COM-480 Data Visualization
Milestone 1

Auguste Baum, Yanis Berkani & Clément Petit

23 April 2021

1 Dataset

As three real music enthusiasts, we've always wanted to learn more and do some scientific research in the music domain. Accordingly, the dataset we chose is a collection of music data from a Kaggle competition[1].

More precisely, the data consists of music features for about 600 000 tracks and 1.1 million artists from the Spotify streaming service, spanning 100 years (from 1922 to 2021). The data are divided into 2 main datasets:

- **Tracks:** name, date of release, popularity, duration, energy, tempo...
- **Artists:** name, followers, popularity, genres.

Note that you can find out a lot more details about the features in our `EDA.ipynb` notebook in our Github repository[2] (works best on Firefox).

As often on Kaggle the data is quite clean, but it is still very important to do some cleaning and processing to make the data fit our project's main goals. Hence, we started by studying and understanding the two datasets before merging the information. Then, we cleaned the merged dataset; for example we handled strange track names, we removed completely silent tracks, we dropped duplicates and so on. Finally, we processed the data to best fit our project main ideas. For instance, one of them is to visualize the evolution of popular genres, hence we computed a dataframe containing the top 10% songs in terms of popularity along with their genre for each year.

[1] https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks
[2] https://github.com/com-480-data-visualization/data-visualization-project-2021-vizbrains

2 Problematic

With **Music Trends**, we would like to visually depict the evolution of music throughout the years. Typically, we would like to answer questions such as:

- What kind of music was trending in a given period of time? e.g. what did our parents/grand-parents listen to?
- Which artists/songs were the most popular at some point in time? Which of those were pioneers in their style ?
- How did popular artists/songs evolve through time?
- How did music evolve over the course of the last century ? e.g. is it more energetic and 'danceable' than before?

Seeing how the dataset contains many uncommon metrics about a specific song, such as '*danceability*', '*speechiness*', and more common metrics like '*popularity*', and that it covers a large period of time, our work might really bring some interesting information about music. For instance, it could be used as a data explorer that helps casual listeners ascertain what traits they tend to look for in music, putting their tastes in perspective and allowing them to know which styles they could look for in the future.

Such visualizations would clearly be interesting for any music enthusiast as well as for people who like to see the evolution of trends, hence it is quite wide-reaching.

3 Exploratory Data Analysis

We chose to use the `pandas` Python library for the Exploratory Data Analysis. The notebook containing our EDA can be found in our Github repository. As already mentioned in section 1, we first got insights about the datasets by displaying some basic statistics, then we merged the two datasets before cleaning and processing the merged dataset.

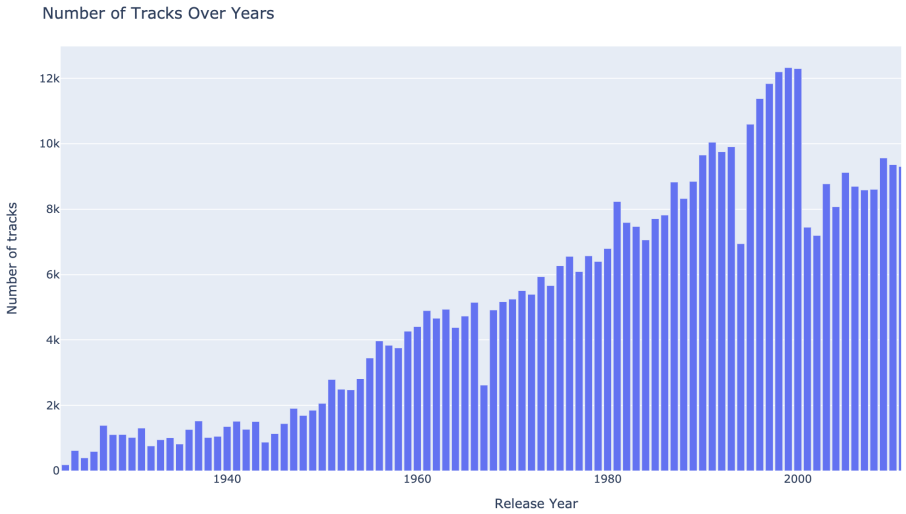
We write some of findings here.

First, the number of songs per year in our dataset is not uniform, as seen in figure 1a.

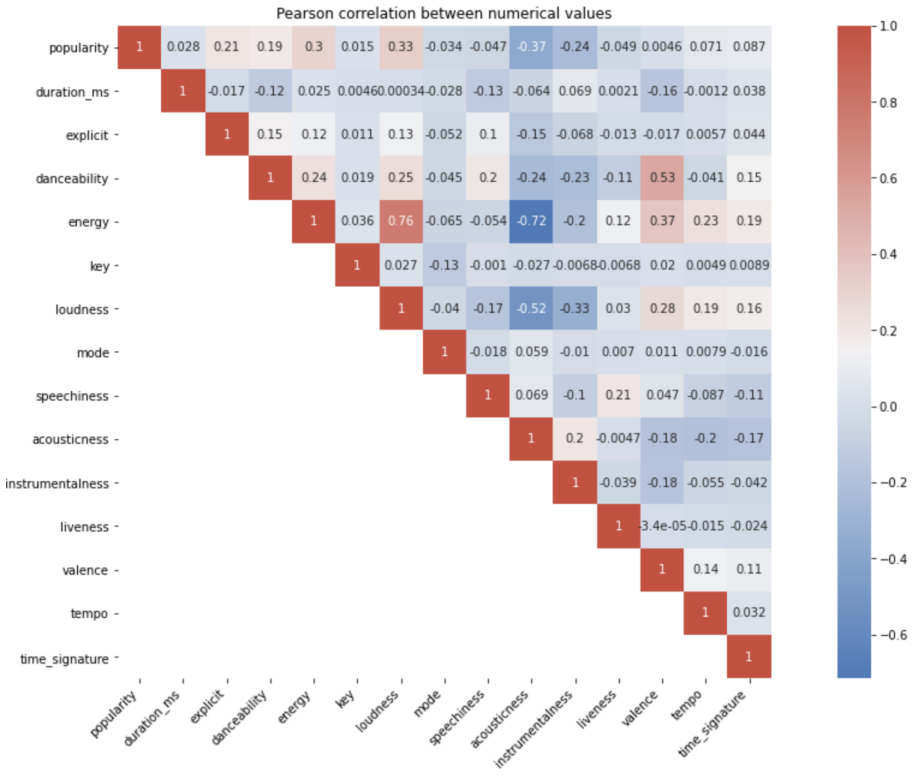
Second, in general the different features don't appear highly correlated, though there are some exceptions (see figure 1b).

4 Related Work

There are a lot of research that are conducted on musical data as it is a very interesting and popular topic. However, we believe our approach to be original as it will bring new visual representations of the evolution of music over the last century, rather than a textual description.



(a) The number of songs by year in the dataset



(b) The Pearson correlation of each pair of features

Figure 1: Notable data analysis findings

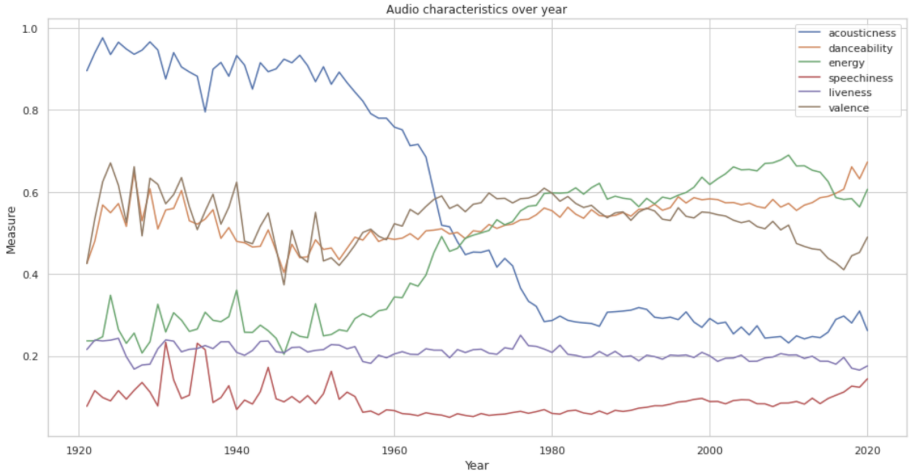
Moreover, since our dataset comes from a Kaggle competition, people have done many things with the dataset we chose. In general, it was used to perform recommendation or predictions, but some have also analyzed the trends of songs over time. However, we did not find work that were similar to what we have in mind. For instance, we wish to find trending genres by selecting popular songs first and then analyzing their genre whereas others consider popular genres as genres that appear the most. Also, they do not go further than "who are the most popular artists and tracks?" whereas we would like to actually study popular artists and tracks evolution. Finally, we will especially focus on the visualization aspect, which was not the case for the related work that we found. You can find some examples of the analysis we found on Kaggle in figure 2.[3]

As you can see below, we have very diverse sources of inspiration:

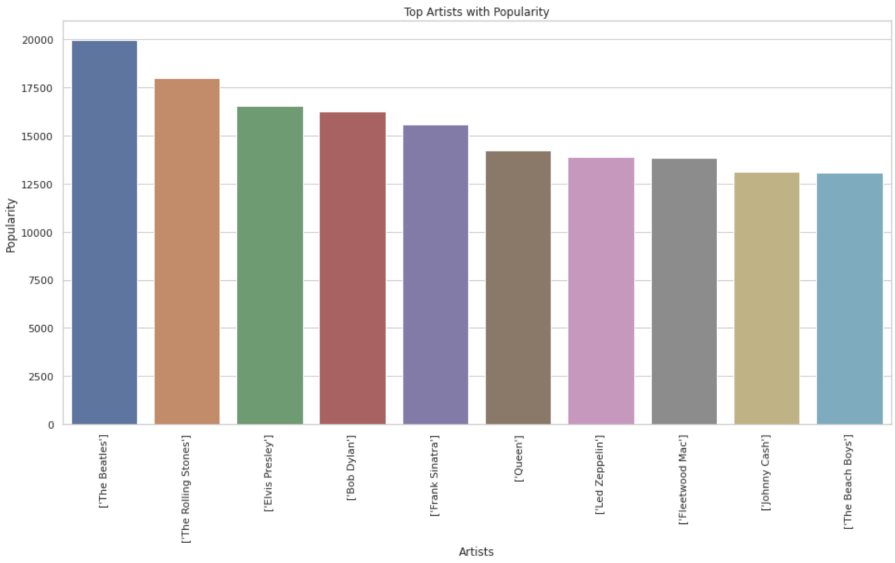
- The DataIsBeautiful **subreddit** for visualizations that effectively convey information is one of our principal source of inspiration. A few examples that are not related to music but that we found particularly interesting are this video about the evolution of the smartphone market, this radial bracket of the UEFA Champions League 2020/21, this chart about **Google** trends in 2020, this chart about **Disney's** live action movies vs animated movies...
- A **Youtube** video showing the evolution of the 15 richest people in the world: Top 15 Richest People In The World (1997-2019)
- A nice and short textual article describing music trends by decade since 1940.
- Another article giving three interesting ways to look at the data behind music.
- An example of last year's projects also about musical data but focusing on few well known artists and their songs: Hit Artist Analyzer on Github

Note that none of us have already explored this dataset in another context.

[3] https://www.kaggle.com/mohitkr05/spotify-data-visualization



(a) How the feature averages change with the year



(b) The overall top artists by popularity

Figure 2: Some graphs made by Kaggle user mohitkr05