Simon Leahy
Data Wrangling
Springboard Data Science Career Track
6/18/2020

Overarching Question: What are the attributes that most contribute to the offensive success of a soccer team, and what attributes and range of attributes amongst teammates increase the likelihood of exceeding expected level of success based on overall offensive skill?

The data cleaning requirements of this project involved both traditional data wrangling functions and ad hoc mathematical applications to manipulate the data into the format necessary to address the question. The first step of importing the csv formatted data required cleaning, as the dataset included characters native to languages other than English. In order to allow for non-English characters, the encoding argument needed to be set to 'latin1'. This was applied to the players skillset csv, players' teams csv, and teams' success csv. The next requirement to cleanse the data was to properly weight teams' success based on the league each team was in. The imported team success data listed all of the goals scored within the 2017/2018 season, but that only includes interleague games. It was necessary to adjust the success metric of goals scored to account for differing difficulties, as otherwise teams with players with lower overall skillsets would yield disproportionately high success metrics. Though the exact weighting is subjective, there are many sources that convey a metric that identify the strength and difficulty of each league. The weighting was in the form of a dictionary, with league as keys and the key league weight divided by best league weight (i.e. Spain:1662,England:1660.4/1662) rendering each league weight to be multiplied to goals scored as value ranging from 1 to 0.95. The notebook was built in such a way that a different rating system could be easily implemented. After each team was assigned a weighted success metric, the teams were concatenated by row (axis=0) to create a dataframe of all teams and their new offensive success variable.

The next step, to satisfy the scope of the project, was to extract and merge the top five offensively inclined players (as defined by average of offensive skillsets) with their respective teams in order to generate both an average and standard deviation value of those five players skill sets for each team. This was done using sort_values to perform a two level sort on players by their team and then offensive average, groupby and apply to cut the players data set off at five for each team, pivot_table to aggregate

the five players' skillset by average and standard deviation for each team, and merge to join the the new aggregation of players at the team level with each team's weighted success metric. Additionally, column names were wrangled using the rename function and list comprehension to apply "_avg" and "_std" to each offensive attribute where necessary. As mentioned earlier, some teams from the teams data had non-English characters, and thus did not match the english characters in the team field of the players dataset. These rows were dropped using .dropna, as they comprised a small portion of the data. Outliers did not exist within any of the generated data frames and did not need to be addressed.