Project 1 Presentation Notes

Write up

* Intro
* Data cleaning
* Things we tried that didn’t work
* Data story (3 q)
* Regression
* Conclusion + call to action
* Limitations + future work
* Works cited

Data cleaning:

Our initial study data included information we did not use for this particular analysis, including the user’s primary streaming service, whether or not the user listened to music while working, whether or not they considered themselves instrumentalists or composers, and if they spoke more than one language. We figured we needed more information to be able to use any of these columns in objective data analyses. We also could not compare changes in mental health rankings over time since this data only includes a timestamp of when each response was submitted to the survey, so we dropped that column as well.

We dropped rows with data we deemed unreliable, such as the respondent who said he listened to music with a “999999999.0” BPM and 3 rows that claimed they listened to music 24 hours/day. We also dropped eight rows with null values for Music Effects and one row with no Age listed so we could make sure that we were all using the same data for our various analytics.

Q3: Does the frequency of listening to different genres correlate with different mental health disorders? Can we look at BPM versus disorder quantity?

With data from over 700 respondents and 16 different music genres, we wanted to see if there was a correlation between users claiming a high rank of each disorder and the genres they listened to. The Frequency columns all contained string values to describe how often the respondent listened to each genre. I replaced those values with integers, with “Never” as 0, “Rarely” as 1, “Sometimes” as 2, and “Very frequently” as 3. I took our main dataframe and separated it into four different dataframes with one for each genre where the data was filtered by mental health rankings of 7/10 or higher. For each disorder dataframe, I summed the integer values for each genre’s Frequency in order to compare them objectively. This graph shows that distribution: A screenshot of a graph

Description automatically generated

In these same four dataframes, I also looked at the genres each respondent listed as their favorite. Interestingly, the results are somewhat similar but actually vary quite a bit with the first graph. A screenshot of a computer

Description automatically generated

According to these values (donut charts?), the three genres most listed as favorites among users with high disorder rankings are Rock, Pop, and Metal. But the frequency each of these is listened to doesn’t exactly match, as Metal is listened to 61% as much as Rock, which aligns with favoritism, but Metal is listened to 63% as much as Pop, even though it’s just under Pop in favoritism. This could indicate that listeners of Metal are more likely to report having poor mental health, though we don’t have enough information here to determine if listening to Metal music *causes* poor mental health.

It is worth noting that the numbers represented in each bar of the Frequency graph do contain some repeat respondents, as many of them reported higher rankings for more than one disorder. Also, the numbers represent an aggregate of the data since each Frequency value was changed to an integer and summed. So there are not exactly 1632 different people in this data who listen to Rock music frequently. Rather, 1632 is the total of aggregated responses to the frequency of listening to Rock music among users with 7 or higher disorder rankings. This specific data encompasses the responses of 417 users, so the Frequency graph shows that many of these users listen to multiple types of music, and many of them also have high ranks for multiple mental health disorders.

Since the top favorite genres among our respondents with high mental disorder ranks are rock, pop, and metal, I thought there might be a relationship between high BPM and disorder rank. To evaluate this, I created groups of BPM counts for my initial dataframe with null values for BPM dropped. I used the average disorder rank for each BPM group for each disorder to create the following heatmap.

A table of pink and purple squares

Description automatically generated

Disorder rank is indicated by the color bar. There were no average values above 7.5 when all respondents who indicated a BPM value were included. The heatmap clearly shows that disorder rank is similar for almost all users within each individual disorder, despite their average BPM grouping. Since increasing the average BPM does not necessarily increase the disorder rank, we cannot say that BPM really makes much difference to disorder rank. We can glean more information from looking at the favorite genres and frequencies of listening for each of those when looking for trends in mental health status.

Limitations:

This dataset had quite a few limitations, and we developed some suggestions for improvements for future study. Foremost, we can’t judge mental health improvement with listening to music over time from this data. The same users reporting in over a set time period would offer more capability of establishing relationships with both frequency of genre usage and hours listened/day. Second, the mental health data was very subjective. Users were allowed to choose their rank for each disorder from 0 to 10, so it follows that each user might interpret their numbers differently. If these results were concentrated over a specific geographic location, for instance, the United States, we could compare their music statistics from the survey with actual published mental health statistics for objective data.

The BPM section was not specific at all. It was optional, so about 100 users did not give a number, and we do not know if each user interpreted this number to be an average, the most common, or the highest BPM they listened to. Also, BPM is not something common music listeners are necessarily aware of, so many of these numbers could have been estimates. This is probably why we could not correlate anything with BPM. Additionally, the original data asked if users listened to music during work or not, but it did not ask for occupation or when else they listen. Clarifying this might make that part of the data more significant. We would also like to see gender of the respondent included, as we would have liked to be able to compare how frequencies of listening to certain genres affect men and women differently.

Lastly, we have no basis for the results in the music effects section. Was the reported effect from listening to certain genres or listening for certain periods of time? Many of our respondents listed high ranks for multiple disorders, so does their reported effect apply to a certain disorder or multiples? It might have been useful to ask for ranks for each disorder over time, or after increasing or decreasing the amount of time the user listened to music each day. If we ran this survey in the future, we would clarify the questions to cover these limitations and hopefully collect more objective data.

Works Cited:

* Kaggle dataset: [Music & Mental Health Survey Results (kaggle.com)](https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results)
  + Thank you Professor Booth for helping us find this!
* [Obsessive-Compulsive Disorder (OCD) - Psychiatric Disorders - Merck Manual Professional Edition (merckmanuals.com)](https://www.merckmanuals.com/professional/psychiatric-disorders/obsessive-compulsive-and-related-disorders/obsessive-compulsive-disorder-ocd)
* Basic information regarding different music genres and mental health disorders: Wikipedia.com
* Data Analysis:
  + Matplotlib.org documentation
  + Xpert learning AI
  + TA Mike Wenner: troubleshooting ☺
* Formatting:
  + Coolors.co
  + https://www.remove.bg/upload

Value Totals: 994 Metal, 1632 Rock, 1573 Pop

Metal = 61% Rock, Metal = 63% Pop