**2020 DB Term Project Report 02組**

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

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

// pitcher\_four\_seam.sql和pitcher\_two\_seam.sql

因跟pitcher\_FF.sql、pitcher\_FT.sql的內容重疊所以沒加進來

// ejection\_info\_player.sql和ejection\_info\_team.sql

是用來查詢使用者指定data的sql，所以沒加進來

// 所有球種球速分析等合起來後再加入

/\*\*\*剩下部分

database： database的maintain(應該不會再有其他更新?)、如何連結database和application(瑋廷救我)

application：application的interface和function及實踐方法(瑋廷救我)

other：project的進度表(預期和實際狀況)、遭遇的問題\*\*\*/

# Data

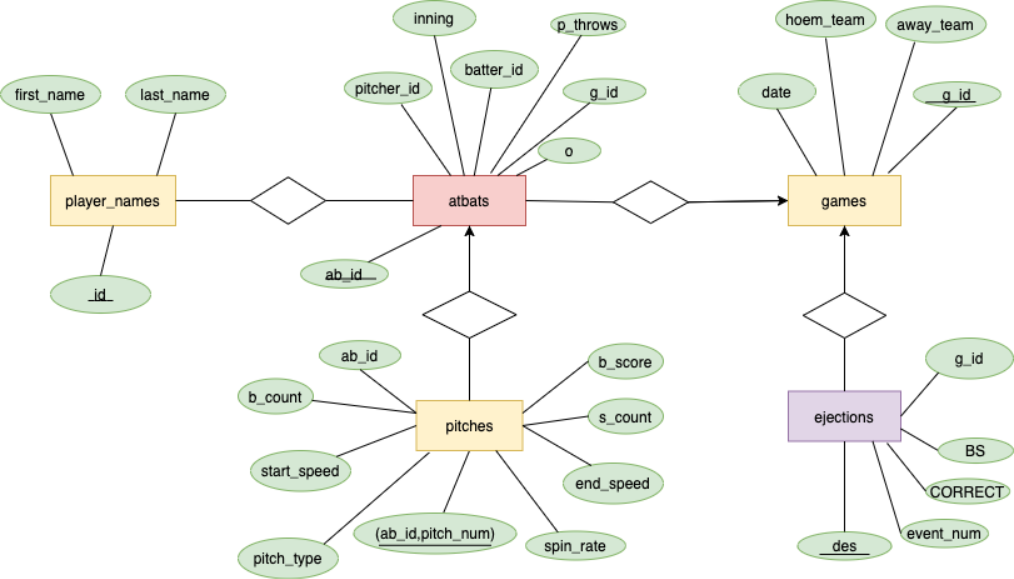
### ****Source****

[Kaggle Pitch Data 2015-2018](https://www.kaggle.com/pschale/mlb-pitch-data-20152018#ejections.csv)

### ****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](http://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****



以上僅列出重要的attribute，因為實在太多了。

### Other Tables Extend From The Tables Above

### 命名方式分成batter\_\*, pitcher\_\*, team\_\*, other,其中有些後面有per\_game表示這是用單場資料group by,後面有per\_years表示是年度資料做group by。

### 若有create\_table字樣，表示此table為其他已寫出來的小table組合而成。

* **batter\_babip** (每年度各打者的BABIP)
  + **year** 年度
  + **first\_name** first name
  + **last\_name** last name
  + **hits** 安打數
  + **home\_run** 全壘打數
  + **flyout** 飛球出局數
  + **groundout** 滾球出局數
  + **BABIP** 場內安打率
* **batter\_create\_table** (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
  + **H** 安打及全壘打次數
  + **1B** 一壘安打數
  + **2B** 二壘安打數
  + **3B** 三壘安打數
  + **HR** 全壘打數
  + **AVG** 打擊率
  + **OBP** 上壘率
  + **OPS** 整體攻擊指數
  + **K** 三振次數
  + **K9** 三振率
  + **BB** 保送次數
  + **BB9** 保送率
* **batter\_homerun**(每年度所有打者的全壘打排行榜，其中打者的打擊率須超過50)
  + **Year** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **event** result of the atbat
  + **cnt** 全壘打數
* **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
* **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
* **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
* **pitcher\_create\_table2** (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
  + **Double** 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
  + **G/F** 滾飛比(滾地球跟飛球的比例)
  + **FF** 球種為FF的數量
  + **CH** 球種為CH的數量
  + **EP** 球種為EP的數量
  + **FC** 球種為FC的數量
  + **CU** 球種為CU的數量
  + **PO** 球種為PO的數量
  + **FS** 球種為FS的數量
  + **FT** 球種為FT的數量
  + **IN** 球種為IN的數量
  + **KC** 球種為KC的數量
  + **KN** 球種為KN的數量
  + **SC** 球種為SC的數量
  + **SI** 球種為SI的數量
  + **SL** 球種為SL的數量
  + **UN** 球種為UN的數量
* **ff\_velocity\_avg\_years** (每年度各選手的直球均速)
  + **years** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **avg\_ff\_velocity** 直球均速
* **ff\_velocity\_max\_years** (每年度各選手的直球極速)
  + **years** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **max\_ff\_velocity** 直球極速
* **pitcher\_babip** (每年度各投手的BABIP)
  + **years** 年度
  + **first\_name** first name
  + **last\_name** last name
  + **hits** 安打數
  + **home\_run** 全壘打數
  + **flyout** 飛球出局數
  + **groundout** 滾球出局數
  + **BABIP** 場內被安打率
* **pitch\_type\_ratio**(每年度各選手的各球種使用比例)
  + **year** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **pitch\_type** 球種
  + **ratio** 使用比例
* **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
* **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
* **payoff\_pitch\_ratio** (每年度, 2好球後各投手的球種使用比例及當下的好壞球數)
  + **year** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **ball\_strike** 好壞球數
  + **pitch\_type** 球種
  + **ratio** 使用比例
* **pitcher\_pitch\_number** (每年各投手投球總數)
  + **years** 年度
  + **first\_name** first name
  + **last\_name** last name
  + **pitch\_num** 投球數
* **pitcher\_spin\_strike\_rate** (轉速與好球率的關係)
  + **year** 年度
  + **first\_name** first name
  + **last\_name** last name
  + **pitcher\_throws** 投手站位
  + **batter\_stand** 打者站位
  + **pitch\_type** 球種
  + **Spin** 轉速
  + **strike\_ratio** 好球率
* **pitcher\_take\_strike\_rate** (每年度各投手的奪好球率)
  + **year** 年度
  + **first\_name** first name
  + **last\_name** last name
  + **pitch\_type** 球種
  + **ratio** 奪好球率(swing\_strike + call\_strike + 非兩好球時的界外)/某球種總球數
  + **total** 總球數
* **strike\_ratio\_for\_every\_pitch\_type** (每年度各投手各球種的使用比例及當下的好壞球數)
  + **year** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **ball\_strike** 好壞球比數
  + **pitch\_type** 球種
  + **ratio** 使用比例
* **pitcher\_strkiout\_ratio\_for\_every\_pitch** (每年度各投手所有的三振中各球種使用比例)
  + **year** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **ball\_strike** 好壞球比數
  + **pitch\_type** 球種
  + **ratio** 使用比例
* **batter\_**(與每年各打者保送次數及保送率)
  + **Year** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **BB** 保送次數
  + **PA** 打數
  + **BB\_rate** 保送率
* **pitcher\_hit\_avg** (每年度各投手的被打擊率)
  + **years** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **hits** 安打數
  + **home\_run** 全壘打數
  + **atbat\_or\_other** 總球數
  + **AVG** 打擊率
* **pitcher\_whip** (每年度各投手的每局被上壘率)
  + **years** 年度
  + **first\_name** first name
  + **last\_name** last name
  + **walk** 保送次數
  + **hits** 安打數
  + **total\_innings** 總投球局數
  + **WHIP** 每局被上壘率
* **pitcher\_fip** (每年度各投手的FIP=被全壘打數x13+死球x3-三振x2)/投球局數)
  + **years** 年度
  + **id** 對應player\_id
  + **first\_name** first name
  + **last\_name** last name
  + **home\_run** 全壘打數
  + **walk** 保送次數
  + **hit\_by\_pitch** 觸身球數
  + **strikeout** 三振數
  + **IP** 投球局數
  + **fip\_const** 計算FIP用的常數
  + **FIP** FIP
* **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

### ****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cerate\_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。

### ****How Do We Connect Our Database To Our Application****

# Application

### ****Interface****

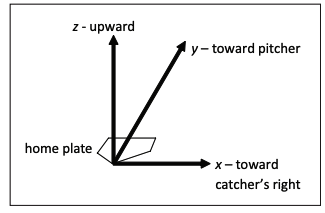
### ****Function****

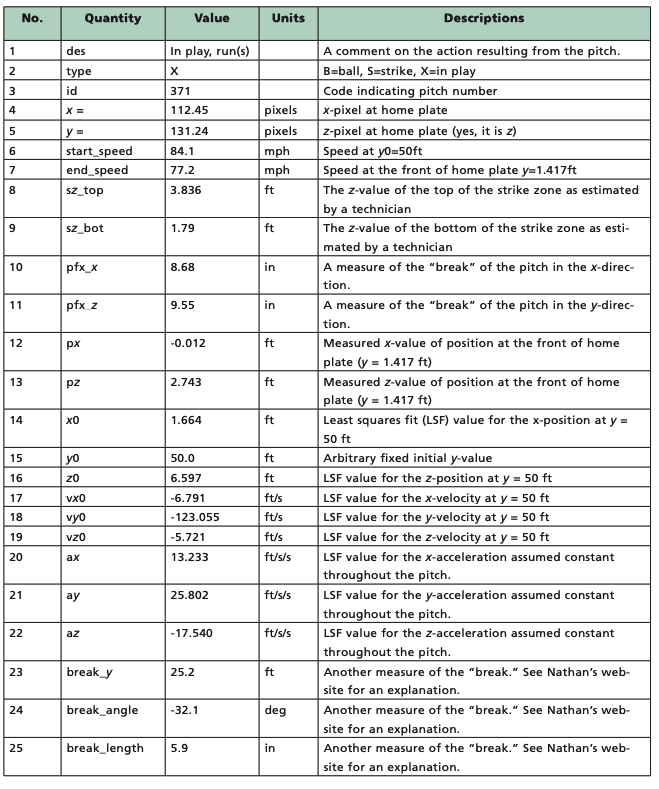
# Others

### ****Progress****

### ****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。所以造成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。這時必須要用if(event=”Single”,1,0) as Single的方式，group by時再用sum(Single)，才能正確顯示出0支1壘安打的打者id。
4. 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>)



但是y,z方向的定義很亂。目前正由組員王昶淵研究中。

### ****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，將新建的table做視覺化呈現。

Proposal負責口頭報告及展示Grafana前端頁面。

影片拍攝。

李嘉盛：

Teams\_\*及其他類query的撰寫，完成陳煜盛指派的任務，報告的撰寫。

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。