CSE 564: Final Project Proposal

By Amirhossein Najafizadeh & Iliya Mirzaei

Spring 2025

Project Title

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

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.

Research Questions

To guide our analysis, we frame the following central research questions:

1. Is there a statistical association between genre listening patterns (e.g. BPM, frequency of genres) and self-reported mental health indicators such as anxiety, depression, insomnia, and OCD?

2. Do individuals with higher levels of stress, OCD, or insomnia exhibit unique genre or feature preferences compared to the general population?
3. Can we identify distinct clusters or listener profiles that capture co-occurring mental health traits and music preferences?
4. To what extent can music consumption behaviors and genre preferences predict self-reported mental health scores?
5. Can your music taste guess your mental health score?
6. Are anxious people secretly jazz fans?
7. What genre is the unofficial soundtrack to insomnia?
8. Can we build the ultimate 'mental reset' playlist?
9. Who is the happiest headbanger?
10. Is foreign-language music an emotional shield?
11. What is your 'emotional audio fingerprint'?
12. Who listens to the most mentally cursed playlists?
13. Can we predict how bad your year was from your top genre + listening hours?
14. Is video game music the new therapy?
15. Can we generate your mental health-safe playlist using AI?

Datasets

We are working with six different datasets to understand the relationships between the following elements:

- Mental health and Music genres [4][6]
- Music genres and Songs [7][8]
- Songs and Audio features [5][9]

The first three datasets [7][8] help us explore how different music genres are connected to various songs. These datasets include information about which songs belong to which genres. The other datasets provide details about the features of each song [5][9], such as tempo, energy, and mood, which can help us analyze their connection to mental health using the datasets published by the National Library of Medicine [4][6]. The common link between mental health dataset, music genres and specific songs is the genre column, while the song name acts as the key connecting the features to the genre.

- **Age** – The age of the participant [4][6].
- **Hours per day** – How many hours of music the person listens to daily [4].
- **While working** – Whether the person listens to music while working [4].

- **Genre** – The music genres the person prefers [4][9].
- **Foreign language** – Is the song in a foreign language [4].
- **Genre Frequency** – How often a genre is listened to [4].
- **Anxiety Level** – Self-reported level of anxiety [4][6].
- **Depression Level** – Self-reported level of depression [4][6].
- **Insomnia Level** – Self-reported level of sleep difficulties [4][6].
- **OCD Level** – Self-reported level of obsessive-compulsive disorder symptoms [4][6].
- **Artist** – The name of the artist who performed the song [7][8].
- **BPM** – Beats per minute, representing the speed of the song [5][9].
- **Song Energy** – A measure of intensity and activity in the song [5][9].
- **Song Dance-ability** – How suitable the song is for dancing [5][9].
- **Song Loudness (dB)** – Overall loudness of the song in decibels [5][9].
- **Song Live-ness** – Likelihood that the song was recorded live [5][9].
- **Song Valence** – A measure of the musical positivity of the song [5][9].
- **Song Duration** – Length of the song in milliseconds [5][9].
- **Song Acoustic-ness** – How acoustic or non-electronic the song is [5][9].
- **Song Speech-ness** – Presence of spoken words in the song [5][9].
- **Song Popularity** – How popular the song is, based on listener data [5][9].
- **Song Mode** – Indicates whether the song is in a major or minor key [5][9].
- **Song Tempo** – Speed of the song, usually in BPM [5][9].
- **Song Year** – The year the song was released [7][8].

By analyzing these features, we aim to discover patterns and connections between people's mental health and the music they listen to.

Approach

In this section, we will explain the steps we took to prepare our data and perform the analysis. First, we start with raw data, which isn't ready for analysis yet. Our goal is to process and clean this data so that it can be used for meaningful insights. Once the data is ready, we will filter it to reduce the size from 45,000 items down to about 1,000. This step helps us focus on the most relevant data and makes the analysis more manageable. Next, we will analyze these sampled data points using the important features that we've selected. We will explore relationships between the variables and try to understand patterns or trends.

Data Wrangling

1. **Filling Missing Values**: The original dataset, which looks at the relationship between mental health and music genres, has some missing information. For each feature (or column), up to 30% of the data is missing, and about 10% of the entire dataset is incomplete overall. To fix this, we will use **Linear Regression** which helps us predict the missing values based on the patterns in the existing data. By doing this, we are not just guessing randomly — we are using logical relationships within the data to make smart predictions. This way, we can complete the dataset in a meaningful and reliable way.

2. **Normalization & Standardization**: The music-related attributes in our dataset are numbers on different scales, which makes them hard to compare directly. On the other hand, the mental health scores are on the same scale but have much larger values than the music features. To fix this, we'll first normalize the mental health scores by converting them to a range between 0 and 1. This helps bring them closer in scale to the music data. After that, we will standardize the entire dataset using **z-score standardization**. This method adjusts all the values so they have the same scale — with a mean of 0 and a standard deviation of 1. Doing this makes the data more consistent and easier to compare, which is important for accurate analysis and modeling.

3. **Resolve Inconsistencies**: The original datasets for songs and genres use different names or labels for the same music genres. This creates confusion and leads to duplicate entries. To fix this problem, we will use **linguistic algorithms** that can recognize and match similar or identical genre names, even if they are written slightly differently. By matching and merging genres that mean the same thing, we can clean up the data, remove duplicates, and make the dataset smaller and more accurate.

Merging Datasets

To combine the data from mental health and music attributes, we need to link them together in a meaningful way. First, we will use the genre key to connect the mental health dataset with the songs dataset. Once the datasets are linked by genre, we will then use the song name to connect the music attributes (such as BPM, song valence, and energy) to the dataset. By doing this, we'll create a unified dataset where both mental health data and music-related data are connected and can be analyzed together.
Here is a simplified example of what the final dataset will look like:

Age	Genre	Anxiety	Depression	BPM	Song Valence	Song Energy
18	Pop	0.5	0.2	0.23	0.01	0.5
32	Rock	0.9	0.9	0.86	0.01	0.72

Data Filtering & Dimension Reduction

To make our data easier to analyze and work with, we will first use **Principal Component Analysis (PCA)**. PCA helps us identify the most important features (or attributes) that explain the most variance in our data. This will allow us to focus on the key features that are most relevant for clustering. Next, we will use the **knee finding** method to determine the optimal number of features to keep — those with the highest variance. This helps us reduce the dimensionality of our data while retaining the most important information.
After reducing the dimensions, we will perform **K-Means** clustering to group our data into clusters based on their similarities. To find the best number of clusters, we will again use the knee finding method. This method helps us identify the point where adding more clusters

doesn't improve the model much, which gives us the ideal number of clusters. Once we have the best number of clusters, we will assign each data point to a cluster and sample the data according to the distribution of the clusters. This ensures that we have a good representation of all clusters in the final dataset. This method will reduce the size of our data from 45,000 entities to 1,000.

Visualization Methods

1. **PCP Plot** to visualize the relationship between all features
2. **Radar Plot** to visualize the relationship between music attributes and mental health levels
3. **Heat-map Matrix** to visualize the relationship between mental health levels and songs' features
4. **Scatter Plot** to visualize the attributes variance
5. **Histograms / Bar Charts** to visualize the different features values (like Age, Hours per day, Genre frequency, Mental health levels, Songs attributes)
6. **Pie Charts** to visualize the categorical attributes (like Sex, While working, Foreign language)
7. **Box Plots** to visualize the different mental health features values

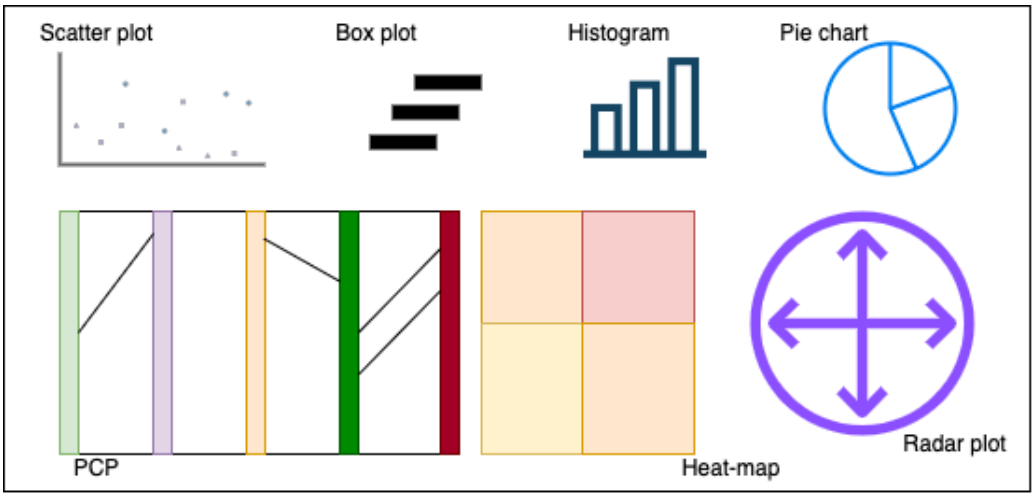

Figure 1: Simple design of the expected dashboard

Implementation Details

- Using **Python-Flask** as our back-end application.
- Using **React.js** as our front-end application.
- Using **D3.js** to plot our charts.

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-music-relationship-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/apple-music-dataset-10000-tracks-uncovered

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