PROJECT: HYPOTHESIS TESTING WITH MEN'S AND WOMEN'S SOCCER MATCHES

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.

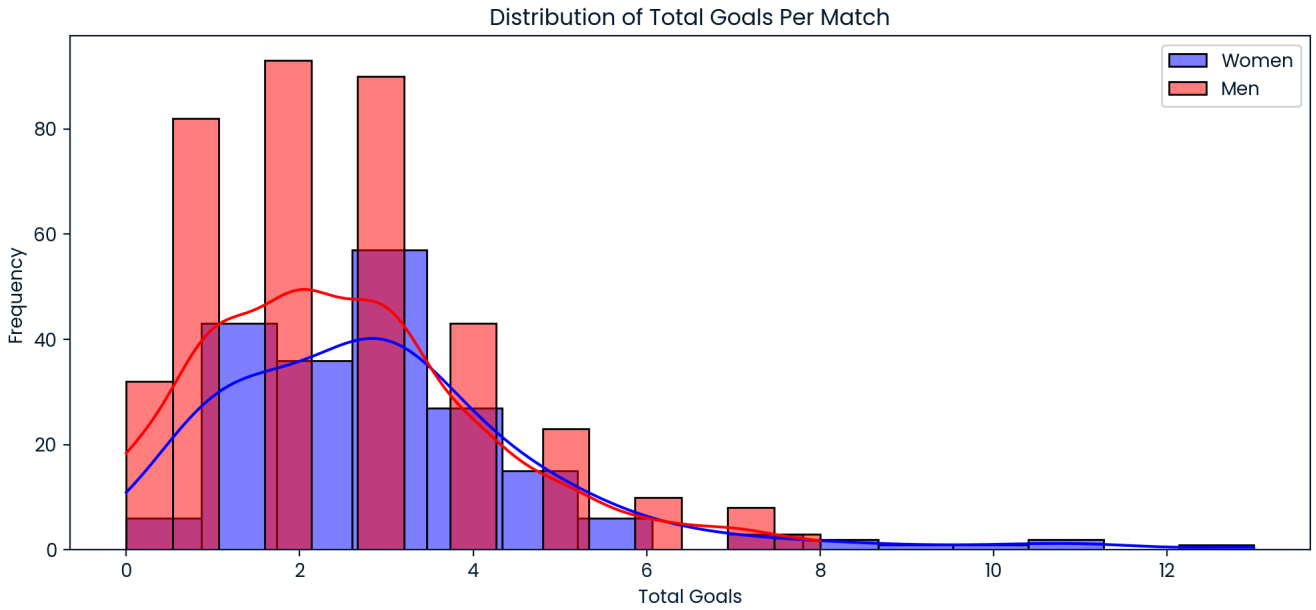
H_A : The mean number of goals scored in women's international soccer matches is greater than men's.

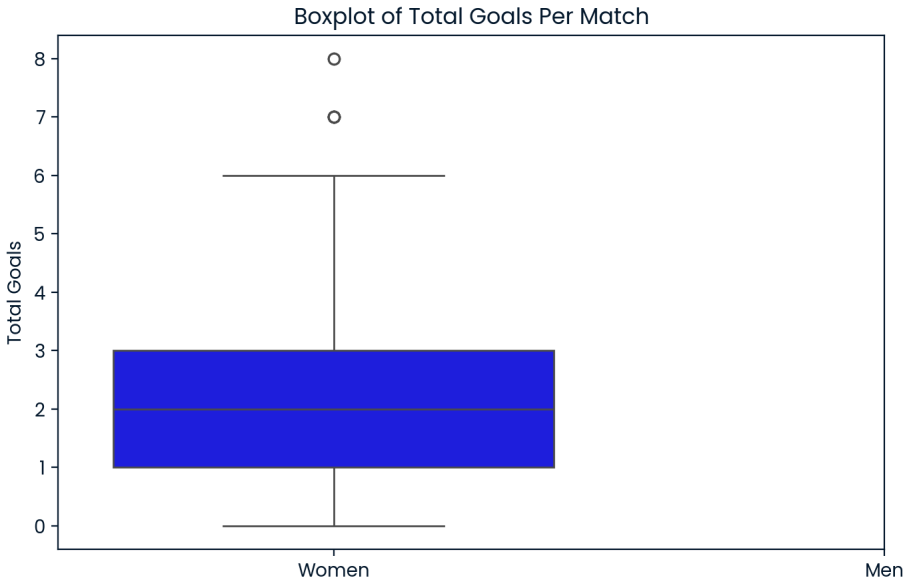
	U...	date	home_team	away_team	h...	a...	tournamer
1600	1600	2003-09-20T00:00:00.000	Nigeria	North Korea	0	3	FIFA W
1601	1601	2003-09-20T00:00:00.000	Norway	France	2	0	FIFA W
1602	1602	2003-09-20T00:00:00.000	Germany	Canada	4	1	FIFA W
1603	1603	2003-09-20T00:00:00.000	Japan	Argentina	6	0	FIFA W
1604	1604	2003-09-21T00:00:00.000	United States	Sweden	3	1	FIFA W
1605	1605	2003-09-21T00:00:00.000	Brazil	South Korea	3	0	FIFA W
1606	1606	2003-09-21T00:00:00.000	Australia	Russia	1	2	FIFA W
1607	1607	2003-09-21T00:00:00.000	China PR	Ghana	1	0	FIFA W
1609	1609	2003-09-24T00:00:00.000	Norway	Brazil	1	4	FIFA W
1610	1610	2003-09-24T00:00:00.000	France	South Korea	1	0	FIFA W
1611	1611	2003-09-24T00:00:00.000	Germany	Japan	3	0	FIFA W
1612	1612	2003-09-24T00:00:00.000	Canada	Argentina	3	0	FIFA W
1613	1613	2003-09-25T00:00:00.000	Sweden	North Korea	1	0	FIFA W
1614	1614	2003-09-25T00:00:00.000	United States	Nigeria	5	0	FIFA W
1615	1615	2003-09-25T00:00:00.000	Ghana	Russia	0	3	FIFA W
1616	1616	2003-09-25T00:00:00.000	China PR	Australia	1	1	FIFA W
1620	1620	2003-09-27T00:00:00.000	South Korea	Norway	1	7	FIFA W

Rows: 200

	U...	date	home_team	away_team	h...	a...
25164	25164	2002-05-31T00:00:00.000	France	Senegal	0	1
25165	25165	2002-06-01T00:00:00.000	Germany	Saudi Arabia	8	0
25166	25166	2002-06-01T00:00:00.000	Republic of Ireland	Cameroon	1	1
25167	25167	2002-06-01T00:00:00.000	Uruguay	Denmark	1	2
25168	25168	2002-06-02T00:00:00.000	Argentina	Nigeria	1	0
25169	25169	2002-06-02T00:00:00.000	England	Sweden	1	1
25170	25170	2002-06-02T00:00:00.000	Paraguay	South Africa	2	2
25171	25171	2002-06-02T00:00:00.000	Spain	Slovenia	3	1
25172	25172	2002-06-03T00:00:00.000	Brazil	Turkey	2	1
25173	25173	2002-06-03T00:00:00.000	Croatia	Mexico	0	1
25174	25174	2002-06-03T00:00:00.000	Italy	Ecuador	2	0
25175	25175	2002-06-04T00:00:00.000	China PR	Costa Rica	0	2
25176	25176	2002-06-04T00:00:00.000	Japan	Belgium	2	2
25177	25177	2002-06-04T00:00:00.000	South Korea	Poland	2	0
25178	25178	2002-06-05T00:00:00.000	Germany	Republic of Ireland	1	1
25179	25179	2002-06-05T00:00:00.000	Russia	Tunisia	2	0

Rows: 384






```
Shapiro-Wilk test for women: W=0.8491013050079346, p-value=3.8905201759850683e-13
Shapiro-Wilk test for men: W=0.9266489744186401, p-value=8.894154401688226e-13
At least one distribution is non-normal. Use Wilcoxon-Mann-Whitney test.
```