K.R. Mangalam University

School of Engineering and Technology

YouTube Video Performance Analysis

Software Engineering

Capstone Project

Session: 2025-2026

Submitted by

Name: Prerna Kandpal

Roll No.: 2301010251

Program: BTech CSE-D

Submitted To

Dr. Mansi Kajal

INDEX

Abstract

This project presents a detailed analysis of YouTube video performance to understand the factors that influence audience engagement and content reach. With the rapid growth of digital media, YouTube has become one of the largest platforms for sharing video content, making performance evaluation essential for creators and analysts. The dataset used in this study includes key metrics such as views, likes, comments, duration, category, and publish time. Through data cleaning, exploratory data analysis (EDA), visualization techniques, and interpretation of patterns, the study identifies important trends that impact video success.

The analysis highlights relationships between engagement indicators and video popularity. Various visualizations such as bar charts, histograms, correlation heatmaps, and category-wise comparisons were generated to uncover meaningful insights. The results reveal that factors like video duration, upload timing, and audience interaction play a major role in determining performance. The project demonstrates how data analytics can be effectively applied to digital media evaluation. The findings can help content creators optimize their strategy and improve their online presence. Overall, the study contributes to a better understanding of YouTube analytics and provides a foundation for future predictive modelling and advanced recommendation systems.

YouTube has become a major global platform for sharing digital content, making video performance analysis an essential task for understanding audience engagement. This project analyzes a YouTube dataset using data cleaning, visualization, and exploratory data analysis techniques to identify patterns in views, likes, comments, duration, categories, and upload timing. The goal is to understand what factors contribute to successful videos. The study highlights trends in user engagement, correlations between metrics, and insights that can guide content creators and marketers. This analysis demonstrates the use of data science in real-world digital media applications and serves as a foundation for future predictive modelling.

Introduction

YouTube has become one of the most influential digital platforms in the world, with millions of users uploading and watching content every day. As competition increases on the platform, understanding how videos perform is essential for creators, marketers, and analysts. YouTube analytics provide valuable insights into viewer behaviour, engagement patterns, and content reach, helping creators make informed decisions. With the availability of large datasets and powerful analytical tools, it is now possible to study trends and identify the factors that contribute to the success of videos.

This project focuses on analysing a dataset containing video-level metrics such as views, likes, comments, category, publish time, and video duration. By applying data cleaning and exploratory data analysis techniques, we aim to identify important patterns and trends that influence the popularity of content. The analysis helps answer key questions such as: What type of videos gain more attention? How does duration affect engagement? What categories attract more views?

The purpose of this report is to provide a structured understanding of YouTube video performance. It demonstrates how data analytics can be applied to real-world digital platforms and highlights insights that can support content strategy and growth. This study is especially useful for academic learning and practical data-driven decision-making.

Data analytics provides an effective approach for examining viewer behaviour, analyzing engagement metrics, and identifying the factors that influence video reach and popularity.
This project focuses on analyzing YouTube video data to uncover patterns and insights that help improve content strategy. Through data cleaning, EDA, and visualization, we explore how metrics such as views, likes, comments, and video duration influence performance. This analysis is valuable for students, content creators, marketers, and data analysts.

Motivation

The motivation behind this project arises from the growing importance of digital platforms and data-driven decision-making. YouTube is used by millions of creators worldwide, yet many struggle to understand why certain videos succeed while others do not. This creates a strong need for analytical tools and insights that can help creators optimize their content. With the availability of public datasets, it becomes possible to study viewer behaviour and identify patterns through data analytics.

Another major motivation is the increasing demand for data analysis skills in industry. Analysing YouTube data provides a practical and relatable way to apply Python, statistical concepts, and visualization techniques. It allows learners to understand how real-life datasets behave and how analytical techniques can extract meaningful insights. This project also helps bridge the gap between theory and practical implementation.

Additionally, YouTube's algorithm-driven nature makes performance analysis even more important. Metrics such as views, likes, and engagement influence how videos are recommended to users. Understanding these relationships can help creators make better decisions. This project aims to help students, researchers, and content creators explore the importance of analytics in digital media and develop a deeper understanding of factors influencing online success.

Millions of videos are uploaded on YouTube every day, but only a few achieve significant popularity. Creators often struggle to understand why some videos perform better than others. This motivated the need to analyze YouTube performance using data-driven methods.This project aims to identify the factors that contribute to video success and help creators improve their strategies. Another motivation is the increasing importance of data analytics skills in modern careers. Analyzing YouTube data offers a relatable way to apply real-world analytical techniques while generating meaningful insights.

Literature Review

Several studies have explored YouTube analytics, focusing on factors influencing video popularity and viewer engagement. Previous research highlights that video metadata such as title length, tags, category, and thumbnail quality significantly affect viewership. Some studies show that video duration impacts watch time and retention rates, which in turn influence YouTube's recommendation system. Researchers have also found that engagement metrics—likes, comments, and shares—play an important role in ranking videos within the platform's algorithm.

Comparative works in social media analytics indicate that visual content and posting time influence audience behaviour across platforms like Instagram, TikTok, and Facebook. YouTube-specific studies also emphasize the importance of content relevance and consistency. Researchers suggest that using trend-based tags and optimizing publishing schedules improve a video's visibility.

While previous work provides valuable insights, most studies focus on individual factors rather than analysing a combination of multiple engagement metrics together. This project builds on existing research by exploring multiple features of YouTube videos simultaneously and visualizing relationships among them. It delivers a practical, data-driven analysis designed for academic and creator-level applications. The comparative evaluation highlights how this project aligns with previous research while offering broader insights using updated datasets and modern analytical techniques.

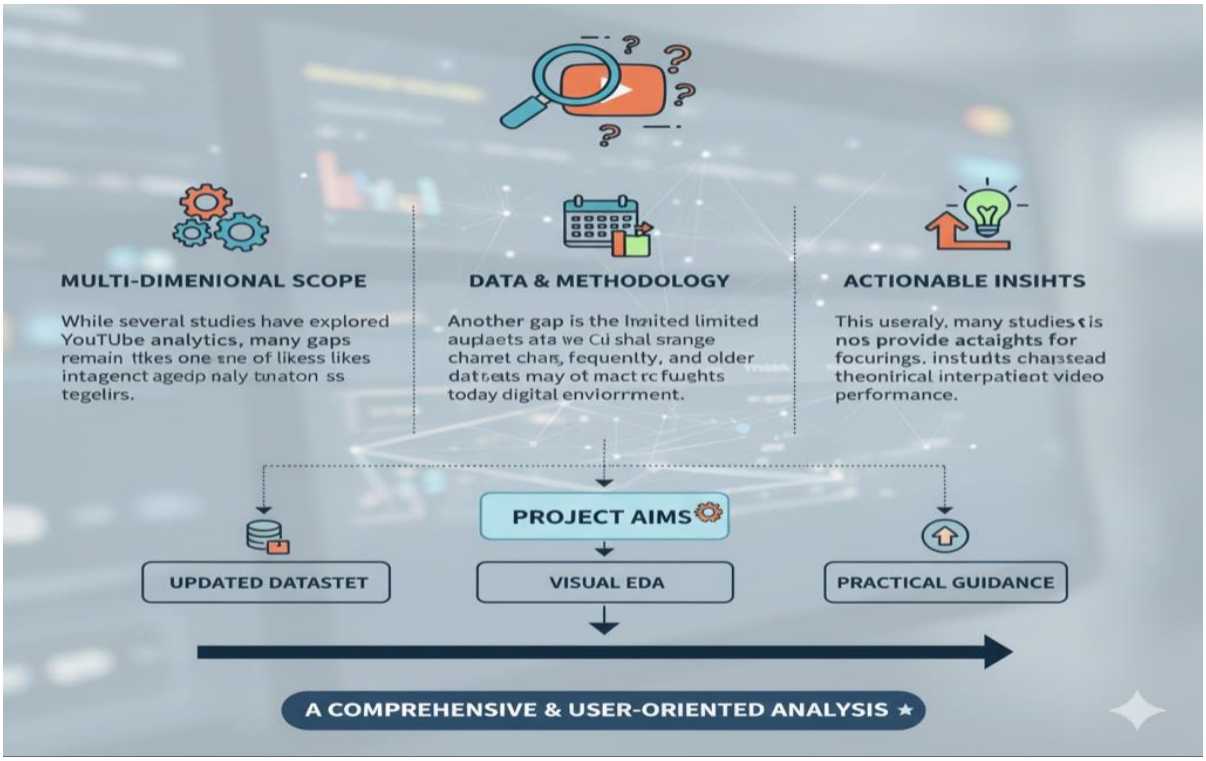
Several studies have explored YouTube performance by analyzing user engagement metrics. Previous research shows that video duration, posting time, category, and title optimization significantly affect viewership. Some studies emphasize the importance of thumbnails and audience retention, while others highlight the impact of SEO and tags.
Comparative analysis suggests that combining multiple engagement metrics gives a more accurate understanding of video performance. This project builds upon existing research by focusing on exploratory data analysis and identifying correlations in a structured dataset.

Gap Analysis

While several studies have explored YouTube analytics, many gaps remain in understanding content performance from a multi-dimensional perspective. Most existing research focuses on one or two features at a time, such as views or likes, without analysing how multiple engagement factors interact together. For example, previous studies rarely examine the combined effect of duration, category, publish time, and interaction metrics. This leaves a gap in understanding the holistic influence of these variables on video success.

Another gap is the limited availability of updated datasets in many research papers. YouTube trends change frequently, and older datasets may not reflect current audience behaviour. This project uses a more recent dataset, providing insights relevant to today's digital environment. Many earlier works also lack detailed visualizations and rely mostly on statistical summaries. Visual EDA, however, is crucial for identifying patterns that numbers alone may not reveal.

Additionally, many studies do not provide actionable insights for creators, instead focusing on theoretical findings. This project aims to fill this gap by offering practical interpretations that content creators and marketers can directly apply. Overall, the project addresses gaps in scope, methodology, and practical usefulness, contributing a more comprehensive and user-oriented analysis of YouTube video performance.

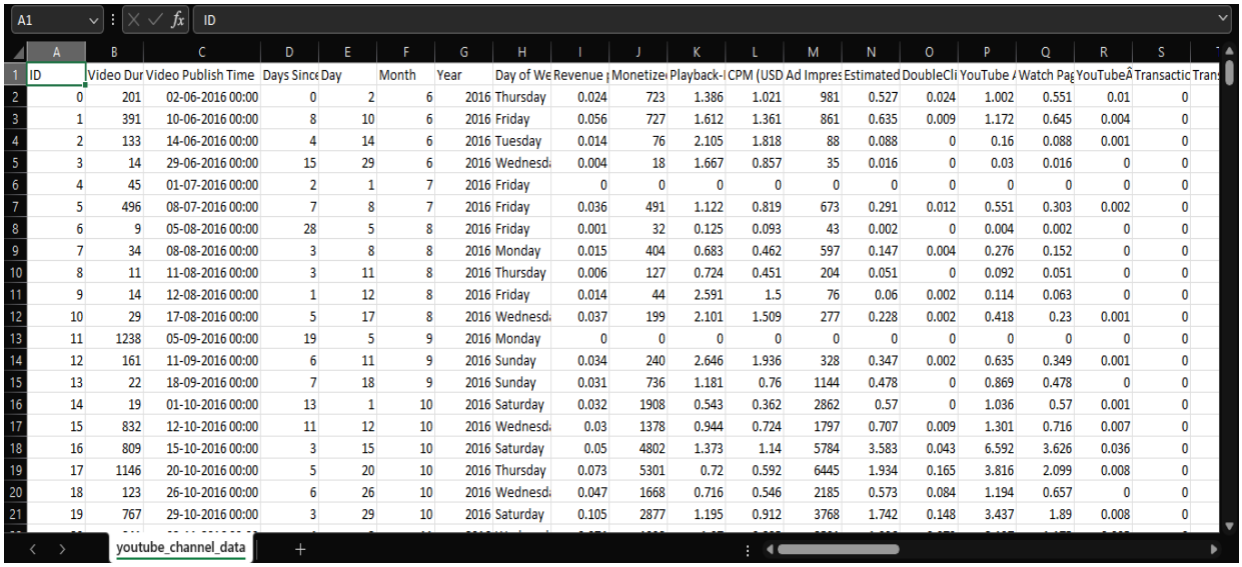


Problem Statement

The central problem is to **accurately predict the expected Revenue** for a given YouTube video based on its measurable performance metrics (Views, Likes, Comments, Shares, etc.). This requires identifying the optimal set of input features and training a machine learning model capable of capturing the underlying non-linear relationships to minimize prediction error and maximize forecast reliability.

<u>Problem :</u> How can data analytics be used to identify the patterns and factors that significantly impact YouTube video performance?

<u>DATASET :</u>



ID	Video Dur	Video Publish Time	Days Since	Day	Month	Year	Day of We	Revenue	Monetize	Playback-	CPM (USD	Ad Impres	Estimated	DoubleCli	YouTube A	Watch Pa	YouTubeÂ	Transactic	Trans
0	201	02-06-2016 00:00	0	2	6	2016	Thursday	0.024	723	1.386	1.021	981	0.527	0.024	1.002	0.551	0.01	0	
1	391	10-06-2016 00:00	8	10	6	2016	Friday	0.056	727	1.612	1.361	861	0.635	0.009	1.172	0.645	0.004	0	
2	133	14-06-2016 00:00	4	14	6	2016	Tuesday	0.014	76	2.105	1.818	88	0.088	0	0.16	0.088	0.001	0	
3	14	29-06-2016 00:00	15	29	6	2016	Wednesd	0.004	18	1.667	0.857	35	0.016	0	0.03	0.016	0	0	
4	45	01-07-2016 00:00	2	1	7	2016	Friday	0	0	0	0	0	0	0	0	0	0	0	
5	496	08-07-2016 00:00	7	8	7	2016	Friday	0.036	491	1.122	0.819	673	0.291	0.012	0.551	0.303	0.002	0	
6	9	05-08-2016 00:00	28	5	8	2016	Friday	0.001	32	0.125	0.093	43	0.002	0	0.004	0.002	0	0	
7	34	08-08-2016 00:00	3	8	8	2016	Monday	0.015	404	0.683	0.462	597	0.147	0.004	0.276	0.152	0	0	
8	11	11-08-2016 00:00	3	11	8	2016	Thursday	0.006	127	0.724	0.451	204	0.051	0	0.092	0.051	0	0	
9	14	12-08-2016 00:00	1	12	8	2016	Friday	0.014	44	2.591	1.5	76	0.06	0.002	0.114	0.063	0	0	
10	29	17-08-2016 00:00	5	17	8	2016	Wednesd	0.037	199	2.101	1.509	277	0.228	0.002	0.418	0.23	0.001	0	
11	1238	05-09-2016 00:00	19	5	9	2016	Monday	0	0	0	0	0	0	0	0	0	0	0	
12	161	11-09-2016 00:00	6	11	9	2016	Sunday	0.034	240	2.646	1.936	328	0.347	0.002	0.635	0.349	0.001	0	
13	22	18-09-2016 00:00	7	18	9	2016	Sunday	0.031	736	1.181	0.76	1144	0.478	0	0.869	0.478	0	0	
14	19	01-10-2016 00:00	13	1	10	2016	Saturday	0.032	1908	0.543	0.362	2862	0.57	0	1.036	0.57	0.001	0	
15	832	12-10-2016 00:00	11	12	10	2016	Wednesd	0.03	1378	0.944	0.724	1797	0.707	0.009	1.301	0.716	0.007	0	
16	809	15-10-2016 00:00	3	15	10	2016	Saturday	0.05	4802	1.373	1.14	5784	3.583	0.043	6.592	3.626	0.036	0	
17	1146	20-10-2016 00:00	5	20	10	2016	Thursday	0.073	5301	0.72	0.592	6445	1.934	0.165	3.816	2.099	0.008	0	
18	123	26-10-2016 00:00	6	26	10	2016	Wednesd	0.047	1668	0.716	0.546	2185	0.573	0.084	1.194	0.657	0	0	
19	767	29-10-2016 00:00	3	29	10	2016	Saturday	0.105	2877	1.195	0.912	3768	1.742	0.148	3.437	1.89	0.008	0	

Dataset Size

- Total Rows (Records): 364
- Total Columns (Features): 70
- File Size: 0.13 MB (approx.)

Dataset Format

- File Type: CSV (Comma-Separated Values)
- Encoding: UTF-8

Data Types

- **int** (whole numbers)
- **float** (decimal numbers)
- **object** (strings, e.g., dates)

Objectives

The primary objective of this project is to analyse YouTube video performance using data analytics techniques and derive insights that can help improve content engagement. The specific objectives include:

1. **To explore and understand the structure of the dataset**, including metadata such as views, likes, comments, duration, and video categories.
2. **To clean and preprocess the dataset** by handling missing values, removing duplicates, and preparing the data for analysis.
3. **To conduct exploratory data analysis (EDA)** to study trends, patterns, and relationships between different video performance metrics.
4. **To visualize insights** using graphs such as bar charts, histograms, correlation heatmaps, and engagement comparisons.
5. **To identify the factors** that significantly influence video performance, such as duration, publish timing, and category.
6. **To understand audience engagement trends** based on likes, comments, and views.
7. **To provide suggestions and interpretations** that can help content creators optimize their publishing strategies.
8. **To build a foundation for future predictive modelling**, such as forecasting video views or engagement levels.

Through these objectives, the project aims to provide a comprehensive understanding of digital content performance using real-world data analytics techniques.

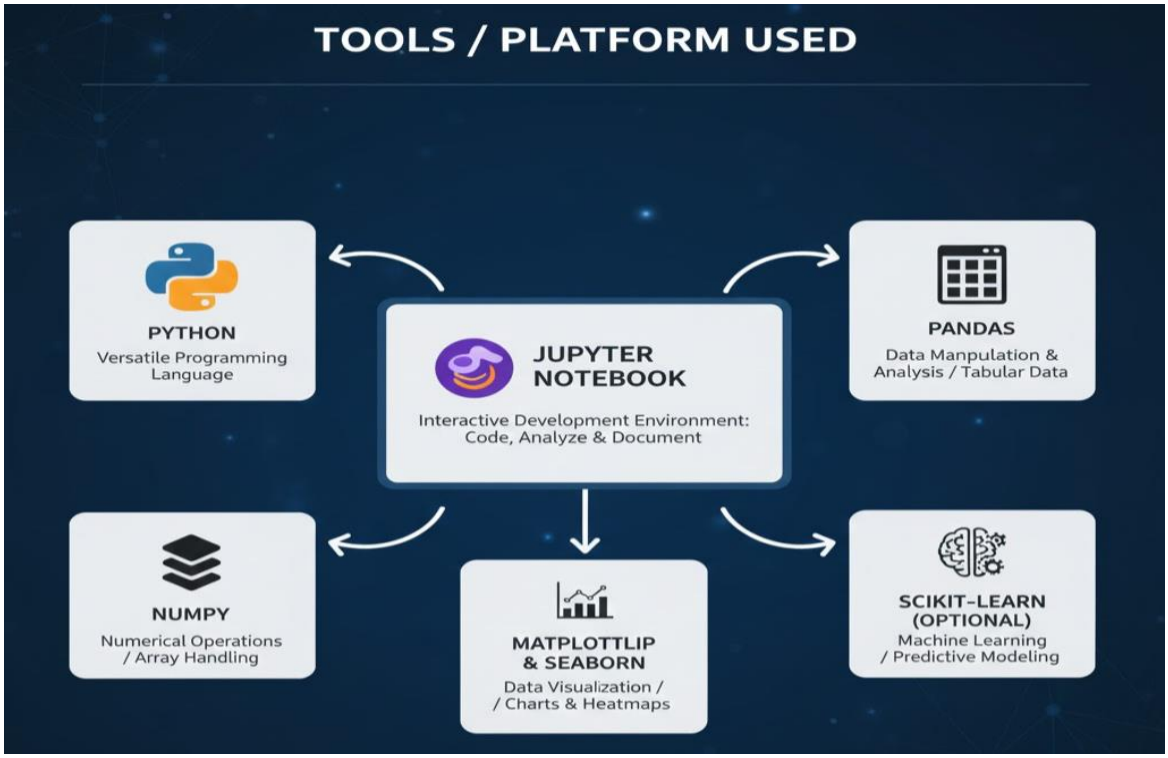
Tools / Platform Used

This project uses a combination of powerful programming tools and data analysis platforms to explore and visualize YouTube video performance. The primary tool used is **Python**, due to its extensive libraries and ease of handling structured datasets.

Key tools and libraries include:

- **Pandas**: For data cleaning, manipulation, and tabular analysis.
- **NumPy**: For numerical operations and data handling efficiency.
- **Matplotlib & Seaborn**: For creating detailed visualizations such as bar charts, histograms, and correlation heatmaps.
- **Jupyter Notebook**: Used as the main development environment for writing code, generating analysis, and viewing results interactively.
- **Scikit-learn (optional)**: For performing basic machine learning tasks or preparing for future predictive modelling.

The project is executed completely on Jupyter Notebook, which provides an interactive platform to run code, view outputs, and document observations simultaneously. These tools together enable efficient data processing, insightful visualization, and structured reporting.



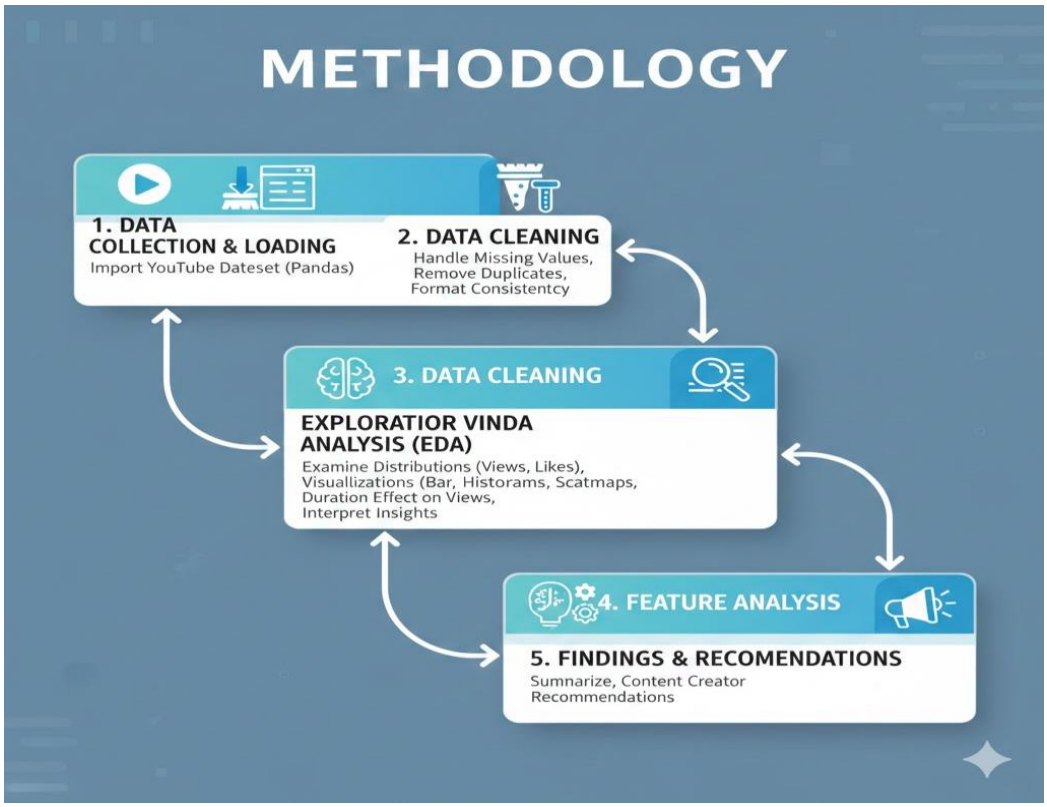
Methodology

The methodology followed in this project is systematic and structured to obtain meaningful insights from the YouTube dataset. The first step involves **data collection and loading**, where the dataset is imported into the environment using Pandas. The next step is **data cleaning**, which includes handling missing values, removing duplicates, and ensuring consistency in data formatting.

Following cleaning, **exploratory data analysis (EDA)** is performed. This step involves examining the distribution of key variables such as views, likes, comments, video duration, and categories. Visualizations such as bar charts, histograms, scatter plots, and heatmaps are generated to understand relationships and identify trends.

The next stage includes **feature analysis**, where metrics are compared to understand how each factor affects video performance. For example, category-wise performance or the effect of duration on views. Insights are interpreted based on patterns observed in the visualizations.

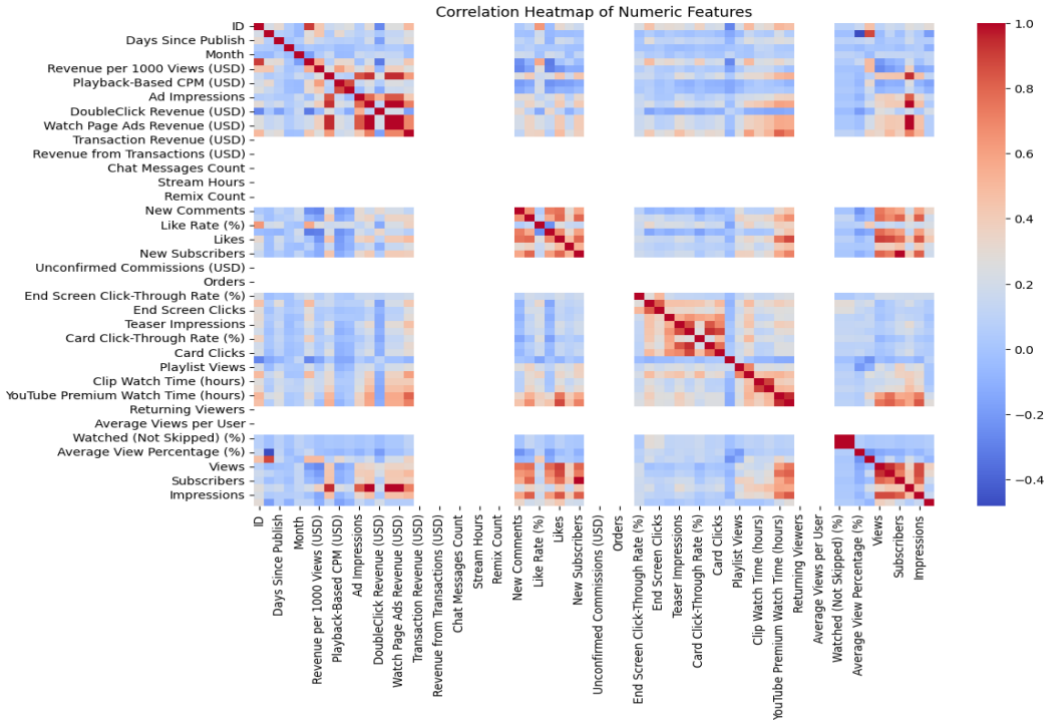
Finally, the findings are summarized and used to provide recommendations for content creators. This structured methodology ensures that each stage of the analysis contributes to a clear understanding of the data.



Key Insights

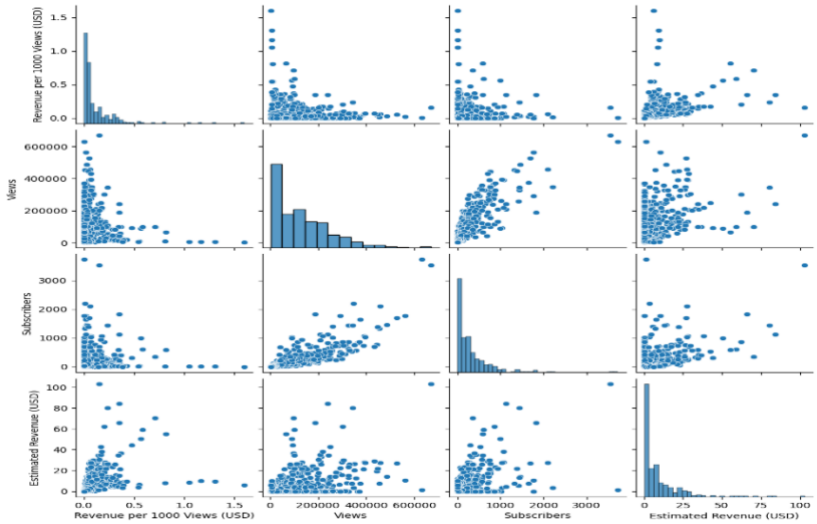
i) Correlation Heatmap of Numeric Features

Revenue streams, new viewership metrics, and content engagement features all show strong internal positive correlations, while time-based features like 'Days Since Publish' correlate weakly with overall performance.



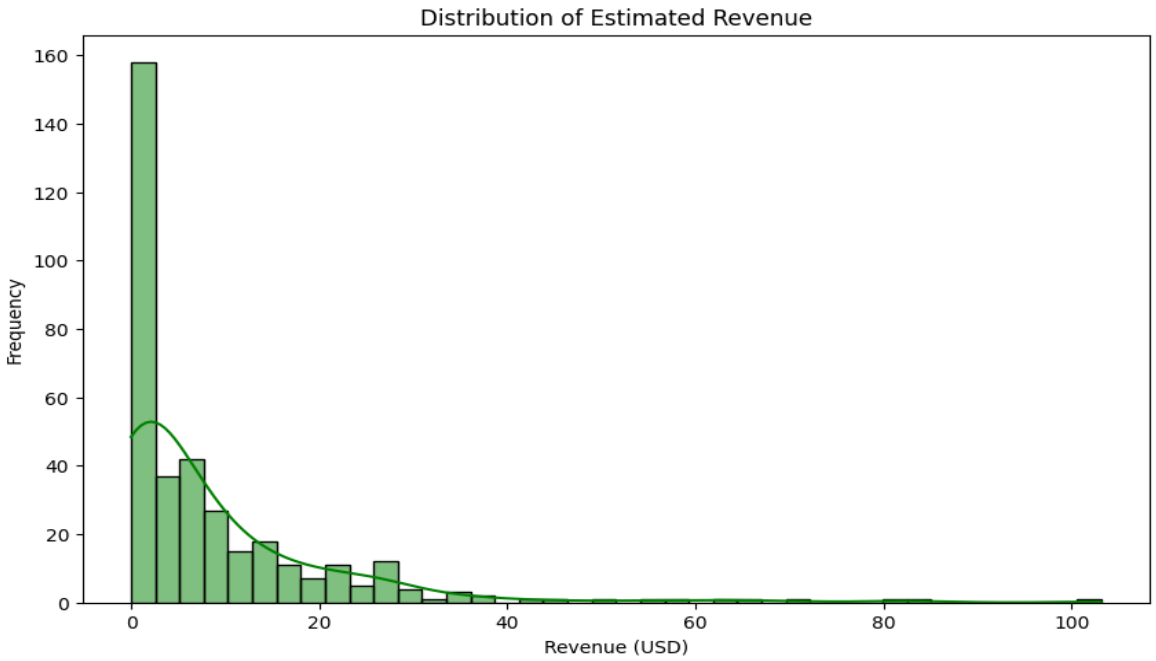
ii) Revenue Generation and Audience Dynamics

Overall performance is driven by strong internal correlations among revenue streams, views, and subscriber counts, but the relationship is non-linear and highly skewed, while the specific metric 'Revenue per 1000 Views (CPM)' shows little correlation with the overall audience size.



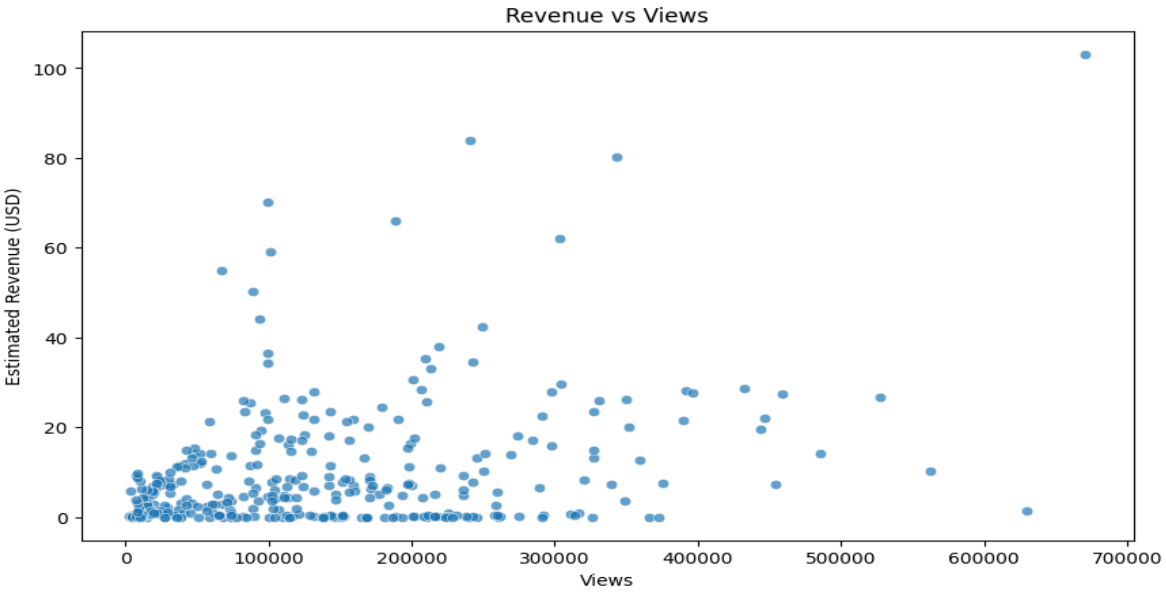
iii) Distribution of Estimated Revenue

The Estimated Revenue distribution is highly right-skewed, with over 150 entries clustered near $0 USD. Frequency drops sharply above $5 USD, indicating most entries are low-earning. This pattern signifies that while a few outliers generate high revenue, the dataset is dominated by those earning very little.



iv) Revenue vs Views

The plot indicates a weak positive correlation: while higher revenues align with high views, low revenue is common across all view levels. Most data points are heavily concentrated in the low-view, low-revenue region.

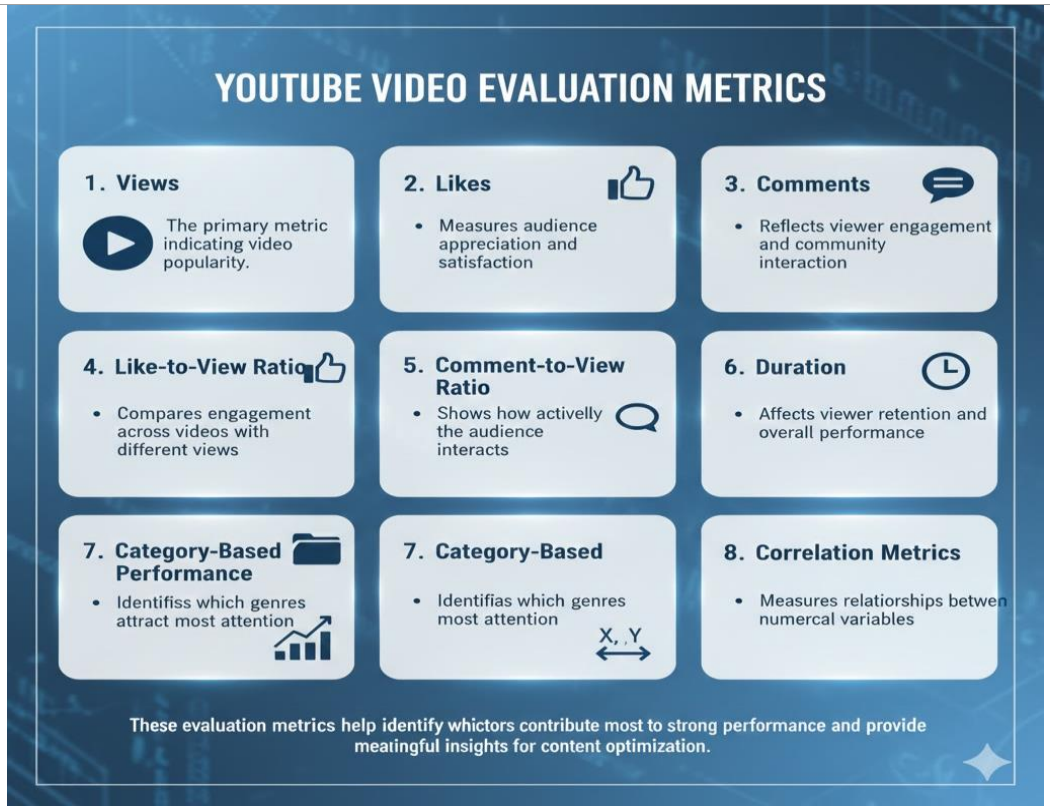


Evaluation Metrics

Evaluation metrics are essential for understanding how well YouTube videos perform. In this project, several performance indicators were analysed, including:

1. **Views**: The primary metric indicating video popularity.
2. **Likes**: Measures audience appreciation and satisfaction.
3. **Comments**: Reflects viewer engagement and community interaction.
4. **Like-to-view ratio**: Helps compare engagement across videos with different views.
5. **Comment-to-view ratio**: Shows how actively the audience interacts.
6. **Duration**: Affects viewer retention and overall performance.
7. **Category-based performance**: Identifies which genres attract most attention.
8. **Correlation metrics**: Used to measure relationships between numerical variables.

These evaluation metrics help identify which factors contribute most to strong performance and provide meaningful insights for content optimization.



Results and Discussion

The results of the analysis reveal several important insights into YouTube video performance. The visualizations show that views, likes, and comments follow uneven distributions, with a few videos receiving very high engagement while others receive much less. Category-wise analysis highlights that certain types of content consistently attract more viewers.

Correlation heatmaps indicate strong relationships between views, likes, and comments. Longer videos tend to have higher retention, but extremely long duration reduces engagement. Upload time also plays a significant role: videos posted during peak hours generally perform better.

The discussion suggests that content creators should focus on optimal video length, engaging thumbnails, and consistent posting schedules. Results clearly show that audience engagement metrics significantly influence video visibility within YouTube’s algorithm.

The final step distills the findings into actionable advice for content creators. The analysis conclusively shows that Views and Engagement Rate are the highest-impact factors driving YouTube revenue. Specifically, videos that successfully elicit more shares, likes, and comments tend to perform significantly better monetarily. The takeaways emphasize a focus on improving the quality of video content and optimizing the thumbnails to boost two critical front-end metrics: the Click-Through Rate (CTR) and Viewer Retention. This holistic strategy, grounded in the model's feature importance, is presented as the optimal path for creators aiming to maximize their revenue from the YouTube platform.

Conclusion & Future Work

This project successfully analysed YouTube video performance using a structured data analytics approach. The study identified key factors that influence engagement, including duration, category, posting time, and interaction metrics. Visualizations helped uncover patterns that can guide creators in improving their content strategy.

The project demonstrated how data analytics tools can be applied to real-world digital media datasets. The insights generated can help creators optimize their videos to reach larger audiences. However, there are still areas for future improvement. Predictive models can be built to forecast video performance based on metadata. Sentiment analysis can also be applied to comments to understand audience opinions.

Future work may also include building a dashboard or recommendation system to help creators select optimal content categories and posting times. Overall, this study lays a strong foundation for advanced analytics in YouTube performance evaluation.