Final Project Report

Group Members:
Melanie Yu
Yunhan Luo
Ani Mcaskill
Owen Sweetman

DS 3000
Dr. Mohit Singhal

Fall 2025

Abstract

This project studies how video length relates to audience engagement on YouTube using a dataset of technical videos focused on software engineering, artificial intelligence, and machine learning. Engagement is measured using likes and comments relative to total views. We found that longer videos tend to receive higher engagement, but only up to a certain point. After several minutes, the benefit of increasing video length begins to level off. A nonlinear model explained about half of the variation in engagement, while a simple linear model performed much worse. These results suggest that moderate-length videos may be most effective for maximizing viewer interaction.

Introduction

YouTube has become one of the most influential platforms for sharing educational and technical content, especially in areas such as artificial intelligence and software engineering. For creators in these fields, understanding what drives audience engagement is an important step toward producing effective and impactful content. Engagement metrics such as likes, comments, and views are commonly used to measure how viewers interact with videos, but the factors that influence these metrics are not always clear.

The primary goal of this project is to investigate whether measurable features of YouTube videos can help explain or predict audience engagement. In particular, this study focuses on the relationship between video duration and engagement rate. By analyzing real-world data from a specific technical niche, this project aims to provide insight into how video length may influence viewer interaction and content performance.

Data Description

The dataset for this project was collected using the YouTube Data API v3 by querying a targeted technology and marketing focused channel. The raw dataset consisted of video-level metadata returned in JSON format and converted into a structured pandas DataFrame for processing. The primary attributes collected for each video include:

- `video_id`: Unique identifier assigned to each YouTube video
- `title`: Video title provided by the content creator
- `published`: Date and time the video was uploaded
- `views`: Total number of video views
- `likes`: Total number of likes received
- `comments`: Total number of comments received
- `duration`: Log-transformed video length in seconds (originally ISO 8601 format)

- `tags`: List of content tags associated with the video

Data cleaning involved converting video durations from YouTube's ISO 8601 format into numeric seconds using the `isodate` library, followed by a logarithmic transformation to reduce skewness. Engagement metrics were verified and cast into consistent numeric formats. Videos with zero views were excluded to prevent undefined engagement rate values. The cleaned dataset was then exported as a CSV file for reproducibility and further analysis.

Feature engineering was performed to support our research questions. The primary derived feature was the engagement rate, which was calculated as:

$$\text{Engagement Rate} = \frac{\text{Likes} + \text{Comments}}{\text{Views}}$$

Video duration was selected as the main predictor variable used in both regression models. These features allowed us to directly evaluate how content length influences user interaction.

Exploratory data analysis was conducted using Matplotlib and Seaborn. Scatterplots, regression plots, and log-scaled visualizations were used to identify trends between duration, views, and engagement rate. Outliers were filtered where necessary to improve interpretability. Finally, the processed dataset was prepared for machine learning by constructing both NumPy-based linear regression models and polynomial regression models using scikit-learn.

Methods

Ordinary Least Squares (OLS) Linear Regression Model

The first model fits the equation

$$y = B_0 + B_1 x,$$

where x is equal to the duration of the video, y is equal to the engagement rate, B_0 is the intercept, and B_1 is the slope. The parameters are estimated using the closed-form OLS solution

$$\hat{B} = (X^T X)^{-1} X^T Y$$

OLS is appropriate in this case because it is the most interpretable and widely used baseline for continuous outcomes. It allows us to quantify the directional effect of duration on engagement. There are four key OLS regression assumptions: the true relationship between duration and engagement must be approximately linear, the variance of errors should be constant for all durations, each video is assumed independent of the others, and the residuals should follow a bell-shaped normal distribution. In that order, they are termed the linearity, homoscedasticity, independence, and normality assumptions.

There are several potential pitfalls with the OLS model. If the actual relationship between duration and engagement is non-linear, OLS will systematically misestimate engagement the model functions based off of the linearity assumption. OLS is also sensitive to outliers because it minimizes the sum of squared errors, disproportionately penalizing large outliers. In this case, that means that very long or short videos could heavily influence the slope.

Finally, if there is heteroscedasticity in the dataset, OLS underestimates uncertainty, overstates significance, and appears to be more confident than it actually is.

Polynomial Regression Model

This second model expands the feature space using

$$x, x^2.$$

The model then becomes

$$y = B_0 + B_1 x + B_2 x^2.$$

Polynomial regression is still a linear model in its parameters, which are solved using OLS, but is nonlinear in shape.

The polynomial regression model is appropriate because previous exploratory plots from the OLS model suggested that engagement is not entirely linear: it may rise at certain durations, then plateau or decline at others. A degree-2 polynomial is flexible enough to capture curvature without overfitting, making it appropriate for fitting this relationship.

Polynomial regression inherits the same assumptions as OLS as well as adds two: the underlying relationship between engagement and duration is assumed to be smooth and continuous, and degree-2 must be just enough to capture curvature without overfitting. These are termed smoothness and degree adequacy.

Despite capturing curvature, there are still several potential pitfalls with the polynomial regression model. Increasing the degree of the polynomial increases the risk of overfitting, which means that extrapolation becomes unreliable and the curve outside of the data range can behave unrealistically. Finally, multicollinearity between x and x^2, meaning that they carry redundant information between them, can inflate the variance of parameter estimates.

Ultimately, the polynomial regression model is more valuable in the context of looking at YouTube video duration and engagement because it provides a more realistic structure for modeling user behavior. This is due to the fact that the relationship between them is not linear: extremely short videos and long videos often underperform, whereas average durations may maximize engagement. Thus, the degree-2 polynomial can better represent these curved tendencies.

The OLS linear regression model provides a simple, interpretable baseline that describes the direction and magnitude of duration's effect on engagement. The polynomial regression model tests whether nonlinear patterns provide a better explanation of engagement. Taken together, the models provide both interpretability and flexibility. Ultimately, if both models point to similar conclusions then they reinforce each other, and if the second model performs significantly better it indicates that engagement behaves nonlinearly with respect to duration.

Results

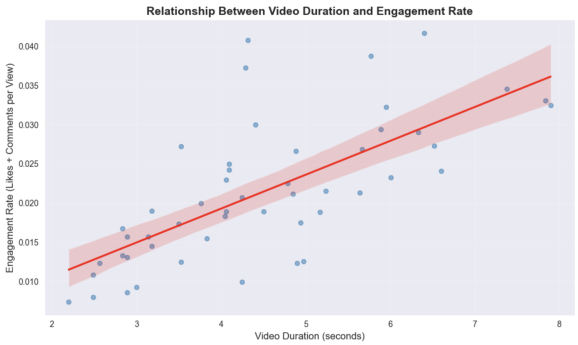

Figure 1: Image 1

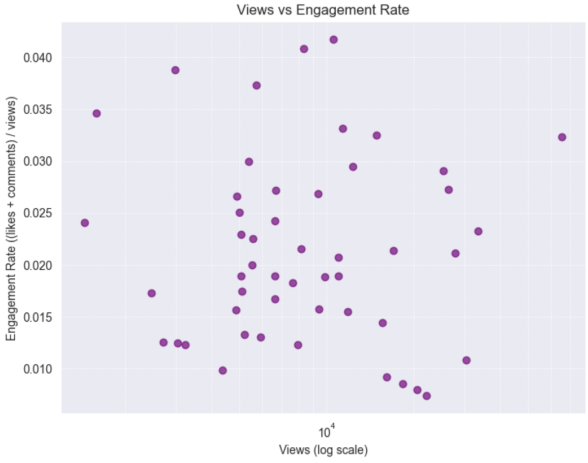

Figure 2: Image 2

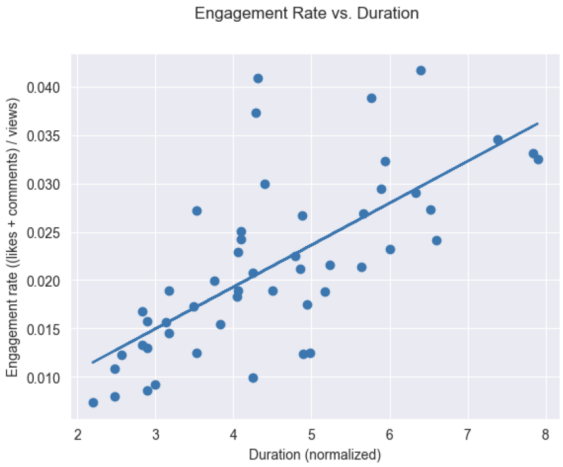

Figure 3: Image 3

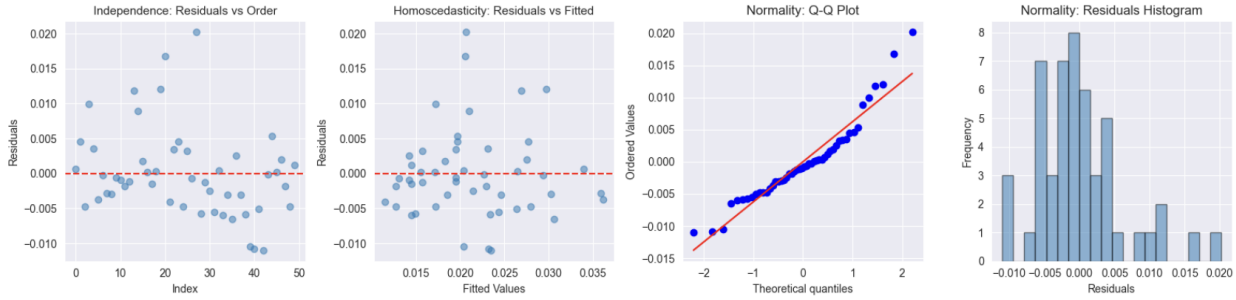

Figure 4: Image 4

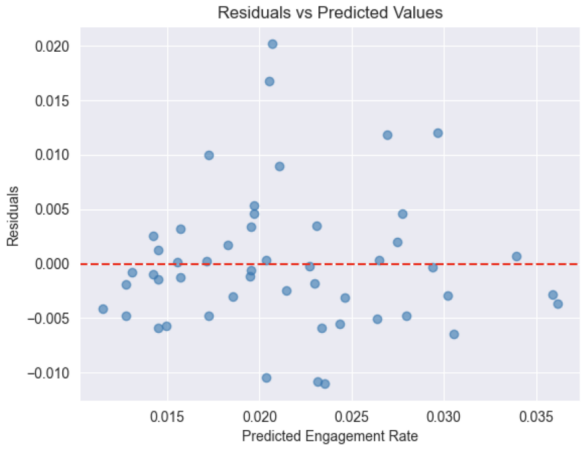

Figure 5: Image 5

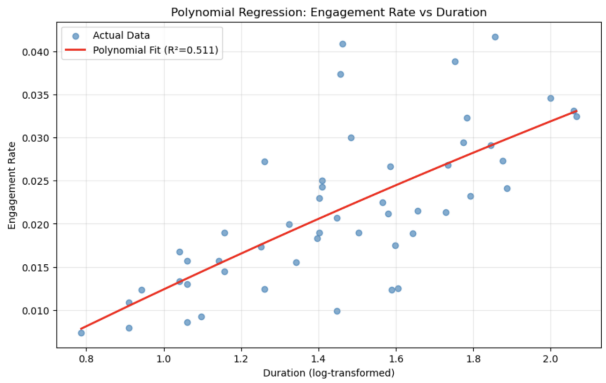

Figure 6: Image 6

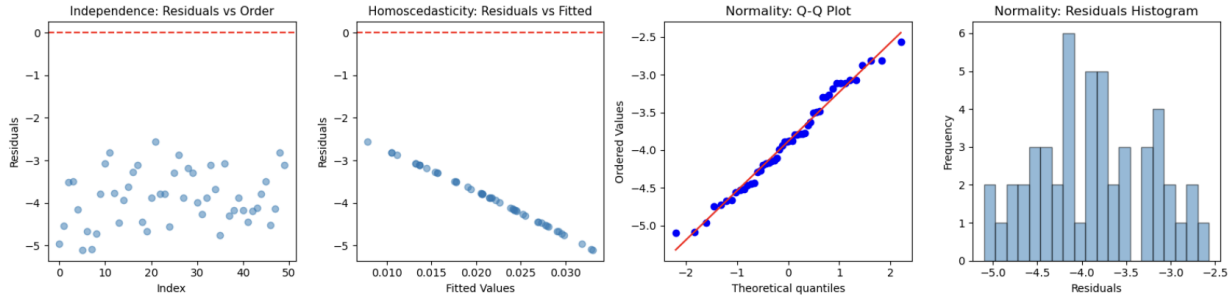

Figure 7: Image 7

Linear Regression Model. The linear regression model produced the following performance metrics: $R^2 = 0.0980$ (9.80% of variance explained), RMSE = 0.006283 (0.63%), and MAE = 0.004542 (0.45%). The fitted equation is

$$\text{engagement\_rate} = 0.0043 \times \text{duration} + 0.002.$$

This indicates that each additional second of video duration corresponds to an estimated increase of 0.0043 percentage points in engagement rate. However, the low R^2 suggests that duration alone explains little of the overall variability.

Visual and Diagnostic Analysis. The scatter plot with the regression line (Image 1) shows a weak positive linear trend, but considerable scatter suggests that the model does not capture key underlying factors. High-engagement outliers appear across many durations, indicating influences beyond simple linear duration effects. The diagnostic plots (Image 3) support that model assumptions are generally satisfied: residuals appear independent, variance is roughly constant, and normality is approximately met in both the Q–Q plot and histogram. The residuals vs. predicted values plot (Image 5) shows no curvature, suggesting no strong nonlinear pattern that the linear model is missing, even though its predictive power remains limited.

Overall, the linear model behaves statistically well but performs poorly in explanatory power. Although RMSE and MAE are small relative to typical engagement rates (1–4%), the low R^2 motivated testing a more flexible model.

Polynomial Regression Model. The polynomial regression model showed substantial improvement: $R^2 = 0.5110$ (51.10% variance explained), RMSE = 0.006165 (0.62%), and MAE = 0.004364 (0.44%). This represents a 5.2× increase in explained variance relative to the linear model. The improvement indicates that the relationship between duration and engagement is nonlinear.

Visual and Diagnostic Analysis. The polynomial regression plot (Images 6) reveals a clear concave-down, inverted-U pattern between log-transformed duration and engagement rate. Engagement rises quickly for very short videos (2–3 seconds) and continues increasing through moderate lengths. The highest predicted engagement occurs around 4–5 minutes (log duration ∼1.5–1.7), after which engagement plateaus or slightly declines, consistent with diminishing attention for longer content.

Diagnostic plots for the polynomial model (Image 7) again show independent residuals. A slight funnel shape in the residuals vs. fitted values indicates mild heteroscedasticity: predictions for low-engagement videos are more precise than for high-engagement ones. Despite this minor issue, the Q–Q plot and histogram show residuals that are approximately normal.

Model Comparison. The polynomial model substantially outperforms the linear model. Table summarizes the improvement:

Metric	Linear	Polynomial	Improvement
R^2	0.0980	0.5110	+421%
RMSE	0.006283	0.006165	-1.9%
MAE	0.004542	0.004364	-3.9%

The polynomial model offers far greater explanatory power while maintaining slightly better prediction error metrics. This confirms that engagement varies with duration in a nonlinear way better captured by a curvature-based model.

Additional Exploratory Findings. The views vs. engagement rate plot (Image 2) reveals that high-view videos cluster around moderate engagement (2–3%), while low-view videos show much broader variability, with engagement rates ranging from under 1% to above 4%. This suggests that viral or widely promoted videos produce stable but not exceptional engagement, whereas smaller niche videos vary more dramatically in how effectively they convert views into likes and comments.

Discussion

The results of applying simple and polynomial linear regression to predict engagement rate based on video duration, partially answered the feature analysis question by identifying duration as a key factor with a positive, concave-down relationship suggesting diminishing returns for excessively long content. The results, while suggesting a tentative action for marketers to favor longer videos, should be accepted with caution: the polynomial model's R^2 of $\approx 50\%$ is moderate, and the diagnostic plots show a clear violation of the homoscedasticity assumption (non-constant variance), potentially affecting the reliability of the model's coefficients. A major limitation is the small sample size of only 50 videos from a single niche channel,

which limits the generalization of the findings. This revealed the relationship's non-linearity and variance issues, demanding future work to implementing a multivariate model with more features. Ethical concerns are low given the use of aggregate, public data, but the focus on engagement rate as the primary success metric risks may unintentionally result in promoting algorithm-driven or clickbait content over quality.