


Project D18: KAGGLE-SPOTIFY DATA

HW10

Team members: Jan Markus Rokka, Urmi Tari, Reena Seeba

Link of the repository: <https://github.com/JanMarkusRokka/SPOTIFY-DATA.git>

Task 2. Business understanding

For our dataset we have picked the “ Spotify Tracks Dataset” from Kaggle. This dataset consists of a large list of songs from the popular streaming platform Spotify, along with 20 parameters describing each track. In the dataset’s description there is a list of examples describing different ways to utilize the data. Our goals are quite similar. We are planning to:

1. build a model to predict danceability based on the song’s other parameters.
2. find the most and least popular genres based on averages.
3. find the parameters which affect the track’s popularity the most.

Our project could be of use to music producers to help make their next song a success (in terms of popularity). It could also be helpful for music platforms like Spotify to generate better personalized playlists. Additionally the process of analyzing this data would develop our own data mining skills and help us consolidate what we have learned so far.

Our project will be considered a success when we manage to finalize our first data model to the point that it predicts relatively well at least 70% of the time. The second goal will be considered met after we find the “most” and “least” popular genres based on averages calculated on the given dataset. The third venture will be considered successful after we find which parameter(s) are most correlated with a high popularity value.

For data analysis we will be using Jupyter Notebook along with Python. As for the dataset we are planning to use all of the songs as well as most of the given parameters.

Possible risks and contingencies concerning the project include:


1. tuning our model on the test data and therefore causing our model to be bad at generalizing.
2. overfitting.
3. relying too much on a single method.
4. picking a biased sample.

Here are some column names from our selected dataset, which might require an explanation (derived from Kaggle):

- **Popularity** - “The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.”
- **Danceability** - “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”
- **Energy** - “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.”
- **Key** - “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”
- **Mode** - “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** - “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.”
- **Acousticness** - “A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic”
- **Instrumentalness** - “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”
- **Liveness** - “Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.”
- **Valence** - “A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.”
- **Time signature** - “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.”

Task 3. Business understanding

Our data requirements are such that we require a dataset that describes songs from many different genres and with differing levels of popularity. The dataset should also include various columns describing the song’s popularity, genre, tempo along with some columns potentially derived from other columns such as “energy”. Our selected dataset, “ Spotify Tracks Dataset”, seems to match our requirements quite well. The data is also consistent, because it has been taken from Spotify.

As has been described in the previous task, the data types that are necessary are:

- track id (number) - for distinguishing the tracks
- popularity (number between 0 and 100) describing the popularity of the song - used to predict danceability
- genre (String) the genre of the track - used to find the most and least popular ones in the dataset
- duration_ms (number) describing the duration of the track - to see, if it affects the popularity of the track and to predict danceability
- explicit (boolean value or unknown) whether or not the track has explicit lyrics - to see, if it affects the popularity of the track and to predict danceability
- energy (number between 0.0 and 1.0) represents a perceptual measure of intensity and activity - to see, if it affects the popularity of the track and to predict danceability
- key (integer map) the key the track is in, integers map to pitches using Pitch Class notation - to see, if it affects the popularity of the track and to predict danceability
- loudness (number) the loudness of track in dB - to see, if it affects the popularity of the track and to predict danceability
- mode (1 or 0) indicates the modality (major - 1 or minor - 0) of a track - to see, if it affects the popularity of the track and to predict danceability
- speechiness (number between 0.0 to 1.0) presence of spoken words in a track - to see, if it affects the popularity of the track and to predict danceability
- acousticness (number between 0.0 to 1.0) confidence measure whether the track is acoustic - to see, if it affects the popularity of the track and to predict danceability
- instrumentalness (number between 0.0 and 1.0) describing the instrumentalness of the track - to see, if it affects the popularity of the track and to predict danceability
- liveness (number between 0.0 and 1.0) presence of an audience in the recording - to see, if it affects the popularity of the track and to predict danceability
- valence (number between 0.0 and 1.0) describing the musical positiveness conveyed by a track - to see, if it affects the popularity of the track and to predict danceability
- tempo (number) the overall estimated tempo of a track in beats per minute - to see, if it affects the popularity of the track and to predict danceability

- time_signature (number between 3 to 7) notational convention to specify how many beats are in each bar (or measure) - to see, if it affects the popularity of the track and to predict danceability
- track_genre (String) the genre the track belongs to - used to see if it affects danceability;

All these necessary types and values of data are in the dataset and will be used to complete the task. We might also use the column artists (contains the name of the artist) in relation to the third goal. We will see if there is a correlation between the artist and the track's popularity. Our main goal is to analyze track features to understand what makes certain songs popular. Since we are not planning on doing language analysis, we currently might not be able to get any useful results out of analyzing the track's and/or artist's and/or album's name.

To verify the data we analyzed it and decided that the ratio between repetitive values and unique ones is not going to affect our analysis in a negative way.

Task 4. Planning our project

List of tasks:

Task	Estimate time spent (hours)	Responsible
Data preparation + HW10	10 + 10 + 10	Reena, Jan, Urmi
Build a model to predict danceability based on the song's other parameters	10	Jan
Find the most and least popular genres based on averages.	7	Urmi
Find the parameters which affect the track's popularity the most.	10	Reena
Analyzing the results of the 1st goal + writing a summary of the results	10	Jan
Analyzing the results of the 2nd goal + writing a summary of the results	7	Urmi
Analyzing the results of the 3rd goal + writing a summary of the results	10	Reena
Poster design	10	Urmi

The tools that we are planning to use for data analysis will be Jupyter Notebook along with Python. Methods we are planning to use for the purposes of data analysis include regression analysis (3rd goal), cluster analysis (1st goal) and basic statistical analysis formulas (mean, sum, etc.) (2nd goal).