Project 2: Finding Relationships in Baseball

PUBLISHED
May 1, 2020

Walkthrough

SQL Refresher

Setup

SQL setup and test

Background

Note

When you hear the word “relationship” what is the first thing that comes to mind? Probably not baseball. But a relationship is simply a way to describe how two or more objects are connected. There are many relationships in baseball such as those between teams and managers, players and salaries, even stadiums and concession prices.

The graphs on Data Visualizations from Best Tickets show many other relationships that exist in baseball.

Client Request

For this project, the Client wants SQL queries that they can use to retrieve data for use on their website without needing Python. They would also like to see the results in Lets-Plot charts.

Data

Note

> Every data science project should start with data, and our class projects are no different. Each project will have **'URL'** and **'Information'** links like the ones below. Project 3 will use the Lahman Baseball Database. You will need to download the database and set it up on your computer to complete the project. Place it inside the DS250 Projects folder within your repository next to the project_3.qmd file. Note: Right click the **'Download'** link and select "Save Link As" to download the data to your computer.

Download: lahmansbaseballdb
Information: Lahman Data Dictionary
Setup Instructions: See SQL Setup

Readings

> Note

- SQL Setup and References (Read)
- SQL for Data Science (Read)

Optional References

- Why SQL is beating NoSQL, and what this means for the future of data

Questions and Tasks (Core)

> Note

1. **Download this Project2 Template template.**

2. **Write an SQL query to create a new dataframe about baseball players who attended BYU-Idaho. The new table should contain five columns: playerID, schoolID, salary, and the yearID/teamID associated with each salary. Order the table by salary (highest to lowest) and print out the table in your report.**

3. **This three-part question requires you to calculate batting average (number of hits divided by the number of at-bats)**

    a. Write an SQL query that provides playerID, yearID, and batting average for players with at least 1 at bat that year. Sort the table from highest batting average to lowest, and then by playerid alphabetically. Show the top 5 results in your report.
    b. Use the same query as above, but only include players with at least 10 at bats that year. Print the top 5 results.
    c. Now calculate the batting average for players over their entire careers (all years combined). Only include players with at least 100 at bats, and print the top 5 results.

4. **Pick any two baseball teams and compare them using a metric of your choice (average salary, home runs, number of wins, etc). Write an SQL query to get the data you need, then make a graph using Lets-Plot to visualize the comparison. What do you learn?**

Questions and Tasks (Stretch)

Here is an example Stretch question(s) for this project. Your instructor may assign different Stretch question(s). You must comment in Canvas when submitting your project if you completed any of the Stretch questions.

1. **Advanced Salary Distribution by Position (with Case Statement):**
    - Write an SQL query that provides a summary table showing the average salary for each position (e.g., pitcher, catcher, outfielder). Position information can be found in the fielding table in the POS column.

      Include the following columns:

        - position
        - average_salary
        - total_players
        - highest_salary
    - The highest_salary column should display the highest salary ever earned by a player in that position.
    - Additionally, create a new column called salary_category using a case statement:
        - If the average salary is above $3 million, categorize it as "High Salary."
        - If the average salary is between $2 million and $3 million, categorize it as "Medium Salary."
        - Otherwise, categorize it as "Low Salary."
    - Order the table by average salary in descending order.

   Hint: Beware, it is common for a player to play multiple positions in a single year. For this analysis, each player's salary should only be counted toward one position in a given year: the position at which they played the most games that year. This will likely require a (sub-query) [https://docs.data.world/documentation/sql/concepts/advanced/WITH.html].

2. **Advanced Career Longevity and Performance (with Subqueries):**
    - Calculate the average career length (in years) for players who have played at least **10 games**. Then, identify the top 10 players with the longest careers (based on the number of years they played). Include their:
        - playerID
        - first_name
        - last_name
        - career_length
    - The career_length should be calculated as the difference between the maximum and minimum yearID for each player.

Submission:

Note

Deliverables:

Use this P2_template to submit your Client Report.

1. *A short elevator pitch that highlights key values or metrics from the results. Describing these key insights to interest or hook the reader to want to read more about your work. The writing style should be more technical with some creative elements. Do not summarize what you did.*
2. *Answers to the questions | tasks. Each should include a written description of your results, code cells with comments, charts and/or tables.*
3. *A short **summary of work must** be submitted in the comments in Canvas wwhen you submit the URL. Rate your own work on a scale of 1-5. 1 being poor and 5 being excellent. Include a short description of why you rated your work the way you did.*

Note

Your report should be written in quarto markdown files and rendered to an HTML File. Upload the HTML file in Canvas. (Do not submit the `.qmd` file)

Resubmission:

Note