MUSIC AND MENTAL HEALTH ANALYSIS

PUSL2076 Data Programming in R coursework

Name	Student ID
H.G.D.N.Nimeka	10899185
T.M.D.Semini Nethranjalee Peiris	10899484
W.Oshani Primalsha	10899485

R Project

Introduction

Music and mental health have an unbreakable connection, since music is a powerful measure for emotional expression and well-being. Music has been scientifically connected to the release of neurotransmitters that affect mood and stress levels. Music, whether by listening, playing, or composing, can help ease the symptoms of depression, anxiety, and stress.

The Music and Mental Health dataset is a result of a survey that collect data from individuals from different ages, cultures. With the help of this dataset, we hope to provide a complete statistical understanding of the link between music and mental health. Ideally, the findings could help apply Music Therapy with more understanding or simply offer amazing insights into the mind.

This data set has 736 rows and 33 columns.

The dataset includes 33 columns:

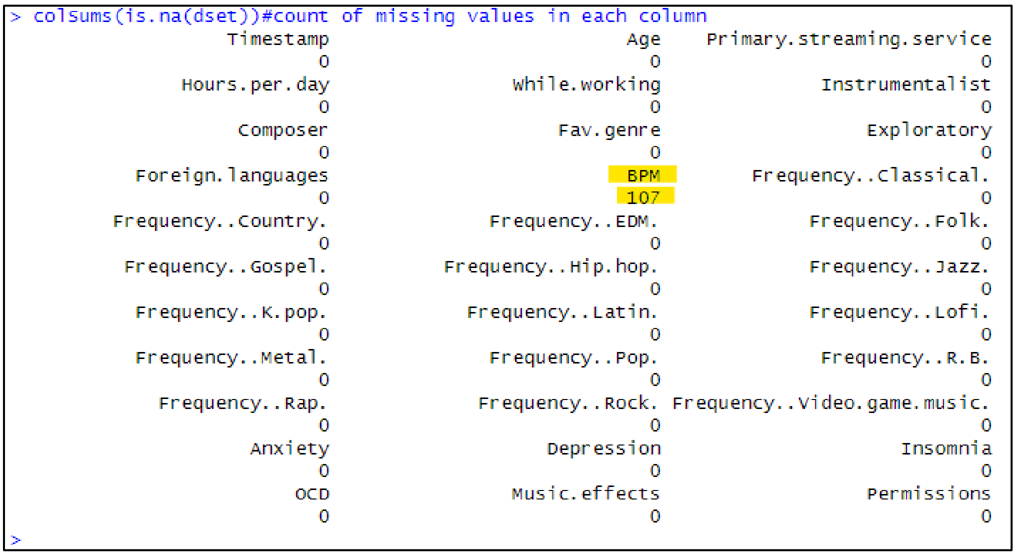
Column Name	Description	Data Type	Data Range
Timestamp	Date and Time the details submitted	Date and time	10/11/2022 15:46 - 9/9/2022 7:48
Age	Individuals age	Numeric	10-89
Primary Streaming Service	Individuals primary streaming service	Character	Spotify, Pandora, YouTube Music, Apple Music, Other streaming service, I do not use a streaming service
Hours per day	Number of hours per day that individual listen to music	Numeric	0-24
While working	Whether the individual listen to music while working or learning	Character	Yes/no
Instrumentalist	Whether the individual play an instrument	Character	Yes/no
Composer	Whether the individual compose music	Character	Yes/no
Fav Genre	Individuals’ favorite music genre	Character	Rock, Pop, Metal, Classical, Video game music
Exploratory	Whether the individual explore new music/artist	Character	Yes/no
Foreign languages	Whether the individual listen to music that is in foreign language, that he/she is not fluent in	Character	Yes/no

BPM	Beats per minute of favorite genre	Numeric	0-1000
Frequency Classical	How frequently does the individual listen to classical music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Country	How frequently does the individual listen to Country music	Character	Rarely , Sometimes , Never , Very frequently
Frequency EDM	How frequently does the individual listen to EDM music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Folks	How frequently does the individual listen to Folks music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Gospel	How frequently does the individual listen to Gospel music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Hiphop	How frequently does the individual listen to Hiphop music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Jazz	How frequently does the individual listen to Jazz music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Kpop	How frequently does the individual listen to Kpop music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Latin	How frequently does the individual listen to Latin music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Lofi	How frequently does the individual listen to Lofi music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Metal	How frequently does the individual listen to Metal music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Pop	How frequently does the individual listen to Pop music	Character	Rarely , Sometimes , Never , Very frequently
Frequency R&B	How frequently does the individual listen to R&B music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Rap	How frequently does the individual listen to Rap music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Rock	How frequently does the individual listen to Rock music	Character	Rarely , Sometimes , Never , Very frequently
Frequency Video game music	How frequently does the individual listen to Video game music	Character	Rarely , Sometimes , Never , Very frequently
Anxiety	Anxiety level reported individual	Numeric	0-10
Depression	Depression level reported individual	Numeric	0-10
Insomnia	Insomnia level reported individual	Numeric	0-10
OCD	OCD level reported individual	Numeric	0-10
Music effect	How does music effect on individuals' mental health	Character	Improve , No effect , Worsen
Permission	Permission to publicize data	Character	I understand

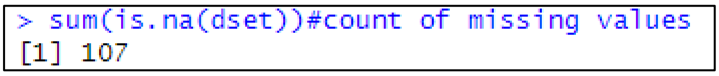
Data preparation

01. Handling null values

Total missing values in the dataset-107

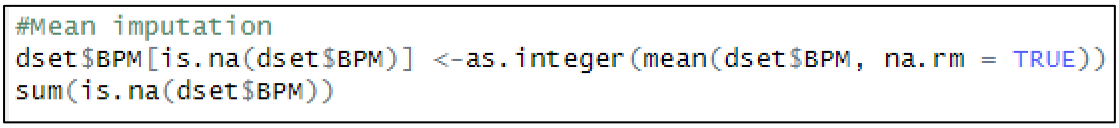


```
> colSums(is.na(dset))#count of missing values in each column
                 Timestamp                          Age     Primary.streaming.service 
                         0                            0                             0 
             Hours.per.day                while.working              Instrumentalist 
                         0                            0                             0 
                  Composer                    Fav.genre                  Exploratory 
                         0                            0                             0 
         Foreign.languages                          BPM       Frequency..Classical. 
                         0                          107                             0 
       Frequency..Country.               Frequency..EDM.             Frequency..Folk. 
                         0                            0                             0 
        Frequency..Gospel.           Frequency..Hip.hop.             Frequency..Jazz. 
                         0                            0                             0 
         Frequency..K.pop.             Frequency..Latin.             Frequency..Lofi. 
                         0                            0                             0 
         Frequency..Metal.               Frequency..Pop.              Frequency..R.B. 
                         0                            0                             0 
           Frequency..Rap.              Frequency..Rock.  Frequency..Video.game.music. 
                         0                            0                             0 
                   Anxiety                   Depression                     Insomnia 
                         0                            0                             0 
                       OCD                Music.effects                  Permissions 
                         0                            0                             0 
> 
```



```
> sum(is.na(dset))#count of missing values
[1] 107
```

Here only BPM column contains all null values. So use the mean in this column which is 123.3764 to solve the problem. Mean imputation is the most suitable to solve this issue.

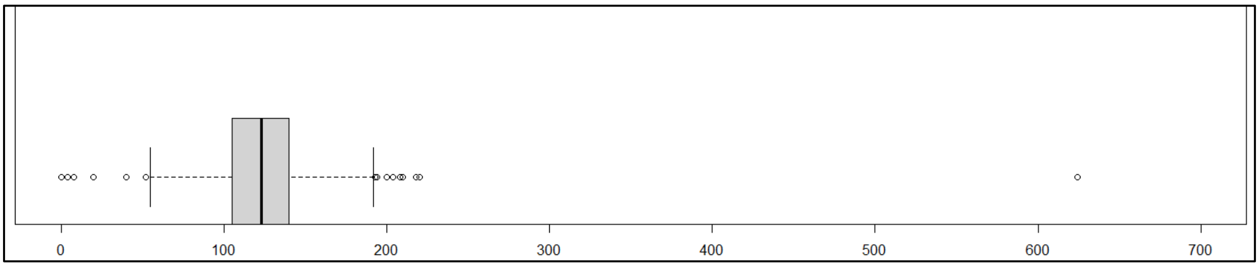


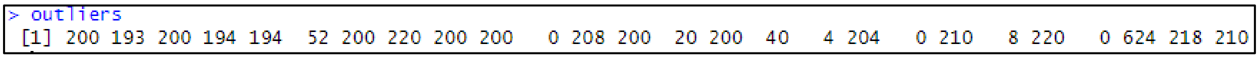
```
#Mean imputation
dset$BPM[is.na(dset$BPM)] <-as.integer(mean(dset$BPM, na.rm = TRUE))
sum(is.na(dset$BPM))
```

I. Outliers

In order to ensure that statistical analyses are strong, handling outliers is an essential step in the data preparation process. First, we searched for any outliers in numerical data using visualization tool which is boxplots.

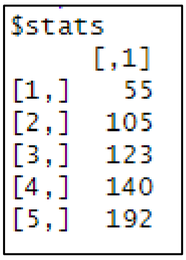
The only outlier problem in this data set is in the BPM column. Here is the boxplot to give a good overview.



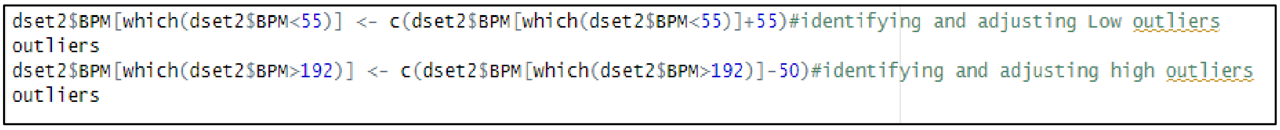


```
> outliers
 [1] 200 193 200 194 194  52 200 220 200 200   0 208 200  20 200  40   4 204   0 210   8 220   0 624 218 210
```

We used a systematic approach in our study of the company dataset to identify and fix outliers.

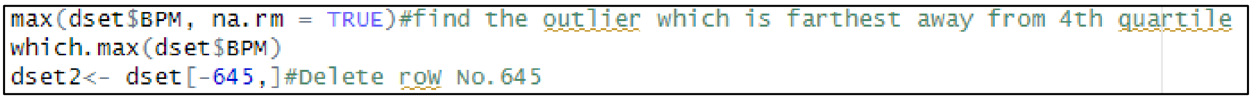


```
$stats
     [,1]
[1,]   55
[2,]  105
[3,]  123
[4,]  140
[5,]  192
```



```
dset2$BPM[which(dset2$BPM<55)] <- c(dset2$BPM[which(dset2$BPM<55)]+55)#identifying and adjusting Low outliers
outliers
dset2$BPM[which(dset2$BPM>192)] <- c(dset2$BPM[which(dset2$BPM>192)]-50)#identifying and adjusting high outliers
outliers
```

Low outliers in the BPM column (<55 because the beginning of the 1st quartile range is 55) are adjusted by adding 55, while high outliers (>192 because the end of the 3rd quartile range is 192) are adjusted by subtracting 50.



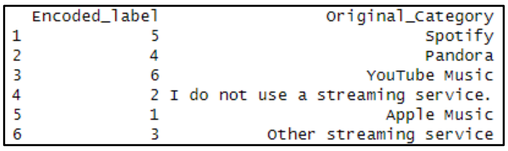
```
max(dset$BPM, na.rm = TRUE)#find the outlier which is farthest away from 4th quartile
which.max(dset$BPM)
dset2<- dset[-645,]#Delete row No.645
```

The maximum value of the BPM column of the record is displayed in line 645 with the value of 624 bpm. This is the outlier that deviates furthest from the fourth quartile. Since we can remove 5% records from the dataset that record 645 is removed to create a new dataset called "dset2" in order to reduce its influence.

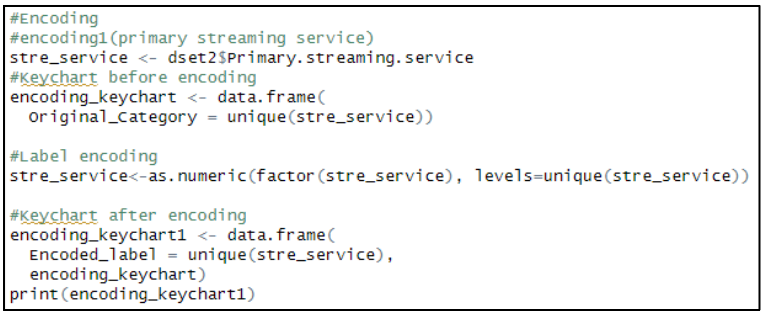
This targeted removal, and decreasing of the possible effects of outliers guarantees that the outlier, does not have a major impact on analyses or later modeling attempts, leading to a more robust and reliable dataset for statistical processing.

II. Encoding

Label encoding is a technique for converting categorical variables into numerical representations, making it easier to train machine learning models. The process includes extracting the column, creating an encoding key chart, and using the factor function for label encoding. The resulting Primary Streaming Service and Fav.genre columns are converted to numeric labels, simplifying categorical data representation.

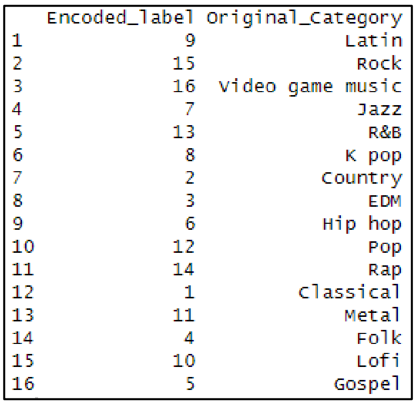


```
  Encoded_label                Original_Category
1             5                          Spotify
2             4                          Pandora
3             6                    YouTube Music
4             2 I do not use a streaming service.
5             1                      Apple Music
6             3          Other streaming service
```

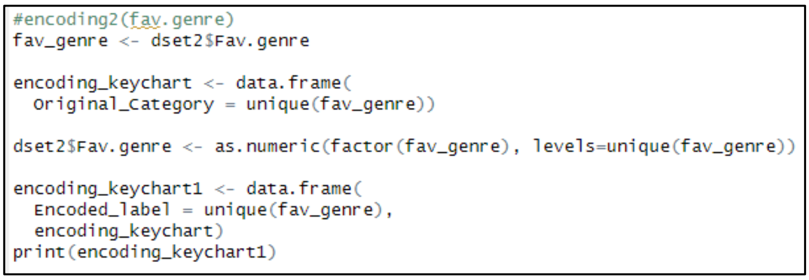


```
#Encoding
#encoding1(primary streaming service)
stre_service <- dset2$Primary.streaming.service
#Keychart before encoding
encoding_keychart <- data.frame(
  Original_Category = unique(stre_service))

#Label encoding
stre_service<-as.numeric(factor(stre_service), levels=unique(stre_service))

#Keychart after encoding
encoding_keychart1 <- data.frame(
  Encoded_label = unique(stre_service),
  encoding_keychart)
print(encoding_keychart1)
```

The encoding key chart provides a visual guide to understanding the correspondence between original and encoded values.



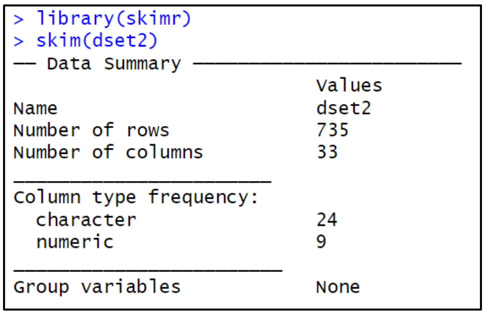
```
   Encoded_label Original_Category
1              9             Latin
2             15              Rock
3             16  Video game music
4              7              Jazz
5             13               R&B
6              8             K pop
7              2           Country
8              3               EDM
9              6           Hip hop
10            12               Pop
11            14               Rap
12             1         Classical
13            11             Metal
14             4              Folk
15            10              Lofi
16             5            Gospel
```



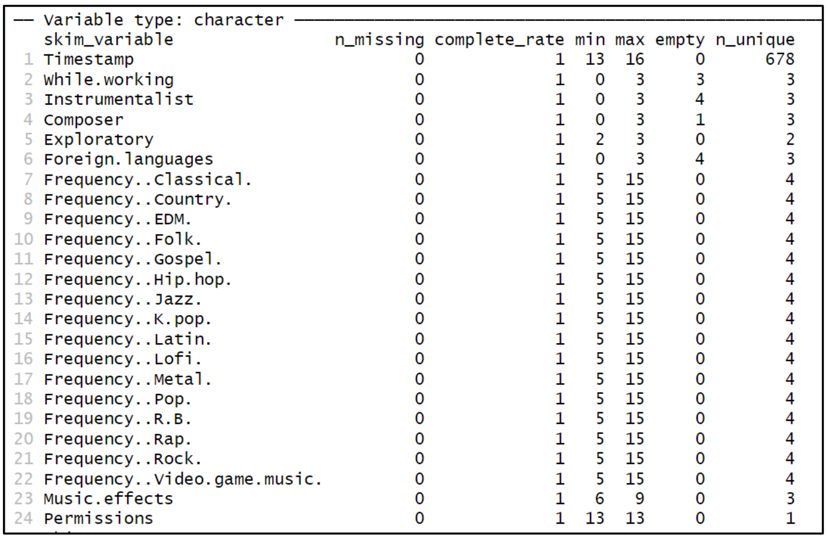
```
#encoding2(fav.genre)
fav_genre <- dset2$Fav.genre

encoding_keychart <- data.frame(
  Original_Category = unique(fav_genre))

dset2$Fav.genre <- as.numeric(factor(fav_genre), levels=unique(fav_genre))

encoding_keychart1 <- data.frame(
  Encoded_label = unique(fav_genre),
  encoding_keychart)
print(encoding_keychart1)
```

Getting a summary of prepared dataset

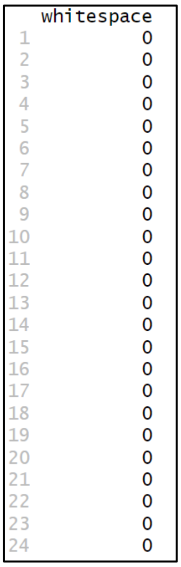
To get a clear idea about the finalized dataset we can use “skimr” package. The “skimr” is a package in R is designed to provide a short yet informative summary of a dataset's important characteristics of a dataset.



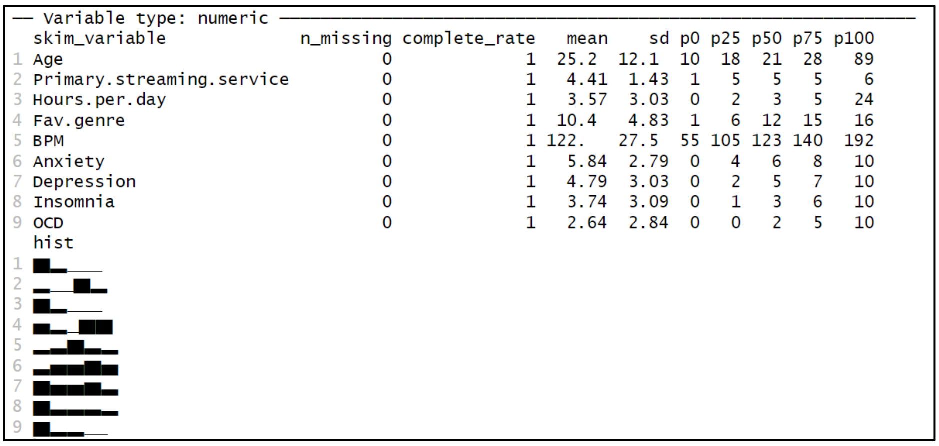
```
> library(skimr)
> skim(dset2)
── Data Summary ────────────────────────
                           Values
Name                       dset2
Number of rows             735
Number of columns          33
_______________________
Column type frequency:
  character                24
  numeric                  9
________________________
Group variables            None
```



```
── Variable type: character ─────────────────────────────────────────
   skim_variable                 n_missing complete_rate min max empty n_unique
 1 Timestamp                             0             1  13  16     0      678
 2 While.working                         0             1   0   3     3        3
 3 Instrumentalist                       0             1   0   3     4        3
 4 Composer                              0             1   0   3     1        3
 5 Exploratory                           0             1   2   3     0        2
 6 Foreign.languages                     0             1   0   3     4        3
 7 Frequency..Classical.                 0             1   5  15     0        4
 8 Frequency..Country.                   0             1   5  15     0        4
 9 Frequency..EDM.                       0             1   5  15     0        4
10 Frequency..Folk.                      0             1   5  15     0        4
11 Frequency..Gospel.                    0             1   5  15     0        4
12 Frequency..Hip.hop.                   0             1   5  15     0        4
13 Frequency..Jazz.                      0             1   5  15     0        4
14 Frequency..K.pop.                     0             1   5  15     0        4
15 Frequency..Latin.                     0             1   5  15     0        4
16 Frequency..Lofi.                      0             1   5  15     0        4
17 Frequency..Metal.                     0             1   5  15     0        4
18 Frequency..Pop.                       0             1   5  15     0        4
19 Frequency..R.B.                       0             1   5  15     0        4
20 Frequency..Rap.                       0             1   5  15     0        4
21 Frequency..Rock.                      0             1   5  15     0        4
22 Frequency..Video.game.music.          0             1   5  15     0        4
23 Music.effects                         0             1   6   9     0        3
24 Permissions                           0             1  13  13     0        1
```



```
   whitespace
 1          0
 2          0
 3          0
 4          0
 5          0
 6          0
 7          0
 8          0
 9          0
10          0
11          0
12          0
13          0
14          0
15          0
16          0
17          0
18          0
19          0
20          0
21          0
22          0
23          0
24          0
```



```
── Variable type: numeric ───────────────────────────────────────────────────
  skim_variable            n_missing complete_rate   mean    sd p0 p25 p50 p75 p100
1 Age                              0             1  25.2  12.1  10  18  21  28   89
2 Primary.streaming.service        0             1   4.41  1.43  1   5   5   5    6
3 Hours.per.day                    0             1   3.57  3.03  0   2   3   5   24
4 Fav.genre                        0             1  10.4   4.83  1   6  12  15   16
5 BPM                              0             1 122.   27.5  55 105 123 140  192
6 Anxiety                          0             1   5.84  2.79  0   4   6   8   10
7 Depression                       0             1   4.79  3.03  0   2   5   7   10
8 Insomnia                         0             1   3.74  3.09  0   1   3   6   10
9 OCD                              0             1   2.64  2.84  0   0   2   5   10
  hist
1 ▇▂▁▁▁
2 ▂▁▁▇▂
3 ▇▂▁▁▁
4 ▅▂▁▇▇
5 ▁▂▇▂▁
6 ▂▆▆▇▆
7 ▇▅▅▆▃
8 ▇▃▃▃▂
9 ▇▂▂▁▁
```

Data Visualization

Users age distribution:

Age limit	No. of Individuals
10-20	339
20-30	247
30-40	76
40-50	26
50-60	26
60-70	15
70-80	5
80-90	1

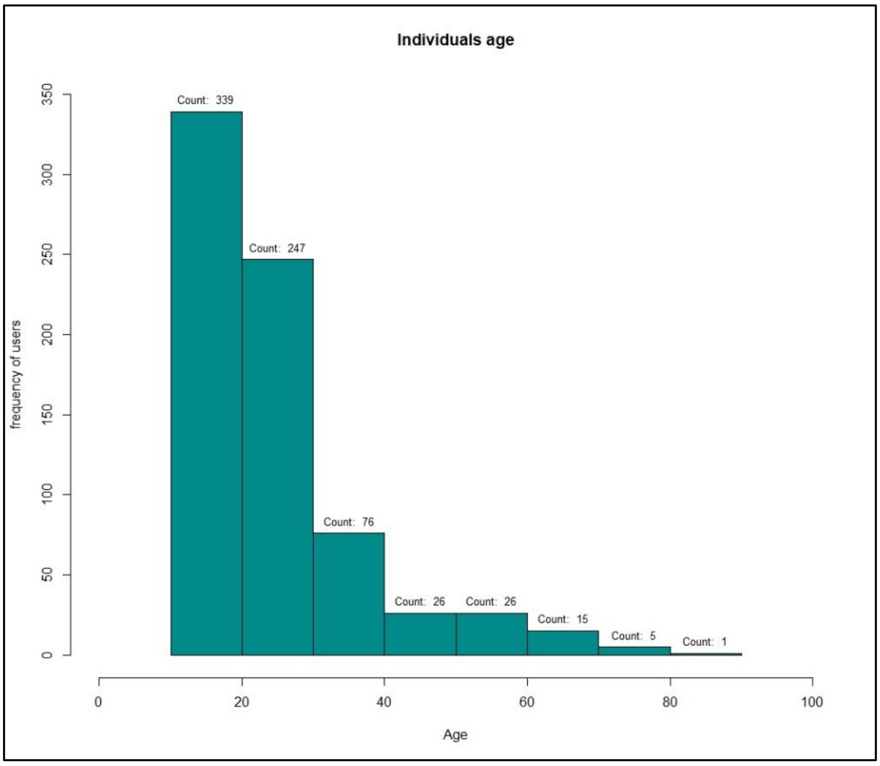


Figure 1-Individuals Age Distribution

According to this histogram, 46.12% of the people in the sample are between the ages of 10 and 20, while 33.6% are between the ages of 20 and 30.

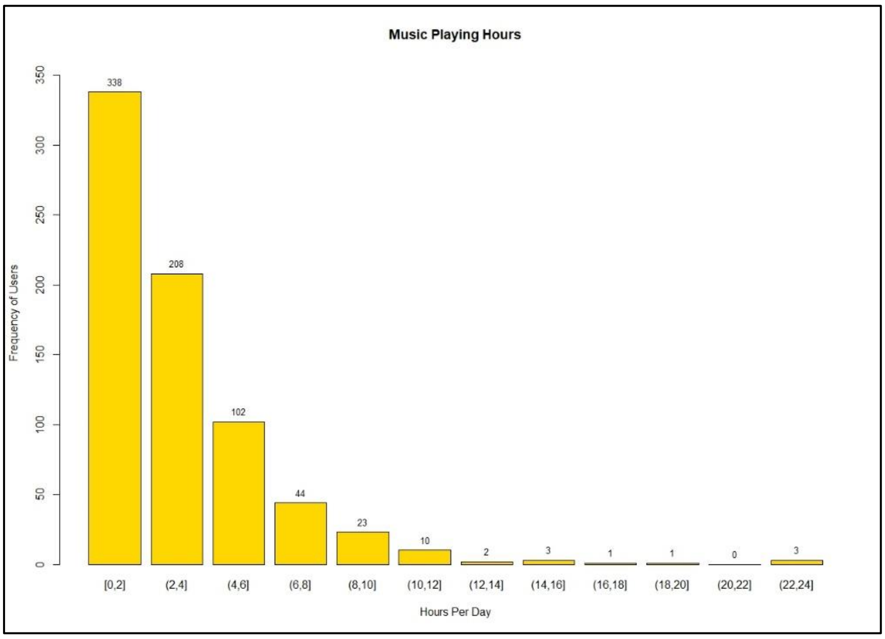
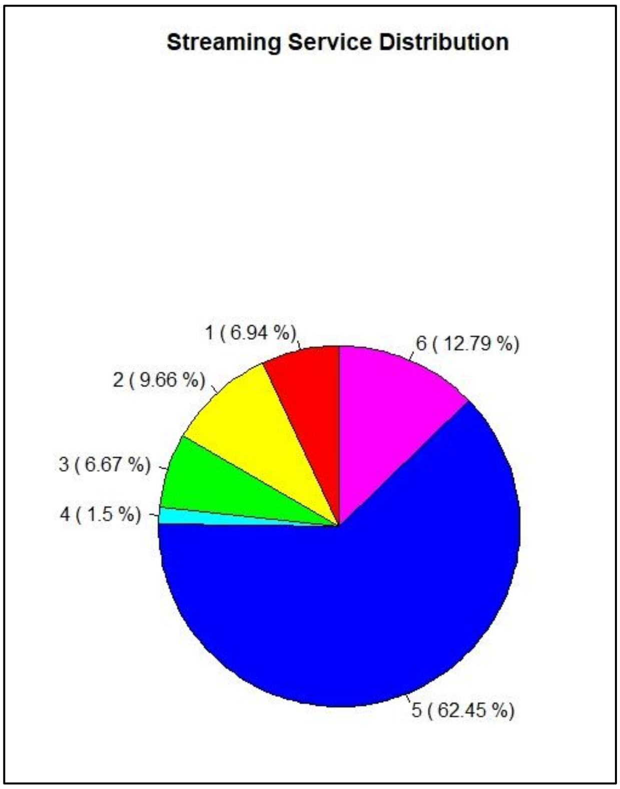


Figure 2 – Music Playing Hours Distribution

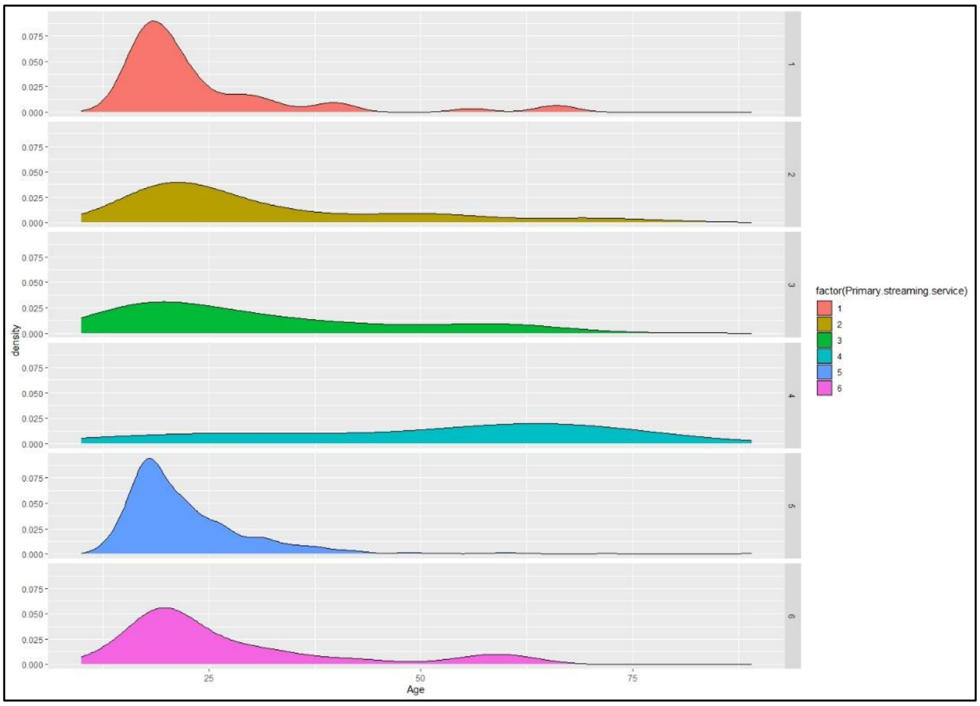
As this bar chart shows 45.98% of the overall sample listen to music for 0 to 2 hours, 28.29% listen to 2 to 4 and 13.87% listen 4 to 6.

Streaming service Distribution



Encode No.	Original Category	User Count
1	Apple Music	51
2	I do not use a streaming service	71
3	Other streaming service	49
4	Pandora	11
5	Spotify	459
6	Youtube Music	94

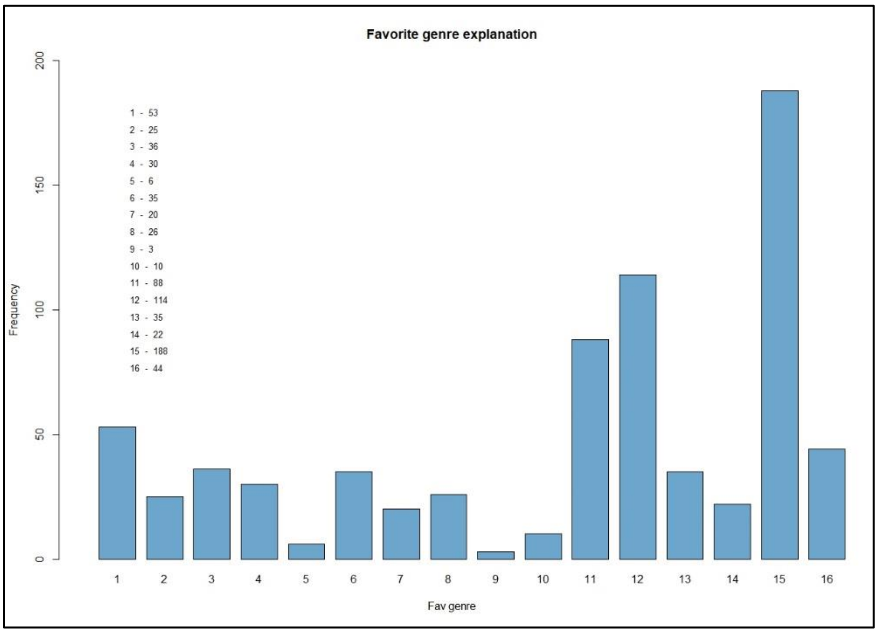
This pie chart shows the primary streaming service that the person uses to listen to music. According to this, 62.45% use Spotify(5) as their primary streaming service, while Pandora(4) being the least preferred option, with a usage rate as low as 1.5% from the surveyed individuals. 12.79% with the count of 94 people uses Youtube Music(6) and 6.94% uses with the count of 51 people uses Apple Music(1) for stream music. 9.66% of surveyed individuals does not use a streaming service(2).



This density plot explains the ages of different streaming service users. This highlight that many who prefer Spotify(5) are under the age of 25, while fewer individual who choose Pandora(4) are between the ages of 50 and 75 from the surveyed individuals.

Favorite genre distribution

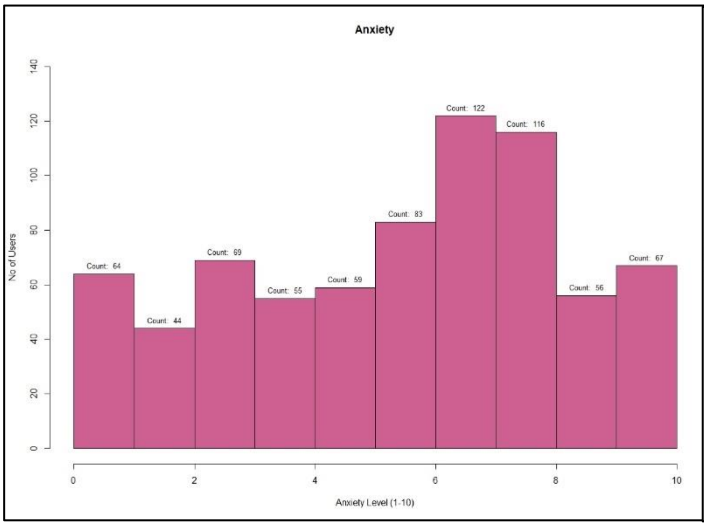
This bar chart shows the favorite music genre of the surveyed individuals.

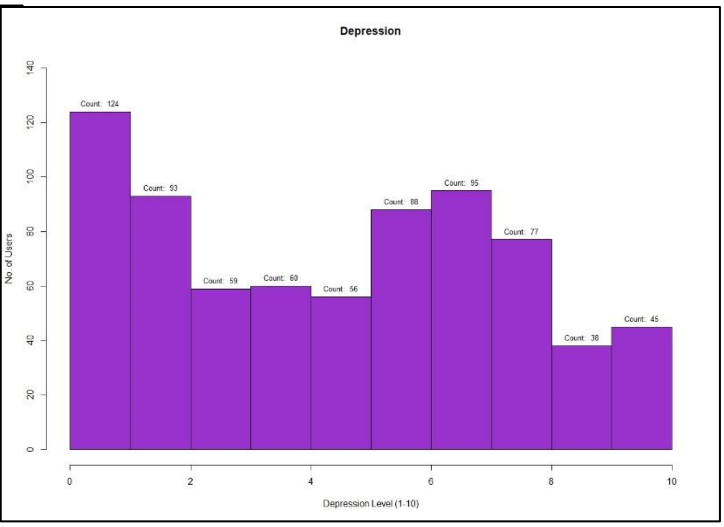


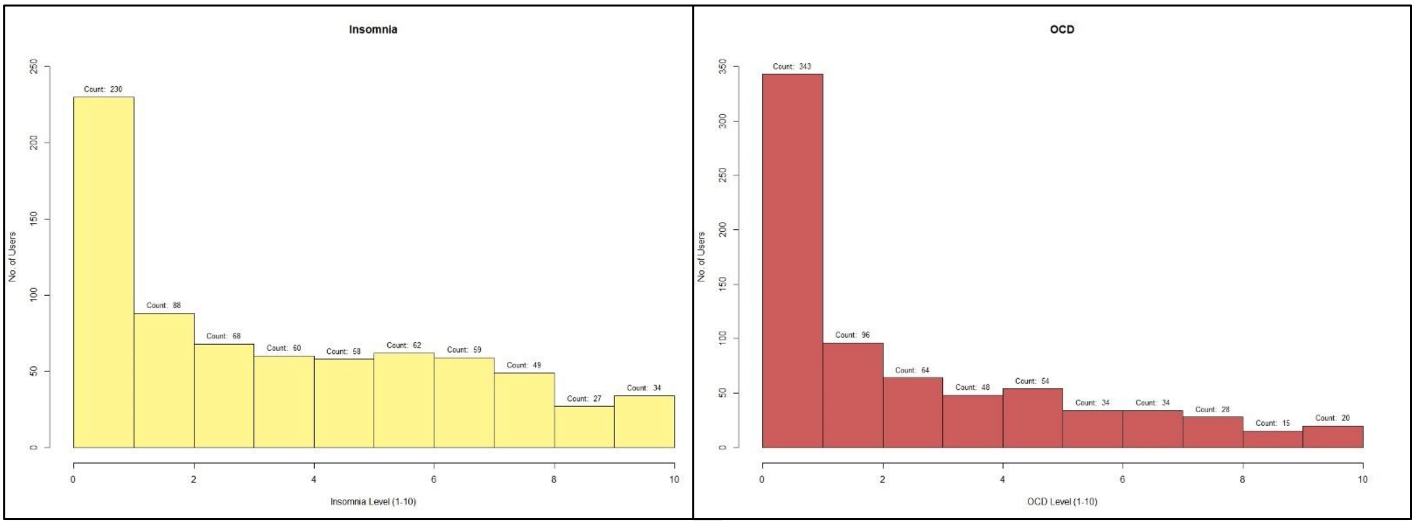
Encode No.	Original Category	User Count
1	Classical	53
2	Country	25
3	EDM	36
4	Folk	30
5	Gospel	6
6	HipHop	35
7	Jazz	20
8	Kpop	26
9	Latin	3
10	Lofi	10
11	Metal	88
12	Pop	114
13	R&B	35
14	Rap	22
15	Rock	188
16	Video Game Music	44

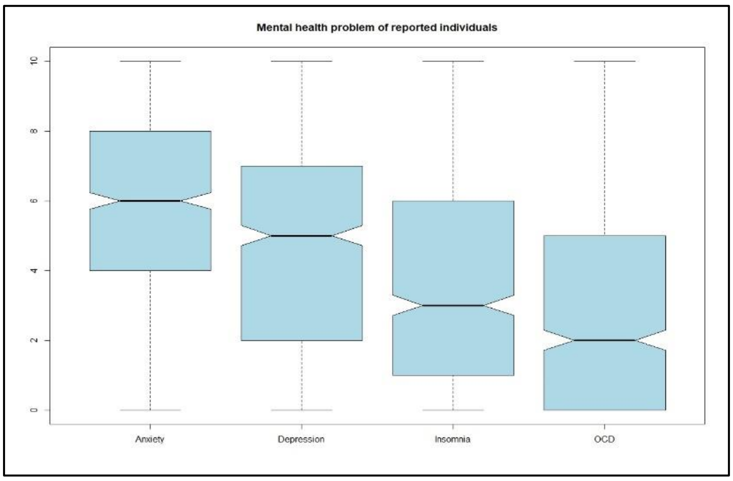
Exploring of Mental Health: Analysis of OCD, Depression, Anxiety, and Insomnia Levels

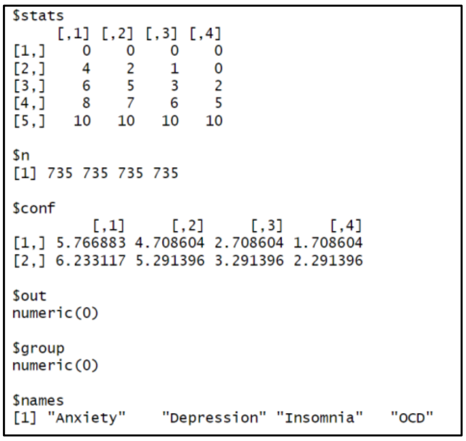
The below 4 plots focuses on the anxiety, depression, insomnia, and OCD levels of the surveyed individuals on a scale of 1 to 10. Through dedicated bar graphs for each condition, it visually represents the distribution of reported levels and the number of people in each level, offering a comprehensive view of individuals' mental health experiences.









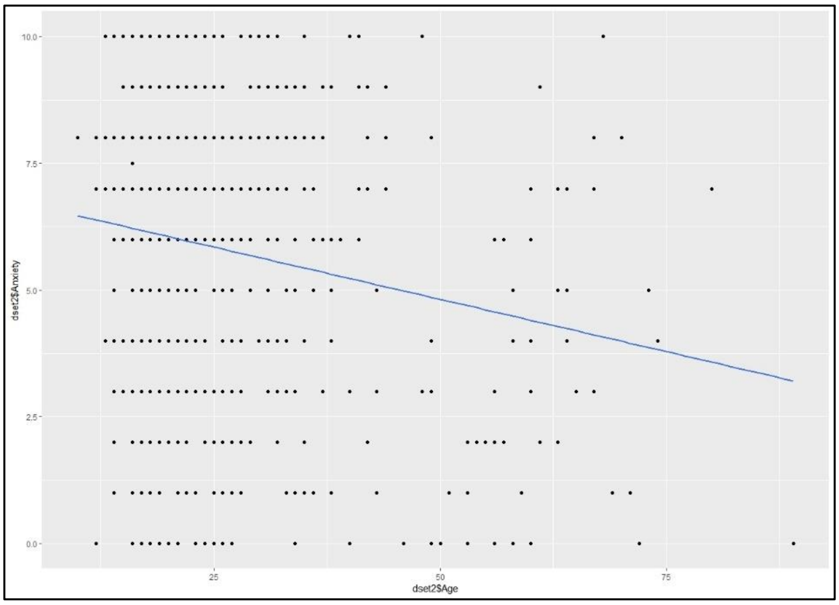


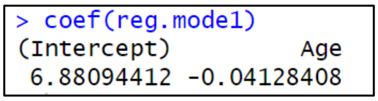
```
$stats
     [,1] [,2] [,3] [,4]
[1,]    0    0    0    0
[2,]    4    2    1    0
[3,]    6    5    3    2
[4,]    8    7    6    5
[5,]   10   10   10   10

$n
[1] 735 735 735 735

$conf
         [,1]     [,2]     [,3]     [,4]
[1,] 5.766883 4.708604 2.708604 1.708604
[2,] 6.233117 5.291396 3.291396 2.291396

$out
numeric(0)

$group
numeric(0)

$names
[1] "Anxiety"    "Depression" "Insomnia"   "OCD"
```

In order to farther study about mental health indicators, this grouped boxplot of anxiety, depression, insomnia, and OCD levels of individuals, each boxplot reveals the distribution and central tendencies of these mental health indicators.

Mental Health – Age Analysis - “ggplot2” package: Analyze how mental health problems occur with age

Anxiety – Age:



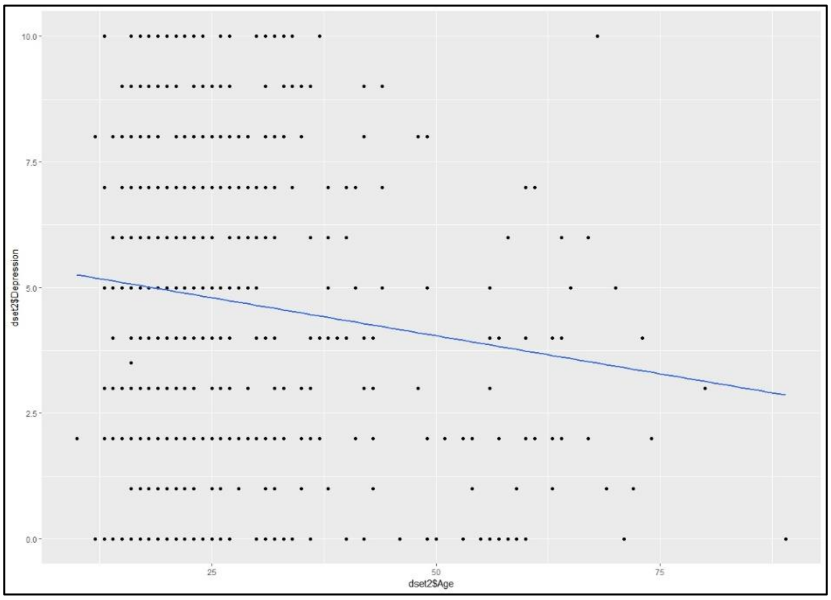


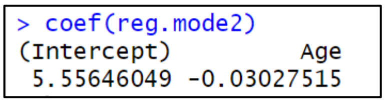
```
> coef(reg.mode1)
(Intercept)         Age
 6.88094412 -0.04128408
```

Figure 1 : Anxiety – Age

This scatter plot visualizes the relationship between anxiety levels and age. It includes a linear regression line to present the trend. The beginning anxiety level is represented by the model's intercept - 6.88, while the coefficient - (-0.04) reveals that anxiety decreases slightly with increasing age. This proves that anxiety tends to slightly reduce with age.

Depression – Age:



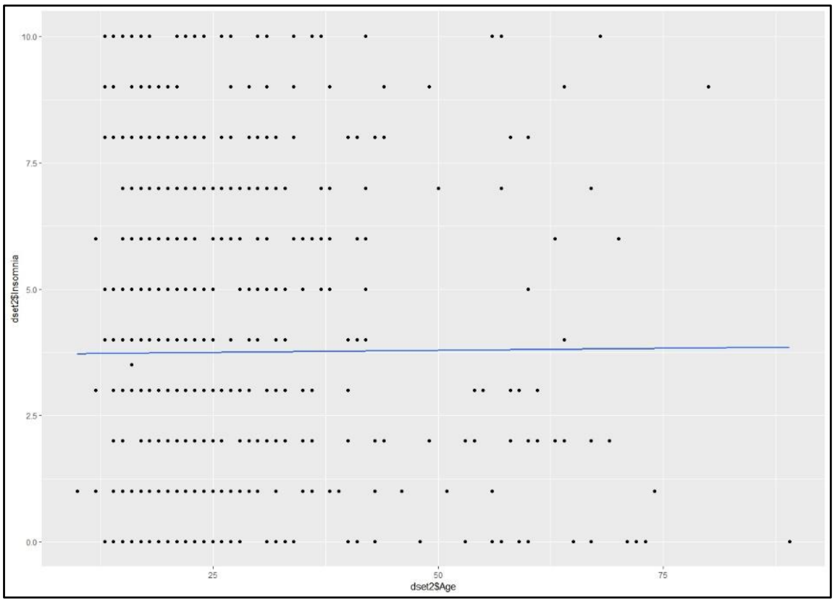


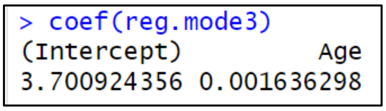
```
> coef(reg.mode2)
(Intercept)         Age
 5.55646049 -0.03027515
```

Figure 2 : Depression – Age

This plot visualizes the relationship between Depression levels and age. In here, in the linear regression line the intercept is 5.55 and the negative slope is -0.03 meaning that when age increase by a unit depression level will decrease by 0.03.

Insomnia – Age:



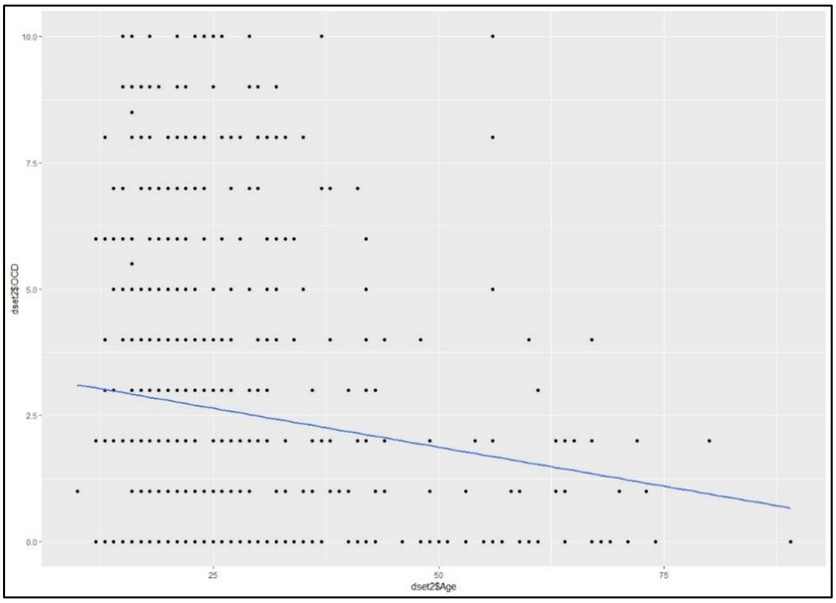


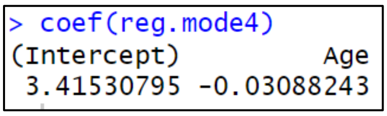
```
> coef(reg.mode3)
(Intercept)         Age
3.700924356 0.001636298
```

Figure 3 : Insomnia – Age

This graph reveals the prevalence of insomnia in people of various ages. With these data, the linear regression line indicates an intercept of approximately 3.70 and a minimal positive slope (0.0016) for 'Age.' This means that there is a slight increase in insomnia for every age unit increase.

OCD – Age:



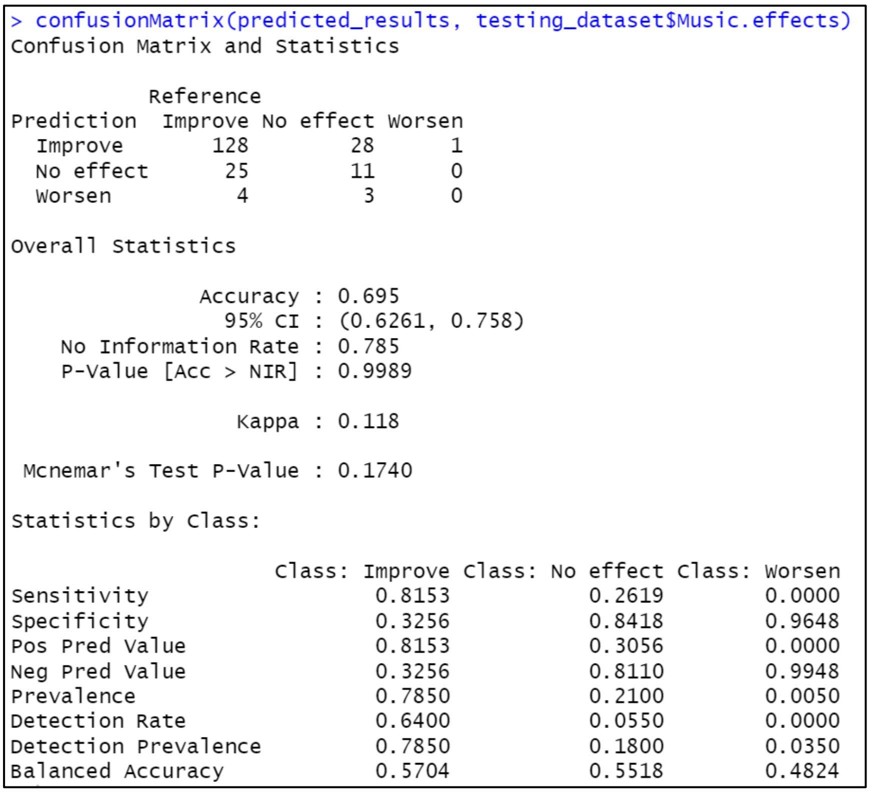


```
> coef(reg.mode4)
(Intercept)         Age
 3.41530795 -0.03088243
```

Figure 4 : OCD – Age

This graph reveals the prevalence of OCD in people of various ages. With these data, the linear regression line indicates an intercept of approximately 3.41 and a negative slope of -0.03 for 'Age.' This means that there is a decrease in OCD for every age unit increase.

Data Classification: Confusion matrix



```
> confusionMatrix(predicted_results, testing_dataset$Music.effects)
Confusion Matrix and Statistics

           Reference
Prediction  Improve No effect Worsen
  Improve       128        28      1
  No effect      25        11      0
  Worsen          4         3      0

Overall Statistics

               Accuracy : 0.695
                 95% CI : (0.6261, 0.758)
    No Information Rate : 0.785
    P-Value [Acc > NIR] : 0.9989

                  Kappa : 0.118

 Mcnemar's Test P-Value : 0.1740

Statistics by Class:

                     Class: Improve Class: No effect Class: Worsen
Sensitivity                  0.8153           0.2619        0.0000
Specificity                  0.3256           0.8418        0.9648
Pos Pred Value               0.8153           0.3056        0.0000
Neg Pred Value               0.3256           0.8110        0.9948
Prevalence                   0.7850           0.2100        0.0050
Detection Rate               0.6400           0.0550        0.0000
Detection Prevalence         0.7850           0.1800        0.0350
Balanced Accuracy            0.5704           0.5518        0.4824
```

Here, the Naive Bayes classifier was trained using the dataset with the target variable “music effects”, with the aim of predicting whether listening to music improves, worsens, or has no effect at all. And the confusion matrix reveals the model's performance.

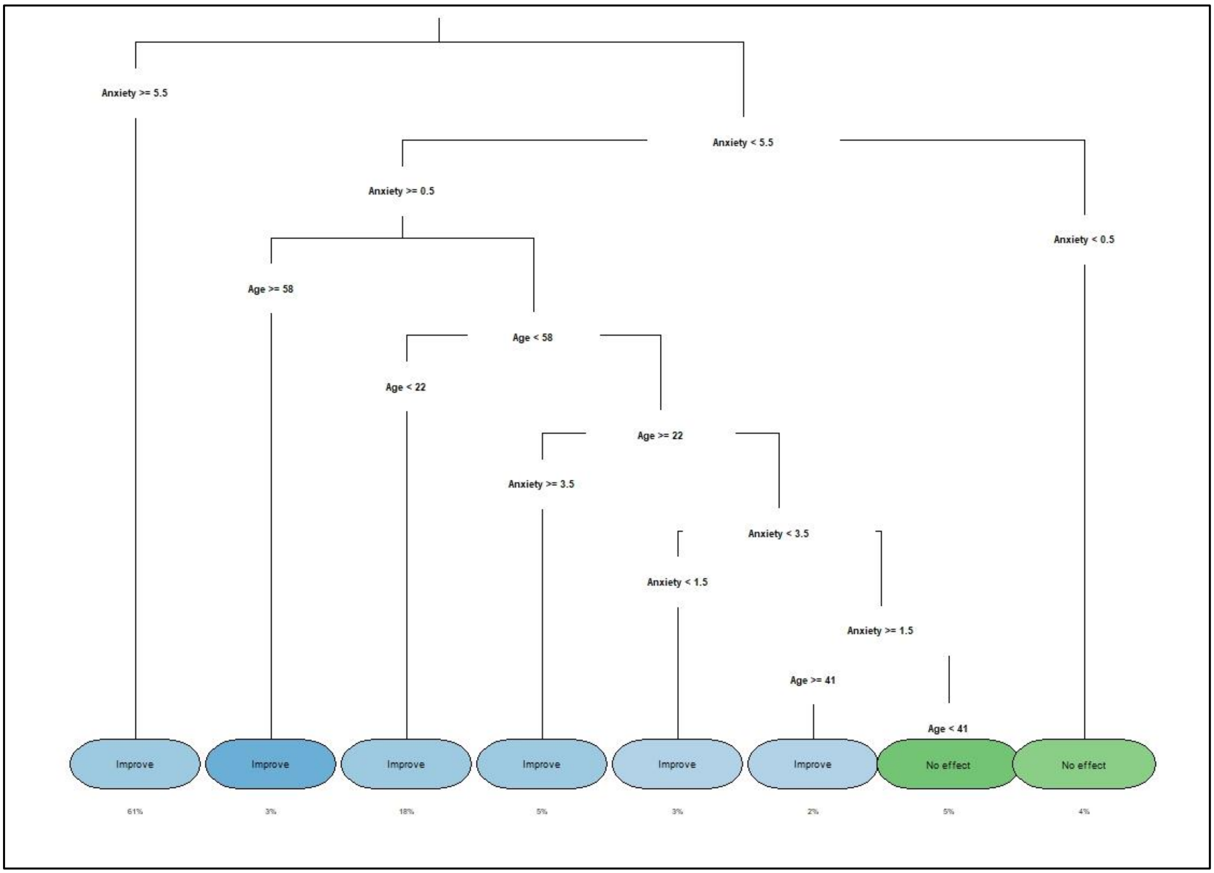
The number of correctly predicted incidents among all instances is 69.5%, which is the overall accuracy of the Naive Bayes model. For the “Improve” class, the sensitivity (true positive rate) is 81.53%; for the “No Impact” class it is 26.19%; and for the “Worsen” class it is 0%. For each of the associated classes, the specificity (true negative rate) is 32.56%, 84.18% and 96.48%. These measurements show that different classes have different success rates when it comes to the model's ability to recognize instances.

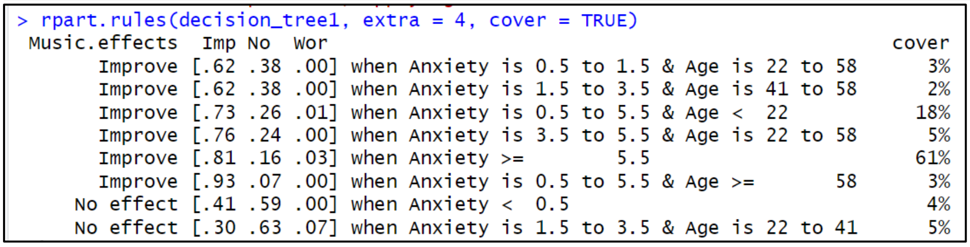
Data Classification: Decision Tree

The decision tree is a powerful tool for data analysis and decision making, providing an understandable and visual representation of the complex relationships within dataset. using “rpart” and “rpart.plot” packages, which are crucial for modelling and visualizing decision trees. “rpart” facilitates the construction of decision trees based on recursive partitioning, while “rpart.plot” improves the visualization of these trees.

Decision tree Analysis of music effects:

Age - Anxiety

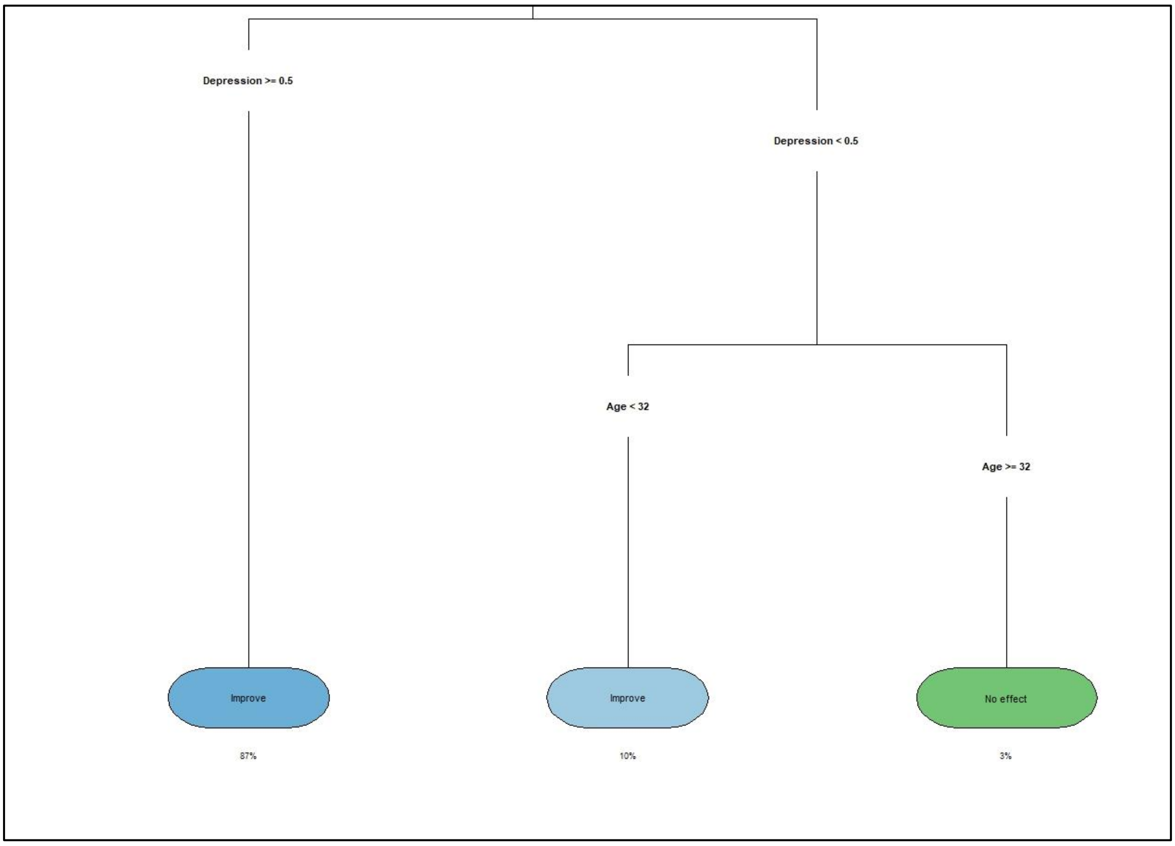


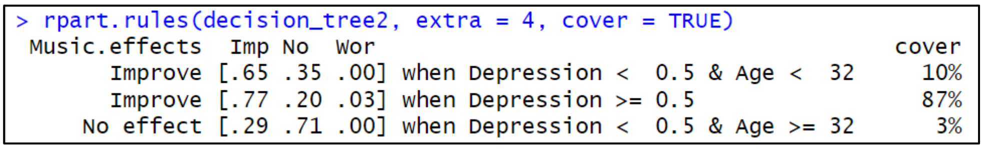


```
> rpart.rules(decision_tree1, extra = 4, cover = TRUE)
 Music.effects  Imp No  Wor                                              cover
       Improve [.62 .38 .00] when Anxiety is 0.5 to 1.5 & Age is 22 to 58     3%
       Improve [.62 .38 .00] when Anxiety is 1.5 to 3.5 & Age is 41 to 58     2%
       Improve [.73 .26 .01] when Anxiety is 0.5 to 5.5 & Age <  22          18%
       Improve [.76 .24 .00] when Anxiety is 3.5 to 5.5 & Age is 22 to 58     5%
       Improve [.81 .16 .03] when Anxiety >=        5.5                      61%
       Improve [.93 .07 .00] when Anxiety is 0.5 to 5.5 & Age >=       58     3%
     No effect [.41 .59 .00] when Anxiety <  0.5                              4%
     No effect [.30 .63 .07] when Anxiety is 1.5 to 3.5 & Age is 22 to 41     5%
```

The output of “rpart.rules” describes decision rules that produced from the decision tree model. Here, the given rules clearly define conditions that lead to improvement or have no effect at all. Each of them is linked to specific age and anxiety areas and also, the percentages show the distribution of results by calculating the frequency of each rule in the data set. For example, those who have anxiety levels 5.5 or over 5.5 are likely to improve 61% of the time, while those between the ages of 22 and 58 who have anxiety levels between 0.5 and 1.5 are likely to improve 3% of the time.

Age - Depression





```
> rpart.rules(decision_tree2, extra = 4, cover = TRUE)
 Music.effects  Imp No  Wor                                cover
       Improve [.65 .35 .00] when Depression <  0.5 & Age <  32    10%
       Improve [.77 .20 .03] when Depression >= 0.5                87%
     No effect [.29 .71 .00] when Depression <  0.5 & Age >= 32     3%
```

This shows decision rules for predicting Music effects taking variables Depression and Age. Specifically, individuals with a depression level greater than 0.5 have an 87% chance of improvement, while people with a depression level less than 0.5 and age under 32 have a 10% chance of improvement. People with depression level less than 0.5 but age 32 years or older also have a 3% chance of no effect.

Age – Insomnia & Age - OCD

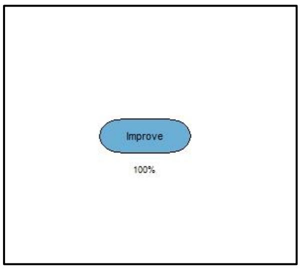
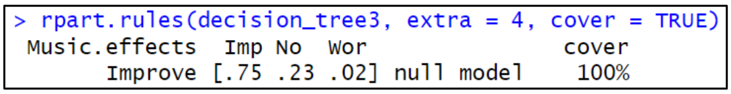


Figure 1: Age-Insomnia(decision_tree3)



```
> rpart.rules(decision_tree3, extra = 4, cover = TRUE)
 Music.effects  Imp No  Wor            cover
       Improve [.75 .23 .02] null model    100%
```

These trees shown indicate a significant improvement in music effects with a probability of 75% under conditions where a null model occurs. This suggests that the wide range of music effects may not be fully explained by current predictors such as insomnia and OCD. The large coverage of the decision tree of 100% shows the dominance of the null model.

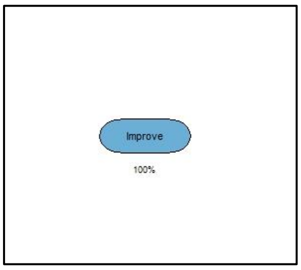
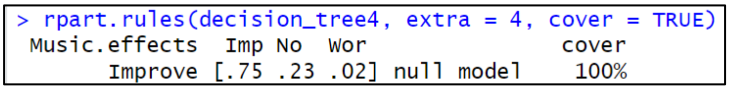


Figure 2: Age-OCD(decision_tree4)



```
> rpart.rules(decision_tree4, extra = 4, cover = TRUE)
 Music.effects  Imp No  Wor            cover
       Improve [.75 .23 .02] null model    100%
```

Conclusion

Overall, our thorough examination of the dataset using a variety of machine learning methods, including decision trees, Naive Bayes classifiers, and different visualization techniques, offers a statistical understanding about how music and mental health's deep connection. With Naive Bayes classifier giving 69.5% overall accuracy in predicting whether music improves, worsens, or has no effect on mental health, we can say music is indeed powerful measure for emotional expression and well-being.

Contribution

Data preparation	H.G.D.N.Nimeka	10899185
Data visualization and regression	W.Oshani Primalsha	10899485
Data classification	T.M.D.Semini Nethranjalee Peiris	10899484
Data analytics through models	H.G.D.N.Nimeka	10899185