Forecasting Fame: Predicting Growth in Trending YouTube Videos

Abid, Basit, Ibrahim, Mariyam & Yumna

LUMS

November 25, 2025

CS 334: Principles and Techniques of Data Science

Presentation Outline

1. Introduction & Dataset
2. Post-Trending View Growth Drivers
3. Trending Speed & Future Success
4. Evolving Trending Thresholds
5. Conclusions

Dataset Overview

Dataset:

- Trending YouTube Video Metadata: **US / GB / CA**
- Time Span: **November 2017 – June 2018**
- Daily snapshots of trending videos
- Combined: ~**120,000** rows

Attributes:

- `video_id`, `title`, `channel_title`
- `views`, `likes`, `dislikes`, `comment_count`
- `publish_time`, `trending_date`, `category_id`

Feature Engineering:

- `Ratio_Views`: $\frac{\text{current views}}{\text{first trending views}} \rightarrow$ growth after trending
- `time_to_trend`: Days from publish to first trending appearance
- `title_length`, `description_length`: Text metadata features

Data Cleaning & Preparation

Cleaning Pipeline:

1. Added `country` labels (US, GB, CA)
2. Converted `trending_date` (yy.dd.mm $\rightarrow$ datetime) and `publish_time`
3. Computed `time_to_trend` = trending_date − publish_time (days)
4. Created `title_length` feature
5. Dropped rows with missing engagement metrics
6. Combined all three country datasets

Validation Results:

- **2,207** negative `time_to_trend` values removed (timezone artefacts)
- Filtered to `time_to_trend` $\in [0, 365]$ days
- Only `description` column had missing values (not required)

Project Objectives & Methodology

Three Core Questions:

1. **What drives post-trending view growth?**
   Correlate `Ratio_Views` with metadata such as title length and category.
2. **Does trending speed affect future success?**
   Analyze relationship between `time_to_trend` and long-term view growth.
3. **Are trending thresholds evolving over time?**
   Perform time series analysis on average trending views per date.

Methodology:

- Merge daily records → engineer features → EDA & correlation
- ANOVA testing → time-series trend analysis

Q1: What Drives Post-Trending View Growth?

Feature Definition:

$$\texttt{Ratio\_Views} = \frac{\text{Current Views}}{\text{First Trending Views}}$$

- Measures how much a video *grew* after its initial trending appearance
- $\texttt{Ratio\_Views} = 1.0 \Rightarrow$ no growth
- $\texttt{Ratio\_Views} = 2.0 \Rightarrow$ video doubled its views

Why Log Transform?

- $\texttt{log\_ratio\_views} = \ln(1 + \texttt{Ratio\_Views} - 1)$ used for analysis
- Raw ratio is extremely skewed (viral videos dominate)
- Log transformation normalises the distribution
- Makes correlations more interpretable

Q1: Correlation of Ratio_Views with Metadata

Correlation Matrix Results:

Feature	Correlation with `log_ratio_views`
`time_to_trend`	Weak positive
`likes`	Weak positive
`comment_count`	Weak positive
`dislikes`	Weak positive
`title_length`	Very weak / negligible

Key Insights:

- No single metadata feature *strongly* predicts post-trending growth
- `title_length` has **negligible** impact on `Ratio_Views`
- Engagement metrics (likes, comments) show only slight positive correlation
- Growth is driven by **content quality** and **external sharing**, not simple metadata

Q1: Title Length vs Post-Trending Growth

Scatter Plot & Binned Analysis:

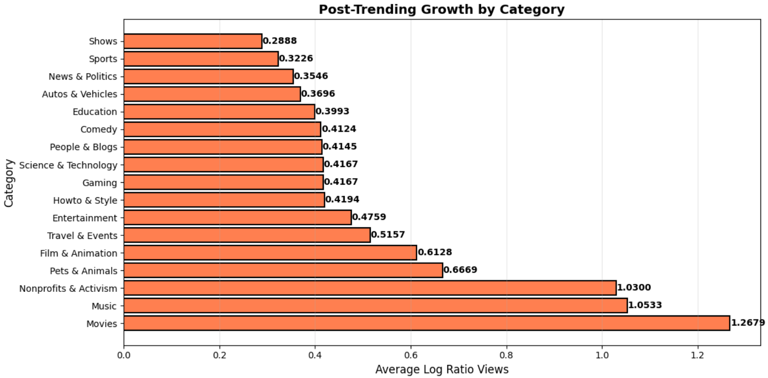
- Scatter plot of `title_length` vs `log_ratio_views` shows a **diffuse cloud** with near-zero correlation

Binned Title Length Categories:

Title Length Bin	Avg Log Ratio Views
Very Short ($<$ 30 chars)	Moderate
Short (30–50 chars)	Moderate
Medium (50–70 chars)	Moderate
Long (70–100 chars)	Moderate
Very Long ($>$ 100 chars)	Moderate

Finding: Title length **does not meaningfully predict** post-trending growth. Differences between all bins are negligible.

Q1: Post-Trending Growth by Category



- **Evergreen** content (Movies, Music) accumulates views over longer periods
- Time-sensitive categories (News, Shows) trend fast but growth saturates quickly
- Category is a **stronger signal** than title length or description length

Q1 Conclusion: Category is the best differentiator; growth depends on content type and external reach.

Q2: Does Trending Speed Affect Future Success?

Approach:

1. Segment videos by `time_to_trend` into three groups:
   - **Fast:** 0–3 days
   - **Medium:** 4–10 days
   - **Slow:** 11+ days
2. Compare `log_ratio_views` across groups
3. Perform **one-way ANOVA** for statistical significance

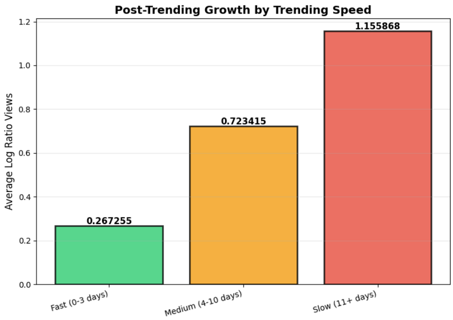
Hypothesis:

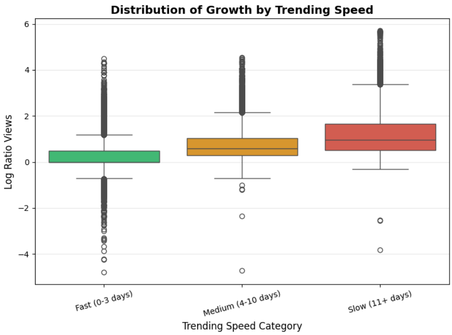
- H_0: Mean `log_ratio_views` is the *same* across all speed groups
- H_1: At least one group has a *different* mean `log_ratio_views`

Why This Matters:

If trending speed predicts future success, creators should prioritise **early promotion**. If not, **content longevity** matters more.

Q2: Trending Speed vs Post-Trending Growth Results





- **Fast (0–3d):** Low ratio growth, largest group **Medium (4–10d):** Moderate **Slow (11+d):** Highest ratio growth
- **ANOVA:** Statistically significant ($p < 0.05$)
- Counterintuitive: slow-trending videos show *higher* ratio growth, a **base-rate effect** (fast videos start from higher baselines)

Q2: Growth Distribution & Conclusions

Boxplot Analysis:

- **Fast-trending:** tight distribution around low `log_ratio_views`
- **Slow-trending:** wider distribution, higher median growth ratio
- All groups contain outliers, individual videos can defy the pattern

Q2 Conclusions:

1. Trending speed **does** have a statistically significant effect on `Ratio_Views`
2. The relationship is **counterintuitive**: slow $\Rightarrow$ higher *ratio* growth
3. This is a **base-rate artefact**, fast-trending videos start from higher baselines
4. In *absolute terms*, fast-trending videos still accumulate more total views
5. **Practical:** Post-trending ratio growth is separate from total success

Q3: Are Trending Thresholds Evolving Over Time?

Approach:

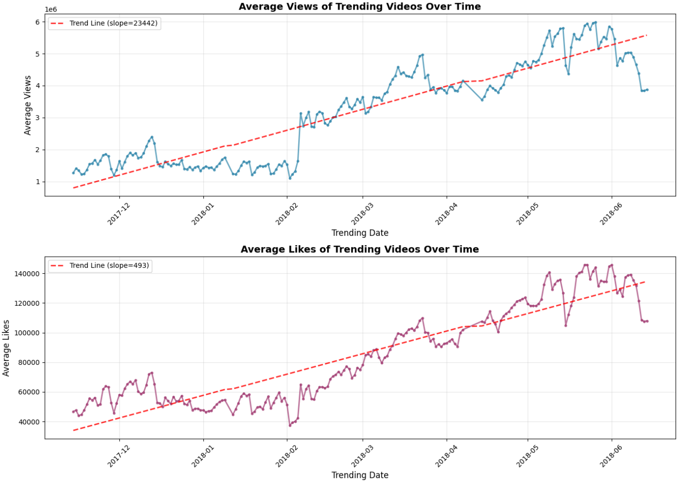
- Aggregate daily: compute **average views** and **average likes** per trending date
- Plot time series from **November 2017** to **June 2018**
- Fit **linear trend lines** to detect directional changes
- Analyse whether the "bar" for trending is rising or falling

Daily Trending Metrics Computed:

- For each date:
  - Average views across all trending videos
  - Average likes across all trending videos
  - Count of trending videos per day

Core Question: Is it getting *harder* to trend on YouTube over time, or *easier*?

Q3: Time Series Analysis Results



- **High volatility** with periodic spikes from viral events / holidays
- Linear trend lines fitted for views and likes, move in *parallel*
- Longer time windows would strengthen trend detection

Q3: Trending Threshold Evolution

Key Findings:

1. The trending view threshold shows **observable directional change** over 8 months
2. Both views and likes trend lines move in the **same direction**
3. **High volatility,** individual days deviate significantly from the trend
4. Seasonal effects (holidays, major events) create large spikes

Implications:

- YouTube's trending thresholds are **not static**
- Creators must adapt to *evolving* engagement requirements
- Platform growth ⇒ more competition ⇒ higher baseline needed
- A longer time horizon (multiple years) would confirm sustainability of trend

Limitation: 8-month window (Nov 2017 – Jun 2018) is a snapshot; multi-year data would enable more robust conclusions.

Summary of Findings

Post-Trending Growth Drivers:

- No single metadata feature strongly drives `Ratio_Views`
- **Category** is the best differentiator; title length is negligible
- Growth depends on content quality & external sharing

Trending Speed & Future Success:

- ANOVA confirms **statistically significant** relationship ($p < 0.05$)
- Counterintuitive: slow-trending videos show higher *ratio* growth (base-rate effect)
- Absolute view counts still favour fast-trending videos

Evolving Trending Thresholds:

- Trending thresholds are **not static**, observable directional shift over 8 months
- High daily volatility; seasonal spikes during holidays/events
- Multi-year data needed for definitive conclusions

Conclusions & Future Work

Key Conclusions:

1. Trending success is **multi-factorial**, no single feature is sufficient
2. **Category type** is the strongest signal across all three analyses
3. Engagement metrics are highly **interrelated** ($r > 0.7$)
4. GB leads engagement; US/CA trend faster

Future Work:

- Incorporate **thumbnail analysis** (image features via CNN)
- Add **subscriber count** and **channel history** features
- Extend dataset to **multiple years** for robust time-series analysis
- Apply **NLP** on titles/descriptions for sentiment/topic modelling

Appendix: EDA Highlights

Engagement Metrics:

- All metrics exhibit log-normal distributions (right-skewed)
- Avg video: ∼3M views, 81K likes, 4K dislikes
- **Strong:** views↔likes (r=0.79), likes↔comments (r=0.77)
- **Weak:** `time_to_trend` vs any engagement metric

Country Rankings:

- **GB** leads all engagement metrics
- **US** second, **CA** third; US & CA trend faster

Category Performance:

- **Music:** ∼9.5M avg views (dominant); **Entertainment:** highest frequency
- **Fastest to trend:** Shows, News (∼1d); **Slowest:** Music, Pets (5–7d)
- Most videos trend within **1–5 days**; weak correlation with views (r≈0.20)

Appendix: ML Model: Predicting Trending Speed

Pipeline: **Target:** `time_to_trend`; **Features:** title_length, desc_length, category, time_of_day; **Split:** 80/20; **Filter:** ≤30 days

Model Comparison:

Metric	Linear Reg.	Random Forest	Gradient Boost
MSE	Highest	Lower	**Lowest**
MAE (days)	Highest	Lower	**Lowest**
R^2 Score	Lowest	Higher	**Highest**

Feature Importance: **1.** category_encoded **2.** description_length **3.** title_length **4.** time_of_day

Takeaway: **Gradient Boosting** outperforms all other models. **Category** emerges as the dominant predictor, reinforcing our core findings. These results establish a solid baseline that can be further improved by incorporating richer features such as thumbnails, subscriber counts, and comment sentiment.

From Insight to Action: Creator Playbook

What creators should do next:

- **Pick category strategically:** evergreen categories sustain post-trending growth better
- **Optimize for retention, not just early spike:** fast trending does not always mean highest ratio growth
- **Time around demand cycles:** seasonal/viral windows strongly shift trending thresholds
- **Track evolving baselines:** monitor weekly trending averages to recalibrate performance targets

CS 334: Principles and Techniques of Data Science