An Exploratory Analysis on Music and Mental Health

Report submitted to the

SDM COLLEGE(Autonomous)

in partial fulfilment of the degree of

MASTER OF SCIENCE

IN

STATISTICS

by

NAVYA KAMATH (214005)

Under the supervision of

Ms. Shwetha Kumari

Department of Postgraduate Studies

In Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE (Autonomous)

UJIRE-574240

2023

SRI DHARMASTHALA MANJUNATHESHWARA COLLEGE (AUTONOMOUS), UJIRE-574240

Department of P.G. Studies in Statistics

CERTIFICATE

Certified that this is the bonafide record of project work done by Navya Kamath during year 2022-23 as a part of M.Sc (Statistics) third semester course work.

Reg. No: 214005

Project Guide: Ms.Shwetha Kumari Head of the Dept: Prof.Shanthi Prakash

Place: Ujire

Date:

ACKNOWLEDGEMENTS

Firstly, I would like to thank our Principal Dr.A.Jayakumar Shetty and our Dean Dr.Vishwanatha P for their support. It is my pleasure to thank Prof.Shanthi Prakash, HOD of Statistics, for his support.

I would also like to thank Ms.Shwetha Kumari, Asst.Professor, Department of Statistics, SDM College, Ujire, Dr.Savitha Rao, Asst.Professor, Department of Statistics, SDM College, Ujire, Ms.Supriya S.P. , Asst.Professor, Department of Statistics, SDM College, Ujire, Mr.Pradeep, Asst.Professor, Department of Statistics, SDM College, Ujire, Ms.Swathi Shetty, Asst.Professor, Department of Statistics, SDM College, Ujire for their kind help, encouragement and support.

Finally to all who helped me in many ways, I say, “Thank You!”.

Navya Kamath

Contents

1 Chapter 1

Introduction

1.1 Introduction

Music is a crucial element of everyday life and plays a central role in all human cultures. It is listened to and played by person of all ages, races and ethnic backgrounds. Music is not simply entertainment, from scientific study's it is proven that music can influence both physical as well as mental wellbeing of a human being.

Metal health diseases such as depression and anxiety can have devasting consequences both for patients and their families. Symptoms can be severe and make an person too weak, leaving individuals alone and isolated. Relationships among family and friends may suffer and the person might not receive the support needed to manage the disease. Music can not only improve symptoms associated with mental illness but it can also provide an environment for social interaction. It can help an individual to express emotions while producing state of mental relaxation.

Studies on people suffering from mental illness have shown a visible improvement after using music as a primary tool. Other studies have also shown how music is beneficial in improving heart rate, motor skills, brain stimulation and immune system enhancement.

Mental and physical illness can be expensive because of the medications and psychological treatment regimen. Interventions using music offers music-based activities in both therapeutic environment(Music therapy) with the support of a trained professional, and non-therapeutic setting, providing an atmosphere that is positive and supportive to treat symptoms associated with various disorders.

1.2 Literature Review

Lander Roman (2022) in his report regarding "Effect of Music on Well-being and Mental-health" used SOEP-Core V37 database of the Socio-Economic Panel by the German Institute for Economic Research, DIW Berlin. Under this study the Propensity Score Matching method is used to show how much of a effect does music have on the stress and depression.

Xinlei Dong, Xin Kang and Xiaolei Ding (2022) in their research about "Influence and Analysis of Music Teaching Environment Monitoring on Student's Mental Health using Data Mining technology" have collected data from students regarding Music Education system. Based on those data, using various line charts and radar chart researchers have shown that music health, whether it be formal music instruction or artistic practise related music, it is significant way to cultivate college student's psychological wellness.

1.3 Objectives

1. To identify which streaming service is most used.
2. To identify which genre of music is preferred the most.
3. To identify the percentage of music listeners while working.
4. To identify the hour spent per day by the respondents to listen music.
5. To identify the which age-group listen to music the most.
6. To determine the most preferred genre of music for different age-groups.
7. To determine the association between Composer and Instrumentalist.
8. To determine the association between Exploratory and Foreign language music listeners.
9. To compare mean BPM of different genre of music.
10. To determine the severity of various mental illness.
11. To identify the effect of music on mental illness.

1.4 Scope of the Survey

As today's generation is suffering from stress and anxiety, through this survey we can see how effective the music is to deal with these mental health issues. We can understand which age group is suffering the most and how their mental health has been improved after listening to various genre of music. Through this survey we can see how music can be used as an alternative to therapy and other medical treatment for mental health issues as is it cost effective.

2 Chapter 2

Methodology

2.1 Materials and Methods

A secondary data has been collected from the website of "Kaggle". The dataset contained features like age, primary streaming service, hours per day, while working, instrumentalist, composer, fav genre, exploratory, foreign language, BPM, anxiety, depression, insomnia, OCD etc of the music listeners. The data consists of 31 columns and 610 rows.

2.2 About the data

- **Age :** It denotes the respondent's age and the range is between 10 to 89. The measurement are given in years.
- **Primary streaming service :** It stands for respondent's primary streaming service in which he listen's music.
- **Hours per day :** It denotes the number of hours the respondent listen to music per day.
- **While working :** It denotes whether the respondent listen to music while working or not.
- **Instrumentalist :** It denotes whether the respondent plays instrument or not.
- **Composer :** It denotes whether the respondent compose music or not.
- **Fav genre :** It stands for the respondent's favourite genre of music.
- **Exploratory :** It denotes whether the respondent actively explore new artists or not.
- **Foreign language :** It denotes whether the respondent listen to music in a language he/she are not fluent in or not.
- **BPM :** It stands for the heart rate of the respondent while listening to their favourite genre of music.
- **Frequency [Classical] :** It denotes how frequently the respondent listen to classical music.
- **Frequency [Country] :** It denotes how frequently the respondent listen to classical music.

- **Frequency [EDM] :** It denotes how frequently the respondent listen to EDM music.
- **Frequency [Folk] :** It denotes how frequently the respondent listen to folk music.
- **Frequency [Gospel] :** It denotes how frequently the respondent listen to gospel music.
- **Frequency [Hip hop] :** It denotes how frequently the respondent listen to hip hop music.
- **Frequency [Jazz] :** It denotes how frequently the respondent listen to Jazz music.
- **Frequency [K pop] :** It denotes how frequently the respondent listen to K pop music.
- **Frequency [Latin] :** It denotes how frequently the respondent listen to latin music.
- **Frequency [Lofi] :** It denotes how frequently the respondent listen to lofi music.
- **Frequency [Metal] :** It denotes how frequently the respondent listen to metal music.
- **Frequency [Pop] :** It denotes how frequently the respondent listen to pop music.
- **Frequency [R&B] :** It denotes how frequently the respondent listen to R&B music.
- **Frequency [Rap] :** It denotes how frequently the respondent listen to rap music.
- **Frequency [Rock] :** It denotes how frequently the respondent listen to rock music.
- **Frequency [Video game music] :** It denotes how frequently the respondent listen to video game music.
- **Anxiety :** It denotes the self-reported anxiety, on a scale of 0 to 10.
- **Depression :** It denotes the self-reported depression, on a scale of 0 to 10.
- **Insomnia :** It denotes the self-reported insomnia, on a scale of 0 to 10
- **OCD :** It denotes the self-reported OCD, on a scale of 0 to 10
- **Music effect :** It denotes the effect of music on respondent's mental health condition.

2.3 Statistical Technique used for Data Analysis

The programming language 'Python' and open source software 'R' has been used to carry out the analysis of the data. The statistical methods considered in order to carry out the analysis are as follows :

2.3.1 Bar Chart

A bar chart is a way of summarizing a set of categorical data (continuous data can be made categorical by auto-binning). The bar chart displays data using a number of bars, each bar representing a particular category. The height of each bar is proportional to a specific aggregation (for example the sum of the values in the category it represents). The categories could be something like an age-group or geographical location. It is also possible to colour or split each bar into another categorical column in the data, which enables you to see the contribution from different categories to each bar or group of bars in the bar chart.

2.3.2 Pie Chart :

A pie chart is a type of graph that represents the data in the circular graph. The slices of pie show the relative size of the data, and it is a type of pictorial representation od data. A pie chart requires a list of categorical variables and numerical variables. Here the term 'pie' represents the whole and the 'slices' represents the parts of the whole. To find out the composition of something, Pie-chart works the best at that time.

2.3.3 Histogram :

A histogram is a graphical representation of a grouped frequency distribution with continuous classes. It is an area diagram and can be defined as a set of rectangles with bases along with the intervals between class boundaries and with areas proportional to frequencies in the corresponding classes. In such representations, all the rectangles are adjacent since the base covers the intervals between class boundaries.

The heights of rectangles are proportional to corresponding frequencies of similar classes and for different classes, the heights will be proportional to corresponding frequency densities.

2.3.4 Box Plot :

A box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. In addition to the box on a box plot, there can be extending from the box indicating variability outside the upper and lower quartiles. Outliers that differ significantly from the rest of the dataset may be plotted as individual points beyond the whiskers on the box-plot. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacings in each subsection of the box-plot indicate the degree of dispersion and skewness of the data, which are usually described using the five-number summary. In addition, the box-plot allows one to visually estimate various L-estimators, notably the interquartile range, midhinge, range, mid-range, and trimean. Box plots can be drawn either horizontally or vertically.

2.3.5 Frequency Curve :

Frequency curve is a graph of frequency distribution where the line is smooth. It is just like a frequency polygon. In the polygon the line is straight, but in the curve the line is smooth. Frequency curve is an area diagram. The X-axis is marked with class intervals. The Y-axis is marked with frequencies. The histogram is drawn. The midpoint of the top of the rectangles are joined by a smooth line. The beginning and end of the curve should touch the X-axis at the mid points of first and last class interval. The area of the curve is equal to that of histogram. The frequency curve is divided into 3 types based on the shape of the curve. They are Normal distribution curve, Positively skewed distribution curve and Negatively skewed distribution curve. The Normal distribution curve is asymmetrical and has an inverted bell shape. Positively skewed distribution curve is asymmetrical.

The low values of the variables have the highest frequencies. Negatively skewed distribution curve is also asymmetrical. High values of the variables have the highest frequencies.

2.3.6 Chi – square test of Independence :

A chi-square test of independence used to determine whether two categorical variables are related. If two variables are related, the probability of one variable having a certain value is dependent on the value of the other variable. The test compares the observed frequencies to the frequencies you would expect if the two variables are unrelated. When the variables are unrelated, the observed and expected frequencies will be similar.

When you want to perform a chi-square test of independence, the best way to organize your data is a type of frequency distribution table called a contingency table. A contingency table, also known as a cross tabulation or crosstab, shows the number of observations in each combination of groups. It also usually includes row and column totals.

The contingency table is used to calculate the expected frequencies (E) :

$$\frac{(\text{Row } r \text{ total} \times \text{Column } c \text{ total})}{\text{Grand total}}$$

The chi-square test of independence evaluates null hypothesis (H_0) that variable 1 and variable 2 are not related in the population vs. the alternative hypothesis (H_1) that variable 1 and variable 2 are related in the population.

Pearson's chi-square formula is used to calculate the test statistic:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

where X^2 is the chi-square test statistic, O is the observed frequency and E is the expected frequency. If the X^2 value is greater than the critical value, then the difference between the observed and expected distributions is statistically significant. The data allows you to reject the null hypothesis that

the variables are unrelated and provides support for the alternative hypothesis that the variables are related.

2.3.7 One-way Analysis of Variance (ANOVA) :

A key statistical test in research fields including biology, economics and psychology, Analysis of Variance (ANOVA) is very useful for analysing datasets. It allows comparisons to be made between three or more group of data. The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

One-way ANOVA is a statistical method to test the null hypothesis (H_o) that three or more population means are equal vs. the alternative hypothesis (H_1) that at least one mean is different.

ANOVA determines whether the groups created by the levels of the independent variable are statistically different by calculating whether the means of the treatment levels are different from the overall mean of the dependent variable.

If any of the group means is significantly different from the overall mean, then the null hypothesis id rejected.

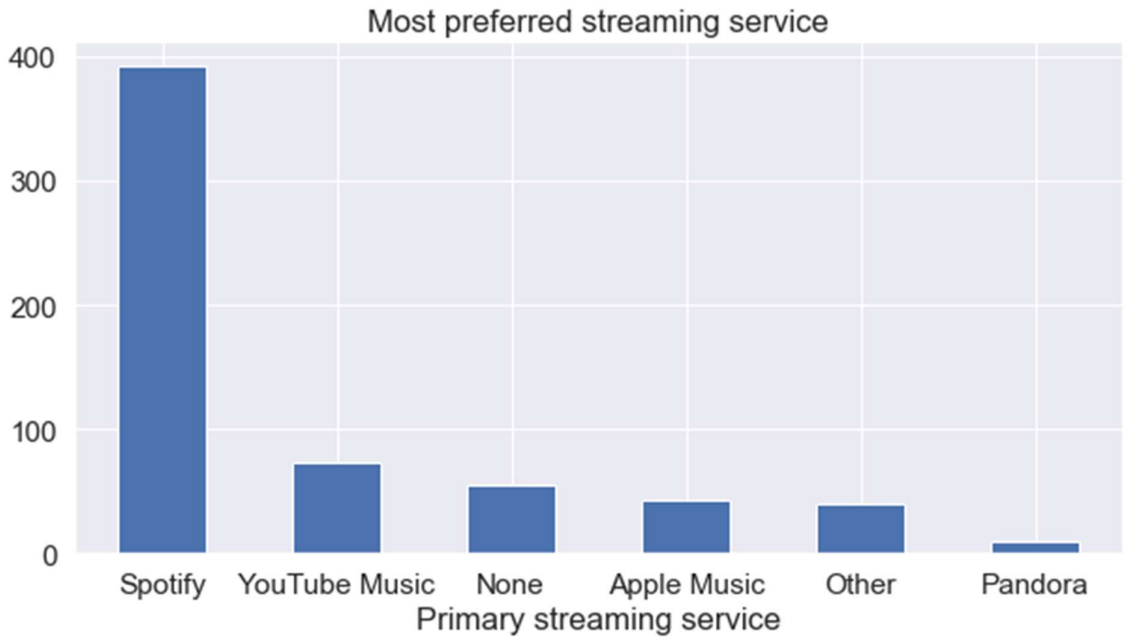
ANOVA uses the F-test for statistical significance. This allows for comparison of multiple means at once, because the error is calculated for the whole set of comparisons rather than for each individual two way comparison (which would happen with t-test).

The F-test compares the variance in each group mean from the overall group variance. If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-value, and therefore a higher likelihood that the difference observed is real and not due to chance.

3 Chapter 3

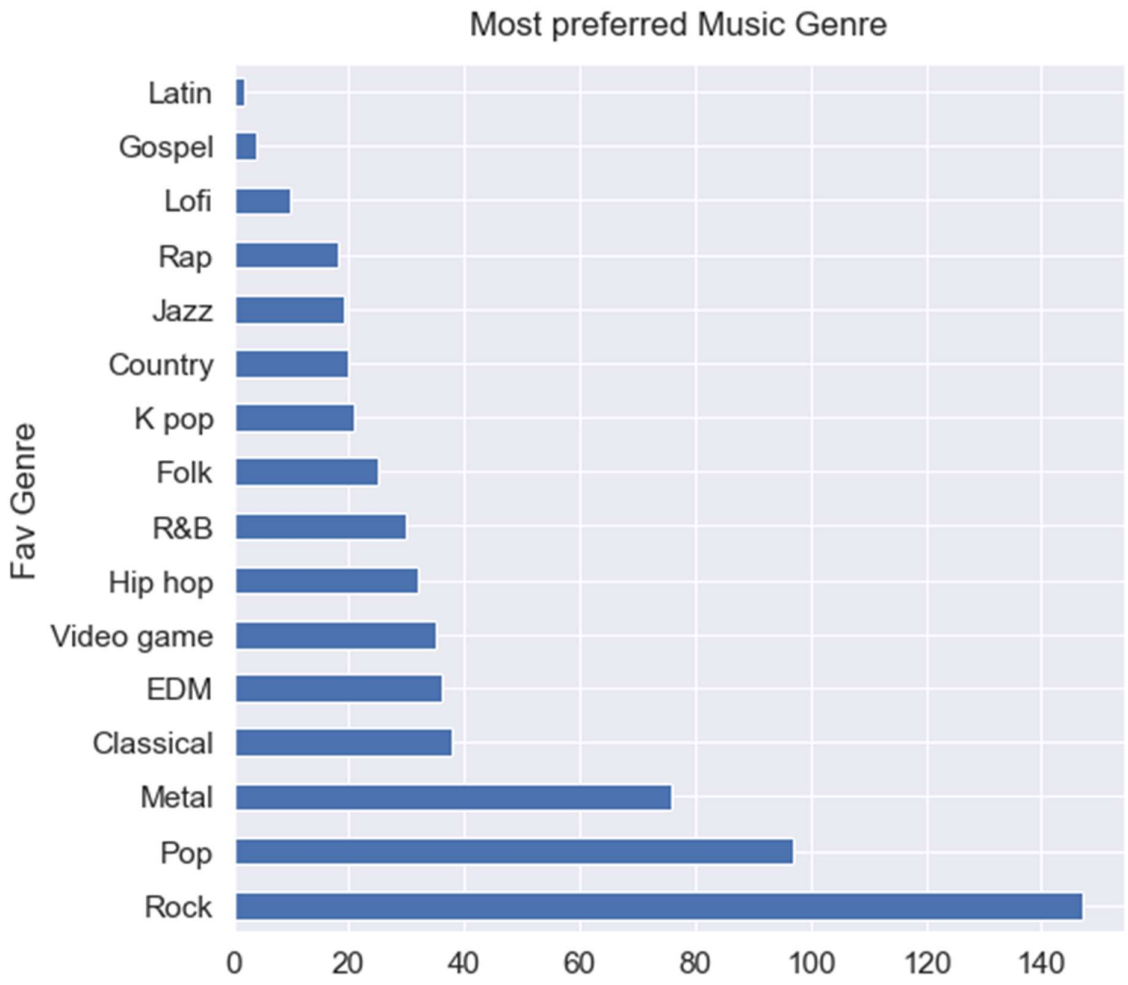
Results and Discussion

3.1 Analysis on the most used streaming service :



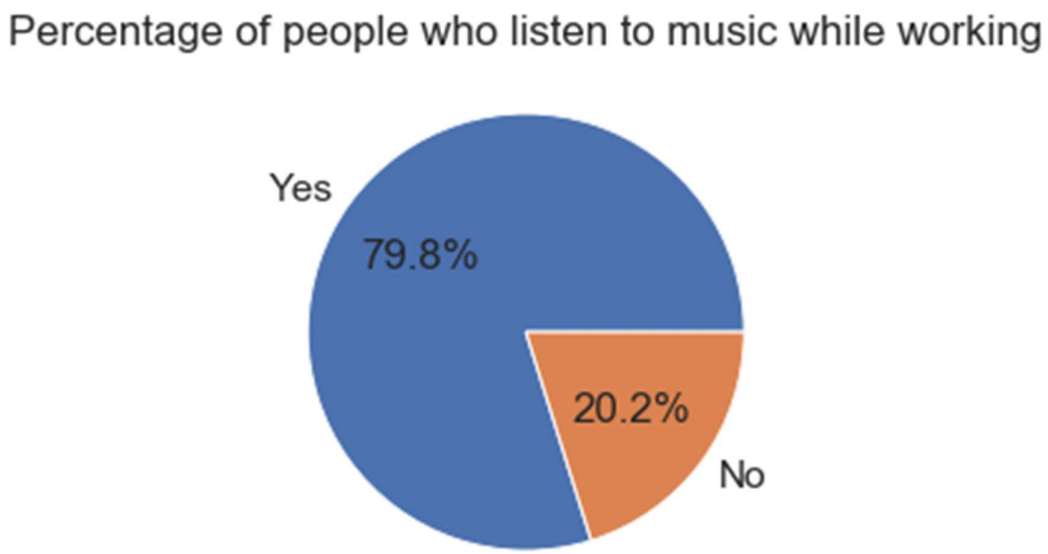
Here we can observe highest used streaming service is Spotify and Pandora is the least used streaming service.

3.2 Analysis on the most preferred Genre of music :



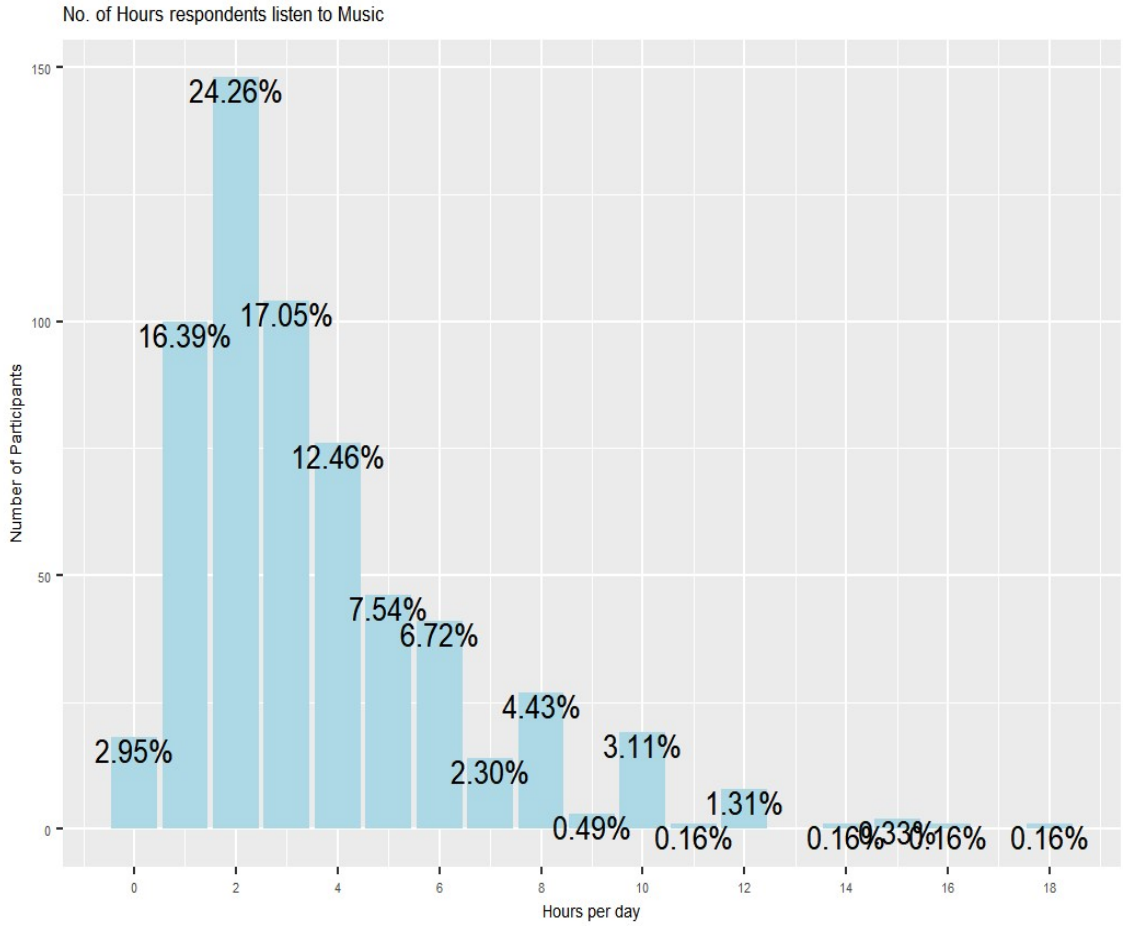
Here we can observe that highest preference is given to Rock music. Then listeners also prefer Pop and Metal music. Amongst all the other genre Latin and Gospel are the least preferred genre of music.

3.3 Analysis on music listener's while working :



Here we can observe that maximum percentage of people which is 79.8% listen to music while they are working and remaining 20.2% listen to music in their free time.

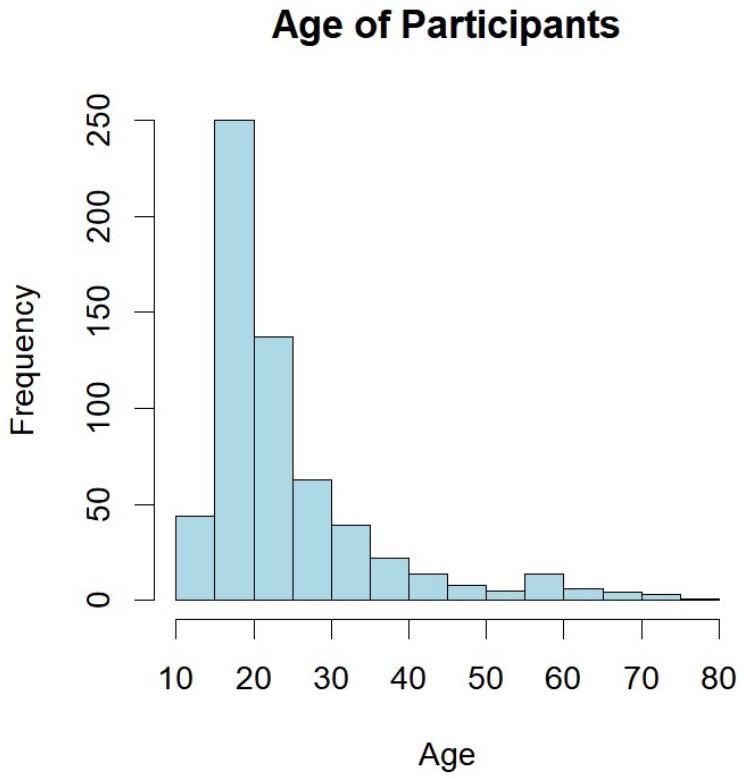
3.4 Analysis on hours spent by the respondents to listen music :



Here we can observe that maximum hours spent on listening to the music is between 1 to 3 hours. That around 57.7% respondents spend 1 to 3 hours each day to listen music.

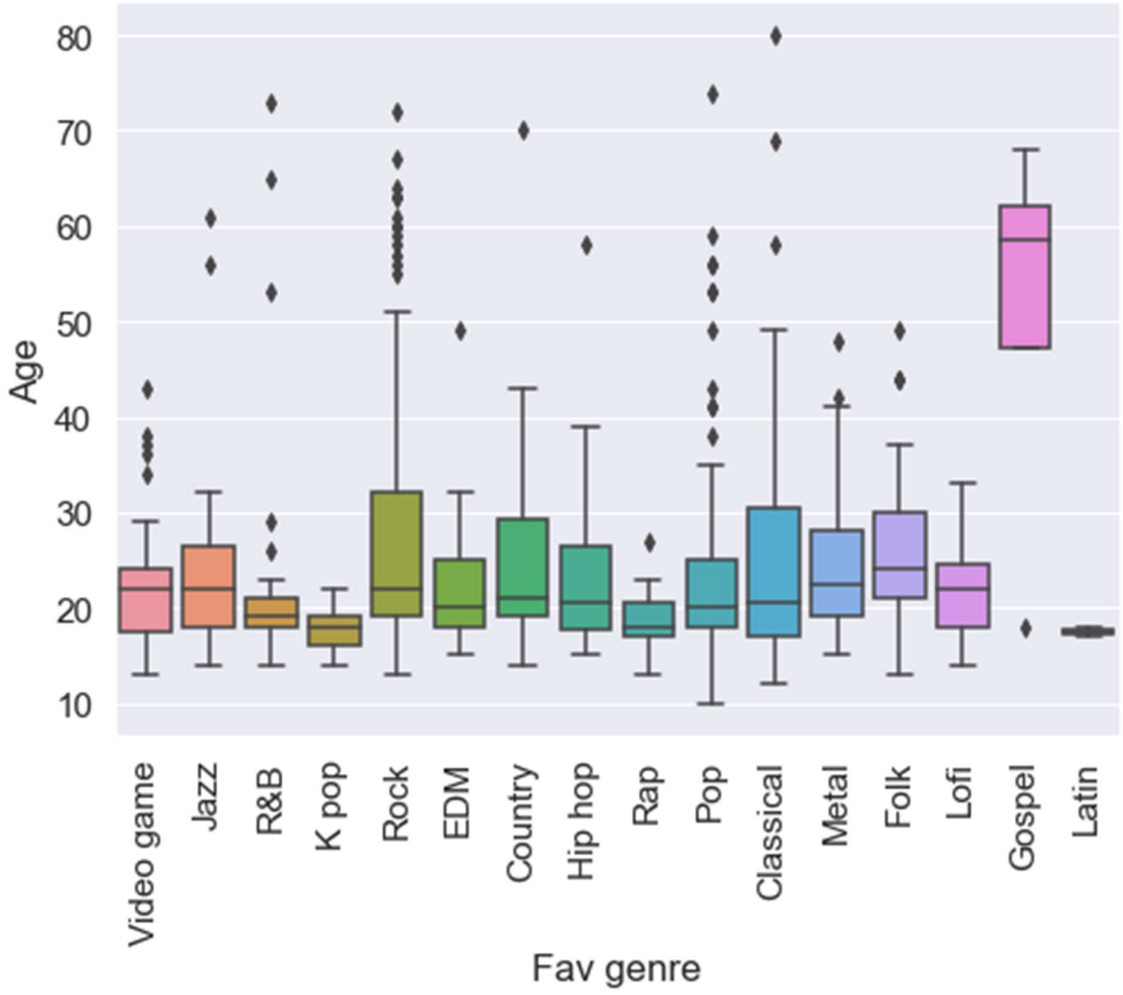
3.5 Analysis on the which age group listen to music the most :

Age group	Age	Total
Children	9 - 16	81
Young Adults	17 - 30	413
Middle-aged Adults	30 - 60	102
Old Adults	60 +	14



Here we can observe that most of the music listeners are between the age of 15 to 25, which is the age-group of young adults.

3.6 Analysis on the preferred genre of music for different age groups :



Here we can observe that Gospel music is listened by the people who are aged above 50. And Latin and K pop music is listened by the people below the age of 20. Other genre is listened by the people between the age group 15 to 30 with outliers indicating some of the listener are of other age-group.

3.7 To test association between listeners who are Instrumentalist and Composer using Chi - square test of Independence :

The hypothesis are as follows :

H_0: The variable instrumentalist and composer are independent of each other.

H_1: The variable instrumentalist and composer are dependent of each other.

The contingency table is :

Composer / Instrumentalist	Composer	Non Composer	Total
Instrumentalist	82	116	198
Non Instrumentalist	25	387	412
Total	107	503	610

We obtain p-value as 2.0716144371238075e-26.

Conclusion :

Here we can observe that the obtained p-value is less than 0.05. Hence we reject H_0.

Therefore we can conclude that variable instrumentalist and composer are dependent on each other.

Therefore there is a significant relationship between listener who are instrumentalist and composer.

3.8 To test association between listener who listen to Exploratory and Foreign language music using Chi - square test of Independence :

The hypothesis are as follows :

H_0: The variable exploratory and foreign language are independent of each other.

H_1: The variable exploratory and foreign language are dependent of each other.

The contingency table is :

Foreign Language / Exploratory	Foreign Language	Non Foreign Language	Total
Exploratory	279	170	449
Non Exploratory	65	96	161
Total	344	266	610

We obtain p-value as 2.79415534275113e-06.

Conclusion :

Here we can observe that the obtained p-value is less than 0.05. Hence we reject H_0.

Therefore we can conclude that variable exploratory and foreign language are dependent on each other.

Therefore there is a significant relationship between listener who listen to exploratory and foreign language music.

3.9 Univariate Analysis of factor mean BPM of different Music Genre using One-Way Analysis of Variance :

The mean BPM of respondents are classified according to the Music genre.

The hypothesis are as follows :

H_o : There is no significant difference between mean BPM of music genres.

H_1 : There is a significant difference between the mean BPM of music genres.

The ANOVA table is obtained as follows :

Source	D.o.f	SSq	MSSq	F-obs	P-value
Genre	6	39224	6537	6.147	3.27e-06
Residuals	454	482804	1063	-	-
Total	460	522028	-	-	-

Where D.o.f is degrees of freedom, SSq is sum of squares, MSSq is mean sum of squares and F-obs is calculated value of F-statistic.

Conclusion :

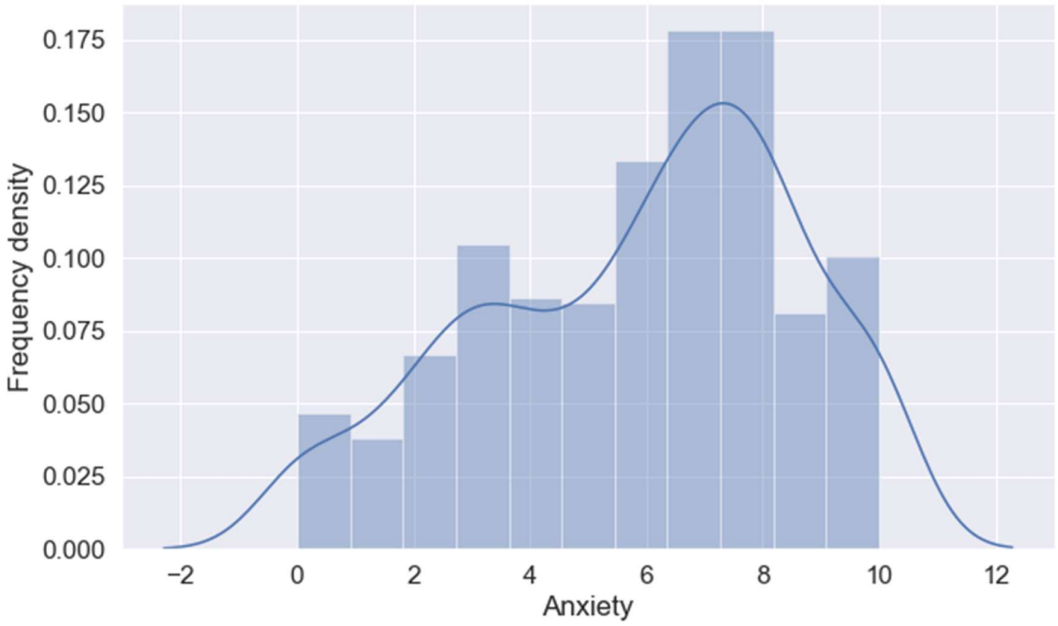
Here we observe that the obtained p-value is less than 0.05. Hence we reject H_o.

Therefore can conclude that there is a significant difference between the mean BPM of music genres (Rock, Pop, Metal, Classical, EDM, Video game music, Hiphop)

From post hoc Tukey test we can conclude that metal-classical, hiphop-edm, metal-hiphop, pop-metal, rock-metal are the music genre whose mean BPM have significant difference.

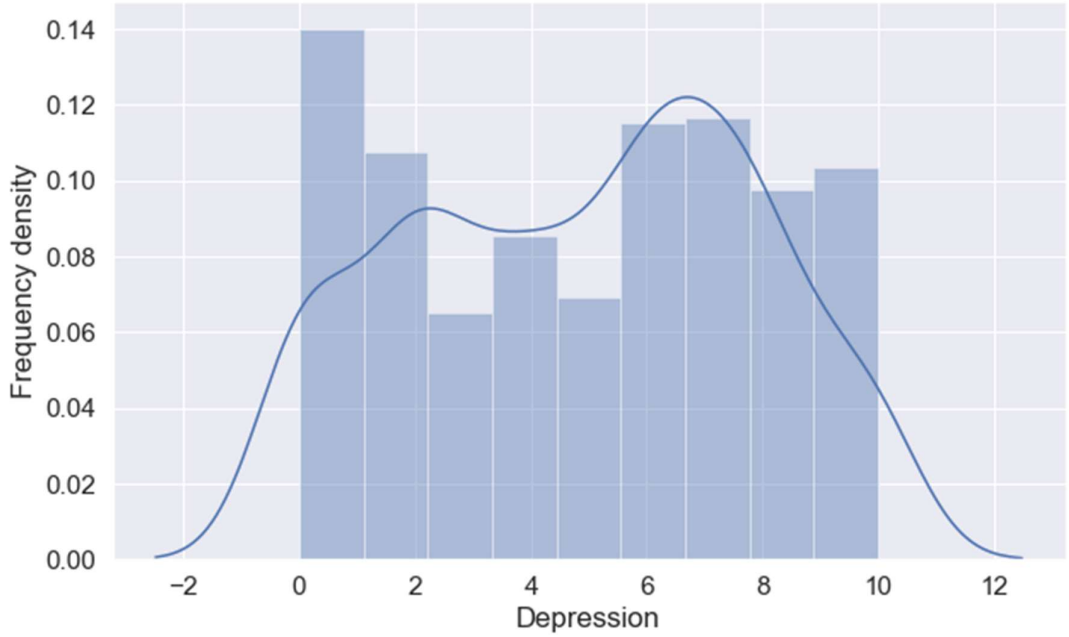
3.10 Analysis of respondents Mental health distribution :

Anxiety :



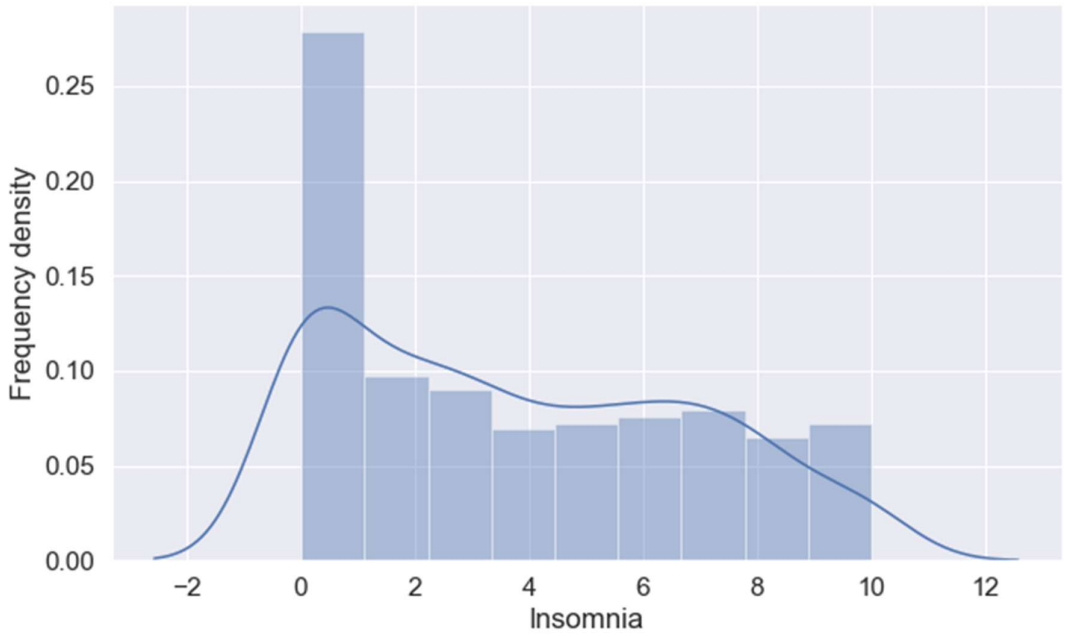
Here we can observe that Anxiety is not reduced after listening to music.

Depression :



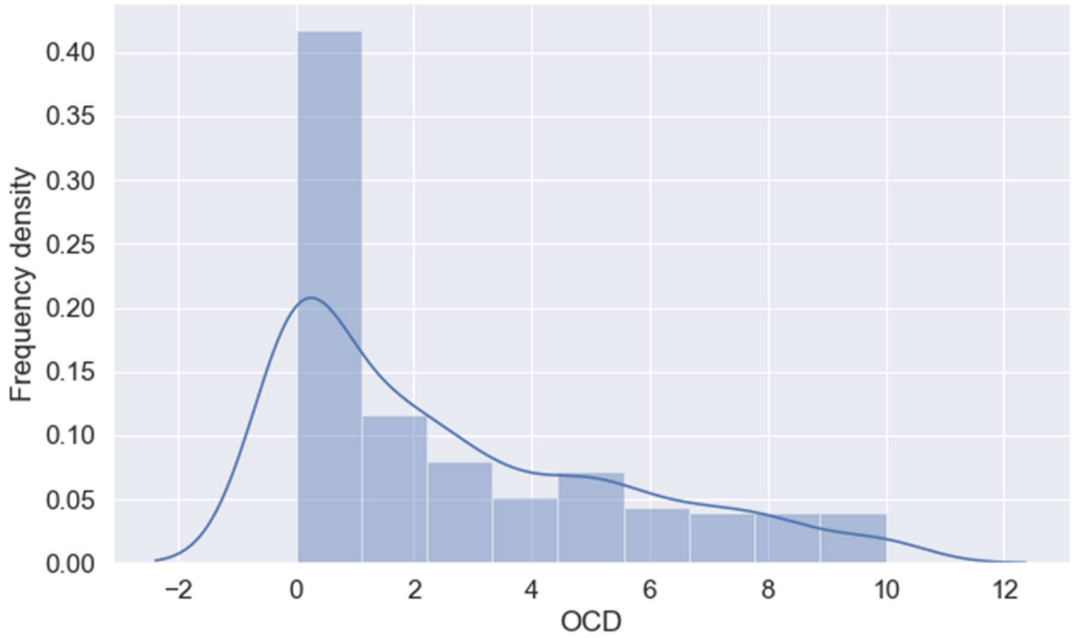
Here we can observe that Depression is moderate after listening to music.

Insomnia :



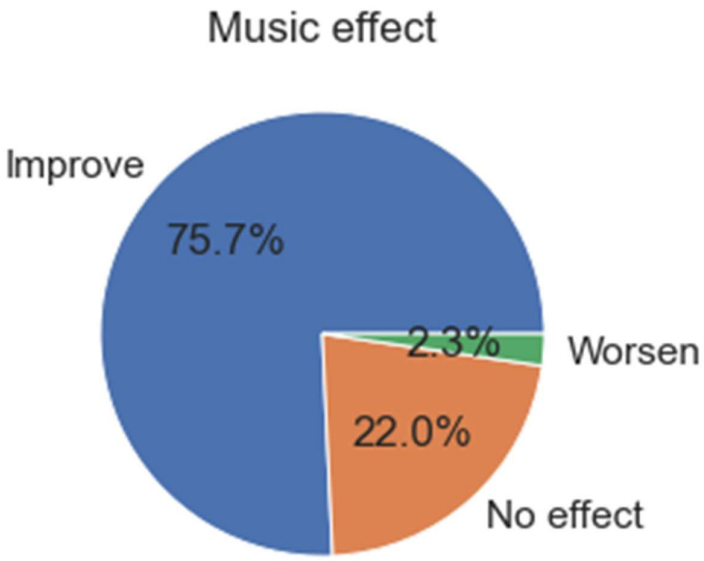
Here we can observe that Insomnia has been reduced after listening to music.

OCD :



Here we can observe that OCD has been reduced after listening to music.

3.11 Analysis on the effect of music on the Mental health :



Here we can observe that according to 75.7% there is improvement in their mental health condition after listening to music, for 22% listeners music did not help in anyway. And for 2.3% listners their condition got worse after listening to music.

4 Chapter 4

Conclusion

Following are the overall conclusions,

1. Spotify is the highest used streaming service and Pandora is the least used streaming service.
2. Rock, Pop and Metal music are widely preferred music's and Gospel and Latin are the least preferred music.
3. Almost 80% of people listen to music while they are working.
4. On the daily basis maximum number of people spend 1 to 3 hours each day to listen music.
5. The maximum music listeners belong to young-adult age group which is between 17 to 30.
6. People above the age of 50 years prefer listening to the Gospel music where people less than 20 years of age listen to Latin and K pop music.
7. From the Analysis we can see that there is a significant relationship between listener who are instrumentalist and composer.
8. From the Analysis we can see that there is a significant relationship between the listener who listen to exploratory music and foreign language music.
9. From the Analysis we can see that there is a significant difference between the mean BPM of music genres.
10. From the graph's we can observe that insomnia and OCD are not experienced by the music listeners while Anxiety is some what severely experienced and depression being moderate among the listeners.
11. Around 76% find music improving their mental illness, 22% find no effect in the music and around 2% find their condition getting worse after listening to music.

5 Chapter 5

Summary

A project entitled “An Exploratory Analysis on Music and Mental Health” has been done. A secondary data has been collected from the website of “Kaggle”. The data contained features like age, primary streaming service, categories of different genre of music, mental health conditions etc. The data consists of 31 columns and 610 records.

The analysis and interpretation of the data is done by using some of the statistical methods like graphical method, one-way analysis of variance and chi-square test of independence.

From the results obtained, we came to know that there is a significant relationship between the instrumentalist and composer, and also there is a significant relationship between the listeners who listen to exploratory and foreign language music. Also we came to know that there is a significant difference between mean BPM while listening to various genre of music.

6 Chapter 6

Bibliography

1. https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results
2. https://www.hindawi.com/journals/jeph/2022/1120156/
3. https://diposit.ub.edu/dspace/bitstream/2445/189148/1/TFM-ECO_Roman_2022.pdf
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8566759/
5. https://www.freecodecamp.org/news/drop-list-of-rows-from-pandas-dataframe/5.
6. https://www-geeksforgeeks-org.cdn.ampproject.org/v/s/www.geeksforgeeks.org/how-to-create-frequency-tables-in-python/
7. https://www.statology.org/anova-unequal-sample-size/
8. https://www.geeksforgeeks.org/how-to-extract-a-column-from-r-dataframe-to-a-list/
9. https://www.geeksforgeeks.org/anova-test-in-r-programming/

7 Chapter 7

Appendix

Python code for Analysis

7.1 Bar chart of streaming service

```
sns.set(font_scale = 1.4)
data['Primary streaming service'].value_counts().plot(kind = 'bar', figsize = (10,5), rot = 0)
plt.xlabel('Primary streaming service')
plt.title('Most preferred streaming service');
```

7.2 Pie chart of music listener's while working

```
total = data['While working'].value_counts().values.sum()
def fmt(x):
    return '{:.1f}%'.format(x, total*x/100)
data['While working'].value_counts().plot(kind='pie',autopct=fmt)
plt.ylabel('')
plt.title("Percentage of people who listen to music while working")
```

7.3 Frequency curve of mental health distribution of respondents

```
plt.figure(figsize=(10,6))
plt.xlabel('Anxiety')
plt.ylabel('Frequency density')
sns.distplot(a=data['Anxiety'],hist=True)
```

7.4 Chi-square test of independence

To test the association between the listeners who are instrumentalist and composer.

```
data_crosstab1=pd.crosstab(data['Instrumentalist'],data['Composer'],margins=False)
from scipy.stats import chi2_contingency
stat, p, dof, expected = chi2_contingency(data_crosstab1)
alpha = 0.05
print("p value is " + str(p))
if p <= alpha:
    print("Dependent (reject Ho)")
```

```
else:
   print("Independent (We do not reject Ho)")
```

7.5 Boxplot for preferred genre of music for different age groups

```
plt.figure(figsize=(8,6))
sns.boxplot(x=data['Fav genre'],y=data['Age'])
plt.xticks(rotation=90)
```

R code for Analysis

7.6 Histogram for age group of music listeners

```
hist(data$Age, main="Age of Participants", xlab="Age",
col="LightBlue")
```

7.7 One-way ANOVA

To compare the mean BPM of various music genre.

```
Anova=aov(bpm~genre)
Anova
#summary of the test
summary(Anova)
#Post hoc test
TukeyHSD(Anova)
```