BIOS 669 Project Proposal

Madhuri Raman

March 24th, 2024

https://www.kaggle.com/datasets/azminetoushikwasi/ucl-202122-uefa-champions-league?select=goalkeeping.csv

For my 669 project, I will be working with the UCL datasets in the link above. UCL stands for the UEFA Champions League which is the most prestigious tournament in European soccer, and this data all pertains to the 2021-2022 season (which my favorite team won ☺).

There are 8 separate datasets that about matches and players during that season which are: `attacking`, `defending`, `goalkeeping`, `attempts`, `disciplinary`, `distribution`, and `key\_stats`. Each dataset contains a different subset of the total player base for that tournament depending on their relevance to the dataset. For example, `attacking` contains a list of all players with at least one non-zero attacking stat, `goalkeeping` contains a list of all goalkeepers and their relevant stats, etc. There are around 750 players in total, so that is the maximum number of rows in any given dataset, and there are about 10 columns (more or less) in each dataset for different relevant metrics in soccer.

The goal of this project is to work with these datasets to answer a bunch of interesting questions that could inform predictions for future seasons. In terms of 669 skills, I will be using PROC SQL and various JOIN methods to combine datasets whenever needed for answering a particular question. These may come in handy for creating analysis datasets with derived variables. Additionally, I will most likely have to use lookup table methods. The first step, however, will be to investigate each dataset for data cleaning issues and make any relevant adjustments to the data.

Some questions of interest (more to come):

1. Produce a table of teams by number of total fouls, yellow cards, red cards, and total disciplinary actions in general from highest to lowest. Show the top three teams and bottom three teams of this table. (using `disciplinary` dataset)
2. For each team, what are the most common types of goals scored in the tournament? Specifically, using the `goals` dataset, consider left foot vs. right foot vs. header vs. other, and inside area vs. outside area.
3. Which goalkeeper(s) had the most saves per match? The most conceded goals per match? (these will require creating derived variables from the `goalkeepers` data)
4. Combine `distribution` and `attacking` data to see which players had the most assists per game and most passes completed per game. Compare these two lists to see if they are the same players or different. Basically, are players who assist the most goals also the best at passing in general during a game?
5. Which position covers the most distance on average? (using `key\_stats`)
6. Come up with more interesting questions that can be solved by joining more datasets.

Kinsey’s feedback/ideas:

* PROC REPORT will likely be the nicer choice for any tables you want to format
* any "Table 1" type of info that would be helpful to describe the 750 players?
* a lot of your questions would look nice as figures too