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.

```
# Start your code here!
import pandas as pd
```

```
import matplotlib.pyplot as plt
import pingouin
from scipy.stats import mannwhitneyu
men = pd.read_csv("men_results.csv")
women = pd.read_csv("women_results.csv")
men["date"] = pd.to_datetime(men["date"])
mensubset = men[(men["date"]>"2002-01-01")&(men["tournament"].isin(["FIFA World Cup"]))]
women["date"] = pd.to_datetime(women["date"])
womensubset = women[(women["date"]>"2002-01-01")&(women["tournament"].isin(["FIFA World Cup"]))]
men
women
```

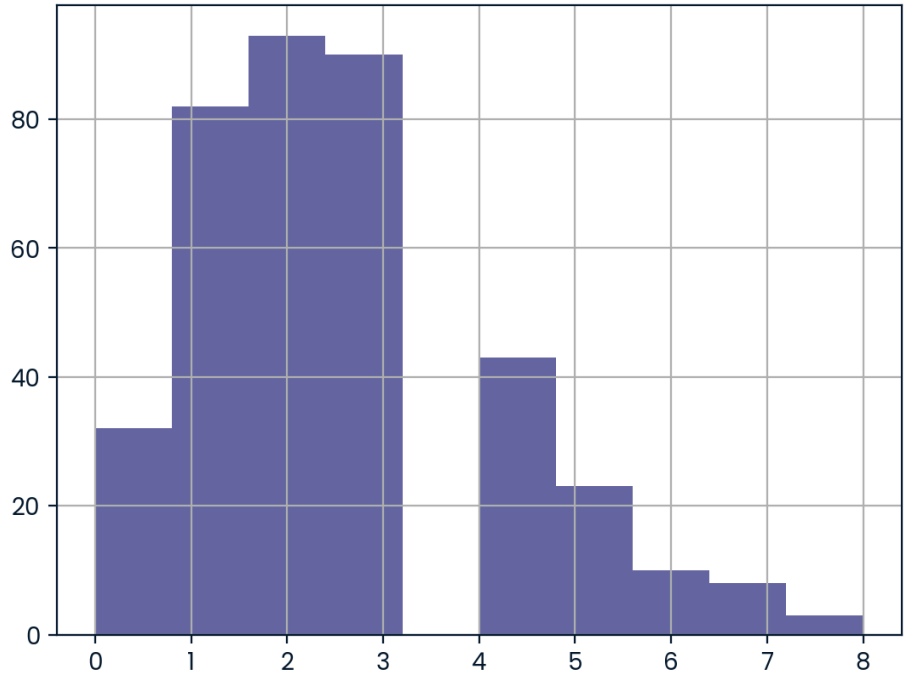
	U...	date	home_team	away_team	h...	a...
0	0	1969-11-01T00:00:00.000	Italy	France	1	
1	1	1969-11-01T00:00:00.000	Denmark	England	4	
2	2	1969-11-02T00:00:00.000	England	France	2	
3	3	1969-11-02T00:00:00.000	Italy	Denmark	3	
4	4	1975-08-25T00:00:00.000	Thailand	Australia	3	
5	5	1975-08-25T00:00:00.000	Hong Kong	New Zealand	0	
6	6	1975-08-27T00:00:00.000	Thailand	Singapore	3	
7	7	1975-08-27T00:00:00.000	New Zealand	Malaysia	3	
8	8	1975-08-29T00:00:00.000	Australia	Singapore	3	
9	9	1975-08-29T00:00:00.000	Hong Kong	Malaysia	0	
10	10	1975-08-31T00:00:00.000	New Zealand	Australia	3	
11	11	1975-08-31T00:00:00.000	Thailand	Malaysia	3	
12	12	1975-09-02T00:00:00.000	Australia	Malaysia	5	
13	13	1975-09-02T00:00:00.000	New Zealand	Thailand	3	
14	14	1977-08-02T00:00:00.000	Taiwan	Indonesia	5	
15	15	1977-08-03T00:00:00.000	Thailand	Singapore	2	

Rows: 4,884

```
mensubset["group"] = "men"
womensubset["group"] = "women"
mensubset["goals_scored"] = mensubset["home_score"] + mensubset["away_score"]
womensubset["goals_scored"] = womensubset["home_score"] + womensubset["away_score"]
```

```
mensubset["goals_scored"].hist()
plt.show()

#not a normal data, hence we use Mann- Whitney U test
```



```
men_women = pd.concat([womensubset, mensubset], axis = 0, ignore_index=True)
men_women
```

	U...	date	home_team	away_team	h...	a...	tournam...
0	1600	2003-09-20T00:00:00.000	Nigeria	North Korea	0	3	FIFA World Cup
1	1601	2003-09-20T00:00:00.000	Norway	France	2	0	FIFA World Cup
2	1602	2003-09-20T00:00:00.000	Germany	Canada	4	1	FIFA World Cup
3	1603	2003-09-20T00:00:00.000	Japan	Argentina	6	0	FIFA World Cup
4	1604	2003-09-21T00:00:00.000	United States	Sweden	3	1	FIFA World Cup
5	1605	2003-09-21T00:00:00.000	Brazil	South Korea	3	0	FIFA World Cup
6	1606	2003-09-21T00:00:00.000	Australia	Russia	1	2	FIFA World Cup
7	1607	2003-09-21T00:00:00.000	China PR	Ghana	1	0	FIFA World Cup
8	1609	2003-09-24T00:00:00.000	Norway	Brazil	1	4	FIFA World Cup
9	1610	2003-09-24T00:00:00.000	France	South Korea	1	0	FIFA World Cup
10	1611	2003-09-24T00:00:00.000	Germany	Japan	3	0	FIFA World Cup
11	1612	2003-09-24T00:00:00.000	Canada	Argentina	3	0	FIFA World Cup
12	1613	2003-09-25T00:00:00.000	Sweden	North Korea	1	0	FIFA World Cup
13	1614	2003-09-25T00:00:00.000	United States	Nigeria	5	0	FIFA World Cup
14	1615	2003-09-25T00:00:00.000	Ghana	Russia	0	3	FIFA World Cup
15	1616	2003-09-25T00:00:00.000	China PR	Australia	1	1	FIFA World Cup

Rows: 584

```
men_women_subset = men_women[["goals_scored", "group"]]
men_women_subset
```

	goal…	
0	3	women
1	2	women
2	5	women
3	6	women
4	4	women
5	3	women
6	3	women
7	1	women
8	5	women
9	1	women
10	3	women
11	3	women
12	1	women
13	5	women
14	3	women
15	2	women
16	8	women

Rows: 584

```
mwpivot = men_women_subset.pivot(columns="group", values="goals_scored")
mwpivot
```

0		3
1		2
2		5
3		6
4		4
5		3
6		3
7		1
8		5
9		1
10		3
11		3
12		1
13		5
14		3
15		2
16		8

Rows: 584

```
result = pingouin.mwu(x=mwpivot["women"], y=mwpivot["men"], alternative="greater")
result
```

		alt...	p-val	RBC	CLES
MWU	43273	greater	0.0051066098	-0.1269010417	0.5634505208

Rows: 1

```
pvalue = result["p-val"].values[0]
pvalue
```

```
0.005106609825443641
```

```
if pvalue <= 0.01:
 answer = "reject"
else:
 answer = "failed to reject H0"
```

```
result_dict = {"p_val": pvalue, "result": answer}
result_dict
```

```
{'p_val': 0.005106609825443641, 'result': 'reject'}
```