Sean Ryan

DS121

Final Project

**Introduction**

In this project, I observed a dataset featuring different songs on Spotify and performed a regression analysis to find the R2 value for different characteristics of the songs to see if there was any correlation between those characteristics and the song’s popularity on the platform. The characteristics I tested for were tempo, danceability and instrumentalness (frequency of vocals).

**Data**

My project features a dataset containing more than 114,000 songs that can be found on the music platform Spotify. The dataset also gives ratings for different elements of songs, such as tempo or energy. For this project I focused on comparing the popularity rating for each Spotify track to 3 different other variables from the dataset to test for correlation. The popularity rating is a score given to each song between 0 and 100 where a larger number means a more popular song. The score is based on the total number of plays and number of recent plays it has on Spotify. The variables I tested the popularity score against were; tempo, danceability, and instrumentalness. The tempo attribute is the average beats per minute(bpm) of the song. The danceability attribute gives a score for each song between 0 and 1 where a larger number classifies a song as more “danceable”, as the creator of the dataset described, “based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity”. Finally, the instrumentalness attribute also gives a score for each song between 0 and 1, but in this case, a larger score means a song is more likely to contain no vocals in the song.

**Methodology**

For the data I chose, my plan was to run an R2 test to test for correlation between popularity and the other three chosen variables: tempo, danceability and instrumentalness. I did this by creating an algorithm and running for each of the three variables against popularity to see if any of these variables had a consistent effect on how popular the song is on Spotify. The algorithm I created finds the residual sum of squares and divides it by calculating the total sum of squares and subtracting that from 1. My hypotheses for these tests are that songs that have higher danceability will be more popular, songs with a lower instrumentalness score will be more popular and tempo will not have an effect on popularity of the song. My goal is to see if any of these variables possess significant correlation to popularity.

**Analysis**

*See FinalProject.ipynb*

**Results**

After performing the analysis, I was unable to prove that tempo, danceability or the instrumentalness of a song has an effect on its popularity on Spotify. All three tests resulted in an R2 value nearly equal to zero, indicating that there is no correlation. Therefore the hypothesis I was able to reject is that there is a definite correlation between tempo and popularity of a song. However, because of the results of the analysis, it was not possible to determine that danceability or instrumentalness had any effect. On the other hand, the algorithm seemed to function as expected since when calculating R2, the line of best fit for each of the graphs depicted a line with a nearly flat slope so there seemed to be a widespread of attributes when comparing them to popularity.

**Conclusion**

In this experiment, I tested for correlation between popularity and three other variables: tempo, danceability and instrumentalness from a dataset containing tracks on Spotify. Originally, I hypothesized that danceability and instrumentalness would have an effect on the songs popularity since these scores were calculated from observing critical elements of songs such as the rhythm and vocals. However, performing this experiment led to results that cannot determine if there is any significant correlation between these variables. The results produced from this test may be a sign of popularity taking into account far more factors than the ones I tested for and therefore a wide range of tempos and rhythms in songs can be popular or unpopular.

**References**

Dataset: <https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset>

<https://www.investopedia.com/terms/r/r-squared.asp>

<https://colab.research.google.com/drive/1a8rRbi33dQ1BE9EqFt1kJbrrTdJ00j7u>