Views, Likes, and Stardom: Insights from YouTube Trending Categories

🎯 BUSINESS PROBLEM

YouTube engagement metrics are heavily influenced by a small group of star creators, which makes it difficult to determine whether a category's high performance reflects genuine audience interest or is simply driven by a few popular channels. By analyzing engagement both with and without the impact of star bias, we aim to identify which categories truly sustain audience engagement and which depend primarily on star power. These insights are critical for platforms, advertisers, and emerging creators to make informed and strategic decisions.

🔍 OBJECTIVES

The goal of this analysis is to evaluate YouTube engagement across categories by comparing metrics with and without the influence of star creators. Specifically, we aim to:

- Identify categories where engagement is driven by genuine audience interest versus those dominated by a few star creators.
- Provide actionable insights for advertisers to allocate budgets more effectively across categories.
- Guide new creators toward categories where they can grow organically without being overshadowed by established stars.

📚 DATASET

This project is based on Rishav Sharma's YouTube Trending Dataset, a widely used dataset that captures daily records of videos appearing on YouTube's trending page across multiple countries

Source: [Link to kaggle dataset]

Scope: 200,000+ trending videos collected from 2017 onwards, across india

Granularity: Each row represents a video's performance on a given day when it appeared on the trending list

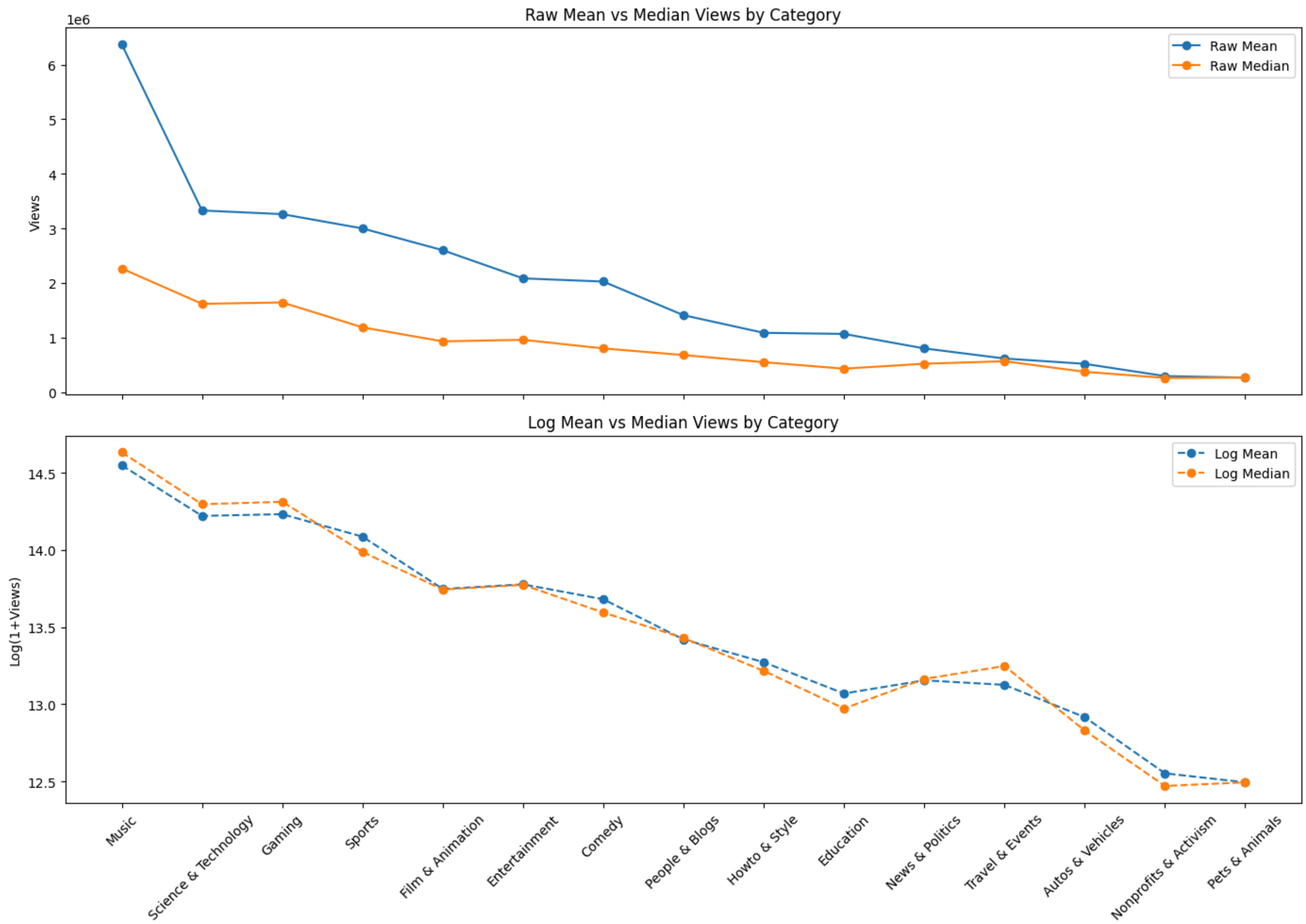
🛠️ TOOLS AND TECHNIQUES USED

- **Python** → Core language for data analysis
- **DuckDB** → SQL-style queries on large-scale datasets within Python
- **Pandas & NumPy** → Data cleaning and wrangling
- **Matplotlib & Seaborn** → Data visualization for category-level insights
- **SciPy & Statsmodels** → Hypothesis testing (ANOVA, Tukey HSD)
- **Google Colab** → Interactive analysis & reproducibility

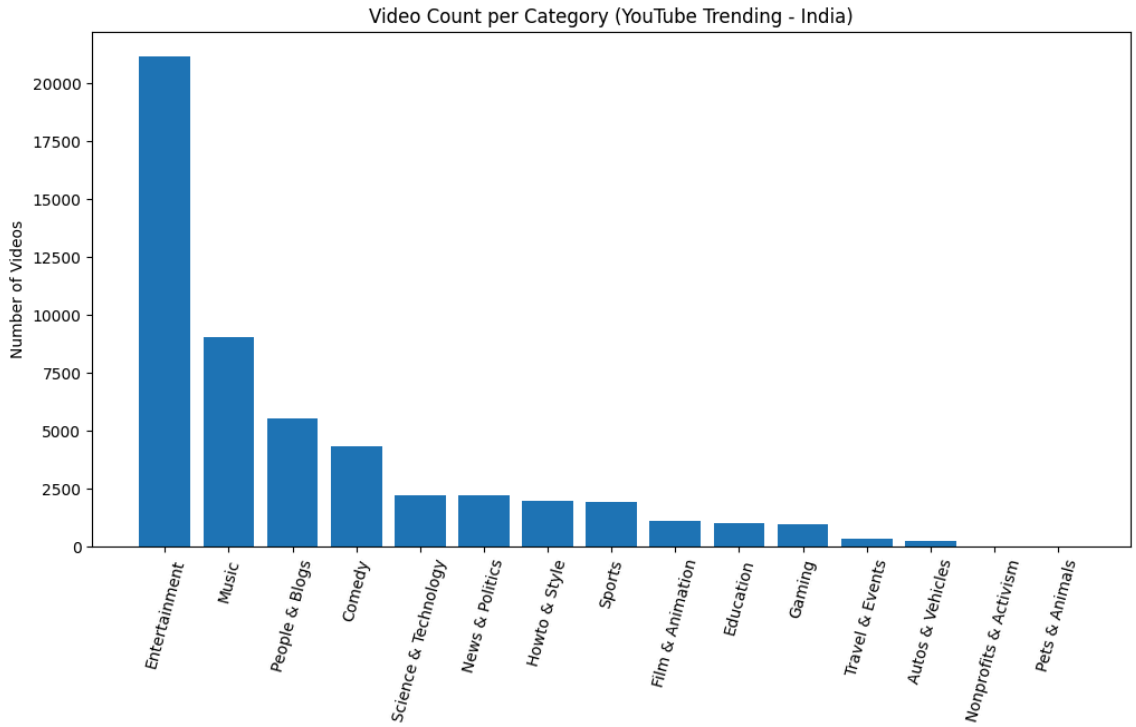
📊 ANALYSIS AND FINDINGS

During the initial analysis, I observed a large deviation between the average and median views across categories. This indicated the presence of extreme outliers (very high-view videos), which made the raw data highly skewed.

To address this, I applied a logarithmic transformation to the view counts. This step reduced the impact of extreme values and allowed for more balanced comparisons across categories.



📋Category Distribution



As shown in the figure above, the dataset is dominated by Entertainment and Music categories, which together account for the majority of trending videos. Categories such as People & Blogs, Comedy, and Science & Technology also contribute significantly, while niches like Nonprofits & Activism and Pets & Animals have very limited representation.

TOTAL VIEWS BY CATEGORY

category_name varchar	total_views int128
Music	57701430840
Comedy	8738342375
News & Politics	1759857817
Film & Animation	2855738209
Education	1069316740
Nonprofits & Activism	4712343
Entertainment	44139446423
Gaming	3102124725
People & Blogs	7819025476
Sports	5813862516
Howto & Style	2156190553
Science & Technology	7360693599
Autos & Vehicles	133894081
Travel & Events	202222182
Pets & Animals	266674
15 rows	2 columns

YouTube in India is still heavily dominated by music and entertainment content, while other genres like education, gaming, and news are growing steadily

Statistical Test: ANOVA on Views Across Categories

To test whether certain categories consistently attract significantly higher views (i.e., the stardom effect), I performed a one-way ANOVA across all content categories.

- **Null Hypothesis (H_0):** All categories have the same mean view count.
- **Alternative Hypothesis (H_1):** At least one category has a significantly different mean view count.

Results:

- F-statistic: 231.66
- p-value: 0.0 ($p < 0.05$)

Since the p-value is extremely small, we reject the null hypothesis.

💬This means there is strong statistical evidence that some categories attract disproportionately higher views compared to others

Post-Hoc Analysis: Tukey's HSD Test

Since the ANOVA showed strong evidence that not all categories have equal view counts, I conducted a Tukey's HSD (Honestly Significant Difference) test to determine which specific categories differ significantly.

- **Null Hypothesis (H_0):** The mean views between two categories are equal.
- **Alternative Hypothesis (H_1):** The mean views between two categories are significantly different.

Key Findings:

- Music has significantly higher views than almost every other category.
- Entertainment, Film and animation & Gaming also record higher views compared to Education, News & Politics, Travel, and Nonprofits.
- This confirms that viewership is highly concentrated in star-driven categories such as Music, Entertainment, Film & Animation, and Gaming, rather than being evenly spread across all categories.

Stardom Concentration Analysis

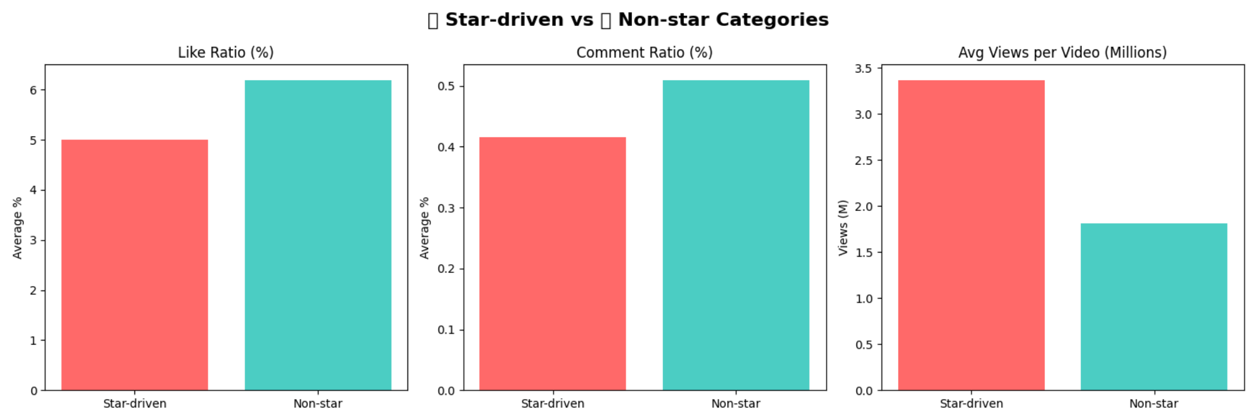
To check whether a handful of categories dominate YouTube, we defined star-driven categories: Music, Entertainment, Gaming, and Film & Animation.

Results show that these four categories account for:

- 75.46% of total views
- 73.51% of total engagement (likes + comments)

This clearly indicates that the majority of YouTube attention and interaction is concentrated in a few star categories, while the remaining categories share less than 25% of the views and engagement.

Views vs Engagement: Star vs Non-star Categories



While non-star categories achieve better engagement ratios, their reach (views) is limited.

Star-driven categories capture the majority of visibility and attention, even if engagement per viewer is slightly lower.

Together with the earlier finding (75%+ of views and engagement share), this strengthens the case that YouTube stardom is heavily concentrated in a few categories.

📌 Final Conclusions

- Stardom dominates trending content – Music, Entertainment, Gaming, and Film & Animation make up ~75% of total views among trending videos.
- ANOVA shows strong category-level differences – Music trending videos attract significantly more views than almost every other category, with Entertainment, Gaming, and Film & Animation also outperforming Education, News & Politics, Travel, and Nonprofits.
- Engagement patterns differ – While star-driven trending videos get the highest visibility, Non-star trending videos (Education, News & Politics, How-to, Nonprofits) show higher like and comment ratios, suggesting stronger viewer interaction relative to views.
- Reach vs. depth trade-off – On YouTube’s trending list, star-driven categories achieve mass reach, whereas non-star categories achieve deeper engagement despite lower average views.

🕵️Engagement Index by Category

category_name varchar	engagement_index double
Travel & Events	0.0940191269422659
Comedy	0.0911672609989718
Pets & Animals	0.08591013747121953
Gaming	0.07727131280964211
Music	0.061851219788573275
People & Blogs	0.05792879347820148
Science & Technology	0.0539221654402137
Education	0.05215157297546843
Howto & Style	0.04432868276276137
Autos & Vehicles	0.04302031842617449
Entertainment	0.0407636166017351
Nonprofits & Activism	0.03677045580086169
Film & Animation	0.03459135353817721
Sports	0.032266545430638456
News & Politics	0.02590430292699038
15 rows	2 columns

Engagement Index=(Total Likes/Total Views)+(Total Comments/Total Views)

- Likes / Views → captures how often viewers show positive appreciation.
- Comments / Views → captures how often viewers actively engage in discussion.
- Combining both gives a normalized measure of *audience interaction per view*.