CareerFoundry - Data Analytics Program

Alina Racu

Data Immersion

6.1 Sourcing Open Data

Data Source

I sourced two external data sets with information related to music from Spotify, the worlds largest music streaming service provider with over 381 million monthly active users as of 2021.

The data sets were downloaded from Kaggle platform:
https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking/activity
https://www.kaggle.com/rodolfofigueroa/spotify-12m-songs

Data Collection

Data set **"tracks_2017"**: The data has been retrieved from Spotify's regional charts: https://spotifycharts.com/regional. The database was last time updated December 12, 2018.

Data set **"tracks_features"**: The process of assessing song features is automatized so I consider the data to be reliable.

Data Contents

Data set **"tracks_2017"**: The data set contains information on the daily ranking of the 200 most listened songs in 53 countries from 2017 and early 2018 by Spotify users. The data spans from January 1, 2017 to January 9, 2018. Each row contains a ranking position on a specific day for a song. For instance, the first 200 rows present the ranking for the 1st of January in Argentina. The following 200 rows will contain the ranking for the 2nd of January in Argentina.

Since the data set did not have a sufficient number of variables required by the project, I decided to merge it with another data set, "tracks_features", that includes features of songs such as loudness, danceability, tempo, key, etc.

Data set **"tracks_features"**: The data set is essentially a music catalog containing audio features for over 1.2 million songs, which were obtained with the Spotify API.

A description of what the audio features entail can be found on Spotify's website:
https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features.

Limitations

Data set **"tracks_2017"**: Spotify data was missing on very few days, so the database is not perfectly complete.

Data set **"tracks_features"**: As song features are assessed automatically, the data is as good as the model processing it and the human inputs that went into that model.

In addition, unfortunately, the "tracks_2017" data set did not have a "track ID" variable, like the set, which prompted me to merge the files on key variable "track" (song name) and "artist". These variables of string type did not fully overlap, probably due to inconsistencies, which led to a smaller merged data set.

Choice of the Data Set

I have chosen these data sets as I have always been interested in music and music theory. I enjoy the Spotify platform on a daily basis listening to my favourite songs or exploring new ones. And finally, I appreciated the wealth and variety of data the two data sets contain which is in line with the project brief's criteria.

Data Profile

Data set **"tracks_2017"**: 3,441,197 rows and 7 columns.

Columns: Position, Track Name, Artist, Streams, URL, Date, Region.

Data set **"tracks_features"**: 1,204,025 rows and 24 columns.

Columns: track_id, track_name, album, album_id, artists, artist_ids, track_number, disc_number, explicit, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, time_signature, release_year, release_date.

Data Cleaning & Consistency Checks

Dataframe **"tracks_2017"**:

- Mixed data types: changed data type in columns "Track Artist", "Artist", "URL" to string;
- Dropping columns: dropped column "URL";
- Renaming columns: renamed columns to be in line with the style of the data set "tracks_features" which had a higher number of variables. Also renamed columns "region" to "country" as well as the country short names to full names in this column;
- Missing values: 657 values missing in variables "track_name" and "artist"; removed the rows containing this missing data;
- Duplicates: none.

Dataframe **“tracks_features”**:

- Subsetting: due to the size of the data set I created a subset of the songs that were released between 2016 and 2018 as newer songs are most likely to be popular in top 200 rankings; the new dataframe has 154,902 rows.
- Mixed data types: no need to change any data types;
- Dropping columns: dropped columns “album_id", “artist_ids”, “track_number”, “disc_number”, “key”;
- Renaming columns: renamed columns , “name” to “track_name”, “artists” to “artist”, “year” to “release_year”; removed brackets in values of column “artist”
- Missing values: none;
- Duplicates: none.

Data Combination

I merged the two dataframes using an inner join on column “track_name” adn “artist”. The new dataframe has 538,930 and 23 columns.

The main limitation of the data set is that the key column used for the inner join was track name rather than track ID, which may be prone do inconsistencies. Thus, after the inner join, the merged dataframe ended up with a much lower number of rows. Nonetheless, I am curious of how this data is shaped and what insights I can get out of it.

Descriptive Statistics

	position	streams	danceability	energy	key	loudness	mode	speechiness
count	538930.000000	5.389300e+05	538930.000000	538930.000000	538930.000000	538930.000000	538930.000000	538930.000000
mean	87.384931	6.902732e+04	0.678937	0.661905	5.657640	-5.872234	0.500633	0.086928
std	59.529680	2.682843e+05	0.147243	0.144795	3.543962	1.879910	0.500000	0.076716
min	1.000000	1.001000e+03	0.143000	0.027900	0.000000	-34.475000	0.000000	0.022900
25%	34.000000	3.608000e+03	0.606000	0.565000	2.000000	-6.714000	0.000000	0.042700
50%	81.000000	9.865000e+03	0.696000	0.698000	6.000000	-5.975000	1.000000	0.059700
75%	138.000000	3.437200e+04	0.773000	0.757000	9.000000	-4.781000	1.000000	0.098900
max	200.000000	9.891056e+06	0.951000	0.991000	11.000000	1.162000	1.000000	0.954000

	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	release_year
count	538930.000000	538930.000000	538930.000000	538930.000000	538930.000000	538930.000000	538930.000000	538930.000000
mean	0.177669	0.003071	0.147666	0.542591	114.637002	213616.170933	3.997551	2016.921942
std	0.221555	0.032487	0.090369	0.211654	23.239277	36297.344581	0.120032	0.664400
min	0.000003	0.000000	0.021900	0.039400	60.309000	46227.000000	1.000000	2016.000000
25%	0.023200	0.000000	0.092800	0.381000	97.092000	199773.000000	4.000000	2016.000000
50%	0.076700	0.000000	0.109000	0.523000	116.073000	213264.000000	4.000000	2017.000000
75%	0.242000	0.000016	0.154000	0.675000	129.923000	228520.000000	4.000000	2017.000000
max	0.990000	0.961000	0.972000	0.982000	207.581000	552240.000000	5.000000	2018.000000

Exploratory Questions

What kind of songs tend to reach the top 10 ranking of Spotify? What kind of songs have the highest number of streams? How long are songs staying in the top ranking? Are there any specific features that make some songs more popular than other?

How is song popularity distributed around regions? Are there certain features that make some songs more popular in one region than in another?

Appendix: Explanations on Variable of the Data set "tracks_features"

acousticness number<float>

A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

>= 0 <= 1

---

analysis_url string

A URL to access the full audio analysis of this track. An access token is required to access this data.

---

danceability number<float>

Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

---

duration_ms integer

The duration of the track in milliseconds.

---

energy number<float>

Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

---

id string

The Spotify ID for the track.

---

instrumentalness number<float>

Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

---

key integer

The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.

>= -1 <= 11

---

liveness number<float>

Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

---

loudness number<float>

The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.

---

mode integer

Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

---

speechiness number<float>

Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

---

tempo number<float>

The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

---

time_signature integer

An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".

>= 3 <= 7

---

track_href string

A link to the Web API endpoint providing full details of the track.

---

type string

The object type.

Allowed value: "audio_features"

---

uri string

The Spotify URI for the track.

---

valence number<float>

A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

>= 0 <= 1