Problem Set 5

Beth Felkner

February 29, 2024

1 Task 1: Scraping from a website with no API

This dataset is about the 2023 MLB Hall of Fame Inductees. It includes data both about the Hall of Fame voting (number of votes received, number of years on the ballot, etc) for each player and about career statistics for each player (WAR, RBIs, ERA for pitchers, etc.) I think it will be useful to me on my project for this class, because I am thinking I want to do something about machine learning prediction for which players will or won't get into the HoF. I used the same online tutorial we used in class, from Grant McDermott, as well as consulting the script we'd created together during that class period (with the 100 meters data from Wikipedia).

2 Task 2: Scraping from a website with an API

I used the MLB website, which has a hidden endpoint API like the rugby examples in Grant McDermott's tutorial, and I chose specifically to look at the page for one of my favorite players who is also a 2024 HoF inductee, Adrian Beltre. I followed McDermott's example of how to find the API from the Network pane of the Inspector. I used the jsonlite package to get the data into R. When I first pulled the data into R I got a very complex multi-layer nested list with tons of data, but with help from the tutorial and your office hours I was able to extract a useful dataframe that has yearly stats across his career. One thing that jumped out to me and I found interesting when looking at the data is that Beltre only played 124 games during our 2011 World Series run season. Looking it up confirms that he was battling hamstring injuries throughout the season, but I do not recall that in my memories of the season (although I was of course only 10 so I gues my memory of it is shaky), and it is especially impressive to me that we made the World Series with our star player out for much of the year.