Spotify Dataset

This month we are analyzing the tracks of Spotify!

Can we identify what makes a hit track?

Structure

**Primary**:

* – **id** (Id of track generated by Spotify)

**Numerical**:

* – **acousticness** (Ranges from 0 to 1)
* – **danceability** (Ranges from 0 to 1)
* – **energy** (Ranges from 0 to 1)
* – **duration\_ms** (Integer typically ranging from 200k to 300k)
* – **instrumentalness** (Ranges from 0 to 1)
* – **valence** (Ranges from 0 to 1)
* – **popularity** (Ranges from 0 to 100)
* – **tempo** (Float typically ranging from 50 to 150)
* – **liveness** (Ranges from 0 to 1)
* – **loudness** (Float typically ranging from -60 to 0)
* – **speechiness** (Ranges from 0 to 1)

**Dummy**:

* – **mode** (0 = Minor, 1 = Major)
* – **explicit** (0 = No explicit content, 1 = Explicit content)

**Categorical**:

* – **key** (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on…)
* – **time*signature*** (The predicted timesignature, most typically 4)
* – **artists** (List of artists mentioned)
* – **artists** (Ids of mentioned artists)
* – **release\_date** (Date of release mostly in yyyy-mm-dd format, however precision of date may vary)
* – **name** (Name of the song)

# **Win £50 Amazon Voucher!**

A winning entry will be chosen based on best visualization and successfully following the entry rules – winning a £50 Amazon voucher code!

The top 5 entries will also receive 2 data books from our sponsors Packt!

## **How to submit your entry:**

* + Follow Onyx Data on LinkedIn
  + Share a LinkedIn post that contains both a direct @ mention to **@Onyx Data**, and the hashtag **#dataDNA** (it’s OK if you already follow Onyx Data)
  + In your post, share an image of your visualization or dashboard (remember, it must be a ***single image***)
  + Tag, mention, and invite 5 connections to view your post or play along (optional)

Spotify is one of the newest innovations to have come to audio listening and experience with over 125 million subscribers. Though the service has recently begun, it dominates Apple Music and Amazon music in the audio streaming market. From music, they have extended the audio service to Podcasts, Audiobooks, and so on. Spotify Trends helps any content creator/musician in order to understand what listeners prefer and how to compete in this immensely growing market.

### Project Overview

The two purposes of this project are:

1. **Build an ML model**— To predict the popularity of any song by analyzing various metrics in the dataset. This prediction helps any content creator/musician to understand what Spotify listeners prefer to hear more nowadays, which is key in order to compete in the market.

To attain this objective, it's important to start by doing exploratory analysis and achieve a few insights from data. Find out which features are highly correlated with the popularity attribute. The next step is to test different model algorithms and pick the best model based on key evaluation metric (R2 Score).

1. **Build a content-based recommendation system**that can suggest artists for any users. This helps users listen to songs based on their music preferences.

**Data**

The dataset is taken from the Kaggle website. The data used in this was collected from Spotify’s Web API. This is basically a computer algorithm that Spotify has that can estimate various aspects of the audio file. More info on various attributes used in the dataset is found on this [Spotify Developer](https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/" \t "https://www.kdnuggets.com/2021/04/_blank)page.

Some of the key attributes present in each event in the data are:

* **Key**— The estimated overall key of the track. Integers map to pitches using standard pitch class notation. Ex: 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.
* **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.
* **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 the 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.
* **Instrumentalness**— Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
* **Loudness**— The overall loudness of a track in decibels (dB). Values typical range between -60 and 0 dB.
* **Valence**— A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive.
* **Tempo**— The overall estimated tempo of a track in beats per minute (BPM).
* **Popularity**— The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.