Baseball Data Scraping

Noah Jones

10/12/2021

Part 1. Scrape baseball-reference.com with rvest

```
library(rvest)
library(dplyr)
# starting page
teampage <- read_html("http://www.baseball-reference.com/teams/")
teams <- teampage %>%
  html_nodes(".left a") %>%
  html_text()

session <- html_session("http://www.baseball-reference.com/teams/")

# create a table called baseball that contains all of the teams' franchise histories

baseball <- data.frame()
for(i in seq_along(teams)) {
  team_name <- teams[i]
  team_history <- session %>%
    follow_link(team_name) %>%
    read_html() %>%
    html_node("#franchise_years") %>%
    html_table()
  team_history$GB <- as.character(team_history$GB)
  team_history$current_name <- team_history$Tm[1]
  if(i==1) {
    baseball <- team_history
  } else {
    baseball <- full_join(baseball, team_history)
  }
}


# at the end, be sure to print out the dimensions of your baseball table
dim(baseball)
```

```
## [1] 2744   22
```

```
# also print the head of the table
head(baseball)
```

```
##   Year                   Tm      Lg   G  W   L Ties  W-L% pythW-L%   Finish
## 1 2021 Arizona Diamondbacks NL West 162 52 110    0 0.321    0.377 5th of 5
## 2 2020 Arizona Diamondbacks NL West  60 25  35    0 0.417    0.458 5th of 5
## 3 2019 Arizona Diamondbacks NL West 162 85  77    0 0.525    0.541 2nd of 5
## 4 2018 Arizona Diamondbacks NL West 162 82  80    0 0.506    0.533 3rd of 5
## 5 2017 Arizona Diamondbacks NL West 162 93  69    0 0.574    0.594 2nd of 5
## 6 2016 Arizona Diamondbacks NL West 162 69  93    0 0.426    0.424 4th of 5
##     GB        Playoffs   R  RA Attendance BatAge PAge #Bat #P
## 1 55.0                 679 893  1,043,010   28.9 28.5   64 41
## 2 18.0                 269 295              29.1 27.7   45 26
## 3 21.0                 813 743  2,135,510   28.7 28.6   45 27
## 4  9.5                 693 644  2,242,695   29.2 29.6   49 30
## 5 11.0 Lost NLDS (3-0) 812 659  2,134,375   28.3 28.7   45 23
## 6 22.0                 752 890  2,036,216   26.7 26.4   50 29
##            Top Player           Managers         current_name
## 1     E.Escobar (2.3) T.Lovullo (52-110) Arizona Diamondbacks
## 2      Z.Gallen (2.5)  T.Lovullo (25-35) Arizona Diamondbacks
## 3       K.Marte (6.9)  T.Lovullo (85-77) Arizona Diamondbacks
## 4 P.Goldschmidt (5.5)  T.Lovullo (82-80) Arizona Diamondbacks
## 5 P.Goldschmidt (6.3)  T.Lovullo (93-69) Arizona Diamondbacks
## 6      J.Segura (6.4)     C.Hale (69-93) Arizona Diamondbacks
```

Some light text clean up

```
## [1] "Lengths (21, 20) differ (comparison on first 20 components)"
## [2] "13 element mismatches"
```

```
## [1] TRUE
```

Part 2. dplyr to summarize the baseball data

```
# Printing a summary table of our scraped data

baseball_summary <- baseball %>%
  filter(Year %in% 2001:2020) %>%
  group_by(current_name) %>%
  summarise("Wins" = sum(W), "Losses" = sum(L), "Runs" = sum(R), "Runs Allowed" = sum(RA), "Win Pct" = 
  arrange(desc(‘Win Pct‘))
print(baseball_summary, n=30)
```

```
## # A tibble: 30 x 6
## # Groups:   current_name [30]
##    current_name           Wins Losses  Runs ‘Runs Allowed‘ ‘Win Pct‘
##    <chr>                 <int>  <int> <int>          <int>     <dbl>
##  1 New York Yankees       1832   1303 16187          13838     0.584
##  2 St. Louis Cardinals    1747   1388 14767          13081     0.557
##  3 Los Angeles Dodgers    1738   1400 14042          12468     0.554
##  4 Boston Red Sox         1731   1406 16249          14303     0.552
##  5 Atlanta Braves         1675   1460 14319          13274     0.534
##  6 Oakland Athletics      1674   1463 14469          13296     0.534
##  7 Los Angeles Angels     1666   1472 14604          13838     0.531
##  8 Cleveland Indians      1616   1520 14772          14113     0.515
##  9 San Francisco Giants   1608   1527 13471          13315     0.513
## 10 Philadelphia Phillies  1600   1537 14351          14174     0.510
## 11 Minnesota Twins        1595   1544 14582          14524     0.508
## 12 Chicago Cubs           1593   1543 14056          13528     0.508
## 13 Houston Astros         1578   1559 14103          13851     0.503
## 14 Texas Rangers          1570   1569 15523          15630     0.500
## 15 Washington Nationals   1549   1587 13734          13828     0.494
## 16 Toronto Blue Jays      1548   1589 14771          14576     0.493
## 17 New York Mets          1540   1596 13602          13752     0.491
## 18 Chicago White Sox      1540   1598 14243          14607     0.491
## 19 Arizona Diamondbacks   1538   1600 14127          14366     0.490
## 20 Seattle Mariners       1531   1607 13603          14089     0.488
## 21 Tampa Bay Rays         1525   1612 13874          14361     0.486
## 22 Milwaukee Brewers      1521   1617 13872          14445     0.485
## 23 Cincinnati Reds        1472   1666 13853          14927     0.469
## 24 Miami Marlins          1470   1665 13341          14393     0.469
## 25 San Diego Padres       1469   1670 12954          14060     0.468
## 26 Colorado Rockies       1465   1675 15371          16064     0.467
## 27 Detroit Tigers         1455   1678 14165          15273     0.464
## 28 Pittsburgh Pirates     1423   1710 13124          14644     0.454
## 29 Baltimore Orioles      1404   1732 14023          15632     0.448
## 30 Kansas City Royals     1379   1759 13622          15524     0.439
```

3. Regular expressions to extract values in the Managers Column

```r
# Using regular expressions to extract first and last names

managers_data <- str_match_all(baseball$Managers, "([A-Z]\\.[^\\(]+) \\((\\d+)-(\\d+)")

names <- character(0)
wins <- numeric(0)
losses <- numeric(0)

# Extracting the data we want from the matrices into vector form
for(i in seq_along(managers_data)){
  for(j in seq_along(1:nrow(managers_data[[i]]))){
    names <- append(names, managers_data[[i]][j,2])
    wins <- append(wins, as.numeric(managers_data[[i]][j,3]))
    losses <- append(losses, as.numeric(managers_data[[i]][j,4]))
  }
}
# Using the vectors to create a tibble, and then using dplyr to get the desired result
managers <- tibble(
  Name = names,
  Wins = wins,
  Losses = losses
)

managers %>%
  mutate(Games = Wins + Losses) %>%
  group_by(Name) %>%
  summarise(Games = sum(Games), Wins = sum(Wins), Losses = sum(Losses), Win_Pct = sum(Wins)/sum(Games))
  arrange(desc(Games))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 592 x 5
##    Name        Games  Wins Losses Win_Pct
##    <chr>       <dbl> <dbl>  <dbl>   <dbl>
##  1 C.Mack       7679  3731   3948   0.486
##  2 T.La Russa   5255  2821   2434   0.537
##  3 B.Cox        4505  2504   2001   0.556
##  4 D.Baker      4500  2406   2094   0.535
##  5 B.Harris     4377  2158   2219   0.493
##  6 J.McGraw     4373  2583   1790   0.591
##  7 J.Torre      4323  2326   1997   0.538
##  8 B.Bochy      4032  2003   2029   0.497
##  9 S.Anderson   4028  2194   1834   0.545
## 10 G.Mauch      3939  1902   2037   0.483
## # ... with 582 more rows
```