**YouTube Data Analysis using API Key**

Kaviyarasu S

*Coimbatore Institute of Technology*

Coimbatore*, India*

Abstract— This project focuses on analyzing YouTube data using data science techniques to gain valuable insights into user behavior, content trends, and performance metrics. The objective is to identify and analyze the factors that impact likes, subscribers, and views on YouTube, providing content creators with data-driven solutions to enhance their channel's popularity and expand their viewership. The YouTube dataset, accessed through the YouTube API key, allows for the exploration of the platform's dynamic nature, capturing the latest trends and patterns. Through the analysis, content creators can understand content preferences, engagement patterns, influential factors, audience demographics, channel growth strategies, seasonality, engagement with subscribers, and competitor analysis. The advantages of this analysis include insights into user behavior, optimized content strategy, identification of popular trends, data-driven decision making, enhanced audience engagement, competitive insights, real-time analysis, and evidence-based collaboration. By leveraging these insights, content creators can make informed decisions, improve their content strategies, and foster audience engagement on YouTube.

1. Introduction

Many YouTube content creators struggle with the challenge of limited likes, subscribers, and views on their channels. Understanding the factors that influence video performance and user engagement is crucial for channel growth and attracting a larger audience. However, without insights into these key factors, content creators find it difficult to develop effective strategies to improve their channel performances. The problem at hand is to identify and analyse the factors that impact likes, subscribers, and views on YouTube, and devise data-driven solutions to help content creators enhance their channel's popularity and expand their viewership.

1. Dataset Description

YouTube, the world's leading online video sharing platform since its establishment in 2005, attracts millions of users daily with its diverse collection of videos spanning entertainment, education, music, gaming, and more. For this project, we selected the YouTube dataset to address the challenges faced by one of our team members who owns a YouTube channel and seeks to enhance its performance by gaining more likes, subscribers, and views. By analyzing this dynamic dataset, we aim to uncover valuable insights and patterns that can assist not only our team member but also other content creators in enhancing audience engagement and expanding their reach. Leveraging the YouTube API and an API key, we accessed up-to-date information on videos, channels, comments, and relevant metadata. This direct integration with the official YouTube platform ensures the authenticity and reliability of our dataset. By utilizing the YouTube API key, we have the opportunity to explore this vast world of YouTube data, providing valuable insights for content creators, marketers, and platform administrators to make data-driven decisions, optimize strategies, and improve user experiences.

III. Data Preprocessing

Data preprocessing is a vital step in preparing the YouTube dataset for analysis. It involves cleaning, integrating, transforming, selecting relevant features, detecting outliers, and normalizing the data. These steps ensure data accuracy, compatibility, efficiency, and integrity, enabling content creators to gain valuable insights and make informed decisions to improve their channel's performance.

IV. Descriptive Statistics

Descriptive statistics provide key insights into the dataset, including measures of central tendency (mean, median, mode), variability (standard deviation, range), distribution shape (skewness, kurtosis), percentiles, and outliers. These statistics summarize the data, giving content creators a clear understanding of its typical values, spread, distribution shape, and exceptional data points. By analysing descriptive statistics, content creators can make informed decisions and draw meaningful conclusions from the dataset.

V. Visualization Techniques

Visualization techniques were employed to provide a visual representation of the dataset. Bar chart, scatter plot, histogram, bubble plot, distribution of dataset, year vise plotting, day vice plotting.

VI. Correlation Analysis

Correlation analysis was performed to examine the relationships between different variables in the dataset. The relation between likes counts and views count, duration and views are analyzed.

VII. Performance Analysis

Understanding the Factors for YouTube Channel Success. We analyzed the impact of video durations, content preferences, and viewer engagement on likes and views. By optimizing video length, creating content aligned with viewer interests, and considering influential factors, content creators can attract more likes, subscribers, and views. We also looked at seasonal trends, studied successful channels, and provided real-time insights to enhance engagement and maximize channel performance.

VIII. Future content prediction

Develop a predictive model to forecast the performance of upcoming YouTube content. By analysing factors such as video duration, upload day, and historical data trends, this model can provide content creators with valuable insights into the expected likes, views, and engagement for future videos. With this information, creators can make informed decisions about content creation and optimize their strategies to maximize reach and audience engagement. This predictive capability empowers content creators to plan their content pipeline effectively and tailor their approach to attract a larger audience and drive channel growth..

VIII. Data Preparation

Before building predictive models, it is important to get the data ready for analysis. This involves cleaning the data by removing any irrelevant or duplicate information. We also handle missing values by filling in the gaps with appropriate values. Additionally, we engineer new features or transform existing ones to make them more useful for prediction. Normalization techniques are applied to ensure that all the data is on a consistent scale. By performing these steps, we ensure that the data is in a suitable format for training the models and obtaining accurate insights about future content performance.

VIII. ML Algorithms

For our analysis, we employed two machine learning algorithms: linear regression and random forest. Linear regression establishes a linear relationship between input features (duration and publish day) and target variables (views and likes). Random forest combines multiple decision trees to predict views and likes, considering non-linear relationships and interactions. By comparing their performance, we gained insights into the factors influencing video popularity, empowering content creators to enhance their channel's performance.

VIII. Model Training

we use duration and publish day as the input data (features) and views and likes as the target variables. The goal is to train models that can accurately predict the number of views and likes based on the duration of the video and the day it is published. We employ machine learning algorithms such as regression or classification models, depending on the nature of the target variables. The dataset is split into training and testing sets, with the training set used to train the models and the testing set used to evaluate their performance. The models learn the patterns and relationships between the input data and target variables during the training process. By training these models, we can gain insights into the factors that influence the number of views and likes on YouTube and use them to optimize content strategy and improve channel performance.

VIII. Model Evaluation

To assess the performance of our trained models, we employed various evaluation metrics. These metrics provide insights into how well the models can predict the number of views and likes based on the input features. Common evaluation metrics used include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R2) score.

IX. Comparison with channels

In our project, we compared several YouTube channels to see how well they were doing and find out what made them successful. We looked at important numbers like views, likes, and subscribers to understand which channels were doing better than others in the same category. This helped us identify strategies that worked and contributed to their success.

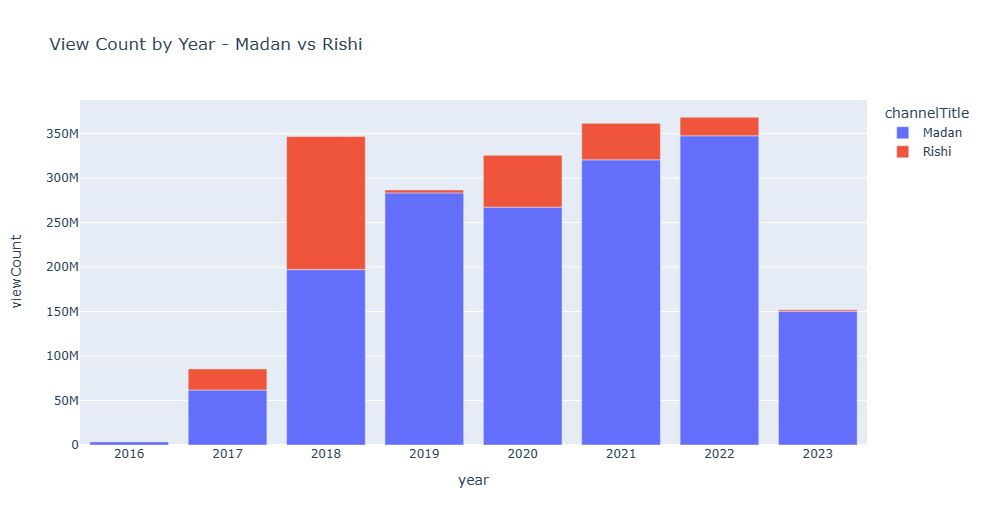
X. Dashboard

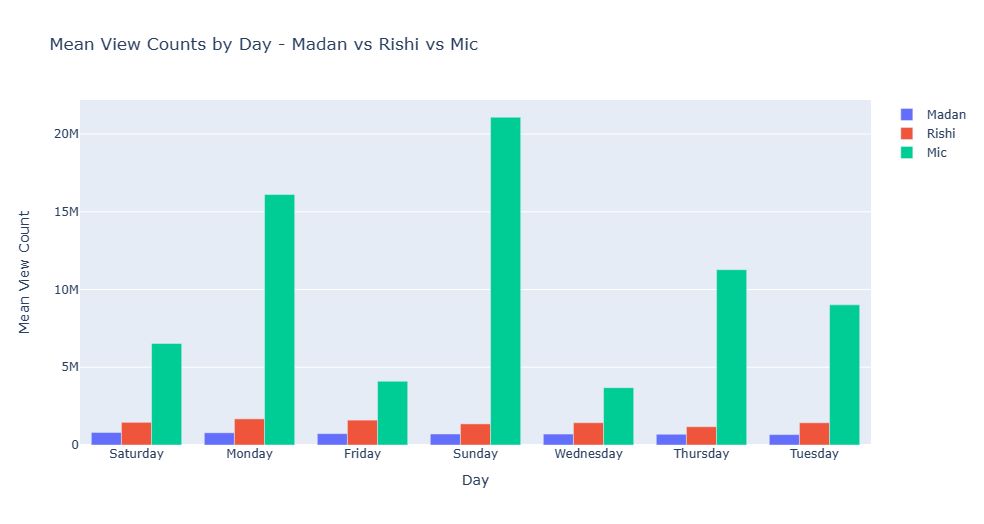
In addition to our analysis and modelling, we also developed an interactive dashboard using Excel to visualize and explore the YouTube dataset. With that dashboard content creators can easily navigate through various metrics, charts, and visualizations.

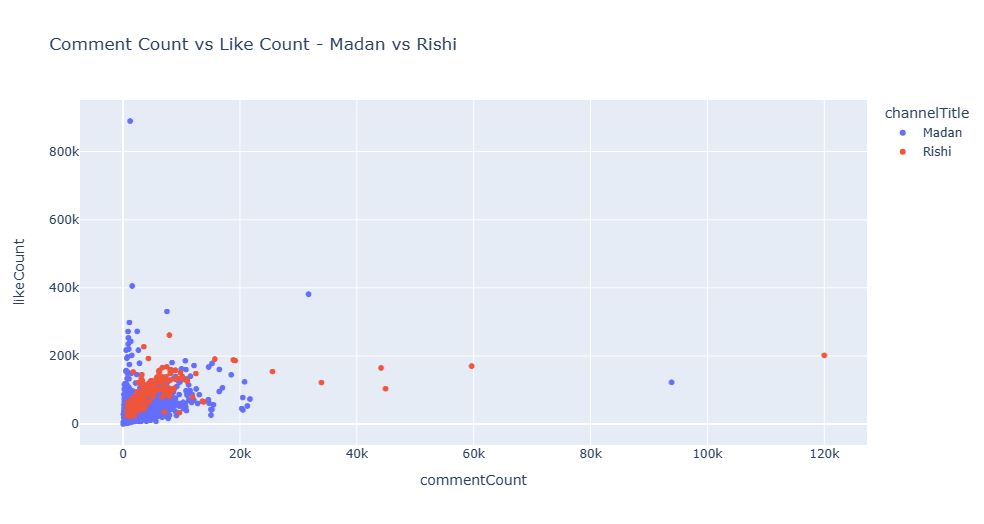
XI. Conclusion

In conclusion, our YouTube data analysis offers valuable insights and practical strategies for content creators to improve their channels. By understanding how viewers behave, what content they like, and how to engage them, creators can optimize their strategies, attract more subscribers, and succeed on YouTube. Applying these insights and making data-driven decisions can help creators build a loyal audience, increase their reach, and achieve long-term success on the platform. With the right strategies and a deep understanding of their audience, content creators can thrive on YouTube's dynamic platform.

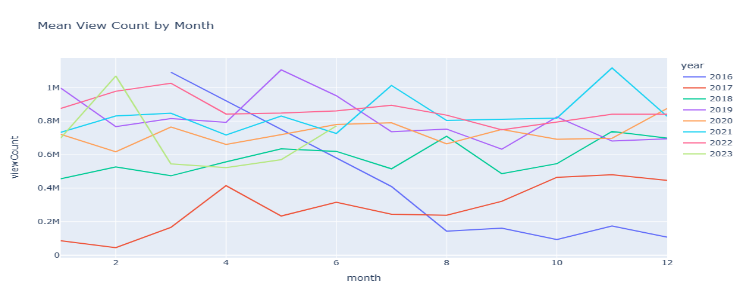
XII. Sample Visualization

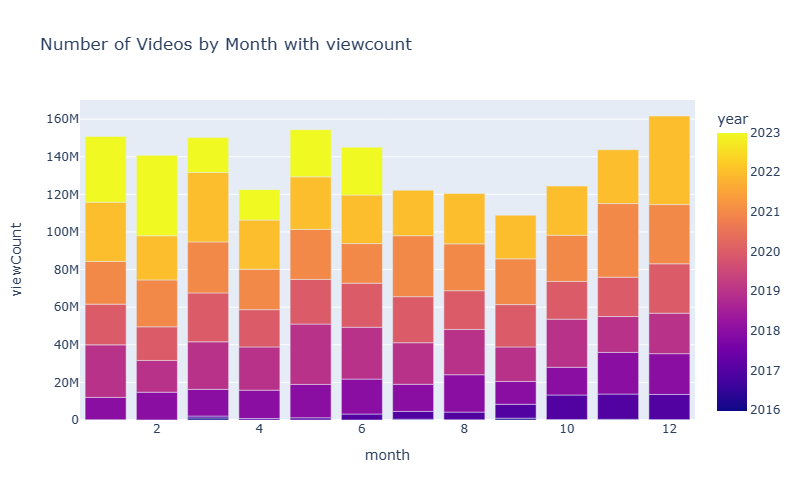


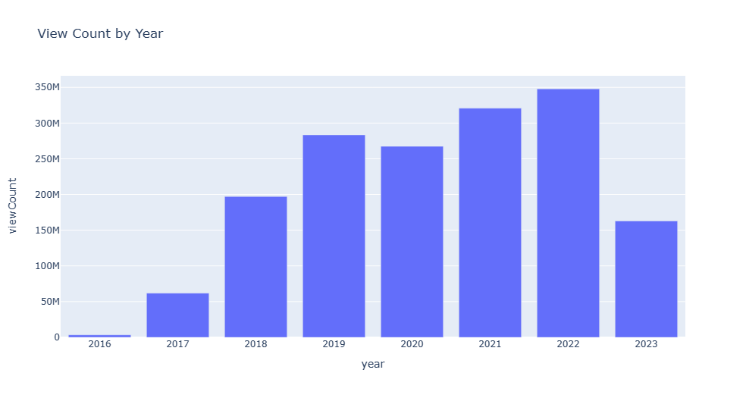


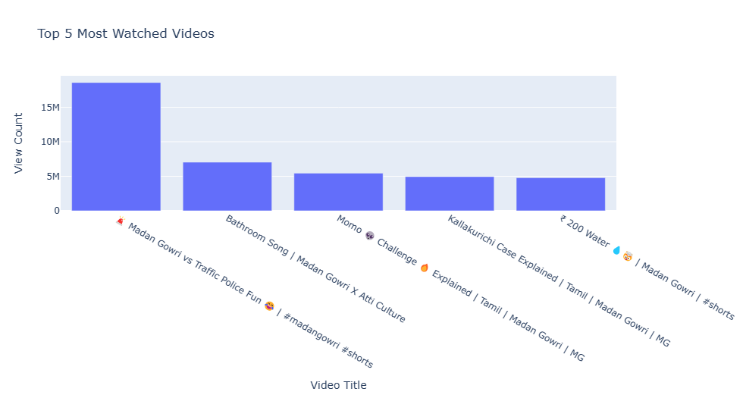


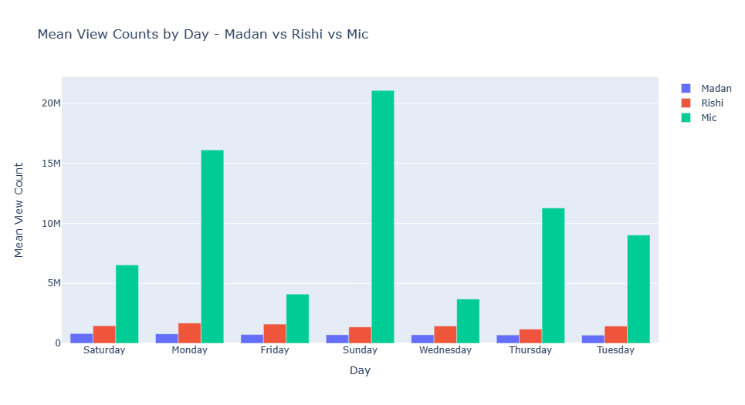


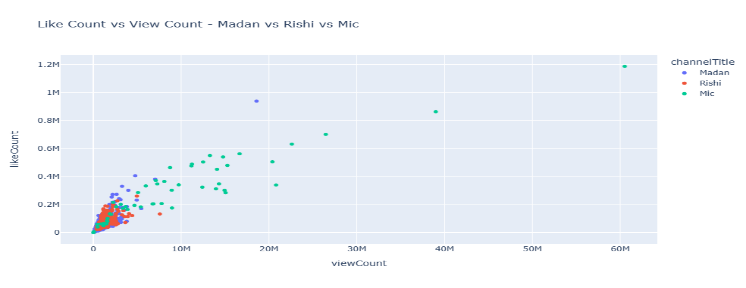












XIII.Dashboard Creation:

