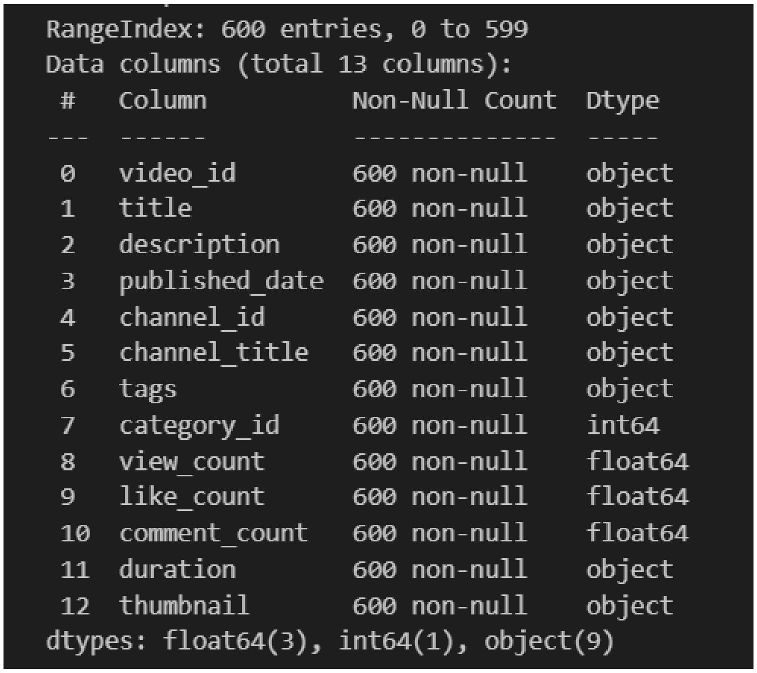
1) **Data Reading/Inspection:**
   - I will be checking the data for any null/duplicate values. Mean/median/mode and skewness will be measured. Furthermore, I will be applying log transformation to fix the skewness and representing it through boxplots.

2) **Feature Engineering**
   - Conversion of ISO 8601 format of date and duration into datetime object and seconds, respectively. This will be done to enhance readability.
   - Engagement metrics such as engagement(like+comments/view_count), engagement velocity(likes+comments/days_since_publish), columns will be made to further analyse the factors effecting view count.
   - Tag count column, to count the number of tags. Later, we will examine how it effects a video's performance.

3) **Exploratory Visual Analysis**
   - Bubble plot with four variables `engagment`, 'view_count' , 'tag_count' and 'duration'
   - Plot of 'engagement velocity' and 'days_since_published', this would help us figure how quickly do videos gain engagement.
   - Line plot of engagement and views
   - Heatmap for key video metrics

4) **Hypothesis and their testing**
   - Does category of duration effects views?
   - Do tag count effect views?

Summary of dateset

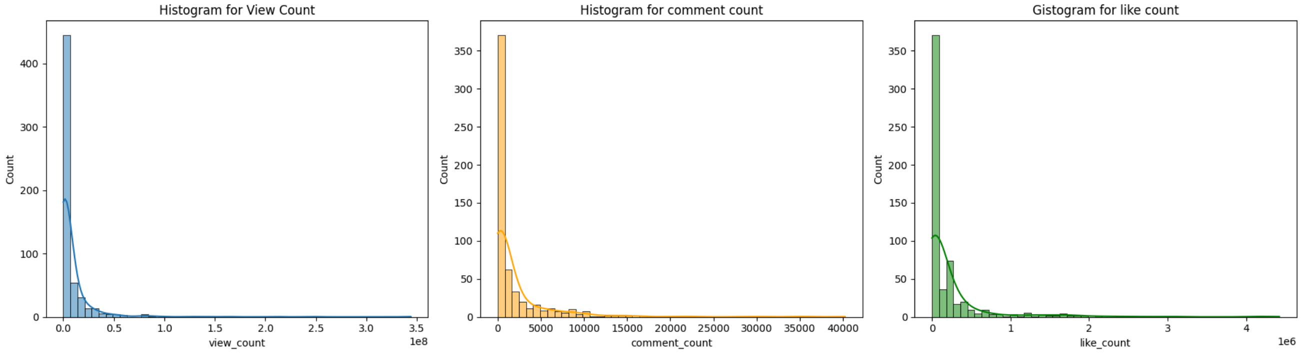


```
RangeIndex: 600 entries, 0 to 599
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   video_id        600 non-null    object 
 1   title           600 non-null    object 
 2   description     600 non-null    object 
 3   published_date  600 non-null    object 
 4   channel_id      600 non-null    object 
 5   channel_title   600 non-null    object 
 6   tags            600 non-null    object 
 7   category_id     600 non-null    int64  
 8   view_count      600 non-null    float64
 9   like_count      600 non-null    float64
 10  comment_count   600 non-null    float64
 11  duration        600 non-null    object 
 12  thumbnail       600 non-null    object 
dtypes: float64(3), int64(1), object(9)
```

Exploratory Visual Insights

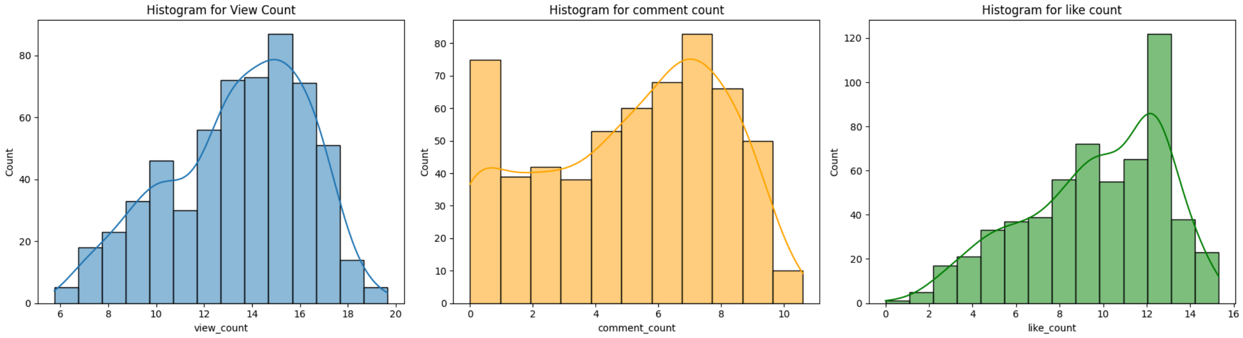
Histrogram of view count, comment count and like count

Before log transformation:

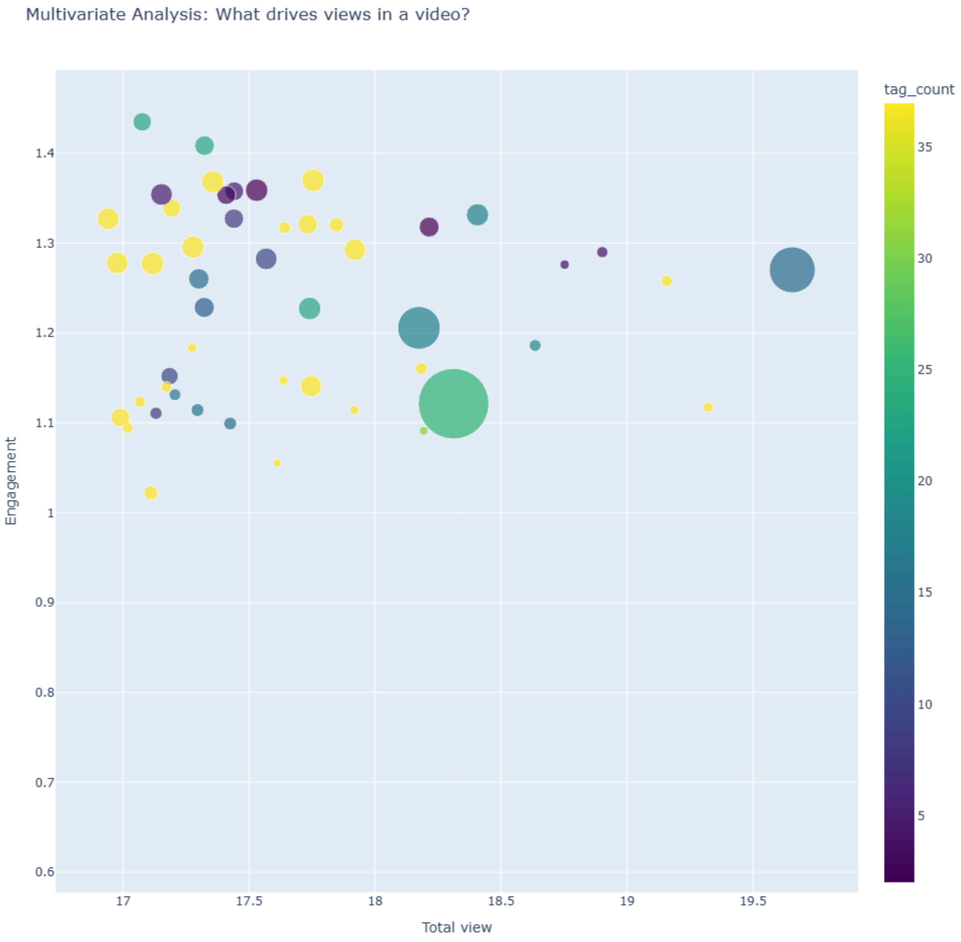


Heavily right skewed

After Log transformation:



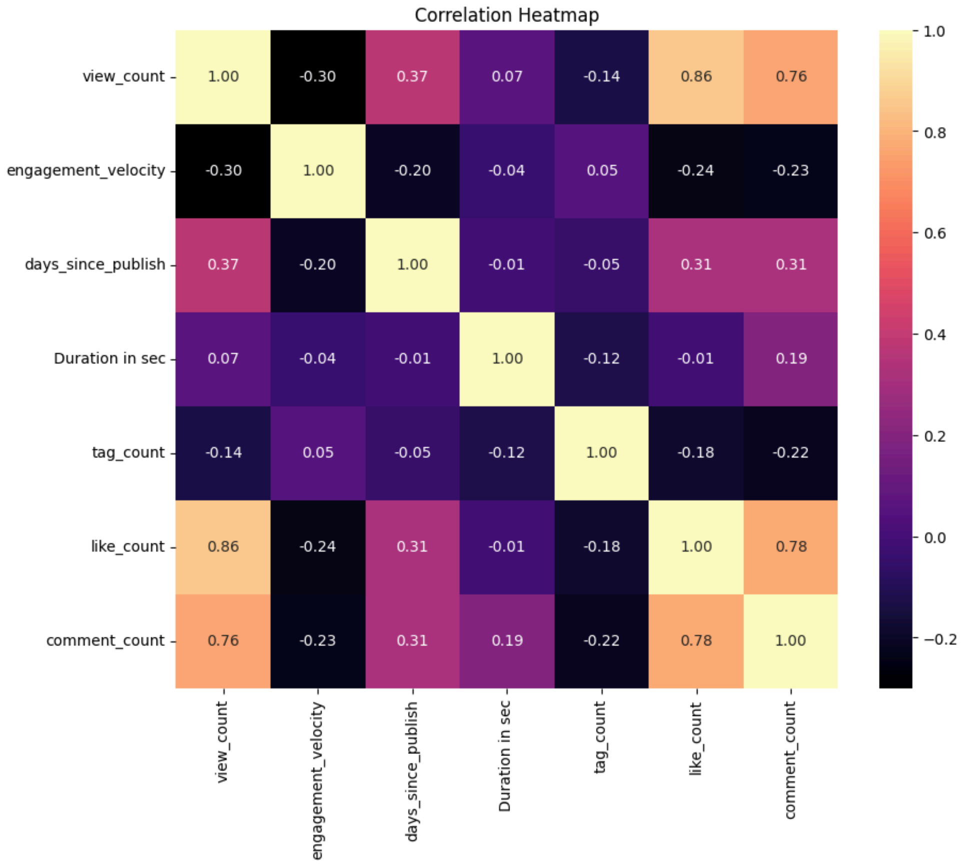
Identifying Factors affecting view count - Bubble plot:



- Some high viewed videos show similar engagement as mid viewed videos.

- Low engagement, high views (1.1): Likely boosted by external traffic (ads/shares).
- High engagement, fewer views (1.3+): Niche, highly engaged audience.
- Larger bubbles don't consistently align with high views or high engagement.Both small (short) and large (long) bubbles appear across view counts, reinforcing that duration doesn’t dictate virality.

Heat map for key video metrics



- Likes and views are highly correlated (r=0.86), same goes for comments and views (r=0.78), suggesting engaged viewers do both

.

- Likes and enagement velocity show negative correlation, means a trending video does not gurantee many likes.
- Viiews and duration have near zero correlation(0.07), debunking 'longer videos = more views' as a universal rule.

- Tag count and view counts have a negative correlation.

Hypothesis Testing:

1. First hypothesis

H0: Duration category has no effect on view count
H1: Duration category has effect on view count

I used Kruskal-Wallis for the following reasons:

- It is a non-parametric alternative to ANOVA
- Used when comparing two or more independent groups (categorical) against a numeric variable
- It does not assume normality or equal variances

Result and Interpretation:

The p-value: 0.0767, therefore we fail to reject the null hypothesis. This suggests that there is no statistically significant difference in view counts across different video duration categories.

2. The second hypothesis

H0 : There is no significant correlation between the number of tags and view count.
H1 : There is a significant correlation between the number of tags and view count

I used Spearman's test for two main reasons:

- Non-Linear Relationship:

Even after log-transforming view_count, the relationship with tag_count might not be perfectly straight/linear.

- Better for Count Data:

Tag_count is a whole number (like 5, 10, 15 tags), and Spearman works best for discrete/ranked data.

Result and Interpretation:

The results of spearman test are:
Correlation: -0.11368145819753249, P Value: 0.005954593666120997

Therefore Null hypothesis is rejected. This indicates a statistically significant correlation between tag count and view count however the correlation is weak and negative as (Spearman's ρ = -0.113,)

EDA Insights and Limitations

Insights

- Higher engagement (likes/comments) correlates with higher viewership.

- Kruskal-Wallis test indicated no statistically significant difference in view counts across duration categories (Short, Medium, Long). However, short videos (<60s) dominated the dataset, possibly due to platform trends favoring concise content.

- Tag Count Influence: Spearman test showed weak correlation between tag count and view count, indicating tags may not significantly impact reach.

- Engagement Velocity Trends: Highest engagement occurs within the first few days after publishing, followed by a decline.

- High-view videos did not always have the highest engagement, suggesting algorithmic promotion may play a bigger role in virality than raw engagement metrics.

Limitations

- The analysis is based on only 600 videos, while YouTube hosts billions of videos across diverse categories.
- Findings may not generalize to the entire platform due to potential sampling bias.

Conclusion

This analysis reveals that engagement (likes/comments) strongly correlates with viewership, but video duration and tag count have negligible effects. Short-form content dominates the dataset.

Key Takeaways for Content Creators:

- Prioritize engagement: Encouraging likes and comments may boost visibility.
- Optimize early momentum: Maximize engagement in the first few days post-publishing.
- Experiment freely with duration: Neither short nor long videos guarantee virality—optimize for content quality over strict length.
- Avoid tag spam: Use relevant tags, but don't over-invest; algorithmic promotion likely matters more.

Next Steps:

- Expand Dataset – Gather more videos to see bigger trends
- Creator-Specific Factors – Investigate channel size, subscriber base, and posting frequency.
- Predictive Modeling – Build a regression/ML model to predict virality.