````
---
title: "The Effects of Music on Mental Health: A Data-driven Analysis"
output:
  pdf_document: default
  html_document: default
date: "2024-04-30"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## load packages
```{r}
library(dplyr)
library(tidyr)
library(ggplot2)
library(corrplot)
library(readr)
library(pheatmap)
library(cluster)
library(factoextra)
```

## loading data
```{r}
data <- read.csv("C:/Users/cgonz/Dropbox/My PC (LAPTOP-
CUJK30BH)/Downloads/mxmh_survey_results.csv")

head(data)
```
## Data summary
```{r}
summary(data)

```
## 1.   What correlations exist between different music genres and their self-reported
mental health conditions?
```{r}
favgenre_data <- data %>%
  count(Fav.genre) %>%
  arrange(desc(n))  # arranges frequency in descending order

ggplot(favgenre_data, aes(x=Fav.genre, y=n)) +
  geom_bar(stat="identity", fill="red") +
  labs(title="Favorite Genres by Popularity", x="Genre", y="Count") +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```

```{r}
this_data <- data %>%
  mutate(across(starts_with("Frequency"), ~case_when(
    . == "Never" ~ 0,
    . == "Rarely" ~ 1,
    . == "Sometimes" ~ 2,
    . == "Very frequently" ~ 3,
    TRUE ~ NA_real_
  )))

genre_columns <- names(data)[grepl("^Frequency", names(data))]

data_long <- data %>%
  pivot_longer(cols = genre_columns, names_to = "Genre", values_to = "Frequency") %>%
  group_by(Genre, Frequency) %>%
````
````

```
  summarise(
    Anxiety_Median = median(Anxiety, na.rm = TRUE),
    Depression_Median = median(Depression, na.rm = TRUE),
    Insomnia_Median = median(Insomnia, na.rm = TRUE),
    OCD_Median = median(OCD, na.rm = TRUE),
    .groups = 'drop'
  )

# function to create bar plots for each condition
plot_condition <- function(data, condition, title, ylab) {
  ggplot(data, aes(x = Genre, y = get(condition), fill = factor(Frequency))) +
    geom_bar(stat = "identity", position = position_dodge(width = 0.7), width = 0.6) +
    scale_fill_manual(values = c("darkred", "red", "orange", "pink"),
                      labels = c("Never", "Rarely", "Sometimes", "Very Frequently")) +
    labs(title = title, x = "Genre", y = ylab, fill = "Listening Frequency") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

# Plotting each condition
plot_condition(data_long, "Anxiety_Median", "Relation between Anxiety & Genre Frequency",
"Median Anxiety Score")
plot_condition(data_long, "Depression_Median", "Relation between Depression & Genre
Frequency", "Median Depression Score")
plot_condition(data_long, "Insomnia_Median", "Relation between Insomnia & Genre
Frequency", "Median Insomnia Score")
plot_condition(data_long, "OCD_Median", "Relation between OCD & Genre Frequency", "Median
OCD Score")
```

Correlation matrix
```{r}
mental_health_vars <- c("Anxiety", "Depression", "Insomnia", "OCD")
genre_cols <- grep("Frequency", names(this_data), value = TRUE)


cor_matrix <- cor(this_data[genre_cols], this_data[mental_health_vars],
use="complete.obs")

pheatmap(cor_matrix, display_numbers = TRUE, title = "Correlation Heatmap")
```

"Which age group experiences the most mental health issues?"
```{r}
ggplot(data, aes(x=Age)) +
  geom_histogram(binwidth = 1, fill="steelblue", color="black") +
  ggtitle("Age Distrubution of Respondents") +
  xlab("Age") +
  ylab("Count")

```

```{r}
this_data$AgeGroup <- cut(this_data$Age, breaks=c(0, 18, 35, 50, 65, Inf), labels=c("0-
18", "19-35", "36-50", "51-65", "66+"))

ggplot(this_data, aes(x=AgeGroup, y=Anxiety, fill=AgeGroup)) +
  geom_bar(stat="summary", fun="mean") +
  labs(title="Average Anxiety Score by Age Group")

ggplot(this_data, aes(x=AgeGroup, y=Depression, fill=AgeGroup)) +
  geom_bar(stat="summary", fun="mean") +
  labs(title="Average Depression Score by Age Group")

ggplot(this_data, aes(x=AgeGroup, y=Insomnia, fill=AgeGroup)) +
```

```
  geom_bar(stat="summary", fun="mean") +
  labs(title="Average Insomnia Score by Age Group")

ggplot(this_data, aes(x=AgeGroup, y=OCD, fill=AgeGroup)) +
  geom_bar(stat="summary", fun="mean") +
  labs(title="Average OCD Score by Age Group")
```

ANOVA
```{r}
# Performing ANOVA for each mental health variable
anova_anxiety <- aov(Anxiety ~ AgeGroup, data=this_data)
summary(anova_anxiety)

cat("\nDEPRESSION\n")
anova_depression <- aov(Depression ~ AgeGroup, data=this_data)
summary(anova_depression)

cat("\nINSOMNIA\n")
anova_insomnia <- aov(Insomnia ~ AgeGroup, data=this_data)
summary(anova_insomnia)

cat("\nOCD\n")
anova_OCD <- aov(OCD ~ AgeGroup, data=this_data)
summary(anova_OCD)
```

Correlation tests
We need to see if Hours per Day is normally distributed:
```{r}
shapiro.test(this_data$`Hours.per.day`)
```
This implies our data is normally distributed (low p value). Since this subset of our data is normally distributed, we're going to be performing a Pearson correlation test.

```{r}
cat("ANXIETY\n")
cor.test(this_data$Hours.per.day, this_data$Anxiety, method="pearson")

cat("\nDEPRESSION\n")
cor.test(this_data$Hours.per.day, this_data$Depression, method="pearson")

cat("\nINSOMNIA\n")
cor.test(this_data$Hours.per.day, this_data$Insomnia, method="pearson")

cat("\nOCD\n")
cor.test(this_data$Hours.per.day, this_data$OCD, method="pearson")

```

Is there a correlation between self-reported mental health and identifying with being a musician (i.e. instrumentalist or composer)?
```{r}
data$Is_Musician <- ifelse(data$Instrumentalist == "Yes" | data$Composer == "Yes", "Yes", "No")

summary_stats <- data %>%
  group_by(Is_Musician) %>%
  summarise(across(c(Anxiety, Depression, Insomnia, OCD),
                   list(mean = ~mean(.x, na.rm = TRUE),
                        sd = ~sd(.x, na.rm = TRUE))))

```

T-test for: "Is there a correlation btwn being a musician and mental health issues?"

```{r}
# T-test for Anxiety
t_test_anxiety <- t.test(Anxiety ~ Is_Musician, data = data)
print(t_test_anxiety)

# T-test for Depression
t_test_depression <- t.test(Depression ~ Is_Musician, data = data)
print(t_test_depression)

# T-test for Insomnia
t_test_insomnia <- t.test(Insomnia ~ Is_Musician, data = data)
print(t_test_insomnia)

# T-test for OCD
t_test_ocd <- t.test(OCD ~ Is_Musician, data = data)
print(t_test_ocd)
```

MODELS

Supervised: Multiple Regressions & Diagnostic Plots
```{r}
cat("ANXIETY\n")
# predicting anxiety from Age, Hours per Day, and Classical Music Frequency
lm_model <- lm(Anxiety ~ Age + Hours.per.day + Frequency..Classical., data=this_data)
summary(lm_model)

par(mfrow = c(2, 2))  # Set up the plotting area for multiple plots
plot(lm_model, which = 1:4)  # Plots for residuals

cat("\nDEPRESSION\n")
# predicting depression from Age, Hours per Day, and Classical Music Frequency
lm_model <- lm(Depression ~ Age + Hours.per.day + Frequency..Classical., data=this_data)
summary(lm_model)

par(mfrow = c(2, 2))  # Set up the plotting area for multiple plots
plot(lm_model, which = 1:4)  # Plots for residuals
```

Unsupervised: K-means Clustering
```{r}
data_selected <- data %>%
  select(Age, Hours.per.day, Anxiety, Depression, Insomnia, OCD) %>%
  na.omit()

# save row indices of data used for clustering
row_indices <- row.names(data_selected)

# standardize data
data_scaled <- scale(data_selected)

# perform clustering (k = 4)
set.seed(123)  # Ensure reproducibility
kmeans_result <- kmeans(data_scaled, centers = 4, nstart = 25)

# data frame has a 'Cluster' column initialized with NA
data$Cluster <- NA

valid_indices <- which(complete.cases(data[, c("Age", "Hours.per.day", "Anxiety",
"Depression", "Insomnia", "OCD")]))

# assign the cluster results back to the original data using the valid indices
data$Cluster[valid_indices] <- kmeans_result$cluster

fviz_cluster(kmeans_result, data = data_scaled, geom = "point", ellipse.type = "convex",
```

```
            palette = "jco", ggtheme = theme_minimal())
```