DATA 607 Project 1

Kenan Sooklall

2021-02-10

Description

The goal of this project is to parse chess tournament results into:

Player's Name, Player's State, Total Number of Points, Player's Pre-Rating, and Average Pre Chess Rating of Opponents

```
library(readr)
library(stringr)
library(dplyr)
library(ggplot2)
```

Read the file with **read lines**, the data can be obtained from my github

```
lines <- read_lines("https://raw.githubusercontent.com/ksooklall/CUNY-SPS-Masters-DS/main/DATA_607/proje
lines[1:7]
```

```
## [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] "-----------------------------------------------------------------------------------------"
```

Prase the file skipping rows that don't contains any data and populate two vectors - Player data (player_rows) - Match data (mdf)

```
player_rows = c()
mdf = c()

for (i in seq(5, 195, 3)) {
  row1 <- unlist(strsplit(lines[i], '\\|'))
  player_num <- as.numeric(gsub(' ', '', row1[1]))
  player_name <- str_to_title(str_trim(row1[2]))
  total_points <- str_trim(row1[3])

  row2 <- unlist(strsplit(lines[i+1], '\\|'))
  players_state <- str_trim(row2[1])
  players_pre_rating <- unlist(str_extract_all(row2[2], "[[:digit:]]+"))[2]
  player_rows <- rbind(player_rows, c(player_num, player_name, players_state, total_points, players_pre
  temp_df <- data.frame(row1[4:10])
  temp_df$player_num <- player_num
  temp_df$players_pre_rating <- players_pre_rating

  mdf <- rbind(mdf, temp_df)
}
```

Aggregate the player_rows data into a dataframe

```
df <- data.frame(player_rows)
colnames(df) <- c(’player_num’, ’player_name’, ’player_state’, ’total_points’, ’players_pre_rating’)
head(df)
```

```
##   player_num         player_name player_state total_points players_pre_rating
## 1          1            Gary Hua           ON          6.0               1794
## 2          2     Dakshesh Daruri           MI          6.0               1553
## 3          3        Aditya Bajaj           MI          6.0               1384
## 4          4 Patrick H Schilling           MI          5.5               1716
## 5          5          Hanshi Zuo           MI          5.5               1655
## 6          6         Hansen Song           OH          5.0               1686
```

Aggregate the match rows data into a dataframe

```
colnames(mdf) <- c(’wl_opponent_id’, ’player_num’, ’players_pre_rating’)
mdf$wl <- sapply(strsplit(as.character(mdf$wl_opponent_id), ’ ’), ’[’, 1)
mdf$opponent_id <- sapply(mdf$wl_opponent_id, function(x)gsub(’\\s+’, ’ ’, x))
mdf$opponent_id <- as.numeric(sapply(strsplit(as.character(mdf$opponent_id), ’ ’), ’[’, 2))
mdf$players_pre_rating <- as.numeric(mdf$players_pre_rating)
mdf <- mdf[, c(’player_num’, ’wl’, ’opponent_id’, ’players_pre_rating’)]
head(mdf)
```

```
##   player_num wl opponent_id players_pre_rating
## 1          1  W          39               1794
## 2          1  W          21               1794
## 3          1  W          18               1794
## 4          1  W          14               1794
## 5          1  W           7               1794
## 6          1  D          12               1794
```

Calculate the averages

```
final_cols <- c(’player_name’, ’player_state’, ’total_points’, ’players_pre_rating’, ’avg’)
avg_pre <- mdf %>% group_by(opponent_id) %>% summarise(avg = as.integer(mean(players_pre_rating)), .grou
df <- merge(df, avg_pre, by.x="player_num", by.y="opponent_id")[, final_cols]
head(df)
```

```
##                  player_name player_state total_points players_pre_rating  avg
## 1                   Gary Hua           ON          6.0               1794 1605
## 2                  Anvit Rao           MI          5.0               1365 1554
## 3 Cameron William Mc Leman           MI          4.5               1712 1467
## 4             Kenneth J Tack           MI          4.5               1663 1506
## 5          Torrance Henry Jr           MI          4.5               1666 1497
## 6               Bradley Shaw           MI          4.5               1610 1515
```

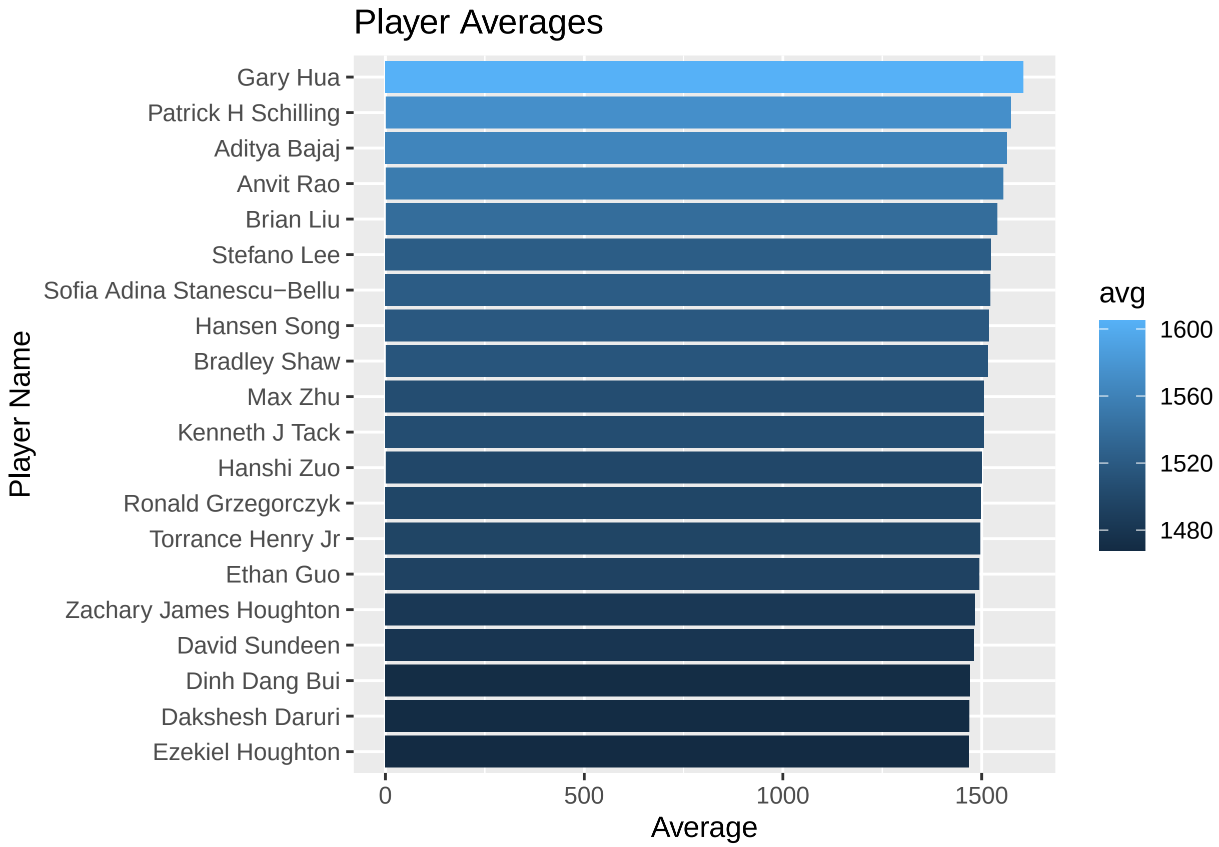
Save the result into a csv for further use

```r
write.csv(df, ’chess_data.csv’)
```

Exploratory data analysis

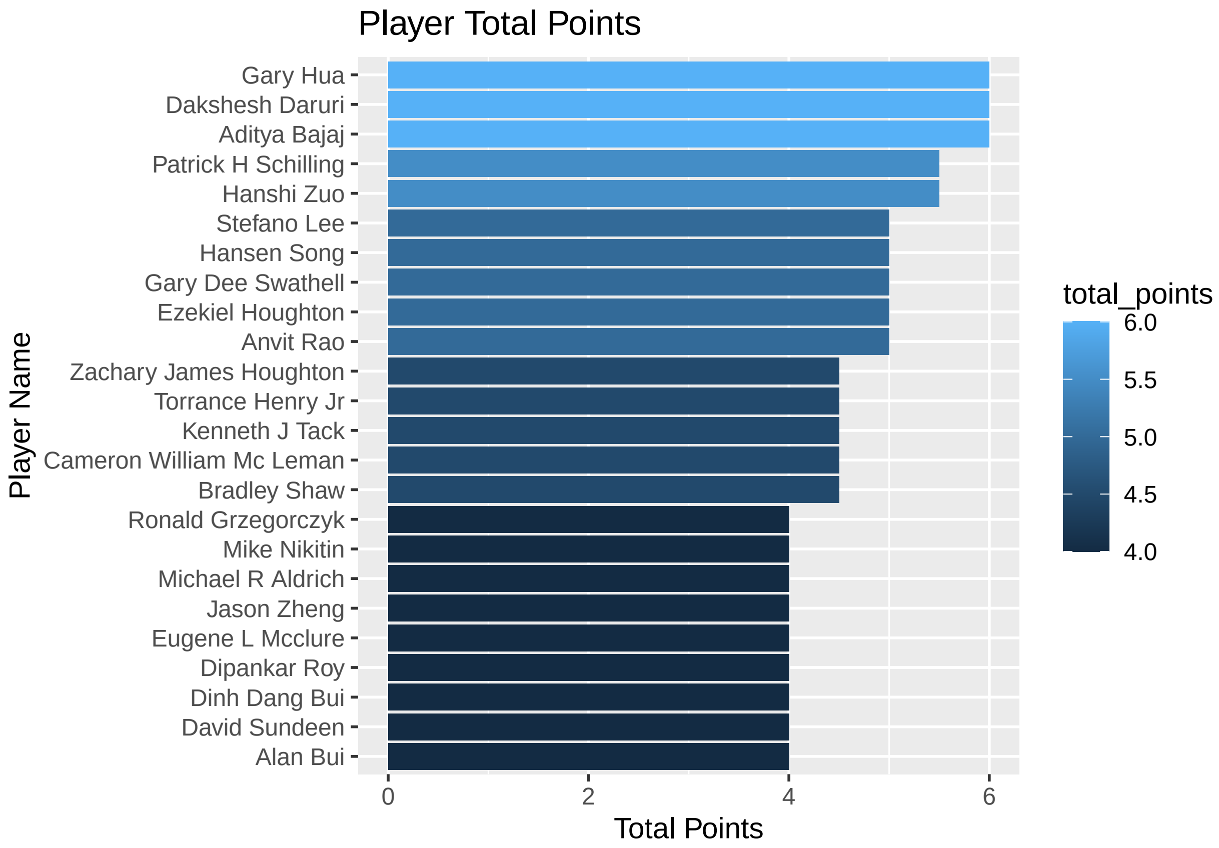
The best player by average

```
df %>% top_n(n=20, avg) %>% ggplot(aes(x=reorder(player_name, avg), y=avg, fill=avg)) + geom_col() + coo
```



The best player by total points

```
df$total_points<-as.numeric(df$total_points)
df %>% top_n(n=20, total_points) %>% ggplot(aes(x=reorder(player_name, total_points), y=total_points, fi
```



Did anyone stick out Boxplot of both playyer avg and points