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.

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.

Value Totals: 994 Metal, 1632 Rock, 1573 Pop

Metal = 61% Rock, Metal = 63% Pop