Big Data Analytics — Assignment 1

YouTube Trending Videos Analysis Using Graph Database

Students: Hassan Rais - Muhammad Adeel
Registration No.: 2022212 - 2022331

Course: Big Data Analytics
Assignment 1
Date: November 2, 2025

Table of Contents

1. Executive Summary
2. Introduction
3. Dataset Description
4. Methodology & Project Phases
5. Data Preprocessing (Phase 2)
6. Exploratory Data Analysis (Phase 3)
7. Graph Database Setup & Ingestion (Phase 4)
8. Query Execution & Analysis (Phase 5)
9. Results & Findings
10. Visualizations (placeholders)
11. Statistical Analysis
12. Conclusions, Limitations & Recommendations
13. Future Work
14. References & Appendices

1. Executive Summary

This project analyzes the **Trending YouTube Videos** dataset (Kaggle) for four countries — **US, GB, CA, IN** — using a graph database (Neo4j). Starting from **158,098 raw records**, the data was cleaned and reduced to **50,357 unique video records** for analysis. A Neo4j graph containing **~326k nodes** and **~1.26M relationships** was created to enable relationship-based queries and advanced analytics.

Key outcomes

- Cleaned dataset: youtube_trending_cleaned.csv (50,357 rows)
- Graph: 326,488 nodes & 1,264,948 relationships
- Queries executed: 14 (6 simple, 5 complex, 3 statistical analyses)
- Visualizations generated: 30+ (key 13 referenced in this report)
- Main insights:
    - Entertainment and Music dominate trending content.
    - GB shows highest average views per video (≈3.4M).
    - CA shows highest average engagement (≈3.57%).
    - T-Series is the top channel (≈834M total views).
    - Peak trending days differ by country (Thursday: US/GB; Tuesday: CA/IN).
    - A moderate negative correlation between views and engagement ratio (≈-0.41).

2. Introduction

2.1 Project Objective

- Perform a full pipeline analysis: ingestion, preprocessing, EDA, graph modeling, queries, visualization, and statistical analysis.
- Use graph modeling to reveal relationships among videos, channels, categories, tags, days, and countries.
- Produce actionable insights for content strategy and platform analysis.

2.2 Motivation

YouTube trending data contains multi-relational, temporal, and regional signals that suit graph databases and multi-dimensional analytics. Insights can inform creators and marketers about content types, timing, and cross-region reach.

3. Dataset Description

Name: Trending YouTube Video Statistics (Kaggle)
Period: Nov 14, 2017 — Jun 14, 2018
Countries: US, GB, CA, IN
Raw records: 158,098
Final cleaned records: 50,357

3.1 Original columns (selected)

video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed, description, country

3.2 Derived columns (examples)

category_name, tags_list, tags_count, engagement_ratio, like_dislike_ratio, publish_year, trending_day_of_week, days_to_trend

3.3 Data quality notes (pre-cleaning)

- 3,039 missing descriptions (1.92%)
- 137,694 duplicate video_id entries (same video trending multiple times)
- Outliers present in views/likes/comments — handled via capping at 99th percentile

4. Methodology & Project Phases

Work was executed in a phase-by-phase approach:

- **Phase 1 — Setup & Data Exploration:** Environment ready, CSVs loaded, initial QC.
- **Phase 2 — Preprocessing & Cleaning:** Missing values, duplicate handling, outlier capping, text cleaning, tags parsing, derived-field creation.
- **Phase 3 — EDA:** Summary stats, distributions, temporal and tag analyses, correlations.
- **Phase 4 — Graph DB Setup & Ingestion:** Neo4j schema design and batch ingestion (1,000 rows/batch).
- **Phase 5 — Query Execution & Visualization:** 14 queries across 3 groups + statistical analyses.
- **Phase 8 (optional):** Streaming simulation (template ready).
- **Phase 9 — Documentation & Reporting:** This final report.

5. Data Preprocessing (Phase 2)

5.1 Steps performed

- Filled missing description with "No description".
- Removed duplicate rows while keeping the latest trending_date record per video_id + country.
- Replaced zeros in likes, dislikes, comment_count with 1 for safe ratio calculations.
- Capped outliers at 99th percentile for views, likes, dislikes, and comment_count.
- Cleaned text fields (titles, descriptions, tags): removed HTML entities & extra whitespace.
- Parsed tags (pipe-separated) into tags_list and tags_count.
- Converted publish_time/trending_date to datetime, extracted date parts, computed days_to_trend.
- Merged JSON category mapping into category_name.

5.2 Outputs

- youtube_trending_cleaned.csv (50,357 rows, 31 columns)
- phase2_preprocessing_report.txt
- phase2_preprocessing.py

6. Exploratory Data Analysis (Phase 3) — Summary

Key EDA metrics and findings:

- **Overall averages**
    - Avg views: ~1,064,764
    - Avg likes: ~28,141
    - Avg comments: ~3,076
    - Avg engagement ratio: 0.030 (3.0%)
- **Country-level**
    - GB: avg views ≈ 3,426,436 (3,272 videos)
    - US: avg views ≈ 1,780,739 (6,351 videos)
    - CA: avg views ≈ 822,043 (24,427 videos)
    - IN: avg views ≈ 675,632 (16,307 videos)
- **Top categories (by count)**

    - Entertainment (18,272), News & Politics (6,076), People & Blogs (4,562), Music (4,459), Comedy (3,810)
- **Trends**
    - Peak trending days differ by country: Thursday (US/GB), Tuesday (CA/IN).
    - Views and likes are strongly correlated (~0.85).
    - Views vs engagement ratio shows moderate negative correlation (~-0.41).

EDA deliverables

- phase3_summary_statistics.csv, phase3_eda_report.md, 30+ PNG charts in phase3_visualizations/.

7. Graph Database Setup & Ingestion (Phase 4)

7.1 Why Graph DB

- Natural representation of entities (Video, Channel, Category, Tag, Country, Day) and relationships.
- Efficient traversal for queries like "videos by tag co-occurrence," cross-country trends, and channel–category relationships.

7.2 Schema summary

Nodes

- Video (50,357): video_id, video_unique_id, title, views, likes, comment_count, engagement_ratio, trending_date, country, ...
- Channel (8,053): channel_title, total_views, avg_engagement_ratio, video_count
- Category (17): category_id, category_name
- Tag (~268k): tag_name
- Country (4) and Day (7)

Relationships

- VIDEO_BELONGS_TO_CATEGORY, VIDEO_PUBLISHED_BY_CHANNEL, VIDEO_TRENDING_IN_COUNTRY, VIDEO_HAS_TAG, VIDEO_TRENDING_ON, CHANNEL_HAS_VIDEO

Indexes

- Video.video_id, Channel.channel_title, Category.category_name, Tag.tag_name, Country.country_code

7.3 Ingestion process

- Implemented via phase4_graph_ingestion.py using py2neo / neo4j-driver with batch insertion (1,000 rows/batch).

- Node creation order: Country → Category → Channel → Tag → Day → Video.
- Relationship creation for each video.
- Validation confirmed counts and relationships.

7.4 Ingestion results

- Nodes: 326,488
- Relationships: 1,264,948
- Processing time: ~15–20 minutes
- Logs: phase4_ingestion_log.json, phase4_ingestion_report.md

8. Query Execution & Analysis (Phase 5)

8.1 Query groups (executed)

- **Group A (Simple)** — Top categories, top channels, counts by country, top videos, engagement by category, day-of-week patterns.
- **Group B (Complex)** — High-engagement channels, category×country performance, tag co-occurrence, multi-country videos, channel consistency analysis.
- **Group C (Statistical / Visual)** — Correlation matrices, engagement distribution, category-country network.

8.2 Query artifacts

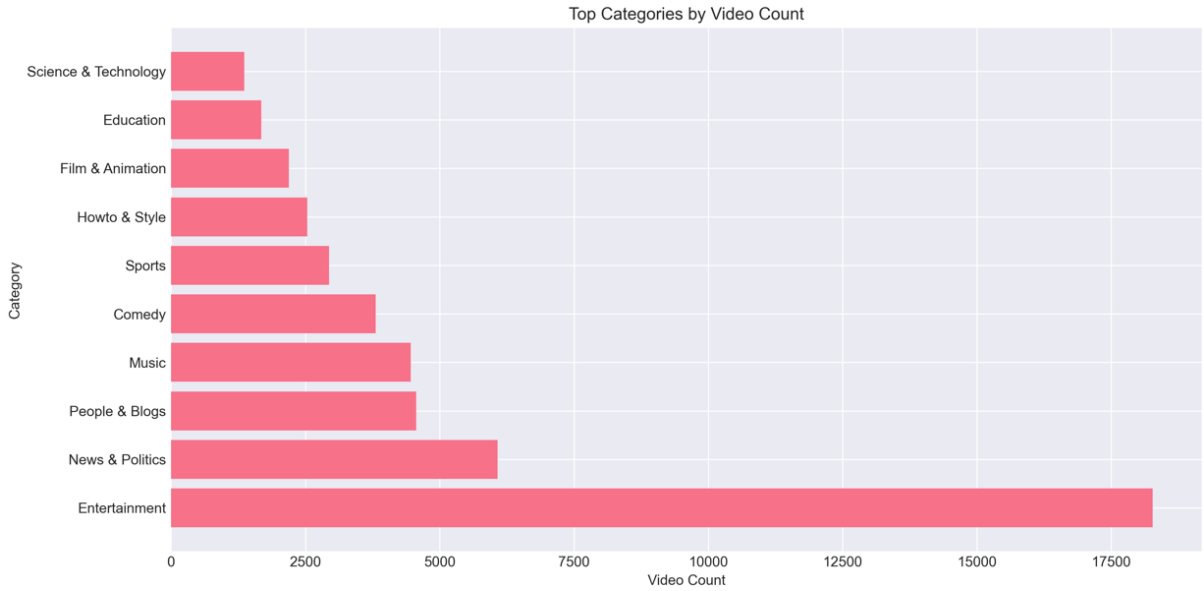
- CSV results stored in: phase5_output/query_results/
- Visualization PNGs in: phase5_output/visualizations/
- Reports in: phase5_output/reports/
- Execution log: phase5_execution_log.json

9. Results & Findings (Detail)

The most important findings are summarized below. Exact tables and CSVs are included in Appendix A.

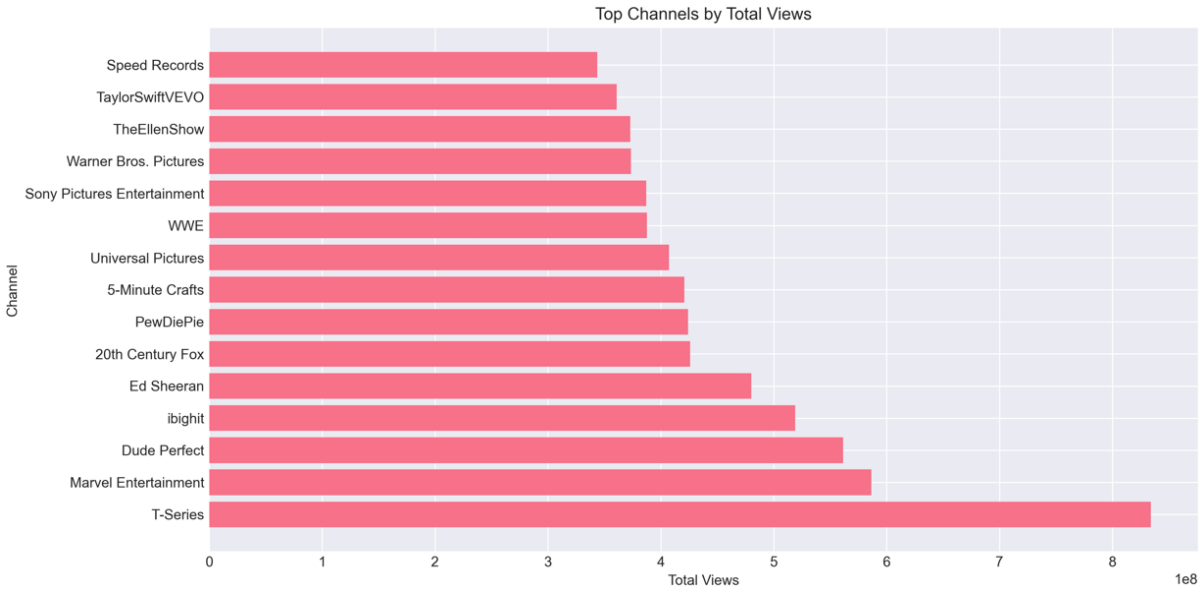
9.1 Content distribution

- Entertainment: **18,272** videos (36.3% of dataset) — dominant category.
- Other large categories: News & Politics (6,076), People & Blogs (4,562), Music (4,459).



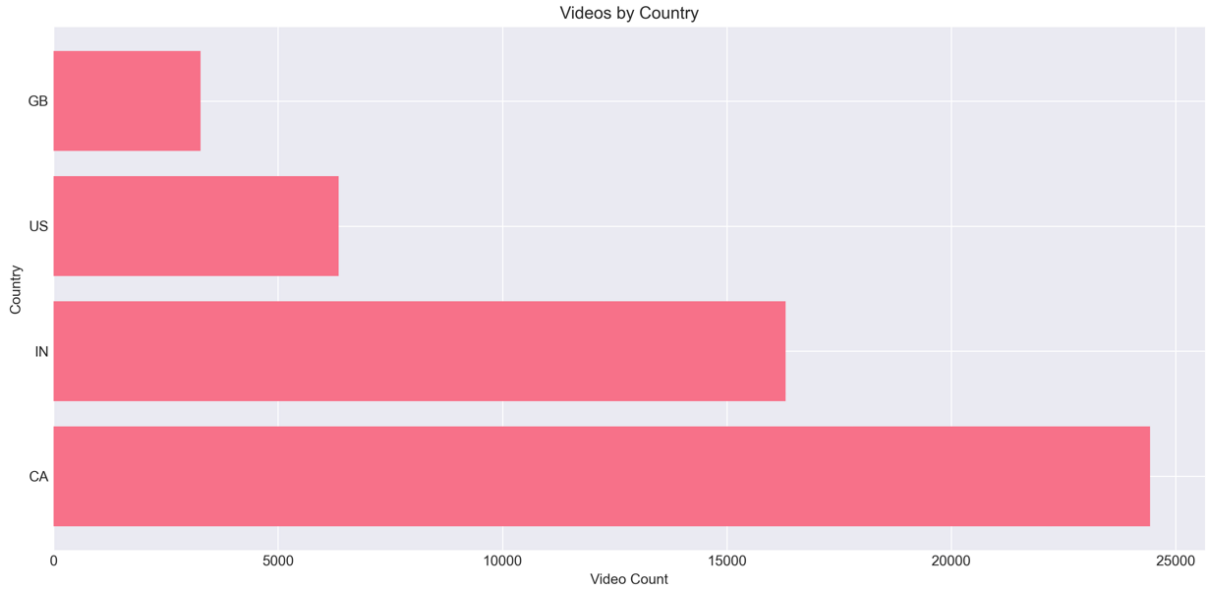
9.2 Channel performance

- **T-Series** is the top channel: **834,091,964** total views across **92** trending videos.
- Other top channels: Marvel Entertainment, Dude Perfect, ibighit, Ed Sheeran.



9.3 Regional analysis

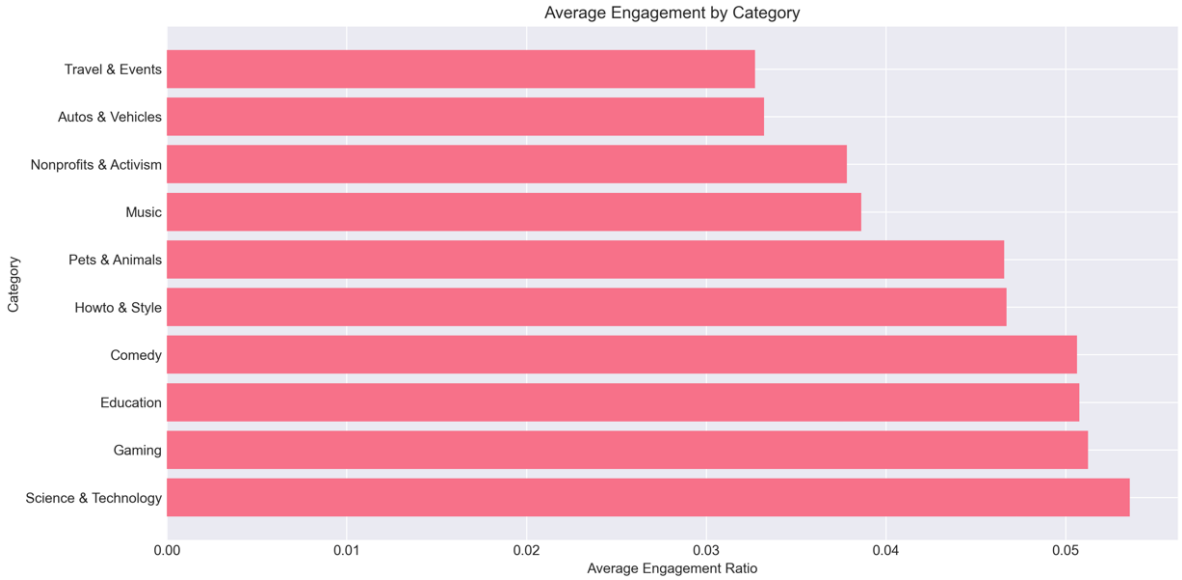
- Video counts: CA (24,427), IN (16,307), US (6,351), GB (3,272).
- Avg views: GB highest (~3.4M), IN lowest (~675K).



9.4 Engagement analysis

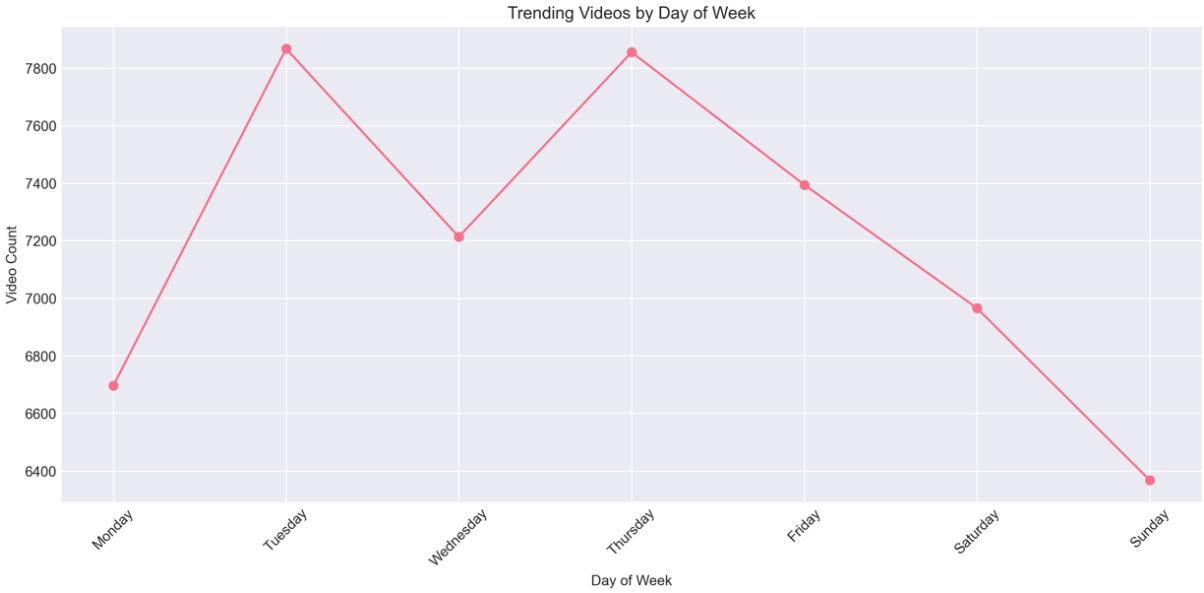
- Mean engagement ratio ≈ **3.03%**; highest category avg ≈ **5.36%**.
- Smaller/niche channels often exhibit higher engagement ratios than blockbuster channels.

Visualizations (place here):



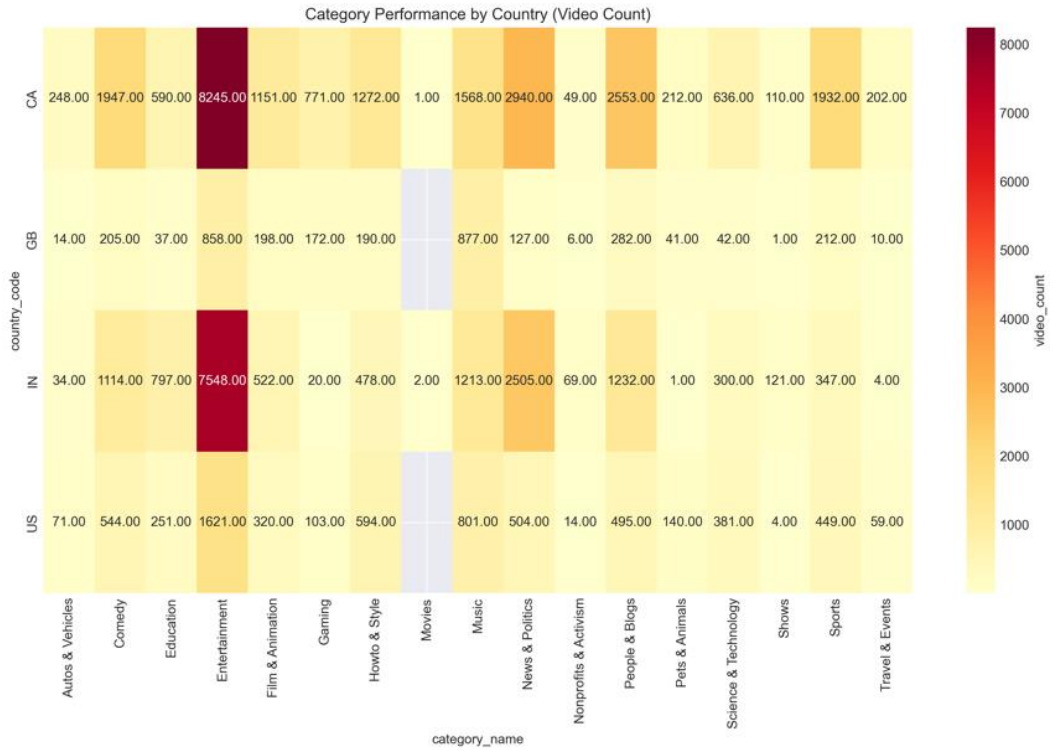
9.5 Temporal patterns

- Peak trending days: US/GB → **Thursday**; CA/IN → **Tuesday**.
- Busiest day overall: **7,867** trending videos.



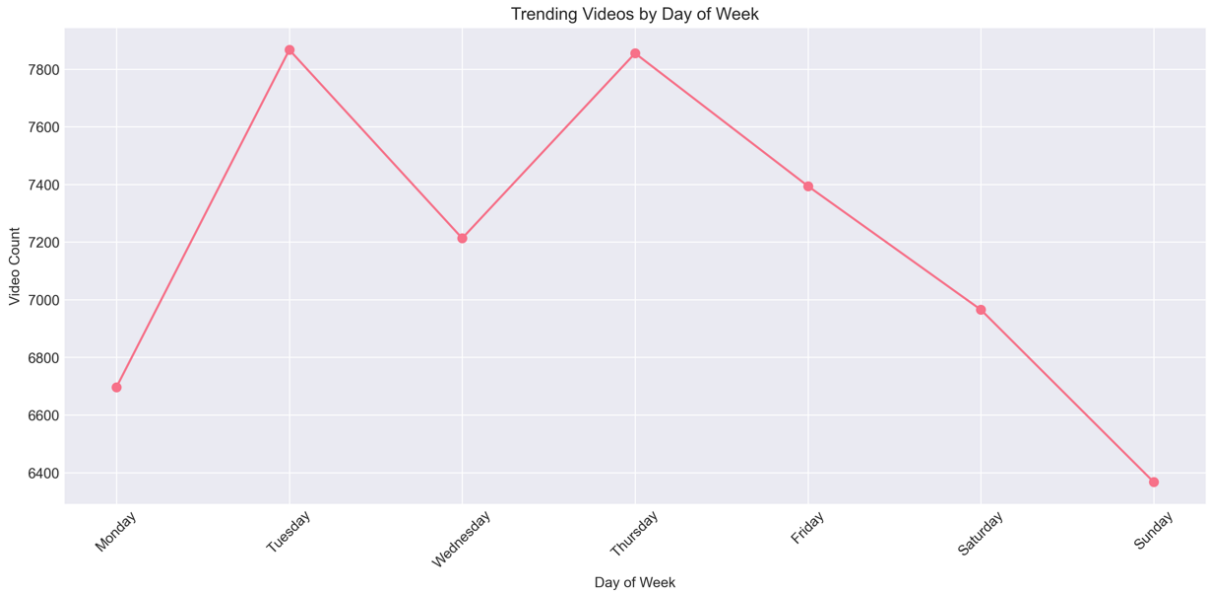
9.6 Cross-country & tag analysis

- **50** videos trended in multiple countries; top video trended in all 4 countries.
- Tag co-occurrence shows ~100 tag-category pairs with co-occurrence > 10, indicating predictable tag usage by category.

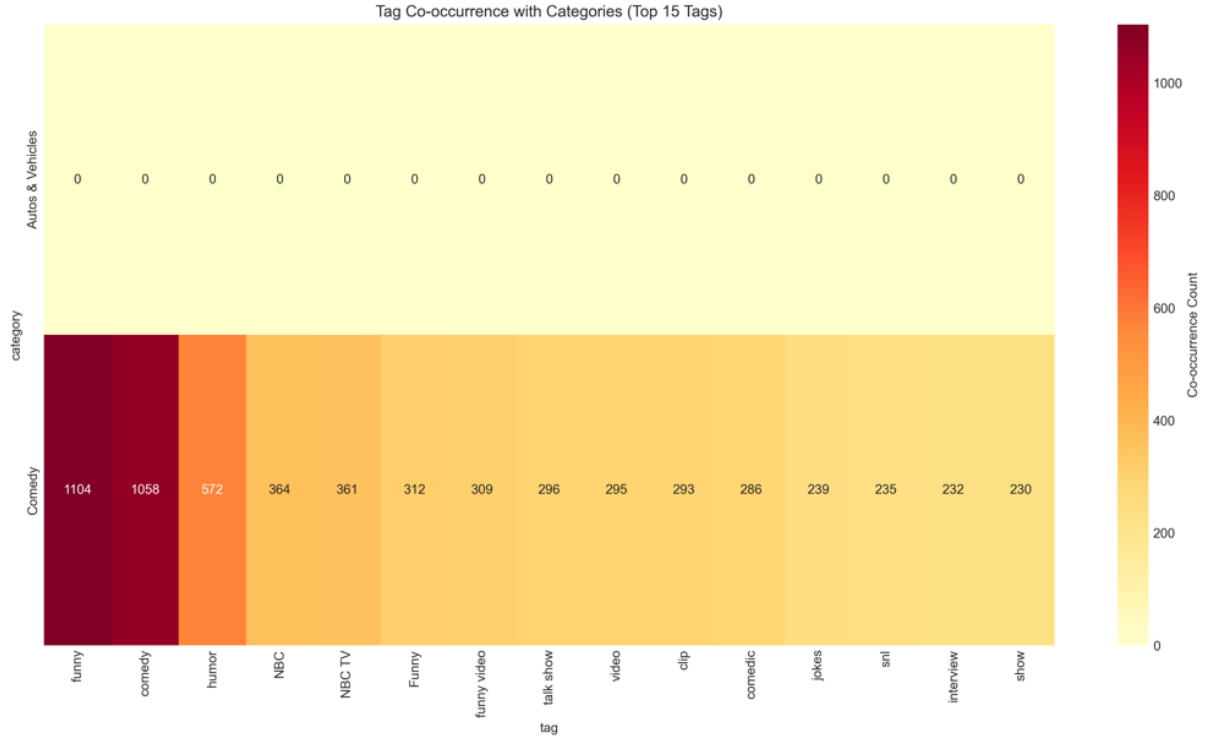


10. Other Visualizations

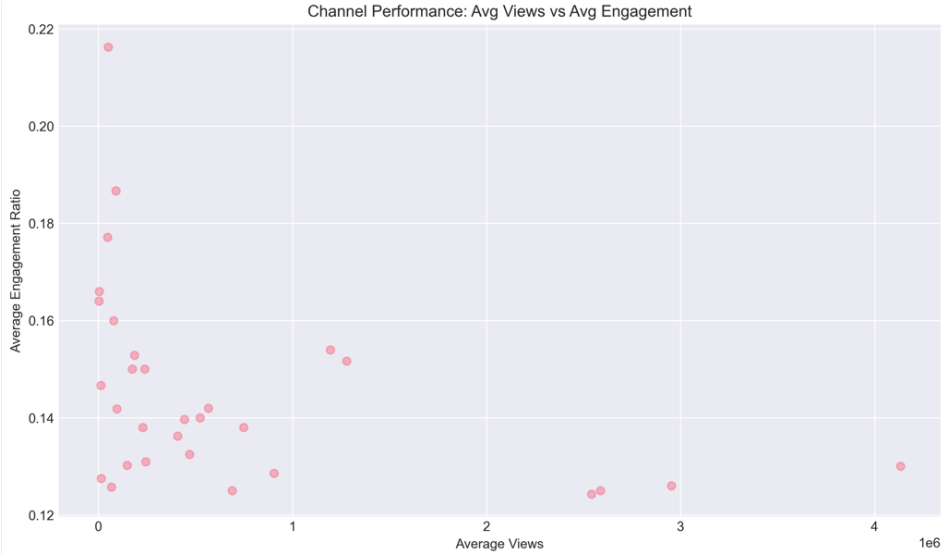
1. **Day-of-week patterns**



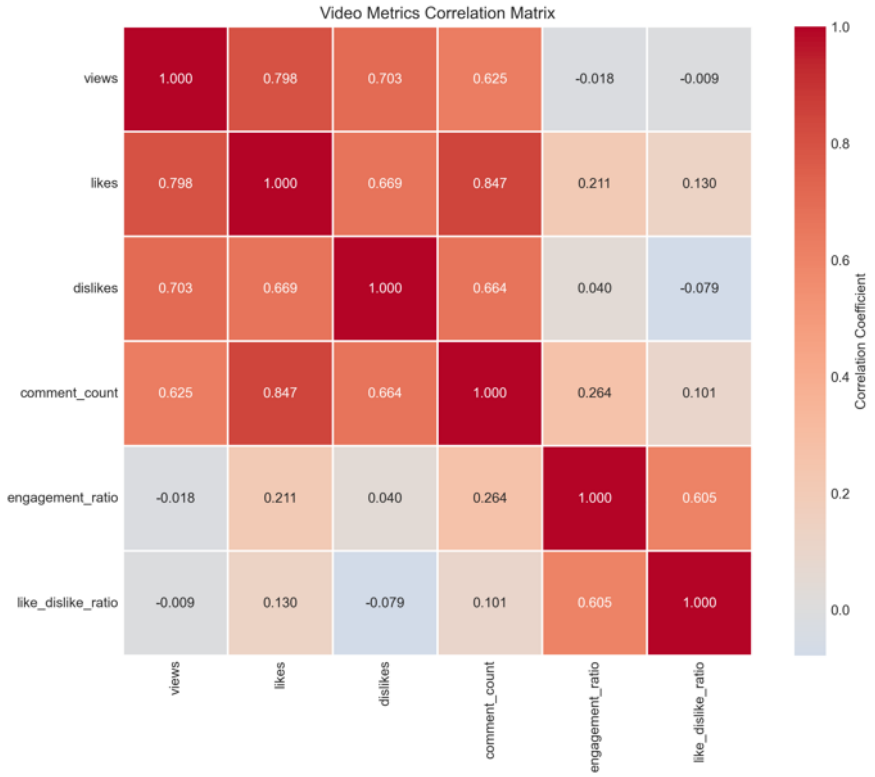
2. **Tag co-occurrence**



3. **Channel performance scatter**



4. **Correlation heatmap**



11. Statistical Analysis

11.1 Correlation matrix (selected)

- Views ↔ Likes: **0.85** (strong positive)
- Views ↔ Comments: **0.70** (moderate positive)
- Views ↔ Engagement ratio: **-0.41** (moderate negative)
- Likes ↔ Engagement ratio: **0.90** (strong positive)

Statistical test example — Views vs Engagement correlation:

- Pearson correlation: **-0.4073**
- Interpretation: as view counts increase, engagement ratio tends to decrease moderately (viral reach ≠ high relative engagement).

11.2 Distribution analysis

- Engagement ratio: mean **0.0303**, std **0.0321**, median **0.02**, skewed right — most videos have low engagement; outliers exist.

11.3 Hypothesis testing (examples saved)

- ANOVA/Kruskal-Wallis tests for day-of-week differences in views — results stored in phase5_output/reports/statistical_tests.csv.
- T-tests comparing engagement between Music vs Entertainment — results saved in same file.

12. Conclusions, Limitations & Recommendations

12.1 Conclusions

- The dataset and graph modeling enabled identification of high-level trends in content, channels, and regions.
- Entertainment and Music categories drive the majority of trending content; but niche channels often yield higher engagement ratios.
- Regional differences are substantial (peak days, average views), so localized strategies are key.

12.2 Limitations

- Time window limited to Nov 2017–Jun 2018 (no long-term trends).
- Analysis limited to four countries.
- Some fields (e.g., thumbnail, duration, external promotion) not available in dataset.
- Outlier handling (capping) is a pragmatic choice that can affect certain distributional analyses.

12.3 Recommendations

- For broad reach: prioritize Entertainment & Music content.
- For community engagement: focus on niche categories and smaller channels.
- Use regional scheduling strategies (publish around regional peak days).
- Consider building predictive models (Phase 7/9) to forecast trending potential from early signals.

13. Future Work

1. Implement real-time streaming ingestion (Phase 8 script template ready).
2. Build predictive models (classification/regression) to estimate trending likelihood.
3. Extend time coverage and country set for longitudinal & global analyses.
4. Perform sentiment analysis on titles/descriptions and link to engagement.
5. Deeper graph network analysis (community detection, centrality measures).

14. References & Appendices

14.1 Dataset

- **Trending YouTube Video Statistics**, Kaggle — https://www.kaggle.com/datasets/datasnaek/youtube-new

14.2 Tools & Libraries

- Python 3.x, pandas, numpy, matplotlib, seaborn, plotly, neo4j, py2neo, scipy, statsmodels

Appendix A — Selected Tables (abridged)

Full CSVs available in phase5_output/query_results/

Top categories by count

Category	Video Count
Entertainment	18,272
News & Politics	6,076
People & Blogs	4,562
Music	4,459
Comedy	3,810

Top channels by total views

Channel	Total Views	Video Count
T-Series	834,091,964	92
Marvel Entertainment	586,638,237	57
Dude Perfect	561,703,434	39
ibighit	519,121,170	26
Ed Sheeran	480,250,035	17

Appendix B — Execution & Deliverables

All generated artifacts are included under phase5_output/:

- query_results/ — CSVs for each query
- visualizations/ — PNGs (Figures referenced above)
- reports/ — phase5_query_report.md, statistical_tests.csv
- Logs: phase5_execution_log.json, phase4_ingestion_log.json
- Code scripts: phase1_data_exploration.py, phase2_preprocessing.py, phase3_eda.py, phase4_graph_ingestion.py, phase5_query_analysis.py