Assignment 1 - NFL FastR 2024

Robert Morson

# Introduction

This assignment uses the **nflfastR** dataset for the 2024 NFL season.  
The dataset includes both **quantitative variables** (numbers such as yards gained and EPA)  
and **categorical variables** (labels such as play type and team).

# Step 1: Load packages

knitr::opts\_chunk$set(echo = TRUE)  
library(nflfastR) # For NFL play-by-play data  
library(dplyr) # For data manipulation

## Warning: package 'dplyr' was built under R version 4.3.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2) # For plotting  
  
# Load 2024 play-by-play data  
pbp\_data <- load\_pbp(2024)  
  
# Select a mix of quantitative and categorical variables  
df <- pbp\_data %>%  
 select(yardline\_100, yards\_gained, epa, play\_type, posteam) %>%  
 filter(!is.na(yardline\_100))  
  
# Display first few rows  
head(df)

## ── nflverse play by play data ──────────────────────────────────────────────────

## ℹ Data updated: 2025-09-03 05:24:26 EDT

## # A tibble: 6 × 5  
## yardline\_100 yards\_gained epa play\_type posteam  
## <dbl> <dbl> <dbl> <chr> <chr>   
## 1 35 0 0.258 kickoff ARI   
## 2 70 3 -0.201 run ARI   
## 3 67 22 2.03 pass ARI   
## 4 45 9 0.754 pass ARI   
## 5 36 2 -0.0296 run ARI   
## 6 34 2 -0.248 run ARI

# Quantitative variables  
summary(df[, c("yardline\_100", "yards\_gained", "epa")])

## yardline\_100 yards\_gained epa   
## Min. : 1.00 Min. :-24.000 Min. :-12.68910   
## 1st Qu.:30.00 1st Qu.: 0.000 1st Qu.: -0.58549   
## Median :48.00 Median : 1.000 Median : -0.03376   
## Mean :47.23 Mean : 4.202 Mean : 0.01328   
## 3rd Qu.:67.00 3rd Qu.: 6.000 3rd Qu.: 0.59600   
## Max. :99.00 Max. : 98.000 Max. : 8.54123   
## NA's :80 NA's :1

# Categorical variables  
table(df$play\_type)

##   
## extra\_point field\_goal kickoff no\_play pass punt   
## 1302 1166 2949 2829 20006 2119   
## qb\_kneel qb\_spike run   
## 437 75 15044

table(df$posteam)

##   
## ARI ATL BAL BUF CAR CHI CIN CLE DAL DEN DET GB HOU IND JAX KC   
## 1342 1396 1531 1597 1328 1403 1401 1447 1451 1421 1479 1419 1554 1352 1284 1595   
## LA LAC LV MIA MIN NE NO NYG NYJ PHI PIT SEA SF TB TEN WAS   
## 1490 1396 1380 1381 1459 1352 1346 1391 1316 1749 1454 1375 1330 1460 1362 1686

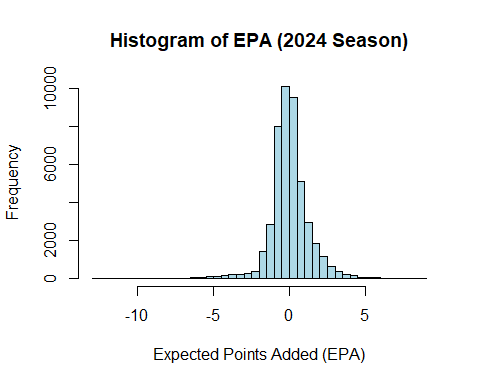
# Create a new variable: log of yards gained  
df$log\_yards <- log(df$yards\_gained + 1)

## Warning in log(df$yards\_gained + 1): NaNs produced

head(df$log\_yards)

## [1] 0.000000 1.386294 3.135494 2.302585 1.098612 1.098612

hist(df$epa, breaks=50, main="Histogram of EPA (2024 Season)",   
 xlab="Expected Points Added (EPA)", col="lightblue")



ggplot(df, aes(x = yardline\_100, y = yards\_gained, color = play\_type)) +  
 geom\_point(alpha = 0.5) +  
 labs(title = "Yardline vs Yards Gained (2024 Season)",  
 x = "Yardline (distance to endzone)",  
 y = "Yards Gained") +  
 theme\_minimal()

## Warning: Removed 80 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

