What makes songs popular ?

Analysis on Spotify music

2020 1/15

adv. M.-S. Chen	adv. M.-L. Lo	stud.林修同	stud.吳彬睿	stud.蔣思陽	stud.何欣育
NTU EE	NTU EE	NTU EE	NTU CE	NTU EE	NTU DS
mschen@ntu.edu.tw	mllo2003@.ntu.edu.tw	stlin@arbor.ee.ntu.edu.tw	brwu@arbor.ee.ntu.edu.tw	syjiang@arbor.ee.ntu.edu.tw	hyho@arbor.ee.ntu.edu.tw

1. Introduction

在當今的時代, 互聯網的快速發展使得數據的獲取變得越來越容易, 熱愛聽音樂的人們也可以透過Youtube、Apple Music、Spotify等音樂串流平台隨時享受音樂的美妙。而電腦的計算能力的大幅度提升, 資料容易儲存的特質, 讓從大筆資料中尋找深層次特徵成為可能, 其總結出來的規律, 可以為用戶和開發者們提供更加清晰的指導意義。

近幾年來, Spotify於歐美地區起家開始興起, 透過他友善的使用者介面與穩定的音樂品質, 已漸漸普及至全世界, 成為世界上擁有龐大使用者的音樂串流平台之一。Spotify也熱衷於提供他所擁有的部分數據, 但會針對歌手、品牌、開發者給予不同層面的訊息, 因此能獲取到十分詳細的數據資料。從一個資料科學家可以拿到的資料來看, 除了歌名、歌手、發行日期的資訊以外, 資料訊息, 特別的是有量化後的音樂資訊, 比如歌曲的活力、舞動的程度、給人正向樂觀的感覺之程度等等, 都是非常重要的音樂屬性。與此同時, 根據Spotify內部演算法，其也會提供一個重要的指標來判斷歌曲熱門與否。

一般而言, 一首的音樂作品, 其受歡迎程度與否應該取決於大眾對其的評價。可「眾口難調」, 好的音樂作品也會由於推廣不利而被埋沒, 所以對於資料科學家們會而言, 更加關注音樂本身的特質, 比如音高, 時長等, 同時, 對於歌手等創作人而言, 若其能根據一些深層次受歡迎的規律和特性, 其會對未來的創作提供更好的幫助。

2. Data Description

透過Spotify Web API下載2000年至2019年的歌曲資訊共20萬筆，18個變數，各變數描述如(表2.1)所示。

(表2.1)變數基本敘述表

變數名稱	變數描述	資料類型	變數範圍
Song Name	歌曲名稱	Nominal	Lover, Home, Stay,....
Artist Name	歌手名字	Nominal	Taylor Swift,...
Release Date	發行日期	Nominal	2019-12-20, 2019-12-19...
Key	歌曲的平均音高	Nominal	0 = C, 1 = C#, -1 if no key
Mode	歌曲的模式(大調或小調)	Nominal	1 = Major, 0 = Minor
Genres	歌曲的種類	Nominal	C-pop, K-pop, hip-hop...
Tempo	歌曲的節奏(BPM)	Numerical	[0, 240]
Duration(ms)	歌曲長度(單位:毫秒)	Numerical	[7229, 4725264]
Danceability	歌曲適合跳舞的程度，根據節奏、拍子強度和整體規律性等音樂元素計算而來。	Numerical	[0.0, 1.0]
Acousticness	歌曲為原聲、純傳統樂器演奏版本的信賴程度。數字越大，較高的機率為原聲歌曲。	Numerical	[0.0, 1.0]
Energy	歌曲的活力、強度。	Numerical	[0.0, 1.0]
Liveness	歌曲的現場感，數字越大，代表歌曲為現場演出的機率較高。	Numerical	[0.0, 1.0]
Loudness	歌曲的音量大小(單位:dB)	Numerical	[-60.0, 2.0]
Speechiness	歌曲包含單詞的多寡程度，數字越小，則屬於純音樂的機率較大。	Numerical	[0.0, 1.0]
Valence	歌曲傳遞正向能量的程度，分數越高，歌曲聽起來較令人感到快樂，開朗。	Numerical	[0.0, 1.0]
Instrumentalness	歌曲無包含人聲的程度，“Ooh” 和“aah”會當作樂器，而其他單詞則會當作人聲。	Numerical	[0.0, 1.0]
Artist Popularity	歌手的熱門程度	Numerical	[0, 100]
Popularity	歌曲熱門程度	Numerical	[0, 100]

3. Data Preprocessing

3.1 檢查遺失值與重複資料

資料中有些樣本的資訊記載不完全, 並且有出現歌曲重複的現象, 因此將這些樣本皆予以刪除, 剩餘181034筆。

3.2 移除新歌與無法分辨歌手的歌曲

基於Spotify並不會即時更新歌曲熱門程度的數據, 而是隔一段時間才更新一次, 因此在我們下載數據的當下, 有1432筆的新歌是沒有歌曲熱門程度的數據, 所以移除這些樣本。而因為有些歌曲為多人演唱, 但在「歌手」這個屬性中, 會被記錄為Variety Artist, 也就無法分辨這首歌真正演唱者為何, 因此將這些樣本予以刪除, 剩餘176047筆。

3.3 檢查離群值

逐一針對數值型變數進行離群值檢查，發現「Duration」這個屬性的變異較大，有許多的離群值，因此移除這些離群值，剩餘166146筆。

3.4 標準化

從(表2.1) 可看到有些變數的數值都較其他變數大了許多(如:Duration和Tempo), 因此在後續分析, 針對數值型變數進行標準化(Normalization) , 會在兩個狀況下使用不同的標準化:

(1) 基本分析

使用Min-Max Normalization, $y = \frac{x-min}{max-min}$

(2) 找尋Important Factors和實作Clustering

使用Standard deviation Normalization, $y = \frac{x-\bar{x}}{std(x)}$

4. Methodology and Implementation

4.1 基本定義

流行歌曲:針對Popularity, 上四分位(75%)及以上的歌曲

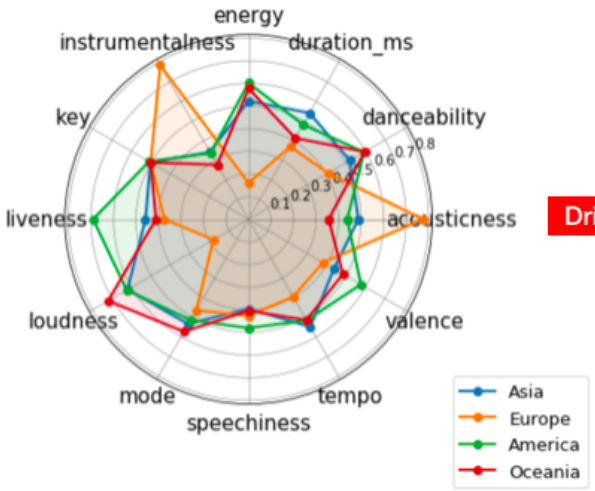
不流行歌曲:針對Popularity, 下四分位(25%)及以下的歌曲

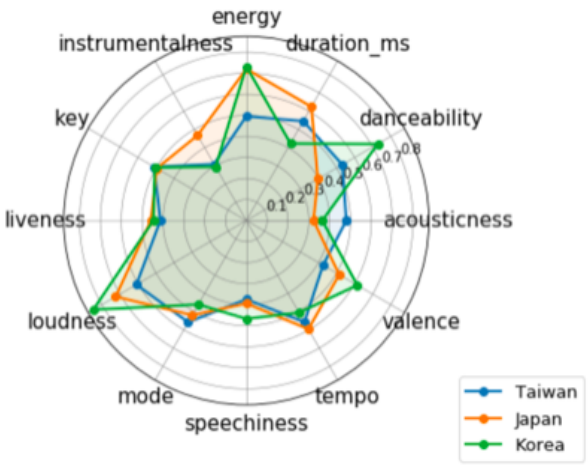
4.2 基本分析

我們針對不同的音樂屬性, 將資料標準化之後, 對資料進行一些基本分析。

4.2.1 雷達圖

我們首先針對地區資料進行統計分析, 畫出了以下 (圖4.1) 的的地區分佈的雷達圖。

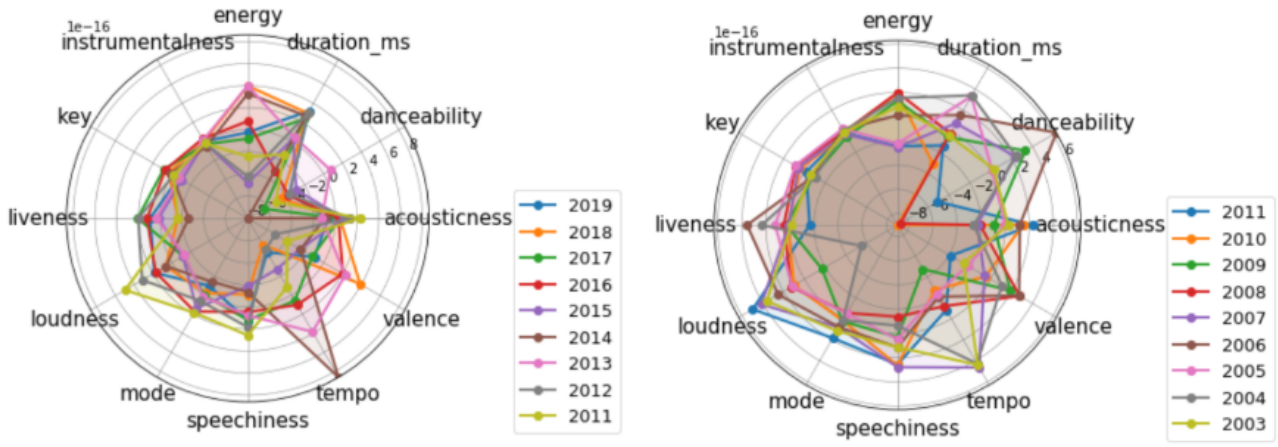




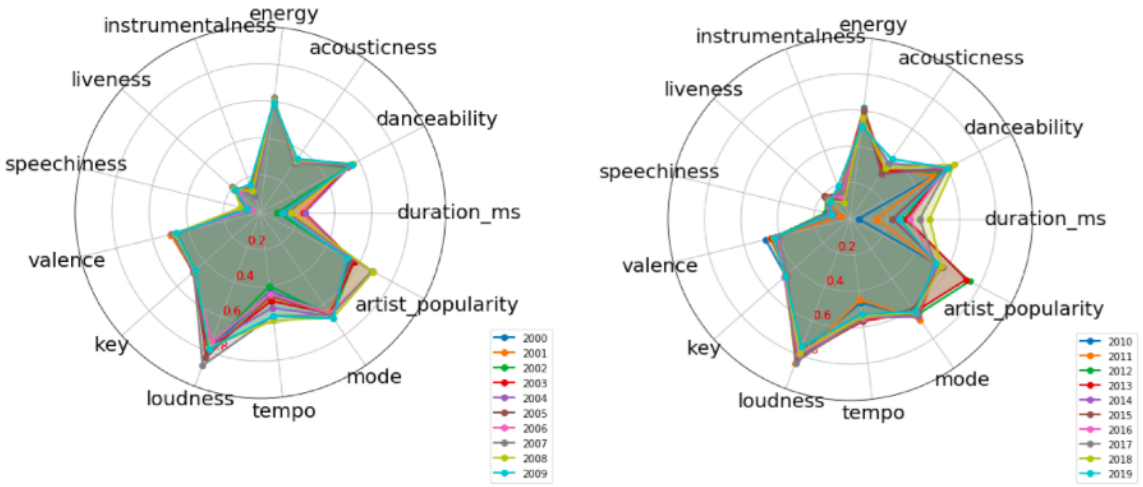
(圖4.1a) (各大洲的流行歌音樂屬性分布)　　(圖 4.1b) (亞洲的流行歌音樂屬性分布)

如(圖4.1a)所示, 畫出亞洲、歐洲、美洲和大洋洲四個地區的雷達圖。其中不難發現, 各個大洲之間音樂愛好大不相同, 這可能是由於文化、日常喜好的不同而導致的, 所以我們進一步將資料細分, 拿出亞洲的東亞部分, 分別畫出台灣、日本、韓國三個地區的雷達圖(圖4. 1b), 從中可以發現, 其差異性明顯小於不同大陸之間, 但是仍存在較大的不同。

其次針對年份資料進行統計分析, 畫出了以下 (圖4.2) 的的時間分佈的雷達圖。



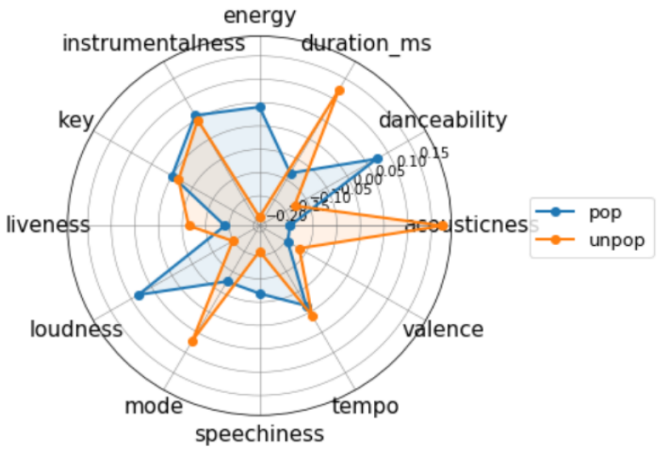
(圖4.2a) 每年音樂屬性分佈雷達圖



(圖4.2b) 每年流行音樂屬性分佈雷達圖

從(圖4.2a)不難看出每一年的出產的歌他們的流行的重點都會有所不同，如2009年說唱(speechness)類型的歌曲會比較流行、2008年節奏感(tempo)的音樂會比較流行。但是從(圖4.2b)觀察來看，流行歌曲的音樂屬性都十分相似，通常他們的歌曲時間長度都不會太長；現場感 (liveness) 不會太強，因為現場感的音樂相比如專輯而言，比較吵雜，聽者能會喜歡聲音相對乾淨的mv版本；最後如果想要成為一首流行歌，說唱(speechiness)類型的歌曲也許不是一個好選擇，從雷達圖分布看來，流行歌曲中說唱的佔比相較之下會低一點。

最後針對流行資料進行統計分析，畫出了以下 (圖4.3) 的的時間分佈的雷達圖。



(圖4.3) 2000-2019 熱門跟冷門歌曲的音樂屬性差異

(圖4.3)是 2000~2019年所有歌曲中的流行與不流行的資料比對。從這張雷達分布圖來看，我們可以很輕易地看出pop跟unpop的歌曲屬性差異，詳細差別如(表4.1)所示。

(表4.1) 流行歌曲與不流行歌曲對比表

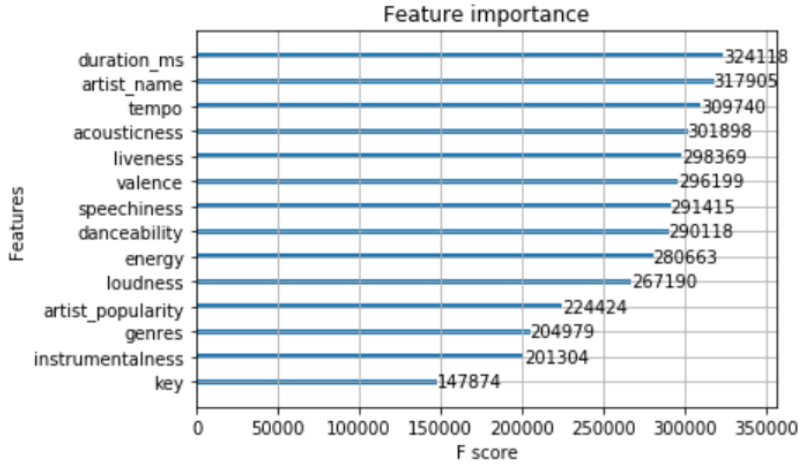
	POP	UNPOP
可舞蹈性 danceability	高	低
原聲性 acoustiness	低	高
正能量性 valence	稍低	稍高
節拍性 tempo	稍弱	稍強
說唱性 speechiness	高	低
大小調 mode	偏小調	偏大調
音量大小 loudness	大聲	小聲
現場感 liveness	低	高
音高 key	接近	
無人聲程度 instrumentalness	接近	
活力 energy	強	弱

4.3 Feature importance (XGBoost)

我們通過XGBoost來進行迴歸分析，從而希望找到影響音樂潮流的最主要的特徵。我們放入2000-2019的所有歌曲，將熱門程度（popularity）作為自變數，其他變數則作為依變數。

4.3.1 迴歸結果

RMSE: 9.629213 Baseline: 15.819152093671578



(圖4.4) XGBoost迴歸結果分析

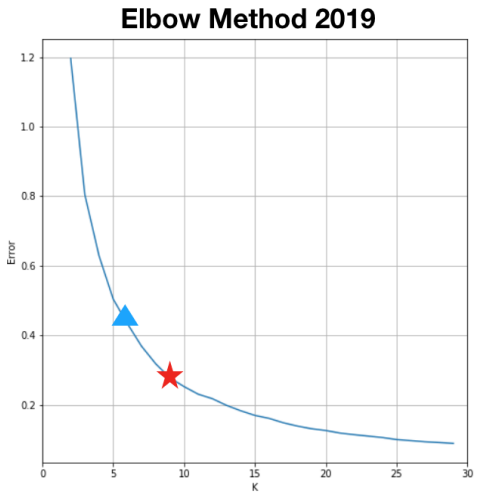
4.3.2 結果分析

從(圖4.4)不難看出, duration_ms、artist_name會是影響popularity的重要因素, 其中duration_ms也與我們從上述雷達分布圖所做的分析相呼應。然而歌曲的key屬性相對來說則不是那麼重要, 因為key對應到的差異可能是男女歌手的分別, 所以也符合客觀上的認知。

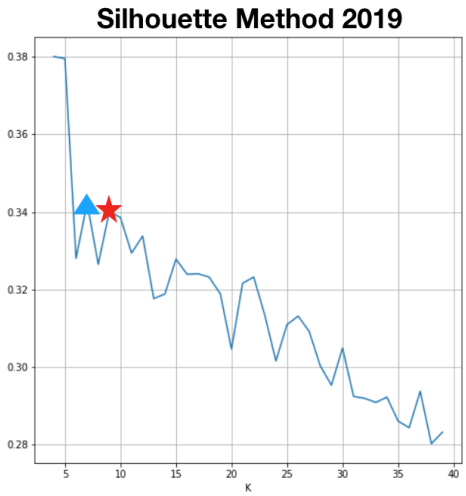
4.4 Clustering(K-means)

為了發現流行音樂之間, 是否可以透過音樂屬性來做分群, 進而找出歌曲符合哪種音樂屬性的分布可以有機會成為流行歌曲, 我們對每一年的所有流行歌曲透過音樂屬性當作向量來做分群, 找出那年流行的音樂屬性分布, 並且透過Elbow Method 和 Silhouette Method找出適當的分群群數(1~30), 使得群內變異數最小(透過 Elbow Method 計算)、群間變異數最大(透過 Silhouette Method 計算)。

在以下例子中, 我們針對2019年的流行音樂進行分群, 透過以上方法可以分別得到下面兩張圖(圖4.5a和圖4.5b):



(圖4.5a) Elbow Method



(圖4.5b) Silhouette Method

從圖中我們可以看出 k = 9 的時候, 雖然在 Silhouette Method中, k = 7時(群間變異數值相對較高, 且群內變異數相對較低) 是比較好的選擇, 但是k = 9在 Elbow Method中群內變異數所減少的幅度比較大, 而且k = 9 跟k = 7在 Silhouette Method中所計算出的群間變異數其實差異不大, 所以最後我們選擇 k = 9 來當作我們分群的群數。

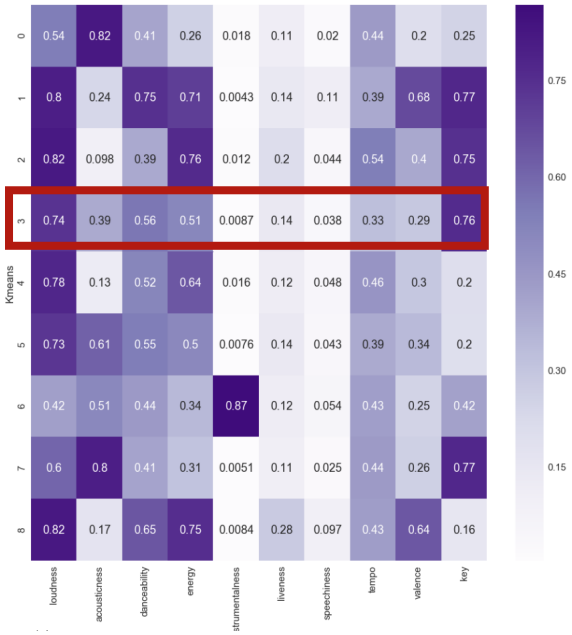
4.4.1 推薦系統

得到分群結果後, 我們希望能透過比對歌曲(可能是非流行音樂或者不是分群該年的歌曲)與欲推薦年份中每個流行音樂分群之間的關係, 推薦最相似分群中與歌曲具有相似音樂屬性分布的流行歌。

以下範例為我們隨機從2018年選出一首歌(圖4.6a)，並且透過其音樂屬性分布，找出它和2019年的流行音樂分群中最相似的群(圖4.6b)，從群中再找出與它最相似的前15首流行歌作為推薦結果。

不染 (電視劇《香蜜沉沉燼如霜》主題曲)

loudness	0.750248
acousticness	0.502049
danceability	0.348613
energy	0.45815
instrumentalness	0
liveness	0.0804355
speechiness	0.0199081
tempo	0.504866
valence	0.329091
key	0.636364



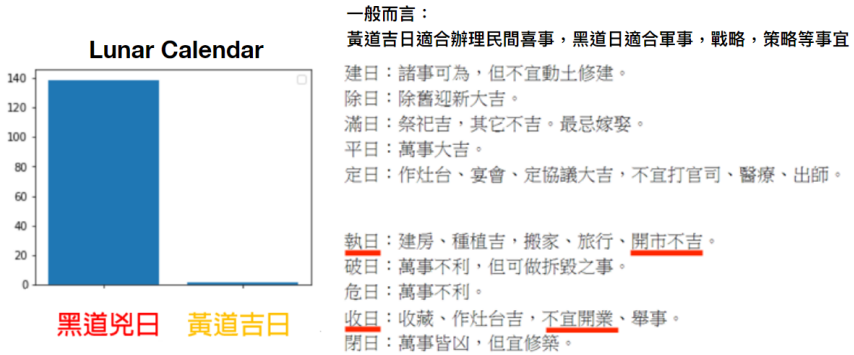
(圖4.6a)歌曲屬性　　(圖4.6b)2019年的流行音樂分群

從最後推薦歌曲的音樂屬性分布，可以看出推薦的歌曲與所選的歌曲確實有較高的相似度，而推薦歌曲之間確實也跟所屬分群的音樂屬性高度相似。推薦結果與其音樂屬性分布分別如(圖4.7a)和(圖4.7b)所示。

32	將故事寫成我們
779	水深火熱
499	暫時愛著我
550	時間停止
409	不敢
478	清醒 (戲劇《淺情人不知》片尾曲)
57	我比從前想你了 - 電視劇《我們不能是朋友》片尾曲
98	寂寞考 - 大人中演唱會 LIVE
39	囚
752	給親愛的你
434	平行時空
268	別再叫我哥
136	初戀
303	失憶鎮
755	Hold On

(圖4.7a) 推薦的15首歌曲　　(圖4.7b)推薦的15首歌曲之分群資訊

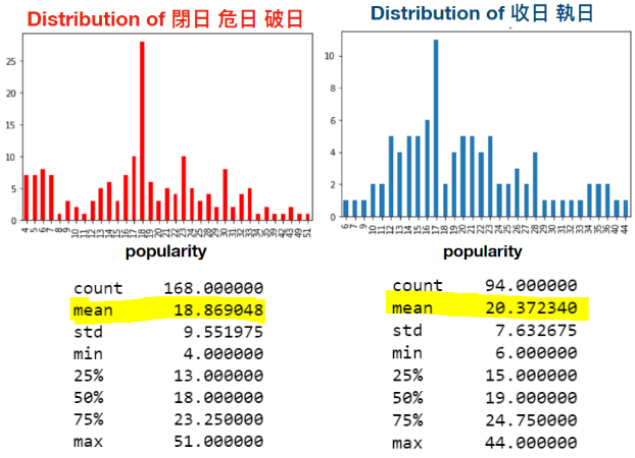
4.5 迷信分析



(圖4.8) 黑道兇日和黃道吉日的區別

(圖4.8) 是我們把2019台澎金馬發行的歌曲的日去拿去統計, 我們發現歌手在發行專輯時, 他們都會挑準黑道兇日來發行他們的專輯, 相比之下在黃道吉日發行的專輯數量是非常少的, 因此我們稍微地去研究了一下黑道兇日跟黃道吉日的差別, 原來在我們認知中的黃道吉日對於歌手來說並不是這們一回事，相較於適合辦理民間喜事的黃道吉日，是和戰略陰謀的黑道兇日是更適合歌手們去發行專輯打入市場這片戰場的。

接著我們進一步去分析黑道兇日裡面的建日。在黑道兇日中有執日、收日這兩個日子市開是不吉, 不適合開業的, 因此我們把所有在執日、收日的歌曲分為一類, 剩下的分成一類, 接著再去對他們最後的流行程度popularity去做分析, 到底會不會因為日子不好導致歌曲比較不火紅呢？



(圖4.9) 熱門程度分布圖

如(圖4.9), 我們把圖畫出來後, 接著對他們做簡單的數值分析，我們發現這兩者基本上是沒有差異的, 他們popularity的distribution相似, 而mean也是很接近, 反倒讓人跌破眼鏡的是在這些在不適開業發行的歌曲平均起來popularity翻而還高了1%, 因此我們可以認定這只是迷信而已, 這些好日子壞日子, 不會是一個重要的因素。

5. Conclusion

經過我們的分析，可以得知各個地區音樂類型差異很大，即使是地緣關係相近，風格依舊迥異。透過Regression，我們發現歌曲長度、歌手姓名會影響歌曲是否popular，而歌曲的key、instrumentalness 較無相關。此外我們對歌曲長度做進一步分析，透過計算流行歌曲與非流行歌duration_ms的均值，我們推薦如果要成為popular的歌曲，歌曲時間長度不要超過四分鐘會是比較好的選擇。我們利用Clustering(K-Means)，來得到華語地區每年度流行音樂的特徵分布，進而做到推薦歌曲的功能。最後我們也發現台灣藝人偏好在「黑道凶日」發歌，但發行日期並不會影響其歡迎程度。

6. Reference

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

[2] Spotify's "This Is" playlists: the ultimate song analysis for 50 mainstream artists https://www.freecodecamp.org/news/spotifys-this-is-playlists-the-ultimate-song-analysis-for-50-mainstream-artists-491882081819/?fbclid=IwAR0m6g9g4HrPOS-i3x1-d1sVtN-KX3e0NP7LLuqVugZUCi6UjqS64zdrMYs

[3]Determining The Optimal Number Of Clusters: 3 Must Know Methods https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

[4]黃道吉日，吳亦凡Antares、王源一样同步发行，网友：你们不一样 https://sh.qihoo.com/mob/transcoding?url=9c5dd67703274a9b7&cota=1&sign=360_e39369d1

[5]日曆 月曆 年曆 黃道吉日 黃曆 農民曆 http://www.bestday123.com/