Project Proposal: The Viral Gap: Comparative Analysis of Artist Popularity on YouTube vs. TikTok

1. Name of the Project and Team Members

Project Title: The Viral Gap: Comparative Analysis of Artist Popularity on YouTube vs. TikTok

Team Members: Henry Yu, Rohit Praveen

2. What problem are you trying to solve?

Research Question: "How does artist popularity correlate between long-form legacy media (YouTube) and short-form viral content (TikTok), and which artists bridge the gap between passive consumption and active user engagement?"

Traditional popularity metrics (like streaming numbers) often fail to capture "virality." We hypothesize that there is a distinction between **"Consumer Popularity"** (high passive viewership on YouTube) and **"Viral Relevance"** (high content creation volume on TikTok). By analyzing **10 distinct artists** across diverse genres (Pop, Rap, R&B, Latin, Alt), we aim to identify:

- **Platform Dominance:** Which artists are sustained by loyal fanbases vs. algorithmic trends.
- **The "Viral Gap":** Quantifying the difference between how an audience *watches* an artist versus how much they *create* with them.

3. How will you collect data and from where?

We are implementing a dual-source pipeline to compare official metrics against User Generated Content (UGC) volume.

(1) YouTube Data API v3 (Official Source)

- **Method:** Authenticated API requests via google-api-python-client.
- **Metrics:** Channel Total Views, Subscriber Count.
- **Role:** Represents "Passive Consumption" (historical, monetized popularity).

(2) TikTok Hashtag Scraping (Custom Engineering)

- **Challenge:** TikTok aggressively blocks standard requests and redirects desktop users to "Shop" pages, hiding view counts.
- **Solution:** We pivoted to a **Selenium WebDriver** approach. We added a "human-in-the-loop" feature where the script pauses, allowing the user to manually solve CAPTCHAs or close pop-ups, before programmatically extracting the data.
-
- **Metrics: Hashtag Post Count** (e.g., #artistname count).
- **Role:** Represents "Active Engagement" (how many users are creating content about the artist).

4. What analysis will you do and what visualizations will you create?

Data Analysis Methodology:

- **Normalization:** Since YouTube Views are in the billions and TikTok Posts are in the millions, we will use **Min-Max Scaling** to normalize all data to a 0–1 scale for fair comparison.
- **Combined Popularity Index:** We will calculate a weighted score for every artist to determine an overall winner:
  $\text{Score} = (0.55 \times \text{NormYouTube}) + (0.45 \times \text{NormTikTok})$
  (YouTube is weighted higher as it represents verified consumption, while TikTok represents viral buzz).

Visualizations (using Matplotlib & Seaborn):

1. **Platform Dominance Scatter Plot:** A quadrant analysis plotting *Normalized YouTube Views* (X) vs. *Normalized TikTok Posts* (Y) to visually cluster artists into "Viral Stars," "Legacy Giants," or "Dual Threats."
2. **The "Viral Gap" Bar Chart:** Comparing the Combined Popularity Index ranking to identify which artists over-perform on social media relative to their video views.
3. **Scale Comparison Chart:** A dual-axis or log-scale chart comparing raw Billions (Views) vs. Millions (Posts) to visualize the magnitude difference between consumption and creation.