Data 607 Project 1

Jiaxin Zheng

2024-10-09

Introduction:

In this project we take chess tournament results text file into a .csv file. 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

Data Cleaning Strategy: 1. read the data from text file 2. removing all the |'s , -'s and NA columns 3. separate the rows into more clean and combine as a readable data table 4. separate the USCF ID / Rtg (Pre->Post) 5. Calculating the opposing player rating

```
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
```

```
url <- "https://raw.githubusercontent.com/Jennyjjxxzz/Data-607_Project1/refs/heads/main/tournamentinfo.
tournament <- read.delim(url, header = FALSE, sep = "\n")
```

```
str(tournament)
```

```
## 'data.frame':    196 obs. of  1 variable:
##  $ V1: chr  "----------------------------------------------------------------------------------------
```

```
head(tournament)
```

```
##                                                                                            V1
## 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    |
cols <- c('player_num','name','total_pts','round_1','round_2','round_3','round_4',
          'round_5','round_6','round_7','NA')
```

```
df_tournament <- read.csv(url, sep="|", header = FALSE, skip = 3, col.names = cols)
dashes <- "-----------------------------------------------------------------------------------------"
df_tournament <- df_tournament %>% filter(player_num != dashes)
```

```
df1 <- df_tournament %>%
  filter(row_number() %% 2 == 1)
```

```
df2 <- df_tournament%>%
  filter(row_number() %% 2 != 1)
```

```
combine_df<- cbind(df1, df2)
head(combine_df)
```

```
##   player_num                              name total_pts round_1 round_2
## 1          1   GARY HUA                             6.0     W  39    W  21
## 2          2   DAKSHESH DARURI                      6.0     W  63    W  58
## 3          3   ADITYA BAJAJ                         6.0     L   8    W  61
## 4          4   PATRICK H SCHILLING                  5.5     W  23    D  28
## 5          5   HANSHI ZUO                           5.5     W  45    W  37
## 6          6   HANSEN SONG                          5.0     W  34    D  29
##   round_3 round_4 round_5 round_6 round_7 NA. player_num
## 1   W  18   W  14   W   7   D  12   D   4  NA         ON
## 2   L   4   W  17   W  16   W  20   W   7  NA         MI
## 3   W  25   W  21   W  11   W  13   W  12  NA         MI
## 4   W   2   W  26   D   5   W  19   D   1  NA         MI
## 5   D  12   D  13   D   4   W  14   W  17  NA         MI
## 6   L  11   W  35   D  10   W  27   W  21  NA         OH
##                                 name total_pts round_1 round_2 round_3 round_4
## 1  15445895 / R: 1794   ->1817          N:2     W       B       W       B
## 2  14598900 / R: 1553   ->1663          N:2     B       W       B       W
## 3  14959604 / R: 1384   ->1640          N:2     W       B       W       B
## 4  12616049 / R: 1716   ->1744          N:2     W       B       W       B
## 5  14601533 / R: 1655   ->1690          N:2     B       W       B       W
## 6  15055204 / R: 1686   ->1687          N:3     W       B       W       B
##   round_5 round_6 round_7 NA.
## 1   W       B       W      NA
## 2   B       W       B      NA
## 3   W       B       W      NA
## 4   W       B       B      NA
## 5   B       W       B      NA
## 6   B       W       B      NA
```

```
combine_df<- combine_df %>%
  subset(select=c(1:10, 12:13))

colnames(combine_df)<- c("Player_num", "Name", "Total_Points", "Round1", "Round2", "Round3", "Round4",
view(combine_df)
```

```
combine_df <- combine_df %>%
  mutate(Pre_Rating = str_extract(Opponent_Info, "(?<=R: )\\d+"),  # Extract the Pre-Rating
         Post_Rating = str_extract(Opponent_Info, "(?<=->)\\d+"))  # Extract the Post-Rating


combine_df <- combine_df %>%
  rowwise() %>%
  mutate(Average_Opponent_Rating = mean(as.numeric(c(str_extract(Round1, "(?<=R: )\\d+"),
                                                     str_extract(Round2, "(?<=R: )\\d+"),
                                                     str_extract(Round3, "(?<=R: )\\d+"),
                                                     str_extract(Round4, "(?<=R: )\\d+"),
                                                     str_extract(Round5, "(?<=R: )\\d+"),
                                                     str_extract(Round6, "(?<=R: )\\d+"),
                                                     str_extract(Round7, "(?<=R: )\\d+"))), na.rm = TRU

#columns for the cleaned dataframe
final_df <- combine_df %>%
  select(Name, State, Total_Points, Pre_Rating, Average_Opponent_Rating)

view(final_df)
head(final_df)
```

```
## # A tibble: 6 x 5
## # Rowwise:
##   Name                      State Total_Points Pre_Rating Average_Opponent_Rat~1
##   <chr>                     <chr> <chr>        <chr>                       <dbl>
## 1 " GARY HUA              ~ "   ~ "6.0  "      1794                          NaN
## 2 " DAKSHESH DARURI       ~ "   ~ "6.0  "      1553                          NaN
## 3 " ADITYA BAJAJ          ~ "   ~ "6.0  "      1384                          NaN
## 4 " PATRICK H SCHILLING   ~ "   ~ "5.5  "      1716                          NaN
## 5 " HANSHI ZUO            ~ "   ~ "5.5  "      1655                          NaN
## 6 " HANSEN SONG           ~ "   ~ "5.0  "      1686                          NaN
## # i abbreviated name: 1: Average_Opponent_Rating
```

```
write.csv(final_df, "tournament_cleaned.csv", row.names = FALSE)
```