All data came from the Weehoop package which can be viewed at [https://CRAN.R-project.org/package=wehoop](https://cran.r-project.org/package=wehoop).

Exercises

Code is in Red

Question Solution are in Blue

The `wnba\_data` data set consists of these 10 variables: game\_id, season, season\_type, game\_date, team\_id, team\_display\_name, team\_winner, opponent\_team\_id, team\_home\_away. Our goal is to create a two way table for a given team containing their mid-season record, final record, whether they would make playoffs based on their mid-season record, and whether they actually made playoffs.

1. Load in the `wnba\_data` data set.

wnba\_data <- read\_csv("kevari20/data/wnba\_data.csv")

1. What do you notice about the team IDs in this data set? Do they all belong to a valid team or are some not needed? (Hint: Might need to use the `distinct()` function).

wnba\_data %>% select(team\_id, team\_display\_name) %>% distinct() %>% arrange(team\_id)

The IDs above 20 are not necessary because they don't represent teams playing in the regular season.

1. Filter out the IDs we won’t be using.

wnba <- wnba\_data %>% filter(team\_id <= 20)

1. Now, let's make sure each team ID is associated with the correct team name. Use a `select()` statement with both `team\_id` and `team\_name` and then use the `distinct()` function. Which team IDs are repeated? What might this mean?

wnba %>% select(team\_id, team\_display\_name) %>% distinct() %>% arrange(team\_id)

This means that the three names with ID 3 and the three names with ID 17 are the same teams who switched their name between seasons.

1. For the IDs you found, rename the teams so that the same IDs all have the most recent team name. You can create a new variable called `team\_name`.

wnba <- wnba %>% mutate(team\_name = case\_when(team\_id == 3 ~ 'Dallas Wings',

team\_id == 17 ~ 'Las Vegas Aces',

TRUE ~ team\_display\_name))

1. Now that our team names are correct, we can look at games played. Create a new data set called `reg\_season` that only has data for regular season games (season\_type == 2) and one called `playoffs` that only has data for playoff games (season\_type == 3).

reg\_season <- wnba %>% filter(season\_type == 2)

playoffs <- wnba %>% filter(season\_type == 3)

1. To calculate win percentage at the mid-way point, we need to know how many games are played in a season. Use the `tally()` function to count the number of games played by each team within each season. What do you notice?

reg\_season %>% group\_by(season, team\_id) %>% tally()

The number of games played is inconsistent.

1. Do some Google searching on how many games were played by WNBA teams during these seasons. You might find that the number of regular season games has fluctuated since 2020, but there is still a problem. Can you tell what it is?

Not every team has the same number of games recorded for each season.

1. Let’s look into the 2008 season. There are 4 teams that played 33 games instead of 34. Find out who these teams are and Google their season statistics. Did they actually only play 33 games? Why is this a problem?

reg\_season %>% filter(season == 2008) %>% group\_by(team\_id) %>% tally()

No. They all played 34 games. This is a problem because our analysis will not be accurate if there is missing data.

1. The data was scraped from the ESPN website. Go to <https://www.espn.com/wnba/team/schedule/\_/name/atl/season/2008> and click on the first two scores recorded in the 'RESULT' Section. What is different about the pages these links take you to? How might this be causing the problem?

The second game of the season played on 05/23 does not have any team box score data, unlike the other games. This, along with other missing game data, likely caused issues when scraping the data from the webpage.

1. What are some ways we could solve this problem if we still wanted to create the table originally indicated? Choose a way and make a table with your new data.

- We could look up the games and fill in the missing data by adding rows to the dataset.

- We could remove the seasons with missing data from the data set.

