Project 1: Chess tournament cross-tables

Jie Zou

2021-02-27

Main Data Processing

Read txt file $warn = FALSE$: don't show the warnings while reading the file

```
file<-readLines("tournamentinfo.txt", warn = FALSE)
head(file)
```

```
## [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    |"
```

Data Re-organizing as we can see that the file has lots of useless dashes, and we don't care about the titles. Therefore, we are going to read the file by lines. To do so, we get the **sequence of lines** that we need(e.g. line 5, 6, 8, 9, 11, 12, etc).

```
line1 <- c(seq(5, length(file), by = 3))
line2 <- c(seq(6, length(file), by = 3))
head(line1)
```

```
## [1]  5  8 11 14 17 20
```

```
head(line2)
```

```
## [1]  6  9 12 15 18 21
```

split each data entry into two lines, **line1** will contains *[pair num], [player name], [total]* and [rounds], where **line2** will contains *[state], [USCF ID / Rtg (pre->post)], [letter result]*

```
head(file[line1])
```

```
## [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(file[line2])
```

```
## [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    |"
```

Data Extracting we extract the name from line1 we've read

([|]).*?\1: "|" follow by any characters or spaces and finish with "|"

```
name <- str_extract(file[line1], "([|]).*?\\1")
head(name)
```

```
## [1] "| GARY HUA                         |" "| DAKSHESH DARURI                |"
## [3] "| ADITYA BAJAJ                      |" "| PATRICK H SCHILLING            |"
## [5] "| HANSHI ZUO                        |" "| HANSEN SONG                    |"
```

we can see that the extracted data has "|" which we don't need. So, I **replace "|" with ""**

```
name<-str_replace_all(name, "[|]", "")
head(name)
```

```
## [1] " GARY HUA                         " " DAKSHESH DARURI                "
## [3] " ADITYA BAJAJ                     " " PATRICK H SCHILLING            "
## [5] " HANSHI ZUO                       " " HANSEN SONG                    "
```

After that, the data still contains extra spaces in both beginning and the end of strings. For the sake of aesthetics, I **remove the spaces** from two sides. (Well, that is totally fine if you don't want to bother with spaces)

```
name<-str_trim(name)
head(name)
```

```
## [1] "GARY HUA"            "DAKSHESH DARURI"     "ADITYA BAJAJ"
## [4] "PATRICK H SCHILLING" "HANSHI ZUO"          "HANSEN SONG"
```

The same procedure is used to extracting **state**, **total points** and **pre-ratings**

```
state<-str_trim(str_replace_all(str_extract(file[line2], ".{3}[|]"), "[|]", ""))
head(state)
```

```
## [1] "ON" "MI" "MI" "MI" "MI" "OH"
```

```
total_pts<-str_extract(file[line1], "\\d+\\.\\d+")
head(total_pts)
```

```
## [1] "6.0" "6.0" "6.0" "5.5" "5.5" "5.0"
```

```
pre_rating <- str_trim(str_replace_all(str_extract(file[line2], ":.\\d*.+?[-]"), ":|[-]|P\\d+", ""))
head(pre_rating)
```

```
## [1] "1794" "1553" "1384" "1716" "1655" "1686"
```

Date Reformation create the data frame from the data we just extracted above

```
tournament<-data.frame(name, state, total_pts, pre_rating)
```

Sub-data Processing

The purpose of sub data In my opinion, to calculate the average pre chess rating of opponents, it is easier to bind the pair num of player and pair num of opponents. (if you are not sure what I am talking about, here is my approach)

Approach (1). **extract all numeric number** in line1

```
rounds<-str_extract_all(file[line1], "\\d+")
r = rounds # make a copy
head(r)
```

```
## [[1]]
##  [1] "1"  "6"  "0"  "39" "21" "18" "14" "7"  "12" "4"
##
## [[2]]
##  [1] "2"  "6"  "0"  "63" "58" "4"  "17" "16" "20" "7"
##
## [[3]]
##  [1] "3"  "6"  "0"  "8"  "61" "25" "21" "11" "13" "12"
##
## [[4]]
##  [1] "4"  "5"  "5"  "23" "28" "2"  "26" "5"  "19" "1"
##
## [[5]]
##  [1] "5"  "5"  "5"  "45" "37" "12" "13" "4"  "14" "17"
##
## [[6]]
##  [1] "6"  "5"  "0"  "34" "29" "11" "35" "10" "27" "21"
```

(2). we know that the first three numeric numbers represent [pair num] and [total points], and there is no need to use these here, because all I care about is the pair number of opponents. I **eliminate unnecessary numbers and create a new data frame**

```
r1<- data.frame()
for(i in r){
  a<-i[4]
  b<-i[5]
  c<-i[6]
  d<-i[7]
```

```
  e<-i[8]
  f<-i[9]
  g<-i[10]
  r1<-rbind(r1, c(a,b,c,d,e,f,g))
}
names(r1)<-c("1","2","3","4","5","6","7")
head(r1)
```

```
##    1  2  3  4  5  6  7
## 1 39 21 18 14  7 12  4
## 2 63 58  4 17 16 20  7
## 3  8 61 25 21 11 13 12
## 4 23 28  2 26  5 19  1
## 5 45 37 12 13  4 14 17
## 6 34 29 11 35 10 27 21
```

(3). Now, we know that each cell correspond to an opponent, each row is corresponding to the opponents whose player play against with. Therefore,

i. we loop through the whole dataset, and find the pre_ratings are associated with individual opponents with the same pair num.

ii. set up a counter to count the number of opponents that player has played against with.

iii. I sum them up, take the mean(sum/count) and round them to whole number

iv. store the data into variable

```
avg_pre_rating<-NULL

for( i in 1:nrow(r1)){
  count<-0
  total<-0
  for (j in 1:ncol(r1)){
    temp <- r1[i, j]
    if(!is.na(temp)){
      total<-total+as.integer(tournament$pre_rating[as.integer(temp)])
      count<-count+1
    }
  }
  avg_pre_rating<-c(avg_pre_rating, round(total/count, 0))
}

avg_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
```

Data Merging

merge the sub data which we just calculated into the main data set

```
tournament<-tournament%>%mutate(avg_pre_rating = avg_pre_rating)
```

Export Data

```
write.csv(tournament, "p1_chess.csv", row.names = FALSE)
```