Mehra_Kai_Assignment-2

Kai Mehra

2023-02-07

Libraries

```
# Load the {tidyverse}, {ggplot2}, and {ggpubr} libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(ggpubr)
```

Loading Data

```
# load the high_elo_opening.csv dataset as chess_data
chess_data <- read_csv("../Data/high_elo_opening.csv")
```

```
## Rows: 1884 Columns: 24
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): opening_name, side, ECO, moves_list, move1w, move1b, move2w, move...
## dbl  (11): num_games, perf_rating, avg_player, perc_player_win, perc_draw, p...
## date  (1): last_played_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(chess_data)
```

```
## [1] 1884   24
```

The chess_data website is Kaggle dataset containing information on high-level chess games played in 2017 and 2018. The dataset is organized around the openings played in a large number of games. The dataset contains 1884 openings, however many of these are variations of the same base opening. The dataset contins information on the average rating of the players in the game, the percentage of white or black winning or a draw, and the moves making up the opening.

Research Question

In the game of chess, the player using the white piece gets to play the first move. Traditionally, it is believed that this gives the player playing white an advantage. This idea has been backed up using chess engine analysis which are based on complicated computer algorithms and systems that evaluate positions. From the starting position, white is given an advantage of around points, which is equivalent to white having 0.5 more pawns than black. While not a massive advantage, this can strongly influence play at the highest level. Therefore, top-level players with the white pieces generally go for the win, while players with the black pieces often play for a draw. However, if a player with black attempts to go for the win, it is believed that they have to make a riskier (riskier in the sense that they could lose easily) opening choice than a player with white.

In this analysis, I wanted to test the validity of these beliefs and try to answer if white truly has an advantage over black in the game of chess? Does white's privilege to move first contribute to it having more openings and options to win games without taking as many risks as black? I hypothesize that these ideas are true due to the high-stakes nature of competitive chess.

Variables of Interest

The main variables of interest are the opening name, number of games, the average rating of the players in the game, percent that the white player won, percent that the black player one, percent that the game ended in a draw, and the moves making up the opening, specifically the first move for white and black. # Data Wrangling

```
# select the variables of interest
chess_openings <-
  chess_data %>%
  select(opening_name,
         num_games,
         avg_player,
         perc_white_win,
         perc_draw,
         perc_black_win,
         move1w,
         move1b)
```

```
# filter out incorrectly input data and N/A values
chess_openings <-
  chess_openings %>%
  filter(move1w != "00E+03" &
         move1w != "00E+04") %>%
  filter(!is.na(move1b))
```

```r
# Filter out games by lower rated players, and separate the opening into its
# base name and corresponding variations
chess_openings <-
  chess_openings %>%
  filter(avg_player >= 2000) %>%
  separate(opening_name, into = c("name", "variation_1", "variation_2"),
           sep = ",", convert = TRUE)
```

```
## Warning: Expected 3 pieces. Additional pieces discarded in 242 rows [16, 34, 53, 62,
## 130, 131, 153, 154, 156, 166, 169, 170, 171, 177, 178, 213, 214, 215, 218, 254,
## ...].
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 671 rows [1, 2, 3, 4, 8,
## 10, 11, 21, 22, 23, 24, 25, 41, 42, 44, 46, 59, 60, 63, 65, ...].
```

I only wanted to consider games by players rated 2000 and above according to the ELO (a system of rating chess players_. Players at or above this rating are generally professional or at least semi-professional players who compete at high-level tournaments. Lower rated players do not play as well as high rated players, and higher-rated players usually play well-studied and theoretically advantageous openings.

Openings have many different variations, often 2 or three nested variations. To simplify the analysis, I mainly focused on the base openings as a collective rather than individual variations.

```r
# Filter out N/A values
chess_openings <-
  chess_openings %>%
  filter(!is.na(perc_white_win) &
         !is.na(perc_black_win) &
         !is.na(perc_draw))

# Add a variable called mode_result
chess_openings <-
  chess_openings %>%
  mutate(mode_result = case_when(
    (perc_white_win > perc_black_win & perc_white_win > perc_draw) ~ "white",
    (perc_black_win > perc_white_win & perc_black_win > perc_draw) ~ "black",
    (perc_draw >= perc_white_win & perc_draw >= perc_black_win) ~ "draw")) %>%
  filter(!is.na(mode_result)) # remove N/A values
```

I created the variable mode_result which represents the most likely result for that specific opening. Essentially, the mode_result was whichever result occured in the highest percentage of games for that opening.

```r
# creating the chess_openings_simplified data frame
chess_openings_simplified <-
  chess_openings %>%
  group_by(name) %>% # group observations by opening name
  # calculate the mean win, draw, and number of games for all openings
  #irrespective of variation
  summarise(mean_white_win = round(mean(perc_white_win, na.rm = T), 2),
            mean_black_win = round(mean(perc_black_win, na.rm = T), 2),
            mean_draw = round(mean(perc_draw, na.rm = T), 2),
            total_num_games = sum(num_games)) %>%
```

```
  # adding the mode_result variable but instead on an condesed opening scale
  mutate(mode_result = case_when(
    (mean_white_win > mean_black_win & mean_white_win > mean_draw) ~ "white",
    (mean_black_win > mean_white_win & mean_black_win > mean_draw) ~ "black",
    (mean_draw >= mean_white_win & mean_draw >= mean_black_win) ~ "draw")) %>%
  # filtering N/A values
  filter(!is.na(mode_result))
```

I created the chess_openings_simplified data frame which condensed all of the variations of every opening under the base opening name. I computed the mean win and draw percentages and the total number of games played for each condensed opening. I then added the same mode_result variable as before, but it was calcluated on the mean win and draw percentages.

Additional Data Wrangling and Analysis

Expected Value of each opening

```
chess_openings_simplified <-
  chess_openings_simplified %>%
  # create the white_ev and black_ev variables
  mutate(
    white_ev = ((1 * mean_white_win + .5 * mean_draw)/100),
    black_ev = ((1 * mean_black_win + .5 * mean_draw)/100)
  )
```

In most chess tournaments, wins are worth 1 point and 0.5 points for draws. Thus, I calculated the expected value of each opening for white and black based on the win and draw percentages. An expected value close to 0 means the opening is terrible for that player, as they are expected to lose while an expected value close to 1means the opening is great for that player, as they are expected to win. Most openings fall around 0.5 which predicts a draw in the game.

Calculating overall mean win and draw rates

```
# calculating overall summary statistics for the full dataset
ov_mean_white_win <- round(mean(chess_openings$perc_white_win, na.rm = TRUE), 2)
ov_mean_black_win <- round(mean(chess_openings$perc_black_win, na.rm = TRUE), 2)
ov_mean_draw <- round(mean(chess_openings$perc_draw, na.rm = TRUE), 2)

ov_mean_white_win
```

```
## [1] 39.45
```

```
ov_mean_black_win
```

```
## [1] 29.98
```

```
ov_mean_draw
```

```
## [1] 30.57
```

```
# calculating overall expected values for the simplified dataset
ov_white_ev <- round(mean(chess_openings_simplified$white_ev, na.rm = TRUE), 3)
ov_black_ev <- round(mean(chess_openings_simplified$black_ev, na.rm = TRUE), 3)

ov_white_ev
```

```
## [1] 0.544
```

```
ov_black_ev
```

```
## [1] 0.456
```

I calculated important summary statistics for the dataset to use as comparative values. Overall, white has the highest likelihood of winning, while drawing is more likely than black winning. White also has an expected value nearly 0.1 points higher.

Risky Openings

```
# creating the risky_opening binary variable
chess_openings_simplified <-
  chess_openings_simplified %>%
  mutate(
    risky_opening = (mean_white_win >= mean_draw & mean_black_win >= mean_draw)
  ) %>%
  filter(!is.na(risky_opening)) # removing N/A variables
```

I defined a risky opening as one where the likelihood of losing is less than the likelihood of drawing. A risky opening involves a player playing in such a way that they give themselves a greater shot at winning while accepting that they have a higher likelihood of losing.

Table of best openings for White and Black by Expected Value

```
chess_openings_simplified %>%
  arrange(-white_ev) %>% # arranging by highest expected values
  slice(1:5)
```

```
## # A tibble: 5 x 9
##   name           mean_~1 mean_~2 mean_~3 total~4 mode_~5 white~6 black~7 risky~8
##   <chr>            <dbl>   <dbl>   <dbl>   <dbl> <chr>     <dbl>   <dbl> <lgl>
## 1 Neo-Grünfeld ~    55.2    22.0    22.8    1582 white     0.666   0.334 FALSE
## 2 Latvian Gambit    55.2    26.8    18.0     856 white     0.642   0.358 TRUE
## 3 Budapest Defe~    49.6    26.8    23.6    7777 white     0.614   0.386 TRUE
## 4 Elephant Gamb~    52.6    30      17.4     593 white     0.613   0.387 TRUE
```

```
## 5 Catalan Openi~    43.1    20.9    36.0   36178 white     0.611   0.389 FALSE
## # ... with abbreviated variable names 1: mean_white_win, 2: mean_black_win,
## #   3: mean_draw, 4: total_num_games, 5: mode_result, 6: white_ev, 7: black_ev,
## #   8: risky_opening
```

```
chess_openings_simplified %>%
  arrange(-black_ev) %>% # arranging by highest expected values
  slice(1:5)
```

```
## # A tibble: 5 x 9
##   name           mean_~1 mean_~2 mean_~3 total~4 mode_~5 white~6 black~7 risky~8
##   <chr>            <dbl>   <dbl>   <dbl>   <dbl> <chr>     <dbl>   <dbl> <lgl>
## 1 Grob Opening      29      49.7    21.3     246 black     0.396   0.604 TRUE
## 2 Indian Game D~    23      38.8    38.2     492 black     0.421   0.579 FALSE
## 3 King's Knight~    33.6    44.4    22.0     710 black     0.446   0.554 TRUE
## 4 Danish Gambit     31.4    38.5    30.1     366 black     0.464   0.536 TRUE
## 5 Mieses Opening    36.9    42.8    20.3     236 black     0.470   0.530 TRUE
## # ... with abbreviated variable names 1: mean_white_win, 2: mean_black_win,
## #   3: mean_draw, 4: total_num_games, 5: mode_result, 6: white_ev, 7: black_ev,
## #   8: risky_opening
```

The tables above show the top 5 openings by expected value by white and black. White clearly has an advantage as the 5th best opening by expected value is greater the the expected value for black.

Difference in Expected Value conditional on white or black winning

```
chess_openings_simplified %>%
  filter(mode_result == "white") %>% # selecting openings favoring white
  summarise(mean(white_ev-black_ev))
```

```
## # A tibble: 1 x 1
##   `mean(white_ev - black_ev)`
##                         <dbl>
## 1                       0.108
```

```
chess_openings_simplified %>%
  filter(mode_result == "black") %>%  # selecting openings favoring white
  summarise(mean(black_ev-white_ev))
```

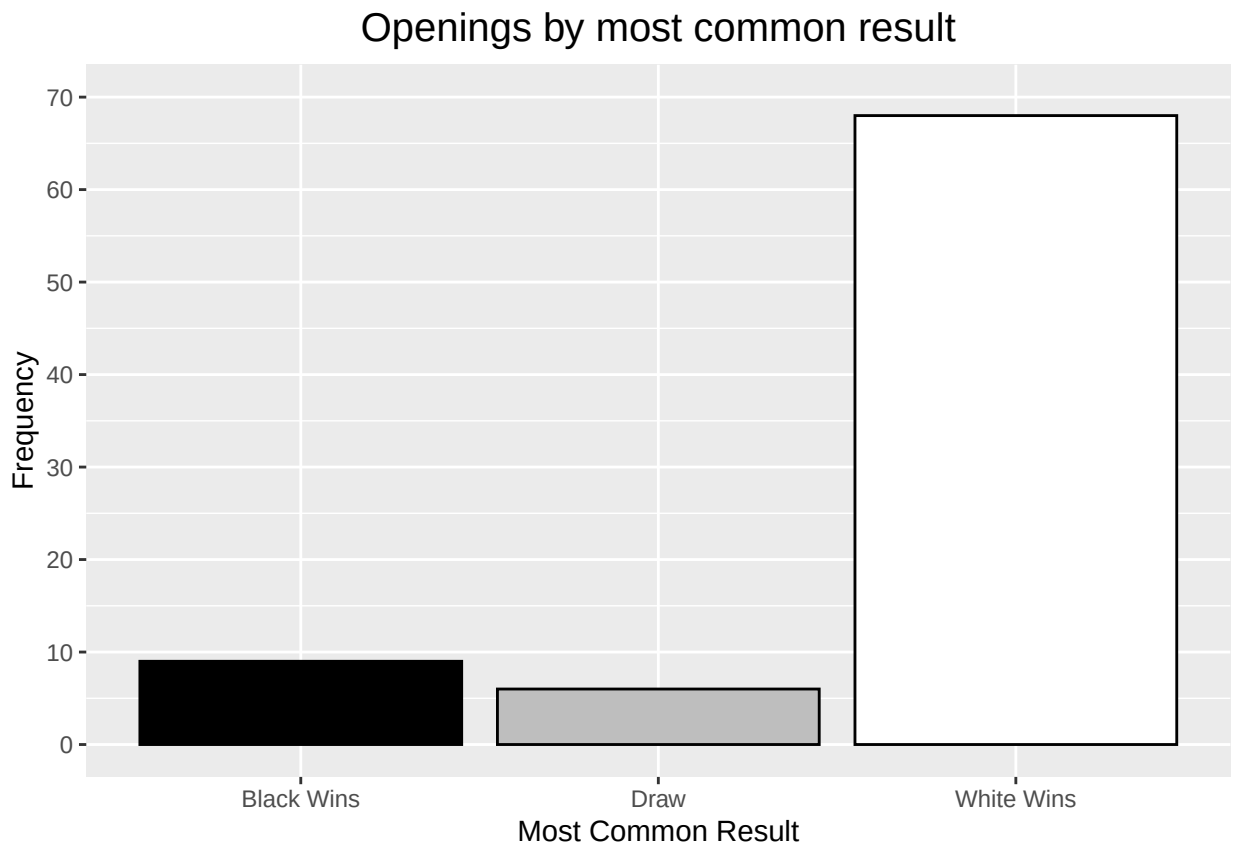
```
## # A tibble: 1 x 1
##   `mean(black_ev - white_ev)`
##                         <dbl>
## 1                      0.0749
```

These values show that for openings where white is expected to win, players with white have around a 0.1 point advantage over players with black For openings where black is expected to win, players with black have only a 0.07 point advantage over players with white. Thus, when white is playing a favorable opening they more advantage than black playing a favorable opening.

Visualizations

Visualization 1: Histogram by most common result

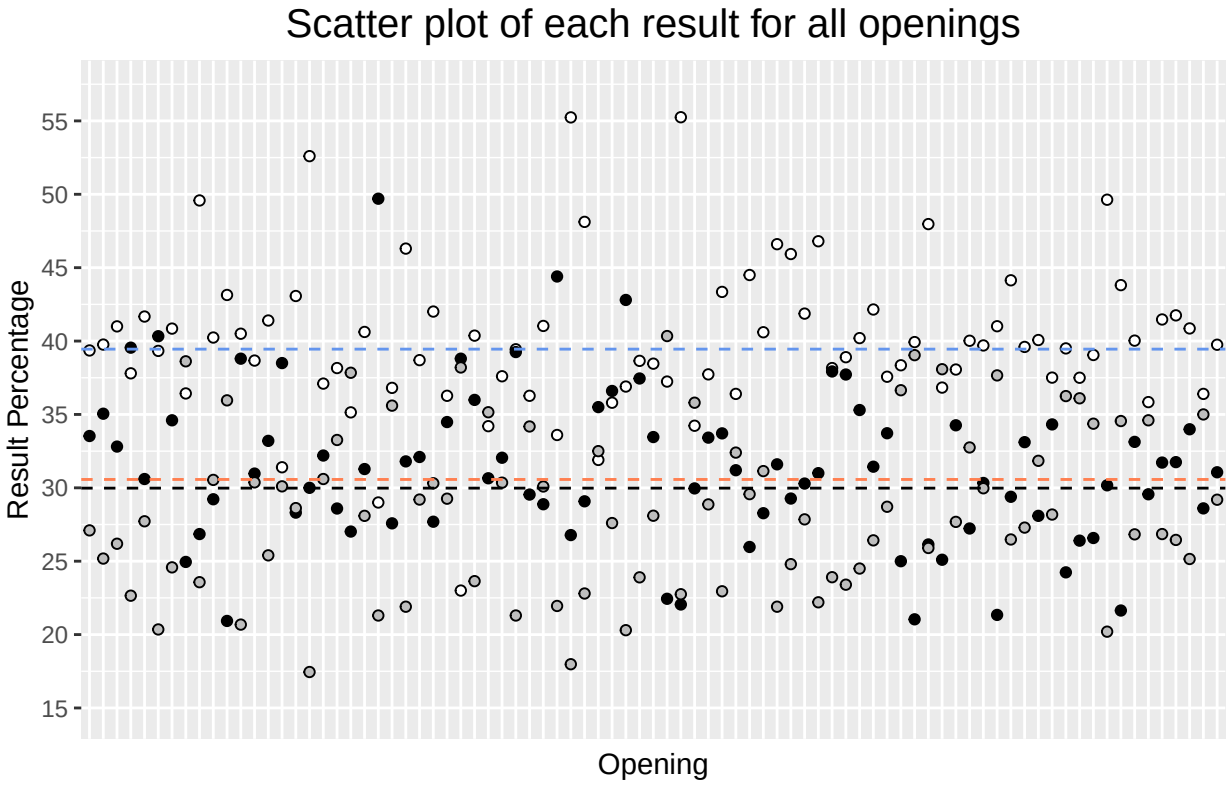
```
chess_openings_simplified %>%
  ggplot() +
  geom_bar(
    aes(
      x = mode_result,
      fill = factor(mode_result)),
    color = "black"
  ) + # bar plot colored by the mode result
  scale_fill_manual(
    values = c("black",
               "grey",
               "white")
  ) + # manually filling colors
  labs(
    x = "Most Common Result",
    y = "Frequency",
    title = "Openings by most common result"
  ) + # better labels and titles
  scale_x_discrete(
    labels = c("Black Wins", "Draw", "White Wins")
  ) + # better labels
  scale_y_continuous(
    limits =c(0, 70),
    breaks = seq(0, 70, 10)
  ) + # better axis scale
  theme(
    plot.title = element_text(size = 15, hjust = 0.5),
    legend.position = "none"
  ) # center the title
```



Visualization 2: Scatterplot of openings by percentage of each result

```
chess_openings_simplified %>%
  arrange(mean_white_win) %>%
  ggplot() +
  geom_point(
    aes(
      x = name,
      y = mean_white_win
    ),
    color = "black",
    fill = "white",
    shape = 21
  ) + # scatterplot average percent of white winning for each opening
  geom_point(
    aes(
      x = name,
      y = mean_black_win
    ),
    color = "black"
  ) + # adding points for average percent of black winning for each opening
  geom_point(
    aes(
      x = name,
```

```
      y = mean_draw
    ),
    color = "black",
    fill = "grey",
    shape = 21
  )  + # adding points for average percent of draws for each opening
  geom_hline(
    yintercept = ov_mean_white_win,
    linetype = 2,
    color = "cornflowerblue"
  ) + # adding overall mean white winning percentage for context
  geom_hline(
    yintercept = ov_mean_black_win,
    linetype = 2,
    color = "black"
  ) + # adding overall mean black winning percentage for context
   geom_hline(
    yintercept = ov_mean_draw,
    linetype = 2,
    color = "coral"
  ) + # adding overall mean draw percentage for context
  theme(
    plot.title = element_text(size = 15, hjust = 0.5),
    axis.ticks.x = element_blank(),
    axis.text.x = element_blank()
  ) + # center title and remove x ticks and labels
  labs(
    x = "Opening",
    y = "Result Percentage",
    title = "Scatter plot of each result for all openings",
    caption = "Horizontal lines represent overall percent frequency of each result. Blue = white, Coral
"
  ) + # better labels
  scale_y_continuous(
    limits = c(15, 57),
    breaks = seq(15, 57, 5)
  ) # better y axis scale
```



Horizontal lines represent overall percent frequency of each result. Blue = white, Coral = draw, Black = black.

Visualization 3: Most popular opening move bar chart

Most popular opening move by white colored by result

```
white_move1 <-
  ggplot(chess_openings) +
geom_bar(
  aes(
    x = move1w,
    fill = factor(mode_result)),
  color = "black",
  position = "dodge" # putting bars next to eachother
  ) + # barplot
  labs(
    x = "Move 1 for white",
    y = "Frequency",
    title = "White move 1",
    fill = "Most common result"
  ) + # better labels
  scale_fill_manual(
    values =c("black", "grey", "white"),
    labels=c("Black Wins", "Draw", "White Wins")
  ) + # filling by result
  scale_y_continuous(
```

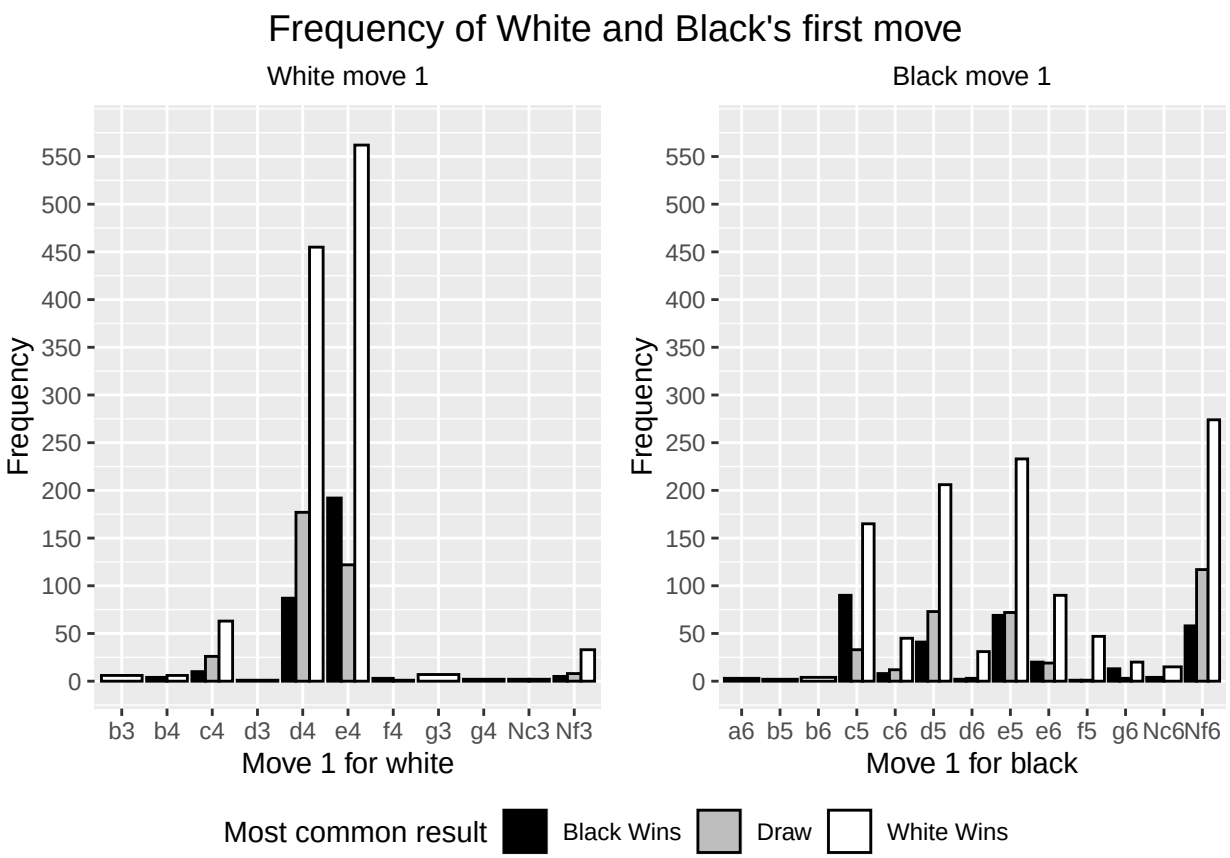
```
    limits = c(0, 575),
    breaks = seq(0, 575, 50)
  ) + # better scale
  theme(
    plot.title = element_text(size = 10, hjust = 0.5)
  ) # centering the title
```

Most popular opening move by black colored by result

```
black_move1 <-
  ggplot(chess_openings) +
geom_bar(
  aes(
    x = move1b,
    fill = factor(mode_result)),
  color = "black",
  position = "dodge"
  ) + # bar plot
  labs(
    x = "Move 1 for black",
    y = "Frequency",
    title = "Black move 1",
    fill = "Most common result"
  ) + # better labels
  scale_fill_manual(
    values =c("black", "grey", "white"),
    labels=c("Black Wins", "Draw", "White Wins")
  ) + # filling by result
  scale_y_continuous(
    limits = c(0, 575),
    breaks = seq(0, 575, 50)
  ) + # better scaling
  theme(
    plot.title = element_text(size = 10, hjust = 0.5),
    legend.position = "bottom"
  ) # centering the tile and putting the legend on the bottom
```

```
combo_white_black_move1 <-
  ggarrange(
  white_move1, black_move1,
  ncol = 2,
  common.legend = TRUE,
  legend = "bottom"
) # combing the two plots above into one

annotate_figure(combo_white_black_move1, top = text_grob(
  "Frequency of White and Black's first move",size = 14))
```




```
# better title placement
```

Visualization 4: Bar plot of risky or not risky openings for white and black

White

```
white_risk <-
  chess_openings_simplified %>%
  filter(mode_result == "white") %>% # results where white wins
  ggplot() +
  geom_bar(
    aes(
      x = risky_opening,
      y = (..count..)/sum(..count..), # percentage
      fill = factor(risky_opening))
  ) + # barplot
  scale_fill_manual(
    values = c("cornflowerblue", "coral")
  ) + # fill by risky or not
  labs(
    x = "Opening Risk",
    y = "Proportion of openings",
    title = "White"
  ) + # better labels
```

```
  scale_y_continuous(
    limits = c(0 ,1),
    breaks = seq(0, 1, .1)
  ) + # better scale
  scale_x_discrete(
    labels = c("Safe Opening", "Risky Opening")
  ) + # better labels
  theme(
    plot.title = element_text(size = 12, hjust = 0.5),
    legend.position = "none"
  ) # center plot and no legend
```

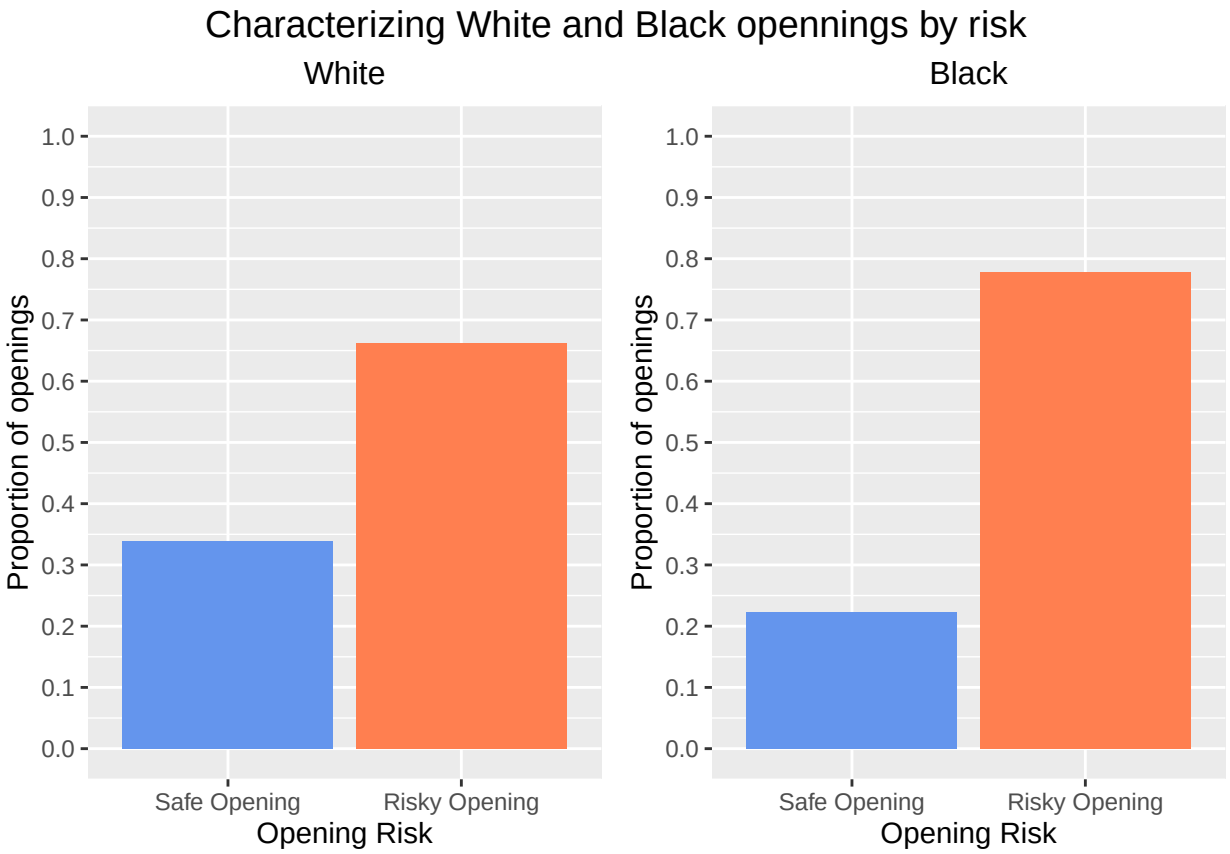
Black

```
black_risk <-
  chess_openings_simplified %>%
  filter(mode_result == "black") %>% # black wins
  ggplot() +
  geom_bar(
    aes(
      x = risky_opening,
      y = (..count..)/sum(..count..), # percentage
      fill = factor(risky_opening))
  ) + # bar plot
  scale_fill_manual(
    values = c("cornflowerblue", "coral")
  ) + # fill by risky or not
  labs(
    x = "Opening Risk",
    y = "Proportion of openings",
    title = "Black"
  ) + # better labels
  scale_y_continuous(
    limits = c(0 ,1),
    breaks = seq(0, 1, .1)
  ) + # better scale
  scale_x_discrete(
    labels = c("Safe Opening", "Risky Opening")
  ) + # better labels
  theme(
    plot.title = element_text(size = 12, hjust = 0.5),
    legend.position = "none"
  )  # center plot and no legend
```

```
risk_plot <- ggarrange(white_risk, black_risk,
          ncol = 2) # combining plots into one
```

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

```
annotate_figure(risk_plot, top = text_grob(
  "Characterizing White and Black openings by risk",size = 14))
```

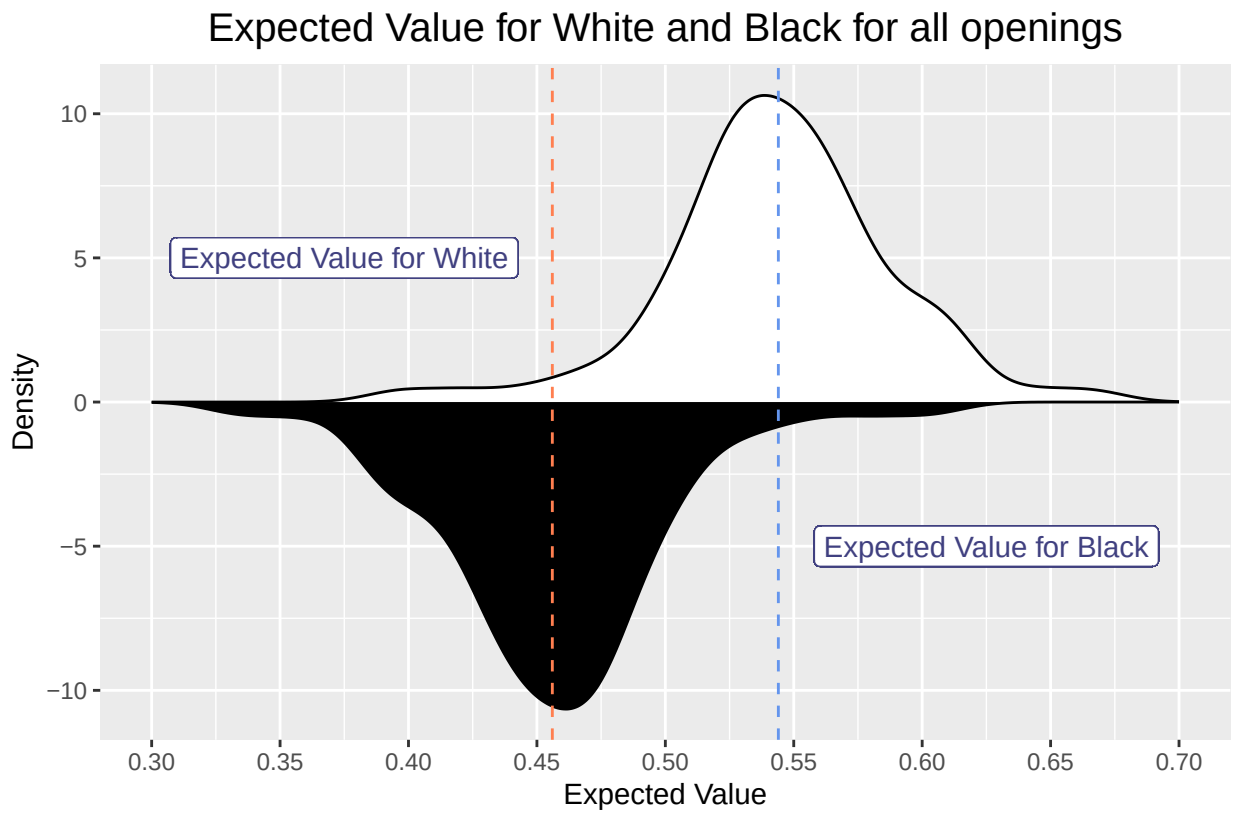



```
# fixing the title
```

Visualization 5: Openings by Expected Value

```
chess_openings_simplified %>%
  ggplot() +
  geom_density(
    aes(
      x = white_ev,
      y = ..density..),
    fill="white"
    ) + # density plot
  geom_label(
    aes(
      x=0.375,
        y=5,
        label="Expected Value for White"),
    color="#404080"
    ) + # manually placing a label
  geom_density(
```

```
  aes(
    x = black_ev,
    y = -..density..), # black values upside dwon
  fill= "black"
  ) + # density plot
geom_label(
  aes(x=0.625, y=-5,
      label="Expected Value for Black"
      ),
  color="#404080") + # manually placing a label
scale_x_continuous(
  limits = c(0.3, .7),
  breaks = seq(0.3, .7, 0.05)
) + # better scale
geom_vline(
  xintercept = ov_white_ev,
  linetype = 2,
  color = "cornflowerblue"
) + # average white expected value for context
geom_vline(
  xintercept = ov_black_ev,
  linetype = 2,
  color = "coral"
) + # average black expected value for context
labs(
  x = "Expected Value",
  y = "Density",
  title = "Expected Value for White and Black for all openings",
  caption = "The vertical lines indicate the overall average expected value for white and black. Blue
theme(
  plot.title = element_text(size = 15, hjust = 0.5)
) # center the title
```



The vertical lines indicate the overall average expected value for white and black. Blue = white, coral = black

Discussion

The the summary statistics at the beginning of the analysis show that white wins nearly 40% of games while draws and black wins happen only about 30% of the time. Further, the average expected value for white of 0.544 was nearly a tenth greater than the expected value for black of 0.456. These summary statistics give the indication that white does have an advantage.

The five visualizations support the idea that white has an advantage by considering the overall results, simplicity of first move, risk of opening, and density of expected value.

The barplot shows a simple breakdown of openings by their most common result. A vast majority of openings in the data set are most commonly won by white rather than a draw or black win. Openings favoring black barely beat out draws, but they are both dominated by openings favoring white. This gives strong credence to the idea that white has an advantage by going first.

The scatterplot of each opening shows the percent frequency of each opening for each of the three results. The horizontal dashed lines show the average win percentage for white (in blue), percentage of draws (in coral), and win percentage for black (in black). While the specific opening names are not listed in the plot, it is clear that white is favored in a plurality of openings, and their average win percentage is much higher than the other results. Further, in openings favoring a draw, a white win is often the next most common result. In fact, there is only one opening where a white win is the least likely ("Indian Game Defense") however it has been played less than 500 times compared to thousands for other openings. Overall, this plot continues to reinforce the idea that it is easier to win with the white pieces.

The next barplot shows the frequency of the first moves for white and black, and it is broken down by the most commons result. White primarily plays one of three moves (pawn to e4, d4, or c4), and for all of these

openings white is strongly favored. Comparatively, black plays a vast variety of moves, but all of them still favor white. The famous "Sicilian Defense" which starts with pawn to c5 is known in chess communities for being the best opening to play when black needs a win. The data backs this up, but white is still favored in this opening. White has a much more streamlined simple plan to get a win, while players with the black pieces have to study many more openings to try and counteract the white advantage.

The next bar plot characterizes openings favoring white and black respectively by risk. Again, a risky opening is one where you are more likely to lose than draw. Openings which favor white are most of the time risky, with around 65% of them being considered risky openings. However, openings which favor black are considered risky openings nearly 80% of the time. This shows that black has to take more risks if they want to win games, while white can be a bit more secure.

Finally, the double density plot shows the distributions of the expected value for white and black. White players have a strong advantage that is clearly shown in their distribution and average expected value. This final plot shows that it is not just a few amazing openings for white contributing to their advantage, rather most openings favor white and contribute to them performing well.

Overall, this analysis confirms traditional chess thinking. The player using the white pieces is favored in the vast majority of openings, and they have a much higher expected value of tournament points when playing with white rather than black. Players with black have to play riskier openings if they try to win, and they overall are at a distinct disadvantage when playing the games. This analysis confirms the thinking the large, high-stakes tournament organizers employ when they ensure that players get equal chances to play with the white and black pieces in each round.