國立成功大學
National Cheng Kung University

大數據分析期末報告

音樂串流平台使用習慣分析

學生:黃郁喬、陳玟樺、周培倫 撰

指導教授:黃韻勳

中華民國 111 年 8 月

目錄

第一章 緒論

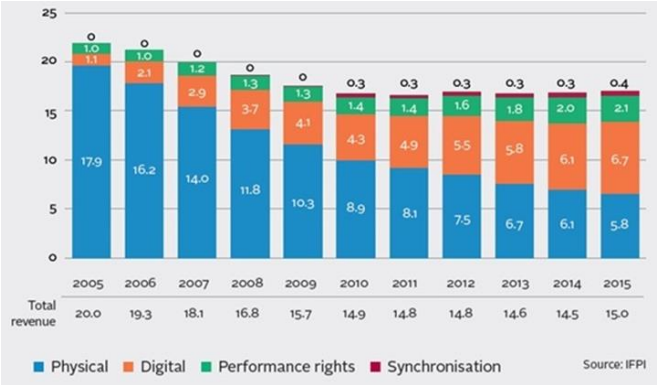
1.1 前言

聽音樂始終是現代人的重要日常活動之一，不管是搭乘大眾運輸工具時、在家裡放鬆做家務時、晚上睡前或是洗澡時，任何時間點都可以隨時來首悅耳的音樂，對大學生族群來說更是生活中不可或缺的一部分。

根據網路調查，大部分民眾每天平均花費 1.7 小時在聽音樂，以一日 16 小時的活動來說，每日約佔了 1 成的時間，可見聽音樂已成為各世代在家中休息時的消遣活動（約 5～6 成）。其中，年輕世代較常在通勤或學校、公司的休息時間聽音樂，20 世代聽音樂的時間最長（平均 2.2 小時／天）；成熟世代則多在工作時或家中休息、做家事時享受音樂。

以另一方面來看，在 IFPI（International Federation of the Phonographic Industry，國際唱片業協會）的 2016 全球音樂報告中可以看到 2015 年實體唱片業利潤僅剩 5.8 億美元，而數位音樂利潤則是逐年上升，達到 6.7 億美元的水準。2015 為數位音樂產值首次超過實體音樂的一年，其中串流服務的快速成長，更是營收增加的重要驅力。

這些說明了音樂對生活影響的重要性，也是這份報告所研究的主要內容，以 Python 語言做分析研究報告。



IFPI 國際唱片業協會(2016)

https://www.ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2016

1.2 研究目的

本研究我們將所蒐集的大量數據進行主題式的分析以及討論，目的探討音樂和生活的相關內容。

主要以將這些大數據經過一系列的流程萃取出潛在的重要資訊，將所收集的資訊進行基礎的前置資料處理，進行進階的方式做分析討論，如；關連規則、約略集合理論、決策樹分析、主成分分析、集群分析、迴歸分析、貝氏分類及類神經網路等，並透過目前主流大數據處理分析工具的 Python 程式語言設計大數據分析模型。

第二章 研究方法

2.1 資料前處理

資料是資料科學中的基石，沒有好的資料，就難以產生好的資料價值，然而，在真實的世界中，資料往往沒有想像中的「乾淨」。在實務中，資料會有資料缺失（Imcomplete/Missing data）、雜訊（Noise）、離異值等等的問題。

資料前處理泛指的是在分析演算法之前，先對資料進行處理，讓資料在格式上比較標準一致，為的是讓演算法不會因為資料產生的瑕疵而錯誤判斷。

2.1.1 不正確的資料

確認資料的有效範圍及驗證資料的合理性。

2.2 資料視覺化

資料視覺化（Data visualization）被許多學科視為與視覺傳達含義相同的現代概念。它涉及到資料的視覺化表示的建立和研究。

為了清晰有效地傳遞資訊，資料視覺化使用統計圖形、圖表、資訊圖表和其他工具。可以使用點、線或條對數字資料進行編碼，以便在視覺上傳達定量資訊，而有效的視覺化可以幫助使用者分析和推理資料和證據，它使複雜的資料更容易理解和使用。使用者可能有特定的分析任務，以及該任務要遵循的圖形設計原則。

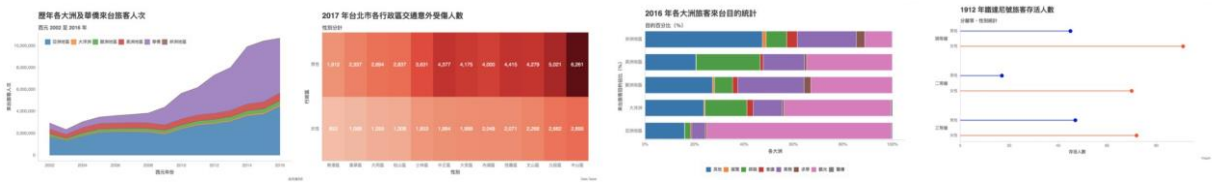


Fig.1

2.2.1 長條圖

條形圖（bar chart），或條圖（bar graph），常稱為長條圖，又稱為柱狀圖、棒形圖，是一種以長方形的長度為變量的統計圖表。長條圖用來比較兩個或以上的價值，只有一個變量，通常利用於較小的數據集分析。長條圖亦可橫向排列，或用多維方式表達。

繪製長條圖時，長條柱或柱組中線須對齊項目刻度。相較之下，折線圖則是將數據代表之點對齊項目刻度。在數字大且接近時，兩者皆可使用波浪形省略符號，以擴大表現數據間的差距，增強理解和清晰度。

2.2.2 散佈圖

散布圖可以用來表示實驗中的連續自變數和另一個連續應變數之間的關係，也可以用來表示二個連續自變數之間的關係。若系統中存在參數，在實驗中會刻意增加或減少其數值，此參數即為自變數，若是自變數和應變數的散布圖，一般會將自變數放在橫軸，應變數放在縱軸。若兩個參數都是自變數，可將任一個放在橫軸，此時，散布圖可以看出其相關性的程度。

散布圖可以推測二個參數中許多不同種類的相關性，配合一定的信賴區間。以體重及身高為例，可能會將體重放在 y 軸，將身高放在 x 軸。相關性可能是正相關、負相關、無相關性。若散布圖有從左下到右上分布的圖形，表示兩者正相關，若散布圖有從左上到右下分布的圖形，表示兩者負相關。為了研究兩參數之間的關係，可以在散布圖上繪製擬合線。趨勢線的方程式就是參數相關性的方程式。若是線性相依，繪製最適曲線的程序即為線性迴歸，保證在有限時間內有正確的解。針對任意的相關性關係，不存在通用、可以產生正確解的最適曲線產生程序。若是要確認兩組參數之間是否有非線性的關係，也可以用散布圖來觀察。可以在散布圖中加上平滑曲線（Local regression）來達到此一機能。若數據可以表示為簡單關係的混合模型表示，其關係在視覺上上會是以疊加模式來表示。

2.2.3 圓餅圖

圓餅圖，或稱餅狀圖，是一個劃分為幾個扇形的圓形統計圖表，用於描述量、頻率或百分比之間的相對關係。在圓餅圖中，每個扇區的弧長大小為其所表示的數量的比例。這些扇區合在一起剛好是一個完全的圓形。顧名思義，這些扇區拼成了一個切開的餅形圖案。

在一些特定情況下，圓餅圖可以很有效地對訊息進行展示。特別是在想要表示某個大扇區在整體中所占比例，而不是對不同扇區進行比較時，這一方法十分有效。圓餅圖在扇區所占比例達到母體的 25%或 50%時，可以很好地達到展示的目的。但通常，可能更多情況會採用其它圖表如條形圖或圓點圖，或非圖表的方法如表格來表達訊息。

2.3 回歸模型

回歸模型(regression model)對統計關系進行定量描述的一種數學模型。

迴歸分析係建立一個或多個自變數(或稱解釋變數) 對某一個應變數(或稱被解釋變數)關係模式。

2.3.1 簡單回歸(Simple regression)

描述一個自變數對一個應變數的關係 ，Y=β0 +β1X1，母體迴歸線可代表兩變數間的線性關係，由於母體迴歸線無法得知，而改以樣本迴歸線 yˆi = β0 + β1xi 估計母體迴歸線。

2.3.2 線性回歸(Linear Regression)

線性回歸 (Linear regression) 是統計上在找多個自變數(independent variable)和依變數(dependent variable)之間的關係建出來的模型。只有一個自變數和一個依變數的情形稱為簡單線性回歸(Simple linear regression)，大於一個自變數的情形稱為多元回歸(multiple regression)。一般回歸分析的介紹都會以簡單線性回歸為例子來說明，在此文章兩種我都會說明和公式推導。

2.3.2.1 均方根誤差(Mean square error)

除判定係數(determinant of coefficient)外，在機器學習領域中，最常被用於判別迴歸模型好壞之方法，即計算均方誤差， 均方誤差在計算「預測值與實際值間差異的均方值」，均方值就是先平方在取平均，公式可表示如下:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

2.4 羅吉斯回歸

感知器演算法(Perceptron LearningAlgorithm) 能夠成功進行二元分類(Binary classification)，但只能知道分類結果為 A 類或是 B 類，卻無法得知區分為 AT 類及 B 類的機率為何？

羅吉斯迴歸類似線性迴歸分析，主要在探討依變數與自變數之間的關係。線性迴歸中的依變數(Y)通常為連續型變數，但羅吉斯迴歸所探討的依變數(Y)主要為類別變數，特別是分成兩類的變數，並可得知分為此兩類的機率分別為多少。

2.5 混淆矩陣

在機器學習領域和統計分類問題中，混淆矩陣（confusion matrix）是可視化工具，特別用於監督學習，在無監督學習一般叫做匹配矩陣。矩陣的每一列代表一個類的實例預測，而每一行表示一個實際的類的實例。

在預測分析中，混淆矩陣是具有兩行兩列的表，該表報告假陽性，假陰性，真陽性和真陰性的數量。這不僅可以進行正確分類的分析，還可以進行更詳細的分析。對於分類器的真實性能，準確性不是可靠的指標，因為如果數據集不平衡，它將產生誤導性結果。

2.6 決策樹分析

決策論中，決策樹（Decision tree）由一個決策圖和可能的結果組成，用來創建到達目標的規劃。決策樹建立並用來輔助決策，是一個利用像樹一樣的圖形或決策模型的決策支持工具，包括隨機事件結果，資源代價和實用性。它是一個算法顯示的方法。決策樹的另一個使用是作為計算條件概率的描述性手段。

機器學習中，決策樹是一個預測模型；他代表的是對象屬性與對象值之間的一種映射關係。樹中每個節點表示某個對象，而每個分叉路徑則代表某個可能的屬性值，而每個葉節點則對應從根節點到該葉節點所經歷的路徑所表示的對象的值。決策樹僅有單一輸出，若欲有複數輸出，可以建立獨立的決策樹以處理不同輸出。數據挖掘中決策樹是一種經常要用到的技術，可以用於分析數據，同樣也可以用來作預測。

一個決策樹包含三種類型的節點：

決策節點：通常用矩形框來表示

機會節點：通常用圓圈來表示

終結點：通常用三角形來表示

2.6 群集分析

在各種領域的研究中,若面臨到必須將所看到的資料分成幾個「有意義」的組時,則群集分析則是提供這樣「分類」的工作;主要目標是依照收集的 p 個變數(X~X)將 n 個個體分幾個群,使群內個體間離的近(或相似),不同群的個體離的遠(或差異性大)。(群集分析方法也可以對變數做分群)。一般群集分析的分類法有「聯合分群法」(Joining)(或稱樹形分群法)、「雙向聯結法」(Two-WayJoining)或稱集區分群法)、「K 組平均數分群法』(K-Means Clustering)。

第三章 資料來源與處理

3.1 資料匯入

將 google 表單收集的資料，轉成 excel 檔案並匯入 python。

```
In [4]: # 匯入需要的模組

        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt

        %matplotlib inline
        plt.rcParams['font.family'] = 'SimHei'

In [5]: # 匯入資料集
        import pandas as pd
        df = pd.read_excel('music_bigdata__new_excel.xlsx')
        df
```

Out[5]:

	性別	年齡(歲)	每月可支配所得(區間)	每月平均所得(實際)	你是否有使用過音樂串流平台(包含免費試用)?	你最常使用哪種音樂串流平台?	其他	你每個月預估花多少錢在音樂平台上面？	你選擇音樂串流平台時，可接受的最高價錢範圍為何?	你最常聽下列哪一種語言的音樂?	...	當你使用音樂串流平台時，其人性介面對你的重要程度為何?	當你使用音樂串流平台時，其購買價格高低對你的重要程度為何?	當你使用音樂串流平台時，其歌曲風格對你的重要程度為何?	當你使用音樂串流平台時，其排行榜功能對你來說重要程度為何?	當你使用音樂串流平台時，其推薦歌曲功能對你來說重要程度為何?	當你使用音樂串流平台時，其有無MV對你來說重要程度為何?	當你使用音樂串流平台時，其有無歌曲demo(尚未正式發行的歌曲樣本)對你來說重要程度為何?	當你使用音樂串流平台時，其音樂應用程式能否在別的裝置上同步對你來說重要程度為何?	我不認真回答這份問卷?	備註
0	生理女	17	3,000 – 6,000	3500	是	KKBOX	NaN	50	0–100/月	華語	...	3	4	3	3	4	2.0	2.0	2	否	NaN
1	生理女	20	3,000 – 6,000	4500	是	YOUTUBE MUSIC	NaN	80	0–100/月	華語	...	4	4	2	3	4	2.0	1.0	1	否	NaN
2	生理女	38	$9,000以上	40000	是	YOUTUBE MUSIC	NaN	180	150–200/月	日語	...	4	4	3	3	3	2.0	3.0	3	否	NaN
3	生理男	29	$9,000以上	35000	是	SPOTIFY	NaN	150	150–200/月	西洋	...	3	3	4	2	3	1.0	3.0	4	否	NaN
4	生理男	24	6,000 – 9,000	8500	是	SPOTIFY	NaN	130	100–150/月	華語	...	4	4	4	2	2	1.0	2.0	3	否	NaN

Fig. 2

3.2 資料前處理

3.2.1 資料清理

我們的表單中有詢問填寫者，是否有認真作答，因此我們可以透過此欄位的回答結果，有效的去除不認真回答的結果，另外表單中有一個欄位詢問是否有使用音樂平台的習慣，若他的回答為否、沒有使用過音樂平台，則對此份報告研究分析內容不具任何意義，故將資料刪除。

```
In [4]: # 資料前處理---篩掉不認真的回答、沒有使用過音樂平台
df = df[df['我不認真回答這份問卷?'] == '否']
df = df[df['你是否有使用過音樂串流平台(包含免費試用)?'] == '是']
df
```

Out[4]:

性別	年齡(歲)	每月可支配所得(區間)	每月平均所得(實際)	你是否有使用過音樂串流平台(包含免費試用)?	你最常使用哪種音樂串流平台?	其他	你每個月預估花多少錢在音樂平台上面？	你選擇音樂串流平台時，可接受的最高價錢範圍為何?	你最常聽下列哪一種語言的音樂?	...	當你使用音樂串流平台時，其人性介面對你的重要程度為何?	當你使用音樂串流平台時，其購買價格高低對你的重要程度為何?	當你使用音樂串流平台時，其歌曲風格對你的重要程度為何?	當你使用音樂串流平台時，其排行榜功能對你來說重要程度為何?	當你使用音樂串流平台時，其推薦歌曲功能對你來說重要程度為何?	當你使用音樂串流平台時，其有無MV對你來說重要程度為何?	當你使用音樂串流平台時，其有無歌曲demo(尚未正式發行的歌曲樣本)對你來說重要程度為何?	當使音串平時其樂厓式否別裝上步你說要度
0	女	17	3,000–6,000	3500	是	KKBOX	NaN	50	0–100/月	華語	...	3	3	3	3	3	3	3
1	女	20	3,000–6,000	4500	是	YOUTUBE MUSIC	NaN	80	0–100/月	華語	...	3	3	3	3	3	3	3
2	女	38	$9,000以上	40000	是	YOUTUBE MUSIC	NaN	180	150–200/月	日語	...	3	3	3	3	3	3	3
3	女	29	$9,000以上	35000	是	SPOTIFY	NaN	150	150–200/月	西洋	...	3	3	3	3	3	3	3
4	女	24	6,000–9,000	8500	是	SPOTIFY	NaN	130	100–150/月	華語	...	3	3	3	3	3	3	3
5	女	22	6,000–9,000	7000	是	APPLE MUSIC	NaN	75	0–100/月	日語	...	3	3	3	3	3	3	3
6	女	23	6,000–9,000	8300	是	SPOTIFY	NaN	165	0–100/月	華語	...	3	3	3	3	3	3	3
7	女	21	$9,000以上	23000	是	YOUTUBE MUSIC	NaN	250	200–250/月	華語	...	3	3	3	3	3	3	3
8	女	27	$9,000以上	33000	是	KKBOX	NaN	175	150–200/月	華語	...	3	3	3	3	3	3	3
9	女	30	$9,000以上	52000	是	YOUTUBE MUSIC	NaN	270	$300以上/月	華語	...	3	3	3	3	3	3	3
10	女	18	6,000–9,000	6000	是	KKBOX	NaN	99	0–100/月	華語	...	3	3	3	3	3	3	3

Fig. 3

3.2.2 資料整合

我們將表單中，想分析的資料擷取出來，並整合成我們想要的數據。

```
In [5]: # 資料前處理-選取要分析的欄位數據
df = df[['年齡(歲)', '每月可支配所得(區間)', '每月平均所得(實際)', '你最常使用哪種音樂串流平台?', '你選擇音樂
df
```

Out[5]:

	年齡(歲)	每月可支配所得(區間)	每月平均所得(實際)	你最常使用哪種音樂串流平台?	你選擇音樂串流平台時，可接受的最高價錢範圍為何?	你每個月預估花多少錢在音樂平台上面？	你最常聽下列哪一種語言的音樂?
0	17	3,000–6,000	3500	KKBOX	0–100/月	50	華語
1	20	3,000–6,000	4500	YOUTUBE MUSIC	0–100/月	80	華語
2	38	$9,000以上	40000	YOUTUBE MUSIC	150–200/月	180	日語
3	29	$9,000以上	35000	SPOTIFY	150–200/月	150	西洋
4	24	6,000–9,000	8500	SPOTIFY	100–150/月	130	華語
5	22	6,000–9,000	7000	APPLE MUSIC	0–100/月	75	日語
6	23	6,000–9,000	8300	SPOTIFY	0–100/月	165	華語
7	21	$9,000以上	23000	YOUTUBE MUSIC	200–250/月	250	華語
8	27	$9,000以上	33000	KKBOX	150–200/月	175	華語
9	30	$9,000以上	52000	YOUTUBE MUSIC	$300以上/月	270	華語
10	18	6,000–9,000	6000	KKBOX	0–100/月	99	華語

Fig. 4

3.3 資料視覺化

3.3.1 長條圖

我們將每月可支配所得以及使用平台可接受最高價的範圍，作長條圖的分析，目的在了解大家在使用平台的接受範圍。

```
# 長條圖
x = df.loc[:, '每月可支配所得(區間)']
y = df.loc[:, '你選擇音樂串流平台時，可接受的最高價錢範圍為何?']
plt.bar(x,y,align='center')
plt.figsize=(6,4)
plt.xlabel('每月可支配所得', fontsize = 12)
plt.ylabel('可接受音樂平台的最高價錢範圍', fontsize = 12)
plt.title('每月支配所得與音樂平台花費關係', fontsize = 18)
# plt.legend(loc = "upper left")
plt.show()
```

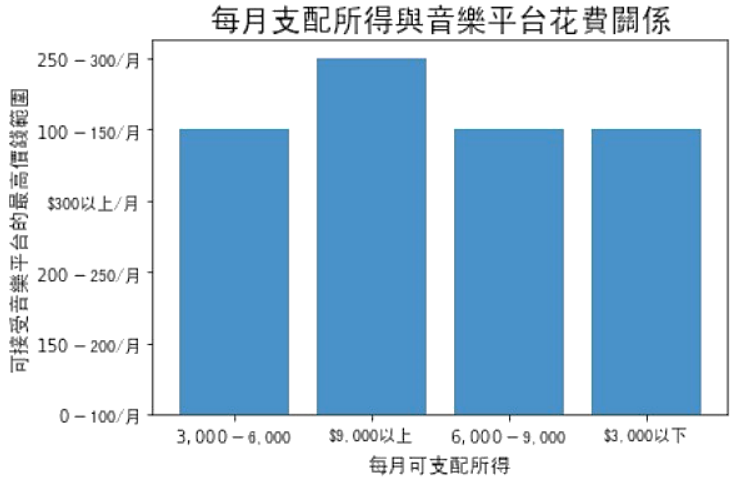


Fig.5

3.3.2 散佈圖

此次研究，我們將年齡作為 X 軸，每個月的平台花費作為 Y 軸，如圖 Fig.5，目的在了解各個年齡層對平台使用的花費分布情形，幫助我們了解數據大致的走向，以利我們做後續分析。同理，我們以年齡作為 X 軸，每個月的平均所得作為 Y 軸，如圖 Fig.6，目的在了解所的與年齡的相關性，根據 Fig.5、6 可發現，兩者皆呈正相關。

```
# 資料視覺化
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# 進行散佈圖分析(或相關係數分析)，確認變數間的關係
df.corr()

%matplotlib inline

plt.rcParams['font.family']='SimHei' #顯示中文

# 年齡(歲) & 你每個月預估花多少錢在音樂平台上面？  散佈圖
plt.rcParams['axes.unicode_minus']=False
df.plot(kind='scatter',title='散佈圖',figsize=(6,4),x='年齡(歲)',y='你每個月預估花多少錢在音樂平台上面？',marker='+')

#_____________________________________________________________________
# 年齡(歲) & 每月支配所得(實際) 散佈圖
plt.rcParams['axes.unicode_minus']=False
df.plot(kind='scatter',title='散佈圖',figsize=(6,4),x='年齡(歲)',y='每月平均所得(實際)',marker='+')

#_____________________
# 你每個月預估花多少錢在音樂平台上面？ & 你最常使用哪種音樂串流平台? 散佈圖
plt.rcParams['axes.unicode_minus']=False
df.plot(kind='scatter',title='散佈圖',figsize=(6,4),x='你每個月預估花多少錢在音樂平台上面？',y='你最常使用哪種音樂串流平台?',marke


# 年紀 & 音樂平台
plt.rcParams['axes.unicode_minus']=False
df.plot(kind='scatter',title='散佈圖',figsize=(6,4),x='年齡(歲)',y='你最常使用哪種音樂串流平台?',marker='+')
```

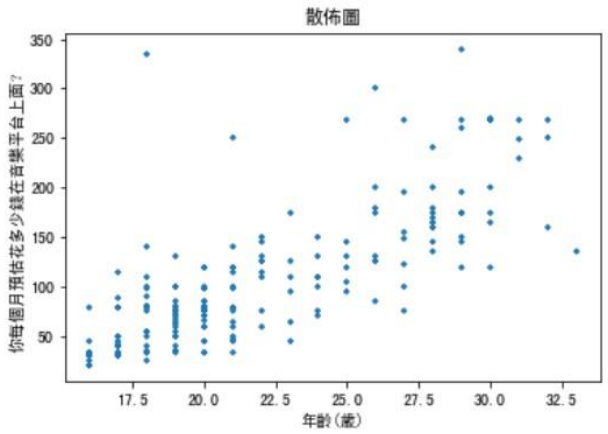


Fig.6

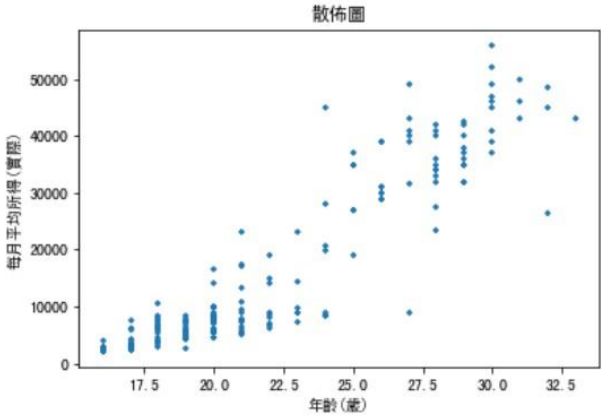


Fig.7

3.3.3 圓餅圖

我們將各種音樂平台的使用數據視覺化，目的在了解哪一種音樂平台目前最普及，將資料畫成圓餅圖可看出 KKbox、Spotify、YouTube 為目前最多人使用的音樂平台，因此也會是此篇研究的重點方向。

```
# 圓餅圖 音樂平台佔有率
data_platform = df.loc[:, "你最常使用哪種音樂串流平台?"]
data_kk = data_platform[data_platform == 'KKBOX'].count()
data_sp = data_platform[data_platform == 'SPOTIFY'].count()
data_youtube = data_platform[data_platform == 'YOUTUBE MUSIC'].count()
data_apple = data_platform[data_platform == 'APPLE MUSIC'].count()
data_others = data_platform[data_platform == '其他'].count()

data_circle_all_labels = [data_kk, data_sp, data_youtube, data_apple, data_others]
labels = ['KKBOX','SPOTIFY','YOUTUBE', 'APPLE MUSIC','其他']
plt.pie(data_circle_all_labels, labels=labels, autopct='%.1f%%')
plt.show()
```

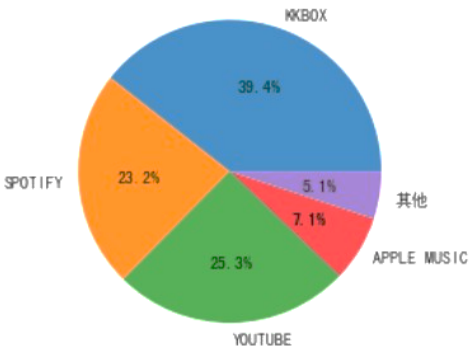


Fig.8

第四章 主要研究結果與討論

4.1 回歸分析

4.1.1 簡單線性回歸

為了探討年齡和每個月在平台上的花費，我們利用簡單的線性回歸來作初步的分析， X 係數值為 12.11，可知回歸函數為正相關的線性函數，R 平方表示在所有自變數下，解釋應變數的變異百分比或其預測解釋能力，可以用來代表線性回歸模式的適合度，我們所得的值是 0.6 左右，也就是說此線性回歸模型的適合度佳，可利用此方式進行分析。

```
# 簡單線性迴歸 1 年齡(歲) 你每個月預估花多少錢在音樂平台上面？
from sklearn.model_selection import train_test_split
X = df[['年齡(歲)']]
y = df[['你每個月預估花多少錢在音樂平台上面？']]

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.2, random_state=85)
X_train
```

```
# 簡單線性回歸
import matplotlib.pyplot as plt
import numpy as np
from sklearn import linear_model # 回歸模式套件
plt.style.use('ggplot')

# 設定 linear regression 物件
regr = linear_model.LinearRegression()

# 以 training datasets 的資料建立線性迴歸模型
regr.fit(X_train, y_train)

# 計算 R2
r_squared = regr.score(X_train, y_train)

print('截距項', regr.intercept_)
print('X係數', regr.coef_)
print('R平方:',r_squared)
```

```
截距項 [-157.39799181]
X係數 [[12.11085596]]
R平方: 0.6292589129684851
```

Fig. 9

4.1.2 均方根誤差

均方根誤差（RMSE）是迴歸預測模型的兩種主要效能指標之一，可衡量預測值和實際值之間的平均差異，藉此估計預測模型預測目標值的準確度。我們所得的值為 49.98，因此可以知道回歸方程預估的平均花費和我們的數據中大家預估的花費落差 49.98 元，此落差值可知此回歸方程的預測在合理範圍內。

```
# rmse(均方根誤差) - 衡量迴歸方程式好壞
rmse = np.sqrt(((regr.predict(X_test) - y_test) ** 2). mean ( )).round(4)
print("均方根誤差 (root-mean-square error, RMSE) = " , rmse )
```

```
均方根誤差 (root-mean-square error, RMSE) =  你每個月預估花多少錢在音樂平台上面？    49.9856
dtype: float64
```

Fig.10

4.1.3 平均絕對誤差

絕對平均誤差 MAE，誤差的絕對值取平均，也就是誤差離真實值的平均距離，對離群值沒那麼敏感，而我們所得的值為 26.17，可知我們所預測的誤差值和大家的預估花費約誤差 26.17 元。

```
# mae
mae = (abs(regr.predict(X_test) - y_test).mean()).round(4)
print("平均絕對誤差(MAE) = ", mae)
```

```
平均絕對誤差(MAE) =  你每個月預估花多少錢在音樂平台上面？    26.172
dtype: float64
```

Fig.11

4.2 羅吉斯回歸分析

羅吉斯迴歸是一種統計模型，用於判斷事件發生的機率，如圖 Fig.12。它可顯示不同特徵之間的關係，然後計算某個結果的發生機率。它和線性迴歸模型很相似，但是適合二分因變數的模型。我們將資料先進行轉換，並進行了年齡和每月所得的相關分析，根據 Fig.14 我們所繪出的羅吉斯函數圖型，將我們的資料作二元分類，可以知道年齡越大、收入越高的人越傾向使用 Spotify；同理，年齡越小、收入越低的人越傾向使用 KKbox。

並且，我們將預測結果和實際數據逐一比對，可以發現只有兩組數據判斷錯誤，如圖 Fig.13，因此可以知道我們的羅吉斯模型非常適合此數據的分析。

```
# 羅吉斯回歸 變數: 語言 年齡 應變數: 平台
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
data_reg = df
data_reg = data_reg[data_reg['你最常使用哪種音樂串流平台?'].isin(['KKBOX', 'SPOTIFY'])]
target_class = data_reg['你最常使用哪種音樂串流平台?'].replace({"KKBOX":0, "SPOTIFY":1})
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
x = pd.DataFrame(data_reg[['年齡(歲)', '每月平均所得(實際)']])
y = target_class
data_reg= pd.concat([x,y], axis=1)
data_reg
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
def sigmoid(z):
    return 1.0 / (1.0 + np.exp(-z))
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
z = np.arange(-7, 7, 0.1)
phi_z = sigmoid(z)

plt.plot(z, phi_z)
plt.axvline(0.0, color='k')
plt.ylim(-0.1, 1.1)
plt.xlabel('z')
plt.ylabel('$\phi (z)$')

plt.yticks([0.0, 0.5, 1.0])
ax = plt.gca()
ax.yaxis.grid(True)
plt.show()
```

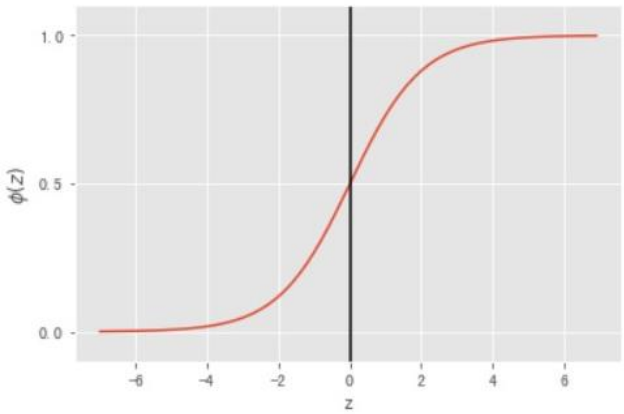


Fig.12

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(data_reg[['年齡(歲)', '每月平均所得(實際)']], data_reg[['你最常使用哪種音樂串
X_train
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train_std,y_train['你最常使用哪種音樂串流平台?'].values) #y_train為一個 Dataframe 無法直接到 fit 作運算，所以要把值取出來
```

```
LogisticRegression()
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
lr.predict(X_test_std)
```

```
array([0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0], dtype=int64)
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
y_test['你最常使用哪種音樂串流平台?'].values
```

```
array([0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0], dtype=int64)
```

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
#有兩個參數, i控制實際值, v控制羅吉斯迴歸模型分類值, 逐一比對!
error = 0
for i, v in enumerate(lr.predict(X_test_std)):
    if v!= y_test['你最常使用哪種音樂串流平台?'].values[i]:
        error+=1
print(error)
```

```
2
```

Fig. 13

```
# 羅吉斯回歸 變數: 所得 年齡 應變數: 平台(kk sp)
from matplotlib.colors import ListedColormap

def plot_decision_regions(X, y, classifier, resolution=0.02):

    # setup marker generator and color map
    markers = ('s', 'x', 'o', '^', 'v')
    colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
    cmap = ListedColormap(colors[:len(np.unique(y))]) #對於一維陣列，unique函式去除其中重複的元素，並按元素由小到大排列

    # plot the decision surface
    x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1  #取出 "年齡(歲)" 的最大值與最小值
    x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1  #取出 "每月平均所得(實際)" 的最大值與最小值
    xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution),  #使用座標向量創造出座標矩陣，矩陣長度為x1_max-x1_min
                           np.arange(x2_min, x2_max, resolution))  #矩陣寬度為x2_max-x2_min
    Z = classifier.predict(np.array([xx1.ravel(), xx2.ravel()]).T) #ravel函數主要創造一個一維的row vector (1Xn矩陣)
    Z = Z.reshape(xx1.shape)                                       #將兩個1Xn矩陣合併為1Xn後，在進行轉置，之後每個row就代表一個座
    plt.contourf(xx1, xx2, Z, alpha=0.4, cmap=cmap)                #Contourf用於繪製等高線圖
    plt.xlim(xx1.min(), xx1.max())
    plt.ylim(xx2.min(), xx2.max())

    for idx, cl in enumerate(np.unique(y)): #若只有兩類，會回傳[0,1]
        plt.scatter(x=X[y == cl, 0],        #逐一確認哪些點為0，哪些點為1
                    y=X[y == cl, 1],
                    alpha=0.6,
                    c=cmap(idx),      #將第一種顏色指定給該類別下所有的點，第二種顏色指定給該類別下的點
                    edgecolor='black',
                    marker=markers[idx],
                    label=cl)
```

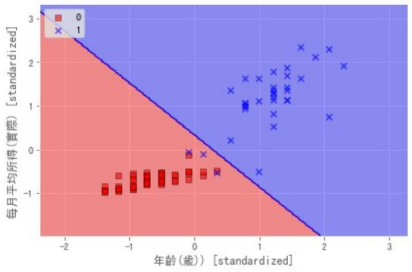


Fig. 14

4.3 混淆矩陣

通過這個矩陣可以方便地看出機器是否將兩個不同的類別混淆，根據我們所做的混淆矩陣分析，可以看出僅有兩筆資料預測錯誤，誤把 KKbox 判斷為 Spotify，而剩下資料則判斷正確，因此可以得到我們的模型判斷正確率為 0.92，如圖 Fig.15，總結此模型的判斷正確率很高，適用於此數據分析。

```
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0, probability=True)
classifier.fit(X_train, y_train)
classifier.predict(X_test)
```

```
array([1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0,
       0, 0, 1], dtype=int64)
```

```
error = 0

for i, v in enumerate(classifier.predict(X_test)):
    if v!= y_test.values[i]:
        error+=1
print(error)
```

```
2
```

```
# 顯示出混淆矩陣
from sklearn.metrics import confusion_matrix
y_pred = classifier.predict(X_test)
confusion_matrix(y_test, y_pred)
```

```
array([[14,  0],
       [ 2,  9]], dtype=int64)
```

```
X = df_confusion.iloc[:, 0:2].values
y = df_confusion.iloc[:, 2].values
```

```
sns.lmplot('年齡(歲)','每月平均所得(實際)',data=df_confusion, fit_reg=False, hue ='你最常使用哪種音樂串流平台?')
```

```
# 分類結果的正確率
from sklearn.metrics import accuracy_score
y_pred=classifier.predict(X_test)
accuracy_score(y_test, y_pred)
```

```
0.92
```

Fig.15

4.4 決策樹分析

我們的決策數分析第一層分類使用年齡，可以知道月收入低於 14250 的人傾向使用 KKbox；大於月收入高於 14250 的人傾向使用 Spotify，第二層分類我們使用每月實際平均所得，因為我們使用二元分析，故第二層分類無意義，如圖 Fig.16。

```
# 決策樹
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.datasets import load_breast_cancer
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import pandas as pd
import numpy as np
from sklearn import tree
```

```
df_tree = df[["年齡(歲)", "每月平均所得(實際)", "你最常使用哪種音樂串流平台?"]]
# 資篩選 轉換 平台(kkbox : 0, spotify: 1))
df_tree = df_tree[df_tree['你最常使用哪種音樂串流平台?'].isin(['KKBOX', 'SPOTIFY'])]
target_class = df_tree['你最常使用哪種音樂串流平台?'].replace({"KKBOX": 0, "SPOTIFY": 1})
df_tree = df_tree.drop(columns=["你最常使用哪種音樂串流平台?"])
df_tree.insert(2, column = "KKBOX OR SPOTIFY", value = target_class)
df_tree
```

```
fn=['年齡(歲)','每月平均所得(實際)']
cn=['KKBOX', 'SPOTIFY']
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (4,4), dpi=400)
tree.plot_tree(clf,
               feature_names = fn,
               class_names=cn,
               filled = True);
fig.savefig('imagename.png')
```

```
clf = DecisionTreeClassifier(max_depth = 1,
                             random_state = 0)
clf.fit(X_train, Y_train)

# clf.predict(X_test)
```

```
DecisionTreeClassifier(max_depth=1, random_state=0)
```

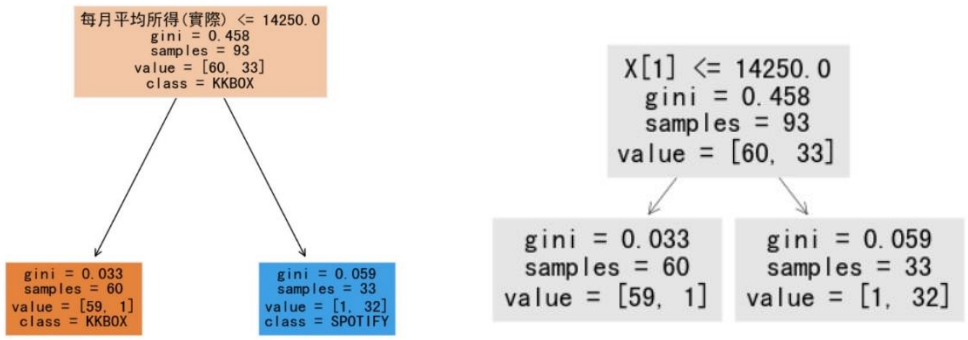


Fig.16

4.5 群集分析

我們先篩選出特徵資料，分別為年齡以及每月實際收入，並將資料標準化，根據群集分析的結果顯示可以看出，我們分析的結果分為兩個類別，分別為 Spotify、KKbox，而兩個族群的質心可以看出，KKbox 的年齡和實際收入相較於 Spotify 的資料都比較低。

```
# 取出特徵資料
features = df[["年齡(歲)", "每月平均所得(實際)"]]
features
```

```
# 以最小-最大的標準化方法，分別針對 年齡(歲) 與 每月平均所得(實際) 進行標準化。
from sklearn.preprocessing import MinMaxScaler
min_max = MinMaxScaler()
MinMax_df = df_cluster
MinMax_df.iloc[:,0:2] = min_max.fit_transform(MinMax_df.iloc[:,0:2])
MinMax_df
```

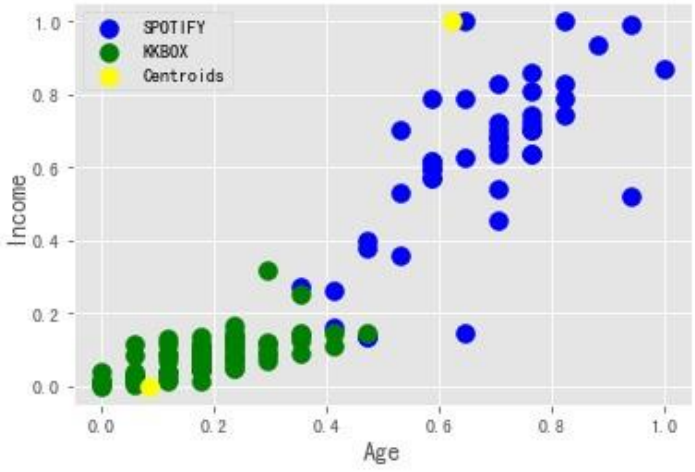
```
# 使用 KMean法 針對最小-最大的標準化後的資料進行分群，
# 同樣設定群集數目為 2、random_state 為 0
from sklearn.cluster import KMeans
cluster = KMeans(n_clusters=2, random_state=0)
MMdf = MinMax_df.iloc[:,[1,2]]
model = cluster.fit(MMdf)
MinMax_df['cluster'] = model.labels_
```

```
y_kmeans = cluster.fit_predict(MMdf)

plt.scatter(MinMax_df[MinMax_df["cluster"]==1]["年齡(歲)"], MinMax_df[MinMax_df["cluster"]==1]["每月平均所得(實際)"], s = 100,
plt.scatter(MinMax_df[MinMax_df["cluster"]==0]["年齡(歲)"], MinMax_df[MinMax_df["cluster"]==0]["每月平均所得(實際)"], s = 100,
plt.scatter(cluster.cluster_centers_[:, 0],  cluster.cluster_centers_[:, 1], s = 100, c = 'yellow', label = 'Centroids')
plt.xlabel('Age' , fontsize = 14)
plt.ylabel('Income', fontsize = 14)
plt.legend()
plt.show()
```

	年齡(歲)	每月平均所得(實際)
0	17	3500
1	20	4500
2	28	40000
3	29	35000
4	29	36000
...	...	...
199	20	5400
200	26	30000
201	28	32000
202	16	2500
203	30	56000

198 rows × 2 columns



第五章 結論與未來研究方向

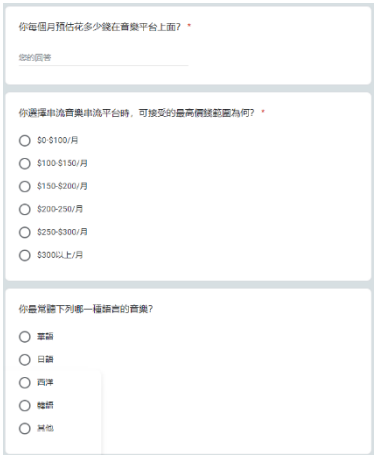
首先我們將蒐集的資料做基礎的處理，並將資料做視覺化，而得到的圓餅圖，我們判斷以市占率較高的 KKbox、Spotify 為此篇研究的重點方向。

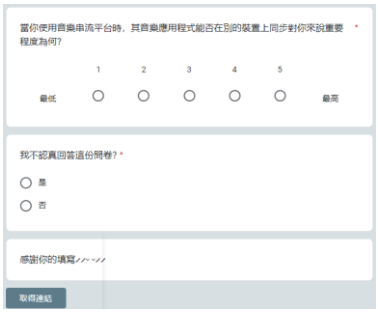
我們將處理過的資料用不同模型做各項分析，回歸模型、簡單線性回歸、羅吉斯回歸、混淆矩陣、決策樹分析、群集分析……等。透過回歸模型，我們可以知道所預測的資料為正相關，而且各項數值可以顯示，線性回歸模型的適合度佳。羅吉斯回歸則幫助我們將資料作二元分類，可以知道年齡越大、收入越高的人越傾向使用的平台得到簡單的小結論。為了判斷我們羅吉斯回歸的正確性，我們使用了混淆距陣來確認我們判讀錯誤的資料數，以及此模型判讀的正確率，我們得到高正確率，可以知道我們使用的模型，適用於我們蒐集到的資料判讀。最後我們以決策數分析以及群集分析做為此篇研究的收尾，決策樹分析所得的圖，可以讓我們得出明確的分類標準，年齡大於 22.5 歲的人，傾向於與用 KKBOX，最後透過群集分析可以看出兩個平台的使用者其年齡跟每月實際收入的分布狀態，將資料明確的分為兩類。

為來研究方向會以更多平台的綜合分析為主，增加復回歸的分析、更多元的分類來判讀決策數和更多特徵資料的群集分析。分析結果將更加貼近所蒐集到的數據以及實際裝況。

第六章 附件

6.1 問卷調查表單





第七章 參考資料

https://magazine.feg.com.tw/magazine/tw/magazine_detail.aspx?id=11770

https://zh.wikipedia.org/zhtw/%E6%95%B0%E6%8D%AE%E5%8F%AF%E8%A7%86%E5%8C%96

https://zh.wikipedia.org/zh-tw/%E6%9D%A1%E5%BD%A2%E5%9B%BE

https://zh.wikipedia.org/zh-tw/%E9%A5%BC%E5%9B%BE

https://zh.m.wikipedia.org/zhtw/%E6%8A%98%E7%B7%9A%E5%9C%96

https://zh.m.wikipedia.org/zhtw/%E6%95%A3%E5%B8%83%E5%9B%BE

https://ithelp.ithome.com.tw/articles/10186204

https://zh.wikipedia.org/zh-tw/%E5%86%B3%E7%AD%96%E6%A0%91

https://zh.m.wikipedia.org/zhhant/%E6%B7%B7%E6%B7%86%E7%9F%A9%E9%98%B5

https://www.yongxi-stat.com/logistic-regression/

https://moodle.ncku.edu.tw/pluginfile.php/991888/mod_resource/content/2/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%20-%20%E7%BE%85%E5%90%89%E6%96%AF%E8%BF%B4%E6%AD%B8.pdf

https://moodle.ncku.edu.tw/pluginfile.php/991888/mod_resource/content/2/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%20-%20%E7%BE%85%E5%90%89%E6%96%AF%E8%BF%B4%E6%AD%B8.pdf

https://moodle.ncku.edu.tw/pluginfile.php/984313/mod_resource/content/1/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%20-%20%E6%84%9F%E7%9F%A5%E5%99%A8%E6%BC%94%E7%AE%97%E6%B3%95.pdf