

Understanding What Makes a Song a Hit

Diane Lin, Stephanie You

Introduction and Data

Project Motivation

Our project stems from a shared passion for music and a curiosity about the factors influencing song popularity. In an era dominated by viral trends on platforms like TikTok, where snippets of songs can propel them to stardom, we were intrigued by the possibility of predicting song popularity using quantitative metrics. According to the SongTown music podcast, a part of a song being a ‘hit’ may be attributed to the lyrics, but the most important factor is the song’s ‘fit’ with the artist and the producer. This realization led us to consider not just the lyrical content and melody of songs but also how well they align with the artist’s style and the producer’s vision.

An article released by The Guardian talks about forgotten bands seeing the limelight again thanks to Gen Z. For example, Keane’s “Somewhere Only We Know” was a part of the band’s 2004 debut single. It wasn’t until last summer when member Tim Rice-Oxley noticed that “suddenly there were lots of much younger people” in the audience, due to TikTok. However, “Somewhere Only We Know” didn’t need the boost, as it was huge at the time of Keane’s debut.

Most TikTok users won’t explore an artist’s work beyond their viral hit, but if just a small percentage of tens of millions of new listeners become genuine fans, it can make a huge difference. Songs usually go viral when they become attached to a certain format or message, but there are also examples where songs become viral due to being attached to a timeless genre. According to Today, Icelandic musician Laufey released her debut album that garnered almost 700K followers and over 18M likes on Tiktok. Her discography, which has been likened to the soundtracks of Disney or a Studio Ghibli film, has encouraged younger generations to fall in love with jazz and classical music. She now boasts almost 3 million monthly listeners on Spotify.

GenZ’s obsession with nostalgia allows us to beg the question: what causes a song to go viral or experience bursts of virality? Are factors that make a song popular the same in the past as they are now?

Data

In our pursuit to understand the dynamics of song popularity, we turn to the Million Song Dataset as our guide. There were several curators for this dataset—the Million Song Dataset used a company called the Echo Nest to derive data points about one million contemporary songs. It was also a collaboration between the Echo Nest and LabROSA (laboratory working towards intelligent machine listening). It was collected in 2011. Though it is now 2024, old songs that were popular have resurfaced on TikTok, showing that data collected 13 years ago is still relevant. The link for it is here: <http://millionsongdataset.com/faq/>

Research Question

How effectively can the combination of artist hotness, artist familiarity, song duration, tempo, fade times, and genre predict a song's hotness? Further, can we find sufficient evidence that an artist's familiarity differs based on genre?

Relevant Variables

Song Hotness: Our response variable. Indicates the hotness/popularity of a song between 0 and 1, with 1 being the highest value. Continuous numeric variable.

Artist Hotness: Indicates how much 'buzz' the artist is getting when the song was downloaded, on a scale of 0 to 1, with 1 being the highest value. Continuous numeric variable.

Artist Familiarity: Indication of how well known the artist is, on a scale of 0 to 1, with 1 being the highest value. Continuous numeric variable.

Song Duration: Duration of a song, in seconds. Continuous numeric variable.

Tempo: Tempo in BPM of a song. Continuous numeric variable.

Start of Fade Out: Start time of the fade out, in seconds, at the end of a song. Continuous numeric variable.

End of Fade In: Time of the end of the fade in, at the beginning of the song. Continuous numeric variable.

Genre: Array string of genres the artist is associated with. See **data cleaning** for more. Categorical variable.

artist_familiarity	song_hotness	artist_hotness	duration	tempo	fade_out	fade_in	gen_genre
0.6510457	0.6045007	0.4017237	209.6061	129.738	198.699	0.066	pop
0.6364236	0.4051157	0.4479355	491.1277	119.826	486.034	0.000	rock
0.7071998	0.6841362	0.5134633	290.2983	150.062	285.605	0.145	rock
0.4354158	0.4952936	0.3062423	124.8649	137.522	115.740	0.000	rock
0.8093238	0.5939431	0.4880198	450.8469	130.586	426.104	0.467	rock
0.6610802	0.5101193	0.4427658	190.4844	102.428	177.029	0.258	rock

Cleaning Process

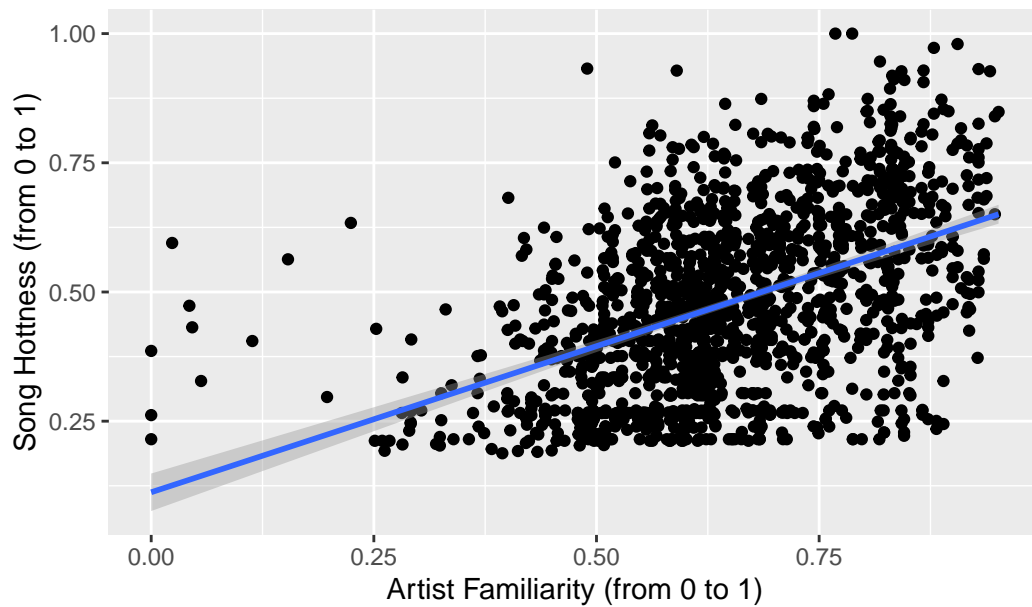
We narrowed our focus to four primary genres: rap, rock, pop, and country, capturing a broad spectrum of musical styles while minimizing complexity. We further simplified the genre landscape by generalizing subgenres (ex: Classical rock) into overarching categories (Rock), and creating a new variable, `gen_genre`, with those overarching categories. Then, we filtered for all `song.hottnesss` scores that were less than zero, because those made no sense according to the documentation for this (when checking the values that were less than 0, there were 6 rows, with a `song.hottnesss` of -1. We concluded that it was a placeholder for datapoints with a null value for song hotness).

EDA

We hypothesized two of the most important factors in determining song popularity was artist familiarity and genre, as people are more inclined to listen to the artists that are ‘cool’ at the time, and based on personal experience, 2011 was dominated by pop music. Thus, we created scatter plots to visualize the relationships.

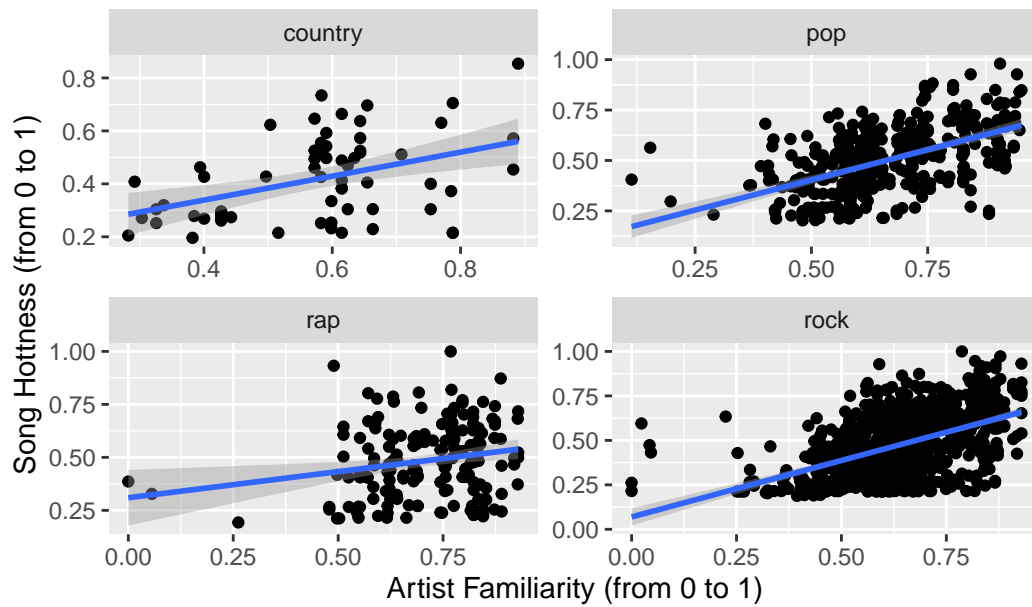
```
`geom_smooth()` using formula = 'y ~ x'
```

Artist familiarity does not necessarily imply song hottness

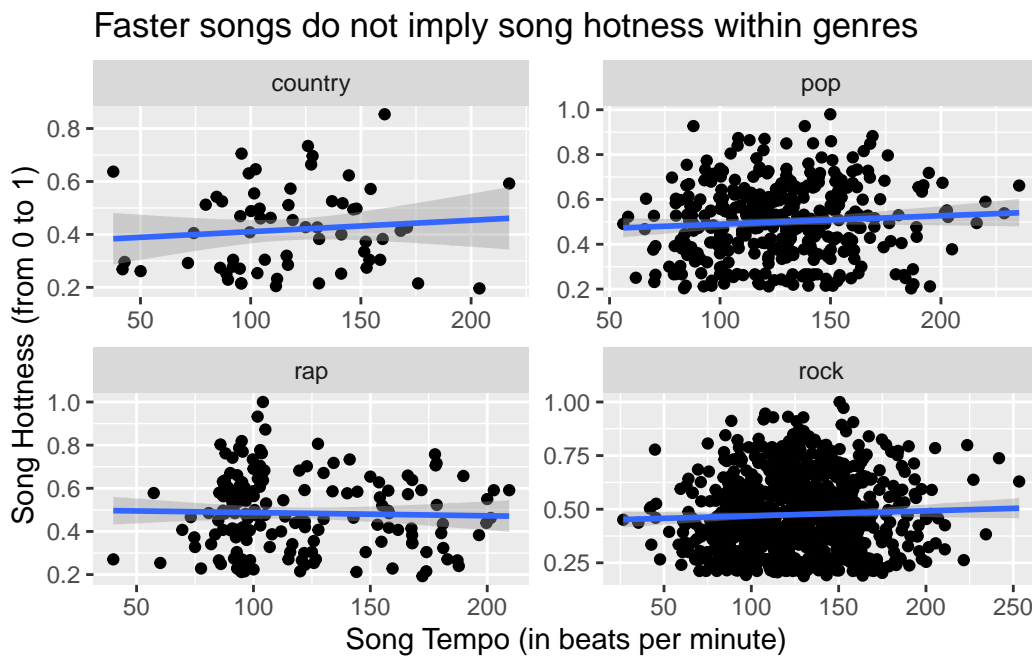
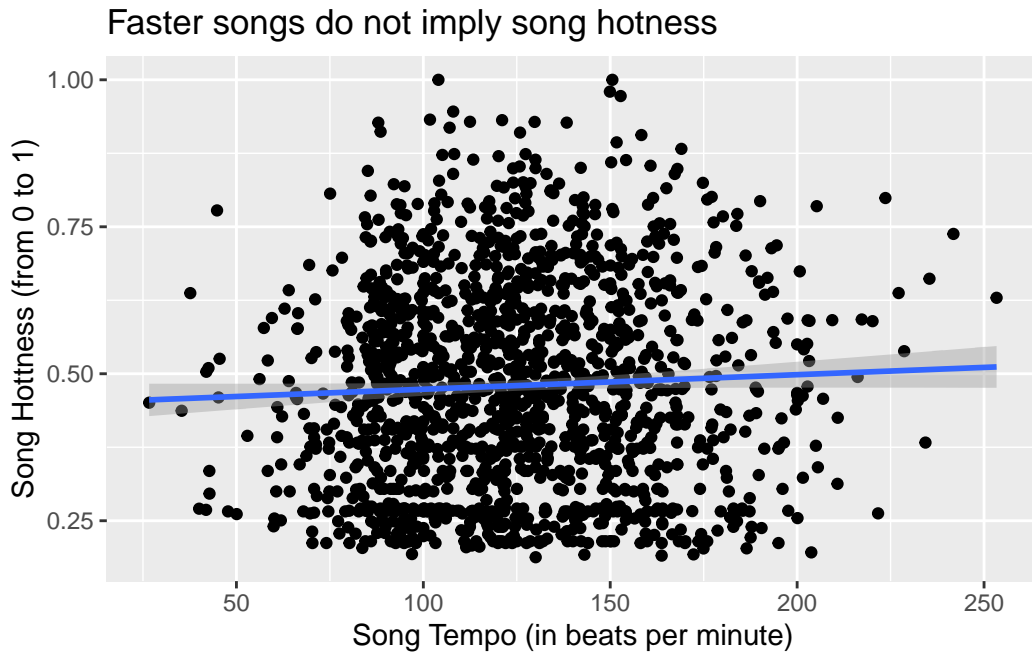


`geom_smooth()` using formula = 'y ~ x'

Song hottness may depend on genre



Because pop and rap songs have a reputation of being faster, we wanted to explore the relationship between song tempo, genre, and song hotness. We found that faster songs do not imply song hotness.



Methodology

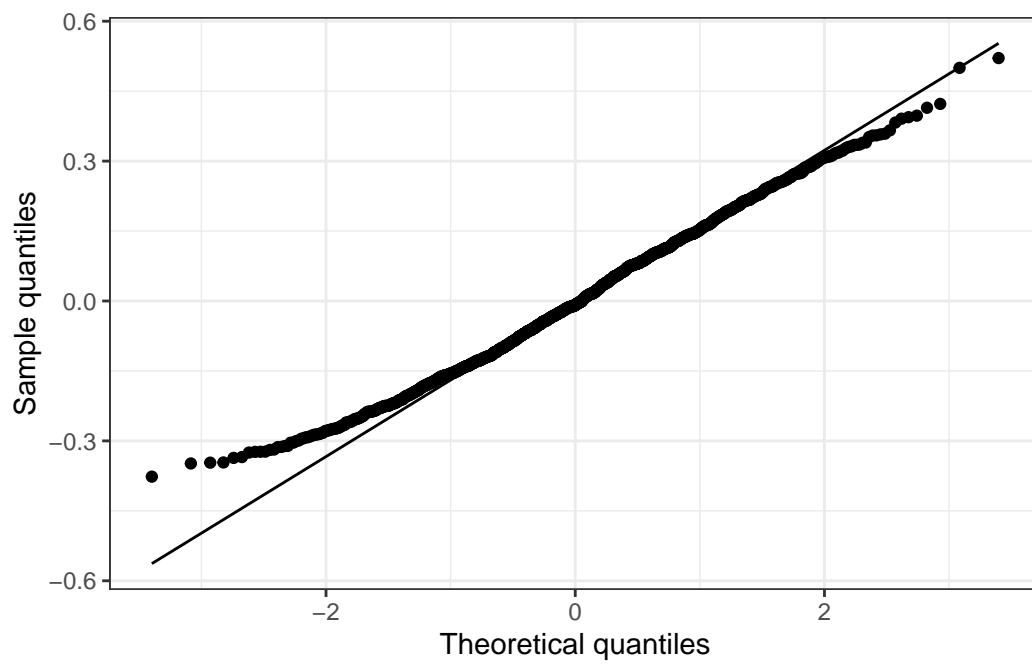
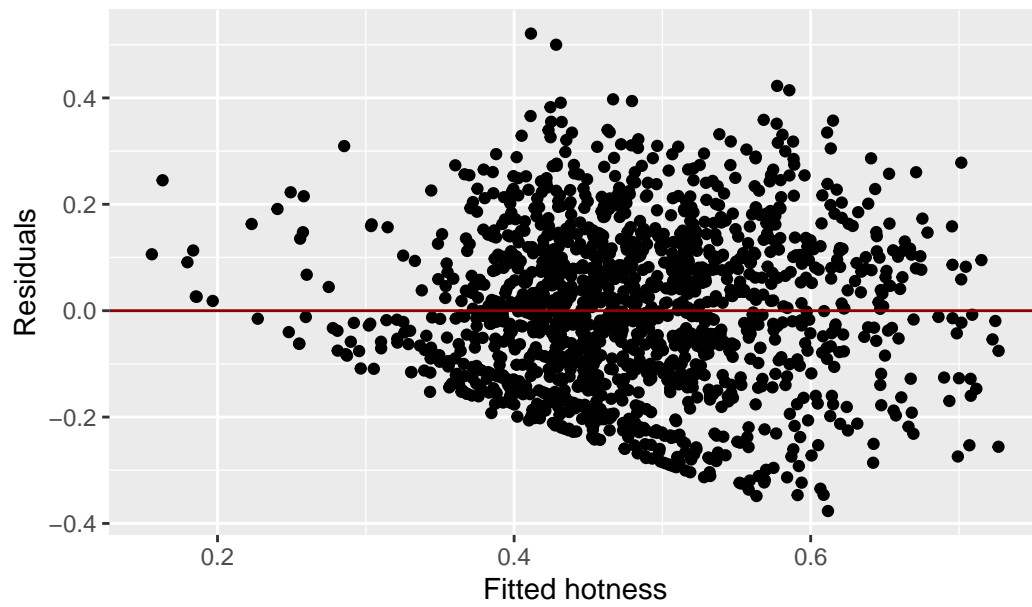
To get a comprehensive understanding of how well (or poorly) the predictors we had picked could predict song hotness, we decided to create more than one model. Since our response variable, song hotness, is a continuous numeric variable, we will be using a linear regression model.

First, we fit a regular linear model, using artist familiarity, artist hotness, song duration, genre, start of fade out, end of fade in, and song tempo as the predictors.

term	estimate	std.error	statistic	p.value
(Intercept)	0.0295918	0.0297706	0.9939914	0.3203916
artist_familiarity	0.3302057	0.0438741	7.5262112	0.0000000
artist_hotness	0.3349121	0.0483736	6.9234428	0.0000000
duration	0.0011337	0.0006368	1.7802836	0.0752373
as.factor(gen_genre)pop	0.0446009	0.0208737	2.1367068	0.0327876
as.factor(gen_genre)rap	0.0102870	0.0228824	0.4495582	0.6530957
as.factor(gen_genre)rock	0.0284675	0.0199278	1.4285280	0.1533537
fade_out	-0.0010607	0.0006506	-1.6302884	0.1032561
fade_in	0.0001040	0.0025412	0.0409441	0.9673460
tempo	0.0002053	0.0001187	1.7302405	0.0837985

Metric	Value
R-squared	0.2547013
Adjusted R-squared	0.2547013

No evidence of linearity



Model Assumptions

It is also important to discuss the assumptions of linear regression, like independence, linearity, normal distribution, and variance. We assume independence is satisfied, as knowing the information about one song does not tell us anything about another song. This makes sense, as each song has its own audio features, like tempo, fade times, and duration. Linearity is violated since we see non-symmetry in the residual plot around the horizontal axis. Constant variance seems to be violated as the variability of the residuals appears to increase for larger predicted values of hotness. Normality is satisfied, as there is not a large deviation from what is expected.

Possible Interactions

Different music genres cater to distinct audience expectations. Pop is the primary genre streamed on music listening platforms. However, in nicher genres, familiarity plays a bigger role than the style of the song. For instance, a listener may not be inclined to listen to heavy rock, but may know big names in the rock industry, like Blink-182. Thus, we have reason to believe that there is a relationship between genre and artist familiarity that is not merely additive, making it a significant factor to consider in models predicting song hotness.

term	estimate	std.error	statistic	p.value
(Intercept)	0.1262965	0.0774617	1.6304392	0.1032247
artist_familiarity	0.1778224	0.1342279	1.3247800	0.1854517
artist_hotness	0.3148694	0.0487192	6.4629380	0.0000000
duration	0.0011431	0.0006356	1.7985879	0.0722907
as.factor(gen_genre)pop	-0.0686405	0.0843354	-0.8138987	0.4158359
as.factor(gen_genre)rap	0.0901281	0.0964823	0.9341411	0.3503858
as.factor(gen_genre)rock	-0.1018558	0.0799402	-1.2741493	0.2028135
fade_out	-0.0010779	0.0006494	-1.6598735	0.0971549
fade_in	0.0004380	0.0025341	0.1728514	0.8627922
tempo	0.0002148	0.0001183	1.8152281	0.0696942
artist_familiarity:as.factor(gen_genre)pop	0.1916982	0.1389785	1.3793369	0.1680024
artist_familiarity:as.factor(gen_genre)rap	-0.0819153	0.1525650	-0.5369209	0.5914042
artist_familiarity:as.factor(gen_genre)rock	0.2197761	0.1335606	1.6455160	0.1000790

Metric	Value
R-squared	0.2613998
Adjusted R-squared	0.2613998

The artist hotness and genre term interaction contrasts that of artist familiarity and genre. In mainstream genres, like pop, an artist's hotness might significantly boost song popularity due to the genres' reliance on media exposure and trend cycles. Conversely, in nicher genres,

where music depth is more valued, the style of the music may trump artist hotness. Therefore, we have reason to believe that incorporation artist hotness and genre as an interaction term allows us to see how genre influences the relationship between an artist's market presence and a song's success, and how an artist's presence influences the relationship between genre and a song's success.

term	estimate	std.error	statistic	p.value
(Intercept)	0.1099767	0.0598905	1.8362953	0.0665175
artist_familiarity	0.3274338	0.0439129	7.4564404	0.0000000
artist_hotness	0.1558652	0.1263305	1.2337891	0.2174803
duration	0.0011633	0.0006371	1.8259348	0.0680645
as.factor(gen_genre)pop	-0.0064840	0.0637866	-0.1016516	0.9190472
as.factor(gen_genre)rap	-0.0302681	0.0806514	-0.3752951	0.7074956
as.factor(gen_genre)rock	-0.0783122	0.0602994	-1.2987230	0.1942444
fade_out	-0.0010939	0.0006510	-1.6804495	0.0930841
fade_in	0.0002392	0.0025400	0.0941696	0.9249873
tempo	0.0002102	0.0001187	1.7706322	0.0768308
artist_hotness:as.factor(gen_genre)pop	0.1194331	0.1354856	0.8815185	0.3781825
artist_hotness:as.factor(gen_genre)rap	0.1007304	0.1682126	0.5988278	0.5493808
artist_hotness:as.factor(gen_genre)rock	0.2406068	0.1288563	1.8672486	0.0620674

Metric	Value
R-squared	0.2577649
Adjusted R-squared	0.2577649

Comparing Models

To compare the models, we analyzed the adjusted r-squared values. Our linear model performed rather poorly, with an Adjusted R-squared of 0.2501. However, we wanted to run more than one model before deciding if these predictors were able to predict song hotness. Thus, we decided to fit another linear model, with an interaction term of artist familiarity and genre. We did this based on our exploratory data analysis that showed that there was some correlation between artist familiarity and song hotness for songs of certain genres. This model performed slightly better than our original linear model, with an adjusted R Squared of 0.2614. We also ran the linear model with an interaction term of artist hotness and genre. This model had an r squared value of 0.2517.

Addressing Violations

One concern we had was that linearity was violated for both models. The residual plots for both showed that the residuals were not randomly scattered, and there was a definite pattern towards the smaller residuals. While our choice to fit a linear regression model may seem

counterintuitive given the violation of the linearity assumption, it was a pragmatic decision based on several considerations. Firstly, linear regression is a widely used and well-understood modeling technique, making it accessible and interpretable for our analysis. Despite its reliance on the assumption of linearity, linear regression can still provide valuable insights and predictive accuracy under certain conditions. Specifically, the violation of the linearity assumption doesn't invalidate the entire model. Linear regression models can still yield reasonable results even when the relationship between the predictors and the response variable is not strictly linear. Furthermore, we are more concerned with the general performance of our model ; if it performed well or poorly is more of an indicator of if our predictors are good predictors, rather than if the model is a valid model or not.

Results

By the methodology above, we chose the model with the largest r^2 value – the second model, with an interaction term of artist familiarity and genre.

term	estimate	std.error	statistic	p.value
(Intercept)	0.1262965	0.0774617	1.6304392	0.1032247
artist_familiarity	0.1778224	0.1342279	1.3247800	0.1854517
artist_hotness	0.3148694	0.0487192	6.4629380	0.0000000
duration	0.0011431	0.0006356	1.7985879	0.0722907
as.factor(gen_genre)pop	-0.0686405	0.0843354	-0.8138987	0.4158359
as.factor(gen_genre)rap	0.0901281	0.0964823	0.9341411	0.3503858
as.factor(gen_genre)rock	-0.1018558	0.0799402	-1.2741493	0.2028135
fade_out	-0.0010779	0.0006494	-1.6598735	0.0971549
fade_in	0.0004380	0.0025341	0.1728514	0.8627922
tempo	0.0002148	0.0001183	1.8152281	0.0696942
artist_familiarity:as.factor(gen_genre)pop	0.1916982	0.1389785	1.3793369	0.1680024
artist_familiarity:as.factor(gen_genre)rap	-0.0819153	0.1525650	-0.5369209	0.5914042
artist_familiarity:as.factor(gen_genre)rock	0.2197761	0.1335606	1.6455160	0.1000790

We have insufficient evidence to suggest differential song hotness based on artist familiarity by genre, as the p-value for the interaction term is greater than our significance level of 0.05.

Given our p-value of 1.4×10^{-10} for artist hotness, which is less than $\alpha = 0.05$, we can see that there is sufficient evidence to suggest that there is a linear relationship between artist hotness and song hotness, which controlling for the other variables in our model. For every one unit increase in artist hotness, we expect an average increase of 0.3148694 in song hotness , while controlling for the other variables in our model.

Discussion

Our models performed poorly, indicating that our predictors weren't good predictors. This is evident in the r^2 values, which were in the 0.2 range. r^2 values range from 0 to 1, where 0 indicates that none of the variation in the response variable is accounted for by the predictors. This may be partly attributed to data limitations, such as the dataset's bias towards mainstream genres and lack of representation for niche or emerging music styles. Furthermore, as discussed before, we had concerns with the violation of the linearity assumption, but decided to look at our models regardless. Future work could include a broader range of genres, and take into account more predictor values, such as song key and loudness.

In our final model, we were not surprised to find that there was insufficient evidence to suggest song hotness based on artist familiarity by genre. Especially in the present day, we don't see people listening to songs merely because they are by well-known artist names. Instead, we tend to focus on trends, specifically artist hotness, which is why on TikTok, where smaller artists like wave2earth, Dasha, and DJO gain fame through their stylistic choices, but are also easily forgotten as new trends with different types of music emerge. In 2011, the same principle seems to have been true; as trends that are influenced by our world came and went, they influenced the type of music that was popular each year.

Appendix

<http://millionsongdataset.com/>

<https://songtown.com/on-songwriting/what-makes-a-great-song-a-hit-song/>

<https://www.theguardian.com/music/2023/dec/11/viral-hit-tiktok-gen-z-mother-ladytron-keane-miguel>

<https://www.today.com/popculture/music/laufey-jazz-classical-music-gen-z-rcna56020>