Data Wrangling Project Proposal

Introduction

The streaming music industry has dramatically transformed how consumers interact with music, with platforms like Spotify offering easy access to millions of songs. As of April 2024, Spotify has reached over 615 million monthly active users globally [1], making it one of the largest music streaming platforms in the world. This vast user base and the platform's unique data insights provide an invaluable resource for analyzing song characteristics and their impact on popularity. Our project seeks to explore this data to better understand the features that define popular music on Spotify, focusing on attributes such as danceability, energy, and lyrical content. Specifically, we aim to categorize song genres using advanced natural language processing (NLP) techniques and analyze how musical features, such as danceability and energy, correlate with a song's popularity, measured by streaming counts. By leveraging Spotify's Web API and publicly available datasets, we aim to uncover key trends that shape the digital music landscape and offer valuable insights into the drivers of music popularity and genre classification on streaming platforms.

Data Description

The Spotify Most Streamed Songs Dataset, sourced from Kaggle [2], contains detailed information on the most streamed songs on Spotify from various years, including attributes such as song name, artist, release year, streams, danceability, energy, and other musical features like valence and tempo. The dataset spans multiple years, capturing data from the early 2000s to more recent releases, allowing for an in-depth exploration of trends in musical popularity over time. This rich dataset provides a foundation for analyzing the relationship between musical characteristics and song popularity, as well as exploring genre classification based on song attributes.The Spotify Web API [3] allows access to an extensive array of metadata about tracks, albums, and artists. It includes track features such as acousticness, energy, valence, tempo, and lyrics. This API provides access to global data, making it possible to analyze popular music trends from different regions. We will use this API to extract additional song features not included in the Kaggle dataset, with a particular focus on song lyrics for natural language processing (NLP) to categorize songs by genre. The combination of historical streaming data and metadata from the API offers a comprehensive view of the evolution of music consumption.

Aim 1: Categorization of Song Genres Using Lyrics

1.1 Accessing Lyrics through Genius API and Spotify API:

We will use the **Spotify Web API** to gather song metadata, focusing on details like genre, artist, and album. This metadata helps us categorize songs correctly. For the lyrics, we will use the **Genius API** to download song lyrics, which provide a detailed textual representation of a song’s content. These lyrics are essential for identifying patterns and emotions that can help us categorize the songs into their respective genres.

1.2 NLP Analysis for Genre Categorization:

Using **Natural Language Processing (NLP)** techniques, we will analyze the lyrics of songs to categorize them into genres like R&B, Pop, and Rap. We will apply sentiment analysis to understand the emotional tone of the lyrics, as well as other techniques like topic modeling and word embeddings to identify linguistic features that define each genre. This analysis will help us reveal patterns in the lyrics that distinguish one genre from another.

Aim 2: Relationship Between Danceability, Energy, and Popularity

2.1 Correlation Analysis:

We will explore the relationship between key musical features such as **danceability** and **energy** and their popularity, measured by **streaming counts**. We hypothesize that songs with higher levels of danceability and energy are more likely to be popular, as these characteristics could make songs more engaging to listeners. By analyzing the correlation between these features and streaming counts, we aim to understand how they contribute to a song's success.

2.2 Visualization of Trends:

To better illustrate the relationship between danceability, energy, and popularity, we will create visualizations such as **scatter plots** and **heatmaps**. These graphical tools will help us highlight trends and correlations between these musical features and streaming popularity, making it easier to identify patterns and interactions between features.

Concluding Remarks

Our goal was to visually present the defining characteristics of the most popular songs on Spotify, focusing on genre classification through lyrical analysis and examining the relationship between musical features like **danceability**, **energy**, and **streaming popularity**. Using **Python** and **R**, we cleaned, organized, and analyzed the data to extract meaningful insights. Our analysis revealed that while we expected **danceability** and **energy** to correlate strongly with popularity, the results showed weak **negative correlations** with streaming counts, suggesting that other factors may influence a song’s success more than initially anticipated. Additionally, we discovered that **positive emotions** dominate across genres like **Rap**, **Pop**, and **R&B**, highlighting the importance of emotional resonance in listener engagement. By combining these insights, we can better understand how specific song attributes impact a song's popularity and genre classification.

References

[1] Stacy.Goldrick@groupsjr.com. (2024, April 23). *Spotify Reports First Quarter 2024 Earnings*. Spotify. https://newsroom.spotify.com/2024-04-23/spotify-reports-first-quarter-2024-earnings/

[2] Abdullah, M. (2024). *Spotify Most Streamed Songs*. Kaggle.com. https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs

[3] Spotify. (n.d.). *Web API | Spotify for Developers*. Developer.spotify.com. Retrieved October 12, 2024, from https://developer.spotify.com/documentation/web-api