Statistics Project:
To what extent can a song's popularity be explained by its audio features?

Nicholas Tiveron
Ginevra Cepparulo
Cecilia Iacometta

Second semester 2022

1 Introduction

This project is based on the study of popularity of songs and its relation to some musical metrics. The Spotify API permitted us to create a data set based on the features we wanted to observe so that we could proceed with the data analysis and understand what influences the popularity of songs.
This topic is very relevant for musicians, music listeners as well as music streaming platforms such as Spotify and the music industry as a whole. As music listeners ourselves, we became intrigued by the factors that characterize the popularity of songs since it often happens that famous songs are quite similar to one another.
We have chosen to use the Spotify API in order to address our research question for several reasons. As consumers we are aware of the fact that the majority of the listeners have a Spotify account on which they follow artists, playlists, look top 200 daily hits, and hunt for music via a search bar. Moreover, we found out that Spotify counts more than 200 million global users and so it totally dominates the music streaming market. All of the benefits of our research illustrate the importance of data analysis for multinational companies with a global scope and for the single individual alike.

2 Research question

This paper will analyze Spotify-generated audio analysis metrics, over a sample of songs obtained from 10 most popular musical genres during the year 2021 in Italy, in order to find whether there exist some audio analysis features that can explain song popularity. It is believed that this endeavor is of particular importance given that Spotify pays record companies loyalties based on how many streams, and hence how popular the songs they produce are. Furthermore, the conclusions drawn from this statistical investigation could reveal interesting insights about musical tastes of the population from which the data is drawn. The focus of this study is an in-sample analysis on the degree to which Spotify's audio features can predict song popularity for already developed data. This is a necessary milestone for out-of-sample predictions. For this reason the research question is: *To what extent can a song's popularity be explained by its audio features?*

3 Variables and Hypothesis

Attribute	**Scale**	**Explanation**
Acousticness	0 - 1	Likelihood of a track being acoustic
Danceability	0 - 1	Suitability for dancing
Duration	ms	Length of the track
Energy	0 - 1	Measure of intensity and activity
Instrumentalness	0 - 1	Likelihood of having no vocals
Liveness	0 - 1	Likelihood of detecting an audience in the recording
Loudness	-60 - 0 (dB)	Loudness of a track
Mode	0 or 1	Indicates whether a song is Major (1) or minor (0)
Speechiness	0 - 1	Presence of spoken words in the track
Tempo	BPM	Estimated tempo of a track
Key	0 - 11	Key of the track, in pitch class notation
Valence	0 - 1	The musical positivity conveyed

Table 1: Features tracked

3.1 The dependent variable

Song popularity can be measured in several ways for example stream count or peak position on billboard charts. This investigation takes the approach of using Spotify's own popularity metric. Spotify has its own algorithm for determining this metric, they explain that it's mostly based on the total number of plays the track has and how recent those plays are. Meaning that songs that are being played a lot now have a higher popularity than songs that were played a lot in the past.

3.2 The independent variables

These attributes were extracted from the spotify database through the Spotify search API. All of their audio analysis features were selected as well some additional features; song name, respective artists and genres. Spotify keeps secret the algorithms to compute some of these metrics, for related business reasons. Given this as well as the recentness to which it was possible to capture some of these features, hypotheses will be based on intuition rather than theory.

3.2.1 Acousticness

Acousticness is a metric calculated by Spotify that measures the level of acousticness of a song and ranges from 0.0 to 1.0. Given the high current trend in electronic music in Italy in recent years it is believed that acousticness will show negative relationship with popularity.
Acousticness is negatively correlated to song popularity.

3.2.2 Danceability

Danceability is a metric calculated by Spotify and ranges from 0.0 to 1.0. It measures how suitable a track is for dancing. It could be said that danceability will be inversely related to acousticness. In fact, since electronic music is generally also danceable, it is believed that danceability will show a positive relationship with popularity.
Danceability is positively correlated to song popularity.

3.2.3 Duration

Duration is a discrete variable that measures the length in milliseconds (ms) of a song. In most modern musical genres lengthy songs are thought to bore the listener. Most songs today are within a restricted range; those that last beyond generally have lower popularity.
Duration is negatively correlated to song popularity.

3.2.4 Energy

Energy is a measure from 0.0 to 1.0 calculated by Spotify. It reflects how active and intense a song is. Spotify explains that it is based on dynamic range, perceived loudness, timbre, onset rate, and general entropy. Since today music is in many times an accompaniment to physical activities it makes sense that many popular songs are energetic.
Energy is positively correlated to song popularity.

3.2.5 Instrumentalness

Instrumentalness is calculated by Spotify to range between 0.0 and 1.0. This metric shows the level of instrumental music that the song contains. The least vocals a song contains the more instrumental it is. Given that lyrics allow the person to sing along to the song and hence to remember it more easily it is expected to be inversely related to song popularity.
Instrumentalness is negatively correlated to song popularity.

3.2.6 Key

Key is a measure of the key of a song. It is mapped to Pich Class using standard notation (for example, 0 = C, 4= E, 7 = G, 11 = B). In order to make a hypothesis about key it is necessary to have a musical theory background. Experts say that G,C and E are most commonly used and will therefore be seen in popular songs.
G, C and E are positively correlated to song popularity.

3.2.7 Liveness

Liveness is a metric that ranges between 0.0 and 1.0 and it detects the present of an audience in the audio recording. Since the presence of an audience and hence background noise is thought to lower the quality of the audio recording, it is believed that liveness will be negatively correlated with song popularity.
Liveness is negatively correlated to song popularity.

3.2.8 Loudness

Loudness is a discrete variable that measures the loudness of a song in the unit of decibels (db). Literature shows that values range from -60db to 0db. It is thought that loudness is the most common way to show expressiveness so we could then assume that the louder the song the more popular.
Loudness is positively correlated to song popularity.

3.2.9 Mode

Mode is a binary variable that represents the modality of the song; minor (0) or major(1). Songs written in major scales are cheerful whereas ones written in minor scales tend to be sad. It is believed that most people prefer to listen to happy music.
The Major mode is positively correlated to song popularity.

3.2.10 Speechiness

Speechiness is a variable calculated by Spotify that measures the presence of spoken words in a track. It ranges from 0.0 to 1.0, so the more speech-like the track the closer speechiness is to 1.0. It is believed that a medium level of speechiness is most popular nowadays and hence that speechiness is negatively correlated to song popularity.
Speechiness is negatively correlated to song popularity.

3.2.11 Tempo

Tempo is a metric that measures the average tempo of a track in beats per minute (Bpm). From music theory, we know that it describes the speed of the beat in a piece of music. The tempo is related to the feel or genre of the music. For example ballads generally have slow tempos while dance music have high ones. Given the assumption that there are both hip-hop and ballad tracks in the top charts, we believe that this variable will not show a correlation to song popularity.
Tempo is not correlated to song popularity.

3.2.12 Valence

Valence is a metric calculated by Spotify that ranges from 0.0 to 1.0 and measures how happy or sad the mood of a song is. The more cheerful it is the closer its valence value is to 1.0. Since it is believed that cheerful songs are more popular, it is hypothesized to behave positively with respect to song popularity.
Valence is positively correlated to song popularity.

4 Methodology

In this section we will explore the methodology of our research, from the data collection and transformation to a discussion of the models used and their assumptions. Firstly, it is imperative that we cover the basics of our sample, starting from its characteristics and why we decided to utilize it.

4.1 Definition of the sample

Because of the dominance of digital music in that field, we picked Spotify which is the most used platform as our criterion for determining the popularity of songs and all the other parameters that are automatically detected. Specifically, using the Spotify API which grants free access to developers to the Spotify database, we were able to select the top 50 songs (as established by Spotify's own metric) for each of the 10 most popular genres in the Italian market for the year 2021; one important detail is that the same song could be labeled with multiple genres meaning that some songs were duplicates and had to be removed and therefore the total number of songs that were considered was in fact 468, a bit less than 500.

4.2 Data collection

To compose the dataset we used the Spotify Search API which allows the user to input a series of parameters such as the genre, the year and the market and obtain a list of results which satisfy the criteria. We created a script which obtains all this information and uses it to make up a dataframe, thanks to the Spotify lightweight Python library. With the audio features API request we were able to obtain all the songs' features, which constitute the explanatory variables that we then used in the analysis process: the only exception is "popularity" which is one of the features provided but we used it as the response variable to address our research question. Thus, the dataframe we created has as rows all the 468 songs and as columns all their features.

4.3 Data transformation

By plotting different features against song popularity and against each other, we noticed some non linear relationships. Hence we decided to transform certain explanatory variables so that we could capture the linear relationships present and apply the analysis methods learned: for example by squaring the energy values and plotting them against acousticness we were able to obtain a linear graph. Additionally, to let our model be more free from assumptions, we included some interaction terms such as energy*instrumentalness to represent multiplicative effects and not be burdened by an additive assumption.

4.4 Assumptions of the linear model

The main technique that we used was Ordinary Least Squares regression which is used to estimate the coefficients of the linear regression equations which describe the relationships between the independent variable, popularity, and the dependent variables, which are the song features that have been selected. OLS achieves this by minimizing the sum of squared residuals, which are the differences between the observed values and the predicted ones.
This model however has five main assumptions to guarantee its effectiveness: linear relationship, multivariate normality, no multicollinearity if possible otherwise very little, no auto-correlation and homoscedasticity.
The assumption of linearity was checked by the entire analysis process which concludes that it is respected by conducting several tests like the t-test and looking at various statistics such as the p-value and the F-statistic. The distribution of popularity was plotted and it showed to closely resemble a normal distribution.
To address the problem of multicollinearity we checked that each of the explanatory variables we chose did not have high correlation coefficients with any other variable and at the end when we added more variables which inevitably had higher correlation values between them, we made sure to account for that by adding the interaction terms.
Since auto-correlation is only an issue in the case of time series, we didn't have such a problem in our analysis. Finally, we graphed several observed against residuals plots to check if any variables that we had added would have had an effect on the variance of the residuals thus violating homoscedasticity, however they all showed close to no sign of heteroscedasticity, except for the final model indicating that for that one the last assumption was not met and therefore that the linear model is not suitable for that specific choice of variables as we will discuss later.

5 Analysis

5.1 Correlation analysis

In order to select which variables to consider for our regression model we performed correlation analysis. In particular, we picked the 5 variables that had the highest correlation coefficient with respect to popularity.

- **Danceability**: 0.32
  Significant r value, was chosen.
- **Energy**: 0.24
  Significant r value, was chosen.
- **Loudness**: 0.29
  Significant r value, was chosen.
- **Speechiness**: 0.0026
  Non-significant r value, wasn't chosen.
- **Acousticness**: -0.22
  Significant r value, wasn't chosen.
- **Instrumentalness**: -0.3
  Significant r value, was chosen.
- **Liveness**: 0.072
  Non-significant r value, wasn't chosen.
- **Valence**: 0.27
  Significant r value, was chosen.
- **Tempo**: 0.1
  Significant r value, wasn't chosen.
- **Duration**: -0.063
  Non-significant r value, wasn't chosen.

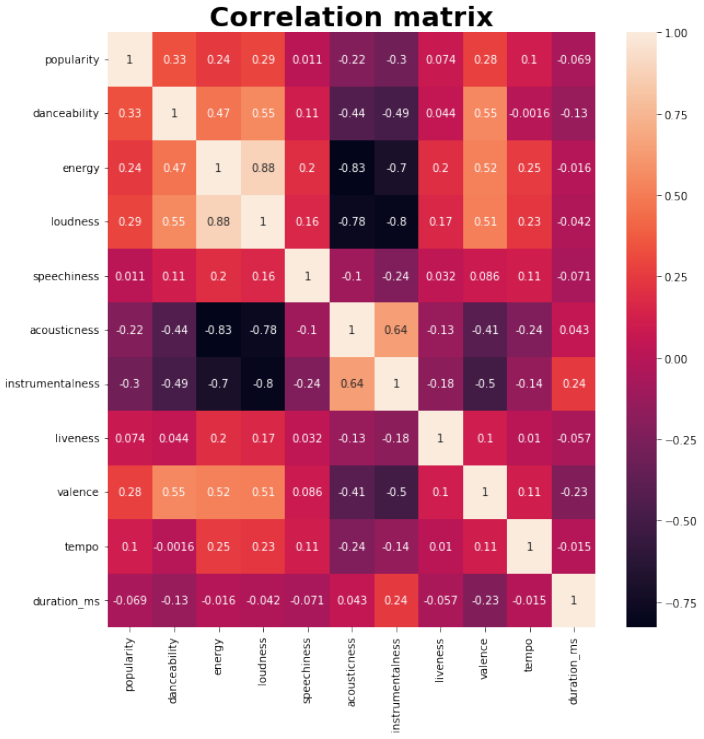

Figure 1: Correlation Matrix

5.2 Regression analysis

After determining which variables are significant what is left to ascertain is whether a combination of these features can give us a reliable model. In this section we will explore different results we obtained with various iterations of the OLS linear model.
The first attempt we made was by creating a linear model that had all the 5 selected variables as the explanatory variables, added by the forward stepwise regression method since we noticed that after adding each variable the explanatory power coefficient R-squared improved.

OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.141
Model:	OLS	Adj. R-squared:	0.132
Method:	Least Squares	F-statistic:	15.14
Date:	Mon, 05 Sep 2022	Prob (F-statistic):	8.68e-14
Time:	21:21:32	Log-Likelihood:	-1806.8
No. Observations:	468	AIC:	3626.
Df Residuals:	462	BIC:	3651.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	54.4574	5.234	10.405	0.000	44.173	64.742
danceability	15.5186	4.273	3.631	0.000	7.121	23.916
loudness	0.1236	0.196	0.631	0.528	-0.261	0.508
instrumentalness	-5.6404	2.898	-1.946	0.052	-11.336	0.055
valence	4.6986	3.033	1.549	0.122	-1.262	10.659
energy	-3.1330	4.792	-0.654	0.514	-12.549	6.283

Omnibus:	19.819	Durbin-Watson:	1.314
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.254
Skew:	-0.487	Prob(JB):	2.42e-05
Kurtosis:	3.377	Cond. No.	140.

Figure 2: OLS Regression Results 1

Since we had previously observed a high correlation between certain variables, to account for this issue without removing any variables we decided to add the 3 interaction terms and we observed a dramatic increase in the adjusted R-squared coefficient to signify that this model was significantly better than the one without the interaction terms.

OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.179
Model:	OLS	Adj. R-squared:	0.165
Method:	Least Squares	F-statistic:	12.54
Date:	Mon, 05 Sep 2022	Prob (F-statistic):	2.51e-16
Time:	21:21:34	Log-Likelihood:	-1796.1
No. Observations:	468	AIC:	3610.
Df Residuals:	459	BIC:	3648.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.4898	6.790	10.529	0.000	58.147	84.833
danceability	13.3468	4.329	3.083	0.002	4.840	21.853
loudness	1.1853	0.431	2.750	0.006	0.338	2.032
instrumentalness	-44.5254	11.761	-3.786	0.000	-67.638	-21.413
valence	3.4644	3.040	1.139	0.255	-2.511	9.439
energy	-13.1253	6.521	-2.013	0.045	-25.940	-0.311
energy:loudness	0.5053	0.599	0.843	0.399	-0.672	1.683
instrumentalness:loudness	-1.9264	0.579	-3.327	0.001	-3.064	-0.788
instrumentalness:energy	37.1574	12.678	2.931	0.004	12.243	62.072

Omnibus:	23.567	Durbin-Watson:	1.368
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.533
Skew:	-0.511	Prob(JB):	1.73e-06
Kurtosis:	3.561	Cond. No.	448.

Figure 3: OLS Regression Results 2

The empirical model that determines popularity can thus be specified as: $popularity = 72.85 + 12.21 * danceability + 1.25 * loudness - 46.00 * instrumentalness + 3.51 * valence - 13.48 * energy + 0.50 * energy * loudness - 2.01 * instrumentalness * loudness + 38.12 * instrumentalness * energy$.
The results tell us that popularity increases by 12.21 units if a song is most danceable and so has value 1.0 for danceability. If a song is most instrumental, popularity will decrease by 46.00. If a track is most happy and so has a valence of 1.0 then popularity will be positively affected through an increase of 3.51. If a track is most intense and active and so has energy 1.0 its popularity will decrease by 13.48. For every 1 unit increase in loudness (increased dcb value) popularity will increase by 1.25%. Then there are the interaction terms. The effect of energy for different values of loudness is positive with coefficient 0.5. The effect of instrumentalness for different values of loudness is -2.01 and the effect of instrumentalness for different values of energy is 38.12.
Now we shall interpret the output of our model using standardized coefficients and taking into account that a 1% increase for continuous variables measured between 0.00 and 1.00 is divided by 100 for correct interpretation. The results tell us that popularity increases by 0.1761% for every 1% increase in danceability. If instrumentalness increases by 1% popularity will decrease by 0.1227%. If a track increases by 1% in valence, popularity will be positively affected by an increase of 0.0654%. For every 1% increase in energy, popularity will decrease by 0.2347%. For every 1% increase in loudness, popularity will increase by 0.7282%. Then there are the interaction terms. A 1% increase in energy will increase the effect of loudness on popularity by 0.0706%. The effect of a 1% increase in instrumentalness will increase the effect of loudness on popularity by 0.3514%. If instrumentalness increases by 1% the effect of energy on popularity will increase by 0.2292%.

OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.180
Model:	OLS	Adj. R-squared:	0.166
Method:	Least Squares	F-statistic:	12.58
Date:	Tue, 06 Sep 2022	Prob (F-statistic):	2.22e-16
Time:	17:34:12	Log-Likelihood:	-617.67
No. Observations:	468	AIC:	1253.
Df Residuals:	459	BIC:	1291.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1837	0.068	-2.701	0.007	-0.317	-0.050
danceability	0.1761	0.057	3.090	0.002	0.064	0.288
loudness	0.7282	0.201	3.624	0.000	0.333	1.123
instrumentalness	-0.1227	0.078	-1.566	0.118	-0.277	0.031
valence	0.0654	0.056	1.166	0.244	-0.045	0.176
energy	-0.2347	0.103	-2.279	0.023	-0.437	-0.032
energy:loudness	0.0706	0.084	0.839	0.402	-0.095	0.236
instrumentalness:loudness	-0.3514	0.106	-3.327	0.001	-0.559	-0.144
instrumentalness:energy	0.2292	0.078	2.933	0.004	0.076	0.383

Omnibus:	23.344	Durbin-Watson:	1.368
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.228
Skew:	-0.509	Prob(JB):	2.02e-06

Figure 4: Standardized OLS

In spite of this, when we plotted the observed against fitted for this model we did not see a convincing linear trend, which was confirmed by the fact that in the observed vs residuals plot the residuals showed a linear relationship, indicating that the model as constructed wasn't a great linear fit for the data. This is confirmed by the Mallows C_p statistic which is higher than the threshold of $p+1$ (with p equal to the number of explanatory variables used) designating a biased model. Since in this model the p-value was very close to 0 but when looking at the individual p-values two variables (valence and the influence term of energy and loudness) had them much greater than the accepted threshold, it means that we have to reconsider this model but without those two variables, since the null hypothesis is not completely refuted because of them.
In the second model obtained by removing the variables with a large p-value from the ones of the previous model, we detected a slight decrease in the adjusted R-squared value accompanied by a Mallows C_p statistic that was still higher than the accepted value to consider the model unbiased.

OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.175
Model:	OLS	Adj. R-squared:	0.165
Method:	Least Squares	F-statistic:	16.35
Date:	Mon, 05 Sep 2022	Prob (F-statistic):	4.23e-17
Time:	21:21:37	Log-Likelihood:	-1797.2
No. Observations:	468	AIC:	3608.
Df Residuals:	461	BIC:	3637.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	72.3177	6.660	10.858	0.000	59.230	85.406
danceability	15.2639	3.981	3.834	0.000	7.440	23.087
loudness	1.4182	0.344	4.122	0.000	0.742	2.094
instrumentalness	-47.0232	11.395	-4.127	0.000	-69.416	-24.631
energy	-14.2320	5.693	-2.500	0.013	-25.420	-3.044
instrumentalness:loudness	-2.2148	0.487	-4.549	0.000	-3.172	-1.258
instrumentalness:energy	35.4014	12.624	2.804	0.005	10.594	60.208

Omnibus:	24.071	Durbin-Watson:	1.353
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.853
Skew:	-0.528	Prob(JB):	1.48e-06
Kurtosis:	3.513	Cond. No.	438.

Figure 5: OLS Regression Results 3

Overall, our best model was actually the first one even though it still was not a particularly great linear fit for the data. This could be explained by the fact that in the correlation matrix none of the variables displayed a very strong correlation with popularity. Since the t statistics proclaim those same variables to be significant we hence conclude that there may be better variables to answer to the goal of our research.

6 Conclusions

This research aimed to answer the question: *To what extent can a song's popularity be explained by its audio features?*. Through an analysis of the data it attempted to construct a linear model that would explain song popularity through audio analysis features. The initial findings about correlation analysis showed that some song features were related to popularity. Even more, several correlations between dependent variables were observed. While this gave hope for the production of a successful model, linear regression analysis resulted in a linear model constructed from a selection of features through a stepwise selection procedure having an adjusted R squared value of 16.40%. Given our model, it is concluded that it was not possible to explain a song's popularity by the analysis of its audio features.
The limited explanatory power of the model could be explained by several factors. For instance, it could be due to the metric used to describe a song's Popularity. Moreover, it could be related to the sample from which the data was taken, which spans several genres. Literature from Herremans showed that a genre focused analysis yielded conclusive results. In fact, genres can be identified through their audio features, therefore popular songs from different genres would likely show different characteristics.

6.1 Implications

Given the predictive power of our model, it was not possible to gather insights about song popularity in the sample taken. However it adds to the current literature available regarding Hit Song Science in exploring how song attributes can be used to predict the successfulness of a business inventure for Spotify. Further research could be conducted to observe whether other models other than OLS might yield better results. Moreover it forms a starting point for predicting, through the use of models such as Random Forests or SVM, song popularity for out-of-sample data.
Taking into account the academic implications that this research has brought, finding from further research might mean a significant difference for music streaming platforms such as Spotify as well as the music Industry in general. Specifically, it could incentivize new techniques for business planning creating new opportunities for value creation. Record companies might also become interested in buying popularity prediction software to use for example in the selection of new artists to invest in. This, however, begs the question of whether it could potentially stagnate the development of music as well as innovation and change in other arts and in general of ideology .

7 Works Cited

- Nijkamp, Rutger. Prediction of Product Success: Explaining Song Popularity by Audio Features from Spotify Data, University of Twente, 10 July 2018, https://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf.Accessed 6 Sept. 2022.
- Çimen, Ahmet, and Enis Kayış. A Longitudinal Model for Song Popularity Prediction, Ozyegin University, 2021, https://www.scitepress.org/Papers/2021/106077/106077.pdf. Accessed 6 Sept. 2022.
- Suh, Brendan Joseph. International Music Preferences: An Analysis of the Determinants of Song Popularity on Spotify for the U.S., Norway, Taiwan, Ecuador, and Costa Rica, Claremont Colleges, 29 Apr. 2019, https://core.ac.uk/download/pdf/216833222.pdf. Accessed 6 Sept. 2022.
- "F Statistic / F Value: Simple Definition and Interpretation." Statistics How To, 20 Nov. 2021, https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/. Accessed 6 Sept. 2022.
- Fogarty, Taylor, et al. "Predicting the Future of Music." Towards Data Science, 24 May 2019, https://towardsdatascience.com/predicting-the-future-of-music-c2ca274aea9f. Accessed 6 Sept. 2022.
- Frost, Jim. "How to Interpret p-Values and Coefficients in Regression Analysis." Statistics By Jim, Jim Frost, 22 July 2022, https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/. Accessed 6 Sept. 2022.
- Pham, James, et al. Predicting Song Popularity, http://cs229.stanford.edu/proj2015/140_report.pdf. Accessed 6 Sept. 2022.
- Watts, Cameron. "Extracting Song Data from the Spotify API Using Python." Extracting Song Data from the Spotify API Using Python, Towards Data Science, 10 Feb. 2022, https://towardsdatascience.com/extracting-song-data-from-the-spotify-api-using-python-b1e79388d50. Accessed 6 Sept. 2022.