1. Introduction

In this notebook, I visualized music tracks data from the Kaggle Spotify datasets, which contain more than 170 thousand tacks from 30,000 artists over the last 100 years. This notebook has essential data exploration analysis and also tries to find out how music has evolved across the years.

As far as I am concerned, the analysis has some limitations:

- The scope of Spotify's tracks collection has more records in recent years than in the early 20s-50s.
- The popularity score only reflects Spotify users' preference (likely to be relatively young user groups.)
- The popularity is scored in the current time. It only shows how people nowadays like old music, not necessarily represent how the classic track's popularity back in the time (a good feature for this will be the record sales for the album).

The datasets has the following features and definitions:

- **duration_ms:**
  The length of the track in milliseconds (ms)
- **artists:**
  The list of artists credited for the production of the track
- **year:**
  The release year of track
- **key:**
  The primary key of the track encoded as integers between 0 and 11
- **mode:**
  The binary value representing whether the track starts with a major (1) chord progression or not (0)
- **release_date:**
  The date of release of the track in yyyy-mm-dd, yyyy-mm, or even yyyy format
- **acousticness:**
  The relative metric of the track being acoustic
- **danceability:**
  The relative measurement of the track being danceable
- **energy:**
  The energy of the track
- **instrumentalness:**
  The relative ratio of the track being instrumental
- **liveness:**
  The relative duration of the track sounding as a live performance
- **loudness:**
  Relative loudness of the track in the typical range [-60, 0] in decibel (dB)
- **speechiness:**
  The relative length of the track containing any kind of human voice

- **valence:**
  The positiveness of the track
- **tempo:**
  The tempo of the track in Beat Per Minute (BPM)
- **name:**
  The title of the track
- **popularity:**
  The popularity of the song lately, default country = US

Import datasets

Kaggle has provided 4 datasets:

- data: the main dataset that based on each track
- data_by_artist: The average of features group by each artist
- data_by_genres: The average of features group by each genres, this is the only dataset that has genres information
- data_by_year: The average of features group by each year

```
Dimensions of 'data'
```

Out[2]: (174389, 19)

Out[3]:

	acousticness	artists	danceability	duration_ms	energy	explicit	
0	0.991000	['Mamie Smith']	0.598	168333	0.224	0	0cS0A1fUEUd1EW3FcF
1	0.643000	["Screamin' Jay Hawkins"]	0.852	150200	0.517	0	0hbkKFIJm7Z05H8ZI9
2	0.993000	['Mamie Smith']	0.647	163827	0.186	0	11m7laMUgmOKqI3oYz
3	0.000173	['Oscar Velazquez']	0.730	422087	0.798	0	19Lc5SfJJ5O1oaxY0f
4	0.295000	['Mixe']	0.704	165224	0.707	1	2hJjbsLCytGsnAHfdɜ

Clean up the format of artists column:

Out[4]:
```
0                        Mamie Smith
1                Screamin Jay Hawkins
2                        Mamie Smith
3                     Oscar Velazquez
4                               Mixe
                          ...
174384    DJ Combo, Sander-7, Tony T
174385                   Alessia Cara
174386                      Roger Fly
174387                   Taylor Swift
174388                      Roger Fly
Name: artists, Length: 174389, dtype: object
```

```
Dimensions of 'data_by_artist'
```

Out[5]: (32539, 15)

Out[6]:

	artists	acousticness	danceability	duration_ms	energy	instrumentalness	livene
0	"Cats" 1981 Original London Cast	0.598500	0.470100	267072.000000	0.376203	0.010261	0.2830
1	"Cats" 1983 Broadway Cast	0.862538	0.441731	287280.000000	0.406808	0.081158	0.3152
2	"Fiddler On The Roof" Motion Picture Chorus	0.856571	0.348286	328920.000000	0.286571	0.024593	0.3257
3	"Fiddler On The Roof" Motion Picture Orchestra	0.884926	0.425074	262890.962963	0.245770	0.073587	0.2754
4	"Joseph And The Amazing Technicolor Dreamcoat"...	0.510714	0.467143	270436.142857	0.488286	0.009400	0.1950

```
Dimensions of 'data_by_genres'
```

Out[7]: (3232, 14)

Out[8]:

	genres	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	
0	21st century classical	0.754600	0.284100	3.525932e+05	0.159580	0.484374	0.168580	-2
1	432hz	0.485515	0.312000	1.047430e+06	0.391678	0.477250	0.265940	-1
2	8-bit	0.028900	0.673000	1.334540e+05	0.950000	0.630000	0.069000	-
3	[]	0.535793	0.546937	2.495312e+05	0.485430	0.278442	0.220970	-1
4	a cappella	0.694276	0.516172	2.018391e+05	0.330533	0.036080	0.222983	-1

```
Dimensions of 'data_by_year'
```

Out[9]: (102, 14)

Out[10]:

	year	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	lou
0	1920	0.631242	0.515750	238092.997135	0.418700	0.354219	0.216049	-12.
1	1921	0.862105	0.432171	257891.762821	0.241136	0.337158	0.205219	-16.
2	1922	0.828934	0.575620	140135.140496	0.226173	0.254776	0.256662	-20.
3	1923	0.957247	0.577341	177942.362162	0.262406	0.371733	0.227462	-14.
4	1924	0.940200	0.549894	191046.707627	0.344347	0.581701	0.235219	-14.

Dimensions of 'data_w_genres'

Out[43]: (32539, 16)

Out[11]:

	artists	acousticness	danceability	duration_ms	energy	instrumentalness	livene
0	"Cats" 1981 Original London Cast	0.598500	0.470100	267072.000000	0.376203	0.010261	0.2830
1	"Cats" 1983 Broadway Cast	0.862538	0.441731	287280.000000	0.406808	0.081158	0.3152
2	"Fiddler On The Roof" Motion Picture Chorus	0.856571	0.348286	328920.000000	0.286571	0.024593	0.3257
3	"Fiddler On The Roof" Motion Picture Orchestra	0.884926	0.425074	262890.962963	0.245770	0.073587	0.2754
4	"Joseph And The Amazing Technicolor Dreamcoat"...	0.510714	0.467143	270436.142857	0.488286	0.009400	0.1950

2. Exploratory Data Analysis (EDA)

Data Distribution

```
feature types:
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 174389 entries, 0 to 174388
Data columns (total 19 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   acousticness      174389 non-null  float64
 1   artists           174389 non-null  object 
 2   danceability      174389 non-null  float64
 3   duration_ms       174389 non-null  int64  
 4   energy            174389 non-null  float64
 5   explicit          174389 non-null  int64  
 6   id                174389 non-null  object 
 7   instrumentalness  174389 non-null  float64
 8   key               174389 non-null  int64  
 9   liveness          174389 non-null  float64
 10  loudness          174389 non-null  float64
 11  mode              174389 non-null  int64  
 12  name              174389 non-null  object 
 13  popularity        174389 non-null  int64  
 14  release_date      174389 non-null  object 
 15  speechiness       174389 non-null  float64
 16  tempo             174389 non-null  float64
 17  valence           174389 non-null  float64
 18  year              174389 non-null  int64  
dtypes: float64(9), int64(6), object(4)
memory usage: 25.3+ MB
```

numeric value description:

Out[13]:

	acousticness	danceability	energy	liveness	loudness	m
count	174389.000000	174389.000000	174389.000000	174389.000000	174389.000000	174389.00(
mean	0.499228	0.536758	0.482721	0.211123	-11.750865	0.70
std	0.379936	0.176025	0.272685	0.180493	5.691591	0.45
min	0.000000	0.000000	0.000000	0.000000	-60.000000	0.00(
25%	0.087700	0.414000	0.249000	0.099200	-14.908000	0.00(
50%	0.517000	0.548000	0.465000	0.138000	-10.836000	1.00(
75%	0.895000	0.669000	0.711000	0.270000	-7.499000	1.00(
max	0.996000	0.988000	1.000000	1.000000	3.855000	1.00(

Most of them range from 0-1 but some came from their own range. Loudness can go from -60 to 3.85, Tempos are from 0 to 243

Let's hear some tracks

What are those Maximum and Minimum sounds like? It is a music dataset, let's hear them!

The loudest tracks:

Out[14]:

	artists	name	loudness
0	Apocolothoth	Sold	3.855
1	The Stooges	Your Pretty Face Is Going to Hell - Alternate ...	3.744
2	Wolfchilde	Weight of Years	3.367
3	The Stooges	Death Trip - Iggy Pop Mix	1.963
4	DYING SPASM	drag	1.830

One of the loudest track: 'Your Pretty Face Is Going to Hell'(turn your volume down)

Out[15]:

0:00 / 4:54

The most energetic tracks:

Out[16]:

	artists	name	energy
0	Rain Recordings	Forest Rain	1.0
1	Creatress	Steady Forest Rain	1.0
2	Epic Soundscapes	Heavy Rain	1.0
3	Rain Sounds ACE	Moderate Rain	1.0
4	Darkthrone	Transilvanian Hunger - Studio	1.0

Surprisingly, the top 4 energetic tracks with the highest popularity scores are all rain sound recordings. It might be due to how Spotify's algorithm calculates the energy score.

Out[17]:

0:00 / 2:50

The most danceable tracks:

Out[18]:

	artists	name	danceability
0	Tone-Loc	Funky Cold Medina	0.988
1	Spooner Street, Rio Dela Duna, Leonardo La Mark	Cool - Leonardo La Mark Remix	0.987
2	Pitbull, Trina, Young Bo	Go Girl	0.986
3	Tone-Loc	Funky Cold Medina - Re-Recorded	0.985
4	Nilla Pizzi	O mama mama - Remix 2014	0.985

One of the top five tracks:

Out[19]:

0:00 / 4:17

The happiest (highest valence) tracks:

Out[20]:

	artists	name	valence
0	Montez de Durango	Pasito Duranguense	1.000
1	Spongebob Squarepants	Electric Zoo	1.000
2	Raffi	Les Petites Marionettes	1.000
3	Raymond Scott	Chatter	1.000
4	8-Bit Arcade	2000 Light Years Away (8-Bit Computer Game Ver...	0.997

Out[21]:

0:00 / 3:20

The clamest (lowest valence) tracks:

Out[22]:

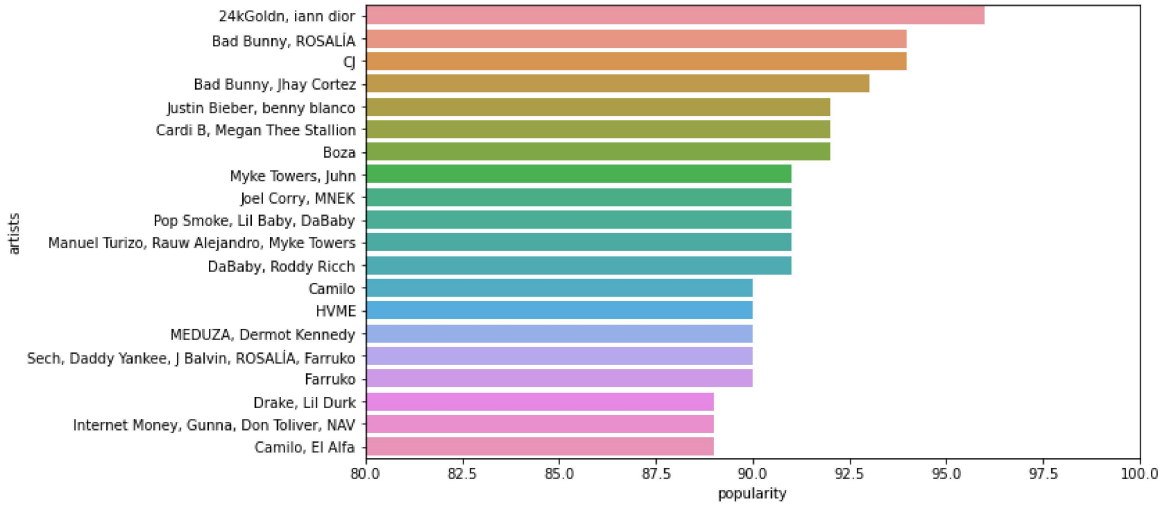
	artists	name	valence
0	Erik Eriksson, White Noise Baby Sleep, White N...	Clean White Noise - Loopable with no fade	0.0
1	Granular	White Noise - 500 hz	0.0
2	Granular	White Noise - 145 hz	0.0
3	High Altitude Samples	Soft Brown Noise	0.0
4	Water Sound Natural White Noise	Deep Sleep Recovery Noise	0.0

Seems the most non-positive tracks are those white noise recording that help people sleep:

Out[23]:

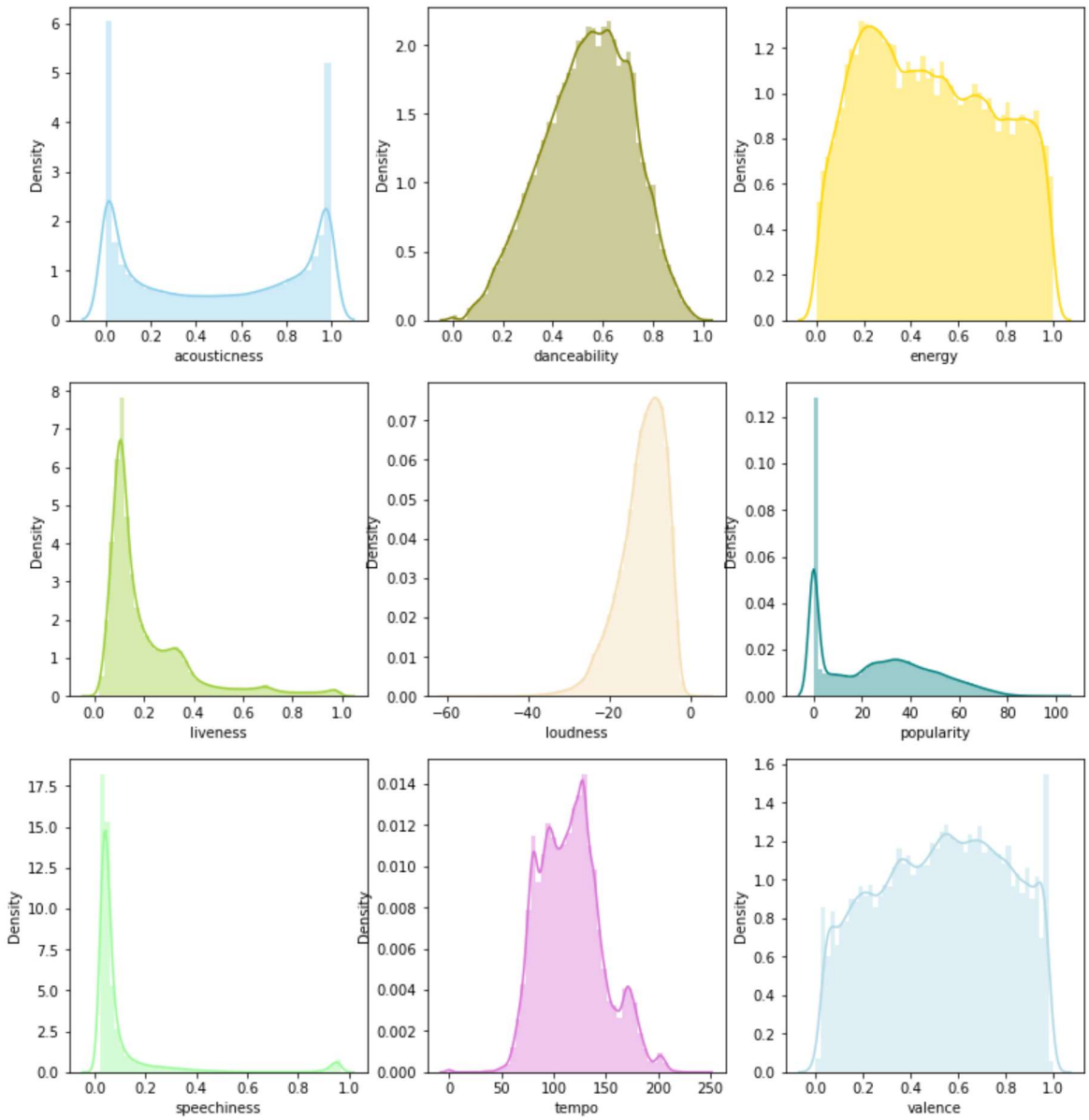
0:00 / 1:32

The most popular artists:



Feature Distrubutions

Out[25]: <AxesSubplot:xlabel='valence', ylabel='Density'>

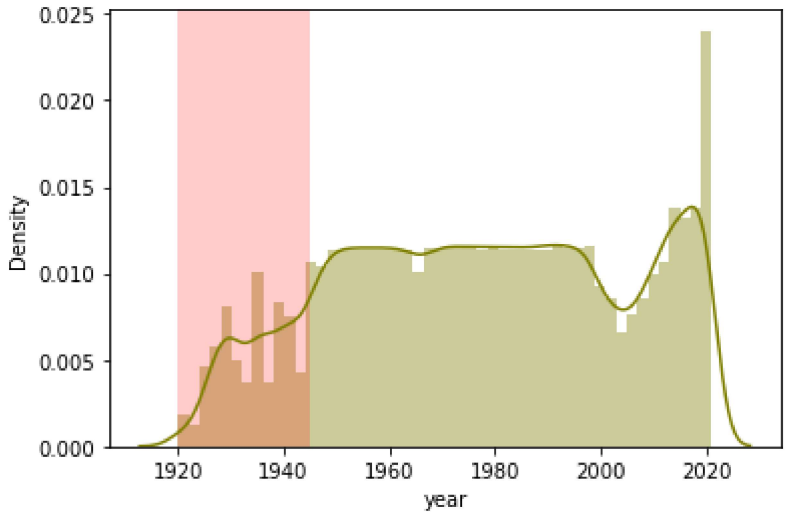


From those charts, we can see:

- Most of the tracks are not live music.
- Most of the tracks are low in 'speechness.' For example, Raps have a certain 'speech' level while Parlour music does not.
- Danceability's distribution seems like a normal shape
- Valence(sentiment) is balanced in general, but with a high number of records near 1.0 - very positive song. Generally, there are more positive songs than negative songs in this collection
- Acuousticness is polarized. Most of the tracks are either pure instrumental or vocal, which make sense.
- Many tracks have '0' popularity. Most of the popularity scores sit from 20 to 60.

Let's examine the distribution by years. The chart below shows that fewer tracks are available before 1945, and the number varies significantly among those years. On the other hand, a large amount of collection shows up from 2019 to the present.

Out[26]: `<matplotlib.patches.Polygon at 0x1c771a79e88>`



Popularity - What are those 0s score?

There are a large number of records with '0' popularity scores in the previous histogram matrix. What are those tracks that are being '0' popular? Do they have some shared characteristics? Let's take a look at those and see if there is any pattern.

some 0 popularity tracks:

Out[27]:

	artists	name	popularity
0	Alessia Cara	A Little More	0
1	DJ Combo, Sander-7, Tony T	The One	0
2	Alessia Cara	A Little More	0
3	Roger Fly	Together	0
4	Roger Fly	Improvisations	0

0s popularity record numeric value summary:

Out[28]:

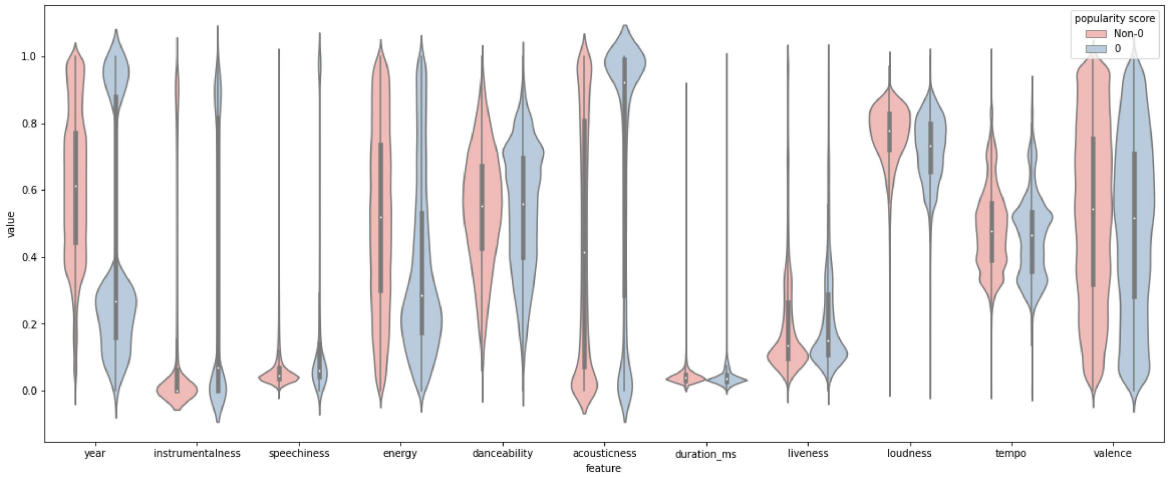
	acousticness	danceability	duration_ms	energy	explicit	instrumentalnes
count	40905.000000	40905.000000	4.090500e+04	40905.000000	40905.000000	40905.00000
mean	0.673725	0.534144	2.325309e+05	0.370634	0.080137	0.34423
std	0.389875	0.184625	2.024124e+05	0.262042	0.271508	0.39159
min	0.000000	0.000000	4.937000e+03	0.000000	0.000000	0.00000
25%	0.284000	0.393000	1.523080e+05	0.173000	0.000000	0.00001
50%	0.918000	0.552000	1.909470e+05	0.285000	0.000000	0.06770
75%	0.987000	0.687000	2.528240e+05	0.531000	0.000000	0.81600
max	0.996000	0.987000	5.338302e+06	1.000000	1.000000	0.99900

Non-0s popularity record numeric value summary:

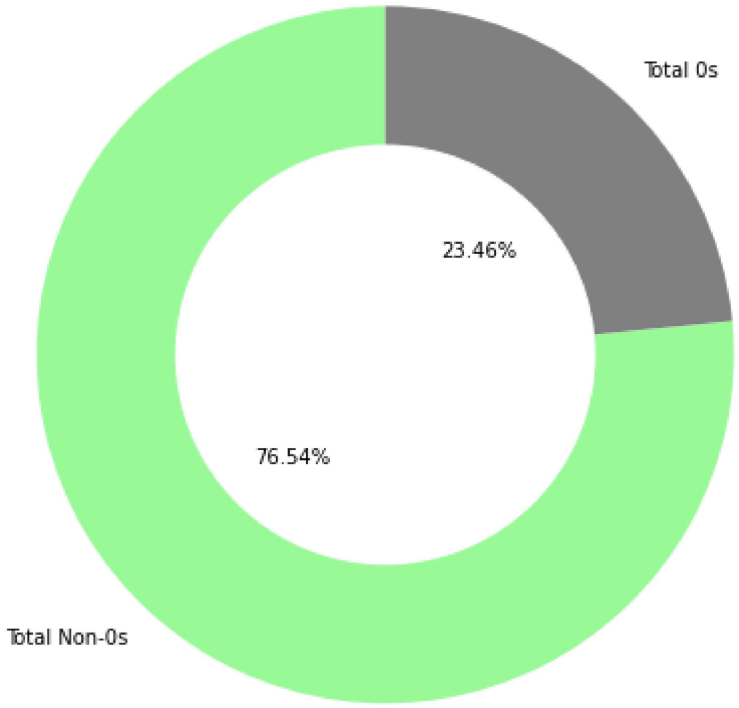
Out[29]:

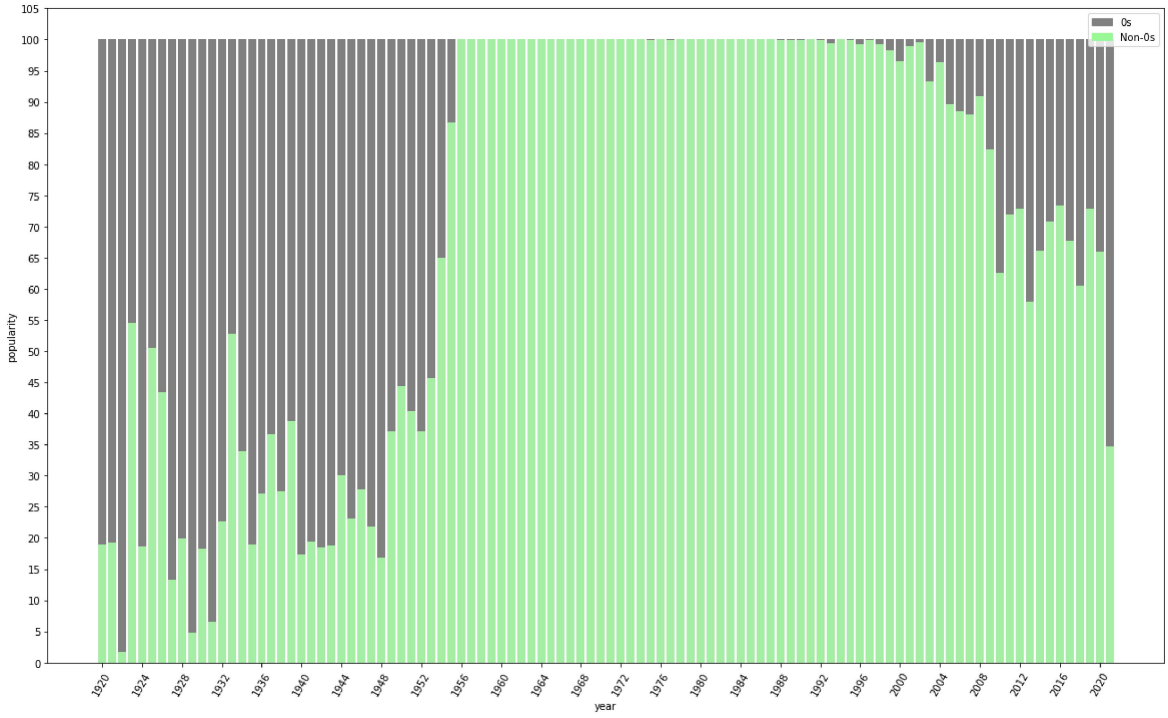
	acousticness	danceability	duration_ms	energy	explicit	instrumenta
count	133484.000000	133484.000000	1.334840e+05	133484.000000	133484.000000	133484.0
mean	0.445756	0.537559	2.328956e+05	0.517069	0.064457	0.1
std	0.360302	0.173297	1.273368e+05	0.266594	0.245566	0.3
min	0.000000	0.000000	1.470800e+04	0.000000	0.000000	0.0
25%	0.070700	0.421000	1.696658e+05	0.299000	0.000000	0.0
50%	0.412000	0.547000	2.124000e+05	0.519000	0.000000	0.0
75%	0.805000	0.663000	2.679730e+05	0.737000	0.000000	0.0
max	0.996000	0.988000	4.892761e+06	1.000000	1.000000	1.0

If we compare the average of the 0s popular with the other data, the 0s one are generally older(year 1960 vs. 1982), less energetic(0.37 vs. 0.51), more instrumental(0.34 vs. 0.15), and more speeches involved(0.19 vs. 0.08.)



The violin plots present similar results: Less energetic and more acoustic tracks are more likely to have a 0 popularity score. For year, it is obvious that the 0 popular tracks appear to be either older or newer than the average.

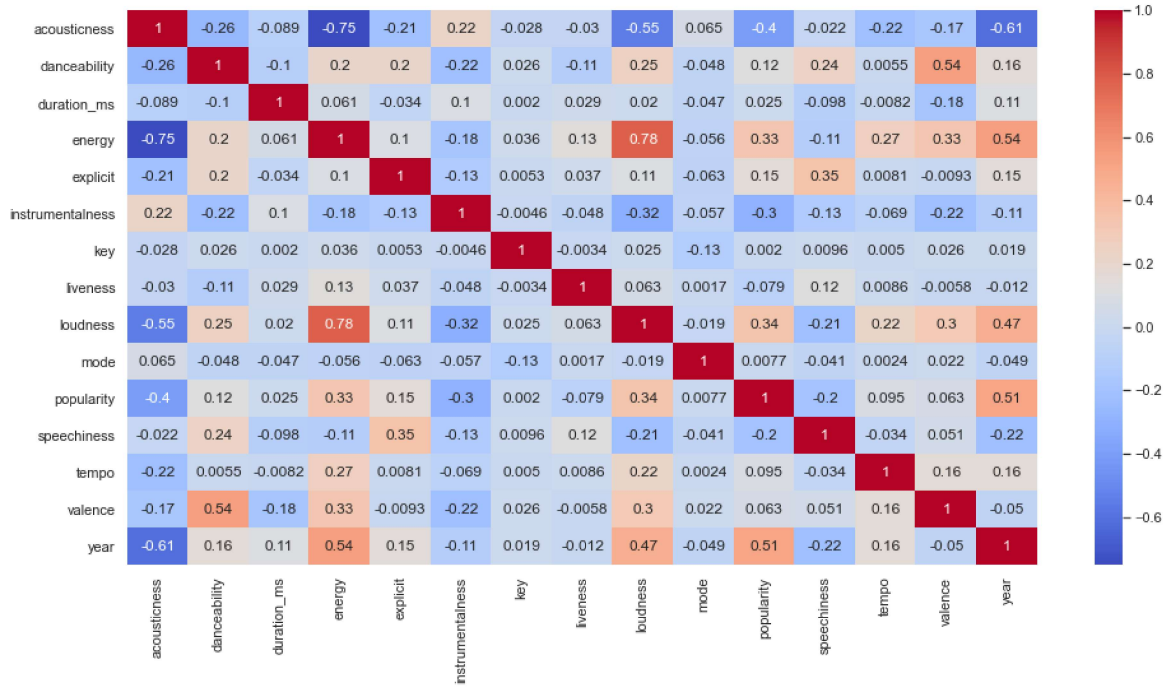

Total 0s
23.46%
76.54%
Total Non-0s


0s
Non-0s
popularity
year

In total, we have about a quarter of tracks(23.46%) that are 0 popular. The bar charts above show that the tracks before 1955 take up the majority of those 0s score. One reason might be that they are 'too old' for people to listen to nowadays. It seems that there is a 'Golden Age' period from 1960 to 2000, where more than 95% of tracks back then have scored some popularity. Nonetheless, the latest tracks that are in 2021 might be too new to have a popularity score assigned.

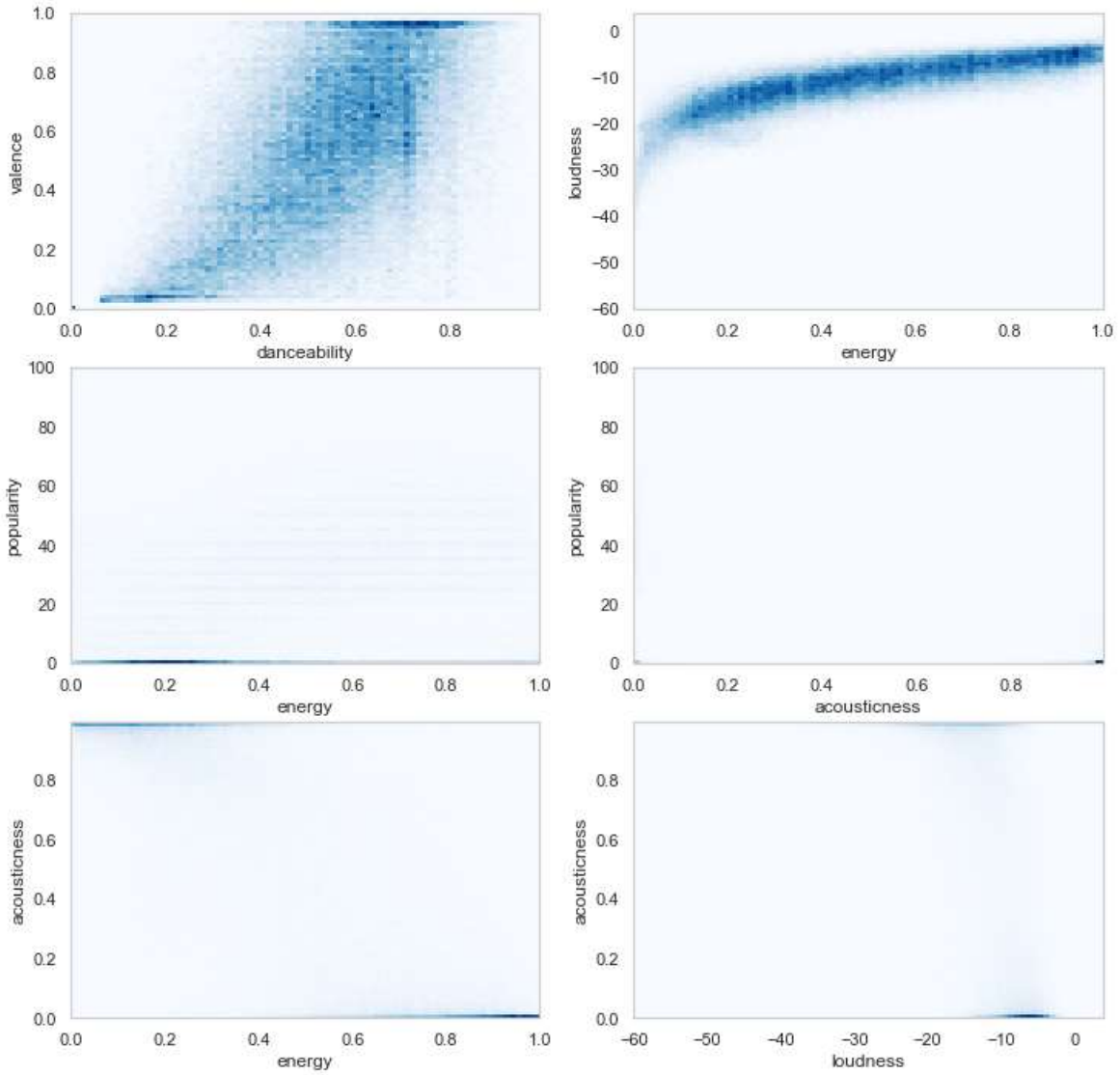
Feature Correlations

Out[32]: <AxesSubplot:>



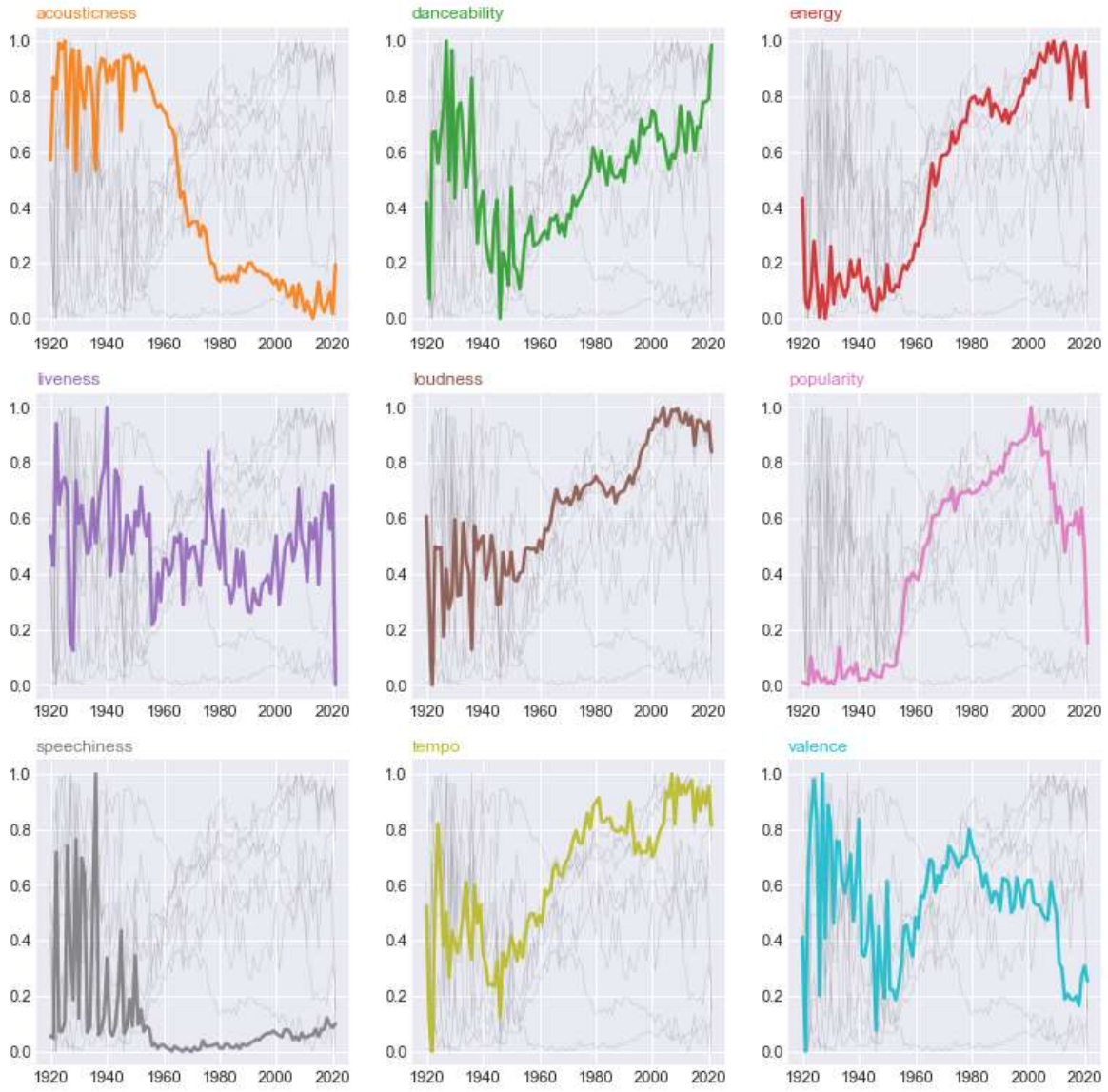
If we look at the popularity column, we notice that it is highly correlated with some of the features. Spotify users seem to prefer the more energetic, louder, less acoustic tracks.
There is also a high correlation of 0.51 with feature 'year.' We notice that the music track has become less acoustic, more energic, and louder over the years.
Some other correlation pairs are inline with what we expected, such as high energy tracks are louder, non-acoustic tracks(R&R, electronic, etc.) are more energetic and louder.

Based on the heatmap, I plot some high correlation feature pairs to examine their relationship in 2d histogram. The darker of bins in the charts, the more points lay at the same position. As we can see, Danceability/Valence, Energy/Lounderness shows a strong positive linear relationship, while others are not very obvious.



Music Trend Analysis

Numeric Attributes Trend Over The Years



Lines are volatile before 1950 due to the fewer available records. Some general trends spotted are music becomes louder, faster pace, more energetic ,and easier to dance to. There are also more vocal tracks instrumental over the years. It seems that tracks from 1970 to 2000 are more positive, which might lead to the high popularity scores from that period.

The analysis below summarizes tracks' characteristics into different periods (decades). I wonder if every generation of music has its traits that we could recognize and the artists representing that generation.

Assign decades to the dataset:

Out[36]:

	artists	name	year	decades
0	Mamie Smith	Keep A Song In Your Soul	1920	20s
1	Screamin Jay Hawkins	I Put A Spell On You	1920	20s
2	Mamie Smith	Golfing Papa	1920	20s
3	Oscar Velazquez	True House Music - Xavier Santos & Carlos Gomi...	1920	20s
4	Mixe	Xuniverxe	1920	20s
...	...	...	...	...
174384	DJ Combo, Sander-7, Tony T	The One	2020	2010 & newer
174385	Alessia Cara	A Little More	2021	2010 & newer
174386	Roger Fly	Together	2020	2010 & newer
174387	Taylor Swift	champagne problems	2021	2010 & newer
174388	Roger Fly	Improvisations	2020	2010 & newer

174389 rows × 4 columns

The most popular artists of each generation:

Out[37]:

	decades	artists	popularity	rank
0	2010 & newer	24kGoldn, iann dior	96.0	1.0
69	90s	Frank Sinatra, B. Swanson Quartet	84.0	1.0
72	80s	Grover Washington, Jr., Bill Withers	83.0	1.0
115	2000s	Shakira, Wyclef Jean	82.0	1.0
531	60s	The Beach Boys, Mark Linett, Sweet, Larry Walsh	75.0	1.0
742	70s	Earth, Wind & Fire, The Emotions	73.0	1.0
3569	50s	Gayla Peevey	62.0	1.0
4764	40s	Bing Crosby, The Andrews Sisters	58.6	1.0
16436	20s	Benny Goodman, Peggy Lee	37.0	1.0
18700	30s	Richard Himber and his Orchestra, Johnny Mercer	34.0	1.0

The six radar charts below show six music generations' features, from the 60s to 2010, following with the most popular artist of that time. The grey hexagon underneath shows the time period's average, and the colored one is the artist's characteristics

Now let's Combine Key and Mode in the dataset:

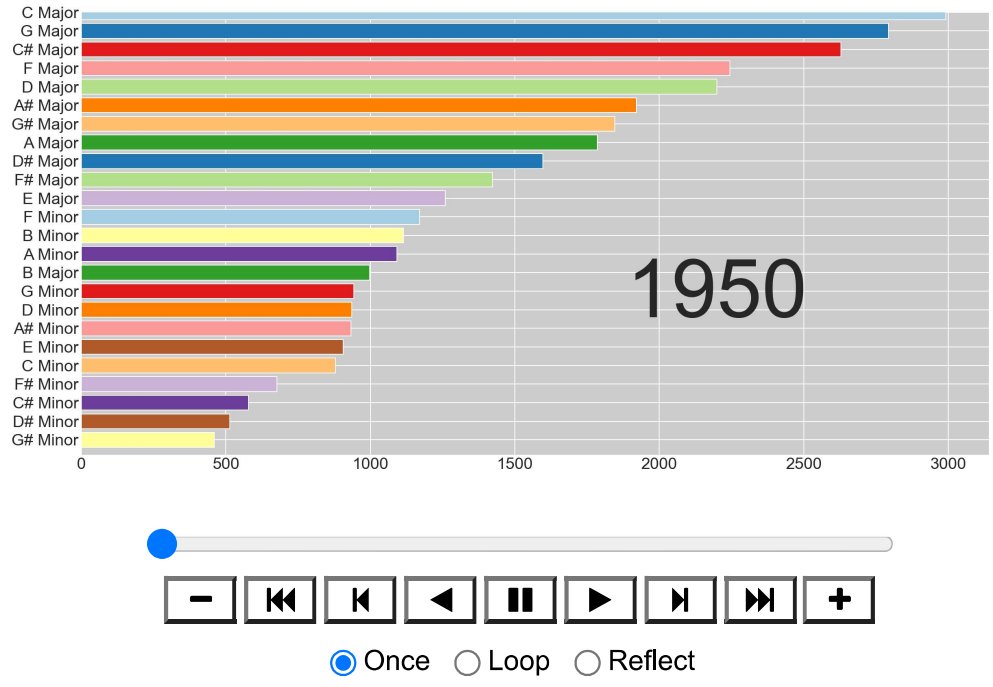
Out[39]:

	name	key	mode	key_mode
0	Keep A Song In Your Soul	5	0	F Minor
1	I Put A Spell On You	5	0	F Minor
2	Golfing Papa	0	1	C Major
3	True House Music - Xavier Santos & Carlos Gomi...	2	1	D Major
4	Xuniverxe	10	0	A# Minor
...	...	...	...	...
174384	The One	6	0	F# Minor
174385	A Little More	4	1	E Major
174386	Together	4	0	E Minor
174387	champagne problems	0	1	C Major
174388	Improvisations	7	1	G Major

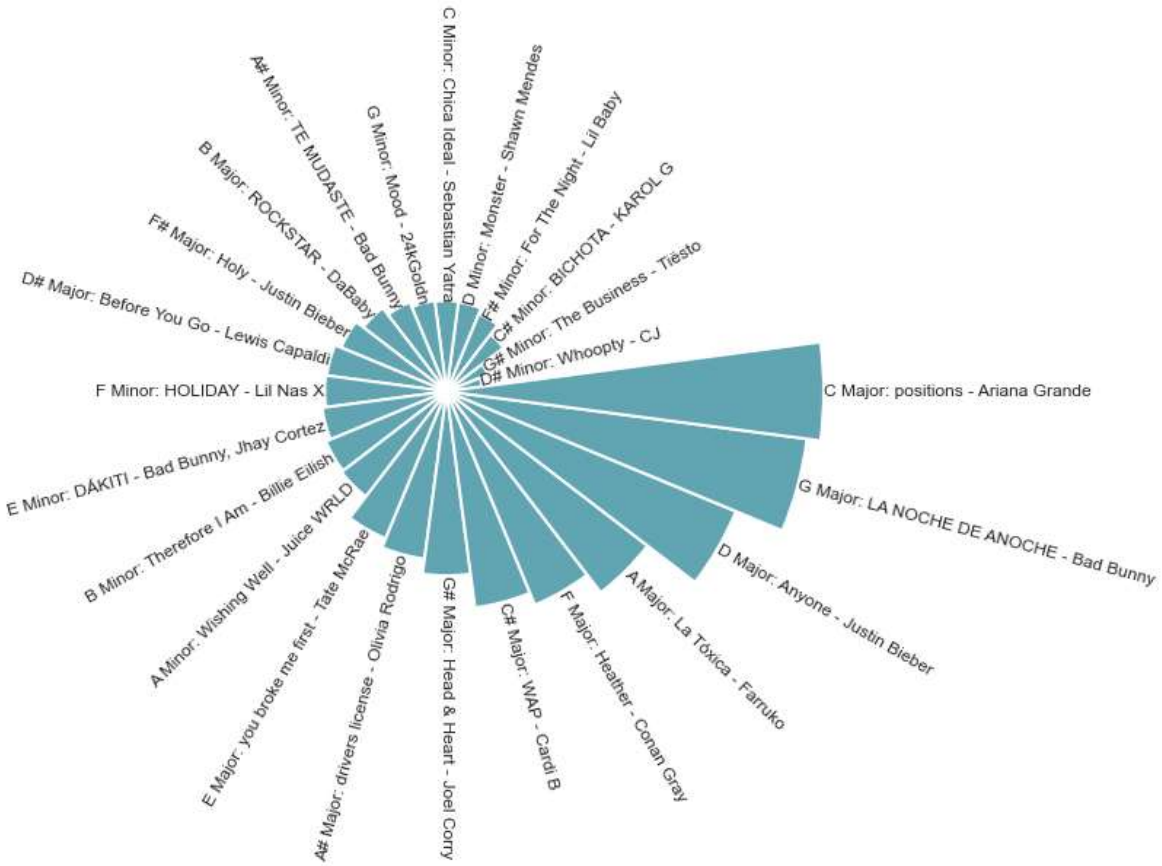
174389 rows × 4 columns

Race chart animation of music keys being used over the years:

Out[40]:



The most popular song of each key:



To be continued...