Milestone 1

Dataset

For this project we will use the dataset “Spotify Top 200 charts (2020 – 2021)” found on the website Kaggle, it is all the songs that have been on the Top 200 weekly charts of the platform Spotify in 2020 and 2021. This data set contains mainly 3 types of data: texts, dates and numbers. More precisely, it contains 23 columns which represents:

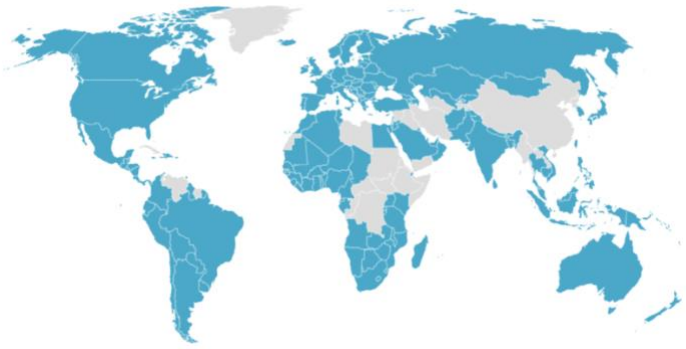
- Highest charting Position: The highest position that the song has been.
-Number of Times Charted: The number of times that the song has been on Top 200 chart.
-Week of Highest Charting: The week when the song had the Highest Position.
-Song Name: Name of the song.
-Song iD: The song ID provided by Spotify.
-Streams: Approximate number of streams the song has.
-Artist: The main artist/ artists involved in making the song.
-Artist Followers: The number of followers the main artist has on Spotify.
-Genre: The genres the song belongs to.
-Release Date: The date that the song was released.
-Weeks Charted: The weeks that the song has been on in the chart.
-Chord: The main chord of the song instrumental.
-Tempo: The overall estimated tempo of a track in beats per minute (BPM).
-Popularity
-Danceability
-Acousticness
-Energy
-Instrumentalness
-Liveness
-Loudness
-Speechiness
-Valence

The last 9 variables are scores defined by the platform Spotify going from 0 to 1 and are used to define more subtill characteristic of the songs. The data is well organized no parsing, or any other preprocessing would be needed.

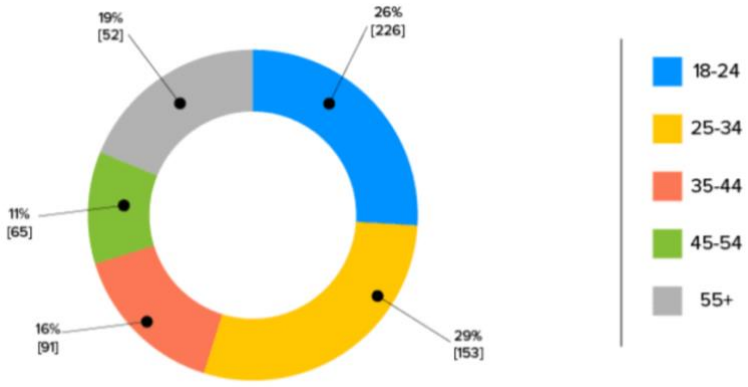
Problematic

From this dataset we could start by showing the ranking of the song that stay the longest and the highest in the top chart of Spotify in 2020 and 2021. But more interestingly we could derive correlation with the type of song and the time it stays in the top chart. For example, which characteristic between Energy and Instrumentalness is the more significative to predict the success of a song. We could even try to find if there is a common pattern shared by big hits in 2020 and 2021.
Of course, these numbers would concern the target audience of the platform Spotify.

Here we can see the country where Spotify is available:

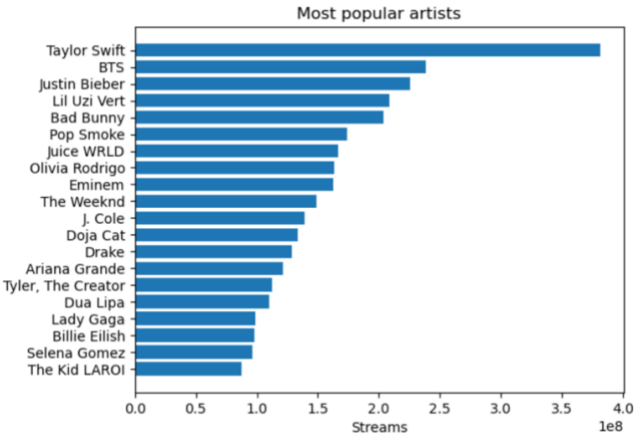


Another important statistic is the age of the typical user of the platform:

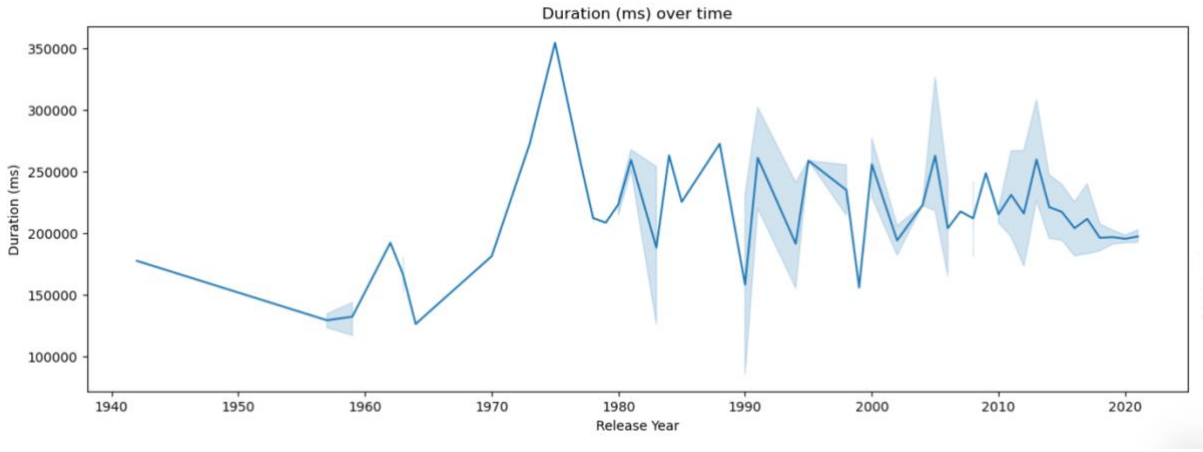


Exploratory Data Analysis

We have done some preprocessing and exploratory data analysis that can be found in the notebook “EDA_spotify.ipynnb” in the GitHub repository.
There were format and type issues that we fixed, and we plotted a few graphs to study the distribution of some of the many variables.
Per example, we can see below the most popular artists on the top 200:



Or the evolution of the duration of a song over the years:



Related works

Researching for some related works that have already been done with the data, we mainly found data explorations made for a preprocessing to achieve machine learning tasks such as a classification or prediction. If visualizations were used, they were mostly simple and highlighted only a few numbers of features. The purpose was only to get the main idea and proceed to a clear analysis between artists and popularity and genre of the music present in the top charts.

With our visualization, we would like to push those visualizations and analysis further by giving a better idea and feeling around the songs and catch the mood of the music present on the top charts. We would like to add complete illustrations playing with colors to give the best representation of the emotion that a song can give. It would also be interesting to add visualizations with connections between genre, keys, and the sentiment that it may procure to the listener.

Our inspiration and main idea came from these articles, where we found the direction of the analysis very interesting, but the proper visualization still needed to be done.

- https://towardsdatascience.com/predicting-popularity-on-spotify-when-data-needs-culture-more-than-culture-needs-data-2ed3661f75f1
  They take a more important look of what causes the popularity, which is directly linked to the feeling that a music gives.

- https://towardsdatascience.com/spotify-sentiment-analysis-8d48b0a492f2
  They proceed to a sentiment analysis, taking into account the audio valence which is some kind of measure to determine how positive/negative a music sound.

Spotify already created such a visualization, *Spotify Audio Features,* but we could not properly understand and did not find it intuitive. It is sadly the only interesting and innovative

illustration that we could find on the subject. But we would like to take it as a starting point and just make something revealing the same information but in a clearer way and after applying it for some individual songs, reveal the bigger picture of the top chart.