1. **Data Preparation**

To conduct the analysis, weekly leaderboard data was collected from three major music platforms, as identified by Decision Lab statistics (2024): Apple Music, Spotify, and Zing MP3. While YouTube Music is a strong competitor, the platform does not provide Vietnamese-specific data, making it unsuitable for this analysis.

**Data Sources**

**1. Apple Music (**[**Top Charts**](https://www.top-charts.com/songs/all-genres/vietnam/apple-music)**)**

* Top Charts is a third-party website that archives weekly Apple Music data. Since the official Apple Music website only displays the current leaderboard, this platform serves as a credible alternative for historical data.
* Data collection was automated using a script to download files for multiple weeks. (Scrape\_Apple\_Music\_W45\_2019\_to\_W52\_2024.py)
* The dataset includes **rank**, **song name**, and **artist name**.
* While data accuracy was validated by comparing two weeks of Top Charts data with Apple Music’s official leaderboard, potential inaccuracies due to third-party sourcing remain a consideration.
* Sampling bias may exist since Apple Music requires a paid subscription, limiting the dataset to users willing to pay for the service. This could result in an overrepresentation of iPhone users and individuals more inclined to purchase albums or concert tickets.
* The data is publicly accessible, mitigating licensing concerns.

**2. Spotify (**[**Spotify Charts**](https://charts.spotify.com/charts/overview/global)**)**

* Data was sourced from Spotify’s official website and downloaded using an automated script. (Scrape\_Spotify\_W11\_2019\_to\_W52\_2024.py)
* The dataset includes **rank**, **song name**, **artist name**, and **streams**.
* Spotify offers both free and premium tiers. Streams are counted after 30 seconds of playback, meaning rankings reflect a broader demographic but may favor users who prefer free streaming services over other platforms.
* The data is publicly available, ensuring compliance with licensing requirements.
* Spotify has experienced significant growth, particularly among younger audiences, making its data a valuable component of this analysis.

**3. Zing MP3 (**[**Zing MP3**](https://zingmp3.vn/zing-chart-tuan/Bai-hat-Viet-Nam/IWZ9Z08I.html)**)**

* Data was sourced from Zing MP3’s official website. Since no direct download option exists, web scraping was used to automate data collection while ensuring compliance with ethical guidelines. (Scrape\_ZingMP3\_W32\_2018\_to\_W52\_2024)
* The dataset includes **rank**, **song name**, and **artist name**.
* Sampling and ranking biases may exist, as some popular songs on other platforms appear absent from Zing MP3’s charts, while lesser-known artists frequently occupy top positions. This suggests that platform-specific user behavior influences rankings.
* The data is publicly available, negating licensing concerns.
* As Vietnam’s oldest music streaming platform, Zing MP3’s extensive user base makes its data a valuable addition to this analysis.

1. **Data Processing**

Errors were identified in the downloaded Apple Music and Spotify data, as well as the scraped Zing MP3 dataset.

**1. Data Issues**

**a. Apple Music & Spotify**

* Weekly data files were stored separately and required consolidation.
* Vietnamese diacritical marks led to inconsistencies in artist and song names.
* Multiple artist names were stored within a single cell.
* No genre information was provided, but it is a key variable for analysis.
* The **Source File** column contained combined year-week values that required reformatting.

**b. Zing MP3**

* Web scraping inconsistencies resulted in missing weeks.
* Multiple artist names were stored within a single cell.
* No genre information was provided.
* The **Week** column contained incorrect date formatting.

**2. Solutions**

**a. Apple Music & Spotify**

**a.1. Merging Files**

* A script was used to consolidate all weekly song data into a single dataset. (Merge.py)
* This solution was fully effective.

**a.2. Fixing Encoding Issues in Artist Names**

* Encoding errors caused artist names to display incorrectly.
* Extracted all artist names into a separate text file, de-duplicated them, and converted them back to UTF-8 encoding.
* While this solution resolved most issues, some names remained incomplete due to missing vowels.

**a.3. Splitting Artist Names**

* The =TEXTSPLIT() function in Excel was used to separate artist names into multiple columns.
* The first artist was stored in the **1st Artist** column, followed by additional artist columns.
* This solution was fully effective.

**a.4. Fetching Artist Genres via Spotify API**

* The **1st Artist** column was used to query Spotify’s API for genre information.
* Retrieved genre data was stored in a secondary sheet and manually adjusted for accuracy.
* Some bias was introduced as Spotify assigns genres based on playlist classification rather than specific songs. Additionally, artists span multiple genres, leading to potential inconsistencies.

**a.5. Processing Source File and Weekly Dates**

* The **Source File** column was extracted and de-duplicated.
* A separate sheet was used to manually input the starting date, with the following dates calculated using =(Previous Cell) + 7.
* This solution was fully effective.

**b. Zing MP3**

* Missing weeks were manually added to complete the dataset.
* Although time-intensive, this solution ensured data completeness.

With these data preparation steps completed, the dataset is now structured for in-depth analysis.

1. **Appendices**

https://www.decisionlab.co/blog/vietnam-music-streaming-industry-q1-2024

https://support.spotify.com/us/artists/article/how-we-count-streams/