STAT 1341
Midterm Project
Due: November 7, 2021

Project Overview

Your midterm project is an open-ended project where you (and one partner if you choose) will select a unique topic or question of interest to you in the field of sports analytics to research, analyze, and answer. The final product will consist of a written report, the R script containing the code, and any data used to analyze the question. These materials must be submitted no later than November 7, 2021 on Canvas. A list of suggestions for possible research topics is presented on the third page. You may choose one of these suggestions, expand or modify one (or more) of them, or pick a topic of your own choosing. Students who have previously taken STAT 1223 and used a sports related topic for that project may not reuse the same idea for this project.

Project Proposal

Prior to beginning your project, you must first submit a proposal of your plan that is no longer than one double-spaced page. This must include the following information:

- The research question you hope to answer
- A description of why you believe you research question is relevant/important
- Where you are going to obtain your data from

Do not submit this proposal until you are certain that you can obtain the data that you need! Do some research into where you can find the data online or attempt to reproduce/modify some of the code provided in class. You may be able to make use of some datasets I provide you with in class, but you may also need to collect your own.

Students who are working in pairs need to submit only one copy of this proposal between them with both names included. Once your proposal has been graded and approved, you may begin working on the actual project. If I believe that your proposed project is lacking, we will need to discuss how to revamp the research question to make it more rigorous. The project proposal is due by **Friday, October 15, 2021**; however, the earlier you submit the proposal, the earlier I can grade it and the earlier you can get started on the actual project.

Project Requirements

Your paper must contain the following subsections:

- **Introduction**: Provide some background information about why your topic is relevant and why sports teams or sports statisticians could benefit from knowing this research. Describe any other research you know of that has been done in this area. Make sure you include the research question that you intend to answer.
- **Data Collection**: Include a description of how you obtained your final dataset used in the analysis. (Did you find an existing dataset online? Scrape the data using R? Use a dataset from class?) Specify any data cleaning you had to do. (Removal of incomplete observations, removal of extraneous variables, combining multiple datasets into one, etc.)
- **Descriptive Statistics**: Summarize your data graphically and numerically. Include graphs (histograms, scatterplots, etc.) and tables of summary statistics (mean, standard deviation, median, etc.). If you are modeling, pay careful attention to the response variable. Discuss what you see in the graphs and summary statistics. Every graph or table you insert should be discussed in the text of your paper, even if just for a sentence or two.
- **Inferential Statistics**: Answer your research question. Provide a thorough analysis of your data using inferential statistics. Describe any methods that you used (multiple regression, logistic regression, model selection, chi-squared analysis, etc.), and include the model if applicable. Include test statistics, p-values, confidence intervals, etc. Consider analyzing the residuals and exploring unusual observations or results.

- **Discussion/Conclusion**: Explain how what you learned may be relevant to the sport. (What are the practical implications that players/coaches/the front office could learn? Are your results generalizable to other leagues or other sports? If so, then to what extent?) Include a paragraph on any limitations of your project and ideas for future research. (Were there things you simply couldn't accomplish given the time and resources? Where could this project go from here?)

Your R code must:

- Run with no errors (Warnings are fine)
- Be readable and organized with no excess lines of unnecessary code
- Include clear comments describing what each line/chunk does

The dataset(s) that you submit must:

- Contain all necessary values needed for the code to run
- Not contain extraneous columns of data

Paper Formatting and Length

If you are working independently, the final paper must be between 1500 and 2000 words. The paper for groups of two must be between 2500 and 3000 words. (You can exceed these upper limits by a bit if need be, but be sure to hit the minimum word count.) Do not include any R code in your paper. Make sure the paper is double-spaced and use either Arial or Times New Roman 12 pt. font.

Grading

Your midterm project will be graded according to the following criteria:

Component	Points
Proposal	**5**
Data	**5**
R Code	**15**
Reproducibility	5
Readability	5
Documentation	5
Paper	**75**
Introduction	10
Data Collection	5
Descriptive Statistics	15
Inferential Statistics	20
Discussion/Conclusion	15
Grammar, Formatting, and Word Count	5
Overall Impression and Quality	5
Total	100

When you are ready to submit, upload to Canvas (1) the final paper, (2) the R code, and (3) the dataset(s) you used to Canvas. As with the proposal, students working in pairs need make only one submission that includes the paper, R code, and datasets, being sure to include both partners' names on the final product. Both partners will receive the same grade.

Project Ideas

- How has home field/court/ice advantage changed over time? Did COVID impact the home team's advantage?
- Are pitchers predictable in the types of pitches they throw?
- Should teams take advantage of the 2-for-1 in basketball?
- How has the addition of the "loser" point in hockey impacted the way teams approach the game?
- What is the probability that a player will be elected to the Hall of Fame in a given sport?
- Do make-up calls exist in MLB/NFL/NBA/NHL?
- Can an expected points model accurately predict player performance in the NBA?
- How has the shift impacted baseball?
- Who is the worst pitcher to throw a no-hitter/perfect game?
- What impact does aging have on an athlete in a given sport?
- Is there a difference in expected goals models in different soccer leagues?
- Does referee bias exist?