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:

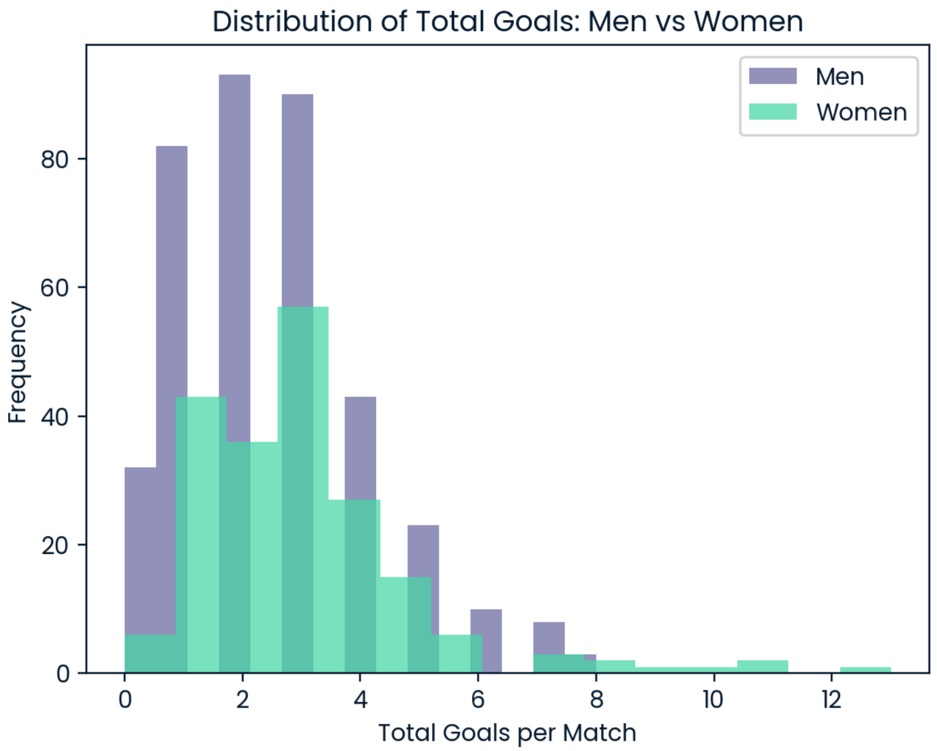
H_0 : The mean number of goals scored in women's international soccer matches is the same as men's.

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

```
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import shapiro, ttest_ind, mannwhitneyu

# Exploratory Data Analysis
men_results = pd.read_csv('men_results.csv')
women_results = pd.read_csv('women_results.csv')
print(men_results.info())
print(women_results.info())

# Convert 'date' to datetime
men_results['date'] = pd.to_datetime(men_results['date'])
women_results['date'] = pd.to_datetime(women_results['date'])

# Filter by date and tournament
men_wc = men_results[(men_results['date'] >= '2002-01-01') &
                     (men_results['tournament'] == 'FIFA World Cup')]

women_wc = women_results[(women_results['date'] >= '2002-01-01') &
                         (women_results['tournament'] == 'FIFA World Cup')]

# Calculate total goals per match
men_wc['total_goals'] = men_wc['home_score'] + men_wc['away_score']
women_wc['total_goals'] = women_wc['home_score'] + women_wc['away_score']

# Plot histograms
plt.hist(men_wc['total_goals'], bins=15, alpha=0.7, label='Men')
plt.hist(women_wc['total_goals'], bins=15, alpha=0.7, label='Women')
plt.xlabel('Total Goals per Match')
plt.ylabel('Frequency')
plt.title('Distribution of Total Goals: Men vs Women')
plt.legend()
plt.show()

# Test for normality using Shapiro-Wilk
sw_men = shapiro(men_wc['total_goals'])
sw_women = shapiro(women_wc['total_goals'])

print(f"Shapiro-Wilk p-value (Men): {sw_men.pvalue:.4f}")
print(f"Shapiro-Wilk p-value (Women): {sw_women.pvalue:.4f}")

# Decide which test to use based on normality
if sw_men.pvalue > 0.05 and sw_women.pvalue > 0.05:
    print("\n✅ Both distributions appear normal: using unpaired t-test.")
    test_stat, p_val = ttest_ind(women_wc['total_goals'], men_wc['total_goals'], alternative='greater')
else:
    print("\n⚠️ At least one distribution is not normal: using Mann-Whitney U test.")
    test_stat, p_val = mannwhitneyu(women_wc['total_goals'], men_wc['total_goals'], alternative='greater')

# Print test results
print(f"Test statistic: {test_stat:.4f}")
print(f"One-tailed p-value: {p_val:.4f}")

# Final decision based on alpha = 0.10
alpha = 0.10  # 10% significance level
```

```
if p_val < alpha:
    print("🎉 Conclusion: Women’s matches have significantly more goals on average (p < 0.10).")
else:
    print("📉 Conclusion: No significant difference at the 10% level.")

# Store results in a dictionary
result_dict = {"p_val": p_val, "result": result}

# Optional: print the dictionary
print(result_dict)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44353 entries, 0 to 44352
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  44353 non-null  int64 
 1   date        44353 non-null  object
 2   home_team   44353 non-null  object
 3   away_team   44353 non-null  object
 4   home_score  44353 non-null  int64 
 5   away_score  44353 non-null  int64 
 6   tournament  44353 non-null  object
dtypes: int64(3), object(4)
memory usage: 2.4+ MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4884 entries, 0 to 4883
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  4884 non-null   int64 
 1   date        4884 non-null   object
 2   home_team   4884 non-null   object
 3   away_team   4884 non-null   object
 4   home_score  4884 non-null   int64 
 5   away_score  4884 non-null   int64 
 6   tournament  4884 non-null   object
dtypes: int64(3), object(4)
memory usage: 267.2+ KB
```



```
Shapiro-Wilk p-value (Men): 0.0000
Shapiro-Wilk p-value (Women): 0.0000

⚠️ At least one distribution is not normal: using Mann-Whitney U test.
Test statistic: 43273.0000
One-tailed p-value: 0.0051
🎉 Conclusion: Women’s matches have significantly more goals on average (p < 0.10).
{'p_val': 0.005106609825443641, 'result': 'reject'}
```