codecademy

Off Platform Project: The Best Of Baseball Awards

In this project, we will be looking at a massive baseball database containing information about players, teams, managers, salaries, and just about anything you might want to know about baseball. This dataset contains data from 2019 all the way back to 1871. Let's see what interesting facts we can learn from this database!

Before you get started, make sure you have gone through the necessary setup steps found in our article about setting up PostgreSQL locally. You will need to set up a PostgreSQL server as well as a client to connect to your server. As described in the article linked above, we recommend using Postbird for your client.

This project is part of the Design Databases With PostgreSQL Skill Path. To learn the skills needed to complete this project, check out the lessons on Queries, Aggregate Functions, and Multiple Tables.

Step 1 — Downloading The Data

In the files that you downloaded to begin this project, you will find a file called `baseball_database.sql`. In your PostgreSQL client, create a new database (we named ours `baseball`), and then open this `.sql` file. This should run all of the PostgreSQL commands to completely set up your tables and populate the tables with the data.

If you're using Postbird, you can create a new database under the "Select database" menu, by choosing "Create Database". Then you can then import the .sql file by selecting "Import .sql file" from the "File" menu.

A note about this data: This is a remarkable dataset that was put together by Sean Lahman, among others. This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License. We ported Sean's work to a PostgreSQL database to work for this project.

Step 2 — Investigate The Data

Now that we've got the data in a database, it's time to take a look at what we're working with. Note that there is a *lot* of data here — 29 tables to be exact. If you're curious about any of the fields in any of these tables, you can consult the `readme.txt` file provided in the downloaded files. You can also find more detailed documentation on Sean Lahman's site.

For now, let's take a look at some of the most important tables. Take a look at the first few rows of the `people`, `batting`, `pitching`, and `teams` table. You can do this by writing some SQL queries or by using your client's GUI to look at the table's contents. (In Postbird, this is the "Content" tab). Make note of the fields there, and how they relate to each other. Start to think about how these tables connect. For example, how might you link a player's name to the team they played on in a given year.

Part 3 - Handing Out Awards

Now that we've got a sense of what the data looks like, let's use our querying skills to hold our own baseball awards show — The Best of Baseball Awards. For each of these awards, write a query to find the award winner. If you get stuck on any of these queries, you can look at our solution in the `solutions.sql` file provided. We've also given you a hint on how you might start writing your queries for each of these awards.

Heaviest Hitters

This award goes to the team with the highest average weight of its batters on a given year.

Hint: We need to join three tables together — The `people` table contains the weights of each player. We can link those players to the year and team they were batting for in the `batting` table. Finally, the `batting` table has a `team_id` field, but we want the actual team name. We can link the `batting` table to the `teams` table to find the name of the team. We'll need to use the `GROUP BY`, `AVG()`, and `DESC` to find the average weight of the players on these teams.

Shortest Sluggers

This award goes to the team with the smallest average height of its batters on a given year. This query should look very similar to the one you wrote to find the heaviest teams.

Hint: Once again we need to join the `people`, `batting`, and `teams` tables. This time, use your aggregate functions on the `height` column. Also make sure to use `ASC` this time to get the shortest teams first.

Biggest Spenders

This award goes to the team with the largest total salary of all players in a given year.

Hint: You'll mostly be using the `salaries` table for this one. Use `SUM()`, `GROUP BY` and `ORDER BY` to sum every player's salary for a given team on a given year. You'll want to group by both `teamid` and `yearid`. If you want to get the real name of the team rather than just the `teamid`, you'll need to `JOIN` with the `teams` table.

Most Bang For Their Buck In 2010

This award goes to the team that had the smallest "cost per win" in 2010. Cost per win is determined by the total salary of the team divided by the number of wins in a given year. Note that we decided to look at just teams in 2010 because when we found this award looking across all years, we found that due to inflation, teams from the 1900s spent much less money per win. We thought that looking at all teams in just the year 2010 gave a more interesting statistic.

Hint: This should look very similar to your last query. You'll still need to join the `salaries` and `teams` table. This time you want to divide the sum of the players salaries by the `w` column from the `teams` table. Because of

this, you'll also need to add `w` to the `GROUP BY` clause. Finally, we added the `round()` function to the number we're reporting to make our output a little more readable.

Priciest Starter

This award goes to the pitcher who, in a given year, cost the most money per game in which they were the starting pitcher. Note that many pitchers only started a single game, so to be eligible for this award, you had to start at least 10 games.

Hint: You'll need to connect the `salaries` table and the `pitching` table. The column you're interested in the `pitching` table is `gs` (for "games started"). When you join these two tables, you'll want to make sure the `playerid`, `yearid`, and `teamid` all match — it is possible for one player to play on multiple teams in a given year. Make sure to use a `where` clause to ensure the pitcher has started in at least 10 games. Finally, you may want to join with the `people` table to get the player's full name.

Part 4 — Create Your Own Award

Come up with your own award and write a query to find the winner. We'd love to see your creativity! If you come up with a fun award, get in touch with us on Twitter (@Codecademy) to tell us your award name and winner. Use the hashtag #TheBestOfBaseballAwards to see what other people have found!

To get you started, here were some award names that we thought could be fun:

- Bean Machine: The pitcher most likely to hit a batter with a pitch
- Canadian Ace: The pitcher with the lowest ERA who played for a team whose stadium is in Canada
- Worst of the Best: The pitcher or batter inducted into the hall of fame with the worst career stats (you can decide what stat to look at)