Chess Analysis Project 1

Mike Birlew

2023-02-11

Introduction

Chess is a complex game of strategy with several factors involved that lead to a win. The raw data set comes from kaggle's website https://www.kaggle.com/datasets/datasnaek/chess. This data set includes over 20000 games that were collected from an online chess simulator containing user data from Lichess.org. This project will begin to look into patterns that players may use to influence there chances of winning specifically move sets, known as opening moves, that have been defined as standard opening plays consisting of a specific set of moves.

Data Cleaning

To analyze the data for move strategy comparison, the raw data has been cleaned from sixteen columns to seven columns. The winner column has been filtered of all games that are not defined as a win leaving 6,325 total games. The standardized opening moves sets (opening_eco) has been factored to contain 272 different opening moves.

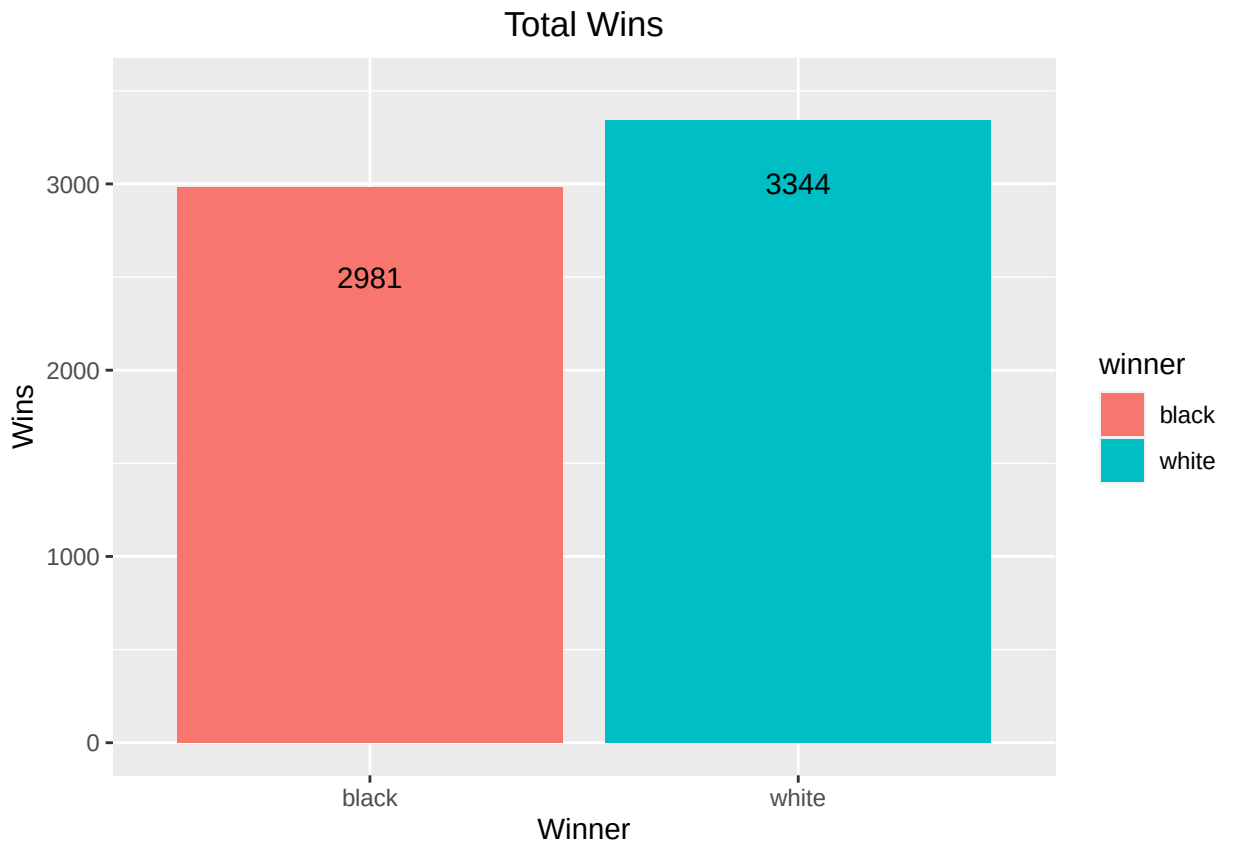
Table 1: Table 1. Chess Data Annotation

Parameter	Data_type	Discription
rated	number	Games are rated or unrated in play
turns	integer	The amount of turns for a game
winner	character	Winner of the match by color
opening_eco	factor	Opening move assigned 1-272
opening_name	character	Name of opening move used
opening_ply	integer	Number of moves in the opening phase
winner2	number	Dummy variable assigned to winner color (1=White, 0=Black)
game_type	number	Dummy variable assinged to type of game (1=Timed, 0=Un-timed)

(Table 1.) the data set being utilized contains three numeric parameters which include a 'rated' parameter that determines whether a game is added to a persons total game score, the 'game type' that is used to signify whether a game was used under Fischer rules, a typed of timed play and a 'winner2' dummy variable assigned to which color won the game. The two integer parameters are 'turns' indicating the number of total turns played in the game and an 'opening_ply' which indicates how many moves were included in each opening move set. The character variables are 'winner' displaying which color won the game and the 'opening name' which is the name of the opening moved used to begin the game. The 'opening echo' is a factored parameter for each type of moves set.

Data Analysis

When looking for advantages in play fig.1 shows that 'white' wins more often than 'black'. In the game of chess 'white' always moves first. This is a well studied phenomenon in chess and the consensus that follows is called 'First-move advantage' https://en.wikipedia.org/wiki/First-move_advantage_in_chess It generally follows that with the first move 'black' is reacting to white as the game continues. Out of 6,325 games 'white' wins 3,344 of them. This 53% advantage is consistent with other studies that have been done on 'First-move Advantage'.



(Figure.1) The total number of games in the data set (6,325) divided into wins between 'white' and 'black' shows white winning a 53% of games in what is called 'First-move Advantage'.

Outside of First-move advantage, are there any other factors that influence the game, especially any that 'black' can use to overcome 'white' always moving first and having First-move advantage? Table 2 shows a summary of the total wins and percentage of wins per each move set. There are some interesting numbers that come out of this table for 'Black Wins'. For each move set, there is a set of moves that have a larger advantage of winning the game despite 'black' winning a slightly less percentage of the games overall. 'Black Wins' shows a larger median and mean when it comes to specific move sets winning games against 'white'. There is also a particular move set that, when used, can increase 'black' wins. This is seen at the 'Black Wins' Maximum value with a total of 291 wins at 9.4% against 'White Wins' Maximum win move of only 171 at 5.1% of wins. When normalized against all games played (6,325), the Maximum wins of the top two move sets is 'black' at 4.4% and 'white' at 2.7% for a 1.7% advantage over 'white' when both players are using the most winning move set.

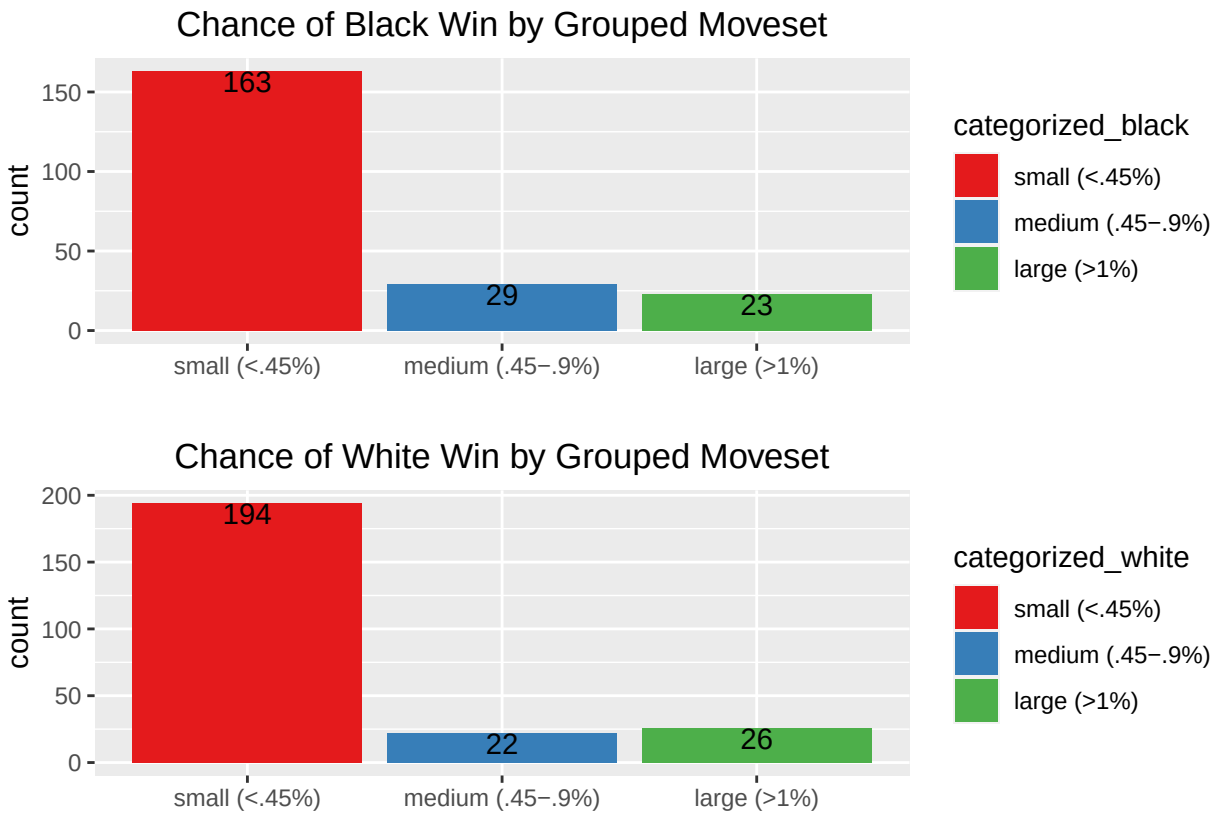
Table 2. White/Black Win & Percent to Move Sets

	Black Wins	Percent to Win		White Wins	Percent to Win
	Min. : 1.00	Min. :0.03355		Min. : 1.00	Min. :0.02990
	1st Qu.: 1.00	1st Qu.:0.03355		1st Qu.: 1.00	1st Qu.:0.02990
	Median : 4.00	Median :0.13418		Median : 3.00	Median :0.08971
	Mean : 13.87	Mean :0.46512		Mean : 13.82	Mean :0.41322
	3rd Qu.: 12.00	3rd Qu.:0.40255		3rd Qu.: 11.00	3rd Qu.:0.32895
	Max. :281.00	Max. :9.42637		Max. :171.00	Max. :5.11364

(Table 2.) Shows a comparison of total moves to wins against 'black' and 'white', particularly intersting is the opening move set that increases wins when used by 'black' to 9.4% . When normalized against 'white' wins the max win percent is still higher at 8.4% compared to 'white' max percent win at 5.1% giving 'black' a 3.3% advantage over the strongest move for 'white.

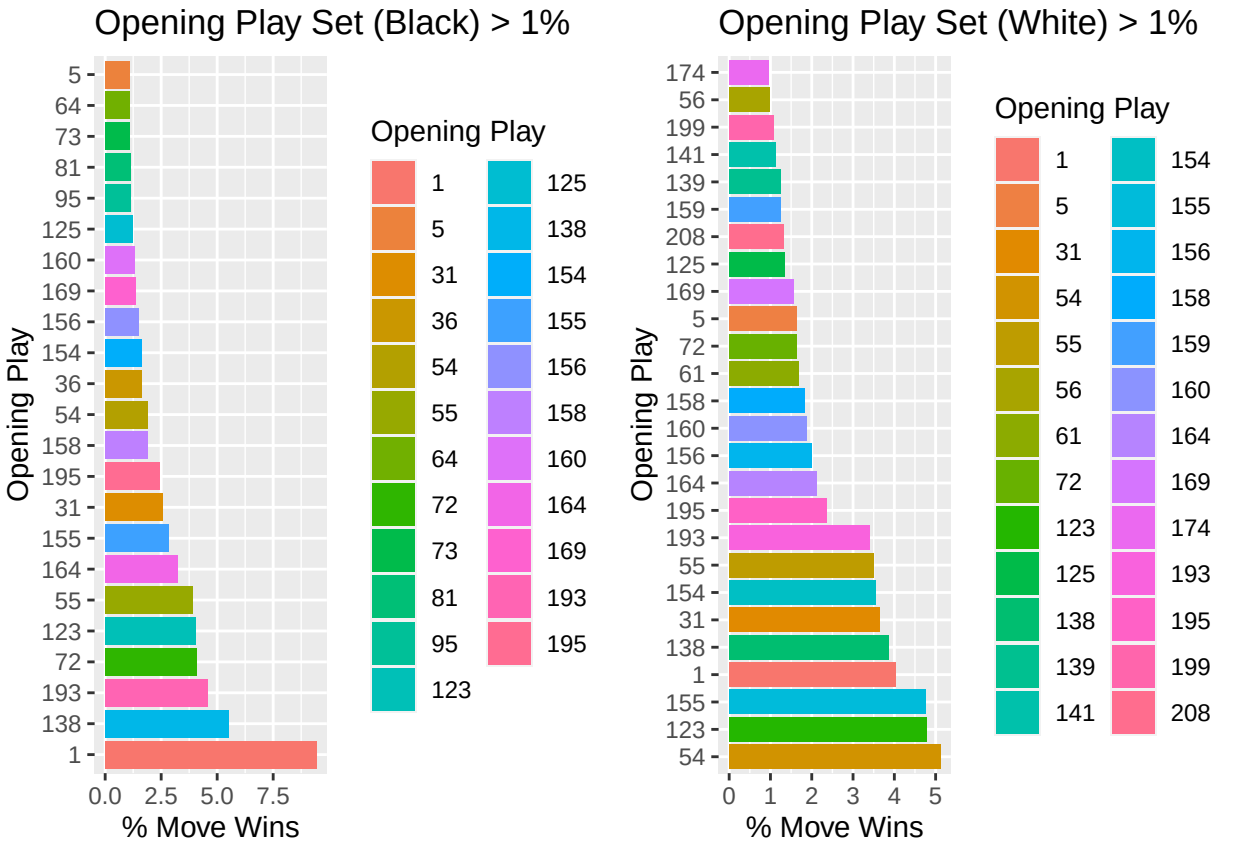
Using the information from Table 1., the data can further be categorized into the most successful move sets for winning based on the total wins of each color. Using the combined mean of both 'white' and 'black' the categories are separated into chances of win at Small (<.45%), Medium (.45 - .9%) and Large (>1%). Figure 1. Shows out of all the wins for a particular color, how many move sets have the highest chance of winning a match. 'White' shows a higher count for 'small' and 'large' chances of winning between move sets but interestingly 'black' has move sets that give a better medium advantage within the categories. If there is a possibility of overcoming the ~3% First-move advantage that 'white' has, it may lie in these opening moves.

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
```



(Fig.1) Shows a categorized approach of looking at the move sets that have the highest chance of influencing wins out of all wins within the group. Out of the 272 different move sets listed, white has a total of 48 that are above .45% win ratio with black having 52 above .45% win ratio.

Exploring 'white' and 'black' win ratios further and looking into identifying all top move sets greater than 1% can show the top three winning moves. Figure 1. shows two moves in particular for 'black' that out performs the rest of the opening plays. Move 1 refers to A00 Polish (Sokolsky) opening which consists of over 9% of all wins by 'black'. The second opening move set that has the highest win ratio is also 'black' with C20 Kings pawn game at 5.5%. 'White' has the third highest win ration with opening move 54 which is Owens Defense.



Appendix

```
library(tidyverse)
library(ChessPack)
library(ggplot2)
library(knitr)
library(readr)
library(cowplot)
library(kableExtra)
anno.table <- data.frame( Parameter = c('rated', 'turns', 'winner', 'opening_eco',
                                        'opening_name', 'opening_ply', 'winner2',
                                        'game_type'),
                          Data_type = c('number', 'integer', 'character', 'factor',
                                        'character', 'integer', 'number', 'number'),
                          Discription = c('Games are rated or unrated in play',
                                          'The amount of turns for a game', 'Winner of the match by col
                                          'Opening move assigned 1-272', 'Name of opening move used',
                                          'Number of moves in the opening phase',
                                          'Dummy variable assigned to winner color (1=White, 0=Black)',
                                          'Dummy variable assinged to type of game (1=Timed, 0=Un-timed
```

```
kable(anno.table, valign = 't',
caption = 'Table 1. Chess Data Annotation') %>%
  kable_styling(latex_options = 'hold_position')
# turn opening plays into factors
chess.small<- chess.small %>%
  mutate(opening_eco=factor(opening_eco))
# wins vs opening moves
opening.wins <- chess.small %>% count(opening_eco, winner, sort = TRUE)%>%
  arrange(n)
# plot overall wins between black and white
ggplot(opening.wins, aes(x=winner,  y= n, fill = winner)) +
  geom_col() +
  annotate('text', x=1,y=2500, label='2981') +
  annotate('text', x=2,y=3000, label='3344') +
  ylim(0,3500) +
  labs(x = "Winner", y = "Wins", title = "Total Wins") +
  theme(plot.title = element_text(hjust = 0.5))

# create number of matches black/white win by play
pivot_games <- pivot_wider(opening.wins, names_from = winner, values_from = n)
# creates data frame for black wins drops na
# adds total games and percent of moves to wins
black_win <- select(pivot_games, opening_eco, black) %>% drop_na() %>%
  mutate('total_games' = sum(black)) %>%
  mutate('percent_move_win'= (black/total_games)*100)

# creates data frame for white wins drops na
# adds total games and percent of moves to wins
white_win <- select(pivot_games, opening_eco, white) %>% drop_na() %>%
  mutate('total_games' = sum(white)) %>%
  mutate('percent_move_win'= (white/total_games)*100)
# create table for percent and game win camparison
table_black <- summary(black_win[c('black', 'percent_move_win')])
table_white <-summary(white_win[c('white', 'percent_move_win')])

# knit table_black, table_white together for comparison
kables(
  list(
    # change column names for comparison
    kable(table_black, col.names = c('Black Wins', 'Percent to Win'),
          valign = 't'),
    kable(table_white, col.names = c('White Wins', 'Percent to Win'),
          valign = 't'))) %>%
  kable_styling(latex_options = 'hold_position')
# categorize for black to win chance small, medium, large
categorized_black <- cut(black_win$percent_move_win, breaks = c(0, .45,.94, 9.5),
                         labels = c("small (<.45%)", "medium (.45-.9%)","large (>1%)"))
# categorize for white to win chance small, medium, large
categorized_white <- cut(white_win$percent_move_win, breaks = c(0, .45,.94, 5.5),
                         labels = c("small (<.45%)", "medium (.45-.9%)","large (>1%)"))
# create black category d.f.
final_set_black <- data.frame(black_win, categorized_black)
# create white category d.f.
```

```
final_set_white <- data.frame(white_win, categorized_white)
# plot black grouped wins
fig_1.1 <- ggplot(final_set_black, aes(x = categorized_black,fill = categorized_black)) +
  geom_bar() +
  labs(title = 'Chance of Black Win by Grouped Moveset', x = " ") +
  # add counts to bars
  geom_text(aes(label=..count..), stat = 'count',vjust= 1) +
  # color bars for categories
  scale_fill_brewer(palette="Set1") +
  # adjust title
  theme(plot.title = element_text(hjust = 0.5))
# plot white grouped wins
fig_1.2 <- ggplot(final_set_white, aes(x = categorized_white, fill = categorized_white)) +
  geom_bar() +
  labs(title = 'Chance of White Win by Grouped Moveset',x = " ") +
  # add counts to bars
  geom_text(aes(label=..count..), stat = 'count',vjust=1) +
  # color bars for categories
  scale_fill_brewer(palette="Set1") +
  # adjust title
  theme(plot.title = element_text(hjust = 0.5))
# combine white, black, plots
plot_grid(fig_1.1,fig_1.2,nrow = 2)
# subset white d.f. top moves > .95%
most_win_white <- subset(final_set_white, percent_move_win > .95)
#head(most_win_white)
# subset white d.f. medium moves
med_win_white <- subset(final_set_white, percent_move_win > .45 & percent_move_win < .95)
# subset black d.f. top moves > .95%
most_win_black <- subset(final_set_black, percent_move_win > .95)
#head(most_win_black)
# subset black d.f. medium moves
med_win_black <- subset(final_set_black, percent_move_win > .45 & percent_move_win < .95)
# create plot for black top moves > .95%
# refactor opening_eco for display continuity
fig_2.1 <- ggplot(most_win_black,aes(fct_rev(fct_reorder(opening_eco,
                                                         percent_move_win)),
                                     percent_move_win)) +
  geom_col(aes(fill = opening_eco)) +
  labs(x='Opening Play', y= '% Move Wins',
       title = 'Opening Play Set (Black) > 1%', legend = 'Opening Play') +
  # change legend title
  guides(fill=guide_legend(title ='Opening Play')) +
  # flip coordinates for clarity
  coord_flip()
# create plot for white top moves > .95%
# refactor opening_eco for display continuity
fig_2.2 <- ggplot(most_win_white,aes(fct_rev(fct_reorder(opening_eco,
                                                         percent_move_win)),
                                     percent_move_win)) +
  geom_col(aes(fill = opening_eco)) +
  labs(x='Opening Play', y= '% Move Wins',
       title = 'Opening Play Set (White) > 1%', legend = 'Opeing Play') +
```

```
  # change legend title
  guides(fill=guide_legend(title ='Opening Play')) +
  # flip coordinates for clarity
  coord_flip()
# plot top moves white/black for comparison
plot_grid(fig_2.1,fig_2.2,ncol = 2)
```