DATA-437 In-Class Assignment November 10, 2021

Consider the `Teams.csv` data file. This data file contains seasonal stats for major league teams going back to the first professional season in 1871. We are interested in studying the relationship between wins and runs for recent seasons, so we focus our exploration on season since 2001. **In Python**, answer the following:

1. (3 points) Using pandas, read the csv file and create a data-frame called `teams`.
2. (4 points) Suppose that one is interested in relating the proportion of wins with the runs scored and runs allowed for all teams. Towards this goal, the relevant fields of interest in this dataset are the number of games played G, the number of team wins W, the number of losses L, the total number of runs scored R, and the total number of runs allowed RA. We create a new data-frame called `my_teams` containing only the above five columns plus the information on the team (teamID), the season (yearID), and the league (lgID).
3. (5 points) Compute the runs differential (`RD` = `R - RA`), winning percentage (`Wpct` = `W / (W` + `L)`), and the dummy variable `League` that takes the value 0 when `lgID = NL`, and 1 otherwise.
4. (6 points) Build a linear model in which `RD` and `League` are the predictor variables, and `Wpct` is the target variable. Is `League` significant? If not, remove it from the model. Compute the RMSE of this model.
5. (6 points) Using $k = 1.85$, predict the `Wpct` using the Pythagorean formula. Compute the RMSE of this model.
6. (3 points) What model would you use to predict `Wpct`? Be specific.
7. (4 points) Based on your answer for part (6), predict the `Wpct` of a team with `R = 730` and `RA = 750`.