



A Study on Music Listening Habits and Different Mental Health Conditions

Venura Shenan | 15405

02/07/2024

Abstract

Mental health conditions like anxiety, depression, insomnia, and obsessive-compulsive disorder (OCD) are becoming more common and affect many people. These conditions can be mild or severe, requiring different treatment approaches. One treatment that has gained attention is music therapy, particularly active listening to music, which has been linked to positive mental health outcomes. Previous research suggests that listening to music can reduce symptoms and improve well-being.

This study aims to explore the potential connections between music listening habits and the mentioned mental health conditions. By analyzing a comprehensive dataset, I sought to find specific music listening behaviors that correlate with the severity and presence of these mental health issues. The analysis revealed several interesting relationships, indicating that certain listening habits might be linked to changes in mental health status.

To further investigate these findings, we developed and trained a predictive model to classify mental health status based on individuals' music listening habits. While the model was able to accurately identify certain classes for specific conditions, its overall performance was not optimal. This highlights the complexity of predicting mental health status solely based on music listening habits.

The results of this study show the potential of music listening habits as indicators of mental health status but also stress the need for more advanced and refined predictive models. Future research should focus on including additional variables and using more advanced analytical techniques to improve the predictive power of such models. This study adds to the growing evidence on the link between music therapy and mental health, offering valuable insights for both clinical practice and future research.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 5 |
| 2 | Literature Review | 6 |
| 3 | Data | 7 |
| 4 | Theory and Methodology | 9 |
| 4.1 | Exploratory Data Analysis (EDA) | 9 |
| 4.2 | Data Preprocessing | 9 |
| 4.3 | Machine Learning Model: Random Forest Classifier | 9 |
| 4.4 | Enhancing Performance: K-Means Clustering | 9 |
| 4.5 | Model Evaluation | 10 |
| 5 | Exploratory Data Analysis | 11 |
| 5.1 | Dataframe Overview | 11 |
| 5.2 | Outliers | 12 |
| 5.3 | Correlation | 14 |
| 5.4 | Distributions of Numeric Variables | 15 |
| 5.5 | Listening Habits | 15 |
| 5.6 | Genres | 17 |
| 5.7 | General Behaviours | 19 |
| 5.8 | Mental Health Conditions | 20 |
| 6 | Advanced Analysis for Predictive Modeling | 27 |
| 6.1 | Model 1 - Random Forest Classifier | 27 |
| 6.2 | Model 2 - Random Forest Classifier with Reduced Response Class Counts | 30 |
| 6.3 | Model 3 - Random Forest Classifier with only 2 Response Classes | 33 |
| 6.4 | Model 4 - Random Forest Classifier with 2 Response Classes and Cross Validation | 35 |
| 7 | General Discussion and Conclusions | 39 |

List of Figures

| | | |
|----|---|----|
| 1 | Dataset Preview | 8 |
| 2 | Dataset Preview | 8 |
| 3 | Dataframe head | 11 |
| 4 | Dataframe Info | 11 |
| 5 | Statistics of Numeric Variables | 12 |
| 6 | Boxplots of Numeric Variables | 13 |
| 7 | Heatmap | 14 |
| 8 | Distribution of Age | 15 |
| 9 | Streaming Platform | 15 |
| 10 | Age vs Listening Habits | 16 |
| 11 | Genre Analysis | 18 |
| 12 | General Behaviours | 19 |
| 13 | Mental Health Conditions | 20 |
| 14 | Mental Health and Favourite Genre | 21 |
| 15 | Mental Health and Listening Times | 22 |
| 16 | Effects of Music Listening | 23 |
| 17 | Mental Health and Musical Background | 24 |
| 18 | Mental Health Condition of Individuals who have Worsened Conditions | 25 |
| 19 | Mental Health Condition of Individuals who have Improved Conditions | 26 |
| 20 | Elbow Plots for Target Variables | 30 |
| 21 | Natural Clusters within Target Variables | 30 |

List of Tables

| | | |
|----|--|----|
| 1 | Correlation Matrix of Variables | 14 |
| 2 | Classification Report for Anxiety - Model 1 | 28 |
| 3 | Classification Report for Depression - Model 1 | 28 |
| 4 | Classification Report for Insomnia - Model 1 | 29 |
| 5 | Classification Report for OCD - Model 1 | 29 |
| 6 | Classification Report for Anxiety - Model 2 | 31 |
| 7 | Classification Report for Depression - Model 2 | 31 |
| 8 | Classification Report for Insomnia - Model 2 | 31 |
| 9 | Classification Report for OCD - Model 2 | 32 |
| 10 | Classification Report for Anxiety - Model 3 | 33 |
| 11 | Classification Report for Depression - Model 3 | 34 |
| 12 | Classification Report for Insomnia - Model 3 | 34 |
| 13 | Classification Report for OCD - Model 3 | 34 |
| 14 | Best Parameters and CV Accuracy for Each Target Variable | 36 |
| 15 | Classification Report for Anxiety - Model 4 | 36 |
| 16 | Classification Report for Depression - Model 4 | 36 |
| 17 | Classification Report for Insomnia - Model 4 | 37 |
| 18 | Classification Report for OCD - Model 4 | 37 |

1 Introduction

Mental health is a pressing global issue affecting millions worldwide. Approximately 1 in 4 people suffer from mental health conditions, with depression alone affecting 350 million individuals globally. Alarming, the global suicide count stands at 700,000 deaths annually, surpassing malaria-related deaths. Suicide ranks as the fourth leading cause of death among 15-29 year olds, underscoring the urgency for effective interventions and treatment options [1].

Despite the widespread prevalence of mental health challenges, stigma and discrimination persist, affecting 9 out of 10 individuals with mental health conditions. Access to psychological care remains limited, with only 50% of individuals in high-income countries and a staggering 90% in low- and middle-income countries unable to access necessary treatment [1].

Music therapy, known for its global accessibility, presents a promising avenue for mental health treatment. Both active participation, such as playing instruments or singing, and passive listening to music are integral components of music therapy interventions. If we studied how mental health conditions differ with various music listening habits, we can use this results to ease the conditions more leisurely.

Objectives

- Explore the Impact Patterns of Different Music Genres on Mental Health Conditions: Examine the effects of various music genres on mental health issues such as anxiety, depression, insomnia, and obsessive-compulsive disorder (OCD).
- Examine the Impact Patterns of Listening Habits on Mental Health Conditions: Investigate how the frequency and duration of music listening sessions influence mental health outcomes. Determine optimal listening times to maximize therapeutic benefits for different mental health issues.
- Utilize Machine Learning Techniques to Predict Mental Health Conditions using Listening Habits: Develop machine learning models capable of predicting the severity of mental health conditions based on individual listening habits.
- Further Development: Apply findings to the development of a personalized music streaming platform that offers tailored music recommendations to assist users with mental health concerns.

Significance of the Study

This study aims to address gaps in mental health care by analyzing individual-level data on music preferences, listening habits, and mental health outcomes using machine learning techniques. By tailoring interventions effectively, this approach could potentially bridge the gap in mental health care services globally. The development of personalized music therapy has the potential to enhance mental health outcomes and overall well-being. **It is important to note that while music therapy offers promising benefits, it should complement, rather than replace, traditional mental health treatments.**

2 Literature Review

- According to a study conducted among Nigerian undergraduate students, music listening habits significantly influence mental health outcomes. Employing a cross-sectional survey design, the research focused on 400 students from tertiary institutions in Moor Plantation, Ibadan, Nigeria. The study revealed that a majority of students listen to music frequently, with Hip hop and Gospel being the most preferred genres. Positive effects on mood, such as increased happiness and relaxation, were commonly reported, alongside music's role in stress reduction and concentration enhancement. Interestingly, 95.25% of students used music as a coping mechanism during emotional distress. Despite these benefits, a small minority (2%) noted negative impacts on mental health associated with their music habits. Moreover, the study highlighted a high acceptance of music therapy among students, with 93% expressing openness to its potential benefits. These findings underscore music's potential as a beneficial tool for mental health enhancement among Nigerian undergraduates, emphasizing the need for further research and potential guidelines for optimizing music-based interventions.[2]
- Melissa Monfared has conducted an in-depth EDA and Relationship Analysis on the same dataset which provides some useful insights. [3]
- According to a project conducted by Catherine Rasgaitis on this dataset, a predictive model was developed to forecast mental health conditions using the dataset's variables, achieving up to an 18% accuracy while encompassing all response classes within the target variables. In this Study she has used 3 models namely Model 1A, Model 1B and Model 2. Model 1A was trained using Complete Rankings, MultiOutput + AdaBoost + RFC. Model 1B was trained using Binary Classification, MultiOutput + AdaBoost + RFC. Model 2 was trained using Complete Rankings, LazyPredict + XGBoost [4]

3 Data

The dataset was retrieved from Kaggle. According to the dataset description, data collection was managed via a Google Form. Respondents were not restricted by age or location. The form was posted in various Reddit forums, Discord servers, and social media platforms. Posters and "business cards" were also used to advertise the form in libraries, parks, and other public locations. The form was relatively brief to encourage completion, with more challenging questions (such as BPM) left optional for the same reason. Thus, the results of this research can be scaled to a worldwide population.

Variables

- **Timestamp:** Date and time when the form was submitted.
- **Age:** Respondent's age.
- **Primary streaming service:** Respondent's primary streaming service.
- **Hours per day:** Number of hours the respondent listens to music per day.
- **While working:** Whether the respondent listens to music while studying/working.
- **Instrumentalist:** Whether the respondent plays an instrument regularly.
- **Composer:** Whether the respondent composes music.
- **Favorite genre:** Respondent's favorite or top genre.
- **Exploratory:** Whether the respondent actively explores new artists/genres.
- **Foreign languages:** Whether the respondent regularly listens to music with lyrics in a language they are not fluent in.
- **BPM:** Beats per minute of respondent's favorite genre.
- **Frequency [Classical]:** Frequency of listening to classical music.
- **Frequency [Country]:** Frequency of listening to country music.
- **Frequency [EDM]:** Frequency of listening to EDM music.
- **Frequency [Folk]:** Frequency of listening to folk music.
- **Frequency [Gospel]:** Frequency of listening to Gospel music.
- **Frequency [Hip hop]:** Frequency of listening to hip hop music.
- **Frequency [Jazz]:** Frequency of listening to jazz music.
- **Frequency [K pop]:** Frequency of listening to K pop music.
- **Frequency [Latin]:** Frequency of listening to Latin music.
- **Frequency [Lofi]:** Frequency of listening to lofi music.
- **Frequency [Metal]:** Frequency of listening to metal music.
- **Frequency [Pop]:** Frequency of listening to pop music.
- **Frequency [R&B]:** Frequency of listening to R&B music.
- **Frequency [Rap]:** Frequency of listening to rap music.

- **Frequency [Rock]:** Frequency of listening to rock music.
- **Frequency [Video game music]:** Frequency of listening to video game music.
- **Anxiety:** Self-reported anxiety level (scale of 0-10).
- **Depression:** Self-reported depression level (scale of 0-10).
- **Insomnia:** Self-reported insomnia level (scale of 0-10).
- **OCD:** Self-reported OCD level (scale of 0-10).
- **Music effects:** Whether music improves or worsens respondent's mental health conditions.

| | Timestamp | Age | Primary streaming service | Hours per day | White coding | Instrumental | Composer | Fan game | Esplanade | Foreign language | BPM | Frequency (Classical) | Frequency (Country) | Frequency (EDM) | Frequency (Pop) |
|----|-------------------|-----|---------------------------|---------------|--------------|--------------|----------|------------------|-----------|------------------|-----|-----------------------|---------------------|-----------------|-----------------|
| 1 | 0/7/2022 10:29:50 | 16 | Spotify | 3 | Yes | Yes | Yes | Yes | Yes | Yes | 116 | Rarely | Never | Rarely | Never |
| 2 | 0/7/2022 10:37:31 | 63 | Pandora | 1.5 | Yes | No | No | Rock | Yes | No | 119 | Sometimes | Never | Never | Rarely |
| 3 | 0/7/2022 21:25:18 | 18 | Spotify | 4 | No | No | No | Video game music | No | Yes | 132 | Never | Never | Very frequently | Never |
| 4 | 0/7/2022 14:48:48 | 61 | YouTube Music | 2.5 | Yes | Yes | Yes | Jazz | Yes | Yes | 84 | Sometimes | Never | Never | Rarely |
| 5 | 0/7/2022 21:54:47 | 10 | Spotify | 4 | Yes | No | No | R&B | Yes | No | 107 | Never | Never | Rarely | Never |
| 6 | 0/7/2022 21:58:50 | 18 | Spotify | 5 | Yes | Yes | Yes | Jazz | Yes | Yes | 86 | Rarely | Sometimes | Never | Never |
| 7 | 0/7/2022 22:30:28 | 18 | YouTube Music | 3 | Yes | Yes | No | Video game music | Yes | Yes | 86 | Sometimes | Never | Rarely | Sometimes |
| 8 | 0/7/2022 22:18:59 | 21 | Spotify | 1 | Yes | No | No | K-pop | Yes | Yes | 96 | Never | Never | Rarely | Rarely |
| 9 | 0/7/2022 22:33:05 | 18 | Spotify | 8 | Yes | No | No | Rock | No | No | 94 | Never | Very frequently | Never | Sometimes |
| 10 | 0/7/2022 22:44:03 | 15 | I use a streaming service | 1 | Yes | No | No | R&B | Yes | Yes | 105 | Rarely | Rarely | Rarely | Rarely |
| 11 | 0/7/2022 22:51:15 | 18 | Spotify | 3 | Yes | Yes | No | Country | Yes | No | | Never | Very frequently | Never | Never |
| 12 | 0/7/2022 23:39:32 | 19 | YouTube Music | 8 | Yes | No | No | EDM | Yes | No | 125 | Rarely | Never | Very frequently | Never |
| 13 | 0/7/2022 23:49:00 | 17 | Spotify | 3 | Yes | No | No | Hip-hop | Yes | Yes | | Rarely | Never | Rarely | Never |
| 14 | 0/7/2022 23:12:03 | 19 | Spotify | 2 | Yes | No | No | Country | Yes | No | 88 | Never | Very frequently | Rarely | Sometimes |
| 15 | 0/7/2022 23:18:06 | 18 | Spotify | 4 | Yes | Yes | No | Jazz | Yes | Yes | 148 | Very frequently | Rarely | Never | Never |
| 16 | 0/7/2022 23:19:52 | 17 | Spotify | 2 | No | No | No | Pop | Yes | Yes | | Rarely | Rarely | Never | Never |
| 17 | 0/7/2022 23:39:41 | 18 | Spotify | 8 | Yes | No | No | Hip-hop | Yes | Yes | 103 | Never | Never | Never | Never |
| 18 | 0/7/2022 23:39:48 | 18 | Spotify | 12 | Yes | No | Yes | Hip-hop | Yes | Yes | 120 | Rarely | Never | Sometimes | Rarely |
| 19 | 0/7/2022 23:48:56 | 17 | Spotify | 24 | No | No | No | R&B | Yes | No | 88 | Rarely | Never | Never | Never |
| 20 | 0/7/2022 23:41:58 | 15 | Spotify | 3 | No | No | No | Hip-hop | No | No | 130 | Never | Never | Never | Never |
| 21 | 0/7/2022 23:41:58 | 17 | Apple Music | 4 | Yes | No | No | Hip-hop | Yes | Yes | 138 | Rarely | Never | Sometimes | Rarely |
| 22 | 0/8/2022 0:28:02 | 17 | Apple Music | 4 | Yes | No | No | Pop | Yes | No | 135 | Never | Rarely | Rarely | Never |
| 23 | 0/8/2022 1:39:02 | 18 | I use a streaming service | 5 | Yes | No | No | R&B | Yes | Yes | 116 | Rarely | Rarely | Sometimes | Very frequently |
| 24 | 0/8/2022 1:19:48 | 18 | Spotify | 2 | Yes | No | No | Pop | Yes | No | 79 | Rarely | Never | Never | Rarely |
| 25 | 0/8/2022 4:13:11 | 16 | Other streaming service | 3 | Yes | Yes | Yes | Rock | Yes | Yes | 84 | Rarely | Rarely | Never | Rarely |
| 26 | 0/8/2022 4:38:14 | 18 | Spotify | 2 | No | No | No | Pop | Yes | Yes | 109 | Sometimes | Rarely | Rarely | Very frequently |
| 27 | 0/8/2022 4:40:38 | 14 | Spotify | 18 | Yes | Yes | Yes | Rock | Yes | Yes | 136 | Sometimes | Rarely | Rarely | Rarely |
| 28 | 0/8/2022 5:55:51 | 18 | YouTube Music | 6 | Yes | Yes | Yes | Pop | No | Yes | 101 | Sometimes | Rarely | Rarely | Rarely |
| 29 | 0/8/2022 5:16:30 | 17 | Spotify | 2 | Yes | Yes | No | Pop | Yes | Yes | 126 | Never | Very frequently | Rarely | Rarely |
| 30 | 0/8/2022 5:30:27 | 17 | Apple Music | 5 | Yes | No | No | Pop | Yes | No | 143 | Rarely | Never | Never | No |
| 31 | 0/8/2022 10:30:22 | 20 | Apple Music | 5 | Yes | No | No | Rock | Yes | Yes | | Never | Rarely | Rarely | Very frequently |
| 32 | 0/8/2022 10:38:05 | 19 | Spotify | 2 | Yes | No | No | Classical | No | No | 120 | Very frequently | Never | Never | Never |
| 33 | 0/8/2022 10:54:30 | 18 | Spotify | 6 | Yes | No | No | Jazz | Yes | Yes | | Never | Never | Never | Rarely |
| 34 | 0/8/2022 10:59:53 | 17 | Spotify | 4 | No | No | No | Rock | Yes | No | 142 | Rarely | Rarely | Rarely | Very frequently |
| 35 | 0/8/2022 11:08:51 | 18 | Spotify | 1 | Yes | No | No | Classical | No | No | 75 | Very frequently | Never | Never | Rarely |
| 36 | 0/8/2022 11:13:25 | 18 | Spotify | 5 | Yes | Yes | Yes | Pop | Yes | No | 103 | Sometimes | Rarely | Sometimes | Never |
| 37 | 0/8/2022 11:25:48 | 21 | Spotify | 4 | Yes | No | No | Pop | No | Yes | 89 | Never | Never | Rarely | Never |
| 38 | 0/8/2022 11:27:18 | 17 | Other streaming service | 3 | Yes | No | No | Pop | Yes | No | | Never | Never | Sometimes | Never |
| 39 | 0/8/2022 11:38:38 | 26 | Other streaming service | 0.5 | No | No | No | Rock | No | Yes | 140 | Rarely | Never | Rarely | Sometimes |
| 40 | 0/8/2022 11:39:21 | 20 | Spotify | 4 | Yes | No | No | EDM | Yes | No | 161 | Rarely | Rarely | Very frequently | Sometimes |
| 41 | 0/8/2022 11:50:31 | 23 | YouTube Music | 2 | Yes | Yes | Yes | Video game music | Yes | No | 80 | Sometimes | Rarely | Never | Rarely |

Figure 1: Dataset Preview

| Frequency (Latin) | Frequency (Lofi) | Frequency (Metal) | Frequency (Pop) | Frequency (R&B) | Frequency (Rock) | Frequency (Soul) | Frequency (Soviet) | Frequency (Video game music) | Anxiety | Depression | Insomnia | OCD | Music effects | Permissions |
|-------------------|------------------|-------------------|-----------------|-----------------|------------------|------------------|--------------------|------------------------------|---------|------------|----------|-----|---------------|-------------|
| Very frequently | Rarely | Never | Very frequently | Sometimes | Very frequently | Never | Sometimes | | 3 | 0 | 1 | 0 | | Understand |
| Sometimes | Rarely | Never | Sometimes | Sometimes | Rarely | Very frequently | Rarely | | 7 | 2 | 2 | 1 | | Understand |
| Never | Sometimes | Sometimes | Rarely | Never | Rarely | Rarely | Very frequently | | 7 | 7 | 10 | 2 | No effect | Understand |
| Very frequently | Sometimes | Never | Sometimes | Never | Never | Never | Never | | 8 | 7 | 10 | 3 | Improve | Understand |
| Sometimes | Sometimes | Never | Sometimes | Very frequently | Very frequently | Never | Rarely | | 7 | 2 | 5 | 9 | Improve | Understand |
| Rarely | Very frequently | Rarely | Very frequently | Very frequently | Very frequently | Very frequently | Never | | 6 | 8 | 7 | 7 | Improve | Understand |
| Rarely | Never | Rarely | Rarely | Rarely | Never | Sometimes | Never | | 4 | 8 | 6 | 0 | Improve | Understand |
| Never | Sometimes | Never | Sometimes | Sometimes | Rarely | Never | Rarely | | 5 | 3 | 5 | 3 | Improve | Understand |
| Never | Never | Very frequently | Never | Never | Never | Very frequently | Never | | 2 | 0 | 0 | 0 | Improve | Understand |
| Rarely | Rarely | Never | Sometimes | Sometimes | Rarely | Sometimes | Sometimes | | 2 | 2 | 5 | 1 | Improve | Understand |
| Never | Never | Never | Rarely | Rarely | Never | Never | Never | | 7 | 7 | 4 | 7 | No effect | Understand |
| Rarely | Rarely | Never | Rarely | Rarely | Sometimes | Rarely | Rarely | | 1 | 0 | 0 | 1 | Improve | Understand |
| Never | Very frequently | Never | Sometimes | Sometimes | Rarely | Rarely | Never | | 9 | 3 | 2 | 7 | Improve | Understand |
| Never | Never | Never | Rarely | Never | Never | Very frequently | Never | | 2 | 1 | 2 | 0 | Improve | Understand |
| Sometimes | Rarely | Sometimes | Sometimes | Never | Never | Sometimes | Rarely | | 6 | 4 | 7 | 0 | Improve | Understand |
| Rarely | Rarely | Rarely | Very frequently | Never | Sometimes | Sometimes | Rarely | | 7 | 5 | 4 | 1 | Worsen | Understand |
| Never | Never | Never | Never | Sometimes | Very frequently | Never | Rarely | | 6 | 8 | 4 | 3 | Improve | Understand |
| Never | Never | Sometimes | Sometimes | Rarely | Sometimes | Very frequently | Never | | 5 | 7 | 10 | 0 | Improve | Understand |
| Rarely | Never | Sometimes | Rarely | Sometimes | Very frequently | Very frequently | Never | | 7 | 5 | 0 | 3 | Improve | Understand |
| Never | Never | Rarely | Rarely | Sometimes | Very frequently | Rarely | Never | | 7 | 3 | 0 | 2 | Improve | Understand |
| Sometimes | Rarely | Rarely | Very frequently | Rarely | Very frequently | Sometimes | Sometimes | | 6 | 9 | 3 | 0 | Improve | Understand |
| Never | Rarely | Rarely | Rarely | Never | Very frequently | Sometimes | Never | | 10 | 10 | 2 | 4 | Improve | Understand |
| Never | Rarely | Rarely | Rarely | Very frequently | Very frequently | Sometimes | Rarely | | 6 | 7 | 5 | 4 | Improve | Understand |
| Sometimes | Never | Never | Very frequently | Never | Never | Never | Rarely | | 3 | 3 | 6 | 5 | No effect | Understand |
| Rarely | Never | Sometimes | Never | Sometimes | Rarely | Very frequently | Sometimes | | 10 | 6 | 8 | 10 | Improve | Understand |
| Never | Sometimes | Never | Very frequently | Sometimes | Sometimes | Sometimes | Never | | 7 | 4 | 2 | 5 | Improve | Understand |
| Never | Very frequently | Never | Very frequently | Very frequently | Very frequently | Very frequently | Sometimes | | 8 | 6 | 10 | 5 | Improve | Understand |
| Never | Never | Never | Never | Sometimes | Sometimes | Never | Never | | 3 | 2 | 1 | 2 | Improve | Understand |
| Very frequently | Rarely | Rarely | Very frequently | Rarely | Never | Sometimes | Sometimes | | 6 | 6 | 4 | 5 | Improve | Understand |
| Never | Never | Sometimes | Never | Never | Never | Sometimes | Never | | 8 | 2 | 1 | 5 | Improve | Understand |
| Sometimes | Sometimes | Very frequently | Sometimes | Sometimes | Very frequently | Rarely | Rarely | | 7 | 7 | 2 | 0 | Improve | Understand |
| Never | Sometimes | Never | Very frequently | Never | Never | Very frequently | Very frequently | | 4 | 4 | 4 | 3 | No effect | Understand |
| Never | Sometimes | Very frequently | Very frequently | Never | Never | Sometimes | Sometimes | | 9 | 8 | 2 | 3 | Improve | Understand |
| Never | Rarely | Rarely | Very frequently | Rarely | Sometimes | Very frequently | Never | | 5 | 6 | 6 | 1 | Improve | Understand |
| Never | Never | Rarely | Never | Never | Never | Never | Rarely | | 0 | 0 | 0 | 0 | No effect | Understand |
| Rarely | Sometimes | Rarely | Very frequently | Sometimes | Sometimes | Rarely | Never | | 3 | 2 | 1 | 1 | Improve | Understand |
| Rarely | Never | Never | Very frequently | Never | Rarely | Never | Never | | 6 | 8 | 0 | 2 | Improve | Understand |
| Never | Rarely | Never | Very frequently | Sometimes | Very frequently | Very frequently | Never | | 2 | 0 | 4 | 0 | No effect | Understand |
| Never | Never | Never | Rarely | Rarely | Never | Very frequently | Never | | 9 | 8 | 1 | 0 | No effect | Understand |
| Never | Sometimes | Rarely | Very frequently | Very frequently | Sometimes | Very frequently | Rarely | | 7 | 2 | 8 | 6 | Improve | Understand |
| Rarely | Rarely | Rarely | Sometimes | Sometimes | Sometimes | Very frequently | Sometimes | | 4 | 2 | 3 | 4 | Improve | Understand |
| Never | Sometimes | Very frequently | Sometimes | Rarely | Sometimes | Very frequently | Sometimes | | 7 | 6 | 5 | 0 | Improve | Understand |

Figure 2: Dataset Preview

4 Theory and Methodology

4.1 Exploratory Data Analysis (EDA)

To investigate the relationship between music listening habits and mental health conditions, this study began with a comprehensive Exploratory Data Analysis (EDA). The EDA process involved examining distributions, correlations, and patterns within the dataset encompassing music listening habits, demographic variables, and mental health conditions. Insights were generated to identify potential relationships and dependencies among variables through visualizations and statistical summaries.

4.2 Data Preprocessing

After EDA, the dataset underwent preprocessing to prepare it for modeling. Numeric variables were scaled using the Standard Scaler technique to normalize features, ensuring that variables with larger scales did not dominate during model training. For categorical variables, One-Hot Encoding was applied to nominal variables, creating binary columns for each unique category. Ordinal variables were encoded using Label Encoding, assigning numerical labels based on the order of categories.

4.3 Machine Learning Model: Random Forest Classifier

The primary machine learning model employed in this study was the Random Forest Classifier. Random Forest is an ensemble learning method that combines multiple decision trees during training. Each decision tree is trained on a random subset of the data and features, which helps to reduce overfitting and improve generalization. The final prediction is determined by aggregating predictions from all trees through a voting mechanism, where the mode of predictions is taken for classification tasks. [5]

Mathematically, the prediction \hat{y} from a Random Forest can be represented as:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_n(x))$$

where $T_i(x)$ denotes the prediction of the i -th decision tree.

4.4 Enhancing Performance: K-Means Clustering

To further enhance model performance, K-Means clustering was applied to the target variable (mental health conditions) to find the optimal number of factor levels. K-Means is an unsupervised clustering algorithm that partitions data into K clusters based on similarity in mental health condition patterns. The algorithm initializes with random centroids and iteratively updates them to minimize the sum of squared distances within clusters.

The objective function of K-Means is to minimize:

$$\sum_{k=1}^K \sum_{x \in C_k} ||x - \mu_k||^2$$

where C_k represents the set of points in cluster k , and μ_k is the centroid of cluster k .

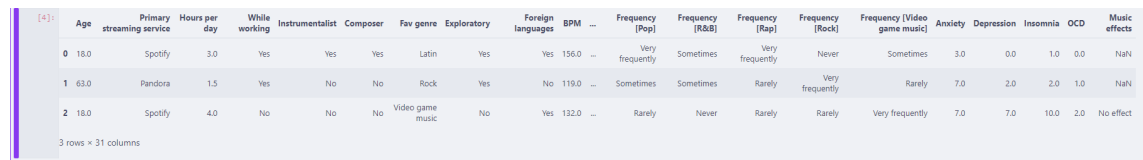
4.5 Model Evaluation

Model Evaluation is a critical aspect of every machine learning project as it is the step where we can get an idea of how well the trained model performs. In this study to evaluate the trained models, classification reports were used. The used metrics within the classification reports are as follows

- **Precision:** Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates how many of the predicted positive cases are actually positive. It is calculated as the ratio of true positives to the sum of true positives and false positives.
- **Recall:** Recall (or sensitivity) measures the proportion of true positives that are correctly identified by the model out of all actual positives. It indicates how well the model can identify true positive cases. It is calculated as the ratio of true positives to the sum of true positives and false negatives.
- **F1-score:** The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both measures. It is particularly useful when the class distribution is imbalanced. The F1-score is calculated as $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.
- **Macro average:** The macro average for metrics such as the F1-score and precision represents the unweighted average of the metric across all classes. This treats each class equally regardless of its size.
- **Weighted average:** The weighted average for metrics such as the F1-score and precision gives more weight to classes with a higher number of instances. This provides a balanced evaluation metric considering the class distribution, offering a more comprehensive performance measure in imbalanced datasets.

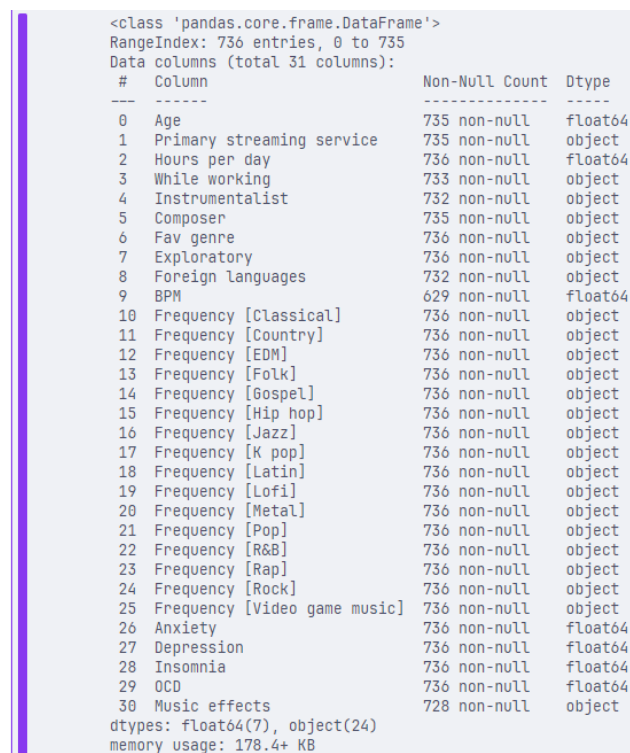
5 Exploratory Data Analysis

5.1 Dataframe Overview



| | Age | Primary streaming service | Hours per day | While working | Instrumentalist | Composer | Fav genre | Exploratory | Foreign languages | BPM | Frequency [Pop] | Frequency [R&B] | Frequency [Rap] | Frequency [Rock] | Frequency [Video game music] | Anxiety | Depression | Insomnia | OCD | Music effects |
|---|------|---------------------------|---------------|---------------|-----------------|----------|------------------|-------------|-------------------|-------|-----------------|-----------------|-----------------|------------------|------------------------------|---------|------------|----------|-----|---------------|
| 0 | 18.0 | Spotify | 3.0 | Yes | Yes | Yes | Latin | Yes | Yes | 136.0 | Very frequently | Sometimes | Very frequently | Never | Sometimes | 3.0 | 0.0 | 1.0 | 0.0 | NaN |
| 1 | 63.0 | Pandora | 1.5 | Yes | No | No | Rock | Yes | No | 119.0 | Sometimes | Sometimes | Rarely | Very frequently | Rarely | 7.0 | 2.0 | 2.0 | 1.0 | NaN |
| 2 | 18.0 | Spotify | 4.0 | No | No | No | Video game music | No | Yes | 132.0 | Rarely | Never | Rarely | Rarely | Very frequently | 7.0 | 7.0 | 10.0 | 2.0 | No effect |

Figure 3: Dataframe head



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 736 entries, 0 to 735
Data columns (total 31 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Age                                       735 non-null    float64
1   Primary streaming service               735 non-null    object
2   Hours per day                           736 non-null    float64
3   While working                           733 non-null    object
4   Instrumentalist                          732 non-null    object
5   Composer                                735 non-null    object
6   Fav genre                               736 non-null    object
7   Exploratory                             736 non-null    object
8   Foreign languages                       732 non-null    object
9   BPM                                      629 non-null    float64
10  Frequency [Classical]                   736 non-null    object
11  Frequency [Country]                     736 non-null    object
12  Frequency [EDM]                         736 non-null    object
13  Frequency [Folk]                        736 non-null    object
14  Frequency [Gospel]                      736 non-null    object
15  Frequency [Hip hop]                     736 non-null    object
16  Frequency [Jazz]                        736 non-null    object
17  Frequency [K pop]                       736 non-null    object
18  Frequency [Latin]                       736 non-null    object
19  Frequency [Lofi]                        736 non-null    object
20  Frequency [Metal]                       736 non-null    object
21  Frequency [Pop]                         736 non-null    object
22  Frequency [R&B]                         736 non-null    object
23  Frequency [Rap]                         736 non-null    object
24  Frequency [Rock]                        736 non-null    object
25  Frequency [Video game music]            736 non-null    object
26  Anxiety                                 736 non-null    float64
27  Depression                              736 non-null    float64
28  Insomnia                               736 non-null    float64
29  OCD                                     736 non-null    float64
30  Music effects                           728 non-null    object
dtypes: float64(7), object(24)
memory usage: 178.4+ KB
```

Figure 4: Dataframe Info

Here we can see all the columns and their relevant data type and non null record count

| | count | mean | std | min | 25% | 50% | 75% | max |
|---------------|-------|--------------|--------------|------|-------|-------|-------|-------------|
| Age | 736.0 | 2.520109e+01 | 1.204776e+01 | 10.0 | 18.0 | 21.0 | 28.0 | 89.0 |
| Hours per day | 736.0 | 3.572758e+00 | 3.028199e+00 | 0.0 | 2.0 | 3.0 | 5.0 | 24.0 |
| BPM | 736.0 | 1.358818e+06 | 3.686048e+07 | 0.0 | 105.0 | 120.0 | 140.0 | 999999999.0 |
| Anxiety | 736.0 | 5.836957e+00 | 2.792710e+00 | 0.0 | 4.0 | 6.0 | 8.0 | 10.0 |
| Depression | 736.0 | 4.794837e+00 | 3.029564e+00 | 0.0 | 2.0 | 5.0 | 7.0 | 10.0 |
| Insomnia | 736.0 | 3.737772e+00 | 3.088797e+00 | 0.0 | 1.0 | 3.0 | 6.0 | 10.0 |
| OCD | 736.0 | 2.635870e+00 | 2.840047e+00 | 0.0 | 0.0 | 2.0 | 5.0 | 10.0 |

Figure 5: Statistics of Numeric Variables

Here we can see some important statistics about the numeric variables. The mean age of the respondents is 25. The average listening time of respondents is 3.5 hours. The max BPM value suggests that there are garbage values in the dataset which need cleaning. The average anxiety and Depression levels are higher than the average OCD and Insomnia levels.

5.2 Outliers

As depicted in Figure 6 below, there are noticeable outliers in several variables. The following steps were taken to manage these outliers effectively:

Capping the Age Variable

To ensure the relevance and accuracy of our study, the age variable was capped between 5 and 75 years. This range was selected based on the following considerations:

- **Minimum Age:** Mental health conditions can manifest as early as age 5.
- **Maximum Age:** Considering the average global life expectancy of around 71 years, an upper limit of 75 years was chosen to generalize findings effectively.

Removing Extreme Values in BPM Columns

Extreme values in the Beats Per Minute (BPM) columns were removed to maintain dataset integrity. This step is crucial to eliminate anomalies that could skew our analysis and to focus on realistic and clinically relevant data points.

Addressing Outliers in Listening Times

Outliers in listening times can significantly distort results and lead to inaccurate conclusions. Therefore, identified outliers were removed to ensure that our analysis reflects typical listening behaviors and provides meaningful insights into the impact of music therapy on mental health conditions.

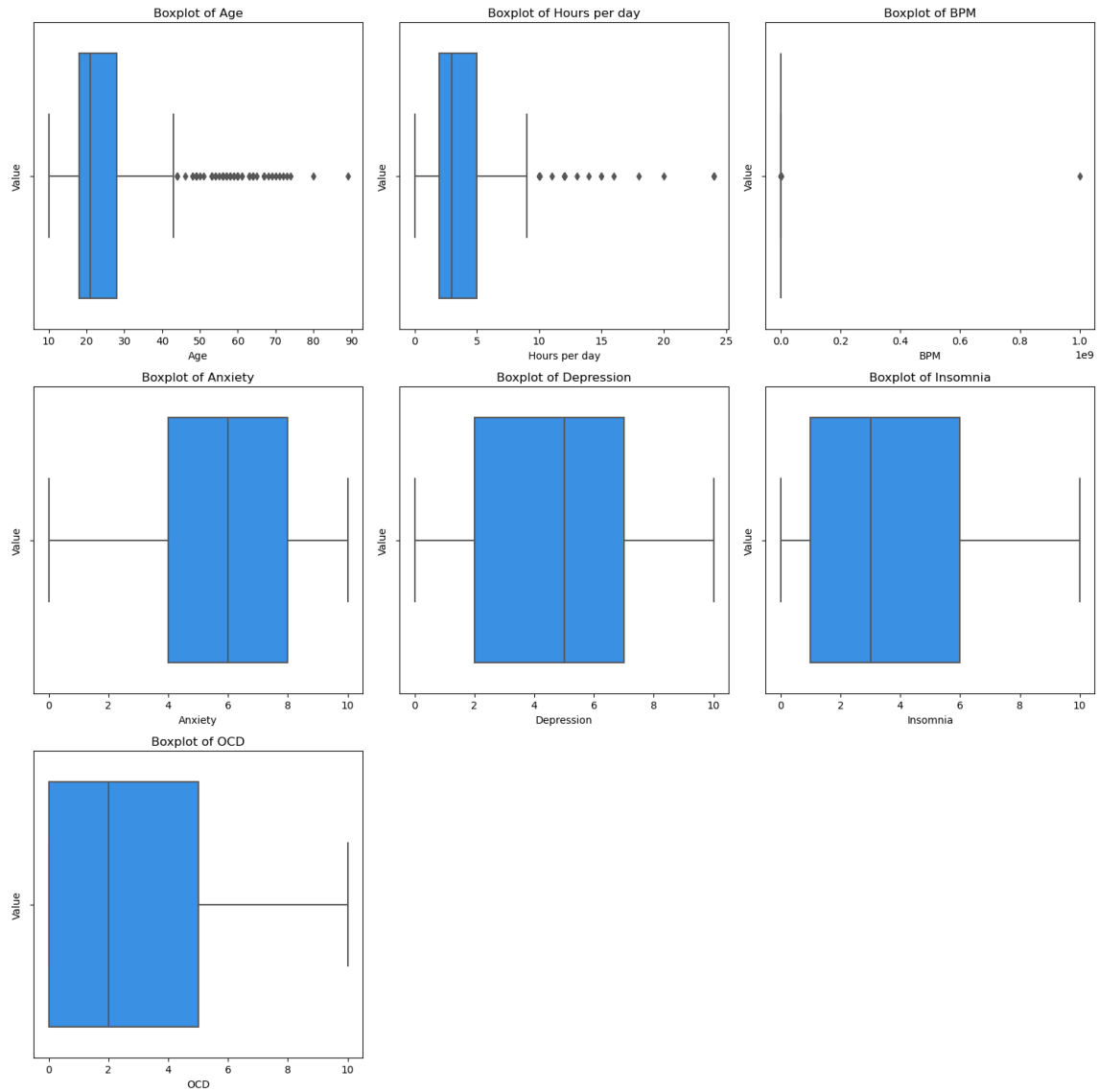


Figure 6: Boxplots of Numeric Variables

5.3 Correlation

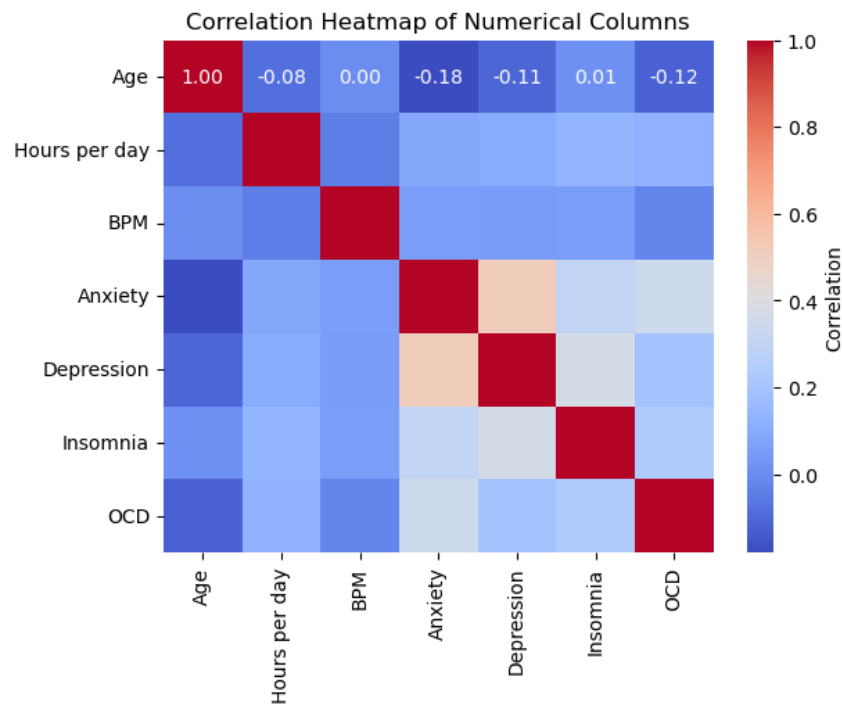


Figure 7: Heatmap

Table 1: Correlation Matrix of Variables

| | Age | Hours per day | BPM | Anxiety | Depression | Insomnia |
|---------------|--------|---------------|--------|---------|------------|----------|
| Age | 1.000 | -0.084 | 0.003 | -0.181 | -0.111 | 0.011 |
| Hours per day | -0.084 | 1.000 | -0.046 | 0.084 | 0.096 | 0.135 |
| BPM | 0.003 | -0.046 | 1.000 | 0.051 | 0.046 | 0.050 |
| Anxiety | -0.181 | 0.084 | 0.051 | 1.000 | 0.511 | 0.298 |
| Depression | -0.111 | 0.096 | 0.046 | 0.511 | 1.000 | 0.366 |
| Insomnia | 0.011 | 0.135 | 0.050 | 0.298 | 0.366 | 1.000 |
| OCD | -0.121 | 0.120 | -0.020 | 0.344 | 0.190 | 0.233 |

The correlation matrix reveals some interesting relationships between the variables. Age seems to have a slight negative correlation with factors like depression, anxiety, and OCD, suggesting these issues may decrease as people get older. The number of hours per day a person spends has a small positive correlation with anxiety, depression, insomnia and OCD, implying these mental health concerns tend to rise as daily listening time increases. The strongest relationship is between anxiety and depression, which have a moderate positive correlation, indicating these two issues often go hand-in-hand.

5.4 Distributions of Numeric Variables

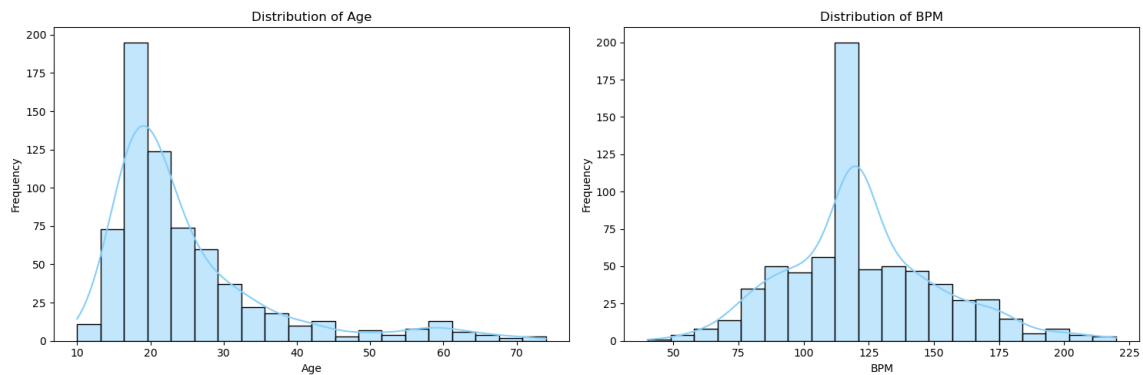


Figure 8: Distribution of Age

Age is negatively skewed. This may be due to the fact that data was collected on an online survey distributed via social media platforms. BPM seems to have a normal distribution.

5.5 Listening Habits

Streaming Platform

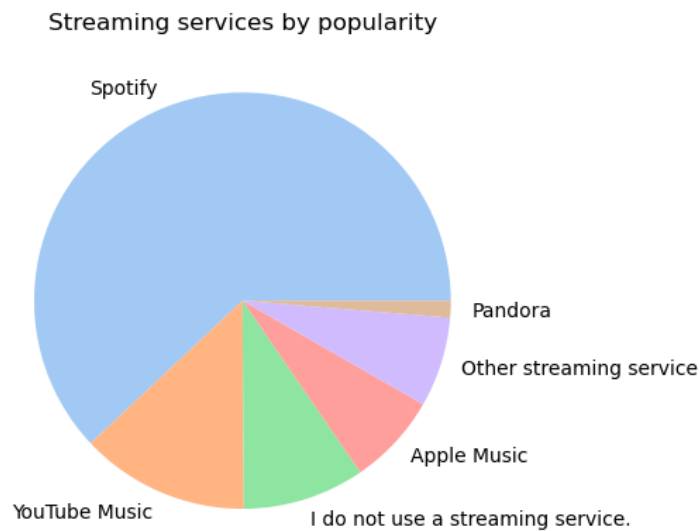


Figure 9: Streaming Platform

Spotify is the most used streaming service while you tube is the second. Pandora is the least used service.

Analysis of Age in Relation to Listening Habits

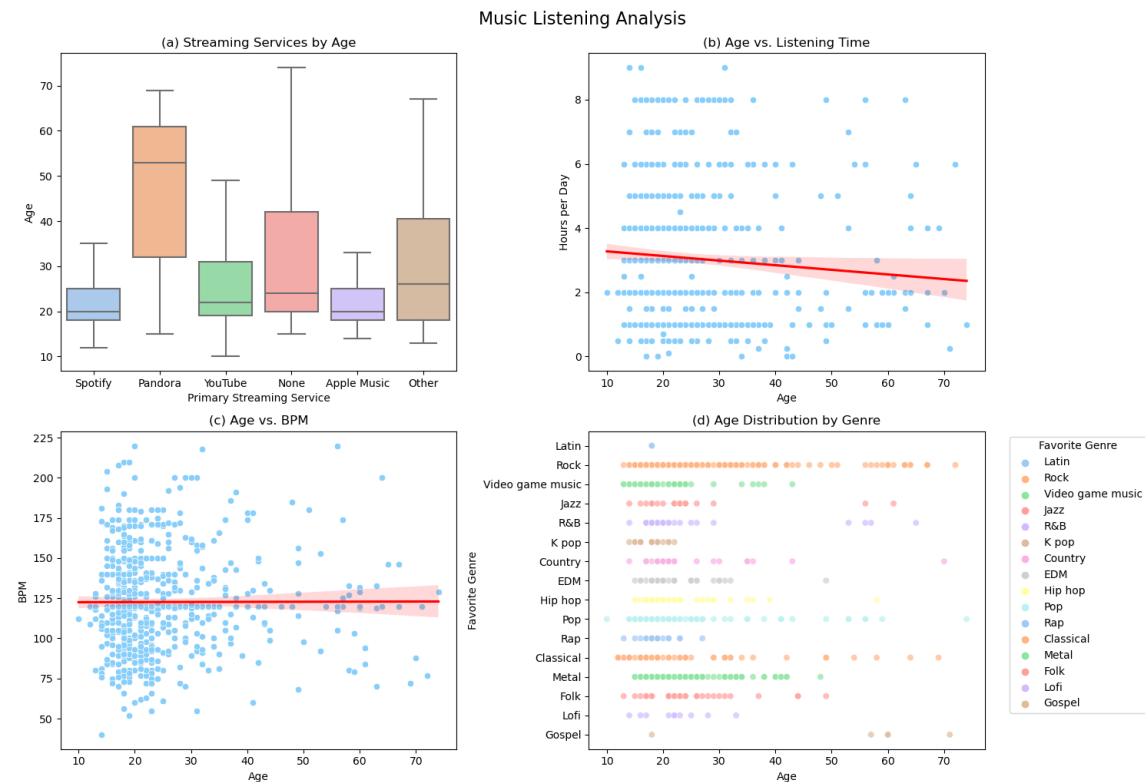


Figure 10: Age vs Listening Habits

Figure 11 reveals notable trends in music streaming preferences, listening habits, and genre choices across different age groups. Subplot (a) shows that major streaming services like Spotify, Apple Music, and YouTube Music are predominantly used by individuals in their 20s and 30s, while Pandora is mainly favored by those over 30. Subplot (b) highlights a slight decrease in average daily listening time with age. Younger individuals tend to spend more time listening to music daily, whereas older age groups show a gradual decline. Subplot (c) indicates that while the mean BPM of preferred songs remains consistent, younger listeners enjoy a broader range of tempos, suggesting a greater openness to different musical styles. Older listeners prefer more consistent tempos. Subplot (d) examines genre popularity among age groups. Rock is popular across all ages, while genres like K-pop, Lofi, and Rap are favored by younger listeners. Interestingly, younger individuals also listen to Folk music more than older ones. Overall, this analysis highlights the diverse musical habits and preferences across age groups, with younger people showing more versatility in their listening choices and older individuals displaying more consistent habits.

5.6 Genres

As we can see in Figure 11 below, some useful insights can be observed when analyzing the genres individuals listen to. Subplot (a) shows that Rock is the most preferred genre among the individuals surveyed. Following Rock, Pop and Metal take the second and third spots on the list of preferred genres, respectively. On the other end of the spectrum, Jazz emerges as the least preferred genre among the participants.

Subplot (b) provides an interesting perspective on how tempo (BPM) vary across different genres. Metal stands out as the genre with the widest range of BPM, indicating a significant diversity in tempo within this genre. In contrast, Gospel music has the narrowest BPM range. When considering the overall picture, it becomes evident that most genres have a median tempo falling within the range of 100 to 140 BPM. This suggests a general preference for mid-tempo music across various genres.

The subplot (c) delves into the listening frequencies of different genres, uncovering some intriguing patterns. Surprisingly, Pop music, despite being popular, is the genre most individuals have never listened to. Gospel music, though listened to by very few, has the least number of individuals who have never listened to it at all. This indicates a small but dedicated listener base for Gospel music. Additionally, when considering all genres, the proportion of individuals who never or rarely listen to a specific genre is higher than those who occasionally or very frequently listen to it. This highlights a trend where a majority of individuals exhibit a broader but less frequent engagement with various music genres.

In conclusion, these insights offer a comprehensive understanding of music preferences and listening habits among individuals. The data suggest that while certain genres like Rock, Pop, and Metal dominate in terms of preference, there is a diverse range of listening habits that vary significantly across genres. The analysis of BPM ranges further enriches our understanding of the rhythmic diversity within these genres, while the examination of listening frequencies sheds light on the engagement levels with different types of music.

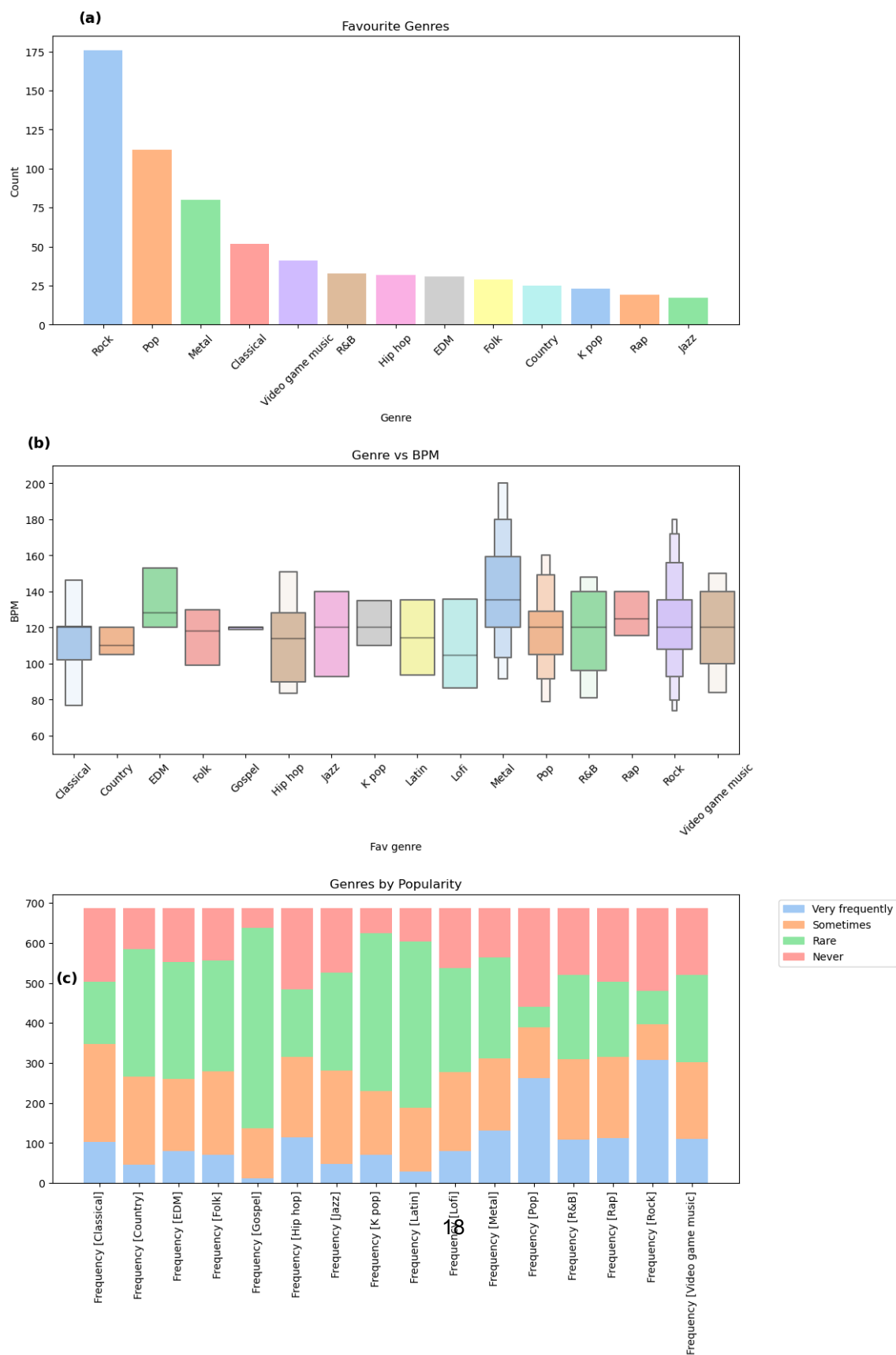


Figure 11: Genre Analysis

5.7 General Behaviours

As we can see in Figure 12 below, subplot (a) shows that most individuals listen to music while working. This suggests that music is commonly used as a tool to enhance concentration or create a more enjoyable work environment. Subplots (b) and (c) indicate that a significant portion of respondents have not got a musical background, such as playing an instrument or composing music. This suggests that many people only consume music and not engage in musical activities, which could influence their listening preferences and behaviors.

Furthermore, Subplot (d) and (e) reveals that the majority of individuals actively explore new music to discover new tastes and interests. This indicates a strong trend towards musical exploration and a desire for variety in listening habits.

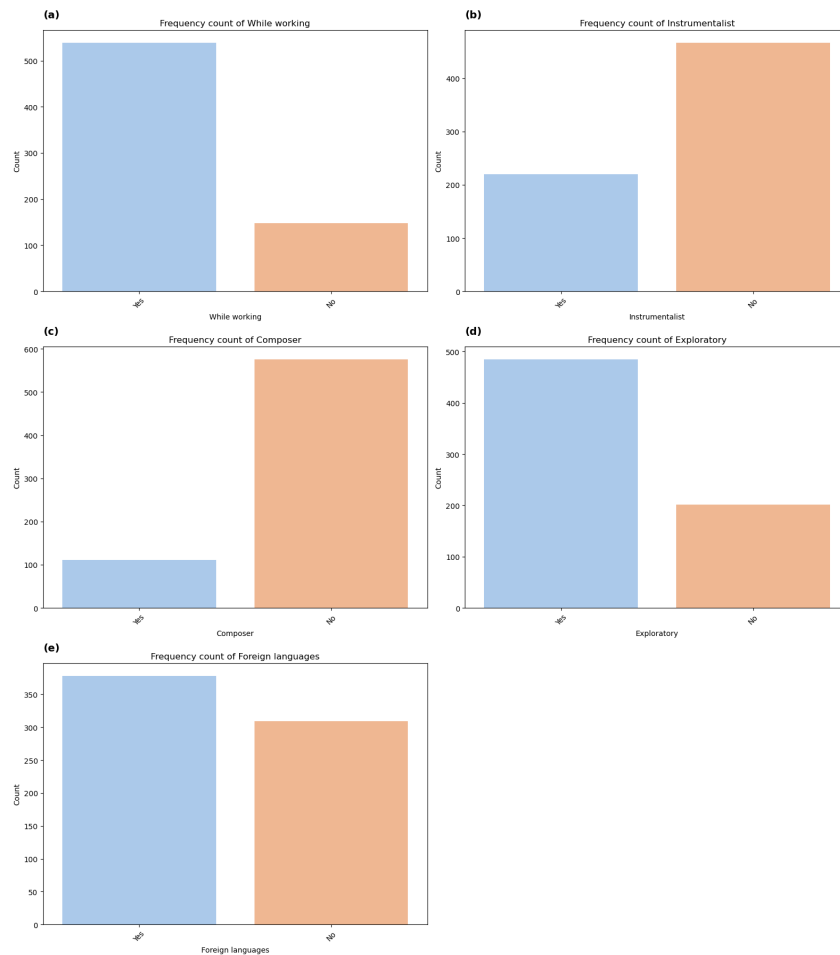


Figure 12: General Behaviours

5.8 Mental Health Conditions

Distribution of Mental Health Conditions

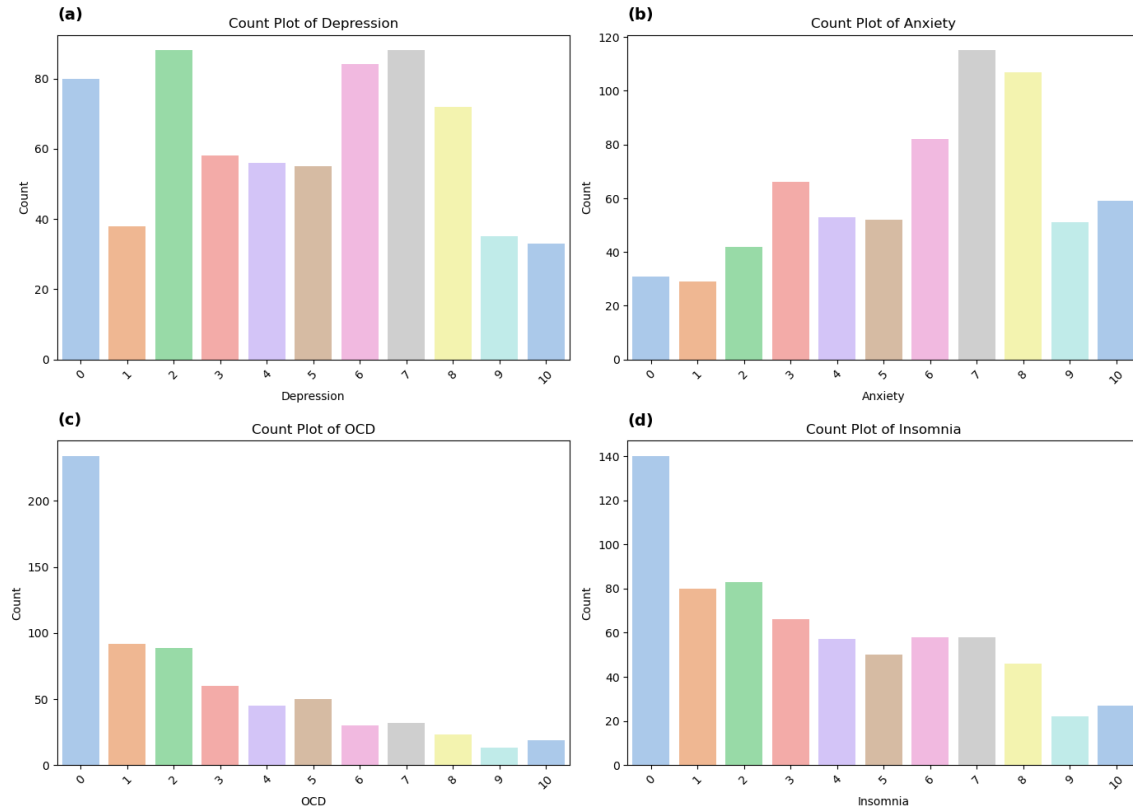


Figure 13: Mental Health Conditions

As we can see in Figure 13 above, the scales for Depression and Anxiety are widely distributed among individuals compared to OCD and Insomnia. Anxiety exhibit a positively skewed distribution, with a larger number of individuals reporting higher scales, indicating more severe cases for these conditions in the sampled population. Conversely, OCD and Insomnia show more uniform distributions with fewer extreme scores. OCD, in particular, tends to have a negatively skewed distribution, where fewer individuals report higher scales compared to Depression and Anxiety. This suggests that OCD symptoms are less prevalent or severe in the sampled population compared to Depression and Anxiety. These insights into the distribution of mental health scales provide a clear understanding of how different conditions are perceived among individuals in the dataset.

Mental Health and Favourite Genre

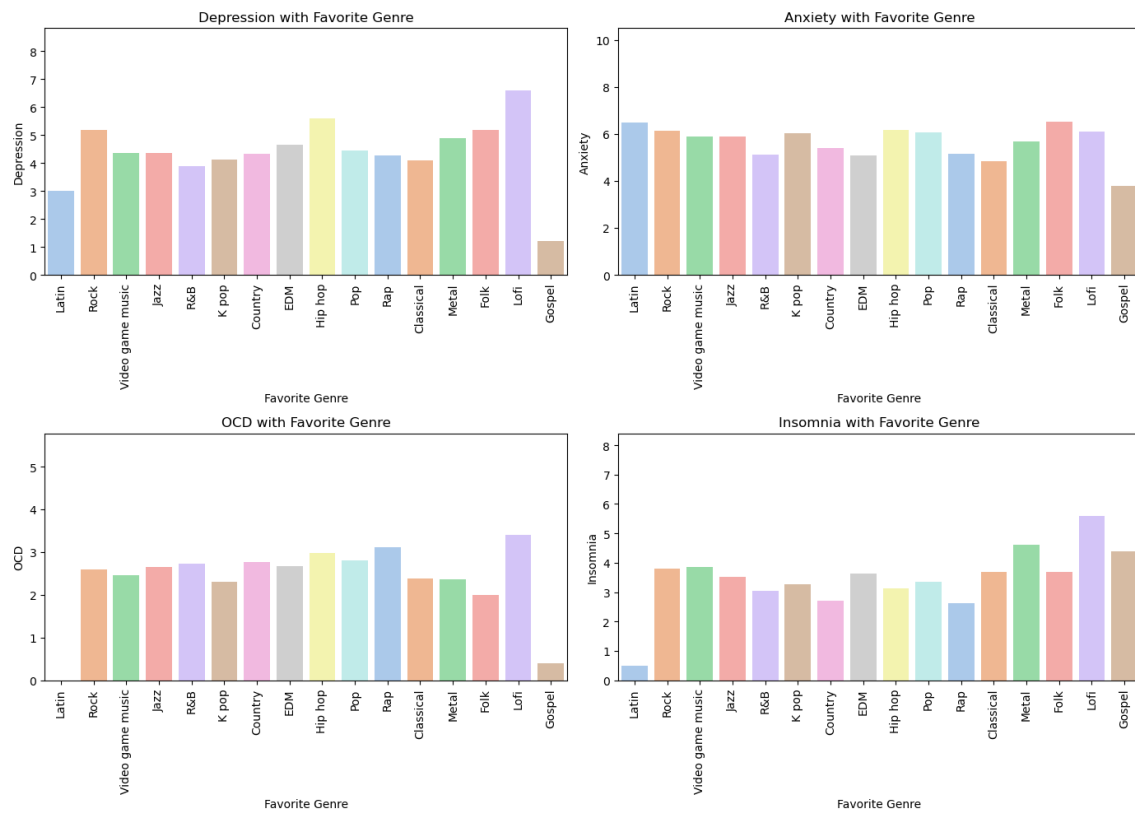


Figure 14: Mental Health and Favourite Genre

As we can see in Figure 14 above, individuals who have Gospel as their favorite genre tend to report lower levels of Depression, Anxiety, and OCD compared to those with other favorite genres. Conversely, individuals who favor Lofi music show higher rates of Depression, Insomnia, and OCD. Apart from these observations, there are no significant patterns evident in the graphs for other genres.

Mental Health and Listening Times

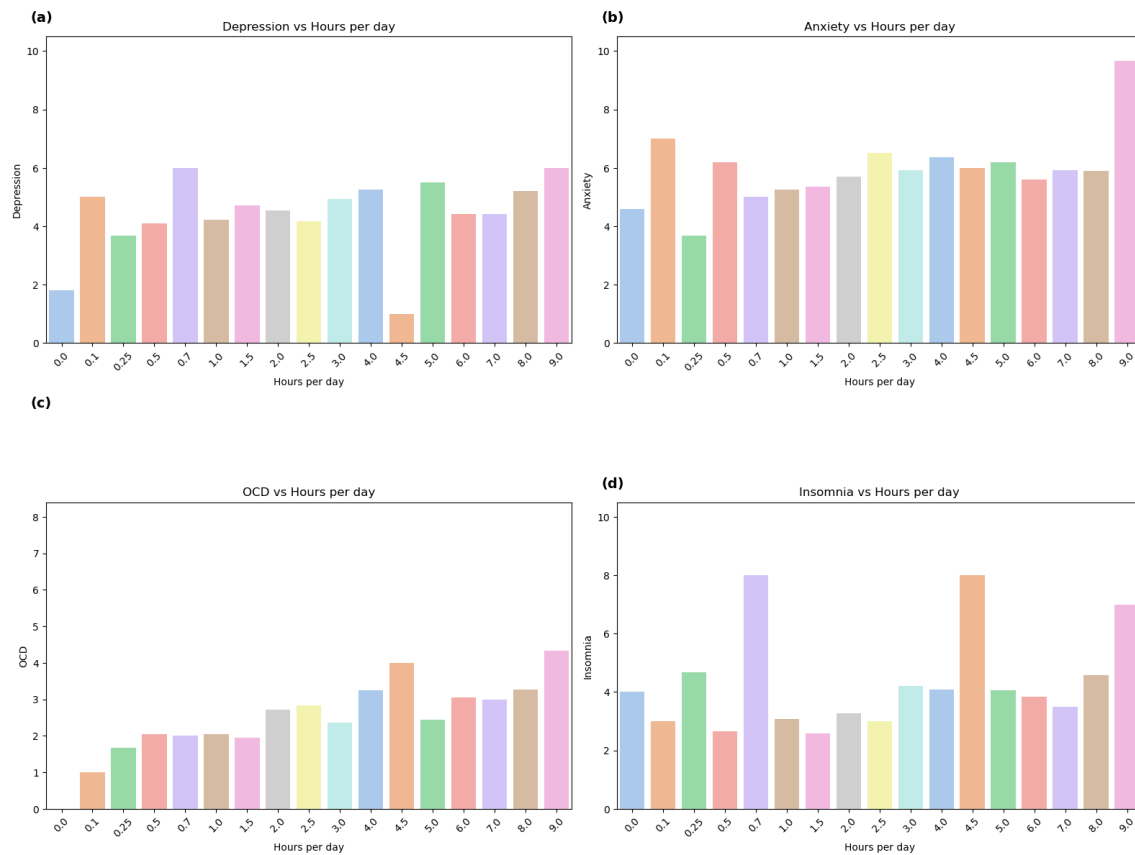


Figure 15: Mental Health and Listening Times

There appears to be a slight increase in the reported scales with increased listening time. This trend is more pronounced for OCD, while other conditions exhibit some irregular patterns. However, this increase is not substantial. Overall, no clear pattern emerges beyond these observations

Effects of Music Listening

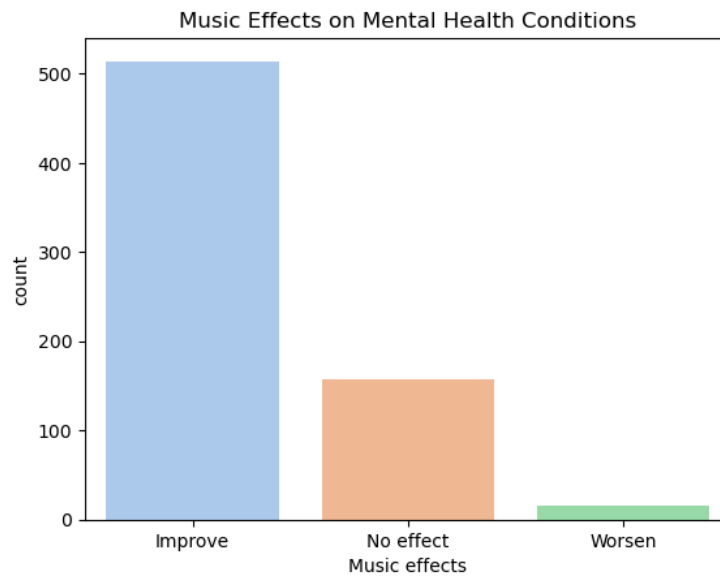


Figure 16: Effects of Music Listening

Here we can see that a majority of the individuals have noticed that their conditions are improving. While there are also a minority of individuals who have noticed worsened conditions. There are also some people who haven't seen any improvement or worsening of their conditions.

Mental Health and Music Background

As we can see in Figure 17 below, there are no significant patterns of mental health conditions being changed due to the musical background of the respondents. The only somewhat significant observation is that the Insomnia rate of Individuals who compose music is slightly higher.

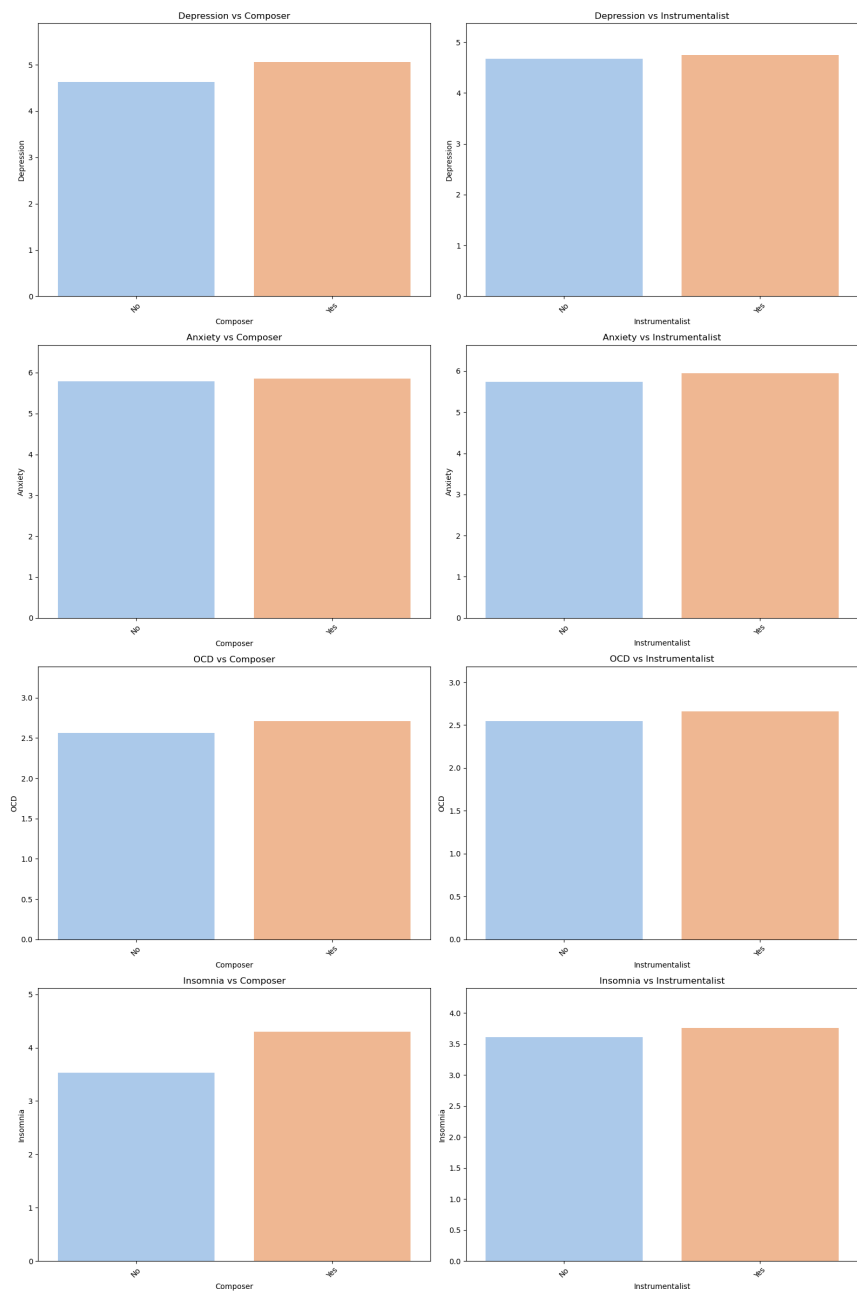


Figure 17: Mental Health and Musical Background

Subset Analysis (Conditions Worsened)

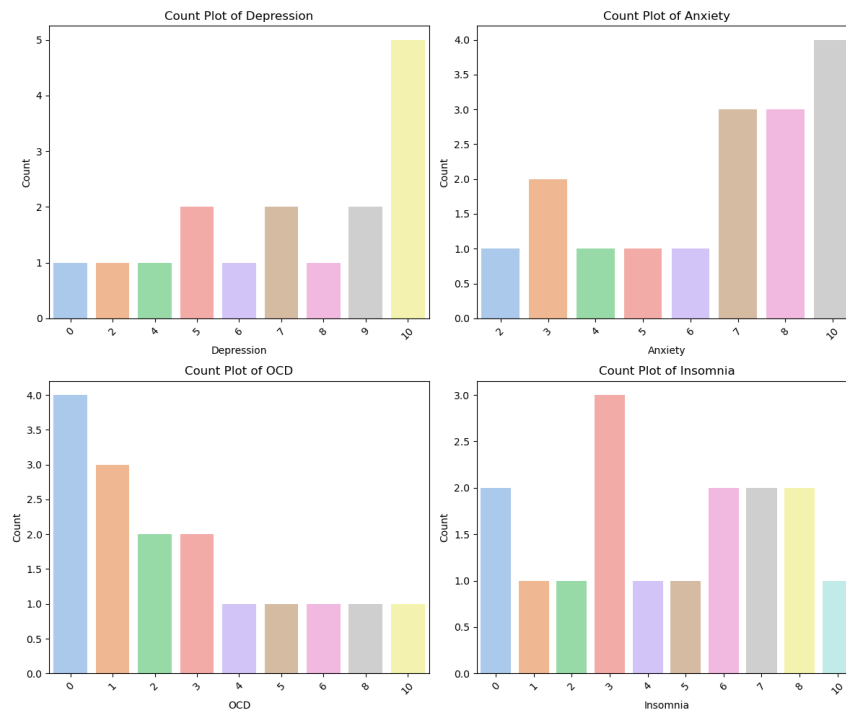


Figure 18: Mental Health Condition of Individuals who have Worsened Conditions

The data frame was filtered to analyse the individuals who have reported that their conditions have worsened with listening to music more deeply. Here we can see that this subset of individuals exhibits extreme cases of Depression and Anxiety. Interestingly, their OCD scales are lower, while Insomnia scales do not depict a clear pattern. This implies that extreme cases of Depression and Anxiety should always be treated with professional healthcare procedures.

Subset Analysis (Conditions Improved)

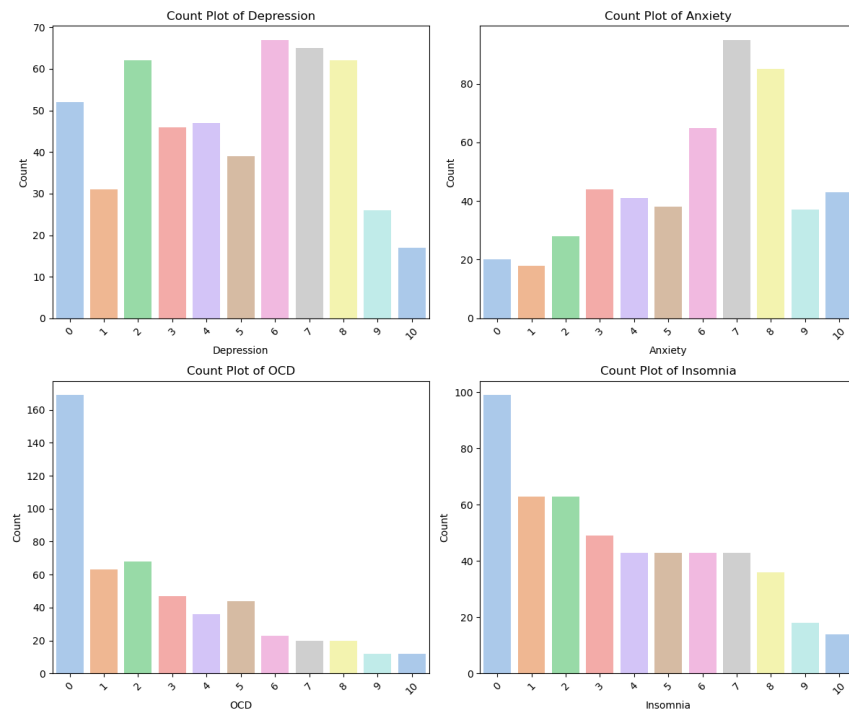


Figure 19: Mental Health Condition of Individuals who have Improved Conditions

The data frame was filtered to analyse the individuals who have reported that their conditions have improved with listening to music more deeply. Here we can see that this subset of individuals exhibits general cases of Depression and Anxiety compared to the ones who reported worsened conditions. OCD and insomnia levels are also lower. This implies that general cases of depression and anxiety can be improved by listening to music.

6 Advanced Analysis for Predictive Modeling

This advanced analysis was conducted to develop a machine learning module aimed at predicting mental health conditions based on other variables. The initial step involved defining the variables and their respective data types as follows:

- **Target Variables:**

- Anxiety
- Depression
- Insomnia
- OCD

- **Feature Columns:**

- **Numerical Variables:** Age, BPM, Hours per day
- **Nominal Variables:** Primary streaming service, While working, Instrumentalist, Composer, Fav genre, Exploratory, Foreign languages, Music effects
- **Ordinal Variables:** Frequency [Classical], Frequency [Country], Frequency [EDM], Frequency [Folk], Frequency [Gospel], Frequency [Hip hop], Frequency [Jazz], Frequency [K pop], Frequency [Latin], Frequency [Lofi], Frequency [Metal], Frequency [Pop], Frequency [R&B], Frequency [Rap], Frequency [Rock], Frequency [Video game music]

As for the next step in preparing the data for machine learning, several transformations are applied to ensure the features are in a suitable format for modeling. Numeric columns undergo scaling, which normalizes their values to a standard range to prevent features with larger scales from dominating the model. Ordinal columns are encoded to preserve their inherent order, facilitating meaningful comparisons within the data. Nominal columns are converted into a binary format through one-hot encoding, creating separate binary features for each category and allowing categorical data to be effectively utilized in machine learning algorithms without implying false ordinal relationships between categories. Then the dataset was split in to training and testing.

6.1 Model 1 - Random Forest Classifier

As for the initial modeling, separate RandomForestClassifier models are trained for each target variable in the dataset. By iterating through each target variable, we define a RandomForestClassifier, fit it with training data, and then store each trained model in a dictionary. Once trained, we proceed to evaluate these models by predicting outcomes on the test dataset. This methodical approach ensures that we can accurately assess the performance and predictive capabilities of each model relative to its respective target variable, facilitating targeted insights into the data's predictive patterns.

Results

Table 2: Classification Report for Anxiety - Model 1

| Class | Precision | Recall | F1-score | Support |
|---------------------|------------------|---------------|-----------------|----------------|
| 0 | 0.00 | 0.00 | 0.00 | 7 |
| 1 | 0.50 | 0.17 | 0.25 | 6 |
| 2 | 0.00 | 0.00 | 0.00 | 11 |
| 3 | 0.00 | 0.00 | 0.00 | 15 |
| 4 | 0.00 | 0.00 | 0.00 | 12 |
| 5 | 0.09 | 0.10 | 0.10 | 10 |
| 6 | 0.23 | 0.19 | 0.21 | 16 |
| 7 | 0.08 | 0.28 | 0.13 | 18 |
| 8 | 0.03 | 0.06 | 0.04 | 18 |
| 9 | 0.00 | 0.00 | 0.00 | 13 |
| 10 | 0.00 | 0.00 | 0.00 | 12 |
| Accuracy | | | 0.08 | 138 |
| Macro avg | 0.08 | 0.07 | 0.07 | 138 |
| Weighted avg | 0.07 | 0.08 | 0.06 | 138 |

Table 3: Classification Report for Depression - Model 1

| Class | Precision | Recall | F1-score | Support |
|---------------------|------------------|---------------|-----------------|----------------|
| 0 | 0.25 | 0.26 | 0.26 | 19 |
| 1 | 0.00 | 0.00 | 0.00 | 6 |
| 2 | 0.14 | 0.17 | 0.15 | 18 |
| 3 | 0.10 | 0.07 | 0.08 | 14 |
| 4 | 0.14 | 0.08 | 0.11 | 12 |
| 5 | 0.29 | 0.22 | 0.25 | 9 |
| 6 | 0.07 | 0.05 | 0.06 | 22 |
| 7 | 0.17 | 0.50 | 0.26 | 12 |
| 8 | 0.11 | 0.14 | 0.12 | 14 |
| 9 | 0.00 | 0.00 | 0.00 | 4 |
| 10 | 0.00 | 0.00 | 0.00 | 8 |
| Accuracy | | | 0.15 | 138 |
| Macro avg | 0.11 | 0.14 | 0.12 | 138 |
| Weighted avg | 0.13 | 0.15 | 0.13 | 138 |

Table 4: Classification Report for Insomnia - Model 1

| Class | Precision | Recall | F1-score | Support |
|---------------------|------------------|---------------|-----------------|----------------|
| 0 | 0.20 | 0.56 | 0.29 | 27 |
| 1 | 0.00 | 0.00 | 0.00 | 13 |
| 2 | 0.20 | 0.09 | 0.13 | 22 |
| 3 | 0.00 | 0.00 | 0.00 | 14 |
| 4 | 0.00 | 0.00 | 0.00 | 8 |
| 5 | 0.00 | 0.00 | 0.00 | 13 |
| 6 | 0.08 | 0.12 | 0.10 | 8 |
| 7 | 0.25 | 0.09 | 0.13 | 11 |
| 8 | 0.00 | 0.00 | 0.00 | 12 |
| 9 | 0.00 | 0.00 | 0.00 | 4 |
| 10 | 0.00 | 0.00 | 0.00 | 6 |
| Accuracy | | | 0.14 | 138 |
| Macro avg | 0.07 | 0.08 | 0.06 | 138 |
| Weighted avg | 0.10 | 0.14 | 0.09 | 138 |

Table 5: Classification Report for OCD - Model 1

| Class | Precision | Recall | F1-score | Support |
|---------------------|------------------|---------------|-----------------|----------------|
| 0 | 0.36 | 0.92 | 0.52 | 49 |
| 1 | 0.33 | 0.13 | 0.19 | 15 |
| 2 | 0.00 | 0.00 | 0.00 | 15 |
| 3 | 0.00 | 0.00 | 0.00 | 11 |
| 4 | 0.00 | 0.00 | 0.00 | 8 |
| 5 | 0.00 | 0.00 | 0.00 | 11 |
| 6 | 0.00 | 0.00 | 0.00 | 7 |
| 7 | 0.00 | 0.00 | 0.00 | 6 |
| 8 | 0.00 | 0.00 | 0.00 | 6 |
| 9 | 0.00 | 0.00 | 0.00 | 5 |
| 10 | 0.00 | 0.00 | 0.00 | 5 |
| Accuracy | | | 0.34 | 138 |
| Macro avg | 0.06 | 0.10 | 0.06 | 138 |
| Weighted avg | 0.16 | 0.34 | 0.20 | 138 |

The observed poor performance of the model can be attributed to several factors that commonly affect machine learning models. Firstly, class imbalance in the target variables is a significant issue where one class is heavily represented compared to others. This imbalance can cause the model to bias predictions towards the majority class, resulting in lower accuracy and predictive power for minority classes. To overcome this, Reducing the class count was considered next.

6.2 Model 2 - Random Forest Classifier with Reduced Response Class Counts

Instead of considering 10 factor levels for each target variable, K-means clustering was carried out using elbow plots to identify possible clusters within them. As can be seen from the elbow plots, each target variable was found to have an optimal number of 3 clusters. This optimal number was determined to best represent the natural groupings within the data. Subsequently, the target variables were divided into three distinct clusters labeled as "Low," "Medium," and "High." This clustering approach was adopted to better identify and capture the inherent patterns and natural groupings within the training set, and to reduce the complexity associated with handling a large number of classes. After establishing these clusters, the same modeling procedure was carried out.

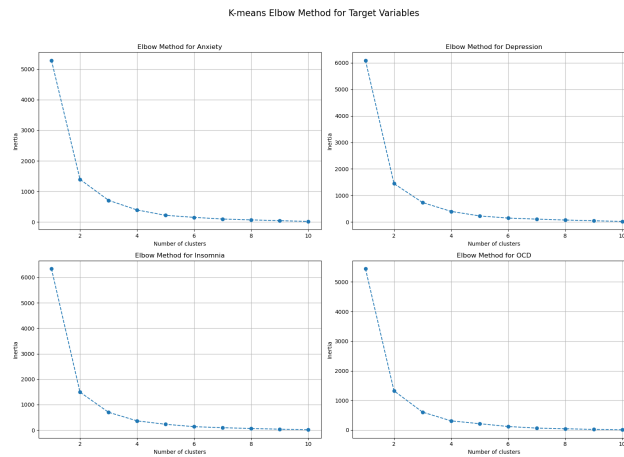


Figure 20: Elbow Plots for Target Variables

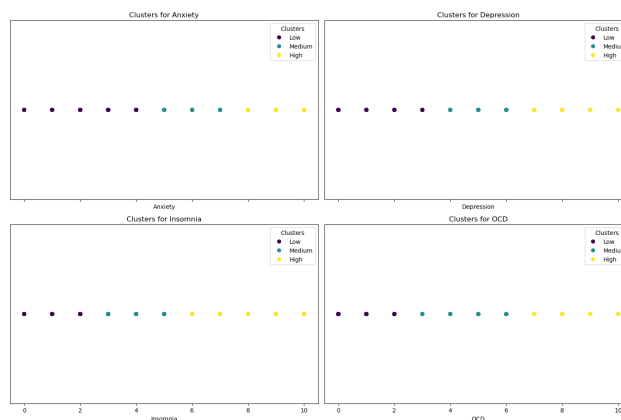


Figure 21: Natural Clusters within Target Variables

Results

Table 6: Classification Report for Anxiety - Model 2

| | precision | recall | f1-score | support |
|---------------------|------------------|---------------|-----------------|----------------|
| High | 0.42 | 0.37 | 0.40 | 43 |
| Low | 0.56 | 0.37 | 0.45 | 51 |
| Medium | 0.39 | 0.59 | 0.47 | 44 |
| accuracy | 0.44 | | | |
| macro avg | 0.46 | 0.45 | 0.44 | 138 |
| weighted avg | 0.46 | 0.44 | 0.44 | 138 |

Table 7: Classification Report for Depression - Model 2

| | precision | recall | f1-score | support |
|---------------------|------------------|---------------|-----------------|----------------|
| High | 0.42 | 0.66 | 0.51 | 38 |
| Low | 0.53 | 0.58 | 0.55 | 57 |
| Medium | 0.50 | 0.19 | 0.27 | 43 |
| accuracy | 0.48 | | | |
| macro avg | 0.48 | 0.47 | 0.45 | 138 |
| weighted avg | 0.49 | 0.48 | 0.45 | 138 |

Table 8: Classification Report for Insomnia - Model 2

| | precision | recall | f1-score | support |
|---------------------|------------------|---------------|-----------------|----------------|
| High | 0.47 | 0.41 | 0.44 | 41 |
| Low | 0.47 | 0.71 | 0.56 | 62 |
| Medium | 0.38 | 0.09 | 0.14 | 35 |
| accuracy | 0.46 | | | |
| macro avg | 0.44 | 0.40 | 0.38 | 138 |
| weighted avg | 0.45 | 0.46 | 0.42 | 138 |

Table 9: Classification Report for OCD - Model 2

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.00 | 0.00 | 0.00 | 22 |
| Low | 0.57 | 0.99 | 0.73 | 79 |
| Medium | 0.00 | 0.00 | 0.00 | 37 |
| accuracy | | 0.57 | | |
| macro avg | 0.19 | 0.33 | 0.24 | 138 |
| weighted avg | 0.33 | 0.57 | 0.42 | 138 |

Classification Report for Anxiety - Model 2

- **High Anxiety:** The precision is 0.42, recall is 0.37, and the F1-score is 0.40. This indicates moderate identification accuracy for high anxiety cases.
- **Low Anxiety:** The precision is 0.56, recall is 0.37, and the F1-score is 0.45. This suggests slightly better balance between precision and recall.
- **Medium Anxiety:** The precision is 0.39, recall is 0.59, and the F1-score is 0.47. The model is better at identifying medium anxiety cases.
- **Overall Metrics:** The accuracy is 0.44. The macro and weighted averages show moderate performance across all classes.

Classification Report for Depression - Model 2

- **High Depression:** The precision is 0.42, recall is 0.66, and the F1-score is 0.51. This shows reasonable identification of high depression cases.
- **Low Depression:** The precision is 0.53, recall is 0.58, and the F1-score is 0.55. The model has a balanced identification for low depression.
- **Medium Depression:** The precision is 0.50, recall is 0.19, and the F1-score is 0.27. The model struggles with identifying medium depression cases.
- **Overall Metrics:** The accuracy is 0.48. The macro and weighted averages indicate balanced performance across all classes.

Classification Report for Insomnia - Model 2

- **High Insomnia:** The precision is 0.47, recall is 0.41, and the F1-score is 0.44. This indicates moderate identification for high insomnia cases.
- **Low Insomnia:** The precision is 0.47, recall is 0.71, and the F1-score is 0.56. The model identifies low insomnia cases well.
- **Medium Insomnia:** The precision is 0.38, recall is 0.09, and the F1-score is 0.14. The model struggles with medium insomnia cases.

- **Overall Metrics:** The accuracy is 0.46. The macro and weighted averages show moderate performance across all classes.

Classification Report for OCD - Model 2

- **High OCD:** The precision, recall, and F1-score are all 0.00, indicating the model fails to identify high OCD cases.
- **Low OCD:** The precision is 0.57, recall is 0.99, and the F1-score is 0.73. The model performs excellently for low OCD cases.
- **Medium OCD:** The precision, recall, and F1-score are all 0.00, indicating the model fails to identify medium OCD cases.
- **Overall Metrics:** The accuracy is 0.57. The macro averages indicate poor performance, while the weighted averages show moderate performance due to strong identification of low OCD cases.

We can see a significant improvement in results; nevertheless, given the study's objectives, this level of performance is still considered sub-optimal. In an attempt to boost our outcomes, I strategically reduced the class count (factor levels) within the frequency columns to three using a properly created mapping dictionary. Despite these efforts, did not observe the anticipated significant enhancements in performance as it gives almost the same classification report

As the next step Binary Classification of the target variable was considered using Kmeans clustering.

6.3 Model 3 - Random Forest Classifier with only 2 Response Classes

Following a similar approach as in Model 2, K-means clustering was applied to segment the target variable into two distinct classes: 'Low' and 'High'. Subsequently, the exact modeling procedure was executed.

Results

Table 10: Classification Report for Anxiety - Model 3

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.58 | 0.88 | 0.70 | 77 |
| Low | 0.57 | 0.20 | 0.29 | 61 |
| accuracy | | | 0.58 | |
| macro avg | 0.58 | 0.54 | 0.50 | 138 |
| weighted avg | 0.58 | 0.58 | 0.52 | 138 |

Table 11: Classification Report for Depression - Model 3

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.56 | 0.77 | 0.65 | 69 |
| Low | 0.63 | 0.39 | 0.48 | 69 |
| accuracy | 0.58 | | | |
| macro avg | 0.59 | 0.58 | 0.56 | 138 |
| weighted avg | 0.59 | 0.58 | 0.56 | 138 |

Table 12: Classification Report for Insomnia - Model 3

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.59 | 0.19 | 0.28 | 54 |
| Low | 0.64 | 0.92 | 0.75 | 84 |
| accuracy | 0.63 | | | |
| macro avg | 0.61 | 0.55 | 0.52 | 138 |
| weighted avg | 0.62 | 0.63 | 0.57 | 138 |

Table 13: Classification Report for OCD - Model 3

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.33 | 0.02 | 0.04 | 48 |
| Low | 0.65 | 0.98 | 0.78 | 90 |
| accuracy | 0.64 | | | |
| macro avg | 0.49 | 0.50 | 0.41 | 138 |
| weighted avg | 0.54 | 0.64 | 0.52 | 138 |

Classification Report for Anxiety - Model 3

- **High Anxiety:** The precision is 0.58, recall is 0.88, and the F1-score is 0.70. This indicates good identification accuracy for high anxiety cases.
- **Low Anxiety:** The precision is 0.57, recall is 0.20, and the F1-score is 0.29. This suggests that the model struggles with identifying low anxiety cases.
- **Overall Metrics:** The accuracy is 0.58. The macro and weighted averages show moderate performance across all classes, with a better balance in identifying high anxiety cases.

Classification Report for Depression - Model 3

- **High Depression:** The precision is 0.56, recall is 0.77, and the F1-score is 0.65. This shows reasonable identification of high depression cases.
- **Low Depression:** The precision is 0.63, recall is 0.39, and the F1-score is 0.48. The model struggles with identifying low depression cases.

- **Overall Metrics:** The accuracy is 0.58. The macro and weighted averages indicate balanced performance across all classes.

Classification Report for Insomnia - Model 3

- **High Insomnia:** The precision is 0.59, recall is 0.19, and the F1-score is 0.28. This indicates moderate identification for high insomnia cases.
- **Low Insomnia:** The precision is 0.64, recall is 0.92, and the F1-score is 0.75. The model identifies low insomnia cases well.
- **Overall Metrics:** The accuracy is 0.63. The macro and weighted averages show moderate performance across all classes, with better identification for low insomnia cases.

Classification Report for OCD - Model 3

- **High OCD:** The precision is 0.33, recall is 0.02, and the F1-score is 0.04, indicating poor identification accuracy for high OCD cases.
- **Low OCD:** The precision is 0.65, recall is 0.98, and the F1-score is 0.78. The model performs excellently for low OCD cases.
- **Overall Metrics:** The accuracy is 0.64. The macro averages indicate poor performance, while the weighted averages show moderate performance due to strong identification of low OCD cases.

We can see a significant improvement when we only considered 2 response classes to each target variable. As the next step of improving performance hyper parameter tuning with cross validation was considered.

6.4 Model 4 - Random Forest Classifier with 2 Response Classes and Cross Validation

Hyper Parameter Tunings were given to the model 3 using the following parameter grid.

```
param_grid = {
    'n_estimators': [50, 100, 150],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

Apart from this modeling procedure was carried out the same.

Results

There was a slight improvement of the results with the implementation of cross validation. The results and interpretations are as follows.

Table 14: Best Parameters and CV Accuracy for Each Target Variable

| Target Variable | Best Parameters | Best CV Accuracy |
|-----------------|---|------------------|
| Anxiety | {max_depth: 20, min_samples_leaf: 2, min_samples_split: 5, n_estimators: 100} | 0.6114 |
| Depression | {max_depth: 10, min_samples_leaf: 4, min_samples_split: 2, n_estimators: 150} | 0.6128 |
| Insomnia | {max_depth: None, min_samples_leaf: 2, min_samples_split: 2, n_estimators: 100} | 0.6274 |
| OCD | {max_depth: 10, min_samples_leaf: 4, min_samples_split: 2, n_estimators: 50} | 0.6958 |

Table 15: Classification Report for Anxiety - Model 4

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.58 | 0.90 | 0.70 | 77 |
| Low | 0.58 | 0.18 | 0.28 | 61 |
| accuracy | 0.58 | | | |
| macro avg | 0.58 | 0.54 | 0.49 | 138 |
| weighted avg | 0.58 | 0.58 | 0.51 | 138 |

Table 16: Classification Report for Depression - Model 4

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.56 | 0.75 | 0.64 | 69 |
| Low | 0.62 | 0.41 | 0.49 | 69 |
| accuracy | 0.58 | | | |
| macro avg | 0.59 | 0.58 | 0.57 | 138 |
| weighted avg | 0.59 | 0.58 | 0.57 | 138 |

Table 17: Classification Report for Insomnia - Model 4

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.76 | 0.24 | 0.37 | 54 |
| Low | 0.66 | 0.95 | 0.78 | 84 |
| accuracy | 0.67 | | | |
| macro avg | 0.71 | 0.60 | 0.57 | 138 |
| weighted avg | 0.70 | 0.67 | 0.62 | 138 |

Table 18: Classification Report for OCD - Model 4

| | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| High | 0.00 | 0.00 | 0.00 | 48 |
| Low | 0.65 | 1.00 | 0.79 | 90 |
| accuracy | 0.65 | | | |
| macro avg | 0.33 | 0.50 | 0.39 | 138 |
| weighted avg | 0.43 | 0.65 | 0.51 | 138 |

Classification Report for Anxiety - Model 4

- **Best CV Accuracy:** 0.6114
- **Test Accuracy:** 0.58
- **Interpretation:** The best CV accuracy indicates the model performed reasonably well during the cross-validation phase. However, the test accuracy is slightly lower, suggesting some overfitting or variability in the test data. The classification report show that the model is better at identifying the "High" anxiety class compared to the "Low" class.

Classification Report for Depression - Model 4

- **Best CV Accuracy:** 0.6128
- **Test Accuracy:** 0.58
- **Interpretation:** Similar to Anxiety, the best CV accuracy is higher than the test accuracy, indicating a potential overfitting issue. The model performs reasonably well in identifying the "High" depression class, but struggles with the "Low" class, as seen in the precision, recall, and F1-scores.

Classification Report for Insomnia - Model 4

- **Best CV Accuracy:** 0.6274
- **Test Accuracy:** 0.67
- **Interpretation:** For Insomnia, the test accuracy is higher than the best CV accuracy, suggesting that the model generalizes well to the test data. The confusion matrix and classification report indicate that the model is effective at identifying the "Low" insomnia class, but has more difficulty with the "High" class.

Classification Report for OCD - Model 4

- **Best CV Accuracy:** 0.6958
- **Test Accuracy:** 0.65
- **Interpretation:** The best CV accuracy is higher than the test accuracy, indicating potential overfitting. The model performs poorly in identifying the "High" OCD class, as reflected in the precision, recall, and F1-scores of 0. The "Low" class is predicted more accurately, but the imbalance between the classes is evident.

7 General Discussion and Conclusions

- **Exploratory Data Analysis (EDA):** The EDA provided valuable insights into various aspects of music listening behaviors and their association with mental health. Key observations included prevalent music listening platforms, typical listening habits, and how these habits vary with demographic factors such as age. Notably, the analysis highlighted the relationship between music listening and mental health conditions. For instance, individuals with mild to moderate levels of anxiety and depression reported improvements in their conditions through music listening. Conversely, those with severe anxiety and depression experienced a worsening of symptoms, underscoring the importance of professional treatment for individuals diagnosed with extreme cases of these conditions.
- **Advanced Analysis and Predictive Modeling:** The advanced analysis, while not yielding exceptional performance, demonstrated a general efficacy in predicting mental health conditions based on music listening habits. The best-performing model showed reasonable accuracy in predicting high anxiety, high depression, low insomnia, and low OCD classes. However, the performance was hampered by class imbalance within the dataset. Techniques such as oversampling and undersampling were not employed due to the limited dataset size, which could lead to repetitive samples. Similarly, synthetic methods like SMOTE were avoided to maintain the dataset's originality and integrity.
- **Conclusions:** This study provides a comprehensive understanding of the interplay between music listening habits and mental health conditions. The findings suggest that music, an accessible and universally available medium, can play a beneficial role in alleviating symptoms of certain mental health conditions. However, for severe mental health issues, professional intervention remains crucial. Overall, this study highlights the potential of music as a complementary therapeutic tool, promoting mental well-being among the general population.

References

- [1] World Health Organization. (2021). Mental health. Retrieved from <https://www.government.nl/topics/mhpss/mhpss-worldwide-facts-and-figures#:~:text=Mental%20health%20in%20general,people%20worldwide%20suffer%20from%20depression.>
- [2] Effects of Listening to Music on Mental Health of Nigerian Undergraduate Students. Retrieved from <https://journalindj.com/index.php/INDJ/article/view/390>
- [3] Mental Health & Music Relationship Analysis & EDA. Retrieved from <https://www.kaggle.com/code/melissamonfared/mental-health-music-relationship-analysis-eda#Visualization>
- [4] Predicting Mental Health from Music Taste. Retrieved from <https://www.kaggle.com/code/catherinerasgaitis/predicting-mental-health-from-music-taste#Model-2:-Complete-Rankings,-LazyPredict-+-XGBoost>
- [5] RandomForestClassifier from scikitlearn <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>