A Bayesian Hierarchical Model For Predicting Song Popularity[*]

James Hubbs

April 26, 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 feature predictors

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.

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.

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. Though music is often conceptualized in a decade-by-decade fashion, so decade is a natural choice for grouping 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.

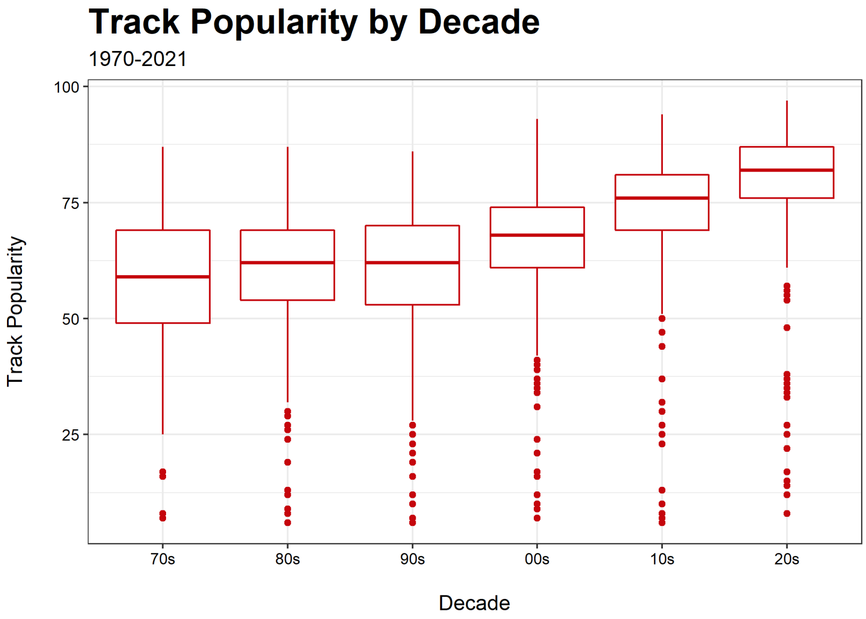
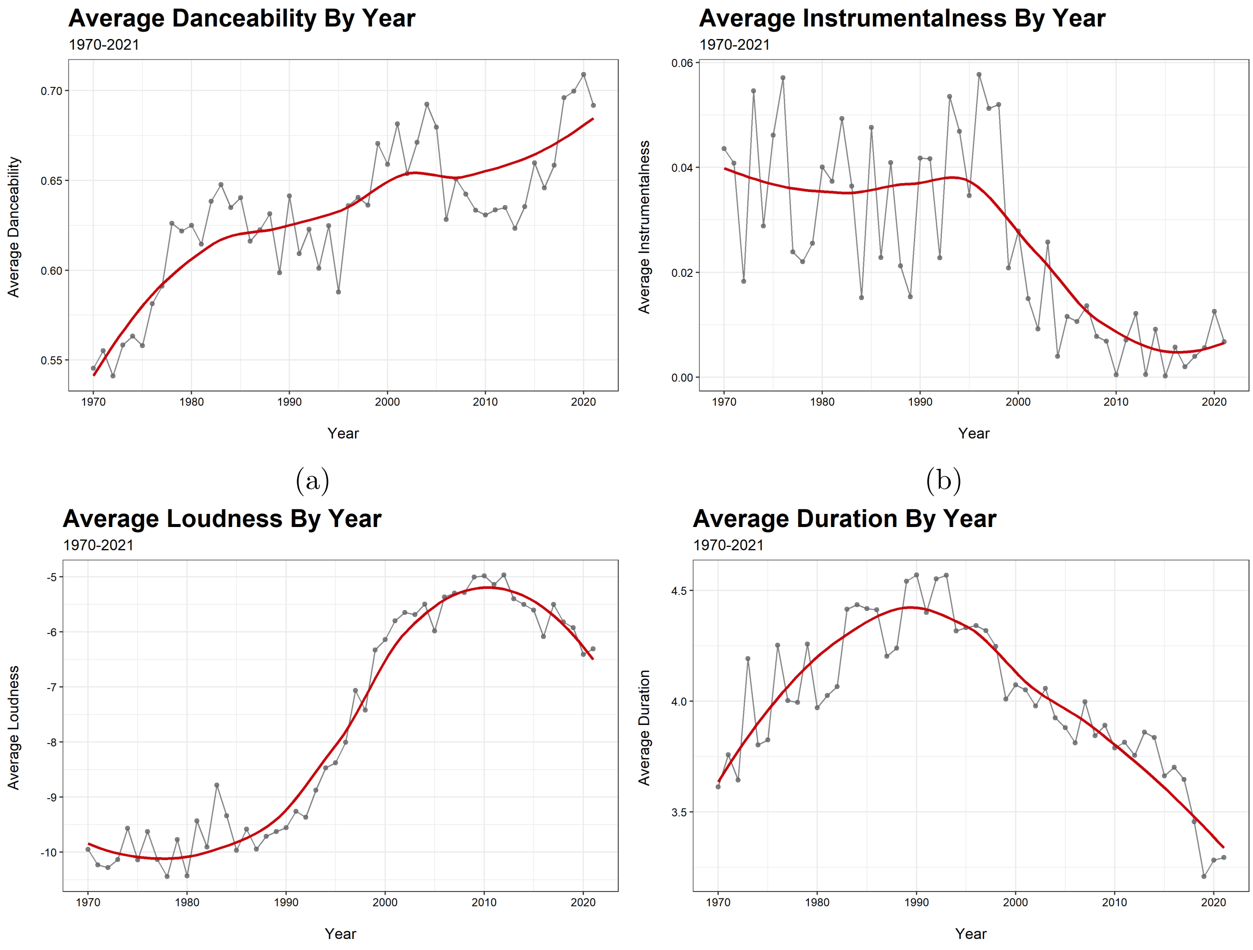


Figure 1: Track popularity over time. Since popularity is measuring present popularity, songs from more recent time periods have greater popularity. Note that "20's" includes only 2020 and 2021.

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

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



(a)

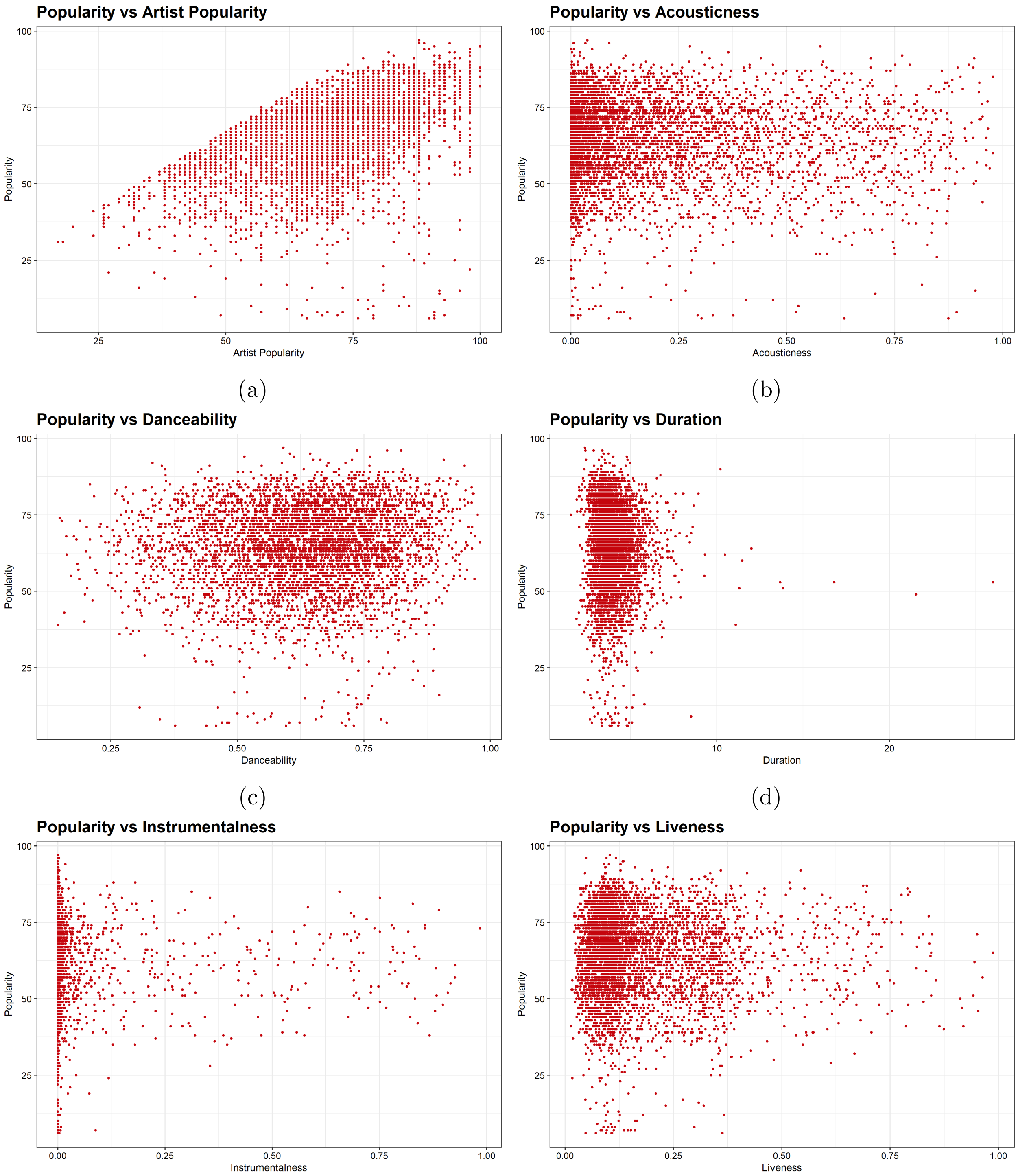
(b)

(c)

(d)

Figure 2: Trends in danceability, instrumnetalness, loudness, and duration metrics by year. Popular songs were, on average, getting louder and louder from about 1990 up until around 2010. We also observe a sharp decline in average instrumentalness around 1995. 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.

The primary interest is in the relationship between the predictors and popularity. I consider 10 predictors. Visual inspection suggests, surprisingly, that these relationships are weak:


Popularity vs Artist Popularity
Popularity
100
75
50
25
25
50
75
100
Artist Popularity
Popularity vs Acousticness
Popularity
100
75
50
25
0.00
0.25
0.50
0.75
1.00
Acousticness
Popularity vs Danceability
Popularity
100
75
50
25
0.25
0.50
0.75
1.00
Danceability
Popularity vs Duration
Popularity
100
75
50
25
10
20
Duration
Popularity vs Instrumentalness
Popularity
100
75
50
25
0.00
0.25
0.50
0.75
1.00
Instrumentalness
Popularity vs Liveness
Popularity
100
75
50
25
0.00
0.25
0.50
0.75
1.00
Liveness

(a)

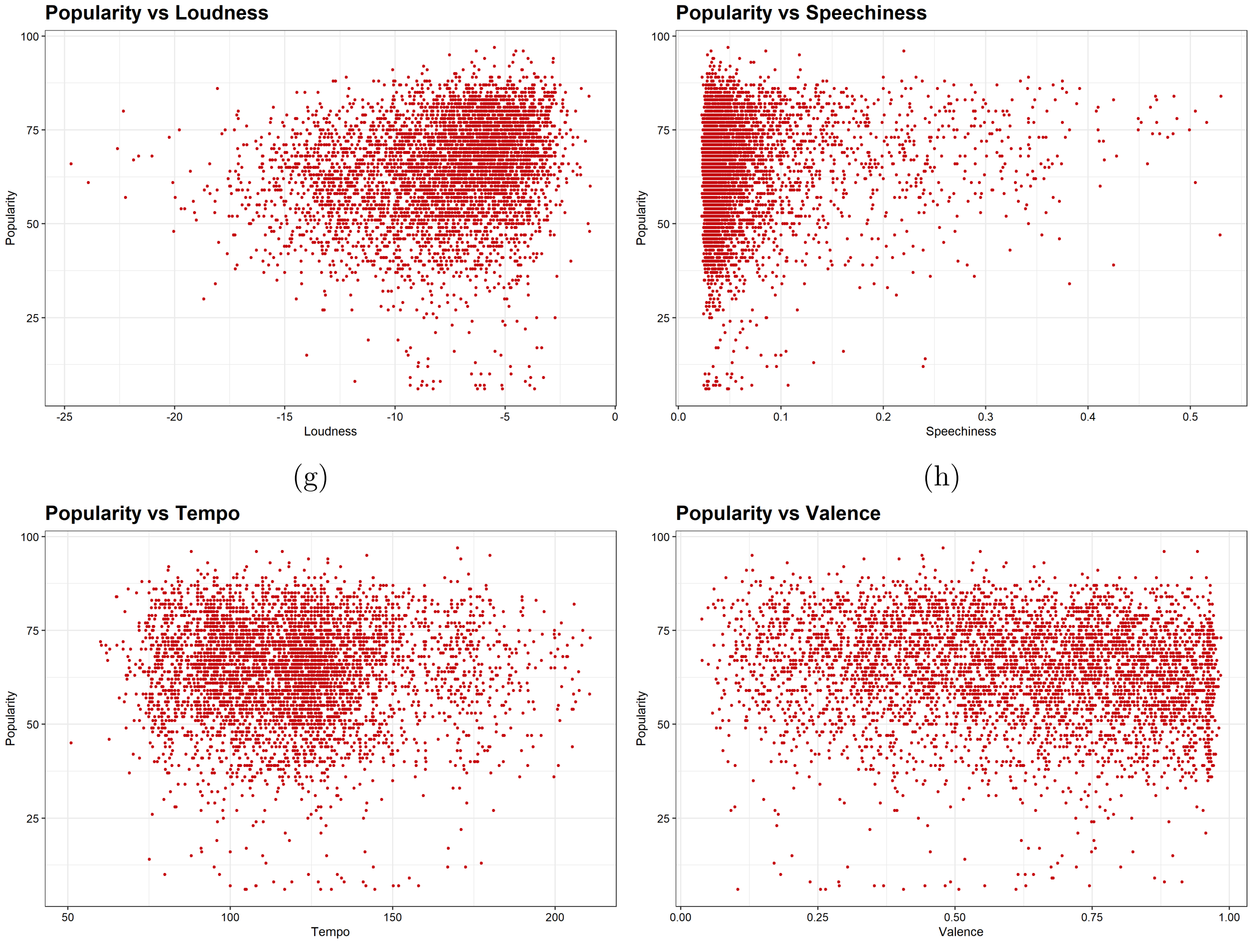
(b)

(c)

(d)

(e)

(f)



(g) (h)

(i) (j)

Figure 3: Response vs Predictors

TODO: brief commentary.

Also need to explain why artist popularity is considered a predictor, since it isn't an audio feature. The idea is that we basically want to control for artist popularity. A super popular artist could put out literally anything and it will be popular regardless of audio features.

3 Modeling

I estimate a standard linear hierarchical model with $J = 10$ predictors and $T = 6$ decades by which the data are grouped.

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

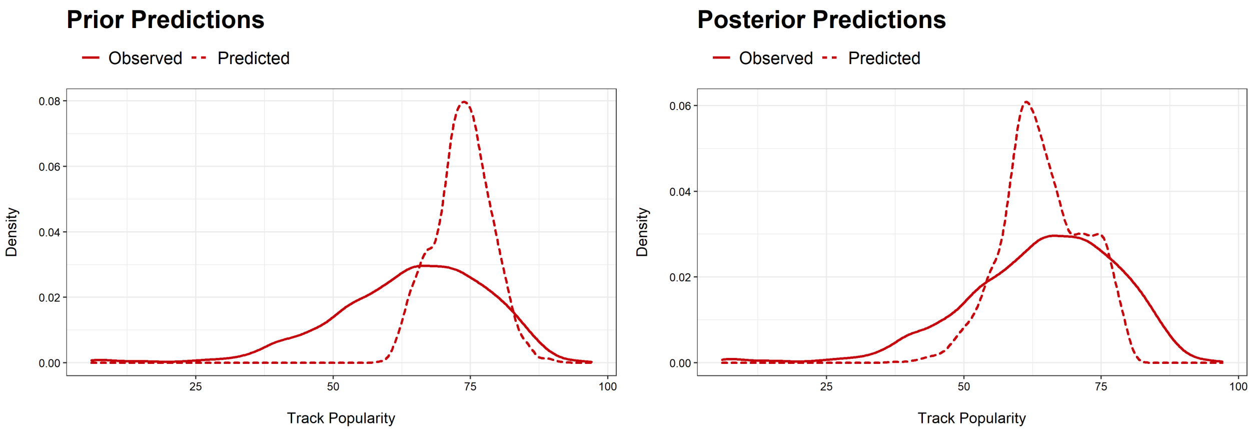
$$
\begin{aligned}
Y_{it} \mid \alpha_t \ \beta_{jt} \ \sigma^2 &\sim N(\mu_{it}, \sigma^2) \text{ where } \mu_{it} = \alpha_t + x_i^T \beta_{jt} \\
\alpha_t \mid \mu_\alpha \ \tau_\alpha^2 &\sim N(\mu_\alpha, \tau_\alpha^2) \\
\beta_{jt} \mid \mu_j \ \tau_\beta &\sim N(\mu_{\beta_j}, \tau_\beta^2) \\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

The hyperparameter $\bar{\mu}_{\beta_j}$ 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 relatiomship between popularity and danceability to be positive). The relatively small absolute value of 0.10 was chosen so as to not be overly strong.

TODO: More commentary on the model (e.g., pooled variance for coefficients–variability of the effect of one predictor is essentially the same as the variability of the effect of some other predictor. Means however vary by predictor, since of course the effect of one could differ greater from the effect of another.)

Also TODO: Add some kind of numerical summary of the model. (posterior means of coefficients, MAPE, RMSE, etc.)

Here are some more figures that I intend to include (and comment on). This is all *very* incomplete:



(a)

(b)

Figure 4: Posterior looks pretty bad but it does update the prior reasonably

Why's there so much space here ahhh

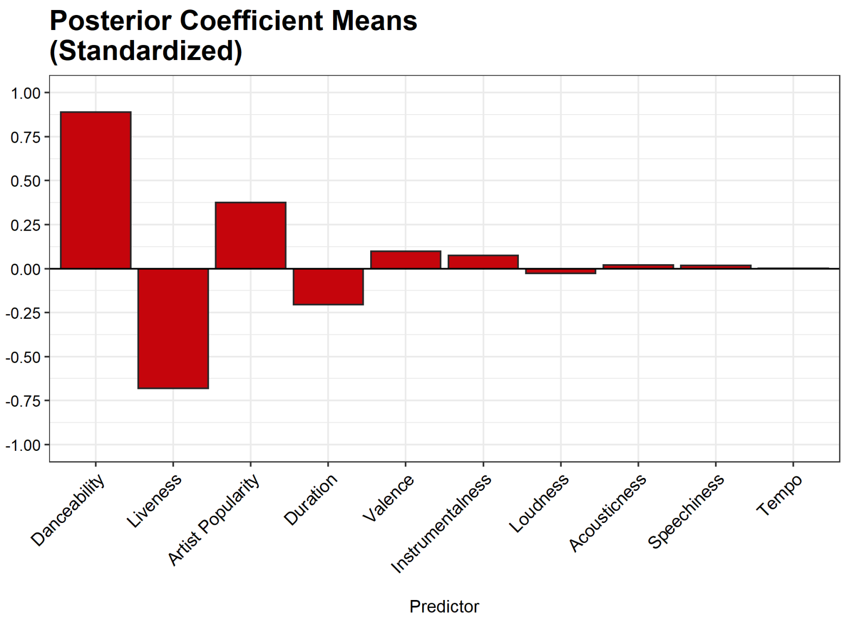


Figure 5: Danceability has largest effect on popularity.

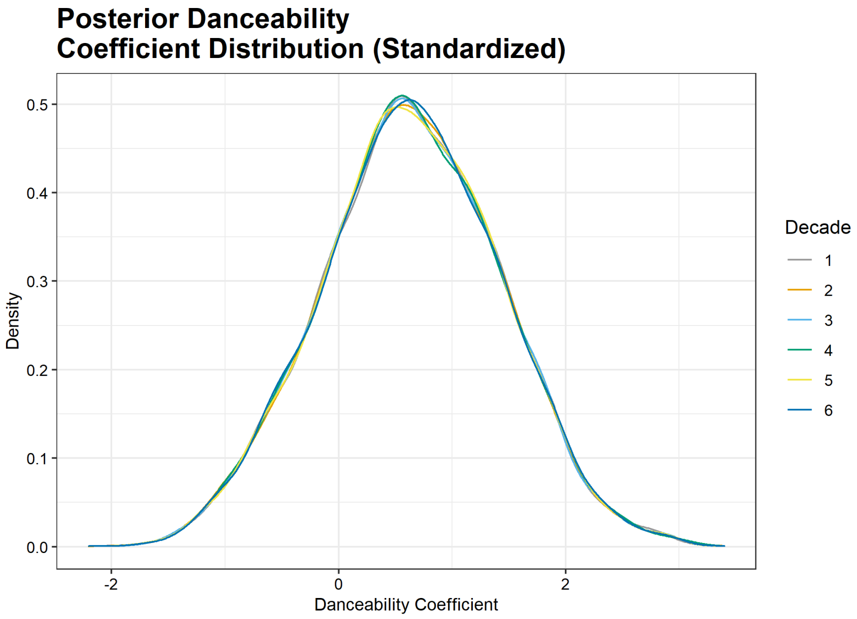


Figure 6: Danceability effect not variable across decades (not true for all predictors). We really like dancey songs regardless of decade!

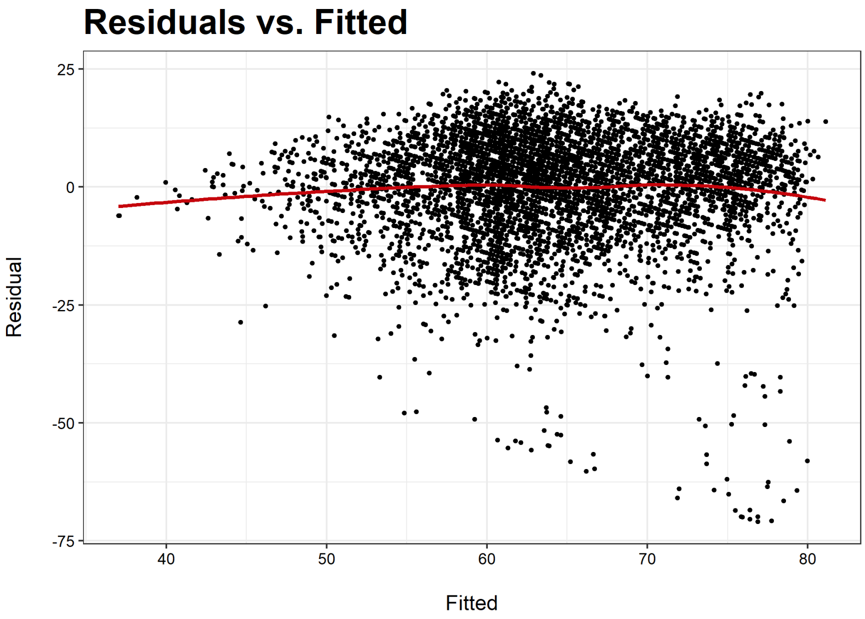


Figure 7: Probably don't care too much about the residuals, but this shows that there is a systemic problem of predicting low-popularity songs. Have to sympathize with the model a bit, though. Why does "Money For Nothing", a song with hundreds of millions of plays on Spotify, have a popularity score under 10? Also should note that fitted here refers to posterior mean predictions.

4 Discussion

TODO.

References