HOMEWORK #6 Due **Tues OCT 10**

INSTRUCTIONS:

A) *For each answer,* include a:
   1. summary output table and/or graphs, as appropriate
   2. short statement about how you handled assumptions and those outcomes
   3. short answer that clearly answers the question, based on the results.

B) Provide your code in an Appendix, organized so that we can relate it to questions

C) Submit a pdf (with your name in the file name).

It's Baseball Playoffs Season! Go Guardians! So here's a baseball data set, from 2022. MLBbattingdata2022.csv lists batting statistics for the top 130 baseball players in the 2022 season. Variables listed for each batter are:

- GP = games played
- AB = at bats
- R = runs scored
- H = hits
- AVG = batting average
- HR = home runs
- OBP = on base percentage

1. [3 pts] Since scoring runs is the reason for batting, which *single variable* above most plausibly predicts runs scored (R) by a batter?

2. [1 pt.] Show the model for that most plausible variable (from above) and report its coefficient of determination.

3. [1 pt.] Discuss how well model assumptions were met, and show evidence to back up your argument.

4. Now get creative: what *combination* of multiple predictors most plausibly predicts runs scored Runs Scored *AND* meets model assumptions? In other words, use multiple combinations of predictors from above to find the most plausible model to predict R. Specifically,
[2 pts.] explain your rationale for your hypothesized predictor combinations, and
[1 pts.] show evidence for your selection of a most-plausible model, and
[1 pts.] show the output for that model, and
[1 pts.] show how well that model met regression assumptions.