Lab 1

Due on 02/02 at 11:59 pm

Instructions: This lab report needs to be professional. Only report relevant and finalized code. Your writing should be concise and void of spelling errors. Use code chunk options to hide unnecessary messages/warnings. Your report should be reproducible. Reports that involve simulations need to have the random seed specified so that simulation results are reproducible. You are allowed to work on this lab assignment in groups of 2-3. You still need to submit an individual lab report if you do work in a group, and you need to list your collaborators.

Question 1 In lecture it was demonstrated that baseball is a game of offense, pitching, and defense with a regression model that considered expected run differential as a function of explanatory variables OPS, WHIP, and FP. Do the following:

- Fit a similar regression model with runs as the response variable. Report problems with this model. Investigate problematic residuals to discover what went wrong. Fix the problem with this model by adding categorical variable(s) to the list of explanatory variables. Briefly explain what went wrong.
- We can significantly improve the regression model in the notes through a principled rescaling of OPS, WHIP, and FP. Split the Teams data frame by `yearID` and, for each year, create variables `OPSscale = OPS/avgOPS`, `WHIPscale = avgWHIP/WHIP`, and `FPscale = FP/avgFP` which require you to first create league average variables `avgOPS`, `avgWHIP`, and `avgFP`. Fit the linear regression model with runs differential as the response and explanatory variables `OPSscale`, `WHIPscale`, and `FPscale`, and report relevant output. Why does this model perform so much better than the model in the notes? Support your answer. Hint: functions `split`, `do.call`, and `lapply` are useful.

Question 2 Choose 3 batters and 3 pitchers that have played in at least 10 seasons and do the following:

- Display the seasonal statistics for these players. The following statistics should be included for batters (derivations of unconventional statistics are in parentheses): year, G, AB, R, H, X2B, X3B, HR, RBI, SB, CS, SBpct (SB / (SB + CS)), BB, SO, OBP, SLG, OPS. The following statistics should be included for pitchers: year, W, L, IPouts, H, ER, HR, BB, HBP, SO, ERA, WHIP, SOper9 (SO / IP * 9), SOperBB (SO / BB). These statistics can be found in or computed from statistics that are found in the `Batting` and `Pitching` dataframes in the `Lahman` package.
- Create career stat lines for each of the players that you selected. Be careful about how these statistics are calculated.
- Provide a plot for career trajectories for one batting and one pitching statistic of your choice. These are two separate graphics, one for the batters and one for the pitchers. The graphics that you produce should display the trajectories of the 3 batters and the 3 pitchers. Provide interesting commentary on your graphic.

Question 3 Problem 2 on page 28 of Analyzing Baseball Data with R

Question 4 Problem 3 on page 29 of Analyzing Baseball Data with R