Chess Analysis: Initial Report

Kyle Sherman

3/28/2022

Section 1: Introduction

There are 3 phases of a standard chess game with white getting the first move: The opening, the middlegame, and the endgame. While the middlegame and endgame do play an important role, the opening sets the state for the flow and development of the rest of the game. A good opening will give a player an advantage when entering the middle game, prepare pieces to launch an attack, and set the defense all at the same time. They play such an important role that many professional players will spend years studying and perfecting different openings to see which ones will impact their game.

Like many others around the world, I am an avid fan of the game and play it often – though I am far from the definition of a good player. Over the years, and hundreds of losses, I have found myself curious of the following questions which will serve as my primary research for this project: 1. Is there an implicit bias in favor of the white piece as a result of getting the first move? 2. Which opening's have the highest win rate, and which have the lowest win rate? 3. Is there a correlation between the above and the player's ELO (skill rating)?

I have acquired a dataset from Kaggle consisting of more than 20,000 chess games played on the 2nd largest online chess platform 'Lichess.com'. The uploader of the data collected it using Lichess.com' public access API. The observations contain all the necessary data from over 20,000 chess games which should be enough to find any potential correlations relating to my research questions.

The following are the included variables:

```
Rated (T/F)
Start Time
End Time
Number of Turns
Game Status
Winner
Time Increment
White Player Rating
Black Player Rating
```

All Moves in Standard Chess Notation
Opening Eco (Standardized Code for any given opening)
Opening Name
Opening Ply (Number of moves in the opening phase)

Section 2: Data

readme.txt

Dataset of Chess Games on the Lichess Platform

Data Collected 9/21/2019 Dataset Obtained from https://www.kaggle.com/datasets/datasnaek/chess original source https://database.lichess.org/

About the dataset:

This dataset is a collection of more than 20,000 chess games and their associated data. The data was collected using the public API from Lichess, the second largest online chess platform. The data was collected by by observing the games of chess teams on the platform with over 1,500 players each.

```
Variable List:
id (character): The ID of the game
rated (logical): True if the game was a rated game; False if it was unrated
created_at (double): UNIX time when the game started
last_move (double): UNIX time when the game ended
turns (double): The total number of turns in the game
victory_status (character): The type of victory; out of time; mate; resignmation; draw
winner (character): White or Black won the game
increment_code (character): Total length of the game
white_id (character): ID of the white player
white_rating (double): ELO Rating of the white player at the time of the game
black_id (character): ID of the black player
black_rating (double): ELO Rating of the black player at the time of the game
moves (character): List of each move made in the game
opening_eco (character): 3 digit code for the opening
opening_name (character): full opening name
opening_ply (double): number of moves in the opening
```

glimpse()

```
> glimpse(chess_games)
Rows: 20,058
Columns: 16
$ id             <chr> "TZJHLljE", "l1NXvwaE", "mIICvQHh", "kWKvrqYL", "9tXo1AUZ", "MsoD
$ rated          <lgl> FALSE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, FA
$ created_at     <dbl> 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.
$ last_move_at   <dbl> 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.5e+12, 1.
$ turns          <dbl> 13, 16, 61, 61, 95, 5, 33, 9, 66, 119, 39, 38, 60, 31, 31, 43, 52
$ victory_status <chr> "outoftime", "resign", "mate", "mate", "mate", "draw", "resign",
$ winner         <chr> "white", "black", "white", "white", "white", "draw", "white", "bl
$ increment_code <chr> "15+2", "5+10", "5+10", "20+0", "30+3", "10+0", "10+0", "15+30",
$ white_id       <chr> "bourgris", "a-00", "ischia", "daniamurashov", "nik221107", "trel
$ white_rating   <dbl> 1500, 1322, 1496, 1439, 1523, 1250, 1520, 1413, 1439, 1381, 1381,
$ black_id       <chr> "a-00", "skinnerua", "a-00", "adivanov2009", "adivanov2009", "fra
$ black_rating   <dbl> 1191, 1261, 1500, 1454, 1469, 1002, 1423, 2108, 1392, 1209, 1272,
$ moves          <chr> "d4 d5 c4 c6 cxd5 e6 dxe6 fxe6 Nf3 Bb4+ Nc3 Ba5 Bf4", "d4 Nc6 e4
$ opening_eco    <chr> "D10", "B00", "C20", "D02", "C41", "B27", "D00", "B00", "C50", "B
$ opening_name   <chr> "Slav Defense: Exchange Variation", "Nimzowitsch Defense: Kennedy
$ opening_ply    <dbl> 5, 4, 3, 3, 5, 4, 10, 5, 6, 4, 1, 9, 3, 2, 8, 7, 8, 8, 5
```

Section 3: Data analysis plan

This section is very preliminary and is subject to change. I am still learning a lot about this dataset and how to solve the questions I am asking, so it should be expected that this section will be changing quite significantly over time.

Does white have a first move advantage?

To answer this question, I will need to use the winner, white_rating, black_rating variables. It is important to note that the ratio between the white and black rating will likely play a factor in the outcome of the game as it would be hard for a player of low skill to defeat a skilled player regardless of any first move bias. To counter this, I will filter out any games where the ELO difference is outside of a +- 250 points boundary. I will also need to filter the victory type to only include games which result in a mate or a draw; if a game ends as a result of time running out or a resignation, then there is a good chance that the result could be impacted from external factors and could skew the outcome. I think it would also be beneficial to group the white team win/loss frequency by ELO rating as it will allow me to get a better insight into whether the skill level of the player will change the respective player's ability to utilize any start bias that may exist. To answer this question, I will be looking at the win/loss ratio of the white team grouped by ELO rating.

What openings have the highest and lowest average win rate grouped by player ELO rating?

This question is going to require the most work to solve. There are many factors which could have an impact on the outcome, so it will be important to consider any outliers. In order to ensure the integrity of the findings, I will need to filter out any of the non-rated games because players will typically play unrated when they want to try new openings that they don't know or they want to mess around, which would have an impact on the legitimacy of the outcome. Additionally, I will want to filter out any games that don't end in a mate or a draw. Lastly, I will filter out any games where the ELO difference is outside of a +- 250 points boundary. Anything outside of this range is going to lead to significantly unpredictable variability in prediction accuracy. Because there are many different openings, it would likely be best to select only the top 5 most played openings for each ELO group. The least played openings are going to have such a small ratio of the total game pool that the impact of a single game will be too significant. The win rate for each opening selected will be the win/loss ratio for the respective opening.