Data Visualization - Milestone 1

Inès Kahlaoui (310587), Romain Berquet (316122) & Antoine Munier (314500)

7th April 2023

Dataset

For our data visualization project, the KIBRAM project, we wanted to study music trend and decided to make use of Spotify's Web API, which makes an excellent source of data for music-related tasks as it contains data and metadata about basically every track on the platform.

We created a baseline application to be able to collect Spotify's Web API data and metadata.

We make use of Spotify's Web API to collect the following information available as of April 2023 :

- Artists on the platform
  This dataset contains information about various artists on the platform, notably, for a given artist : name, reference image, number of followers on the platform, popularity score (0-100), music genre
- Audio Tracks on the platform
  This dataset contains information about various audio musical tracks on the platform. We chose not include podcast / audiobooks / shows / episodes in this dataset as they are not relevant for the purpose of our project. Relevant features are : name, popularity, album, artist
- Albums on the platform
  This dataset contains information about various music albums on the platform, notably, for a given album : name, type, artists, release date

This dataset is relatively clean, as it is updated every few days, but requires some pre-processing steps to check for any outliers or anomalies in the data that could affect our visualizations. We notably need to double down on the size of the data (which is relatively large), delete doublons and NaN or undefined values and select relevant features for the purpose of our project.

We also collected metadata of each track, artist and album for probable future usage.

Problematic

In this project, we want to explore the different relationships between a music's success and its audio representation. To do so, we want to create data visualizations that showcase insights into the most popular music on Spotify in 2023.

To get more specific, we want to understand (see) what makes a song and an artist popular, by creating links between the said song's (or artist's) features.

As a baseline, and as said before in the Dataset section, we will restrict ourselves to analyze the top 200 artists and top 500 tracks of 2023 on Spotify based on a popularity score given by the platform, ranging from 0 to 100. We also selected from Spotify's albums dataset only the albums which artists figured in the top 200 artists or in the top 500 tracks.

Our focus will be on analyzing trends and patterns among the top artists, tracks, and albums and identifying any relationships between them. We also want to offer a new way of experiencing music via data visualization. This would mean incorporating music features in it (like playable audio files). We thus want to create a data visualization on Spotify's data to answer the following questions :

- How do audio features and the popularity of a song relate to each other?
- What types of music genres are popular nowadays? How does the genre of a music relates to a song's audio features?

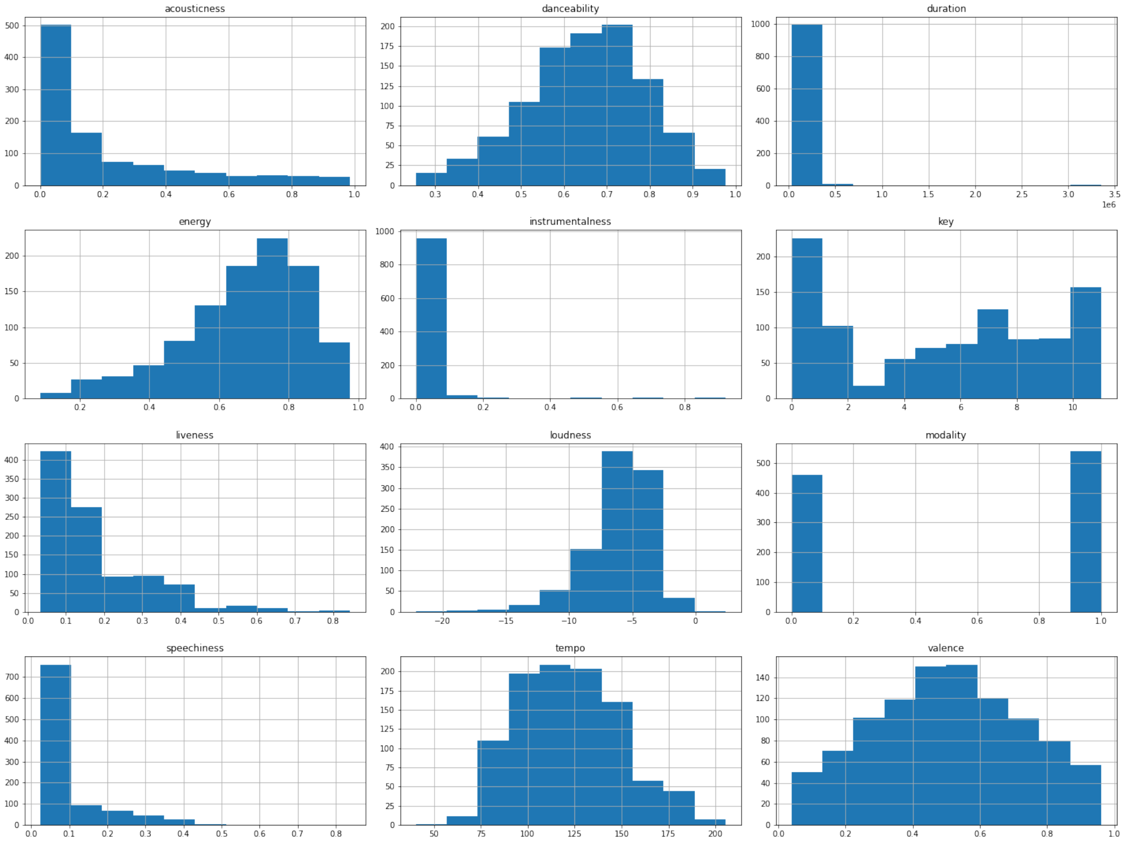
- Who dominates the top charts on Spotify ? Are small artists really far behind ?

This project targets a large set of people. On one hand, music enthusiasts and Spotify users, may be curious about or wish to understand the current music landscape. On the other, some industry professionals might want to get insights about current trends or make use of this data to predict future music trends.

Exploratory Data Analysis

As a first approach, we decided to focus on particular track's / artist's / album's features, cited in the 'Dataset' section, but keep as a resource the Spotify's Web API requests to collect additional features (if they allow us to obtain a better visualization).

As most of the artists, albums and tracks features are non-mathematical, we decided to focus on as a first exploration into the audio features. We found some interesting distribution of features, with some more diverse than others.

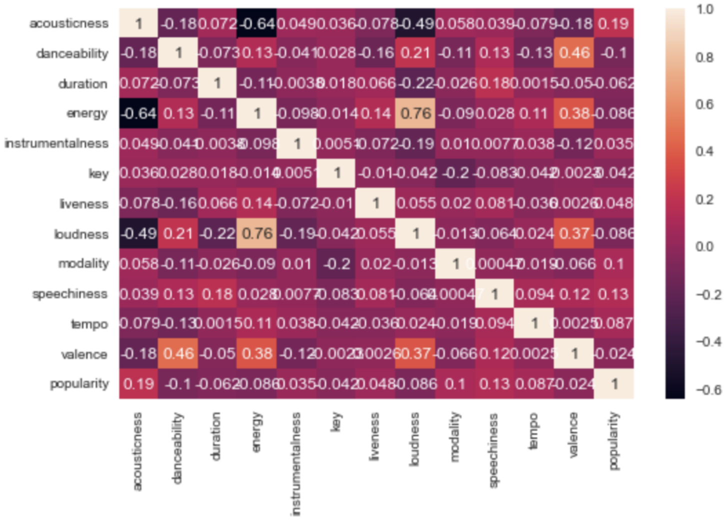


We also considered studying the correlation between features (see figure above).

This would allow us to use machine learning techniques to cluster the data based on sensory audio features. The resulting cluster visualisation would lead to a more sensorial experience when navigating through it, considering that music would be played during the visualisation. On the same note, we decided to collect the Spotify URL and URI for each track in our dataframe for future use. This would for example to play 30-second snippets of a given track.

With some machine learning techniques, notably Principal Analysis Component and K-means clustering, we found some alternative representation and visualization of music. At this stage of the project, the meaning of this representation by component analysis is yet unknown, but we keep it as resource for the future to explore it further.

We also realized that it would be more insightful to encourage a visualization that allows users to compare a selected number of tracks / artists / albums. This would allow for more user interaction and would also serve as a tool tailored to the user's preferences.



Related work

Our first inspiration for this project came from Kirell Benzi's interactive fulldome installation: the Jazz Luminaries. This piece of art represents the essence of our project : convey music, i.e. multimodal music.

Several projects make use of Spotify's data to study different topics. We particularly found out a similar (but overall different) project made in the context of the Data Visualisation course (cf. GitHub). We believe though that we are treating Spotify's API in a different way than this project.

We also refer to this article that gave us some insights about how to deal with the Spotify data by studying Spotify 2021 Trends. This also reassured us in the fact that the data would be clean enough for us to achieve good visualisations.