Project Proposal: TikTok Trends Analysis

Wenjing Huang (whuang08@usc.edu)
USC ID: 3860016877

Revised: December 15, 2025

1. What problem are you trying to solve?

Short-form video platforms like TikTok generate massive amounts of content, but only a small fraction of videos achieve ”viral” status. While creators often focus on content quality, the quantitative relationship between reach (views) and active engagement (likes, comments) remains unclear at scale.

This project aims to solve the problem of **quantifying engagement patterns** in high-performing short-form videos. Specifically, I will address the research question: *”What is the statistical relationship between view counts and like counts for trending content, and does this relationship follow a predictable linear or non-linear pattern?”* Understanding this helps creators and marketers benchmark their performance against typical viral engagement rates.

2. How will you collect data and from where?

I will collect data using the **Kaggle API** rather than manual web scraping. This decision was made to ensure access to a large ($N \approx 50,000$), standardized, and legally compliant dataset that avoids the anti-bot protections found on live social media sites.

- **Source:** The ”YouTube Shorts and TikTok Trends 2025” dataset hosted on Kaggle.
- **Method:** I will use the `kagglehub` Python library to authenticate and programmatically download the dataset within the `get_data.py` script.
- **Dataset Size:** Approximately 50,000 aggregated trend records.
- **Data Fields:** I will extract columns including `view_count`, `like_count`, `comment_count`, and trend metadata.

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

The analysis will focus on descriptive statistics and correlation analysis to measure the intensity of user engagement.

Planned Analysis:

- **Data Cleaning:** I will normalize column names, handle missing values, and convert string-based metrics into numeric types using `pandas`.
- **Descriptive Statistics:** I will calculate the mean, median, and maximum values for views and likes to understand the scale of the data.
- **Correlation Analysis:** I will compute the Pearson correlation coefficient between views, likes, and other engagement metrics to determine how strongly they co-vary.

Planned Visualizations:

- **Virality Scatter Plot:** A log-log scatter plot displaying `view_count` vs. `like_count`. This will visualize the "shape" of virality and highlight outliers that have high views but low engagement (or vice versa).
- **Correlation Heatmap:** A color-coded matrix showing the correlation strength between all numeric variables (views, likes, comments, etc.).