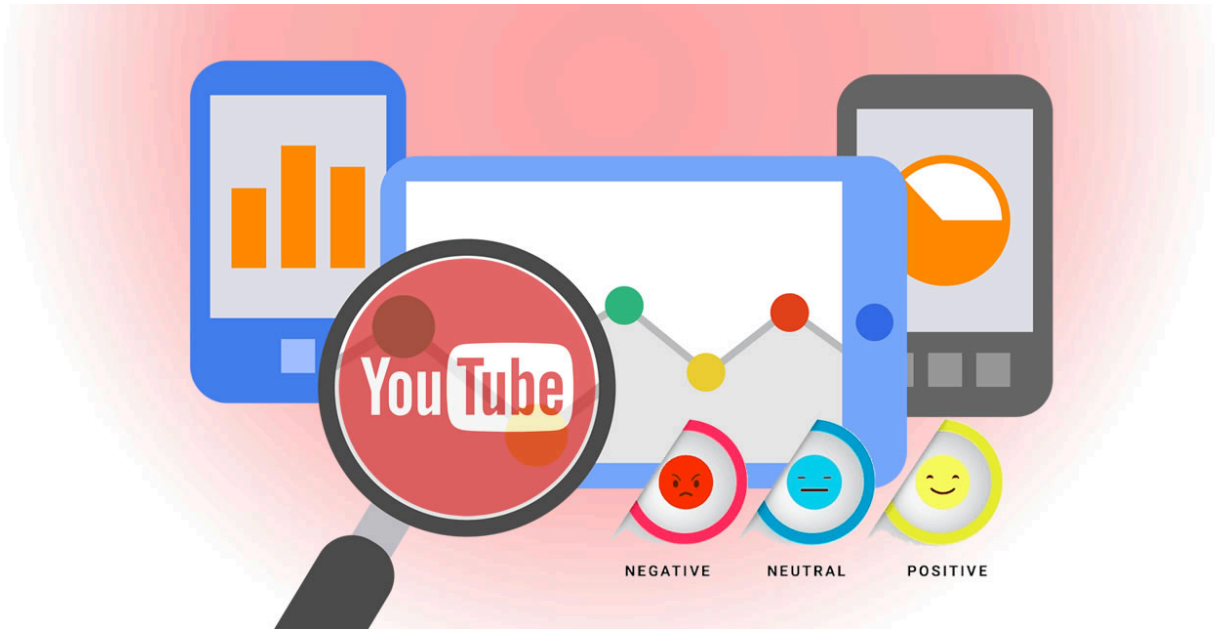
Analyzing Learning Trends in YouTube Shorts:
A Real-Time Research Study



Author: Devika Kadam

Affiliation: Arizona State University

Project: YouTube Shorts Learning Trend Analyzer

Supervisor: Professor Xiao

TABLE OF CONTENTS

1. ABSTRACT

This study examines how learning-focused YouTube Shorts perform and what factors influence their reach, engagement, and viral behavior. Using real-time metadata across eight educational categories, the project applies detailed exploratory analysis, feature engineering, outlier detection, category scoring, and two forms of recommendation systems. The results show that concise videos (20–60 seconds), clear titles, and category relevance consistently drive higher performance. Metadata length (tags, descriptions) has little impact, while category and duration patterns strongly influence viewer engagement. Viral Shorts share clear structural traits, and certain categories such as Study Skills and Finance dominate high-performing clusters. The research provides actionable insights for creators, educators, and analysts and offers a complete workflow for understanding educational short-form content dynamics.

2. INTRODUCTION

Short-form platforms like YouTube Shorts have become a major medium for sharing fast, accessible educational content. Viewers expect quick value, and creators often experiment with different structures, titles, and posting times—but many of these decisions are based on assumptions rather than data.

This research aims to investigate which factors actually drive performance for learning-focused YouTube Shorts. The goal is to understand how duration, category choice, posting time, titles, descriptions, and other metadata influence reach and engagement. The project also identifies viral patterns and builds simple recommendation systems to demonstrate how data-driven insights can support content strategy.

3. DATASET DESCRIPTION

The dataset consists of real YouTube Shorts from eight learning categories: Study Skills, Finance, Tech Skills, AI Tools, Mental Health, Career Tips, Soft Skills, and Travel Learning. For each video, the following information was available:

- Title, description, and tags
- View count, likes, comments
- Duration in seconds
- Published date, hour, and weekday
- Category labels
- Engagement rate
- Title and description word counts

This dataset was expanded through feature engineering to support deeper analysis, including time-based features, normalized scores, duration buckets, and outlier flags.

4. Methodology

The methodology for this research was designed to create a complete end-to-end analytical workflow that converts raw YouTube Shorts metadata into meaningful insights. Each step from cleaning the dataset to building recommendation systems was executed systematically to ensure accuracy, reliability, and interpretability. The methods used were chosen because they are appropriate for short-form social media data, which tends to be highly skewed and behavior-driven.

The workflow consisted of data cleaning, feature engineering, exploratory data analysis, viral outlier detection, category scoring, and recommendation system development. Together, these steps form a structured analytical pipeline that mirrors real research and industry standards.

4.1 Data Cleaning

Data cleaning focused on ensuring that the dataset was accurate, complete, and ready for analysis. Since YouTube metadata often contains missing fields, inconsistent formatting, or irregular text, this step was essential.

- All timestamps (publish date and time) were converted from string format to Python datetime objects.
  This allowed accurate extraction of posting hour, weekday, and month.
- Missing or invalid fields were handled to prevent errors in calculations and visualizations.
- Titles and descriptions were standardized by removing leading/trailing spaces and normalizing text.
  This ensured correct word counts and better quality for text-based analysis.
- Empty or unusable rows were removed to maintain data integrity.

Data cleaning created a solid foundation, making sure that every subsequent analysis step relied on clean and trustworthy data.

4.2 Feature Engineering

Feature engineering played a central role in enriching the dataset. YouTube Shorts metadata alone does not fully capture viewer behavior or content structure, so new variables were created to better understand performance patterns.

4.2.1 Engagement Rate

Engagement rate was calculated using:
(likes + comments) ÷ view count × 100

This metric allows fair comparison between videos regardless of their total views. A video with fewer views but high interaction may have stronger audience impact than a viral video with weak interaction. Engagement rate helped identify meaningful viewer responses beyond raw popularity.

4.2.2 Category and Sub-category Fields

The original category column was supplemented with a more detailed sub-category field to capture niche topics such as:

- “Study hacks”
- “Motivation for students”
- “Packing tips”
- “AI tools for productivity”

This helped break down broad themes into more specific groups, enabling more nuanced analysis.

4.2.3 Time-Based Features

Several new time variables were created by extracting components from the publish timestamp:

- published_month
- published_weekday

- published_hour
- hour_bucket (Late Night, Morning, Afternoon, Evening)

These variables made it possible to analyze posting behavior and performance patterns across different times and days. Understanding time trends is important because viewer activity varies throughout the day.

4.2.4 Duration and Text Metrics

Content structure strongly influences viewer attention on short-form platforms. To study this, several structural features were engineered:

- Duration buckets (10–20 sec, 21–40 sec, 41–60 sec, etc.)
- title_length (words)
- description_length (words)
- tag_count

These features enabled analysis of whether shorter titles, longer descriptions, or certain duration ranges performed better.

4.2.5 Normalized Scores

Raw view counts and engagement rates have large variations and cannot be compared directly across categories. Therefore, each metric was normalized between 0 and 1 to compute:

- view_score
- engagement_score

These formed the foundation for category ranking, allowing an objective evaluation of which categories perform best overall.

Feature engineering transformed basic metadata into meaningful analytical variables that helped uncover deeper behavioral patterns.

4.3 Outlier Detection (Viral Identification)

Identifying viral videos required a method that works well for skewed data distributions. The Interquartile Range (IQR) method was chosen because it is robust, simple, and suitable for distributions that contain extreme values.

The process involved:

1. Calculating the 25th percentile (Q1) and 75th percentile (Q3).
2. Computing IQR = Q3 – Q1.
3. Setting the viral threshold at Q3 + 1.5 × IQR.
4. Labeling any video above this threshold as a viral outlier.

This allowed the research to isolate truly exceptional performance and compare viral vs. non-viral videos. Viral analysis helped reveal structural traits that are common among high-performing Shorts.

4.4 Analytical Procedure

After engineering all variables, the dataset was explored using a structured blend of descriptive, visual, and qualitative analysis.

4.4.1 Descriptive Statistics

Basic statistics such as mean, median, range, and variance were computed for views, likes, comments, engagement rate, title length, and duration.
This offered a baseline understanding of dataset characteristics before deeper exploration.

4.4.2 Exploratory Data Analysis (EDA)

EDA formed the core of the research. It included:

- Category-wise comparisons
- Duration vs. views and engagement relationships
- Title/description length patterns

- Time-of-day and weekday performance
- Viral vs. non-viral comparisons
- Correlation analysis
- Scatterplots to explore multi-variable interactions
- Histograms and boxplots to understand distributions

Each visualization was paired with interpretation to explain what the pattern means and why it matters.

4.4.3 Comment Analysis (Qualitative)

Viewer comments were reviewed to understand how audiences responded to the content. Comments were categorized into:

- Appreciation / Positive feedback
- Confusion or unclear parts
- Questions or requests for more information
- Repeated themes

This provided context for understanding viewer satisfaction and clarity.

4.4.4 Scoring System (Category Ranking)

To objectively rank categories, the normalized view_score and engagement_score were combined using the formula:

Final Score = (view_score + engagement_score) ÷ 2

This scoring system helped identify which learning categories consistently generate the strongest results, balancing both popularity and engagement.

5. OUTLIER DETECTION (VIRAL VIDEO IDENTIFICATION)

To identify “viral” Shorts, an Interquartile Range (IQR) method was applied to the view count distribution. This method is appropriate for skewed social media data where a small number of videos receive disproportionately high views.

The steps were:

1. Calculate Q1 and Q3.
2. Compute the IQR (Q3 – Q1).
3. Determine the viral threshold as Q3 + 1.5 × IQR.
4. Label any video exceeding this threshold as viral.

This method avoids assumptions of normality and focuses on true statistical extremes. The viral cluster was later compared with the rest of the dataset to identify structural commonalities.

```
Q1: 2194.75
Q3: 661096.75
IQR: 658902.0
Upper Bound for Outliers: 1649449.75

Number of Viral Shorts Detected (IQR): 64
```

	video_id	main_category	view_count	engagement_rate
51	cQO0lqMLBkk	Career Tips	1769537	0.0636
53	piWdncMDyNs	Career Tips	21579593	0.0377
93	NMuaxhPbP6Y	Career Tips	2467359	0.0826
100	b5BsywkETfU	Finance	29305157	0.0445
106	1OF53QNbMrE	Finance	8274423	0.0473
107	0GxSnWSp3VM	Finance	30094889	0.0320
114	RIOcs8stB6w	Finance	3676678	0.0361
117	AIOWO943HtM	Finance	6775348	0.0276
122	b_X9JYZKsuI	Finance	1787021	0.0239
124	T8msuRYVeRA	Finance	3958383	0.0184

Figure: Viral YouTube Shorts identified using the IQR-based outlier detection method.

6. ANALYTICAL PROCEDURE

The analytical procedure combined descriptive statistics, graphical exploration, and qualitative assessment to understand how each feature influenced the performance of learning-focused YouTube Shorts. This step was essential for translating the cleaned and engineered dataset into clear insights.

6.1 Descriptive Statistics

Basic statistics such as mean, median, maximum, and minimum values were calculated for views, likes, comments, duration, and engagement rate. This helped establish a baseline understanding of how the videos in the dataset performed overall, and revealed how spread out the values were.

6.2 Exploratory Data Analysis (EDA)

EDA included:

- View distributions
- Engagement rate patterns
- Duration vs. views relationships
- Category-wise comparisons
- Time-of-day performance
- Title and description length effects
- Correlation analysis
- Scatterplots for multi-feature understanding
- Boxplots to compare across groups

Each chart was followed by factual interpretation in later sections.

6.3 Comment Analysis (Qualitative Review)

Viewer comments were reviewed to understand the audience's reactions to educational Shorts. Comments were grouped into themes such as appreciation, confusion, repeated

questions, and requests for more detail. This qualitative step added context to the numerical findings and highlighted the clarity and usefulness of the videos from a viewer's perspective.

6.4 Recommendation Scoring System

Normalized view scores and engagement scores were combined to produce a final score for each category. This scoring method allowed the study to objectively compare different learning categories and determine which ones consistently attracted the strongest viewer response. The final rankings are presented in Section 8.

Normalized view_score and engagement_score were combined to generate:
Final Score = (view_score + engagement_score) ÷ 2.
This created an objective ranking of categories.

This analytical procedure ensured a balanced combination of numerical, visual, and qualitative insights, allowing for a full understanding of how educational Shorts behave on YouTube.

7. RESULTS AND INTERPRETATION

7.1 Duration Patterns

Duration emerged as one of the strongest predictors of performance. Videos between 20 and 60 seconds consistently achieved higher views and better engagement. Very long Shorts (above 100 seconds) performed poorly regardless of category. This confirms that learning content in short form must remain direct and concise to maintain viewer interest.

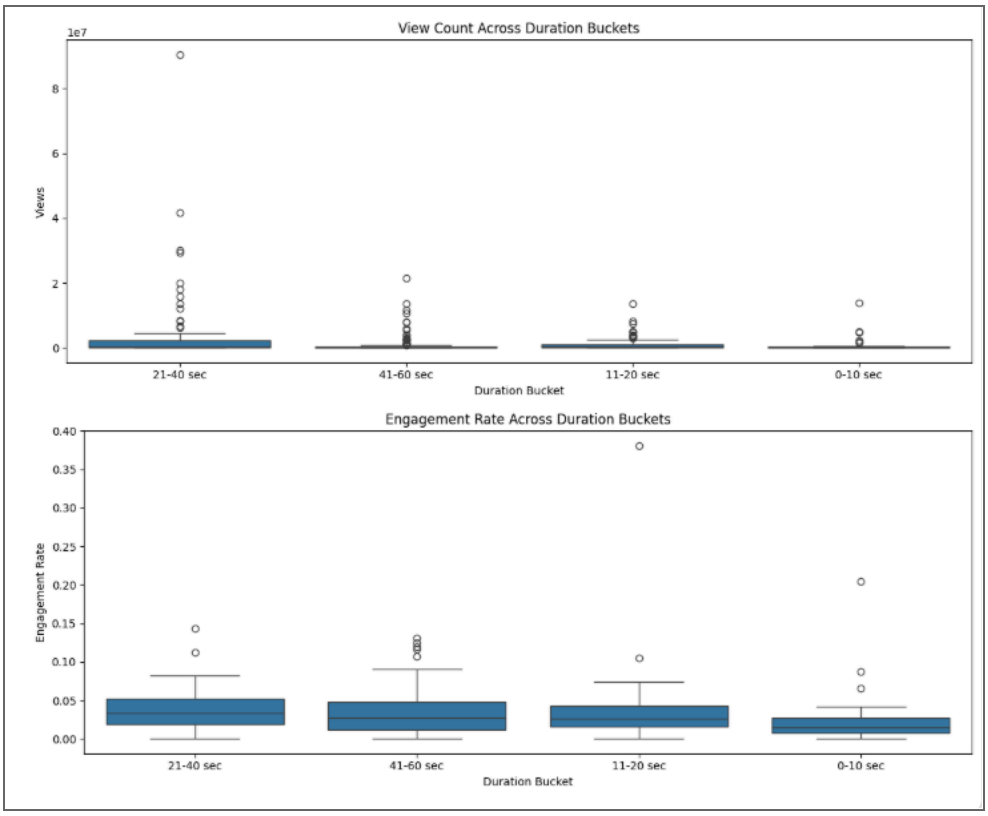


Figure: Duration Distribution & Duration vs Views Scatterplot

7.2 Category Performance

Study Skills, Finance, and Career Tips dominated high-performing segments. Tech Skills and AI Tools performed moderately but showed promising engagement. Soft Skills and Travel Learning showed weaker and inconsistent performance.

The results suggest that categories offering practical, high-value information attract stronger viewer attention.

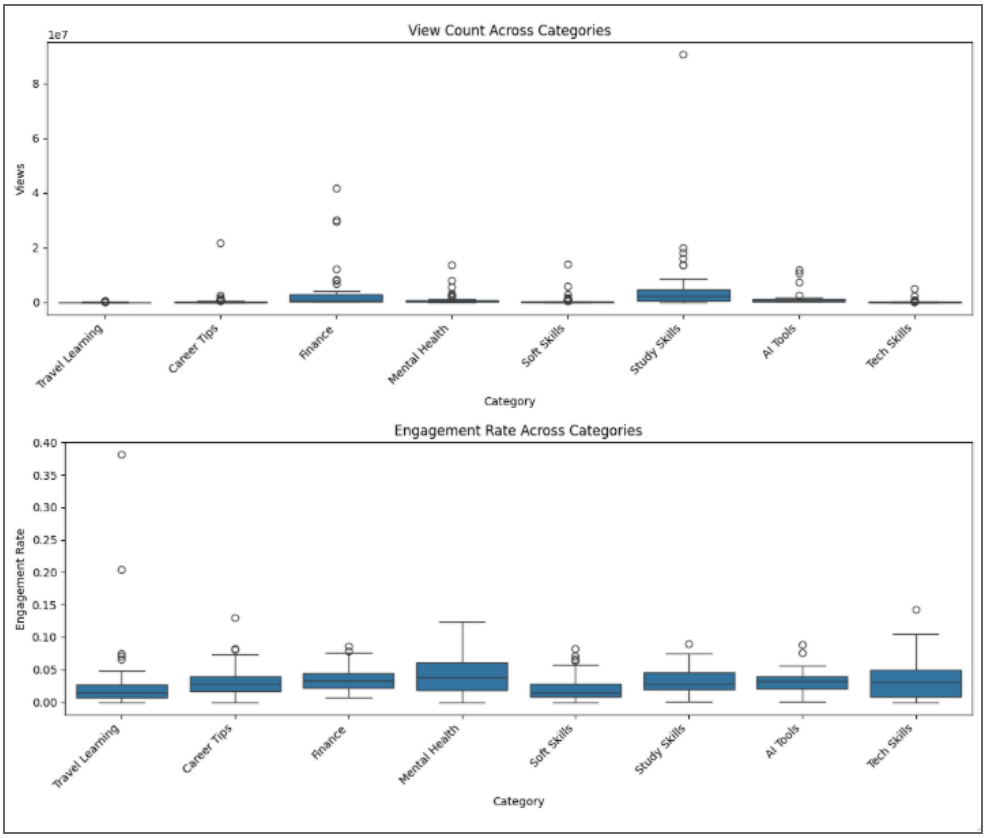


Figure: Average Views by Category + Engagement Rate by Category Boxplot

7.3 Effect of Posting Time

Afternoon and Evening posting windows consistently produced higher views, likely due to user availability. Engagement rate remained stable across days, suggesting that viewer interaction depends more on content quality than posting time.

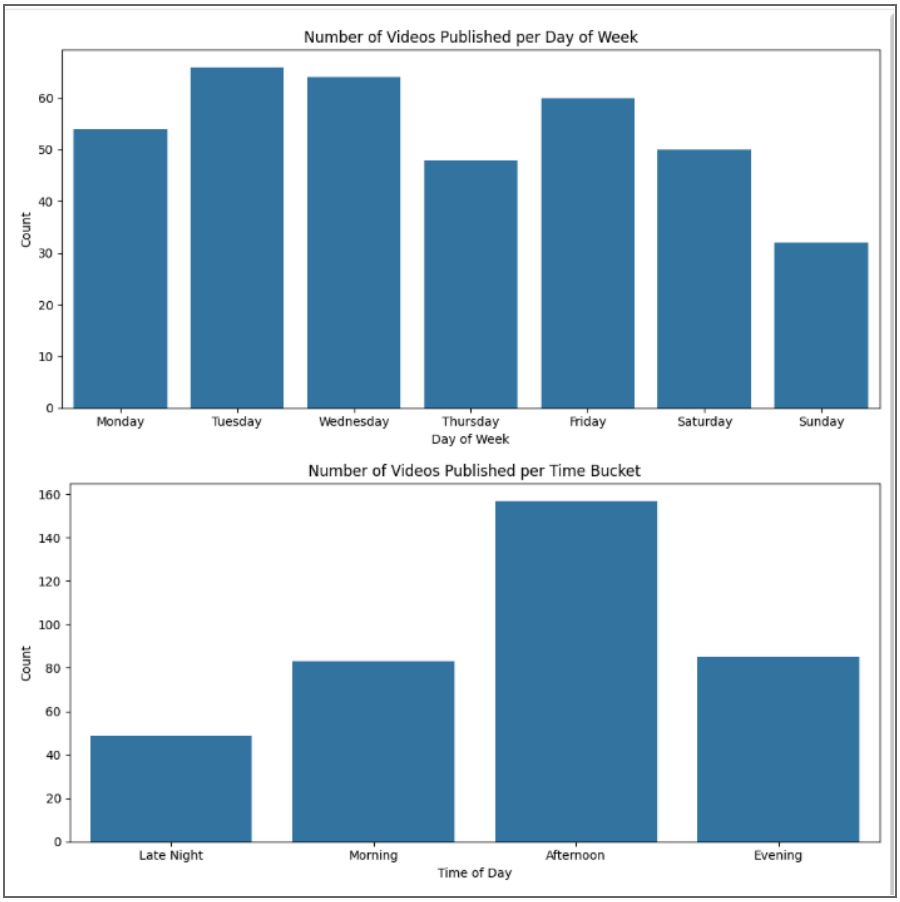


Figure: Posting Hour Distribution + Hour Bucket Countplot

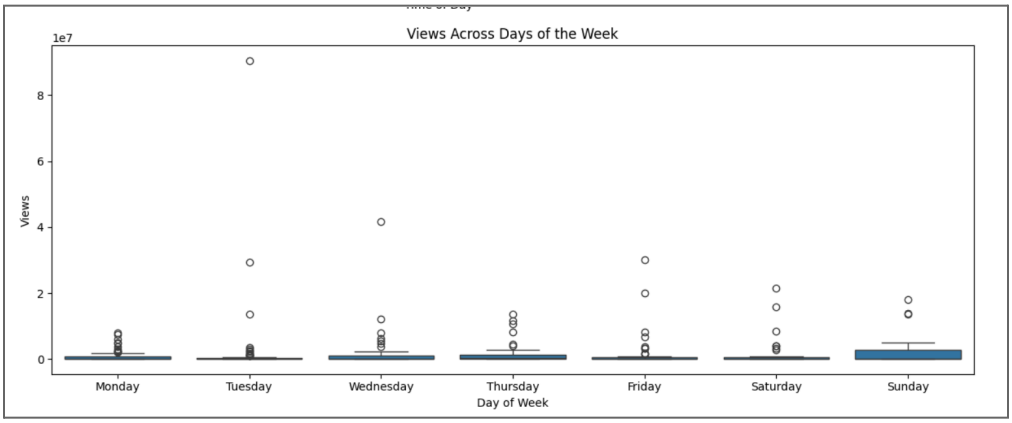


Figure: Views by Weekday Boxplot

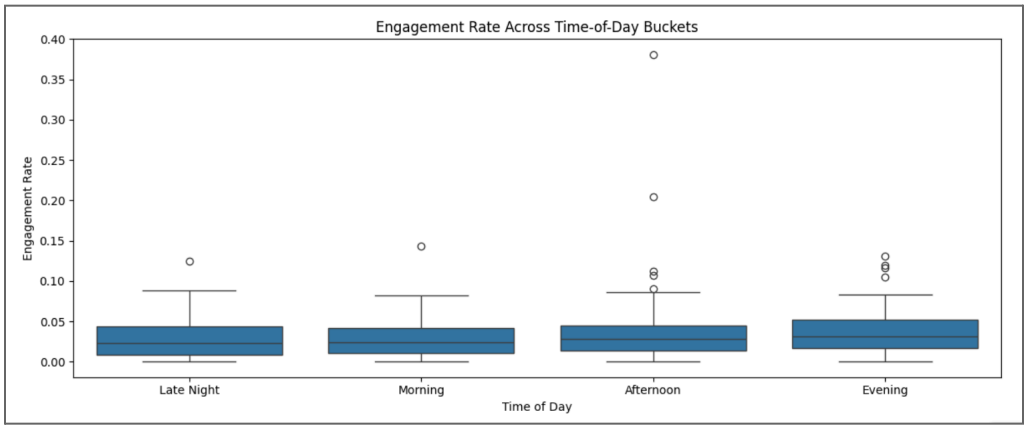


Figure: Engagement by Hour Bucket

7.4 Metadata Effects

Title length, and description length showed very weak or no correlation with performance. This indicates that metadata quantity does not improve visibility. Only clarity matters—simple titles perform better than overly long ones.

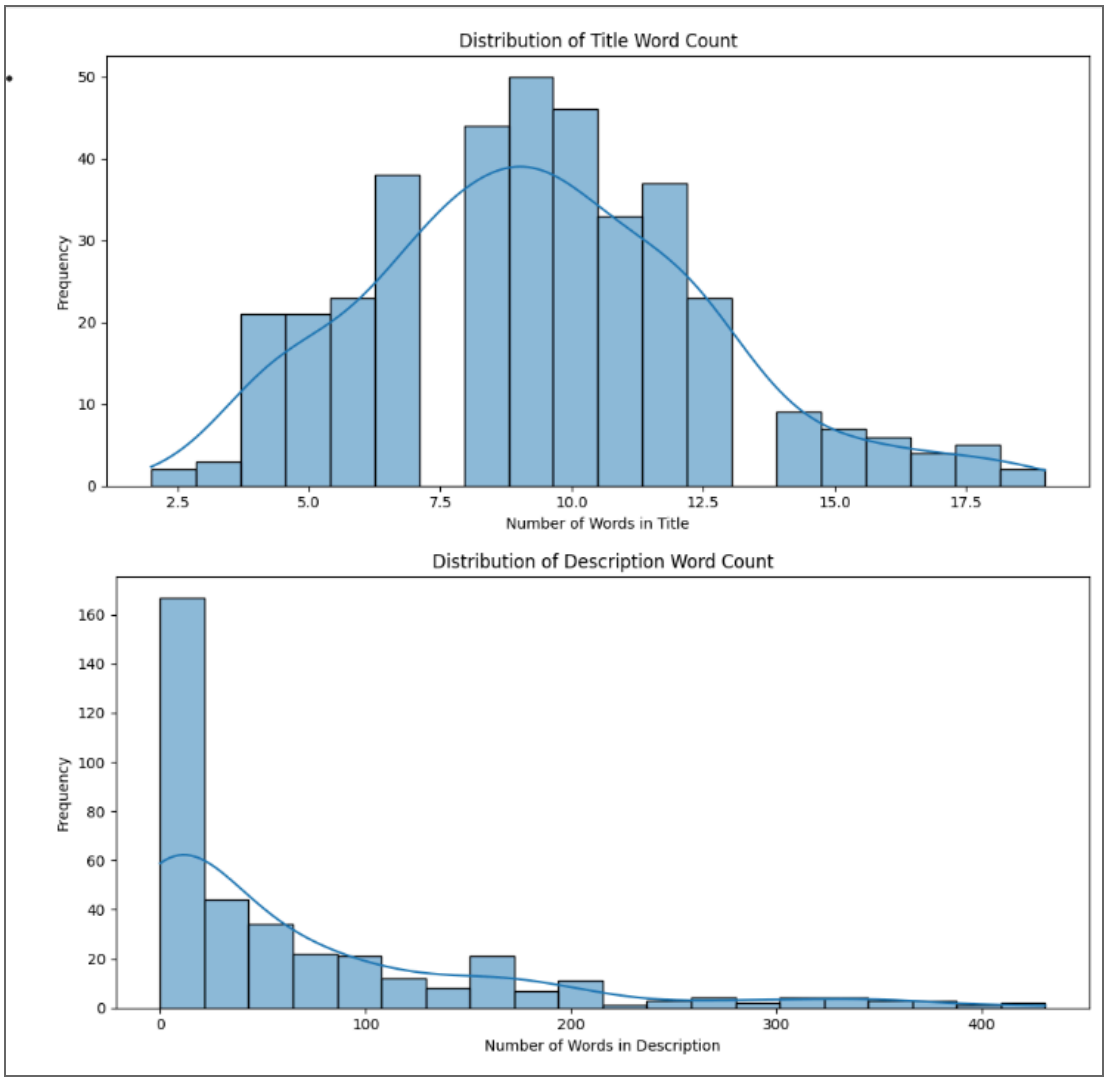


Figure: Title Length Histogram + Description Length Histogram

7.5 Correlation Analysis

Views, likes, and comments formed a tight cluster, meaning they grow together. Engagement rate behaved independently, showing that popularity and viewer interaction measure different outcomes. Duration showed weak linear correlation but strong non-linear behavior—mid-range duration performs best.

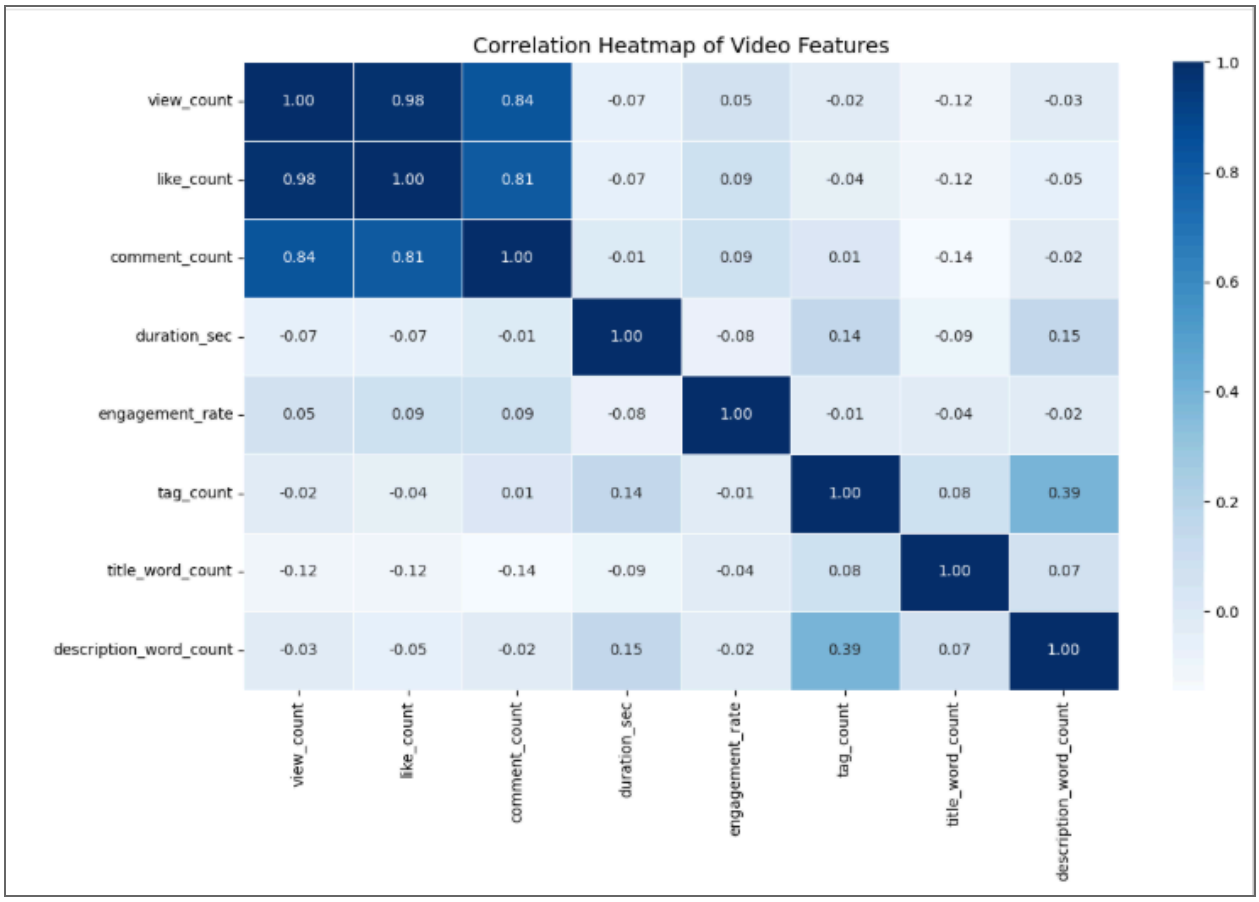


Figure: Correlation Heatmap

7.6 Viral Video Characteristics

Viral Shorts shared clear patterns:

- Duration between 20–40 seconds
- Strong, practical categories (Finance, Study Skills)
- Clear and simple titles
- Short descriptions
- High viewer intent topics

This confirms that viral performance is not random; it follows predictable content structures.

```
Q1: 2194.75
Q3: 661096.75
IQR: 658902.0
Upper Bound for Outliers: 1649449.75

Number of Viral Shorts Detected (IQR): 64
```

	video_id	main_category	view_count	engagement_rate
51	cQO0IqMLBkk	Career Tips	1769537	0.0636
53	piWdncMDyNs	Career Tips	21579593	0.0377
93	NMuaxhPbP6Y	Career Tips	2467359	0.0826
100	b5BsywkETfU	Finance	29305157	0.0445
106	1OF53QNbMrE	Finance	8274423	0.0473
107	0GxSnWSp3VM	Finance	30094889	0.0320
114	RIOcs8stB6w	Finance	3676678	0.0361
117	AIOWO943HtM	Finance	6775348	0.0276
122	b_X9JYZKsuI	Finance	1787021	0.0239
124	T8msuRYVeRA	Finance	3958383	0.0184

Figure: Viral Outlier Table (IQR Results)

```
Number of Viral Shorts Detected (Z-score): 5
```

	video_id	main_category	view_count	engagement_rate	view_zscore
53	piWdncMDyNs	Career Tips	21579593	0.0377	3.227041
100	b5BsywkETfU	Finance	29305157	0.0445	4.477019
107	0GxSnWSp3VM	Finance	30094889	0.0320	4.604796
135	lKXiyApvKjI	Finance	41791423	0.0465	6.497267
253	lZ2hkZGzeGE	Study Skills	90474424	0.0434	14.374058

Figure: Extreme Outliers (Z-score Table)

7.7 Multivariate Insights

When combining multiple variables, category and duration together created the strongest performance patterns. Metadata elements such as tags or description length had near-zero predictive power. Categories such as Study Skills clustered tightly in the mid-duration, high-performance region.

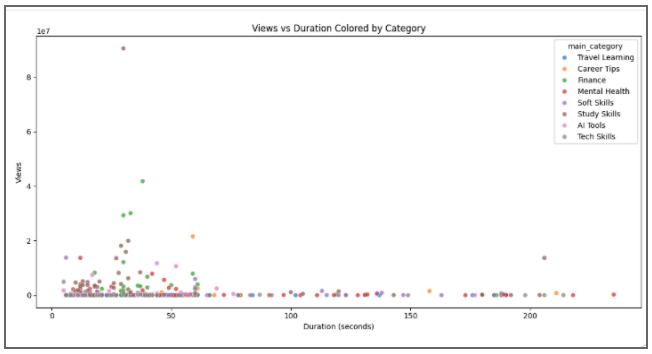


Figure: Duration vs Views by Category

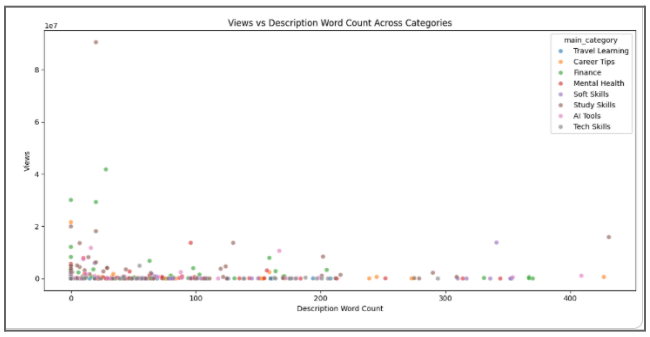


Figure: Duration vs Views by Category

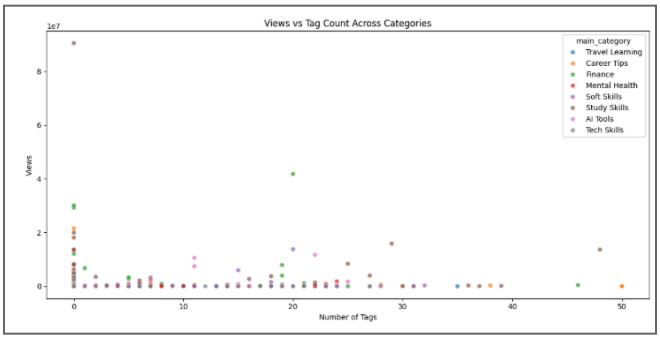


Figure: Tag Count vs Views by Category

7.8 Engagement Rate Analysis

Engagement rate shows how strongly viewers interact with a video compared to how many people watched it. The distribution showed that many videos had low engagement, while only a small group had very high engagement. Categories like Mental Health and AI Tools had fewer total views but much stronger interaction. This confirms that high views do not always mean high engagement.

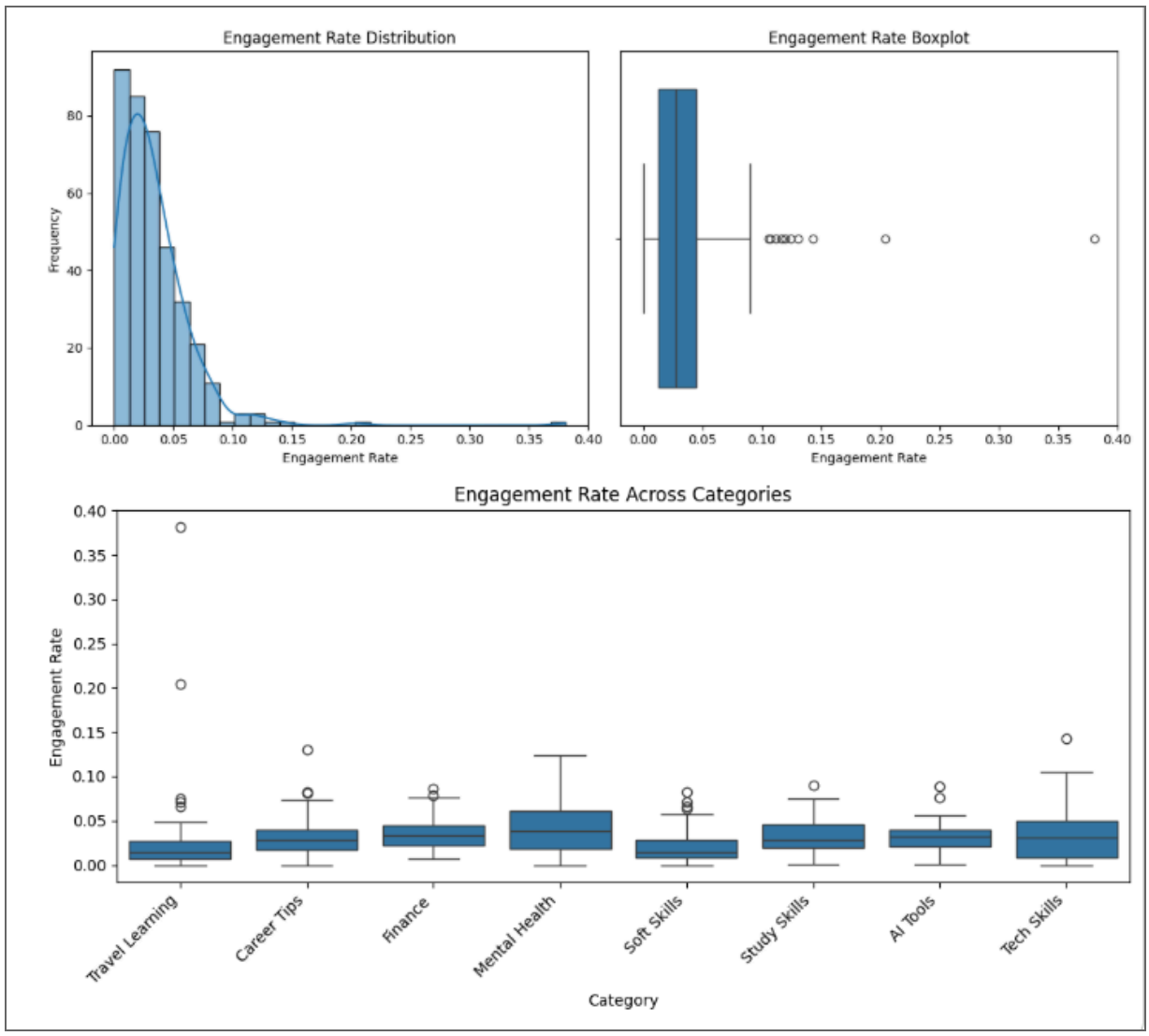


Figure: Engagement Rate Distribution Plot

8. CATEGORY IMPACT SCORING

Using normalized view and engagement values, categories were ranked:

1. Study Skills
2. Finance
3. Mental Health
4. AI Tools
5. Career Tips
6. Tech Skills
7. Travel Learning
8. Soft Skills

Study Skills ranked highest due to consistent high performance and stable engagement. Finance ranked second due to strong reach. Mental Health ranked high based on very high engagement quality, even with a smaller sample.

9. RECOMMENDATION SYSTEMS

9.1 Rule-Based Recommender

Based on EDA insights, the following recommendations were developed:

- Keep videos between 20 and 60 seconds
- Use short, clear titles (7–12 words)
- Keep descriptions concise
- Post during Afternoon or Evening
- Focus on categories like Study Skills, Finance, and AI Tools

These suggestions summarize the strongest patterns found in the data.

9.2 TF-IDF Content-Based Recommender

A simple content-based recommendation system was created using TF-IDF vectorization and cosine similarity. Titles and descriptions were combined into a text vector to find videos with similar themes. This demonstrates how platforms recommend related content based on text similarity.

10. Discussion

The findings suggest that structural choices—such as video duration, category selection, and posting time—significantly influence performance. Metadata quantity does not play a major role. Viral videos tend to follow similar structural patterns, proving that virality in learning Shorts is not random. The results align with viewer expectations in a short-form environment where quick value delivery is essential.

The combination of quantitative metrics and qualitative comment analysis provides a more complete view of viewer behavior. The scoring model and recommendation systems demonstrate how insights can be converted into practical tools for creators.

11. CONCLUSION

This research provides a clear understanding of how learning-focused YouTube Shorts behave and what drives their success. Concise duration, clear value, category relevance, and strategic timing were the strongest factors behind high performance. Metadata elements such as long descriptions or excessive tags do not affect visibility.

Viral videos follow predictable patterns and appear mostly in categories where viewers seek quick, practical knowledge. The results of this study offer valuable guidance for content creators and researchers and demonstrate an end-to-end analytical workflow suitable for academic and professional environments.

12. LIMITATIONS AND FUTURE WORK

The dataset did not include watch-time or retention data, which would have provided a deeper understanding of viewer attention. Demographic information such as age, country, or audience type was also unavailable. Viral thresholds depend on distribution properties and may vary by category.

Future improvements may include integrating retention curves, building dashboards, expanding the dataset globally, using advanced embedding-based recommenders, or forecasting which topics will trend next.