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

Dieu-Anh Le, Rohith Raj Srinivasan

UC DAVIS

Background & Motivation

- Music is a **ubiquitous** part of daily routines, influencing activities like **work, exercise, and relaxation**.
- Music significantly **impacts mood and stress levels**, making it a potential tool for emotional regulation.
- With rising mental health issues, exploring music effects on conditions like **anxiety and depression** is crucial.
- Understanding the complex correlation of musical preferences can lead to **personalized music therapy**, enhancing its effectiveness.
- Analyzing a **rich dataset on musical habits and mental health** can provide empirical insights into their interplay, guiding future research.

Research Questions

1. Besides streaming music, are “musicians” associated with better mental health conditions?

2. What are the individual characteristics associated with mental well-being?

3. Can individual characteristics predict the music effects? If so, how accurately?

Dataset Overview

Dataset (Music & Mental Health Survey Results)

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	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)

Raw data collected via a Google form

736 entries
33 columns

Data source: https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results/data

Quantitative summary (Numerical)

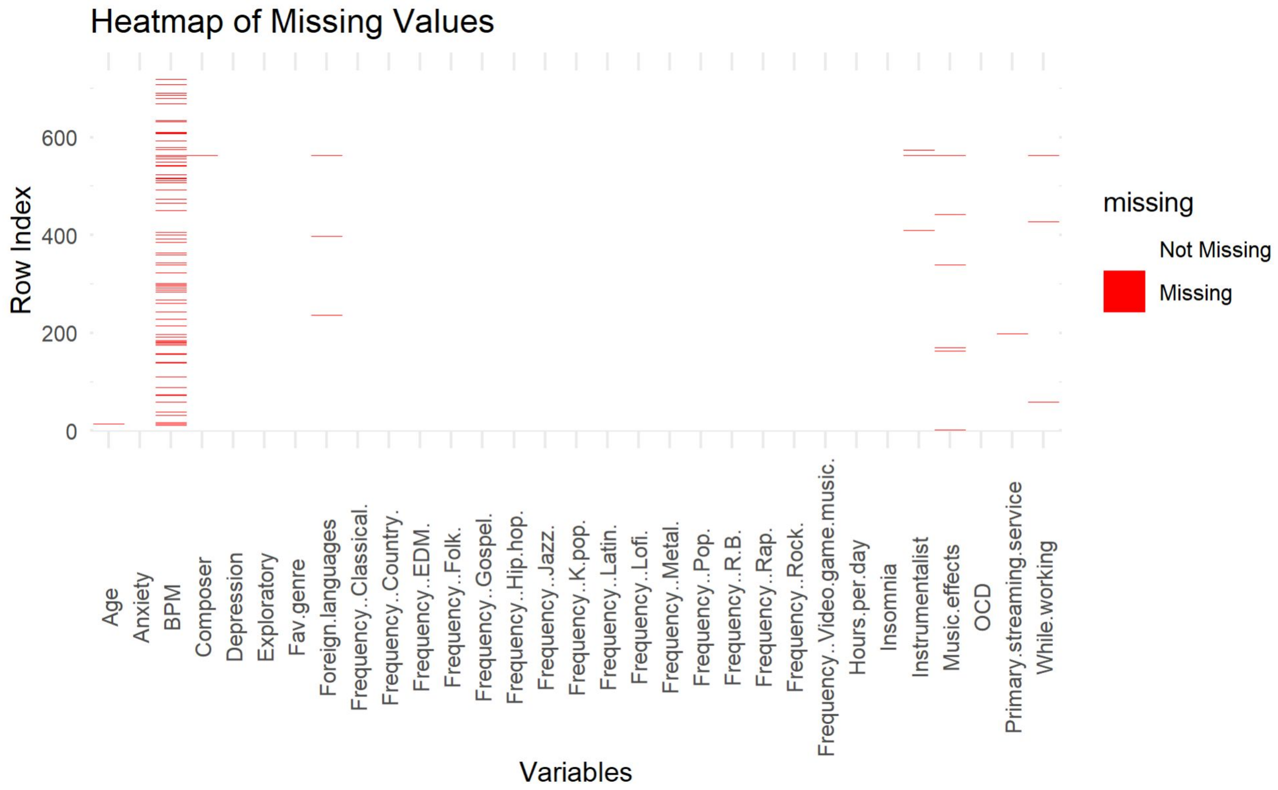
```
Numerical Summary:
               count          mean           std   min    25%    50%    75%  \
Age            735.0  2.520680e+01  1.205497e+01  10.0   18.0   21.0   28.0
Hours per day  736.0  3.572758e+00  3.028199e+00   0.0    2.0    3.0    5.0
BPM            629.0  1.589948e+06  3.987261e+07   0.0  100.0  120.0  144.0
Anxiety        736.0  5.837636e+00  2.793054e+00   0.0    4.0    6.0    8.0
Depression     736.0  4.796196e+00  3.028870e+00   0.0    2.0    5.0    7.0
Insomnia       736.0  3.738451e+00  3.088689e+00   0.0    1.0    3.0    6.0
OCD            736.0  2.637228e+00  2.842017e+00   0.0    0.0    2.0    5.0

                       max  missing_values   mode
Age                   89.0               1   18.0
Hours per day         24.0               0    2.0
BPM            999999999.0             107  120.0
Anxiety               10.0               0    7.0
Depression            10.0               0    7.0
Insomnia              10.0               0    0.0
OCD                   10.0               0    0.0
```

Quantitative summary (Categorical)

Column	Num Unique Categories	Most Frequent Category	Count of Most Frequent
Primary streaming service	7	Spotify	458
While working	3	Yes	579
Instrumentalist	3	No	497
Composer	3	No	609
Fav genre	16	Rock	188
Exploratory	2	Yes	525
Foreign languages	3	Yes	404
Frequency [Classical]	4	Rarely	259
Frequency [Country]	4	Never	343
Frequency [EDM]	4	Never	307
Frequency [Folk]	4	Never	292
Frequency [Gospel]	4	Never	535
Frequency [Hip hop]	4	Sometimes	218
Frequency [Jazz]	4	Never	261
Frequency [K pop]	4	Never	416
Frequency [Latin]	4	Never	443
Frequency [Lofi]	4	Never	280
Frequency [Metal]	4	Never	264
Frequency [Pop]	4	Very frequently	277
Frequency [R&B]	4	Never	225
Frequency [Rap]	4	Rarely	215
Frequency [Rock]	4	Very frequently	330
Frequency [Video game music]	4	Never	236
Music effects	4	Improve	542

Data Preprocessing

Missing values and NA's



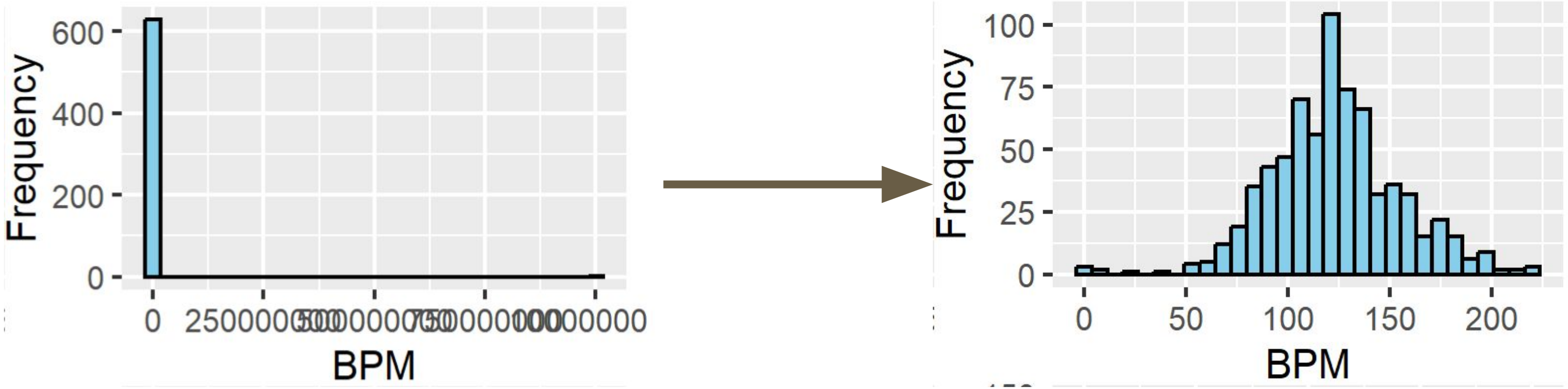
Approaches:

1. Drop
2. Regression modelling

BPM Prediction

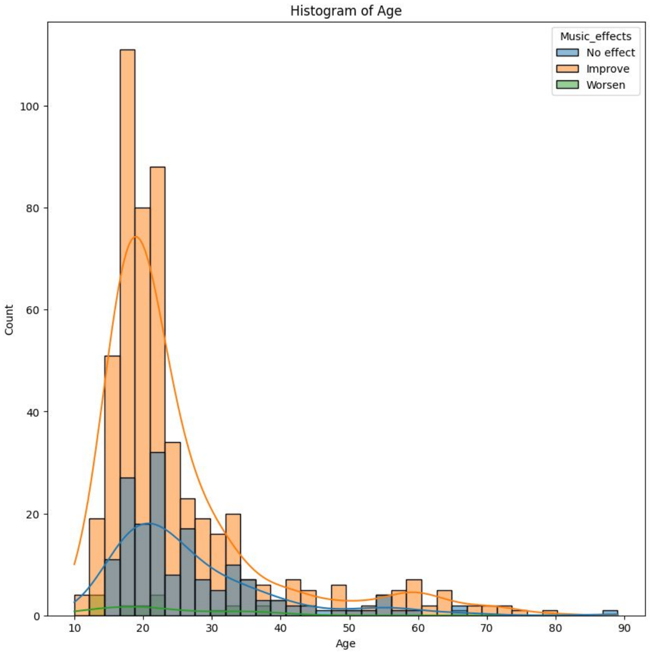
1. Drop non-BPM NA values
2. Drop two highest values of BPM which we suspected to be outliers
3. Forward stepwise selection based on AIC for BPM modelling
4. Predict missing BPM values

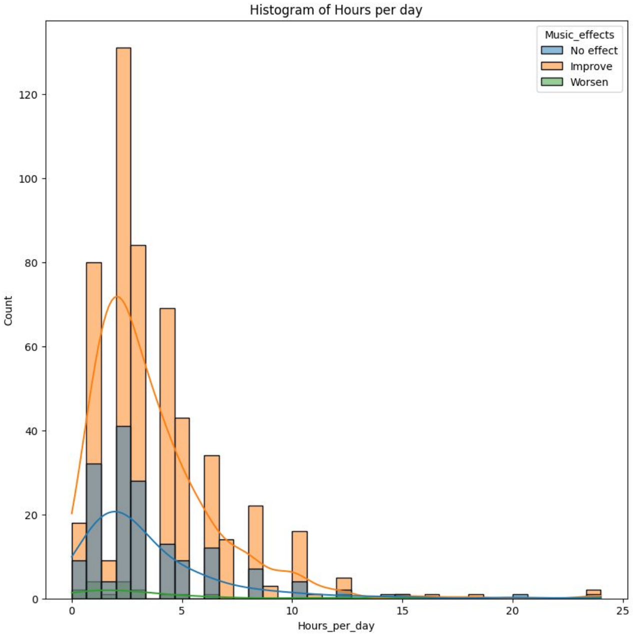
Approximately Normal!

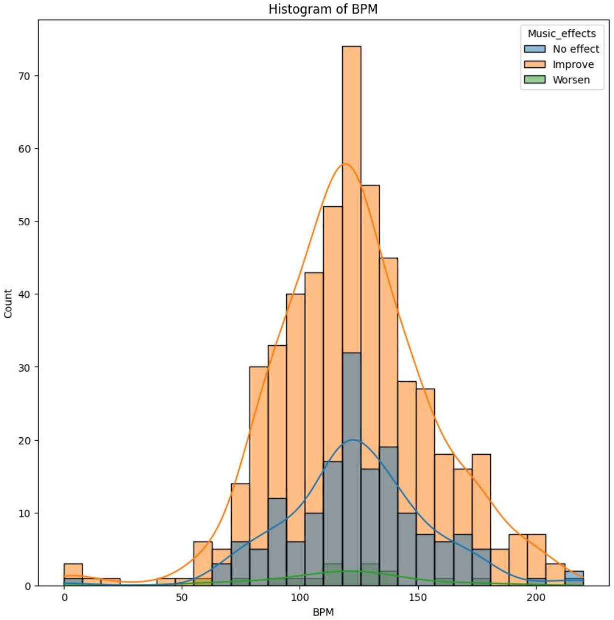


EDA + Descriptive Statistics

Music effects (Age, Hours_per_day, BPM)



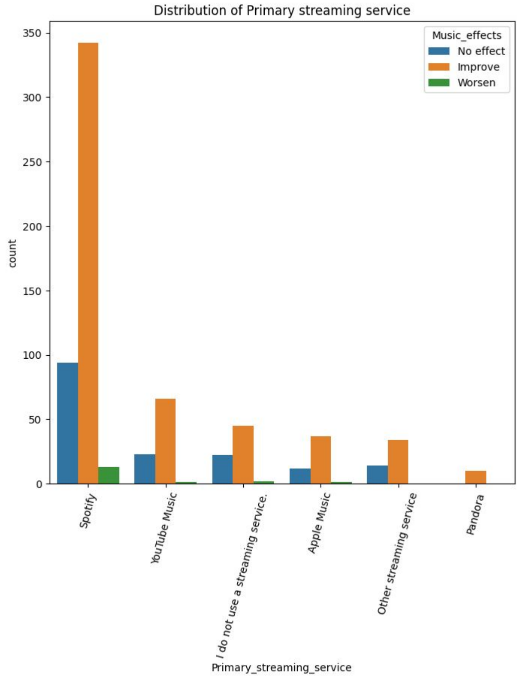


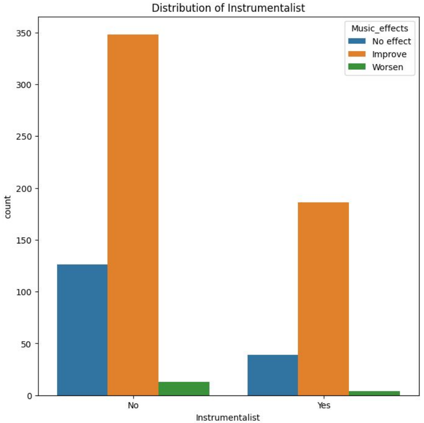


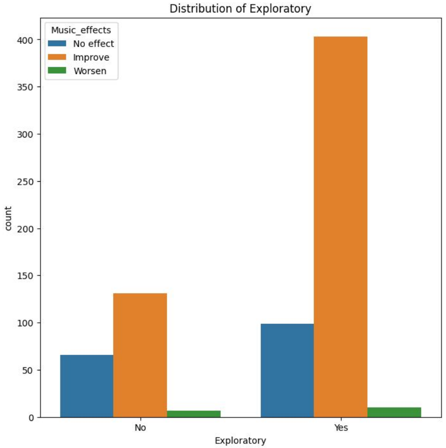
Age: 15-35 most common

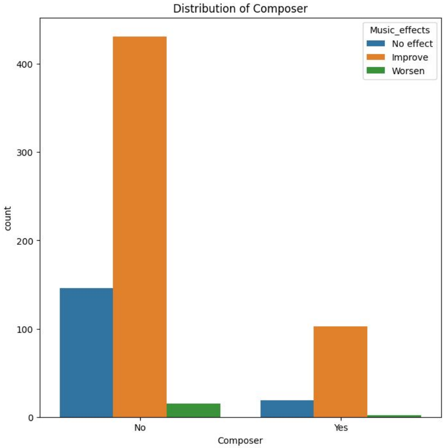
Hours_per_day: <10 most common

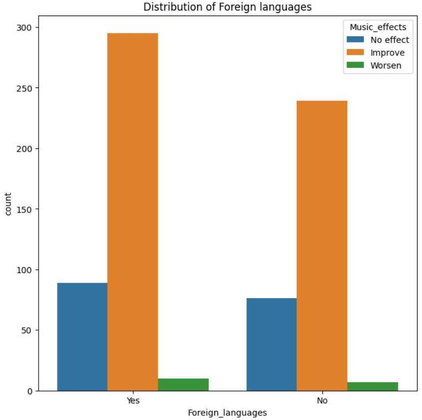
Music effects (Categorical columns)

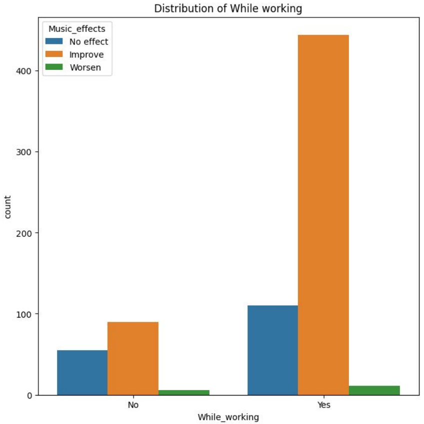






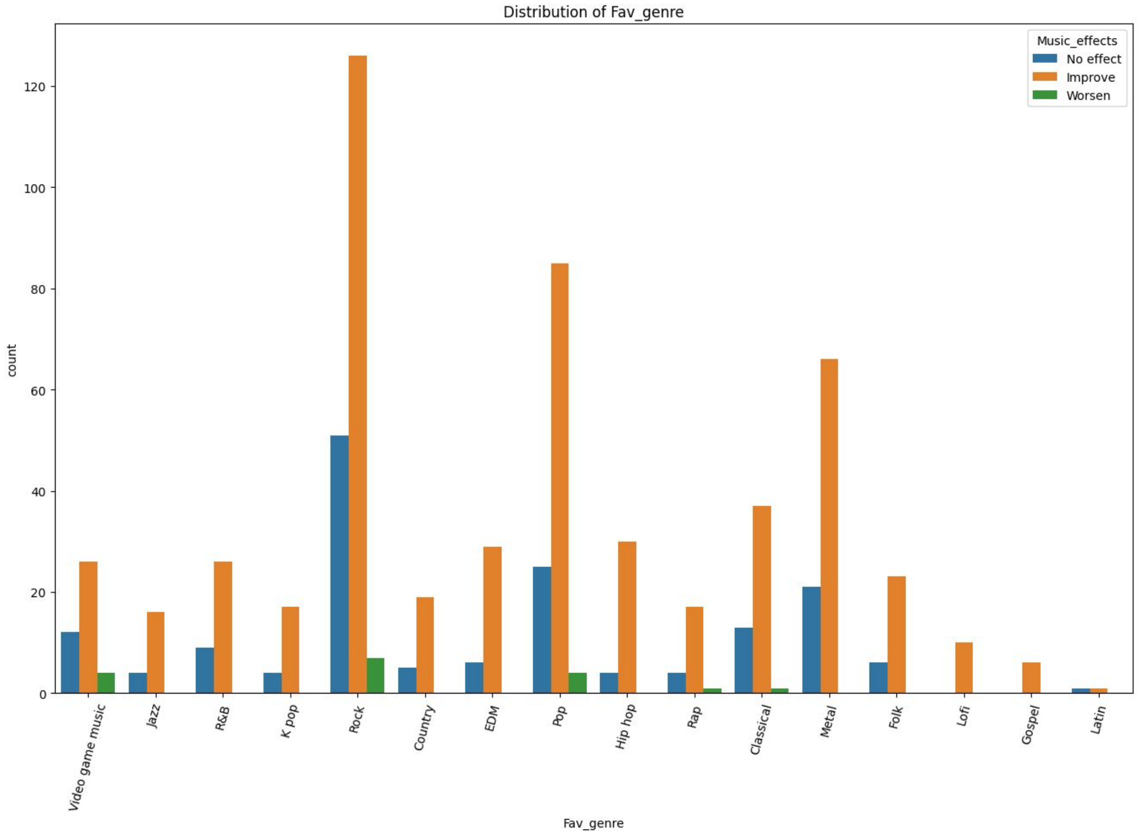






Observation: For each categorical variable, levels are not approximately equal in size!

Music effects (Fav_genre)



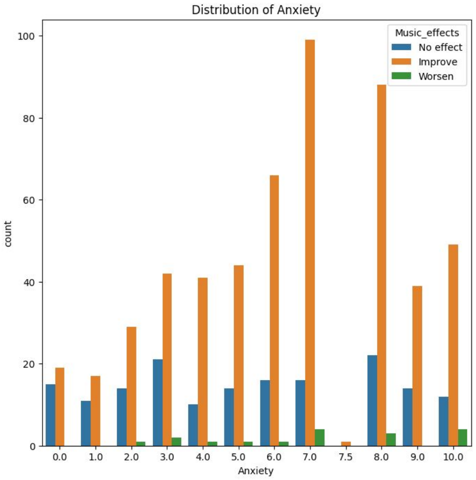
Top 3 fav_genre

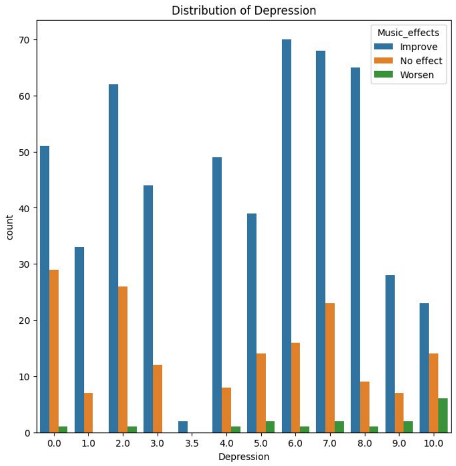
- Rock
- Pop
- Metal

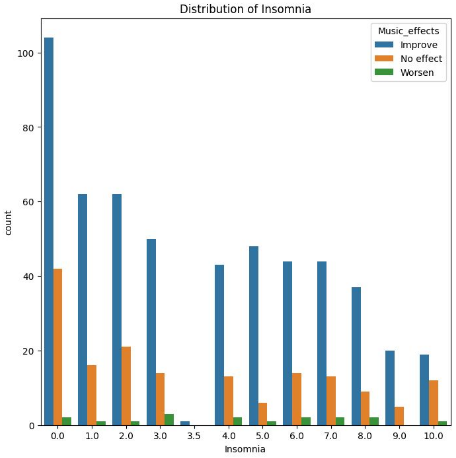
Music effects (Anxiety, Depression, Insomnia, OCD)

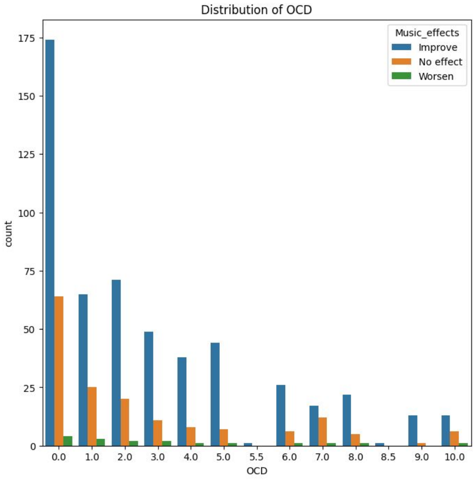
Observations:

- Music has positive effects across all self-reported levels of mental health conditions.
- Depression and anxiety turned out to be most common mental health conditions.





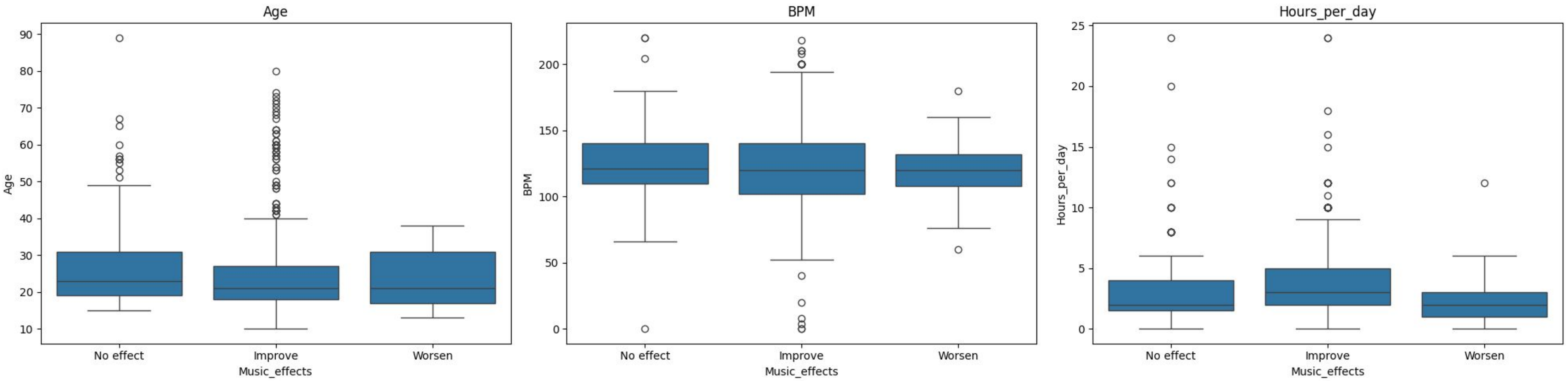




Music effects (Frequency_<Genre>)

Majority of the listeners across all music genres, regardless of how frequently they listen to, experienced positive music effects on their mental health status!

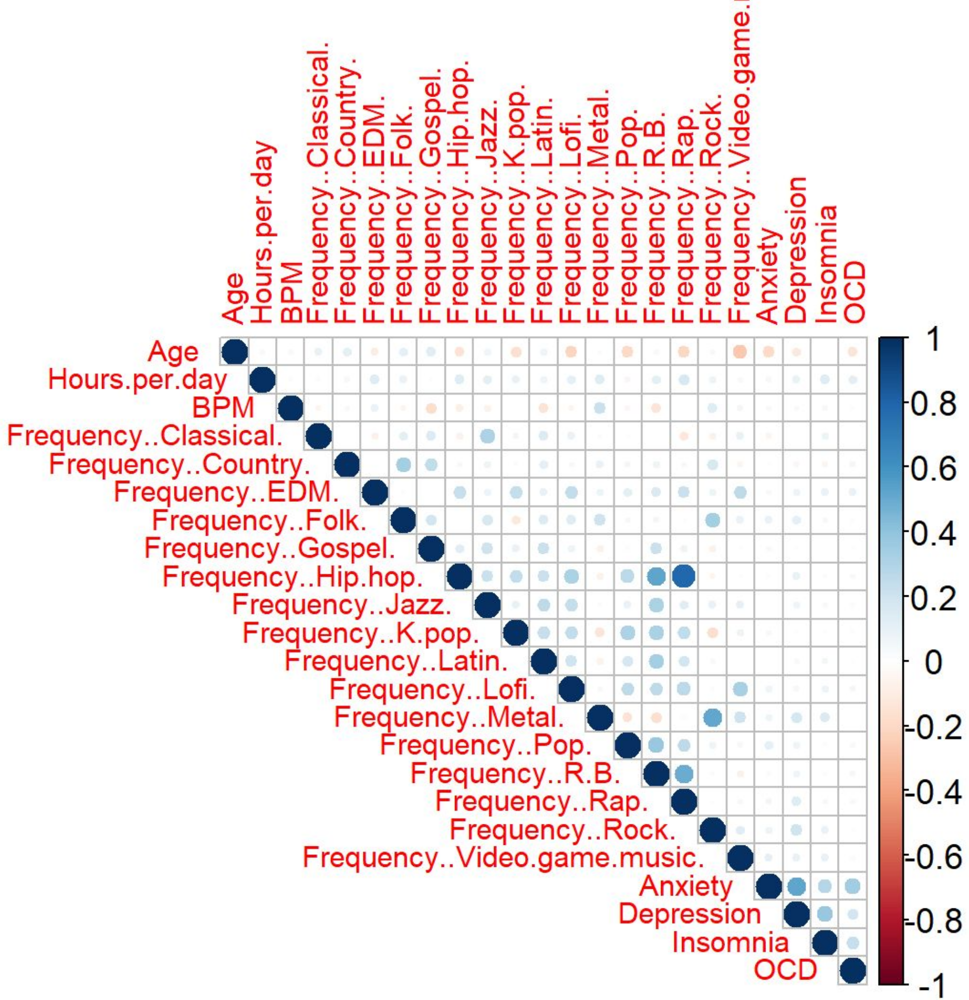
Outlier detection



We considered these as True Outliers and retain them in the dataset.

Correlation Matrix

No Collinearity between variables!
(except *Frequency of Hiphop* and *Rap*)



Comparative Analysis

Research Question 1:
Besides streaming music, are "musicians" associated with better mental health conditions?

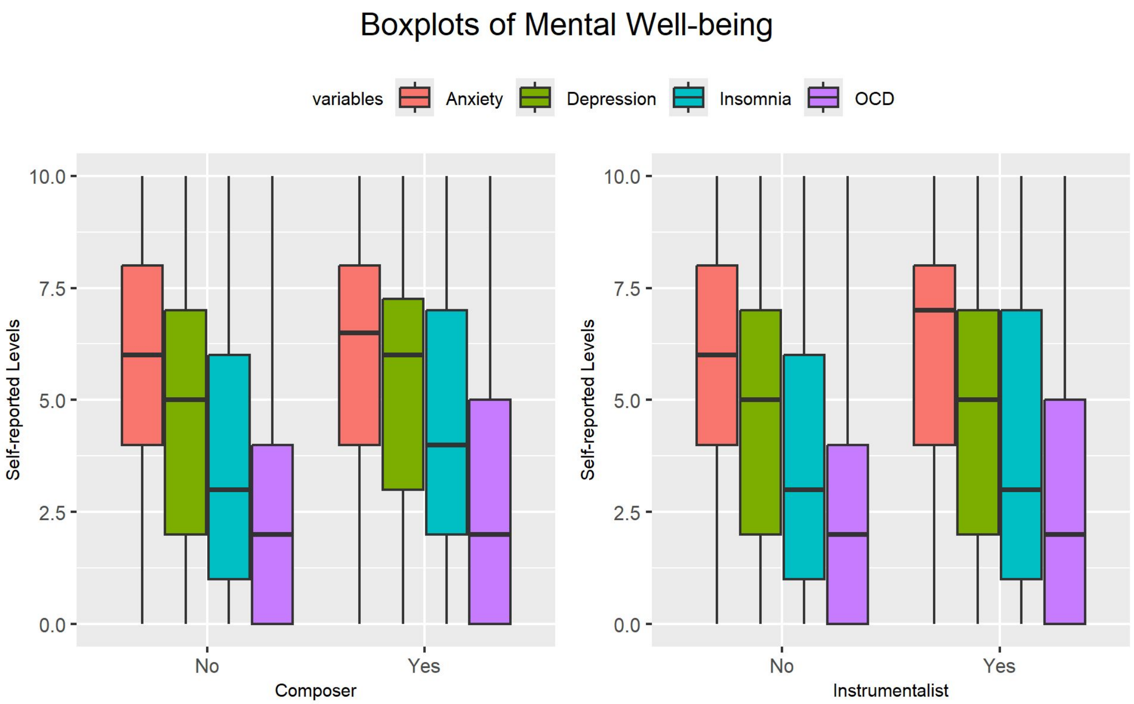
Approach

1. T-test between:
   a. Composer vs Non-Composer
   b. Instrumentalist vs Non-Instrumentalist
2. Anova-tests between:
   a. Composer and Instrumentalist
   b. Composer but Not-Instrumentalist
   c. Not-Composer but Instrumentalist
   d. Neither Composer nor Instrumentalist

Testing hypothesis:

H0: There is no significant difference in the average self-reported level of mental health conditions across the two groups.

Composer and Instrumentalist



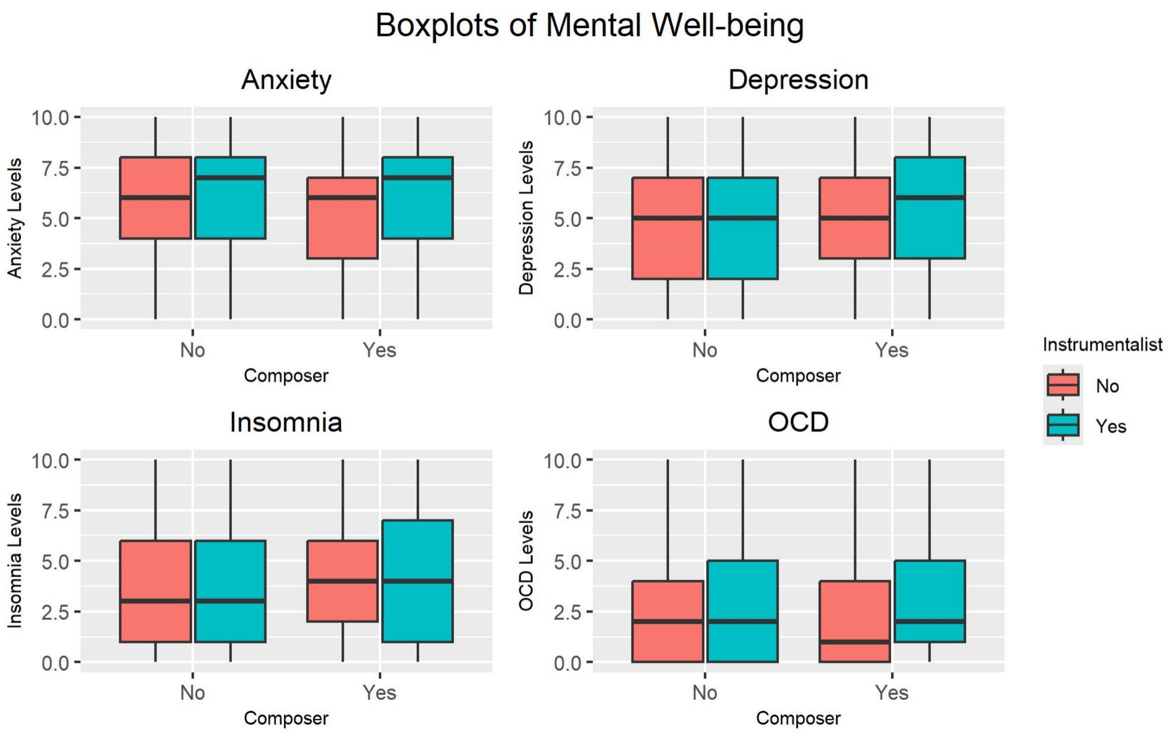
Groups:

1. Composer - Non-Composer

2. Instrumentalist - Non-Instrumentalist

No Statistical Difference !

Composer * Instrumentalist (Interaction!)



Groups:

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

No Statistical Difference !

Inferential Modelling

Research Question 2:
What are the individual characteristics associated with mental well-being?

Approach

Linear Modelling compare and contrast across different models for robustness

1. 80% training set, 20% testing set
2. Bi-directional stepwise selection (AIC, BIC, Adjusted R-squared, Cp)
3. Regularization regression (Ridge, Lasso, Elastic Net)
4. Final model: compare and decide

Non-Linear Modelling explore the non-linear relationship between predictor and response variables

1. GAM

Feature Selection (Test MSE)

	Anxiety	Depression	Insomnia	OCD
AIC	7.927725484	9.473937	10.01085	7.66598
BIC	8.218664192	9.590871	10.40131	7.617343
Cp	8.032062473	9.571132	10.0822	7.638846
Adjusted R2	8.24567377	9.6052	10.11025	7.896986
Ridge	8.167785671	9.372826	9.848177	7.858744
Lasso	8.217512657	9.583269	9.893856	7.731363
Elastic Net	8.231344735	9.394042	9.860527	7.73137

Anxiety Model (14 variables)

Akaike Information Criterion (AIC)

+ Low test error
- Low R-squared :(

```
Residuals:
    Min      1Q  Median      3Q     Max 
-6.7344 -1.8890  0.3237  1.9208  5.8462 

Coefficients:
                               Estimate Std. Error t value               Pr(>|t|)    
(Intercept)                    6.091570   0.426440  14.285 < 0.0000000000000002 ***
Age                           -0.053558   0.009696  -5.523         0.0000000509 ***
Frequency_MetalRarely          0.494780   0.293194   1.688             0.092053 .  
Frequency_MetalSometimes       1.136955   0.329741   3.448             0.000607 ***
Frequency_MetalVery_frequently 0.140160   0.314872   0.445             0.656394    
Music_effectsImprove           0.802866   0.265138   3.028             0.002574 ** 
Music_effectsWorsen            1.546798   0.739497   2.092             0.036915 *  
Frequency_FolkRarely           0.790218   0.280860   2.814             0.005071 ** 
Frequency_FolkSometimes        0.901312   0.315320   2.858             0.004416 ** 
Frequency_FolkVery_frequently  0.490316   0.386423   1.269             0.205017    
Frequency_LofiRarely          -0.308453   0.280455  -1.100             0.271878    
Frequency_LofiSometimes        0.528040   0.308456   1.712             0.087471 .  
Frequency_LofiVery_frequently -0.543342   0.379990  -1.430             0.153306    
ExploratoryYes                -0.431413   0.262391  -1.644             0.100703    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.618 on 561 degrees of freedom
Multiple R-squared:  0.1184,	Adjusted R-squared:  0.09796 
F-statistic: 5.795 on 13 and 561 DF,  p-value: 0.000000004476
```

Depression Model (30 variables)

	s0
(Intercept)	3.627035549
Age	-0.007837104
Hours_per_day	0.016891055
While_working	.
Instrumentalist	.
Composer	0.128856552
Exploratory	.
Foreign_languages	0.060682824
BPM	.
Frequency_Classical	.
Frequency_Country	-0.034873667
Frequency_EDM	0.025628363
Frequency_Folk	0.063369812
Frequency_Gospel	0.001477032
Frequency_Hip_hop	0.015892477
Frequency_Jazz	.
Frequency_K_pop	.
Frequency_Latin	0.019971645
Frequency_Lofi	.
Frequency_Metal	0.133197617
Frequency_Pop	0.024296322
Frequency_R_B	0.002614535
Frequency_Rap	0.066143555
Frequency_Rock	0.149327443
Frequency_Video_game_music	0.057194271
Music_effects	-0.047504075
Fav_genreClassical	-0.097079749
Fav_genreCountry	.
Fav_genreEDM	0.067498777
Fav_genreFolk	.
Fav_genreGospel	.
Fav_genreHip_hop	0.173268957
Fav_genreJazz	.
Fav_genreK_pop	-0.053932118
Fav_genreLatin	.
Fav_genreLofi	0.830382973
Fav_genreMetal	.
Fav_genrePop	-0.020537331
Fav_genreR&B	-0.227895525
Fav_genreRap	-0.127505970
Fav_genreRock	0.187123244
Fav_genreVideo_game_music	.
Primary_streaming_serviceSpotify	0.069314316
Primary_streaming_serviceYouTube_Music	-0.024431666
Primary_streaming_serviceI_do_not_use_a_streaming_service.	-0.105412744
Primary_streaming_serviceApple_Music	0.111164815
Primary_streaming_serviceOther_streaming_service	.
Primary_streaming_servicePandora	.

Insomnia Model (5 variables)

(Intercept)	3.02942318
Age	0.01804476
Hours_per_day	0.17726028
While_working	0.02080210
Instrumentalist	0.21135026
Composer	0.07844320

Although Ridge yields the lowest test error, inclusion of all variables reduces the model interpretability, hence Elastic Net was used instead!

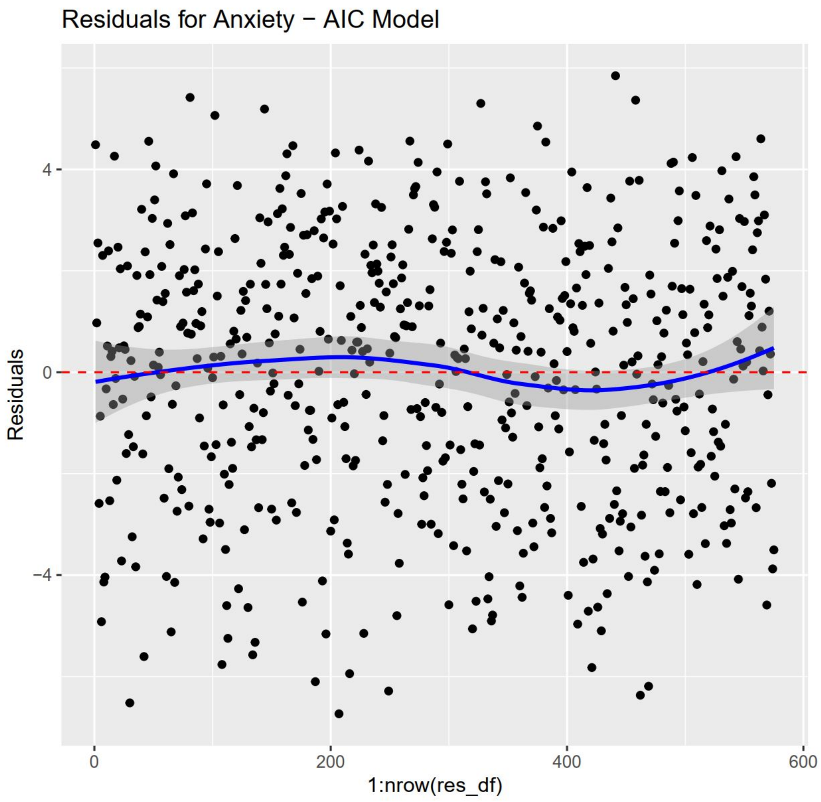
Note that there were not much variation in test error across all models.

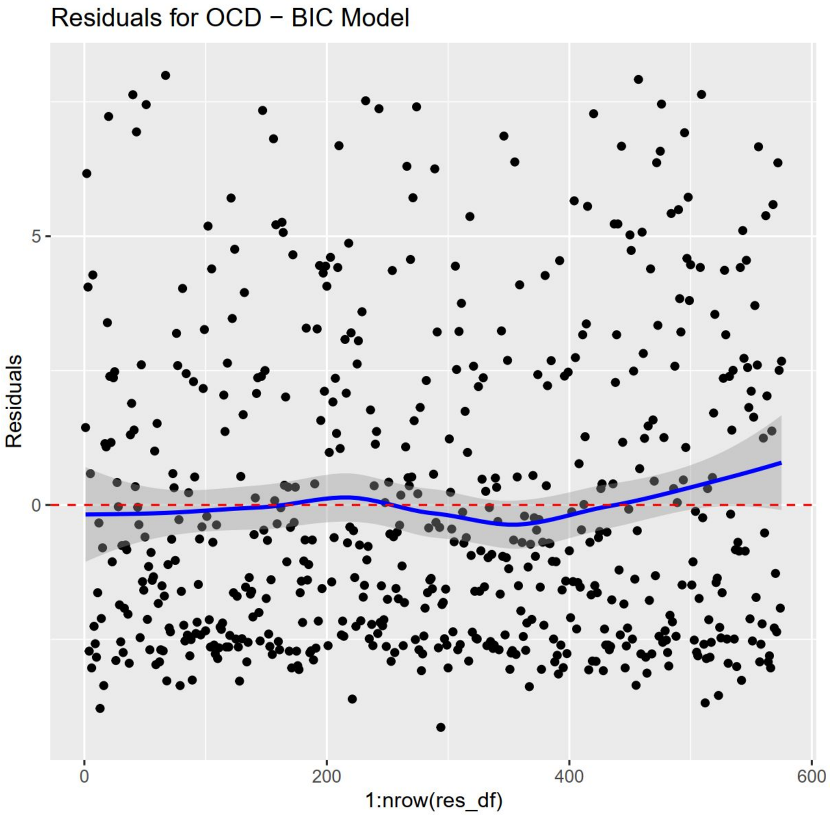
OCD Model (2 variables)

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.1389 -2.3926 -0.7564  1.8519  7.9927 

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   2.844962   0.313092   9.087 < 0.0000000000000002 ***
Age          -0.025687   0.009933  -2.586              0.00996 ** 
Hours_per_day  0.112706   0.038891   2.898              0.00390 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.823 on 572 degrees of freedom
Multiple R-squared:  0.02678,	Adjusted R-squared:  0.02338 
F-statistic: 7.871 on 2 and 572 DF,  p-value: 0.0004244
```

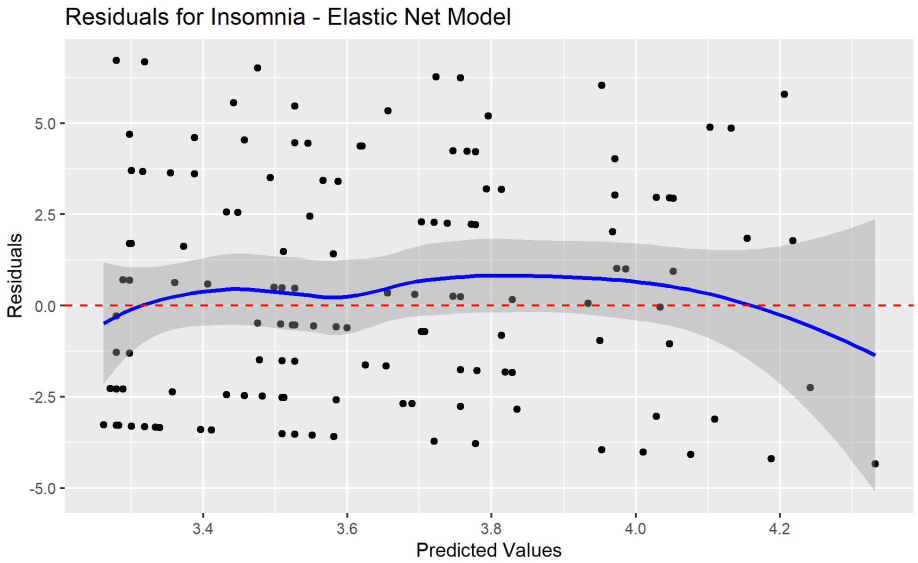
Music preferences have almost no correlation with OCD (?)

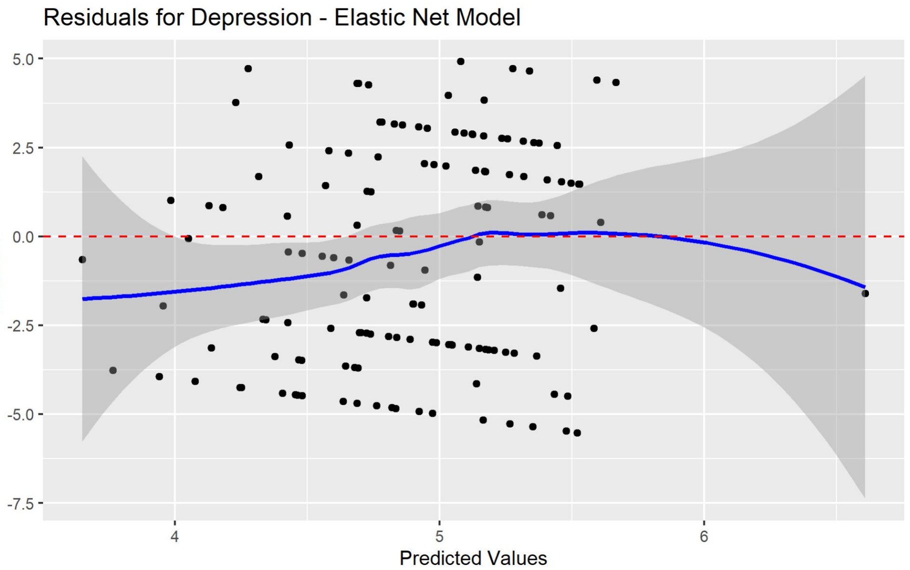
Residual Plots (Anxiety and OCD)



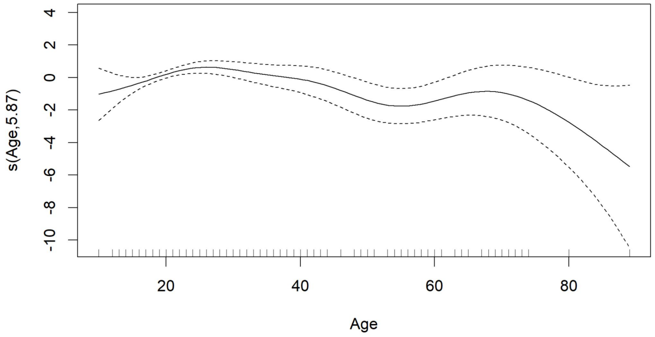


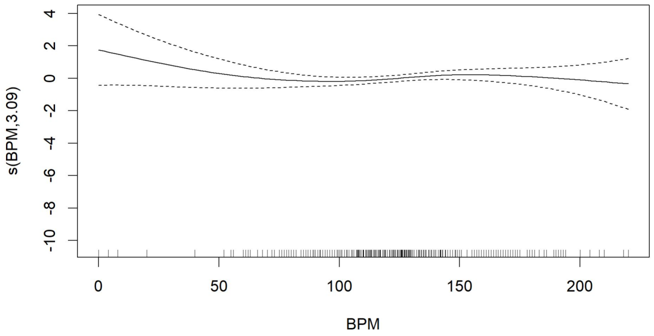
Residual Plots (Insomnia and Depression)

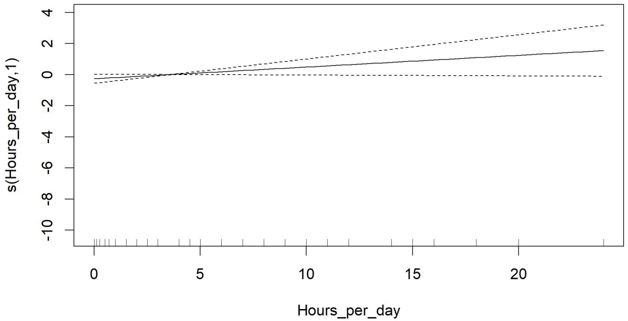




Linearity Assessment between predictors and Depression using GAM







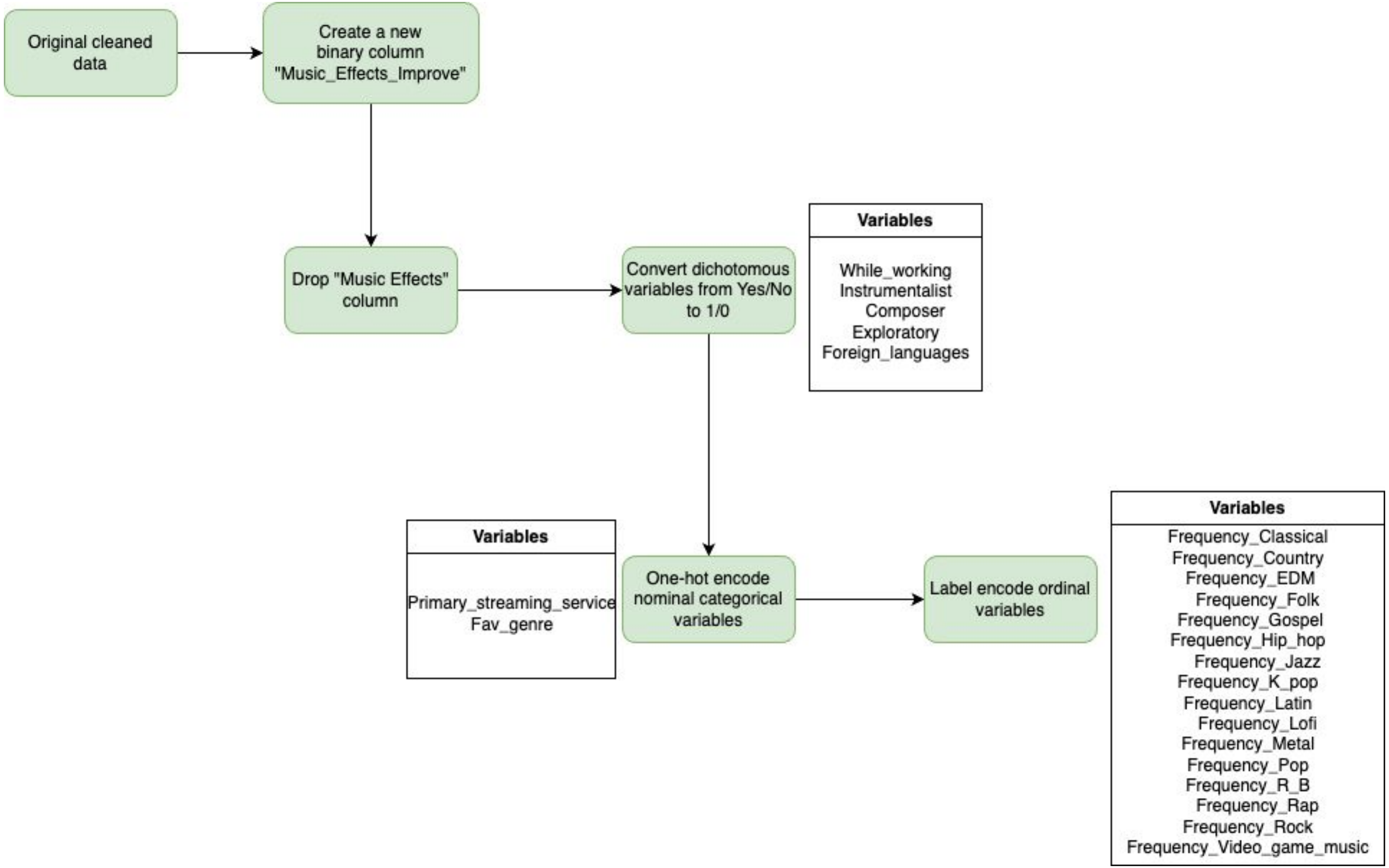
Classification Modelling

Research Question 3:
Can individual characteristics predict the music effects? If so, how accurately?

Approach

- Data preprocessing
- Splitting the data: training = 80%, testing = 20%
- Training models on all features and reporting the performance
- Feature selection:
  - Stepwise forward selection (AIC, BIC)
  - Regression (Lasso, Ridge, Elastic Net)
- Re-train all the models on the subset of features selected, and report performance
- Compare and decide

Data Preprocessing for Parametric methods



Model performance without feature selection

Parametric models

Model	Accuracy	Test Error	
Logistic Regression	0.7153	0.2847	No penalty
LDA	0.7083	0.2917	
Random Forest (non-parametric)	0.7500	0.2500	n_estimators = 100
Neural Network 1	0.7083	0.2917	2 hidden layers (64 * 32) 5 epochs
Neural Network 2	0.7222	0.2778	3 hidden layers (64 * 32 * 16) 5 epochs

Feature selection - Stepwise Forward Selection

```
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
```

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
```

Bayesian Information Criterion (BIC)

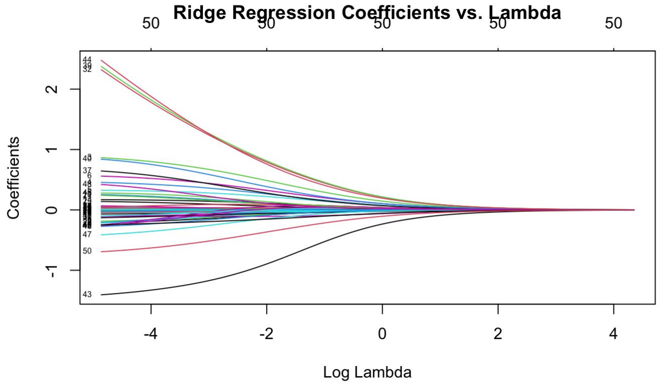
While_working +
Exploratory +
Anxiety +
Frequency_R_B

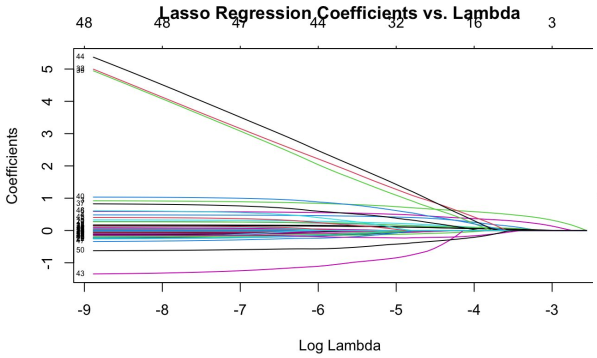
Model performance with feature selection

Parametric models

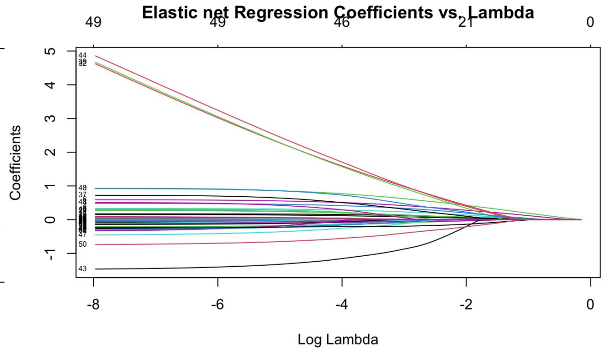
Model	Accuracy	Test Error	
Logistic Regression	0.7778	0.2222	No penalty
LDA	0.7639	0.2361	
Random Forest (non-parametric)	0.7361	0.2639	n_estimators = 100
Neural Network 1	0.7569	0.2431	2 hidden layers (64 * 32) 5 epochs
Neural Network 2	0.7778	0.2222	3 hidden layers (64 * 32 * 16) 5 epochs

Feature selection - Regularization





32 zeroed out



36 zeroed out

While_working
Instrumentalist
Composer
Exploratory
Frequency_Country
Frequency_Gospel

Frequency_R_B
Anxiety
Primary_streaming_service_PandoraTrue
Fav_genre_GospelTrue

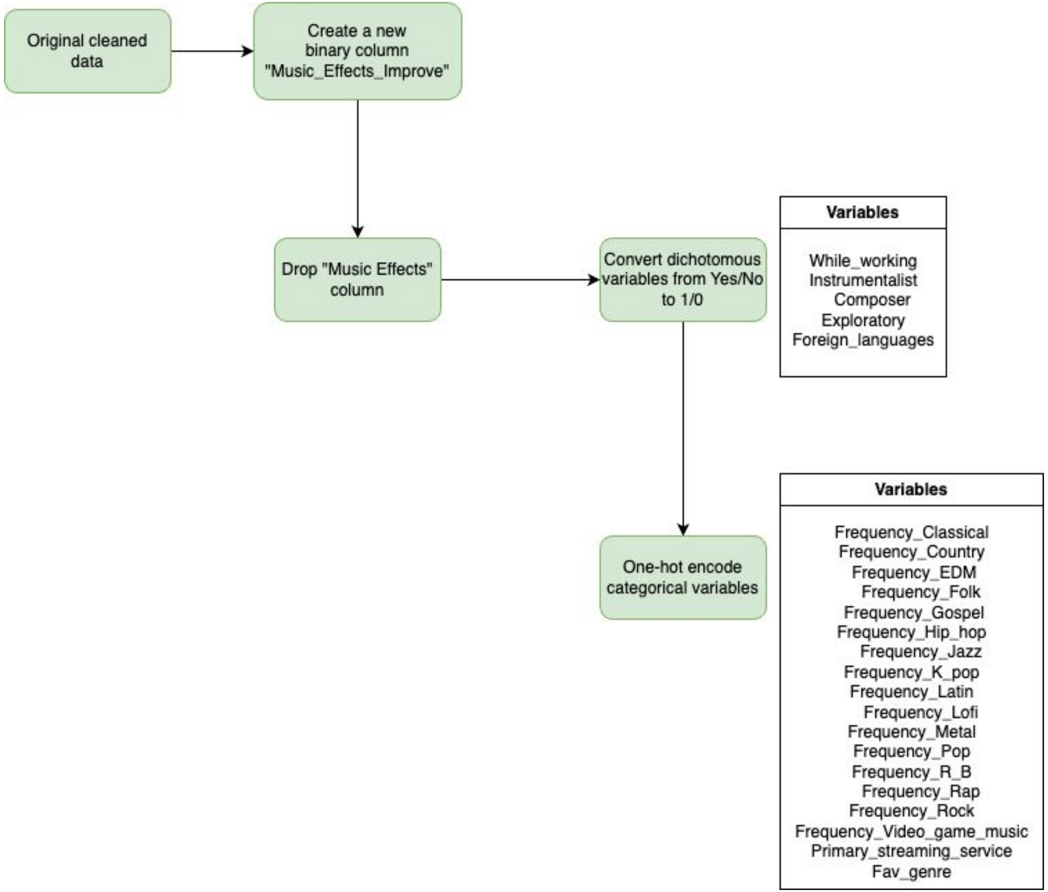
Fav_genre_Hip.hopTrue
Fav_genre_LofiTrue
Fav_genre_RockTrue
Fav_genre_Video.game.musicTrue

Model performance with feature selection

Parametric models

Model	Accuracy	Test Error	
Logistic Regression	0.7569	0.2431	No penalty
LDA	0.7361	0.2639	
Random Forest (non-parametric)	0.7083	0.2917	n_estimators = 100
Neural Network 1	0.7222	0.2778	2 hidden layers (64 * 32) 5 epochs
Neural Network 2	0.7500	0.2500	3 hidden layers (64 * 32 * 16) 5 epochs

Data Preprocessing for Non-parametric methods



Model performance

Non-Parametric models

Model	Accuracy	Test Error	
KNN	0.7778	0.2222	Best k = 17 in [1, 30]
Random Forest	0.7431	0.2569	n_estimators = 100
SVC 1 (parametric)	0.7222	0.2778	Kernel = Linear
SVC 2	0.7500	0.2500	Kernel = Polynomial
SVC 3	0.7570	0.2430	Kernel = Radial Basis Function (RBF)

Research Questions - Answered

1. Besides streaming music, are “musicians” associated with better mental health conditions?

   No, there were no difference across all groups of Composer and Instrumentalists.

2. What are the individual characteristics associated with mental well-being?

   Anxiety ~ Age + Frequency_Metal + Music_effects + Frequency_Folk + Frequency_Lofi + Exploratory

   Depression ~ !(While_working + Instrumentalist + Explanatory + some_Frequency + some_Fav_genre + some_Primary_streaming_service)

   Insomnia ~ Age + Hours_per_day + While_working + Instrumentalist + Composer

   OCD ~ Age + Hours_per_day

   However, R-squared values are very low, indicating limited variations in mental health levels were explained by individual characteristics.

3. Can individual characteristics predict the music effects? If so, how accurately?

   Yes, our models were able to predicts music effects from the individual characteristics with an accuracy rate ~ 0.78.

THANK YOU!

Learning & Knowledge

- Quantitative analysis of dataset
- Plotting complex relationships
- Handling missing values
- Outlier detection
- Data pre-processing
- Linear regression (interaction terms, significance, eval. metrics)
- Non-linear regression (GAM)
- Stepwise forward selection (AIC, BIC, AdjR2, Cp)
- Regularization (Ridge, Lasso, Elastic Net)
- Logistic regression
- Linear discriminant analysis
- K nearest neighbors
- Random Forest
- Support vector machine
- Neural networks