Problem Set 4

William Miller

February 18, 2025

1 Data I'm interested in scraping

I am interested in scraping pro tennis data from the ATP tour. I grew up playing competitive tennis and am very interested in statistically evaluating the strategy tips I received when I was playing junior tournaments. I was given advice like your first serve percentage is more important than your number of aces/service winners, your performance on "momentum points" (most often at the 30-15 score) was a key determinant of success, that your number of approaches and volleys was important for maintaining a lead, and that limiting unforced errors is more important than hitting winners. Because of the data revolution that has recently taken tennis by storm, there are many websites I could use to procure this data. The official ATP tour forbids web scraping in their official terms of service and does not provide a public API, but there are numerous services that do. The one that looks most appealing is called Tennis Abstract. It appears to be free and currently has data on over 15,000 matches and 2 million individual points.

2 Questions from previous problems

- Question 5 (R exercise part 1)
  - Calling `class(mydf)` the following: `"tbl_df" "tbl" "data.frame"`
  - Calling `class(mydf$date)` returns `"character"`
- Question 6 (R exercise part 2)
  - After many attempts I was unable to figure out how to get sparklyr to install, even after including Java 11 in my bash profile script. I have submitted some code that I believe would work (I tested it on my computer) and that I hope is satisfactory to complete this assignment. I still think the issue might be the "cli" package but I am not totally sure.
  - To answer the questions raised in problem 6, the class of df1 is a tibble/dataframe, and I believe df would be a sparkdataframe, but I cannot confirm this. For the second part, from what I can gather, spark clusters are case insensitive and replace all special characters with "_" which is why the columns change from Sepal.Length to Sepal_Length