Week4Assignment

Bin Lin

2016-9-23

str(raw_data)

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 with the following information for all of the players:: Player's Name, Player's State, Total Number of Points, Player's Pre-Rating, and Average Pre Chess Rating of Opponents

First of all, we can go to the url website of the original data: https://bbhosted.cuny.edu/bbcswebdav/pid-25842547-dt-content-rid-119781398_1/courses/SPS01_DATA_607_01_1169_1/SPS01_DATA_607_01_1169_1_ImportedContent_20160815114002/SPS01_DATA_607_01_1162_1_Impo (https://bbhosted.cuny.edu/bbcswebdav/pid-25842547-dt-content-rid-

119781398_1/courses/SPS01_DATA_607_01_1169_1/SPS01_DATA_607_01_1169_1_ImportedContent_20160815114002/SPS01_DATA_607_01_1162_1_Impo

This website can really give us an idea about how this dataset is structured. It seems to me everything is seperated by an "|". Therefore, the first step I am trying to do is to load the "stringr" package which I am going to use for regular expression. I loaded the raw data from my local machine and set the seperator to be "|".

```
library(stringr)
raw_data <- read.csv("C:/Users/blin261/Downloads/tournamentinfo.txt", header = FALSE, stringsAsFactors = FALSE, sep = "|")</pre>
```

The following codes just help me to explore the raw data.

```
head(raw_data)
##
                                                                  V1
## 1 -----
## 2
                                                                Pair
## 3
## 5
                                                                  1
## 6
##
                         V2 V3
                                 V4
                                      V5
                                          ۷6
## 1
## 2 Player Name
                            Total Round Round Round Round
## 3 USCF ID / Rtg (Pre->Post)
                            Pts
                                 1
                                      2
                                         3
## 4
## 5 GARY HUA
                            6.0 W 39 W 21 W 18 W 14 W
## 6 15445895 / R: 1794 ->1817
                            N:2 W
                                         W
##
       V10 V11
## 1
## 2 Round Round NA
## 3 6
            NA
## 4
            NΑ
## 5 D 12 D 4 NA
## 6 B
       W
```

```
class(raw_data)

## [1] "data.frame"

rownames(raw_data)
```

```
[1] "1"
                     "3"
                                  "5"
                                               "7"
               "2"
                            "4"
                                         "6"
                                                      "8"
                                                           "9"
                                                                   "10" "11"
##
    [12] "12"
                "13"
                     "14"
                            "15"
                                   "16"
                                         "17"
                                               "18"
                                                      "19"
                                                            "20"
                                                                   "21"
                                                                         "22"
                      "25"
                                   "27"
                                                                         "33"
    [23] "23"
               "24"
                            "26"
                                         "28"
                                               "29"
                                                      "30"
                                                            "31"
                                                                   "32"
##
    [34] "34"
               "35"
                      "36" "37" "38"
                                         "39"
    [45] "45"
                      "47" "48"
                                  "49"
##
               "46"
                                         "50"
                                               "51"
                                                      "52"
                                                            "53"
                                                                   "54"
                                                                         "55"
               "57"
                      "58" "59"
                                         "61"
##
    [56] "56"
                                   "60"
                                               "62"
                                                      "63"
                                                                   "65"
                                                                         "66"
    [67] "67"
                     "69" "70" "71"
                                         "72" "73"
##
               "68"
                                                      "74"
                                                                   "76"
    [78] "78"
                     "80" "81" "82"
                                               "84"
               "79"
                                         "83"
                                                     "85"
                                                                  "87"
##
                                                           "86"
                                                                         "88"
    [89] "89"
               "90"
                     "91" "92"
                                   "93"
                                         "94"
                                               "95"
## [100] "100" "101" "102" "103" "104" "105" "106" "107" "108"
## [111] "111" "112" "113" "114" "115" "116" "117" "118" "119" "120" "121"
## [122] "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132" 
## [133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143"
## [144] "144" "145" "146" "147" "148" "149" "150" "151" "152" "153" "154"
## [155] "155" "156" "157" "158" "159" "160" "161" "162" "163" "164" "165"
## [166] "166" "167" "168" "169" "170" "171" "172" "173" "174" "175" "176"
## [177] "177" "178" "179" "180" "181" "182" "183" "184" "185" "186" "187"
## [188] "188" "189" "190" "191" "192" "193" "194" "195" "196"
```

```
colnames(raw_data)
```

```
This code really helps me visualize the raw data, because the raw data contains lots of "——", which really causes a lot of confusion. More importantly, this subset tells me which variables are located at which specific rows.
```

[1] "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8" "V9" "V10" "V11"

```
raw_data[-seq(1,196,3), 1:10]
```

##		V1		V2	V3		V4		V5		V6		V7
##		Pair	Player Name		Total					Rou	nd		
##		Num	USCF ID / Rtg (Pre->Post	:)	Pts		1		2	3			4
##		1 ON	GARY HUA 15445895 / R: 1794 ->1	.817	6.0 N:2	W	39	w B	21	W W	18	w B	14
##		2	DAKSHESH DARURI	.027	6.0	W			58		4		17
##	9	MI	14598900 / R: 1553 ->1	.663	N:2	В		W		В		W	
##		3	ADITYA BAJAJ		6.0	L	8		61		25		21
##		MI		.640	N:2	W		В		W	_	В	26
##		4 MI	PATRICK H SCHILLING 12616049 / R: 1716 ->1	744	5.5 N:2	W	23	В	28	W W	2	W B	26
##		5	HANSHI ZUO	.,	5.5	W			37		12		13
##	18	MI		.690	N:2	В		W		В		W	
##	20	6	HANSEN SONG		5.0	W	34	D	29	L	11	W	35
##		ОН		.687	N:3	W		В		W		В	
##		7 MI	GARY DEE SWATHELL 11146376 / R: 1649 ->1	.673	5.0 N:3	W	57	W B	46	W W	13	W B	11
##		8	EZEKIEL HOUGHTON	.073	5.0	W	3		32		14		9
##		MI	15142253 / R: 1641P17->1	.657P24	N:3	В		W		В		W	
##		9	STEFANO LEE		5.0	W	25	L	18	W	59	W	8
##		ON	·	.564	N:2	W		B		W		В	
##		10 MI	ANVIT RAO 14150362 / R: 1365 ->1	.544	5.0 N:3	D W	16	L W	19	W B	55	W B	31
##		11	CAMERON WILLIAM MC LEMAN		4.5	D	38		56		6		7
##		MI		.696	N:3	В		W		В		W	
##		12	KENNETH J TACK		4.5	W	42	W	33	D	5	W	38
##		MI		.670	N:3	W		В		W	_	В	
##		13 MI	TORRANCE HENRY JR 15082995 / R: 1666 ->1	662	4.5	W	36		27		7	D B	5
##		M1 14	BRADLEY SHAW	.662	N:3 4.5	B W	54	W	44	B W	8		1
##		MI		.618	N:3	W		В		W		W	-
##		15	ZACHARY JAMES HOUGHTON		4.5	D	19		16	W	30	L	22
##	48	MI	15619130 / R: 1220P13->1	.416P20	N:3	В		В		W		W	
##		16	MIKE NIKITIN		4.0	D	10		15	Н		W	39
##		MI 17	10295068 / R: 1604 ->1 RONALD GRZEGORCZYK	.613	N:3 4.0	B W	48	W	41		26	B ı	2
##		MI		610	N:3	W		В	41	W	20	В	
##		18	DAVID SUNDEEN		4.0	W	47		9		1		32
##	57	MI	11342094 / R: 1600 ->1	.600	N:3	В		W		В		W	
##		19	DIPANKAR ROY		4.0	D	15		10		52		28
##		MI 20		.570	N:3	W	40	В	40	W	22	В	11
##		MI	JASON ZHENG 14529060 / R: 1595 ->1	.569	4.0 N:4	L W		w B	49	W	23	w B	41
##		21	DINH DANG BUI		4.0	W			1		47		3
##	66	ON	15495066 / R: 1563P22->1	.562	N:3	В		W		В		W	
##		22	EUGENE L MCCLURE		4.0	W	64		52		28		15
##		MI	12405534 / R: 1555 ->1 ALAN BUI	.529	N:4	W	4	В	42	W	20	В	58
##	72	23 ON	15030142 / R: 1363 ->1	371	4.0	L B		W		L B			20
	74	24	MICHAEL R ALDRICH		4.0		28		47	W	43	L	25
##	75	MI	13469010 / R: 1229 ->1	.300	N:4	В		W		В		В	
	77		LOREN SCHWIEBERT			L			53				24
	78 80	MI 26	12486656 / R: 1745 ->1	.681		B W		W	10	B W		W	,
	80 81	ON	MAX ZHU 15131520 / R: 1579 ->1	.564	3.5 N:4	W B	49	W	40	w B		L W	4
	83	27	GAURAV GIDWANI		3.5		51		13				37
	84	MI	14476567 / R: 1552 ->1	.539	N:4	W		В		W		В	
	86	28	SOFIA ADINA STANESCU-BEL		3.5		24			W			19
	87	MI	14882954 / R: 1507 ->1	.513	N:3	W		M		В		W	3.4
	89 90	29 MI	CHIEDOZIE OKORIE 15323285 / R: 1602P6 ->1	508P12	3.5 N:4	W B	50	D W	6	L B		L W	34
	90 92	30	GEORGE AVERY JONES	.500712			52		64				55
	93	ON	12577178 / R: 1522 ->1	444	-	W		В		В		W	
	95	31	RISHI SHETTY		3.5	L	58	D	55	W	64	L	10
	96	MI	15131618 / R: 1494 ->1		2 -	В		W		В		W	
	98 99	32 ON	JOSHUA PHILIP MATHEWS 14073750 / R: 1441 ->1		3.5 N:4	W	61	L B		W W		L B	18
	101	33	JADE GE	در			60						36
	102	MI	14691842 / R: 1449 ->1	.421		В		W		В		W	
	104	34	MICHAEL JEFFERY THOMAS		3.5			W	60	L	37	W	29
	105	MI	15051807 / R: 1399 ->1	400	2 -	B		W		В		B	_
	107 108	35 MI	JOSHUA DAVID LEE 14601397 / R: 1438 ->1	392	3.5	L W	46	L W	38	W B		L W	6
	110	36 36	14601397 / K: 1438 ->1 SIDDHARTH JHA		3.5		13		57				33
	111	MI	14773163 / R: 1355 ->1	.367	N:4	W		В		W		В	
##	113	37	AMIYATOSH PWNANANDAM		3.5	В		L	5	W	34	L	27
	114	MI	15489571 / R: 980P12->1	.077P17	2.0	_		В		W		W	
	116 117	38 MI	BRIAN LIU 15108523 / R: 1423 ->1	439		D W	11	W B		W W		L W	12
	119	39	JOEL R HENDON	. ,,,,		L			54				16
	120	MI	12923035 / R: 1436P23->1	413	N:4					В		W	-
			/MCDC/DATACO7/M/ Is										

```
## 122
         40
              FOREST ZHANG
                                             3.0 W 20 L 26 L 39 W 59
## 123
         ΜI
              14892710 / R: 1348 ->1346
                                                         В
              KYLE WILLIAM MURPHY
## 125
         41
                                              3.0
                                                   W
                                                      59 L
                                                           17 W 58 L 20
## 126
              15761443 / R: 1403P5 ->1341P9
## 128
         42
              JARED GE
                                             3.0
                                                   L 12 L
                                                            50 L 57 D 60
## 129
         ΜI
              14462326 / R: 1332
                                 ->1256
                                                              В
## 131
              ROBERT GLEN VASEY
                                             3.0
                                                     21 L 23 L 24 W
         43
## 132
         ΜI
              14101068 / R: 1283
                                 ->1244
                                                         В
## 134
              JUSTIN D SCHILLING
                                             3.0
                                                         L
                                                           14 L 32 W
              15323504 / R: 1199
## 135
         MΙ
                                 ->1199
                                                              В
## 137
         45
              DEREK YAN
                                                       5 L
                                                           51 D 60 L
## 138
         ΜI
              15372807 / R: 1242 ->1191
                                                         В
## 140
         46
              JACOB ALEXANDER LAVALLEY
                                              3.0
                                                   W 35 L
                                                             7 L 27 L
              15490981 / R: 377P3 ->1076P10
## 141
         ΜI
## 143
         47
              ERIC WRIGHT
                                             2.5
                                                   L 18 W 24 L 21 W 61
## 144
         ΜI
              12533115 / R: 1362 ->1341
              DANIEL KHAIN
## 146
                                                   L 17 W
         48
                                                           63 H
                                                                    D
## 147
         ΜI
              14369165 / R: 1382 ->1335
## 149
         49
              MICHAEL J MARTIN
                                             2.5
                                                   L 26 L
                                                            20 D
                                                                63 D
## 150
         MΙ
              12531685 / R: 1291P12->1259P17
                                                              В
## 152
              SHIVAM JHA
                                                      29 W
                                                           42 L
                                                                33 W
              14773178 / R: 1056 ->1111
## 153
         ΜI
                                                         В
## 155
         51
              TEJAS AYYAGARI
                                             2.5
                                                   L 27 W
                                                           45 L 36 W
## 156
              15205474 / R: 1011
                                  ->1097
         ΜI
## 158
         52
              ETHAN GUO
                                             2.5
                                                   W 30 D 22 L 19 D
## 159
              14918803 / R: 935
                                  ->1092
         ΜI
                                             N:4
                                                   В
                                                         W
                                                              В
## 161
              JOSE C YBARRA
                                                   Н
                                                           25 H
         53
                                             2.0
                                                         L
                                                                    L
## 162
         ΜI
              12578849 / R: 1393
                                  ->1359
## 164
              LARRY HODGE
                                             2.0
                                                   L 14 L 39 L 61 B
         54
## 165
         ΜI
              12836773 / R: 1270
                                  ->1200
                                                         В
## 167
         55
              ALEX KONG
                                                      62 D
                                                           31 L
                                                                10 L
              15412571 / R: 1186
## 168
         ΜI
                                  ->1163
                                                         В
                                                              W
## 170
         56
              MARISA RICCI
                                              2.0
                                                         L
                                                            11 L 35 W
              14679887 / R: 1153
## 171
                                  ->1140
         ΜI
                                                         В
## 173
         57
              MICHAEL LU
                                             2.0
                                                      7 L
                                                           36 W 42 L
                                                                       51
              15113330 / R: 1092
## 174
         ΜI
                                  ->1079
## 176
              VIRAJ MOHILE
                                             2.0
                                                             2 L 41 L
         58
                                                   W
                                                     31 L
## 177
         ΜI
              14700365 / R: 917
                                  -> 941
              SEAN M MC CORMICK
## 179
         59
                                             2.0
                                                   L 41 B
                                                              L
                                                                  9 L
## 180
         ΜI
              12841036 / R: 853
                                  -> 878
                                                              В
## 182
         60
              JULIA SHEN
                                                   L
                                                     33 L
                                                           34 D 45 D
              14579262 / R: 967 -> 984
## 183
         MΙ
                                                         В
                                                              В
## 185
         61
              JEZZEL FARKAS
                                                   L 32 L
                                                             3 W
                                                                54 L
## 186
              15771592 / R: 955P11-> 979P18
         ON
                                                   В
                                                        W
                                                              В
                                                                    W
## 188
         62
              ASHWIN BALAJI
                                             1.0
                                                   W 55 U
                                                              U
                                                                    U
## 189
         ΜI
              15219542 / R: 1530 ->1535
              THOMAS JOSEPH HOSMER
## 191
         63
                                             1.0
                                                   L
                                                       2 L 48 D 49 L 43
## 192
              15057092 / R: 1175 ->1125
## 194
         64
              BEN LI
                                             1.0
                                                   L 22 D 30 L 31 D 49
## 195
         ΜI
              15006561 / R: 1163 ->1112
                                                   В
                                                         W
                                                              W
         ٧8
                   V10
##
               V9
## 2
      Round Round Round
##
  3
        5
              6
          7 D 12 D
## 5
## 6
## 8
      W
         16 W
              20 W
## 9
## 11
         11 W
              13 W
## 12 W
            В
## 14
      D
          5 W
               19 D
## 15 W
            В
## 17 D
          4 W
              14 W
## 18
      В
## 20 D
         10 W
              27 W
## 21 B
## 23
      L
          1 W
               9 L
## 24
      В
## 26
              28 W
## 27 B
##
  29
                7 W
         26 L
## 30 W
            В
                 В
## 32 D
          6 W
              25 W
## 33 W
            В
## 35 L
          3 W
              34 W
## 36 B
## 38 H
            D
               1 L
## 39
## 41 W
        33 L
               3 W
                    32
## 42 W
            W
## 44 D 27 L
               5 W 31
## 45
```

```
## 47 W 54 W 33 W 38
## 48 B
         В
## 50 L 2 W 36 U
## 51 W
## 53 W 23 W 22 L 5
## 54 W
         В
## 56 L 19 W 38 L 10
## 57 B
## 59 W 18 L 4 L
## 60 W
              В
## 62 W 28 L 2 L
## 63 W
         В
## 65 W 40 W 39 L
## 66 W
## 68 H L 17 W 40
## 69
          W
## 71 L 17 W 37 W 46
## 72 B
         W
## 74 W 60 W 44 W 39
## 75 W
        W
            В
## 77 D 34 L 10 W 47
## 78 B
              В
         W
## 80 L
        9 D 32 L 11
## 81 B
## 83 D 14 L 6 U
## 84 W
## 86 L 20 L 8 D 36
## 87 B
          В
## 89 W 52 W 48 U
## 90 W
         В
## 92 L 31 W 61 W 50
## 93 W
        В
            В
## 95 W 30 W 50 L 14
## 96 B
## 98 W 51 D 26 L 13
## 99 W
## 101 L 13 L 15 W 51
## 102 B
## 104 D 25 L 11 W 52
## 105 W
        В
## 107 W 57 D 52 W 48
         В
## 108 B
## 110 H
          L 16 D 28
## 111
             В
## 113 H
        L 23 W 61
## 114
          В
        L 18 L 15
## 116 H
## 117
## 119 W 44 L 21 L 24
## 120 B
         W
## 122 L 21 W 56 L 22
## 123 B
             W
         W
## 125 X
          U
               U
## 126
## 128 D 61 W 64 W 56
## 129 W
         W
## 131 W 59 L 46 W 55
## 132 B
         В
## 134 L 39 L 24 W 59
## 135 W
         В
## 137 W 63 D 55 W 58
## 138 W
          В
## 140 W 64 W 43 L 23
## 141 B
        W
## 143 L
        8 D 51 L 25
## 144 W
         B W
         L 29 L 35
## 146 H
## 147
## 149 W 58 H U
## 150 B
## 152 H
         L 31 L 30
## 153
          В
## 155 L 32 D 47 L 33
## 156 B
        W W
## 158 L 29 D 35 L 34
## 159 B
         W
            В
## 161 U
         W 57 U
## 162
## 164 L 15 L 59 W 64
## 165 W
         B W
## 167 B
          D 45 L 43
```

```
## 168
                 В
## 170 H
           L 40 L 42
## 171
           В
## 173 L 35 L 53 B
## 174 W
          В
## 176 L 49 B
                 L 45
## 177 W
## 179 L 43 W 54 L 44
## 180 W
## 182 L 24 H
## 183 B
## 185 D 42 L 30 L 37
## 186 B
          W
## 188 U
## 189
## 191 L 45 H
## 192 B
## 194 L 46 L 42 L 54
## 195 W
          В
```

Then I am going to extract the information I need and create vectors with appropriate data types.

```
player_number <- as.numeric(raw_data[seq(5,195,3), "V1"])
head(player_number)</pre>
```

```
## [1] 1 2 3 4 5 6
```

```
player_name <- raw_data[seq(5,195,3),"V2"]
player_name <- str_trim(player_name)
head(player_name)</pre>
```

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

```
player_state <- raw_data[seq(6,195,3), "V1"]
player_state <- str_trim(player_state)
head(player_state)</pre>
```

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

```
total_number_of_points <- as.numeric(raw_data[seq(5,195,3), "V3"])
head(total_number_of_points)</pre>
```

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

I used two steps of regular expression techniques to draw the pre-rating data from the raw data

```
pre_rating <- raw_data[seq(6,195,3), "V2"]
head(pre_rating)</pre>
```

```
for (i in 1:length(pre_rating))
{
    #I want any string that starts with "R" followed by a space and followed by more than 1 digits.
    pre_rating[i] <- str_extract(pre_rating[i], "R:[:space:]+[:digit:]+")

#After I got the string containing pre-rating, I get rid of the "R" and space, then convert these digits from string i
nto numeric values.
    pre_rating[i] <- str_extract(pre_rating[i], "[[:digit:]]+")
}
head(pre_rating)</pre>
```

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

The last variable I want is the average of opponents' rating for each individual players. The first step invloves getting those opponents' player numbers, which can be found on columns 4 to 10. I only need to extract strings that refer to numeric values from these columns. After I got those strings, I convert them into numeric values and store them into a variable called opponent_number. Then I found their corresponding pre-rating values, use the mean function to get the average of these opponents' pre-rating values.

```
## [1] 1605.286 1469.286 1563.571 1573.571 1500.857 1518.714 1372.143
## [8] 1468.429 1523.143 1554.143 1467.571 1506.167 1497.857 1515.000
## [15] 1483.857 1385.800 1498.571 1480.000 1426.286 1410.857 1470.429
## [22] 1300.333 1213.857 1357.000 1363.286 1506.857 1221.667 1522.143
## [29] 1313.500 1144.143 1259.857 1378.714 1276.857 1375.286 1149.714
## [36] 1388.167 1384.800 1539.167 1429.571 1390.571 1248.500 1149.857
## [43] 1106.571 1327.000 1152.000 1357.714 1392.000 1355.800 1285.800
## [50] 1296.000 1356.143 1494.571 1345.333 1206.167 1406.000 1414.400
## [57] 1363.000 1391.000 1319.000 1330.200 1327.286 1186.000 1350.200
## [64] 1263.000
```

Then I put all the vectors together and create a data frame called chess_tournament.

```
chess_tournament <- data.frame(player_number, player_name, player_state, total_number_of_points, pre_rating, average_opponen
t_point)
head(chess_tournament)</pre>
```

```
##
  player_number
                          {\tt player\_name~player\_state~total\_number\_of\_points}
## 1
                1
                             GARY HUA
                                                ON
## 2
                2
                      DAKSHESH DARURI
                                                ΜI
                                                                      6.0
## 3
                3
                         ADITYA BAJAJ
                                                мт
                                                                      6.0
## 4
                4 PATRICK H SCHILLING
                                                ΜI
                                                                      5.5
## 5
                5
                          HANSHI ZUO
                                                                      5.5
                                                MΙ
## 6
                6
                          HANSEN SONG
                                                                       5.0
## pre_rating average_opponent_point
## 1
          1794
                             1605.286
## 2
          1553
                             1469.286
## 3
          1384
                             1563.571
## 4
          1716
                             1573.571
## 5
          1655
                             1500.857
## 6
          1686
                             1518.714
```

In the end, I export the output into a csv file in my working directory.

```
write.csv(chess_tournament, file = "chess_tournament_result", row.names=FALSE)
```