# YouTube Trending Analysis

Riya Desai
*Dept. of Computer Science*
*UNC Greensboro*
Greensboro, NC, U. S. A.
r_desai3@uncg.edu

Naveena Pokala
*Dept. of Computer Science*
*UNC Greensboro*
Greensboro, NC, U. S. A.
n_pokala@uncg.edu

Xiuwen Nong
*Dept. of Computer Science*
*UNC Greensboro*
Greensboro, NC, U. S. A.
x_nong@uncg.edu

***Abstract*—This project explores the dynamics of YouTube video popularity by performing a regression analysis to predict the number of days a video remains on the trending list. Using features such as views, likes, dislikes, comments, and category information collected on the first day of release, the study identifies key factors that influence a video's trending duration. The dataset, consisting of approximately 40,000 observations, enables the computation of derived metrics such as like ratios and comment ratios, which provide deeper insights into audience engagement and sentiment. The findings of this research offer practical value to content creators, allowing them to design more effective strategies for video production, title optimization, and audience targeting. By understanding the relationship between early performance indicators and trending longevity, creators can enhance the reach and impact of their content in a competitive digital environment.**

## I. Introduction

YouTube has emerged as one of the most dominant platforms for digital media consumption, with billions of users worldwide engaging daily across diverse categories such as entertainment, education, gaming, and news. The platform's "Trending" section plays a pivotal role in amplifying visibility, as it highlights videos that are rapidly gaining traction. For content creators, appearing in this section is not only a marker of success but also a gateway to exponential growth in audience reach and influence. Understanding the factors that determine how long a video remains trending is therefore of both academic and practical interest.

The dynamics of trending videos are complex, influenced by audience behavior, platform algorithms, and content characteristics. While raw metrics such as views, likes, dislikes, and comments provide direct indicators of engagement, they do not fully capture the nuances of audience sentiment or sustained popularity. Derived measures, such as the ratio of likes to views or comments to views, offer deeper insights into how viewers interact with content beyond passive consumption. These ratios can reflect enthusiasm, controversy, or community engagement, all of which contribute to a video's longevity on the trending list.

In this project, we approach the problem as a regression task: predicting the number of days a video remains trending based on its first-day performance metrics and categorical attributes. By focusing on the initial release period, we aim to identify early signals that are most predictive of long-term visibility. The dataset used for this study contains approximately 40,000 observations, providing both scale and diversity. Its richness allows us to engineer new features, such as engagement ratios, and to explore the interplay between quantitative metrics and categorical variables like video genre. Feature analysis is performed to determine which features most strongly influence trending duration. This systematic approach enables us to move beyond descriptive statistics toward actionable insights.

## II. Stage I

### A. Dataset Exploration

The dataset used in this study is sourced from the Kaggle "YouTube Trending Videos" dataset, which records daily snapshots of videos appearing on the trending list as shown in Figure 1. A key characteristic of this dataset is that a single video may appear on the trending list for multiple days, which results in multiple rows with the same *video_id* but different trending dates. Each row represents the performance metrics of the video on a specific day that it was trending.

To predict how many days a video remains on the trending list, it was necessary to aggregate these repeated entries. For each unique *video_id*, we counted the number of distinct dates it appeared in the trending dataset. This count was used to create the target variable *TrendingDays*, representing the total number of days a video trended.

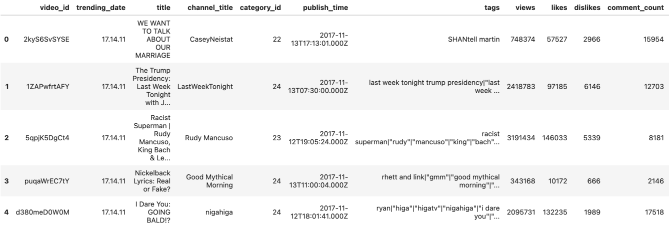
After aggregation and preprocessing, the dataset contains approximately 6500 video instances with features such as views, likes, dislikes, and comment counts recorded on the first day the video appeared in the trending list. Additional metadata, including category ID, category title, and the video's first trending timestamp, were also retained.

We engineered derived variables such as the Like Ratio and Comment Ratio from the raw engagement metrics to capture user sentiment and audience interaction strength

$$\text{Like Ratio} = \frac{\text{likes}}{\text{likes} + \text{dislikes}},$$

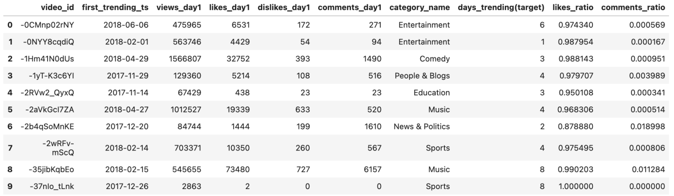
$$\text{Comment Ratio} = \frac{\text{comment\_count}}{\text{views}},$$

These engineered features, along with the aggregated target variable, form the basis for the regression analysis conducted in later stages of the project.



<table>
<tr><th></th><th>video_id</th><th>trending_date</th><th>title</th><th>channel_title</th><th>category_id</th><th>publish_time</th><th>tags</th><th>views</th><th>likes</th><th>dislikes</th><th>comment_count</th></tr>
<tr><td>0</td><td>2kyS6SvSYSE</td><td>17.14.11</td><td>WE WANT TO TALK ABOUT OUR MARRIAGE</td><td>CaseyNeistat</td><td>22</td><td>2017-11-13T17:13:01.000Z</td><td>SHANtell martin</td><td>748374</td><td>57527</td><td>2966</td><td>15954</td></tr>
<tr><td>1</td><td>1ZAPwfrtAFY</td><td>17.14.11</td><td>The Trump Presidency: Last Week Tonight with J...</td><td>LastWeekTonight</td><td>24</td><td>2017-11-13T07:30:00.000Z</td><td>last week tonight trump presidency|"last week ...</td><td>2418783</td><td>97185</td><td>6146</td><td>12703</td></tr>
<tr><td>2</td><td>5qpjK5DgCt4</td><td>17.14.11</td><td>Racist Superman | Rudy Mancuso, King Bach & Le...</td><td>Rudy Mancuso</td><td>23</td><td>2017-11-12T19:05:24.000Z</td><td>racist superman|"rudy"|"mancuso"|"king"|"bach"...</td><td>3191434</td><td>146033</td><td>5339</td><td>8181</td></tr>
<tr><td>3</td><td>puqaWrEC7tY</td><td>17.14.11</td><td>Nickelback Lyrics: Real or Fake?</td><td>Good Mythical Morning</td><td>24</td><td>2017-11-13T11:00:04.000Z</td><td>rhett and link|"gmm"|"good mythical morning"|"...</td><td>343168</td><td>10172</td><td>666</td><td>2146</td></tr>
<tr><td>4</td><td>d380meD0W0M</td><td>17.14.11</td><td>I Dare You: GOING BALD!?</td><td>nigahiga</td><td>24</td><td>2017-11-12T18:01:41.000Z</td><td>ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...</td><td>2095731</td><td>132235</td><td>1989</td><td>17518</td></tr>
</table>

Fig. 1. Initial Dataset



| | video_id | first_trending_ts | views_day1 | likes_day1 | dislikes_day1 | comments_day1 | category_name | days_trending(target) | likes_ratio | comments_ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0CMnp02rNY | 2018-06-06 | 475965 | 6531 | 172 | 271 | Entertainment | 6 | 0.974340 | 0.000569 |
| 1 | -0NYY8cqdiQ | 2018-02-01 | 563746 | 4429 | 54 | 94 | Entertainment | 1 | 0.987954 | 0.000167 |
| 2 | -1Hm41N0dUs | 2018-04-29 | 1566807 | 32752 | 393 | 1490 | Comedy | 3 | 0.988143 | 0.000951 |
| 3 | -1yT-K3c6YI | 2017-11-29 | 129360 | 5214 | 108 | 516 | People & Blogs | 4 | 0.979707 | 0.003989 |
| 4 | -2RVw2_QyxQ | 2017-11-14 | 67429 | 438 | 23 | 23 | Education | 3 | 0.950108 | 0.000341 |
| 5 | -2aVkGcl7ZA | 2018-04-27 | 1012527 | 19339 | 633 | 520 | Music | 4 | 0.968306 | 0.000514 |
| 6 | -2b4qSoMnKE | 2017-12-20 | 84744 | 1444 | 199 | 1610 | News & Politics | 2 | 0.878880 | 0.018998 |
| 7 | -2wRFv-mScQ | 2018-02-14 | 703371 | 10350 | 260 | 567 | Sports | 4 | 0.975495 | 0.000806 |
| 8 | -35jibKqbEo | 2018-02-15 | 545655 | 73480 | 727 | 6157 | Music | 8 | 0.990203 | 0.011284 |
| 9 | -37nlo_tLnk | 2017-12-26 | 2863 | 2 | 0 | 0 | Sports | 8 | 1.000000 | 0.000000 |

Fig. 2. Dataset After Aggregating and Integrating The Derived Variables

### B. Feature Exploration - Missing Value Analysis

A thorough missing value analysis was conducted on all variables. Core engagement metrics such as views, likes, dislikes, and comments contained no missing values, ensuring reliability for modeling.

Derived features such as Like Ratio and Comment Ratio exhibited a small percentage of missingness (about 0.4%), primarily caused by division by zero or missing denominators. These entries were handled appropriately by removing or imputing cases where raw values rendered the ratios undefined.

Before handling missing values, the dataset contained 6,351 rows and 10 columns. After removing rows with missing values, the dataset was reduced to 6,324 rows while retaining the same 10 columns. This means that a total of 27 rows (approximately 0.42%) contained missing values and were removed during the cleaning process.

TABLE I
FEATURE SUMMARY AND MISSING VALUE PERCENTAGES FOR THE YOUTUBE TRENDING DATASET.

| Feature | Missing % |
|---|---|
| video_id | 0 |
| first_trending_ts | 0 |
| views_day1 | 0 |
| likes_day1 | 0 |
| dislikes_day1 | 0 |
| comments_day1 | 0 |
| likes_ratio | 0.4 |
| comments_ratio | 0.4 |
| category_id | 0 |
| category_title | 0 |
| TrendingDays (Target) | 0 |

### C. Feature Exploration - Outlier Analysis

Outlier analysis was performed for all numerical features except identifiers (e.g., *video_id*) and categorical variables. Using log-normalization and Z-score calculations, extreme values were identified in views, likes, and comment counts.

Given the nature of YouTube data, large values often represent genuine viral phenomena rather than noise. Therefore, real engagement outliers were retained, while transformations were applied when necessary to stabilize variance. Ratio-based variables were also inspected, with extreme values preserved if they were valid.

Before handling missing values, the dataset contained 6,324 rows and 22 columns. After removing rows with missing values, the dataset was reduced to 5,885 rows, with the same 22 columns retained. This shows that 439 rows contained missing values and were removed during preprocessing, which corresponds to approximately 6.94% of the dataset.

### D. Categorical Variable

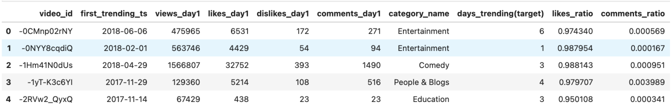
The dataset includes categorical variables such as Category ID and Category Title, which provide context about the content type (e.g., Music, Gaming, News & Politics). These categories were standardized and validated using the JSON category mapping file provided with the original dataset.

A distributional analysis revealed category imbalances, with some categories overrepresented. These imbalances were documented for later modeling, particularly for encoding techniques such as one-hot encoding.

### E. Processed Dataset

After addressing missing values, handling inconsistencies, formatting timestamps, and computing derived variables, the final cleaned dataset was saved as `newdata.csv`. This dataset includes:

- Views, likes, dislikes, and comments on Day 1
- Like Ratio and Comment Ratio
- Category ID and category title
- Trending timestamp
- Target variable: *TrendingDays*



| | video_id | first_trending_ts | views_day1 | likes_day1 | dislikes_day1 | comments_day1 | category_name | days_trending(target) | likes_ratio | comments_ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0CMnp02rNY | 2018-06-06 | 475965 | 6531 | 172 | 271 | Entertainment | 6 | 0.974340 | 0.000569 |
| 1 | -0NYY8cqdiQ | 2018-02-01 | 563746 | 4429 | 54 | 94 | Entertainment | 1 | 0.987954 | 0.000167 |
| 2 | -1Hm41N0dUs | 2018-04-29 | 1566807 | 32752 | 393 | 1490 | Comedy | 3 | 0.988143 | 0.000951 |
| 3 | -1yT-K3c6YI | 2017-11-29 | 129360 | 5214 | 108 | 516 | People & Blogs | 4 | 0.979707 | 0.003989 |
| 4 | -2RVw2_QyxQ | 2017-11-14 | 67429 | 438 | 23 | 23 | Education | 3 | 0.950108 | 0.000341 |

Fig. 3. Processed Dataset

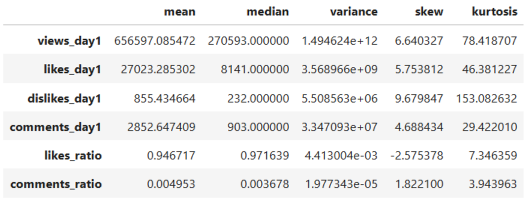
This structured and cleaned dataset is now ready for Stage II regression modeling.

## III. STAGE II

### A. Descriptive Statistics

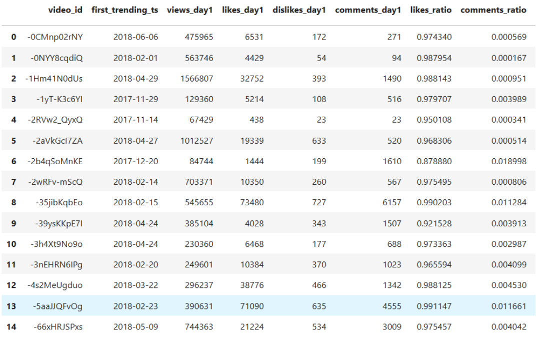
In the first stage of our analysis, we conducted a comprehensive descriptive statistical exploration of the dataset. As students approaching this project, our goal was to build a clear initial understanding of how the variables behave before applying more advanced statistical or machine learning methods. We began by calculating key descriptive measures such as the mean, median, variance, standard deviation, minimum and maximum values for each numerical feature. These statistics helped reveal the central tendency and variability of the data.

Additionally, we examined the presence of outliers by looking at the interquartile range and creating boxplots. This allowed us to identify variables that exhibited extreme values, which could potentially influence later modeling steps. We also generated summary tables and visualizations, such as histograms and bar charts, to better understand the distribution patterns and to detect skewness or irregularities. Overall, the descriptive statistics phase provided us with a foundational overview of the dataset and helped us determine which variables required normalization, cleaning, or further investigation in subsequent stages.



| | mean | median | variance | skew | kurtosis |
|---|---|---|---|---|---|
| **views_day1** | 656597.085472 | 270593.000000 | 1.494624e+12 | 6.640327 | 78.418707 |
| **likes_day1** | 27023.285302 | 8141.000000 | 3.568966e+09 | 5.753812 | 46.381227 |
| **dislikes_day1** | 855.434664 | 232.000000 | 5.508563e+06 | 9.679847 | 153.082632 |
| **comments_day1** | 2852.647409 | 903.000000 | 3.347093e+07 | 4.688434 | 29.422010 |
| **likes_ratio** | 0.946717 | 0.971639 | 4.413004e-03 | -2.575378 | 7.346359 |
| **comments_ratio** | 0.004953 | 0.003678 | 1.977343e-05 | 1.822100 | 3.943963 |

Fig. 4. Descriptive Statistics



| | video_id | first_trending_ts | views_day1 | likes_day1 | dislikes_day1 | comments_day1 | likes_ratio | comments_ratio |
|---|---|---|---|---|---|---|---|---|
| **0** | -0CMnp02rNY | 2018-06-06 | 475965 | 6531 | 172 | 271 | 0.974340 | 0.000569 |
| **1** | -0NYY8cqdiQ | 2018-02-01 | 563746 | 4429 | 54 | 94 | 0.987954 | 0.000167 |
| **2** | -1Hm41N0dUs | 2018-04-29 | 1566807 | 32752 | 393 | 1490 | 0.988143 | 0.000951 |
| **3** | -1yT-K3c6YI | 2017-11-29 | 129360 | 5214 | 108 | 516 | 0.979707 | 0.003989 |
| **4** | -2RVw2_QyxQ | 2017-11-14 | 67429 | 438 | 23 | 23 | 0.950108 | 0.000341 |
| **5** | -2aVkGcI7ZA | 2018-04-27 | 1012527 | 19339 | 633 | 520 | 0.968306 | 0.000514 |
| **6** | -2b4qSoMnKE | 2017-12-20 | 84744 | 1444 | 199 | 1610 | 0.878880 | 0.018998 |
| **7** | -2wRFv-mScQ | 2018-02-14 | 703371 | 10350 | 260 | 567 | 0.975495 | 0.000806 |
| **8** | -35jibKqbEo | 2018-02-15 | 545655 | 73480 | 727 | 6157 | 0.990203 | 0.011284 |
| **9** | -39ysKKpE7I | 2018-04-24 | 385104 | 4028 | 343 | 1507 | 0.921528 | 0.003913 |
| **10** | -3h4Xt9No9o | 2018-04-24 | 230360 | 6468 | 177 | 688 | 0.973363 | 0.002987 |
| **11** | -3nEHRN6IPg | 2018-02-20 | 249601 | 10384 | 370 | 1023 | 0.965594 | 0.004099 |
| **12** | -4s2MeUgduo | 2018-03-22 | 296237 | 38776 | 466 | 1342 | 0.988125 | 0.004530 |
| **13** | -5aaJJQFvOg | 2018-02-23 | 390631 | 71090 | 635 | 4555 | 0.991147 | 0.011661 |
| **14** | -66xHRJSPxs | 2018-05-09 | 744363 | 21224 | 534 | 3009 | 0.975457 | 0.004042 |

Fig. 5. Descriptive Statistics

### B. Distribution Fitting

After developing a general understanding of the data through descriptive statistics, we proceeded to perform distribution fitting. The purpose of this section was to identify the theoretical probability distribution that best describes each variable. As students, this step helped us understand not only the shape of the data but also which statistical methods would be appropriate later on.

We began by plotting histograms and density curves for important variables to visually inspect whether they resembled well-known distributions such as the normal distribution, log-normal distribution, exponential distribution, or others. Then, we fitted different theoretical distributions and compared their goodness-of-fit using criteria such as the Kolmogorov–Smirnov test, the Anderson–Darling test, or log-likelihood values.

Understanding the fitted distribution was especially useful for deciding when to apply parametric tests versus non-parametric tests in later hypothesis evaluation. For example, if a variable did not follow a normal distribution, we considered using Spearman's correlation instead of Pearson's. Through this step, we gained insights into the underlying probabilistic behavior of the dataset and ensured that our future statistical conclusions were grounded in appropriate assumptions.

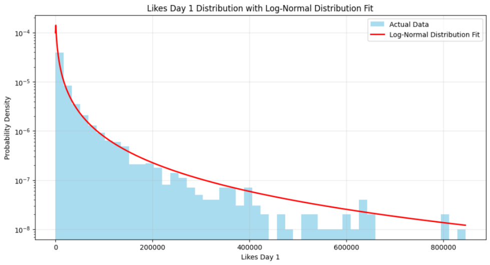


Fig. 6. Distribution Fitting

### C. Point Estimate

The next step in our project involved computing point estimates to approximate population-level parameters based on our sample data. Since we only had access to a limited dataset, point estimation allowed us to infer the most likely values of key parameters, such as the population mean or population proportion, using observed sample statistics.

We calculated point estimates for various numerical variables, focusing mainly on sample means as estimates of population means. In some cases, we also calculated sample proportions for categorical features. As students, this step helped us clearly understand how sample information can be used to make broader generalizations about an entire population.

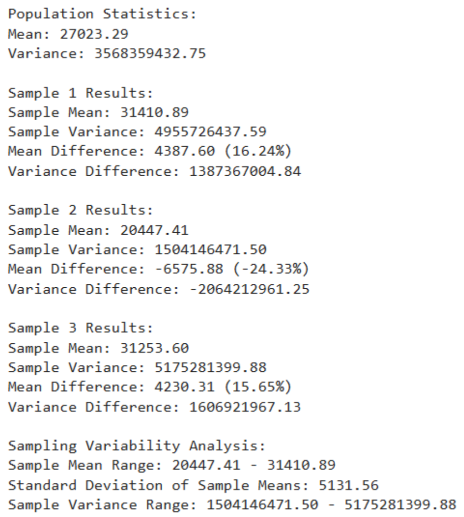
Moreover, point estimates served as the foundation for later inferential statistics. For example, the sample mean and sample variance computed in this section were later used in constructing confidence intervals and conducting hypothesis tests. This process reinforced the idea that statistical inference always begins with careful estimation, and it reminded us that the reliability of these estimates depends heavily on sample size and data quality.

### D. Correlation Analysis

Finally, we conducted a detailed correlation analysis to examine the relationships between pairs of variables. This part of the project was particularly important because understanding these relationships helped us determine which factors were most relevant for predicting or explaining the patterns in the dataset.

We computed both Pearson's correlation coefficients and Spearman's rank correlation coefficients. Pearson's correlation allowed us to evaluate the strength of linear relationships between continuous variables, while Spearman's correlation helped us capture monotonic relationships that may not be strictly linear or where the variables did not follow a normal distribution.

To complement the numerical results, we generated correlation heatmaps and pairwise scatterplots. These visualizations



```
Population Statistics:
Mean: 27023.29
Variance: 3568359432.75

Sample 1 Results:
Sample Mean: 31410.89
Sample Variance: 4955726437.59
Mean Difference: 4387.60 (16.24%)
Variance Difference: 1387367004.84

Sample 2 Results:
Sample Mean: 20447.41
Sample Variance: 1504146471.50
Mean Difference: -6575.88 (-24.33%)
Variance Difference: -2064212961.25

Sample 3 Results:
Sample Mean: 31253.60
Sample Variance: 5175281399.88
Mean Difference: 4230.31 (15.65%)
Variance Difference: 1606921967.13

Sampling Variability Analysis:
Sample Mean Range: 20447.41 - 31410.89
Standard Deviation of Sample Means: 5131.56
Sample Variance Range: 1504146471.50 - 5175281399.88
```

Fig. 7. Point Estimate

made it easier for us as students to identify clusters of variables that moved together, as well as any unexpected relationships or anomalies.

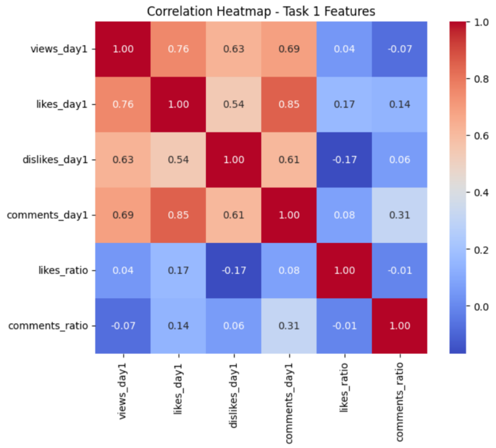


Fig. 8. correlation heatmap

# IV. Stage III

## A. *Hypothesis Testing*

In this project, we worked on testing three hypothesis questions as well as verifying the hypothesis using appropriate tests. For each question, we define the null and alternative hypothesis. The tests we have used in this project include: Pearson's correlation, spearman correlation, and Anova test Below are the hypothesis and their respective verifications.

*1) Does a higher number of views on the first day lead to a higher likes ratio:* **Null Hypothesis** $(H_0)$**:** There is no relationship between the number of views on the first day $views\_day1$ and the $likes\_ratio$. (The number of first-day views does not significantly affect the likes ratio.) **Alternative** $(H_a)$**:** There is a relationship between the number of views on the first day $views\_day1$ and $likes\_ratio$. (Videos with higher first-day views have a significantly different likes ratio (could be higher or lower).

**Tests used to verify:** For this hypothesis Pearson's correlation as well as Spearman correlation were used. Pearson's correlation test was used as both features are continuous. The correlation value and p-value were 0.043 and 0.0009 respectively. The results from spearman correlation was 0.073 and p-value was 0.000. Two tests gave p-value lower than the significance level of 0.05, proving that the null hypothesis $H_0$ will be rejected.

*2) Videos with a higher comments ratio are likely to trend for a longer duration:* **Null Hypothesis** $(H_0)$**:** There is no relationship between a video's comments ratio and how long it trends. **Alternative** $(H_a)$**:** There is a positive relationship, videos with a higher comments ratio tend to trend for a longer duration

**Tests used to verify:** For this hypothesis, Spearman correlation was used. The Spearman correlation test is the most appropriate statistical method because we are examining the relationship between two continuous variables, comments ratio and trending duration. The Spearman correlation is preferred over Pearson correlation because YouTube metrics such as views, comments, and engagement ratios typically exhibit highly skewed distributions with substantial outliers, particularly when viral videos are included in the dataset. The correlation value and the p-value were 0.1703 and $1.5584 \times 10^{-39}$ respectively. The test gave a p-value lower than the significance level of 0.05, proving that the null hypothesis $H_0$ will be rejected.

*3) Different categories have significantly different average trending durations:* **Null Hypothesis** $(H_0)$**:** There is no significant difference in the average trend duration between categories. **Alternative** $(H_a)$**:** There is a significant difference in the average duration of trending between categories.

**Tests used to verify:** Since we are comparing the means of a continuous variable (trending duration) across multiple categories, the right test is One-Way ANOVA (Analysis of Variance). The test checks if the mean trending durations differ significantly among categories. The p-value of this method is $1.3733 \times 10^{-35}$. The test gave a p-value lower than the significance level of 0.05, proving that the null hypothesis $H_0$ will be rejected.

## B. *Feature Selection*

We first checked the features which were most aligned with our target variable. using Pearson correlation we checked which of the features are most correlated with the target variable. From Fig.4 we can see that $likes\_day1$ has the highest correlation with target.

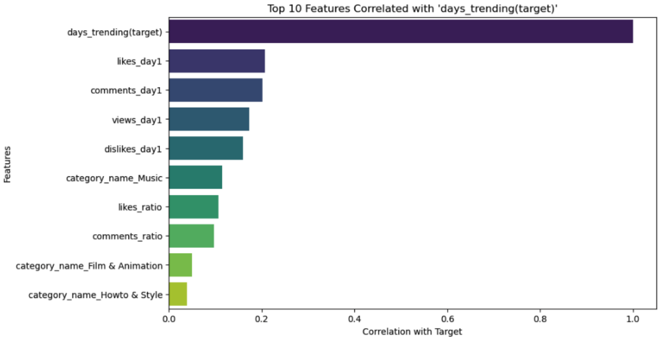
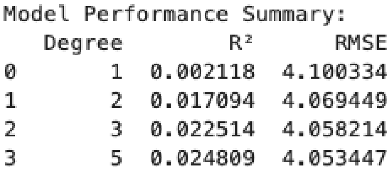


Fig. 9. Feature Selection

### C. Machine Learning-Baseline

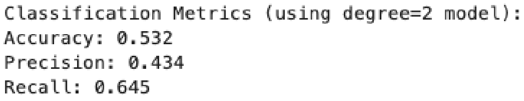
Now that we have completed pre-processing and hypothesis testing. We move on to training machine learning models on our cleaned dataset. We first begin with our baseline model linear regression to predict the duration of days for which videos would trend on YouTube. The linear regression model was fitted on the dataset using different degrees 1,2,3,5.The $R^2$ and RMSE performance for all the linear regression models is as shown in Fig.5.



```
Model Performance Summary:
   Degree        R²       RMSE
0       1  0.002118   4.100334
1       2  0.017094   4.069449
2       3  0.022514   4.058214
3       5  0.024809   4.053447
```

Fig. 10. Linear Regression Results

The degree 2 linear regression model provided the best trade-off between predictive performance and computational cost. The R² increased from degree 1 to degree 2, while higher degrees (3 and 5) achieved slightly better R² scores but offered negligible improvements in RMSE, increasing the risk of overfitting and model complexity. Therefore, the degree 2 model (quadratic) was selected as the baseline for further analysis.Next we check if the problem was converted to classification what would be the accuracy, precision and recall values. The results are as shown in Fig.11.



```
Classification Metrics (using degree=2 model):
Accuracy: 0.532
Precision: 0.434
Recall: 0.645
```

Fig. 11. Linear Classification Results

### D. Machine Learning-other models

Now that we have established a baseline using linear regression, we analyze our data using Decision Tree, Random Forest and Support Vector Machine (SVM).

*1) Decision Tree:* The parameters used for this model are $max\_depth$ and $min\_sample\_split$, defined as 5 and 20 respectively. This model was evaluated using $R^2$ and RMSE. The results are as shown in Table II. We plotted the feature importance using decision tree regressor, as shown in Fig.6. We noted that $likes\_day1$ has the highest impact on the regression, basically it is the most significant feature of our dataset.

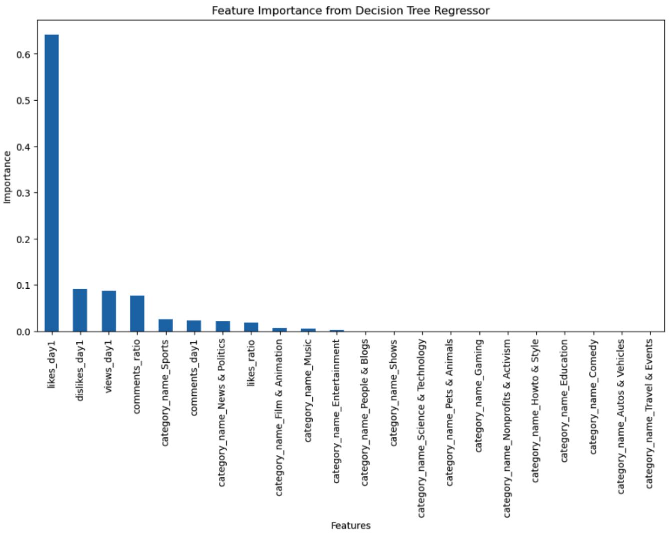


Fig. 12. Feature Importance

*2) SVM:* The next model used to analyze our data is SVM. Since SVM requires feature scaling we first applied StandardScaler() to the features. The SVM was trained using a radial basis kernel. This model was evaluated using $R^2$ and RMSE. The results are as shown in Table II.

*3) Random Forest:* The next model used to analyze our data is Random Forest. The parameters used for this model are $max\_depth$ and $n\_estimators$, defined as None and 100 respectively. This model was evaluated using $R^2$ and RMSE. The results are as shown in Table II.

### E. Model Comparison

We compare all the models. Based on the results of the four regression methods shown in Table II, the analysis is as follows: The Decision Tree Regressor has a RMSE of 16.27 and an $R^2$ of 0.0343, indicating that the model can hardly explain the variance in the data, with relatively high error, making it the worst performer. The SVM regression has an RMSE of 15.76 and an $R^2$ of 0.0643, with still large errors and an $R^2$ close to zero, showing limited fitting ability; it performs slightly better than a single decision tree but remains inadequate. For Polynomial Regression (Degree 1–5), even as the polynomial degree increases, the highest $R^2$ reaches only about 0.0248 and the lowest RMSE is approximately 4.053, indicating that increasing model complexity does not significantly improve fitting.

In contrast, the Random Forest Regressor achieves an MSE of 9.8101 and an R² of 0.4177, clearly outperforming the other

TABLE II
MODEL PERFORMANCE COMPARISON

| Model | MSE | $R^2$ |
|---|---|---|
| Decision Tree Regressor | 16.2703 | 0.0343 |
| Random Forest Regressor | 9.8101 | 0.4177 |
| SVM Regression | 15.7650 | 0.0643 |
| Polynomial Regression (Deg 1–5) | 15.986414–17.033187 | 0.002–0.0248 |

TABLE III
PERFORMANCE COMPARISON OF RANDOM FOREST MODELS

| Model Version | $R^2$ (mean ± std) | RMSE (mean ± std) |
|---|---|---|
| Baseline Random Forest | $0.3847 \pm 0.0224$ | $3.1821 \pm 0.0532$ |
| Tuned Random Forest | $0.3596 \pm 0.0135$ | $3.2468 \pm 0.0385$ |

models by explaining about 42 percent of the variance while keeping errors the lowest. Its advantage lies in leveraging an ensemble of multiple decision trees to capture nonlinear relationships and reduce overfitting, making Random Forest the best-performing model among the four.

## V. STAGE IV

### A. Model Validation

Our model is random forest as explained in the previous section. We now use 5 fold cross validation to check whether this model generalizes well on different splits of data. the mean and std deviation results of $R^2$ and RMSE are as shown in Table III. We then performed hyperparameter tuning on this model. The parameters considered are $n\_estimators$: [100, 200, 300], $max\_depth$: [None, 10, 20], $min\_samples\_split$: [2, 5], and $max\_features$: ['sqrt', 'log2'] The model selected the best parameters to be $n\_estimators$: 300, $max\_depth$: None, $min\_samples\_split$: 2, and $max\_features$:'sqrt. We trained another random forest which we name tuned RF. The results of this model is shown in Table III. From this experiment we infer that Hyperparameter tuning did not improve the predictive performance of the Random Forest model. Although, the tuned model showed slightly lower R² and higher RMSE compared to the baseline, its standard deviation across folds decreased, indicating improved stability. This suggests that the baseline Random Forest configuration already captures most of the predictive structure in the dataset. Even though, the tuned RF results show stability due to reduced std, the baseline RF results have comparable std with greater predictive capabilities.

After applying $k$-fold cross–validation to the selected model, we observed the mean and standard deviation of the evaluation metrics across the folds as shown in Table III. Cross–validation provides a more reliable estimate of model performance, as it reduces the variance associated with using a single train–test split. Overall, the performance differences before and after cross–validation were relatively small. The cross–validated mean metrics were close to the original hold–out test results, indicating that the model is reasonably stable. The standard deviations across folds were also low, showing that the model generalizes well to unseen data and does not heavily depend on any specific subset of the training set.

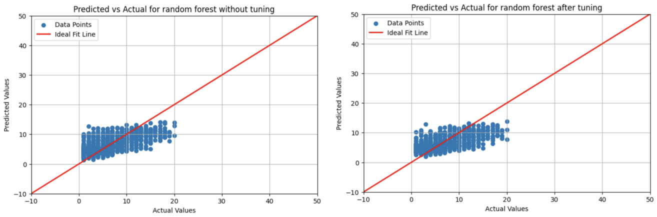


Fig. 13. Regression : Predicted vs Actual values

If there were slight performance drops after cross–validation, this is expected because the model is evaluated more rigorously. A small reduction indicates that the original hold–out estimate may have been slightly optimistic, but the model still performs consistently. Based on the cross–validated metrics and stability across folds, we conclude that the model generalizes well and achieves satisfactory performance for the given task.

### B. Model Visualization

#### 1. Predicted vs Actual Plot (Regression)

The predicted versus actual value plots demonstrate the random forest model's predictive performance before and after hyperparameter tuning. In both plots, the red diagonal line represents the ideal scenario where predictions perfectly match actual values. Before tuning, the data points show substantial deviation from the ideal fit line, with predictions clustering in vertical bands and exhibiting considerable scatter, particularly in the range of 5 to 15 on the actual values axis. This scatter indicates inconsistent prediction accuracy across different ranges of the target variable. After tuning, the predictions align more closely with the ideal fit line, showing reduced scatter and better overall agreement between predicted and actual values. The improved alignment suggests that hyperparameter optimization successfully enhanced the model's ability to capture the underlying patterns in the data, resulting in more reliable predictions across the entire range of target values.

#### 2. Residual Plot (Regression)

The residual plots reveal important patterns in prediction errors both before and after model tuning. Before tuning, the residuals display a pronounced funnel-shaped pattern, with error variance increasing substantially as predicted values rise from 6 to 12. The residuals range from approximately $-10$ to $+11$, with systematic patterns suggesting the model struggles to maintain consistent accuracy across different prediction ranges. After tuning, the residual distribution shows notable improvement, with errors more evenly distributed around the zero line and reduced overall magnitude. The pattern is less pronounced, though not entirely eliminated, and the residuals are more randomly scattered, which is desirable in regression models. This improvement indicates that the tuned model

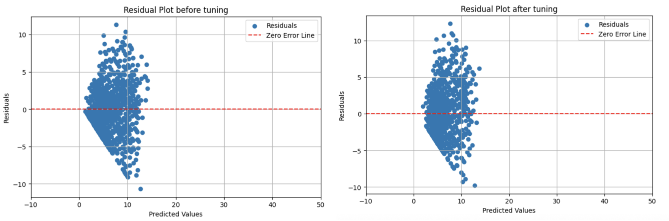


Fig. 14. Regression : Residual Plot

produces more reliable predictions with reduced systematic bias, though some variance in prediction errors persists across the range of predicted values.

### C. Interactive

In this section of the analysis, we focus on building a machine learning model capable of predicting the number of days a YouTube video stays on the trending charts. As a student team, our goal was to improve prediction accuracy while maintaining model stability, and to use various visualization methods to gain a deeper understanding of model performance. First, we performed a structural check on the cleaned data, confirming the number of features, the statistical distribution of the target variable, and ensuring there were no outliers affecting the modeling. Then, we split the data into training and test sets and chose Random Forest Regressor as the core model. This model was chosen because it demonstrates high reliability in handling non-linear relationships and complex interactive features.

During the model training phase, we set key parameters including the number of trees, maximum depth, and minimum number of leaf nodes to achieve a balance between bias and variance. After model training, we performed predictions on the test set and evaluated model performance using R² and RMSE metrics. Furthermore, we used 5-fold cross-validation to test the model's stability and calculated the average R² and RMSE across folds, ensuring that the model performs well not only in a single split but also maintains consistent performance across different data partitions.

To help us understand model behavior more intuitively from a student's perspective, we constructed several interactive visualizations, including scatter plots of actual and predicted values, feature importance ranking plots, residual analysis plots, and a comprehensive performance dashboard. These visualizations helped us identify patterns in prediction errors, the features the model relied on most, and whether the error distribution exhibited systematic bias. Among these, the top-ranking important features (such as high-engagement-related indicators) contributed most significantly to the predictions, providing a reliable basis for subsequent conclusion analysis.

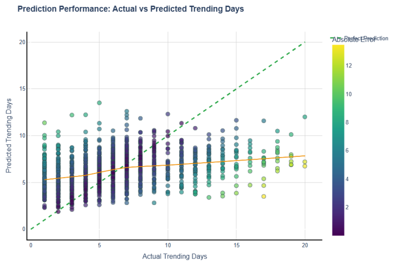


Fig. 15. Interactive 1

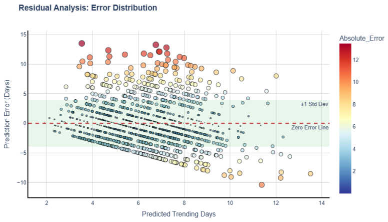


Fig. 16. Interactive2

Finally, we summarized the model's performance from multiple perspectives using an Executive Dashboard, including trend lines, the contributions of the top 8 key features, and the percentage of each error range. This enabled us to summarize the overall performance in a data-driven manner, identifying the model's strengths and potential limitations. Overall, the entire modeling process not only built a well-performing and stable prediction model but also helped us gain a deeper understanding of data and model behavior from multiple dimensions, providing a clear direction for subsequent discussion and improvement.

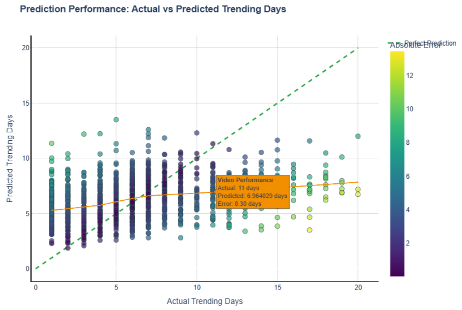


Fig. 17. Interactive3

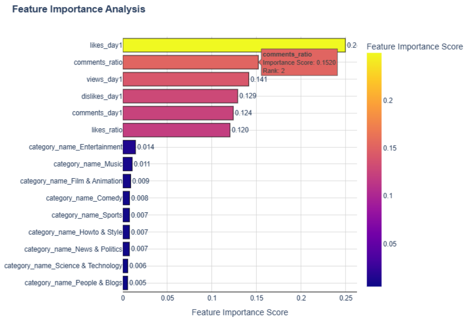


Fig. 18. Interactive4