EXPLORATORY DATA ANALYSIS ON YOUTUBE TRENDING VIDEOS

I. ABSTRACT

YouTube is one of the largest video-sharing platforms globally, with millions of users and content creators. Understanding the dynamics of trending videos can offer valuable insights for marketers, creators, and data scientists. This project focuses on performing Exploratory Data Analysis (EDA) on YouTube's trending video dataset using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn. The analysis uncovers patterns in video categories, publication times, user engagement metrics (likes, dislikes, comments), and correlation trends. This research identifies what attributes contribute to a video's likelihood to trend and provides a data-driven foundation for content optimization.

II. INTRODUCTION

The goal of EDA is to summarize the main characteristics of a dataset, often using visual methods. YouTube, as a leading video content platform, provides data-rich insights that are ideal for EDA. This project aims to examine a publicly available dataset of trending YouTube videos, extracting information to answer questions like:

- What categories trend the most?
- How do likes and comments correlate with views?
- Does upload time affect trending potential?

The dataset is pre-processed and analyzed using Python, allowing for interactive visualizations and statistical insights. The results highlight critical factors that influence video performance on YouTube.

III. SYSTEM REQUIREMENTS

A. Hardware Requirements

- Processor: Intel i5 or higher
- RAM: 4 GB minimum
- Storage: 1 GB free space
- OS: Windows/Linux/macOS

B. Software Requirements

- Python 3.8 or higher

- Jupyter Notebook
- Libraries: pandas, numpy, seaborn, matplotlib, plotly
- IDE: VS Code / JupyterLab / Google Colab

IV. DATASET DETAILS

- Source: Kaggle (YouTube Trending Video Statistics)
- Fields Included:
    - video_id, title, channel_title, category_id
    - publish_time, views, likes, dislikes, comment_count
    - tags, description, country

The dataset includes daily records of videos that trended in a particular country, capturing their metadata and engagement statistics.

V. METHODOLOGY

A. Data Cleaning

- Removed null values and duplicates
- Converted publish_time to datetime format
- Mapped category_id to category names via a separate JSON file

B. Feature Engineering

- Extracted publish_day, publish_hour from publish_time
- Created like_ratio = likes / views
- Created comment_ratio = comments / views

C. Analysis Techniques

- Univariate analysis for individual feature distributions
- Bivariate analysis for correlations
- Temporal analysis for hour/day-based trends
- Visualizations using Seaborn, Matplotlib, and Plotly

VI. IMPLEMENTATION MODULES

1. **Data Loading and Preprocessing**
    - Loaded CSV files using Pandas
    - Checked for missing values
    - Standardized column formats
2. **Descriptive Statistics**
    - Used .describe(), .value_counts(), and .groupby() to summarize data
    - Plotted bar charts, histograms, and box plots
3. **Correlation Analysis**
    - Heatmap for correlation between views, likes, dislikes, and comments
    - Scatter plots for visualizing relationship strength
4. **Category-Level Analysis**
    - Aggregated view count per category
    - Ranked categories by trending frequency
5. **Time-Based Analysis**
    - Determined most popular upload hours and days
    - Compared performance of videos based on upload time

VII. RESULTS AND DISCUSSION

- Top Trending Categories: Entertainment, Music, and Sports consistently dominated the trending lists.
- Correlation:
    - Strong correlation between views and likes
    - Moderate correlation between views and comment count
- Timing Impact:
    - Videos uploaded between 6 PM–9 PM showed a higher probability of trending

    - Weekends showed a spike in video engagements
- Engagement Ratios: Videos with high like/view and comment/view ratios had better visibility on the trending page.

These insights suggest that not only content category but also engagement metrics and timing play a significant role in trend dynamics.

VIII. CONCLUSION

The EDA on YouTube Trending Videos has revealed key patterns in video performance, viewer engagement, and temporal trends. These findings offer content creators actionable insights to optimize video publishing strategies. The project demonstrates the power of data-driven analysis using Python and standard data science libraries. Through this analysis, trends were visualized and interpreted to provide value for stakeholders in media, marketing, and digital content creation.

IX. FUTURE SCOPE

- Sentiment Analysis on comments to assess viewer reaction
- Predictive Modeling to estimate trend potential of new videos
- Geo-based Analysis by integrating trending data across countries
- Dashboard Development using Plotly Dash or Streamlit for real-time visualization
- API Integration for real-time data fetching from YouTube Data API