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
women = pd.read_csv("women_results.csv")
men = pd.read_csv("men_results.csv") #convert to readable dataframe from csv for
both
print(men.info())
print(women.info()) #find out the column names for each dataframe and datetype
women["date"] = pd.to_datetime(women["date"])
men["date"] = pd.to_datetime(men["date"])#convert time for both men and women to
datetime format since date is listed as object type
women_sub = women[(women["date"] > "2002-01-01") & (women["tournament"].isin(["FIFA
World Cup"]))]
men_sub = men[(men["date"] > "2002-01-01") & (men["tournament"].isin(["FIFA World
Cup"]))] #filter datae for both subsets so that we see data for tournaments that
occurred in the FIFA world cup and after the date of 2002-01-01
women_sub['total_goals']= women_sub['home_score']+women_sub['away_score']
men_sub['total_goals']= men_sub['home_score']+men_sub['away_score'] #created new
column that gave us the total goals of each match
men_sub["group"] = "men"
women_sub["group"] = "women" #since we want to combine these two subsets together,
we created a group in each subset that seperates the matches by men and women once
combined
men_and_women = pd.concat([women_sub, men_sub], axis=0, ignore_index=True) #used
.concat to combine the two subsets into one
men_and_women_sub= men_and_women[['total_goals', 'group']] #we only want total goals
and group as we want to find out if average goals is higher, lower, or same
depending on group men or women
men_and_women_wide = men_and_women_sub.pivot(columns= 'group', values='total_goals')
#since the data is independent and unpaired, we will use wmw test. need to pivot
data before we can do so.
wmw_test = pingouin.mwu(x=men_and_women_wide['women'], y=men_and_women_wide['men'],
alternative='greater')
print(wmw_test) #p value is less than the 0.1 significance level we created, heence
we reject the null hypothesis
result_dict = {'p_val': 0.005107, "result": "reject"}
print(result_dict)
```

```
---  ------      --------------  -----
 0   Unnamed: 0  44353 non-null  int64
 1   date        44353 non-null  object
```

```
 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
None
       U-val alternative     p-val       RBC      CLES
MWU  43273.0     greater  0.005107 -0.126901  0.563451
{'p_val': 0.005107, 'result': 'reject'}
```