**Introduction**

Data for a particular sport is often stored across numerous locations. For example, in NCAA Division I Softball, batting statistics for each season are typically stored in separate tables. (See for example the statistics hosted by <https://d1softball.com/>)

Suppose we are interested in tracking the statistics of players across multiple seasons. A common way to prep the data to do this is to use join statements to merge each seasons data into one table.

This module looks at some simple batting stats over two seasons through the use of joining functions for a small subset of NCAA Division 1 Softball players’ statistics for the 2021 and 2022 seasons. (This is only a small window of a much bigger dataset).

Data

**batting2021**

|  |  |  |  |
| --- | --- | --- | --- |
| **Player** | **team** | **R** | **H** |
| Aaliyah Swan | Cal State Northridge | 8 | 20 |
| Abbey Latham | Ole Miss | 25 | 53 |
| Bella Rocco | Boise State | 16 | 42 |
| Carson Fischer | Northern Colorado | 11 | 15 |
| Drew Dudley | Austin Peay | 10 | 29 |
| Emily Gant | Boston University | 31 | 36 |

**batting2022**

|  |  |  |  |
| --- | --- | --- | --- |
| **Player** | **team** | **R** | **H** |
| Aaliyah Swan | Cal State Northridge | 16 | 23 |
| Abbey Latham | Ole Miss | 36 | 45 |
| Bella Rocco | Boise State | 10 | 29 |
| Emily Gant | Boston University | 40 | 55 |
| Lexi Osowski | Austin Peay | 42 | 64 |
| Mikayla Allee | Ole Miss | 36 | 28 |

**Methods**

Recall the six main joining functions available in the dplyr package from R. Namely,

1. left\_join
2. right\_join
3. full\_join
4. inner\_join
5. semi\_join
6. anti\_join

Exercises

Which of the joining functions are mutating joins and which are filtering joins?

When using a join function, which variable would you want to use as your key and why looking at the datasets above?

Draw the dataset you would create when using a left\_join of batting2022 to batting2021.

Are you keeping the player names from batting2022 or batting2021?

Would the dataset created by a right\_join of batting2022 to batting2021 be identical to the dataset created in question 3? Why or why not?

What is the difference between a full\_join and inner\_join? Answer based on what players would be included in each dataset.

How many rows would result from a full\_join of batting2021 and batting2022?

6. Looking at the new table created below, what join function do you think created it and why?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Player** | **team** | **R.x** | **H.y** | **R.x** | **H.y** |
| Aaliyah Swan | Cal State Northridge | 8 | 20 | 16 | 23 |
| Abbey Latham | Ole Miss | 25 | 53 | 36 | 45 |
| Bella Rocco | Boise State | 16 | 42 | 10 | 29 |
| Emily Gant | Boston University | 31 | 36 | 40 | 55 |

.

7.  Draw the dataset you would create when using an anti\_join of batting2022 to batting2021.

8. If we wanted to see if players' performances improved or not from the 2021 season to the 2022 season, which of the previously mentioned join functions would be the most practical? Explain.