Spotify Track Feature Analysis 1921-2020

By Shuki Santana-Molk

1. Introduction

For many years, since the start of the commercial internet, the music industry has struggled to find its place in that world.

It seemed that in a world where file sharing services allow people to download music, regardless of the legality of its source, the music industry is obsolete.

Starting the 2000's, various services have tried to resolve this problem by offering music services of various technologies and financial models.

Services like Deezer (est. 2007), Pandora (est. 2000), Spotify (est. 2008), iTunes (est.2001) and Apple Music (est. 2015) offer users a vast diversity of legal music, for no or a very small charge.

As of the beginning of 2020, Spotify holds the biggest market share (32%) of all streaming services.

Source: https://www.midiaresearch.com/blog/music-subscriber-market-shares-q1-2020

As part of its activity, Spotify allows users to mine data using API service, to better understand its features, and even allow the capable users to design a personalized algorithm that would "beat" the Spotify algorithm and predict which songs will fit their musical taste.

In this Analysis however, we are going to focus on general features of songs from 1921 to 2020, and try to determine if any trends or insights can be extracted from the database.

2. The database

The database is taken from the "Spotify Dataset 1921-2020, 160k+ Tracks" dataset that can be found here:

https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

Direct download (~38 MB):

https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks/download

In order to better understand the characteristics Spotify assigns to each track, please read this Audio Features Documentation and Track Features Documentation.
Please notice - Not every feature is included in our database.

3. Dataset description:

1. The dataset consists of 5 files: 1 containing raw data (data.csv) and 4 other files containing aggregated data. We will be using only the data.csv as the database for this research.
2. The database consists of a shy of 170K rows and of 19 columns that describe tracks (songs - see 3.4) in different manners:
   - General data i.e. 'name', 'id' (Spotify unique key for each track), 'artists'.
   - Measurable technical data i.e. 'duration', 'tempo', 'release date', 'key', 'loudness'.
   - Soft and mutable features data i.e. 'acousticness', 'liveness', 'valence', 'popularity'.
3. There are no NULLs in the database.
4. The database contains information about "tracks", because not all tracks are songs. For example, some tracks contain "white noise" for meditation.
   Throughout this paper, we will be using the terms "track" and "song" interchangeably.

4. Data preparation

1. Reducing the file size:

   Firstly, only essential columns are loaded: as we are going to be looking at trends, not analyzing particular songs, columns like 'name' (of the track), 'id' or 'artists' are irrelevant.

   Secondly, as the file is based on Spotify data, it is already well constructed in the sense of having all data stored in a numerical way (where applicable) i.e. 'keys' are represented as numbers - 0 instead of C, 1 instead of C#, 2 instead of D etc.

   Boolean columns i.e. 'mode' or 'explicit' are also stored as o and 1.

   So no conversion is needed. However, the default of Pandas "read_csv" function, imports all integers and floats as int64 and float64, while for our needs, smaller data types are sufficient. Hence, after reviewing the needs for each column, smaller data types are used.

   The above includes converting the duration from milliseconds to seconds.

   By conducting all these downsizing, we manage to reduce the database to less than 20% of its original size.

2. Rearranging the columns:

The columns in the database are sorted alphabetically, we change this order to match a more logical order of measurable data first and only then soft and mutable data.

5. Analysis description.

Data was analyzed and visualized in Python, using Jupyter Notebook. The notebook is annexed to this report.

The study is divided into 4 research questions:

1. General overview and exceptional phenomena.
2. The way different songs' characters changed over the years:
   - 2.1. Explicit songs - **Extracting insights from rogue data**
   - 2.2. Length
   - 2.3. Loudness (is The Loudness War a real thing?)
   - 2.4. Tempo
   - 2.5. Speechiness - explicit correlation
3. Popularity of musical keys.
4. What makes a song popular over different time periods.

5.1. General overview and exceptional phenomena

Firstly, we conduct a quick summary of all the main statistical attributes for each feature. This can reveal some basic insights regarding our database.

While some statistical analysis doesn't make too much sense over some of the features (i.e. average year or average key), other do.

For example, the ‘mode’ feature which represents if a song is in a minor or a major key - only has two options (0 for minor or 1 for major). By calculating its average (0.708556) we actually get the percentage of major songs (and by deduction - the percentage of minor songs) of the total number of songs. In this case, about 7 out of 10 songs are in a major key.

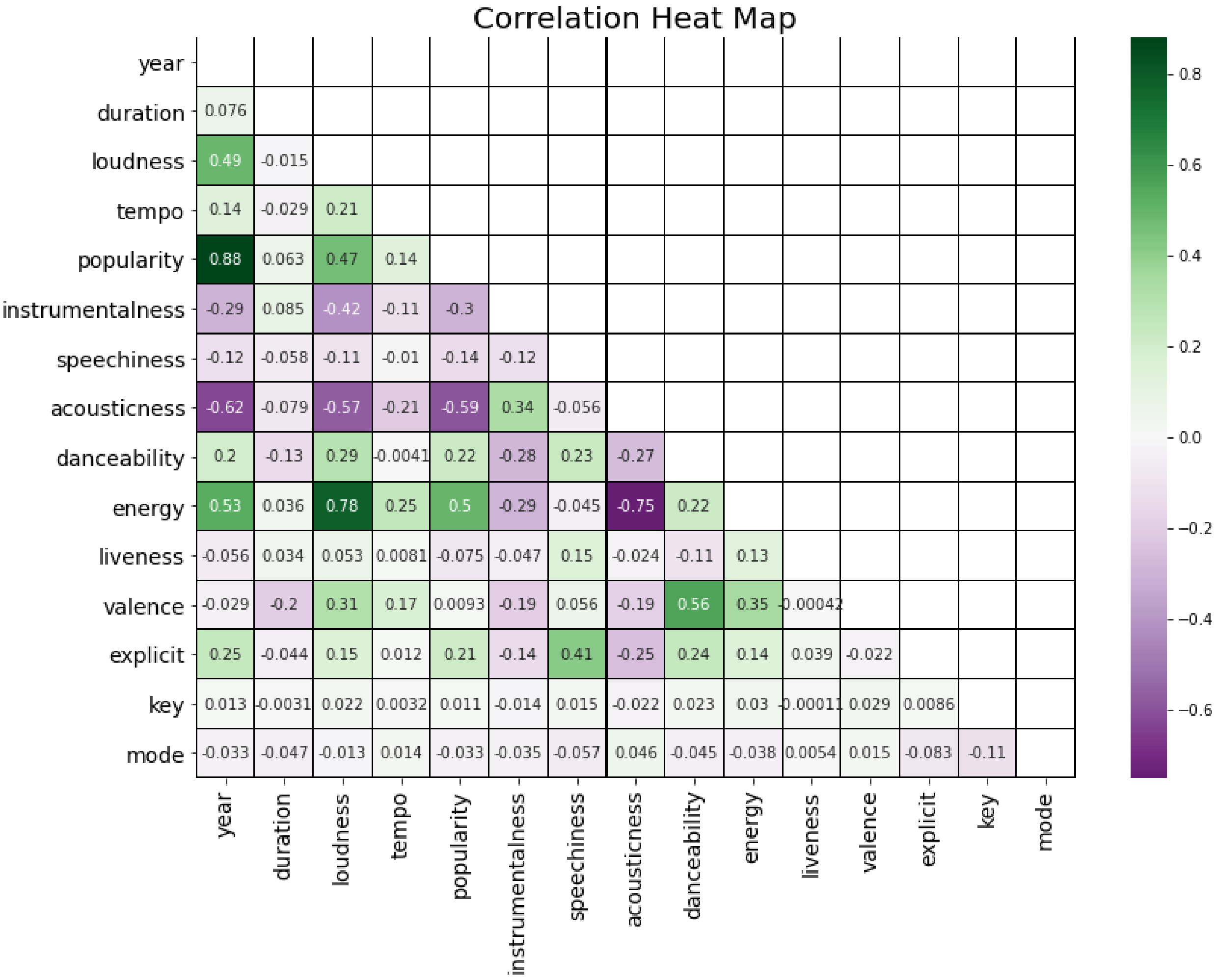
Another example - most songs are not very instrumental - by the 70% percentile we still have only 0.0239 instrumentalness rate, only when we reach the 90% percentile we get 0.831 instrumentalness rate.

	year	duration	loudness	tempo	popularity	instrumentalness
count	169909	169909	169909	169909	169909	169909
mean	1977.223	231.404		116.525		0.161987
std	25.593	121.320	0	30.744	0	0.309082
min	1921	5	-60	0	0	0
10%	1942	137	-18.9219	79	0	0
30%	1961	179	-13.5078	97	19	2.03E-06
50%	1978	209	-10.4766	114	33	0.000204
70%	1995	250	-7.78125	130	45	0.023895
90%	2012	336	-5.01563	161	60	0.831055
max	2020	5404	3.855469	244	100	1

	speechiness	acousticness	danceability	energy	liveness
count	169909	169909	169909	169909	169909
mean	0.094055				0.206665
std	0.14978	0	0	0	0.176758
min	0	0	0	0	0
10%	0.030106	0.009033	0.298096	0.130981	0.072998
30%	0.036591	0.147949	0.447021	0.303955	0.10498
50%	0.045013	0.491943	0.547852	0.480957	0.13501
70%	0.064575	0.831055	0.641113	0.663086	0.223022
90%	0.192017	0.98291	0.759766	0.867188	0.407959
max	0.969238	0.996094	0.987793	1	1

	valence	explicit	key	mode
count	169909	169909	169909	169909
mean		0.084863	5.200519	0.708556
std	0	0.278679	3.515257	0.454429
min	0	0	0	0
10%	0.159058	0	0	0
30%	0.366943	0	2	1
50%	0.543945	0	5	1
70%	0.708008	0	8	1
90%	0.887207	0	10	1
max	1	1	11	1

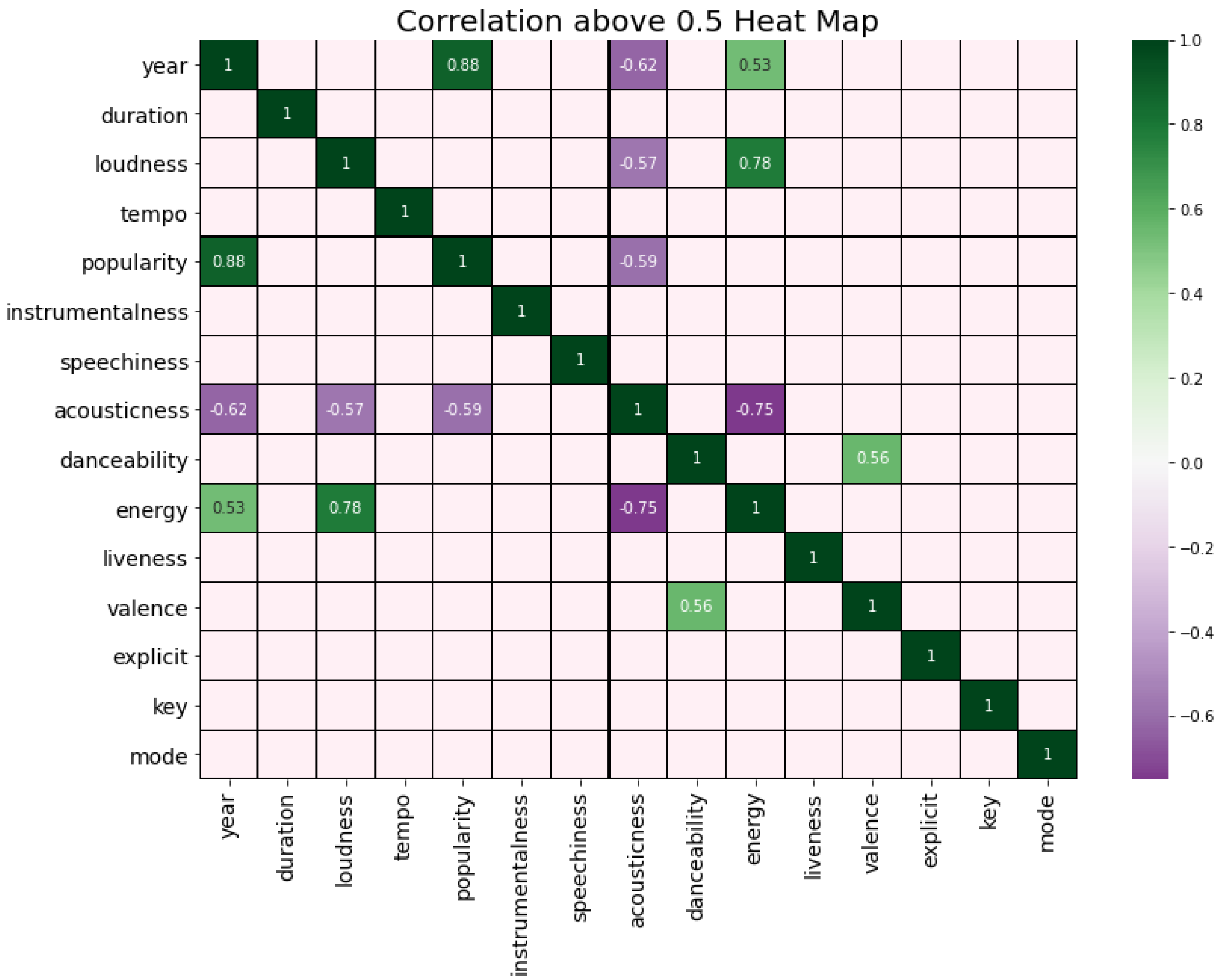
Another elementary approach to analyzing data is by examining the correlations between the different features.



The stronger the correlation is between 2 features, the darker the color gets.

We can see most correlations are quite light, meaning - weak correlations.

Let us see only correlations higher than 0.5.
Notice that correlation can be positive (The higher one feature goes, so does the other), or negative (When one feature is higher, the other is lower and vice versa). Thus, a correlation stronger than 0.5 is also a correlation lower than (-0.5).



Below are the correlations in descending order:

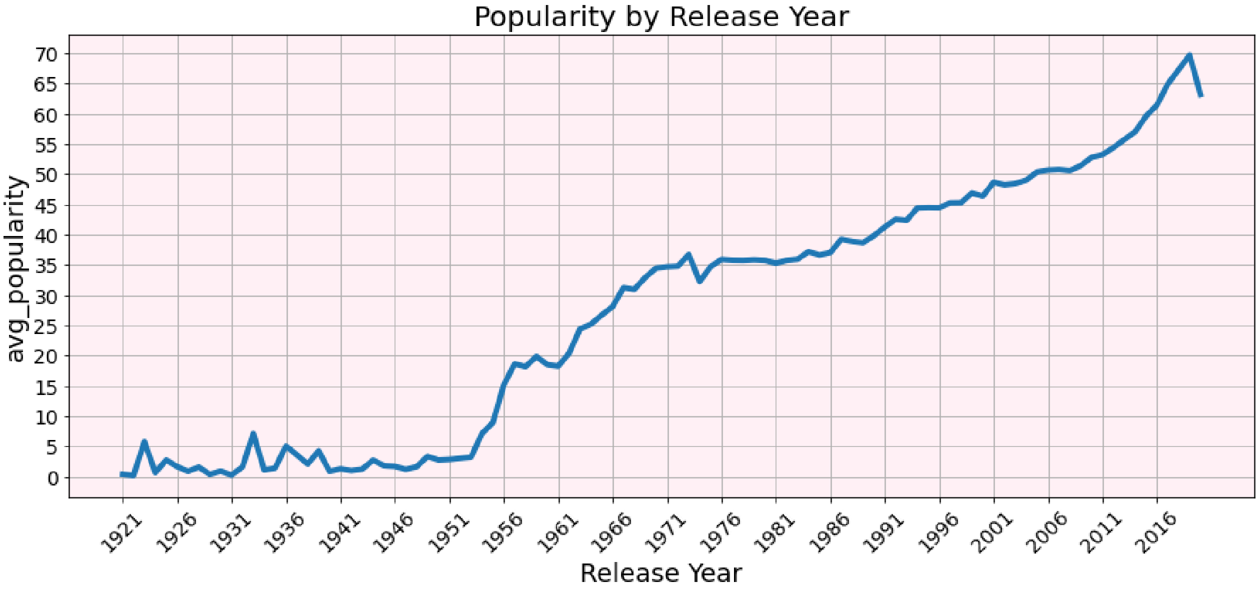
Features	**Correlation**
'year' / 'popularity'	0.8807241644750167
'loudness' / 'energy'	0.7829806054198414
'acousticness' / 'energy'	(-0.7502863679586607)
'year' / 'acousticness'	(-0.6245588720553954)
'popularity' / 'acousticness'	(-0.5933536569832886)
'loudness' / 'acousticness'	(-0.5670788590244086)
'danceability' / 'valence'	0.5602416447727726
'year' / 'energy'	0.5324195814270654

We will not examine every correlation, but will analyze those above 0.6.

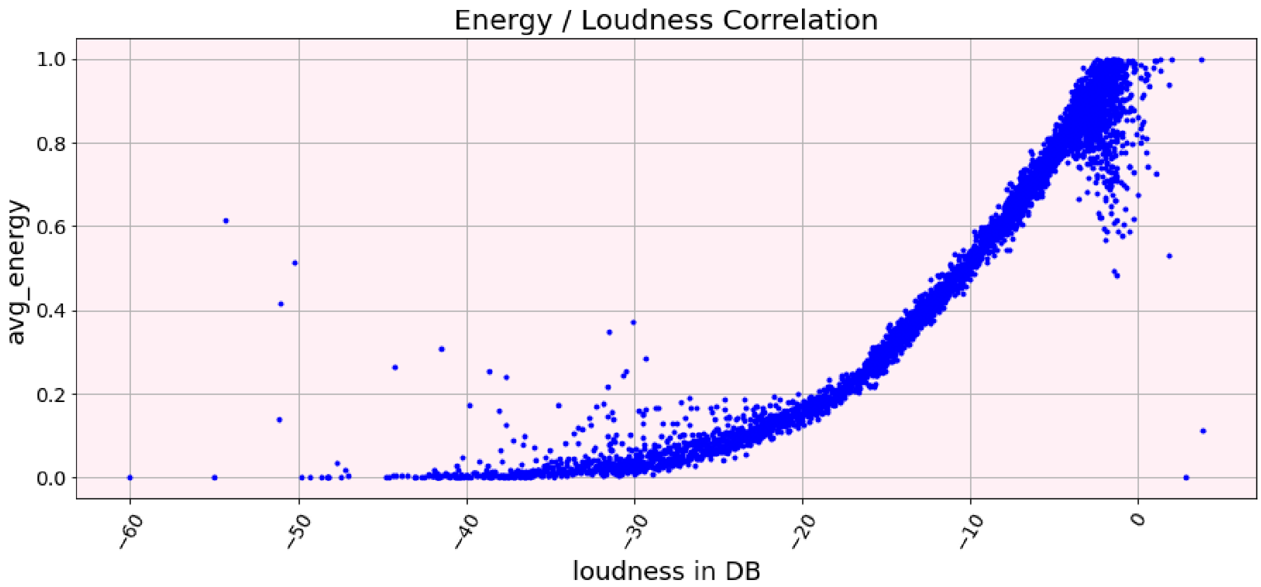
From this list we can learn that the strongest correlation (0.88) is between 'year' and 'popularity' which makes sense: most people would listen either to recent songs and / or songs from their youth.

Less people would listen to songs from the 1920's or 1930's.

In fact, let's see popularity as a function of release year:



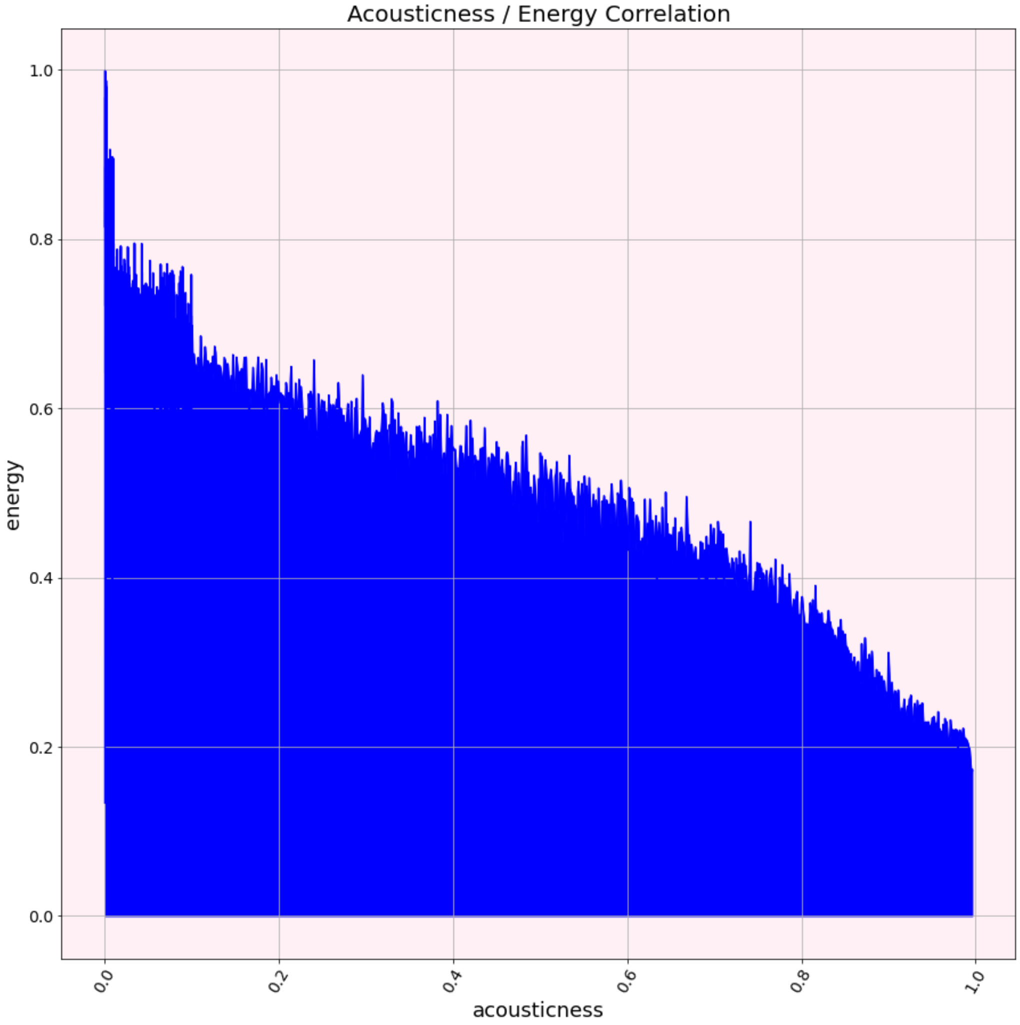
The 2nd strongest correlation is between ‘loudness’ and ‘energy’.
This is actually a known relationship which lead to what is known as "The loudness war".
Below is the visualization of this correlation:



We can see that the loudness / energy correlation is well established in our dataset.

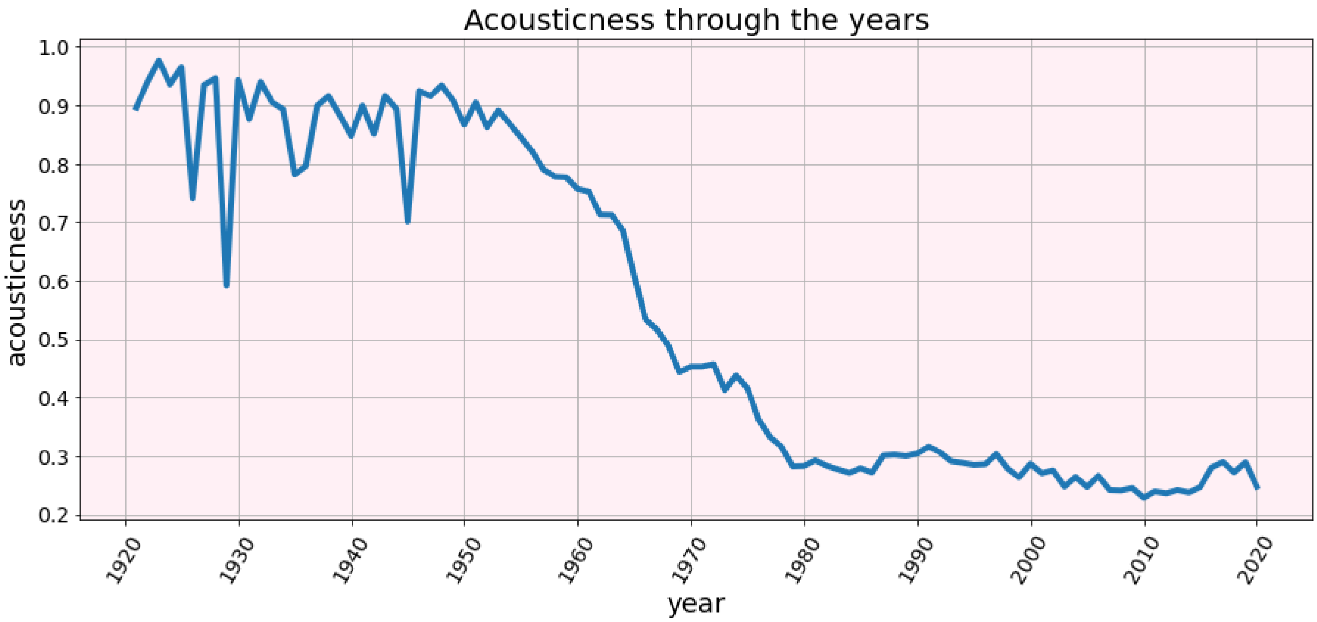
However, this correlation might be a result of the way Spotify algorithm calculates the energy level.
While DB (Loudness) is a measurable value, the energy level is a matter of debate, and as far as we know, the energy rate given to each track might be a calculated result of its loudness.

The 3rd strongest correlation is ‘acousticness’ / ‘energy’.



We can see there’s a steady decline in the energy level as acousticness level is going up.

The 4th strongest correlation, and the last we are going to analyze, is ‘year’ / ‘acousticness’, which is also a negative one, with a value of approx. (-0.62).



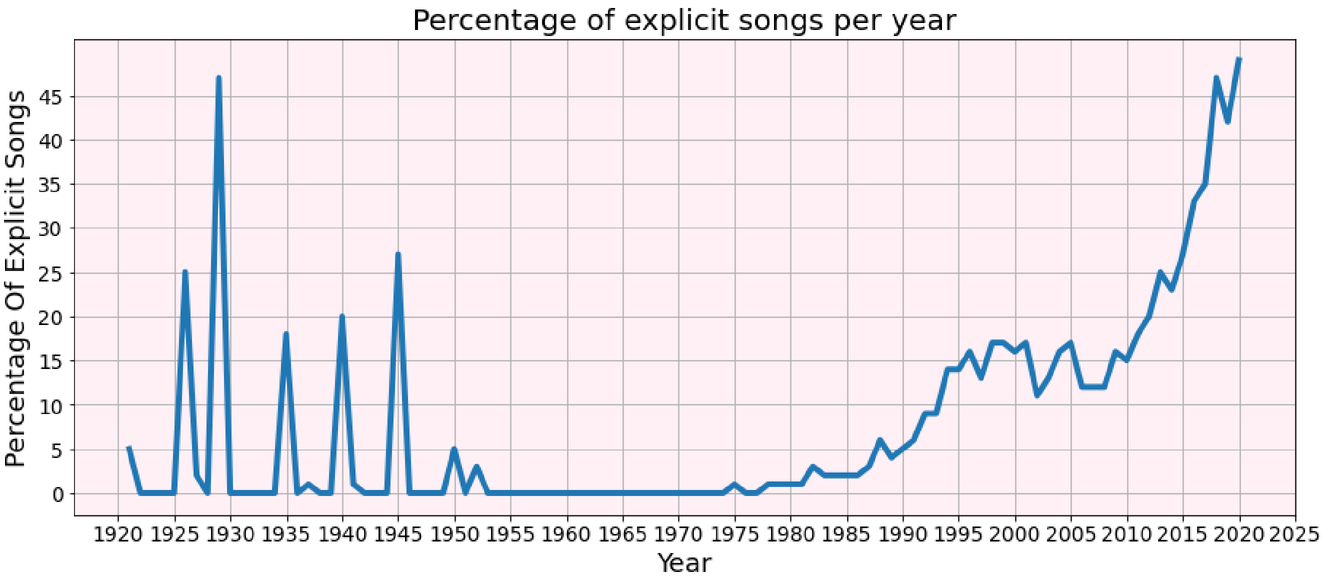
WOW, what a drop!
It is quite visible that acousticness has significantly decreased during the 50's (birth of Rock 'n' Roll and the solid-body electric guitar), 60's and 70's, and stayed roughly the same since the 80's.

5.2. <u>The way different songs' characters changed over the years</u>

5.2.1. Explicit (explicit songs per year)

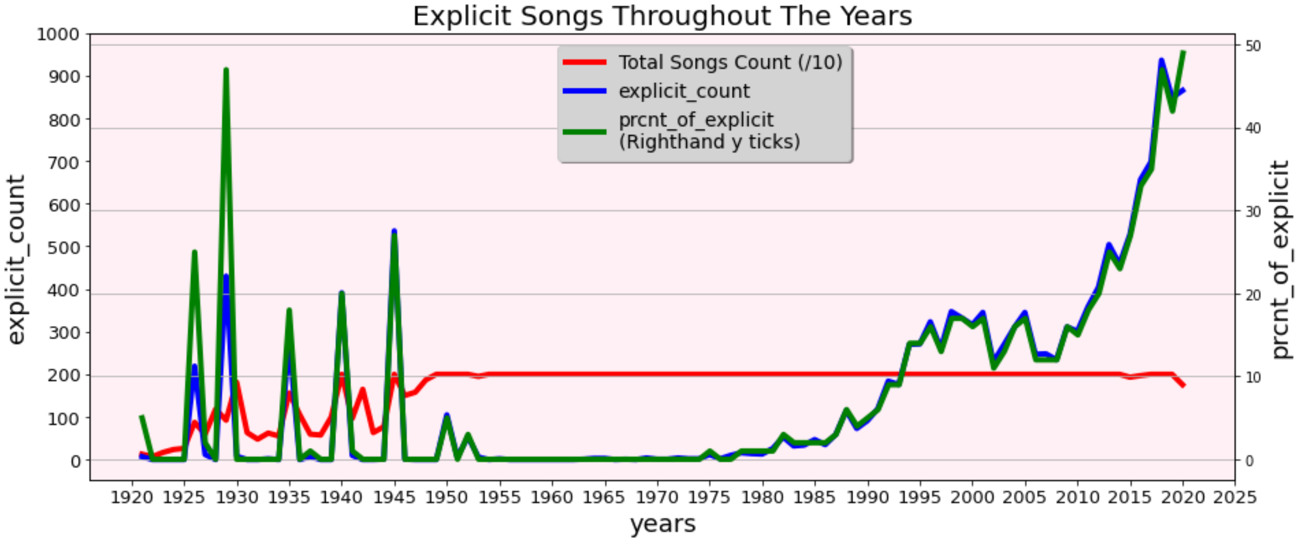
The explicit feature is a very straight forward one - it indicates whether a song contains explicit lyrics or not.

Considering that the explicit values are either 0 or 1, by calculating the yearly average we get the percentage of explicit songs on each year.



It is clear there's something strange about the peaks in the 20's-40's.

Let's compare it with yearly songs count and also with the percentage of explicit songs for each year.



The correlation (where exists) between the total number of songs (in red) and the number of explicit songs (in blue) is expected, but strangely enough, the percentage (in green) is also higher in years with more songs!

This calls for an investigation.

One option is simply to conclude that our database is rogue.

However, we can also offer another explanation:

As we've already seen, some older songs were re-released over the years, either due to re-releasing an entire album or catalogue, or a compilation.

For example, The Beatles’ "past Masters" (songs that were originally released only as singles), guardians of the galaxy soundtrack, essential jazz hits, etc.

So when a record label (and that would usually be one of the "big three": Universal Music Group, Sony Music Entertainment, or Warner Music Group) make their catalogue available on Spotify, they can decide not to upload an old song if they have a better re-released version of it, which is usually a better sounding one.

Better sound quality ➡ better chance users will add it to their personal playlist ➡ more times the song is played ➡ more royalties to the record label.

But do all old songs have re-released versions?

Naturally, re-releasing a song occurs when there is a demand - successful artists, songs that were hits at their time, etc.

We know that mainstream media during the 20's-50's was rather conservative (arguably, still is) and explicit songs could not become hits.

That does not mean that artists didn't have explicit songs back then - just that they remained esoteric, and consequently, there was no reason for the record label to re-release them.

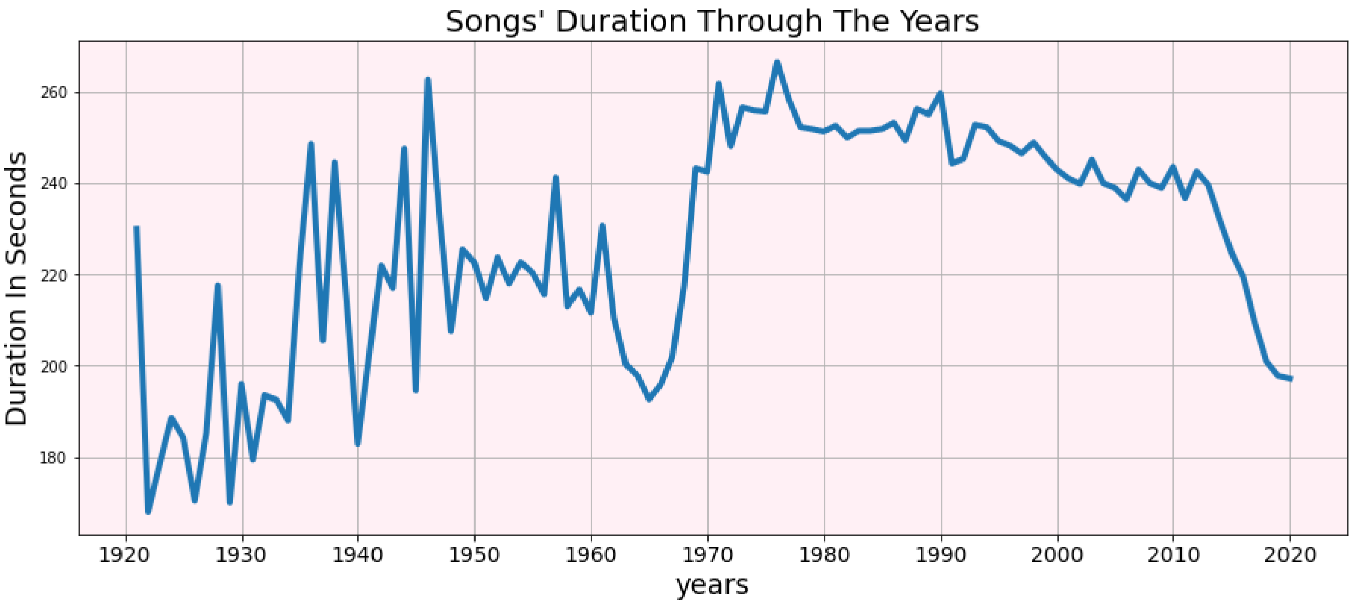
Considering all the above, we can assume that to begin with, the percentage of explicit songs from this time period is higher than the later years (in the general Spotify database) because "clean" songs from that era would only appear under their newer release date, so the more songs from these years were sampled into our database - the higher the chance was to draw an explicit one.

This is how come we get a higher 'explicit' percentage on years were the general songs count is higher.

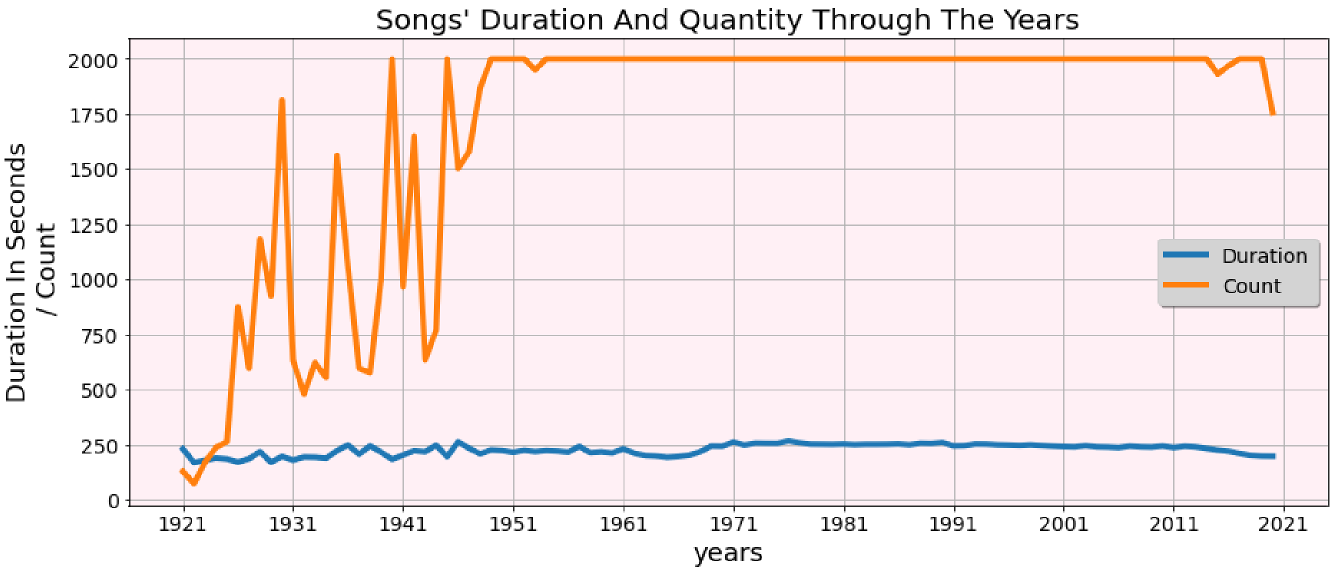
Starting the mid 40's however, we see a steady count of ~2,000 songs per year, and indeed for 3 decades or so, very few songs contains explicit lyrics, while starting the 80's, we notice an increase of songs with explicit lyrics.

5.2.2. Length (duration)

We calculate the average length of songs for each year.

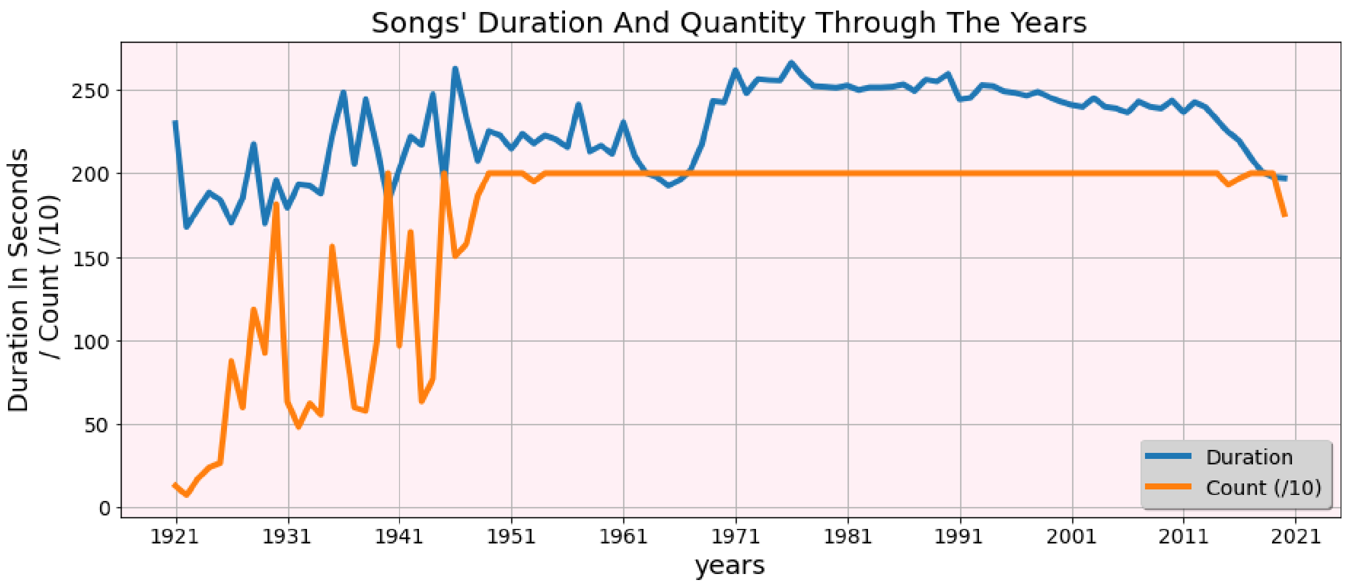


We can see that during the last part of the 60's and the beginning of the 70's there has been a leap in songs' duration from ~200 seconds (the classic 3:20 length) to ~260 seconds, which is about 1 minute (or 30%) of an increase!
However, as our database is not the complete Spotify database, but only a sample of it - it would be wise to see how many songs from each year are sampled.



Obviously, the two scales are too far apart from each other.

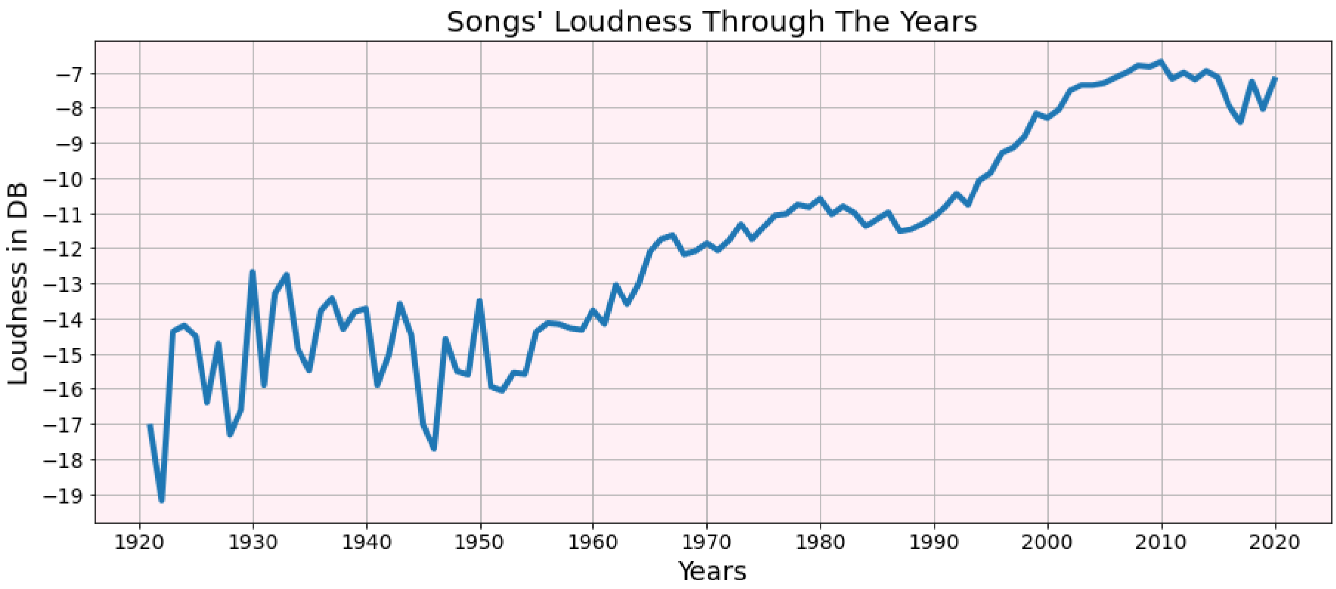
In this case, the exact number is not as important as seeing the trend, so we will "bring down" the count plot by dividing it by 10 instead of adding a 2nd 'y' ticks axis.



This looks ok.
Although song count is bumpy during the 20's-50's, from the beginning of the 50's onwards we have a steady count of roughly 2,000 songs per year, so it's clear the leap we spotted in the 60's-70's is not a matter of uneven songs quantity.

5.2.3. Loudness (is "The Loudness War" a real thing?)

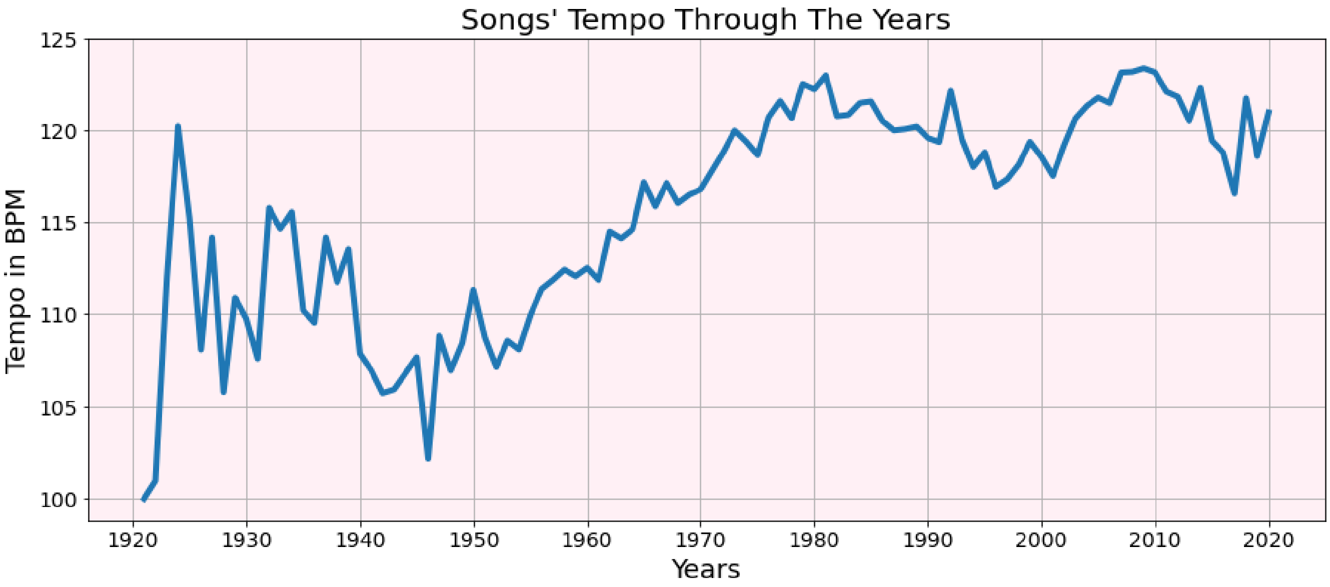
We calculate the average loudness of songs for each year



So yes, loudness does increase over the years and the loudness war is indeed raging on.
One more observation we can make, is that the line is bumpy until we get to the mid-50's, so we can assume the unevenness we've seen in number of songs for these years - affects our result in this case.

5.2.4. Tempo

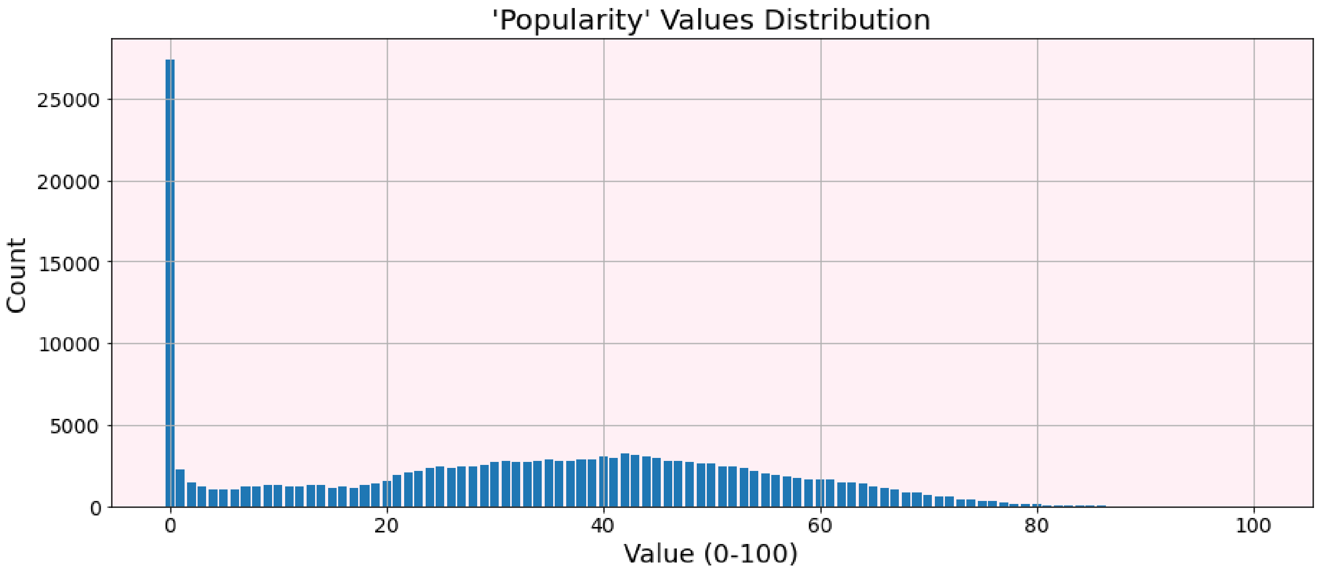
The tempo of a song is measured in BPM (Beats Per Minute), and we calculate the average BPM of songs for each year.



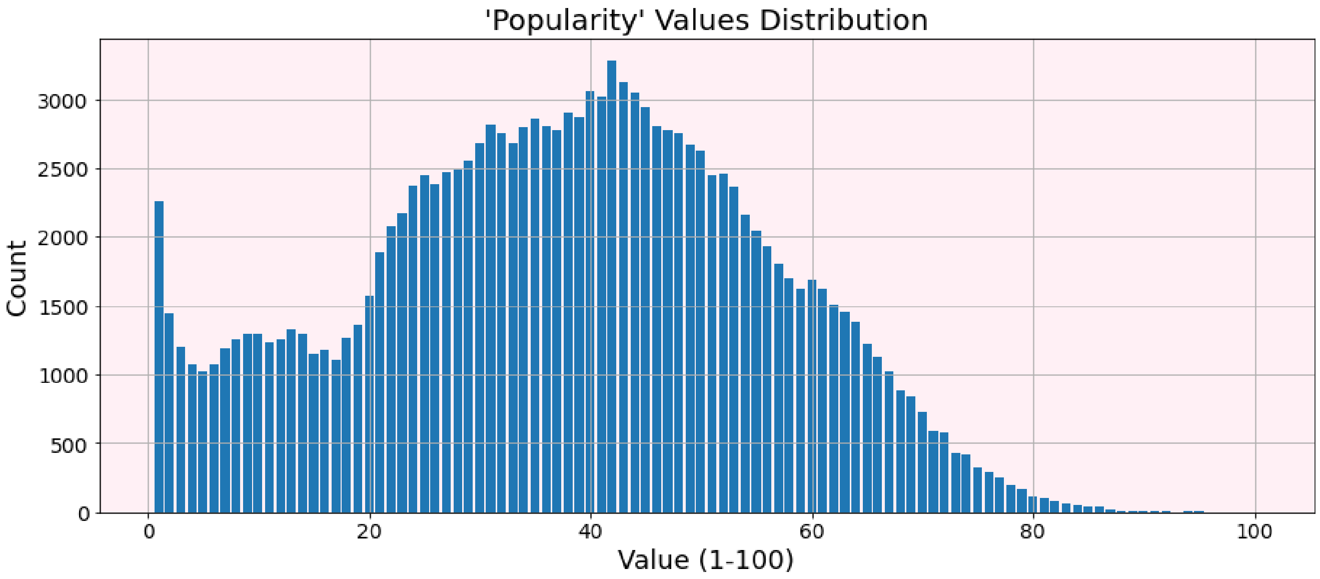
Indeed we can see that songs are getting faster as years go by.
But what about tempo of popular songs?
We've already seen there is no noticeable correlation between popularity and tempo (0.14). Could that mean that if we examine tempo of only popular songs - we will see a relatively narrow range of BPM through the years, or maybe even prescribe an "acceptable" range for tempo of a popular song?

For this, we need to decide what a popular song is.

Firstly, let's see the distribution of popularity (how many songs have popularity of value 0, 1, 2...100)

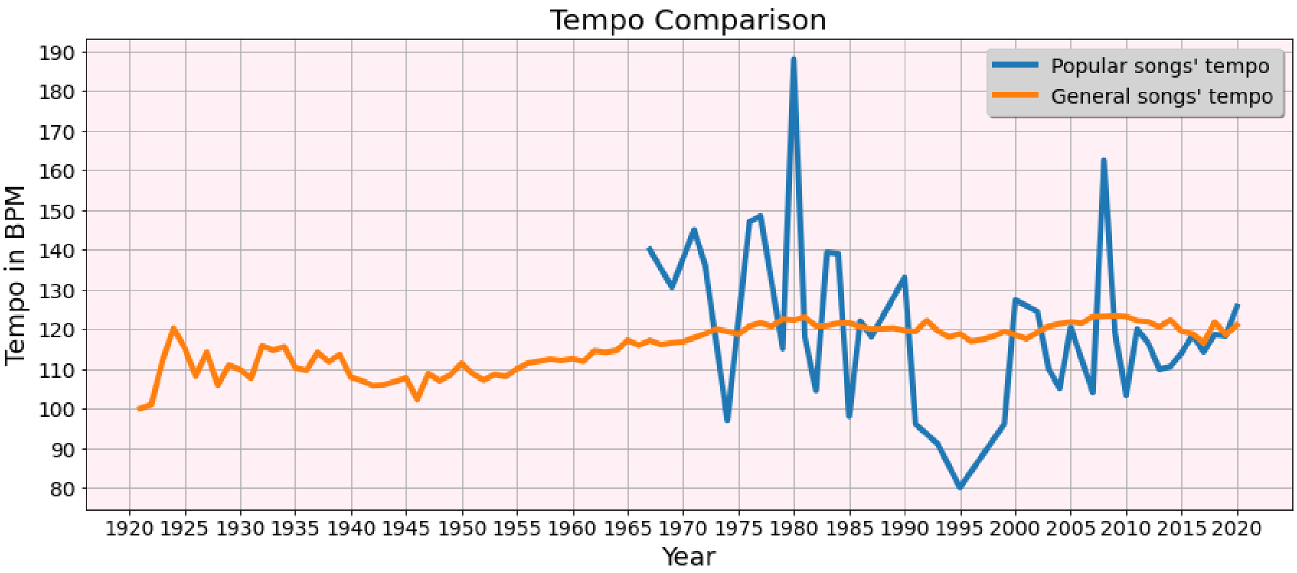


We can see the vast majority of songs have a popularity value = 0.
These are the "long tail" of Spotify. Let's take them out to get a better view of songs that has popularity value of at least 1.



Not quite a bell curve, is it? More like a "It's Lonely at the Top" curve, which is what we could expect - It is no secret that out of hundreds of thousands of songs being published every year, only very few become hits.
For our purposes, we will examine songs with popularity rating of 80 or higher, which are less than 0.5% of all the songs.

That is why we will compare it with the general (all songs) tempo curve.



There are two conclusions we can draw from this plot:

1. Older songs are less popular than new songs - we don't even see songs from before the mid 60's making it into the popular (above 80) curve.

This is actually a phenomenon we've already encountered while examining the 'year' / 'popularity' correlation.

2. While average tempo of songs in general is usually between 100-120 BPM, hit songs vary from 80 - 190 BPM - this actually corresponds with the insignificant correlation (0.14) we've seen between tempo and popularity in the Heat Map.

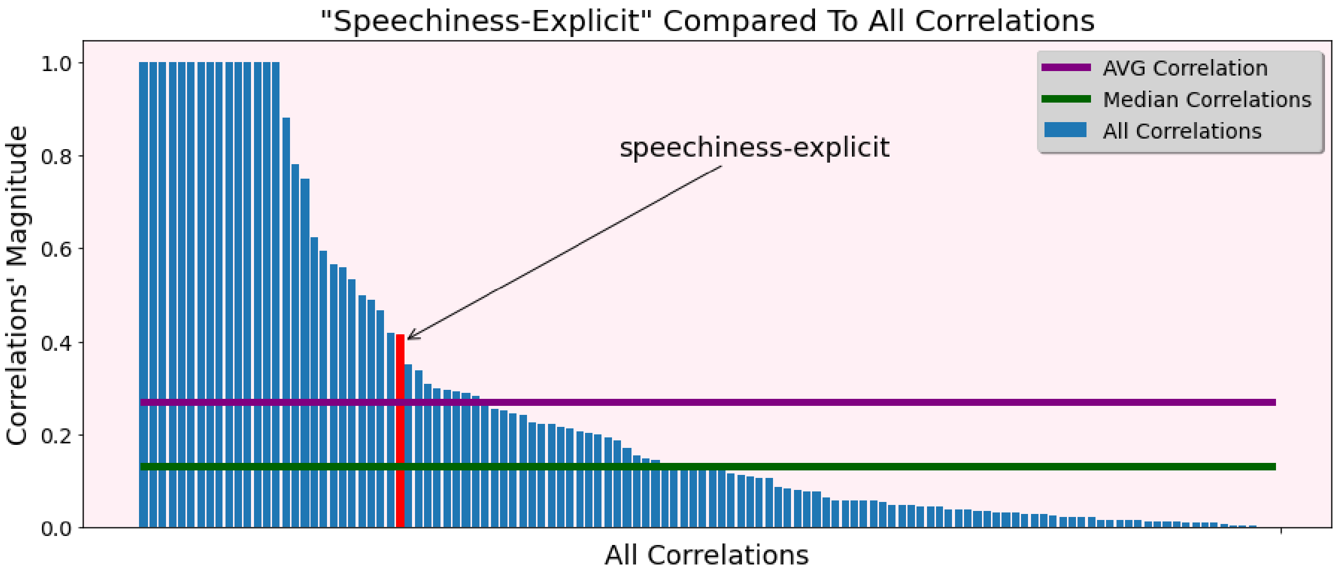
.

5.2.5. Speechiness - explicit correlation

We could already see using the Heat Map, that the correlation (0.41) between the two features is not too high, although it is not a very low one either.
But is it indeed such a medial correlation?

Perhaps when compared with other correlations we will find it actually stands out?

Let's see its position amongst all correlation.



It seems like the ‘speechiness’ - ‘explicit’ correlation does stand out a little above other correlations, but that doesn’t change the fact that it is not a very strong correlation in its own right.

Meaning - we cannot confidently predict that a song that has a high speechiness value will be classified as a song with explicit lyrics.

5.3. Popularity of musical keys

Before diving into the musical key analysis, a quick musical theory lesson:

In western music, every scale can be played in 7 modes. In a nutshell, each mode contains the same notes of the scale, in the same order, but each mode starts with a different note.

I.e. C major Consists of these notes: C, D, E, F, G, A, B (Do, Re, Mi, Fa, Sol, La, Si)

The 2nd mode is called Dorian, and D Dorian Consists of: D, E, F, G, A, B, C

The 3rd mode is called Phrygian, and E Phrygian Consists of: E, F, G, A, B, C, D

And so on and so forth...

The most known mode aside from the major mode, is of course the minor mode, which is the 6th mode. In our example, if we go 6 notes from C, we will get this scale: A, B, C, D, E, F, G, which is A minor.

So basically, C major is equivalent to A minor - they are different modes of the same scale.

The same logic goes to G major ==> E minor, Eb major ==> C minor, etc.

Now that we know that, we need to check - does the 'mode' feature in our database represents a

minor scale of the 'key' feature,
i.e., when key = 0 (C) and mode = 0 (minor), the musical key is C minor,
or, the 'mode' feature means that the song is in the minor mode of that 'key',
i.e., when key = 0 (C) and mode = 0 (minor), the musical key is A minor?
The way to determine this, is to listen to some songs from the database and see the way they are represented in the 'key' and 'mode' columns.

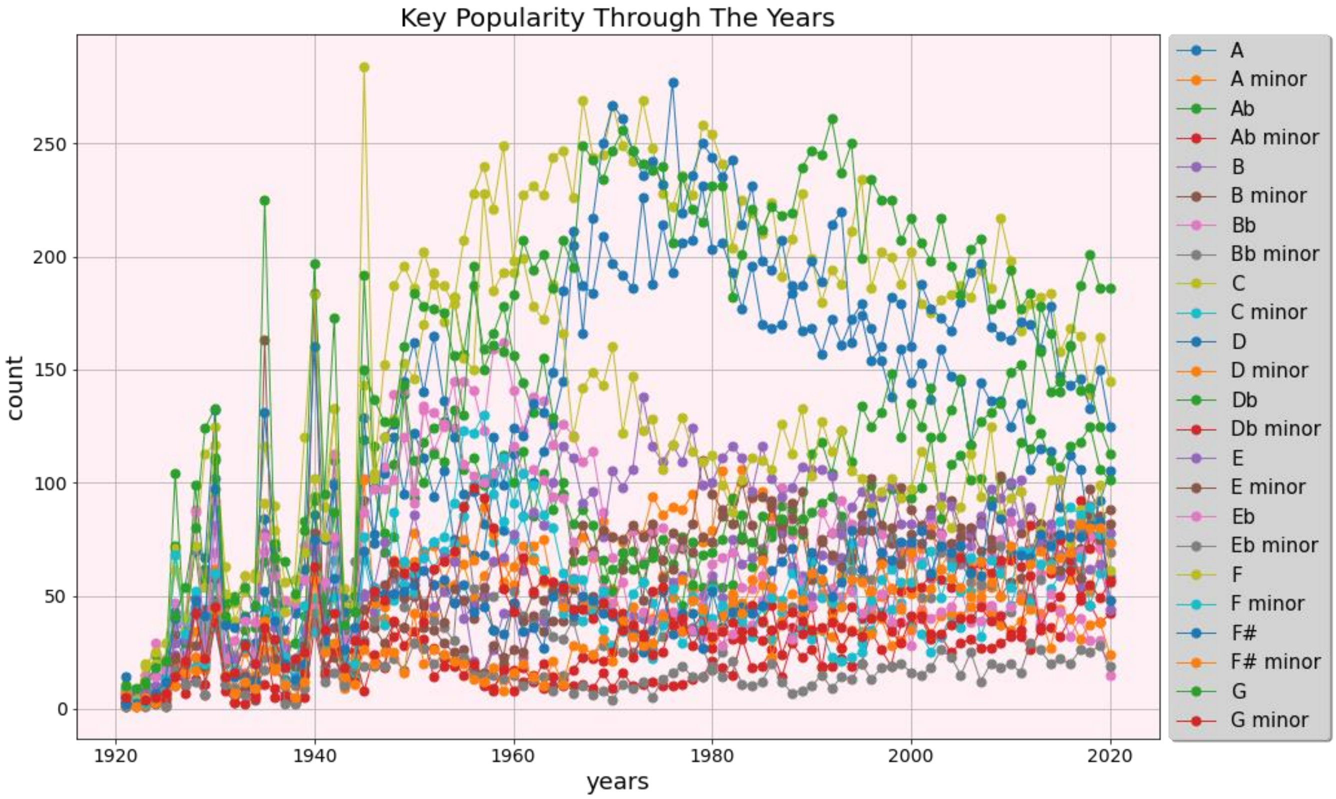
Here are some sample songs:
- The Beatles' “When I'm Sixty Four”. ‘Key’ = 2, ‘mode’ = 1 is played in Db major.
- “Where the Wild Roses Grow” by Nick Cave and Kylie Minogue. ‘Key’ = 7, ‘mode’ = 0 is played in G minor.
- Chopin's "Waltz No.19 in A minor, Op.posth". ‘Key’ = 9, ‘mode’ = 0 is, as the name suggests, played in A minor.
- “Buddha of Suburbia” by David Bowie. ‘Key’ = 2, ‘mode’ = 1 is played in D major.

We can conclude then, that the key represents the pitch of the scale, and the mode determines whether this is a major or a minor scale on that very pitch.

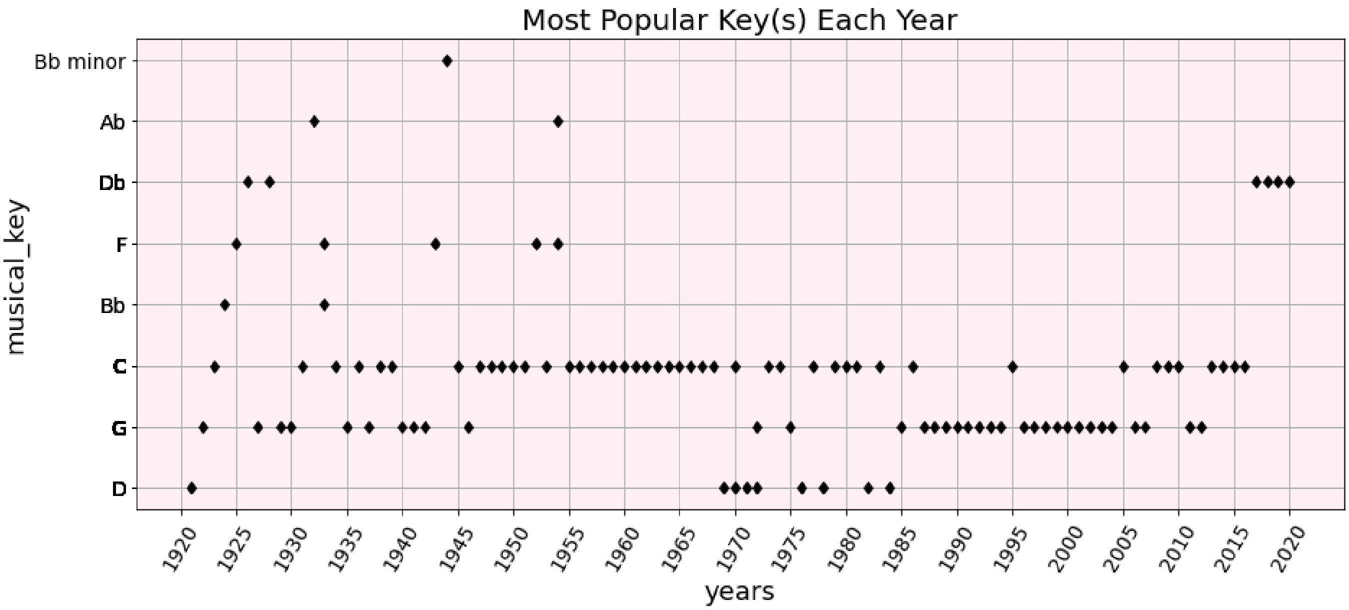
Naturally, in order to truly evaluate the popularity of each key, we cannot use an average function, the appropriate method would be to count how many songs were released in each key every year.

It’s worth mentioning that there is an inherent inaccuracy in the database, as ‘release date’ and respectively ‘year’ do not necessarily represent the original release date of the song, but rather the release date of the specific track, i.e., “Eleanor Rigby” which was originally released in 1966, has a release date of 2014 in our database because it is from the 2014 re-release of the “Yellow Submarine” album.

Naturally, in order to truly evaluate the popularity of each key, we cannot use an average function, what we really need to do is to count how many songs were recorded in each key.



It's clear to say, this is way too messy, and also not quite what we wanted to see - we want the top key(s) in each year.

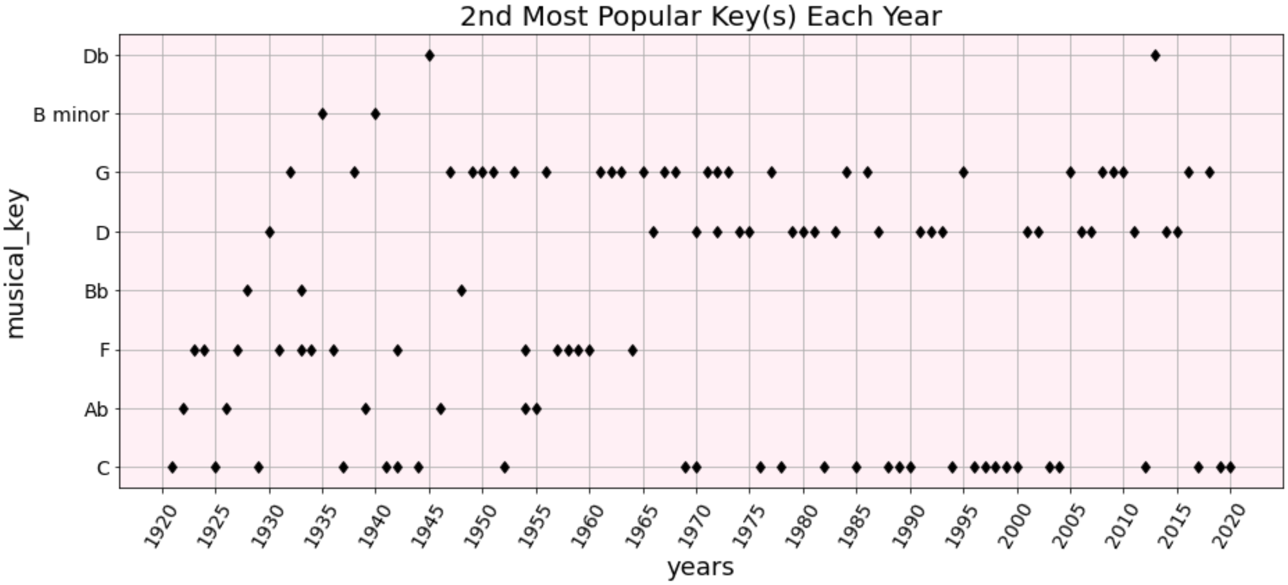


There are two clear winners here!
C (Major) and G (Major) are the dominating keys throughout the years.
This makes sense, as both are basic keys and are easy to play on both the keyboard and the guitar.

Furthermore, only 8 keys (out of 24) ever made it to be the most popular key of the year.

That makes one think - what if the popularity of these keys dims other popular keys?

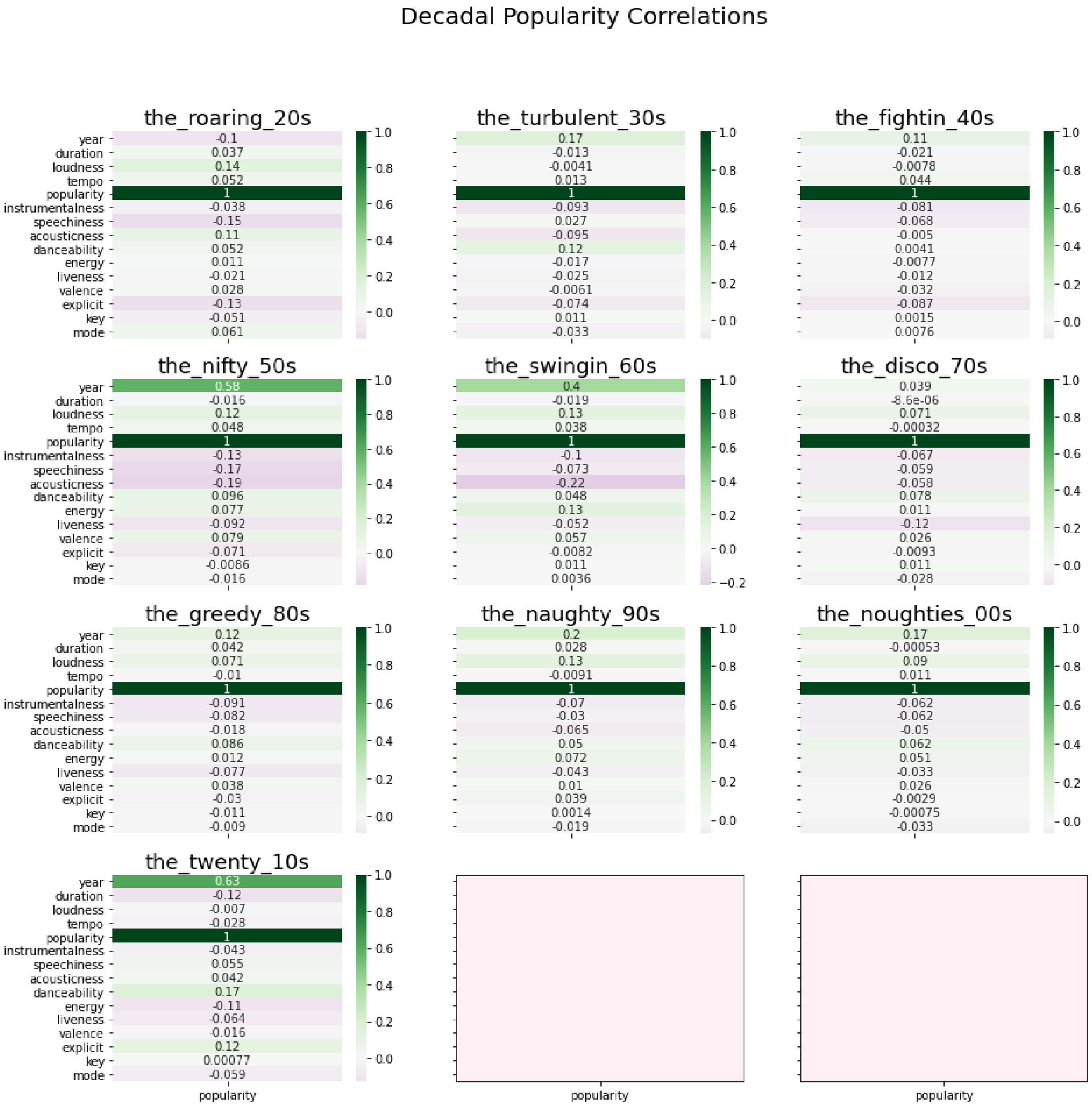
It could be that these 8 keys just “switch places”, and on years where G is the most popular key, the 2nd most popular is C, but perhaps by looking for the 2nd most popular key of each year, we can find a little more diversity of popular keys.



Clearly C Major and G Major are still the dominant keys, but this is a little more diverse.
It is worth mentioning that F major is considered an easy key for keyboard players (only 1 "black key"), but not for guitar players, and we can see that it is quite present until 1964, but not from 1965 and on which is compatible with the establishment of Rock music as a guitar driven genre.

5.4. What makes a song popular in different periods

To answer this question we will break down correlations between ‘popularity’ the other features in every decade.



Even a quick glance can reveal that there is no significant correlation between popularity and any other feature (ignoring a few weak correlation with the year in some cases), regardless of the decade.

Pity, but apparently it takes some other, non-measurable qualities, for a song to become a hit.

6. Summary

Before summarizing, we should remind that this database contains only a fraction of over 60 million tracks available on Spotify (as of Oct 2020. Source: https://newsroom.spotify.com/company-info) and accordingly, all conclusions are only valid to this database, and not the whole of the Spotify tracks database.

The strongest and most noticeable conclusion, is that, unfortunately, we cannot safely predict the behavior of certain features as a result of others. Out of 105 possible pairs for the 15 features we analyzed, only 8 pairs had correlation stronger than 0.5.

The only parameter that does makes a difference is time - it is clear several features did change through the years.

There is however one main feature that we do not have in this database in spite of its importance - the genre(s) associated with each track.

When we talk about music we sometimes tend to forget that music is not of homorganic nature - each genre holds various and different characteristics.

Having a genre indication for each track might have improve the prediction of correlations, as long as the correlated features are of tracks that share the same genre.