Spotify: What Makes a Song Popular?

Picture in your mind an owner of a record label. Let's say they want to create the next hit song. Like a good owner, they want to create music that generates the most amount of profit possible. Although this record label holder may be hypothetical, my interest in this topic is real.

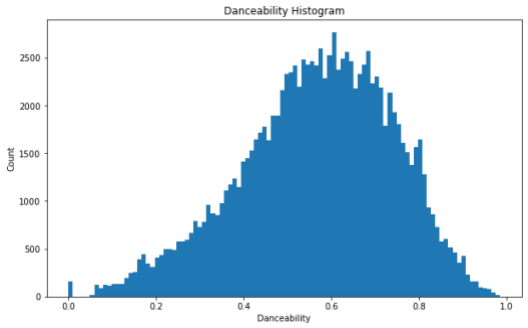
Like the title suggests, I want to investigate what makes music popular.

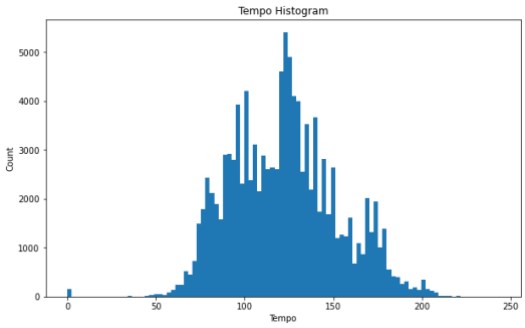
To answer this, I analyzed a Spotify dataset containing 114,000 songs. The rows of the dataset corresponded to one unique song per row, while the columns corresponded to information and statistics about each song. Examples include the name of the artist, name of the song, etc. From this point forward, I will be referring to the columns as "features." Information about each feature can be viewed in the appendix at the end. I will be making frequent references to the terms listed.

Spotify measures popularity by assigning a number between 0 to 100 corresponding to the number of plays a song gets on its app. This value is how I will be operationalizing "popularity" for the purposes of this analysis.

To begin, I took features with numerical data (no categorical or Boolean columns) and determined which were normally distributed. This was important as many statistical tests rely on the assumption that the data used is normalized. Determining what tests were available for me to use was an important step.

I plotted the 10 numerical features and found that of them (popularity, duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo), only danceability and tempo could even remotely be construed as normal distributions.






```
KS Statistic: 0.5505590466687807
P-value: 0.0
```

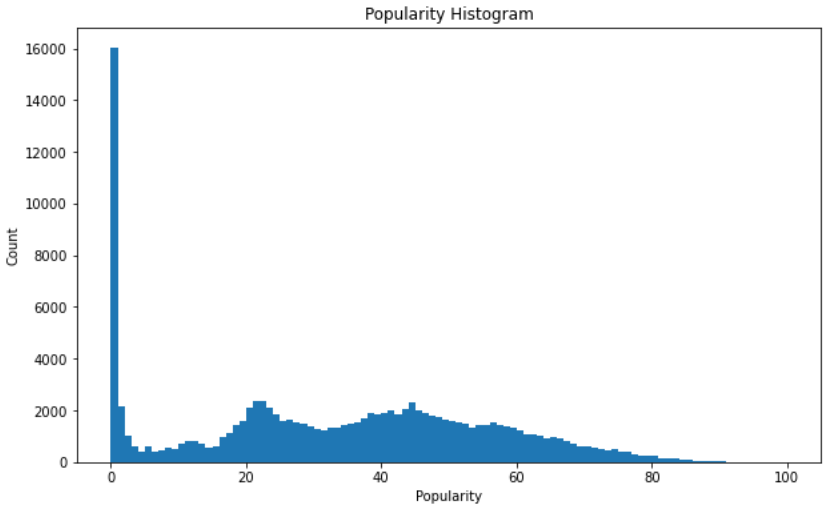
```
KS Statistic: 0.9986228070175439
P-value: 0.0
```

I ran two Kolmogorov–Smirnov tests to determine a goodness of fit to a normal distribution. Danceability did not have much similarity with a normal distribution. The tempo KS statistic suggested that the distribution is very close to a normal distribution. However, I have several issues with this value. The tempo histogram has several smaller peaks: i.e. at 100 bpm and 175 bpm – whereas a true normal distribution only has one peak at the mean. Additionally, there are

large gaps in the tempo distribution. A true normal distribution would have a continuous line with no breaks.

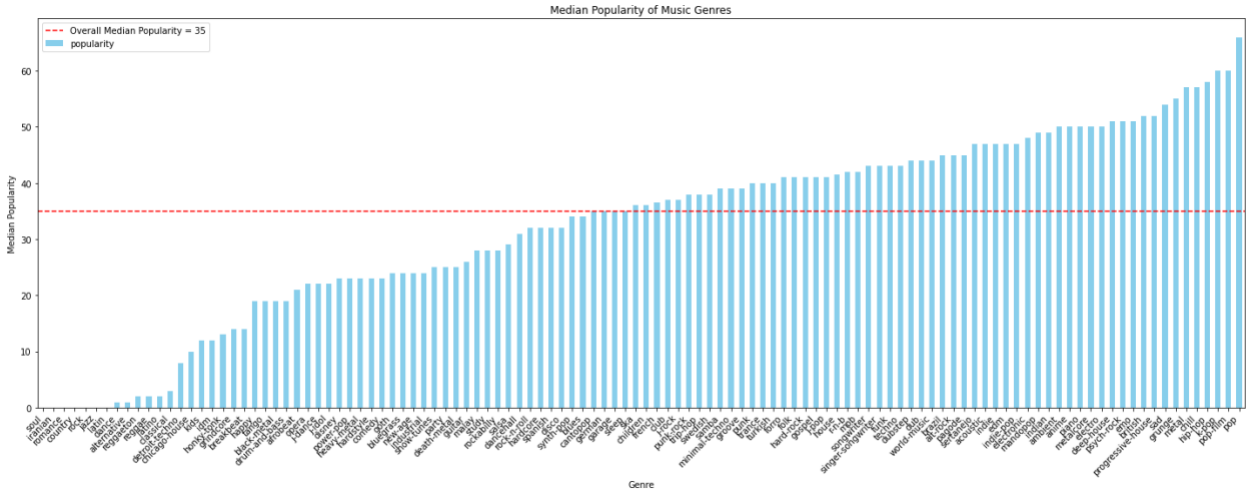
Therefore, I chose to not blindly trust the result and treated that none of the 10 features were normally distributed.

Curiously, popularity is not normally distributed.



A large portion of songs have exactly 0 popularity. This was not due to lots of songs having missing data. This was due to Spotify giving thousands of songs a popularity score of zero.

To understand what types of songs were especially poorly rated by Spotify, I sorted the songs by music genre and plotted their probability in histograms.




```
Genres with a median popularity less than 10:
['soul', 'iranian', 'romance', 'country', 'rock', 'jazz', 'latin', 'dance', 'alternative', 'reggaeton',
'reggae', 'latino', 'classical', 'detroit-techno']
Genres with a median popularity greater than 50:
['deep-house', 'psych-rock', 'emo', 'british', 'progressive-house', 'sad', 'grunge', 'metal', 'chill', 'hip-
hop', 'k-pop', 'pop-film', 'pop']
```

These results carried some implications for the overall dataset. Of the 114,000 songs in the dataset, 16,020 of them have a popularity value of 0. Any statistical tests or prediction models based on the original dataset would be affected by the skewed data. I did not to irresponsibly ignore inconvenient data. With this, I chose to include all songs from the dataset in my calculations below.

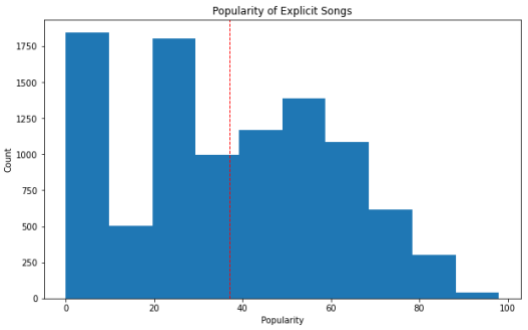
Visually, it appeared to me that there were significant differences in popularity between music genres. To put numbers behind my belief, I ran a chi-squared test.

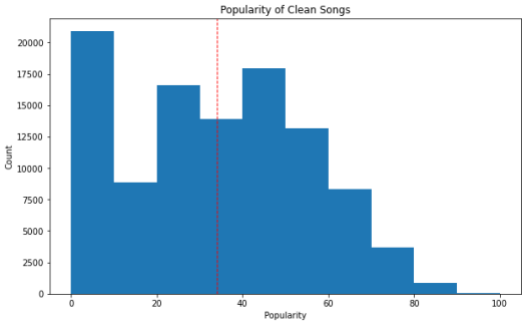
```
Chi-squared Statistic: 39476.450142072164
P-value: 0.0
Degrees of Freedom: 113
With 113 degrees of freedom, a Chi-squared score of 39476.450142072164, and a p-value of 0.0, there is a
significant difference between categories of music.
```

Next, I investigated if explicitly rated songs are more popular than not explicit (clean) songs. I separated the songs into two respective groups.

The next step was to determine what test to use to determine if the differences between the two groups were significant. I selected the Mann-Whitney U Test. There were two primary reasons why I chose this test. First, the number of explicit and clean songs is unequal. Equal group size is a condition that is required for many other tests, but the Mann-Whitney U Test is flexible to this situation. Additionally, the popularity of explicit and clean songs is not normally distributed which is acceptable under the Mann-Whitney U Test (but unacceptable for many other tests). Given said qualities about the data, this test was appropriate for the situation.

The p-value of this test is far below the conventional alpha level of 0.05. Therefore, there is a statistically significant difference in popularity between songs that are explicit vs. songs that are not explicit. Explicit songs are more popular than clean songs, but only slightly so.

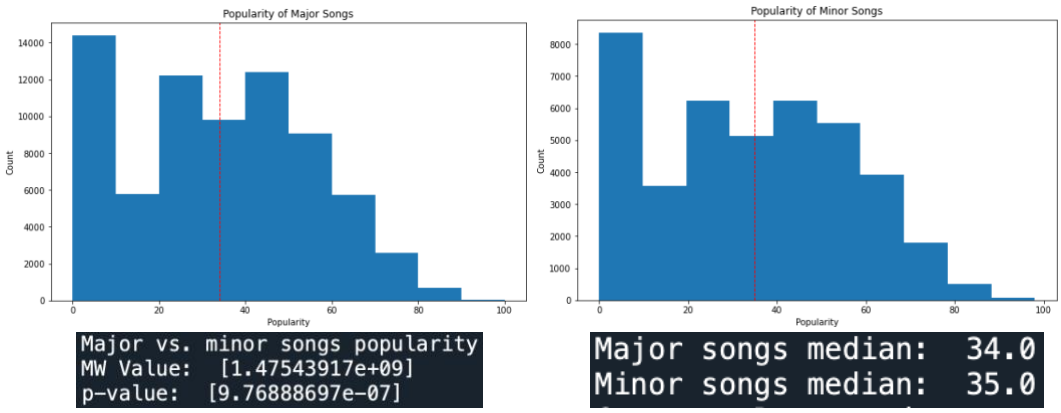





```
Explicit vs. clean songs popularity
MW Value:  550173096.0
p-value:  6.142226598728333e-42
```

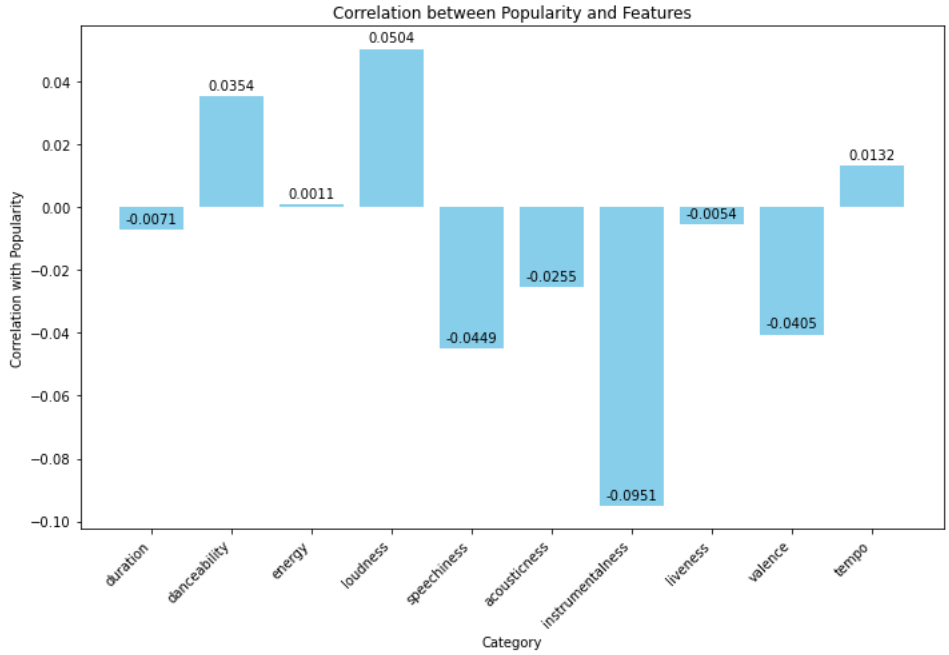
```
Explicit songs median:  37.0
Clean songs median:  34.0
```

Next, I wanted to determine whether songs in major or minor key are more popular. I ran the same process: I split the data into major and minor songs, and ran another Mann-Whitney U test. As before, the group size of major and minor songs is unequal along with neither graph being normally distributed, making the Mann-Whitney U Test a good selection. Once again, the p-value is far below our established alpha level of 0.05, indicating that the results are statistically significant. Songs in minor key are more popular than songs in major key.

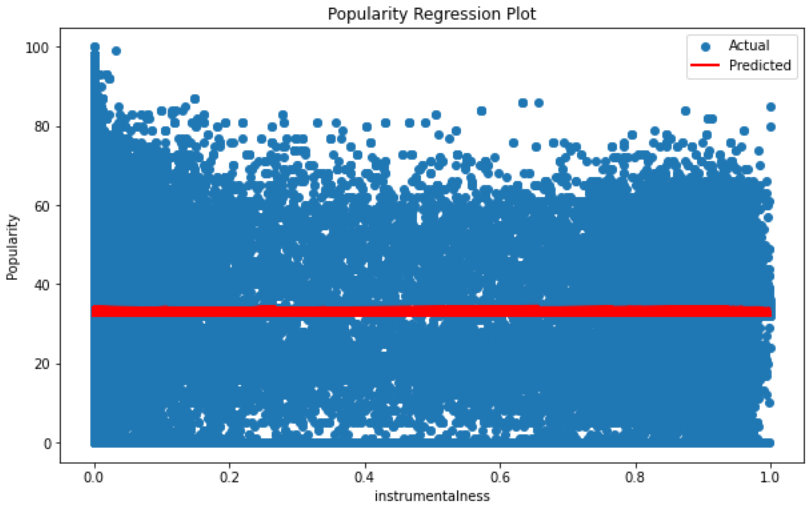


So far, I have identified three factors that could impact the popularity of a song: the genre of a song, whether a song's lyrics are explicit or clean, and if the song is in major or minor. The next criteria I was curious about was whether any individual feature was a good predictor of popularity. To begin, I ran a correlation calculation between the ten numerical features and popularity.

All ten numerical features had a weak correlation with popularity. The largest positive or negative correlation was between popularity and instrumentalness (-0.0951). My interpretation of these results: None of the features alone in a vacuum are indicative of how popular a song is.



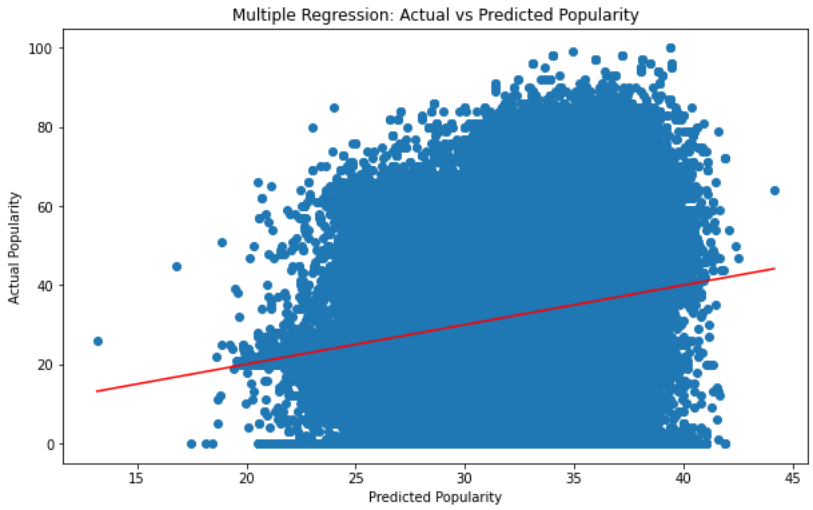
I wanted to confirm my above results and ran linear regressions to use the above variables to predict popularity. The results were consistent with my earlier takeaway: none of the above features were accurate predictors of popularity.




```
Greatest R-squared category: 'instrumentalness' R-squared of 0.00905146803208523
```

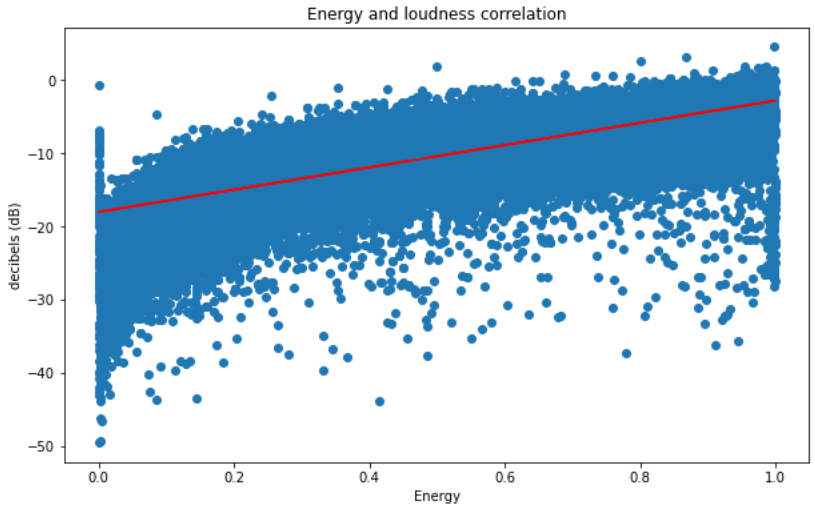
Even instrumentalness, the "best" predictor, was extremely inaccurate.

I was curious to see if using the variables simultaneously would be a better predictor of popularity and ran a multiple regression. The results yielded a prediction model that was marginally better at predicting popularity.




```
R-squared for multiple regression using all 10 categories: 0.022585186073760033
Difference in R-squared values:  0.013533718041674803
```

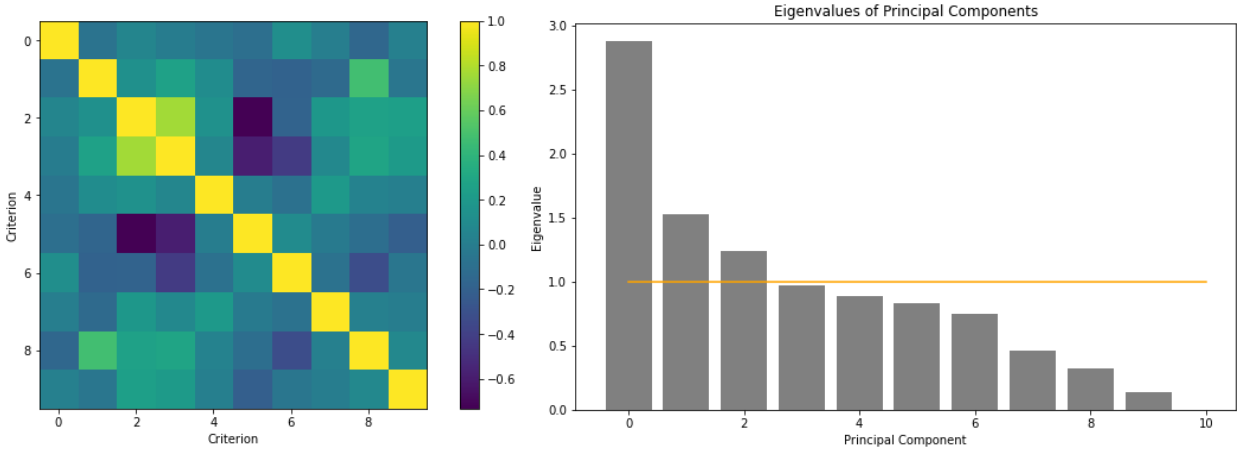
The multiple regression r-squared value was only marginally greater than its single linear regression counterpart. I suspected this difference could be attributed to multicollinearity. To confirm my suspicions, I ran a linear regression and correlation calculation of two similar features: energy and loudness. My intuition was correct: these two variables were very correlated with each other.



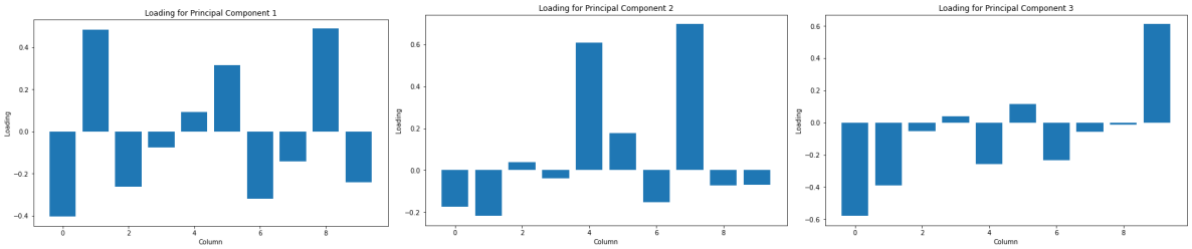

```
Correlation between Energy and Loudness: 0.761689959890819
```

Because I found these two features to be highly correlated with each other, I decided to run a principal component analysis on my 10 numerical features. By using this dimensionality reduction method, instead of working with 10 dimensions of varying correlation to each other, I could identify the "true dimensions" responsible for determining popularity.

I started off by creating a correlation matrix of the 10 numerical features. I also Z-scored the data and performed a PCA on the 10 features. I then created a scree plot below:



There are a couple of methods that I considered to choose the number of principal components to extract. For example, I could have used the elbow method and extracted 1 component. Or, I could have taken the number of principal components accounting for 90% of the variance (7 components). In the end, I chose to use the Kaiser criterion: the components with an eigenvalue > 1. This meant that I had to extract 3 principal components.



In order to determine the meaning of each principal component, I decided to use ±0.4 as my threshold for a "significant column." Any column mentioned below will be beyond this threshold.

For the first principal component: This principal component points towards features 1 and 8, which correspond to danceability and valence respectively. It also points away from feature 0, which is duration. Overall, this component seems to reflect how "energetic" and "fun" a song is. The negative presence of duration indicates that despite this, everything gets boring eventually.

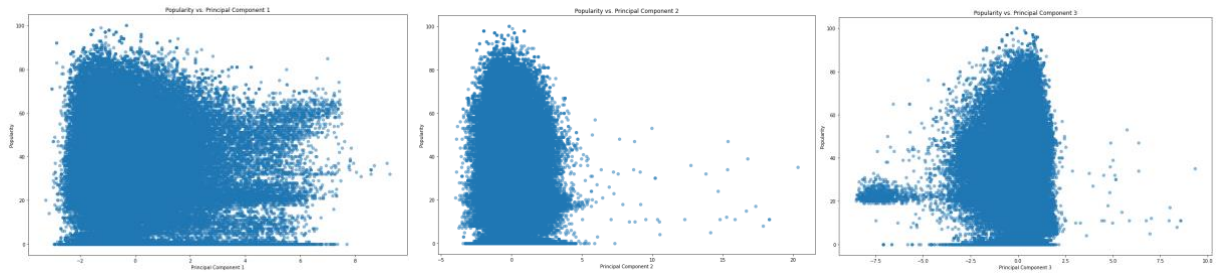
For the second principal component: This principal component points towards features 4 and 7, which correspond to speechiness and liveness respectively. Specifically, the presence of speechiness and live lyrics have an impact on this component. This component reflects the lyricism of a song.

For the third principal component: This principal component points away from feature 0 and towards feature 9 which correspond to duration and tempo respectively. The absence of duration and presence of tempo both have an effect. This component seems to reflect the "chronoception" of a song. A uniquely human trait is the ability to perceive the passage of time. If a song is sufficiently entertaining, humans are not dissatisfied by time passing. But if a song bores the listener, they suddenly believe that time is moving very slowly and can't wait for the end.

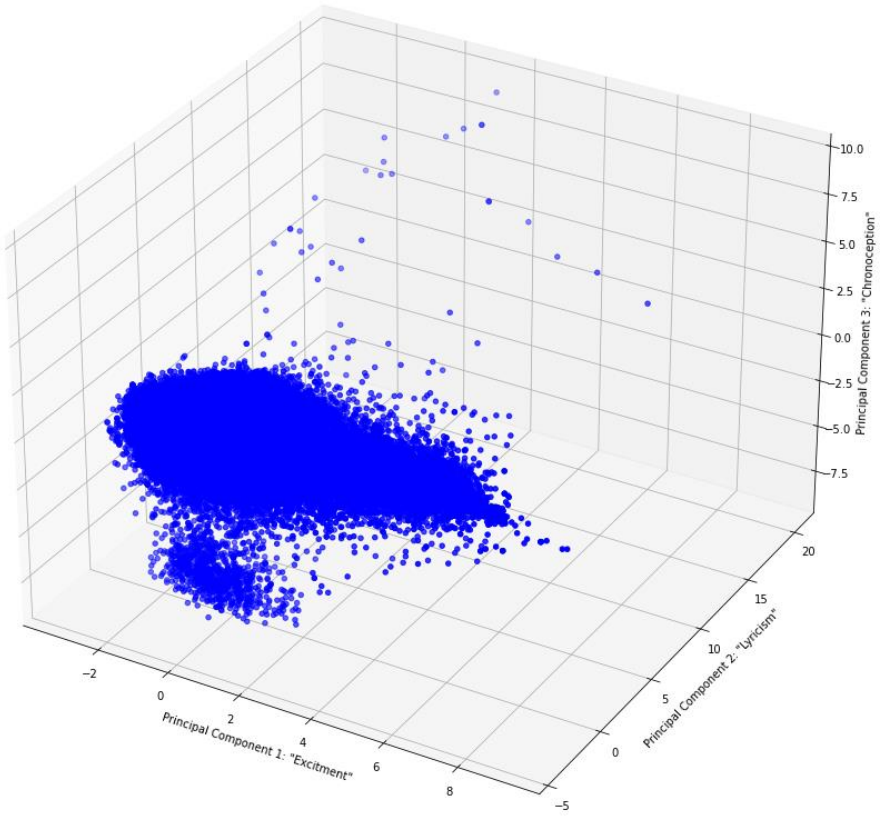
I reduced my 10 original dimensions into three calculated dimensions: we now can evaluate a song's popularity in three calculated dimensions: "excitement," "lyricism," and "chronoception."

The process of calculating this PCA was also helpful in building greater context on my above results. For example, I determined that minor songs are significantly more popular by major songs. However, they were only more popular by a rating of 1 point. While this is a statistically significant difference, these groups *did not have drastically different ratings* – indicating that a song's key is just one piece of the puzzle to a song's popularity.

After this, I rotated the data to display the relationship between the three principal components.



3D Plot of First Three Principal Components



I originally started this analysis to answer my question: “What makes a song popular?” After finishing said analysis, I have determined that song popularity is a much more complicated question than a simple linear regression between one variable and popularity.

The popularity of a song can be measured in the 3 aforementioned dimensions. There were other results from my analysis that have a small effect on popularity (i.e. genre, type of lyrics, and key signature).

However, there are other factors not present in this dataset that could also add another dimension to popularity. For example, how would advertisements affect a song’s popularity? Songs that receive more financial backing are more likely to be heard by the general public and therefore

have more chances to become popular. I would like to further investigate concepts like these in the future.

In the end, music is a complicated entity. While there are many quantifiable traits of songs: key signature, beats per minute, etc., it is difficult to truly measure how music speaks to us. People may find it challenging to explain why they are drawn to certain types of music but not others. There's beauty in that: with so much variety in music – there is truly something for everyone.

But a hypothetical record label owner may not care.

Appendix

Descriptions some of the columns/features:

artist(s) – the artist(s) who are credited with creating the song.
album_name – the name of the album.
track_name – the title of the specific song.
popularity – An integer calculated by Spotify. Depending on the number of plays a song gets, a value is assigned from 0 (least popular) to 100 (most popular).
duration – this is the duration of the song in milliseconds.
explicit – a Boolean variable that is true/false depending if a song has explicit lyrics.
danceability – An integer calculated by Spotify that quantifies how easy it is to dance to the song. A value from 0 (very difficult to dance to) to 1 (very easy to dance to).
energy - A value calculated by Spotify to quantify the "intensity of a song. Ranges from a 0 (slow, soft) to 1 (fast, "hard").
key – the key of the song from A to G# (mapped to categories 0 to 11).
loudness – average volume of a track in decibels.
mode – a Boolean variable that is 1 = if the song is in major, 0 – song is in minor. Further reading can be found here.
speechiness – A value calculated by Spotify to quantify how much of a song is spoken. Ranges from 0 (fully instrumental songs) to 1 (songs consisting entirely of words).
acousticness – A value calculated by Spotify that varies from 0 (song contains exclusively synthesized sounds) to 1 (song features exclusively acoustic instruments like acoustic guitars, pianos or orchestral instruments).
instrumentalness – A value calculated by Spotify that is the inverse of speechiness, varying from 1 (for songs without any vocals) to 0.
liveness - A value calculated by Spotify. It tries to quantify how likely the song was recorded in a studio without a live audience (values close to 0) vs the recording was live in front of an audience (values close to 1).
valence - A value calculated by Spotify. Quantifies how "uplifting" a song is. Songs with positive moods have values close to 1 and negative moods have values close to 0.
tempo – speed of the song in beats per minute.
time_signature – how many beats there are in a measure (usually 4 or 3).
track_genre – genre assigned by Spotify, e.g. "blues" or "classical."