**Global YouTube Statistics**

**1. PROJECT BY**

Athashri Ajay Gangan

**2. Introduction**

The rapid growth of YouTube as a leading video-sharing platform has transformed how content is consumed, created, and monetized. With millions of videos uploaded daily, understanding the factors that drive video performance, viewer engagement, and channel growth is essential for content creators, marketers, and analysts. This project aims to conduct an advanced statistical analysis of YouTube data, exploring various factors such as views, likes, comments, and subscriber growth.

**3. Objectives**

The primary objectives of this project are:

* Data Exploration: To comprehensively explore the YouTube dataset, identifying key features such as video titles, categories, views, likes, dislikes, comments, and subscriber counts. Understanding these features will help in recognizing patterns and correlations.
* Data Preprocessing: To clean and preprocess the data by addressing missing values, outliers, and inconsistencies. This step ensures that the data is reliable and ready for analysis.
* Feature Analysis and Selection: To identify significant features that impact video performance and viewer engagement. This involves statistical analysis and feature selection techniques to highlight the most influential factors.
* Predictive Modeling: To develop predictive models that can classify videos based on their popularity or predict future engagement metrics such as likes, views, or comments.
* Data Visualization: To create visual representations of the data, making insights accessible and understandable for a broader audience.
* Reporting and Recommendations: To compile findings into a comprehensive report, providing actionable recommendations for content creators to enhance their channel performance.

**4. Scope of Work**

The project will involve the following tasks:

* Data Exploration: Understanding the dataset’s structure, identifying types of data (categorical, numerical), and exploring the relationships between features such as video category, upload time, and engagement metrics.
* Data Preprocessing: Handling missing values through imputation or removal, addressing outliers, and standardizing numerical features to ensure accurate analysis.
* Feature Selection: Analyzing correlations between features to identify the most relevant variables for predictive modeling.
* Data Visualization: Creating visualizations such as scatter plots, bar charts, and heatmaps to illustrate key findings and highlight relationships between features.
* Result Interpretation and Reporting: Analysing model results to understand the impact of different features on video engagement and compiling insights into a detailed report.

**5. Methodology**

The project will follow a structured approach:

* Data Collection: The dataset will be sourced from YouTube’s API or a relevant data repository, containing attributes such as video titles, views, likes, dislikes, comments, and more.
* Data Preprocessing: Handling Missing Data and techniques like mean, median imputation, or removal of incomplete records will be used.
* Outlier Detection
* Exploratory Data Analysis (EDA): Conduct descriptive statistics to summarize the dataset. Create visualizations to explore the relationships between features.
* Feature Selection: Perform correlation analysis and dimensionality reduction techniques to identify significant variables.
* Predictive Modelling: Develop and evaluate models using machine learning algorithms to predict video performance metrics.
* Visualization: Use advanced visualization tools to present data findings clearly.
* Reporting: Compile results, visualizations, and recommendations into a comprehensive report.

**6. Tools and Technologies**

The project will utilize the following tools and technologies:

* **Programming Language:** Python
* **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
* **IDE:** Jupyter Notebook or any Python-compatible Integrated Development Environment (IDE)
* **Data Source:** Kaggle datasets

**7. Expected Outcomes**

* Identification of key factors influencing video popularity and engagement.
* Predictive models capable of forecasting future performance based on current data.
* Insightful visualizations showcasing trends and patterns within the data.
* A comprehensive report with actionable recommendations for optimizing YouTube content strategies.

**8. Timeline**

The project is expected to be completed within a [specific timeframe, e.g., 4 weeks], with the following milestones:

* Week 1: Data Collection and Preprocessing
* Week 2: Exploratory Data Analysis and Feature Selection
* Week 3: Model Building and Evaluation
* Week 4: Visualization
* Week 5: Reporting and Final Submission

**9. Conclusion**

This YouTube Data Analysis project aims to identify what drives video performance and viewer engagement on YouTube. By exploring and analysing the data, we gain insights into the key factors that make videos popular. The project uses data preprocessing, statistical analysis, and predictive modeling to find patterns and trends.

The insights and recommendations from this analysis will help content creators improve their strategies, optimize video performance, and better connect with their audience. Ultimately, this project highlights the importance of data-driven decisions in boosting success on YouTube.