

# 大數據期末報告

## — 音樂串流平台使用習慣分析 —

組員：黃郁喬、陳玟樺、周培倫



# 目錄

## CONTENTS

**01** • 研究目的

**02** • 研究方法

**03** • 資料來源與處理

**04** • 研究結果與討論

**05** • 結論與未來研究方向

## PART 01

# 研究目的

*Research purposes*



# 研究目的

- 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
- 在IFPI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
- 數位音樂利潤則是逐年上升，達到6.7億美元的水準
- 將所蒐集的大量數據分析及討論，目的是分析音樂平台背後使用的族群及其使用習慣，以利廣告的投放及拓展客群
- 主要以『年齡』、『所得』、『每月平台花費』來分析使用族群及其消費習慣

# 資料來源（Google 表單）

---

- 年齡（歲）
- 每月平均所得（實際）
- 每個月預估花多少錢在音樂平台上面
- 選擇串流音樂串流平台時，可接受的最高價錢範圍為何？
- 最常聽下列哪一種語言的音樂？
- 廣告資訊來源

# 廣告資訊來源統計

- KKBOX

使用kkbox的族群其廣告資訊來源大多自電信推薦

你最常使用哪種音樂串流平台? 廣告資訊來源		
count	78	78
unique	1	3
top	KKBOX	電信推薦
freq	78	53

- SPOTIFY

使用spotify的族群，廣告資訊來源大多來自朋友推薦

你最常使用哪種音樂串流平台? 廣告資訊來源		
count	46	46
unique	1	2
top	SPOTIFY	朋友推薦
freq	46	36

## PART 02

# 研究方法

*Research methods*



# 資料前處理

不正確的資料

遺漏值  
(Missing Data)



# 資料視覺化

長條圖  
(Bar Chart)

散佈圖  
(Scatter Diagram)

圓餅圖  
(Pie Chart)

# 回歸模型(Regression model)

簡單回歸  
(Simple regression)

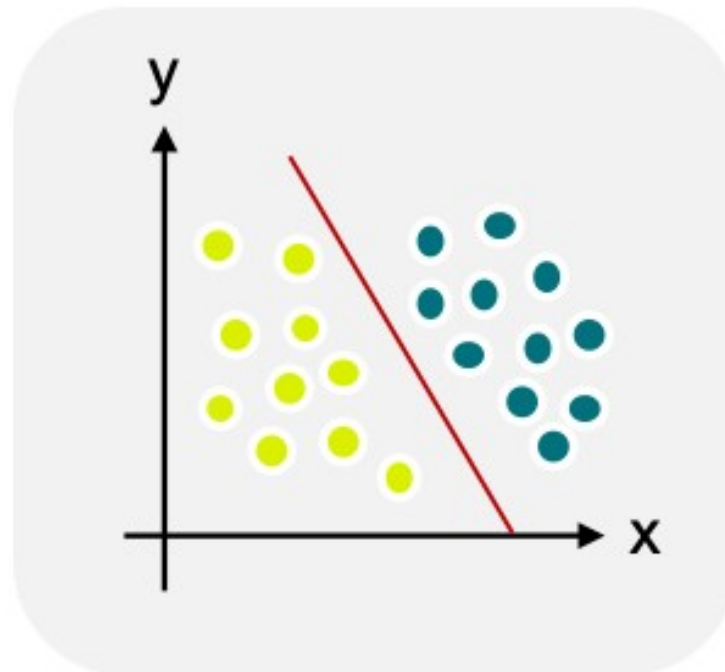
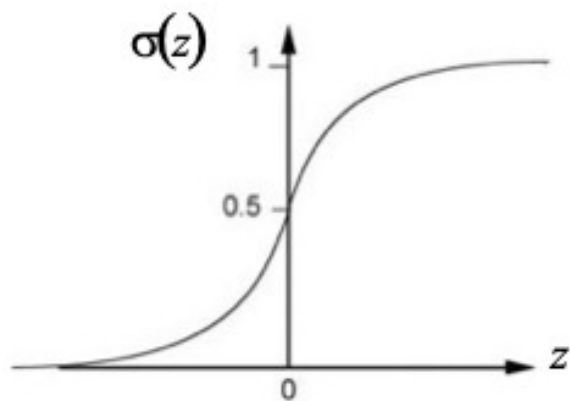
線性回歸  
(Linear Regression)

→ 均方根誤差(Mean square error)

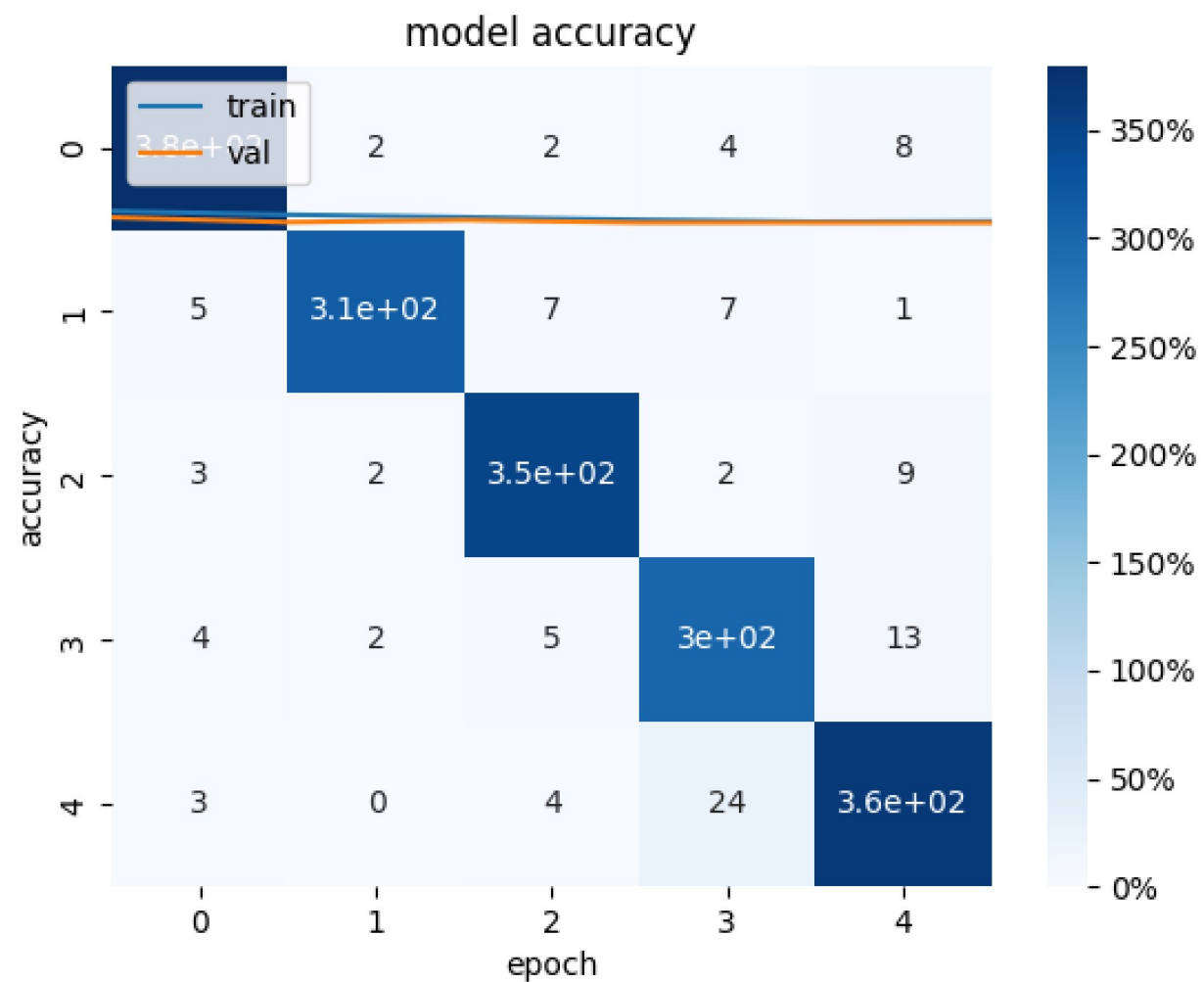
# 羅吉斯回歸

## (Logistic Regression Model)

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

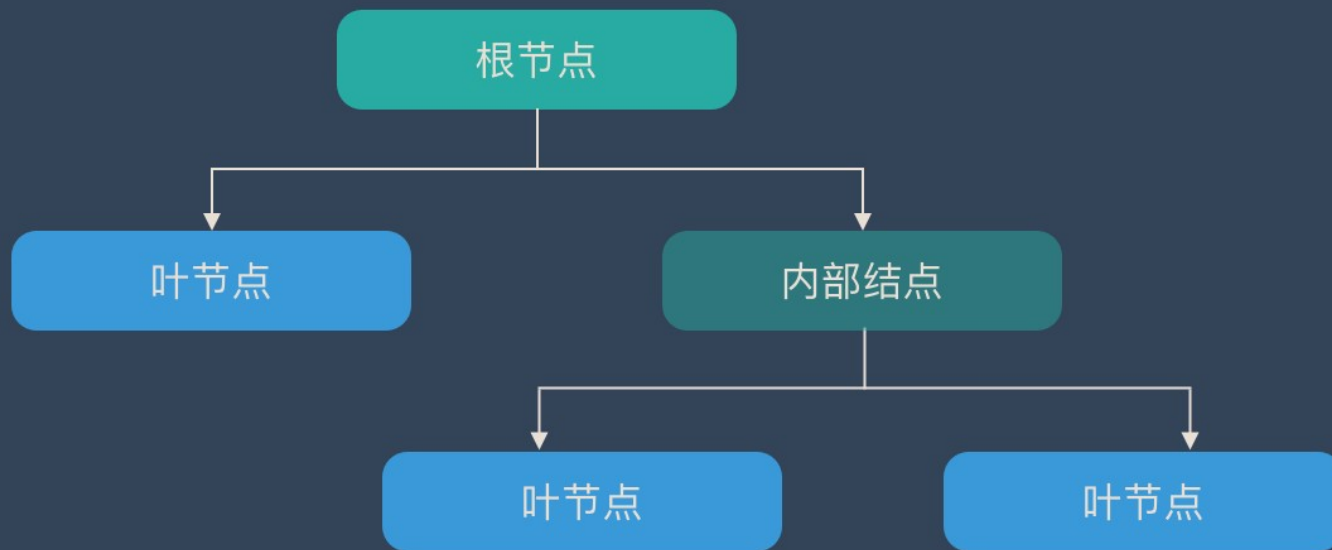


# 混淆矩陣



# 决策树分析(Decision tree)

## 决策树的结构



## PART 03

# 資料來源與處理

*Data source and processing*



# 資料前處理 — 資料清理

- 表單中有詢問填寫者是否有認真作答
- 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。

# 資料前處理 — 資料整合

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
```

4

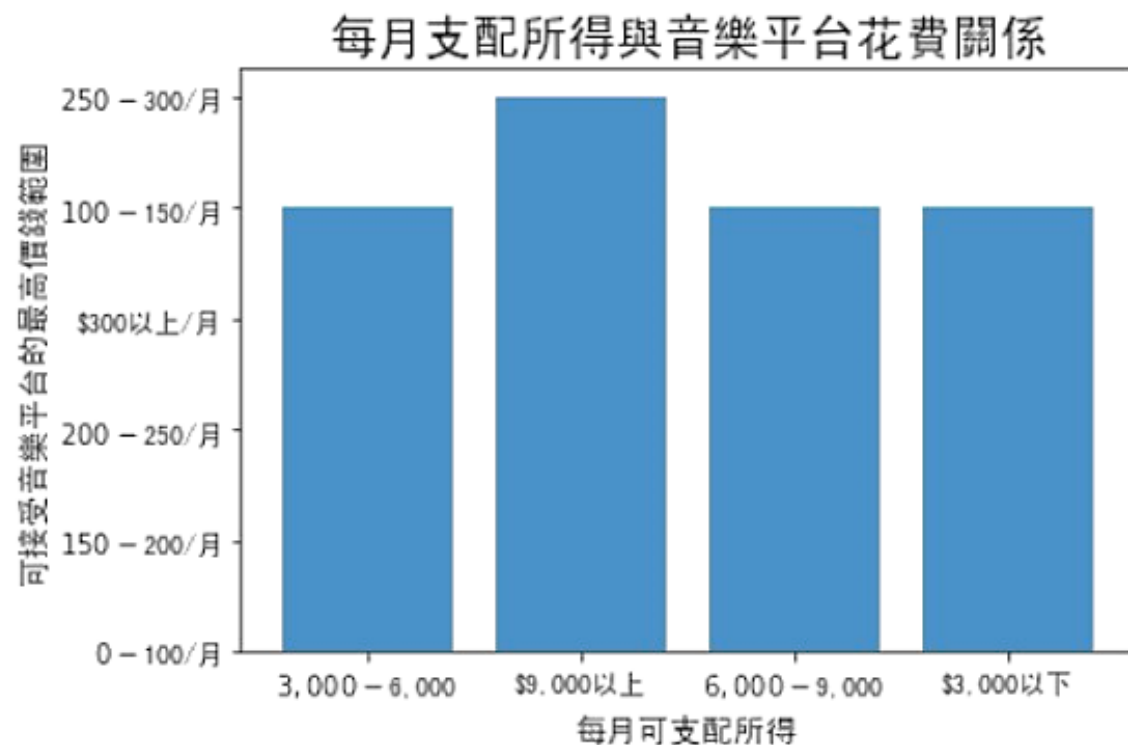
Out[5]:

	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 在音樂平台上呢?	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



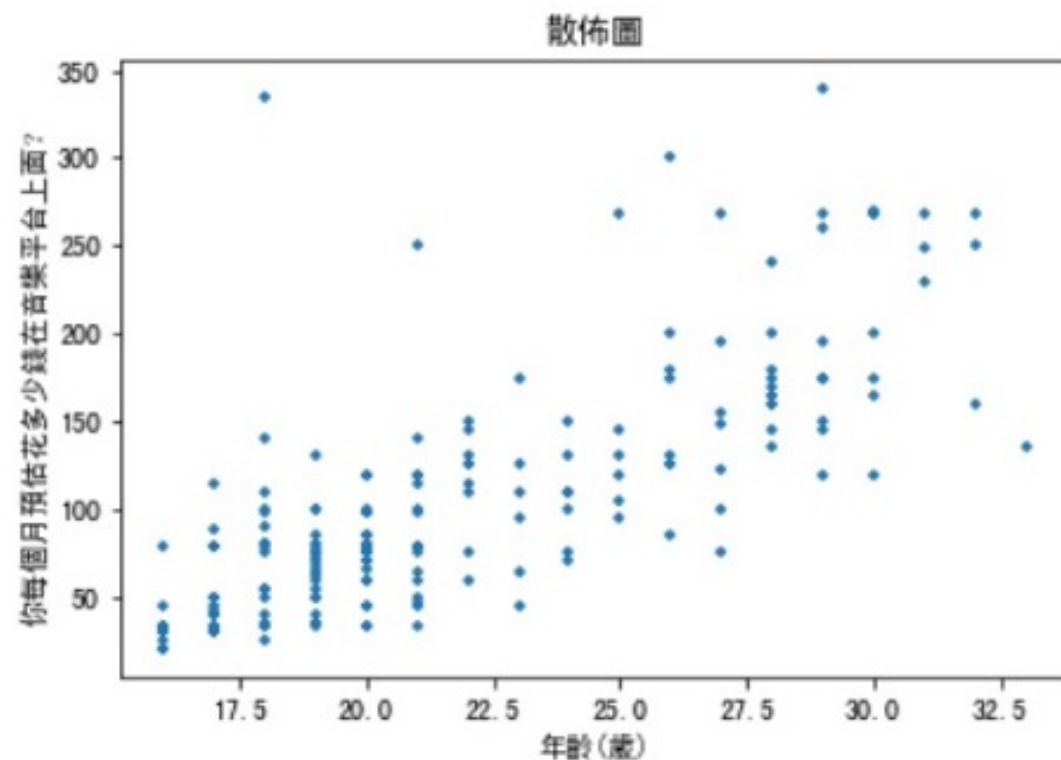
# 資料前處理 — 資料視覺化

長條圖：X軸 → 每月所得 / Y軸 → 可接受之平台使用價格



# 資料前處理 — 資料視覺化

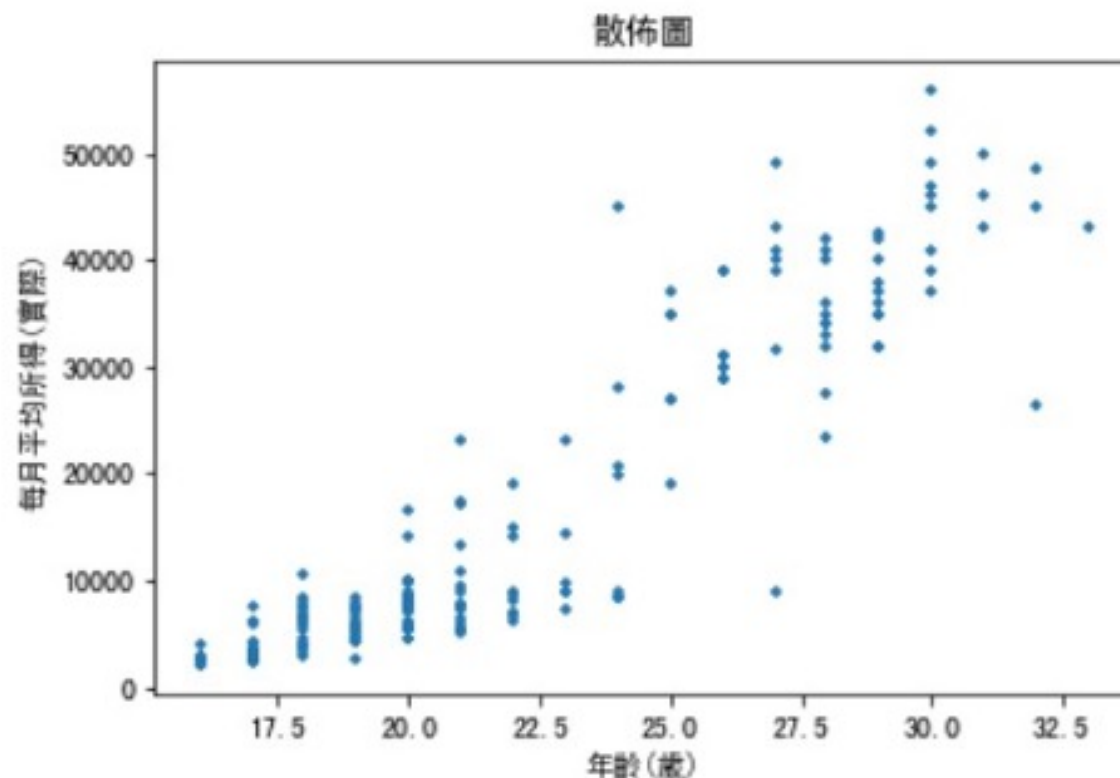
散佈圖：X軸 → 年齡 / Y軸 → 每個月的平台花費



正相關

# 資料前處理 — 資料視覺化

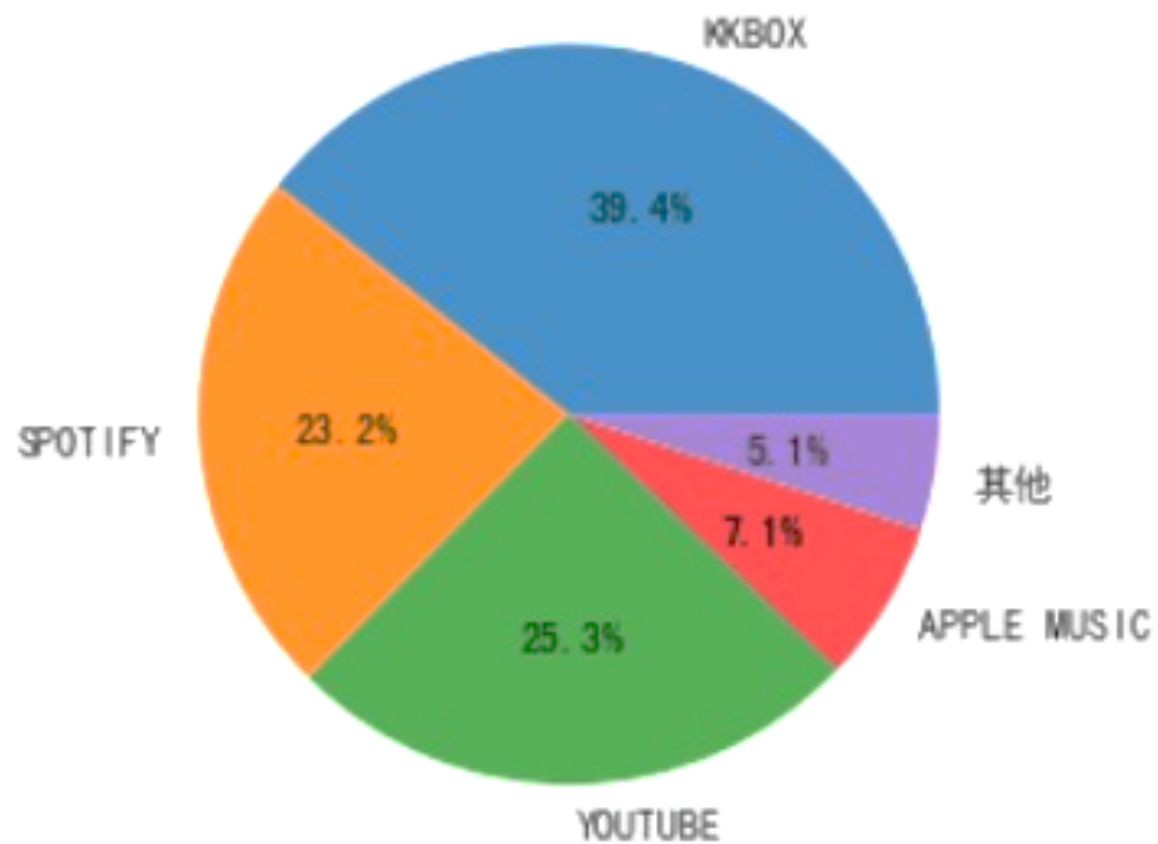
散佈圖：X軸 → 年齡 / Y軸 → 平均所得



正相關

# 資料前處理 — 資料視覺化

圓餅圖



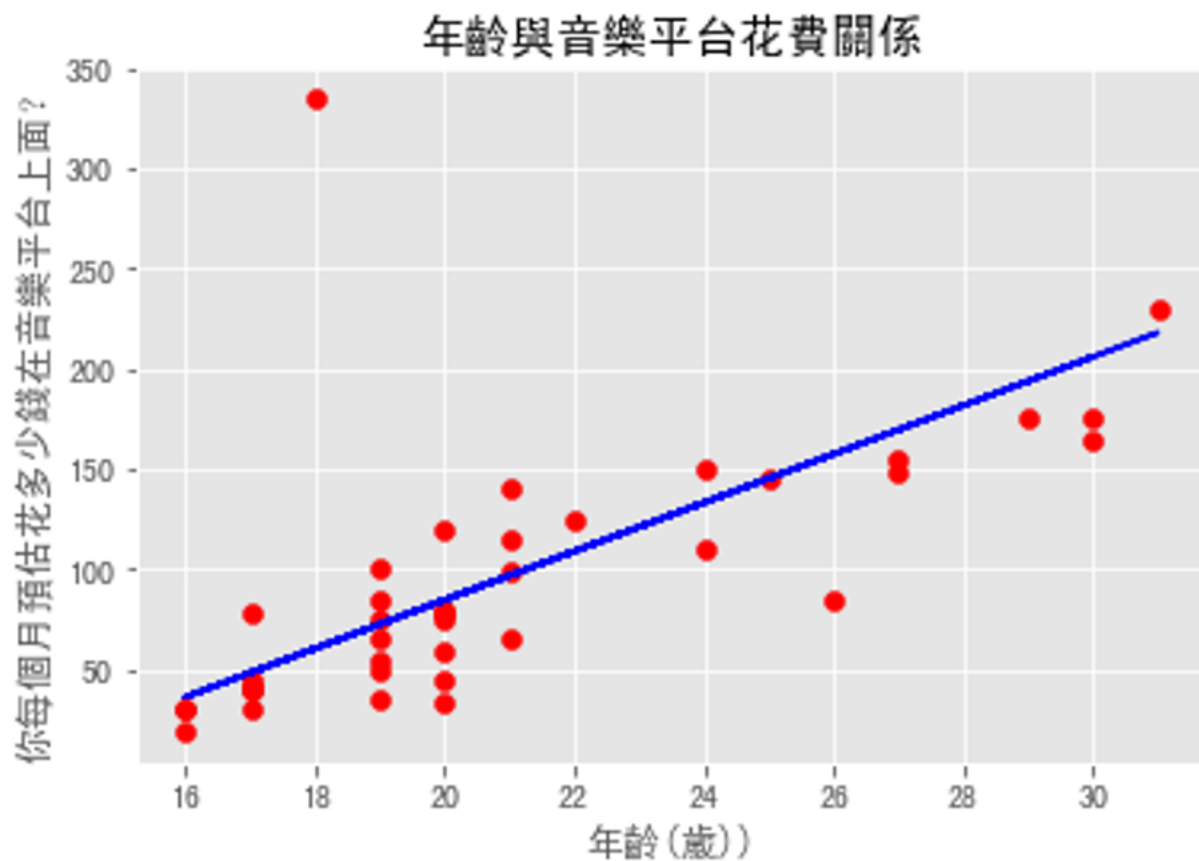
## PART 04

# 研究結果與討論

*Findings and Discussion*

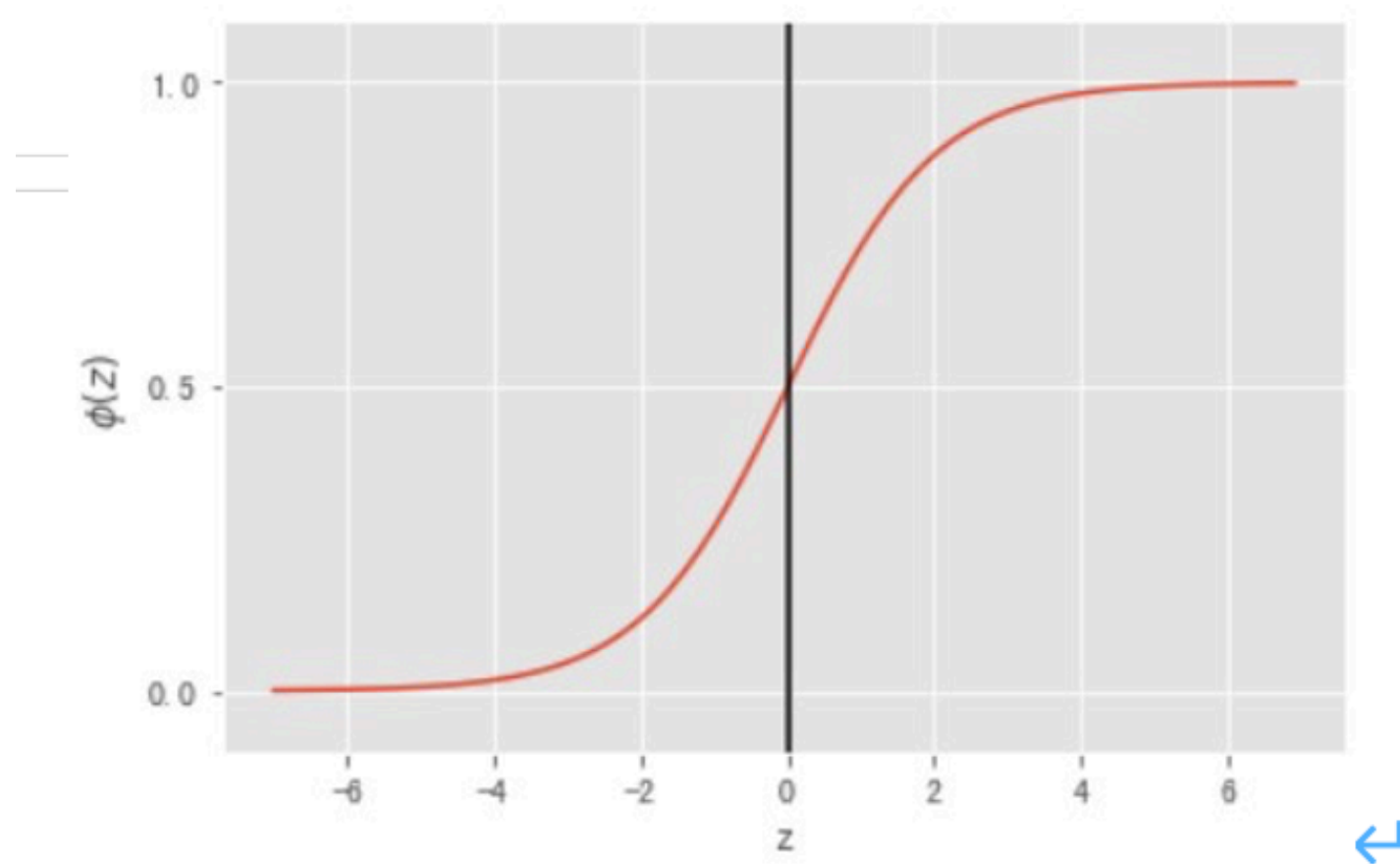


# 回歸分析 — 簡單線性回歸



- $R$  平方：0.629
- 準確率：36.75%
- 平均絕對誤差：26.17

# 羅吉斯回歸



# 羅吉斯回歸

- 資料分割

訓練集：75%

測試集：25%

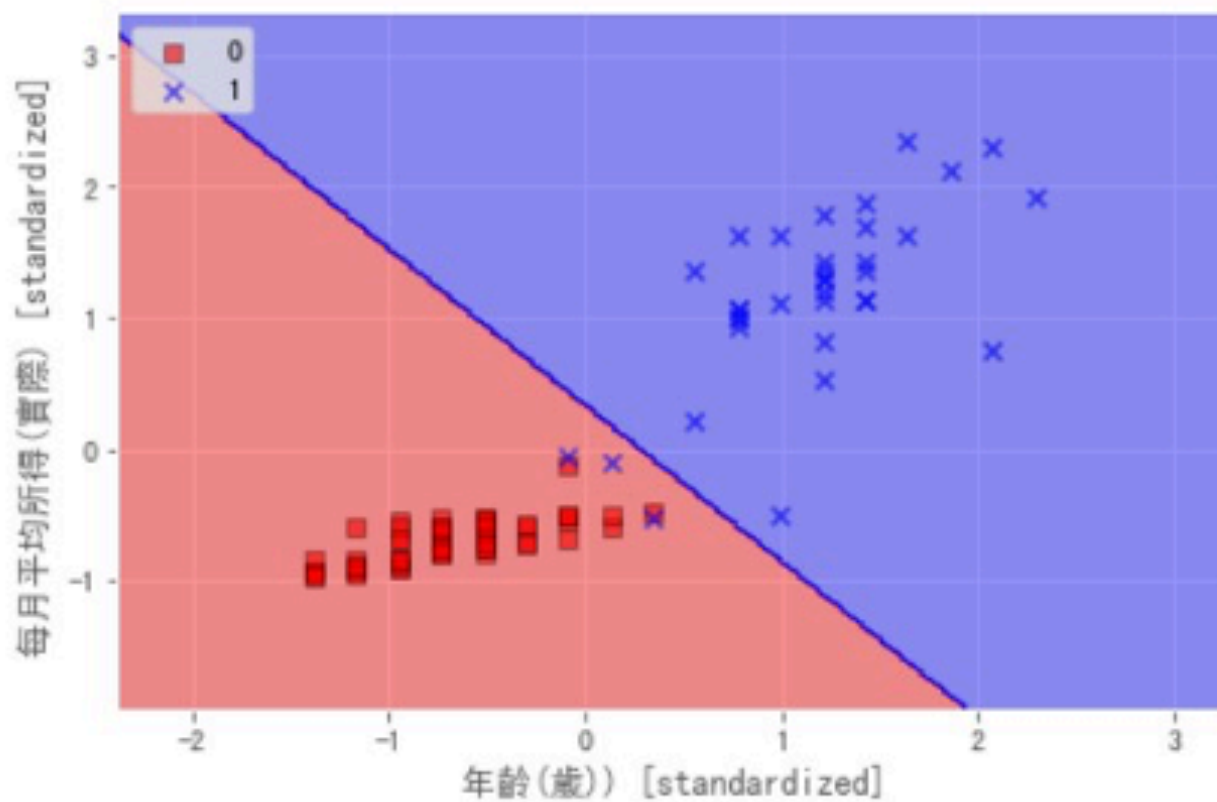
- 錯誤個數

總資料：124個

錯誤個數：2個



# 羅吉斯回歸



# 混淆矩陣

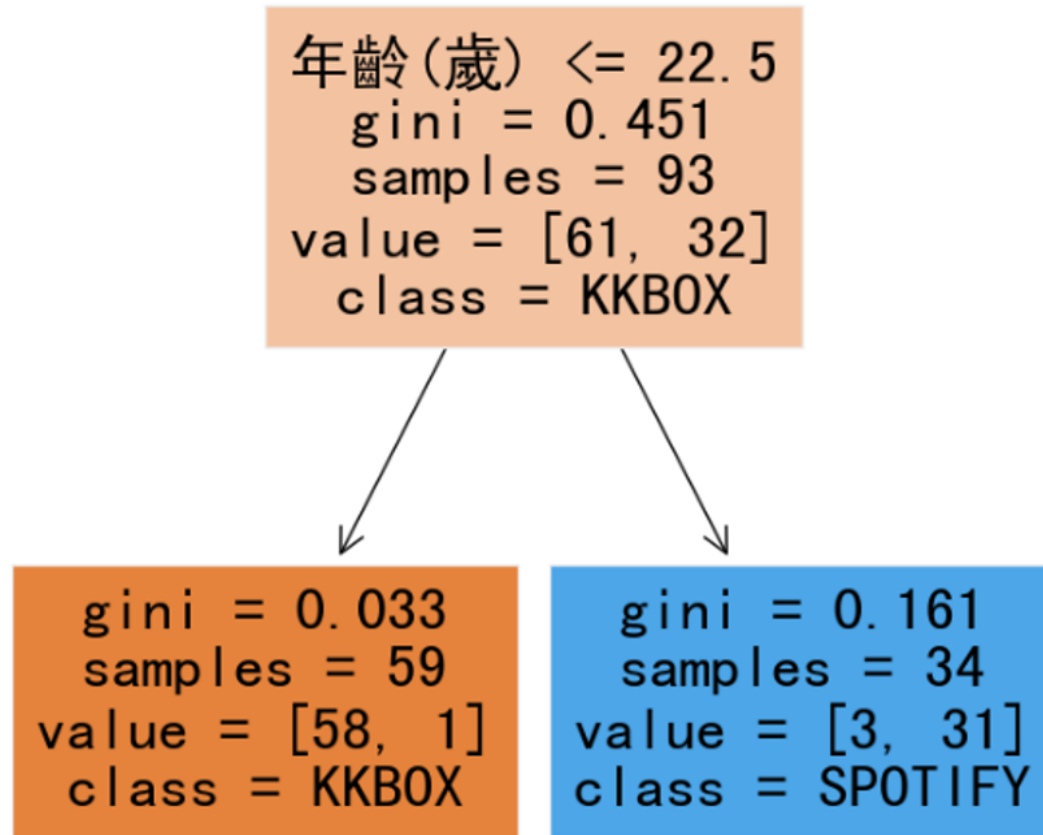
---

	KKBOX	SPOTIFY
KKBOX	14	0
SPOTIFY	2	9

---

- 模型中有兩筆資料預測錯誤，實際結果為SPOTIFY，但預測結果為KKBOX
- 正確率：92%

# 決策樹分析



• 特徵重值：年齡 > 收入

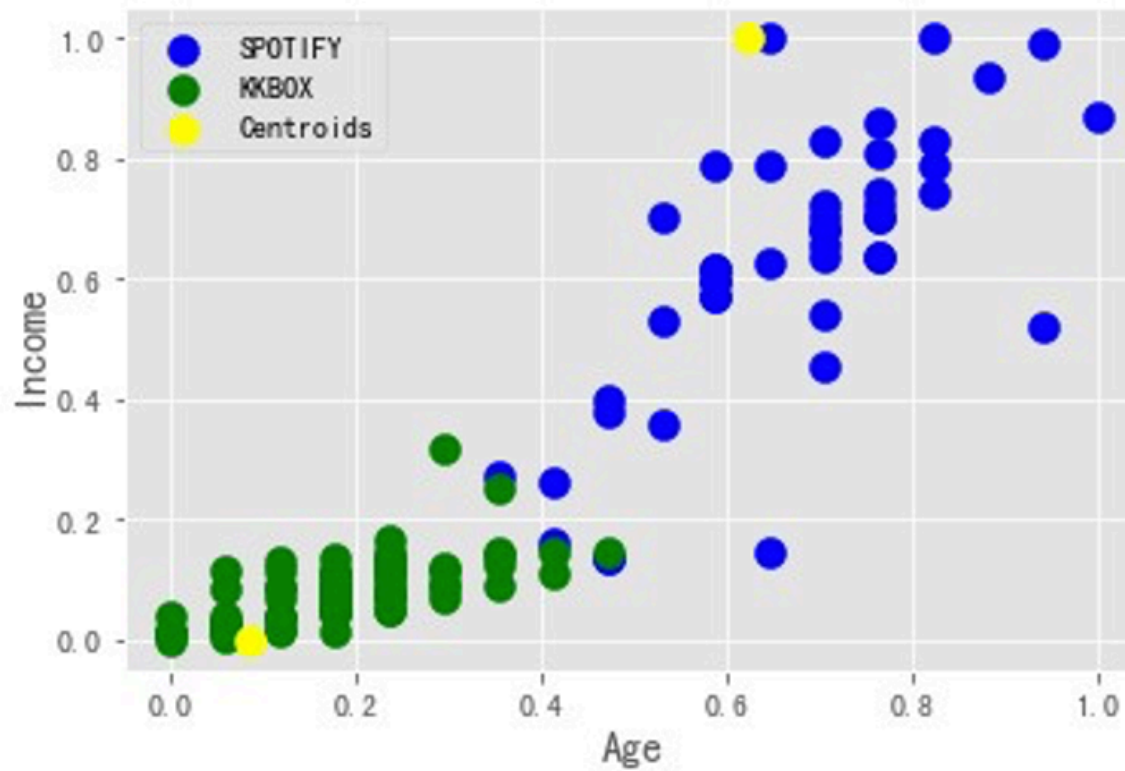
	0
年齡	1.0
收入	0.0

• 準確率：94%

# 群集

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 x 2 columns



## PART 05

# 結論與未來研究方向

*Conclusion and future research directions*



# 結論與未來研究方向

- 以KKbox、Spotify為此篇研究的重點方向
- 年齡  $\leq 22.5$  歲的族群傾向使用KKBOX，根據表單調查KKBOX的廣告資訊來源大多來自電信推薦
- 年齡  $> 22.5$  歲以上的族群傾向SPOTIFY，其廣告資訊來源大多來自朋友推薦
- 未來研究方向：
  - 最近興起的podcast也是值得注意的點
  - 加入更多平台的研究來增加資料的多元性



Q & A

•

•

•

•