Introduction to Database Systems

Individual Homework 1: SQL tasks in MySQL

1. Introduction

In this homework you need to practice some basic usages of MySQL, including creating databases, creating tables, loading csv files, and using MySQL command to find the answer of tasks. After this homework, you will be capable of querying and analyzing your data by MySQL from zero to one.

There is a dataset called MLB game data in this homework. It contains data from the start of the 2016 season through the end of the 2021 postseason. You will need to create the database by the data and implement some queries for the questions in task C. The data is modified from Kaggle Dataset(feel free to google them).

For the first dataset, you need to create a database based on our setting, and load the csv files into your created database. There are 12 questions you need to solve by SQL. Among the questions, there are 6 advanced problems which you might need to spend more time on. Read the following content for more details.

2. Tasks

A. Create Tables

First, download the MLB game data from here.

You can refer to this page to see the meaning of the columns.

Then you should create tables based on the following setting. Notice that you **must** make the detail of your tables the same as our description, including 'table name', 'attribute name', 'attribute type', 'primary key', 'foreign key', 'null'.

Please **paste the screenshot of your tables by using the `describe` command to your report**, it will take 5% of your grades in this homework.

Table Name	Attribute Name	Type	Primary Key	Foreign Key	NULL
games	Game	int	YES		NO
	away	char(3)			NO
	home	char(3)			NO
	away_score	tinyint			
	home_score	tinyint			

	Date	datetime			NO
inning	Game	int	YES	games (Game)	NO
	Inning	char(3)	YES		NO
	Runs	tinyint			
	Hits	tinyint			
	Errors	tinyint			
hitters	Game	int	YES	games (Game)	NO
	Team	char(3)			NO
	AB	tinyint			
	R	tinyint			
	H	tinyint			
	RBI	tinyint			
	BB	tinyint			
	K	tinyint			
	num_P	tinyint			
	Position	varchar(20)			
	Hitter_Id	mediumint	YES		NO
pitchers	Game	int	YES	games (Game)	NO
	Team	char(3)			NO
	IP	float			
	H	tinyint			
	R	tinyint			
	ER	tinyint			
	BB	tinyint			
	K	tinyint			
	HR	tinyint			

	PC_ST	varchar(10)			
	Pitcher_Id	mediumint	YES		NO
pitches	Pitch_Id	mediumint	YES		NO
	Game	int		games (Game)	NO
	EventId	smallint			NO
	Num	tinyint			NO
	Inning	char(3)			
	Pitcher	varchar(35)			
	Pitch	varchar(50)			
	_Type	varchar(20)			
	MPH	smallint			
players	Id	mediumint	YES		NO
	Name	varchar(20)			

Please **answer the following question in your report**, it will take 10% of this homework.

1. (2%) What is the difference between type “char” and type “varchar”?
   譯：變數型態 “char” 和 “varchar” 有什麼不同？
2. (2%)What is the purpose of setting “Foreign Key”, and trying to explain it with the tables, “games” and “inning”?
   譯:“Foreign Key”的設置目的為何，試著用“games”以及“inning”兩張表格解釋，“Foreign Key”的設置可能會帶來什麼限制？
3. (3%) How many bytes should it take for “tinyint”, “smallint”, “mediumint”, “int”? (e.g. 8 bytes for “bigint”)
   And what’s the range they can express? (e.g. from -1000 to 1000)
   譯：“tinyint”, “smallint”, “mediumint”, “int” 各需要多少bytes來儲存？(e.g. 8 bytes for “bigint”)
   還有他們的表示範圍可以從哪裡到哪裡？(e.g. from -1000 to 1000)
4. (3%) What do you think about this DB schema? If you can change this table architecture, how would you modify it and why?
   譯：你對這資料庫架構有什麼想法？如果你可以修改這架構，你會怎麼改？為什麼？

B. Load CSV Data

After creating the database, you need to load the downloaded csv files into your database.

Here we don't restrict the method you use, but you have to check the data is loaded successfully by yourself. The following number is the data records for each table.

Table Name	# of Data Records
games	13413
inning	238096
hitters	334336
pitchers	116551
pitches	3879438
players	3341

C. Query Tasks

In this part, here are 12 query tasks you need to write. **Please read the following rules carefully**. You are only allowed to use **one** query (**one delimiter**) to find the answer, and you **don't** have to explain your SQL. Note that the **column names and the order** of your query answers should be **the same as our examples**.

For **task 11 and 12**, take **screenshots** of your queries, and **write your analysis** into the report. Try to explain what your queries are doing, why you write these queries, what's the meaning of the result, what's your conclusion, etc.

For homework submission, please write every query task into a single `sql` file, named as "1.sql", "2.sql", etc.

1. (5%) How many games ended with a score difference of 10 points or more?
   譯：請找出有多少場比賽，其比賽結束時分差達到10分以上(包含10分)？
   Hint: use the "count" function

cnt
558

2. (5%) Find the games with the most innings spent in this dataset, list its game id and the number of innings used, and arrange them in ascending order of game id? (The top half inning is viewed as one inning, e.g. if the game ends early in top 9 innings, it will be regarded as nine innings. The innings should be showed in integer format)

譯：請找出這份資料中所花費局數最多的比賽，列出其比賽編號與使用局數並依照比賽編號順序由小到大排列？ (只有上半場仍算做一局，例如：9局上半提前結束視為進行九局，請用整數表示局數)

Hint: use the “order by” function

Game	num_innings
360701114	19
370905102	19

3. (5%) Please find the top three pitchers with the longest total pitching innings and their pitch id as well as total innings during the 2021 season (4/1-11/30). List out in decreasing order of the total innings. Round the probability to the first decimal place.

譯：請找出2021年賽計期間(4/1～11/30)前三名總投球局數最長的投手與他們的總投球局數，以出場總局數由高到低做排序，四捨五入到小數點後1位。

Hint: use the “between”, ”group by” function

Pitcher_Id	Pitcher	tol_innings
39251	W. Buehler	222.5
5403	A. Wainwright	209.8

4. (5%) Find the top three players consuming the highest average number of pitches per plate appearance(P/PA), regardless of the data that the player with zero plate appearance in a game as well as the player with the total plate appearance less than 20 in the dataset. Show the hitter's name, average of P/PA, average of hits, average of base on balls, average of strikeouts, and total plate appearance . Please output in descending order of the average of P/PA. Round the probability to the fourth decimal place.

(The plate appearance: BB + K + AB)

(e.g. The P/PA of a hitter in two games are respectively 5 balls and 4 balls per game, and his average number of P/PA is 4.5)

譯：請找出前三名平均單場每打席消耗球數最高的選手(P/PA)，不考慮當場比賽中打席數為零的資料與總打席數小於20的選手，並且呈現其打者名稱、平均單場每打席消耗球數、平均打數、平均保送數、三振數。 請依照平均單場P/PA由高到低輸出，並且四捨五入到小數後第四位。

(打席數算法請用保送次數+三振次數+打數)

(舉例：某球員在兩場比賽中每打席消耗球數分別為5球與4球，其平均每打席消耗球數為4.5)

Hitter	avg_P/PA	avg_AB	avg_BB	avg_K	tol_PA
E. Butler	4.2101	1.3913	0.0870	0.3478	42
C. Quantrill	4.1852	1.5000	0.0556	0.3333	34

5. (5%) Find the shortest time interval between the games of each team in the month with the most games. Please list the team, month, and time interval. (The format of the time interval should be “hh:mm”)
譯：請問總比賽場次最多的月份中，各個隊伍最快多久有一場比賽，請列出隊伍名稱、月份、時間間隔。 (當月月份輸出格式為"yyyy-mm"、時間間隔輸出格式為"hh:mm")
Hint: use the “lag” function, the “over” clause with “partition by” function

Team	The_month	time_interval
ARI	2017-08	17:00
ATL	2017-08	03:25

6. (5%) Find the types of balls whose speed cannot exceed 95 (MPH). Please output according to the dictionary order of the types of balls.
譯：請找出球速無法超過95(MPH)的球種，請依球種字典順序輸出。
Hint: use the “distinct”, “not in” function

Type
Eephus Pitch
Forkball

7. (10%) Find out the players with the highest hit rate average per game among the five teams with the highest winning rate in 2021 (only consider players with a total batting over 100 in 2021 and not consider the data that players in the game with zero batting). Please list the team’s name, the hitter’s name, hit rate average per game, the number of total hits, and the winning rate of the team to which they belong, and rank in descending order of the winning rate.(Round the probability to the fourth decimal place.) (hit rate: H / AB)
譯：請找出2021年間勝率最高的五隻球隊中，其平均每場比賽打擊率最高的選手（只考慮當年度總打數超過100的選手且不考慮當場比賽中打數為零的選手資料）列出球隊名稱、打者名稱、平均每場比賽打擊率、總打數以及其所屬隊伍的勝率並且依照勝率由高到低排序。（四捨五入至小數點後第4位）（打擊率：H / AB）

Team	Hitter	avg_hit_rate	tol_hit	win_rate
SF	B. Crawford	0.2940	503	0.6527
LAD	T. Turner	0.3167	258	0.6437

8. (10%) Find out the average strikeout rate, PC-ST value per game between traded and non-traded pitchers from 2020 to 2021 (only consider pitchers with more than 50 total innings in the two years), please list the average strikeout rate per game in 2020, the average strikeout rate per game in 2021, the average PC-ST value in 2020, and the average PC-ST value in 2020 (Round the average PC-ST value to the fourth decimal place) (For the strikeout rate calculation, please use the number of 9*K/IP)
譯：請找出2020到2021年間，有被交易過與沒被交易過的投手之間平均單場比賽的三振率、投球數與好球數（只考慮兩年內總投球局數超過50局的投手），請列出兩種投手的數量、2020年間平均每場比賽三振率、2021年間平均每場比賽三振率、2020年間平均每場比賽"總投球數-好球數"、2021年間平均每場比賽"總投球數-好球數"（須四捨五入至小數點第四位並按照範例中的格式）(三振率算法請用 9 * 三振數 / 投球局數).

Pitcher	cnt	2020_avg_K/9	2021_avg_K/9	2020_PC-ST	2020_PC-ST
Changed	133	12.8063	12.1091	40.8364-25.6936	42.0674-26.9533
Unchanged	221	11.9625	11.8112	53.1891-33.6151	52.7626-33.9636

9. (10%) During the 2021 season, we wonder what the least difference of the hit rate between the two teams is, so that the probability that the team can win is over 95%? (The difference of the hit rate should be rounded down to two decimal places after the minus calculation. The win rate should be rounded to four decimal places)
(e.g. In two matches, A and B, The difference of the hit rate in A is 12% and the team with higher hit rate wins the game. The difference of the hit rate in B is 0.10 and the team with higher hit rate loses the game. Then, when the hit rate difference over 12% the win probability is 100%)
(e.g. In two matches, A and B, The difference of the hit rate in A is 12% and the team with higher hit rate wins the game. The difference of the hit rate in B is 0.12 and the team with higher hit rate loses the game Then, when the hit rate difference over 12% the win probability is 50%)
譯：請找出2021年間當某場比賽中隊伍打擊率比對方最少高出多少時，則此隊伍的獲勝概率超過九成伍，請列出此最小打擊率差值與在此打擊率差

下的隊伍勝率勝率？（打擊率差請在計算完兩隊差值後無條件少捨去到小數點後第二位，勝率請四捨五入至小數後四位）

（例：A, B兩場賽事中，A賽事中兩隊打擊率差為0.12且打擊率較高的那隊獲勝，B賽事中兩隊打擊率差為0.10且打擊率較高的那隊輸球，輸出結果為當最小打擊率差為0.12時，隊伍勝率為100%。）

（例：A, B兩場賽事中，A賽事中兩隊打擊率差為0.12且打擊率較高的那隊獲勝，B賽事中兩隊打擊率差為0.12且打擊率較高的那隊輸球，輸出結果為當最小打擊率差為0.12時，隊伍勝率為50%。）

Here is the sample of win rate over 90%

hit_rate_diff	win_rate
0.09	0.9089

10. (15%) Find the top five players who most frequently win the title of winning-pitcher in 2021, as well as their frequency, and rank them from high to low according to the frequency (winning-pitcher: only consider the starting-pitchers who have thrown 5 innings or more, the score is higher than the opponent when he is off the field, and the game is never tied until his team win) (starting-pitchers: The pitcher pitches the first of the ball in his team.)
找出最常獲得勝投投手稱號的前五名球員，以及其勝投率，並依照勝投率由高排到低（勝投：只考慮投滿5局以上的先發投手，在其離場的下半局結束後比分贏過對手，並且下場後不曾被追平或落後直至比賽獲勝）（先發投手：該隊伍投出第一球的選手即為此隊伍之先發投手）

Hint: there might be some problems you meet in this question

1. Due to the first name only in pitches data, you need to **first find the starting-pitchers, then using the “like” function to match the pitcher’s name between pitches data and pitchers data**. please use “pitches.Pitcher like concat('% ', first_pitcher.first_pitcher_name, '%')”.

譯：由於在pitches的資料中只使用選手的部份名字，你需要**先找出先發投手的名單，再利用"like"函式去找到pitchers資料中相對應的球員資料**。 請使用以下搜尋條件“pitches.Pitcher like concat('% ', first_pitcher.first_pitcher_name, '%')”

2. Due to the first name only in pitches data, there might be two pitchers viewed as starting-pitchers in a team in your query. Just directly view both of them as the starting-pitchers in following computing.

譯：由於在pitches的資料中只使用選手的部份名字，你可能會在先發投手的資料中發現同一隊伍中有兩人的稱呼相同。請直接將兩者視為先發投手進

行後面的計算。

game	Inning	first_pitcher	Game	Team	Pitcher
401228971	T1	Keller	401228971	PIT	M. Keller
401228971	T1	Keller	401228971	PIT	K. Keller
401228971	B1	Mikolas	401228971	STL	M. Mikolas

3. Remember the home team defends first. Therefore, the pitcher in the home team should only exist in the rows with Inning equaling “T1”. Two pitchers with the same first name in the two teams viewed as starting-pitchers should not happen in your query.

譯：請注意在棒球規則中，主場球隊會先防守，因此在”T1”時的投手應該屬於主場球隊，在這次的資料中不應該出現以下情況。

game	Inning	first_pitcher	Game	Team	Pitcher
401229321	B1	Rogers	401229321	WSH	J. Rogers
401229321	B1	Rogers	401229321	MIA	T. Rogers
401229321	T1	Rogers	401229321	WSH	J. Rogers
401229321	T1	Rogers	401229321	MIA	T. Rogers

Pitcher	winning-pitcher
L. Castillo	0.3491
C. Archer	0.3366
T. Roark	0.3333

11. (10%) Does the home team with home advantage have much more opportunity to win the game, or the team with a higher average score of the whole team’s players? (The score can be the evaluation of the pitch score or the hit score of the team, or the error of the teams.) **Answer your own view with one SQL query.** (The method you evaluate the team score, and the correlation between the method and the wins will be used as the basis for scoring)

比賽通常都會有所謂的「主場優勢」，但是主場優勢也只是優勢，不能保證為隊伍帶來勝利。你認為擁有主場優勢的隊伍比較容易贏，還是隊伍依照「其最後一次測量attribute」的平均程度較高的隊伍比較容易贏？（此題為開放式答案，**請用一個SQL找出的結果闡述你的觀點並解釋你所使用的方法**。程度可以是隊伍投手、打擊者的近期表現或者隊伍近期的失誤情況等等。）

（此題的將會考量到所使用的評價方法，統計結果與隊伍獲勝之間的關聯性強弱作為評分依據。）

12. (10%) If your team wants to trade a relief pitcher today (the relief pitchers that can be considered here are players who have served no more than 10 starts in the data), I would like to ask you, which players will be your first choice. **You can answer by your own view with multiple SQL queries. Focus on observation to the dataset and explain your analysis.**
    若今天你的球隊想交易一位後援投手（這裡可以考慮的後援投手為在資料中擔任先發次數不超過10場的選手），想請問你，哪幾位選手會是你的首選，請利用SQL分析這份資料並且解釋（此題為開放式答案，**可以使用多個SQL，重點請著重在對資料的分析與發想**）

3. Grading

TA will run your ".sql" file automatically. The environment will be **Ubuntu 22.04** and **MySQL 8.0.23**, which is the same as the environment you had prepared in Hw0. Make sure your code can run in this environment correctly.
Following is the grading policy.

Plagiarism is not allowed! You will get a huge **penalty** if we find that.

Description	Score(%)
Part1 table screenshots	5
Question answering	10 (bonus)
SQL tasks	95
Total	100 + 10

4. Discussion

TAs had opened a channel **HW1 討論區** on Teams of the course, you can post questions about the homework on the channel. TAs will answer questions as soon as possible.

Discussion rules:

1. Do not ask for the answer to the homework.
2. Check if someone has asked the same question before asking.
3. We encourage you to answer other students' questions, but again, do not give the answer to the homework. Reply the messages to answer questions.
4. Since we have this discussion forum, do not send email to ask questions about the homework unless the questions are personal and you do not want to ask publicly.

5. Submission

1. The deadline of this homework is **3/31 (Fri.) 23:55:00**.
2. You should put your `pdf` and `sql` files into one folder, each should be named as "HW1_XXXXXXX.pdf", "1.sql", "2.sql" …. And the folder should be named as "HW1_XXXXXXX" where XXXXXXX is your student ID.

   Then compress your folder into one `zip` file. Submit it to the New E3 System with the format **HW1_XXXXXXX.zip** where XXXXXXX is your student ID.

   We **only accept one zip file**, wrong format or naming format cause -10 points to your score (after considering late submission penalty).
3. Late submission lead to score of (original score)*0.85^{days}, for example, if you submit your homework right after the deadline, you'll get (original score)*0.85 points.
4. If there is anything you are not sure about submission, ask in the discussion forum.