Tentative Baseball project

Baseball player database: Using web scraping from baseball-reference.com, get the details of all top 20 hitters from each team each year from 2001 to 2020 grabbing all their batting data. Make a database that allows for all kinds of graphs and tables showing the top or bottom players under a category or a multitude of categories. The idea is to make transformative queries that bypass what is currently possible on baseball reference. (Example: Who has the most rbis given a minimum of x games played in the AL, not including the Yankees and Red Socks only looking at 2004-2007 and 2011.) High priority on making an easy to use interface for users. Also implement direct player comparisons of up to 10 players a time. Do this with direct player search and fuzzy search to catch spelling errors or players with the same name. Allow for specific statistical comparisons and bring in multiple graphs to back up comparisons. (Example: Compare # of average RBIs of Alex Rodriguez, Manny Machado, Mike Trout, and Pedro Martinez. This would show line graphs of RBIs over their years played in the tracked time period, the total RBIs in a table, and box and whisker plots for each player.) Potentially, I will create a new database that tracks and adds 2021 baseball data and then at the end of the year merge it with the above database.

Baseball team records: Using web scraping, I will grab team records of all teams from 2001 to 2020 and add it to a separate database. I will then look for trends in win/loss numbers over the years independent of teams. I will then analyze any outliers, looking through historical evidence of why some teams may have done so much better or worse than the average win rate for the x ranked team in the division. I will back up this analysis with player data from my player database if applicable. Then I will do some sort of predictive analysis to find the expected win rates of the best team in AL, NL etc.

Machine learning aspect: I will then web scrape several lists of prospect rankings over the years of my player database. My goal will be to find out who is the best and most accurate ranker at figuring out who will make it into the MLB. I will also sum up lists and try to find out from a generic prospect list, how likely any prospect is to make it into the MLB given their rankings across multiple rankers lists. I will then try to break down what rankers excel at in predictions. (Example: X ranker excels at predicting catchers. X also overvalues players coming from X school.)