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

**Content**

**Abstract**

1. **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。

本研究的主要目標包括：

* 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。

1. **Literature review and related works**

本研究將使用資料集為Kaggle 平台上公開之「World University Rankings」的資料集，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，這也為本報告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）

|  |  |  |
| --- | --- | --- |
| 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，因此成為本報告進行分析兩者數據的重要依據。

1. **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。對於本報告將使用「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

|  |  |  |
| --- | --- | --- |
| 題目 | 問題說明 | 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. 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 |
| 3. 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) |
| 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) |

1. **Proposed models (approaches)**

本報告旨在探討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Ｘ，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。

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。本研究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. 首先，透過統計摘要（如平均值、標準差、最大值、最小值等）與視覺化方法（如分布圖、盒鬍圖），全面檢視資料的基本結構與變數間的關聯性。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。
  3. 移除缺漏資料

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。

* 1. 清除不一致或抵觸的資料
  2. 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。

1. 變數關聯性分析（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變數對排名的影響程度，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

本研究選定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。

* 1. 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）計算變數間線性相關程度，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變數組合，便於後續特徵挑選與變數縮減。

* 1. 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 R² 值）評估模型適配性與預測力。

* 1. 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。

1. 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，本階段著重於變數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，擷取各變數對應的回歸係數（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、向後剃除與雙向選擇）進行特徵選擇，篩選出對預測結果具高度解釋力之核心變數，避免冗餘資訊干擾模型準確性。

* 1. 繪製特徵圖示
  2. 為提升變數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。

1. 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，本研究進一步針對「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。透過控制變數的方式（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。

* 1. 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。

1. 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，本研究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。

1. 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，本研究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變數進行群聚分析，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

本研究首先以 Z-score 標準化方式處理各項特徵變數（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。

* 1. 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？

1. **Experiments**
2. 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。

從散佈圖與回歸線中可觀察到，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。R² 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。

1. 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 為變數進行回歸分析，結果顯示該指標與得分呈現中度負相關。這代表 就業表現佳的學校在世界排名上有一定優勢，但相較於論文引用，影響力略低。

這反映出部分世界排名指標較偏重學術表現，但就業力仍在部分排名機制（如 QS）中扮演加分角色。

1. 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）

透過多元線性回歸分析多個教育變數對排名的貢獻度後發現：

* 最顯著的三個指標依序為：citations、broad\_impact、alumni\_employment

其中 citations 的回歸係數最大，且在所有樣本中幾乎皆為顯著變項，顯示其為預測排名的關鍵變數；broad\_impact 亦反映整體影響力，與 citations 呈強烈共變。

1. 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。
* 然而，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，可能受限於排名演算法對語言與期刊來源的偏好。
* 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. 高專利 vs. 高學術影響力大學比較

分析方法： 分組比較分析

根據 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。

1. 學術 vs. 創新型大學的分類

分析方法： K-Means 與階層式分群

以 citations、patents、publications 為特徵變數進行分群，成功區分出三類型大學：

1. 學術導向型（高引用與發表）
2. 創新導向型（高專利）
3. 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。

1. **Conclusion**

Please include what the project has done and whether the objectives are achieved. What do you get in the project and what can be done in the future.

1. **Others（Future Work）**

Please state the workload and role of each member in your team for the project. Besides, show the timeline for the project and check whether all the proposed works have been done.

(文字雲....)

**Reference**