Chess Games Analysis

(COMP3125 Individual Project)

Dylan Patel

# Introduction

Chess is a timeless game of strategy and intellect, with millions of games played daily worldwide. Analyzing large datasets of chess games can reveal patterns in player behavior, popular strategies, and the impact of rating differences on game outcomes. This report utilizes the Lichess dataset from 2013 to explore three key questions:

1. Which openings are most commonly played in different rating ranges?
2. How often do players win via checkmate versus timeouts at various rating levels?
3. What is the average rating difference between opponents in the dataset?

Understanding these patterns can provide valuable insights into chess strategy and skill development across different player levels.

# Datasets

## Source of dataset

The dataset is sourced from Lichess.org, a popular online chess platform. Specifically, the PGN file contains games played in 2013 through January. Lichess datasets are widely used in the chess community for research due to their credibility and extensive game records, which include metadata such as player ratings, game outcomes, and chosen openings.

## Character of the datasets

The Lichess dataset consists of chess games stored in PGN (Portable Game Notation) format. Each game entry includes rich metadata describing the players, game details, and results. Below are the key characteristics:

* Format: PGN (Portable Game Notation).
* Game Metadata: Each game includes the following information:
  + Event: Specifies the type of game (e.g., "Rated Classical game").
  + Site: URL of the game on Lichess.org.
  + Result: Outcome of the game (e.g., "1-0" for White wins).
  + Player Ratings: Ratings of both players before the game (WhiteElo, BlackElo) and their respective rating changes (WhiteRatingDiff, BlackRatingDiff).
  + Opening: Name of the opening played, including variations and ECO (Encyclopaedia of Chess Openings) codes (e.g., "Queen's Pawn Game: Colle System, Anti-Colle", ECO: D04).
  + Time Control: Indicates the time control used (e.g., "480+2" for 8 minutes plus a 2-second increment per move).
  + Termination: Describes how the game ended (e.g., "Normal" for a typical conclusion).
  + Moves: A complete list of moves made during the game.
* Size and Scope:
  + The dataset includes thousands of games played in late 2012 and early 2013, representing a wide range of player ratings and skill levels.
* Preprocessing and Cleaning:
  + Metadata such as player ratings, opening names, and game terminations were extracted using Python regular expressions.
  + Games were grouped into predefined rating ranges for analysis.
  + Data cleaning excluded incomplete or corrupt entries to ensure consistency in results.

This structured dataset allows for detailed exploration of player behavior and strategy across varying skill levels.

# Methodology

#### **Question 1: Popular Openings by Rating Range**

* **Approach**: Extracted player ratings and openings from the PGN metadata. Games were grouped into rating ranges:
  + Rating <1200
  + Rating 1200–1400
  + Rating 1400–1600
  + Rating 1600–1800
  + Rating >1800
* **Analysis**: Counted occurrences of each opening in each range and identified the top 10 for visualization using bar plots.

#### **Question 2: Game Termination by Rating Range**

* **Approach**: Termination metadata was extracted to classify games as wins via checkmate or timeouts.
* **Analysis**: Calculated the percentage of each termination type within each rating range. Results were visualized using grouped bar charts.

#### **Question 3: Rating Difference Distribution**

* **Approach**: Extracted player ratings and computed absolute differences between opponents.
* **Analysis**: Calculated the average and median rating differences. Displayed the distribution using a histogram with overlaid average and median lines.

# Result

#### **Popular Openings by Rating Range**

The analysis revealed distinct preferences for openings across rating levels.

* Players below 1200: Van Kruijs Opening
* Players between 1200-1400: Van Kruijs Opening
* Players between 1400-1600: Van Kruijs Opening
* Players between 1600-1800: Owen defense
* Players above1800: Caro-kann defense

Bar plots showcased the top 10 openings for each rating range, highlighting variations in strategy across skill levels.

#### **Game Termination by Rating Range**

* **Checkmate** was the predominant termination type in games of all ratings.
* 70% of termination is due to checkmate in all ratings while 30% is due to time forfeit.

Grouped bar charts effectively illustrated these trends.

#### **Rating Difference Distribution**

* **Average Rating Difference: 154.17**
* **Median Rating Difference: 121.00**  
  The histogram showed most games occurred between players with relatively small rating differences, ensuring fair matchups.

# Discussion

While the analysis provided meaningful insights, some limitations were observed:

1. Dataset Scope: The dataset covers games only from 2013, limiting the generalizability to current trends.
2. Data Quality: Incomplete or noisy metadata in some games required exclusion.
3. Future Enhancements:
   * Expand the analysis to include more recent datasets.
   * Explore additional metrics, such as move accuracy or time usage.
   * Apply machine learning models to predict game outcomes based on openings and player ratings.

# Conclusion

This report analyzed chess games from the 2013 Lichess dataset, uncovering trends in opening preferences, game termination types, and rating differences. These findings contribute to understanding player behavior and strategy across skill levels. Future research can build upon these results to develop tools for player training and AI strategy enhancement.

##### Acknowledgment

Thanks to Lichess.org for providing access to their dataset and the chess community for their contributions to open research.

##### References

Lichess.org PGN Dataset, 2013.

.