CLUSTERING AND FITTING

Assignment 2

Name: Shiva Sai Mallavarapu

Student ID: 23069222

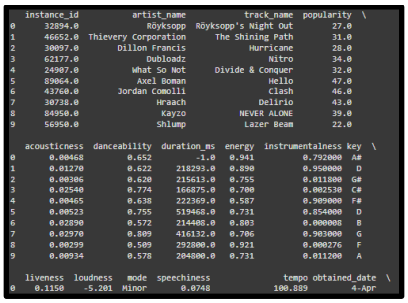
Dataset Link: ***https://www.kaggle.com/datasets/vicsuperman/prediction-of-music-genre***

GitHub Repo: **https://github.com/Shiv-web-lab/CLUSTERING-AND-FITTING-.git**

1. Introduction

The report centres on using a second dataset from a site like Kaggle to perform clustering and fitting methods. A Music Genre prediction dataset (***https://www.kaggle.com/datasets/vicsuperman/prediction-of-music-genre***) has been taken from Kaggle to perform this data analysis technology with data pre-processing, data visualisation and machine learning model fitting strategies. The analysis consists of k-means clustering and linear regression analysis. The report contains different graphical plots in words and some graphical representations using packages of Python.

2. Results



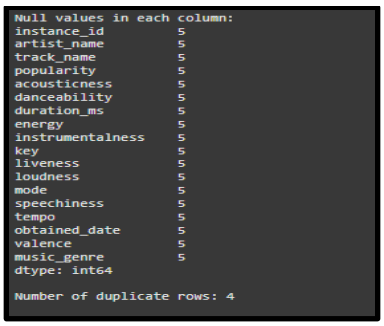
	instance_id	artist_name	track_name	popularity	\
0	32894.0	Röyksopp	Röyksopp's Night Out	27.0	
1	46652.0	Thievery Corporation	The Shining Path	31.0	
2	30097.0	Dillon Francis	Hurricane	28.0	
3	62177.0	Dubloadz	Nitro	34.0	
4	24907.0	What So Not	Divide & Conquer	32.0	
5	89064.0	Axel Boman	Hello	47.0	
6	43760.0	Jordan Comolli	Clash	46.0	
7	30738.0	Hraach	Delirio	43.0	
8	84950.0	Kayzo	NEVER ALONE	39.0	
9	56950.0	Shlump	Lazer Beam	22.0	

	acousticness	danceability	duration_ms	energy	instrumentalness	key	\
0	0.00468	0.652	-1.0	0.941	0.792000	A#	
1	0.01270	0.622	218293.0	0.890	0.950000	D	
2	0.00306	0.620	215613.0	0.755	0.011800	G#	
3	0.02540	0.774	166875.0	0.700	0.002530	C#	
4	0.00465	0.638	222369.0	0.587	0.909000	F#	
5	0.00523	0.755	519468.0	0.731	0.854000	D	
6	0.02890	0.572	214408.0	0.803	0.000068	B	
7	0.02970	0.809	416132.0	0.706	0.903000	G	
8	0.00299	0.509	292800.0	0.921	0.000276	F	
9	0.00934	0.578	204800.0	0.731	0.011200	A	

	liveness	loudness	mode	speechiness	tempo	obtained_date	\
0	0.1150	-5.201	Minor	0.0748	100.889	4-Apr	

Figure 1: Importing dataset

Through data loading, different aspects of the dataset's characteristics become clear including the nature of the columns and rows, missing, and duplicated entries essential for cleaning the data for data visualization, clustering, and regression modelling.



```
Null values in each column:
instance_id          5
artist_name          5
track_name           5
popularity           5
acousticness         5
danceability         5
duration_ms          5
energy               5
instrumentalness     5
key                  5
liveness             5
loudness             5
mode                 5
speechiness          5
tempo                5
obtained_date        5
valence              5
music_genre          5
dtype: int64

Number of duplicate rows: 4
```

Figure 2: Detecting null rows

Identification of 5 null values in each column shows that data cleaning steps need to be made to address the incompleteness of a dataset. Deleting rows with void rows has been conducted by using the 'dropna()' function.

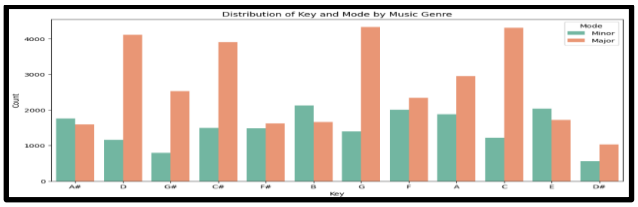


Figure 3: Bar diagram

In order to partly compare and contrast its findings, the bar diagram of the distribution of key and mode in the music showcases patterns in composition by genre and identifies frequently used key-mode pairs. This insight helps in analysing the genre attributes that would provide assistance in features of music, recommending systems, and contents based on machine learning analysis of musical features.

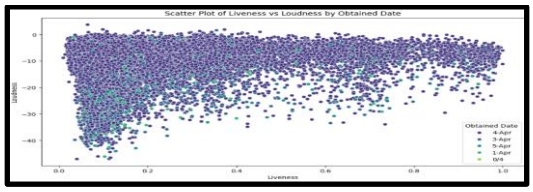


Figure 4: Scatter chart

The scatter chart displayed below shows a relationship between liveness and loudness where clusters as well as outliers can be identified. It shows trends in the songs' dynamics to help in sorting genres and creating the right order for songs in a playlist.

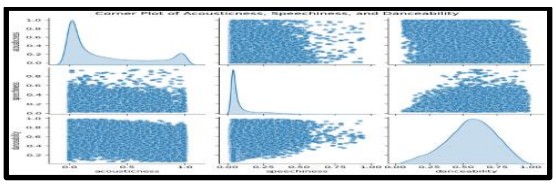


Figure 5: Corner or pair plot

The corner chart illustrates the correlations and tendencies between 'acousticness', 'speechiness' and 'danceability' to music. It assists in recognizing how these characteristics are related and provides information on the features of a song as well as its genre.

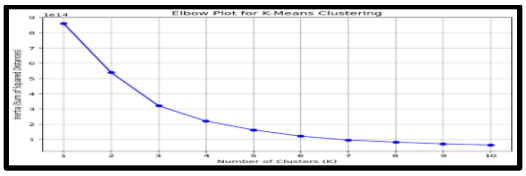


Figure 6: Elbow plot for k-means clustering

The result obtained from using K-means clustering demonstrates how the music data can effectively be classified using features to achieve a set of clusters. By applying K-means the model groups the data points and finds the clusters into which it is possible to subgroup the music tracks based on certain similarities that may refer to different genres or styles.

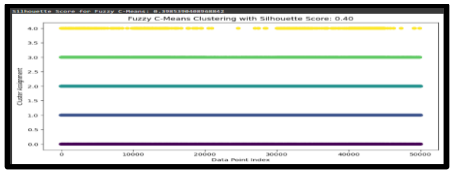


Figure 7: Fuzzy C-means clustering

The results obtained from employing the Fuzzy C-means clustering show more flexible Music data grouping as many of the data points are allowed to be a member of more than one cluster and with different levels of membership.

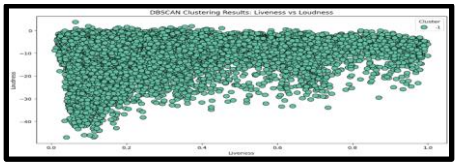


Figure 8: DBSCAN clustering algorithm

It demonstrates the identification of core samples, and noise points as well as the formation of the clusters depending on the density. DBSCAN helps to discover characteristics of the data including outliers and separate clusters within the music dataset.

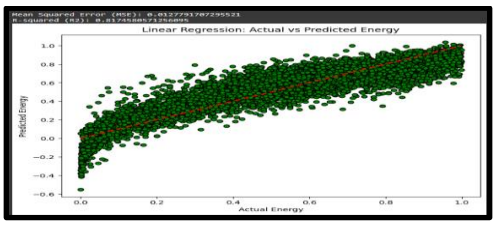


Figure 9: Linear regression model

The results of line fitting using linear regression evaluate efficiently with mean squared error or, MSE ‘0.0128’ and R-squared or R2 ‘0.8175’, and hence the model is considered a good one for prediction. The low MSE shows that there is very little difference between the predicted and the actual energy values and the high R2 means that the model accounts for ‘81.75%’ of the variance existing in the current data set.

3. Conclusion

In the report, when using the data set, the clustering and fitting methods were shown to apply to the music genre prediction. Data pre-processing and visualisation of the dataset for improving feature selection have been perfectly covered. Comparing the results obtained with K-means, Fuzzy C-means, and DBSCAN the quantitative analysis showed the presence of different clusters and outliers. Additional strengths of graphical outcomes were enhancing comprehension of the musical features.