**Grade: 96%**

**MUSIC ANALYSIS FOR GENRE CLASSIFICATION & SONG RECOMMENDATION**

**-**

**Group 8**

**-**

**Aditya Uday Tornekar**

**Nisha Rangnani**

**Rishabh Mahesh Upadhye**

**Shreyas Raghavan Sadagopan**

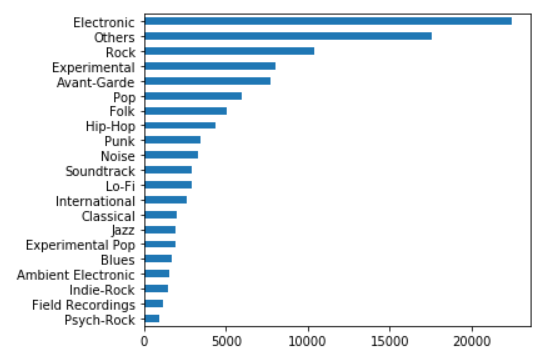
**OBJECTIVE -** We aim to perform analysis on a music dataset to classify audios in various genres while creating clusters and association rules for song recommendations. The user over here will be allowed to select the songs as per his/her taste and would be recommended to a list of songs based on the selection.

**DATA SET DESCRIPTION -** The dataset features 106,574 music tracks and their 518 attributes, which include song title, album, artist, genre, play count, duration, description, etc. The dataset was obtained from UCI Machine Learning and can be found on [FMA: A Dataset For Music Analysis Data Set](https://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis). The dataset zip folder consists of four different CSV files, containing information such as metadata per music track (title, artist, tags, etc.), genre (genre name, parent genre), features (common features extracted), and audio features provided by Spotify for a subset of tracks. We can predict genre, artist, or year the track was released from this dataset. We can use artist ID, type, tags, and other audio characteristics such as spectral bandwidth, centroid, contrast, etc. What surprises and fascinates me the most about this data is the number of unique features each song has, something to which we usually never pay attention.

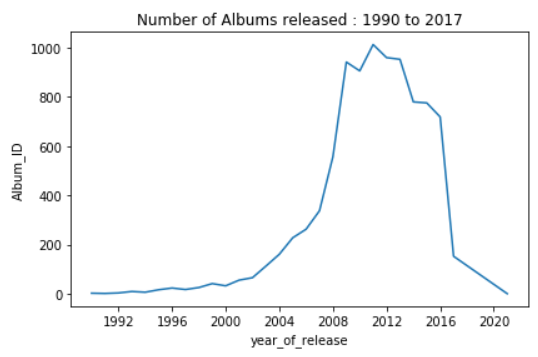
**PRELIMINARY DATA EXPLORATION**

Genre being our dependent variable, we plotted it to see its distribution. We can see that Electronic tracks are

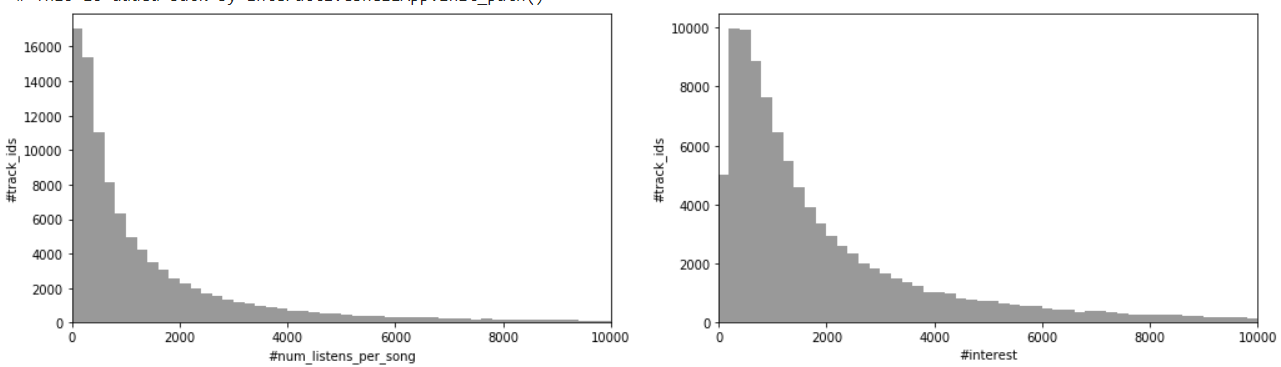
most preferred.

****

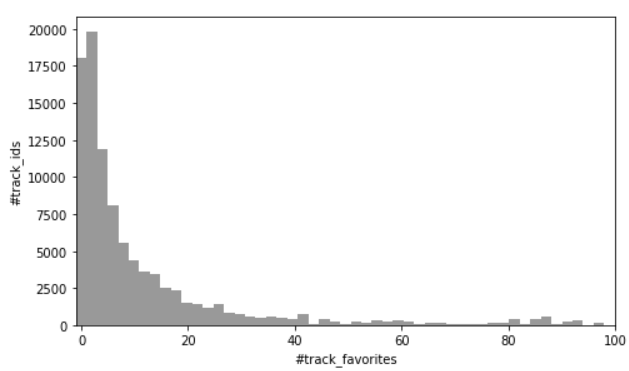
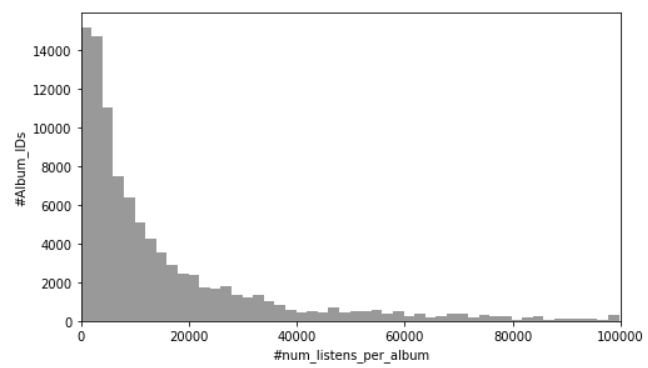
Other attribute visualizations:

****

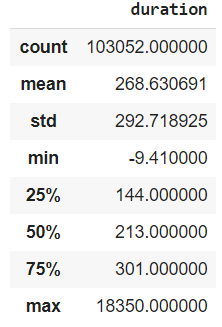
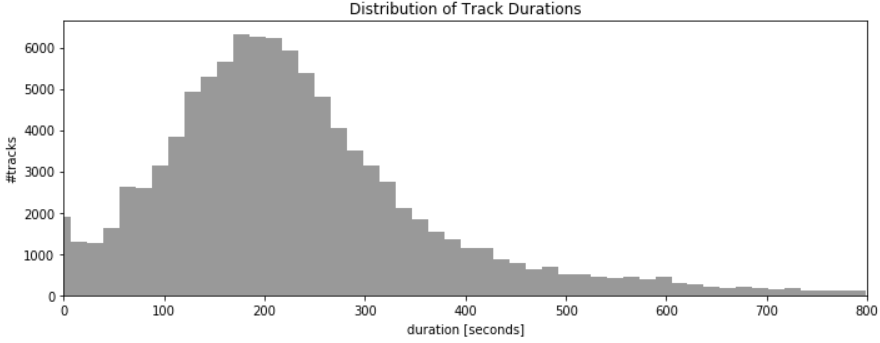
Most albums were released between 2011 and 2013.

****

Track IDs 1 to 2000 were listened to more than 2000 times, higher track IDs were listened to less than 2000 times. Track IDs 1 to 2000 also have an interest score of more than 3000.

****

Album IDs 0 to 2000 have been listened to more than 20,000 times; other track IDs were listened to less than 20,000 times.

****

Tracks are of about 269 seconds on average, up to a maximum of 18,350 seconds. We can see that as

the duration of the song increases over 200 seconds, the number of tracks keeps decreasing.

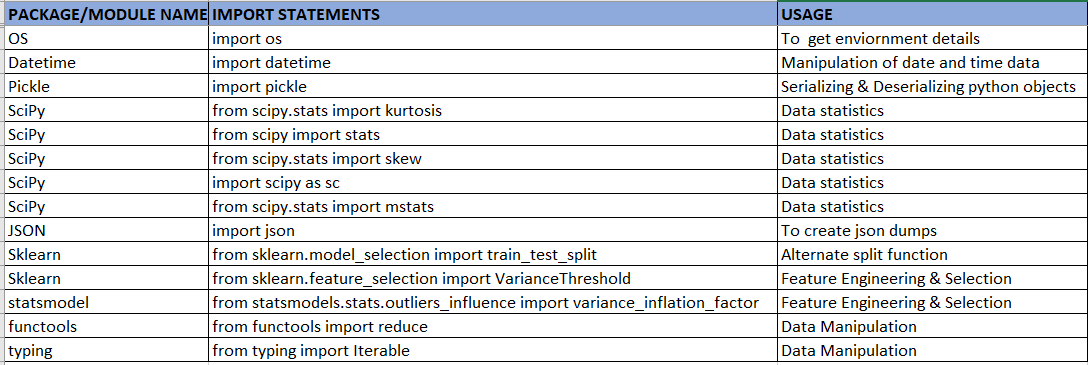
**PREDICTIONS -** Our problem includes the **classification** of the songs as per their genres, and based on these classifications, we would be **recommending** songs to the users. There have been many cases when a track is uploaded to the album, but getting its genre is tedious.

* As it is a multi-class problem, we plan to implement the following algorithms for predicting the track genres: k-Nearest Neighbors, Naive Bayes, Random Forest, Gradient Boosting, Neural Nets. Based on the observations of different metrics we obtain; we would select the champion model and use it.
* We would be recommending the users the songs based on the selection of the users: We would be following a Hybrid approach of recommendation, which would include Content-Based as well as Collaborative filtering.

**INFERENCE -** By this project, we are aiming to successfully classify the songs based on their genres. It would be interesting to observe based on the variables obtained like the number of listens per song, interests, track favorites, and duration of how the categorization would be affected.

* The users can listen to genres that they like, and song recommendations can be provided using this genre preference
* Genre classification helps in building buckets of songs, which makes it easier to search/recall a song
* Song recommendation systems are useful for user engagement

**NON-SPARK PACKAGES (Excluding NumPy, Pandas, Matplotlib, Seaborn, PySpark)**

****