DATA 607 Project 1

Samuel I Kigamba

September 22, 2019

Project 1

In this project, youre given a text file with chess tournament results where the information has some structure. Your job is to create an R Markdown file that generates a .CSV file (that could for example be imported into a SQL database) with the following information for all of the players: Players Name, Players State, Total Number of Points, Players Pre-Rating, and Average Pre Chess Rating of Opponents For the first player, the information would be: Gary Hua, ON, 6.0, 1794, 1605 1605 was calculated by using the pre-tournament opponents ratings of 1436, 1563, 1600, 1610, 1649, 1663, 1716, and dividing by the total number of games played. If you have questions about the meaning of the data or the results, please post them on the discussion forum. Data science, like chess, is a game of back and forth The chess rating system (invented by a Minnesota statistician named Arpad Elo) has been used in many other contexts, including assessing relative strength of employment candidates by human resource departments. You may substitute another text file (or set of text files, or data scraped from web pages) of similar or greater complexity, and create your own assignment and solution. You may work in a small team. All of your code should be in an R markdown file (and published to rpubs.com); with your data accessible for the person running the script.

Step 1

Download the chess .txt file from the class link provided (or from any other source per the instruction above) and uploaded it into a repository on github (in my case https://raw.githubusercontent.com/igukusamuel/DATA-607-Project-1/master/tournamentinfo.txt. Use RCurl to load the txt data into R from Github or other link of your choosing and perform the below data manipulations using regular expressions.

```
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.5.3
```

```
## Loading required package: bitops
```

```
chess_T <- getURL("https://raw.githubusercontent.com/nabilahossain/Class-IS607/master/Project%201/tourna
#chess
```

Step 2

Use regular expressions to pull the (a) player's name and (b) the state. Load the stingr package and use the str_extract_all and the str_replace_all functions to extract and to format/clean the data from the .txt file

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

```
name <- unlist(str_extract_all(chess_T, "\\| [[:upper:]- ]{4,} \\|"))
name <- str_replace_all(name, pattern = "(\\| )|([[:space:]]{1,}\\|)", replacement = "")
head(name[25:30])
```

```
## [1] "LOREN SCHWIEBERT"           "MAX ZHU"
## [3] "GAURAV GIDWANI"             "SOFIA ADINA STANESCU-BELLU"
## [5] "CHIEDOZIE OKORIE"           "GEORGE AVERY JONES"
```

```
state <- unlist(str_extract_all(chess_T, "\\|[[:space:]]{1,}[[A-Z]]{2} \\|"))
state <- str_replace_all(state, pattern = "(\\|[[:space:]]{1,})|([[:space:]]{1,}\\|)", replacement = ""
head(state, 10)
```

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

Step 3

Use regular expressions and the stingr functions mentioned in step 2 to extract (a) player's points and (b) pre-rating.

```
total_points <- unlist(str_extract_all(chess_T, "\\|[[:digit:].[:digit:]]{3}[[:space:]]{1,}\\|"))
total_points <- str_replace_all(total_points, pattern = "(\\|)|([[:space:]]{1,}\\|)", replacement = "")
head(total_points, 10)
```

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

```
pre_rating <- unlist(str_extract_all(chess_T, "[:] [[:alnum:] ]{2,9}\\-\\>"))
pre_rating <- str_replace_all(pre_rating, pattern = "(\\: )|(\\s{1,}\\-\\>)|([O-Q]\\d{1,2})|(\\-\\>)", r
pre_rating <- as.numeric(pre_rating)
head(pre_rating, 15)
```

```
##  [1] 1794 1553 1384 1716 1655 1686 1649 1641 1411 1365 1712 1663 1666 1610
## [15] 1220
```

Step 4

Extract, using same functions as above, each players number, needed later to join/combine tables and then create a table with (a) player's number, (b) name, (c) state, (d) total points and (e) pre-rating as the headers

```
player_num <- unlist(str_extract_all(chess_T, "\\d{1,2}\\s\\|"))
player_num <- str_replace_all(player_num, pattern = "(\\s\\|)", replacement = "")
player_num <- as.numeric(player_num)
table_1 <- data.frame(player_num = player_num, name = name, state = state, total_pts = total_points, pre
head(table_1)
```

```
##   player_num                name state total_pts 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
```

Step 5

Extracting the first row from the txt file and creating a table named rounds with 10 columns. Here only extract the information found in the first line of each player and save it in the table of rounds.

```
r_1 <- unlist(str_extract_all(chess_T, "\\d{1,2}\\s\\|(.+?)\\|[[:upper:]]{1}\\s{1,4}[[:digit:]]{0,3}\\|
r_1 <- str_replace_all(r_1, pattern = "(\\|[[:upper:]]{1} )", replacement = "\\|")
head(r_1, 10)
```

```
##  [1] "1 | GARY HUA                       |6.0  | 39| 21| 18| 14|  7| 12|  4|\n"
##  [2] "2 | DAKSHESH DARURI                 |6.0  | 63| 58|  4| 17| 16| 20|  7|\n"
##  [3] "3 | ADITYA BAJAJ                    |6.0  |  8| 61| 25| 21| 11| 13| 12|\n"
##  [4] "4 | PATRICK H SCHILLING             |5.5  | 23| 28|  2| 26|  5| 19|  1|\n"
##  [5] "5 | HANSHI ZUO                      |5.5  | 45| 37| 12| 13|  4| 14| 17|\n"
##  [6] "6 | HANSEN SONG                     |5.0  | 34| 29| 11| 35| 10| 27| 21|\n"
##  [7] "7 | GARY DEE SWATHELL               |5.0  | 57| 46| 13| 11|  1|  9|  2|\n"
##  [8] "8 | EZEKIEL HOUGHTON                |5.0  |  3| 32| 14|  9| 47| 28| 19|\n"
##  [9] "9 | STEFANO LEE                     |5.0  | 25| 18| 59|  8| 26|  7| 20|\n"
## [10] "10 | ANVIT RAO                       |5.0  | 16| 19| 55| 31|  6| 25| 18|\n"
```

```
r_1 <- str_replace_all(r_1, pattern = "([[:space:]]{2})|(\\|\n)", replacement = "")
rounds <- data.frame(str_split_fixed(r_1, "\\|", 10))
names(rounds) <- c("num", "name", "total_pts", "round1", "round2", "round3", "round4", "round5", "round
tail(rounds, 4)
```

```
##    num                  name total_pts round1 round2 round3 round4 round5
## 61 61         JEZZEL FARKAS        1.5     32      3     54     47     42
## 62 62         ASHWIN BALAJI        1.0     55
## 63 63   THOMAS JOSEPH HOSMER       1.0      2     48     49     43     45
## 64 64                BEN LI        1.0     22     30     31     49     46
##    round6 round7
## 61     30     37
## 62
## 63
## 64     42     54
```

Step 6

Install package reshape2 to combine players opponents information from the 7 columns into 1 and use the subset function to eliminate missing information. combine all the seven rounds of information extracted in table rounds in step 5 and create a second table named table2.

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.5.3
```

```
r_3 <- data.frame(rounds[c(1, 4:10)])
r_3$num <- str_replace_all(r_3$num, pattern = "\\s{1,}(\\d{1,2})", replacement = "\\1")
r_4 <- melt(r_3,id.vars="num", variable.name = "rounds", value.name = "opponent_number" )
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
tail(r_4)
```

```
##     num rounds opponent_number
## 443 59  round7              44
## 444 60  round7
## 445 61  round7              37
## 446 62  round7
## 447 63  round7
## 448 64  round7              54
```

```
table_2 <- subset(r_4, opponent_number != " ")
table_2$num <- as.numeric(table_2$num)
tail(table_2)
```

```
##     num rounds opponent_number
## 439  55 round7              43
## 440  56 round7              42
## 442  58 round7              45
## 443  59 round7              44
## 445  61 round7              37
## 448  64 round7              54
```

Step 7

install package sqldf to join table 1 and table 2 to get the opponents pre-ratings and name set this data into table 3.

```
library(sqldf)
```

```
## Warning: package 'sqldf' was built under R version 3.5.3
```

```
## Loading required package: gsubfn
```

```
## Warning: package 'gsubfn' was built under R version 3.5.3
```

```
## Loading required package: proto
```

```
## Warning: package 'proto' was built under R version 3.5.3
```

```
## Loading required package: RSQLite
```

```
## Warning: package 'RSQLite' was built under R version 3.5.3
```

```
table_3 <- sqldf("select t_2.num as 'player_num', t_1.name as 'opponent_name', t_2.rounds, t_2.opponent
       left join table_1 t_1
        on t_2.opponent_number =  t_1.player_num
       order by t_2.num asc")
head(table_3)
```

```
##   player_num     opponent_name rounds opponent_number opponent_pre_rating
## 1          1     JOEL R HENDON round1              39                1436
## 2          1     DINH DANG BUI round2              21                1563
## 3          1     DAVID SUNDEEN round3              18                1600
## 4          1      BRADLEY SHAW round4              14                1610
## 5          1 GARY DEE SWATHELL round5               7                1649
## 6          1    KENNETH J TACK round6              12                1663
```

Step 8

use stats package and the aggregate function to find each players average pre-rating and set it as table 4. Use the subset fuction to acomplish this.

```
library(stats)
table_4 <- aggregate(opponent_pre_rating ~ player_num, data = table_3, FUN = 'mean')
head(table_4)
```

```
##   player_num opponent_pre_rating
## 1          1            1605.286
## 2          2            1469.286
## 3          3            1563.571
## 4          4            1573.571
## 5          5            1500.857
## 6          6            1518.714
```

Step 9

Using the sqldf package join table 1 and table 4 to obtain a table of the required information and in the reqired format. use the format fuction to round off the decimals to your units of choice.

```
Chess_Tournament <- sqldf("select t_1.name as 'Player_Name', t_1.state as 'Player_State', t_1.total_pts
      from table_1 t_1 left join table_4 t_4
         on t_4.player_num =  t_1.player_num")
Chess_Tournament$Opponents_Average_Pre_Rating <- format(round(Chess_Tournament$Opponents_Average_Pre_Ra
head(Chess_Tournament)
```

```
##           Player_Name Player_State Total_Points Player_Pre-Rating
## 1            GARY HUA           ON          6.0              1794
## 2     DAKSHESH DARURI           MI          6.0              1553
## 3        ADITYA BAJAJ           MI          6.0              1384
## 4 PATRICK H SCHILLING           MI          5.5              1716
## 5          HANSHI ZUO           MI          5.5              1655
## 6         HANSEN SONG           OH          5.0              1686
```

```
##   Opponents_Average_Pre_Rating
## 1                       1605.3
## 2                       1469.3
## 3                       1563.6
## 4                       1573.6
## 5                       1500.9
## 6                       1518.7
```

Step 10

Finally save a .csv file into local drive and upload into github repository. See my uploaded file under https://github.com/igukusamuel/DATA-607-Project-1/blob/master/DATA_607_Project_1.csv

```
write.csv(Chess_Tournament, file = "C:/Users/iguku/Google Drive/R and SQL/DATA 607 Project 1/DATA_607_P
```