CODE FIRST GIRLS G6
Autumn Wade
Amy Minns
Isobel Wilson
Isha Ibrahim
Vilma Sinani
Wenjia Knight

CFG Degree Spring 2023 Data 2 - Group 6 Project Report

1.0 INTRODUCTION

The aim of this project is to examine the relationship between COVID-19 lockdowns and the mood of music listened to in the UK. Our key objectives were to use the skills we have built throughout the course to see what kind of relationship (if any) there was between these factors in the hopes that others (such as record labels, chart companies and mental health professionals in the NHS) may be able to use our data and findings to build upon this work.

1.1 BACKGROUND

Music Psychology

Music psychology is a relatively new field of study that aims to understand the impact music has on humans both cognitively and physically, and to ascertain if various psychological states can be influenced or induced by music application.

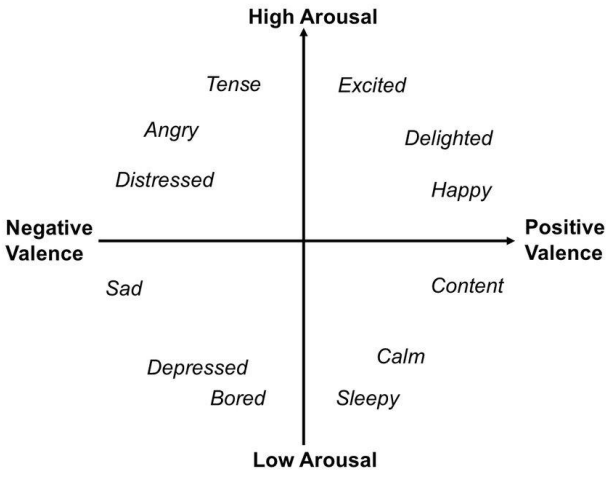
Research shows that humans like to listen to music that matches how we feel (Hunter et al., 2011). This phenomenon is called "emotional congruence". Feeling is what makes us human and we like to feel; whether negative or positive, we just want to feel.

Listening to music stimulates the brain to release dopamine. Dopamine is also known as 'the feel-good hormone' and is a key component of the human pleasure system. It is released by the brain whenever you listen to music that moves you, and music will move you more with emotional congruence, i.e. feel congruent to the emotion you are feeling at the time. The aim of the project is to apply these theories of music psychology on a small scale.

The algorithm

We will be attempting to classify the mood of the music data using Russell's circumplex model of emotion. (Russell, Lewicka 1980)

Russell theorised that emotion could be mapped on a two dimensional plane using arousal and valence as dimensions. In this model, different measures of arousal and valence represent different moods. We used the audio track valence and energy data for classification and we focused on four larger mood classification areas: happy, sad, calm, and agitated. This is a key limitation of the project. In the future, this model could be adapted to include a larger range of moods. The model excludes any sentiment analysis of lyrics and it is also worth noting that musical preferences and sound expectations can be culture dependent. For example, in non-western music there may not be correlation between positive/negative emotions and valence/energy; hence, the western definition of 'happy' may not be interpreted as such elsewhere.



Furthermore, emotions are complex. No algorithm can capture emotions with complete accuracy, but we should be able to get some basic approximations.

The Questions

By taking the daily streaming data of the UK over the last three years and applying an algorithm to ascertain a 'mood' for each of the tracks, we aim to create a UK 'mood' that can be referenced by date across the last five years. This algorithm attempts to answer the following questions:

- Has the mood of the music people in the UK listen to changed in relation to the COVID-19 pandemic timeline?
- Is there a correlation between the mental state of the UK population and the mood of music they choose to listen to?

To answer the first question, we cross referenced mood data with the unique timeline of the UK COVID response, seeing if a change in the 'mood' of tracks can be found before, during, and after COVID, and the various lockdowns. To further test the mood algorithm we attempted to answer the second question. We also tried to correlate the track mood data with the NHS mental health service referrals data we have collected.

Data does not have feelings, but humans do.

What could this data be used for?

As stated above, data does not have feelings, but the humans that data is being collected from do.
If we do find a correlation between the mood of the music streamed by the general population of the UK and the mental health state of the UK, we suggest this data could be used in two very different ways.

1. Marketing and consumer data research

Emotion data is seen as the next big thing. For a business to gain an insight into the consumers emotional state would enable much more customised experiences, products or stories.
Imagine that after having a bad day at work you listen to some melancholic music on your way home in the car, that data is streamed to your tv streaming service which, when you get home, suggests some mood appropriate TV shows for you. Or rather, you get an email from your grocery provider suggesting you may want to rebuy the particular chocolate bar that it knows you purchased the last time you listened to 'sad' music.
Alternatively, a company may want their product to have positive 'happy' connotations, so after a day of listening to happy up-beat music this company may choose to show you an advertisement for their product. The next time you listen to happy music, you may then be reminded of the product. This is of course a very 'big data' use and probably a massive invasion of privacy and data laws.

2. NHS mental health services funding distribution

This use is much more altruistic and could help people.
The NHS has been known to be struggling, especially the mental health sector, where demand, unfortunately, far outweighs supply. Waiting lists for access to mental health services can be years in some areas. We speculate that this data could be used to predict,regionally, areas with a heightened listening of 'sad' music giving the NHS a way to spread the funding for services more evenly based on emotional need. Or perhaps give an idea on mental health trends to also aid in provision of services.
Again, this could be an invasion of privacy. The use of data laws may dictate what this data can and cannot be used for, but these are merely suggestions.

1.2 STEP SPECIFICATIONS

Step 1: Initial ideas and discussions

During the initial brainstorm meeting, Amy put forward the idea of looking into the mood of music and people's mental health. This idea was welcomed by the group and discussions on how we could execute the project included:

- Identifying Spotify API as the API to use, as one of the project requirements is to use at least one API. It was not decided how Spotify API would be used at this stage.
- Looking at different moods for people in different countries/regions.
- Looking for correlations between low mental health levels in areas heavily affected by COVID.

- Searching government COVID data as another source of data in CSV format.
- Looking at pre and post COVID moods of music, focusing on different areas depending on local restrictions. For example, England had a different tier system than Wales, which means different areas of the country had different measures in place.
- Could this project be used to predict future trends in popular music? Record label companies might be interested in our project.
- Using classification models to label songs into different mood categories.
- Perform sentiment analysis on song lyrics to help assess the overall mood conveyed by a song.

We agreed that we would research which datasets were available to see what kind of questions we might be able to answer. We also further researched the utility of Spotify API.

Step 2: Data gathering

At the second meeting, everyone was finding it difficult to gather suitable datasets to give insight on what music was popular at a specific point in time and to assess the mood of the population. There are available datasets on Kaggle and Google data such as 'Top 50 Spotify songs - 2019', 'Top Spotify songs from 2010-2019', 'Top 100 songs (2021)', 'Billboard "The Hot 100" Songs (dated 04/08/1958 - 06/11/2021)'. There was no dataset that could give us the full history of popular songs from 2018 - 2023 (before, during, and after COVID). However, Wenjia was able to produce a few datasets which included:

- A dataset of popular songs in the UK for the past 5 years from 27/04/2018 to 21/04/2023 (uk_top_singles_chart.csv).

Wenjia used a Python web scraping package to scrape the Official Charts website and saved each week's Top 40 Singles into a CSV file (uk_top_singles_chart.csv). This generated a dataset that has 10440 rows of records and contains the Top 40 popular songs in the UK every single week for 261 weeks in total. Columns include date of the chart, position of the track in the chart, title of the track, artist of the track, label company which produced the track, and ISRC number (International Standard Recording Code) as the unique identifier of the track.

Amy suggested that Spotify also has a Charts page that has daily and weekly charts with a downloadable CSV. This would be a better option as you can get daily top streaming songs as a more accurate indication of what people are listening/streaming on multiple devices. And you can specify the location to cities which would help the project look into different areas within the country. However, the web page was written in such a way that it was impossible to scrape the data. Wenjia did attempt a Python script to automate the process, but it did not work.

- A dataset of audio features of all those popular songs which made to Top 40 singles chart from 27/04/2018 to 21/04/2023 (track_audio_features.csv).

In order to get this dataset, from uk_top_singles_chart.csv Wenjia extracted all the unique ISRC numbers as we do not need to make duplicate API requests on songs which repeatedly appeared in the Top 40 charts. Then, by using the ISRC number from unique_isrc.csv file as the query at Spotify API 'search' endpoint, the Spotify track ID was obtained for each track and saved in a separate CSV file (isrc_spotify_ids.csv). Then, at the Spotify API's 'Get Tracks Audio Features' endpoint, 'track_spotify_id' from isrc_spotify_ids.csv was used to form the query and save all the audio features of a track into track_audio_features.csv. Later on, the classification of the mood of each track was based on these features. Several other APIs were explored to see if additional datasets could be procured from other sources. Apple Music API requires a paid Developer Program account. Deezer API and Shazam API do not store the detailed track audio features we needed for this project. Therefore, only Spotify API was used for this project.

- A dataset on NHS Mental Health referrals from 2018 to 2023.

It was very difficult to get this dataset. Autumn found the NHS Mental Health Services Monthly Statistics Power BI dashboard. The new referrals section of this dashboard provided the figures we needed. Unfortunately, we could not export the data from this Power BI so we had to search for the source data

which had led to these figures. The most promising file was on this link called 'MHSDS Time_Series_data_Apr_2016_JanPrf_2023.csv'. This seemed to provide the data on which the Power BI was based as there is a statement on this link which states "The underlying data has also been made available in CSV format. Please refer to 'MHSDS Monthly: Time Series Data for Selected MHSDS Measures, April 2016 to Performance January 2023 CSV' available under 'Resources'." However, after trying multiple times to count the occurrences of the specific reference given to new referrals (MHS32) and also by filtering this reference and then summing up the correlating values, we could not reproduce the figures on the Power BI dashboard. Due to the time constraints of this project, NHS_manual_data.xlsx was created manually with the data from the Power BI dashboard.

Step 3: Redefine the questions.
We then returned to our ideas documented in the group project folder after the initial meeting. With the available datasets we gathered and having in mind the limitations of these datasets, we were able to narrow down on the two questions we aimed to answer with this project as outlined in the background section.

Step 4: Further processing and cleaning of the data
From track_audio_features.csv, we have 'acousticness', 'danceability', 'duration_ms', 'energy', 'instrumentalness', 'key', 'liveness', 'loudness', 'mode', 'speechiness', 'tempo', 'time_signature', 'valence' columns.We needed to establish a way of using these features to classify the mood of a track. Amy has expertise in music and she developed a machine learning model (a decision tree model) and algorithms (method detailed in ML_spotify.ipynb and Spotify data mood allocation.ipynb) to categorise the tracks into four different moods - happy, sad, calm, and agitated based on Russell's dimensional mood classification model.

For each track, an additional column called 'mood' was added onto track_audio_features.csv and a new dataset track_audio_features_plus_mood_analysis.csv was created ready for the next stage.

Autumn and Isobel then cleaned the NHS data by changing some column data types, ensuring there were no missing values and splitting date column into date and month columns for easier analysis.

Wenjia joined track_audio_features_plus_mood_analysis.csv back together with the original dataset uk_top_singles_chart.csv and created one cleaned dataset popular_tracks_and_moods.csv that has Top 40 popular tracks for each week and the mood of each track.
The two cleaned datasets NHS_manual_data_cleaned_reformat.csv and popular_tracks_and_moods.csv were then handed over to Isha and Vilma for data analysis.

On another note, Wenjia created a SQL database using various CSV files obtained during the process of web scraping and making API requests. This database can be used by people who want to know what song was popular or how many songs of an artist have made it to Top 40 at any given time covered by the period in the original data.

Step 5: Data analysis and visualisation
Vilma analysed moods of the tracks in different years within the track moods dataset and produced visualisation looking into the changes of moods during covid lockdowns. Isha analysed music mood preference, mental health referrals and their correlations between the years 2018-2022 and also produced detailed visualisations. The Jupyter Notebooks were shared with everybody in the group. A few suggestions were made so different kinds of charts could be created to enhance the visualisation. Final versions of the Jupyter Notebooks were then produced and uploaded to the group repository.

1.3 IMPLEMENTATION AND EXECUTION

As we had four team members who are starting roles as data engineers (Autumn, Wenjia, Isobel and Amy) and two team members who are aspiring data analysts/scientists (Isha and Vilma), we felt it was best to play to these strengths. The engineers would work on the collection and cleaning of the data and then the analysts would work on the analysis and visualisation of the data.

We also decided that it was best for someone to act as a Scrum Master/Project Leader to ensure we all stay on track. Autumn took on this role. She created a Notion board for the group and assigned tasks to everyone and gave deadlines for each task. Through the course of the project she reviewed the Notion board and made adjustments where necessary based on the requirements of the team.

This worked well as, at the beginning of the project, we had enough resources to effectively collect and clean the data. However, when it came to the analysis and visualisation part of the project, it became clear that we needed more team members to work on the visualisation in order to meet the project deadline. Especially since we had originally started as a group of 7 and had arranged to have 3 analysts but ended up with only 2. To combat this issue, Isobel and Amy started to take on some extra work in the visualisation section of the project as they felt they had more to contribute to the project in this way, so Autumn assigned them to the visualisation and analysis tasks along with the other data analysts.

This is an example of how our team effectively used good communication to ensure each team member felt they had been heard. We also made sure to regularly check in with each other and to keep everyone in the loop when they couldn't attend our meetings. In doing so, we could resolve any arising issues, such as individual availability and time constraints, quickly and collaboratively, to ensure the best outcome for the time we had to complete the project.

In addition to the Notion board we used to stay on track of the tasks we needed to complete, we also used Google Drive to store all our draft documentation, Google Jamboards to share ideas, Google Collab to work together on writing documents such as the week 2 homework, GitHub to store out final completed project so each team member's contributions could be easily seen.

The Python libraries we have used throughout our project are: Pandas for data analysis and cleaning, NumPy for working with arrays, Seaborn, Matplotlib and Plotly for data visualisation, Scikit-learn for machine learning, Spotipy for the Spotify API, Xlrd for reading historical xls files, Mysql-connector-python for the SQL database andBeautifulsoup4 for data scraping from web-pages.

1.4 RESULT REPORTING

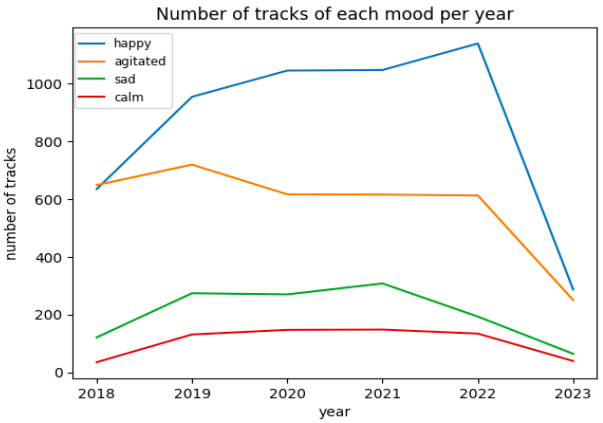
Exploration of the data confirmed our hypothesis - there is correlation between the mood of music preferred by the UK population and the number of mental health referrals.

The initial exploration was performed on the music mood data to observe the changes in music mood over the years and in preference of music mood over time. This exploration showed that:

- Happy music remained the most streamed type of music, but the levels did change pre-pandemic (45.1%), during the pandemic (48.5%), and post-pandemic (52.2%).
- The popularity of sad music increased from pre-pandemic to post-pandemic period, from a streaming popularity of 11.4% to 14%, and showed a steep decline post-pandemic, to 10.3%.
- Music that can be classified as 'agitated' remained the second most popular type of music listened to. However, the popularity of this type of music dropped over the pandemic, from 38.8% to 29.8%, and increased slightly again after the pandemic to 30.7%.

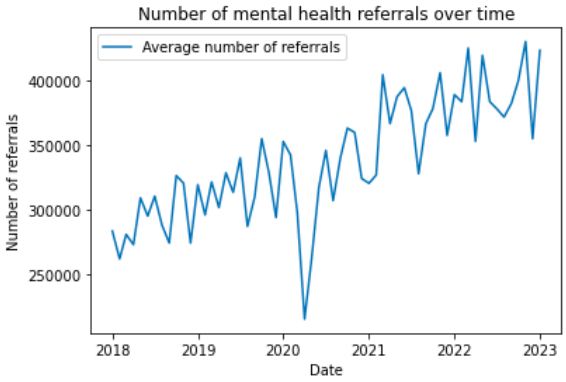
- Calm music is the least listened to type of music, but the popularity of this type of music increased over the pandemic, from 4.7% to 7.5%. Post pandemic, however, the popularity of calm music dropped to 6.1%.

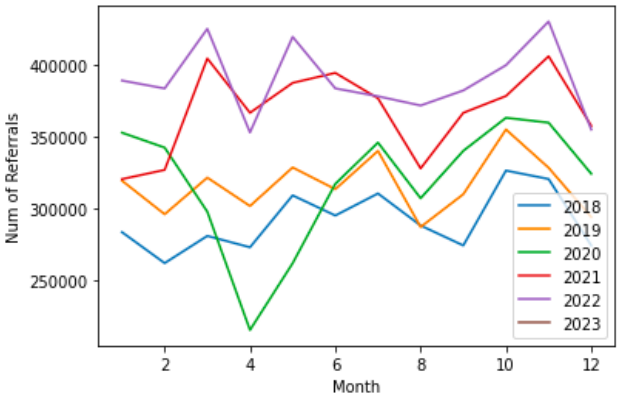
The below graph shows the general trend of popularity of music moods over the years. Note that the data for 2023 is not representative of the full year.



Exploration of the mental health data showed a general upward trend in the number of mental health referrals over time. This suggests that although the pandemic may have impacted the number of mental health referrals, it is more likely that the number of referrals are impacted by a number of external factors that we have not looked at. Since the aim of the research was to look at the correlation between the music mood preference and the number of mental health referrals, the causation behind the patterns noted in mental health referral numbers do not affect our findings.

Our data exploration also showed that the number of mental health referrals have seasonal patterns. This discovery was useful in exploring anomalies in the data, such as the steep valley observed in March 2023. While the direct cause is unknown , this may speculatively be due to the fact that NHS services began to predominantly focus on COVID-19 from this point onwards. In other words , this dip does not reflect fewer people having mental health issues but rather reflects fewer people having access to services.

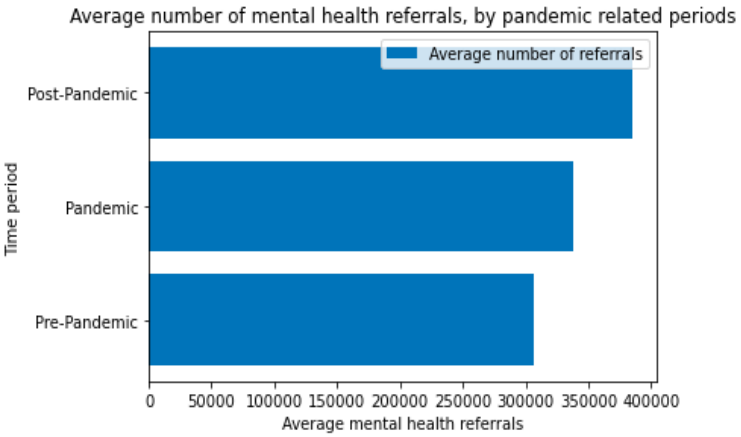




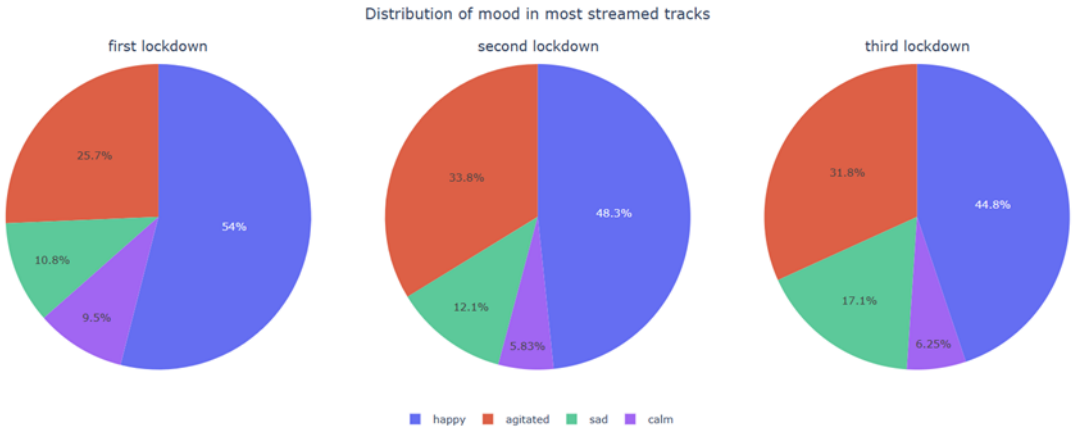
This also displayed tumultuous patterns in the number of referrals after the pandemic season, as opposed to the pre-pandemic years that show clear and predictable patterns. This can be seen in the seasonality chart of the number of mental health referrals above. Even though the post-pandemic period follows some of the largest general trends such as peak periods, it is difficult to extrapolate and anticipate the demand for mental health services by looking at the patterns alone, potentially affecting mental health service providers and their ability to anticipate demand.

Analysis of average number of mental health referrals by pandemic related periods show continuous increase in the number of mental health referrals as well. This is important for mental health service providers to understand. We initially expected that this graph would look different, showing a decrease in

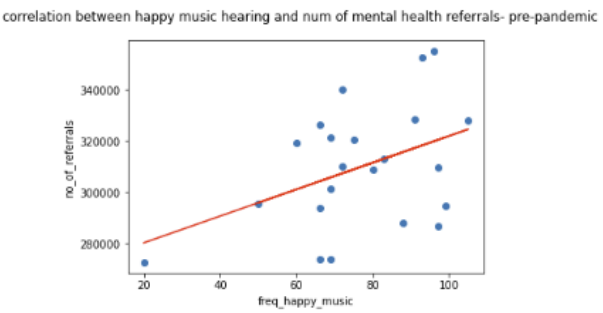
the number of referrals after the pandemic. Further time based data would be beneficial to observe in the seasonality and the level of demand for mental health services decreases in the future.
Correlation exploration of both mental health data and music mood data show promising results and correlations.

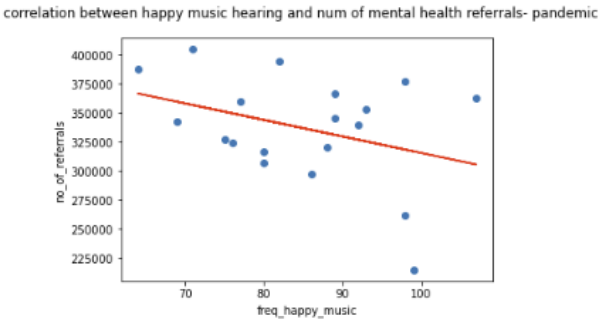


1. The music mood preferences are not only influenced by the general passage of the pandemic period, but also influenced by the events within the pandemic period. Below is a representation of how music preferences changed over the first, second, and third lockdown of the pandemic:

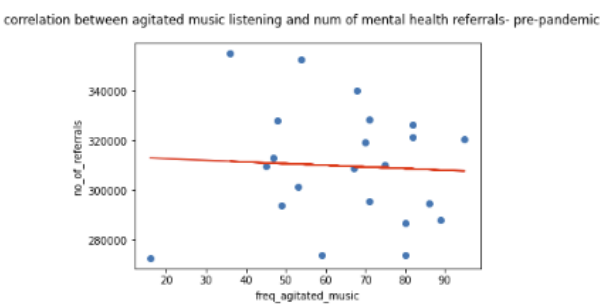


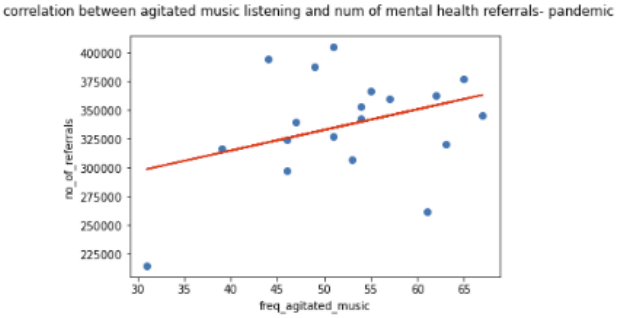
2. Pre-pandemic, there is a positive correlation between the amount of happy music streamed and the number of mental health referrals, but this is the opposite during the pandemic. The lesser happy music is streamed, the higher the number of mental health referrals.



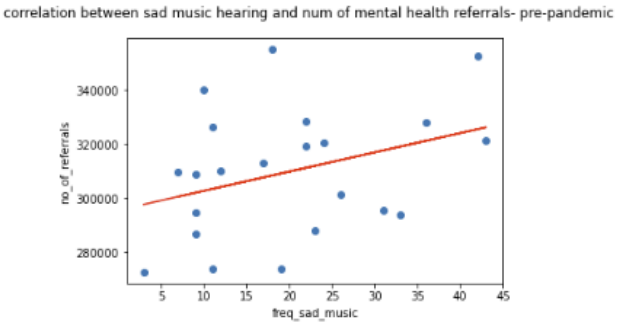


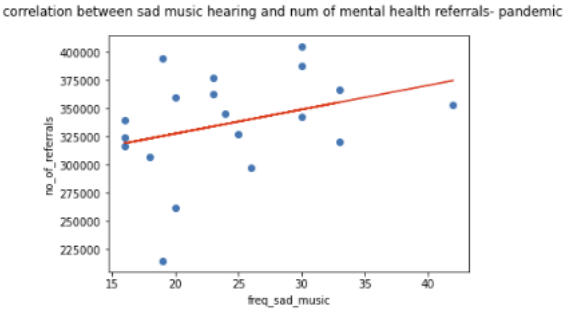
3. There is almost no correlation between the streaming of agitated music and the number of mental health referrals in the pre pandemic period. However, there is a significant positive correlation between the two factors during the pandemic period.





4. There is a positive correlation between the popularity of sad music streamed and the number of mental health referrals, both during the pre-pandemic and the pandemic period.





1.5 CONCLUSION

This correlational data can be advantageous to services such as online mental health regulatory application based services, that could potentially use frequency of type of music streamed by individuals to assess risk level of self harm, for example. The correlations confirm that music preferences classified by mood even on a small scale such as this research can be beneficial in assessing the mental health standing of a population, and possibly of individuals. With sufficient data, it may be possible to build a predictive model that helps mental health service providers equip themselves better for the changes in demand for their services during tumultuous periods of time.

It is also important to consider the shortcomings of this research in the practical application of our project

- Correlation data is limited, as we only have monthly mental health referral data. Weekly or daily data from the NHS would be ideal.
- There is only a few months worth of music data for 2023.
- This data will not suffice to create a predictive model predicting number of referrals based on music mood preference of the population as we do not have sufficient data points.
- Our model of emotion is very limited; given more time we could build a model which incorporates sentiment analysis.
- There are a myriad of other factors that impact the amount of referrals, as the correlations observed here are weak - they do not demonstrate statistical significance.

References:

Misery loves company: Mood-congruent emotional responding to music. By Hunter, Patrick G.; Schellenberg, E. Glenn; Griffith, Andrew T. (2011)

"A circumplex model of affect". Journal of Personality and Social Psychology. (Russell, Lewicka ,1980)