EVALUATING THE IMPACT OF DIFFERENT ANALYTICS METRICS ON YOUTUBE VIDEO VIEWS USING STATISTICAL METHODS

A Thesis

by

NUMAN AHMAD

Submitted to the Honors College Department
of East Texas A&M University
in partial fulfillment of the requirements
for the degree of
BACHELORS OF SCIENCE
May 2025

EVALUATING THE IMPACT OF DIFFERENT ANALYTICS METRICS ON YOUTUBE VIDEO VIEWS USING STATISTICAL METHODS

A Thesis

by

NUMAN AHMAD

Approved by:

Advisor:	Nahid Hasan, Ph.D.
Committee:	Nahid Hasan, Ph.D. Kent Montgomery, Ph.D.
Head of Department:	Abdullah Arslan, Ph.D.
Dean of the Honors College:	Erin L. Webster Garrett, Ph.D.

ABSTRACT

EVALUATING THE IMPACT OF DIFFERENT ANALYTICS METRICS ON YOUTUBE VIDEO VIEWS USING STATISTICAL METHODS

Numan Ahmad, BS
East Texas A&M University, 2025

Advisor: Nahid Hasan, Ph.D.

Many videos are posted on YouTube daily, with some achieving millions of views and others not being pushed by the algorithm at all. In this study, we aim to examine a YouTube channel's analytics data to explore the metrics that have the greatest influence on viewership outcomes. Using statistical analysis such as correlation and regression analysis, we assessed the effects of watch time, average percentage viewed, average view duration, impressions, and click-through rates on viewership. We found that impressions, click-through rates, and average percentage viewed were not statistically significant for the first twenty-four hours after posting. However, over time these analytics became statistically significant for the models. The data provided a predictive model that can explain approximately 98% of the total variability in viewership. This research helps to explain the predictive relationship between YouTube's metrics and video success.

Keywords: Click-through rates, audience retention, average percentage viewed, regression model, social media strategy

ACKNOWLEDGEMENTS

I would like to begin by thanking my supervisor Dr. Nahid Hasan for his willingness to aid me in this academic endevour. Without his contributions, I would not have been able to complete this work. Both the Computer Science and Mathematics departments at East Texas A&M University was exceptionally accommodating to my interest in these areas and to fostering my curiosity.

I would also like to thank Dr. Kent Montgomery for serving in my thesis advisory committee, as well as Dr. Erin L. Webster Garrett, Dean of the Honors College, for giving me the opportunity of working on this thesis.

Finally, I would like to thank my family, particularly my mother Huma Siddiqui, my father Abrar Ahmad, and my siblings for their continued support of my academic aspirations. Without them, this would truly be impossible.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

Chapter 1

INTRODUCTION

With social media use projected to surpass six billion users by 2028, understanding the elements that drive video views across platforms like YouTube is crucial for content creators seeking to expand their audience and achieve success in an increasingly competitive environment [19].

This thesis investigates the complex relationship between YouTube's key analytics such as impressions, watch time, click-through rates, and viewership to evaluate their impact on YouTube video views. By exploring the key metrics involved in viewership outcomes, we aim to provide actionable insights for optimizing content to increase engagement, offering practical, real-world applicability for creators striving to maximize success on the platform.

To understand video performance on YouTube, it is important to first grasp the concept of **impressions (IMP)**. Impressions reflect the number of times a video's thumbnail is shown to users across the platform. Whether the thumbnail is seen in search results, recommendations, or browsing features, each instance where it appears on a screen counts as an impression [25]. An impression turns into a video view when a YouTube video begins playing.

The percentage of impressions that result in clicks on a video is called **click-through rate (CTR)**, which is calculated by dividing the number of clicks by the total number of impressions. This metric is essential for understanding how well a video's thumbnail and title convert exposure into actual views. A higher CTR indicates that the thumbnail and title are engaging enough to draw users into watching the video [23].

While impressions and CTR provide insight into how effectively a video attracts viewers, **Average Percentage Viewed (APV)** measures the percentage of a video that maintains viewer engagement when it is actually being watched. APV is a measure of Audience Retention, and it serves as an important indicator of how well content resonates with viewers, further influencing the platform's algorithm [6]. Similarly, **Average View Duration (AVD)** measures the total watch time of a video divided by the total number of plays, including replays [16].

Higher retention rates in both of these metrics not only suggest viewer satisfaction but also increase the likelihood of a video being promoted by YouTube's recommendation algorithm to both **new viewers** and **returning viewers** of a YouTube channel [7]. When viewers are engaged with the content for a longer time, the channel and video are considered to be more favorable, and they are both further promoted within the platform's ecosystem with more impressions [23]. This type of growth is often analyzed in **watch time**, which measures the total amount of time people spend watching a particular video [16, 23].

An important analytic that captures the holistic success of channel content as it is being positively promulgated through this type of positive engagement—is **Average Views per Viewer (AVV)**, which is defined as how many times a unique viewer has watched a channel's videos on average [27].

To analyze the impact of these variables on video performance, we use **linear regression models**. Regression analysis is a statistical method used to model and analyze relationships between a dependent variable and one or more independent variables [28].

By estimating the strength and direction of these relationships through **correlations**, linear regression models help determine how these key metrics affect the growth of YouTube channels at various points in time; in this study we perform analyses at 1, 7, 14, 21, 28, and 90 days after the day of publishing.

Similarly, a **log-linear model** is a statistical method used to examine relationships between data points by modeling the logarithm of the expected values of the dependent variable as a linear function of the independent variables [15]. This approach is particularly useful when analyzing data that often exhibit skewed distributions. By applying a log transformation, which can help address non-linear variables, this type of model will help

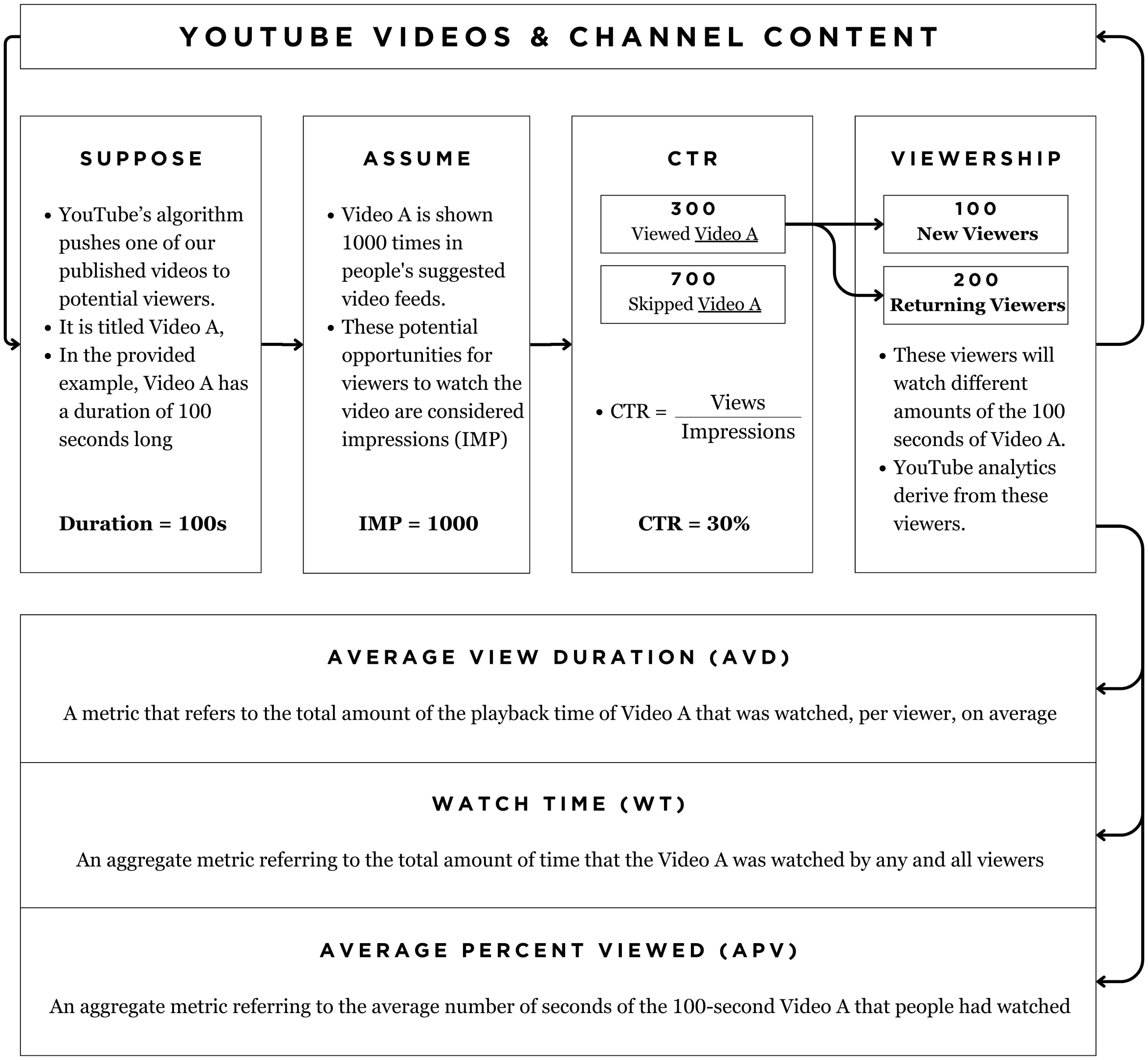
YOUTUBE VIDEOS & CHANNEL CONTENT
SUPPOSE
• YouTube's algorithm pushes one of our published videos to potential viewers.
• It is titled Video A,
• In the provided example, Video A has a duration of 100 seconds long
Duration = 100s
ASSUME
• Video A is shown 1000 times in people's suggested video feeds.
• These potential opportunities for viewers to watch the video are considered impressions (IMP)
IMP = 1000
CTR
300
Viewed Video A
700
Skipped Video A
• CTR = Views / Impressions
CTR = 30%
VIEWERSHIP
100
New Viewers
200
Returning Viewers
• These viewers will watch different amounts of the 100 seconds of Video A.
• YouTube analytics derive from these viewers.
AVERAGE VIEW DURATION (AVD)
A metric that refers to the total amount of the playback time of Video A that was watched, per viewer, on average
WATCH TIME (WT)
An aggregate metric referring to the total amount of time that the Video A was watched by any and all viewers
AVERAGE PERCENT VIEWED (APV)
An aggregate metric referring to the average number of seconds of the 100-second Video A that people had watched

Figure 1.1. Example Analytics Derivative Pathway from YouTube Video Uploads

identify how analytics like impressions, watch time, CTR, AWT, and AVD influence video views. This modeling approach allows for a clearer understanding of the relationships between variables such as exponential growth or decay patterns, and it can offer insights into how small changes in key metrics can have compounding effects on long-term video success.

Finally, to enhance the analysis of video performance, this research integrates machine learning with regression models. Machine learning allows for identifying complex, non-linear relationships between key metrics like CTR, Audience Retention, APV, and overall viewership on improving predictive accuracy. Drawing our data and previous studies, these models learn from historical data to forecast future performance and adapt to changing trends [9]. This approach automates feature extraction, optimizes content strategies, and scales with growing datasets, providing creators with deeper insights for refining their strategies in response to evolving viewer behavior and platform dynamics.

In summary, this thesis aims to offer a deeper understanding of the complex relationship between YouTube analytics and content creation techniques. By leveraging data-driven insights, creators can refine their approach to video production, improving performance and fostering sustained growth on the platform.

Chapter 2
LITERATURE REVIEW

In a recent study, Wilson [23] analyzed the relative importance of click-through rates (CTR) and watch time for YouTube video views, using a dataset of newly released videos and log-linear models. The primary aim was to test whether CTR or average watch time per view had a stronger association with views per subscriber, controlling for channel authority. The research proposed that CTR would have a significantly larger effect on views than watch time per impression, a hypothesis that reflects growing interest in understanding how algorithmic recommendations prioritize different metrics (3.5). Although Wilson provides valuable insights into YouTube's algorithmic priorities, this proposed log-linear model only accounts for views per impression and CTR values. This research disregards the other analytics metrics that perhaps play a more significant role in YouTube video views.

Previously, Covington *et al.* [5] developed a recommendation system for YouTube, incorporating deep learning models. The research focused on two stages: candidate generation and ranking, which helped optimize YouTube's recommendation system based on user history, search tokens, and video demographics. The candidate generation model retrieved relevant videos from a large corpus, while the ranking model scored them based on watch time and CTR. This two-stage deep learning approach addressed challenges like scale, freshness of content, and noise in historical user behavior. The methodology in this study involved using TensorFlow for training models with billions of parameters across vast datasets, providing insight into YouTube's underlying recommendation algorithms. This thesis builds on Covington's model by digging deeper into how CTR and audience retention interact, aiming to uncover new ways to optimize YouTube's recommendation system over time.

Park *et. al.* [16] conducted a data-driven study of YouTube's view duration focused on how collective preferences and audience reactions impact video watch time . They hypothesized that view duration correlates with other video engagement metrics such as view counts, likes, and comment sentiment. Using two datasets—one with 1,125 randomly sampled YouTube videos and another tracking individual viewer behavior via a Chrome extension—the researchers applied a mixed-effects model to analyze these relationships. A key methodological tool was the use of Linguistic Inquiry and Word Count (LIWC) software to measure the sentiment of video comments [21]. The study's methodology provided a robust framework for understanding how user engagement and sentiment correlate with view duration across different video categories. The connections Park identified between sentiment, engagement, and watch time inspire this thesis to explore how metrics like APV and audience retention can help predict long-term video performance.

In another study that utilized a dataset of over 37 million videos, research offered an in-depth examination of video popularity metrics on YouTube [4]. The methodology focused on the correlations between popularity metrics such as view count, comments, ratings, and favorites. By using linear regression models, they estimated view counts based on these other metrics, controlling for various content-related factors. Disregarding the video's contents, this research contributed a systematic approach to understanding video popularity by focusing on how different metrics interact. This research approach to understanding how metrics drive popularity provides a strong foundation, which this thesis will expand by integrating machine learning models for more accurate predictions.

Borghol *et al.* [3] further extended this content-agnostic approach by analyzing the popularity dynamics of "clone" videos—videos with nearly identical content. The methodology employed multivariate regression to determine how factors like uploader characteristics, total previous views, and video age influenced popularity. By controlling for content variability, this study demonstrated the importance of non-content factors, such as the rich-get-richer phenomenon, in driving video popularity. This research offered a novel perspective on how video popularity evolves based on uploader networks and video age, especially for new content, providing insights into YouTube's recommendation and search biases. By showing how non-content factors influence video popularity, the research highlights areas this thesis will explore further—bridging both content and engagement metrics like CTR and APV.

Researchers have also recently investigated how YouTubers adjust their speech across different video formats, focusing on Phil Lester's speech patterns [12]. Through mixed-model regression, the study compared Lester's speech in solo vlogs, collaborative vlogs, gaming videos, and live streams. The hypothesis was that his speech would become less formal in unscripted formats, like live streams or gaming videos. Results supported this, showing that he spoke more carefully in scripted solo vlogs while adopting a more relaxed style in spontaneous contexts. This suggests that creators adjust their presentation to suit different video formats and audience expectations. However, since the study focuses on just one vlogger, its findings may not fully capture how other YouTubers shift their speech [12]. While Lee focuses on speech patterns, this thesis extends that idea by examining how video formats and presentation styles impact important performance metrics like watch time and retention.

Morgan *et al.* [15] set out to understand how attention patterns unfold in YouTube videos by collecting data from YouTube's API. They tracked 1,460 trending videos and 4,250 recently uploaded ones every six hours over two months, following up for a year to confirm trends. By analyzing views, comments, and likes as the main metrics to see the growth of ”trending” and ”recently uploaded” videos, they detected differences in audience engagement levels that could alter the trajectory of a video's performance in the future. Their hypothesis was that early fluctuations in engagement could predict whether a video would become more popular over time. They found that most videos stabilized quickly unless unexpected events sparked renewed interest. However, the study focuses only on trending and recent videos, which limits how well the findings apply to other types of content on the platform [15]. The findings on engagement patterns offer key insights that this thesis will build on, broadening the analysis to cover a wider variety of content types and their impact on performance.

A peer-reviewed book chapter had also recently explored how predictive analytics using time-series datasets from Twitter and Facebook can forecast outcomes like sales and public sentiment [10]. Methods include linear regression, autoregressive models, LASSO regression, and machine learning techniques such as decision trees. LASSO regression helps select key variables and prevent overfitting, relevant to this thesis's focus on identifying which YouTube metrics (e.g., CTR, APV, captions) influence performance. However, predictive models face challenges with sudden behavioral shifts—a concern for YouTube, where algorithm changes affect engagement. Pre-processing data is also complex, which this thesis can potentially address by automating feature extraction through a machine learning algorithm. The need for continuous model monitoring further informs this thesis's approach to tracking dynamic video metrics over time [10]. The challenges that this literature identified with predictive models—like handling shifts in behavior—inform this thesis's efforts to design models that track dynamic metrics such as CTR, APV, and retention.

Zhou *et al.* [28] explored whether YouTube advertising influences sales by analyzing data from a company's ad budget and sales performance. They used simple linear regression models to examine the relationship between ad spending and sales, with t-tests and F-tests to verify the results. Their hypothesis was straightforward: increased ad spending would lead to higher sales. The analysis confirmed that spending more on YouTube ads does improve sales, underscoring the value of the platform for marketers. That said, the authors acknowledge that their use of a linear model may overlook more complex relationships, and they suggest future research explore non-linear models for deeper insights [28]. The reasearcher's findings on ad spending underscore this thesis's goal of exploring how broader YouTube metrics, like CTR and audience retention, can provide deeper insights into video growth and optimization.

Bhatnagar and Urolagin [2] employed a deep learning framework to investigate the factors influencing the popularity of YouTube videos, with an emphasis on prediction and analysis. Their research centered on feature extraction from videos and the use of neural networks to forecast video performance. To develop their model, they created a dataset containing various genres and video attributes such as views, likes, and scene quality. By applying supervised learning techniques, they trained a sequential classifier to categorize videos based on these extracted features. The study's deep learning model demonstrated how predictive analytics can identify which videos are likely to achieve popularity, leveraging performance metrics like video duration and scene quality to optimize predictions. The research underscores the role of deep learning in enhancing video analytics. This literature proposes that these models can be extended beyond content creation to applications in media strategy. This thesis builds upon this framework by incorporating both engagement metrics and analytics, helping forecast YouTube video popularity across a range of content types.

Lau *et al.* [11] conducted a predictive analysis of YouTube views by utilizing a dataset comprising publicly available YouTube statistics. The study introduced two regression models—Ordinary Least Squares (OLS) and Online Gradient Descent—to estimate view counts based on various available metrics. These models utilized variables such as the number of comments, ratings, and favorites, which were found to correlate highly with views, while excluding content-based factors. Lau's hypothesis was that OLS would outperform Online Gradient Descent due to its gradient adjustment capabilities. Results indicated that OLS was more accurate in predicting view counts, yet its performance was limited by the simplicity of linear regression, which overlooks non-linear relationships between variables. This thesis builds on Lau's findings by incorporating machine learning models that handle non-linear interactions for enhanced accuracy.

Pinto *et al.* [17] proposed early view patterns as predictors of long-term popularity, using a dataset of YouTube videos with varying initial view trajectories. They developed two predictive models—a multivariate linear regression model and a model based on daily popularity growth rates—to test if initial popularity surges could forecast future view counts. The hypothesis was that early engagement metrics would be reliable indicators of long-term success, particularly for content experiencing a popularity peak. Findings supported this, showing up to a 20% improvement over baseline models. However, these models struggled with content experiencing sustained engagement over time, which this thesis seeks to overcome by incorporating continuous metrics such as audience retention and average percentage viewed.

Mekouar *et al.* [14] applied logistic and stepwise regression to predict YouTube video popularity, using a dataset that included video views and interaction statistics. Their variables focused on engagement-related features, with a particular interest in how these metrics could predict whether a video would reach a specific popularity threshold. Mekouar hypothesized that stepwise regression would improve prediction accuracy by selectively including impactful variables. Results showed promising accuracy; however, the model was limited by its binary prediction framework, which this thesis expands upon by predicting continuous popularity metrics like average view percentage and total watch time.

Rowe [18] used a multiple linear regression model to analyze audience increase on YouTube, leveraging data from the YouTube API to track changes in subscriber counts and engagement. Key variables included the number of channels a user subscribed to, views per video, and video frequency, with the hypothesis that user engagement behaviors directly influence subscriber growth. The study found that frequent content sharing and engagement significantly boosted subscriber numbers. Nonetheless, Rowe's approach did not consider non-behavioral factors such as content format, an aspect this thesis addresses by incorporating metrics related to video presentation styles alongside engagement behaviors.

Szabo and Huberman [20] explored the long-term popularity of YouTube and Digg content, collecting daily view counts from the YouTube API to model popularity trajectories. They applied a linear regression model that used early engagement metrics, hypothesizing that early activity levels could predict long-term view growth. Findings confirmed a strong correlation between early and eventual popularity, although predictions were more accurate for rapidly consumed content than for videos with prolonged interest. This thesis extends Szabo and Huberman's work by integrating variables such as audience retention and click-through rate, aiming to improve predictions for videos with varied consumption patterns.

A 2022 study on YouTube data analysis [8] combined linear regression and neural networks to predict video views, using a dataset of videos with detailed engagement and content characteristics. Their methodology compared the effectiveness of regression and neural network models in predicting popularity based on factors like view count, likes, and video duration. The study hypothesized that neural networks would outperform linear models in handling complex, non-linear data. While results indicated that neural networks provided better accuracy, the model required extensive computational resources. This thesis builds on these findings by implementing a hybrid approach, combining linear models for initial predictions and machine learning for refining predictions based on complex interactions among variables.

2.1. Hypothesis

In this paper, we want to test the assertion that: High average percentage viewed (APV), when combined with high CTR, leads to increased views on a channel over time.

Chapter 3
METHODOLOGY

In this chapter, we explored the methods used to analyze our YouTube dataset. First, we described the dataset we collected from the channel's YouTube Studio profile. Then we discussed several statistical methods, such as correlation, multiple linear regression, and log-linear models to perform the necessary analysis. We assessed the YouTube channel's viewership by utilizing these extracted analytics data with these methods.

3.1. Software and Tools

For data compilation and statistical analysis, we used the following software. This allows us to share our files with others to review our work.

- R Version 4.4.2 with R Studio Version 2024.09.1+394
- GitHub: https://github.com/thenumanahmad/YouTube-Analytics [1]
- YouTube Channel: https://www.youtube.com/@theibnahmad [22]

3.2. Data Description

In this study, we used a novel dataset, which is manually extracted from the channel's YouTube Studio profile, providing a detailed and longitudinal view of video engagement metrics at various time points [1]. This dataset not only presents insights into general viewer behavior but includes an analysis of study videos regarding the Medical College Admissions Test, allowing for a targeted investigation into how "study with me"-styled content performs.

This data contains various metrics collected from a YouTube channel, capturing the performance and viewer engagement across multiple videos at specific time points after the day of posting: 1, 7, 14, 21, 28, and 90 days. The dataset captures this information for 152 videos at these time points. Exported as CSV files, this data captures the following metrics:

- **Video Details**: Information about the video itself, such as:
  - *Content*: The title or identifier for each video.
  - *Date Published*: The date the video was uploaded to YouTube.
- **Engagement Metrics**: Measures of viewer interaction with the video, including:
  - *Comments Added*: The number of comments received, indicating viewer interest and engagement in discussion or feedback.
  - *Shares*: The number of times the video was shared, which reflects its appeal and potential to spread among viewers.
  - *Likes (vs. Dislikes)*: The ratio of likes to dislikes, providing an overall sentiment of viewer approval.
  - *Dislikes* and *Likes*: Separate counts for dislikes and likes, giving a more detailed view of audience reaction.
- **Viewer Interactions**: Metrics indicating viewer loyalty and reach, including related to changes in the channel's subscriber base, such as:
  - *Subscribers Lost* and *Subscribers Gained*: Counts of subscribers who either unsubscribed or joined after each video was published. This helps to understand how content may affect subscriber retention and growth.
  - *Subscribers*: The total number of subscribers at the time of data collection, serving as a reference for the channel's audience size.
  - *Returning Viewers*: The number of viewers who had previously watched content on the channel, helping to measure repeat viewership.
  - *Unique Viewers*: The count of distinct individuals who watched each video, giving a measure of unique reach.

- **Content Engagement**: Indicators of how well viewers engaged with the video content itself, such as:
    - *Viewed (vs. Swiped Away)*: The proportion of viewers who chose to watch the video rather than skip it, offering insight into the video's appeal in the feed.
    - *Shown in Feed*: The number of times the video appeared in users' feeds, reflecting its visibility.
    - *Average Percentage Viewed*: The average percentage of the video that viewers watched, indicating depth of engagement.
    - *Average View Duration*: The average time viewers spent watching the video, which helps assess the content's ability to retain attention.
- **Performance Indicators**: Key measures of video reach and overall success, including:
    - *Views*: Total view count for each video, representing its reach.
    - *Watch Time (Hours)*: The cumulative hours viewers spent watching the video, a key metric that often correlates with YouTube's recommendation algorithms.
    - *Impressions*: The number of times the video's thumbnail was shown to users, reflecting its exposure.
    - *Impressions Click-Through Rate*: The percentage of impressions that led to a view, providing insight into the effectiveness of the thumbnail and title in attracting viewers.

These metrics together provide a comprehensive picture of each video's engagement, reach, and performance on YouTube. By analyzing these variables, this study aims to identify which factors most strongly influence a video's success. The dataset serves as a foundation for understanding patterns in viewer behavior and performance trends, and will support the use of correlation and regression models to highlight significant predictors of video success.

3.3. Correlation

In our analysis, we employ correlation to examine the relationships between each independent variable and the dependent variable, the number of views. Correlation quantifies the degree and direction of association between two variables, represented by the correlation coefficient. For two variables X_i and Y, the correlation coefficient $r_{X_i,Y}$ is defined as:

$$r_{X_i,Y} = \frac{\sum_{j=1}^{n}(X_i^{(j)} - \bar{X}_i)(Y^{(j)} - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(X_i^{(j)} - \bar{X}_i)^2 \sum_{j=1}^{n}(Y^{(j)} - \bar{Y})^2}} \tag{3.1}$$

where $X_i^{(j)}$ and $Y^{(j)}$ represent individual observations of the variables X_i and Y, with $\bar{X}_i$ and $\bar{Y}$ as their respective means. The correlation coefficient $r_{X_i,Y}$ ranges from -1 to 1, where values close to 1 suggest a strong positive association, values close to -1 indicate a strong negative association, and values around 0 imply little to no linear relationship.

In this study, correlation analysis serves as a preliminary step, helping to identify the initial associations between each predictor and the dependent variable. By examining these relationships, we can discern which independent variables may warrant further investigation in the regression model. This approach ensures that variables with notable correlations are considered in our subsequent linear and log-linear models, thereby enhancing the ability to interpret and robustness of our results.

3.4. Regression Model

Consider a set of independent variables X_1, $X_2, \ldots, X_n$ of size n and an dependent variable Y. Then we may fit the following linear model:

$$Y = Y(\mathbf{X}) = \sum_{i=1}^{k} \beta_i X_i + \epsilon, \tag{3.2}$$

where $X_0 = 1$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T \in \mathbb{R}^{k+1}$ is a vector of coefficients (parameters), and $\epsilon \in \mathbb{R}$ is the error term with 0 mean and constant variance.

Because there is an intercept term in Eq. (3.2), we considered that all independent variables (predictors) are centered i.e., $E(X_i) = 0$ for $i = 1, \ldots, k$, without any loss of generality.

In our study, the dependent variable is the measured number of views, and the independent variables are the numbers of returning viewers, new viewers, total watch time in hours, click-through rates, number of impressions, average view duration, average percentage viewed, and average views per viewer.

We want to identify the most influential (or important) predictors X_i in the linear model (3.2). First, we compute the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[Y_i - \hat{Y}(\mathbf{X}_i)]^2}{\sum_{i=1}^{n}[[Y_i - E[Y]]^2}, \tag{3.3}$$

where $\hat{Y}(.)$ is the predicted value of the model that can be computed using the given data. R^2, in general, represents the percentage of responses explained by the fitted model.

3.5. Log-Linear Models

To further examine the relationships between variables, we use log-linear models—a type of statistical model that captures multiplicative effects by assuming a logarithmic relationship between the expected value of the response variable and a linear combination of predictors. This approach is especially effective for analyzing count data or when the response variable grows exponentially with changes in predictor variables.

In a log-linear model, we transform the expected value of the response variable Y by taking its natural logarithm, expressed as:

$$\log(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{3.4}$$

where $E(Y)$ represents the expected value of Y, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients that correspond to each predictor variable $X_1, X_2, \ldots, X_n$. This transformation allows each coefficient β_i to be interpreted as the multiplicative effect on Y for a one-unit increase in X_i, while holding other variables constant.

In our literature review, we discussed Wilson's [23] research and the proposal that CTR and IMP play important roles in viewership outcomes. This is represented by the following equation, which indicates the mathematical relationship between these variables:

$$E\left[\frac{WT}{IMP}\right] = CTR \times E[WT], \tag{3.5}$$

where $E\left[\frac{WT}{IMP}\right]$ is the expected minutes of watch time per impression, CTR represents the click-through rate, measuring the number of views per impression, and $E\left[WT\right]$ represents the expected watch time minutes watched per view.

After taking log on both sides of Eq. (3.5), we obtain the log-linear version of the above model, as follows:

$$\begin{aligned}
& E\left[\frac{WT}{IMP}\right] = CTR \times E[WT] \\
\implies & \log(E(Y)) = \log\left(E\left[\frac{WT}{IMP}\right]\right) + \log(IMP) - \log(E[WT]) \\
\implies & \log(E(Y)) = \beta_0 + \beta_1 \cdot \log(E[WT]) + \beta_2 \cdot \log(IMP) + \beta_3 \cdot \log(E[WT])
\end{aligned}$$

In our study, log-linear models are used to analyze metrics where changes are better understood in terms of proportional or multiplicative effects—such as impressions or view counts—rather than simple additive increases. By applying log-linear models, we can uncover and plot the exponential relationships between variables, offering a nuanced perspective that complements our linear regression analysis (4.1).

Chapter 4
RESULTS

In this chapter, we analyzed YouTube's analytics dataset (described in Section 3.2 in Chapter 3). The results are shown and discussed throughout this chapter.

4.1. Exploratory Data Analysis (EDA)

The dataset was analyzed using a log transformation to stabilize variance and address skewness, providing a clearer understanding of the data's distribution and interquartile ranges across time points. Box-and-whisker plots were used to visualize the transformed data, highlighting outliers and the central tendencies for key metrics such as Watch Time (WT), Click-Through Rate (CTR), and Impressions (IMP).

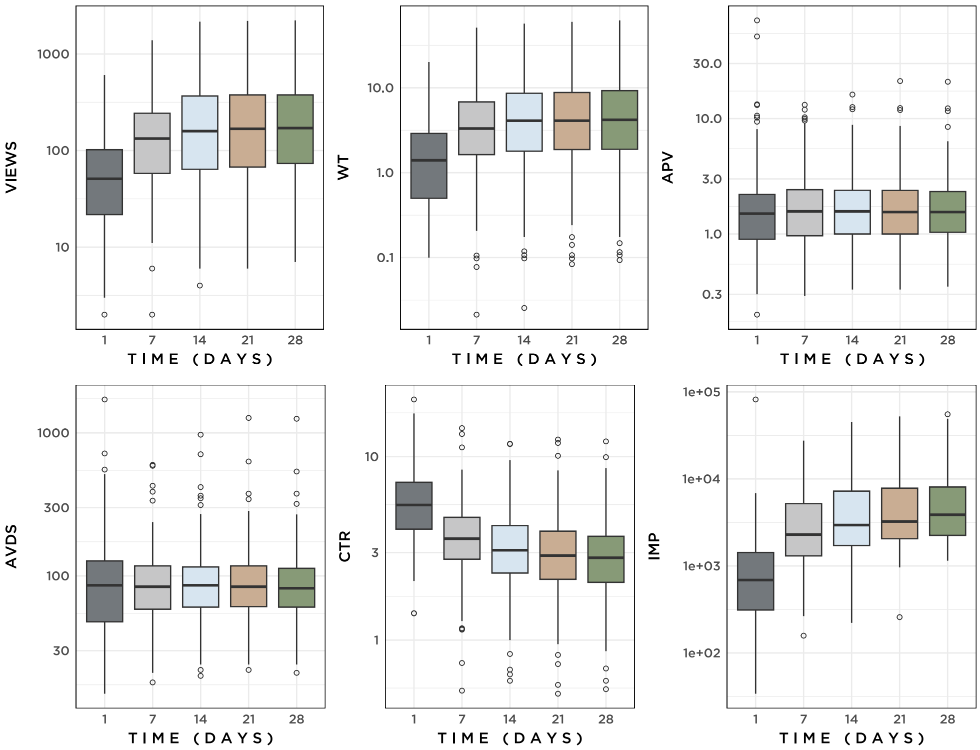

Figure 4.1. Box-and-Whisker Plots of Log-Linear Transformations of YouTube Analytics Data at Specified Time Points After the Day of Posting.

Figure 4.1 showed that engagement metrics like Average Percentage Viewed (APV) and Average Duration Viewed (AVDS) exhibited minor increases in variability over time.

This suggests that viewer behavior becomes more diverse as videos remain online. On the other hand, Views, WT, and IMP increase with a higher rate in the first day after posting. Conversely, CTR decreases over time. This foundational analysis provided a robust framework for deeper exploration of relationships and predictors.

4.2. Correlation

Correlation analysis at various time points offered insights into the evolving relationships between engagement metrics. **Figure** 4.2 demonstrates the pairwise correlations between variables for Day 1 and Day 7. From the figure for day 1, we observed that the Pearson correlation coefficient between Views and WT is 0.92. It indicates that on the first day after posting, there is a very strong correlation between the number of views and the amount of watch time per video. Similarly, a moderate correlation was observed between APV and AVDS with $r = 0.79$. The correlation between CTR and APV at Day 1 was notably low at $r = 0.05$ but fluctuated between $r = 0.023$ and $r = 0.27$ in all later datasets. However, CTR and Impressions showed weaker associations during this period ($r = 0.05$), suggesting that their influence might manifest later in the content life cycle.

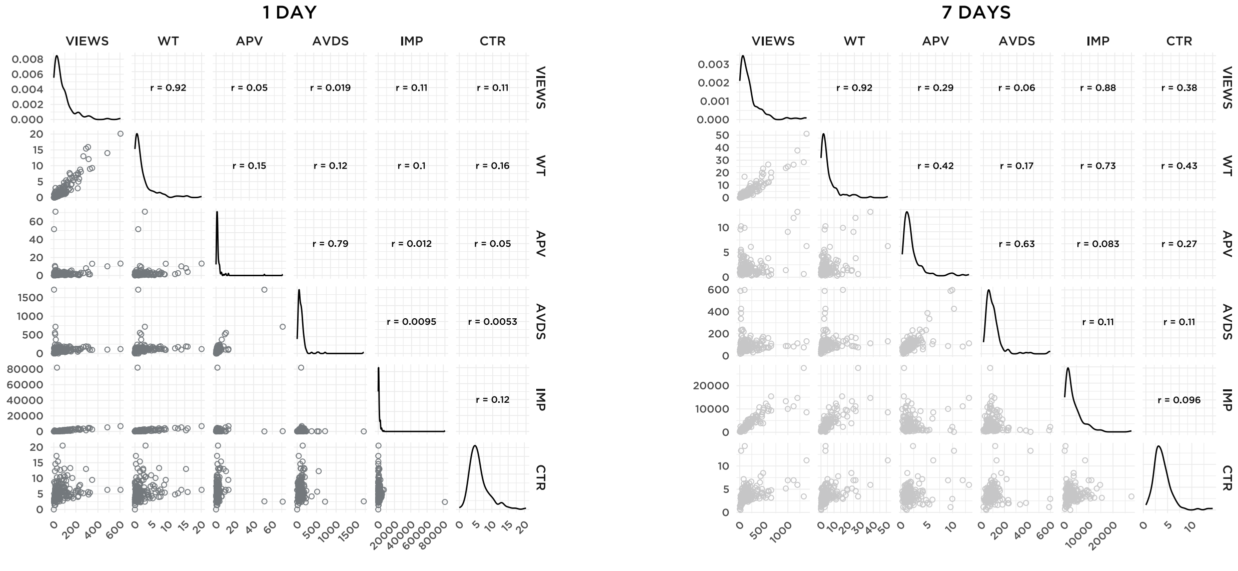

Figure 4.2. Correlation Plots of YouTube Analytics Data at 1 and 7 Days After the Day of Posting

Correlations from Day 1 are unique because it shows most of the lowest correlations across the variables. Also, many of the correlations at this time point show the greatest deviation from the stabilization of these values at later points in Days 14, 21, and 28 (see **Figure** 4.2). This is exemplified in many of the correlations, such as: APV with WT; AVDS with APV; IMP with Views, APV, and AVDS; and, lastly, CTR with Views, WT, APV, and AVDS. At later time points, it is frequently found that these correlations increase and stabilize at much higher values.

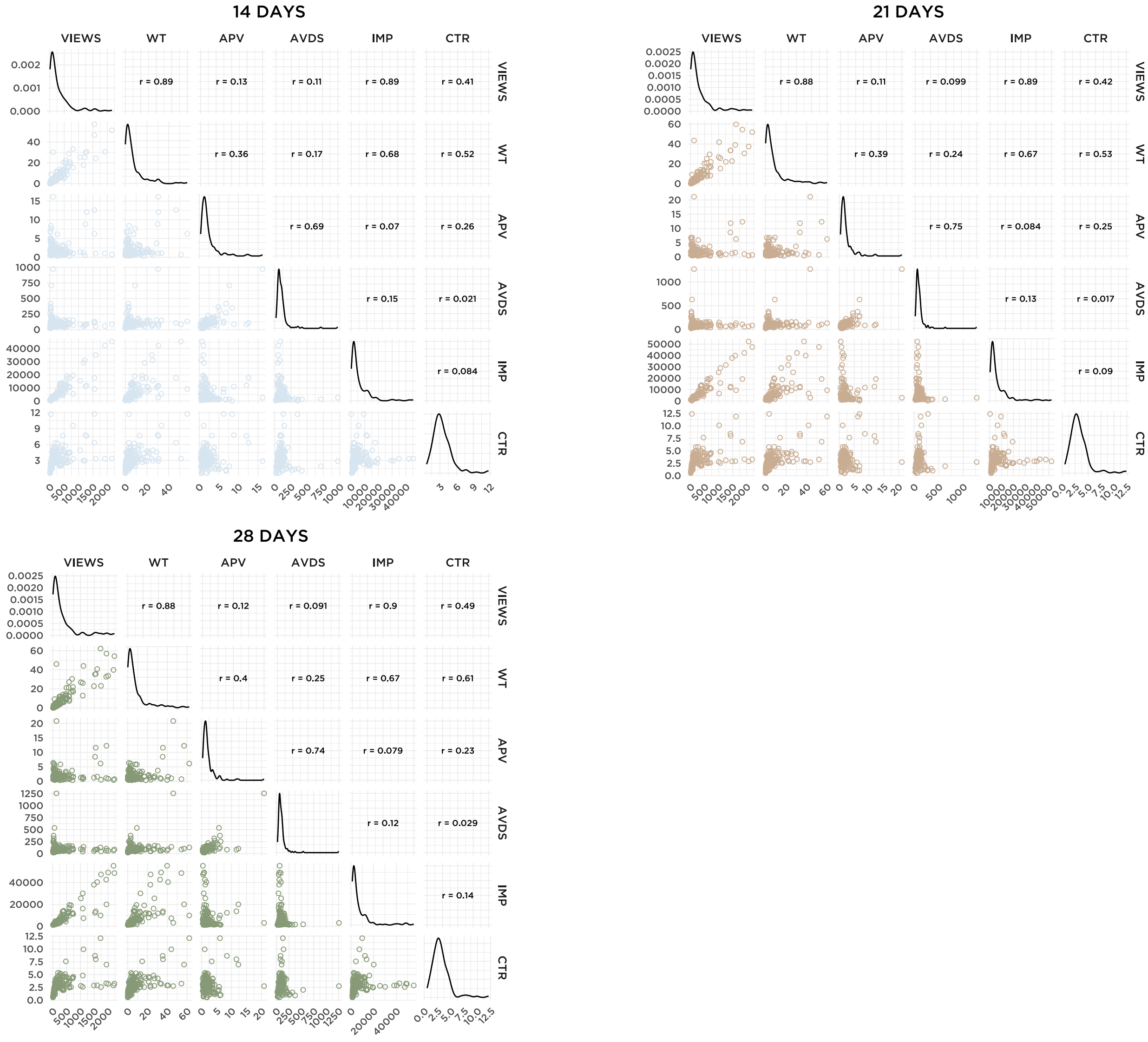

Figure 4.3. Correlation Plots of YouTube Analytics Data at 14, 21, and 28 Days After the Day of Posting

Conversely, AVDS with APV had the highest correlation at one day after posting. The data in later time points maintained a Pearson correlation coefficient between $[0.63 - 0.79]$. This indicates the importance of AVDS and it's impact on APV throughout a video's life on YouTube. Moreover, the correlation betwen WT and Views was highest at this time point as well.

By Day 7, stronger correlations emerged between CTR, WT, and APV, highlighting CTR's growing role in driving audience engagement. For example, CTR's correlation with WT and APV increased from $r = 0.16$ and $r = 0.05$ on Day 1 to $r = 0.43$ and $r = 0.27$ on Day 7, respectively. This progression underscores the dynamic nature of engagement variables and their differing impacts over time. Day 7 also showed the greatest correlations between certain metrics. For example, APV with Views and WT, IMP with WT, and CTR with APV were highest at this time point. Additionally, the correlation between IMP and AVDS was notably high at Day 7. The relationship between CTR and the other variables remained consistent through Days 14, 21, and 28, with CTR showing an increasingly strong correlation with overall performance metrics.

Day 14 in particular exhibits the highest correlation between AVDS and Views, which peaked at $r = 0.11$ and hovered around $r = 0.09$ thereafter. Similarly, the correlation between IMP and AVDS peaked around $r = 0.15$ and stagnated slightly below thereafter. Interestingly, however, CTR had the lowest correlation with IMP at this time point. At Day 21, IMP and APV experienced the highest correlation with $r = 0.084$.

Day 28 is more notable, as it exhibits the greatest correlations for many of the metrics. Specifically, at 28 days, the correlation between AVDS with WT, IMP with Views, and CTR with Views, WT, AVDS, and IMP reaches its peak. However, the correlation for CTR with IMP in particular is negligible because the correlation generally remained very weak ($r < 0.15$) for all datasets recorded.

4.3. Multiple Linear Regression of YouTube Data

Regression analysis was conducted to evaluate the predictive influence of key metrics (APV, AVDS, WT, IMP, and CTR) on video performance at different time points. Table 4.1 shows the detailed results of this model for 1, 7, 14, 21, and 28 days after posting.

On Day 1, WT (with estimate = 25.8035, $p-$value < 0.0001) and AVDS (with estimate = 0.0994, $p-$value= 0.0005) emerged as significant predictors in this model. This reflects their importance in the immediate post-upload period. However, APV (with p-value = 0.3514), IMP (with p-value =0.6960), and CTR (with p-value = 0.3486) were not statistically significant. This indicates limited influence at this early stage. The model explained 87.32% of the variance in viewership ($R^2 = 0.8732$). This high metric demonstrates strong predictive strength.

By Day 7, all variables, including CTR, became highly significant (with $p-$value < 0.0001), with CTR's influence notably increasing (with estimate = 11.5273). The model's predictive accuracy improved substantially as R^2 increased from 0.8732 to 0.9711. This indicates that the metrics provided a robust explanation of video performance at this stage. This trend continued through 14 and 21 days after posting, where all variables maintained high levels of significance. CTR's estimate rose further to 20.4164 on Day 14 from -0.8388 on Day 1. This demonstrates CTR's growing influence in video engagement over time. The models for these time points explained nearly all variance, with R^2 values of 0.9832 (Day 14) and 0.9793 (Day 21).

By Day 28, CTR and APV emerged as the most influential predictors of long-term engagement, with estimates of 38.9986 and 27.9557, respectively. Both metrics are considered highly significant (p $<$ 0.0001). These findings highlight CTR's critical role in driving sustained engagement, especially in the later stages of a video's lifecycle. The model achieved a very high goodness-of-fit ($R^2 = 0.9832$), indicating that the model fits very well with the data.

Table 4.1. Regression Model Results at Different Time Points

Time after posting	Predictor	Estimate	Std. Error	p-value	Multiple R^2
Day 1	Intercept	32.7663	6.4209	0.0000***	0.8732
	APV	0.5783	0.6186	0.3514	
	AVDS	0.0994	-0.0278	0.0005***	
	WT	25.8035	0.8375	0.0000***	
	IMP	0.0002	0.0004	0.6960	
	CTR	-0.8388	0.8921	0.3486	
Day 7	Intercept	-19.1312	10.9149	0.0818	0.9711
	APV	13.7187	2.3769	0.0000***	
	AVDS	-0.5634	0.0561	0.0000***	
	WT	17.9711	0.9134	0.0000***	
	IMP	0.0298	0.0016	0.0000***	
	CTR	11.5273	2.1433	0.0000***	
Day 14	Intercept	-46.7240	14.6842	0.0018**	0.9832
	APV	17.1186	3.1928	0.0000***	
	AVDS	-0.6607	0.0638	0.0000***	
	WT	18.0029	1.0104	0.0000***	
	IMP	0.0283	0.0011	0.0000***	
	CTR	20.4164	3.6214	0.0000***	
Day 21	Intercept	-35.4863	15.5135	0.0236*	0.9793
	APV	22.6868	3.6341	0.0000***	
	AVDS	-0.8765	0.0691	0.0000***	
	WT	18.4979	1.0266	0.0000***	
	IMP	0.0264	0.0011	0.0000***	
	CTR	21.5877	4.0138	0.0000***	
Day 28	Intercept	-87.4238	18.5141	0.0000***	0.9832
	APV	27.9557	3.6838	0.0000***	
	AVDS	-0.9063	0.0734	0.0000***	
	WT	15.6930	1.1449	0.0000***	
	IMP	0.0284	0.0010	0.0000***	
	CTR	38.9986	5.2616	0.0000***	

Chapter 5

DISCUSSION

The results reveal that metrics influencing video performance evolve over time. In the early stages, Watch Time and Average Duration Viewed are the most significant predictors, but metrics like Click-Through Rate and Average Percentage Viewed gain increasing importance in the later stages.

Regression models demonstrate excellent predictive accuracy, particularly after Day 7, with R^2 values consistently exceeding 0.97. By Day 28, CTR emerges as the most dominant factor. This underscores its essential role in determining the long-term success of a YouTube video. These findings emphasize the importance of focusing on specific engagement metrics at different stages to optimize video performance.

Our dataset in sum consists of 152 data points. Although we were able to show strong correlations between variables, our analysis of YouTube Video Performance may be strengthened by exploring larger datasets. We recommend future work to focus on YouTube datasets with a larger number of data points.

Moreover, although our dataset's variables have strong correlations and relationships, perhaps manual exports of data from the YouTube channel resulted in skewed data.

For a future study, we recommend implementation of a data protocol, such as YouTube Data API v3.0, which will automatically scrape a YouTube' channel's data to enhance the results of our study and create an automatic extraction protocol to ensure accurate and reliable data that is not subject to human-induced error.

REFERENCES

[1] Ahmad, N. (2024). *YouTube Analytics* (Version 1.0) [Data set]. GitHub. https://github.com/thenumanahmad/YouTube-Analytics

[2] Bhatnagar, D., & Urolagin, S. (2021). YouTube Video Popularity Analysis and Prediction Using Deep Learning. International Journal of Computational Intelligence in Control, 13(2), 127–134.

[3] Borghol, Y., Ardon, S., Carlsson, N., Eager, D., & Mahanti, A. (2012). The untold story of the clones. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 286, 1186–1194. https://doi.org/10.1145/2339530.2339717

[4] Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010). A first step towards understanding popularity in YouTube. In Proceedings of INFOCOM.

[5] Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems (pp. 191–198).

[6] Dascalu, C. G., Antohe, M. E., Boiculese, V. L., & Moscalu, M. (2019). New Technologies in Medical Education: The Potential of Video Resources – YouTube Channeling. In *eLearning and Software for Education* (Vol. 1, pp. 255–263). eLSE 2019. Carol I National Defence University Publishing House. https://doi.org/10.12753/2066-026x-19-035

[7] Hoiles, W., Aprem, A., & Krishnamurthy, V. (2017). Engagement and popularity dynamics of YouTube videos and sensitivity to meta-data. IEEE Transactions on Knowledge and Data Engineering, 29(7), 1426–1437. https://doi.org/10.1109/tkde.2017.2682858

[8] Fan, J., & Lian, T. (2022). YouTube data analysis using linear regression and neural network. In *2022 International Conference on Big Data, Information and Computer Network (BDICN)*. https://doi.org/10.1109/BDICN55575.2022.00055.

[9] Khan, M., & Malik, A. (2022). Researching YouTube: Methods, tools, and analytics. In *The SAGE Handbook of Social Media Research Methods* (Vol. 0, pp. 651–663). SAGE Publications Ltd. https://doi.org/10.4135/9781529782943

[10] Lassen, N. B., la Cour, L., & Vatrapu, R. (2022). Predictive analytics with social media data. In N. Buus Lassen, L. la Cour, and R. Vatrapu (Eds.), *The SAGE Handbook of Social Media Research Methods* (pp. 1–30). SAGE Publications Ltd. https://doi.org/10.4135/9781529782943

[11] Lau, T. R., Afizah Afif, Z., Saedudin, R. D. R., Mustapha, A., & Razali, N. (2019). A regression approach for prediction of YouTube views. *Bulletin of Electrical Engineering and Informatics*, *8*(4), 1502–1506. https://doi.org/10.11591/eei.v8i4.1630.

[12] Lee, S. (2017). Style-shifting in vlogging: An acoustic analysis of “YouTube Voice.” *Lifespans & Styles: Undergraduate Working Papers on Intraspeaker Variation*, **3(1)**, 29–32. https://doi.org/10.2218/ls.v3i1.2017.1826

[13] Liu, S. J. (2023). Manipulating Space and Time in Visual Media (Order No. 30742183). Available from ProQuest One Academic. (2901812433). https://login.proxy.tamuc.edu/login?url=https://www.proquest.com/dissertations-theses/manipulating-space-time-visual-media/docview/2901812433/se-2

[14] Mekouar, S., Zrira, N., & Bouyakhf, E.-H. (2017). Popularity prediction of videos in YouTube as case study: A regression analysis study. In *Proceedings of the 2017 International Conference on Big Data and Cloud Applications (BDCA)*. https://doi.org/10.1145/3090354.3090406.

[15] Morgan, J. S., Barjasteh, I., Lampe, C., & Radha, H. (2014). The entropy of attention and popularity in YouTube videos. *Michigan State University*. https://arxiv.org/abs/1412.1185

[16] Park, M., Naaman, M., & Berger, J. (2021). A data-driven study of view duration on YouTube. Proceedings of the International AAAI Conference on Web and Social Media, 10(1), 651–654. https://doi.org/10.1609/icwsm.v10i1.14781

[17] Pinto, H., Almeida, J. M., & Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of YouTube videos. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*, 365–374. https://doi.org/10.1145/2433396.2433443.

[18] Rowe, M. (2011). Forecasting audience increase on YouTube. In *Workshop on User Profile Data on the Social Semantic Web, ESWC 2011*. Heraklion, Greece: Extended Semantic Web Conference. https://oro.open.ac.uk/28845/.

[19] Statista. (2024). Number of social media users worldwide from 2017 to 2028 (in billions) [Graph]. In Statista. Retrieved October 14, 2024, from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[20] Szabo, G., & Huberman, B. A. (2008). Predicting the popularity of online content. *arXiv preprint arXiv:0811.0405*. http://arxiv.org/abs/0811.0405v1.

[21] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, **29(1)**, 24–54. https://doi.org/10.1177/0261927X09351676

[22] theibnahmad. (n.d.). *YouTube channel*. YouTube. Retrieved November 22, 2024, from https://www.youtube.com/@theibnahmad

[23] Wilson, L. T. (2023). The relative importance of click-through rates (CTR) versus watch time for YouTube views. International Journal of Pervasive Computing and Communications, 19(4), 573–595. https://doi.org/10.1108/IJPCC-10-2021-0269

[24] Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities.

[25] Woo, B. K. P. (2019). The Click-Through Rate as a Measure of Dementia Health Promotion on YouTube. Innovations in Clinical Neuroscience, 16(7-08), 11. https://pmc.ncbi.nlm.nih.gov/articles/PMC6850496/

[26] YouTube. (2012). YouTube now: Why we focus on watch time. http://youtubecreator.blogspot.com/2012/08/youtube-now-why-we-focus-on-watch-time.html

[27] Wever, G. S., Elliott, I., McCaul, J., Workman, M., Laubscher, M., Dunn, R. N., & Held, M. (2020). Viewership footprint for a low-resource, student-centred collaborative video platform to teach orthopaedics in southern Africa. *South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde*, *110*(6), 532–536. https://doi.org/10.7196/SAMJ.2020.v110i6.14348.

[28] Zhou, Y., Ahmad, Z., Alsuhabi, H., Yusuf, M., Alkhairy, I., & Sharawy, A. M. (2021). Impact of YouTube Advertising on Sales with Regression Analysis and Statistical Modeling: Usefulness of Online Media in Business. *Computational Intelligence and Neuroscience*, **2021**, 1–10. https://onlinelibrary.wiley.com/doi/full/10.1155/2021/9863155

APPENDICES

APPENDIX A

R code for Figures and Tables From Chapter 4

A.1. R code to make Figure 4.1

In this section, we share the **R** code to make Figure 3.1, a scatterplot of viewership metrics against each of our variables. We used this to show how much influence each analytic has on viewership, which pertains to the boxplots indicated:

```
> library(ggplot2)
> library(dplyr)
> library(showtext)
> library(patchwork)
> showtext_auto()
> create_boxplot <- function(metric) {
+     data1 <- read.csv("24.csv") %>% select(!!sym(metric))
+     data2 <- read.csv("7.csv") %>% select(!!sym(metric))
+     data3 <- read.csv("14.csv") %>% select(!!sym(metric))
+     data4 <- read.csv("21.csv") %>% select(!!sym(metric))
+     data5 <- read.csv("28.csv") %>% select(!!sym(metric))
+     data_combined <- bind_rows(
+         data1 %>% mutate(TimePoint = factor("1", levels = c("1", "7",
"14", "21", "28"))),
+         data2 %>% mutate(TimePoint = factor("7", levels = c("1", "7",
"14", "21", "28"))),
+         data3 %>% mutate(TimePoint = factor("14", levels = c("1", "7",
"14", "21", "28"))),
+         data4 %>% mutate(TimePoint = factor("21", levels = c("1", "7",
"14", "21", "28"))),
+         data5 %>% mutate(TimePoint = factor("28", levels = c("1", "7",
"14", "21", "28")))
+     )
+ data_combined <- data_combined %>% rename(Metric = !!sym(metric))
+ colors <- c("#73787c", "#c6c6c7", "#d7e5f0", "#c9ad93", "#879a77")
+ ggplot(data_combined, aes(x = TimePoint, y = Metric, fill =
+         TimePoint)) +
+         geom_boxplot(outlier.shape = 1) +
+         scale_y_log10() +
+         scale_fill_manual(values = colors) +
+         labs(x = "T I M E   ( D A Y S )", y = metric) +
+         theme_minimal() +
+         theme(
+             axis.title = element_text(face = "bold", size = 12),
+             axis.text = element_text(face = "bold", size = 10),
+             panel.border = element_rect(color = "black", fill = NA),
+             legend.position = "none"
+         )
+ }
> plot_views <- create_boxplot("VIEWS")
> plot_wt <- create_boxplot("WT")
> plot_apv <- create_boxplot("APV")
> plot_avds <- create_boxplot("AVDS")
> plot_ctr <- create_boxplot("CTR")
> plot_imp <- create_boxplot("IMP")
> final_plot <- (plot_views | plot_wt | plot_apv) /
+     (plot_avds | plot_ctr | plot_imp)
> print(final_plot)
```

A.2. R code to make Figures 4.2 and 4.3

R code used to make Figure 4.2 and 4.3. The correlation matrices in this appendix illustrate the relationships between key YouTube performance metrics, including Click-Through Rate (CTR), Impressions (IMP), Audience Retention, Average Percentage Viewed (APV), Views (VIEWS), and Watch Time (WT).

```
> library(ggplot2)
> library(GGally)
> library(showtext)
> library(magick)
> library(qpdf)
> showtext_auto()
> day_colors <- c("1 DAY" = "#73787c", "7 DAYS" = "#c6c6c7", "14 DAYS"
= "#d7e5f0", "21 DAYS" = "#c9ad93", "28 DAYS" = "#879a77")
> panel.cor <- function(data, mapping, digits = 2, prefix = "r = ", ...)
+ {
+   x <- eval_data_col(data, mapping$x)
+   y <- eval_data_col(data, mapping$y)
+   r <- abs(cor(x, y, use = "complete.obs"))
+   txt <- paste0(toupper(prefix), " ", format(r, digits = digits))
+   ggally_text(
+     label = txt,
+     mapping = aes(),
+     xP = 0.5, yP = 0.5,
+     family = "Gotham Bold", fontface = "bold", size = 2.5,
+     color = "black"
+   ) + theme_void()
+ }
> create_and_save_pairs_plot <- function(data, title, filename) {
+   p <- ggpairs(
+     data,
+     columns = c("VIEWS", "WT", "APV", "AVDS", "IMP", "CTR"),
+     lower = list(continuous = wrap("points", size = 1.5, shape = 21,
      color = day_colors[title], fill = NA)),
+     upper = list(continuous = wrap(panel.cor)),
+     diag = list(continuous = wrap("densityDiag"))
+   ) +
+     theme_minimal(base_size = 12) +
+     theme(
+       text = element_text(family = "Gotham Bold", face = "bold"),
+       axis.text.x = element_text(family = "Gotham Bold", face = "bold",
        size = 8, angle = 45, hjust = 1),
+       axis.text.y = element_text(family = "Gotham Bold", face = "bold",
        size = 8),
+       axis.title = element_text(family = "Gotham Bold", face = "bold",
        size = 12),
+       plot.title = element_text(family = "Gotham Bold", face = "bold",
        size = 14, hjust = 0.5),
+       panel.grid.major = element_line(size = 0.3),
+       panel.grid.minor = element_line(size = 0.2)
+     ) +
+     ggtitle(title)
+   ggsave(filename, plot = p, width = 6, height = 6)
+ }
> data_24 <- read.csv("24.csv")
> data_7 <- read.csv("7.csv")
> data_14 <- read.csv("14.csv")
> data_21 <- read.csv("21.csv")
> data_28 <- read.csv("28.csv")
> create_and_save_pairs_plot(data_24, "1 DAY", "plot_1day.pdf")
> create_and_save_pairs_plot(data_7, "7 DAYS", "plot_7days.pdf")
> create_and_save_pairs_plot(data_14, "14 DAYS", "plot_14days.pdf")
> create_and_save_pairs_plot(data_21, "21 DAYS", "plot_21days.pdf")
> create_and_save_pairs_plot(data_28, "28 DAYS", "plot_28days.pdf")
> pdf_combine(c("plot_1day.pdf", "plot_7days.pdf", "plot_14days.pdf",
"plot_21days.pdf", "plot_28days.pdf"),
+ output = "combined_plots_multipage.pdf")
> images <- image_read_pdf("combined_plots_multipage.pdf")
> final_image <- image_montage(images, geometry = "x6", tile = "2x3")
> image_write(final_image, path =
    "combined_single_page_plots_larger_canvas.pdf", format = "pdf")
```

A.3. R code to make Table 4.1

R code used to extract data from our dataset (including outliers) for Table 4.1, which is the correlation matrices demonstrating the r values between variables at the specified time points.

```
> data24 = read.csv("24.csv")
> datanew = data24[, c("VIEWS", "WT", "APV", "AVDS", "IMP", "CTR")]
> cor(datanew)
> summary(datanew)
> model24 = lm(VIEWS ~ ., data = datanew)
> summary(model24)
> data7 = read.csv("7.csv")
> datanew = data7[, c("VIEWS", "WT", "APV", "AVDS", "IMP", "CTR")]
> cor(datanew)
> summary(datanew)
> model24 = lm(VIEWS ~ ., data = datanew)
> summary(model24)
> data14 = read.csv("14.csv")
> datanew = data14[, c("VIEWS", "WT", "APV", "AVDS", "IMP", "CTR")]
> cor(datanew)
> summary(datanew)
> model24 = lm(VIEWS ~ ., data = datanew)
> summary(model24)
> data21 = read.csv("21.csv")
> datanew = data21[, c("VIEWS", "WT", "APV", "AVDS", "IMP", "CTR")]
> cor(datanew)
> summary(datanew)
> model24 = lm(VIEWS ~ ., data = datanew)
> summary(model24)
> data28 = read.csv("28.csv")
> datanew = data28[, c("VIEWS", "WT", "APV", "AVDS", "IMP", "CTR")]
> cor(datanew)
> summary(datanew)
> model24 = lm(VIEWS ~ ., data = datanew)
> summary(model24)
```

VITA

Numan Ahmad was born in Johnson City, Tennessee, on February 27, 2002. He attended elementary school in Lewisville Independent School District, and graduated from Brighter Horizons Academy near the top of his class in 2021. While attending, he was awarded with Academic Excellence awards in the following AP courses: Language and Composition, U.S. History, Biology, Calculus, and World History. He was also recognized as an AP Scholar and granted the National Merit Scholarship Commendation award for his final years of high school.

The following August of 2021, he attended East Texas A&M University and is set to graduate with Bachelor of Science in Computer Science in May 2025. Numan made the President's List for the semesters of Spring 2024 at East Texas A&M University.

Numan was also awarded an abundance of monetary scholarships as a result of his academic excellence, including: A full ride scholarship by East Texas A&M University via the Honors College, Naderkhani Endowed Scholarship Fund ($1000), Castle Hills Community Scholarship ($2000) Gwen Smith Greek Council Award ($500), Michael O'Malley MD Memorial Rural Youth Physician's Scholarship ($1000), and has helped the Student Honors Council with obtaining university funding ($300).

Numan Ahmad
Department of Computer Science
East Texas A&M University
P.O. Box 3011
Commerce, TX 75429-3011
nahmad2@leomail.tamuc.edu