UC Berkeley MIDS
DATASCI-200
December 11, 2023

Project 2: Beats & Bytes (Decoding the Secret Sauce of Chartbusters

Team: Laura Lubben, Rachel Burgess, Bryce Loomis, Sammy Cayo
Github Repo: UC-Berkeley-I-School/project2_lubben_cayo_burgess_loomis

Executive Summary

In the ever-evolving world of digital music streaming, Spotify emerges as a critical player, offering a diverse range of songs to a global audience. This project delves into the intriguing dynamics of song popularity on Spotify, aiming to discern what musical characteristics contribute to a track's success. The analysis aims to understand the complex interplay between a song's attributes and its appeal to listeners.

The dataset, sourced from various playlists, categorizes songs by genre and subgenre, providing a comprehensive view of each track's musical dimensions. These include danceability, energy, loudness, tempo, valence, and content-related attributes like speechiness, acousticness, and instrumentalness. Crucial information such as release dates, artist names, and album details enrich our analysis.

Our research pivots on answering the following critical questions:

- **Musical Attributes**: How do factors like musical characteristics correlate with a song's popularity on Spotify?

- **Genre and Trends**: Are there noticeable trends in popularity across different genres and over time?

Overview of the Data:

Our dataset has 32,833 songs from various genres between the years 1957 and 2020. The first 11 columns of the dataset primarily contain information about the track, detailing the song’s name, unique identifier, as well its album and artist. Since the dataset is a collection of different playlists, it also contains unique identifiers and names of each playlist as well as their genre and subgenres. The remaining 12 columns are notable parameters that distinguish each track by musical feature.

These features include danceability, energy, key, loudness, mode speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration. We’ve included a deeper overview of each musical feature.

Musical Features:

- **Danceability:** How suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable
- **Energy:** A measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy
- **Key:** Estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D ♭ , 2 = D, and so on. If no key was detected, the value is -1.
- **Loudness:** Overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of track
- **Mode:** Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- **Speechiness:** Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.
    - *Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.*
- **Acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
- **Instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- **Tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **Liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- **Valence:** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- **Duration**: Duration of song in milliseconds

The dataset also categorizes each track by playlist genre and subgenre. There are 6 unique playlist genres in the data set, and 24 unique playlist subgenres for further classification. The additional subgenres gives us great flexibility for trend analysis.

Data Preprocessing and Exploratory Analysis

The foundational stage of our Spotify dataset project involved data preprocessing and exploratory analysis. The data was loaded into a Pandas DataFrame for efficient manipulation and analysis. Using Pandas in the data cleansing phase, we dropped entries that were missing track, album, or artist name data. Next, we removed any tracks missing release date information.

With a clean dataset, we created a separate DataFrame “df_playlist” for playlist analysis before dropping all duplicate tracks from the original “df” dataframe. Since a number of songs were included in multiple playlists–for instance, the song “Closer” by The Chainsmokers has 10 separate entries in different playlists–we wanted to remove these for track and genre analysis but maintain the option to analyze them as part of their broader playlists.

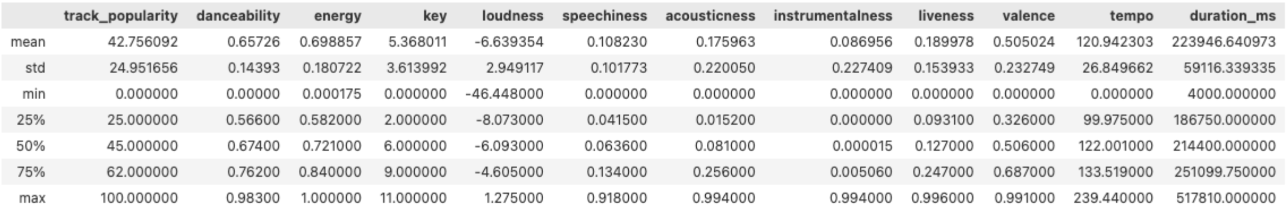
As our last step in preprocessing, we formulated two new date columns. We created a ‘year_only’ column that stored just the year the track’s album was released as well as a ‘decade’ column with the track’s decade. We used these later in trend and genre time analysis.

The exploratory analysis phase was marked by a deep dive into the dataset, employing Python's visualization libraries, Matplotlib and Seaborn. We created various visualizations, including correlation plots and histograms, to gain initial insights into the relationships between musical features and song popularity. This phase was critical in identifying potential variables for more in-depth analysis, such as trends over time as seen in Appendix 1-3.

Top Songs and Musical Feature Analysis

Following the initial exploration, we analyzed the top songs in our dataset. Selecting standout tracks like "Blinding Lights" by The Weeknd, we delved into their musical attributes, such as danceability, energy, and tempo, and how these correlated with their high popularity scores. "Blinding Lights," for example, scores a 98 in popularity, with features such as danceability = 0.513, energy = 0.796, key = 1, loudness -4.075, mode = 1, speechiness = 0.0629, acousticness 0.00147, instrumentalness = 0.000209, liveness = 0.0938, valence = 0.345, temp = 171.017, and duration = 201573 milliseconds. Summary statistics for each musical feature, showing similar patterns can be seen in Table 1 below.

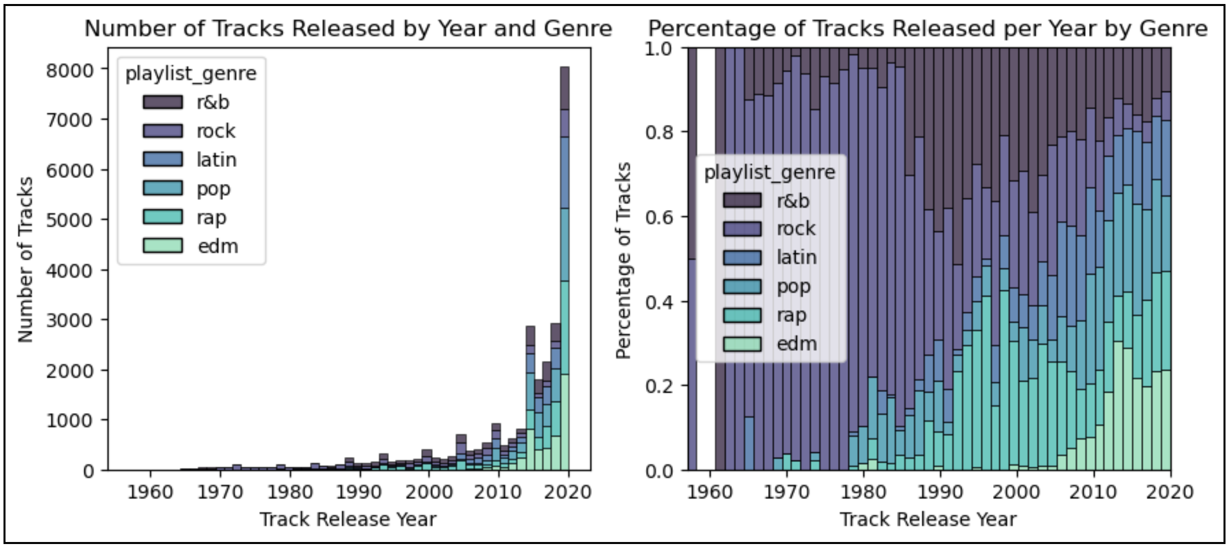
Table 1



	track_popularity	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms
mean	42.756092	0.65726	0.698857	5.368011	-6.639354	0.108230	0.175963	0.086956	0.189978	0.505024	120.942303	223946.640973
std	24.951656	0.14393	0.180722	3.613992	2.949117	0.101773	0.220050	0.227409	0.153933	0.232749	26.849662	59116.339335
min	0.000000	0.00000	0.000175	0.000000	-46.448000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	4000.000000
25%	25.000000	0.56600	0.582000	2.000000	-8.073000	0.041500	0.015200	0.000000	0.093100	0.326000	99.975000	186750.000000
50%	45.000000	0.67400	0.721000	6.000000	-6.093000	0.063600	0.081000	0.000015	0.127000	0.506000	122.001000	214400.000000
75%	62.000000	0.76200	0.840000	9.000000	-4.605000	0.134000	0.256000	0.005060	0.247000	0.687000	133.519000	251099.750000
max	100.000000	0.98300	1.000000	11.000000	1.275000	0.918000	0.994000	0.994000	0.996000	0.991000	239.440000	517810.000000

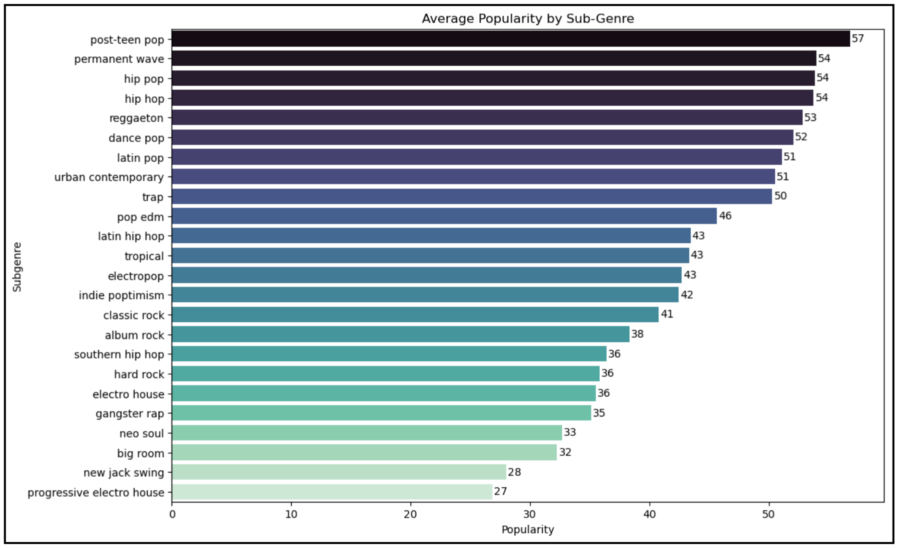
With this knowledge, we employed statistical techniques to compute summary statistics for each musical feature across the dataset, providing an overarching understanding of prevailing musical trends. In Graph 1 below, you can see the number of tracks released per year.

Graph 1



Grouping these features by playlist genre and subgenre, we were able to draw comparisons and identify patterns that might influence a song's popularity—focusing on identifying the top five songs for each decade from our Spotify dataset, as seen in Figure 4 in the Appendix. This task involved segmenting the dataset by decade to ensure a clear temporal distinction. Analyzing the popularity of songs within each era, we ranked them based on their popularity metrics. This methodology provided us with a focused snapshot of musical trends and preferences across different periods, offering insightful reflections on the evolving landscape of music over time.

Graph 2



From these insights, as seen in Graph 2 to the right and in Appendix 5-7, post teen-pop had the highest popularity, followed by permanent wave, with the lowest popularity going to new jack swing and progressive electric house. In order to present our findings effectively, we generated a bar chart that visually represented the average popularity of subgenres.

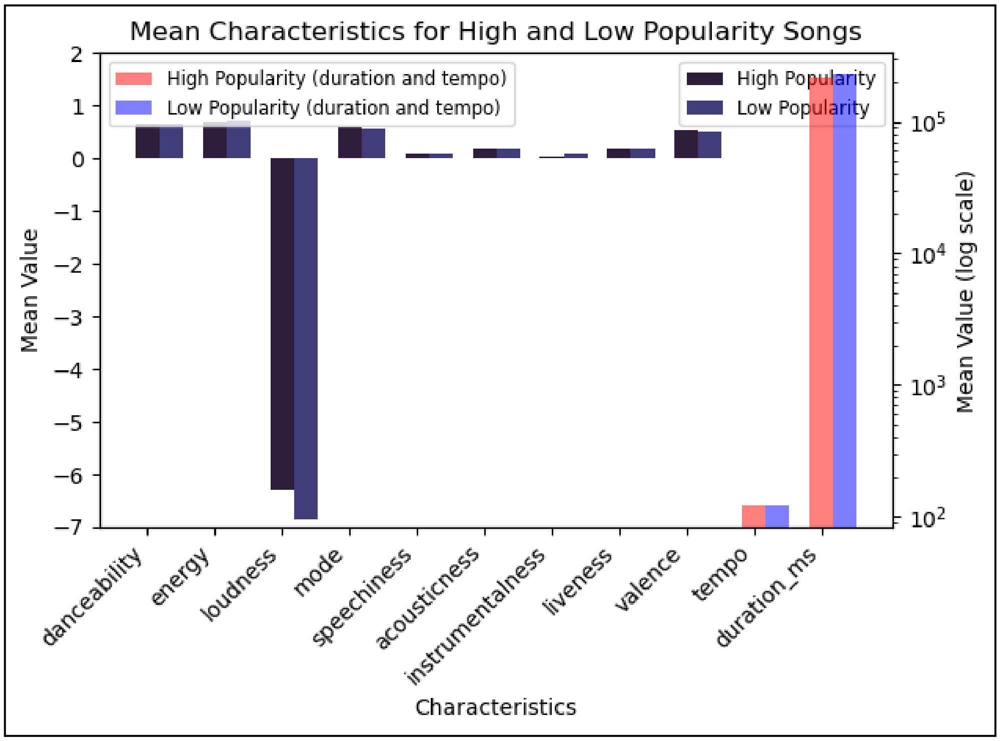
Each bar in the chart represented a subgenre, with the length of the bar corresponding to the average popularity. We added annotations to each bar to enhance clarity, displaying the count of tracks associated with the respective subgenre. This analysis offered valuable insights into the landscape of subgenre popularity. By identifying the highest average popularity subgenres and understanding their prevalence, we gained a nuanced understanding of musical trends within our dataset. These findings can inform various aspects of the music industry, from marketing strategies to playlist curation, and provide a deeper appreciation of listener preferences across different subgenres.

Genre and SubGenre Analysis

Next, we explore the characteristics of songs in two distinct ways: by comparing songs of high and low popularity and by analyzing correlations between audio features and song popularity. We first established a threshold for high popularity to differentiate high-popularity songs from low-popularity songs. This threshold was set at our dataset's 75th percentile of track popularity.

We divided the dataset into two subsets: high-popularity songs, where the track popularity exceeds the defined threshold, and low-popularity songs, where the track popularity falls below the threshold. We identified specific audio characteristics, including danceability, energy, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration, as potential influencers of song popularity. We computed the mean values of these characteristics for both high and low-popularity song subsets, as seen in Graph 3 below.

Graph 3

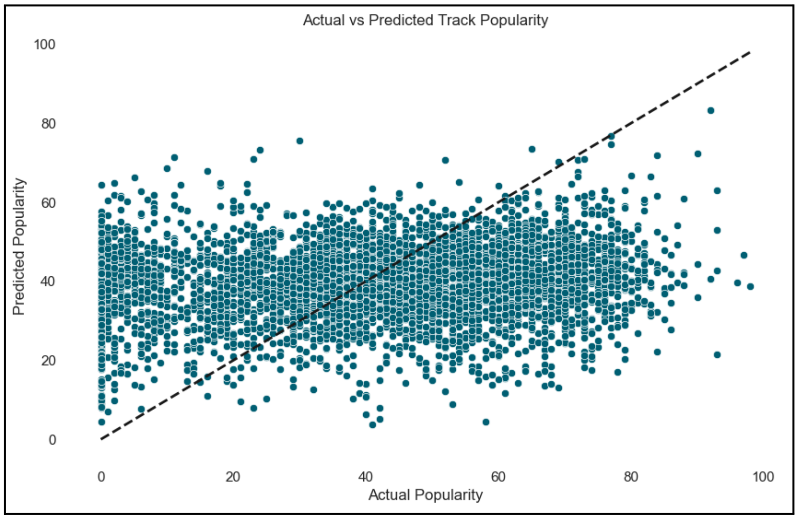


Our correlation analysis revealed that the relationships between individual audio features and song popularity are generally weak, suggesting that a song's popularity is influenced by a combination of factors not captured by individual characteristics. The weak correlations emphasize the complexity of musical preferences and the need to consider multiple attributes when determining a song's potential for popularity.

The analysis aimed to find a better model for determining the combination of variables that lead to track popularity. The initial attempt using linear regression yielded unsatisfactory results with a high Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and a low R-squared (R2) score, indicating poor model fit. The linear regression model could not capture the complex relationship between the audio features and track popularity.

Subsequently, a feature selection approach was implemented using SelectKBest with f_regression to identify the top features contributing to tracking popularity. The selected features were then used to build a new model. However, this model also resulted in a high MSE, RMSE, and a low R2 score, suggesting that feature selection alone did not significantly improve the model's performance. A Random Forest Regressor model was then employed, showing performance improvements. The Random Forest model had a lower MSE, RMSE, and a higher R2 score than the linear regression models. It demonstrated that combining selected audio features had a better predictive power for tracking popularity. The feature importance analysis in the Random Forest model highlighted the key features contributing to track popularity, with "duration_ms," "acousticness," "danceability," "energy," and "instrumentalness" being the top influential factors. In summary, the Random Forest Regressor model outperformed the linear regression models and provided better insights into the combination of features that influence track popularity. However, there is still room for improvement, especially for tracks with very high popularity. As seen in Graph 4 to the right, further model refinement and feature engineering could enhance predictive accuracy.

Graph 4



Comparative Analysis of Genre-Specific Musical Attributes

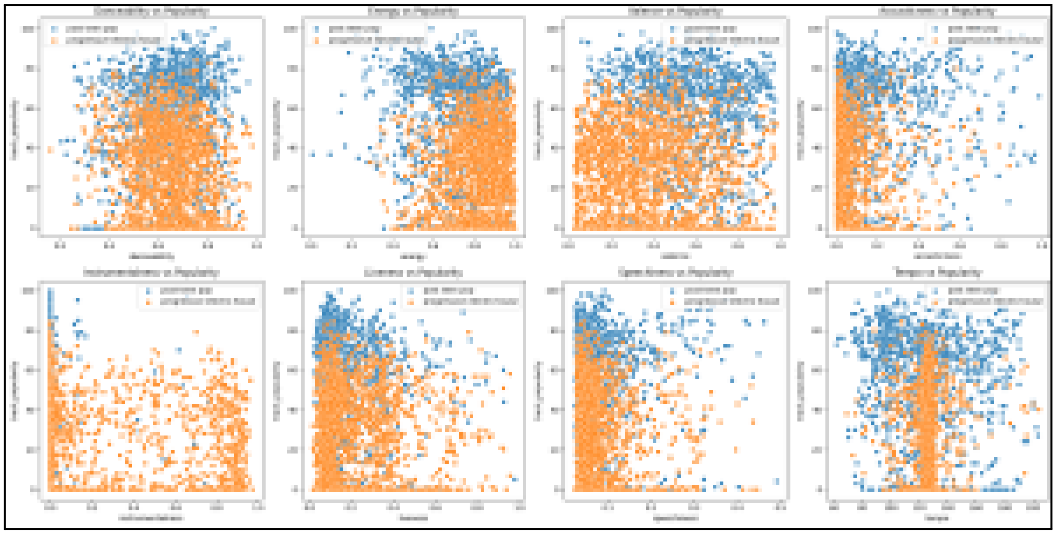
In a focused segment of our analysis, we concentrated on contrasting two significantly different genres – post-teen pop, which emerged as the most popular genre in our dataset, and progressive electro-house, identified as the least popular. This comparative study aimed to unravel how various musical attributes impact popularity within genres that are poles apart in their musical styles and listener appeal.

We began by isolating a subset of data for each genre. For post-teen pop, the most popular genre, we plotted a range of musical attributes against the popularity scores of the tracks. This exercise revealed intriguing patterns, notably in terms of tempo. We observed that tempo in post-teen pop displayed a broad spectrum of variance, indicating a diversity in the rhythm and speed of songs within this genre that resonates with its vast audience appeal.

In stark contrast, as seen in Graph 5 below, our analysis of the progressive electro-house genre provided fascinating insights despite its lower popularity. The tempo of tracks in this genre tended to cluster around a specific range, aligning with the typical characteristics expected of electro-type music. This clustering suggests a more uniform style within the genre, potentially contributing to its niche appeal.

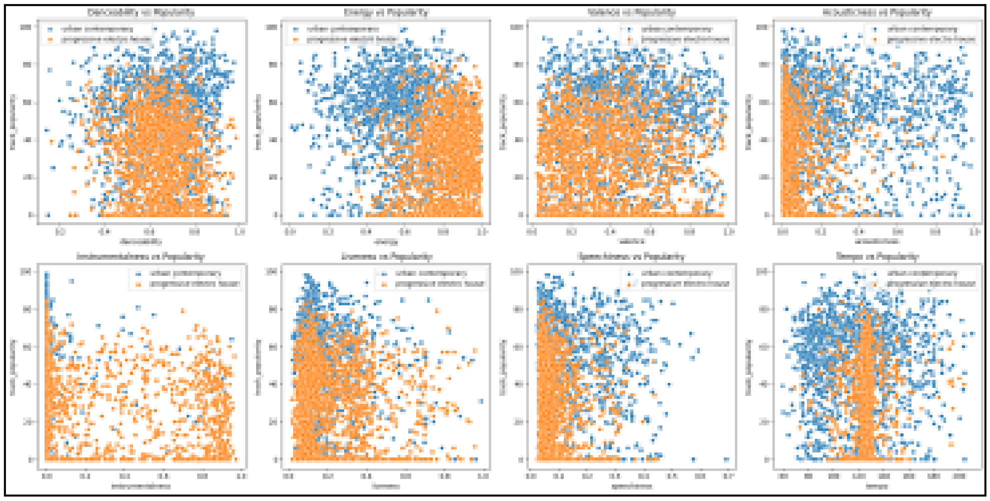
Further broadening our genre comparative analysis, we juxtaposed urban contemporary with progressive electro-house. These two genres, vastly different in their musical essence, offered a rich ground for exploring how various attributes contribute to popularity. This comparison sheds light on the nuances of genre-specific preferences, highlighting how certain musical features are favored in one genre over another. Through this detailed genre-specific analysis, we gained a deeper understanding of how different musical attributes, particularly tempo, play a crucial role in defining a genre's popularity.

Graph 5



The variance in tempo in post-teen pop, as opposed to the more uniform tempo in progressive electro house, underscores the diversity within popular genres and the specificity of less popular ones. This study segment enriched our understanding of the relationship between musical attributes and popularity. It emphasized the importance of considering genre-specific dynamics when analyzing music trends and preferences.

Graph 6



Conclusion

The deep dive into Spotify's extensive music database has provided enlightening answers to our fundamental research questions about song popularity. We embarked on this journey, aiming to unravel what makes a song popular and whether it is possible to predict a song's popularity based on its musical characteristics.

Our analysis revealed that combined musical attributes play a significant role in determining a song's popularity: duration_ms, acousticness, danceability, energy, and instrumentalness. For instance, more danceable and louder tracks were consistently favored, aligning with current trends in listener preferences, whereas factors like a key had almost no effect. This finding resonates with our initial hypothesis about the impact of certain musical features on a song's appeal.

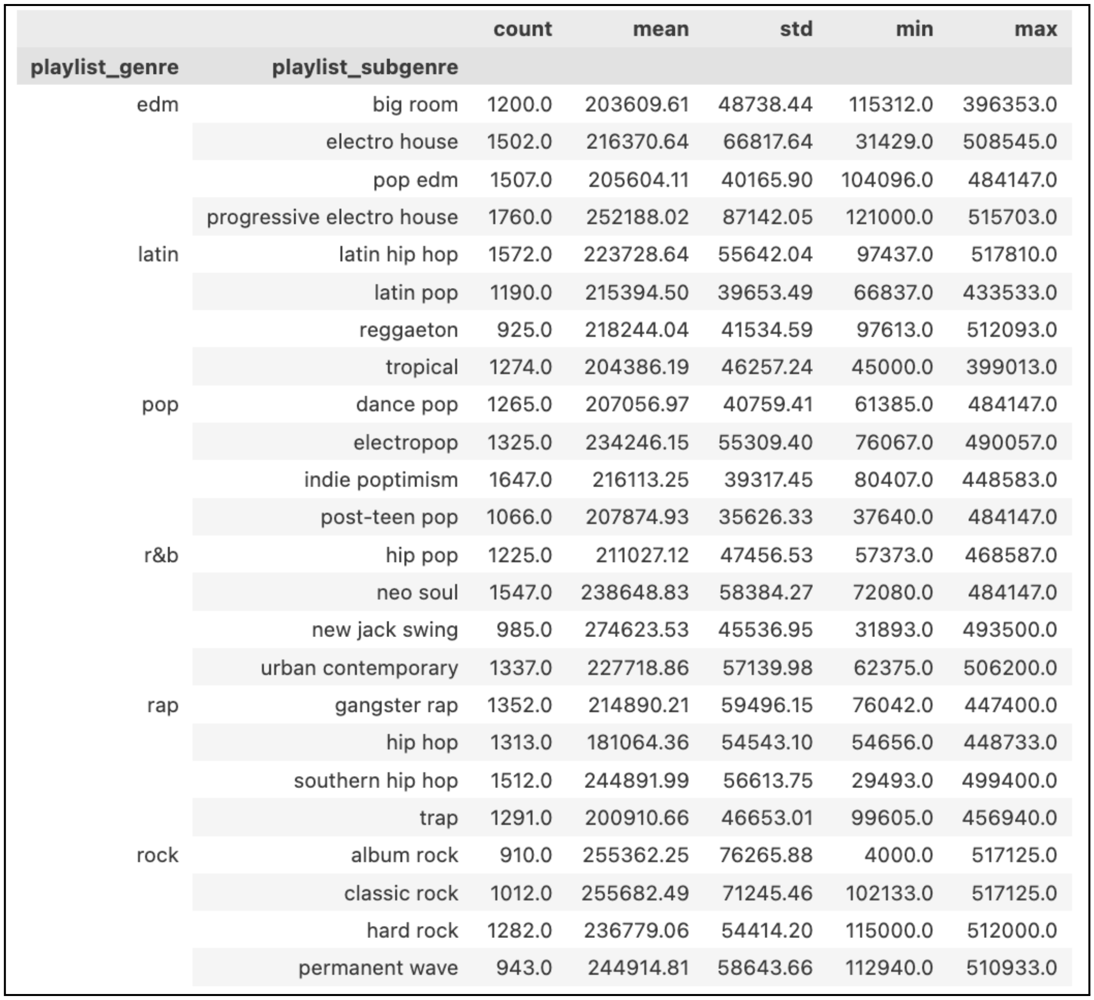
Furthermore, our exploration into genre-specific trends uncovered nuanced insights. The contrast between genres like post-teen pop and progressive electro-house highlighted how different musical attributes are valued differently across genres. This was particularly evident in the tempo variance within these genres, suggesting a correlation between genre-specific attributes and popularity.

The study also delved into predictive modeling, employing regression analysis and machine learning techniques like Random Forest Regressors. While these models provided a basis for predicting song popularity, the complex interplay of various features indicated that song popularity is influenced by a multifaceted combination of factors, not just individual characteristics.

In conclusion, our research has shed light on the intricate relationship between a song's musical characteristics and popularity. While specific attributes correlate with popularity, the comprehensive nature of music and its subjective appeal to diverse audiences make predicting song popularity a challenging, yet fascinating endeavor. Our findings offer valuable insights for artists, producers, and the music industry, emphasizing the importance of considering a blend of musical features and genre-specific trends in the quest to create popular music.

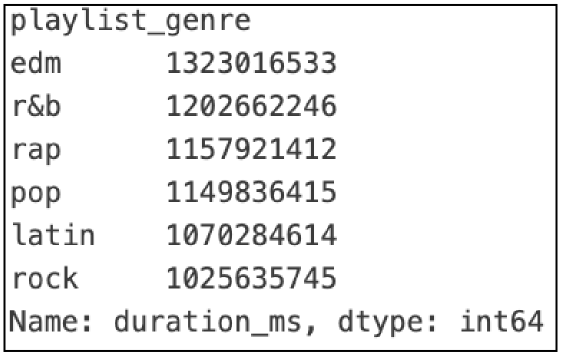
Appendix:

1.



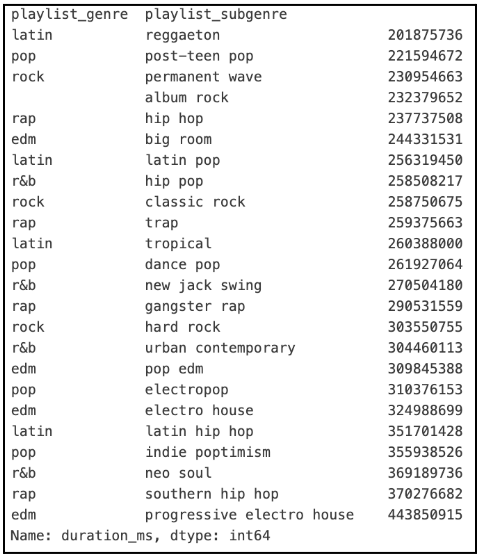
		count	mean	std	min	max
playlist_genre	playlist_subgenre					
edm	big room	1200.0	203609.61	48738.44	115312.0	396353.0
	electro house	1502.0	216370.64	66817.64	31429.0	508545.0
	pop edm	1507.0	205604.11	40165.90	104096.0	484147.0
	progressive electro house	1760.0	252188.02	87142.05	121000.0	515703.0
latin	latin hip hop	1572.0	223728.64	55642.04	97437.0	517810.0
	latin pop	1190.0	215394.50	39653.49	66837.0	433533.0
	reggaeton	925.0	218244.04	41534.59	97613.0	512093.0
	tropical	1274.0	204386.19	46257.24	45000.0	399013.0
pop	dance pop	1265.0	207056.97	40759.41	61385.0	484147.0
	electropop	1325.0	234246.15	55309.40	76067.0	490057.0
	indie poptimism	1647.0	216113.25	39317.45	80407.0	448583.0
	post-teen pop	1066.0	207874.93	35626.33	37640.0	484147.0
r&b	hip pop	1225.0	211027.12	47456.53	57373.0	468587.0
	neo soul	1547.0	238648.83	58384.27	72080.0	484147.0
	new jack swing	985.0	274623.53	45536.95	31893.0	493500.0
	urban contemporary	1337.0	227718.86	57139.98	62375.0	506200.0
rap	gangster rap	1352.0	214890.21	59496.15	76042.0	447400.0
	hip hop	1313.0	181064.36	54543.10	54656.0	448733.0
	southern hip hop	1512.0	244891.99	56613.75	29493.0	499400.0
	trap	1291.0	200910.66	46653.01	99605.0	456940.0
rock	album rock	910.0	255362.25	76265.88	4000.0	517125.0
	classic rock	1012.0	255682.49	71245.46	102133.0	517125.0
	hard rock	1282.0	236779.06	54414.20	115000.0	512000.0
	permanent wave	943.0	244914.81	58643.66	112940.0	510933.0

2.



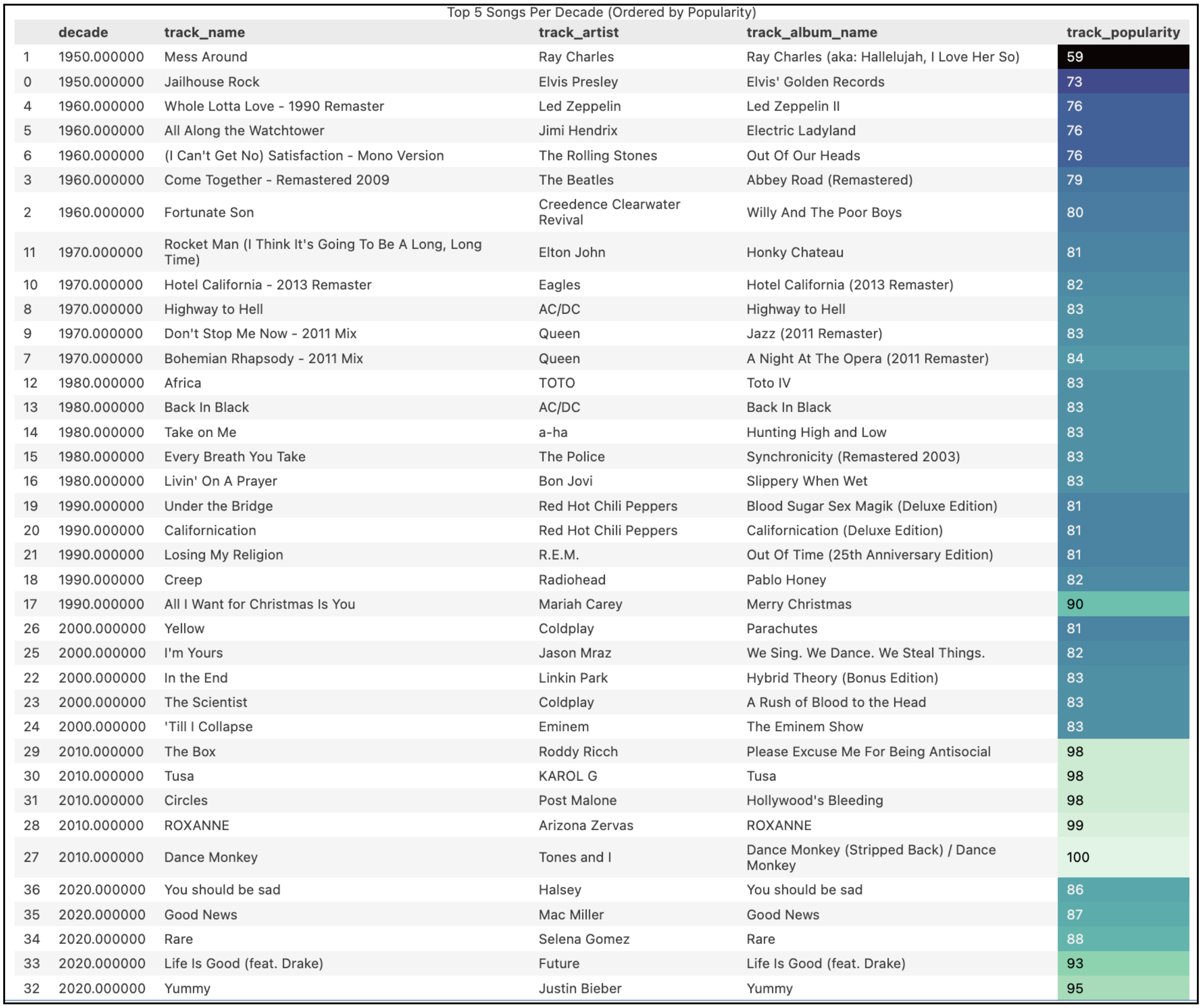
```
playlist_genre
edm      1323016533
r&b      1202662246
rap      1157921412
pop      1149836415
latin    1070284614
rock     1025635745
Name: duration_ms, dtype: int64
```

3.



```
playlist_genre  playlist_subgenre
latin           reggaeton                    201875736
pop             post-teen pop                221594672
rock            permanent wave               230954663
                album rock                   232379652
rap             hip hop                      237737508
edm             big room                     244331531
latin           latin pop                    256319450
r&b             hip pop                      258508217
rock            classic rock                 258750675
rap             trap                         259375663
latin           tropical                     260388000
pop             dance pop                    261927064
r&b             new jack swing               270504180
rap             gangster rap                 290531559
rock            hard rock                    303550755
r&b             urban contemporary           304460113
edm             pop edm                      309845388
pop             electropop                   310376153
edm             electro house                324988699
latin           latin hip hop                351701428
pop             indie poptimism              355938526
r&b             neo soul                     369189736
rap             southern hip hop             370276682
edm             progressive electro house    443850915
Name: duration_ms, dtype: int64
```

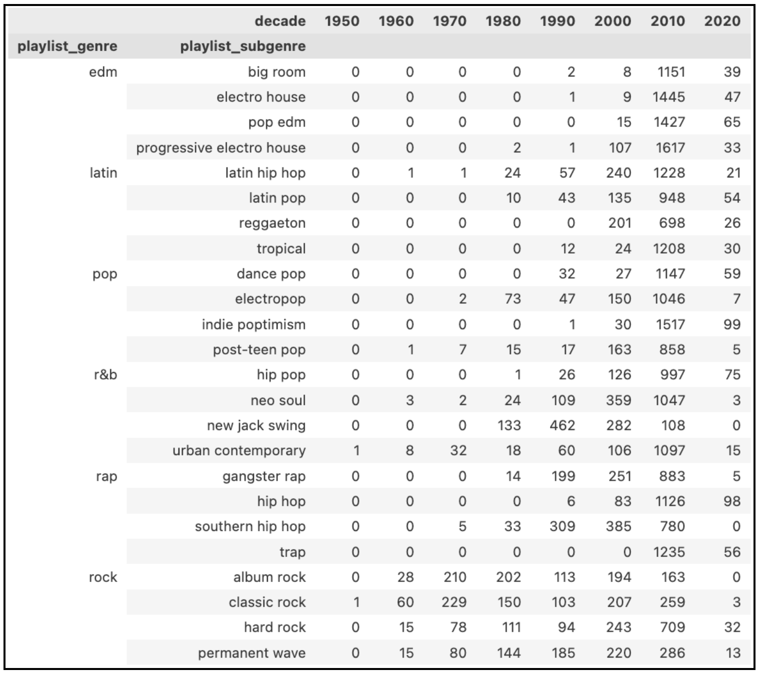
4.



Top 5 Songs Per Decade (Ordered by Popularity)

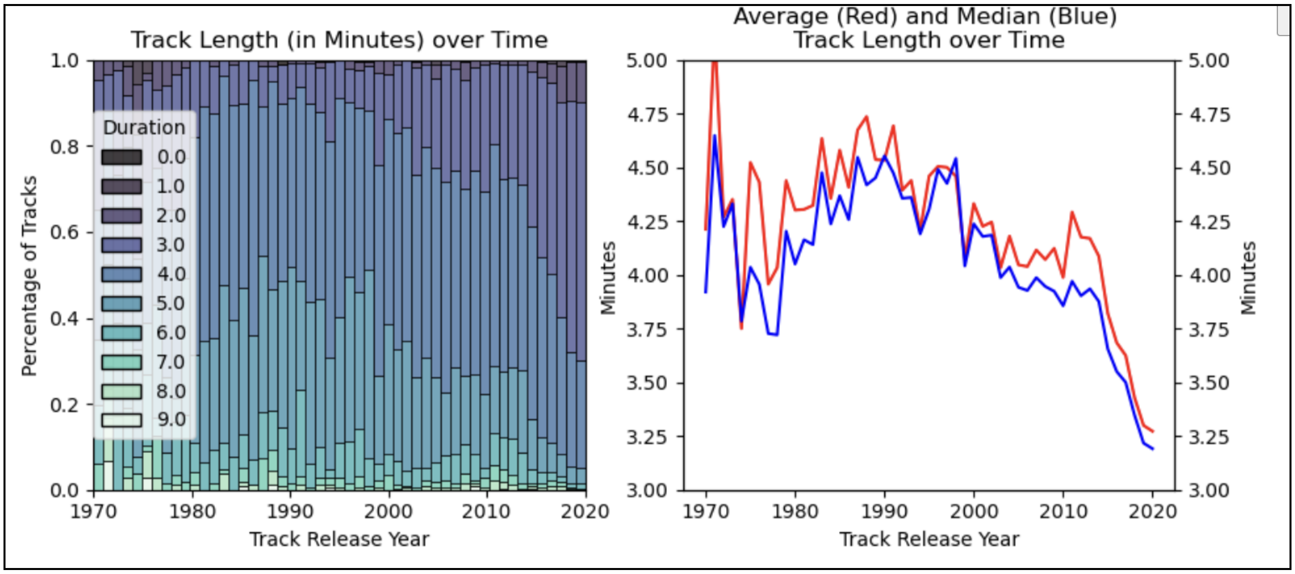
	decade	track_name	track_artist	track_album_name	track_popularity
1	1950.000000	Mess Around	Ray Charles	Ray Charles (aka: Hallelujah, I Love Her So)	59
0	1950.000000	Jailhouse Rock	Elvis Presley	Elvis' Golden Records	73
4	1960.000000	Whole Lotta Love - 1990 Remaster	Led Zeppelin	Led Zeppelin II	76
5	1960.000000	All Along the Watchtower	Jimi Hendrix	Electric Ladyland	76
6	1960.000000	(I Can't Get No) Satisfaction - Mono Version	The Rolling Stones	Out Of Our Heads	76
3	1960.000000	Come Together - Remastered 2009	The Beatles	Abbey Road (Remastered)	79
2	1960.000000	Fortunate Son	Creedence Clearwater Revival	Willy And The Poor Boys	80
11	1970.000000	Rocket Man (I Think It's Going To Be A Long, Long Time)	Elton John	Honky Chateau	81
10	1970.000000	Hotel California - 2013 Remaster	Eagles	Hotel California (2013 Remaster)	82
8	1970.000000	Highway to Hell	AC/DC	Highway to Hell	83
9	1970.000000	Don't Stop Me Now - 2011 Mix	Queen	Jazz (2011 Remaster)	83
7	1970.000000	Bohemian Rhapsody - 2011 Mix	Queen	A Night At The Opera (2011 Remaster)	84
12	1980.000000	Africa	TOTO	Toto IV	83
13	1980.000000	Back In Black	AC/DC	Back In Black	83
14	1980.000000	Take on Me	a-ha	Hunting High and Low	83
15	1980.000000	Every Breath You Take	The Police	Synchronicity (Remastered 2003)	83
16	1980.000000	Livin' On A Prayer	Bon Jovi	Slippery When Wet	83
19	1990.000000	Under the Bridge	Red Hot Chili Peppers	Blood Sugar Sex Magik (Deluxe Edition)	81
20	1990.000000	Californication	Red Hot Chili Peppers	Californication (Deluxe Edition)	81
21	1990.000000	Losing My Religion	R.E.M.	Out Of Time (25th Anniversary Edition)	81
18	1990.000000	Creep	Radiohead	Pablo Honey	82
17	1990.000000	All I Want for Christmas Is You	Mariah Carey	Merry Christmas	90
26	2000.000000	Yellow	Coldplay	Parachutes	81
25	2000.000000	I'm Yours	Jason Mraz	We Sing. We Dance. We Steal Things.	82
22	2000.000000	In the End	Linkin Park	Hybrid Theory (Bonus Edition)	83
23	2000.000000	The Scientist	Coldplay	A Rush of Blood to the Head	83
24	2000.000000	'Till I Collapse	Eminem	The Eminem Show	83
29	2010.000000	The Box	Roddy Ricch	Please Excuse Me For Being Antisocial	98
30	2010.000000	Tusa	KAROL G	Tusa	98
31	2010.000000	Circles	Post Malone	Hollywood's Bleeding	98
28	2010.000000	ROXANNE	Arizona Zervas	ROXANNE	99
27	2010.000000	Dance Monkey	Tones and I	Dance Monkey (Stripped Back) / Dance Monkey	100
36	2020.000000	You should be sad	Halsey	You should be sad	86
35	2020.000000	Good News	Mac Miller	Good News	87
34	2020.000000	Rare	Selena Gomez	Rare	88
33	2020.000000	Life Is Good (feat. Drake)	Future	Life Is Good (feat. Drake)	93
32	2020.000000	Yummy	Justin Bieber	Yummy	95

5.

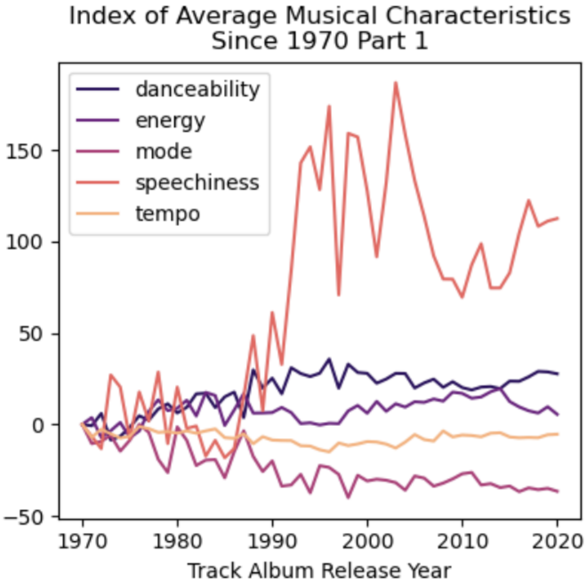


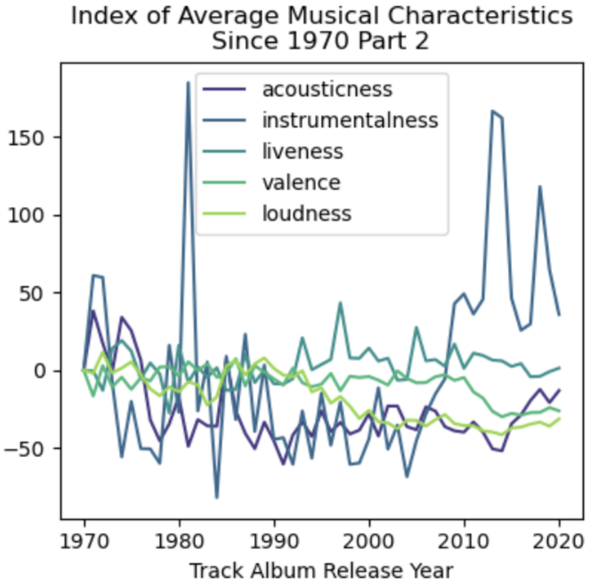
	decade	1950	1960	1970	1980	1990	2000	2010	2020
playlist_genre	playlist_subgenre								
edm	big room	0	0	0	0	2	8	1151	39
	electro house	0	0	0	0	1	9	1445	47
	pop edm	0	0	0	0	0	15	1427	65
	progressive electro house	0	0	0	2	1	107	1617	33
latin	latin hip hop	0	1	1	24	57	240	1228	21
	latin pop	0	0	0	10	43	135	948	54
	reggaeton	0	0	0	0	0	201	698	26
	tropical	0	0	0	0	12	24	1208	30
pop	dance pop	0	0	0	0	32	27	1147	59
	electropop	0	0	2	73	47	150	1046	7
	indie poptimism	0	0	0	0	1	30	1517	99
	post-teen pop	0	1	7	15	17	163	858	5
r&b	hip pop	0	0	0	1	26	126	997	75
	neo soul	0	3	2	24	109	359	1047	3
	new jack swing	0	0	0	133	462	282	108	0
	urban contemporary	1	8	32	18	60	106	1097	15
rap	gangster rap	0	0	0	14	199	251	883	5
	hip hop	0	0	0	0	6	83	1126	98
	southern hip hop	0	0	5	33	309	385	780	0
	trap	0	0	0	0	0	0	1235	56
rock	album rock	0	28	210	202	113	194	163	0
	classic rock	1	60	229	150	103	207	259	3
	hard rock	0	15	78	111	94	243	709	32
	permanent wave	0	15	80	144	185	220	286	13

6.



7.


Index of Average Musical Characteristics Since 1970 Part 1


Index of Average Musical Characteristics Since 1970 Part 2