Joining Functions Worksheet

Softball relies on a player's ability to score runs by stepping up to home plate and batting. The statistics that are collected can determine the batting order, what pitches a pitcher throws and even where the infielders and outfielders stand. This worksheet looks at some simple batting stats over two seasons through the use of joining functions (this is only a small window of a bigger dataset). When there are more than one datasets, it is important to be able to merge them in cases where you want to compare. In this case, we want to compare batting statistics over two seasons. The complete datasets can be used to do further calculations and comparisons.

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

Joining functions:

1. left_join
2. right_join
3. full_join
4. inner_join
5. semi_join
6. anti_join

Questions

1. Which of the joining functions are mutating joins and which are filtering joins?
   Mutating joins: left, right, full, and inner
   Filtering joins: semi and anti

2. When using a join function, which variable would you want to use as your key and why looking at the datasets above?
   The variable you would want to use is Player and not team because there are multiple players on Ole Miss and there are two players from Austin Peay, but they aren't the same players

3. Draw the dataset you would create when using a left_join of batting2022 to batting2021.

Player	team	R.x	H.x	R.y	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
Carson Fischer	Northern Colorado	11	15	NA	NA
Drew Dudley	Austin Peay	10	29	NA	NA
Emily Gant	Boston University	31	36	40	55

   a. Are you keeping the player names from batting2022 or batting2021?
      We keep the player names from batting2021 and fill in the rest of the non-matching keys with NA.

4. 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?
It would not be identical because the left_join keeps the "left" dataset in the R code which is batting2021. The code for the right_join would look similar, but it keeps the "right" dataset players which would be from batting2022.

5. What is the difference between a full_join and inner_join? Answer based on what players would be included in each dataset.
A full_join would include all of the players even if there isn't a matching key. The missing variables if there isn't a matching key would be filled with the value NA. An inner_join includes only players with matching keys and all of the variables from both datasets.

a. How many rows would result from a full_join of batting2021 and batting2022?
There would be 8 rows.

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

An inner_join was used because its only players with matching keys and its the variables from both datasets.

7. Draw the dataset you would create when using an anti_join of batting2022 to batting2021.

Player	team	R	H
Carson Fischer	Northern Colorado	11	15
Drew Dudley	Austin Peay	10	29

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.

We would need to use a filtering join because we would want to have the batting statistics from both seasons to compare. A left join of batting2022 to batting2021 would keep only the players who were in the 2021, but also the 2022 season. This would allow us to make easy comparisons between players that we know played in both seasons.