PS6

Colton Steele

March 20, 2023

1 Answers

Data Cleaning and Transformation: The data I scraped was fairly clean in the sense that there weren't errors in the data as I scraped it using softballR which just scrapes the NCAA website. However, I did have to transform the data in a few ways. After scraping the game level data into a data frame, I ran a loop to calculate the win count and loss count for each team along with the runs for, runs against, and run differential. Additionally, I calculated the winning percentage of all the opponents for each team in my data set to approximate a strength of schedule value. The end goal with all of this is to create an ELO rating for college softball as it's something I haven't ever seen. I think it would be interesting to create it and see how my ELO system fares against actual results.

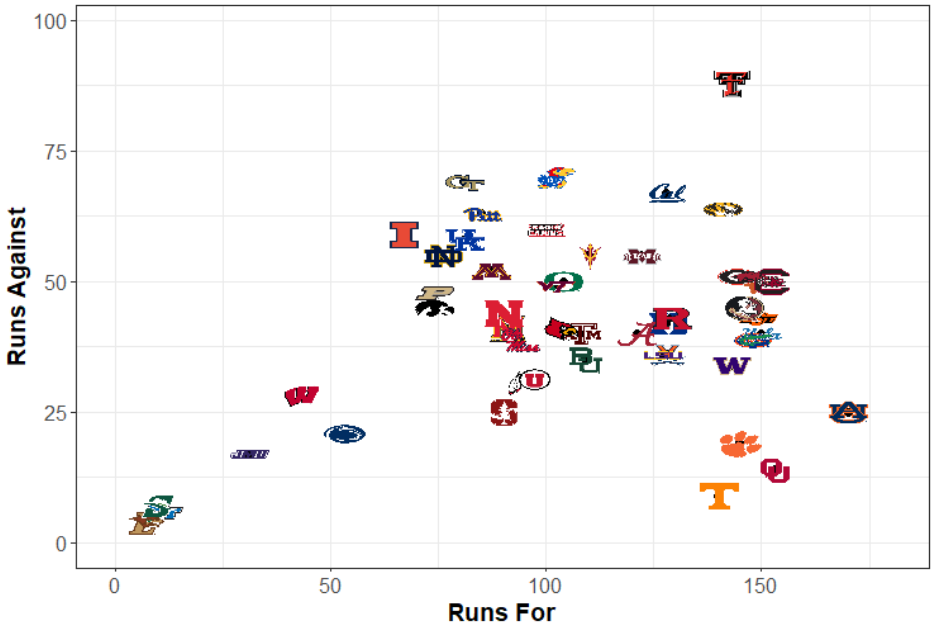

Figure 1: Run Differential

This image shows the top 25 teams in the country for NCAA softball by run differential with their university logo representing their dot on the scatterplot. The further right a team is, the more runs they score as a team. The further up they are, the more runs they give up. Thus, the further down and right a team gets, the better their run differential is while up and to the left means the worse their run differential is.

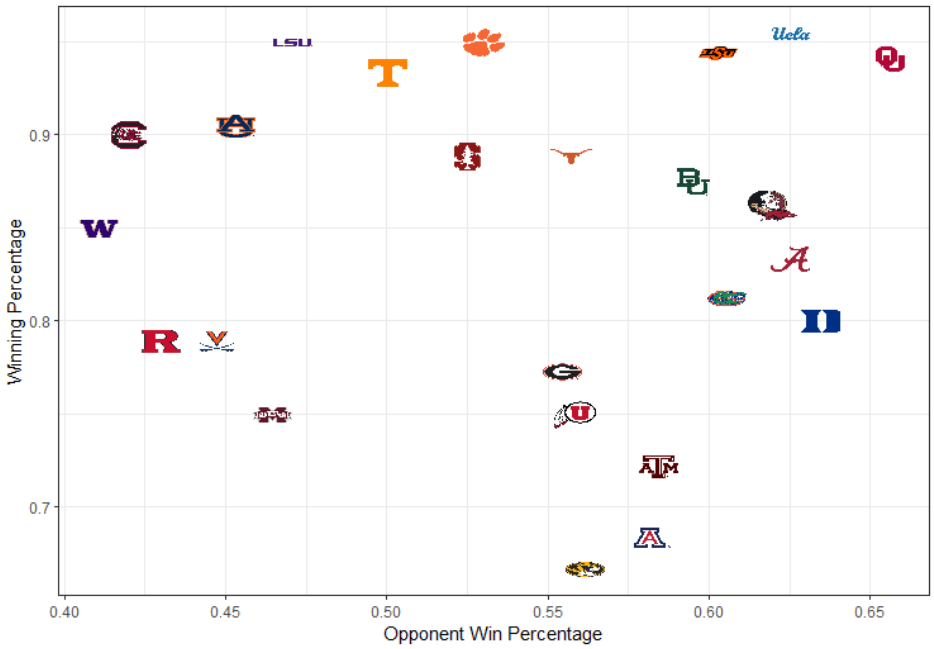

Figure 2: Strength of Record

This image portays the top 25 teams (still by run differential) and compares their winning percentage to the cumulative winning percentage of all of their opponents on the season. This image gives us a good idea of whether a teams win percentage is high because they are a genuinely good team or if they have had a fairly easy schedule up to this point. Essentially, this graph is attempting to portray each teams strength of record.

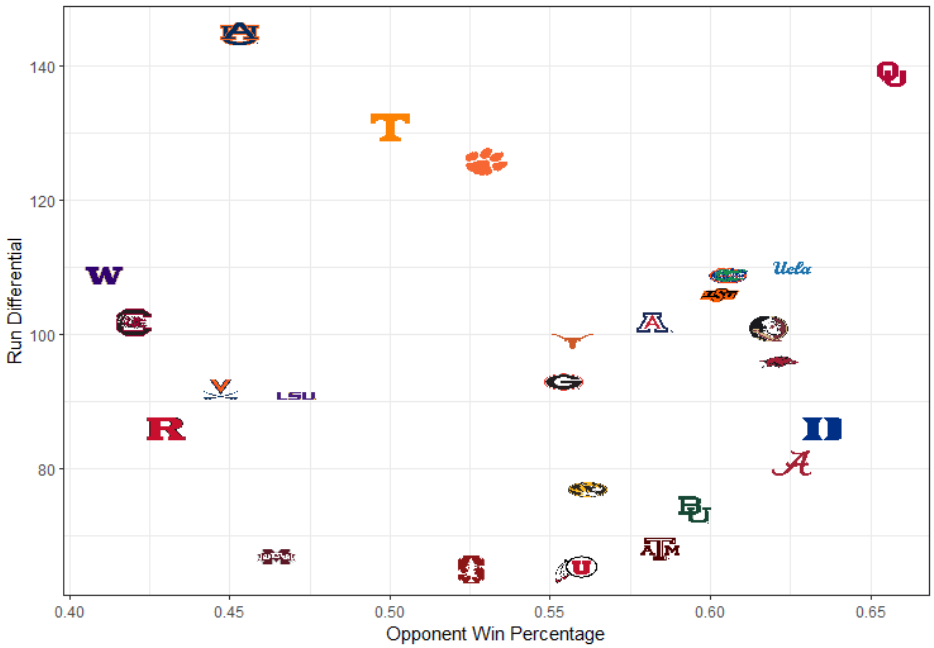

Figure 3: Run Differential by Opponent Difficulty

This image plots the run differential of the top 25 teams by run differential against the winning percentage of their opponents. This image, similar to the previous one, shows how strong a teams run differential is and if they have just faced low quality opponents.