What's in a Song?: Song Popularity and Auditory Characteristics

By Emily Bansemer and James Doucette

1. Introduction

What makes a song popular? This is a question that has both clear financial incentives -- giving you the edge on creating the next big hit -- but also has interesting implications into psychology and to answering the question of why we enjoy about music in the first place. To go about answering this question we utilized the Spotify API and a dataset of 260,000 songs from 26 Spotify genres to see what musical features correlate with popularity. We wanted to not only look at obvious features like genre, year and artist that already have intuitive and well established links to popularity. Rather we wanted to examine underlying compositional and audio features, like key or danceability, and their more subtle relationship to song popularity. We also wanted to analyze how the relevance of these features changed depending upon genre. What characteristics might define popularity in already popular genres? And are popular features typically characteristic of the style of the genre? Our hope was to find subtle features that drive popularity and influence how much people want to listen to a song. With this preliminary, correlative analysis, we might begin to probe these interesting psychological questions about what drives our music taste.

2. Literature Review

We found several articles interested in the same question as us, and will discuss two of them in this paper. The first of these articles is "What makes a song likeable?". It analyzes many of the same features as us, but unlike us, it was interested in the relationships between all of the variables. They found positive relationships between danceability and valence, between energy and loudness and some others. Another significant difference is that their dataset was just the 100 most popular songs from 2017. Their results were that these popular songs had high danceability and energy, and low acousticness, liveness and speechiness.A problem I see with these results is that there is no comparison with less popular songs. The fact that the most popular songs have low acousticness could just be indicative of most songs on Spotify having low acousticness, rather than any relationship between acousticness and popularity. From the way their results were gathered, there is no way to tell.

The second of article is "Why Do Some Songs Become Popular". Their analysis differed from ours in two key ways. Firstly they were focused on what distinguished the huge hits, rather than general trends with popularity. Secondly, they analyzed the lyrics of the songs rather than the audio features, looking at things like which words and topics were included. Their results were that the very top songs had lyrics that were noticeably different from typical songs of their genre, but not the most different. While fresh, they also remained somewhat familiar. Additionally, they found that the strength of this depended on genre. For genres like pop and dance music, where lyrics are traditionally less important, the lyrics were less different from typical songs.

3. Models and Methods

The first step in our process was data collection. To do the analysis we wanted, we needed a large group of songs and all of the features associated with them. Spotify makes it somewhat difficult to get large lists of song ids, but we found a dataset containing 260,000 songs. The only problem was that it was missing some of the features we were interested in. However, getting the features from a song id is a straightforward retrieval task, just one that takes a lot of time due to limits on API calls.

The next step in the process was to clean the dataset. The dataset contained duplicates of several songs, so we removed all but one instance of songs that had the same release date, artist and title. We noticed that many things were weird about averages for the year 2020 -- it turned out that our dataset had just one song from 2020, so we removed this. Some of the features contained the information we wanted, but not in the correct form. The release date was a string of variable length, as for some songs it contained month day and year, and for others only the year, which we converted into just an integer representing the year. After reading the information Spotify gives about each feature, we saw that some of the variables that are stored as continuous are actually measures of how confident they are that the song is of the type described. For example acousticness is not a measure of how acoustic the song is, but how confident Spotify is that the song is acoustic. For this reason we converted this variable from the confidence that the song is acoustic to Spotify's prediction if the song is acoustic or not. We did the same thing for the variables speechiness, liveness and instrumentalness.

After completing our dataset, it was time to look into the relationship between each feature and song popularity. We were curious about both what impact they had, as well as the statistical significance of this impact. For the continuous variables we chose to use Spearman Correlation. The reason we chose this over Pearson is because many of our variables are not in any kind of natural units. Popularity, which we compare every song to is not a raw amount like number of streams; it is a number from 0 to 100 that Spotify calculates using some unknown formula. This means that even if there was a linear relationship between one of the features and number of plays, it would not be captured in the relationship between that feature and the popularity data we have available. Because it is computed on the ranks of the data, Spearman only looks at the strength of monotonic relationship, meaning that the relationship would still be captured here. The formula for Spearman is below.

$$\frac{\sum\left(x_i-\mu_x\right)\left(y_i-\mu_y\right)}{\sqrt{\sum\left(x_i-\mu_x\right)^2\cdot\sum\left(y_i-\mu_y\right)^2}}$$

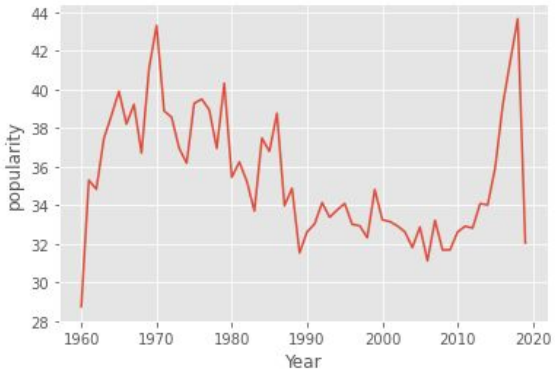
First convert each variable into its rank: the position it would be in if the dataset were in order. Next compute the Pearson Correlation (right) with these new variables.

For the discrete variables, to measure direction we just looked at the difference in the means when the filter was applied and when it was not. To get the significance of this difference, we used a Kruskal-Wallis H Test. While our first instinct was to use a z test, we realized that this was not ideal for similar reasons as Pearson correlation. It compares means rather than medians, and for this reason is less robust.

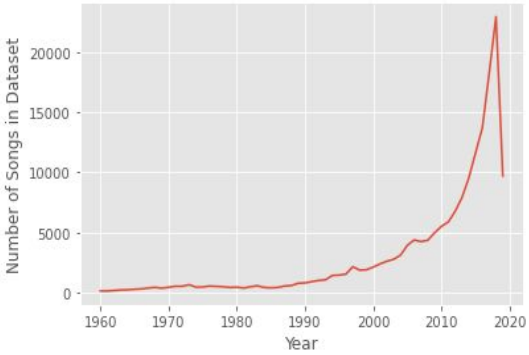
There were several confounding variables we wanted to control for in our analysis. The first of these is the year a song was released. In our initial look we saw that year was related to many other variables.

As can be seen in the graphs above, songs released more recently are generally faster, sadder, shorter and more energetic than songs released longer ago. We predicted that songs released more recently would be popular, and wanted to make sure that a trends were because of the feature itself, not because that feature is related with release year. An example is that if we found fast songs to be more popular, we would not know if these songs are more popular because they are fast, or because they are expected to be newer.

We actually did not find a clear linear relationship between a songs release year and its popularity. Instead we got a random and erratic relation, pictured on the right. This seems weird at first -- but makes sense when considering that our dataset is not a random sample of the songs on Spotify.



Our dataset contains the 10,000 most popular songs from 26 genres. This means each year should have same average popularity, just with different amounts of songs from each year. We did find this to be the case, pictured on the right.



While the average popularity by year doesn't have a clear pattern, it is still not flat. This combined with the fact that many features are strongly related to year meant that we still wanted to control for year when looking for relationship with popularity. To do this, we used a partial correlation, with year as a covariate. The formula for this is below.

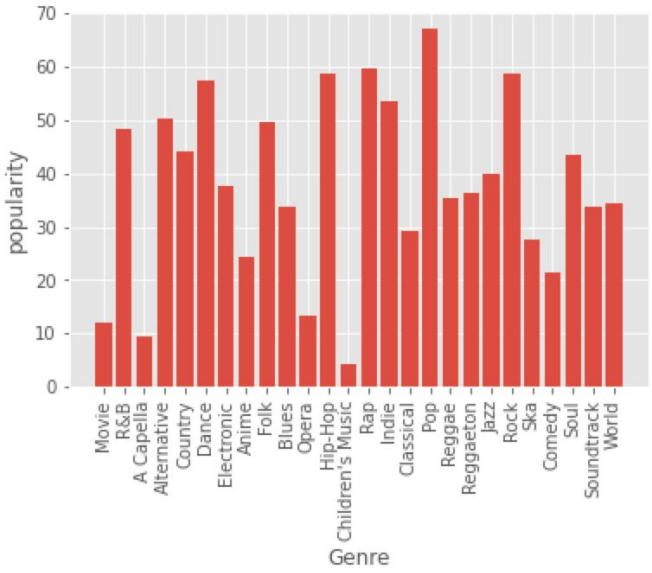
$$\mathbf{w}_X^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (x_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$

$$\mathbf{w}_Y^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$

$$e_{X,i} = x_i - \langle \mathbf{w}_X^*, \mathbf{z}_i \rangle$$
$$e_{Y,i} = y_i - \langle \mathbf{w}_Y^*, \mathbf{z}_i \rangle$$

Then the Spearman correlation is computed on e_X,i and e_Y,i

The second variable we wanted to control for is genre.

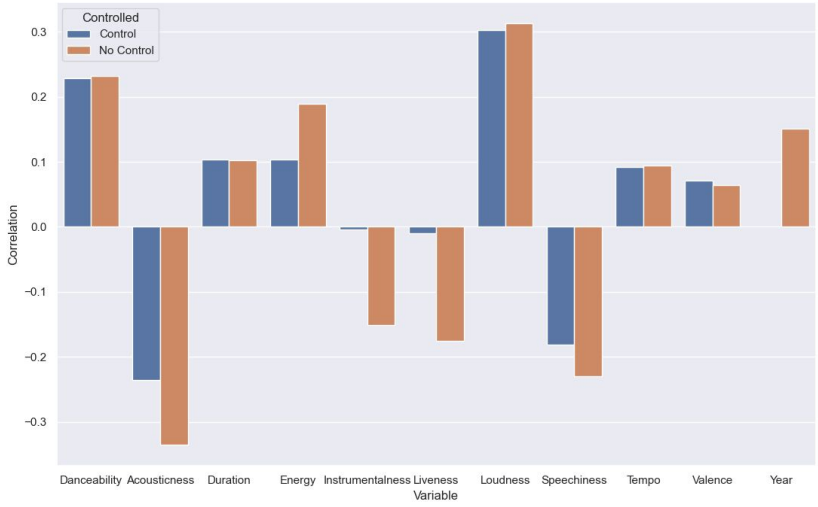


As mentioned above, our dataset is made from the 10,000 most popular songs from 26 genres. This causes there to be an approximately equal amount of songs from each genre, leading to vastly different average popularities. These range from Pop at 67 to Children's Music at 4

To control for genre, we initially wanted to use a partial correlation like we did with year. The problem with this is that while the genres were represented by numbers in the dataset, the values of these numbers were not meaningful. They were simply an enumeration. Because of this, we did not use a partial correlation. Instead, at the times where we wanted to control for genre, we split up the dataset by genre did our analysis on each of these subsets.

4. Results

First we wanted to get an idea of what underlying, continuous musical factors relate to popularity across genres.



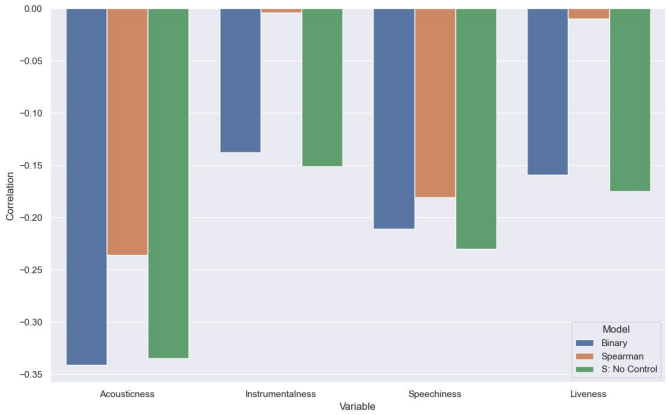
We looked at Spearman correlations with and without controlling for year. Year itself is relatively predictive (r=.151). As such, year has a large underlying effect on the majority of our variables' correlations.

Looking specifically at correlations, controlling for year, it is apparent that none of these correlations are particularly high. Loudness has the highest partial correlation of r=.303, but even that is not indicative of a strong correlation. This makes sense as we are looking at an incredibly wide diversity of music. No one musical variable can capture all these differences and there are likely large genre differences (see below). Additionally, people listen to a wide diversity of music with a wide diversity of features such that one feature can't be fully predictive. It is likely a combination of features that make a song popular: musical composition, artist, genre and otherwise. So these relatively small correlations are expected as they are only predicting a small portion of popularity.

That being said it is interesting to point out that danceability is largely unaffected by the control and remains the second highest positive correlation (r=.228). Danceability then seems to be an important feature across years in making a song popular. This makes sense also considering the Spotify dataset we chose is probably being used for parties and social events. But it's also interesting to think we just like loud (r=.303), danceable songs.

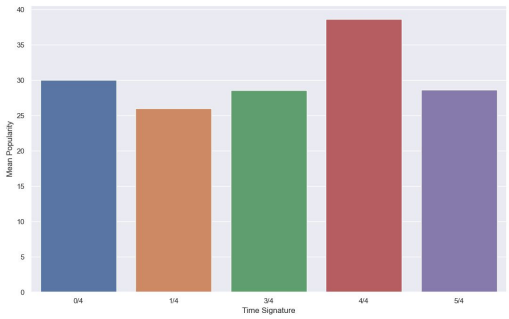
Acousticness has the strongest negative correlation with (r=-.236). Though this was unexpected it does make sense given it like is negatively correlated with positively predictive features like danceability and loudness. Speechiness also has a strong negative correlation (r=-.181). This makes sense as people are more often going to come to Spotify for music, not spoken word.

We also looked at speechiness, acousticness, liveness and instrumentalness as discrete features, given they represent confidence intervals and not continuous values.



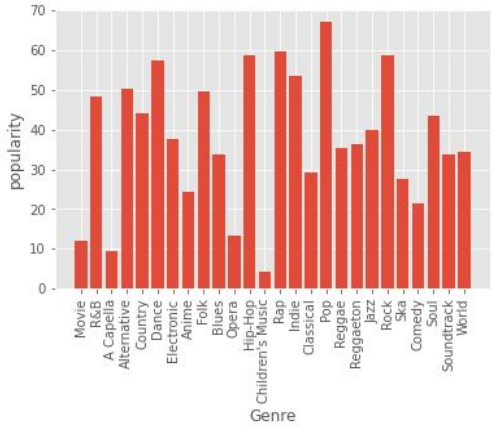
As seen in the graph the point biserial correlation does as good if not better than the uncontrolled Spearman correlation. However, we did not control for year when correlating the discrete values. The correlations, then, with this discrete metric, if controlled by year, will likely look similar to the controlled Spearman correlations.

We also used a Kurskal-Wallis H Test to look at other discrete variables such as release month, key, mode (major vs minor) and time signature. All these differences were significant, as we have enough data to back up even a very weak relationship. However, the means between classes were very close for all of these variables besides time signature, indicating a weak relationship. For that reason, Time signature is the only one of these which we will discuss.



The 4/4 time signature has a high level of popularity likely because it is the most common time signature and overwhelmingly common in popular genres like pop. The other time signatures likely have higher variability as well.

Furthermore, looking between genres also points to how crucial genre is in determining popularity. And how crucial genre is in dictating what features make something popular.



As predicted and as seen in the graph genres like pop, rock, and rap have much higher popularities than genres like a cappella, children's music, or opera. These differences were shown to be significant, like everything else we computed, but by looking at how different they are, it is clear that the relationship between genre and popularity is strong. This indicates that a song being a pop song is highly correlated with popularity, while a song being children's music will not lead to popularity. Given that Spotify is used largely by teens to 20 somethings this makes sense.

We also examined individual indicators of popularity within genres. It is first notable that the variables in our dataset are more predictive for some genres than for others. For example, popularity is incredibly unpredictable for the genres ska and soundtrack. They have only one correlation above .04 each (for ska speechiness (r=-.96) and for soundtrack year (r=.052)). By contrast most genres have a mix of correlations between .03 and .1 and one or two above .1. This makes sense given that ska is a really nebulous genre which sounds very different musically than a lot of other genres; it is likely that the musical variables we are looking at don't really capture the music of ska and as such popularity. Soundtrack also makes sense because sound wise it's not a unifying genre, the genre just indicates whether a song was in a movie or show. This would lead to less unifying predictors of popularity.

On the other hand, the popularity of anime music is very well predicted by our musical variables. Most genres have maybe one or two correlations less than -.1 or greater than.1 but anime has (nearly) 6: duration r=.184, energy r=.15, instrumentalness r=-.206, loudness r=.229, valence r=.105, acousticness r=-.094. This makes sense given similarity in stylings between popular anime openings that are likely defining the genre. Additionally, the musical variables Spotify has are very characteristic of the genre. Anime music is known for long, loud, positive, high energy songs to open for popular shounen.

There are also some individual values within genres that were interesting to examine. The strongest correlation in all the data comes from the correlation between classical music popularity and year (r=-.386). This makes sense since the bulk of classical music was made hundreds of years ago, especially the really popular ones by, for example, Bach or Beethoven.

Additionally, popularity within more modern genres like rap and reggaeton are both highly correlated with year (r=.257 and r=.261 respectively). Popularity within opera is

predicted by loudness pretty well (r=.101) which makes sense intuitively. And the only genre for which acousticness is strongly, positively correlated with popularity is indie (r=.142); since acousticness is characteristic of the genre this does make sense.

There was also a strong negative correlation between comedy popularity and valence (r=-.139) which is surprising because this indicates more sad or angry songs are actually more popular. This could indicate a flaw in how valence is calculated; Spotify isn't transparent about how it comes up with its variables. Or it could be an interesting commentary on a proclivity towards dark comedy.

It was also surprising to see a staggeringly low correlation between danceability and popularity for both dance and electronic music (r=.072 and r=.06 respectively). This could be related to the danceability variable itself being inaccurate; it isn't clear how it is calculated so it could be unrepresentative of actual danceability.

Interestingly some of the more popular genres like pop, rock, hip-hop and country didn't have any correlations above .1 or below -.1. They still had correlations relatively higher than ska or soundtrack though. This might be because they're more broad, all encompassing genres than, say, anime. In any case it's interesting that popularity in these genres isn't well captured by these musical variables when we thought, for example, energy and danceability would characterize pop pretty well. Further research could look into what makes pop music pop music to see what features within the genre do actually lead to the quite large popularity of the genre.

5. Discussion

Through this project we hoped to utilize the Spotify API to look at large scale trends and predictors of popularity. Particularly, we hoped to look at music level composition features and audio characteristics to see what characteristics of the song itself predict popularity. What we found is that there are a number of mildly correlated characteristics but no incredibly strong predictors of popularity. We were expecting this given the vast number of variables that go into a song's popularity, like artist, genre, or release year (which we controlled for). Additionally, people are simply interested in different things and each song has a wide breadth of characteristics; there is no singular component that creates a song that will be popular. We didn't plan on finding highly correlated predictors, rather we hoped to find (and did find) smaller predictors that trend towards popularity: namely danceability, loudness, (lack of) acousticness, and (lack of) speechinesss.

Within genre, the factors that predicted popularity were highly variable. Some genres' sound characteristics, like anime, were simply better captured by the variables in the data. On the other hand some genres are more abstract and less unified, leading to less predictive power, e.g. ska and soundtrack. Surprising characteristics we found could be due to faults in the way Spotify constructs these variables, like, for example, danceability not highly correlating with popularity in dance or electronic genres.

With more time it would be interesting to see how obvious features, like artist, also play into popularity and interact with the correlations we discovered. The dataset we utilized also is

specific to Spotify users, so to be clear we are looking at popularity on a particular subset of the population, whose taste could be very different from people who choose not to use Spotify. As such popularity data from different sources could reveal trend information from a larger, more diverse swath of the population. Additionally, popularity as a metric is very temporally local; it is a measure of streams during a specific time period rather than overall streams. Looking at more temporally broad popularity data could produce very different results and reveal more about popularity at different time points.

It is also important to note that the variables we were looking at were not objective measures of these songs' features. Rather they were Spotify's estimations of features like energy and even loudness. Future research could also utilize different measures to see how different variables and differently measured variables change these popularity predictions.

It should also be noted that when cleaning our data we removed duplicate songs that appeared in multiple genres. This has the very real potential of skewing which songs appear in what genre. This missing data could then change the characteristics of popularity within a genre. There could also be systematic popularity differences between songs that appear in one genre and songs that appear in multiple. Future analysis could examine songs under multiple genres to see how the popularity of a song might differ between genres. Additionally, analysis could see how genre popularity changes when these duplicate songs are included. Also it is unclear how Spotify even assigns genre so analysis into what makes a song a part of a genre could also be worthwhile.

This preliminary, correlative analysis can't say with certainty that these variables like danceability or loudness cause popularity. However, the auditory features we found could have implications for what people like to listen to. Maybe there is something to the psychological appeal of danceable and loud songs. Future research will have to delve into this mechanistic question in greater depth.

Bibliography

https://towardsdatascience.com/what-makes-a-song-likeable-dbfdb7abe404
https://www.psychologytoday.com/us/blog/finding-new-home/201806/why-do-some-songs-become-popular