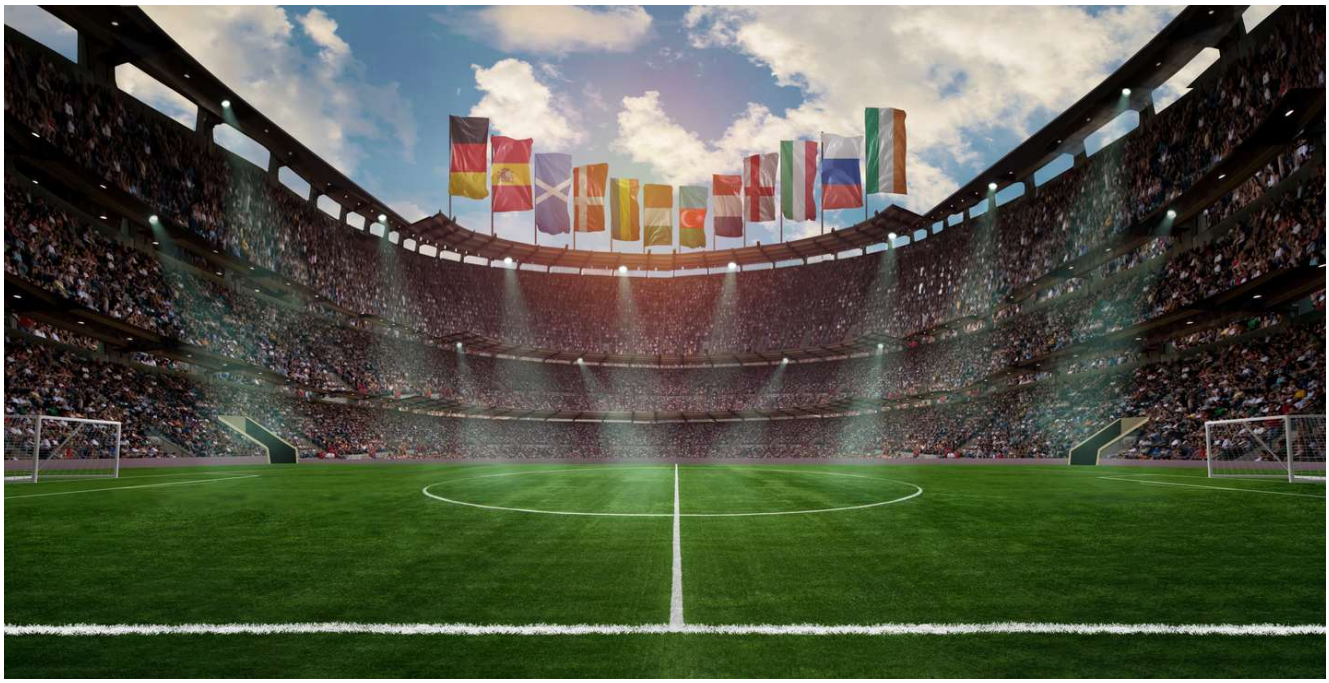
PROJECT: HYPOTHESIS TESTING WITH MEN'S AND WOMEN'S SOCCER MATCHES



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.

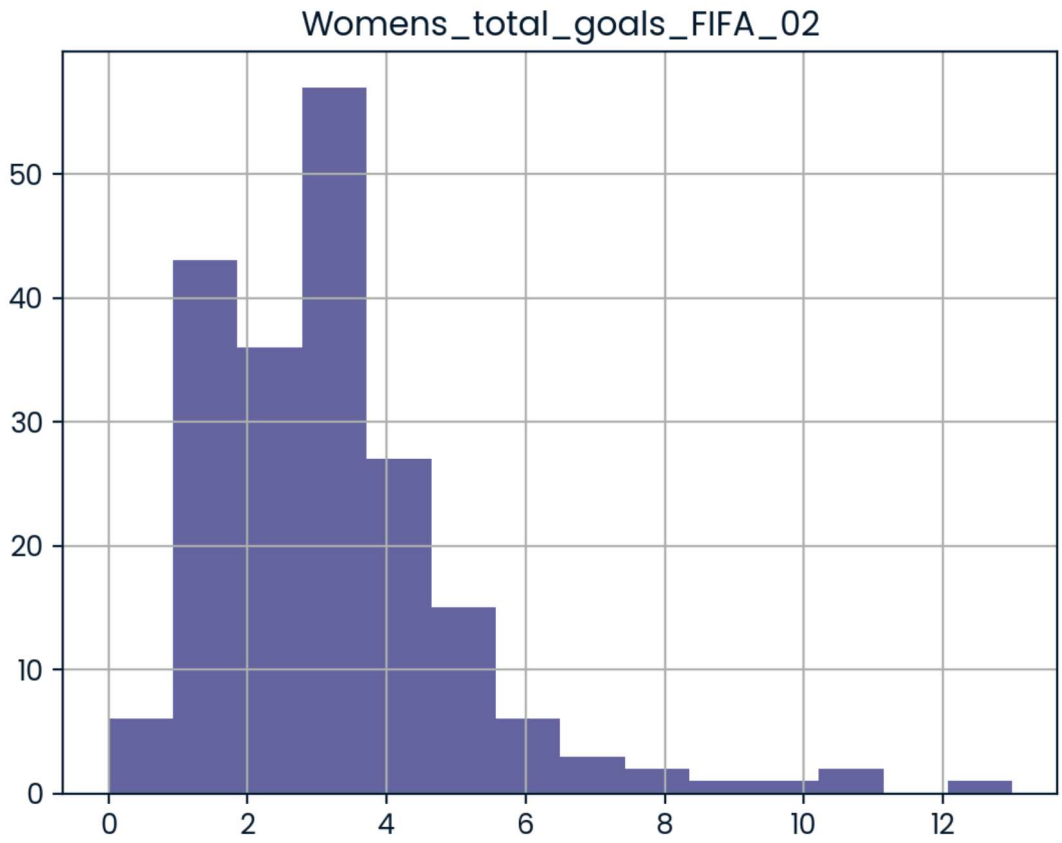
H_A : The mean number of goals scored in women's international soccer matches is greater than men's.

```
import pandas as pd
women_results = pd.read_csv("women_results.csv")
men_results = pd.read_csv("men_results.csv")

#Checking info on both tables, no null values found, both csvs contain identical
columns

#convert date from object to datetime
women_results["date"] = pd.to_datetime(women_results["date"])
men_results["date"] = pd.to_datetime(men_results["date"])

# Filter the dataframe for dates after 2002-01-01 and for games where the tournament
is FIFA World Cup
FIFA_02_women_results = women_results[(women_results["date"] > "2002-01-01") &
(women_results["tournament"] == "FIFA World Cup")]
FIFA_02_men_results = men_results[(men_results["date"] > "2002-01-01") &
(men_results["tournament"] == "FIFA World Cup")]

#FIFA_02_men_results contains 384 observations, FIFA_02_women_results contains 200
observations

#Checking the mean of total goals for Men's and Women's FIFA World Cup matches
(2002-01-01 and earlier)
FIFA_02_women_results["total_goals"] =
FIFA_02_women_results["home_score"]+FIFA_02_women_results["away_score"]
FIFA_02_men_results["total_goals"] =
FIFA_02_men_results["home_score"]+FIFA_02_men_results["away_score"]
print("Men's total goals mean: " +
str(round(FIFA_02_men_results["total_goals"].mean(), 2)))
print("Women's total goals mean: " +
str(round(FIFA_02_women_results["total_goals"].mean(), 2)))

#Women's mean total goals are higher than Men's mean total goals. Is the difference
meaningfully different?

#Determining type of hypothesis test
#Two independent groups -> unpaired two-sample test
#Checking distribution for hypothesis test determination
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(1)
FIFA_02_women_results.hist("total_goals", bins=14)
plt.title("Womens_total_goals_FIFA_02")
plt.show()

plt.figure(2)
```

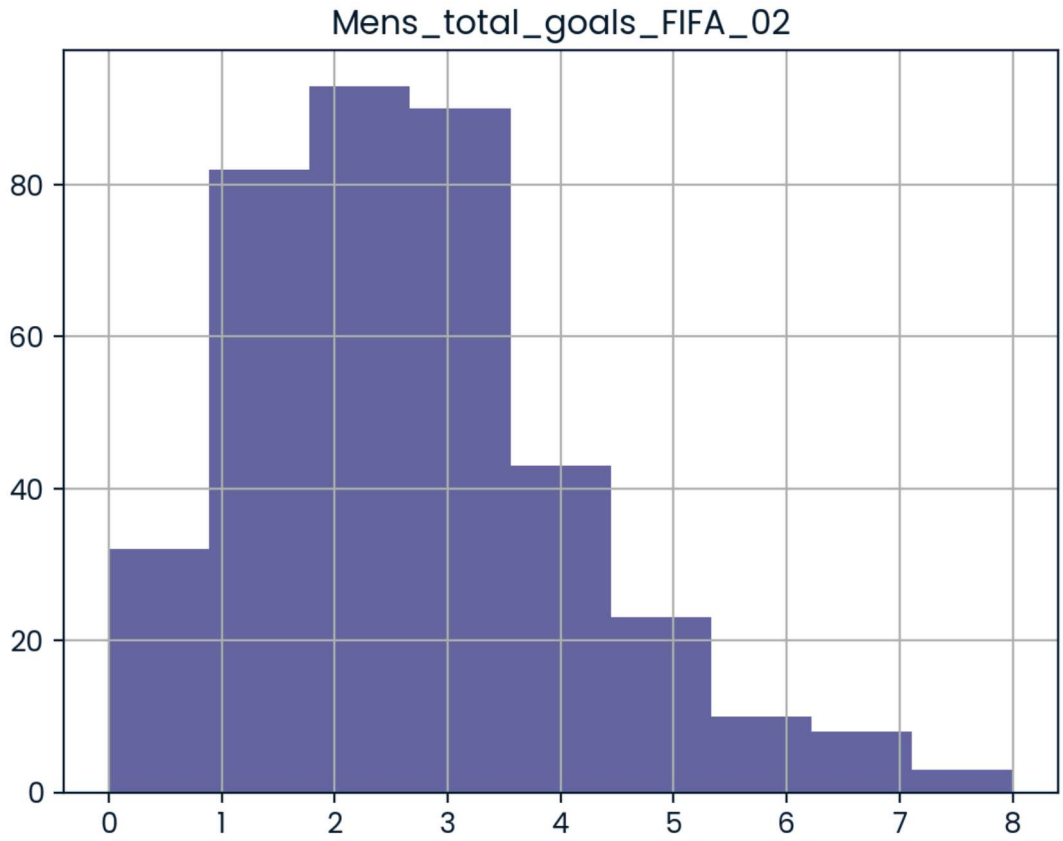
```
FIFA_02_men_results.hist("total_goals", bins=9)
plt.title("Mens_total_goals_FIFA_02")
plt.show()

#Distribution does not appear normal, confirm with Shapiro-Wilk normality test
from scipy.stats import shapiro
import numpy as np

womens_shapiro = shapiro(FIFA_02_women_results["total_goals"])
print("Womens Shapiro p-value: " + str(womens_shapiro.pvalue))

mens_shapiro = shapiro(FIFA_02_men_results["total_goals"])
print("Mens Shapiro p-value: " + str(mens_shapiro.pvalue))

alpha = 0.01

if womens_shapiro.pvalue > alpha:
    print("FIFA_02_women_results normally distributed")
else:
    print("FIFA_02_women_results not normally distributed")

if mens_shapiro.pvalue > alpha:
    print("FIFA_02_men_results normally distributed")
else:
    print("FIFA_02_men_results not normally distributed")

#Both subsets of data not normally distributed, proceed with Wilcoxon-Mann-Whitney
test at 10% significance level
alpha_mwu = 0.1
import pingouin
mann_whitney_u_test = pingouin.mwu(x=FIFA_02_women_results["total_goals"],
y=FIFA_02_men_results["total_goals"], alternative="greater")

p_val = np.array(mann_whitney_u_test["p-val"])

if p_val > alpha_mwu:
    result = "fail to reject"
else:
    result = "reject"

result_dict = {"p_val": p_val, "result": result}
print(result_dict)
```

```
Men's total goals mean: 2.51
Women's total goals mean: 2.98
```

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




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



```
Womens Shapiro p-value: 3.8905201759850683e-13
Mens Shapiro p-value: 8.894154401688226e-13
FIFA_02_women_results not normally distributed
FIFA_02_men_results not normally distributed
{'p_val': array([0.00510661]), 'result': 'reject'}
```

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