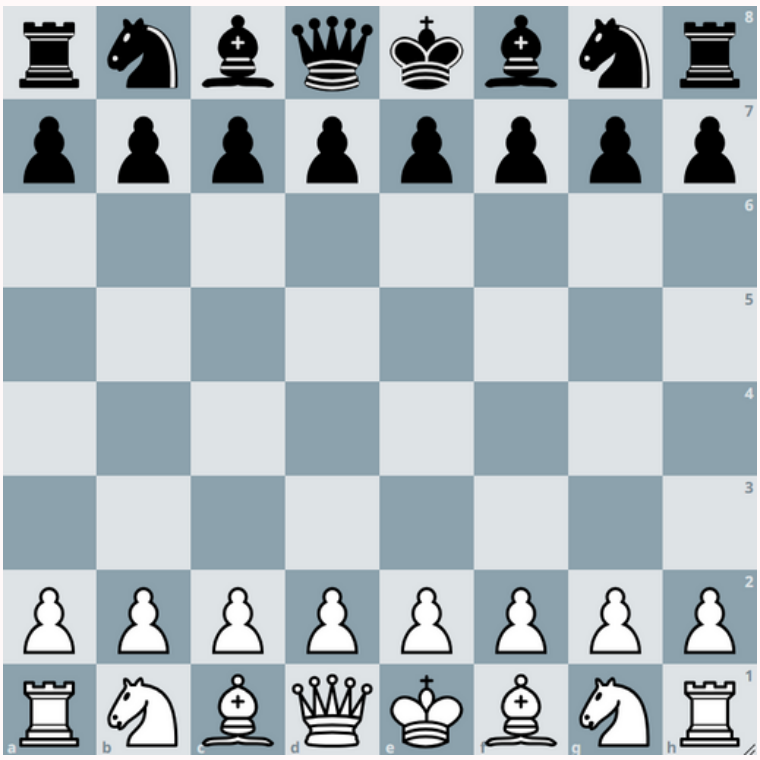
CHESS DATA ANALYSIS



UĞUR GÜNAL

MOTIVATION

As an chess addict, In my pursuit of chess mastery, I've unknowingly amassed a wealth of data through nearly 20,000 online games. Motivated by a desire to understand my strengths, weaknesses, and the determining factors behind each game's outcome, I've embarked on a journey of self-discovery.

This substantial dataset serves as a mirror, reflecting my chess journey and providing an opportunity to uncover patterns that may shape my playing style. I'm particularly intrigued by the influence of time control on my performance—does it enhance my strategic thinking or introduce vulnerabilities that need addressing?

Furthermore, I aim to dissect my performance across different board positions. Pinpointing areas where success is elusive will guide targeted improvements and contribute to a more versatile playing style.

This exploration is an open-ended adventure; I expect to unearth insights that may have previously gone unnoticed. The goal is not just statistical analysis but a continual quest to enhance strategic thinking and refine tactical skills in the ever-evolving landscape of chess.

MY DATASET

My chess data set is taken from the online chess platform “Lichess”

FORMAT OF DATASET

Initially, the dataset was in PGN format, which is not favorable for constructing a data frame. Therefore, it is parsed to obtain a usable format.

PGN DATA:

```
[Event "Rated Blitz game"]
[Site "https://lichess.org/bHdUS35O"]
[Date "2023.12.09"]
[White "Alekin2"]
[Black "ugurjoe"]
[Result "0-1"]
[UTCDate "2023.12.09"]
[UTCTime "21:05:10"]
[WhiteElo "2316"]
[BlackElo "2242"]
[WhiteRatingDiff "-7"]
[BlackRatingDiff "+7"]
[Variant "Standard"]
[TimeControl "180+0"]
[ECO "A08"]
[Termination "Time forfeit"]

1. Nf3 d5 2. d3 Nf6 3. g3 e6 4. Bg2 Be7 5. O-O O-O 6. Nb
15. f3 Bf5 16. Rxe5 Bxd3 17. Ne6 Qd6 18. f4 Nxe6 19. cxd
```

CLEARING DATA

Since some of the information in my dataset is unnecessary, I have removed them.

such as: *game link, player names...*

I have also added some extra information for further analysis.

such as: *number of moves, selected player's color(that's me in this case)*

DATA FRAME:

Event	Date	PlayerColor	PlayerRating	OpponentRating	RatingDifference	Result	ECO	Termination	Moves	NumberOfMoves
Rated Blitz game	2023.12.09	Black	2242	2316	-74	1.0	A08	Time forfeit	[g1f3, d7d5, d2d3, g8f6, g2g3, e7e6, f1g2, f8e...	34
Rated Blitz game	2023.12.09	Black	2235	2322	-87	1.0	C65	Normal	[e2e4, e7e5, g1f3, b8c6, f1b5, g8f6, d2d3, f8c...	17
Rated Blitz game	2023.12.09	White	2241	2228	13	0.0	A54	Time forfeit	[d2d4, g8f6, c2c4, d7d6, b1c3, e7e5, d4e5, d6e...	32
Rated Blitz game	2023.12.08	Black	2246	2277	-31	0.0	D02	Normal	[d2d4, d7d5, g1f3, g8f6, c2c4, e7e6, g2g3, c7c...	36
Rated Blitz game	2023.12.06	White	2240	2232	8	1.0	E91	Time forfeit	[d2d4, g8f6, c2c4, g7g6, b1c3, f8g7, e2e4, d7d...	52

EXPLORATORY DATA ANALYSIS

After filtering some of the games with missing information
there are *17,293* games left

```
df.shape
```

```
(17293, 11)
```

There are 11 columns in my dataframe

```
df.dtypes
```

```
Event                object
Date                 object
PlayerColor          object
PlayerRating          int64
OpponentRating        int64
RatingDifference      int64
Result              float64
ECO                  object
Termination          object
Moves                object
NumberOfMoves         int64
dtype: object
```

COLUMNS

Before the analysis, I will explain some of the columns since their names are not that descriptive

EVENT

Time format of the games in this case there are only 2 format: blitz and rapid.

Blitz: Time Control that 2 sides have 3 minutes.

Rapid: Time Control that 2 sides have 10 minutes.

ECO

Eco is abbreviation FOR *Encyclopedia of chess openings*. It is a Code that shows which opening is played in the game.

TERMINATION

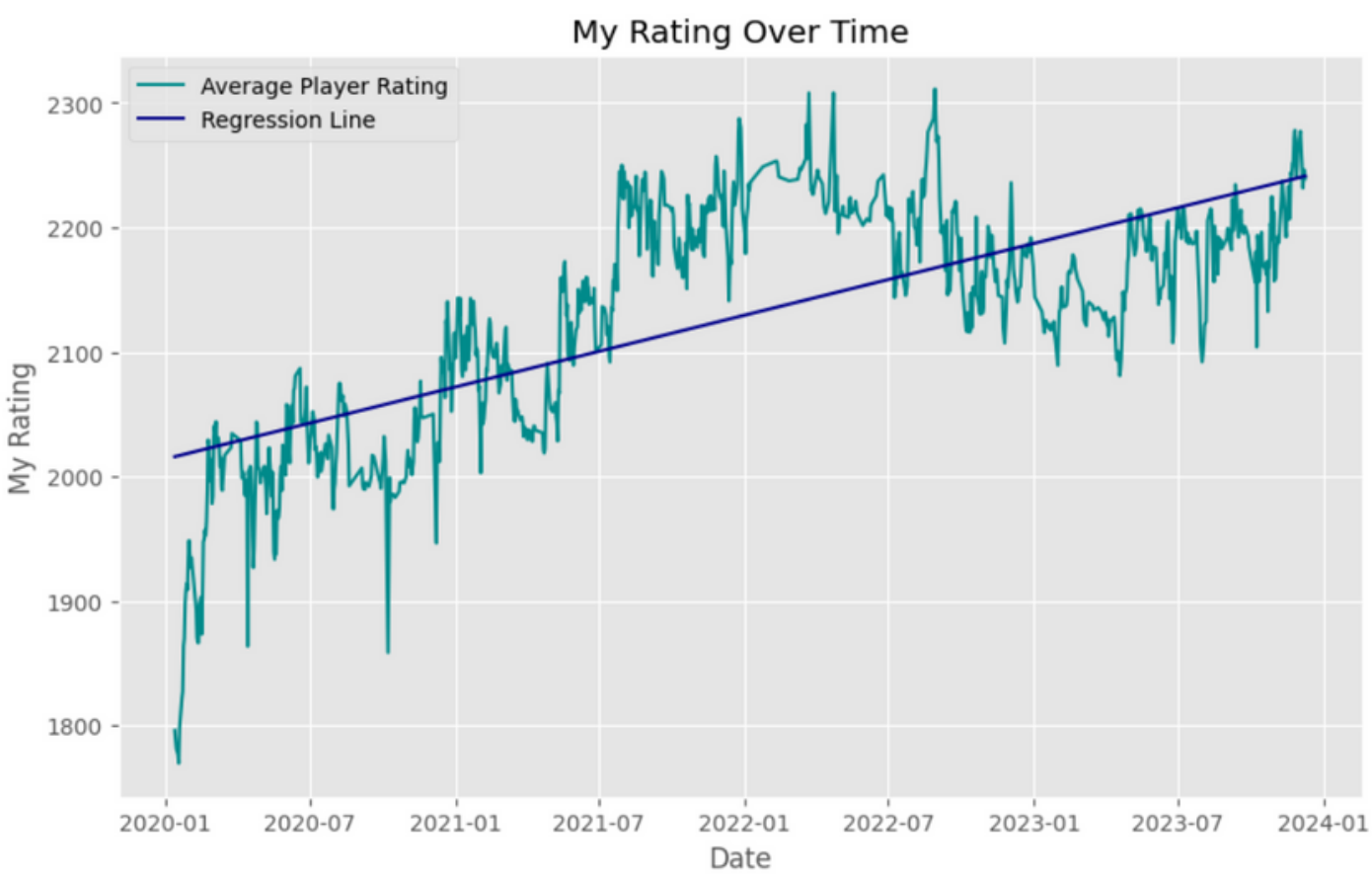
Shows the reason for the game ending. In this case, there are only 2 types: normal and time forfeit.

Time Forfeit: If the time for one of the sides has expired.

Normal: Any other scenario.

RATING OVER TIME

First of all, I've evaluated my development from 2020 to 2024, a period in which I regularly engaged in online chess



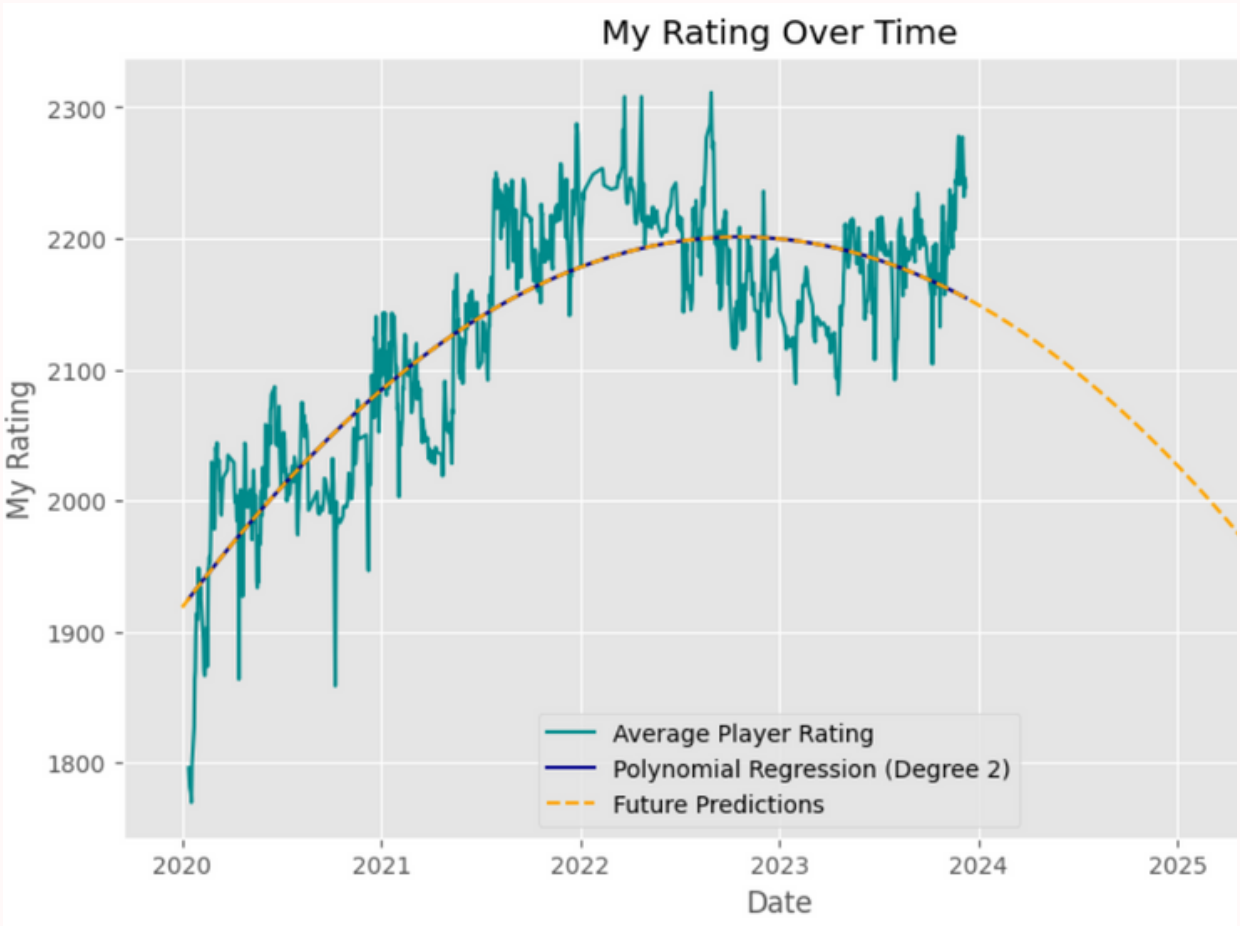
Also I have plotted a regression line to see average Trend of my rating over time.

```
The slope of the regression line is: 0.15801375375892915
```

On average, I have gained NEARLY **0.16 rating** per day according to linear regression

RATING OVER TIME

While linear regression is a valuable tool for predictions, it is not the only approach. There are also other machine learning algorithms, such as polynomial regression,

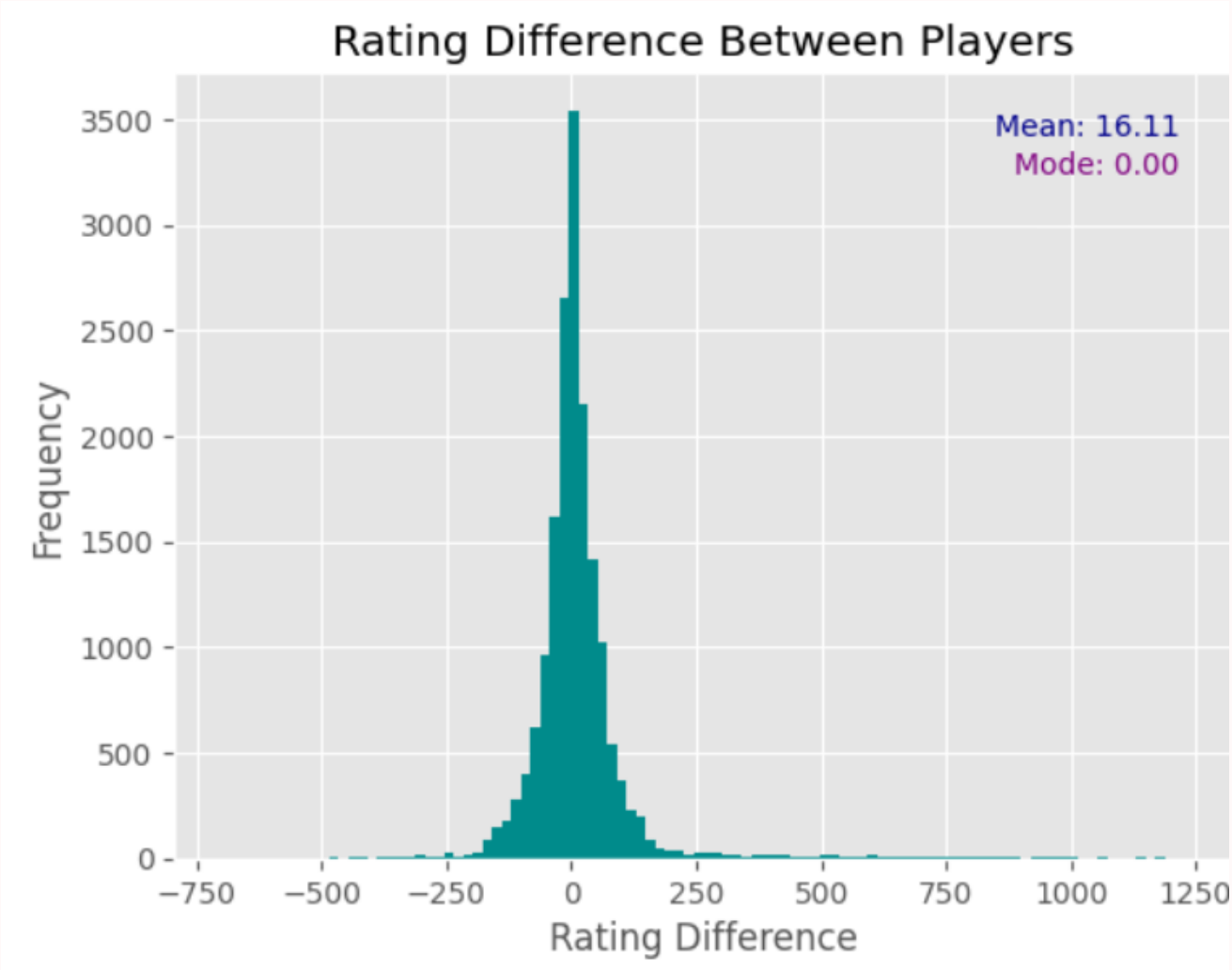


I have implemented a Polynomial Regression model to analyze the same graph over time.

The results were rather pessimistic; I am inclined to continue trusting linear regression

RATING DIFFERENCE

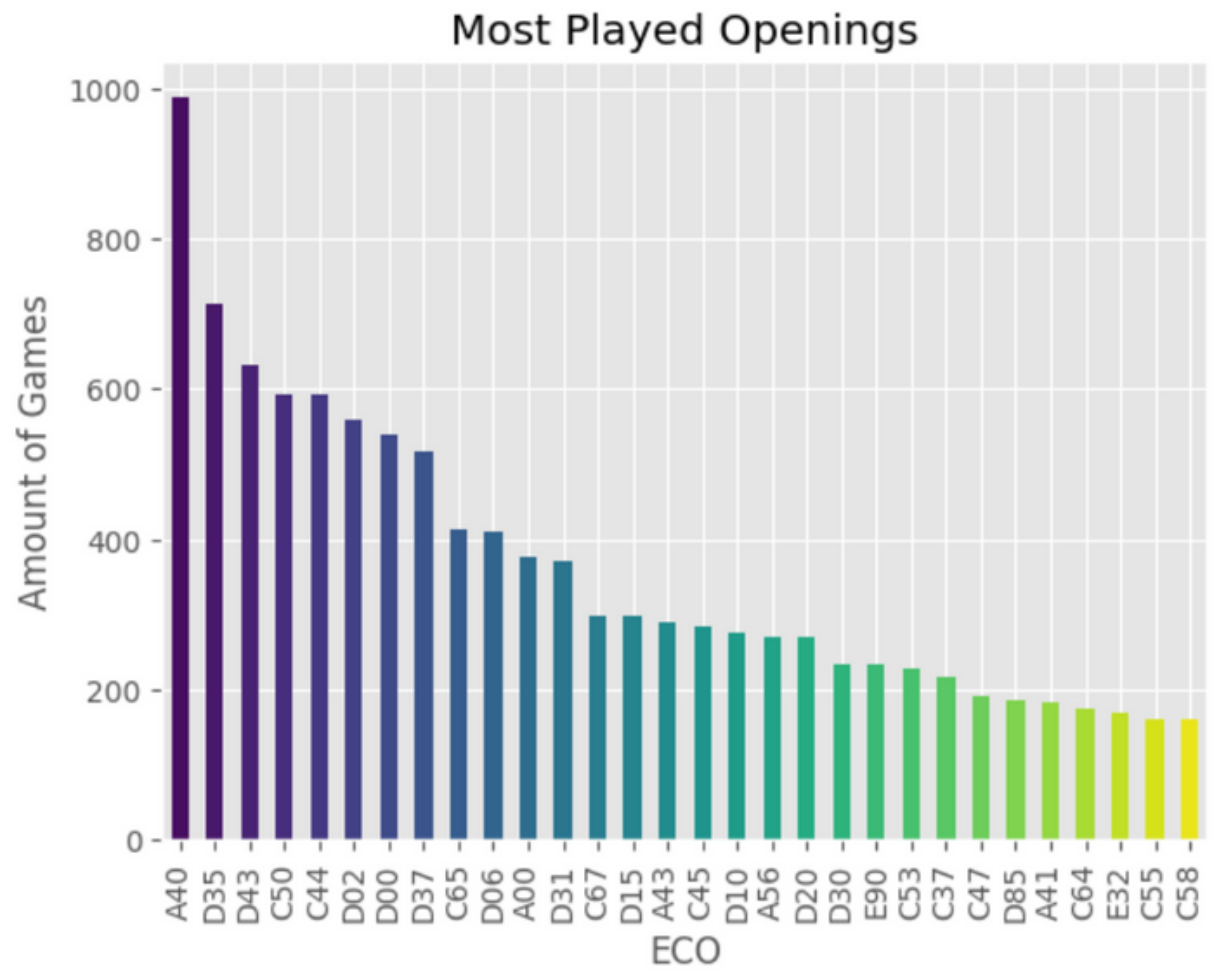
LETS LOOK IF THERE THERE are HUGE DIFFERENCEs BETWEEN PLAYERS RATINGs



As anticipated, the **mode is 0**. This aligns with expectations, considering that many online chess platforms prioritize facilitating competitive games, often resulting in matchups where players have similar skill levels.

OPENING ANALYSIS

As previously mentioned, each opening is uniquely identified by its ECO code. Let's examine the distribution of openings.

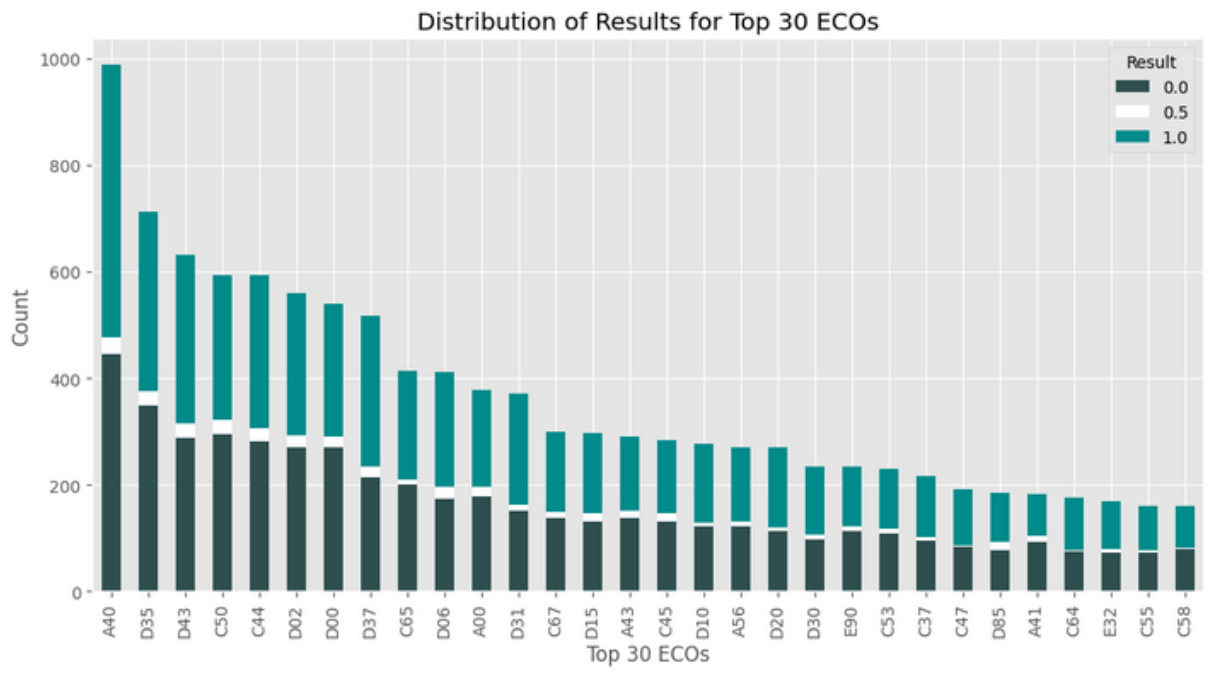


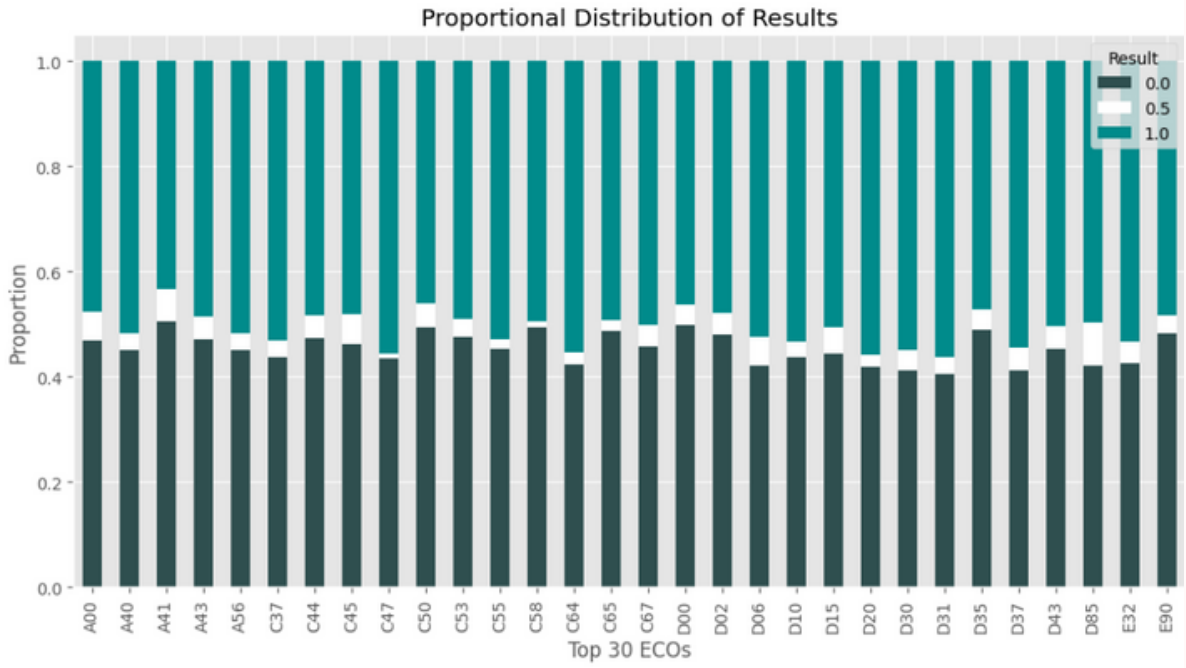

```
There are 285 unique ECOs in the DataFrame.

One of the top 30 ECOs played in the 62.64% of games.
```

For a more focused and representative examination, I have limited the opening analysis to the **top 30 ECO** codes, which constitute the majority of the games.

OPENING ANALYSIS





Despite varying game counts for each opening, there is a consistent pattern in the distribution of results (win, draw, loss). This uniformity suggests that the number of games played in a particular opening does not exert a substantial influence on the outcome.

POSITION ANALYSIS

Despite the limited impact of openings on game results, a more specific analysis can be conducted by examining the most common positions.

Here are the 4 most common positions in
Move Count 10

Total Games Played: 122
Success Rate: 53.28%
Win Rate: 46.72%
Lose Rate: 40.16%
Draw Rate: 13.11%

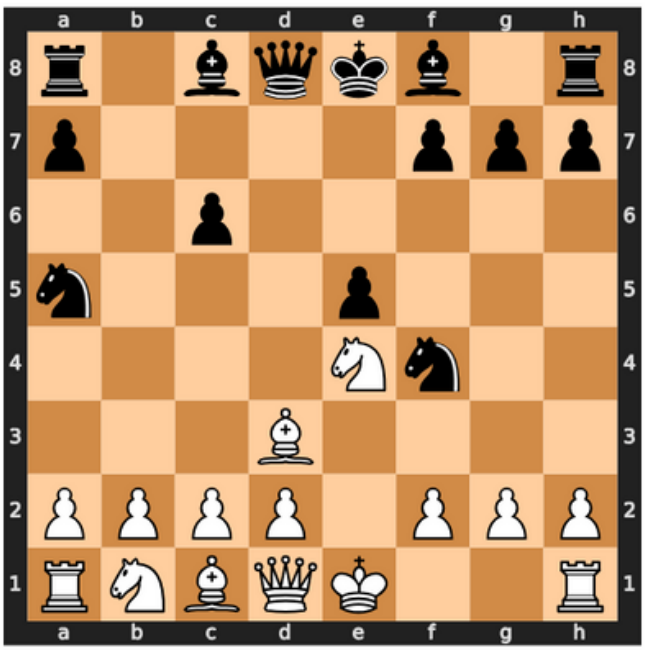
Total Games Played: 121
Success Rate: 53.31%
Win Rate: 51.24%
Lose Rate: 44.63%
Draw Rate: 4.13%

Total Games Played: 83
Success Rate: 49.40%
Win Rate: 48.19%
Lose Rate: 49.40%
Draw Rate: 2.41%

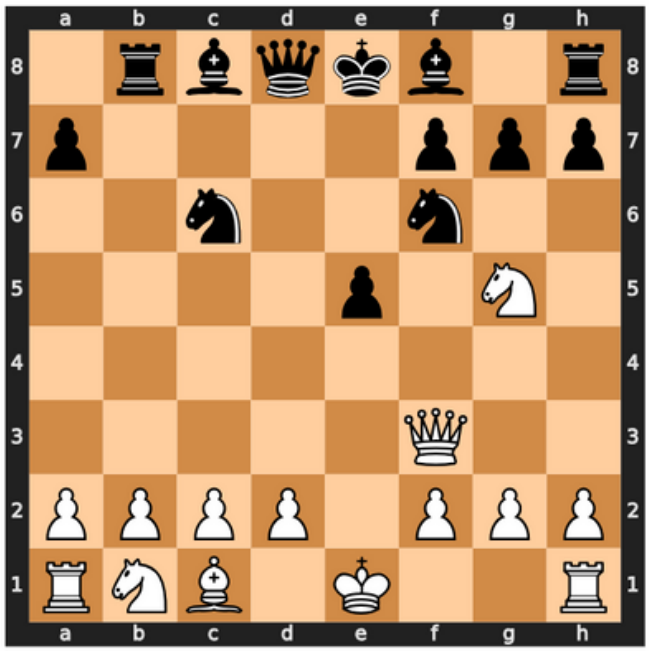
Total Games Played: 117
Success Rate: 44.87%
Win Rate: 42.74%
Lose Rate: 52.99%
Draw Rate: 4.27%

POSITION ANALYSIS

With the increasing depth of moves I managed to find positions with more extreme results.
This is also a result of the smaller number of games.



```
Total Games Played: 9
Success Rate: 34.57%
Win Rate: 33.33%
Lose Rate: 55.56%
Draw Rate: 11.11%
====================
```



```
Total Games Played: 14
Success Rate: 78.57%
Win Rate: 78.57%
Lose Rate: 21.43%
Draw Rate: 0.00%
====================
```

In **move count 18**, I have identified two positions where I achieved either a remarkably high win rate or a significant loss rate.

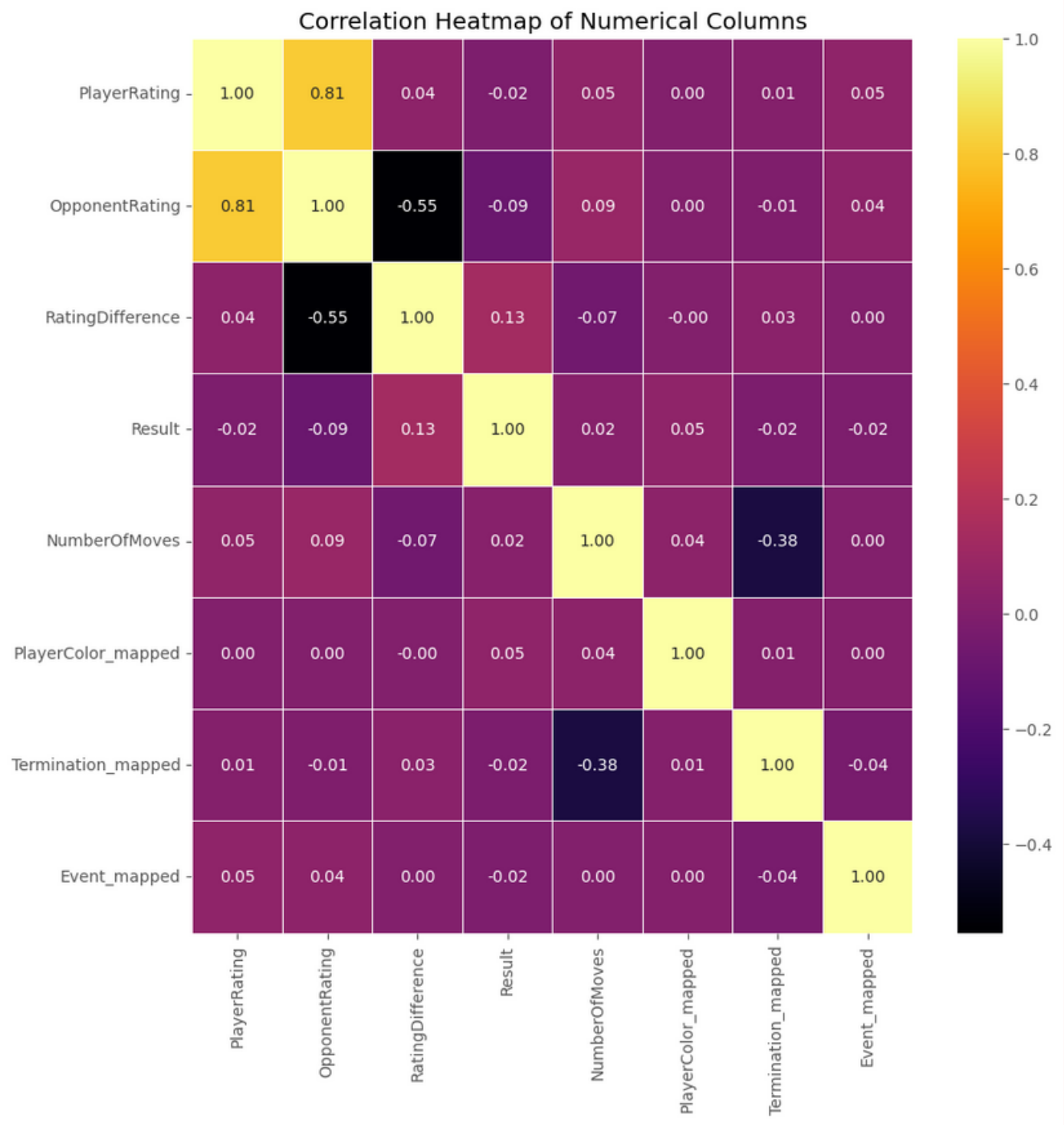
Success Rate: That is a parameter I have created to compare my success in a particular positions. It is calculated by the following formula:

SuccessRate = (WinRate * 1) + (DrawRate * 0.5)

HEAT MAP

Okay, let's explore the parameters that exhibit correlations with each other.

In order to utilize non-numeric parameters such as **Event**, **Player Color**, and **Termination**, I have mapped them to integer values.



HEAT MAP ANALYSIS

Factors Affecting Game Results

As expected, there is a correlation between Rating Difference and Result, as the Rating Difference mirrors the disparity in the chess skills between players.
But this correlation is not strong as I have predicted.
And maybe it can be said that there is weak correlation between Player Color and Result. It is also expected because white side has a slight advantage in game of chess

Correlation Between Ratings

As mentioned earlier, online chess platforms tend to pair individuals with similar skill levels. Therefore, it is not surprising to observe a strong correlation between Opponent Rating and Player Rating.

Termination & Number of Moves

Most of the games in my database are played under the time control of 3+0 (Blitz), indicating that each side has a total of 3 minutes for their moves. As the game progresses, there is an increased likelihood of one side losing on time. Therefore this strong correlation is very reasonable

CONCLUSIONS

There isn't any result that can be labeled as extraordinary, but it is quite satisfying to observe that my data analysis aligns with theoretical expectations. There isn't a singular, obvious parameter for game results to be identified, which is natural, as progressing in chess requires a comprehensive understanding and improvement across all necessary skills to elevate one's rating. Focusing on specific openings is not the key for the progress.

But concentrating on particular positions with a high lose rate can provide players with valuable insights into their deficiencies