CSC343 Term Project - Data Cleaning

The changes based on the comments from phase 3 are highlighted

High-Level Overview of Raw Data

The first dataset selected has the detailed statistics of 18,207 real-world soccer players, saved in a .csv file. It stores each player's basic information such as age, weight, height, nationality, club, URL to the photo, position, and jersey number, where they match with the player's real-world data. In addition, it also has plenty of attributes that reflect the player's in-game behaviours and abilities, including an overall rating (indicating how valuable this player is), the rating when this player is playing for each position in soccer, player's international reputation, as well as the numerical evaluation of soccer skills including passing, shooting, and physical strength. These attributes will be reflected during in-game matches and can be treated as an objective rating for their real-world soccer skills. Overall, there are a total of 89 attributes for each of the 18,207 players in this dataset. Besides the dataset on players' statistics, we include two additional datasets, one containing the languages spoken in each country of the world, and the other fabricated one containing the number of players playing in each role (goalkeepers, attackers, midfielders, defenders) in the lineup formation of standard soccer match.

Data Cleaning Steps & Decisions Made:

On the player data:

1. Drop the columns on irrelevant attributes or information, including jersey number, the international reputation, URLs to player and club photos, etc.
2. Some attributes such as "Skill Movements" and "Weak Foot" contain special characters in value, the clean-up process removes the special characters and converts the values to integers.
3. The "Work Rate" attribute represents if a player's contribution on offence and defence. In the raw data, the rate on offence and defence are listed in the same column, using a "/" to separate (one example can be "High/Medium", meaning that the offence work rate for this soccer player is "High", and the defence work rate is "medium"). The clean-up process separates the work rate into two attributes: "attackRate" and "defendRate".
4. The raw data lists the ratings when each player is playing for each of the positions in soccer, however, some positional ratings are not necessary. For instance, no coach will ever make an attacker play as a goalkeeper. Therefore, for each player, only relevant positional ratings are kept. One example is that, if a player is a defender, then only the player's rating when he is playing as a defender will be kept after cleaning-up process.

On the Country-languages data:

1. In the old schema, we only keep one official language for each country. To address the cases where some countries have multiple official languages, we changed the language table from Language(country, language) to Language(country, language1, language2, language3) where languge2 and language3 can be NULL if the country does not have as many official countries. The reason for having three slots is that, In our data, the maximum number of official languages a country has is 3
2. Load the .csv file in python, read each row and save to a dictionary. Then impute missing countries in the player nationality if not already exists in countrylanguage dataset, to satisfy the foreign key constraints.
3. Remove repeated rows of countries to make each country unique.

On the Formation data:

1. The data originally breaks up the midfield positions into two categories, in our cleaning-up process, we merge the two categories into one. Therefore, in the final table, there will only be one attribute for the number of midfielders needed for a specific formation of lineup.
2. “Number of Goalkeepers” is one of the columns in raw data, we eliminate this as in a standard soccer match there will always be exactly one goalkeeper in any lineup.