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

College football has the most widely followed recruiting cycle of any college sport. The top high school players are often being looked at years before they graduate. One of the top recruited positions are wide receivers. Wide receivers are responsible for catching passes from the quarterback and running down the field to gain the most yards as possible. They are an integral part of the offense.

We will be investigating recruiting data from 2013 to 2019 for high school wide receivers. We are interested in cleaning the data and analyzing the impact some of them 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

**Exercises**

To begin, load in the necessary packages and the \*\*college\_FB.csv\*\* file. The data contains recruiting information for Division 1 college football recruits and statistics for those who played in the NFL in 2023.

library(tidyverse)

library(here)

football <- read.csv(here("football.csv"))

1. 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.

# data set with the NFL players

nfl\_2023 <- football %>%

filter(!is.na(Tm))

# data set without the NFL players

non\_nfl2023 <- football %>%

filter(is.na(Tm))

2. The new data set now contains players that competed in the NFL in 2023 including players not in the recruiting class from 2013 to 2019. Filter out those players using the `Ranking` and `Tm` variable.

nfl\_recruits\_2023 <- football %>%

filter(!is.na(Tm),

!is.na(Ranking))

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 20)? What are some possible reasons why lower ranked players were playing in 2023 over higher ranked players?

nfl\_2023 %>%

filter(Ranking < 100) %>%

summarise(n())

In the 2023 NFL there were 20 wide receivers who were a part of the top 100 in their high school recruiting class. Some possibilities that explain why lower ranked players were playing over higher ranked players include injury and larger improvements in college.

4. Let's practice splitting and joining tables using dplyr. 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.

2013-14: 364.32

2015-2016: 310.49

2017-2019: 442.75

nfl\_13\_14 %>%

summarise(mean(Yds))

nfl\_15\_16 %>%

summarise(mean(Yds))

nfl\_17\_19 %>%

summarise(mean(Yds))

The recruits from 2017-2019 had the highest mean number of receiving yards in the NFL in 2023, while recruits from 2015-2016 had the lowest mean number of receiving yards.

6. 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: 1037.08

2015-16: 805.23

2017-19: 809.41

nfl\_13\_14 %>%

summarise(mean(Ranking))

nfl\_15\_16 %>%

summarise(mean(Ranking))

nfl\_17\_19 %>%

summarise(mean(Ranking))

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

7. Use `bind\_rows()` to merge the three data frames for the 3 different groups of years back together so they are all in one data set again.

nfl\_13\_16<- bind\_rows(nfl\_13\_14, nfl\_15\_16)

nfl\_full <- bind\_rows(nfl\_13\_16, nfl\_17\_19)

8. Write the code to make a simple scatterplot of recruiting ranking vs. receiving yards in the NFL that matches the one shown below. What is the general trend?

A graph showing the difference between receiving yards

Description automatically generated

ggplot(data = nfl\_full,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.