

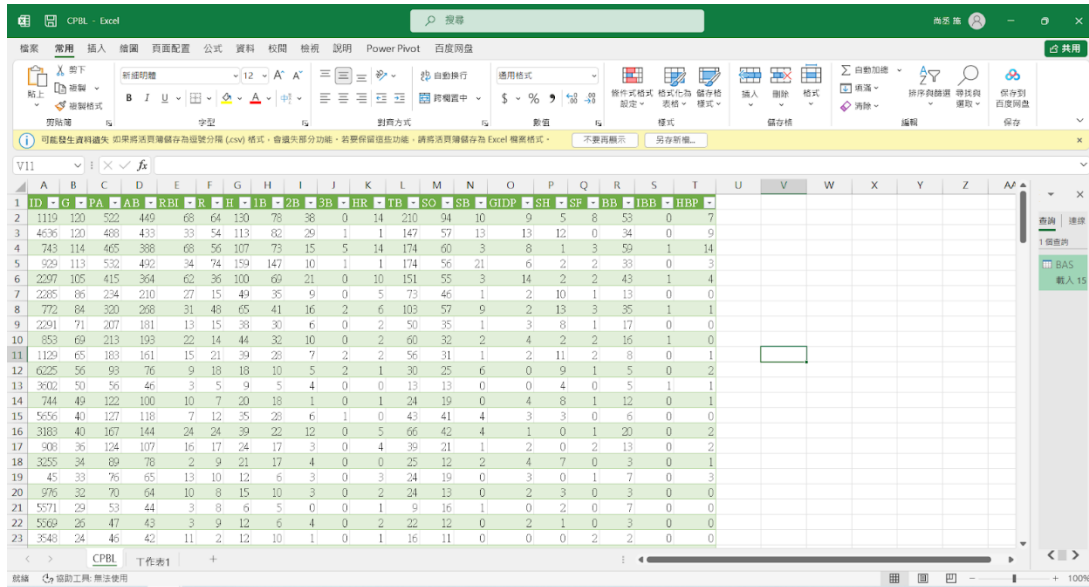
顧客關係管理 期末報告

《利用weka對資料集進行分群分析-以2021年中華職棒打者為例》

組員:企管三 施尚丞

1. 蒐集資料

首先至github尋找可用之資料。



	ID	G	PA	AB	R	RBI	H	1B	2B	3B	HR	TB	SO	SB	GDP	SH	SF	BB	IBB	HBP
1	1119	120	522	449	68	64	130	78	38	0	14	210	94	10	9	5	8	53	0	7
2	4636	120	408	433	33	54	113	82	29	1	1	147	57	13	13	12	0	34	0	9
3	743	114	465	368	68	56	107	73	15	5	14	174	60	3	8	1	3	59	1	14
4	929	113	532	492	34	94	159	147	10	1	1	174	56	21	6	2	2	33	0	3
5	2297	105	415	364	62	36	100	69	21	0	10	151	55	3	14	2	2	43	1	4
6	2235	86	234	210	27	15	49	35	9	0	5	73	46	1	2	10	1	13	0	0
7	772	84	320	268	31	43	65	41	16	2	6	103	57	9	2	13	3	35	1	1
8	2291	71	207	181	13	15	38	30	6	0	2	50	35	1	3	8	1	17	0	0
9	853	69	213	193	22	14	44	32	10	0	2	60	32	2	4	2	2	16	1	0
10	1129	65	183	161	15	21	39	28	7	2	2	56	31	1	2	11	2	8	0	1
11	6225	56	93	76	9	18	18	10	5	2	1	30	25	6	0	9	1	5	0	2
12	3602	50	56	46	3	5	9	5	4	0	0	13	13	0	0	4	0	5	1	1
13	744	49	122	100	10	7	20	18	1	0	1	24	19	0	4	8	1	12	0	1
14	5656	40	127	118	7	12	35	28	6	1	0	43	41	4	3	3	0	6	0	0
15	3183	40	167	144	24	24	39	22	12	0	5	66	42	4	1	0	1	20	0	2
16	908	36	124	107	16	17	24	17	3	0	4	39	21	1	2	0	2	13	0	2
17	3255	34	89	78	2	9	21	17	4	0	0	25	12	2	4	7	0	3	0	1
18	45	33	76	65	13	10	12	6	3	0	3	24	19	0	3	0	1	7	0	3
19	976	32	70	64	10	8	15	10	3	0	2	24	13	0	2	3	0	3	0	0
20	5571	29	53	44	3	8	6	5	0	0	1	9	16	1	0	2	0	7	0	0
21	5069	26	47	43	3	9	12	6	4	0	2	22	12	0	2	1	0	3	0	0
22	3548	24	45	42	11	2	12	10	1	0	1	16	11	0	0	0	2	2	0	0

上圖資料為2021年有在中華職棒一軍出賽之選手紀錄。

- 欄位屬性保留：
球員編號(ID)、出賽數(G)、打席(PA)、打數(AB)、得分(R)、打點(RBI)、安打(H)、一壘安打(1B)、二壘安打(2B)、三壘安打(3B)、全壘打(HR)、壘打數(TB)、被三振數(SO)、盜壘數(SB)、雙殺打(GDP)、犧牲觸擊(SH)、高飛犧牲打(SF)、四壞球(BB)、敬遠(IBB)
- 刪除欄位屬性：
HBP(觸身球)、滾地球出局(GO)、飛球出局(FO)、盜壘失敗(CS)。上述屬性若打擊數多，得到此數據的機會呈現正相關，與分群較無顯著價值，因此排除欄位。

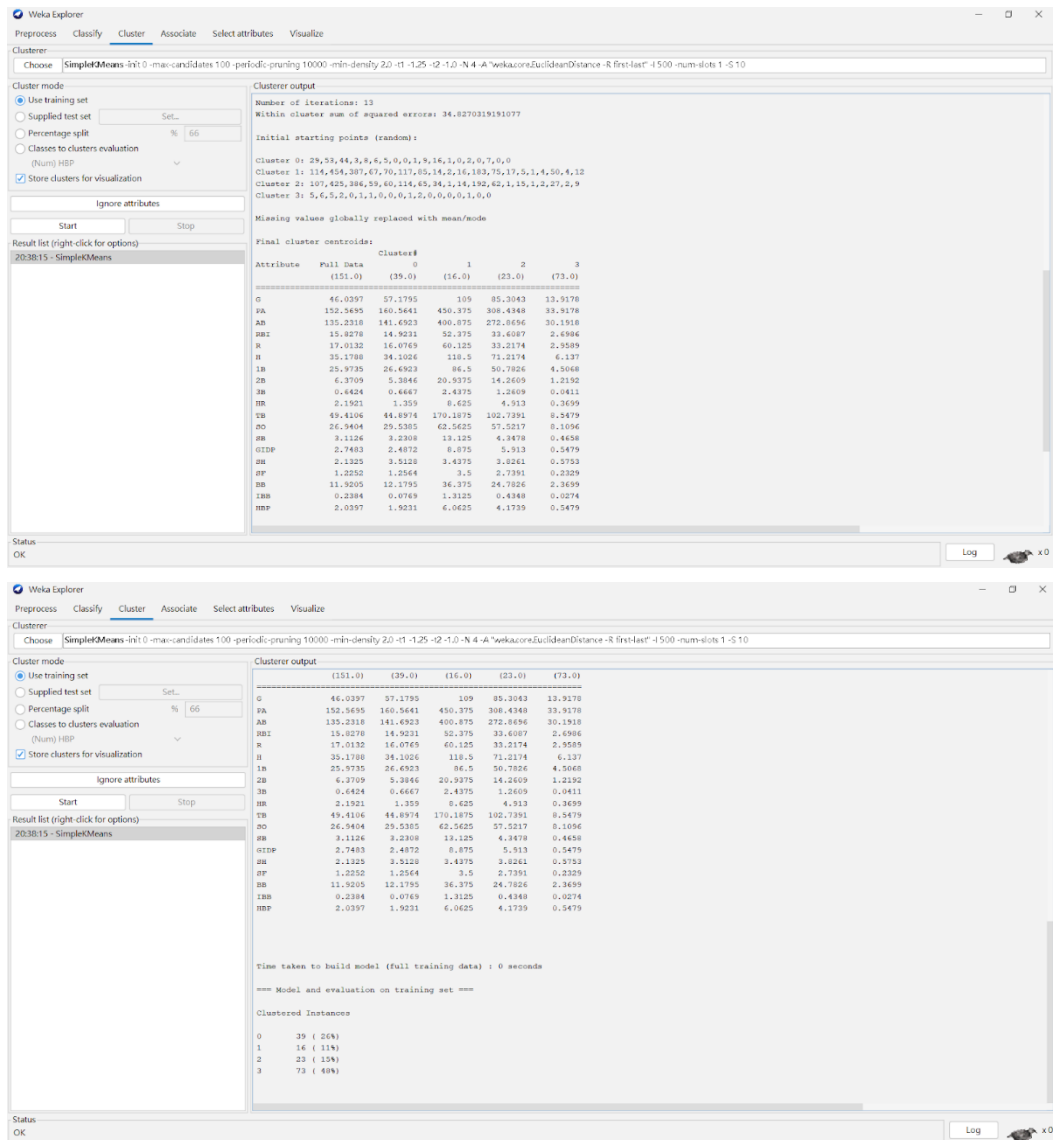
1. Data Mining

使用K-Means分析方式，判斷其分群之結果，以及相同分群內的相似性、不同分群間的差異性。將現有的Dataset區分為4群，故在一開始設定K-Means的參數時，將「numClusters」設定為4。

預設參數

Dataset	CPBL
分群數	4
忽略欄位	ID

使用K-Means所得到的分析結果如下圖：



2. 結論

□ 分群: 依照數據結果中顯示的上場表現, 由高到低排列。

● Cluster1: 全明星(11%)

Cluster1的欄位屬性表現上, 所有數據都表現優異, 包括出賽數、安打數等, 可以看到SH欄位在前3個群組中數據最低為3.4375, 擁有更多的打席數下卻拿到較少的觸擊機會, 顯示教練團更加信任這群選手

的攻擊能力，而不希望只有短程的推進。在IBB的欄位可以看到這群球員受到故意四壞球(敬遠)的機會遠大於其他選手，表示其他球隊欲到時寧可將他們送上一壘也不要遭到更猛烈的攻擊。

- Cluster2: 一線先發球員(15%)

Cluster2的欄位屬性表現上，所有數據表現也在聯盟平均之上，出賽穩定，但攻擊能力較明星球員不足。另外一線球員雖然打席數少了明星球員142個，但三振數(SO)卻只有少5個，即便可以穩定出賽，顯現在選球方面和擊球確實度還是需要多加著墨。

- Cluster0: 板凳&新秀潛力球員(26%)

Cluster0的欄位屬性表現上，雖然拿到接近聯盟平均的出賽機會、打席數，但進攻表現卻低於聯盟平均，包括打點(RBI)、安打(H)，同時也沒有好的長打能力，例如全壘打(HR)、壘打數(TB)，三振數卻高於聯盟平均(26.9404)的29.5385次。這群球員可能需要更多的實戰經驗累積，以在關鍵時刻得以發揮實力，甚至往一線球員邁進。

- Cluster3: 一軍輪替末端&新秀球員(48%)

Cluster3的欄位屬性表現上，所有數據皆大幅低於平均，表示可能實力還尚未受到教練團的認可，或是因隊上傷病得到短暫的出賽次數，拿到的機會不多。這些球員大多數的技術都尚未純熟，也有可能是老將能力下滑導致，新球員就必須在二軍磨練，老將則將面臨被淘汰的命運。

- 有趣的數據

- 盜壘(SB):

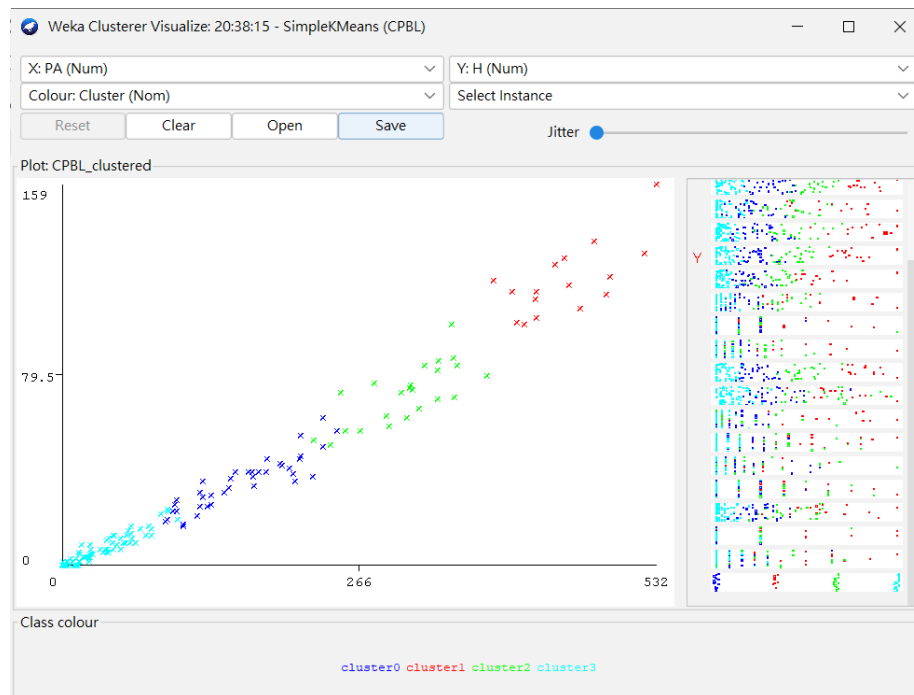
在盜壘方面的數據，Cluster1(明星球員)的盜壘數為13.13次，而Cluster0(一線球員)和Cluster2(板凳球員)的盜壘數分別為3.23次和4.35次，與聯盟平均3.11次相差不多，表示不只打擊好，腳程快、壘間判斷力佳也是成為明星球員的條件之一。

- 高飛犧牲打(SF):

高飛犧牲打成功次數可以展現一個球員的戰術執行能力和擊球確實度，在明星球員和一線先發球員的高飛犧牲打次數為3.5次及2.74次，板凳球員僅符合平均值的1.25次，表示有得分機會時教練相對不信任板凳新秀上場，一軍輪替末端的機會更是稀少。

3. 心得

經由上述的分群結果可以發現，穩定出賽的人數僅占全體的26%，接近一半的人(Cluster3)在一軍只有短短的亮相機會，這份資料甚至不包含從未上過一軍的二軍農場選手。



X軸:打席數(PA)/Y軸:安打數(H)

我們可以將打席數理解為球員所得到的機會，安打數則為球員將機會轉化為貢獻的次數，上圖可以直觀觀察到，紅點和綠點的人數稀少，但深藍點和淺藍點卻高密度的在原點周圍聚集。由此可見，職棒的環境具有高度競爭力，不符合戰力的球員極有可能遭到汰換，高薪資的報酬下同時存在險惡的高風險，若要生存下去，選手勢必要付出一定的努力外，也要三思自己的能力，否則在被環境淘汰後的路將會相當狹隘。

4. 參考資料

<https://github.com/ldkrsi/cpbl-opendata/tree/master/CPBL/batting>