Project 1

Folorunsho Atanda

2023-09-23

Objective: Extract information from a text file to create a csv. Text file contains a chess tournament information. csv file is going to have **Players Name**, **Player's State**, **Total Number of Points**, **Player's Pre-Rating**, and **Opponent's Average Pre-Rating**

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
url <- "https://raw.githubusercontent.com/folushoa/Data-Science/Data-607/Project%201/tournamentinfo.txt
columnNames <- c("Pair Num", "Player Name", "Total Pts", "Round 1", "Round 2",
                 "Round 3", "Round 4", "Round 5", "Round 6", "Round 7")
tournament_info <- fread(url, skip = 4, sep = "|", fill = TRUE,
                         col.names = columnNames, drop = 11)
```

```
rows_to_read <- seq_len(nrow(tournament_info)) %% 3 != 0
tournament_info <- tournament_info[rows_to_read, ]
head(tournament_info, 5)
```

```
##    Pair Num                 Player Name Total Pts Round 1 Round 2 Round 3
## 1:        1                    GARY HUA       6.0   W  39   W  21   W  18
## 2:       ON 15445895 / R: 1794   ->1817       N:2       W       B       W
## 3:        2              DAKSHESH DARURI       6.0   W  63   W  58   L   4
## 4:       MI 14598900 / R: 1553   ->1663       N:2       B       W       B
## 5:        3                ADITYA BAJAJ       6.0   L   8   W  61   W  25
##    Round 4 Round 5 Round 6 Round 7
## 1:   W  14   W   7   D  12   D   4
## 2:       B       W       B       W
## 3:   W  17   W  16   W  20   W   7
## 4:       W       B       W       B
## 5:   W  21   W  11   W  13   W  12
```

The text file had two lines that contained info for one player, so I decided to create two table. Table one contains the player info from the first line. Table two contains the player info from the second line.

```
player_info_1_rows <-seq_len(nrow(tournament_info)) %% 2 != 0
player_info_2_rows <-seq_len(nrow(tournament_info)) %% 2 == 0

player_info_1 <- tournament_info[player_info_1_rows, ]
head(player_info_1, 5)
```

```
##    Pair Num          Player Name Total Pts Round 1 Round 2 Round 3 Round 4
## 1:        1             GARY HUA       6.0   W  39   W  21   W  18   W  14
## 2:        2      DAKSHESH DARURI       6.0   W  63   W  58   L   4   W  17
## 3:        3         ADITYA BAJAJ       6.0   L   8   W  61   W  25   W  21
## 4:        4 PATRICK H SCHILLING       5.5   W  23   D  28   W   2   W  26
## 5:        5           HANSHI ZUO       5.5   W  45   W  37   D  12   D  13
##    Round 5 Round 6 Round 7
## 1:   W   7   D  12   D   4
## 2:   W  16   W  20   W   7
## 3:   W  11   W  13   W  12
## 4:   D   5   W  19   D   1
## 5:   D   4   W  14   W  17
```

```
player_info_2 <- tournament_info[player_info_2_rows, ]
head(player_info_2, 5)
```

```
##    Pair Num                 Player Name Total Pts Round 1 Round 2 Round 3
## 1:       ON 15445895 / R: 1794   ->1817       N:2       W       B       W
## 2:       MI 14598900 / R: 1553   ->1663       N:2       B       W       B
## 3:       MI 14959604 / R: 1384   ->1640       N:2       W       B       W
## 4:       MI 12616049 / R: 1716   ->1744       N:2       W       B       W
## 5:       MI 14601533 / R: 1655   ->1690       N:2       B       W       B
##    Round 4 Round 5 Round 6 Round 7
## 1:       B       W       B       W
## 2:       W       B       W       B
```

```
## 3:       B       W       B       W
## 4:       B       W       B       B
## 5:       W       B       W       B
```

Clean up **player__info__1** table

```
player_info_1$`Pair Num` <- NULL
```

```
player_info_1 <- player_info_1 %>% separate_wider_regex(cols = "Round 1",
                 patterns = c("[a-zA-Z]", "\\ *", "Opponent 1" = "[0-9]*"))
player_info_1 <- player_info_1 %>% separate_wider_regex(cols = "Round 2",
                 patterns = c("[a-zA-Z]", "\\ *", "Opponent 2" = "[0-9]*"))
player_info_1 <- player_info_1 %>% separate_wider_regex(cols = "Round 3",
                 patterns = c("[a-zA-Z]", "\\ *", "Opponent 3" = "[0-9]*"))
player_info_1 <- player_info_1 %>% separate_wider_regex(cols = "Round 4",
                 patterns = c("[a-zA-Z]", "\\ *", "Opponent 4" = "[0-9]*"))
player_info_1 <- player_info_1 %>% separate_wider_regex(cols = "Round 5",
                 patterns = c("[a-zA-Z]", "\\ *", "Opponent 5" = "[0-9]*"))
player_info_1 <- player_info_1 %>% separate_wider_regex(cols = "Round 6",
                 patterns = c("[a-zA-Z]", "\\ *", "Opponent 6" = "[0-9]*"))
player_info_1 <- player_info_1 %>% separate_wider_regex(cols = "Round 7",
                 patterns = c("[a-zA-Z]", "\\ *", "Opponent 7" = "[0-9]*"))
```

```
player_info_1$`Total Pts` <- as.numeric(player_info_1$`Total Pts`)
player_info_1$`Opponent 1` <- as.numeric(player_info_1$`Opponent 1`)
player_info_1$`Opponent 2` <- as.numeric(player_info_1$`Opponent 2`)
player_info_1$`Opponent 3` <- as.numeric(player_info_1$`Opponent 3`)
player_info_1$`Opponent 4` <- as.numeric(player_info_1$`Opponent 4`)
player_info_1$`Opponent 5` <- as.numeric(player_info_1$`Opponent 5`)
player_info_1$`Opponent 6` <- as.numeric(player_info_1$`Opponent 6`)
player_info_1$`Opponent 7` <- as.numeric(player_info_1$`Opponent 7`)
head(player_info_1, 5)
```

```
## # A tibble: 5 x 9
##   `Player Name`  `Total Pts` `Opponent 1` `Opponent 2` `Opponent 3` `Opponent 4`
##   <chr>                <dbl>        <dbl>        <dbl>        <dbl>        <dbl>
## 1 GARY HUA               6            39           21           18           14
## 2 DAKSHESH DARU~         6            63           58            4           17
## 3 ADITYA BAJAJ           6             8           61           25           21
## 4 PATRICK H SCH~         5.5          23           28            2           26
## 5 HANSHI ZUO             5.5          45           37           12           13
## # i 3 more variables: `Opponent 5` <dbl>, `Opponent 6` <dbl>,
## #   `Opponent 7` <dbl>
```

Clean up **player__info__2** table

```
player_info_2$`Total Pts` <- NULL
player_info_2$`Round 1` <- NULL
player_info_2$`Round 2` <- NULL
player_info_2$`Round 3` <- NULL
player_info_2$`Round 4` <- NULL
player_info_2$`Round 5` <- NULL
```

```
player_info_2$`Round 6` <- NULL
player_info_2$`Round 7` <- NULL
```

```
player_info_2 <- player_info_2 %>% separate_wider_regex(cols = "Player Name",
                 patterns = c("[0-9]*","\\ ", "\\/", "\\ ", "[R]","\\:", "\\ *",
                 "Player Rating" = "[0-9]*", "[P]?", "[0-9]*", "\\ *", "\\-",
                 "\\>", "\\ *", "[0-9]*", "[P]?", "[0-9]*"))
```

```
player_info_2$`Player Rating` <- as.numeric(player_info_2$`Player Rating`)
```

```
colnames(player_info_2) <- c("Player State", "Player Rating")
head(player_info_2, 5)
```

```
## # A tibble: 5 x 2
##   `Player State` `Player Rating`
##   <chr>                    <dbl>
## 1 ON                        1794
## 2 MI                        1553
## 3 MI                        1384
## 4 MI                        1716
## 5 MI                        1655
```

The opponent number in **player_info_1** correspond to the the player row info in **player_info_1** and **player_info_2**. So I replaced the opponent number in **player_info_1** with the player pre-rating from **player_info_2**.

```
for (i in 1:nrow(player_info_1)){
  for (j in 3:9){
    if (is.na(player_info_1[i, j])){
      next
    }
    else{
      player_info_1[i, j] <- player_info_2[as.numeric(player_info_1[i, j]), 2]
    }
  }
}
```

```
colnames(player_info_1) <- c("Player Name", "Total Pts", "Opponent 1 Rating",
         "Opponent 2 Rating", "Opponent 3 Rating", "Opponent 4 Rating",
         "Opponent 5 Rating", "Opponent 6 Rating", "Opponent 7 Rating")
head(player_info_1, 5)
```

```
## # A tibble: 5 x 9
##   `Player Name`       `Total Pts` `Opponent 1 Rating` `Opponent 2 Rating`
##   <chr>                     <dbl>               <dbl>               <dbl>
## 1 GARY HUA                      6                1436                1563
## 2 DAKSHESH DARURI               6                1175                 917
## 3 ADITYA BAJAJ                  6                1641                 955
## 4 PATRICK H SCHILLING         5.5                1363                1507
## 5 HANSHI ZUO                  5.5                1242                 980
## # i 5 more variables: `Opponent 3 Rating` <dbl>, `Opponent 4 Rating` <dbl>,
## #   `Opponent 5 Rating` <dbl>, `Opponent 6 Rating` <dbl>,
## #   `Opponent 7 Rating` <dbl>
```

create table with **Player Name**, **Player State**, **Player Total Points**, **Player Rating**, and **Opponent Avg Rating**.

```
player_info <- tibble("Player Name" = player_info_1$`Player Name`,
                "Player State" = player_info_2$`Player State`,
                "Player Total Points" = player_info_1$`Total Pts`,
                "Player Rating" = player_info_2$`Player Rating`,
                "Opponent Avg Rating" = signif(rowMeans(player_info_1[, 3:9],
                 na.rm = TRUE), 4))
head(player_info, 5)
```

```
## # A tibble: 5 x 5
##   `Player Name`       `Player State` `Player Total Points` `Player Rating`
##   <chr>               <chr>                          <dbl>           <dbl>
## 1 GARY HUA            ON                               6              1794
## 2 DAKSHESH DARURI     MI                               6              1553
## 3 ADITYA BAJAJ        MI                               6              1384
## 4 PATRICK H SCHILLING MI                               5.5            1716
## 5 HANSHI ZUO          MI                               5.5            1655
## # i 1 more variable: `Opponent Avg Rating` <dbl>
```