Does a BU Students Major Impact the Variety and Amount of Music they Consume in the Year 2022

Arya Daryanani

Link to Code:

https://colab.research.google.com/drive/1RqmeWuYPM1-aIyx9ZU4iiZ9f3JjYhaNn?usp=sharing

Link to Results:

https://docs.google.com/spreadsheets/d/1RhfFzLSITFVViL6LHmzarlh6t_fEiGc2MXPKvWHTUdM/edit?usp=sharing

Introduction

Stress is something all students will encounter during their experience. No matter what one's academic level or standing might be, stress is something that creeps its way when a student is left to face an overwhelming task with high stakes. Being a university student at Boston University is no exception to this, where there are many times students must face challenging moments to persevere through their higher education journey. However, desiring to test the degrees of these stress levels, as STEM focused majors, we wondered if STEM degrees typically felt more pressure in comparison to non-STEM students.

Upon discussing this issue, we both concluded that while there are many ways to relieve stress healthily, listening to music can typically become the most convenient method, and as long as someone has a working device they can easily grant themselves a release of stress. Curious, we discovered that music is able to reduce stress in a variety of ways. Listening to music, keeping in mind that it is enjoyable and according to personal preference, affects the body's stress hormones (adrenaline and cortisol) by reducing them. As well as increases the production of dopamine, which reduces emotions of depression or anxiety. Another impact music has on the body, using a scientific and strategic approach, is listening to music that follows approximately 60 beats per minute, which can force the brain to synchronize with the beat and leave it feeling calmer.

Applying our newfound knowledge on the matter to the project we produced two hypotheses. Firstly, we believe that students who are avidly pursuing a STEM degree would favor listening to not only more music but also music that is sounds intenser than those who declared a non-STEM major. Secondly, students who rank their stress level on the higher end of

the spectrum consume more music in comparison to those who claim to be managing lower stress levels.

Methodology

The first thing we did was send out a survey collecting results, this was posted on the class's Piazza and on the BU graduating class of 26's snapchat story to accumulate results that could potentially reach non-STEM students.

After a significant amount of time, we began observing our results. Firstly we looked at the spreadsheet directly, only to become overwhelmed by our responses and not be able to draw any conclusions by facing the unorganized spreadsheet.

Beginning the project we downloaded and imported the spreadsheet as a CSV file using the pandas library, as it was the format and library we were most comfortable with using. The next plan of action was to create a sense of cohesiveness throughout the data we collected, especially regarding the fields we planned to utilize later. This proved to be a crucial aspect as our survey accepted personalized user responses for a selection of questions, where many people could have the same response but because of capitalization and extra punctuation, they would be documented as completely different values. To instill a sense of order, switching all letters to uppercase and ridding of unnecessary punctuation for the fields 'Field of Study' and 'Top Genres' ended up with different values being separated by a space to avoid any confusion. Another thing that was a common occurrence with responses was the use of commas for the 'Minutes Listened' field, which needed to be replaced with nothing as this would not allow the values of the row to be treated as an integer but a string instead.

Next, we had to assure that retrieving specific data and differentiating whether or not the response was from a STEM student was efficient and easily adaptable, thus, the use of functions. We wrote two functions that could be called based on the field passed through the parameter's data type, as well as splits and returns the row's results into two separate lists (STEM and non-STEM) to be used when needed. Both functions have similar methods, differing only by the typecast on values getting appended to the list. The methods adapt the run-time complexity of O(n) as the relationship of the list growing is in terms of n, where n represents the number of times the data frame is looped over.

Using the scipy.stats library, we ran a T-test on a list consisting of total minutes listened to by STEM students and a list consisting of total minutes listened to by non-STEM students to determine whether there is a difference in music consumption between the two groups.

The next step was to create three separate word clouds to get a generalized idea of the most popular responses from students. By utilizing the word cloud library and especially utilizing the WordCloud, STOPWORDS and ImageColorGenerator functions The first word cloud consists of all the responses and the other two being divided by STEM and non-STEM students respectfully.

We then ran two chi square tests using the scipy.stats library, again, to test if the stress levels of a student was independent of the heaviness level of the music listened to amongst STEM and non-STEM students. Running these allowed us to test our second hypothesis of whether or not students experiencing more stress would lean to consuming intenser music genres.

To gain a better understanding of our findings and the responses, we used various graphs to visualize the relationship between the amount of music consumed, the level of stress, and the heaviness of music listened to, in order to further seek any correlation between the fields present. We used functions from the sklearn.linear.model library and matplotlib.pyplot library to plot a linear regression model. Assigning the y axis a scale that started at 0, end at 300000, and incremented by 50000 to represent the 'Minutes Listened' field, as well as assigning the x axis a scale that started at 0, ended at 5, and incremented by 0.5 to represent the 'Stress Levels' field. This was made as a scatter plot, where we then used linear regression to depict the line of best fit amongst the responses in order to observe any correlation–negative or positive–and its strength. Judging by the behavior of the data points surrounding the line, we then used more built in methods of LinearRegression from sklearn.linear model, alongside the sklearn.metrics library's r2_score function to find the linear coefficient and r2 correlation score to further analyze the data.

The next graphs we plotted employed usage of the seaborn library. The first graph amongst this selection was a cat plot in bar form; assigning the x and y axis a scale that started at 0, ended at 5 and incremented by 0.5 and 1 respectively, the x axis belonged to the "Music Heaviness" field while the y axis belonged to the "Stress Levels" field. The cat plot averaged the stress levels per music heaviness according to each stress level. We opted for this plot type because seaborn's catplot utilizes a 95% confidence level, which helps us as data scientists understand which data points have a 95% chance of falling within a given range, and organize our data into outliers and normative points. The data provides "error bars" which help data scientists understand the outliers in a given dataset. The next two seaborn graphs we used were a joint plot and a 3D scatterplot to observe the relationship between the three fields of 'Minutes

Listened', 'Stress Level' and 'Music Heaviness'. Using seaborn's technology to our advantage, we depicted higher levels of music heaviness with darker hues. We set our x-axis to extend from 0-5, since the highest one could rate their stress was 5, and our y-axis had a range from 0 to 350,000.

Finally we found the correlation between the total minutes listened to and the total stress level using the numpy library to determine whether or not the fields were reliant on one another.

Results

Analyzing our t test data, we ended up with a p value of 0.5943532889416374, which ends up being less than 0.05. Thus, concluding that there is no relationship between the quantity of music consumed and whether or not the student was a STEM major. This implies that a person's minutes consuming music is solely reliant on their lifestyles, hobbies and preference as non-STEM majors listened to approximately the same amount of music STEM majors listened to.

Observing the word clouds, in particular the two separating responses and preferences of STEM and non-STEM majors; similar genres were reoccurring, such as pop music reigning over as the most popular amongst the two categories.

Conducting the chi-square tests were to draw out any correlations between stress levels and music intensity consumed amongst students of different majors. For STEM majors the p value was 0.8241363164424579, and for the non-STEM majors the p value was 0.8712628916824271. While both results ended up having somewhat similar values, they did exceed 0.05, thus showing that there was ultimately no correlation between the fields whether or not the student was a STEM major, and that the stress levels and music heaviness are completely independent from one another.

Taking a closer look at the graphs we were able to draw several conclusions from seeing a visualized representation of our findings. Firstly, to further determine the relationship between total minutes of music consumed and the level of stress we ran a linear regression model through our data. We ended up with the linear coefficient of -1236.70588235 and the r2 score of 0.0005524843156681225. The linear coefficient we yielded indicated that there was a negative correlation between the two fields, thus, allowing us to understand that as the amount of stress increases, the amount of music consumed decreases. The r-squared value, which determines the strength of the relationship, being the rounded value of 0.0005, indicated an extremely weak relationship between the two fields. This means that despite there being somewhat an inverse

relationship between the minutes listened and the stress levels, there isn't any strength to it, so anomalies can vastly occur. By also graphing the line of best fit, we were able to reinforce and further prove the negative correlation with its decreasing slope. Alongside the lines many outliers which again further proved the almost non-existent relationship between the two fields. For our cat plot that drew the relationship between the stress levels and music heaviness, and despite each field being independent of one another, when plotting all of the results and not differentiating them by the students major, we noticed that the most intense music ranked as a 4 was consumed by students who ranked their stress level a 2. While students who ranked their stress level as a 1 typically consumed the least intense music at a ranking of approximately 2.25. While it may appear that there is a pattern here we believe that this trend was a pure coincidence as proven by our chi test the fields are independent from one another.

When observing for any patterns using the joint plot and 3D scatter plot we were able to justify that despite any patterns that may have appeared when comparing two fields together, in the large scale of responses each one was sporadic.

Our correlation between the minutes listened, stress level and music heaviness we landed with the values [-1.40479654e-06 1.35308988e-01] 2.6500765858792623 array([[1. , -0.02350498, -0.06776877],
[-0.02350498, 1. , 0.12047043],
[-0.06776877, 0.12047043, 1.]])

This further consolidated our previous findings that there is no relationship between stress and the amount of minutes listened to, since the first coefficient is less than 0. However, the other coefficient is greater than 0 which also implies that there may be a relationship between stress levels and music heaviness as witnessed from comparing the chi tests and cat plot results. The intercept implies that when the values of the amount of music consumed and stress levels reach a joint 0, students will still lean towards listening to heavier music.

Conclusions

Since we aimed to find statistically significant data, we formally hypothesized that the average music consumed between STEM and non-STEM students would be vastly different. However, the lack of correlation and insignificant p-value fails to reject the null hypothesis. As a result, all our data points to the notion that there is no significant difference between the average minutes of music consumed between the two sample populations. For our second hypothesis in which we tested whether one sample group (STEM students) was more stressed than the other

(non-STEM students), we also did not find any statistically significant evidence. As a result, we fail to reject this null hypothesis and can conclude that both sample populations do not have a significant difference.

While our data failed to reject our null hypotheses, it is not a strong indicator of the true mean. Out of an estimated population of 16,872 undergraduates, only 37 responded to our survey. Our sample set reflects 0.002% of our population, which is insignificant. Since our dataset was so limited, it may have skewed our results and showed a biased result, which is not a reflection of the true mean. To overcome this, we need to increase our sample size and run multiple tests to ensure that our results stay consistent among our control groups.

As data scientists, we can try to circumvent such issues in the future by expanding our reach and promoting on various platforms, specifically those catered to the entire BU community. By using platforms such as Piazza and Snapchat, we are inadvertently using a biased sampling method, since a majority of the population is not present in these communities.

Although our results have proven to be statistically insignificant, we learned that BU's students' music tastes share a variety of similarities despite their field of study and that the academic rigor of the institute applies across all colleges. We also came to the realization that visualization methods increase the accessibility of our data and conclusions exponentially, as it is easy to decipher for the common person. Adopting methods of colorful, informative visualization breaks the silos between data and the public, and helps us as a people understand each other and our similarities better.