

專案作品

# 展店選址一把抓



姓名：何宗育



# 大綱

---

**01** | 專案介紹

**02** | 平台建置


**03** | 資料收集及前處理

**04** | 資料分析

**05** | 機器學習

**06** | 專案成果及商業應用

**07** | 總結



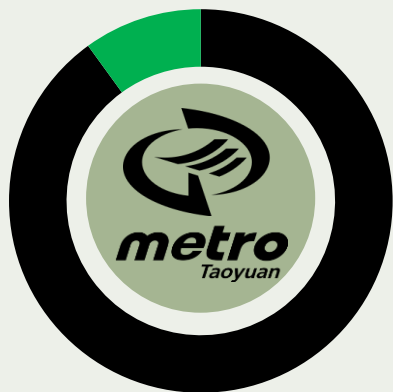
# /01 專案介紹

---

- 研究動機
- 研究目的
- 使用工具

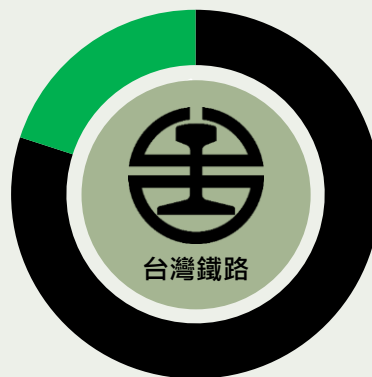
台灣109年軌道(日載客量)總計 **247萬**  
運量帶動周邊發展形成**軌道經濟**

5萬人次/日



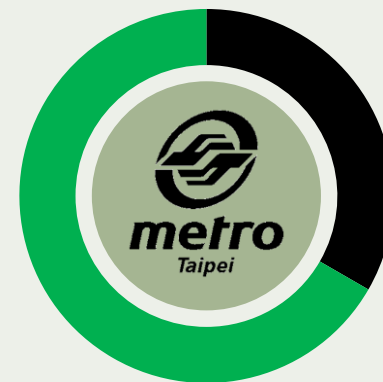
桃園捷運

55萬人次/日



台鐵

187萬人次/日



台北捷運

專案介紹

平台建置

資料收集  
及前處理

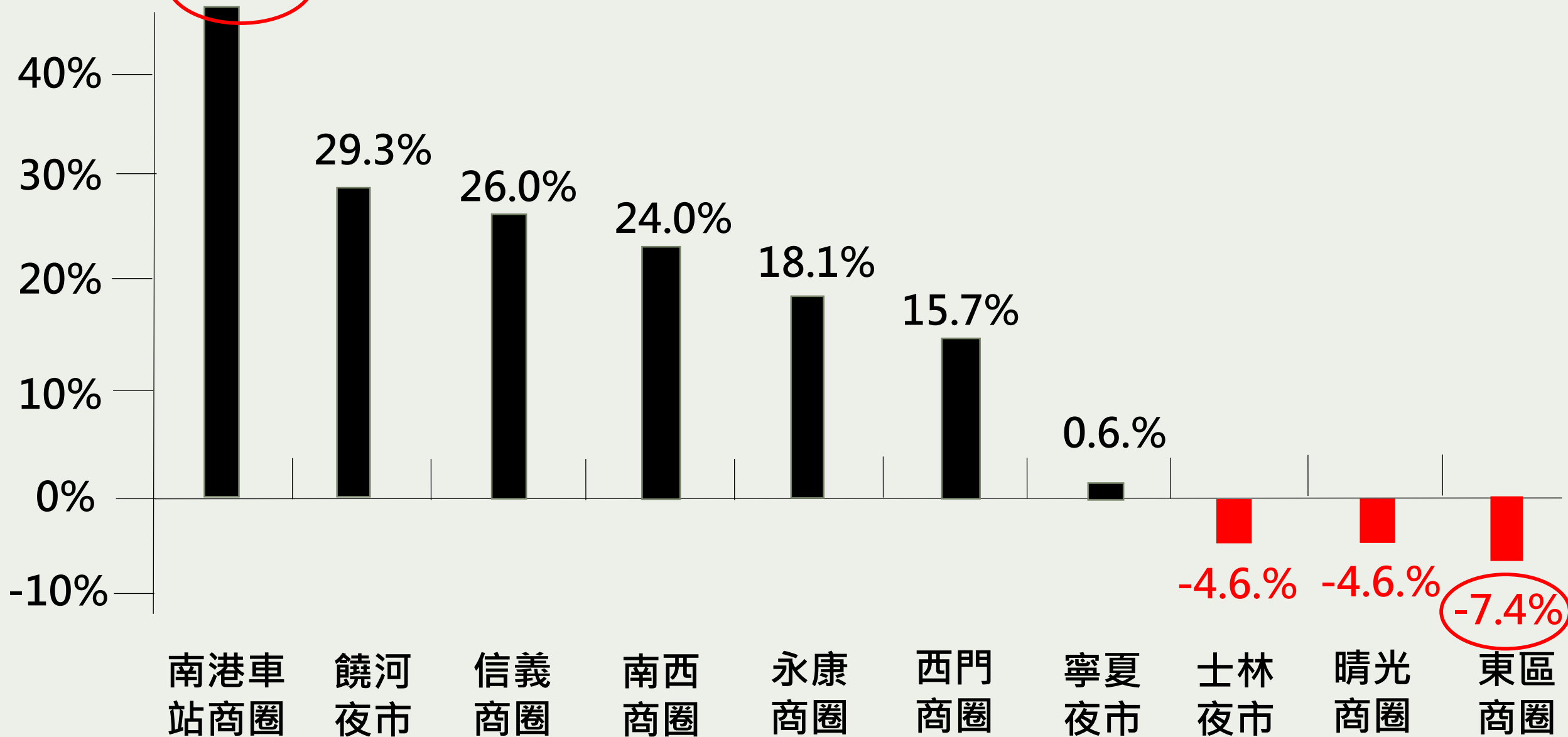
資料分析

機器學習

專案成果及  
商業應用

研究動機 89.2%

## 近五年台北市十大商圈鄰近捷運站進出人次變化



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

## 運用大數據分析

結合不同領域多元的資料加以整合分析一個地點的環境特色

預測軌道運輸站點周圍最適合**展店之因素**



商家數量



站點流量



店面租金

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 工具列表

## 資料收集及爬取

python™

se Selenium

BeautifulSoup

## 平台建置

ubuntu

hadoop

APACHE  
Spark™

lab

## 資料分析與清洗

python™

NumPy

pandas

## 機器學習

python™

scikit  
learn

## 視覺化

matplotlib

seaborn

## 網頁互動式呈現

HTML CSS JS

DB

Power BI

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用



## /02 平台建置

---

- Hadoop叢集架構
- Hadoop運作流程
- 節點配置



# Hadoop叢集架構

應用



程序管理



分散式  
檔案系統



虛擬機  
與系統



vmware



ubuntu

專案介紹

平台建置

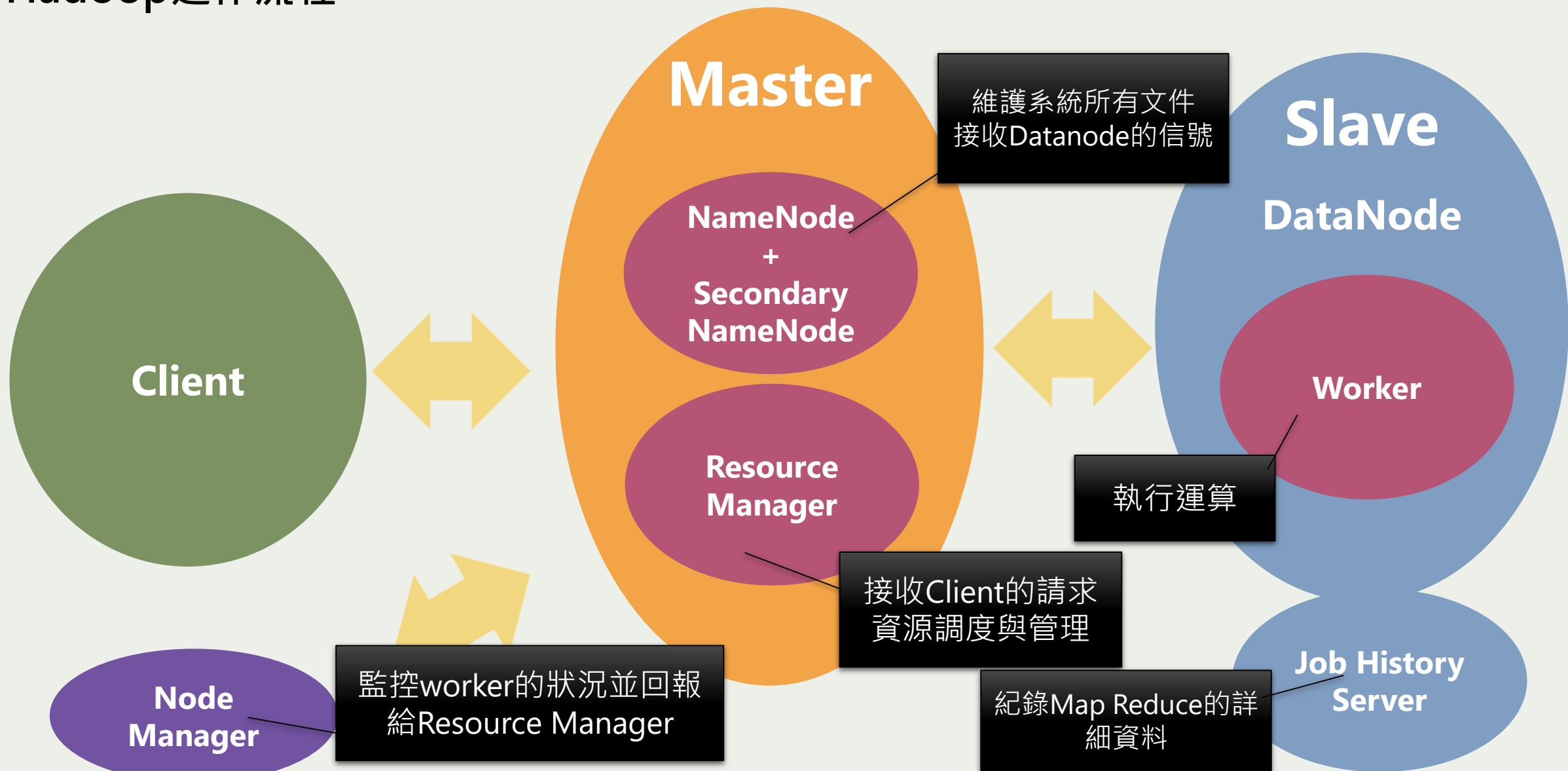
資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# Hadoop運作流程



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

## 節點配置



BDSE36

NameNode

BDSE169

Resource  
Manager



BDSE99

Job History  
Server



BDSE78  
BDSE79

DataNode /  
Worker



BDSE99  
BDSE100

DataNode /  
Worker



BDSE29  
BDSE30

DataNode /  
Worker



BDSE37  
BDSE170

DataNode /  
Worker



BDSE190  
BDSE191

DataNode /  
Worker

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

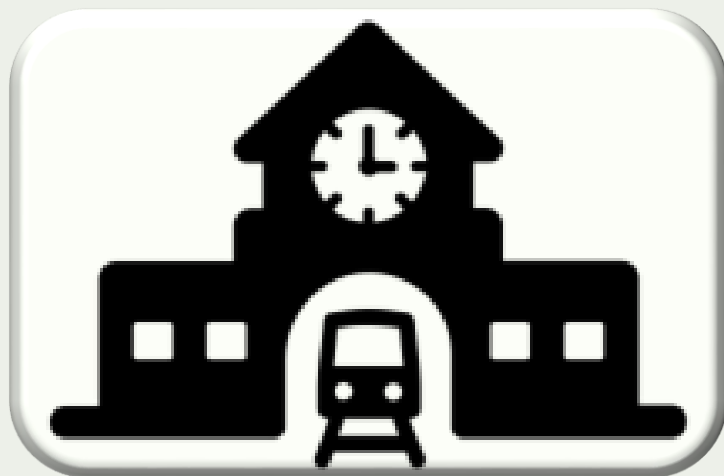


## /03 資料收集及前處理

---

- 資料收集
- 資料前處理

## 資料介紹



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用



政府資料開放平臺  
DATA.GOV.TW



內政資料開放平臺  
OPEN DATA

專案介紹

平台建置

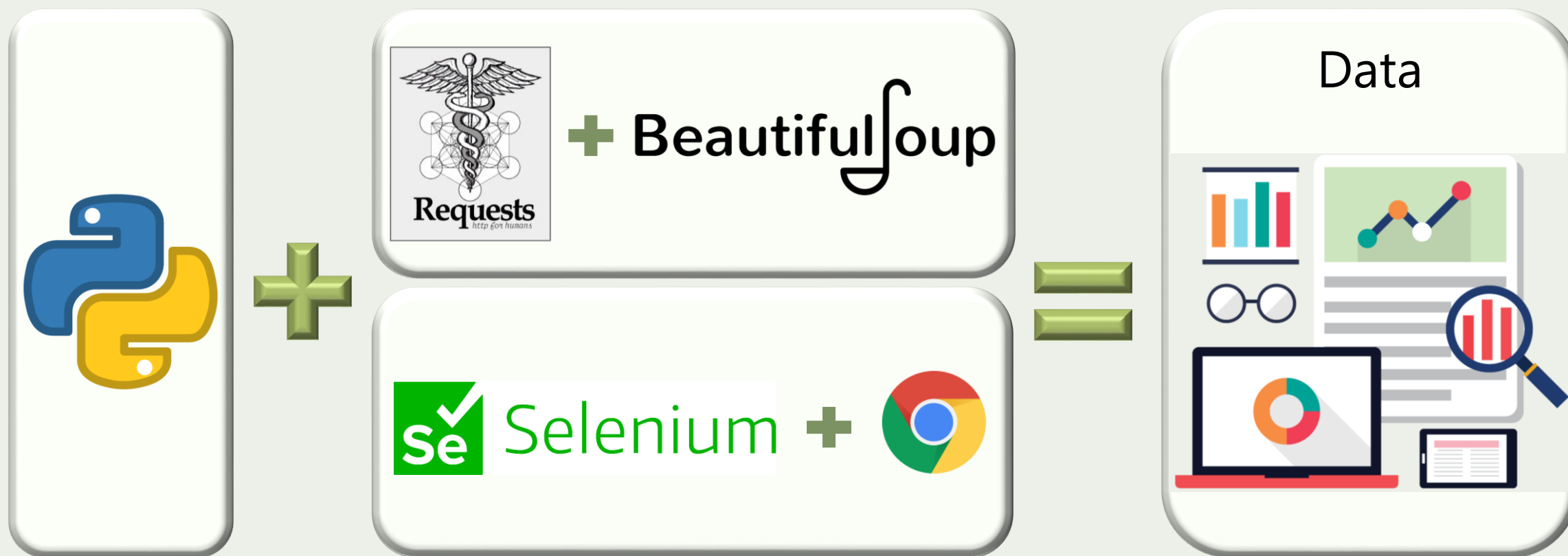
資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

## 爬蟲-工具&方法



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 爬蟲-問題

## 困難點

網頁彈跳視窗載入慢，以致程式抓取不到頁面內容



## 解決方式

增加緩衝時間，讓彈跳窗跳出後再定位

z z z



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用



# 爬蟲-問題

## 困難點

網頁有防爬蟲機制，  
爬取頻率過高時連線失敗



## 解決方式

增加隨機休息時間，  
模擬真人操作，降低  
被擋爬蟲機率

z z z

```
ConnectionError: HTTPSConnectionPool(host='rent.591.com.tw', port=443): Max retries exceeded with url: /rent-detail-10006047.html (Caused by NewConnectionError('<urllib3.connection.HTTPSConnection object at 0x0000017F89841250>: Failed to establish a new connection: [WinError 10060] 連線嘗試失敗，因為連線對象有一段時間並未正確回應，或是連線建立失敗，因為連線的主機無法回應。'))
```



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 爬蟲-問題

## 困難點

部分網址失效，以致連線失敗，程式中斷



## 解決方式

程式中增加異常處理



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 爬蟲-問題

## 困難點

部分連結內無資料，  
程式抓取失敗而中斷



## 解決方式

程式中增加異常處理



專案介紹

平台建置

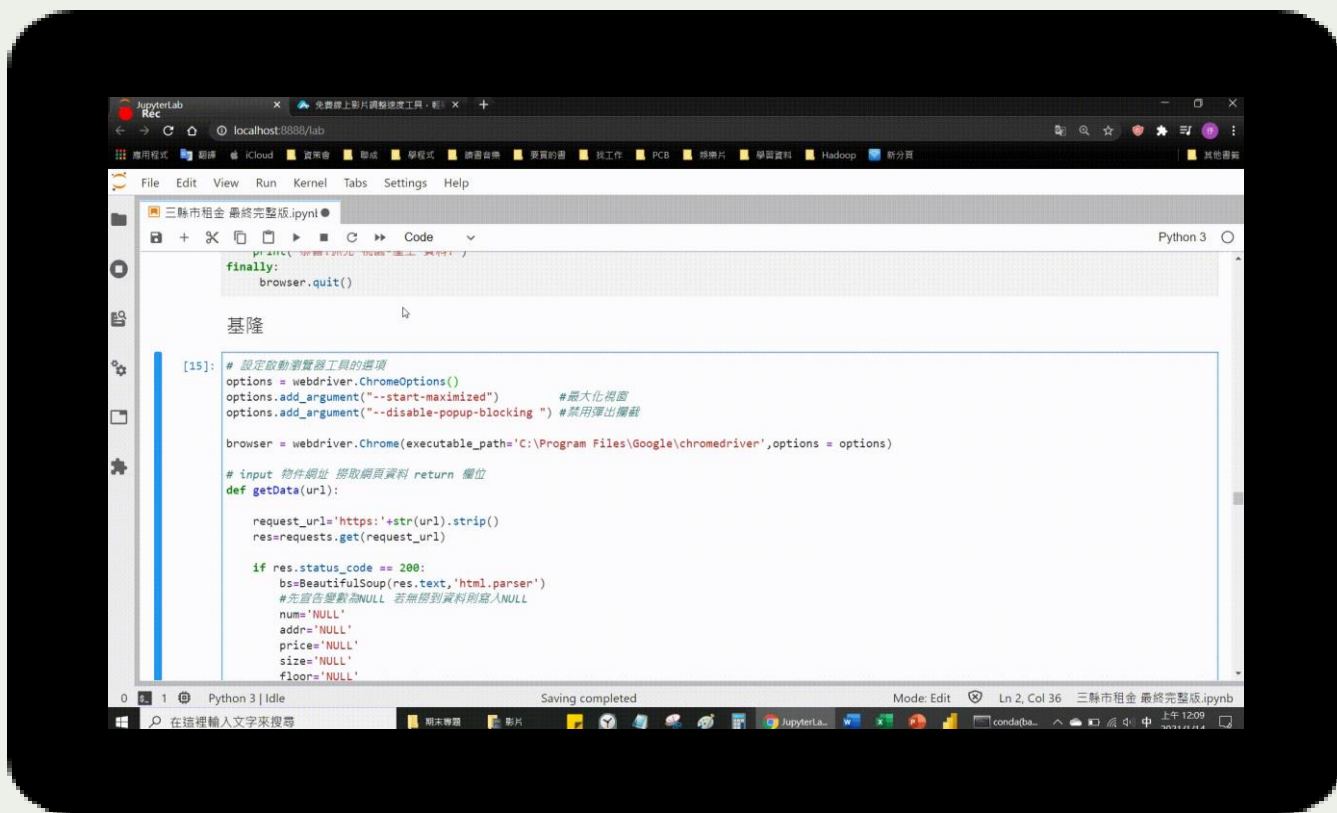
資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

## 爬蟲-演示&成果



	store_name	store_no	addr
0	全家基隆車站店	18683	基隆市仁愛區中山一路16之1號
1	全家基隆鑫仁二店	19621	基隆市仁愛區仁二路198號
2	全家基隆仁三店	17147	基隆市仁愛區仁三路59號
3	全家基隆廟口	1288	65歲以上人口
4	全家基隆仁四	6725	20-24歲人口
未婚人口	有偶人口	出生數	死亡數
2034	2010	38	30
4081	4729	85	97
2293	2452	29	45
3530	3959	50	36
2027	2243	31	30
1886	2114	28	32
2869	3401	43	46
4864	4766	72	37
Country			
基隆市			
基隆市			
基隆市	中正區	信義里	1001701003
基隆市	中正區	義重里	1001701004
基隆市	中正區	港通里	1001701005
基隆市	中正區	中船里	1001701006
基隆市	中正區	博士人口	碩士人口
基隆市	中正區	59	500
基隆市	中正區	129	1366
基隆市	中正區	71	521
基隆市	中正區	164	1165
基隆市	中正區	1379	3639
基隆市	中正區	514	1241
基隆市	中正區	727	1683
基隆市	中正區	1125	235
基隆市	中正區	1194	239
基隆市	中正區	181	367
基隆市	中正區	269	18
基隆市	中正區	188	5
(R10063405)	新北市淡水區水源街二段	25,000	33.22坪
None	None	None	None
(R9855297)	新北市板橋區光武街	40,000	31.82坪
404	404	404	404
(R9954796)	新北市中和區景安路	220,000	56.2坪
(R9954763)	新北市中和區景安路	220,000	56.2坪
(R9954699)	新北市中和區景安路	220,000	56.2坪
(R8587075)	新北市板橋區民生路	15,000	2坪
None	None	None	None
(R9912902)	新北市蘆洲區中正路	28,000	17坪
(R9952951)	新北市蘆洲區中正路	26,000	15坪
(R10109265)	新北市三重區大同北路	18,000	30坪
None	None	None	None
店面(店舖)	店面(店舖)	店面(店舖)	店面(店舖)
2020-12-08	2020-12-08	2020-12-08	2020-12-08
2020-12-20	2020-12-20	2020-12-20	2020-12-20
2021-01-11	2021-01-11	2021-01-11	2021-01-11
2021-01-11	2021-01-11	2021-01-11	2021-01-11
2020-11-25	2020-11-25	2020-11-25	2020-11-25
2021-01-03	2021-01-03	2021-01-03	2021-01-03
2021-01-03	2021-01-03	2021-01-03	2021-01-03
2021-02-13	2021-02-13	2021-02-13	2021-02-13
2021-02-13	2021-02-13	2021-02-13	2021-02-13

# 資料前處理



## 資料探索

判斷蒐集資料可用性



## 資料清洗

- 站點
- 鄰里資料
- 租金
- 流量
- 設施
- 商家數量



## 距離計算

- 站點    v.s
- 鄰里辦公室
  - 租金物件
  - 設施數量
  - 商家數量

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 資料探索 - 初步判斷資料可用性

## 初步蒐集資料



鄰里資料

機能設施資料

站點資料

店家數量資料

租金資料

流量資料

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 資料清洗 – 站點資料

站點資料



地址資訊整理，新增欄位

車站營運狀態調整

車站種類	車站狀態	車站編號	車站名稱	車站位置	lat	lng
tp_mrt	activated	BL01	頂埔	236040新北市土	24.96012	121.4205
tp_mrt	activated	BL02	永寧	236036新北市土	24.96682	121.43613
tp_mrt	activated	BL03	土城	236017新北市土	24.97313	121.44432
tp_mrt	activated	BL04	海山	236023新北市土	24.985305	121.44873



車站種類	車站狀態	車站編號	車站名稱	縣市	行政區	轉乘狀態	轉乘捷運	轉乘台鐵	轉乘高鐵	車站位置	lat	lng
tp_mrt	1	BL01	台北捷運頂埔站	新北市	土城區	N	1	0	0	新北市土城區中	24.96012	121.4205
tp_mrt	1	BL02	台北捷運永寧站	新北市	土城區	N	1	0	0	新北市土城區中	24.96682	121.43613
tp_mrt	1	BL03	台北捷運土城站	新北市	土城區	N	1	0	0	新北市土城區金	24.97313	121.44432
tp_mrt	1	BL04	台北捷運海山站	新北市	土城區	N	0	0	0	新北市土城區海	24.985305	121.44873

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

## 資料清洗 – 站點資料

站點資料



合併捷運、台鐵、高鐵轉乘站點

tp_mrt	1	BL12	台北捷運台北車站
tp_mrt	1	R10	台北捷運台北車站
ty_mrt	1	A1	桃園捷運台北車站
train	1	1000	台鐵台北站
thsr	1	Taipei	高鐵台北站



TpMrt_TyMrt_Train_Thsr	1	R10_BL12_A1_1000_Taipei	台北車站
------------------------	---	-------------------------	------

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用



# 資料清洗 – 鄰里資料

鄰里資料



## 困難點

各行政區各里數量年間變化

字元亂碼，不同年間使用不同編碼方式

## 解決方法

對照行政院統一之村里代碼，  
依縣市、區、里名稱比對合併

資料轉換，以站點為單位，填入需要欄位

縣市代碼	縣市名稱	鄉鎮市區代碼	鄉鎮市區名稱	村里代碼	村里名稱
63000	臺北市	63000010	松山區	63000010-002	莊敬里
63000	臺北市	63000010	松山區	63000010-003	東榮里
63000	臺北市	63000010	松山區	63000010-004	三民里



Station_Name	Num_Village	Income_Gross	Income_Average	Population
頂埔站	30	3059.16	843.21	12895
永寧站	49	5171.97	855.15	21287
土城站	63	7041.97	872.39	28175
海山站	105	10821.72	882.97	42791

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 資料清洗 – 鄰里資料

鄰里資料



## 困難點

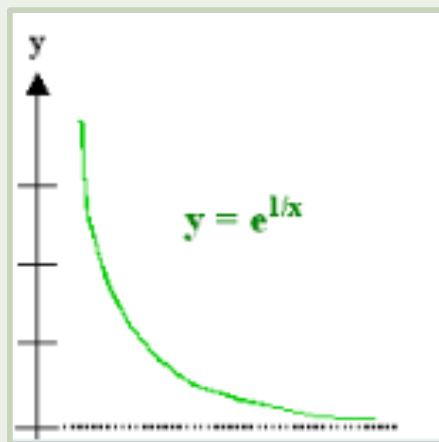
各行政區各里數量年間變化

字元亂碼，不同年間使用不同編碼方式

## 解決方法

對照行政院統一之村里代碼，  
依縣市、區、里名稱比對合併

依各里辦公室與站點位置距離  
計算**權重**，**距離越遠權重越小**



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 資料清洗 - 租金

租金資料



問題

以坪數、租金資訊取得「單位租金」

缺值處理

型別轉換

編號	地址	月租金	坪數	單位租金	樓層	房屋型態
( R10191372 )	台北市中正區開...	95000	17.3	5491	1F/5F	店面 ( 店鋪 )
( R9874162 )	台北市大安區信...	786830	143.1	5498	1F/12F	電梯大樓

處理

Drop 坪數、租金、地址缺值資料

數值欄位： object → int

日期欄位： object → data

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 資料清洗 - 租金

租金資料



問題

缺值處理

型別轉換

處理

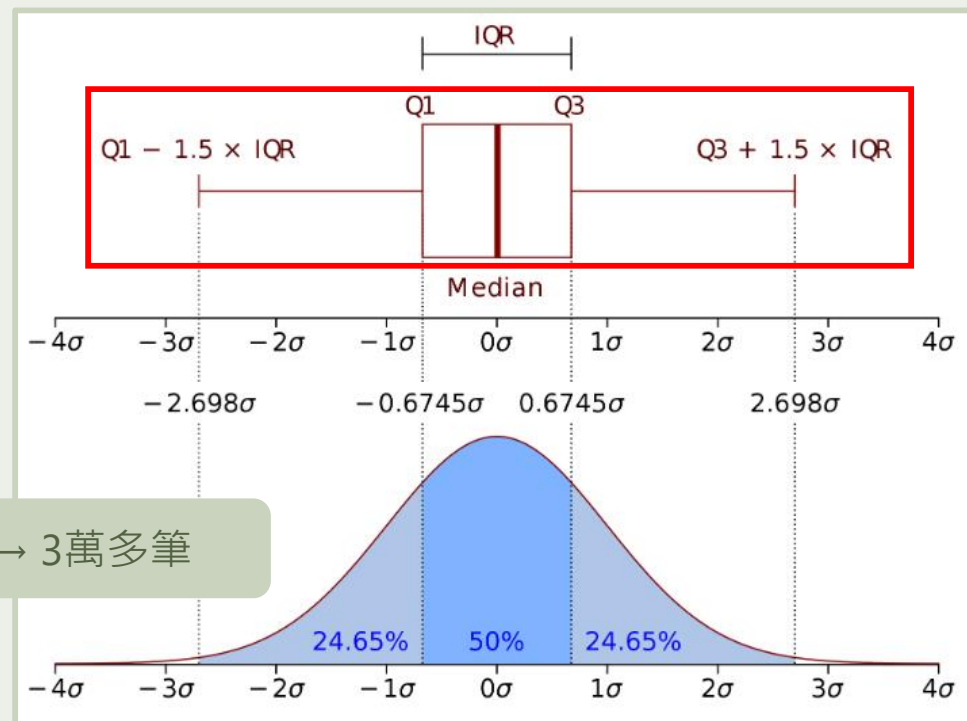
Drop 坪數、租金、地址缺值資料

數值欄位: object  $\rightarrow$  int

日期欄位: object  $\rightarrow$  data

離群值處理

5萬多筆  $\rightarrow$  3萬多筆



專案介紹

平台建置

資料收集  
及前處理

資料分析

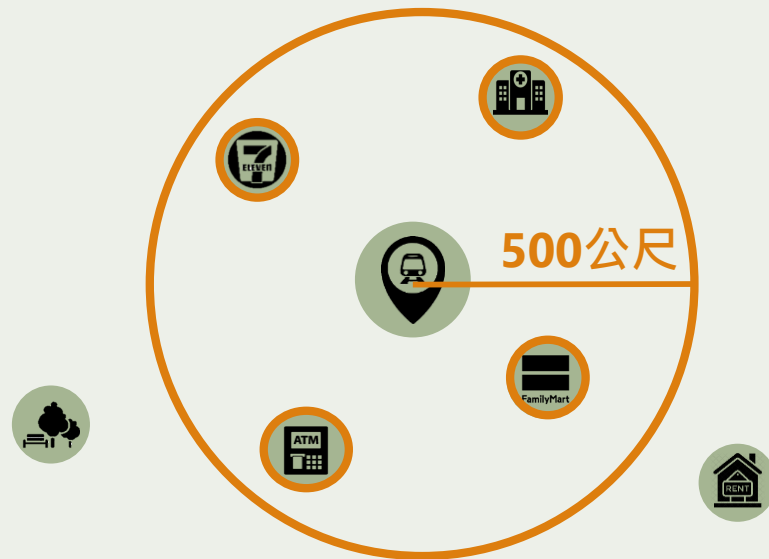
機器學習

專案成果及  
商業應用

## 資料前處理 - 距離計算

GeoPy

資料經緯度



主資料

站點資料

里辦公室

機能設施

租金物件

主資料加入各  
站點周圍500公  
尺內目標數量

合併站點周遭  
商家數量資訊

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 資料前處理 - 綜合大表

## 特徵值

## 預測項目

車站種類	中等教育數量	幼年人口%	基礎教育人口%	車站流量
營運狀態	高等教育數量	壯年人口%	未受教育人口%	房租中位數
車站編號	總教育機構數量	老年人口%	男性人口%	咖啡店數量
轉乘狀態	旅遊景點數量	青壯年人口%	已婚人口%	飲料店數量
轉乘數量	公園數量	中壯年人口%	移入人口%	酒吧數量
ATM數量	村里辦公室數量	扶幼比%	出生率%	健身房數量
便利商店數量	綜合所得總額	扶老比%	週間流量	旅館數量
醫療院所數量	平均綜合所得	高等教育人口%	週末流量%	
基礎教育數量	總人口	中等教育人口%		

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

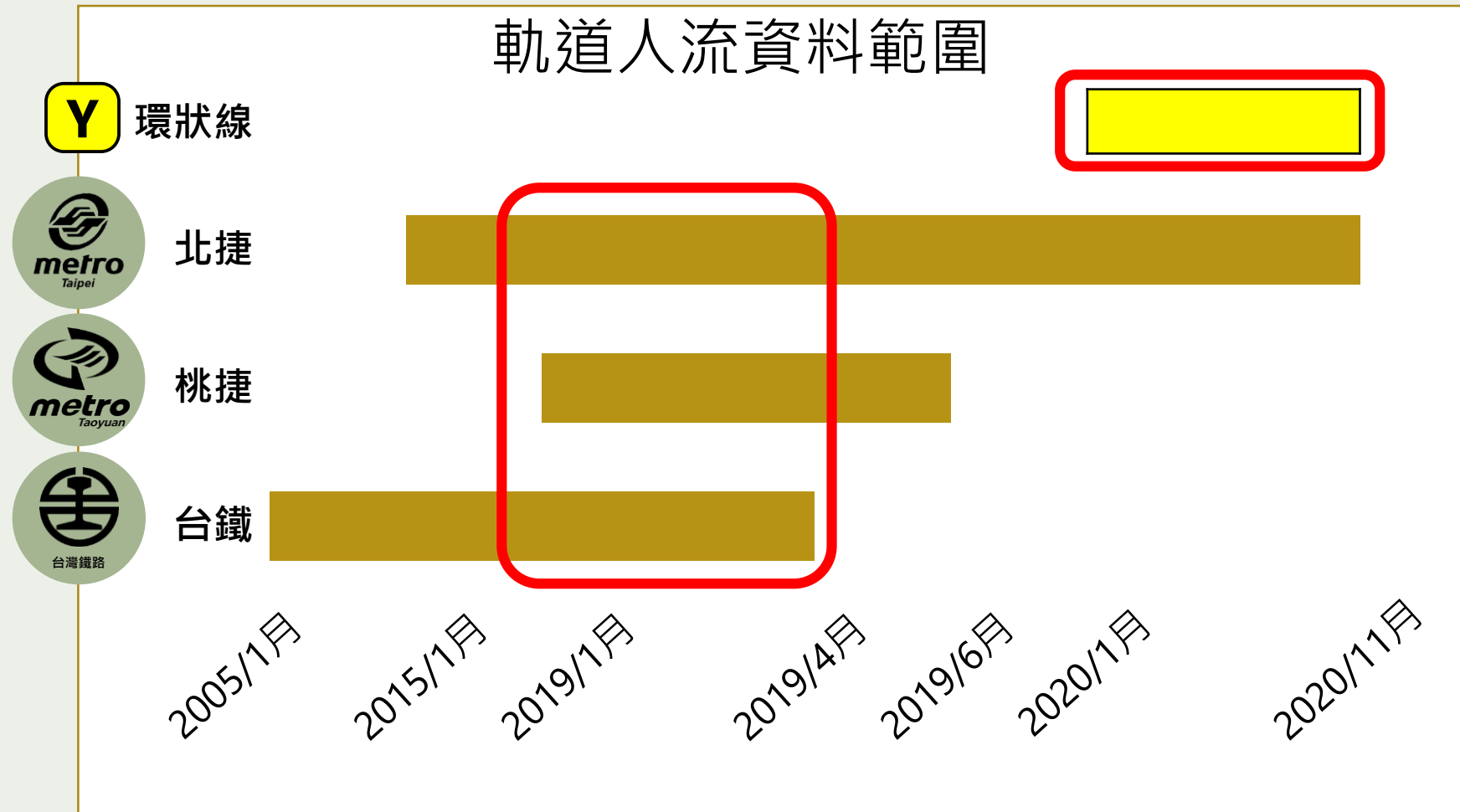


## /04 資料分析

---

- 資料探索
- 相關性分析
- 參數挑選

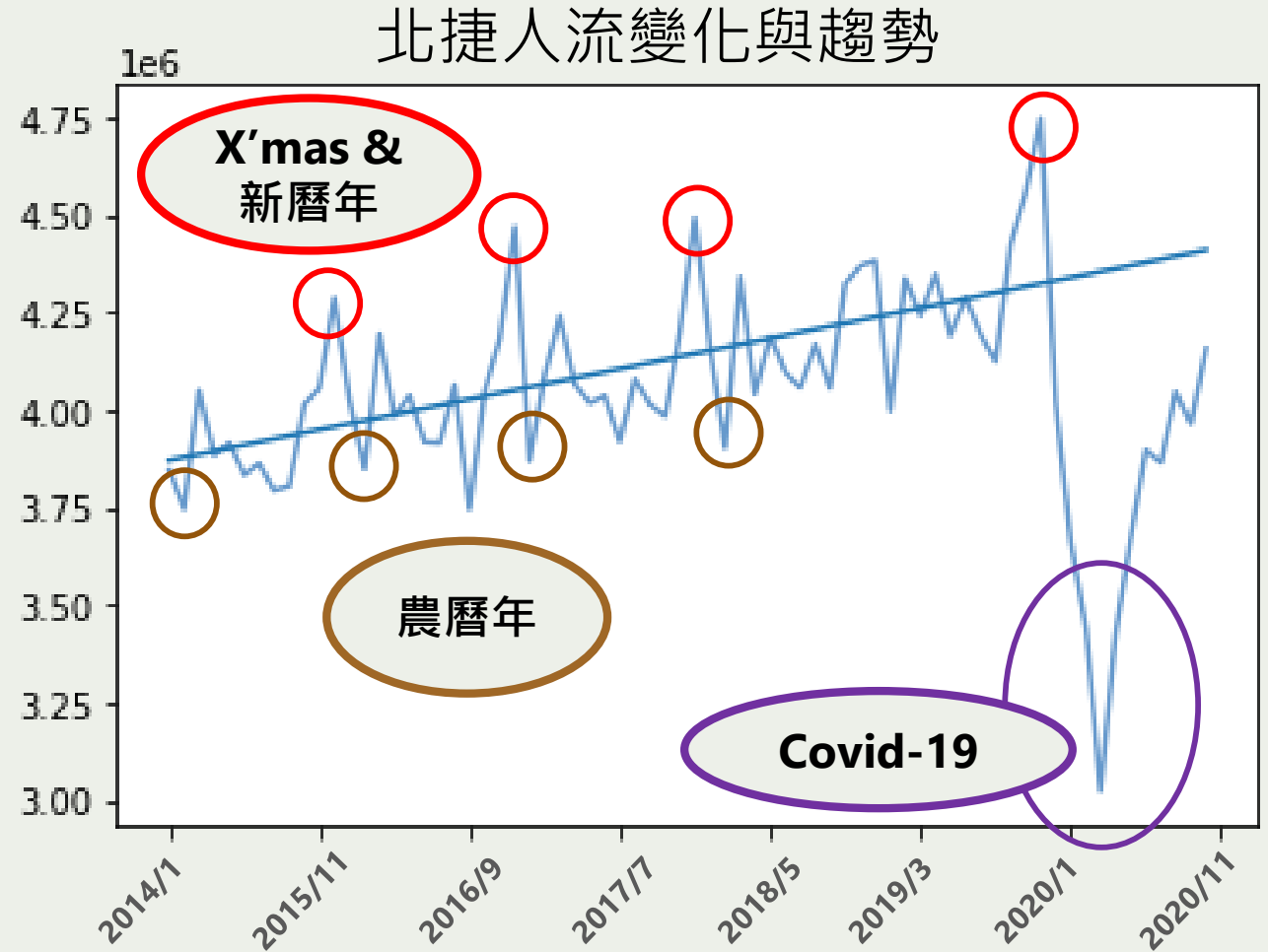
# 資料探索(1/3)



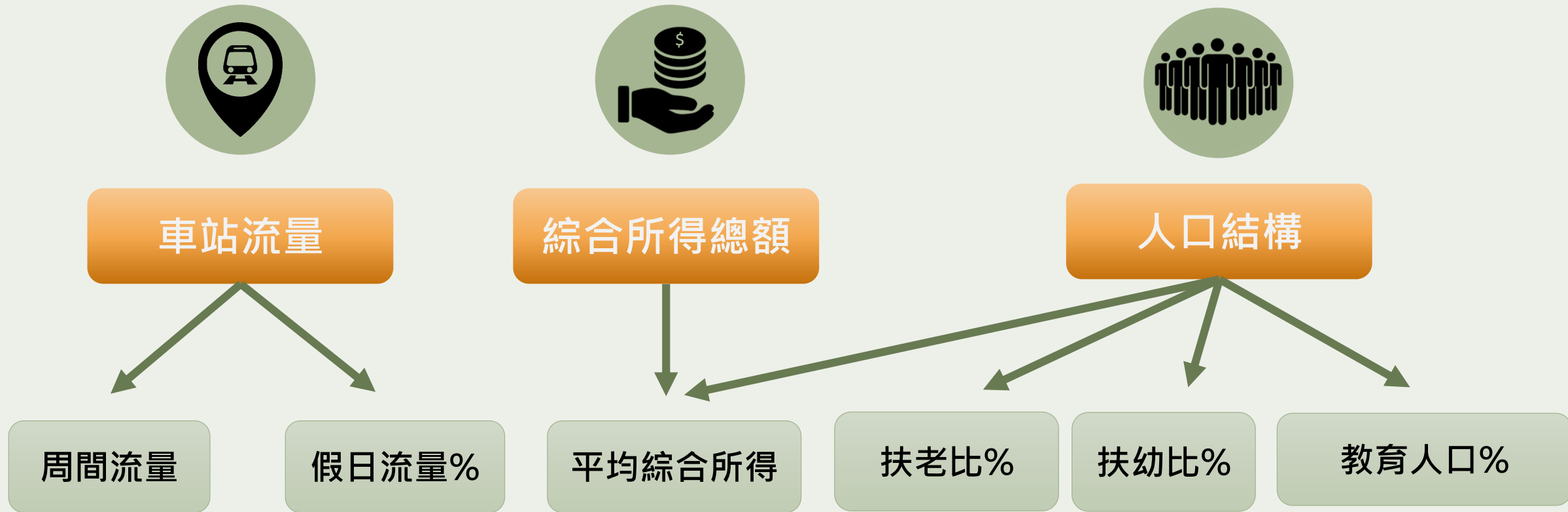


## 資料探索(2/3)

- 降低模型影響
- 避免極端值
  - 2020新冠肺炎疫情影响
  - 聖誕及跨年
  - 農曆年



# 資料探索(3/3) - 資料延伸



開店、租金、流量

之影響因子



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 相關性分析(1/3)



- 縱/橫軸為特徵
- 紅 → 正相關
- 藍 → 負相關
- 淺色 → 無相關

	便利商店數量	醫療院所數量	綜合所得總額	總人口	壯年人口比%
咖啡店數量	0.69	0.71	0.73	0.60	-0.60

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用

# 相關性分析(2/3)



- 縱/橫軸為特徵
- 紅 → 正相關
- 藍 → 負相關
- 淺色 → 無相關

	收入總和	收入平均	高等教育比	壯年比	中等教育比
租金中位數	0.74	0.70	0.66	-0.69	-0.65

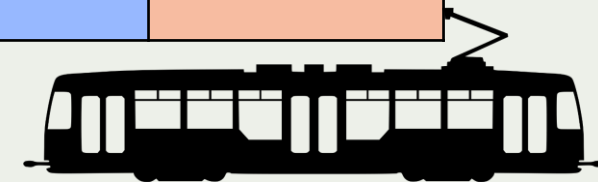


# 相關性分析(3/3)



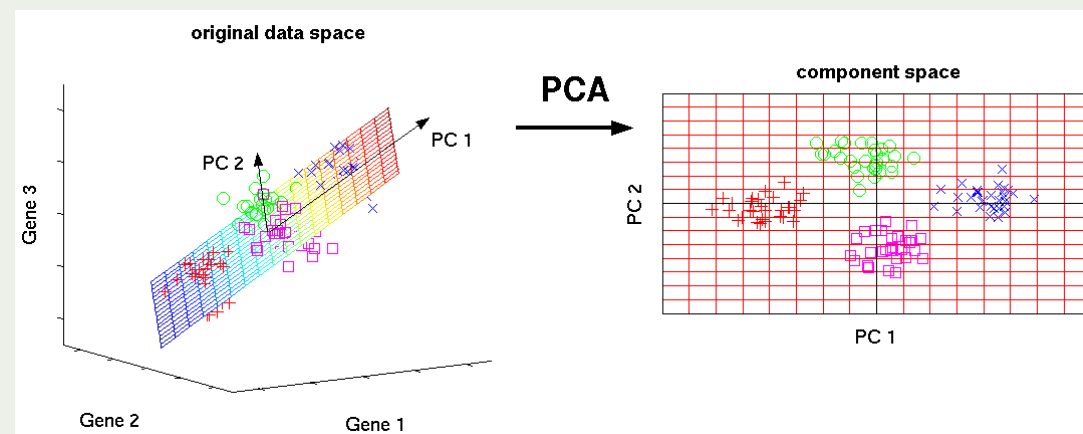
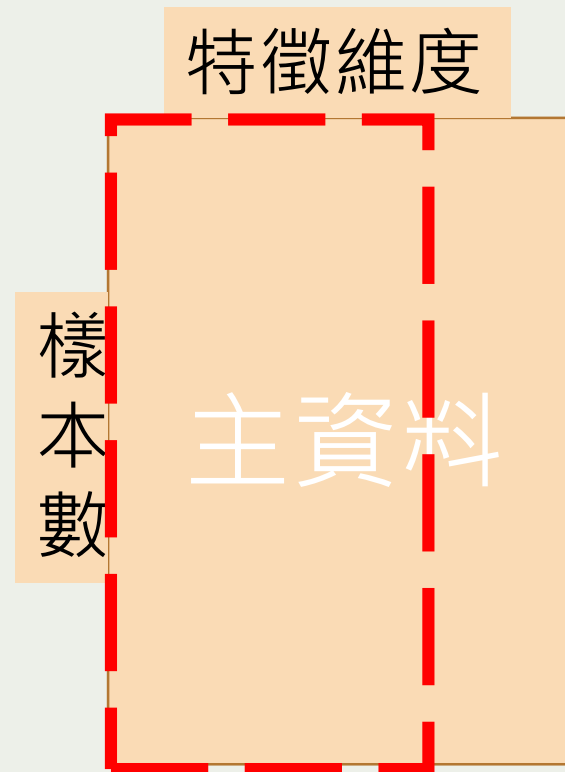
- 縱/橫軸為特徵
- 紅 → 正相關
- 藍 → 負相關
- 淺色 → 無相關

	ATM數量	便利商店數量	綜合所得總額	總人口	扶幼比%	扶老比%
車站流量	0.59	0.55	0.50	0.46		
周末流量%					-0.40	0.43

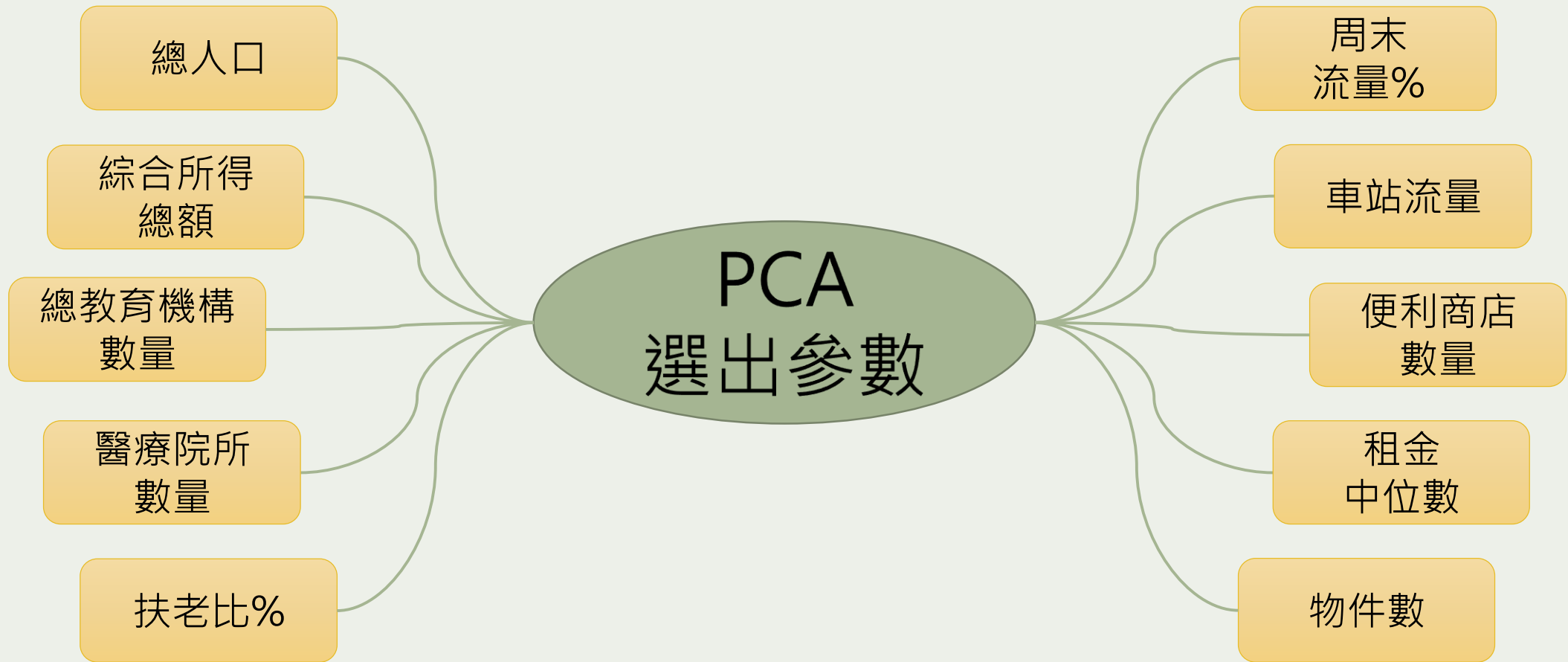


# 參數挑選(1/2)

- 理想的資料
  - 樣本數  $\gg$  特徵維度
  - 互不干擾
- 主成分分析PCA(非監督式)  
(Principal Component Analysis)
- 功能:
  - 降維
  - 去貢獻值



## 參數挑選(2/2)







## /05 機器學習

---

- 模型訓練
- 績效評估
- 效能促進
- 評估結果討論



## 流程

模型訓練



效能評估



效能促進



■ 挖掘重要因子  
■ 預測

- 商家數量
- 站點流量
- 租金價格

專案介紹

平台建置

資料收集  
及前處理

資料分析

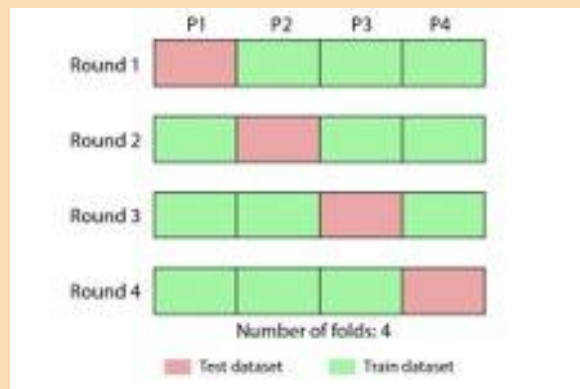
機器學習

專案成果及  
商業應用

## 模型績效評估、效能促進

### 交叉驗證

- 將訓練樣本拆分
- 估算績效均值



### 評估指標

#### 迴歸模型

- $R^2$ (決定係數): 0~1
  - 模型的解釋力

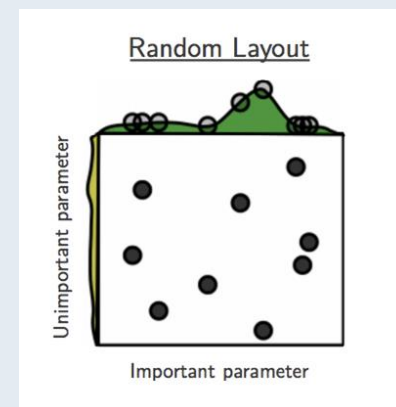
#### 分類模型

- Accuracy (分類正確率) : 0~1
  - 是否正確預測分類項目

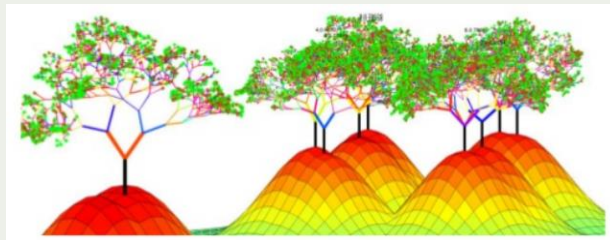
### 參數調整

#### 隨機搜尋法

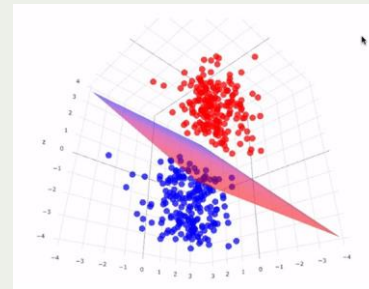
- 排列組合隨機抽樣
- 取得最佳模型參數



## 模型訓練



@Deepak George

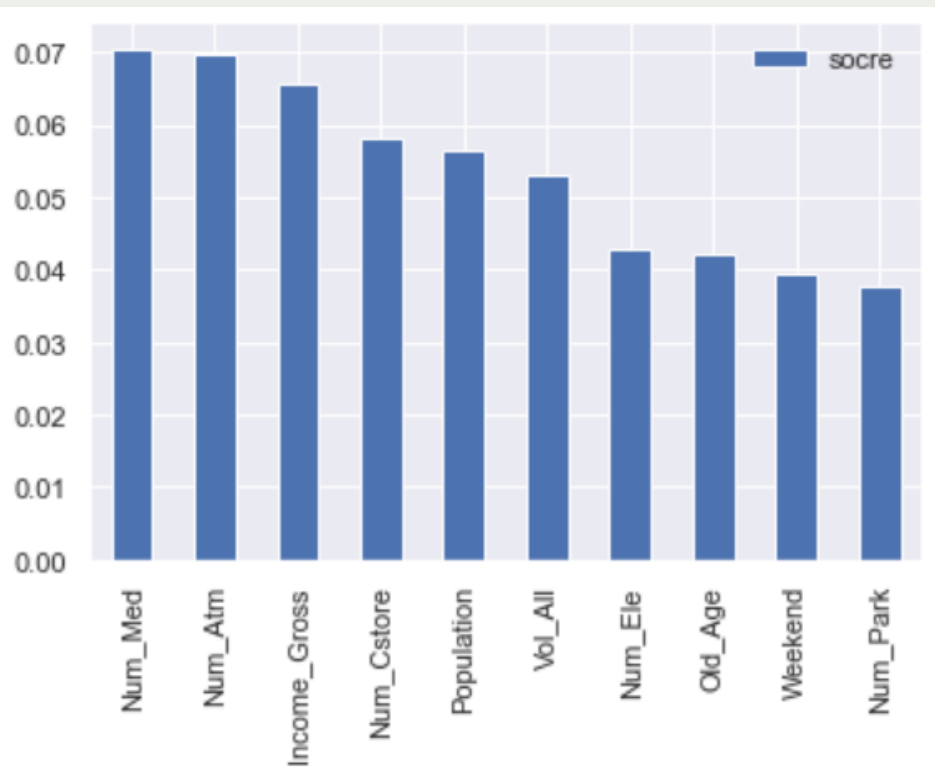


	Linear Reg 線性迴歸	RF 隨機森林	GBDT 梯度提升樹	SVM 支持向量
特點	<ul style="list-style-type: none"><li>運算速度快</li><li>處理可線性分類問題</li></ul>	<ul style="list-style-type: none"><li>決策樹具可解釋性</li><li>結合多個決策樹不易過度擬合</li></ul>	<ul style="list-style-type: none"><li>集成多個預測器根據錯誤分類不斷更新權重</li></ul>	<ul style="list-style-type: none"><li>易於處理非線性分類問題</li><li>高維度資料集表現良好</li></ul>
特徵重要性	V(Lasso)	V	V	
分類	SGDClassifier	RFClassifier	GBClassifier	SVC
迴歸	Lasso	RFRegressor	GBRegressor	SVR

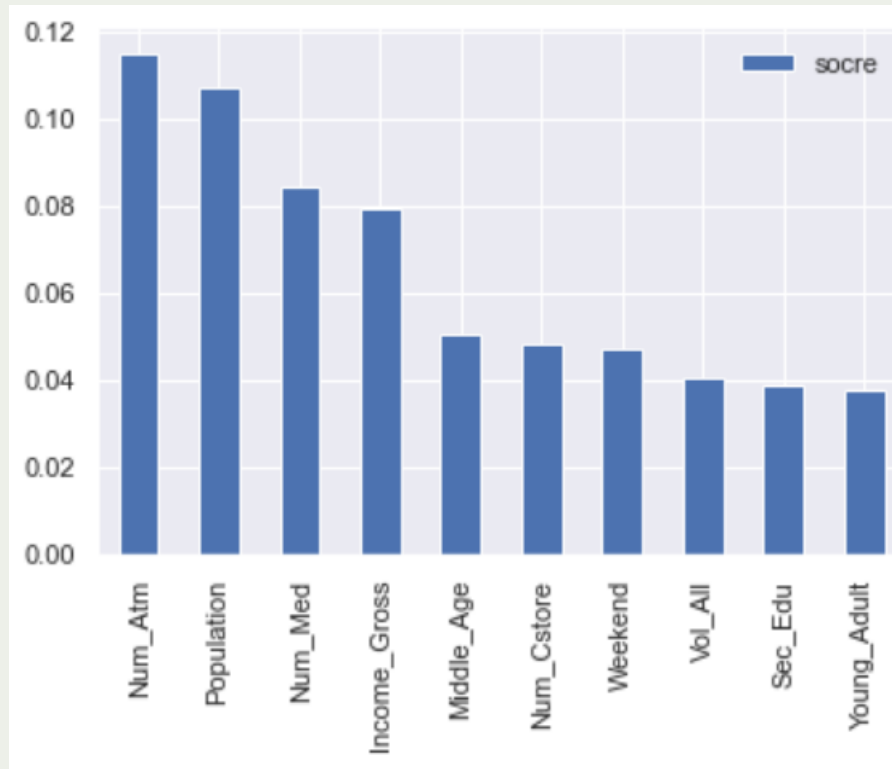
## 商家數量



### 隨機森林



### 梯度提升樹



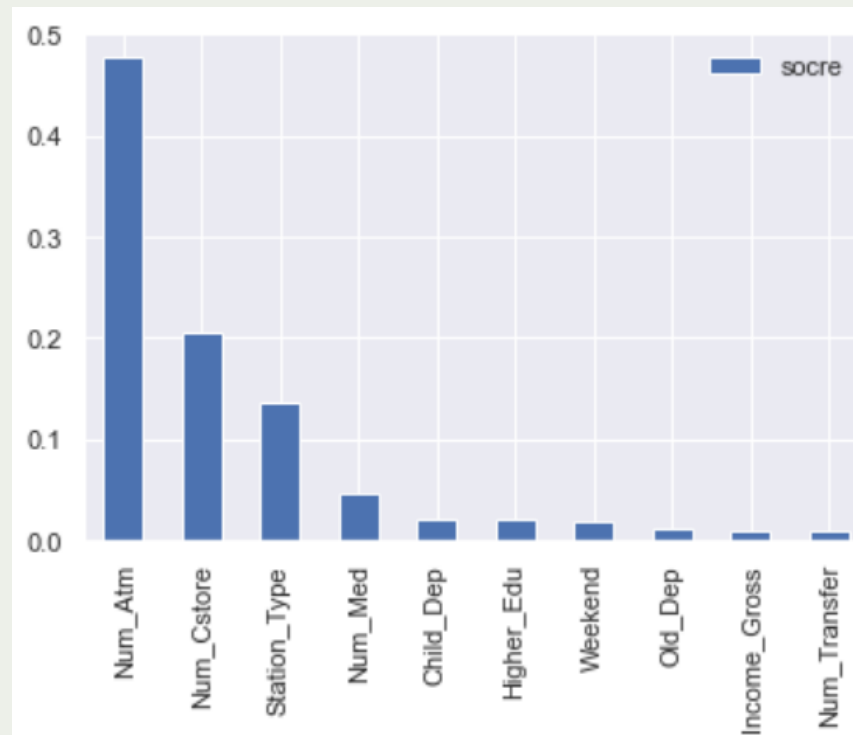
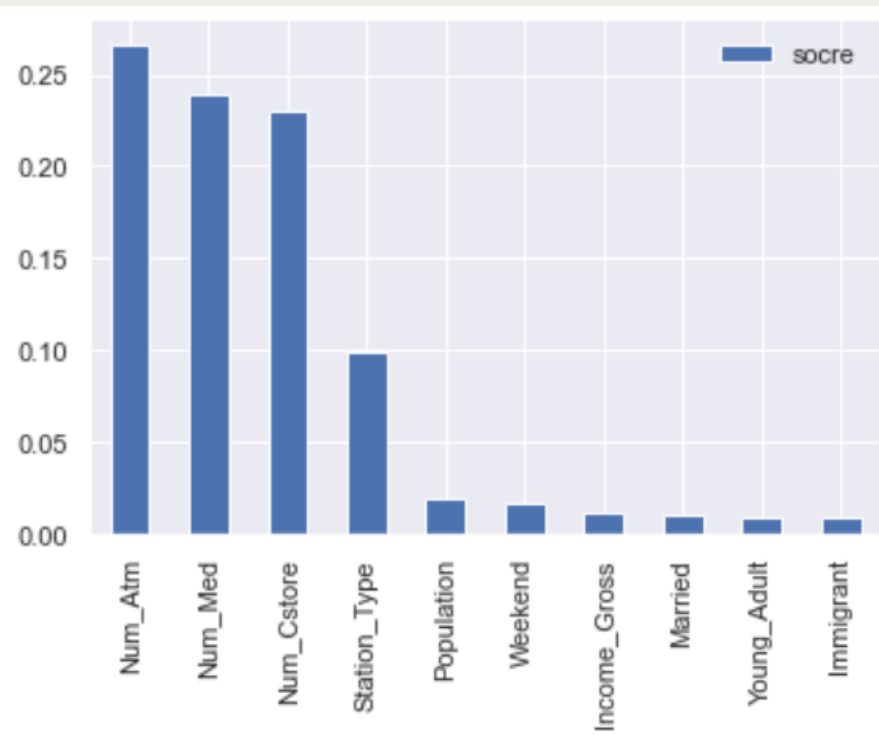
- ✓ 民生設施數量
  - 醫療院所、ATM、便利商店、公園
- ✓ 綜合所得總額
- ✓ 總人口數
- ✓ 車站流量

## 站點流量



隨機森林

梯度提升樹



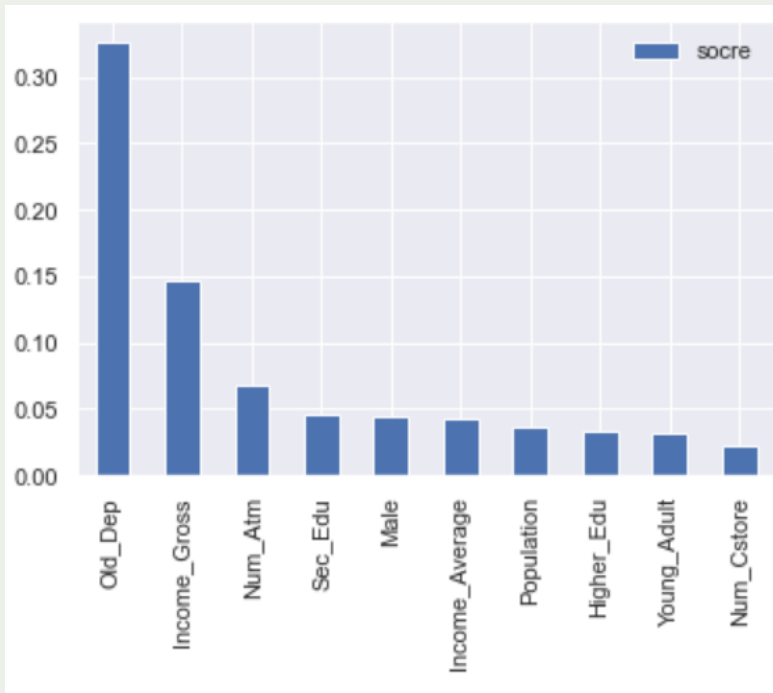
- ✓ 民生設施數量
  - ATM、醫療院所、便利商店
- ✓ 車站類型

# 租金價格

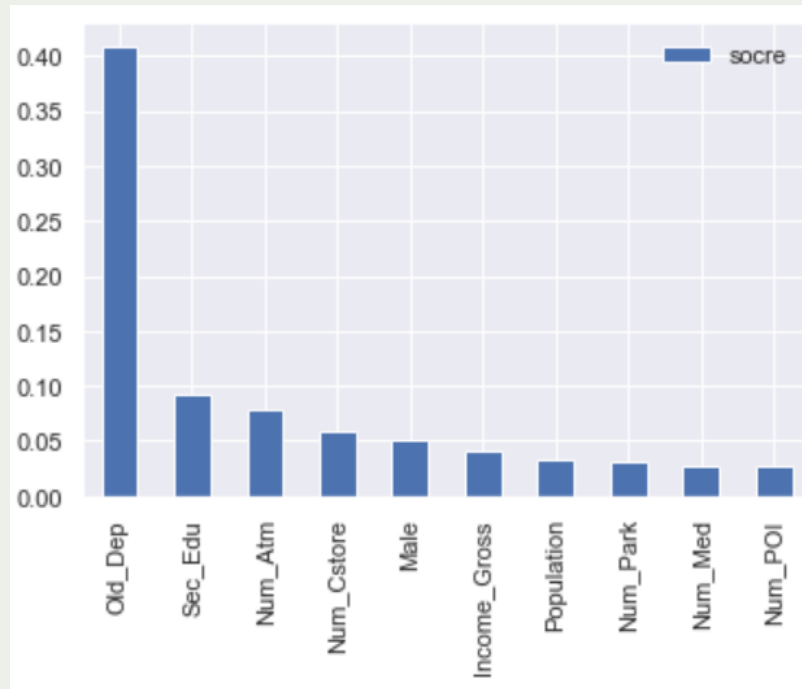


## 重要特徵值挖掘

### 隨機森林



### 梯度提升樹



- ✓ 人口特徵 (扶老比)
- ✓ 綜合所得總額
- ✓ ATM數量
- ✓ 教育 (中等學校數量)

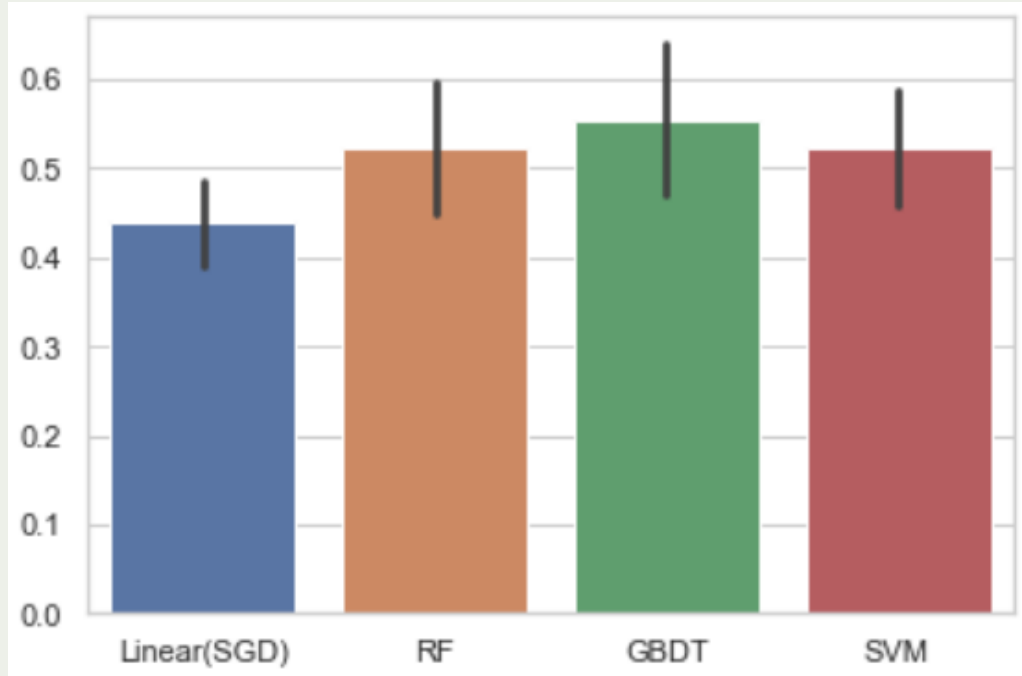
## 商家數量



- 將商家依數量等份分為5級

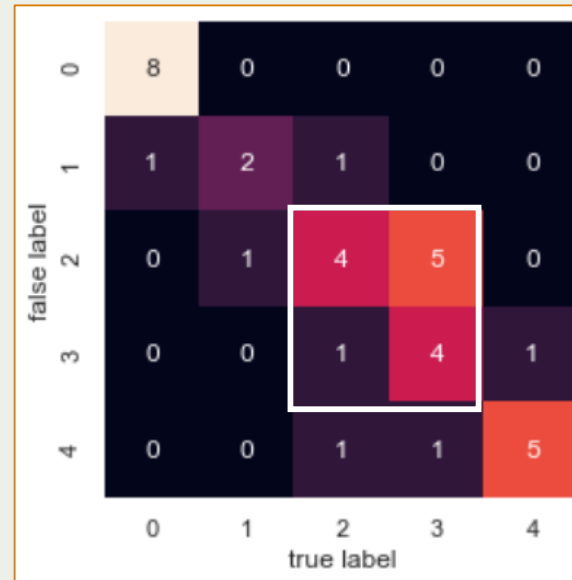
訓練資料集

Accuracy (分類正確率)



測試資料集

Accuracy = 0.66



✓ 最佳參數模型  
分類正確率  
提升至6成以上

✓ 中高數量等級  
仍有改進空間



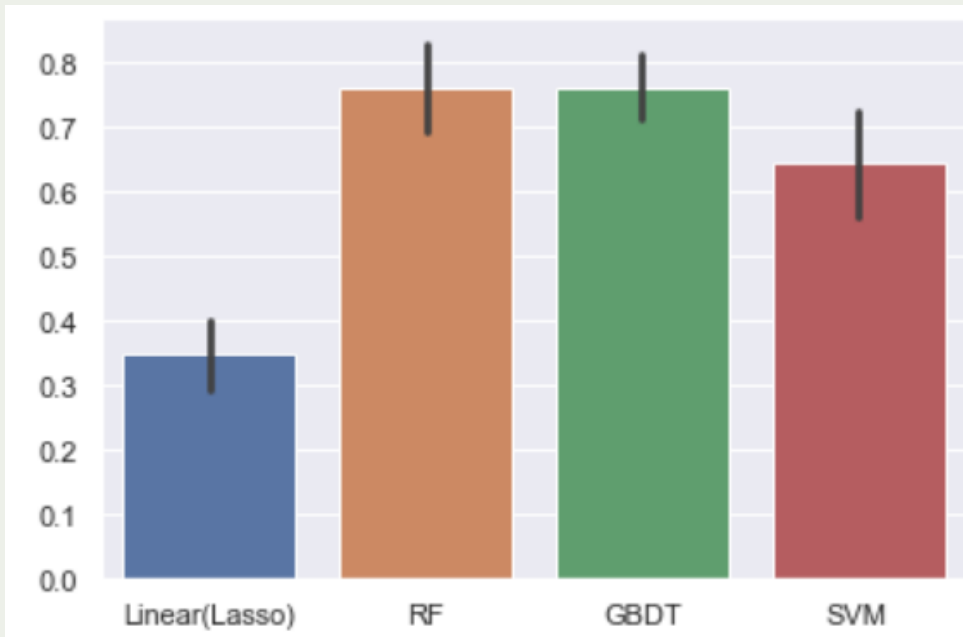
## 站點流量



- 取對數轉換(Log)縮小級距

訓練資料集

$R^2$ (模型解釋力)



測試資料集

$R^2$  : 0.70

	實際	預測
count	173.000000	173.000000
mean	29.551497	23.971770
std	41.130682	24.106028
min	0.018000	0.165752
25%	4.698000	5.244086
50%	19.011000	20.783638
75%	44.154000	33.423596
max	315.115000	175.851382

流量(千分位)

✓ 取對數轉換後  
模型預測力明顯提升

✓ 在中低運量站點  
表現較佳

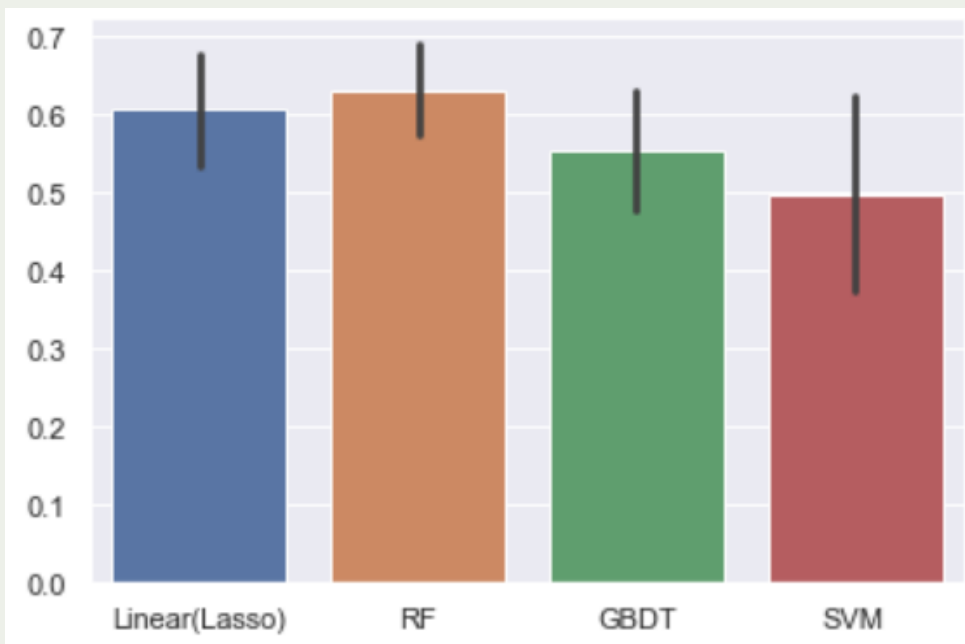
✓ 需針對鬧區與偏鄉  
選取更具解釋力之特徵

## 租金價格



## 模型績效評估

訓練資料集

 $R^2$ (模型解釋力)

測試資料集

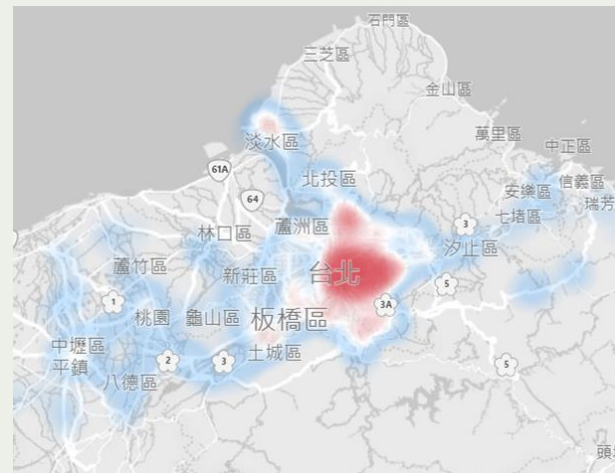
 $R^2$  : 0.78

RMSE : 323 (誤差範圍。台幣)

	實際	預測
count	265.000000	292.000000
mean	1365.550943	1349.750000
std	745.410822	666.214048
min	367.000000	495.000000
25%	900.000000	918.750000
50%	1201.000000	1164.000000
75%	1562.000000	1520.000000
max	5000.000000	4173.000000

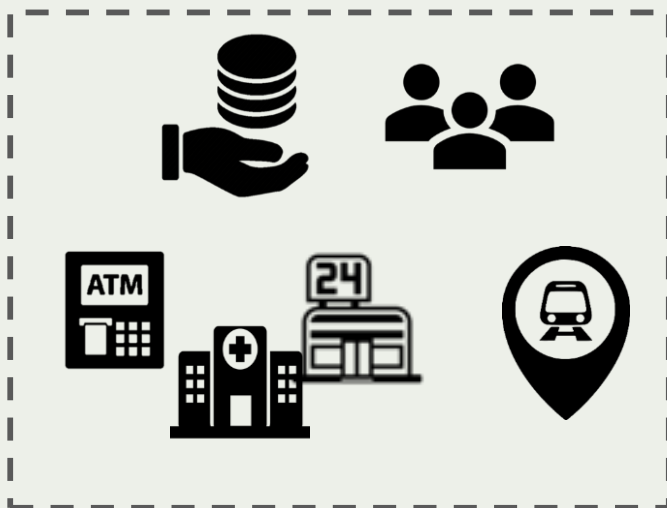
✓ 最佳參數模型  
解釋力近8成

✓ 預測資料分布  
與實際值接近



# 預測結果小結

■ 大數據結合機器學習模型，提供商家選址**參考依據**



# 模型效能改進

## ■ 商家型態分類

- 連鎖、經營主題

## ■ 獲取更直接的資訊

- 站點乘客結構、店家營業額

## ■ 再擴大特徵選取

- 如土地使用型態、建物比例、厭惡設施等

## ■ 拓展至全台軌道系統以增加樣本





## /06 專案成果及商業應用

---

■ 視覺化呈現

■ 未來展望

# 視覺化DEMO



<https://github.com/belle3759/group1>

專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

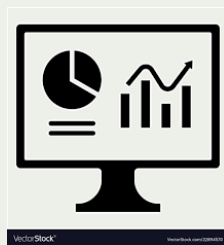
專案成果及  
商業應用

# 未來展望

目前成果



加入軌道周圍資訊，  
減少資訊不對稱



整合性儀表板，視覺化  
呈現各類分析資訊

未來



結合旅客輪廓、消費者  
資料，建立推薦系統

專案介紹

平台建置

資料收集  
及前處理

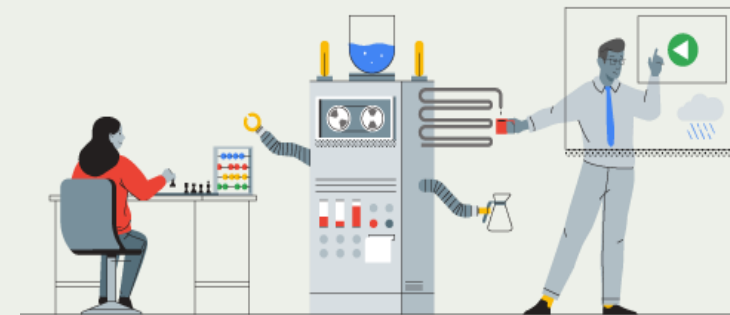
資料分析

機器學習

專案成果及  
商業應用

## 參考資料

- <https://news.cnyes.com/news/id/4307607>
- <https://data.gov.tw/>
- <https://data.moi.gov.tw/MoiOD/default/Index.aspx>
- <https://data.tycg.gov.tw/>
- <https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/>
- <https://www.invespcro.com/blog/the-use-of-machine-learning-and-artificial-intelligence-in-conversion-optimization/>
- <https://cs230.stanford.edu/ection/9/>
- [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html)



專案介紹

平台建置

資料收集  
及前處理

資料分析

機器學習

專案成果及  
商業應用



**THANK YOU**