**European Soccer Datasets**

Serie A Dataset

This dataset contains data on all soccer matches in the top tiers of Italian football since 1934. In total there are 9 variables with 26, 164 matches. The data is not too messy but the FT variable will need to split into home goal and away goal. Managing teams moving up and down tiers will need to be handled. This dataset will need to be aggregated and joined with a dataset that describes the wealth of the teams.

Premier League Dataset

This dataset contains data on all soccer matches in the top tiers of English football since 1888. In total there are 194,000 matches recorded. The FT variable needs to be split for ease of use. A way to handle moving between tiers needs to be introduced. This dataset will be aggregated and joined with data that describes the wealth of each team.

La Liga Dataset

This dataset contains data on all soccer matches in the top tiers of Spanish football since 1928. In total there are 24,295 matches recorded. The FT variable needs to be split and thought needs to be put into handling movement between tiers. This dataset will be aggregated and joined to data that describes the wealth of teams.

Champs Dataset

This dataset contains dataset on all soccer matches in the champions league since 1955. The FT and HT variables need to split so that we can analyse goals easier. This dataset will be aggregated and joined to data that describes the wealth of teams.

Forbes top 20 clubs

<https://www.forbes.com/soccer-valuations/list/#tab:overall>

Web scraping will be used to collect the data for the top 20 richest clubs in football.

The data shows the most valued soccer clubs as measured by Forbes from 2007 to now. It will be used in conjunction with the other soccer dataset, to compare how the valuations influence the performance of the clubs. The data is mostly tidy, but needs to be joined by year, country and league.

**NBA**

The first type of dataset, “Performance” contains a teams overall performance in the 2018-19 season, record for each month, conference, division and overall record, while the second type of dataset, “Draft” contains data about the teams draft position and their corresponding draft picks with their college statistics

In the NBA there exists a salary cap and therefore most of the teams in the league spend about the same amount on their teams. There is wrangling to be done to make this data tidy such as gathering some of the variables in the first dataset while also joining the datasets together to analyse easier. In the first dataset as well due to the way the win loss record is shown, many of the observations in the table have been automatically switched to dates rather than the record format that is used and must be fixed up in order to analyse.