# Project Milestone 2:

# Data Analysis on Spotify Music Data

Lauren Jun, Kayla Levy, Yousuf Qaum, Drew Wilenzick

**Table of Contents:**

## Datasets & Question

For our final project, we analyzed song performance on Spotify using a dataset of the most streamed songs in 2023. This dataset included song titles, artist names, release dates, streaming statistics, and several Spotify-specific music characteristics. Initially, our questions centered on song performance across different platforms (including minimal data from Apple Music and Deezer), the impact of time of year on song popularity, and the relationship between Spotify's music characteristics and song performance. Our primary focus in this investigation is now on the latter. Spotify's music characteristics include danceability, valence, energy, acousticness, instrumentalness, liveness, and speechiness[1], each scaled from 0 to 1. As avid Spotify users, we sought to derive insights from these attributes and learn more about the songs we stream. Using our original dataset, we developed a linear regression model incorporating these attributes along with other data from Spotify's API (e.g., key, playlists). However, the original dataset's limitation to the top 1,000 songs restricted our analysis to very popular tracks. To enhance our study, we utilized a larger dataset of over 40,000 songs, which included attributes and a classification of each song as a "hit" or "flop." Our logistic regression analysis explored how these attributes influenced a song's success. This expanded data set enabled a more comprehensive analysis of the factors contributing to a song's popularity.

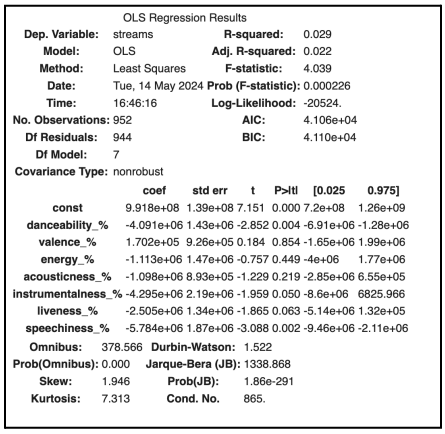
Our project goal aimed to determine if it is possible to predict the success of a new song in terms of streams using its Spotify attributes, thereby potentially aiding music producers and record label executives in optimizing releases, discovering new talent, and aligning strategies with evolving music consumption trends. Despite the complexity of modeling song performance due to various ways people discover music, our findings could be valuable for (1) optimizing music to listener preferences, (2) identifying new artists and songs in the music industry, and (3) understanding personal music tastes in comparison to mainstream trends.

## Analysis #1: Linear Regression

For this investigation, we aimed to explore how individual music attributes influence the stream counts of songs on Spotify. Musical attributes are expressed as percentages of the song's characteristics. For instance, valence describes the musical positiveness conveyed by a track. A track with high valence sounds more positive, while tracks with low valence sound more negative (e.g., sad, depressed, or angry). A full explanation of each musical attribute can be found in the appendix. We regressed these musical attributes on streams, resulting in the regression shown in Figure 1.

From the results, we observe that danceability and speechiness have p-values well below 0.05, making them statistically significant predictors of streams. Instrumentalness, with a p-value of 0.05, can also be considered a significant predictor. Interestingly, valence has a very high p-value of 0.854, indicating that the tone of the song may not significantly influence stream numbers. The model's $R^2$ value is 0.029, which is quite low and suggests that the model does not explain much of the variability in stream counts.

We then focused on a trimmed set of attributes—danceability, instrumentalness, liveness, and speechiness—because most of these were identified as statistically significant predictors in the initial regression analysis, aiming to improve the model's explanatory power and eliminate noise from less relevant attributes. In this refined model, all attributes except liveness remained statistically significant, reinforcing their importance to a song's stream count. Although the $R^2$ value is slightly lower, this is expected due to the fewer variables. These low $R^2$ values could be due to violations of some key linear regression assumptions, which we will check below.

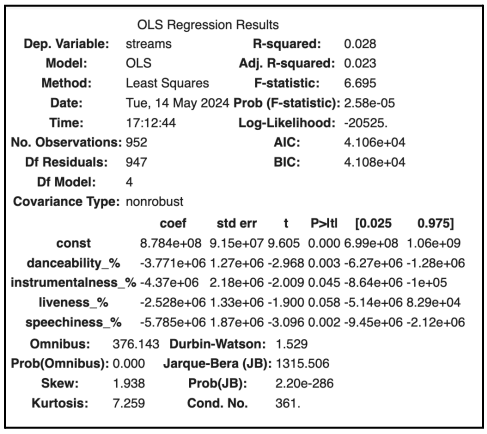


OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | streams | **R-squared:** | 0.029 |
| **Model:** | OLS | **Adj. R-squared:** | 0.022 |
| **Method:** | Least Squares | **F-statistic:** | 4.039 |
| **Date:** | Tue, 14 May 2024 | **Prob (F-statistic):** | 0.000226 |
| **Time:** | 16:46:16 | **Log-Likelihood:** | -20524. |
| **No. Observations:** | 952 | **AIC:** | 4.106e+04 |
| **Df Residuals:** | 944 | **BIC:** | 4.110e+04 |
| **Df Model:** | 7 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 9.918e+08 | 1.39e+08 | 7.151 | 0.000 | 7.2e+08 | 1.26e+09 |
| **danceability_%** | -4.091e+06 | 1.43e+06 | -2.852 | 0.004 | -6.91e+06 | -1.28e+06 |
| **valence_%** | 1.702e+05 | 9.26e+05 | 0.184 | 0.854 | -1.65e+06 | 1.99e+06 |
| **energy_%** | -1.113e+06 | 1.47e+06 | -0.757 | 0.449 | -4e+06 | 1.77e+06 |
| **acousticness_%** | -1.098e+06 | 8.93e+05 | -1.229 | 0.219 | -2.85e+06 | 6.55e+05 |
| **instrumentalness_%** | -4.295e+06 | 2.19e+06 | -1.959 | 0.050 | -8.6e+06 | 6825.966 |
| **liveness_%** | -2.505e+06 | 1.34e+06 | -1.865 | 0.063 | -5.14e+06 | 1.32e+05 |
| **speechiness_%** | -5.784e+06 | 1.87e+06 | -3.088 | 0.002 | -9.46e+06 | -2.11e+06 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 378.566 | **Durbin-Watson:** | 1.522 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1338.868 |
| **Skew:** | 1.946 | **Prob(JB):** | 1.86e-291 |
| **Kurtosis:** | 7.313 | **Cond. No.** | 865. |



OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | streams | **R-squared:** | 0.028 |
| **Model:** | OLS | **Adj. R-squared:** | 0.023 |
| **Method:** | Least Squares | **F-statistic:** | 6.695 |
| **Date:** | Tue, 14 May 2024 | **Prob (F-statistic):** | 2.58e-05 |
| **Time:** | 17:12:44 | **Log-Likelihood:** | -20525. |
| **No. Observations:** | 952 | **AIC:** | 4.106e+04 |
| **Df Residuals:** | 947 | **BIC:** | 4.108e+04 |
| **Df Model:** | 4 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 8.784e+08 | 9.15e+07 | 9.605 | 0.000 | 6.99e+08 | 1.06e+09 |
| **danceability_%** | -3.771e+06 | 1.27e+06 | -2.968 | 0.003 | -6.27e+06 | -1.28e+06 |
| **instrumentalness_%** | -4.37e+06 | 2.18e+06 | -2.009 | 0.045 | -8.64e+06 | -1e+05 |
| **liveness_%** | -2.528e+06 | 1.33e+06 | -1.900 | 0.058 | -5.14e+06 | 8.29e+04 |
| **speechiness_%** | -5.785e+06 | 1.87e+06 | -3.096 | 0.002 | -9.45e+06 | -2.12e+06 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 376.143 | **Durbin-Watson:** | 1.529 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1315.506 |
| **Skew:** | 1.938 | **Prob(JB):** | 2.20e-286 |
| **Kurtosis:** | 7.259 | **Cond. No.** | 361. |

**[Figure 1 (Left) ] Linear Model Predicting Streams Based on Musical Attributes**

**[Figure 2 (Right) ] Linear Model Predicting Streams Based on A Curated Selection of Musical Attributes**

To conclude, our analysis identified danceability, instrumentalness, and speechiness as the most valuable predictors of song streams on Spotify, indicating that tracks suited for dancing, with fewer vocals, and some presence of spoken words tend to attract more listeners. By optimizing these attributes, we may be able to predict the potential popularity of a song more accurately. Although the model's $R^2$ value suggests room for improvement, focusing on these key attributes may provide valuable insights for artists and producers aiming to maximize their streaming numbers.

We then explored how the number of playlists a song is featured in and the speechiness percentage influence stream counts on Spotify. We selected speechiness because it was the most statistically significant variable in our previous regressions and is relevant since most songs include vocals. Additionally, we included the number of playlists a song is featured in due to its clear logical connection to stream counts. While it is intuitive that a song featured in many playlists would garner more streams, quantifying this relationship helps us better understand the magnitude of its impact.

Our results, shown in Figure 3, indicate that both variables are statistically significant, with an $R^2$ value of 0.626, which is a substantial improvement over our previous regressions. The coefficient for speechiness is -2.379e+06, indicating that the more spoken words in a song, the fewer streams it would have on average. However, because speechiness ranges from 0 to 1, this coefficient suggests that the maximum potential decrease in streams due to speechiness is around 2.37 million, which is relatively small compared to the billions of streams these songs can and have achieved.

Additionally, when we plotted predicted versus actual streams, we observed that the model's data points initially cluster closely along the 1:1 diagonal line on the left side, indicating accurate predictions for songs with lower stream counts. However, as the stream counts increase, the model's predictions become increasingly inaccurate and variable, deviating significantly from the diagonal line. This pattern suggests that the model may be influenced by omitted variables, non-linear relationships, or heteroscedasticity, which could be contributing to the observed discrepancies.

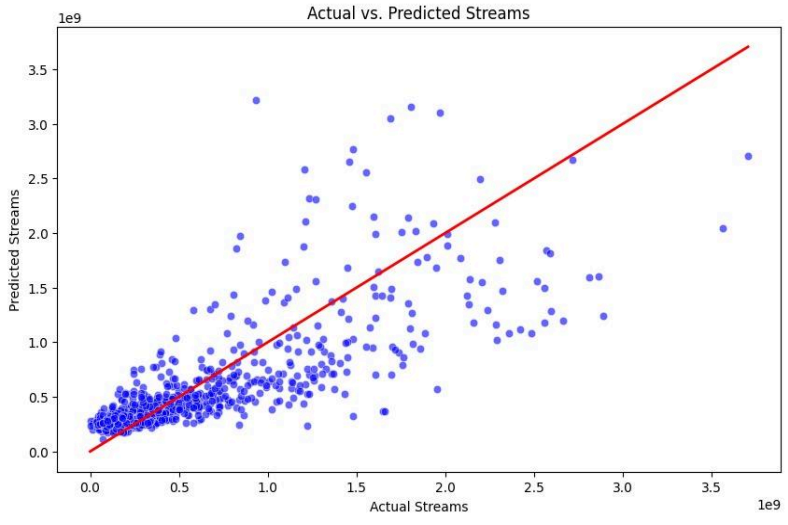
OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | streams | **R-squared:** | 0.626 |
| **Model:** | OLS | **Adj. R-squared:** | 0.625 |
| **Method:** | Least Squares | **F-statistic:** | 792.6 |
| **Date:** | Tue, 14 May 2024 | **Prob (F-statistic):** | 3.83e-203 |
| **Time:** | 21:11:03 | **Log-Likelihood:** | -20071. |
| **No. Observations:** | 952 | **AIC:** | 4.015e+04 |
| **Df Residuals:** | 949 | **BIC:** | 4.016e+04 |
| **Df Model:** | 2 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 2.449e+08 | 1.82e+07 | 13.454 | 0.000 | 2.09e+08 | 2.81e+08 |
| **in_spotify_playlists** | 5.639e+04 | 1430.890 | 39.412 | 0.000 | 5.36e+04 | 5.92e+04 |
| **speechiness_pct** | -2.379e+06 | 1.14e+06 | -2.086 | 0.037 | -4.62e+06 | -1.41e+05 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 203.754 | **Durbin-Watson:** | 1.692 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1673.614 |
| **Skew:** | 0.734 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 9.328 | **Cond. No.** | 1.53e+04 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.53e+04. This might indicate that there are strong multicollinearity or other numerical problems.




**[Figure Three (Left) ] Linear Model Predicting Streams Based on Spotify Playlists and Speechiness**

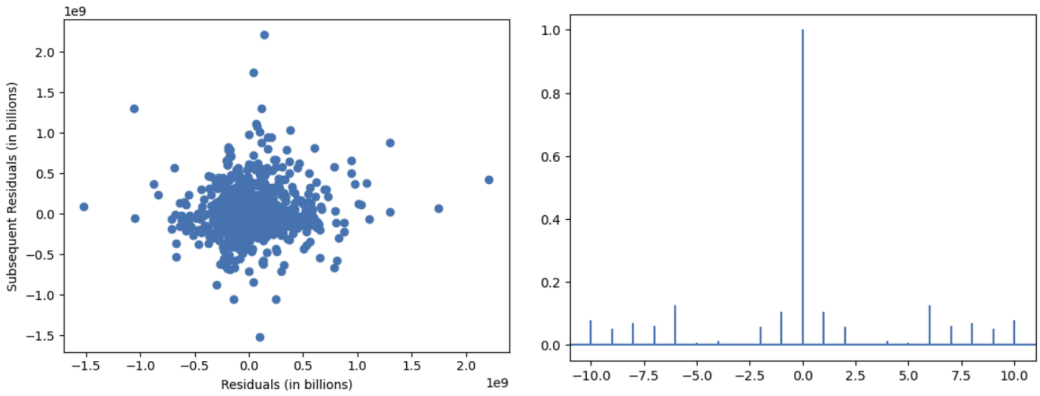
**[Figure Four (Right) ] Prediction Accuracy Plot of Regression Using # of Spotify Playlists, Speechiness (0-1)**

In our discussion of the results, we first acknowledge the significant bias in our initial dataset, which included only the top-performing songs on Spotify. This inherent bias limited our ability to generalize findings to a broader range of songs, prompting us to use a second, larger dataset for more comprehensive analysis. Despite the improved $R^2$ value of 0.626 and the statistical significance of both variables, our model still shows inaccuracies for higher stream counts, suggesting the need to address potential omitted variables and non-linear relationships. The relatively small impact of speechiness relative to stream counts highlights the complexity of predicting song popularity on Spotify – our third method of analysis aims to more accurately understand if these attributes are related to a song's success.

## Analysis #2: Checking the Assumptions of Linear Regression

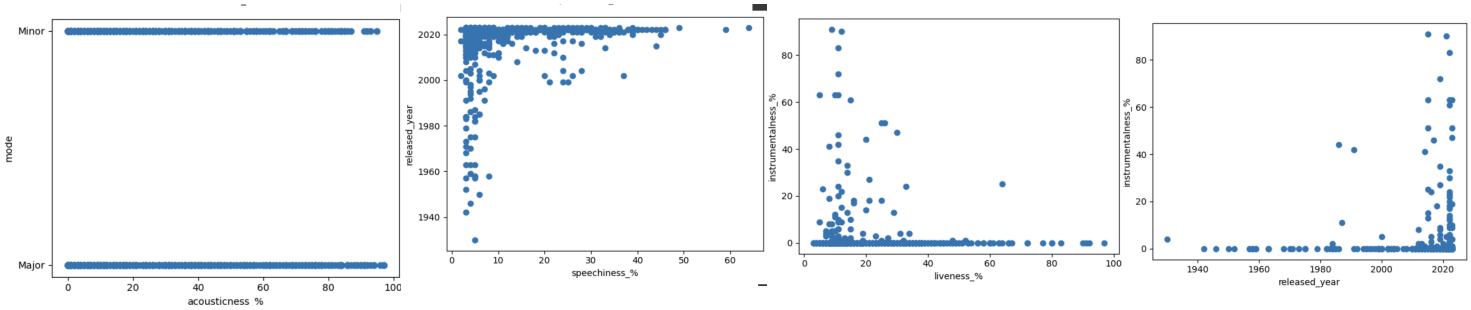
### 1. Checking if residuals are mutually independent

In order to check assumptions of linear regression, we first checked the residuals for mutual independence using two different methods. The first method involved plotting errors to detect any patterns (i.e. the graph to the left below). The absence of a pattern indicates mutual independence. The second method utilized the autocorrelation function (ACF) (i.e. the graph to the right below), where we assessed how frequently autocorrelations appeared outside the test bounds. Apart from the spike at 0.0, there were few autocorrelations outside the test bounds. The ACF plot displays the correlation coefficients between the residuals at different lags. Thus, identifying autocorrelation is crucial for ensuring the validity of the regression model.



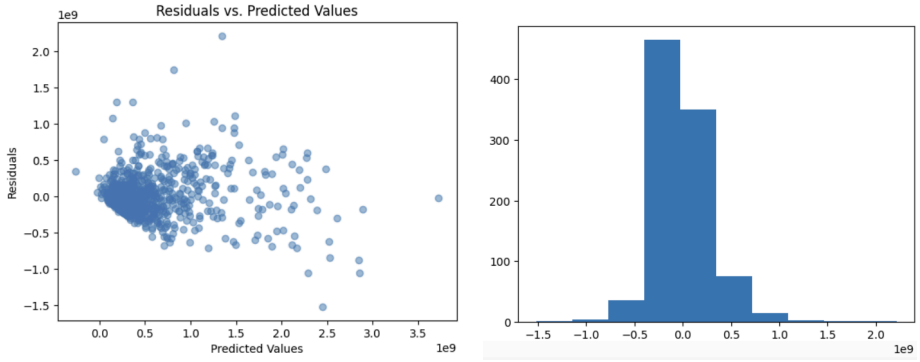

### 2. Independent of the covariates

We assessed the independence of the covariates by testing linearity for any relationship between the residuals and the x values. In the plots below generated from our code, we observed no discernible pattern, indicating independence of the covariates from the predictors.



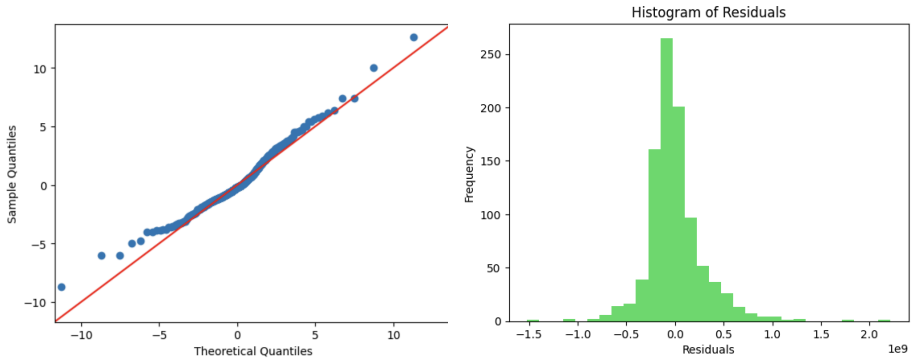

### 3. Plotting residuals against the predicted values – checking the assumption of homoscedasticity

Additionally, plotting the residuals against the predicted values is significant for checking the assumption of homoscedasticity, which states that the variance of the residuals should be constant across all levels of the predictors. In the graph below, we can visually inspect if the variability of the residuals changes as the predicted values change. The change in the spread of the residuals as the predicted values increase or decrease can indicate a violation in heteroscedasticity, thus violating the assumption of constant variance.




### 4. Checking if residuals are normally distributed

We also modeled a Q-Q plot and a Histogram of Residuals to check for normality of residuals. Normally distributed residuals are indicated by a Q-Q plot that is nearly a straight line. However, in our Q-Q plot, we can see that a pattern exists (i.e. could potentially indicate right skewness). The Q-Q plot compares the distribution of the residuals to a theoretical normal distribution. If the residuals are normally distributed, we can infer that the points on the Q-Q plot will fall approximately along a diagonal line. So this condition may be violated.




After testing our dataset against the assumptions, we found that while some passed, others failed. Specifically, the assumptions of having mutually independent residuals and independence of covariates were met. However, the assumptions of normally distributed residuals and homoscedasticity were violated. As a result, we should consider using a new dataset for testing logistic regression, including both top songs and less popular ones, to ensure broader representation and validity in our analysis.

## Analysis #3: Logistic Regression - What makes a hit song?

The guiding question of our project has been how a song's attributes reflect and impact its success in the industry and on Spotify. Earlier, we ran a linear regression on Spotify's top songs, aiming to analyze the effect of those attributes. For our logistic regression, we aimed to determine if and how attributes of a song make it a hit or a flop. To do so, we researched and found a larger dataset that included the same song attributes as scraped by Spotify's API, but instead of including only the top 1000 songs from a single year, included data on 5000-9000 songs from each decade from the 1960s to the 2010s. In addition, this dataset had a variable identifying the song as a "flop" (0) or a "hit" (1) which we used as our dependent variable **y** (note not all songs are included, just those that can be classified as 0/1):

**Y = 1 (hit)** if this song has been featured in the weekly list (Issued by Billboards) of Hot-100 tracks in that decade at least once and is, therefore, a 'hit'

**Y = 0 (flop) if:** The track does not appear in the 'hit' list of that decade, The track's artist does not appear in the 'hit' list of that decade, The track belongs to a genre that could be considered non-mainstream and/or avant-garde (these are documented by Spotify and are in our dataset), The track's genre does not have a song in the 'hit' list.

The independent variables we used are similar to those in our linear regression analysis, including danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.

Logit Regression Results

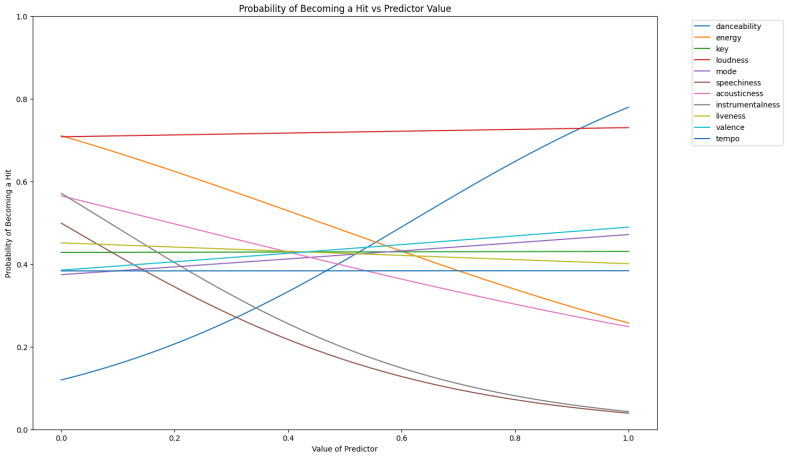
| | | | | | | |
|---|---|---|---|---|---|---|
| **Dep. Variable:** | target | | **No. Observations:** | | 41106 | |
| **Model:** | Logit | | **Df Residuals:** | | 41090 | |
| **Method:** | MLE | | **Df Model:** | | 15 | |
| **Date:** | Tue, 14 May 2024 | | **Pseudo R-squ.:** | | 0.2390 | |
| **Time:** | 17:10:06 | | **Log-Likelihood:** | | -21683. | |
| **converged:** | True | | **LL-Null:** | | -28493. | |
| **Covariance Type:** | nonrobust | | **LLR p-value:** | | 0.000 | |
| | **coef** | **std err** | **z** | **P>\|z\|** | **[0.025** | **0.975]** |
| **const** | 0.4827 | 0.175 | 2.760 | 0.006 | 0.140 | 0.826 |
| **danceability** | 3.3016 | 0.093 | 35.625 | 0.000 | 3.120 | 3.483 |
| **energy** | -1.9241 | 0.100 | -19.233 | 0.000 | -2.120 | -1.728 |
| **key** | 0.0101 | 0.003 | 3.031 | 0.002 | 0.004 | 0.017 |
| **loudness** | 0.1072 | 0.004 | 24.285 | 0.000 | 0.099 | 0.116 |
| **mode** | 0.3882 | 0.026 | 14.830 | 0.000 | 0.337 | 0.439 |
| **speechiness** | -3.2196 | 0.157 | -20.567 | 0.000 | -3.526 | -2.913 |
| **acousticness** | -1.4013 | 0.053 | -26.459 | 0.000 | -1.505 | -1.298 |
| **instrumentalness** | -3.3709 | 0.067 | -50.268 | 0.000 | -3.502 | -3.239 |
| **liveness** | -0.2026 | 0.069 | -2.921 | 0.003 | -0.339 | -0.067 |
| **valence** | 0.3583 | 0.060 | 5.956 | 0.000 | 0.240 | 0.476 |
| **tempo** | 0.0021 | 0.000 | 4.848 | 0.000 | 0.001 | 0.003 |

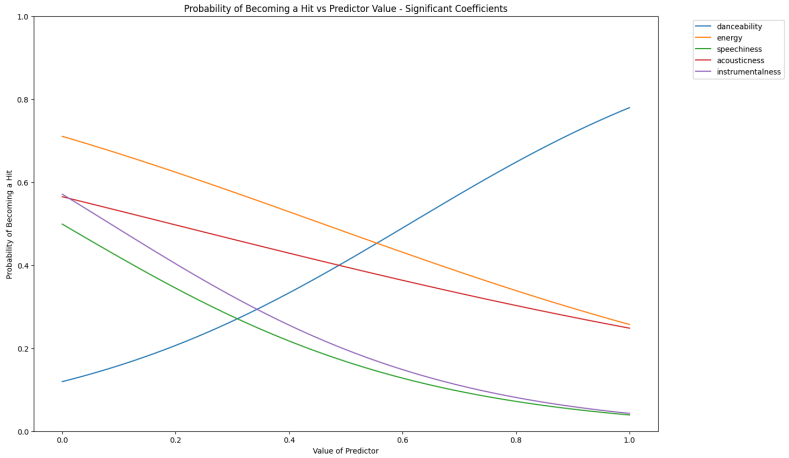
**[Figure Five] (Right) Logistic Regression Model Trained on 1960s-2010s flop or hit data**

Upon analyzing this model, we see that our p-values are quite low, suggesting that our predictors are significant.

First, we visualized, based on our coefficients, the implied probability of a song being a “hit” given the value of each covariate (ranging from 0 to 1) while keeping the remaining attributes equal to the average (note: the lines are independent, for each line, the rest of the attributes are average other than the noted one). We then replotted the significant relationships. We the most significant increase in probability when raising danceability, and a significant decrease when increasing acousticness and speechiness.

**[Figure Six ] Implied probability from modifying covariates (Left - All, Right - Significant)**



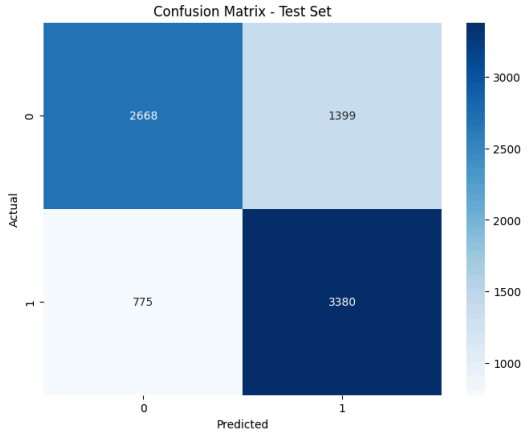

We summarize the significant relationships and consider the possible reasons for the relationships we saw.

| **Attribute** | **Coefficient** | **P-value** | **Possible Explanation** |
|---|---|---|---|
| Danceability | 3.3016 | < 0.001 | Higher danceability significantly increases odds of a hit; suggests popularity of danceable rhythms. |
| Acousticness | -1.401 | < 0.001 | More acoustic songs are less likely to be hits, suggesting a preference for electronic/synthesized music. |

| Attribute | Coefficient | P-value | Possible Explanation |
|---|---|---|---|
| Energy | -1.9241 | < 0.001 | Higher energy levels are associated with lower odds of a song being a hit; may indicate that excessively energetic songs are less appealing. |
| Speechiness | -3.2196 | < 0.001 | Higher speechiness drastically reduces odds of a hit, indicating preference for musical elements over spoken words. |
| Instrumentalness | -3.3709 | < 0.001 | Higher instrumentalness significantly decreases odds of a song being a hit. |

We also trained a model using a train-test split of .8/.2 and derived results based on our model. Our test data showed an accuracy of 73.6%. This was exciting to us, as we could predict whether these songs were hits or flopped without actually knowing the artist or genre of the song.

**[Figure Seven] Confusion Matrix of Logistic Regression as a Heat Map**




The confusion matrix shows that the model predicted true positives very well, given that our model's false negative rate was only 775/(775+3380) = 0.187 (18.7%). Our false positive rate was 1399/(2668+1399) = 0.344 (34.4%). We likely would have a more successful model given more song-specific predictors, and we could also consider other models that would consider other considerations like social media. Still, a 73.6% accuracy rate was very interesting and leads us to believe that song attributes could be used in an analysis to predict whether a new song would succeed.

**Conclusion:** Our motivation was to analyze the predictive power of Spotify attributes. To do so, we utilized linear regression analysis, checked regression assumptions, and created a fairly successful (74%) logistic regression analysis to classify a song as a hit or flop. Our findings offer valuable implications for music producers, record label executives, and enthusiasts alike, aiming to enhance music promotion strategies, discover new talent, and understand evolving music consumption trends.

**Appendix:**

**Source:** [https://developer.spotify.com/documentation/web-api/reference/get-audio-features](https://developer.spotify.com/documentation/web-api/reference/get-audio-features)

[1] Musical Attributes:

- Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- Danceability: 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. A value of 0.0 is least danceable and 1.0 is most danceable.
- Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- Instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).