Final Project

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.

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 high 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.

Cleaning Process

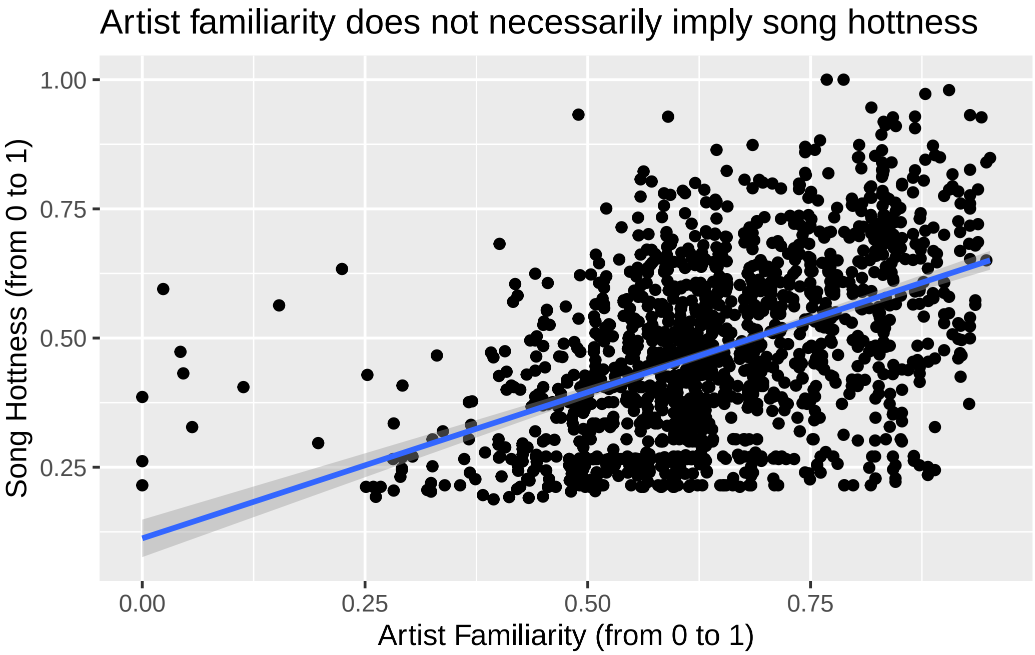
We narrowed our focus to four primary genres: rap, rock, pop, and country, capturing a broad spectrum of musical styles while minimizing complexity. We further simplified the genre landscape by generalizing subgenres (ex: Classical rock) into overarching categories (Rock), and creating a new variable, gen_genre, with those overarching categories. Then, we filtered for all song.hotttnesss 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.hotttnesss of -1. We concluded that it was a placeholder for datapoints with a null value for song hotness).

```
# A tibble: 6 x 36
  artist.familiarity artist.hotttnesss artist.id artist.latitude artist.location
               <dbl>             <dbl> <chr>               <dbl>           <dbl>
1              0.651             0.402 ARXR32B1~             0                 0
2              0.636             0.448 ARD842G1~            43.6               0
3              0.707             0.513 ARYKCQI1~             0                 0
4              0.435             0.306 AR47JEX1~            37.8               0
5              0.809             0.488 ARPQ4Z01~            29.4               0
6              0.661             0.443 ARV1JVD1~            35.9               0
# i 31 more variables: artist.longitude <dbl>, artist.name <chr>,
#   artist.similar <dbl>, artist.terms <chr>, artist.terms_freq <dbl>,
#   release.id <dbl>, release.name <dbl>, song.artist_mbtags <dbl>,
#   song.artist_mbtags_count <dbl>, song.bars_confidence <dbl>,
#   song.bars_start <dbl>, song.beats_confidence <dbl>, song.beats_start <dbl>,
#   song.duration <dbl>, song.end_of_fade_in <dbl>, song.hotttnesss <dbl>,
#   song.id <chr>, song.key <dbl>, song.key_confidence <dbl>, ...
```

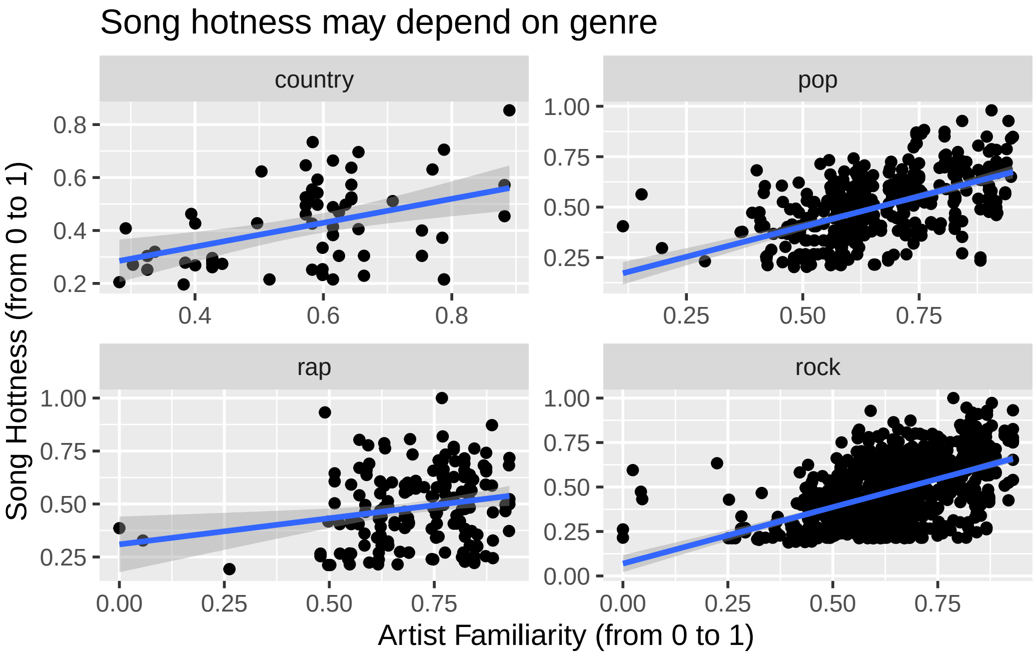
EDA

We hypothesized two of the most important factors in determining song popularity was artist familiarty 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.

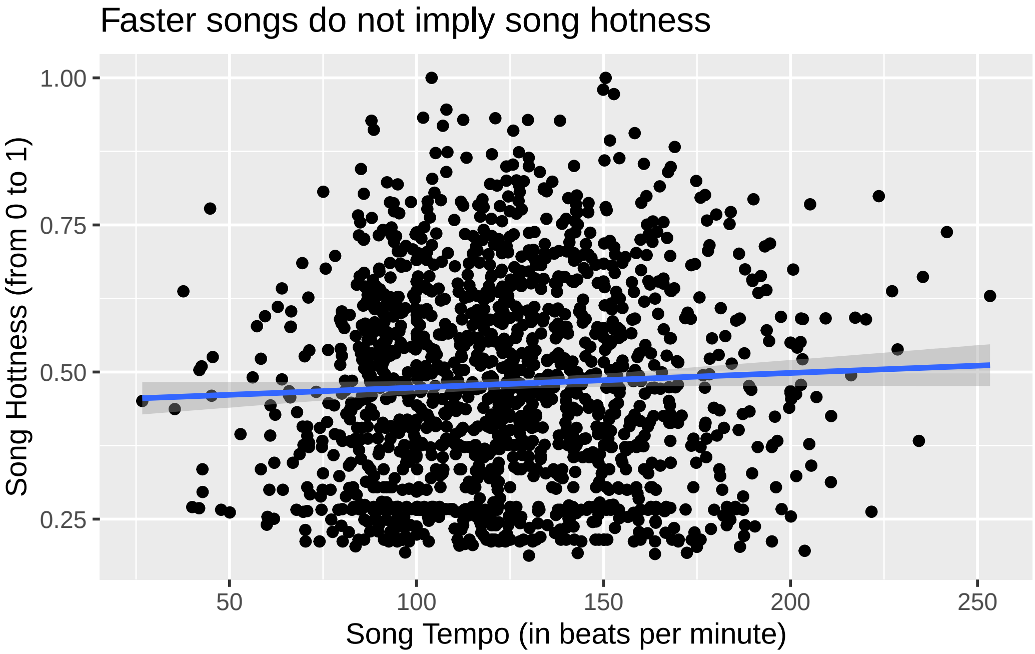
```
`geom_smooth()` using formula = 'y ~ x'
```

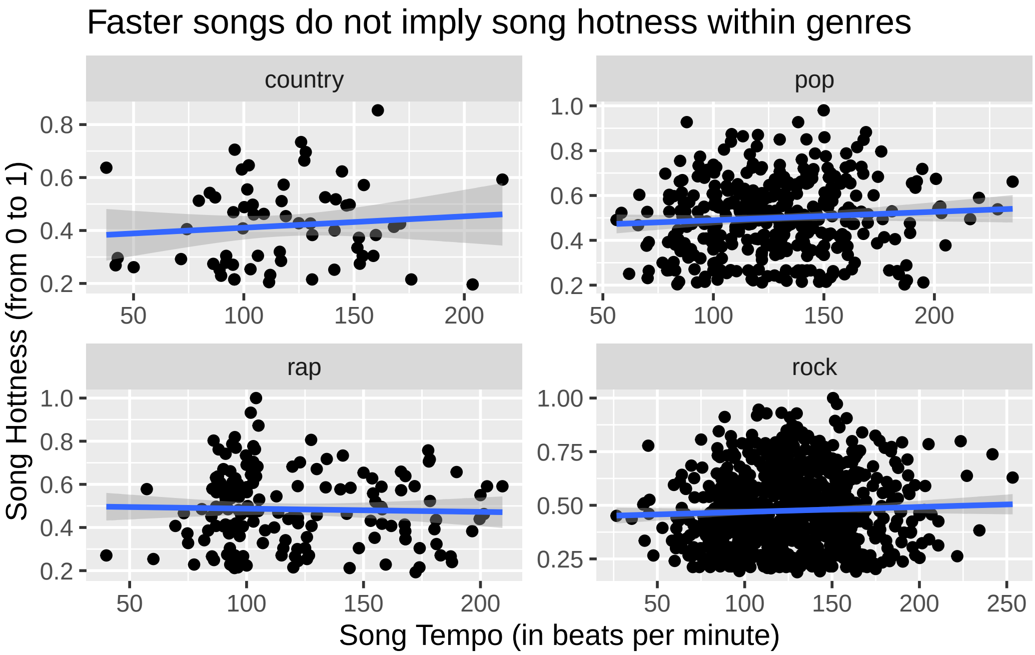


```
`geom_smooth()` using formula = 'y ~ x'
```



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

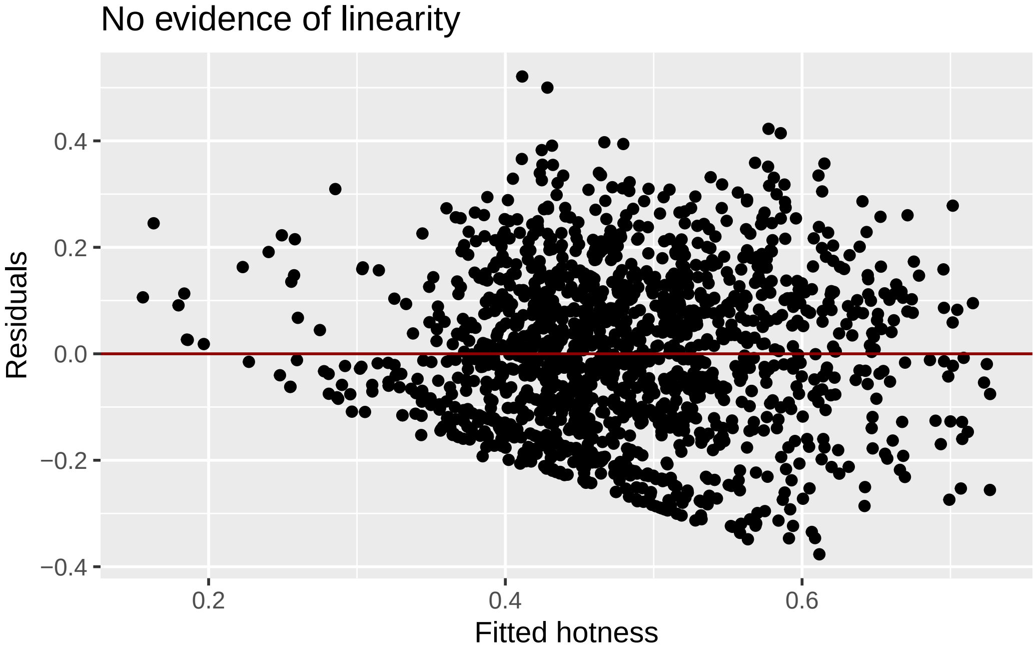
To get a comprehensive understanding of how well (or poorly) the predictors we had picked could predict song hotness, we decided to create more than one model. Since our response variable, song hotness, is a continuous numeric variable, we will be using a linear regression model.

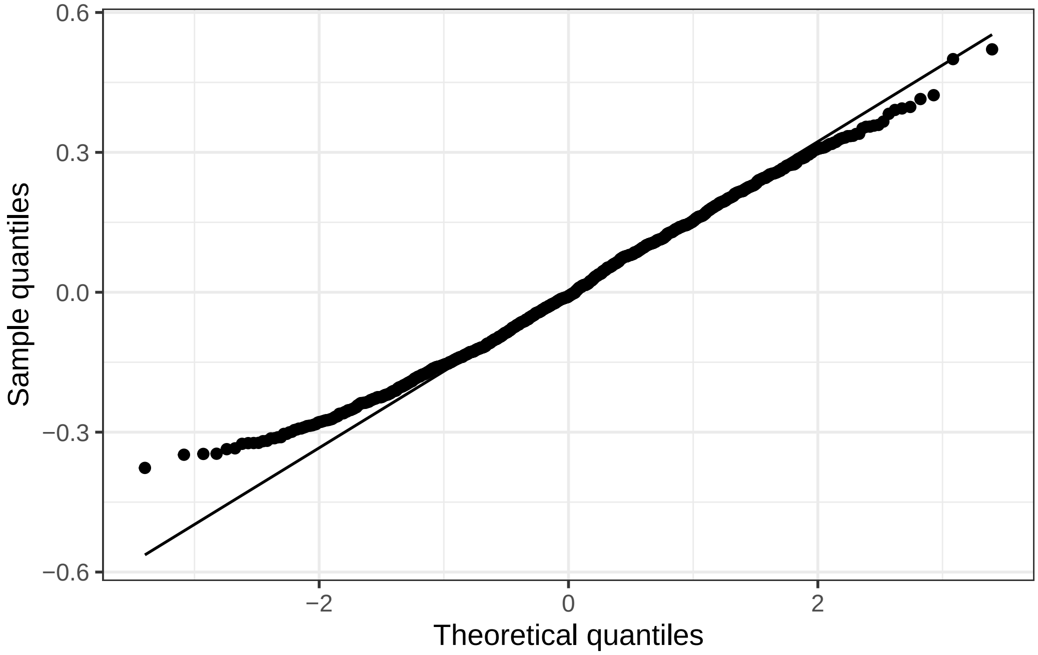
First, we fit a regular linear model, using artist familiarity, artist hottness, song duration, genre, start of fade out, end of fade in, and song tempo as the predictors.

```
# A tibble: 1,471 x 36
   artist.familiarity artist.hotttnesss artist.id          artist.latitude
                <dbl>             <dbl> <chr>                        <dbl>
 1              0.651             0.402 ARXR32B1187FB57099             0
 2              0.636             0.448 ARD842G1187B997376            43.6
 3              0.707             0.513 ARYKCQI1187FB3B18F             0
 4              0.435             0.306 AR47JEX1187B995D81            37.8
 5              0.809             0.488 ARPQ4Z01187FB3A736            29.4
 6              0.661             0.443 ARV1JVD1187B9AD195            35.9
 7              0.718             0.479 ARS10WB1187B99EEAD             0
 8              0.570             0.412 AROEL1B1187B988B90             0
 9              0.643             0.501 AR3793X1187FB50CB3             0
```

```
10              0.751             0.524 ARDGB6U1187FB3AD07            51.5
# i 1,461 more rows
# i 32 more variables: artist.location <dbl>, artist.longitude <dbl>,
#   artist.name <chr>, artist.similar <dbl>, artist.terms <chr>,
#   artist.terms_freq <dbl>, release.id <dbl>, release.name <dbl>,
#   song.artist_mbtags <dbl>, song.artist_mbtags_count <dbl>,
#   song.bars_confidence <dbl>, song.bars_start <dbl>,
#   song.beats_confidence <dbl>, song.beats_start <dbl>, ...

# A tibble: 10 x 5
   term                      estimate std.error statistic  p.value
   <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)               0.0296    0.0298      0.994  3.20e- 1
 2 artist.familiarity        0.330     0.0439      7.53   9.09e-14
 3 artist.hotttnesss         0.335     0.0484      6.92   6.58e-12
 4 song.duration             0.00113   0.000637    1.78   7.52e- 2
 5 as.factor(gen_genre)pop   0.0446    0.0209      2.14   3.28e- 2
 6 as.factor(gen_genre)rap   0.0103    0.0229      0.450  6.53e- 1
 7 as.factor(gen_genre)rock  0.0285    0.0199      1.43   1.53e- 1
 8 song.start_of_fade_out   -0.00106   0.000651   -1.63   1.03e- 1
 9 song.end_of_fade_in       0.000104  0.00254     0.0409 9.67e- 1
10 song.tempo                0.000205  0.000119    1.73   8.38e- 2


Call:
lm(formula = song.hotttnesss ~ artist.familiarity + artist.hotttnesss + 
    song.duration + as.factor(gen_genre) + song.start_of_fade_out + 
    song.end_of_fade_in + song.tempo, data = music_filter)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37666 -0.11609 -0.00803  0.10539  0.52091 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.0295918  0.0297706   0.994   0.3204    
artist.familiarity        0.3302057  0.0438741   7.526 9.09e-14 ***
artist.hotttnesss         0.3349121  0.0483736   6.923 6.58e-12 ***
song.duration             0.0011337  0.0006368   1.780   0.0752 .  
as.factor(gen_genre)pop   0.0446009  0.0208737   2.137   0.0328 *  
as.factor(gen_genre)rap   0.0102870  0.0228824   0.450   0.6531    
as.factor(gen_genre)rock  0.0284675  0.0199278   1.429   0.1534    
```

```
song.start_of_fade_out   -0.0010607  0.0006506  -1.630   0.1033
song.end_of_fade_in       0.0001040  0.0025412   0.041   0.9673
song.tempo                0.0002053  0.0001187   1.730   0.0838 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1512 on 1461 degrees of freedom
Multiple R-squared:  0.2547,    Adjusted R-squared:  0.2501
F-statistic: 55.48 on 9 and 1461 DF,  p-value: < 2.2e-16
```





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.

```
# A tibble: 13 x 5
```

```
   term                                    estimate std.error statistic  p.value
   <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                              1.26e-1  0.0775       1.63  1.03e- 1
 2 artist.familiarity                       1.78e-1  0.134        1.32  1.85e- 1
 3 artist.hotttnesss                        3.15e-1  0.0487       6.46  1.40e-10
 4 song.duration                            1.14e-3  0.000636     1.80  7.23e- 2
 5 as.factor(gen_genre)pop                 -6.86e-2  0.0843      -0.814 4.16e- 1
 6 as.factor(gen_genre)rap                  9.01e-2  0.0965       0.934 3.50e- 1
 7 as.factor(gen_genre)rock                -1.02e-1  0.0799      -1.27  2.03e- 1
 8 song.start_of_fade_out                  -1.08e-3  0.000649    -1.66  9.72e- 2
 9 song.end_of_fade_in                      4.38e-4  0.00253      0.173 8.63e- 1
10 song.tempo                               2.15e-4  0.000118     1.82  6.97e- 2
11 artist.familiarity:as.factor(gen_genre~  1.92e-1  0.139        1.38  1.68e- 1
12 artist.familiarity:as.factor(gen_genre~ -8.19e-2  0.153       -0.537 5.91e- 1
13 artist.familiarity:as.factor(gen_genre~  2.20e-1  0.134        1.65  1.00e- 1
```

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.

```
# A tibble: 13 x 5
   term                                    estimate std.error statistic  p.value
   <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                              1.10e-1  0.0599      1.84   6.65e- 2
 2 artist.familiarity                       3.27e-1  0.0439      7.46   1.52e-13
 3 artist.hotttnesss                        1.56e-1  0.126       1.23   2.17e- 1
 4 song.duration                            1.16e-3  0.000637    1.83   6.81e- 2
 5 as.factor(gen_genre)pop                 -6.48e-3  0.0638     -0.102  9.19e- 1
 6 as.factor(gen_genre)rap                 -3.03e-2  0.0807     -0.375  7.07e- 1
 7 as.factor(gen_genre)rock                -7.83e-2  0.0603     -1.30   1.94e- 1
 8 song.start_of_fade_out                  -1.09e-3  0.000651   -1.68   9.31e- 2
 9 song.end_of_fade_in                      2.39e-4  0.00254     0.0942 9.25e- 1
10 song.tempo                               2.10e-4  0.000119    1.77   7.68e- 2
11 artist.hotttnesss:as.factor(gen_genre)~  1.19e-1  0.135       0.882  3.78e- 1
12 artist.hotttnesss:as.factor(gen_genre)~  1.01e-1  0.168       0.599  5.49e- 1
13 artist.hotttnesss:as.factor(gen_genre)~  2.41e-1  0.129       1.87   6.21e- 2
```

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 0.255. 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.

```
# A tibble: 13 x 5
   term                              estimate std.error statistic  p.value
   <chr>                                <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                        1.26e-1  0.0775       1.63  1.03e- 1
 2 artist.familiarity                 1.78e-1  0.134        1.32  1.85e- 1
 3 artist.hotttnesss                  3.15e-1  0.0487       6.46  1.40e-10
 4 song.duration                      1.14e-3  0.000636     1.80  7.23e- 2
 5 as.factor(gen_genre)pop           -6.86e-2  0.0843      -0.814 4.16e- 1
 6 as.factor(gen_genre)rap                   9.01e-2  0.0965      0.934 3.50e- 1
 7 as.factor(gen_genre)rock                 -1.02e-1  0.0799     -1.27  2.03e- 1
 8 song.start_of_fade_out                   -1.08e-3  0.000649   -1.66  9.72e- 2
 9 song.end_of_fade_in                       4.38e-4  0.00253     0.173 8.63e- 1
10 song.tempo                                2.15e-4  0.000118    1.82  6.97e- 2
11 artist.familiarity:as.factor(gen_genre~  1.92e-1  0.139       1.38  1.68e- 1
12 artist.familiarity:as.factor(gen_genre~ -8.19e-2  0.153      -0.537 5.91e- 1
13 artist.familiarity:as.factor(gen_genre~  2.20e-1  0.134       1.65  1.00e- 1
```

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.

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. In the future, we will 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.