How to Make a Hit Song Today

Using Regression to Predict Song Popularity

Timothy Regis

December 22, 2020

To Do

- Figure Numbers
- Read Over
- Better Formatting
- ReadMe and GitHub Setup
- Comment Code

Abstract

Making a hit song can often seem like an impossible task to the average person as the components of a final composition can be extremely complex. In an attempt to uncover these secrets, this paper applies a generalized regression model to a dataset of Spotify's music catalog, and analyzes the effects of different musical characteristics on the song's popularity. The findings of this analysis suggest that the danceability, energy, valence, and speechiness of a song are all strong predictors of a song's popularity today, with danceability and energy showing significant positive correlations and the latter two suggesting negative returns to popularity. While this cannot be considered a 'golden rule' for songmaking, these results can help interested musicians incorporate new factors into their music to better appeal to wider audiences.

Keywords

Generalized Linear Regression, Musical Attributes, Spotify Music Catalog, Song Popularity

Introduction

The musical industry of today's world is incredibly complex. Differing tastes have given rise to a seemingly endless pool of genres to choose from. This has made it increasingly difficult to uncover just what exactly goes into making a hit song that resonates with the greatest number of listeners. At times, it seems like only a small and exclusive group holds the secret to this as only a select few appear to top the charts on a regular basis. It can easily be assumed that this is a pure fame game, that is, the most famous and flashy artists garner the largest listening bases and thus will take the lead in a popularity contest. However, this assumption ignores both what got the artist to this point, and why people are still listening.

Whether it be the heaviest of heavy metal, the lightest pieces of classical music, or anything in between, a song can take off in popularity for a number of different reasons. To gain a more thorough and logical understanding of this concept, Spotify has computed numerical values for a wide variety of musical factors

that can be studied and compared to a song's popularity. This includes things like a song's; valence, energy, danceability, and many more specific traits. All of these characteristics result in songs that either gain massive popularity or succumb to total irrelevancy, never to be streamed again.

In this paper, this exact process was carried out. Using a generalized linear regression model in R [R Core Team (2020)], I conducted an analysis on the effects of a song's musical traits on its overall popularity to be able to provide a rudimentary framework for creating a 'hit' song in today's world. To accomplish this, I have made use of a large dataset containing data on over 170,000 songs on Spotify's catalog. The regression model was then applied to this dataset using its popularity measure as the dependent variable and a variety of characteristics as independent predictor variables.

The results of this analysis suggest that a song's popularity is strongly influenced by its danceability and energy level, with more energetic and danceable songs showing much higher popularity. Furthermore, a song's valence and speechiness are both strong predictors of its popularity as well, however, these qualities appear to be negatively correlated with the song's popularity.

An important area not covered by this analysis is the use of advertising space across Spotify's platforms. On all versions, Spotify picks a few artists or specific songs to advertise on a user's homepage. This makes it difficult to find a definitive answer as these advertisements can pull in very large groups of listeners and sway the popularity in the artist's favor. Moreover, based on a user's listening history, Spotify will often re-suggest these songs or others made by the same or similar artists providing the idea that a song's popularity may not be a case of linear growth due to a suggestion bias. While this makes it more difficult to uncover the truth behind what makes a song most popular, the results that have been generated provide a good initial idea of the most important factors in a song's overall popularity. As such, these findings can be taken advantage of by future artists that hope to give their music a slight edge over their competition.

To begin, Section 2 provides an overview of the dataset used for the analysis and the according modifications that were required. Section 3 lays out the development of the regression model used for this paper along with a brief explanation on the chosen model. In Section 4, the results of this model analysis are displayed along with their interpretations and a validation process is performed on the model. Section 5 features a discussion of these results and the conclusion that these results bring us to. Additionally, in this section, the weaknesses of this analysis are discussed as well as the areas for improvements and future work. Finally, Section 6 is an Appendix section featuring the code used for this analysis, and Section 7 provides references for all sources used for this paper.

2 Data Discussion

For the purposes of this analysis, one extremely large dataset containing information on over 170,000 songs in Spotify's catalog was used. This data was gathered using Spotify's Web API software and was set up and organized by a user on Kaggle. Kaggle is an online data science community that allows users to find and create datasets, statistical models, and data science challenges.

Spotify's Web API toolkit provides access to their massive catalog of data on many songs in their libraries. This data contains a number of calculated values based on the audio features of individual songs, surrounding categories like moods, properties, musical contexts, and much more. To collect the data, this requires setting up a developer account on Spotify's API platform, and then using a package like spotifyr [CITATION] to request this data. Unfortunately, requesting this data directly from Spotify provides information in a very raw format that requires extensive data cleaning in order to create a workable dataset. As such, there are very few large scale datasets available for this data as it is most often used for smaller-scale research. Luckily, however, a "self-taught data scientist and music enthusiast" on Kaggle [named Yamac Eren Ay], realized this problem and decided to tackle it by creating their own large scale and up to date dataset that covers a very large variety of musical characteristics on over 170,000 songs. They created 5 separate datasets to accomplish this. The one used for the purposes of this study was the basic data dataset which features information on each individual track. This contains data covering the tracks' artists, name, year, release date, duration, and a wide range of musical attribute values such as danceability, energy, key, popularity, and more. The

other datasets contain this same information but are instead categorized by a song's artists, genres, and year, where musical attribute values are computed as means for each category.

In order to study the relationship between a song's popularity and its musical characteristics, I have selected a few of these attribute values that I felt would be the most influential. Selected: popularity (how popular it is), valence (measure of happiness), danceability (measure of dance ableness), energy (measure of energeticness), instumentalness (how many instrument sounds included), speechiness (measure of spoken words included), tempo (a song's tempo).

To begin with cleaning this data, I first had to examine the types of songs included in the dataset. Upon inspection, this dataset appears to have included non-musical tracks along with regular music. This may cause a problem with the predictions of the model due to the fact that their calculated attribute values are much different that those of regular music. The issue with cleaning these songs out of the data, however, is that there is no formal designation for the type of track listed. In an attempt to work around this, I used details from Spotify's Web API documents that provided some insight on the values that these irregular tracks are given. One attribute that stands out as a definer for track type is speechiness, which determines the overall presence of spoken works in a track. In their documentation for this variable, tracks that are awarded scores greater than 0.66 are most likely made entirely of spoken works, these would be things like; audiobooks, podcasts, poetry, etc. A rough and easy solution I used was to simply remove songs with a score greater than 0.66, this works, but it may also mean that we lose some "actual" musical tracks that are just exceptionally speechy. As there is no clear way to work around this, this solution is what was used to continue.

Further exploration into the dependent variable, song popularity, reveals some very important information on its construction. As explained in their documents, Spotify computes a song's popularity based on the recent listening history of all users. This means that, naturally, songs from older eras will have disproportionately lower popularity scores than songs released in the recent past. Thus, the predicted values generated by the model will be suggestive of what makes a song most popular today, rather than an absolute definition of how to make the most popular song of all time. Moreover, due to this style of calculation, this leaves the popularity variable with many zero score songs, likely due to their high age or irregular track type. These zero scores will likely result in less strong predictions as their attribute values are still calculated as constants. Due to this bias, the decision was made to remove all songs with a popularity score of zero, while this may result in losing some songs that were really just a horrible mix of tunes, the increased performance of the model greatly outweighs this loss.

This decision, however, does raise some concerns with what exact value we should decide to make the boundary. It is likely that songs with popularity scores anywhere between 1 and 10 also have irregularities in their style, and thus should not be included, but as popularity score increases, the likelihood of more "proper" songs ending up cut out of the model will increase as well. Thus, zero is the safe choice that was used.

To begin exploration, using ggplot from the tidyverse package [Wickham et al., (2019).], the distributions of the variables were plotted to see how they are set up and check for any more irregularities.

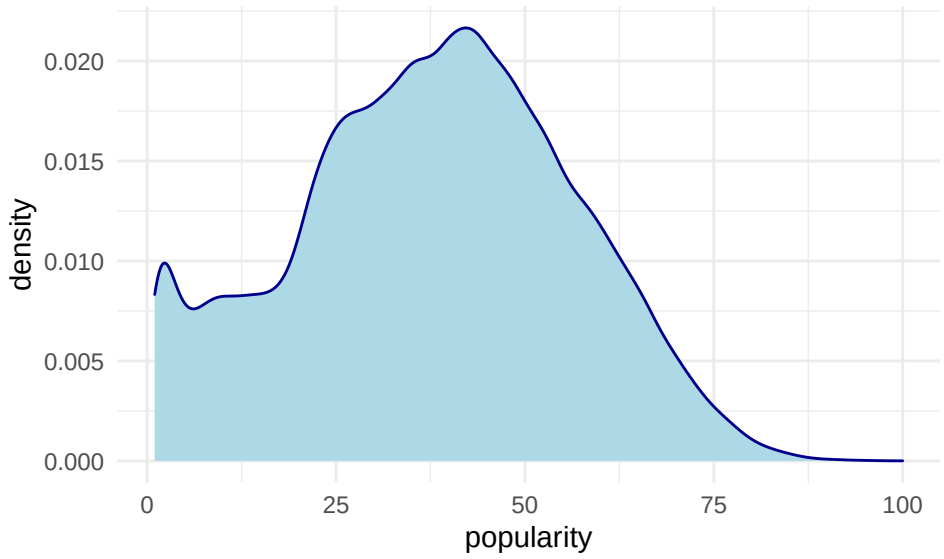


Figure ____ plots the density distribution of song popularity from the data. As we can see, this distribution is somewhat bell shaped, but grows extremely large as values approach zero. As was mentioned in the earlier data discussion, this is likely due to the number of older or irregular tracks that were not cut out of the final data. This will likely cause some problems with the model's predictive capabilities and must be kept in mind as we progress. Besides this issue, the data appears to be significantly clustered between scores of about 20 to 60, with very few songs reaching scores past 75. This observation helps to highlight the extremely competitive environment the music industry is in currently.

```
mean(data$popularity)
```

```
## [1] 37.82478
```

```
sd(data$popularity)
```

```
## [1] 18.26483
```

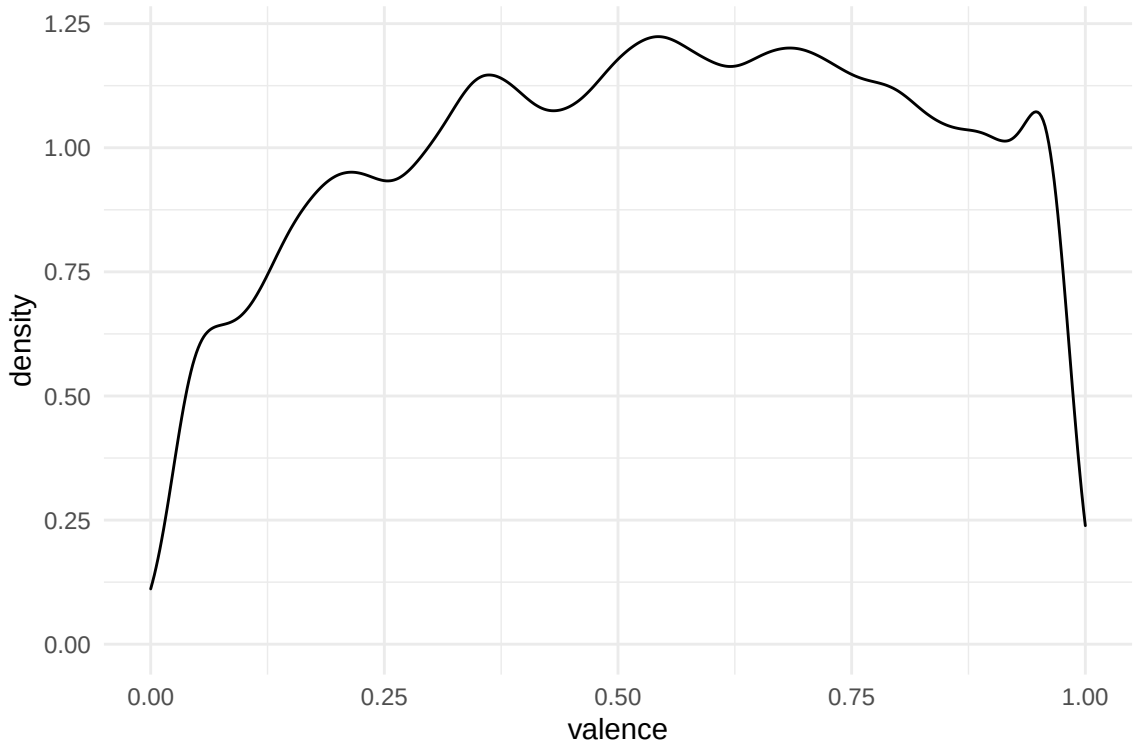
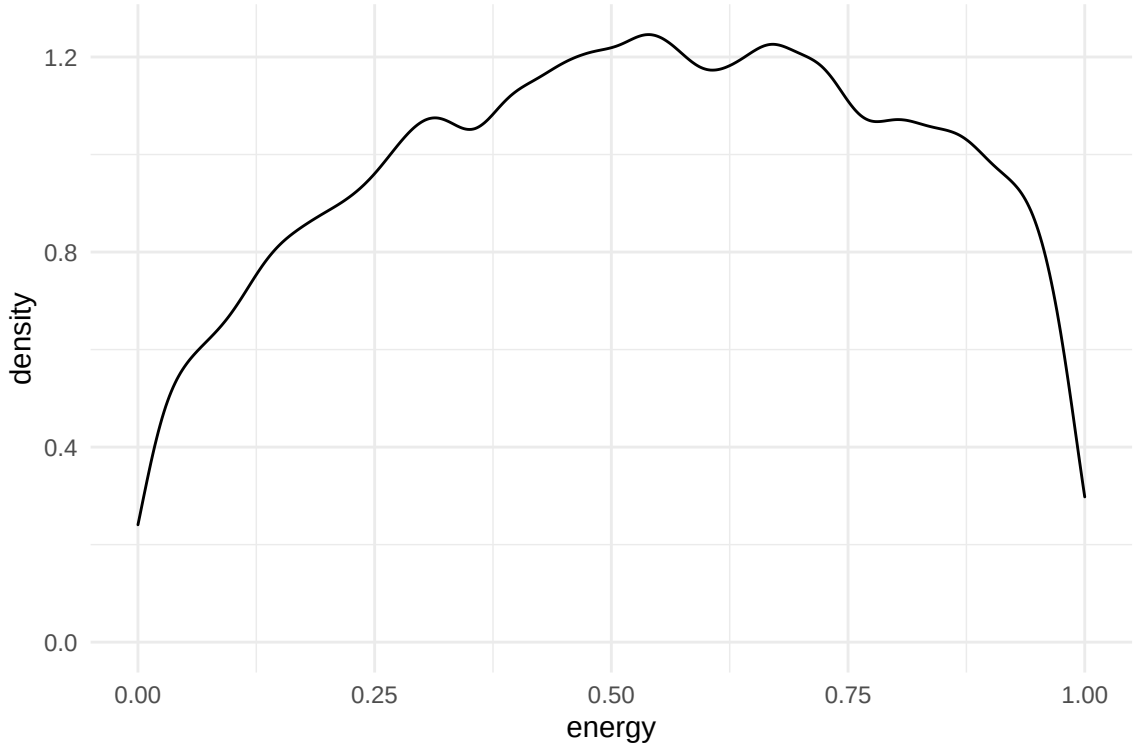


Figure _____ plots the distribution of valence scores given to songs in the dataset. A song's valence is a measure of the musical positiveness the track conveys, where moody and depressing songs score values closer to 0, and upbeat and cheerful songs score closer to 1. The distribution is very even but falls off as values get closer to either extreme. This is unsurprising given the wide variety of moods that artists choose to centre their music around.



In Figure _____, the distribution of energy scores for songs is displayed. A song's energy score is measured

from 0 to 1, and pertains specifically to its intensity and activity, with faster and louder songs earning scores closer to 1 . The distribution here appears to be highest around middle values suggesting somewhat of a balance between high and low energy attributes that most artists aim for. As was expected, this distribution quickly falls off as we approach either extreme.

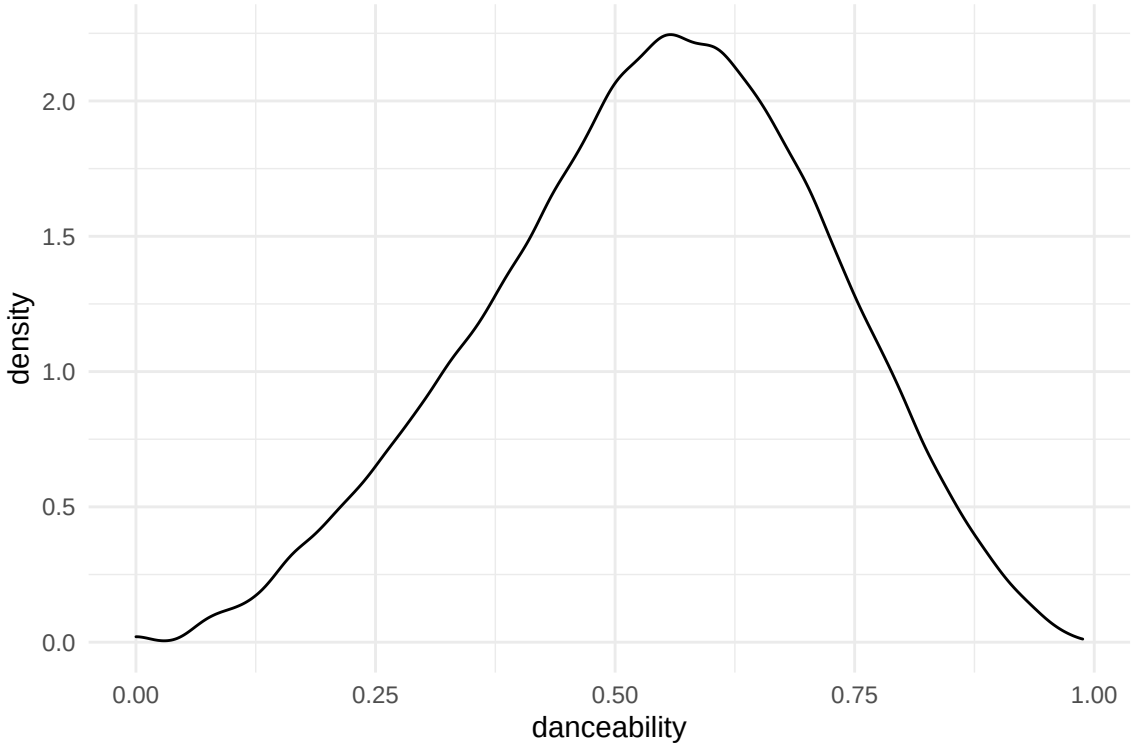
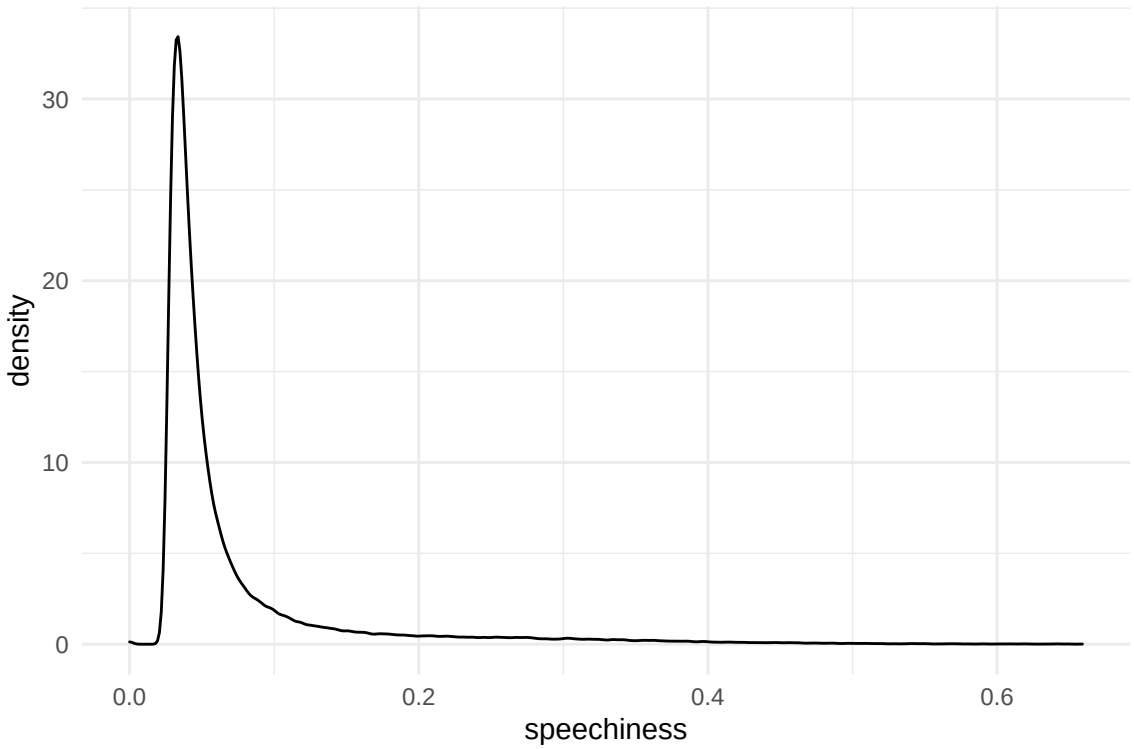


Figure _____ displays the distribution of danceability scores given to the songs. Danceability is a rather interesting variable that tracks how suitable a song is for dancing to. This value ranges from 0 to 1 and is based on a variety of musical attributes like tempo, rhythm stability, beat strength, and regularity. As we can see, the majority of songs have scores around 0.55 and resemble somewhat of a normal distribution as we go out from this point.



In Figure _____, the distribution for a song's speechiness is plotted. As mentioned earlier, speechiness measures the amount of spoken word featured in a track, and Spotify explains that tracks scoring over 0.66 are mostly entirely spoken word. Due to this, these tracks were removed from the final data and thus we see the distribution here from zero to less than 0.66. As we can see, the largest amount of this distribution is centered around 0.05, suggesting that most of the songs featured are largely instrumental. This may be partially due to the number of irregular tracks that were not cut out of the final data, however, it is difficult to cut out these values due to the extensive list of music genres that also feature little to no spoken words.

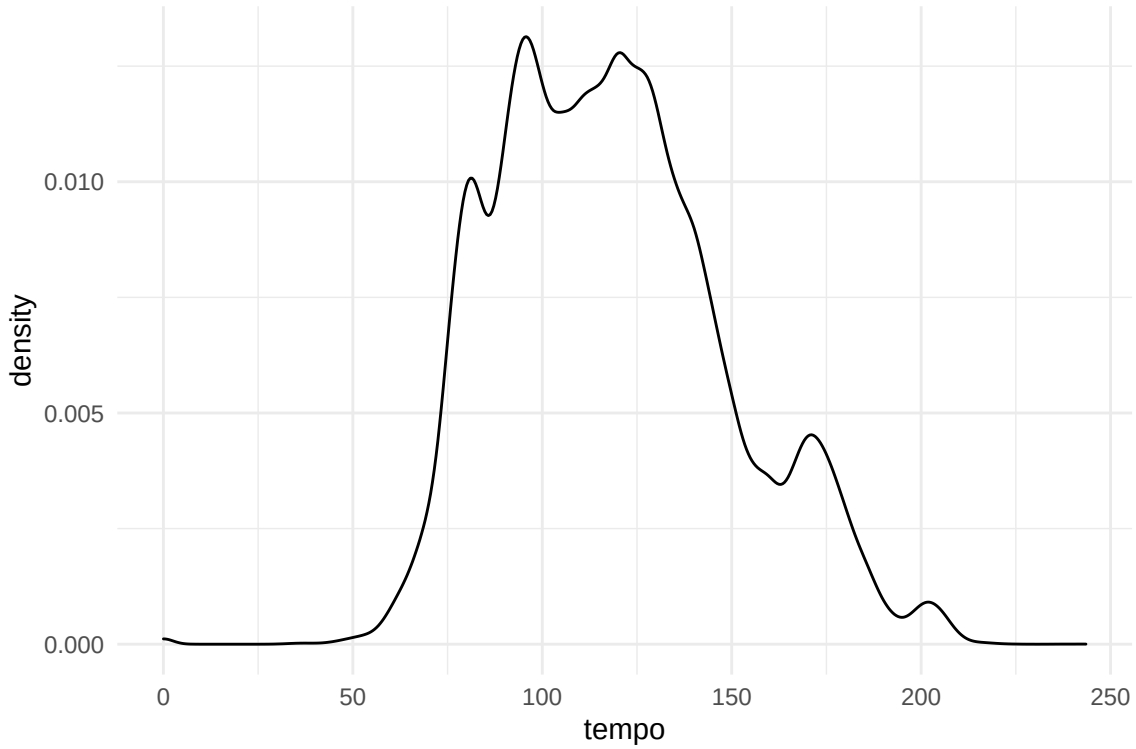


Figure ____ plots the distribution of the final variable of interest, tempo. This value simply tracks the overall tempo of a track, or speed and pace, in beats per minute. The distribution is heavily grouped around middle values, with very few songs featuring tempos below 60 or above 180. This observation is unsurprising when we consider what songs featuring these irregular tempos would actually sound like.

3 Model Development

For the purposes of this study, a generalized linear regression model of the negative binomial family was used to analyze the relationship between our dependent and independent variables. A generalized linear model is a generalization of ordinary least squares linear regression which removes the strict assumptions made by the linear model. This model then allows us to link our dependent variable through a specified probability distribution, which we refer to as the family of our regression model. Thus, unlike in linear regression, we do not need to have a normally distributed response variable.

This decision was made for a couple of reasons. First, since we are dealing with real and very messily distributed data, a regular linear regression model would not be sufficient as our data would not be able to satisfy any of its harsh restrictions. This includes things like a normally distributed error terms, linear relationships between variables, and homoscedasticity, all of which the Spotify data lacked. Thus, a generalized regression model was highly favoured here. To determine the family of the model this requires examining the design of our dependent variable, popularity. Popularity is measured as count data, which works specifically well with poisson style distributions. The issue with a poisson family, however, is that it requires equal dispersion, that is equivalent mean and variance values. This is not something that our popularity variable satisfies as we find a mean of 37.8 and variance of 18.3, thus we need to find a similar distribution that doesn't make this same requirement. This leaves us with a few options such as, negative binomial, or quasipoisson. These distributions both have somewhat similar requirements, but, after testing both models, the negative binomial model was favoured. The testing of the quasipoisson model is featured in Section 4.2, along with further model validation.

Another option that was considered at this step was to user a zero-inflated negative binomial regression model.

This would allow for the inclusion of the disproportiate amount of songs with a popularity score of zero, and thus would allow for a more complete analysis. However, given the irregular calculation of popularity score, it was decided that removing the zero score songs would be a better option. Thus, a negative binomial family was chosen for the regression model.

To create the model, I have chosen four predictor variables to measure against a song's popularity; danceability, speechiness, energy, and valence. The results of using these predictors in the generalized regression model selected are displayed in Section 4 below.

Some other options for predictor variables I had considered were instrumentalness, which predicts whether a song is entirely instrumental or not, and tempo, which simply tracks a song's tempo. These were not included in the final model, however, due to their distributions and significance. Since instrumentalness tracks the odds that a song contains no vocal sounds, this leaves an extremely high proportion of songs with values very close to, or exactly at zero. This leaves very little for our model to use to make predictions with and thus it was not considered a useful predictor. Furthermore, after including tempo into the final model, its coefficient estimate is extremely low, suggesting very little importance to the model, as such, this variable was discarded as well.

4 Model Results

4.1 Summary

Table _____

term	estimate	std.error	statistic	p.value
(Intercept)	3.04	0.01	597.67	0
danceability	0.85	0.01	84.60	0
speechiness	-0.10	0.02	-5.12	0
energy	0.86	0.01	143.78	0
valence	-0.64	0.01	-93.21	0

This section identifies the core predictions made by the regression model.

Table _____ displays the output from the model using summary statistics, this includes the coefficient's estimate, standard error, t-value, and p-value.

This table provides us with our first look at the final regression model we will be using:

$$\mathbf{log(P) = B0 + B1}\boldsymbol{D} \mathbf{+} \boldsymbol{B2}\mathbf{S + B3}\boldsymbol{E} \mathbf{+} \boldsymbol{B4}\mathbf{V}$$

Where: P represents the song's popularity. D represents the song's danceability score. S represents the song's speechiness level. E represents the song's energy level. V represents the song's valence. B0 represents the intercept of the model function. And the values B1 through B4 represent the coefficient estimates for each predictor.

To interpret both the impact and the strength of our predictor variables, we will pay attention to three key columns of this regression summary; Estimate, t-value, and p-value. The coefficient estimates tells us the log change in a song's popularity generated by a change in the predictor's value. Next, the t-values will tell us whether or not we can reject the null hypothesis, that our coefficient estimate is actually zero. Here, to conform with a 95% confidence interval, we will be looking for absolute values greater than 1.96. Lastly, a predictor's p-value works with the t-value to confirm our rejection of the null hypothesis. Keeping in line with the same 95% confidence interval, we require values less than 0.05. Independent variables that satisfy both of the t and p-value conditions we have described above will, as a result, allow us to reject the null hypothesis, and thus, prove to be significant predictors of a song's popularity.

Estimates: As we can see, danceability and energy both have positive coefficients in our model. This means that gearing music toward these characteristics, that is, music that is quite energetic and easy to dance to,

will result in higher levels of overall popularity. With an estimate of 0.86, energy appears to be the most significant factor in boosting song popularity, with a slight edge over danceability, which sits at an estimate of 0.85. On the other hand, both speechiness and valence have negative coefficients, suggesting that songs that include mostly spoken words, or are very upbeat and happy, are much less likely to reach significant levels of popularity. Valence appears to be the more significant driving factor here with an estimate of -0.64, whereas speechiness produces an estimate of -0.10.

P and t-values: When analyzing our chart we can clearly see that for all predictors, both their p and t-values satisfy the conditions needed on a 95% confidence interval. This means that we can comfortably reject the null hypothesis that their true coefficient values are equal to zero. Thus, we can move on to validating these results and discussing their meaning on a larger scale.

4.2 Model Validation

To validate the regression model I performed an MSPE and MSRes test on the data. This first involved splitting the data into two separate subsets. I decided to take out approximately 30% of the dataset, at 40,000 observations and used this set for MSPE calculations. The remaining 101,322 observations were then used for MSRes calculations. The goal of these calculations is to have similar mean squared errors between the two datasets.

Running these tests provide an MSPE of 0.31 and an MSRes of 0.8, thus giving us a difference of 0.49. While this is not a very large absolute distance, this is approximately a 158% percent difference which suggests a fairly high level of error with the model. This was somewhat expected given the wide variety factors that influence a music popularity as discussed in the introduction. As such, this result must remembered when considering the final predictions made by the model. Regardless, the final results of the model are still valuable when considering the relative effects of the selected musical attributes on a song's popularity.

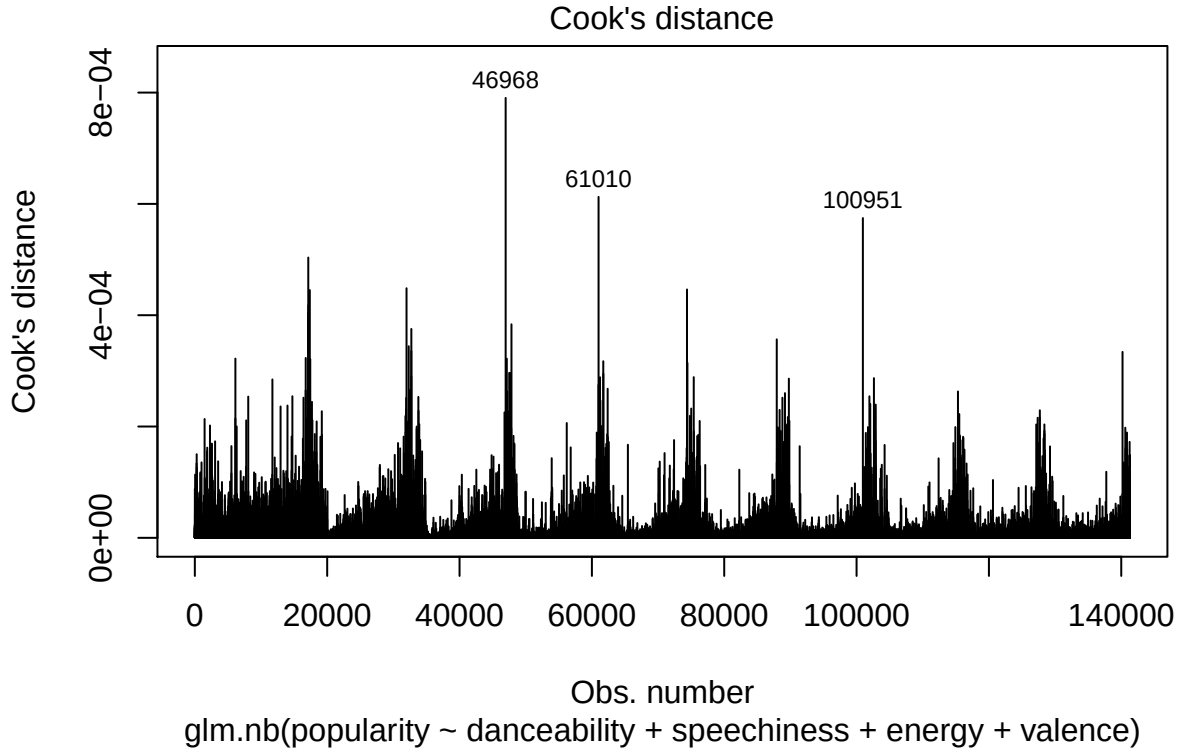


Figure ____ plots the cook's distance of all observations computed by the regression model. Cook's distance is an estimate of the influence of outlying observations in a regression analysis. When data points result in overly large errors this can negatively affect our model's predictions as they are considered to influence the

final estimate computed. To evaluate the results of this test we are looking for values greater than 0.00003. As we can see, there are a number of observations that fail this test. This will likely result in our model miscalculating a song's popularity score and so we must also keep this mind when considering the final results of the model.

5 Discussion

Song popularity is an extremely difficult thing to predict given the extensive amount of factors that go into creating a musical piece. The number of options an artist has when creating a song is almost infinite and these unique choices are impossible to totally account for. When we also consider the work that goes into things like advertising and branding, the list of factors that influence popularity grows even larger. Despite this, I have made my own attempt at uncovering some of these secrets using a statistical analysis in R and will discuss the results of this process below.

The use of the regression analysis on the Spotify data provides some very important information regarding how the chosen attributes affect song popularity. Firstly, it was found that both danceability and energy have strong positive correlations with song popularity. This suggests that songs that are easy to dance to and are full of energy have higher chances of gaining attraction. This result is unsurprising, naturally, songs that are easy to dance to often resonate more with listeners and can allow for a better connection to the music. These things are extremely important when it comes to the success of a song as they encourage listeners to share the music with others, exponentially boosting the song's popularity. Furthermore, a song that includes very little energy is likely to lose the attention of listeners quickly _____ whereas highly energetic songs, filled with beats that drive a person's ____ will be able to keep them engaged for much longer periods of time. Between these two factors, energy had a slightly stronger coefficient suggesting that listeners gravitate more toward a song's energy, rather than its ability to make them get up and dance.

On the other hand, a song's speechiness and valence are strongly negatively correlated with song popularity. This suggests that more somber music, driven more by beats than by words, are more likely to reach higher levels of fame. This is quite interesting given the widespread popularity of pop music which often features both heavy vocal usage and positive intentions. However, when we consider the extensive list of genres that usually feature no vocals, this result begins to make sense. For example, a large portion of electronic music uses this method as well as many _____. The negative correlation on valence is extremely interesting as well. This is somewhat surprising as we most often see happier songs taking the largest advertising space. However, again when we consider the massive groups of fans behind artists that specialize in these "sad" songs, this begins to make sense. This is likely due to the connection that is built between an artist and their fans when they can connect through their hardships.

While these results are not absolute, they do suggest significant trends behind song popularity. This can be extremely useful to up-and-coming musicians that hope to find new elements to include in their music. With this knowledge, they can begin to prioritize certain aspects of their songs that help give them a boost in popularity.

Limitations

There were many limitations when it came to analyzing this data, stemming both from the design of the variables, as well as their distributions. Spotify's Web API toolkit is an extremely useful tool, however, they made many decisions when calculating attribute values for songs. The issue with this is that these decisions are not entirely clear, the choices made in assigning certain values to songs are left out of their discussion documents and thus we are left to make an analysis with limited background information. This is likely to have resulted in miscalculated predictions in the model, but due to this lack of information, there is no easy way of retrieving better data to study from. Furthermore, these decisions that the Spotify developers made have resulted in some unusual distributions. This can be seen in variables like instrumentalness and speechiness where their designs lead to heavily skewed distributions centered very close to zero. These

irregularities cause issues with the model's predictive ability by adding more bias to the final results. Again, howeever, since these decisions were made by Spotify and are not extensively documented, there is very little room to work around these problems.

Another prominent limitation of this analysis was the limited use of explanatory variables. As regression models tend to run into overfitting when too many variables are included, this meant that the amount used in the final model had to be relatively low. Since there is such a massive list of musical factors that may affect popularity, this required cutting a majority of these out of the model. This is likely to lead to less strong predictive capabilities as the model will be unable to get a broader understanding of influences of song popularity.

Future Work

As a result of the expansive list of musical attributes that Spotify calculates for its songs, this analysis leaves a lot of room for future work. An easy start to this would be to incorporate additional attributes into the model, as well as create new models with alternate predictors. Some examples of explanatory variables that might be useful to the model include things like instrumentalness, tempo, and release year, as explained throughout the paper. This would allow for us to obtain a wider understanding of the musical factors that influence a song's popularity. Additionally, this can allow for stronger predictive capability of the model if a more optimal combination is found.

Due to Spotify's creation of the API data, using secondary datasets to further study the results is very difficult as this data has been exclusively set up by Spotify's team. This means that the majority of future work must take place using this data and making further attempts to extrapolate more information from it. Regardless, the massive amount of information that Spotify has already provided on their music leaves many opportunities for future work by itself and can lead us to stronger answers regarding how to make the most popular song of today.

6 Appendix

Code for this study can be found at:

7 References

Eren Ay, Yamac, 2020, 'Spotify Dataset 1921-2020, 160k+ Tracks', [Dataset], accessed via Kaggle kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks, as at 21 December 2020.

Angelo Canty and Brian Ripley (2020). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686