Import the Data

Before we jump into the data it can be useful to give a little background so that you can better understand the features. Since the dawn of statistics practitioners have been trying to find advantages when it comes to games. Much of this was motivated by gambling—here we will look at the results from this tournament in a different way. We are going to ask the simple question

Was the tournament setup in a fair way?

Of course the findings from an investigation centering around this question could be used to strategically place bets, but lets assume that we are simply interested in whether or not the tournament organizers did an adequate job. The reason for doing this is to prepare for the AAVAIL data that is coming. This exercise is an important reminder that you do not have to wait until the day that data arrive to start your work.

There are 32 teams, each representing a single country, that compete in groups or pools then the best teams from those groups compete in a single elimination tournament to see who will become world champions. This is by far the world’s most popular sport so one would hope that the governing organization FIFA did a good job composing the pools. If for example there are 8 highly ranked teams then each of those teams should be in a different pool.

In our data set we have more than just rank so we can dig in a little deeper than that, but first let’s have a look at the data.

```
DATA_DIR = os.path.join("..","data")
df = pd.read_csv(os.path.join(DATA_DIR, 'worldcup-2018.csv'))
df.columns = [re.sub("\s+","_",col.lower()) for col in df.columns]
df.head()
.
```

```
team  group previous_appearances  previous_titles previous_finals
  previous_semifinals current_fifa_rank first_match_against match_index
  history_with_first_opponent_w-l history_with_first_opponent_goals
  second_match_against  match_index.1 history_with_second_opponent_w-l
  history_with_second_opponent_goals  third_match_against match_index.2
  history_with_third_opponent_w-l history_with_third_opponent_goals unnamed:_19
0 Russia  A 10  0 0 1 65  Saudi Arabia  1 -1.0  -2.0  Egypt 17  NaN NaN Uruguay
  33  0.0 0.0 NaN
1 Saudi Arabia  A 4 0 0 0 63  Russia  1 1.0 2.0 Uruguay 18  1.0 1.0 Egypt 34  -5
  .0  -5.0  NaN
2 Egypt A 2 0 0 0 31  Uruguay 2 -1.0  -2.0  Russia  17  NaN NaN Saudi Arabia  34
  5.0 5.0 NaN
3 Uruguay A 12  2 2 5 21  Egypt 2 1.0 2.0 Saudi Arabia  18  -1.0  -1.0  Russia
  33  0.0 0.0 NaN
4 Porugal B 6 0 0 2 3 Spain 3 -12.0 -31.0 Morocco 19  -1.0  -2.0  Iran  35  2.0
  5.0 NaN
```

To limit the dataset for educational purposes we create a new data frame that consists of only the following columns:

- team
- group
- previous_appearances
- previous_titles
- previous_finals
- previous_semifinals
- current_fifa_rank