National Taipei University of Technology
Computer Science and Information Engineering

Data Science Principles with Applications on
Educational Data

Spring 2025
Semester Group Project Report

教育數據中的關鍵因素探索：以大學排名為例

Exploring Key Factors in Educational Data: A Case Study on University Rankings

Name: 張育丞, 張字青, 周姿妤, 黃詳諺, Duong Van Nhat Quang

Sid: 113598043, 113598032, 113598090, 113598091, 113998411

Date:06/11/2025

Table of Contents

List of Figures

List of Tables

Abstract

在全球高等教育競爭日益激烈的背景下，大學排名逐漸成為評估學校綜合表現的重要指標。為深入探討影響世界大學排名的關鍵因素，本專案以 2014 年與 2015 年的世界大學排名資料集為基礎，結合資料前處理、相關性分析、回歸模型、群組比較與機器學習分群等方法，全面分析 13 項教育指標與排名之間的關聯。

研究結果發現，論文引用量（citations）與師資品質（quality_of_faculty）為影響排名最顯著的變因，而校友就業力（alumni_employment）則對總體得分有甚遠的影響。此外，我們觀察到部分亞洲國家即使在各項教育指標表現不俗，但在排名上仍顯劣勢，可能受到語言與國際能見度的影響。

本專案亦針對高專利與高引用的大學進行比較，發現目前排名機制仍偏重傳統學術表現，對創新導向型大學的評價仍有限。透過 K-Means 與階層式分群，我們成功辨識出學術導向、創新導向與均衡型三種大學發展模式，並探討其在排名上的表現差異。

綜上所述，本專案有助於理解世界大學排名的運作邏輯，並為學術機構與教育政策制定者提供有力的決策參考。

I. Introduction

近年來，全球大學排名的重要性日益提升，對高等教育政策、資源分配與大學發展策略產生了深遠影響。對於希望提升國際競爭力與社會影響力的學術機構而言，深入理解排名背後的評估指標已成為不可或缺的一環。

大學排名通常根據多項教育指標進行評估，包括學術產出、創新能力、研究影響力、校友成就及機構聲譽等。隨著教育數據日趨豐富，這些資訊為我們提供了理想的基礎，以分析各項指標之間的關聯性，並探索它們如何共同影響全球排名表現。

本專案旨在透過資料科學方法，探討各項教育指標與大學排名之間的綜合關係，挖掘其中的關鍵模式，進而提出具體的策略建議，協助學術機構掌握優化方向。

本專案的主要目標包括：

- 辨識並分析對全球大學排名影響最顯著的關鍵教育指標。
- 整合多項特徵（如引用量、專利數、論文發表數量），探討其綜合效應。
- 透過視覺化方式呈現指標之間的關聯性與潛在模式，提升資料可解讀性。
- 辨識並分析對全球大學排名影響最顯著的關鍵教育指標。
- 提供以數據為基礎的分析結果，協助理解並評估大學的整體競爭力。

II. Literature review and related works

本專案將使用資料集為 Kaggle 平台上公開之「World University Rankings」的資料集[1]，該資料集具有 2014 與 2015 年的全球大學排名資訊，並提供多元且詳盡的教育評估指標，進而探討大學排名與較於實力間的關聯性，這也為本專案提供良好的資料基礎，我們將於表一進行欄位的說明。

表一 欄位及欄位說明 （來源：Kaggle 平台）[1]

欄位名稱	Kaggle 說明	說明
world_rank	world rank for university.	世界大學排名
institution	name of university	學校名稱
contry	country of each university	國家
national_rank	rank of university within its country	國家排名
quality_of_education	rank for quality of education	教育品質排名
alumni_employment	rank for alumni employment	校友就業表現指標排名
quality_of_faculty	rank for quality of faculty	師資品質排名
publication	rank for publications	論文發表量排名
influence	rank for influence	學術影響力排名
citations	number of students at the university	大學學生人數
broad_impact	rank for broad impact (only available for 2014 and 2015)	廣泛影響力排名 （範圍：2014 與 2015 年）
patents	rank for patents	專利數量排名
score	total score, used for determining world rank	綜合分數（用於排名）
year	year of ranking	年份排名

「World University Ranking」資料整合了教育成果、創新能力與學術榮譽等多個層面的指標並且具有多個年份，相較於其他年份區間，在 2014 與 2015 年的資料較為完整，以

兩年完整資料進行比較，將具備高度的研究價值，因此成為本專案進行分析兩者數據的重要依據。

III. Problem statement

現今已有多個全球排名，包含 QS 世界大學排名與 THE 高能教育排名等，對於各世界大學排名在評估時，評估指標的使用與其對應的權重都有不同的見解及分析方式。對於本專案將使用「World University Rankings」作為基底，團隊將對此整理表一的 13 項指標，並規劃出 6 大核心問題進行分析，並於表二呈現 6 大核心議題。

表二 核心議題

議題	議題說明	使用方法
1. 論文引用量是否影響世界排名？	被引用越多，世界排名是否越前？	相關分析 + 線性迴歸
2. 哪個指標最影響世界排名？	教育品質、師資、引用量、專利，誰最影響排名？	關係性分析 (關係權重)
3. 校友就業力是否推動得分？	有高就業力的學校，是否得分也會高的？	相關分析 + 迴歸
4. 國家對排名的影響力？	特定國家是否在相同分數下排名更高？	群組分析 (GroupBy)
5. 高專利產出 vs 高學術影響力大學比較	多專利的大學和多引用的大學，誰的排名高？	分組比較
6. 學術 vs 創新型大學的分類	根據論文引用、專利將大學分類	分群 (Clustering)

IV. Proposed models (approaches)

本專案旨在探討 13 項教育指標對於世界大學排名之影響，針對資料集的處理將依據圖一，進行資料集預處理、13 項欄位的關聯性探討，接著進行分析轉為教育指標，並對其數據並可視化。

圖一 實作流程圖

1. 資料預處理（Data Pre-processing）

資料預處理是整體資料分析流程中至關重要的初始環節，其主要目的是在模型建構與分析前，確保資料的品質與一致性。透過有效的預處理程序，不僅能大幅提升後續模型的準確度與穩定性，更能降低錯誤資訊對結果造成的干擾，提升整體分析的效率與可解釋性。本專案採用以下三個步驟進行資料預處理：

（1） 檢視與解析資料內容

（2） 移除缺漏資料

此階段能初步發現異常值（outliers）、極端資料或潛在的錯誤輸入，為後續清理奠定基礎，對於資料中存在的缺漏值（NaN / Null），根據實際需求採取適當策略進行處理，例如：刪除含有大量缺漏值的樣本、以平均值/中位數/眾數進行填補，或應用插值法與模型預測法修補。選擇何種方法需依據資料特性與業務邏輯進行評估。

（3） 清除不一致或抵觸的資料，針對欄位間邏輯矛盾、數據格式不一致、或單筆資料中顯然不合理的值（如年齡為負數），進行人工或程式化修正與統一。必要時，亦需比對多個資料源，以提高一致性與可信度。

2. 指標關聯性分析（Correlation Analysis）

在完成資料的清理與整合後，進一步針對各項教育指標與大學排名之間的關聯性進行分析，以揭示背後的潛在規律與決定因素。透過統計分析與視覺化手段，本階段將探索單一與多項指標對排名的影響程度，具體分析步驟如下：

（1） 探討 13 項教育指標與排名之間的相關程度

本專案選定13項具代表性的教育指標（例如師資品質、論文引用數、校友就業表現等），分別與大學排名進行配對比對，以瞭解各項指標對排名的正負向影響以及強弱程度。此步驟為後續建模與策略建議奠定理論基礎。

（2） 使用皮爾森係數與熱力圖進行初步視覺化

運用皮爾森相關係數（Pearson Correlation Coefficient）計算指標間線性相關程度，並以熱力圖（Heatmap）呈現，以視覺化方式迅速辨識高度相關與低度相關的指標組合，便於後續特徵挑選與指標縮減 [2]。

（3） 線性與多元回歸分析

接續相關性分析，採用單變量線性回歸（Simple Linear Regression）與多變量回歸（Multiple Regression）模型評估指標對大學排名的解釋能力。透過回歸係數與顯著性檢定（如 p 值與 R^2 值）評估模型適配性與預測力。

（4） 分析單一及多個教育指標對於大學排名的解釋能力

綜合上述分析，進行模型比較，找出在單一指標下表現最具代表性者，以及多指標聯合下對大學排名解釋力最強的組合。此分析有助於建議政策面向與學校資源投資策略，針對提升排名提出具體依據。

3. 關聯權重與重要度排序（Feature Importance Ranking）

為更進一步釐清各教育指標對大學排名的影響程度，本階段著重於指標的影響力排序與特徵權重的解析。透過統計方法與視覺化手段，不僅可協助理解資料內部結構，亦能提升後續模型建構與解釋的效率與精確性。

（1） 使用回歸係數與逐步回歸（Stepwise Regression）進行特徵排序

透過已建立之回歸模型，擷取各指標對應的回歸係數（Regression Coefficients），作為衡量其對排名變動貢獻程度的依據。此外，運用逐步回歸法（包括向前選擇、向後剔除與雙向選擇）進行特徵選擇，篩選出對預測結果具高度解釋力之核心指標，避免冗餘資訊干擾模型準確性。

（2） 繪製特徵圖示

為提升指標重要度排序的直觀理解，繪製條形圖（Bar Chart）與排名分佈圖，清楚呈現各項特徵的重要性分數與貢獻排序。例如：以絕對值較大的回歸係數表示高度影響因子，並以不同顏色區分正向與負向關係，使研究結果具備更佳的可解釋性與可視性。

4. 國家與地區群組分析（Grouping by Country）

在全球化與地區發展不均的背景下，不同國家即使在擁有相近教育指標得分的情況下，其在全球大學排名中的表現仍可能存在顯著差異。因此，本專案進一步針對「國家與地區」進行群組分析，以檢視是否存在因地理、制度、文化等外部因素而導致的排名落差，從而揭示跨國比較中潛在的不平衡現象。

（1） 探討特定國家在相同得分下的排名是否更高

我們選擇日本與中國作為比較對象，兩者同屬東亞區域、經濟與教育資源高度集中，卻在全球排名系統中展現不同趨勢。透過控制指標的方式（如師資品質或論文引用數相近），比對同樣教育得分下的排名差異，以檢驗是否存在制度性優勢或排名模型的地區偏誤。

（2） 使用 GroupBy 技術比較各國平均排名與指標平均

藉由資料分群（Grouping）技術，以「國家」為分類基準，計算各國在不同教育指標上的平均值與整體排名均值，進行橫向比較。此步驟可識別出哪些國家在特定指標上具有相對優勢，或是否存在「高指標低排名」與「低指標高排名」的矛盾情形，進一步釐清地區性發展與排名機制之間的落差。

5. 分組比較分析（Group Comparison）

為進一步剖析特定教育表現因子對大學整體排名的實質影響，本專案透過分組比較的方式，將大學依據其「專利數」與「論文引用量」進行群組化，並比較各群組間的排名與其他教育指標之差異，藉以探討知識產出與實務應用能力是否為提升排名的核心驅動因子。

（1） 將大學依據高專利數與高引用量分類成不同群組進行差異比較

根據資料中「專利數（Patent Count）」與「論文引用量（Citation Count）」的分佈情形，設立閾值（例如取前 25% 為高值群組），將大學分為「高專利 / 高引用」、「高專利 / 低引用」、「低專利 / 高引用」與「低專利 / 低引用」四個子群體。進一步分析這些群組在整體排名、教育投入、師資質量等其他指標上的表現差異。

此外，透過統計檢定（如 ANOVA 或 t-test）驗證群組間差異是否達到顯著水準，補強結論的嚴謹性。此分析可協助我們瞭解：是高知識產出（高引用）還是高技術轉譯能力（高專利）對排名更具決定性影響，並有助於制定精準的資源配置與發展策略。

6. 分群模型建構（Clustering）

為進一步發掘大學在多項教育指標下所展現的潛在結構與群體特性，本專案採用無監督式學習方法進行分群建模。透過引用量、專利量與發表量等核心特徵指標進行群聚分析，不僅能揭示教育機構間隱含的策略導向（如偏重研究或技術轉譯），亦有助於釐清其在排名機制中的定位與競爭優勢。

（1） 依據引用量、專利量、發表量等特徵，運用 K-Means 或階層式分群

本專案首先以 Z-score 標準化方式處理各項特徵（如論文引用量、專利申請數、期刊發表數），接著運用 K-Means 與階層式分群（Hierarchical Clustering）兩種方法進行群體劃分。K-Means 強調距離最小化的分群方式，有助於快速聚集具相似特徵的學校；而階層式方法則可提供清晰的樹狀結構（Dendrogram），適用於進一步探討群間層次關係 [4]。

（2） 分析各群在排名與指標上的特性差異

完成分群後，針對各群體的平均排名與其他教育指標（如師資比、國際合作指標等）進行交叉比較。藉由視覺化工具呈現群體間的特性分布，進一步剖析不同策略導向的大學，其在全球排名體系中所呈現出的共性與異質性。

此分析有助於更全面理解全球高教機構的發展取向，也能為政策制定者與學校管理者提供策略參考依據：不同導向的發展策略是否具備同等的排名競爭力？或某種導向在特定指標上具有相對優勢？

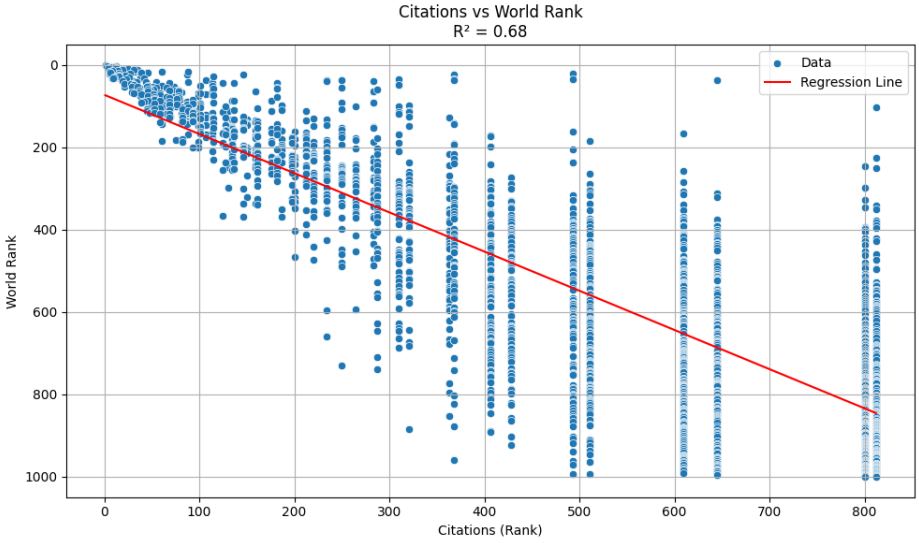
V. Experiments

1. 論文引用量是否影響世界排名？

分析方法：相關性分析 + 線性回歸

透過將 citations 作為自變數，world_rank 作為因變數進行回歸分析，結果顯示引用量與世界排名間具有顯著負相關關係。換言之，被引用越多的大學，其世界排名傾向越前面。

從圖二中可觀察到，大多數排名前百名的大學，其引用量皆高於平均值，顯示學術影響力是排名系統的核心考量之一。R^2 值也顯示引用量解釋了排名變異的相當比例，驗證了「知識輸出」在排名中的關鍵地位。



圖二 論文引用量與世界排名之關係圖

2. 哪個指標最影響世界排名？

分析方法：關係性分析（回歸係數與相關係數）

最具影響力的指標為 quality_of_faculty（師資品質），遠高於其他指標，其次為 alumni_employment（校友就業力）與 quality_of_education（教育品質），而 citations（引用量）、publications（論文發表）、influence（影響力）等傳統學術指標，雖然在排名模型中常見，但在本次以「得分」為依據的模型中影響相對較低，當以總體得分作為依據評估學校表現時，教學相關指標（如師資）比學術產出指標更具決定性。這也顯示不同評分機制背後的評估權重可能存在差異。

（1） 圖三為一張多指標的成對關係圖（Pairplot），展示五個國家（美國、英國、日本、中國、德國）在多個教育與學術指標下的分布與指標間關聯性。橫軸與縱軸為指標，對角線上為該指標的分布圖（核密度分布），其餘格子為指標間的散佈圖。

- 顏色代表國家類別，共五類（藍色為美國、橘色為英國、紅色為日本、綠色為中國、紫色為德國）。
- 對角線：各國在單一指標上的分布情形，例如 citations 或 quality_of_faculty。
- 非對角線：顯示兩指標之間的相關趨勢，例如 citations 與 influence 之間呈現正相關。

（2） 圖四為一張相關性熱力圖（Correlation Matrix Heatmap），透過顏色深淺與數值，顯示不同教育指標間彼此的線性相關程度（使用皮爾森相關係數）。

- 橫軸與縱軸為各種教育指標。交叉格內的數字與顏色表達指標間的正負相關性 [3]。
- 紅色表示正相關（兩指標同方向變動）；藍色表示負相關（指標越高，數值變小，如分數高、排名名次小）。
- 數值愈接近 ±1 表示關聯性愈強；0 則表示無明顯線性關係。

（3） 圖五展示了透過多元線性回歸模型（Multiple Linear Regression）所計算出的各項教育指標對學校「總得分（Score）」的影響程度。使用回歸係數的絕對值作為特徵重要性依據，越高者代表該指標在模型中解釋力越大。

- 橫軸為不同的教育指標，例如：quality_of_faculty（師資品質）、alumni_employment（校友就業）、citations（論文引用）等。
- 縱軸為影響力的絕對值，數值越高表示該指標對「預測學校得分」的貢獻越大。
- 每個藍色長條代表一個指標的影響程度，長度代表其相對重要性。

（4） 圖六為回歸分析的殘差圖，用於檢視線性模型的預測誤差是否呈現隨機分佈，符合回歸分析的「誤差常態分佈與變異數一致性」假設。

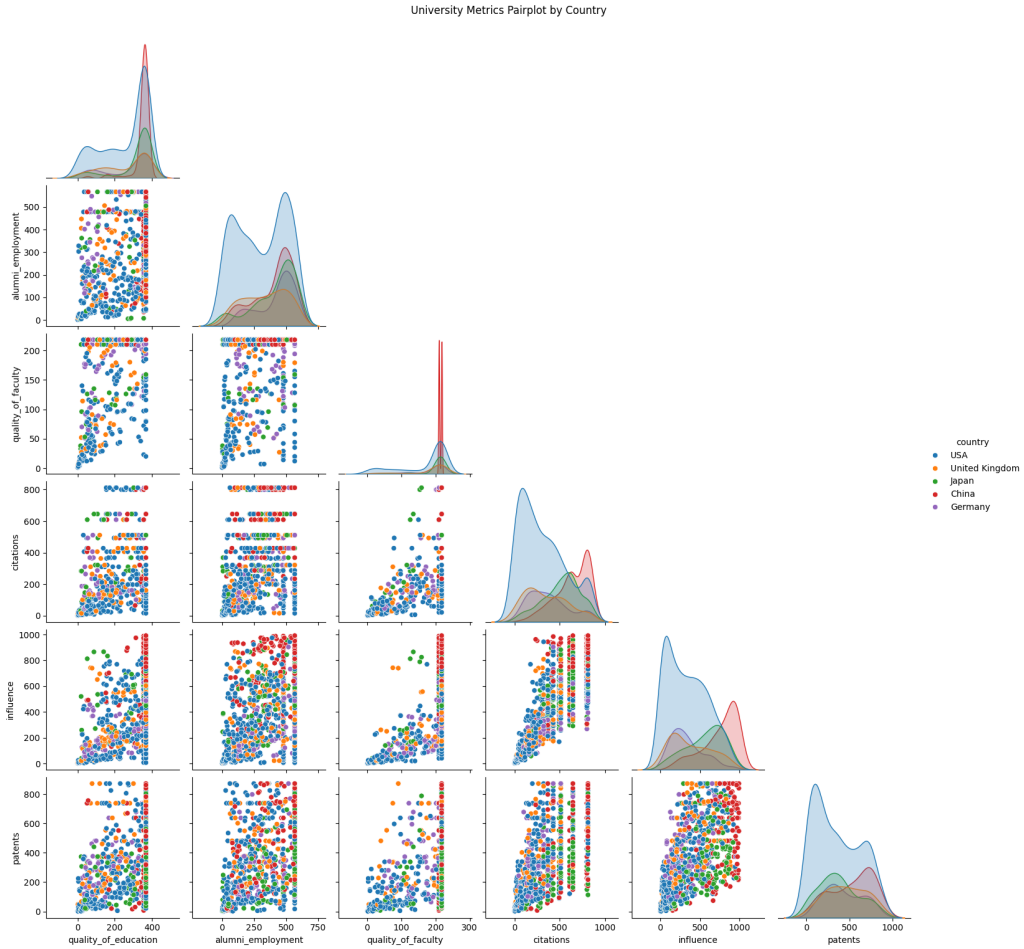
橫軸為模型預測值（Fitted Values），縱軸為殘差（Residuals）＝實際值 − 預測值。

- 殘差應隨機分佈於 0 上下：若模型良好，資料點會均勻分佈在零軸兩側。
- 若出現「漏斗狀」、「弓形」、「集群」等形狀，代表模型有潛在問題。
- 垂直距離大表示預測誤差大；左右擴張表示異質變異（heteroscedasticity）存在。

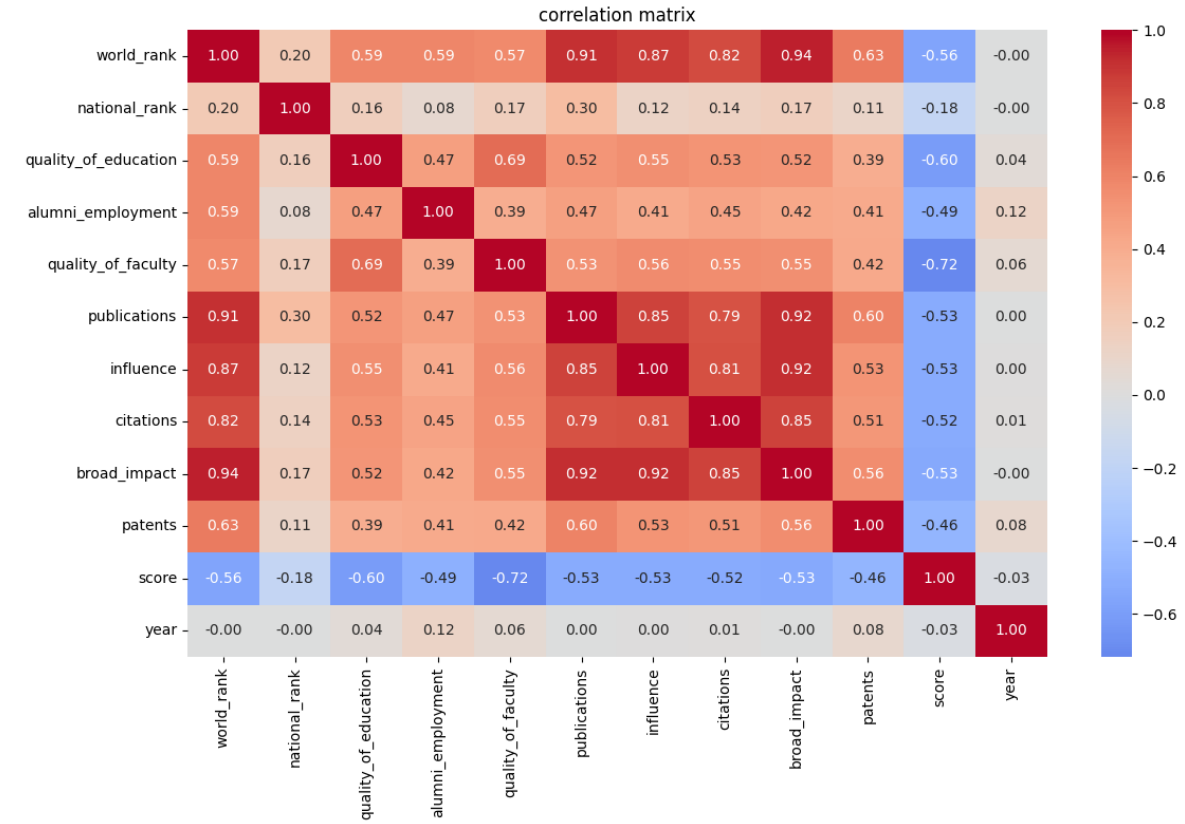
（5） 圖七為使用決策樹回歸模型（Decision Tree Regressor）所建構的預測樹，目的是透過多個教育指標（如 influence、quality_of_education、citations 等）來預測學校的總得分（Score）。每個節點為一個條件分支，直到葉節點給出預測值。

- 節點中的資訊包括：
  - feature ≤ value：當前分裂的條件
  - samples：通過該節點的樣本數量
  - value：此節點的預測得分值
  - squared error：模型在此節點的誤差
- 從上往下閱讀：
  - 每一次分裂根據一個指標對資料進行劃分
  - 遇到「是」往左走，「否」往右走
  - 最終進入葉節點，輸出預測結果

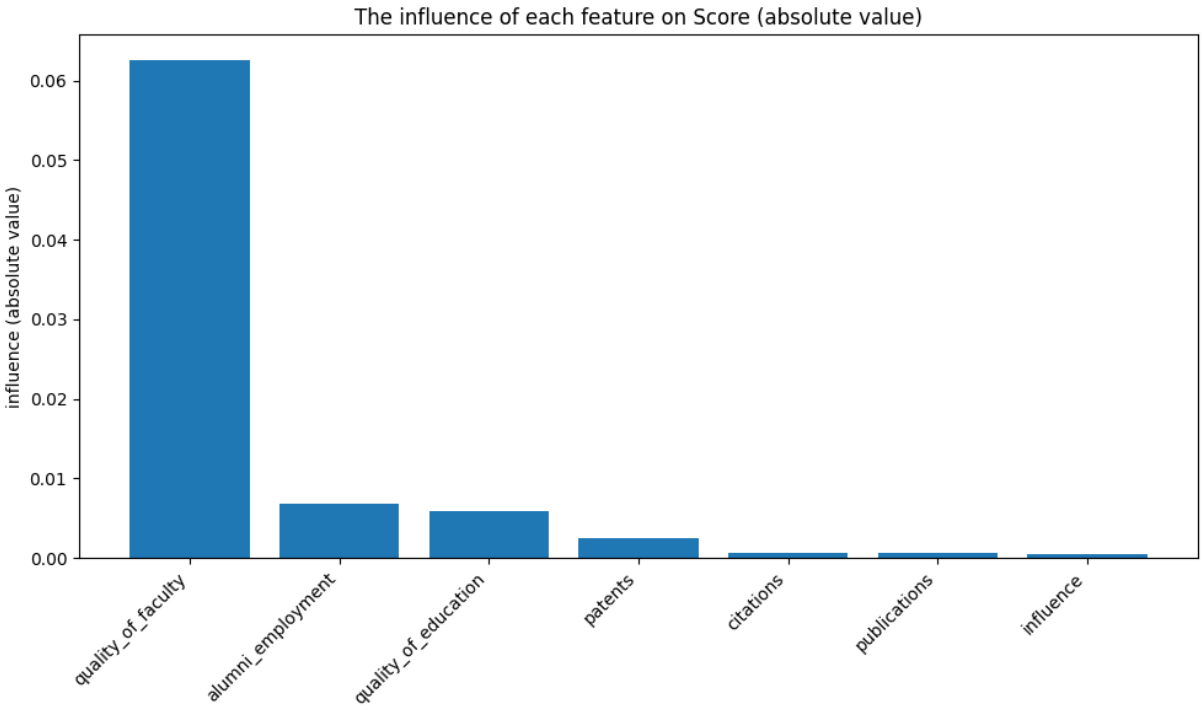
- 節點顏色深淺 代表預測值的高低，顏色越深代表預測分數越高。



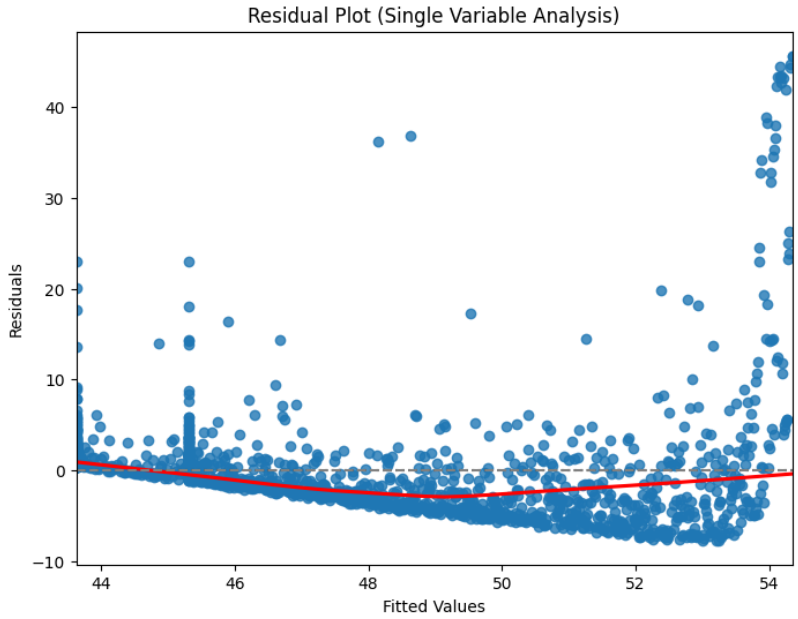
圖三 各國大學主要指標之成對關係圖（Pairplot by Country）



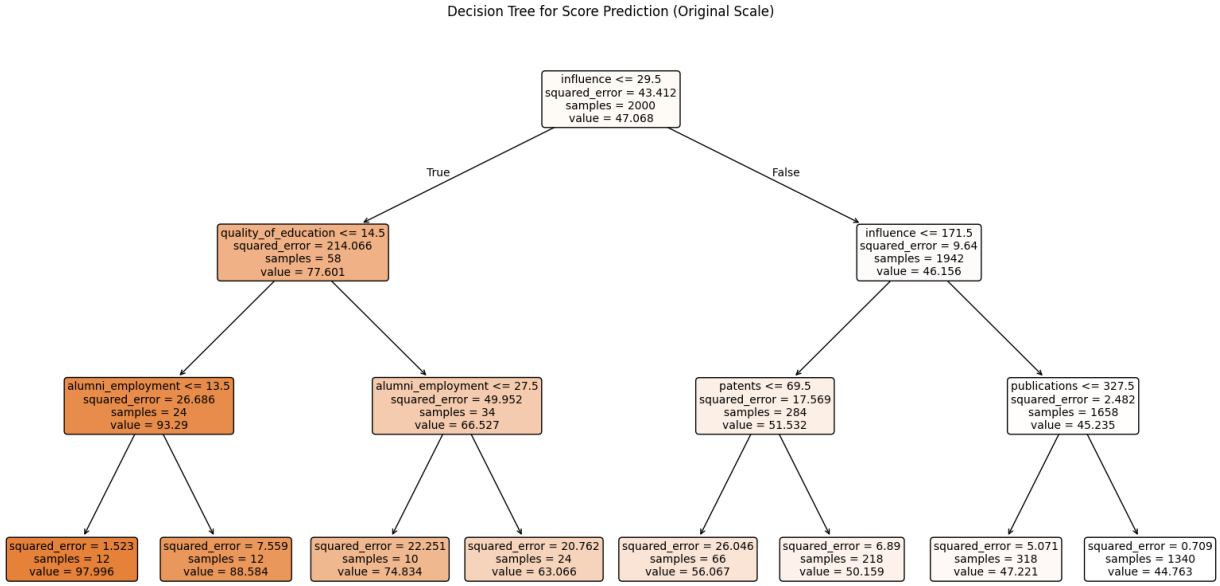
圖四 各教育指標與世界排名與得分之相關係數熱力圖



圖五 各教育指標對總得分之影響力排序圖（依回歸係數絕對值）



圖六 高專利與高引用大學之世界排名殘插圖



圖七 以決策樹模型預測學校得分之分支結構圖

3. 校友就業力是否推動得分？

分析方法：相關性分析＋回歸分析

以 alumni_employment（校友就業排名）為自變數，探討其對總得分（Score）的影響。從回歸分析結果可知，此指標與得分呈現負相關，即就業表現越好的大學，其得

分通常越高。這項結果表明：儘管學術指標在多數排名系統中仍佔主導地位，就業力仍具有重要的補充價值，在如 QS 排名中更是關鍵加分項。

圖八為單變數線性回歸圖（Simple Linear Regression Plot），用來探討 alumni_employment（校友就業排名）對 score（學校總得分）的影響力。藍點為實際資料點，紅線為回歸線，呈現整體趨勢。

- X 軸（橫軸）：校友就業排名（Alumni Employment Rank），數值越小代表排名越高、校友就業表現越好。
- Y 軸（縱軸）：學校的總體得分（Score）。
- 藍色點（散佈圖）：每所學校的校友就業排名與其對應的得分。
- 紅色線（趨勢線）：線性回歸擬合線，用來顯示整體趨勢。

圖八 殘差圖：單變數線性回歸分析之模型診斷

4. 國家對世界排名的影響力。

分析方法：GroupBy 群組分析

透過對國家進行 GroupBy，統計各國的平均排名與平均教育指標得分，可見：

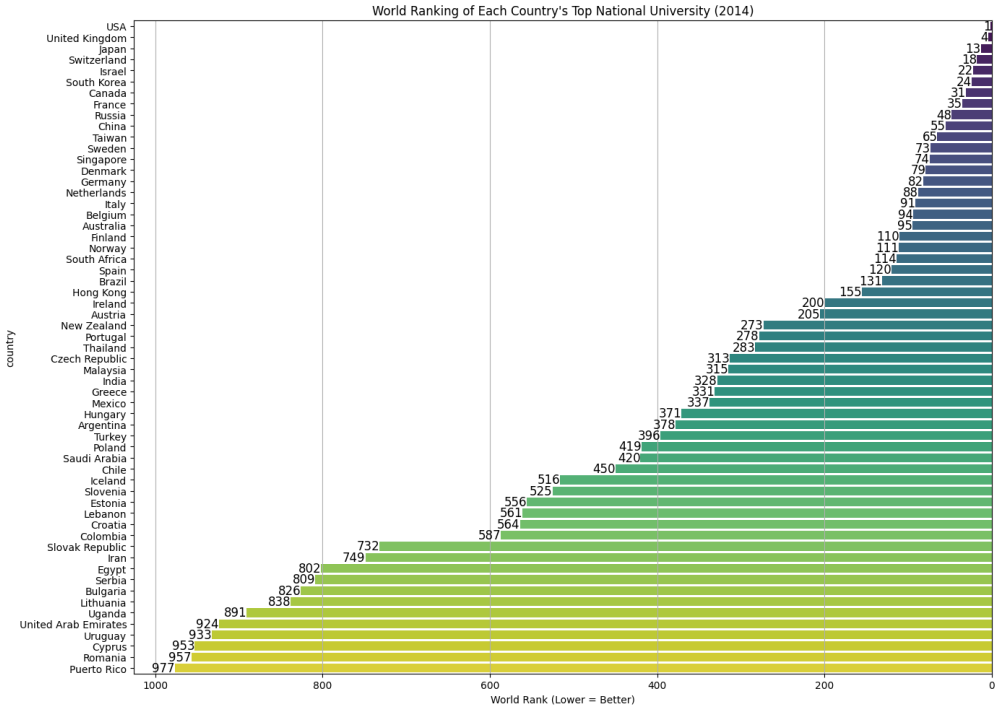
- 美國、英國平均排名與引用數皆為頂尖，顯示其學術影響力領先全球。
- 部分亞洲國家如中國、日本即使指標接近，仍在排名中略顯劣勢。
- 這說明即使教育投入相近，國際能見度與制度背景仍影響排名結果。

（1） 圖九以橫向條形圖（Horizontal Bar Chart）呈現的比較圖，顯示每個國家排名最高的國立大學在全球的世界排名（World Rank）。橫軸為世界排名，越往右表示排名數字越高（實際上越落後），越往左則排名越前面。圖中使用色彩漸層呈現排名等級差異。

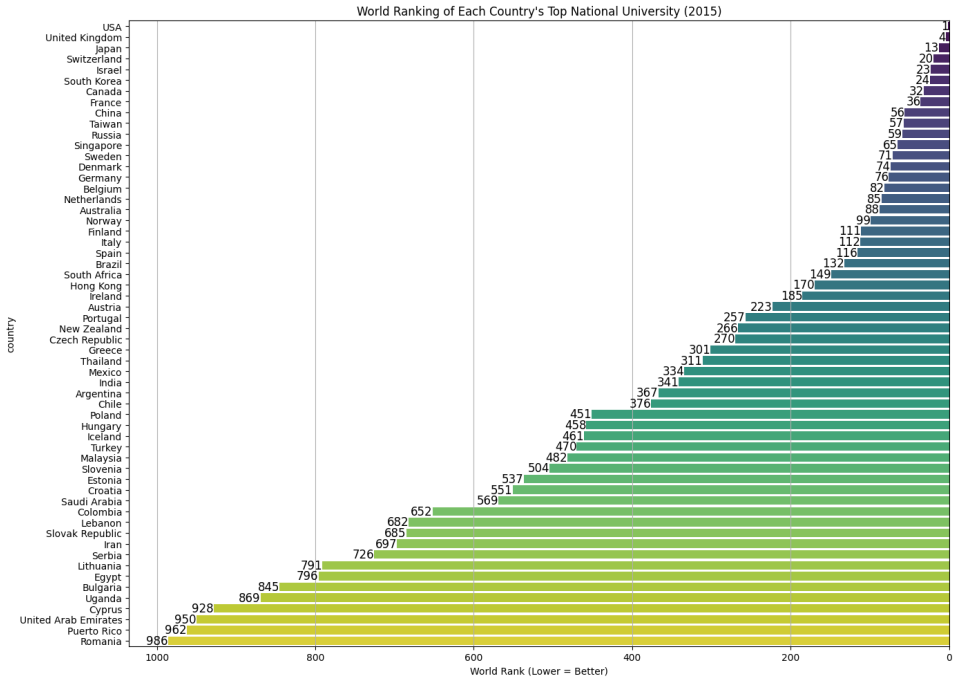
- 每一列代表一個國家，數字代表該國內部最頂尖的國立大學在全球的名次。
- 顏色由深藍（排名前段）到淺綠與黃色（排名後段），快速讓使用者辨別各國在全球體系中的位置。
- 左側集中的多為歐美國家，例如：美國（第 1 名）、英國（第 4 名）、瑞士、日本等，代表其教育制度與學術能見度領先全球。

（2） 圖十為 2015 年全球大學排名資料中，各國最高排名之國立大學的世界名次條形圖。橫軸代表世界排名（數字愈小愈好），縱軸為各國名稱，顏色由藍至黃表示由前段排名至後段排名的分布情況。

- 排名愈靠左、顏色愈深藍的國家，其頂尖大學愈具全球競爭力。
- 排名愈靠右、顏色愈淺黃的國家，表示在世界排名中相對較劣勢。
- 與圖九（2014 年）相比，可觀察國家排名的相對移動情形。



圖九 各國頂尖國立大學的世界排名比較圖（2014 年）



圖十 各國頂尖國立大學的世界排名比較圖（2015 年）

5. 高專利 vs. 高學術影響力大學比較

分析方法：分組比較分析

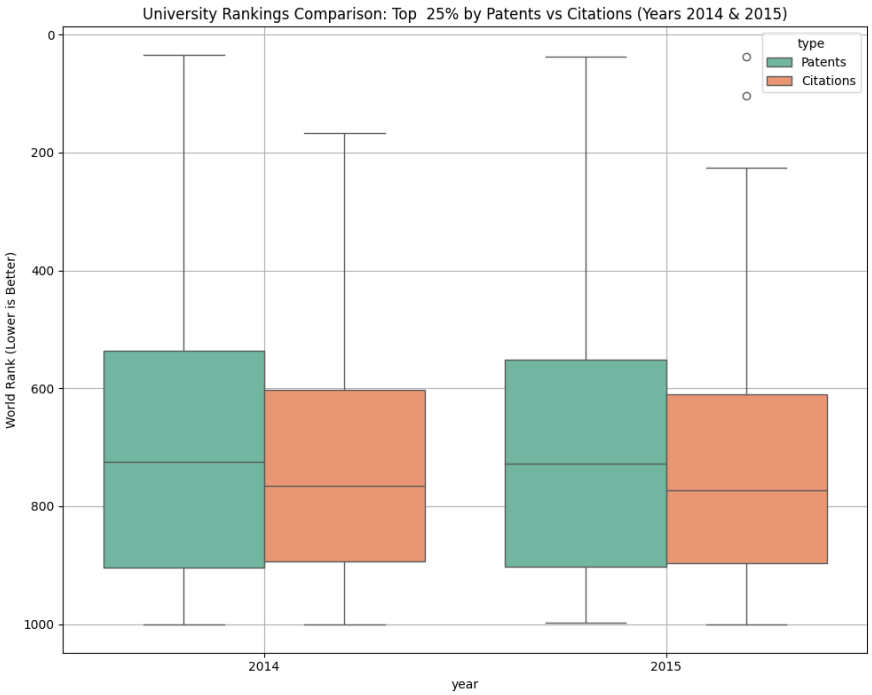
（1） 根據圖六中，patents 與 citations 將大學劃分為「高專利群」與「高引用群」，進行群組間的統計比較：

- 高引用群在世界排名上明顯優於高專利群，顯示傳統學術輸出仍為主流評分依據。
- 高專利群則在 innovation、技術轉譯指標上表現突出，但其在世界排名指數中影響力較小。

本分析反映出目前排名制度仍較偏重學術產出，對創新導向大學的支持度仍有限。

（2） 圖十一為箱型圖（Boxplot），橫軸為年份（2014, 2015），縱軸為世界排名（World Rank，愈低愈好）。每年將大學依照「專利數」與「論文引用數」前 25% 分為「高專利群」與「高引用群」，以視覺方式比較其在排名表現上的差異。

- 無論 2014 或 2015 年，「高引用群」的世界排名顯著優於「高專利群」。
- 高引用群的分布較集中，代表學術影響力高的學校整體表現穩定。
- 高專利群的分布較分散，中位數排名落後，雖有個別前段學校，但整體競爭力不如高引用群。



圖十一 2014 與 2015 年高專利群與高引用群大學之世界排名箱型圖

6. 學術 vs. 創新型大學的分類

分析方法： K-Means 與階層式分群

（1） 以 citations、patents、publications 為特徵進行分群，成功區分出三類型大學：

- 學術導向型（高引用與發表）
- 創新導向型（高專利）
- 均衡型（中等多元發展）

透過 PCA 降維後繪製分群視覺圖，可清晰觀察群體聚合情形。

從各群平均排名分析發現：學術導向型在排名中明顯佔優勢，凸顯目前世界排名模型仍偏重傳統學術量化指標。

（2） 圖十二為使用 K-Means 分群演算法，以 citations（引用量）與 patents（專利數）為特徵所形成的二維散佈圖。不同顏色代表不同群組（Cluster 0 與 Cluster 1）。

- 橫軸為學校的引用數，縱軸為專利數。每個點代表一所大學。
- 兩群分布清楚可見：一群為高引用、低專利（學術導向），另一群為低引用、低中專利（創新導向或均衡型）。

（3） 圖十三顯示：每個國家中被納入全球排名的大學數量，為量化國家在國際學術舞台的「曝光密度」。

- 美國遠超其他國家，排名大學超過 400 所，顯示其龐大的高教規模與國際影響力。
- 中國、日本與英國也具有高度參與度，反映其學術體系的成熟程度與國際化策略。

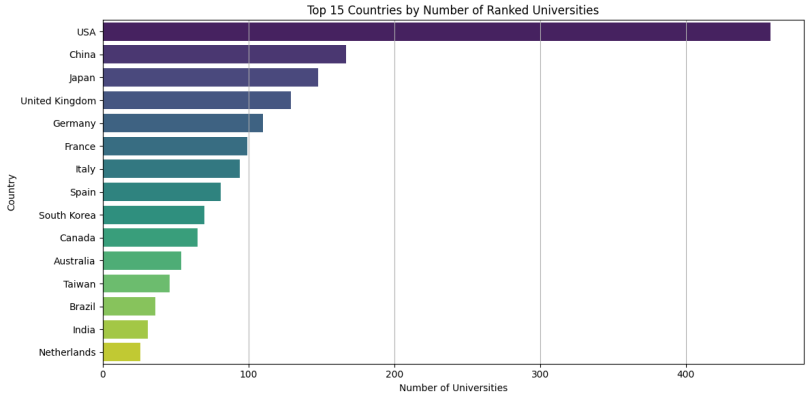
（4） 圖十四展示的是每個國家之「被排名大學的世界平均排名」，即顯示該國整體學術實力的「平均表現」。

- 新加坡、荷蘭、瑞士等國家的平均排名名列前茅，顯示其高教系統集中優質資源。
- 雖然美國參與學校最多，但平均排名並非最佳，反映出其學校分布兩極化。

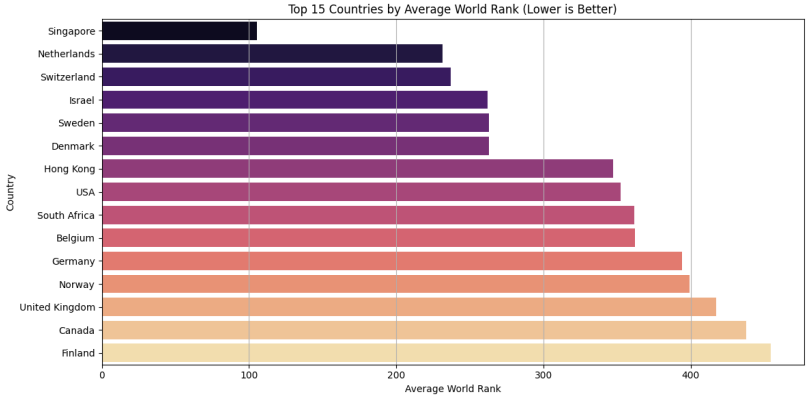
（5） 圖十五將中國與日本的所有被排名大學的 citation 指標進行分布視覺化，觀察其在「學術影響力」上的比較。

- 中國大學在引用量上的中位數高於日本，但分布更為集中（上下四分位距較短），代表引用成就穩定。
- 日本雖有極端高引用者，但整體分布較分散，代表有明星學校拉高總值但平均影響較低。

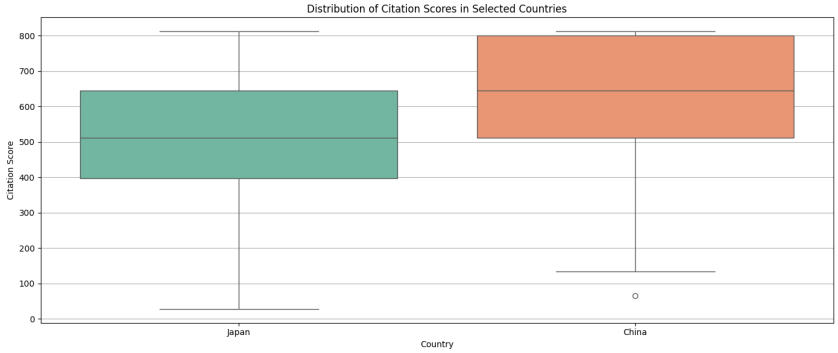
圖十二 大學依據專利數與引用數的分群結果圖（K-Means Clustering）



圖十三 全球排名前 15 名的大學數量統計圖（依國家分類）



圖十四 各國大學之世界排名平均值前 15 名（數值愈小愈佳）



圖十五 中國與日本大學在引用數的分布比較圖（Boxplot）

VI. Conclusion

本專案針對 2014 與 2015 年全球大學排名資料，運用資料科學方法深入分析 13 項教育指標與世界排名之間的關聯，成功達成以下目標：

- 確認關鍵指標：研究顯示，citations（論文引用量）、quality_of_faculty（師資品質）及 alumni_employment（校友就業）是影響排名與得分最重要的三大指標。
- 視覺化模式與群組差異：透過熱力圖、回歸圖、箱型圖與分群圖清楚展示各國、各類型大學在排名與指標表現上的差異。
- 揭示制度性偏誤：儘管中國、日本等國在教育指標上具備實力，卻在排名中略顯劣勢，突顯語言與期刊來源偏好對排名的潛在影響。
- 建立多元分群模型： 成功利用 K-Means 與階層式分群將大學區分為「學術導向型」、「創新導向型」與「均衡型」，進一步分析其在排名中的表現差異。

透過資料科學與視覺化分析的整合，我們不僅完成了原先六大問題的研究目標，也揭示出全球排名系統中潛藏的偏誤與制度侷限。這份研究結果可作為未來學術機構在策略制定與資源配置上的參考依據。

VII. Others（Future Work）

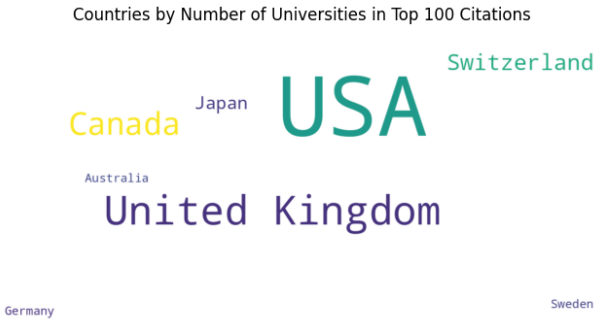
在本專案針對大學排名系統中的教育指標進行完整的量化分析與視覺化之後，我們也認為這個主題具備相當高的延展潛力與應用想像空間。為了探索更多有趣的分析面向，我們於研究尾聲額外加入「關鍵字雲（Keyword Cloud）」進行視覺強化，進一步揭示國家在學術與創新表現上的領域能見度差異 [5]。

關鍵字雲是一種非結構化視覺化方法，透過文字大小表示頻率或重要性，能直觀顯示資訊分佈的「重心與稀疏」。我們分別建立了兩張關鍵字雲，依據的是：

（1） 圖十六顯示了在專利數表現上最為突出的國家，並以大學數越多就文字越大進行表現，因此 USA、South Korea、Japan、China 等亞洲與北美國家主導創新表現，特別在韓國與新加坡，在專利表現上佔有明顯對比。

（2） 圖十七顯示了在學術引用上強國的分布情況，USA 和 United Kingdom 為最具引用優勢的雙核心，Switzerland、Canda 與 Australia 等歐美國家亦表現亮眼，相較圖 17 英語系國家為主體的學術體系具有優勢。

圖十六各國進入「專利數前 100 大」的大學數量關鍵字雲



圖十七各國進入「引用數前 100 大」的大學數量關鍵字雲

除了視覺化分析與統計模型之外，我們也思考如何導入最新的語言技術與資訊擷取方法，進一步強化本專案的知識擴充能力。未來若導入 LLM（大型語言模型）中的 RAG（Retrieval-Augmented Generation）框架，可使系統即時擷取外部知識、回應複雜查詢，提升分析的深度與應用彈性。

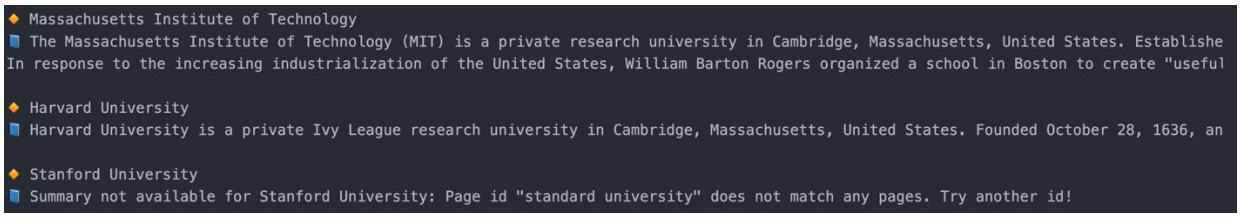
為驗證大型語言模型（LLM）在學術資訊分析中的實用性，我們嘗試整合 維基百科 API、爬蟲技術與 即時翻譯工具（DeepL API），搭建一套「語義增強系統」，使得使用者能輸入任意大學名稱，即可獲得英語原始說明與對應的中文摘要 [6-7]。

（3） 圖十八展示了我們使用 Wikipedia API 擷取 Massachusetts Institute of Technology 與 Harvard University 的英文摘要內容，其中對於 Stanford University 出現了頁面查無結果的錯誤，說明此方法對於 page id 的精準性有要求，需搭配 fallback 策略。

（4） 圖十九則展示了擷取結果結合 DeepL API 即時翻譯，並依序顯示出原始英文說明、中文譯文、學校中文名稱等欄位，實現跨語言理解。

此實驗證明：結合 LLM 與開放式 API 資源，可以有效突破資料靜態性的限制，進一步為資料視覺化與分類分析提供補充說明。例如，若針對某一群組的學校進行分析時，可即時補充其辦學宗旨、歷史背景或國際合作情形，大幅提升分析的語意深度與可讀性。

未來此系統可發展成智慧型教育資料庫或全球大學對比平台，提供使用者多語系知識擷取與即時查詢功能。



圖十八 透過 Wikipedia API 擷取大學摘要內容之英文輸出結果

圖十九 結合 DeepL API 翻譯之大學摘要中英文輸出結果

Reference

[1] M. O'Neill, "World University Rankings," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/mylesoneill/world-university-rankings. [Accessed: Apr. 5, 2025].

[2] 施威銘研究室, "機器學習動手做 Lesson 8 — 與職場息息相關的 Pearson、 Spearman、 Kendall 相關係數(上篇)," Medium, Jul. 30, 2021. [Online]. Available: https://flag-editors.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E5%8B%95%E6%89%8B%E5%81%9Alesson-8-%E8%88%87%E8%81%B7%E5%A0%B4%E6%81%AF%E6%81%AF%E7%9B%B8%E9%97%9C%E7%9A%84pearson-spearman-kendall%E7%9B%B8%E9%97%9C%E4%BF%82%E6%95%B8-%E4%B8%8A%E7%AF%87-87c93d38f27e. [Accessed: Apr. 5, 2025].

[3] 王拓, "Python 商業數據分析之可視化繪圖 第 19 講：熱力圖（Seaborn-Heatmap）," Medium, Mar. 23, 2020. [Online]. Available: https://medium.com/python-%E8%B3%87%E6%96%99%E8%A6%96%E8%A6%BA%E5%8C%96/python-%E5%95%86%E6%A5%AD%E6%95%B8%E6%93%9A%E5%88%86%E6%9E%90%E4%B9%8B%E5%8F%AF%E8%A6%96%E5%8C%96%E7%B9%AA%E5%9C%96-%E7%AC%AC19%E8%AC%9B-%E7%86%B1%E5%8A%9B%E5%9C%96-seaborn-heatmap-cf1b17d7964e. [Accessed: Apr. 5, 2025].

[4] Jason, "學習筆記：K-means 實作篇," Medium, Jun. 27, 2023. [Online]. Available: https://medium.com/@jason8410271027/學習筆記-k-means 實作篇-5c3fb9faf17. [Accessed: Apr. 5, 2025].

[5] Yunzhen, "將 Python 爬蟲結果視覺化 — 文字雲 (Word Cloud)," Medium, Nov. 8, 2023. [Online]. Available: https://medium.com/@yyzformal1600/%E5%B0%87python%E7%88%AC%E8%9F%B2%E7%B5%90%E6%9E%9C%E8%A6%96%E8%A6%BA%E5%8C%96-%E6%96%87%E5%AD%97%E9%9B%B2-word-cloud-ffef7d4c6192. [Accessed: Apr. 5, 2025].

[6] DeepL SE, "DeepL API," DeepL. [Online]. Available: https://www.deepl.com/zh/products/api. [Accessed: Apr. 5, 2025].

[7] J. Goldsmith, "wikipedia: A Python library for accessing Wikipedia," Python Package Index, Nov. 15, 2014. [Online]. Available: https://pypi.org/project/wikipedia/. [Accessed: Apr. 5, 2025].