DATA 300 3 Homework 4 Solution

Aadarsha Gopala Reddy

November 1, 2022

Contents

```
# load the libraries
library(rvest)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
```

```
library(dplyr)
library(xml2)
library(RSelenium)
```

```
# get the page html
page <- read_html("https://www.espncricinfo.com/player/team/india-6")

# output the html text to a text file
write_html(page, "html_text.txt")
```

```
# get the text from the page itself
compact_name <- page %>%
  html_nodes(".ds-text-compact-l.ds-font-medium.ds-text-ui-typo") %>%
  html_text()
name <- page %>%
  html_nodes(".ds-text-tight-l") %>%
  html_text()
age <- page %>%
  html_nodes(".ds-text-tight-m.ds-font-regular.ds-text-ui-typo-mid") %>%
  html_text()

# get the nodes with this specific class
# ".ds-inline-flex.ds-items-start.ds-leading-none"
# and then get the href attribute
links <- page %>%
  html_nodes(".ds-inline-flex.ds-items-start.ds-leading-none") %>%
  html_attr("href") %>%
```

```
  paste("https://www.espncricinfo.com", ., sep = "")

# keep only the links which start with "https://www.espncricinfo.com/player/"
# (keep only links to player pages)
links <- links[grepl("https://www.espncricinfo.com/player/", links)]
```

```
# get the data for each player
more_info <- function(player_link) {
  player_page <- read_html(player_link)
  player_info <- player_page %>%
    html_nodes(".ds-p-4") %>%
    html_text()
  player_info <- as.data.frame(
    player_info,
    stringsAsFactors = FALSE
  )
  return(player_info)
}

# call the function for link in the links list
player_info <- lapply(links, more_info)
```

```
# convert the list to a data frame
player_info_dataframe <- do.call(rbind, player_info)

# Keep only the rows containing "Full Name"
player_info_dataframe <- player_info_dataframe[
  grepl("Full Name", player_info_dataframe$player_info),
]

# export as csv
write.csv(
  player_info_dataframe,
  "player_info_dataframe.csv",
  row.names = FALSE
)
```

```
# duplicate the data frame
player_info_dataframe2 <- read.csv("player_info_dataframe.csv")
# export as csv file "player_info3.csv"
write.csv(
  player_info_dataframe2,
  "player_info3.csv",
  row.names = FALSE
)

# loop through each row in the dataframe
for (i in seq_len(nrow(player_info_dataframe2))) {
  # remove text "TEAMS" and after
  player_info_dataframe2[i, 1] <-
    gsub("TEAMS.*", "", player_info_dataframe2[i, 1])

  ########################################
```

```r
# put text "RELATIONS" and after, into a new column
player_info_dataframe2$relations[i] <-
  strsplit(player_info_dataframe2$x[i], "RELATIONS")[[1]][2]
# remove text "RELATIONS" and after, from the original column
player_info_dataframe2[i, 1] <-
  gsub("RELATIONS.*", "", player_info_dataframe2[i, 1])

########################################

# put text "Playing Role" and after, into a new column
player_info_dataframe2$playing_role[i] <-
  strsplit(player_info_dataframe2$x[i], "Playing Role")[[1]][2]
# remove text after "Playing Role" from the original column
player_info_dataframe2[i, 1] <-
  gsub("Playing Role.*", "", player_info_dataframe2[i, 1])

########################################

# put text "Fielding Position" and after, into a new column
player_info_dataframe2$fielding_position[i] <-
  strsplit(player_info_dataframe2$x[i], "Fielding Position")[[1]][2]
# remove text "Fielding Position" and after, from the original column
player_info_dataframe2[i, 1] <-
  gsub("Fielding Position.*", "", player_info_dataframe2[i, 1])

########################################

# put text "Bowling Style" and after, into a new column
player_info_dataframe2$bowling_style[i] <-
  strsplit(player_info_dataframe2$x[i], "Bowling Style")[[1]][2]
# remove text "Bowling Style" and after, from the original column
player_info_dataframe2[i, 1] <-
  gsub("Bowling Style.*", "", player_info_dataframe2[i, 1])

########################################

# put text "Batting Style" and after, into a new column
player_info_dataframe2$batting_style[i] <-
  strsplit(player_info_dataframe2$x[i], "Batting Style")[[1]][2]
# remove text "Batting Style" and after, from the original column
player_info_dataframe2[i, 1] <-
  gsub("Batting Style.*", "", player_info_dataframe2[i, 1])

########################################

# put text "Also Known As" and after, into a new column
player_info_dataframe2$aka[i] <-
  strsplit(player_info_dataframe2$x[i], "Also Known As")[[1]][2]
# remove text "Also Known As" and after, from the original column
player_info_dataframe2[i, 1] <-
  gsub("Also Known As.*", "", player_info_dataframe2[i, 1])

########################################
```

```r
# put text "Age" and after, into a new column
player_info_dataframe2$age[i] <-
  strsplit(player_info_dataframe2$x[i], "Age")[[1]][2]
# remove text "Age" and after, from the original column
player_info_dataframe2[i, 1] <-
  gsub("Age.*", "", player_info_dataframe2[i, 1])

########################################

# put text "Born" and after, into a new column
player_info_dataframe2$born[i] <-
  strsplit(player_info_dataframe2$x[i], "Born")[[1]][2]
# remove text "Born" and after, from the original column
player_info_dataframe2[i, 1] <-
  gsub("Born.*", "", player_info_dataframe2[i, 1])

# born format: "MMM DD, YYYY, City, State"
# convert to seperate location and date columns
# extract the date into new column
player_info_dataframe2$born_date[i] <-
  strsplit(player_info_dataframe2$born[i], ",")[[1]][1]
# extract the year into new column
player_info_dataframe2$born_year[i] <-
  strsplit(player_info_dataframe2$born[i], ",")[[1]][2]
# combine date and year into born_date column
player_info_dataframe2$born_date[i] <-
  paste(
    player_info_dataframe2$born_year[i],
    player_info_dataframe2$born_date[i],
    sep = "-"
  )

# calculate age from date of birth and today's date;
# replace the age column; round to 3 decimal places
player_info_dataframe2$age[i] <-
  round(as.numeric(
    difftime(
      Sys.Date(),
      as.Date(
        player_info_dataframe2$born_date[i],
        format = "%Y-%B %d"
      ),
      units = "auto"
    )
  ) / 365, 3)

# extract the city into new column
player_info_dataframe2$born_city[i] <-
  strsplit(player_info_dataframe2$born[i], ",")[[1]][3]
# extract the state into new column
player_info_dataframe2$born_state[i] <-
  strsplit(player_info_dataframe2$born[i], ",")[[1]][4]
# combine city and state into new column
```

```
  player_info_dataframe2$born_location[i] <-
    paste(player_info_dataframe2$born_city[i],
      player_info_dataframe2$born_state[i],
      sep = ","
    )

  ########################################

  # put text "Full Name" and after, into a new column
  player_info_dataframe2$full_name[i] <-
    strsplit(player_info_dataframe2$x[i], "Full Name")[[1]][2]
  # remove text "Full Name" and after, from the original column
  player_info_dataframe2[i, 1] <-
    gsub("Full Name.*", "", player_info_dataframe2[i, 1])
}

# remove redundant columns
# (first column, born_date, born_year, born_city, born_state)
player_info_dataframe2 <- player_info_dataframe2[
  ,
  -c(1, 9, 11, 12, 13)
]

# add the players' links to the dataframe
player_info_dataframe2$links <- links

# reorder the columns
player_info_dataframe2 <- player_info_dataframe2[
  ,
  c(10, 6, 1, 8, 7, 9, 4, 5, 3, 2, 11)
]

# export to csv
write.csv(player_info_dataframe2, "player_info2.csv", row.names = FALSE)

#output the dataframe in a neat format
player_info_dataframe2
```

```
##                          full_name            aka               relations
## 1             Mayank Anurag Agarwal           <NA>                    <NA>
## 2                   Arshdeep Singh           <NA>                    <NA>
## 3               Ravichandran Ashwin           <NA>                    <NA>
## 4                       Avesh Khan     Avesh Khan                    <NA>
## 5                Simran Dil Bahadur           <NA>                    <NA>
## 6                     Taniya Bhatia           <NA>                    <NA>
## 7             Yastika Harish Bhatia           <NA>                    <NA>
## 8                       Ekta Bisht           <NA>                    <NA>
## 9        Jasprit Jasbirsingh Bumrah           <NA>                    <NA>
## 10           Yuzvendra Singh Chahal           <NA>                    <NA>
## 11      Deepak Lokandersingh Chahar  Deepak Chahar      RD Chahar (cousin)
## 12              Rahul Desraj Chahar           <NA>      DL Chahar (cousin)
## 13                    Harleen Deol           <NA>                    <NA>
## 14                   Shikhar Dhawan           <NA>                    <NA>
## 15          Ruturaj Dashrat Gaikwad           <NA>                    <NA>
```

```
## 16    Rajeshwari Shivanand Gayakwad           <NA>                    <NA>
## 17           Richa Manabendra Ghosh           <NA>                    <NA>
## 18            Jhulan Nishit Goswami          Babul                    <NA>
## 19                Dayalan Hemalatha           <NA>                    <NA>
## 20              Deepak Jagbir Hooda           <NA>                    <NA>
## 21 Ishan Pranav Kumar Pandey Kishan           <NA>                    <NA>
## 22             Shreyas Santosh Iyer           <NA>                    <NA>
## 23      Venkatesh Rajasekaran Iyer Venkatesh Iyer                     <NA>
## 24  Ravindrasinh Anirudhsinh Jadeja           <NA>                    <NA>
## 25     Krishnakumar Dinesh Karthik         Dinesh                    <NA>
## 26         Harmanpreet Kaur Bhullar           <NA>                    <NA>
## 27                      Virat Kohli           <NA>                    <NA>
## 28                    Kuldeep Yadav           <NA>                    <NA>
## 29          Bhuvneshwar Kumar Singh           <NA>                    <NA>
## 30        Smriti Shriniwas Mandhana           <NA>                    <NA>
## 31               Sabbhineni Meghana           <NA>                    <NA>
## 32                     Meghna Singh           <NA>                    <NA>
## 33             Mohammed Shami Ahmed           <NA> Mohammed Kaif (brother)
## 34                   Mohammed Siraj           <NA>                    <NA>
## 35             Kiran Prabhu Navgire           <NA>                    <NA>
## 36           Hardik Himanshu Pandya           <NA>     KH Pandya (brother)
## 37            Rishabh Rajendra Pant           <NA>                    <NA>
## 38            Axar Rajeshbhai Patel   Akshar Patel                    <NA>
## 39             Harshal Vikram Patel           <NA>                    <NA>
## 40                     Poonam Yadav           <NA>                    <NA>
##             born_date    age                    born_location
## 1    1991-February 16 31.729              Bangalore, Karnataka
## 2    1999-February 05 23.753             Guna, Madhya Pradesh
## 3   1986-September 17 36.148  Madras (now Chennai), Tamil Nadu
## 4    1996-December 13 25.901            Indore, Madhya Pradesh
## 5    1999-December 13 22.901                      New Delhi,NA
## 6    1997-November 28 24.942               Chandigarh, Punjab
## 7    2000-November 01 22.014    Baroda (now Vadodara), Gujarat
## 8    1986-February 08 36.753             Almora, Uttar Pradesh
## 9    1993-December 06 28.923                      Ahmedabad,NA
## 10       1990-July 23 32.299                    Jind, Haryana
## 11     1992-August 07 30.255              Agra, Uttar Pradesh
## 12     1999-August 04  23.26            Bharaatpur, Rajasthan
## 13       1998-June 21 24.381               Chandigarh, Punjab
## 14   1985-December 05 36.932                        Delhi,NA
## 15    1997-January 31 25.767               Pune, Maharashtra
## 16       1991-June 01 31.441               Bijapur, Karnataka
## 17  2003-September 28 19.107            Siliguri, West Bengal
## 18   1982-November 25 39.962                    Nadia, Bengal
## 19  1994-September 29  28.11  Madras (now Chennai), Tamil Nadu
## 20      1995-April 19 27.556                       Rohtak,NA
## 21       1998-July 18 24.307                     Patna, Bihar
## 22   1994-December 06 27.923                       Mumbai,NA
## 23   1994-December 25 27.871            Indore, Madhya Pradesh
## 24   1988-December 06 33.926          Navagam-Khed, Saurashtra
## 25       1985-June 01 37.444  Madras (now Chennai), Tamil Nadu
## 26      1989-March 08 33.674                       Punjab,NA
## 27   1988-November 05 34.011                        Delhi,NA
## 28   1994-December 14 27.901             Kanpur, Uttar Pradesh
```

```
## 29   1990-February 05 32.759            Meerut, Uttar Pradesh
## 30       1996-July 18 26.307  Bombay (now Mumbai), Maharashtra
## 31       1996-June 07 26.419          Krishna, Andhra Pradesh
## 32       1994-June 18 28.392            Bijnor, Uttar Pradesh
## 33  1990-September 03 32.184            Amroha, Uttar Pradesh
## 34      1994-March 13 28.658                    Hyderabad,NA
## 35             NA-NA   <NA>                          NA,NA
## 36    1993-October 11 29.077               Choryasi, Gujarat
## 37    1997-October 04 25.093            Haridwar, Uttarakhand
## 38    1994-January 20   28.8                   Anand, Gujarat
## 39   1990-November 23 31.962                  Sanand, Gujarat
## 40     1991-August 24 31.211              Agra, Uttar Pradesh
##                          bowling_style  batting_style fielding_position
## 1                   Right arm Offbreak Right hand Bat              <NA>
## 2                 Left arm Medium fast  Left hand Bat              <NA>
## 3                   Right arm Offbreak Right hand Bat              <NA>
## 4                 Right arm Fast medium Right hand Bat              <NA>
## 5                       Right arm Medium  Left hand Bat              <NA>
## 6                                  <NA> Right hand Bat              <NA>
## 7                Slow Left arm Orthodox  Left hand Bat              <NA>
## 8                Slow Left arm Orthodox  Left hand Bat              <NA>
## 9                         Right arm Fast Right hand Bat              <NA>
## 10                      Legbreak Googly Right hand Bat              <NA>
## 11                    Right arm Medium Right hand Bat              <NA>
## 12                      Legbreak Googly Right hand Bat              <NA>
## 13 Right arm Offbreak, Legbreak Googly Right hand Bat              <NA>
## 14                  Right arm Offbreak  Left hand Bat              <NA>
## 15                  Right arm Offbreak Right hand Bat              <NA>
## 16               Slow Left arm Orthodox Right hand Bat              <NA>
## 17                    Right arm Medium Right hand Bat      Wicketkeeper
## 18                    Right arm Medium Right hand Bat              <NA>
## 19                  Right arm Offbreak Right hand Bat              <NA>
## 20                  Right arm Offbreak Right hand Bat              <NA>
## 21                                 <NA>  Left hand Bat      Wicketkeeper
## 22 Right arm Offbreak, Legbreak Googly Right hand Bat              <NA>
## 23                      Right arm Medium  Left hand Bat              <NA>
## 24               Slow Left arm Orthodox  Left hand Bat              <NA>
## 25                  Right arm Offbreak Right hand Bat      Wicketkeeper
## 26                  Right arm Offbreak Right hand Bat              <NA>
## 27                    Right arm Medium Right hand Bat              <NA>
## 28                 Left arm Wrist spin  Left hand Bat              <NA>
## 29                    Right arm Medium Right hand Bat              <NA>
## 30                  Right arm Offbreak  Left hand Bat              <NA>
## 31                    Right arm Medium Right hand Bat              <NA>
## 32                    Right arm Medium Right hand Bat              <NA>
## 33                      Right arm Fast Right hand Bat              <NA>
## 34                Right arm Medium fast Right hand Bat              <NA>
## 35                                 <NA> Right hand Bat              <NA>
## 36                Right arm Medium fast Right hand Bat              <NA>
## 37                                 <NA>  Left hand Bat      Wicketkeeper
## 38               Slow Left arm Orthodox  Left hand Bat              <NA>
## 39                      Right arm Medium Right hand Bat              <NA>
## 40                      Legbreak Googly Right hand Bat              <NA>
##           playing_role
```

```
## 1               Batter
## 2               Bowler
## 3   Bowling Allrounder
## 4               Bowler
## 5   Bowling Allrounder
## 6  Wicketkeeper Batter
## 7  Middle order Batter
## 8               Bowler
## 9               Bowler
## 10              Bowler
## 11              Bowler
## 12              Bowler
## 13 Middle order Batter
## 14      Opening Batter
## 15              Batter
## 16              Bowler
## 17 Wicketkeeper Batter
## 18              Bowler
## 19    Top order Batter
## 20          Allrounder
## 21 Wicketkeeper Batter
## 22    Top order Batter
## 23          Allrounder
## 24          Allrounder
## 25 Wicketkeeper Batter
## 26          Allrounder
## 27    Top order Batter
## 28              Bowler
## 29              Bowler
## 30      Opening Batter
## 31      Opening Batter
## 32              Bowler
## 33              Bowler
## 34              Bowler
## 35 Middle order Batter
## 36          Allrounder
## 37 Wicketkeeper Batter
## 38  Bowling Allrounder
## 39              Bowler
## 40              Bowler
##                                                                  links
## 1         https://www.espncricinfo.com/player/mayank-agarwal-398438
## 2        https://www.espncricinfo.com/player/arshdeep-singh-1125976
## 3     https://www.espncricinfo.com/player/ravichandran-ashwin-26421
## 4             https://www.espncricinfo.com/player/avesh-khan-694211
## 5        https://www.espncricinfo.com/player/simran-bahadur-1204925
## 6          https://www.espncricinfo.com/player/taniya-bhatia-883423
## 7         https://www.espncricinfo.com/player/yastika-bhatia-960715
## 8             https://www.espncricinfo.com/player/ekta-bisht-442048
## 9         https://www.espncricinfo.com/player/jasprit-bumrah-625383
## 10      https://www.espncricinfo.com/player/yuzvendra-chahal-430246
## 11         https://www.espncricinfo.com/player/deepak-chahar-447261
## 12         https://www.espncricinfo.com/player/rahul-chahar-1064812
## 13          https://www.espncricinfo.com/player/harleen-deol-960845
```

```
## 14         https://www.espncricinfo.com/player/shikhar-dhawan-28235
## 15      https://www.espncricinfo.com/player/ruturaj-gaikwad-1060380
## 16   https://www.espncricinfo.com/player/rajeshwari-gayakwad-709635
## 17          https://www.espncricinfo.com/player/richa-ghosh-1212830
## 18         https://www.espncricinfo.com/player/jhulan-goswami-53932
## 19     https://www.espncricinfo.com/player/dayalan-hemalatha-961107
## 20          https://www.espncricinfo.com/player/deepak-hooda-497121
## 21          https://www.espncricinfo.com/player/ishan-kishan-720471
## 22          https://www.espncricinfo.com/player/shreyas-iyer-642519
## 23        https://www.espncricinfo.com/player/venkatesh-iyer-851403
## 24       https://www.espncricinfo.com/player/ravindra-jadeja-234675
## 25         https://www.espncricinfo.com/player/dinesh-karthik-30045
## 26      https://www.espncricinfo.com/player/harmanpreet-kaur-372317
## 27           https://www.espncricinfo.com/player/virat-kohli-253802
## 28         https://www.espncricinfo.com/player/kuldeep-yadav-559235
## 29     https://www.espncricinfo.com/player/bhuvneshwar-kumar-326016
## 30       https://www.espncricinfo.com/player/smriti-mandhana-597806
## 31    https://www.espncricinfo.com/player/sabbhineni-meghana-556529
## 32          https://www.espncricinfo.com/player/meghna-singh-709839
## 33        https://www.espncricinfo.com/player/mohammed-shami-481896
## 34        https://www.espncricinfo.com/player/mohammed-siraj-940973
## 35 https://www.espncricinfo.com/player/kiran-prabhu-navgire-1289983
## 36         https://www.espncricinfo.com/player/hardik-pandya-625371
## 37          https://www.espncricinfo.com/player/rishabh-pant-931581
## 38            https://www.espncricinfo.com/player/axar-patel-554691
## 39         https://www.espncricinfo.com/player/harshal-patel-390481
## 40          https://www.espncricinfo.com/player/poonam-yadav-630972
```