作品名稱：美國職棒數據初探

作　　者：張閔翔 經濟二 B11303055

摘要

此份報告主要會分成兩個研究主題，首先，我整理了今年(2023 年)每一場大聯盟比賽的資料，研究進場看球的觀眾人數，是否會對於主場或客場的球隊在表現上有明顯的影響。再者，我會利用今年每一顆職棒打者打出去的球，去解析要打出什麼類型的球比較有機會形成安打。

壹、 研究動機

根據心理學的研究指出，當一個場地的觀眾人數愈多的時候，能夠更凝結主場球隊的向心力，甚至可以讓所有人的心跳同步，不過也有可能只會單純增加主場球隊的壓力，因此從實際的數據來分析會比較客觀。至於想要研究擊球方面的數據，是因為平常在看棒球的人很喜歡去評價擊球品質的好壞，但是大多都沒有理性上的依據，所以我決定去利用官方提供的資料，定義「擊球品質」。

貳、 研究目的

1. 研究主場及客場球隊在不同觀眾人數的場次、勝場和得分之差異。
2. 利用數據分析各種擊球條件形成的結果好壞。

參、 研究過程及方法

這次使用的軟件是 pybaseball，它是利用爬蟲的方式，或是直接抓取網站提供的 CSV 檔之後，幫使用者抓下數據並轉換成 Pandas 的 DataFrame。此份報告主要擷取的數據是由 Statcast 和 Baseball Reference 這兩個網站所提供的。

我們先來進行數據初探，圖一是今年紅襪隊所有進行的比賽，只要是能夠客觀記錄下來的比賽相關數據都已經呈現在這個 data frame 裡面了，但此份報告只

需要主客場、勝負、得分及觀眾人數，所以稍微整理一下之後，我們得到了每支球隊在主場比賽時的得分、對手得分還有觀眾人數，以圖二為示。

	Date	Tm	Home_Away	Opp	W/L	R	RA	Inn	W-L	Rank	GB	Win	Loss	Save	Time	D/N	Attendance	cLI	Streak	Orig. Scheduled
1	Thursday, Mar 30	BOS	Home	BAL	L	9.0	10.0	9.0	0-1	5.0	1.0	Gibson	Kluber	Bautista	3:10	D	36049.0	.98	-1	None
2	Saturday, Apr 1	BOS	Home	BAL	W-wo	9.0	8.0	9.0	1-1	2.0	1.0	Jansen	Bautista	None	3:04	D	29062.0	.92	1	None
3	Sunday, Apr 2	BOS	Home	BAL	W	9.0	5.0	9.0	2-1	2.0	1.0	Houck	Irvin	None	2:44	D	27886.0	.97	2	None
4	Monday, Apr 3	BOS	Home	PIT	L	6.0	7.0	9.0	2-2	3.0	2.0	Underwood	Crawford	Bednar	2:57	N	28369.0	.92	-1	None
5	Tuesday, Apr 4	BOS	Home	PIT	L	1.0	4.0	9.0	2-3	4.0	3.0	Contreras	Pivetta	Bednar	2:36	N	28842.0	.84	-2	None
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
158	Wednesday, Sep 27	BOS	Home	TBR	L	0.0	5.0	9.0	76-82	5.0	23.0	Glasnow	Bello	None	2:26	N	34559.0	.00	-4	None
159	Thursday, Sep 28	BOS	@	BAL	L	0.0	2.0	9.0	76-83	5.0	24.0	Kremer	Sale	Wells	2:30	N	27543.0	.00	-5	None
160	Friday, Sep 29	BOS	@	BAL	W	3.0	0.0	9.0	77-83	5.0	23.0	Pivetta	Means	Whitlock	2:24	N	28192.0	.00	1	None
161	Saturday, Sep 30	BOS	@	BAL	L	2.0	5.0	9.0	77-84	5.0	24.0	Zimmermann	Winckowski	None	2:54	N	43150.0	.00	-1	None
162	Sunday, Oct 1	BOS	@	BAL	W	6.0	1.0	9.0	78-84	5.0	23.0	Houck	Coulombe	None	2:44	D	36640.0	.00	1	None

162 rows × 20 columns

▲圖一 今年所有的比賽數據（以紅襪隊為例）

	Home_Away	W/L	R	RA	Attendance
1	Home	L	9.0	10.0	36049.0
2	Home	W	9.0	8.0	29062.0
3	Home	W	9.0	5.0	27886.0
4	Home	L	6.0	7.0	28369.0
5	Home	L	1.0	4.0	28842.0
...	...	...	...	...	...
154	Home	W	3.0	2.0	37102.0
155	Home	L	0.0	1.0	33392.0
156	Home	L	2.0	3.0	33399.0
157	Home	L	7.0	9.0	34094.0
158	Home	L	0.0	5.0	34559.0

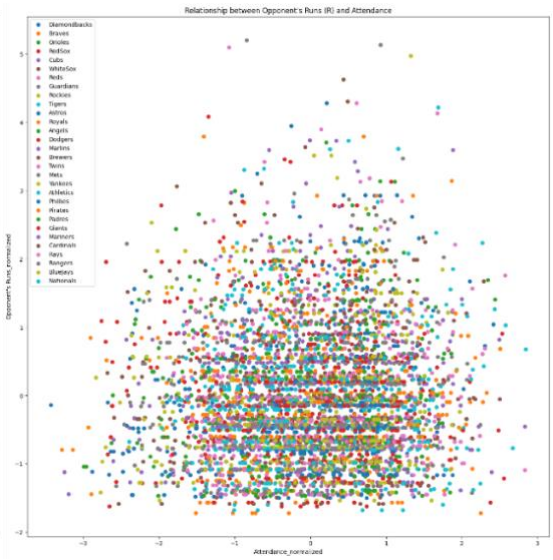
81 rows × 5 columns

▲圖二 整理後 data frame（共 30 個）

至於擊球的數據部分，我搜集了今年有效（打進場內）的擊球，在全部 73 萬顆投手所投出來的球之中，有 12 萬 5 千多筆的資料符合資格。若先把所有擊球的數據按照該球的擊球初速和系統計算出的預期加權上壘率（註 1）繪製成圖，可以發現在大約 95 英里之前的擊球數值大致相同，不過一旦超過 95 英里，預期加權上壘率呈現明顯的線性成長。有趣的是，大概在 115 英里的地方，有個很明顯的下降，因此我們進一步去結合擊球仰角的數據，此分析在下個章節會有詳細的說明。

肆、　研究結果與討論

由圖三和圖四可見，橫軸為各球隊觀眾人數標準化之後的數據，而縱軸則是得分標準化後的數據，我們發現現場觀眾人數對於主客隊的得分並沒有什麼效果，主場的相關係數為-0.023，客場的相關係數為 0.045。



▲圖三　主隊觀眾人數 vs.得分　　▲圖四　客隊觀眾人數 vs.得分

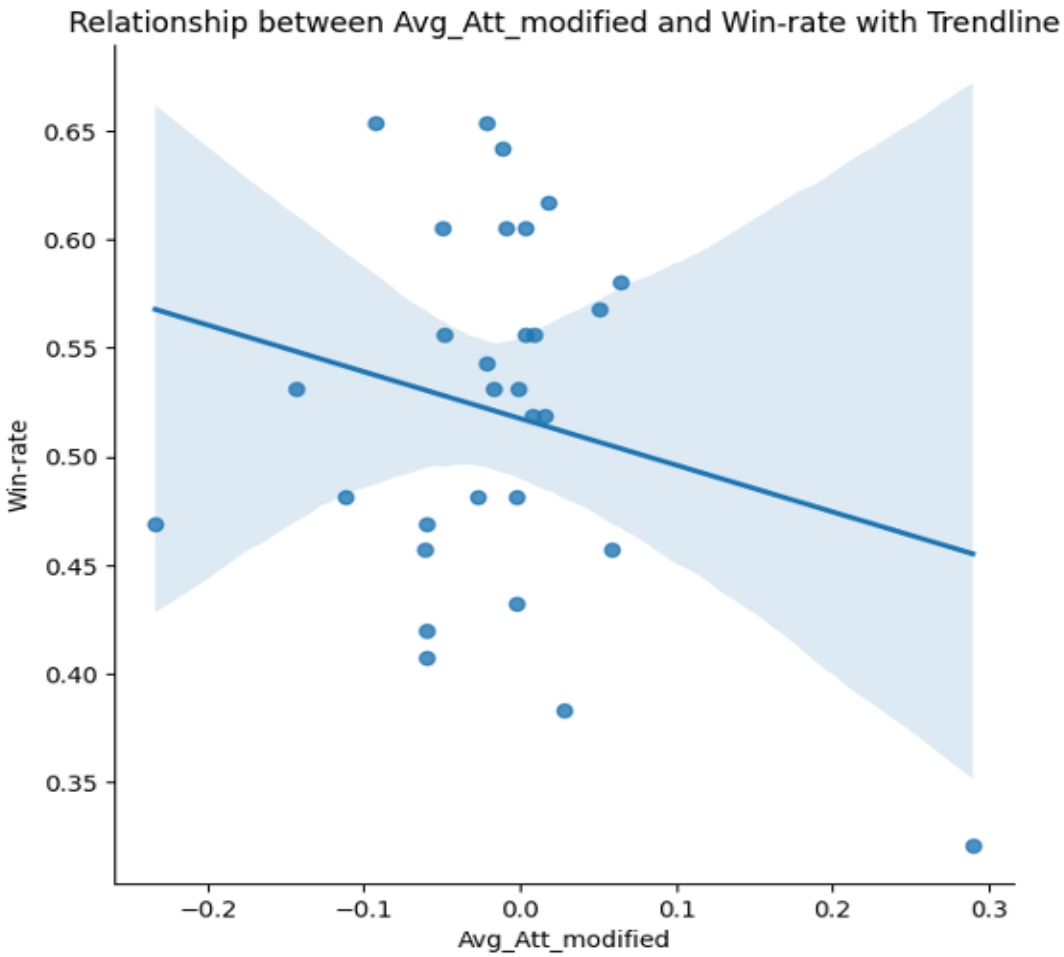
不過得分的多寡並不會直接影響球隊戰績，因為每一場比賽的關鍵應該在於勝利（更直接的說法是，贏 1 分和贏 10 分是一樣的意思），所以我們應該要進一步分析今年每支球隊在贏球還有輸球的平均觀眾人數，並且觀察到了以下的現象。我發現今年的 30 個主場觀眾人數，勝場的平均多於敗場的只有 11 隊，而其中有打進季後賽的更只有 4 支球隊，而敗場平均人數比較多的隊伍則有 8 支晉級到季後賽。我們進一步將數據繪製成圖七，可以看到橫軸的部分是我將各支球隊勝場的平均觀眾人數減掉敗場的平均觀眾人數，除以該支球隊的總平均人數之後，再去跟球隊季賽勝率做比較。可以發現雖然相關係數仍然只有大約 -0.219，但藉由這張圖，和剛剛提到各支球隊晉級季後賽的現象，已經可以說明觀眾人數愈多或許真的對主場球隊有微弱的負面影響。

	Team	Avg_Attendance_W	Avg_Attendance_L	Avg_Attendance	Avg_Att_modified	Wins	Losses	Win-rate
0	Diamondbacks	22579.91	26059.11	24212.12	-0.143697	43	38	0.531
1	Braves	39243.87	39683.59	39401.30	-0.011160	52	29	0.642
2	Orioles	23438.31	24635.03	23911.09	-0.050049	49	32	0.605
3	RedSox	32952.38	33023.50	32989.26	-0.002156	39	42	0.481
4	Cubs	34312.58	34196.75	34261.10	0.003381	45	36	0.556
5	WhiteSox	21770.63	21177.27	21405.49	0.027720	31	50	0.383
6	Reds	22045.58	27920.23	25164.22	-0.233453	38	43	0.469
7	Guardians	23680.55	23319.03	23513.69	0.015375	42	39	0.519
8	Rockies	33225.65	31331.50	32196.73	0.058831	37	44	0.457
9	Tigers	20250.34	21526.52	20946.44	-0.060926	37	44	0.457
10	Astros	37143.36	38184.67	37683.30	-0.027633	39	42	0.481
11	Royals	15567.24	16527.77	16136.44	-0.059526	33	48	0.407
12	Angels	31556.79	33521.33	32599.69	-0.060262	38	43	0.469
13	Dodgers	47024.34	48028.18	47371.35	-0.021191	53	28	0.654
14	Marlins	14673.24	13938.57	14355.79	0.051176	46	35	0.568
15	Brewers	31540.08	31433.84	31498.11	0.003373	49	32	0.605
16	Twins	25024.81	23469.35	24371.90	0.063822	47	34	0.580
17	Mets	32720.98	33297.16	32994.29	-0.017463	43	38	0.531
18	Yankees	41010.79	40699.03	40862.70	0.007629	42	39	0.519
19	Athletics	12297.54	9320.29	10275.95	0.289730	26	55	0.321
20	Phillies	38019.79	38364.22	38157.56	-0.009026	49	32	0.605
21	Pirates	18965.13	21213.90	20131.16	-0.111706	39	42	0.481
22	Padres	39982.34	40873.81	40389.56	-0.022072	44	37	0.543
23	Giants	30205.76	31691.50	30866.09	-0.048135	45	36	0.556
24	Mariners	33352.64	33043.03	33215.04	0.009321	45	36	0.556
25	Cardinals	39955.74	40057.39	40013.47	-0.002540	35	46	0.432
26	Rays	17212.55	18858.43	17781.49	-0.092561	53	28	0.654
27	Rangers	31492.54	30916.68	31272.15	0.018415	50	31	0.617
28	BlueJays	37291.12	37325.95	37307.46	-0.000934	43	38	0.531
29	Nationals	22231.38	23616.28	23034.96	-0.060122	34	47	0.420

▲圖五 各隊觀眾人數調整後之數據

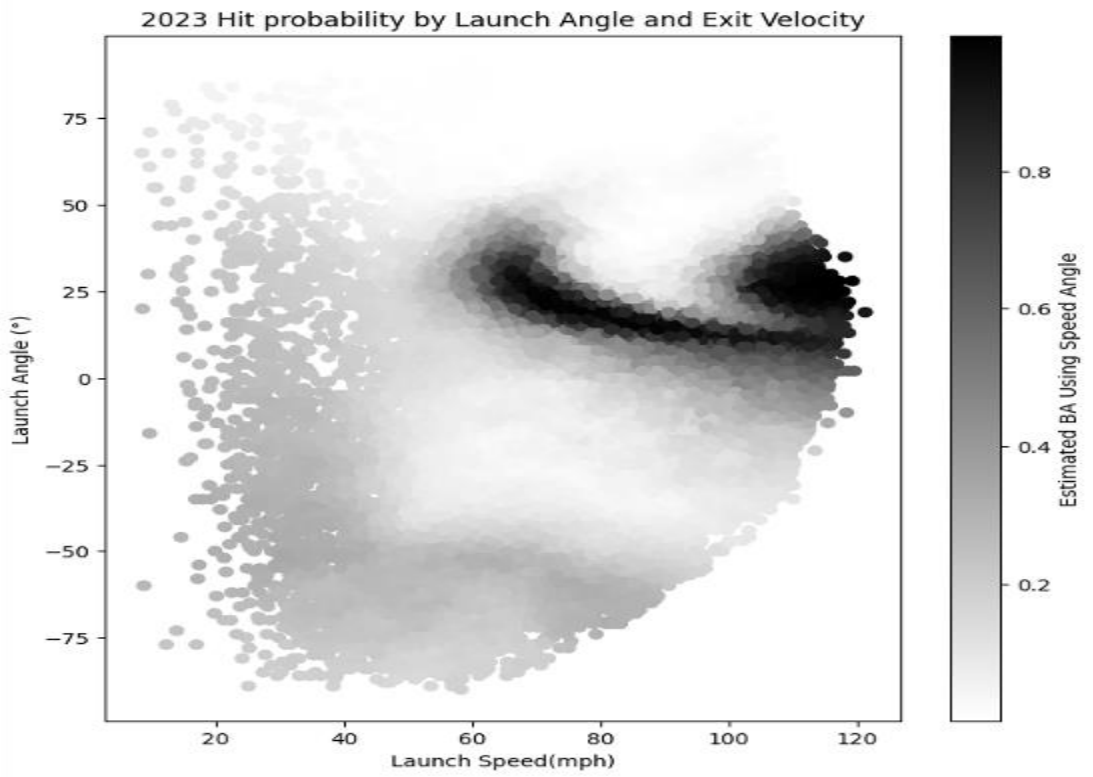
```
W Count: 11
L Count: 19

4          Cubs
5       WhiteSox
7      Guardians
8       Rockies
14      Marlins
15      Brewers
16        Twins
18      Yankees
19    Athletics
24     Mariners
27     Rangers
```

勝場觀眾人數多於敗場之隊伍 圖六▲

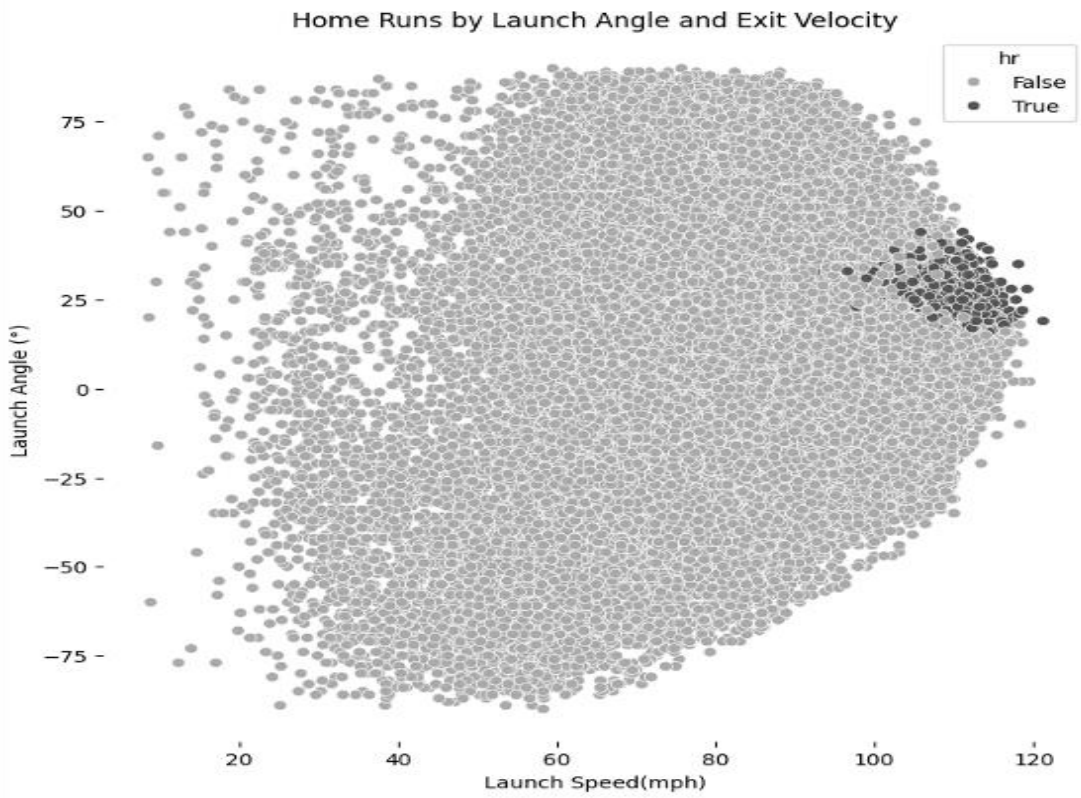


▲圖七 觀眾人數 vs. 勝率

有關打擊數據的部分，我們延續上一章節，加入擊球仰角的數據進行分析，並發現在固定的擊球仰角之下（大約 25 度上下），只要擊球初速超過大約 60 英里，即有相當高的預期打擊率。再來觀察到的是圖八最黑的部分大約在擊球初速 100 到 120 英里，擊球仰角 20 到 40 度之間的一個範圍，我們可以發現在這個範圍中，預期的打擊率幾乎要趨近於 1，也就是說達成這樣條件的擊球，幾乎可以確定是一支安打。我也整理了今年所有全壘打球，恰好幾乎全部都落在前述的範圍之中（圖九），因此可以說明只要滿足此擊球條件，就有非常高的機會形成全壘打。



▲圖八 擊球初速 vs.仰角之預期打擊率



▲圖九 今年全壘打的擊球初速及仰角

伍、 結論

1. 主場觀眾人數對於主隊及客隊的得分數並沒有太顯著的影響

2. 贏球場次觀眾人數較多的隊伍，並不會比較有機會晉級季後賽

3. 主場觀眾人數，整體來看與主場勝率呈微弱負相關

4. 擊球初速愈高會提升整體打擊率，不過在某些條件下仍然可能會降低

5. 形成全壘打最佳條件約在 20-40 度的擊球仰角和 100-120 英里的初速

陸、 參考文獻

1. The New Science Of Hitting
https://fivethirtyeight.com/features/the-new-science-of-hitting/

2. Python 棒球數據分析套件 pybaseball 介紹
https://ithelp.ithome.com.tw/m/users/20163024/ironman/6694

3. 大聯盟小百科-預期加權上壘率 https://reurl.cc/Z9nD9Q