臺 北 市 立 大 學

都會產業經營與行銷學系

大四畢業專題

2018-2019 NCAA 一級籃球聯賽隊伍商業價值聚類分析

Cluster Analysis of 2018-2019NCAA Division I Men's Basketball Team's Business Value

大 專 生：朱寧

指導教授：莊旻達 博士

中 華 民 國 108 年 1 月

中文摘要

美國大學對體育的重視是廣為人知的，體育成績除了是學校吸引菁英人才的一個管道也是一種盈利模式，其盈利能力與職業團體聯盟不相上下，尤其是美式足球與籃球。全美大學體育協會 NCAA 是美國大學體育的管理機構，根據 NCAA 最新財務報表，2019 年的『三月瘋』就帶來了 8.675 億美元的受益，男子籃球一級聯賽中賺的就佔 NCAA 所有收益中百分之七十五，隨者 NCAA 商業化擴大，除了學校方面大力的投資推廣球隊外，各個運動品牌也爭相贊助其最優秀球隊，增加品牌知名度及商品銷售量。然而市場上除了勝率可以評斷球隊之的差異外，沒有一套系統可以評價球隊的商業價值，因此本論文想用機器學習聚類分析的從收益表現及勝率去將球隊分群，找出各群內球隊對應可用的商業價值，提供想從 NCAA 籃球獲利者其投資參考依據。

本論文將研究運用機器學習 K-means 研究以下三點

(一) 如何將球隊收益、支出及勝率中分群，找出各群內相關性質及群外相異性質

(二) 勝率跟球隊收益及支出是否有直接關係

(三) 各聯盟是否有存在水平差距

Abstracts

Sports and competitiveness are part of the US culture. People take sports competitions seriously, thus colleges can abstract more elite student by their superb sports tournament performance. Also, college sport event is a huge business, even its profitability could be compared with professional sports leagues, especially for American football and basketball. The National Collegiate Athletic Association (NCAA) is a non-profit association that organizes the sports programs of many colleges and universities in the US. Citing the newest financial report of NCAA ,''the March Madness''(which refers to that time of year (usually mid-March through the beginning of April) when the men's and women's college basketball tournaments are held.) bring approximately 800 millions revenue that account for 75 percent in total revenue of NCAA. With the extension of commercialization , many university and their can get enormous amount of money by television and licensing rights, moreover, many sports brands are competing to sponsor those basketball teams have well performance in order to increase their brand awareness and product sales. However, except the win-lose rate can compare NCAA basketball team's performance, there is no a system can review these basketball team's business value. Therefore, the author plans to use cluster analysis of machine learning to divide NCAA basketball team into different group and find out the business value in each team. Ultimately, offered this information to those people who search for profits from NCAA basketball team.

This dissertation will study the following three points from cluster analysis

1. How to divide all basketball teams into different groups according to revenues, expenses and win-loss ratio to find similar characteristics in the groups and the differences among the groups？

2. Is there a correlation among income, expenses and win-loss ratio？

3. Is there gap among all conferences of the NCAA basketball team？

目錄

第一章 緒論

一、研究動機

不懂體育沒辦法社交，對運動賽事的狂熱是美國文化的一部分。相比華人的「唯有讀書高」及「白富美」的價值觀，運動的重視程度貫穿整個社會價值觀，人的綜合競爭力來自強壯的體魄，從升學上可以看到在體育表現優異的學生可較容易申請名校及獲得獎學金，即使課業表現一般;也可以從政治人物口中聽到他們大談在學校運動賽事那些自豪的經歷，某種程度上，美國人將體育上的成就視為一種競爭力強獲一種人生上的成功，所以美國人願意花大筆時間及金錢在投資運動賽事上面，更因如此運動賽事成為巨大的商業生意，甚至大學體育賽事的吸金程度可與職業體育聯盟去做相比。

美國國家大學體育協會（英語：National Collegiate Athletic Association，縮寫：NCAA]）是美國一家非盈利組織，管理著 1,281 個大專院校、聯盟或個體體育組織；同時還負責組織美國和加拿大許多大學或學院的體育賽事，每年從電視轉播、門票及周邊商品等可創造將近 10 億美元的收入。而這些收入的最大的來源來自 NCAA 第一級別的男子籃球錦標賽，佔了總收入的 80%。最後這些收入將被分配給協會內美國的各個院校或聯盟。有此可知，各大學院校都想在這體育領域上取得優秀成績能得到名聲與金錢上的巨大成功，根據美國研究院的一項調查顯示：所有加入了美國國家大學體育學會(NCAA)第一分區的大學，他們投入在訓練運動員上的費用，是教學費用的三到六倍。想必而知，許多運動品牌廠商早就嗅到商機，每年都會相繼爭取跟優秀球隊的合作以搏取市場廣大關注度及炒熱

聯名商機，因此廠商如何選擇學校球隊為他們的合作對象相當重要，數據分析的手法可以幫助他們做更精準的決策，使利益最大化。

所以同樣身為籃球的狂熱份子，即同樣身為隸屬管理學院下的學生，想運用大學課堂上所學的統計推論去將數據與商業融合，並找出有趣的洞見及發展出一套分類模式，去提升職場上所需求運用數據的能力。

二、研究目的

本研究目的除了提升自我數據分析能力外，也想用量化的角度去探討全世界最值錢的大學體育賽事的市場生態，找出有價值的資訊，不只提供運動品牌在選球隊合做時有數據評量的依據去做判斷，為也提供未來學校有更好的策略。

近年來大數據當道，大大改變了職業運動的生態，ＮＢＡ許多球隊運用統計及數據探勘將球員表現量化，甚至使用人工智慧追蹤球員場上動態，立即時分析其球員細微動作去預估得分機率，提供球隊教練做現狀評估與決策去獲得更大奪勝的機會。然而 NCAA 在公開數據整理上及運用資料上遠不及ＮＢＡ，所以資料搜集的難度較具有挑戰性，另外在ＮＣＡＡ研究除了比賽結果勝分差預測外，相關研究少，在 NCAA 是一塊大餅有十億美元的情況下，如果能從數據更精確地分析球隊商業價值表現，進而製作一個分類模型，不僅能讓大學球隊審視自身價值有數據客觀的參考，也讓市場上各品牌更容易鎖定所想投資標的去發展自身策略。

三、 研究架構

本文共分五章。各章節內容依序如下：第一章為緒論，說明本文的研究動機、研究目的以及研究章節架構及流程。第二章則是文獻回顧，本章將搜集有關 NCAA 收益及相關教練薪酬的文獻，以利判斷之後機器學習分所運用之變數。第三章張則是研究方法，本章將使用的機器學習分類方法進一步探討 NCAA 分群結果等問題，並且詳細定義使用的各項變數以及實證資料的來源與處理方式。第四章則是研究結果，首先是將分群所得視覺化結果加以分析，再來則是各分群結果進行解釋與探討。第五章則是結論與未來研究方向，本章將作一個總結，並且針對ＮＣAA 球隊表現模型提出未來相關問題。

第二章 文獻方法回顧

一、 群集分析(cluster analysis)的概念

群集分析(cluster analysis)主要在考量個體與個體間其中的差異性，使用群集分析法去做資料間簡化與分類，也就是只把全部的資料分成 n 個群集，使各個群集內的資料具有高度相似性，而不同群集之間具有高度的不相似性；以統計的角度來描述，這樣的情況即是組內差異小、組間差異大的意思。

另外群集分析是一種邏輯程序，利用客觀的計量方法，將資料依某些屬性歸集在各群體中，使同一個群集內的資料具有相同的特性(homogeneity)，而不同群集之間卻有顯著的差異性(heterogeneity)。(Jain, Murty & Flynn,1999)

使用幾何圖形來看，同一群內的分子應當聚集在一起，而不同群集的分子應當彼此遠離，然而單靠圖形來決定群集，將會產生兩個問題，第一個問題是變數多時不易畫圖；另一個問題是將會有主觀或難以辨別的現象存在，所以群集分析提供一個數值的分類法，將資料照所收集的 p 個變數、N 個個體分成幾個群，根據 N 個觀察點的距離（相似度）形成距離矩陣（相似係數矩陣），將最接近的觀察點視為一個集群，並使群內個體間距離相近；不同群其個體間距離較遠，因此群集分析最主要的工作是定義各個體之間的距離、以及各群之間的距離，依照這些指標將所有觀察點最佳分割成數個集群。(P. Berkhin,2002)

二、個體間距離與相似性的量度

距離的量度(continuous data)：

距離的衡量是以點與點之間的距離為其測度，在連續型的資料中，一個最簡個簡單的測量方法就是使用曼哈頓距離，曼哈頓距離（Manhattan Distance）或計程車幾何是由十九世紀的赫爾曼・閔可夫斯基所創詞彙 ，是種使用在幾何度量空間的幾何學用語，用以標明兩個點在標準座標系上的絕對軸距總和。(Paul E. Black,1998)

例如在平面上，坐標（x1, y1）的點 P1 與坐標（x2, y2）的點 P2 的曼哈頓距離為： $|x_1 - x_2| + |y_1 - y_2|$.

要注意的是，曼哈頓距離依賴座標系統的旋轉，而非系統在座標軸上的平移或映射。

另外大多數較常採用的方法為歐幾里德距離(Euclidean distance)，在數學中，歐幾里得距離是歐幾里得空間中兩點間“普通”（即直線）距離。使用這個距離，歐氏空間成為度量空間。例如在 M 維空間中觀察了 N 個個體(observation)，每個個體有 M 個變數(variable)，則第 i 個個體與第 j 個個體之間的歐幾里德距離為量度個體兩兩之間的距離，將可建構一個 N×N 的距離矩陣，距離矩陣的第 i 行第 j 列代表第 i 個個體與第 j 個個體之間的距離，因此，這樣建構出來的距離矩陣將會是一個對角線數值為 0（因為自己與自己的距離為 0）的對稱矩陣（第 i 個個體到第 j 個個體的距離，與第 j 個個體到第 i 個個體的距離相等），如果各變數的衡量單位不同，在計算歐幾里德距離前宜將各變數之數值予以標準化，使其平均數為 0，標準差為 1。（Elena Deza,2009）

$$d_{ij} = \sqrt{\sum_{p=1}^{M} (x_{ip} - x_{jp})^2} \qquad i, j = 1, 2, \ldots, N$$

此外，量度距離的尺度除了歐幾里德距離(Euclidean distance) 外，還有很多不同的尺度及衡量方式，如：閔可夫斯基距離（Minkowski distance），此法定義 i, j 兩個個體之間的距離為：

$$d_{ij} = \left(\sum_{p=1}^{M} (\left| x_{ip} - x_{jp} \right|)^n \right)^{1/n} \qquad i, j = 1, 2, \ldots, N; \quad n = 1, 2, \ldots, \infty$$

其實，這是歐幾里德距離更一般化(general)的型式，在閔可夫斯基距離（Minkowski distance）中，若 n=2 即是上述所指稱的歐幾里德距離；若 n=1 則為最上述所說的曼哈頓距離（Manhattan distance），Manhattan distance 是以絕對長度來量度兩點之間的距離。(Renato Cordeiro de Amorim&Boris Mirkin, 2011)

三、群集分析的方法

群集分析依目的之不同，主要區分為階層群集方法 (hierarchical clustering method) 和非階層群集方法 (non-hierarchical clustering method) 兩種。茲分別簡述如下：

(一) 階層群集法 (hierarchical clustering method)

透過一種階層架構的方式，將資料層層反覆地進行分裂或聚合，以產生最後的樹狀結構這些群集可以用樹形圖(tree diagram or dendrogram) 來表示，可以看出群集之間的階層關係。要決定群集與群集（或個體）之間的距離，計算兩群間的距離有很多種方法 ，而本篇論文所使用的方法為 Ward(1963) 提出的華德法(Ward method)又稱最小變異數法 其實就是變異數分析中的組間平方和（between sum of square），同樣地，假設群集 G1、G2 依次有 n1、n2 個體，G1 與 G2 兩群距離是用 G1 群中心點 $\bar{g}_1$ 到兩群合併中心點 $\bar{\bar{g}}$ 的距離平方乘上 G1 群的個體數目 n1，再加上 G2 群中心點 $\bar{g}_2$ 到兩群合併中心點 $\bar{\bar{g}}$ 的距離平方乘上 G2 群的個體數目 n2，即

$$d(G1,G2) = n1 \cdot \left\| \bar{g}_1 - \bar{\bar{g}} \right\|^2 + n2 \cdot \left\| \bar{g}_2 - \bar{\bar{g}} \right\|^2$$

其中 $\bar{g}_1$， $\bar{g}_2$ 的定義如重心法所述，而合併中心點 $\bar{\bar{g}}$ 之定義為：

$$\bar{\bar{g}} = \frac{n1}{n1+n2}\bar{g}_1 + \frac{n1}{n1+n2}\bar{g}_2$$

(二) 非階層群集法 (non-hierarchical clustering method)

非階層群集法不分析其階層關係，主要探討可分割成的群集數目，以集群及裡面的個體，此法不像階層群集一樣可畫出樹形圖。此外，非階層群集分析方法有許多種，如：sequential threshold, parallel threshold, optimizing partitioning, K-means algorithm，目前本文僅針對最常使用的 K 組平均法（K-means algorithm）進行介紹。術語「k-均值」於 1967 年才被 James MacQueen [6 首次使用。

K-means 分群法 的主要概念是先隨機選取數量 n 個資料點，將之分別視為其 K 個群中心，再將 所有資料點找出對應之最近的群中心，群集產生後重新計算新的群中心，反覆疊代直到群中心不變為止。

$$\arg\min_{\mu} \sum_{c=1}^{K} \sum_{i=1}^{n_c} \|x_i - \mu_c\|^2 \Big|_{x_i \in S_c}$$

K-means 演算法(J. A. Hartigan & M. A. Wong,1979)

1. 隨機指派群集中心：在訓練組資料中「隨機」找出 K 筆紀錄來作為初始種子(初始群集的中心)

$$\mu_c^{(0)} \in R^d, c = 1,2,\dots,K$$

2. 產生初始群集：計算每一筆紀錄到各個隨機種子之間的距離，然後比較該筆紀錄究竟離哪一個隨機種子最近，然後這筆紀錄就會被指派到最接近的那個群集中心，此時就會形成一個群集邊界，產生了初始群集的成員集合。

3. 產生新的質量中心：根據邊界內的每一個案例重新計算出該群集的質量中心，利用新的質量中心取代之前的隨機種子，來做為該群的中心

$$S_c^{(t)} = \left\{ x_i : \left\| x_i - \mu_c^{(t)} \right\| \leq \left\| x_i - \mu_{c^*}^{(t)} \right\|, \forall i = 1, \ldots, n \right\}$$

4. 變動群集邊界：

指定完新的質量中心之後，再一次比較每一筆紀錄與新的群集中心之間的距離，然後根據距離，再度重新分配每一個案例所屬的群集

$$\mu_c^{(t+1)} = \frac{sum(S_c^{(t)})}{n_c} = \sum_{i=1}^{n_c} x_i \Big|_{x_i \in S_c^{(t)}}$$

5. 持續反覆 3, 4 的步驟，一直執行到群集成員不再變動為止

$$S_c^{(t+1)} = S_c^{(t)}, \forall c = 1, \ldots, K$$

(三) 二階段群集法 (two-stages clustering technique)

二階段群集法是結合階層群集法與非階層群集法，此法第一階段先用「階層群集法」來做分群，以決定分群個數；爾後，第二階段再以「非階層群集法」（如：K-means method）進行群集，採用二階段群集法之目的是因為在第一階段「階層群集法」中，一旦兩個個體被分在同一群，往後就永遠在同一群內，故先採用「階層群集法」來決定分群數目，再藉由「非階層群集法」做分層，便可避掉上述的缺點。(D. Arthur & S. Vassilvitskii , 2006:)

第三章 研究方法

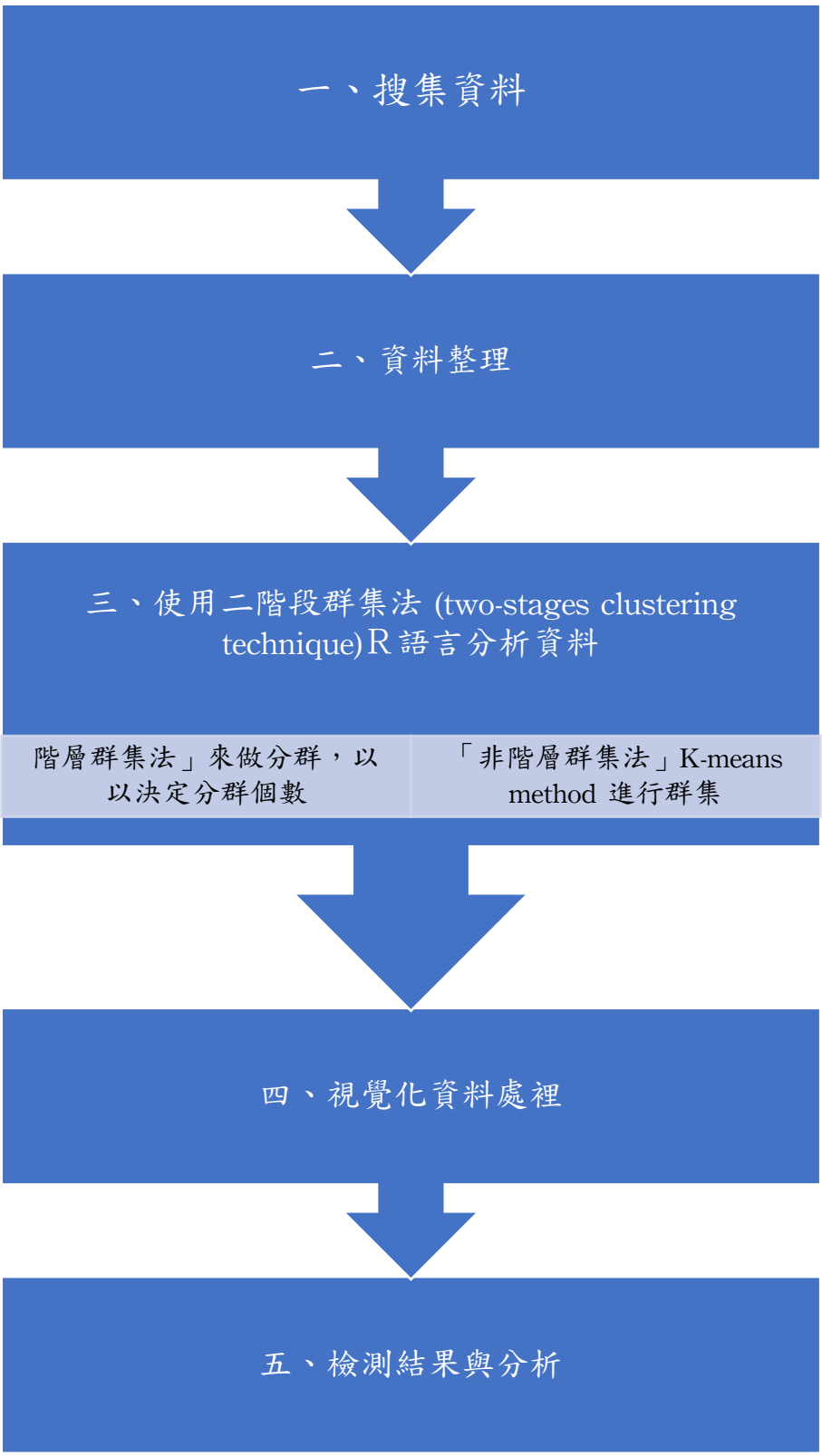

一、搜集資料
二、資料整理
三、使用二階段群集法 (two-stages clustering technique) R 語言分析資料
階層群集法」來做分群，以以決定分群個數
「非階層群集法」K-means method 進行群集
四、視覺化資料處裡
五、檢測結果與分析

圖 1

一、搜集資料

本論文的數據原始資料來自兩處，其一是來自美國教育部(US Department of Education, Office of Postsecondary Edu- cation) 的 Equity in Athletes Data Analysis。其二是來自 Sports Reference 公司旗下的 basketball reference 網站，根據 SimilarWeb 的報導指出，Sports Reference 旗下的 Basketball-Reference 已經成為僅次於 NBA.com ，光去年就有十億了瀏覽量，不管球迷、球員擊球平等任何想要分析運動賽事的人都會使用這個網站。因此此份論文的資料來源兩者皆具有可信度去做此次分析。

(一)Database 介紹

1.The Equity in Athletics Data Analysis

是來自美國教育部 Office of Postsecondary Education（ＯＰＥ）提供的體育數據分析工具。此分析工具旨在提供大眾快速的體育數據客製化報告。這屆數據來自 OPE Equity in Athletics Disclosure 網站的資料集. 由於《公平體育披露法》Equity in Athletics Disclosure Act (EADA)的要求, 因此所有參加聯邦學生資助的機構都要每年提交校內體育數據和他校際合作比賽的資料。

2.Sports References

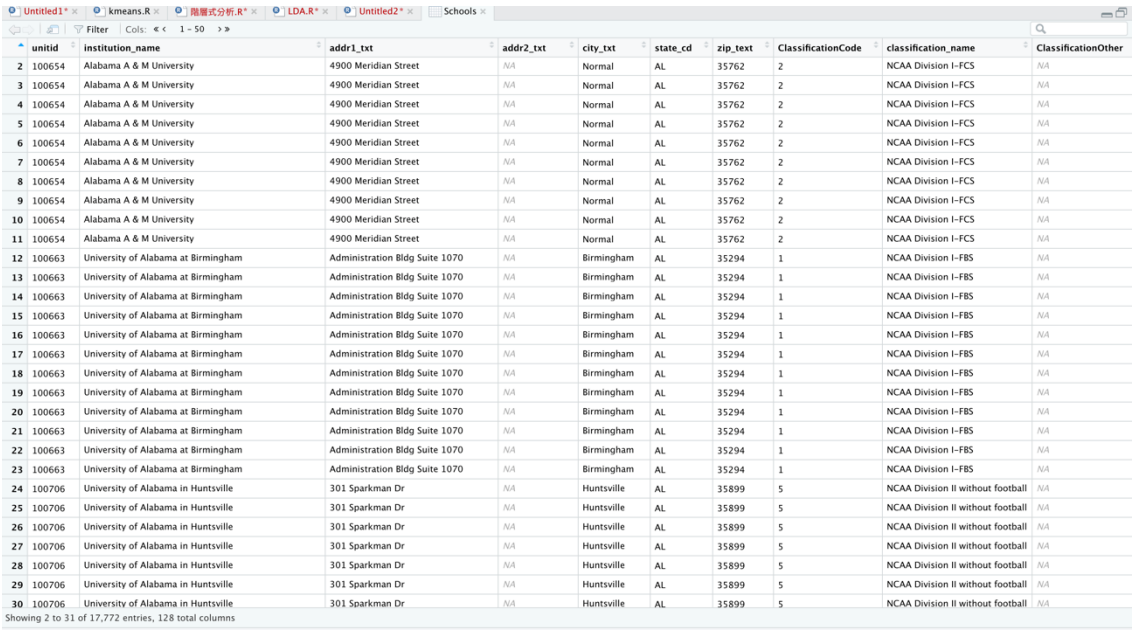
Sports Reference，LLC 是一家位於賓夕法尼亞州的公司，經營著幾個與體育有關的網站，包括 Sports-Reference.com，棒球的 Baseball-Reference.com，籃球的 Basketball-Reference.com，冰球的 Hockey-Reference.com，Pro- Football-Reference.com 用於美式足球，而 FBref.com 用於協會足球（足球）。該網站還包括有關大學橄欖球，

大學籃球和奧運會的部分。「我每天都會使用 Baseball-Reference ，」美國職棒大聯盟明尼雙城隊總管 Thad Levine 如此說道，他們甚至雇用了專業的數據分析專員。很多團隊也是如此，他們利用 Basebal Reference 大量蒐集資料，並經過整理轉化為對自己球隊最為有利的數據資料。

二、資料整理

資料整理為本篇論文最為費時的環節，由於美國教育部網站即使可以客製化資料變數還是很多，資料量還是很大，所以要處有一些 missing 值 及不需要用到的欄位，所以要透過觀察去取捨該如何處理，而每一筆資料的屬性不同造成的影響也不一樣。另外本篇資料是由兩份不同的資料及合併因此原先用 Excel 的 VLOOKUP 去比對兩組資料，發現他們對於大學名稱資料的紀錄方式是很雜亂的 例如: 教育部的資料有些會寫全名和有些寫簡寫沒有統一格式的規定。所以沒辦法使用 VLOOKUP 去輕易合併兩組資料。因此，本論文是依照字母排序，三百筆資料依序依序比對去合併，由於對於美國大學的別名不是那麼清楚，大約花了 3~5 小時去做資料的整理。

1. 首先使用 R 來看 The Equity in Athletics Data Analysis 的原始資料大約原始資料大約有一萬多筆，128 個變數，因此要用篩選功能來選擇我要使用的變數-學校、男子籃球隊收入及支出。



Untitled1* | kmeans.R | 階層式分析.R* | LDA.R* | Untitled2* | Schools

Filter | Cols: « < 1 – 50 > »

	unitid	institution_name	addr1_txt	addr2_txt	city_txt	state_cd	zip_text	ClassificationCode	classification_name	ClassificationOther
2	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
3	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
4	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
5	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
6	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
7	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
8	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
9	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
10	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
11	100654	Alabama A & M University	4900 Meridian Street	NA	Normal	AL	35762	2	NCAA Division I–FCS	NA
12	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
13	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
14	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
15	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
16	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
17	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
18	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
19	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
20	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
21	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
22	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
23	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070	NA	Birmingham	AL	35294	1	NCAA Division I–FBS	NA
24	100706	University of Alabama in Huntsville	301 Sparkman Dr	NA	Huntsville	AL	35899	5	NCAA Division II without football	NA
25	100706	University of Alabama in Huntsville	301 Sparkman Dr	NA	Huntsville	AL	35899	5	NCAA Division II without football	NA
26	100706	University of Alabama in Huntsville	301 Sparkman Dr	NA	Huntsville	AL	35899	5	NCAA Division II without football	NA
27	100706	University of Alabama in Huntsville	301 Sparkman Dr	NA	Huntsville	AL	35899	5	NCAA Division II without football	NA
28	100706	University of Alabama in Huntsville	301 Sparkman Dr	NA	Huntsville	AL	35899	5	NCAA Division II without football	NA
29	100706	University of Alabama in Huntsville	301 Sparkman Dr	NA	Huntsville	AL	35899	5	NCAA Division II without football	NA
30	100706	University of Alabama in Huntsville	301 Sparkman Dr	NA	Huntsville	AL	35899	5	NCAA Division II without football	NA

Showing 2 to 31 of 17,772 entries, 128 total columns

圖 2

2. 勝率跟聯盟的資料是從 Sports References(圖 3 及圖 4)複製下來整理。

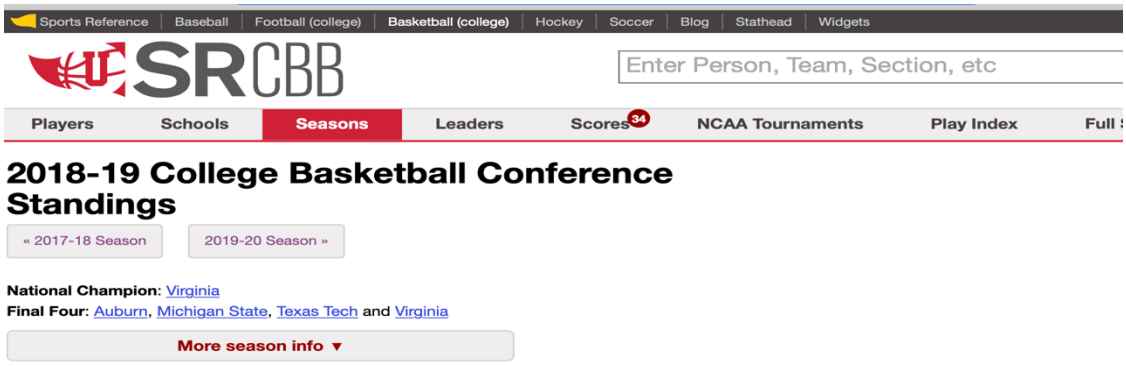
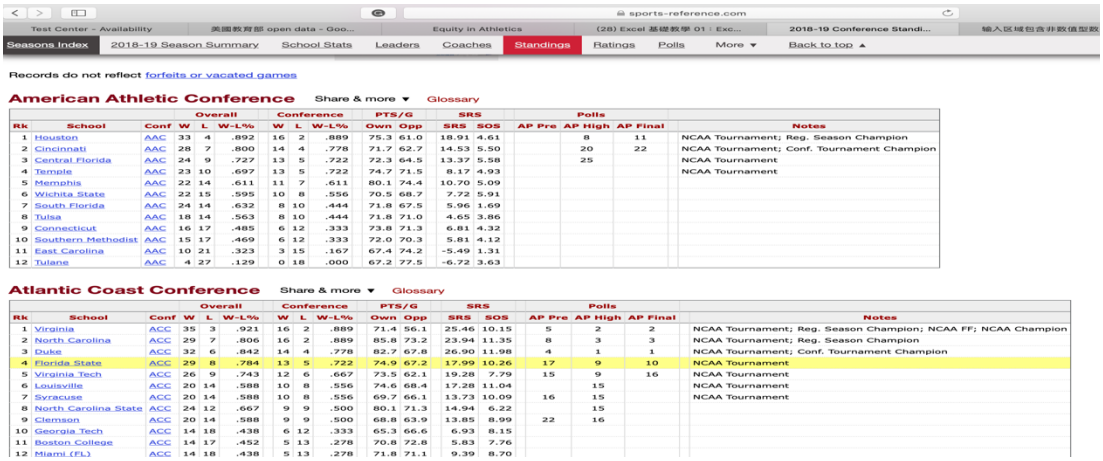


圖 3



Records do not reflect forfeits or vacated games

American Athletic Conference Share & more Glossary

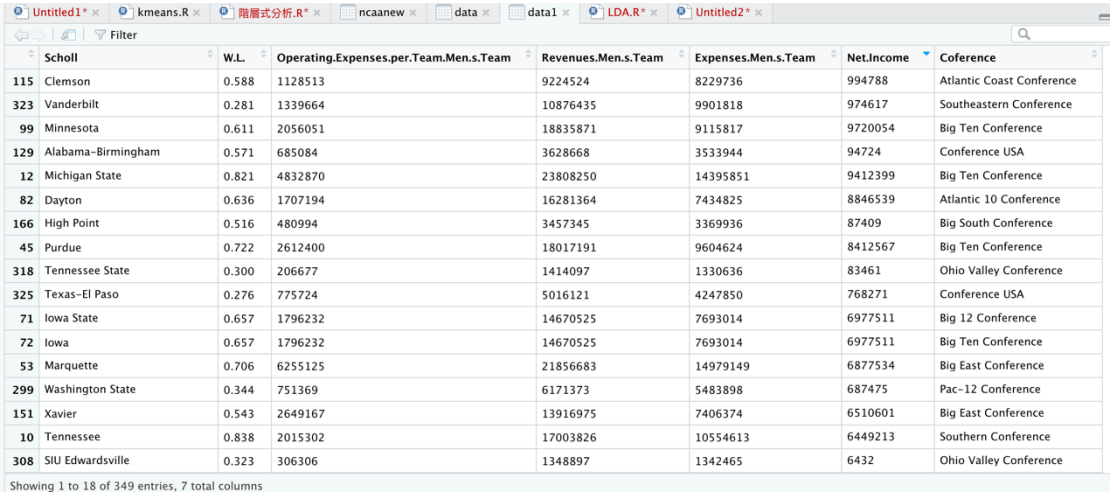
Rk	School	Conf	Overall W	L	W-L%	Conference W	L	W-L%	PTS/G Own	Opp	SRS	SOS	Polls AP Pre	AP High	AP Final	Notes
1	Houston	AAC	33	4	.892	16	2	.889	75.3	61.0	18.91	4.61		8	11	NCAA Tournament; Reg. Season Champion
2	Cincinnati	AAC	28	7	.800	14	4	.778	71.7	62.7	14.53	5.50		20	22	NCAA Tournament; Conf. Tournament Champion
3	Central Florida	AAC	24	9	.727	13	5	.722	72.3	64.5	13.37	5.58		25		NCAA Tournament
4	Temple	AAC	23	10	.697	13	5	.722	74.7	71.5	8.17	4.93				NCAA Tournament
5	Memphis	AAC	22	14	.611	11	7	.611	80.1	74.4	10.70	5.09				
6	Wichita State	AAC	22	15	.595	10	8	.556	70.5	68.7	7.72	5.91				
7	South Florida	AAC	24	14	.632	8	10	.444	71.8	67.5	5.96	1.69				
8	Tulsa	AAC	18	14	.563	8	10	.444	71.8	71.0	4.65	3.86				
9	Connecticut	AAC	16	17	.485	6	12	.333	73.8	71.3	6.81	4.32				
10	Southern Methodist	AAC	15	17	.469	6	12	.333	72.0	70.3	5.81	4.12				
11	East Carolina	AAC	10	21	.323	3	15	.167	67.4	74.2	-5.49	1.31				
12	Tulane	AAC	4	27	.129	0	18	.000	67.2	77.5	-6.72	3.63				

Atlantic Coast Conference Share & more Glossary

Rk	School	Conf	Overall W	L	W-L%	Conference W	L	W-L%	PTS/G Own	Opp	SRS	SOS	Polls AP Pre	AP High	AP Final	Notes
1	Virginia	ACC	35	3	.921	16	2	.889	71.4	56.1	25.46	10.15	5	2	2	NCAA Tournament; Reg. Season Champion; NCAA FF; NCAA Champion
2	North Carolina	ACC	29	7	.806	16	2	.889	85.8	73.2	23.94	11.35	8	3	3	NCAA Tournament; Reg. Season Champion
3	Duke	ACC	32	6	.842	14	4	.778	82.7	67.8	26.90	11.98	4	1	1	NCAA Tournament; Conf. Tournament Champion
4	Florida State	ACC	29	8	.784	13	5	.722	74.9	67.2	17.99	10.26	17	9	10	NCAA Tournament
5	Virginia Tech	ACC	26	9	.743	12	6	.667	73.5	62.1	19.28	7.79	15	9	16	NCAA Tournament
6	Louisville	ACC	20	14	.588	10	8	.556	74.6	68.4	17.28	11.04		15		NCAA Tournament
7	Syracuse	ACC	20	14	.588	10	8	.556	69.7	66.1	13.73	10.09	16	15		NCAA Tournament
8	North Carolina State	ACC	24	12	.667	9	9	.500	80.1	71.3	14.94	6.22		15		
9	Clemson	ACC	20	14	.588	9	9	.500	68.8	63.9	13.85	8.99	22	16		
10	Georgia Tech	ACC	14	18	.438	6	12	.333	65.3	66.6	6.93	8.15				
11	Boston College	ACC	14	17	.452	5	13	.278	70.8	72.8	5.83	7.76				
12	Miami (FL)	ACC	14	18	.438	5	13	.278	71.8	71.1	9.39	8.70				

圖 4

3. 最終合併兩者整理生成論文所使用的資料集



Untitled1* | kmeans.R | 階層式分析.R* | ncaanew | data | data1 | LDA.R* | Untitled2*

Filter

	Scholl	W.L.	Operating.Expenses.per.Team.Men.s.Team	Revenues.Men.s.Team	Expenses.Men.s.Team	Net.Income	Coference
115	Clemson	0.588	1128513	9224524	8229736	994788	Atlantic Coast Conference
323	Vanderbilt	0.281	1339664	10876435	9901818	974617	Southeastern Conference
99	Minnesota	0.611	2056051	18835871	9115817	9720054	Big Ten Conference
129	Alabama-Birmingham	0.571	685084	3628668	3533944	94724	Conference USA
12	Michigan State	0.821	4832870	23808250	14395851	9412399	Big Ten Conference
82	Dayton	0.636	1707194	16281364	7434825	8846539	Atlantic 10 Conference
166	High Point	0.516	480994	3457345	3369936	87409	Big South Conference
45	Purdue	0.722	2612400	18017191	9604624	8412567	Big Ten Conference
318	Tennessee State	0.300	206677	1414097	1330636	83461	Ohio Valley Conference
325	Texas-El Paso	0.276	775724	5016121	4247850	768271	Conference USA
71	Iowa State	0.657	1796232	14670525	7693014	6977511	Big 12 Conference
72	Iowa	0.657	1796232	14670525	7693014	6977511	Big Ten Conference
53	Marquette	0.706	6255125	21856683	14979149	6877534	Big East Conference
299	Washington State	0.344	751369	6171373	5483898	687475	Pac-12 Conference
151	Xavier	0.543	2649167	13916975	7406374	6510601	Big East Conference
10	Tennessee	0.838	2015302	17003826	10554613	6449213	Southern Conference
308	SIU Edwardsville	0.323	306306	1348897	1342465	6432	Ohio Valley Conference

Showing 1 to 18 of 349 entries, 7 total columns

圖 5

初始資料集摘要：

表 1

變數名稱：	School	學校
	W.L.	勝率
	Operating.Expense.per.Team.Men.sTeam	營業費用
	Revenues.Men.s.Team	球隊收入
	Expenses.Men.s.Team	球隊支出
	Net.Income	淨利
	Conference	聯盟
欄位數：	7	
列位數：	349	

4. 使用 R 內建指令做標準化、清除 NA 值及排除不需要用到的變數，因此會生產出一個新的資料集 xxx 供二階段群集法分析使用

圖 6

三、使用二階段群集法 (two-stages clustering technique)

R 語言分析資料

採用二階段群集法之目的是因為在第一階段「階層群集法」中，一旦兩個個體被分在同一群，往後就永遠在同一群內，故先採用「階層群集法」來決定分群數目，再藉由「非階層群集法」做分層，便可避掉上述的缺點，會使分群結果更加穩定。

(一) 階層式分析- (hierarchical clustering method)

.本篇論問先使用歐式距離透過個體間距離與相似性的量度，看結果會大致分成幾群，再使用 hierarchical clustering method 階層架構中的華德法方式，將資料層層反覆地進行分裂或聚合，以產生最後的樹狀結構這些群集可以用樹形圖(tree diagram or dendrogram) 來表示，可以比對看出群集之間的而本篇論文所使用的方法為 Ward(1963) 提出的華德法又稱最小變異數法去決定之後非階層式分析中機器學習-kmeans 需要用的分群個數。圖 7 圖 8 圖 9。

```
E.dist <- dist(x = xxx, method = "euclidean")#計算歐幾里得距離
# 將以上資料間距離作為參數投入階層式分群函數：hclust()
# 使用歐式距離進行分群
h.E.cluster <- hclust(E.dist)
plot(h.E.cluster, xlab="歐式距離",family="黑體-繁 中黑")

dev.off()
plot(hclust(E.dist, method="ward.D2"), xlab = "華德法: Ward's Method")  # 華德法
```

圖 7

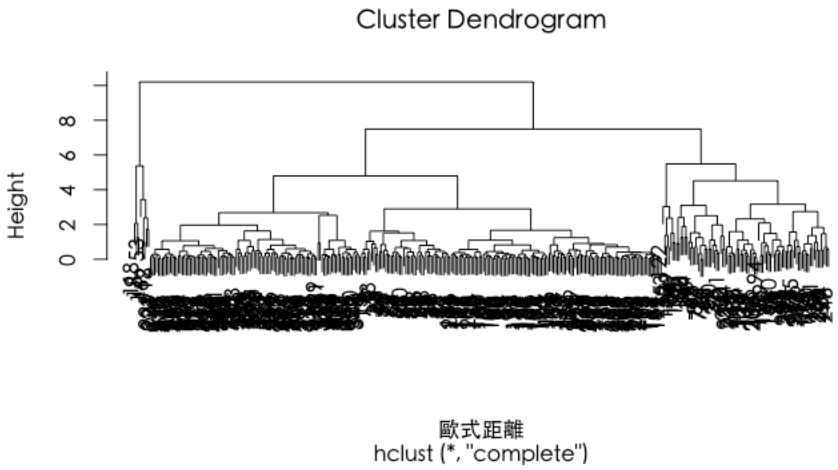


圖 8

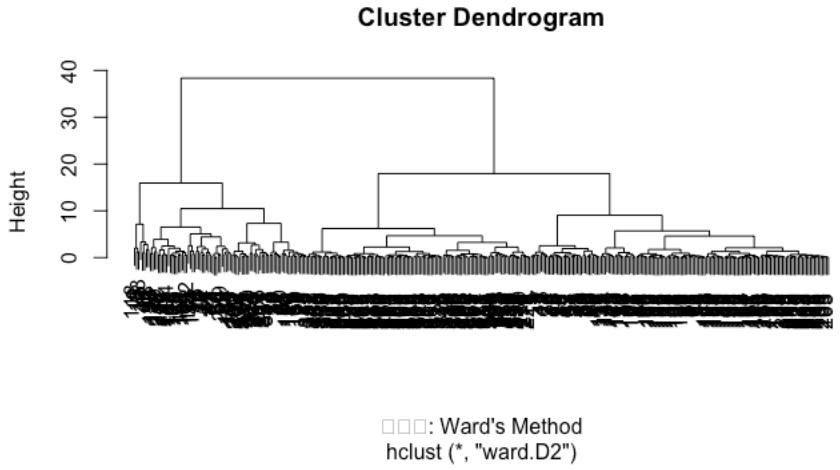


圖 9

很明顯從圖 8 及圖 9，在第三個階層中可以看出華德法跟歐式距離可以把全部資料分為五筆。

因此為了供讀者再看清楚，使用聚合演算法將兩個方法再搭配起來並劃出五個資料群。圖 10 及圖 11。

```r
#我們聚焦在採用歐式距離搭配華德最小變異聚合演算法
dev.off()
par(family="黑體-繁 中黑")
plot(hclust(E.dist, method="ward.D2"), xlab = "華德法: Ward's Method")
rect.hclust(tree = hclust(E.dist, method="ward.D2"), k = 5, border = "red")
```

圖 10

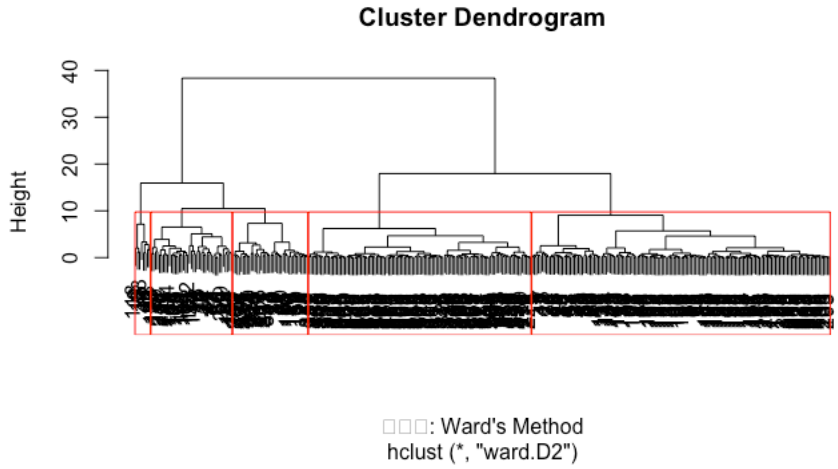


圖 11

(二) 非階層式分析-K means 法

接者使用 Kmeans 將分組結果合併到原始資料去，是先隨機選取數量 1000 個資料點(第一次產生時亂數時, 從當下 時間 (current time), 生成一個 種子 (seed) 出發, 不斷迭代更新產生隨機均等分配亂數 (uniform random number),所以不同時間下執行 R, 啟用不同的種子, 內部的隨機種子會跟者改變了, 設定好模 擬亂數是不會重復的,)，將之分別視為其 5 個群中心，再將 所有資料點找出對應之最近的群中心，群集產生後重新計算新的群中心，反覆疊代直到群中心不變為止並將結果合併至原始資料中。圖 12～圖 15。

```r
set.seed(1000)
km3<-kmeans(xxx, centers=5) #kmeans分群
km3$size #各群size
head(km3) #各群相關係數
ggpairs(xxx,columns = 1:4,mapping=aes(colour=as.character(km3$cluster)))
ncaanew<-cbind(data1, cbind(km3$cluster)) #將分群的結果，與原資料data1合併

names(ncaanew)[8] <- "clustergroup" #重新命名
```

圖 12

圖 13 為各群資料大小

```
> km3$size #各群size
[1]  15 104  70  82  78
```

圖 13

圖 14 為各群相關係數

```
$centers
        W.L. Operating.Expenses.per.Team.Men.s.Team Revenues.Men.s.Team Expenses.Men.s.Team
1  1.0719900                             3.1086504           3.1737460           2.7674210
2 -0.2030123                            -0.5143350          -0.4845196          -0.5469261
3  0.3873531                             1.0740176           1.0104320           1.2648771
4  0.9845576                            -0.3311190          -0.3743677          -0.3742252
5 -1.3181412                            -0.5277999          -0.4775416          -0.5446914
```

圖 14

圖 15 為重新合併的資料集，多出 clustergroup 分類完的變數，為之後視覺化資料分析做使用。

Revenues.Men.s.Team	Expenses.Men.s.Team	Net.Income	Coference	clustergroup
1107986	2126312	(1018326)	Sun Belt Conference	5
2122220	2140282	(18062)	Horizon League	5
6910271	7095493	(185222)	Atlantic Coast Conference	3
2238825	2438136	(199311)	Big West Conference	5
596992	815704	(218712)	Southwestern Athletic Conf.	5
8484584	10798893	(2314309)	Big 12 Conference	3
1833808	2094896	(261088)	Metro Atlantic Athletic Conference	2
1274977	1301285	(26308)	Big Sky Conference	2
2064588	2392346	(327758)	America East Conference	4
11959355	15468381	(3509026)	Pac-12 Conference	3
6769220	7143168	(373948)	Pac-12 Conference	3
2198323	2604565	(406242)	Sun Belt Conference	2
4153208	8331273	(4178065)	Atlantic Coast Conference	3
1974451	2452114	(477663)	Big Sky Conference	4
6275440	6901843	(626403)	Pac-12 Conference	3
3879707	4594025	(714318)	American Athletic Conference	4
1591652	1667122	(75470)	Sun Belt Conference	2

Showing 1 to 18 of 349 entries, 8 total columns

圖 15

四、視覺化資料處裡

接者我們使用新的資料集來進行視覺化資料分析去找出可以衡量球隊商業價值的 insight。本文使用圖 16 下 package 去撰寫圖 17 的語法。

```
library(factoextra)
library(cluster)
library(ggplot2)
library(ggfortify)
library(GGally)
```

圖 16

```
#看成績分布群框
  ggplot(ncaanew,aes(x = clustergroup, y = W.L.)) +
  geom_jitter() +
  stat_summary(fun.y = median, colour = "red", geom = "point", size = 5)
#看支出分布群框
  ggplot(ncaanew,aes(x = clustergroup, y =Expenses.Men.s.Team)) +
  geom_jitter() +
  stat_summary(fun.y = median, colour = "red", geom = "point", size = 5)
#看收入分布群框
  ggplot(ncaanew,aes(x = clustergroup, y =Revenues.Men.s.Team)) +
  geom_jitter() +
  stat_summary(fun.y = median, colour = "red", geom = "point", size = 5)
#看收入支出的關西
  qplot(Revenues.Men.s.Team,Expenses.Men.s.Team, data = ncaanew, color=as.character(clustergroup))
#聯盟差距
  qplot(Revenues.Men.s.Team,Coference, data = ncaanew, color=as.character(clustergroup))
#勝率跟收入的分佈
  qplot(W.L.,Revenues.Men.s.Team , data = ncaanew, color=as.character(clustergroup))
#聯盟表現
  qplot(W.L.,Coference, data = ncaanew, color=as.character(clustergroup))
#支出分佈
  qplot(Expenses.Men.s.Team , data = ncaanew, color=as.character(clustergroup))
```

圖 17

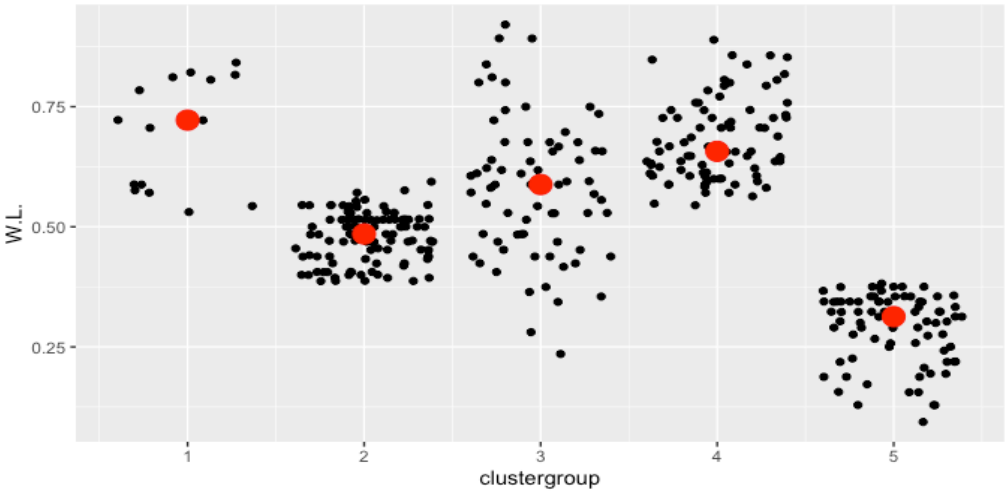


圖 18

從圖 18 可以看出來第一群及第三群的球隊數勝率比較分散，然而第三、第四及第五群的球隊數勝率比較集中。再者，第一群平均勝率跟第四群的平均勝率較高，第二群的平均勝率明顯表現普通集中在 0.5，而最後第五組勝率表現明顯差勁全都低於 0.5。

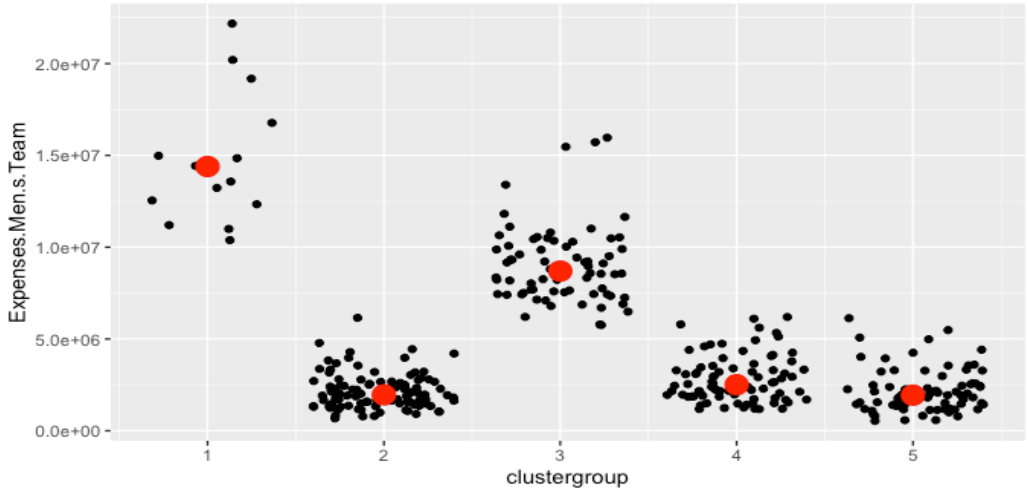


圖 11

圖 19 在說明各群學校球隊的支出狀況，很明顯的第一群支出遠遠高於其他群，區間落在大約 1000 萬美元至大約 2400 萬美元，第二高

支出的來自第三群，區間位於 500 萬美元至 1500 萬美元。至於其他三群，支出都未逾五百萬美元以下。

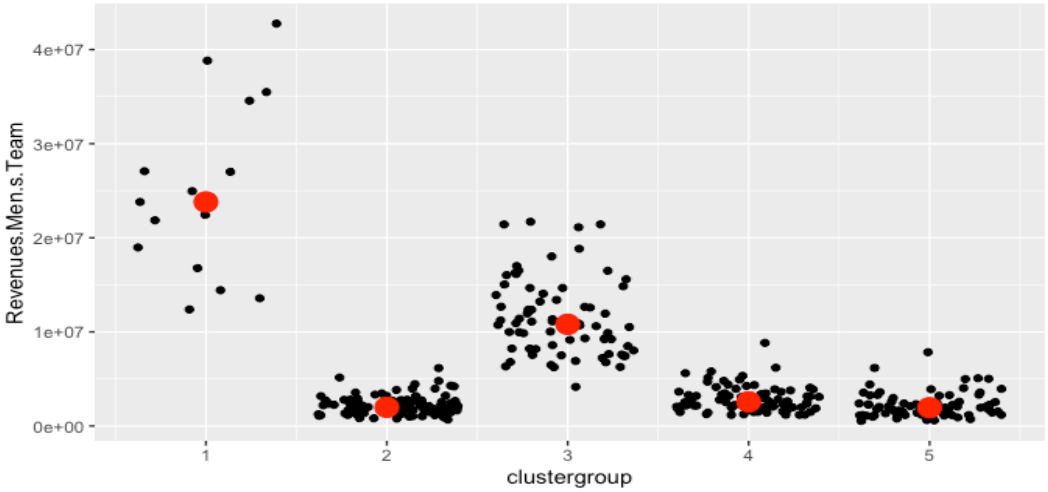


圖 2

圖 20 可以發現收入跟支出有很大的相似，其中最引人注意的是第一群。他們平均支出大約落在 1500 萬美元，而平均收入大約落在 2300 萬美元，可以大致推估第一群的球隊多有所賺錢，再者有四所學校球隊收入高達 3500 萬美元至 4200 萬美元遠遠超越其他學校。除了第三群平均收入有 1000 萬美元，第二、四及五群都低於五百萬美元。

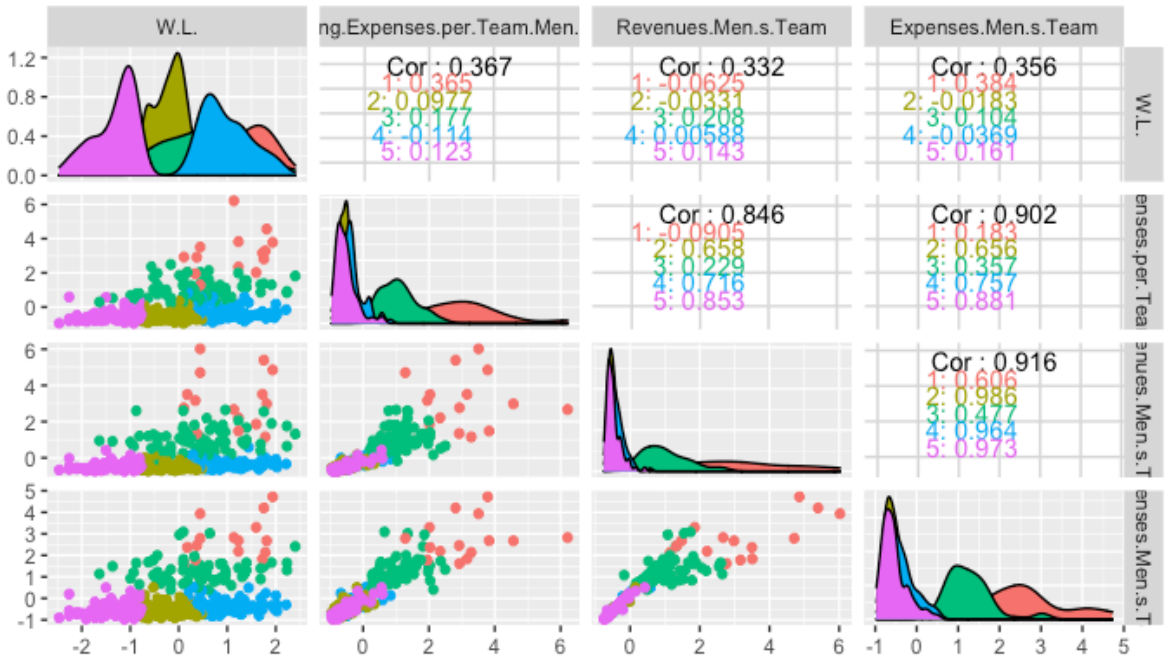


圖 21

圖 21 在說明資料間變數之間相關性，從勝率可以看出它與收入及支出相關性不大（大致落於 3%），也就是指勝率對於球隊收入跟支出沒有直接太大的影響，球隊會不會贏球跟學校所投資在球員、設備及教練等沒有很強的相關性，另外球隊的勝率跟所賺的轉播金、門票收入及周邊商品等也沒有很強的影響。至於支出跟收入的顯著的正相關（大約 91.6%），可以推測學校願意投資在球隊身上越多所賺的就越多，反之亦然。

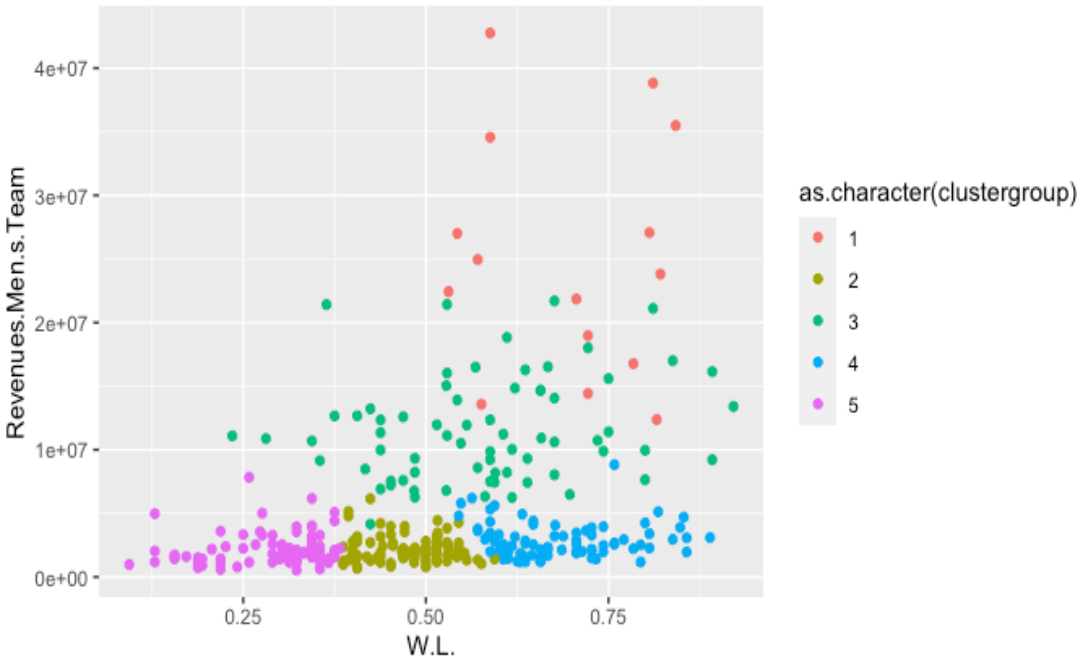


圖 22

圖 22 是從圖 21 勝率及球隊收入中拉出來的小圖，可以看出第一群勝率就算沒到 0.75 也可以取得巨額收益，然而第四群擁有最多勝率超過 0.75 的球隊數但卻沒賺到什麼錢。

圖 23

圖 23 中是從圖 21 中拉出來的小圖，除了高收入及高支出的第一群跟第三群外，其他群收入跟支出的球隊趨近一條回歸線成正比。

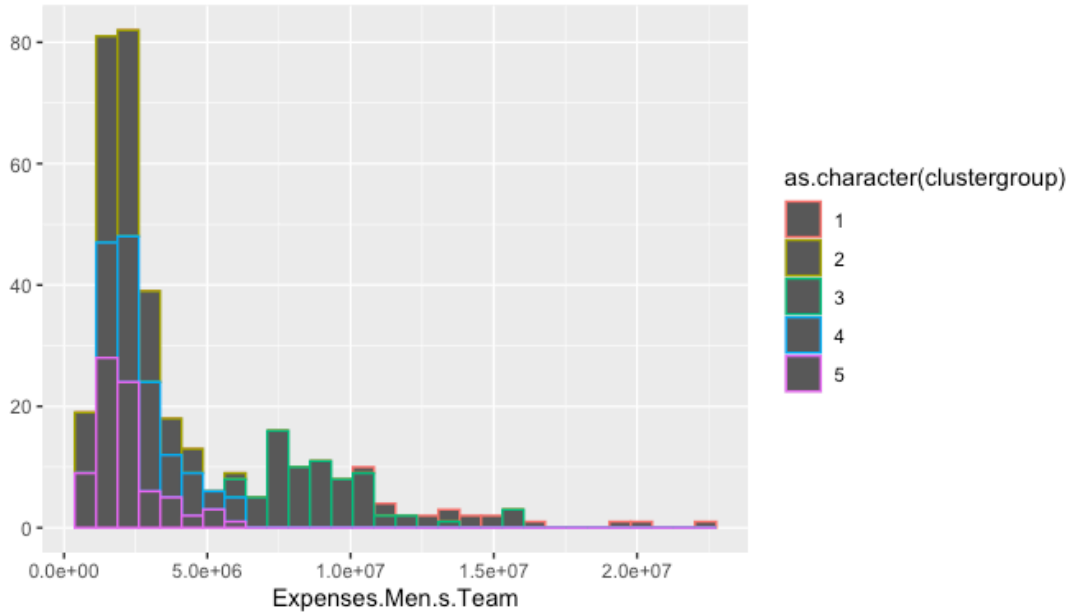


圖 24

圖 24 中可以更清楚看清楚各群的支出分佈。大部分多數的球隊的支出都分佈在 200~300 萬美元，而少數第一群跟第三群球隊支出分佈在 700 萬美元在 2000 萬美元。

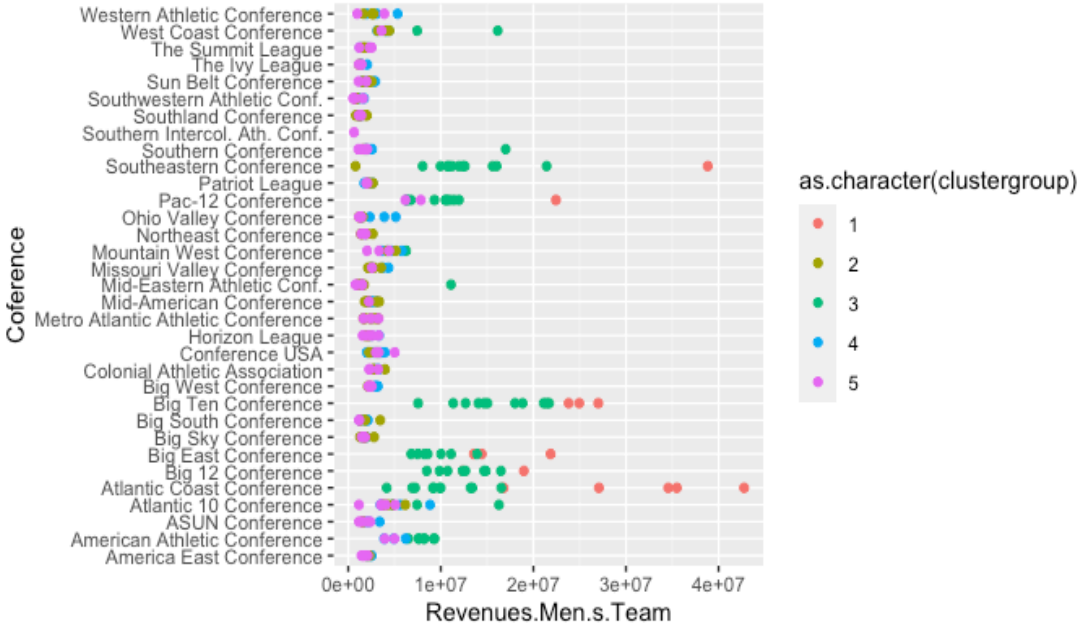


圖 25

圖 25 可以看出各個聯盟之間的收益差異，第一群跟第三群明顯集中在較特定且有名的聯盟例如：Big10、Big12、ACC 和 Big East。

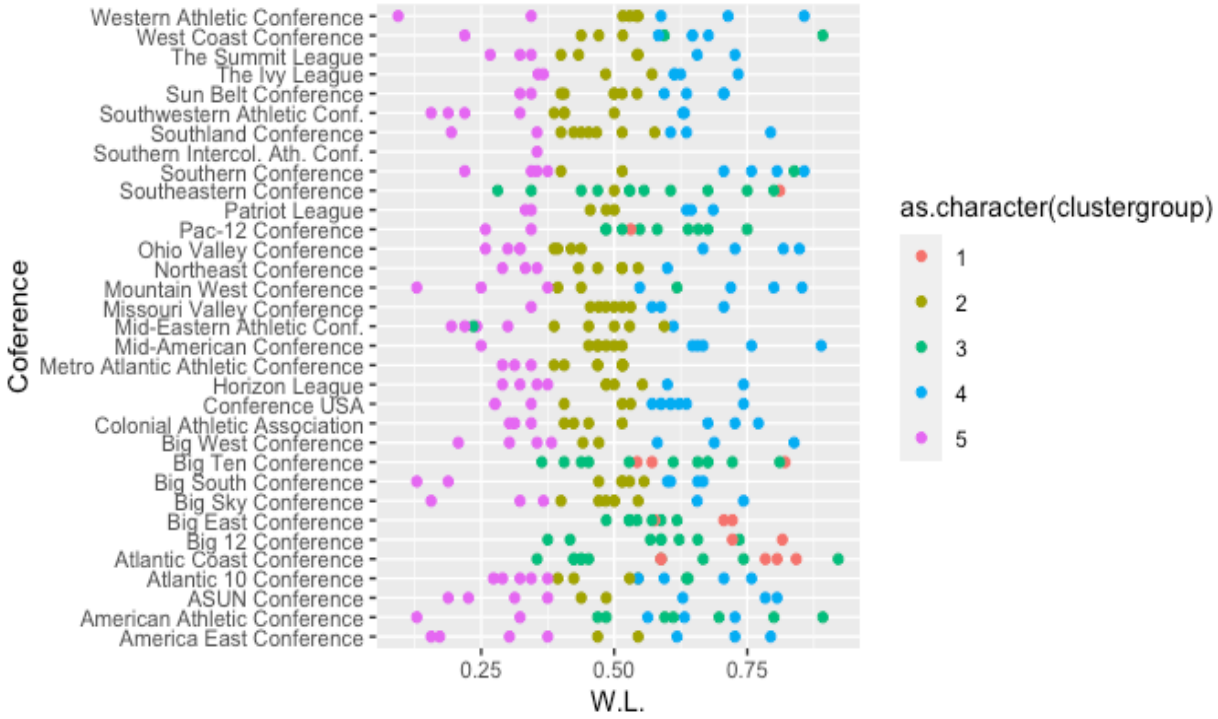


圖 26

圖 26 可以明顯看出個聯盟球隊勝率表現分佈差不多，至於勝率大於 0.75 的最多者為 ACC（其中多為第一群傳統籃球名校 Duke、Louisville、Gatech 及第三群 Virginia、Florida State，次多為 Southern Conference 三個來自第四群 East Tennessee State、Furman、Abilene Christian 皆一不是著名的籃球名校，，一個來自第三群 Tennessee 籃球隊不受注目，但 NBA 球員產出率還不差錯)

五、檢測分群結果與分析-各群定義及說明

(一)第一群：天之驕子球隊

表 2

W.L.		Revenues.Men.s.Team		Expenses.Men.s.Team	
平均數	0.695133333	平均數	24994579.4	平均數	14750531
標準誤	0.030066413	標準誤	2429894.2	標準誤	897106.34
中間值	0.722	中間值	23808250	中間值	14395851
眾數	0.722	眾數	#N/A	眾數	#N/A
標準差	0.116446718	標準差	9410939.76	標準差	3474477.9
變異數	0.013559838	變異數	8.8566E+13	變異數	1.207E+13
峰度	-1.804054235	峰度	-0.7208405	峰度	0.087458
偏態	-0.178401238	偏態	0.47528424	偏態	0.8967131
範圍	0.311	範圍	30376349	範圍	11794843
最小值	0.531	最小值	12382899	最小值	10383630
最大值	0.842	最大值	42759248	最大值	22178473
總和	10.427	總和	374918691	總和	221257959
個數	15	個數	15	個數	15

1. 天之驕子球隊定義原因：群內的球隊都是三高球隊，勝率高、收入高及支出高，顯示他們球隊表現優異，有龐大轉播金支持、門票收入或周邊商品銷售佳，再者他們也願意花大筆錢聘請教練人才及投資籃球訓練設備。

2. 指標分析

- 平均勝率：大約 0.7 為五群中最高。
- 總收入：約 3.7 億美元
- 平均收入：約 2500 萬美元為五群之中最高，大約為補習班球隊收入的兩倍及平庸之輩、待開發模範生、不及格球隊的十倍。
- 總支出：約 2.2 億美元，15 支球隊的總支出相近於平庸之輩 105 支球隊的總支出。

- 群內收入差：約 3000 萬美元，表示有極端值球隊收入遠遠超過所有學校的收入。例如：Louisville（ACC）-約 4300 萬美元、Kentucky（SEC）-約 3900 萬美元、Duke(ACC) -約 3500 萬美元、Syracuse(ACC)-約 3450 萬美元為天之驕子賺最多的前四名，然而最低 Texas Tech 除了沒超越 Duke 的勝率外，收入僅約為 1000 萬美元。

(二)第二群：平庸之輩球隊

表 3

W.L.		Revenues.Men.s.Team		Expenses.Men.s.Team	
平均數	0.47853846	平均數	2187474.99	平均數	2160566.99
標準誤	0.00505773	標準誤	96959.3579	標準誤	94061.5558
中間值	0.485	中間值	1988544	中間值	1959669.5
眾數	0.5	眾數	#N/A	眾數	#N/A
標準差	0.05157894	標準差	988795.316	標準差	959243.417
變異數	0.00266039	變異數	9.7772E+11	變異數	9.2015E+11
峰度	-0.9010265	峰度	2.40014671	峰度	2.38665179
偏態	-0.1840477	偏態	1.32906634	偏態	1.27018377
範圍	0.207	範圍	5476203	範圍	5476203
最小值	0.387	最小值	676673	最小值	676673
最大值	0.594	最大值	6152876	最大值	6152876
總和	49.768	總和	227497399	總和	224698967
個數	104	個數	104	個數	104

1. 平庸之輩球隊定義原因：勝率、收入及支出都表現平平，其學校知名度也較低

2. 指標分析

平均勝率：0.5。

- 勝率變異數：接近 0，表示群內大家勝率集中相似接近於 0.5，表現都平平。
- 總收入：約 2.3 億美元

- 平均收入：約 218 萬美元
- 總支出：約 2.2 億美元
- 平均收入：約 216 萬美元
- 峰度：收入跟支出的峰度都約 2.多為高尖峰，表示有極端傾向。

(三)第三群：補習班型球隊

表 4

W.L.		Revenues.Men.s.Team		Expenses.Men.s.Team	
平均數	0.57882857	平均數	11507606.5	平均數	9042928.9
標準誤	0.01735143	標準誤	490846.919	標準誤	250104.492
中間值	0.588	中間值	10805137.5	中間值	8697842.5
眾數	0.676	眾數	14670525	眾數	7693014
標準差	0.14517246	標準差	4106719.96	標準差	2092524.31
變異數	0.02107504	變異數	1.6865E+13	變異數	4.3787E+12
峰度	-0.0700732	峰度	0.07005509	峰度	2.70609775
偏態	0.16569601	偏態	0.76330115	偏態	1.37238734
範圍	0.686	範圍	17548660	範圍	10206060
最小值	0.235	最小值	4153208	最小值	5760815
最大值	0.921	最大值	21701868	最大值	15966875
總和	40.518	總和	805532455	總和	633005023
個數	70	個數	70	個數	70

1. 補習班球隊定義原因：他們平均支出接近 900 萬，大於平庸之輩、待開發模範生及無價值球隊平均支出的加總，另外勝率差大表現球隊資質差距極大，有如高中補習班同學付一樣高額的學費但是成績落差極大。在補習班球隊中，有本季度冠軍 Virginia(ACC)，但知名度不及 ACC 聯盟球隊，被視為黑馬。然而同樣在 ACC 聯盟裡的 Wake Forest 雖然勝率差約 0.36，但他們收入也有高達 1100 萬美元大於勝率高的待開發模範生球隊，至於知名度他們也誕生出許多知名 NBA 球員球員例如：馬刺明星大前鋒「石佛」Tim Duncan 及明星控衛「CP3」Crist Paul。

2. 指標分析

- 平均勝率：約 0.58。
- 最大勝率：約 0.92
- 最小勝率：約 0.24
- 勝率差：約 0.7 為五群中最大，群內有本季度冠軍 Virginia 也有表現極差的球隊如 South Carolina State。
- 總收入：約 8 億美元
- 平均收入：約 1100 萬美元
- 總支出：約 6 億美元
- 支出峰度分別：約為 2.7，表示有球隊願意花較多的錢投資在球隊上。
- 支出偏態：1.3 稍微往右偏的峰態，表示群內分佈傾向更高的投資金額
- 平均支出：約 900 萬美元

(四)第四群：待開發模範生球隊

表 5

W.L.		Revenues.Men.s.Team		Expenses.Men.s.Team	
平均數	0.68028049	平均數	2874206.46	平均數	2816593.16
標準誤	0.00935053	標準誤	150328.307	標準誤	138118.835
中間值	0.657	中間值	2562265	中間值	2507275.5
眾數	0.727	眾數	#N/A	眾數	#N/A
標準差	0.08467268	標準差	1361280.71	標準差	1250719.25
變異數	0.00716946	變異數	1.8531E+12	變異數	1.5643E+12
峰度	-0.52258	峰度	3.58313424	峰度	0.22368871
偏態	0.60195056	偏態	1.49391091	偏態	0.9179823
範圍	0.344	範圍	7655491	範圍	5017882
最小值	0.545	最小值	1178426	最小值	1178426
最大值	0.889	最大值	8833917	最大值	6196308
總和	55.783	總和	235684930	總和	230960639
個數	82	個數	82	個數	82

1. 待開發模範生球隊定義原因：很明顯的多數落在這群的球隊，勝率都表現很好，但是極多數都不具籃球名校知名度，收入及支出都遠低於補習班球隊，然而待開發模範生球隊也培養了一些傳奇 NBA 球星，例如：籃球之神-Michael Jordan 來自 North Carolina-Greensboro、NBA 最火紅球員之一的 Steven Curry 來自 Davidson 學院、最受矚目新人 NBA 選秀榜眼 Ja Morant 來自 Murry Stat. 及台灣人最熟悉的林書豪來自 Harvard。
2. 指標分析

- 平均勝率：約 0.68。為五群中平均次高著。
- 最大勝率：約 0.889
- 眾數勝率：約 0.7
- 總收入：約 2.4 億美元

- 平均收入：約 290 萬美元，比平庸之輩球隊多七十萬美元，但始終遠低於補習班球隊的 1000 萬美元
- 總支出：約 2.3 億美元
- 平均支出：約 280 萬美元

(五)第五群：不及格球隊

表 6

W.L.		Revenues.Men.s.Team		Expenses.Men.s.Team	
平均數	0.28910256	平均數	2230978.26	平均數	2169055.72
標準誤	0.00831092	標準誤	149494.879	標準誤	133264.864
中間值	0.313	中間值	1939588	中間值	1939588
眾數	0.344	眾數	#N/A	眾數	#N/A
標準差	0.07340009	標準差	1320303.02	標準差	1176963.41
變異數	0.00538757	變異數	1.7432E+12	變異數	1.3852E+12
峰度	-0.2541629	峰度	3.97950098	峰度	1.53112916
偏態	-0.8680672	偏態	1.72752909	偏態	1.26620072
範圍	0.288	範圍	7306988	範圍	5598197
最小值	0.094	最小值	533743	最小值	533743
最大值	0.382	最大值	7840731	最大值	6131940
總和	22.55	總和	174016304	總和	169186346
個數	78	個數	78	個數	78

1. 不及格定義原因：三低球隊，最高勝率者只有 0.38 ，都不具有知名度
2. 指標分析

- 平均勝率：約 0.3。
- 最大勝率：約 0.4
- 平均收入：約 223 萬美元
- 平均支出：約 216 萬美元

第四章 研究結果

本論文使用二階段群集法 (two-stages clustering technique)將球隊收益、支出及勝率中分成五群，其第一群為天之驕子型隊有著極高的收入、支出及名聲，第二群為平庸之輩型球隊三者表現普同群內高度相似，第三群為補習班型球隊，其特色為平均支出高約為 900 多萬美元左右，勝率差距極大，平均收入也高達 1000 萬左右，群內大家花的錢賺的錢差不多，但各球隊比賽成績差距極大。第四群為待開發模範生球隊，平均勝率約 0.68。為五群中平均次高著，NCAA 前十名勝率第四群就佔了五隊，然而多不為籃球名校，總收益跟總支出表現跟平庸之輩球隊差不多 200 萬到 300 萬美元遠遠低於天之驕子型球隊，不過同時也產生了幾位 NBA 明星球員。第五群為無價值球隊，三低，最高勝率也只有 0.38，多數為不知名學校。

NCAA 勝率與收入及支出相關性不大（相關係數大約為 3%），大多數球隊的排名跟學校所投資在球員、設備及教練等沒什麼關係，另外球隊的勝率跟所賺的轉播金、門票收入及周邊商品等也沒有很強的影響。至於支出跟收入有顯著的正相關（大約 91.6%），可以推測學校願意投資在球隊身上越多所賺的就越多，反之亦然。

各個聯盟之間具有顯著的收益差異，多數聯盟球隊收入落在 500 萬美元以下，然而其中六位知名聯盟收益皆高於 1000 萬美元，例如：東南聯盟（Southeastern Conference）太平洋 12 聯盟（Pacific-12 Conference）大十聯盟 (Big Ten Conference)、大東聯盟（Big East Conference）12 聯盟（Big 12 Conference）及大西洋海岸聯盟(Atlantic Coast Conference)。至於聯盟比賽勝率，多數聯盟分布平均，但少數

值得注意的是大十聯盟、大東聯盟、12 聯盟及大西洋海岸聯盟全為天之驕子球隊型及補習班型球隊並且多數球隊勝率都偏高，可顯示知名度高的聯盟資源豐富，再者也要特別注意的是南方聯會（Southern Conference）勝率差距最大一半為低於 0.5 的無價值球隊跟平庸之輩，而有將近四所勝率超越 0.75 的表現且只有大西洋海岸聯盟可以相比，由此可推論即使沒有知名度、沒有資源，還是可以擁有好成績被別人看到的。

第五章 結論與建議

一、結論

透過此次論文二階段群集法可以得知 NCAA 球隊收益分布狀況進而去評價分析出更細節的資訊並將其分群。從商業的角度來看，第一群-天之驕子隊球隊實力、吸金力如此驚人，也願意砸錢投資社備、人才等，光是 15 隊的總收入 3.7 億美元就能超越平庸之輩 104 隊總收入 2.3 億美元，大型運動品牌如：Nike、Adidas 都爭相搶著合作，另外也是 NBA 球探挖取優秀新秀的重要之地例如:轟動全美的 Duke 三王 Zion Williamson、Barrett 和 Reddis 及帶領 Texa-tech 進入總冠軍賽的 Jarrett Culver，天之驕子隊就為 2019NBA 包含選秀前十名中的四個。第二群-平庸之輩球隊，所有數據皆為普通、平庸且幾乎非名校，因此商業價值不高，但適合在地品牌去合作。第三群為補習班型球隊，即使他們勝率差距極大，他們的關注度級學校重視程度可以從收入約 1100 萬美元及支出 900 萬美元看到，以勝率來看十名裡有四所來自補習班型球隊其中還包含冠軍球隊，另外 NBA 選秀中也包含了四位，本群總收入約約 8 億美元及總收入約 6 億美元，為

NCAA 裡的一塊大餅，在這裡除了龍頭大型運動品牌開發外也適合其他品牌發展例如: Under Armour。第四群為待開發模範生球隊，雖然他們資金不豐沛，吸金力不高，但他們最大的商業價值來自於他們傳奇性球員，例如：Steven Curry 帶領運動品牌新星 UA 成功崛起、NIKE 靠Michael Jordan 系列根據富比世報告每年的球鞋贊助收益達到近 1.3 億美元。因此如果運動品牌在 NCAA 待開發模範生球隊中有幸贊助淺力新星，其往後收益指日可待甚至可能會打倒龍頭品牌。第五群為無價值球隊，三低-勝率低、收入低及支出低，沒什麼人關注分配到的轉播金也較少且也不願意花錢整頓，除了在地品牌會贊助外，商業價值不高。

二、建議

此次論文資料搜集中，發現美國教育部很早有立法要求學校體育所有相關數據並有固定格式整合供國家研究及公開大眾使用，相較之下台灣學校體育賽事的資料比較分散且也沒有直接公開球隊收入、支出及教練薪資的資料集，如果之後政府有資源統一格式、協助各級學校建立全國體育資料庫，除了能作為研究使用也可以為本土運動品牌更熟悉台灣體育生態進而改善長年台灣體育經濟的積弱不振，資料的公開透明，可以使更精爭更公平也能帶來新型態的體育經濟產值。

第六章參考文獻

資料文獻

（1） Jain, Murty and Flynn(1999): *Data Clustering: A Review, ACM Comp. Surv.,*

（2） P. Berkhin(2002), *Survey of Clustering Data Mining Techniques, Accrue Software*

（3） Paul E. Black(1998), *Dictionary of Algorithms and Data Structures，NIST*

（4） Deza, Elena(2009). *Encyclopedia of Distances. Springer. 2 194.*

（5） Renato Cordeiro de Amorim, Boris Mirkin(2011). *Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering[J] Pattern Recognition*

（6） MacQueen, J. B. (2009)*Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press: 281297.1967 . MR 0214227. Zbl 0214.46201.*

（7） J. A. Hartigan (1975) *"Clustering Algorithms". Wiley.*

（8） J. A. Hartigan and M. A. Wong (1979) *"A K-Means Clustering Algorithm", Applied Statistics, Vol. 28, No. 1, p100-108.*

（9） D. Arthur，S. Vassilvitskii (2006): *"How Slow is the k-means Method?," Proceedings of the 2006 Symposium on Computational Geometry (SoCG).*