# Data 607: Project 1

Mohammad Zahid Chowdhury

2025-02-23

## # Load necessary libraries:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.4.2
```

```
library(purrr)
library(tidyr)
```

## Read in the raw data & clean up the data with symbols in the headers:

```
rawdata <- readLines("C:/Users/zahid/OneDrive/Desktop/Data Science/DATA 607 - Project/zc1.txt", warn = F

df1 <- rawdata %>%
  str_replace_all("->", ">>") %>%  ## removes arrows
  str_replace_all("-{3,}", "") %>%  # removes "-" repeated 3 or more times
  discard(~ . == "") %>%  # removes empty strings
  .[-(1:2)]  # removes the first two elements of the character vector
```

## I organized the information using the players performance and players background information

## determined the format by looking at the contents in the record

```
data.format1 <- keep(df1, ~ str_detect(str_sub(.x, 1, 6), "[0-9]"))
head(data.format1,5)
```

```
## [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|"
```

```
data.format2 <- keep(df1, ~ str_detect(str_sub(.x, 1, 6), "[A-Z]{2}"))
head(data.format2,5)
```

```
## [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    |"
```

## Define the players performance numericla pattern:

```
extract_fields <- function(data, start, end, func = identity) {
  func(str_sub(data, start, end))
}
```

## Apply data processing and extraction for data.format1 (numeric player data):

```
players_performance <- data.frame(
  player_num  = map_dbl(data.format1, extract_fields, 1, 6, as.numeric),
  player_name = map_chr(data.format1, extract_fields, 8, 40, str_trim),
  total_pts   = map_dbl(data.format1, extract_fields, 42, 46, as.numeric),
  round1      = map_chr(data.format1, extract_fields, 48, 52),
  round2      = map_chr(data.format1, extract_fields, 54, 58),
  round3      = map_chr(data.format1, extract_fields, 60, 64),
  round4      = map_chr(data.format1, extract_fields, 66, 70),
  round5      = map_chr(data.format1, extract_fields, 72, 76),
  round6      = map_chr(data.format1, extract_fields, 78, 82),
  round7      = map_chr(data.format1, extract_fields, 84, 88),
  stringsAsFactors = FALSE
```

```
)

# preview the output
head(players_performance, 3)
```

```
##   player_num     player_name total_pts round1 round2 round3 round4 round5
## 1          1        GARY HUA         6  W  39  W  21  W  18  W  14  W   7
## 2          2 DAKSHESH DARURI         6  W  63  W  58  L   4  W  17  W  16
## 3          3    ADITYA BAJAJ         6  L   8  W  61  W  25  W  21  W  11
##   round6 round7
## 1  D  12  D   4
## 2  W  20  W   7
## 3  W  13  W  12
```

## Determine the pattern for players info:

```
players_info <- data.format2 %>%
  tibble::enframe(name = NULL, value = "data.format2") %>%
  mutate(
    player_state = sub("^\\s+|\\s+$", "", substr(data.format2, 1, 6)),  # extract first "word" (player
    uscf_id = str_extract(data.format2, "\\d+"),  # extract first numeric value (USCF ID)
    pre_rating = as.numeric(str_extract(data.format2, "(?<=R: )\\d+")),  # extract number after "R: "
    post_rating = as.numeric(str_extract(data.format2, "(?<=>>)\\d+"))  # extract number after ">>"
  ) %>%
  select(-data.format2)


# preview the output
head(players_info, 3)
```

```
## # A tibble: 3 x 4
##   player_state uscf_id  pre_rating post_rating
##   <chr>        <chr>         <dbl>       <dbl>
## 1 "ON "        15445895       1794        1817
## 2 "MI "        14598900       1553        1663
## 3 "MI "        14959604       1384        1640
```

## Obtain all the data:

```
allresults <- cbind(players_performance, players_info)
```

## determining the average pre-chess rating per opponent:

```
head(allresults)
```

```
##   player_num         player_name total_pts round1 round2 round3 round4 round5
## 1          1            GARY HUA       6.0  W  39  W  21  W  18  W  14  W   7
## 2          2     DAKSHESH DARURI       6.0  W  63  W  58  L   4  W  17  W  16
## 3          3        ADITYA BAJAJ       6.0  L   8  W  61  W  25  W  21  W  11
## 4          4 PATRICK H SCHILLING       5.5  W  23  D  28  W   2  W  26  D   5
## 5          5          HANSHI ZUO       5.5  W  45  W  37  D  12  D  13  D   4
## 6          6         HANSEN SONG       5.0  W  34  D  29  L  11  W  35  D  10
##   round6 round7 player_state  uscf_id pre_rating post_rating
## 1  D  12  D   4           ON  15445895       1794        1817
## 2  W  20  W   7           MI  14598900       1553        1663
## 3  W  13  W  12           MI  14959604       1384        1640
## 4  W  19  D   1           MI  12616049       1716        1744
## 5  W  14  W  17           MI  14601533       1655        1690
## 6  W  27  W  21           OH  15055204       1686        1687
```

```
colnames(allresults)
```

```
##  [1] "player_num"   "player_name"  "total_pts"    "round1"       "round2"
##  [6] "round3"       "round4"       "round5"       "round6"       "round7"
## [11] "player_state" "uscf_id"      "pre_rating"   "post_rating"
```

```
average_calc <- allresults %>%
  select(player_num, starts_with("round")) %>%
  pivot_longer(cols = starts_with("round"), names_to = "round", values_to = "outcome_opp") %>%
  mutate(round = as.numeric(str_replace(round, "round", "")),
         outcome = str_extract(outcome_opp, "^\\w+"),
         opponent_num = as.numeric(str_extract(outcome_opp, "\\d+$"))) %>%
  select(player_num, round, outcome, opponent_num) %>%
  left_join(select(allresults, player_num, total_pts, player_state, post_rating), by = "player_num") %>%
  left_join(select(allresults, player_num, pre_rating), by = c("opponent_num" = "player_num")) %>%
  arrange(player_num, round) %>%
  rename(opponent_pre_rating = pre_rating)


write.csv(allresults, "allresults.csv", row.names=FALSE)
write.csv(average_calc, "average_calc.csv", row.names=FALSE)
```