Harmonies of the Mind: Investigating the Interplay Between Musical Preferences and Mental Well-being

Dieu-Anh Le, Rohith Raj Srinivasan

Code: https://github.com/rohis06/TTP_Project

1. Introduction

Music therapy has gained attention as a promising intervention for enhancing mental health and well-being. By harnessing the therapeutic potential of music, music therapy offers a non-invasive and accessible approach to addressing mental health conditions. Research has demonstrated its effectiveness in improving mood, reducing stress, and promoting relaxation, making it a valuable adjunct to traditional psychotherapeutic interventions [1, 2]. Despite the growing recognition of music therapy's benefits, there remains a gap in understanding how individual music preferences may influence its outcomes. Music is a deeply personal and subjective experience, with individuals gravitating toward specific genres, artists, and songs based on their unique preferences, backgrounds, and emotional associations. These preferences may not only reflect personal taste but also serve as coping mechanisms or sources of emotional resonance [3]. However, the relationship between music preference and mental health outcomes within the context of music therapy has been relatively underexplored [4]. While some studies have investigated the effects of specific music genres or styles on mood regulation and emotional well-being, few have examined how individual differences in music taste may impact the effectiveness of music therapy interventions. This report delves into the statistical analysis and modeling techniques used to explore these relationships, aiming to uncover insights into how music might affect mental well-being and the potential individual characteristics that could predict these effects.

2. Background and Motivation

Music plays an important role in our daily routines, influencing activities like work, exercise, and relaxation. It significantly impacts mood and stress levels, making it a potential tool for emotional regulation. With rising mental health issues, it is crucial to explore music's effects on an individual's mental health. Understanding this interplay between music preference and mental well-being could have significant implications for the development of personalized interventions and the optimization of music therapy practices. By identifying which music genres or styles are most closely associated with positive mental health outcomes, clinicians and therapists may be better equipped to tailor their interventions to individual preferences, thereby enhancing engagement, efficacy, and client satisfaction. Motivated by this gap in research, our project aims to analyze the MxMH dataset available on Kaggle to uncover potential correlations between respondents' music preferences and self-reported mental health indicators.

With limited longitudinal data, we are unable to explore the treatment effects of music therapy on mental health conditions. Nonetheless, through rigorous data analysis and statistical modeling, we seek to explore the complex correlation between music preferences and mental well-being, ultimately contributing to the advancement of evidence-based practices in music therapy and mental health care.

3. Research Questions

To systematically investigate the interplay between musical preferences and mental well-being, we have formulated three key research questions:

RQ1: Besides streaming music, are "musicians" associated with better mental health conditions? To address this question, we will compare the mental health conditions of musicians and non-musicians. Comparative testing approaches, including t-tests and Analysis of Variance (ANOVA), will be employed to determine if there are significant differences in mental health outcomes between these groups. Detailed methodologies and findings related to this question are outlined in Section 5.1.

RQ2: What are the individual characteristics associated with mental well-being? This question seeks to identify specific individual characteristics that correlate with mental well-being. We will perform regression modeling and inferential analysis using various feature selection techniques to pinpoint the most influential factors. The approach and results of this analysis are discussed in Section 5.2.

RQ3: Can individual characteristics predict the music effects? If so, how accurately? The final research question explores the predictive power of individual characteristics on the effects of music. Using advanced machine learning models such as Linear Discriminant Analysis (LDA), Random Forest, and Neural Networks, we will assess the accuracy of these predictions. The methodologies and results of this exploration are detailed in Section 5.3.

Our project contributed the following findings:

- Provided a comprehensive insight of the dataset, identifying early patterns and trends between potential predictors and response variables through detailed exploratory data analysis.
- Used stepwise modeling in data preprocessing phase to predict missing values for beats per minutes (BPM) of respondent favorite genre without affecting the multicollinearity aspect of the model.
- Comparative analysis suggested no statistical difference between musicians and non-musicians
- Inferential modeling revealed that different favorite music genres and listening frequencies are associated with different mental health conditions.
- Generalized Additive Modeling explored the non-linear relationship between age and hours of listening to music per day and levels of mental health.
- Predictive modeling assessed the effectiveness of individual characteristics in the dataset in predicting music effectiveness on mental health.

4. Dataset

4.1. Overview

Our dataset is a mental health survey that was conducted through Google Forms with no age or geographical location restriction, sourced from Kaggle. The form was distributed through online platforms such as Reddit, Discord, and other social media channels, as well as offline venues, such as libraries, parks, and public spaces, via

posters and "business cards" to reach a diverse demographic. The survey was designed to be brief and open-ended, encompassing a range of inquiries regarding both music and mental health, with the overarching objective of exploring the potential association between musical preferences and mental well-being. The dataset contains 736 entries with 33 different variables. The variable "Frequency_<Genre>" represents 16 distinct variables corresponding to the frequency of listening to each of the 16 music genres. Table 1 provides an overview of our dataset, with the last five variables being response variables and the remaining variables serving as predictors.

Variable	Data Type	Description
Age	Continuous numeric	Age of survey respondents
Primary_streaming_service	Categorical	Where they primarily stream from
Hours_per_day	Continuous numeric	Number of hours listening to music per day
While_working	Boolean	Do they listen to music while working or studying?
Instrumentalist	Boolean	Do they play an instrument regularly?
Composer	Boolean	Do they compose music?
Fav_genre	Categorical	Favorite or top genre from 16 genres listed below
Exploratory	Boolean	Do they actively listen to new artists or genres?
Foreign_language	Boolean	Do they listen to music in languages that they are not fluent in?
BPM	Continuous numeric	Beats per minute of their favorite genre
Frequency_<Genre>	Categorical / Discrete numeric	Subjective ranking {"Never," "Rarely," "Sometimes," "Very Frequently"} on how often they listen to each of the genres {Classical, Country, EDM, Folk, Gospel, HipHop, Jazz, Kpop, Latin, Lofi, Metal, Pop, RB, Rap, Rock, Video game music}
Anxiety	Discrete numeric	Subjective ranking on anxiety level (scale 0 (none) to 10 (severe))
Depression	Discrete numeric	Subjective ranking on depression level (scale 0 (none) to 10 (severe))
Insomnia	Discrete numeric	Subjective ranking on insomnia level (scale 0 (none) to 10 (severe))
OCD	Discrete numeric	Subjective ranking on OCD level (scale 0 (none) to 10 (severe))
Music_effects	Categorical	Does music improve or worsen their health condition? (Improve / No effect / Worsen)

Table 1: Description of all variables in the dataset

4.2. Data Preprocessing

Before delving into the analysis, we first need to preprocess the data. Figure 1 shows the heatmap of all missing entries from the survey. In total, there were 129 missing entries, 107 of which belonged to the BPM column. Table 2 and 3 provides more details on these missing values along with the summary statistics for each of the explanatory variables. Additionally, some categorical variables intended to be boolean have more than two unique values with the third category being empty spaces (""), which we considered as missing values.

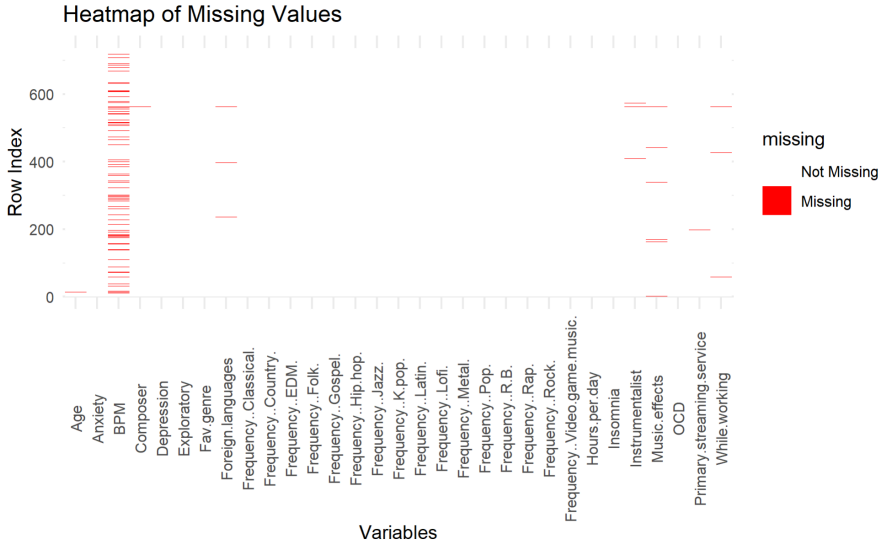


Figure 1: Heatmap of missing values

	Min	Q1	Median	Q3	Max	Mean	Mode	Missing Count
Age	10	18	21	28	89	25.2	18	1
Hours_per_day	0	2	3	5	24	3.57	2	0
BPM	0	100	120	144	999999999	1589948	120	107
Anxiety	0	4	6	8	10	5.84	7	0
Depression	0	2	5	7	10	4.80	7	0
Insomnia	0	1	3	6	10	3.74	0	0
OCD	0	0	2	5	10	2.64	0	0

Table 2: Summary statistics of numerical variables

	Unique Categories	Number of Unique Categories	Most Common Category	Most Frequent Category Count	Missing Count		Unique Categories	Number of Unique Categories	Most Common Category	Most Frequent Category Count	Missing Count
Primary_streaming_service	Spotify, YouTube_Music, Apple_Music, Pandora, Other_streaming_service, I_do_not_use_a_streaming_service	7	Spotify	458	1	**Fav_genre**	Classical, Country, EDM, Folk, Gospel, Hip_hop, Jazz, K_pop, Latin, Lofi, Metal, Pop, R&B, Rap, Rock, Video_game_music	16	Rock	188	0
While_working	Yes, No	3	Yes	579	3	**Frequency_Country**	Never, Rarely, Sometimes, Very_frequently	4	Never	343	0
Instrumentalist	Yes, No	3	No	497	4	**Frequency_EDM**	Never, Rarely, Sometimes, Very_frequently	4	Never	307	0
Composer	Yes, No	3	No	609	1	**Frequency_Folk**	Never, Rarely, Sometimes, Very_frequently	4	Never	292	0
Exploratory	Yes, No	2	Yes	525		**Frequency_Gospel**	Never, Rarely, Sometimes, Very_frequently	4	Never	535	0
Foreign_languages	Yes, No	3	Yes	404	0	**Frequency_Hip_hop**	Never, Rarely, Sometimes, Very_frequently	4	Sometimes	218	0
Music_effects	Improve, No_effect, Worsen	4	Improve	542	4	**Frequency_Jazz**	Never, Rarely, Sometimes, Very_frequently	4	Never	261	0
Frequency_Classical	Never, Rarely, Sometimes, Very_frequently	4	Rarely	259	0	**Frequency_Latin**	Never, Rarely, Sometimes, Very_frequently	4	Never	443	0
Frequency_Video_game_music	Never, Rarely, Sometimes, Very_frequently	4	Never	236	0	**Frequency_K_pop**	Never, Rarely, Sometimes, Very_frequently	4	Never	416	0
Frequency_R&B	Never, Rarely, Sometimes, Very_frequently	4	Never	225	0	**Frequency_Lofi**	Never, Rarely, Sometimes, Very_frequently	4	Never	280	0
Frequency_Rap	Never, Rarely, Sometimes, Very_frequently	4	Rarely	215	0	**Frequency_Metal**	Never, Rarely, Sometimes, Very_frequently	4	Never	264	0
Frequency_Rock	Never, Rarely, Sometimes, Very_frequently	4	Very_frequently	330	0	**Frequency_Pop**	Never, Rarely, Sometimes, Very_frequently	4	Very_frequently	277	0

Table 3: Summary statistics of categorical variables

After a careful examination of the dataset, we suspected that the maximum value of BPM, as seen in Table 2, was a typo and will treat it as an outlier for removal in the data preprocessing phase. Upon further inspection, we also decided to drop the second highest BPM value for a similar reason. Due to the high number of missing entries in BPM, dropping all these values would drastically reduce the size of our dataset. Therefore, we decided to drop all the non-BPM missing values. We then divided the dataset into two sets: one with missing BPM entries and one without. We ran a forward stepwise selection procedure on the dataset without missing BPM entries to select the best subset of predictors (excluding the response variables Anxiety, Depression, Insomnia, OCD, and Music_effects). This model was then used to regress and predict the missing BPM values, which were subsequently filled into the original dataset. Figure 2 shows the summary of the model used for BPM prediction.

```
Call:
lm(formula = BPM ~ Composer + Foreign_languages + Frequency_EDM +
    Frequency_Folk + Frequency_Gospel + Frequency_Hip_hop + Frequency_Latin +
    Frequency_Metal + Frequency_Rap, data = bpm.not.na)

Residuals:
     Min       1Q   Median       3Q      Max
-137.781  -21.699   -1.586   20.523   89.535

Coefficients:
                   Estimate Std. Error t value             Pr(>|t|)
(Intercept)         128.861      6.246  20.632 < 0.0000000000000002 ***
ComposerYes          -8.284      3.512  -2.359               0.0187 *
Foreign_languagesNo  -4.036      2.803  -1.440               0.1504
Frequency_EDM         2.675      1.280   2.090               0.0370 *
Frequency_Folk       -2.069      1.378  -1.502               0.1337
Frequency_Gospel     -4.796      1.959  -2.448               0.0147 *
Frequency_Hip_hop    -4.537      2.089  -2.172               0.0302 *
Frequency_Latin      -3.838      1.619  -2.370               0.0181 *
Frequency_Metal       5.599      1.220   4.591          0.00000538 ***
Frequency_Rap         3.025      2.023   1.495               0.1354
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.61 on 604 degrees of freedom
Multiple R-squared:  0.09806,   Adjusted R-squared:  0.08462
F-statistic: 7.296 on 9 and 604 DF,  p-value: 0.0000000004044
```

Figure 2: BPM model summary

Some of the factors influencing the beats per minute (BPM) of an individual's favorite music genre include whether the respondent is a composer, whether they listen to music in foreign languages, and the frequency of listening to certain music genres such as EDM, folk, gospel, hip hop, Latin, metal, and rap. Composers and those who do not listen to foreign language music tend to prefer music with lower BPM, while genres like EDM, metal, and rap are associated with higher BPM. The p-value of the F-statistic indicates that at least one of the model coefficients is significant, suggesting that the model provides a better prediction than a naive comparison to the average BPM. However, the R^2 value indicates that less than 10% of the variation in BPM is explained by the variables in the model, implying that other factors influencing BPM were not captured by this model.

Additionally, although we used all other variables as predictors for BPM, we did not detect any high collinearity between BPM and any other variables in the dataset (discussed further in Section 4.4). Furthermore, the predicted BPM values are approximately normally distributed, as illustrated in Figure 3 in Section 4.3.1. This is the first step in data preprocessing. Further data cleaning and manipulation will be explained as we proceed in answering the research questions.

4.3. Exploratory Data Analysis (EDA)

In this section, we used various techniques to visualize the underlying relationships between music effects and other predictors. Through the following plots and graphs, we aim to gain a better understanding of our dataset, setting a foundation for addressing the research questions in the subsequent sections. By examining these visual representations, we can identify patterns, trends, and potential correlations that will inform our further analysis.

4.3.1. Music effects (Age, Hours_per_day, and BPM)

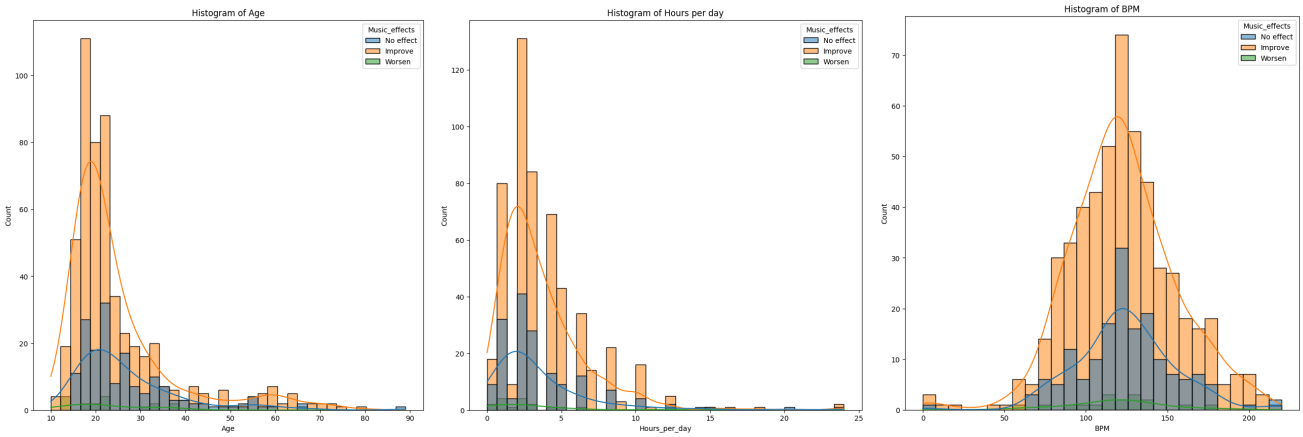


Figure 3: Music effects (Age, Hours_per_day, and BPM)

As seen in Figure 3, the age histogram reveals that the most common age group among the respondents is 15-35 years old. This age range has the highest frequency, indicating that a significant portion of the respondents falls within this demographic. The distribution appears to be skewed towards younger ages, with a gradual decline in frequency as age increases beyond 35. This observation suggests that the dataset may be more representative of a younger population's perspectives on the effects of music on mental health. The histogram for the number of hours per day respondents listen to music shows a distinct pattern. The majority of respondents report listening to music for less than 10 hours per day, with the highest frequency observed in the lower range of hours. As the number of hours increases beyond 10, the frequency of respondents decreases substantially, indicating that it is less common for individuals in this dataset to listen to music for extended periods throughout the day.

4.3.2. Music effects (Categorical columns)

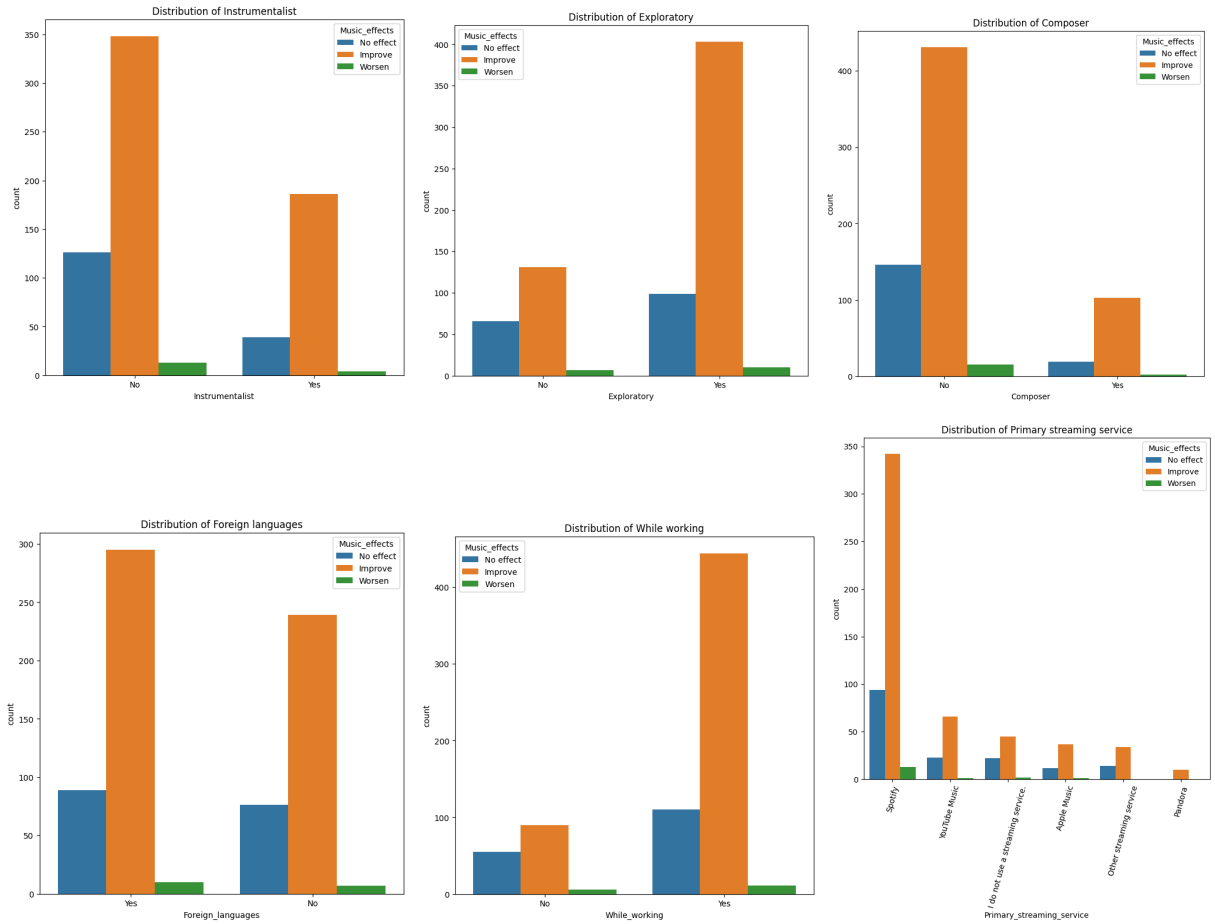


Figure 4: Music effects with Categorical columns

From Figure 4, some of the observations that we can make are:

- Most of the respondents are not instrumentalists.
- Most of the respondents actively listen to new artists and genres (exploratory).
- Most of the respondents are not composers.
- There's a good mix of respondents who listen and don't listen to Foreign languages.
- Most of the respondents listen to music while working
- Most of the respondents use Spotify as the music streaming platform.

4.3.3. Music effects (Fav_genre)

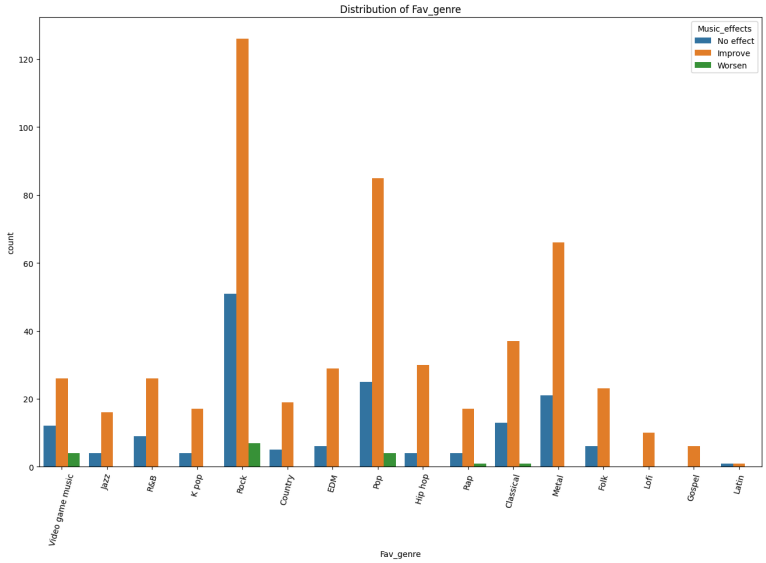


Figure 5: Music effects (Fav_genre)

As seen in Figure 5, Rock, Pop, and Metal were the most popular favorite genres of music among the respondents, and this observation aligns with the finding that the primary age group of the respondents falls between 15 and 35 years old. These genres are typically associated with the preferences of younger demographics, particularly teenagers and young adults.

4.3.4. Music effects (Anxiety, Depression, Insomnia, OCD)

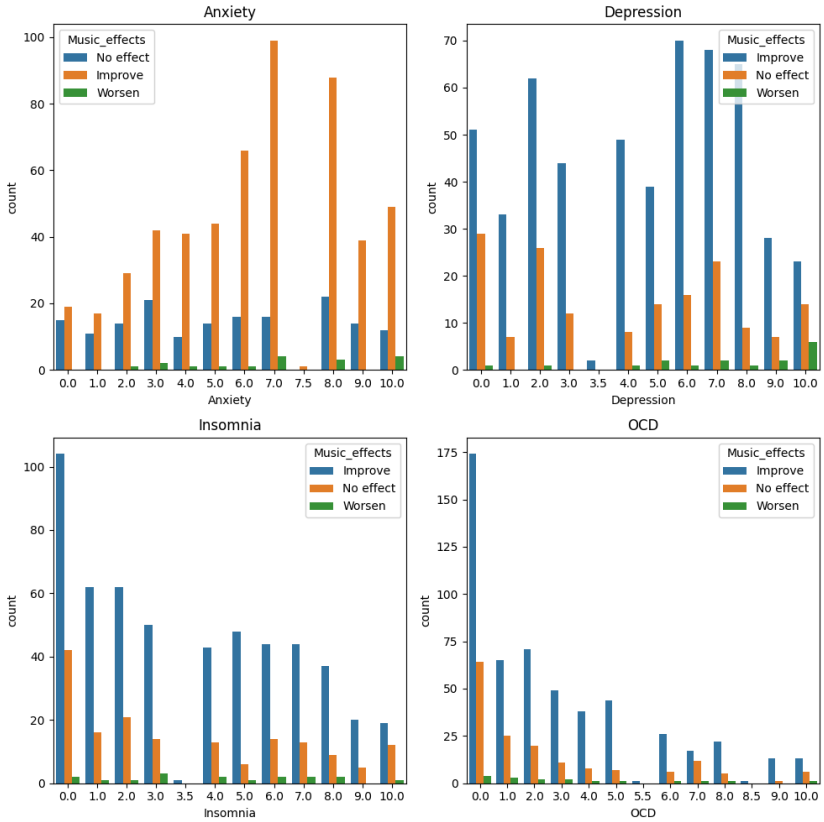


Figure 6: Music effects (Anxiety, Depression, Insomnia, OCD)

From Figure 6, it is evident that music had a positive impact on various mental health conditions reported by the respondents. Depression and anxiety emerged as the most prevalent mental health issues among the participants. This finding highlights the potential therapeutic benefits of music in alleviating symptoms associated with depression and anxiety, which are among the most common mental health disorders worldwide. The positive effects observed in the data suggest that incorporating music into treatment plans or self-care routines could be a valuable complementary approach for individuals struggling with these conditions.

4.3.4. Music effects (Frequency_<Genre>)

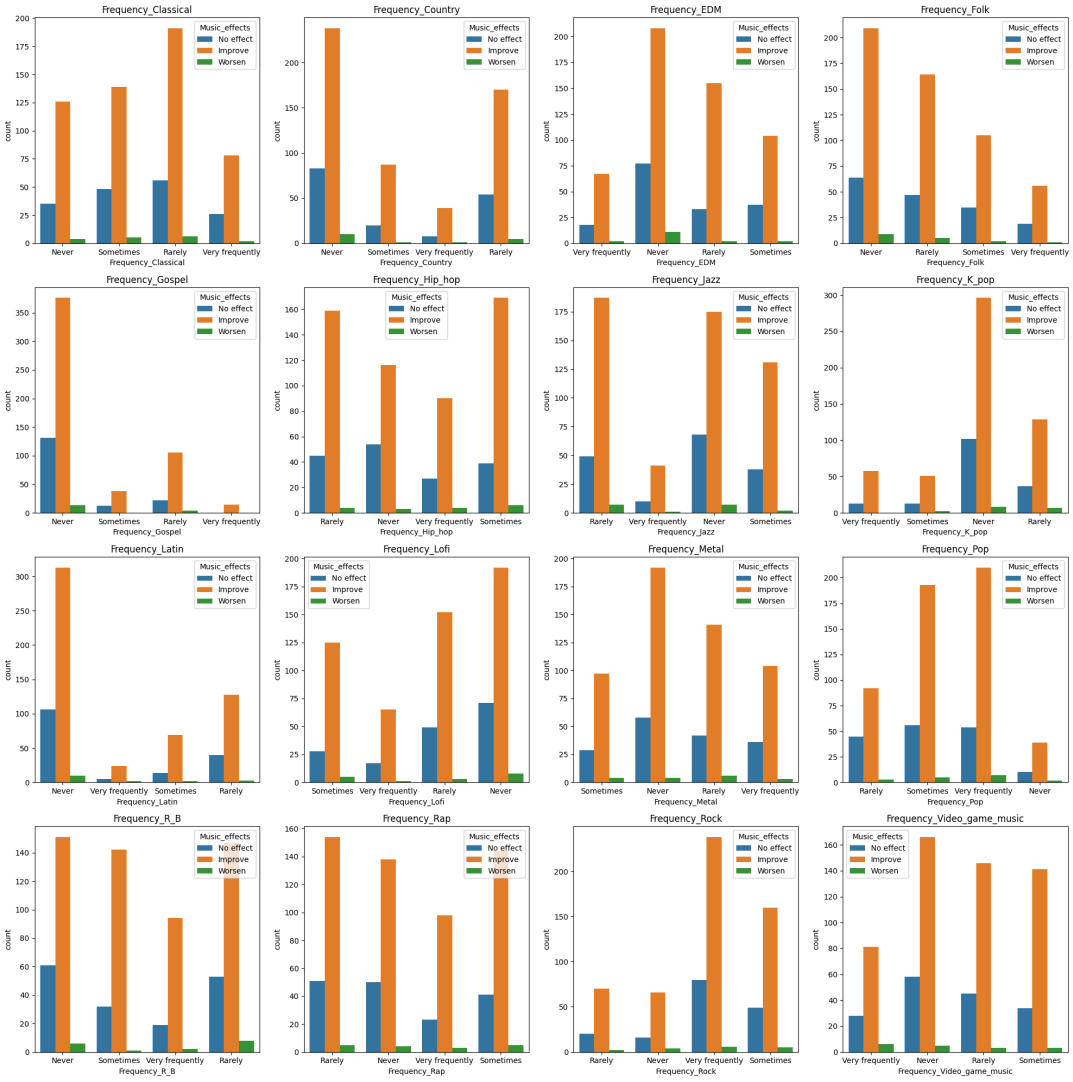


Figure 7: Music effects with Frequency_<Genre>

From Figure 7, we can observe that the majority of listeners across all music genres, regardless of how frequently they listen, experienced positive effects of music on their mental health status. This trend is consistent across different genres, indicating a general beneficial impact of music on mental well-being. Notably, even occasional listeners reported improvements in their mental health, suggesting that even limited exposure to music can have a positive influence. This finding highlights the potential of music as a therapeutic tool for enhancing mental health across diverse populations and preferences.

4.4. Outlier Detection

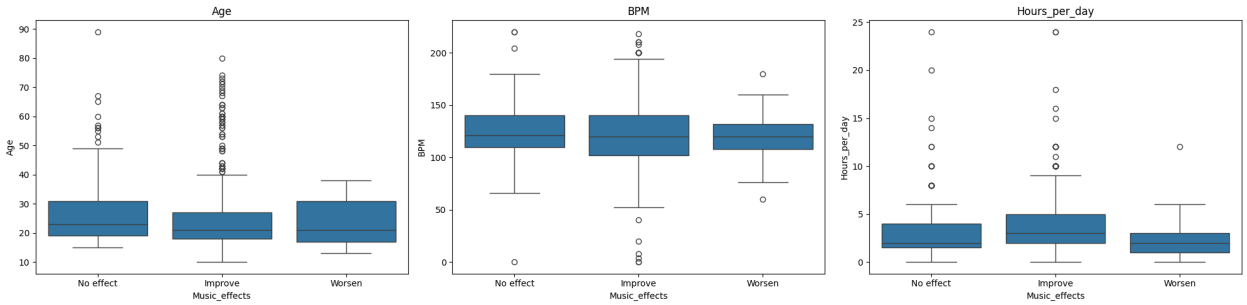


Figure 8: Outlier detection in Age, BPM, and Hours_per_day

In our dataset, box plots for *Age*, *BPM*, and *Hours_per_day* revealed several outliers. We decided to retain these outliers, treating them as true outliers, to ensure the integrity and completeness of our analysis. Outliers can provide valuable insights into the variability within the population and help avoid bias in our results. Additionally, modern robust statistical methods allow for effective handling of outliers, ensuring our conclusions remain reliable and comprehensive.

4.4. Collinearity detection

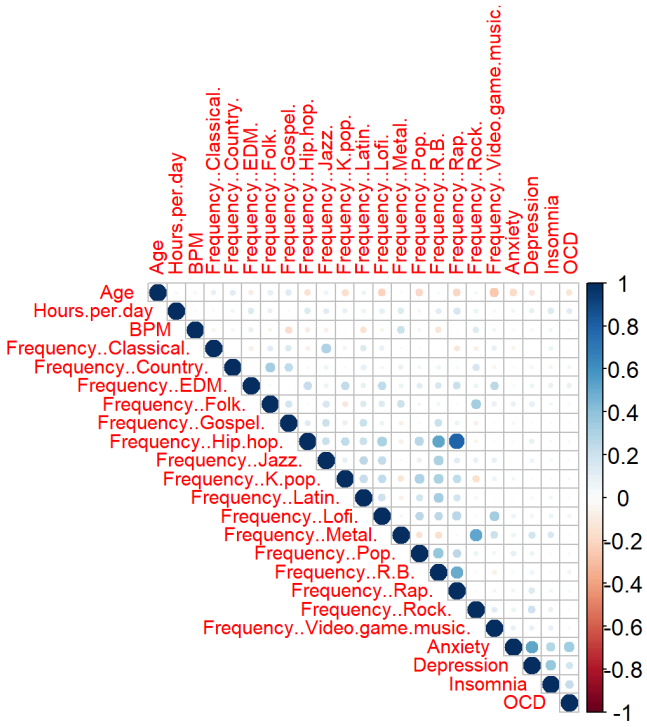


Figure 9: Correlation matrix

On plotting the correlation matrix for our dataset, as shown in Figure 9, we observed that there was no significant collinearity between most of the variables. This indicates that our variables are relatively independent, which is beneficial for avoiding multicollinearity issues in our analyses. However, we did notice moderately significant positive collinearity between *Frequency_Hiphop* and *Frequency_Rap*. This relationship suggests that individuals who frequently listen to hip-hop music also tend to listen to rap music, which could be due to the similarities and overlap in the audience and cultural aspects of these genres.

5. Approach and Results

5.1. Comparative Analysis

To answer the first research question, we used comparative analysis to assess the differences between different groups of respondents who make music. Figure 10 shows an overview of the approach in exploring the first research question. First, we employed a t-test to see whether there is any significant difference between the musician and non-musicians groups. Then, we conducted an ANOVA test to explore the differences across all four groups as well as the interaction between Composer and Instrumentalist. Each of these tests was replicated for each of the mental health conditions, allowing us to thoroughly investigate the potential impact of musical involvement on mental well-being.

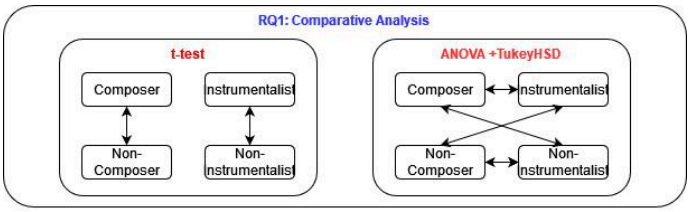


Figure 10: Research Question 1 approach overview

5.1.1. T-tests

We performed two t-tests to compare the mental health outcomes of : (1) Composer vs Non-Composer and (2) Instrumentalist vs Non-Instrumentalist. Both tests aimed to determine whether there is a significant difference in mental health conditions between the two groups of those who make/play music and those who do not. Such a pair of t-tests is then replicated for each of the four mental health conditions. Each t-test was conducted for the various mental health conditions included in the study, such as anxiety, depression, insomnia, and OCD. The results of these t-tests provided initial insights into whether composing or playing an instrument is associated with better mental health outcomes.

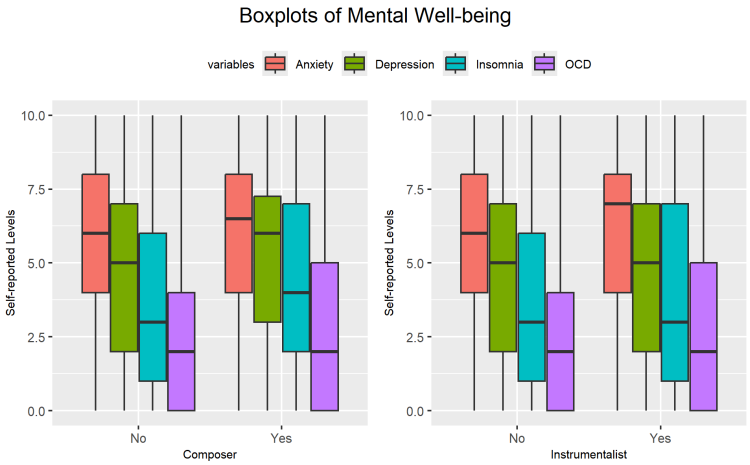


Figure 11: Boxplot for mental health condition between Yes/No Composer and Instrumentalist.

Figure 11 shows the boxplot summarizing the self-reported levels of mental health conditions. At a glance, there seems to be no difference in the distribution of Q1/median/Q3 across each group of musicians for each mental health condition. Though, the median level is slightly higher for both Composer and Instrumentalist across all four mental health levels. At a glance, there seems to be no difference in the distribution of the first-quartile, median, and third-quartile across each group of musicians for each mental health condition.

	Composer	Non-Composer	t-value	df	p-value	95% Confidence Intervals
Anxiety	5.8629	5.838682	0.0854	173.36	0.9320	[-0.5353, 0.5838]
Depression	5.1613	4.722973	1.4703	178.28	0.1432	[-0.1591, 1.0171]
Insomnia	4.3710	3.578547	2.5720	175.53	0.0109	[0.1813, 1.3972]
OCD	2.7419	2.609797	0.4701	178.50	0.6389	[-0.4225, 0.6868]

Table 4: t-test results for differences in mental health levels between Composer and Non-Composer

	Instrumentalist	Non-Instrumentalist	t-value	df	p-value	95% Confidence Intervals
Anxiety	5.9913	5.7731	0.98483	455.54	0.3252	[-0.2172, 0.6535]
Depression	4.8493	4.7752	0.3112	464.52	0.7558	[-0.3944, 0.5427]
Insomnia	3.8537	3.6509	0.8084	429.97	0.4193	[-0.2903, 0.6958]
OCD	2.7314	2.5862	0.6451	462.67	0.5192	[-0.2971, 0.5875]

Table 5: t-test results for differences in mental health levels between Instrumentalist and Non-Instrumentalist

Tables 4 and 5 display the summary of the t-test results for the two groups. At a 5% significance level, Composer groups have significantly higher levels of insomnia than Non-Composers. For all other groups, the rest of the p-values are very across all mental health conditions, indicating that there is almost no statistical difference in the mean self-reported level of mental health condition.

5.1.2. ANOVA Model

Given our suspicion that there might be interactions between being a Composer and an Instrumentalist, we further conducted ANOVA tests to explore these potential interactions. The ANOVA tests were performed on the following four groups:

1. Composer and Instrumentalist
2. Composer and Non-Instrumentalist
3. Non-Composer and Instrumentalist
4. Non-Composer and Non-Instrumentalist

The purpose of these ANOVA tests was to assess whether the combination of composing and playing an instrument has a unique impact on mental health outcomes, beyond the individual effects of each activity. Each ANOVA test was replicated for each of the mental health conditions to provide a comprehensive understanding of how these musical activities interact to influence mental well-being.

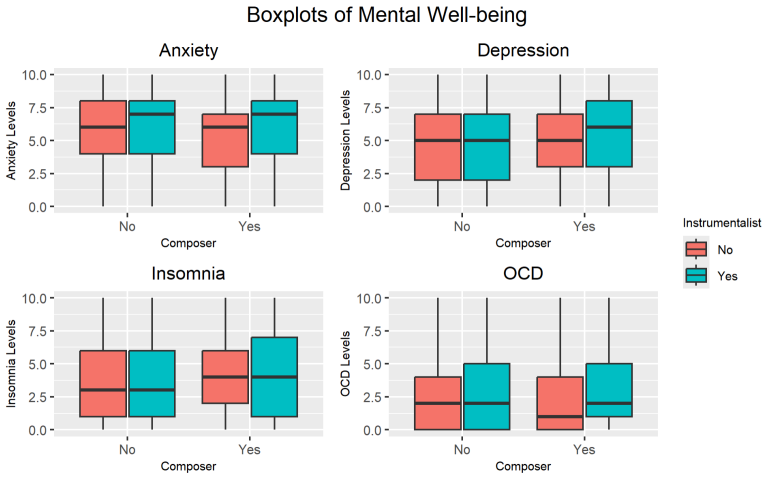


Figure 12: Boxplot for mental health conditions with interaction between Composer and Instrumentalist.

Similar to the boxplots in Figure 11, there was little to no difference in the first quartile, median, and third quartile across all four mental health conditions. Table 6 summarizes the results from the ANOVA tests. As discussed earlier, there is a statistically significant difference between the two groups of Composers. However, the high p-value of the interaction term indicates that there is no significant interaction effect between being a Composer and an Instrumentalist on all four self-reported mental health conditions. Furthermore, we conducted the TukeyHSD test (results shown in Appendix A), and found no significant differences in the mental health levels between any of the four groups mentioned earlier.

	Anxiety		Depression		Insomnia		OCD	
	F-value	p-value	F-value	p-value	F-value	p-value	F-value	p-value
Instrumentalist	0.9532	0.3292	0.0939	0.7593	0.6777	0.41065	0.4038	0.5254
Composer	0.1141	0.7356	2.1655	0.1416	6.2014	0.01299	0.0533	0.8174
Instrumentalist:Composer	0.8448	0.3583	0.7177	0.3972	0.1308	0.7177	0.5673	0.4516

Table 6: ANOVA results for differences in mental health levels with interactions between Composer and Instrumentalist

5.2. Inferential Modeling

To answer the second research question, we employed various techniques to explore both the linear and non-linear relationships between different variables in an individual mental health condition. We began with feature selection for linear modeling through stepwise selection with different criterias followed by regularization models, and selected the model with the lowest average MSE obtained through cross-validation. Finally, we use Generalized Additive Modeling (GAM) to assess any potential non-linear relationships between the explanatory variables and each of the response variables (Anxiety, Depression, Insomnia, OCD).

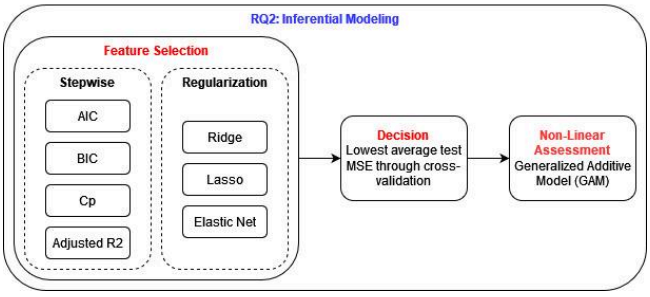


Figure 13: Research Question 2 approach overview

5.2.1. Stepwise Feature Selection

Due to the large number of variables in the dataset, we employed various techniques to select the optimal subset of features. The first technique used is bi-directional stepwise selection, which combines both forward and backward selection. This approach was applied based on four different criteria to select the best model: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mallow's Cp, and Adjusted R^2. Each criterion helps balance model fit and complexity in different ways, ensuring that the selected features contribute meaningfully to the model without overfitting. All models summary can be found on our GitHub repository through the link at the beginning of this report.

5.2.2. Regularization Regression

To further support the robustness of subset selection via the stepwise method, we also performed regularization with different penalty terms through Ridge, Lasso, and Elastic Net regressions. For each model, we employed cross-validation for parameter tuning to find the value of lambda (and alpha if applicable) that minimizes the mean squared error (MSE).

Before performing regularization regression, we had to preprocess categorical variables into the desired format. Boolean variables such as Composer, Instrumentalist, While_working, Exploratory, and Foreign_languages were converted to 0 and 1 representations for No and Yes, respectively. All Frequency_<Genre> variables were represented by discrete numeric scales ranging from 1 (Never) to 4 (Very frequently). For variables that were neither boolean nor could be represented by discrete numerics, such as Fav_genre and Primary_streaming_service, we used one-hot encoding to transform these columns before passing them to the models. This preprocessing ensured that the data was in a suitable format for the regularization techniques, allowing for effective feature selection and parameter estimation.

5.2.3. Model Evaluation

By combining stepwise selection and regularization methods, we aimed to identify the most important predictors and improve the overall accuracy and generalizability of our models for predicting mental health outcomes. Our main metric for model evaluation is the average Mean Squared Error (MSE) for each model obtained through cross-validation. However, it is important to note that Ridge regression does not perform feature selection. Therefore, even if Ridge regression yields the lowest average MSE, we avoid selecting the model it returns due to its inability to reduce the number of predictors. Our dataset has 26 explanatory variables after cleaning, and not performing feature selection would severely impair the model's interpretability. Additionally, since there was no collinearity observed among our explanatory variables, Ridge regression's advantage in handling multicollinearity does not apply here.

	Anxiety	Depression	Insomnia	OCD
AIC	7.3315	8.3762	8.9396	7.7769
BIC	7.3821	8.5426	9.1372	7.8648
Cp	7.0697	8.1766	8.7767	7.7769
Adjusted R2	7.4756	8.7802	9.2611	7.9589
Ridge	7.3341	8.2718	9.0516	7.9376
Lasso	7.2170	8.2789	9.2942	7.8593
Elastic Net	7.2547	8.2833	8.9354	7.8688
Best Model	Cp	Cp	Cp	Cp
Second Best Model	Lasso	Ridge -> Lasso	Elastic Net	Lasso

Table 7: Table summary of average MSE for different mental health conditions for different feature selection technique

As shown in Table 7, the model selected using Mallow's Cp yields the lowest average MSE for all of the response variables. The models with the second lowest average MSE are selected from Lasso regression for Anxiety and OCD, Ridge regression for Depression, and Elastic Net regression for Insomnia. These results indicate that while Mallow's Cp provides the most accurate predictions overall, regularization techniques also offer competitive performance, particularly for specific mental health conditions.

	RSE	df	R2	Adjusted R2	F-stat	p-value	Model
Anxiety	2.709	690	0.0879	0.0549	2.661	0.0000	Anixety ~ Age + Primary_streaming_service + Fav_genre + Exploratory + Foreign_languages + Frequency_Folk + Frequency_Pop
Depression	2.905	694	0.1026	0.0755	3.780	0.0000	Depression ~ Age + Fav_genre + Frequency_Country + Frequency_Folk + Frequency_Metal + Frequency_Rap + Frequency_Rock
Insomnia	3.008	695	0.0746	0.0479	2.800	0.0000	Insomnia ~ Hours_per_day + Fav_genre + Frequency_Classical + Frequency_Country + Frequency_EDM + Frequency_Metal
OCD	2.800	710	0.0400	0.0332	5.917	0.0000	OCD ~ Age + Hours_per_day + Foreign_languages + Frequency_Country + Frequency_EDM

Table 8: Summary of models with minimum average MSE for each mental health condition

Anxiety. In this model, age has a negative correlation with anxiety levels. Specifically, as age increases by one year, anxiety levels decrease by 0.04 units on average. Regarding categorical variables, the choice of primary streaming service and favorite genre shows different directions and magnitudes of correlation with anxiety levels. For instance, on average, those using Apple Music as the primary streaming service were associated with a 1.09 unit higher in anxiety levels while using YouTube Music is associated with a 0.35 unit lower in anxiety levels

compared to those using Spotify. Different favorite music genres also show varied correlations with anxiety levels. Respondents who prefer Latin, R&B, and Rap genres are associated with lower anxiety levels than those who prefer Classical music. Conversely, those who prefer other music genres tend to have higher anxiety levels compared to Classical music listeners. Those who explore new music genres and artists (ExploratoryYes) or do not listen to foreign language music also show lower anxiety levels compared to their counterparts. Additionally, a higher frequency of listening to Folk or Pop music is associated with higher anxiety levels.

When comparing this to the second-best model selected by Lasso regression, variables such as Age, Foreign_languages, Frequency_Pop, and Frequency_Folk have similar magnitudes and directions in both models. However, there are some discrepancies in the correlation sign across different genres of favorite music between the models selected by Cp and Lasso. Additionally, Frequency_video_game is included in the regularized model but not selected by the stepwise model.

Depression. Similar to the Anxiety model, age has a negative correlation with depression levels, with every one-year increase in age resulting in a 0.02 unit decrease in depression levels on average. Favorite genres again impact depression levels differently, with a base level of Classical music. For instance, genres such as Gospel, Jazz, Kpop, Latin, Metal, Pop, R&B, and Rap are associated with lower depression levels compared to Classical music. Most coefficients are low, except for Fav_genreLofi, which is associated with a 1.84 unit increase in depression levels compared to Classical music on average. Additionally, as the frequency of listening to Country music increases by one unit, depression levels decrease by 0.417 units on average. In contrast, genres like Folk, Metal, Rap, and Rock show an increase in depression levels with increased listening frequency.

Compared to the Lasso-selected model, variables that were absent in the Cp-selected model such as hours of listening to music per day, listening to foreign language music, and primary streaming service are also correlated with higher depression levels. Although Ridge regression had the second-lowest average MSE, we chose Lasso as the second-best model due to Ridge's lack of feature selection. Variables present in both Ridge and Lasso models have similar signs and magnitudes.

Insomnia. As with the previous two mental health conditions, different favorite genres affect insomnia levels differently. Those who prefer Gospel or Lofi music are associated with 2.33 and 1.95 unit increases in insomnia levels, respectively, compared to those who prefer Classical music. However, despite being statistically significant in all other three mental health models, age is not an important factor influencing insomnia levels. Genres such as Classical, Country, EDM, and Metal are statistically significant at 10% confidence level: increased listening to Country music decreases insomnia levels, while increased listening to Classical, EDM, and Metal increases insomnia levels. Additionally, an increase of one hour per day in music listening duration is associated with a 0.13 unit increase in insomnia levels on average.

All variables selected by the stepwise regression with Cp criteria are also included in the Elastic Net regression model. However, the Elastic Net model includes more terms, such as Hours_per_day, Composer, Exploratory, BPM, and Frequency_Latin/Lofi/Rap/Rock/Video_game_music, which are all positively correlated with insomnia levels.

OCD. Similar to the Anxiety and Depression models, age has a statistically significant effect on OCD levels. As respondents age by one year, their OCD levels decrease by 0.03 units on average. Other variables such as Hours_per_day, Foreign_languages, Frequency_Country, and Frequency_EDM are positively correlated with OCD levels. The second-best model for OCD (Lasso) includes similar variables with similar magnitudes and directions. However, the Lasso model also includes additional variables such as While_working, indicating that

those who listen to music while working have higher OCD levels than those who do not, and Fav_genreGospel, indicating that people who prefer Gospel music are likely to have lower OCD levels on average.

Detailed results for both the best and second-best models discussed above can be found in Appendix B. Despite being chosen as the best models that optimize the variance and bias trade-offs, the R^2 values for all these models are very low, indicating that less than 11% of the variations in mental health levels are explained by the model predictors. However, the p-value for the F-statistics is approximately 0, indicating that at least one coefficient is statistically significant. The residual plots in Appendix B show that the residuals are scattered around 0 without any clear patterns or obvious outliers, further supporting the conclusion that there is little bias in the model.

5.2.4. General Additive Model (GAM)

Lastly, we used GAM to explore any non-linear relationship between predictors and explanatory variables. For each mental health condition, we used the best model selected from section 5.2.3, smoothed all non-categorical variables and applied a GAM fit. We then assess the GAM model summary and plots, for any variables without a clear non-linear relationship with the response variable, we re-run the GAM fit without smoothing those variables.

Anxiety and Depression. The GAM plots for Anxiety and Depression both indicate a nonlinear relationship between age and Anxiety/Depression levels. Figure 14 and 15 show the plots for smoothed Age variables. As age increases, the general trend for Anxiety/Depression levels decreases, at a certain age, there is an upward rising trend. For instance, in the ranges 10-20, 25-45, and 70-80, as age increases, Anxiety/Depression levels decrease, but in other age ranges, the Anxiety/Depression levels in fact increase. In addition for both models, the EDF term is greater than 5, indicating a complex pattern between age and the response variable which strongly suggest a non-linear relationship. The parametric coefficients for these two models are similar in direction and magnitude as that in the best model (selected by Cp discussed in section 5.2.3.) and hence follow the same interpretation.

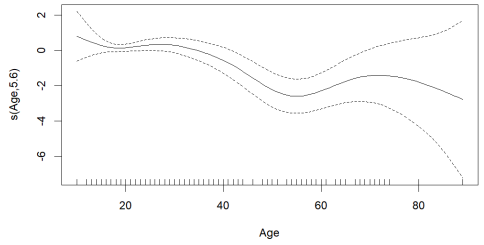


Figure 14. GAM plot for Anxiety

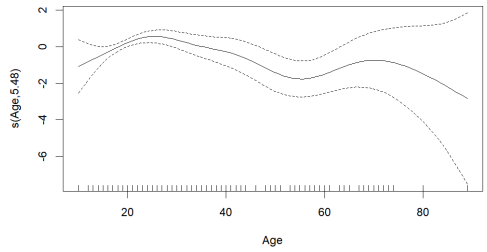


Figure 15. GAM plot for Depression

Insomnia and OCD. After running GAM on Insomnia and OCD best models, we found that the smoothed variables illustrated by Figure 16 and 17 are approximately linear with the corresponding mental health condition , hence we re-run the model removing the smoothed terms. The EDF terms for these models are about 1, indicating weak to no nonlinear relationship. Hence, the final model is in fact the same as the linear model selected by stepwise using Cp in section 5.2.3.

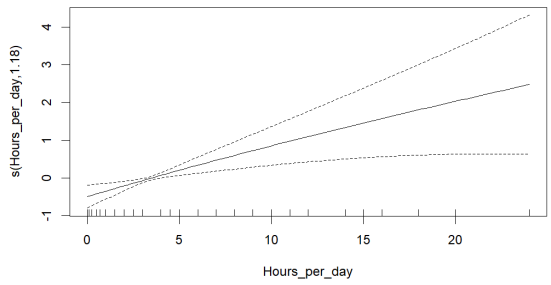


Figure 16. GAM plot for Insomnia

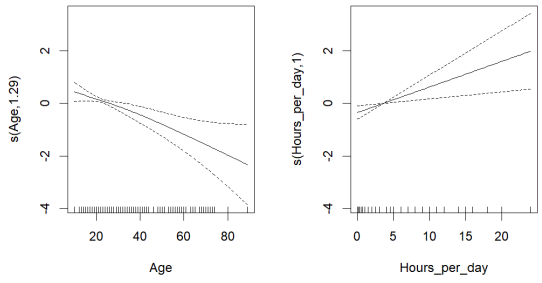


Figure 17. GAM plot for OCD

In general, across all GAM plots, we also observe that the entries are concentrated toward the lower (left) end, hence there's greater standard deviation towards the right tail. Besides, majority of the variables in this dataset are categorical, hence although GAM analysis did provide some insight to non-linearity of the continuous numerical variables, it does not provide much details on other variables in the model (apart from the parametric coefficients that have the same interpretation as the Cp best model discussed in section 5.2.3.)

5.3. Classification Modeling

The classification modeling process begins with data preprocessing, ensuring the data is clean and in a format that the ML models can understand. Next, the dataset is split into training (80%) and testing (20%) subsets to evaluate model performance accurately. Initially, all available features are used to train various classification models, and their performance is reported. Feature selection techniques are then employed to identify the most relevant predictors. Stepwise forward selection methods, such as AIC and BIC, are used to iteratively add variables that improve the model fit. Additionally, regularized regression methods like Lasso, Ridge, and Elastic Net are applied to shrink or eliminate less important features. After identifying the optimal subset of features through these selection methods, all the classification models are retrained using only the selected features.

5.3.1. Data Preprocessing for Parametric ML models

As illustrated in Figure 18, the previously cleaned data required additional preprocessing steps before running the parametric machine learning (ML) models. We first created a new binary column, *Music_Effects_Improve* (the response variable), derived from the *Music_Effects* column. This new column indicates whether music helps improve or does not improve mental health conditions, as our focus is on the same. To achieve this, we retained the 'Improve' entries and combined the 'No effect' and 'Worsen' entries into a single class. Subsequently, we dropped the original 'Music_Effects' column since it was no longer needed.

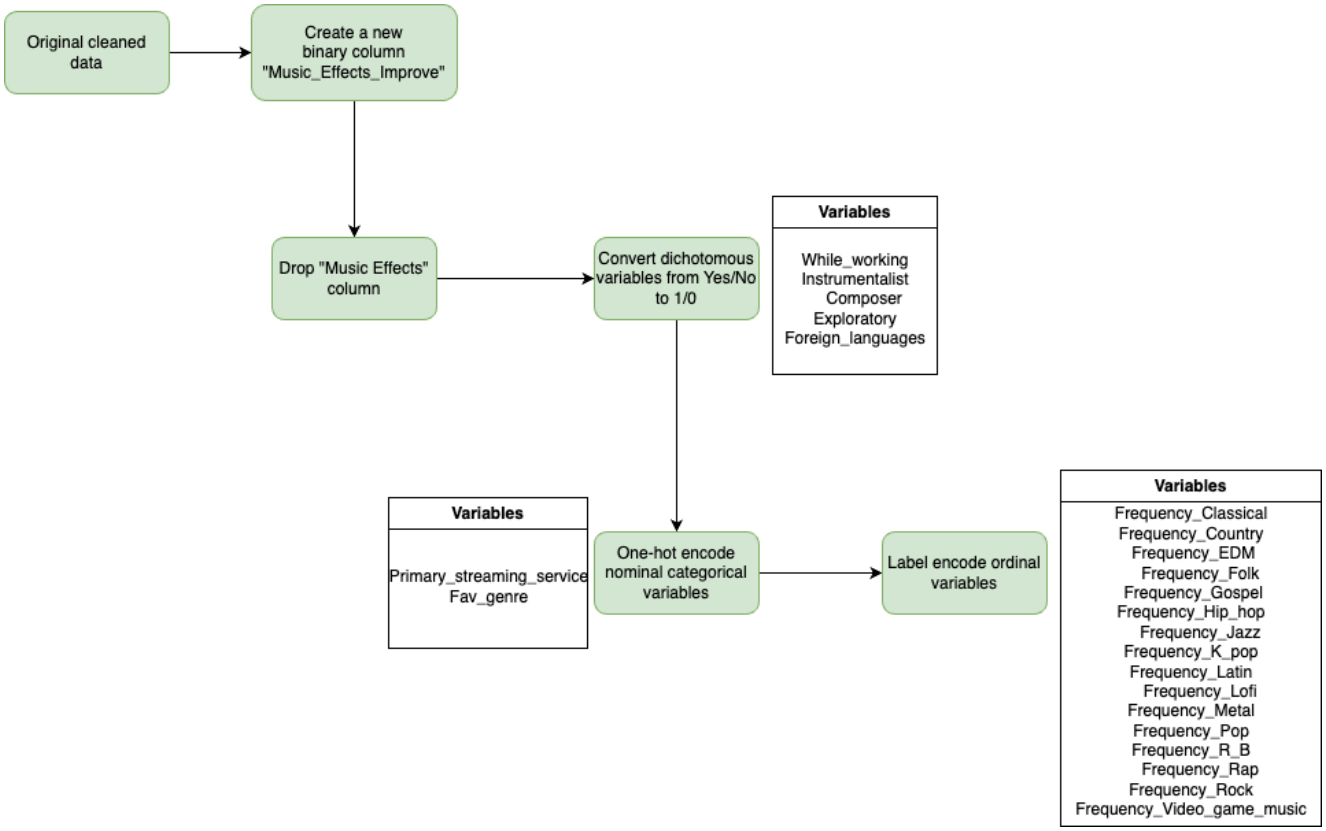


Figure 18: Data preprocessing for parametric ML models

After this step, we converted the dichotomous variables from 'Yes/No' to '1/0', as machine learning models perform better with numerical data. Finally, we applied one-hot encoding to the nominal categorical variables (those without an inherent ordering among their categories) and label encoding to the ordinal variables (categorical variables with an inherent ordering among their categories).

5.3.2. Performance of parametric models without feature selection

The parametric models that we worked with are Logistic Regression, Linear Discriminant Analysis (LDA), and two different Neural Networks, one with 2 hidden layers and the other with three hidden layers. The performance of these models in terms of the accuracy and the test error by including all the features in the dataset and not performing any sort of feature selection is shown in Table xx. Just for comparison, we have also added a Random Forest model (a non-parametric model) to understand how it behaves when compared to other parametric models.

	Accuracy	Test Error	TPR	TNR	FPR	FNR	Notes
Logistic Regression	0.7153	0.2847	0.8909	0.1471	0.8529	0.1091	No penalty
LDA	0.7083	0.2917	0.9	0.0882	0.9118	0.1	
Random Forest (non-parametric)	0.75	0.25	0.9818	0	1	0.0182	n_estimators = 100
Neural Network 1	0.7014	0.2986	0.8909	0.0882	0.9118	0.1091	2 hidden layers (64 * 32) 5 epochs
Neural Network 2	0.7222	0.2778	0.9091	0.1176	0.8824	0.0909	3 hidden layers (64 * 32 * 16) 5 epochs

Table 9: Performance of parametric models without any feature selection
(TPR = True Positive Rate, TNR = True Negative Rate, FPR = False Positive Rate, FNR = False Negative Rate)

As noticed in the Table 9, the accuracies of all the models ranged between 0.70 and 0.75, indicating that these classification models were able to do the classification task at hand decently. The Neural Network architecture with three hidden layers performed the best, with an accuracy of 0.7222.

5.3.3. Stepwise Forward Selection

To perform feature selection and run our parametric models, we utilized the Stepwise Forward Selection technique. By employing stepwise forward selection, we aimed to find the best model by considering important metrics such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), while disregarding the Adjusted R-squared since it is not applicable for Logistic Regression. These metrics helped us assess the model's goodness-of-fit and complexity. The results are shown in Figure 19 and Figure 20, indicating the best models obtained with the AIC and BIC metrics, respectively. Upon close observation of these results, we noticed a few strong predictors for each metric, with some common predictors between them.

```
glm(formula = Music_effects_Improve ~ While_working + Exploratory +
    Anxiety + Frequency_R_B + Fav_genre_Hip.hop + Instrumentalist +
    Fav_genre_Country + Fav_genre_Lofi + Primary_streaming_service_Pandora +
    Fav_genre_EDM + Frequency_Folk + Frequency_Country + Frequency_Rap +
    Frequency_Classical, family = binomial, data = train_data)

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -0.8592     0.3705  -2.319 0.020396 *
While_working                           0.7403     0.2364   3.131 0.001744 **
Exploratory                             0.7312     0.2256   3.240 0.001194 **
Anxiety                                 0.1281     0.0371   3.454 0.000551 ***
Frequency_R_B                           0.3376     0.1128   2.992 0.002767 **
Fav_genre_Hip.hopTrue                   2.1524     1.0692   2.013 0.044105 *
Instrumentalist                         0.6364     0.2444   2.604 0.009212 **
Fav_genre_CountryTrue                   0.7961     0.8546   0.931 0.351596
Fav_genre_LofiTrue                     15.3555   784.3435   0.020 0.984380
Primary_streaming_service_PandoraTrue  15.0937   928.2635   0.016 0.987027
Fav_genre_EDMTrue                       0.8500     0.5282   1.609 0.107584
Frequency_Folk                         -0.2186     0.1103  -1.982 0.047497 *
Frequency_Country                       0.2578     0.1384   1.862 0.062558 .
Frequency_Rap                          -0.2083     0.1170  -1.780 0.075104 .
Frequency_Classical                    -0.1859     0.1124  -1.653 0.098238 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 19: Akaike Information Criterion (AIC)

```
Call:
glm(formula = Music_effects_Improve ~ While_working + Exploratory +
    Anxiety + Frequency_R_B, family = binomial, data = train_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.77813    0.31053  -2.506  0.01222 *
While_working  0.69075    0.22567   3.061  0.00221 **
Exploratory    0.65360    0.21151   3.090  0.00200 **
Anxiety        0.10733    0.03511   3.057  0.00224 **
Frequency_R_B  0.24589    0.09823   2.503  0.01231 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 20: Bayesian Information Criterion (BIC)

However, we decided to proceed with the predictors determined by the BIC metric for three reasons. First, we were interested in reducing the model's complexity, and BIC performs well in this regard by tending to penalize complex models more heavily than AIC, resulting in fewer predictors compared to AIC. Second, by choosing BIC, we could eliminate the problem of overfitting. Lastly, since our goal was to identify the strong predictors, BIC was preferred. Finally, we settled on *While_working*, *Exploratory*, *Anxiety*, and *Frequency_R_B* as the most robust predictors to be used in order to re-run the machine learning models and evaluate their performance.

	Accuracy	Test Error	TPR	TNR	FPR	FNR	Notes
Logistic Regression	0.7778	0.2222	0.9818	0.1176	0.8824	0.0182	No penalty
LDA	0.7639	0.2361	0.9545	0.1471	0.8529	0.0455	
Random Forest (non-parametric)	0.7361	0.2639	0.9091	0.1765	0.8235	0.0909	n_estimators = 100
Neural Network 1	0.7778	0.2222	0.9636	0.1765	0.8235	0.0364	2 hidden layers (64 * 32) 5 epochs
Neural Network 2	0.7639	0.2361	0.9818	0.0588	0.9412	0.0182	3 hidden layers (64 * 32 * 16) 5 epochs

Table 10: Performance of parametric models with feature selection (stepwise forward selection)

The performance of all the models using only the predictors selected by stepwise forward selection is shown in Table 10. Two notable observations can be made from the table. Firstly, the accuracy of all the parametric models has improved by approximately 0.05, indicating that fitting the model with only the strong predictors is beneficial rather than using all the predictors. This improvement in accuracy can be attributed to the fact that including less significant predictors can introduce noise and complicate the model, leading to overfitting or underfitting. Secondly, we can infer from the table that the performance of the Random Forest model dropped after feature selection. This observation aligns with the expectations, as Random Forest is a non-parametric model. Unlike parametric models, which can benefit from feature selection by reducing model complexity and mitigating overfitting, Random Forest models are inherently robust to irrelevant or redundant features. By randomly selecting subsets of features during the tree-building process, Random Forest models can automatically identify and utilize the most important features, making explicit feature selection less beneficial or even detrimental to their performance.

5.3.4. Regularization

In addition to stepwise forward selection, we performed feature selection through regularization techniques, including Ridge, Lasso, and Elastic Net. The coefficients versus Lambda graphs for all these techniques are provided in the Appendix D. The Lasso regularization method zeroed out 32 out of the 50 predictors, effectively selecting the remaining 18 as the most relevant features. Similarly, the Elastic Net regularization method zeroed out 36 of the predictors, retaining 14 as the most influential features. We proceeded with the non-zero predictors obtained through the Elastic Net regularization method.

	Accuracy	Test Error	TPR	TNR	FPR	FNR	Notes
Logistic Regression	0.757	0.243	0.9364	0.1765	0.8235	0.0636	No penalty
LDA	0.7361	0.2639	0.9182	0.1471	0.8529	0.0818	
Random Forest (non-parametric)	0.7083	0.2917	0.8727	0.1765	0.8235	0.1273	n_estimators = 100
Neural Network 1	0.7083	0.2917	0.9	0.0882	0.9118	0.1	2 hidden layers (64 * 32) 5 epochs
Neural Network 2	0.7014	0.2986	0.8818	0.1176	0.8824	0.1182	3 hidden layers (64 * 32 * 16) 5 epochs

Table 11: Performance of parametric models with feature selection (regularization)

An interesting observation from the regularization results is that the four predictors chosen through the stepwise forward selection (*While_working*, *Exploratory*, *Anxiety*, and *Frequency_R_B*) were also present in the non-zero predictors identified by the Elastic Net regularization. This consistency across different feature selection methods reinforces the notion that these predictors are indeed strong and influential in the context of our problem.

The performance of the models using the predictors selected through regularization techniques followed a similar trend as the results obtained from stepwise forward selection. However, the accuracy improvement was not as substantial as the one achieved through stepwise forward selection. Based on these findings, we can conclude that the performance of the parametric models, in descending order, is as follows:

1. Stepwise forward selection
2. Regularization (Lasso, Elastic Net)
3. No feature selection

For our dataset, feature selection through stepwise forward selection proved to be the most effective approach to predicting whether a particular combination of features helps improve mental health conditions. This method yielded the highest accuracy improvement among the parametric models, outperforming both regularization techniques and the models without any feature selection.

5.3.5. Data Preprocessing for Non-Parametric ML models

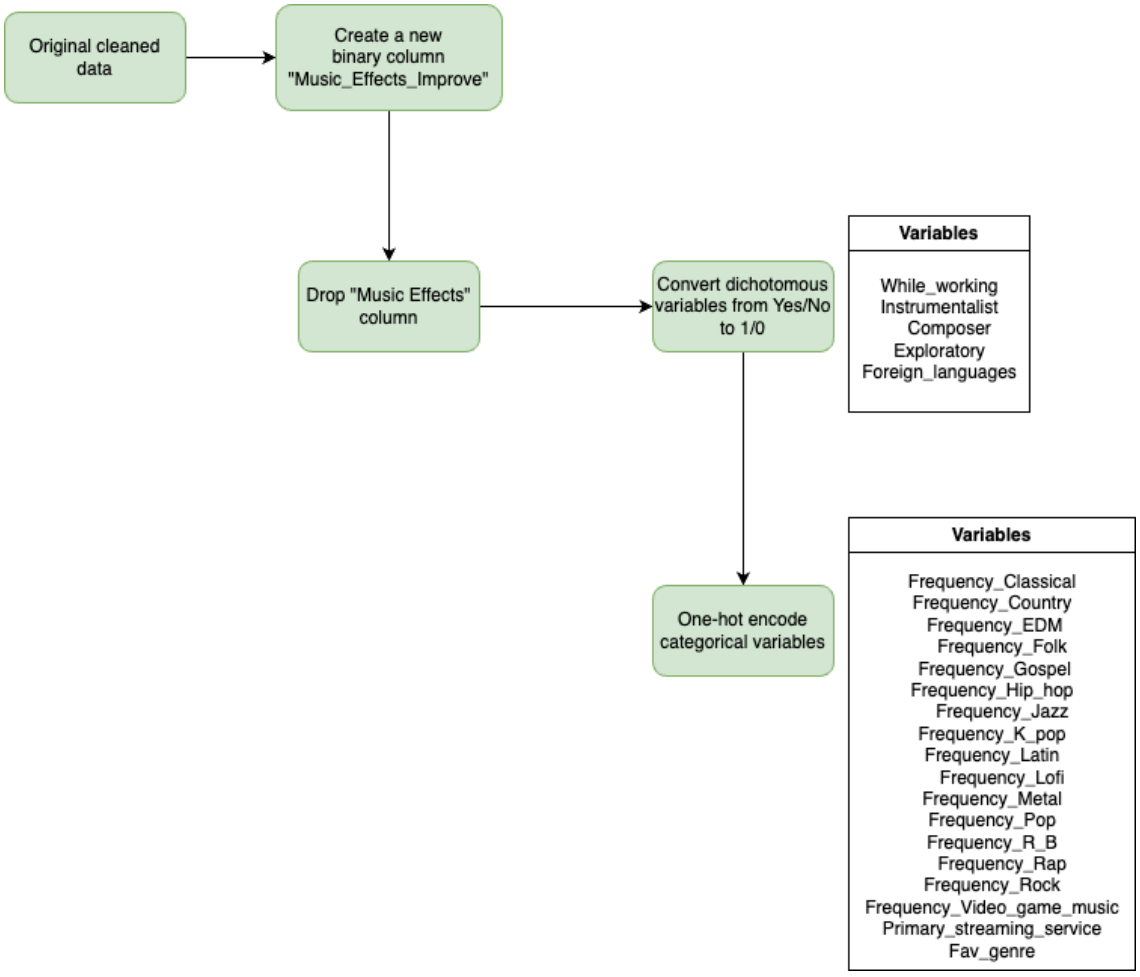


Figure 21: Data preprocessing for non-parametric ML models

Since some non-parametric ML models, such as KNN, rely heavily on distance metrics, the key distinction in the data preprocessing process for these models, as illustrated in Figure 21, is the need to one-hot encode all categorical variables, irrespective of whether they are nominal or ordinal. This step ensures that the categorical data is appropriately transformed into a format suitable for distance-based computations, thereby maintaining the accuracy and effectiveness of the model.

5.3.6. Performance of Non-parametric ML models

The non-parametric models we employed in our analysis are K-Nearest Neighbors (KNN), Random Forest, and Support Vector Classifier with three different kernels (linear, polynomial, and radial basis function). For these non-parametric models, we made the decision to include all features present in the dataset when evaluating their performance. This decision was based on the understanding that non-parametric models can often benefit from the inclusion of all available features, as they have an inherent ability to handle and effectively utilize high-dimensional data. The performance of these non-parametric models, in terms of accuracy and test error, is presented in Table 12.

	Accuracy	Test Error	TPR	TNR	FPR	FNR	Notes
KNN	0.7778	0.2222	0.9909	0.0882	0.9118	0.0091	Best k = 17 in [1, 30]
Random Forest	0.743	0.257	0.9727	0	1	0.0273	n_estimators = 100
SVC 1 (parametric)	0.7222	0.2778	0.8909	0.1765	0.8235	0.1091	Kernel = Linear
SVC 2	0.75	0.25	0.9818	0	1	0.0182	Kernel = Polynomial
SVC 3	0.757	0.243	0.9909	0	1	0.0091	Kernel = Radial Basis Function (RBF)

Table 12: Performance of non-parametric models

Among the non-parametric models evaluated, K-Nearest Neighbors (KNN) with k = 17 achieved the highest accuracy of 0.7778. The Random Forest model, with 100 estimators, followed closely with an accuracy of 0.743. The Support Vector Classifier (SVC) with a Radial Basis Function (RBF) kernel performed the best among the three SVC models, attaining an accuracy of 0.7570. The SVC with a polynomial kernel had an accuracy of 0.7500, while the linear SVC model exhibited the lowest performance with an accuracy of 0.7222.

6. Findings

6.1. Comparative Analysis

Based on the t-tests, we found that the Insomnia level was significantly higher among Composers compared to Non-Composer, this was further supported by the ANOVA results. However, there were no significant differences between Instrumentalists and Non-Instrumentalists across all mental health conditions. Additionally, the high p-value of the interaction term from the ANOVA result indicates that there was no significant interaction effect between being both a Composer and an Instrumentalist on mental health conditions. Lastly, TukeyHSD analysis also found no significant difference between the four groups of (1) Composer and Instrumentalist, (2) Composer and Non-Instrumentalist, (3) Non-Composer and Instrumentalist, and (4) Non-Composer and Non-Instrumentalist for any of the mental health conditions.

6.2. Inferential Modeling

After running various regression modeling, we found that Age plays a major role in influencing the anxiety, depression, and OCD levels. Specifically, as respondents get one year older, they are associated with lower levels of anxiety, depression, and/or OCD. It's interesting to see that the platform they primarily stream music on can also influence their anxiety level, but not other mental health conditions. Certain favorite genres are also

associated with different levels of mental health conditions. Latin, R&B, and Rap consistently show lower anxiety and depression levels, indicating a generally positive impact on mental health. Lofi is associated with higher depression and insomnia levels. EDM is associated with higher insomnia and OCD levels, despite a positive effect on depression levels in some cases. Gospel shows lower depression and OCD but higher insomnia, indicating genre-specific effects vary by condition. Classical music listeners show higher anxiety and insomnia but not significantly higher depression or OCD levels.

6.3. Classification Modeling

Based on our classification modeling results, we observed that all the models we implemented demonstrated comparable accuracy and test error. Among both parametric and non-parametric approaches, K-Nearest Neighbors (KNN) and a Neural Network with two hidden layers emerged as the top performers, achieving the highest accuracy of 0.778. Further analysis of the confusion matrices revealed that the True Positive Rate (TPR) and False Negative Rate (FNR) fell within desirable ranges. However, the True Negative Rate (TNR) and False Positive Rate (FPR) deviated somewhat from optimal levels. This suggests a slight bias in our models towards the 'Improve' class, likely due to an imbalance in the dataset with a higher proportion of 'Improve' samples. This insight highlights the importance of considering class distribution and potentially employing techniques to address class imbalance in future iterations of our model development process.

7. Limitations and Future Works

7.1. Dataset Limitations

Self-selection Bias. The dataset relies on voluntary survey responses, which may introduce self-selection bias. Individuals who choose to participate might have a particular interest in music or mental health, and hence could skew the results.

Limited Demographic Diversity. Clinical therapy effectiveness largely depends on individual characteristics. This dataset lacks individual demographic factors such as gender, socio-economic status, cultural background, etc. All these variables could influence both music preferences and mental health status, which limits our ability to understand and control for the confounding variables.

Measurement Bias. The survey results rely heavily on respondents' subjective evaluation of music effects on their mental health condition. Depending on how the questions were phrased, the response may be biased and prone to inaccuracies.

7.2. Future Works

Longitudinal Study. Future research can conduct studies over time to analyze the causality or effects of musical therapy on mental health conditions. This would provide better understanding of how musical preferences and engagement overtime impact mental health outcomes.

Intervention Study. By collecting more demographic data with more diverse and representative samples that can be generalized to a broader population, researchers have more data to design intervention studies to test the efficiency of music therapies tailored to individuals or groups of individuals with shared characteristics. This would help with clinical applications. For instance, these studies could examine the weight of the impact on music-making versus passive listening respondents.

8. Conclusion

Our study aimed to investigate the relationships between music listening habits, individual characteristics, and mental health outcomes. Through rigorous analysis, we have uncovered several key findings that contribute to our understanding of these complex interactions as follows:

RQ1: Besides streaming music, are "musicians" associated with better mental health conditions? While composing music displays some significant effects on mental health, playing an instrument does not, and there's no significant interaction effects between composing and playing an instrument.

RQ2: What are the individual characteristics associated with mental well-being? Different favorite genres and listening frequencies have varied impacts on mental health conditions such as anxiety, depression, insomnia, and OCD. While some genres like Latin, R&B, and Rap generally correlate with lower mental health issues, others like Lofi and EDM tend to show higher levels of certain conditions. Age is associated with all mental health conditions except for insomnia, with older individuals generally reporting lower levels of anxiety, depression, and OCD. Additionally, listening to more music per day is associated with higher levels of insomnia and OCD.

RQ3: Can individual characteristics predict the music effects? If so, how accurately? Yes, our models were able to predict music effects from the individual characteristics with an accuracy rate of ~ 0.78.

In conclusion, while we found no significant association between being a musician and improved mental health, we did identify several individual characteristics that correlate with various mental health conditions. However, the low explanatory power of these associations suggests that mental health is influenced by a complex interplay of factors beyond those captured in our study. Notably, our models demonstrated a promising ability to predict the effects of music on individuals based on their characteristics, achieving a relatively high accuracy rate. These findings provide valuable insights for future research and potential applications in personalized music therapy and mental health interventions.

9. Learning and knowledge

Throughout the course of our project, we leveraged a comprehensive array of statistical and machine learning techniques learned in class to analyze and derive insights from our dataset. Our approach began with thorough quantitative analysis, including the visualization of complex relationships within the data. We addressed challenges such as missing values and outlier detection, ensuring robust data pre-processing. Our modeling efforts spanned both linear and non-linear regression techniques, including the implementation of generalized additive models (GAM) and the exploration of interaction terms in linear regression. We employed stepwise forward selection methods, utilizing criteria such as AIC, BIC, adjusted R-squared, and Mallows' Cp to refine our models. Additionally, we explored regularization techniques like Ridge, Lasso, and Elastic Net to enhance model performance and prevent overfitting. For classification tasks, we implemented logistic regression, linear discriminant analysis, and K-nearest neighbors. We also delved into more advanced algorithms such as Random Forests, Support Vector Machines, and Neural Networks. By applying this diverse toolkit of methods, we were able to successfully extract meaningful patterns, make accurate predictions, and draw valuable conclusions from our dataset, ultimately fulfilling the project objectives.

10. References

[1] Lee, J., & Thyer, B. A. (2013). Does Music Therapy Improve Mental Health in Adults? A Review. *Journal of Human Behavior in the Social Environment*, *23*(5), 591–603. https://doi.org/10.1080/10911359.2013.766147

[2] Legge, Alexander W. "On the neural mechanisms of music therapy in mental health care: Literature review and clinical implications." Music Therapy Perspectives 33.2 (2015): 128-141.

[3] Tríona McCaffrey, Jane Edwards, “Music Therapy Helped Me Get Back *Doing* ”: Perspectives of Music Therapy Participants in Mental Health Services, *Journal of Music Therapy*, Volume 53, Issue 2, Summer 2016, Pages 121–148, https://doi.org/10.1093/jmt/thw002

[4] Rahman, Jessica Sharmin, et al. "Towards effective music therapy for mental health care using machine learning tools: human affective reasoning and music genres." Journal of Artificial Intelligence and Soft Computing Research 11.1 (2021): 5-20.

APPENDIX A. TukeyHSD Results

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Anxiety ~ group, data = mt.data)

$group
               diff        lwr       upr     p adj
INC-IC   -0.1056697 -1.0746615 0.8633221 0.9922701
NIC-IC   -0.7216117 -2.1797401 0.7365166 0.5795842
NINC-IC  -0.2577959 -1.0816554 0.5660635 0.8517309
NIC-INC  -0.6159420 -2.0064200 0.7745360 0.6644189
NINC-INC -0.1521262 -0.8492885 0.5450360 0.9432740
NINC-NIC  0.4638158 -0.8297175 1.7573491 0.7923717
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Insomnia ~ group, data = mt.data)

$group
               diff        lwr       upr     p adj
INC-IC   -0.9174630 -1.9870641 0.1521381 0.1217839
NIC-IC   -0.1338661 -1.7433903 1.4756581 0.9965281
NINC-IC  -0.7969443 -1.7063441 0.1124556 0.1093525
NIC-INC   0.7835968 -0.7512530 2.3184466 0.5538323
NINC-INC  0.1205187 -0.6490291 0.8900665 0.9778091
NINC-NIC -0.6630781 -2.0909176 0.7647613 0.6296653
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Depression ~ group, data = mt.data)

$group
                diff        lwr       upr     p adj
INC-IC   -0.68764931 -1.7378754 0.3625768 0.3317692
NIC-IC   -0.38494838 -1.9653173 1.1954205 0.9233052
NINC-IC  -0.48413100 -1.3770578 0.4087958 0.5021312
NIC-INC   0.30270092 -1.2043462 1.8097481 0.9549672
NINC-INC  0.20351831 -0.5520897 0.9591263 0.8995526
NINC-NIC -0.09918262 -1.5011579 1.3027926 0.9978531
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = OCD ~ group, data = mt.data)

$group
               diff        lwr       upr     p adj
INC-IC   -0.24506291 -1.2355307 0.7454049 0.9199856
NIC-IC   -0.51548452 -2.0059298 0.9749608 0.8097460
NINC-IC  -0.27276123 -1.1148801 0.5693577 0.8382514
NIC-INC  -0.27042161 -1.6917172 1.1508740 0.9613332
NINC-INC -0.02769832 -0.7403119 0.6849153 0.9996408
NINC-NIC  0.24272329 -1.0794790 1.5649256 0.9650564
```

APPENDIX B. Models Summary

B.1. Best Models (from Stepwise Selection)

Anxiety

```
Call:
lm(formula = Cp.formula, data = mt.data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0033 -1.9236  0.4508  2.0237  5.6651

Coefficients:
                                                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)                                         5.714278   0.644441   8.867 < 0.0000000000000002 ***
Age                                                -0.046410   0.009679  -4.795             0.000002 ***
Primary_streaming_serviceYouTube_Music             -0.353418   0.323126  -1.094              0.27445
Primary_streaming_serviceI_do_not_use_a_streaming_service. -0.142279   0.380892  -0.374              0.70886
Primary_streaming_serviceApple_Music                1.094774   0.411080   2.663              0.00792 **
Primary_streaming_serviceOther_streaming_service    0.035270   0.427138   0.083              0.93422
Primary_streaming_servicePandora                    0.520898   0.928397   0.561              0.57493
Fav_genreCountry                                    0.190902   0.682068   0.280              0.77965
Fav_genreEDM                                        0.170174   0.601040   0.283              0.77716
Fav_genreFolk                                       1.360302   0.681931   1.995              0.04646 *
Fav_genreGospel                                     1.110424   1.222984   0.908              0.36422
Fav_genreHip_hop                                    0.927430   0.608251   1.525              0.12778
Fav_genreJazz                                       0.871871   0.726645   1.200              0.23061
Fav_genreK_pop                                      1.048604   0.723131   1.450              0.14749
Fav_genreLatin                                     -0.776399   1.964900  -0.395              0.69287
Fav_genreLofi                                       0.661410   0.951050   0.695              0.48701
Fav_genreMetal                                      0.767765   0.489009   1.570              0.11686
Fav_genrePop                                        0.709473   0.483150   1.468              0.14244
Fav_genreR&B                                       -0.073366   0.604089  -0.121              0.90337
Fav_genreRap                                       -0.192328   0.701506  -0.274              0.78404
Fav_genreRock                                       1.110842   0.440307   2.523              0.01186 *
Fav_genreVideo_game_music                           0.723117   0.572439   1.263              0.20694
ExploratoryYes                                     -0.365331   0.247440  -1.476              0.14028
Foreign_languagesNo                                -0.304255   0.218612  -1.392              0.16444
Frequency_Folk                                      0.206688   0.116918   1.768              0.07754 .
Frequency_Pop                                       0.178113   0.131103   1.359              0.17473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.709 on 690 degrees of freedom
Multiple R-squared:  0.08793,   Adjusted R-squared:  0.05489
F-statistic: 2.661 on 25 and 690 DF,  p-value: 0.00002395
```

Depression

```
Call:
lm(formula = Cp.formula, data = mt.data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1544 -2.3435  0.1279  2.2969  7.5602

Coefficients:
                           Estimate Std. Error t value   Pr(>|t|)
(Intercept)                3.286410   0.612579   5.365 0.00000011 ***
Age                       -0.023098   0.009853  -2.344    0.01935 *
Fav_genreCountry           0.614801   0.795749   0.773    0.44002
Fav_genreEDM               0.089273   0.661993   0.135    0.89277
Fav_genreFolk              0.371028   0.733283   0.506    0.61303
Fav_genreGospel           -0.383618   1.288170  -0.298    0.76594
Fav_genreHip_hop           0.827082   0.708648   1.167    0.24356
Fav_genreJazz             -0.308955   0.779164  -0.397    0.69184
Fav_genreK_pop            -0.767986   0.772190  -0.995    0.32030
Fav_genreLatin            -0.595890   2.109354  -0.282    0.77765
Fav_genreLofi              1.840943   1.019107   1.806    0.07128 .
Fav_genreMetal            -0.413807   0.595707  -0.695    0.48751
Fav_genrePop              -0.117233   0.513516  -0.228    0.81948
Fav_genreR&B              -0.701309   0.674963  -1.039    0.29915
Fav_genreRap              -1.114873   0.803139  -1.388    0.16554
Fav_genreRock              0.268798   0.511914   0.525    0.59969
Fav_genreVideo_game_music -0.042653   0.610011  -0.070    0.94428
Frequency_Country         -0.417150   0.144660  -2.884    0.00405 **
Frequency_Folk             0.180408   0.131582   1.371    0.17080
Frequency_Metal            0.326136   0.142558   2.288    0.02245 *
Frequency_Rap              0.351846   0.129600   2.715    0.00680 **
Frequency_Rock             0.305591   0.145222   2.104    0.03571 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.905 on 694 degrees of freedom
Multiple R-squared:  0.1026,    Adjusted R-squared:  0.07548
F-statistic:  3.78 on 21 and 694 DF,  p-value: 0.00000003486
```

Insomnia

```
Call:
lm(formula = Cp.formula, data = mt.data)

Residuals:
   Min     1Q Median     3Q    Max
-5.634 -2.497 -0.290  2.466  7.646

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.84883    0.73102   2.529 0.011656 *
Hours_per_day               0.13046    0.03890   3.354 0.000841 ***
Fav_genreCountry           -0.05436    0.88016  -0.062 0.950771
Fav_genreEDM                0.05691    0.76794   0.074 0.940942
Fav_genreFolk               0.55847    0.75111   0.744 0.457410
Fav_genreGospel             2.33078    1.31057   1.778 0.075769 .
Fav_genreHip_hop            0.05008    0.71294   0.070 0.944018
Fav_genreJazz              -0.03052    0.82358  -0.037 0.970445
Fav_genreK_pop             -0.18522    0.81264  -0.228 0.819771
Fav_genreLatin              0.02218    2.17947   0.010 0.991884
Fav_genreLofi               1.95471    1.06125   1.842 0.065916 .
Fav_genreMetal              0.61847    0.66391   0.932 0.351892
Fav_genrePop                0.15937    0.56550   0.282 0.778162
Fav_genreR&B               -0.34831    0.69696  -0.500 0.617402
Fav_genreRap               -1.32897    0.80924  -1.642 0.100994
Fav_genreRock               0.34930    0.55000   0.635 0.525578
Fav_genreVideo_game_music   0.40925    0.65972   0.620 0.535235
Frequency_Classical         0.27445    0.13215   2.077 0.038193 *
Frequency_Country          -0.27039    0.14324  -1.888 0.059483 .
Frequency_EDM               0.21241    0.12347   1.720 0.085812 .
Frequency_Metal             0.26911    0.13765   1.955 0.050979 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.008 on 695 degrees of freedom
Multiple R-squared:  0.07457,   Adjusted R-squared:  0.04794
F-statistic:   2.8 on 20 and 695 DF,  p-value: 0.00004786
```

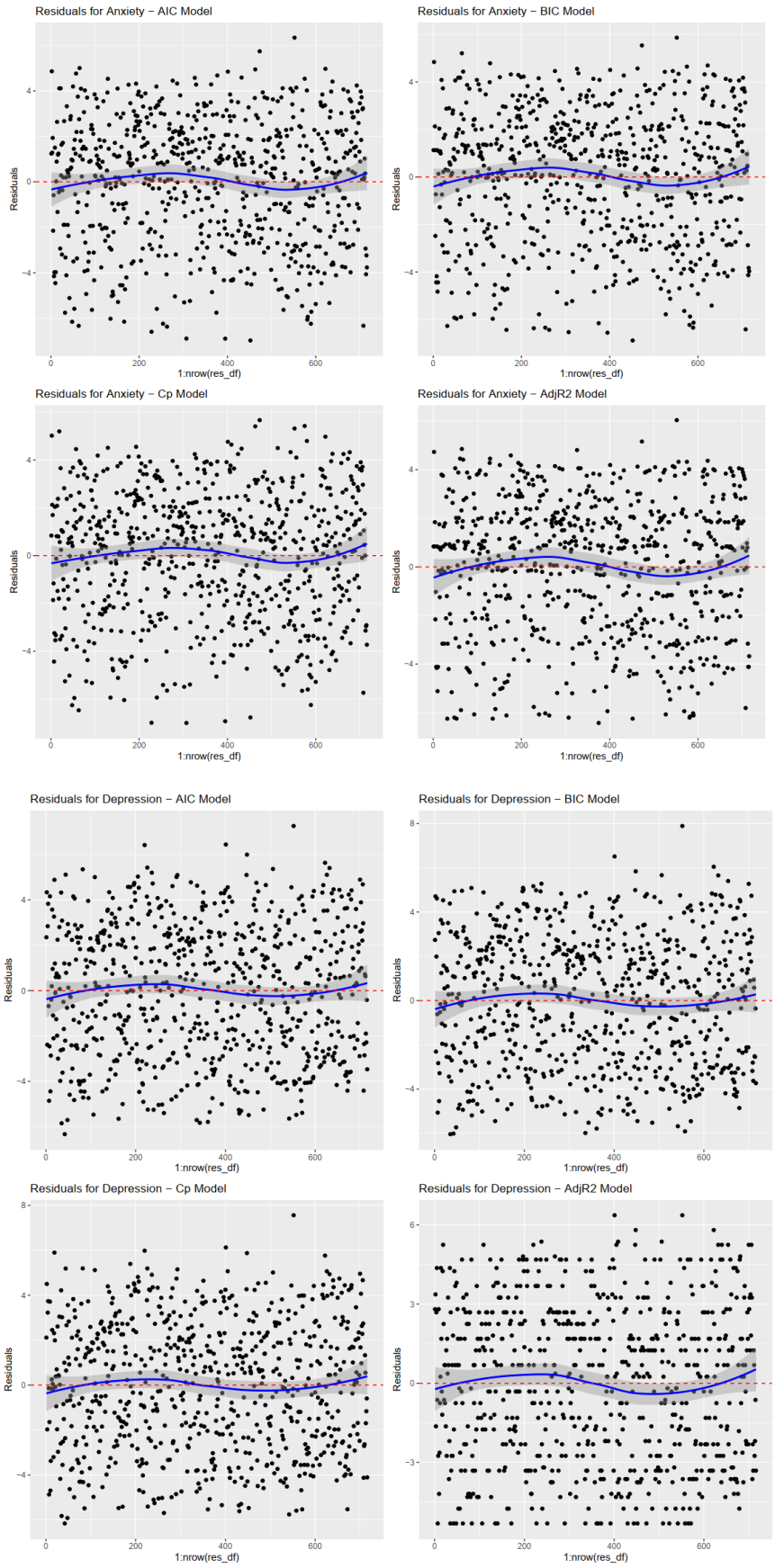
OCD

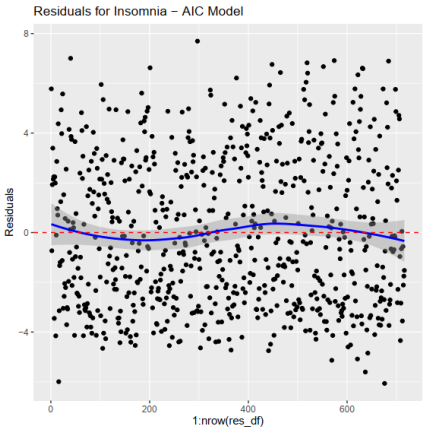
```
Call:
lm(formula = Cp.formula, data = mt.data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9495 -2.3237 -0.8061  1.6689  8.3051

Coefficients:
                   Estimate Std. Error t value     Pr(>|t|)
(Intercept)         2.23701    0.40253   5.557 0.0000000388 ***
Age                -0.03253    0.00892  -3.647     0.000285 ***
Hours_per_day       0.09531    0.03511   2.714     0.006799 **
Foreign_languagesNo 0.33044    0.21523   1.535     0.125153
Frequency_Country   0.19070    0.11453   1.665     0.096335 .
Frequency_EDM       0.18669    0.10216   1.827     0.068048 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.8 on 710 degrees of freedom
Multiple R-squared:   0.04,     Adjusted R-squared:  0.03324
F-statistic: 5.917 on 5 and 710 DF,  p-value: 0.00002271
```

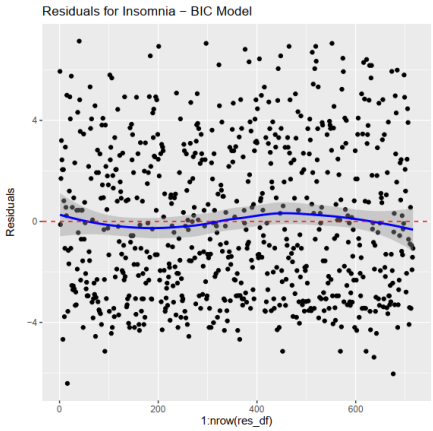
B.2. Second Best Models (from Regularization)

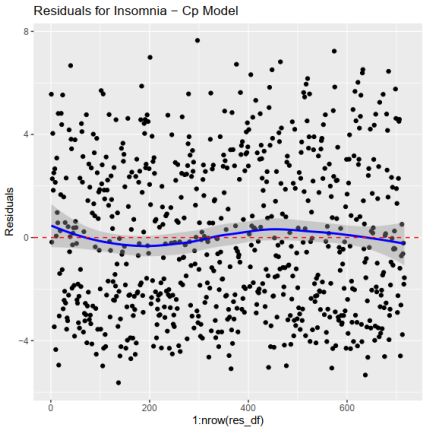
	Coefficients				
	Anxiety (Lasso)	Depression (Ridge)	Depression (Lasso)	Insomnia (Elastic Net)	OCD (Lasso)
(Intercept)	5.808934615	3.091857157	3.038053714	2.492342903	2.65841381
Age	-0.032610015	-0.0080242048	-0.01291815	x	-0.01903494
Hours_per_day	x	0.0223734783	0.019955393	0.0504017554	0.05153144
While_working	x	0.0713566994	x	x	0.10627452
Instrumentalist	x	0.0233342325	x	x	x
Composer	x	0.1225572195	x	0.2326863629	x
Exploratory	x	-0.015393048	x	0.0224354609	x
Foreign_languages	0.094397264	0.1313320883	0.093583467	x	x
BPM	x	0.0006151842	x	0.0004971068	x
Frequency_Classical	x	0.0069819022	x	0.0800396944	x
Frequency_Country	x	-0.0901102795	-0.128312067	-0.065593247	0.03877106
Frequency_EDM	x	0.0401128547	x	0.0618049018	0.05741244
Frequency_Folk	0.127022815	0.0728972198	0.082408924	x	x
Frequency_Gospel	x	-0.0034651358	x	x	x
Frequency_Hip_hop	x	0.052233081	x	x	x
Frequency_Jazz	x	0.044443494	0.003061034	x	x
Frequency_K_pop	x	-0.0364320183	x	x	x
Frequency_Latin	x	0.0274182767	x	0.0539025872	x
Frequency_Lofi	x	-0.0118572815	x	0.042946347	x
Frequency_Metal	x	0.1184173299	0.198945635	0.1335038368	x
Frequency_Pop	0.07567455	0.0660802192	0.035016689	x	x
Frequency_R_B	x	0.0486879217	0.024873173	x	x
Frequency_Rap	x	0.0846189879	0.15071084	0.0199282648	x
Frequency_Rock	x	0.1377177658	0.287138085	0.0291213083	x
Frequency_Video_game_music	0.043276451	0.0602331757	0.020768331	0.0523125341	x
Music_effects	0.241738285	-0.0601876623	x	x	x
Fav_genreClassical	-0.333809193	-0.1211684415	x	x	x
Fav_genreCountry	x	-0.0510354273	x	-0.2718928946	x
Fav_genreEDM	-0.006493569	0.0531988044	x	x	x
Fav_genreFolk	0.102187503	0.1220501704	x	x	x
Fav_genreGospel	x	-0.4305577944	x	0.6695119264	-0.13673361
Fav_genreHip_hop	x	0.362797287	0.555105811	x	x
Fav_genreJazz	x	-0.1968808219	x	x	x
Fav_genreK_pop	x	-0.3351960792	-0.06984781	-0.0055337286	x
Fav_genreLatin	x	-0.2459213461	x	x	x
Fav_genreLofi	x	0.6027312777	0.766722434	0.5673125035	x
Fav_genreMetal	x	0.0243530561	x	0.2266193183	x
Fav_genrePop	x	-0.1027775214	x	-0.0237737727	x
Fav_genreR&B	-0.118404196	-0.3535841761	-0.162784363	-0.1916405382	x
Fav_genreRap	-0.166384402	-0.3360154625	-0.06581367	-0.5779764573	x
Fav_genreRock	0.19743309	0.1975318108	0.136752282	0.0042600748	x
Fav_genreVideo_game_music	x	-0.0582843423	x	x	x
Primary_streaming_service Spotify	x	0.1665274109	0.165353955	x	x
Primary_streaming_service YouTube_Music	-0.063129489	-0.2546075408	-0.237502352	x	x
Primary_streaming_service I_do_not_use_a_streaming_service.	x	-0.1490488666	x	x	x
Primary_streaming_service Apple_Music	0.482863732	0.1596547449	x	x	x
Primary_streaming_service Other_streaming_service	x	-0.095496841	x	0.1376997906	x
Primary_streaming_service Pandora	x	-0.1727043573	x	-0.3538434112	x

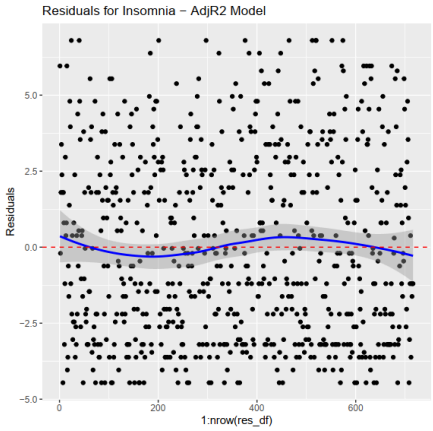
B.3. Residuals Plots

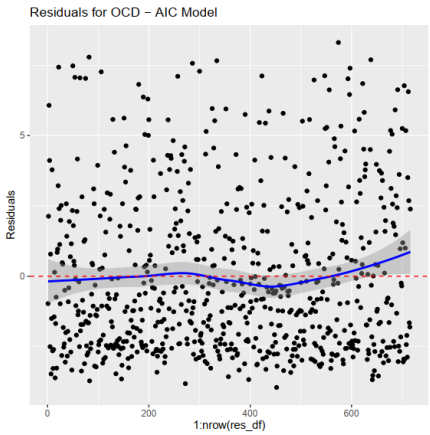


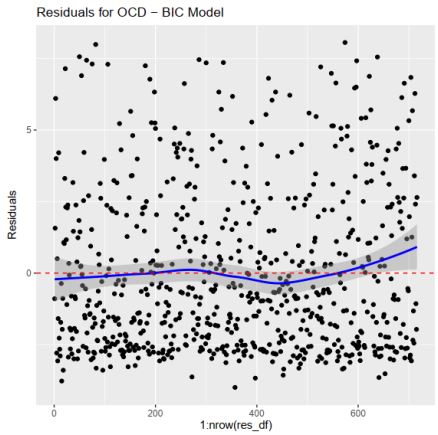

Residuals for Insomnia – AIC Model
Residuals
1:nrow(res_df)

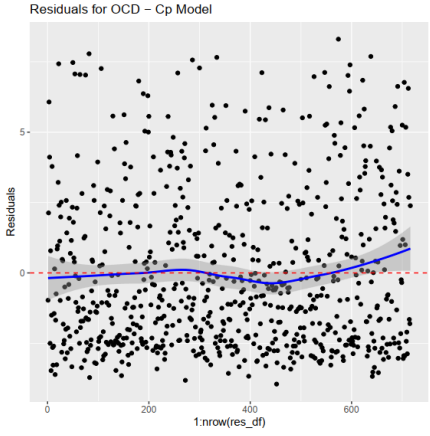

Residuals for Insomnia – BIC Model
Residuals
1:nrow(res_df)

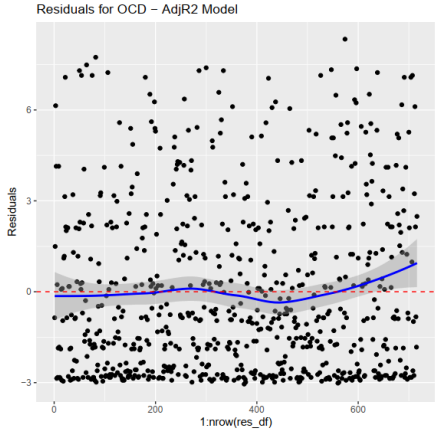

Residuals for Insomnia – Cp Model
Residuals
1:nrow(res_df)


Residuals for Insomnia – AdjR2 Model
Residuals
1:nrow(res_df)


Residuals for OCD – AIC Model
Residuals
1:nrow(res_df)


Residuals for OCD – BIC Model
Residuals
1:nrow(res_df)

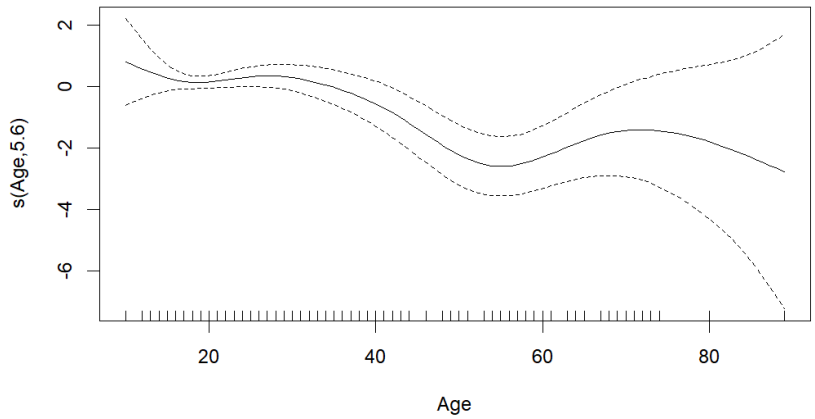

Residuals for OCD – Cp Model
Residuals
1:nrow(res_df)


Residuals for OCD – AdjR2 Model
Residuals
1:nrow(res_df)

B.4. General Additive Model (GAM)

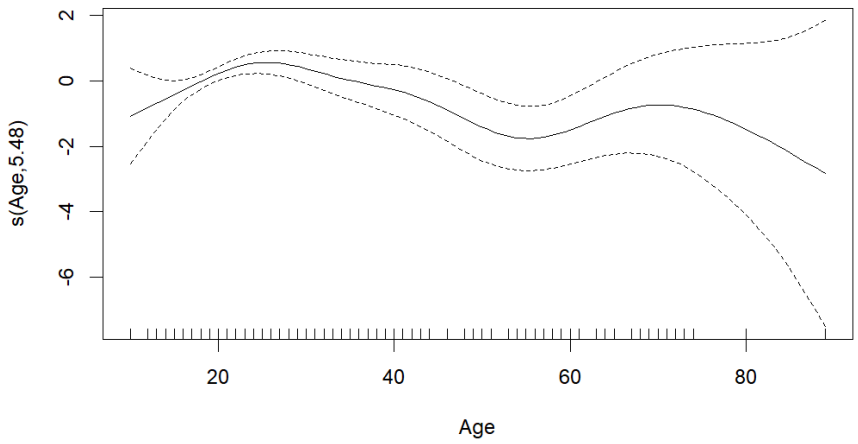
Anxiety

```
Family: gaussian
Link function: identity

Formula:
Anxiety ~ s(Age) + Primary_streaming_serviceYouTube_Music + Primary_streaming_serviceI_do_not_use_a_streaming_service. +
    Primary_streaming_serviceApple_Music + Primary_streaming_serviceOther_streaming_service +
    Primary_streaming_servicePandora + Fav_genreCountry + Fav_genreEDM +
    Fav_genreFolk + Fav_genreGospel + Fav_genreHip_hop + Fav_genreJazz +
    Fav_genreK_pop + Fav_genreLatin + Fav_genreLofi + Fav_genreMetal +
    Fav_genrePop + Fav_genreR.B + Fav_genreRap + Fav_genreRock +
    Fav_genreVideo_game_music + Exploratory + Foreign_languages +
    Frequency_Folk + Frequency_Pop

Parametric coefficients:
                                                  Estimate Std. Error t value          Pr(>|t|)
(Intercept)                                        4.30323    0.56946   7.557 0.000000000000133 ***
Primary_streaming_serviceYouTube_Music            -0.28628    0.32243  -0.888             0.3749
Primary_streaming_serviceI_do_not_use_a_streaming_service. -0.09412    0.38038  -0.247             0.8046
Primary_streaming_serviceApple_Music               1.09934    0.40855   2.691             0.0073 **
Primary_streaming_serviceOther_streaming_service   0.13386    0.42785   0.313             0.7545
Primary_streaming_servicePandora                   0.42836    0.93042   0.460             0.6454
Fav_genreCountry                                   0.10048    0.68168   0.147             0.8829
Fav_genreEDM                                       0.14011    0.60057   0.233             0.8156
Fav_genreFolk                                      1.31649    0.67875   1.940             0.0528 .
Fav_genreGospel                                    1.38196    1.23545   1.119             0.2637
Fav_genreHip_hop                                   0.90676    0.60656   1.495             0.1354
Fav_genreJazz                                      0.88325    0.72490   1.218             0.2235
Fav_genreK_pop                                     1.22098    0.72044   1.695             0.0906 .
Fav_genreLatin                                    -0.59859    1.95051  -0.307             0.7590
Fav_genreLofi                                      0.58025    0.94507   0.614             0.5394
Fav_genreMetal                                     0.65310    0.49222   1.327             0.1850
Fav_genrePop                                       0.78068    0.48271   1.617             0.1063
Fav_genreR.B                                       0.08005    0.60543   0.132             0.8949
Fav_genreRap                                      -0.13782    0.70574  -0.195             0.8452
Fav_genreRock                                      1.09561    0.44037   2.488             0.0131 *
Fav_genreVideo_game_music                          0.68933    0.56927   1.211             0.2264
Exploratory                                       -0.31024    0.24649  -1.259             0.2086
Foreign_languages                                  0.24260    0.21858   1.110             0.2674
Frequency_Folk                                     0.18018    0.11723   1.537             0.1247
Frequency_Pop                                      0.16692    0.13032   1.281             0.2007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F    p-value
s(Age) 5.601  6.724 5.332 0.00000933 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0704   Deviance explained = 10.9%
GCV = 7.5414  Scale est. = 7.219     n = 716
```



Depression

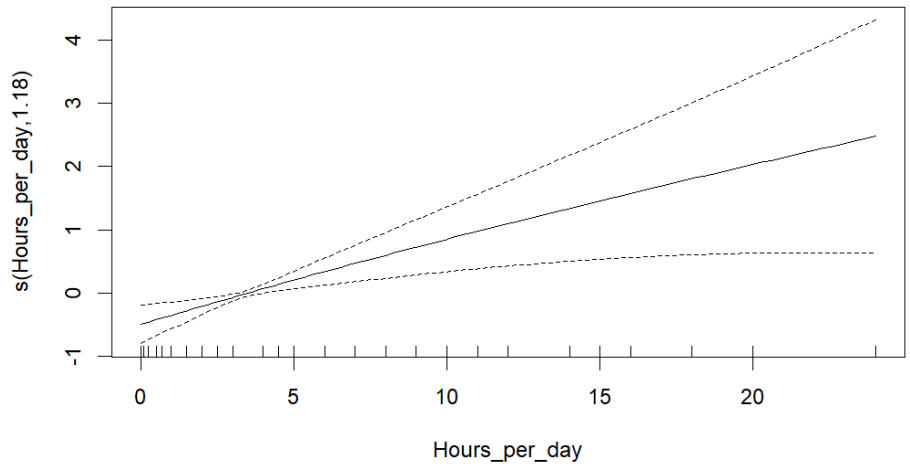
```
Family: gaussian
Link function: identity

Formula:
Depression ~ s(Age) + Fav_genreCountry + Fav_genreEDM + Fav_genreFolk +
    Fav_genreGospel + Fav_genreHip_hop + Fav_genreJazz + Fav_genreK_pop +
    Fav_genreLatin + Fav_genreLofi + Fav_genreMetal + Fav_genrePop +
    Fav_genreR.B + Fav_genreRap + Fav_genreRock + Fav_genreVideo_game_music +
    Frequency_Country + Frequency_Folk + Frequency_Metal + Frequency_Rap +
    Frequency_Rock

Parametric coefficients:
                           Estimate Std. Error t value     Pr(>|t|)
(Intercept)                 3.09480    0.55587   5.567 0.0000000371 ***
Fav_genreCountry            0.43008    0.78987   0.544       0.58628
Fav_genreEDM               -0.09907    0.65809  -0.151       0.88039
Fav_genreFolk               0.26421    0.72742   0.363       0.71655
Fav_genreGospel            -0.22035    1.30149  -0.169       0.86561
Fav_genreHip_hop            0.63538    0.70404   0.902       0.36711
Fav_genreJazz              -0.43168    0.77368  -0.558       0.57706
Fav_genreK_pop             -0.69493    0.76645  -0.907       0.36489
Fav_genreLatin             -0.41451    2.08745  -0.199       0.84266
Fav_genreLofi               1.63702    1.00914   1.622       0.10522
Fav_genreMetal             -0.62141    0.59316  -1.048       0.29518
Fav_genrePop               -0.18979    0.51081  -0.372       0.71034
Fav_genreR.B               -0.76779    0.67399  -1.139       0.25502
Fav_genreRap               -1.10181    0.80202  -1.374       0.16995
Fav_genreRock               0.23684    0.50812   0.466       0.64128
Fav_genreVideo_game_music  -0.15915    0.60481  -0.263       0.79252
Frequency_Country          -0.41833    0.14371  -2.911       0.00372 **
Frequency_Folk              0.13180    0.13088   1.007       0.31427
Frequency_Metal             0.32033    0.14133   2.266       0.02373 *
Frequency_Rap               0.36776    0.12883   2.855       0.00444 **
Frequency_Rock              0.23135    0.14496   1.596       0.11096
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(Age) 5.477    6.6 3.649 0.00154 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0965   Deviance explained = 12.9%
GCV = 8.5612  Scale est. = 8.2447    n = 716
```



Insomnia

```
Family: gaussian
Link function: identity

Formula:
Insomnia ~ s(Hours_per_day) + Fav_genreCountry + Fav_genreEDM +
    Fav_genreFolk + Fav_genreGospel + Fav_genreHip_hop + Fav_genreJazz +
    Fav_genreK_pop + Fav_genreLatin + Fav_genreLofi + Fav_genreMetal +
    Fav_genrePop + Fav_genreR.B + Fav_genreRap + Fav_genreRock +
    Fav_genreVideo_game_music + Frequency_Classical + Frequency_Country +
    Frequency_EDM + Frequency_Metal

Parametric coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.32329    0.73215   3.173  0.00157 **
Fav_genreCountry          -0.05334    0.88004  -0.061  0.95168
Fav_genreEDM               0.06039    0.76793   0.079  0.93735
Fav_genreFolk              0.55707    0.75101   0.742  0.45848
Fav_genreGospel            2.33748    1.31050   1.784  0.07492 .
Fav_genreHip_hop           0.04743    0.71287   0.067  0.94698
Fav_genreJazz             -0.03387    0.82354  -0.041  0.96721
Fav_genreK_pop            -0.18836    0.81255  -0.232  0.81676
Fav_genreLatin             0.01687    2.17936   0.008  0.99382
Fav_genreLofi              1.94782    1.06121   1.835  0.06686 .
Fav_genreMetal             0.61635    0.66384   0.928  0.35349
Fav_genrePop               0.16071    0.56544   0.284  0.77632
Fav_genreR.B              -0.35270    0.69692  -0.506  0.61296
Fav_genreRap              -1.31517    0.80986  -1.624  0.10484
Fav_genreRock              0.35068    0.54994   0.638  0.52390
Fav_genreVideo_game_music  0.41323    0.65971   0.626  0.53127
Frequency_Classical        0.27384    0.13214   2.072  0.03860 *
Frequency_Country         -0.27215    0.14327  -1.900  0.05791 .
Frequency_EDM              0.21091    0.12350   1.708  0.08812 .
Frequency_Metal            0.26856    0.13764   1.951  0.05144 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                   edf Ref.df    F p-value
s(Hours_per_day) 1.179  1.336 7.78 0.00206 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0482   Deviance explained = 7.51%
GCV = 9.3214  Scale est. = 9.0457    n = 716
```

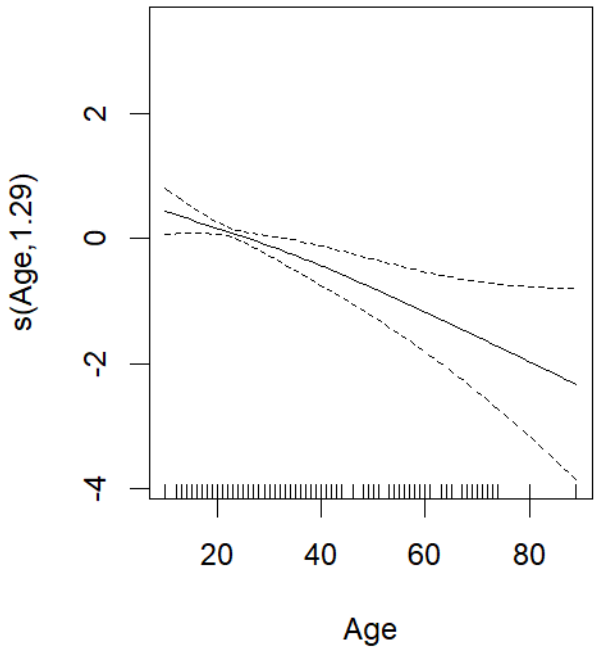
```
Family: gaussian
Link function: identity

Formula:
Insomnia ~ Hours_per_day + Fav_genreCountry + Fav_genreEDM +
    Fav_genreFolk + Fav_genreGospel + Fav_genreHip_hop + Fav_genreJazz +
    Fav_genreK_pop + Fav_genreLatin + Fav_genreLofi + Fav_genreMetal +
    Fav_genrePop + Fav_genreR.B + Fav_genreRap + Fav_genreRock +
    Fav_genreVideo_game_music + Frequency_Classical + Frequency_Country +
    Frequency_EDM + Frequency_Metal

Parametric coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.84883    0.73102   2.529 0.011656 *
Hours_per_day              0.13046    0.03890   3.354 0.000841 ***
Fav_genreCountry          -0.05436    0.88016  -0.062 0.950771
Fav_genreEDM               0.05691    0.76794   0.074 0.940942
Fav_genreFolk              0.55847    0.75111   0.744 0.457410
Fav_genreGospel            2.33078    1.31057   1.778 0.075769 .
Fav_genreHip_hop           0.05008    0.71294   0.070 0.944018
Fav_genreJazz             -0.03052    0.82358  -0.037 0.970445
Fav_genreK_pop            -0.18522    0.81264  -0.228 0.819771
Fav_genreLatin             0.02218    2.17947   0.010 0.991884
Fav_genreLofi              1.95471    1.06125   1.842 0.065916 .
Fav_genreMetal             0.61847    0.66391   0.932 0.351892
Fav_genrePop               0.15937    0.56550   0.282 0.778162
Fav_genreR.B              -0.34831    0.69696  -0.500 0.617402
Fav_genreRap              -1.32897    0.80924  -1.642 0.100994
Fav_genreRock              0.34930    0.55000   0.635 0.525578
Fav_genreVideo_game_music  0.40925    0.65972   0.620 0.535235
Frequency_Classical        0.27445    0.13215   2.077 0.038193 *
Frequency_Country         -0.27039    0.14324  -1.888 0.059483 .
Frequency_EDM              0.21241    0.12347   1.720 0.085812 .
Frequency_Metal            0.26911    0.13765   1.955 0.050979 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0479   Deviance explained = 7.46%
GCV = 9.3218  Scale est. = 9.0484    n = 716
```

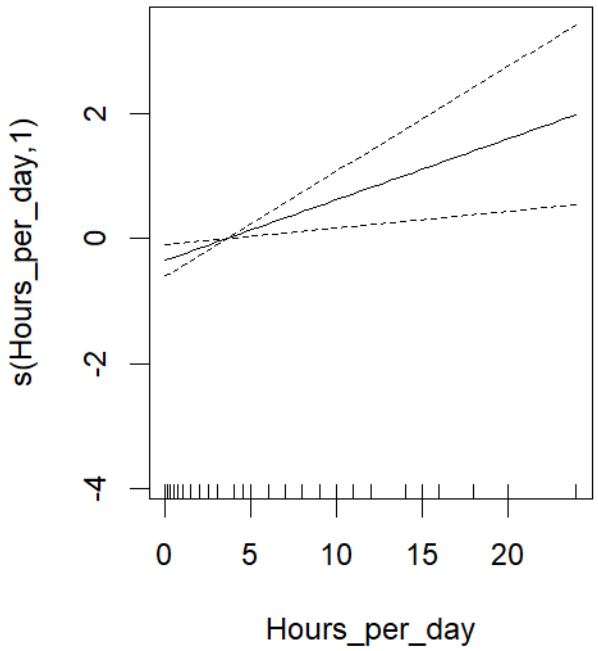


OCD

```
Family: gaussian
Link function: identity

Formula:
OCD ~ s(Age) + s(Hours_per_day) + Foreign_languages + Frequency_Country +
    Frequency_EDM

Parametric coefficients:
                  Estimate Std. Error t value       Pr(>|t|)
(Intercept)         2.1003     0.3257   6.449 0.000000000209 ***
Foreign_languages  -0.3363     0.2154  -1.561          0.1189
Frequency_Country   0.1895     0.1145   1.655          0.0984 .
Frequency_EDM       0.1833     0.1023   1.792          0.0735 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df     F p-value
s(Age)        1.294  1.534 9.813 0.00118 **
s(Hours_per_day) 1.000  1.000 7.585 0.00604 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0338   Deviance explained =  4.1%
GCV = 7.9057  Scale est. = 7.8362    n = 716
```

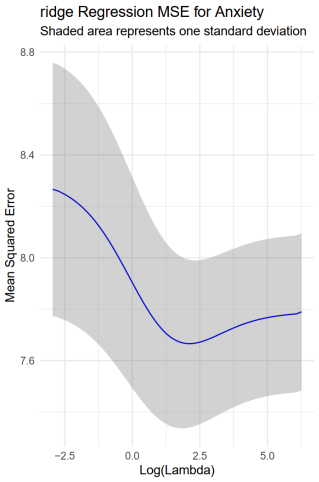
```
Family: gaussian
Link function: identity

Formula:
OCD ~ Age + Hours_per_day + Foreign_languages + Frequency_Country +
    Frequency_EDM

Parametric coefficients:
                  Estimate Std. Error t value       Pr(>|t|)
(Intercept)        2.56745    0.40635   6.318 0.000000000466 ***
Age               -0.03253    0.00892  -3.647        0.000285 ***
Hours_per_day      0.09531    0.03511   2.714        0.006799 **
Foreign_languages -0.33044    0.21523  -1.535        0.125153
Frequency_Country  0.19070    0.11453   1.665        0.096335 .
Frequency_EDM      0.18669    0.10216   1.827        0.068048 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0332   Deviance explained =    4%
GCV =  7.907  Scale est. = 7.8407    n = 716
```

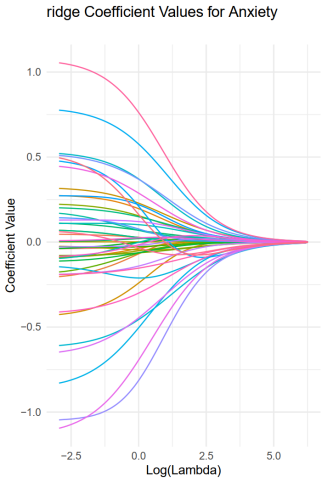


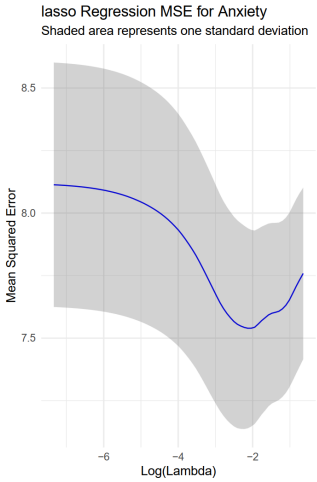


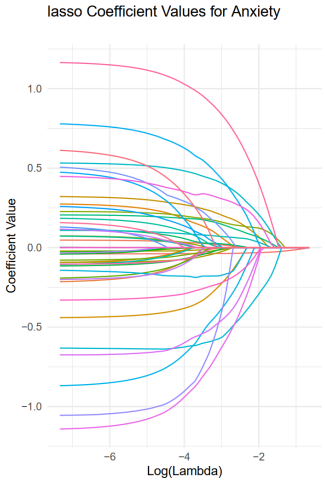
APPENDIX C. Regularization Regression Plots (RQ2)

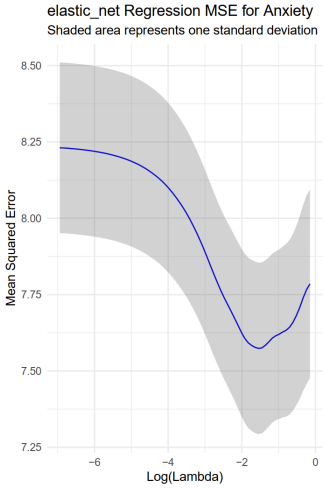
C.1. Anxiety Models

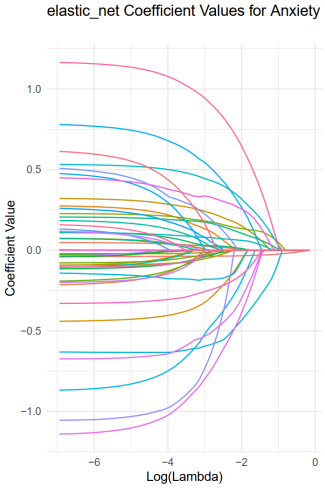




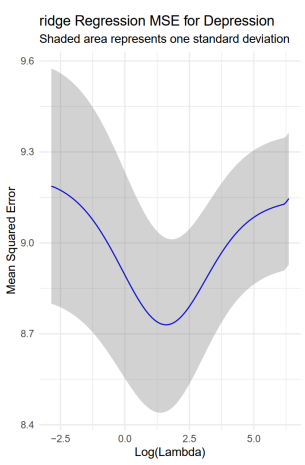


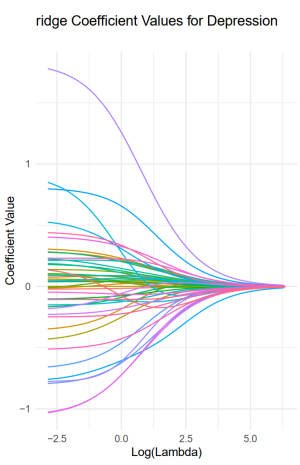


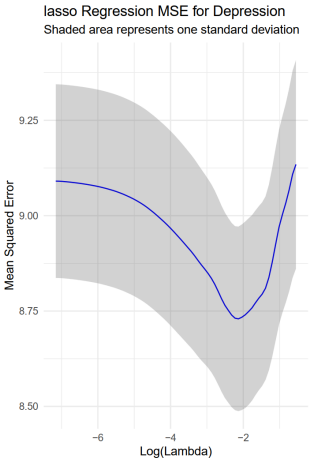


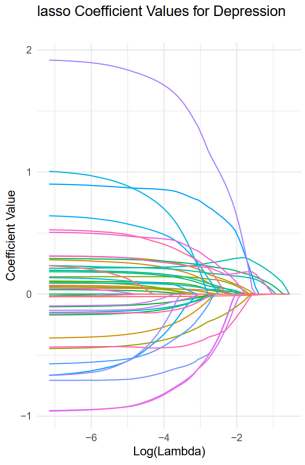


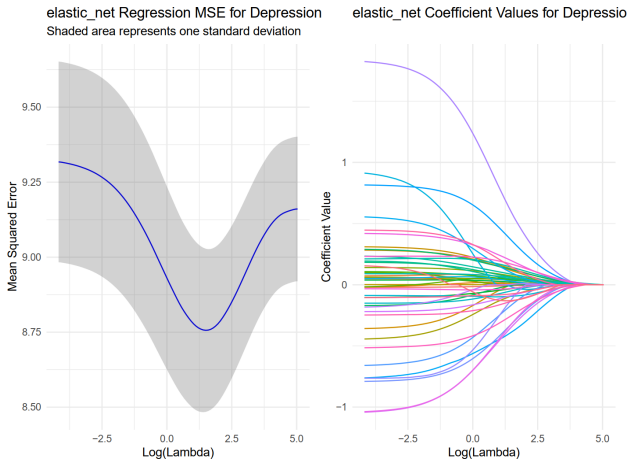
C.2. Depression Models



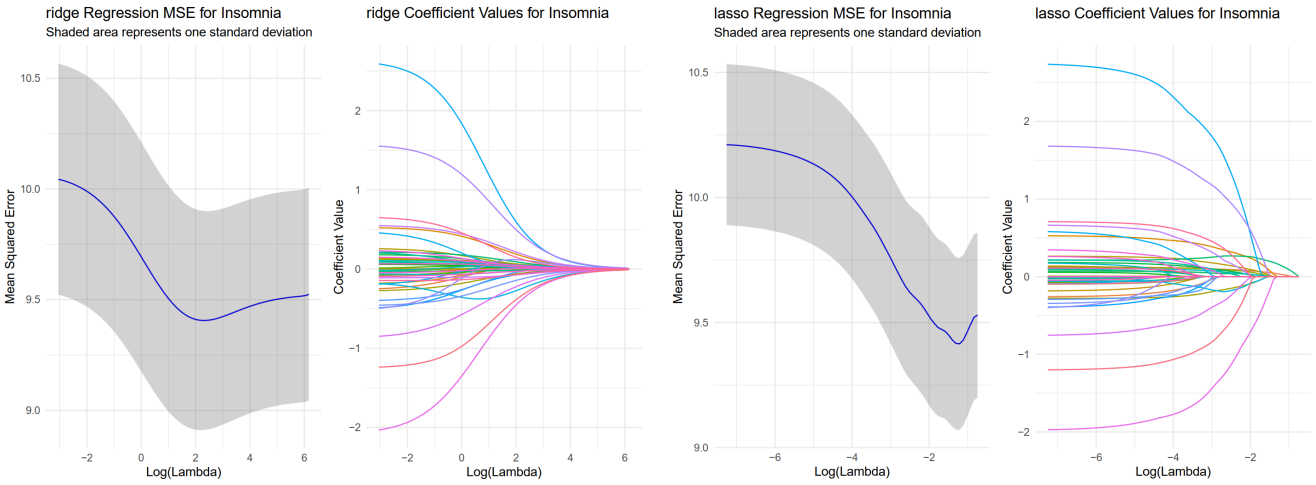


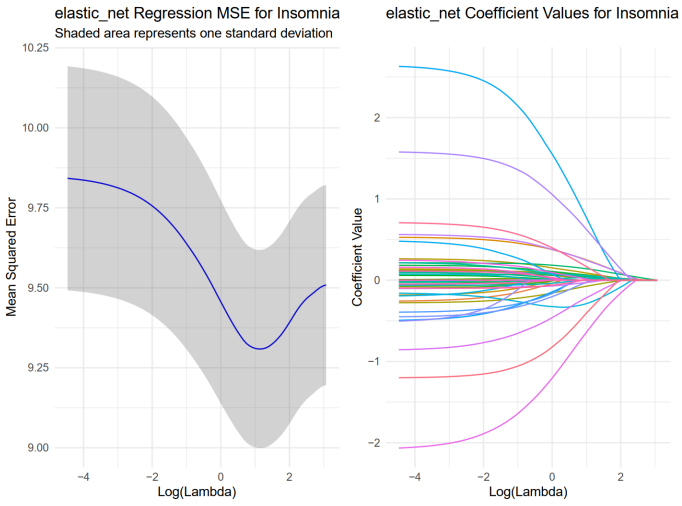




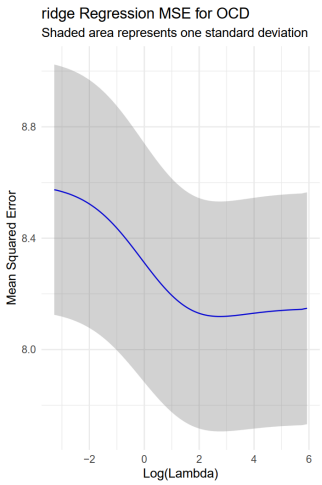


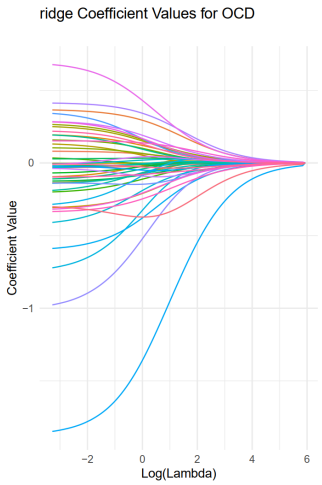
C.3. Insomnia Models

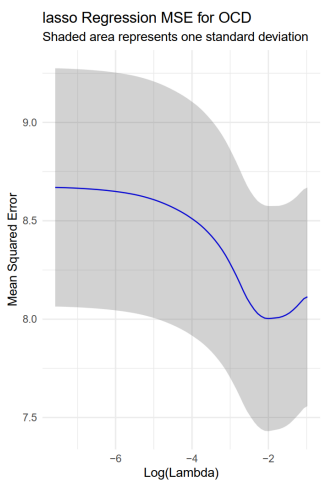


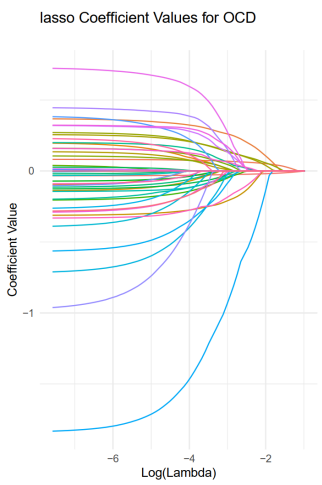


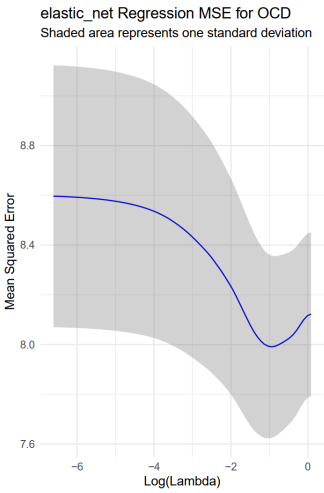
C.4. OCD Models

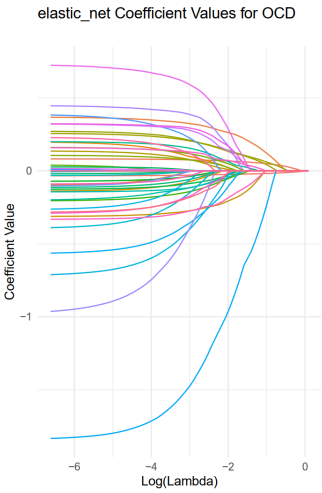












APPENDIX D. Regularized Models (RQ3)