**Premier League Dataset**

You have been provided with a dataset of premier league teams with corresponding data. The filename is **premier league.csv.** The function **colnames()** will give you the variables being investigated.

1. What is the average and standard deviation of the number of goals (*goals*) scored in the premier league over the season 2006 -2017?
2. Do teams that score more goals have more wins ? Draw a scatterplot of *wins* against *goals* to investigate this ?
3. “Attack is the best form of defence” has been a philosophy of many teams such as the Dutch national team who believe the number of attacks makes it harder for the opposing team to score goals. Let us test this hypothesis now. Plot the number of goals scored (*goals*) and total\_scoring\_att (*total\_scoring\_att*) against goal\_conceded (*goals\_conceded*).

Create a new dataset with the teams of **Manchester United, Liverpool, Arsenal, Tottenham, Chelsea**, and **Manchester City**. Select only the following variables: *team, season,* *wins, losses, goals, ontarget\_scoring\_att, total\_scoring\_att, total\_red\_card, total\_pass, goals\_conceded*. Call the dataset data1 .

1. Arsenal, Chelsea and Tottenham Hotspur are the three London teams. Draw a time series plot for the number of goals- (goals) and comment.
2. Among all the teams under consideration, which team seems to be more consistent in goal scoring. Draw a visualisation to find this out.
3. Create a new variable called “goal\_difference” and “league points”.

Goal\_difference = goals – goals\_conceded.

League\_points = (wins\*3) + (38 – wins-losses)

PS: The goal difference is the difference between the goals scored and goals conceded.

Total league points are usually computed as (Number of Wins\*3) + (Number of draws\*1)+(Number of losses\*0). In this dataset, the number of draws are absent. But we can work this out knowing that the total number of games = 38. Hence The number of draws = 38 – wins – losses.

Draw a scatterplot and explain your findings.

1. Plot a Box Plot for the number of passes (*total\_pass*) for each team in dataset A. Which team seems to be more entertaining? (Entertaining here is defined as teams having high number of passes).
2. On separate plots for each team, draw a scatter plot of goals scored (*goals*) against goal conceded (*goals\_conceded*).

Now we can consider the original dataset again

1. Draw time series plots for the logarithm total number of red cards (*total\_red\_card*) and total number of yellow cards (*total\_yel\_card*) for each season. (We use logarithm here because it is very likely that the total number of yellow cards will exceed the total number of red cards by much for each season which will make the scales uneven).

**Second Tutorial**: You have been provided with a dataset on football players in different leagues in Europe. Think of some interesting questions which visual techniques can help you answer. In this work, there is no right or wrong answer, but you are allowed to be as creative as much as you can.