Kingdom

Andy Vo

Abstract— **Chess is a popular strategy-based board game between two players and is derived from the ancient Indian game, Chaturanga. It consists of 32 pieces, 16 pieces for each player: one king, one queen, two rooks, two knights, two bishops, and eight pawns. The board is a square chessboard with an eight-by-eight grid of 64 squares. The objective of the game is to "checkmate" the opponent's king, where the king is under attack and can not move any pieces for its escape. Organized competitive chess began in the 19th century, where the matches and the moves of each player were recorded. I collected PGN files and analyzed official chess matches among the top Chess masters throughout history using Python and PGN Mentor's PGN database to record data such as how many pieces they won, how many pieces they lost, the averages of the two former statistics against total matches, and the overall net result of pieces throughout their career.**

I. OBJECTIVES

The main objective was to analyze official chess matches among high-ranked chess players using Python and PGN Mentor's PGN database to gather data on statistics such as how many pieces they won and how many pieces they lost. Additional statistics I explored for the project include net gain of pieces, average moves before capturing a piece, and the averages of how many pieces they won and lost in comparison to the number of matches in their career. The question this project seeks to answer is what does the path to a top Chess player look like in the context of the Chess battlefield?

II. MOTIVATION

After the introduction of computers, the world of chess was thoroughly explored through many individuals studying chess theory. People were able to develop engines like Stockfish that can analyze any given moment in a game of chess to compute the next best set of possible moves for the player and the opponent's best response. Through these engines, many people were able to analyze important statistics that contribute to the progression of the professional scene such as openings, strategies, and sequences of moves. However, I was more interested in statistics that may hold no meaningful value, but satisfy the curiosity of the mind. On their journey to become Grandmaster in the realm of chess, how many pieces did they win, how many pieces did they lose, how many pieces did they survived with? We wanted to find the answer for many high-ranked chess players, so the goal is to uncover this information.

III. DISCUSSION OF DATA

Portable Game Notation (PGN) is a standard plain text format for chess games that contain information such as moves and the players' respective colors. This format can be seen in many chess databases, as such, we will be analyzing matches that are in this format. The data is ordered by the following structure: Event, Site, Date, Round, White, Black, Result, and Movetext. Event is the name of the tournament or match event. Site has the location of the match and is in City, Region COUNTRY format. Date is the starting date of the game and is in YYYY.MM.DD format. Round represents the specific round in terms of the organization of the event. For example, if the semi-finals of a chess tournament were the 29th match, the round would contain the number 29. White is the player who went first and represents the white side. Black is the player who went second and represents the black side. Result is the result of the match in the format: White's score, dash, and Black's score. It can also have a * to represent an ongoing match. Movetext records the actual sequence of moves in the chess match. The moves are recorded by the move number and are in Standard Algebraic Notation. SAN is the standard for recording chess moves and consists of the abbreviation for a piece, an x if a capture took place, and the name of the final square the piece moved to in two-character algebraic notation. There are also optional tags people can use for clarity's sake. An example of a PGN file can be seen in Figure 1:

A. Acquisition of Data

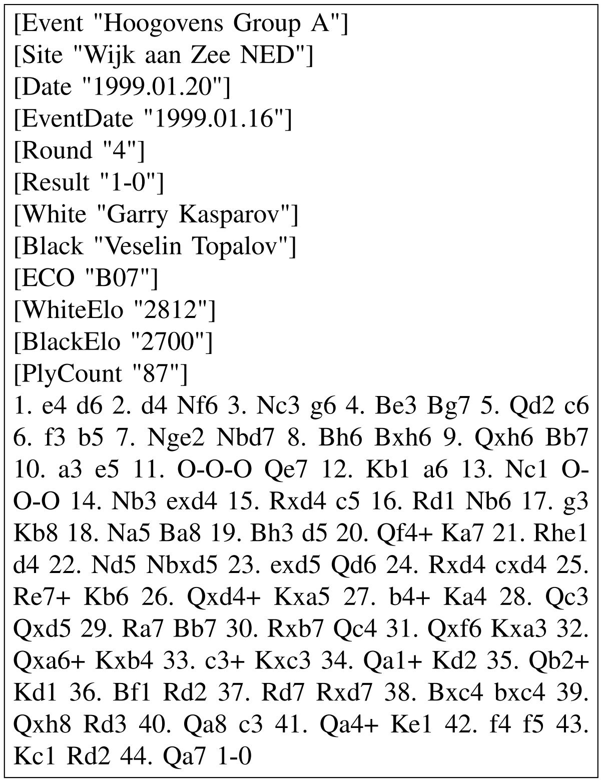
In order to obtain the data for this project, I searched the internet for various Chess PGN databases that contain the majority of the official matches for the top Chess players in history. After my research, I found PGN Mentor, a PGN viewer tool that also has downloadable PGN files of the top Chess masters in history. The files themselves contained over hundreds if not thousands of Chess games in PGN format. Based on a list from Chess.com, I was able to choose 19 Chess masters for my dataset and obtained 47,347 PGN files.

B. Cleaning and Parsing the Data

After collecting thousands of Chess games, I wanted to clean up the data that was unnecessary for the project. Specifically, I aimed to collect data on who played as white, who played as black, what their ratings were, the result of the match, and the moves list. As such, I extracted and parsed the input using a Python-based simple, open-sourced parser called PGN parser. Using this library and with the information cleanly processed, I collected the data I needed and organized the data into a class structure I built.

C. Integration of Data

After collecting the data, I created various algorithms to scan the data and extract the necessary information from



```
[Event "Hoogovens Group A"]
[Site "Wijk aan Zee NED"]
[Date "1999.01.20"]
[EventDate "1999.01.16"]
[Round "4"]
[Result "1-0"]
[White "Garry Kasparov"]
[Black "Veselin Topalov"]
[ECO "B07"]
[WhiteElo "2812"]
[BlackElo "2700"]
[PlyCount "87"]
1. e4 d6 2. d4 Nf6 3. Nc3 g6 4. Be3 Bg7 5. Qd2 c6
6. f3 b5 7. Nge2 Nbd7 8. Bh6 Bxh6 9. Qxh6 Bb7
10. a3 e5 11. O-O-O Qe7 12. Kb1 a6 13. Nc1 O-
O-O 14. Nb3 exd4 15. Rxd4 c5 16. Rd1 Nb6 17. g3
Kb8 18. Na5 Ba8 19. Bh3 d5 20. Qf4+ Ka7 21. Rhe1
d4 22. Nd5 Nbxd5 23. exd5 Qd6 24. Rxd4 cxd4 25.
Re7+ Kb6 26. Qxd4+ Kxa5 27. b4+ Ka4 28. Qc3
Qxd5 29. Ra7 Bb7 30. Rxb7 Qc4 31. Qxf6 Kxa3 32.
Qxa6+ Kxb4 33. c3+ Kxc3 34. Qa1+ Kd2 35. Qb2+
Kd1 36. Bf1 Rd2 37. Rd7 Rxd7 38. Bxc4 bxc4 39.
Qxh8 Rd3 40. Qa8 c3 41. Qa4+ Ke1 42. f4 f5 43.
Kc1 Rd2 44. Qa7 1-0
```

Fig. 1. Kasparov vs. Topalov, Wijk ann Zee 1999

them. For example, in order to collect the number of pieces won, I scanned through the moves list of each game, and develop a method to count the number of pieces won. Due to the simplicity of PGN format and the official Chess rules, I was able to determine that the first move in a turn will always be White's move, while the second move in a turn will always be Black's move. As such, once I determine the player's side, I looked for any x's that symbolized a piece has been captured. Inversely, I also measured any pieces lost in this algorithm, which is also dependent on what color the player I am currently analyzing is representing. Also in this algorithm, I recorded the number of moves there were before a piece has been captured on the researched player's side. In order to do so, I had a counter that counts the number of moves before an x has been scanned. After obtaining this data, I developed more algorithms to obtain the data concerning net gain of pieces and the averages of how many pieces they won and lost per game. Finally, I visualize the data into bar graphs to compare the data against each player.

D. *Validation of Data*

In order to validate the data, I cross-referenced the PGN games with Chess.com official database for games played by the top Chess players in history. Also, after obtaining a new data metric, I would print out the data and check the math for each one based on number of games played and moves there are.

IV. Responsibilities and Roles of Each Member

A. *Andy Vo*

- Found different chess databases to analyze the matches.
- Developed algorithms and data structures for the data.
- Created a visualization of results from the data.

All members were responsible for the final presentation and research paper.

V. Algorithms and Models

In order to parse the PGN files into readable data, I used a Python-based, open-sourced PGN parser to clean up the data I needed for the project. I parsed every PGN file before I analyzed the data. For the extraction of data I wanted, I created a simple algorithm that scans the entire moves list for the specific PGN and records the necessary information. Afterwards, I also used simple division and subtraction algorithms to calculate data metrics such as the net gain of chess pieces and the averages of how many pieces they won and lost.

VI. Results

Figures 2-7 visualize the data I collected for the project. Specifically, in order, they represent pieces won in a player's career, average pieces won based on the amount of games played, pieces lost, average pieces lost, average moves before capturing a piece, and overall piece net gain.

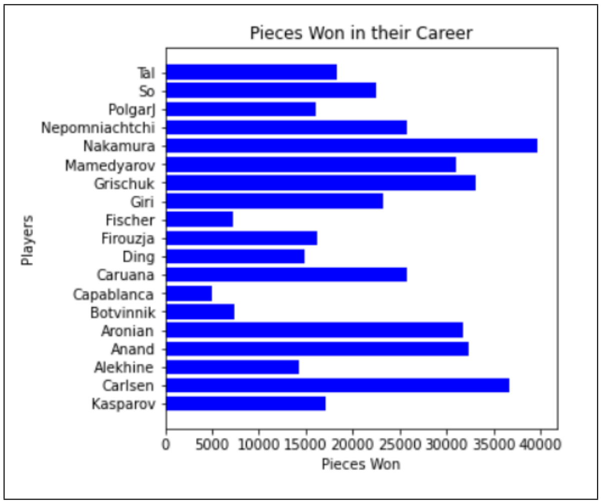


Fig. 2. Pieces Won in their Career

Figure 2 represents how many pieces the Chess players have won throughout their entire careers. Although this metric is skewed by time and matches played, it was interesting to see the results of different play styles.

Figure 3 represents the average number of pieces won in the Chess masters' careers. This metric met my expectations of being similar for each player due to the overall skill and intellect of these Chess masters.

Figure 4 represents how many pieces the Chess players have lost throughout their entire careers. This metric is also skewed by time and matches played, and the graph shows

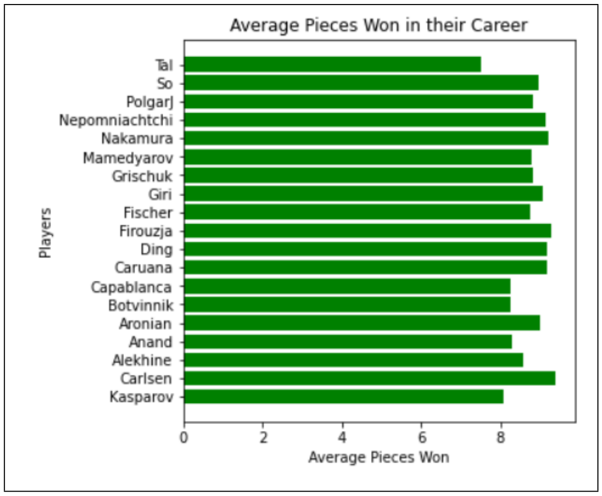


Fig. 3. Average Pieces Won in their Career

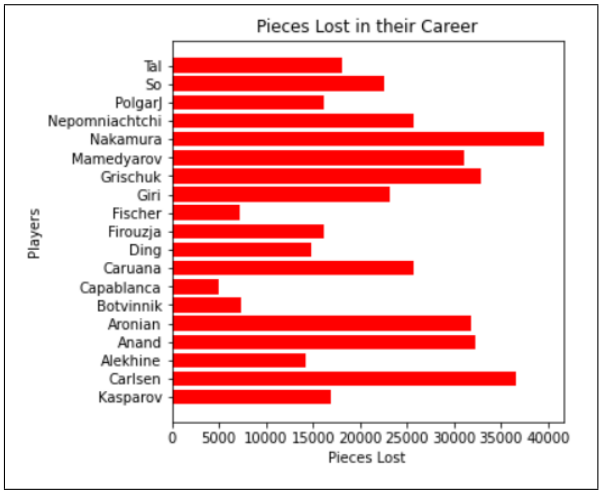


Fig. 4. Pieces Lost in their Career

that this data is parallel to the number of pieces won based on the overall shape of the graphs.

Figure 5 represents the average number of pieces lost in the Chess masters' careers. This metric also met my expectation of being similar due to their skills.

Figure 6 represents the average number of moves before capturing a piece. This metric seems to be heavily skewed by the amount of games played, but it does offer an insight on players who prefer to develop their board over making the first attack.

Figure 7 represents the overall piece net gain for the Chess masters. This metric would have been interesting to compare with their win/loss record to see if they can be correlated.

VII. Primary Issues

The main primary issue was finding a dataset that contained a large portion of the matches of top Chess players in history. While Chess.com has a database with downloadable PGN files for these masters, they had to be downloaded individually, which was unappealing to me to use as a dataset. Although, I did plan to use it if there were no other choices. Luckily, I found a dataset after weeks of research

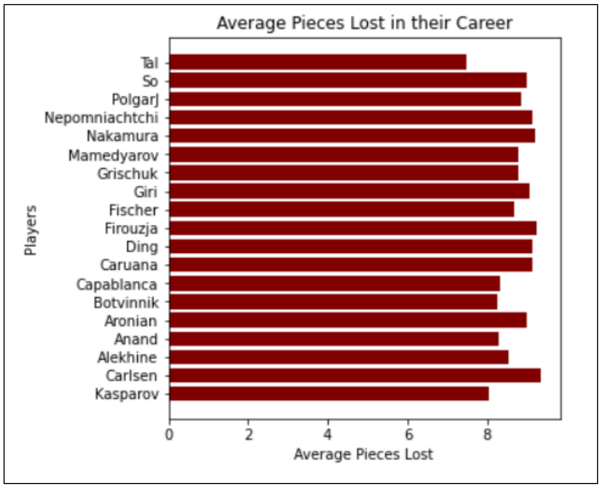


Fig. 5. Average Pieces Lost in their Career

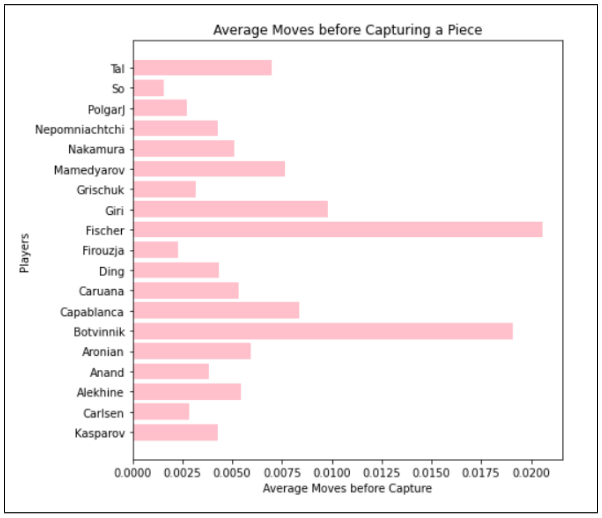


Fig. 6. Average Moves before Capture

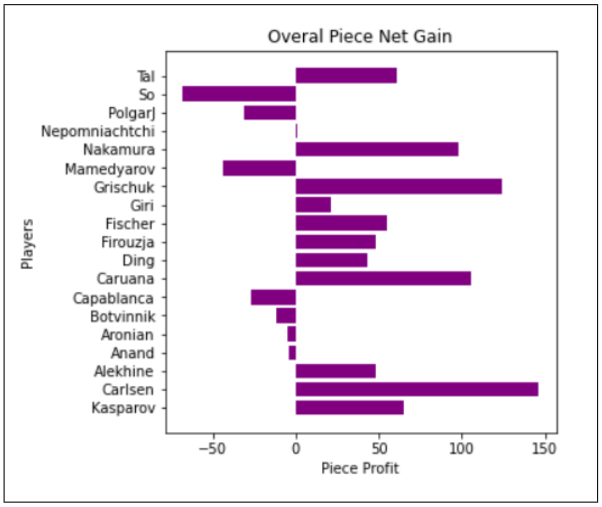


Fig. 7. Overall Piece Net Gain

that contained tons of PGN files for the games I intended to analyze. If I were to use the Chess.com database, then I would imagine scraping would have been a primary issue as well.

VIII. Future Work

For future improvements to this project, I would like to work with various Chess bots such as Stockfish to obtain more specific information such as the number of sacrifices used and favored openings for each player. I would also like to improve further on the data analysis to analyze more data metrics such as win rate and number of resigns.

IX. TIMELINE OF MILESTONES

The milestones for this project were to research chess databases, clean and parse the data, develop the code to analyze the data, and analyze and visualize the data. We also have some optional statistics that we wanted to explore, but ultimately, due to time constraints, they were not implemented.

We had around six weeks to complete our project. The following table was the timeline for our project:

TABLE I

Timeline of Milestones

Week	Milestones
Week 1	Research Chess Databases
Week 2	Clean and Parse the Data
Week 3	Develop Data Structures and Algorithms to Analyze the Data
Week 4	Optional Additional Statistical Research
Week 5	Analyze and Visualize the Data
Week 6	Final Report and Presentation

X. CONCLUSIONS

Overall, the data obtained and analyzed met my expectations in terms of how the players would compare to each other. Although variables such as time and games played skewed some of our results, the information seemed to match the overall assumptions of the data. Through this project, I was able to collect various statistics that can symbolize the path of a Grandmaster such as how many pieces they won, how many pieces they lost, and the overall net gain of pieces in their career. Once further built upon, this data should provide insight on how different playstyles can be reflected based on their data.