Data visualization project

Mohamed El Hadi Ferdjouni
Remini Anis
Laidi Zakaria

January 2025

Contents

1 Introduction **3**

2 Description of dataset **3**

3 Data pre-treatment **3**

4 Charts **4**
4.1 Pie chart 4
4.2 Normal curve 4
4.3 Bar graph 4
4.4 Scatter plot 4
4.5 Progress bar 4

5 Time control and players rating **5**
5.1 Time control: 5
5.2 Players rating 5

6 Games length and win rates **6**
6.1 Games length 6
6.2 Win rates 6

7 Opening choice impact **6**

8 Chess Game Length by Rating Gap **6**

Introduction

1 Introduction

This report presents an in-depth analysis of a dataset containing over 20,000 online chess games played on the Lichess platform. Chess is a game of strategy, and analyzing large datasets allows us to uncover patterns in player behavior, opening choices, game length, and win rates. By leveraging various data visualization techniques, we aim to extract meaningful insights that provide a better understanding of how different factors—such as time control, rating differences, and game length—affect the outcomes of chess games.

The dataset includes essential details such as player ratings, the opening moves played, the number of moves per game, and the method of victory (e.g., checkmate, resignation, or timeout). To ensure effective analysis, we performed necessary data cleaning and preprocessing, removing unnecessary variables and introducing new ones such as average rating and rating difference. Through the use of pie charts, bar graphs, scatter plots, and other visualization tools, we will highlight key findings and trends that emerge from the dataset.

2 Description of dataset

- Link for the data set: Original dataset.
- Number of variables : 17.
- Number of observations : 20058.

3 Data pre-treatment

Before starting the process of visualizing the data, we performed a cleaning of the data set, removed columns that we considered unimportant for this particular task including the rated variable and the IDs of the players, etc. We introduced new variables called average_rating and rating_diff to use in our graphs, even though they are not explicitly added to the data set they were calculated in the website.

Techniques and methods

4 Charts

4.1 Pie chart

The "pie chart" is also known as a "circle chart", dividing the circular statistical graphic into sectors or sections. Each sector denotes a proportionate part of the whole. To find out the composition of something, Pie chart works the best at that time.

Formula: $(GivenData/Totalvalueof Data)360^\circ$

4.2 Normal curve

A continuous probability distribution that is symmetric about the mean, depicting that data near the mean are more frequent in occurrence than data far from the mean.
We define Normal Distribution as the probability density function of any continuous random variable for any given system.

4.3 Bar graph

The pictorial representation of data in groups, either in horizontal or vertical bars where the length of the bar represents the value of the data present on the axis. They are usually used to display or impart the information belonging to 'categorical data' i.e., data that fit in some category.

4.4 Scatter plot

Scatter plots are the graphs that present the relationship between two variables in a data set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

4.5 Progress bar

A visual representation that shows the advancement of a task or process. While commonly associated with computing, progress indicators are also used in various non-computer-related fields to depict progress.

Results and insights

5 Time control and players rating

5.1 Time control:

We categorized the time control chosen for each game under three main categories:

- Rapid : More than 10 min for each player.
- Blitz : Between 3 min and 10 min for each player.
- Bullet : Less than 3 min for each player.

There do not exist a category for classical chess games because these are extremely rare in online chess games.

Category	Rapid	Blitz 3	Bullet
Number of games	15988	3939	131

Table 1: Time control

We notice a huge disparity between categories,Rapid time control is most used because it gives players time to think carefully about their moves but it is not too long that it gets boring. Blitz comes second because it is a swift time control that makes the games short and light. Bullet is rarely chosen due to the time being too short that it takes high time management skill to perform in this category.

5.2 Players rating

In order to shed light on players ratings, we assigned each game a variable named average_rating that represents the sum of both players' rating divided by two.
We note that the biggest area in the graph was for the intermediate and advanced level, We deduced that it is due to the fact that masters dedicate their time more to over the board chess, and beginners mostly struggle with the game which make them play less.

6 Games length and win rates

6.1 Games length

There is no way to find out the exact time spent in games, so we measured it using the number of moves played, we divided the number of moves to groups and we got the following

Category	< 20	$]20-40]$	$]40-60]$	$]60-80]$	$]80-100]$	> 100
Number of games	1795	4148	5462	3911	2206	2536

Table 2: Games length

This suggests that most games are of moderate lengths between 20 and 60 moves,that does not change the fact that there are many long/short games.

6.2 Win rates

An approximately even distribution with a slight favor of the white side for the winner appears from the data, with most games ending with resignation, which is suitable for online games where inexperienced players mostly give up fighting in losing positions and rather start again.

7 Opening choice impact

We implemented a bar chart to showcase the number of draws and wins with both color for each opening, and visualized the five most played openings. The data indicate the impact of a strong opening on the result of the game, for example a challenging opening like the Sicilian defense has a higher win rate for black, or theoretical openings like the Italian game that has a balanced win rate with a slight advantage for white etc.

8 Chess Game Length by Rating Gap

Rendering the scatter plot with more than 20k rows of data took a toll on the machines so we divided the data into bins to visualize them more clearly, we made the two types of bins, bins for the rating gap, the size of each one was 100, for the number of moves bins the size was 20, and we counted the number of cases that fell in each combination of bins. The result demonstrates that most games have a small to medium margin when it comes to rating difference, to be more precise between [0,400] points of elo, we notice also that the higher the rating difference the shorter the game gets, with a few outliers here and there .

Conclusion

Through this analysis, we have gained valuable insights into the dynamics of online chess. The findings show that rapid time control is the most popular choice among players, and that games tend to be of moderate length, typically lasting between 20 and 60 moves. Additionally, we observed that rating differences influence game length, with larger disparities often leading to shorter games. The analysis of opening choices confirmed that certain openings yield better win rates depending on the color played, highlighting the strategic importance of preparation.

These findings not only enhance our understanding of online chess trends but also provide useful information for players looking to optimize their gameplay strategies. Future studies could incorporate additional variables, such as player experience levels or deeper move-by-move analysis, to further refine our understanding of online chess behavior.

(Definitions and such were sourced from GeeksForGeeks And Byju's.)