Music according to Spotify

Breaking down the most popular search requests

by the biggest streaming platform

By Eduardo Beltrán Herrera

IBM Machine Learning Specialization

12/24/2020

Introduction

Being the biggest music streaming platform in the world, Spotify has become the biggest signifier of trends in popular music and in many ways, pop culture. The swedish company now has massive amounts of data from music fans all around the world: from popular artists by country, to listening habits of each individual customer, Spotify surely has a big enough sample to describe the listening habits of the general population.

Thankfully, Spotify makes most of this data publicly available, and so we have tasked ourselves with analyzing their database and giving a general overview of what is popular now, how has music changed over the years, and what will probably be popular in the near future.

The dataset was acquired from *Kaggle.com*, where a search query returning around 4000 songs per year was shared by the community members. It is unclear how the algorithm chooses which 4000 songs to return, but we can assume it has to do with recent popularity (the presence of Christmas music is a big clue).

The dataset contains information about each song (artists involved, release date of the song, tempo, duration, loudness, and popularity) as well as some interesting measures that are obtained via the algorithm (danceability, valence, acousticness, energy and liveness). While these measures are bound to be imperfect, we believe they will at least give us a general feel of how music "feels" within each year. The dataset information spans from 1920 to 2020, and it was last updated on **November 25, 2020**. The update date is relevant, as the popularity score is based on how much a song is listened to and search for at the time of the data extraction.

Our goal is to clean the data and reduce it's size to only include music (as podcasts and other audio-based media can be found in the set), do a quick overview of how music has changed overtime, and try some hypothesis about the future of music, and the popularity of the tracks yet to come.

Data cleansing

The original dataset contains **155,978 observations**, which include around 4000 songs per year starting from 1920 until 2020 8the years before 1944 vary a lot in quantity, some only containing as little as a couple hundred results).

For this analysis, the data before 1945 was dropped to reduce the set size and focus on more modern music. To try and filter out audiobooks or podcasts, all records with a "speechiness" of 0.75 or higher were also filtered. Then, duplicates for artists and song names were eliminated, to try and account for re-releases of the same song by the same artists.

After further observation of the data, lots of outliers were observed in the song duration feature. The outliers were eliminated using the interquartile range multiplied by 1.5:

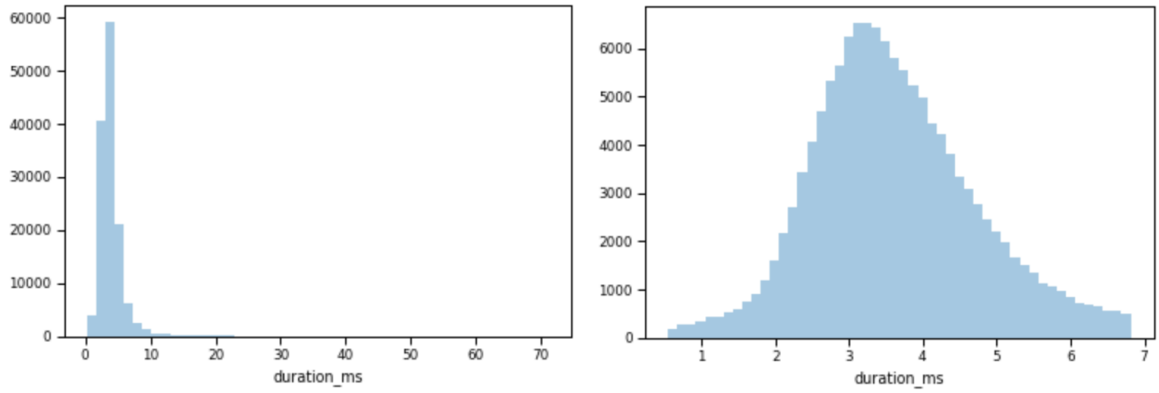


Figure 1. *Before and after of histogram for the song duration*

All these transformations reduced the dataset to **129,471 observations**.

Some features of the dataset were also dropped, allowing us to focus on the more objective data, although some of the algorithmic measures were kept:

1. **'valence'** – how *happy* the music is (score from 0 to 1)
2. **'year'** – release year of the song
3. **'artists'** - all artists involved with the song
4. **'danceability'** – how danceable the song is, according to its energy, bpm, etc. (score from 0 to 1)
5. **'duration_ms'** – duration of the song in miliseconds
6. **'explicit'** – if the song contains swear words or explicit content (0 = non-explicit, 1 = explicit)
7. **'loudness'** – how loud the song is in dB
8. **'name'** – name of the song
9. **'popularity'** – how popular the song is relatively to all other songs, at the time of the extraction (score from 0 to 100)
10. **'tempo'** – beats per minute of the song

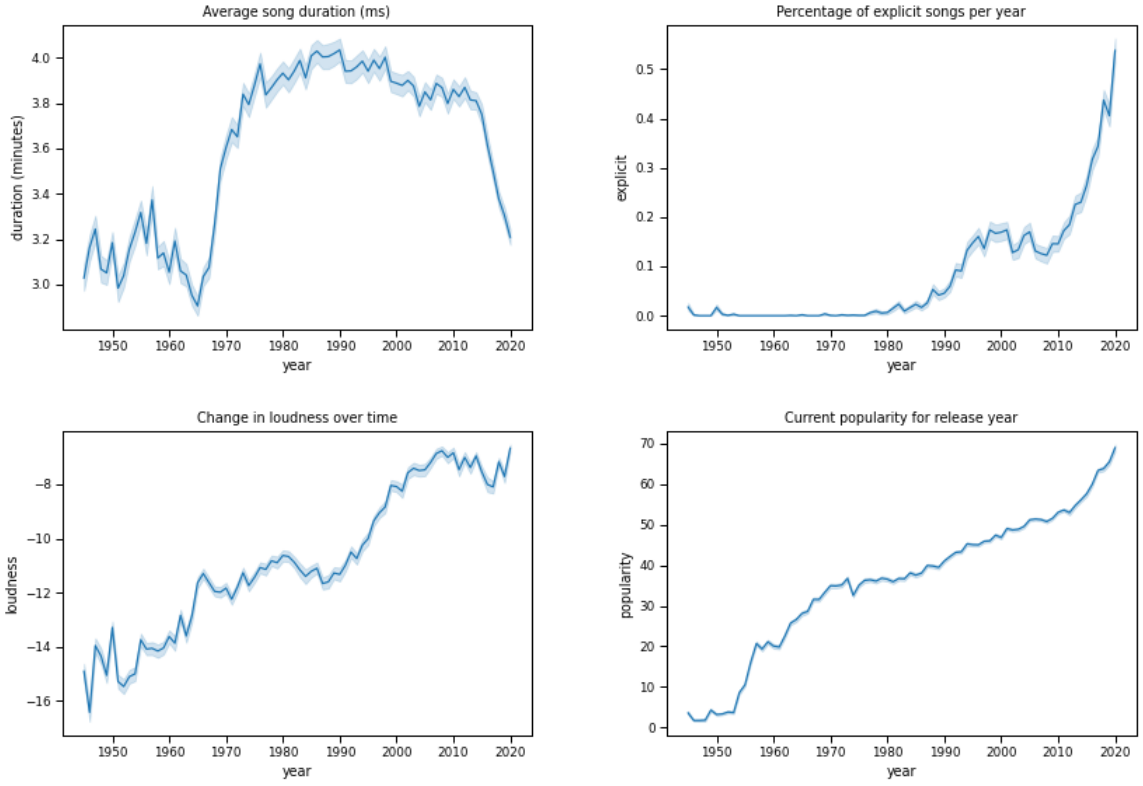
valence	year	artists	danceability	duration_ms	explicit	loudness	name	popularity	speechiness	tempo
0.763	1983	['ZZ Top']	0.611	273907	0	-5.357	Legs - 2008 Remaster	61	0.0378	125.398
0.159	1957	['Thelonious Monk']	0.526	558267	0	-19.983	Functional	8	0.0793	136.973
0.961	1964	['Dusty Springfield']	0.602	158200	0	-8.021	I Only Want To Be With You	58	0.0290	133.000
0.328	1968	['The Vogues']	0.171	166301	0	-12.494	Turn Around, Look at Me	41	0.0313	180.861
0.572	1967	['Jimi Hendrix']	0.479	163800	0	-5.167	Fire	52	0.1150	153.450
0.715	1949	['Lehman Engel']	0.627	132867	0	-9.415	Show Me (from My Fair Lady) - Voice	0	0.0389	109.283
0.848	1980	['Steve Winwood']	0.678	220293	0	-6.861	Second-Hand Woman	27	0.0326	132.111

Table 1. *Sample of the data after selecting features*

The dataset contains no empty or null values, and so no further action was taken to try and fix these issues. All features were also of the correct data type, and there are no categorical variables that require codification or encoding.

Analysis: music over the years

Features were graphed along time to try and find changing characteristics in music trends:



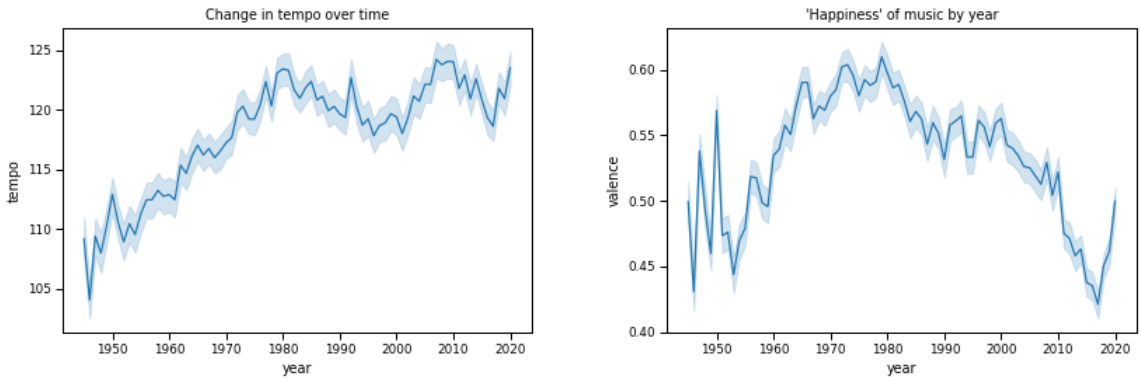


Figure 2. All features through time

From the above figures, we can make the following observations:

- The average song duration increased drastically in the 1970's, from 3 minutes to almost 4. The duration dipped again in the mid-2010s.
- The percentage of explicit songs has consistently increased through time, with a massive uptick in 2010.
- The average loudness of music has also increased consistently and has remained at around -8 dB since the 2000s.
- More recent music is also more popular.
- Average tempo got higher from the 1950s to 1980 and has remained between 115 and 125 bpm.
- The perceived happiness of music picked at the late 1970s and early 1980s; it then dipped into its lowest score in the mid-2010s, later bumping up again until today.

We then obtained the correlation coefficient for all variables, and plotted them in a heat map:

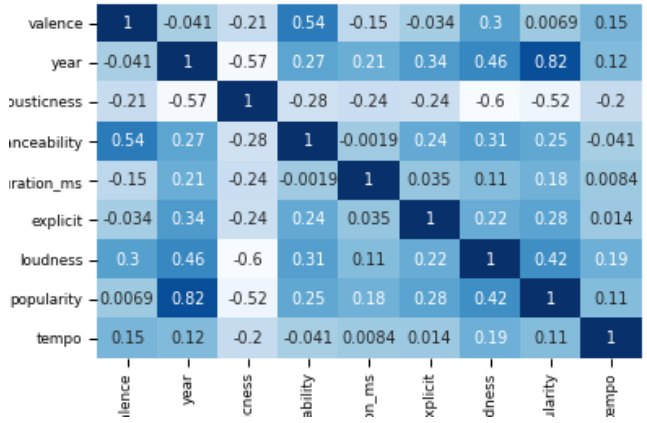


Figure 3. *Correlation matrix*

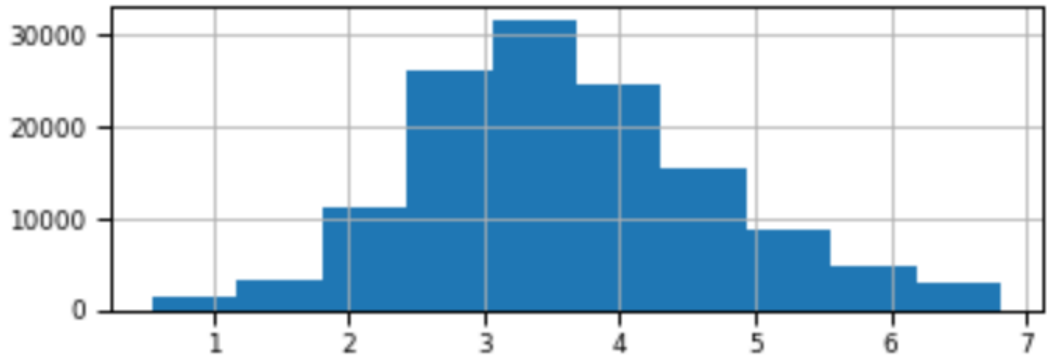
The only strong correlation between variables in between the release year and popularity, with **0.82**. All other features have a mid to low correlation.

Hypothesis testing: guessing the new biggest hit

We will proceed with a few hypothesis tests using the dataset:

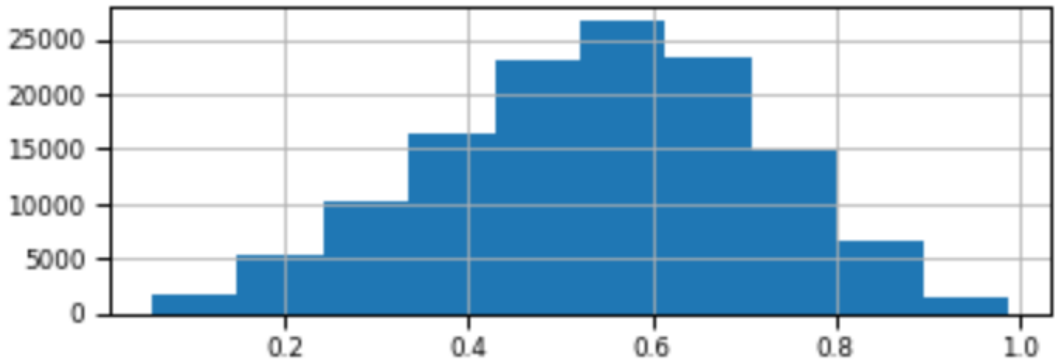
1. **The probability of getting a song of duration lower than 2.5 minutes**
   Assuming a normal distribution for the duration feature, we can estimate the probability of randomly selecting a song of 2.5 minutes or less:

```
Mean =  3.616617886502239
Std =  1.1137574562874817
Prob of duration < 2.5 minutes = 15.803461%
```



2. **The probability of getting a danceability score of 0.6 or higher**
   Assuming a normal distribution for the danceability feature, we can estimate the probability of randomly selecting a song of danceability of 0.6 or higher:

```
Mean =  0.5413757799043908
Std =  0.17417645179178232
Prob of danceability > 0.6 = 36.821696%
```



3. **The probability of getting at least 2 explicit songs from a shuffle of 10**
   Using a binomial distribution, we can estimate the probability of getting 2 or more explicit songs from a sample of 10 songs.

```
Proportion of explicit songs in dataset = 0.085092%
n = 10
k = 2
Probability of getting at least 2 explicit songs from a sample of 10 = 20.687001%
```

The hypothesis tests showed that the probability of getting a short song is around **15.8%**, the probability of getting a danceable song is around **38.8%** and the probability of getting 2 or more explicit songs out of a random selection of 10 is **20.7%**.

Conclusion

The data set is very comprehensible and contains enough data for further analysis and machine learning applications. However, the current features lack correlation between them, and so the quality of certain predictive models would be severely impacted in accuracy.

The most important feature that is currently missing from the set is the **song genre**, which would allow for a more complex analysis of the evolution of music tastes across time. Other measures of popularity would be good for complementing the current feature, like the average # of reproductions per day, or the country which accounts for the most plays of a song.

Overall, Spotify allows us to get a glimpse of how music has changed over the years, and we can use the information to better understand where the music industry is headed.