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

Project Proposal – Project 1

Dancing Through The Decades

Objectively understanding Australian taste in music

Presentation date: 21-JAN-2021

**Group**:

Brett Scotland

Darren McMurtie

Jesse Tan

Kailyn Yong

Will Bobzin

**Premise**:

~~When our group was formed, we immediately started bouncing different topics that were of interest. Soon after we discovered that~~ we are all passionate about Music~~”, which comes hardly as a surprise since it’s not only multi-billion dollar industry but a genuine passion for most people~~. Far from the days of vinyl albums and cassettes, streaming is king in 2021. Last year, 286 million people were active users of Spotify. 130 million of those enjoyed it enough to pay for a premium service, removing advertisements and other limitations.

Being a streaming platform, Spotify has access to a wealth of information not only of user’s listening habits but also on the music itself. Through algorithms, Spotify measures, classifies and segments music to ~~try and~~ understand and predict the taste of each user. To this purpose, it utilizes categories such as *tempo, danceability, liveness* or *speechiness.*

~~Luckily, some of that information is accessible through the Spotify API.~~ We were particularly interested in using Spotify’s musical analysis to ~~try and~~ understand what ~~do~~ popular songs have in common.

To measure popularity, we looked at Australia’s renowned music listener poll: Triple Js Hottest 100. Australian radio station Triple J holds a yearly contest, where listeners vote for their favourite songs of the year.

Though emphasis is placed on Australian and alternative music ~~from the world~~, it serves as a useful reference point as yearly votes are counted in the millions. We believe that concentrating on Triple J Hottest 100 serves a double purpose by adding a “local twist” to this analysis, while also limiting the scope to a manageable level.

**The Data**:

We have ~~in our hands~~ a data file extracted from Spotify, with information of more than 160.000 songs released between 1921 and 2020. The data includes the usual music information (Artist, song, duration, date of release, etc.) as well as the technical analysis described above.

We also have a smaller dataset containing the Triple J Hottest 100 ranking for each year between 1993 and 2017. This dataset includes artist, song title and country of origin of the artist, among other categories.

**The Questions**:

* 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 elements in all triple j winners to determine commonalities – For example: Liveness – Re-read the doco, liveness just means that there is a live audience. Perhaps replace by energy (intensity) or valence (how “happy” the track is? – Will , loudness, danceability. Hypothesis: explore first. My guess would be high speechiness, mid valence, low danceability. Happy to discuss. - - Will I agree here, you could do one separate about Liveness, but I think we have enough here.

Educated guess: Winners of top 100 will show low danceability, high liveness.

Compare mean danceability (and other relevant variables (liveness, danceability etc.)) in Triple J winners to population mean for these variables using a single sample t-test to determine which, if any, variables are relevant in the creation of a Triple J winner.

* 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. How to implement: retrieve month for each winner, chi square test on that data to confirm. By grouping months into yearly quarters and using a chi squared test after establishing the frequency at which Triple J winners’ release date falls into each quarter the time of year where winners are more likely to arise might be ascertained if there is such a time. A comparison to the population song release distribution will also indicate that if there is a statistically significant result that it is not due to a simple increase in the volume of songs released at that time of year. Yes think this one is straight forward
* Australian 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 – maybe remove? Sounded like 1 variable was enough - Will as they will choose songs they’ve enjoyed live and that showcase similar accents to theirs. Compare mean values between group and universe, implement a test to see if it’s statically significant. The use of single sample t tests to compare the mean of speechiness and liveness to their respective population means will determine if Triple J voters are drawn to these characteristics in Australian songs.
* Do Aussies like long songs? We will compare average duration for top 100 songs for each year, against the average duration of all songs released that year. This will tell us if Aussies prefer relatively short songs. Hypothesis: COMPLETE A year by year comparison of the average duration of the songs preferred by Aussies (i.e. the Triple J hottest 100 songs) to the population mean of the average duration of all songs released in that year according to Spotify will determine if Aussies like long songs. This analysis will require multiple single sample t-tests to perform.

**The Challenges**:

* Mashing is relatively 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 heavily skewed towards the tastes of young Australians and favors Australian artists and alternative music. While we can use the available data to extract trends about Australian music tastes, no valid conclusions 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.

I know we cannot get them all to match, but Ryan said this can be mentioned as one of the challenges and he would not hold it against us.

**The 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>
* Triple J Top 100 Database - <https://github.com/majames/hottest100/>

**The Tools:**

* Jupyter Notebooks
* MatplotLib
* Python
* ?