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) after `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.

```
# Import necessary libraries
library(tidyverse)
```

Hidden output

```
# Load men and women results
men <- read_csv("men_results.csv")
women <- read_csv("women_results.csv")
```

Hidden output

```
# Some EDA to understand the results
str(men)
str(women)
```

Hidden output

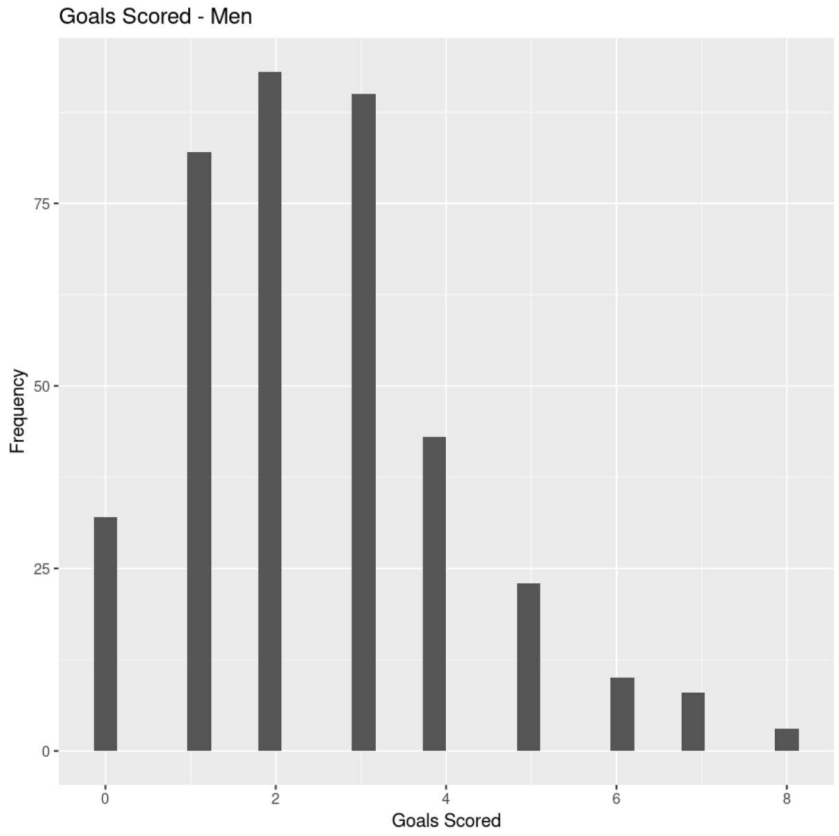
```
# Filter matches / date and total goals scored in a single variable
men <- men %>%
          filter(tournament == "FIFA World Cup", date > "2002-01-01") %>%
          mutate(goals_scored = home_score + away_score)

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

Some histogram plots (men and women results)

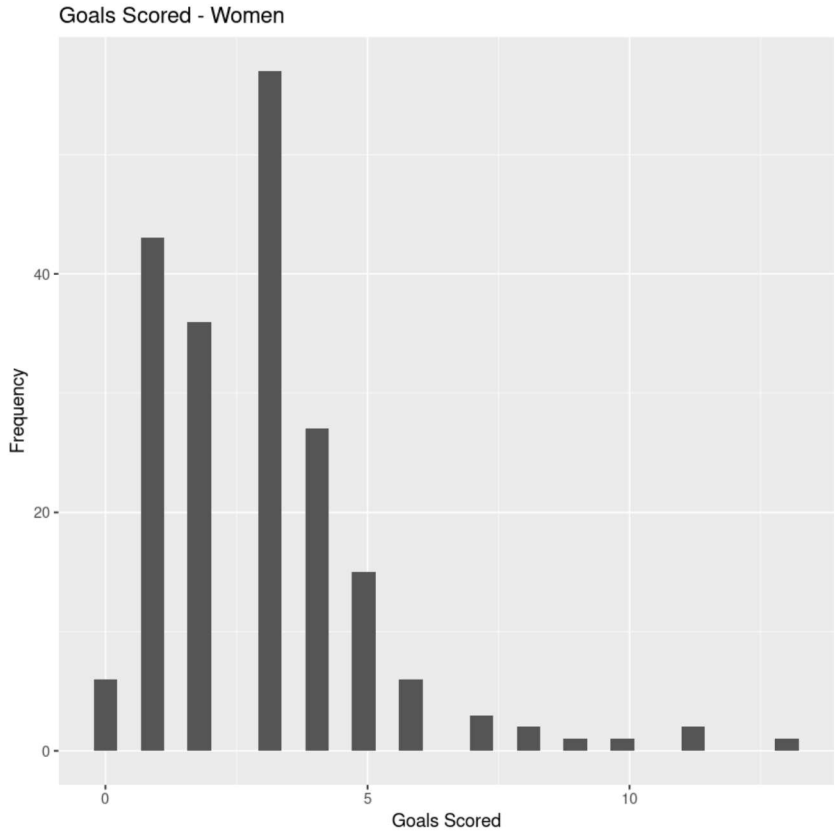
```
men_plot <- ggplot(men, aes(x = goals_scored)) +
  geom_histogram(bins = 30) +
  ggtitle("Goals Scored - Men") +
  xlab("Goals Scored") +
  ylab("Frequency")

men_plot
```




```
women_plot <- ggplot(women, aes(x = goals_scored)) +
  geom_histogram(bins = 30) +
  ggtitle("Goals Scored - Women") +
  xlab("Goals Scored") +
  ylab("Frequency")

women_plot
```



In the plots we can see that the goals results are not normally distributed (and also unpaired two-sample)

In this case, a test that can be used is the Wilcoxon-Mann-Whitney

```
# Calculate the test results
test_results <- wilcox.test(
  x = women$goals_scored,
  y = men$goals_scored,
  alternative = "greater"
)

test_results
```

```
    Wilcoxon rank sum test with continuity correction

data:  women$goals_scored and men$goals_scored
W = 43273, p-value = 0.005107
alternative hypothesis: true location shift is greater than 0
```

```
# Calculate the p_val and result based on the previous test results
p_val <- round(test_results$p.value, 4)
alpha <- 0.10
result <- ifelse(p_val <= alpha, "reject", "fail to reject")

result
```