1. Problem Statement

By recommending videos based on users' preferences, YouTube's recommendation algorithm aims to increase users' watch time on the platform. Instead of focusing on what is uploaded, YouTube focuses on what viewers are watching. View counts are the primary factor that determines YouTube's revenue. As part of this targeted livestream, I will analyze interactive data and content weight to understand how different types of data, including likes, follows, shares, comments, and watch time, affect view counts.

This dataset includes several months (and counting) of daily trending YouTube videos. Data is provided for regions such as the US, GB, DE, CA, and FR (United States, Great Britain, Germany, Canada, and France, respectively), with up to 200 trending videos listed per day. It now also includes data from RU, MX, KR, JP, and IN (Russia, Mexico, South Korea, Japan, and India, respectively) over the same period.

The dataset contains the video title, channel name, publication time, tags, views, likes, dislikes, description, and comment count.

I will analyze the direction into the main two parts of the interactive data and work weight.

Interactive data: the algorithm description inside the two directions of the factors mentioned, namely, positive and negative factors, if the positive factors are good, the work of the recommended traffic will be good the more, and vice versa will reduce the recommended.

Work weight: the algorithm of de-emphasis and dispersion and other processing is actually in the shortlisted works for weighted sorting, it is through a variety of coefficients to calculate the results.

2. Articulation of value

The value of this analytics project lies in its ability to provide insight into how YouTube's recommendation algorithm drives viewership. In this project, we examine the impact of different interactive elements (likes, shares, comments, watch time), as well as the weighting mechanism for the algorithm, in order to gain a better understanding of how user engagement and video ranking are affected. As a result of the findings, creators can optimize content strategies to increase exposure and watch time, while YouTube receives valuable feedback on its recommendation system's performance. Furthermore, by examining data across multiple regions, the project reveals how cultural and regional differences affect algorithmic behavior, which could inform more personalized and region-specific recommendation strategies.

This analysis has significant potential not only for content creators who aim to increase engagement, but also for advertisers and platform designers to maximize viewer retention and platform profitability. By analyzing the balance between user satisfaction and monetization goals of the platform, the project can inform all stakeholders in the YouTube ecosystem of better-informed strategies.

3.Calculation of the potential economic value

- **Document and footnote your assumptions**
- **Show how you calculated the value**

To calculate the potential economic value of YouTube dataset, we'll estimate the economic benefits in terms of how this data could be used by different stakeholders, such as YouTube itself, content creators, advertisers, or third-party data analytics companies. Let's break this down step by step:

Assumptions:

Monetization of YouTube views: We assume YouTube generates $0.0018 per view based on average CPM (Cost Per Thousand Impressions) earnings. This rate is common across advertisers and is subject to variance by region and ad type.

View growth due to optimization: A successful use of the dataset for better recommendations could lead to an increase in watch time and views. We assume a conservative 2% increase in viewership for optimized content.

Number of videos impacted: The dataset includes up to 200 trending videos per day across 10 regions. We will assume this dataset can help optimize recommendations for 20% of these videos, or 40 videos per day.

The dataset can be used for sentiment analysis, video categorization, RNNs, popularity prediction, and long-term statistical trends, each offering potential value to creators, marketers, and YouTube.

Estimate the impact on views for YouTube content optimization:

- Number of trending videos per year
- Videos optimized (20% of trending videos)
- Average views per trending video
- Total views impacted
- Increase in viewership due to optimization

- Revenue per view (CPM rate of $0.0018)
- The potential additional revenue YouTube could generate by optimizing recommendations using this dataset.

By analyzing what factors drive video popularity and applying ML models (like RNNs), creators could optimize their content for higher engagement. If the dataset increases the success rate of trending videos by even 1%, it could add substantial revenue for creators, though this is more difficult to quantify precisely without creator-specific data.

Analyzing how video popularity evolves over time, broken down by region and category, could provide significant value for platforms like YouTube, which are always looking to refine their recommendation algorithms and content categorization.

With this dataset, YouTube could potentially generate an additional $262,800 annually by optimizing video recommendations through improved understanding of interactive data and content weight. Furthermore, this dataset has broader applications for sentiment analysis, machine learning, and statistical trend analysis, which could further enhance user engagement and ad revenue, though those values are more challenging to quantify directly without deeper data exploration.

4.Project plan: Build a 13-week plan. Identify the steps, identify the weeks, and what you will do in each step (you can look at the syllabus to build this plan)

Week 1: Project Kickoff & Problem Definition

- Define the project objectives and key research questions.
- Finalize the problem statement, focusing on YouTube recommendation algorithms and interactive data.
- Set up the tools, environment, and project framework.

Week 2: Data Collection & Exploration

- Review and explore the dataset (e.g., video views, likes, comments).
- Perform an initial analysis to understand trends and detect any missing or inconsistent data.
- Set up a data pipeline for analysis.

Week 3: Data Cleaning & Preprocessing

- Clean the dataset: handle missing values, format inconsistencies, and irrelevant fields.
- Normalize variables (views, likes, comments) for consistent analysis.
- Visualize initial insights to verify data readiness.

Week 4: Descriptive Statistics & Initial Insights

- Conduct descriptive statistics to understand correlations between factors like likes and views.
- Identify which data factors might drive video popularity.
- Begin documenting early insights.

Week 5: Sentiment Analysis of Comments

- Clean and process YouTube comments for sentiment analysis.
- Apply natural language processing (NLP) techniques to analyze comment sentiment (positive/negative).
- Analyze how comment sentiment affects video performance.

Week 6: Categorizing YouTube Videos

- Use clustering algorithms (K-means, etc.) to categorize YouTube videos based on data like views, likes, and comments.
- Explore categories and how they relate to video popularity.

Week 7: Machine Learning Model for Comment Generation

- Train a Recurrent Neural Network (RNN) to generate YouTube comments.
- Evaluate how well the model generates realistic comments.
- Document findings for further analysis.

Week 8: Analyzing Factors Influencing Video Popularity

- Conduct regression analysis to identify which factors affect video popularity the most.
- Test hypotheses from earlier analysis to confirm or reject them.
- Document the key findings.

Week 9: Investigating Weighted Sorting in YouTube's Algorithm

- Analyze how YouTube’s algorithm might weight factors like views and likes when ranking videos.
- Simulate a weighted sorting process to see the impact on recommendations.
- Document insights from this analysis.

Week 10: Regional Analysis

- Compare video performance across regions (US, GB, DE, etc.).
- Identify regional trends and differences in recommendation algorithms and video engagement.
- Visualize regional data for insights.

Week 11: Long-term Trends Analysis

- Perform a time-series analysis to understand how video popularity changes over time.
- Identify trends in views, likes, and comments over months.
- Document key insights and prepare visualizations.

Week 12: Final Model Tuning & Evaluation

- Fine-tune the machine learning models (e.g., sentiment analysis, RNN).
- Validate models using cross-validation or other techniques.
- Finalize model results and prepare documentation.

Week 13: Final Report and Presentation

- Compile all analysis and insights into a final report.
- Create a presentation summarizing key findings and conclusions.
- Practice and prepare for the final presentation.

5.Discuss the dataset: describe the dataset you've found, where you've found it from, and how will this dataset solve the problem statement?

The dataset I am using for this project is a comprehensive collection of trending YouTube videos from multiple regions, including the US, GB, DE, CA, FR, RU, MX, KR, JP, and IN. It includes several months of daily trending video data, with up to 200 videos listed per day in each region. The dataset provides key video attributes such as video title, channel name, publication time, tags, views, likes, dislikes, description, and comment count, allowing for detailed analysis.

This dataset was sourced from YouTube's publicly available trending data, which tracks the most popular videos in each region daily. By leveraging this data, I can analyze how different types of interactive data (likes, comments, shares, watch time) and content weight affect a video's view count. The dataset's rich attributes will allow me to explore factors influencing YouTube's recommendation algorithm, such as the impact of user engagement on video ranking, regional differences in viewing behavior, and how video metadata affects popularity.

This dataset directly addresses the problem statement by providing a solid foundation for understanding how YouTube's recommendation algorithm functions. It will help identify which interactive data factors contribute positively or negatively to video recommendations, and how content is weighted and sorted in the algorithm. Through statistical analysis and machine learning, this dataset can offer valuable insights for content creators, marketers, and even YouTube itself on how to optimize content for higher visibility and engagement.

6.Identify the type of modeling that will solve the problem you identified: supervised or unsupervised. If supervised, is this a classification or regression problem? If it is classification, is it binary or multi class? If it is unsupervised, what type of unsupervised learning is it?

I will use unsupervised Learning to analyze and categorizing YouTube videos into distinct groups based on their features (e.g., engagement metrics, region, or video type). Because unsupervised learning techniques like K-means clustering or hierarchical clustering will be employed. This will help discover natural groupings in the data, enabling insights into how different types of videos perform and are recommended.

Appendix

https://www.kaggle.com/datasets/datasnaek/youtube-new?resource=download