(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: [-60, 0] 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

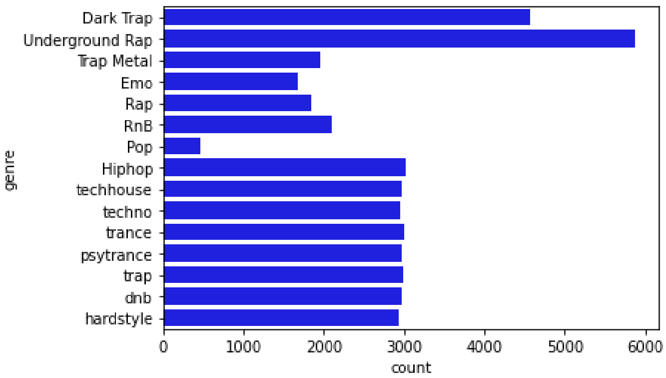
Acousticness	Meaning: The value that describes how acoustic a song is. Higher values mean that the song is most likely to be an acoustic one. Range: [0,1] Missing value: No Noise: No
Instrumentalness	Meaning: The relative ratio of the track being instrumental. Higher values mean that the song contains 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: [50,150] 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, I use obj.drop to delete unimportant features.

Secondly, I use obj.isnull().sum() to check whether the data collection having missing value or not.

Thirdly, I check the number of the different “genre”. (In figure1)



▲Figure1

▶Figure 2

```
genre
Dark Trap          5875
Emo                5875
Hiphop             5875
Pop                5875
Rap                5875
RnB                5875
Trap Metal         5875
Underground Rap    5875
dnb                5875
hardstyle          5875
psytrance          5875
techhouse          5875
techno             5875
trance             5875
trap               5875
dtype: int64
```

Fourthly, since this dataset has the imbalance problem, I use “RandomOverSampler” from imblearn.over_sampling package to solve this problem. (In figure 2)

Fifthly, I use the function LabelEncoder() and StandardScalar() to normalize the feature.

Finally, I cut my data collection into four parts, x_train, x_test, y_train and y_test, respectively to finish my experiment.

(3) **請執行 Random Forest，並列出最佳分類的結果。結果包括 Imbalance 處理(Over-Sampling、 Under-Sampling) 、Cross-Validation、Random Forest 參數、Accuracy、 Confusion Matrix、哪些類別的音樂彼此之間比較不易分別、Feature Importance、運用哪些方法提升分類準確率。(執行 Output Accuracy 的畫面，請截圖）**

I use the “RandomOverSampler” from imblearn.over_sampling package to solve the imbalance problem and the result in figure 2.

From the part of Cross-Validation, I use the “cross_val_score” and setup the cv=5 to compute the score and the result in the following table.

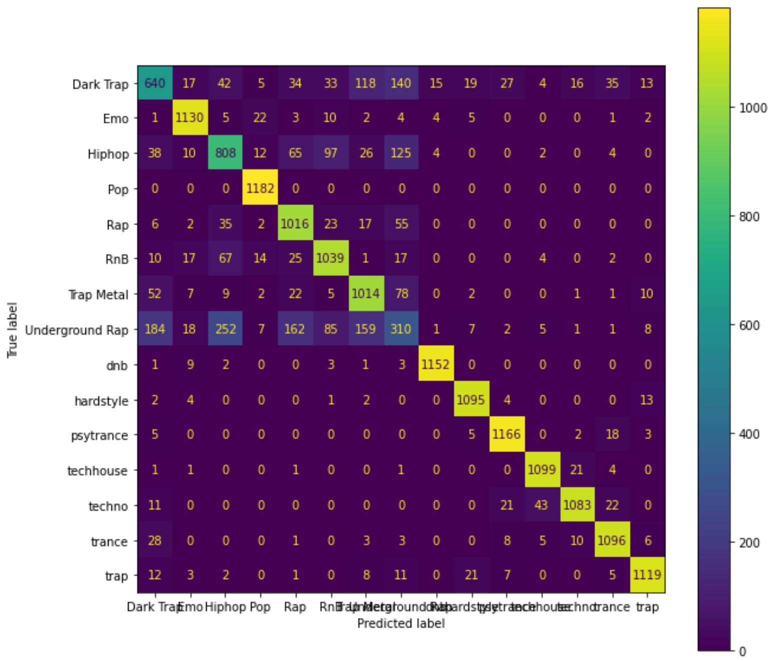
K	1	2	3	4	5
score	0.8187234	0.81764539	0.87483688	0.89548936	0.91302128

The following table is the parameters that I setup in my experiment.

Parameters	Value
criterion	entropy
n_estimators	10
random_state	3
n_jobs	2
The other parameters	default

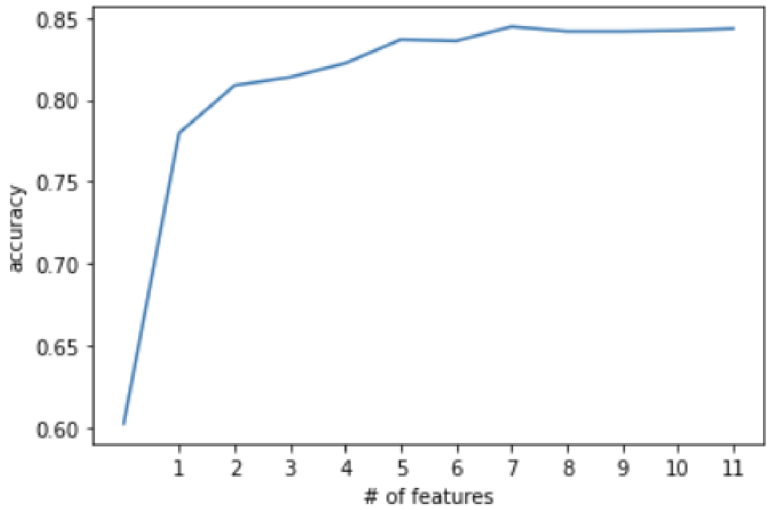
The following figure is the result of accuracy and confusion matrix.

```
Accuracy Score = 0.8375269555095723
Confusion Matrix:
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7fc0b7c54410>
```



By the excel Pivot Report, I figure out the genre “Underground Rap” is the most difficult to cluster. For example, having 177 Underground Rap in the Dark Trap, having 152 Underground Rap in the Trap Metal, and so on.

In the next problem, I use the “CatBoostRegressor, Pool, EShapCalcType, EFeaturesSelectionAlgorithm” from “catboost” package to look for the important features. I setup the parameters, num_features_to_select, from one to twelve and using accuracy as my index to decide how many features I want.



In final problem, according to the above figure, I use the eight features as my features, including danceability, energy, key, loudness, speechiness, instrumentalness, valence and tempo. Using this technique, I have the accuracy 0.84442863811257 is better than choosing all features. Hence, I improve this classification.

(4) 請執行 Support Vector Machine，並列出最佳分類的結果。結果包括 Imbalance 處理(OverSampling、Under-Sampling)、Cross-Validation、SVM 參數、Accuracy、 Confusion Matrix、 哪些類別的音樂彼此之間比較不易分別、Feature Importance、運用哪些方法提升分類準確率。 （執行 Output Accuracy 的畫面，請截圖）

I use the “RandomOverSampler” from imblearn.over_sampling package to solve the imbalance problem and the result in figure 2.

From the part of Cross-Validation, namely, called k-fold CV, I use the “cross_val_score” and setup the cv=3 to compute the score and the result in the following table.

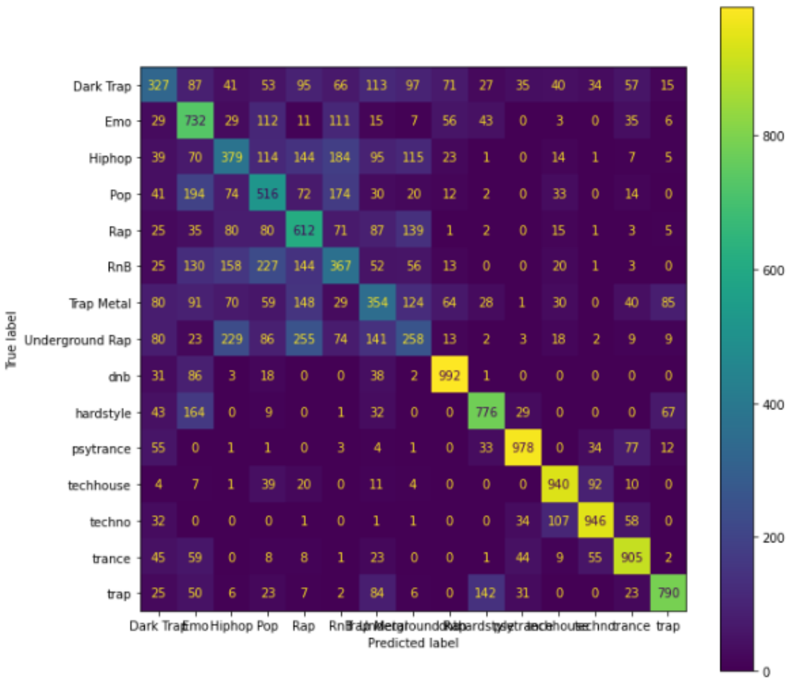
K	1	2	3
Score	0.55339574	0.56810213	0.5627234

The following table is the parameters that I setup in my experiment.

Parameters	Value
kernel	linear
C	1
random_state	42
The other parameters	default

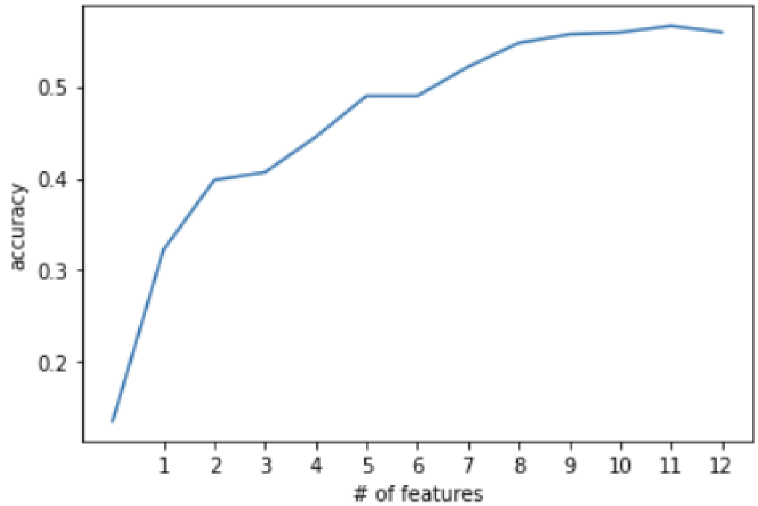
The following table is the result of accuracy and confusion matrix

```
Accuracy Score = 0.5592542657007442
Confusion Matrix:
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7fdbbb7a6690>
```



By the excel Pivot Report, I figure out the genre “RnB” (139), “Underground Rap” (228) and Hiphop(391) can not clearly split. Besides, the “psytrance” can split easily.

In the next problem, I use the “CatBoostRegressor, Pool, EShapCalcType, EFeaturesSelectionAlgorithm” from “catboost” package to look for the important features. I setup the parameters, num_features_to_select, from one to twelve and using accuracy as my index to decide how many features I want.



In final problem, according to the above figure, I use the eleven features as my features, including danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. Using this technique, I have the accuracy 0.5642582687097452 is better than choosing all features. Hence, I improve this classification.

(5) 請根據 Linear SVM 的 Feature Importance，選出 Top-N 重要的 Features，並運用這些 Features 重新執行作業二的 Clustering，觀察效果是否有提升。(執行 Output 效果的畫面，請截圖)

From the problem four, I delete the feature time_signature, so I choose the eleven features in my experiment. (Top-11).

In this selection, I do the five times experiments, including K-means clustering, Hierarchical clustering, DBSCAN, GMM and birch and I use the six functions as my index to observe whether this method is better than before or not. Besides, since all the clustering is unsupervised, I label the “genre” as my correct answer in my experiments. In order to satisfy the consistency, the number of clustering decide by the elbow diagram. Namely, setting n_cluster=3 in K-means, choosing n_clusters = 3 in Hierarchical clustering, choosing eps = 2, min_samples = 10 in DBSCAN, choosing n_components=13 in GMM, and choosing n_clusters=3 in Birch clustering.

This table is the result of the K-means. According to the table, all the index is better than before, so the K-means clustering is improvement.

Statistics	12 features	11 features
Rand Index	0.6869803334548603	0.687322219810357
Normalized Mutual Information	0.27263503968041736	0.27422102531477643
Adjusted Mutual Information	0.272505204073346	0.27409146914715327
V-measure	0.2726350396804173	0.27422102531477643
Fowlkes-Mallows Scores	0.310132635528717	0.31136208633253515
Silhouette Coefficient	0.1241779920506309	0.13229700437768982

```
rand Index: 0.687322219810357
Normalized Mutual Information: 0.27422102531477643
Adjusted Mutual Information: 0.27409146914715327
V-measure: 0.27422102531477643
Fowlkes-Mallows Scores: 0.31136208633253515
Silhouette Coefficient: 0.13229700437768982
```

This figure is the result of the K-means using 11 features.

This table is the result of the Hierarchical Clustering. According to the table, the feature “time_signature” play the no important roles in this experiment, because the three index is not improve and the three index is improve in my experiments.

Statistics	12 features	11 features
Rand Index	0.6683292477400535	0.6349808203824221
Normalized Mutual Information	0.22931345594632954	0.2525502336763705
Adjusted Mutual Information	0.22917501014712952	0.25241282415192756
V-measure	0.22931345594632954	0.2525502336763704
Fowlkes-Mallows Scores	0.296143103258433	0.31020982462792496
Silhouette Coefficient	0.10129300699842342	0.08654225502939954

```
rand Index: 0.6349808203824221
Normalized Mutual Information: 0.2525502336763705
Adjusted Mutual Information: 0.25241282415192756
V-measure: 0.2525502336763704
Fowlkes-Mallows Scores: 0.31020982462792496
Silhouette Coefficient: 0.08654225502939954
```

This figure is the result of the Hierarchical Clustering using 11 features.

This table is the result of the DBSCAN clustering. According to the table, the feature “time_signature” plays the important role, because if I do not use this feature some index will become zero.

Statistics	12 features	11 features
Rand Index	0.4496019614155614	0.4905386215403915
Normalized Mutual Information	1.6567008922942938e-15	0.0
Adjusted Mutual Information	3.4012731976784166e-15	3.434715439758919e-17
V-measure	1.6567008922942924e-15	0.0
Fowlkes-Mallows Scores	0.6705236471710461	0.7003846240034054
Silhouette Coefficient	0.0804763098194222	0.10715349106849564

```
rand Index: 0.4905386215403915
Normalized Mutual Information: 0.0
Adjusted Mutual Information: 3.434715439758919e-17
V-measure: 0.0
Fowlkes-Mallows Scores: 0.7003846240034054
Silhouette Coefficient: 0.10715349106849564
```

This figure is the result of the DBSCAN using 11 features.

This table is the result of the GMM clustering. According to the table, five index is better than before, so the GMM clustering is improvement by this method.

Statistics	12 features	11 features
Rand Index	0.832593847207826	0.7999745472731431
Normalized Mutual Information	0.17969857215491994	0.19909620248715318
Adjusted Mutual Information	0.17900768025064157	0.1983700077731238
V-measure	0.17969857215491994	0.1990962024871532
Fowlkes-Mallows Scores	0.16905267769622698	0.19864981591118044
Silhouette Coefficient	0.02134537575448128	0.0261791728696776

```
rand Index: 0.8303133293704442
Normalized Mutual Information: 0.20664903333331139
Adjusted Mutual Information: 0.20598050857134337
V-measure: 0.20664903333311396
Fowlkes-Mallows Scores: 0.18874536603914377
Silhouette Coefficient: 0.0261791728696776
```

This figure is the result of the GMM using 11 features.

This table is the result of the birch clustering. According to the table, in the birch clustering, the feature “time_signature” play the key point.

Statistics	12 features	11 features
Rand Index	0.5572622767164677	0.5342536307461073
Normalized Mutual Information	0.08476237662452793	0.03925297967905736
Adjusted Mutual Information	0.08471665590849174	0.03920767595267153
V-measure	0.08476237662452793	0.03925297967905736
Fowlkes-Mallows Scores	0.4609514980507308	0.39719438290289577
Silhouette Coefficient	0.11062887813553333	0.0953481531797912

```
rand Index: 0.5342536307461073
Normalized Mutual Information: 0.039252979679057366
Adjusted Mutual Information: 0.03920767595267153
V-measure: 0.039252979679057366
Fowlkes-Mallows Scores: 0.39719438290289577
Silhouette Coefficient: 0.0953481531797912
```

This figure is the result of the Birch clustering using 11 features.