UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

Music and Mental Health

Group 3 Block 3 Epic
Ellice Nelson - 23371323
Alisia Kazimierek - 23387076
Nanda - 23070854

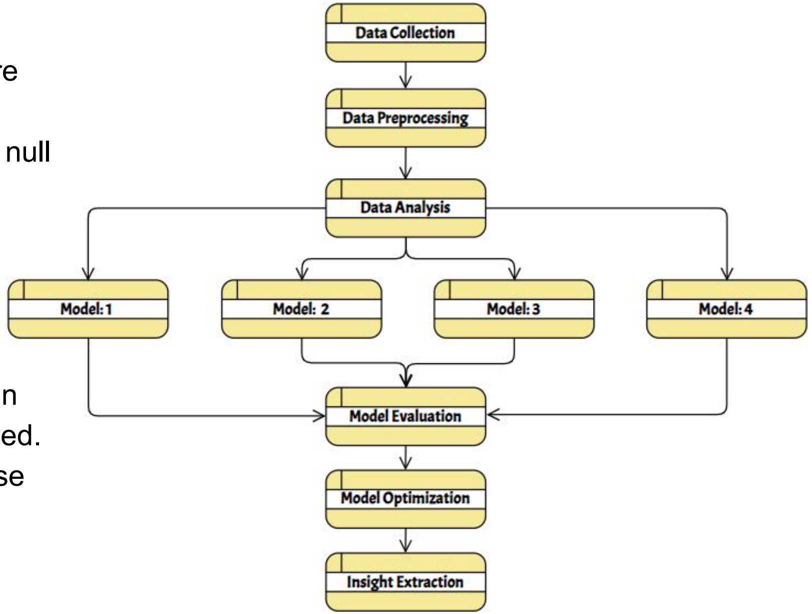
Our Research Question

Can we predict an individual's **mental health range** based on **music preferences**? If yes, which genres have the greatest effect on mental health?

AMERICAN MUSIC THERAPY ASSOCIATION®

Data Science Workflow

Data Collection - Finding a suitable dataset in the domain we are looking for.
Data Preprocessing - Remove irrelevant columns, outliers, and null values.
Data Analysis - Visually explore our data and understand the importance/impact of outliers
Models - Implement possibly suitable predictive models for our dataset to answer research questions.
Model Evaluation - See how effective/accurate our models are in answering the research questions and how they could be improved.
Model Optimization - Employ optimization techniques to increase the accuracy of our models.
Insight Extraction - Format our findings in a comprehensible, concise manner.



Dataset source

Our chosen dataset is titled **‘Music & Mental Health Survey Results’**, which we will refer to as **MXMH**. It originally contained 33 features and 736 samples. This level of samples and features were necessary to bring about conclusive insights. It also had many null values sporadically in each column.

Data Preprocessing

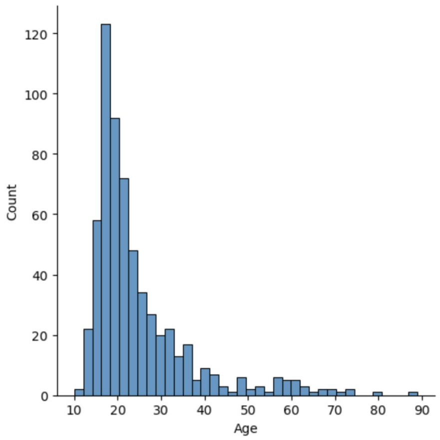
1. **Remove null values**
   - Dropna() function

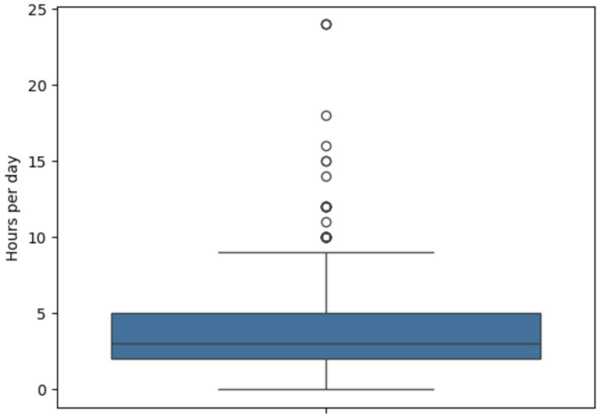
2. **Removed outliers:**
   - Discovered during data analysis
   - Avoid biased results
   - Avoid impacting our statistical analyses
   - Reduce inaccurate predictions.

3. **Drop Irrelevant columns:**
   - ‘Permissions’, ‘Timestamp’, ‘While working’, ‘Instrumentalist’, ‘Composer’, ‘Exploratory’.
   - Drop(axis=1) function.

Locating Outliers

Method 1: We used visual aids like boxplots and histograms as they allow for easy identification of data outside the typical range of values.





Locating Outliers

Method 2: Z-scores
Z-scores standardise data and finds values that are far from the average . We used a z-score threshold of 3 as it identified data points that fall outside 3 standard deviations.

```
from scipy.stats import zscore

z_scores = zscore(dataset['BPM'])
threshold = 3
outlier_indices = (z_scores > threshold) | (z_scores < -threshold)
outlier_values = dataset['BPM'][outlier_indices]
print(outlier_values)
# using 3 because 99.7% of the data falls within 3 standard deviations of the mean
```

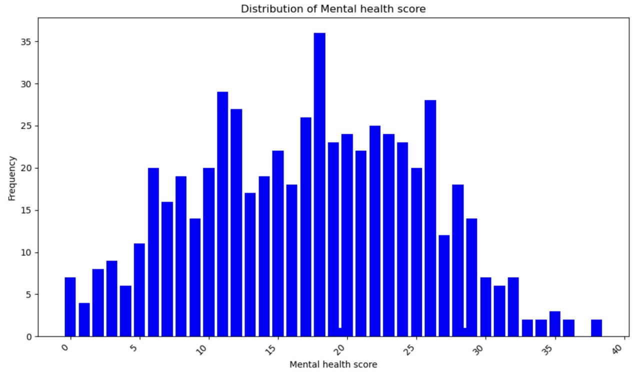
```
568    999999999.0
```

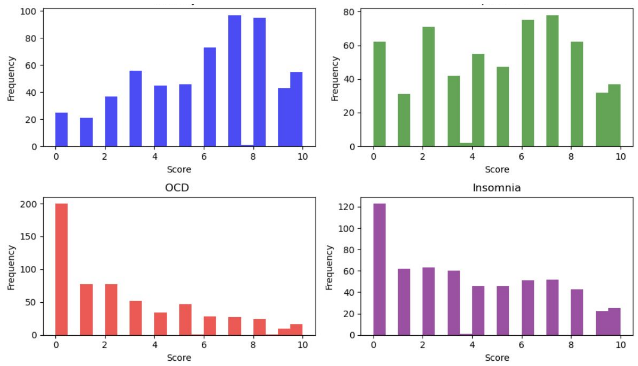
Data Exploration/ Descriptive Analysis

Mental health

- Most scores are within the range of 5 to 30.
- Majority report moderate levels of mental health scores.
- Few individuals report extreme score
- Mean score: 17.212

Anxiety: Left-skewed indicates elevated anxiety levels
Depression: Multiple peaks, generally high levels of depressive symptoms
OCD: right-skewed distribution, most report 0 as their level of OCD
Insomnia: interesting pattern. 50 respondents for each level between 1 and 9 but 125 reported a score of 0 while 25 reported a score of 10. Suggests either severe sleep disturbance or none at all.

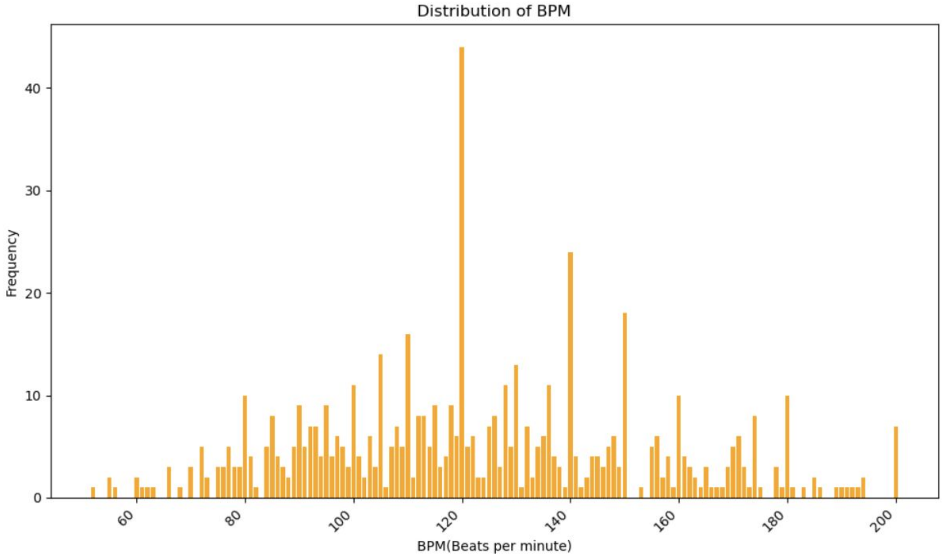




Data Exploration/ Descriptive Analysis

BPM

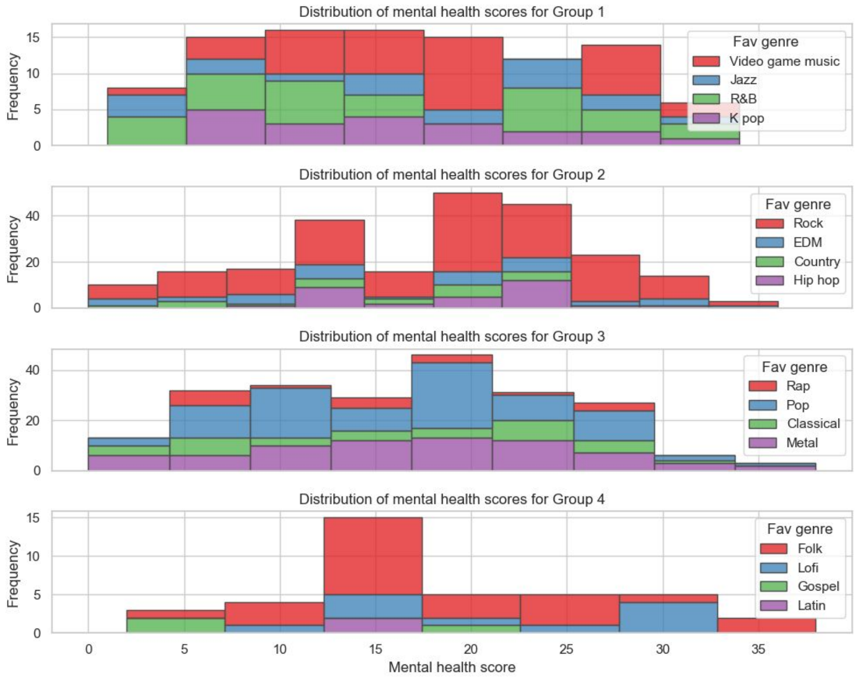
- Overwhelming amount favour music with 120 bpm
- Various peaks, indicating a range of preferred BPMs
- Diverse musical tastes evident



Data Exploration/ Descriptive Analysis

<u>Impact of Music preferences on Mental Health</u>

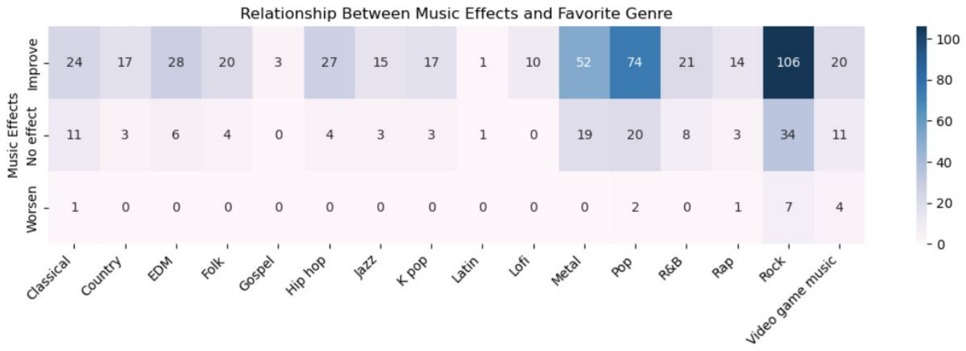
- Created a mental health score(max score:40) by adding an individual's levels of anxiety, OCD, depression and insomnia.
- Lowest mean mental health score observed was 'Gospel' genre with 10.67.
- Highest mean mental health score observed was 'Lo Fi' genre with 21.70



Data Exploration/ Descriptive Analysis

Impact of Music preferences on Mental Health

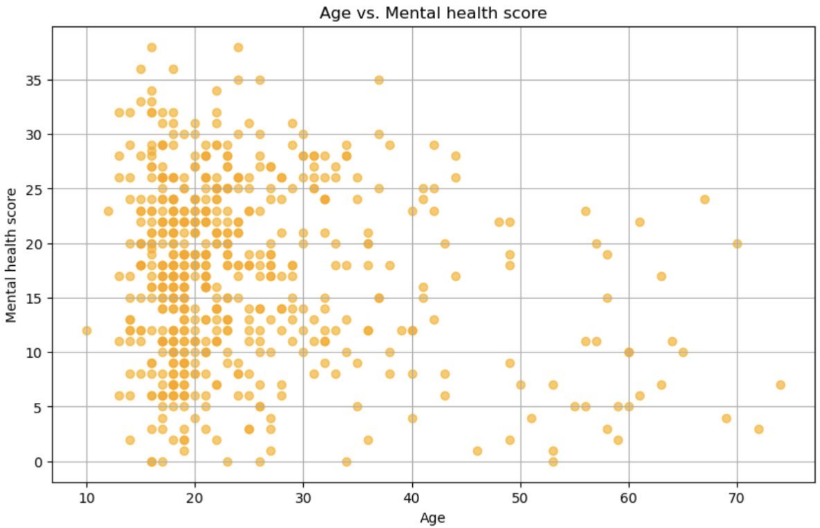
- Heatmap enables us to see how individuals perceive the impact of music on their mood
- Individuals who believed music improved their mood often favour genre like Rock, Pop and Metal but weren't limit to those genres.
- Suggests energetic and uplifting music could positively influence mental health.
- Small portions of these genres believe music has no effect



Data Exploration/ Descriptive Analysis

Age and Mental Health Analysis

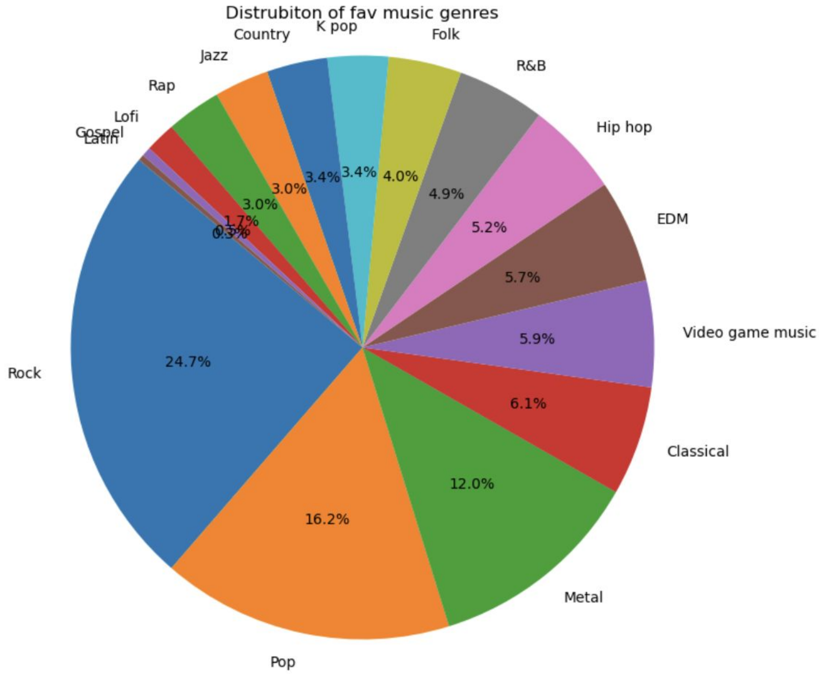
- Younger individuals tend to exhibit higher mental health scores
- Below the age of 30, 6 individuals have a mental health score of 35 or above.
- While the highest mental health score among individuals above 50 is 24
- Suggests a potential correlation between age anc



Data Exploration/ Descriptive Analysis

Music genres

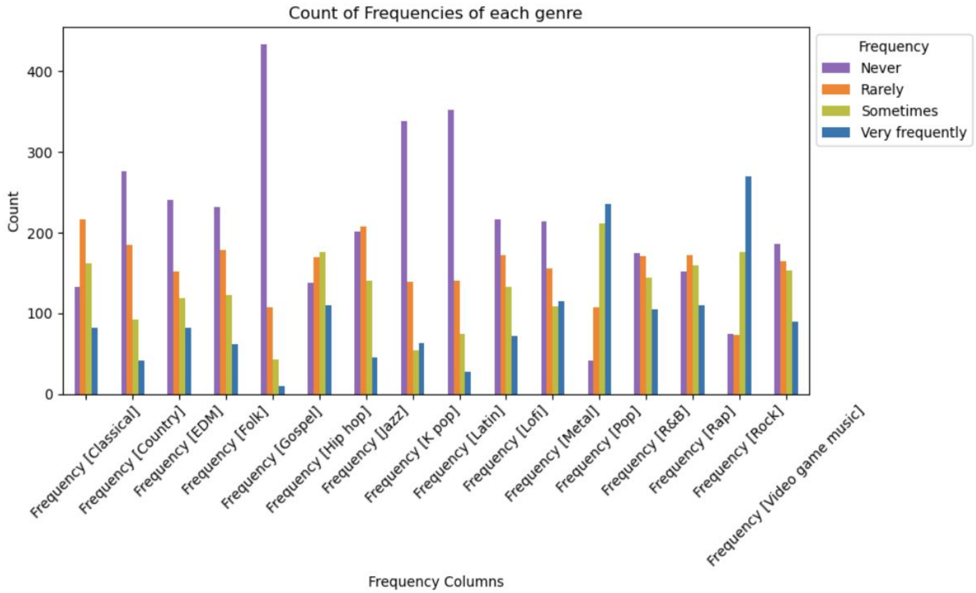
- Most favoured is Rock with 24.7%
- Genre with best mental health, Gospel, only favoured by 0.5% of respondents.
- Genre with worst mental health, Lofi , favoure by 1.7% of respondents.
- Variety in favourite music indicates a diverse mus tastes among our respondents.



Data Exploration/ Descriptive Analysis

Frequency of listening to Music Genres

- Metal and Rap are very frequently listened to by respondents.
- Genres like Folk and K-pop have the highest statistics for never being listened to.
- Analysing this frequency gives us more knowledge into the relationship between music genres and the effect it has on mental health .



OLTP and OLAP Questions

OLTP Q1

What favourite genre group experiences the most anxiety?

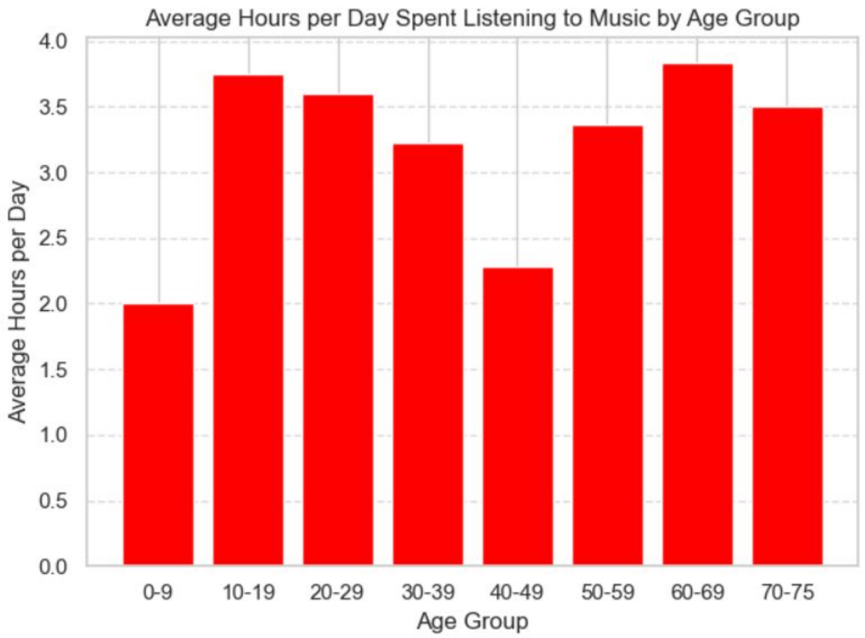
- K-Pop 6.65 avg

Fav genre	Anxiety	Depression	Insomnia	OCD
Classical	5.055556	4.527778	3.916667	2.444444
Country	5.700000	4.100000	2.600000	2.600000
EDM	5.294118	4.970588	4.000000	2.676471
Folk	6.583333	5.416667	4.125000	2.333333
Gospel	4.000000	1.333333	4.666667	0.666667
Hip hop	6.193548	6.064516	3.709677	2.741935
Jazz	5.833333	4.722222	4.000000	2.222222
K pop	6.650000	4.100000	3.350000	2.600000
Latin	5.000000	4.500000	4.500000	2.500000
Lofi	6.100000	6.600000	5.600000	3.400000
Metal	5.436620	5.098592	4.380282	2.225352
Pop	6.067708	4.390625	3.260417	2.963542
R&B	5.310345	4.172414	2.931034	2.724138
Rap	5.166667	3.888889	2.333333	3.333333
Rock	6.163265	5.425170	3.921769	2.670068
Video game music	6.457143	4.600000	4.542857	2.628571

OLTP Q2

What age group likes to listen to music the most?

- 60-69 age group



OLTP Q3

How many 18 year olds experience some form of mental illness?

- All 67 of the 18 year olds experience mental health illnesses

```

cleanedDataset.loc[(cleanedDataset['Age'] == 18) & (cleanedDataset['Mental health score'] == 0)]

```

Age	Primary streaming service	Hours per day	Fav genre	Foreign languages	BPM	Frequency [Classical]	Frequency [Country]	Frequency [EDM]	Frequency [Folk]	...	Frequency [R&B]	Frequency [Rap]	Freq [

0 rows × 28 columns

#E ✖ OLTP ✖

67 18 year olds experience some form of mental health challenge

OLTP Q4

How many people‘s favourite genre is rock?

- 147

Fav genre	
Rock	147
Pop	96
Metal	71
Classical	36
Video game music	35
EDM	34
Hip hop	31
R&B	29
Folk	24
K pop	20
Country	20
Jazz	18
Rap	18
Lofi	10
Gospel	3
Latin	2

OLTP Q5

How many people in the largest favourite genre group experience some form of mental illness?

- 146/147 Rock listeners

```
1 cleanedDataset.loc[(cleanedDataset['Fav genre'] == 'Rock') & (cleanedDataset['Mental health score'] != 0)]
```

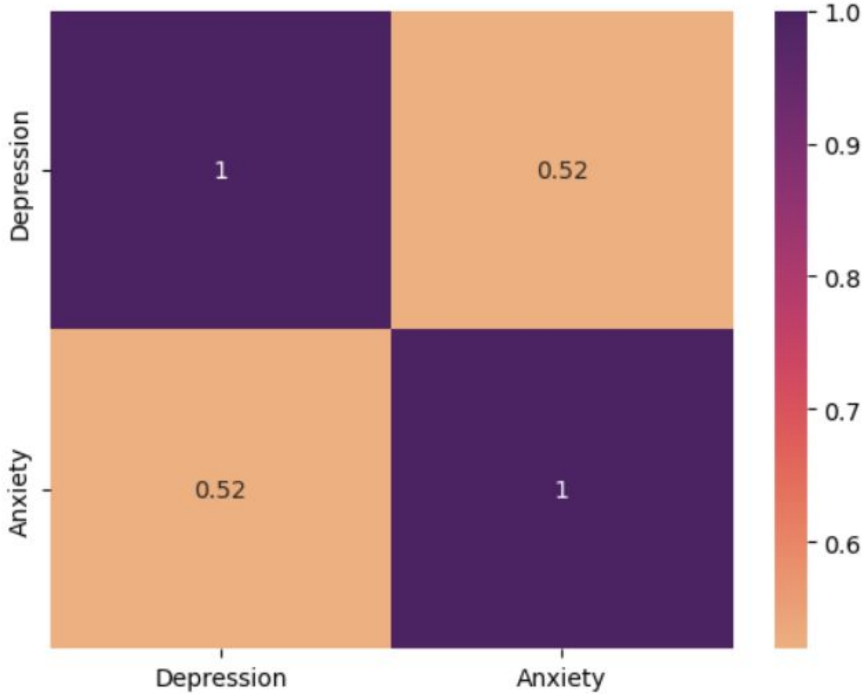
	Age	Primary streaming service	Hours per day	Fav genre	Foreign languages	BPM	Frequency [Classical]	Frequency [Country]	Frequency [EDM]	Frequency [Folk]	...	Frequency [R&B]	Frequency [Rap]	Frequenc [Rock
8	19.0	Spotify	6.0	Rock	No	94.0	Never	Very frequently	Never	Sometimes	...	Never	Never	Ver frequentl
24	16.0	Other streaming service	3.0	Rock	Yes	84.0	Rarely	Rarely	Never	Rarely	...	Sometimes	Rarely	Ver frequentl
26	14.0	Spotify	12.0	Rock	Yes	136.0	Sometimes	Sometimes	Rarely	Rarely	...	Very frequently	Very frequently	Ver frequentl
33	17.0	Spotify	4.0	Rock	Yes	142.0	Rarely	Rarely	Rarely	Very frequently	...	Rarely	Sometimes	Ver frequentl
38	26.0	Other streaming service	0.5	Rock	Yes	140.0	Rarely	Never	Rarely	Sometimes	...	Never	Never	Ver frequentl
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
674	17.0	Spotify	5.0	Rock	No	110.0	Very frequently	Rarely	Rarely	Sometimes	...	Very frequently	Very frequently	Ver frequentl
697	16.0	Spotify	3.0	Rock	Yes	90.0	Rarely	Never	Never	Rarely	...	Sometimes	Sometimes	Ver frequentl
701	30.0	YouTube Music	1.0	Rock	Yes	115.0	Sometimes	Rarely	Rarely	Rarely	...	Rarely	Rarely	Ver frequentl
710	16.0	Spotify	8.0	Rock	No	120.0	Very frequently	Never	Rarely	Sometimes	...	Rarely	Never	Ver frequentl
731	17.0	Spotify	2.0	Rock	Yes	120.0	Very frequently	Rarely	Never	Sometimes	...	Never	Rarely	Ver frequentl

146 rows × 28 columns

OLAP Q1

What is the correlation between depression and anxiety?

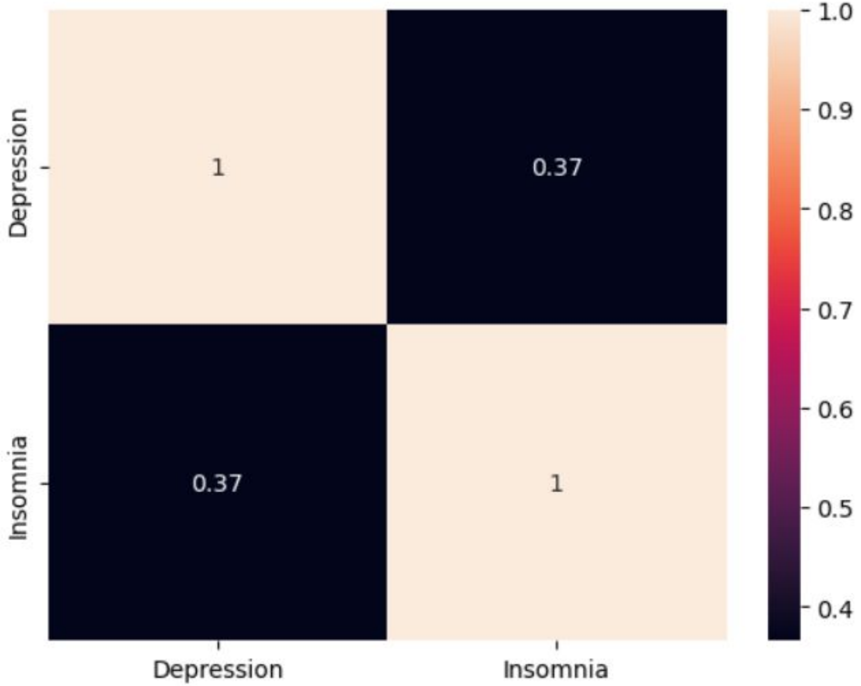
- 0.52



OLAP Q2

What is the correlation between depression and insomnia?

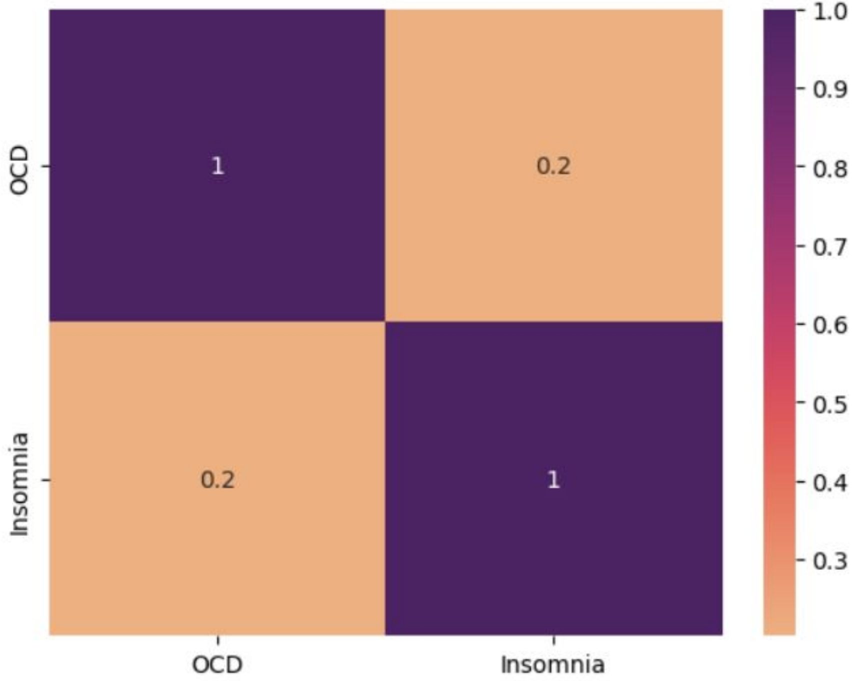
- 0.37



OLAP Q3

What is the correlation between OCD and insomnia?

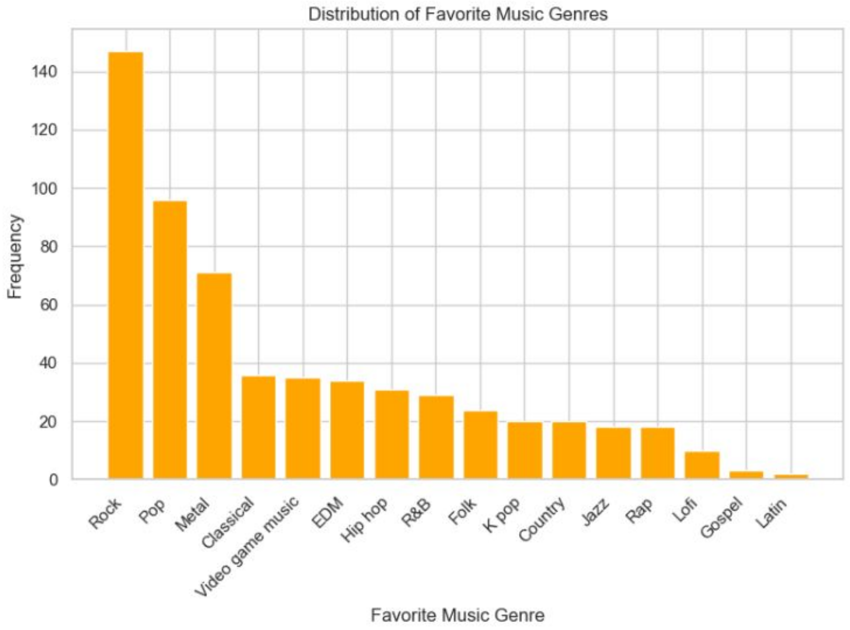
- 0.2



OLAP Q4

What are the top 3 genres in the dataset?

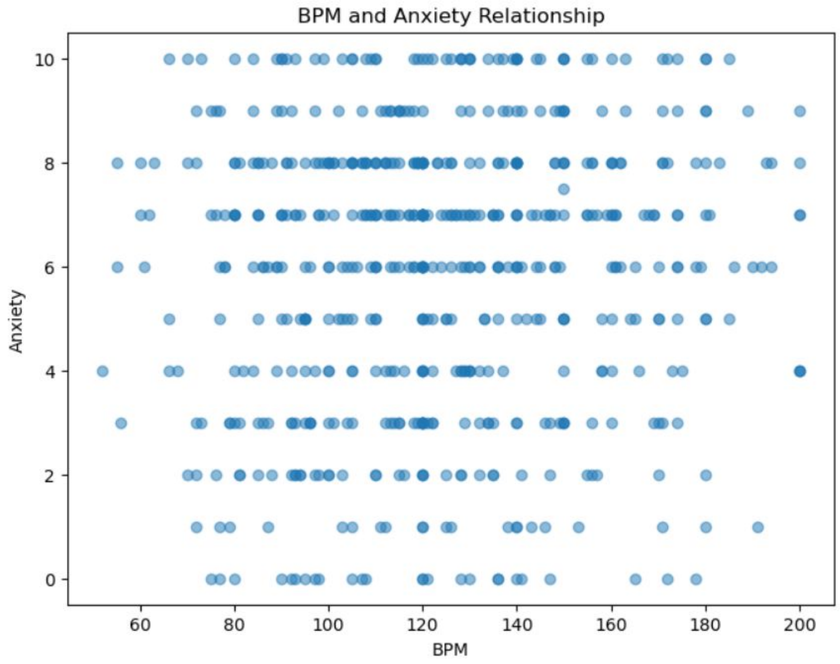
- Rock
- Pop
- Metal



OLAP Q5

What is the relationship between BPM and anxiety?

- Undefined



OLAP Q6

How does music preferences (genre of music) relate to mental health outcomes?

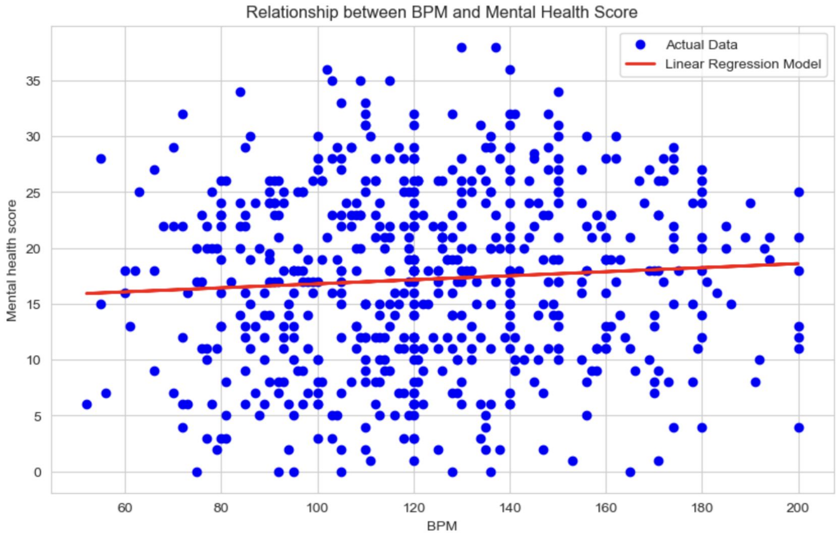
- Certain genres have higher mental health scores

Average Mental Health Score for each Favorite Music Genre Combination:

	Music Genre Combination	Favorite Genre	Mental health score
0	0_0_0_0_0_0_0_0_0_0_0_0_0_0_0_1	Gospel	10.666667
1	0_0_0_0_0_0_0_0_0_0_0_0_0_0_1_0	Classical	15.944444
2	0_0_0_0_0_0_0_0_0_0_0_0_0_1_0_0	Rock	18.180272
3	0_0_0_0_0_0_0_0_0_0_0_0_1_0_0_0	Rap	14.722222
4	0_0_0_0_0_0_0_0_0_0_0_1_0_0_0_0	Lofi	21.700000
5	0_0_0_0_0_0_0_0_0_0_1_0_0_0_0_0	Latin	16.500000
6	0_0_0_0_0_0_0_0_0_1_0_0_0_0_0_0	Jazz	16.777778
7	0_0_0_0_0_0_0_0_1_0_0_0_0_0_0_0	Folk	18.458333
8	0_0_0_0_0_0_0_1_0_0_0_0_0_0_0_0	Hip hop	18.709677
9	0_0_0_0_0_0_1_0_0_0_0_0_0_0_0_0	K pop	16.700000
10	0_0_0_0_0_1_0_0_0_0_0_0_0_0_0_0	R&B	15.137931
11	0_0_0_0_1_0_0_0_0_0_0_0_0_0_0_0	EDM	16.941176
12	0_0_0_1_0_0_0_0_0_0_0_0_0_0_0_0	Pop	16.682292
13	0_0_1_0_0_0_0_0_0_0_0_0_0_0_0_0	Metal	17.140845
14	0_1_0_0_0_0_0_0_0_0_0_0_0_0_0_0	Video game music	18.228571
15	1_0_0_0_0_0_0_0_0_0_0_0_0_0_0_0	Country	15.000000

OLAP Q7

How does the BPM of music relate to the mental health of individuals? Are there specific ranges of BPM associated with lower (or higher) levels of mental health issues?



- No observable relationship

Predictive Models Applied

Models we chose

- Multiple Linear Regression
- Random Forest Regression
- Naive Bayes Classification
- Multi-Layer Perceptron Classification

Multiple Linear Regression

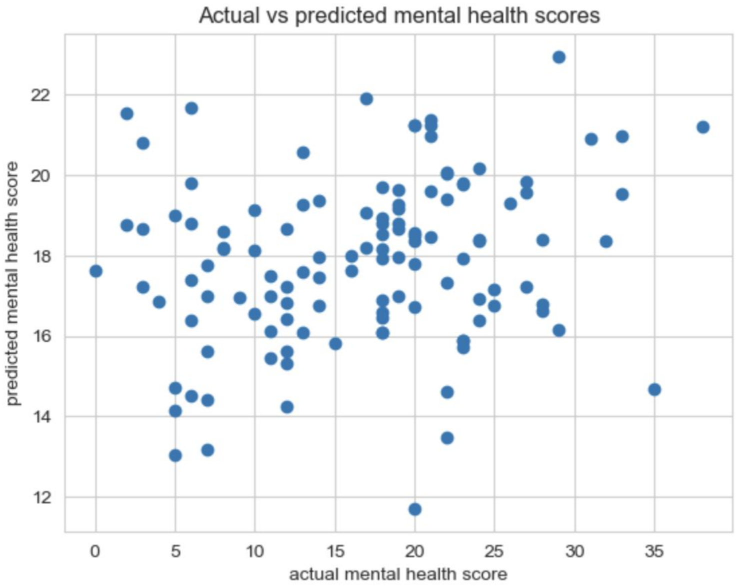
- Multiple linear regression is a **supervised learning algorithm** and a **statistical method** used for regression tasks.It is used to model the relationship between a **dependent variable** (the outcome we want to predict) and **one or more independent variables** (the predictors).

- **Input features:** 21 columns
- **Target:** Mental health score column
- **Initial train/test split:** 80% for training, 20% for testing

Pre-Optimization Results of Multiple Linear Regression

Performed poorly on the evaluation metrics we used.

```
Mean Square Error: 64.03927155313792
R-squared: 0.0266002128865655
Mean Absolute Error: 6.323278465656104
Root Mean Square Error: 8.002454095659527
```



Optimization of Multiple Linear

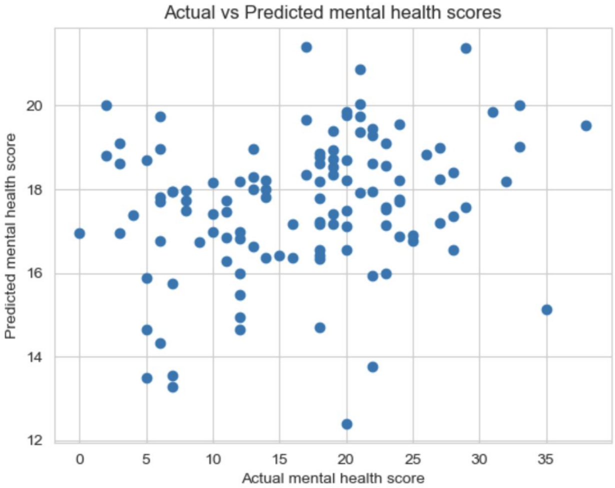
Hyperparameter tuning: Used GridSearchCV to search through the best parameter values from the given set of the grid of parameters. It uses cross-validation to find the combination that minimises the mean squared error(MSE).

Lasso Model: optimises the multiple linear regression by adding a regularisation term that penalises the absolute size of the coefficients to prevent overfitting and enhance the accuracy by selecting only the most relevant features for prediction.

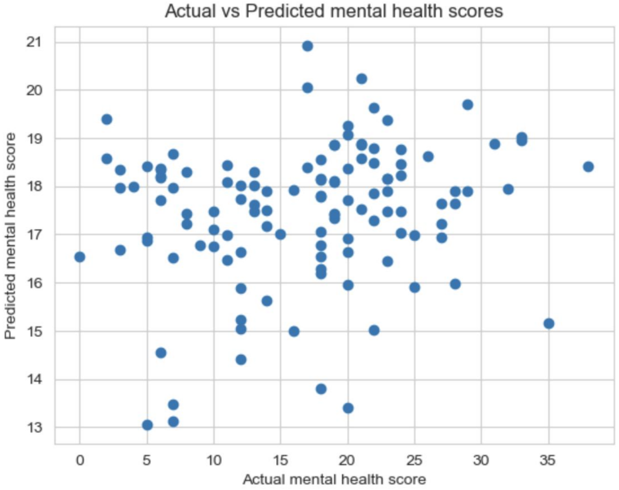
Elastic Net model: It combines the strengths of Lasso and Ridge regression techniques. It keeps the feature selection quality from the lasso penalty as well as the effectiveness of the ridge penalty. By adjusting its parameters, Elastic Net finds the right balance between these penalties , making the model more accurate and stable for predicting outcomes.

Post Optimization Results of Multiple Linear Regression

```
Lasso - Mean Squared Error: 63.91864586613501
Lasso - R-squared: 0.02843372871732064
```




```
Elastic Net - Mean Squared Error: 62.535714237376
Elastic Net - R-squared: 0.04945435122559794
```



Both lasso and Elastic Net techniques slightly improved the performance metrics compared to the original model. Elastic Net in particular indicated a better model fit.

Random Forest Regression

- A **supervised learning algorithm** used for regression tasks, where the goal is to predict a continuous output variable based on input features.
- Constructs **multiple decision trees** during training and outputs the average prediction of the individual trees.

- **Input features:** 23 columns
- **Target:** ‘Mental health score’ column.
- **Initial Train/Test split:** 60%/20%

Pre-Optimization Results of Random Forest Regression

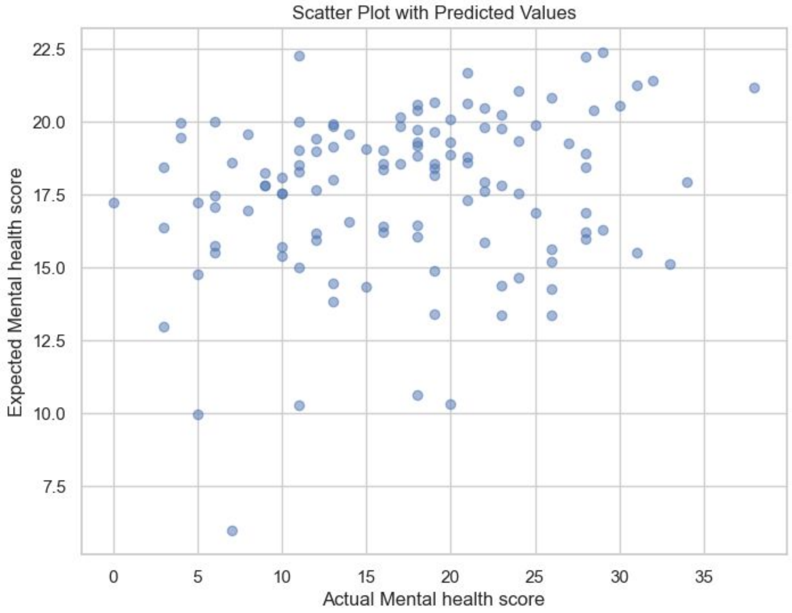
Poor results in all categories

Mean Absolute Error (MAE): 6.540

Mean Squared Error (MSE): 63.510

Root Mean Squared Error (RMSE): 7.969

R-squared (R2): 0.024



Optimization of Random Forest

1. **Hyperparameter Tuning**
   - This meant using grid search to find the optimal hyperparameters (min_sample_split, max_depth, bootstrap, etc.) for the model based on our data.
   - Apply best parameters to the model

2. **Unimportant Feature Removal**
   - Use feature importance attribute on the model to find out what features are most essential
   - Remove the non-essential features
   - Recreate model with this noise now reduced

3. **Train/Validation/Test Split**
   - Change from 60/20 train/test split to 60/20/20 train/valid/test split

Post-Optimization Results of Random Forest Regression

1. Better after hyperparameter training.
2. Better after unimportant features removal combined with hyperparameter tuning.
3. Better after train/valid/test split

Mean Absolute Error (MAE): 6.368

Mean Squared Error (MSE): 61.469

Root Mean Squared Error (RMSE): 7.840

R-squared (R2): 0.056

Naive Bayes Classification

- A **supervised learning algorithm** used for classification tasks, where the goal is to predict the class label(category) of input data based on input features.
- Applies **Bayes' Theorem** during training and selects the class with the highest probability of being the class' label.

- **Input features:** 23 features
- **Target:** 'Mental health score' category columns.
- **Train/Test/Validation split:** 70%/15%/15%

Pre-Optimization Results of Naive Bayes Classification

Accuracy: 42.25352112676056%

Classification Report for Gaussian Naive Bayes Classifier

	precision	recall	f1-score	support
Zero Category	0.00	0.00	0.00	3
Very Low	0.42	1.00	0.59	60
Low	0.00	0.00	0.00	42
Moderate	0.00	0.00	0.00	36
High	0.00	0.00	0.00	1
accuracy			0.42	142
macro avg	0.08	0.20	0.12	142
weighted avg	0.18	0.42	0.25	142

Optimization of Naive Bayes

1. **Split testing set into testing and validation set**
   - Additional step in diagnosing overfitting
   - Avoids Data Leakage from testing set, thus preventing overly optimistic performance metric

2. **Improved Data Pre-processing**
   - Used functions like SimpleImputer and StandardScaler to pre-process data more efficiently
   - Used IQR Method to remove outliers from BPM column
   - Implemented pipeline data structure to ensure consistency

Post Optimization Results of Naive Bayes Classification

```
Naive Bayes Testing Set Metrics:
Accuracy: 86.51685393258427%
```

```
Naive Bayes Validation Set Metrics:
Accuracy: 87.77777777777777%
```

Classification Report for Gaussian Naive Bayes Classifier:

	precision	recall	f1-score	support
Zero Category	1.00	1.00	1.00	1
Very low	0.89	0.80	0.84	40
Low	0.77	0.96	0.85	24
Moderate	0.95	0.87	0.91	23
High	1.00	1.00	1.00	1
accuracy			0.87	89
macro avg	0.92	0.93	0.92	89
weighted avg	0.87	0.87	0.87	89

Table comparing actual value to predicted value

Multi-Layer Perceptron Classification

- A **supervised learning algorithm** used for classification tasks, where the goal is to predict the class label(category) of input features
- Uses a **neural network** that passes input data through multiple nodes, calculates an output, and converts the output value into a probability with softmax, and selects the label with the highest probability.

- **Input features:** 23 features
- **Target:** ‘Mental health score’ category columns.
- **Train/Test/Validation split:** 70%/15%/15%

Pre-Optimization Results of Multi-Layer Perceptron Classification

```
MLP Classifier Testing Set Metrics:
Accuracy: 92.13483146067416%

MLP Classifier Validation Set Metrics:
Accuracy: 90.0%
```

Classification Report for MLP Classifier:

	precision	recall	f1-score	support
Zero Category	1.00	1.00	1.00	1
Very Low	0.93	0.93	0.93	40
Low	0.96	0.96	0.96	24
Moderate	0.88	0.91	0.89	23
High	0.00	0.00	0.00	1
accuracy			0.92	89
macro avg	0.75	0.76	0.76	89
weighted avg	0.91	0.92	0.92	89

Optimization of Multi-Layer Perceptron

1. **Hyperparameter Tuning**
- Used scikit-learn's GridSearchCV to find best parameters for the MLP Classifier model
- Improved model's performance but makes it take longer to run due to iterating through all grid elements

2. **Implemented Data pre-processing pipelines**
- Implemented pipeline data structure to ensure consistency
- Similar implementation to Naive-Bayes, using SimpleImputer and StandardScaler

Post Optimization Results of Multi-Layer Perceptron Classification

```
MLP Classifier Testing Set Metrics:
Accuracy: 95.50561797752809%
```

```
MLP Classifier Validation Set Metrics:
Accuracy: 96.66666666666667%
```

Classification Report for MLP Classifier:

	precision	recall	f1-score	support
Zero Category	1.00	1.00	1.00	1
Very Low	0.95	0.97	0.96	40
Low	1.00	0.96	0.98	24
Moderate	0.92	0.96	0.94	23
High	0.00	0.00	0.00	1
accuracy			0.96	89
macro avg	0.77	0.78	0.78	89
weighted avg	0.95	0.96	0.95	89

Conclusion

Our goal:
Predict an individual's mental health score based on their music taste preferences.

Outcome:
We **cannot** reliably **predict mental health score** from music taste in this dataset alone.

- More important factors not contained in this dataset.
- We **can predict** an individual’s **mental health category** (very low, low, moderate or high mental health score) with ~95% accuracy.