2025-06-09

Problem 2

```
library(rvest)
library(tidyverse)
```

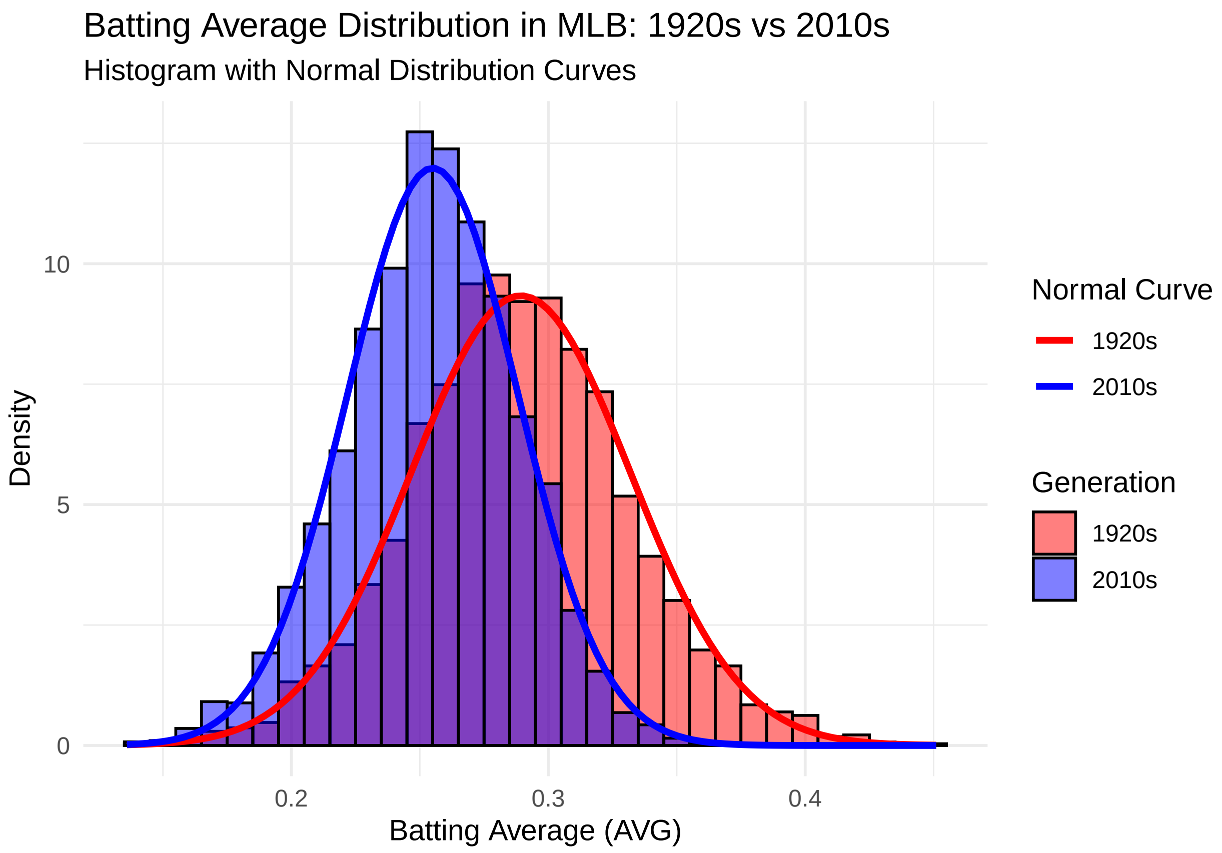
```
lst <- list()
i <- 1
for(year in c(1920:1929, 2010:2019)){
  url <- paste0("https://www.fangraphs.com/leaders/major-league?startdate=&enddate=&month=0&season1=", 
  print(url)
  page <- read_html(url)
  tables = html_nodes(page, "table")
  df <- html_table(tables[[9]])
  df$Season <- year
  lst[[i]] <- df
  i <- i+ 1
}
data <- do.call("rbind", lst)
save(data, file="data.RData")
```

```
load("data.RData")
data$Generation <- ifelse(data$Season<=1929, "1920s", "2010s")
data <- data %>%
  select(Generation, `AVGAVG - Batting Average (H/AB)`, `PAPA - Plate Appearances`) %>%
  dplyr::filter(`PAPA - Plate Appearances` > 150) %>%
  dplyr::select(Generation, `AVGAVG - Batting Average (H/AB)`) %>%
  rename(AVG="AVGAVG - Batting Average (H/AB)")
# calculate the mean and standard deviation for each generation.
stats <- data %>%
  group_by(Generation) %>%
  summarise(Mean=mean(AVG),
            SD=sd(AVG))
stats
```

```
# A tibble: 2 x 3
  Generation  Mean     SD
  <chr>      <dbl>  <dbl>
1 1920s      0.289 0.0427
2 2010s      0.255 0.0333
```

```
ggplot(data, aes(x = AVG)) +
  geom_histogram(aes(y = ..density.., fill = Generation),
                 alpha = 0.5, color = "black", binwidth = 0.01, position = "identity") +
    stat_function(fun = function(x) dnorm(x, mean = stats$Mean[stats$Generation == "1920s"],
                                          sd = stats$SD[stats$Generation == "1920s"]),
                aes(color = "1920s"), size = 1.2) +
stat_function(fun = function(x) dnorm(x, mean = stats$Mean[stats$Generation == "2010s"],
                                      sd = stats$SD[stats$Generation == "2010s"]),
              aes(color = "2010s"), size = 1.2) +

scale_fill_manual(values = c("1920s" = "red", "2010s" = "blue")) +
scale_color_manual(values = c("1920s" = "red", "2010s" = "blue"),
                   name = "Normal Curve") + labs(
  title = "Batting Average Distribution in MLB: 1920s vs 2010s",
  subtitle = "Histogram with Normal Distribution Curves",
  x = "Batting Average (AVG)",
  y = "Density",
  fill = "Generation"
) +
theme_minimal()
```


Batting Average Distribution in MLB: 1920s vs 2010s
Histogram with Normal Distribution Curves
Density
Batting Average (AVG)
Normal Curve
1920s
2010s
Generation
1920s
2010s