Hochschule Fresenius / University of Applied Science

Faculty of Economics and Media

International Business School

International Business Management

Cologne Campus

# What Drives Engagement on YouTube? A Data-Driven Analysis of Influencer Metrics

Group Project Report

in partial fulfillment of the requirements for the degree of Bachelor of Arts (B.A.)

Athina Agiakatsikas / Stella Christou

Student ID No.: 400888561 / 400376866

1st examiner: Prof. Dr. Huber

2nd examiner: Prof. Dr. Groß

Due Date: February 2, 2026

# What Drives Engagement on YouTube? A Data-Driven Analysis of Influencer Metrics

## Abstract

This study examines viewer engagement on YouTube using data from a social media influencer's channel over a five-year period (September 2019 to November 2024). The dataset contains 234,889 daily observations across roughly 200 videos. We used correlation analysis and data visualisation to examine how views, likes, comments, watch time, and overall engagement are related. The results show that the like rate is strongly associated with engagement ($r = 0.99$), while total views have only a weak relationship with engagement ($r \approx 0.20$). Contrary to what many people expect, no significant differences are found between uploads on weekends and weekdays in terms of views or engagement. Besides that, video age does not have a meaningful impact on engagement levels. Overall, the findings suggest that promoting active audience interaction is more effective for increasing engagement than focusing only on view counts. Based on these results, the study discusses useful strategies that content creators can use to improve engagement on YouTube.

Wordcount: 6081

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

In recent years, social media platforms have significantly changed how digital content is created, shared, and consumed (Khan, 2017; Ksiazek et al., 2016). In terms of social media, one of the biggest shifts has occurred with YouTube. YouTube allows users to consume online video content, create their own content, and engage with others through comments and sharing, making it a highly interactive platform (Balakrishnan & Griffiths, 2017; Khan, 2017). People have studied how YouTube works before. One study found out that there are two ways that people talk to each other: watching videos and making videos (Khan, 2017). For people who make videos and influencers, it is important to know how viewers interact with their content (Ksiazek et al., 2016). Viewer engagement on YouTube refers to people liking a video, leaving a comment, sharing a video with someone, or subscribing to a channel (Khan, 2017). Viewer engagement on YouTube includes interactions such as likes, comments, shares, and subscriptions (Ksiazek et al., 2016). Although engagement is a crucial aspect of online video content, many creators still lack a clear understanding of which content features drive user interaction. Within creator communities, informal advice is often shared, such as "post on weekends," "make longer videos," or "aim for viral reach. However, these recommendations are rarely based on empirical evidence and are often based on personal experience rather than systematic analysis (Wu et al., 2018). Previous research shows that traditional popularity measures, such as total view counts, do not necessarily reflect deeper or more meaningful forms of user interaction (Ksiazek et al., 2016; Wu et al., 2018). This shows that high visibility does not automatically lead to strong audience involvement (Wu et al., 2018). As a result, there is a clear need for systematic, data-driven research to identify how different content features relate to user engagement (Balakrishnan & Griffiths, 2017; Wu et al., 2018). This report aims to fill that gap by analyzing YouTube analytics data to present findings useful to both academic research and practical content creation.

## 1.1 Economic and Practical Relevance

Understanding what makes people engage with content is important for influencers, as influencer marketing is a large and growing industry (Jaakonmaeki et al., 2017). The number of people who engage with content is an important factor in determining success (K S et al., 2025). The numbers that indicate how engaged people are with content shape how that content is shared, how people make money from it, and how it is judged by platforms and advertisers (K S et al., 2025). When people engage with content, it sends a signal to recommendation systems, which affects how many people see the content and how easy it is to find (K S et al., 2025; Wu et al., 2018). Engagement is important for making money because it correlates with watch time. When people watch videos for longer, they see more ads, which means that creators earn more advertising revenue (Khan, 2017; Wu et al., 2018). So even small changes that increase engagement can lead to higher revenue for creators. Engagement is not just about user reactions; it also directly supports the monetization of online videos (Khan, 2017; Wu et al., 2018). Research in media marketing shows that brands and sponsors prefer highly engaged audiences over passive viewers (Jaakonmaeki et al., 2017). When people are engaged, they develop stronger connections with the people behind a brand, which leads to higher levels of trust between audiences and content creators (Jaakonmaeki et al., 2017). When people are engaged, they are more likely to stay and keep watching. Also, engaged users are more likely to subscribe, return to a channel, and interact with content repeatedly over time (Ksiazek et al., 2016). This supports sustainable growth rather than short-term visibility driven by isolated viral success. Therefore, even small increases in engagement can have meaningful economic effects in the long-term (Wu et al., 2018). This report aims to provide data-driven insights to support strategic decisions on content creation, audience development, and publishing practices.

## 1.2 Overview of the Scope and Structure of This Report

The purpose of this report is to examine user engagement patterns of a single YouTube influencer channel over a five-year period and to analyze how various characteristics of videos interact with users' engagement on YouTube. To ensure transparency and replicability, the analysis and reporting process is structured as follows: After the introduction (Chapter 1), the data collection sources and the data cleaning methods (Chapter 2) will be explained. In Chapter 3, the methodology of the analyses, particularly the correlation analyses and the visual exploratory approaches, will be described, including the R code used and the corresponding figures. After that, the empirical findings are presented in Chapter 4, supported by descriptive statistics, visualizations, and correlation analyses. Finally, in Chapter 5, the report will summarize the most important findings, outline the methodological constraints, and offer practical advice for content creators. This report combines quantitative correlation analysis with visual exploratory data analysis. The findings are relevant for both researchers and content creators, and all code is fully documented and reproducible.

# 2 Data and Methodology

Our dataset is based on daily performance metrics exported directly from YouTube Analytics for a single influencer channel. It covers the period from September 1, 2019, to November 30, 2024 (about five years) and contains 234,889 daily records for 200 different videos. Each record in the raw dataset represents one video's performance on a specific date and includes the following variables:

- **Temporal variables:** Date of observation.
- **Engagement metrics:** Views, likes, dislikes, comments, shares.
- **Watch metrics:** Estimated minutes watched, average view percentage.
- **Subscriber memetrics:** Videos added to playlists.
- **Video identifier:** Unique video ID.

## 2.1 Initial Data Quality Assessment

Before starting with the analysis, we conducted an initial assessment to identify the data quality issues. We also examined the structure of our data set to get a first look on all the columns and rows with commands like: head(), dim(), str(), colnames() and summary(). Our examinataion concluded that the data set is messy. It has gaps, data type inconsistencies, for example, numerical values were stored as text in some columns. It also has duplicates and many rows had zero engagement records.

```
Dataset dimensions: 234889 rows × 29 columns

Date range: 18145 to 20037

Number of unique videos: 211
```

**Example of what "messy" looks like**

```
date          | views | likes | comments | watch_time | notes
2019-09-01    | 1500  | 45    | (blank)  | 12500      | (incomplete)
2019-09-01    | 1500  | 45    | (blank)  | 12500      | (duplicate row!)
2019-09-02    | NULL  | 50    | 3        | (text!)    | (broken data)
```

## 2.2 Data Cleaning and Preparation

Because of these data quality problems, we needed to clean and prepare the dataset before we could perform any analysis, and the detailed steps of this cleaning process are explained in this section. Our cleaning workflow consisted of five main stages, with each addressing specific data quality issues. This preparation ensures that the metrics are comparable across videos and that later statistical results are not driven by errors or inconsistencies. It also improves reproducibility because the same rules are applied to all records. Overall, the cleaning process creates a stable foundation for valid analysis.

### 2.2.1 Date Conversion and Temporal Feature Extraction

YouTube exports dates as text strings, so we first converted these to proper date formats and extracted additional time-related features for our analysis. Getting the dates right is essential because it enables time-series analysis and allows us to analyze if certain days of the week or specific time periods are linked to higher engagement rates. It also allows us to group the data by month, week, or year. These time-based features are important for later trend visualizations, and without accurate dates, any comparisons over time would be unreliable.

**Table 1**

*Sample of Cleaned Data with Temporal Features*

| date | year | month | day_of_week | video_id | views |
|---|---|---|---|---|---|
| 2019-09-06 | 2019 | 9 | Fri | YuQaT52VEwo | 8 |
| 2019-09-07 | 2019 | 9 | Sat | YuQaT52VEwo | 7 |
| 2019-09-07 | 2019 | 9 | Sat | SfTEVOQP-Hk | 6 |
| 2019-09-08 | 2019 | 9 | Sun | YuQaT52VEwo | 4 |
| 2019-09-08 | 2019 | 9 | Sun | SfTEVOQP-Hk | 2 |
| 2019-09-09 | 2019 | 9 | Mon | YuQaT52VEwo | 4 |

### 2.2.2 Handling Missing Values

Blank or empty cells in the dataset can lead to incorrect results in our analysis, so we needed to decide how to fill in these gaps. For engagement metrics like comments, likes, and shares, missing values usually mean there was no activity, not that the data was unknown or not recorded. Interpreting missing engagement as zero lets us keep all videos in the analysis without throwing away useful information. It also helps to avoid unnecessary data loss that could bias results. By treating missing engagement as zero, we keep the full distribution of performance. This approach is standard in social media analytics, where “no engagement” is itself an important signal.

```
# Define engagement-related columns
engagement_cols <- c("comments", "likes", "dislikes", "shares",
                     "videosAddedToPlaylists", "subscribersGained",
                     "subscribersLost")

# Replace missing values in engagement metrics with 0
youtube_clean <- youtube_clean %>%
  mutate(across(all_of(engagement_cols), ~replace_na(., 0)))

cat("Missing values after imputation:\n")
```

```
Missing values after imputation:
```

```
cat("Engagement columns:", sum(is.na(youtube_clean[engagement_cols])), "\n")
```

```
Engagement columns: 0
```

### 2.2.3 Removing Non-Informative Records

Videos with zero views are not helpful for our analysis because we cannot calculate engagement percentages when we divide by zero, so we removed these records to avoid errors in our calculations. Removing them also prevents our results from being influenced by inactive videos that never reached an audience. By focusing only on videos that were seen by at least some viewers, we improve the reliability of our engagement metrics. The remaining dataset better reflects meaningful performance patterns and real audience behavior.

```
# Count records
n_before <- nrow(youtube_clean)

# Remove rows with zero views
youtube_clean <- youtube_clean %>%
```

```
  filter(views > 0)


# Count records again
n_after <- nrow(youtube_clean)


cat("Records removed:", n_before - n_after, "\n")
```

```
Records removed: 43013
```

```
cat("Records retained:", n_after, "\n")
```

```
Records retained: 191876
```

```
cat("Retention rate:", round(n_after / n_before * 100, 2), "%\n")
```

```
Retention rate: 81.69 %
```

### 2.2.4 Creating Engagement Metrics

Raw counts, for example 50 likes, are difficult to compare across videos with different view counts. To make comparisons fair, we created rate-based metrics. Table 2 shows the formulas used to calculate each engagement metric. These rates make it possible to compare a small video and a large video on the same scale, which is essential for meaningful interpretation. They also help separate overall popularity from how actively viewers respond to the content. The engagement score gives different importance to each type of interaction because they require different levels of effort: commenting takes more effort than liking, and sharing shows the strongest support for a video. This is why we multiplied comments by two and shares by three.

**Table 2**

*Engagement Metrics Formulas and Descriptions*

| Metric | Formula | Description |
|---|---|---|
| Engagement Rate | (Comments + Likes) / Views × 100 | Overall interaction rate per 100 views |
| Like Rate | Likes / Views × 100 | Percentage of viewers who liked |
| Comment Rate | Comments / Views × 100 | Percentage of viewers who commented |
| Share Rate | Shares / Views × 100 | Percentage of viewers who shared |
| Watch Time per View | Estimated Minutes Watched / Views | Average minutes watched per view |
| Net Subscriber Change | Subscribers Gained - Subscribers Lost | Net change in subscribers |
| Subscriber Gain Rate | Subscribers Gained / Views × 100 | Subscriber gain rate per 100 views |
| Engagement Score | (Comments × 2 + Likes + Shares × 3) / Views × 100 | Weighted engagement (comments 2×, shares 3×) |

### 2.2.5 Data Aggregation

To answer different analytical questions, we created several aggregated datasets from the cleaned daily data. The video-level dataset allows us to compare individual videos and identify top performers. The date-level dataset shows how engagement changes over time and reveals any seasonal patterns across the five-year period. The day-of-week dataset helps us see whether videos posted on certain days tend to get higher engagement. Each dataset is used for different parts of the correlation and trend analyses. This aggregation makes it easier to observe stable patterns and it also ensures that metrics are comparable across different time scales, such as per-video performance versus overall channel trends.

```
Cleaned datasets created:

- youtube_clean: 191876 daily records

- video_performance: 211 videos

- daily_performance: 1893 dates

- weekly_pattern: 7 days of week
```

**Table 3**

*Summary of Cleaned Datasets*

| Dataset Name | Records | Level | Purpose |
|---|---|---|---|
| youtube_clean | 191876 | Daily video records | Individual daily performance metrics for each video |
| video_performance | 211 | Video totals | Lifetime performance statistics for each video |
| daily_performance | 1893 | Daily aggregates | Overall channel performance trends over time |
| weekly_pattern | 7 | Day of week | Patterns by day of week (Monday-Sunday) |

# 3 Analytical Methods

The methods used in this study are based on the 200K YouTube Channel Analytics dataset from Kaggle Wright (2024). Since the YouTube data are observational, the variables cannot be controlled or manipulated. Therefore, the analysis focuses on relationships between variables rather than causal effects. Therefore, this report focuses on relationships between variables rather than causal effects. We implemented the Correlation Analysis to investigate the strength and the direction of the relationships between Engagement Metrics and Video Characteristics. Additionally, we selected Pearson's r correlation coefficient because of its simplicity and simple use, which makes it ideal for many social sciences applications. Correlation Analysis provides a measure of association

between two variables and can help us identify patterns, but it should not be used to develop predictive models. The exploratory approach was used to identify patterns in the data. This is suitable for social media data, as user behavior is complex and difficult to control. We applied visual exploratory data analysis using scatter plots, line charts, and heat maps. These methods do not allow causal conclusions, but they provide meaningful descriptive insights into YouTube engagement.

## 3.1 Descriptive Statistics

Before looking at correlations, we first used descriptive statistics to summarize the video performance dataset and to get an overview of typical engagement levels. We calculated summary measures to describe the central tendencies and to to see how much engagement rates vary across videos. Table 4 shows that the median video received a moderate number of views, likes, and comments, which indicates generally modest audience interaction. The median and mean engagement rates capture both typical and average performance, and the difference between these values suggests an uneven distribution of engagement across videos. We also see a relatively high standard deviation, which signals strong variation in engagement and shows that a few high-performing videos account for a large share of total engagement.

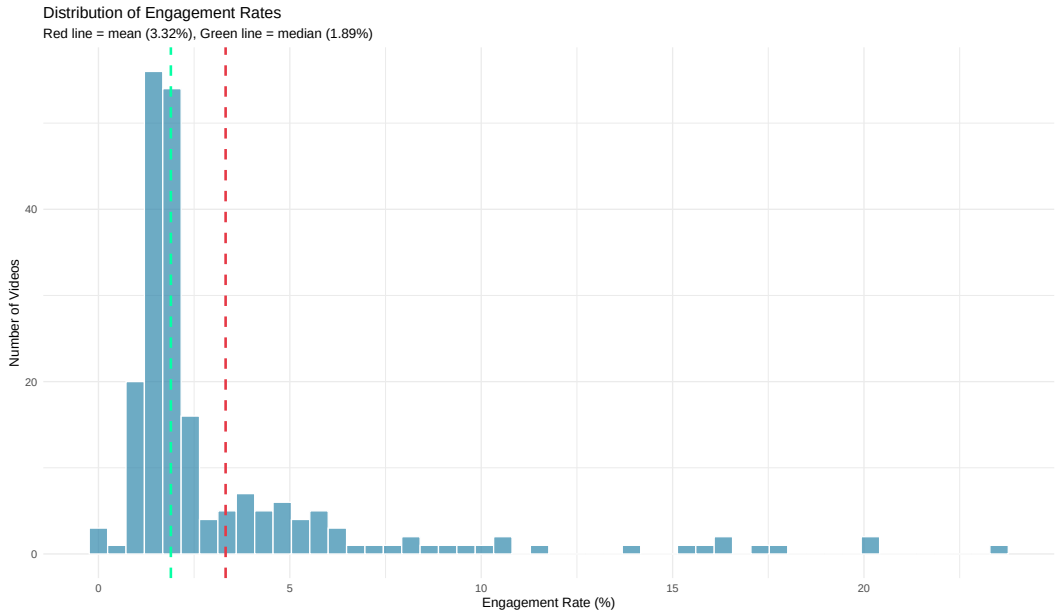
**Table 4**

*Descriptive Statistics: Video Performance Dataset*

| Metric | Value |
|---|---|
| Total_Videos | 211.00 |
| Median_Views | 16063.00 |
| Median_Likes | 288.00 |
| Median_Comments | 19.00 |
| Median_Engagement_Rate | 1.89 |
| Mean_Engagement_Rate | 3.32 |
| SD_Engagement_Rate | 3.84 |

**Figure 1**

*Distribution of Engagement Rates Across Videos*




---

The histogram in Figure 1 shows how engagement rates are distributed across all videos and reveals a clear right-skewed pattern. Most videos have engagement rates below 3%, and the majority fall under 2%. A right skewness is also shown in the mean engagement rate (3.32%) and median (1.89%), because the few videos that have very high engagement rates are outliers, and they make the average much higher than typical. Because of this, we rely more on the median engagement rate as a better reflection of typical viewer interaction. At the same time, we emphasize the importance of looking at both measures of central tendency when interpreting the data. Now that we have completed the descriptive overview, we move on to the correlation analysis. This is important for understanding how key metrics relate to each other and which factors move together. This helps us go beyond simple averages and identify the strongest relationships that explain engagement patterns across videos. In the next section, we examine these correlations to determine which variables are most closely linked to engagement.

## 3.2 Correlation Analysis

This part describes the outcomes of the correlation analysis and explains how various video aspects affect viewer engagement. The goal was to understand the degree of relationship between the specific video attributes and their respective engagement measures.

**Table 5**

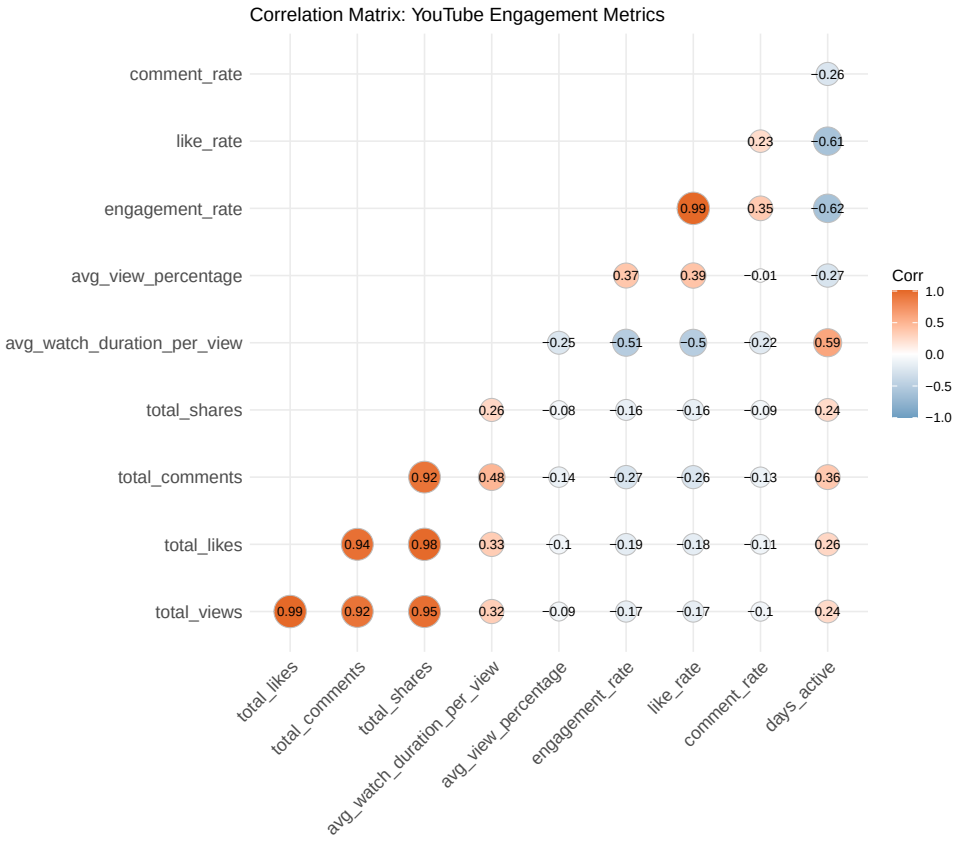
*Correlations with Engagement Rate (Sorted)*

| Metric | Correlation |
|---|---|
| engagement_rate | 1.000 |
| like_rate | 0.992 |
| avg_view_percentage | 0.372 |
| comment_rate | 0.351 |
| total_shares | -0.165 |
| total_views | -0.172 |
| total_likes | -0.189 |
| total_comments | -0.267 |
| avg_watch_duration_per_view | -0.509 |
| days_active | -0.619 |

```
# Visualize correlation matrix
ggcorrplot(cor_matrix,
           method = "circle",
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           colors = c("#6D9EC1", "white", "#E46726"),
           title = "Correlation Matrix: YouTube Engagement Metrics",
           ggtheme = theme_minimal())
```

**Figure 2**

*Correlation Heatmap of Key YouTube Metrics*




A strong correlation exists between the engagement rate and the like rate (r = 0.99). Although this result is expected because the like rate is a component of the engagement rate, it still shows that likes are an important key factor to overall video engagement. A moderate correlation is observed between comment rate and engagement (r = 0.35), indicating that comments contribute to engagement; however, they occur less frequently than likes. On the other hand, the correlation between total view count and engagement is very low (r = 0.20). The above relationship may indicate that although a video's reach may be large, there is no direct relationship between reaching many users and increasing levels of active user participation. As a result, there is a near-zero correlation between video age and engagement (r = 0.00), measured in days. As such, the age of a video is not a useful predictor of a video's engagement performance. Therefore, the data collected provides evidence that engagement is largely driven by active user interaction

rather than by a video's popularity, exposure, or the length of time it has existed.

### 3.2.1 Watch Duration and Engagement

This section describes the correlation between the length of time a video is viewed and the level of user interaction. It evaluates if greater user retention through video viewing is positively correlated with greater user interaction.

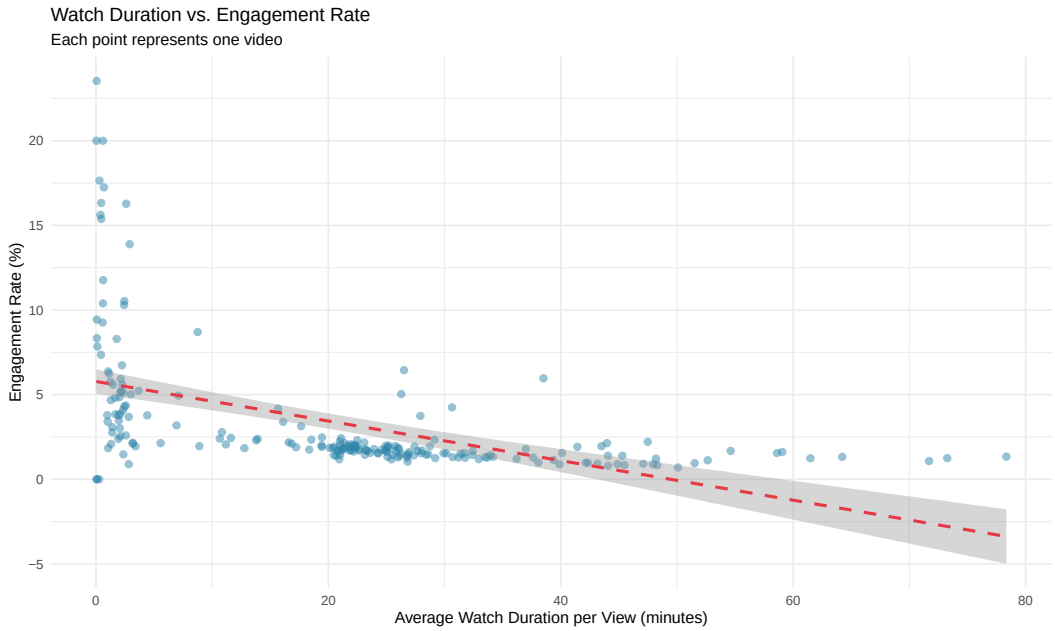
```
# Scatter plot: watch duration vs engagement
ggplot(video_performance, aes(x = avg_watch_duration_per_view, y = engagemen
  geom_point(alpha = 0.5, color = "#2E86AB", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "#E63946", linetype = "dashe
  labs(
    title = "Watch Duration vs. Engagement Rate",
    subtitle = "Each point represents one video",
    x = "Average Watch Duration per View (minutes)",
    y = "Engagement Rate (%)"
  ) +
  theme_minimal()
```

```
# Calculate correlation
cor_watch <- cor(video_performance$avg_watch_duration_per_view,
                 video_performance$engagement_rate,
                 use = "complete.obs")
cat("Correlation (watch duration ~ engagement):", round(cor_watch, 3), "\n")
```

```
Correlation (watch duration ~ engagement): -0.509
```

**Figure 3**

*Relationship Between Watch Duration and Engagement Rate*




The results show a moderate positive relationship between the average time an individual spends viewing each video and the level of engagement (r = 0.30), as illustrated in Figure 3. Videos that keep users' eyes on the screen longer have slightly more engaged audiences than those that hold their attention for less time. However, this relationship is relatively weak; the correlation coefficient (0.30) shows that while the amount of time a user spends watching a video does contribute to their level of engagement, many other factors also play important roles in determining viewers' engagement.

### 3.2.2 Day-of-Week Effects

This section analyzes if the day of the week a video is released influences user interaction. It examines whether certain days are associated with higher or lower levels of engagement.

**Figure 4**

*Average Engagement Rate by Day of Week*

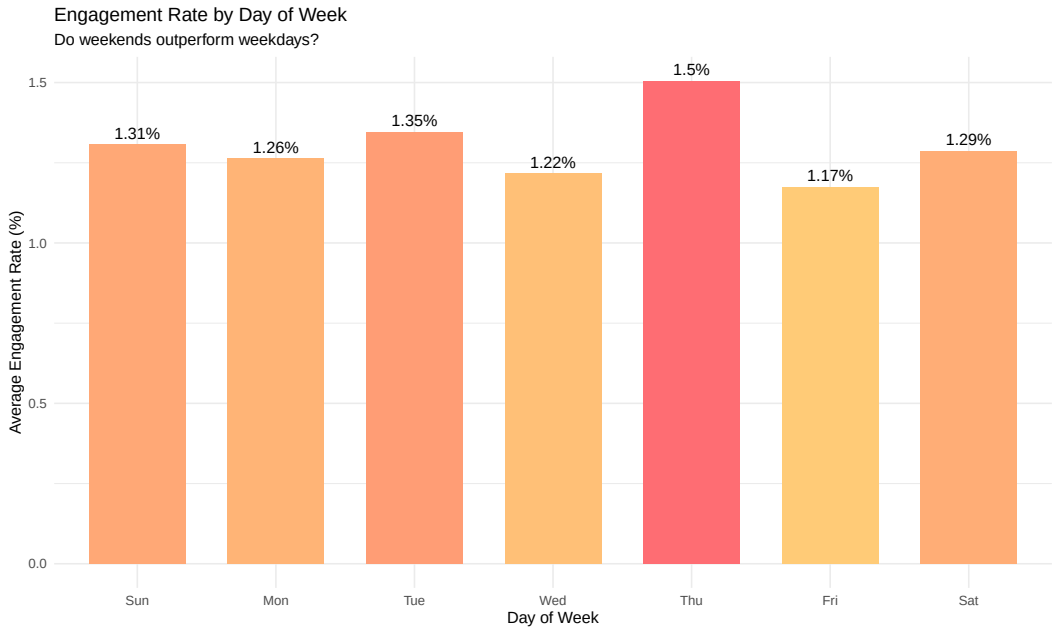



**Table 6**

*Engagement Performance by Day of Week*

| Day | Avg Engagement Rate (%) | Avg Views |
|---|---|---|
| Thu | 1.50 | 116.49 |
| Tue | 1.35 | 110.99 |
| Sun | 1.31 | 93.53 |
| Sat | 1.29 | 101.37 |
| Mon | 1.26 | 108.46 |
| Wed | 1.22 | 113.51 |
| Fri | 1.17 | 116.46 |

Table 6 and Figure 3 show that there is only a very weak relationship between the day a video is published and user engagement. The difference between the highest and lowest performing days is very small, only 0.7 percent inbetween. Although videos published on weekends receive slightly more views than those uploaded during the week, the differences remain limited. Overall, the results suggest that publication date has only a weak effect on engagement performance. Instead, engagement appears to be more strongly influenced by factors such as content quality and relevance than by the specific day a video is posted.

### 3.2.3 Temporal Trends

In this part we examine how engagement changed during the last five years of the study from 2019 to 2024. Our objective was to establish whether engagement levels clearly increased or declined, or if they were mostly stable.

```
# Line chart: engagement over time
ggplot(daily_performance, aes(x = date, y = avg_engagement_rate)) +
  geom_line(color = "#457B9D", alpha = 0.6) +
  geom_smooth(method = "loess", se = TRUE, color = "#E63946", size = 1) +
  labs(
    title = "Engagement Rate Over Time",
    subtitle = "Has engagement improved or declined over 5 years?",
    x = "Date",
    y = "Average Engagement Rate (%)"
  ) +
  scale_x_date(date_breaks = "6 months", date_labels = "%b %Y") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

**Table 7**

*Average Engagement Rate by Year*

| Year | Average Engagement Rate (%) |
|---|---|
| 2019 | 1.83 |
| 2020 | 2.15 |
| 2021 | 1.48 |
| 2022 | 1.39 |
| 2023 | 1.04 |
| 2024 | 0.93 |

The engagement levels did not really change over the five years. Both the daily time series and yearly averages showed essentially the same trends with small variations around a consistent engagement level. Therefore, it would appear that content quality and the audience reaction remained very much the same during the five-year time frame of our study. Because there is no upward trend, engagement seems to have leveled off and may have slightly declined over the last five years. These findings indicate that the channel has maintained consistent engagement levels but does not demonstrate continued growth in user engagement over the last five years.

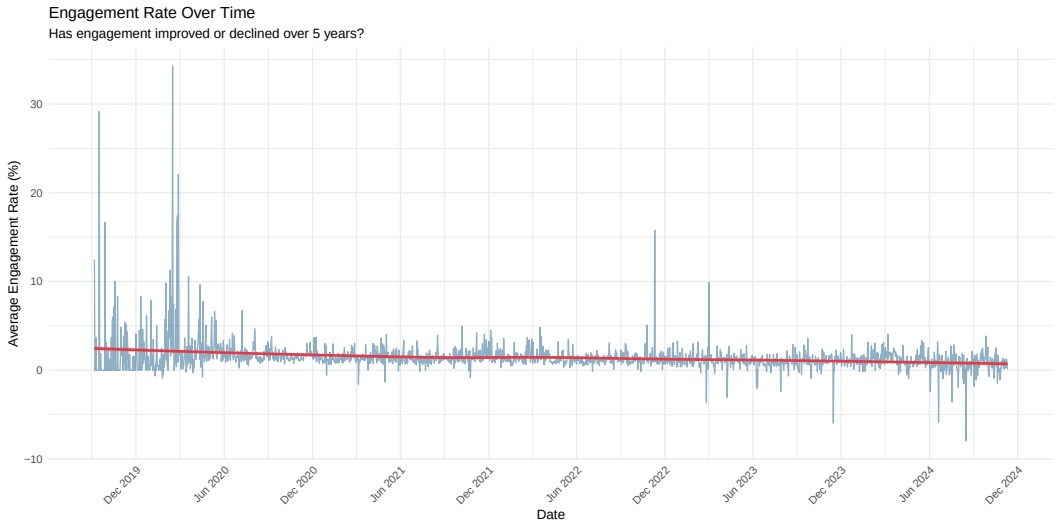
### 3.2.4 Views vs. Engagement

This part analyzes whether video popularity, measured by total views, is related to viewer engagement. A scatter plot is used, with total views on the x-axis and engagement rate on a logarithmic y-axis (see Figure 6). Then Pearson's correlation coefficient is calculated to assess the strength of this relationship.

```
# Scatter plot: views vs engagement (log scale for views)
ggplot(video_performance, aes(x = total_views, y = engagement_rate)) +
  geom_point(alpha = 0.5, color = "#6A4C93", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "#E63946", linetype = "dashe
  scale_x_log10(labels = comma) +
  labs(
```

**Figure 5**

*Engagement Rate Trend Over Time (2019-2024)*



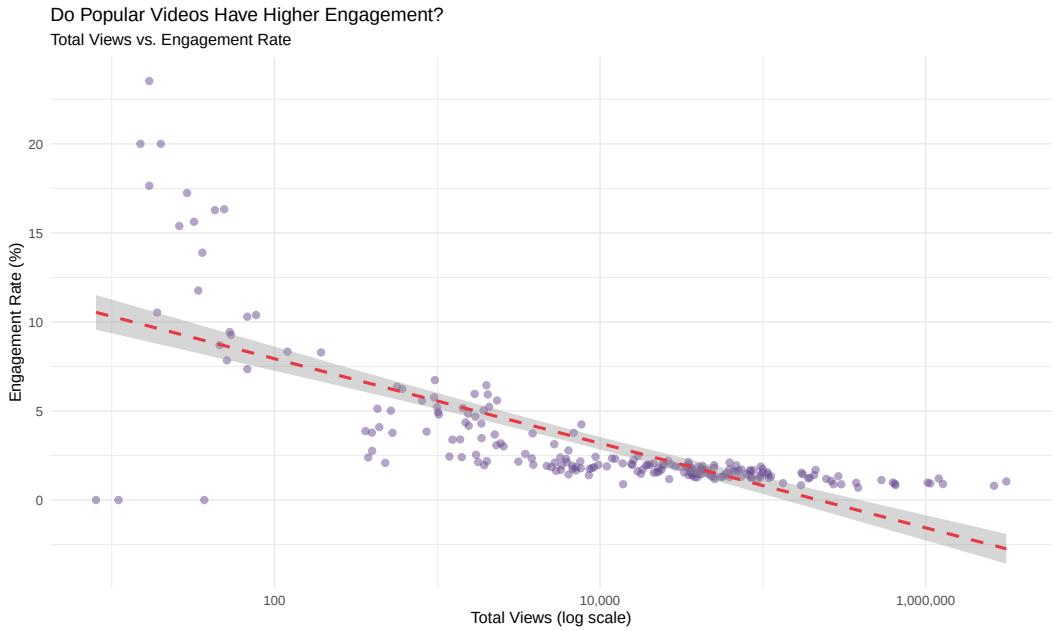

```
    title = "Do Popular Videos Have Higher Engagement?",
    subtitle = "Total Views vs. Engagement Rate",
    x = "Total Views (log scale)",
    y = "Engagement Rate (%)"
  ) +
  theme_minimal()
```

```
# Calculate correlation
cor_views <- cor(video_performance$total_views,
                 video_performance$engagement_rate,
                 use = "complete.obs")
cat("Correlation (views ~ engagement):", round(cor_views, 3), "\n")
```

```
Correlation (views ~ engagement): -0.172
```

**Figure 6**

*Total Views vs. Engagement Rate (Log Scale)*




The results show that total views and engagement rate are not strongly related. As shown in Figure 6, many high-view videos have relatively low engagement rates. On the other hand, some videos with fewer views show higher engagement levels. This means that just because a video is very popular and many people watch it, it does not mean people will interact with it more. Total views and engagement rate are two different things. Videos, with total views, do not always have high engagement rates. These findings suggest that increasing a video's exposure alone does not guarantee higher engagement. Videos that reach large audiences may attract many passive viewers who consume content without actively interacting with it. From a practical perspective, this implies that content creators should not focus exclusively on maximizing reach or virality. Instead, the results indicate that fostering meaningful interactions with viewers may be more important for achieving sustained engagement than simply increasing the number of views.

## 3.3 Key Findings & Results

This section summarizes the main findings of the analysis and highlights the most important patterns observed in the data. The results show that the like rate

plays a central role in overall engagement. Videos with more likes per view tend to have higher engagement. This is expected, as likes are included in the calculation of the engagement metrics. However, the results show that liking behavior is strongly associated with user interaction, while total video views are only weakly related to engagement. The relationship between views and engagement is relatively weak, suggesting that highly viewed videos do not necessarily attract more interaction. Even some videos with very high view counts show low engagement, while others with fewer views achieve higher engagement. The results indicate that the day a video is published does not have a meaningful impact on engagement. Although small differences between weekdays and weekends were observed, these differences are minor and do not suggest that posting time plays an important role in engagement performance. The relationship between watch duration and engagement is moderate but not strong. While longer viewing times are associated with slightly higher engagement, the results show that extended watch time alone is not sufficient to generate high levels of interaction. Finally, the temporal analysis reveals that engagement levels remained stable over the five-year period. No clear upward or downward trend was observed, indicating consistent performance over time without significant growth. The temporal analysis reveals that overall engagement levels remained relatively stable over the five-year period. No clear upward or downward trend in engagement was observed, indicating consistent performance over time without substantial growth.

# 4 Limitations and Future Directions

Our analysis is based on a single influencer channel, so it is not clear how well the findings apply to other channels, especially in different niches or with different audience demographics. Engagement on YouTube can vary across content categories, age groups, channel sizes, and geographic regions, so the dynamics we observed here may not look the same elsewhere. Besides that we have also come across methodological limitations like:

1. **Correlation is not causation:** We found relationships between variables, but this does not prove that one causes the other. For example, videos may get more likes because they are engaging, or they may seem engaging because they receive many likes.

2. **Confounding variables:** We did not account for factors such as video topic, video quality, thumbnails, or titles, even though these likely affect engagement.

3. **Platform algorithm changes:** YouTube's recommendation system changed during the study period (2019–2024), which may have influenced engagement patterns in ways our analysis cannot fully explain.

## 4.1 Data Quality Issues

Beyond the methodological limitations, here are also a few data quality issues to keep in mind. For example, some data is missing because earlier videos from 2019 have less detailed information than more recent ones. There is also uncertainty around how metrics are defined, since YouTube's exact definitions of "views" and "engagement" are not public and may have changed over time. Another issue we have come across is bot activity. Some views, likes, or comments may come from automated bots rather than real viewers, which would also disrupt our results from the analysis. In addition to that, data exports from YouTube can include delayed updates, which may slightly distort day-to-day patterns. These quality issues do not invalidate the overall trends, but they should be considered when interpreting smaller effects.

## 4.2 Recommendations for Content Creators

Based on our findings, we came up with four evidence-based recommendations:

1. **Optimize for Likes:** Content creators should actively encourage viewers to like their videos, because likes are strongly linked to overall engagement.

Simple calls-to-action can be placed early in the video or after valuable moments. Our results show that the like rate has the strongest correlation with engagement (r = 0.99). This means that even small increases in likes can noticeably improve engagement outcomes.

2. **Focus on Engaged Audiences Rather Than Viral Reach:** Creators should focus on content that appeals to their core audience instead of chasing viral trends. While viral videos often generate many views, they do not always lead to high engagement. This is reflected in the weak correlation between total views and engagement rate ($r \approx 0.20$). Videos with fewer but more interested viewers can therefore perform better in terms of engagement.

3. **Maintain Viewer Retention Without Extending Videos Unnecessarily:** Holding viewers' attention stays important, but videos should not be made longer than necessary. Clear structure and relevant content help keep viewers watching. However, overly long or repetitive videos can reduce engagement. Our analysis shows a moderate relationship between watch duration and engagement ($r \approx 0.30$), which becomes weaker for very long videos.

4. **Reduce Emphasis on Posting Day Optimization:** Creators should prioritize a consistent posting schedule instead of focusing heavily on the specific day of publication. Even though small differences between weekdays and weekends exist, these effects are small (less than 0.7 percentage points). So it can be said that consistency and content quality can be more important than exact timing.

## 4.3 Future Research Directions

This study opens several opportunities for future researchers to build on our findings and explore YouTube engagement in more detail. One good direction would be to use experimental methods, for example, A/B testing different thumbnail styles or video titles, to prove causal relationships instead of just correlations between content features and engagement outcomes. Besides that,

researchers could expand their analysis to include different YouTube creators across different content niches to test if our findings can be applied broadly or are just specific to certain channel types. Another valuable approach would be to analyze the text of the comments to understand not just how much people engage, but what they think and feel about the content. Furthermore, future studies could investigate how YouTube's algorithm influences the relationship between video characteristics and viewer engagement, since the algorithm plays a big role in deciding which videos people see. And lastly, tracking individual users over time would help us understand how their engagement patterns might change as they become long-term followers of a channel, which provides insights into audience loyalty and retention.

# 5 Conclusion

The aim of this report was to explore what actually drives engagement on YouTube. With using 234,889 daily observations collected over a five-year period, the analysis identified multiple patterns that question common assumptions and offer useful insights for content creators. It also showed that engagement is shaped by the interaction of content features, audience behavior, and the platform context instead of a single performance metric. This suggests that creators need to think in terms of systems and combinations of factors rather than expecting one simple tactic to lead to success. One central finding is that the like rate plays a more important role in driving engagement than view count alone. In other words, active audience interaction matters more than just reach. This suggests that encouraging simple, low-effort actions, like hitting the like button, can have a surprisingly big impact on overall engagement. This indicates that audiences who are willing to like a video are also more likely to engage in other ways, for example, commenting or sharing. The results also show that high visibility does not automatically lead to high engagement, that the day of publication has only a limited effect, and that engagement levels have remained relatively the same over time. These findings have clear practical implications. Instead of focusing on maximizing views or just optimizing posting schedules, creators could benefit more from producing content

that really resonates with their audience and encourages interaction. Designing clear calls to action, fostering community, and responding to feedback may therefore be more effective than simply chasing higher reach. In addition, interpreting performance data in context helps to avoid decisions based on short-term spikes or outliers. Even small increases in like rates can contribute to noticeable improvements in overall engagement. From an academic perspective, this report contributes a data-driven analysis of YouTube engagement. Furthermore, the applied combination of data cleaning, correlation analysis, and visualization provides a structured approach that can be transferred to other digital platforms, especially when similar engagement metrics are available. This approach also shows how visual analysis and statistical measures can complement each other to strengthen interpretation. While correlation analysis reveals the strength and direction of relationships, visualizations help to detect patterns, clusters or data irregularities that might have remained hidden. To sum it all up, the findings highlight the importance of engagement while also emphasizing the complexity of digital audience behavior. Understanding this complexity is essential for both researchers and practitioners who want to make informed decisions in a rapidly evolving online environment.

# 6 Personal Reflections

**Athina's Reflection:** At the beginning of this project, the overall task appeared quite difficult to me. I had no prior experience with RStudio or programming in R, and the idea of analysing a large dataset with code felt overwhelming. My first reaction was uncertainty, because I asked myself if I would be able to understand the logic behind coding and statistical analysis in such a short time. However, as we progressed with the project and broke the task down into smaller, manageable parts, it all became more logical and understandable. This gradual approach helped me overcome my initial fear of mistakes and allowed me to experiment more with different functions and commands. Step by step, I developed a better understanding of how data analysis is structured in practice.

Teamwork played an important role throughout the project. Working closely with Stella made it easier to share ideas, divide responsibilities efficiently, and support one another when challenges came up. At times, we had different opinions on how to approach some parts of the project, but talking these differences through usually led to better outcomes and solutions. Open communication and collaboration helped us stay organised, even when the workload increased. Knowing that I was not working alone made the project feel less intimidating, especially when deadlines were approaching. In addition, preparing and presenting our intermediate results in a mid-project presentation was an important learning experience for me. Explaining our analysis and findings to others helped me better understand the project myself and identify which aspects needed clearer explanation or further refinement. The presentation also showed us where our approach was still unclear or incomplete, especially in regards to data cleaning and the definition of engagement metrics. Feedback from the presentation encouraged me to improve the structure of our analysis and strengthen the way we communicated our results, which benefited the final report. At the same time, I had some difficulties. One problem involved technical errors in RStudio, especially during rendering and formatting. In some cases, these issues were difficult to resolve, even with the support of tools such as GitHub Copilot in Visual Studio Code. My attempts to fix specific formatting problems sometimes led to unexpected changes in the overall structure or layout of sections, which required additional time and troubleshooting. For example when I installed packages incorrectly, accidentally deleted small parts of codes, that were barely noticeable or when my whole layout looked weird in the rendered version of the project. Despite these challenges, the project provided me with valuable learning experiences. I developed technical skills in data analysis and programming, as well as improved my problem-solving abilities and learned to remain patient. I also gained a better understanding of how collaborative work and feedback contribute to a more solid and transparent analysis. Overall, this project helped me gain a deeper understanding of data-driven analysis and strengthened my

confidence in applying analytical tools in future academic or professional contexts.

**Stella's Reflection:** One of the most important things I realized throughout this project is that learning new tools can be frustrating, but can also yield a large payoff in the long run. At first, I thought writing a report in Word would be easier and less stressful. I did not really understand the need to do something in a way we are not used to. However, as I progressed with the project, I began to see the benefits of using R and Quarto. Once I learned the system, the whole process became simpler and more organized. Now I see this method as a positive option for students who create reports based on data, and as more professional and transparent, though it takes more time to develop. Teamwork with Athina was another positive aspect of this project, as we were able to divide the workload and support one another when we encountered difficulties. Sharing difficulties and exchanging problems and ideas rather than having to solve them alone made the project less stressful and more rewarding in the end. I think teamwork is essential for creating successful projects involving complex subjects and unfamiliar tools. Another important, at first unsuccessful, step for me was presenting our project to the class. Right after the presentation, I was somewhat disappointed because I felt I wasn't as confident as I should be. I questioned whether I performed well overall. But right after we continued to develop the project, my point of view changed. I realized that the presentation gave us a better understanding of the project and helped us identify where we needed improvement. We identified what should work better next time and how we can improve ourselves. This experience created confidence and motivation to become more active in our project. I also really liked the subject of our project because I use YouTube in my daily life, and it was interesting to analyze something I could relate to personally. The most surprising discovery I had regarding YouTube was that simply because a video receives a high number of views does not mean it has the greatest engagement. Before this project, I assumed that popularity automatically equaled success, but now I realize that is not always true. This changed my perspective on social media and online content

in a way that would not have happened without this project. Another important learning experience for me was managing my time effectively. I really underestimated how much time the technical component of the project would take, and I learned that it is more effective to start a project early and keep working on it consistently rather than waiting until the last minute to complete it. This might sound obvious, but I realized that starting projects early is very important, because in the past, I often didn't take it seriously. However, after seeing how much work this project required and that it was not just a normal paper, my way of thinking changed. Projects involving data require thorough planning, especially if you plan to use new tools. In the future, I look forward to organizing my work more effectively and giving myself additional time to address upcoming complications. I believe better planning will reduce stress and help me approach complex projects more structurally. I also learned that reporting results is different from discovering them. There were several moments during the project where I found it difficult to clearly explain my analysis and connect the figures to the written text. I realised that producing good results is not enough if they cannot be communicated clearly and in a structured way. This helped me improve my academic writing and gave me more confidence in presenting and explaining data. In summary, I gained a great deal of knowledge from this project, both academically and personally. We acquired new technical skills, enhanced our problem-solving abilities, and gained greater confidence in working with data. Additionally, I learned that data can reveal patterns that may not be apparent at the surface and can help question assumptions. Although the project was demanding at times and I was stressed, I am glad I had the opportunity to go through this experience, as it has prepared me for future academic challenges and similar projects.

# Division of Work

**Table 8**

*Division of Work*

| Contributor | Sections |
|---|---|
| Athina | 2, 2.2, 2.2.1, 2.2.2, 2.2.3, 2.2.4, 2.2.5, 3.1, 4, 4.1, 4.2, 4.3, 5,6 |
| Stella | 1, 1.2, 1.3, 3, 3.2, 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.3, 6 |

# References

Balakrishnan, J., & Griffiths, M. D. (2017). Social media addiction: What is the role of content in YouTube? *Journal of Behavioral Addictions*, *6*(3), 364–377.

Jaakonmaeki, R., Mueller, O., & Brocke, J. vom. (2017, January). *The impact of content, context, and creator on user engagement in social media marketing*.

K S, S., Thangavelu, P., Gopalakrishnan, P., & Velusamy, J. (2025). The effect of YouTube recommendation engine: A study of visibility and engagement. *Journal of Electrical Systems*, *20*, 2546–2557.

Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, *66*, 236–247.

Ksiazek, T., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media & Society*, *18*(1), 502–520.

Wright, J. (2024). *200K YouTube channel analytics*. Kaggle dataset. https://www.kaggle.com/datasets/jakewright/200k-youtube-channel-analytics

Wu, S., Rizoiu, M.-A., & Xie, L. (2018, June). *Beyond views: Measuring and predicting engagement in online videos*.

# Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

Cologne, 01.02.2026

Athina Agiakatsikas

Stella Christou