

2020 DB Term Project Report 02 組

Data

Source

[Kaggle Pitch Data 2015-2018](#)

Five Tables From Kaggle

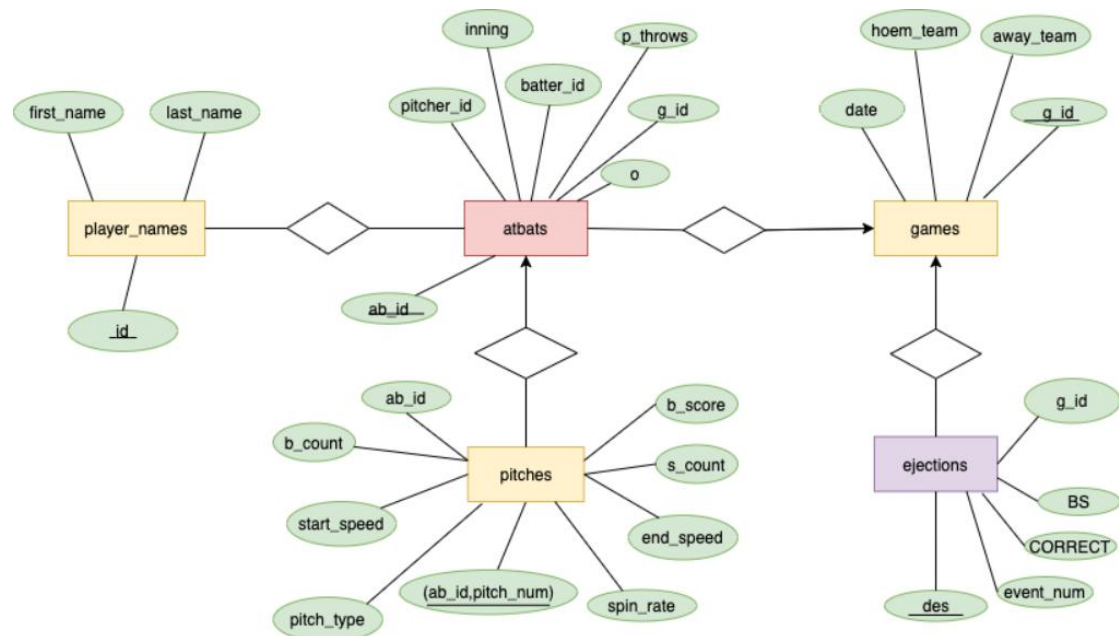
- **games**
 - **attendance** number of fans who attended (NOTE: for first game of doubleheaders, value is often erroneously 1 or 0. This comes directly from XML files. This data may not be recorded for those games; MLB gameday pages do not report attendance for these game)
 - **away_final_score** final score for the visiting team
 - **away_team** three letter abbreviation for away team; third letter often indicates league(national vs american)
 - **date** date of game
 - **elapsed_time** length of game, in minutes
 - **g_id** game ID. Matches with game_id in atbats.csv
 - **home_final_score** final score for the home team
 - **home_team** three letter abbreviation for home team; third letter often - - indicates league(national vs american)
 - **start_time** start time of game
 - **umpire_1B** umpire of 1B
 - **umpire_2B** umpire of 2B
 - **umpire_3B** umpire of 3B
 - **umpire_HP** umpire of HP
 - **venue_name** name of stadium
 - **weather** description of weather
 - **wind** description of wind
 - **delay** length of delay before game, in minutes
- **ejections**

- **ab_id** foreign key for atbats.csv, may be unreliable (ejection happened before, after, during atbat)
- **des** Human readable, in format
- **event_num** event number for ejection (from xml file; many event_nums are skipped)
- **g_id** foreign key for games.csv
- **player_id** foreign key for player_names.csv
- **date** directly from games.csv
- **BS** 'Y' if ejection was for arguing balls and strikes, empty otherwise
- **CORRECT** Whether the ejection was correct (only for BS ejection). From closecallsports.com
- **team** team for player ejected
- **is_home_team** whether that team is the home team-
- **pitches.** (Pitch-level data, including lots of information about the trajectory of the pitch. Match up with atbats.csv for complete picture of game situation. Data comes from unlabeled xmls from MLB website, so the meaning of some fields is not clear.)
 - **px** x-location as pitch crosses the plate. X=0 means right down the middle
 - **pz** z-location as pitch crosses the plate. Z=0 means the ground
 - **start_speed** Speed of the pitch just as it's thrown
 - **end_speed** Speed of the pitch when it reaches the plate
 - **spin_rate** The pitch's spin rate, measure in RPM
 - **spin_dir** Direction in which pitch is spinning, measured in degrees
 - **break_angle**
 - **break_length**
 - **break_y**
 - **ax**
 - **ay**
 - **az**
 - **sz_bot**
 - **sz_top**
 - **type_confidence** Confidence in pitch_type classification. Goes up to 2 for some reason.
 - **vx0**
 - **vy0**

- **vz0**
- **x**
- **x0**
- **y**
- **y0**
- **z0**
- **pfx_x**
- **pfx_z**
- **nasty**
- **zone**
- **code** Records the result of the pitch. See dataset description for list of codes and their meaning
- **type** Simplified code, S (strike) B (ball) or X (in play)
- **pitch_type** Type of pitch. See dataset description for list of pitch types
- **event_num** event number, used for finding when exactly ejections happen.
- **b_score** score for the batter's team
- **ab_id** at-bat ID. Matches up with atbats.csv
- **b_count** balls in the current count
- **s_count** strikes in the current count
- **outs** number of outs (before pitch is thrown)
- **pitch_num** pitch number (of at-bat)
- **on_1b** True if there's a runner on first, False if empty
- **on_2b** True if there's a runner on second, False if empty
- **on_3b** I don't know
- **atbats.** (This file lists the information that cannot change over the course of an at-bat)
 - **ab_id** at-bat ID. First 4 digits are year. Matches with ab_id in pitches.csv
 - **batter_id** player ID of the batter. Given by MLB, player names found in player_names.csv
 - **event** description of the result of the at-bat
 - **g_id** game ID. First 4 digits are year
 - **inning** inning number
 - **o** number of outs after this at-bat
 - **p_score** score for the pitcher's team

- **p_throws** which hand pitcher throws with. Single character, R or L
- **pitcher_id** player ID of the pitcher. Given by MLB, player names found in player_names.csv
- **stand** which side batter hits on. Single character, R or L
- **top** True if it's the top of the inning, False if it's the bottom
- **player_names** (Matches names with player's ID)
 - **id** matches with batter_id and pitcher_id
 - **first_name** first name
 - **last_name** last name

ER Diagram



Other Tables Extend From The Tables Above

- **batter_create_table_per_game2** (每場比賽各打者的各種資料)
 - **years** 年度
 - **id** 對應 player_id
 - **first_name** first name
 - **last_name** last name
 - **g_id** game ID
 - **PA** 擔任打席次數

- **AB** 打數
- **Single** 一壘安打數
- **DDouble** 二壘安打數
- **Triple** 三壘安打數
- **HR** 全壘打數
- **K** 三振次數
- **BB** 四壞球次數
- **HBP** 觸身球次數
- **IBB** 故意四壞球次數
- **SF** 高飛犧牲打次數
- **GDP** 滾地球雙殺次數
- **Date** 日期
- **batter_create_table_per_year** (每年度各打者的各種資料)
 - **years** 年度
 - **id** 對應 player_id
 - **first_name** first name
 - **last_name** last name
 - **PA** 擔任打席次數
 - **AB** 打數
 - **Single** 一壘安打數
 - **DDouble** 二壘安打數
 - **Triple** 三壘安打數
 - **HR** 全壘打數
 - **K** 三振次數
 - **BB** 四壞球次數
 - **HBP** 觸身球次數
 - **IBB** 故意四壞球次數
 - **DP** 滾地球雙殺次數
 - **AVG** 打擊率
 - **OBP** 上壘率
 - **WOBA** 打者加權指數
 - **SLG** 長打率
 - **OPS** 整體攻擊指數
 - **BABIP** 場內安打率
 - **SF** 高飛犧牲打次數
 - **ROE** 野手失誤次數
- **pitcher_create_table_per_game** (每場比賽各投手的各種資料)
 - **years** 年度

- **id** 對應 player_id
- **first_name** first name
- **last_name** last name
- **g_id** game ID
- **Single** 一壘安打數
- **DDouble** 二壘安打數
- **Triple** 三壘安打數
- **HR** 全壘打數
- **IP** 投球局數
- **Pitch_per_Game** 用球數
- **K** 三振數
- **BB** 保送數
- **HBP** 觸身球數
- **DP** 雙殺數
- **Ground** 滾地球數
- **Fly** 飛球數
- **ground_fly_ratio** 滾飛比(滾地球跟飛球的比例)
- **date** 日期
- **pitcher_create_table_per_year2** (每年度各投手的各種資料)
 - **years** 年度
 - **id** 對應 player_id
 - **first_name** first name
 - **last_name** last name
 - **IP** 投球局數
 - **pitch_num** 用球數
 - **Single** 一壘安打數
 - **DDouble** 二壘安打數
 - **Triple** 三壘安打數
 - **HR** 全壘打數
 - **H9** 平均 9 局被安打次數
 - **K** 三振數
 - **K9** 平均 9 局三振次數
 - **BB** 保送數
 - **BB9** 平均 9 局保送次數
 - **IBB** 故意四壞球次數
 - **HBP** 觸身球數
 - **DP** 雙殺數
 - **Ground** 滾地球數

- **Fly** 飛球數
- **ground_fly_ratio** 滾飛比(滾地球跟飛球的比例)
- **FIP** FIP
- **BABIP** 場內被安打率
- **WHIP** 每局被上壘率
- **pitch_type_create_table_per_game** (每場比賽各投手的各球種資料)
 - **years** 年度
 - **id** 對應 player_id
 - **first_name** first name
 - **last_name** last name
 - **g_id** game ID
 - **pitch_type** 球種
 - **use_count** 使用次數
 - **use_ratio** 使用率
 - **strike_count** 好球次數
 - **strike_ratio** 好球率
 - **v0_avg** 初速平均
 - **v_delta_avg** 速度變化平均
 - **spin_rate_avg** 轉速平均
 - **Date** 日期
- **pitch_type_create_table_per_year** (每年度各投手的各球種資料)
 - **years** 年度
 - **id** 對應 player_id
 - **first_name** first name
 - **last_name** last name
 - **pitch_type** 球種
 - **use_count** 使用次數
 - **use_ratio** 使用率
 - **strike_count** 好球次數
 - **strike_ratio** 好球率
 - **v0_avg** 初速平均
 - **v_delta_avg** 速度變化平均
 - **spin_rate_avg** 轉速平均
- **team_final_score** (所有對戰的比分)
 - **g_id** game ID
 - **date** 日期
 - **year** 年度
 - **home_team** home team

- **away_team** away team
 - **home_final_score** 主隊最終分數
 - **away_final_score** 客隊最終分數
- **team_region_status** (所有年度, 所有聯盟分區的戰績)
 - **year** 年度
 - **team** 隊伍
 - **League** 聯盟
 - **Division** 分區
 - **win** 勝場
 - **total** 總比賽場次
 - **win_rate** 勝率
- **team_opponent** (某球隊每場比賽的相關資訊)
 - **g_id** game ID
 - **home_or_away** 該隊伍在這場比賽是主隊或客隊
 - **opponent** 對手
 - **score** 該隊伍在這場比賽的分數
 - **opponent_score** 對手在這場比賽的分數
- **team_first_inning_run_ratio** (所有球隊每一年的首局得分率, 首局總得分數)
 - **year** 年度
 - **team** 隊伍
 - **total_game** 比賽場次數
 - **scoring_rate** 首局得分率
 - **total_score** 首局總得分數
- **team_LLRR** (左投或右投對左打或右打的打擊率)
 - **year** 年度
 - **pitcher_batter_stand** 投手和打者的站位
 - **cnt_baserun** 安打數
 - **cnt_atbat** 打數
 - **AVG** 打擊率
- **ejection_game**(單場比賽驅逐出場次數大於等於 3 次的場次)
 - **date** 日期
 - **g_id** game ID
 - **home_team** home team
 - **away_team** away team
 - **ejection_cnt** 驅逐出場次數
- **ejection_max_player**(每年度驅逐出場次數最多的選手)
 - **year** 年度

- **id** 對應 **player_id**
 - **first_name** first name
 - **last_name** last name
 - **cnt** 驅逐出場次數
- **ejection_team**(每年度所有隊伍的驅逐出場次數，若次數為 0 的話不會顯示在 table 裡)
 - **year** 年度
 - **team** 隊伍
 - **cnt** 驅逐出場次數

Data Normalization

我們為了分析更多資料於是由原始的 table 中再延伸了許多 table，而過程中因想增加可識別度及方便被其他 table 使用的關係，所以有滿多 table 都有部分功能相依的狀況(如主鍵為<year, id>時，**first_name** 及 **last_name** 只相依於 id)，因此我們的 data 在包含新增的 table 的情況下並沒有滿足 2NF

Database

Database We USE

MySQL

How Do We Maintain Our Database

因為使用者在查詢資料不會想花很多時間，所以我們建很多臨時 table。

分為：打者單場數據、投手單場數據、打者全年數據、投手全年數據、投球球種全年數據、投球球種單場數據。

單場數據用(g_id,id)當 primary key、全年數據用(years,id)當 primary_key

有了這些 key，在新建 table 時會快很多

我們用單場數據的加總產生全年數據，但不是用 create_table_per_game 產生 create_table_per_year，因為在 sum 的過程非常耗時。所以是先把小的

*_per_game sum 成 *_per_year。

若資料庫發生問題也能用備用的資料來快速恢復

How Do We Connect Our Database To Our Application

我們會在報告的 Application 的部分中一併說明

Application

Interface、Function

我們使用 Grafana 的 Open Source 資料視覺化平台，架在宿舍電腦中的虛擬機中，可以從固定 IP 訪問，亦有申請域名如下。

<http://mlb.nctu.me:3000>

Database 如上述說明，是使用 MySQL，也是架在同一台虛擬機中，藉由 3306 Port，並開一個唯獨的使用者 GrafanaReader，讓 Grafana 能夠執行。

```
mysql> SHOW GRANTS FOR GrafanaReader;  
+-----+  
| Grants for GrafanaReader@% |  
+-----+  
| GRANT USAGE ON *.* TO 'GrafanaReader'@'%' |  
| GRANT SELECT ON `MLB`.* TO 'GrafanaReader'@'%' |  
| GRANT SELECT ON `MLB`.`games` TO 'GrafanaReader'@'%' |  
| GRANT SELECT ON `MLB`.`player_names` TO 'GrafanaReader'@'%' |  
| GRANT SELECT ON `MLB`.`atbats` TO 'GrafanaReader'@'%' |  
| GRANT SELECT ON `MLB`.`ejections` TO 'GrafanaReader'@'%' |  
| GRANT SELECT ON `MLB`.`pitches` TO 'GrafanaReader'@'%' |  
+-----+
```

Name	MySQL	Default	<input checked="" type="checkbox"/>
------	-------	---------	-------------------------------------

MySQL Connection

Host	localhost:3306		
Database	MLB		
User	GrafanaReader	Password	configured
TLS Client Auth		<input type="checkbox"/>	With CA Cert
Skip TLS Verify		<input type="checkbox"/>	

如此一來，Grafana 即可讀取 MySQL 的資料。

在 Grafana 內部，我們使用兩種方法進行 query。

第一種是針對運算量較高的 query，我們會先在 MySQL 跑過一次後，建立相對應的 Table，在網站上再針對特定年分、選手、勝率等資料進行 query，方法如下。

```
SELECT * FROM(
SELECT years AS year, name, PA, AB AS atbat, Single AS 1B, DDouble AS 2B, Triple AS 3B, HR, K, BB, IBB, AVG, OBP, SLG, OPS, BABIP FROM(
SELECT *, CONCAT(first_name, ' ', last_name) AS name
FROM batter_create_table_per_year
)AS T
)AS T1
WHERE year in ($year)
AND name in ($player)
AND PA BETWEEN $PA_s AND $PA_e
AND atbat BETWEEN ($atbat_s) AND ($atbat_e)
AND 1B BETWEEN $1B_s AND $1B_e
```

圖中的 batter_create_table_per_year 是一個已經建立過的 table，我們再根據使用者輸入的數值，對這個 table 進行 query (第二次 query)以滿足使用者進一步縮小範圍的需求，又能避免 query 花費過長時間。

接下來展示面向使用者的 UI，當連上網站後會出現的主畫面如下。

年分	選手	打擊數	打擊率	一壘安打數	二壘安打數	三壘安打數	全壘打數	三振數	四壞球數	故意四壞球數	打擊率	上壘率	長打率	整體攻擊指數	場內安打率
2015	Alex Rodriguez	622	0.325	75	22	1	33	145	84	5	0.2495	0.3553	0.4838	0.8311	0.3670
2015	Aramis Ramirez	517	0.275	68	31	1	17	68	31	3	0.2458	0.2959	0.4232	0.7114	0.3630
2015	Adrian Beltre	629	0.268	109	32	4	18	84	41	4	0.2870	0.3339	0.4325	0.7831	0.4489
2015	Carlos Beltran	531	0.278	78	34	1	19	84	45	2	0.2762	0.3371	0.4707	0.8059	0.4109
2015	Jayson Werth	578	0.331	44	16	1	12	84	38	0	0.2205	0.3016	0.3857	0.6800	0.3720
2015	Michael Cuddyer	408	0.279	69	18	1	10	87	24	0	0.2386	0.3088	0.5905	0.6920	0.4335
2015	A.J. Pierzynski	456	0.277	88	24	1	9	37	19	2	0.2998	0.3394	0.4300	0.7357	0.4232
2015	Jimmy Rollins	566	0.326	76	24	3	13	86	44	0	0.2221	0.2827	0.3938	0.6385	0.3323
2015	Ichiro Suzuki	442	0.302	79	5	6	1	51	31	1	0.2264	0.2760	0.2761	0.5521	0.3321
2015	Chase Utley	423	0.373	48	21	2	6	64	32	4	0.2118	0.2861	0.3432	0.6080	0.3176
2015	Marlon Byrd	544	0.266	72	25	5	23	144	29	2	0.2470	0.2984	0.4326	0.7275	0.4113
2015	Matt Holliday	577	0.275	43	16	1	4	49	36	5	0.2795	0.3035	0.4105	0.7860	0.4412
2015	Joe Mauer	667	0.303	111	34	2	10	109	67	12	0.2648	0.3373	0.3794	0.7167	0.3984
2015	Justin Morneau	183	0.269	36	10	3	3	25	13	2	0.3077	0.3607	0.4556	0.9163	0.4375
2015	Miguel Cabrera	512	0.300	98	28	1	18	82	77	15	0.3372	0.4395	0.5326	0.9681	0.4792
2015	Brandon Phillips	625	0.290	140	19	2	12	68	27	1	0.2832	0.3264	0.3932	0.7148	0.4363
2015	Omar Infante	455	0.240	65	25	7	2	66	9	0	0.2205	0.2530	0.3188	0.5819	0.3532

主畫面中可以看到上方是輸入欄，中間最大一塊是呈現數據的 Table，最右邊則是連結到不同類別的 Query，其他畫面的排版跟主畫面皆相同。

打者	投手	其他
場次別總覽 ☆	場次別總覽 ☆	場次別的對戰隊伍及比分 ☆
年度別總覽 ☆	場次球種別總覽 ☆	場次別雙方得分紀錄 ☆
	年度別總覽 ☆	場次別驅逐出場 ☆
	年度球種別總覽 ☆	年度別隊伍得分數與勝率 ☆
		年度球隊遭驅逐出場次數 ☆
		Table of Contents

圖片中央的 Table 會先列出我們的 Query 結果，並根據數值進行視覺化上色、改動，如圖中會根據每個數值的差到優、低到高進行標色淺到深，PR 60 與 PR 85 是我們預設的變色點，在打者部分，分別會是 淺綠 < 綠 < 深綠，應用上來說，假如點擊「整體攻擊指數」，系統會根據該數值進行排序，根據點擊次數，會在高到低、低到高、取消，三個狀態之間改變，如圖片中是根據整體攻擊指數進行高到低排序，輔以顏色標記後，我們可以看到整體攻擊指數高的球員，在打擊率、上壘率、長打率、場內安打率都是有較佳的表現(多為深色)。

打者年度資料															
年度	選手	打席數	打數	一壘安打數	二壘安打數	三壘安打數	全壘打數	三振數	四壞球數	故意四壞球數	打擊率	上壘率	長打率	整體攻擊指數	場內安打率
2017	J.D. Martinez	489	432	57	26	3	45	128	53	8	0.3032	0.3763	0.6898	1.0661	0.4751
2018	Mookie Betts	616	522	96	47	5	32	91	81	8	0.3448	0.4367	0.6379	1.0632	0.4884
2017	Mike Trout	507	402	62	25	3	33	90	94	15	0.3060	0.4418	0.6294	1.0594	0.4412
2017	Aaron Judge	678	542	75	24	3	52	207	127	11	0.2841	0.4218	0.6273	1.0432	0.4722
2016	David Ortiz	626	537	82	48	1	38	84	80	15	0.3147	0.4010	0.6213	1.0207	0.4107
2018	J.D. Martinez	653	573	106	37	2	43	146	69	11	0.3281	0.3997	0.6248	1.0199	0.5070
2017	Bryce Harper	493	421	77	27	1	29	98	68	11	0.3183	0.4118	0.5938	1.0056	0.4861
2015	Paul Goldschmidt	695	567	109	38	2	33	148	118	29	0.3210	0.4352	0.5697	1.0034	0.5103
2018	Christian Yelich	653	576	110	34	7	36	134	68	2	0.3247	0.4012	0.5955	0.9875	0.4840
2017	Charlie Blackmon	725	644	127	35	14	37	134	65	9	0.3307	0.3972	0.6009	0.9857	0.4706
2016	Joey Votto	678	557	116	34	2	29	119	108	15	0.3250	0.4336	0.5494	0.9771	0.4648
2015	Mike Trout	683	576	93	32	6	41	158	92	14	0.2986	0.4012	0.5885	0.9765	0.4580
2016	Mike Trout	681	549	107	32	5	29	136	116	12	0.3151	0.4405	0.5501	0.9759	0.4800
2017	Freddie Freeman	515	441	70	35	2	28	94	65	14	0.3061	0.4019	0.5850	0.9753	0.4315
2016	Daniel Murphy	582	531	107	47	5	25	57	35	10	0.3465	0.3900	0.5951	0.9731	0.4569
2015	Miguel Cabrera	512	430	98	28	1	18	82	77	15	0.3372	0.4395	0.5326	0.9681	0.4792
2018	Max Muncy	481	395	50	17	2	35	130	79	6	0.2633	0.3909	0.5823	0.9648	0.4035

↑ 打者區是以綠色為基底。

↑ 投手區是以藍色為基底。

<div> 其他 / 場次別雙方得分紀錄 </div>		<div> Last 6 hours </div>		<div> </div>		<div> </div>			
Panel Title									
比賽場次	主場或客場	對手	己方得分		對手得分				
201501491	away	tex	<div><div></div></div>	21	<div><div></div></div>	5			
201501943	away	atl	<div><div></div></div>	20	<div><div></div></div>	6			
201702140	away	tex	<div><div></div></div>	16	<div><div></div></div>	7			
201700909	home	bal	<div><div></div></div>	16	<div><div></div></div>	3			
201600030	home	hou	<div><div></div></div>	16	<div><div></div></div>	6			
201501913	away	atl	<div><div></div></div>	15	<div><div></div></div>	4			
201800318	home	min	<div><div></div></div>	14	<div><div></div></div>	1			
201700926	home	bal	<div><div></div></div>	14	<div><div></div></div>	3			
201700329	home	bal	<div><div></div></div>	14	<div><div></div></div>	11			
201601905	home	bal	<div><div></div></div>	14	<div><div></div></div>	4			
201501021	home	det	<div><div></div></div>	14	<div><div></div></div>	3			
201500666	home	kca	<div><div></div></div>	14	<div><div></div></div>	1			
201500077	home	bos	<div><div></div></div>	14	<div><div></div></div>	4			
201702185	home	bal	<div><div></div></div>	13	<div><div></div></div>	5			
201701870	away	det	<div><div></div></div>	13	<div><div></div></div>	4			
201701192	away	hou	<div><div></div></div>	13	<div><div></div></div>	4			
201601920	home	bal	<div><div></div></div>	13	<div><div></div></div>	5			
201601623	home	cle	<div><div></div></div>	13	<div><div></div></div>	7			
201501963	away	bos	<div><div></div></div>	13	<div><div></div></div>	8			
			<div><div></div></div>		<div><div></div></div>				

打書

場次別總覽

年度別總覽

投手

場次別總覽

場次按隊別總覽

年度別總覽

年度按隊別總覽

其他

場次別的對戰隊伍及比分

場次別雙方得分紀錄

場次別離隊出場

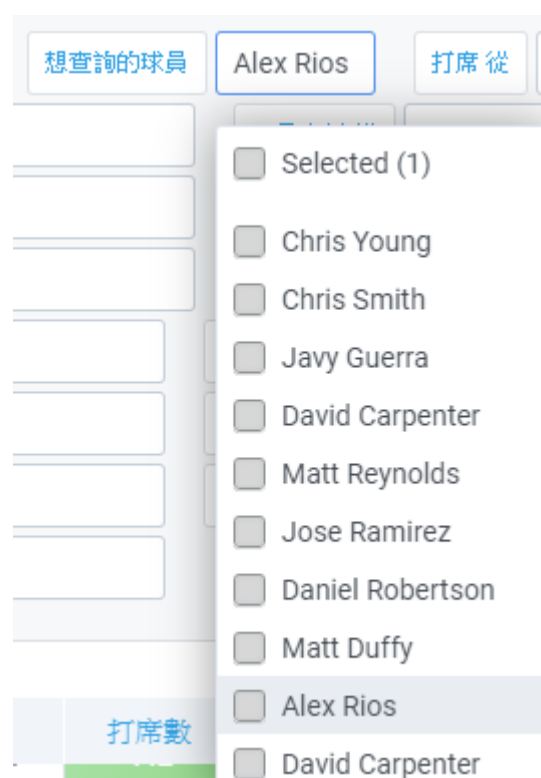
年度別隊伍得分數與勝率

年度按隊連續離隊出場次數

其他區會以適合的方式進行視覺化，如勝場數排行、得分比較等。
而主畫面上方是提供使用者縮小範圍的輸入欄。



這些數值都是以變數寫入上述第二次 query 內，會根據使用者輸入的數值不同，進行相對應的 query，例如使用者可以輸入球員姓名，輸入欄相對應的球員就會被反白出來，避免使用者輸入的球員不在我們的資料集中或是打錯字，算是一種防呆機制。



選取完之後，系統會將使用者選取的數值代入上述說的第二次 Query 中相對應變數的地方，然後刷新 Table，如下圖，可以見到原本很多 row 的 table，只剩下選取的球員的數據，其他不符條件的數據都會被屏除，應用上可以讓使用者查詢自己關注的球員、年分、數值排行等，而不會因龐大的數

據量而分散注意力。

想查詢的年份

All

想查詢的球員

Alex Rodriguez + Aramis Ramirez

打席 從

0

到

800

打數 從

0

席 到

700

一壘安打 從

0

支 到

170

二壘安打 從

0

支 到

56

三壘安打 從

0

支 到

15

全壘打數 從

0

轟 到

52

被三振數 從

0

到

216

被四壞球保送數 從

0

到

143

被故意四壞球保送數 從

0

到

29

打擊率 從

0

到

0.3484

上壘率 從

0

到

0.4527

長打率 從

0

到

0.6898

攻擊指數 從

0

到

1.1

場內安打率 從

0

到

0.5472

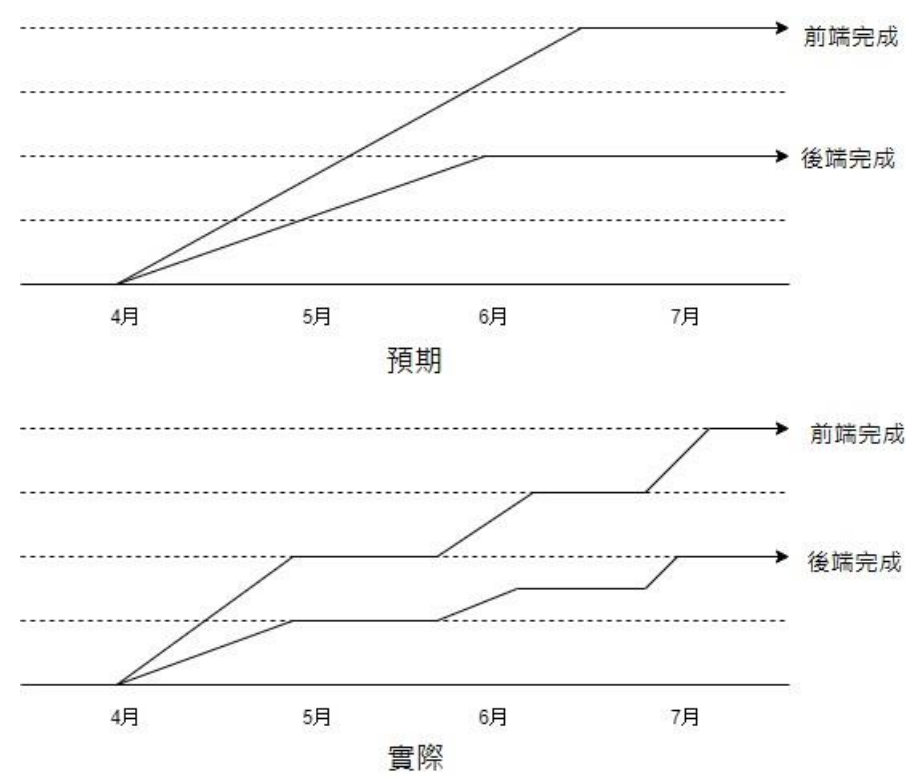
打者年度資料

年度	選手	打席數	打數	一壘安打數	二壘安打數	三壘安打數	全壘打數	三振數	四壞球數	故意四壞球數	打擊率	上壘率	長打率	整體攻擊指數	場內安打率
2015	Alex Rodriguez	622	525	75	22	1	33	145	84	5	0.2495	0.3553	0.4838	0.8311	0.3670
2015	Aramis Ramirez	517	476	68	31	1	17	68	31	3	0.2458	0.2959	0.4232	0.7114	0.3650

欄位中都已經預先手動填入所有數值的 max 和 min，因此使用者可以知道數值的分布範圍，可以避免填入偏離資料集的數值而查不到相對應的數據。

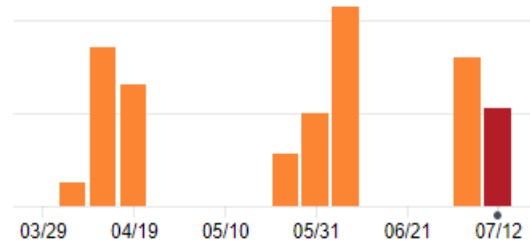
Others

Progress



我們在 4/21 訂出題目並找好 source，4/23 建好基礎的 table
計畫在 6 月中前完成後端開發並與前端連結，6 月底前就能完成 project

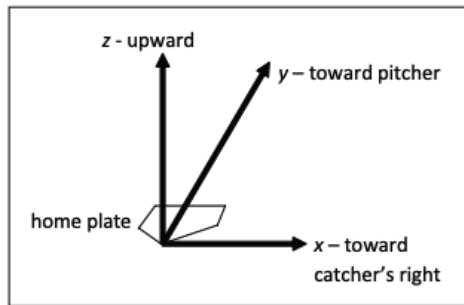
但實際上因段考期間大家都很忙，中途斷斷續續的進度停滯了一段時間
最後收尾及驗收的部分到了 7 月才漸漸完成



↑從 Github 的 commit 數量也能看得出來

Problems and Solves

1. 命名問題：因為 batter 跟 pitcher 的 query 是不同人寫的，命名方式有些不同
→ 後續要不斷確認這個 column 的名稱，而且不能亂改。
2. Create table 速度：太多 table join 在一起會讓 query 速度極慢，可能吃完晚餐都還沒好
→ 這時用 `alter table /*table_name*/ add primary key(/*column_name*/);` 就能加快非常多速度。
3. 在 group by 時，如果先篩選出特定 row，比如說某打者打一壘安打的 row，先篩選，再對打者 id 做 group by 並 count(*), 會沒辦法顯示一壘安打是 0 的打者 id，因為已經被篩掉了
→ 這時可用 `if(event="Single",1,0) as Single` 的方式，group by 時再用 `sum(Single)`，來正確顯示出 0 支 1 壘安打的打者 id。
或者先把所有打者 id 列出來後與一壘安打 left join，再用 ifnull 把 count 換成 0
4. 比較 ratio 時可能會有 data 筆數過少導致數值過度異常
→ 把 data 筆數也加入條件篩選
5. 原始資料意義不明：pitcher 這個 table 紀錄了每球的很多數據，但是某 column 定義不明，導致我們想要查縱向位移量時，不知道要用哪個 column
→ 我們找到了這篇論文，下面兩張圖是 column 的單位及定義 (<http://baseball.physics.illinois.edu/KaganPitchfx.pdf>)



No.	Quantity	Value	Units	Descriptions
1	des	In play, run(s)		A comment on the action resulting from the pitch.
2	type	X		B=ball, S=strike, X=in play
3	id	371		Code indicating pitch number
4	x =	112.45	pixels	x-pixel at home plate
5	y =	131.24	pixels	z-pixel at home plate (yes, it is z)
6	start_speed	84.1	mph	Speed at y0=50ft
7	end_speed	77.2	mph	Speed at the front of home plate y=1.417ft
8	sz_top	3.836	ft	The z-value of the top of the strike zone as estimated by a technician
9	sz_bot	1.79	ft	The z-value of the bottom of the strike zone as estimated by a technician
10	px_x	8.68	in	A measure of the "break" of the pitch in the x-direction.
11	px_z	9.55	in	A measure of the "break" of the pitch in the y-direction.
12	px	-0.012	ft	Measured x-value of position at the front of home plate (y = 1.417 ft)
13	pz	2.743	ft	Measured z-value of position at the front of home plate (y = 1.417 ft)
14	x0	1.664	ft	Least squares fit (LSF) value for the x-position at y = 50 ft
15	y0	50.0	ft	Arbitrary fixed initial y-value
16	z0	6.597	ft	LSF value for the z-position at y = 50 ft
17	vx0	-6.791	ft/s	LSF value for the x-velocity at y = 50 ft
18	vy0	-123.055	ft/s	LSF value for the y-velocity at y = 50 ft
19	vz0	-5.721	ft/s	LSF value for the z-velocity at y = 50 ft
20	ax	13.233	ft/s/s	LSF value for the x-acceleration assumed constant throughout the pitch.
21	ay	25.802	ft/s/s	LSF value for the y-acceleration assumed constant throughout the pitch.
22	az	-17.540	ft/s/s	LSF value for the z-acceleration assumed constant throughout the pitch.
23	break_y	25.2	ft	Another measure of the "break." See Nathan's website for an explanation.
24	break_angle	-32.1	deg	Another measure of the "break." See Nathan's website for an explanation.
25	break_length	5.9	in	Another measure of the "break." See Nathan's website for an explanation.

6. 無法確認 sql 的結果是否正確
→ 與其他現有網站交叉比對
7. data 豐富度比不上其他網站
→ 分析其他有趣的資料(如驅逐出場次數等)

Contribution

陳煜盛：

唯一棒球迷，需要解釋術語、傳統、負責計劃研究方向、指派工作。

負責 `batter_*.sql` 的編寫及 `pitcher_*_per_game` 的部分。

整合 `proposal` PPT 與口頭報告所有後端部分。

洪瑋廷：

唯一前端負責人，負責 `Grafana` 的排版設計、`query` 資料根據使用者需求進行二次 `query` 並視覺化。

維護架設 `MySQL` 和 `Grafana` 的 `Server`。

`Proposal` 負責口頭報告及展示 `Grafana` 前端頁面。

影片拍攝。

報告中 `application` 相關部分的撰寫。

李嘉盛：

`Teams_*` 及其他類 `query` 的撰寫

完成陳煜盛指派的任務

報告其餘部分的撰寫。

王昶淵：

`Pitcher_*.sql` 的撰寫

負責完成陳煜盛指派的任務

研究 `pitches column` 的物理意義。

Repository and Discussion Channel

Github: https://github.com/brianchennn/Database_Team_Project

HackMD: https://hackmd.io/GFo9xYfMRNCeQgr_60Z1gg?view