Spotify Analysis

Shao Yan Chen

June ,2022

1 研究動機與目的

在現今的時代中，音樂已經成為人們生活中不可或缺的一環，而 Spotify 為目前最受歡迎和使用最廣泛的流行音樂平台，每月使用的用戶約為 3.45 億人次。它提供了來自世界各地的各種歌曲和類型，因此想藉由這學期所學的資料分析方法，來探討不同類型的音樂會有怎樣的特性，也希望能更好地了解 Spotify 上受歡迎的歌曲背後的原因。透過 Kaggle 平台上的開放資料，以 Spotify Track DB 這個資料集為例，根據資料集中的 18 個變數，探討音樂類型與不同特徵之間的關係。

2 資料介紹及資料預處理

2.1 資料來源

https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db

2.2 變數介紹

研究包含 232725 筆資料與 18 個變數，變數介紹如表 1 所示。

表 1: 變數介紹

變數名稱	變數屬性	變數說明
genre	類別型	音樂類型
artist_name	類別型	表演者姓名
track_name	類別型	曲目名稱
track_id	類別型	曲目編號
popularity	連續型	受歡迎程度 (等級 0 ∼ 100)
acousticness	連續型	原聲樂器占有比例 (0 ∼ 1 之間)
danceability	連續型	可跳舞等級 (0 ∼ 1 之間)
duration_ms	連續型	音樂持續時間 (秒)
energy	連續型	能量 (按照歌曲的強度和活力來判別，0 ∼ 1 之間)
Instrumentalness	連續型	樂器性 (0 ∼ 1 之間，0.5 以上為純音樂)
key	類別型	調性 (A、A#、C、C#)
liveness	連續型	現場感 (0 ∼ 1 之間，越高可能為 LIVE 的音檔)
loudness	連續型	響度
mode	類別型	調式 (Major(大調)、Minor(小調) 兩類)
speechiness	連續型	某個曲目出現多少話語 (0 ∼ 1 之間)
tempo	連續型	平均音速
time_signature	類別型	拍子記號
valence	連續型	情緒值 (0 ∼ 1 之間，越靠近 0 情緒越悲傷，越靠近 1 越開心)

2.3 資料預處理

研究資料包含 232725 個資料，18 個變數，由於表演者名稱、曲目編號、曲目名稱這三個變數無法進行解釋，故將此移除，接著將 mode 改成 0、1 變數，大調改成 1、小調改成 0，key 依照順序改成 1 ∼ 12，最後透過 popularity (受歡迎程度) 定義一個新的二元變數，命名為 popularity_group ，將受歡迎程度大於或等於 57 的曲目歸類為" 流行"，因此將被歸類為 1，而曲目受歡迎程度低於 57 的將被歸類為 0。

3 描述性統計分析

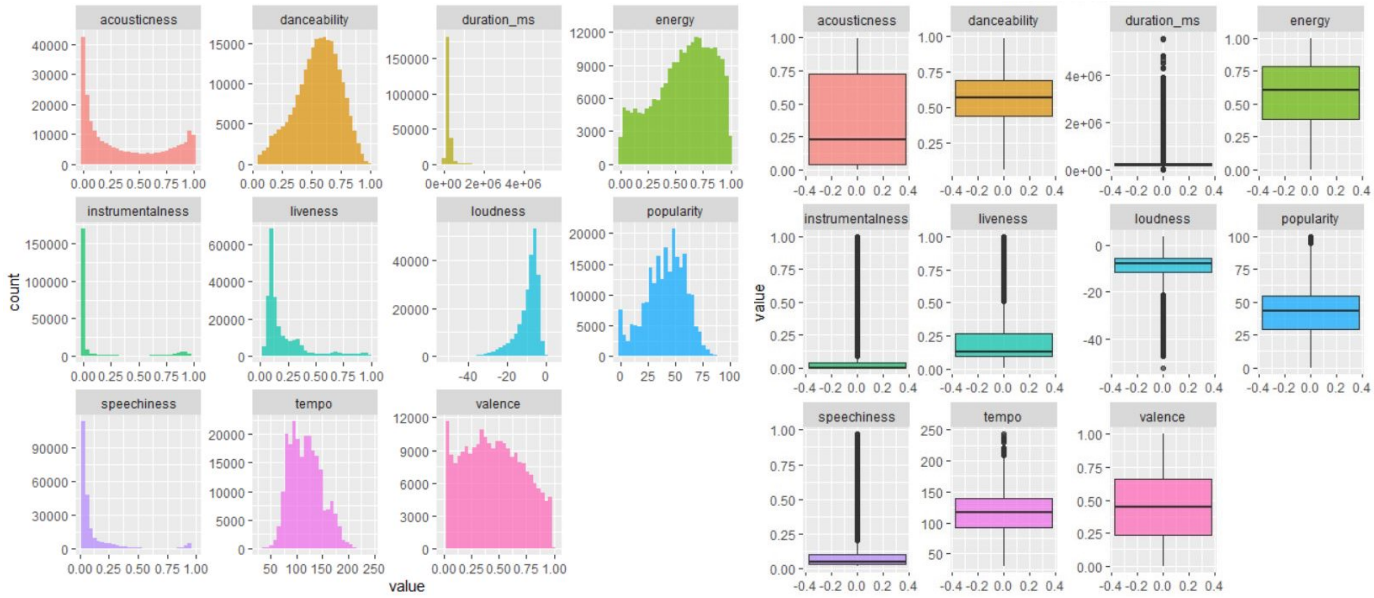


圖 1: 連續型變數之直方圖與盒狀圖

圖 1 為資料預處理後連續型變數之直方圖與盒狀圖，從直方圖中，我們觀察到

(1) 大多數歌曲的時長約為 2.5 到 4 分鐘

(2) 許多觀察值的樂器性不大於 0.1，約佔資料的 80%

(3) 大多數歌曲的響度在 -5dB 和 -10db 之間

(4) 大多數音樂的語言性小於 0.25，表示更多語言性的歌曲可能不太大眾

而從盒狀圖中，我們觀察到歌曲的時長、樂器性、現場感、響度、語言性皆有較多的離群值。

根據最受歡迎的歌曲出現在流行歌曲列表中的頻率彙整出最流行的音樂類型，如圖 2 所示，由圖 2 可知，Pop 為最受歡迎歌曲中最流行的音樂類型。

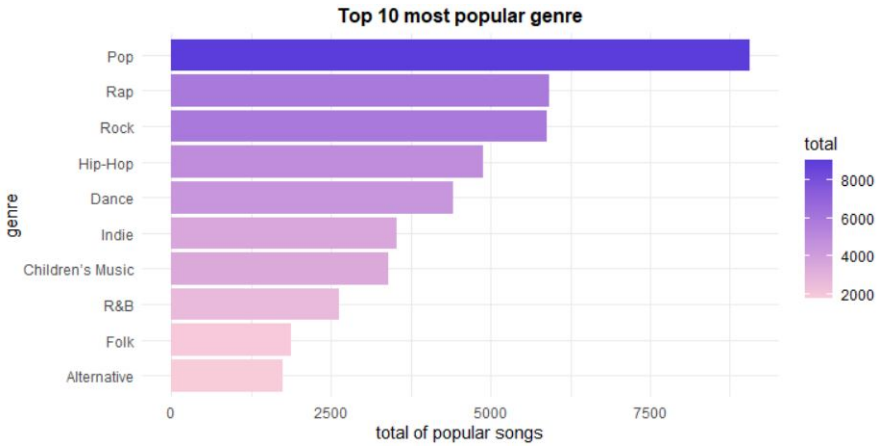


圖 2: 最受歡迎歌曲的音樂類型 Top 10

接著根據最受歡迎的歌曲出現在流行歌曲列表中的頻率彙整出最流行的音樂調性與調式，如圖 3 所示，由圖 3 可知，C# 為最受歡迎歌曲中最流行的音樂調性，而 Major(大調) 為最受歡迎歌曲中最流行的音樂調式。

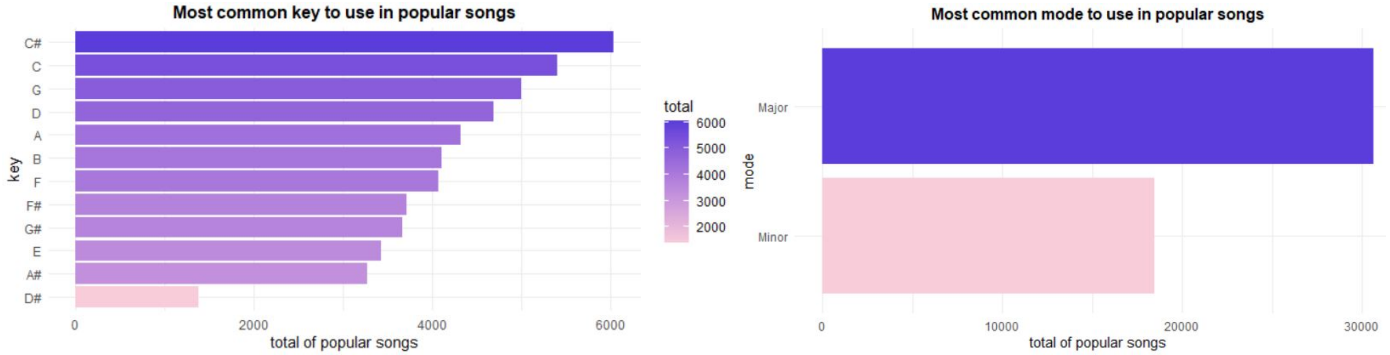


圖 3: 最受歡迎歌曲的音樂調性與調式

4 Unsupervised Learning(非監督式學習)

4.1 Clustering Analysis(分群分析)

在做分群分析之前，由於 duration_ms 變數與其他變數的範圍差距過大，因此必須對所有連續型變數進行標準化，接著由於資料筆數過大，電腦設備不夠完善，無法對這麼龐大的資料進行分群，因此將透過隨機抽樣的方式選取進一步要分析的歌曲以減少資料集中的觀察量。由於分群分析屬於非監督式學習演算法，我們選擇刪去了資料集中的類別型變數進行分群分析。分群分析的目的是使組內的總變異最小化，並使組間的總變異最大化。

4.1.1 Optimal number of clusters(決定最佳的分群數目)

我們必須找到一個 n，使得當資料分成 n 群時，組內的總變異 (SSE) 最小，那麼我們可以說 n 是最佳的分群數目，由圖 4 可知，分三群為最佳的分群數目。

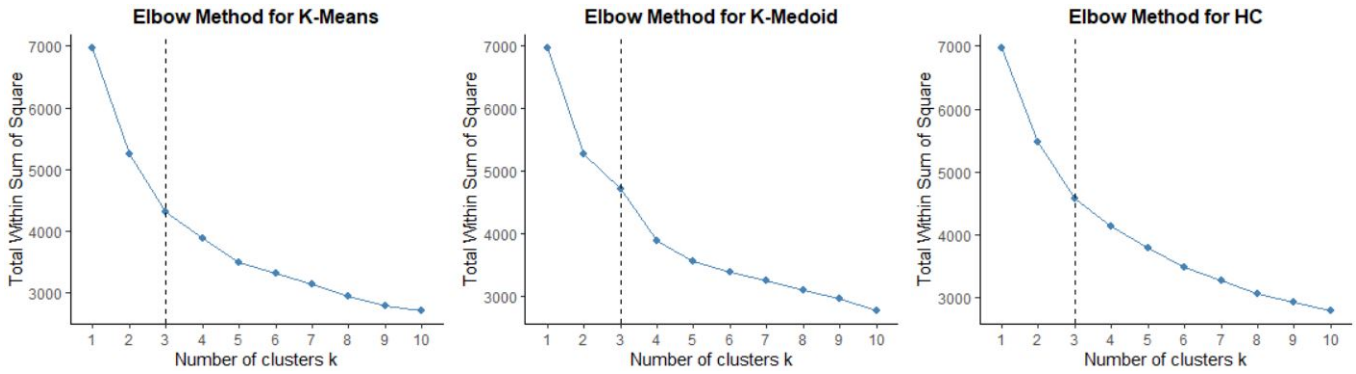


圖 4: 最佳的分群數目

4.1.2 K-means

K-means 透過 Euclidean distance (歐式距離) 來測量群集之間的相似性，其基本的概念是用群集中心 (mean) 來表示群集，它的目標是使集群內差異的平方誤差最小化，呈現結果如圖 5 與表 2 所示，每群個數分別為 530、38、130。

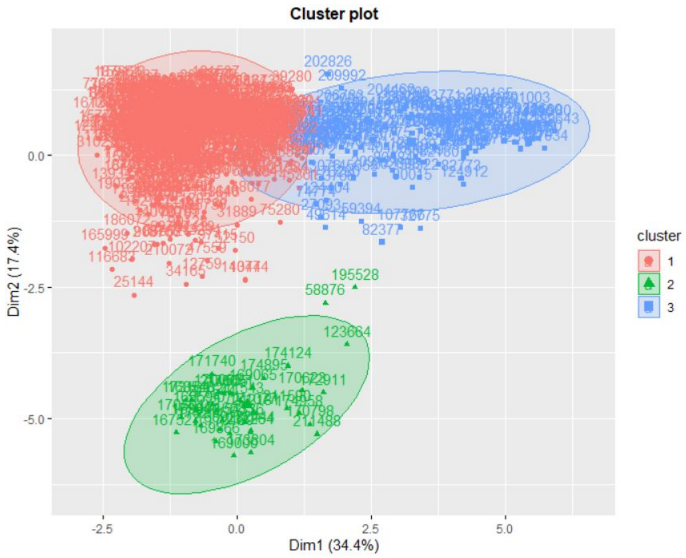


圖 5: K-means

表 2: K-means

	acousticness	danceability	duration_ms	energy	instrumentalness
1	-0.4206763	0.27008540	-0.04321850	0.3453361	-0.2755902
2	1.2535968	0.09960124	0.47896787	0.2502691	-0.4613934
3	1.3486288	-1.13023160	0.03619252	-1.4810643	1.2584291
	liveness	loudness	speechiness	tempo	valence
1	-0.1040773	0.4083632	-0.1679883	0.1505782	0.2535793
2	2.5265602	-0.4038073	3.7420224	-0.7382727	-0.1044746
3	-0.3142177	-1.5468294	-0.4089468	-0.3980929	-1.0032844

4.1.3 K-medoid

PAM(Partitioning Around Medoids) 是一種常用的 K-medoids 演算法。中心點可以定義為與群集中所有對象的平均相似度最小的群集。與 K-means 相比，K-medoid 對離群值更穩健 (robust)，因為它最小化了成對差異的總和，而不是距離的總和，呈現結果如圖 6 所示。

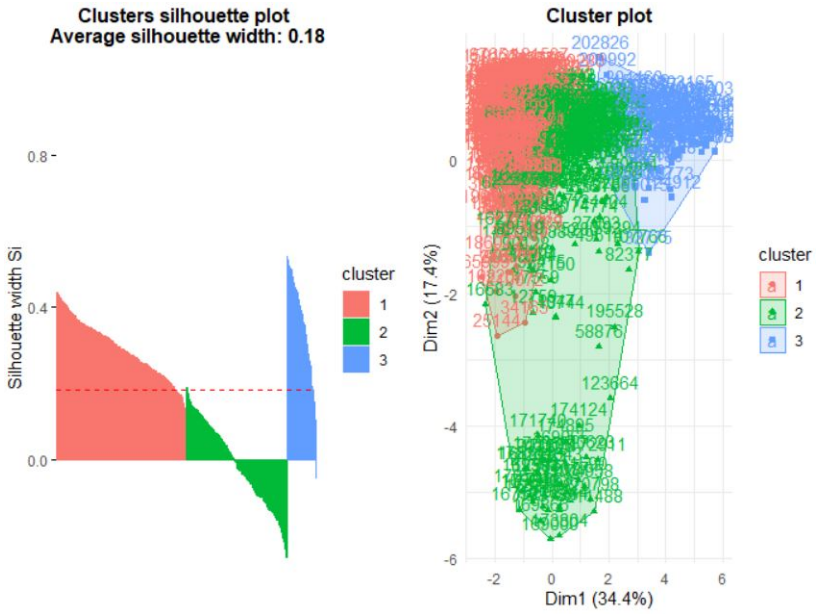


圖 6: K-medoid

4.1.4 Hierarchical Clustering (階層式分群)

我們使用 Euclidean distance (歐式距離) 和 Ward method (華德法) 進行階層式分群，組別之間的差異定義為當我們合併它們時，平方和將會增加多少，而 Ward method (華德法) 的目的是合併集群，使得組內變異的增加是最小的，呈現結果如圖 7 所示。

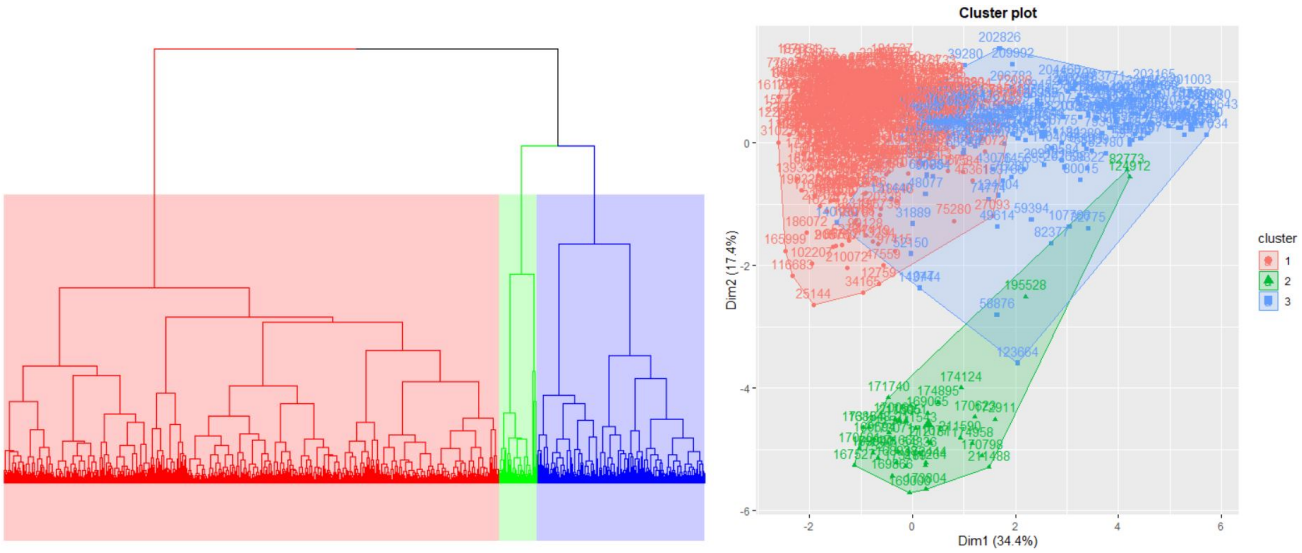


圖 7: Hierarchical Clustering

4.1.5 Conclusion of the Cluster Analysis

每個集群中的音樂特徵

- 第 1 群：具有高度可舞性、能量、響度、平均音速、情緒值的歌曲，但原聲性和持續時間最低。
- 第 2 群：具有高度原聲性、能量、現場感和語言性的歌曲，持續時間相對較長，但樂器性和平均音速最低。
- 第 3 群：具有高度原聲性、樂器性的歌曲，但可舞性、能量、現場感、響度、語言性和情緒值最低。

4.2 Principal Component Analysis (PCA)

利用原變數之間的線性組合來達成保留原資料的資訊並降低維度，這樣對資料的視覺話會有很大的幫助。首先將連續型的變數挑選出來，分別是 popularity、acousticness、tempo、danceability、duration_ ms、energy、liveness、loudness、speechiness、instrumentalness 及 valance。做完 PCA 後共有 11 個主成分，表 3 為各主成分的標準差、解釋比例及累積解釋比例。

表 3: 各主成分的標準差、解釋比例及累積解釋比例

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6
Sd	1.90012	1.30768	1.08224	0.99917	0.928290	0.869917
p.o.v	0.32822	0.15546	0.10647	0.09089	0.078338	0.068796
c.p	0.32822	0.48368	0.59015	0.68105	0.759390	0.828186
	Comp7	Comp8	Comp9	Comp10	Comp11	
Sd	0.798658	0.696691	0.612528	0.526065	0.338787	
p.o.v	0.057986	0.044125	0.034108	0.025158	0.010434	
c.p	0.886173	0.930298	0.964407	0.989565	1	

接著要選取主成分個數，將透過 permutation test 進行選取，如表 4 所示，從檢定結果可知前三個主成分的 p-value 極小，故選取前三個主成分，其可以解釋的變異為 59.02%。

表 4: Permutation test

0	0	0	1	1	1	1	1	1	1	1

表 7 為各主成分與各連續變數之關係，由表可知:

- 第一主成分與 popularity、danceability、energy、loudness、tempo 及 valance 成正比，若為正是比較有活力及節奏感的音樂。
- 第二主成分與 liveness、speechiness 成反比，若為正可能是配樂或伴奏。
- 第三主成分與 duration_ms 成正比與 valence 成反比，若為正可能是較悲傷的音樂類型。

表 5: 各主成分與各連續變數之關係

	Comp1	Comp2	Comp3
popularity	0.236	0.298	
acousticness	-0.420	-0.188	-0.209
danceability	0.334		-0.451
duration_ms			0.594
energy	0.446		0.247
instrumentalness	-0.322	0.183	
liveness		-0.619	0.254
loudness	0.467		0.153
speechiness			-0.646
tempo	0.157	0.149	0.259
valence	0.324		-0.412

4.3 Correspondence Analysis (CA)

將類別變數從列連表的數據轉換為圖表，可以利用圖形去觀察出數據面不易得出的結論。此資料的 mode 變數只分成 major 與 minor，為了畫出變數間的二維關係，因此只對 genre、key 來做，呈現結果如圖 8 所示。

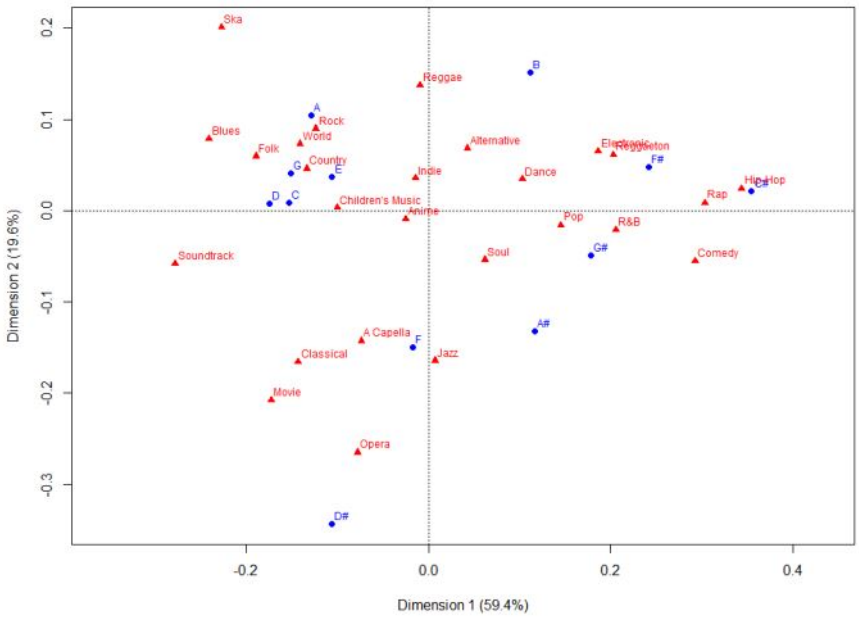


圖 8: Correspondence Analysis Factor Map

2 個維度可以解釋 79.01%，可以看出 Rap 和 Hip-Hop 較相似，大多是 C#、F#、G#，而 Eletronic 和 Reggaeton 這類型的音樂較相似，大多是 F#、A、C、D、G、E 的音樂較相似，D# 與其他的音樂差異最大。

4.4 Factor Analysis (FA)

觀察到的變數可以由無法觀測到的因子的線性組合組成，希望能找到最少的因子來解釋原始變數。可使用最大因子數為 6，計算後可解釋 63.9%，雖然因子數越多可以解釋的比例越大，但會使模型變得十分複雜且難以解釋，在權衡後決定選擇因子數為 4。使用 varimax 所得到的結果，可以解釋 54.4%，但變數間還是有些重疊，因此再嘗試用 promax 的方法，使用 promax 方法因為不是正交

旋轉，因此解釋能力稍微下降到 50.1%，但將 cutoff point 設定在 0.4 後僅有 loudness 會有些微重疊。

表 6: Factor Analysis

	Factor1	Factor2	Factor3	Factor4
popularity			0.525	
acousticness	-0.615			
danceability			0.610	
duration_ms				-0.131
energy	0.957			
instrumentalness			-0.554	
liveness		0.557		
loudness	0.600			
speechiness		0.922		
tempo	0.226			
valence				0.989

表 6 為因素分析表，由表可知:

- Factor1 由 acousticness、energy、loudness、tempo 組成
- Factor2 由 liveness、speechiness 組成
- Factor3 由 popularity、danceabilit、instrumentalness 組成
- Factor4 由 duration_ms、valance 組成

表 7: 各變數被解釋的比例

popularity	acousticness	danceability	duration_ms	energy	instrumentalness
0.2993	0.6709	0.5560	0.0259	0.995	0.3732
liveness	loudness	speechiness	tempo	valence	
0.3640	0.8080	0.8207	0.0737	0.995	

表 7 為各變數被解釋的比例，可以看出 tempo 及 duration_ms 不論是在解釋比例以及 factor loadings 都不太理想，從 correlation matrix 可以看出這兩個變數跟其他變數的相關性相當低，可能是此原因導致。

5 Supervised Learning(監督式學習)

5.1 Classification Analysis

針對此資料，在分類前將資料量過少的類別進行刪除，並把相同風格音樂類型合併，最後 y(genre) 的類別剩下 Rock / Country / Hip-Hop / Reggae / Comedy / Electronic / Jazz / R&B / World / Pop / Soul / Classical / Dance / Blues / Ska ，共 15 類。

5.1.1 Linear Discriminant Analysis (LDA)

- 簡介：對資料進行投影，找出分類效果最佳的投影方向。
- 結果：進行 10-CV 得出 Accuracy 約 46%，對測試集預測率約 46%。
- 說明：其預測正確率不夠高，考量到資料間存在非線性關係，故以線性切割較難成功分類。

5.1.2 Quadratic Discriminant Analysis (QDA)，僅供參考，此資料不適用

- 簡介：LDA 變體，當不同分類樣本的斜方差矩陣不同時，可使用二次判別。資料須為多維常態分配。
- 結果：執行多類別 QDA 得到 10-CV 約 0.45，對測試集預測率約 0.45。
- 說明：因資料並非多維常態分配，使用 QDA 預測的結果並沒有比 LDA 好。

5.1.3 Decision Tree

- 介紹：根據不同演算法 (本分析使用 CART) 找出節點，進行投票分類。
- 結果：並執行 10-CV 訓練出的樹如圖 9 所示，測試集預測率 0.42：

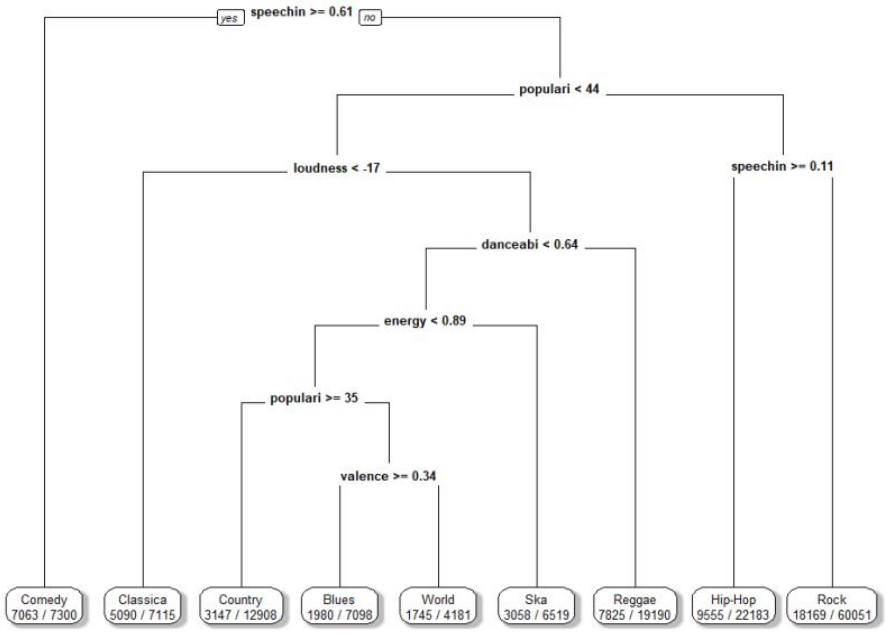


圖 9: Decision Tree

- 說明：
  a. 對於 Comedy 的分類非常準確，其餘也有像 Ska, Classica 這種風格較為明確的音樂類型，在預測上成功率也比較高。
  b. 值得注意的是這棵樹有判斷出 popularity 高的類別有 Hop-hop, Rock，顯現出主流音樂大多是嘻哈、搖滾音樂。
  c. 根據最後畫出來的樹，在判斷上重要的變數有：speechiness, popularity, loudness, danceability, energy, valence，這些變數通常也是人們判斷音樂類型時會去注意的特徵。

考量到音樂類型大部分僅有些微差異，透過變數一層一層篩選上比較難精準分類，故決策樹預測效果沒有特別好。

5.1.4 Multi-nomial Logistic Regression

- 簡介：以 sigmoid 作為 link function 進行迴歸分析。
- 結果：10-CV 約 0.47，預測率約 0.48。
- 說明：是此次預測效果最好的模型，奈何跟上面諸多模型一樣，受限於音樂類型的相似性而較難有好表現。

5.1.5 Nearest Neighbors(NN)

- 簡介：針對附近 k 個資料進行分類，直到整體資料分類不再變動為止。
- 結果：進行 10-CV，選出 k 的最佳參數值為 $k = 10$，建模後預測率僅約 0.18。
- 說明：資料過於相近，影響 KNN 模型分類效能。

5.1.6 Support Vector Machines (SVM)

- 簡介：找出超平面將不同類別的資料分割，因其較複雜需較長的學習時間。
- 結果：此資料難以找到最佳解平面，設定 $\text{iters} = 1000$ 的情況下，估計出預測率 0.06 的模型。
- 說明：若時間成本足夠，iters 拉高相信可以取得更好的模型。

5.1.7 Conclusion of Classification Analysis

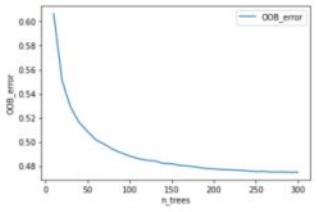
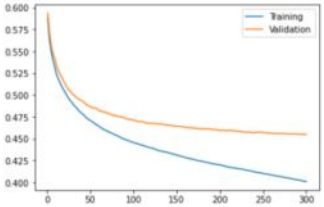
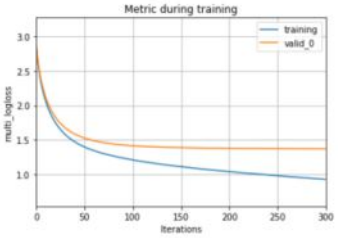
整體來說個別模型預測率都很低，可能與某些音樂本身界定上較模糊有關，且每首音樂皆可被劃分到多種音樂類型中，導致雖特徵相近但類別不同，最終模型難以分類。另外，就單一方法來對如此複雜的資料進行建模，或許較難建立良好的模型預測，因此接著將嘗試使用 ensembler, randomforest 等方法嘗試找出較好的模型。

6 Ensemble Methods

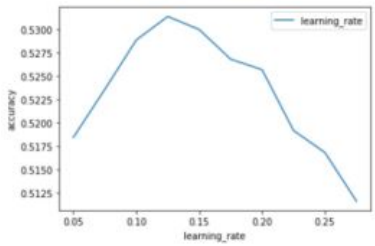
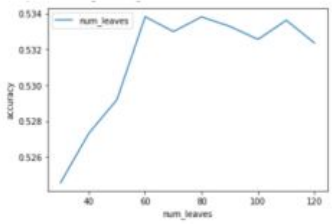
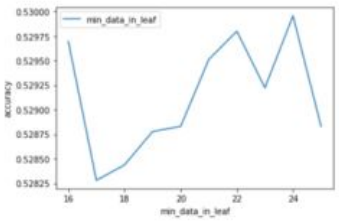
Ensemble Methods 代表的是一種協同合作的概念，透過不同的簡易分類器對於同一個樣本點進行預測，之後再共同投票決定預測結果。該方法不僅能夠大幅地降低較複雜的模型可能帶來的高 Variance 問題，同時也能夠降低簡易模型可能的 Bias 問題。本次研究我們將採用以下三種實務上常用的 Ensemble Methods:

1. Random Forest: 透過 Bagging 的方式建構決策樹，再透過多棵樹狀模型共同決策。
2. XGBoost: 一種結合 Bagging 和 Boosting 的演算法，能夠透過抽取重要特徵以及樣本的方式建構新的決策樹 (Bagging)，並且利用模型時刻迭代修正，建構出不同參數的決策樹模型 (Boosting)。
3. LGBM: 一種優化 XGBoost 的演算法，透過 Leaf-Wise 演算法來優化每棵決策樹的建構，以此大幅優化運算效率與模型建構。

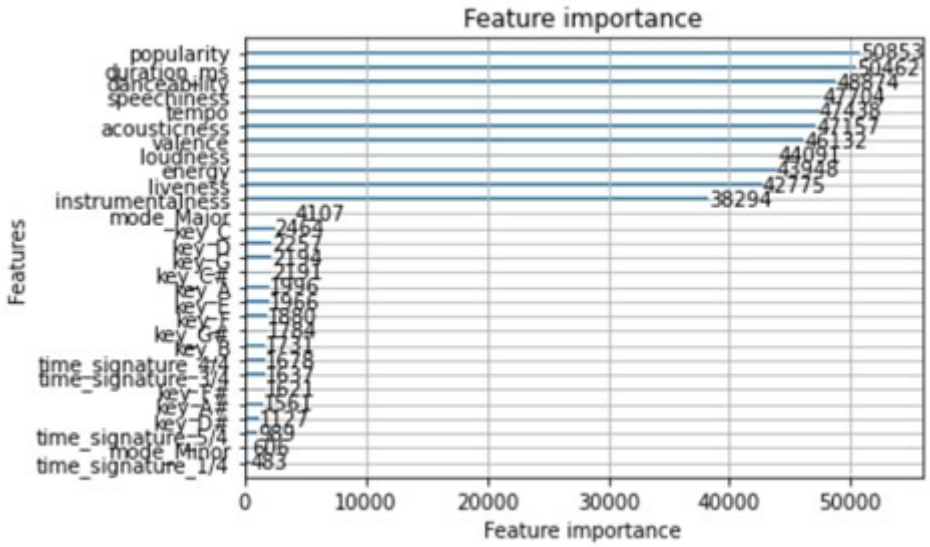
首先我們可以看到這組資料中的 Genre 總共有 26 種風格的樂曲，並且有些單一曲目會被歸類至多個類別，例如 Roy Woods 所演唱的 How I Feel 歌曲中融合了饒舌以及嘻哈的元素，並且以靈魂樂和 R&B 曲風進行編曲；亦或是像 Tove Lo 所演唱的 Habits 是一首具有 R&B 風格的舞曲流行樂，這些音樂通常融合了多種元素在裏頭，因此若單純使用一對一類別的 Ensemble 模型可能無法將其準確分類，因此我們將這些曲目自成一類新變數 (Others)，並且我們去掉樣本數過少 (<0.1%) 的 A'Capella 歌曲。接著我們將類別變數轉換唯獨熱編碼並對所有變數進行標準化處理丟入以上三種模型中，同時將所有資料五等分進行 Cross Validation 取平均準確率，結果如下 (損失函數圖形僅取其中一次):

Random Forest	XGBoost	LGBM
Accuracy: 0.52185	Accuracy: 0.53486	Accuracy: 0.54962
		
Run Time: 20min20sec	Run Time: 2h35min36sec	Run Time: 25min9sec

以隨機森林來說，我們利用 Out-of-bag 方式計算設定不同棵樹時的平均誤差 (Error)，大約在增加至 200 棵樹之後平均誤差就不再有顯著改善，而 XGBoost 和 LGBM 則是跌代次數 150 以後就逐步收斂了，從結果上來看，準確率 LGBM > XGBoost > RFM，然而，運行效率上 LGBM 遠優於 XGBoost，並且收斂的相當快速，礙於時間成本關係，本研究中我們僅用 LGBM 模型作為 Ensemble Method 的範例進行參數優化，以下為我們針對幾個會大幅影響模型表現的超參數進行優化，包含 learning_rate, num_leaves, min_data_in_leaf，結果如下:

learning_rate	num_leaves	min_data_in_leaf
		
Best parameter: 0.125	Best parameter: 80	Best parameter: 24

從結果來看 learning_rate 對於模型預測準度的影響較為顯著，其他參數的變動對於預測率改善相當有限，並且極有可能僅僅是抽樣誤差，因此我們僅針對 learning_rate 調整為 0.125。接著我們也可以看看各個變數對於我們模型的顯著性，觀察下表我們得知這組資料中所有類別變數對於我們預測的幫助較小，並且連續變數之間的影響其實並沒有差很多，這與我們進行 PCA 和 FA 時得到的結果差不多，所有主成分或是要素的參數組成大小並沒有非常明顯的差異。



透過 Tuning 所得到的結果以及變數顯著性結果，我們將參數重新設定並刪除所有的類別變數重新建模進行 5-folds-CV，如表 8 所示。

表 8: 透過 Tuning 所得到的結果以及變數顯著性結果

	precision	recall	f1-score		precision	recall	f1-score
Alternative	0.29	0.20	0.24	Movie	0.60	0.57	0.59
Anime	0.65	0.57	0.60	Opera	0.78	0.87	0.82
Blues	0.45	0.42	0.43	Others	0.45	0.75	0.57
Children's	0.63	0.44	0.52	Pop	0.29	0.06	0.10
Classical	0.66	0.67	0.67	R&B	0.27	0.18	0.22
Comedy	0.96	0.96	0.96	Rap	0.31	0.06	0.10
Country	0.44	0.49	0.46	Reggae	0.51	0.45	0.47
Dance	0.33	0.07	0.12	Reggaeton	0.59	0.62	0.61
Electronic	0.59	0.63	0.61	Rock	0.26	0.08	0.12
Folk	0.30	0.30	0.30	Ska	0.65	0.61	0.63
Hip-Hop	0.39	0.19	0.26	Soul	0.26	0.17	0.21
Indie	0.22	0.02	0.04	Soundtrack	0.67	0.82	0.74
Jazz	0.44	0.42	0.43	World	0.58	0.47	0.52
Best Overall Accuracy of Ensembles: 0.5522							

相較於原先的模型，最終的模型有了些許改善，最終的總體預測準確率為 0.5522，並且我們可以從該圖形明顯看出哪些類別容易預測，那些則較難抓到特定特徵。

7 結論

本研究為了探討不同類型的音樂的特性，首先先對資料進行預處理，接著對資料進行描述性統計，以了解資料的各項特徵。透過分群分析的手法將不同類型的音樂分開，接著使用主成份分析、對應分析、因素分析研究數據中的結構，最後透過所學的分類演算法以及 Ensemble 對資料進行實作分類，發現對於 Spotify 資料集來說，Ensemble 方法中的 LGBM 演算法預測效果比起其他分類演算法來的好，其預測率約為 55%。