110學年度專題報告競賽

題目：NBA球員表現評估

系所班別：統計系三、四

姓名學號：410778018(葉哲宇)

410878014(周家廉)

410878016(周亨昆)

410878023(呂庭禎)

410878056(簡廷洋)

410878058(許淳鴻)

報告日期：2022/06/17

專題競賽報告分工表

學號	姓名	分工內容
410778018	葉哲宇	主講者、PPT 架構、模型製作、文書
410878014	周家廉	模型製作、文書、PPT 製作、校稿
410878016	周亨昆	模型製作、文書、講稿製作
410878023	呂庭禎	PPT 製作、海報
410878056	簡廷洋	模型製作、文書、PPT 製作、校稿
410878058	許淳鴻	報告影片製作、資料整理、主講者

目錄

1. 引言

1.1 研究背景

籃球是非常普及的運動，北美職業籃球聯賽(National Basketball Association,NBA)在全球非常流行，每年帶來超過 47.5 億美元的收入。目前籃球在全球的運動職業聯賽中排名第 3。

隨著時代前進，大數據不只改變商業與科技世界，它也翻轉了職業運動，更精準剖析比賽、評估球員，讓競爭不再只屬於球場。用統計與數據分析來將球員表現量化，甚至更進一步提供球員場上價值的細微運算，並且協助總裁、教練、球探、球員培訓與醫療團隊做現狀評估與決策。

傳統的籃球數據雖然可以描述球員在場上的進攻得分表現，但防守端卻無法被衡量，而即使有得分、助攻、籃板等數據紀錄表現，事實上籃球作為一個團隊運動，場上各個位置有不同的工作，整體比賽結果也取決於團隊運作的結果，很難用一套固定數據去統一評量各個球員表現，必須考慮戰術執行、防守策略、比賽步調或是球員手感。

隨著 NBA 官方等籃球數據統計網站的完善，分析團隊有足夠完整的資料進行建模與分析，能更細微的分析球員進攻強度，根據進球時防守者位置區分該進球為被防守或未被防守；球員防守能力也能被量化，評估對於每一個防守籃板，負責卡位球員的貢獻程度，即使他不是拿到籃板的球員，根據其所站位置與距離也能量化其表現，對球隊分析團隊而言，這些數據成了勝負關鍵。

利用 NBA 官方網站的統計數據進行分析找出影響各位置球員表現的關鍵變數，尤以現今仍未被發掘之處，藉此提供球隊管理階層在組織球隊或進行球員交易時，以科學的方式評估該選手是否適合，衡量選手加入球隊後能夠為球隊帶來的貢獻，藉此找到真正適合的選手，或許也能發現因為傳統數據而被忽略的球員。

1.2 研究動機

NBA 是北美的男子職業籃球聯盟，也是四大北美職業體育聯賽之一，並被視為全世界水準最高的男子職業籃球賽事。現在這個時期，更是季後賽的階段，許多球隊正如火如荼的競爭者，只為了爭奪至高無上的榮耀-總冠軍。

在這個科學進步的年代，球員間的許多數據都能藉由比賽中的表現來量化呈現。而能在籃球最高的殿堂中競爭且生存的球員，不外乎都具備出色的籃球技巧力和勁爆的體能表現。

然而許多球員雖然擁有非常出色的數據，但是否這些數據能否完整反映在自己的整體表現上，亦或許是否這些數據其實對球隊的影響並不是這麼的重要，一直是個值得談討的問題。比如說一個球員雖然抄截能力很高，那是否這個球員的防守能力就是真的很出色呢?

我們希望透過數據分析來記錄和判斷球員在場上比賽時所累積的攻守技術表現。NBA 會將每場比賽的攻守數據都記錄下來，或許藉由各項數據的觀察及統計，我們能發展出一套模型來判斷、分析各個賽季中的球員是否有足夠的能力可以站穩在 NBA 的賽場上。

1.3 文獻探討

在每場籃球比賽中，都有數據即時紀錄著球員的上場表現。除了有傳統數據之外，近年來進階數據也開始逐漸風行，像是 TS%、USG%及 FP 等等，畢竟在這個數據盛行的年代，傳統數據已經不夠使用了。

GmSc 是 ESPN 作家 John Hollinger 所設計的，它用來評估球員的表現。和 EFF 比起來，GmSc 又多了數據加權計算;PER 也是前者發明的，這個數據也代表了球員上場的表現效率; Dean Oliver 將 Poss 這個數據推廣,此數據代表籃球賽場上的進攻節奏。

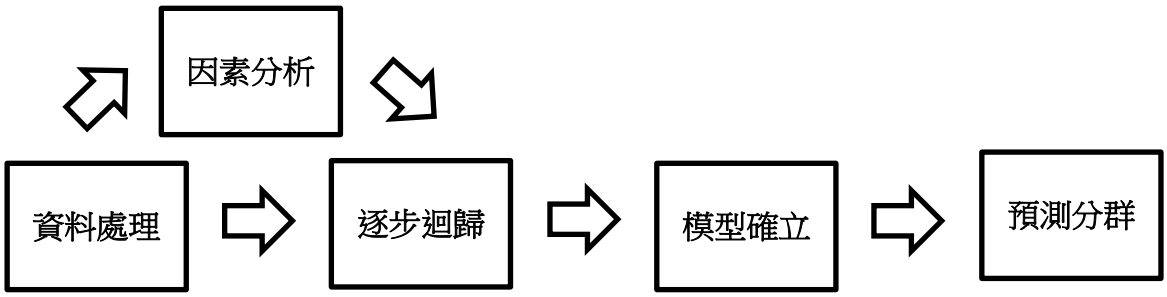
然而，籃球分析相關工作人員對於籃球的見解不盡相同，所以會帶入主觀及經驗去設計進階數據或是去改良既有的進階數據公式，像前面所提及的 GmSc 就是一個很好的例子。由此可見，如何將進階數據有效選擇並正確運用更是一門大學問，我們會在研究中用專業統計技術更適當的進行變數選擇並在實務上呈現。

陳政逸(2013)利用球員能力指標預測 NBA 球賽戰績，首先將球員先分為四種，分別是進攻、防守、團隊及綜合。這四種球員分別有對應的公式，其正確率很高，準確命中該年度最有價值球員及最佳防守球員等獎項，間接代表了這些公式有一定的參考性。我們從中了解了其中公式的變數選用方法及應用，並在研究中我們將組分得更徹底，考慮了球員在場上打的位置不同不能全部都參考同一套標準;黃茂源(2018)NBA 台灣運彩大數據分析與預測，他們以團隊為基礎出發，與我們要研究球員雖然方向主題不一樣，但我們也從中學習觀摩了變數選用及模型建立。我們會在研究中會將資料庫的數量增加，並刪除出場數過少的球員，增加模型的真實及準確率。

以上這些進階數據發明、改良及專業人士對於進階數據的運用及看法，使我們更加奠定用統計方法及工具找到能夠最真實表現球員表現的進階數據。我們在研究中，會嘗試選入更多數據來做為分析依據並擴大資料庫，期盼能找到更好的方法及數據，更能接近球員在場上的真實表現。

2. 研究設計與方法

我們的分析，將依以下流程：



我們採用 NBA 官網中 2017-2022 年所有 NBA 球員的攻守數據。先將所有球員分成三大類，後衛(Guard)、前鋒(Forward)、中鋒(Center)，並分別取其中出賽數前百分之七十五的球員數據使用。NBA 官網中有高達 280 種的變數，由於變數間相關性的限制，故先刪除具有共線性的變數，再挑選出我們在此研究中較有興趣的變數 26 個。

經資料處理過後，挑選 18 個解釋變數；而反應變數則是依照上述所提及的三大類球員分別選取，並在每一項在分為 2 個進攻反應變數、2 個防守反應變數和 2 個綜合反應變數作為判斷標準。

依三大類將 18 個變數則做因素分析將資料降維，用以觀測是否會有更精準的結果，並根據組成變數加以命名與解釋。依三大類進行多元迴歸分析並採用逐步選取法，以觀測解釋變數與反應變數的關係；每類多元迴歸分析中使用 6 個反應變數，解釋變數分為原先挑選的 18 個變數及因素分析結果的因子。接者我們觀察各模型的特色，根據 R 平方挑選較合適的模型。在確立模型後，將資料分為兩種情境：

情境 A：隨機抽取 80% 的資料作為訓練集；另 20% 的資料做為測試集。
情境 B：以年份 2017 到 2020 年的資料做為訓練集；年份 2021 的資料做為測試集。

設定兩情境後，再比較兩情境下預測結果之 MAE 與 RMSE 數值。最後整體資料依迴歸模型得出的關鍵變數，使用 K-means 分群法將其分群並檢視各群之差異與特色。

表 2-1 解釋變數介紹

變數	中文名稱	變數	中文名稱
%FGM	投籃命中占比	%AST	助攻占比
%FGA	投籃出手占比	%TOV	失誤占比
%3PM	三分命中占比	%STL	抄截占比
%3PA	三分出手占比	%BLK	阻攻占比
%FTM	罰球命中占比	%PTS	得分占比
%FTA	罰球出手占比	EFG%	有效命中率
%OREB	進攻籃板占比	TS%	真實命中率
%DREB	防守籃板占比	MIN	上場時間
%REB	籃板占比	DRAFTPOINTS	選秀分數

表 2-2 反應變數介紹

進攻		防守		綜合	
變數	**中文名稱**	**變數**	**中文名稱**	**變數**	**中文名稱**
OFFRTG	進攻效率	CONT2	兩分球出手干擾	FP	范特西分數[1]
FBPS/ PTS OFF TO/ 2ND PTS[2]	快攻得分/透過失誤得分/二波得分	CONT3	三分球出手干擾	USG%	球員使用率

註[1]：范特西分數為球員得分*1+球員籃板*1.2+球員助攻*1.5+球員抄截*3+球員阻攻*3-球員失誤*1。

註[2]：快攻得分為後衛反應變數；透過失誤得分為前鋒反應變數；二波得分為中鋒反應變數。

3. 分析結果

3.1 因素分析

以少數幾個因素來解釋一群相互之間有關係存在的變數，透過因素分析能找出幾群內部相關性高的變數族群。將全部 18 個解釋變數中，依據陡坡圖顯示特徵值大於 1 之因子數，選擇較合適的因子數量，中鋒、後衛、前鋒，三個位置較合適的因子數量分別為 6 個、5 個和 4 個。而各個位置的因子數量對解釋變數可解釋的總變異比例分別是中鋒 82.7%、後衛 78.3%、前鋒 76.0%。

表 3-1 後衛因子命名與組成變數

因子	因子命名	組成變數
FACTOR1	一球在手	%FTM, %FTA, %PTS, %FGM, %FGA, %TOV, %AST
FACTOR2	高控衛型	%REB, %DREB, %OREB, %BLK
FACTOR3	三分射手型	%3PA, %3PM
FACTOR4	命中率	EFG%, TS%
FACTOR5	緊密防守	DRAFTPOINTS, %STL, MIN

表 3-2 前鋒因子命名與組成變數

因子	因子命名	組成變數
FACTOR1	得分主力	%PTS, %FGA, %FGM, %FTM, %FTA, %TOV, MIN, DRAFTPOINTS
FACTOR2	3D 型	%OREB, %REB, %DREB, %BLK, %3PA, %3PM
FACTOR3	命中率	EFG%, TS%
FACTOR4	無私防守	%AST, %STL

表 3-3 中鋒因子命名與組成變數

因子	因子命名	組成變數
FACTOR1	得分數據	%FTM, %FTA, %PTS, %FGM, %FGA, %TOV
FACTOR2	高砲塔型	%3PA, %3PM, %BLK, %OREB
FACTOR3	籃板魔人	%REB, %DREB
FACTOR4	命中率	EFG%, TS%
FACTOR5	無私防守	%STL, %AST
FACTOR6	長時間	DRAFTPOINTS, MIN

3.2 透過迴歸模型找出影響表現重要變數

3.2.1 後衛（原始變數）

透過多元迴歸模型找出影響表現重要變數，並根據我們所挑選篩選過後的後衛資料，進行逐步迴歸分析，比較各反應變數模型之 R 平方，如表 3-4 所示。

表 3-4 後衛迴歸模型的 R 平方

	OFFRTG	FBPS	CONT2	CONT3	FP	USG%
R 平方	0.68	0.59	0.36	0.34	0.84	0.99

以反應變數的角度來看，CONT2 和 CONT3 模型的 R 平方比較低，以 CONT2 和 CONT3 變數的定義，顯著的解釋變數可能集中於有關所有手段的得分變數，利入戰術的成功執行就容易有空檔可得分，然而 CONT2 和 CONT3 是在被對手防守的情境下，使得得分變的困難，導致 CONT2 和 CONT3 模型表現不如其他四個模型優異。逐步迴歸模型結果（ "✔" 為顯著的解釋變數）如表 3-5 所示。

表 3-5 後衛迴歸模型顯著變數表

	OFFRTG	FBPS	CONT2	CONT3	FP	USG%
%FGM				✔		✔
%FGA	✔	✔			✔	✔
%3PM						✔
%3PA					✔	
%FTM					✔	✔
%FTA	✔	✔				✔
%OREB			✔			
%DREB						
%REB		✔			✔	
%AST		✔			✔	✔
%TOV		✔				✔
%STL		✔	✔		✔	
%BLK	✔		✔	✔	✔	✔
%PTS	✔	✔			✔	✔
%EFG		✔	✔		✔	
TS%	✔			✔	✔	
MIN	✔	✔	✔	✔	✔	✔
DRAFTPOINTS	✔		✔	✔	✔	✔

結果可見，各模型間存在顯著變數上的差異，故以下根據各顯著解釋變數在 6 個模型中顯著次數與標準化係數大小，排序出前 8 大影響表現的變數：

數值越大對後衛表現越加分：%FGA, %FTA, %AST, %STL, %BLK, %PTS, MIN, DRAFTPOINTS

從挑選出 8 大重要影響變數與排序中可以發現一些有趣的事。%FGA, %FTA 和 %PTS 數值越大對後衛表現越加分，表示出後衛對於得分顯現出了侵略性，充分展現對於渴望得分的重要性。

變數%AST 顯示出後衛組織球隊的能力，流暢地讓球導傳起來，有助於製造空檔給隊友，同時也代表球隊戰術成功的執行。MIN 和 DRAFTPOINTS 代表該名球員被受教練團的肯定，是該隊主力球員。

3.2.2 後衛（因子）

後衛因素分析部分，我們使用前述 5 個因子做為反應變數。經迴歸分析後，比較各反應變數模型之 R 平方，如表 3-6 所示。

表 3-6 後衛因子迴歸模型之 R 平方

	OFFRTG	FBPS	CONT2	CONT3	FP	USG%
R 平方	0.39	0.55	0.23	0.26	0.81	0.97

由上表可以發現，與原始變數做迴歸分析的模型結果類似，OFFRTG, CONT2 與 CONT3 模型的 R 平方較其他三者模型低；而 FBPS 則介於中間，表現最好的則是 FP 和 USG%。逐步迴歸模型結果 （ "✔" 為顯著的因子）如表 3-7 所示。

表 3-7 後衛因子迴歸模型顯著變數表

	OFFRTG	FBPS	CONT2	CONT3	FP	USG%
FACTOR1	✔	✔	✔	✔	✔	✔
FACTOR2		✔	✔	✔	✔	
FACTOR3	✔	✔	✔	✔	✔	✔
FACTOR4	✔	✔	✔	✔	✔	✔
FACTOR5	✔	✔	✔	✔	✔	✔

藉由迴歸分析，我們可以發現除了反應變數 FACTOR2(高控後)之外，其餘四個因子在所有模型中皆為顯著，表示此四個因子是影響後衛表現的重要變數；而

FACTOR2(高控後)也在四個模型中顯著，故對於後衛表現也有一定的影響力。其影響方式如下：

數值越大對後衛表現越加分：一球在手、高控後型、三分射手型、命中率、緊密防守

由組成變數可知所有因子的組成項皆為正號，且對六個反應變數皆呈現高度相關；故可知所有因子數值越大，對後衛表現皆具有正向影響。

高控後型(FACTOR2)、三分射手型(FACTOR3)在 USG%模型下皆雖為顯著變數，但 USG%(使用率)模型下其迴歸係數卻皆為負數。推測高控後型(FACTOR2)的迴歸係數為負數原因為，其組成變數為籃板與阻攻，表示後衛在場上時，雖然可以有效提高球隊防守，但卻忽略了傳統對後衛要求的得分與組織，進而降低其使用率。而三分射手型(FACTOR3)的迴歸係數為負數原因為，其組成變數為三分球的出手與進球，表示後衛在場上時，雖然可以有效提高球隊進攻，但卻也因此忽視防守，使其使用率降低。

3.2.3 前鋒（原始變數）

透過多元迴歸模型找出影響表現重要變數，並根據我們所挑選場數篩選過後的前鋒資料，進行逐步迴歸分析，比較各反應變數模型之 R 平方，如表 3-8 所示。

表 3-8 前鋒迴歸模型的 R 平方

	OFFRTG	PTS OFF TO	CONT2	CONT3	FP	USG%
R 平方	0.63	0.82	0.63	0.54	0.91	0.99

以反應變數的角度來看，CONT2 及 CONT3 模型的 R 平方較低;表現最好的是 USG%模型，R 平方高達 1.00。CONT2 及 CONT3 的變數定義分別為為兩分球出手干擾和三分球出手干擾，顯著的解釋變數可能集中於有關防守水準類的變數，但是放入模型之 18 個解釋變數並非全然是此類變數，導致 CONT2 及 CONT3 模型表現不如其他四個模型優異。逐步迴歸模型結果("✔" 為顯著的解釋變數) 如表 3-9 所示。

表 3-9 前鋒迴歸模型顯著變數表

	OFFRTG	PTS OFF TO	CONT2	CONT3	FP	USG%
%FGM	✔					
%FGA	✔	✔			✔	✔
%3PM	✔					
%3PA						
%FTM					✔	
%FTA	✔	✔				✔
%OREB	✔	✔	✔			
%DREB			✔		✔	✔
%REB			✔		✔	
%AST			✔	✔	✔	
%TOV				✔	✔	✔
%STL		✔	✔		✔	
%BLK	✔	✔	✔		✔	
%PTS	✔				✔	✔
EFG%		✔			✔	
TS%	✔					
MIN	✔	✔	✔	✔	✔	✔
DRAFT POINTS						

結果可見，各模型間存在顯著變數上的差異，我們排序出前 10 大影響表現的變數，而其對反應變數的影響方式如下:

數值越大對前鋒表現為越加分: %FTA, %OREB, %DREB, %AST, %STL, %BLK, %PTS, MIN

數值越大對前鋒表現為越扣分: %TOV
數值越大對前鋒表現為不一定: %FGA

從挑選出 10 大重要影響變數與排序中可以發現一些現象。首先是 MIN，上場時間越多，也代表了這個球員受到球隊的肯定及重要性是舉足輕重的。%OREB, %DREB 和%BLK 也是加分項，可以知道前鋒的角色任務，也包含內線的威脅及對籃板球的掌握;透過%PTS, %AST 及%STL 也可看出前鋒任務也可包含得分、抄截及助攻，說前鋒的工作是很全面的一點也不為過，像是湖人隊王牌 LeBron James 及籃網隊王牌 Kevin Durant 在球場上的位子都是前鋒，都是歷史級別的全能球員。至於失誤率在扣分項的原因，是因為作為球員本來就是不要丟失球權造成失誤。

3.2.4 前鋒（因子）

前鋒因素分析部分，我們使用前述 4 個因子做為反應變數。經迴歸分析後，比較各反應變數模型之 R 平方，如表 3-10 所示。

表 3-10 前鋒因子迴歸模型之 R 平方

	OFFRTG	PTS OFF TO	CONT2	CONT3	FP	USG%
R 平方	0.27	0.72	0.33	0.26	0.75	0.97

由上表可知因子模型的 R 平方普遍較原始變數模型低，其中 PTS OFF TO, FP 與 USG%模型的 R 平方明顯較其他三個模型高，這個趨勢與原始變數的結果相似，但是 OFFRTG, CONT2 與 CONT3 模型的 R 平方卻異常的低(遠小於 0.5)，代表此模型配適這可能與因子的組成變數與這三個反應變數的關係有關，不適合進行因素分析。如表 3-11 所示。

表 3-11 前鋒因子迴歸模型顯著變數表

	OFFRTG	PTS OFF TO	CONT2	CONT3	FP	USG%
FACTOR1	✔	✔	✔	✔	✔	✔
FACTOR2	✔	✔	✔	✔	✔	✔
FACTOR3	✔	✔	✔	✔	✔	✔
FACTOR4	✔	✔	✔	✔	✔	✔

Factor1(得分主力), Factor2(3D 型), Factor3(高效率), Factor4(無私防守)四個因子在所有模型中皆為顯著，表示此四個因子是影響投手表現的重要變數。其影響方式如下：

數值越大對前鋒表現越加分：得分主力、高效率、無私防守
數值越大對前鋒表現越扣分：3D 型

大部分模型的迴歸係數皆為正，代表大多因子對前鋒皆有正向的影響，得分主力表示其得分能力的優秀、高效率代表固定時間內能取得的表現好、無私防守則是守備及助攻方面的出色。

只有 3D 型的迴歸係數為負，代表前鋒球員表現越 3D(以三分和防守為主)反而會降低他這幾個反應變數的數據。但在實務上，3D 型代表各項能力都有一定的表現，是優秀的前鋒。

3.2.5 中鋒（原始變數）

透過多元迴歸模型找出影響表現重要變數，並根據我們所挑選場數篩選過後的中鋒資料，進行逐步迴歸分析，比較各反應變數模型之 R 平方，如表 3-12 所示。

表 3-12 中鋒迴歸模型之 R 平方

	OFFRTG	2ND PTS	CONT2	CONT3	FP	USG%
R 平方	0.62	0.74	0.63	0.63	0.97	0.99

以反應變數的角度來看，OFFRTG, CONT2 及 CONT3 模型的 R 平方較低;表現最好的是 USG%模型，R 平方高達 0.99。OFFRTG 的變數定義代表了進攻效率，但是中鋒的主要任務並不全然在進攻上，所以這個模型表現較差；CONT2 及 CONT3 的變數定義分別為為兩分球出手干擾和三分球出手干擾，顯著的解釋變數可能集中於有關防守水準類的變數，但是放入模型之 18 個解釋變數並非全然是此類變數，導致 CONT2 及 CONT3 模型表現也不出色。逐步迴歸模型結果("✔" 為顯著的解釋變數)如表 3-13 所示。

表 3-13 中鋒迴歸模型顯著變數表

	OFFRTG	2ND PTS	CONT2	CONT3	FP	USG%
%FGM				✔		
%FGA	✔	✔				✔
%3PM	✔					
%3PA						
%FTM						
%FTA	✔					✔
%OREB	✔	✔	✔			
%DREB		✔				
%REB		✔		✔	✔	
%AST					✔	✔
%TOV				✔		✔
%STL			✔	✔	✔	
%BLK	✔		✔		✔	
%PTS	✔	✔			✔	
EFG%						
TS%	✔					
MIN	✔	✔	✔	✔	✔	✔
DRAFTPOINTS						

結果可見，各模型間存在顯著變數上的差異，我們排序出 7 大影響表現的變數，而其對反應變數的影響方式如下:

數值越大對中鋒表現為越加分: %OREB, %REB, %STL, %BLK, %PTS, MIN
數值越大對中鋒表現為不一定: %FGA

從挑選出 7 大重要影響變數與排序中可以發現一些現象。首先是 MIN，上場時間越多，也代表了這個球員受到球隊的肯定及重要性是舉足輕重的。%OREB, %REB 和 %BLK 也是加分項，可以知道中鋒的角色任務，也包含內線的威脅及對籃板球的掌握;透過%PTS, %AST 及%STL 也可看出中鋒任務也可包含得分、抄截及助攻。至於得分率在不見得是加分的原因，是因為球員的出手比例不一定會對球隊直接造成正面的影響。

3.2.6 中鋒（因子）

中鋒因素分析部分，我們使用前述 6 個因子做為反應變數。經迴歸分析後，比較各反應變數模型之 R 平方，如表 3-10 所示。

表 3-14 中鋒因子迴歸模型之 R 平方

	OFFRTG	2ND PTS	CONT2	CONT3	FP	USG%
R 平方	0.24	0.58	0.38	0.37	0.76	0.97

由上表發現，因子分析做出的迴歸分析與原始變數做迴歸分析的模型結果存在某些結果是一致的。我們可以發現跟中鋒原始變數一樣，OFFRTG, CONT2 及 CONT3 的 R 平方較其他反應變數都比較小，且甚至可能有過小的問題 (OFFRTG, CONT2 及 CONT3 的 R 平方皆為小於 0.5)。這說明了這三個模型的解釋力可能會有不足的疑慮。逐步迴歸模型結果 （ "✔" 為顯著的因子）如表所示。如表 3-15 所示。

表 3-15 中鋒因子迴歸模型顯著變數表

	OFFRTG	2ND PTS	CONT2	CONT3	FP	USG%
FACTOR1	✔	✔	✔	✔	✔	✔
FACTOR2		✔		✔	✔	✔
FACTOR3	✔	✔	✔	✔	✔	✔
FACTOR4	✔	✔	✔	✔	✔	✔
FACTOR5	✔	✔	✔	✔	✔	✔
FACTOR6	✔	✔	✔	✔	✔	✔

藉由迴歸分析，我們可以發現除了反應變數 FACTOR2(高砲塔型)之外，其餘五個因子在所有模型中階顯著，表示此五個因子是影響中鋒表現的重要變數。其影響方式如下：

數值越大對中鋒表現越加分：得分主力、籃板魔人、真實射手、無私防守、長時間
數值越大對中鋒表現不一定越加分: 高砲塔型

其中先說明一下為甚麼高砲塔型(FACTOR2)變數其數值越大不一定代表越加分，其原因為組成高砲塔變數為%3PA(三分出手比率), %3PM(三分命中比率), %BLK(阻攻比率), %OREB(進攻籃板比率)，但我們經過報表發現%BLK 及%OREB 的組成為負號，這其中說明了若高砲塔型數值越大，背後隱含的意義可能是在球隊中火鍋及進攻籃板的減少。籃球運動競爭是一體兩面的，要獲得勝利的必要條件除了進攻以外，防守也是不可或缺的一部分，故雖然高砲塔數值上升說明了這名中鋒球員進攻能力變強(佔球隊中 3 分球出手比率及 3 分球命中數比率增加)，卻犧牲了自己防守的數據(佔球隊中阻攻比率及進攻籃板比率下降)，所以高砲塔數值越大對中鋒表現不一定越加分。

而我們發現大部分的因子對中鋒表現具有正向的影響，且因為我們將反應變數分成進攻、防守及整體，我們可以觀察到不同因子間對不同反應變數的影響有多大。其中幾個比較有趣的點是，無私防守(FACTOR5)在 CONT2 模型及 CONT3 模型下雖皆為顯著變數，但在 CONT2 模型下其迴歸係數為負，而在 CONT3 模型下其迴歸係數為正，表示如果中鋒在球隊中有較高的抄截及助攻比率，其兩分球出手干擾的次數反而越低，但反過來說，如果中鋒在球隊中有較高的抄截及助攻比率，其三分球出手干擾的次數反而越高。這是藉由迴歸分析下產生較有趣的一個結果。

3.3 模型預測

將資料分為兩種情境來進行預測：情境 A，採隨機抽樣將原始資料分為 80%的訓練集與 20% 的測試集，藉由訓練集的資料來建立模型；情境 B，依年度分類，以 2017, 2018, 2019 及 2020 四年的資料做為訓練集建立模型，2021 年資料做為測試集。並利用平均絕對誤差(MAE)與均方根誤差(RMSE)做為衡量模型預測能力的指標。

3.3.1 後衛（原始變數）

表 3-16 後衛預測模型之 MAE 與 RMSE

反應變數	MAE	RMSE
OFFRTG	(2.075, 2.084)[3]	(2.692, 2.759)
FBPS	(0.432, 0.475)	(0.574, 0.620)
CONT2	(0.647, 0.735)	(0.827, 0.923)
CONT3	(0.548, 0.627)	(0.704, 0.788)
FP	(3.264, 3.999)	(4.294, 5.001)
USG%	(0.140, 0.140)	(0.182, 0.178)

註[3]：表格內容為（情境 A 結果, 情境 B 結果）。

MAE 中，除了在 USG%數值在情境 A 中與情境 B 相同，其餘數值在情境 B 中皆大於情境 A。RMSE 中，除了在 USG%數值在情境 A 中大於情境 B 外，其餘數值在情境 B 中皆大於情境 A。其中 USG% 預測最佳，FBPS, CONT2 和 CONT3 三者次之，OFFRTG 與 FP 誤差則是較大。

3.3.2 後衛（因子）

表 3-17 後衛因子預測模型之 MAE 及 RMSE

反應變數	MAE	RMSE
OFFRTG	(3.027, 3.024)[4]	(3.819, 3.823)
FBPS	(0.460, 0.492)	(0.600, 0.639)
CONT2	(0.716, 0.807)	(0.907, 1.009)
CONT3	(0.596, 0.653)	(0.751, 0.820)
FP	(3.657, 4.381)	(4.644, 5.518)
USG%	(0.782, 0.783)	(0.985, 0.985)

註[4]：表格內容為（情境 A 結果, 情境 B 結果）。

MAE 中，除了在 USG%數值在情境 A 中小於情境 B 外，其餘數值在情境 B 中皆大於情境 A。RMSE 中，除了在 USG%數值在情境 A 中與情境 B 相同，其餘數值在情境 B 中皆大於情境 A。其中 FBPS 與 CONT3 預測最佳，CONT2 和 USG% 兩者次之，OFFRTG 與 FP 誤差則是較大。

3.3.3 前鋒（原始變數）

表 3-18 前鋒預測模型之 MAE 與 RMSE

反應變數	MAE	RMSE
OFFRTG	(2.183,2.208)[5]	(2.893,2.894)
PTS OFF TO	(0.432,0.290)	(0.556,0.397)
CONT2	(0.849,0.860)	(1.142,1.168)
CONT3	(0.450,0.457)	(0.608,0.622)
FP	(1.932,2.121)	(3.141,3.382)
USG%	(0.123,0.125)	(0.166,0.167)

註[5]：表格內容為（情境 A 結果, 情境 B 結果）。

MAE 中，情境 A 的 PTS OFF TO 數值比情境 B 大，其他數值則不大於情境 B。RMSE 也有類似的情況。其中 USG%預測最，PTS OFF TO 與 CONT3 次之,，FP 與 OFFRTG 誤差很大。

3.3.4 前鋒（因子）

表 3-19 前鋒因子預測模型之 MAE 及 RMSE

反應變數	MAE	RMSE
OFFRTG	(3.272,3.266)[6]	(4.115,4.113)
PTS OFF TO	(0.363,0.371)	(0.477,0.488)

CONT2	(1.206,1.210)	(1.579,1.574)
CONT3	(0.604,0.607)	(0.774,0.783)
FP	(3.957,4.053)	(5.057,5.193)
USG%	(0.681,0.706)	(0.869,0.898)

註[6]：表格內容為（情境 A 結果，情境 B 結果）。

MAE 中，情境 A 產生的預測誤差除了 OFFRTG 以外普遍皆比情境 B 小。RMSE 也有類似的情況。而整體模型的比較，CONT2 誤差較大, OFFRTG 跟 FP 誤差很大，其他四個變數中，以 PTS OFF TO 模型預測能力最佳，CONT3 與 USG%次之。

3.3.5 中鋒（原始變數）

表 3-20 中鋒預測模型之 MAE 與 RMSE

反應變數	MAE	RMSE
OFFRTG	(2.310,2.251)[7]	(3.030,2.897)
2ND PTS	(0.393,0.418)	(0.547,0.563)
CONT2	(1.067,1.114)	(1.479,1.591)
CONT3	(0.390,0.404)	(0.509,0.545)
FP	(1.390,1.395)	(1.875,1.884)
USG%	(0.137,0.142)	(0.183,0.188)

註[7]：表格內容為（情境 A 結果，情境 B 結果）。

MAE 中，情境 A 的 OFFRTG 數值比情境 B 大，其他數值則比情境 B 小。RMSE 中，情境 A 的 OFFRTG 數值較情境 B 大，其他數值則比情境 B 小。其中 USG%和 CONTESTED 3PT SHOTS 預測最佳，OFFRTG 與 FP 誤差則是較大。

3.3.6 中鋒（因子）

表 3-21 中鋒因子預測模型之 MAE 及 RMSE

反應變數	MAE	RMSE
OFFRTG	(3.378,3.381)[8]	(4.235,4.249)
2ND PTS	(0.524,0.520)	(0.696,0.692)
CONT2	(1.596,1.623)	(2.067,2.090)
CONT3	(0.390,0.549)	(0.509,0.710)
FP	(4.232,4.571)	(5.368,5.690)
USG%	(0.640,0.665)	(0.802,0.842)

註[8]：表格內容為（情境 A 結果, 情境 B 結果）。

MAE 中，情境 A 的 2ND PTS 數值比情境 B 大，其他數值則比情境 B 小。RMSE 中， 情境 A 的 2ND PTS 數值較情境 B 大，其他數值則比情境 B 小。其中 2ND PTS 和 CONT3 預測最佳，OFFRTG 與 FP 誤差則是較大。

3.4 K-means 分群

3.4.1 後衛分群

根據逐步迴歸選取後留下的顯著變數，找出前 10 大顯著影響表現的解釋變數，並根據陡坡圖（以 Elbow Method）得到最適分群數為 6 群，再依 K-means 分群法做分群。如表 3-22 所示。

表 3-22 後衛分群結果（數值為 cluster mean）

	G1	G2	G3	G4	G5	G6
%FGA	25.3	17.9	19.2	29.1	18.7	15.1
%FTA	24.9	11.3	16.5	39.7	13.3	12.5

%AST	24.2	15.1	36.8	43.2	15.2	16.5
%STL	20.2	18.1	24.7	21.9	18.0	25.6
%BLK	11.5	8.8	9.9	12.2	10.4	21.9
%PTS	24.7	17.5	17.8	30.1	16.6	14.1
EFG%	50.9	54.9	48.9	51.6	47.2	50.0
TS%	55.1	57.3	52.4	57.3	50.0	52.9
MIN	25.9	24.9	22.9	31.6	13.6	19.9
DRAFT POINTS	2.7	2.2	2.4	3.0	2.3	2.2
人數	185	252	244	54	184	138

以後衛來說，G4 的表現明顯高於其他五群。G4 內為在後衛這個位置中能力最出色的一群人，在場上皆有一定的主宰力，這五年來的大家熟知的後衛，例如：Stephen Curry 和 James Harden 皆在這個群集內。

G1, G2 和 G3 則在同一區間內。G1 屬於樣樣皆通的萬金油角色，在各項數據皆能做出貢獻，卻無法像 G4 一樣在場上產生主導力。G2 屬於替補出發的砍分手，兩項投籃命中率相關數據皆與 G2 在同一水準內，但其他數據則皆較為弱勢。G3 屬於籃球傳統意義上的後衛，得分不是其主要工作，串聯球隊攻勢才是，由其助攻數據 G2 在同一水準內，便可得知。

G5, G6 為能力較弱的兩群。G6 為在阻攻和抄截方面都為六群中最佳，但在普遍對好後衛認知所需要得分和組織都屬中下水準，可以推測其工作為，當對方得分主力手感火燙時，上場負責封鎖。G5 為能力最弱的一個群集，由上場時間便可得知，推測為球隊第三號後衛。因此，各群命名與解釋如下：

G1：地板型後衛，著一定的下限，夠穩定做出貢獻；然而，卻不能幫助球隊提高上限。所有數據皆在中間水準。
G2：板凳暴徒，上場即負責得分，在有限的上場時間跟出手數內，提供高效率的得分貢獻。有六群中，最高的有效命中率跟真實命中率，然而其他能力都較為弱勢。
G3：傳統組織型後衛，場負責主導球隊攻勢。助攻數據明顯突出。
G4：明星後衛，有著不俗的能力，樣樣都行，能夠決定球隊上限。大部分數據皆明顯在高於其他五群。
G5：替補深處，吃掉少數的上場時間。上場時間為六群中最少，其他數據也皆為後段班水準。

G6 : 防守型後衛，上場負責防守，得分和組織皆不在工作範圍內。抄截和阻攻數據皆為六群中最高。

綜合以上結果，後衛能力由強至弱排序為：G4 >>> G1 = G2 = G3 > G6 > G5

3.4.2 前鋒分群

依據迴歸模型找出影響表現重要變數以 K-means 分群法分群，並根據陡坡圖（以 Elbow Method）得到最適分群為 6 群。如表 3-23 所示。

表 3-23 前鋒分群結果（數值為 cluster mean）

	G1	G2	G3	G4	G5	G6
%FGA	16.0	21.1	15.0	28.5	18.1	17.5
%FTA	21.0	26.4	15.6	38.8	16.2	13.9
%OREB	44.2	35.1	36.8	16.6	16.6	15.0
%DREB	27.0	27.0	24.6	25.0	21.2	18.2
%AST	11.2	14.2	12.3	31.0	24.7	12.2
%TOV	16.5	20.2	14.6	30.6	20.9	13.5
%STL	17.7	16.9	20.5	21.0	23.1	18.3
%BLK	54.3	30.5	29.7	17.6	20.4	14.6
%PTS	17.9	22.4	15.3	30.1	17.3	16.8
MIN	17.0	23.5	16.3	33.5	26.0	21.5
人數	107	169	107	63	98	542

從前鋒球員分群中可以觀察到一些現象。首先，MIN, %PTS, %FGA, %FTA 呈現高度同向趨勢,可以推測出說上場時間越多，在場上的得分與各項出手數也會與之一同增加;再者，MIN, %TOV, %AST 也有正向關係，間接說明了球權越多，會造成的失誤數與助攻數也會跟著提升;其他數據方面，%OREB, %DREB, %STL 呈現反向趨勢，也意味著出色的搶籃板與抄截能力很難一個人同時包辦，除非是明星球員。此外，籃板與阻攻能力也有正向關係。因此，各群命名與解釋如下：

G1 : 地板型前鋒，此群的球員雖然不在所屬球隊的主要輪替名單下，但是有不錯的抓籃板及阻攻的能力能幫助到球隊。
G2 : 禁區型前鋒，籃板及阻攻能力在所有群中特別突出，能幫球隊增加禁區的威脅。

G3 : 板凳深處前鋒，此群的球員不在球隊的主要輪替名單中，上場時間為所有群中最少。
G4 : 明星前鋒，樣樣都行，在場上有很大的影響力，並能左右比賽的勝負。大部分數據皆明顯在高於其他五群。
G5 : 主力前鋒，有一定上場時間，此群的球員也有著不錯的抄截及投籃能力。
G6 : 輪替型前鋒，此群的球員各項能力中等，但是在所屬球隊的輪替名單下，並有著一定的上場時間。

綜合以上結果，前鋒能力由強至弱排序為：G4 >> G5 > G2 > G6 == G1 > G3

3.4.3 中鋒分群

依據迴歸模型找出影響表現重要變數以 K-means 分群法分群，並根據陡坡圖（以 Elbow Method）得到最適分群為 5 群。如表 3-24 所示。如表 3-24 所示。

表 3-24 中鋒分群結果（數值為 cluster mean）

	G1	G2	G3	G4	G5
%FGA	19.9	22.6	20.3	14.7	16.7
%OREB	19.9	44.1	27.0	47.4	47.5
%REB	21.2	36.1	28.0	32.1	34.0
%STL	15.4	18.6	17.6	17.5	15.8
%BLK	42.7	49.4	20.4	57.4	34.7
%PTS	19.7	24.6	21.2	17.2	18.7
MIN	20.4	29.6	24.1	16.1	19.3
人數	34	74	57	88	151

以中鋒來說，G2 的能力表現無庸置疑，各項數據都非常出眾，且上場時間最多說明在球隊最受到教練團的信任。

而 G3, G4, G5 的表現也很微妙，雖然 G3 沒有任何一項數據是最頂尖的，但因為得分、出手數及上場時間都是第二突出的，也說明在球隊定位下也可以拿到不少的球權;而 G4 和 G5 雖然在大部份數據都不出色，但在某單項數據則特別突；G4 為阻攻，G5 為籃板(包括總籃板及進攻籃板)，說明這 2 個群集應為同等重要，故最後決定將 G3, G4 及 G5 的能力排名為中等。

最後發現 G1 的各項能力都不出眾，而籃板能力(包括總籃板及進攻籃板)最低更說明了雖然在球隊是中鋒，但搶到籃板的能力卻可能比同隊的後衛、前鋒還少，最後決定將 G1 能力評為最低。因此，各群命名與解釋如下：

G1：中下水準中鋒，有三個數據是各群中最低 (%OREB, REB, STL)，且沒有數據是最高的。
G2：明星中鋒，各項數據都不錯，其中有 5 項數據是最高的 (%FGA, %REB, %STL, %PTS, MIN)，且沒有任何數據是最低的。
G3：平均水準中鋒，雖然沒有任何一項數據是最高的，但有 4 項數據是第二高的(%FGA, %STL, %PTS, MIN)，但阻攻能力是最慘烈的(跟最高的火鍋能力有明顯差異)
G4：阻攻中鋒，進攻能力非常差 (%FGA, %PTS, MIN 是最低的)，但火鍋能力最出眾。
G5：籃板中鋒，雖沒有數據是最低的，但大部分數據都是第二低的 (%FGA, %STL, %BLK, %PTS, MIN)，但籃板能力相當出眾。

結論來說:能力評比為 G2>>>G3=G4=G5>G1

4. 結論

在本次研究中，首先我們從 NBA 官網收集五年的數據，其中包含了 18 個解釋變數以及 8 個反應變數，並將球員分成後衛、前鋒及中鋒三個位置。接著我們篩選掉出場數過少的球員，原因是出場數過少的球員表現穩定性不夠，間接影響到數據上，會造成模型的錯估及失準。

接著我們做因素分析以及逐步迴歸。前者將資料降維並做適當的變數分群，用以觀測是否會有更精準的結果。後者則為多元迴歸分析並採用逐步選取法，以觀測解釋變數與反應變數的關係。我們發現後衛的顯著變數有%FGA, %FTA , %PTS, %AST 及 DRAFTPOINTS，顯著因素有 FACTOR1, FACTOR4 及 FACTOR5;前鋒方面，我們發掘它的顯著變數有 MIN, %PTS, %AST 及%STL，顯著因素有 FACTOR1, FACTOR3 及 FACTOR4；至於中鋒，顯著變數的顯著變數有 MIN, %OREB, %REB, %BLK，顯著因素有 FACTOR3, FACTOR4 及 FACTOR6。簡而言之，原始變數模型和因子模型各有其特殊之處。原始變數模型著重在單一變數的重要性，反觀因子模型，呈現了球員的哪些特色會影響到他們自己的表現。因素分析的結果，有許多因子是透過主觀判斷不易想到的變數組合，我們以科學的方式與客觀的詮釋方法，由一個全新的角度去呈現球員的特質。

再來我們觀察各迴歸模型的特色及差別並確立模型，將資料分為兩種情境去計算 MAE 與 RMSE 的值並藉此衡量模型的好壞。在這邊裡，我們主要目的是比較兩個

情境間及模型間的差異性，而不是絕對的數值大小。從預測結果來看，後衛、前鋒及中鋒三個位置皆以 USG%模型預測表現最佳;整體而言，原始變數模型之預測能力又比因子模型稍好一些。

最後我們依所選取的關鍵變數，採用 K-means 分群法，將其分群並檢視各群之差異與特色。從結果來看，前鋒的 G4、中鋒的 G2 以及後衛的 G4 都是明星級別球員;而 前鋒的 G3、中鋒的 G1 以及後衛的 G5 皆是屬於目前在 NBA 的舞台並不亮眼。而其他群夾在兩者之間並各帶有其特色，例如前鋒的 G1 球員，雖然不是明星級別球員， 但是阻攻能力是一流的。

由此可見我們最後的分群效果出色，群集間各具特色且代表力強。我們這項研究，也提供後人在評估籃球球員表現時，能有更不一樣的觀點與方式去進行探討。

4.1 檢討與反思

研究還有些許進步空間。首先我們要多選一些負面變數。我們的正面變數比負面變數多出不少，會導致防守類模型與進攻類模型相較起來，難有更完整出色的解釋及預測能力。以 CONT2 及 CONT3 模型為例，若我們納入更多的負面解釋變數，像是助攻失誤率、對手投籃命中率及對手透過失誤得分等等，或許能提高防守類模型的解釋及預測能力。

另一方面，我們不應該選擇 DRAFTPOINTS 當作解釋變數。DRAFTPOINTS 是我們以選秀順位當作其中一個評分的變數，但經由我們的分析及預測，發現這個變數使用的效果並不出色。

除此之外，我們應該選擇分析出場數前 25%的球員，會比我們原本設定篩除掉出場數後 25%的球員，理論上結果會更好。可能的原因是出場數在 25%到 75%之間的球員，與出場數在 25%以前的球員相較起來，球季上場次數比較少，導致年度數據容易受到幾次波動較大的數據影響，而無法呈現該球員真實的能力。

最後，有些時候籃球的數據會無法還原出比賽的真實情況。以防守為例，一個球員的抄截和阻攻次數多，並不代表他的防守能力出色，因為好的防守，不單單只有阻攻及抄截，球員的防守判斷和站位也是重要的因素，但這是數據無法展現的。換個角度來看，如果一個球員有好的防守，造成對手沒有出手，但不會有數據去呈現這樣的情形。

籃球是團隊運動，場上的每個人都會影響比賽的走向，數據是難以展現出完整的場上情況的。數據必然有值得參考之處，它能幫助我們快速的了解球員了解比賽，但未必全然是事實，需要更多資訊才能將比賽完全體現。

參考文獻

1. **陳政逸(2013)** 利用球員能力指標預測 NBA 球賽戰績
   https://ndltd.ncl.edu.tw/cgibin/gs32/gsweb.cgi?o=dnclcdr&s=id=%22101FCU05336013%22.&searchmode=basic

2. **黃茂源(2018)** NBA 台灣運彩大數據分析與預測
   https://nccur.lib.nccu.edu.tw/bitstream/140.119/117656/1/100701.pdfsic

3. **Basketball-Reference**
   https://www.basketball-reference.com/

4. **Dean Oliver(2004)** Basketball on Paper: Rules and Tools for Performance Analysis
   https://books.google.com.tw/books?id=ygrQDwAAQBAJ&printsec=frontcover&dq=Dean+Oliver&hl=en&sa=X&ved=2ahUKEwjJnvTif33AhUlEqYKHY10AuMQ6AF6BAgJEAI#v=onepage&q=Dean%20Oliver&f=false

5. **Basketball Calculators**
   https://captaincalculator.com/sports/basketball/

6. **蘇昱禎(2021)** 預測 NBA 球隊能否進季後賽之結果
   http://ir.nptu.edu.tw/bitstream/987654321/21012/1/109NPTU0507001-001.pdf

7. **ADAM FROMAL (2012)** Understanding the NBA: Explaining Advanced Comprehensive Stats and Metrics
   https://bleacherreport.com/articles/1040320-understanding-the-nba-explaining-advanced-comprehensive-stats-and-metrics