

Project 1

Shamecca Marshall

2023-09-22

Loading and reading the data

```
library(stringr)
library(readr)
url = 'https://raw.githubusercontent.com/Meccamarshall/Data607/main/tournamentinfo.txt'
chesstournament = readLines(url)

head(chesstournament)

## [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 |"

tail(chesstournament)

## [1] " 63 | THOMAS JOSEPH HOSMER |1.0 |L 2|L 48|D 49|L 43|L 45|H |U |"
## [2] " MI | 15057092 / R: 1175 ->1125 | |W |B |W |B |B | | |"
## [3] "-----"
## [4] " 64 | BEN LI |1.0 |L 22|D 30|L 31|D 49|L 46|L 42|L 54|"
## [5] " MI | 15006561 / R: 1163 ->1112 | |B |W |W |B |W |B |B |"
## [6] "-----"
```

Deconstructing the Data

After noticing that the first four lines were not apart of the dataset, I decided that it needed to be deconstructed. The first four lines are followed by player info and the games played and then repeated. I am going to seperate the data into two matrices to show a cleaner dataset.

```
cTournament <- matrix(unlist(chesstournament), byrow=TRUE)

c1 <- cTournament[seq(5,length(cTournament),3)]
head(c1)

## [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|"

c2 <- cTournament[seq(6,length(cTournament),3)]
head(c2)
```

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

Capturing the data

I need to capture the values in vectors using regular expressions and string manipulation.

```
# matching first numbers
ID <- as.numeric(str_extract(c1, '\\d+'))

# matching the first combination of a letter, any amount of characters and "/"
Name <- str_extract(c1, '[A-z].{1,32}')

# extracting the name
Name <- str_trim(str_extract(Name, '.+\\s{2,}'))

# matching the first two letters (state) in the second matrix
State <- str_extract(c2, '[A-Z]{2}')

# matching at least 1 number, a period, and 1 number
TotalNumberOfPoints <- as.numeric(str_extract(c1, '\\d+\\.\\d'))

# matching the combination of "R", any characters and "-"
PreRating <- str_extract(c2, 'R:.{8,}-')

# matching first 4 numbers
PreRating <- as.numeric(str_extract(PreRating, '\\d{1,4}'))

# matching all combinations of 1 letter 2 spaces and any numbers
Rounds <- str_extract_all(c1, '[A-Z]\\s{2,}\\d+')

# matching numbers
Rounds <- str_extract_all(Rounds, '\\d+')
```

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

Calculating Average Pre Chess Rating of Opponents

```
AvgOppPreChessRating <- c()

for(i in c(1:length(Rounds))){
  AvgOppPreChessRating[i] <- round(mean(PreRating[as.numeric(Rounds[[i]])]),0)
}
AvgOppPreChessRating
```

```
## [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
```

Constructing the data frame

```
Project1 <- data.frame(ID,Name,State,TotalNumberOfPoints,PreRating,AvgOppPreChessRating)
head(Project1)
```

##	ID	Name	State	TotalNumberOfPoints	PreRating
## 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

##	AvgOppPreChessRating
## 1	1605
## 2	1469
## 3	1564
## 4	1574
## 5	1501
## 6	1519

Writing the CVS file

```
write_csv(Project1, 'ChessTournament.csv' , append = FALSE)
```