

Final Project Report

DSCI 510: Principles of Programming for Data Science

1. Project Overview & Team

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

Team Members:

- [Rohit Praveen] (USC ID: 6393931049) – [GitHub Username: rohit-design]
- [Henry Yu] (USC ID: 1678028702) – [GitHub Username: henryyu528]

Research Question:

"How does artist popularity correlate between relatively long lasting forms of media (YouTube) and relatively quick viral media (TikTok), and which artists bridge the gap between passive consumption and active user engagement?"

Description:

In the modern streaming era, "popularity" is split between passive consumption (watching a music video) and active engagement (creating a dance trend). This project investigates whether "Consumer Popularity" on YouTube translates directly to "Viral Relevance" on TikTok. By analyzing the media relevance of 10 major artists across diverse genres, we quantify the "Viral Gap"—the difference between an artist's viewership and their cultural footprint on social media.

2. Data Collection

Data Collected:

We collected cross-platform performance metrics for 10 distinct artists (Taylor Swift, Bad Bunny, Drake, The Weeknd, Billie Eilish, Post Malone, Kendrick Lamar, Adele, Doja Cat, and NBA YoungBoy).

- **Total Data Points:** 20 primary metrics (10 YouTube Channel Stats + 10 TikTok Hashtag Stats), processed into a ranked dataset.

A. Data Sources & Approach:

1. YouTube (Passive Consumption):

- **Source:** YouTube Data API v3.
- **Method:** We authenticated via [google-api-python-client](#) to fetch the official channel statistics.
- **Key Metric:** `viewCount` (Total lifetime views). This represents the "legacy" popularity of the artist.

2. TikTok (Active Engagement):

- **Source:** TikTok Hashtag Search Pages (<https://www.tiktok.com/tag/{artist}>).
- **Method:** We engineered a custom **Selenium Web Scraper**.
- **Key Metric:** **Post Count** (e.g., "18.0M posts"). This represents "viral" popularity—the volume of user-generated content created about the artist.

B. Changes & Challenges:

- **Challenge (Anti-Scraping):** Our original plan was to use simple `requests` libraries to scrape TikTok. However, TikTok implemented aggressive anti-bot measures that redirected our script to "Shop" or "Login" pages, hiding the data.
- **Resolution (The "Manual-Assist" Scraper):** 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.
- **Metric Shift:** We originally intended to collect TikTok *views*, but TikTok recently hid view counts on many hashtag pages in favor of *post counts*. We adjusted our analysis to use "Post Counts" as a proxy for engagement, which arguably measures "active creation" better than views.

3. Analysis & Visualizations

A. Analysis Techniques & Findings:

We used Min-Max Normalization (`sklearn.preprocessing.MinMaxScaler`) to scale the disparate metrics (YouTube Views in Billions vs. TikTok Posts in Millions) to a uniform 0–1 range.

We then calculated a Combined Popularity Index using a weighted formula:

$$\text{Score} = (0.55 \times \text{NormYouTube}) + (0.45 \times \text{NormTikTok})$$

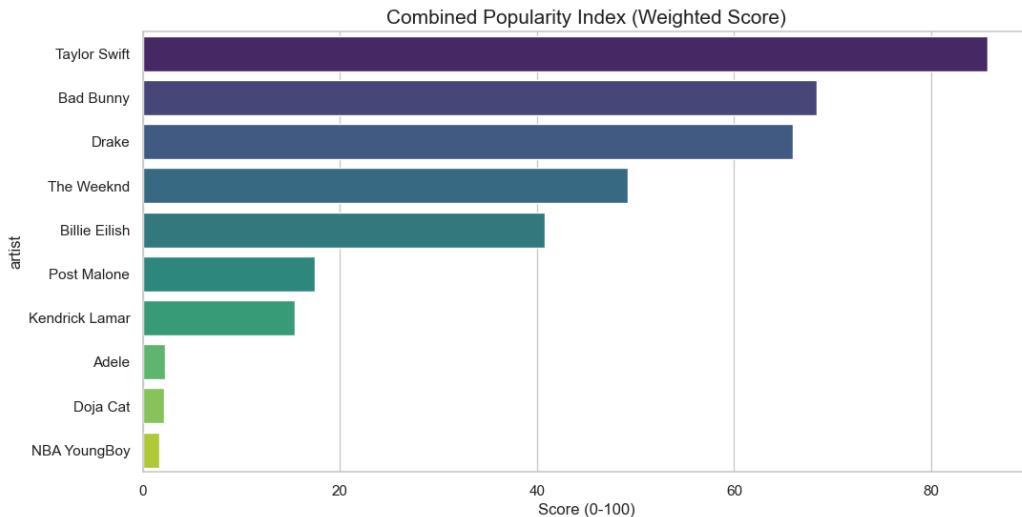
- *Weighting Logic:* YouTube (0.55) is weighted slightly higher as it represents monetized, verified consumption, while TikTok (0.45) represents unverified viral buzz.

Key Findings from the Data:

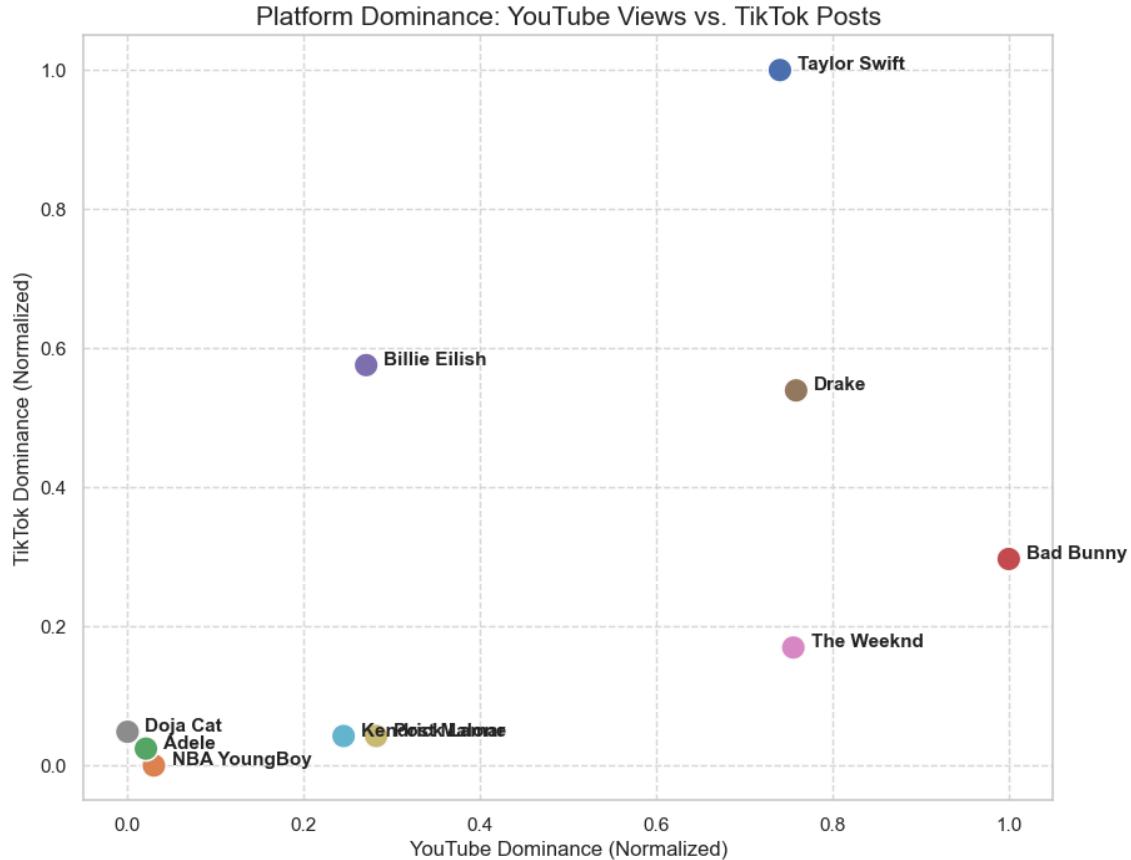
1. **The "Dual Threat" Leader:** Taylor Swift ranked #1 (Score: 85.7) by a significant margin. She dominates TikTok with **18.0M posts** (the highest in the dataset) while maintaining massive YouTube viewership (8.6B).
2. **The YouTube Giant:** Bad Bunny (#2) actually has the **highest YouTube views** in the entire dataset (10.5B), surpassing Taylor Swift. However, his TikTok footprint (6.4M posts) is significantly lower than Swift's, indicating his audience is more about *consumption* than *content creation*.
3. **The Viral Outlier:** Billie Eilish (#5) presents an interesting case. Despite having significantly lower YouTube views (5.3B) compared to The Weeknd (8.7B), she has nearly **3x the TikTok posts** (11.0M vs 4.3M). This identifies her as a "Viral First" artist whose cultural impact is driven by Gen-Z social trends.

B. Figures & Visualizations:

- **Figure 1: Combined Popularity Ranking (Bar Chart)**
 - *Setup:* A seaborn bar chart ranking artists by their final weighted score (0-100).
 - *Meaning:* Visualizes the hierarchy of cross-platform dominance. It clearly separates the "Superstars" (Swift, Bad Bunny, Drake) from the niche giants (NBA YoungBoy).



- **Figure 2: Platform Dominance (Scatter Plot)**
 - *Setup:* X-Axis = Normalized YouTube Views; Y-Axis = Normalized TikTok Posts.
 - *Meaning:* This quadrant analysis reveals the "nature" of an artist's fame.
 - *Top-Right:* Multi-platform Superstars (Taylor Swift).
 - *Bottom-Right:* YouTube-Heavy Artists (Bad Bunny, The Weeknd).
 - *Top-Left:* Viral-Heavy Artists (Billie Eilish).



C. Observations & Conclusion:

The data proves that high streaming numbers do not guarantee high social engagement. While Bad Bunny wins on passive consumption (YouTube), he trails significantly behind Billie Eilish in active users. This suggests that "virality" is a distinct metric from "popularity" even though they may seem very similar. Artists like Taylor Swift who successfully bridge this "Viral Gap" achieve the highest overall cultural dominance and are thus the bigger artist.

D. Impact:

This analysis provides a data-driven framework for understanding modern music marketing. It demonstrates that artists cannot rely solely on legacy platforms; to achieve #1 status, they must foster an environment where fans create content, not just watch it. “Legacy Fame” and “Viral Relevance” are both important for an artist’s success.

4. Future Work

A. Improvements & Expansion:

Given more time, we would expand this project in three key directions:

1. **Fully Autonomous Scraping:** We would implement "stealth" browsing techniques (e.g., rotating residential proxies and browser fingerprinting) to bypass TikTok's anti-bot measures without human intervention.
2. **Temporal Analysis:** Instead of a static snapshot, we would track these metrics over a 4-week period to measure *growth velocity*. (e.g., "Is Kendrick Lamar gaining TikTok posts faster than Drake?").
3. **Sentiment Analysis:** We would scrape the comments from the top YouTube videos to perform Natural Language Processing (NLP), determining if the engagement is positive (fan love) or negative (controversy), adding a qualitative layer to our quantitative ranking.