Final Report

Section 1 – Introduction

Our research is investigating the question: **"What is the correlation between an individual's musical preferences and their mental health state?".**
We're particularly interested in the potential links between music choices and conditions such as insomnia, anxiety, OCD, and depression.

In our society, where mental health issues are increasingly prevalent, understanding such correlations could allow music to be a pivotal tool in early detection or therapeutic strategies. However, this research is not straightforward due to the subjective nature of music preferences and self-reported mental health conditions.

We're taking a detailed approach in this study. Instead of broadly asking if music impacts mental health, we're focusing on specific conditions. **To do this, we're using objective Spotify data to create unique musical attribute scores for each individual.**

Furthermore, we're exploring the possibility of grouping people based on their musical profiles and comparing the mental health scores across these groups. This could highlight if certain musical attributes or genres have a stronger association with mental health states than others.

To reach our goal, we will use classification models, predicting self-reported mental health scores based on musical preferences. We aim to construct four separate models, each one dedicated to a different condition: insomnia, anxiety, OCD, and depression.

Section 2 – Data overview

Our data consists of 2 datasets from Kaggle:

- "Music & Mental Health Survey Results" – A survey of over 700 individuals asking about their music preferences, such as the frequency of hearing rock, pop, jazz etc.... as well as their self-rated mental health scores on insomnia, anxiety, OCD, and depression.

- "Spotify Tracks Dataset" – including information about 10,000 different songs in Spotify. For our purposes we used the attributes for each track that suggest how suitable a track is for dancing, how much energy it has and more to calculate each individual's score.

A total of 9 scores (S) were calculated for each individual (one for each attribute) by this formula:

$$f_i \in \left\{0, \frac{1}{3}, \frac{2}{3}, 1\right\} - \mathit{Grequncy\ genre\ value}$$
$$g_i - \mathit{Calculated\ genre\ mean\ attribute\ score}$$
$$h_i - \text{Total hours per day of listening to music}$$
$$n = 13 - \mathit{Total\ number\ of\ geners\ in\ our\ data}$$

$$S = \frac{\sum_i^n (f_i * h_i * g_i)}{\sum_i^n (f_i * h_i)}$$

The scores calculated did not have very high correlation with the mental health conditions, but some of them did improve our model (see appendix 1).

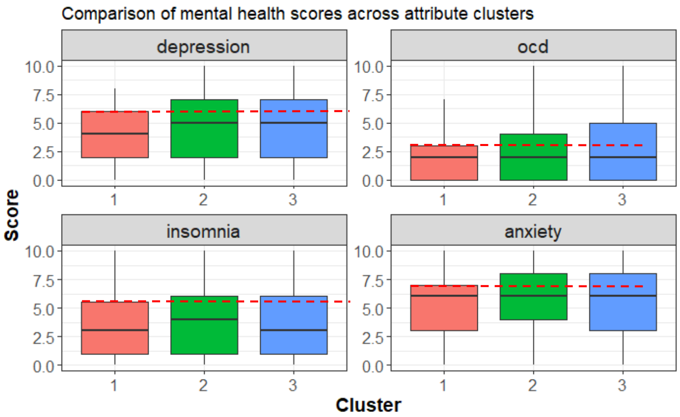
Section 3 – Methods and results

Clustering Approach

We used the k-means clustering algorithm to group the individuals in our dataset by two features sets (Set 1 – Attributes scores, Set 2 – Genre listening frequency value).
The optimal number of clusters for attribute scores was determined to be 3, as per the silhouette coefficient. For genre frequency, the data was divided into two clusters.

Clustering Results

Once the clusters were formed, we compared the distribution of all mental health conditions across these groups. Our ANOVA test results indicated a significant difference in depression scores across set 1 (see appendix 3) Specifically, cluster 1, characterized by high acousticness and low values for all other attributes (See appendix 2), exhibited slightly lower average scores across all mental health conditions.



These results suggest there may indeed be patterns in how certain types of music correlate with an individual's mental state, and shed further light on the relationship between musical preferences and mental health.

The same output for set 2 of clustering did not produce any significant results (see appendixes 4+5).

Classification Approach

We implemented four classification models, categorizing the mental health scores into low (0-3), medium (4-7), and high (8-10) groups. This division allowed for easier interpretation and analysis. The model framework was based on the random forest classification method with a selection of features guided by the Chi-square test's p-values. Also, the models were validated using 5-fold cross validation.

<u>Predicting depression:</u>

```
Prediction  1  2  3
         1 33 26 10
         2 24 34 16
         3  0  2  3
```

```
F1 Score:   0.4000934
Precision:   0.5125734
Recall:   0.4102609
Accuracy:   0.472973
```

<u>Predicting insomnia:</u>

```
           Truth
Prediction  1  2  3
         1 46 13 28
         2  1  2  0
         3 20  7 31
```

```
F1 Score:   0.4291057
Precision:   0.5766284
Recall:   0.4343
Accuracy:   0.5337838
```

<u>Predicting anxiety:</u>

```
           Truth
Prediction  1  2  3
         1 43 26 24
         2  4  7  8
         3 14  9 13
```

```
F1 Score:   0.3696458
Precision:   0.3972993
Recall:   0.3868245
Accuracy:   0.4256757
```

<u>Predicting OCD:</u>

```
           Truth
Prediction  1  2  3
         1 82 39 22
         2  2  2  1
         3  0  0  0
```

```
F1 Score:   0.4047117
Precision:   0.4867133
Recall:   0.341657
Accuracy:   0.5675676
```

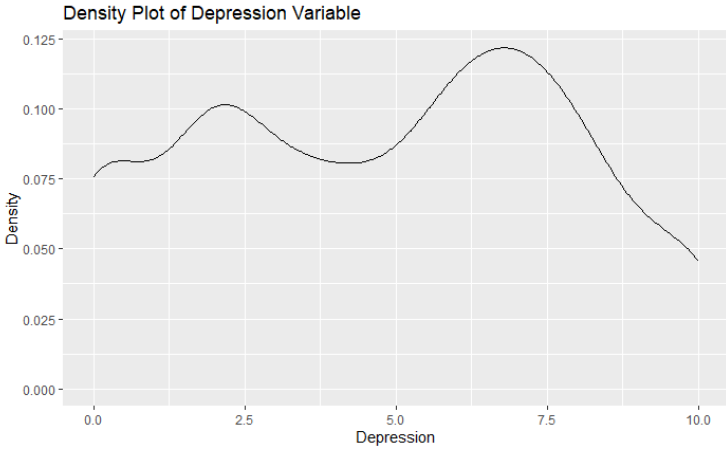
Classification results

To get a better understanding of the results, we have decided group and perform ANOVA tests across the three depression groups, only nine out of thirty-nine tests revealed statistically significant differences, specifically in the frequency of hearing metal and rock music, amount of hours listening per day (see appendix 6). **This highlighted the challenges in differentiating between the groups.**

A noteworthy discovery was the correlation of -0.38 between self-rated high OCD scores and the calculated attribute of speechness. This implies that as the number of spoken words in a track increases, it becomes less likely to be favored by individuals with OCD (See appendix 7).

Linear Regression Approach

This graph shows that the value of depression is distributed almost uniformly. There is no bell curve, and that makes the predictions harder for regression. We confirmed this observation by running a linear regression model and got r^2 of 0.036.



Total mental health score approach

We have calculated a total mental health score for each individual using the formula: $e^{depression} + e^{anxiety} + e^{insomnia} + e^{ocd}$. Classifying 1 if the resulting score was greater than $2 * e^6$ (meaning he rated 2 or more conditions with 6 or more OR one condition 7 or more). These were the results:

```
          Truth
Prediction  1  0
         1 89 39
         0  8 12
```

```
F1 Score:  0.7911111
Precision:  0.6953125
Recall:  0.9175258
Accuracy:  0.6824324
```

This model can identify individuals with a high likelihood of experiencing one or more conditions, but it does not provide specific information about the individual conditions themselves.

Section 4 – Limitations and Future Work

Our project faced several challenges. One of the main limitations is the data itself, which relied on individuals' self-reported mental health state. Each participant may have their unique interpretation and perception of their mental health condition, leading to potentially high variance among responses. Moreover, since we did not have a way to verify these self-reported conditions with a medical professional, we cannot ascertain the accuracy of the mental health states claimed by the participants.

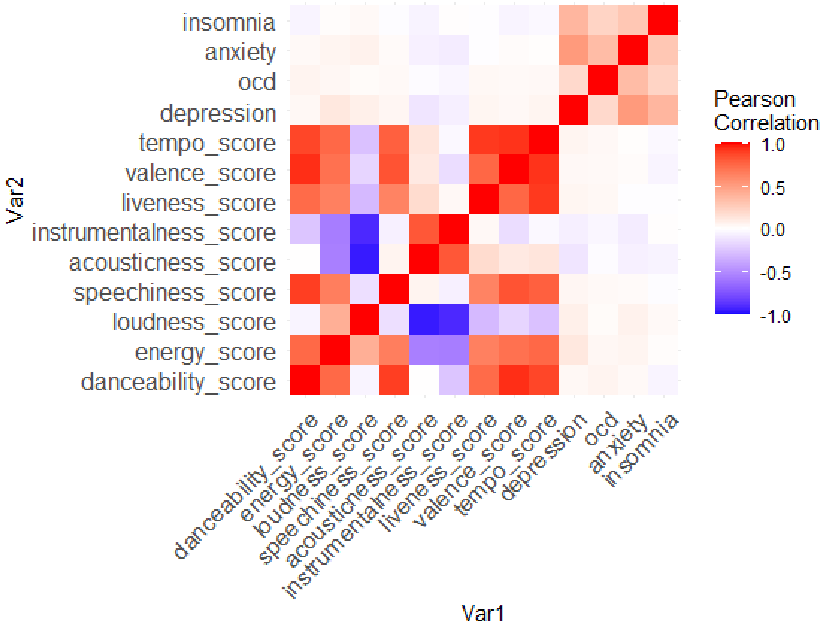
Another issue we encountered is the high invariability of music preferences among individuals. It's crucial, especially in a study like this, to survey diverse groups of people who listen to various types of music to yield more comprehensive and accurate results.

Given more time or resources, we would conduct a larger scale survey that includes a broader population sample from a specific country or region. We would inquire about the participants' music listening history to derive more precise scores for each individual. Furthermore, we would endeavor to reach out to individuals experiencing each of the mental health conditions and administer separate surveys to gain deeper insights.

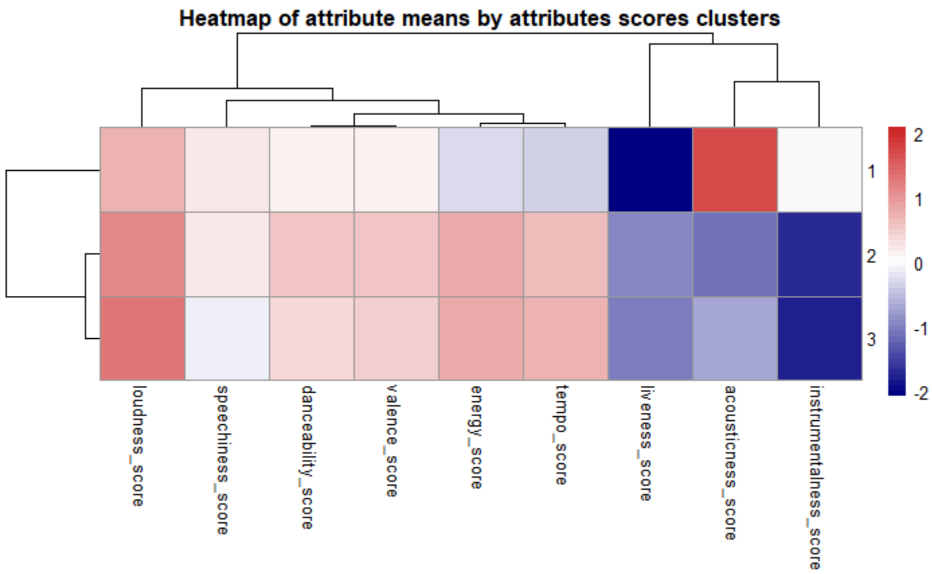
The steps we took in this project provide a foundation for further research. It is our hope that future investigations can address these limitations and continue the important work of exploring the correlation between music preferences and mental health.

Appendix:

1. correlation of calculated scores to mental health.



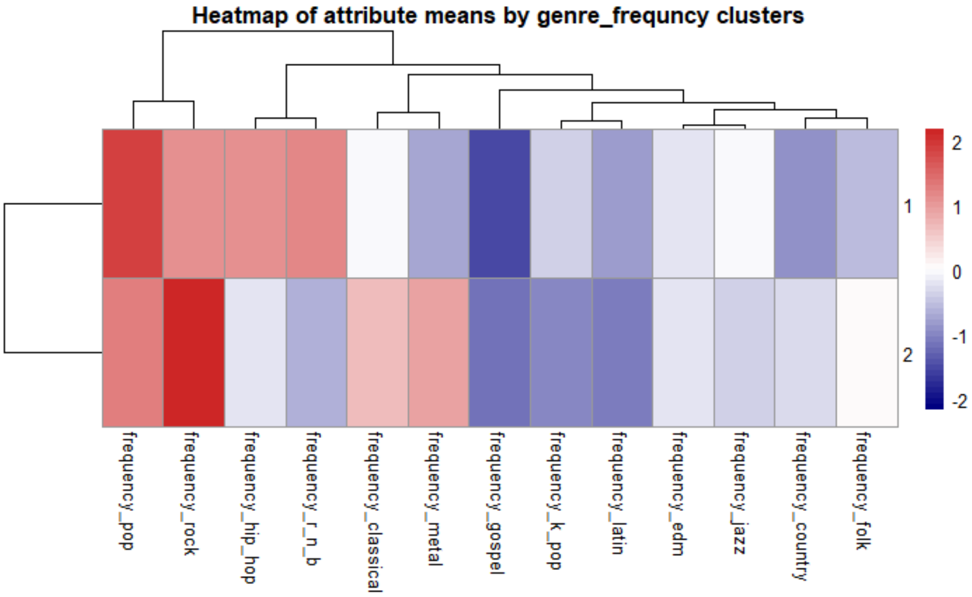
2. Attribute scores clustering:



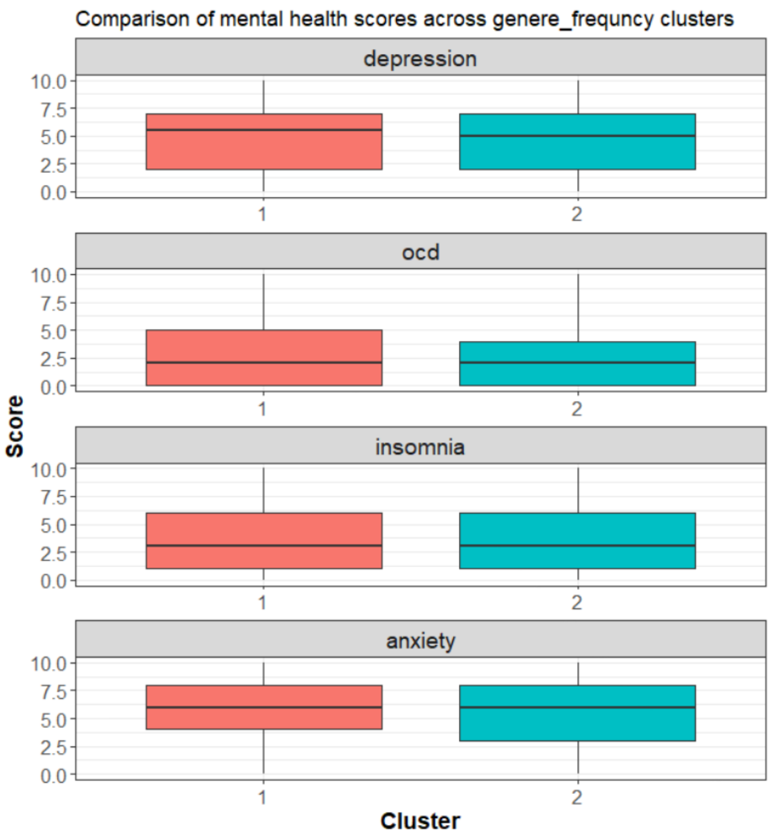
3. Result of ANOVA test for significant difference in depression scores between clusters of set 1:

```
                     Df Sum Sq Mean Sq F value Pr(>F)
as.factor(cluster)    2     82   41.03   4.553 0.0108 *
Residuals           726   6542    9.01
---
```

4. Genre listening frequency clustering (set 2):



5. Results for set2:



6. ANOVA tests results (showing only significant results out 13 tests between each group):

Variance test between low depression group and medium depression group:

```
[1] "frequency_metal"
             Df Sum Sq Mean Sq F value  Pr(>F)
group         1   0.97  0.9740    6.76 0.00961 **
Residuals   475  68.44  0.1441
---
```

```
[1] "frequency_rock"
             Df Sum Sq Mean Sq F value  Pr(>F)
group         1   1.03  1.0288   8.361 0.00401 **
Residuals   475  58.45  0.1231
```

Variance test between medium depression group and high depression group:

```
[1] "hours_per_day"
             Df Sum Sq Mean Sq F value Pr(>F)
group         1     54   54.06   6.361 0.0121 *
Residuals   397   3374    8.50
```

```
[1] "frequency_metal"
             Df Sum Sq Mean Sq F value Pr(>F)
group         1   0.70  0.7048   5.085 0.0247 *
Residuals   397  55.03  0.1386
```

Variance test between low depression group and high depression group:

```
[1] "hours_per_day"
             Df Sum Sq Mean Sq F value Pr(>F)
group         1     50   50.22   4.573 0.0332 *
Residuals   350   3843   10.98
```

```
[1] "frequency_folk"
             Df Sum Sq Mean Sq F value  Pr(>F)
group         1   0.83  0.8254   7.756 0.00565 **
Residuals   350  37.25  0.1064
---
```

```
[1] "frequency_hip_hop"
             Df Sum Sq Mean Sq F value Pr(>F)
group         1   0.75  0.7541   5.957 0.0152 *
Residuals   350  44.31  0.1266
```

```
[1] "frequency_metal"
             Df Sum Sq Mean Sq F value   Pr(>F)
group         1   2.69  2.6912    19.7 1.22e-05 ***
Residuals   350  47.82  0.1366
---
```

```
[1] "frequency_rock"
             Df Sum Sq Mean Sq F value   Pr(>F)
group         1    1.7  1.6976   13.38 0.000293 ***
Residuals   350   44.4  0.1268
---
```

7. Correlation of high OCD score and spoken words in a track:

	TableName <chr>	Feature <chr>	Correlation <dbl>	P_Value <dbl>
rho	high_ocd	speechiness_score	-0.3809579	0.002068169