**Goal distribution between halves**

# Description

In this task we want to find out which proportions of all goals in a soccer game occurs in the respective halves. This is important since if we have a model giving us the expected number of goals in a game, at half time we need to know how many goals are expected to be scored in the remainder of the game.

For a rather large number of games where goals were scored, we have provided a number of variables.

The easiest possible thing to do is to simply disregard all variables and compute the percentage of goals occurring in the respective halves; for the attached data set this would be approximately 44% in first half and 56% in second half.

We are, however, interested in whether we can do better than this by using information about the game provided when it starts.

An example of this could be that we notice that a certain competition has a significantly different distribution (e.g. only 38% of all goals are scored in first half) than the remainder of the data.

The variables are:

* competitionIdx: a number identifying a given competition
* L1: the number of goals expected to be scored by team 1 throughout the entire game
* L2: the number of goals expected to be scored by team 2 throughout the entire game
* Attendance: a coarse description of the size of the crowd
* Weather: a coarse description of the weather observed at the start of the game
* PitchCondition: a coarse description of the condition of the pitch at the start of the game
* GoalsHalf1: The actual total number of goals scored in half 1
* GoalsHalf2: The actual total number of goals scored in half 2

Some things possible worth considering:

* Can some of the variables be disregarded entirely?
* Some competitions occur very rarely in the data set; can we group certain competitions or perhaps even disregard them?
* Do we need to “clean” the data by e.g. removing games having extreme values of L1, L2?
* Given that we conclude a group of games should have a goal distribution of its own, how do we find out whether this fact is due to chance/randomness or is likely to be systematic?

A final, important, remark is that there is no text book solution to this task, so feel very free to generate your own ideas.