MTH443: PROJECT REPORT

Music genre mapping and prediction

Supervisor:

Prof. Amit Mitra

Group members:

Dwija Kakkad (210367)
Praneat Data (210740)
Nihar Janbandhu (210665)
Sai Praneeth Donthu (210900)
Vrinda Sharma (211184)

Contents

Table 1: Type of the features

Feature	Type
Valence	Continuous
Year	Categorical
Acousticness	Continuous
Artist	Categorical
Danceability	Continuous
Duration	Continuous
Energy	Continuous
Explicit	Categorical (binary)
ID	Categorical
Instrumentalness	Continuous
Key	Categorical
Liveliness	Continuous
Loudness	Continuous
Popularity	Categorical
Mode	Categorical
Speechiness	Continuous
Tempo	Continuous

1 Introduction

1.1 Aim

The datasets we used under consideration consists of 85,000 songs produced from 1920 - 2000. The feature vectors include 19 parameters: valence, year, acousticness, artist, danceability, duration, energy, explicit, ID, instrumentalness, key, liveliness, loudness, mode, name, popularity, release date, speechiness and tempo for each song. The type of each feature is given as:

Out of these, we denote the following features as sound features: acousticness', 'danceability', 'energy', 'instrumentalness', 'liveness', and 'valence'. The dataset 'databygenres' gave the genres and their corresponding mean values of a subset of the above features, and the main dataset contained features without the genre. Such features motivated us to perform a genre classification of the songs in the main dataset and a subsequent feature analysis for each genre i.e. songs classified as alberta hip hop tend to also score high on acousticness.

1.2 Exploratory Data Analysis

We plot how the sound features changed in time to get a better idea of the shift of genres over time.

Over the years, most produced songs have shown an increment in their scores of energy, danceability and decrement in liveness and acousticness. Considering the entire data might lead to inaccurate feature prediction in genres because of fluctuation in patterns during 1970s. Further, this will increase model training time due to a very large dataset.
Thus, we considered dataset from 2000-2020 because of uniformity in the features of produced songs and optimal number of data points for ease in model training and accurate predictions. We see that the data from 2020-2021 has sufficient data points to ensure model accuracy.

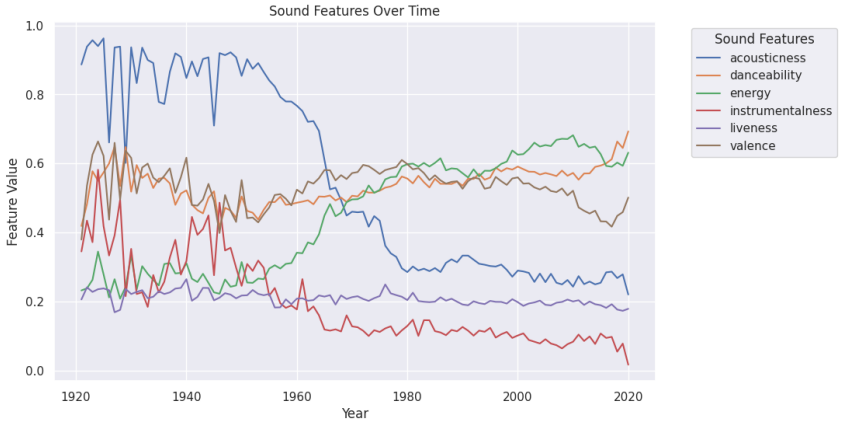


Figure 1: Shift in Trends around 1970s

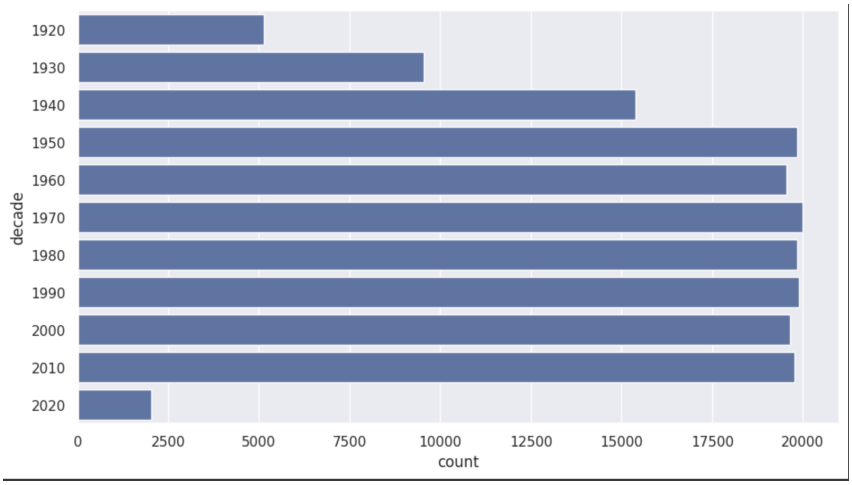


Figure 2: Shift in Trends around 1970s

1.3 Genre Selection

From the genres dataset, top 10 genres were selected based on their scores in popularity and higher frequency. Some sound features were also plotted for each of these genres, to get an estimate of the differentiating qualities and similarities amongst each of the genres. Initially, it seems that energy and danceability are highly correlated and we could safely drop one of the two. To get a better idea of which features to retain, we used PCA.

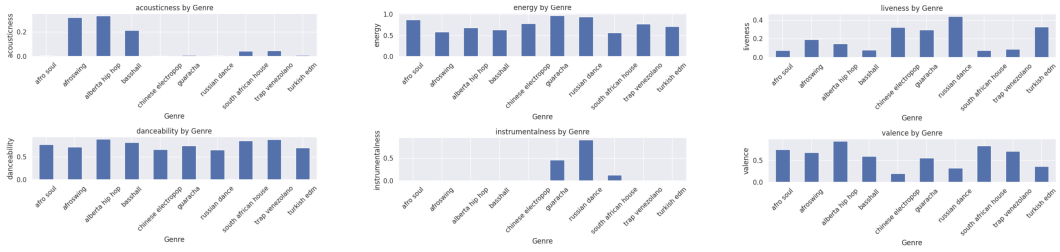
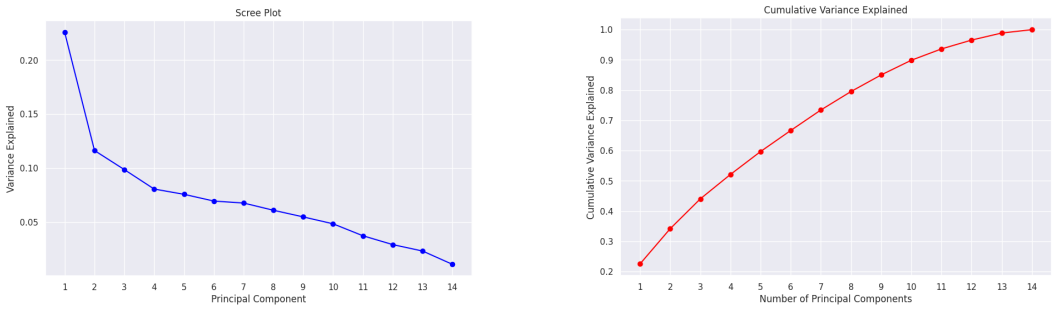


Figure 3: Sound features for genres

2 Genre Classification

2.1 Principal Component Analysis: Feature Reduction

In an attempt to reduce features and retain only the most relevant original sound features, we performed PCA on the main dataset with 14 sound feature columns. Because the range in all the values was high, we first normalized the data before performing PCA.



(a) Scree Plot

(b) Cumulative Variance Explained

Figure 4: Comparison of Scree Plot and Cumulative Variance Explained

We do not see a defined elbow point for the Scree plot, and even for the cumulative variance plot is smoothly increasing. We conclude that the features are all important, and we retain all of them.

2.2 Clustering

We use kmeans clustering on the main dataset (without genres) to cluster based on the sound features. We will then later map the cluster centers to the closest genre in the genres dataset. To find the optimal number of clusters, we calculated the within cluster distance for number of clusters ranging from 0 to 20. The elbow plot is given below:

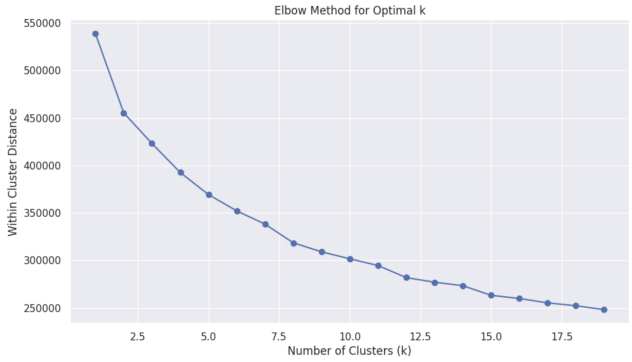


Figure 5: Finding the optimal k

We observe that no particular elbow forms, which is peculiar but not unexpected because the number of genres is very large. To avoid overfitting, and because we are considering only the top 10 genres we take $k = 10$. After running k-means, we then map the cluster centers of the 10 clusters formed to the closest genre in the genre dataset.

2.3 Feature Analysis

To see how well our clustering has worked, we take a particular genre, and consider the artist who has a large number of songs belonging to that genre (for more data points). We then estimate the kernel density for each sound feature for the songs of that artist and compare the plots fo these with the kernel density plots of the features for that genre (obtained from the genre dataset). To demonstrate, we have taken the 'Seattle hip-hop' genre and considered artists 'Eminem' who is the most frequent artist in that genre and 'Frank Sinatra' who has a less number of songs from our chosen genre. The code is flexible and the artist can be changed to check the KDE plots of other artists and compare their similarities to other genres.

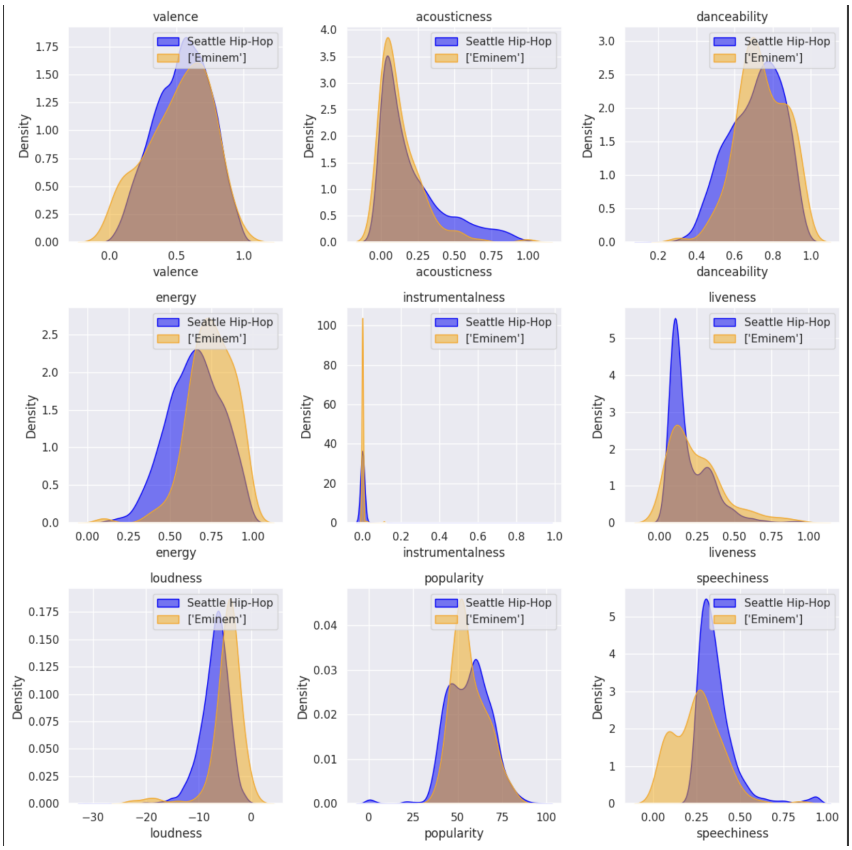


Figure 6: KDE comparison: Artist within the genre

As expected, the KDE plots of the features of Eminem's songs are quite similar to the genre and for Frank Sinatra, the plots differ quite a lot in some features. By this visual inspection, we can assume that our clustering has been satisfactory.

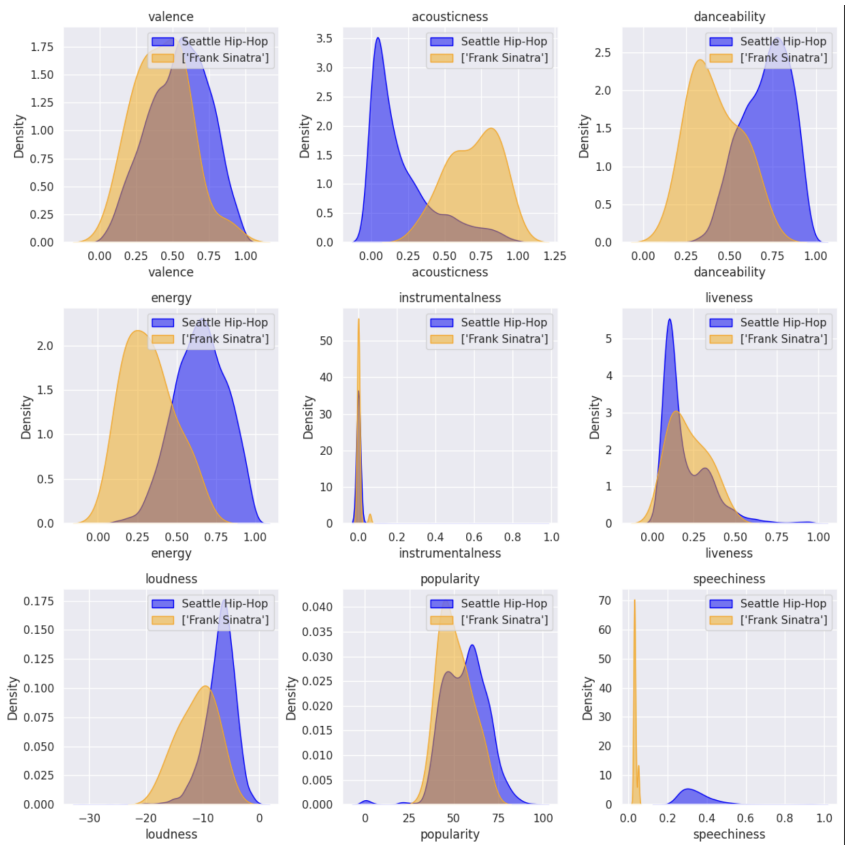


Figure 7: KDE comparison: Artist outside the genre

2.4 Genre Analysis

To analyze if the genres are indeed accurate ways to clusters songs across the year, we analyzed song feature variations for each genre across the selected timespan i.e. 2000-2020. For most genres the mean values of the song features do not fluctuate significantly. Hence, the selected genres will provide an accurate clustering criterion for the songs.

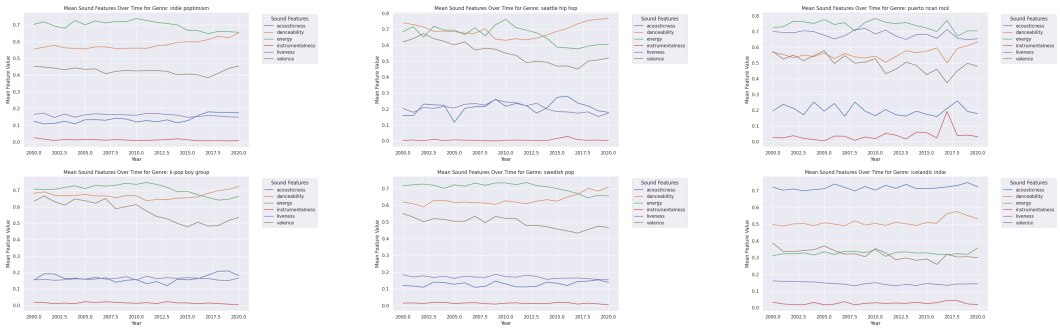


Figure 8: Sound features for genres

3 Discussion

Now that we have clusters ready, we can predict the genre of any new song by assigning it to the nearest cluster center. If we have new user data, we can recommend songs by picking songs from the most frequent genres of the users.