Alex Min, Jon Davis, Sahir Doshi, and Veda Patel
BA222 Project 1
Leder-Luis
29 September 2022

Project 1: Analysis of Spotify Top Chart Songs of 2022

Our data set is a publicly available list that analyzes the most popular Spotify songs of 2022 as they appear on Spotify's official Top 200 global chart. All of our data comes from https://charts.spotify.com/charts/overview/global, and has been organized by Kaggle.com member, Sveta151. It was collected using BeautifulSoup, a Python package to parse through HTML documents. This data set was last updated on September 4th, 2022. We chose this data set because we agreed that music is something that almost everyone enjoys and everyone has different preferences. This variety of preferences could lead to a compelling analysis of peoples' tastes.

The data consist of several variables to determine the musical attributes that tend to resonate more with the public and garner popularity. The variables included in the dataset are peak rank, weeks on chart, danceability, duration, energy, key, loudness, mode, speechiness, acousticness, liveness, and tempo. Most of these variables are numerical with a few exceptions, such as mode which categorizes songs as major or minor scales, and key. Due to the immense popularity of TikTok and its ability to create a constant influx of new dance trends, we decided to focus largely on the danceability and duration of the songs in the dataset. TikTok is an app and social media platform that allows users to post short clips of themselves and their friends dancing to any song for up to 3 minutes, though most trends tend to be less than a minute. As a result, we have seen many songs, both old and new, skyrocket in popularity as users invent new dance routines and styles. Correspondingly, we hypothesized that higher danceability and shorter duration are major factors in songs becoming popular.

Spotify measures danceability by analyzing multiple aspects of a song such as the tempo, bass strength, and stability. Songs are given a rating between 0 and 1 based on these factors. This value is indicative of how easy it is for the average person to dance to the song, with 1 being a maximally danceable value. The duration of the song is measured in milliseconds within the dataset, so we opted to convert it to seconds to conduct a more convenient analysis. We proceeded to use these values in our comparison with popularity, which we measured with the variable peal rank.

Summary Statistics

Our data contained a total of 647 observations. Below are the summary statistics of the relevant features (peak rank, danceability, and duration).

	Peak Rank	Danceability	Duration
Maximum	200*	0.985	613.0272 seconds (10.21712) minutes
Mean	65.92	0.674	203.62986 seconds (3.393831 minutes)
Median	51.50	0.70	196.3884 seconds (3.27314 minutes)
Standard Deviation	56.96	0.151	54.924 seconds (0.9154 minutes)
Range	199	0.792	576.09 seconds (9.6015 minutes)

**Though this value of 200 is the max, it actually represents the lowest popularity. The range for peak rank is 1-200, with 1 being a song that was the most popular song on Spotify at a given time.*

Analysis

In comparing peak rank with danceability, we found a correlation of -.01, meaning that there is almost no correlation. This can be visualized in Figure 1, a scatterplot that places the peak rank of a song on the *y*-axis and its danceability value on the *x*-axis. If the danceability of a song does make it more popular, we would see a much clearer linear relationship with a negative slope in this graph.

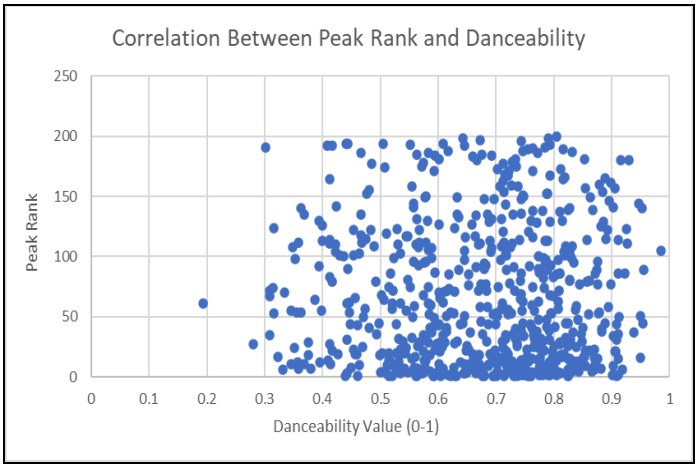

Figure 1

This lack of correlation is observable within the dataset due to some very obvious contradictions to our hypothesis, such as *Intro,* by popular Hip-Hop and R&B artist Drake, which peaked at 61 on the chart but only carries a danceability value of 0.193. This is the lowest danceability value among the dataset, yet the song still had a peak rank very close to the mean. This is evidence that peak rank does not seem to correlate with danceability and a song's popularity is perhaps more dependent on circumstance or the prestige of the artist. However, despite this lack of correlation between peak rank and danceability, it is noticeable that almost all of the points in Figure 1 are shifted to the right, with only a few exceptions falling below a danceability value of 0.3. A histogram explores further into this revelation that may indicate some truth to our hypothesis.

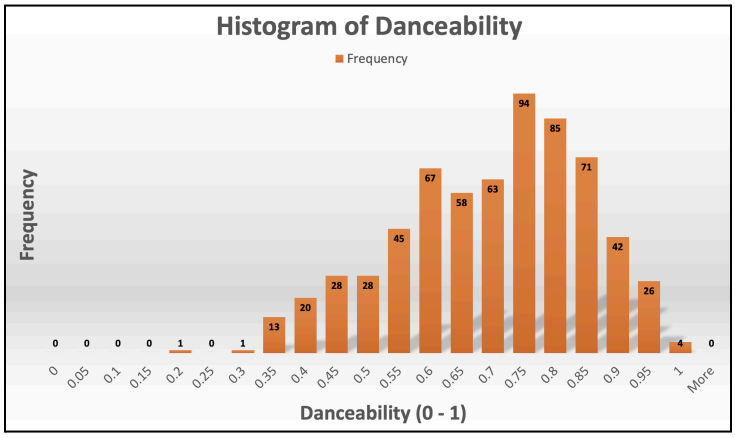

Figure 2

In figure 2, the histogram is skewed to the left, demonstrating that most songs in Spotify's Top 200 charts have relatively higher danceability. If the distribution were normal, and danceability had no effect on the popularity of a song, the mean of danceability within the dataset would be approximately 0.5. However, a vast majority of songs in Spotify's Top 200 carry a

higher value than this. Therefore, although there is no correlation between a song's *peak* rank and danceability, there is a noticeable relationship between danceability and *general* rank. Every song on the dataset was popular enough and received enough recognition to be one of the 200 most popular songs at a given time. Therefore, without a dataset that consists of *all* songs rather than the most popular songs, we cannot render our hypothesis invalid. A more clear correlation might be observable if the dataset consisted of Spotify's Top 1,000 songs rather than Top 200.

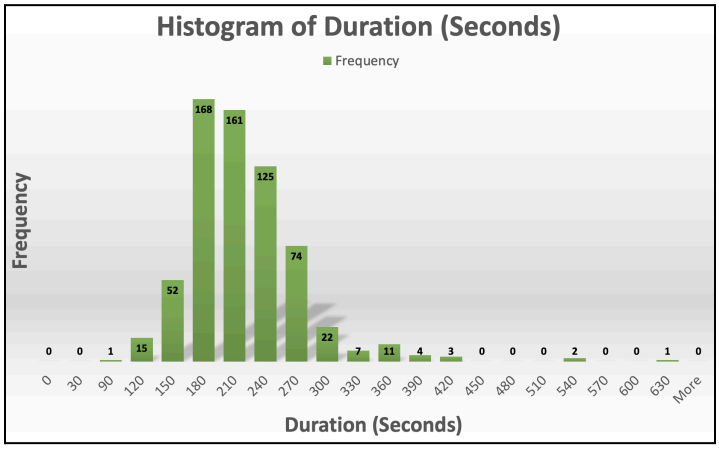

Figure 3

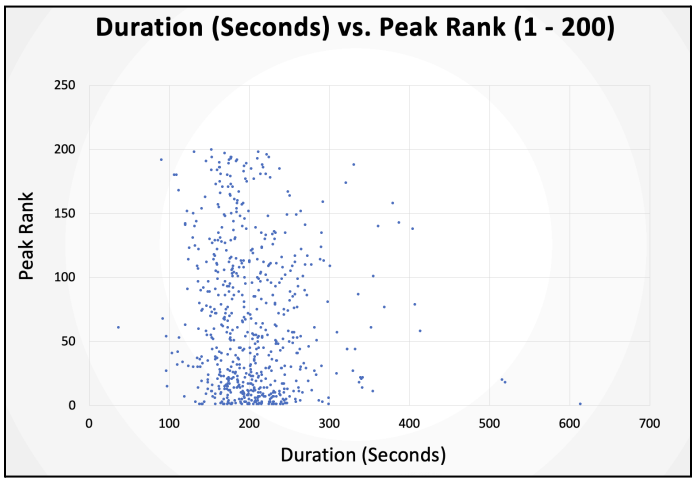

Figure 4

Figure 3 depicts the duration of all the songs on the chart. The chart is skewed right with most songs falling around 150 seconds (2.50 minutes) to 270 seconds (4.50 minutes). Although this histogram may seem indicative of a correlation between a song's length and its popularity, it is mildly misleading due to outliers, such as Taylor Swift's *All Too Well.* The song, which lasts for 613.0272 seconds (10.21712 minutes), pulls the mean, resulting in a histogram that appears to favor shorter songs. Figure 4 displays a more accurate image of this concern, as we can see that almost all data points fall between 100 seconds (1.667 minutes) and 300 seconds (5.00

minutes), without being more condensed around any specific duration value. The correlation between these two variables is almost nonexistent. If our hypothesis that shorter songs are becoming more popular was correct, then there would be a noticeable positive slope in Figure 4.

This dataset gives quality insight into the correlations that we are looking for, because it comes directly from Spotify. Because Spotify is one of the most popular music streaming platforms, the sample is an accurate representation of all popular songs. For the purpose of simplicity, we deleted two variables, time signature and instrumentalness, and ignored several others. In doing so, we did not change our results, however, we did open the door for further questions. Could time signature have an effect on danceability? Does this effect have any correlation with the peak rank of a song? What would happen if we analyzed the weeks on the chart variable as opposed to the peak rank variable as a measure of popularity? We believe this dataset could lead to many further discoveries regarding the qualities of a song that make it popular.

After analyzing this dataset, we found that our hypothesis is **inconclusive**, as there are more variables in play that affect a song's popularity. Though some songs might be more deliberately and successfully constructed for TikTok and its trends, there is no noticeable trendline between either the danceability or the duration of a song and its peak rank. Peak rank may be much more significantly impacted by circumstance, exemplified by the aforementioned *Intro* by Drake. The song may be completely undanceable and unusable on TikTok, but it was the first song on a new album by one of the most popular artists of this generation. Therefore, we can say that TikTok may not have as much of an impact on the popularity of songs as we initially believed, or that the danceability and duration of a song are not the main qualities of trendy TikTok songs.