Spotify Data Analysis - Genre Categorization for Jamaican Music

Data Scientist – Sharlmagne Henry

General objective – Use Spotify's data to determine if there are any quantifiable differences among Jamaican music and determine the subgenres/subcategories using unsupervised learning.

Methodology

Data Collection – Spotify API and python.

Exploratory data analysis (EDA) & Data wrangling – initial data analysis in tandem with data wrangling using Pandas and Seaborn libraries

Modelling – K-Means clustering to classify data based on their similarities in metrics

Statistical data analysis – Detailed analysis to determine the characteristics of each cluster and make assumptions based on the characteristics defined in the descriptive analysis.

Feature analysis (Features excluded)

Release Date

One general hypothesis was that the release date may have a correlation to the subcategories of Jamaican music, but unfortunately, the release date information for each track was not reliable enough to utilize in the study. When analysing the actually release dates of many tracks, it was concluded that release dates on Spotify was not synonymous to the actual release date of the track. Songs released in the 70s, 80s, and 90s were showing inaccurate release dates conflated with upload dates on Spotify.

Tempo

The tempo of the songs was not reliable enough to make general statements. Though tempo may have been a good feature to define the subcategories. It is impossible to know if songs were in double time or not, and if the tempo stated was half or double the number stated. Looking at some songs that were at 120bpm, using field knowledge, I could indicate whether or not that should be 60bpm or 120bpm. However, it was not possible to determine that without prior classification. So, tempo would have unreliably affected the clusters.

Key

The key of the song generated an unfavourable outcome for categorization. The clustering algorithm was prioritizing the key of the song regardless of the other features. Fortunately, it was sacrificial enough to exclude. When re-analysing the distribution of keys in each cluster, it was even among each.

Time Signature

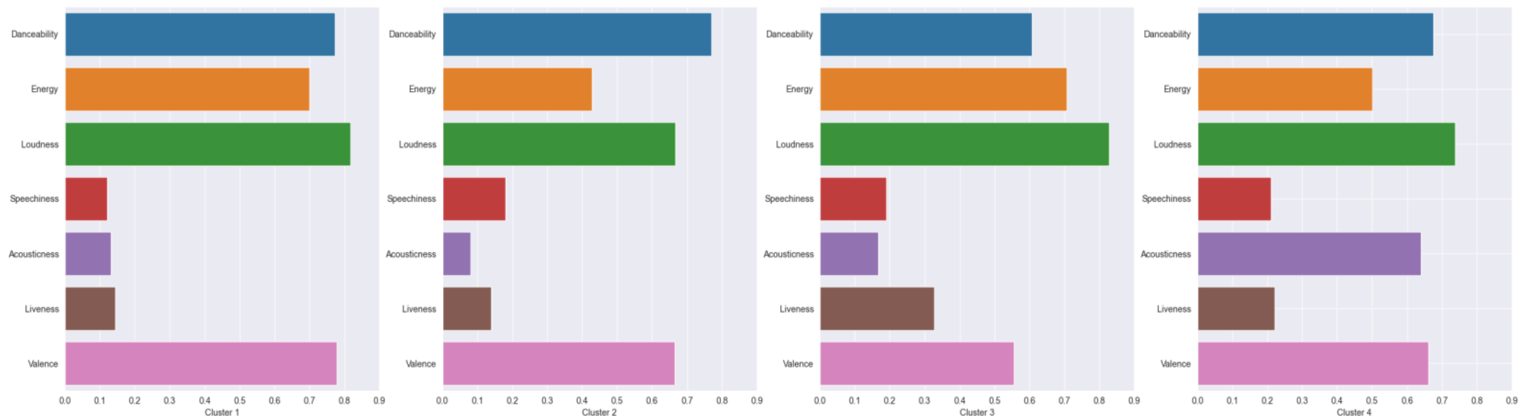
This was unreliable data due to the inaccuracy of the data itself. There were a wide variety of false readings of time signatures of 3s and 5s, when the entirety of Jamaican music can arguably be assumed to be 4/4 or 4 in time signature. This may be a bold claim to make against a computer algorithm, but by manually going through the time signatures of many 3s, and 5s, I personally confirmed it to be an inaccurate reading.

Duration

Initial using duration resulted in a small subcategory of 40 sample, but after analysing the means of this subcategory, many of its features were similar to other subcategories. Therefore, the duration was negatively being prioritized and affecting the clustering algorithm.

Results

Comparison of features between each cluster



Cluster 1 Insights - High Danceability, Loudness & Valence - Moderate Energy - Low in Speechiness, Acousticness and Liveness	**Cluster 2 Insights** - High in Danceability & Valence - Moderate Energy & Loudness - Low in Speechiness, Aousticness and Liveness	**Cluster 3 Insights** - High in Energy Loudness - Moderate Danceability & Valence - Low in Speechiness, Acousticness & Liveness	**Cluster 4 Insights** - High in Loudness - Moderate in Danceability, Energy, Acousticness and Valance - Low in Speechiness and Liveness
ANOVA & Tukey's HSD - 2nd in Danceability (very close/similar to cluster 2) - 2nd in Energy (very close/similar to cluster 3) - 2nd in Loudness (very close/similar to cluster 3) - 4th in Speechiness - 3rd in Acousticness (very close/similar to cluster 2 & 3) - 3rd in Liveness (very close/similar to cluster 2 & 4) - 1st in Valance	**ANOVA & Tukey's HSD** - 1st in Danceability (very close/similar to cluster 1) - 4th in Energy (very close/similar to cluster 4) - 4th in Loudness - 2nd in Speechiness (very close/similar to cluster 3 & 4) - 4th in Acousticness - 4th in Liveness (very close/similar to cluster 1 & 4*) - 3rd in Valance (very close/similar to cluster 4) *Tukey's Only **Plot Only	**ANOVA & Tukey's HSD** - 4th in Danceability - 1st in Energy (very close/similar to cluster 1) ** - 1st in Loudness - 2nd in Speechiness (very close/similar to cluster 2 & 4) - 2nd in Acousticness (very close/similar to cluster 1) - 1st in Liveness - 4th in Valance (very close/similar to cluster 4) * *Tukey's Only **Plot Only	**ANOVA & Tukey's HSD** - 4th in Danceability - 3rd in Energy (very close/similar to cluster 2) * - 2nd in Loudness - 1st in Speechiness (very close/similar to cluster 3 & 4) - 1st in Acousticness - 2nd in Liveness (very close/similar to cluster 1 & 2) * - 1st in Valance (very close/similar to cluster 2 & 3*) *Tukey's Only **Plot Only

As expected, the clusters had more similarities than differences among the different features. Only Cluster 4 had a unique distinction in Acousticness compared to the other clusters.

Cluster sizes are as follows:

Cluster 1 -13656

Cluster 2 -9460

Cluster 3 - 9234

Cluster 4 – 4265

The next step is to find a way to find songs that would fall within on of these clusters and use those songs to conclude the main objective. This would require data from supervised learning with labelled subgenres to cross examine with the clusters. Unfortunately, there is no such data readily available, so the next best option is to use Spotify's API to make recommendations based on a sample from each cluster.

Analysing the differences between the clusters and recommendations, two things need to be taken into consideration.

- The method Spotify used by the API to make recommendations

- Are track features sufficient to make judgement on the genre of music.

Spotify Recommendation

The request made through the Spotify API for song recommendation, does take into consideration the features of the array of tracks submitted. According to the Spotify documentation, *Recommendations are generated based on the available information for a given seed entity (tracks submitted for recommendation) and matched against similar artists and tracks. If there is sufficient information about the provided seeds, a list of tracks will be returned together with pool size details.* The process it goes through to do the aggregation of the array of features is uncertain, but most if not all of the featured are analysed and cross referenced to generate recommendations. Unfortunately, after using my field knowledge about the characteristics of Jamaican music, I can conclude that the recommendations made by Spotify are not uniform enough to make a conclusive judgment on the subcategories of Jamaican music. The songs recommended have a vast mixture of styles that qualify for all the subsets of Jamaican music so no distinction can be made just by using the recommendation.

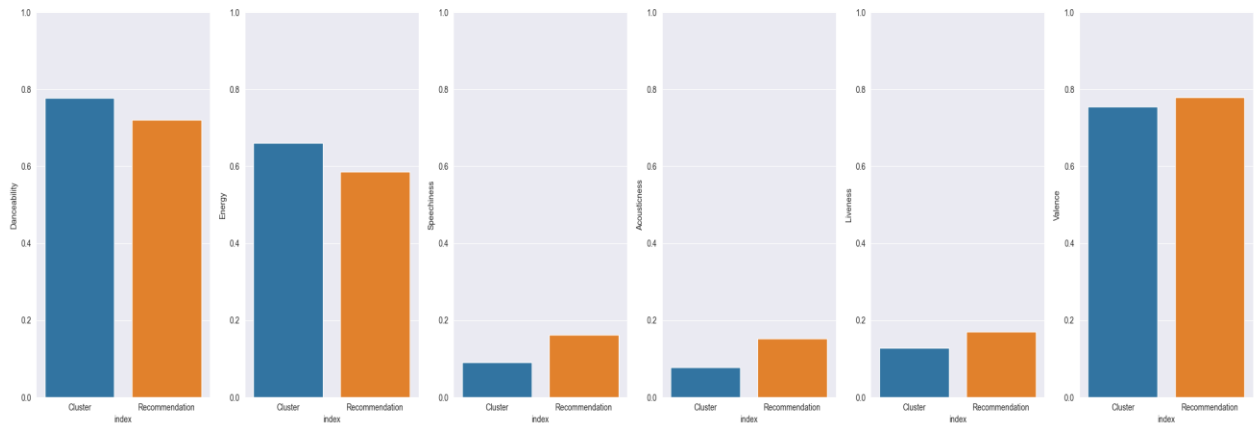
So, are the track features sufficient to determine the subcategories among Jamaican music? Let's take a look at the results. K-means clustering was used to facilitate the unsupervised learning and 4 categories were identified. Visual analysis combined with ANOVA testing and Tukey's HSD were done on each cluster to determine the differences between and among each cluster. Initially, 6 clusters were hypothesized, but there were more overlaps in similarities between the clusters under these circumstances so, the cluster count was optimized to 4. A sample of each cluster was taken and the Spotify API was used to generate recommendations based on each cluster. These recommendations were too diverse to make any conclusive judgement about the subcategories by ear, so two sample t-tests were done to compare them to their respective clusters.

Cluster 1

Two sample T-test revealed dissimilarities in 2 features (Danceability & Energy), but visual analysis confirms 3 dissimilarities (Danceability, Energy, Speechiness & Acousticness).

```
P-value = 0.033226900631662888 - dissimilar Danceability
P-value = 0.001840628199163803 - dissimilar Energy
P-value = 0.14980450652590308 - similar Speechiness
P-value = 0.5914690185065915 - similar Acousticness
P-value = 0.4715544588487689 - similar Liveness
P-value = 0.9889232664955496 - similar Valence
```

Cluster 1 Features vs Spotify Recommendations

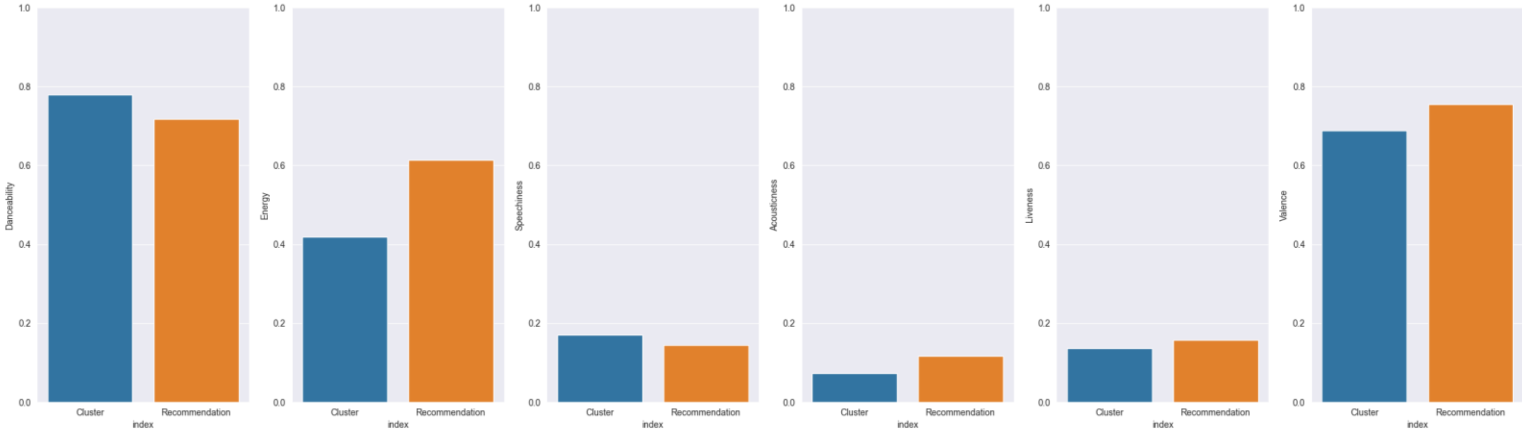


Cluster 2

Two sample T-test revealed dissimilarities in 2 features (Energy & Varience), and visual analysis confirms the same, with a significant dissimilarity in Energy.

```
P-value = 0.06528283143074322 - similar Danceability
P-value = 8.585010312863322e-05 - dissimilar Energy
P-value = 0.23406457167292596 - similar Speechiness
P-value = 0.17348575088594427 - similar Acousticness
P-value = 0.6477890793056936 - similar Liveness
P-value = 0.02988797459004948 - dissimilar Valence
```

Cluster 2 Features vs Spotify Recommendations

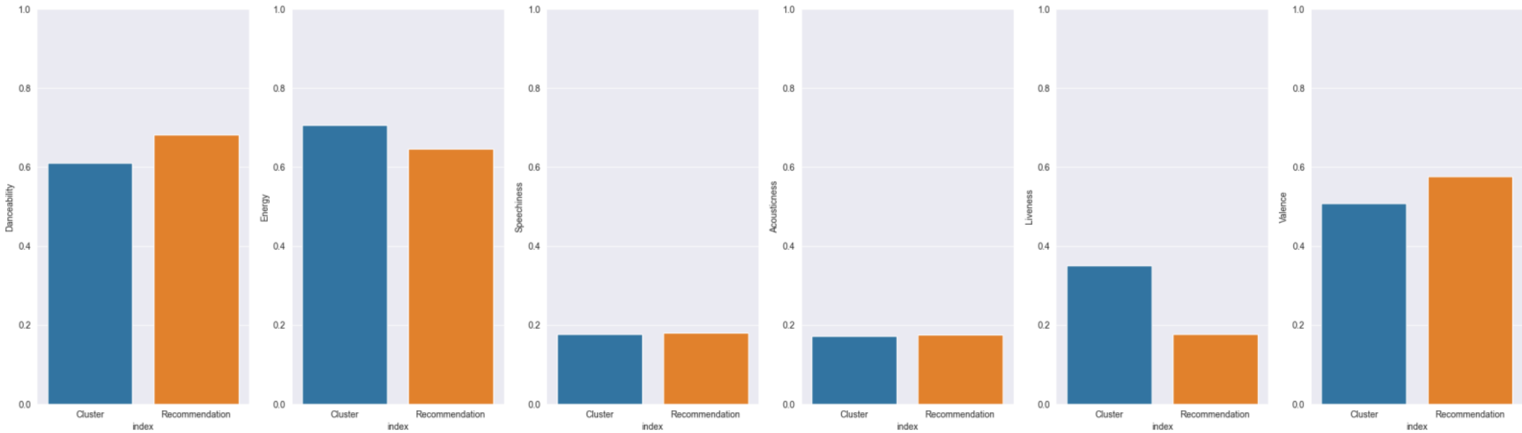


Cluster 3

- Two sample T-test and visual analysis revealed dissimilarities in only 2 features (Danceeability & Liveness), and visual dissimilarities in (Danceability, Energy, Liveness & Valence).

```
P-value = 0.028475241653584598 - dissimilar Danceability
P-value = 0.10538057159705114 - similar Energy
P-value = 0.7349950113269162 - similar Speechiness
P-value = 0.8398651932216186 - similar Acousticness
P-value = 0.001141755360152927 - dissimilar Liveness
P-value = 0.7154523595159802 - similar Valence
```

Cluster 3 Features vs Spotify Recommendations

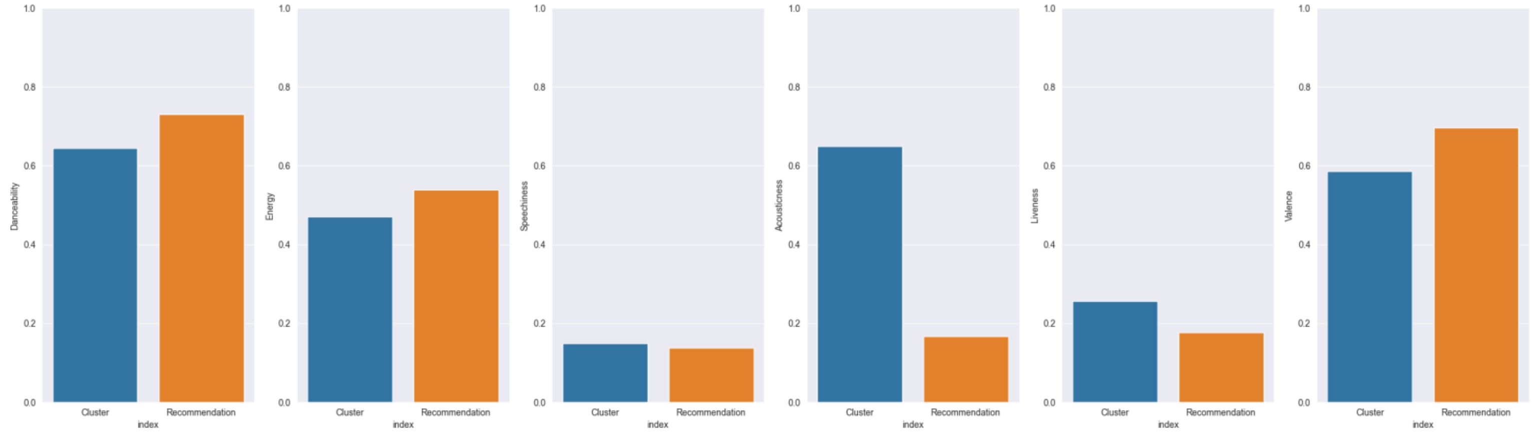


Cluster 4

Two sample T-test revealed dissimilarities in 3 features (Speechiness & Acousticness), but visual analysis confirms dissimilarities in (Danceability, Acousticness, Liveness & Valence) with massive dissimilarity in Acousticness.

```
P-value = 0.08328910409740127 - similar Danceability
P-value = 0.3678224084092947 - similar Energy
P-value = 0.0068567472506601895 - dissimilar Speechiness
P-value = 3.2460601063791004e-10 - dissimilar Acousticness
P-value = 0.27858341313317586 - similar Liveness
P-value = 0.4867328404085496 - similar Valence
```

Cluster 4 Features vs Spotify Recommendations



Going trough all the recommendation compared to the respective cluster there weren't any perfect matches. Out of all the 6 features analysed, on average 4 features were similar to its respective cluster, while 2 features were dissimilar. This represents 66% confidence that the recommendations are similar to the cluster.

Conclusion

The objective of this exercise was to determine if there are any quantifiable differences among Jamaican music and determine the subgenres/subcategories using unsupervised learning. Ultimately, the data used from Spotify to make inferences on the categorization of Jamaican music, was not sufficient to make any conclusive judgement. Categories were assumed, but no generalization could be made about its subjective characteristics. What would improve this exercise? If there were more reliable data on release date of tracks, one could assume that it would improve the clustering of data. In addition, if there was a labelled dataset available to cross reference with the clusters, it would be ideal to confirm if the categories generated are useful or accurate.