

Project stage I report
Dataset Selection & Project Setup

Total: 100 points

Group # 4

Project title/topic : YouTube trending video analysis

Dataset source (url or resource):

<https://www.kaggle.com/datasets/datasnaek/youtube-new/data>

Task 1 Problem framing (10 points)

Your understanding of the project (what you want to study or predict).

Write 4-6 sentences describing: the task (classification/regression/..); who/what benefits from this study; why this dataset fits (size, feature).

In this project, we perform a regression task to predict the number of days a video can be trended based on the number of views, number of likes, dislikes and comments, category name received on its first day of release. From this, we can understand what factors contribute to the video to be trending. This research will be helpful for video content creators. They can use our findings to develop more effective strategies for video content planning and title design, thereby increasing the reach of their videos. This dataset fits because it provides key factors like views, likes, dislikes, and comments. This allows us compute meaningful ratios like likes_ratio and comments ratios which tell us about the video's performance or impact and audience engagement and sentiment. This dataset was selected since it had around 40,000 observations to begin with, the features are such that we can create new features from them for further analysis.

Task 2 Dataset exploration (20 points)

Show the first 10 rows of your dataset; create a dataset dictionary:

Feature	Type: Int / float / ...	examples	Missing %
video_id	String	5qpjK5DgCt4, 2RVw2_QyxQ	0
first_trending_ts	Date	2018-06-06, 2/1/2018	0
views_day1	int	67429, 563746	0
likes_day1	int	438,5214	0
dislikes_day1	int	54,172	0
comments_day1	int	94,1490	0
likes_ratio	int	0.988143008	0.4 %(27 rows)
comments_ratio	int	0.000950979	
Category id	String	2018-06-06	0
Category Title	String	News & Politics	0
Trendingdays (Target variable)	int	6	0

Task 3 Feature exploration (40 points)

Deal with the missing values/outliers; and then create a clean dataset for the next stage analysis.

- Each member works on **at least one important feature** (≥ 3 features per team total).
- In your summary, clearly states each member's contributions.

For this project, we selected nine features suitable for analysis. We established a clear division of labor within the team to improve efficiency and ensure accurate processing of each variable. Riya was primarily responsible for basic

variables related to user interaction, including Views, Likes, Comments, and Dislikes. Naveena was responsible for calculating derived variables from the raw data, including Like Ratio ($\text{likes} / (\text{likes} + \text{dislikes})$) and Comment Ratio ($\text{comment_count} / \text{views}$), and was also responsible for organizing and analyzing Video IDs. Xiuwen was responsible for processing data related to video classification and target variables, including the definition and organization of Category ID, Trending Date, and Target Variable. This division of labor allowed each team member to fully leverage their strengths, ensuring a systematic and efficient data processing process.

Each of us performed a comprehensive cleanup of the dataset and then worked on our own variables, including converting time data formats, deleting and merging columns, calculating the number of missing variables, and counting outliers. Outliers are calculated for all the columns except for video_id and category name using log normalization and z scores.

Task 4 Save cleaned data (10 points)

- Save your final cleaned dataset as **newdata.csv** (UTF-8, includes header).
- Place it in your GitHub repo (e.g., /data/newdata.csv). If data is too large/private, include a script to reproduce it and README with retrieval steps.

Task 5 Delivery (20 points)

- **Notebook:** Use Markdown cells to explain every step (why you did it, not just what). Upload to GitHub.
- Export and upload **HTML or PDF** in canvas, together with project stage I summary.