Predicting Mental Health Using Music

Andrew Shao and Joy Qiu

2025-05-14

Introduction

Mental health has become one of the most pressing challenges in public health, especially among young adults and adolescents in an increasingly complex social and digital world. The burden of mental disorders such as anxiety, depression, insomnia, and obsessive-compulsive disorder (OCD) continues to increase, prompting urgent calls for innovative, accessible approaches to both prevention and early detection. While traditional tools, such as psychological screening, clinical interviews, and therapy, remain central to diagnosis and care, researchers and health professionals are also exploring other indicators of mental well-being, especially those embedded in daily life.

Music is one of the most universal human experiences. It transcends language, permeates culture, and accompanies people through joy, sorrow, stress, and healing. Music is a way for many individuals to process emotions and regulate their mood. Given its deeply personal and affective role, it is reasonable to explore whether music taste can reveal information about one's mental state. Perhaps preferences in genre, listening frequency, or the emotional connection to music can provide insight into one's mental state.

In this project, we explore that idea using machine learning. Using a publicly available dataset that includes individuals' music-listening habits and self-reported mental health scores, we built and evaluated predictive models to determine whether patterns in musical behavior are meaningfully linked to psychological well-being. Rather than assuming a causal relationship, we investigate whether these patterns hold enough signal to make mental health status at least partially predictable.

We aim to assess which algorithms most effectively capture any associations between music habits and mental health and evaluate whether such data could inform future public health strategies or personalized interventions. We aim to contribute a small but meaningful step toward holistic, data-informed mental wellness research.

Related Work

The relationship between music and mental health has gained increasing attention in psychological and behavioral research, particularly regarding how musical engagement can reflect or influence emotional well-being. A prominent study area has been the association between music preferences and personality or mental health traits. For example, Rentfrow and Gosling (2003) demonstrated that individuals' music genre preferences correspond to stable personality characteristics, such as openness to experience and emotional stability. These findings suggest that the type of music someone listens to may provide insights into their internal emotional and mental state.

Additional research has explored how intense or emotionally charged music genres are used for emotion regulation. Sharman and Dingle (2015) specifically investigated the effects of extreme metal music and found that, rather than amplifying anger, listening to this genre actually helped fans process and calm their emotions. This counters the common assumption that heavy or aggressive music worsens emotional states

and instead supports the idea that such genres may serve a cathartic or regulatory role for certain listeners.

Another key contribution comes from Garrido and Schubert (2013), who examined the phenomenon of individuals listening to sad music during periods of emotional distress. Their research suggests that sad music may offer comfort and emotional resonance, providing a space for reflection rather than exacerbating depressive symptoms. This challenges traditional views that link melancholic music to worsening mood and instead frames it as a potential emotional coping tool.

Despite these advances, the field focuses mainly on correlation rather than prediction. Most studies highlight associations between musical behavior and emotional tendencies, but few have applied machine learning to predict mental health status based on those patterns. Our project aims to bridge that gap by using supervised learning models to determine whether musical engagement, particularly genre preferences and listening frequency, contains predictive value for mental health burden. We hope to gather actionable, data-driven insights.

Methods

The original dataset was downloaded from Kaggle and contains 736 observations and 33 variables. The data was collected from a survey posted online to various social media platforms and forums like Reddit and Discord. The original features are participants' anonymous responses to the 31 survey questions, which ask about individuals' music listening habits and preferences, age, and a self-reported measure of how often they experience anxiety, depression, insomnia, and obsessive-compulsive disorder (OCD) on a scale from 1-10. Sixteen features were responses to questions asking how frequently participants listened to different music genres (e.g., rock, pop, metal, EDM, classical). These frequency responses are categorical: with responses being one of "Never," "Rarely," "Sometimes," and "Very frequently". Additionally, there were two features that correspond to a submission timestamp used to identify individuals and a consent indicator.

The central research question of our study was whether music-listening behaviors, such as frequency, genre preference, and time spent listening, can predict self-reported mental health status. To operationalize this, we defined a continuous response variable by summing the four mental health variables from the dataset. The resulting composite mental health score ranged from 0, corresponding to no symptoms, to 40, corresponding to severe combined symptoms (See Figure 1). This variable served as the response variable for all modeling approaches.

Modeling

To address the prediction task, we experimented with a diverse set of regression machine learning algorithms:

- **Linear Regression**
  - **Multiple Linear Regression**
  - **Feature Selection:** Best subset regression, forward stepwise selection, and backward stepwise selection were evaluated to determine the optimal selection method. We also evaluated the following selection criteria: R^2, Mallow's C_p, and Bayesian Information Criterion (BIC).
  - **LASSO Regression**
- **K-Nearest Neighbors (KNN)**
- **Decision Trees**
  - **Pruning**
  - **Random Forests**
  - **Boosting:** We used gradient boosting machines (GBM).

Data and Experiment Setup

Preprocessing

Before applying any algorithms, we cleaned and preprocessed the dataset to ensure consistency and model-readiness. All rows with missing values were removed, reducing the dataset from 736 to 622 complete observations. We treated the genre frequency features as numerical rather than one-hot encoded categories by recoding them as integers from 0 to 3, with zero corresponding to an original response of "Never" and three corresponding to a response of "Very frequently". We also excluded non-numeric or multiple-category string features for simplicity and interpretability purposes. We recoded and dropped these features for simplicity and interpretability, as including them in the analysis would've added more than a hundred dummy variables. This left us with a final feature set of 25 variables.

Model Training

Because the original dataset does not include a predefined test set, we adopted a strategy using nested 10-fold cross-validation (CV) to select parameters as well as evaluate performance robustly. Each model was trained and tested on different data partitions using 10-fold CV. When appropriate, the training folds were partitioned again using 10-fold CV to select optimal hyperparameters. This approach helps prevent overfitting and ensures information leakage does not bias the testing performance results or affect hyperparameter selection.

Model Evaluation

The models were evaluated using 10-fold CV with Mean Squared Error (MSE) as the performance metric for comparison.

Results

Algorithm	Average Training MSE	Average Testing MSE	Lowest Testing MSE
Linear regression	61.602	2.7829354×10^{10}	51.135
LASSO	62.915	66.147	49.565
KNN regression	54.191	72.335	56.123
Pruning	55.796	68.954	54.187
Random forests	12.634	64.538	51.103
Boosting (GBM)	60.841	64.625	51.577

Table 1: Model/Algorithm performance results

Baseline Model: Multiple Linear Regression

Our baseline model using multiple linear regression (MLR) yielded limited predictive power. Across all folds of the nested cross-validation, the average training mean squared error (MSE) was 61.602. The average test MSE ballooned into billions due to the influence of an extreme outlier in the dataset. In all runs, adjusted R^2 values were below 0.1, indicating that the linear model explained less than 10% of the variance in the composite mental health score.

All models had statistically significant F-statistics at the $\alpha = 0.05$ level. Among all predictors, age emerged as the only consistently significant variable, with a negative coefficient, suggesting a weak inverse relationship between age and mental health burden. Younger participants reported higher symptom scores on average. However, the overall model performed poorly, revealing a lack of fit and linear structure in the data.

Linear Regression with Feature Selection

We noticed that the choice of feature selection algorithm (i.e., best subset, forwards and backwards stepwise) had almost no effect on the number of predictors chosen (**Table 2**).

Algorithm	Average Training MSE				Average Testing MSE			
	R^2	Adj R^2	C_p	BIC	R^2	Adj R^2	C_p	BIC
Best subset	65.524	65.524	62.574	63.636	68.543	68.543	71.489	70.42
Forwards	65.524	65.524	62.578	63.637	68.543	68.543	71.444	70.454
Backwards	65.524	65.524	62.574	63.637	68.543	68.543	71.4897	70.454

Table 2: Training and Testing MSE across Feature Selection Methods

Age was selected in every model. In addition, variables like hours of music listening per day and frequency of listening to rock, metal, and EDM were frequently chosen and showed positive coefficients. Notably, Mallow's Cp tended to select larger models with 6–8 predictors, while BIC and adjusted R^2 favored more parsimonious models with 1–3 predictors.
Despite slightly improved interpretability, these feature selection methods did not significantly enhance model performance. Testing MSEs remained high and R^2 values remained low.

LASSO Regression

LASSO regression offered slight improvements in performance and model simplicity compared to simple MLR. It consistently selected between 7 and 13 predictors, retaining core variables such as age, hours of music listening per day, and frequencies of listening to EDM, folk, metal, and rock. The average training MSE was 62.915, and the average testing MSE dropped to 66.147.

The sparsity introduced by L1 regularization reduced model complexity while retaining the most predictive features.

K-Nearest Neighbors (KNN) Regression

The average training MSE was 54.191, and the average testing MSE was 72.335. Notably, the best-performing KNN model achieved a testing MSE of 56.123.

Selected values of K varied significantly across folds, ranging from 2 to 57, indicating sensitivity to partitioning and potential instability in nearest-neighbor structure.

Decision Trees (Pruning)

Using cost-complexity pruning and cross-validation to determine optimal tree size, we found that the tree sizes produced varied between 2 to 6 leaf nodes (**Figure 2** and **Figure 3**). Interestingly, the highest performing tree had only two leaf nodes splitting on age.

The average training MSE was 55.796, while the average testing MSE was 68.954. The best model achieved a testing MSE of 54.187. These trees were often structured around splits on age and frequency of rock, metal, or EDM listening.

Random Forests

Random forests achieved one of the lowest testing MSEs of all models. With an average training MSE of just 12.634 (suggesting overfitting) and an average testing MSE of 64.538, the model demonstrated strong performance at the cost of transparency. The best forest model had a testing MSE of 51.103.

Feature importance rankings, confirmed prior findings: age, hours of listening, and frequency of metal and EDM listening were among the most impactful variables (**Figure 4**).

Boosting

Boosting models had an average training MSE of 60.841 and an average testing MSE of 64.625. The lowest testing MSE observed was 51.577. While boosting performed similarly to random forests, it required significantly longer computation time. Again, age was the most influential predictor (**Figure 5**).

Discussion

This project set out to explore whether patterns in music-listening behavior could be used to predict self-reported mental health outcomes. Our investigation through multiple modeling techniques revealed that while certain features, such as age, hours of music listening per day, and frequency of listening to specific genres, showed consistent associations with mental health scores, the overall predictive performance across models was very poor.

Testing prediction performance was very similar and quite lackluster across the board (**Table 1**). Among all algorithms tested, pruned decision trees would be our recommended algorithm, as it offered the best balance between accuracy and interpretability, achieving comparable testing MSE while being the easiest to interpret. Random forests and boosting yielded slightly better raw performance but at the cost of transparency and computational efficiency. Simple MLR was the worst as it's highly sensitive to outliers.

Age was the most consistently predictive feature across all models, with younger participants tending to report higher mental health symptom burdens. Additionally, frequent listening to genres such as rock, metal, and EDM was positively correlated with higher composite mental health scores. This aligns with prior literature suggesting that these genres are often chosen for emotional intensity or catharsis, and may be preferred by individuals experiencing heightened emotional states. However, causality cannot be inferred from these results.

Many possible explanations exist for the overall poor performance, including the methods utilized in our approach. Our choice to combine the scores for four different mental health disorders, while convenient, may have obfuscated important differences between the disorders. Additionally, our choice to encode frequency ratings numerically is another potential limitation. While the frequency responses are indeed ordered and the encoding allows for smoother integration into certain models (especially the linear regression-based ones), it assumes equal intervals between categories such as "Rarely" and "Sometimes," which probably don't apply to participants' responses. Additionally, we may have omitted features with significant explanatory potential by excluding categorical and text-based variables (e.g., open responses or country of residence). Finally, the models and algorithms used may not be complex enough to capture highly complex relationships between variables.

In our opinion, the more likely source for this issue lies with the dataset in that it contains mostly noise with little or no relationship between the predictors and the response. We believe this to be the case due to the results we observed. Throughout the entire analysis, we noticed that performance did not significantly improve with increased model complexity. Linear regression with best subset selection using R^2 as the criterion produced models with a single predictor (age) and was only marginally outperformed by much more complex algorithms like Random Forest. Another example is with the decision trees produced by pruning; the highest performing tree was one with the smallest possible tree size, with only two leaf nodes (**Figure 2**). A likely explanation for this, if it's indeed true, is that the data is biased due to how it was collected. Since the survey was posted on specific online sites, there is likely to be selection bias and/or information bias. People responding to a random online anonymous survey may be more likely to answer untruthfully or haphazardly. Such bias, if present, could affect the ability to observe actual relationships within the data.

There are thus many approaches to future work. One would be to focus on the relationship between music and specific disorders like depression or OCD individually, instead of aggregating them. Second, consider experimenting with even more complex models/algorithms like deep learning. However, we think it's unlikely

these changes would significantly improve results. The better approach, we believe, would be to change the data being analyzed by either switching to ur supplementing with other data. If not, at the very least, experimenting with the original features, encoding, or other feature engineering strategies is necessary.

In conclusion, our results suggest that while music-listening habits hold some predictive value for mental health, they are insufficient in isolation for accurate prediction. However, they may still serve as a useful component in broader, multimodal assessments that combine behavioral, psychological, and contextual data. More than anything, this study affirms the complexity of how people use and interact with music.

Contributions

This project was a collaborative effort between both group members, with responsibilities divided based on individual strengths and interests.

Andrew Shao led the technical implementation of the machine learning models. He was primarily responsible for data preprocessing, coding the feature selection routines, running nested cross-validation, and training all models including linear regression, LASSO, KNN, decision trees, random forests, and boosting. Andrew also managed the statistical evaluation of model performance and was the point person for debugging and optimizing the R code.

Joy Qiu contributed to the initial exploratory data analysis, created several of the data visualizations, and helped frame the research question and hypotheses. Joy also took the lead on crafting the project presentation, writing the report sections, and interpreting the results in relation to public health relevance and prior literature. Additionally, Joy reviewed the outputs of the models, selected figures to be included in the final documentation, and coordinated the structure and flow of the written report.

References

Garrido, S., & Schubert, E. (2013). Adaptive and maladaptive attraction to negative emotions in music. *Musicae Scientiae*, *17*(2), 147–166. https://doi.org/10.1177/1029864913478305

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, *84*(6), 1236–1256. https://doi.org/10.1037/0022-3514.84.6.1236

Sharman, L., & Dingle, G. A. (2015). Extreme metal music and anger processing. *Frontiers in Human Neuroscience*, *9*, 272. https://doi.org/10.3389/fnhum.2015.00272

Appendix

Figure 1. Mental health score (response variable) distribution

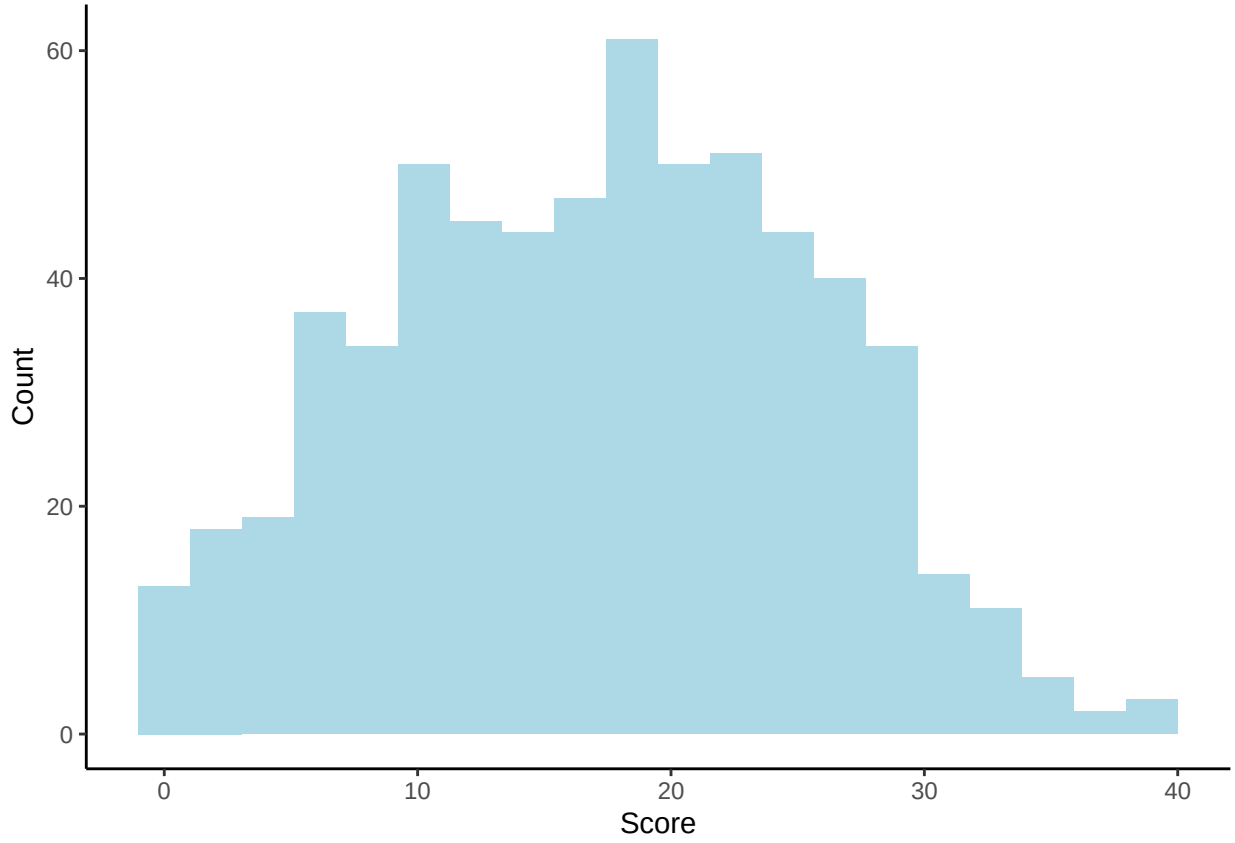

Figure 2. Best performing pruned tree

Figure 3. Most complex pruned tree

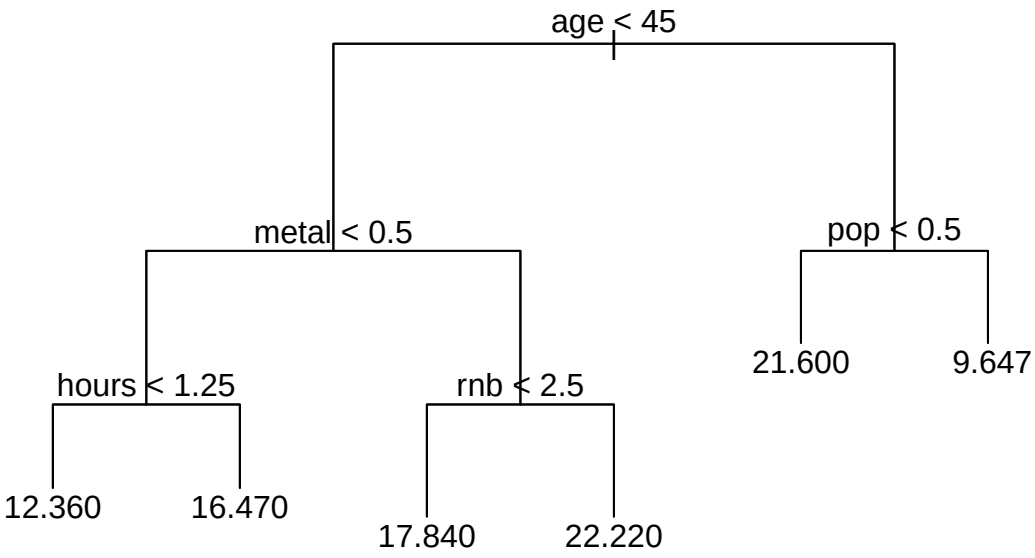

Figure 4. Random forests variable importance plots

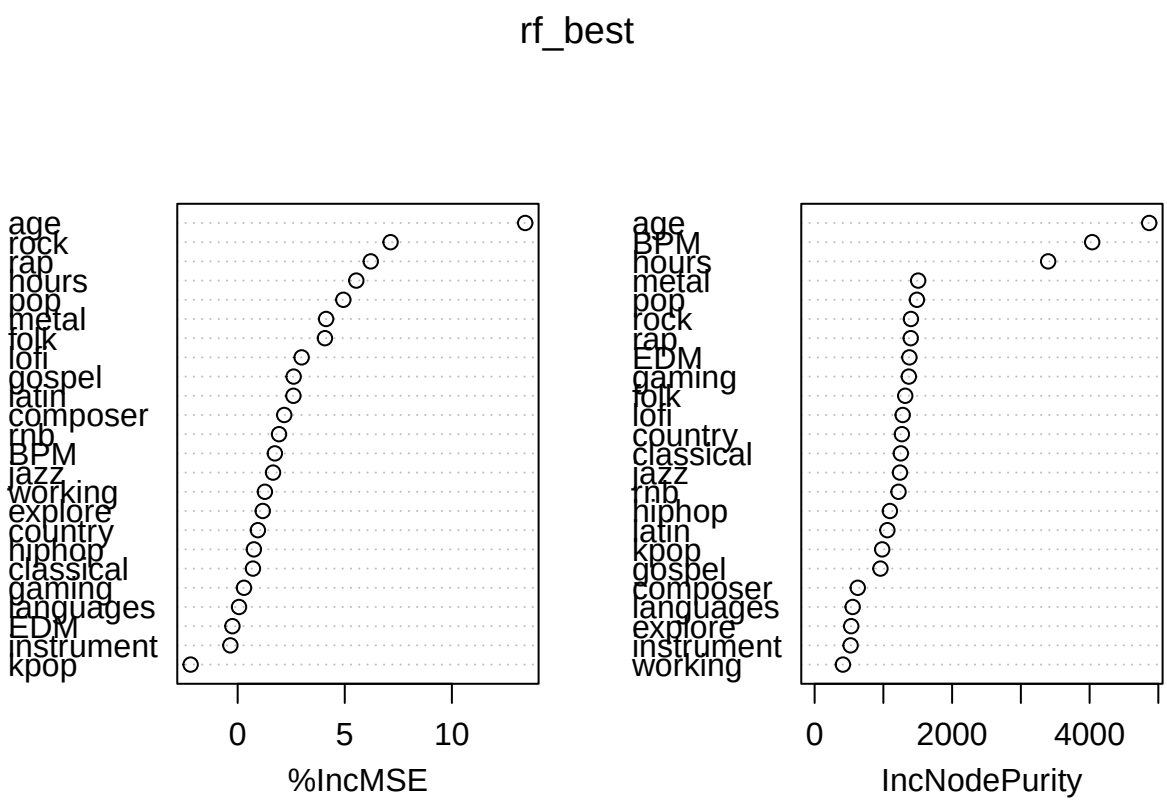

Figure 5. Boosting influence plot

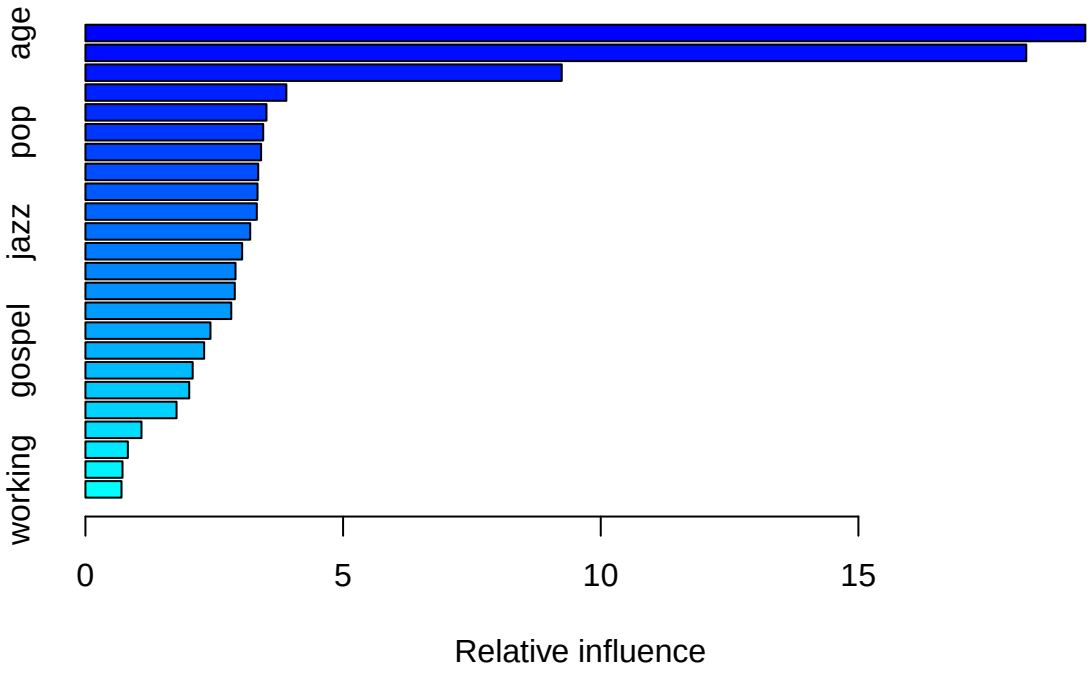

Figure 6. Plot of music genre vs. average depression score

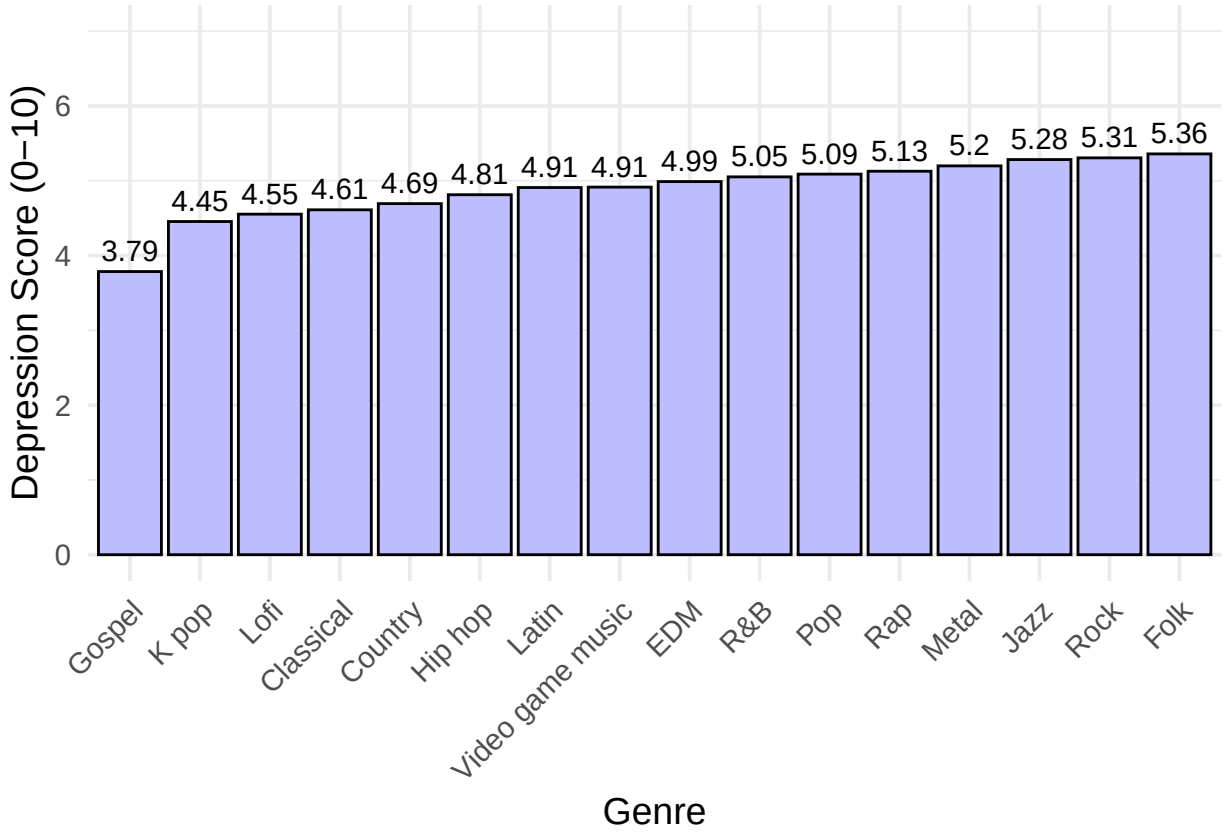

Figure 7. Plot of age group vs. average depression score

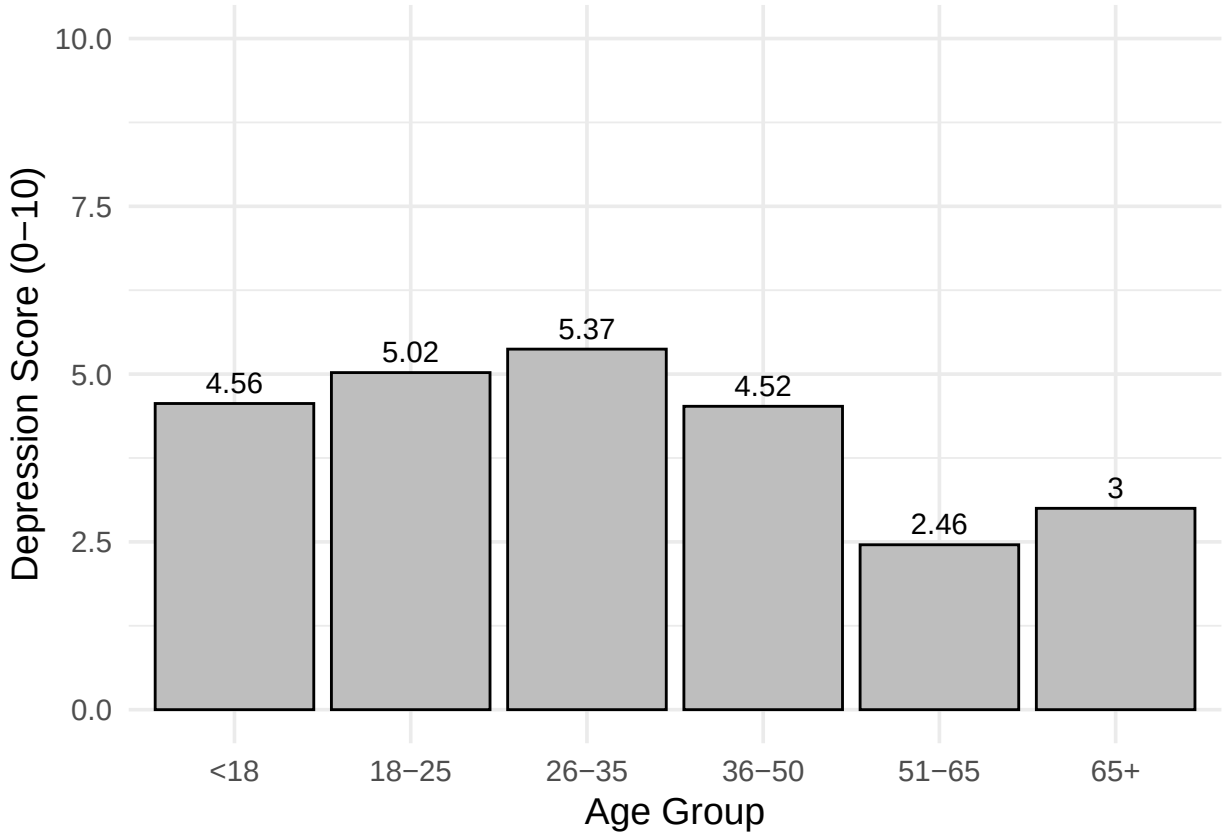

Figure 8. Plot of age groups and their top genres