DATA607 PROJECT 1

Farhod Ibragimov

2025-02-23

Loading libraries:

```
library(tidyverse)
library(readr)
library(data.table)
library(kableExtra)
```

This code cell loading `tournamentinfo.txt` from GitHub URL and stores it into `data`

```
URL <- "https://raw.githubusercontent.com/farhodibr/CUNY-SPS-MSDS/main/DATA607/LAB5/tournamentinfo.txt"
data <- readLines(URL)
```

```
## Warning in readLines(URL): incomplete final line found on
## 'https://raw.githubusercontent.com/farhodibr/CUNY-SPS-MSDS/main/DATA607/LAB5/tournamentinfo.txt'
```

```
head(data)
```

```
## [1] "-----------------------------------------------------------------------------------------"
## [2] " Pair | Player Name                     |Total|Round|Round|Round|Round|Round|Round|Round| "
## [3] " Num  | USCF ID / Rtg (Pre->Post)       | Pts |  1  |  2  |  3  |  4  |  5  |  6  |  7  | "
## [4] "-----------------------------------------------------------------------------------------"
## [5] "    1 | GARY HUA                        |6.0  |W  39|W  21|W  18|W  14|W   7|D  12|D   4|"
## [6] "   ON | 15445895 / R: 1794   ->1817     |N:2  |W    |B    |W    |B    |W    |B    |W    |"
```

This code cell creates a messy table from `data` and stores it into `elo_table` by doing:

1. `merging_lines`function iterates through each line and merges two lines into one. Function grabs first line and appends next line to it, then goes to next line and does same merging as in previous step. The reason i use this because each player's obseravtion is in two lines all over data set.
2. replacing all `"|"` with `","`

```
data[length(data)] <- paste0(data[length(data)], "\n")
data <- data[!grepl("^[-]+$", data)]
head(data)
```

```
## [1] " Pair | Player Name                     |Total|Round|Round|Round|Round|Round|Round|Round| "
## [2] " Num  | USCF ID / Rtg (Pre->Post)       | Pts |  1  |  2  |  3  |  4  |  5  |  6  |  7  | "
## [3] "    1 | GARY HUA                        |6.0  |W  39|W  21|W  18|W  14|W   7|D  12|D   4|"
## [4] "   ON | 15445895 / R: 1794   ->1817     |N:2  |W    |B    |W    |B    |W    |B    |W    |"
## [5] "    2 | DAKSHESH DARURI                 |6.0  |W  63|W  58|L   4|W  17|W  16|W  20|W   7|"
## [6] "   MI | 14598900 / R: 1553   ->1663     |N:2  |B    |W    |B    |W    |B    |W    |B    |"
```

```
merging_lines <- function(data) {
  merged_data <- character(0)
    for (i in seq(1, length(data) - 1, by = 2)) {
        merged_line <- paste0(data[i], "|", data[i + 1])
        merged_data <- c(merged_data, merged_line)
```

```
  }
  return(merged_data)
}
merged_data <- merging_lines(data)
merged_data <- gsub("|", ",", merged_data, fixed = TRUE)
head(merged_data)
```

```
## [1] " Pair , Player Name                    ,Total,Round,Round,Round,Round,Round,Round,Round, , Num
## [2] "    1 , GARY HUA                       ,6.0  ,W  39,W  21,W  18,W  14,W   7,D  12,D   4,,   ON
## [3] "    2 , DAKSHESH DARURI                ,6.0  ,W  63,W  58,L   4,W  17,W  16,W  20,W   7,,   MI
## [4] "    3 , ADITYA BAJAJ                   ,6.0  ,L   8,W  61,W  25,W  21,W  11,W  13,W  12,,   MI
## [5] "    4 , PATRICK H SCHILLING            ,5.5  ,W  23,D  28,W   2,W  26,D   5,W  19,D   1,,   MI
## [6] "    5 , HANSHI ZUO                     ,5.5  ,W  45,W  37,D  12,D  13,D   4,W  14,W  17,,   MI
```

This cell does cleaning and formatting for column names. It creates `elo_table` data frame from `merged_data`.

```
column_names <- strsplit(merged_data[1], split = ",", fixed = TRUE)
column_names <- trimws(column_names)
column_names <- gsub("\\s+", " ", column_names)


print(column_names, quote = FALSE)
```

```
## [1] c(" Pair ", " Player Name ", "Total", "Round", "Round", "Round", "Round", "Round", "Round", "Rou
```

```
print(typeof(column_names))
```

```
## [1] "character"
```

```
#print(length(column_names))
merged_data <- merged_data[-1]


column_names <- c(" Pair ", " Player Name ", "Total", "Round 1", "Round 2", "Round 3", "Round 4", "Roun
print(length(column_names))
```

```
## [1] 22
```

```
#merged_data <- merged_data[-1]
elo_table <- data.frame(merged_data)

elo_table <- elo_table |>
  separate(col = 1,
           into = column_names,
           sep = ",",
           extra = "merge")


head(elo_table)
```

```
##    Pair                      Player Name  Total Round 1 Round 2 Round 3
## 1     1   GARY HUA                         6.0     W  39   W  21   W  18
## 2     2   DAKSHESH DARURI                  6.0     W  63   W  58   L   4
## 3     3   ADITYA BAJAJ                     6.0     L   8   W  61   W  25
## 4     4   PATRICK H SCHILLING              5.5     W  23   D  28   W   2
## 5     5   HANSHI ZUO                       5.5     W  45   W  37   D  12
## 6     6   HANSEN SONG                      5.0     W  34   D  29   L  11
```

```
##   Round 4 Round 5 Round 6 Round 7    State            USCFID  Rtg  Pre  Post
## 1   W  14   W   7   D  12   D   4       ON   15445895 / R: 1794   ->1817
## 2   W  17   W  16   W  20   W   7       MI   14598900 / R: 1553   ->1663
## 3   W  21   W  11   W  13   W  12       MI   14959604 / R: 1384   ->1640
## 4   W  26   D   5   W  19   D   1       MI   12616049 / R: 1716   ->1744
## 5   D  13   D   4   W  14   W  17       MI   14601533 / R: 1655   ->1690
## 6   W  35   D  10   W  27   W  21       OH   15055204 / R: 1686   ->1687
##    Pts     1     2     3     4     5     6     7   __
## 1 N:2   W     B     W     B     W     B     W
## 2 N:2   B     W     B     W     B     W     B
## 3 N:2   W     B     W     B     W     B     W
## 4 N:2   W     B     W     B     W     B     B
## 5 N:2   B     W     B     W     B     W     B
## 6 N:3   W     B     W     B     B     W     B
```

```
#elo_table$pair <- as.numeric(unlist(elo_table$pair))
elo_table <- Filter(function(x)!(all(x == "")), elo_table)
elo_table <- elo_table |>
  mutate(across(everything(), ~trimws(.x))) |>
  mutate(across(everything(), ~gsub("\\s+", " ", .x))) |>
  mutate(across(everything(), ~gsub(" +", " ", .x))) |>
  mutate(across(everything(), ~gsub(" +$", "", .x))) |>
  mutate(across(everything(), ~gsub("^ +", "", .x)))

glimpse(elo_table)
```

```
## Rows: 64
## Columns: 20
## $ ` Pair `                  <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9"~
## $ ` Player Name `           <chr> "GARY HUA", "DAKSHESH DARURI", "ADITYA BAJA~
## $ Total                     <chr> "6.0", "6.0", "6.0", "5.5", "5.5", "5.0", "~
## $ `Round 1`                 <chr> "W 39", "W 63", "L 8", "W 23", "W 45", "W 3~
## $ `Round 2`                 <chr> "W 21", "W 58", "W 61", "D 28", "W 37", "D ~
## $ `Round 3`                 <chr> "W 18", "L 4", "W 25", "W 2", "D 12", "L 11~
## $ `Round 4`                 <chr> "W 14", "W 17", "W 21", "W 26", "D 13", "W ~
## $ `Round 5`                 <chr> "W 7", "W 16", "W 11", "D 5", "D 4", "D 10"~
## $ `Round 6`                 <chr> "D 12", "W 20", "W 13", "W 19", "W 14", "W ~
## $ `Round 7`                 <chr> "D 4", "W 7", "W 12", "D 1", "W 17", "W 21"~
## $ ` State `                 <chr> "ON", "MI", "MI", "MI", "MI", "OH", "MI", "~
## $ ` USCFID  Rtg  Pre  Post ` <chr> "15445895 / R: 1794 ->1817", "14598900 / R:~
## $ ` Pts `                   <chr> "N:2", "N:2", "N:2", "N:2", "N:2", "N:3", "~
## $ ` 1 `                     <chr> "W", "B", "W", "W", "B", "W", "W", "B", "W"~
## $ ` 2 `                     <chr> "B", "W", "B", "B", "W", "B", "B", "W", "B"~
## $ ` 3 `                     <chr> "W", "B", "W", "W", "B", "W", "W", "B", "W"~
## $ ` 4 `                     <chr> "B", "W", "B", "B", "W", "B", "B", "W", "B"~
## $ ` 5 `                     <chr> "W", "B", "W", "W", "B", "B", "B", "B", "W"~
## $ ` 6 `                     <chr> "B", "W", "B", "B", "W", "W", "W", "W", "B"~
## $ ` 7 `                     <chr> "W", "B", "W", "B", "B", "B", "W", "W", "B"~
```

```
head(elo_table)
```

```
##    Pair         Player Name  Total Round 1 Round 2 Round 3 Round 4 Round 5
## 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
##   Round 6 Round 7  State    USCFID  Rtg  Pre  Post   Pts   1   2   3   4   5
## 1    D 12     D 4      ON 15445895 / R: 1794 ->1817   N:2   W   B   W   B   W
## 2    W 20     W 7      MI 14598900 / R: 1553 ->1663   N:2   B   W   B   W   B
## 3    W 13    W 12      MI 14959604 / R: 1384 ->1640   N:2   W   B   W   B   W
## 4    W 19     D 1      MI 12616049 / R: 1716 ->1744   N:2   W   B   W   B   W
## 5    W 14    W 17      MI 14601533 / R: 1655 ->1690   N:2   B   W   B   W   B
## 6    W 27    W 21      OH 15055204 / R: 1686 ->1687   N:3   W   B   W   B   B
##    6   7
## 1   B   W
## 2   W   B
## 3   B   W
## 4   B   B
## 5   W   B
## 6   W   B
```

This cell cleans and format elo_table variable names, splits uscfid___rtg___pre___post into separate variables.

```
colnames(elo_table) <- tolower(colnames(elo_table))
colnames(elo_table) <- trimws(colnames(elo_table))
colnames(elo_table) <- str_replace_all(colnames(elo_table), " ", "_")
print(colnames(elo_table))
```

```
##  [1] "pair"                  "player_name"           "total"
##  [4] "round_1"               "round_2"               "round_3"
##  [7] "round_4"               "round_5"               "round_6"
## [10] "round_7"               "state"                 "uscfid__rtg__pre__post"
## [13] "pts"                   "1"                     "2"
## [16] "3"                     "4"                     "5"
## [19] "6"                     "7"
```

```
elo_table$pair <- elo_table$pair |>
  as.numeric(unlist(elo_table$pair))


head(elo_table)
```

```
##   pair         player_name total round_1 round_2 round_3 round_4 round_5
## 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
##   round_6 round_7 state    uscfid__rtg__pre__post pts 1 2 3 4 5 6 7
## 1    D 12     D 4    ON 15445895 / R: 1794 ->1817 N:2 W B W B W B W
## 2    W 20     W 7    MI 14598900 / R: 1553 ->1663 N:2 B W B W B W B
## 3    W 13    W 12    MI 14959604 / R: 1384 ->1640 N:2 W B W B W B W
## 4    W 19     D 1    MI 12616049 / R: 1716 ->1744 N:2 W B W B W B B
## 5    W 14    W 17    MI 14601533 / R: 1655 ->1690 N:2 B W B W B W B
## 6    W 27    W 21    OH 15055204 / R: 1686 ->1687 N:3 W B W B B W B
```

In this code cell I'm creating tables for players info an players ratings. Players rating table created by splitting uscfid___rtg___pre___post column and its values into separate variables.

```
players_info_table <- elo_table |>
  select(pair, player_name, state, total)



players_rating_table <- elo_table |>
  select(pair, uscfid__rtg__pre__post) |>
  extract(
    col = uscfid__rtg__pre__post,
    into = c("uscf_id", "rating", "new_rating"),
    regex = "(\\d*)\\s*/\\s*R:\\s*(\\d*P?\\d*)\\s*->\\s*(\\d*P?\\d*)"
  )|>
  mutate(
    rating = gsub("P.*", "", rating),
    new_rating = gsub("P.*", "", new_rating),
    across(c("rating", "new_rating"), as.numeric)
  )
#|>
 # mutate(across(c("uscf_id", "rating", "new_rating"), as.numeric))
```

Here I'm creating new rounds and rounds results tables. Rounds results table extracts only characters from rounds table values, such as "W" -won the game, "L" - lost, "D" - draw, and with the rest of characters for not played games.

```
players_rating_table$pair <- elo_table$pair
glimpse(players_rating_table)
```

```
## Rows: 64
## Columns: 4
## $ pair       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ uscf_id    <chr> "15445895", "14598900", "14959604", "12616049", "14601533",~
## $ rating     <dbl> 1794, 1553, 1384, 1716, 1655, 1686, 1649, 1641, 1411, 1365,~
## $ new_rating <dbl> 1817, 1663, 1640, 1744, 1690, 1687, 1673, 1657, 1564, 1544,~
```

```
head(elo_table$player_name[5])
```

```
## [1] "HANSHI ZUO"
```

```
new_row <- data.frame(pair = 0, uscf_id = "0", rating = 0, new_rating = 0, round_1 = 1,
                      round_2 = 1, round_3 = 1,
                      round_4 = 1, round_5 = 1,
                      round_6 = 1, round_7 = 1)



rounds_table <- elo_table |>
  select(pair, starts_with("round")) |>
  mutate(across(starts_with("round"), ~as.numeric(gsub("[^0-9]", "",.))))

#print(rounds_table)
#print(elo_table)

round_results_table <- elo_table |>
  select(pair, starts_with("round")) |>
  mutate(across(starts_with("round"), ~gsub("[^A-Z]", "",.x)))
```

```
rounds_table <- bind_rows(rounds_table, new_row[1])
```

This cell creates opponents_average_rating_table:

1. joining with players_rating_table
2. creates average variable for average ratings of played games opponent players.(this looks little bit complicated. I'm learning to get same results in less amount of code in future)
3. creates final_ratings_table with pair#, players name, player's state, total game points, player's rating, average rating of played opponents.

```
opponents_average_rating_table <- left_join(players_rating_table,
                                            rounds_table,
                                            by = "pair")
opponents_average_rating_table <- rbind(opponents_average_rating_table, new_row)



opponents_average_rating_table <- opponents_average_rating_table |>
  rowwise() |>
  mutate(average = round(mean(
           c(
           opponents_average_rating_table$rating[
             opponents_average_rating_table$round_1[pair]
             ],
           opponents_average_rating_table$rating[
             opponents_average_rating_table$round_2[pair]
             ],
           opponents_average_rating_table$rating[
             opponents_average_rating_table$round_3[pair]
             ],
           opponents_average_rating_table$rating[
             opponents_average_rating_table$round_4[pair]
             ],
           opponents_average_rating_table$rating[
             opponents_average_rating_table$round_5[pair]
             ],
           opponents_average_rating_table$rating[
             opponents_average_rating_table$round_6[pair]
             ],
           opponents_average_rating_table$rating[
             opponents_average_rating_table$round_7[pair]
             ]

                     ),
           na.rm = TRUE))) |>
  ungroup()

glimpse(opponents_average_rating_table)
```

```
## Rows: 65
## Columns: 12
## $ pair       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ uscf_id    <chr> "15445895", "14598900", "14959604", "12616049", "14601533",~
```

```
## $ rating     <dbl> 1794, 1553, 1384, 1716, 1655, 1686, 1649, 1641, 1411, 1365,~
## $ new_rating <dbl> 1817, 1663, 1640, 1744, 1690, 1687, 1673, 1657, 1564, 1544,~
## $ round_1    <dbl> 39, 63, 8, 23, 45, 34, 57, 3, 25, 16, 38, 42, 36, 54, 19, 1~
## $ round_2    <dbl> 21, 58, 61, 28, 37, 29, 46, 32, 18, 19, 56, 33, 27, 44, 16,~
## $ round_3    <dbl> 18, 4, 25, 2, 12, 11, 13, 14, 59, 55, 6, 5, 7, 8, 30, NA, 2~
## $ round_4    <dbl> 14, 17, 21, 26, 13, 35, 11, 9, 8, 31, 7, 38, 5, 1, 22, 39, ~
## $ round_5    <dbl> 7, 16, 11, 5, 4, 10, 1, 47, 26, 6, 3, NA, 33, 27, 54, 2, 23~
## $ round_6    <dbl> 12, 20, 13, 19, 14, 27, 9, 28, 7, 25, 34, 1, 3, 5, 33, 36, ~
## $ round_7    <dbl> 4, 7, 12, 1, 17, 21, 2, 19, 20, 18, 26, 3, 32, 31, 38, NA, ~
## $ average    <dbl> 1605, 1469, 1564, 1574, 1501, 1519, 1372, 1468, 1523, 1554,~
```

```
new_rating_table <- opponents_average_rating_table|>
  select(pair, average) |>
  slice(-n())

final_ratings_table <- players_info_table|>
  left_join(players_rating_table) |>
  left_join(new_rating_table, by = "pair")|>
  select(-new_rating, -uscf_id)
```

```
## Joining with `by = join_by(pair)`
```

```
head(final_ratings_table)
```

```
##   pair         player_name state total rating average
## 1    1            GARY HUA    ON   6.0   1794    1605
## 2    2     DAKSHESH DARURI    MI   6.0   1553    1469
## 3    3        ADITYA BAJAJ    MI   6.0   1384    1564
## 4    4 PATRICK H SCHILLING    MI   5.5   1716    1574
## 5    5          HANSHI ZUO    MI   5.5   1655    1501
## 6    6         HANSEN SONG    OH   5.0   1686    1519
```

```
final_ratings_table |>
  head(10) |>
  kable() |>
  kable_styling(full_width = F)
```

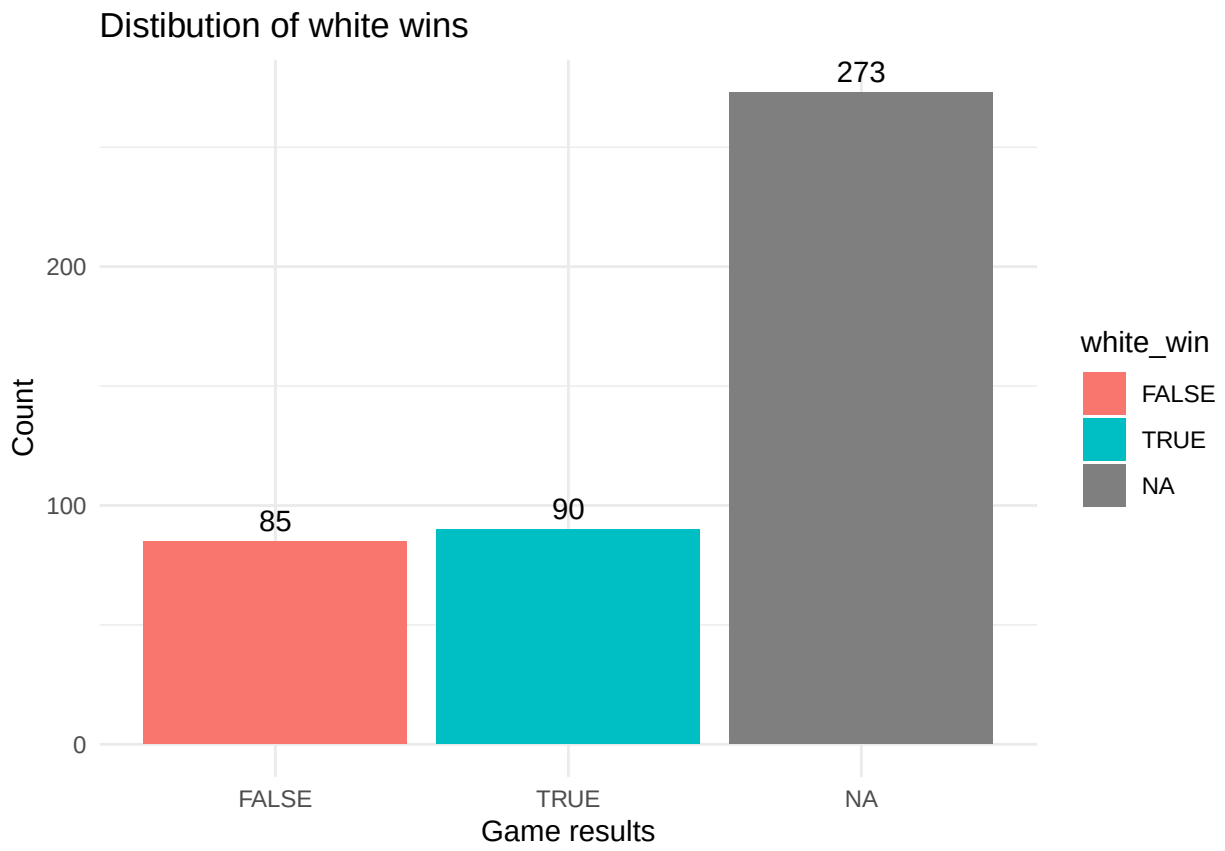
pair	player_name	state	total	rating	average
1	GARY HUA	ON	6.0	1794	1605
2	DAKSHESH DARURI	MI	6.0	1553	1469
3	ADITYA BAJAJ	MI	6.0	1384	1564
4	PATRICK H SCHILLING	MI	5.5	1716	1574
5	HANSHI ZUO	MI	5.5	1655	1501
6	HANSEN SONG	OH	5.0	1686	1519
7	GARY DEE SWATHELL	MI	5.0	1649	1372
8	EZEKIEL HOUGHTON	MI	5.0	1641	1468
9	STEFANO LEE	ON	5.0	1411	1523
10	ANVIT RAO	MI	5.0	1365	1554

This analysis will examine how the outcomes of chess games are influenced by the color of the pieces each player uses. I want to if there is any advantage or disadvantage associated with playing with white or black pieces.

```r
round_results_table_long <- round_results_table |>
  pivot_longer(
    cols = (!pair),
    names_to = "round",
    names_transform = list(round = ~gsub("[^0-9]", "", .x)),
    values_to = "result"
  ) |>
  mutate(result = trimws(result))

rounds_pcs_color_table <- elo_table |>
  select(-(player_name:pts)) |>
  pivot_longer(
    cols = (!pair),
    names_to = "round",
    values_to = "pieces_color"
  ) |>
  mutate(pieces_color = trimws(pieces_color))

wins_by_pcs_color_table <- round_results_table_long |>
  left_join(rounds_pcs_color_table, by = c("pair", "round"))

wins_by_pcs_color_table <- wins_by_pcs_color_table |>
  mutate(white_win = ifelse(result == "W" & pieces_color == "W", TRUE,
                            ifelse(result == "L" & pieces_color == "W", FALSE, NA)))

table(wins_by_pcs_color_table$white_win)
```

```
##
## FALSE  TRUE
##    85    90
```

```r
ggplot(wins_by_pcs_color_table, aes(x = white_win, fill = white_win)) +
  geom_bar(stat = "count") +
  labs(
    title = "Distibution of white wins",
    x = "Game results",
    y = "Count"
  ) +
  theme_minimal() +
  geom_text(stat = "count", aes(x = white_win,
                                label = after_stat(count)),
                                vjust = -0.5)
```



As the bar plot shows, there is very small advantage of the color of starting pieces according to results of this dataset. As a chess players myself I believed that starting with white pieces is advantage. But this analysis puts doubt in it.