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:

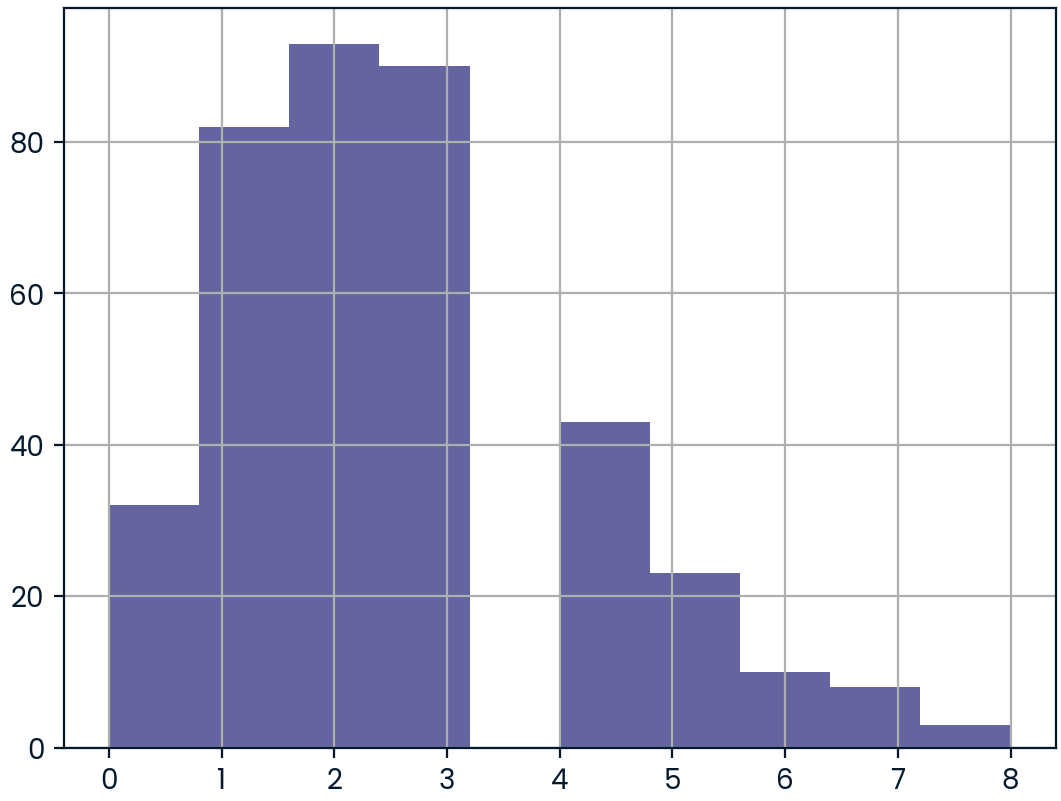
H_0 : The mean number of goals scored in women's international soccer matches is the same as men's.

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")

print(men)
print(women)

men["date"] = pd.to_datetime(men["date"])
men_subset = men[(men["date"] > "2002-01-01") & (men["tournament"].isin(["FIFA World
Cup"]))]

women["date"] = pd.to_datetime(women["date"])
women_subset = women[(women["date"] > "2002-01-01") &
(women["tournament"].isin(["FIFA World Cup"]))]

men_subset["group"] = "men"
women_subset["group"] = "women"

# Corrected column names
men_subset["goals_scored"] = men_subset["home_score"] + men_subset["away_score"]
women_subset["goals_scored"] = women_subset["home_score"] +
women_subset["away_score"]

men_subset["goals_scored"].hist()
plt.show()
plt.clf()

both = pd.concat([women_subset, men_subset], axis=0, ignore_index=True)

both_subset = both[["goals_scored", "group"]]
both_subset_wide = both_subset.pivot(columns="group", values="goals_scored")

result_pg = pingouin.mwu(both_subset_wide["women"], both_subset_wide["men"],
alternative="greater")

results_scipy = mannwhitneyu(x=women_subset["goals_scored"],
y=men_subset["goals_scored"], alternative="greater")

p_val = result_pg["p-val"].values[0]

if p_val <= 0.01:
    result = "reject"
```

```
else:
    result = "fail to reject"

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

```
      Unnamed: 0        date  ... away_score      tournament
0              0  1872-11-30  ...          0        Friendly
1              1  1873-03-08  ...          2        Friendly
2              2  1874-03-07  ...          1        Friendly
3              3  1875-03-06  ...          2        Friendly
4              4  1876-03-04  ...          0        Friendly
...          ...         ...  ...        ...             ...
44348      44348  2022-12-14  ...          1        Friendly
44349      44349  2022-12-14  ...          0        Friendly
44350      44350  2022-12-17  ...          1  FIFA World Cup
44351      44351  2022-12-17  ...          1        Friendly
44352      44352  2022-12-18  ...          3  FIFA World Cup

[44353 rows x 7 columns]
     Unnamed: 0        date  ... away_score        tournament
0             0  1969-11-01  ...          0              Euro
1             1  1969-11-01  ...          3              Euro
2             2  1969-11-02  ...          0              Euro
3             3  1969-11-02  ...          1              Euro
4             4  1975-08-25  ...          2  AFC Championship
...         ...         ...  ...        ...               ...
4879       4879  2022-07-22  ...          0         UEFA Euro
4880       4880  2022-07-23  ...          0         UEFA Euro
4881       4881  2022-07-26  ...          0         UEFA Euro
4882       4882  2022-07-27  ...          1         UEFA Euro
```




```
<Figure size 640x480 with 0 Axes>
```