Unsupervised Learning for Super Basketball League Players

Evan Lin, Alexis Wang, Weichen Chien

*Super Basketball League is the professional basketball league in Taiwan

Traditional Categorization for Basketball Players

Guard (G)

- Point Guard
- Shooting Guard

Forward (F)

- Small Forward
- Power Forward

Center (C)

- Center

*This kind of categorization usually bases on a player's height.

Is it suitable to categorize all players mostly with their height?

Not so good for modern basketball!

*For example, we now see some centers who are supposed to stay under the basket are now shooting threes.

Data Collection

https://sbl.choxue.com/stats

Used Python Selenium for Data Collection

- First, we crawled all players' personal webpage links from the player list of all seasons, and save them in a .txt file.
- Then, we write another crawler to collect each players' game performance by the links in the .txt file.

- Since most players play multiple seasons, we removed duplicate players after the first step.

Example for Original Data on Webpage

陳昱瑞

球 隊 | 臺灣銀行
背 號 | #2
位 置 | G

生 日 | 1995-12-22
年 齡 | 23 歲
身 高 | 185 公分
體 重 | 75 公斤

平均	出場次數	時間(分)	二分	%	三分	%	罰球	%	投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
	----	25	1.5-3.5	42.86%	1.6-5.1	32.07%	1.3-1.6	79.31%	3.1-8.6	36.45%	9.2	3	1.7	1.1	1.3	0.1	2.4
總數據	**出場次數**	**時間(分)**	**二分**	**%**	**三分**	**%**	**罰球**	**%**	**投籃**	**%**	**得分**	**籃板**	**助攻**	**失誤**	**抄截**	**阻攻**	**犯規**
	36	897	54-126	42.86%	59-184	32.07%	46-58	79.31%	113-310	36.45%	331	107	61	41	45	3	87

Data Preprocessing

Data Cleansing

1. Created column names
2. Removed duplicated players (who are traded and appears twice or even thrice)
3. Removed useless columns (e.g. birthday, age)
4. Removed players with incomplete data
5. Parsed and separated columns (e.g. 8-10 for two pointers, we separated them into “two pointer attempt: 10” and “two pointer made: 8”)

Data Preprocessing

Added Columns (dummy variables)

1. Nationality (binary):
    - 1 for Taiwanese Players
    - 0 for Foreign Players
2. National Team Member (binary):
    - 1 for National Team Players
    - 0 for Non-National Team Players
3. Advanced Basketball Statistics:
    - EFF、eFG%、TS%、TOV%

Data Preprocessing

Advanced Statistics: EFF

EFF(Efficiency):

EFF算是最基本、也最常見的進階數據了。為什麼呢？我們先一起來看看EFF的公式長什麼樣子：

EFF = (PTS + TREB + AST + STL + BLK) – (FGA – FGM) – (FTA – FTM) – TO

公式看起來一長串，其實就只是把五大數據：得分、籃板、助攻、抄截、阻攻相加起來，再扣掉投籃不進與罰球不進，最後扣掉失誤而已。

Data Preprocessing

Advanced Statistics: eFG%

eFG%(effective Field Goal percentage，有效命中率)

eFG%的概念很簡單，因為一顆三分球能比一顆兩分球多得1.5倍的分數，所以將FG%中三分命中數乘上1.5，也就是提高1.5倍的權重。

$$eFG\% = \frac{(FGM + 0.5 \times 3PM)}{FGA}$$

這個數據有個很直白的應用，將eFG%乘以出手數再乘2(eFG% × FGA × 2 = 得分)，就是球員藉由投籃出手的得分，因為一次命中就等於得兩分，而命中三分的那部分，在eFG%裡已經先乘好1.5了。

Data Preprocessing

Advanced Statistics: TS%

TS%(True Shooting percentage，真實命中率)

TS%比eFG%多考慮了罰球，用來衡量球員每次進攻機會能夠得到多少分。因為有些進攻機會，本來能夠以投籃出手，但因為被犯規而變成罰球出手(不論兩次或三次)，但這仍然是一次進攻機會，在eFG%裡卻未計入。真實命中率就是希望能考慮到所有的進攻機會，所以比eFG%多考量到罰球的部分，使命中率看起來更"真實"。我們可以從eFG%的公式推導出它與TS%的差異：

$$TS\% = \frac{得分}{2\times(FGA + 0.44\times FTA)}$$

Data Preprocessing

Advanced Statistics: TOV%

TOV%(Turnover Percentage，失誤率)

在談到助攻時，往往會一起談到失誤。助攻透過傳球達成，而傳球是發生失誤的主要原因之一。既然有助攻率，那麼也應該有失誤率。沒錯，還真的有。

$$TOV\% = 100 \times \left[\frac{TOV}{FGA + 0.44 \times FTA + TOV}\right]$$

經過前面那麼多複雜繁瑣的公式洗禮，這個失誤率的公式應該難不倒你了吧？分母就是球員所得到的所有進攻機會，其中有多少進攻機會最後會變成失誤，就是失誤率。

Problem for Using Advanced Basketball Statistics

Warning: Variables are collinear

```
C:\Users\chen0\Anaconda3\lib\site-packages\sklearn\discriminant_analysis.py:388: UserWarning: Variables are collinear.
  warnings.warn("Variables are collinear.")
```

* This will cause the problem of collinear after applying it because these are calculated through the players' game performance. Therefore, we removed the advanced basketball statistics.

Training data

	Nationality	CTteam	Height	Weight	ptsPG	rebPG	astPG	toPG	stlPG	blkPG	pfPG	2ptsPercentage	3ptsPercentage
0	0	0	204	111	19.7	13.1	2.9	2.6	1.5	2.0	2.5	0.4885	0.3529
1	0	0	211	138	16.3	13.9	1.2	2.2	0.7	1.7	3.1	0.7089	0.0000
2	0	0	195	82	17.7	9.6	1.0	2.1	1.4	1.4	1.7	0.5158	0.2447
3	0	0	206	105	22.7	13.0	3.1	4.1	2.4	1.5	3.2	0.4703	0.1594
4	0	0	199	100	15.8	7.1	1.9	1.2	1.7	1.3	3.2	0.5195	0.4000

	ftPercentage	fgPercentage	2ptsM_PG	2ptsA_PG	3ptsM_PG	3ptsA_PG	ftM_PG	ftA_PG	fgM_PG	fgA_PG
0	0.6977	0.4483	5.8	11.9	1.8	5.0	2.7	3.9	7.6	16.9
1	0.4532	0.7089	7.1	10.1	0.0	0.0	2.0	4.3	7.1	10.1
2	0.7546	0.4835	6.4	12.4	0.4	1.7	3.7	4.9	6.8	14.1
3	0.6292	0.4572	8.7	18.6	0.1	0.8	4.8	7.6	8.9	19.4
4	0.7647	0.4860	4.4	8.6	1.3	3.3	2.9	3.8	5.8	11.9

Clustering Steps

- Dimensionality Reduction:
  1. Principal Components Analysis
  2. t-SNE
  3. Linear Discriminant Analysis

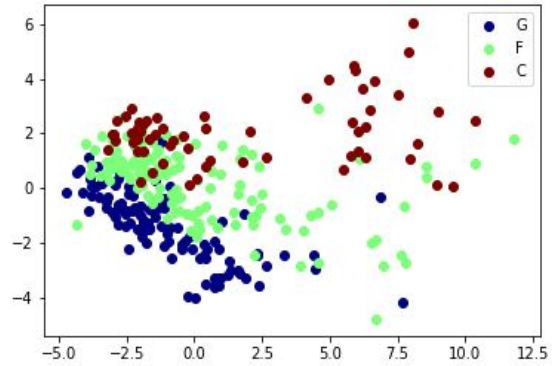
- Unsupervised Learning Algorithms for Clustering:
  1. K-means
  2. Hierarchical Agglomerative Clustering

LDA - different from PCA, LDA is a supervised learning dimensional reduction. Therefore, we plug in the players' position as dummy variables into the data.

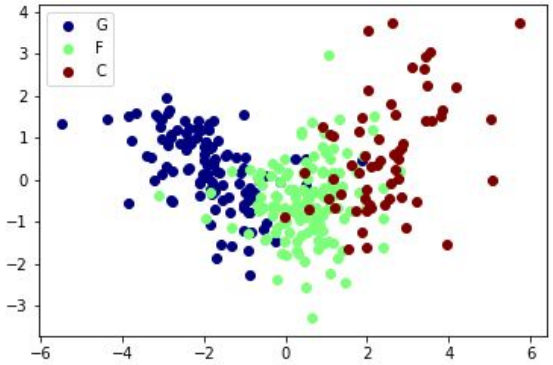
Dimensionality Reduction :

- PCA
- t-SNE
- LDA

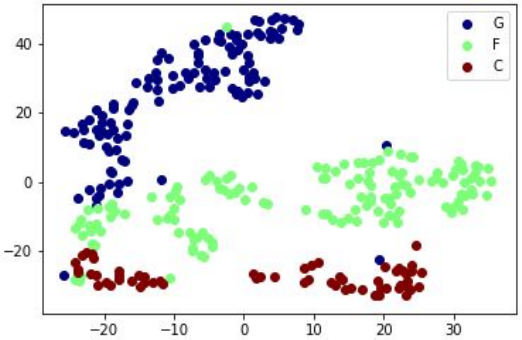
PCA 降維



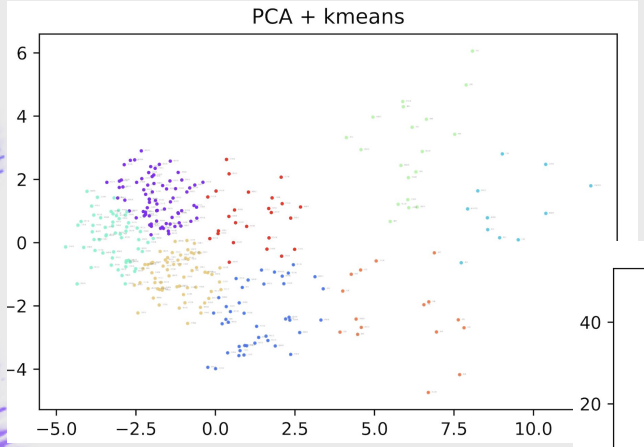
LDA 降維

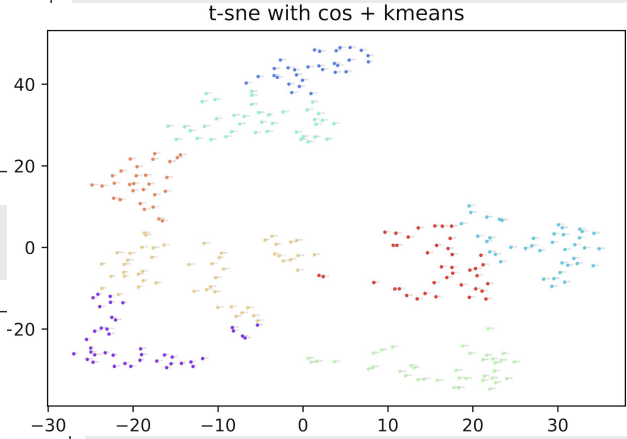


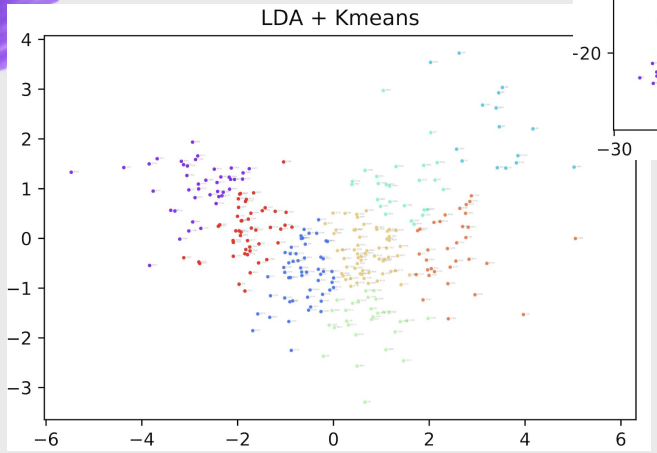
t-SNE 降維



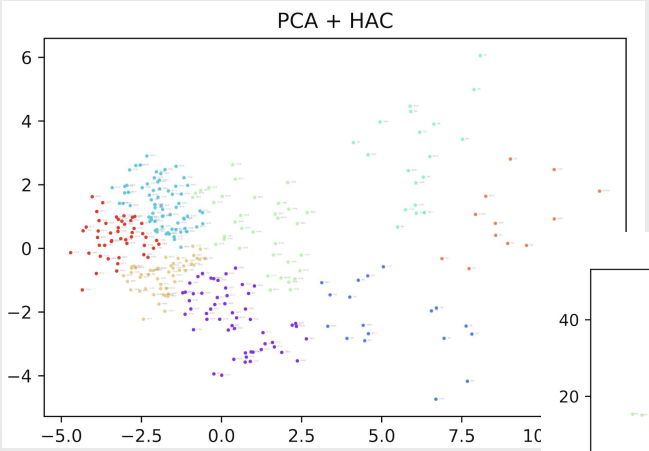
Clustering : K-means

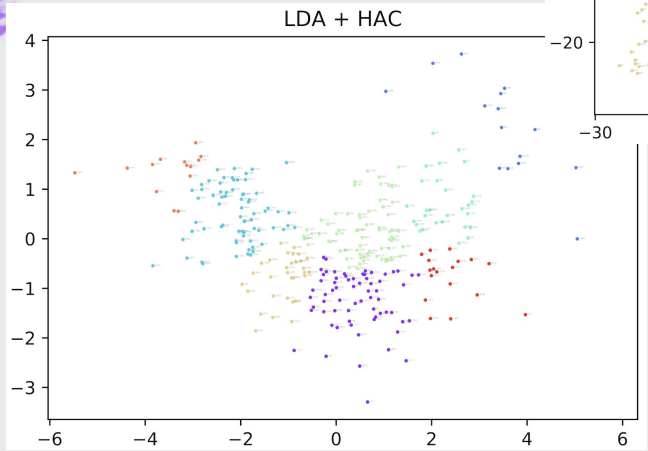


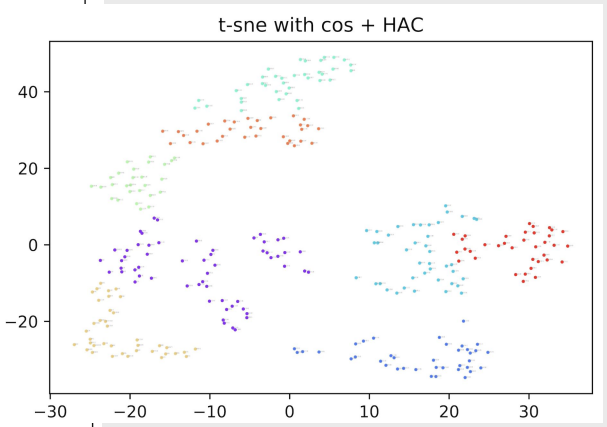




Clustering: Hierarchical Agglomerative Clustering







This Clustering becomes a classification of Starting Players and Bench Players

We found another problem that this kind of clustering becomes a classification between the “starting players who have more time on the court” and
“bench players who do not have much time staying on the court”.
(which is meaningless)

Therefore, we weighted the players’ game performance by time. Making them have the same time playing on the court to 25 mins.

Modified Training Data

	Nationality	CTteam	Height	Weight	pts_25min	reb_25min	ast_25min	to_25min	stl_25min	blk_25min	pf_25min	2ptsPercentage
0	0	0	204	111	15.390625	10.234375	2.265625	2.031250	1.171875	1.562500	1.953125	0.4885
1	0	0	211	138	12.734375	10.859375	0.937500	1.718750	0.546875	1.328125	2.421875	0.7089
2	0	0	195	82	15.258621	8.275862	0.862069	1.810345	1.206897	1.206897	1.465517	0.5158
3	0	0	206	105	16.214286	9.285714	2.214286	2.928571	1.714286	1.071429	2.285714	0.4703
4	0	0	199	100	14.629630	6.574074	1.759259	1.111111	1.574074	1.203704	2.962963	0.5195

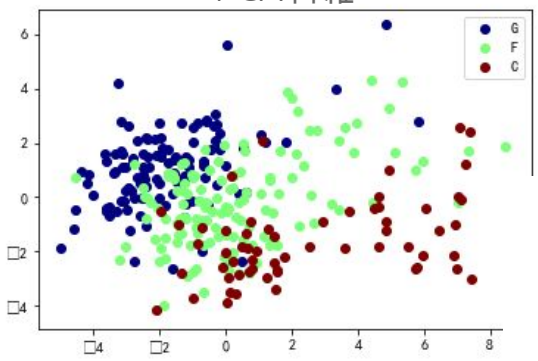
	3ptsPercentage	ftPercentage	fgPercentage	2ptsM_25min	2ptsA_25min	3ptsM_25min	3ptsA_25min	ftM_25min	ftA_25min	fgM_25min	fgA_25min
0	0.3529	0.6977	0.4483	4.531250	9.296875	1.406250	3.906250	2.109375	3.046875	5.937500	13.203125
1	0.0000	0.4532	0.7089	5.546875	7.890625	0.000000	0.000000	1.562500	3.359375	5.546875	7.890625
2	0.2447	0.7546	0.4835	5.517241	10.689655	0.344828	1.465517	3.189655	4.224138	5.862069	12.155172
3	0.1594	0.6292	0.4572	6.214286	13.285714	0.071429	0.571429	3.428571	5.428571	6.357143	13.857143
4	0.4000	0.7647	0.4860	4.074074	7.962963	1.203704	3.055556	2.685185	3.518519	5.370370	11.018519

Dimensionality Reduction:

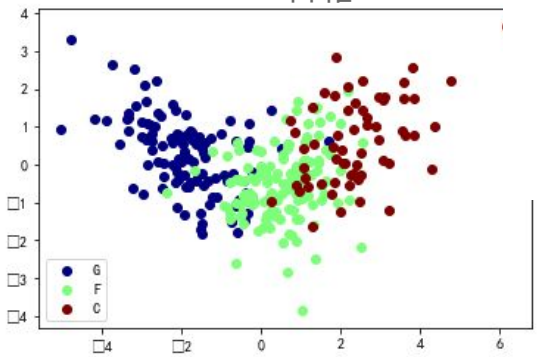
- PCA
- t-SNE
- LDA

(25min)

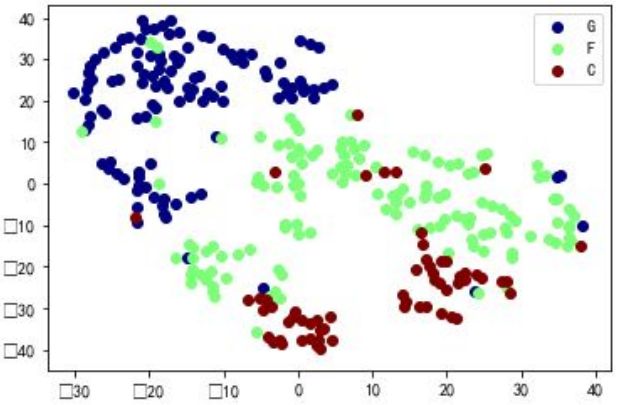
PCA 降維



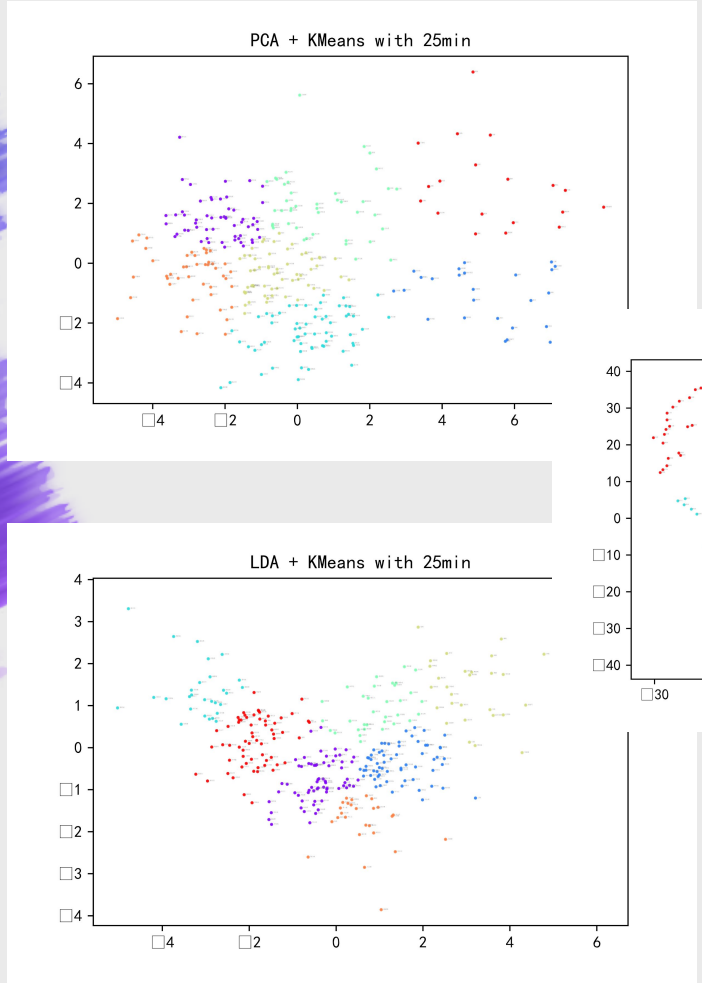
LDA 降維

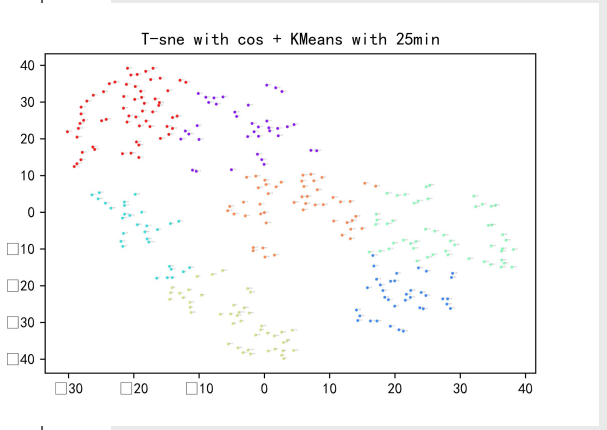


t-SNE 降維

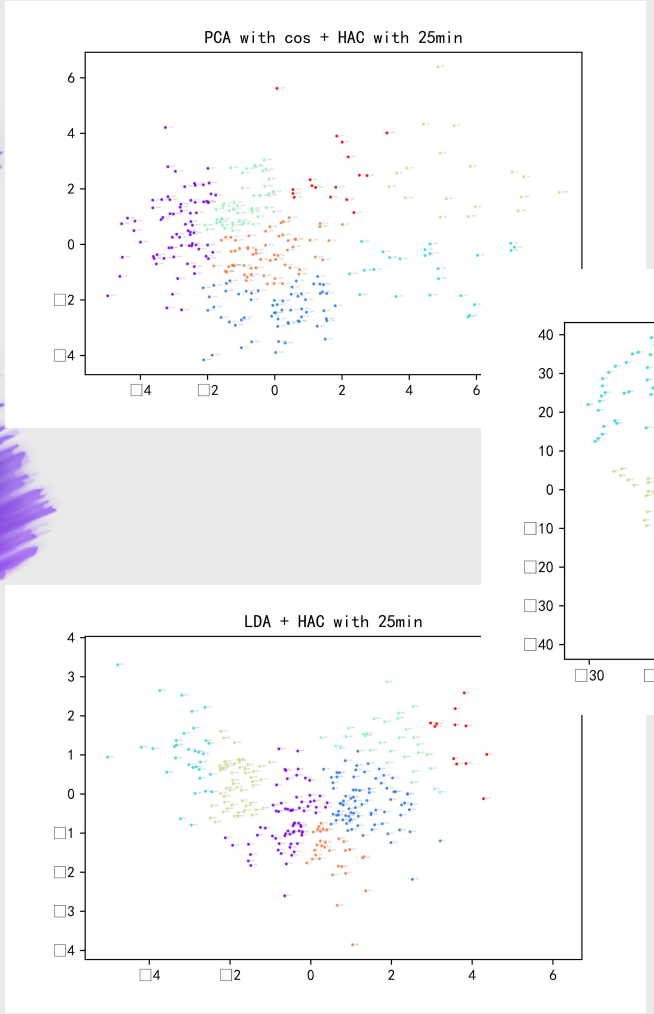


Clustering : K-means (25min)





Clustering: Hierarchical Agglomerative Clustering (25min)

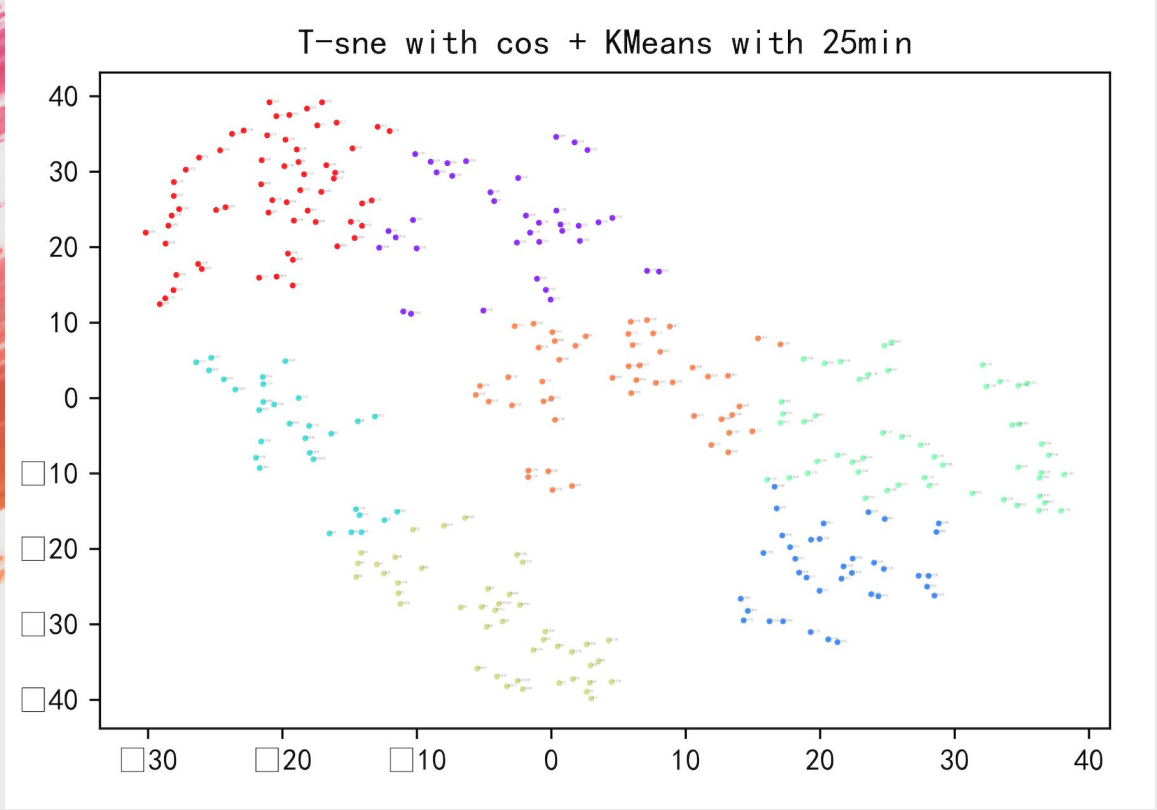


Comparison with Silhouette Score

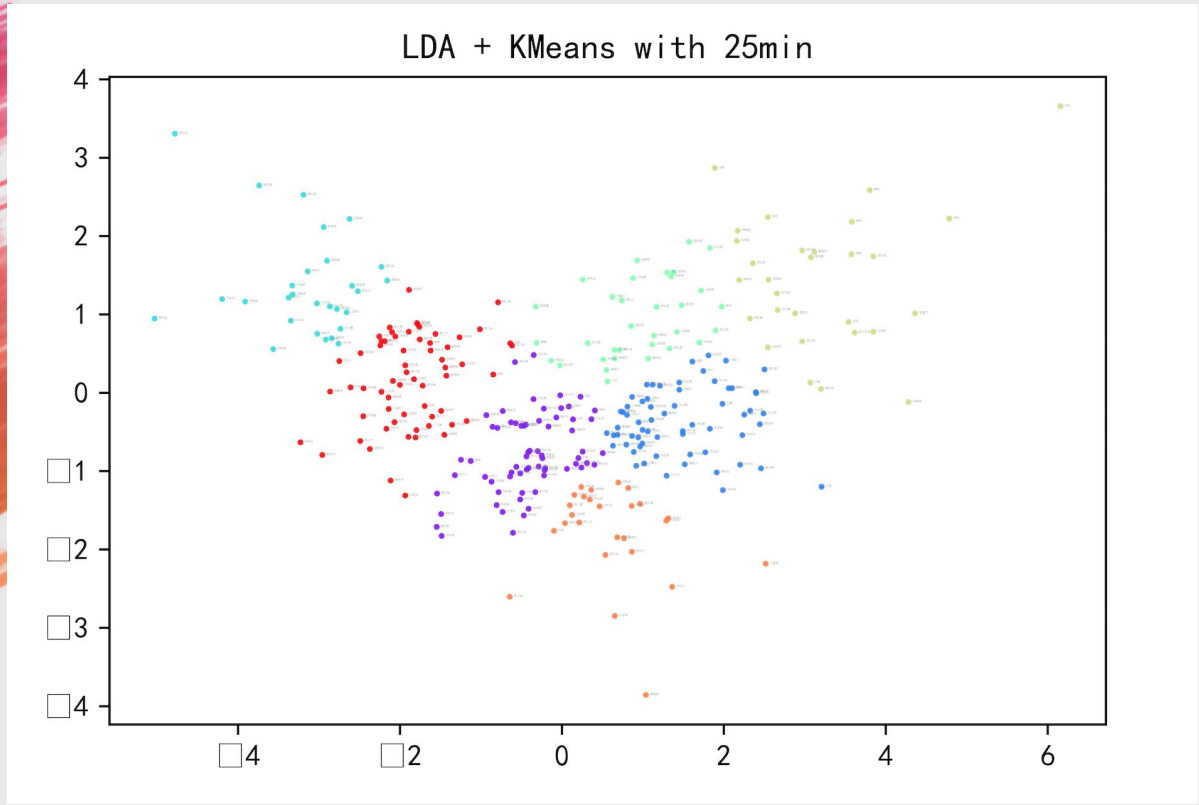
	Kmeans	HAC
PCA	0.356961	0.299498
t-SNE	**0.427757**	0.412008
LDA	0.329420	0.295329

*Finally, we use Silhouette Score, which is a way to represent how well each object has been classified, to choose the best clustering way that we are going to interpret later.

t-SNE + KMeans (25min)



LDA + KMeans (25min)

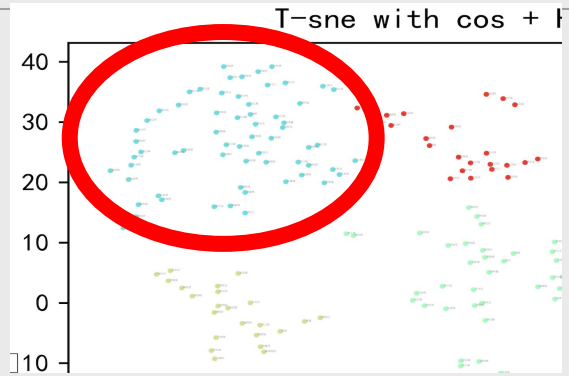


Our Categories

Index	Category
0	Guards who are responsible for organizing offense (similar to point guard)
1	Fixed Spot 3 Point Shooters (Rookies)
2	Swingman（available for both guard and forward）
3	3D Players (3 points +Defense)
4	Players without outstanding statistics
5	Defensive Centers (hard work at painted area）
6	Foreign Players & Best Taiwanese Players (playing at a next level than the other players)

Guards (organizing offense)

Average Height	179.9 cm
Featured Players	陳世念、林韋翰、洪志善
Features	• Lower points per game • Higher assists per game • More turnovers (due to more time handling the ball)

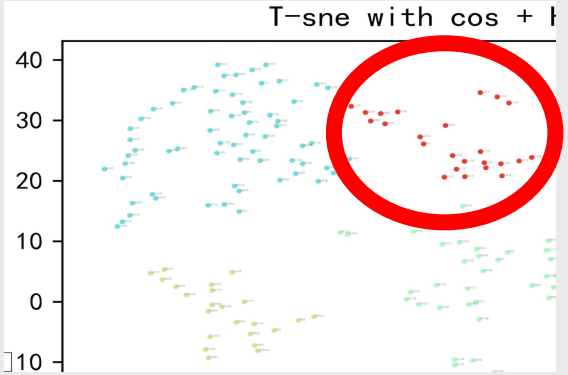


時間(分)	二分	%	三分	%	罰球	%
33	2.1-4.9	42.26%	1.3-3.8	34.37%	1.4-1.9	75.38%

投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
3.4-8.7	38.85%	9.5	3.8	8.8	2.9	1.8	0.1	2.3

Fixed Spot 3 Point Shooters (Rookies)

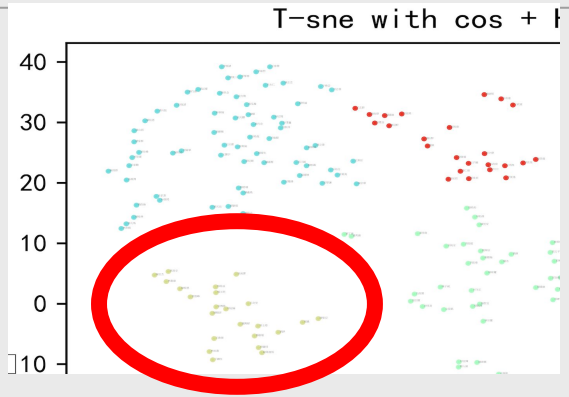
Average Height	182.7 cm
Featured Players	簡佑哲、王子剛、陳昱瑞
Features	• More three point attempts • Less time playing on the court • Less scores • Rookies temp to stay at a spot and wait for any chance for threes.



時間(分)	二分	%	三分	%	罰球	%	投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
25	1.5-3.5	42.86%	1.6-5.1	32.07%	1.3-1.6	79.31%	3.1-8.6	36.45%	9.2	3	1.7	1.1	1.3	0.1	2.4

Swingman（between guard and forward）

Average Height	184.8 cm
Featured Players	周儀翔、劉錚、林書緯 (Jeremy Lin's Brother)
Features	• More attempts on baskets • More score • Mainly play an important role in scoring.



時間(分)	二分	%	三分	%	罰球	%	投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
34	3.7-10	36.67%	1.7-4.3	38.46%	3.9-5.7	68.63%	5.3-14.3	37.21%	16.2	5.6	2.7	2	1.2	0.2	1.4

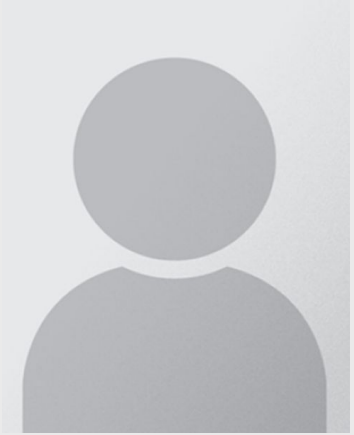
3D Players (3 points + Defense)

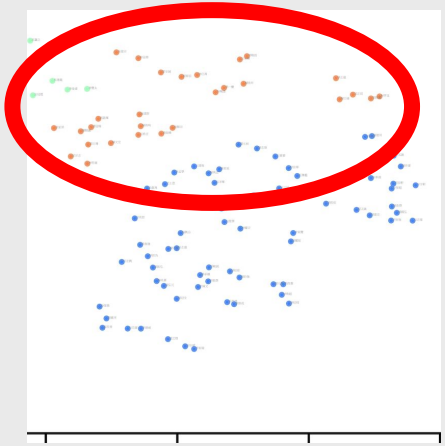
Average Height	191.3 cm
Featured Players	簡浩、呂政儒、陳子威、何守正
Features	• more time on the court • play lots of defense • Shoot much more threes than all the other groups.

時間(分)	二分	%	三分	%	罰球	%	投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
26	2.2-4.9	44.89%	2-5.6	35.89%	1.7-2.3	72.49%	4.2-10.5	40.1%	12.1	3.2	1.2	1.8	0.9	0.2	2.3

Players without outstanding statistics (Forward)

Average Height	192 cm
Featured Players	Unfortunately, we do not recognize any of them….
Features	• Less time on the court... • Minimum points per game… • More time guarding “Gatorade” on the bench rather than a real player...

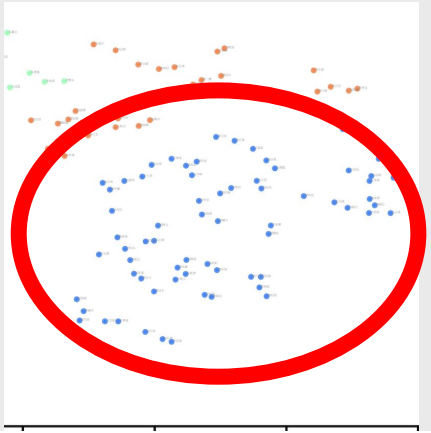




時間(分)	二分	%	三分	%	罰球	%	投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
10	1-2.6	39.13%	0.1-0.2	33.33%	0.1-0.3	50%	1.1-2.8	38.67%	2.4	1.2	1	0.5	0.5	0.1	1

Defensive Center

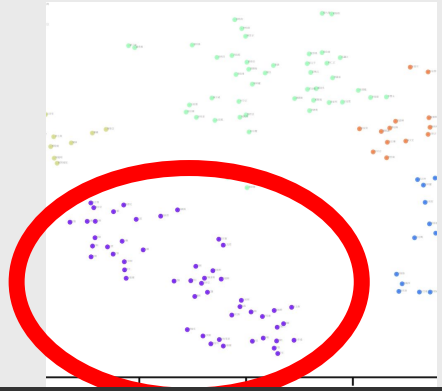
Average Height	196.6 cm
Featured Players	吳岱豪、陳冠全、李德威、蔡文誠
Features	• extremely few three point attempts • more rebounds • more fouls (in order to protect the paint) • Most of them are Taiwanese centers who works hard on guarding foreign players)



時間(分)	二分	%	三分	%	罰球	%	投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
23	3.8-7.4	51.69%	0.4-1	37.04%	1.8-2.6	69.44%	4.2-8.4	50%	10.5	4.8	1.3	1.5	0.6	1	1.6

Foreign Players & Best Taiwanese Players

Average Height	202.8 cm
Featured Players	林志傑、陳信安、曾文鼎 Q.Davis (Naturalized Taiwan)、S. Bular (7’5”)、 OJ.Mayo (Former NBA Player)
Features	• more scores • more rebounds • even much more blocks per game.



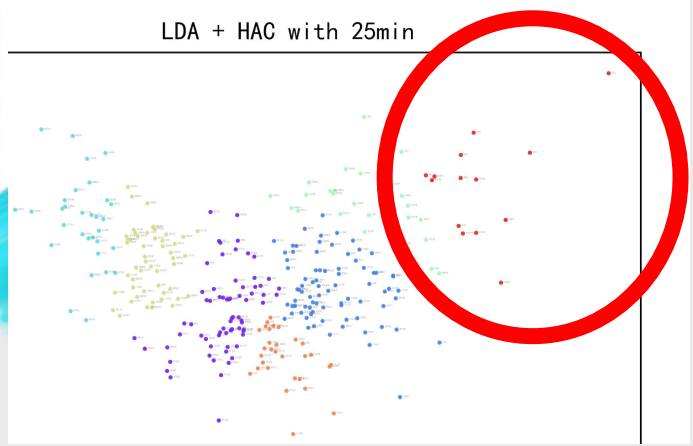
時間(分)	二分	%	三分	%	罰球	%	投籃	%	得分	籃板	助攻	失誤	抄截	阻攻	犯規
33	3.4-5.9	58.49%	3.7-7.6	48.53%	4.2-5.9	71.7%	7.1-13.4	52.89%	22.1	9.7	3.8	3.3	1.7	0.1	2.

Fun Facts

- Taiwanese Players in the Category with Foreign Players
  - 林志傑（Chosen to CBA）
  - 曾文鼎（Chosen to CBA）
  - 陳信安（Chosen to CBA、NBA Summer League Tryout）

* The Taiwanese players in the final group are usually chosen and traded to China to play for their more competitive professional league CBA there.

Fun Facts



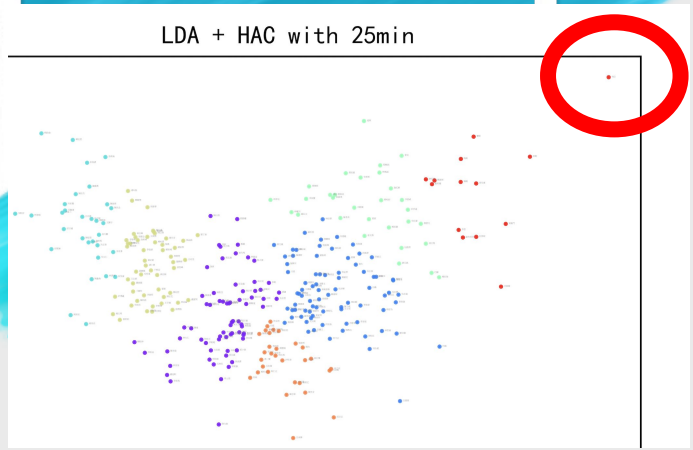
With the combination of LDA with KMeans, the foreign players are separated into “two” groups.

The players in the group with the better performance are the mainly the players “after season 11”.

We infer that this is because:

- Before season 11, SBL has a height limit of 205 cm for the foreign players
- But since season 11, SBL cancelled the height limit and resulted in better players coming to Taiwan.

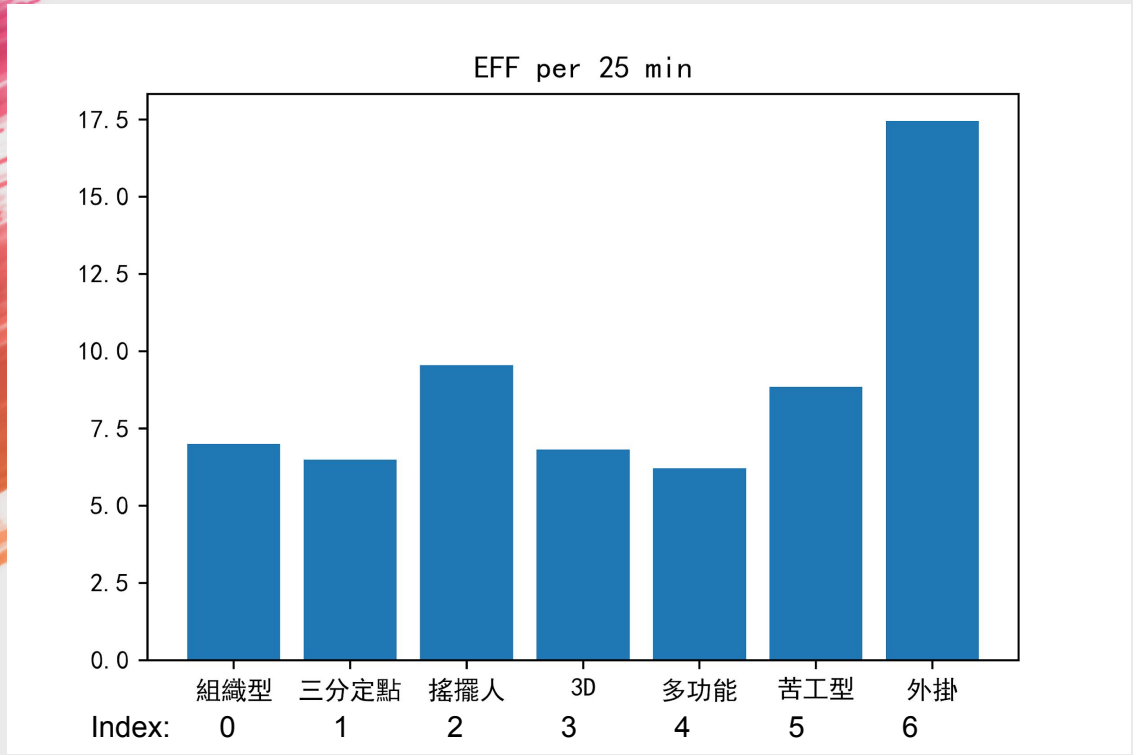
Fun Facts



When we did the visualization after the dimensionality reduction, we can see a point that lies far away from all the other data points.

It is a player called Sim Bular who is from India, and is 226 cm tall (7 feet 5 inches) and he was totally a beast and a nightmare to all the other teams.

Players Average EFF



* We can easily see that the foreign players are having much higher efficiency than the other groups.

Champion Teams' Starting Lineup

S14: 臺北達欣工程
S15: 桃園璞園
S16: 富邦勇士

S14

周儀翔 : Swingman
蘇翊傑 : Guard
林宜輝 : 3D Player
施顏宗 : Defensive C.
S. Bular : Foreign Play

S15

陳堅恩 : Swingman
林金榜 : Forward
簡浩 : 3D Player
吳岱豪 : Defensive C.
Q.Davis : Foreign Player

S16

林書緯 : Swingman
洪志善 : Guard
T.Mitchell : Foreign Player
蔡文誠 : Defensive C.
C.Garcia : Foreign Player

* We took a look at the starting lineup of the champion teams these years and found that they are pretty similar.

THANKS!

Any questions?