Spotify Data Analysis

Paul Galli

2023-12-18

Introduction

This report is based on a Spotify dataset containging over thirty-thousand songs. This dataset has been gathered from Kaggle. In this dataset there are multiple variables. These include, track name, track artist, track popularity, and tempo. There are many different variables in which we will examine. I have chosen this dataset becuase of my strong passion and interest of music. Specifically the questions I answer in this report are which genres contain the most songs. This result was found using a bar graph which helps visualize the number of songs in each genre, what is count of popularity scores in the datset, this was done using a histogram in order to vizualize the amount of each score and to discover whether there are more songs with lower scores or higher scores. We will also discover if there is a normal distribution amongst the scores. We will also view a line graph which will tell us how the popularity of a specific genre has changed over the years. There is a scatter plot which will compare tempo and danceability. Does tempo have an affect on danceability?

Methodology

The dataset was found on Kaggle and there was no need for any preprocessing. The process for analyzing the count of popularity scores was done by counting the amount each score appears in the set and creating a histogram which vizualizes the results. In order to find the results for how many songs each genre has in the set, I found the number of songs each genre contains. This allowed me to create a horizontal bar graph in which the x-axis is the genre and the y-axis is the number of songs. When creating the line graph I needed to do a small bit of preprocessing. In the original dataset when we look at the album release date, it is given in Year-Month-Day format. In order to just contain the Year I preprocessed the dataset to give me only the year rather than the specific date. This allowed me to create a line graph containing each genre, where each line represents a different genre. When we think of music we also think of dancing. One thing I wanted to discover was whether or not tempo has an affect on danceability. I created a scatterplot using data from both tempo and danceability to

see if there are any relations. The final discovery in this project is to see if certain variables have an impact on popularity. Are there certain variables that can make a song more popular? These results were found using a multiple regression model.

Results

Histogram of Popularity Score

Here we can review the results of the histogram. As it appears majority of the scores in this dataset are actaully quite low. Majority of the songs in this dataset have a popularity score less than twenty. The average score is between fifty and sixty. These results may be determnined by the amount of songs in each genre. If the genre that contains the most songs, has many songs with low popularity scores then those results can have a huge impact on the results of the histogram.

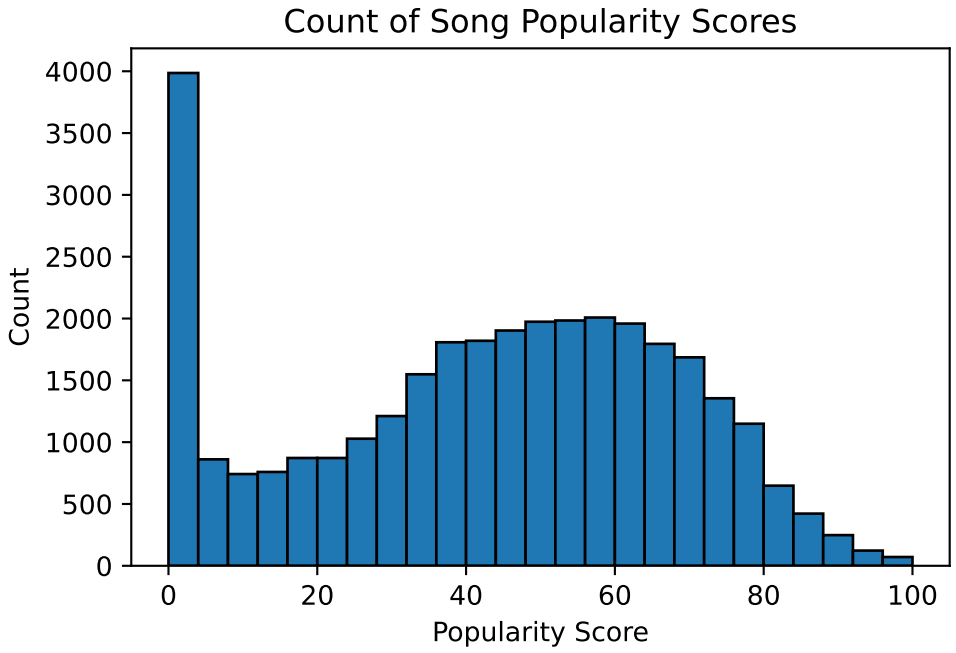

Figure 1: Histogram of Popularity Count

Bar Graph of Song Count for Each Genre

In this dataset there six different genres. Rock, Pop, Latin, Country, EDM, and R&B. My prediction for this was that "pop music" was going to have the most frequent results. That is

becuase “pop music” is short for “popular music”. This is music that considered radio friendly and it’s name is derived from music that is popular amongst listeners. We can see these results are not true. It appears that “EDM” contains the most songs per genre. I believe these results come from many “EDM” songs feature popular singers and artists.

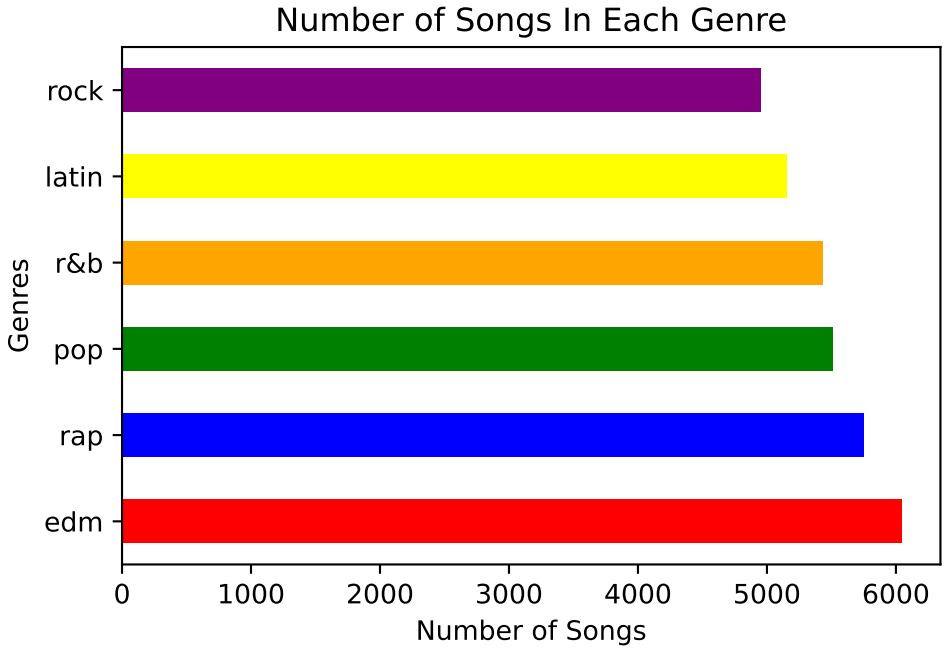

Figure 2: Bar Graph of Song Count in Each Genre

Line Plot of Genre Popularity by Year

Music has been around for many centuries and will continue to be for many more. In this dataset the earliest recorded data is from Nineteen Fifty-Seven. Despite only being a Sixty-Six year difference between now and then, music has changed and evolved over the past six decades. It is intriguing to examine the popularity of each genre. Starting from the earliest date we can see that rock is the only genre to appear until the mid Sixties when “Latin” first appears in the data. The latin songs in the mid-sixties appear to have lower popularity scores with a rise in popularity through to the mid-seventies. The first recorded rap songs show some of the highest popularity scores and then takes a huge dip in the early-seventies. The same can be said about pop music. We see the trend starts with high scores before taking a dip in the mid-seventies. The EDM songs that were made in the late-seventies appear to have low popularity scores. The EDM songs that were produced in the early eighties appear to be the most popular.

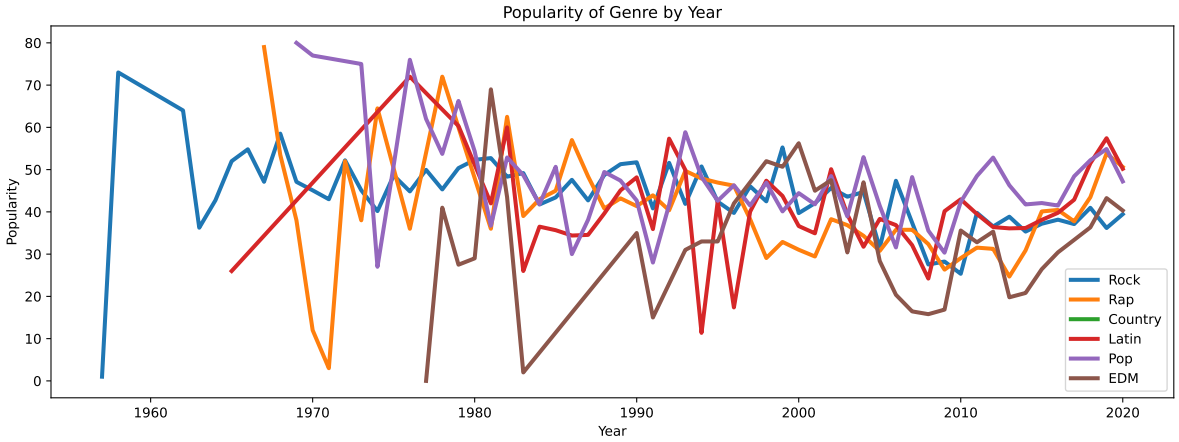

Figure 3: Line Plot of Genre Popularity by Year

Scatter Plot of Tempo and Danceability

When we think of songs that we can dance to, we ususlly think of fast paced songs or even slow songs depending on the type of dancing. The goal of this test was to see if the tempo of a song did have an impact on the danceability of the song. When viewing the scatter plot we can see that songs with tempo's around 120-130 appear to have the most variation in danceability. There is no postive relation in tempo and dancebaility. In fact there is no realtion between tempo and danceability. This does not come as a suprise becuase songs that are slow can be easier to dance to then songs that are faster and vice versa.

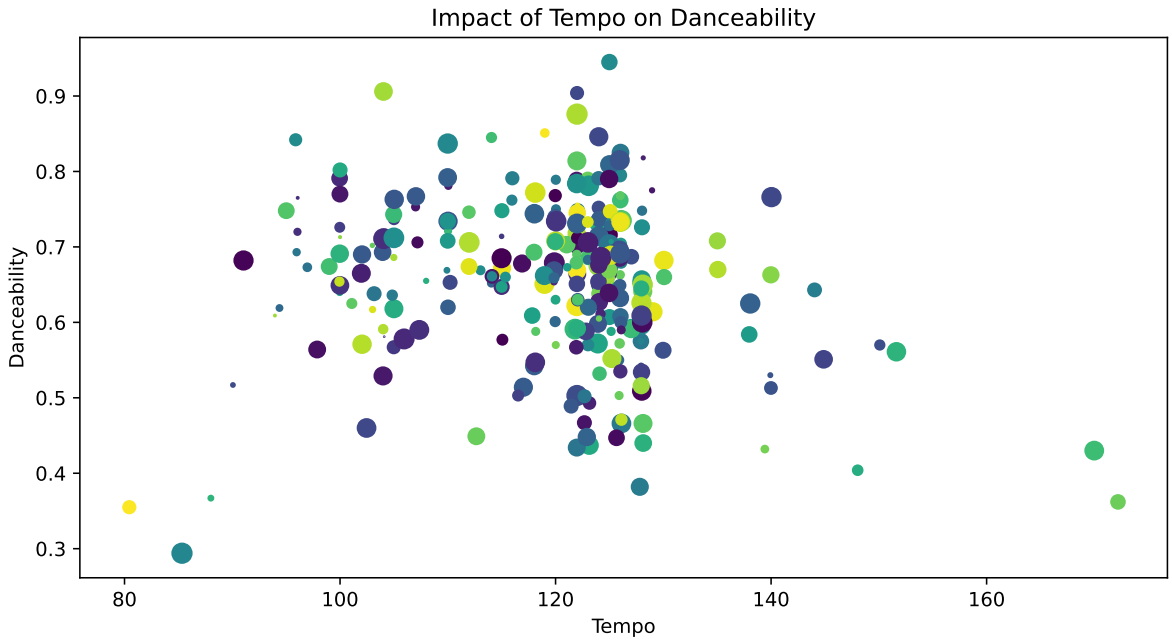

Figure 4: Scatter Plot of Tempo and Danceability

Conclusion

Overall in this project there were some very insightful findings. The interesting increase and decrease of genre popularity throughout each decade, results concluding that this dataset contains majority of songs with low popularity scores, EDM being the genre with the most amount of songs and discovering that tempo does not have an impact on danceability. Some limitations for these results are that the popularity scores are based on when the data was recorded. As time goes on a new or updated dataset should be used in order to produce more accurate results.