Success in Chess

A data driven exploratory data analysis

Jennifer Duan

March 2, 2024

Introduction

Chess has never been more accessible to play. A large contribution towards the community's growth is it's expansion into online services. The top chess players in the world play online matches and stream themselves on sites such as YouTube and twitch.tv. Lichess.org, a popular site for online chess games currently serves over 35 million ranked matches a month.

A single chess game can be broken into three stages: Opening, Mid-game and End-game. The opening consists of the initial set of moves played by both white and black. Traditionally white makes the first move, and it is black's role to respond. The opening "sets the stage" of the game as well as the overall pace of the game. Since openings can be described in approximately 5 turns, it is a common point of study for high level chess players. Most openings in chess have been classified under the Encyclopedia of Chess Openings (or ECO).

By Mid-Game and End-Game both players are now no longer relying on previously studied positions, and in- stead use their general intuition and skill to pick positions that will place their opponent in check.

The Scandanavian Defense: Mieses-Kotroc

Hypothetical Question

The statistical question addressed in this analysis was to investigate the factors influencing the duration of chess games. Specifically, we aimed to understand the relationship between player ratings (both white and black) and the number of turns in a game.

Is there a statistically significant difference between the mean ratings of white and black chess players?

This question was motivated by the curiosity to explore whether players' skill levels, as reflected in their ratings, have a significant impact on the length of chess matches.

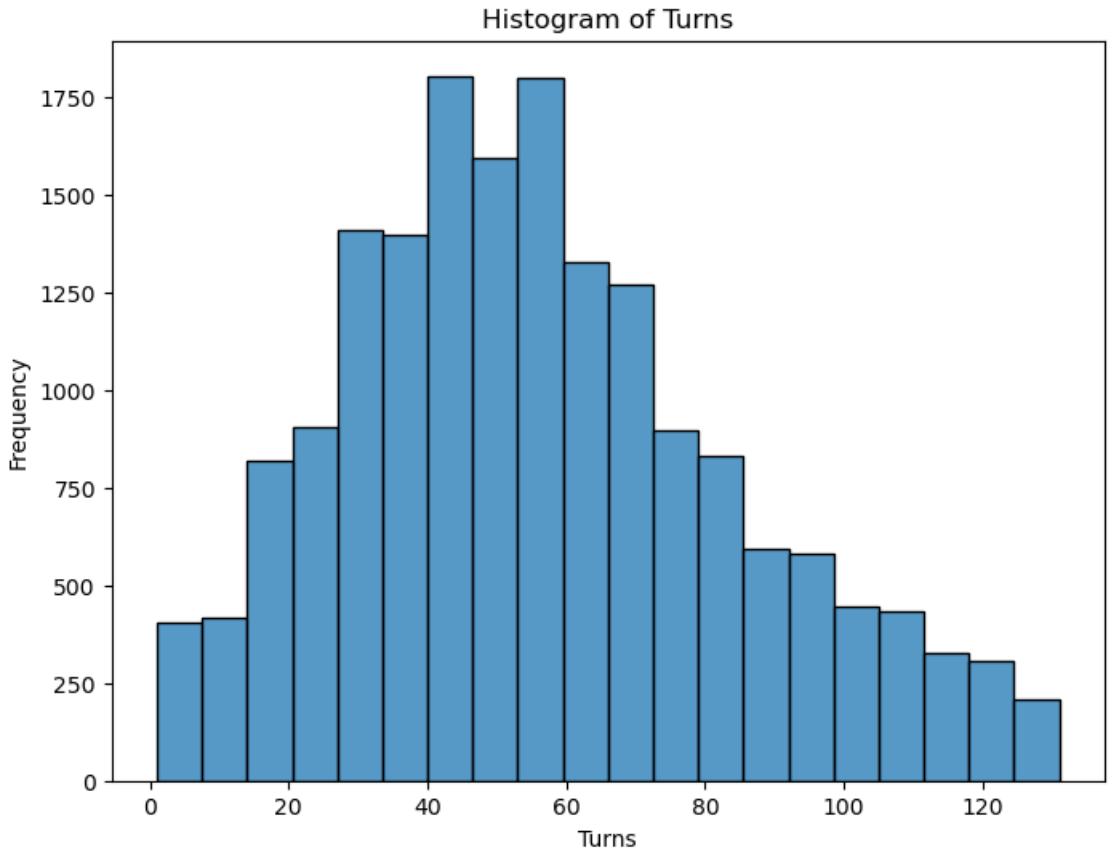
Analysis

To address these research questions, I will analyze a dataset obtained from Kaggle, which contains information about individual chess matches, including player ratings, opening moves, and other 15 relevant variables. This set contains approximately 10,000 games.

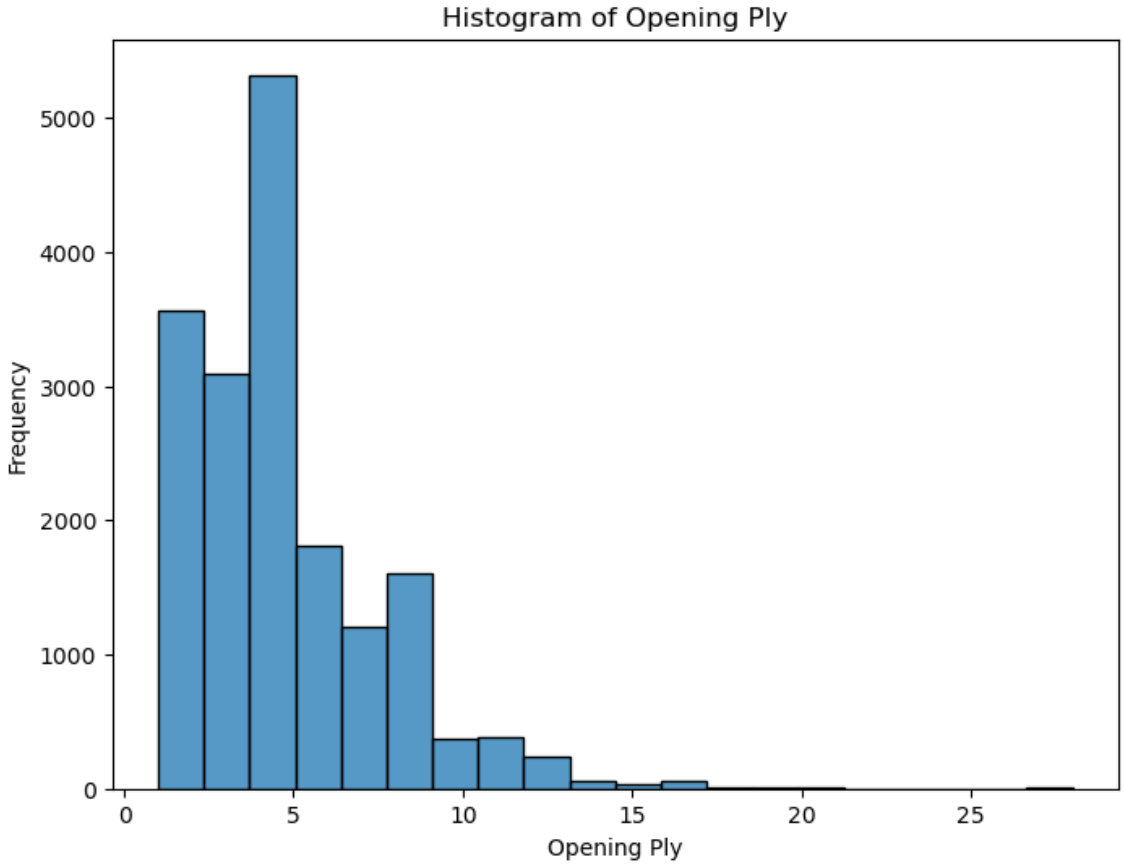
1. Variables Analysis

In the analysis of the variables 'Turns', 'Opening ECO', 'Opening Ply', 'White Rating', and 'Black Rating', outliers can provide valuable insights into potential anomalies or extreme cases within the dataset.

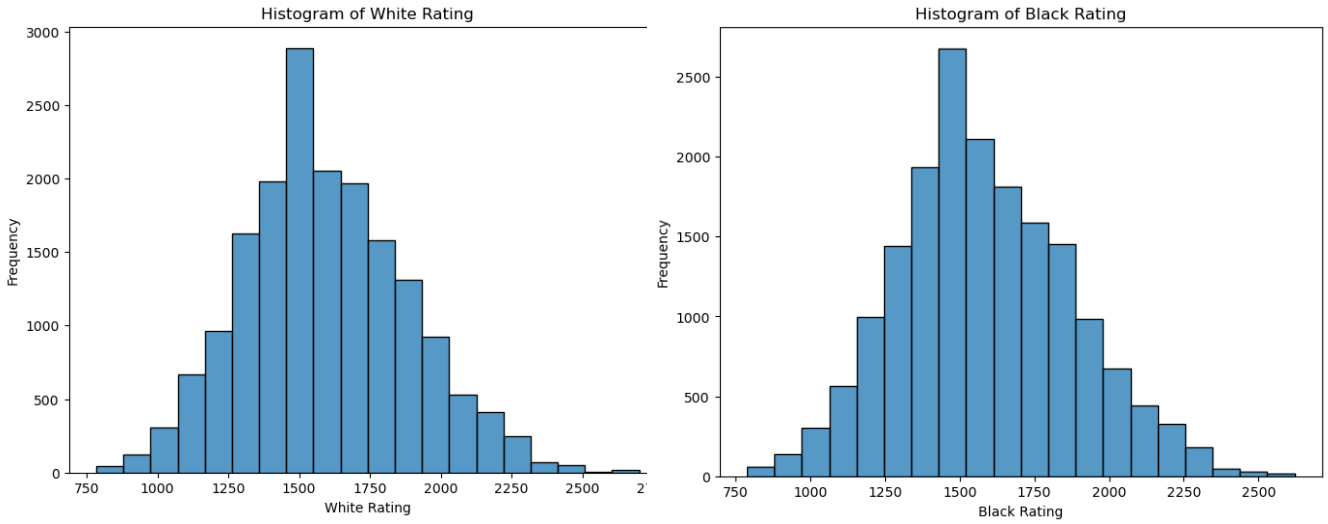
Turns: Outliers in the variable 'Turns' represent games with exceptionally high or low numbers of turns. Games with very few turns may indicate quick resignations or decisive victories, while games with a high number of turns may suggest a prolonged struggle or complex gameplay.



Opening Ply: Outliers in 'Opening Ply' could denote exceptionally short or long opening sequences.



White Rating and Black Rating: Outliers in player ratings could indicate players with exceptionally high or low skill levels compared to the rest of the dataset. These outliers may represent highly skilled grandmasters or inexperienced beginners. Handling outliers in player ratings could involve filtering out players with ratings beyond a certain range considered reasonable for the dataset's context.



I checked the descriptive characteristics about the variables.

Mean Turns: 59.34372619436282

Mean White Rating: 1595.1694961366852

Mean Black Rating: 1586.9201762977473

Spread (Std) Turns: 32.79123920738702

Spread (Std) White Rating: 290.281812118375

Spread (Std) Black Rating: 290.6106468642933

Tails (Skewness) Turns: 0.8965955362314061

Tails (Skewness) White Rating: 0.29543523362766727

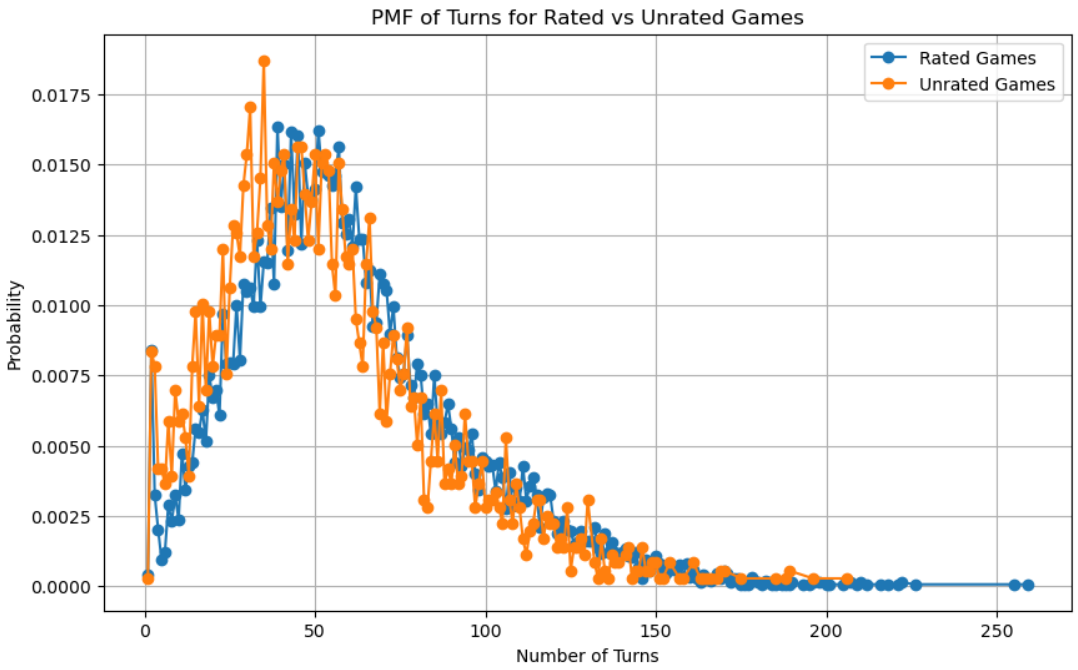
Tails (Skewness) Black Rating: 0.26523329742734925

The summary is:

- The average number of turns in a chess game is around 59.34, indicating the typical duration of a game in the dataset.
- The average rating of white players is approximately 1595.17, while the average rating of black players is approximately 1586.92, providing insight into the average skill levels of players in the dataset.
- The standard deviation of turns is approximately 32.79, indicating variability in the duration of games.
- The standard deviation of player ratings (both white and black) is approximately 290.28 and 290.61, respectively, indicating variability in player skill levels.
- The skewness of the distributions of turns, white player ratings, and black player ratings is positive, indicating slight asymmetry towards higher values in the distributions.

2. Comparison Analysis

I first separate the data into two scenarios: rated games and unrated games based on the 'Rated' column.

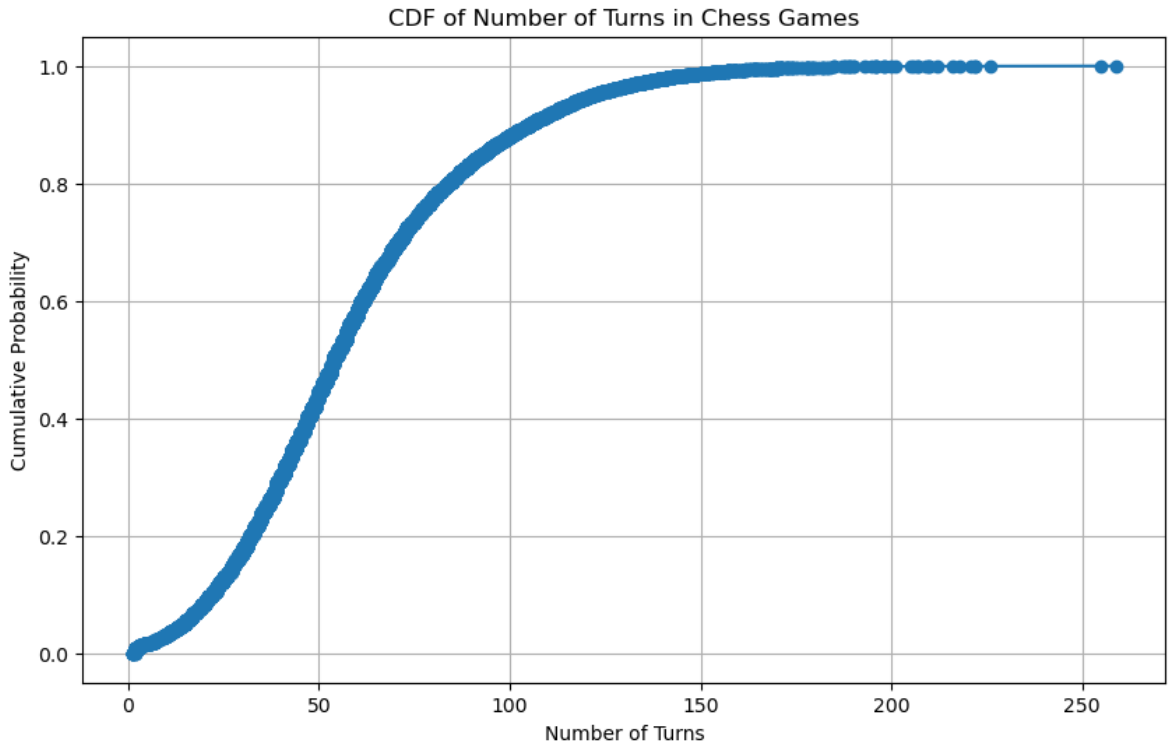


Then, we compute the PMFs for the number of turns ('Turns') in each scenario using the **value_counts()** function normalized by the total number of games in each scenario. From the plot I can see the two scenarios highly related.

3. Cumulative Distribution Function (CDF) for 'Turns',

Cumulative Distribution Function (CDF) helps us understand the distribution of the number of turns in chess games and provides insights into the likelihood of games having a certain number of turns or fewer. It addresses our question by giving us a comprehensive view of the variability in game lengths.

To CDF for 'Turns', which represents the number of turns in a chess game. We can determine the nature and strength of the relationship between 'White Rating' and 'Black Rating'.

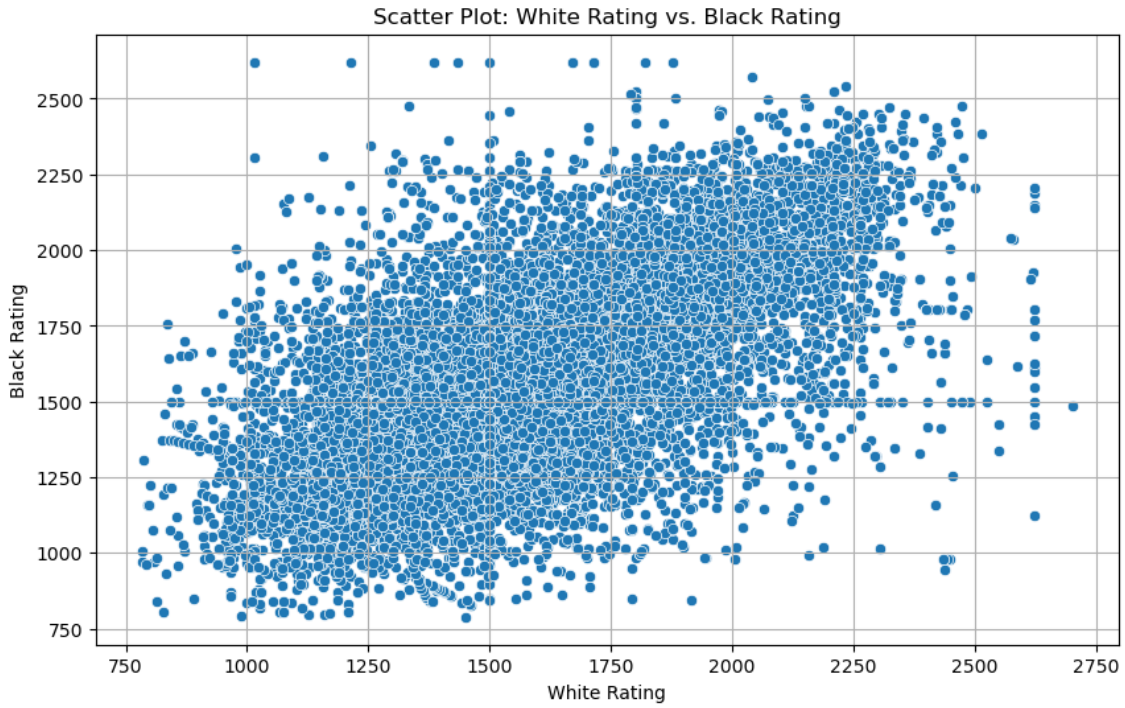


4. Correlation Analysis

The covariance between White Rating and Black Rating is 53197.77. This positive covariance indicates that there is a tendency for both white and black players to have higher ratings simultaneously or lower ratings simultaneously.

The Pearson's correlation coefficient between White Rating and Black Rating is approximately 0.6306. This value indicates a moderately strong positive linear relationship between the ratings of white and black players.

In other words, as the rating of one player increases, the rating of the other player also tends to increase.Overall, both correlation coefficients suggest a significant positive relationship between the ratings of white and black players, implying that players with higher ratings in one group tend to have higher ratings in the other group as well.



5. Regression Analysis

To conduct a regression analysis, let's consider the dependent variable as "Turns" and the explanatory variables as "White Rating" and "Black Rating". We'll use

multiple linear regression to explore how the ratings of white and black players affect the number of turns in a chess game.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Turns   R-squared:                       0.024
Model:                            OLS   Adj. R-squared:                  0.024
Method:                 Least Squares   F-statistic:                     229.5
Date:                Sat, 02 Mar 2024   Prob (F-statistic):           3.48e-99
Time:                        16:41:45   Log-Likelihood:                -89992.
No. Observations:               18378   AIC:                         1.800e+05
Df Residuals:                   18375   BIC:                         1.800e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          29.0410      1.469     19.767      0.000      26.161      31.921
White Rating    0.0052      0.001      4.879      0.000       0.003       0.007
Black Rating    0.0139      0.001     13.115      0.000       0.012       0.016
==============================================================================
Omnibus:                     2294.309   Durbin-Watson:                   1.802
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3443.387
Skew:                           0.916   Prob(JB):                         0.00
Kurtosis:                       4.067   Cond. No.                     1.40e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.4e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In summary, while the ratings of both white and black players have a statistically significant impact on the number of turns in a chess game, the model's explanatory power is limited, and there may be underlying issues such as multicollinearity and deviations from normality that need to be addressed. Further refinement of the model and exploration of additional variables may be necessary to improve its predictive accuracy.

6. Test of hypothesis

Based on the results of the two-sample t-test:

- The t-statistic value is approximately 2.72.
- The p-value is approximately 0.0065.

Since the p-value is less than the significance level (assuming $\alpha = 0.05$), we reject the null hypothesis. This means that there is a significant difference between the mean ratings of white and black players.

In other words, the analysis suggests that there is a statistically significant distinction in the ratings of white and black players in the dataset. This finding could have implications for understanding potential disparities in player performance or skill levels based on racial demographics in chess.