Visualization

Chess Games

Visualization and Analysis of Chess and It's Players

Name: Aaron Lim

Student Number: 46420763

Contents

List of Figures

Introduction

Chess is a popular board game that has been played for centuries and is still growing to this day thanks to sites such as Chess.com, the world's most popular chess site. On January 20th 2023 alone, 31 million games were played with over ten million active members. The chess player-base is huge and is continuing to grow with more people wanting to learn about chess.

I have acquired 2 datasets related to the game, and in this report, the objective is to create visualizations from these datasets that will help us find patterns and trends that may emerge from the data and use these to understand the behavior of chess players, their average performance, the strategies used as well as interesting insights into the game.

Data

Acquiring Data

Two datasets were used. The first dataset is a collection of 60,000 games played on chess.com from 2013 to 2021. It contains a range of games including the various types of time-controls (how much time the players get), different gamemodes which slightly change the rules (e.g. crazyhouse, chess960) and games of various skill level.

The dataset was uploaded to kaggle.com [1] by user ADITYAJHA1504 which was aggregated using a chess.com API.

Features

The dataset contains the following notable features as described on Kaggle:

white_rating	White's ELO rating
black_rating	Black's ELO rating
white_result	Either win or the loss condition (like checkmate, draw, etc.)
black_result	Either win or the loss condition (like checkmate, draw, etc.)
time_class	blitz, bullet, rapid or daily
time_control	Total_time + Time_increment
rules	Either normal chess or other variants (like chess960)
rated	Whether ELO points are at stake
fen	Standard notation for describing a particular board position of a chess game.
pgn	standard plain text format for recording chess games

The other dataset is a complete list of over 400,000 FIDE (The International Chess Federation) rated players which contains player ratings, countries, gender, inactivity, title, birthyear, etc. The dataset was acquired from the official FIDE website [3].

Data Processing

We can obtain even more features from the chess.com dataset by reading the pgn which contains some features we can use. Using a simple python script to parse the pgn, the following features were added:

event	The tournament in which the game was played
start_date	The date when the game started
end_date	The date when the game ended
start_time	The time when the game started
end_time	The time when the game ended
eco	The eco code for the opening used
ecoName	The name of the opening used
moves	The moves made in the game in standard notation

After adding these features, the data was cleaned and was searched for incomplete/ inaccurate data. One such incompletion was found in which multiple games had no moves and were removed. I also made the decision to only analyze standard games of chess (where Rules == "chess") since the other game-modes didn't have many observations and would potentially taint the results.

The FIDE dataset used country codes which weren't consistent with any country code system recognized by geopandas, so the codes were converted to ISO3 format. The dataset contains a flag which indicates whether the person is an active player so in order to keep the data relevant to the present, only active players were considered.

Data Visualization

Rating Distribution

Chess uses the 'Elo rating system' to calculate a player's rating when a game is won or lost based on their rating and their opponents rating. This number is relative to their skill and the difference between ratings is a good indicator of how likely a player will win a match where the higher rated player is expected to win. This rating is also factored in when match-making games so that the difference in rating between players is not too high. Players pride themselves on their rating and will find it interesting to know how much of the player-base have a lower rating than themselves.

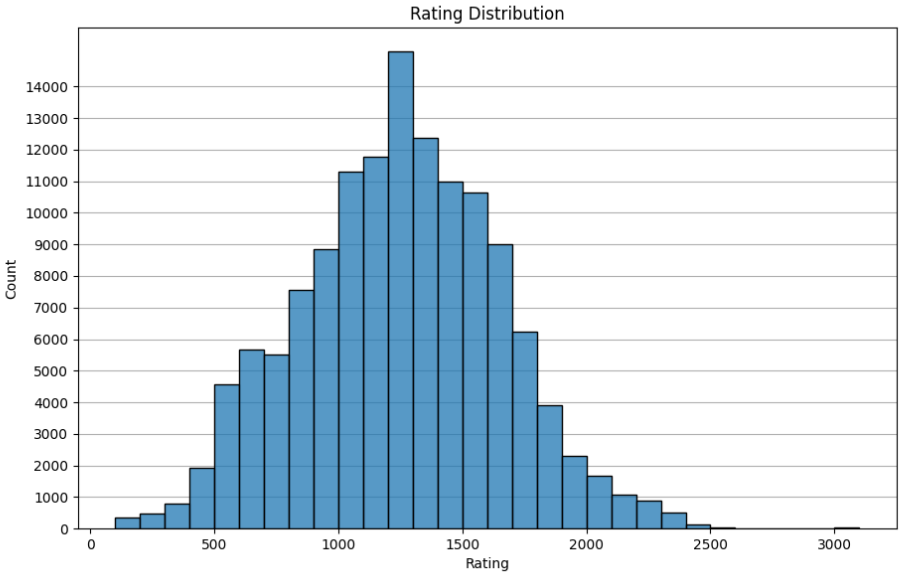


Figure 1: The rating distribution of users

Shown in fig 1, is a histogram with intervals of 100 which outlines the rating distribution. Using this histogram we can see that the rating distribution obtains a slight positive skewness which tells us there are more beginner/ lower rated players than there are advanced players in this dataset. There are some outliers that we can see. The 1200-1300 bin has much more observations than the general shape would suggest. We can also see that for ratings 2600 and above, there are not many observations.

We can see that it appears to follow a normal distribution. To explore this further we can use a Q-Q plot. A Q-Q plot compares the quantiles of two datasets, so if we compare the ratings quantiles to the quantiles of a normal distribution, we can see how similarly they are distributed.

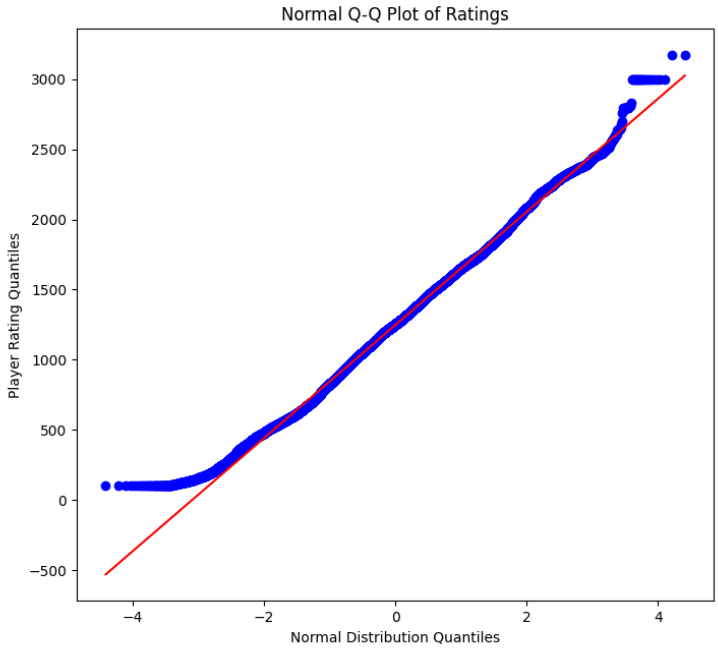


Figure 2: Q-Q Plot comparing the quantiles of the player ratings and a normal distribution

In fig 2, we can see that the quantiles around the center fall directly on the identity line, which is a good sign that the ratings follow a normal distribution as they have comparable quantiles. We can also see that the points trail off at the ends. This is an indication that our data has thin tails - that both ends of the ratings have fewer observations than expected for a normal distribution - but it doesn't trail off by much. From fig 1 and fig 2, there is strong evidence that the player ratings are normally distributed.

Knowing that the ratings are normally distributed, we have a good idea of the rating of the average user. We can find the sample mean:

$$\mu = \frac{\sum x_i}{n} = 1247.28$$

and standard deviation:

$$\sigma = \frac{\sum (x_i - \mu)^2}{n - 1} = 403.72$$

And plot the normal distribution with these sample μ and σ over the histogram:

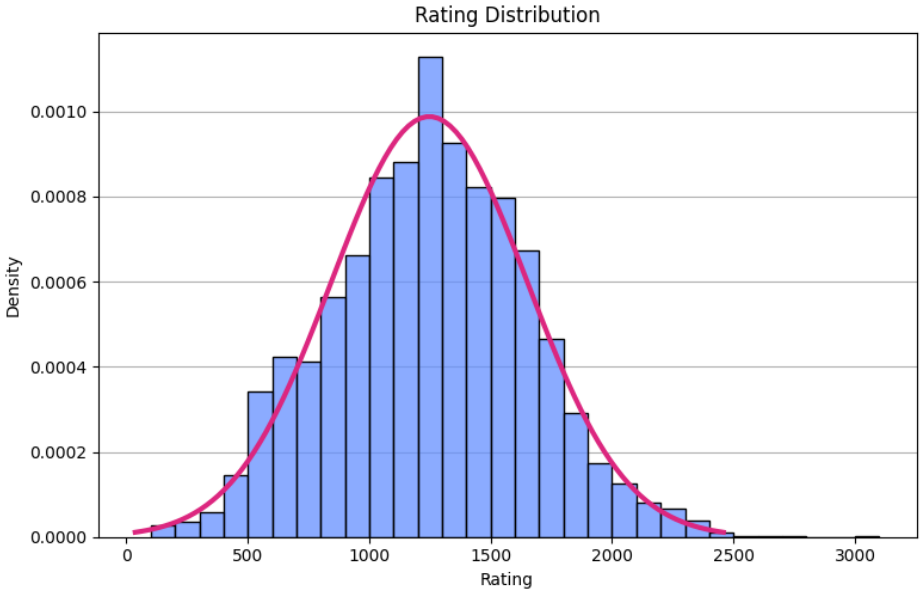


Figure 3: The rating distribution of users against the estimated normal distribution

We can now use this normal distribution to estimate the distribution of the ratings.

Rating Difference

In the previous section, we touched on how higher rated players are expected to win. In this section we will dive deeper and look at how much rating affects the outcome of games. We can first show this by finding the percentage of games where the higher rating wins.

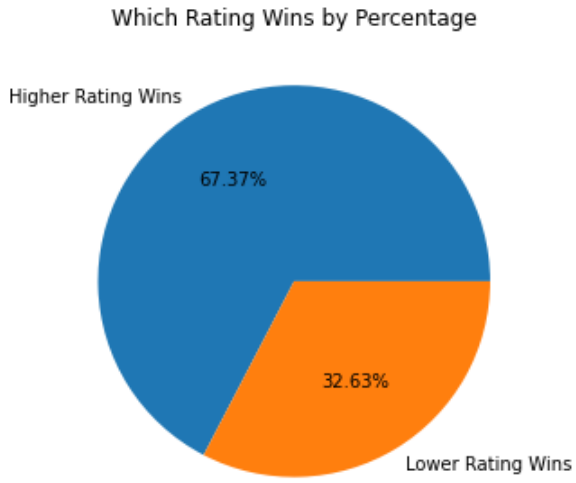


Figure 4: The proportion of games where games are won by the higher/lower rated player

There is a clear indication that if you have a higher rating, the better chance you have of winning. But what this graph does not show is how much the difference between the ratings makes. To investigate further we can show the distribution of rating difference between the winner and the loser. This gives the following histogram:

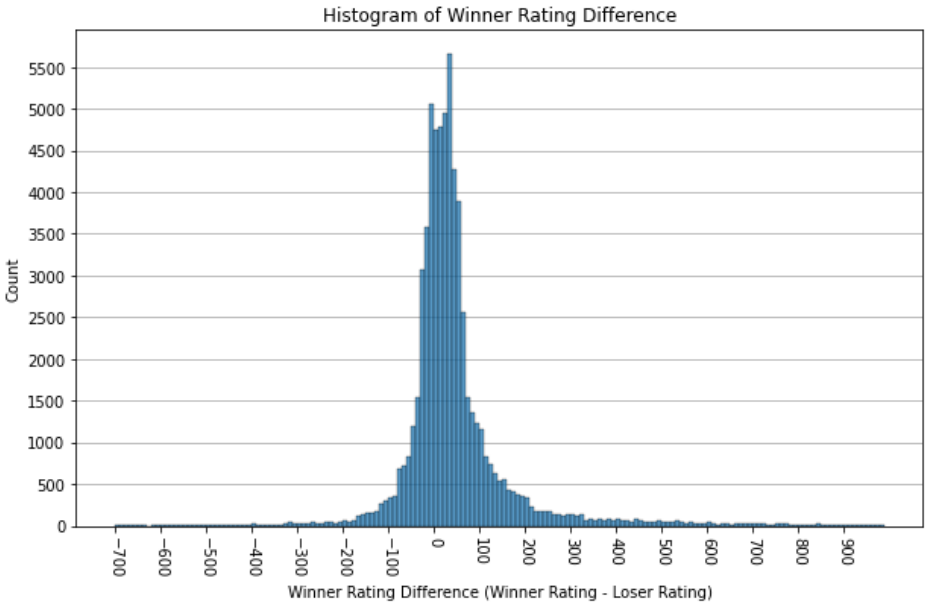


Figure 5: The distribution of the difference between the winners rating and the loser rating

We can see that the data is heavily left-skew, with most of the data on the right side. This means that more often that the higher rated player wins. But this is not the whole story. On the left side we can see that lower rated winners still win as often as higher rated winners for rating differences of around -50 to 50, so lower rated players still have a fighting chance. However, the more negative the difference, the quicker the number of observations decrease than when you go positive.

I wanted to find out how even the games were, depending on the rating difference using the number of moves as a heuristic of the evenness of games. To visualize this, I used a scatter plot to plot the number of moves against the rating difference between white and black for both cases when the higher rated player wins and when the lower player wins. I chose not to use the absolute value so that we can see if high rated or low rated wins differ between colors.

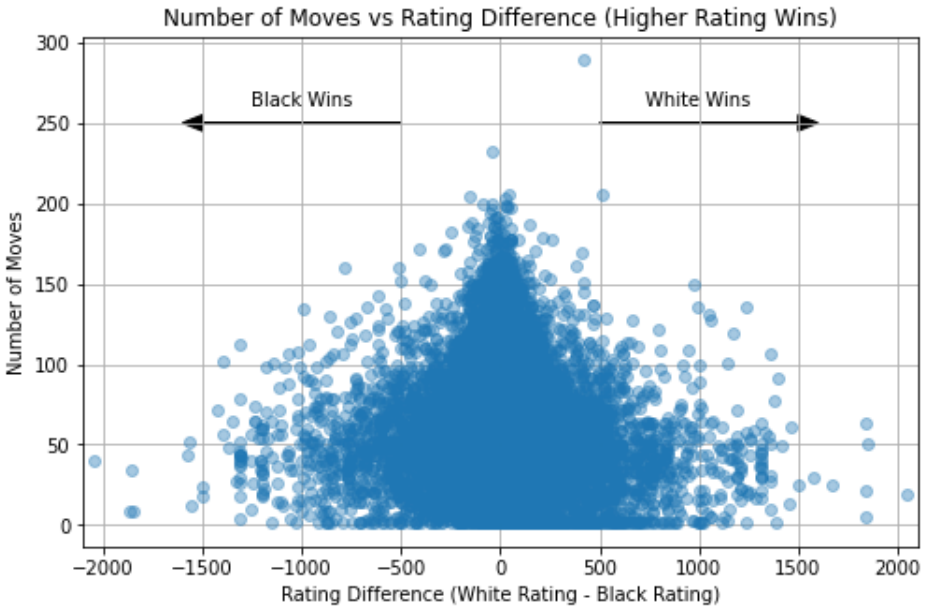


Figure 6: The number of moves vs the difference between whites and black rating where the higher rating won

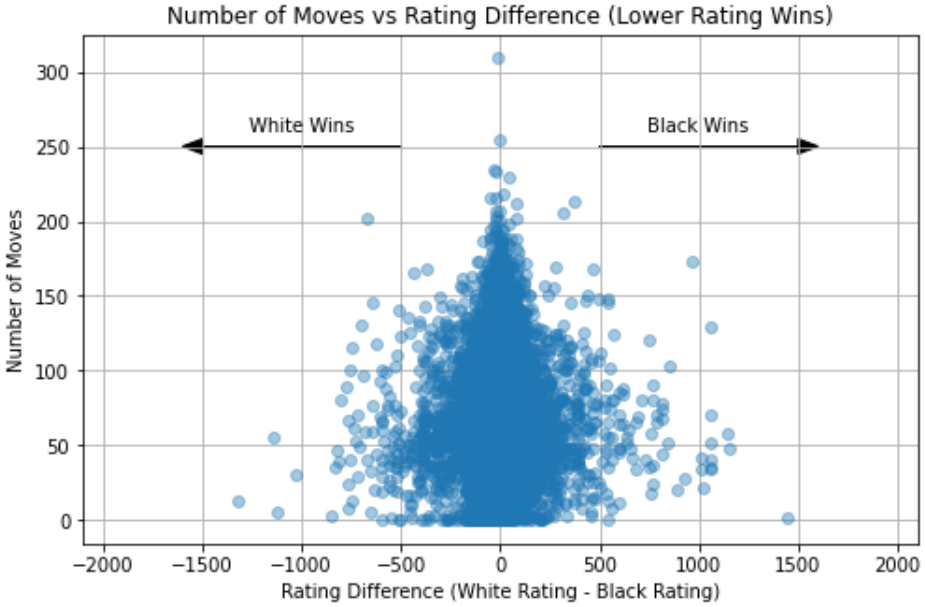


Figure 7: The number of moves vs the difference between whites and black rating where the higher rating won

Since the rating difference is calculated by subtracting blacks rating from white's rating, in fig 6 all data points on the right of the x = 0 are resulted from white wins and subsequently all data points on the left of x = 0 are black wins. Conversely, in fig 7, left of x = 0 are white wins and right are black wins.

We can see that both scatter plots are fairly symmetrical and centered around 0 which tells us that colors have little difference in determining the evenness of the games.

There is an immediate noticeable difference between the two scatter plots. The higher rating winners have a greater spread of points whereas the lower rating wins scatter-plot has a tighter spread, with most of the points contained within a rating difference of -500 to 500. This tells us that it is much harder for the lower rated player to win if the rating difference exceeds 500.

Around x = 0, we tend to see more moves, which indicates that the for small rating differences, the games tend to be more even. We can infer that the 'Elo rating system' implemented by chess.com is doing its job. We also see that as the rating difference grows, the number of moves decreases which tells us that in general, the greater the rating difference, the less likely you are to have close game.

A short-coming of this graph is that it does not tell us about the density of these points, i.e. the distribution of the number of moves over all games. However, there was an attempt to show this by changing the alpha value of the points to appear more opaque so that dense areas appear darker, but it is dubious around the center of the cloud as there are too many points.

Chess Openings

Openings are the strategical moves that you make in the beginning of a chess game. Each opening caters to a different play-style: aggressive, defensive, etc. There are multiple openings with different given names, and for each opening there are multiple variations which are slightly different from each other. This begs the question of which is the best opening and does this change over different ratings.

To investigate this, I wanted to find the win-rates for each popular opening and how effective they are in different Elo ranges. To ensure that the openings had a decent amount of observations, they were only considered if at least 500 games were played. The counts for these openings are seen in the following bar graph:

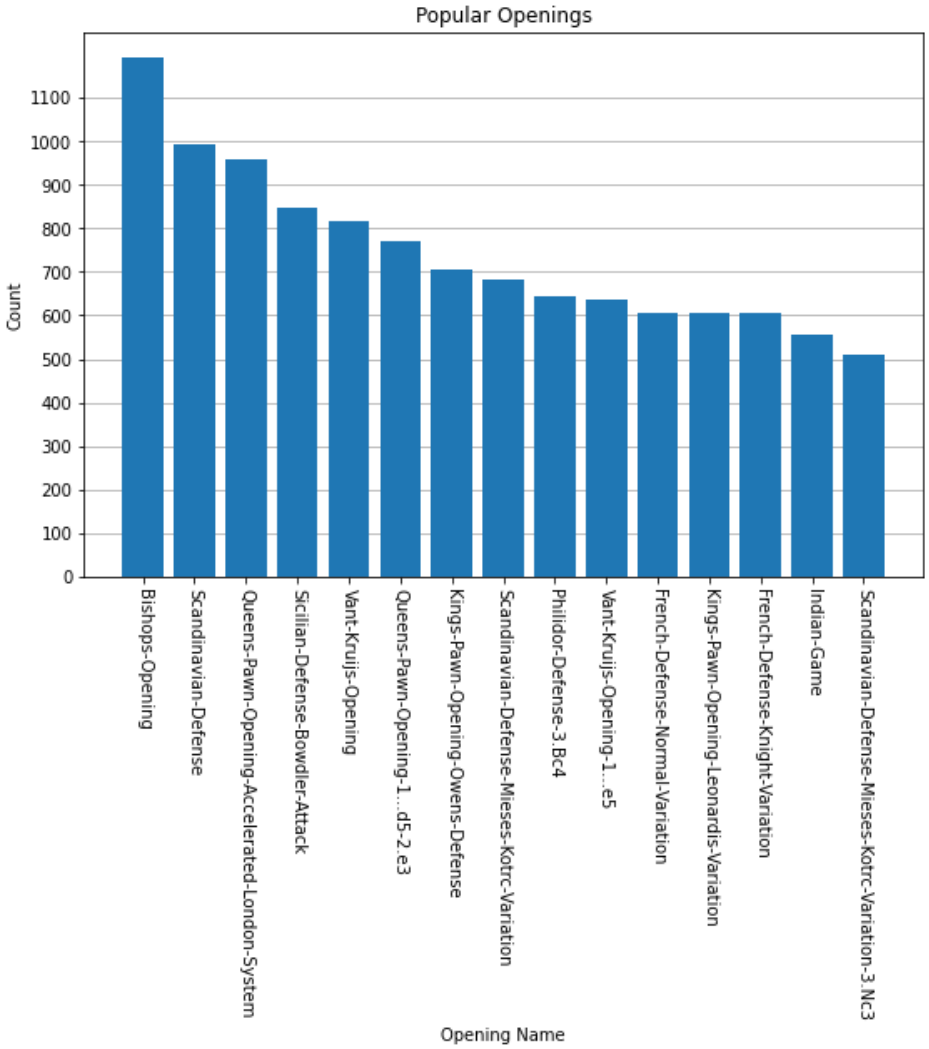


Figure 8: The number of games vs the time of day. The number of games played was polled every 10 minutes

To make sure that each Elo range has the around the same number of observations, I choose the ratings based on the third quantiles: $\{(0, \frac{1}{3}], (\frac{1}{3},\frac{2}{3}], (\frac{2}{3}, 1]\}$ which turned out to be: {(0, 1078], (1078, 1425], 1425+}. Since both players could be categorized in

two separate ranges, the game was considered for a range if at least one of the players fits in that range.

The results can be seen the graphs below:

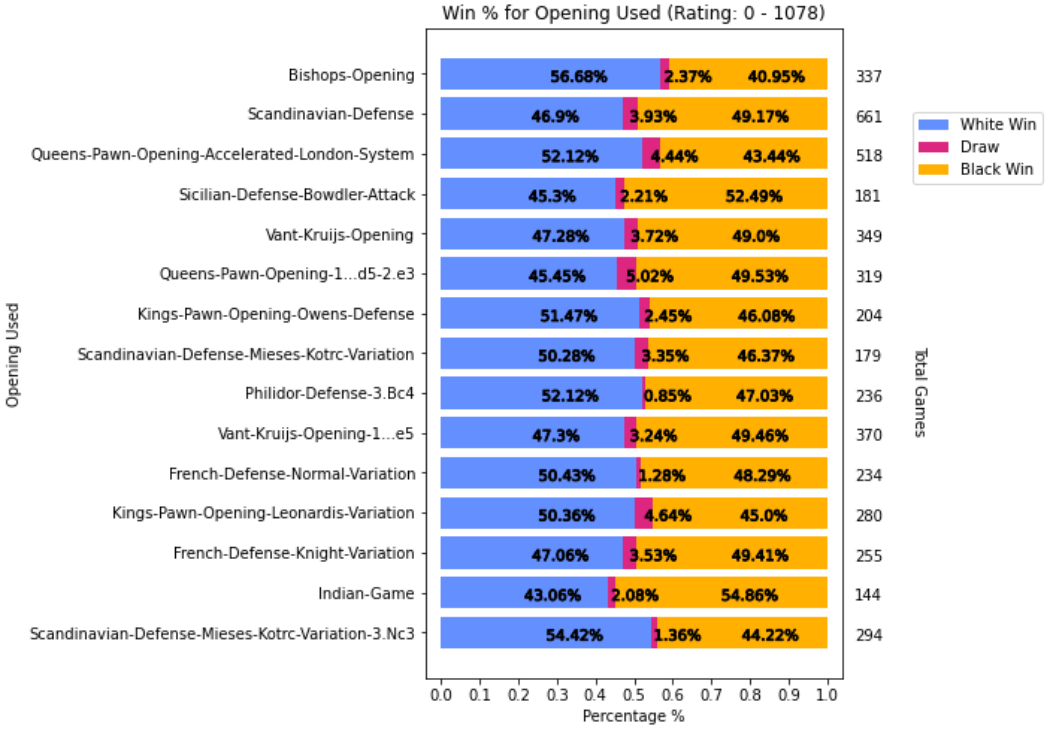


Figure 9: The win rate for each popular opening for ratings 0 - 1078

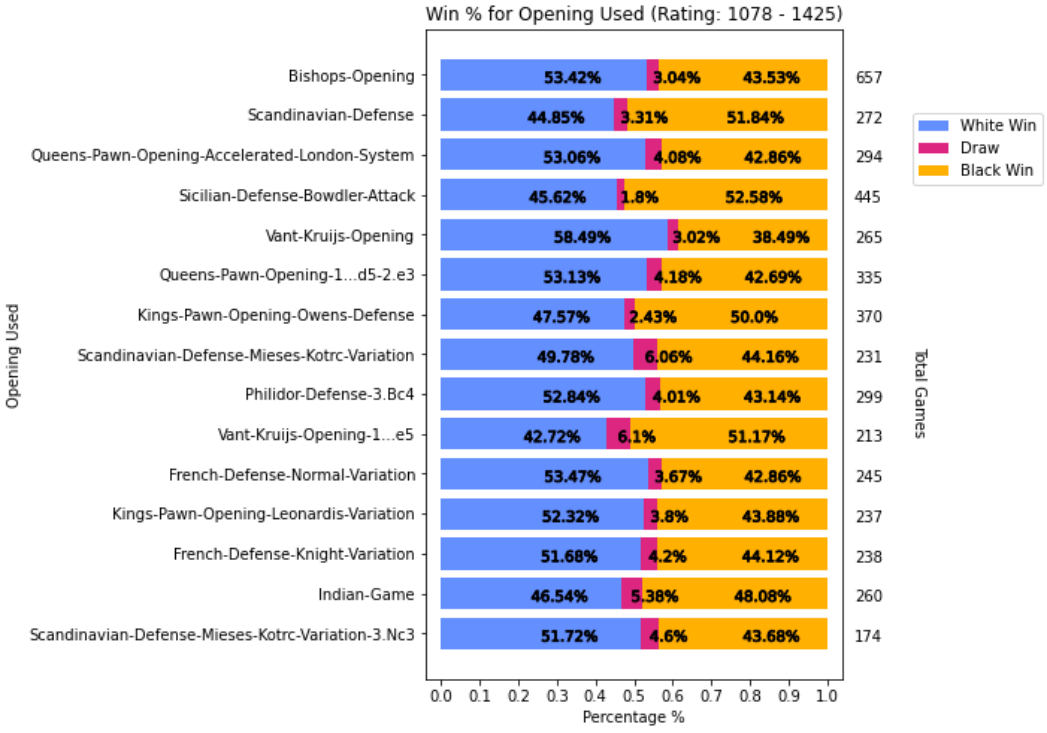


Figure 10: The win rate for each popular opening for ratings 1078 - 1425

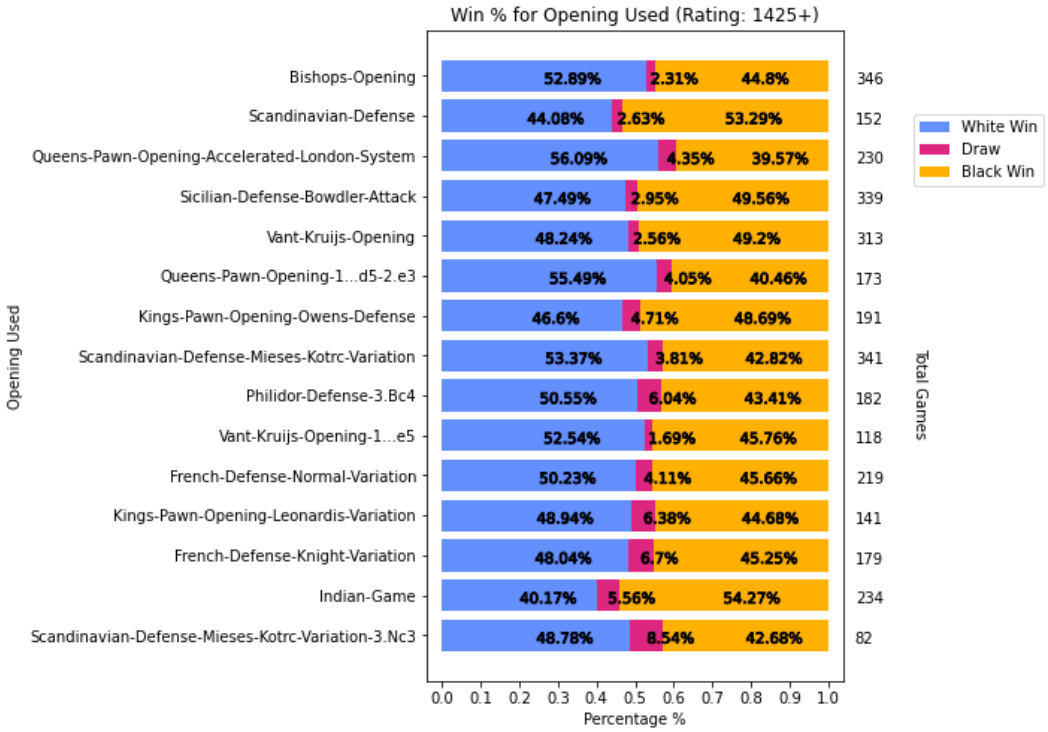


Figure 11: The win rate for each popular opening for ratings 1425+

For most of the openings, the win rate seems fairly consistent over all rating ranges where the win rates don't differ by much. We can also see that for most openings, white has the greater win percentage.

The most popular opening, the Bishop's Opening, performs consistently across all rating ranges, having white win the majority of the time with win percentages in the range 52 - 56%. To increase your odds of winning as black, it might be best to avoid playing into this opening.

The Scandinavian defense is an aggressive opening for black where black plays the queen, the most powerful piece, in their second move. There is a general rule of thumb in chess that the queen shouldn't be brought out too early. This rule of thumb is proven true in the lower ratings where white has an 54% win-rate over black. However, for the higher ratings, it seems that this opening has a more even outcome where around 8% of the games resulted in a draw. This makes sense as it is much easier for a lower rated player to lose their queen to this move early, causing them to lose the game, but for higher rated players, this might result in a trade of queens causing the game to be more even.

Openings with win-rate of over 50% on the color you're playing are your best bets and should be considered to be implemented in your repertoire.

Number of Games Played

In a recent article by chess.com [2], they talked about how their servers are struggling especially at peak hours: 16:00 UTC to 20:00 UTC. In an attempt to visualize this, we can graph the amount of games that are played during the day. To do this, the day was split up in increments of 10 minutes and the number of games was polled for each bin based on the games start time.

The results are shown below:

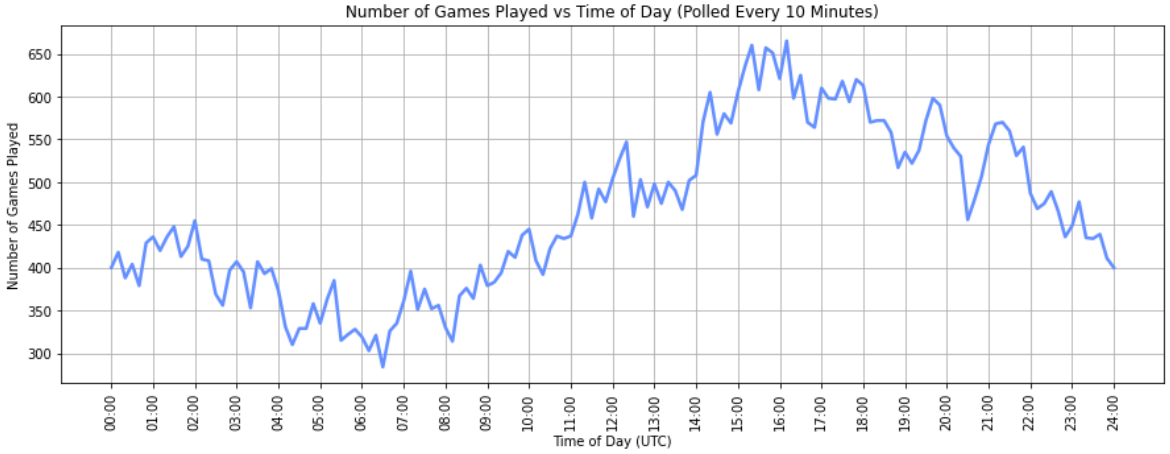


Figure 12: The number of games vs the time of day. The number of games played was polled every 10 minutes

It should be noted that the data in this dataset is a little outdated (2021) and many not reflect the current situation. That being said, the claim made by chess.com seems to be correct in that most players play chess at around 16:00 UTC where around 650 games were played and at its lowest, approximately 300 games were played at around 6:30 UTC. The general shape of the data follows a sin function with an intercept of around 450 games, an amplitude of -150 and a period of 24 hours.

To get the best parameters, I used the curve_fit function in the scipy.optimize library to find the best a and b parameters for the following function where the period is fixed to 144 (6 * 24 minutes in a day).

$$f(x) = a \cdot \sin\left(\frac{2\pi}{144}x\right) + b$$

This gives the parameters a = -124.387, b = 463.993. The fitted function on the plot looks like:

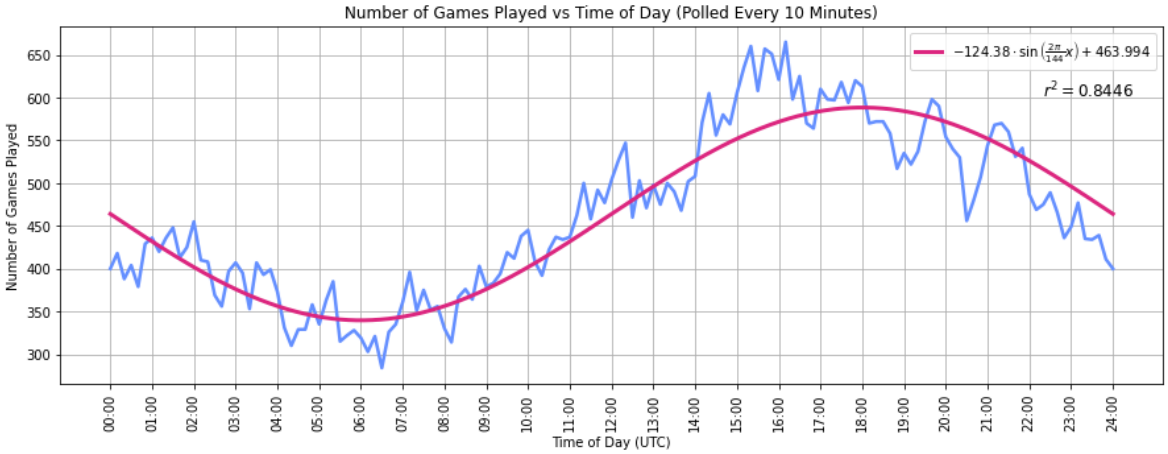


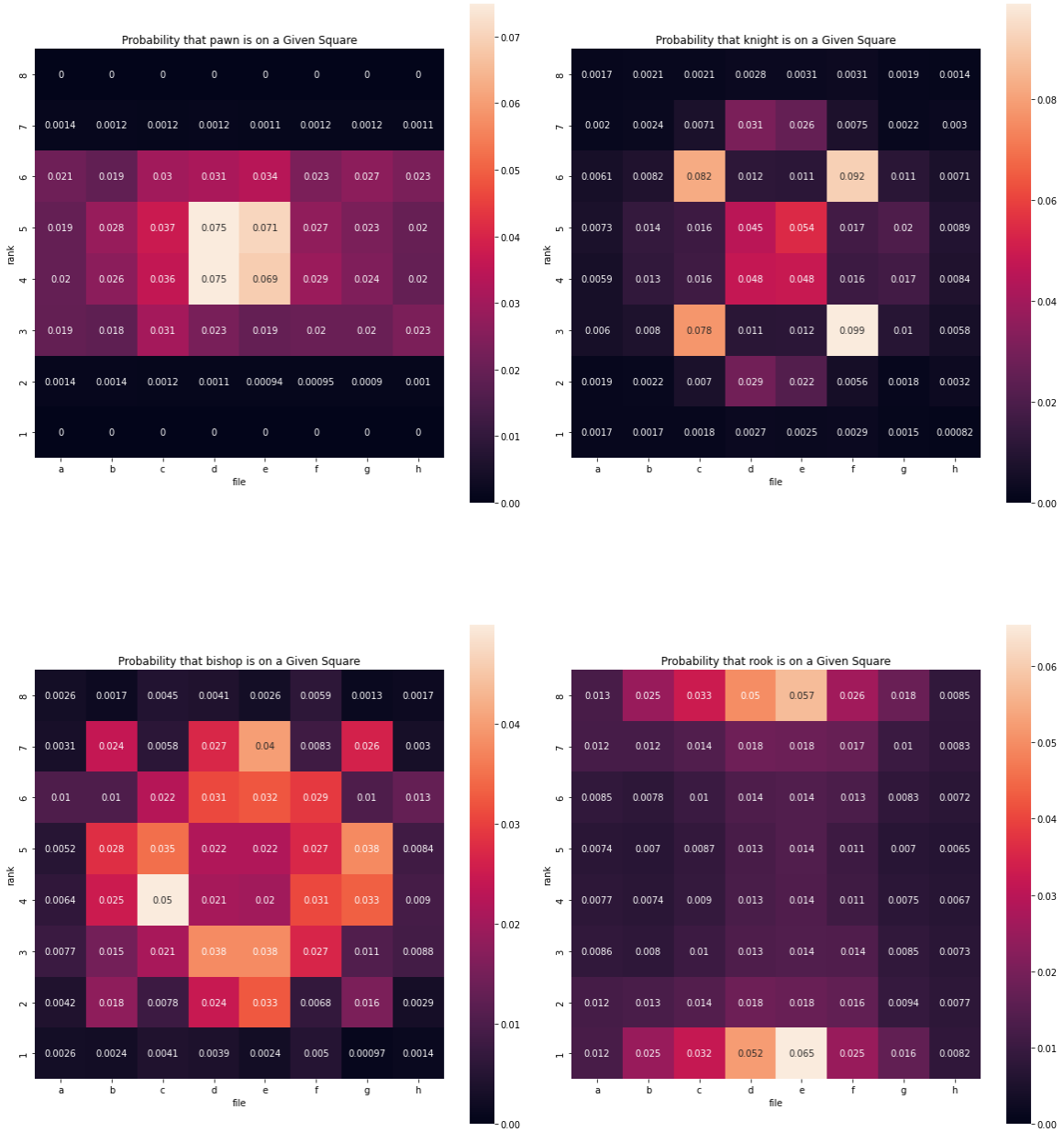
Figure 13: The number of games vs the time of day with fitted sin function

And gives us a R^2 value of 0.8446 which isn't the best but can be used to give a rough estimate.

Piece Behavior

There are 6 pieces in chess which all serve different purposes, and it can be difficult to determine the roles all these pieces play. To explore this, I used a python script to parse through all games and note the squares in which every piece moved to and represented the results in a heatmap. The heatmap plots the probability that the respective piece moves to a given square in a game for both black and white pieces where white starts on the bottom and black on the top.

Squares are noted by their rank (a, b, c, ..., h) and their file (1, 2, 3, ..., 8) as indicated by the axes.



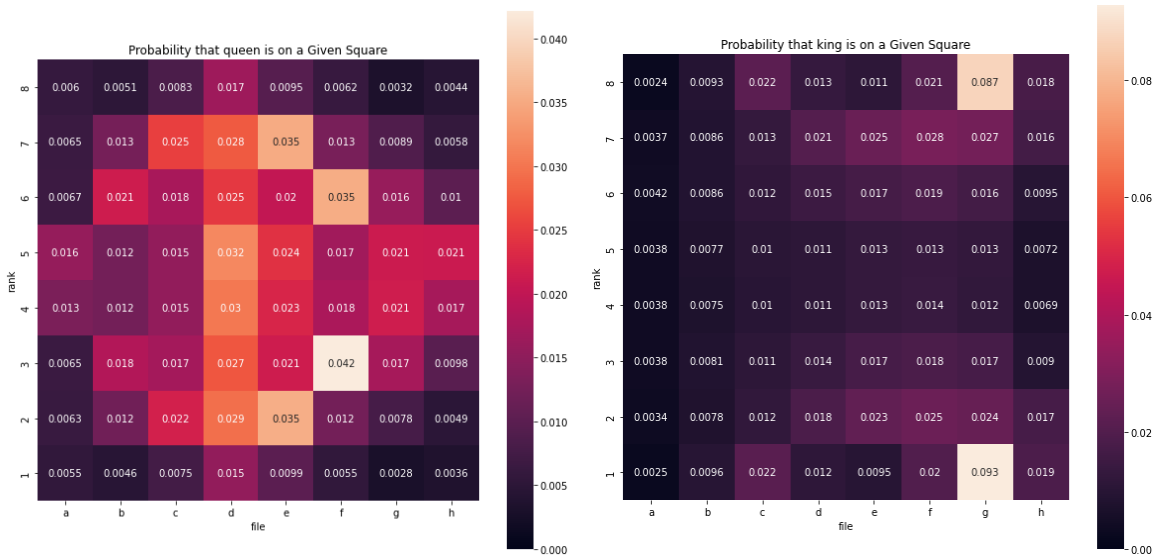


Figure 16: Probability that the piece will move to a square in a game

We can see that pawns are mostly situated in the center of the board with the middle four squares having the most traffic. This makes sense as the role of the pawns are to control the center of the board, preventing the other pieces from coming in.

The knight is mostly placed on the same 12 squares, and like the pawns, tends to be placed in the center of the board. We can see that players tend to bring their right knight to the 3rd rank more often than their left knight. This makes sense because - as we will see later - the king tends to be on the right side of the board and these moves protect the right side.

We can see that the bishop has coverage on most squares of the board except on the corners and edges which makes sense because there, the bishop is the weakest as it covers the least amount of squares. The bishop has a high probability to be on the c4 square which makes sense as it is a move apart of the most popular opening: 'the Bishop's Opening' which was covered previously. We can see that the bishop is used to be a versatile piece which can be placed in many squares.

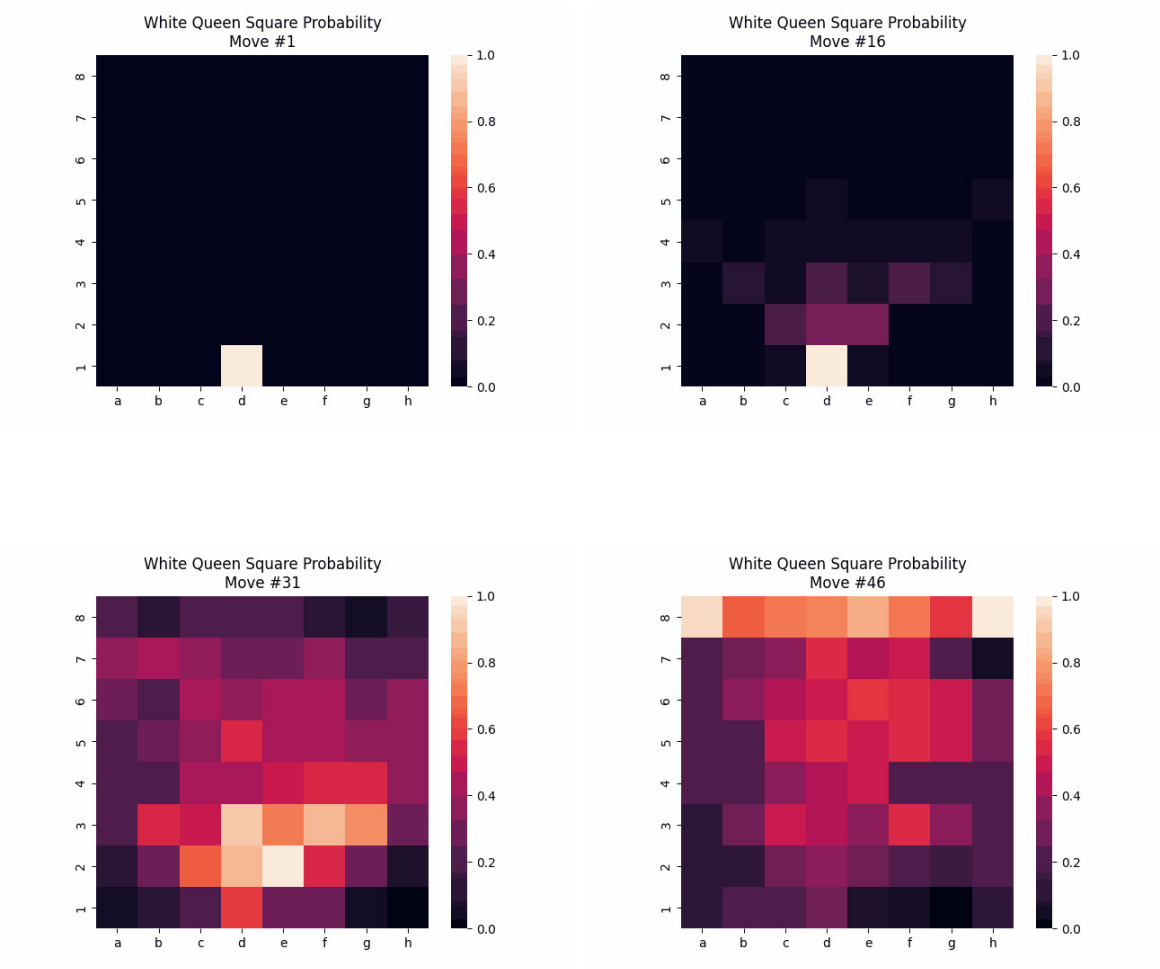
The rook tend to be on the back ranks (1 and 8) and only a small portion of the moves are in the squares at the center of the board. This makes senses as rooks are much safer in the back ranks and usually trapped behind pawns. The most probable move square is e1 ad e8 which makes sense as the rook can be placed on these squares to check the opposing king. The role of the rook is to protect the back ranks as well as attacking the opposing player's back ranks.

Much like the bishop, the queen has a lot of coverage of the board but tends and avoid the corners. The queen's most active squares lie on the d-file which makes senses as this is the file where the queen starts off in the game. The most popular square to place the queen by far is the f3 square which makes sense since it threatens the f7 square, a vulnerable square for black. The queens job is to attack the many squares of the board.

It is evident from the heatmap that the king is by far the least active piece, staying in the corners where it is safe. The most popular squares by far are the g1 and g8 squares which makes sense as this is the square where the king ends up when castling king-side. On the left side, the c1 square is where the king ends up when castling queen-side but is much less popular. This tells us that players much prefer to castle king-side. The king seems to avoid the center squares on the back rank which makes sense as it is prone to the rooks which we saw previously tend to be on these files. The kings job is just to avoid danger.

White Queen Behavior

To investigate the white queen's movement across the game, I created an animation that shows each squares probability of the white queen being on it for each move in the game. This was done using a python script which notes the square the queen is in for each move in the game over all games in the dataset. The resulting animation is attached as 'queen animation.gif'. Here are a couple of notable frames:



We can see that at around move 16, the white queens tend to move out from its staring position. At around 31 moves, the d1 square (the queens starting square) is not longer the most probable square. This is a good indicator of the 'mid game' - the point in

the game after openings are done. At around move 46 we see that a lot of queens are at black back rank. This is due to pawn promotion - when a pawn reaches the other end, your piece can turn into a queen. This is a good indication of the end game - when most of the minor pieces are gone. The visualization gave a good idea of what squares the queen tend to especially in the earlier moves and a gave a good indication of around what move the games turns into the mid-game/ end-game. It is clear that the queen has a lot of control over the board as in the mid/end game, the queen moves to many squares.

Country Rating

FIDE (The International Chess Federation) is the governing body of chess which host multiple tournaments around the world. By participating in these tournaments you get a FIDE rating. Players who have had outstanding performance in chess and have met the requirements are awarded a FIDE title. These titles include (from lowest to highest) Candidate Master, FIDE master, International Master and the highest: Grandmaster. I wanted to find out in which countries, there are the most FIDE rated players, titled players as well as just Grandmasters. Countries with a lot of titled players can be an indication of where the best chess communities and players are located.

Using the FIDE dataset, I only considered players who were marked as active. The following are heatmaps that show the number of FIDE rated players, titled players and Grandmasters for each country:

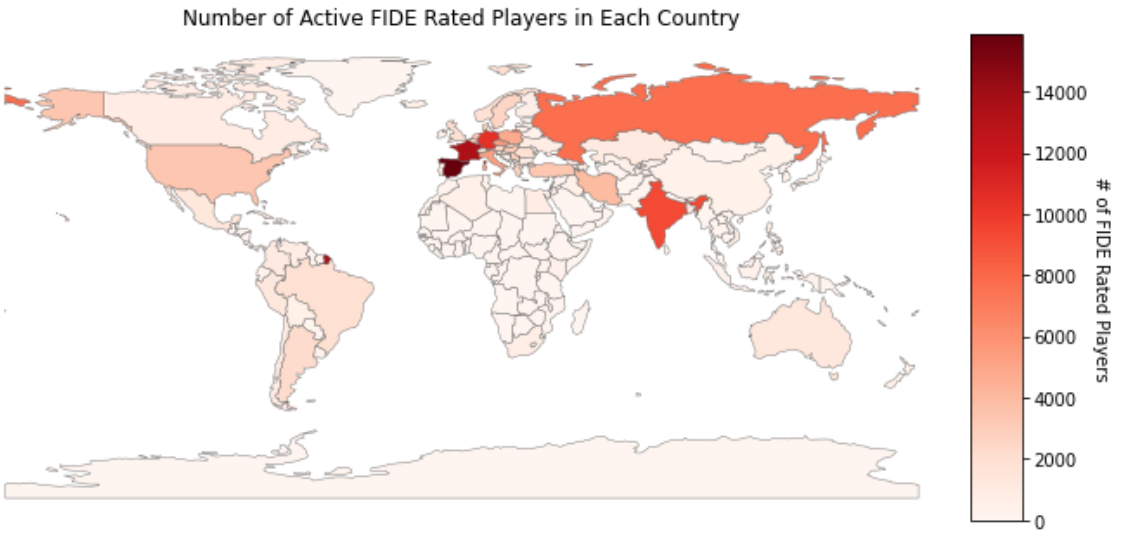


Figure 19: The number of FIDE rated players per country

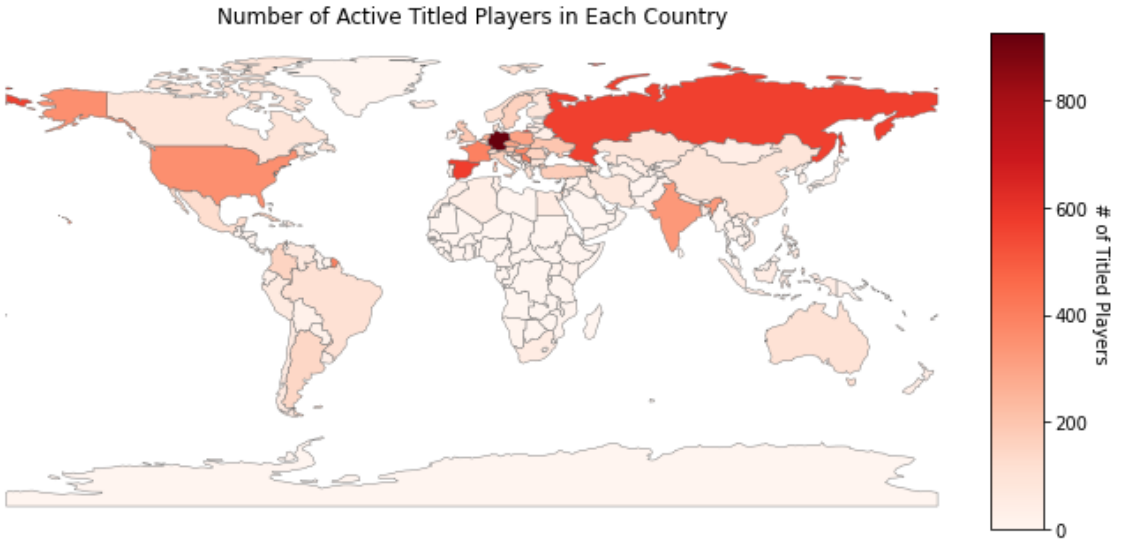


Figure 20: The number of titled rated players per country

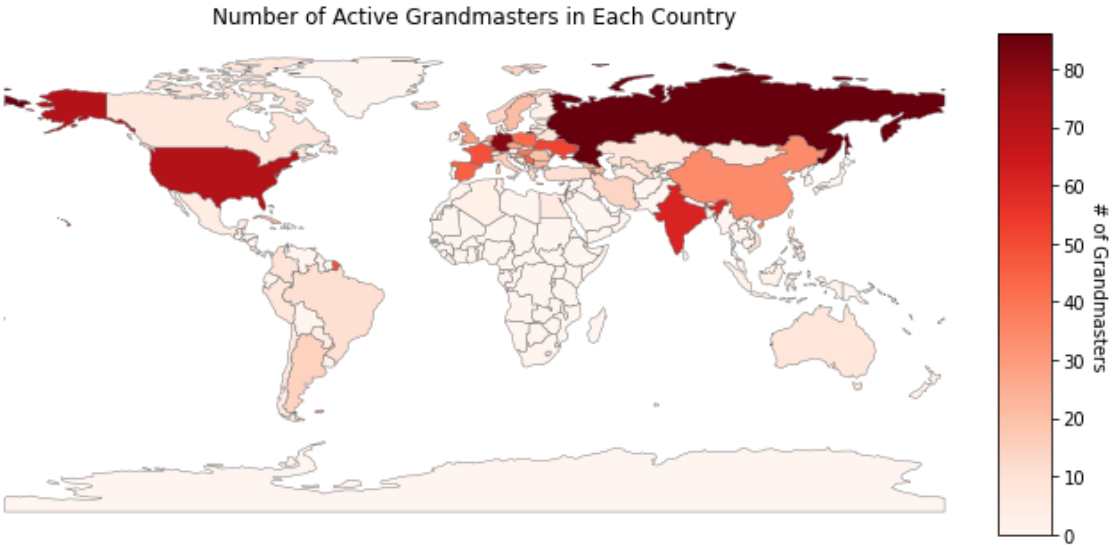


Figure 21: The number of Grandmasters per country

Here, countries with more players are shaded with darker red and countries where there are fewer players are shaded lighter.

There are clearly a great number of FIDE rated players located around Europe, most notably in Spain which has the most players followed by France and Germany which shows us an interesting geographical pattern since these countries are neighbours and creates a line of decreasing number of rated players starting from Spain over to Ukraine.

The country with the most titled players is Germany. An interesting observation, however, is that despite having the most titled players, Germany does not have the most Grandmasters. This belongs to Russia despite have significantly less rated players. Similarly, despite having fewer rated players compared to its neighbours, China has quite a few Grandmasters. These countries have a high density of elite players compared to the other countries.

The drawbacks of these visualizations is that it is hard to see the spread of the ratings. This can be done using a box plot. The following are box plots that outline the spread of ratings in the top 10 countries with the most players for both FIDE rated and titled players.

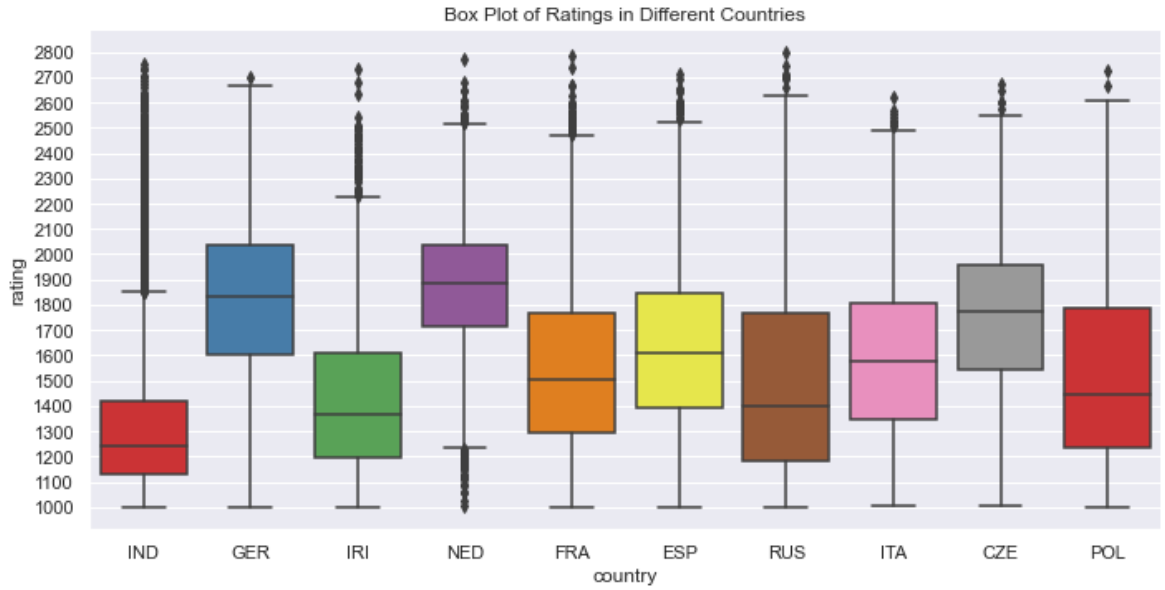


Figure 22: Box Plot of Ratings in The 10 Countries With The Most FIDE Rated Players

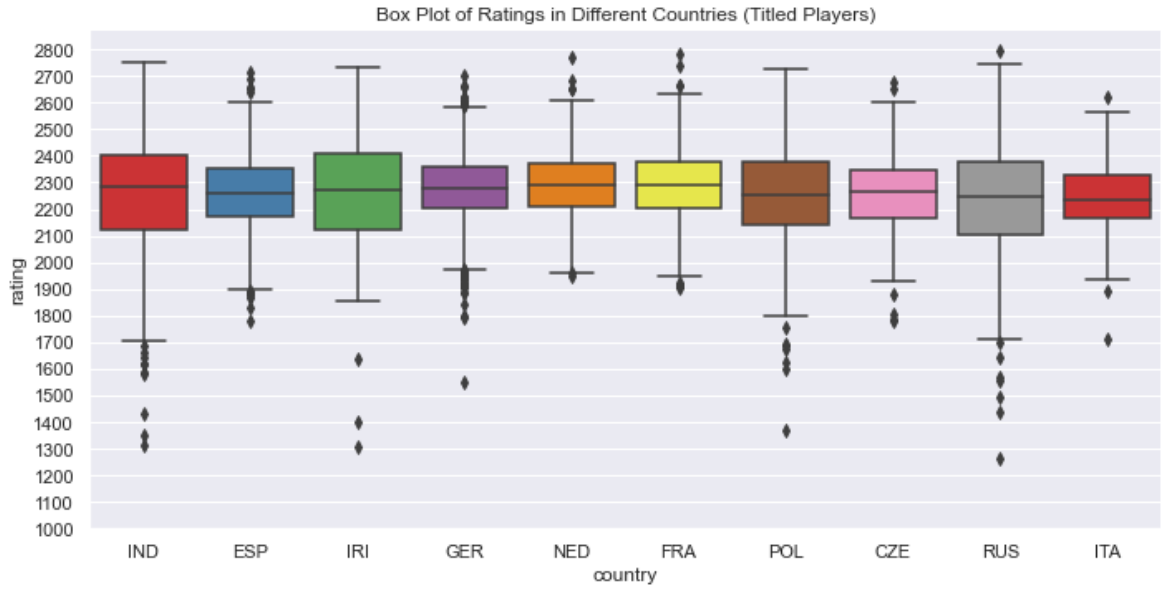


Figure 23: Box Plot of Ratings in The 10 Countries With The Most FIDE Titled Players

These were created in python using the seaborn package where points are outliers if it is less than $Q1 - 1.5 * IQR$ or greater than $Q1 + 1.5 * IQR$ (points within the whiskers contain 99% of the data). This is notable since in fig 22, there are a number of outliers for India (IND) above the maximum whisker. This tells us that India has many rated players but is heavily skewed towards the lower ratings. As we saw in the previous visualizations, Germany has the most FIDE rated players and as seen in fig 22, also has the greatest spread which tells us that Germany has a large variety of players.

In fig 23 we see that titled players across all countries consistently have a median of around 2300 rating. This makes sense as the minimum requirement for being a FIDE Master (3rd highest title out of 4) is having a rating of 2300. An interesting observation is that to qualify for the lowest FIDE title (Candidate Master), you need a minimum rating of 2000, but all countries have lower whiskers that contain ratings below 2000 as well as outliers. This means that a good majority of players have earned a titled but have regressed in skill since then.

Overall, these box plots can be used to provide an overview of the general strength the chess communities in various countries and regions.

Conclusion

To summarize, the visualizations created allow us to find patterns and trends from the data that help us to understand different aspects about chess and its players. Specifically, the relationship between ratings and performance, ratings and geolocation, how pieces are used, and when users use chess.com.

Self-evaluation

I feel like that the visualizations I have created are self-contained, clear to read and has revealed patterns and insights that are interesting to people who are interested in chess. However, I do wish that I had more complex visualizations which I found hard to create given the chess.com dataset's lack of quantitative data. I also wanted to avoid making many histograms/bar graphs so that I had a good variety of visualizations.

I was initially very ambitious with this project where I was going to pull more information from the games using the Lichess API (on a different dataset I was going to use) so that I could get more features to work on, but after making the script I soon found out that the API has a rate limit and would take forever to run on a 20,000 sized dataset. I then thought about analyzing every game on my computer but would take an extremely long time, so I didn't do it. The dataset that was used in this report was then used as it had a little more features and more games.

Pythons pyplot library uses color-blind safe colors already (blue and orange) but for the win-rate bar plot where 3 colors were used, I thought it would be best to find another color palette. So after researching, I used a color-blind safe color palette from the IBM Design Library [4]. I also used a website that simulates color-blindness to see if my plots were legible for all types of color-blindness. In the same vein, all color-bars uses a single color that gradients to white.

I believe that the reports possess the qualities to earn a 5 or better. I have created a variety of interesting visualizations with varying complexity and demonstrated the ability to understand and analyze them with the supporting text. The figures are self-contained and are given appropriate labels and is readable. The method in which the data processed was also discussed.

References

Bibliography

[1] ADITYAJHA1504. 60,000+ chess game dataset (chess.com). URL: https://www.kaggle.com/datasets/adityajha1504/chesscom-user-games-60000-games.

[2] CHESScom. Chess is booming! and our servers are struggling. URL: https://www.chess.com/blog/CHESScom/chess-is-booming-and-our-servers-are-struggling.

[3] International Chess Federation. Standard rating list. URL: https://ratings.fide.com/download_lists.phtml.

[4] David Nichols. Coloring for colorblindness. URL: https://davidmathlogic.com/colorblind/.