What Makes a Song Popular?

Introduction

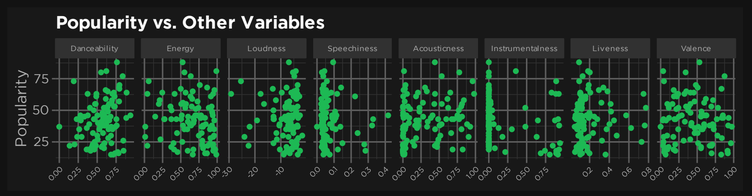
I'm sure we've all wondered what makes a song popular. It seems like some songs go viral seemingly overnight, while others go unnoticed by the general population. I wanted to know what factors had the most influence on a song's popularity, and whether it is possible to predict how popular a song will be without knowing how many people listen to it.

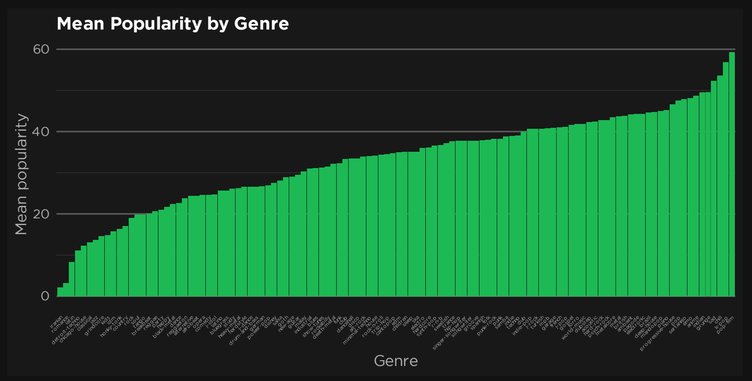
Background Research

I found several studies on this topic. The first study, discussed in the article "Why do songs become popular?" by Arash Emamzadeh in *Psychology Today,* found that songs with content that was more genre-atypical (not like content typically found in the song's genre) tended to be ranked higher. Another study is detailed in the paper "What Makes Popular Culture Popular?: Product Features and Optimal Differentiation in Music" by Noah Askin and Michael Mauskapf. This study found the same results as the first study.

Dataset & Data Exploration

My dataset is from the user Maharshi Pandya on Kaggle. The data was gathered using the Spotify Web API, which can provide JSON metadata about artists, albums, and tracks. Though it is not explicitly specified by Pandya, I believe that the data was current as of October 2022, which was when the dataset was published. The dataset includes 114,000 observations of several variables, such as genre, popularity, and danceability, where each set of observations represents information about 1 track. There are 1,000 tracks from each of 114 genres. To explore the data, I started by creating some scatterplots comparing popularity to some of the other continuous numeric variables in the dataset. I did this to search for any obvious correlation between popularity and one of the other variables. The first group of plots below shows scatterplots of popularity versus several other variables. It does not appear as though any of these variables have a strong correlation with popularity. Next, I created a bar chart comparing the mean popularity of the tracks in each genre. I did this to see if there appeared to be a significant difference between the popularity of songs from different genres. The plot shows that there is a clear difference between different genres, with some being far more popular than others.





Hypothesis

Going into this project, I hypothesized that the factors which contributed the most to a song's popularity would be genre, energy, and danceability. I thought this because these seem like factors which would make a song "fun," which in turn could lead to more popularity.

Inference Tests - Introduction

I decided to run 2 inference tests to see how much influence genre and mode have on popularity. I ran a 1-way ANOVA test to see if the true mean popularity is different for different genres, as well as a 2-sample difference of means t-test to see if the true mean popularity is equal for songs in the major and minor modes.

One-way ANOVA Test for Difference of Means Between Genres

$\mu_n =$ True mean popularity of songs from the n^{th} genre
H_0: $\mu_1 = \mu_2 = \cdots = \mu_{114}$ (i.e., the mean popularity is equal throughout all genres)
H_a: $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_{114}$ (i.e., at least 2 of the means are not equal)
$\alpha = 0.001$

Since the sample means for each group are greater than 30, we can assume that the distributions of sample means are approximately normal. However, since it was not explicitly stated by Pandya, we cannot assume that the samples were selected at random. Proceeding with caution anyways.

$$\bar{\bar{x}} = \frac{\sum(n_i \bar{x}_i)}{N} = \frac{1000(42.483) + 1000(24.399) + \cdots + 1000(41.783)}{114000} = 33.239$$

$$F = \frac{MSTr}{MSE} = \frac{\frac{\sum\left(n_i\left(\bar{x}_i - \bar{\bar{x}}\right)^2\right)}{k-1}}{\frac{\sum\left((n_i - 1)s_i^2\right)}{N-k}} = \frac{127572}{371} = 343.5$$

$p = P(F \geq 343.5) \approx 0$ (on 113 and 113886 degrees of freedom)

Since our p-value (0) is less than α (0.001), we reject the null hypothesis that all the true means are equal. Therefore, we have sufficient evidence to support the alternative hypothesis that at least 2 of the true means are not equal.

2-sample Difference of Means t-test for Mean Popularity by Mode

$\mu_1, \mu_2 =$ True mean popularity of songs in the major and minor mode, respectively
H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$
$\alpha = 0.001$

Conditions for performing test are the same as above.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 4.6709$$

$p = P(t \geq 4.6709) = 3.3003 \times 10^{-6} \approx 0$ (on 84062 degrees of freedom)

Since out p-value (0) is less than α (0.001), we reject the null hypothesis that the true means are equal. Therefore, we have sufficient evidence to support the alternative hypothesis that the means are not equal.

Predictive Model

After figuring out that these factors have a significant influence on the popularity of a song, I wanted to create a predictive model to see if the popularity can be predicted using nothing but audio features. The multiple regression model that I produced is as follows:

$$\widehat{y'} = 5.959 + 0.383x_1 + 1.005x_2 - 1.807x_3 - 0.028x_4 - 2.007x_5 - 2.214x_6 - 2.380x_7 - 2.197x_8 + 1.951x_9 - 1.871x_{10}$$

Where:
$y =$ The song's popularity score from 0 to 100, only for songs with popularity greater than 10
$y' = \sqrt{y}$
$x_1 = 1$ if the song is explicit, 0 otherwise
$x_2 =$ The song's danceability (from 0 to 1)
$x_3 =$ The song's speechiness (from 0 to 1)
$x_4 =$ The song's mode (1 for major, 0 for minor)
$x_5 =$ Whether the song's genre is "breakbeat" (1 if true, 0 if false)
$x_6 =$ Whether the song's genre is "grindcore" (1 if true, 0 if false)
$x_7 =$ Whether the song's genre is "honky tonk" (1 if true, 0 if false)
$x_8 =$ Whether the song's genre is "kids" (1 if true, 0 if false)
$x_9 =$ Whether the song's genre is "pop" (1 if true, 0 if false)
$x_{10} =$ Whether the song's genre is "tango" (1 if true, 0 if false)

These specific genres were chosen because they were found to have the most significant influence on the popularity of the song. The threshold for popularity was chosen to be 10 because there were so many songs with a score of 0 that it might screw up the model.

$R^2 = 0.1696$,
$R^2(adj) = 0.1695$
$F = 1849$ on 10 and 90527 degrees of freedom

F-test for Model Utility

β_n: True population coefficient for x_n
H_0: $\beta_1 = \beta_2 = \cdots = \beta_{10} = 0$ (i.e., all the population slopes are equal to zero)
H_a: $\beta_1 \neq \beta_2 \neq \cdots \neq \beta_{10} \neq 0$ (i.e., at least one of the true slopes is not zero)
$\alpha = 0.001$

No obvious pattern exists in the residuals. The residuals appear to be normally distributed about the regression line. The errors have approximately equal variance throughout the model.

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-(k+1)}} = \frac{\frac{0.1696}{10}}{\frac{1-0.1696}{90538-(10+1)}} = 1849$$

$p = P(F \geq 1849) \approx 0$ (on 10 and 90537 degrees of freedom)

Since our p-value (0) is less than α (0.001), we reject the null hypothesis that the true population slopes are all equal to zero. Therefore, we have sufficient evidence to support the alternative hypothesis that at least one of these slopes is not zero, which indicates a useful model.

Conclusion & Reflection

In conclusion, it appears as though there are factors other than audio features that influence a song's popularity. We know this because of the low R^2 value of 0.1696. This could have been improved by adding more predictors but doing this would have made the model more complex. This process could have been done better by analyzing the content of the songs and comparing it to content typical of the song's genre in the same way that the previous studies on this topic did.