Regional Rhythms: Unveiling Regional Correlations between Audio Features and Song Popularity

Oğuz Ata Çal [*1] **Bora Kargı** [*2] **Karahan Sarıtaş** [*3] **Kıvanç Tezören** [*4]

Abstract

Assessing the impact of audio signals on song popularity constitutes a significant area of inquiry within music studies. Despite the considerable research in this field, the question remains unresolved regarding whether distinct regions demonstrate divergent music preferences based on song audio characteristics. Our study delves into the correlation between audio features and song popularity across distinct regions. Leveraging the Spotify Charts dataset (2017-2021) and audio features from the Spotify Web API, we introduce a popularity metric combining stream proportion and chart rank. Our findings reveal weak correlations between certain audio features and regional preferences. Using an unsupervised clustering technique, we identify groups of regions with similar musical tastes based on these correlations.

1. Introduction

Recent information on global music industry highlights the prominence of digital music in the marketplace. By 2022, streaming, encompassing both subscription and advertising-supported models, claimed the largest market share, surging to 67.0% of the total revenue of US $26.2 billion, from the previous year's 65.5% (IFPI, 2023). Spotify dominates this vast market, experiencing a substantial surge in premium user numbers, from 71 million subscribers in 2017 to 205 million by 2022 (Spotify, 2023).

Spotify's database, due to its unprecedented prominence in the music industry, holds great significance for audio analysis. This paper aims to initiate exploration by analyzing highly streamed songs in each region, focusing on properties like valence and energy levels. The central hypothesis guiding our work is that distinct audio features have varying impacts on music preferences across different regions.

There is research suggesting that audio signals can be used to predict the popularity of songs (Lee & Lee, 2018b). Further studies show that specific audio features, including loudness and duration, exhibit correlations with musical trends (Ni et al., 2011). Several methods are proposed to predict whether a song will be a hit based on audio features, including machine learning (Herremans et al., 2014). Additionally, there are studies on identifying specific patterns in regional music selection based on predefined genres. One particular study delves into country-specific music preferences and develops distinct machine learning models for music recommendation systems (Schedl, 2017). Nevertheless, there seems to be a research gap concerning the impact of audio features in popular songs on diverse regions across the world.

The contributions in this paper are as follows:

- We introduce a reasonable popularity metric (Section 2) that considers both number of streams and a specific period to evaluate regional music preferences.
- We find correlations between audio features and regional music preference using our popularity metric and report results of Spearman's correlation coefficient test (Section 3) with associated p-values.
- We cluster the regions into different groups based on measured feature correlations using Affinity Propagation (Frey & Dueck, 2007) (Section 3).
- The country clusters derived from our analysis are compared with existing literature (Section 4), highlighting similarities. Additionally, we underscore potential avenues for further research in Section 4.

2. Data and Methods

We analyze two large-scale datasets from Spotify, Top 200 daily charts spanning from January 1, 2017, to December 31, 2021, and audio features for each track. Charts are retrieved from the Kaggle dataset (Dave, 2021), and audio features for each track are collected using the Spotify Web API (Spotify, 2023a).

[*]Equal contribution [1]Matrikelnummer 6661014, oguzata.cal@student.uni-tuebingen.de, M.Sc. Machine Learning [2]Matrikelnummer 6673983, bora.kargi@student.uni-tuebingen.de, M.Sc. Machine Learning [3]Matrikelnummer 6661689, karahan.saritas@student.uni-tuebingen.de, M.Sc. Machine Learning [4]Matrikelnummer 6622978, kivanc.tezoren@student.uni-tuebingen.de, M.Sc. Machine Learning.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

The Spotify Charts dataset comprises the title of the tracks, daily rank (either within the Top 200 or Viral 50), date, artist, URL of the track on Spotify, region, chart (Top 200 or Viral 50), trend compared to the records of the previous day (same position, new entry, move up or down), and the number of streams. While the selection of the Top 200 is solely based on the number of daily streams, there are multiple factors affecting the Viral 50 trend list, including the number of shares and the most recent rise in plays. To focus on the popularity of the songs based on streams, we decided to consider only the Top 200 songs for our analysis. Streams are counted when someone listens to the song for 30 seconds or more (Spotify, 2023b). The daily stream count is the cumulative tally of these events throughout the entire day.

The Top 200 Charts dataset includes 91,219 unique songs, 40,176 artists, and 69 unique regions along with the global list. There are different number of data points for different regions, as shown by Figure 1.

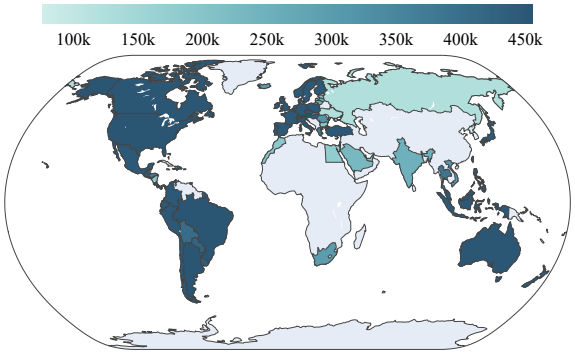

Figure 1: Map shows the number of recorded dates by country, with darker shades indicating more data points and gray indicating no data.

During the preprocessing, a total of 13 tracks were removed from the Charts dataset as no audio features were available from the Spotify API.

Retrieved properties of the tracks include the duration of the track in milliseconds, key (integers map to pitches using standard Pitch Class notation), modality (major or minor), the overall estimated tempo of a track in beats per minute, time signature (number of beats in each bar), and the overall loudness of a track in decibels (dB). We also retrieve the following audio features, all of which take values between 0 and 1: acousticness (inversely proportional to the presence of vocals), danceability, energy, instrumentalness, liveness (presence of an audience), speechiness (presence of spoken words), and valence (musical positiveness) (Spotify, 2023c).

The two datasets are merged to analyze the relationship between popularity and audio features, along with the similarities of musical tastes between countries to discover underlying geographical or cultural links. Links to both the collected audio features and the Spotify Charts dataset are also publicly available on our GitHub repository [1].

In the analysis of the Spotify Top 200 charts, track rankings are ascertained by the volume of streams they accumulate. The ranking of a track t within a specific region r during the time interval (d_s, d_e) is denoted as $rank(t, r, d_s, d_e)$. This ranking is computed by aggregating the streaming data from the Top 200 charts for the designated periods, followed by re-ranking the tracks based on the aggregated amount of streams within the time interval.

To ensure robustness, we utilize *stream proportions* rather than absolute stream counts to gauge *popularity*, accounting for fluctuations in the number of streams influenced by both temporal and regional differences. Stream proportions compare a track's stream count to others within the same chart, mitigating bias from varying time periods and allowing for a fair assessment of popularity amid streaming data variability.

We define our popularity metric for a track t, as *the total amount of stream proportion* within the period (d_s, d_e) in the Top 200 chart of the given region r. This can be formally defined as:

$$p(t, r, d_s, d_e) = \sum_{d=d_s}^{d_e} s(t, r, d)$$

where $s(t, r, d)$ gives us the stream proportion of the track t on the date d for the region r. In our analysis, we use $d_s =$ 2017-01-01 and $d_e =$ 2022-01-01 for all tracks and regions to include all of the data points. The popularity metric we use strongly correlates with a track's total rank, reflecting its overall position in the combined chart computed as *rank*(*t*, *r*, 2017-01-01, 2022-01-01).

3. Results

Based on the formula outlined in Section 2, a popularity metric is computed for each track. Our analysis reveals a notable negative correlation of $\rho = -0.975$ between popularity and total rank, with a significance level of $p < .001$, affirming its appropriateness for our study. This negative correlation stems from the phenomenon where increasing popularity (reflected in higher streams or stream proportion) leads to a decrease in chart rank, resulting in a lower numerical rank assignment.

The correlation between audio features and popularity was evaluated for each country utilizing the SciPy implementation of Spearman's rank correlation method (Virtanen et al., 2020), which conducts a two-sided test, assuming a null

[1] https://github.com/kargibora/DataLiteracy-Regional-Rhythms

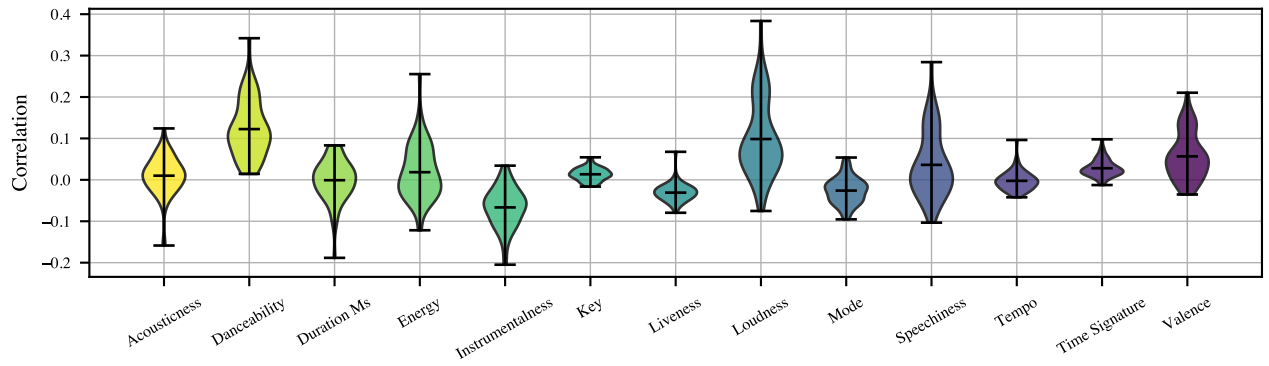

Figure 2: Distribution of the Spearman's correlations between popularity and audio features.

hypothesis of no correlation between features and musical preference.

Figure 2, presents the distribution of these correlations compared to the audio features where the ticks on the plot showcase the maximum, mean, and minimum correlation values for each audio feature.

Figure 3 depicts a scatter plot illustrating these relationships, with the vertical red line representing the $\alpha = 0.05$ cut-off and the horizontal red line indicating the weak correlation limit $|\rho| \geq 0.10$ (Schober et al., 2018).

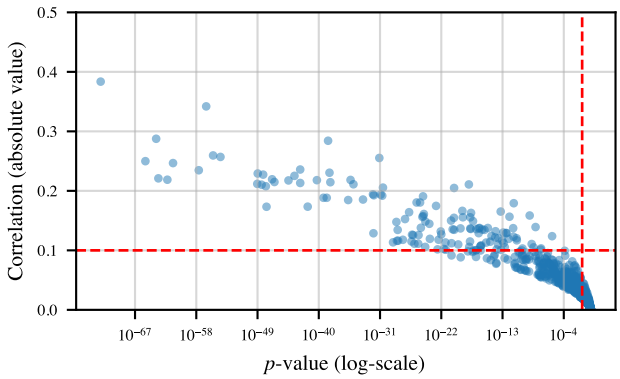

Figure 3: Spearman's correlation vs. p-value plot.

The scatter plot in Figure 4 contrasts loudness and popularity metrics among countries with the lowest (left panel) and highest (right panel) correlation magnitudes. The relationship is depicted through isotonic regression, represented by the red line, which serves as the optimal non-decreasing function minimizing the mean squared error within the dataset. The subplot beneath each scatter plot illustrates the density of loudness values.

We use Affinity Propagation, an unsupervised clustering technique, to identify different groups of countries according to the correlations. This method has been utilized in previous music studies, including those referenced in (Schedl et al., 2017) and (Bauer & Schedl, 2019) as well. Figure 5 illustrates the outcome of applying Affinity Propagation to

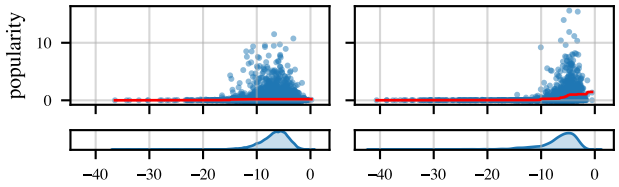

Figure 4: Popularity vs. loudness plots for songs in Taiwan (left) and Nicaragua (right), with density plots for loudness below.

the correlation vectors with a damping effect of 0.5 and a maximum of 200 iterations.

The radar plot in Figure 6 depicts regional song popularity correlations with audio features for countries grouped by affinity propagation clusters.

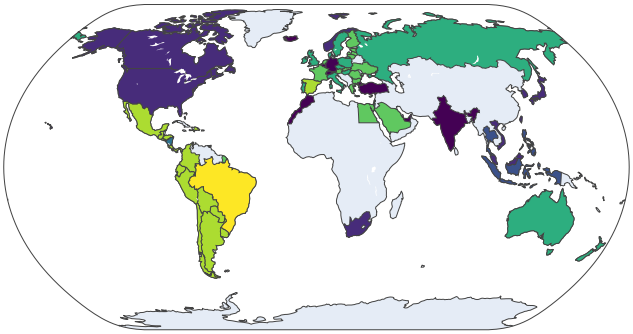

Figure 5: Clustering of the countries based on correlation vectors.

4. Discussion & Conclusion

The correlation analysis presented in Figure 2 reveals relationships between audio features and regional music preferences, corroborating our initial hypotheses concerning the direction of these associations. Musical preferences vary significantly across countries, as there is no universal audio feature that consistently correlates with popularity across all regions. Nonetheless, a majority of the countries ex-

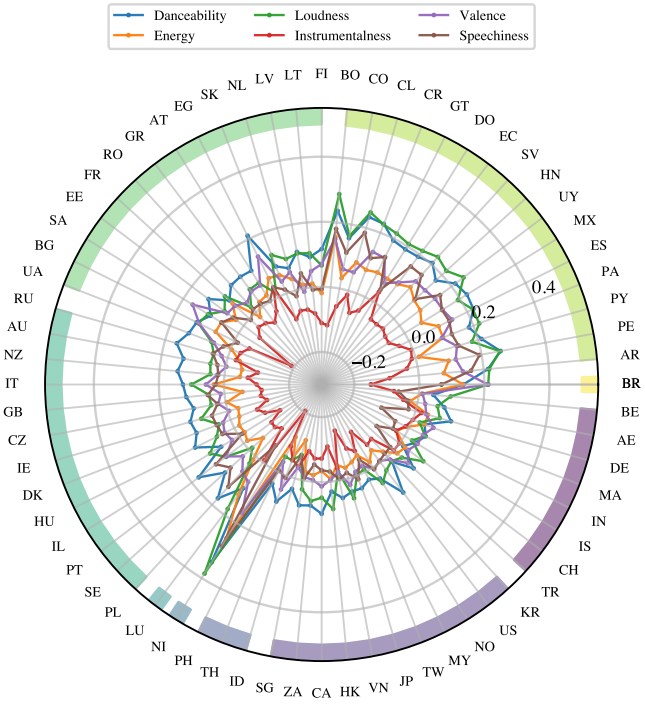

Figure 6: Radar plot of correlations and clustered regions.

hibit a weak-to-moderate correlation between popularity and certain audio features where correlations within the range $0.1 \leq \rho \leq 0.39$ are considered weak (Schober et al., 2018). Attributes such as *acousticness*, *danceability*, *duration*, *energy*, *instrumentalness*, *loudness*, *speechiness*, and *valence* exhibit weak-to-moderate associations with popularity, while *key*, *mode*, *tempo*, *liveness*, and *time signature* demonstrate minimal impact, similar to the findings of (Colley et al., 2022), who explored the correlation between audio signals and global Spotify popularity. Additionally, *danceability* emerges as a positively correlated audio feature for popularity, a trend also supported by the research of (Interiano et al., 2018).

Figure 3 specifically highlights those correlations with magnitudes $|\rho| \geq 0.10$, all of which are associated with p-values less than 0.05. This visualization effectively underscores that the weak correlations identified in our analysis bear statistical significance, affirming the reliability of our findings regarding the impact of various audio features on track popularity.

Figure 4 illustrates Nicaragua as having the highest observed correlation with popularity, exhibiting a steeper best-fit monotonic function compared to the country with the lowest correlation, Taiwan. This contrast highlights the impact of loudness on track popularity across countries with differing correlation degrees.

Based on the correlation vectors of each region, we obtain a set of groups where each group exhibits similar musical preferences, as shown in Figure 5. A notable discovery reveals that Spanish-speaking countries (except Nicaragua) constitute a distinct cluster separate from Portuguese-speaking Brazil, consistent with prior clustering studies such as those by (Schedl et al., 2017) on genre profiles and (Bauer & Schedl, 2019) on artist listener counts.

The radar plot in Figure 6 provides further insight into the similarity of audio feature correlations among countries grouped in the same cluster. Notably, the cluster comprising Spanish-speaking countries, located at the top right of the radar plot, exhibits prominently higher correlations in terms of loudness, danceability, valence, and energy.

Finally, it should be recognized that the songs listened to by Spotify users may not directly reflect the musical preferences of a given region. However, we believe our findings reflect the actual music preferences to a significant degree given the substantial amount of data and Spotify's popularity. Our notebooks are publicly accessible in a repository to facilitate further research and application, allowing for the reproduction of analysis results. We also have included a supplementary folder[2] that contains additional plots and visualizations.

For future research, we can leverage correlation features extracted from regional data to develop hit song prediction models tailored to specific regions, akin to the approach taken by (Herremans et al., 2014). This would allow us to account for regional tastes and preferences more effectively. Furthermore, other conventional acoustic features such as MPEG-7 and Mel-frequency cepstral coefficient (MFCC), as pointed out by (Lee & Lee, 2018a), could be applied to regional data to enhance the depth of the correlation analysis.

Contribution Statement

Our team collectively determined the methods for processing audio feature data and creating charts, establishing correlations with regions through joint discussions. The workload distribution was as follows: Bora Kargı managed data retrieval from the Spotify API, developed a track popularity metric, and investigated country clustering based on audio feature correlations. Oğuz Ata Çal assessed similarity in countries' preferences using their top charts. Karahan Sarıtaş performed exploratory analyses on streams and audio features, utilizing cosine similarity, t-SNE, and PCA techniques based on top charts, and conducted a literature review. Kıvanç Tezören handled data visualization, conducted a literature review on global music preferences, and set up a reproducible repository environment.

[2] https://github.com/kargibora/DataLiteracy-Regional-Rhythms/tree/main/figures

References

Bauer, C. and Schedl, M. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE*, 14(6):1–36, 06 2019. doi: 10.1371/journal.pone.0217389. URL https://doi.org/10.1371/journal.pone.0217389.

Colley, L., Dybka, A., Gauthier, A., Laboissonniere, J., Mougeot, A., Mowla, N., Dick, K., Khalil, H., and Wainer, G. Elucidation of the relationship between a song's spotify descriptive metrics and its popularity on various platforms. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 241–249, 2022. doi: 10.1109/COMPSAC54236.2022.00042.

Dave, D. Spotify charts, 2021. Retrieved from: https://www.kaggle.com/ds/1265407, Accessed on November 20, 2023.

Frey, B. J. and Dueck, D. Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315(5814): 972–976, 2007. doi: 10.1126/science.1136800.

Herremans, D., Martens, D., and Sörensen, K. Dance hit song prediction. *Journal of New Music Research*, 43(3):291–302, July 2014. ISSN 1744-5027. doi: 10.1080/09298215.2014.881888. URL http://dx.doi.org/10.1080/09298215.2014.881888.

IFPI. Global Music Report, 2023. Retrieved from: https://www.ifpi.org/wp-content/uploads/2020/03/Global_Music_Report_2023_State_of_the_Industry.pdf, Accessed on December 6, 2023.

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., and Komarova, N. L. Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science*, 5(1):171274, 2018. doi: 10.1098/rsos.171274. URL http://doi.org/10.1098/rsos.171274.

Lee, J. and Lee, J.-S. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11):3173–3182, November 2018a. ISSN 1941-0077. doi: 10.1109/tmm.2018.2820903. URL http://dx.doi.org/10.1109/TMM.2018.2820903.

Lee, J. and Lee, J.-S. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11):3173–3182, 2018b. doi: 10.1109/TMM.2018.2820903.

Ni, Y., Santos-Rodríguez, R., McVicar, M., and Bie, T. D. Hit song science once again a science? 2011. URL https://api.semanticscholar.org/CorpusID:17320348.

Schedl, M. Investigating country-specific music preferences and music recommendation algorithms with the lfm-1b dataset. *International Journal of Multimedia Information Retrieval*, 6, 03 2017. doi: 10.1007/s13735-017-0118-y.

Schedl, M., Lemmerich, F., Ferwerda, B., Skowron, M., and Knees, P. Indicators of country similarity in terms of music taste, cultural, and socio-economic factors. In *2017 IEEE International Symposium on Multimedia (ISM)*, pp. 308–311, 2017. doi: 10.1109/ISM.2017.55.

Schober, P., Boer, C., and Schwarte, L. A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth Analg*, 126(5):1763–1768, May 2018.

Spotify. Spotify Web API, 2023a. Retrieved from: https://developer.spotify.com/documentation/web-api/, Accessed on November 20, 2023.

Spotify. Spotify FAQ, 2023b. Retrieved from: https://artists.spotify.com/faq/stats, Accessed on December 19, 2023.

Spotify. Spotify for Developers, 2023c. Retrieved from: https://developer.spotify.com/documentation/web-api/reference/get-audio-features, Accessed on December 19, 2023.

Spotify. Shareholder Deck Q3 2023, 2023. Retrieved from: https://s29.q4cdn.com/175625835/files/doc_financials/2023/q3/Shareholder-Deck-Q3-2023-FINAL.pdf, Accessed on December 12, 2023.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.