The relationship between music and depression

2025-11-25

Aasim Ditta
stu#: 1008339246

Hamza Rana
stu#: 1008007988

Nart Machfj
stu#: 1010640834

Abbas Peermohammed
stu#: 1008337499

Samar Chandora
stu#: 1010470358

Abstract

The purpose of this report is to identify if any correlations exist between an individual's music taste and their self-reported depression level. The dataset we used is called: "Music & Mental Health Survey Results". This dataset contains observations from several hundred participants, recording a lot of information about the individual and their music taste. Some notable ones being: their age, favorite genre of music, number of hours they spend listening to music per day and their depression level. In this report we will be taking looking at specific aspects of an individual and compare it to their depression level to see if any relationship/correlation exists.

Introduction

As mentioned before, the dataset we used is called "Music & Mental Health Survey Results", which we found on kaggle. The dataset has a wide variety of subjects, so we could not focus on all of them. The ones chosen to be analyzed for this report were: An individual's age, an individual's favorite genre of music, the number of hours an individual spends listening to music per day, whether or not an individual listens to music while studying/working, whether or not the individual plays an instrument or not, and the individual's depression level.

We could not find when, where and how this dataset was collected.

The purpose of the dataset was identify what, if any, correlations exist between an individual's music taste and their self-reported mental health.

Research Questions:

1. How does age relate to depression levels among participants, and do younger and older age groups differ in their reported depression?

2. Is there a genre of music that is associated with higher levels of depression among listeners?

3. How does the amount of hours spent listening to music per day affect an individual's depression level?

4. Does listening to music while working report better or worse depression?

5. Does playing a musical instrument affect depression levels among individuals.

Variable Codebook

- **Age**
  - Type: Numeric / Continuous
  - Original Values: Whole numbers (10-89)
  - Recoding / Cleaning: Converted to integer; used to create AgeGroup
  - Missing Values: Removed NA entries
- **AgeGroup**
  - Type: Categorical
  - Original Values: N/A
  - Recoding / Cleaning: Derived from Age: <21 -> 'Under 21', >=21 -> '21 and Over'
  - Missing Values: N/A (derived from cleaned Age)
- **Depression**
  - Type: Numeric / Continuous
  - Original Values: 0-10
  - Recoding / Cleaning: Converted to integer
  - Missing Values: NA ignored in calculations
- **Fav_Genre**
  - Type: Categorical
  - Original Values: 16 possible genres
  - Recoding / Cleaning: Trimmed whitespace; dummy variable for Rock created
  - Missing Values: NA ignored in genre-specific analysis
- **Hours per day**
  - Type: Numeric / Continuous
  - Original Values: 0-24 hours
  - Recoding / Cleaning: Used as-is for regression and plots
  - Missing Values: NA ignored in calculations
- **While.working**
  - Type: Binary / Categorical
  - Original Values: Yes / No / Y / N / blanks
  - Recoding / Cleaning: Text standardized; yes/y -> 1, no/n -> 0
  - Missing Values: Removed blank or NA entries
- **Instrumentalist**
  - Type: Binary / Categorical
  - Original Values: Yes / No / blanks
  - Recoding / Cleaning: Trimmed whitespace; filtered to Yes/No
  - Missing Values: Removed NA or invalid entries

Question 1: How does age relate to depression levels among participants, and do younger and older age groups differ in their reported depression?

Data Frame Columns

```
##  [1] "Timestamp"                    "Age"
##  [3] "Primary.streaming.service"    "Hours.per.day"
##  [5] "While.working"                "Instrumentalist"
##  [7] "Composer"                     "Fav.genre"
##  [9] "Exploratory"                  "Foreign.languages"
## [11] "BPM"                          "Frequency..Classical."
## [13] "Frequency..Country."          "Frequency..EDM."
## [15] "Frequency..Folk."             "Frequency..Gospel."
## [17] "Frequency..Hip.hop."          "Frequency..Jazz."
## [19] "Frequency..K.pop."            "Frequency..Latin."
## [21] "Frequency..Lofi."             "Frequency..Metal."
## [23] "Frequency..Pop."              "Frequency..R.B."
## [25] "Frequency..Rap."              "Frequency..Rock."
## [27] "Frequency..Video.game.music." "Anxiety"
## [29] "Depression"                   "Insomnia"
## [31] "OCD"                          "Music.effects"
## [33] "Permissions"
```

A list of all columns were first shown in order to fully understand the dataset. Through looking at the variables, two variables that could possibly be linked are **Age** and **Depression**. Age is a continuous variable ranging across a wide span of different respondents. Depression is a score from 1 to 10, where high numbers signify more severe symptoms of depression.

With these two variables, there are many paths to analayze their relationship. A question that could lead to interesting results is:

"**How does age relate to depression levels among participants, and do younger and older age groups differ in their reported depression**?"

For this scenario, the young group will be considered respondants 21 years of age and younger while the old group will be 21 years of age and older.

Before beginning the analysis, the dataset was loaded and cleaned to remove missing values in the Age column. The Age and Depression variables were both converted to integers so that they could be analyzed properly.

Find Means and Variances of Target Variables

Table 1: Five-number summaries for Age and Depression

Statistic	Age	Depression
Minimum	10	0
Q1	18	2
Median	21	5
Q3	28	7
Maximum	89	10

The five number summaries provide a quick look at the distribution of our variables and give us an initial understanding of the two variables. These include the minimum, first quartile, median, third quartile and maximum.

Looking at the summary for age, we see that it ranges from 10 to 89, giving us meaningful variation for studying age-related trends. The data suggests that 50% of the spread is between 18 years of age and 28 years of age for the age of the respondents.

For depression, the range is obviously from 0 to 10, with a median of 5. the data suggests that 50% of the spread of respondents report a depression score from 2 to 7.

These summaries help us understand the spread and distribution of the variables before peforming futher visual and statistical analyses.

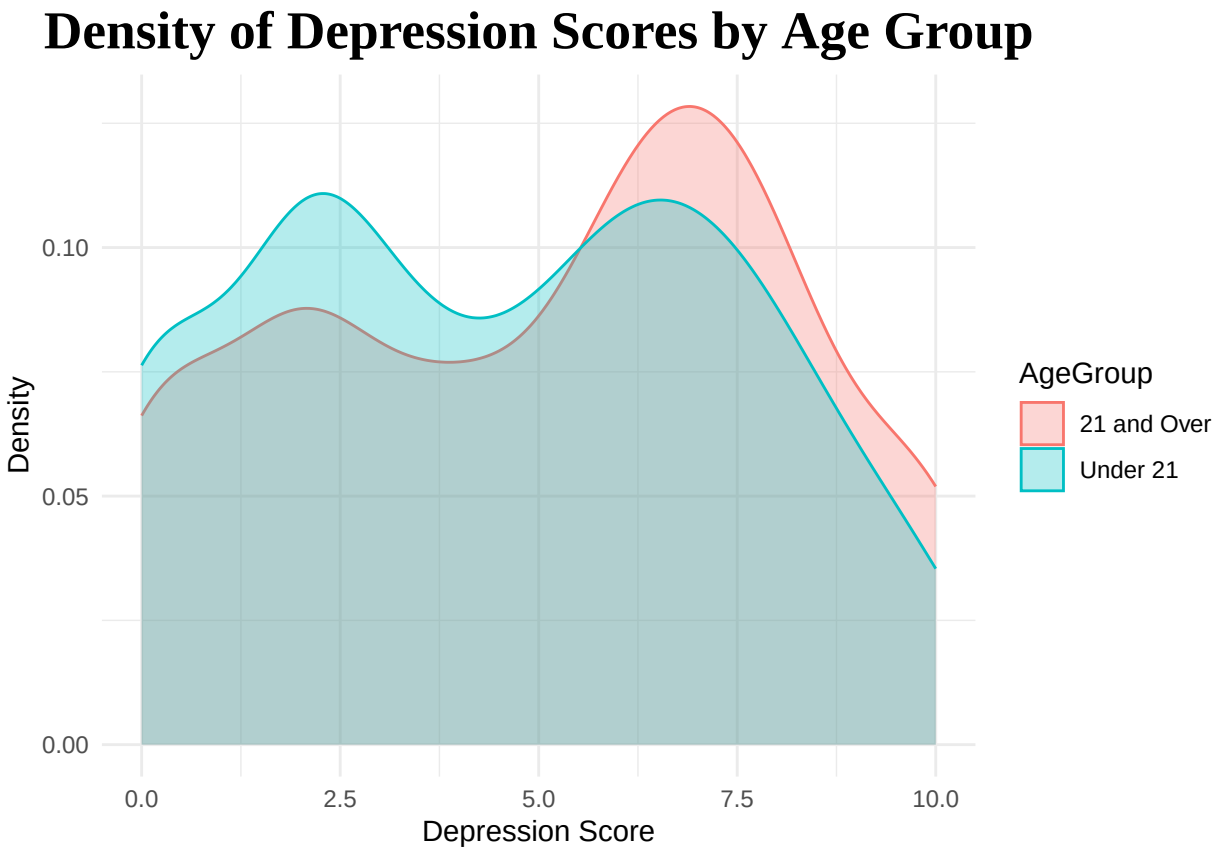

```
## [1] -0.1211921
```

A scatterplot with a regression line was created to visualize the relationship between Age and Depression when it comes to listening to music. As we can see from the graph and the value of the correlation coefficient, the correlation between the two is very weak. With a correlation coefficient of -0.1211921, this shows early signs show that the relationship between the two is uncorrelated.

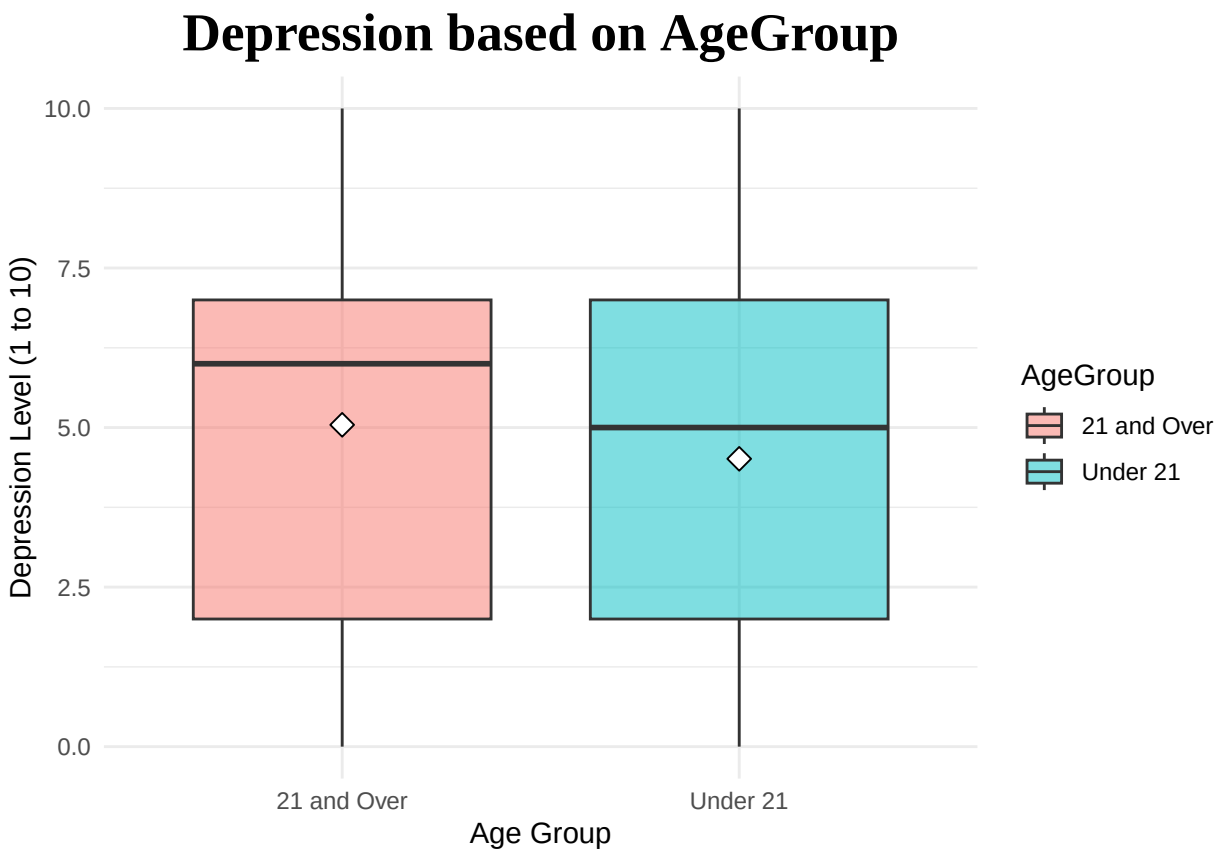
Create Age Group Field

Two age groups were created: **Under 21 and 21 and Over**. The table of counts and percentages shows how many participants are in each group. Looking at the table, we see that the two groups are somewhat balanced, so the choice of the two groups seems reasonable. The two groups also contain enough several hundred people, perfect for performing reliable statistical tests.

AgeGroup	Count	Percent
21 and Over	396	53.87755
Under 21	339	46.12245



The denisty plot shows that the Under 21 group has a slightly higher percentage of respondents on the low depression side, where as the 21 and Over group has a slightly higher percentage of respondents on the higher depression side than the other group. This could possibly lead us to think that older people suffer more depression due to their responsibilites and use music as a support system. Further analysis needs to be made to confirm this.



The boxplot shows that the mean for depression of the 21 and Over group is around 5, where as for the Under 21 group, the mean depression is closer to 4, shown by the white diamond on the boxplots. We see that the median performs similarly, with the older group having a median depression of around 6 and the younger closer to 5. Although the older group performs higher in mean and median, it is surprising to see that the first quartile and third quartile are around the same for both groups.

Check For Equal Variances

Before performing a t-test, we compare variances to decide whether equal-variance assumptions are reasonable. If the variances for the two groups differ noticeably, we use a t-test that does not assume equal variances (Welch's t-test).

```
## [1] 8.854183
```

```
## [1] 9.362709
```

Looking at the variances of the two groups, we see a difference of 0.508526, meaning that we should use the Welch's t-test in order to get the most accurate results.

Peform t-test

A t-test is performed to compare mean depression scores between:

- Under 21
- 21 and Over

Null Hypothesis: Mean of depression among Under 21 is same as Mean of depression among 21 and Over
Alternate Hypothesis: Mean of depression among Under 21 is not the same as Mean of depression among 21 and Over

Display t-test Results

Looking at the results, we can conclude that there is no statistical significant difference between younger and older participants. This conclusion can be made since 0 is not present in the confidence interval

```
## mean in group 21 and Over    mean in group Under 21
##                  5.042929                  4.510324
```

```
## [1] 0.09465402 0.97055560
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 0.0172164
```

```
##        t
## 2.387577
```

Results Summary

Overall, we can see that younger and older people in the dataset report very similar levels of depression, meaning age does not appear to be a big factor in an individual's depression. Together, these analysees suggest that **Age and Depression are statistically independent in the dataset**.

Question 2 Analysis: Is there a genre of music that is associated with higher levels of depression among listeners?

The dataset records an individual's favourite genre of music and their depression level. As stated before, an individual's depression level can range from 0 to 10. Furthermore, an individual can choose from 16 different genres to label as their "favorite genre of music".

So a question that naturally arises is: "Do certain genres of music cause listeners to feel higher levels of depression?"

Music can inflict many different types of emotions on the listener. Feelings of happiness, sadness, excitement, nostalgia, e.t.c. It's only natural that an individual will listen to certain pieces of music, and get feelings of depression. Using this data set we can see if a certain genre is more or less likely to invoke feelings of depression.

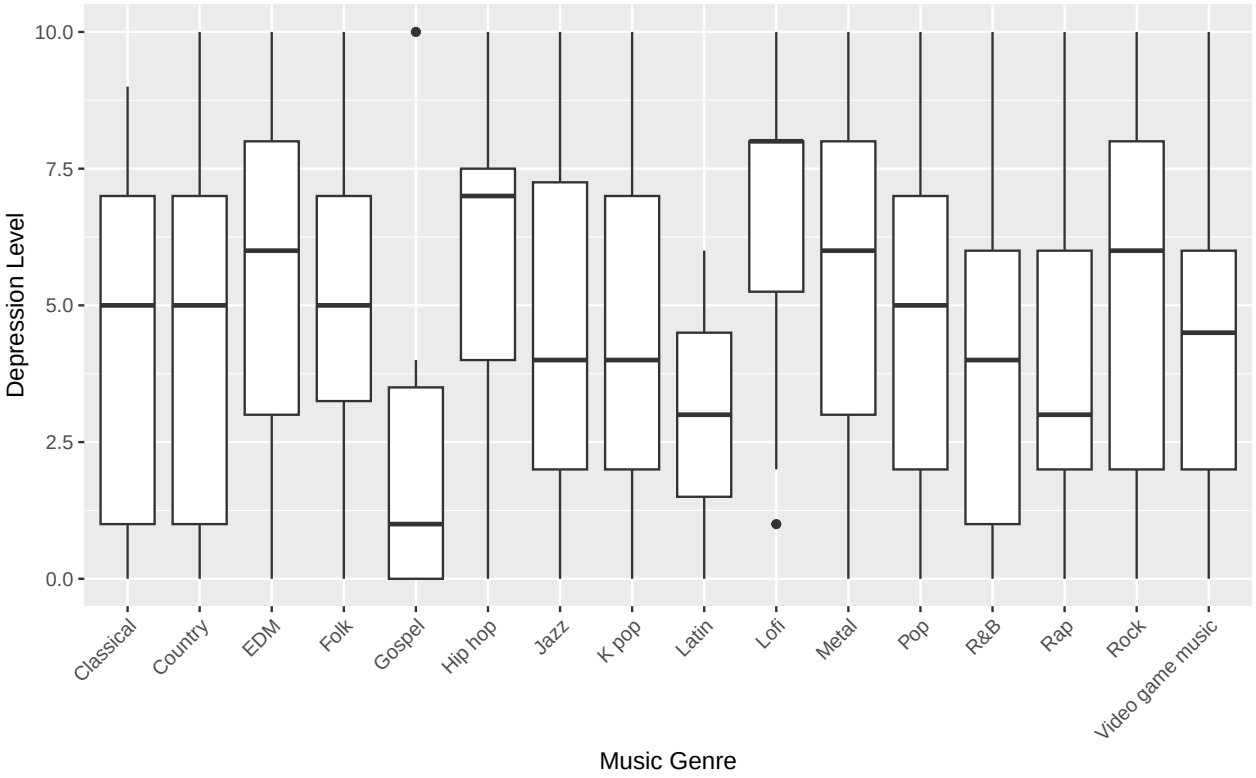
First we need to see what genres of music we have, and the amount of individuals that preferred each genre:

```
##           Fav_Genre Freq
## 1         Classical   53
## 2           Country   25
## 3               EDM   37
## 4              Folk   30
## 5            Gospel    6
## 6           Hip hop   35
## 7              Jazz   20
## 8             K pop   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
```

We can see that the "Rock" genre has a much higher number of enthusiasts than the others, with 188 individuals from the dataset preferring Rock music over any other genre. A large sample size like this can help make a more accurate inference about any relationship that may exist between individuals who prefer Rock music and their depression levels. So our objective should be to analyze the Depression Levels of individuals whose favorite genre of music is Rock in our dataset.

However just finding the number of individuals who prefer each genre is not enough to make an inference, so next we should use boxplots to compare the distribution of depression levels for each genre.

Boxplots:



The boxplot for Rock stands out, as it has a large spread, a relatively central median, and individuals with both low and high depression levels. To analyze the Rock boxplot more thoroughly, we should find its 5 points:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   6.000   5.237   8.000  10.000
```

- Shape: The whiskers on the boxplot are of equal size, with them both equaling 2 points. Since the median is slightly closer to the top of the box, it suggests that there exists a slight amount of left skewness.
- Center: While the median is closer to the top than the bottom, it still is relatively in the center of the box. Since the median is 6, it tells us: 50% of individuals who prefer Rock music have a depression level lower than 6, and 50% of individuals who prefer Rock music have a depression level higher than 6.
- Spread: The middle 50% of depression levels from individuals who prefer Rock music extends across a range of 6 points as they range from a score of 2 (Q1) to 8 (Q3)
- Individually Plotted Points: We can see that there are no individually plotted points, which means there is no outliers.

While the boxplot helps us to understand the sample data, these findings are not enough to make an inference about a real relationship that may exist between an individual preferring Rock Music and their Depression Level.

To make a valid inference, we need to conduct a t-test on the correlation between whether or not an individual favors Rock Music and their Depression Levels.

Correlation test:

To do this, we will use the classical linear regression model, where our explanatory variable will be the Rock Genre, and our response variable will be the depression levels.

The null hypothesis will be: The True Correlation Coefficient equals 0, and the alternative hypothesis will be: The True Correlation Coefficient differs from 0.

To do this we need to set up a "dummy" variable for x. Take individual i: Suppose they prefer rock music, then xi = 1. If they prefer another genre of music, then xi = 0. Then yi is simply the individual's depression level.

```
##
##  Pearson's product-moment correlation
##
## data:  Rock and Depression_Levels
## t = 2.3179, df = 734, p-value = 0.02073
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01305608 0.15654475
## sample estimates:
##        cor
## 0.08524238
```

Our p-value is 0.02, since we are conducting a 95% confidence test, our alpha value is 0.05. Since our alpha value is greater than our p-value, we reject the null hypothesis and conclude that "There exists a relationship between individuals who favor Rock music and their depression levels". However our p-value is not extremely small, so we cant conclude there exists a very strong positive relationship, but we can conclude that there is a statistically significant positive association, as the confidence interval given has a lower limit of 0.013 and upper limit of 0.15.

Conclusion

The Rock genre of music is associated with slightly higher levels of depression among listeners.

Question 3: Is there a relationship between hours spent listening to music per day and depression level?

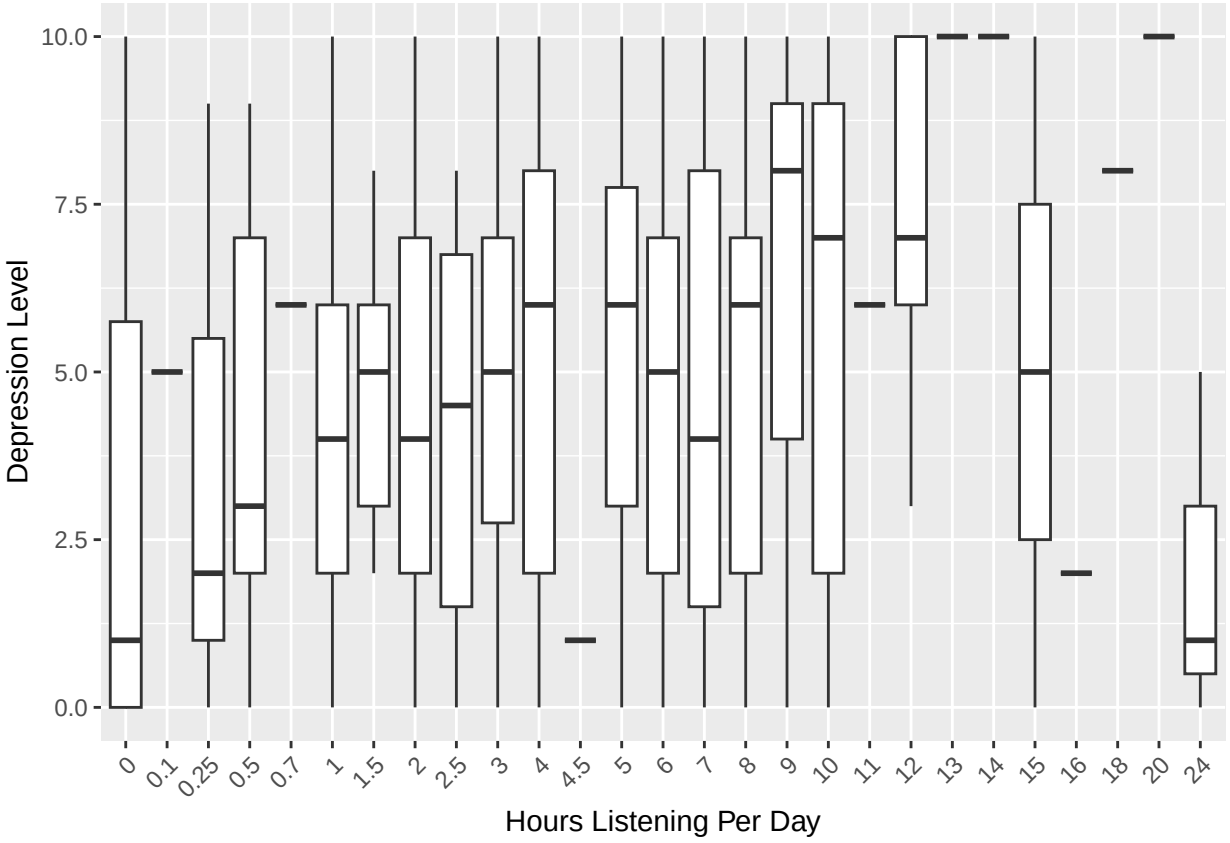
The dataset records how many hours per day each individual listens to music, along with their depression level (ranging from 0 to 10).

A key question is: **Does listening to more music correlate with higher or lower depression levels?**

Music can be used as an escape, and extended time listening may suggest that an individual is more likely to report higher depression levels.

We begin by loading the dataset and displaying the distribution of listening hours:

```
##
##    0  0.1 0.25  0.5  0.7    1  1.5    2  2.5    3    4  4.5    5    6    7    8
##    6    1    3   20    1  117   17  173    6  120   83    1   54   47   15   29
##    9   10   11   12   13   14   15   16   18   20   24
##    3   20    1    9    1    1    2    1    1    1    3
```



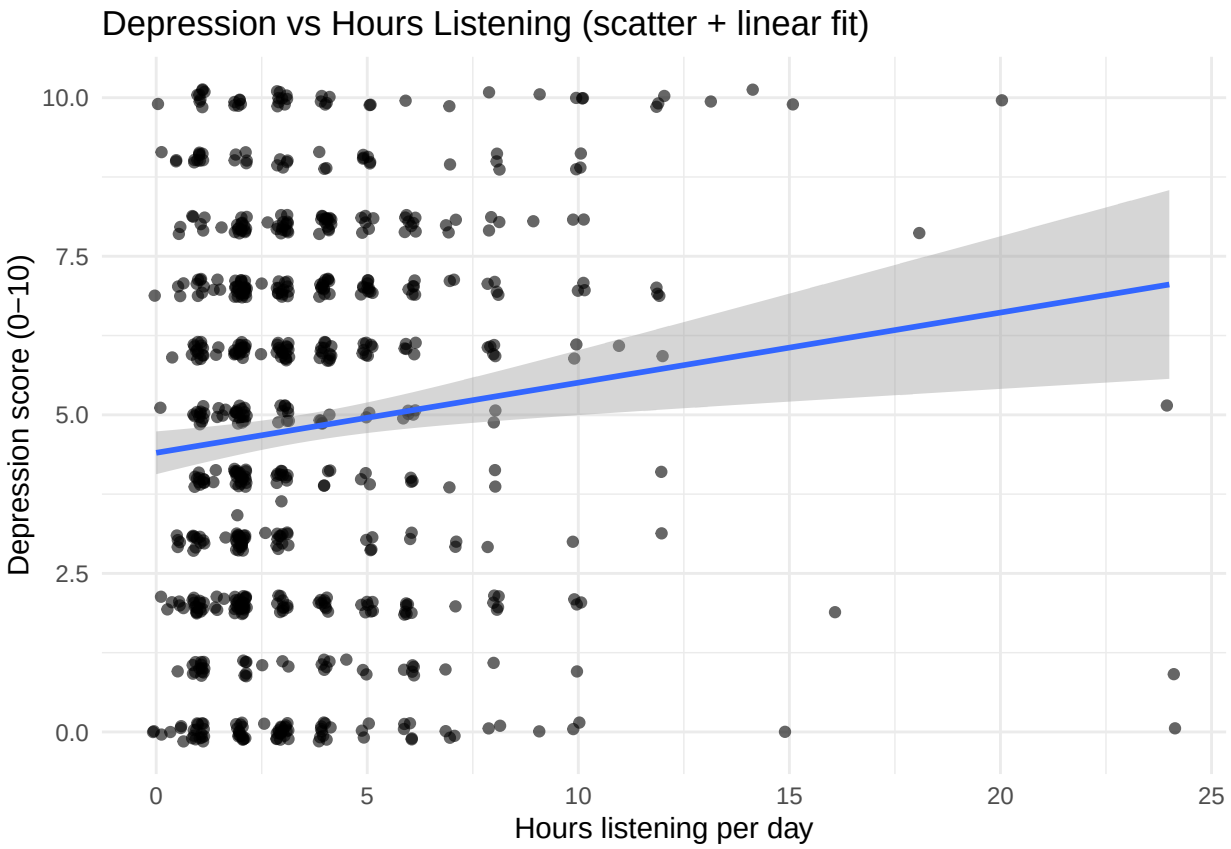
The boxplot shows how depression scores vary across different listening-time groups. This plot only provides an initial visual check for group-level patterns before fitting a regression model. While there may be a slight upward tilt, the visual pattern alone does not indicate a strong relationship.

Correlation test:

To discover an answer to our question, we will use Pearson's product-moment correlation test, where our explanatory variable will be the hours listening, and our response variable will be the depression levels.

The null hypothesis will be: The True Correlation Coefficient equals 0, and the alternative hypothesis will be: The True Correlation Coefficient differs from 0.

```
##
##  Pearson's product-moment correlation
##
## data:  Hours and Depression_Levels
## t = 3.0129, df = 734, p-value = 0.002676
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03856881 0.18134569
## sample estimates:
##       cor
## 0.1105275
```



Analysis and statements

Our p-value is 0.00268. Since we are conducting a 95% confidence test, our alpha value is 0.05. Because our alpha value is greater than our p-value, we reject the null hypothesis and conclude that there is a statistically significant association between individuals time spent listening to music and their reported depression levels. However, our p-value is not extremely small in practical terms, so we cannot conclude there exists a very strong positive relationship.

We can, however, conclude that there is a statistically significant positive association, as the 95% confidence interval for the correlation coefficient is approximately (0.0386, 0.1813). The sample correlation is $r \sim 0.1105$, which indicates a very small positive linear relationship between listening time and depression scores. Because this is observational, cross-sectional data, we do not claim that listening more causes higher depression, as other explanations are possible.

Conclusion

Higher listening time is correlated with slightly higher reported depression levels, but the effect is very small and the association explains only a small proportion of the variability in depression scores.

Question 4: Does listening to music while working report better or worse depression?

Find the Coeffecient of Correlation

```
## [1] 0.05595211
```

Depression and working while studying has a correlation coefficient of 0.05595211. This indicates that the two have little to no linear relationship.

T-Test

We will conduct a two-sample independent t-test comparing depression levels(1-10) between:

- Listens to music while working (Yes)
- Doesn't listen to music while working (No)

A standard t-test would require equal variances between the "yes" and "no" group. The variances for both groups is:

```
## [1] 9.000024
```

```
## [1] 9.606782
```

Since the variances have a difference of 0.606758 a welch's t-test must be conducted

Null Hypothesis: mean of music while working = mean of no music while working

Alternative Hypothesis: mean of music while working =/ mean of no music while working

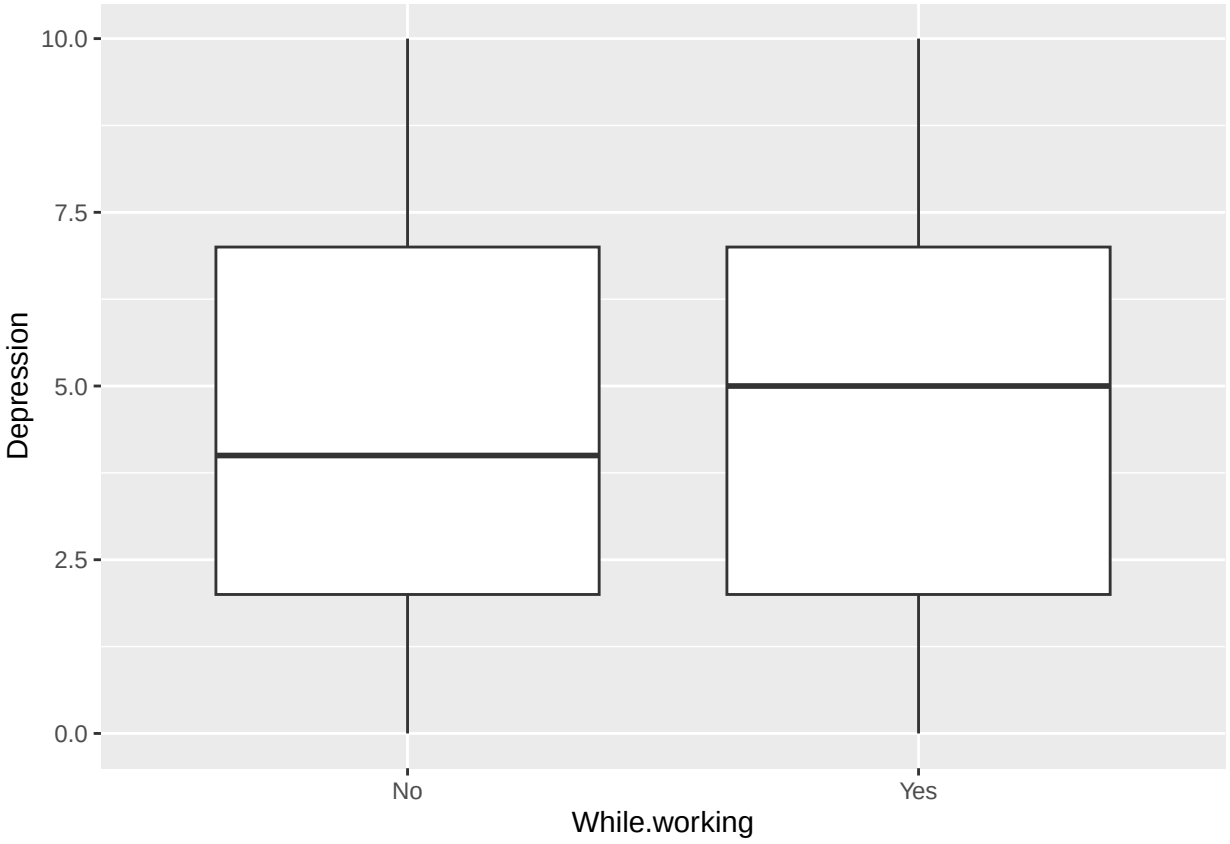
```
##
##  Welch Two Sample t-test
##
## data:  Depression by While.working
## t = -1.4867, df = 234.89, p-value = 0.1384
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.9649850  0.1349377
## sample estimates:
##  mean in group No mean in group Yes
##          4.467532          4.882556
```

This tells us that those who don't listen to music while studying report 4.467532 depression levels on average while those who do report 4.882556 on average.

Although the average depression levels are slightly higher for those who listen to music while studying, the p-value of 0.1385 is greater than 0.05.

Thus we fail to reject the null hypothesis and conclude the data does not provide sufficient evidence that mean depression scores differ between people who listen to music while working and those who do not.

Depression Levels: Music While Working vs. No Music



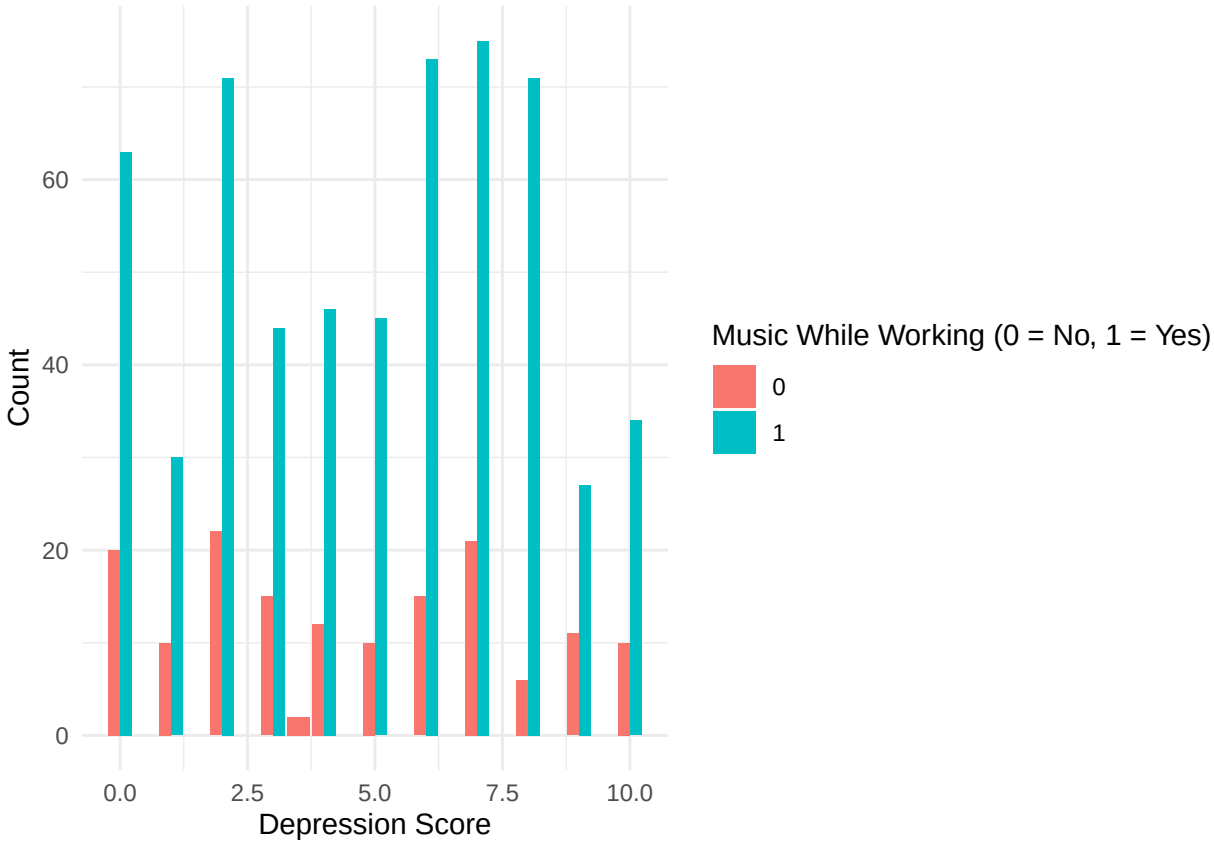
From this boxplot we can see that the median of the data is roughly equivalent. “No” being around 4 and “Yes” being around 5. But from the t-test we can conclude this difference isn’t statistically significant.

The spread of scores is also very similar, both ranging from around 2 to 7.

Both data also have the same min and maxes

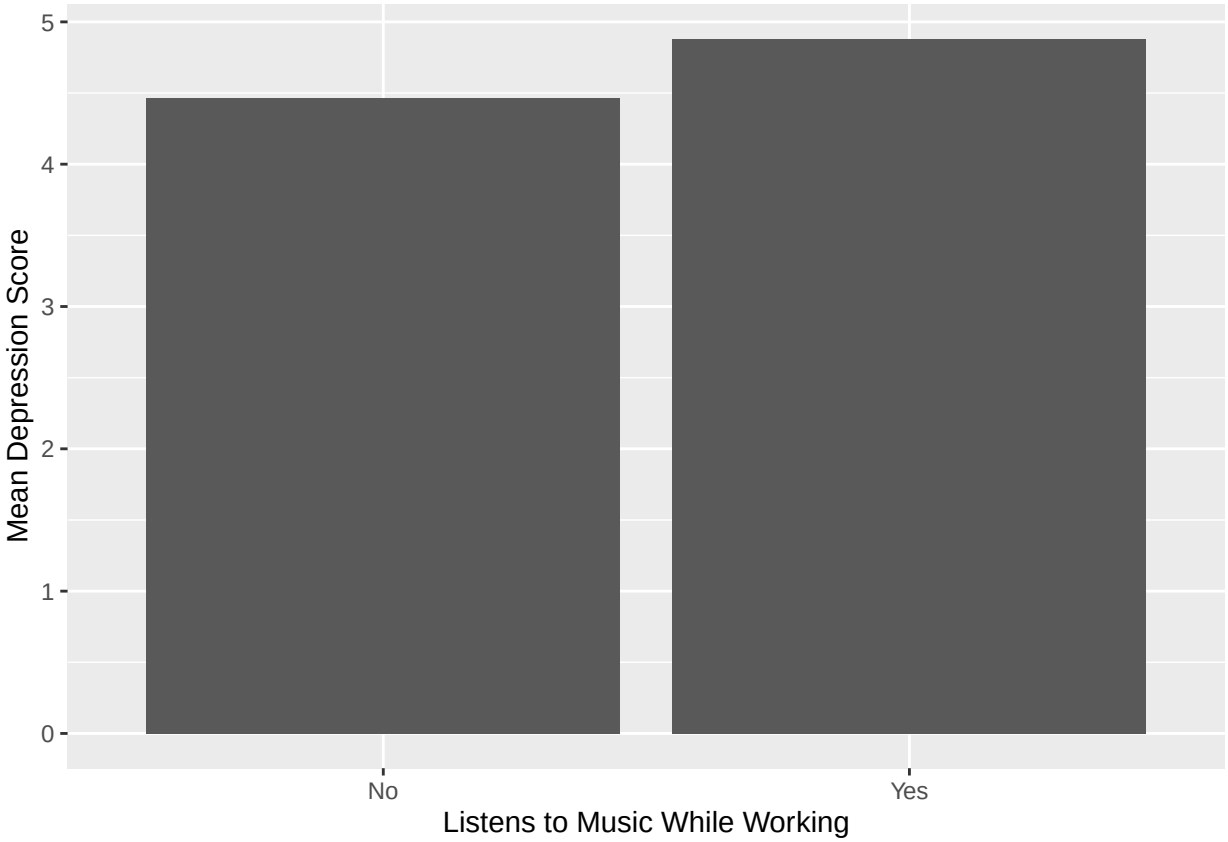
Considering the similarity of the boxplots we can conclude that studying with or without music doesn’t affect depression levels.

Depression Score Frequencies: Music vs. No Music While Working



By viewing a bar graph of depression scores by music while working(blue) and no music while working(red), it is clear that there is a significantly higher count of individuals who work while listening to music at every level of depression. A better method to analyze this is using a mean bar plot, in which the height of the bar represents the mean for each category.

Mean Depression Levels: Music vs. No Music While Working



The mean bar graph indicates that participants who listen to music while working have a slightly higher average depression score than those who do not. However, this difference is minute, and the t-test indicates it is not statistically significant.

Conclusion

Based on the data, listening to music while working does not show a strong relationship with depression scores. The correlation is extremely small (close to 0), and the t-test found no significant difference in average depression levels between students who listen to music and those who do not. The boxplot also shows that the two groups have similar distributions with no major differences. Overall, the evidence suggests that working with or withous music does not predict better or worse depression outcomes in this dataset.

Question 5: Does playing a musical instrument affect depression levels among individuals

Introduction

The purpose of this analysis is to investigate the question:
Does playing a musical instrument affect depression levels among participants in the Music & Mental Health (M&MH) dataset?

Many people believe that playing an instrument may help regulate emotions or reduce stress. To test this idea, I compare depression levels between individuals who identify as instrumentalists and those who do not.

In the sections that follow, I will:

1. Present summary statistics for both groups
2. Visualize the distributions using a boxplot
3. Compare average depression scores using a bar plot with error bars
4. Conduct a two-sample t-test to check for statistical significance
5. Provide a final conclusion based on all findings

Summary Statistics for Depression by Instrumentalist Status

Before performing any statistical test or showing any visuals, it is important to look at basic descriptive statistics.
These give a first look at whether the two groups appear different on average.

Instrumentalist	mean_depression	sd_depression	n
No	4.782828	3.065052	495
Yes	4.823404	2.937323	235

Interpretation

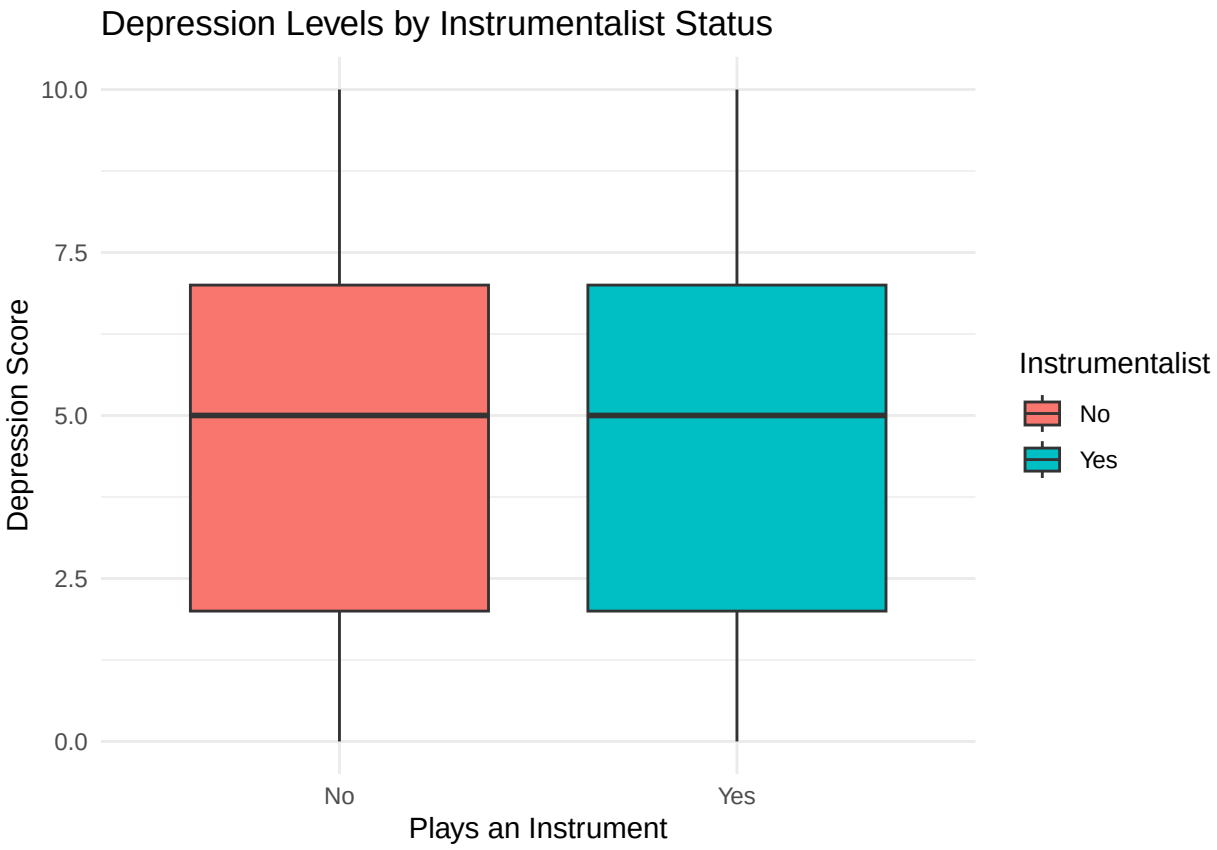
Instrumentalists report an average depression score of about **4.82**,
while non-instrumentalists report **4.78**.

These values are extremely close, suggesting that any difference between the groups is minimal even before running a formal test.

Boxplot of Depression by Instrumentalist Status

To better understand the distribution of depression scores for each group, we use a **boxplot**.
This helps us see differences in:

- medians
- variability
- spread
- possible outliers

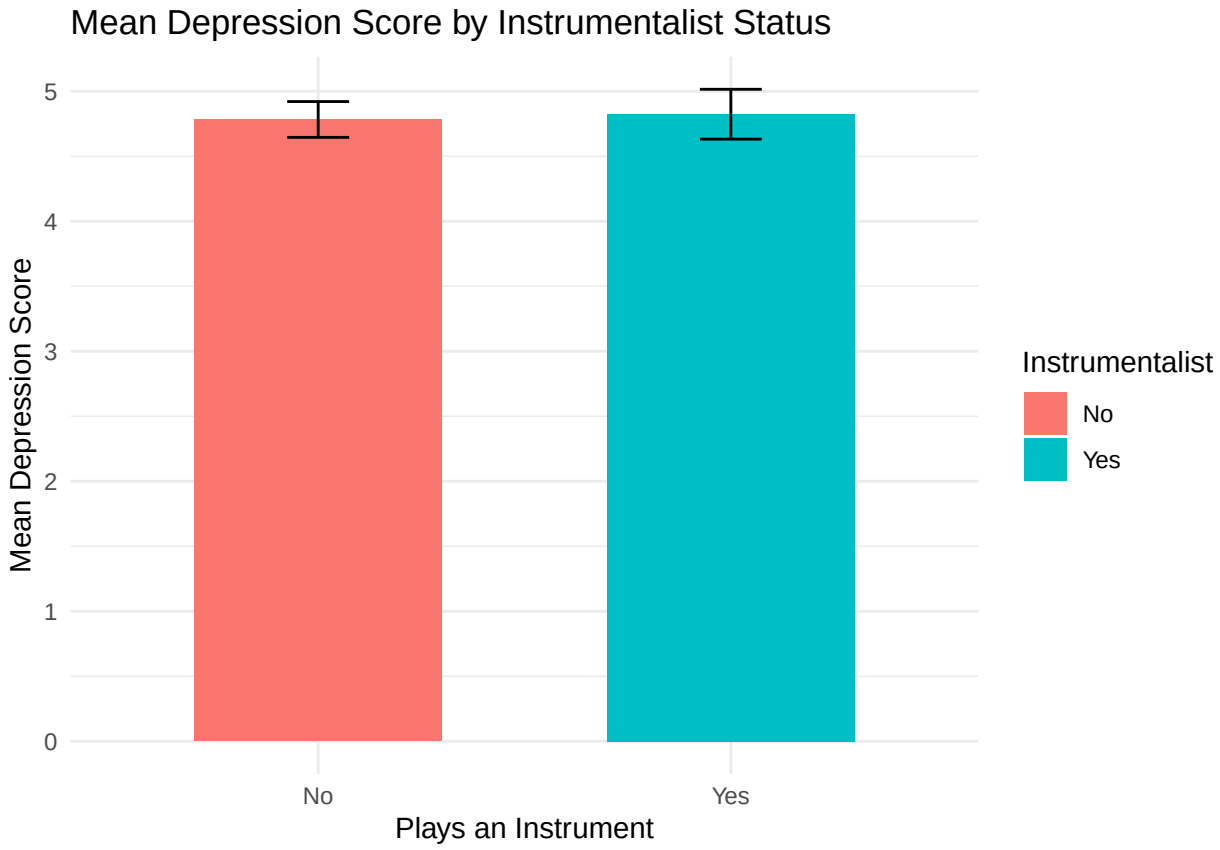


Interpretation

The boxplots for both groups look almost identical.
Their medians, IQRs, and overall spreads overlap heavily, suggesting little difference between instrumentalists and non-instrumentalists.

Mean Bar Plot With Standard Error Bars

Now that we've looked at overall distributions, we focus specifically on comparing average depression levels. This bar plot includes standard error bars, which help us judge whether differences in the sample means might be meaningful.



Interpretation

The difference in means is less than **0.05**, and the error bars overlap completely.
This visually reinforces that depression scores are essentially the same for both groups.

Two-Sample Independent T-Test

To determine whether these tiny differences are statistically significant,
we perform a two-sample independent t-test comparing the mean depression levels between instrumentalists and non-instrumentalists.

Hypotheses

- H_0: The mean depression levels are equal between the two groups
- H_1: The mean depression levels differ

```
##
##  Welch Two Sample t-test
##
## data:  Depression by Instrumentalist
## t = -0.17194, df = 477.96, p-value = 0.8636
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.5042901  0.4231382
## sample estimates:
##  mean in group No mean in group Yes
##          4.782828          4.823404
```

Interpretation

The p-value is **much greater than 0.05**, meaning:

- We fail to reject the null hypothesis
- There is no statistical evidence of a difference in depression scores
- Any observed difference is likely due to random chance

Conclusion

Based on all numerical, visual, and statistical evidence,
there is **no meaningful relationship** between playing a musical instrument and depression levels in this dataset.

Both instrumentalists and non-instrumentalists report nearly identical depression scores,
and the t-test confirms that the difference is not statistically significant.

Final Statement:
Playing a musical instrument does **not** appear to influence depression levels among participants in the M&MH survey.

Conclusion

We looked at how depression relates to several factors in the Music and Mental Health survey, including age, playing a musical instrument, music preferences, listening time, and listening to music while working.

From our analysis of the dataset, several important findings emerged:

- Younger and older participants reported very similar levels of depression, so age does not seem to have a strong effect.
- Whether someone plays an instrument or not does not appear to change their depression levels.
- People who like Rock music reported slightly higher depression levels than others. The effect is small but noticeable.
- Spending more time listening to music is linked to slightly higher depression scores, though the effect is minor.
- Listening to music while working does not seem to make a difference in depression levels; those who listen and those who don't have very similar experiences.

Limitations:

- Dataset only includes 700 participants, possibly not a good representation of the total population.
- Self-reported measures of depression and music habits may introduce bias.
- Some variables may have limited variability or uneven group sizes, which can reduce statistical power.

Future Directions:

- Incorporate additional variables such as gender and income to get a deeper understanding of how they influence depression
- Conduct a long-term study to track changes in depression and music listening habits over several years.
- For relationships that had an effect, perform an advanced analysis to find more infromation on the relationship between the variables and depression

Overall Conclusion:

While age and playing a musical instrument do not appear to affect depression levels, certain music preferences, specifically Rock and higher listening time show small positive associations with depression. Listening to music while working does not appear to influence depression. This shows that although music and mental health are subtly related, music does not directly influence depression.

Appendix — full source + session info

```
knitr::opts_chunk$set(echo = FALSE)
# Load Libraries
library(readr)
library(tidyverse)
library(ggplot2)
library(dplyr)
library(knitr)
cat('\\pagebreak')
cat('\\pagebreak')
# Load in data
#music_data <- read_csv("mxmh_survey_results.csv")
music_data <- read.csv("mdepression.csv")
# Filter out Nulls and Change Data Types
music_data <- music_data %>% filter(!is.na(Age))
# Change age and depression to integer
music_data <- music_data %>% mutate(Age = as.integer(Age), Depression = as.integer(Depression))
# Display all columns of the dataset
names(music_data)
cat('\\pagebreak')
age_summary <- fivenum(music_data$Age)
depression_summary <- fivenum(music_data$Depression)

# Combine into a table
summary_table <- data.frame(
  Statistic = c("Minimum", "Q1", "Median", "Q3", "Maximum"),
  Age = age_summary,
  Depression = depression_summary
)

# Print nicely
kable(summary_table, caption = "Five-number summaries for Age and Depression")
cat('\\pagebreak')

# Scatterplot to visualize correlation
ggplot(music_data, aes(x=Age, y=Depression)) + geom_point() +
  geom_smooth(method=lm, se=FALSE, linetype="dashed",
              color="darkred") + labs(title = "Relationship Between Age And Depression") +
  theme(plot.title = element_text(hjust = 0.5, family = "Times", face = "bold", size=20))

#Calculate correlation of Age and Depression
cor(music_data$Age, music_data$Depression)

cat('\\pagebreak')

# Separate age into two groups to perform analysis
music_data <- music_data %>%
  mutate(AgeGroup = ifelse(Age < 21, "Under 21", "21 and Over"))
```

```r
# Check count of each
my_table <- music_data %>%
  count(AgeGroup, name = "Count") %>%
  mutate(Percent = Count / sum(Count) * 100)

kable(my_table)
cat('\\pagebreak')
ggplot(music_data, aes(x = Depression, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.3) +
  labs(title = "Density of Depression Scores by Age Group",
       x = "Depression Score",
       y = "Density") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, family = "Times", face = "bold", size=20)) +
  scale_fill_manual(values = c("#F8766D", "#00BFC4")) +
  scale_color_manual(values = c("#F8766D", "#00BFC4"))
cat('\\pagebreak')
# Check the spread of the data
ggplot(music_data, aes(x = AgeGroup, y = Depression, fill = AgeGroup)) +
  geom_boxplot(alpha = 0.5) +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "white") +
  labs(title="Depression based on AgeGroup",
                                x ="Age Group", y = "Depression Level (1 to 10)") +
  theme_minimal() + theme(plot.title = element_text(hjust = 0.5, family = "Times", face = "bold", size=
cat('\\pagebreak')
mean_under_21 <- mean(music_data$Depression[music_data$AgeGroup == "Under 21"])
mean_over_21 <- mean(music_data$Depression[music_data$AgeGroup == "21 and Over"])

var_under_21 <- var(music_data$Depression[music_data$AgeGroup == "Under 21"])
var_over_21 <-  var(music_data$Depression[music_data$AgeGroup == "21 and Over"])

var_under_21
var_over_21
result <- t.test(Depression ~ AgeGroup, data = music_data)
result$estimate
result$conf.int
result$p.value
result$statistic
cat('\\pagebreak')
library(readr)
library(ggplot2)
music_data <- read_csv("mdepression.csv")
Fav_Genre = music_data$`Fav genre`
Depression_Levels = music_data$Depression

as.data.frame(table(Fav_Genre))
ggplot(music_data, aes(x = Fav_Genre, y = Depression_Levels)) + labs(x = "Music Genre", y = "Depression
Rock_data <- subset(music_data, Fav_Genre == "Rock")

summary(Rock_data$Depression)
Rock <- ifelse(music_data$`Fav genre` == "Rock", 1, 0)
cor.test(Rock, Depression_Levels)
cat('\\pagebreak')
```

```r
library(tidyverse)
library(readr)
library(ggplot2)


Hours <- music_data$`Hours per day`
Depression_Levels <- music_data$Depression
data <- music_data %>% dplyr::select(Hours = `Hours per day`, Depression)

table(data$Hours)
ggplot(data, aes(x = as.factor(Hours), y = Depression)) +
geom_boxplot() +
labs(x = "Hours Listening Per Day", y = "Depression Level") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

cor.test(Hours, Depression_Levels)
ggplot(data, aes(x = Hours, y = Depression)) +
  geom_jitter(width = 0.15, height = 0.15, alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Depression vs Hours Listening (scatter + linear fit)",
    x = "Hours listening per day",
    y = "Depression score (0-10)"
  ) +
  theme_minimal()
cat('\\pagebreak')
library(tidyverse)

music_data <- read.csv("mdepression.csv")

# Remove blanks
music_data <- music_data[!(is.na(music_data$While.working) | trimws(music_data$While.working) == ""), ]

# Standardize text
music_data$While.working <- tolower(trimws(music_data$While.working))

# Replace yes/no
music_data$While.working[music_data$While.working %in% c("yes", "y")] <- 1
music_data$While.working[music_data$While.working %in% c("no", "n")] <- 0

# Convert to numeric
music_data$While.working <- as.numeric(music_data$While.working)

# Check result
#(mdepression$While.working)


working <- music_data$While.working

depression <- music_data$Depression

cor(working, depression)
mdepressionunf <- read.csv("mdepression.csv", stringsAsFactors = FALSE)
```

```r
#unique(mdepressionunf$While.working)

# 1. Clean the text FIRST
mdepressionunf$While.working <- trimws(mdepressionunf$While.working)
mdepressionunf$While.working <- tolower(mdepressionunf$While.working)

# 2. Remove TRUE blanks or NA
mdepressionunf <- mdepressionunf[
  mdepressionunf$While.working != "" & !is.na(mdepressionunf$While.working),
]

# 3. Convert real yes/no answers
mdepressionunf$While.working <- ifelse(mdepressionunf$While.working == "yes",
                                       "Yes", "No")

yes_group <- mdepressionunf$Depression[mdepressionunf$While.working == "Yes"]
no_group  <- mdepressionunf$Depression[mdepressionunf$While.working == "No"]

var_yes <- var(yes_group, na.rm = TRUE)
var_no  <- var(no_group, na.rm = TRUE)

var_yes
var_no
mdepressionunf %>%
  group_by(While.working) %>%
  summarise(
  mean_depression = mean(Depression, na.rm = TRUE),
  n = n()
)
t.test(Depression ~ While.working, data = mdepressionunf)
  mdepressionunf %>%
  ggplot(aes(x = While.working, y = Depression)) + geom_boxplot()


ggplot(music_data, aes(x = Depression, fill = factor(While.working))) +
  geom_bar(position = "dodge") +
  labs(
    x = "Depression Score",
    y = "Count",
    fill = "Music While Working (0 = No, 1 = Yes)"
  ) +
  theme_minimal()
means <- music_data %>%
  group_by(While.working) %>%
  summarize(mean_depression = mean(Depression, na.rm = TRUE))

# Convert 0/1 into labels
means$While.working <- factor(means$While.working,
                              levels = c(0,1),
                              labels = c("No", "Yes"))

# Bar plot of means
ggplot(means, aes(x = While.working, y = mean_depression)) +
```

```r
  geom_col() +
  labs(x = "Listens to Music While Working",
       y = "Mean Depression Score")
cat('\\pagebreak')
#music_data <- read.csv("music_data_survey_results.csv", stringsAsFactors = FALSE)
music_data$Instrumentalist <- trimws(as.character(music_data$Instrumentalist))
music_data <- music_data[music_data$Instrumentalist %in% c("Yes", "No"), ]
cat("\\pagebreak")
summary_stats <- music_data %>%
  group_by(Instrumentalist) %>%
  summarise(
    mean_depression = mean(Depression, na.rm = TRUE),
    sd_depression   = sd(Depression, na.rm = TRUE),
    n               = n()
  )
kable(summary_stats)
cat("\\pagebreak")
ggplot(music_data, aes(x = Instrumentalist, y = Depression, fill = Instrumentalist)) +
  geom_boxplot() +
  labs(
    title = "Depression Levels by Instrumentalist Status",
    x = "Plays an Instrument",
    y = "Depression Score"
  ) +
  theme_minimal()
cat("\\pagebreak")
summary_df <- music_data %>%
  group_by(Instrumentalist) %>%
  summarise(
    mean_dep = mean(Depression, na.rm = TRUE),
    sd_dep   = sd(Depression, na.rm = TRUE),
    n        = n(),
    se       = sd_dep / sqrt(n)
  )

ggplot(summary_df, aes(x = Instrumentalist, y = mean_dep, fill = Instrumentalist)) +
  geom_col(width = 0.6) +
  geom_errorbar(aes(ymin = mean_dep - se, ymax = mean_dep + se), width = 0.15) +
  labs(
    title = "Mean Depression Score by Instrumentalist Status",
    x = "Plays an Instrument",
    y = "Mean Depression Score"
  ) +
  theme_minimal()
cat("\\pagebreak")
t.test(Depression ~ Instrumentalist, data = music_data)
```