CSE 564: Final Project Report

By Amirhossein Najafizadeh & Iliya Mirzaei

Spring 2025

Project Title

Mood in the Music: A Data-Driven Analysis of How Listening Habits Reflect Mental Health

Code Base

GitHub Repository: Code
YouTube Video: Demo

Background

Music is a universal medium for emotional expression and psychological reflection. Across cultures, people rely on music not only for entertainment, but also as a tool for emotional regulation and mental well-being. Recent interdisciplinary research, covering psychology, neuroscience, musicology, and data science, has increasingly explored this connection, investigating how music affects mood and supports mental health interventions [1].

Empirical studies suggest that music listening can alleviate symptoms of anxiety, depression, and insomnia, and support emotional processing in clinical settings through music therapy [2, 3]. Furthermore, individuals often gravitate toward particular genres, tempos, or lyrical content based on their psychological state, highlighting the bidirectional relationship between music preference and mental health.

Spotify and Apple Music are two of the most popular music streaming apps that people use to listen to songs every day. Because they are widely used around the world, these platforms offer a huge and diverse collection of songs from many different genres — like pop, rock, jazz, classical, and more. What makes these platforms especially useful for our analysis is that they also provide detailed attributes for each song.

This project aims to uncover patterns and correlations between self-reported mental health conditions and the characteristics that people prefer to listen to music. By analyzing large-scale datasets of music preferences, audio features, and mental health survey responses, this work seeks to identify interpretable trends that may inform mental health assessment tools, therapeutic music selection, and personalized wellness recommendations.

Original Datasets and Feature Selection

This project integrates multiple datasets related to music, genre classification, listener behavior, and song-level attributes. The datasets serve complementary purposes and vary in size and structure. Table ?? provides an overview of each dataset, including its intended use, number of records (items), total features, and the number of features selected for analysis.

Name	Type	Usage	Entities	Features
apple_music_dataset [8]	CSV	Songs to Genres	10,000	24
mxmh_survey_results [4]	CSV	Genres to Mental Health	1,000	34
spotify_2000_tops [9]	CSV	Songs to Attributes	2,000	15
spotify_music_dataset [7]	CSV	Songs to Attributes	2,000	18
spotify_song_attributes [6]	CSV	Genres to Attributes	10,080	22
universal_top_spotify_songs [5]	CSV	Country to Listening Behavior	22,000	24

Table 1: Overview of the Original Datasets

Feature Selection and Filtering

In the first step of the preprocessing pipeline, relevant features were selected from each dataset to reduce noise and ensure alignment with the analytical objectives. The filtering process involved selecting key columns and saving the reduced datasets to a temporary directory (`tmp/`). Table ?? provides the details of this filtering.
This filtering step significantly reduced the dimensionality of the original data, preparing the datasets for the downstream steps of transformation, integration, and analysis.

Dataset	Retained	Columns
apple_music_dataset	2	'trackCensoredName', 'primaryGenreName'
spotify_2000_tops	11	'Title', 'Top Genre', 'Beats Per Minute (BPM)', 'Energy', 'Danceability', 'Loudness (dB)', 'Speechiness', 'Acousticness', 'Liveness', 'Valence', 'Popularity'
spotify_music_dataset	13	'song', 'genre', 'popularity', 'danceability', 'energy', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo'
spotify_song_attributes	12	'trackName', 'genre', 'danceability', 'energy', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo'
mxmh_survey_results	11	'Age', 'Hours per day', 'While working', 'Fav genre', 'Exploratory', 'Foreign languages', 'Anxiety', 'Depression', 'Insomnia', 'OCD', 'Music effects'
universal_top_spotify_songs	5	'country', 'tempo', 'danceability', 'energy', 'valence'

Table 2: Selected Features from Original Datasets

Data Wrangling

Decoupling

Several datasets in our project contain a `genre` field in which individual entries are associated with multiple genres. These multi-value genre features are often represented as comma-separated strings within a single cell, such as:

```
Rock, Alternative, Indie
```

To enhance the accuracy and granularity of our analysis, we performed a decoupling step to normalize these values. Each genre was extracted and assigned to its own row, effectively transforming the dataset into a one-to-many relationship between songs and genres. This approach enables a more detailed understanding of genre-specific patterns and ensures compatibility with analytical models that expect one genre per instance.
The following table summarizes the dataset sizes before and after genre expansion:

Dataset	Before Expansion	After Expansion
Spotify Dataset	2,000	3,704
Apple Music Dataset	10,000	11,629

Table 3: Dataset Size Before and After Genre Field Expansion

This preprocessing step was crucial in allowing genre-based clustering, statistical aggregation, and visualization techniques to function correctly, without being misled by overlapping or conflated genre labels.

Normalization

Following the decoupling step, we applied normalization to all numerical features across our datasets. This process was essential to ensure that features with varying scales could be meaningfully compared and analyzed. Several numerical attributes, such as `loudness`, had values ranging approximately between -40 and 40, while others, like `danceability`, ranged between 0 and 2. Without normalization, features with larger numeric ranges would dominate distance-based metrics and skew the results of clustering or dimensionality reduction techniques. To address this imbalance, we normalized each numerical feature to a common scale, typically using Min-Max normalization:

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This transformation maps each value to a range between 0 and 1, allowing for consistent comparison across features. By normalizing the data, we established a consistent foundation for further statistical analysis, clustering, and machine learning tasks within the visualization dashboard.

Fixing Inconsistencies

In this section, we highlight two major inconsistencies present in our datasets. These issues lead to significant duplication and inaccurate information.

Before merging our datasets, we noticed that genres were written differently across them. For example, one dataset had `Rock` and another had `Rock and Roll`. To fix this, we used natural language processing and cosine similarity to compare genre names. This helped us group similar genres under one common name.

The `mxmh_survey_results` dataset does not include a location field, which limits the ability to analyze trends geographically. To resolve this, we leveraged the `universal_top_spotify_songs` dataset, which includes listening trends in 73 countries. We took `tempo`, `danceability`, `energy`, and `valence` for each country. Then, for each respondent in the mental health dataset, we used a regression model to find each country's feature using their BPM (and imputed values for other audio features). The country with the smallest distance was assigned to that respondent. This process effectively mapped user behavior to the closest global listening profile.

This geographic inference process enables the integration of country-level visualizations such as choropleth maps and regional comparisons, enriching the storytelling capacity of the final dashboard.

Now you can view each dataset after we fixed the inconsistencies. For every dataset, we counted how many rows were changed, then calculated the linguistic and regression accuracy. If there were any, we also showed the top 5 most common linguistic changes and how many times each one appeared.

apple_music_dataset

- Total changes: 7,674
- Average best similarity score: 0.6813
- Top 5 genres changed in this dataset following by the number of changes:
  - Hip-Hop: 1,038
  - Pop: 2,223
  - Alternative: 1,111
  - Soundtrack: 688
  - Dance: 157

spotify_2000_tops

- Total changes: 1,994
- Average best similarity score: 0.6782
- Top 5 genres changed in this dataset following by the number of changes:
  - adult standards: 123
  - album rock: 413
  - alternative hip hop: 2
  - alternative metal: 70
  - classic rock: 51

spotify_music_dataset

- Total changes: 3,704
- Average best similarity score: 0.7611
- Top 5 genres changed in this dataset following by the number of changes:
  - pop: 936
  - rock: 162
  - Pop: 697
  - hip hop: 776
  - R&B: 439

spotify_song_attributes

- Total changes: 8,580
- Average best similarity score: 0.4707
- Top 5 genres changed in this dataset following by the number of changes:
  - alt z: 656
  - pop: 602
  - dance pop: 172
  - alternative metal: 150
  - singer-songwriter pop: 164

universal_top_spotify_songs

- Total records analyzed: 22,000+
- Unique Countries in the result: 38
- Average features used for matching: tempo, danceability, energy, valence

Thanks to our language-based method for fixing inconsistencies, we were able to easily find and remove duplicates. Also, by getting rid of outliers and inconsistent entries, the results of our merging in the next step became more accurate.

Data Merging

In this step, we combined all the datasets into one complete dataset to prepare the data for analysis. We used song names to match songs with their attributes, and then used genres to bring together all songs, attributes, and mental health data. Additionally, to enable geographical visualizations, we used the listening behavior dataset from Spotify to infer countries based on musical feature similarity.

- **Step 1: Merging Songs and Genres**
  - Songs in Apple Music dataset: 11,629
  - Songs in Spotify dataset: 3,704
  - Songs after merging: 3,488
- **Step 2: Merging Songs with Their Attributes**
  - Songs in merged dataset: 3,488
  - Songs in Spotify 2000 dataset: 1,994
  - Songs in Spotify attributes dataset: 10,080
  - Songs after merging: 10,522
- **Step 3: Merging with Mental Health Data (Using Genre)**
  - Songs with attributes: 10,522
  - Mental health records: 1,000
  - Final merged records: 520,778
- **Step 4: Country Inference via Audio Similarity**
  - Used Spotify's country-level top song data
  - Extracted average tempo, danceability, energy, and valence per country
  - Matched each respondent to closest country using Euclidean distance based on BPM and placeholder audio features
  - Result: Mental health records enriched with `country` column

In the end, our final dataset includes **23 features** and **520,778 records**. This rich dataset allows us to explore the connection between music attributes and mental health issues.

Filling Missing Values

After creating the final dataset, some values were missing because of the merging process. We cleaned the data by removing duplicate rows and dropping any rows that were completely empty. To fill in the remaining missing values, we used linear regression to make educated guesses based on the patterns in the data.

Sampling

In the final step of our data processing, we aimed to reduce the dataset size from over half a million entities to 2000 data items. To achieve this, we used a clustering method. First, we applied Principal Component Analysis (PCA) to select the best features for clustering. Then, we used the K-means algorithm along with the knee finding method to group the data into 10 clusters. Finally, we selected around 2000 samples based on the size of each cluster to create our final dataset. In the next figures you can see the results of our clustering based on PCA1 and PCA2.

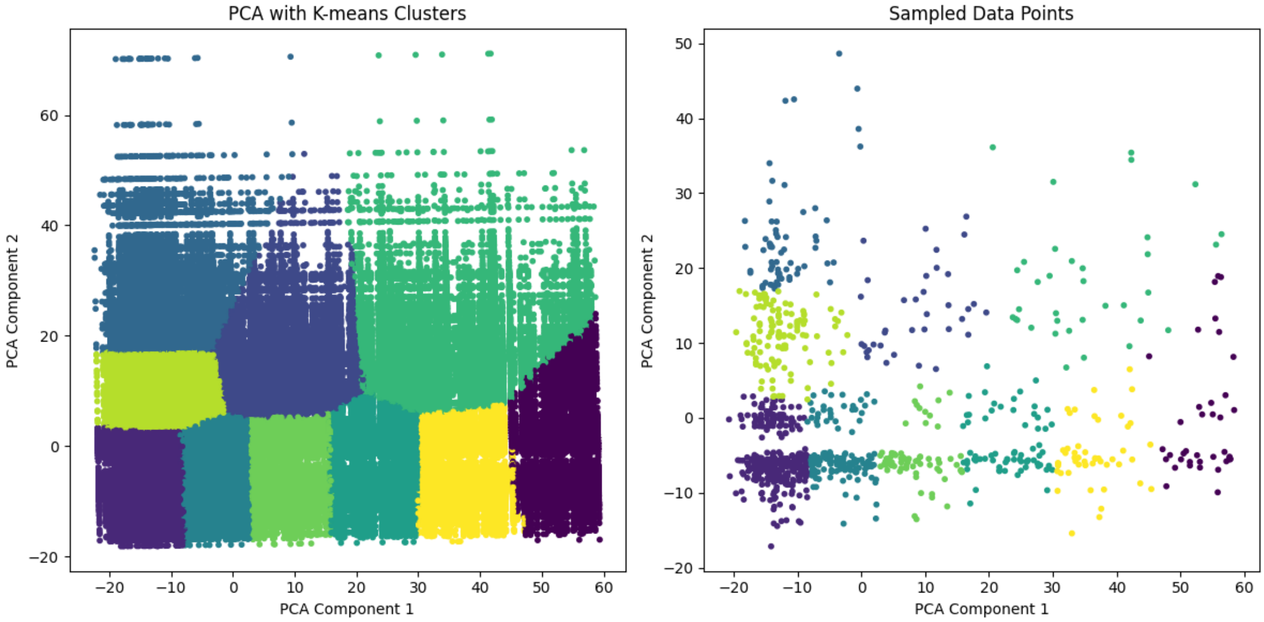


Figure 1: PCA results with 1000 samples

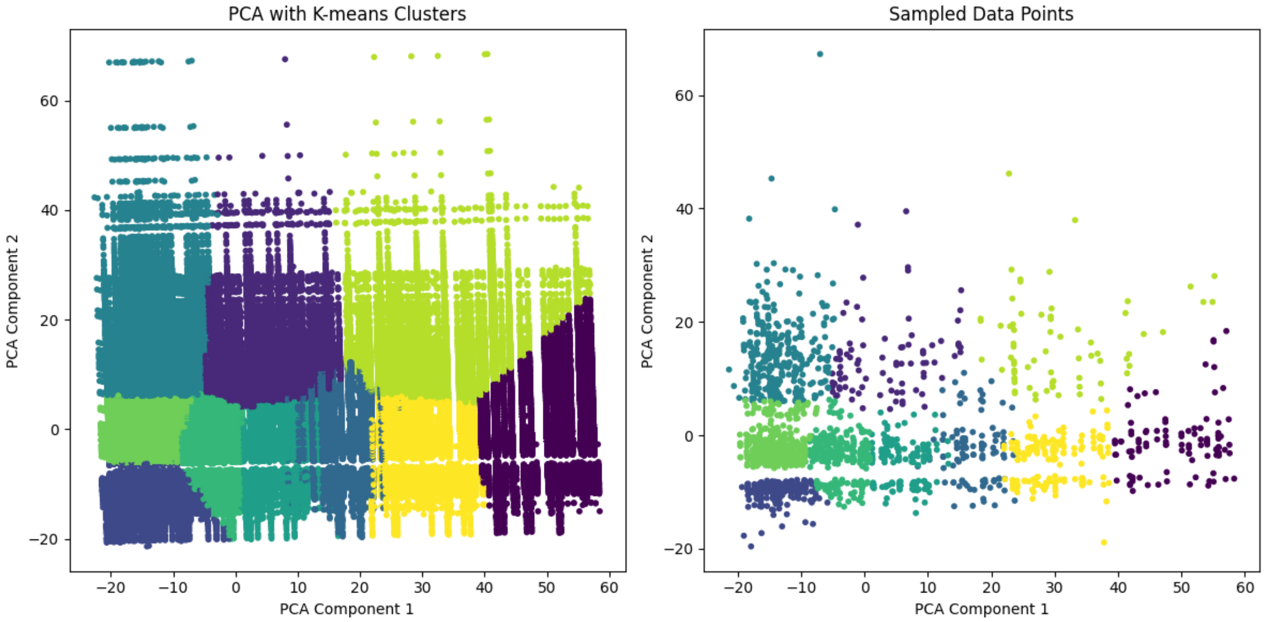


Figure 2: PCA results with normalized data and 2000 samples

Dataset

The final dataset consists of 23 columns and 1995 rows. The columns contain various features related to music attributes and mental health data. Below are the unique values for each column:

Column Name	Number of Unique Values
acousticness	99
bpm	61
danceability	85
energy	98
genre	11
instrumentalness	63
liveness	90
loudness	651
mode	2
popularity	84
speechiness	63
tempo	68
valence	96
age	70
hours	140
while working	2
exploratory	2
foreign	2
anxiety	73
depression	72
insomnia	69
ocd	60
effects	3
country	38

Table 4: Unique Values in Each Column

The following are the top 5 rows of the dataset, showing a snapshot of the music and mental health attributes:

acousticness	bpm	danceability	energy	genre	instrumentalness	liveness	exploratory	foreign	anxiety	depression	insomnia	ocd	effects	country
0.99	0.65	0.40	0.18	K pop	0.00	0.09	Yes	No	0.45	0.74	0.97	0.17	Improve	PK
0.00	0.65	0.51	0.89	Rock	0.01	0.23	Yes	No	0.19	0.56	0.59	0.07	Improve	IT
0.07	0.65	0.82	0.72	Rock	0.16	0.07	Yes	No	0.49	0.63	0.74	0.30	Improve	PK
0.71	0.67	0.61	0.45	Rock	0.00	0.13	Yes	No	0.90	0.80	0.94	0.37	Improve	CL
0.57	0.65	0.24	0.54	Hip hop	0.37	0.09	No	Yes	0.71	0.42	0.17	0.18	No effect	UY

Table 5: Top 5 Rows of the Dataset

Methods

To build our dataset, we utilized Python and shell scripts. For dataset reading and manipulation, we leveraged the `Pandas` library. For linguistic processing, we used the following imports from the `scikit-learn` library:

- `from sklearn.feature_extraction.text import TfidfVectorizer`
- `from sklearn.metrics.pairwise import cosine_similarity`

For Principal Component Analysis (PCA) and clustering, we used the following modules from `scikit-learn`:

- `from sklearn.decomposition import PCA`
- `from sklearn.cluster import KMeans`

To build our dashboard, we used Python Flask for the backend to handle data processing and API endpoints. On the frontend, we used D3.js along with HTML templates to create interactive visualizations and display the results in a user-friendly way.

We implemented the following D3.js visualizations:

- **Correlogram**: The correlogram helps us identify patterns by showing how different features in the dataset relate to one another. It visualizes correlation values as a grid of colored cells, where the color intensity and direction (positive or negative) indicate the strength and type of relationship. For example, it reveals how certain musical attributes like energy or tempo may be positively or negatively correlated with specific mental health issues such as anxiety or depression.
- **Choropleth Map**: This map is used to present geographic data by shading countries based on a specific metric. In our case, each country is colored based on the number of available records, giving an overview of data distribution across regions. We also integrated brushing and linking functionality—when a country is selected, it dynamically updates other charts on the dashboard to reflect only data from that region, enabling localized analysis.
- **Parallel Coordinates Plot (PCP)**: The PCP is used to analyze how multiple variables behave across different records. Each line represents a data point (such as a song or survey entry), and each axis corresponds to a different feature. This plot is especially useful for spotting trends, clusters, and outliers across high-dimensional data. For example, we can examine how different levels of mental health conditions align with musical preferences or listening behavior patterns.
- **Stacked Bar Chart**: This chart compares multiple categories by stacking segments of data within each bar. We used it to compare how different mental health conditions (like stress, depression, and anxiety) are distributed across various music genres. Each mental health issue is categorized into three levels—low, medium, and high—so we can observe genre preference trends relative to the severity of the condition.
- **t-SNE Plot**: t-SNE (t-distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique used to visualize high-dimensional data in a two-dimensional space. This plot helps us identify clusters and neighborhood structures among records. In our project, we used t-SNE to project the dataset and highlight how songs or individuals group based on similarities in their mental health status and associated musical features.
- **Word Cloud**: The word cloud displays the most common music genres in a visually intuitive way, with more frequent genres appearing larger. It is used to compare genre distribution across different countries or mental health categories. This visualization makes it easy to spot dominant trends and differences in listening habits among various groups.

Dashboard Overview

Figure 3: The final dashboard includes six views

Stories & Observations

First, we looked at the status of each plot in the United States. Figure 4 shows an overview of their condition.

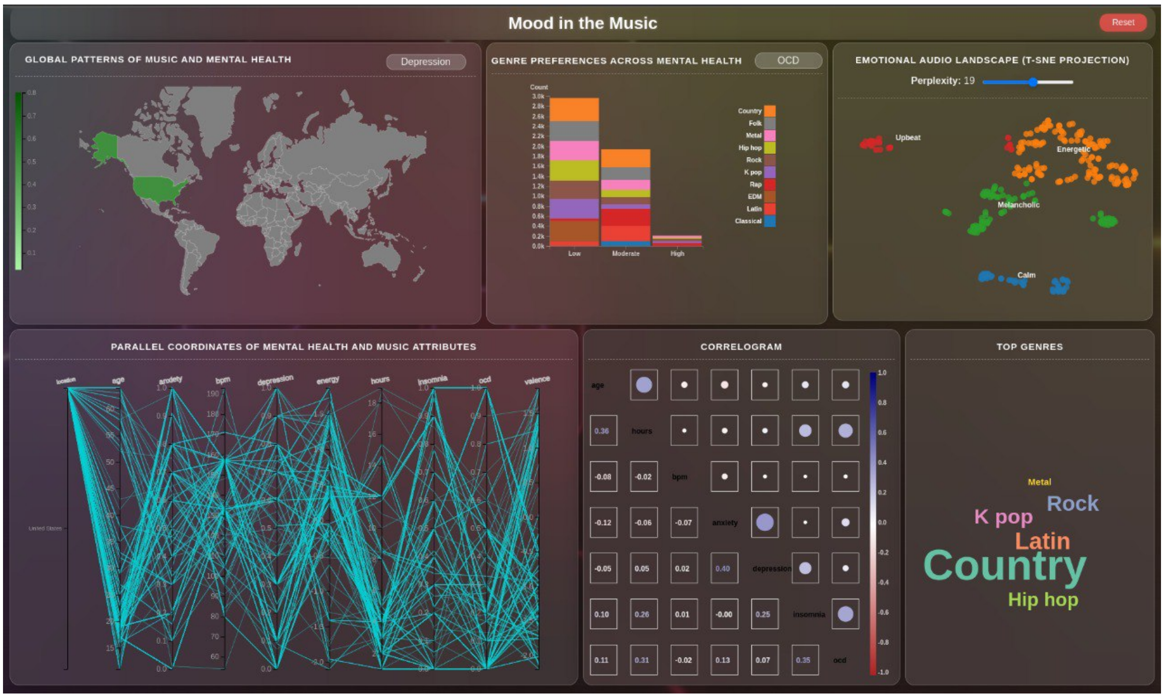
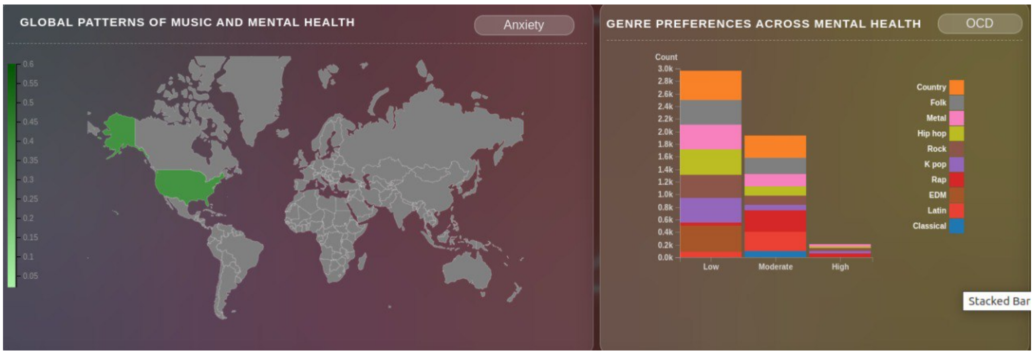
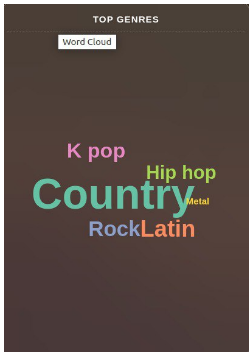


Figure 4: Filtering on USA

As you can see in Fig 5a, the USA tends to have fewer people with high levels of OCD. Americans with medium levels of OCD often listen to rap music. Also, the word cloud shows 5b that the top genre inside USA is country, then it's Hip Hop and K pop.



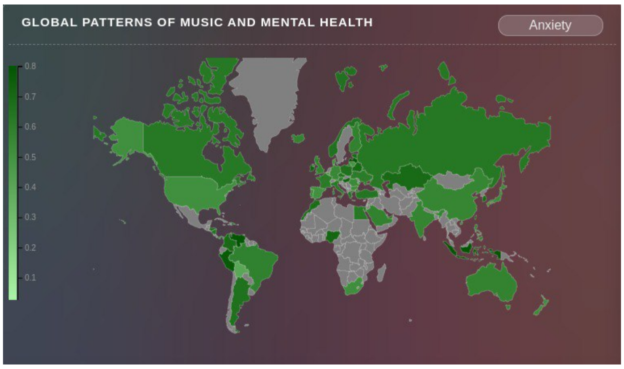
(a) USA genre preferences across mental health (OCD selected)



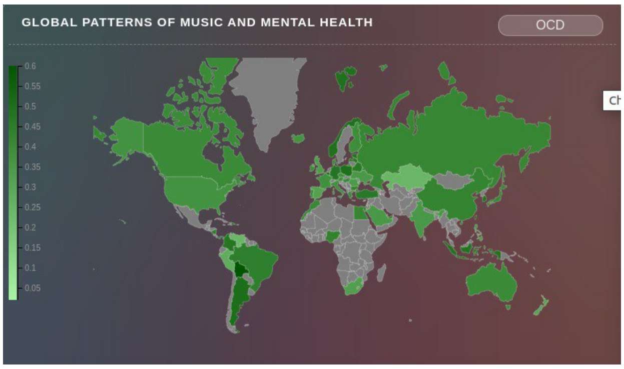
(b) Top genres

Figure 5: Comparison of genre preferences and top genres in the US

Next, we looked at anxiety and OCD levels in different countries based on their music preferences. We found that people in Canada and Russia tend to have higher levels of anxiety, while countries in South America show higher levels of OCD.



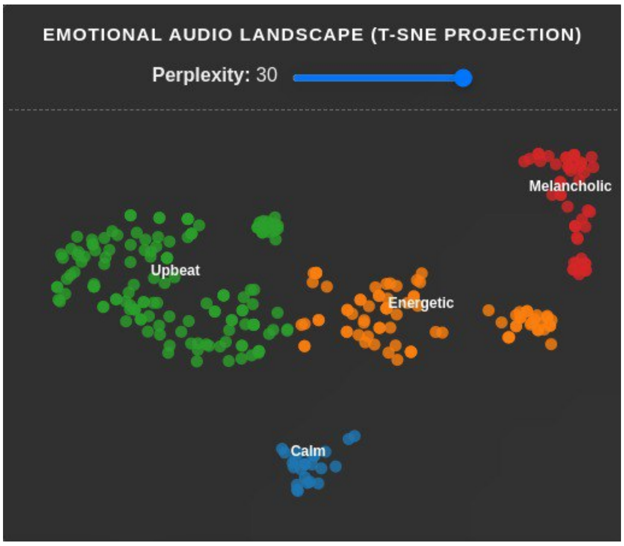
(a) Map of Anxiety



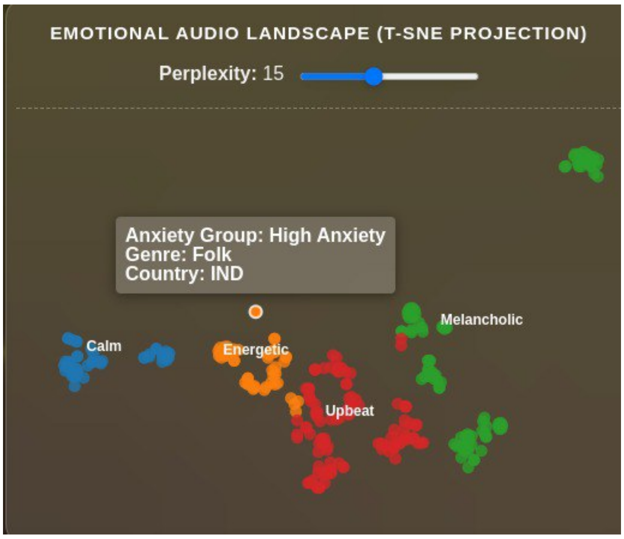
(b) Map of OCD

Figure 6: Comparison of Anxiety and OCD Maps

The t-SNE plots use a perplexity slider to adjust the view: lower values (0–15) highlight subtle emotional patterns within groups, while higher values (16–20) reveal broader emotional trends across mental health groups. Each point represents a surveyed individual, allowing us to see if people with varying anxiety levels prefer different emotional audio zones like Calm, Energetic, Melancholic, or Upbeat.



(a) Global Emotional audio fingerprint of people



(b) T-SNE plot

Figure 7: T-SNE projection

By looking at the Fig 8, our analysis reveals that older individuals tend to listen to music for more hours, and increased music listening time is linked to a higher likelihood of experiencing insomnia or OCD. Additionally, there is a strong positive correlation between anxiety and depression, which is expected as both are common mental health conditions. Similarly, insomnia and OCD also show a strong positive relationship.

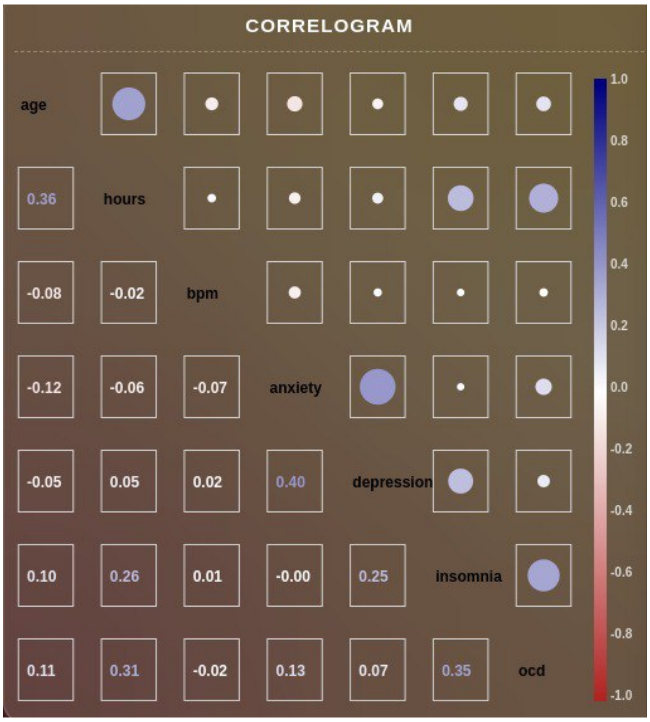


Figure 8: Correlogram Map

The brushed Principal Component Plot (PCP) 9 reveals an interesting trend among users with high insomnia. These individuals tend to listen to high-energy songs for longer periods of time. This suggests a possible link between insomnia and the preference for more stimulating, energetic music, potentially as a way to cope with or counteract feelings of fatigue or restlessness that come with sleep disorders.

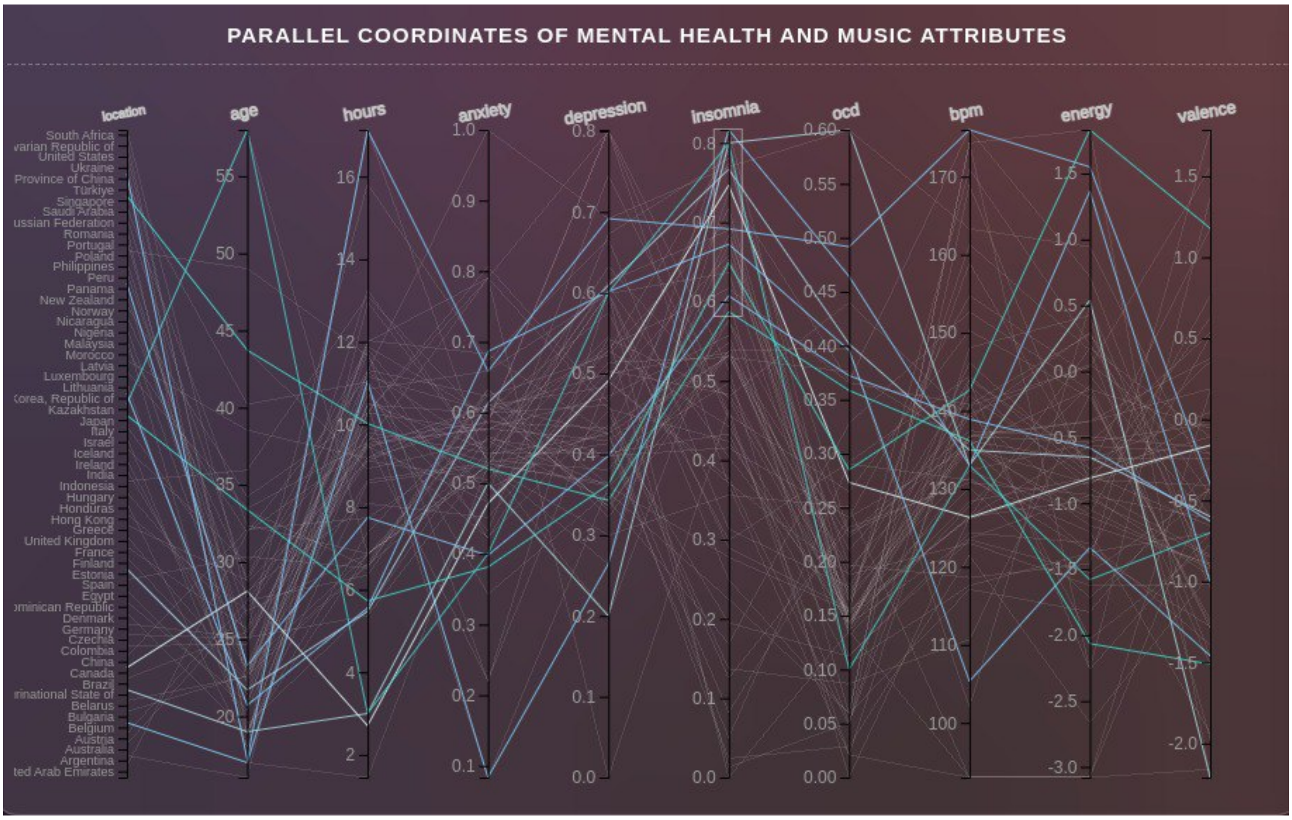


Figure 9: PCP with insomnia brushing

The stacked bar chart 10 provides further insight into the music preferences of users with high insomnia, showing a noticeable shift in the genres they prefer. People with high levels of insomnia overwhelmingly favor genres like classical, Latin, country, and EDM. This shift may reflect the desire for certain types of music that either help them relax (in the case of classical) or stay energized (as seen in EDM). The genre preferences point to how music can play a role in managing the symptoms or emotional states related to insomnia.

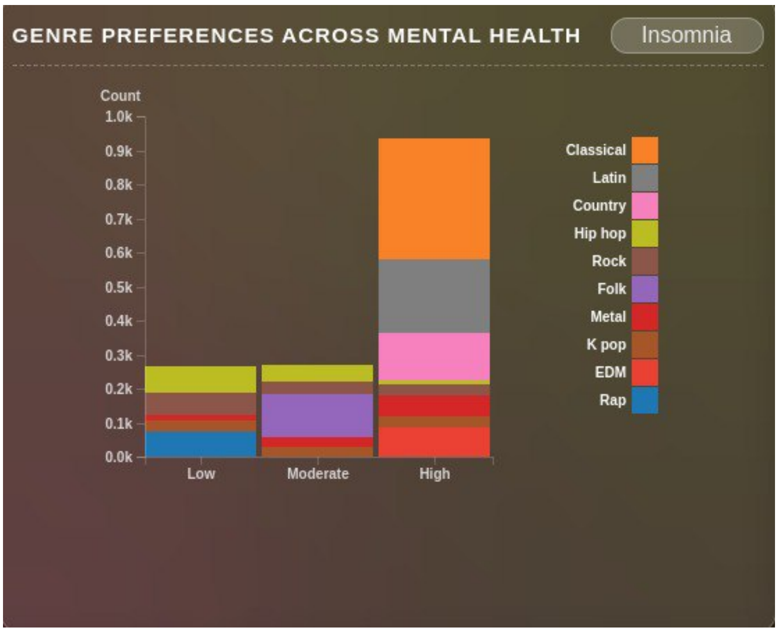


Figure 10: Insight 2 from stacked bar-chart

References

[1] Accelerated construction of stress relief music datasets: National Library of Medicine - https://pmc.ncbi.nlm.nih.gov/articles/PMC11125514

[2] Acoustic Sounds for Wellbeing; A Novel Dataset and Baseline Results: Cornell University - https://arxiv.org/abs/1908.0167

[3] Harnessing Social Music Tags for Characterizing Depression Risk: Cornell University - https://arxiv.org/abs/2007.13159

[4] MxMH dataset of people's music preferences and their self-reported mental health: Kaggle - https://www.kaggle.com/code/melissamonfared/mental-health-musicrelationship-analysis-eda

[5] Songs Attributes: Kaggle - https://www.kaggle.com/datasets/byomokeshsenapati/spotify-song-attributes

[6] The MediaEval Database for Emotional Analysis of Music: DEAM - https://cvml.unige.ch/databases/DEAM

[7] Top 2000s Spotify Songs: Kaggle - https://www.kaggle.com/datasets/iamsumat/spotify-top-2000s-mega-dataset

[8] Top Hits Apple Music: Kaggle - https://www.kaggle.com/datasets/kanchana1990/applemusic-dataset-10000-tracks-uncovered

[9] Top Hits Spotify: Kaggle - https://www.kaggle.com/code/youssefabdelghfar/top-hits-spotify-from-2000-2019