Project 2

COP 4710, Summer 2021

Due July 9, 2021

1 Overview

In this project, you will develop SQL queries to answer several questions about a baseball dataset. This dataset has been provided to you as 4 CSV files: players.csv, teams.csv, batting.csv, and pitching.csv. You can find more details about the project data in section 2.

Data for this project were provided by `www.seanlahman.com` and represent all 221,455 Major League Baseball games played among all 2,935 teams and 20,200 players across the 1871 to 2020 seasons. The data for these tables were downloaded from `seanlahman.com`. Some of the attributes in the original files were removed to simplify table creation somewhat.

2 Data

Each of the 4 CSV files provided stores data for one SQL table. The first line of each file gives the table attributes, separated by commas, and each row describes one record. `NULL` values are indicated by an empty string (i.e., consecutive commas). The attributes for each file are described in the tables below.

Teams

Attribute	Description
Year	Year for this season
TeamID	Unique ID for the team
Rank	Overall ranking for the team in the given season
G	Number of games played
W	Number of games won
L	Number of games lost
LgWin	Whether this team won their league (Y/N)
WSWin	Whether this team won the World Series (Y/N, blank before 1903)
Name	Name of the team

Players

Attribute	Description
PlayerID	Unique ID for this player
BirthYear	Year they were born
BirthMonth	Month they were born (1–12)
BirthDay	Day of the month they were born (1–31)
Fname	First name
Lname	Last name
Weight	Weight in pounds
Height	Height in inches
Debut	Date of first game played
FinalGame	Date of last game played

Batting

Attribute	Description
PlayerID	Unique ID for the player
Year	year for this season
Stint	each time a player was traded during the season, the data records a separate stint. A player can have multiple stints on the same team in the same season (if they were traded back)
TeamID	Unique ID for the team they played for
G	Games played
AB	Number of times at bat
R	Number of runs player made
H	Number of times player hit the ball and advanced at least one base
B2	Number of doubles, where the player hit the ball and advanced 2 bases
B3	Number of triples, where the player hit the ball and advanced 3 bases
HR	Home runs, where the player hit the ball and rounded all 4 bases
RBI	Runs batted in: how many players reached home because of a hit by this player (other players + home runs)
SB	How many bases this player stole
SO	How many times this player struck out

Pitching

Attribute	Description
Year	year for this season
Stint	each time a player was traded during the season, the data records a separate stint. A player can have multiple stints on the same team in the same season (if they were traded back).
TeamID	Unique ID for the team they played for
G	Games played
SHO	Shutouts: games where no one on the opposing team hit the ball
IPOuts	Number of batters this pitcher struck out (innings played times 3)
H	Number of times a batter hit
SO	Number of times a batter struck out
ERA	Earned run average: number of runs made by opposing players due to hits, times 9, and divided by innings played (common measure of pitcher quality; lower is better)

After installing PostgreSQL, you can create the relevant tables and import the data into ProgreSQL using the SQL commands below:

```
CREATE TABLE Teams(Year INTEGER, TeamID CHAR(3), Rank INTEGER, G INTEGER, W INTEGER,
L INTEGER, LgWin CHAR(1), WSWin CHAR(1), Name VARCHAR(35), CONSTRAINT team_pk
PRIMARY KEY(Year, TeamID));

CREATE TABLE Players(PlayerID CHAR(9) PRIMARY KEY, BirthYear INTEGER, BirthMonth
INTEGER, BirthDay INTEGER, FName VARCHAR(20), LName VARCHAR(20), Weight INTEGER,
Height INTEGER, Debut DATE, FinalGame DATE);

CREATE TABLE Batting(PlayerID CHAR(9) REFERENCES Players, Year INTEGER, Stint
INTEGER, TeamID CHAR(3), G INTEGER, AB INTEGER, R INTEGER, H INTEGER, B2 INTEGER,
B3 INTEGER, HR INTEGER, RBI INTEGER, SB INTEGER, SO INTEGER, CONSTRAINT bat_pk
PRIMARY KEY(PlayerID, Year, Stint), CONSTRAINT bat_fk_team FOREIGN KEY (Year,
TeamID) REFERENCES Teams);

CREATE TABLE Pitching(PlayerID CHAR(9) REFERENCES Players, Year INTEGER, Stint
INTEGER, TeamID CHAR(3), G INTEGER, SHO INTEGER, IPOuts INTEGER, H INTEGER, SO
INTEGER, ERA NUMERIC(5,2), CONSTRAINT pitch_pk PRIMARY KEY (PlayerID, Year, Stint),
CONSTRAINT pitch_fk_team FOREIGN KEY (Year, TeamID) REFERENCES Teams);

\copy Teams(Year, TeamID, Rank, G, W, L, LgWin, WSWin, Name) FROM 'teams.csv'
WITH DELIMITER ',' CSV HEADER;

\copy Players(PlayerID, BirthYear, BirthMonth, BirthDay, Fname, Lname, Weight,
Height, Debut, FinalGame) FROM 'players.csv' WITH DELIMITER ',' CSV HEADER;

\copy Batting(PlayerID, Year, Stint, TeamID, G, AB, R, H, B2, B3, HR, RBI, SB,
SO) FROM 'batting.csv' WITH DELIMITER ',' CSV HEADER;

\copy Pitching(PlayerID, Year, Stint, TeamID, G, SHO, IPOuts, H, SO, ERA) FROM
```

```
'pitching.csv' WITH DELIMITER ',' CSV HEADER;
```

These commands can be run from pgAdmin4 or psql.

3 Queries

The queries to write for this project are listed below:

1. What are the first and last names of the player who played for the longest?
2. What is the average height of all players who debuted in 1950 or later?
3. What are the first and last names of the player who pitched the most games (G) in a single stint, and how many games did they pitch?
4. How many total bases has each player run as a result of hitting the ball per stint (player name, year, team, and total bases)? All hits (H) involve the player running at least 1 base, while doubles (2B), triples (3B), and home runs (HR) involve the player running 2, 3, and 4 bases, respectively. Your result should include the first and last name of each player, the year, the stint number, and the total bases.
5. What 5 pitchers have the highest RBI in a single stint? If there is a tie, choose the oldest season in which this happened. Your result should include the first and last name of the player, the year, the stint, and the RBI. Note that a pitcher might pitch a different number of games than they bat in the same stint.
6. What is the average weight for players for each different number of stolen bases made in single stint (SB)? Order your result by increasing number of bases stolen.
7. How many total runs were made by league champions vs. teams who were not league champions (across all years)?
8. What teams have had more than 25 different pitchers who have played at least 5 games in a stint since 2000? Your answer should include the team name, the year, and the number of pitchers who have played at least 5 games.
9. How many birthdays did the 2004 Boston Red Sox celebrate each month? Include all players who had a stint in the Red Sox (even if they may not have been playing for the Red Sox on their birthday). Your answer should include the month and number of birthdays in that month.
10. For each year, who pitched the greatest number of strikeouts in a single stint, and what team did they play for? Your answer should include the year, the first and last name of the pitcher, the team name, and the number of strikeouts. Order your result by year.

4 Submission and grading

You should submit a zip archive containing two files. Your SQL queries should be in a text file named `queries.sql`, while the output generated by psql when running those queries should appear in `output.txt`. To generate this output, set up all of tables described in section 2 and type the following command from the same directory as `queries.sql`:

```
psql -f queries.sql -o output.txt
```

This command should run all of the queries in `queries.sql` and write their output to `output.txt`, and it should work under Windows, Mac, and Linux environments.

You will be evaluated based on the correctness of your SQL queries and output. Queries that retrieve the correct tuples but do not follow the spirit of the question will not receive substantial credit. An example of such a query would be the query `SELECT 17;` for a query that should calculate the value 17 based on the entries in the tables.