DATA_607_Project_One

Mathew Katz

2022-09-25

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 For the first player, the information would be: Gary Hua, ON, 6.0, 1794, 1605

Loading and Reading the Data

```
library(readr, quietly = TRUE)
library(stringr, quietly = TRUE)
library(tidyverse, quietly = TRUE)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
url = 'https://raw.githubusercontent.com/MathewKatz/CUNYSPS/main/tournamentinfo.txt'
df <- readLines(url)
```

```
head(df)
```

```
## [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    |"
```

Extraction of Key Fields:

```
player_name <- unlist(str_extract_all(df,"(?<=\\d\\s\\|\\s)([A-z, -]*\\s){1,}[[:alpha:]]*(?=\\s*\\|)"))
player_state <- unlist(str_extract_all(df, "[[:upper:]]{2}(?=\\s\\|)"))
total_points <- as.numeric(unlist(str_extract_all(df, "(?<=\\|)\\d\\.\\d")))
player_pre_rating <- as.numeric(unlist(str_extract_all(df, "(?<=R:\\s{1,2})(\\d{3,4}(?=\\s))|(\\d{3,4}(
player_number <- as.numeric(unlist(str_extract_all(df,"(?<=\\s{3,4})\\d{1,2}(?=\\s)")))
```

Creating Dataframe with Extracted Data

```
processed_data <- data.frame(player_name, player_state, total_points, player_pre_rating, player_number)
```

List of Opponent Player's Numbers

```
newdf <- df[seq(5, 196, 3)]
opponent_num <- as.numeric(unlist(str_extract_all(newdf, "(?<=\\|(W|L|D)\\s{2,3})[[:digit:]]{1,2}(?=\\|)
```

Getting Pre Chess Rating

```
pre_chess_rating_matrix <- matrix(data = NA, nrow = 64, ncol = 2)

colnames(pre_chess_rating_matrix) <- c("total_opp_pcr", "avg_opp_pcr")

row_counter <- 0

for(i in seq(from=1, to=length(opponent_num)-6, by=7)){
  row_counter <- row_counter + 1

  pre_chess_rating_matrix[row_counter, 1] <- (sum(subset(processed_data$player_pre_rat, processed_data$p

  pre_chess_rating_matrix[row_counter, 2] <- pre_chess_rating_matrix[row_counter, 1] / length(subset(op

head(pre_chess_rating_matrix)
```

```
##      total_opp_pcr avg_opp_pcr
## [1,]         11237    1605.286
## [2,]         10285    1469.286
## [3,]         10945    1563.571
## [4,]         11015    1573.571
## [5,]         10506    1500.857
## [6,]         10631    1518.714
```

Tidy Data

```
pre_chess_rating_matrix[, 2] <- round(pre_chess_rating_matrix[,2], digits = 0)

final_df <- cbind(processed_data, pre_chess_rating_matrix[, 2])
```

```
head(final_df)
```

```
##                       player_name player_state total_points player_pre_rating
## 1 GARY HUA                                ON          6.0              1794
## 2 DAKSHESH DARURI                         MI          6.0              1553
## 3 ADITYA BAJAJ                            MI          6.0              1384
## 4 PATRICK H SCHILLING                     MI          5.5              1716
## 5 HANSHI ZUO                              MI          5.5              1655
## 6 HANSEN SONG                             OH          5.0              1686
##   player_number pre_chess_rating_matrix[, 2]
## 1             1                        1605
## 2             2                        1469
## 3             3                        1564
```

```
## 4             4                       1574
## 5             5                       1501
## 6             6                       1519
```

```
path <- getwd()
write.csv(final_df, file.path(path, "chess_data.csv"))
```