生成式AI大數據的應用

課程講師：曾芷葳

課程設計指導：資訊部 羅順鴻 資訊長

課程演練指導：人資部 賴欣玫 課長

114年8月

目錄

[第一章 大數據觀念及應用 1](#_Toc205194754)

[1.1 大數據的5V概念 1](#_Toc205194755)

[1.2 AI如何分類與分群? 2](#_Toc205194756)

[1.3 監督 vs. 非監督式學習 3](#_Toc205194757)

[第二章 資料處理 4](#_Toc205194758)

[2.1 資料前處理觀念與乾淨資料的價值 4](#_Toc205194759)

[2.2 資料處理步驟 4](#_Toc205194760)

[2.3 Colab操作教學 5](#_Toc205194761)

[2.3.1 Google Colab是什麼? 5](#_Toc205194762)

[2.3.2 Colab有什麼限制? 5](#_Toc205194763)

[2.3.3 如何開啟 Google Colab? 5](#_Toc205194764)

[2.4 範例說明 9](#_Toc205194765)

[2.5 練習活動 10](#_Toc205194766)

[第三章 集群分析 12](#_Toc205194767)

[3.1 集群是什麼? 12](#_Toc205194768)

[3.2 為什麼分群重要? 12](#_Toc205194769)

[3.3 常見應用 12](#_Toc205194770)

[3.4 範例說明 13](#_Toc205194771)

[3.5 練習活動 16](#_Toc205194772)

[第四章 分類分析 18](#_Toc205194773)

[4.1 分類分析是什麼? 18](#_Toc205194774)

[4.2 為什麼要做分類? 18](#_Toc205194775)

[4.3 常見應用 18](#_Toc205194776)

[4.4 範例說明 19](#_Toc205194777)

[4.5 練習活動 21](#_Toc205194778)

[第五章 回歸分析與視覺化 23](#_Toc205194779)

[5.1 迴歸分析是什麼? 23](#_Toc205194780)

[5.2 為什麼會用到迴歸分析? 23](#_Toc205194781)

[5.3 資料視覺化 23](#_Toc205194782)

[5.3.1 散佈圖 23](#_Toc205194783)

[5.3.2 迴歸線圖 23](#_Toc205194784)

[5.4 常見應用 24](#_Toc205194785)

[5.5 範例說明 24](#_Toc205194786)

[5.6 練習活動 27](#_Toc205194787)

[第六章 文字探勘與文字雲 28](#_Toc205194788)

[6.1 斷詞、詞頻與詞雲簡介 28](#_Toc205194789)

[6.1.1 斷詞 28](#_Toc205194790)

[6.1.2 詞頻 28](#_Toc205194791)

[6.1.3 詞雲 28](#_Toc205194792)

[6.2 範例說明 29](#_Toc205194793)

[6.3 練習活動 33](#_Toc205194794)

[第七章 特徵值與目標變數的相關性分析 35](#_Toc205194795)

[7.1 相關性分析簡介 35](#_Toc205194796)

[7.2 分析方法與工具 35](#_Toc205194797)

[7.3 範例說明 36](#_Toc205194798)

[7.4 練習活動 39](#_Toc205194799)

[第八章 常見問答 41](#_Toc205194800)

[附錄 43](#_Toc205194801)

圖目錄

[圖2- 1 Colab畫面-1 6](#_Toc205188990)

[圖2- 2 Colab畫面-2 6](#_Toc205188991)

[圖2- 3 Colab畫面-3 6](#_Toc205188992)

[圖2- 4 Colab畫面-4 7](#_Toc205188993)

[圖2- 5 Colab畫面-5 7](#_Toc205188994)

[圖2- 6 Colab畫面-6 7](#_Toc205188995)

[圖2- 7 Colab畫面-7 8](#_Toc205188996)

[圖2- 8 Colab畫面-8 8](#_Toc205188997)

[圖2- 9 Colab畫面-9 8](#_Toc205188998)

[圖2- 10 資料前處理範例 9](#_Toc205188999)

[圖2- 11 資料前處理範例之詢問ChatGPT指令 9](#_Toc205189000)

[圖2- 12 資料前處理範例之ChatGPT生成程式碼 10](#_Toc205189001)

[圖2- 13 資料前處理範例之結果 10](#_Toc205189002)

[圖2- 14 資料前處理練習 11](#_Toc205189003)

[圖2- 15 資料前處理練習結果 11](#_Toc205189004)

圖3- 1 集群分析範例 13

圖3- 2 集群分析範例之詢問ChatGPT指令 14

圖3- 3 集群分析範例之ChatGPT生成程式碼-1 14

圖3- 4 集群分析範例之ChatGPT生成程式碼-2 15

圖3- 5 集群分析範例之結果 15

圖3- 6 集群分析練習 16

圖3- 7集群分析練習結果 17

圖4- 1 分類分析範例 19

圖4- 2 分類分析範例之詢問ChatGPT指令 19

圖4- 3 分類分析範例之ChatGPT生成程式碼-1 20

圖4- 4 分類分析範例之ChatGPT生成程式碼-2 20

圖4- 5 分類分析範例結果 21

圖4- 6 分類分析練習 22

圖4- 7分類分析練習結果 22

圖5- 1 迴歸分析範例 24

圖5- 2 迴歸分析範例之詢問ChatGPT指令 25

圖5- 3 迴歸分析範例之ChatGPT生成程式碼 26

圖5- 4 迴歸分析範例結果 26

圖5- 5 迴歸分析練習 27

圖5- 6 迴歸分析練習結果 27

圖6- 1 詞雲示意圖 29

圖6- 2 文字探勘及文字雲範例 29

圖6- 3 文字探勘及文字雲範例之詢問ChatGPT指令 30

圖6- 4 文字探勘及文字雲範例之ChatGPT生成程式碼-1 31

圖6- 5 文字探勘及文字雲範例之ChatGPT生成程式碼-2 32

圖6- 6 文字探勘及文字雲範例結果-1 32

圖6- 7 文字探勘及文字雲範例結果-2 33

圖6- 8 文字探勘及文字雲練習 33

圖6- 9 文字探勘及文字雲練習結果 34

圖7- 1 特徵值相關性分析範例 36

圖7- 2 特徵值相關性分析範例之詢問ChatGPT指令 37

圖7- 3 特徵值相關性分析範例之ChatGPT生成程式碼 38

圖7- 4 特徵值相關性分析範例結果 38

圖7- 5 特徵值相關性分析練習 39

圖7- 6 特徵值相關性分析練習結果 40

圖8- 1 常見問題 Q2 41

表目錄

[表1- 1：分類分析例子 2](#_Toc202921530)

[表1- 2：分群分析例子 2](#_Toc202921531)

[表1- 3：監督和非監督的比較表 3](#_Toc202921532)

表7- 1 相關性分析常用工具 36

# 大數據觀念及應用

## 大數據的5V概念

1. Volume（大量性）

大數據的最明顯特徵就是數據量巨大。隨著科技發展，數據產生的速度和規模呈爆炸性成長，從過去的GB、TB，到現在的PB、EB甚至ZB。這種龐大的資料量使得傳統的數據存儲和處理技術難以應付，需要分散式存儲、雲端技術及大數據架構來處理。數據量的增加，也代表著能夠捕捉到更豐富、全面的資訊。

1. Velocity（即時性）

大數據強調的是資料生成與處理的速度。現在許多資料是即時產生的，例如社交媒體貼文、感測器資料、交易記錄等，資料流動速度非常快。企業需要能即時或近即時地取得並分析這些資料，才能快速反應市場變化或監控系統狀態。這也對數據處理技術提出挑戰，例如流式計算、即時分析及自動化決策等。

1. Variety（多樣性）

大數據來自各種不同的資料來源，資料型態多元且複雜，不再只有傳統的結構化資料（例如關聯式資料庫），還包括非結構化資料（如文字、圖片、影片、音訊）、半結構化資料（如JSON、XML）以及感測器產生的時間序列資料等。這種多樣性增加了資料整合和分析的難度，但同時也帶來更多可能的洞察。

1. Veracity（真實性）

資料的品質是大數據分析成功與否的關鍵。由於資料來源複雜，存在雜訊、錯誤、不完整或不一致等問題，這些都會影響分析結果的準確性。如何驗證、清洗和管理資料品質，確保資料真實可靠，是大數據應用必須面對的挑戰。

1. Value（價值性）

大數據的最終目標是從海量且多元的資料中萃取出具有商業價值和實際應用意義的資訊。這些價值可能是提升決策效率、發現新商機、優化流程、降低成本或增強顧客體驗。只有能夠轉換成實際效益的大數據，才能成為企業競爭的利器。

## AI如何分類與分群?

* 分類

分類是指針對每筆資料貼上正確的「標籤」，也就是我們已經知道資料所屬的類別，並希望讓AI學會依據特徵自動判斷未來資料的分類。

舉例來說，若我們擁有一筆包含身高、體重等資訊的資料集，並已標示每筆資料對應的性別，便可訓練模型根據身高體重來預測性別，這就是一種監督式學習中的分類任務。

表1- 1：分類分析例子

|  |  |  |  |
| --- | --- | --- | --- |
| **資料編號** | **身高（cm）** | **體重（kg）** | **性別（標籤）** |
| 1 | 180 | 70 | 男性 |
| 2 | 165 | 55 | 女性 |
| 3 | 175 | 65 | 男性 |
| 4 | 160 | 50 | 女性 |
| 5 | 170 | 60 | 男性 |
| 6 | 158 | 48 | 女性 |
| 7 | 182 | 75 | 男性 |
| 8 | 162 | 52 | 女性 |

* 分群

分群是指AI在沒有事先給定標籤的情況下，依據資料之間的相似性，自動將資料分成不同群組。這是一種非監督式學習（Unsupervised Learning）的方式，常用於探索資料的內在結構或找出隱含的群體特徵。

舉例來說，若我們只有一組資料包含「身高」與「體重」，但不知道這些人是男性還是女性，則AI可以根據資料的分布，自動將其分為兩群（例如：一群身高體重較高、一群較低），以尋找出潛在的相似群體。這樣的方式也常應用於市場分析，例如依購物行為分群找出不同的顧客類型。

表1- 2：分群分析例子

|  |  |  |
| --- | --- | --- |
| **資料編號** | **身高（cm）** | **體重（kg）** |
| 1 | 180 | 70 |
| 2 | 165 | 55 |
| 3 | 175 | 65 |
| 4 | 160 | 50 |
| 5 | 170 | 60 |
| 6 | 158 | 48 |
| 7 | 182 | 75 |
| 8 | 162 | 52 |

## 監督 vs. 非監督式學習

|  |  |  |
| --- | --- | --- |
|  | **監督式學習** | **非監督式學習** |
| **定義** | 使用一組輸入資料和一組對應的配對已標記輸出資料來訓練模型。通常是手動完成標記。 | 可以訓練模型以發現未標記資料中的隱藏模式。 |
| **技術** | 邏輯迴歸、線性迴歸、決策樹、神經網路 | 叢集、關聯規則學習、機率密度和降維 |
| **目標** | 依據已知輸入預測輸出。 | 識別輸入資料點之間有價值的關係資訊，可以將其套用至新的輸入以得出類似的洞察。 |
| **方法** | 盡量減少預測輸出和真實標籤之間的誤差。 | 尋找資料中的模式、相似性或異常。 |

表1- 3：監督和非監督的比較表

# 資料處理

## 資料前處理觀念與乾淨資料的價值

資料前處理（Data Preprocessing）或資料預處理，是指在進行數據分析、建立模型或機器學習之前，對原始數據進行清理、轉換和整理的過程。資料前處理的目的是確保數據的品質和一致性，以及減少不確定性和噪音，讓數據更適合進行後續的分析和建模工作。

原始數據通常會有多個來源，無論是企業的ERP系統、行銷平台的數據資料，抑或業務數據指標等，都會需要先進行前處理。通常最原始的數據資料，不外乎是有錯誤的資料及格式，或是有極端值和重複的資料，這些都是會造成數據品質低落常見的原因。因此，資料前處理的重要性，就是挑出這些不合格的資料，讓龐大的數據資料成為有價值的洞察。

經過前處理的資料，能大幅提升分析的準確性與效率，讓企業能從龐雜的資料中挖掘真正有價值的洞察，進而做出更具依據的決策。

## 資料處理步驟

* 步驟一：資料清洗（Data Cleaning）

資料清洗是清理數據中的錯誤、缺失值和不一致性。企業在進行這個步驟時，必須要填補缺失值、刪除重複或無效的記錄，以及修正錯誤的數值或標籤。

* 步驟二：資料轉換（Data Transformation）

資料轉換是對數據進行轉換，使其更適合進行分析，主要的內容包括特徵縮放（例如標準化）、特徵提取（從原始數據中提取更有意義的特徵）、類別編碼（將類別型數據轉換成數值），以及降維（像是主成分分析）等。舉例來說，一間企業要分析出客戶的消費行為，但這個結果具備多項特徵，其中像是消費頻率、消費品項等資料，就必須在前期轉換為有意義的「值」，才能進行後續分析。

* 步驟三：資料整合（Data Integration）

資料集成是將來自不同來源、不同格式與不同結構的數據整合到統一的數據資料湖中。同常可以看見跨部門、跨系統間的數據來源彙整，以確保數據的一致性和可用性。然而中間可能會面臨到數據資料格式不統一的窘境，因此建議企業必須落實資料清洗的步驟，才能順利執行資料集成中。

* 步驟四：資料歸納（Data Reduction）

最後則是資料歸納，對數據資料的規模進行縮減，以減少數據的維度、大小或複雜度，同時保留主要的特徵和資訊。資料歸納有助於減少運算的成本、降低噪音的影響、提高模型的效率和可解釋性。

## Colab操作教學

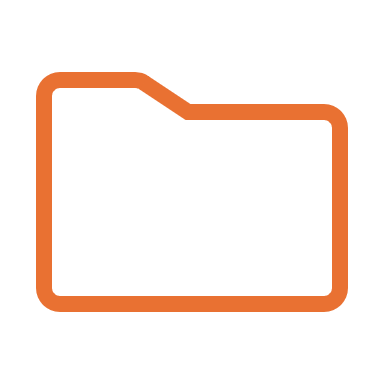
### Google Colab是什麼?

Google Colaboratory又稱為Colab是Google提供服務，讓任何人都可以透過瀏覽器撰寫以及執行 Python 程式碼。程式初學者透過 Colab 學習 Python 語言，可以省去架設環境的困擾。Colab 是基於 Jupyter Notebook 的開發環境且許多常用的套件都已經先安裝好了，因此非常適合作為資料科學開發環境。Colab 也提供免費的運算資源 (GPU)，讓我們可以加速機器學習模型的訓練。

### Colab有什麼限制?

Colab 提供的免費硬體設備 (CPU 與 GPU) 不是最好的，因此在程式的運行效能上，可能比不上你手上的電腦，但是對於資料科學領域的初學者而言已經是相當夠用的！此外，Colab 提供的記憶體 (RAM) 當然也是有限的，如果模型的參數太多或是資料集太大，就會出現記憶體不足的錯誤發生。

### 如何開啟 Google Colab?

1. 進入到任一帳號之Google Drive
2. 在我的雲端硬碟中點選「右鍵」， 選擇「更多」，再點擊「連結更多應用程式」(如圖2- 1)
3. 搜尋欄打上「Colaboratory」並在下方找到對應的名稱按下「安裝」 (如圖2- 2) (按右鍵後有看到Colaboratory，可以跳過第2、3步驟)
4. 回到我的雲端硬碟畫面點選「右鍵」，選擇「更多」，再點擊「Google Colaboratory」(如圖2- 3)
5. 點擊後就會進入到 Colab 頁面(如圖2- 4)
6. 點選右邊的連線（連線到遠端的硬體資源）(如圖2- 5)
7. 回到我的雲端硬碟，點擊「新增」(如圖2- 6)
8. 選擇「資料夾上傳」(如圖2- 7)
9. 在跳出的視窗中，找到「課程檔案」的資料夾並按下「上傳」(如圖2- 8)
10. 選擇左邊功能列的 符號(如圖2- 9)
11. 點擊 掛接至雲端硬碟(如圖2- 9)

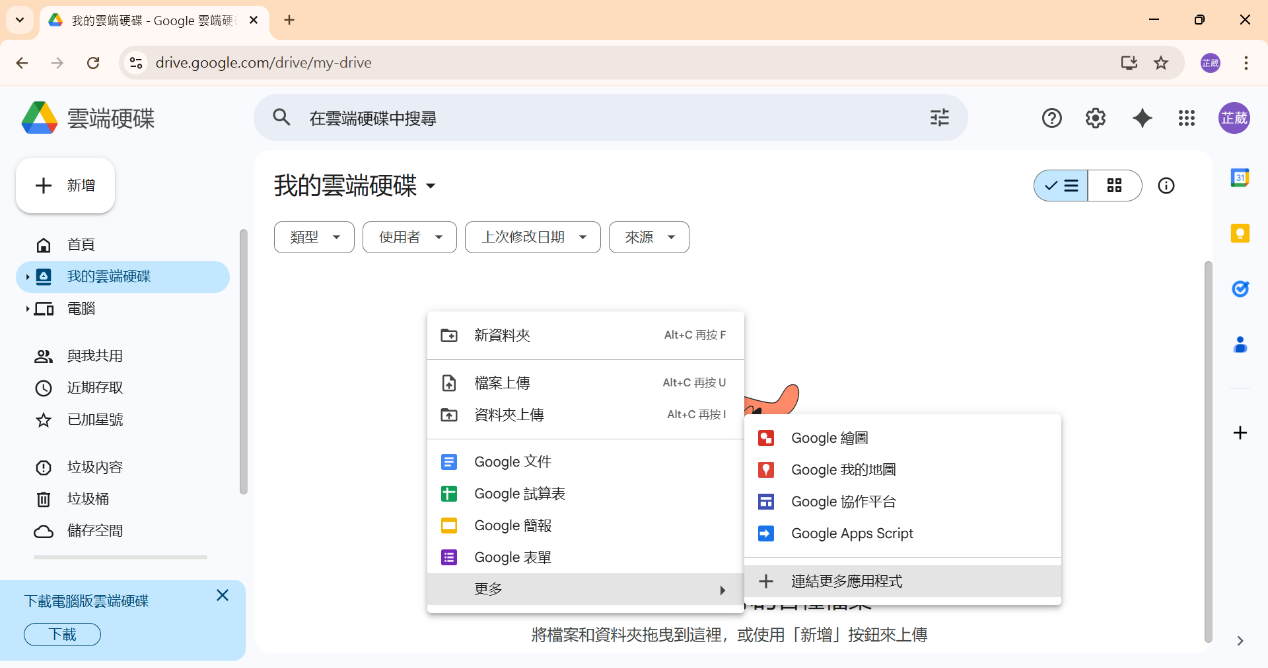


圖2- 1  Colab畫面-1

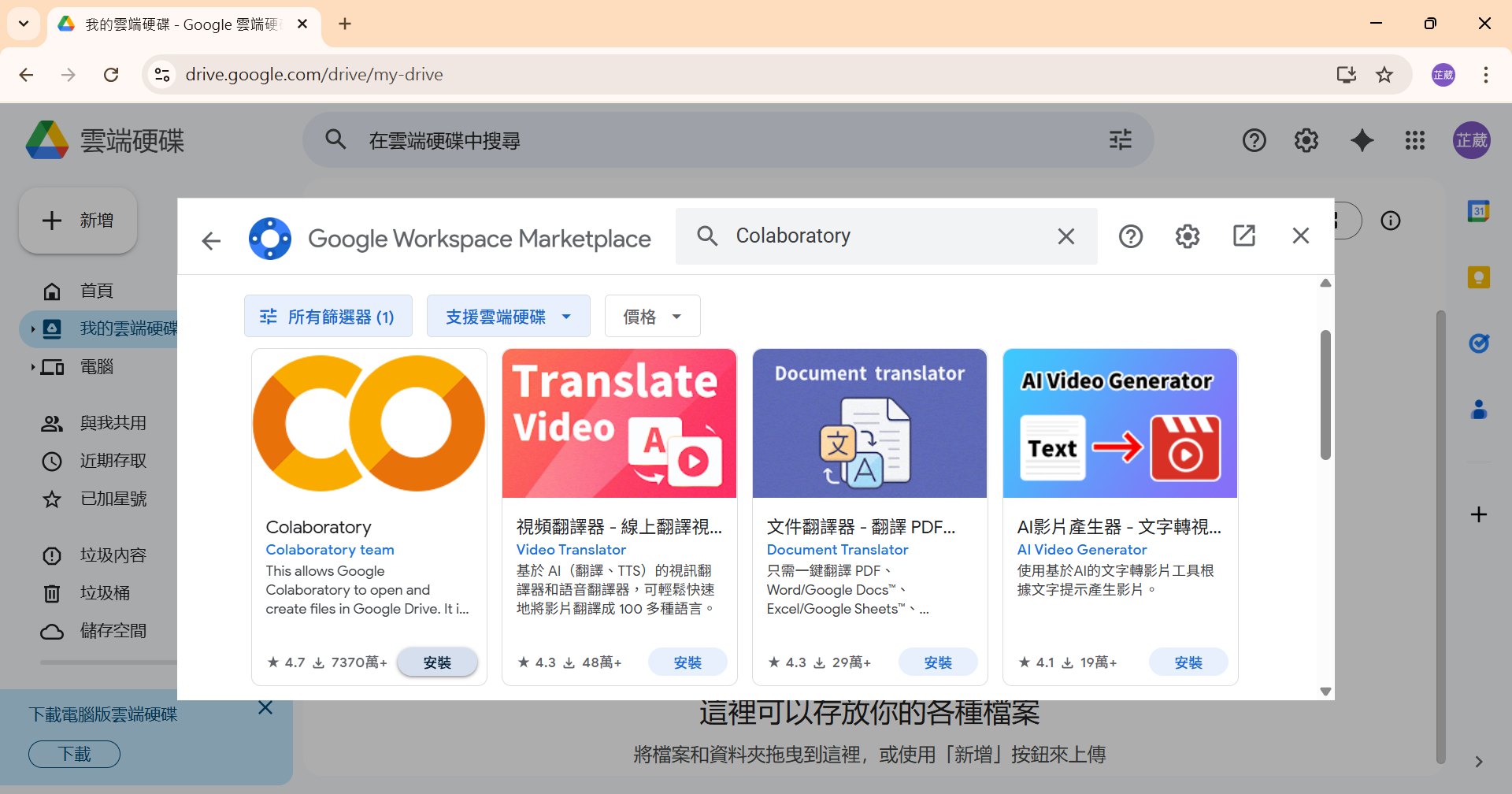


圖2- 2 Colab畫面-2

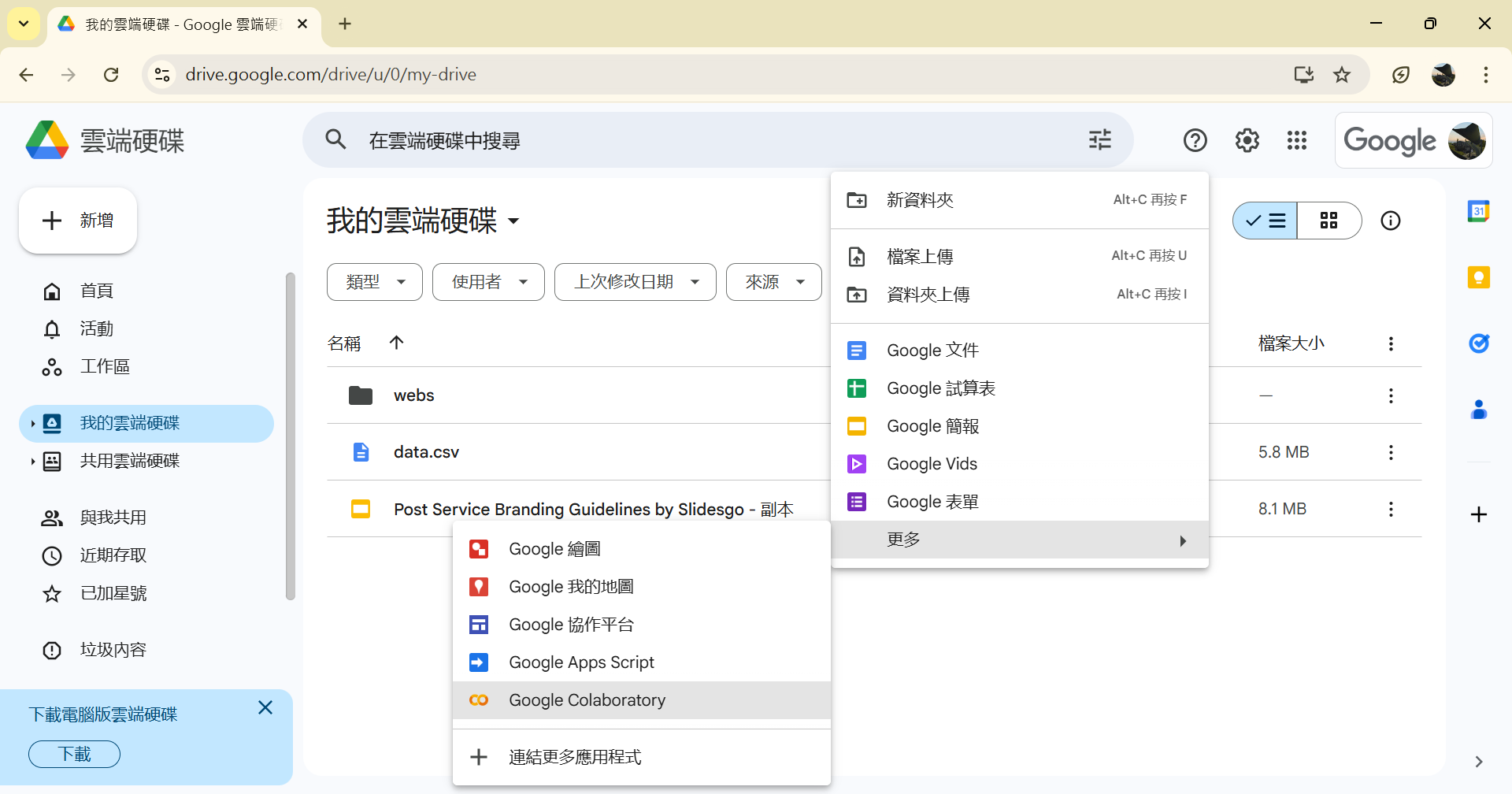


圖2- 3 Colab畫面-3

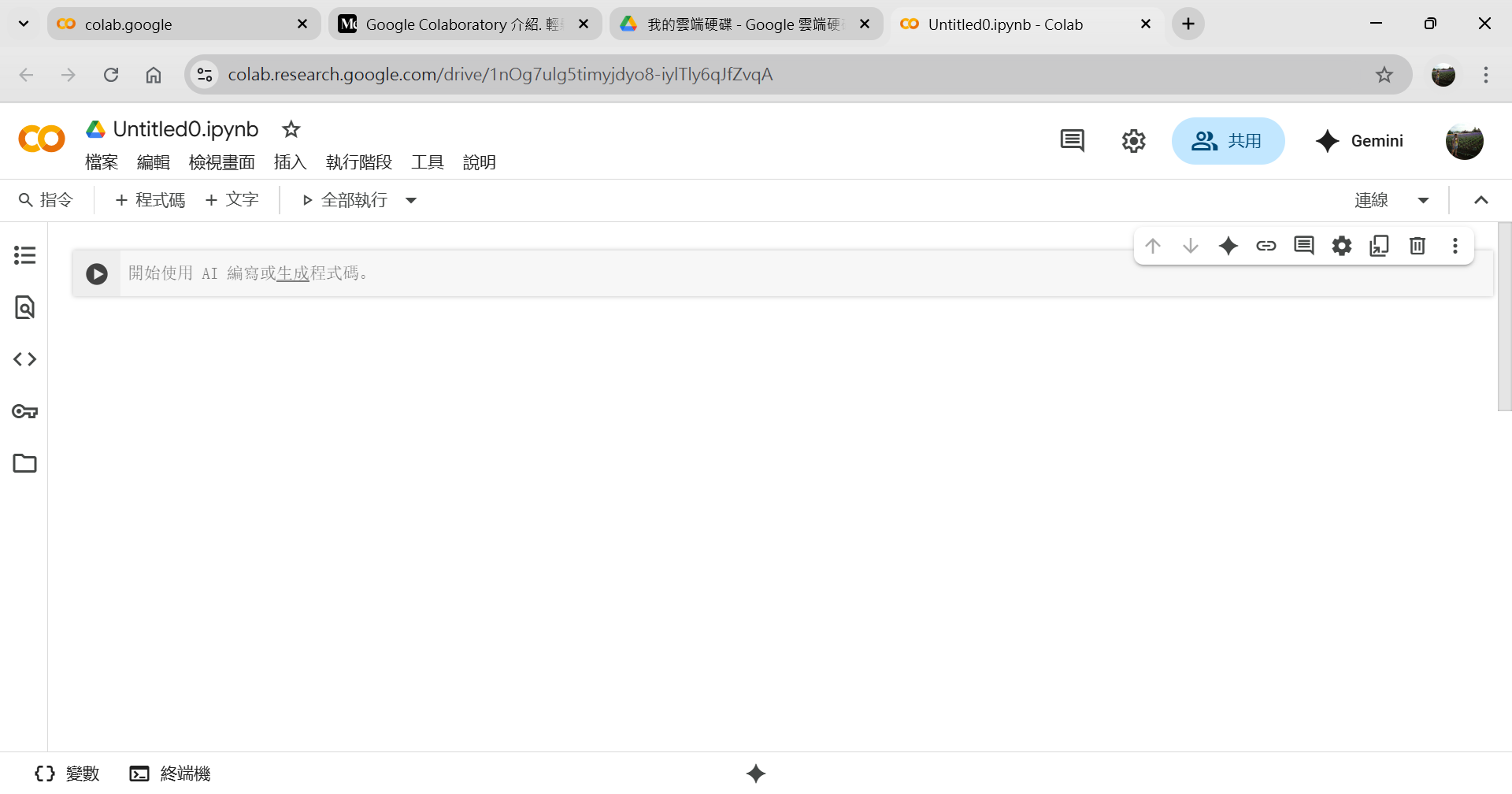


圖2- 4 Colab畫面-4

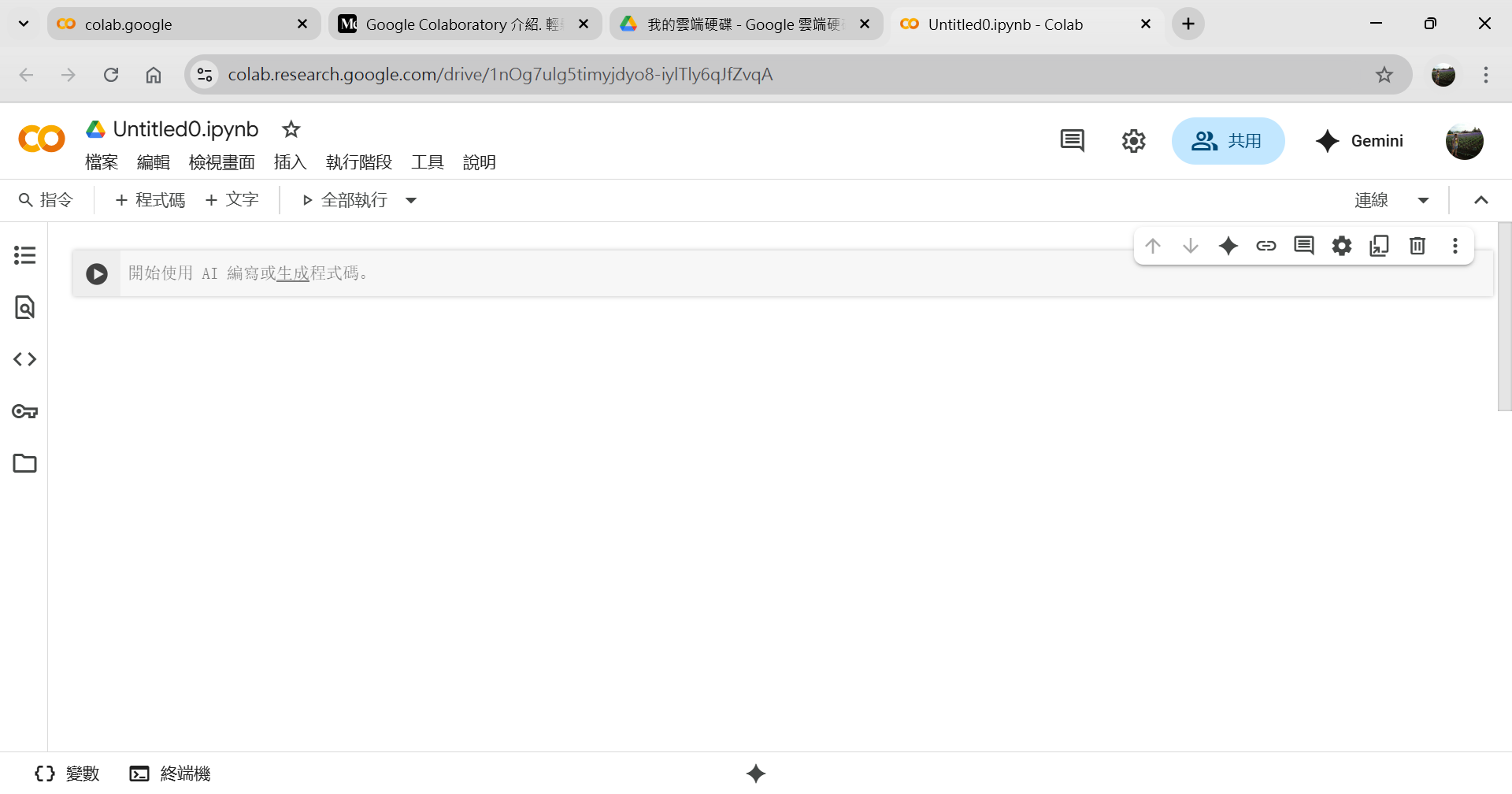


圖2- 5 Colab畫面-5



圖2- 6 Colab畫面-6 

圖2- 7 Colab畫面-7

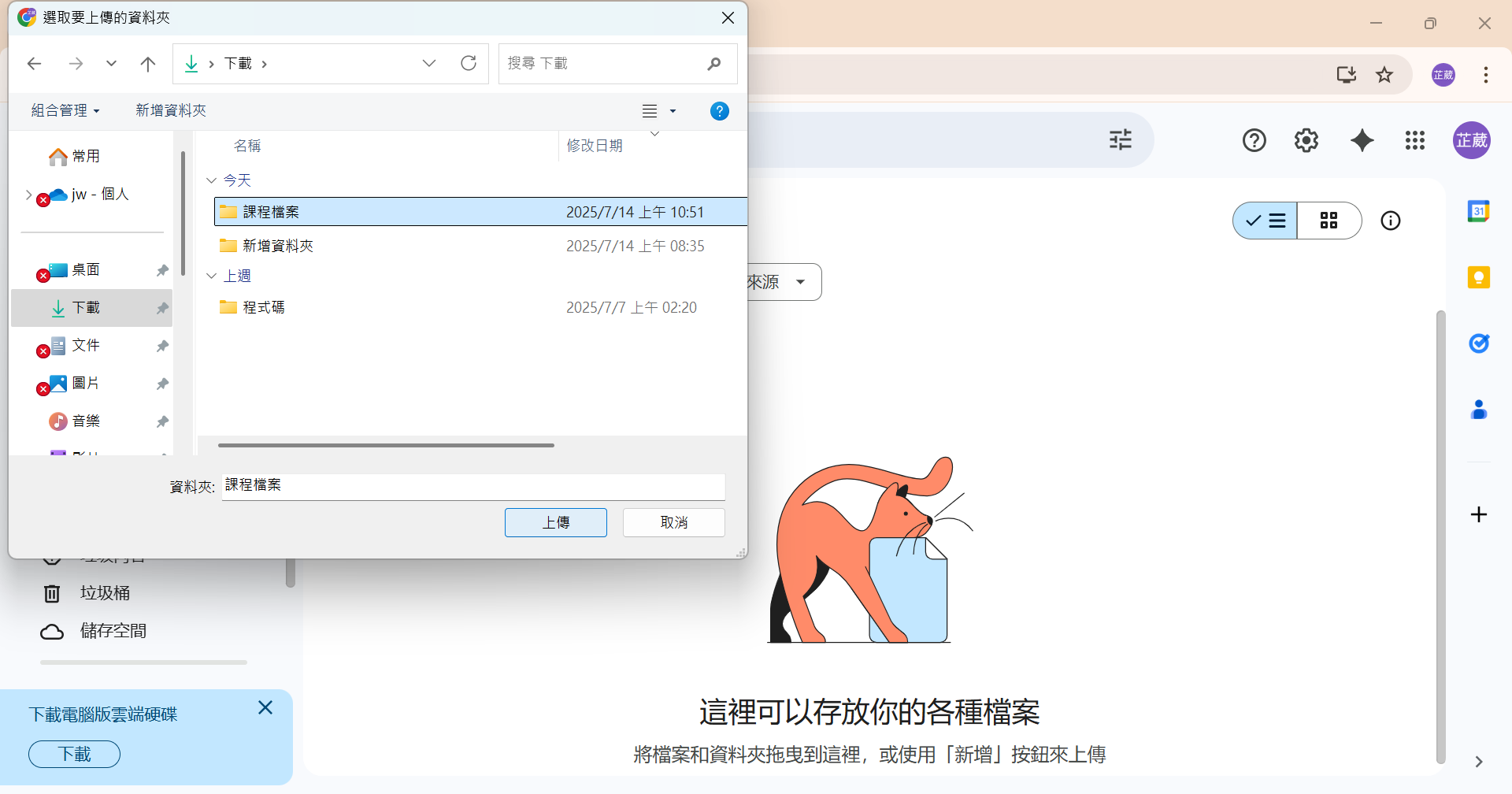


圖2- 8 Colab畫面-8

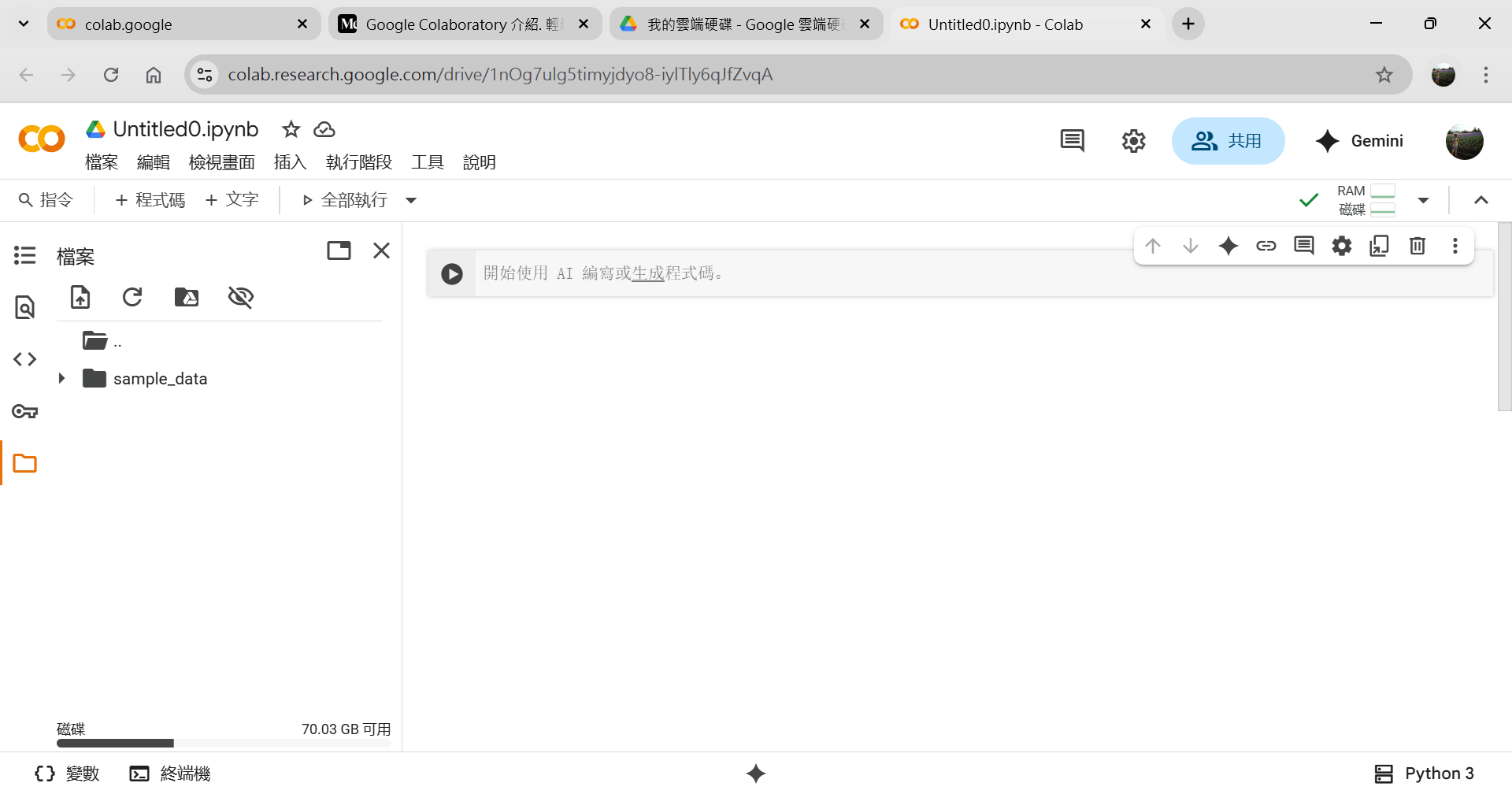


圖2- 9 Colab畫面-9

## 範例說明

* 情境說明

公司準備彙整來自不同部門業務人員所紀錄的客戶訂單，資料來源包含手動輸入的 Excel 表單，也有從舊系統匯出的資料。因為沒有統一格式規範，導致資料裡出現了：不同的日期格式有些訂單金額缺漏甚至有輸入錯誤或亂碼符號。以下是公司彙整的客戶訂單：

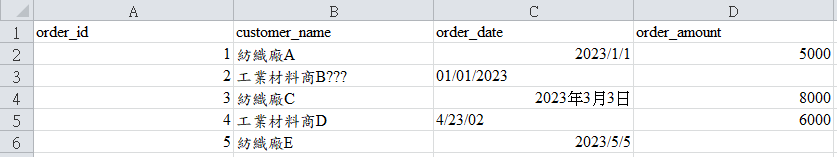


圖2- 10 資料前處理範例

* 操作步驟

1. 將檔案提交給ChatGPT，並輸入提示詞：「公司彙整資料因為沒有統一格式，導致每筆資料格式不太一樣，請幫我把日期格式統一，有些訂單金額缺漏就先以平均值補上保留至個位數，廠商名稱先保留中英文，其餘符號都先刪除，最後要儲存結果，請直接產生完整 Python 程式碼」
2. 將ChatGPT提供的程式碼貼至colab執行，就能看到解決完成的Excel檔案
3. \*檔案為「資料清洗一.xlsx」

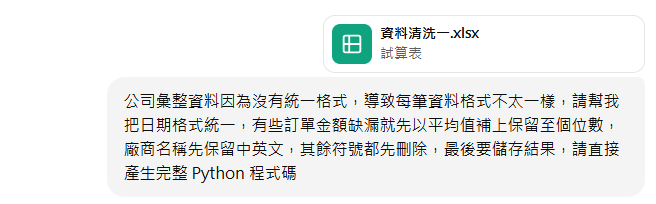


圖2- 11 資料前處理範例之詢問ChatGPT指令



圖2- 12 資料前處理範例之ChatGPT生成程式碼



圖2- 13 資料前處理範例之結果

## 練習活動

* 情境說明：

公司供應鏈部門收到來自不同供應商所回報的原物料進貨資訊，資料來源包含手動輸入的 Excel 表單，也有從系統擷取的資料。由於格式未統一，導致以下問題：

1. 日期格式不一致
2. 單位錯誤或混亂（例如重量單位）
3. 數量欄位遺漏
4. 出現亂碼或多餘符號

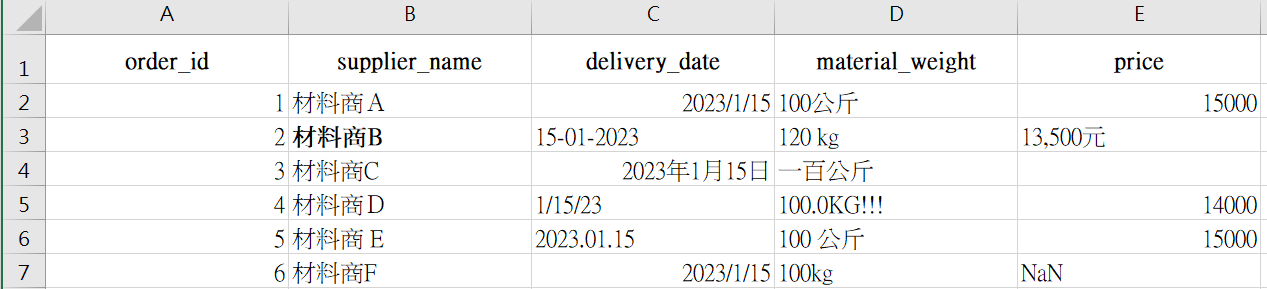


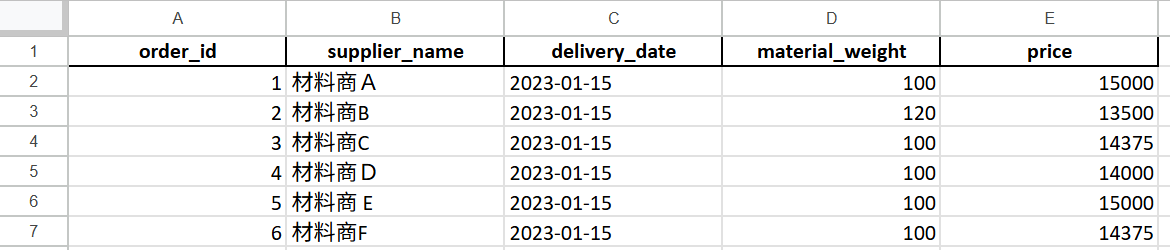
圖2- 14 資料前處理練習

* 任務

利用詢問ChatGPT生成Python程式碼來解決格式不一致的狀況。

\*檔案為「資料清洗二.xlsx」

* 結果

圖2- 15 資料前處理練習結果

# 集群分析

## 集群是什麼?

集群（Clustering）是一種非監督式學習方法，目的是將資料依照特徵相似性進行分組。每一組稱為一個群集（Cluster），群內的資料彼此相似，而群與群之間差異較大。

## 為什麼分群重要?

* 發掘潛在結構：當資料沒有明確標籤時，透過分群能找出資料中的隱含結構或類型。
* 精準行銷：可以根據顧客分群結果，針對不同族群設計個別行銷策略。
* 提升決策效率：幫助公司快速理解資料，進而做出更明智的商業決策。
* 資料簡化：將大量數據歸類，可有效降低分析複雜度。
* 個人化推薦：像是音樂或購物平台，透過使用者行為分群，提供更符合喜好的推薦。

## 常見應用

* 顧客分群（根據消費頻率與金額）

透過分析顧客的購買行為，如購買頻率、每次交易金額與總消費金額等指標，企業可運用集群分析將顧客分為不同群體，如高價值顧客、潛力顧客或低活躍顧客。這有助於制定更精準的行銷策略，如針對高價值客群推送專屬優惠或對低活躍群體進行再行銷以提升黏著度。

* APP 使用行為分群

根據用戶在 APP 中的操作行為（例如使用功能的頻率、停留時間、活躍時間區段等），可以將使用者分群，例如重度使用者、新手用戶或功能特定使用者。這類分群結果能協助產品團隊更了解用戶特性，優化使用者體驗，並針對不同群體推送個性化內容或功能推薦，提高使用率與留存率。

## 範例說明

* 情境說明

這份檔案記錄了不同客戶的訂單金額與產品偏好，公司希望透過這些資料進行客戶分群，找出具有相似消費行為或產品需求的群體，作為後續行銷策略與產品規劃的依據。

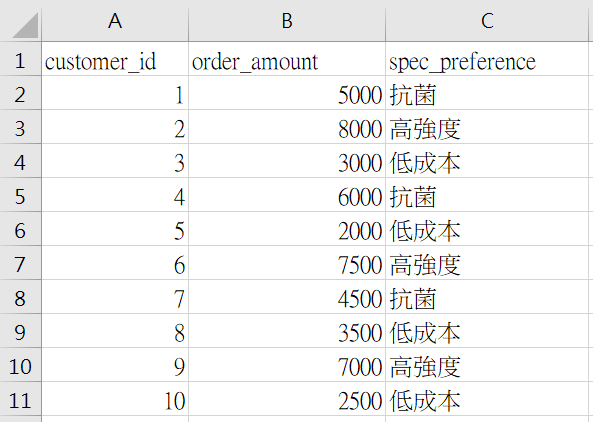


圖3- 1 集群分析範例

* 操作步驟

1. 將檔案提交給ChatGPT，並輸入提示詞：「這份檔案記錄了不同客戶的訂單金額與產品偏好，公司希望找出具有相似消費行為或產品需求的群體，作為後續行銷策略與產品規劃的依據， 請將檔案的資料進行分群並繪製圖表，並直接產生完整Python程式碼。 圖表中文字型已放在 /content/drive/MyDrive/課程檔案/NotoSansCJKtc-Regular.otf，程式碼中請用 fontproperties 明確指定標題與軸標籤的中文字型。請直接回傳完整程式碼，不要再確認字型路徑或環境。」
2. 將ChatGPT提供的程式碼貼至colab執行，就能看到分群好的圖表
3. \*檔案為「集群分析一.xlsx」



圖3- 2 集群分析範例之詢問ChatGPT指令

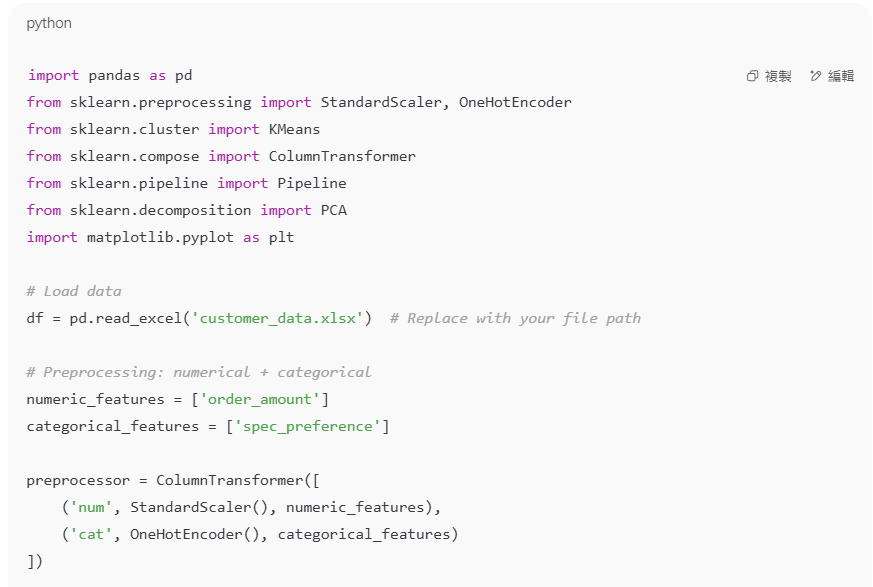


圖3- 3 集群分析範例之ChatGPT生成程式碼-1



圖3- 4 集群分析範例之ChatGPT生成程式碼-2

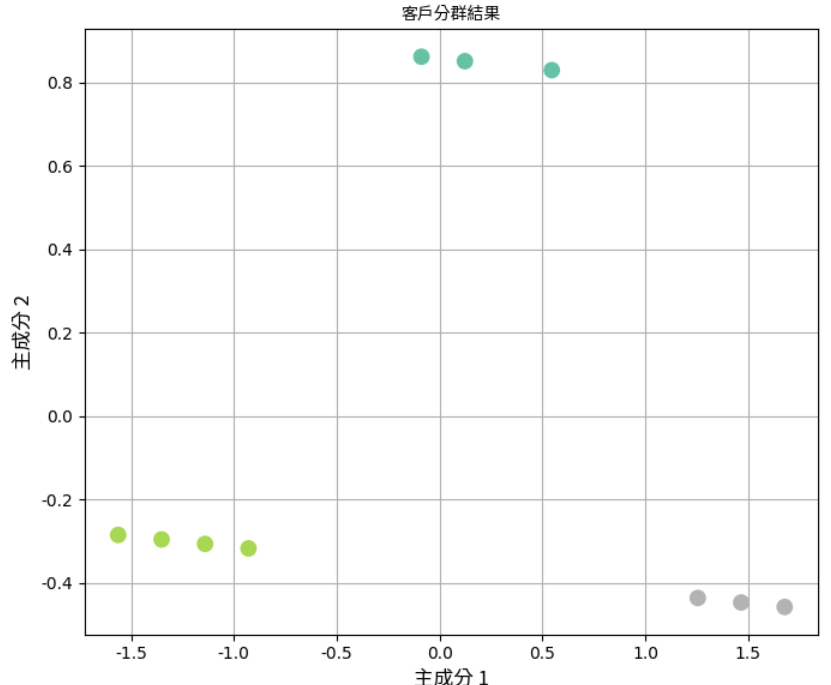


圖3- 5 集群分析範例之結果

## 練習活動

* 情境說明

公司的人資部門希望提升內部訓練課程的參與率與滿意度，於是收集了員工過去參加訓練的紀錄，想透過分群分析找出員工在訓練參與上的偏好，以便後續規劃出更符合不同群體需求的課程內容、精準發送訓練邀請，並預測未來哪些人更有可能積極參與課程，達到提升整體培訓成效的目的。

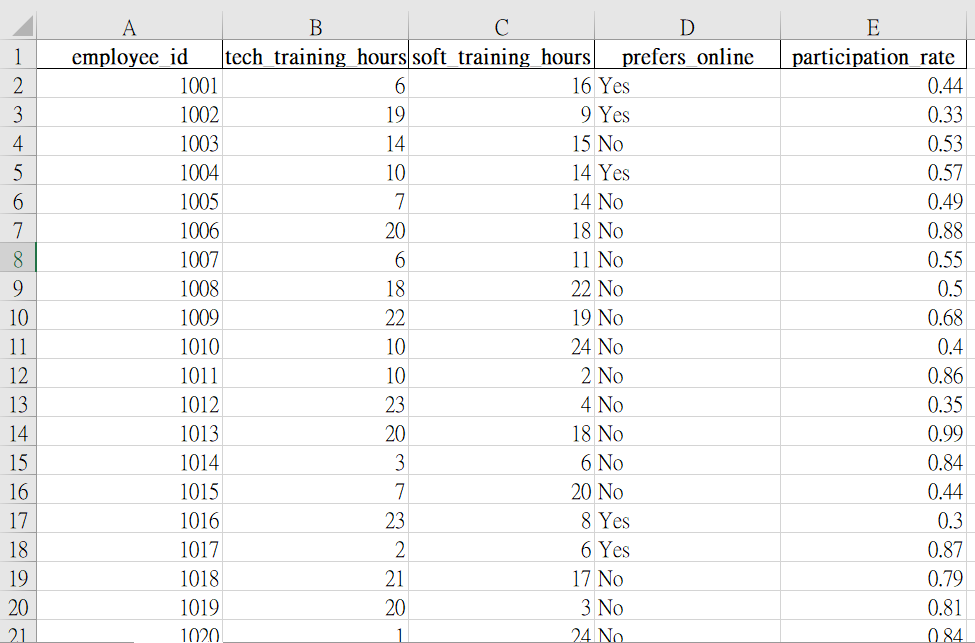


圖3- 6 集群分析練習

* 任務

利用詢問ChatGPT生成Python程式碼來呈現分群圖表供人資部門使用。

\*檔案為「集群分析二.xlsx」

* 結果

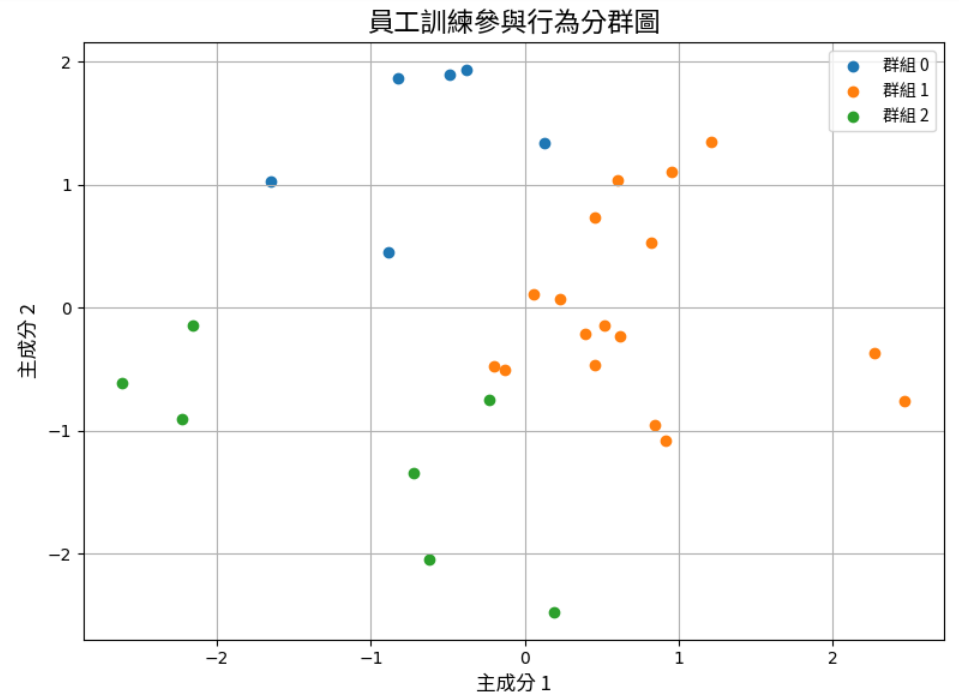


圖3- 7集群分析練習結果

# 分類分析

## 分類分析是什麼?

分類分析是一種常見的監督式學習方法，主要目的是把資料「分到正確的類別中」。當我們已經知道每筆資料屬於哪一類，現在想訓練一個模型，幫助我們判斷新資料應該屬於哪一類。

## 為什麼要做分類?

在資料分析與應用中，分類是一項關鍵的技術，具有多方面的價值與應用效益。分類可以協助我們快速判斷資料的類別，有效整理與理解大量資料，使後續處理與分析更加有條理。透過合適的分類模型與方法，能夠提升預測的準確性與效率，無論是在商業決策、醫療診斷或風險控管等領域，都能發揮明顯效益。

此外，分類也有助於減少人為判斷所造成的錯誤與主觀偏誤，讓分析過程更加客觀與一致，提升決策的公正性與可靠性。最後，透過分類技術的導入，可以讓決策流程更加自動化並具備可規模化的特性，降低人力成本並提升整體作業效率，進而加速企業或組織的數位轉型。

## 常見應用

* 客戶類型分類

透過分析顧客的年齡、性別、購買習慣、使用產品的方式等資料，可運用分類模型將顧客歸類為不同的客戶類型，例如一般用戶、忠誠顧客或高風險客戶。這有助於企業針對不同客戶群體設計更有效的行銷策略，提升服務品質與客戶滿意度。

* 信用卡申請是否核准

銀行或金融機構可根據申請人提供的資訊（如年齡、收入、工作年資、信用評分等）建立分類模型，預測其信用卡申請是否會被核准。這類模型有助於自動化申請流程、提升審核效率，同時也可作為風險控管的重要工具，以降低違約風險。

## 範例說明

* 情境說明

這是一家零售公司彙整的客戶資料，包含基本資訊與消費行為，並依據顧客價值分為高、中、低三種類型。目的是用來預測新客戶可能屬於哪一類型，協助行銷分眾與策略制定。

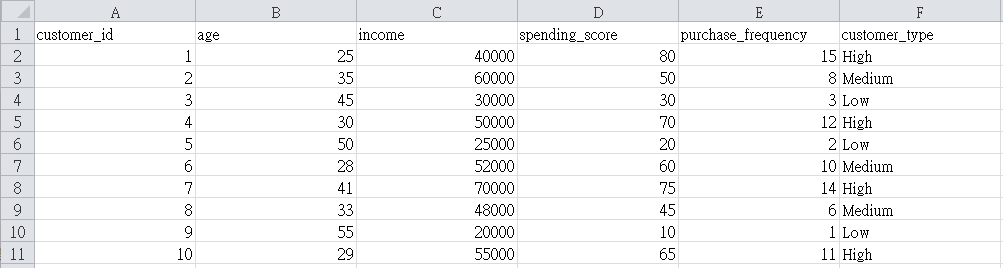


圖4- 1 分類分析範例

* 操作步驟

1. 將檔案提交給ChatGPT，並輸入提示詞：「這是一家零售公司彙整的客戶資料，請使用決策樹分類以協助行銷分眾與策略制定， 並直接產生python程式碼，圖表中文字型已放在 /content/drive/MyDrive/課程檔案/NotoSansCJKtc-Regular.otf ，程式碼中請用 fontproperties 明確指定標題與軸標籤的中文字型。請直接回傳完整程式碼，不要再確認字型路徑或環境。」
2. 將ChatGPT提供的程式碼貼至Colab執行，就能看到分類方法的決策樹
3. \*檔案為「分類分析一.xlsx」



圖4- 2 分類分析範例之詢問ChatGPT指令



圖4- 3 分類分析範例之ChatGPT生成程式碼-1



圖4- 4 分類分析範例之ChatGPT生成程式碼-2

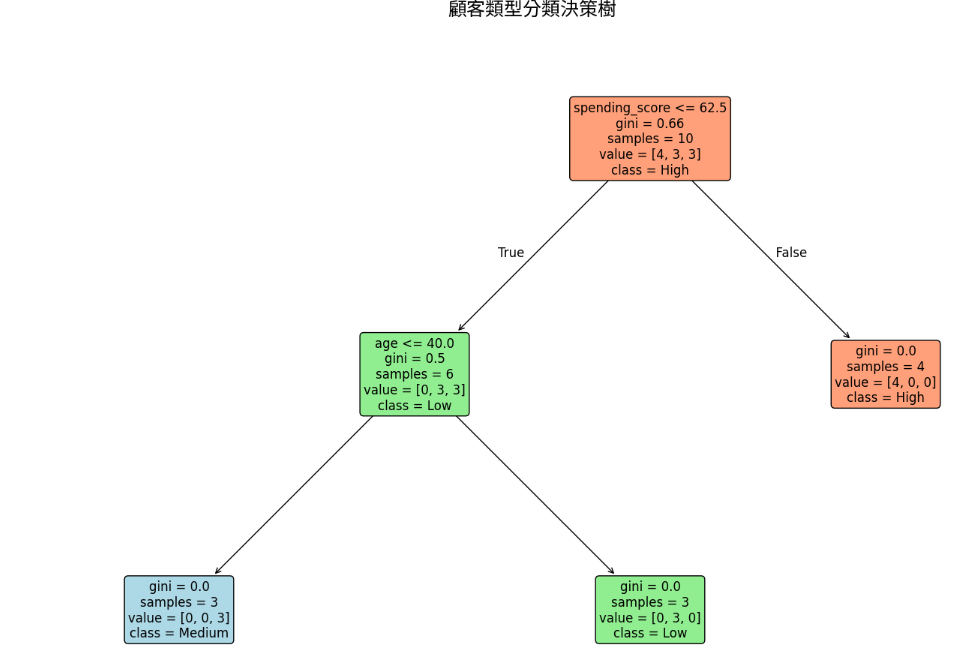


圖4- 5 分類分析範例結果

* 決策數

1. 簡介

決策樹是一種直觀的機器學習算法，它模擬人類的決策過程，透過一系列「是/否」問題來進行分類或預測。

1. 組成元素

* 根節點：樹的起始點，包含所有數據
* 內部節點：表示決策條件（如「年齡 > 30」）
* 分支：連接節點，表示決策結果（True/False）
* 葉節點：最終的分類結果

## 練習活動

* 情境說明

公司設備維修部門希望更有效地管理設備維護資源，因此收集了各設備的維修紀錄。希望將這些資料進行分群分析，以協助公司辨識出常故障且老舊、可能需要汰換的設備，以及穩定運作但使用年限尚短的新設備，同時也能觀察不同設備類型之間在維護行為上的差異。

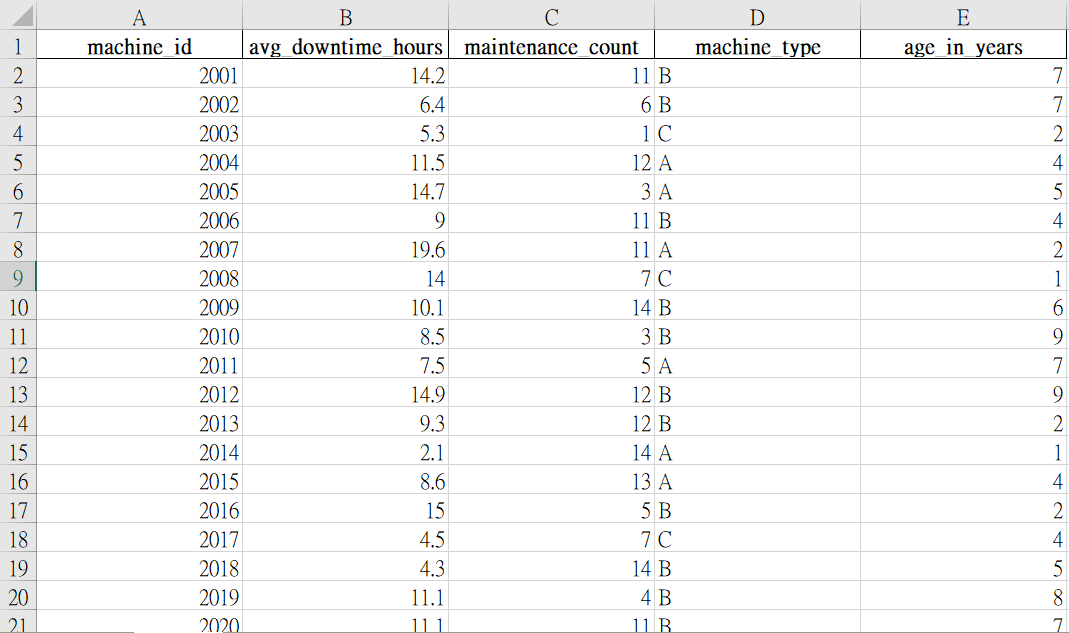


圖4- 6 分類分析練習

* 任務

利用詢問ChatGPT生成Python程式碼來呈現決策樹以提升後續設備維修效率。

\*檔案為「分類分析二.xlsx」

* 結果

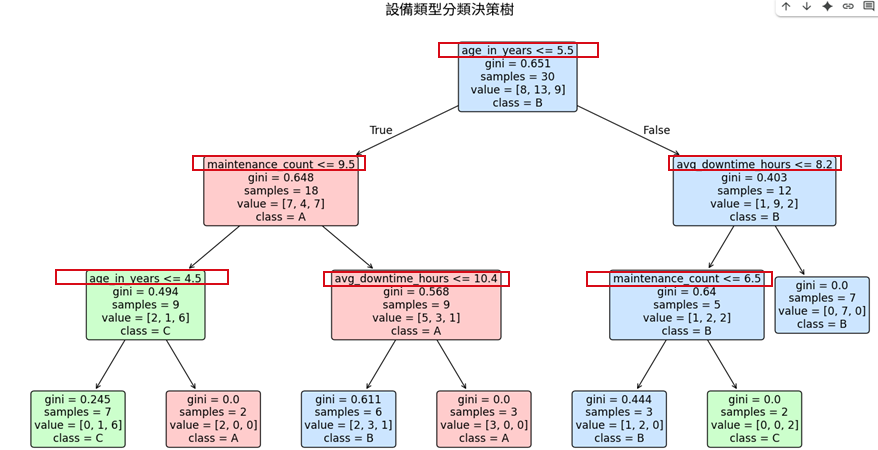


圖4- 7分類分析練習結果

# 回歸分析與視覺化

## 迴歸分析是什麼?

迴歸分析是在找出「兩個或多個變數之間的關係」，特別是當我們想要預測一個數值時，就會用到它，這種方法特別適合用來預測連續型數值，例如，我們想知道「廣告費用」會不會影響「銷售額」，就可以透過迴歸來看看兩者有沒有明確的關聯。

## 為什麼會用到迴歸分析?

* 想預測未來：像是預測房價、預測業績、甚至預測空氣品質都能用迴歸處理。
* 看哪個因素最關鍵：不只是預測，我們也能知道「哪些變數影響最大」。
* 讓決策更有憑據：不是靠感覺，而是靠數據說話，幫助做出更合理的決定。

## 資料視覺化

### 散佈圖

* 定義與功能

散佈圖是用來顯示兩個連續變數之間的關係。圖中的每一個點代表一筆觀測資料，其中橫軸表示一個變數，縱軸表示另一個變數。

* 用途與應用

1. 作為資料探索的第一步，觀察變數間是否有明顯的關聯性。
2. 看出變數間的正相關、負相關或無相關。
3. 協助判斷是否可以進行線性迴歸分析，若點大致排列成斜線狀，代表存在線性關係，適合建立模型。

### 迴歸線圖

* 定義與功能

迴歸線圖是在散佈圖的基礎上，加上一條「最佳線性擬合線」，也就是根據迴歸模型所建立的預測線。這條線代表了「平均趨勢」。

* 用途與應用

1. 直觀呈現變數之間的數學關係。
2. 可幫助了解當自變數變化時，應變數的預測值如何改變。
3. 讓使用者更清楚資料的整體走勢與預測模型的效果。

## 常見應用

* 客戶流失預測

企業可利用歷史資料中的顧客行為指標（如近期期交易次數、客服互動紀錄、使用天數等），建立迴歸模型預測未來一段時間內顧客流失的機率或時間。例如，預測某顧客在未來30天內流失的風險分數。此類預測有助於企業及早介入，例如推送挽留方案或提供優惠，以降低流失率並維持客戶關係。

* 採購金額預測

根據過去的採購記錄、產品需求、季節性因素及市場價格波動等資料，可透過迴歸模型預測未來的採購金額或需求量。這有助於企業在庫存管理與資源分配上作出更精確的規劃，避免庫存過多或短缺的情況，提升營運效率與成本控制能力。

## 範例說明

* 情境說明

為掌握企業規模與客戶購買力之間的關係，公司整理了歷史交易資料，希望透過迴歸分析評估企業規模對訂單金額的影響，作為客戶評估與策略制定的依據。

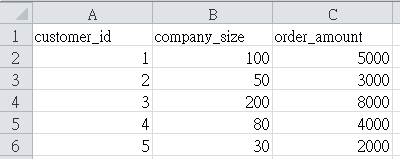


圖5- 1 迴歸分析範例

* 操作步驟

1. 將檔案提交給ChatGPT，並輸入提示詞：「為掌握企業規模與客戶購買力之間的關係，公司整理了歷史交易資料，請用回歸線散佈圖來評估企業規模對訂單金額的影響，並直接產生完整Python程式碼。 圖表中文字型已放在 /content/drive/MyDrive/課程檔案/NotoSansCJKtc-Regular.otf，程式碼中請用 fontproperties 明確指定標題與軸標籤的中文字型。請直接回傳完整程式碼，不要再確認字型路徑或環境。」
2. 將ChatGPT提供的程式碼貼至Colab執行，就能看到線性迴歸散佈圖以利後續客戶評估與策略制定的依據
3. \*檔案為「迴歸分析一.xlsx」



圖5- 2 迴歸分析範例之詢問ChatGPT指令

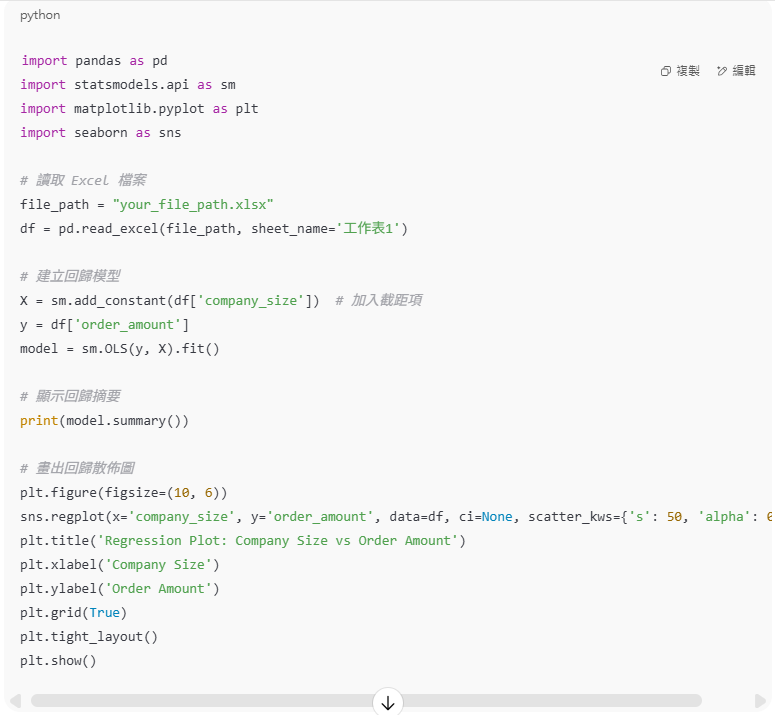


圖5- 3 迴歸分析範例之ChatGPT生成程式碼

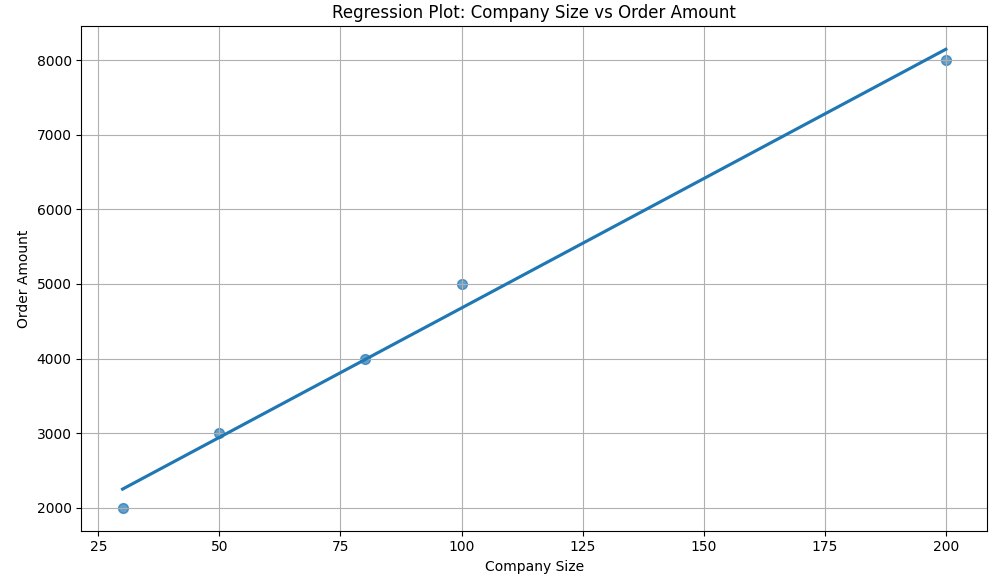


圖5- 4 迴歸分析範例結果

## 練習活動

* 情境說明

一家產品銷售公司每月投入不同金額在三種廣告渠道：電視、廣播與社群媒體，並紀錄當月的總銷售額。公司希望透過迴歸分析找出哪種廣告對銷售最有效，幫助未來更有效配置行銷預算。

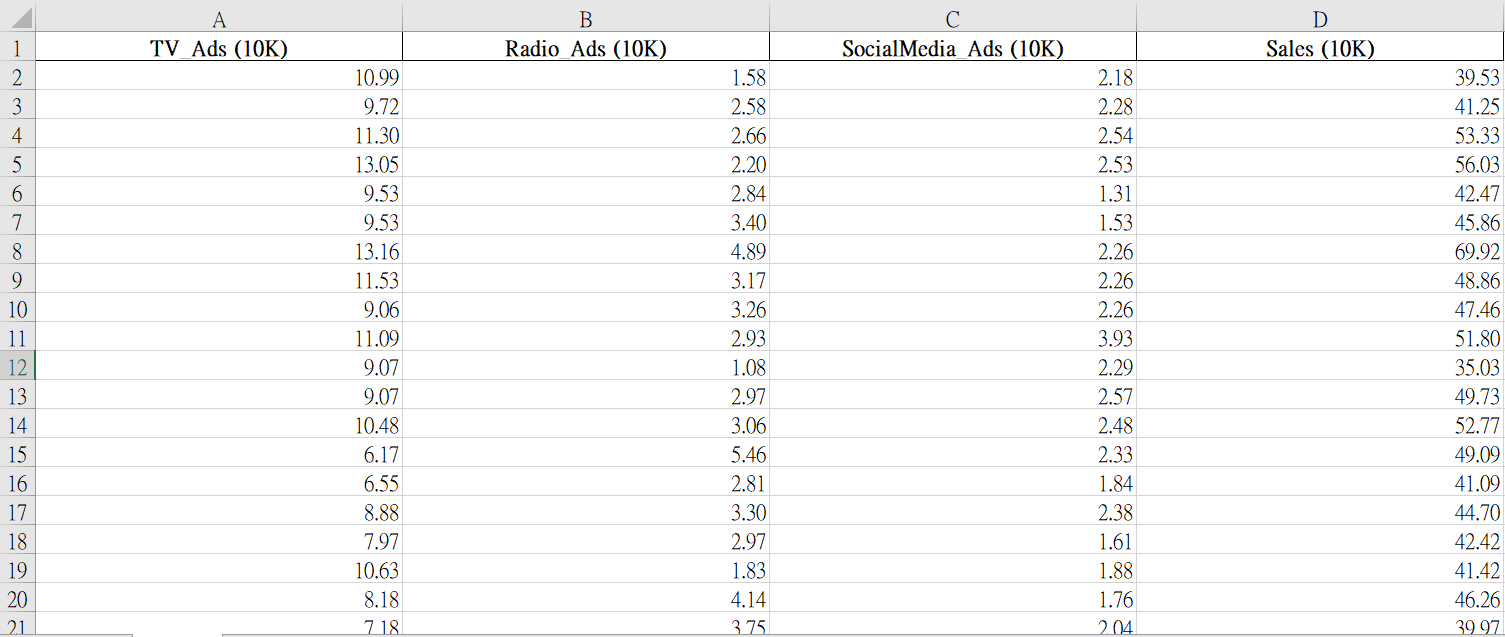


圖5- 5 迴歸分析練習

* 任務

利用詢問ChatGPT生成Python程式碼來呈現迴歸線圖以利後續公司有效進行配置行銷預算。

\*檔案為「迴歸分析二.xlsx」

* 結果

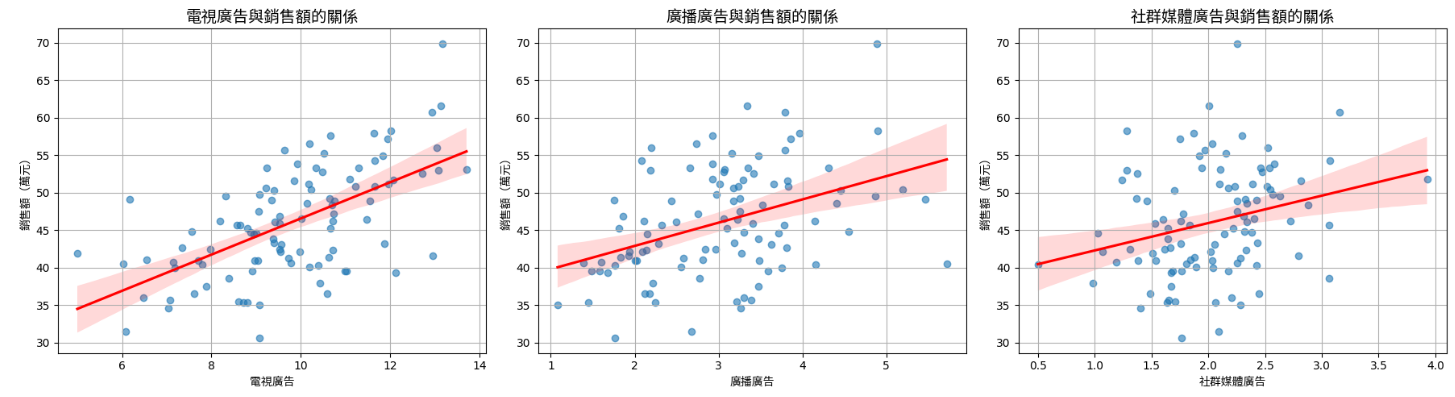


圖5- 6 迴歸分析練習結果

# 文字探勘與文字雲

## 斷詞、詞頻與詞雲簡介

### 斷詞

文字處理的第一步通常是進行「斷詞」，也就是把一整句話切割成一個個有意義的「詞」或「字」。這個步驟在處理中文資料時尤其重要，因為中文不像英文有空格作為詞與詞之間的分隔，我們需要依靠斷詞工具來判斷每個詞語的邊界。例如，句子「我喜歡吃蛋糕」在斷詞後會變成「我／喜歡／吃／蛋糕」，這樣的處理方式能幫助我們更有效地進行後續的分析與應用。

### 詞頻

詞頻指的是計算每個詞在文本中出現的次數，透過這樣的方式，我們可以了解哪些詞最常出現，進一步掌握文章或資料的重點與可能的主題。這個步驟在文字分析中非常實用，能幫助我們快速抓出常見關鍵詞，為後續的分類、標籤或視覺化做準備。

### 詞雲

詞雲是一種將詞頻結果用視覺化方式呈現的工具。在詞雲中，某個詞出現得越頻繁，它的字體就越大、越醒目，讓人一眼就能看出資料中的主要關鍵字。這種圖像化方式不但直觀，也適合在簡報或報告中快速傳達重點內容。

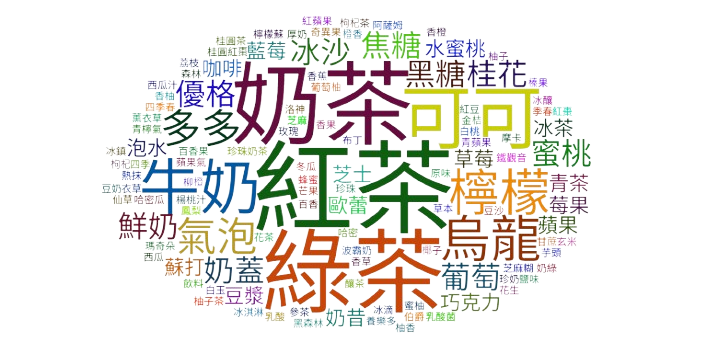


圖6- 1 詞雲示意圖

## 範例說明

* 情境說明

為掌握市場對功能性纖維產品的需求，本公司彙整客戶意見回饋，希望透過文字探勘與文字雲技術，分析常見關鍵字與關注重點，作為後續產品開發與行銷策略的依據。

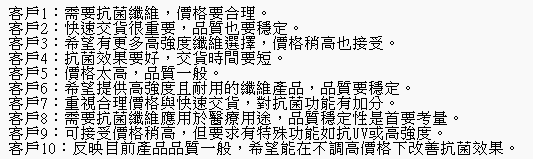


圖6- 2 文字探勘及文字雲範例

* 操作步驟

1. 將檔案提交給ChatGPT，並輸入提示詞：「為掌握市場對功能性纖維產品的需求，本公司彙整客戶意見回饋， 請將回饋內容進行中文斷詞，生成詞雲，突出高頻詞，分析常見關鍵字 ， 並直接產生完整Python程式碼。 圖表中文字型已放在 /content/drive/MyDrive/課程檔案/NotoSansCJKtc-Regular.otf，程式碼中請用 fontproperties 明確指定標題與軸標籤的中文字型。請直接回傳完整程式碼，不要再確認字型路徑或環境。」
2. 將ChatGPT提供的程式碼貼至Colab執行，就能看到線性迴歸散佈圖以利後續客戶評估與策略制定的依據
3. \*檔案為「文字探勘一.txt」



圖6- 3 文字探勘及文字雲範例之詢問ChatGPT指令



圖6- 4 文字探勘及文字雲範例之ChatGPT生成程式碼-1



圖6- 5 文字探勘及文字雲範例之ChatGPT生成程式碼-2



圖6- 6 文字探勘及文字雲範例結果-1



圖6- 7 文字探勘及文字雲範例結果-2

## 練習活動

* 情境說明

某中型企業為了提升員工滿意度與組織效率，透過匿名問卷調查方式，收集各部門同仁對於「工作環境」「管理方式」「溝通效率」等面向的意見回饋。人資部門希望利用文字探勘與文字雲技術，快速掌握員工最常提及的問題與期望，做為後續內部改善計畫的依據。

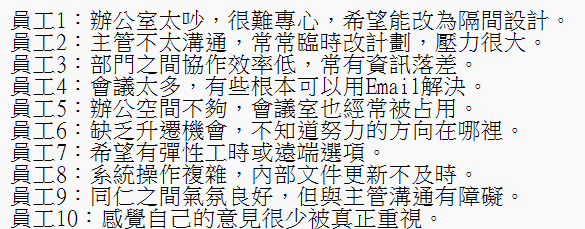


圖6- 8 文字探勘及文字雲練習

* 任務

利用詢問ChatGPT生成Python程式碼來呈現斷詞、詞頻及詞雲結果以利快速掌握內容。

\*檔案為「文字探勘二.txt」

* 結果



圖6- 9 文字探勘及文字雲練習結果

# 特徵值與目標變數的相關性分析

## 相關性分析簡介

在資料分析與建模的初期階段，進行特徵與目標變數的相關性分析是一個重要步驟。這項分析能幫助我們了解變數之間是否存在關聯，進而判斷哪些特徵可能對預測目標變數有助益，哪些則可能無明顯意義或與其他特徵重複。例如，如果某個特徵跟我們要預測的結果關係很密切，那它可能就是個「有幫助的線索」；像這樣的特徵，我們就會特別關注。相反地，如果某個變數跟結果一點關係都沒有，那它可能只是雜訊，反而會影響模型的判斷，我們就可以考慮把它拿掉或重新處理。這樣一來，模型的表現會更好，我們自己也會對資料的樣貌更加清楚。

## 分析方法與工具

* 方法

在執行相關性分析時，常見的方法之一是計算皮爾森相關係數（Pearson Correlation Coefficient），用來衡量兩個變數之間的線性關係。相關係數的數值範圍介於 -1 到 1 之間，其中：

* + 1 表示完全正相關
* -1 表示完全負相關
* 0 則代表沒有線性關聯

若資料中存在非線性或等級關係，則可考慮使用斯皮爾曼等級相關係數（Spearman’s Rank Correlation Coefficient），此方法能捕捉到變數間的單調變化，不受限於線性。

在視覺化方面，我們可以使用橫條圖來呈現各個特徵與目標變數之間的相關程度。這種圖表能夠清楚顯示每個特徵的重要性，讓我們快速判斷哪些變數值得保留或優先考慮。

透過 Python 中的 pandas 和 matplotlib 或 seaborn 等套件，可以方便地計算相關係數並繪製橫條圖，協助我們有效地進行特徵篩選與資料理解。

* 工具

表7- 1 相關性分析常用工具

|  |  |  |
| --- | --- | --- |
| **工具** | **用途** | **解釋方式** |
| pandas.corr() | 計算每個變數與目標變數之間的相關係數（例如：與銷售的關係） | 幫我們算出「哪些變數跟結果有關」 |
| Matplotlib | 畫出**橫條圖**來視覺化相關係數的強弱 | 讓我們「一眼看出誰關係強誰關係弱」 |

## 範例說明

* 情境說明

為了掌握空氣品質的影響因素，資料分析團隊蒐集了各測站的即時監測數據，包含多項污染指標與氣象變數，並希望透過相關性分析，找出與 AQI 變化最相關的特徵，以做為後續預測與決策的依據。

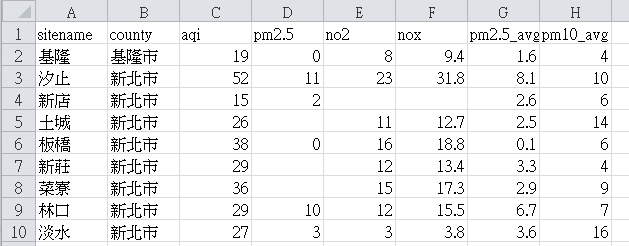


圖7- 1 特徵值相關性分析範例

* 操作步驟

1. 將檔案提交給ChatGPT，並輸入提示詞：「為了掌握空氣品質的影響因素，資料分析團隊蒐集了各測站的即時監測數據，希望找出與 AQI 變化最相關的特徵，以做為後續預測與決策的依據。 請繪製一張橫條圖，呈現各特徵與 AQI 的皮爾森相關係數絕對值，由大到小排序，並加上標籤與標題，並直接產生完整Python程式碼。 圖表中文字型已放在 /content/drive/MyDrive/課程檔案/NotoSansCJKtc-Regular.otf，程式碼中請用 fontproperties 明確指定標題與軸標籤的中文字型。請直接回傳完整程式碼，不要再確認字型路徑或環境。」
2. 將ChatGPT提供的程式碼貼至Colab執行，就能看到依照因素影響相關性高低排序的長條圖
3. \*檔案為「特徵值一.xlsx」

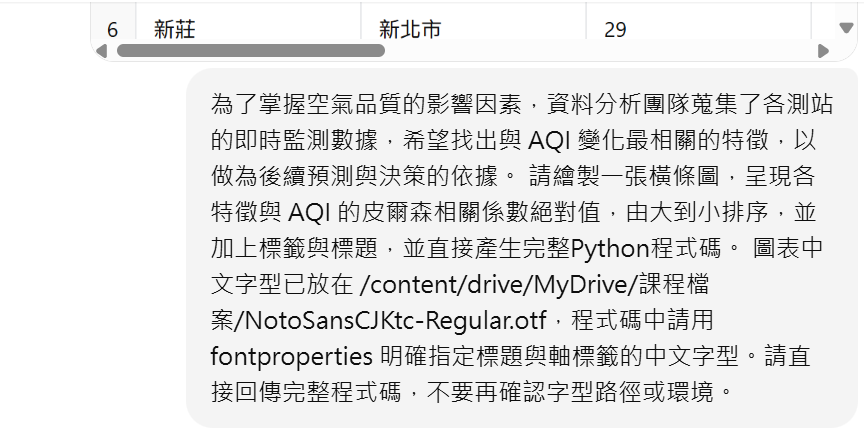


圖7- 2 特徵值相關性分析範例之詢問ChatGPT指令



圖7- 3 特徵值相關性分析範例之ChatGPT生成程式碼

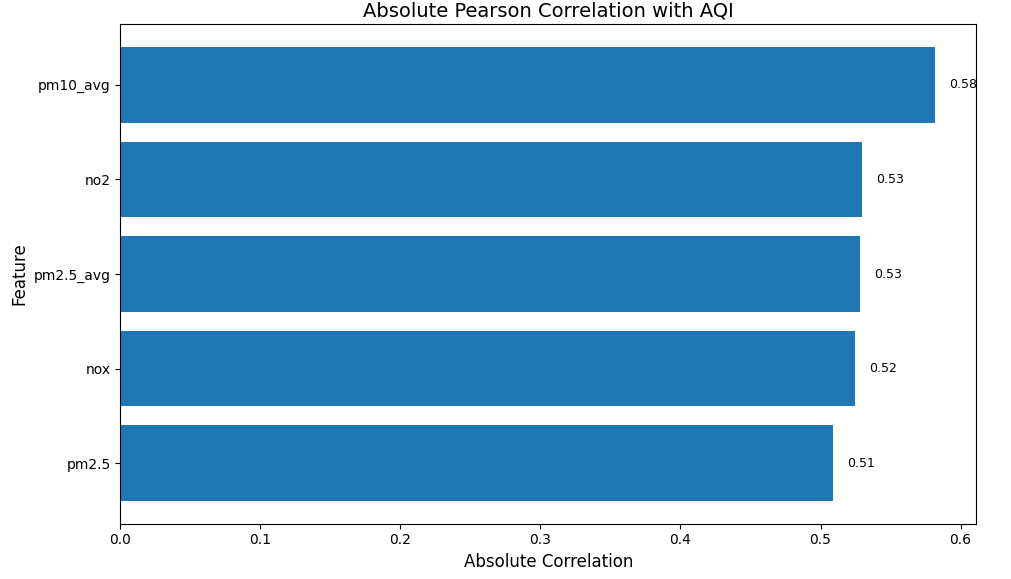


圖7- 4 特徵值相關性分析範例結果

## 練習活動

* 情境說明

某公司人資部門希望了解影響員工績效的主要因素，蒐集了內部員工資料，包含年資、每週工時、受訓時數、出勤率、主管評分等變數，並將「年度績效評分」作為目標欄位，希望找出與績效表現最相關的特徵值，以協助未來的人才培育與績效預測。

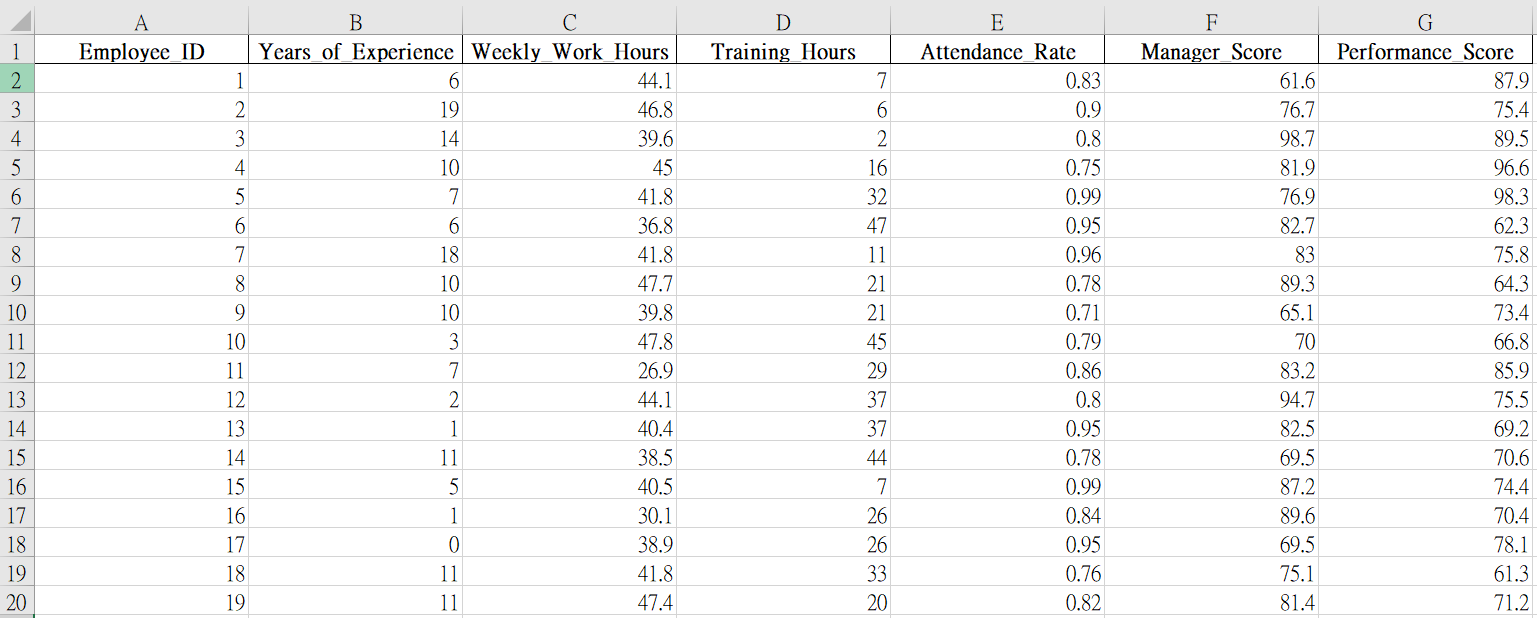


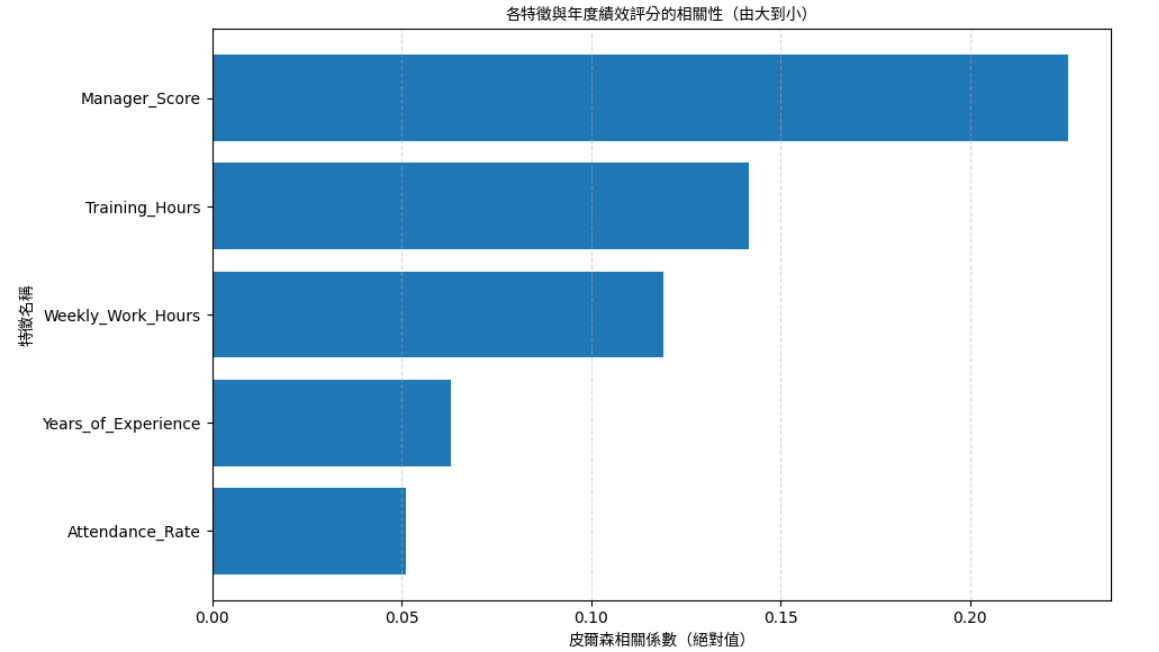
圖7- 5 特徵值相關性分析練習

* 任務

利用詢問ChatGPT生成Python程式碼來呈現影響員工績效因素的相關性高低長條圖。

\*檔案為「特徵值二.xlsx」

* 結果



每週工時

年資

出勤率

受訓時數

主管評分

圖7- 6 特徵值相關性分析練習結果

**#補充：迴歸 vs. 相關性分析**

|  |  |  |
| --- | --- | --- |
| **項目** | **相關性分析（Correlation）** | **迴歸分析（Regression）** |
| **目的** | 看「兩個變數有沒有關係」 | 預測或解釋「一個變數會不會受另一個影響」 |
| **關注焦點** | 關係強度與方向（正相關 or 負相關） | 因果推論與預測（X變Y的能力） |
| **變數關係** | 雙方平等（X 與 Y 地位對等） | 一方當作自變數，另一方是應變數（X 影響 Y） |
| **常見用途** | 探索變數間的關聯（初步了解資料） | 預測、建模、找出影響因素 |
| **輸出數值** | -1 ~ 1 的相關係數（例如：r = 0.8） | 方程式（例如：Y = 2X + 5） |
| **適合資料類型** | 數值型變數（連續型） | 數值型應變數，自變數可為數值或類別（視模型而定） |
| **使用時機** | * 想知道兩變數之間有沒有「線性關係」 * 初步分析前的探索 | * 想知道某變數是否會「被預測」、「被影響」 * 建立模型做預測或解釋變異來源 |

# **常見問答**

Q1. 為什麼我用一樣的提示詞，但產出的程式碼跟範例不一樣？

A：ChatGPT是**生成式模型**，不是記憶體模型。即使你輸入的提示詞跟範例一模一樣，也可能因下列原因導致產出不同：

* ChatGPT 根據當前上下文進行生成（如果之前聊了別的東西會有影響）；
* 同一個任務有多種寫法，它會選擇不同方式來解題；
* 模型版本更新會帶來些微行為變化。

建議：若希望產出一致，可加上語句如「請與範例三相同」、「欄位名稱與圖X相同」、「不要改變欄位順序」來收斂結果。

Q2. 為什麼圖表中文字還是英文或出現亂碼，明明我字型路徑也設好了？

A：這通常是因為「雖然有載入字型，但沒有在title()、xlabel()、ylabel() 裡加上 fontproperties 參數」導致的。

解法示例：

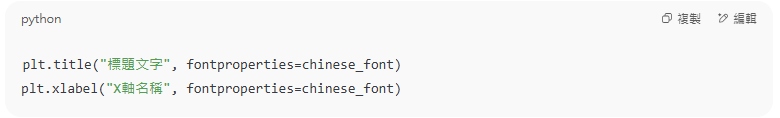


圖8- 1 常見問題 Q2

Q3. 為什麼程式碼會出錯提示欄位找不到？

A：可能：

* 忘了確認 Excel 表單中的欄位名稱與提示詞中一致；
* 使用了含空格或中英混合的欄位名稱，但沒有加 encoding 或錯誤拼寫；
* 有時 ChatGPT 推測了欄位名，但實際資料中沒那個欄位。

建議：可在提示詞中加註實際欄位名，例如：  
「欄位名稱為『訂單金額』『產品種類』，請直接用這些欄位名寫程式」。

Q4. 為什麼我分群的結果和圖表與範例不同，群數也變了？

A：因為 ChatGPT 常使用 KMeans 分群，而 K 值（群的數量）預設會選3、4、5等常見數，但如果你的資料特徵差異不明顯，分群結果會與範例不同。

建議：可在提示詞中補充「請依照資料自動決定最佳群數」或「請設定群數為3」來明確限制。

Q5. ChatGPT 回傳程式碼時出現省略或沒畫圖，是什麼原因？

A：如果提示詞中沒有強調「要完整程式碼」「直接產出圖表」，模型可能會：

* 只產出部分處理邏輯（如資料清洗後停止）；
* 沒有附上 plt.show() 或圖表設定。

建議：加入提示詞如「並完整畫出圖表」「請附上 plt.show()」、「請包含所有必要 import 與執行語句」。

Q6. 同樣是做「分類」，有時模型用 Logistic Regression、有時用決策樹，差別在哪？

A：

* 若你**沒指定模型類型**，ChatGPT 會自由選擇；
* Logistic Regression 適用於二分類且解釋性高；
* 決策樹則直觀可視化，適合多分類或非線性邏輯。

建議：提示詞中可補上「請使用決策樹進行分類」來明確控制方法。

Q7. 我有自己的資料，可以用這些練習套用嗎？會有什麼要注意的？

A：當然可以，但要注意：

* 資料結構與欄位需明確；
* 欄位名稱最好以英文命名避免亂碼；
* 數值欄請先轉換為數值型別；
* 若欄位名稱不同，要在提示詞中明確列出。

小技巧：先用 df.info() 和 df.head() 確認結構，再輸入給 ChatGPT。

# 附錄

* 範例及練習檔案<https://drive.google.com/drive/folders/1oMrBKoWFJF5ZDxApTezqzvdHOILjhiBl?usp=sharing>

