Hypothesis Testing with Men's and Women's Soccer Matches

Figure 1: A soccer pitch for an international match.

You're working as a sports journalist at a major online sports media company, specializing in soccer analysis and reporting. You've been watching both men's and women's international soccer matches for a number of years, and your gut instinct tells you that more goals are scored in women's international football matches than men's. This would make an interesting investigative article that your subscribers are bound to love, but you'll need to perform a valid statistical hypothesis test to be sure!

While scoping this project, you acknowledge that the sport has changed a lot over the years, and performances likely vary a lot depending on the tournament, so you decide to limit the data used in the analysis to only official `FIFA World Cup` matches (not including qualifiers) since `2002-01-01`.

You create two datasets containing the results of every official men's and women's international football match since the 19th century, which you scraped from a reliable online source. This data is stored in two CSV files: `women_results.csv` and `men_results.csv`.

The question you are trying to determine the answer to is:

> Are more goals scored in women's international soccer matches than men's?

You assume a **10% significance level**, and use the following null and alternative hypotheses:

H_0 : The mean number of goals scored in women's international soccer matches is the same as men's.

H_A : The mean number of goals scored in women's international soccer matches is greater than men's.

Exploratory data analysis

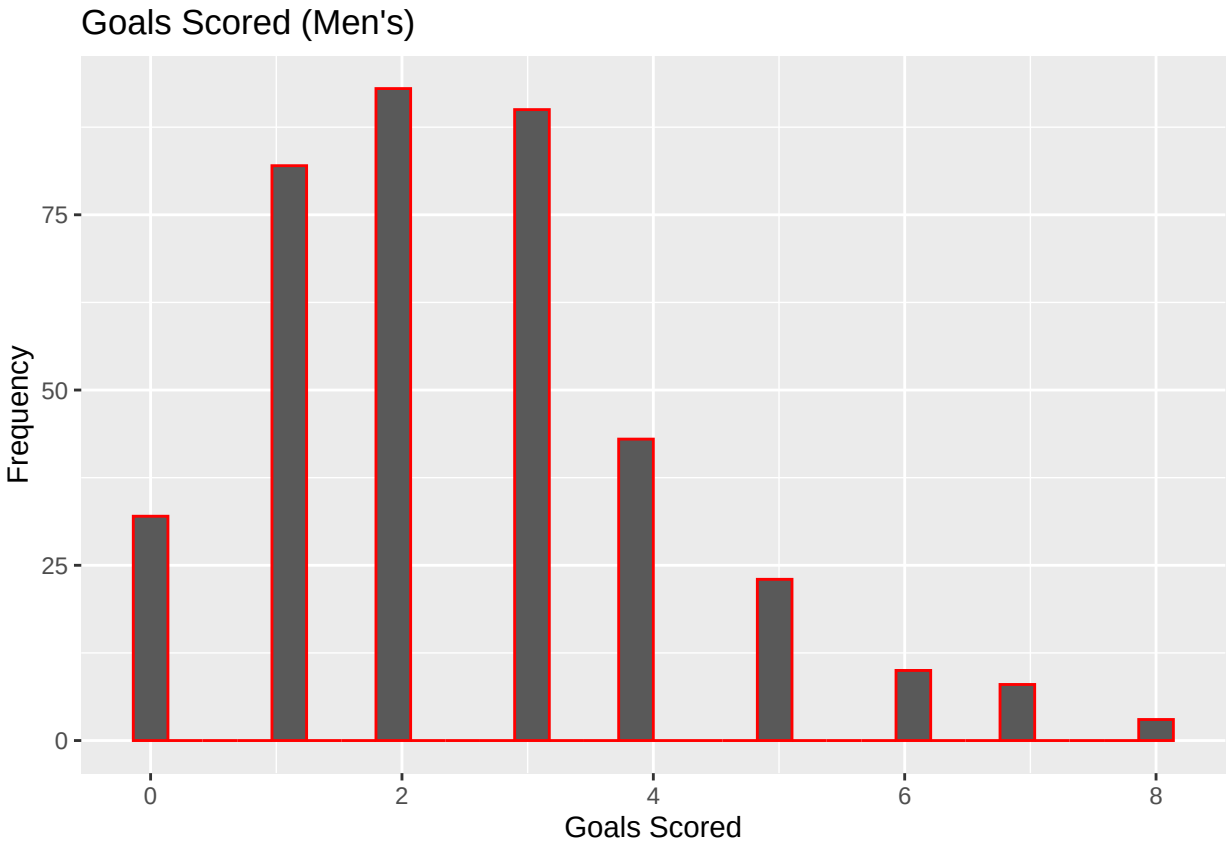
```
library(tidyverse)
library(gridExtra)
## Load men's and women's datasets
```

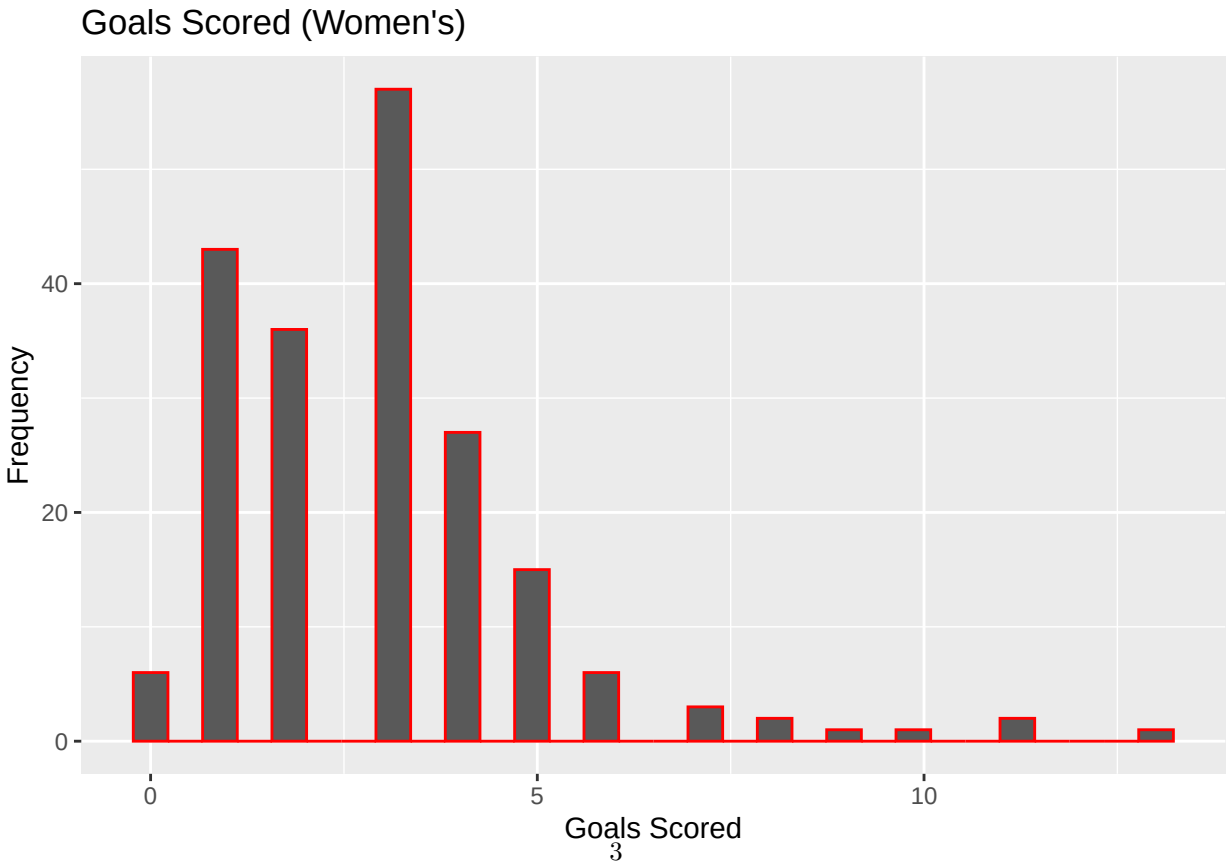
```
men <- read.csv("men_results.csv")
women <- read.csv("women_results.csv")
```

Filtering the data

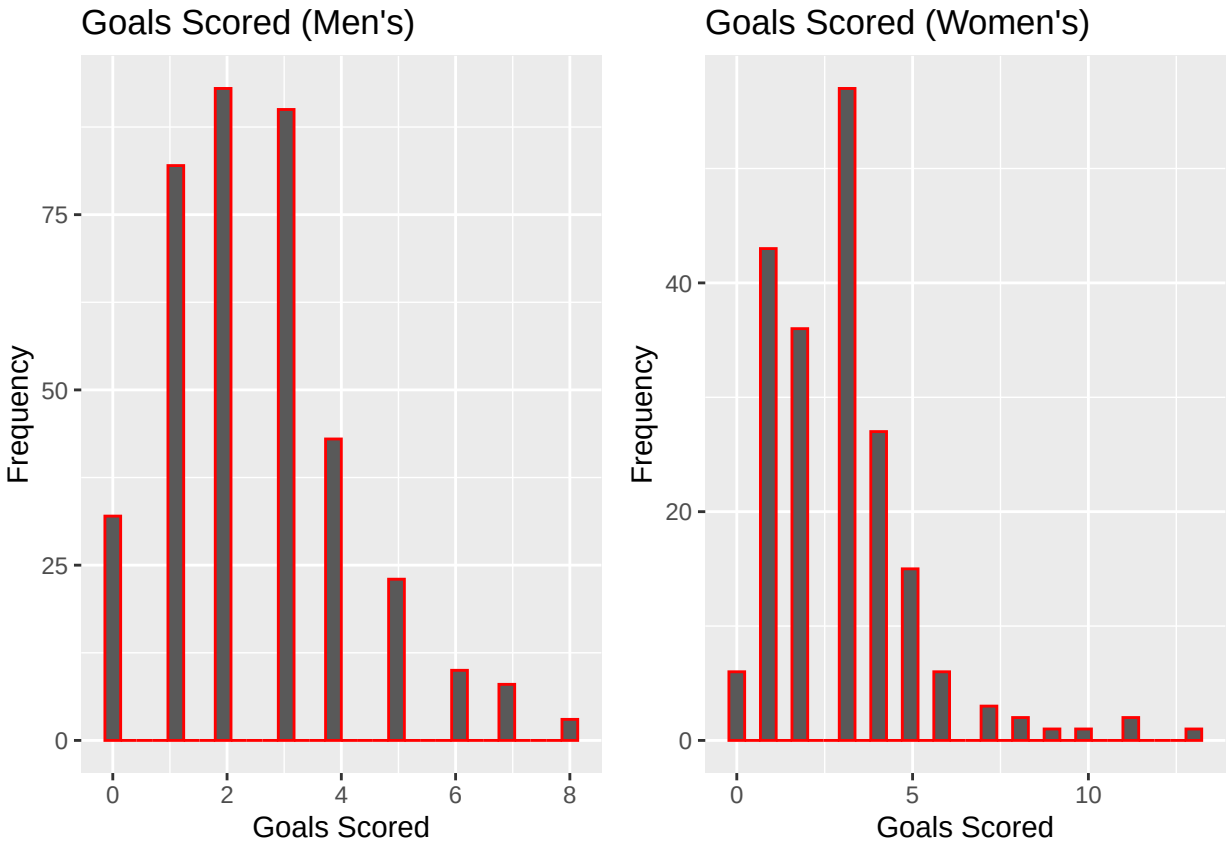
```
men <- men %>%
            filter(tournament == "FIFA World Cup", date > "2002-01-01") %>%
                mutate(goals_scored = home_score + away_score)
women <- women %>%
            filter(tournament == "FIFA World Cup", date > "2002-01-01") %>%
                mutate(goals_scored = home_score + away_score)
```

Normality tests





```
## Goals scored is not normally distributed, so use Wilcoxon-Mann-Whitney test of two groups
grid.arrange(men_plot, women_plot, nrow = 1)
```



Choosing the correct hypothesis test

```r
## Run a Wilcoxon-Mann-Whitney test on goals_scored vs group
test_results <- wilcox.test(
  x = women$goals_scored,
  y = men$goals_scored,
  alternative = "greater"
)
```

Interpreting the result of the hypothesis test

```r
p_val <- round(test_results$p.value, 4)
result <- ifelse(p_val <= 0.01, "reject", "fail to reject")

# Create the result data frame
result_df <- data.frame(p_val, result)
result_df
```

```
##    p_val result
## 1 0.0051 reject
```