DATA607Project1

Erick Hadi

2024-10-06

DATA 607 Project 1

In this project, you're given a text file with chess tournament results where the information has some structure. Your job is to create an R Markdown file that generates a .CSV file (that could for example be imported into a SQL database) with the following information for all of the players: Player's Name, Player's State, Total Number of Points, Player's Pre-Rating, and Average Pre Chess Rating of Opponents

Load the data

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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 error
```

```r
library(dplyr)
library(tidyr)
```

```r
gitlink <- "https://raw.githubusercontent.com/ErickH1/DATA607Project1/refs/heads/main/tournamentinfo.tx
chess_raw_data <- readLines(gitlink)
```

```
## Warning in readLines(gitlink): incomplete final line found on
## 'https://raw.githubusercontent.com/ErickH1/DATA607Project1/refs/heads/main/tournamentinfo.txt'
```

```r
head(chess_raw_data, 7)
```

```
## [1] "-----------------------------------------------------------------------------------------"
## [2] " Pair | Player Name                     |Total|Round|Round|Round|Round|Round|Round|Round|"
## [3] " Num  | USCF ID / Rtg (Pre->Post)       | Pts |  1  |  2  |  3  |  4  |  5  |  6  |  7  |"
## [4] "-----------------------------------------------------------------------------------------"
## [5] "    1 | GARY HUA                        |6.0  |W  39|W  21|W  18|W  14|W   7|D  12|D   4|"
## [6] "   ON | 15445895 / R: 1794   ->1817     |N:2  |W    |B    |W    |B    |W    |B    |W    |"
## [7] "-----------------------------------------------------------------------------------------"
```

Data Processing

Inserting text data into a matrix to make data capturing and processing easier.

```
chess_data_matrix <- matrix(unlist(chess_raw_data), byrow=TRUE)
matrix_1 <- chess_data_matrix[seq(5,length(chess_data_matrix),3)]
matrix_2 <- chess_data_matrix[seq(6,length(chess_data_matrix),3)]

head(chess_data_matrix,10)
```

```
##       [,1]
##  [1,] "-----------------------------------------------------------------------------------------"
##  [2,] " Pair | Player Name                     |Total|Round|Round|Round|Round|Round|Round|Round|"
##  [3,] " Num  | USCF ID / Rtg (Pre->Post)       | Pts |  1  |  2  |  3  |  4  |  5  |  6  |  7  |"
##  [4,] "-----------------------------------------------------------------------------------------"
##  [5,] "    1 | GARY HUA                        |6.0  |W  39|W  21|W  18|W  14|W   7|D  12|D   4|"
##  [6,] "   ON | 15445895 / R: 1794   ->1817     |N:2  |W    |B    |W    |B    |W    |B    |W    |"
##  [7,] "-----------------------------------------------------------------------------------------"
##  [8,] "    2 | DAKSHESH DARURI                 |6.0  |W  63|W  58|L   4|W  17|W  16|W  20|W   7|"
##  [9,] "   MI | 14598900 / R: 1553   ->1663     |N:2  |B    |W    |B    |W    |B    |W    |B    |"
## [10,] "-----------------------------------------------------------------------------------------"
```

```
head(matrix_1)
```

```
## [1] "    1 | GARY HUA                        |6.0  |W  39|W  21|W  18|W  14|W   7|D  12|D   4|"
## [2] "    2 | DAKSHESH DARURI                 |6.0  |W  63|W  58|L   4|W  17|W  16|W  20|W   7|"
## [3] "    3 | ADITYA BAJAJ                    |6.0  |L   8|W  61|W  25|W  21|W  11|W  13|W  12|"
## [4] "    4 | PATRICK H SCHILLING             |5.5  |W  23|D  28|W   2|W  26|D   5|W  19|D   1|"
## [5] "    5 | HANSHI ZUO                      |5.5  |W  45|W  37|D  12|D  13|D   4|W  14|W  17|"
## [6] "    6 | HANSEN SONG                     |5.0  |W  34|D  29|L  11|W  35|D  10|W  27|W  21|"
```

```
head(matrix_2)
```

```
## [1] "   ON | 15445895 / R: 1794   ->1817     |N:2  |W    |B    |W    |B    |W    |B    |W    |"
## [2] "   MI | 14598900 / R: 1553   ->1663     |N:2  |B    |W    |B    |W    |B    |W    |B    |"
## [3] "   MI | 14959604 / R: 1384   ->1640     |N:2  |W    |B    |W    |B    |W    |B    |W    |"
## [4] "   MI | 12616049 / R: 1716   ->1744     |N:2  |W    |B    |W    |B    |W    |B    |B    |"
## [5] "   MI | 14601533 / R: 1655   ->1690     |N:2  |B    |W    |B    |W    |B    |W    |B    |"
## [6] "   OH | 15055204 / R: 1686   ->1687     |N:3  |W    |B    |W    |B    |B    |W    |B    |"
```

Extracting Chess Data

Using Regex and string manipulation to extract relevant information into vectors.

```
ID <- as.numeric(str_extract(matrix_1, '\\d+'))

Name <- str_trim(str_extract(str_extract(matrix_1, '[A-z].{1,32}'), '.+\\s{2,}'))

State <- str_extract(matrix_2, '[A-Z]{2}')

Total_Points <- as.numeric(str_extract(matrix_1, '\\d+\\.\\d'))
```

```
Pre_Rating <- as.numeric(str_extract(str_extract(matrix_2, 'R:.{8,}-'), '\\d{1,4}'))

Rounds <- str_extract_all(matrix_1, '[A-Z]\\s{2,}\\d+')
Rounds <- str_extract_all(Rounds, '\\d+')
```

```
## Warning in stri_extract_all_regex(string, pattern, simplify = simplify, :
## argument is not an atomic vector; coercing
```

Calculate Avg Opponent Rating

Calculating avg opponent rating using pre rating and rounds vectors. Instead of for loops utilized sapply.

```
Avg_Opp_Pre_Rating <- sapply(Rounds, function(x) round(mean(Pre_Rating[as.numeric(x)]), 0))
Avg_Opp_Pre_Rating
```

```
##  [1] 1605 1469 1564 1574 1501 1519 1372 1468 1523 1554 1468 1506 1498 1515 1484
## [16] 1386 1499 1480 1426 1411 1470 1300 1214 1357 1363 1507 1222 1522 1314 1144
## [31] 1260 1379 1277 1375 1150 1388 1385 1539 1430 1391 1248 1150 1107 1327 1152
## [46] 1358 1392 1356 1286 1296 1356 1495 1345 1206 1406 1414 1363 1391 1319 1330
## [61] 1327 1186 1350 1263
```

Inserting Extracted Data Into Chess Data Frame

```
chess_data <- data.frame(ID,Name,State,Total_Points,Pre_Rating,Avg_Opp_Pre_Rating)
head(chess_data)
```

```
##   ID                Name State Total_Points Pre_Rating Avg_Opp_Pre_Rating
## 1  1            GARY HUA    ON          6.0       1794               1605
## 2  2     DAKSHESH DARURI    MI          6.0       1553               1469
## 3  3        ADITYA BAJAJ    MI          6.0       1384               1564
## 4  4 PATRICK H SCHILLING    MI          5.5       1716               1574
## 5  5          HANSHI ZUO    MI          5.5       1655               1501
## 6  6         HANSEN SONG    OH          5.0       1686               1519
```

Export Data to CSV

```
write.csv(chess_data, "chesstournamentinfo.csv")
```