project 1, Data607

Mehdi Khan

September 21, 2017

load the data

```
library(stringr)
tournamentInfo <- read.table("C:\\temp\\tournamentinfo.txt", sep = "|",
    stringsAsFactors = FALSE, fill = TRUE)
head(tournamentInfo)
```

```
##                                                                                           V1
## 1 -----------------------------------------------------------------------------------------
## 2                                                                                     Pair 
## 3                                                                                     Num  
## 4 -----------------------------------------------------------------------------------------
## 5                                                                                        1 
## 6                                                                                       ON 
##                                   V2    V3    V4    V5    V6    V7    V8
## 1                                                                       
## 2  Player Name                      Total Round Round Round Round Round 
## 3  USCF ID / Rtg (Pre->Post)         Pts    1     2     3     4     5   
## 4                                                                       
## 5  GARY HUA                         6.0   W  39 W  21 W  18 W  14 W   7 
## 6  15445895 / R: 1794   ->1817      N:2   W     B     W     B     W     
##      V9   V10 V11
## 1              NA
## 2 Round Round  NA
## 3   6     7    NA
## 4              NA
## 5 D  12 D   4  NA
## 6 B     W      NA
```

data clean up and modifications

```
## location of rows that has no data but only the dashed lines. A
## pattern match consisting of only dashes is used to locate the rows,
## the length 6 is arbitrary:
removeRows <- str_detect(tournamentInfo$V1[1:nrow(tournamentInfo)], "[-]{6,}")

## removing the rows with only dashed lines:
tournamentInfo <- tournamentInfo[!removeRows, ]

## removing first and second rows with unneccasry information:
tournamentInfo <- tournamentInfo[-1:-2, ]

## removing empty column:
tournamentInfo$V11 <- NULL
```

```r
## getting the indexes of alternate rows that have the state names
## (abbreviation of states in usa with two upper case letters ) and
## ratings:
rowval <- grep("[A-Z]{2}", tournamentInfo$V1[1:nrow(tournamentInfo)])

## adding three additional columns and removing spaces from some
## columns:
addColumns <- c("state", "pre_rating", "avg_opponent")
tournamentInfo[, addColumns] <- NA
tournamentInfo$V2 <- str_trim(tournamentInfo$V2, side = "both")
tournamentInfo$V3 <- str_trim(tournamentInfo$V3, side = "both")

## removing spaces on both sides from the values in first column:
tournamentInfo$V1 <- str_trim(tournamentInfo$V1, side = "both")

## Populating state column with extracted values from every second rows:
tournamentInfo[rowval - 1, ]$state <- with(tournamentInfo, str_extract(V1[rowval],
    "[A-Z]{2}"))

## Populating pre_rating column with extracting values from every second
## rows:
tournamentInfo[rowval - 1, ]$pre_rating <- with(tournamentInfo, as.numeric(str_trim(str_sub(str_extract
    "(R:[\\s]*([0-9]+))"), 3), side = "both")))


## Populating oppo_rating column with the average Pre Rating of
## Opponents. First opponents player numbers are collected, opponents
## who actually played a game were considered. Then the pre ratings for
## each of the opponents were found and finally their average ratings
## were calculated:
for (i in 1:length(rowval)) {

    playerno <- str_trim((str_sub((str_extract_all(tournamentInfo[rowval[i] -
        1, 4:10], "[WLD]{1}[:space:]*[0-9]{1,}")), 2)), side = "both")
    avg_rate <- as.integer((sum(as.numeric(tournamentInfo[tournamentInfo$V1 %in%
        playerno, 12])))/length(playerno))

    tournamentInfo[rowval[i] - 1, ]$avg_opponent <- avg_rate
}


## deleting all the rows that are no longer needed:
tournamentInfo <- tournamentInfo[-rowval, ]

## renaming required columns:
names(tournamentInfo)[2:3] = c("name", "total_pts")

## removing columns that are no longer needed:
tournamentInfo <- subset(tournamentInfo, select = c(2, 3, 11, 12, 13))

## removing row names to reflect row numbers:
row.names(tournamentInfo) <- c()
```

```
## reordering column names:
tournamentInfo <- tournamentInfo[c(1, 3, 2, 4, 5)]

head(tournamentInfo)
```

```
##                   name state total_pts pre_rating avg_opponent
## 1             GARY HUA    ON       6.0       1794         1605
## 2      DAKSHESH DARURI    MI       6.0       1553         1469
## 3         ADITYA BAJAJ    MI       6.0       1384         1563
## 4  PATRICK H SCHILLING    MI       5.5       1716         1573
## 5           HANSHI ZUO    MI       5.5       1655         1500
## 6          HANSEN SONG    OH       5.0       1686         1518
```

Saving the data in a csv file:

```
write.csv(tournamentInfo, "C:\\temp\\chessPlayerInfo.txt", row.names = FALSE)
```