Understanding Song Popularity using Spotify Dataset

C. DeWilde, D. Lee, G. Doran, J. Wang, and S.C. Vegunta

August 7, 2025

Introduction

Most of us have heard a new song for the first time and thought something like, *Wow! Who is this?!*

For many artists, "making it big" with a hit record is their ultimate dream. For thousands more employed across today's global music industry, producing hit songs, promoting them, distributing them, and organizing live events around them is not a dream — it's their everyday life.

As a cornerstone of the entertainment sector, today's $30 billion global music industry[1] is a powerful economic engine. In terms of entertainment spending, music revenue surpassed cinema and box office revenue for the first time globally in 2023; overall, 2023 was the biggest year in global music revenue in over two decades, with an additional 4.8% revenue growth in 2024[2]. The world's top streaming artists, such as Taylor Swift and Bad Bunny, have earnings upwards of hundreds of millions of U.S. dollars from streaming alone[3]. In an era with expanding demand and more access to music than ever before, there are growing financial incentives to nailing what makes a song popular.

So, what makes a song a hit? This report explores that elusive idea with data from the Spotify API. Spotify maintains a rich variety of quantitative musical descriptors of its catalogue, such as rhythm, valence, energy, and danceability. We investigate if and how a song's popularity on Spotify, the world's most widely used audio streaming service[4], can be described by its other musical features.

Our analysis explores the following research questions:

1. How well does danceability explain songs' popularity?
2. Does a model using danceability and loudness explain songs' popularity better than danceability alone?
3. Do danceability, loudness, and speechiness explain songs' popularity better than danceability and loudness, or danceability alone?

Data Set and Features Description

The dataset we are using for this project originates from TidyTuesdays, an open source weekly data science project that "provides real-world datasets so that people can learn to work with data"[5]. We specifically chose a dataset originally uploaded in January 2020 called "Spotify Songs". This dataset was originally compiled with spotifyr, "a R wrapper for pulling track audio features and other information from Spotify's Web API in bulk"[6].

The "Spotify Songs" dataset contains over 30,000 rows, where each unit of observation represents a track (or song) sampled from Spotify. The dataset was authored by data scientist Kaylin Pavlik, who sampled songs in Spotify's catalogue from six genres: edm, latin, pop, r&b, rap, and rock. Importantly, this API data pull provides valuable insights into the tracks that aren't available to the average Spotify user publicly. It's "likely that Spotify uses these features to power predictive products such as Spotify Radio,"[7] says Pavlik. Features include:

[1] Rechardt, L., "IFPI looks at a decade of digital transformation in the music industry," WIPO, April 23, 2025. Url: https://www.wipo.int/web/wipo-magazine/articles/ifpi-looks-at-a-decade-of-digital-transformation-in-the-music-industry-73661. Accessed: Aug. 5, 2025.

[2] Smirke, R., "IFPI Global Report 2024: Music Revenues Climb 10% to $28.6 Billion," Billboard, Url: https://www.billboard.com/business/business-news/ifpi-global-report-2024-music-business-revenue-market-share-1235637873/. Accessed: Aug. 5, 2025.

[3] IFPI Global Music Report 2025: State of the Industry. International Federation of the Phonographic Industry (IFPI), 2025. Url: https://www.ifpi.org/wp-content/uploads/2024/03/GMR2025_SOTI.pdf. Accessed: Aug. 5, 2025.

[4] "About Spotify," Spotify. Url: https://newsroom.spotify.com/company-info/. Accessed: Aug. 5, 2025.

[5] Harmon, J., "TidyTuesday," GitHub, Jan. 21, 2020 [Data Set]. Url: https://github.com/rfordatascience/tidytuesday/tree/main?tab=readme-ov-file. Accessed: Aug. 5, 2025.

[6] Thompson, C.; Parry, J.; Phipps, D.; Wolff, T., "spotifyr," spotifyr. Url: https://www.rcharlie.com/spotifyr/. Accessed: Aug. 5, 2025.

[7] K. Pavlik, "Understanding + classifying genres using Spotify audio features," Kaylin Pavlik, Dec. 20, 2019, https://www.kaylinpavlik.com/classifying-songs-genres/. Accessed: Aug. 5, 2025.

- Danceability: "Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity"[8]. Danceability is rated on a scale from 0 to 1.
- Loudness: Average overall loudness of a track in decibels. Decibels are measured on the log scale.
- Speechiness: "Detects the presence of spoken words in a track"[9]. Speechiness is rated on a scale from 0 to 1.

These variables will be vital to deriving insights as to why songs become popular.

The outcome variable that we are interested in is the "popularity" variable. This variable is a private metric that scores each track on a scale from 0 to 100 based on the number of streams, recency of streams, listener engagement, and playlisting[10]. This popularity score plays a huge role in how Spotify's algorithm shares and displays tracks to users.

Since the popularity is significantly affected by the genre of the track, for the purposes of this analysis we will focus only on tracks within the "pop" genre category.

Operationalization

To operationalize the concept of a "successful" or "hit" song, this analysis uses the Spotify popularity metric, interpreting higher popularity scores as indicative of broader commercial success.

Research into what influences songs' popularity has likened the virality of music to the contagious spread of actual viruses[11]. Just as viruses spread through close contact between people, often friends and family, the songs that make it to the top of the charts tend to share similar characteristics and are heavily influenced by proximity and affiliation. There is a recursive aspect to chart-topping performance; songs that score highest in popularity are most often in the Pop category. For this reason, this analysis is restricted to songs in the Pop genre.

We were interested in whether popularity could be described by other metrics of a song, and operationalized this by selecting other descriptive musical features from the data. Brief analysis, including printing summary statistics and simple scatterplots, revealed that danceability, loudness, and speechiness appeared to be positively correlated with popularity by visual cues alone (see Figure 1d), and warranted further investigation.

Kaylin Pavlik's extensive study[12] of the Spotify Songs dataset evaluated which features were most indicative of different genres. She found patterns: low danceability distinguished tracks in the rock category, for example, and EDM songs had a distinct combination of high tempo and low valence.

It is possible that danceability, loudness, and speechiness are more strongly associated with pop music than popularity itself. By design, pop is constructed using musical attributes that contribute to mass appeal. Does modern pop music include rap features because they are popular, which contributes to speechiness? Or is speechiness indicative of popularity because pop songs with a rap feature attract both audiences? Danceability and loudness may also be common characteristics of pop because they lend popularity to a song. Future analysis calls for research into the popularity of music, the industry of songwriting, and the curation of pop as a genre.

Data Wrangling and Visualization

Before deriving insights from the dataset, we made efforts to standardize data formatting and do sanity checks to ensure the validity of the data. These steps are as follows:

1. Check variable name standardization to snake_case for ease of use.
2. Ensure there are no duplicate tracks in the dataset.
    - We found that there were 4477 rows that were duplicate tracks. These rows were deleted.

[8] Harmon, J., "TidyTuesday," GitHub, Jan. 21, 2020 [Data Set]. Url: https://github.com/rfordatascience/tidytuesday/blob/main/data/2020/2020-01-21/readme.md. Accessed: Aug. 5, 2025.

[9] Harmon, J., "TidyTuesday," GitHub, Jan. 21, 2020 [Data Set]. Url: https://github.com/rfordatascience/tidytuesday/blob/main/data/2020/2020-01-21/readme.md. Accessed: Aug. 5, 2025.

[10] '&Tilly', "Spotify Popularity Score guide: what it is and when and why it matters for artists" SubmitHub, May 2025. Url: https://www.submithub.com/story/spotify-popularity-score-guide. Accessed: Aug. 5, 2025.

[11] Rosati, D. P., Woolhouse, M. H., Bolker, B. M., & Earn, D. J. D. (2021), "Modelling song popularity as a contagious process," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 477*(2253), 20210457. Url: https://doi.org/10.1098/rspa.2021.0457.

[12] K. Pavli, "Understanding + classifying genres using Spotify audio features," Kaylin Pavlik, Dec. 20, 2019, https://www.kaylinpavlik.com/classifying-songs-genres/. Accessed: Aug. 5, 2025.

3. Make sure variables are all the correct data type and format.
    - All variables were properly categorized as numerical/categorical. Converted duration variable from ms to min.
4. Check for missing values in dataset
    - No missing values were found in the dataset.
5. Check for zeros in columns where it would indicate poor data collection.
    - There are zeros in track_popularity, key, mode, and instrumentalness. However, these zeros are expected from these variables, so none of these zeros indicate poor data collection.
6. Filter database to remove outlier rows from the features that we're interested in.

Boxplot statistic analysis (summarized in Table 2, Appendix A) of the dataset features show outliers, including the features of interest (danceability, loudness, and speechiness). These outliers and their corresponding records in the dataset features of interest were removed for further analysis presented here.

Figure 1, in Appendix A, shows the dataset features' densities with and without outliers, correlation matrix, and scatter plots with standard error envelops. With the outliers in place, the danceability, loudness, and speechiness feature distributions are skewed (see Figure 1a, Appendix A), while with the outliers removed, the same features distributions are less skewed and more normal (see Figure 1b, Appendix A).

The results in the feature correlation matrix (in Figure 1c, Appendix A) and scatter plots (in Figure 1d, Appendix A) show that the selected features (anceability, loudness, and speechiness) have the highest positive correlations with the popularity feature. Also, the standard error envelops (red envelops surrounding the fitted-linear-regression red line in Figure 1d, Appendix A) of the selected features are mostly narrower, suggesting higher precision, reduced variability, and confidence in the data sample.

Model Specifications

To explore what characteristics contribute to the popularity of pop songs, three (3) linear regression models of increasing complexity as follows were selected:

- Model 1: Includes only *danceability* as the predictor. This baseline model estimates the average effect of rhythmic suitability on popularity.
- Model 2: Adds *loudness*, capturing whether perceived sound intensity contributes additional explanatory power.
- Model 3: Further includes *speechiness*, measuring the prevalence of spoken words in a track.

The dataset was split into two (2) sets as follows using random sampling (and setting a seed for reproducibility): 30% of the dataset was used as the *training set* and the remaining 70% of the dataset was used as a *testing set*. The seeding allows reproducibility of the analysis results following each analysis rerun.

All predictors (or independent features) have values that are continuous, and no transformations were applied, as the corresponding scatterplots (see Figure 1d, Appendix A) suggested their linear relationships with the dependent feature, *track popularity*. Correlation analysis (see Figure 1c, Appendix A) showed limited multi-collinearity, supporting the inclusion of all three (3) independent features.

The testing set was used to evaluate performance using Root Mean Square Error (RMSE). If RMSE drops noticeably between models, for example, comparing Model 1 and 2, it shows the added feature improves the model. This train-test workflow not only provides a set of clear coefficients to interpret, but it also provides a concrete measure of how much each audio feature improves the model's ability to explain a song's popularity.

Results and Discussion

Table 1 in Appendix A shows a comparative analysis and performance of the three (3) models.

The F-statistics of all three (3) models show that they have an overall significance, indicating that at least one independent variable has a non-zero coefficient. Similarly, the t-statistics and associated p-values of all three (3) models' coefficients also show that the selected variables (*danceability*, *loudness*, and *speechiness*) significantly predict the dependent variable (*track popularity*), holding other variables constant. However, since the maximum R^2 values of the three (3) models is 0.02, it suggests that the selected independent features in the three (3) models only explain a maximum of 2% of the dependent variable. Root Mean Square Error (RMSE), which measures the average difference between the values predicted by the model using the testing dataset and the actual observed values, for each model is also shown in Table 1 below. These RMSEs are high (with an average prediction error of 24.85 of the track popularity score, which has a scale 0–100), and they reduce very slightly with each model of increasingly complexity (i.e., moving from Model 1 to 3).

The above results, therefore, suggest that even though the developed models are statistically significant (i.e., the relationships are real and not just random noise), they are very poor models in terms of explaining the variance of the dependent variable (*track popularity*).

Figure 2 in Appendix A shows the diagnostic plots for all three (3) models (Model 1–3). The figure shows that all three (3) models' have similar characteristics as follows:

- Linearity appears to be acceptable.
- Normality of residuals is also largely acceptable, with some heavy tail characteristics.
- There is some evidence of heteroscedasticity—a key concern.
- Some points have higher leverage, requiring further investigation whether Cook's distance is also high.

Table 1: Linear Regression Results

	Dependent variable:		
	Song Popularity		
	(1)	(2)	(3)
Danceability	15.53 (5.73)	16.47 (5.71)	15.34 (5.71)
	$t = 2.71$	$t = 2.89$	$t = 2.69$
	$p = 0.01$	$p = 0.004$	$p = 0.01$
Loudness		1.28 (0.32)	1.14 (0.32)
		$t = 4.02$	$t = 3.53$
		$p = 0.0001$	$p = 0.0005$
Speechiness			68.91 (26.61)
			$t = 2.59$
			$p = 0.01$
Constant	37.21 (3.74)	44.21 (4.11)	40.27 (4.38)
	$t = 9.94$	$t = 10.75$	$t = 9.20$
	$p = 0.00$	$p = 0.00$	$p = 0.00$
Model Test RMSE	25.03	24.78	24.76
Observations	1,398	1,398	1,398
R^2	0.01	0.02	0.02
Adjusted R^2	0.005	0.02	0.02
Residual Std. Error	25.25 (df = 1396)	25.12 (df = 1395)	25.07 (df = 1394)
F Statistic	7.34^{***} (df = 1; 1396)	11.79^{***} (df = 2; 1395)	10.13^{***} (df = 3; 1394)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Analysis GitHub Repository

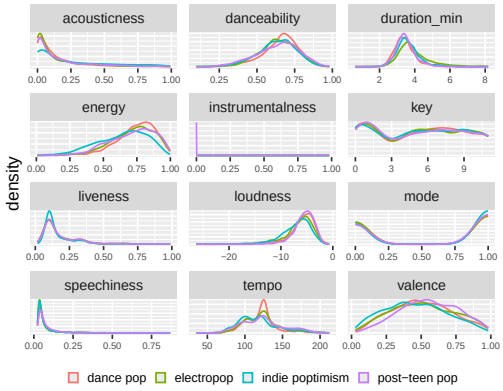
Relevant files used in performing the analysis summarized here can be accessed via the GitHub repository at the weblink as follows: https://github.com/mids-w203/Lab2DeWildeDoranLeeVeguntaWong.

Appendix A: Dataset Statistics and Visualization

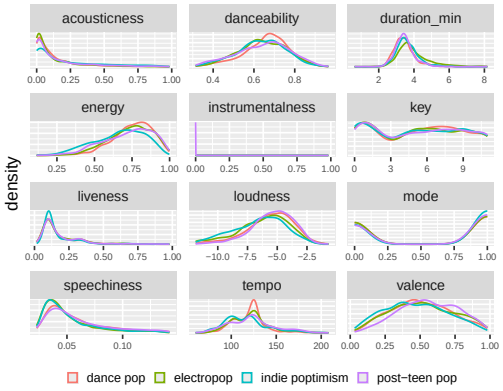
Table 2: Dataset Feature Stats

	Feature	Min	Q1	Median	Q3	Max	Outlier_Cnts
1	track_popularity	0	31	52	68	100	0
2	danceability	0.31	0.56	0.65	0.73	0.96	82
3	energy	0.24	0.59	0.73	0.83	1.00	69
4	key	0	2	5	9	11	0
5	loudness	-11.86	-7.51	-5.84	-4.57	-0.70	216
6	mode	0	0	1	1	1	0
7	speechiness	0.02	0.04	0.05	0.08	0.14	580
8	acousticness	0.0000	0.02	0.08	0.23	0.55	477
9	instrumentalness	0	0	0.0000	0.002	0.01	1,177
10	liveness	0.02	0.09	0.12	0.22	0.41	313
11	valence	0.03	0.34	0.50	0.67	0.98	0
12	tempo	62.48	102.99	120.02	130.09	170.12	287
13	duration_min	2.02	3.17	3.52	3.94	5.10	266

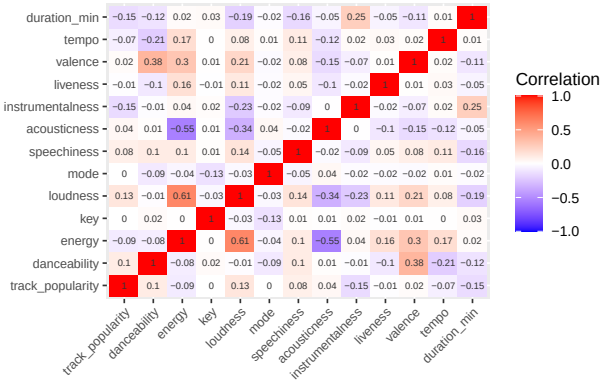
Note: Outliers are defined by the 1.5 IQR rule



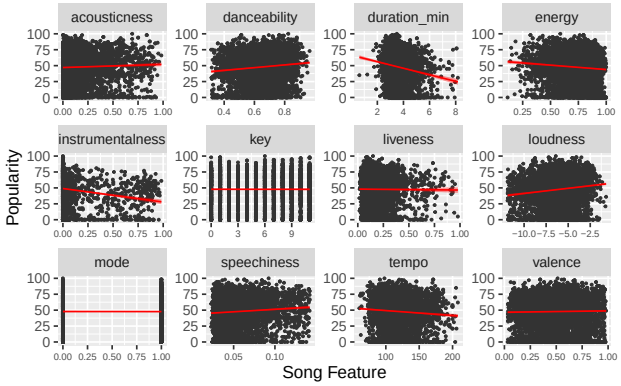
(a) Density Plots: with Outliers



(b) Density Plots: without Outliers in Selected Features

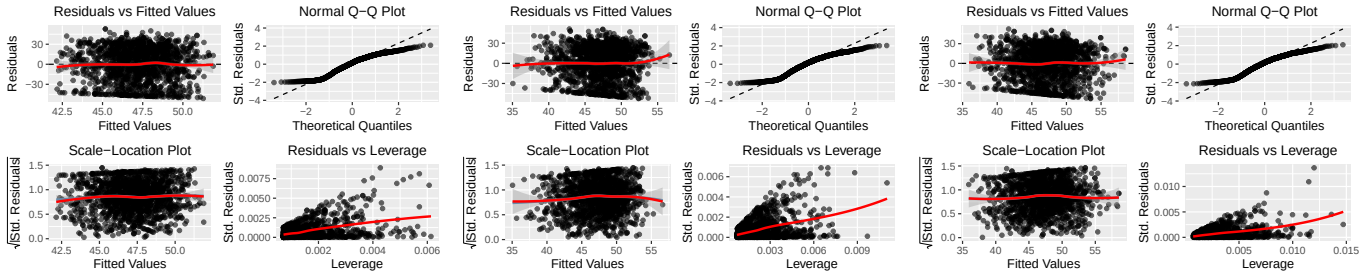


(c) Feature Correlation Matrix



(d) Scatter Plots with Std. Error Envelops

Figure 1: Dataset (for Pop Genre) Features' Summary



(a) Model 1 (b) Model 2 (c) Model 3

Figure 2: Diagnostic Plots

Appendix B: Additional Information

Dataset Source

R-code to access the dataset directly into rstudio is as follows:

```
# r-code to access the dataset
tuesdata <- tidytuesdayR::tt_load('2020-01-21')
ss <- tuesdata$spotify_songs
```

Weblink to access the TidyTuesday Spotify Songs dataset: https://github.com/rfordatascience/tidytuesday/blob/main/data/2020/2020-01-21/readme.md.

Model Specifications & Takeaways

- Model 1 (*Popularity* *Danceability*): A one point increase in danceability raises popularity by 15.53 points ($p<0.01$), but with RMSE = 25.03 and $R^2 = 0.01$, it captures only a tiny slice of the overall variance.
- Model 2 (*Popularity* *Danceability* + *Loudness*): Adding loudness (coef = 1.28, $p<0.01$) boosts danceability's effect to 16.47 and cuts RMSE to 24.78 ($R^2 = 0.02$), showing that track intensity meaningfully improves fit.
- Model 3 (*Popularity* *Danceability* + *Loudness* + *Speechiness*): Introducing speechiness (coef = 68.91, $p<0.01$) further trims RMSE to 24.76 without raising R^2, suggesting spoken versus sung content matters but yields diminishing returns on explained variance.

For the presented analysis, the *daceability*, *loudness*, *speechiness* features were selected because the researchers of this analysis believe that the selected features have a strong effect on a song's popularity.

- Danceability measures "how suitable a track is for dancing based on a combination of musical elements", which the researchers hypothesize is highly relevant to listener enjoyment, and thus popularity.
- Loudness is the average overall loudness of a track in decibels. We hypothesize that "louder" songs are more engaging and attention-grabbing, thus leading to more popularity.
- Finally, speechiness measures the presence of spoken words in a track. We hypothesize that more words/lyrics in a song means that the song has more chance to captivate listeners, thus leading to more popularity.

Residuals vs Fitted Values Plots

Please see Figure 2 for details.