KLE Society's
KLE Technological University

Exploratory Data Analysis

(22ECAC210)

Course Project Report on

"Chess Games Data"

Submitted in partial fulfilment of the requirement for the award of

Degree of Bachelor of Engineering
in
Computer Science and Engineering

Submitted By

Girish Badamikar – 01FE21BCS168
Rishi Hiremath – 01FE21BCS304
Karthik K Noolvi – 01FE21BCS163

Submitted To

SCHOOL OF COMPUTER SCIENCE & ENGINEERING,
KLE TECHNOLOGICAL UNIVERSITY
HUBLI – 580 031 (India).

Abstract

This abstract summarizes the key findings and insights obtained from conducting an Exploratory Data Analysis (EDA) on a chess games dataset. The dataset comprises various attributes related to chess games, such as player ratings, outcomes, time controls, and other game-specific information. The purpose of this EDA was to gain a comprehensive understanding of the dataset, identify patterns, and derive meaningful insights.

Table of Content:

List of tables

List of Figures

Chapter 1. Introduction

Overview of EDA Project:

This project involves conducting Exploratory Data Analysis (EDA) on a chess games dataset. The goal is to gain insights into player performance, game outcomes, and the impact of time controls on chess games. The analysis includes data preprocessing, descriptive statistics, and visualizations.

Importance of EDA in Data Analysis:

EDA is essential as it helps to understand data patterns, detect outliers, and explore relationships between variables. It provides a preliminary understanding of the dataset, enabling better decision-making and guiding further analyses.

Objectives of the Course Project:

- To find which is the most popular opening moves in chess games.
- Aim to provide valuable insights into the dynamics of chess games and potentially discover interesting relationships between various game attributes, player behavior, and game outcomes.
- To find the distribution of results based on the winner is black/white.
- To predict BlackElo ratings with the help of result and WhiteElo ratings

2. Data Collection

Dataset for Course Project:

The dataset used in the course project contains information on chess games, including player ratings, outcomes, time controls, and other game-specific details.

Data Collection:

The data was collected from kaggle website which had a large database consisting of all the informations of several chess games.

Data Pre-processing:

Data pre-processing involved handling missing values, converting data types, and addressing outliers to ensure data quality and consistency. Descriptive statistics and data visualizations were used to gain insights into the dataset.

The dataset consisted of 6256184 rows and 15 columns

After preprocessing it reduced to 6251515 rows and 15 columns

Attribute Information:

Table 2.1: Attribute Information

Event	Event: The name or title of the chess event or tournament in which the game was played. It provides information about the specific competition or occasion.
White	The player or participant who played as the White pieces in the chess game. It typically represents the name or identifier of the White player
Black	The player or participant who played as the Black pieces in the chess game. It typically represents the name or identifier of the Black player.
Result	The outcome or result of the chess game. It can have values such as "1-0" (White wins), "0-1" (Black wins), or "1/2-1/2" (draw).

UTC Date	The date of the chess game in Coordinated Universal Time (UTC) format. It provides the specific day on which the game was played.
UTC Time	The time of the chess game in Coordinated Universal Time (UTC) format. It represents the specific time when the game started.
WhiteElo	The Elo rating of the White player. Elo is a rating system used to gauge the skill level of chess players. It provides an estimation of a player's strength based on their performance in previous games.
BlackElo	The Elo rating of the Black player. Similar to WhiteElo, it represents the estimated skill level of the Black player.
WhiteRatingDiff	The difference in Elo rating for the White player compared to their previous rating. It indicates the change in skill level after the game.
BlackRatingDiff	The difference in Elo rating for the Black player compared to their previous rating. It represents the change in skill level after the game.
ECO	Encyclopedia of Chess Openings (ECO) code. It categorizes the opening moves played in the game, providing a standardized classification system for chess openings.
Opening	The specific name or description of the opening played in the chess game. It provides details about the initial moves made by both players at the beginning of the game.

TimeControl	The time control or time format used in the chess game. It specifies the time limits or regulations governing the pace of the game, such as "Blitz," "Rapid," or "Standard."
Termination	The reason or condition under which the chess game was terminated. It can indicate factors like checkmate, stalemate, resignation, time control violation, or other termination conditions
AN	Algebraic Notation (AN) is a standardized notation system used to represent chess moves in a human-readable format. It describes the moves made during the game, following the rules of algebraic chess notation.

3. Data Exploration

Descriptive Statistics:

Descriptive statistics provide an overview of the dataset. Here are some key metrics:

- Mean, median, and standard deviation of player ratings.
- Count and percentage of game outcomes (wins, draws, losses).
- Distribution of time control settings (e.g., rapid, blitz, classical).

Visualization:

The data is visualized through various plots, graphs, and charts, including:

Scatter plots to explore relationships between player ratings and rating differences.

Bar plots to display the distribution of game outcomes and time control settings.

Hat map to study the correlation between the attributes.

Identifying Patterns, Trends, and Outliers:

The visualizations help identify patterns, trends, and potential outliers in the data:

Patterns in player performance based on ratings and rating differences.

Trends in game outcomes, indicating changes in winning probabilities over time.

Outliers representing rare occurrences or extreme game results.

By combining descriptive statistics and data visualization, we gain valuable insights into player behavior, game dynamics, and the impact of time controls on chess games.

4. Data Cleaning

Addressing Missing Data:

Missing data points were identified and handled appropriately. Techniques such as imputation or removing rows with missing values were used to ensure data completeness.

Handling Outliers:

Outliers, if present, were identified using statistical methods. Depending on the nature of the outliers, they were either corrected, replaced with suitable values, or considered as valid data points if they were not errors.

Data Transformation and Normalization:

Data transformation techniques, such as logarithmic or power transformations, were applied to achieve a more normal distribution of skewed data. Normalization techniques like min-max scaling or z-score normalization were used to bring different attributes to a similar scale, ensuring fair comparisons during analysis.

5. Feature Selection

Dimensionality Reduction Techniques: Dimensionality reduction techniques, such as Principal Component Analysis (PCA) or feature selection methods like Recursive Feature Elimination (RFE), were applied to reduce the number of features while retaining meaningful information.

Rationale for Feature Selection: The rationale behind feature selection was to eliminate redundant or less informative features, reducing computational complexity, and enhancing model performance. Relevant features were chosen based on their impact on game outcomes and player performance

6. Data Analysis

Most popular openings:

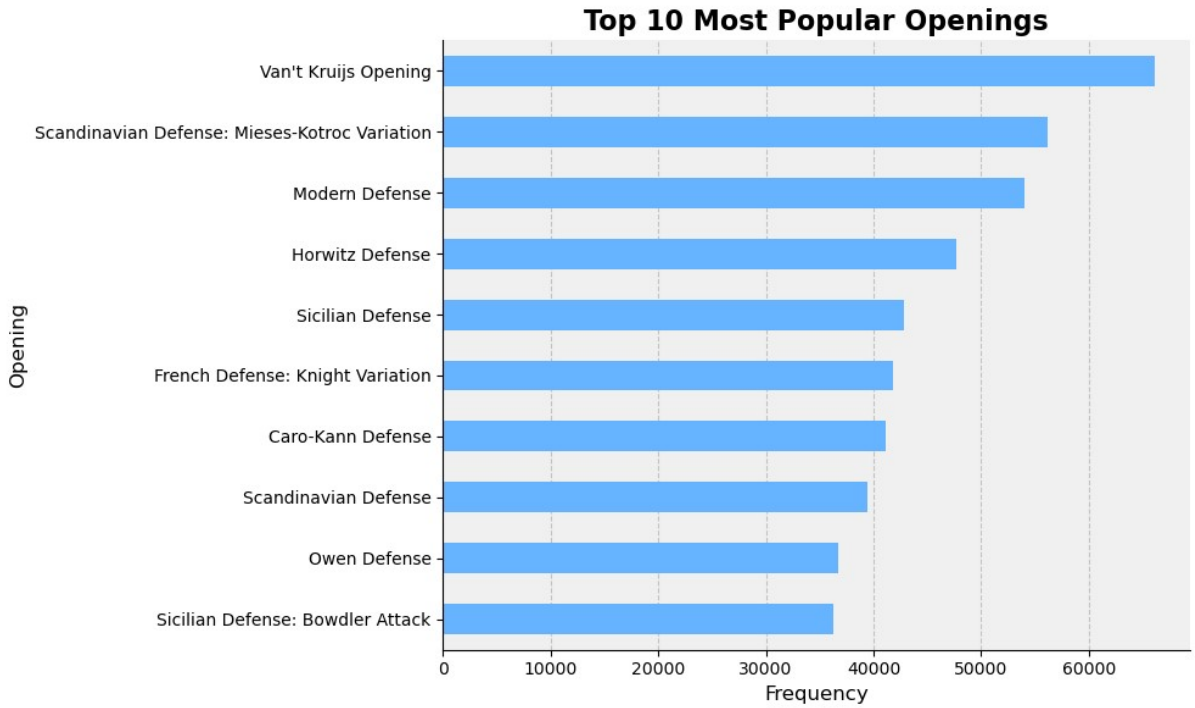

Fig 6.1: Popular Openings

Inference:

The most popular openings is Vant's Krujis Opening followed by the Scandinavian Defense. Most of the players wish to play these openings as they are the safest ones**.**

Distribution Of Results:

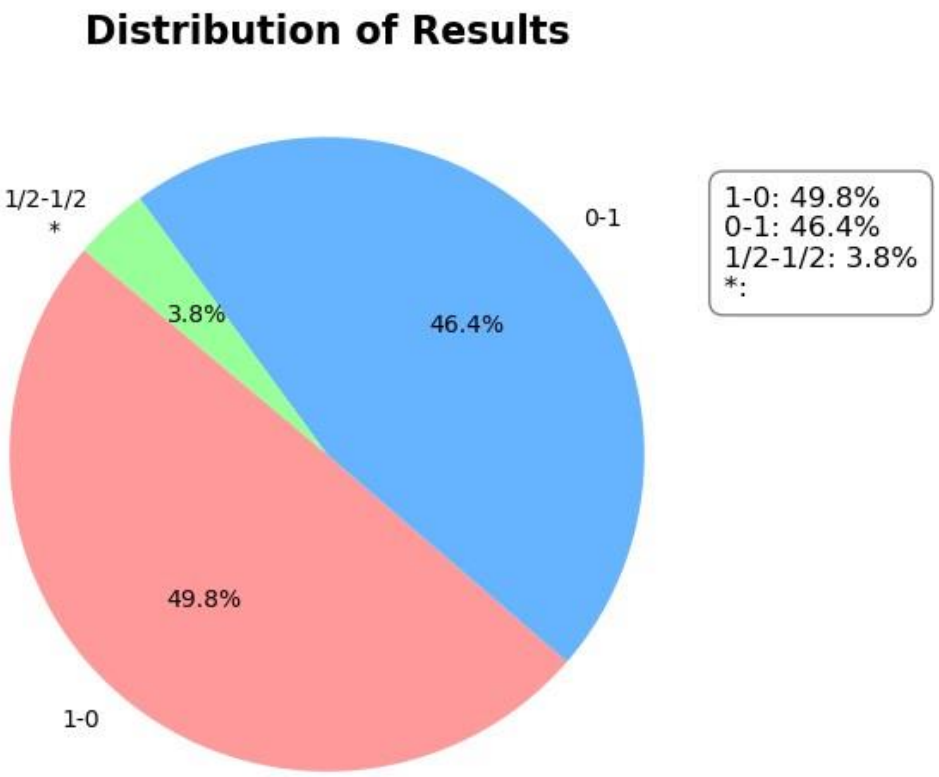

Fig 6.2: Distribution of Results

Inference:

We see that nearly 49.8% of matches are won by white, 46.4% by black and the remaining 3.8% matches ended in a draw

We can say that there is a very slight chance of white winning the match due to the advantage that white plays first.

Disrtibution of White Elo Ratings:

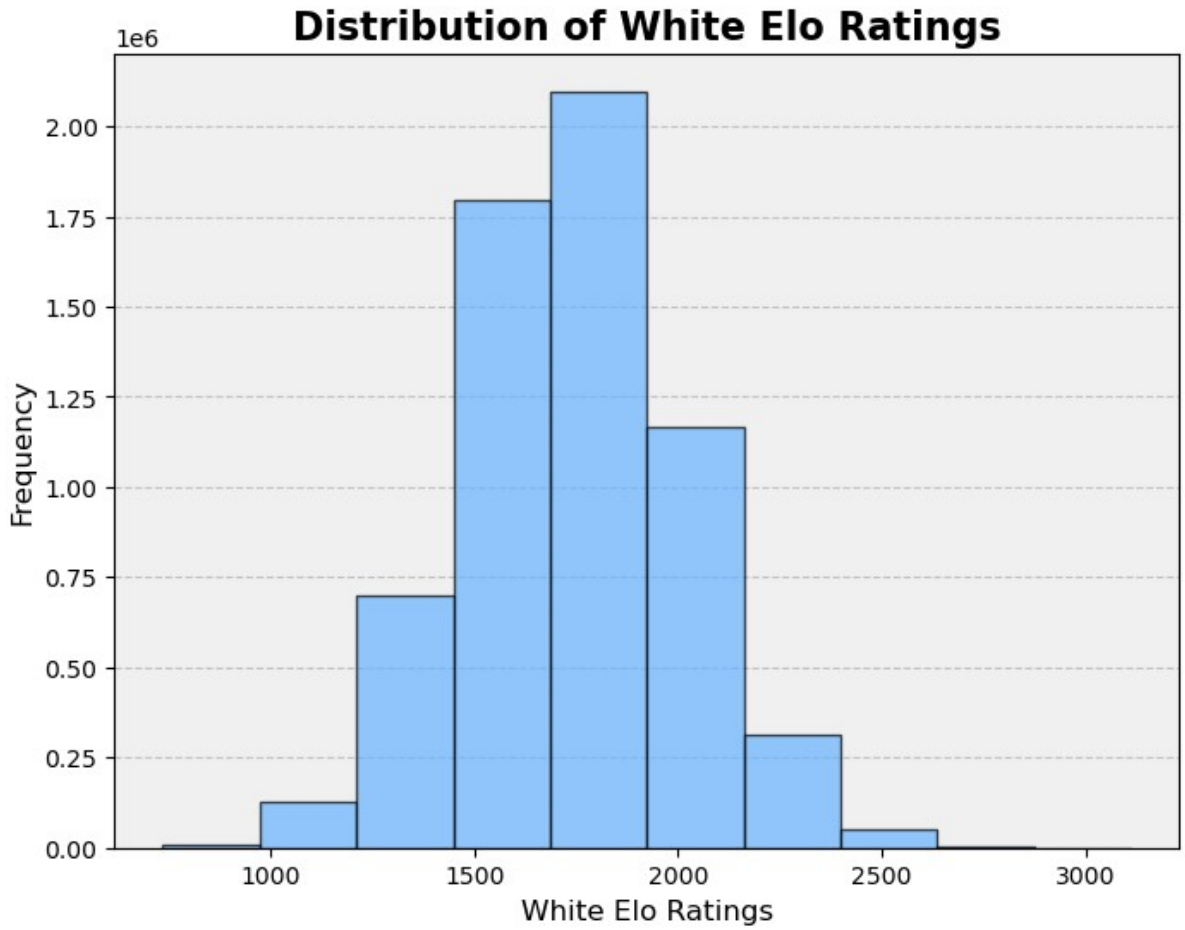

Fig 6.3: White Elo Ratings

Inference:

We can observe from the above graph that most of the players have a white Elo ratings in the range of 1500-2000. A rating above 2000 is considered to be very good for a player. We see that there are very few players with ratings of above 2000.

Rating difference vs Player Rating

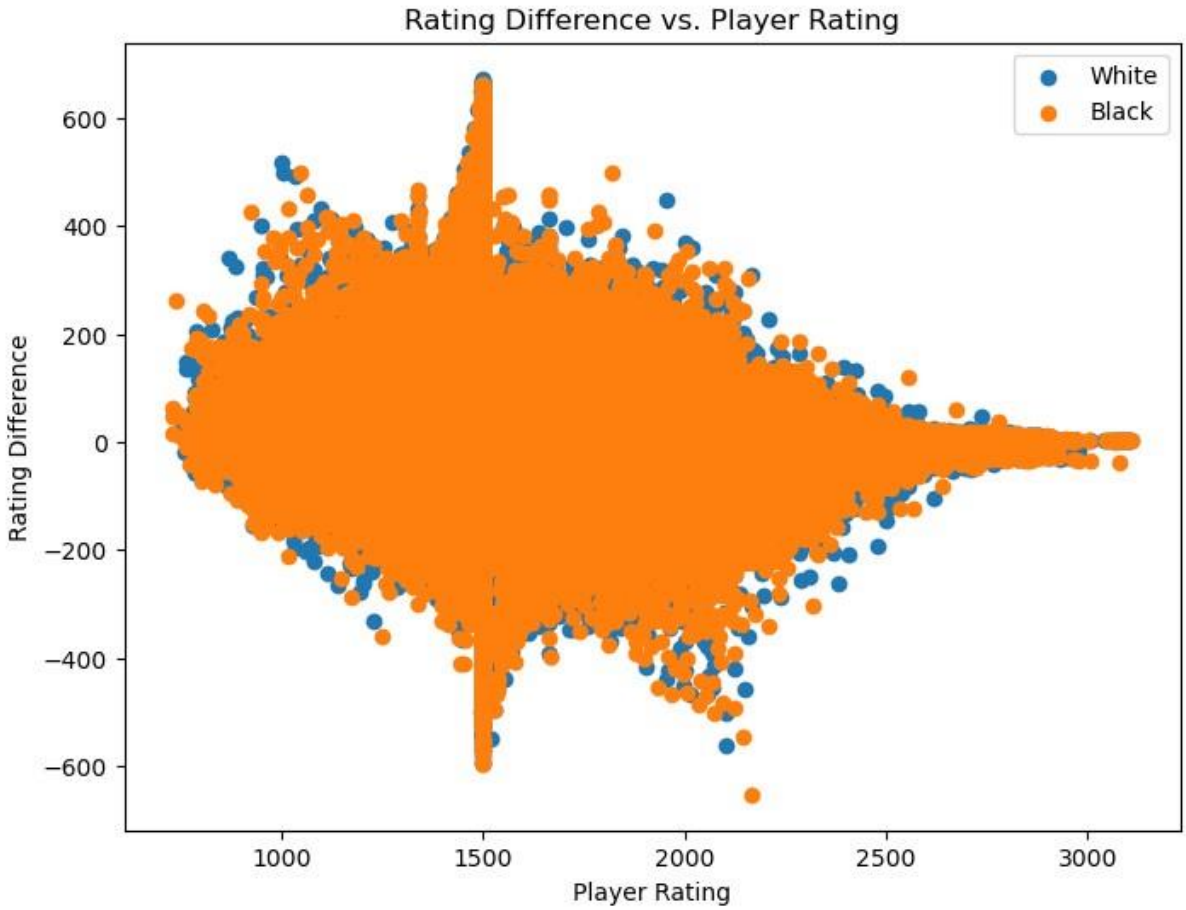

Fig 6.4: Rating difference vs Player rating

Inference:

We see that for both white and black the trend is similar. We see that rating differences has a huge variation for a player rating of around 1500. As we go further with player ratings, we see a decrease in the rating differences which means that here the chances of the result being a draw is more as the rating difference values are nearer to zero.

Event Frequencies

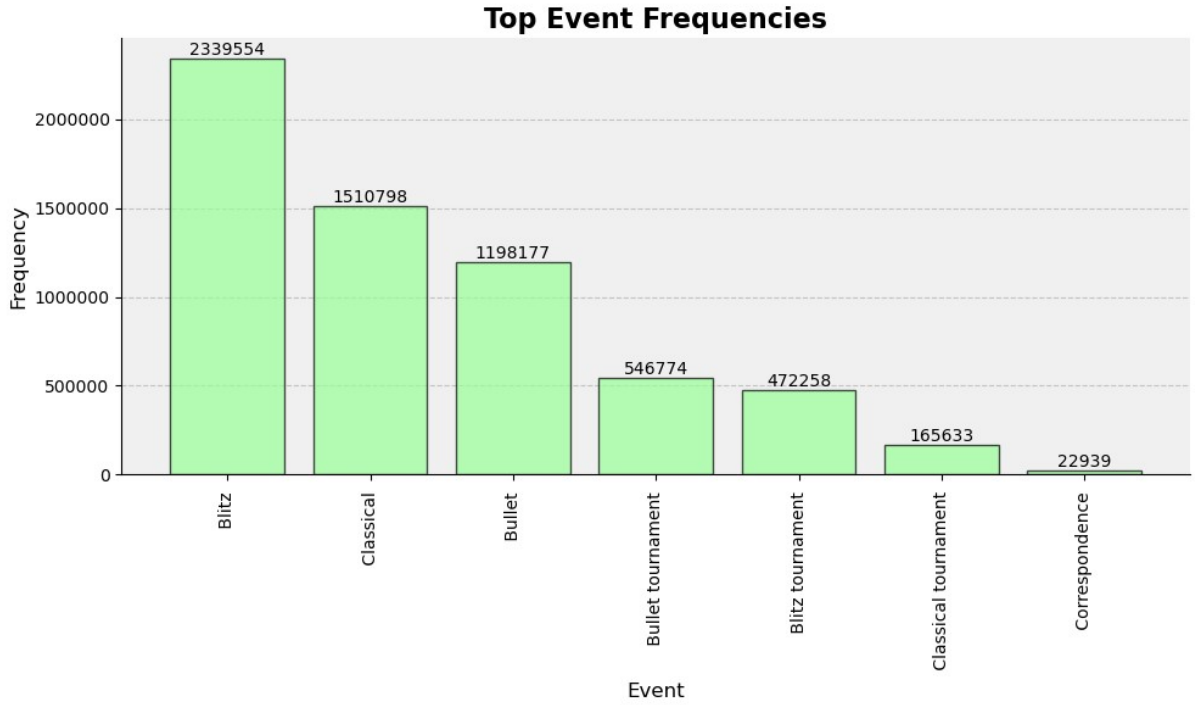

Fig 6.5: Event Frequencies

Inference:

We observe that Blitz(10min or less) games are the most played followed by classical(90 min) and Bullet(3min or less).

Correlation between attributes:

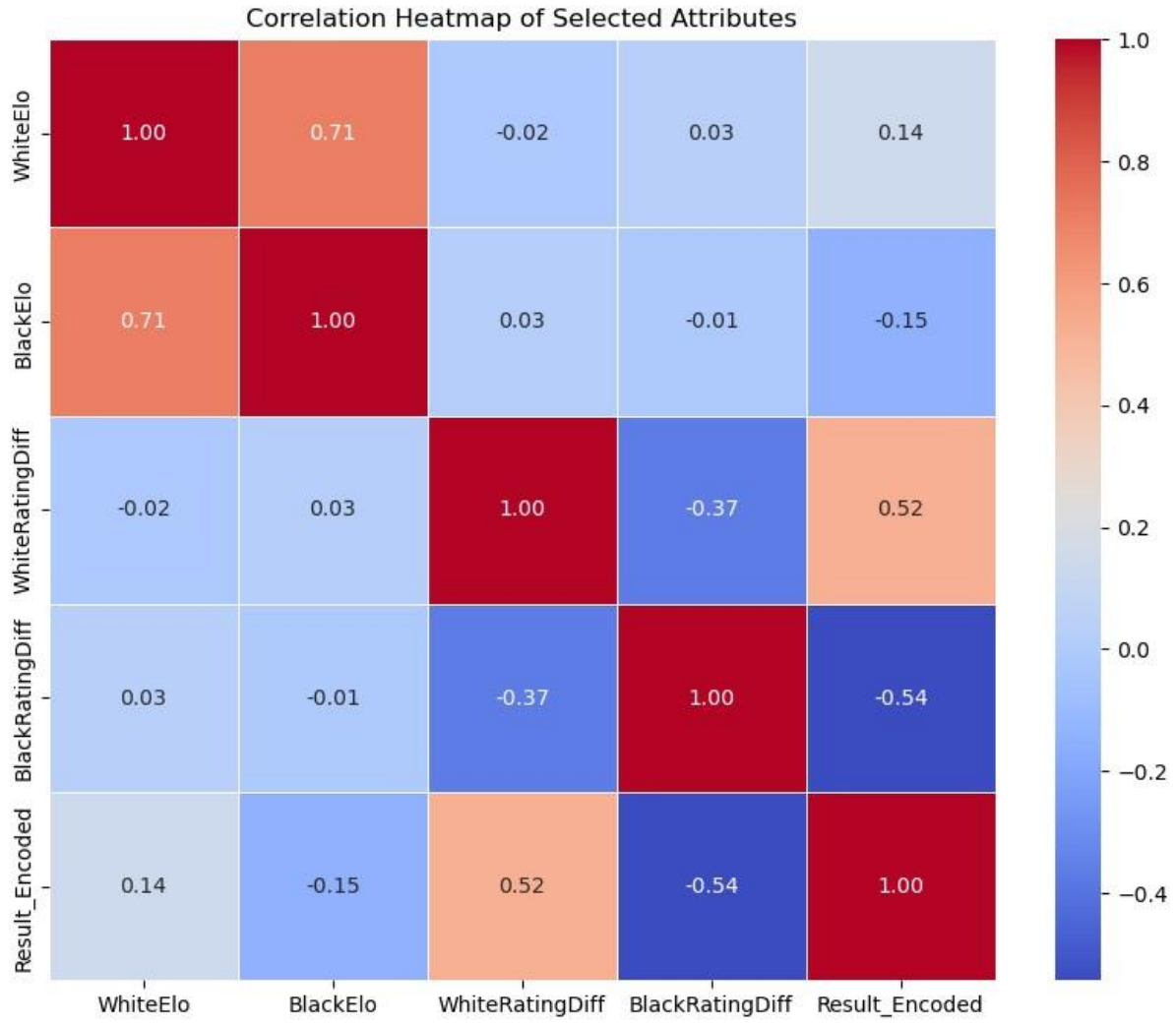

Fig 6.5: Correlation between Attributes

Inference:

We observe that there is a positive correlation between result(result is assumed to be 1-0) and white rating diff and also between white elo and result. This is because when white wins his rating increases and the black's rating decreases.
We see a negative correlation between the result and BlackElo and BlackRatingDiff.
The values nearer to zero indicate that there is no correlation between those attributes.

Prediction of Blackelo using Linear Regression:

Inference:

We got a value of 0.56 for r2 which is okay for prediction
We predicted the value of Blackelo by giving White elo and ratings as input

Clustering of WhiteElo , BlackElo, WhiteRatingDiff and BlackRatingDiff:

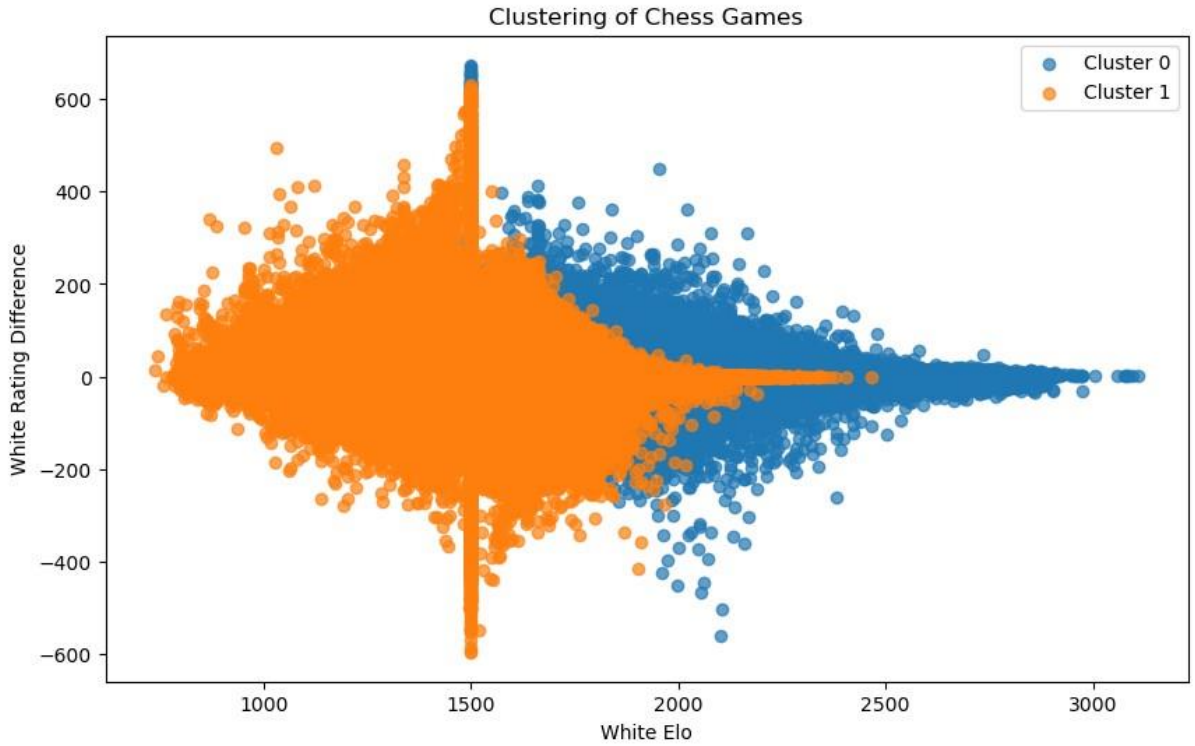

Fig 6.6: Clustering

Inference:

Cluster 0 represents a group of chess games where White players have higher Elo ratings ('WhiteElo'). This could indicate games involving more experienced or skilled players. The horizontal axis in the scatter plot represents the 'WhiteElo' values, which indicates the Elo rating of the White player.
If Cluster 0 is skewed towards the right side of the plot, it means that the majority of games in this cluster have higher 'WhiteElo' values (higher-rated White players).
As 'WhiteElo' is often considered a measure of player skill, higher values suggest players with more experience or better performance.

7. Insights and Findings

Key Insights:

- Popular Openings: Vant’s Krujis Opening and the Scandinavian Defense are the most favored openings among players.
- Match Outcomes: White and Black both win approximately 46.4% of games, with draws accounting for 3.8%. White's slight advantage could be due to the first-move advantage.
- Player Elo Ratings: Most players have Elo ratings between 1500 and 2000, while fewer players possess ratings above 2000, indicating a concentration of players in the intermediate skill range.
- Rating Differences: Rating differences show substantial variation around 1500 player rating. As player ratings increase, rating differences decrease, leading to a higher likelihood of draws.
- Game Types: Blitz games (10 minutes or less) are the most played, followed by classical games (90 minutes) and Bullet games (3 minutes or less).
- Significant Findings and Correlations:
- Positive Correlations: There is a positive correlation between the match result and White Elo as well as White Rating Diff, indicating that White's victory results in increased Elo rating and a favorable rating difference.
- Weak Negative Correlations: The match result shows weak negative correlations with Black Elo and Black Rating Diff, suggesting a less pronounced impact on Black's Elo and rating difference.

Limitations and Challenges:

- Incomplete Data: The analysis could be affected by missing or incomplete data, potentially leading to biased results.
- Simplified Models: The predictive model's moderate accuracy (R-squared value of 0.56) indicates that the current model may not fully capture all factors influencing match outcomes, necessitating more sophisticated models for precise predictions.
- Selection Bias: The data might have selection bias, as it represents only specific games available in the dataset, limiting the generalizability of findings to all chess games.
- Causal Inferences: The analysis is limited to identifying correlations and associations between attributes but does not establish causality.

8. Recommendations

Actionable Recommendations:

- Opening Strategy Education: Organize training sessions or resources that focus on teaching players effective strategies for Vant's Krujis Opening and the Scandinavian Defense, as they are the most popular openings. This can help players make informed decisions at the beginning of the game, potentially improving their performance.
- Elo Rating Ranges: Introduce skill-based tournaments or matchmaking systems that group players based on Elo rating ranges, such as 1500-2000 and above 2000. This can create a more balanced and competitive playing environment, providing players with challenging opponents and enhancing their overall gaming experience.
- Game Balance: Implement mechanisms to address the slight advantage of the first-move advantage for White players. This could include adjusting time control, modifying the scoring system, or introducing innovative game rules to balance the match outcomes between White and Black players.

Potential Areas for Further Analysis or Investigation:

- Opening Performance Analysis: Explore how specific players perform with different openings and identify correlations between opening choices and match outcomes. This analysis could provide insights into player strengths and weaknesses and inform personalized training approaches.
- Time Control Analysis: Investigate how different time controls impact match results and player satisfaction. Understanding players' preferences for different time controls could help tailor chess event formats to accommodate a diverse player base.
- Game Length and Elo Rating: Examine the relationship between game length and Elo rating differences to determine if certain time controls favor specific player skill levels. This analysis could lead to optimizing time controls for better fairness and player engagement.

Potential Impact of Implementing the Recommendations:

- Enhanced Player Engagement: By offering training resources on popular openings and implementing skill-based tournaments, players are likely to be more engaged and motivated to improve their gameplay, leading to a more active and vibrant chess community.
- Improved Competitive Balance: Balancing match outcomes between White and Black players could make chess matches more exciting and competitive, attracting a broader audience and fostering a more inclusive chess environment.

- Data-Driven Decision Making: Utilizing data analysis for personalized recommendations and matchmaking can optimize player experiences, leading to higher player retention and satisfaction.
- Fostering Skill Development: Implementing Elo-based matchmaking systems can help players face challenging opponents at their skill level, providing opportunities for skill development and growth in the chess community.

9. Conclusion

- **Popular Openings**: Vant’s Krujis Opening and the Scandinavian Defense are the most popular openings among players.
- **Match Outcomes**: White wins approximately 46.4% of matches, Black wins 46.4%, and the remaining 3.8% end in a draw. White's slight advantage could be due to playing first.
- **White Elo Ratings**: Most players have White Elo ratings in the range of 1500-2000. There are fewer players with ratings above 2000, which is considered good for a player.
- **Rating Differences**: Rating differences show significant variation around 1500 player rating. As player ratings increase, rating differences decrease, indicating a higher likelihood of draws.
- **Game Types**: Blitz games (10 minutes or less) are the most played, followed by classical games (90 minutes) and Bullet games (3 minutes or less).
- **Correlations**: There is a positive correlation between the result and White Elo and White Rating Diff. As White wins, their rating increases and Black's rating decreases. A negative correlation is observed between the result and Black Elo and Black Rating Diff, but the values are closer to zero, indicating weaker correlations.
- **Prediction**: A predictive model with an R-squared value of 0.56 indicates a moderate level of prediction accuracy for predicting Black Elo using White Elo and ratings as input.
- **Cluster 0**: Cluster 0 represents games with higher-rated White players, potentially involving more experienced or skilled players.

These findings provide valuable insights for players to strategize and improve their gameplay. However, further analysis and modeling may be needed for more accurate predictions or deeper exploration.

10. References

1. https://www.kaggle.com/code/sumeetpachauri/dm-chess-data

2. https://www.digitalocean.com/community/tutorials/exploratory-data-analysis-python
3. https://www.journalofexpertise.org/articles/volume6_issue1/JoE_6_1_Brancaccio_Gobet.pdf
4. https://www.kaggle.com/code/adityajha1504/eda-history-of-chess/notebook