Top Bops and Pop Flops: What Makes a Hit Song?

Eric Lee

Jun 9, 2024

Contents

Summary

Noting the evolving nature of the music industry in the modern era and the common desire among artists to create widely popular songs, this report investigates the factors that make a viral hit and how those factors have changed over time. To conduct this exploration, this report jointly uses data from the Billboard Hot 100 and audio features from the Spotify API to gather diverse features for all charting songs from 2000 to 2019, with two response variables: a song's peak chart ranking and total time spent on the charts. This data is preprocessed and cleaned to adjust for errors and outliers, and preliminary analysis is conducted on each feature with visualizations. This report uses forward selection to determine the significant features for the two response variables, and uses linear models and F-tests to verify the significance of the selected features. Further visualizations are plotted and analyzed both in aggregate and year-by-year for each such feature. Such analysis shows that the most important factors in determining chart performance pertain to the character, production, genre, and collaboration in the making of each song. The impact of these features shift over time with relatively steady trends, and impact peak performance and chart longevity in different ways. These results hold greater significance in enabling budding artists and the music industry to create targeted music more likely to become successful in the context of recent trends.

Introduction

The music industry is a dynamic and ever-evolving landscape, characterized by the constant pursuit of creating hit songs that capture the attention of audiences worldwide. With the rise of digital streaming platforms and social media, the process of achieving viral success has become increasingly complex. Although countless songs are written and released by aspiring pop stars every day, only a select few are successful in ascending to the top of the charts.

In particular, we encounter a wide range of interesting phenomena. Although the top charts mainly feature well-established pop stars, there are scattered instances of indie artists making their first big break. Some artists enjoy remarkable consistency and longevity on the charts, while others only feature as one-hit wonders before fading into obscurity. While songs of different genres, backgrounds, and styles achieve virality, many of the most popular songs adhere to familiar structures and conventions. Moreover, while English songs dominate the charts, there are also appearances from songs in international languages like Korean pop and Latin bachata. These phenomena all lead to one central question: **what are the factors make a song a viral hit in the modern era?**

Significance

The answer to this question holds significance in many regards. First, the ability to release a chart-topping hit is critically important to the livelihoods of budding artists and established stars alike. Producing a song requires time and nontrivial fixed costs, so maximizing the popularity of each song is necessary for artists to sustain themselves. Specifically, a viral hit can bring an artist not only high streaming numbers that translate directly to earnings, but also a level of exposure that brings fame, partnerships, and brand deals.

Second, these insights are crucial in elevating the music of artists who do not fit the mold of the status quo, such as indie artists, international singers, and stylistic pioneers. Discovering the factors behind hit songs and applying that knowledge to bring exposure to unique songs and artists can bring innovation to the music industry while elevating underrepresented voices in a saturated field.

Third, viral hits often transcend mere commercial success and shape popular culture, so understanding the recipe behind viral hits can yield insights about cultural dynamics, social discourse, and popular ideas. Especially since chart-topping hits are often continuously played in phones, cars, and stores all over the world, their influence is extremely pervasive. Moreover, on social media platforms, these hits govern the top trends with regard to language, dance, and humor.

Context and Novelty

Given the significance of this topic, there have been many prior efforts in this field to predict and analyze the popularity of songs based on various audio features. These articles are valuable in informing a strong foundation and logical approach to exploring this question, but they leave an evident gap in the literature investigating the post-2000 era of streaming and social media jointly using data about both top charts and audio features.

An early effort in 2017 was Nijkamp's comprehensive analysis[1] of the relationship between the audio features and stream counts of 1000 songs sourced from Spotify. In this analysis, Nijkamp focused on the impact of audio features in isolation, discovering that certain features were positively or negatively correlated to higher stream counts. While this feature-wise approach offers insights with regard to the relative effects of each audio feature on stream counts, the

[1] Rutger Nijkamp (2024) *Prediction of Product Success: Explaining Song Popularity by Audio Features from Spotify Data*, University of Twente

analysis is limited with regard to evaluating the significance of each feature and their trends over time. A further limitation is that the feature of interest—stream counts—was taken from Spotify and thus excludes other important factors in popularity such as stream counts on other platforms, plays on social media, and plays on radio. In my project, I aim to expand on this article's analysis with additional evaluation of multi-feature regression models and use data from the Billboard charts[2] whose metric for song popularity factors in plays and streams from other platforms.

A related 2020 analysis[3] expands on this research by investigating trends in the featured songs on the Billboard charts. In particular, the article finds that in recent decades, the songs that debut on the charts have become more homogeneous in style with a trend toward hits with higher energy and shorter duration. These insights are notable and point to the need for more focused exploration into the trends dictating the popular hits of the last two for musical features beyond song duration and energy. Further, while this analysis solely focused on insights to be gleaned from Billboard data alone, I look to delve deeper by combining this dataset with Spotify data in order to analyze a more comprehensive set of audio features.

Many other research efforts investigating this question used the Million Song Database provided by Columbia University[4], such as articles by Nasreldin[5] and Pham et al[6]. This dataset is a comprehensive dataset of one million songs published until 2011, containing information about audio features and metadata for each song. These articles corroborate in identifying genre and artist familiarity as prominent features in determining popularity, informing my hypotheses for my analysis. However, this dataset is inherently limited due to its scope in only including songs up to 2011 (missing more than a decade of music up to 2024) and its lack of information regarding performance on the charts. Therefore, I aim to glean novel insights about the last decade of music that these articles are unable to address with a more updated and comprehensive dataset including detailed features up to 2019.

Hypotheses

Considering the complexity of the phenomenon of virality and multivariate nature of the dataset, it is prudent to consider multiple hypotheses with regard to different facets of the research question. Namely, exploring virality entails not only investigating differences between outcome variables, but also examining the impact of four broad factors: characteristic, production, genre, and collaboration.

First, considering that the success of a track is determined by not

[2] Sean Miller (2024) *Billboard Hot 100 Weekly Charts with Spotify Audio Features*, Kaggle

[3] Azhad Syed *Hot or Not: Analyzing 60 Years of Billboard Hot 100 Data*, Toward Data Science

[4] Eric Lee *Data Dictionary*

[5] Mohamed Nasreldin (2018) *Song Popularity Predictor*, Medium

[6] Pham et al. (2015) *Predicting Song Popularity*, Stanford Department of Computer Science

only its rise but its longevity, it is of interest to examine the tradeoff between the two outcome variables: peak position and total weeks on the charts. Especially in recent years, many songs experience a meteoric but short-lived rise to the top of the charts. Meanwhile, other songs are able remain radio classics for many months without cracking the upper echelon of the charts. Thus, I hypothesize that the set of significant factors in determining peak performance is different than the set of significant factors impacting chart longevity.

Second, it is important to consider how the characteristic of each song impacts its performance. The highest ranking songs on the charts often gain popularity on social media platforms (ex. TikTok) and other large events (ex. nightclubs, football games, etc.), where catchy tunes easily compatible with singing and dancing seem to be most rewarded. However, such characteristics seem less crucial for long-lasting hits, as there are many examples of slower ballads that are renowned as pop classics. Therefore, I hypothesize that short songs with an energetic characteristic are more likely to reach the top of the charts, while there is less of a definitive association with regard to chart longevity.

Third, as a complement to the characteristic of each song, it is prudent to examine the impact of audio features contributing to song production. Especially with on streaming platforms where songs are repeated many times, strong production with regard to volume and instrumentation are increasingly important for high replayability value. My hypothesis is that such production audio features are strongly related to chart longevity, and also associated with peak ranking to some extent.

Fourth, noting that genre is closely tied to the style and construction of each song, it is relevant to see whether certain styles make a song more conducive to become a hit. Observing the dominance of pop music in American culture, I hypothesize that songs in the pop genre have higher peaks and longer stays on the charts. Further, I hypothesize that songs in other recognized non-pop genres like rap and country perform better on the charts than songs who are not in any commonly recognized genre.

Fifth, it is also critical to consider other factors such as the effects of collaborations with popular artists and the recognition gained from appearing on the charts to begin with. Many songs on the charts feature multiple artists, with numerous examples of partnerships between established pop stars and rising up-and-comers. In considering the combined pull of multiple artists in addition to instances where well-established stars serve as featured artists, I hypothesize that songs including more collaborations are more likely to reach success on the charts in comparison to songs by solo artists.

Data

To investigate this research question, I gathered a dataset with all songs featured on Billboard's Hot 100 chart from Jan 1, 2000 to Dec 31, 2019 (a two-decade period representative of the streaming era of music) and their features informing song characteristic, production, genre, and collaboration. Specifically, I did this by combining the data from a public dataset[7], which was created by scraping chart data from Billboard and querying the Spotify API for audio features. For purposes of reproducibility, see the references in the margin to access the Python code[8] used for the data processing and the creation of the resulting dataset[9] (CSV file).

Creating New Variables

In the original dataset, each song is associated with a list of genres. To allow numerical analysis on this data to investigate my hypotheses about genres, I added new indicator variables for popular genres to encode this text data as binary (1 = belongs to genre, 0 = does not belong to genre). In particular, I created new variable for 7 known genres: Pop, Country, Rock, R&B, Indie, Rap/Hip-Hop, and Blues. Songs that did not belong to any of the above 7 genres were categorized as OtherGenre, and songs that were not tagged by Spotify as having any genre were classified as NoGenre. Note that songs can be tagged with more than one of the known genres, but is strictly in one or more known genres, OtherGenre, or NoGenre.

Similarly, to convert the textual list of artists into numerical data about the collaboration, I created two new variables for each song using batch operations and string parsing. The Collaborators variable encapsulates the number of collaborators on each song (0 if the song is a solo track), and the FamousCollaborators is a binary variable indicating whether any collaborator has had a previous hit on the charts (1 if so, 0 if not).

Final Dataset

As a result of the data processing steps described above, I arrived at a final dataset with 8,664 songs (rows) and 34 variables (columns). As an overview, the variables are detailed in the following table. For more detailed information regarding these variables and reproducibility notes, please refer to the detailed data dictionary[10].

[7] Sean Miller (2024) *Billboard Hot 100 Weekly Charts with Spotify Audio Features*, Kaggle

[8] Eric Lee *Data Processing.ipynb*

[9] Eric Lee *Processed Data.csv*

[10] Eric Lee *Data Dictionary*

Variable	Type	Range	Description
Song	string	–	Name of the song
Performer	string	–	List of artists on the song
SongID	string	–	Unique ID identifying each song
Month	int	{1, . . . , 12}	Month the song reached its peak position
Day	int	{1, . . . , 31}	Day the song reached its peak position
Year	int	{2000, . . . , 2019}	Year the song reached its peak position
PeakPosition	int	{0, . . . , 100}	Peak position on Hot 100 chart (rank 1 is best)
WeeksOnChart	int	{1, . . . , 87}	Total number of weeks on Hot 100 chart
Duration	float	[0, 1]	Duration of song in milliseconds
Danceability	float	[0, 1]	Score for song danceability
Energy	float	[0, 1]	Score for song intensity and activity
Key	int	{1, . . . , 11}	Key of song by Pitch Class notation
Loudness	float	[-23.023, 0.175]	Loudness of song in decibels
Mode	float	{0, 1}	0 = minor mode, 1 = major mode
Speechiness	float	[0, 1]	Score for presence of spoken words in song
Acousticness	float	[0, 1]	Score for confidence the song is acoustic
Instrumentalness	float	[0, 1]	Score for reliance on instrumentals in song
Liveness	float	[0, 1]	Score for likelihood recording is live
Valence	float	[0, 1]	Score for musical positiveness in song
Tempo	float	[48.72, 213.74]	Tempo of song in beats per minute
Time Signature	int	{1, . . . , 5}	Score for song intensity and activity
Pop	int	{0, 1}	Indicator if song belongs to pop genre
Country	int	{0, 1}	Indicator if song belongs to country genre
Rock	int	{0, 1}	Indicator if song belongs to rock genre
Blues	int	{0, 1}	Indicator if song belongs to blues genre
R&B	int	{0, 1}	Indicator if song belongs to R&B genre
RapHipHop	int	{0, 1}	Indicator if song belongs to rap/hip-hop genre
Indie	int	{0, 1}	Indicator if song belongs to indie genre
OtherGenre	int	{0, 1}	Indicator if song belongs none of the above genres
NoGenre	int	{0, 1}	Indicator if song is tagged with no genres
FamousCollaborator	int	{0, 1}	Indicator if a collaborator has previous hit
Collaborators	int	{0, 1}	Total number of collaborators

Quality Control

An inspection of the data shows no obvious nonsense values. Additionally, the reputability of the sources (Billboard and Spotify) gives confidence that the chance for errors in the data is relatively low. Even in the case where there were inherent errors in the original data tables, the data processing described above addresses many of these concerns. Namely, since the merge between tables was anchored on the unique key `SongID`, songs with incorrectly spelled IDs or mismatches between tables were excluded from the final dataset. Further, given that all rows representing songs outside of the years 2000-2019 was sliced out of the final dataset, there is a reduced chance of error caused by limitations in tracking or technology.

Although there is low likelihood for errors in the dataset, it should be noted that some rows in the final table have `NA` values for certain audio features due to limitations in accessing the Spotify API. There are 7,849 rows of complete data (without `NA` values for any variable) out of the 8,664 rows in the final dataset, meaning there are 815 rosws with at least one missing value.

Given that a very high percentage (90.5%) of the rows in the final dataset have values for each variable, all models in the analysis conducted below simply exclude the rows with missing values. Additionally, in Figure 1 it can be seen that the 815 songs with at least one missing value are spread relatively evenly across each year, with no year contributing more than 6% to the missing values. Since there is no noticeable trend in this distribution, these missing values can attributed to randomness as opposed to a meaningful confounding variable.

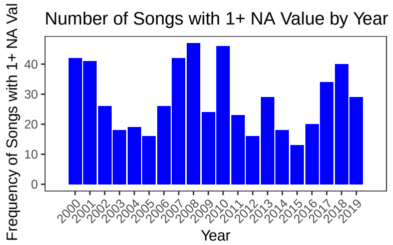


Figure 1: Distribution of songs with one or more N/A values per year.

Thus, the full linear model conducted on all variables utilizes 90.5% of the original dataset, with all other models focusing on a subset of the variables utilizing at least 7,849 rows. Since analyses will be conducted with a shifting focus on each variable, it makes sense to leave `NA` values in this main `data` table while removing the relevant rows in individual operations by variable below.

Summary of Variables

Note that in analyzing the variables, chart performance (`PeakPerformance`) and chart longevity (`WeeksOnChart`) are considered the outcome variables while the other numeric/boolean variables are considered the explanatory variables.

Outcome Variables

In this dataset, there are two variables to quantify virality and success on the charts, namely peak position and longevity. Specifically, peak position refers to the highest rank a given song achieves throughout its stay on the Hot 100 charts. The first step is to investigate the distribution of the peak positions of songs across this time period in Figure 2. As expected, there are many more songs that peak near the bottom of the charts, and increasingly fewer that reach the very top of the charts near the top position.

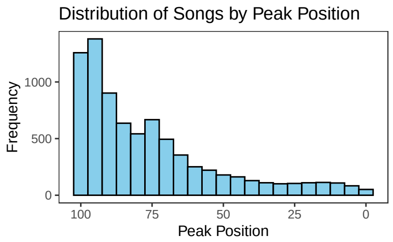


Figure 2: Distribution of songs by peak position.

In turn, the next step is to investigate the distribution of songs by longevity, namely in weeks remaining on the charts. The longevity of a song is defined by the cumulative sum of the weeks it spends on the Hot 100 before being removed. The general trend of the histogram in Figure 3 is unsurprising, that the majority of songs fall off the charts very quickly after debut, with fewer and fewer remaining for a very long time (50+ weeks) as seen by the heavy right tail. Although, there is a very noticeable spike at 20 weeks, indicating that many songs in the database fell off of the charts after exactly 20 weeks. This unusual behavior can be attributed to Billboard's chart policy that songs which are "descending are removed from the Billboard Hot 100... after 20 weeks and if ranking below No. 50" as a way to combat chart inertia and feature new songs gaining popularity and momentum[11]. This policy does not necessitate any major changes in our analysis, as it represents the natural phenomenon of "slipping off the charts" due to obscurity. However, it is wise to remain cognizant that the 20 week mark is likely to appear very frequently, and represents a song that could have potentially lasted slightly longer than 20 weeks (before officially falling off).

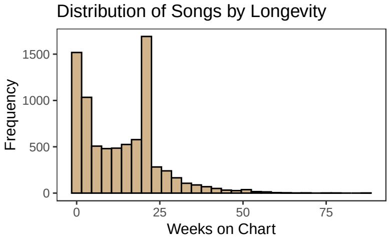


Figure 3: Distribution of songs by total number of weeks on chart.

Audio Features

Next, with a focus on gaining insight into our first hypothesis regarding the importance of different audio features, this section explores the distributions of each audio feature alongside its relationship with the outcome variables.

Before delving into individual audio features, it is prudent to examine the boxplots for each audio feature with relation to peak chart performance and chart longevity in Figure 4 as an exploratory overview. First, consider the audio features generated by Spotify as scores between 0 and 1. This display gives a preliminary look at the general center and spread of each of these variables. Notably, `instrumentalness` seems to be at 0 for the vast majority of songs, while `mode` seems to be at 1 for most songs. Additionally, the centers for `acousticness`, `liveness`, and `speechiness` are relatively low, indicating that these features have a lower general score for recent songs that reach the charts. No outliers are noticed here, and all data falls within 0 and 1 (inclusive), as expected.

Outliers

Similarly, the summary of the center and spread of the other audio features (not between 0 and 1) are within the boxplots below. Apart from the observations of the different centers of each features, there are a few notable outliers that need to be addressed.

First, as seen in the Duration boxplot in Figure 5, there is a song

[11] Billboard (2024) *Billboard Charts Legend: Recurrent Rules*, Billboard

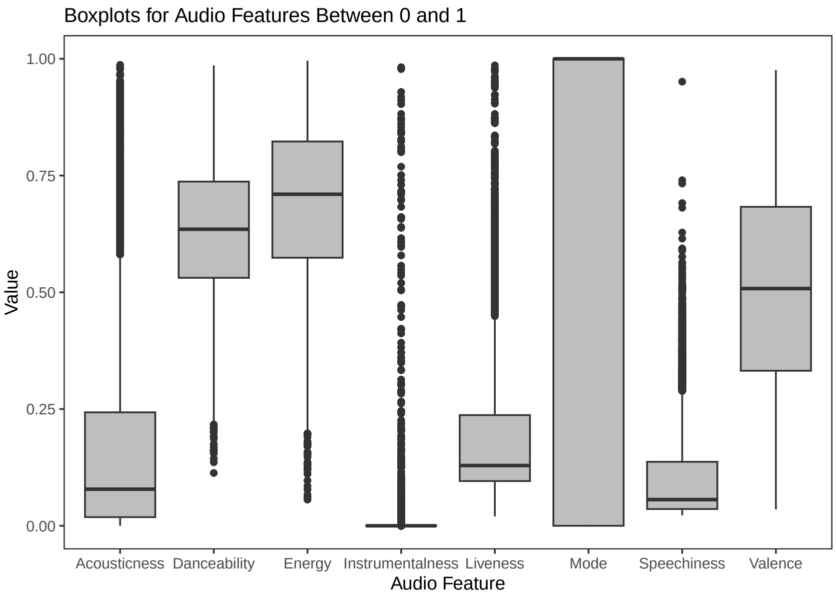


Figure 4: Boxplots for audio features with scores between 0 and 1.

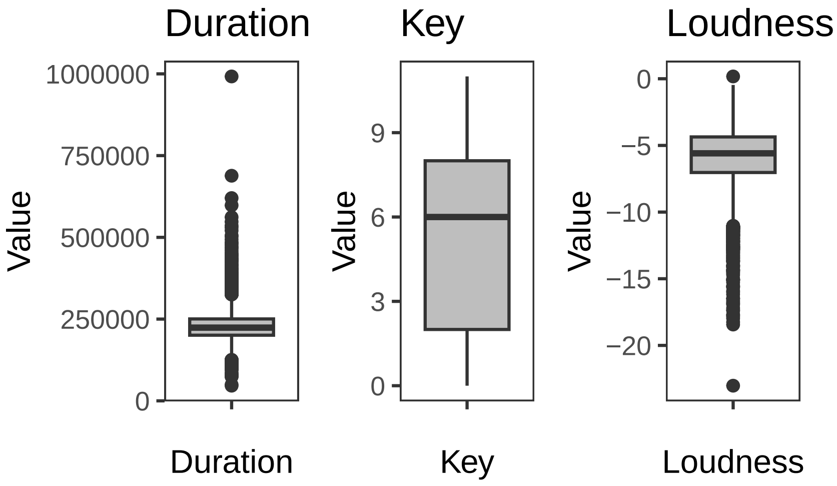


Figure 5: Boxplots for audio features duration, key, and loudness.

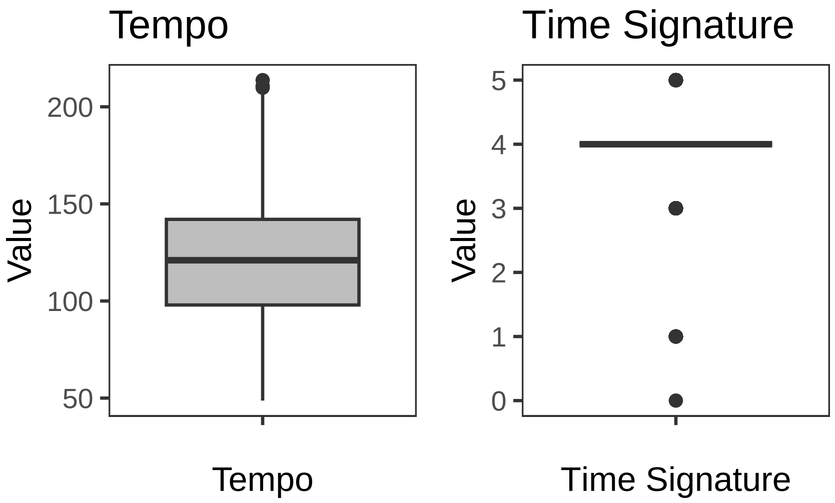


Figure 6: Boxplots for audio features tempo and time signature.

that is extremely long. This outlier is R. Kelly's song, "Trapped in the Closet". Checking with the original Spotify track[12] shows that this song is indeed 16 minutes (992160 ms), so this outlier is not born from error.

Second, there is a very notable outlier near the bottom of the Loudness boxplot (Figure 5. This outlier is Billie Eilish's song, "Listen Before I Go" with -23.023 dB. This value is out of range of Spotify's typical loudness levels, so this value is reverted to `NA` out of an abundance of caution.

Third, there is a seemingly nonsense value in the Time Signature boxplot (Figure 6) showing a song with time signature 0 (which is not possible). This point turns out to be "Imma Be" by "The Black Eyed Peas", which indeed has 0 as the value for `TimeSignature` in the data table. This is clearly an error, as the song has a time signature of 4 beats per bar per other sources [13]. Therefore, we change the value for this song from 0 to 4 for `TimeSignature`.

In addition to analysis with outliers, it should be noted that the distribution for all variables is approximately normal in Figure 38 (Appendix), with the exception of Instrumentalness, Acousticness, Liveness, and Speechiness (which are skewed with a heavy tail). Thus, it is justified to apply a log transformation to these four variables such that the resulting distribution is much more normal, seen in Figure 39 (Appendix). Given that these transformed features being approximately normal, we gain confidence in their application within a linear model.

[12] R. Kelly (2005) *Trapped In The Closet*, Spotify

[13] SongBPM *Song Metrics*, Imma Be

Genres

As described above, a percentage of songs were not labeled with any data, and these songs are classified as NoGenre. For the other songs which have at least one labeled genre, they are labeled as KnownGenre if they fit into Pop, Country, Rock, R&B, Blues, RapHipHop, or Indie. The bulk of the songs fit into at least one of the aforementioned known genres, and that hence these 7 genres encompass the vast majority of the songs that end up on the Hot 100 charts. Songs that do not fit in one of these genres are classified as OtherGenre. The following table displays the distribution of songs across genres (noting that each song can belong to multiple genres, but only one of KnownGenre, OtherGenre, and NoGenre). More detailed boxplots for each of these genre variables can be found in Figures 40 and 41 (Appendix).

Genre	% of Songs in Genre	% of Songs Not in Genre
Pop	65.6%	34.4%
Country	18.1%	81.9%
Rock	20.5%	79.5%
R&B	13.0%	87.0%
Indie	2.2%	97.8%
Rap/Hip-Hop	38.9%	61.1%
Blues	0.3%	99.7%
OtherGenre	7.7%	92.3%
NoGenre	4.6%	95.4%

Collaboration

A minority of songs (19.9%) feature famous collaborators, and Figure 42 (Appendix) shows that the average peak position of all songs without famous collaborators (75.3505) is only slightly worse than the average peak position of all songs with famous collborators (74.6282). This is preliminary evidence that the presence of famous collaborators only has a minimal effect on peak performance.

In addition to considering the effects of the presence of famous collaborators, it is also important to consider the total number of collaborators on a given song (who may or may not be famous or previously featured on the charts). As in Figure 42 (Appendix), the vast majority of songs do not have collaborators, and for the ones who do, the number of collaborators is usually one and sometimes two. Instances of three or more collaborators are a rare occurrence on the charts.

Determining Significant Features

Following the initial examination of the variables, it makes sense to explore linear models to further identify the most positively impactful variables on peak chart performance and chart longevity. The first step is running a baseline linear model on all features in the dataset (only including rows without `NA` values) to extract insights about the full model. Then, this section explores forward selection for feature selection and F-tests with reduced linear models only incorporating the selected features to analyze the significance of the results. The results are then validated with principal component analysis.

Models and Feature Selection

Variables	PeakPosition	WeeksOnChart
	Coefficient	Coefficient
	(Std. Error)	(Std. Error)
Intercept	7.324e+01 (6.31e+0)	-8.383e+00 (2.91e+0)
Duration	-8.319e+06 (6.134e-6)	2.110e-05 (2.83e-6)
Danceability	-1.159e+00 (2.35e+0)	8.046e+00 (1.08e+0)
Energy	8.815e+00 (2.77e+0)	-1.295e+00 (1.27e+0)
Key	-1.027e-02 (7.69e-2)	1.366e-02 (3.54e-2)
Loudness	9.040e-02 (1.83e-1)	2.061e-01 (8.44e-2)
Mode	6.193e-03 (6.124e-01)	1.483e-01 (2.820e-01)
Speechiness	2.201e-01 (4.842e-01)	-1.260e+00 (2.23e-1)
Acousticness	3.362e-01 (2.49e-1)	-2.850e-01 (1.15e-1)
Instrumentalness	8.457e-01 (5.28e-1)	-5.013e-01 (2.43e-1)
Liveness	-1.254e+00 (4.62e-1)	-8.759e-01 (2.13e-1)
Valence	1.993e+00 (1.54e+0)	3.728e+00 (7.08e-1)
Tempo	-2.982e-03 (9.63e-3)	-5.402e-03 (4.43e-3)
TimeSignature	-1.554e+00 (1.03e+0)	8.280e-01 (4.74e-1)
OtherGenre	7.458e+00 (1.77e+0)	-1.924e+00 (8.17e-1)
NoGenre	4.744e-01 (2.49e+0)	1.575e+00 (1.15e+0)
Pop	1.872e+00 (7.68e-1)	7.247e-01 (3.54e-1)
Country	1.400e+01 (9.68e-1)	2.133e+00 (4.46e-1)
Rock	6.407e+00 (7.93e-1)	3.534e+00 (3.65e-1)
Blues	1.970e+00 (4.78e+0)	6.709e-01 (2.20e+0)
R&B	6.653e+00 (8.62e-1)	3.256e+00 (3.97e-1)
RapHipHop	3.437e+00 (7.92e-1)	-2.800e-01 (3.65e-1)
Indie	1.002e+01 (1.87e+0)	-1.709e+00 (8.61e-1)
FamousCollaborator	-3.109e-01 (1.18e+0)	1.445e-02 (5.42e-1)
Collaborators	2.094e+00 (7.49e-1)	5.380e-02 (3.45e-1)

We begin by training two baseline linear models on all features in the dataset, with one predicting `PeakPosition` and another predicting `WeeksOnChart` (only on rows with no `NA` values). The coefficients and corresponding standard errors for each variable in these multiple linear models are in the above table.

The next step is to conduct feature selection using forward selection (with default metric AIC) for both models, with the selected features and associated correlations shown in the table below.

Variables	PeakPosition Coefficient	WeeksOnChart Coefficient
Intercept	65.299	-0.109
Duration	—	0.002
Danceability	—	8.332
Energy	8.397	—
Key	—	—
Loudness	—	0.143
Mode	—	—
Speechiness	—	-1.380
Acousticness	—	-0.231
Instrumentalness	—	-0.545
Liveness	-1.213	-0.091
Valence	2.124	3.544
Tempo	—	—
TimeSignature	-1.640	—
OtherGenre	7.635	-1.166
NoGenre	—	—
Pop	1.760	0.764
Country	13.981	2.185
Rock	6.330	3.575
Blues	—	—
R&B	6.492	3.279
RapHipHop	3.369	—
Indie	10.130	-1.662
FamousCollaborator	—	—
Collaborators	1.870	—

The result of the forward selection is the two subsets of selected features with respect to peak position and total weeks on the charts. The selected variables of significance in the peak position model are Energy, Liveness, Valence, Time Signature, OtherGenre, Pop, Country, Rock, R&B, RapHipHop, Indie, and Collaborators. Meanwhile, the significant variables determined by forward selection in the weeks on chart model are Duration, Danceability, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, OtherGenre, Pop, Country, R&B, and Indie.

In the lens of the four attributes defined in the hypothesis, it is clear that the peak position outcome variable is predicted mostly by variables falling into three categories: characteristic (Energy, Valence, Time Signature) and genre (Pop, Country, Rock, R&B, RapHipHop, Indie, OtherGenre), and collaboration (Collaborators). Evidently, genre plays a crucial part in determining peak position, as the majority of the genre variables are designated as significant by forward selection. Notice that the audio features associated with song production (Loudness, Speechiness, Acousticness) were largely excluded from selection.

In considering the WeeksOnChart model, it becomes apparent that the chosen variables sort well into three categories: characteristic (Danceability, Valence), production (Loudness, Speechiness, Acousticness, Instrumentalness, Liveness), and genre (OtherGenre, Pop, Country, Rock, R&B, Indie). Noticeably, both of the variables related to Collaboration were excluded in the forward selection process. Additionally, there are much fewer variables related to characteristic in this significant set of variables, with many more production variables included.

Seeing that the two sets of chosen significant features for the peak ranking model and chart longevity model are noticeably different offers direct support for the first hypothesis that the factors determining these two outcome variables are different. Further, the subsequent hypotheses are supported by the inclusion of many genre variables for both models, primarily variables contributing to song characteristic being selected for the peak position model, and many audio features contributing to song production selected as significant in the chart longevity model.

F-Tests

To provide further evidence in favor of these results and dissuade critiques about potential inconsistencies from forward selection, I validate the significance of the selected variables with F-tests.

Using these selected features, I trained new reduced models using the complements of these feature subsets for the peak position model and weeks on chart model. Then, I conducted an F-test of the two full models against both of the reduced models. The results of the F-test are in the following table:

Outcome Variable	Degrees of Freedom	F Value	Probability > F
PeakPosition	13	34.037	< 2.2e-16
WeeksOnChart	15	32.33	< 2.2e-16

For both models, the p-value is near 0, which supports rejecting the null hypothesis. In other words, this result suggests that there is a statistically significant difference between the full model and reduced model and that the missing variables (the features selected from forward selection) contribute significantly to explaining the variance in both full models beyond what is explained in the reduced models.

Thus, these results offer evidence that the identified subsets of selected variables are significant, lending support to the established hypotheses. This statistical analysis helps disprove alternative explanations about randomness or inaccuracies within the selected sets of significant features.

Analysis of Correlation

Another step of validation to check is the correlation between each variable. To perform this analysis, I generated a correlation matrix and plotted the pair-wise correlations between selected features in a heatmap.

In observing the mostly light yellow heatmap in Figure 7, it is clear to note that the majority of correlation values between variable pairs is low. However, there two squares in the heatmap indicating high correlation that must be addressed. The most obvious pair of highly correlated features is Energy and Loudness, which have a correlation of 0.713. Additionally, two squares that are slightly darker represent RapHipHop and Speechiness which have a correlation of 0.513. All other features have low to very moderate correlation with value 0.4 or lower.

As energy overlaps with volume, it makes sense to consider the impact of removing Loudness from the linear model. Similarly, given that Speechiness is a measure of the presence of human speech in a song, the redundancy of this feature with regard to the genre feature of RapHipHop comes into question. These features only overlap in the WeeksOnChart model, so I conducted an F-test to examine if the model suffers as a result of removing the features Speechiness and Loudness from that model.

First, when conducting an F-test with a reduced model predicting WeeksOnChart removing the predictor Loudness, the resulting p-value is 0.0043. This is statistically significant, so we reject the null

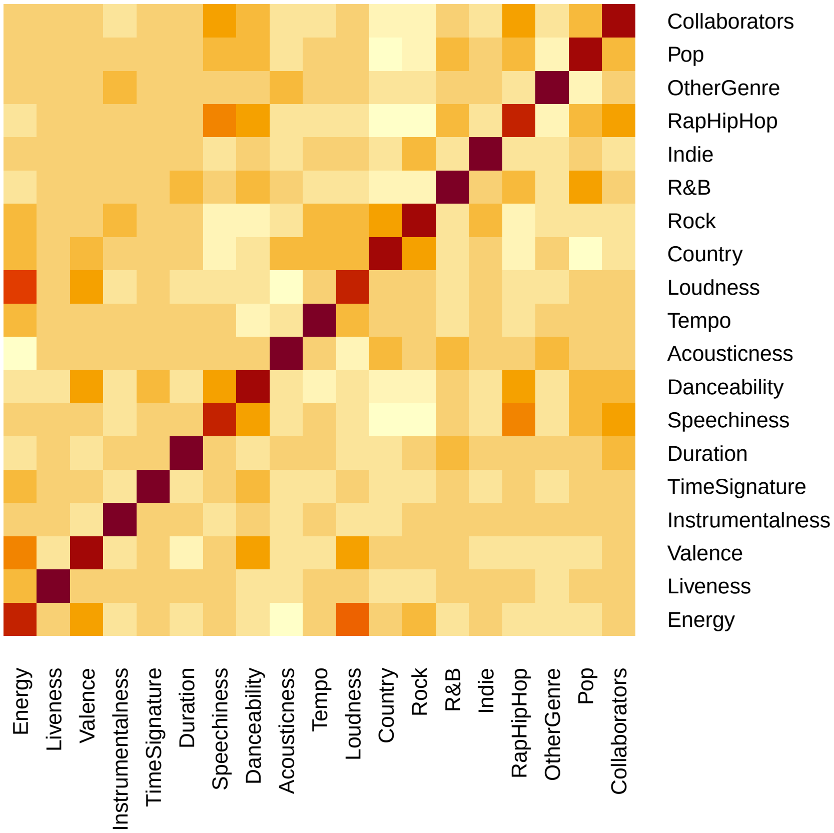


Figure 7: Heatmap for correlation between all variables.

hypothesis and note that removing Loudness from the original model significantly reduces the quality of the model fit. Similarly, I repeated this F-test procedure with a reduced model predicting WeeksOnChart with the predictor Speechiness removed. The resulting p-value is 1.502e-09, which is extremely small and indicates that removing Speechiness likewise reduces the model fit significantly.

Model	Degrees of Freedom	F Value	Probability > F
PeakPosition (no Loudness)	1	8.453	3.654e-03
WeeksOnChart (no Speechiness)	1	38.977	4.526e-10

Therefore, by the results of these F-tests, it makes sense to include Speechiness and Loudness in the model, and that these prediction variables contribute meaningfully in predicting `WeeksOnChart`. This insight and the low correlation values for all other pairwise combinations of the selected features, confers confidence that this set of features is indeed significant with low redundancy. While the collinearity between the identified variables should be recognized a limitation in the linear model, this analysis justifies the inclusion of these variables with regard to significance and offers support against alternative explanations criticizing the significance of the selected set of variables.

Principal Component Analysis

To further validate the selected set of features, I conducted a principal component analysis for each linear model (with the selected features for PeakPosition and WeeksOnChart respectively), and plot the points along the first two principal components.

First, in examining the principal component analysis for `PeakPosition` in Figure 8, it becomes clear that while the points are relatively clustered together, the higher ranked songs (from 1-50) cluster toward the left hand side of the plot while the lower ranked songs (from 51-100) cluster toward the right hand side of the plot. There is not a full separation between the two groups, which we attribute to the limitation that the first two principal components only capture 18.18% and 11.59% of the variance respectively. This indicates high dimensionality in the data which may not be able to be easily expressed on such a two-dimensional plot. Thus, while it is not possible to actively conclude that this principal component analysis validates this set of variables as a significant one for prediction of peak position,

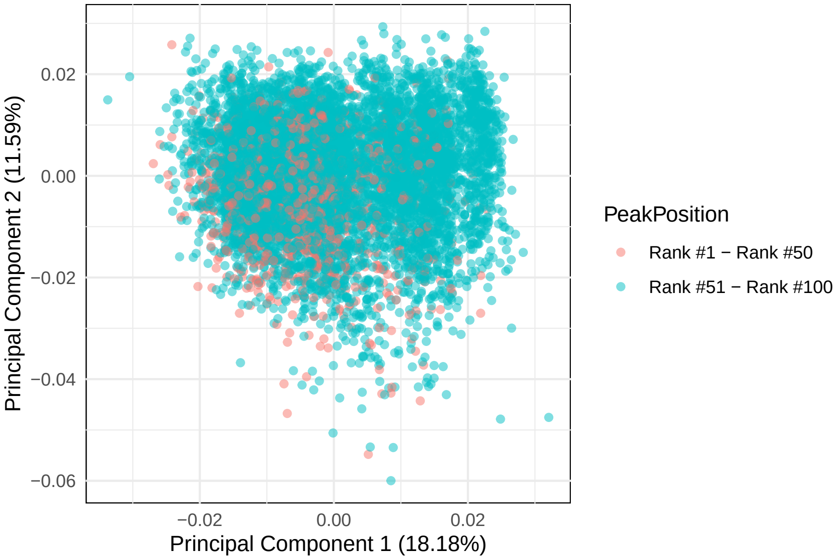


Figure 8: Principal component analysis for the PeakPosition variables.

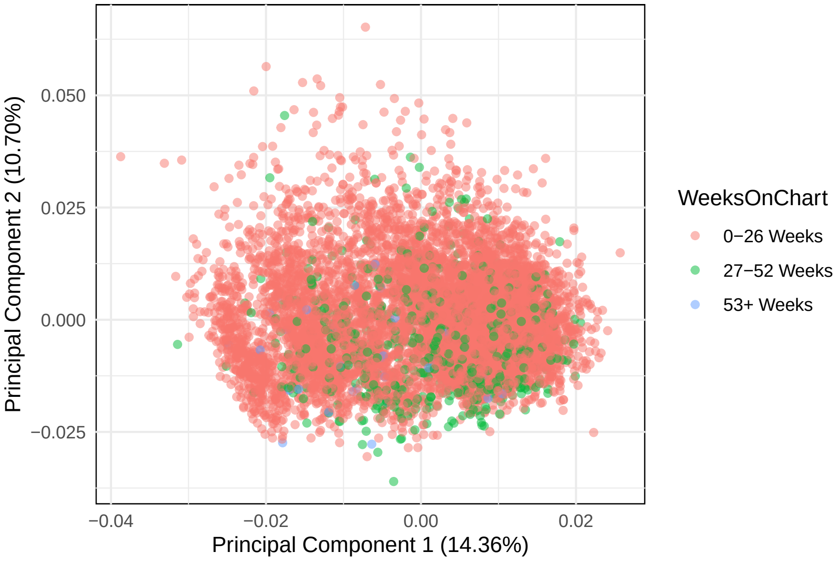


Figure 9: Principal component analysis for the WeeksOnChart variables.

the clustering we see (albeit with overlap) lends support to this set of variables.

Similarly, with the set of significant variables in determining WeeksOnChart, there is similar overlapping in points but slight clustering behavior in Figure 9. In particular, in this PCA, the songs that stayed on the charts for over half a year (green and blue) cluster toward the bottom right side of the plot while the songs with shorter stays are more toward the center and top of the plot. Here, there is a similar issue as the first two principal components only capture 14.36% and 10.7% of the variance respectively. Similarly, while acknowledging this limitation, the visual presence of some grouping in classifying a song chart longevity can be interpreted as some degree of support toward the selected set of features (even in light of the high dimensionality).

Analysis of Significant Features

With the analysis above validating the determined sets of significant variables, the next step is to investigate each such variable with multiple pieces of analysis incorporating many factors. In particular, for each such variable in the following section, some subset of **six** pieces of analysis will join to create a corroborated conclusion about the relation between that song and virality as a whole:

1. **Average value of variable by year.** Visualizing the change in average variable split by year gives insight into trends over time.

2. **Coefficient of variable in linear models predicting chart peak, separated by year.** After partitioning the dataset by year and training 20 linear models (one for each year between 2000 and 2019) based on the selected features to predict peak position, consider the plot showing the change in coefficients with regard to the particular variable by year (gauging how the magnitude and sign of the coefficients change year over year).

3. **Coefficient of variable in linear models predicting chart longevity, separated by year.** From the same procedure as in **(3)**, a similar plot is generated for weeks on chart.

4. **Distribution of songs in top quartile of the variable, by year.** Consider the distribution of songs across years in the top quartile (taken across the whole dataset) of the given variable, so the resulting plot informs whether more or fewer songs are trending toward having that variable being high as time passes.

5. **Average variable value in top quartile, by year.** Consider the average variable value of the songs in the top quartile of that variable

by year to observe changes in absolute value over time. For a given varible (ex. energy), this is to account for the scenarios where it could be possible that less and less songs have high energy, but the average value of the high energy songs is actually increasing.

As a general note for this section, keep in mind that `PeakPosition` has the highest ranking at 1 and the lowest ranking at 100, such that lower/negative coefficients are predictive of higher rankings (and not lower ones). Also, note that some features will not have all of the six pieces of evidence described above as certain models do not have certain features and for the binary genre features analysis of quartiles is less relevant (as all values are either 0 or 1).

Insights About Song Characteristic

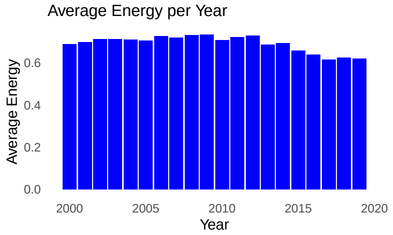


Figure 10: Average energy levels for songs by year.

Energy

A song's energy score represents its level of activity and intensity, heavily contributing to the character of the song. Although there have been fluctuations within the two-decade period, multiple pieces of evidence suggest that the most popular music is trending toward energetic songs.

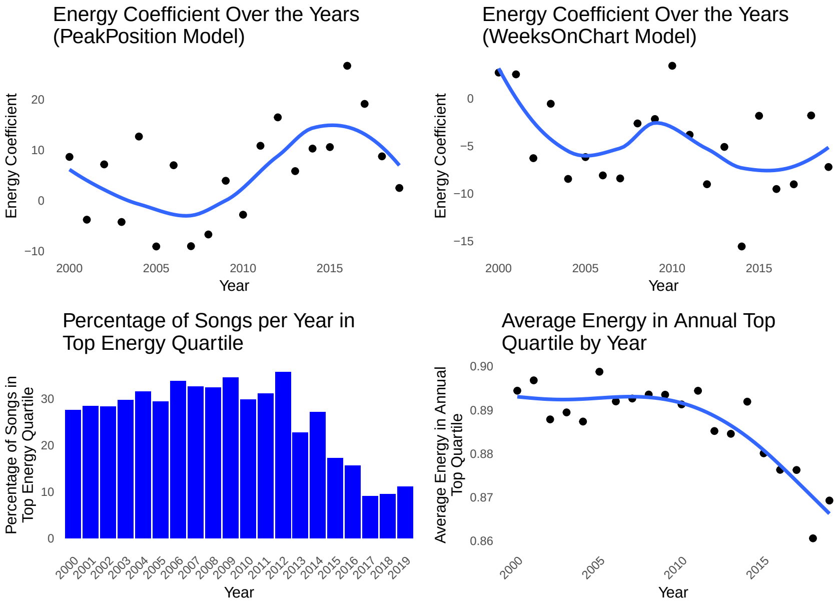


Figure 11: Distribution and average energy per year only for songs in the top energy quartile across the two-decade dataset.

In examining the average energy levels for songs by year in Figure 10, it is clear to see that although the average energy level has been on a general decline since 2011, the average energy is increasing year over year from 2018 onwards. Further, since energy was selected as a significant feature for both linear models, the change in coefficients for the energy variable yields insights about the predictive contribution of energy scores for both peak position and chart longevity. As seen in Figure 11, from 2016 onward, the coefficients for energy trend downward in predicting peak position and upward in predicting chart longevity. Thus, as time approaches 2020 higher energy scores are predictive of both higher peaks and longer stays on the charts. Additionally, the bottom two graphs show that after hitting a low in 2017, the percentage of high energy songs is increasing while the highest energy songs are decreasing in absolute value.

These plots corroborate the conclusion that although energy scores dropped going into the mid-2010s, in recent years the energy level of charting songs is not only increasing but also positively predictive of chart performance with regard to both peak and longevity. The percentage of high energy songs is also increasing, even though the value of those high energy songs is being tempered.

Valence

Another important factor in determining the character of a song is its valence, a measure of musical positiveness. Overall, valence is becoming a less critical factor in achieving chart performance in the modern era, with downward trends. To see this, examine Figure 12 to see that the average valence level of charting songs is decreasing steadily year over year.

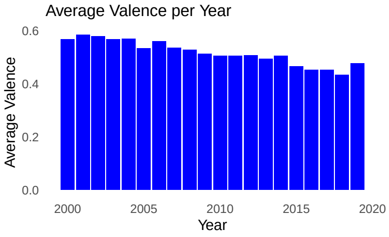


Figure 12: Average valence levels for songs by year.

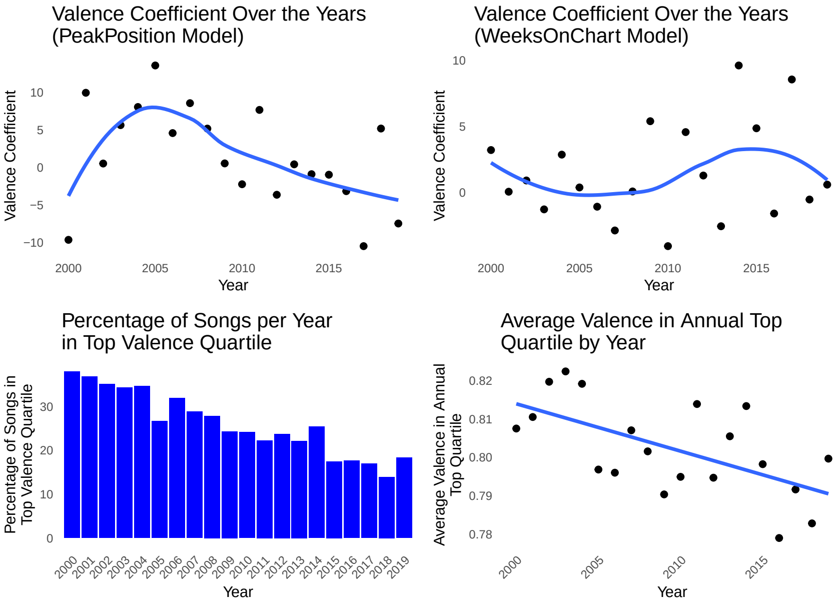


Figure 13: Distribution and average valence per year only for songs in the top valence quartile across the two-decade dataset.

Noting that valence is a significant variable in the linear models for peak position and weeks on chart, it makes sense to consider the coefficients in both models. While there is a slight downward trend in the coefficients for both models in Figure 13, there is a large amount of fluctuation in recent years. This oscillation between positive and negative coefficients indicates a lack of a clear trend approaching the present day with regard to valence. However, there is a clear downward trend over time with regard to the percentage of charting songs by year with high valence. Additionally, not only are high valence songs becoming less common, but the songs with high valence are having lower valence values as well. In other words, while the predictive power of valence with respect to rank and longevity is unclear in recent years, the songs on the charts are moving towards lower valence as a whole.

Danceability

A defining audio feature in upbeat music is danceability, a critical feature selected as significant in determining chart longevity. While there was a clear dip in popularity for danceable sogns near 2010, multiple pieces of analysis show that highly danceable songs are becoming trendy. First, this can be seen in Figure 14, where the average danceability score of charting songs is clearly increasing throughout the 2010s.

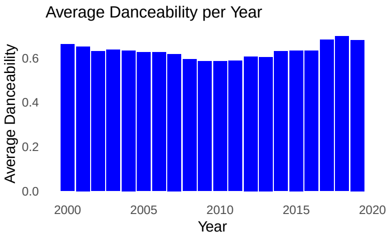


Figure 14: Average danceability levels for songs by year.

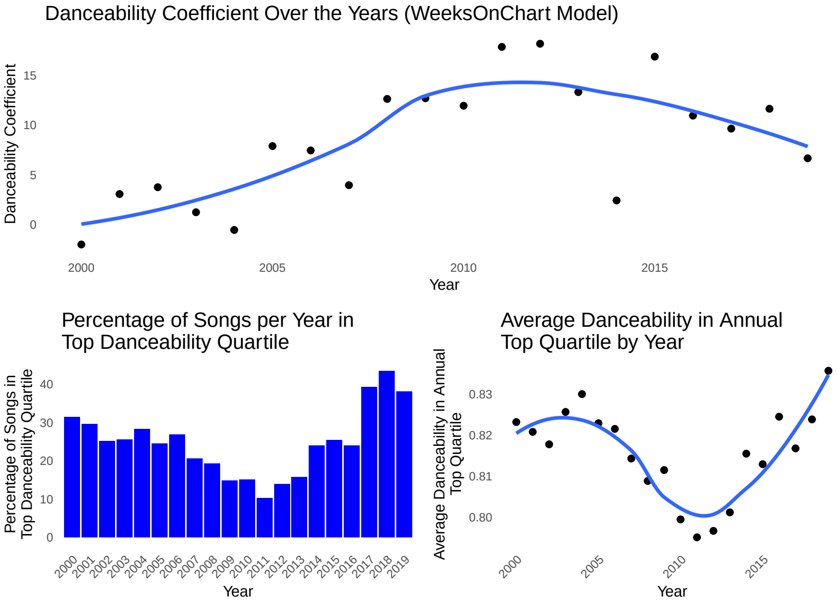


Figure 15: Distribution of average danceability per year only for songs in the top danceability quartile across the two-decade dataset.

Furthermore, in Figure 15, there is a clear increase throughout the

2010s of not only the percentage of highly danceable songs that make the charts, but also the danceability values of those songs as well. However, it is critical to notice that the dowanward trend with regard to the coefficient for danceability in predicting the number of weeks on the chart. This observation indicates that although highly danceable songs are becoming increasingly popular (as seen by their rising frequency on the charts), their longevity is becoming increasingly short-lived in the nature of social media trends and dance fads.

Tempo

Tempo is also a critical audio feature in determining the feel of a song, especially with regard to whether it is a slow or fast song. As a whole, the average tempo of songs in the past two decades has remained relatively static, as seen in Figure 16.

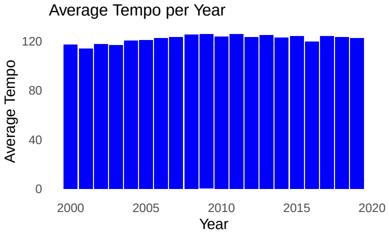


Figure 16: Average tempo for songs by year.

The coefficient for tempo in the linear model predicting chart longevity supports this, with the graph in Figure 17 hovering around 0, fluctuating slightly year by year. The plots for songs in the top tempo quartile add another dimension to the story, showing that in recent years, songs with high tempo are becoming more common while the average tempo is slightly decreasing to roughly 160. Note that the changes in tempo are very small with regard to average values for the overall dataset and the high tempo songs.

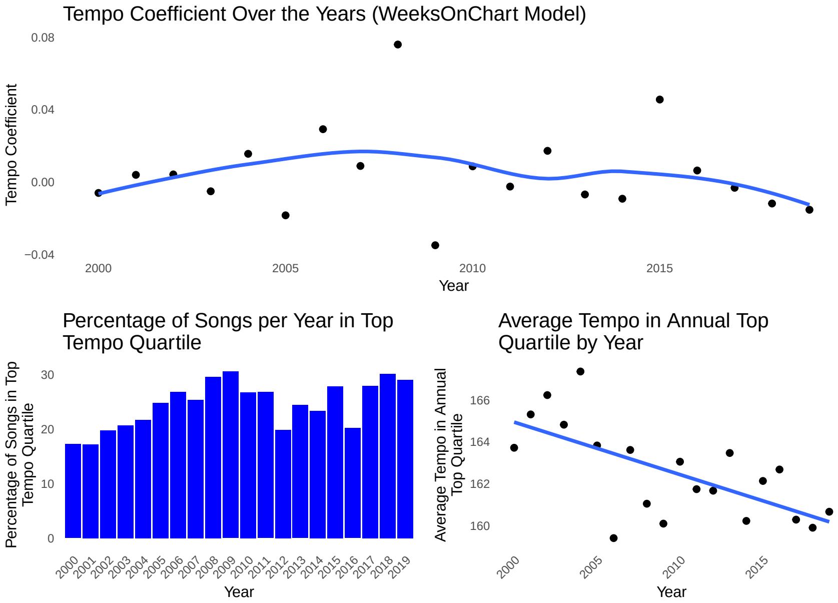


Figure 17: Distribution and average tempo per year only for songs in the top energy quartile across the two-decade dataset.

Given the very small magnitude of change for the tempo values over time along with the relatively flat trend in the coefficient graph, these insights combine to suggest that tempo is a comparatively inelastic trait over time. The charting songs all have a tempo within a relatively narrow range (close to 120) with very minor changes over time.

Duration

The duration of a song is critical with regard to popularity, as the length of a song can easily dictate its playability (ex. on the radio). Overall, for all songs in this two decade period, the clear trend is that popular songs that reach the charts are becoming shorter. This is most evident in Figure 18, where the average song length steadily decreases.

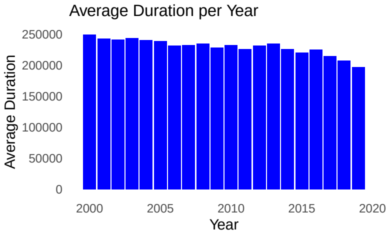


Figure 18: Average duration of songs by year.

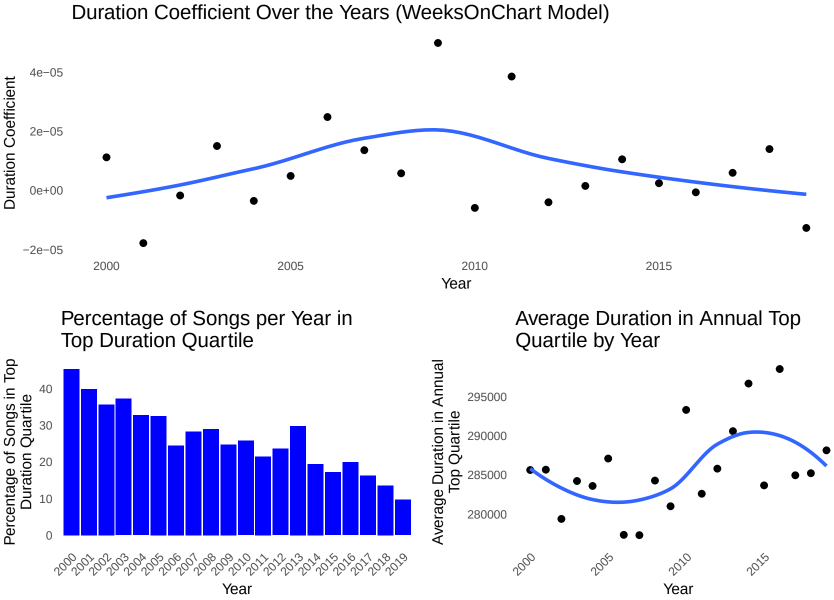


Figure 19: Distribution and average duration per year only for songs in the top duration quartile across the two-decade dataset.

When looking at the longest songs in Figure 19, a similar trend appears to support this insight of popular songs with shorter duration. In particular, very long songs are becoming increasingly uncommon, as seen by the very obvious downward trend from 2000 to 2019. This is further corroborated by examining the average duration of these long songs, which is also shown to be decreasing in recent years with a slight downward trend.

These results are bolstered by the plot illustrating the change in coefficients, where the downward trend starting in 2009 indicates that songs with longer duration become increasingly predicted to have shorter stays on the charts. Therefore, in examining the average and distribution of not only the duration values but also the duration coefficient in the chart longevity model, these pieces of analysis corroborate a trend toward success for shorter songs with regard to making it onto the charts and staying there.

Insights About Song Production

Instrumentalness

The first variable of interest with regard to song production is instrumentalness, which measures the reliance of the song on instrumentals in the background. Examining the change in average instrumentalness score per year in Figure 20 tells us that there is very little change over time, but with a very slight trend toward less instrumentalness in recent years.

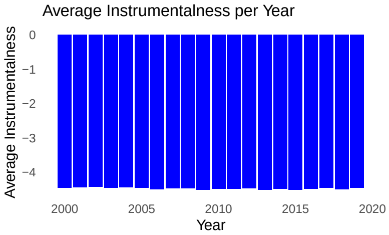


Figure 20: Average log instrumentalness levels for songs by year.

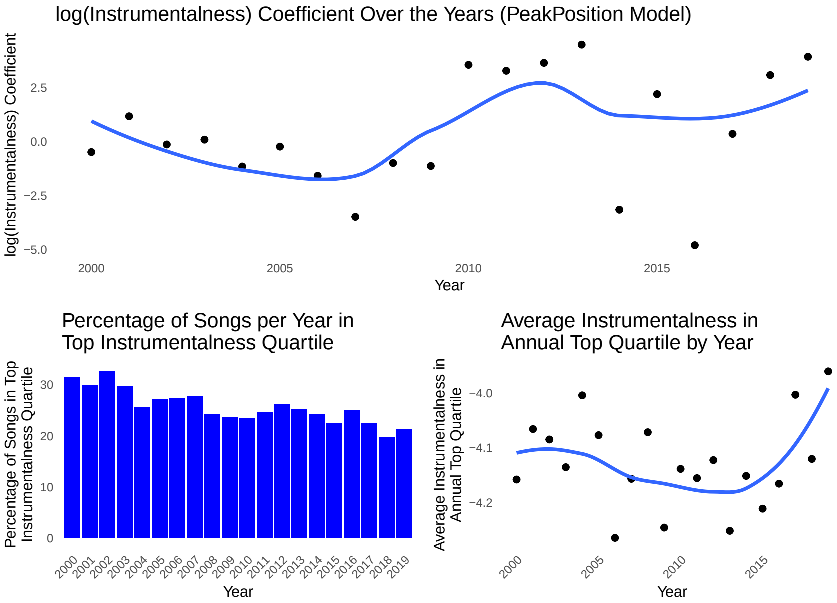


Figure 21: Distribution and average log instrumentalness per year only for songs in the top instrumentalness quartile across the two-decade dataset.

The coefficient for instrumentalness (with log transform) supports this inverse relationship between instrumentalness and chart success, with instrumentalness becoming increasingly predictive of lower chart peaks as seen in Figure 21. The other two plots contribute an interesting insight: over time, while the percentage of highly instrumental songs is decreasing, those highly instrumental songs are increasing in instrumentalness. In other words, very instrumental songs are becoming less common, but more extreme. As a whole, these insights support a trend of popular songs becoming less instrumental over time, with the note that the very instrumental songs that do make it onto the charts are highly specialized (and perhaps fully instrumental).

Acousticness

Another critical audio feature that informs the production style of a song is its level of acousticness as a measure of the involvement of electrical amplification and related effects. Seeing that the average acousticness score (with log transform) has steadily increased year-over-year since 2012 (Figure 22), songs that make it onto the Hot 100 charts on average trend toward acousticness.

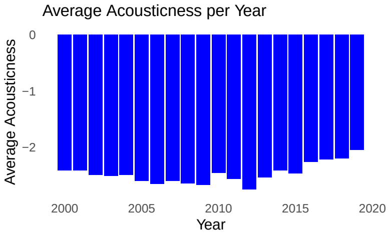


Figure 22: Average log acousticness levels for songs by year.

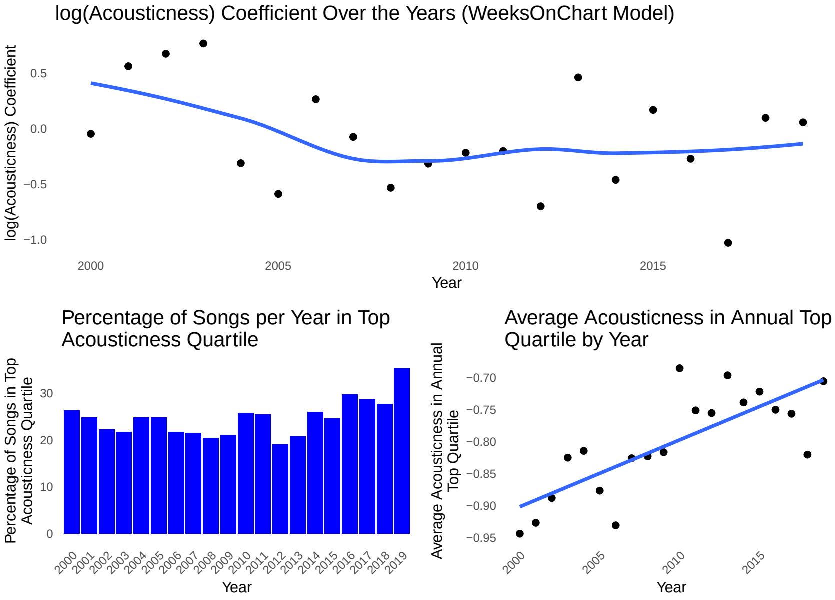


Figure 23: Distribution and average log acousticness per year only for songs in the top acousticness quartile across the two-decade dataset.

This conclusion is echoed by the plots in Figure 23, as not only is the percentage of songs with high levels of acousticness increasing with time, but these highly acoustic songs reaching greater degrees of acousticness as well. The coefficient for acousticness in the model predicting chart longevity fluctuates near 0 with a relatively flat trend line sloping slightly upwards. As a whole, these pieces of evidence all suggest that the popular songs reaching the charts are trending toward higher levels of acousticness.

Loudness

The loudness of a song is a critical factor in its production and final sound, making it a variable of interest. Examining the change in average loudness values per year in Figure 24 shows that, on average, charting songs have been becoming louder from 2009 to 2018 with a slight decrease following. This suggests a positive relationship between loudness and song popularity, which is supported by the analysis of the loudness coefficient in predicting chart longevity.

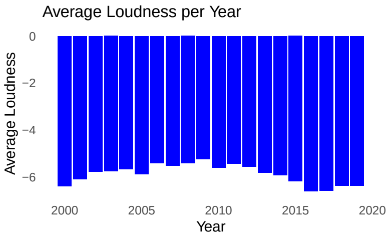


Figure 24: Average loudness levels for songs by year.

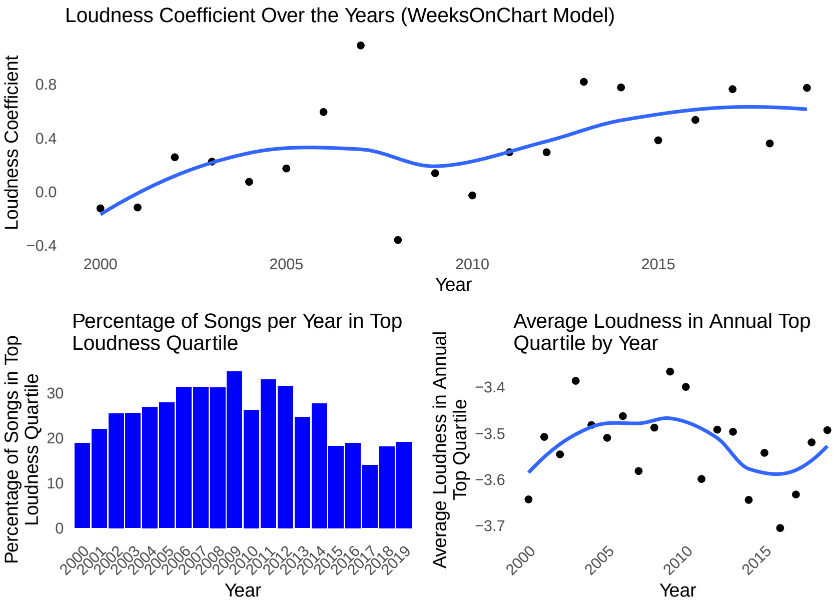


Figure 25: Distribution and average loudness per year only for songs in the top loudness quartile across the two-decade dataset.

In particular, as seen in Figure 25, there is a clear increasing trend demonstrating that loudness becomes increasingly predictive of longer stays on the charts with time. However, the bottom two charts both show a decrease from 2009 until an increase in 2017. This observation implies very loud songs becoming increasingly rare and overall less loud since 2009, with a reversal since 2017. In tandem, these insights suggest a decrease in the popularity of very loud songs throughout the 2010s, but a trend toward loud songs making the charts, staying on the charts for longer, and becoming even louder since 2017.

Liveness

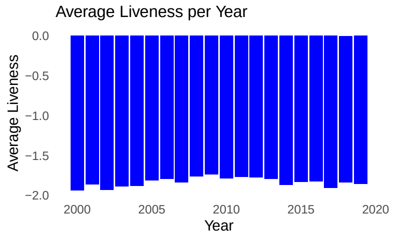


Figure 26: Average log liveness levels for songs by year.

Especially in the modern era of music production, songs can range widely from heavily produced tracks to a cappella live recordings. Consequently, liveness is an important feature in distinguishing the nature of production of each song. Notably, the average liveness levels by year are relatively constant in Figure 26, which is tied to the narrow distribution of liveness values because most songs are not live recordings.

While there is relatively little variation in the average liveness levels over time, Figure 27 shows clear trends in the coefficients for liveness in predicting both peak chart position and longevity. There is a clear downward trend in predicting peak position since 2010, and an equally prominent upward trend in predicting weeks on chart since 2014. These trends suggest that songs with high liveness values are predictive of both higher chart peaks and longer stays on the charts in the past decade.

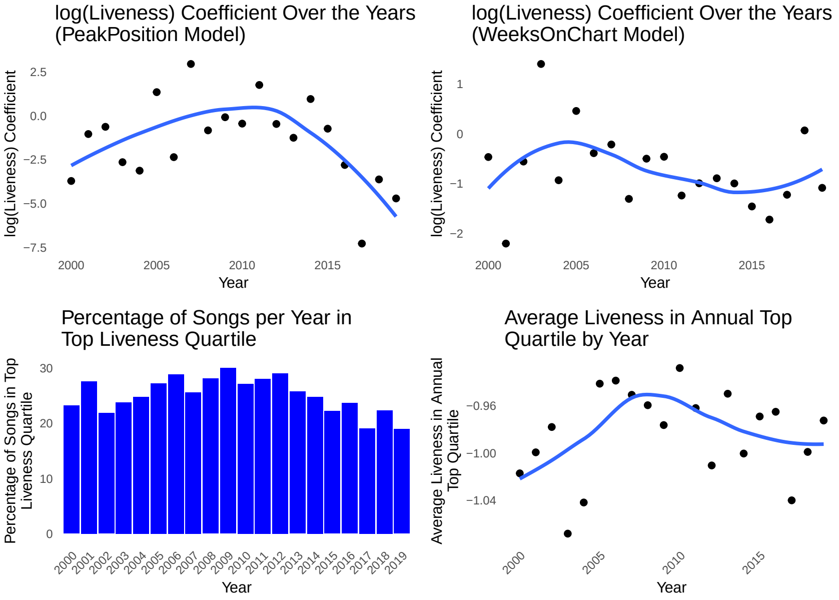


Figure 27: Distribution and average log liveness per year only for songs in the top liveness quartile across the two-decade dataset.

At the same time, observe the downward trend in both the percentage of very live songs and their liveness values since 2009 in the lower half of the figure. These plots indicate convergence toward a lower liveness value, as songs with high liveness values become not only less common, but also less live. When combined with the above insights about predictive power, these trends show that production with liveness is conducive to better chart performance but not when taken to extremes.

Speechiness

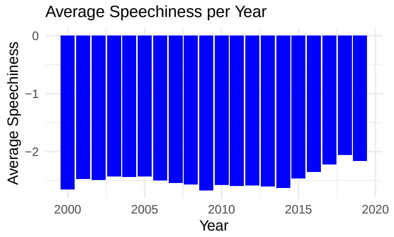


Figure 28: Average log speechiness for songs by year.

The final production feature to consider is speechiness, measuring the presence of spoken words in a track (ex. rap, ad libs, monologues, etc.). Taking a look at the change in average speechiness by year in Figure 28 yields a clear trend toward higher average speechiness scores over time. This insight suggests that songs that make it onto the charts have increasingly high speechiness levels. The lower plots in Figure 29 serve to corroborate the connection between higher speechiness and chart performance, as there are clear trends from 2011 onwards with regard to very speechy songs becoming more common and increasingly speechy with time.

However, the coefficient plot shows an increase in the predictive power of speechiness toward chart longevity in the early 2000s that evens out and fluctuates in more recent years. Combining these insights informs a positive association between high speechability scores and making onto the charts, although this variable is limited in its predictive power with regard to chart longevity.

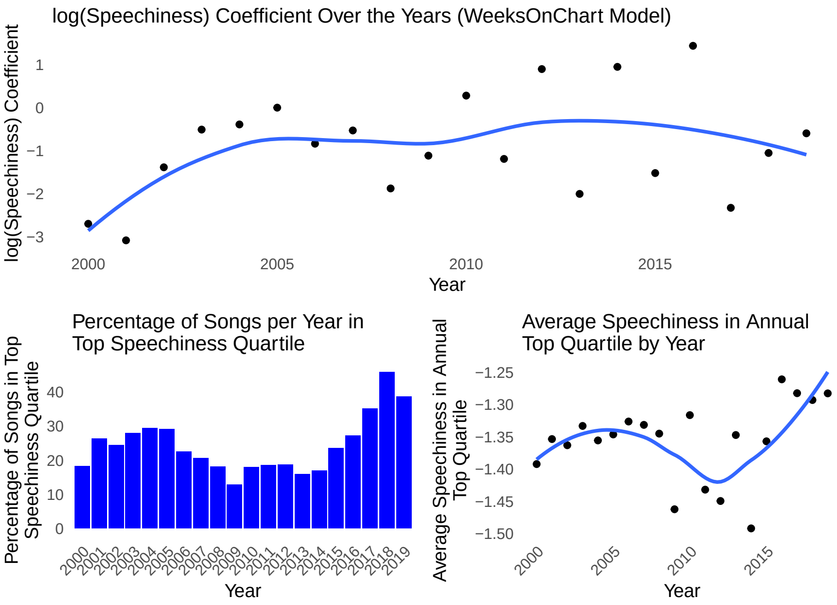


Figure 29: Distribution and average speechiness per year only for songs in the top speechiness quartile across the two-decade dataset.

Insights About Genre

Pop

The type of music most commonly associated with the top charts is pop music, making this genre a critical one for consideration. As seen in Figure 30, for every year within the two decade span, a majority of the songs on the charts belong to the genre. It is particularly interesting to note the at the percentage of pop songs by year shows an increase from 2000 to 2011, with a decrease in frequency approaching recent years. This trend indicates a shift away from the dominance of pop music on the charts.

Furthermore, noting that the pop variable was selected as a significant predictor in the models for both peak position and chart longevity, it is important to consider the trend in coefficients. As seen in the bottom two plots in Figure 30, over time, pop has become increasingly predictive of lower chart peaks and shorter stays on the charts. In putting these observations together, it becomes clear that while pop music remains widely popular in encompassing a majority of songs year over year, adhering to the pop genre has become less critical in recent years to make a chart-topping hit.

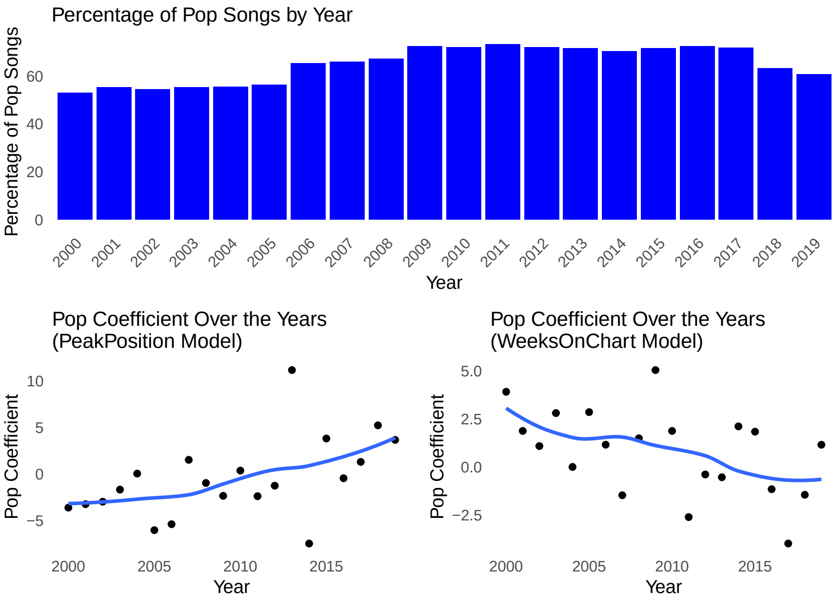


Figure 30: Change in coefficient for pop variable in the linear models trained to predict peak position and total weeks on the charts.

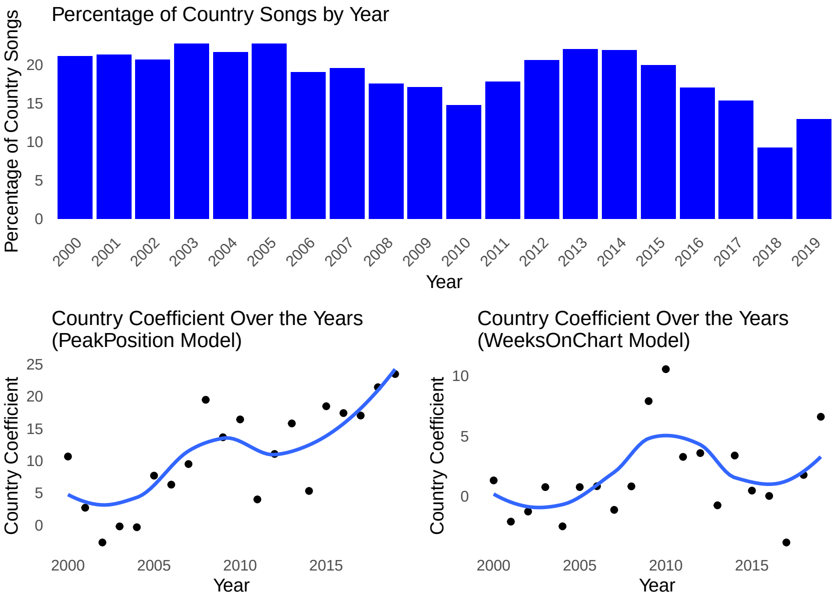


Figure 31: Change in the coefficient for the country variable in the linear models predicting peak position and total weeks on the charts.

Country

Alongside pop, country is one of the most well known genres encompassing niche country tunes as well as country-pop hits. With the genre wedged between fringe and mainstream music, it is interesting to examine country songs with relation to peak chart performance.

Examining the annual change in the percentage of country songs on the charts in Figure 31 shows a somewhat sinusoidal pattern with a recent peak in 2013, followed by a dip until a resurgence in 2018. Note that the overall percentages are lower than pop but nontrivial, between 10% and 20%. Further, the bottom two scatterplots showing the trend in coefficients for both predictive models highlights that the country genre has become increasingly predictive of lower chart peaks but longer stays on the charts in recent years. These insights suggest that country songs are on a slight comeback with regard to both frequency and longevity on the charts, but still generally find difficulty in creating hits at the very top of the charts.

Rock

Another classic and easily recognizable genre is rock, which dominated much of the middle twentieth century. However, the rock genre has faded in popularity since its peak, and the data within this twenty year period shows this. Since 2005, the percentage of songs on the Hot 100 belonging to the rock genre has steadily decreased to 5% since 2018 (Figure 32).

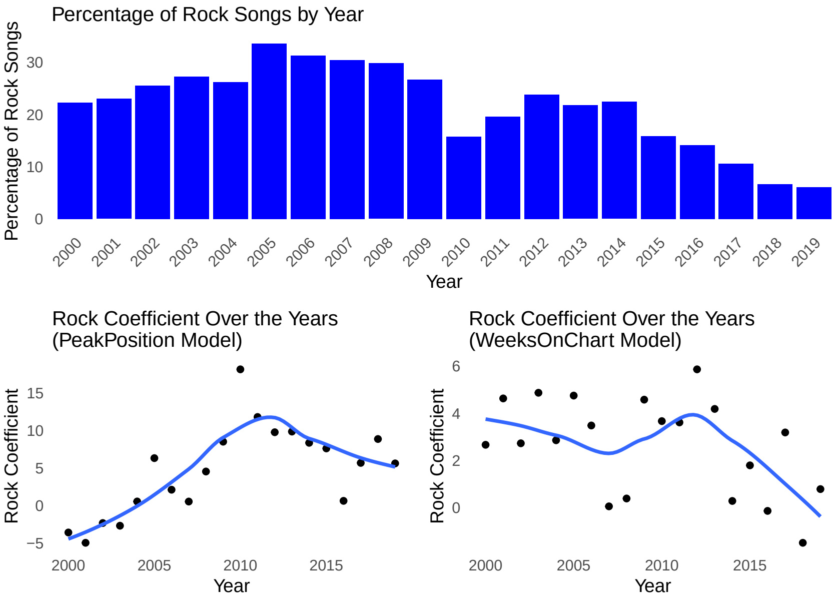


Figure 32: Change in coefficient for rock variable in the linear models trained to predict peak position and total weeks on chart.

Additionally, examining the right coefficient plot reveals that being in the rock genre has become increasingly predictive of short stays on the charts. Meanwhile, the rock variable is negatively predictive of peak position due to the positive coefficients, though to a lesser degree in recent years with a falling trend. Taken together, these insights suggest that rock is becoming increasingly unpopular on the top charts in the modern age, with poor performance in multiple chart metrics.

R&B

The R&B (rhythm and blues) genre has a rich culture with an array of historical hits. However, like rock, this genre of music is on the decline with regard to topping the charts. To see this, note that the percentage of R&B songs that make it onto the charts is significantly decreasing year over year, under 10% since 2017 (Figure 33). The coefficient plots show that creating a song in the R&B genre has become predictive of lower ranks on the charts, with a relatively flat trend in terms of chart longevity.

As a whole, the R&B genre is becoming less popular over time, with the models showing that this genre predictor indicates subpar performance with regard to both absolute rank as well as ability to remain a mainstay on the charts.

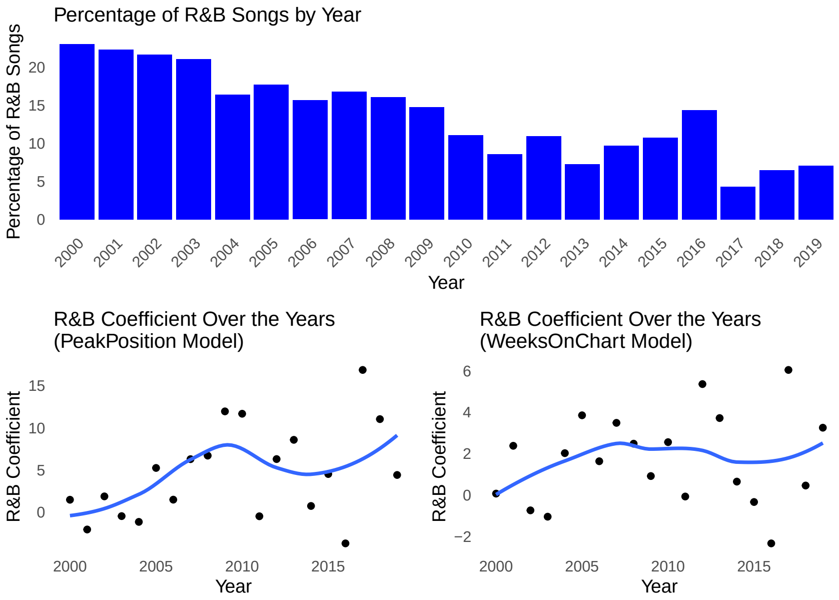


Figure 33: Change in coefficient for R&B variable in the linear models trained to predict peak position and total weeks on charts.

Indie

The indie genre captures a broad variety of songs associated with independent production and a rejection of popular conventions. Consequently, indie music can be portrayed as diametrically opposite to pop music, which is a sentiment echoed by the data. The easiest way to see this is to consider the percentage of indie songs on the charts per year in Figure 34. For each year within the past two decades, the percentage of indie songs reaching the Hot 100 charts has been less than 5%, and even when taking into account the small fluctuations over time, this percentage is decreasing from 2015.

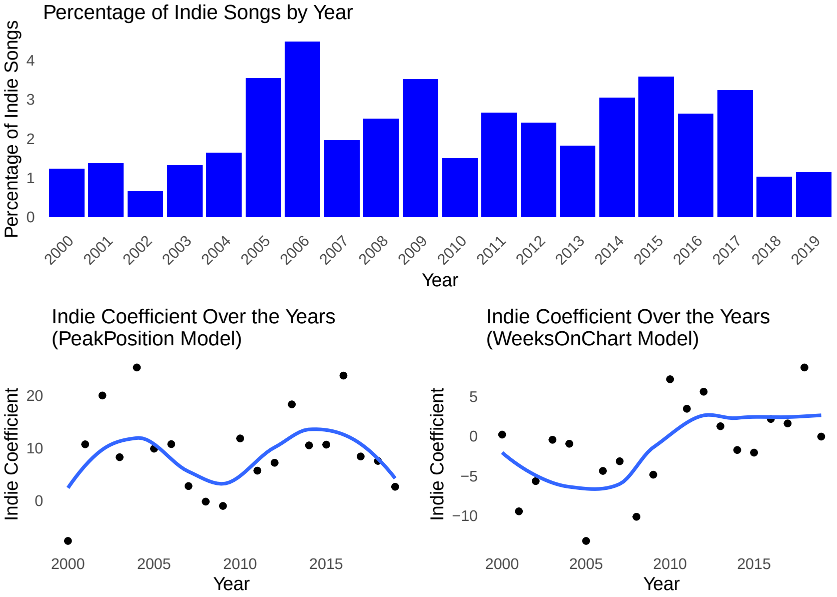


Figure 34: Change in coefficient for indie variable in the linear models trained to predict peak position and total weeks on chart.

The coefficient plots further show the low popularity of the indie genre. The positive values of the coefficient predicting peak position highlights a negative relationship between indie and rank, though the downward dip in recent years indicates that this effect is mellowing. Meanwhile, the coefficient plot predicting weeks on chart is relatively flat in recent years, with positive numbers near zero. These observations show that the indie genre has become unpopular in recent years, though with better chart performance prospects than rock with regard to both ranking and time.

RapHipHop

The final main genre of analysis is rap and hip-hop, a style that grew in popularity in the late twentieth century. The data shows that this trend has continued, with the percentage of rap and hip-hop songs on the charts increasing year over year throughout the 2010s (Figure 35). Moreover, with the rap/hip-hop variable selected as a significant one in both models, the coefficient plots yield significant insights. In particular, for both models, there are clear trends indicating how creating songs in the hip-hop and rap genre are increasingly predictive of higher chart peaks with shorter stays on the charts.

These conclusions paint an interesting picture in showing how songs of this genre are able to successfully climb the charts with percentages competitive with pop, but last much shorter on the charts. This behavior is indicative of viral trends which are able to quickly surge in gaining popularity, but quickly die out.

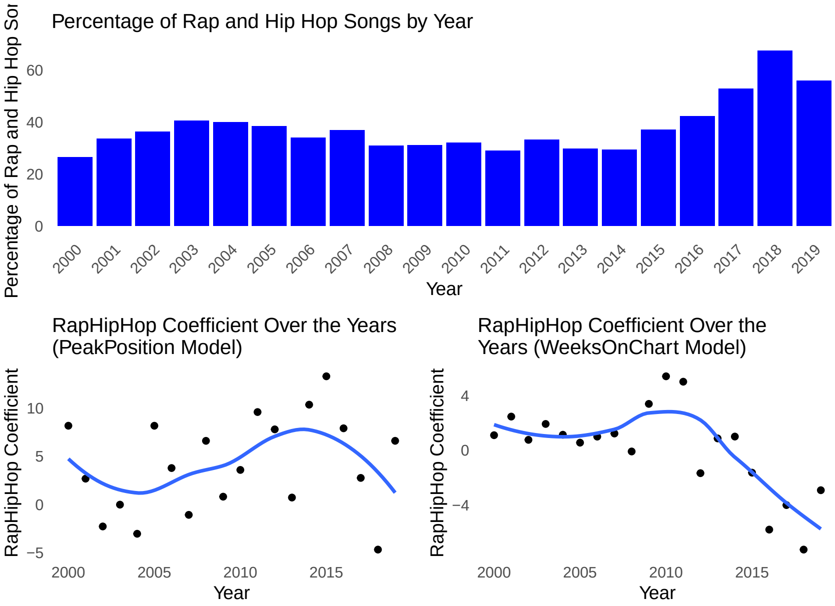


Figure 35: Change in coefficient for rap/hip-hop variable in the linear models trained to predict peak position and total weeks on chart.

OtherGenre

It is also important to consider the OtherGenre variable which serves to group together all songs without any of the other established genre labels seen above. Intuitively, these will be lesser-known and more obscure songs, which is a characterization that the data supports. Specifically, in Figure 36, there is clearly a decrease in the percentage of OtherGenre songs from 2000 to 2019. Furthermore, the coefficient plots show that songs not in any established genre are more predictive of lower chart peaks and shorter stays on the charts. However, it should be noted that the charts coefficient is trending closer to 0 in recent years. In tandem, these observations show that songs not in the defined genres are becoming less frequent on the charts with lower peaks in recent years. However, there is a trend toward such genre-agnostic songs no longer becoming negatively predictive of chart longevity, which points to a paradigm shift.

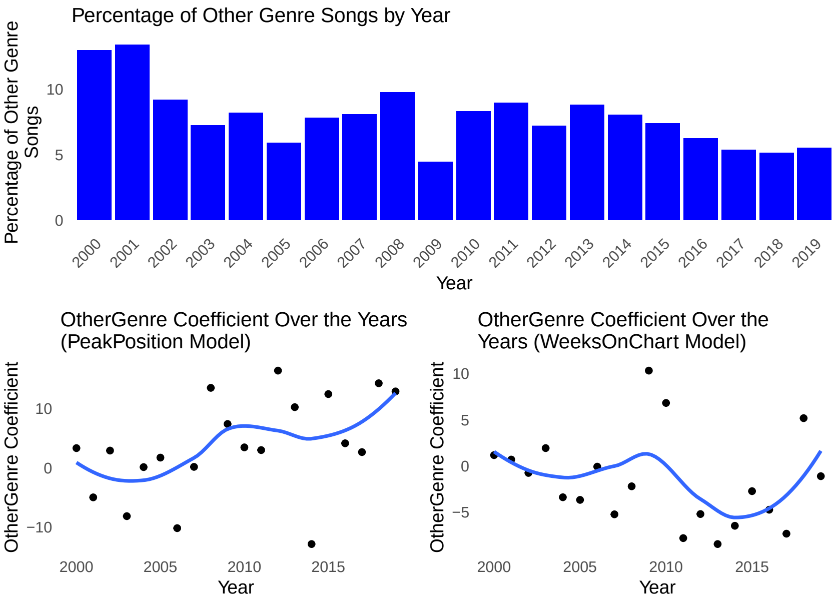


Figure 36: Change in coefficient for OtherGenre in the linear models trained to predict peak position and weeks on the chart.

The percentage of other genre songs per year is highest in the early 2000s, and has relatively steadily decreased since then. In more recent years, roughly 5-10% of songs are not categorized by any of the other existing labels.

Insights About Collaboration

The last major song feature to investigate is collaboration, as collaboration between artists is a common phenomenon—especially for charting songs. To see this, note that the average number of collaborators on songs that make it to the Hot 100 is increasing in a positive trend throughout the 20 year period (Figure 37). Moreover, this observation is corroborated by the insight that the percentage of songs with at least one collaborator is also steadily increasing with time. However, the coefficient plot in predicting peak ranking shows that more collaborators is not necessarily predictive of higher chart peaks.

Clearly, collaboration between artists is a popular phenomenon between artists and becoming increasingly common year by year. While the presence of collaboration is positively associated with the ability to reach the charts, it is not shown to predict better peak performance.

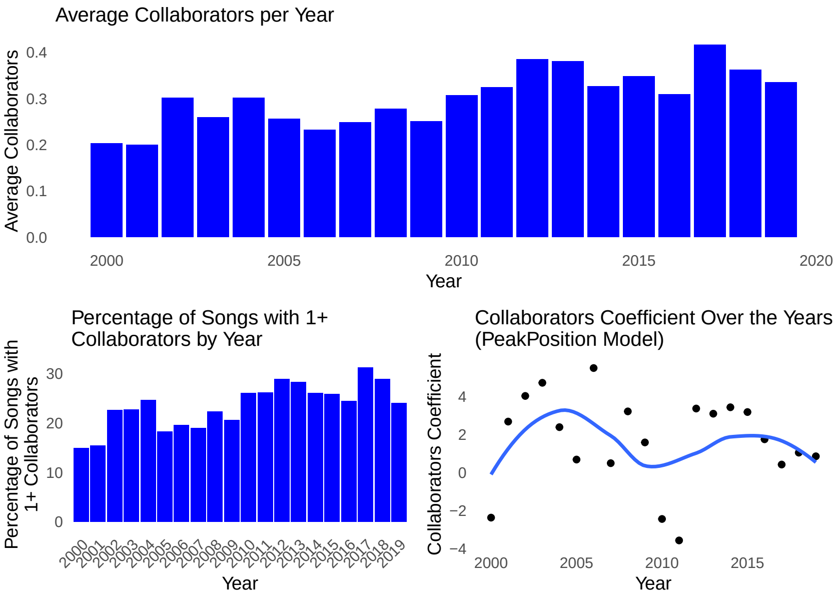


Figure 37: Change in coefficient for number of collaborators in the linear model trained to predict peak position.

Results

After conducting this analysis of my data with regard to feature selection, statistical tests, and visual exploration, a number of insights have become clear which address my opening hypotheses in turn.

1. The features and patterns that impact longevity on the charts (WeeksOnChart) are totally different from those that determine peak ranking on the charts (PeakPosition). This difference is illustrated by the different sets of significant features selected to predict these two outcome variables. Since different features were chosen (with significance validated by F-tests and principal component analysis) for peak position and chart longevity, the factors most important in predicting each must be separate. Further, this distinction can be seen within many of the significant features analyzed. For example, the variable representing the rap/hip-hop genre was predictive of both higher chart peaks and shorter stays on the charts. Therefore, while these two outcome variables are both important to chart performance, they are impacted differently by different song factors.
2. Audio features pertaining to the characteristic of a song are crucial in determining chart success. Specifically, feature selection and the associated validation shows energy, valence, danceability, tempo, and song duration to be significant features in predicting chart performance. Further, as seen by the analysis of significant features and coefficients from the linear models, in recent years,

successful songs are becoming increasingly energetic, short, and danceable. Songs exhibiting these characteristics are more likely to make it onto the charts and rise to the top, but struggle to remain on the charts for a very long time—a typical phenomenon associated with virality.

3. The production of a song and its associated features play a large part in determining success entering and climbing the charts. The development of linear models and feature selection shows that the instrumentalness, acousticness, loudness, liveness, and speechiness of a song are important in predicting chart peak and longevity. In general, chart peaks are shown to have a positive association with liveness and acousticness, and a negative association with instrumentalness and loudness. Moreover, successful songs with regard to both chart peak and longevity are becoming less extreme in their feature values. In other words, these features are converging to a medium and thus moving away from extreme and unique production styles.
4. Genre is seen to play a major role in determining chart performance, with many genre features being selected as significant predictors in the linear models for both chart peak and longevity. As a whole, pop, country, and rap/hip-hop are most indicative of success in entering the charts, while other genres such as indie and rock are much less promising. Songs in the pop and rap/hip-hop genres are trending a viral nature with high peaks but short chart stays. Furthermore, while the charts remain dominated by pop and rap/hip-hop, these genres are becoming less dominant with regard to frequency on the charts—indicating a trend toward more diverse representation of styles and genres on the Hot 100.
5. Overall, collaboration was found to not be as significant of a factor in creating a charting song as initially hypothesized. For the two collaboration variables, only one (number of collaborators) was selected for one model (peak chart position) in the linear model. In turn, this lack of inclusion in the model suggests the collaboration factors are not significantly predictive of chart performance in comparison to the other features. However, it should be noted that the average number of collaborators and percentage of songs with a collaborator are both increasing over time in recent years. Overall, collaboration seems to contribute to increasing the chances of getting a song onto the charts, but its role as a factor is not strongly predictive of success once the song has reached the charts.

Corroborating Resources

These results are corroborated by Viner's 2020 analysis [14] of the im-

[14] Josh Viner (2020) *What Makes a Hit Song: Analyzing Data from the Billboard Hot 100 Chart*, Medium

pact of audio features on Billboard chart data on a smaller timeframe (2010 to 2020) and with relatively elementary analysis (simple averaging of audio features). Although we pursue different methodologies, these respective analyses identify qualities relating to the energetic character and production of a song to be most critical with regard to its success. Furthermore, the aforementioned article by Pham et al.[15] similarly supports the findings above, echoing the conclusion that the most important features in determining chart performance lie in audio features.

In addition, another relevant source to compare results regarding my hypothesis is from the Spotify Newsroom[16] with regard to the impact of collaboration with famous artists on the success of songs. This article is published from data analysts at Spotify, and analyzes a number of top hits throughout the desired time period with further context about each individual song. As seen by the plots and trends demonstrated for the selected significant features, the insights found by Spotify closely mirror those in my analysis. These results add another dimension of nuance to the impact of collaboration on the success of top songs, and help to prove my hypothesis more convincingly.

Conclusion

Limitations

While the results found have been corroborated by multiple sources with in-depth analysis taking multiple variables into consideration, this analysis is inherently limited by three main factors.

1. Although the source of this combined dataset is Spotify and Billboard, large and reputable companies, there are still many other avenues through which people listen to music. Relying on data from Spotify and Billboard means that music more streamed through other platforms (ex. Apple Music, Soundcloud) is not considered in the analysis. In particular, there may be a bias toward established artists, as niche/indie music websites (ex. Soundcloud) or music by smaller artists (ex. covers on YouTube) are likely to be undercounted by Spotify/Billboard.
2. While the dataset I analyzed includes a more recent time period (2000-2019) than other comparable studies, my analysis is limited to music before 2019. Thus, though it is likely the trends found in my analysis are likely to continue in the present day, there is not a strict guarantee that this analysis can be effectively extrapolated to draw conclusions about the music after 2020 without further work on recent years.

[15] Pham et al. (2015) *Predicting Song Popularity*, Stanford Department of Computer Science

[16] Spotify For The Record (2023) *The Crossover Effect: Artist Collaborations Thrive on Spotify*, Spotify

3. In using audio features that are unilaterally determined by Spotify, these scores are subjective in favor of Spotify's aims and at risk of being biased. This is particularly true for scores for subjective features like energy or danceability as compared to objectively determinable features such as tempo and song duration. In particular, this is a notable concern because the principal component analysis conducted shows that while there is a degree of clustering, there is some overlap that can point to redundancy or other flaws inherent in the features.

Ethical Considerations

Noting the limitations of this analysis, we consider the critical ethical issues that may stem as a result.

1. In particular, relying on Spotify and Billboard for analysis may cause representative harms in creating a recipe for popular music. Given that Spotify and Billboard are companies with operations centered in America and mostly working in English, it is likely that these platforms are more accessible in North America than in other places worldwide. Consequently, the scores and rankings found in the dataset are likely to be more representative of the tastes and preferences of urban Western listeners who have better access to Spotify. This means that the results of this analysis is likely biased toward English songs which are popular in America, suggesting that artists emulate this style of music to achieve success on the charts. This may be harmful with regard to failing to adequately represent the listening tastes and musical creations of artists around the world.
2. With this analysis being focused on music only from 2000-2019, the data and resulting analysis may inherently reflect discriminatory tastes prevalent during this period of time. As we move closer to the present day, the music industry has become more diverse with the inclusion of global music, unique singers, and united efforts against problematic musicians and producers (ex. Ke$ha's producer, Dr. Luke). Excluding the popular music from 2020 onwards means excluding the musical tastes of a more conscious population of listeners and may potentially amplify the work of troubled musicians and their beliefs.
3. By relying on Spotify's subjective scores for audio features, this analysis is vulnerable to echoing and amplifying any existing biases within these features. For example, if the models used to determine the scores within Spotify are trained mostly on American music, then the lower scores assigned to other diverse music will influence the findings in this analysis to discourage emulating

other styles of music. Clearly, this can lead to the minimization of music from other countries, cultures, and backgrounds which is a critical representative harm.

Future Research and Broader Significance

To address the limitations and mitigate the ethical concerns discussed above, future work should be done in a similar methodology to this report that incorporates music from 2020 onwards from multiple sources of data beyond Spotify and Billboard. Ideally, a future analysis includes top songs aggregating the features from multiple music libraries across different countries up to 2024 (and beyond).

Despite these limitations, the many modes of statistical analysis with consideration of multiple features and support from corroborating research shows that the results from this analysis holds insights that can be applied as a broad recipe for making a hit song in the current era. These insights are extremely significant in allowing up-and-coming artists make their first hit to secure their livelihood, boost their career, and share the story behind their music. Additionally, by sharing this information about top hits widely, we even the playing field in the music industry to allow smaller artists to succeed while introducing listeners to more diverse music representative of different cultures and styles.

Acknowledgements

I would like to express my gratitude to Professor Chiara Sabatti, Professor Michael Bernstein, and Joonhyuk Lee for their guidance and feedback on the methodology, analysis, and structure for this report.

References

- Azhad Syed *Hot or Not: Analyzing 60 Years of Billboard Hot 100 Data*, Toward Data Science
- Billboard (2024) *Billboard Charts Legend: Recurrent Rules*, Billboard
- Josh Viner (2020) *What Makes a Hit Song: Analyzing Data from the Billboard Hot 100 Chart*, Medium
- Mohamed Nasreldin (2018) *Song Popularity Predictor*, Medium
- Pham et al. (2015) *Predicting Song Popularity*, Stanford Department of Computer Science
- R. Kelly (2005) *Trapped In The Closet*, Spotify
- Rutger Nijkamp (2024) *Prediction of Product Success: Explaining Song Popularity by Audio Features from Spotify Data*, University of Twente
- Sean Miller (2024) *Billboard Hot 100 Weekly Charts with Spotify Audio Features*, Kaggle
- SongBPM *Song Metrics*, Imma Be
- Spotify For The Record (2023) *The Crossover Effect: Artist Collaborations Thrive on Spotify*, Spotify

Appendix

In the appendix below, you can find the histograms and boxplots for the following variables:

- Duration
- Danceability
- Energy
- Key
- Loudness
- Valence
- Tempo
- Time Signature
- Speechiness (with and without log transformation)
- Acousticness (with and without log transformation)
- Instrumentalness (with and without log transformation)
- Liveness (with and without log transformation)
- Time Signature
- Pop
- Country
- Blues
- R&B
- Rap and Hip-Hop
- Indie
- Rock
- OtherGenre
- Mode
- Famous Collaborator
- Number of Collaborators

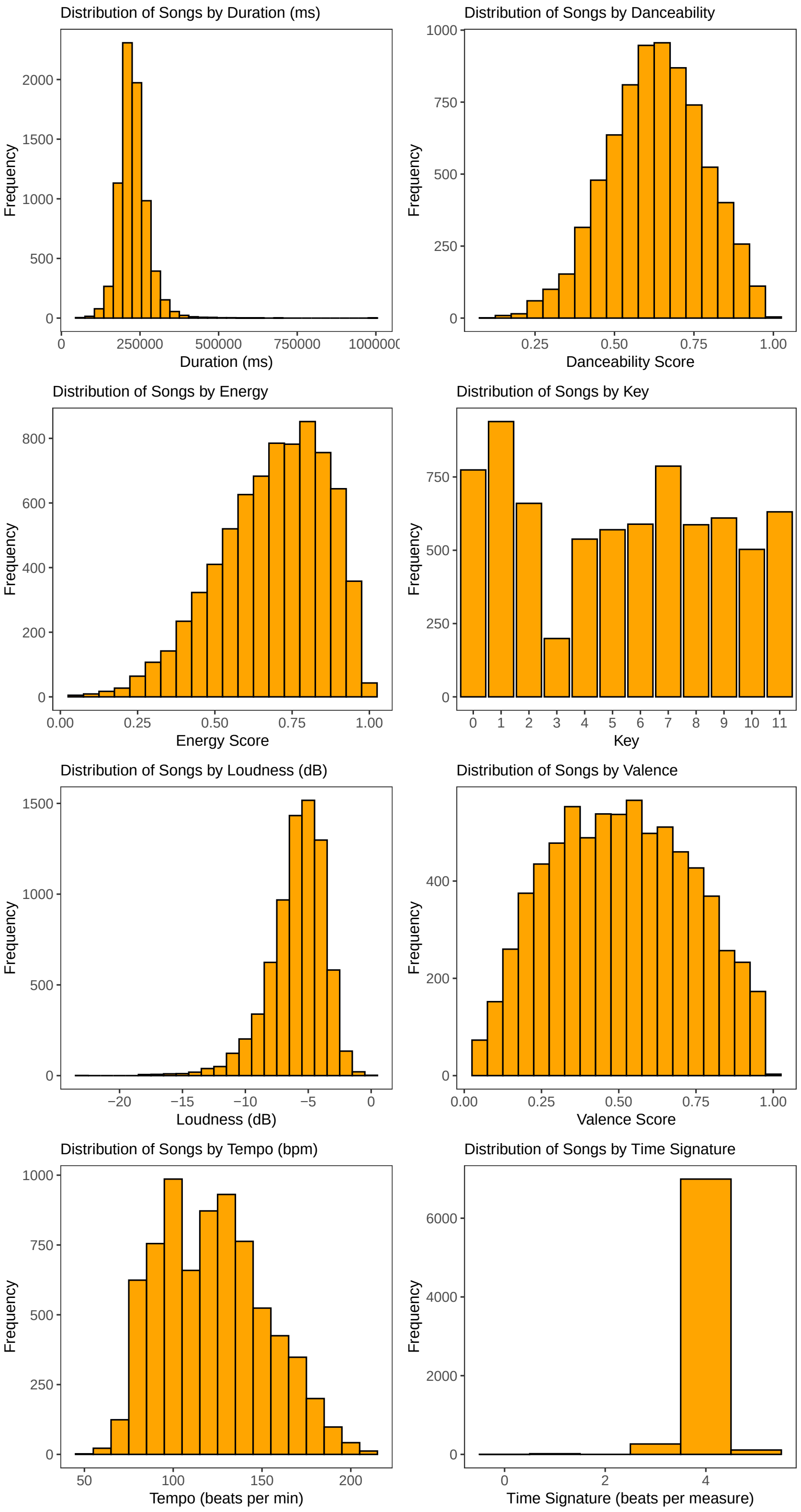


Figure 38: Distribution of songs by duration, danceability, energy, key, loudness, valence, tempo, and time signature.

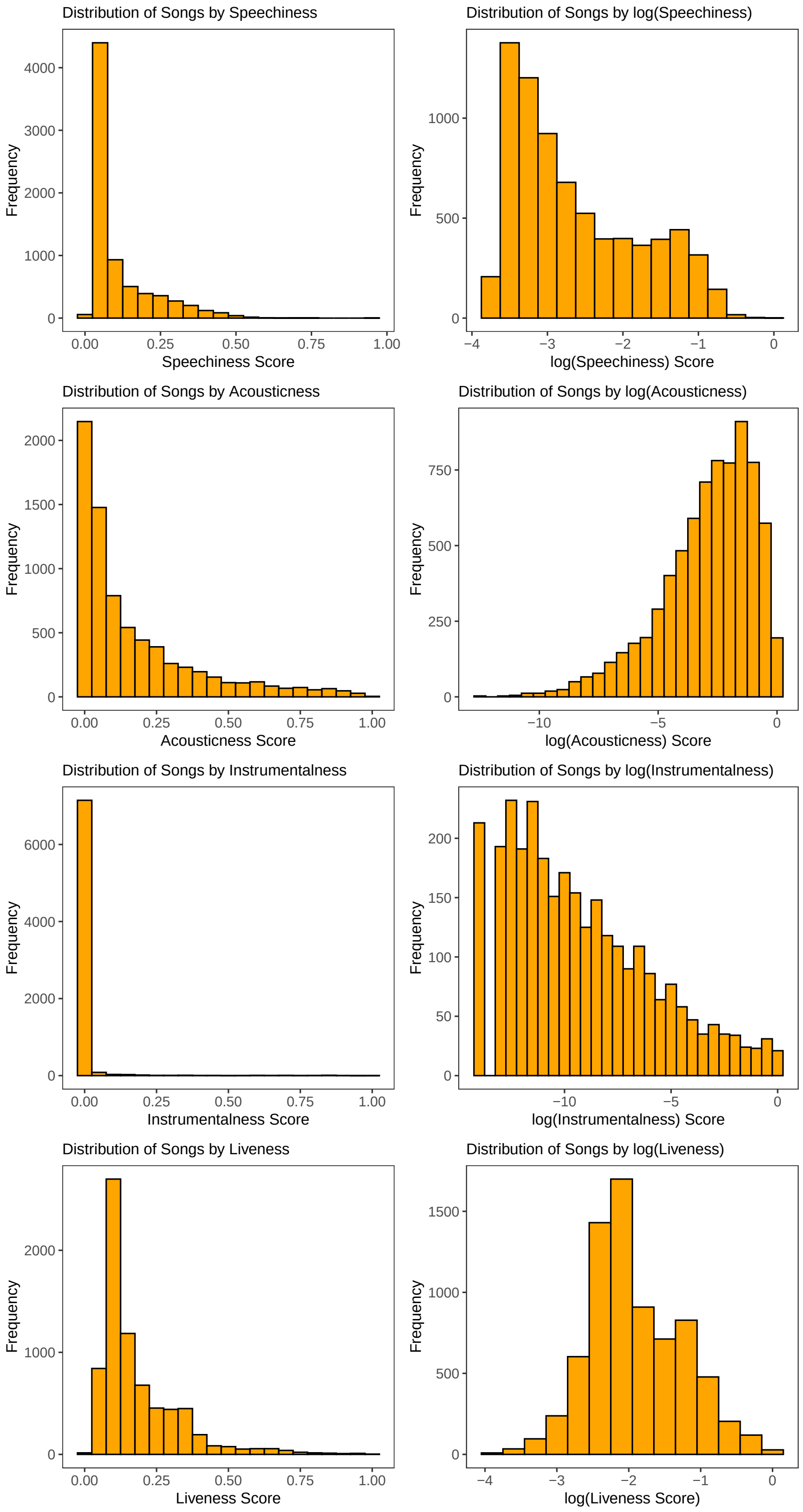


Figure 39: Distribution of songs by speechiness, acousticness, instrumentalness, and liveness with the associated log transformations.

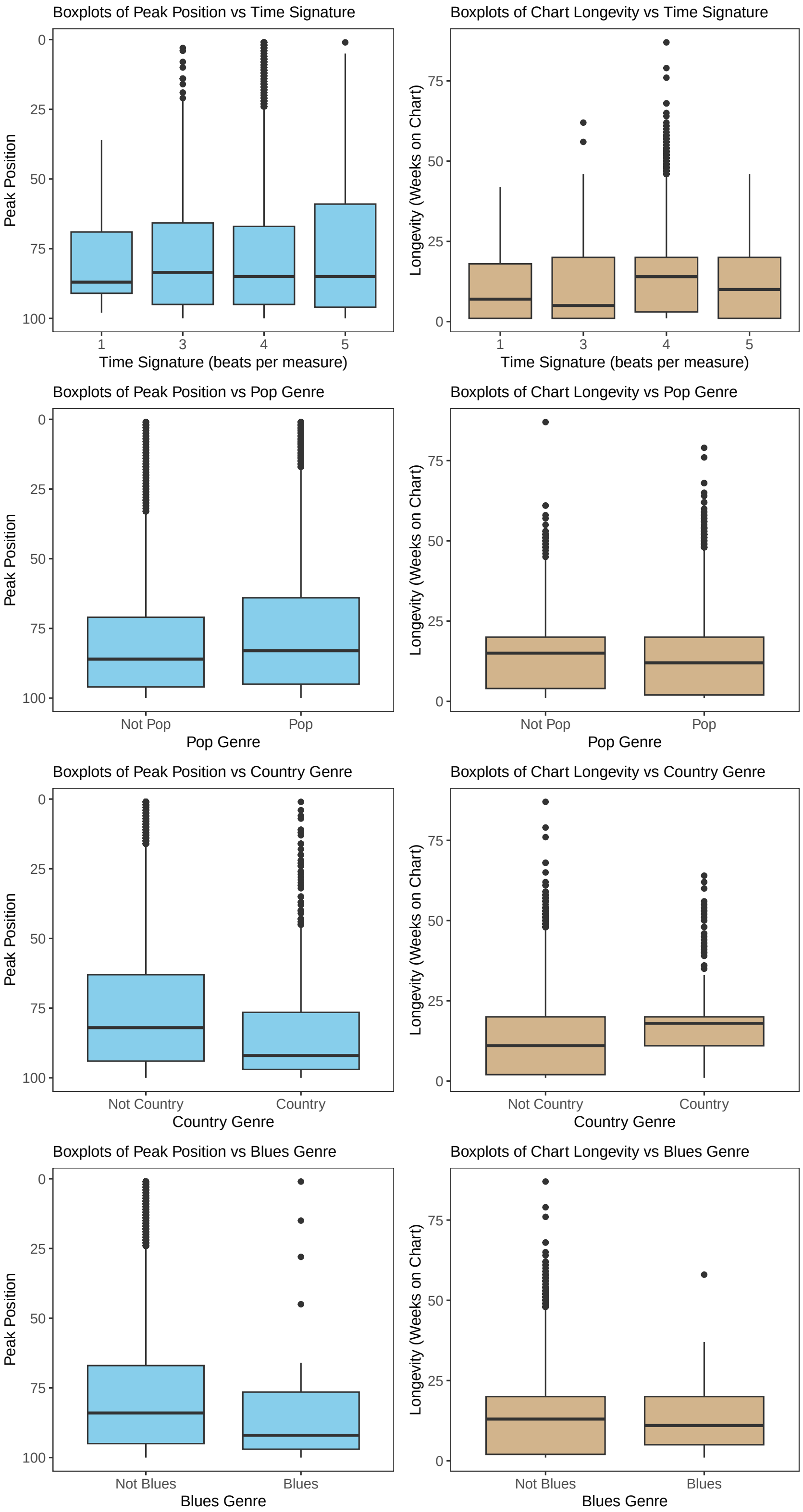


Figure 40: Boxplots of peak position and chart longevity against time signature, pop genre, country genre, and blues genre.

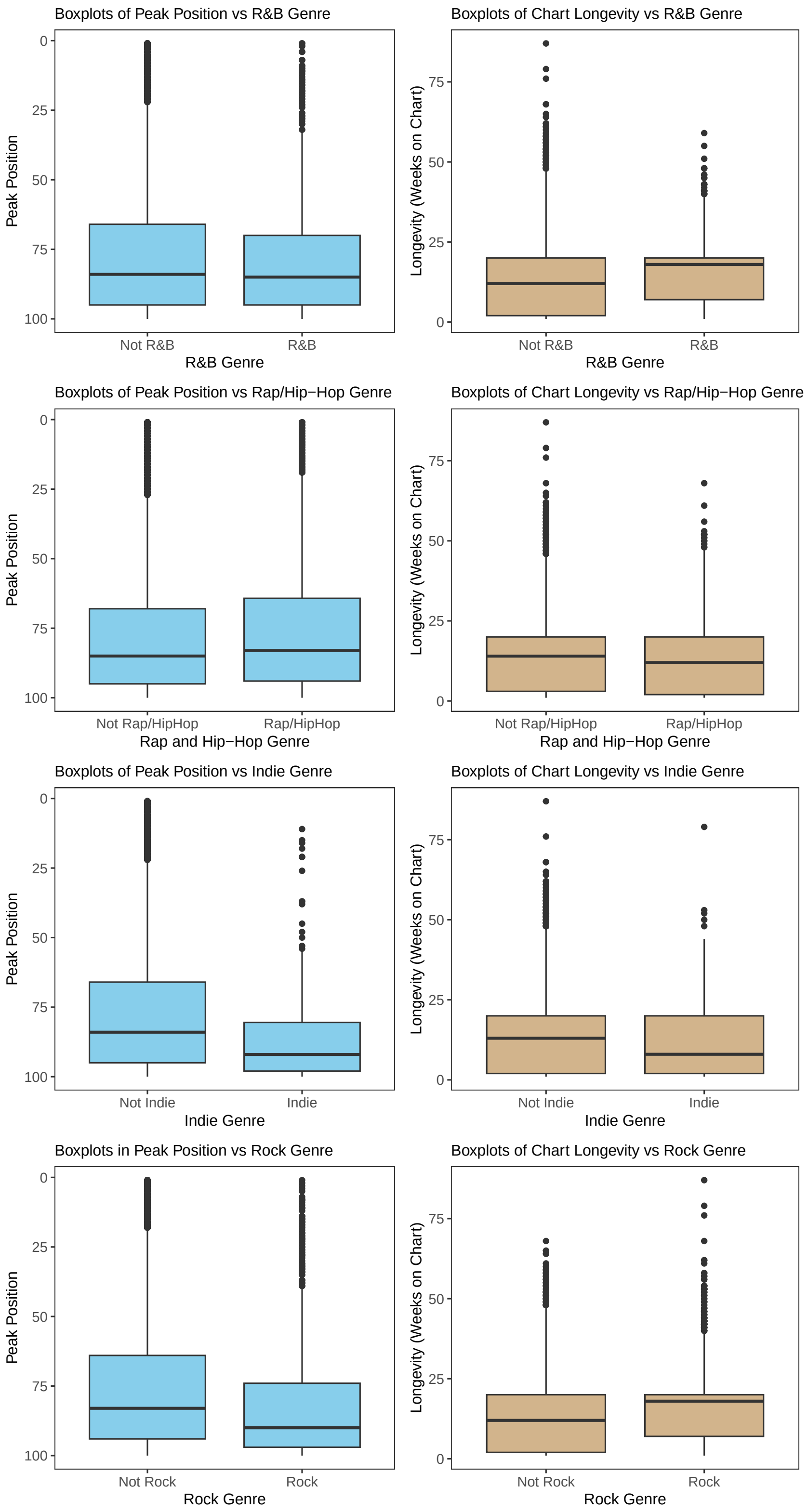
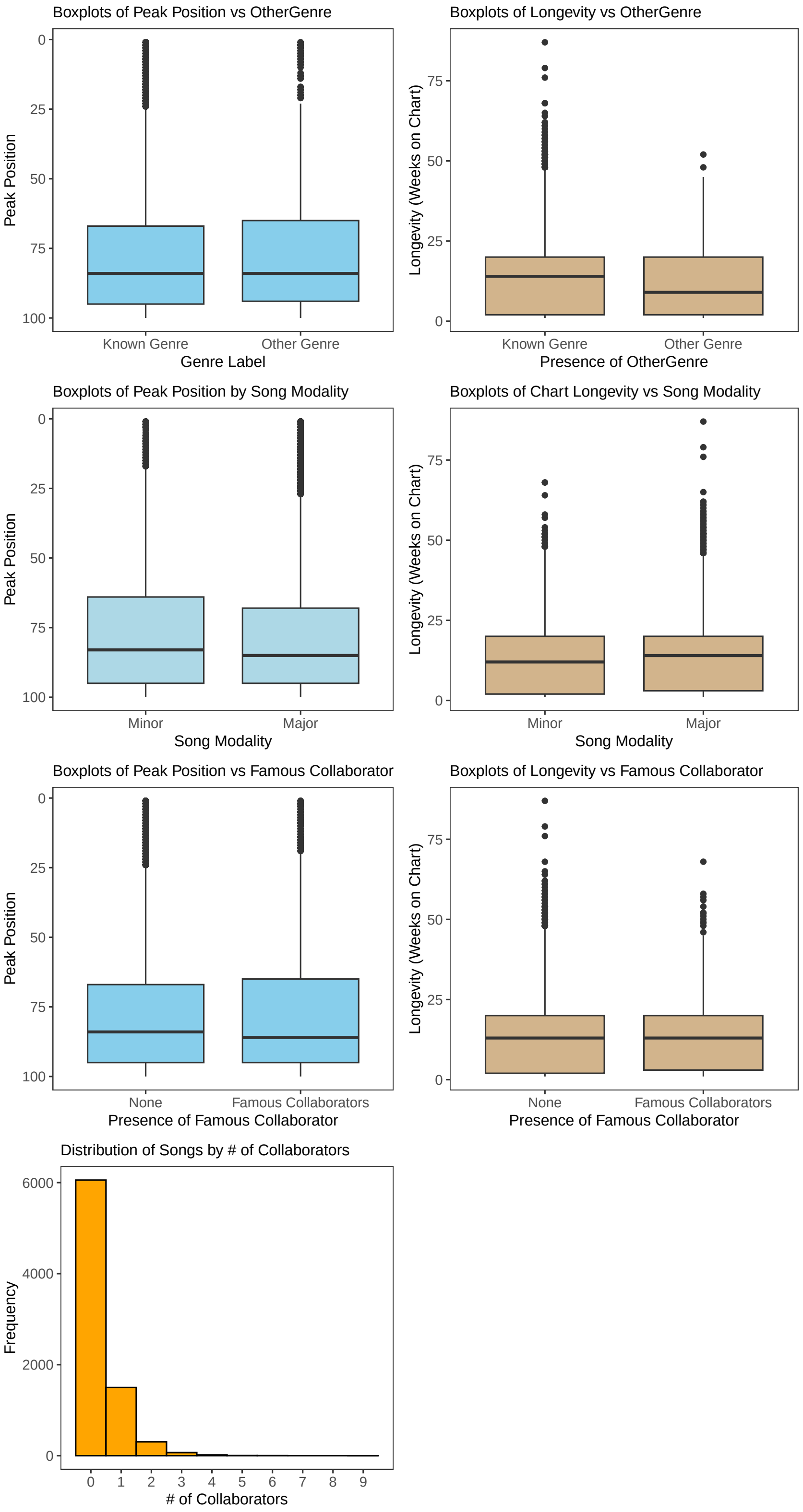


Figure 41: Boxplots of peak position and chart longevity against R&B genre, rap/hip-hop genre, indie genre, and rock genre.



Boxplots of Peak Position vs OtherGenre

Boxplots of Longevity vs OtherGenre

Boxplots of Peak Position by Song Modality

Boxplots of Chart Longevity vs Song Modality

Boxplots of Peak Position vs Famous Collaborator

Boxplots of Longevity vs Famous Collaborator

Distribution of Songs by # of Collaborators

Figure 42: Boxplots of peak position and chart longevity against less popular genres (OtherGenre), modality, and presence of famous collaborators.