Preliminary Draft

Sayan Samanta
DATA 1030 - Hands On Machine Learning

September 23, 2019

1 Introduction

The dataset under consideration is the **The Billboard 200 acoustic data** which encompasses the entire chart from 1963-2019, along with the EchoNest acoustic features of as many songs as available. All the features of the songs were obtained using Spotify's python API.

The dataset has 2 tables:

1. albums - This contains 574,000 rows which include all the albums that in the Billboard 200 starting from 1/5/1963 up to 1/19/2019. The columns include:
   (a) The date of the record.
   (b) The artist performing the song.
   (c) The album in which the song features.
   (d) The rank of the album in the charts.
   (e) The number of track in album
   (f) The total time of all the songs (in miliseconds)

	id	date	artist	album	rank	length	track_length
0	1	None	None	None	None	NaN	NaN
1	2	2019-01-19	A Boogie Wit da Hoodie	Hoodie SZN	1	20.0	185233.800000
2	3	2019-01-19	21 Savage	I Am > I Was	2	15.0	211050.733333
3	4	2019-01-19	Soundtrack	Spider-Man: Into The Spider-Verse	3	13.0	190866.384615
4	5	2019-01-19	Meek Mill	Championships	4	19.0	219173.894737

Figure 1: Sample view of the albums table

Note. that if the song is not available for streaming on Spotify, then number of tracks and length of the album are represented as NaN

2. acoustic features - This contains Spotify EchoNest acoustic data for tracks appearing on Billboard 200 albums from 1/5/1963 to 1/19/2019. Each row contain the following features
   (a) An unique track ID on Spotify
   (b) Name of the track
   (c) The album on which the track appeared.
   (d) The artist

(e) Acousticness of the track
(f) Danceability of the track
(g) Duration of the track (in miliseconds)
(h) Energy of the song
(i) Intrumentalness of the track.
(j) Key
(k) Liveness
(l) Loudness
(m) Mode
(n) Speechiness
(o) Tempo
(p) Time signature
(q) Valence
(r) An unique album id.
(s) Release Date on Spotify

	id	song	album	artist	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
0	0Veyvc3n9AcLSoK3r1dA12	Voices In My Head	Hoodie SZN	A Boogie Wit da Hoodie	0.0555	0.754	142301.0	0.663	0.000000	6.0	0.101	-6.311	0.0	0.427	90.195	4.0	0.207
1	77JzXZonNumWsuXKy9vr3U	Beasty	Hoodie SZN	A Boogie Wit da Hoodie	0.2920	0.860	152829.0	0.418	0.000000	7.0	0.106	-9.061	0.0	0.158	126.023	4.0	0.374
2	18yllZD0TdF7ykcREib8Z1	I Did It	Hoodie SZN	A Boogie Wit da Hoodie	0.1530	0.718	215305.0	0.454	0.000046	8.0	0.116	-9.012	1.0	0.127	89.483	4.0	0.196
3	1wJRveJZLSb1rjhnUHQiv6	Swervin (feat. 6ix9ine)	Hoodie SZN	A Boogie Wit da Hoodie	0.0153	0.581	189487.0	0.662	0.000000	9.0	0.111	-5.239	1.0	0.303	93.023	4.0	0.434
4	0jAfdqv18goRTUxm3ilRjb	Startender (feat. Offset and Tyga)	Hoodie SZN	A Boogie Wit da Hoodie	0.0235	0.736	192779.0	0.622	0.000000	6.0	0.151	-4.653	0.0	0.133	191.971	4.0	0.506

Figure 2: Sample view of the acoustic features table

Since the Spotify python API is opensource. This data can be complemented with other data (such as songs that did not appear on the Billboard 200) for comparison purposes and further analysis.

The current dataset was curated for a piece on the data science analytics website Components. However that piece dealt with the album length alone, there could be plethora of other questions and analysis that can be answered/performed with this data.

2 Objectives

The preliminary questions that one can be interested in evaluating from this dataset is manifold. Some of which are listed below in no particular order:

1. The major characteristic of a 'chart-topping' track, which is an indirect indicator of the 'trend-setter' at a particular time-frame and how the charateristic (or the general listeners taste in music) evolved over time.

2. Measuring this trend over broad time-frames might indicate some of periodic or fickle or any patterns that exists in the listening habits over the years.

3. Comparing adjacent time-frames might give us insight as to major tracks/albums which caused a disruption in the listening choices of people.

4. The rise and fall of genres (subject to further availability of the 'genre' data)

The general type of questions that will be analysed using the data would be a classification problem. Such a classifier might be a good choice to answer some of the questions mentioned above and a few more suject ot the availability of extended data.

Based on just this data, the classifier should be able to predict the time-frame (to be denoted as 'era' hencefoth) of the songs in the test set.

3 Personal Curiosity

A personal curiosity but perhaps not accurately answerable question (hence not promised as a deliverable) might be to prove if the top songs of today would appear in the charts of a previous era or vice versa. Another interesting challenge would be to correlate people's listening choices with the politico-socio-economic condition of the period in which the song was released and how it affected the nature of the songs itself.

4 Future Task

4.1 Preprocessing of the database

Coming Soon ...