A Bayesian Hierarchical Model For Predicting Song Popularity*

James Hubbs

May 5, 2022

1 Introduction

I consider the question of predicting song popularity. What is it that makes certain songs popular? There are seemingly infinitely many possible factors. Of primary interest here, however, is whether or not the particular *audio* qualities of a song are predictive of its popularity. Do things like tempo, rhythm, timbre, and other qualities of the sound help drive popularity?

One might initially assume this to be obviously true. Popular songs tend to leverage a predictable set of tools—most popular songs are in common time and use similar harmonic structures, for example. So rather than analyzing audio features and popularity across all styles of music, I instead focus entirely on songs within the popular framework. Among these songs, can popularity be accurately predicted using features of the audio?

Although pop songs often share some similarities, there remains lots of variation to analyze. For example, Billboard's Year-End Hot 100 Songs chart for 2021[1] features "Leave The Door Open" by Silk Sonic, a soulful callback to late 70s R&B, and "Levitating" by Dua Lipa, a hypermodern electronic dance track. Further, I consider 51 years of popular music beginning in 1970 and ending in 2021—of course, there are large stylistic differences across time.

In this analysis, I show that among select pop songs from 1970-2021, audio features are at best weakly predictive of song popularity.

*Data and code are available on GitHub

[1] https://www.billboard.com/charts/year-end/hot-100-songs/

2 Data Description and Exploration

I consider data from Spotify's Web API. Spotify is one of the largest music streaming services in the world. They're known for their data-forward approach toward streaming—the service makes extensive use of machine learning-based recommendation algorithms, which of course require considerable amounts of data. Some of this data is made available publicly through their web API.[2]

The data were sampled in March 2022. For each year between 1970 and 2021, the Spotify-generated playlists for top hits within a given year were used[3]. Using the top hits playlists ensures that our scope is narrowed to songs within the popular framework. It's also a convenient way to acquire a sample of music from each year in consideration. Each playlist consists of approximately 100 songs, so across 51 years we have a total sample size of about $5,100$.

2.1 Audio Features and Popularity

The following table provides summarized definitions of the primary predictors in consideration. The response variable is popularity, which Spotify describes somewhat vaguely as "The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are." The specifics of this algorithm are not described in their documentation. An important implication here is that we are analyzing *current* popularity. That is, we are not analyzing historical popularity of songs, but rather their popularity at the time of sampling in March 2022.

As with popularity, the below predictors are not defined by Spotify in great detail:

[2]Documentation is available at https://developer.spotify.com/documentation/web-api/
[3]For example, "Top Hits of 2000": https://open.spotify.com/playlist/37i9dQZF1DWUZv12GM5cFk

Table 1: Definitions of audio features

Variable	Spotify Definition (Summarized)
acousticness	A confidence measure from 0.0 to 1.0 of whether 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.
duration_ms	The duration of the track in milliseconds.
instrumentalness	Predicts whether a track contains no vocals.
liveness	Detects the presence of an audience in the recording.
loudness	The overall loudness of a track in decibels (dB).
speechiness	Speechiness detects the presence of spoken words in a track.
tempo	The overall estimated tempo of a track in beats per minute (BPM).
valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

Despite a lack of detail surrounding how these features are constructed by Spotify, it is generally easy to understand them intuitively. For example, Billie Eilish's "Your Power," a soft acoustic folk ballad that essentially uses only voice and acoustic guitar, is one of the highest scoring "acousticness" songs in the entire dataset.

As another example, consider two wildly popular songs from 2021: "All Too Well (10 Minute Version) (Taylor's Version)" by Taylor Swift and "Butter" by BTS. "All Too Well" is a lyrically-driven power ballad that scenically ruminates on the tribulations of a past relationship. "Butter" is an infinitely catchy, synth-driven dance pop song.

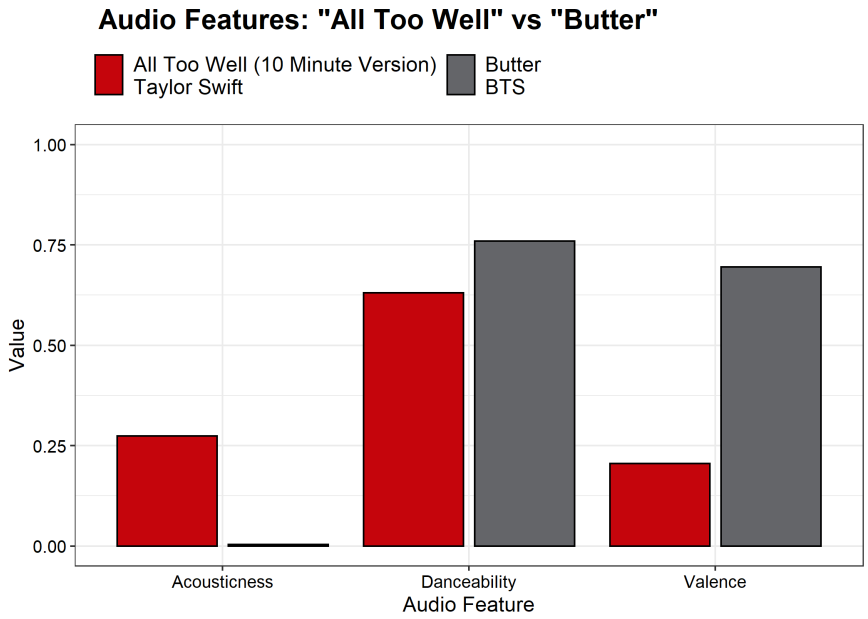


Figure 1: Audio feature comparison between Taylor Swift's "All Too Well" and BTS's "Butter". The upbeat, buoyant vibe of "Butter" leads to a much higher valence than "All Too Well". Similarly, the complete lack of acoustic instrumentation in "Butter" results in a near-zero acousticness score. Although we might expect "Butter" to score radically higher on danceability than "All Too Well", danceability is in fact comparable between songs (though "Butter" does score slightly higher). Both have a steady, consistent beat and use little rhythmic fluctuation.

2.2 Hierarchical Nature of the Data

Songs within a certain period of time are likely to be correlated. Popular songs that were released in the 1970s, for instance, often share some characteristics. The same is true for other time periods. Similarly, since we are measuring present popularity of songs, rather than historic popularity, it may be the case that a listener's relationship with audio features could vary by time period. A listener may enjoy "danceable" songs from the 1980s, but generally dislike more modern songs that score high in danceability (the converse, of course, may also be true for other listeners). It's for these reasons that I view the data through a hierarchical lens, using release decade as the grouping variable. In principle, a more narrow grouping (e.g., by year) could be employed. Since music is often conceptualized in a decade-by-decade fashion, decade is a natural grouping for the data.

2.3 Exploration

There is much to be learned about this dataset through visualization.

The below figure summarizes popularity by decade. Of course, popularity tends to be greater in more recent decades. This is simply reflective of the fact that we're measuring present popularity. In turn, it may also reflect Spotify's user demographics, since over 50% of Spotify users are under the age of 34[4], and younger users will likely prefer newer songs. We also see that across all decades, the median popularity is greater than 50, and this is a product of having sampled generally popular songs, since songs were selected via the "top hits" playlists. Finally, we observe there are many low-valued outliers within each decade—these observations will later prove particularly difficult to predict.

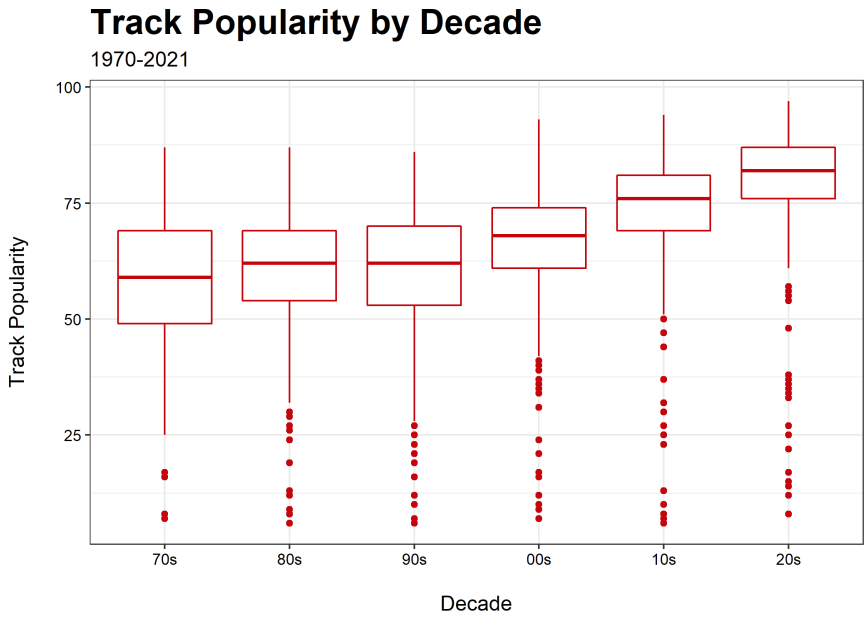


Figure 2: Track popularity over time. Since popularity is measuring present popularity, songs from more recent time periods have greater popularity. Note that "20's" includes only 2020 and 2021.

There are other interesting temporal features of this dataset. In particular, there are prominent trends among some of the predictors:

[4] https://www.businessofapps.com/data/spotify-statistics/

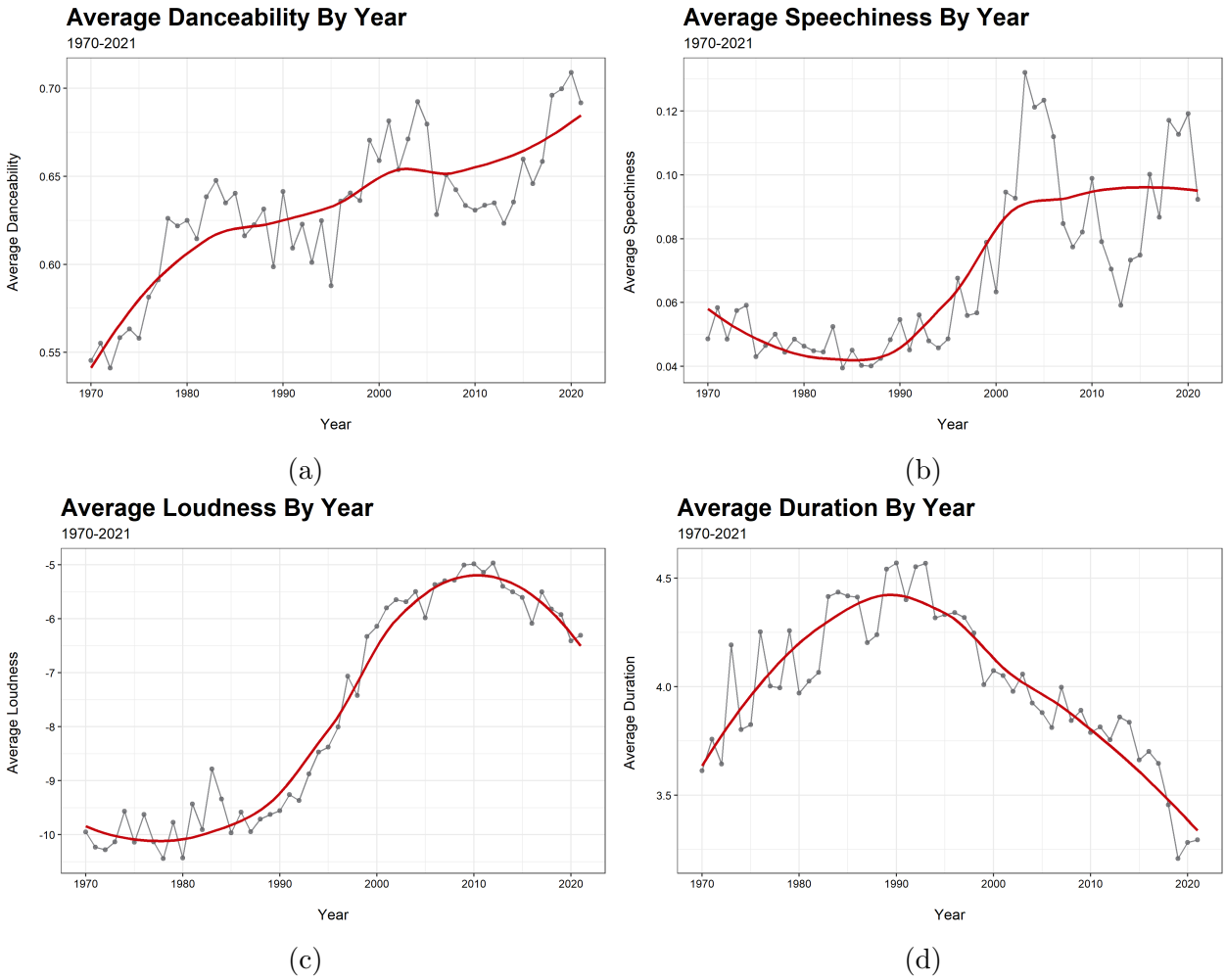


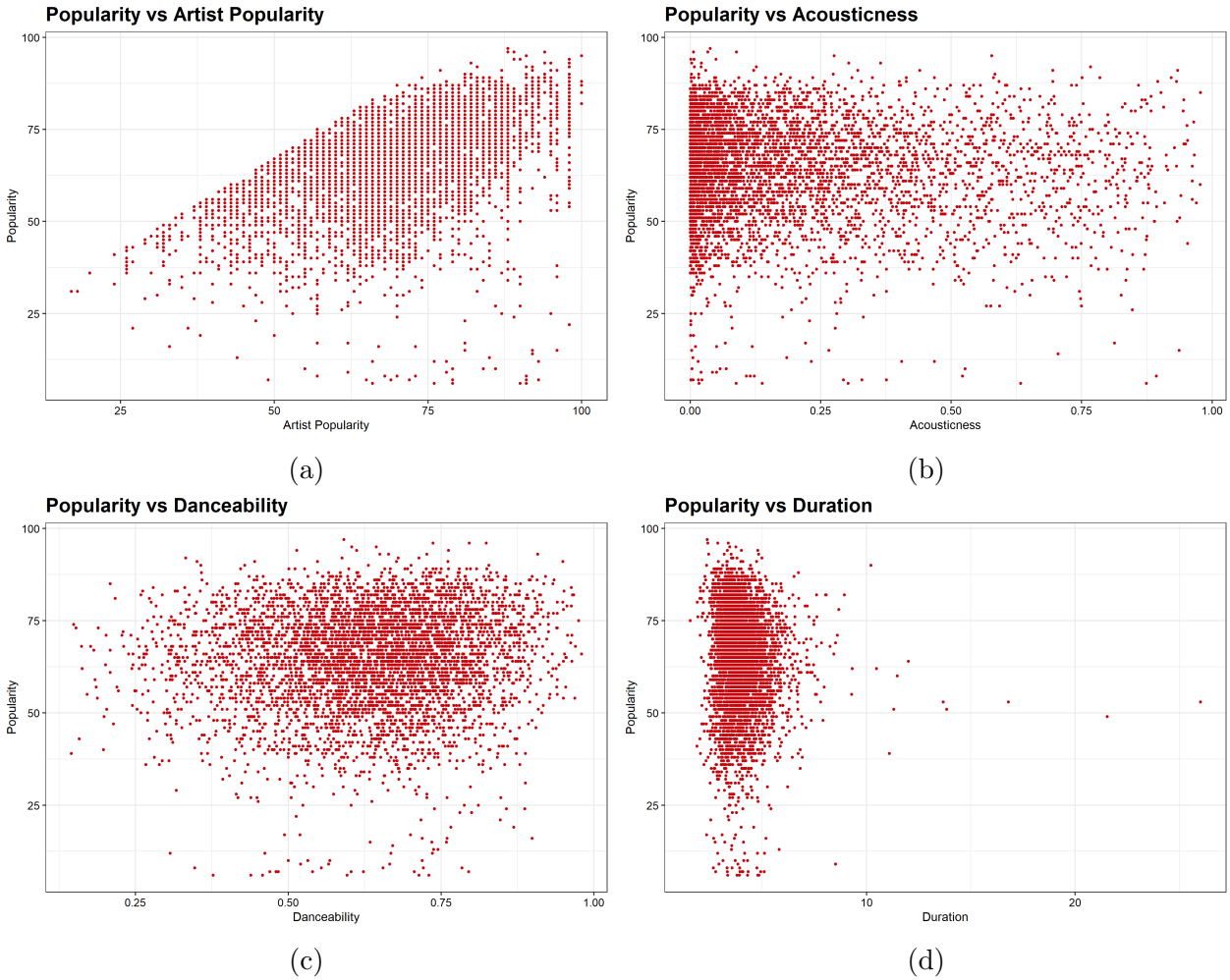
Figure 3: Trends in danceability, speechiness, loudness, and duration metrics by year. Popular songs were, on average, becoming more danceable over time. Similarly, they also were getting louder from about 1990 up until around 2010. Speechiness undergoes a gradual increase starting around 1990, likely corresponding to the global rise in popularity of rap and hip-hop music. These figures also highlight the importance of considering temporal correlation within the data—songs within particularly periods of time are, on average, similar in their audio features.

In addition to the audio features described in Table 1, artist popularity is also considered. As one would expect, this predictor simply measures the overall popularity of a song's artist. Artist popularity is used, in some sense, as a control variable. We ultimately aim to understand the predictive relationship between song popularity and audio features, but it seems likely that some artists could reach such a high level of fame and recognition that their songs become popular independently from their audio features. We will therefor allow artist

popularity to enter the model, serving as a type of control.

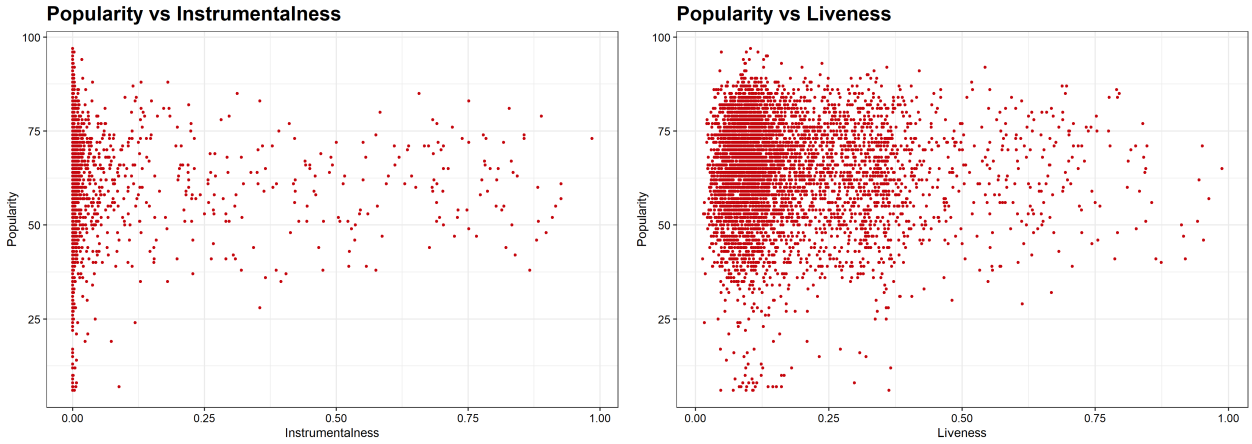
Further, the "energy" feature was omitted from consideration due to reasonably high collinearity with several other features. In particular, its correlation coefficient with loudness, acousticness, and valence was 0.70, -0.60, and 0.40, respectively.

Visual inspection suggests that the relationships between song popularity and each of the 10 predictors in consideration are surprisingly weak:



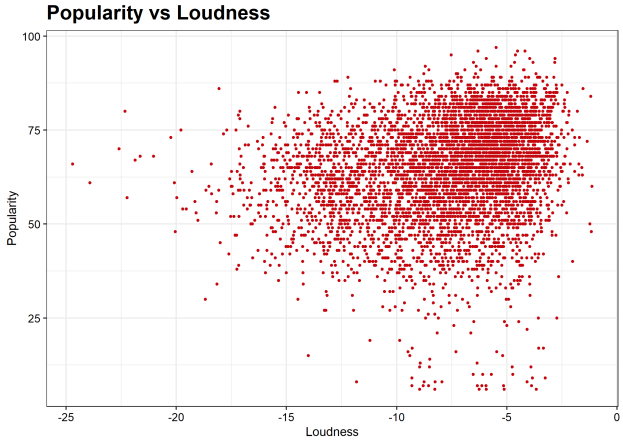
(a) (b)

(c) (d)

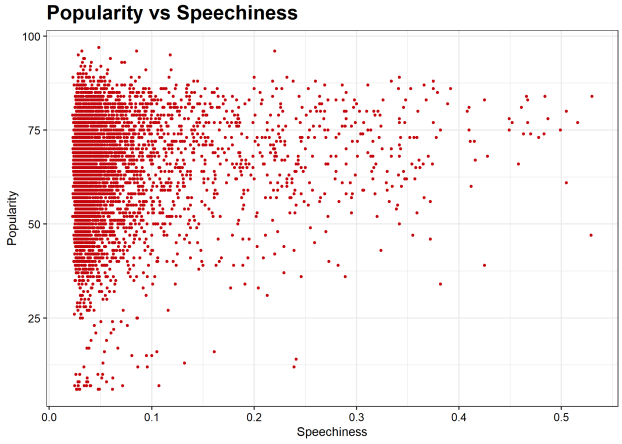


(e)

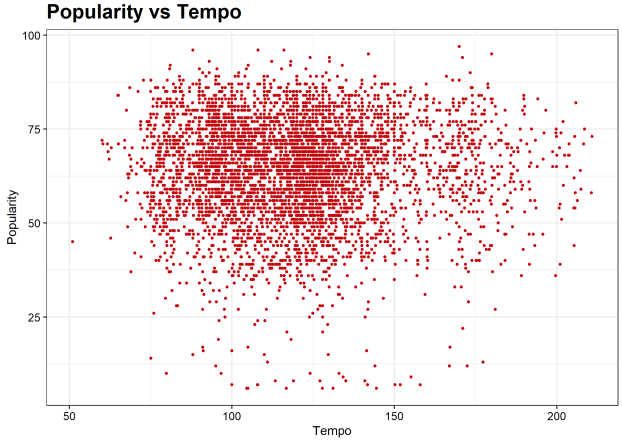
(f)



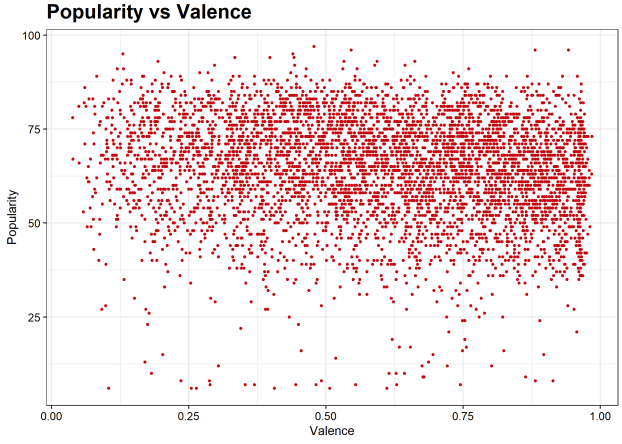
(g)



(h)



(i)



(j)

Figure 4: Response vs predictor scatter plots

Based on these figures, the task of predicting song popularity seems, at first glance, formidable.

3 Modeling

I estimate a standard linear hierarchical model with $P = 10$ predictors and $T = 6$ decades by which the data are grouped. The model was fit using Stan.

Let Y_{it} denote the popularity of song i in decade t.

Then assume

$$
\begin{aligned}
Y_{it} \mid \alpha_t\ \boldsymbol{\beta_t}\ \sigma^2 &\sim N(\mu_{it}, \sigma^2) \text{ where } \mu_{it} = \alpha_t + x_i^T \boldsymbol{\beta_t} \\
\alpha_t \mid \mu_\alpha\ \tau_\alpha^2 &\sim N(\mu_\alpha, \tau_\alpha^2) \\
\beta_{jt} \mid \mu_{\beta_j}\ \tau_\beta^2 &\sim N(\mu_{\beta_j}, \tau_\beta^2) \\
\mu_\alpha &\sim N(50, 5) \\
\mu_{\beta_j} &\sim N(\bar{\mu}_{\beta_j}, 1) \\
\tau_\beta^2 &\sim \text{Inv.Gamma}(1.5, 1) \\
\tau_\alpha^2 &\sim \text{Inv.Gamma}(1.5, 1) \\
\sigma^2 &\sim \text{Inv.Gamma}(1.5, 0.3)
\end{aligned}
$$

The hierarchical structure defined above allows for posterior intercepts and slopes to vary by decade. This is important since each decade has a variable base-level popularity, and we assume the effect of each predictor on popularity may vary by decade. $\boldsymbol{\beta_t}$ is the length-P vector of predictor coefficients for decade t. Within the second layer, β_{jt} is therefor the j-th coefficient in decade t and α_t is the intercept within decade t. The third layer defines priors and their hyperparameters, which were set using a combination of prior predictive checks and a priori beliefs. In particular, the hyperparameters $\bar{\mu}_{\beta_j}$ were set to reflect a priori beliefs regarding the sign of each coefficient. It was set to -0.10 for instrumentalness, acousticness, liveness, and duration, reflecting the belief that the posterior mean for the coefficients associated with these predictors should be negative. For all other predictors, it was set to 0.10 (since, for instance, we expect the relationship between popularity and danceability to be positive). The relatively small absolute value of 0.10 was chosen so as to not be overly strong. Further, note that the β_{jt} coefficients share a variance parameter. That

is, the variability of the effect of one predictor is assumed to be the same as the variability of the effect of some other predictor.

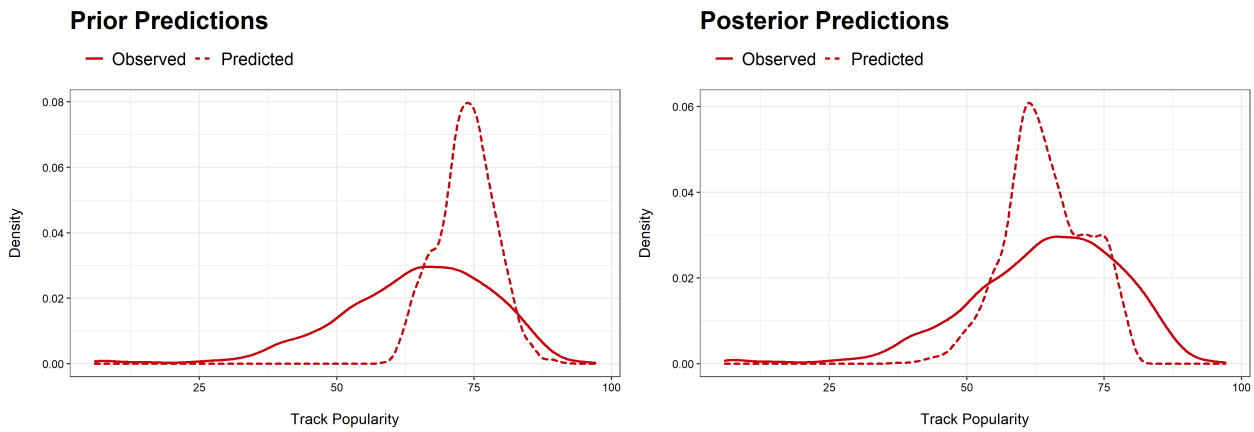
Some numerical summaries are provided below:

Table 2: Numerical Summary Statistics

Quantity	Value
RMSE	11.59
MAPE	19.83%
R^2	0.295
σ^2	11.62
Mean ESS	3918
Mean $\hat{R}$	1.00003

The model is clearly not predicting particularly well; on average, predictions miss by about 20% (though much of this is due to the model's difficulty when predicting low popularity songs—for songs with observed popularity greater than 0.30, the RMSE falls to 9.85 and the MAPE falls to 13.3%). It does, at least, explain approximately 30% of variability. Convergence diagnostics (ESS and $\hat{R}$) were acceptable for all parameters.

To get a better idea of the model's performance, we can see how the posterior predictive distribution updated the prior predictive distribution.



(a) (b)

Figure 5: Prior vs Posterior prediction means. The prior distribution placed almost all density near 75. The posterior pulls the prior leftward to better reflect the observed density, though very little density is placed near very low or very high values of popularity.

The negligible density around small values of popularity for the posterior predictive distribution is reflected in the residual vs fitted figure below, where we see that the model systemically over-predicts songs with low popularity:

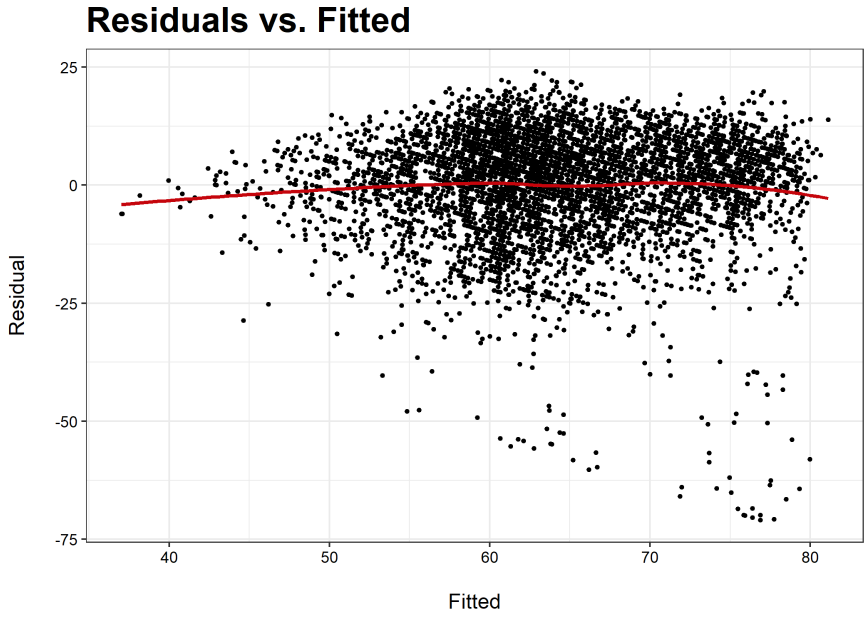


Figure 6: Residuals vs Fitted. Fitted values correspond to posterior means for each posterior predictive distribution.

Despite the indication of weak predictive performance, it's still of interest to assess which predictors are most influential.

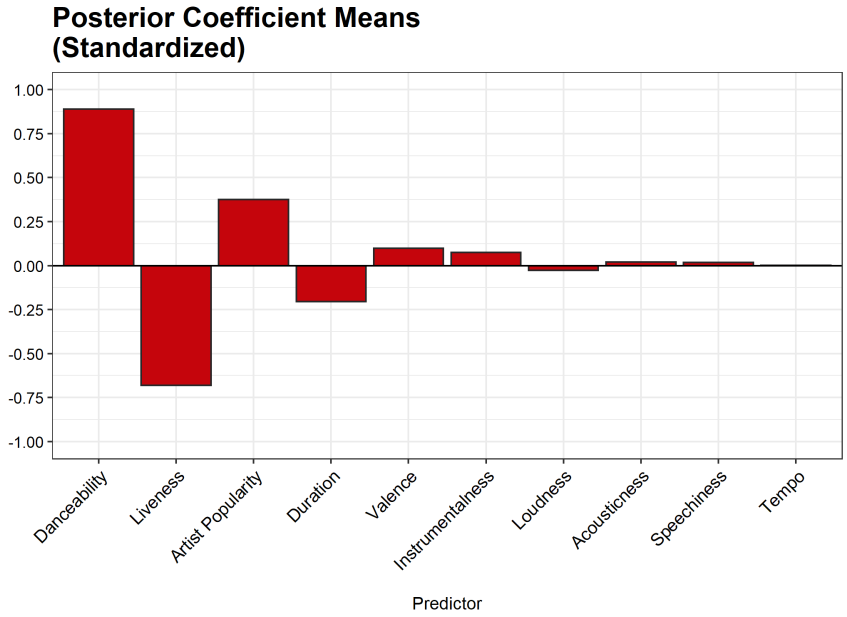
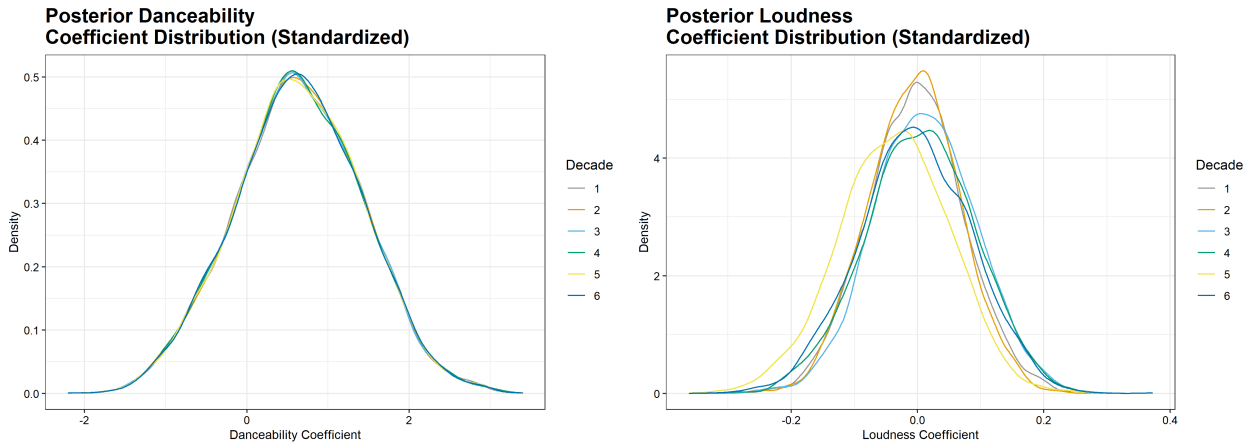


Figure 7: Standardized posterior coefficient means. The data suggest that danceability has the largest effect on popularity.

We see that danceability appears most predictive of popularity—popular songs will typically leverage a pronounced, predictable beat with minimal rhythmic fluctuation.

A hierarchical model with variable intercepts and slopes was used under the assumption that the effect of each predictor on popularity may vary by decade. To verify this assumption, we can plot the posterior distribution for each predictor's coefficient. We find that for some predictors the effect does indeed vary by decade; for others, however, it does not.



(a) (b)

Figure 8: Posterior distributions of danceability and loudness coefficients. The effect of danceability does not vary much at all by decade, whereas the effect of loudness does vary by decade (albeit slightly).

4 Discussion

This analysis has shown that Spotify's audio features are weakly predictive of song popularity. In some sense, it's not particularly surprising that song popularity resists predictive simplicity—a mere 10 predictors is generally insufficient for predicting popularity with confidence. That being said, the audio features that were considered do explain some variability in popularity, with the model suggesting danceability is most predictive.

One key limitation of this analysis should be mentioned. The audio features in consideration are all aggregate measures across an entire song. For instance, loudness measures the overall loudness of the song from beginning to end. But as a hypothetical example, it could be the case that what matters most is not how loud the song is in an aggregate sense, but the contrast in loudness between, say, a song's verse and chorus. These audio features could potentially have greater predictive power if they were available not simply as aggregate summaries, but instead as time-based measures that fluctuate throughout the duration of a song—this is, after all, more consistent with how we naturally perceive music.

Further, there are other (unconsidered) features that may be important when predicting popularity:

- Marketing: Songs that are marketed heavily will, of course, likely rise in popularity.

- Lyrical sentiment: The valence feature considers musical sentiment, but lyrical sentiment may also be predictive of popularity.
- The TikTok effect: Songs that circulate on the social media platform TikTok tend to grow in popularity.[5] How can we measure a song's ability to resonate with TikTok's primary audience?
- Seasonality: Songs that release during (or immediately before) summer may see increased popularity.

These likely-important predictors, however, are clearly outside the scope set out by the primary question of this analysis: Are a song's *audio* features predictive of its popularity? We've found evidence that, in fact, audio features as defined by Spotify are not particularly predictive.

[5] https://www.businessinsider.com/how-tiktok-is-changing-the-music-industry-marketing-discovery-2021-7