Attached are 2 files: *samples.csv* and *test.csv*. *samples.csv* contains data for 3,000 football games, while *test.csv* contains similar data for another 3,000 games. The column headings are as follows:

id = game id
sup = the expected full time goal supremacy (home goals - away goals) implied by bookmaker prices
tot = the expected full time total goals (home goals + away goals)
ht_h = the actual number of goals scored by the home team at half time
ht_a = away at half time
ft_h = home at full time
ft_a = away at full time

We are interested in modelling the half-time/full-time market, which has 9 outcomes corresponding to the combinations of half-time and full-time outcomes: home win, away win and draw. So for example home-home (hh) is home team winning at half time, and still winning at full time; draw-away (da) is draw at half time, away team winning at full time, etc.. The full outcome set then is {hh,hd,ha,dh,dd,da,ah,ad,aa}.

Questions and tasks

1. The data provided give us an expected goal supremacy and total goals as described above. How can we convert these values into expected goals for the home team and away team, respectively?
2. Now that we have expected goals for each team, we want to construct a probability distribution describing the probability of each outcome (number of goals scored) for each team respectively. We want to use the Poisson distribution to accomplish this task; explain why this distribution is an appropriate choice for the type of data we are working with.
3. Suppose that for a particular match, the home team has an expected goals of 1.6, while the away team has an expected goals of 0.9. Assuming independence between the scores of the two teams, what is the probability that the match will finish 0-0?
4. If we can calculate the probabilities of individual *scores*, then we can also calculate the probability of *results* by combining the probabilities of relevant scores together. In the above example, what is the probability of the home team winning the match? (Consider only those scores where neither team scores more than 9 goals.)
5. Using the above, we can produce probabilities of scores and results at full time, but this does not tell us anything about the half time scores. How can we construct expected goals at half time for each team using the data provided in *samples.csv*?
6. For a given team (e.g. the home team), is the number of goals scored at half time independent of the number of goals scored at full time? Write down mathematical expressions to help explain your answer.
7. Using the example from question 3 and your answers from questions 5 and 6, calculate the probability of the home team leading 1-0 at half time and then winning 2-1 at full time.
8. In the last question we calculated the joint probability of a half time score and a full time score; let's call this a *double score*. Now we can apply similar logic to that used in question 4 to calculate the probability of a *double result* as required in the original brief. What is the joint probability of a draw at half time and the away team winning the game (da)?
9. Using your answers to questions 1-8, produce probabilities for all double result outcomes {hh,hd,ha,dh,dd,da,ah,ad,aa} for all matches in *test.csv*. Save your answers in a new file and call it *predictions.csv*.
10. Submit your answers to the above questions, your copy of *predictions.csv* and any scripts or other files you used to generate it.