(1) 請列出每個 Audio Feature 的值域及其意義，同時觀察是否有 missing value 或 noise.

Audio Feature	Description
Danceability	Meaning: The relative measurement of the track be danceable. Range: [0,1] Missing value: No Noise: No
Energy	Meaning: The energy value of the track. Higher values mean that the song is more energetic. Range: [0,1] Missing value: No Noise: No
Key	Meaning: All keys on octave encoded as values with starting on C as 0, C# as 1, etc. Range: [0,1, 2, …, 11] Missing value: No Noise: No
Loudness	Meaning: The overall loudness of a track in decibels (dB). Range: [-33.357, 3.148] Missing value: No Noise: No
Mode	Meaning: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Range: [0,1], where Major is represented by 1 and minor is 0. Missing value: No Noise: No
Speechiness	Meaning: The relative length of the track containing any kind of human voice. Range: [0,1] Missing value: No Noise: No

Acousticness	Meaning: The value that describes how acoustic a song is. Higher values mean that the song is most likely to be an acoustic one. Range: [0,1] Missing value: No Noise: No
Instrumentalness	Meaning: The relative ratio of the track being instrumental. Higher values mean that the song contains more instrumental sounds. Range: [0,1] Missing value: No Noise: No
Liveness	Meaning: Detects the presence of an audience in the recording. Range: [0,1] Missing value: No Noise: No
Valence	Meaning: The positiveness of the track. Higher values mean, the track evokes positive emotions (like joy) otherwise means, it evokes negative emotions (like anger, fear). Range: [0,1] Missing value: No Noise: No
Tempo	Meaning: The tempo of the track in Beat Per Minute (BPM) Range: [57.967,220.29] Missing value: No Noise: No
Time Signature	Meaning: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). Range: [1,5] Missing value: No Noise: No

(2) 如何做分群前的資料前處理(Preprocessing, 包括 Data Clean, Feature Normalization) ?

Firstly, we find that some “features” like type, id, uri, track_href, analysis_url, duration_ms, song_name, Unnamed: 0, title are no important in my experiment, so we use obj.drop to delete these columns.

Secondly, I use obj.isnull().sum() to check whether the data have missing value.

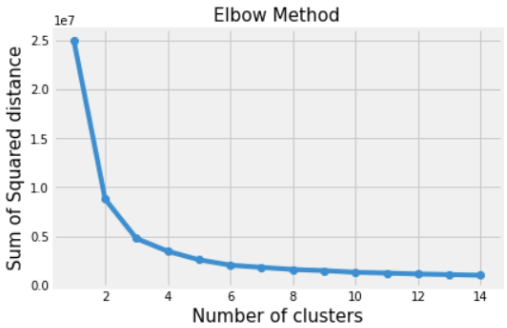
Thirdly, I use LabelEncoder to transform “genre” to digit.

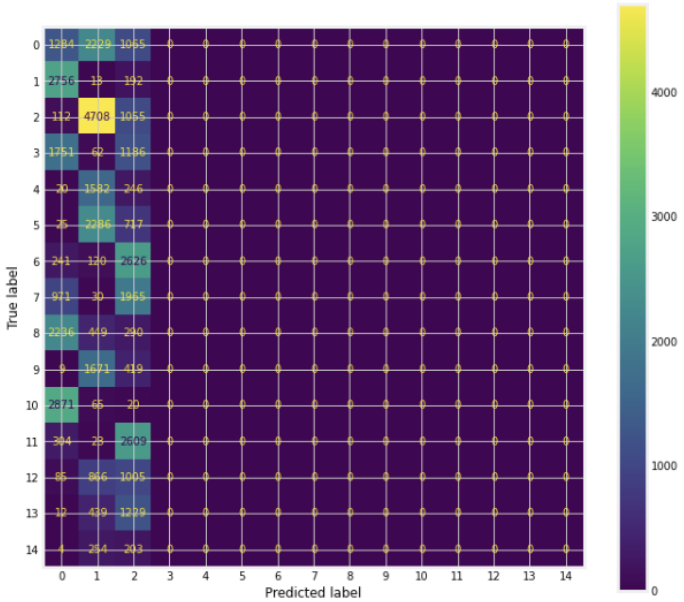
Finally, I make feature normalization with standardScaler().

(3) 請執行 K-means，並列出 K-means 最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

Statistics	Value
參數	Number of clusters
群數	3
每群的數量	0: 12681 1: 14797 2: 14827
主要的音樂類別	2 is Underground Rap
Rand Index	0.6869803334548603
Normalized Mutual Information	0.27263503968041736
Adjusted Mutual Information	0.272505204073346
V-measure	0.2726350396804173
Fowlkes-Mallows Scores	0.310132635528717
Silhouette Coefficient	0.1241779920506309

Elbow Diagram and Confusion Matrix

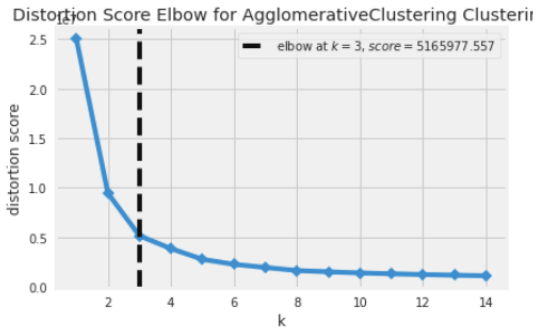


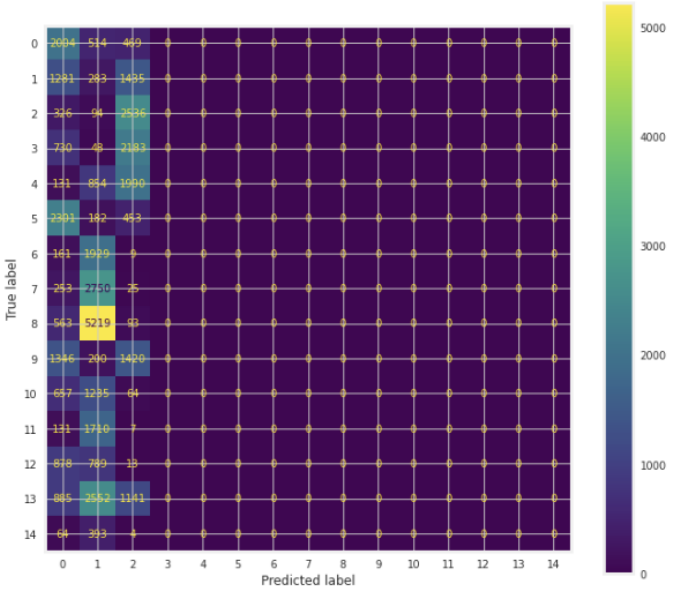


(4) 請執行 Hierarchical Clustering，並列出最佳分群的結果。結果包括 Elbow Diagram、參數、 群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

Statistics	Value
參數	Number of clusters、linkage type、distance
群數	3
每群的數量	0: 11711 1: 11842 2: 18752
主要的音樂類別	2: Underground Rap
Rand Index	0.6683292477400535
Normalized Mutual Information	0.22931345594632954
Adjusted Mutual Information	0.22917501014712952
V-measure	0.22931345594632954
Fowlkes-Mallows Scores	0.296143103258433
Silhouette Coefficient	0.10129300699842342

Elbow Diagram and Confusion Matrix

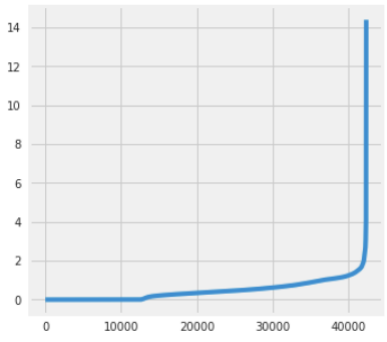


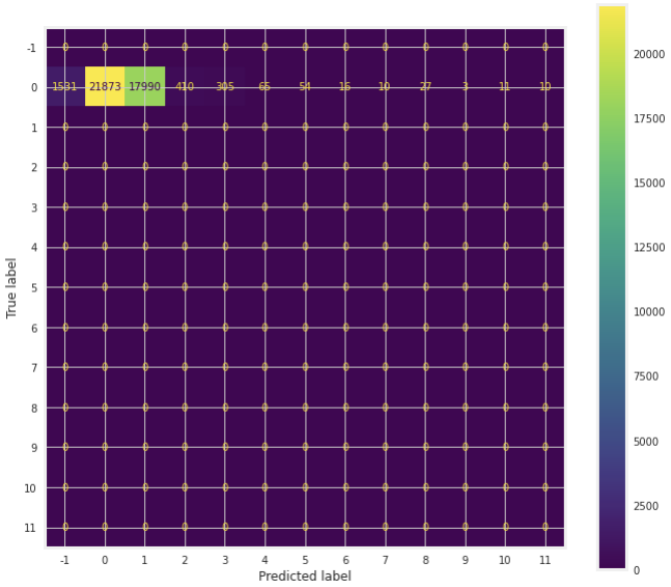


(5) 請執行 DBSCAN，並列出最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

Statistics	Value
參數	neighborhood size
群數	10
每群的數量	0: 21873、1: 17990、2: 410、3: 305、4: 65、5: 54、6: 16、7: 10、8: 27、9: 3、10: 11 Otherwise: 1531
主要的音樂類別	0 is Underground Rap
Rand Index	0.4496019614155614
Normalized Mutual Information	1.6567008922942938e-15
Adjusted Mutual Information	3.4012731976784166e-15
V-measure	1.6567008922942924e-15
Fowlkes-Mallows Scores	0.6705236471710461
Silhouette Coefficient	0.0804763098194222

Elbow Diagram and Confusion Matrix

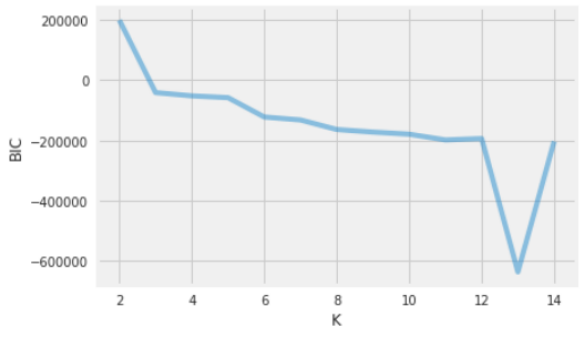


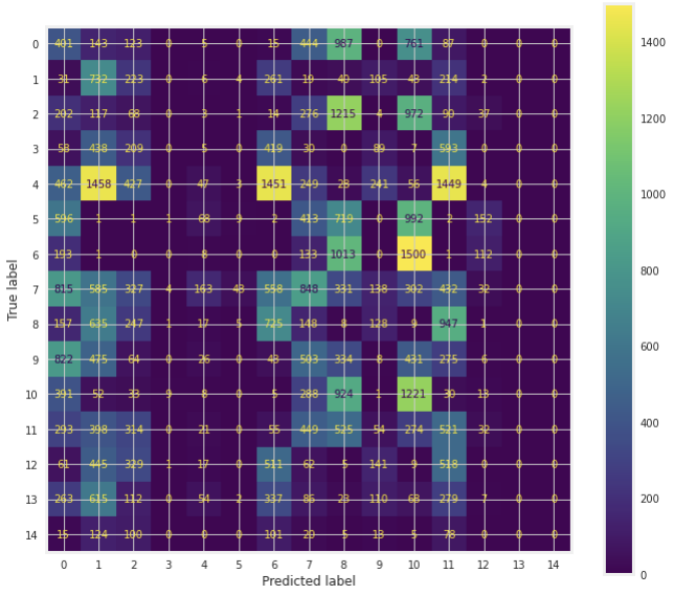


(6) 請執行 GMM，並列出最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

Statistics	Value
參數	Number of clusters
群數	13
每群的數量	0: 509、1: 6805、2:4598、3:314、4:3105、5:2769、6:5466、7:7578、8:905、9:150、10: 1997、11:7982、12:127
主要的音樂類別	11 is Underground Rap
Rand Index	0.832593847207826
Normalized Mutual Information	0.17969857215491994
Adjusted Mutual Information	0.17900768025064157
V-measure	0.17969857215491994
Fowlkes-Mallows Scores	0.16905267769622698
Silhouette Coefficient	0.02134537575448128

Elbow Diagram and Confusion Matrix

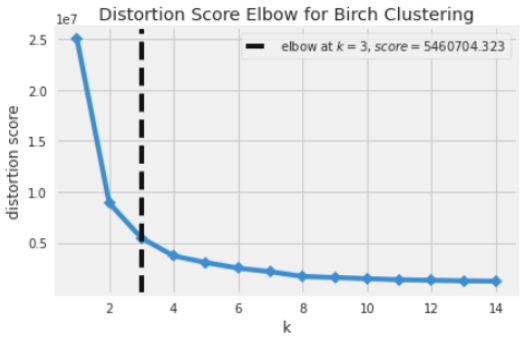


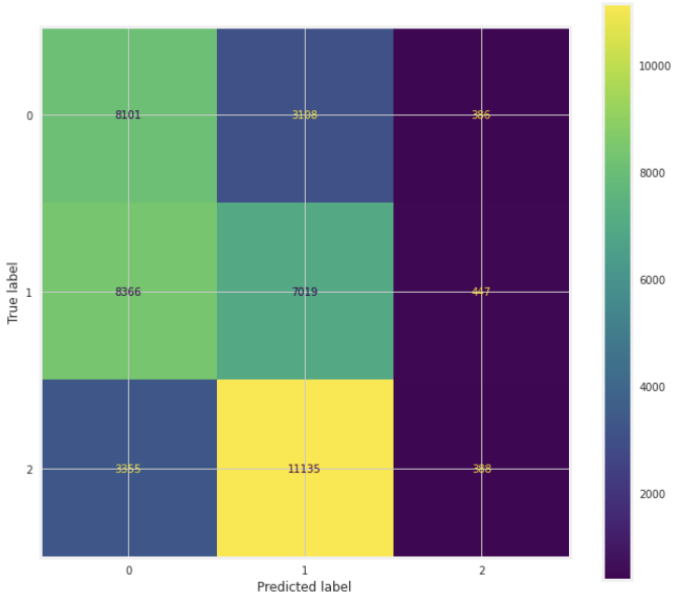


(7) 請執行 BIRCH，並列出最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

Statistics	Value
參數	branching factor, threshold, optional global clusterer
群數	3
每群的數量	0: 21262 1: 19822 2: 1221
主要的音樂類別	0 is Underground Rap
Rand Index	0.5572622767164677
Normalized Mutual Information	0.08476237662452793
Adjusted Mutual Information	0.08471665590849174
V-measure	0.08476237662452793
Fowlkes-Mallows Scores	0.4609514980507308
Silhouette Coefficient	0.11062887813553333

Elbow Diagram and Confusion Matrix





(8) 針對以上的分群方法，比較其分群效果，哪個分群方法效果最好？為什麼？

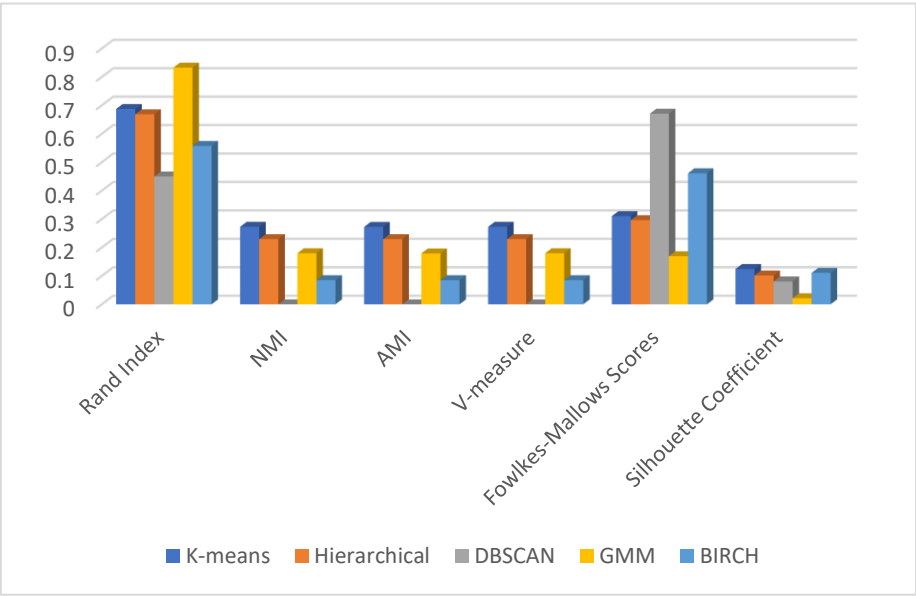
We consider the six factors, including Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient as the indexes.

If we consider the “rand index”, then GMM is best. If we consider the “NMI”, then K-Means is best. If we consider the “AMI”, then K-means is best. If we consider the “V-measure”, then K-means is best. If we consider the “Fowlkes-Mallows Scores”, then DBSCAN is best. If we consider the “Silhouette Coefficient”, then K-means is best. Because the result is different, we use arithmetic mean as index. Namely,

$$AM = \frac{\sum RI + NMI + AMI + VM + MS + SC}{6}$$

Hence, we regard K-means as the best.

Index	K-means	Hierarchical	DBSCAN	GMM	BIRCH
Rand Index	0.68698033	0.668329248	0.449601961	0.832593847	0.557262277
NMI	0.27263504	0.229313456	1.66E-15	0.179698572	0.084762377
AMI	0.2725052	0.22917501	3.40E-15	0.17900768	0.084716656
V-measure	0.27263504	0.229313456	1.66E-15	0.179698572	0.084762377
Fowlkes-Mallows Scores	0.31013264	0.296143103	0.670523647	0.169052678	0.460951498
Silhouette Coefficient	0.12417799	0.101293007	0.08047631	0.021345376	0.110628878
AM	0.32317771	0.292261213	0.20010032	0.260232788	0.23051401



(9) 針對以上的分群方法，哪個分群方法效率最佳？為什麼？

Now, we consider the factor of “time” as the index.

From the table for time complexity, with loss the generality, we should know that the order is BIRCH < GMM < DBSCAN < K-means < Hierarchical Clustering, if the number is so big than the other parameters in time complexity.

However, in my experiments, we compute the time value form draw the elbow diagram, calculate the digit, including Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient to draw the Confusion Matrix, so we have the order is DBSCAN < K-Means < BIRCH < GMM < Hierarchical Clustering.

Hence, we regard “DBSCAN” as the best method in my experiments.

Table for Time Value in my experiment.

The method of clustering	Time value (Unit: Second)
K-means	131.237559
Hierarchical Clustering	1433.456356
DBSCAN	78.479488
GMM	466.816227
BIRCH	320.437908

Table for Time Complexity

The method of clustering	Time complexity
K-means	O(n*n)
Hierarchical Clustering	O(k*n*n)
DBSCAN	O(n*logn)
GMM	O(N*K*D*D*D)
BIRCH	O(n)

Note that:

1. n is the number of objects to be clustered in K-means model.
2. n is the number of objects to be clustered and k is the number of clusters in Hierarchical Clustering model.
3. n is the number of objects to be clustered in DBSCAN model.
4. N is the number of data points; K is the number of Gaussian components and D is the problem dimension in GMM model.
5. n is the number of objects to be clustered in BIRCH model.

(10) 有哪些可能的方法，可以提升分群的效果？

From this selection, I use package named seaborn to find the no important feature. According to the following table, we drop the feature “time_signature”. And then, from the exercise eight, we know the K-means being best in my experiment. Hence, I use new dataset to be clustering in K-Means model.

In the sum, no matter any indexes in my experiment are increase, so we improve the result of clustering.

Statistics	K-Means value	Improve K-Means Value
Rand Index	0.6869803334548603	0.687322219810357
NMI	0.27263503968041736	0.27422102531477643
AMI	0.272505204073346	0.27409146914715327
V-measure	0.2726350396804173	0.27422102531477643
F-M Scores	0.310132635528717	0.31136208633253515
Silhouette Coefficient	0.1241779920506309	0.12435245342511471