How might we find music that benefits mental health?

Kaushal Gianchandani, Abbie Tolon

First off, kudos on the choice of problem! This is indeed an important issue.

While using data science to tackle any social or healthcare related problem, one has to strike a balance between the size of the dataset and the level of details. Constructing datasets with several details is both expensive and time consuming. Moreover, it often relies on building a small sample from a very large population. These small samples may not be representative of the actual population and an analysis based on small samples can be skewed. Alternatively, large datasets that span a big chunk of the population can be used to draw more generalized insights, however, these datasets may not be detailed enough and can often lead to spurious correlations/conclusions.

Keeping this in mind, we propose to work with data provided by Google Trends, which is a website that provides the popularity of different search terms in different geographical domains. Given Google's reach, it is fairly certain that the information conveyed by this data is representative of a large chunk of the population. Furthermore, the Interest in each search term over time is quantified on a scale of 0 to 100, which makes it convenient to compare between two or more search terms. However, Google's privacy policy does not let us access details such as age, gender, ethnicity, education background and other personal details of the users.

We will use Google Trends data to try and identify the effect of different music genres on mental health. In the subsequent part of this handout, We discuss a piecewise outline for the project.

Task 1: Data identification

To begin this analysis, let us compile a list of search terms that you think are relevant to the problem. You can think of it in terms of two questions:

1. What "search term(s)" can best quantify the mental health of a person? For instance, a person who is in a foul mood for multiple days can google "I am depressed. What can I do to be happier?", the key word in this sentence is "depressed" because that is what we are trying to predict. There can be multiple such search terms and we need to make a list of them.
2. List the popular music genres (a simple google search should be enough make this list) Suggestions: Rock, Jazz, Country, Rhythm and blues, Gospel etc

Task 2: Data acquisition

Once you have identified a bunch of relevant search terms, the next step is to make a shared google sheet (let's call it *"B-Day data science"*) in which you document the popularity of each search term relevant to the problem. Here is an example of how to do it:

1. Go to Google Trends
2. Type in the search term, e.g. 'Anxiety'

3. Change the parameters of the search:
    a. country: United States,
    b. time: 2004 - present,
    c. categories: All categories,
    d. Search type: Web Search.

    Note that the interest in the search term anxiety is at an all time high.
4. Scroll down to the “Interest by subregion” section. This data shows the interest in the search term “Anxiety” during the specified time frame (2004 - present) in different states of the USA.
5. Click the download button to download the data as a .csv[1] file (that can be opened on google sheets/excel). After opening the csv file, you need to sort the entries by name such that Alabama is the first entry, followed by Alaska and Arizona. The list should end with Wisconsin and Wyoming as the penultimate and last entries.
    a. To sort: select both the columns simultaneously, right click, go to sort and click on the option Sort A to Z.
6. After sorting the data, copy it to the file *"B-Day data science"*, where the first column is Region, which is sorted alphabetically and the next column contains the number representing the interest in the term “Anxiety” from 0 to 100 for that particular region.
7. Now you need to repeat steps 1 to 6 for the different music genres.
    a. It is a good idea to explicitly type the word ‘music’ after the genre, for instance, look for ‘Rock music’ instead of just ‘Rock’
    b. Always ensure that you sort the data before copying it to the shared file "B-Day data science".
    c. After we have made the combined table and ensured that all the data is sorted, we can delete the extra Region columns. Further, get rid of the blank columns by moving the data around.
8. Once you are finished, your excel file should look like this (the values quoted below are merely placeholders):

Region	Anxiety	Rock	Gospel	Pop	<other music genre>
Alabama	100	54	26	58	…
Alaska	89	57	31	100	…
…	…	…	…	…	…
Wyoming	97	68	100	75	…

Task 3: Visualize the data and draw insights

Based on the data you have collected, we will do the following:

1. Calculate correlations between the possible cause and probable outcome(s). For instance, we can check the popularity of the term “Anxiety” in regions where “Gospel music” is more popular by calculating the correlation between the interest in these two search terms.

[1] csv: Comma-separated values

a. If the number is close to 1, we can say that regions with high interest in “Gospel music” are associated with increased interest in “Anxiety”.
b. If the number is close to 0, we can say that high interest in “Gospel music” has no association with interest in “Anxiety”.
c. If the number is close to -1, we can say that increased interest in “Gospel music” search terms is associated with a decrease in interest in “Anxiety”.

2. After calculating the relevant correlations, we will use these to draw insights regarding how interest in a particular kind of music can be indicative of a person’s mental health.

3. ***[If time permits]*** After identifying the terms with high correlations, we will build a linear regression model to identify the level of anxiety based on the interest in music.

This objective analysis will serve as a basis to answer two questions:
1. Does a particular type of music have any long term effect on our mental state?
2. What kind of music is indicative of good mental health?

Task 4: Identify the limitations of our approach

To conclude our analysis we will try to identify the limitations of our approach. For examples, we encourage you to think about the validity of this statement cited in the second paragraph of this handout: “*information conveyed by this data is representative of a large chunk of the population.*” According to you, is this statement true or false? Give your reasons to support your answer.

Another limitation to note is that we cannot definitively conclude why a person uses a search term. For example, a person might search the term “anxiety” because they are writing a research paper, or they simply want to learn more about the topic, rather than because they themselves have anxiety. Moreover, when comparing the trends between anxiety and music genres, we cannot assume the two are actually related without additional analysis and without controlling for other variables. Instead, this can serve as a good primary analysis for generating additional questions for future research.

Suggestions for your project

1. Instead of showing a bar graph which shows the interest in different kinds of music for all the regions in the US, you could simply show a pie chart that shows the interest in different kinds of music for the state of Utah. Here are the steps to do so:
    a. Make a new sheet in the document “B-Day data science” by clicking on the **+** symbol on the bottom left corner.
    b. Copy Row 1 from Sheet 1 to Sheet 2 as the first row.
    c. Copy Row 46 i.e. the Row corresponding to Utah from Sheet 1 to Sheet 2 as the second row.
    d. Go to Sheet 2 and delete the columns corresponding to “Anxiety”
    e. Select the two rows and draw a chart by clicking on **Insert** and then **Chart**.
    f. You can then change the type of chart by choosing a Pie chart as the **Chart type** in the **Setup** section on the right hand side.

g. You can also tidy up the labels on the chart by simply clicking on them and editing the labels. Google sheets is quite interactive and it can be a fun learning exercise for all of you.

More data sets to work with:

[Note to KG: Insert information on multiple data sets below]