Now Playing

Chart Toppers: Billboard Top 100 Songs

Khushmeet Chandi, Zaid Muqsit, Athena Ru

1:23 2:36

Introduction

Key Objectives

What makes a song popular / successful (as determined by Billboard rankings) in the current musical landscape of the United States of America? How are the features of a song that make it peak higher in the Billboard similar to / different from the features that make a song last longer on the Billboard? How has this evolved over the decades?

Audience

This can inform **music critics** of the trends that are shaping the musical landscape and inform **artists** who are determined to make music that will be a hit with the masses.

Summary

Use song features to **predict** longevity, peak quantile, and decade

- **Model 1:** Random Forest predicting the number of months a song will stay on the Billboard Top 100
- **Model 2:** Cumulative Logit Model predicting the quantile of the Billboard a song will peak in
- **Model 3:** Multinomial Model that predicts what decade a Billboard Top 100 song was released

Conduct **inference** to see what characteristics of songs are related to success.

Data

Sources: Data.World, Billboard.com, Spotify
Compiled in TidyTuesday GitHub Repo

Overview: Top 100 Billboard Songs from August 9th 1958 - May 29th 2021

12,055 unique songs (after cleaning)
33 variables (after engineering)

Data - Feature Engineering

Data preparation

- Drop all NA values from the two datasets and merge them on song_id
- Drop all songs where genre was not given (an empty bracket)

New variables

- debut_position (numerical): chart position when the song first entered the Billboard 100
- months_on_chart (numerical): floor(weeks_on_chart / 4.33) assuming an average of 4.33 weeks in a month
- max_peak_position (numerical): highest position the song peaked in
- peak_quantile (ordered categorical): quantile for max_peak_position
    - 1 for #1-25; 2 for #26-50, 3 for #51-75; 4 for #76-100

Data - Feature Engineering

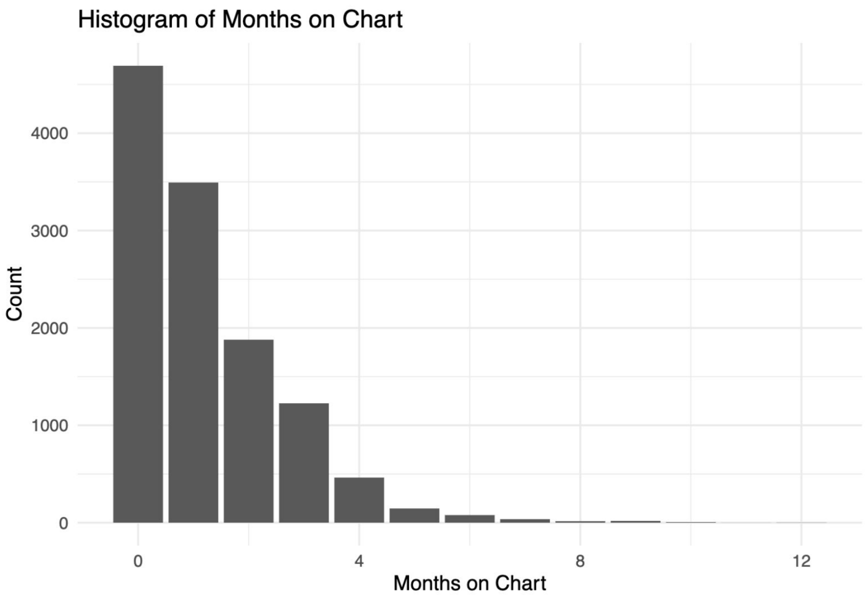
New variables

- decade (ordered categorical): 1950, 1960, 1970, 1980, 1990, 2000, 2010, 2020
- genre_list (list): each sub genre in spotify_genre is replaced with the main genre
  - The main genres are the top genres (five most frequent) in the dataset: pop, rock, soul, country, and rap
  - ("dance pop", "pop", "uk pop") becomes ("pop", "pop", "pop")
- main_genre (categorical): the most frequent top genre in genre_list; the more frequent top genre for tiebreakers; "other" if none of the top genres appear in genre_list
  - "country" is assigned if genre_list is list("country", "country", "country", "rock")
  - “pop” is assigned for ("pop", "rock")

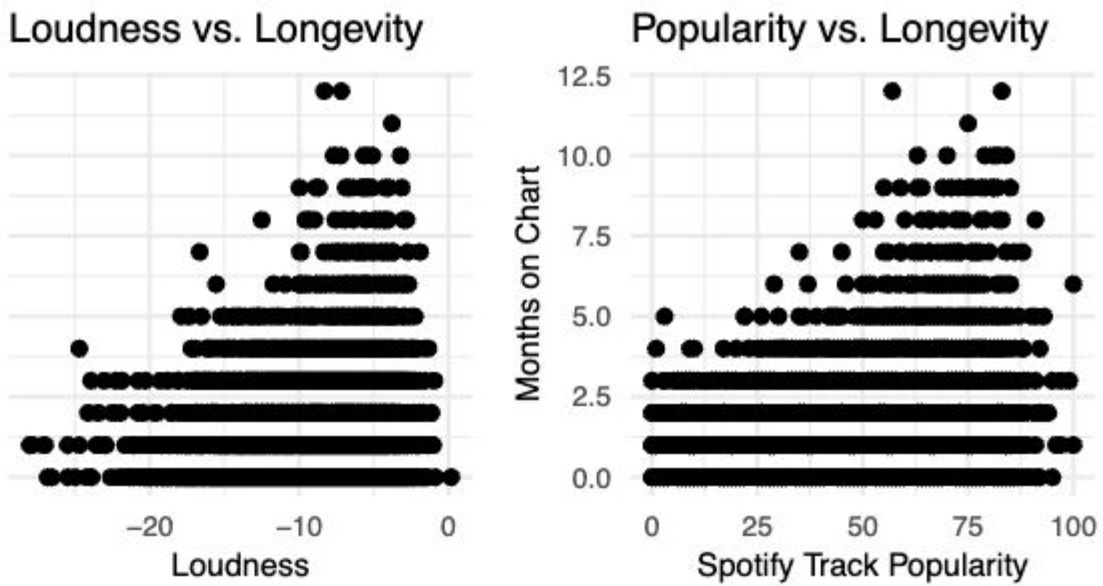
Model 1

Longevity on Chart

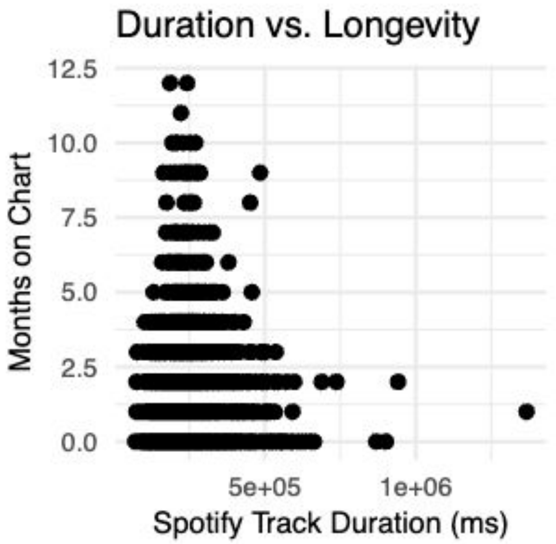
EDA

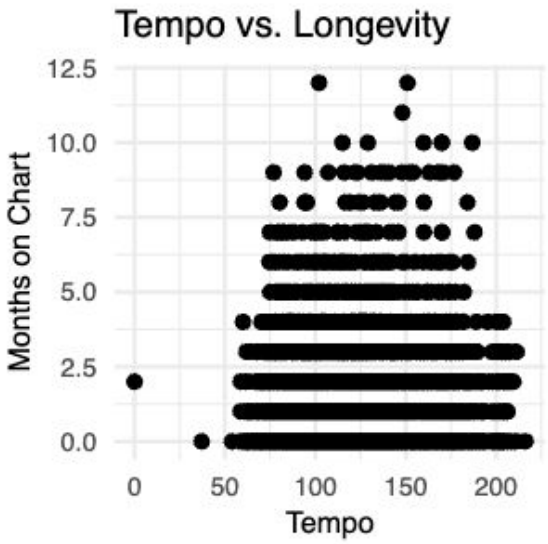

Histogram of Months on Chart
Count
4000
3000
2000
1000
0
0
4
8
12
Months on Chart

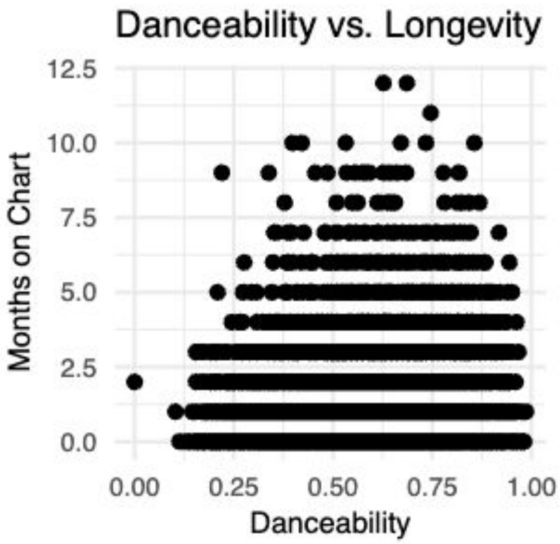
EDA


Loudness vs. Longevity
Popularity vs. Longevity
12.5
10.0
7.5
5.0
2.5
0.0
Months on Chart
–20
–10
0
Loudness
0
25
50
75
100
Spotify Track Popularity

EDA

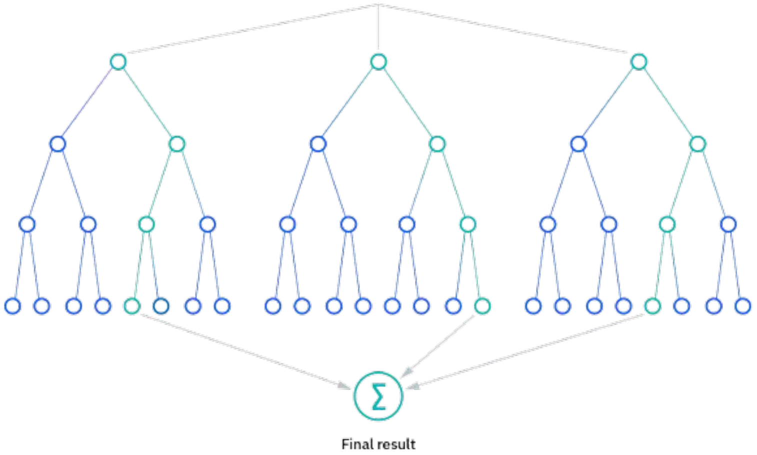

Duration vs. Longevity
Months on Chart
12.5
10.0
7.5
5.0
2.5
0.0
5e+05
1e+06
Spotify Track Duration (ms)


Tempo vs. Longevity
Months on Chart
12.5
10.0
7.5
5.0
2.5
0.0
0
50
100
150
200
Tempo

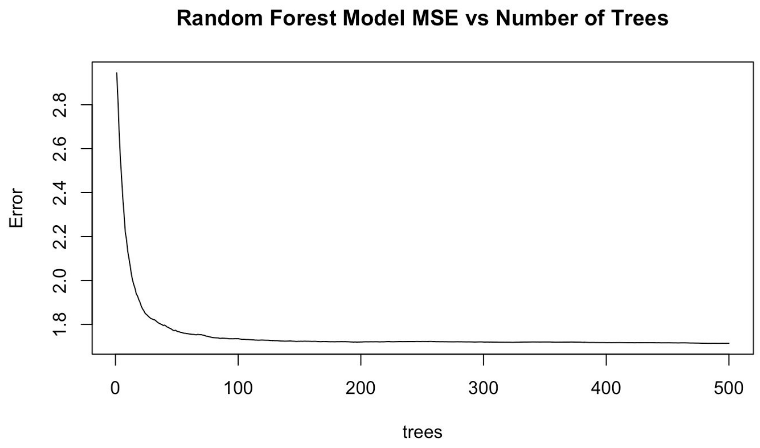

Danceability vs. Longevity
Months on Chart
12.5
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Danceability

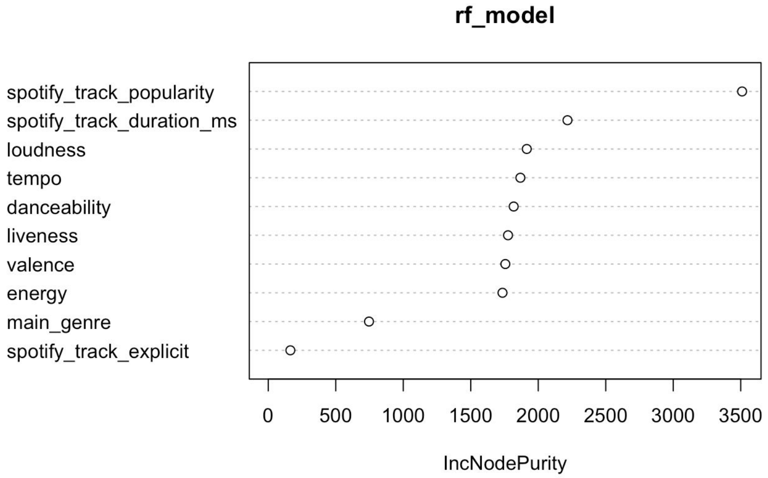
Methodology

Generative Additive Model → Random Forest



Methodology

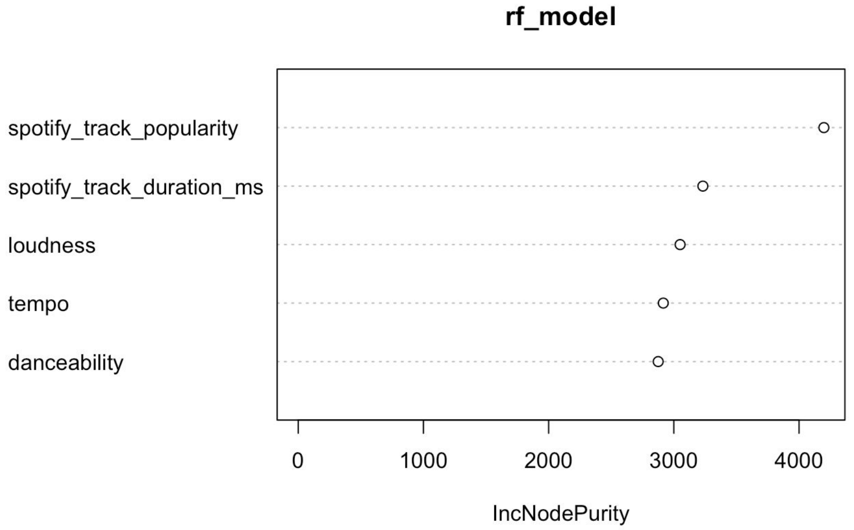




Model Fit & Interpretation

Training R^2: 0.785
Test R^2: 0.131
Test MSE: 1.607
Accuracy: 60.606%

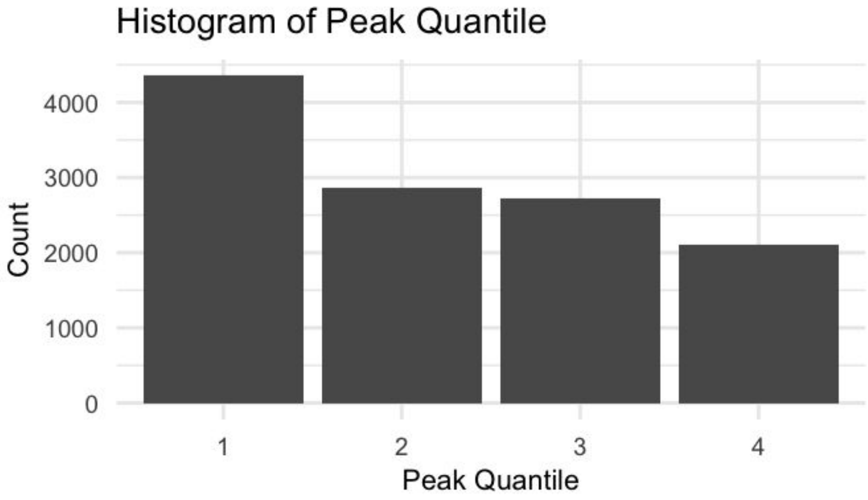
	IncNodePurity
danceability	2864.106
spotify_track_popularity	4119.686
tempo	2940.162
loudness	3124.234
spotify_track_duration_ms	3291.100

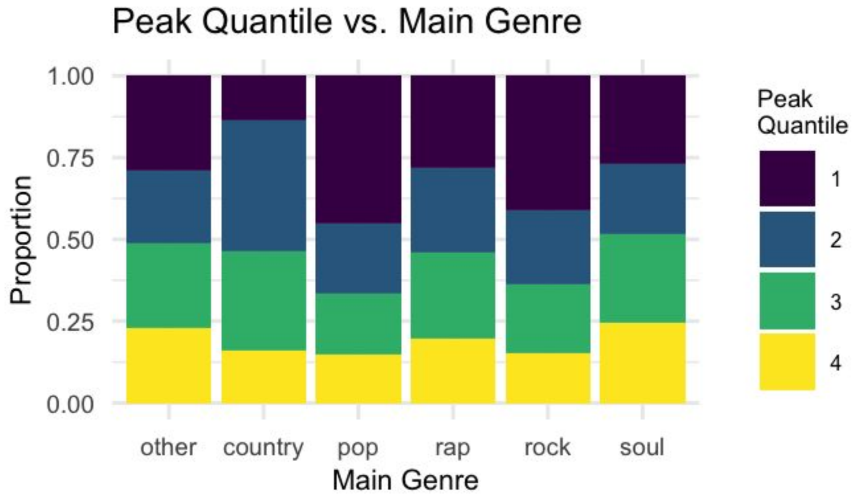


Model 2

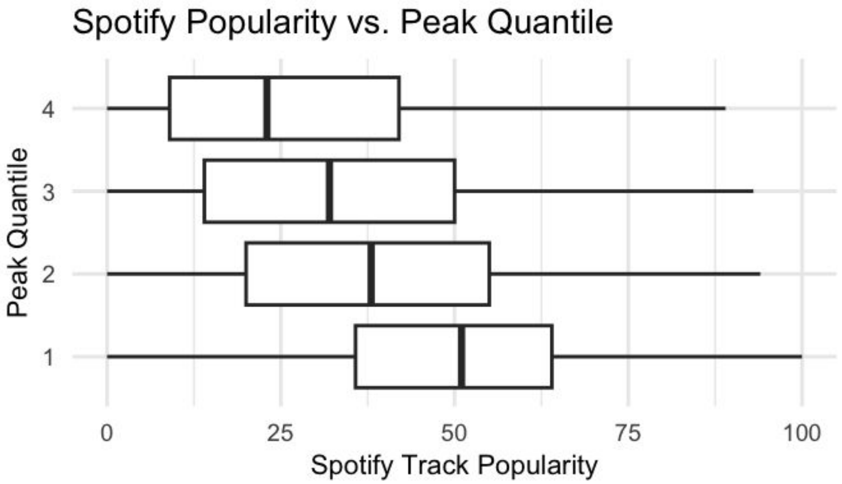
Peak quantile on chart

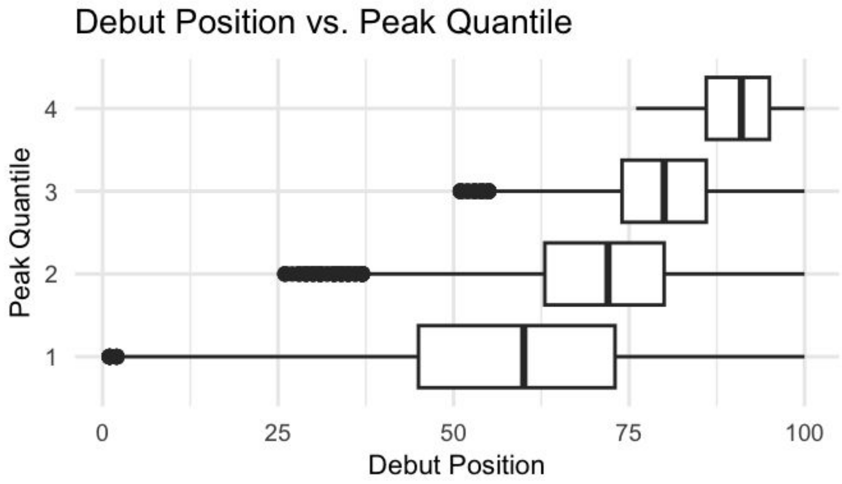
EDA


Histogram of Peak Quantile
Count
4000
3000
2000
1000
0
1
2
3
4
Peak Quantile

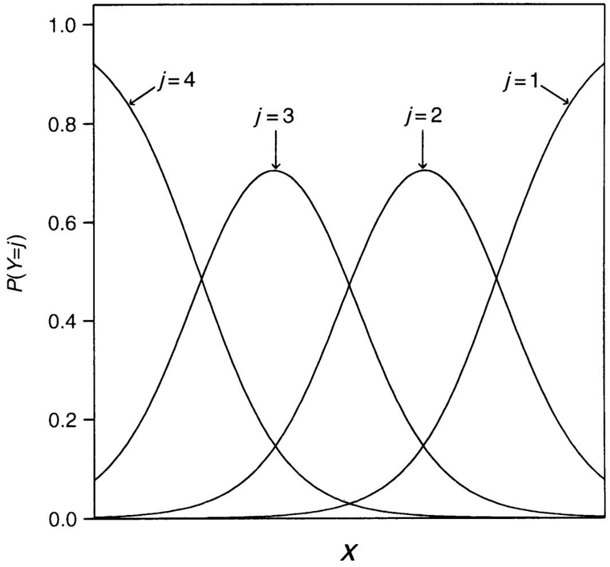

Peak Quantile vs. Main Genre
Proportion
1.00
0.75
0.50
0.25
0.00
other
country
pop
rap
rock
soul
Main Genre
Peak
Quantile
1
2
3
4

EDA


Spotify Popularity vs. Peak Quantile
Peak Quantile
4
3
2
1
0
25
50
75
100
Spotify Track Popularity


Debut Position vs. Peak Quantile
Peak Quantile
4
3
2
1
0
25
50
75
100
Debut Position

Methodology



Cumulative logit model

1. Response (peak quantile) is ordinal categorical with 4 categories
2. Predictors are either continuous or categorical
3. Check multicollinearity
4. **Proportional Odds** property

5-Fold CV on training data (80% of dataset)

Model Fit

Residual Deviance: 4828.967

AIC: 4852.967

Accuracy on testing set: 55.168 %

Statistically significant predictors:

Spotify Track Popularity, Main Genre: Country, Main Genre: Rap, Spotify Track Explicit: TRUE, and Debut Position

	Estimate	2.5 %	97.5 %	OR
spotify_track_popularity	-0.035	-0.039	-0.031	0.966
main_genrecountry	0.419	0.086	0.752	1.520
main_genrepop	0.028	-0.231	0.286	1.028
main_genrerap	0.532	0.036	1.029	1.703
main_genrerock	0.225	-0.041	0.492	1.252
main_genresoul	0.113	-0.198	0.425	1.120
spotify_track_explicitTRUE	1.137	0.733	1.541	3.116
debut_position	0.103	0.096	0.110	1.108
danceability	0.039	-0.516	0.595	1.040

Interpretation

Main Genre: Country and Rap

- Songs with a main genre of country and rap are 1.5 and 1.7 times more likely to to be in an upper quantile (1 or 2), respectively, than songs with a main genre of other

Spotify Track Popularity

- For every one unit increase in Spotify Track Popularity, the odds of peaking in an upper quantile (1 or 2) decreases by 3.4%, holding all else constant. This is counterintuitive.

Debut Position

- For every one unit increase in debut position, the odds of peaking in an upper quantile increases by 10.8%, holding all else constant.

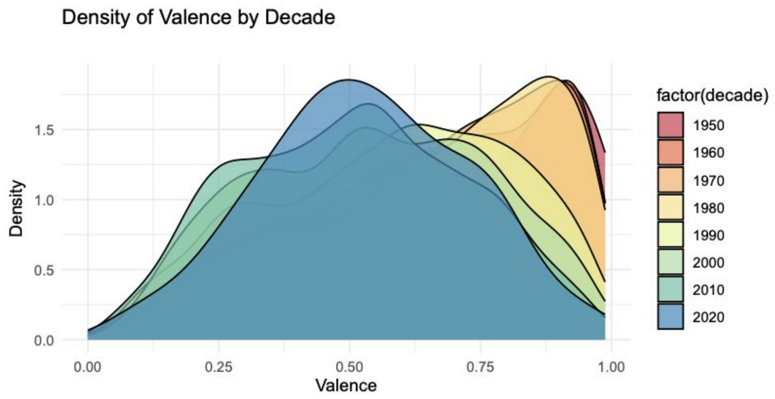
Spotify Track Explicit

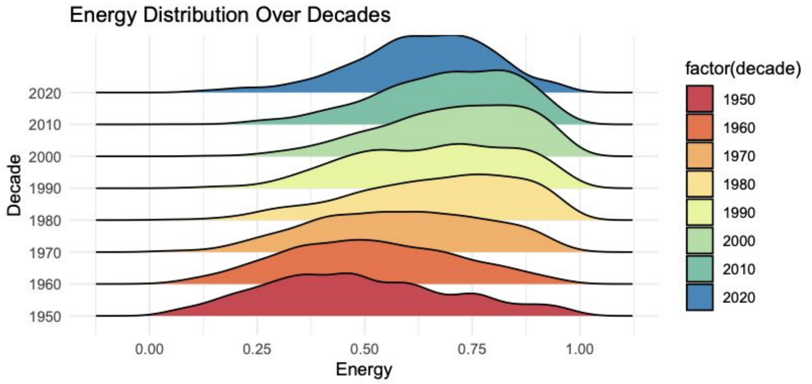
- For explicit songs, the odds of peaking in an upper quantile is 3.116 times that of not explicit songs, holding all else constant.

Model 3

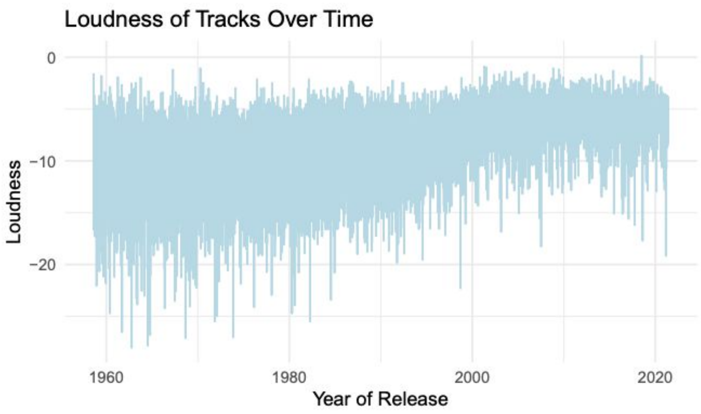
Decade of Song

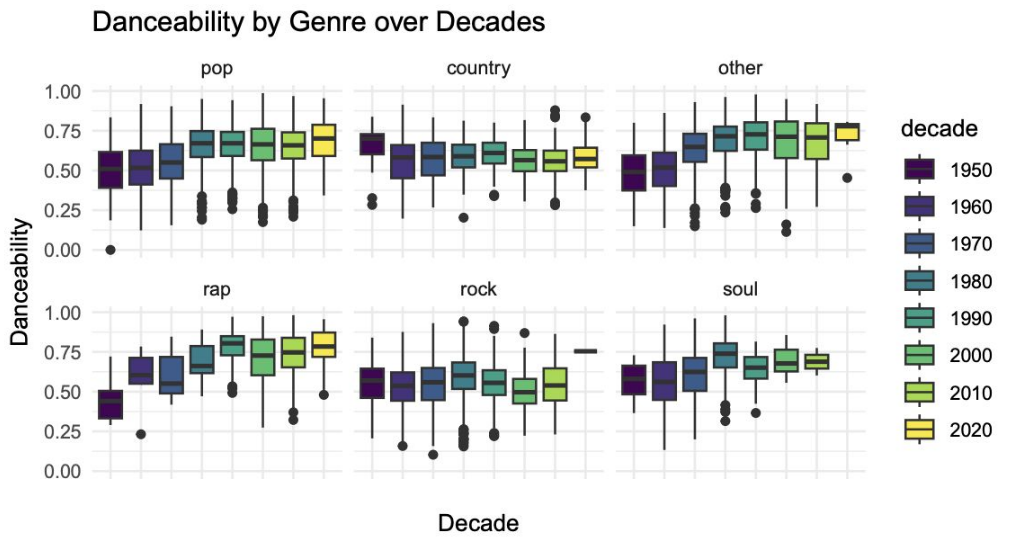
EDA


Density of Valence by Decade
Density
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
Valence
factor(decade)
1950
1960
1970
1980
1990
2000
2010
2020

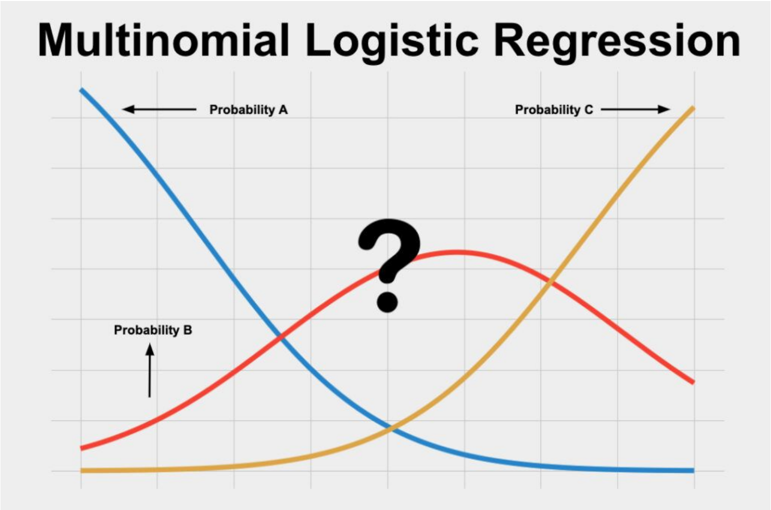

Energy Distribution Over Decades
Decade
2020
2010
2000
1990
1980
1970
1960
1950
0.00
0.25
0.50
0.75
1.00
Energy
factor(decade)
1950
1960
1970
1980
1990
2000
2010
2020

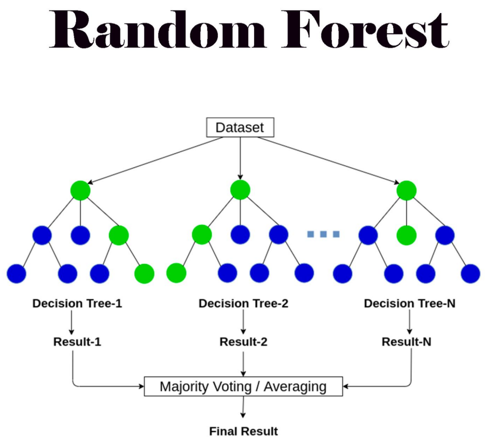
EDA


Loudness of Tracks Over Time
0
-10
-20
Loudness
1960
1980
2000
2020
Year of Release


Danceability by Genre over Decades
pop
country
other
rap
rock
soul
1.00
0.75
0.50
0.25
0.00
Danceability
Decade
decade
1950
1960
1970
1980
1990
2000
2010
2020

Methodology





Interpretation

```
$`1970`
[1] "log(P(1970)/P(1950)) = -5.5288 + 0.51*danceability + 1.1969*energy + -0.2078*loudness +
-4.9755*speechiness + -4.3317*acousticness + -0.1476*instrumentalness + -1.0163*valence +
0.0107*spotify_track_popularity + -0.2628*main_genrecountry + -2.7827*main_genreother +
-9.0596*main_genrerap + 0.6047*main_genrerock + 2.336*main_genresoul +
-3.1058*main_genrecountry:spechiness + 9.2831*main_genreother:spechiness +
-7.3211*main_genrerap:spechiness + 5.1368*main_genrerock:spechiness +
2.2475*main_genresoul:spechiness + 0.1427*main_genrecountry:loudness +
0.0486*main_genreother:loudness + -0.6554*main_genrerap:loudness +
0.1023*main_genrerock:loudness + 0.1487*main_genresoul:loudness +
2.9443*main_genrecountry:acousticness + -0.3741*main_genreother:acousticness +
-4.4072*main_genrerap:acousticness + -0.0851*main_genrerock:acousticness +
0.0142*main_genresoul:acousticness + 1.705*main_genrecountry:danceability +
4.0021*main_genreother:danceability + -9.3302*main_genrerap:danceability +
1.3878*main_genrerock:danceability + 0.8997*main_genresoul:danceability"

$`1980`
[1] "log(P(1980)/P(1950)) = -10.3033 + 7.7816*danceability + 4.0161*energy + -0.2065*loudness
+ -14.3539*speechiness + -5.4317*acousticness + -2.1458*instrumentalness + -4.4328*valence +
0.0225*spotify_track_popularity + -2.4413*main_genrecountry + -2.0745*main_genreother +
3.844*main_genrerap + 1.5594*main_genrerock + -0.1727*main_genresoul +
1.262*main_genrecountry:spechiness + 13.281*main_genreother:spechiness +
-2.8916*main_genrerap:spechiness + 7.8854*main_genrerock:spechiness +
0.2124*main_genresoul:spechiness + -0.0783*main_genrecountry:loudness +
-0.0257*main_genreother:loudness + 0.547*main_genrerap:loudness +
0.1051*main_genrerock:loudness + 0.1629*main_genresoul:loudness +
2.2781*main_genrecountry:acousticness + -1.0437*main_genreother:acousticness +
-2.9391*main_genrerap:acousticness + -0.1998*main_genrerock:acousticness +
0.428*main_genresoul:acousticness + 0.2551*main_genrecountry:danceability +
1.607*main_genreother:danceability + 0.9944*main_genrerap:danceability +
-0.4825*main_genrerock:danceability + 2.0955*main_genresoul:danceability"
```

Model Fit

Accuracy	0.5742739
Kappa	0.4880420
AccuracyLower	0.5542453
AccuracyUpper	0.5941217
AccuracyNull	0.2116183
AccuracyPValue	0.0000000

Table 3: True vs. Predicted Decades (Truth along Side)

	1950	1960	1970	1980	1990	2000	2010	2020
1950	0	0	0	0	0	1	0	0
1960	53	420	115	14	5	2	2	2
1970	2	62	223	105	36	8	1	0
1980	1	12	98	214	91	13	5	0
1990	0	7	20	51	107	40	12	0
2000	0	4	8	14	57	179	47	0
2010	0	2	1	7	19	74	237	25
2020	0	3	1	0	0	0	6	4

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
Class: 1950	0.000	1.000	0.000	0.977	0.000	0.000
Class: 1960	0.824	0.898	0.685	0.950	0.685	0.824
Class: 1970	0.479	0.890	0.510	0.877	0.510	0.479
Class: 1980	0.528	0.890	0.493	0.903	0.493	0.528
Class: 1990	0.340	0.938	0.451	0.904	0.451	0.340

Conclusions

Spotify Track Popularity

- Common predictor across all three models
- Model 1 suggests that artists with higher popularity tend to have songs with higher longevity
- Model 2 indicates higher popularity scores are slightly less likely to peak in an upper quantile.
- Discrepancy may suggest that artists who are more underrated tend to quickly peak for a shorter period of time. While artists who are more well-known and popular tend to stay on the Billboard longer even if they do not peak as high on the Billboard.
- This discrepancy may have been exacerbated by promotions of songs through TikTok and other social media that tends to use clips of songs.
- Short clips might prioritize catchy or memorable segments of a song that make it peak higher yet the overall artist may not be as popular so its fanbase changes or drops quickly once the song's clip is no longer viral.

Conclusions

- Whether the **Track is Explicit** and **Genre** were both significant predictor variables for peak quantile, however, these were not for longevity. This suggests that the genre and whether a song is explicit is more helpful in determining the peak quantile of a song rather than how long it will stay on the Billboard.
- Model 1 and Model 3 offer differing insights into what makes a song popular
  - While a lot of different predictors, including interaction effects, that were significant in predicting decade were included in Model 3, there will little predictors that appeared to be significant in predicting longevity (as seen in LASSO as well) in Model 1
  - This suggests that it more difficult to predict how long a song will stay on the Billboard rather than genre using the current data that we used.

Limitations & Future Work

- Limited to **Billboard 100 Songs**
  - Does not account for songs that do not make the Billboard
  - Future Work would include predicting how to get a song to appear on the Billboard in the first place
- Not representative of **all music**
- **US Bias** (Global Trends, Trends in Other Countries)
- A more general model with separate effects for the different cumulative probabilities (Model 2)

Thank you for listening!