Moving to the top

A Study of the Influence of *Danceability* on Song Popularity

Are easy-to-dance songs more popular?

Everyone loves to listen to music and to dance, whether it is at a club or in the comfort of our homes. But is it true that the most popular songs are also the most *danceable*? And more importantly, to what extent does the *danceability* of a song translate into success [1][2]? The purpose of this analysis is to asses the impact of the easiness to dance on a song's popularity by analyzing a decade worth of data from Spotify's top songs. The findings of this study provide valuable insights that can be applied to various aspects of the music industry.

The data

The data used to conduct this research compiles information about songs that were on the top 100 from Spotify across the years 2010-2019. Some[3] were on the top for multiple years, and seeing that the goal of the study does not involve a temporal analysis, the duplicate information was removed. The data also went through some cleaning before the beginning of the analysis: records with strange values were removed and the variables *artist type* and *genre* were converted into dummies. A description can be found below:

Variable	Description
bpm	Beats per minute
nrgy	How energetic the song is ranging from 0 to 100
dnce	How easy it is to dance the song ranging from 0 to 100
dB	How loud the song is in dB
live	How likely the song is a live recording
val	How positive a song is ranging from 0 to 100
dur	Duration of the song
acous	How acoustic the song is ranging from 0 to 100
spch	How focused on the spoken word the song is
pop	Dummy for 'pop' music genre
rock	Dummy for 'rock' music genre
hip_hop	Dummy for 'hip hop' music genre
rap	Dummy for 'rap' music genre
other	Dummy for 'other' music genre
solo	Dummy for solo song
group	Dummy for group song
Target: popularity	Popularity of the song (not a ranking)

Methodology

Correlations

To begin with, the correlation value between all non dummy variables was analyzed. Some variables appeared to be moderately correlated with the target, revealing potential to be good predictors. Next, 2D scatter plots of each variable against the target were examined to better understand how those were correlated. This was a crucial step seeing that correlation value only reflects linear relationships and it was concluded that a level-log relationship was likely to be more appropriate between the target and the variables *live* and *acous*. It was also possible to determine that the variables *val*, *dur* and *spch* were not relevant for the analysis given the uniform distribution of the popularity of songs across their values.

Functional form

After assessing the plots, it seemed reasonable that the best way to model the relationship between the target and the predictors was not always a level-level relationship. Therefore, a model with a level-level, a level-log and a mixed form model were tested with *other* and *group* as the default omitted dummies. The majority of the variables showed to be significant and this significance was similar across all specifications. However, after careful study, the one that showed the lowest p-values was the the model with the log transformations only on *live* and *acous*, and it was used on the rest of the analysis.

Variables removal

Removing highly correlated variables from the model may result in biased estimates, but keeping them may cause inflation of the standard errors, reducing the estimations precision. In this case the overall estimates became more significant, as well as the R-squared, after removing the variable *dB* (with highest p-value), which was highly correlated with *nrgy*. The dummy variable *solo* estimate showed a fairly high p-value, meaning that the relationship between the variable and the target was probably due to chance, hence it was also removed.

Testing for heteroskedasticity

Both the Breusch-Pagan and the White Special tests showed evidence for the presence of heteroskedasticity (p-values 2.41e-05 and 1.296e-05, respectively). This means that the obtained estimates were most likely biased and inefficient. Knowing this, it was necessary to use a robust OLS estimator.

Functional form misspecification

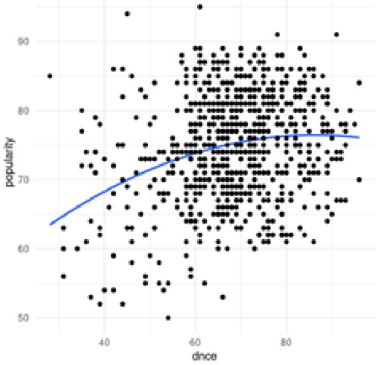
The RESET test with the heteroskedasticity-corrected covariance matrix showed a p-value of 0.0128, hence there is evidence of a functional form misspecification. To address this problem multiple non-linear functions of the independent variables were added to the model until a final best model was reached. This one showed a new p-value of 0.3338 thus being now well specified:

$$\mathit{popularity} = \beta 0 + \beta 1\ \mathit{nrgy} + \beta 2\ \mathit{dnce} + \beta 3 \log(\mathit{live}) + \beta 4 \log(\mathit{acous} + 1) + \beta 5\ \mathit{pop} + \beta 6\ \mathit{rock} + \beta 7\ \mathit{hiphop} + \beta 8\ \mathit{rap} + \beta 9 \log(\mathit{live})^{*}\mathit{rock} + \beta 10\ \mathit{dnce}^2 + u$$

```
                    Estimate Std. Error t value  Pr(>|t|)
(Intercept)       48.5285627  7.2651010  6.6797 4.706e-11 ***
nrgy              -0.0615115  0.0191560 -3.2111 0.0013795 **
dnce               0.8066169  0.2140829  3.7678 0.0001778 ***
log(live)         -0.9117318  0.4246370 -2.1471 0.0321108 *
log(acous + 1)     0.6416060  0.2295817  2.7947 0.0053297 **
pop                1.5417249  0.7508648  2.0533 0.0403978 *
rock              20.1762292  6.9278234  2.9123 0.0036952 **
hip_hop            3.3324448  0.9630884  3.4602 0.0005708 ***
rap                2.8894521  1.3013618  2.2203 0.0266986 *
I(log(live)*rock) -5.5019893  2.4132893 -2.2799 0.0228990 *
I(dnce^2)         -0.0052252  0.0015616 -3.3462 0.0008608 ***
Signif. codes:  0 ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-squared: 0.1408
```

Conclusions

Regressor	*Popularity*
nrgy	1 percentage point increase results in less 0.0615 percentage points in popularity
dnce	1 percentage point increase cause (0.81 - 0.00523*dnce*) percentage points change in popularity
live	1% increase in *live* results in less 0.0091 percentage points in popularity, when the song is not rock. When it is a rock song, the decrease of 0.064 percentage points in popularity.
acous	1 percentage point increase results in more 0.0064 percentage points in popularity
pop	*pop* songs are 1.54 percentage points more popular than songs belonging the category *other*
rock	*rock* songs are 20.18 percentage points more popular than songs belonging the category *other*
hip_hop	*hip_hop* songs are 3.33 percentage points more popular than songs on the category *other*
rap	*rap* songs are *r*2.89 percentage points more popular than songs on the category *other*



Looking at the p-value of the coefficients of *dnce* and *dnce²*, there is evidence that the popularity of a song increases along with *danceability* up to a certain point. However, the turning point (*dnce* = 144) surpasses the range of the variable, meaning that the popularity of song increases with its *danceability* in a quadratic form. In summary, we can conclude that a catchy beat is a key factor for the success of a song .

1. Music we move to: Spotify audio features and reasons for listening, PLOS ONE.Available at: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0275228
2. Musical trends and predictability of success in contemporary songs in and out of the top charts, Royal Society Open Science. Available at: https://royalsocietypublishing.org/doi/10.1098/rsos.171274
3. Spotify Top 100 songs of 2010-2019, Kaggle. Available at: https://www.kaggle.com/datasets/muhmores/spotify-top-100-songs-of-20152019