Data 607 - Project 1

Leticia Salazar

2021-09-19

Project 1

In this project, you're 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:

- Player's Name
- Player's State
- Total Number of Points
- Player's Pre-Rating
- 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 opponent's ratings of 1436, 1536, 1600, 1610, 1649, 1663, 1716, and dividing by the total number of games played.

Objectives: From the cross-tables, choose only the player's opponents and averge pre-rating of their opponents, both for players who played all of the scheduled games (8 points), and for players who had one or more unplayed games (e.g. byes, forfeits) (5 points). Are the average ratings presented to nearest full-point accuracy? (2 points)

Using the provided ELO calculation, determine each player's expected results (number of points), based on his or her pre-tournament rating, and the average pre-tournament rating for all of the player's opponents. Which player scored the most points relative to his or her expected results? (3 extra-credit points)

```
#Load libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(dplyr)
```

```
#Load the dataset

theUrl <- "https://raw.githubusercontent.com/letisalba/Data-607-Project1/main/tournament.txt"

tournament <- read.delim(file = theUrl, header = FALSE, sep = "|", skip = 3) #Separating by "|" and ski

head(tournament)
```

Loading Dataset

```
##                                                                                          V1
## 1 ---------------------------------------------------------------------------------------
## 2                                                                                         1
## 3                                                                                        ON
## 4 ---------------------------------------------------------------------------------------
## 5                                                                                         2
## 6                                                                                        MI
##                                 V2    V3    V4    V5    V6    V7    V8    V9
## 1
## 2  GARY HUA                          6.0   W  39 W  21 W  18 W  14 W   7 D  12
## 3  15445895 / R: 1794   ->1817      N:2   W     B     W     B     W     B
## 4
## 5  DAKSHESH DARURI                   6.0   W  63 W  58 L   4 W  17 W  16 W  20
## 6  14598900 / R: 1553   ->1663      N:2   B     W     B     W     B     W
##     V10 V11
## 1        NA
## 2 D   4  NA
## 3 W      NA
## 4        NA
## 5 W   7  NA
## 6 B      NA
```

```
#Get a glimpse of the dataset
#Note that there's plenty of dashes and NA's which will be removed
glimpse(tournament)
```

```
## Rows: 193
## Columns: 11
## $ V1  <chr> "-----------------------------------------------------------------------~
## $ V2  <chr> "", " GARY HUA                          ", " 15445895 / R: 1794   ->~
## $ V3  <chr> "", "6.0  ", "N:2  ", "", "6.0  ", "N:2  ", "", "6.0  ", "N:2  ", ~
## $ V4  <chr> "", "W  39", "W    ", "", "W  63", "B    ", "", "L   8", "W    ", ~
## $ V5  <chr> "", "W  21", "B    ", "", "W  58", "W    ", "", "W  61", "B    ", ~
```

```
## $ V6  <chr> "", "W  18", "W    ", "", "L   4", "B    ", "", "W  25", "W    ", ~
## $ V7  <chr> "", "W  14", "B    ", "", "W  17", "W    ", "", "W  21", "B    ", ~
## $ V8  <chr> "", "W   7", "W    ", "", "W  16", "B    ", "", "W  11", "W    ", ~
## $ V9  <chr> "", "D  12", "B    ", "", "W  20", "W    ", "", "W  13", "B    ", ~
## $ V10 <chr> "", "D   4", "W    ", "", "W   7", "B    ", "", "W  12", "W    ", ~
## $ V11 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

```
#Removing dashes
tournament2 <- tournament[!grepl("---", tournament[ ,1]), ]
head(tournament2)
```

Data Wrangling

```
##       V1                                V2    V3    V4    V5    V6    V7    V8
## 2     1   GARY HUA                        6.0   W  39 W  21 W  18 W  14 W   7
## 3    ON   15445895 / R: 1794   ->1817      N:2   W     B     W     B     W
## 5     2   DAKSHESH DARURI                 6.0   W  63 W  58 L   4 W  17 W  16
## 6    MI   14598900 / R: 1553   ->1663      N:2   B     W     B     W     B
## 8     3   ADITYA BAJAJ                    6.0   L   8 W  61 W  25 W  21 W  11
## 9    MI   14959604 / R: 1384   ->1640      N:2   W     B     W     B     W
##      V9   V10 V11
## 2 D  12 D   4  NA
## 3 B     W      NA
## 5 W  20 W   7  NA
## 6 W     B      NA
## 8 W  13 W  12  NA
## 9 B     W      NA
```

```
#For one player there is two rows that split between Player's name, points and rounds while the second
#Start by splitting dataset into odd and even rows to be able to clean them individually

#Split dataset into odd row only

odd_rows <- tournament2[seq(1, nrow(tournament2), 2), ]

#Removing columns not needed in odd_rows
odd_rows <- odd_rows[, -c(1, 11)]

#Add column ID named Player Numbers
#This is going to help merge the two datasets later on
odd_rows$Player_Number <- seq(1, 64, length.out = dim(odd_rows)[1])



head(odd_rows)
```

Odd Rows Cleansing

```
##                                  V2    V3    V4    V5    V6    V7    V8    V9
## 2   GARY HUA                     6.0   W  39 W  21 W  18 W  14 W   7 D  12
## 5   DAKSHESH DARURI              6.0   W  63 W  58 L   4 W  17 W  16 W  20
## 8   ADITYA BAJAJ                 6.0   L   8 W  61 W  25 W  21 W  11 W  13
## 11  PATRICK H SCHILLING          5.5   W  23 D  28 W   2 W  26 D   5 W  19
## 14  HANSHI ZUO                   5.5   W  45 W  37 D  12 D  13 D   4 W  14
## 17  HANSEN SONG                  5.0   W  34 D  29 L  11 W  35 D  10 W  27
##      V10 Player_Number
## 2  D   4             1
## 5  W   7             2
## 8  W  12             3
## 11 D   1             4
## 14 W  17             5
## 17 W  21             6
```

```
#Create new column names for even_rows dataset and rename
colnames(odd_rows) <- c("Player_Name", "Points", "Opp_1", "Opp_2", "Opp_3", "Opp_4", "Opp_5", "Opp_6",

#Reorder columns in 'odd_rows' to place Player_Number before Player_Name
odd_rows2 <- odd_rows[c("Player_Number", "Player_Name", "Points", "Opp_1", "Opp_2", "Opp_3", "Opp_4", "
head(odd_rows2)
```

```
##    Player_Number                      Player_Name Points Opp_1 Opp_2 Opp_3
## 2              1  GARY HUA                           6.0   W  39 W  21 W  18
## 5              2  DAKSHESH DARURI                    6.0   W  63 W  58 L   4
## 8              3  ADITYA BAJAJ                       6.0   L   8 W  61 W  25
## 11             4  PATRICK H SCHILLING                5.5   W  23 D  28 W   2
## 14             5  HANSHI ZUO                         5.5   W  45 W  37 D  12
## 17             6  HANSEN SONG                        5.0   W  34 D  29 L  11
##    Opp_4 Opp_5 Opp_6 Opp_7
## 2  W  14 W   7 D  12 D   4
## 5  W  17 W  16 W  20 W   7
## 8  W  21 W  11 W  13 W  12
## 11 W  26 D   5 W  19 D   1
## 14 D  13 D   4 W  14 W  17
## 17 W  35 D  10 W  27 W  21
```

```
#Split dataset into even rows only
even_rows <- tournament2[seq(2, nrow(tournament2), 2), ]
#Remove columns not needed
even_rows <- even_rows[, -c(3:11)]

# Add column ID to match Player Numbers
# This is going to help merge the two datasets later on
even_rows$Player_Number <- seq(1, 64, length.out = dim(even_rows)[1])

head(even_rows)
```

Even Rows Cleansing

```
##        V1                                V2 Player_Number
```

```
## 3     ON   15445895 / R: 1794   ->1817                 1
## 6     MI   14598900 / R: 1553   ->1663                 2
## 9     MI   14959604 / R: 1384   ->1640                 3
## 12    MI   12616049 / R: 1716   ->1744                 4
## 15    MI   14601533 / R: 1655   ->1690                 5
## 18    OH   15055204 / R: 1686   ->1687                 6
```

```
# Create new column names for even_rows dataset and rename
colnames(even_rows) <- c("State", "Ratings_Pre/Post", "Player_Number")

# Reorder columns in even_rows so that Player_Number is before State
even_rows2 <- even_rows[c("Player_Number", "State","Ratings_Pre/Post")]
head(even_rows2)
```

```
##    Player_Number  State                  Ratings_Pre/Post
## 3              1    ON   15445895 / R: 1794   ->1817
## 6              2    MI   14598900 / R: 1553   ->1663
## 9              3    MI   14959604 / R: 1384   ->1640
## 12             4    MI   12616049 / R: 1716   ->1744
## 15             5    MI   14601533 / R: 1655   ->1690
## 18             6    OH   15055204 / R: 1686   ->1687
```

```
#Join odd_rows and even_rows datasets by "Player_Number" to form one table
tournament_subset <- left_join(odd_rows, even_rows2, by = "Player_Number")

#Reorder columns
tournament_subset2 <- tournament_subset[c("Player_Number", "Player_Name", "State", "Points", "Ratings_Pr
head(tournament_subset2)
```

```
##   Player_Number                      Player_Name  State Points
## 1             1  GARY HUA                          ON   6.0
## 2             2  DAKSHESH DARURI                   MI   6.0
## 3             3  ADITYA BAJAJ                      MI   6.0
## 4             4  PATRICK H SCHILLING               MI   5.5
## 5             5  HANSHI ZUO                        MI   5.5
## 6             6  HANSEN SONG                       OH   5.0
##                    Ratings_Pre/Post Opp_1 Opp_2 Opp_3 Opp_4 Opp_5 Opp_6 Opp_7
## 1  15445895 / R: 1794   ->1817      W  39 W  21 W  18 W  14 W   7 D  12 D   4
## 2  14598900 / R: 1553   ->1663      W  63 W  58 L   4 W  17 W  16 W  20 W   7
## 3  14959604 / R: 1384   ->1640      L   8 W  61 W  25 W  21 W  11 W  13 W  12
## 4  12616049 / R: 1716   ->1744      W  23 D  28 W   2 W  26 D   5 W  19 D   1
## 5  14601533 / R: 1655   ->1690      W  45 W  37 D  12 D  13 D   4 W  14 W  17
## 6  15055204 / R: 1686   ->1687      W  34 D  29 L  11 W  35 D  10 W  27 W  21
```

```
# Spliting Ratings_Pre/Post to get Pre and Post ratings alone
split_ratings <- str_split_fixed(even_rows2$`Ratings_Pre/Post`, "-", 2)
colnames(split_ratings) <- c("Pre-Rating", "Post-Rating")

#Remove numbers before "R:"
PreRating <- unlist(str_remove_all(split_ratings, "\\d+\\d [/]"))
```

```
#Extract Player's Pre-Ratings
PreRating <- unlist(str_extract_all(split_ratings, "R: \\d+\\d"))

#Remove "R:"
PreRating2 <- unlist(str_remove_all(PreRating, "R:"))

#Make values as.numeric
PreRatings <- as.numeric(PreRating2)
PreRatings
```

Cleansing Ratings_Pre/Post column

```
##  [1] 1794 1553 1384 1716 1655 1686 1649 1641 1411 1365 1712 1663 1666 1610 1220
## [16] 1604 1629 1600 1564 1595 1563 1555 1363 1229 1745 1579 1552 1507 1602 1522
## [31] 1494 1441 1449 1399 1438 1355 1423 1436 1348 1403 1332 1283 1199 1242 1362
## [46] 1382 1291 1056 1011 1393 1270 1186 1153 1092 1530 1175 1163
```

```
#Define new values
New_PreRatings = c("1794", "1553", "1384", "1716"," 1655", "1686", "1649", "1641", "1411", "1365", "171

#cbind tables together with tournament_subset2
tournament_subset3 <-  cbind(tournament_subset2, New_PreRatings)

head(tournament_subset3)
```

```
##   Player_Number                        Player_Name  State Points
## 1             1  GARY HUA                            ON   6.0
## 2             2  DAKSHESH DARURI                     MI   6.0
## 3             3  ADITYA BAJAJ                        MI   6.0
## 4             4  PATRICK H SCHILLING                 MI   5.5
## 5             5  HANSHI ZUO                          MI   5.5
## 6             6  HANSEN SONG                         OH   5.0
##                     Ratings_Pre/Post Opp_1 Opp_2 Opp_3 Opp_4 Opp_5 Opp_6 Opp_7
## 1  15445895 / R: 1794   ->1817      W  39 W  21 W  18 W  14 W   7 D  12 D   4
## 2  14598900 / R: 1553   ->1663      W  63 W  58 L   4 W  17 W  16 W  20 W   7
## 3  14959604 / R: 1384   ->1640      L   8 W  61 W  25 W  21 W  11 W  13 W  12
## 4  12616049 / R: 1716   ->1744      W  23 D  28 W   2 W  26 D   5 W  19 D   1
## 5  14601533 / R: 1655   ->1690      W  45 W  37 D  12 D  13 D   4 W  14 W  17
## 6  15055204 / R: 1686   ->1687      W  34 D  29 L  11 W  35 D  10 W  27 W  21
##   New_PreRatings
## 1           1794
## 2           1553
## 3           1384
## 4           1716
## 5           1655
## 6           1686
```

```
#Remove Ratings_Pre/Post Column
tournament_subset4 <- tournament_subset3[, -c(5)]

#Reorder column names so that New_PreRatings comes after Points and print new dataset
tournament_subset5 <- tournament_subset4[c("Player_Number", "Player_Name", "State", "Points", "New_PreRa
head(tournament_subset5)
```

```
##   Player_Number                      Player_Name  State Points New_PreRatings
## 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
##   Opp_1 Opp_2 Opp_3 Opp_4 Opp_5 Opp_6 Opp_7
## 1 W  39 W  21 W  18 W  14 W   7 D  12 D   4
## 2 W  63 W  58 L   4 W  17 W  16 W  20 W   7
## 3 L   8 W  61 W  25 W  21 W  11 W  13 W  12
## 4 W  23 D  28 W   2 W  26 D   5 W  19 D   1
## 5 W  45 W  37 D  12 D  13 D   4 W  14 W  17
## 6 W  34 D  29 L  11 W  35 D  10 W  27 W  21
```

```
#Start extracting Opponents columns into matrix to remove W, L, D from the Player Numbers
opponents <- tournament_subset5[6:12]
Opps <- matrix(str_extract(unlist(opponents), "\\d+"), ncol = 7)
head(Opps)
```

Cleansing Opponents columns 1 - 7

```
##      [,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"
```

```
#Name New Columns
opponents_col <- c("Opp1", "Opp2", "Opp3", "Opp4", "Opp5", "Opp6", "Opp7")

#Set as a dataframe and call in columns
Opps_df <- as.data.frame(Opps)
colnames(Opps_df) <- opponents_col

#Replace all NA's with 0
Opps_df[is.na(Opps_df)] <- 0

head(Opps_df)
```

```
##   Opp1 Opp2 Opp3 Opp4 Opp5 Opp6 Opp7
## 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
```

```
#cbind Opps_df together with tournament_subset5
tournament_subset6 <-  cbind(tournament_subset5, Opps_df)

#Remove columns 6 to 12 from tournament _subset6
tournament_subset7 <- tournament_subset6[, -c(6:12)]
head(tournament_subset7)
```

```
##   Player_Number                      Player_Name  State Points New_PreRatings
## 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
##   Opp1 Opp2 Opp3 Opp4 Opp5 Opp6 Opp7
## 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
```

```
#Calculate Opponents Averages with a For Loop
for(i in 1:nrow(tournament_subset7)){
  tournament_subset7$Opp_Avg_Ratings[i] <- round(mean(PreRatings[as.numeric(Opps_df[i,])], na.rm = TRUE)
}

head(tournament_subset7)
```

Computing Opponents Averages

```
##   Player_Number                      Player_Name  State Points New_PreRatings
## 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
##   Opp1 Opp2 Opp3 Opp4 Opp5 Opp6 Opp7 Opp_Avg_Ratings
## 1   39   21   18   14    7   12    4            1593
## 2   63   58    4   17   16   20    7            1639
## 3    8   61   25   21   11   13   12            1665
## 4   23   28    2   26    5   19    1            1574
## 5   45   37   12   13    4   14   17            1581
## 6   34   29   11   35   10   27   21            1519
```

```
#Remove columns 6 to 12 from tournament _subset7
chess_tournament <- tournament_subset7[, -c(6:12)]

#Print final dataset
head(chess_tournament, n = 6)
```

```
##   Player_Number                      Player_Name  State Points New_PreRatings
## 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
##   Opp_Avg_Ratings
## 1            1593
## 2            1639
## 3            1665
## 4            1574
## 5            1581
## 6            1519
```

```
#Write .csv file
write.csv(chess_tournament, file = "chess_tournament.csv")
```

Write .CSV File