Spotify Clarifies

An Analysis of Music Popularity over Time

Britney Brown, Harrison DiStefano, Greg Eastman
Lisa Kaunitz, Tianyang Liu, Jeremy Weidner

06/04/2021

Contents

Abstract

Write something

Introduction

REWRITE TO DESCRIBE QUESTION OF INTEREST

We analyze a database of Spotify tracks that have had attributes assigned to them algorithmically by Spotify. These attributes act as descriptors for each track and allow us to group similar tracks and identify potentially interesting information and trends in the data. We have used this to analyze the songs in the Spotify database for what we found to be the most interesting questions. In the following sections we will cover the data more in depth, describe our methodology for exploring the data, and present our final analyses.

Data Description

We conducted our analysis using a dataset compiling various music features for songs found on Spotify's Web API[1]. This data was obtained by sampling two different subsets of songs, hits and flops, from 1960-2019. Hits are defined as any song that made a top 100 'hit' list any week in a given decade. Flops are defined as the opposite, with additional requirements such as belonging to a genre considered non-mainstream or avant-garde. While the original purpose of this dataset was to predict whether or not a song will be successful, we will be utilizing the musical attributes to analyze the change in music over time. Since this dataset is stratified at the track level, each track acts as a unique unit of observation. With over 40,000 unique songs spread evenly across six decades, it provides plenty of information to analyze the changes in musical attributes across the last several years.

The majority of these variables are generated by Spotify's music analytic algorithm, which is unknown to the general public. Therefore, we have provided documentation to get a fairly robust understanding of the variables. The definitions for all our variables can be found in the attached JSON file as well as in the dataDictionary table when connected to the database instance.

Exploratory Data Analysis

Top Artists per Decade

We begin our exploratory data analysis by checking who the top 5 artists or bands are in each decade. For the purpose of this analysis over time, the top artists are defined in terms of total amount of songs created.

Table 1: Top 5 Artist/Bands with the Most Songs per Decade

Artist	Decade	Total Number of Songs
Traditional	60s	145
P. Susheela	60s	130
Jerry Goldsmith	60s	118
Harry Belafonte	60s	114
Ennio Morricone	60s	96
MPB4	70s	59
Buzzcocks	70s	50
kalapana	70s	49
John Coltrane	70s	44
Vicente Fernández	70s	44
The Cleaners From Venus	80s	47
Malcolm Arnold	80s	45
Nobuo Uematsu	80s	33
Skinny Puppy	80s	33
Running Wild	80s	32
Luis Miguel	90s	28
Madonna	90s	25
Iggy Pop	90s	22
Nobuo Uematsu	90s	19
El Gran Combo De Puerto Rico	90s	17
Toby Keith	00s	27
Tim McGraw	00s	24
Rascal Flatts	00s	24
Iron Maiden	00s	23
Kenny Chesney	00s	23
Drake	10s	50
Glee Cast	10s	41
Taylor Swift	10s	35
Luke Bryan	10s	25
The Weeknd	10s	24

Music Theory Attributes

To get a better understanding of some of the music-related variables in our dataset, we will also explore three very important music theory concepts: tempo, time signature, and key signature.

Tempo

Tempo describes the speed/pace of a song. In general, fast tempos evoke more positive emotions such as happiness and delight while slow tempos evoke negative emotions such as sadness and depression. To get an idea of the artists that tend to convey postive emotions compared to those that channel negative emotions, we will take a look at the top 10 artists with the fastest average tempos as well as the top 10 with the slowest.

Table 2: Top 10 Artist/Bands with the Fastest Average Tempos

Artist	Average Song Tempo	Total Number of Songs
Crass	172.8562	15
Angerfist	157.4122	13
The Lurkers	156.5629	19
Allison	156.0839	10
The Vibrators	150.0379	10
The Dickies	149.5186	19
Avril Lavigne	147.3464	15
Foo Fighters	147.1191	10
Trifonic	146.2458	11
The Nashville Bluegrass Band	145.9536	10

Time Signature

Time signature indicates the rhythm of the song in terms of a beat's duration and the number of beats per measure. While it can get fairly technical, it is important to know that the most common time signature is 4 beats per measure (and is hence referred to as 'common time'). In fact, 88% of our dataset is made up of songs with common time. Compared to the other time signatures, common time has higher averages for danceability and energy but lower averages for acousticness. This time signature also has the highest success rate for hit songs with 43% of common time songs getting classified as a hit. As seen from the average rates in the table below, the success behind this time signature can be attributed to these songs being easier for people to dance to. Therefore, this table indicates that successful hit songs tend to higher levels of dancibility and energy.

Table 3: Summary of Time Signature Characteristics

Time Signature	4.000000e+00	3.0000000	5.0000000	1.0000000	0.0000000
Average Danceability	5.579703e-01	0.4049516	0.3993724	0.4014569	0.1334000
Average Acousticness	3.308935e-01	0.6177392	0.6153970	0.6211903	0.5906667
Average Energy	6.058251e-01	0.3784532	0.3824055	0.3927565	0.4109940
Number of Tracks	0.000000e+00	0.0000000	0.0000000	0.0000000	0.0000000
Number of Hits	1.569300e+04	973.0000000	117.0000000	61.0000000	1.0000000
Percentage of Hit Tracks	4.345000e-01	0.2555000	0.1993000	0.1644000	0.3333000

Key Signature

Key signature describes the combination of sharps and flats that determine the scale of a piece of music. In general, most songs shift from key to key to add variance and intrigue but this dataset simply estimates the overall key. As the table below indicates, the estimated key does not have a large effect on the success rate of songs since the 'Percentage of Hit Tracks' varies around 40% for each key.

Table 4: Summary of Key Signature Characteristics

	Key Signature	Average Danceability	Average Acousticness	Average Energy	Percentage of Hit Tracks
1	C	0.5391916	0.3912941	0.5613325	0.4078
7	C#/Db	0.5757106	0.2853243	0.6194482	0.4621
3	D	0.5225773	0.3628475	0.5817457	0.3876
12	D#/Eb	0.5051904	0.4968816	0.5100628	0.3822
6	E	0.5217508	0.3590309	0.5866685	0.3873
5	F	0.5329451	0.4307504	0.5385227	0.4084
11	F#/Gb	0.5568206	0.2918481	0.6267238	0.4417
2	G	0.5399274	0.3730568	0.5711438	0.4018
10	G#/Ab	0.5477378	0.3751159	0.5693247	0.4398
4	A	0.5261597	0.3525814	0.5924012	0.3834
9	A#/Bb	0.5593870	0.3920532	0.5595066	0.4569
8	B	0.5607593	0.2833212	0.6295437	0.4240

Methodology

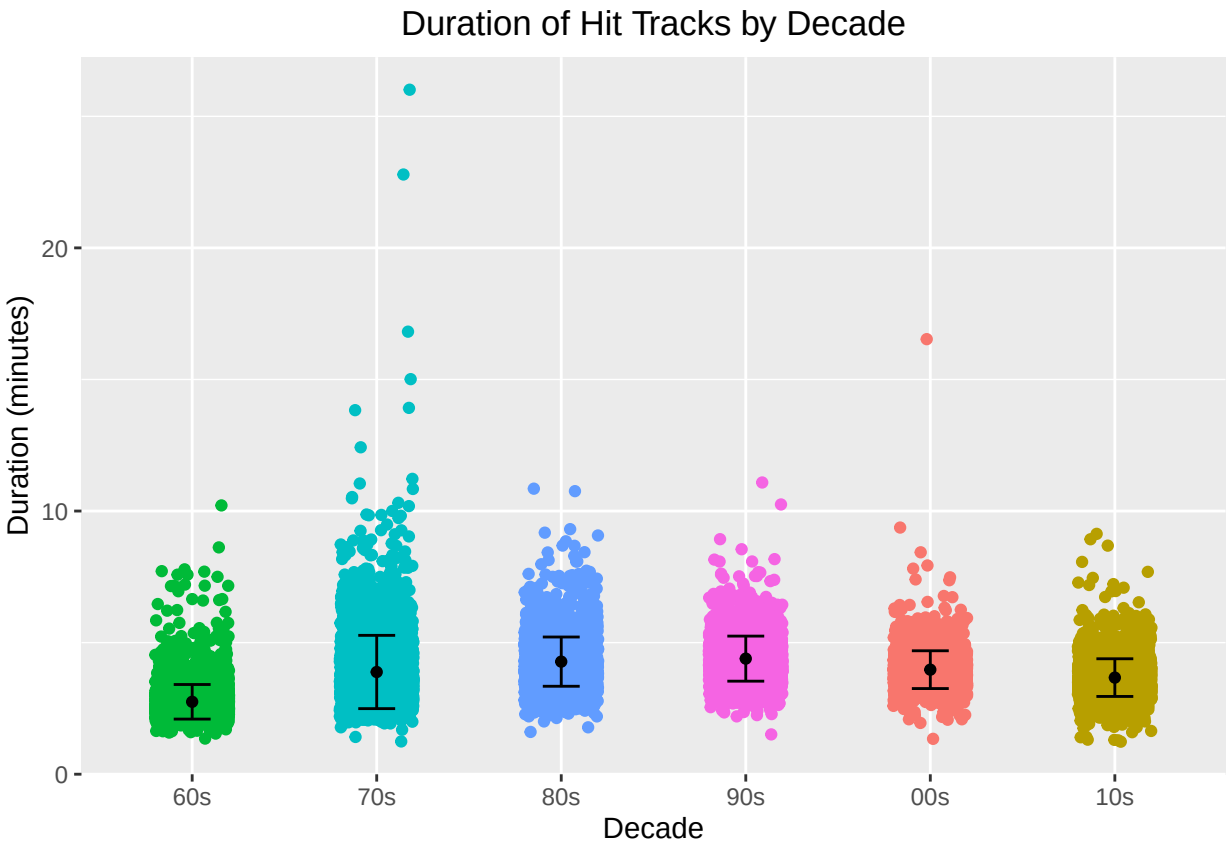
This work uses a variety of analytical techniques to explore the data and answer small questions about the music we all like and listen to. First we will display and discuss what terms appear most often in music titles. The idea of specific pictures and words in titles to get views has become commonplace in online videos and likely is no different in music as well. Although we could note simply display the most common words, the text needed to be cleaned first. Therefore we use the tm[2] package and tidytext[3] to clean the song titles. We made all text lowercase, removed non-alphabet characters, removed all punctuation, removed stop words, and then stemmed the document. Stop words exist simply as determiners to mark nouns like "the' ', or coordinating conjunctions like "but", or prepositions like "in". Although these terms help to form a language, they do not hold any meaning for our song titles. Therefore, we removed them. After that we stemmed the text. This process altered words to help make them return to their roots. An example in our data was changing "remastered", "remaster", and "remastering" to "remast". This allowed us to see each term as the same. Once these steps were completed, we used a word cloud. This was only the first method to learn about the data. Our next method involved multiple sets of univariate analysis. Investigating how the data was distributed could help reinforce the prevalence of musical standards. To do this analysis we used tables to help get a concrete understanding of the bins. In categorical variables this allows us not only to see how the bins are filled, but to immediately compare them to their neighbors. Our final type of analysis is bivariate. We used a visual analysis to see how music changed over the decades and when more sections get added. [I NEED TO SEE WHAT PEOPLE TO TALK ABOUT THIS MORE. I ALSO DON'T KNOW HOW LISA IS TACKLING THE DANCEABILITY STUFF SO THAT WILL NEED TO BE ADDED LATER]

Analysis

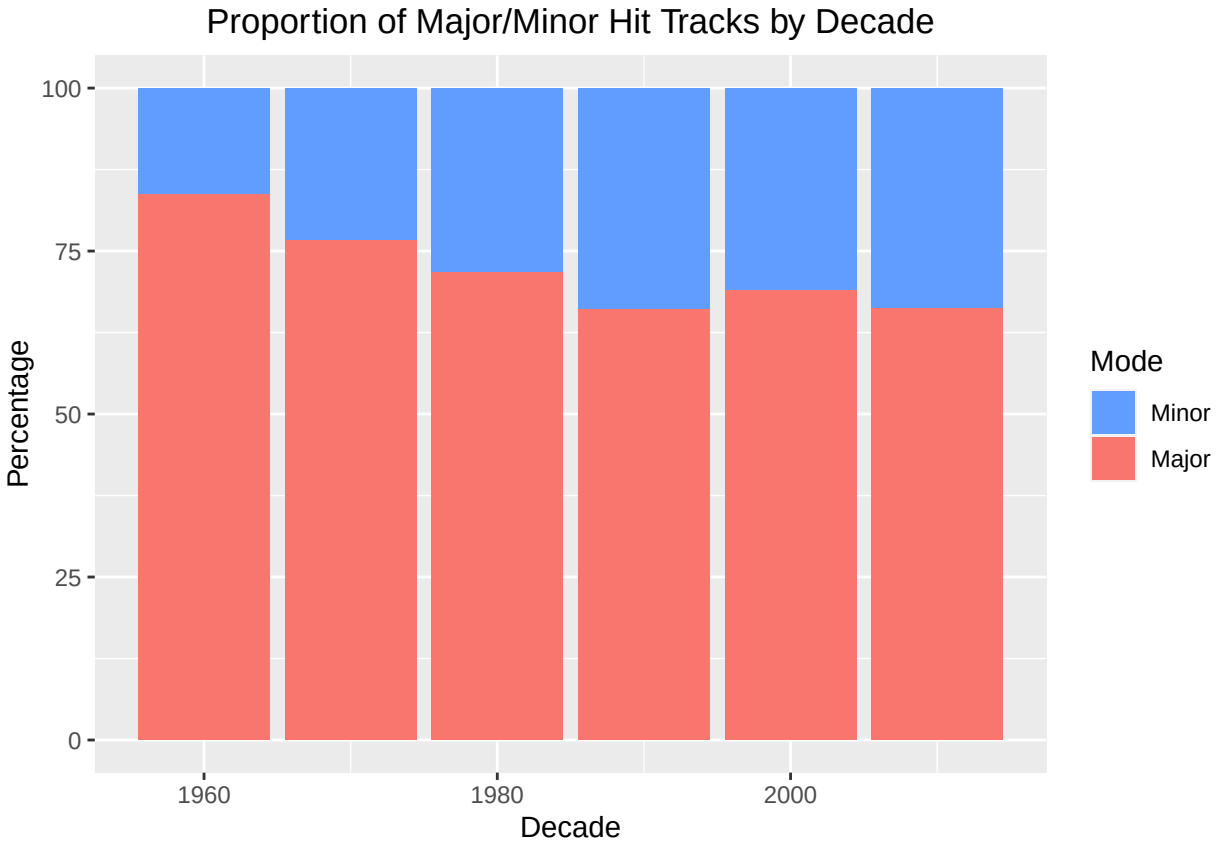
Like music our analysis is multifaceted, but we will begin by looking at the titles of songs. The way that important words were isolated has been discussed, so after cleaning the data we used the wordcloud2[4] package to create the word star below.

There are a number of takeaways from the above plot, but we will start by asking a decades old question, what has love got to do with it? Well it appears love has a lot to do with titles, this word appears over 2000 times, more than the next two most common words combined. This makes sense as it is both a powerful emotion, as well as its own genre in a way. In fact, we see that love songs are not the only emotional genre listed. The word blue, or the blues is often a sadder style of music named after the emotion. This leads us to note that more emotive phrases are popular in titles. Next we note that there is another topic besides feelings that are common in titles.

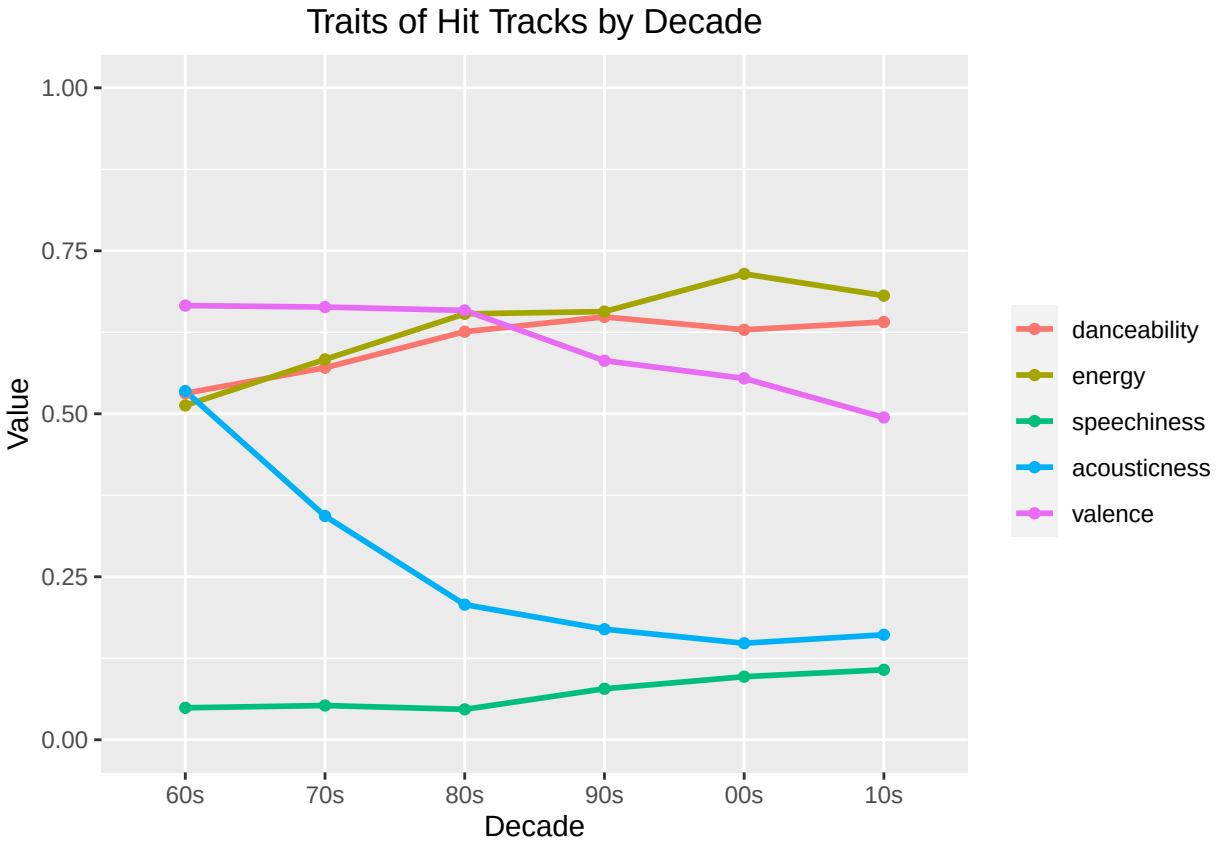
Descriptions of people appear to be the second most common terms in a title. This is reflected by the number of times a title has “man”, “girl”, “babi”, or “boy”. Often people like to talk about themselves and each other, this seems to hold true for our music as well. So when title a piece it seems that the most common practices are to talk about how the artist feels and their relationships to others. Now that we looked at names and titles we will move on to explore the actual music itself.



The duration of hit tracks has mostly stayed under 10 minutes with an average closer to 4 minutes across the decades. The notable exception is in the 1970s, possibly due to the popularity of Progressive Rock during those years with its blending of Rock and Jazz Fusion into long drawn out concept albums and longer-winded tracks.



The proportion of hit songs in the minor mode has increased from about 15% in the 1960s to almost 30% in the 2010s. The minor mode generally sounds more sad, while the major has a happier sound–this may indicate that people have developed more of a taste for sad popular music over the years.



There are a few noticeable trends in the qualities of hit music across the decades as assigned by Spotify. The use of acoustic as opposed to electric instruments of tracks declined significantly from the 60's onward, as more and more electronic instruments were developed and put to widespread use. Danceability, energy, and speechiness (the use of voice) also saw a general upward trend. However valence, or the general noisiness of tracks, has been trending downward since the 80's.

Conclusions

Bibliography

1. FortyTwo102. "Spotify and Billboard Top Hits Data." GitHub. Jan. & Feb., 2020. Accessed May 27, 2021. https://github.com/fortyTwo102/The-Spotify-Hit-Predictor-Dataset.

2. Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54. URL: https://www.jstatsoft.org/v25/i05/.

3. Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

4. Dawei Lang and Guan-tin Chien (2018). wordcloud2: Create Word Cloud by 'htmlwidget'. R package version 0.2.1. https://CRAN.R-project.org/package=wordcloud2