**Cleaning, Merging, and Analyzing Football Data Key**

College football has a widely followed and popular recruiting cycle. The top high school players are often being looked at years before they graduate. One of the top recruited positions is wide receiver. Wide receivers are responsible for catching passes from the quarterback and running down the field to gain as many yards as possible.

We will be investigating recruiting data from 2013 to 2019 for high school wide receivers and information on wide receivers from the 2023 NFL season. We are interested in cleaning the data and analyzing the impact high school recruits had in the NFL in 2023.

**Data Description**

Tm - NFL team the player played for in 2023

Ranking - Recruiting ranking of the player coming out of high school. The lower rankings mean that a player is a top prospect (i.e. a player ranked 1 would be the top prospect in their class)

Year - Year that recruit graduated high school

Yds - Number of receiving yards player had in 2023

NFL\_Position – Position the player played in the NFL

**Exercises**

Assume the data file you will be using in this worksheet will be called “football” and the necessary packages are already loaded. The data contains recruiting information for Division 1 college football recruits and statistics for players who played in the NFL in 2023. Note that some of the high school recruits did not play in the NFL in 2023 and some of the NFL receivers weren't identified in the high school recruiting data.

1. Write the code to create two sets of data, one containing players that competed in the NFL in 2023 and another that contains players who didn't compete in the NFL in 2023. Report the number of players in each dataset. Hint: Players who did not play in the NFL in 2023 will have Tm = NA.

data set with the NFL players. Call it “nfl\_2023”*.*

nfl\_2023 <- football %>%

 filter(!is.na(Tm))

nfl\_2023 %>%

summarise(n())

data set without the NFL players. Call it “non\_nfl\_2023”.

non\_nfl\_2023 <- football %>%

 filter(is.na(Tm))

non\_nfl\_2023 %>%

summarise(n())

2. The first data set now contains players that competed in the NFL in 2023 including players not in the recruiting class from 2013 to 2019. Write the code to filter out the non-recruited players using the *Ranking* variable and the non wide receivers using the *NFL\_Position* variable. Write out the code necessary to find the number of players. Hint: “WR” represents wide receiver in the dataset. Non-ranked players have Ranking = NA in the dataset. Call this dataset “nfl\_recruits\_2023”.

nfl\_recruits\_2023 <- nfl\_2023 %>%

 filter(NFL\_Position == “WR”,

!is.na(Ranking))

nfl\_recruits\_2023 %>%

 summarise(n())

3. Write out the code necessary to find how many wide receivers in 2023 were ranked in the top 100 in their high school recruiting class (there were 17). What are some possible reasons why worse ranked players were playing in 2023 over better ranked players?

nfl\_2023 %>%

 filter(Ranking < 100) %>%

 summarise(n())

In the 2023 NFL there were 17 wide receivers who were a part of the top 100 in their high school recruiting class. Some possibilities that explain why worse ranked players were playing over better ranked players include injury and larger improvements in college.

4. Write the code to make three different new data frames: one being from recruiting years 2013-14, another from 2015-16, and the last from 2017-19.

nfl\_13\_14 <- nfl\_recruits\_2023 %>%

 filter(Year >= 2013 & Year <= 2014)

nfl\_15\_16 <- nfl\_recruits\_2023 %>%

 filter(Year >= 2015 & Year <= 2016)

nfl\_17\_19 <- nfl\_recruits\_2023 %>%

 filter(Year >= 2017 & Year <= 2019)

5. Write the code to get the mean number of receiving yards for 2023 in each recruiting year group. Below are the mean number of receiving yards for each group rounded to the nearest yard.

2013-14: 344

2015-2016: 296

2017-2019: 424

nfl\_13\_14 %>%

 summarise(mean(Yds))

nfl\_15\_16 %>%

 summarise(mean(Yds))

nfl\_17\_19 %>%

 summarise(mean(Yds))

6. Write the code to find the mean rankings for each group of years. Do the lowest ranking years have the highest mean receiving yards? Explain. Note that the mean rankings for each group are shown below.

 2013-14: 1037

 2015-16: 792

 2017-19: 810

nfl\_13\_14 %>%

 summarise(mean(Ranking))

nfl\_15\_16 %>%

 summarise(mean(Ranking))

nfl\_17\_19 %>%

 summarise(mean(Ranking))

No, the group with the best mean ranking (2015-16) actually had the fewest number of mean yards. Years 2017-2019 had the second best mean ranking and the most number of mean receiving yards and years 2013-2014 had by far the worst mean ranking and the second highest number of receiving yards.

7. Using the original “football” dataset, write the code to make a simple scatterplot of recruiting ranking vs. receiving yards in the NFL that is similar to the one shown below. What is the general trend?



ggplot(data = football ,mapping = aes(x = Ranking, y = Yds)) +

 geom\_point() +

 geom\_smooth(method = "lm") +

 labs(title = "Trend of Ranking vs. Receiving Yards") +

 theme\_minimal() +

 theme(plot.title = element\_text(hjust = 0.5))

The general trend is that as a player’s rank number increases their receiving yards in the NFL decreases.