SAL 413 Homework 3

Nolan Pittman, Hunter Geise, Aidan Myers, Ari Glazier

2023-10-18

Overview and Instructions

This problem set is intended to exercise your coding skills. Your styling must conform to the tidyverse style guide. This will be a group assignment. However, you must work only with the people within your group. A group may not seek help from peers in other groups. You may use whatever notes, textbooks, etc. you find helpful for the assignment. Your code should consist of functions found in rvest, httr, stringr, purrr, or very basic R functions. Do not use other tidyverse functions that we have not gone over in class unless directed to do so in the homework. Please compile your document into a PDF and upload the PDF to Blackboard with the filename in the following format: [surname]_hw3.pdf. For example, I would turn in maddox_hw3.pdf. Please include the questions, your code, output, and responses for each question in the pdf. Please start each new question on a new page

Questions

This homework will work on web scraping in R using rvest. Each question will build on the results from the previous question. Read each question carefully as steps to solving the question may be laid out in the question itself.

1. Go to ESPN's website and find the Syracuse men's basketball schedule for the 2022-23 season. Scrape the schedule to find the game IDs for each of the games Syracuse played that season. (The game ID is the 9 digit number at the end of the url link for each game. The first game of the season was against Lehigh, and the game ID is 401482974.) Print out a vector of the game IDs.

```
SUMBB_sch <- str_c("https://www.espn.com/mens-college-basketball/team/schedule/",
"_/id/183/season/2023") %>%
  read_html()

syr_xml <-SUMBB_sch %>%
  html_nodes("a") %>%
  html_attr("href")

syr_xml[str_detect(syr_xml,
                   "https://www.espn.com/mens-college-basketball/game/_/gameId/")] %>%
  na.omit() -> gameids

gameids <- str_extract(gameids, "gameId/\\d+") %>%
  str_replace("gameId/", "")

print(gameids)
```

```
##  [1] "401482974" "401482975" "401482976" "401482977" "401486676" "401482978"
##  [7] "401479672" "401488391" "401482979" "401482980" "401482981" "401482982"
## [13] "401488399" "401500552" "401488407" "401488419" "401488427" "401488431"
## [19] "401488435" "401488446" "401488452" "401488462" "401488463" "401488474"
## [25] "401488483" "401488494" "401488500" "401488509" "401488516" "401488522"
## [31] "401488532" "401514041"
```

2. Write a function that reads in a game ID and will return the box score from that game. Make sure that you include both the home team and away team stats in the returned box score. Ideally what is returned is a list of 2 box scores, one from the home team and one from the road team. (It may also help to include as a third element of the list an indicator of whether Syracuse is the home team for each game at this step. Do not hardcode this in.) The general form of the box score url is: https://www.espn.com/mens-college-basketball/boxscore/_/gameId/(game ID #). You may use the dplyr function bind_cols() to combine the table of player names to the table of player stats. You may need to scrape once for the names table then a second time for the stats table. Also, use the header = TRUE argument within html_table() in order to force the top row of the table into column names.

```
box_scores <- function(id) {
  box_score <- glue(str_c("https://www.espn.com/mens-college-basketball/boxscore/",
  "_/gameId/{id}"))

  boxscore_xml <- read_html(box_score)

  # Determining Syracuse home or not
  SU_home <- boxscore_xml %>%
    html_nodes("table") %>%
    html_table %>%
    pluck(1)

  colnames(SU_home)[colnames(SU_home) == ""] <- "Team"

  SYR_home <- tail(SU_home$Team, 1) == "SYR"

  if (SYR_home) {
    game <- "Home"
  } else {
    game <- "Away"
  }

  # Box Scores
  boxscore_xml %>%
    html_nodes(".Boxscore__ResponsiveWrapper
                .Wrapper:nth-child(1) .Table__Scroller .Table--align-right") %>%
    html_table(header = TRUE) -> away_boxscore_stats

  boxscore_xml %>%
    html_nodes(".Boxscore__ResponsiveWrapper
                .Wrapper:nth-child(1) .Table--fixed-left") %>%
    html_table(header = TRUE) -> away_boxscore_names

  away_full_boxscore <-
    bind_cols(away_boxscore_names, away_boxscore_stats)

  boxscore_xml %>%
    html_nodes(".Wrapper+ .Wrapper .Table__Scroller .Table--align-right") %>%
    html_table(header = TRUE) -> home_boxscore_stats

  boxscore_xml %>%
    html_nodes(".Wrapper+ .Wrapper .Table--fixed-left") %>%
    html_table(header = TRUE) -> home_boxscore_names
```

```
  home_full_boxscore <-
    bind_cols(home_boxscore_names, home_boxscore_stats)

  return(list(away_boxscore = away_full_boxscore,
              home_boxscore = home_full_boxscore,
              Syracuse = game))
}
```

3. Map your function onto each of the game ID obtained in Quesiton 1. The result should be a list of box scores. Print off the first game's box score. No need to print off the box score for every game.

```
scorebox <- map(gameids, box_scores)
scorebox[[1]]
```

```
## $away_boxscore
## # A tibble: 19 x 14
##    starters    MIN   FG    `3PT` FT    OREB  DREB  REB   AST   STL   BLK   TO
##    <chr>       <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 "D. Paroli~ "16"  2-4   0-0   3-4   "0"   "0"   "0"   "0"   "0"   "0"   "3"
##  2 "JT Tan C"  "26"  3-6   0-0   0-0   "3"   "5"   "8"   "3"   "0"   "0"   "1"
##  3 "K. Higgin~ "27"  4-12  1-6   0-0   "0"   "2"   "2"   "3"   "1"   "0"   "1"
##  4 "T. Whitne~ "25"  1-4   0-1   4-4   "0"   "4"   "4"   "2"   "0"   "0"   "4"
##  5 "E. Taylor~ "31"  8-15  4-5   0-0   "2"   "1"   "3"   "1"   "2"   "2"   "0"
##  6 "bench"     "MIN" FG    3PT   FT    "ORE~ "DRE~ "REB" "AST" "STL" "BLK" "TO"
##  7 "B. Momah ~ "6"   2-3   0-0   1-1   "0"   "0"   "0"   "0"   "0"   "0"   "0"
##  8 "B. Chebuh~ "2"   0-0   0-0   0-0   "0"   "0"   "0"   "1"   "1"   "0"   "0"
##  9 "J. Alamud~ "16"  3-6   0-1   2-2   "0"   "2"   "2"   "0"   "0"   "0"   "1"
## 10 "T. Connif~ "1"   0-1   0-0   0-0   "0"   "0"   "0"   "0"   "0"   "0"   "1"
## 11 "H. Adiass~ "13"  0-2   0-0   0-2   "1"   "2"   "3"   "0"   "0"   "0"   "0"
## 12 "B. Reed G" "1"   0-0   0-0   0-0   "0"   "0"   "0"   "0"   "0"   "0"   "0"
## 13 "B. Knostm~ "3"   0-0   0-0   0-0   "0"   "0"   "0"   "1"   "0"   "0"   "0"
## 14 "J. Saigal~ "2"   1-1   1-1   0-0   "0"   "0"   "0"   "0"   "0"   "0"   "0"
## 15 "J. Sincla~ "16"  0-3   0-2   0-0   "0"   "0"   "0"   "3"   "1"   "0"   "0"
## 16 "J. Betlow~ "4"   1-2   1-2   0-0   "0"   "0"   "0"   "1"   "1"   "0"   "0"
## 17 "R. Fenton~ "11"  2-4   1-3   0-0   "0"   "0"   "0"   "1"   "1"   "0"   "1"
## 18 "team"      ""    27-63 8-21  10-13 "8"   "18"  "26"  "16"  "7"   "2"   "12"
## 19 ""          ""    42.9% 38.1% 76.9% ""    ""    ""    ""    ""    ""    ""
## # i 2 more variables: PF <chr>, PTS <chr>
##
## $home_boxscore
## # A tibble: 15 x 14
##    starters    MIN   FG    `3PT` FT    OREB  DREB  REB   AST   STL   BLK   TO
##    <chr>       <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 "C. Bell F" "16"  2-4   1-2   0-0   "0"   "0"   "0"   "0"   "1"   "0"   "0"
##  2 "B. Willia~ "23"  1-4   0-1   0-2   "0"   "3"   "3"   "0"   "0"   "0"   "2"
##  3 "J. Edward~ "28"  5-8   0-0   8-10  "2"   "9"   "11"  "1"   "2"   "4"   "0"
##  4 "J. Mintz ~ "26"  6-9   0-0   4-4   "0"   "0"   "0"   "3"   "1"   "0"   "3"
##  5 "J. Girard~ "25"  5-10  4-6   5-5   "0"   "2"   "2"   "4"   "0"   "0"   "3"
##  6 "bench"     "MIN" FG    3PT   FT    "ORE~ "DRE~ "REB" "AST" "STL" "BLK" "TO"
##  7 "M. Brown ~ "13"  3-5   0-0   1-2   "1"   "2"   "3"   "0"   "1"   "0"   "0"
##  8 "J. Ajak F" "4"   1-1   0-0   1-3   "0"   "0"   "0"   "1"   "0"   "0"   "0"
##  9 "P. Carey ~ "4"   0-0   0-0   0-0   "0"   "1"   "1"   "0"   "0"   "1"   "0"
## 10 "M. Hima C" "8"   1-1   0-0   0-0   "2"   "0"   "2"   "0"   "0"   "2"   "0"
## 11 "J. Taylor~ "20"  1-7   0-1   0-0   "1"   "3"   "4"   "0"   "3"   "0"   "0"
## 12 "Q. Copela~ "11"  2-2   0-0   2-2   "1"   "1"   "2"   "1"   "0"   "0"   "3"
## 13 "S. Torren~ "22"  3-4   1-2   3-3   "0"   "6"   "6"   "3"   "0"   "0"   "2"
## 14 "team"      ""    30-55 6-12  24-31 "10"  "29"  "39"  "13"  "8"   "7"   "13"
## 15 ""          ""    54.5% 50.0% 77.4% ""    ""    ""    ""    ""    ""    ""
## # i 2 more variables: PF <chr>, PTS <chr>
##
## $Syracuse
## [1] "Home"
```

4. Create a table including every player that played for Syracuse along with the number of points they scored for the whole season (NOT per game number). Order the players from the largest scorer to the smallest scorer. Obtain this from the data collected in Question 3. Do not simply scrape the season stats page.

```
hbox <- data.frame()
abox <- data.frame()

for (i in seq_along(scorebox)) {
  syr <- scorebox[[i]]$Syracuse
  if (syr == "Home") {
    hbox <- rbind(hbox, scorebox[[i]]$home_boxscore)
  } else {
    abox <- rbind(abox, scorebox[[i]]$away_boxscore)
  }
}

syr_players <- rbind(hbox, abox)

syr_players <-
  syr_players[!(syr_players$starters %in% c("bench", "",
                                            "team")), ]

players <- split(syr_players, syr_players$starters)

get_PTS <- function(player) player$PTS

PTS_list <- map(players, get_PTS)

total_points_df <-
  map_dfr(names(PTS_list), function(player) {
    tibble(Player = player,
           Points = sum(as.numeric(PTS_list[[player]])))
})

total_points_df[order(-total_points_df$Points), ]
```

```
## # A tibble: 17 x 2
##    Player          Points
##    <chr>            <dbl>
##  1 J. Girard III G    526
##  2 J. Mintz G         521
##  3 J. Edwards C       463
##  4 B. Williams F      215
##  5 C. Bell F          199
##  6 M. Brown F         165
##  7 J. Taylor G        122
##  8 S. Torrence G       74
##  9 Q. Copeland G       42
## 10 M. Hima C           27
## 11 J. Ajak F           14
## 12 A. Clayton G         2
## 13 N. Ruffin G          2
## 14 A. Cordes G          0
```

```
## 15 P. Carey C           0
## 16 S. Feldman G         0
## 17 S. Keating F         0
```