MONU-MEL-DATA-PT-11-2020-U-C

Project Proposal – Project 1

Dancing Through The Decades

Objectively measuring music for the last hundred years using Spotify

Presentation date: 21-JAN-2021

**Group**:

Brett Scotland

Darren McMurtie

Jesse Tan

Kailyn

Will Bobzin

Premise:

Spotify offers a powerful database, containing data for over 160K songs. The database includes some ad-hoc metrics, such as *loudness*, *danceability* and *key* . As a group, we identified “Music” as one of our common interests. When exploring what kind of data project can be developed around that topic, the opportunity to use the Spotify database came up.

After a preliminary exploration, we were impressed by the thoroughness of the database, which has every single song categorized by year of release. This sparked an overarching question that we are keen to answer: **How has music changed over the years?**

Keeping in mind the content of the database and the limited time available, we segmented the main premise into four areas of analysis:

* Which decade has the most *danceable*music? – Comparison between average danceability for each decade, charting how it has evolved. Hypothesis: based on reputation and anecdotal evidence we believe the 70s will beat every other decade in *danceability*. ON HOLD
* Has the duration of songs changed over the years? From the long jazz jams of the 50s to the long power-ballads of the 80s and the short radio-edits of the 90s, we’ll analyze the data to identify the trends in song duration. Hypothesis: COMPLETE
* What is the most popular Key in each decade? And of all times? Something that made us geniuniely curious is wether there is a given key that makes songs popular or if a hit can be written in any key. Furthermore, have the most popular keys changed over time? Hypothesis: COMPLETE
* We are all familiar with the songs that top each decades hit lists. Now, what are the actual most popular songs of the last hundred years? We will look at the popularity information from the database and find out what are the most popular songs of each decades and all times. Hypothesis: COMPLETE

Feedback: Mash with another dataset – Triple J? Can we combine datasets to make more relevant analysis?

**Ideas:**

* What makes a Triple J winner? Mashing data between Spotify and Triple J databases for 1993 to 2017. Get technical/numerical information for each song based on Spotify analytics. Compare some key elemtens in all triple j winners to determine commonalities – For example: Liveness, loudness, danceability.
* What is the perfect time to release a hit? Using the release date informed by spotify, we can find what are the most common release dates for the winners? Hypothesis: perfect launch date is March to June, as it takes a few months for people to get to know a song and start liking it enough to vote for it.
* Australia songs – What kind of Australian songs do people vote for in the Top 100? We will make a subset of Australian artists and measure their speechiness and Liveness and compare those aggregate values with the rest of the song universe in the historical top 100 list. Hypothesis: Top 100 voters will prefer Aussie songs with higher speechiness and liveness as they will choose songs they’ve enjoyed live and that showcase similar accents to theirs.

**Limitations and challenges:**

Mashing is relateively complex as it has to be performed by pivoting *artist* and *title* fields, which have differences in punctuation and usage of special characters. This is particularly challenging in terms of spelling of foreign names (eg. Björk) and multiple artists.

Currently between 2 and 4 million votes are cast for the Australia date top 100. As significant as that number is, it pales in comparison to the 138 million active Spotify subscribers. In other words, the Triple J dataset is heavely skewed towards young Australians. Whil we can use the available data to extract trends about Australian music tastes, none of them could be extrapolated to the Spotify listener base.

Technical data for each song comes from the Spotify algorithmic analysis. While it’s reliable and more importantly, consistent, we have no access to its inner workings and therefore can’t corroborate their accuracy.

**Sources**:

Categories explained - <https://developer.spotify.com/documentation/web-api/reference-beta/#category-tracks>

Database, extracted by a KAggel user from the Spotify API: <https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data.csv>