Decisions in Data Cleaning

Zhenyu Wang, Siwei Tang

November 2020

- **Season Column Modification**
  ISSUE In the schema, values are only allowed to be INTEGER type in season column. However, in the data set, it records as (year/year) (i.e. (2017/2018)), which is actually character type.
  DECISION Intuitively, it is easier to keep year as INTEGER in order to implement any future data analysis more efficiently. Therefore, we choose to keep the beginning year of the season in the column. (That is, modify (2017/2018) to 2017, and type to Integer).

- **Redundant Games in the Champions League and Europa League**
  ISSUE Redundant matches are matches that involve at least one team that is not from the five leagues. The domain of this database only include about La Liga, Premier League, Seria A, Bundasliga, Ligue 1. In the schema, every team in the game has a foreign key constraint to TEAM relation, and this relation only records teams in those five leagues. Therefore, having other teams violate this constraint.
  DECISION For the purpose of data cleaning, redundant matches are removed, only valid matches in between the five leagues are kept.

- **Multiple Team Names**
  ISSUE In the raw dataset, both the official name and the common name of a soccer team are kept. For instance, 'Tottenham Hotspur' and 'Tottenham Hotspur FC'. Since each team has an ID, mutiple team names may lead to multiple IDs for the same team if we do not pay attention to this issue. This definitely violate the schema and the design of this database.
  DECISION To fix this, we only keep the official team name for each team and modify alternative names. This requires a lot of work, because there is no automatic way to do it.

- **TEAM ID**
  ISSUE In the raw dataset, there is no unique ID for teams in all five leagues. Instead, each league has its own ID system. This results in a key constraint violation, that is, two teams from different leagues may have same ID.
  DECISION To solve this issue, we first merge all teams into one table, then rearrange ID system, give each team a unique ID.