# Information Visualization

# CHECKPOINT II: Data cleaning and processing

G52-T

**1. Initial Dataset**

The initial dataset consists of the daily top 200 most streamed tracks on Spotify, for the year of 2019, according to the Spotify Charts. At the Spotify Charts’ website, the chart for each day is available to download in the format of a CSV file. In addition to the information available in each chart we used the information available from the Spotify API endpoints /v1/tracks/{id} and /v1/audio-features/{id} to gather more information for each unique track. Each entry of each the chart contains the position of the track in the chart, the name of the track, the artist, the number of streams, and the URL. The additional data includes its duration, acousticness, danceability, energy, tempo, instrumentality, among others**.**

(from Spotify Charts CSV)

Position,"Track Name",Artist,Streams,URL

1,"Dance Monkey","Tones and I",5180626,https://open.spotify.com/track/1rgnBhdG2JDFTbYkYRZAku

2,"HIGHEST IN THE ROOM","Travis Scott",4423871,https://open.spotify.com/track/3eekarcy7kvN4yt5ZFzltW

3,Señorita,"Shawn Mendes",3916024,https://open.spotify.com/track/6v3KW9xbzN5yKLt9YKDYA2

(from Spotify API /v1/tracks/{id})

{ "album": {...}, "artists": [...], "available\_markets": [...], "disc\_number": 1, "duration\_ms": 207959, "explicit": false, "external\_ids": {...}, "external\_urls": {...}, "href": "https://api.spotify.com/v1/tracks/11dFghVXANMlKmJXsNCbNl", "id": "11dFghVXANMlKmJXsNCbNl", "is\_local": false, "name": "Cut To The Feeling", "popularity": 63, "preview\_url": "https://p.scdn.co/mp3-preview/3eb16018c2a700240e9dfb8817b6f2d041f15eb1?cid=774b29d4f13844c495f206cafdad9c86", "track\_number": 1, "type": "track", "uri": "spotify:track:11dFghVXANMlKmJXsNCbNl" }

(from Spotify API /v1/audio-features/{id})

{ "duration\_ms": 255349, "key": 5, "mode": 0, "time\_signature": 4, "acousticness": 0.514, "danceability": 0.735, "energy": 0.578, "instrumentalness": 0.0902, "liveness": 0.159, "loudness": -11.840, "speechiness": 0.0461, "valence": 0.624, "tempo": 98.002, "id": "06AKEBrKUckW0KREUWRnvT", "uri": "spotify:track:06AKEBrKUckW0KREUWRnvT", "track\_href": "https://api.spotify.com/v1/tracks/06AKEBrKUckW0KREUWRnvT", "analysis\_url": "https://api.spotify.com/v1/audio-analysis/06AKEBrKUckW0KREUWRnvT", "type": "audio\_features" }

**2. Selected/Derived Data**

All the variables from the CSVs were selected. From the first API endpoint were selected: album art URL, explicit and preview URL. From the second one all were selected with the exception of: type, analysis URL, track HREF, URI and id. A *date* variable was added based on the day of the chart. The variable *id* in the resulting dataset was derived from the variable URL obtained from the CSVs in order to make the API requests. No further derived variables were calculated because is the end user that will define the time intervals that would be needed to calculate such variables.

**3. Data abstraction**

The extracted data is tabular and it is stored in JSON format. Each record represents a track in the top 200 of a particular day, thus each record is identified by the position and date attributes. The dataset contains one variable of the type quantitative sequential, the position of the track in the chart; six of the nominal type, the track’s name, artist, URL, id, album art URL and preview URL; two ordinal, the date of the chart data and the key that represents the overall key of the track using the pitch class notation; two binary, the mode that indicates whether that track is in major (represented by 1) or minor (represented by 0) and explicit to indicate if the track has explicit lyrics or not; and 12 variables ratio sequential, streams, the number of streams in that day, duration ms, the track duration in milliseconds, time signature, the number of beats in each bar, loudness, the overall loudness of the track in decibels, tempo, the estimated beats per minute (BPM) of the track, acousticness, a confidence measure of whether the track is acoustic, danceability, which describes how suitable a track is for dancing, energy, a perceptual measure of intensity and activity, instrumentalness, that predicts whether a track contains no vocals, liveness, which detects the presence of an audience in the recording, speechiness, that detects the presence of spoken words in a track and valence, that describes the musical positiveness conveyed by the track, the last seven variables take values between 0.0 and 1.0.

* + - 1. **4. Dataset processing**

1. Since the dataset was created by us, we cleaned the data while gathering it with the questions in mind, that way no further cleaning was necessary afterwards. The necessary information was already considered when creating the dataset, that way most of the cleaning was done before hand. The only variable in the dataset that contains missing values is the track preview URL, that when the value is missing takes the value *null*, that means that for those tracks the end user won’t be able to preview the track, which is not really a problem.
   * + 1. **5. Mapping (Data sample / Questions)**

* “Are music trends affected by the song duration?” can be answered using the variables duration ms and streams in conjunction to check if it is or not.
* “What’s the evolution of the track’s streaming numbers during a period?” can be answered using the date and streams variables and the time period given by the user.
* “How does the features (energy, loudness, tempo, etc) of the track affect its popularity?” can be answered by using streams and the given track feature.