CSC343 Term Project

Siwei Tang, Zhenyu Wang

October 2020

Dataset and Relational Schema

Domain

The domain is European soccer clubs performance at their own countries' leagues and at continental competition in past 10 years. To be more specific, we focus on soccer clubs in five biggest leagues in Europe. These leagues are:

- **Premier League** in England
- **Laliga** in Spain
- **SerieA** in Italy
- **Bundesliga** in Germany
- **Ligue1** in France

We focus on **UEFA Champion League**, and **UEFA Europa League** to measure clubs' performance at continental competition. Notice that performance of any individual player is not covered in the domain.

Dataset

Datasets of this project are retrieved through **direct download** and **API extraction** from open data sources. three open data sources are used, click the bold font title to view the actual data source:

1. **openfootball**
    - This source is for all information about results of domestic leagues.
    - To learn about the data, we need to learn rules of different leagues including championships and relegation rules.
    - The dataset needs to be extracted through API, and use Pandas Package to clean the data. Notice that origin source has one data set for each season, we need to merge different seasons into one table.
2. **fbref**

- This source is for all information about results of European competitions (including result of matches).
- To learn about the data, we need to learn rules of competitions as well as rules in soccer matches.
- The dataset needs to be extracted through API, and use Pandas Package to clean the data. Notice that we need to connect this dataset to datasets about domestic leagues on soccer teams, this may require some manual inspection.

3. **Kaggle**
   - This source is for all information about detailed statistics about matches.
   - To learn about the data, we need to learn rules of competitions as well as rules in soccer matches.
   - The dataset needs to be extracted through direct download, and has already been cleaned. Therefore, we only need to merge the data from this source with other datasets.

Questions

1. Which country has a better soccer club performance at the European level competition in past 10 years? Does this relate to their national team's performance?
2. Does soccer club's performance in its own country's league relate to its performance in European competition in past 10 years?
3. Is there a 'Natural Enemy' for clubs in one country to another country (That is, clubs in one country beat clubs in another country for most times) in past 10 years? And what about competitions between clubs from a country at European's stage?

Schema

The following records basic information about countries and soccer clubs:

- Country(<u>cID</u>, name, NumofTeams)
- Team(<u>tID</u>, cID, name)

The following records result of European competitions:

- ChampionLeague(<u>tID, year</u>, MP, W, D, L, GF, GA, GD, Pts, Rank)
- EuropaLeague(<u>tID, year</u>, MP, W, D, L, GF, GA, GD, Pts, Rank)

The following records result of matches in European competitions:

- ChampionLeagueMatch(Home, Away, Year, Stage, HGoal, AGoal)
- EuropaLeagueMatch(Home, Away, Year, Stage, HGoal, AGoal)

The following records result of domestic competitions:

- PremierLeague(tID, year, MP, W, D, L, GF, GA, GD, Pts, Rank)
- Laliga(tID, year, MP, W, D, L, GF, GA, GD, Pts, Rank)
- SerieA(tID, year, MP, W, D, L, GF, GA, GD, Pts, Rank)
- Bundesliga(tID, year, MP, W, D, L, GF, GA, GD, Pts, Rank)
- Ligue1(tID, year, MP, W, D, L, GF, GA, GD, Pts, Rank)

The following records coefficients that measure performance of soccer clubs by country:

- NationalTeamCoefficient(cID, year, rank, coefficient)
- Countrycoefficient(cID, year, rank, coefficient)

The followings are constraints of this relational schema:

- PremierLeague[tID] $\subseteq$ Team[tID]
- Laliga[tID] $\subseteq$ Team[tID]
- SerieA[tID] $\subseteq$ Team[tID]
- Bundesliga[tID] $\subseteq$ Team[tID]
- Ligue1[tID] $\subseteq$ Team[tID]
- ChampionLeagueMatch[Home] $\subseteq$ Team[tID]
- ChampionLeagueMatch[Away] $\subseteq$ Team[tID]
- EuropaLeagueMatch[Home] $\subseteq$ Team[tID]
- EuropaLeagueMatch[Away] $\subseteq$ Team[tID]
- ChampionLeague[tID] $\subseteq$ Team[tID]
- EuropaLeague[tID] $\subseteq$ Team[tID]
- NationalTeamCoefficient[cID] $\subseteq$ Country[cID]
- CountryCoefficient[cID] $\subseteq$ Country[cID]
- Team[cID] $\subseteq$ Country[cID]

The following records abbreviation of attribute names for better understanding:

- cID: country ID

- tID: team (soccer club) ID
- MP: Number of Matches Played
- W: Win, D: Draw, L: Lose
- GF: Goals For, GA: Goals Against, GD: Goals Difference
- Pts: Points
- HGoal: Home Goals, AGoal: Away Goals