Fall 2020: Math 660/461 - Introduction to Statistical Computing

Assignment 3 **Due: Sunday Oct 18, 2020**

General Instructions:

The data sets are available for download from Canvas as a zip file. Save the file and extract the datasets into your R working directory. Make sure the filenames are unchanged. Make a copy of the Rmarkdown template and name it according to this format: LastNameFirstName-Hw3.Rmd, and put it in your R working directory. You will submit this Rmd file only. If I run your Rmd, I should get a pdf for grading.

Your Rmd file should contain only the **edited** version of your final work, and not be a dump of all the work you did. Hence you should edit your assignment so that it is as concise as possible. Include only what is necessary for supporting your answers. Points will be deducted for handing in unnecessary code or for not editing your output. Use meaningful titles and names for your plots and for your axes.

If you use any material, e.g. books, papers, websites, that helped with your submission, they should be referenced and acknowledged. Such material should only be used to inform your work. If it is found that your work is very similar to work shown on the internet, or other resource, you will get a zero. If students' work are essentially the same, points scored will be divided by the number of students involved.

Information for Assignment 3:

The data for this assignment concerns the 2019-2020 Football season of the English Premier League. The data files are csv files:

standings-1920.csv - final ranking of the teams at the end of the season, with information such as points earned, games won, drawn, lost etc.

teamstats-1920.csv - statistics about the players at the team level

teamgoalkeeping-1920.csv - statistics about the goalkeepers at the team level

players-1920.csv - statistics about the individual players

goalkeepers-1920.csv - statistics about the individual goalkeepers

scoresfixtures-1920.csv - information about individual matches and their results.

Note 1: the *.txt files with the same names except with the added word "glossary", if available, provide some information about what the variables mean. The information provided may not be completely clear, but are the only information available from the website from which the data files were obtained.

Note 2: Use only the files provided on Canvas to do the assignment. I made some changes to some of the data files to make them easier to work with.

The questions below will ask you to extract information from the data files, create new tables, manipulate data, make some plots and so on. You should do it using the tidyverse packages (e.g. use read_csv instead of read.csv), as introduced in Lecture sets 5-7. There is NO modeling or statistical analyses needed.

Since this assignment is about using R code to manipulate data, have your Rmd file show the R code when generating the pdf file. If you are asked to create, for example, a list of away games for a particular team, all you need to do is have your R/Rmd show the R tibble output in your pdf. This means only 10 rows are shown and additional rows are hidden. That is ok. Unless you are specifically asked to keep only certain columns, your final tibble can have any number of columns, as long as the answer is shown.

1. Answer the following:

   (a) How many teams are there?

   (b) How many players are there?

   (c) What is the total number of goals scored?

   (d) What is the average attendance (spectators) at the games?

2. Identify the primary key(s) (if any) for each table. Verify your answers. Note: ignore the "Rk" variable, as it usually just represents the row number.

3. Find a table with a foreign key(s), and write down the name of the table, the foreign key(s) and the associated table(s).

4. For the players and goalkeepers data sets, the "Player" variable has two versions of the player name. Separate this out and keep ONLY the first version of the name. The "Nation" variable also has two abbreviations listed for the nationalities of the players. Separate the values out into two variables (and keep both of them). Assign the resulting dataframes/tibbles to new names different from the original ones.

5. Find, for each team, the mean player age, the age of the youngest player and the age of the oldest player. Find the name of the youngest player for each team. Your output for the latter should only contain the team name, the minimum age, and the name of the player.

6. Add the team performance numbers (specifically, Possession, Assists, Penalty Kicks scored, Save percentage, and Clean sheet percentage) to the standings table. Add also the age statistics found in Q5 above to the standings table. Then make the following plots (A vs B means A is on the y axis):

   (a) Points scored vs Possession

(b) Points scored vs Save percentage

(c) Points scored vs Goal against

(d) Clean sheet percentage vs Save percentage

7. Identify the names of the top 3 referees that refereed the most games. Your R output should only show three referees. Obtain a list of the games they refereed. Your output for this should only contain columns for the week and date the game was played, name of the home and away teams, and the name of the referee.

8. Obtain a list of the games played by the top 3 scoring teams (i.e. the 3 teams that scored the most goals), showing only week, date, home/away teams and the result.

9. Find the 3 days with the most games played. Your R output should only show three days. Then obtain the list of games that were played on these days (you can keep all the columns).

10. Find the 3 days with the most goals scored. Your resulting tibble should only have 3 rows.