CFGDegree Data Specialisation
Final Project (Group 7)

Do music preferences change across different English speaking countries?

Group Members:

Esther González-Hernando
Eziamaka Nwakile
Kaban Mohammadi
Man Zhang
Sarah Newlan-Lewis
Wei Lu

Instructors:

Andreea Avramescu
Georgia Whitton

Date: 28/05/2023

1. INTRODUCTION

A study conducted by the International Federation of the Phonographic Industry[1] shows that fans spend an average of 20.1 hours weekly listening to music. This project intends to provide a better understanding of the differences and similarities in music preferences across these five different countries. For that purpose, we used Spotify's data on music and user preferences to examine the top 50 most streamed tracks globally and in five English-speaking countries from different continents: United Kingdom (UK), United States (US), Australia (AUS), India (IN) and South Africa (SA).

1.1 Aim

To examine differences and similarities in music preferences across tracks in Global (Glob), UK, US, AUS, IN and SA top 50 playlists, as well as investigate how audio features may affect tracks' popularity.

1.2 Objectives

- Using Spotify's API, obtain data for tracks in a total of 6 top 50 playlists.
- Compare and contrast the top tracks across different countries and globally.
- Analyse and visualise data on characteristics of the top tracks.
- Identify any trends or patterns in music preferences between countries.
- Explore any relationships between cultural factors and music preferences.

2. BACKGROUND

Spotify is a digital music streaming service that provides users with access to a large library of songs, podcasts, and other audio content. Since its debut in 2008, Spotify has become one of the most popular music streaming services globally. Given that it is possible to obtain data using an API, we considered Spotify the most adequate data source for our analysis.

2.1 Research Questions

a. Are there any differences in the type of music being listened to across different English-speaking countries?
b. How do the audio features of the tracks contribute to the popularity of tracks? Does this change across different contexts?

2.2 Target Audience

This analysis is relevant for anyone interested in gaining insights into music preferences across the UK, US, AUS, IN, and SA and how they compare to global music patterns. However, we have identified some audiences who may particularly benefit from this analysis. These are: *music industry professionals*, including music producers, record label executives and talent scouts; *artists and songwriters*; *streaming platforms curators*; *music researchers* and *sociologists*; and, *music enthusiasts*.

[1] Available at: https://www.ifpi.org/wp-content/uploads/2022/11/Engaging-with-Music-2022_full-report-1.pdf

3. STEPS SPECIFICATION

The following describes the dataset and tools we used for this project.

3.1. Dataset

The dataset for this project was gathered through requests made to Spotify's API[2], which allowed us to access data from the music platform. The final dataset is available on GitHub and contains over 10 features[3], such as danceability, release date, popularity, the genre of the artist, artist name, instrumentalness, valence, acousticness, and loudness.

3.2. Tools and libraries

This project was executed using Python programming language and Jupyter Notebook. Table 1 provides an overview of the used libraries in this project.

Table 1. Tools and Libraries used in this project

Library	Description/Rationale
Numpy	Numpy was used to help us with the numerical computations and manipulation in the data analysis
Pandas	Pandas provided us with data structures and functions to efficiently handle structured data, such as dataframes. It enabled tasks such as loading, data cleaning, performing calculations, grouping, filtering and merging datasets.
Matplotlib	Matplotlib allowed us to create high-quality scatter plots, bar plots, and histograms to visualise our data in a visually appealing and informative way.
WordCloud	WordCloud helped us create word cloud visualisation, that enabled the generation of world clouds where the size of each word is proportional to its frequency in a given text. For the context of this project, it was the frequency of each of the genres.
Seaborn	Seaborn allowed us to create complex visualisations, such as a heatmap, and provided us with additional statistical functionalities for our analysis.
SciPy	SciPy was used for advanced statistical analysis, mathematical modelling and other scientific computations.
Math	Math was used for performing mathematical calculations and transformations in data analysis projects.
Statistics	Statistics were used for basic statistical analysis and summary statistics in the data analysis.
JSON	JSON was used for reading, writing and manipulating the JSON data in the data analysis.

[2] More information available at: https://developer.spotify.com/documentation/web-api/reference/get-audio-features

[3] See https://developer.spotify.com/documentation/web-api/reference/get-several-audio-features for more information on audio features and variables included in the dataset.

4. IMPLEMENTATION AND EXECUTION

The agile-like methodology was used in the implementation of this project[4]. We had regular meetings to check in on our progress and discussed produced code. The following step was undertaken during the data analysis:

4.1. Data gathering

Spotify's API was the primary data source since it granted us access to a range of music-related data. Data was collected between Saturday 20th and Sunday 21st of May. Team members identified and selected appropriate data sources. Further, they utilised various techniques to gather the necessary data we needed for this project.

4.2. Data cleaning and preprocessing

Once all data was gathered, it was converted into a dataframe (DF) for further analysis. The data went through a cleaning and pre-processing phase. Team members checked if each playlist contained 50 tracks; confirmed that there were no duplicate rows in the DF; and, checked for any empty rows that needed to be removed.

4.3. Analysis and Visualisations

A basic exploratory analysis was conducted to understand the shape of the dataset. Visualisations were then obtained to facilitate the identification of patterns, trends and outliers. For example, outliers were removed when visualising the popularity score against the release date (see Figure 3B with outliers removed). This step ensured that the analysis focused on relevant data.

4.4. Roles

Within the project team, each member had specific roles and responsibilities. Here is a breakdown of the main roles:

- Wei Lu: Worked on the writing of the report, ensured the adequate documentation of the project by creating and updating the readme file in GitHub, and worked on the PowerPoint presentation.
- Esther González-Hernando: Created the code to extract data from Spotify's API, worked on the development of the code for the analysis and visualisation of the data, annotated and commented on the Jupyter Notebooks and supported in the writing of the report.
- Eziamaka Nwakile: Worked on the report and scheduled team meetings.
- Kaban Mohammadi: Worked on the report, scheduled team meetings, created to-do lists, hosted the GitHub local repository and was responsible for the submission of homework in week 2.
- Man Zhang: Worked on the development of the code for the analysis and visualisation of the data, supported the development of the PowerPoint presentation and the drafting of the report.
- Sarah Newlan-Lewis: Worked on the PowerPoint presentation, supported the creation of the readme file in GitHub and worked on the writing of the report.

[4] See Project Activity Log for a detailed overview of members' tasks. Available at: https://docs.google.com/spreadsheets/u/0/d/1vpXu43L-flbeNVOGeDX7NWsrNpQI34jQ/edit

5. RESULTS

The analysis of tracks from the Glob, UK, US, AUS, IN and SA playlists revealed the following results.

Comparison of tracks. Tracks from the US, UK, AUS, IN and SA playlists were compared against those in the Glob top 50 playlist (see Figure 1). Results showed that the US, AUS and the UK were the most similar to the Glob playlist, sharing around 50% of the tracks. On the other hand, SA and IN showed a much lower percentage of similarity with the Glob top 50 playlist. The low level of similarity of these two playlists with the top 50 Glob playlist may be due to the variety of languages other than English spoken in these countries.

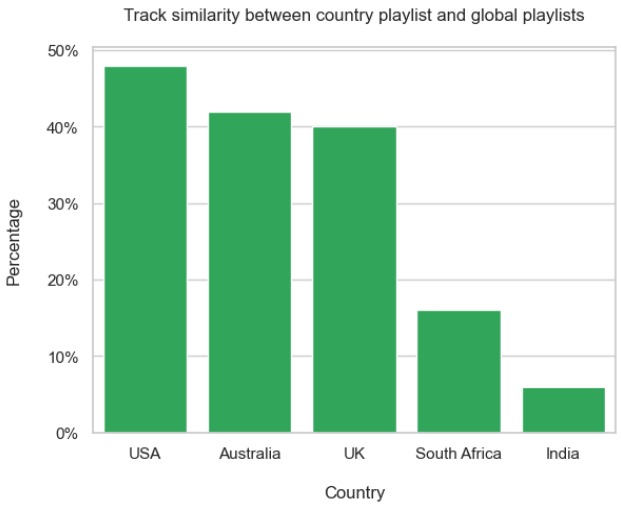
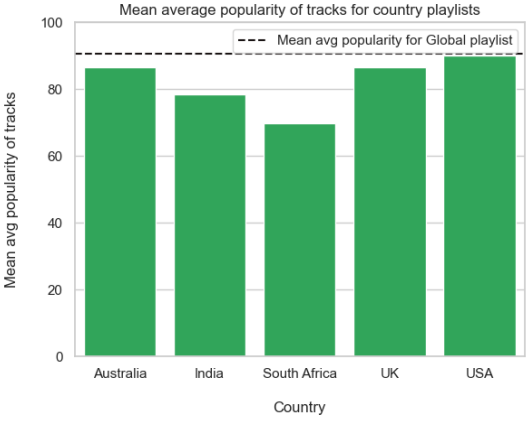


Figure 1. Track similarity between country playlist and global playlists

Results also showed that among the top 5 most popular tracks across playlists, only 'No Role Modelz' and 'Sure Thing' were not featured in the Glob top 50 playlist (see Jupyter notebook). This suggests that these songs might only be popular in predominantly English speaking countries.

Analysis of the tracks' popularity. We identified some variation in the mean average popularity (MAP) of the tracks among the countries' playlists, as well as compared with the Glob top 50 playlist (see Figure 2A). Whereas the MAP of tracks from the US, AUS and UK was very similar to track's MAP of the Glob playlist (e.g., the US top 50 tracks showcased nearly the same MAP as those from the Glob. playlist); the MAP of tracks from SA and IN playlists was considerably lower compared to those of the US, AUS, UK and Glob. top 50 playlists.

A



B

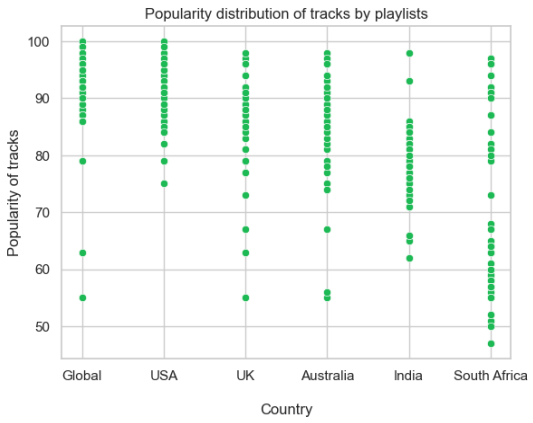


Figure 2. Mean average popularity of tracks for country playlists & Popularity distribution of tracks by playlists

Aligning with these results, the distribution of tracks' popularity grouped by playlists showed how the popularity of tracks from the Glob. and US top 50 playlist are notably high (Figure 2B). However, SA has a very different distribution, showing tracks with relatively low popularity scores. Finally, both in Global, UK and Australia, we can see tracks for which the popularity is notably lower than the rest of the tracks. These points may potentially represent songs that have just made it to the top 50 playlist but for which the popularity has not increased yet.

Exploration of genres and age of the tracks. As can be seen in Figure 3A, the frequency analysis of artist genres across playlists showed 'pop', 'hip-hop' and 'rap' as the most recurring genres in the playlists. Figure 3B showed how most of the tracks in the analysed top 50 playlists have been released between 2020 and 2023.

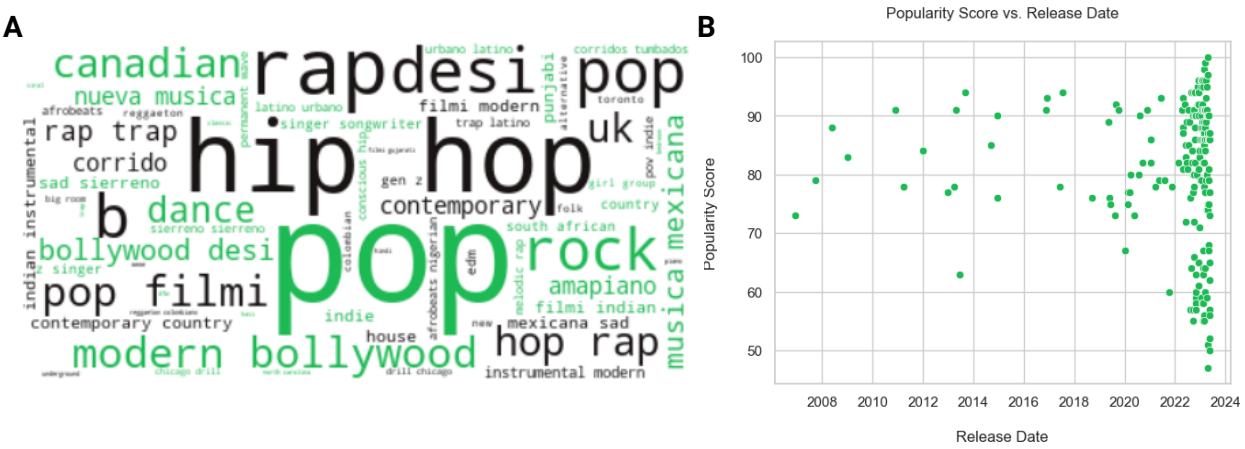


Figure 3: WordCloud and 'Popularity Score vs. Release Date

Analysis of audio attributes. The analysis showed high danceability, energy and valence scores for tracks across all the top 50 playlists (Figure 4). On the other hand, instrumentalness, acousticness, liveness, speachness and liveness attributes were generally low across tracks from all playlists. However, some exceptions to these general trends were found. First, tracks from SA's top 50 playlist showed a noticeable higher danceability mean avg than those of other playlists. This finding suggests that South-African music tends to have a greater emphasis on rhythmic elements, which can make it more suitable for dancing. Further, tracks from IN top 50 playlist showed a considerably higher acousticness mean average than those of other playlists. This finding indicates that Indian music tends to incorporate more organic and unplugged elements, such as acoustic instruments.

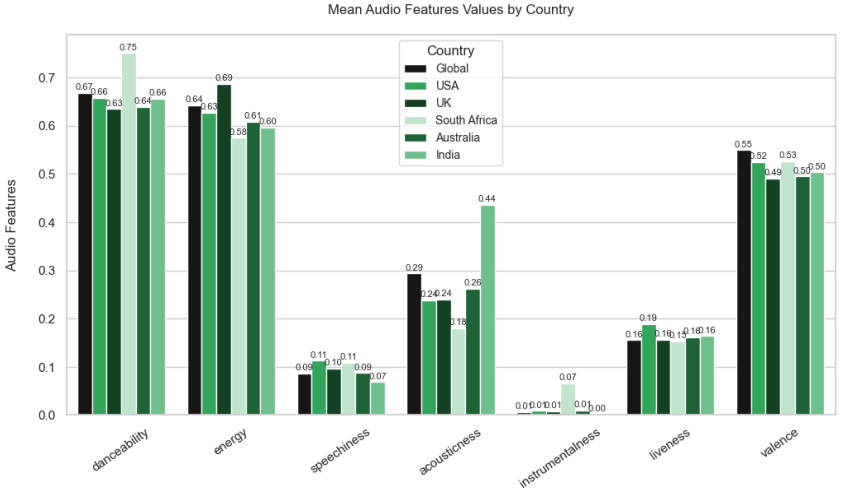


Figure 4 danceability distribution of top 50 tracks in each country

Analysis of potential correlations between audio features and popularity. Considering the results obtained in the analysis of audio attributes and those obtained from a correlation heatmap (see jupyter notebook), we explored further danceability, instrumentalness, acousticness, and valence and their potential correlation with popularity. The analysis showed a skewed distribution for SA between popularity and danceability for the SA playlist (Figure 5A). This finding highlights the significance of instrumental performances in South-African music, as it often integrates more instrumental elements into their music. Moreover, figure 5B also showed particularly high acousticness scores for a great number of the tracks in the IN playlist.

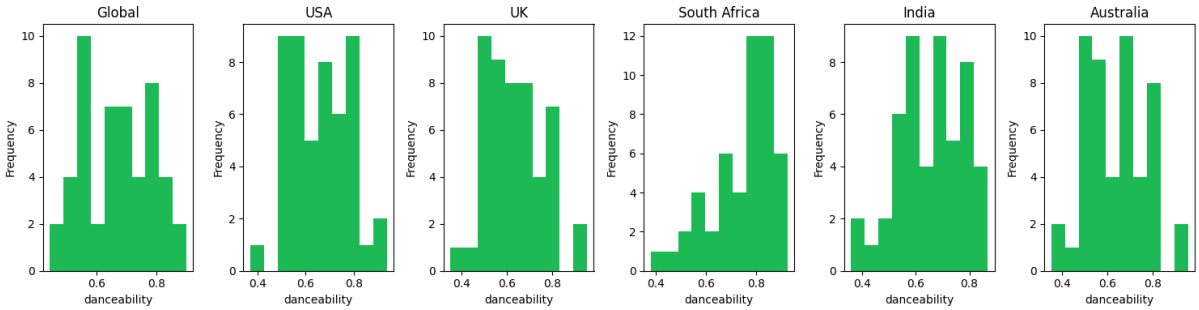


Figure 5A. Frequency - Danceability

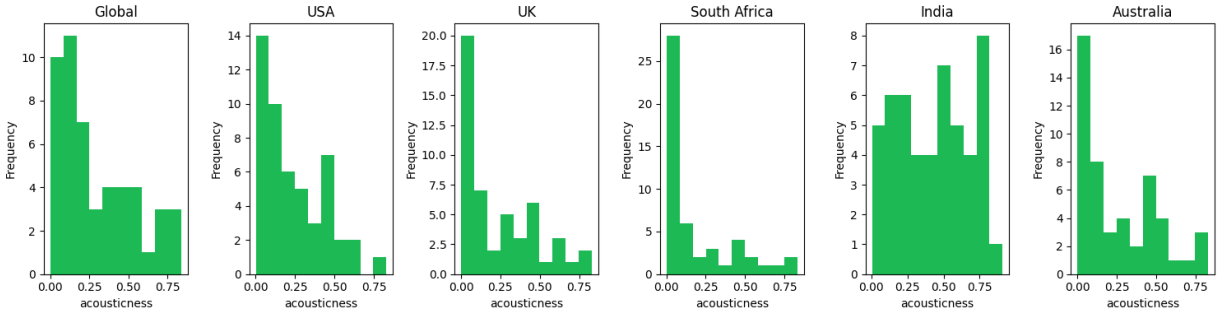


Figure 5B. Frequency - Acousticness

Correlations between selected audio attributes and popularity were further studied using visualisations (see jupyter notebook) and calculating the correlation coefficient between popularity and instrumentalness and popularity and danceability. Visualisations seemed to show a potential correlation between popularity and instrumentalness for IN. Similarly, they pointed towards a potential correlation between popularity and danceability for SA. As can be observed in Table 2, no strong correlations were then found between popularity and instrumentalness nor between popularity and danceability.

Table 2. Calculated correlation coefficients between instrumentalness, danceability and popularity.

Variable 1	Variable 2	Dataset	R coefficient
Popularity	Instrumentalness	All playlists	-0.27
Popularity	Instrumentalness	India	-0.27

Popularity	Danceability	All playlists	-0.22
Popularity	Danceability	South Africa	0.33

Finally, two models were generated using different attributes and inputting popularity as the target value (see jupyter notebook). However, based on the R2 values obtained (both between 0.2 and 0.3), the two models were not very accurate at predicting the popularity of the tracks.

6. PROJECT CHALLENGES

We initially wanted to explore whether TikTok influenced the popularity of songs on Spotify. We soon discovered that TikTok's API could not be used by researchers outside of the US. We then decided to focus the analysis on Spotify since data could be easily accessed through an API.

Additionally, during the extraction of data from Spotify, we encountered difficulties retrieving the music genres of each track since Spotify only provides genre for albums or artists rather than tracks. First, we opted for using album genres since we believed it would better represent the genre of each song. However, we found that we were getting empty lists in our datasets. In the end, we successfully get the genres from the artists.

7. CONCLUSION

In this project, we explored data on the top 50 songs globally and in the UK, US, IN, AUS and SA). Data was gathered using Spotify's API, which allowed us to study differences and similarities between the type of music being listened to across English-speaking countries. Further, it also enabled us to examine the potential influence that audio attributes may have on the popularity of tracks.

Based on the data analysis and visualisation, we observed a high level of similarity between global, UK, US and AUS playlists. Results did not show any strong correlations between danceability, loudness and valence with popularity. Notably, the top 50 songs in SA displayed a higher instrumentalness compared to the other countries. Similarly, acousticness for the top 50 songs in IN far exceeds that of other countries.

Overall, our findings seem to suggest that cultural and context-specific circumstances influence the type of music being listened to in each country. However, based on our data, audio features did not have a strong influence on the popularity of tracks. Our findings are limited and further research using a greater sample of songs and countries may provide further insights into factors that may influence the popularity of songs.