**Learning Goals**

Upon completion of this module students should improve upon

* being able to join two data tables
* being able to anticipate (by sketching) the resulting data table prior to running the code to do so
* identify situations where more than one key is needed

**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 with one row per player (and columns associated with their different statistics for each season).

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**

The data displayed below represent non-random samples taken the full data. **R** is the number of runs scored by the player, **H** is the number of hits. These subsets, as well as the full data (batting2021\_all.csv and batting2022\_all.csv) are available for download on the GitHub repo associated with this module. The full data also includes additional batting statistics.

**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**

For R users, recall that six commonly used joining functions available through the dplyr package are, left, right, full, inne, semi, and anti.

Further details regarding joins can be found at <https://r4ds.hadley.nz/joins.html#how-do-joins-work>

**Exercises**

The first five exercises use the subset data previously displayed.

1. When using a join function on the datasets above, which variable would you want to use as your key and why?
2. Sketch the dataset you would create when using a left\_join of batting2022 to batting2021.
   1. Are you keeping the player names from batting2022 or batting2021?
   2. If you have access to the appropriate technology, conduct the left\_join and check the output with your sketch.
3. 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?

1. If we wanted to see if players' performances improved or not from the 2021 season to the 2022 season, which of the join functions would be the most practical? Explain.
2. Looking at the new table created below, provide the command that would result in this table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |

The remaining exercises use the full datasets. If you have the appropriate technology, check your answer for each by running the appropriate join.

1. Although not seen in the subset of data we used in the first four exercises, when considering multiple seasons of data for all NCAA Division 1 softball players, there is a reasonable chance that two players will have the same name. For now, let’s assume that players with the same name won’t play for the same team. When using a join function, what would you want to use as your key?
2. Suppose you only used the player’s name as the key, explain what would happen to your dataset if you did a full join?
3. What is the difference between a full join and inner join? Answer based on what type of players would be included in each dataset.
4. Describe the dataset you would create when using an anti join of batting2022\_all to batting2021\_all. Provide context as to what this represents for college softball.If you have the appropriate technology, check your answer by running each of these joins.
5. How would that differ if you used an anti join of batting2021\_all to batting2022\_all? Provide context as to what this represents for college softball. If you have the appropriate technology, check your answer by running each of these joins.
6. Suppose you want to analyze trends batting averages across 10 seasons of data. You should assume that all 10 seasons have the same column names.
   1. Briefly explain why repeatedly using an inner join to merge all 10 seasons would not be a good approach.
   2. Briefly explain why repeatedly using use a full join to merge all 10 seasons would not be an ideal approach.
   3. What would be a better solution?

**Conclusions**

By working through the different scenarios, students will become aware of how the choice of join changes the meaning of the resulting table. Additionally, when working with large datasets, students should recognize that care must be taken when choosing their keys as duplicates may not be easy to see in a snippet of the data.