Data Processing (Includes Assessment)

QUESTION 1

Using the column names below create a new dataframe that uses only them.

```
columns = ['team', 'group','previous_appearances','previous_titles','previous_finals',
           'previous_semifinals','current_fifa_rank']

### YOUR CODE HERE

```

To help with this analysis we are going to engineer a feature that combines all of the data in the table. This feature represents the past performance of a team. Given the data we have it is the best proxy on hand for how good a team will perfom. Feel free to change the multiplers, but let's just say that *past_performance* will be a linear combination of the related features we have.

Let X_1,....,X_4 be *previous_titles* , *previous_finals* , *previous_semifinals* , *previous_appearances* and let the corresponding vector α be the multipliers. This will give us,

past_performance = $\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4$

Modify α if you wish. Then add to your dataframe the new feature *past_performance*.

QUESTION 2

Create the engineered feature as a new column

```
alpha = np.array([16,8,4,1])

### YOUR CODE HERE
```

QUESTION 3

Using your choice of tools create one or more **tabular summaries** of the data

QUESTION 4

Check for missing data. Write code to identify if there is any missing data.

QUESTION 5

Come up with one or more plots that investigate the central question… Are the groups comprised in a fair way?

There are a number of ways to use hypothesis testing in this situation. There are certainly reasonable hypotheses tests and other methods like simulation approaches, that we have not discussed, but they would be appropriate here. If you choose to explore some of the methods that are outside the scope of this course then we encourage you to first try the simple approach proposed here and compare the results to any further additional approaches you choose to use.

We could use an ANOVA approach here that would signify a difference between groups, but we would not know which and how many teams were different. As we stated before there are a number of ways to approach the investigation, but lets use a simple approach. We are going to setup our investigation to look at all pairwise comparisons to provide as much insight as possible.

Recall that there are $\frac{(N-1)(N)}{2}$ pairwise connections.

```
N = np.unique(df['group'].values).size
print("num comparisons: ",((N-1)*N) / 2.0)
```

```
num comparisons: 28.0
```

QUESTION 5

1. Choose a hypothesis test
2. State the null and alternative hypothesis, and choose a cutoff value αα
3. Run the test for all pairwise comparisons between teams

QUESTION 6

For all of the *p*-values obtained apply the Bonferroni and at least one other correction for multiple hypothesis tests. Then comment on the results.

There is an allpairtest function in statsmodels that could be used here to combine the work from QUESTION 5 and QUESTION 6.

Generalized Linear Models (GLMs) are an appropriate tool to use here if we wanted to include the results of the tournament (maybe a ratio of wins/losses weighted by the final position in the tournament). *statsmodels* supports R-style formulas to fit generalized linear models. One additional variant of GLMs are hierarchical or multilevel models that provide even more insight into this types of dataset. See the tutorial on multilevel modeling.

An answer key has been provided in the form of an online Jupyter Notebook for your to review upon completion of this exercise.