Basic Info

Title: “Statistics in the Game of Chess”

Names: Solon Grover, Ben Ruckman, Gavin Thomas
Emails: solon.grover@gmail.com, ben.ruckman@live.com, Gsct2002@gmail.com
UIDs: u1331981, u1247760, u1259629

Github Repo

Background and Motivation

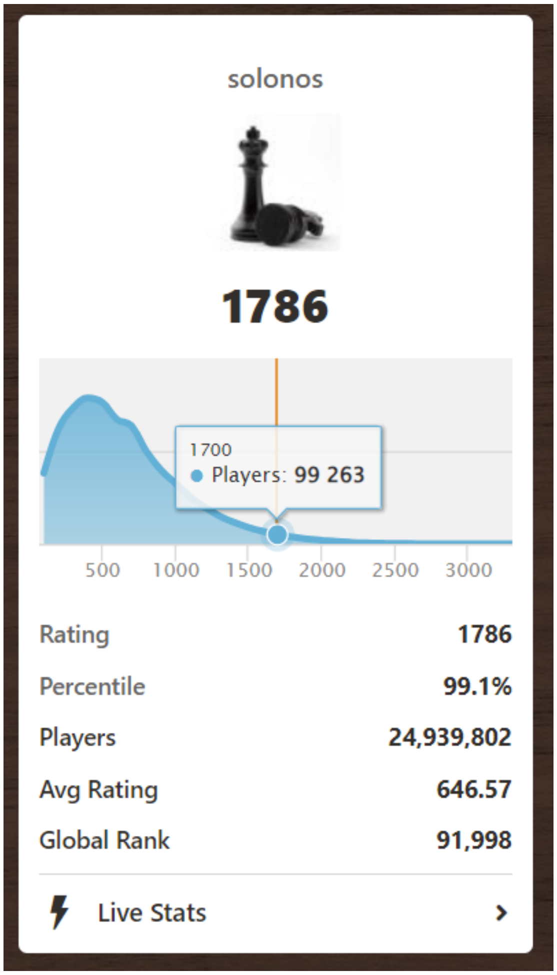
Solon: I’ve been involved in the chess world since elementary school, and lately the game has begun to pick up a bit in popularity, especially after the pandemic. Chess demographics and statistics such as the average age of grandmasters, average peak rating, average rating progression with time, and other related information are things I find to be very interesting as someone who likes to play and study chess regularly. There is a lot of data floating around in the chess world, almost every online game is put into a database and saved. There’s almost infinite variability in the outcomes of a chess game, which can be kind of surprising given that the starting position is, of course, always the same.

Ben: I liked this project idea because I played a lot of chess in Middle School, as well as continuing to play on occasion with family and coworkers today. There’s a lot of statistics in different openings and moves, and I would love to dive deeper into visualization of those statistics!

Gavin: Ever since I heard about IBM’s Deep Blue beating the chess world champion I’ve become much more interested in the game. I have always enjoyed playing but after learning more about that and how complicated the game can be, I wanted to learn more about it. I think that the high amount of game states and statistics and the massive amount of history around the game make it a great candidate for a data visualization project.

Related Work

Chess.com's website contains stats for individual players. Here is an example of a normal curve representing the number of chess.com players at each rating level. The user is able to drag their mouse along the curve, and the exact number of players at that rating threshold is displayed. Other interesting statistics are detailed below the chart, such the player's rating percentile, the total number of players in the whose data is contained in the chart, the average rating, and the global rank of the player.

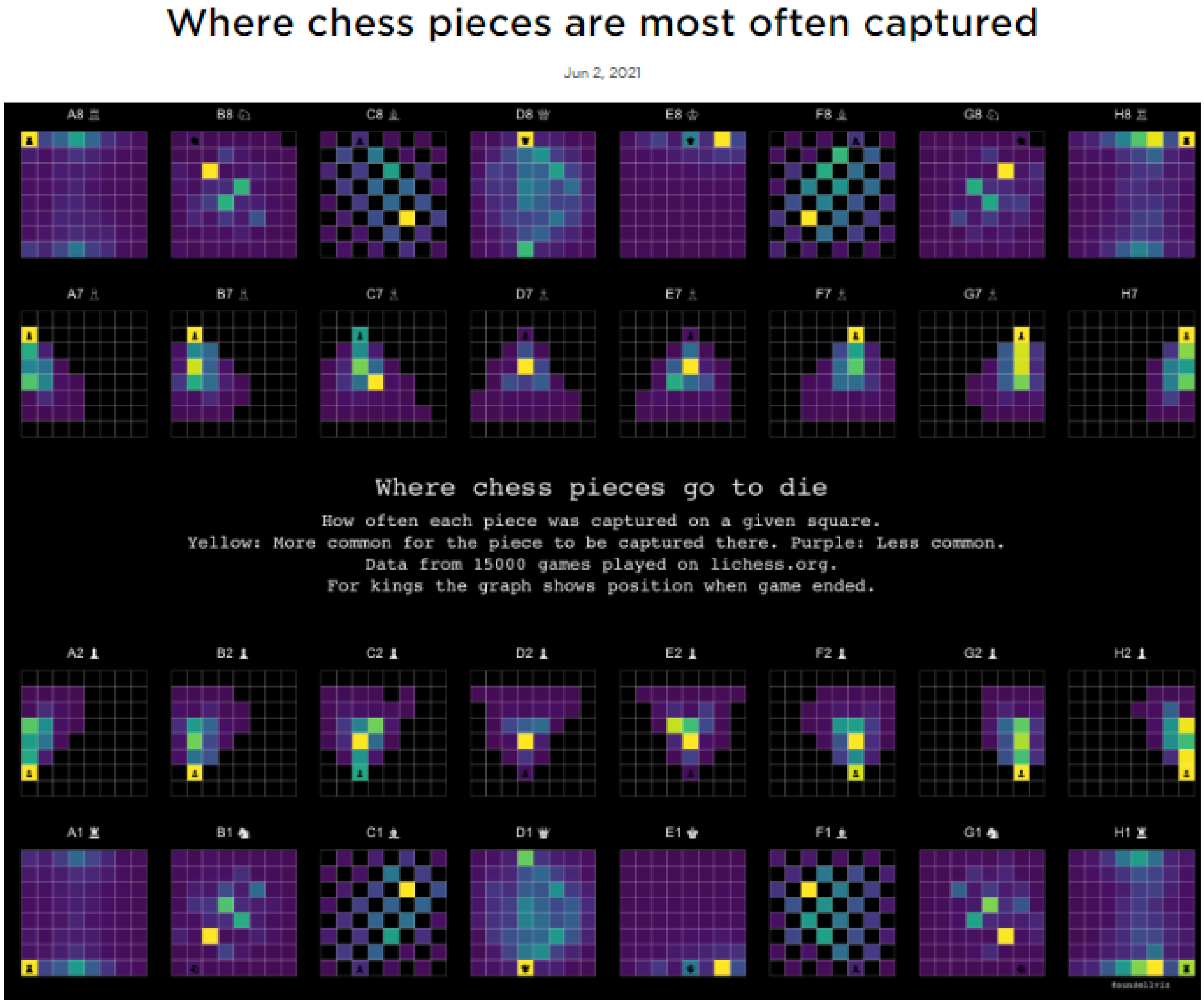


solonos

1786

Rating	1786
Percentile	99.1%
Players	24,939,802
Avg Rating	646.57
Global Rank	91,998

https://flowingdata.com/2021/06/02/where-chess-pieces-are-most-often-captured/
Here is a visualization made which shows where each chess piece most often gets captured. The heat map is something that we’re trying to implement as one of our chess board features where we show the most successful and least successful opening moves for white. Additionally, we could also use this type of visualization to show how common these moves are.


Where chess pieces are most often captured

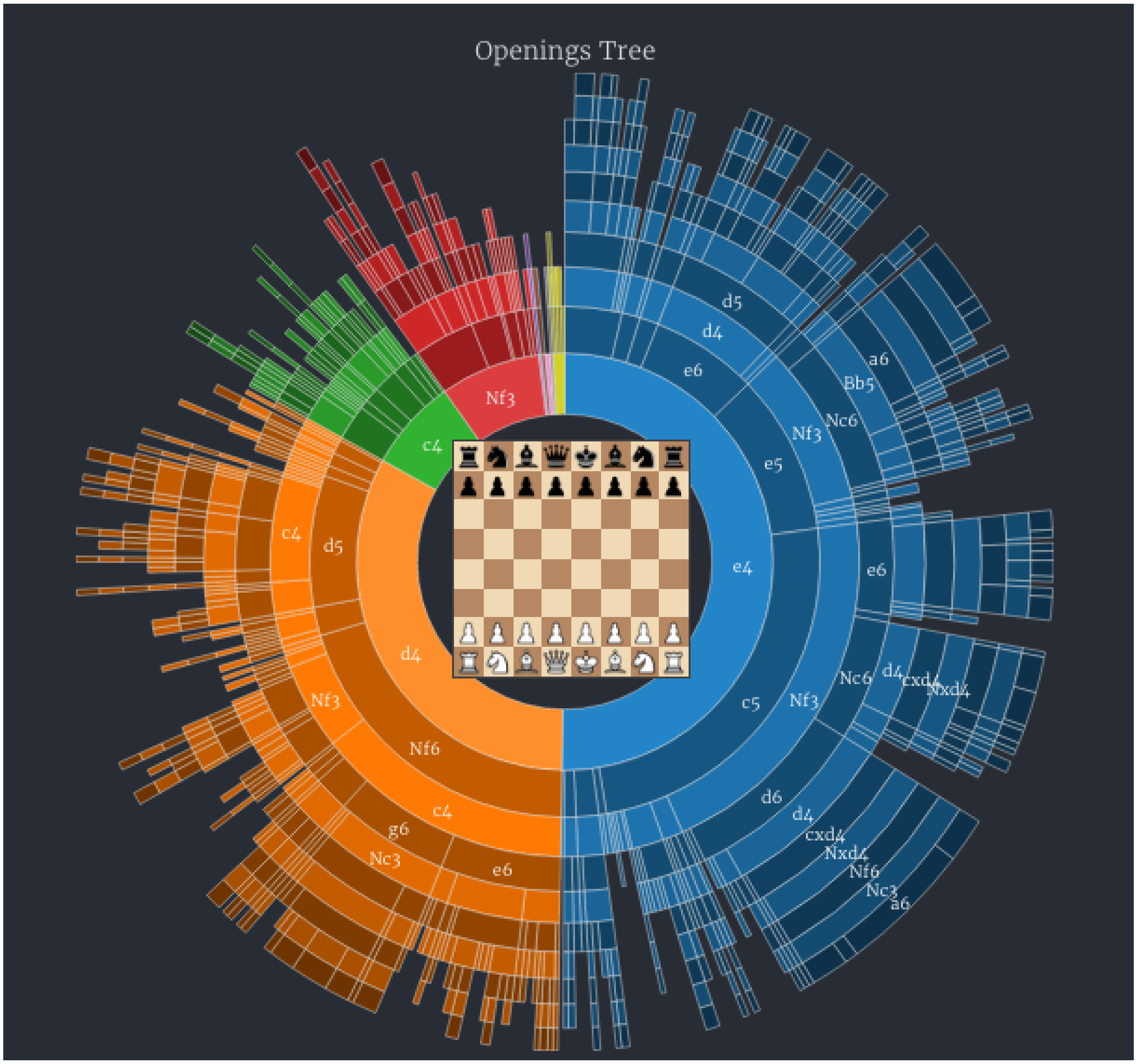
Jun 2, 2021

Where chess pieces go to die

How often each piece was captured on a given square.
Yellow: More common for the piece to be captured there. Purple: Less common.
Data from 15000 games played on lichess.org.
For kings the graph shows position when game ended.

https://www.informationisbeautifulawards.com/showcase/1271-a-visual-look-at-2-million-chess-games

This visualization uses concentric rings and shades of these rings to show the proportion of all openings each move is played. With each level, the number of possible moves branches out quickly, showing how the incredible variety of chess game outcomes looks visually. We were thinking of incorporating some move frequency data in our main visualization, and this gives us some inspiration as to how we could display that data.



Project Objectives

Questions:
-GM questions
-What is the average age that grandmasters achieve their title?
-How many games has the average grandmaster played?
-What is the average IQ of a grandmaster?
-At what age do grandmasters typically start playing chess?
-What is the average age of an individual's peak rating achievement?

-How long does it take to improve 100 elo rating points given an individual's current rating?
-Does playing chess exclusively result in a better or worse rating than performing puzzles, and studying games alongside regular play?
-What does the average player's chess rating look like over the course of their lifetime?
-What is the male to female ratio of chess players?
-What is the black vs white win/draw/loss ratio (general population vs professionals)?
-What are the openings with the highest win ratio? Openings with the worst win ratio? (We could use a heatmap of the board, where the opening moves have a red to green tint based on how likely they are to result in a win)
-What is the average centi-pawn loss per move at different rating levels?
-Where are chess players located?
-How many turns are in games (Distribution of the ratio of games that end at certain move counts)?

Benefits: For players and those interested to gain a better understanding of the demographics and related statistics of the game of chess. There will be some grandmaster data, which most people find interesting, since grandmasters are extreme outliers amongst the general population.

Data

Chess.com has an api to view a ton of chess data.
https://www.chess.com/news/view/published-data-api#pubapi-general

Lichess has a ton of games that we can download to get average stats.
https://database.lichess.org/#standard_games

Large grandmaster/top 100 database about the actual players (birthday, etc).
https://fide.com/

https://www.pnas.org/doi/10.1073/pnas.2006653117
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DZC0MT

Data Processing

We are going to have to do a substantial amount of data cleanup, as chess.com's main way to get stats is going directly through player ids. Then we would want to aggregate those stats, and display them in a nice way to users.

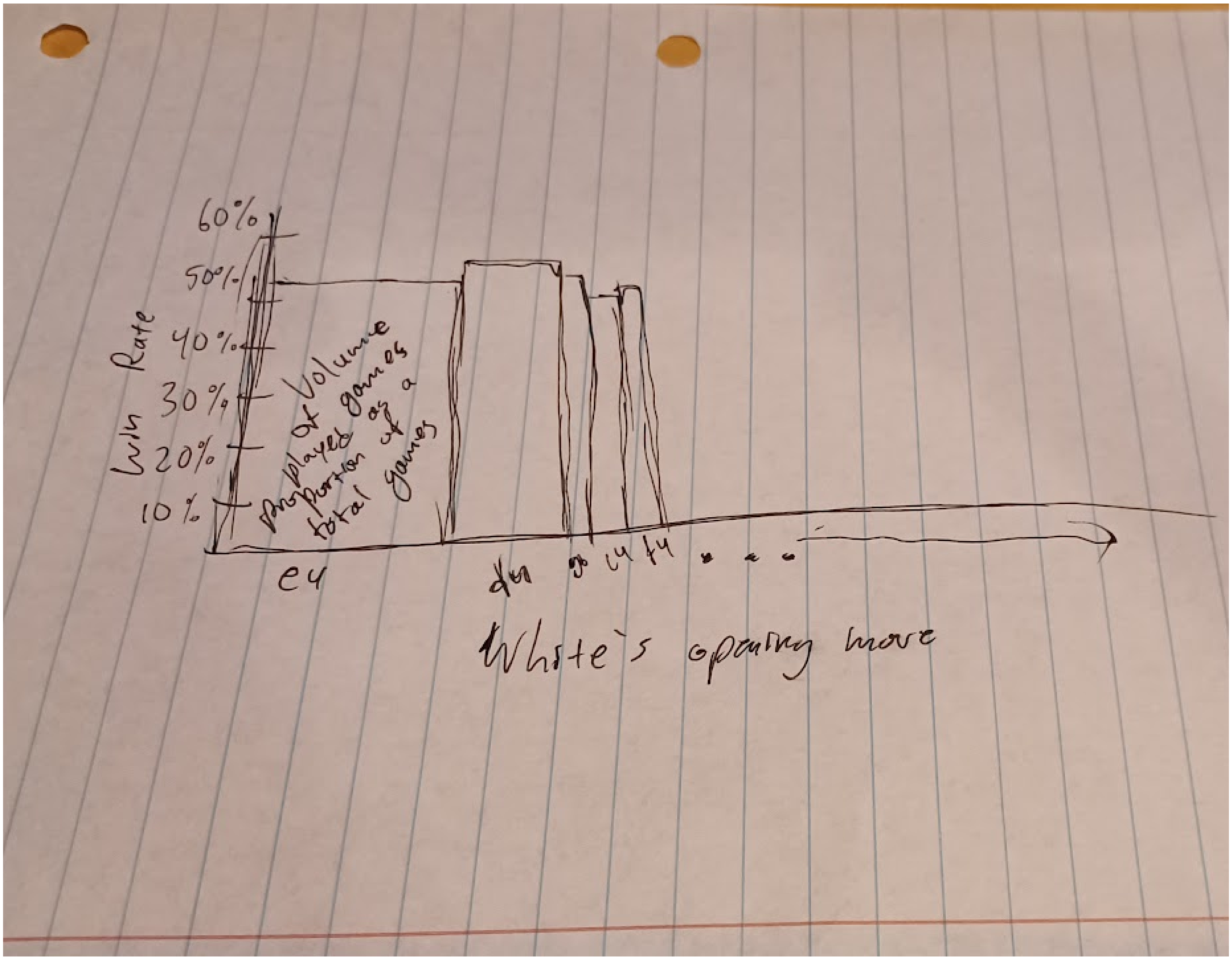
Lichess has a ton of games (almost 5 billion) available to download. There would be a lot of data processing required to get the data/stats we want out of it. We can split up the games based on rating to give us rating specific data that we can use.

Most of our data processing thus far has been processing the lichess data. We are doing this in python, by loading the files into our code, and then creating dictionaries of aggregate statistics per month, and writing these dictionaries to files to be used by the frontend. Another way we have tried processing data is by creating a Python Pandas dataframe with columns for each attribute we need. We can then more easily process the data and perform operations on it to export to the front end.

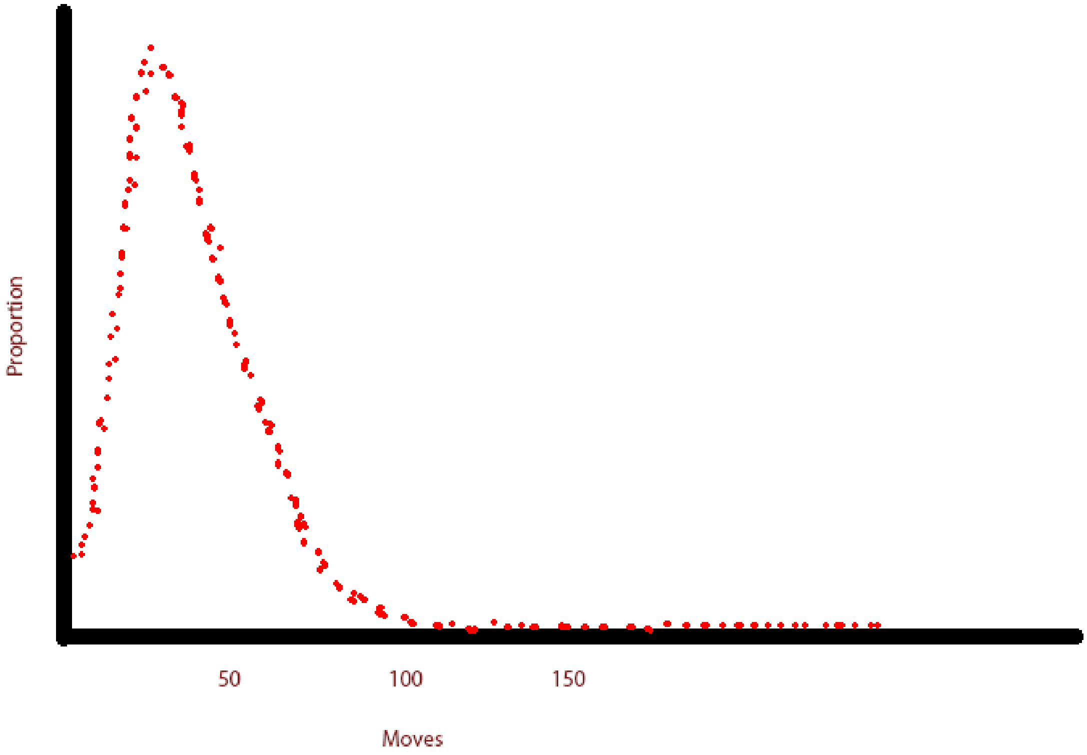
Because of the sheer number of games, and size of the files, it takes a lot of time, memory, and compute to go through all of the games in each file, and we are only up to 2015 (starting in 2013) as of October 28th. It's possible we won't be able to get through all of the data through present day before the end of the project.

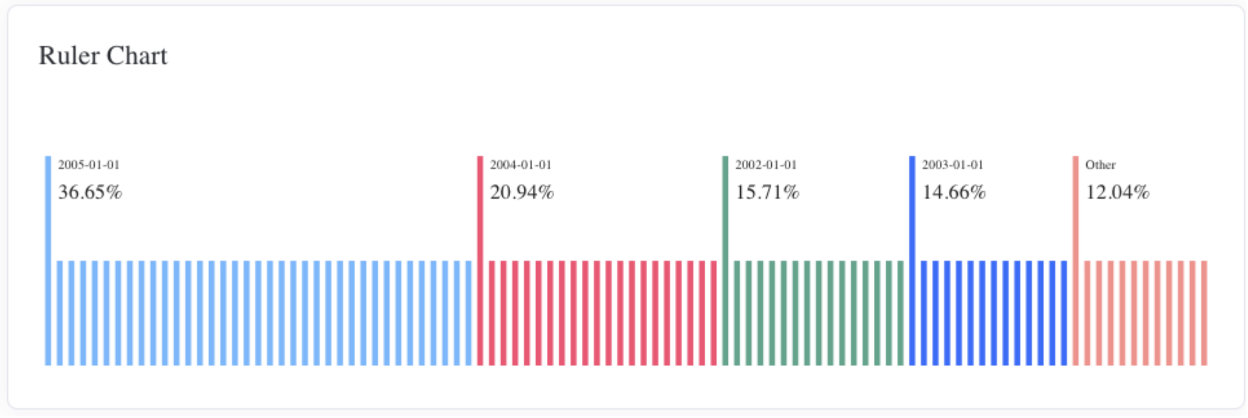
Visualization Design

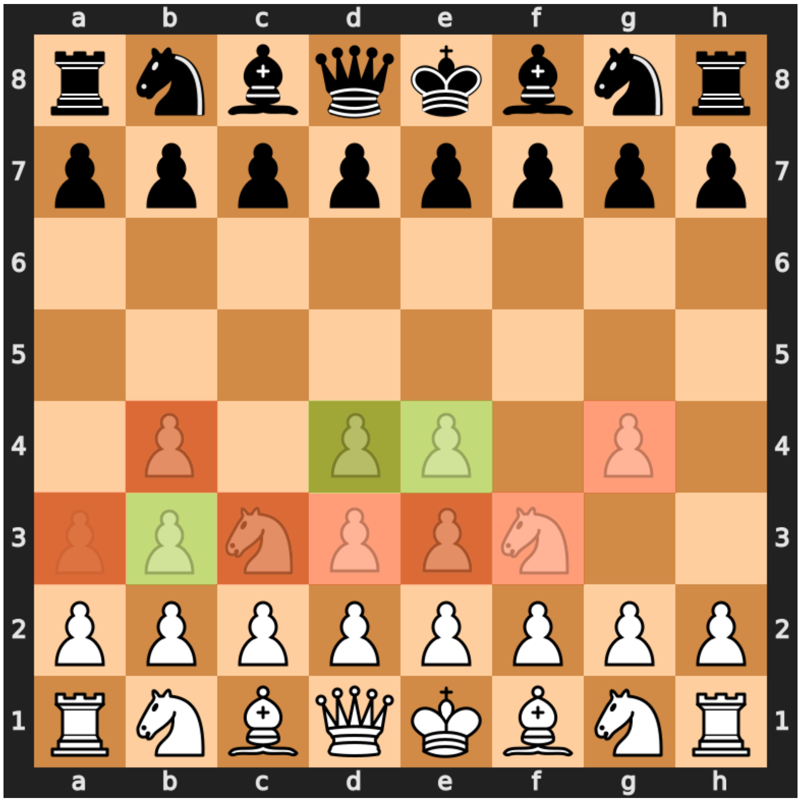
How many turns are in games (Distribution of the ratio of games that end at certain move counts)?

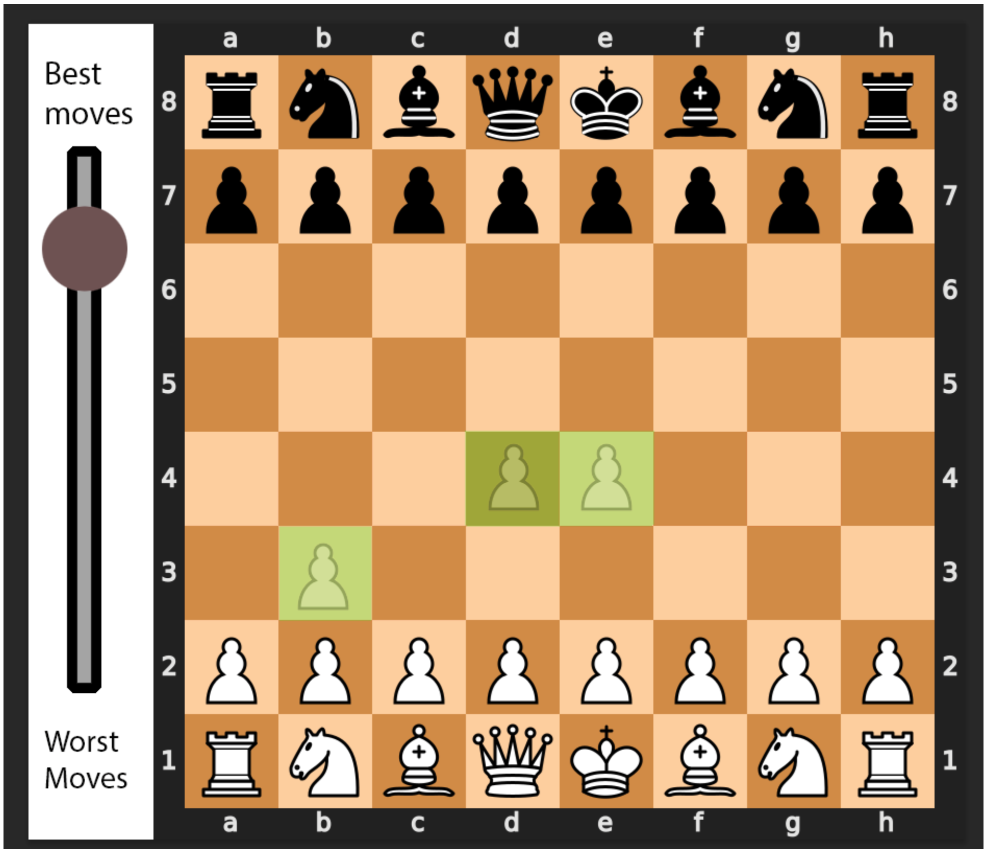


Distribution of the number of moves in completed chess games

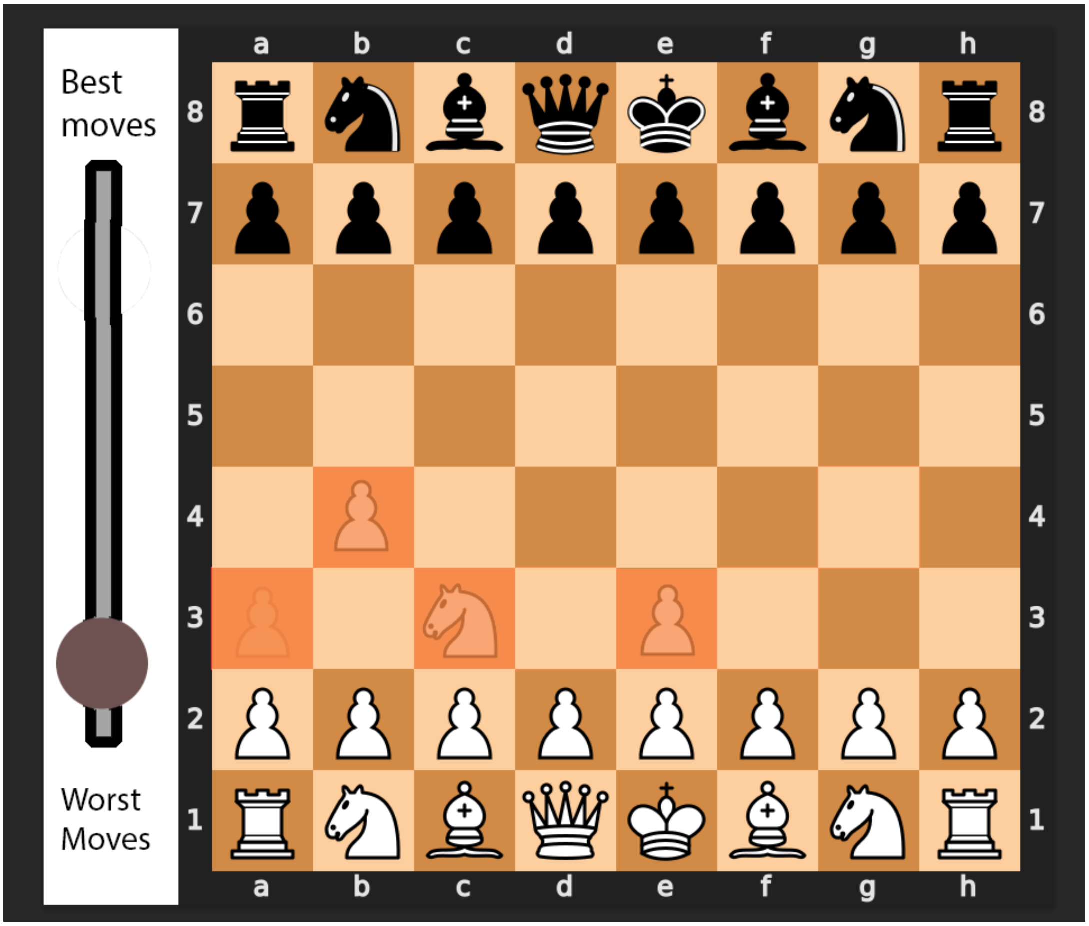








Best moves
Worst Moves

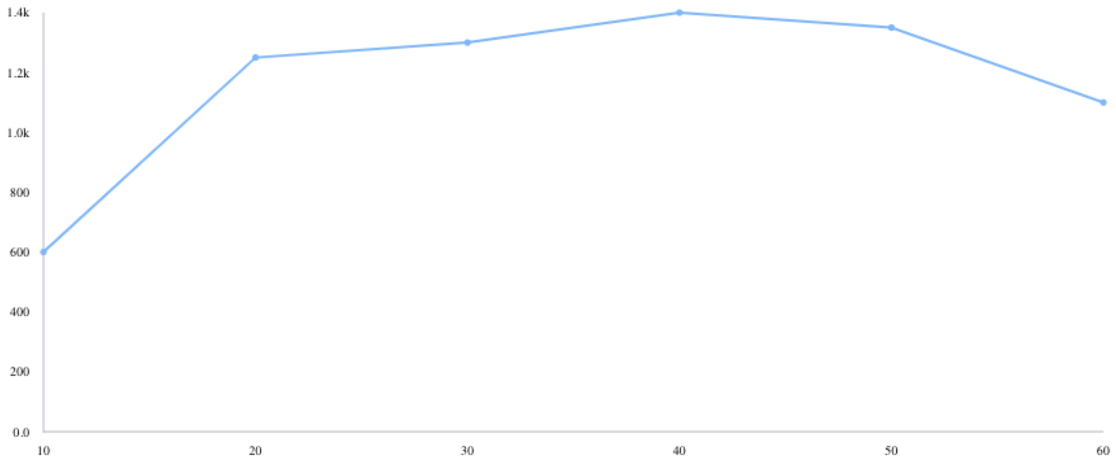


Some visualizations we have might be a heatmap on a chess board, to display “the best openings”, or other interesting stats about chess moves. This can be shown as the opening moves having a red to green tint based on how likely they are to result in a win. We could do multiple boards, perhaps one highlighting all the pawn moves and knight moves in the other. We could also do this with responses from black given a particular white move.

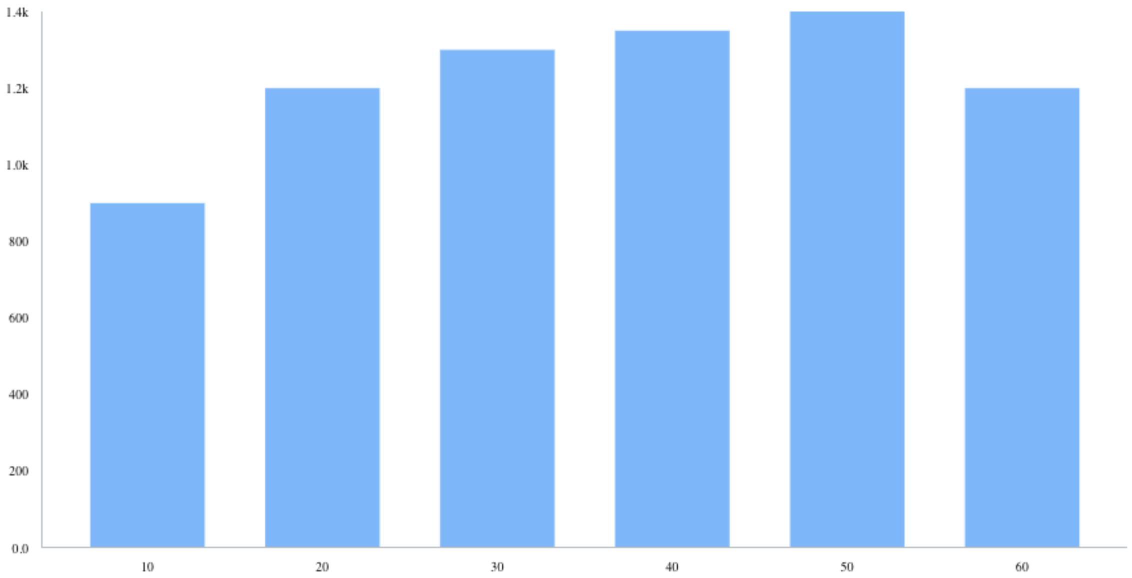
We can add a “Filter by ranking” picker to the side of the chess board visualization, that would allow the user to see what the best opening moves are by ranking.

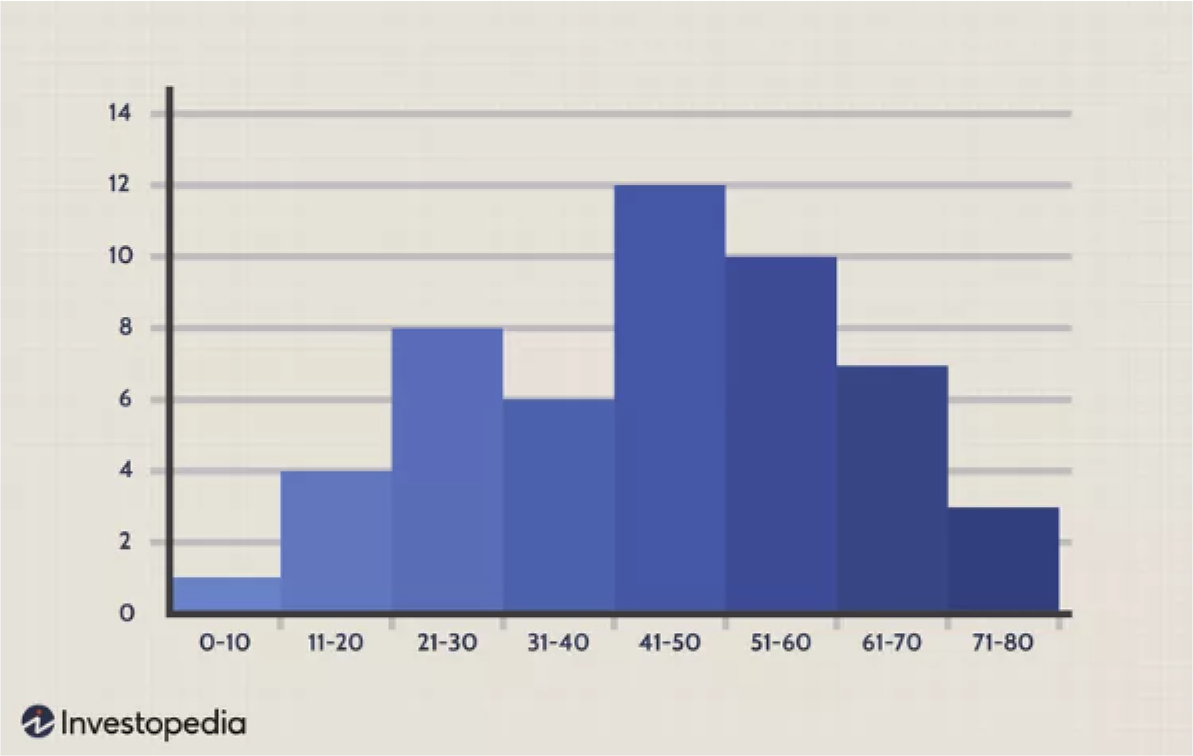
We would also want to use a generic bar/line chart for some of these stats, to visualize the distribution of turn counts, and average rating over the average ages.

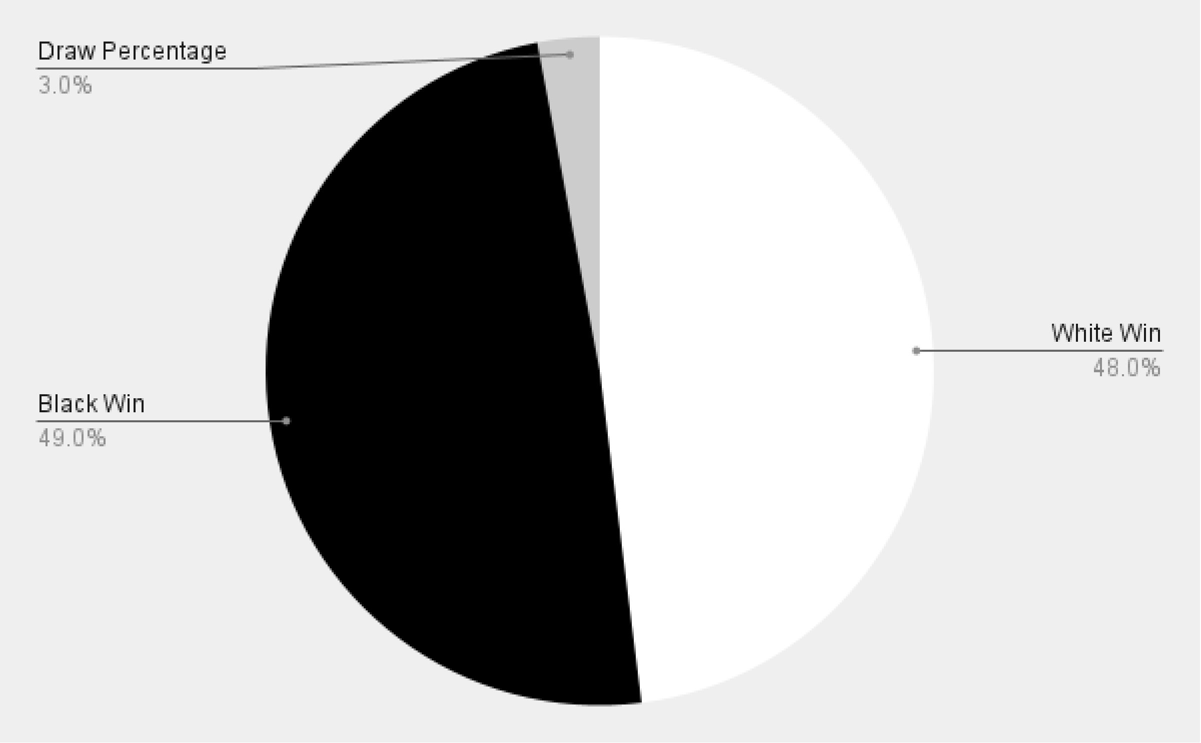
Player Elo by Ages



Average Player Elo by Ages



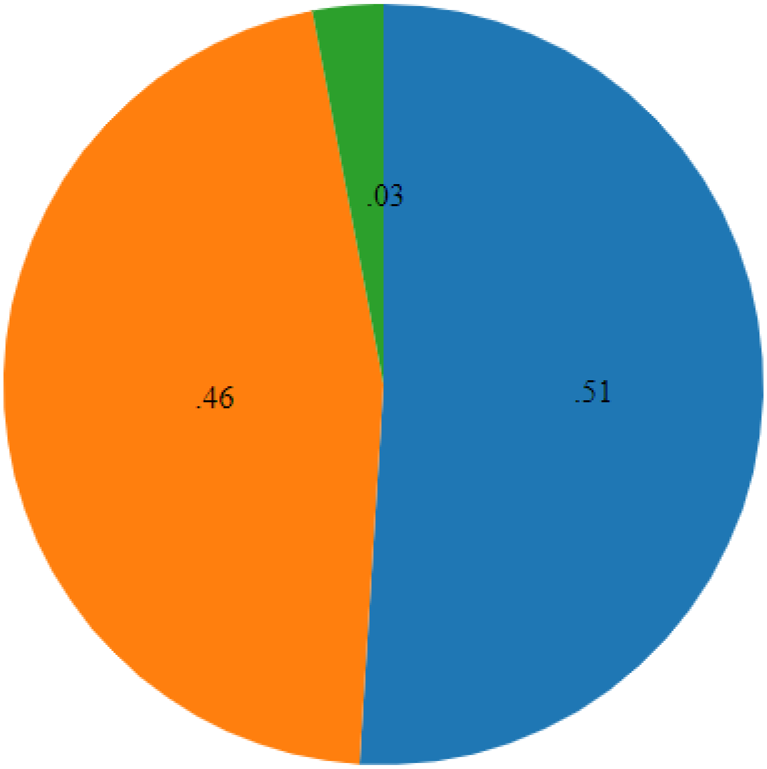

14
12
10
8
6
4
2
0
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
Investopedia


Draw Percentage
3.0%
White Win
48.0%
Black Win
49.0%

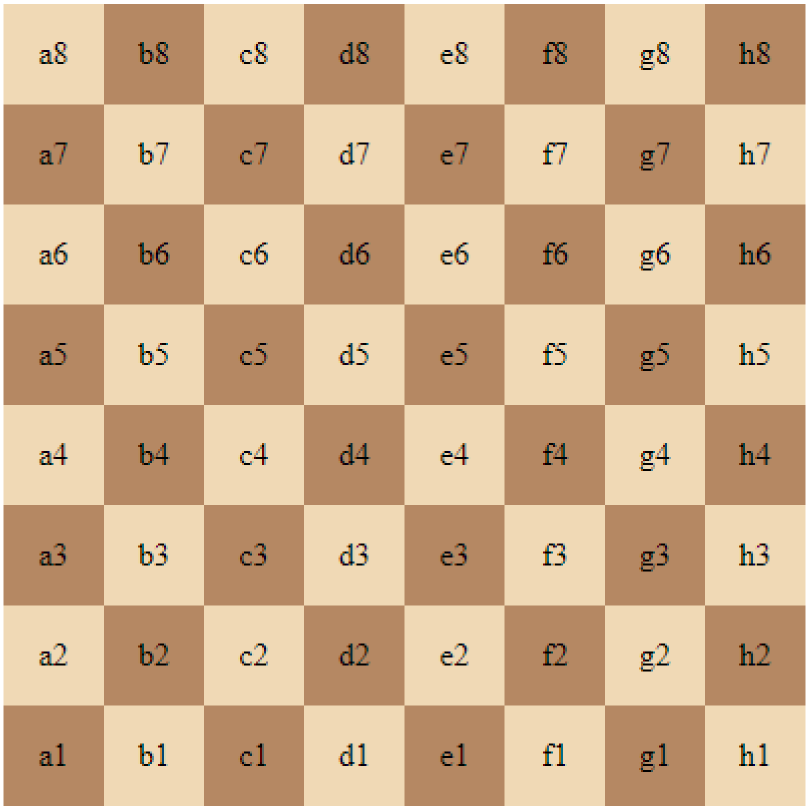
Design Evolution

What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

First iteration of the pie chart!

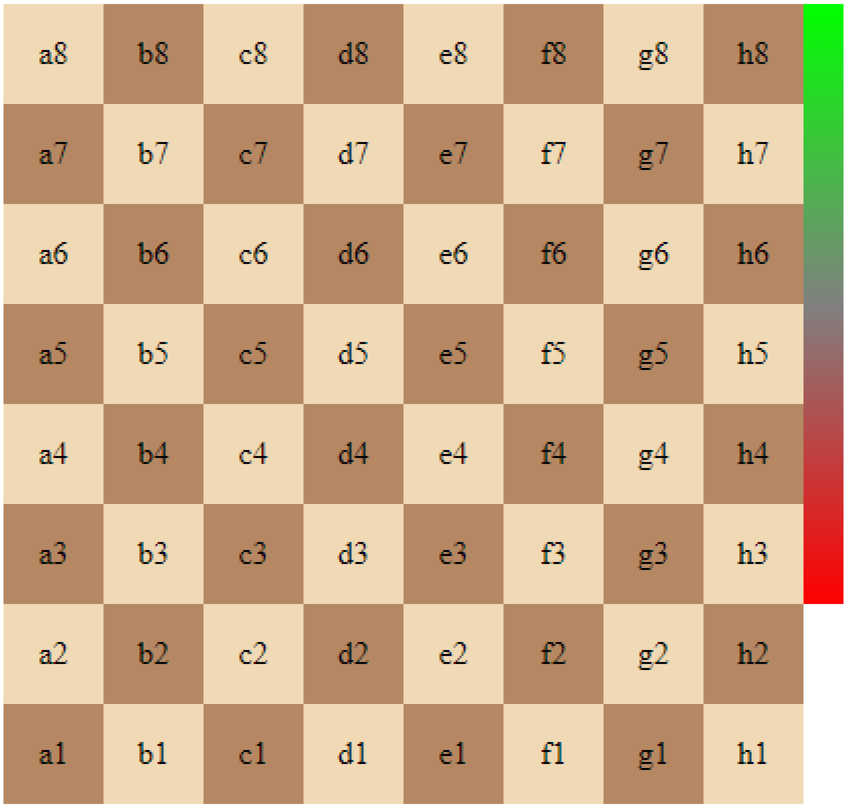


First iteration of the chess board



a8	b8	c8	d8	e8	f8	g8	h8
a7	b7	c7	d7	e7	f7	g7	h7
a6	b6	c6	d6	e6	f6	g6	h6
a5	b5	c5	d5	e5	f5	g5	h5
a4	b4	c4	d4	e4	f4	g4	h4
a3	b3	c3	d3	e3	f3	g3	h3
a2	b2	c2	d2	e2	f2	g2	h2
a1	b1	c1	d1	e1	f1	g1	h1

Chess board with slider
Talking with the TA for our project proposal gave us the idea to show the most popular moves over time periods, as well as to show specific time periods “best moves”



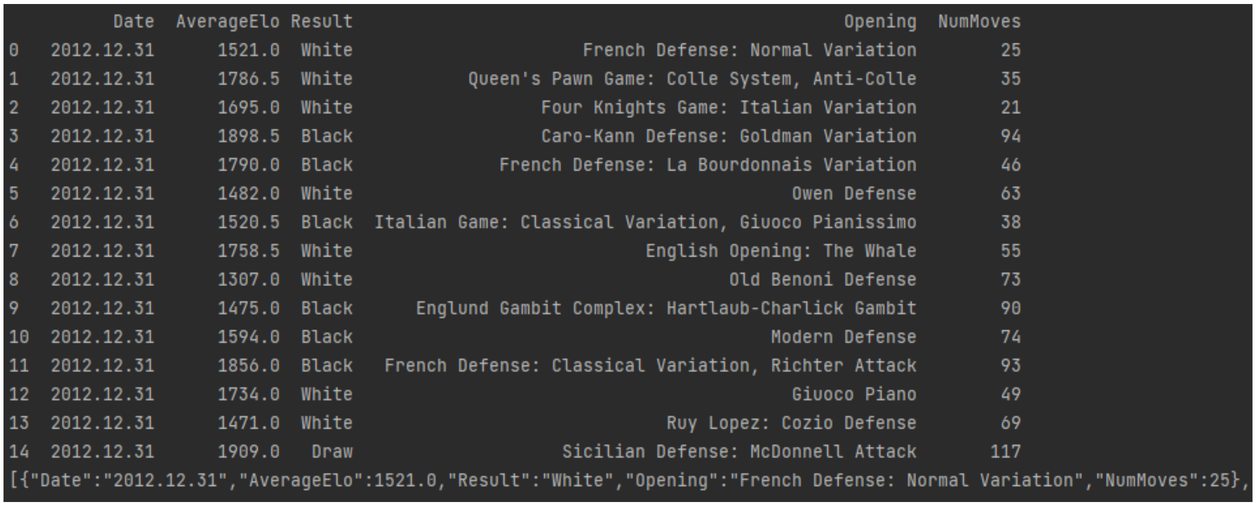
First iteration of aggregated data example

```
 "11, 27": {
   "count": 257,
   "wins": 1
 },
…
```

Current aggregated data format

```
{
 "12, 28": {
   "count": 579,
   "wins": 284,
   "win_percentage": 0.4905008635578584
 },
 "11, 27": {
   "count": 257,
   "wins": 164,
   "win_percentage": 0.6381322957198443
 },
 …
```

Dataframe created to represent each game



```
          Date  AverageElo Result                                            Opening  NumMoves
0   2012.12.31      1521.0  White                        French Defense: Normal Variation        25
1   2012.12.31      1786.5  White              Queen's Pawn Game: Colle System, Anti-Colle        35
2   2012.12.31      1695.0  White                   Four Knights Game: Italian Variation        21
3   2012.12.31      1898.5  Black                    Caro-Kann Defense: Goldman Variation        94
4   2012.12.31      1790.0  Black               French Defense: La Bourdonnais Variation        46
5   2012.12.31      1482.0  White                                           Owen Defense        63
6   2012.12.31      1520.5  Black  Italian Game: Classical Variation, Giuoco Pianissimo        38
7   2012.12.31      1758.5  White                            English Opening: The Whale        55
8   2012.12.31      1307.0  White                                     Old Benoni Defense        73
9   2012.12.31      1475.0  Black      Englund Gambit Complex: Hartlaub-Charlick Gambit        90
10  2012.12.31      1594.0  Black                                         Modern Defense        74
11  2012.12.31      1856.0  Black   French Defense: Classical Variation, Richter Attack        93
12  2012.12.31      1734.0  White                                           Giuoco Piano        49
13  2012.12.31      1471.0  White                               Ruy Lopez: Cozio Defense        69
14  2012.12.31      1909.0   Draw                      Sicilian Defense: McDonnell Attack       117
[{"Date":"2012.12.31","AverageElo":1521.0,"Result":"White","Opening":"French Defense: Normal Variation","NumMoves":25},
```

Length of games aggregated data

```
[
{"length": 53, "frequency": 1877},
{"length": 55, "frequency": 1807},
…
```

Must-Have Features

Visualizations that answer the following questions

-What is the black vs white win/draw/loss ratio (general population vs professionals)?

-What are the openings with the highest win ratio? Openings with the worst win ratio?

-How many turns are in games (Distribution of the ratio of games that end at certain move counts)? Should look something like this

Optional Features

GM questions
-What is the average age that grandmasters achieve their title?
-How many games has the average grandmaster played?
-What is the average IQ of a grandmaster?
-At what age do grandmasters typically start playing chess?

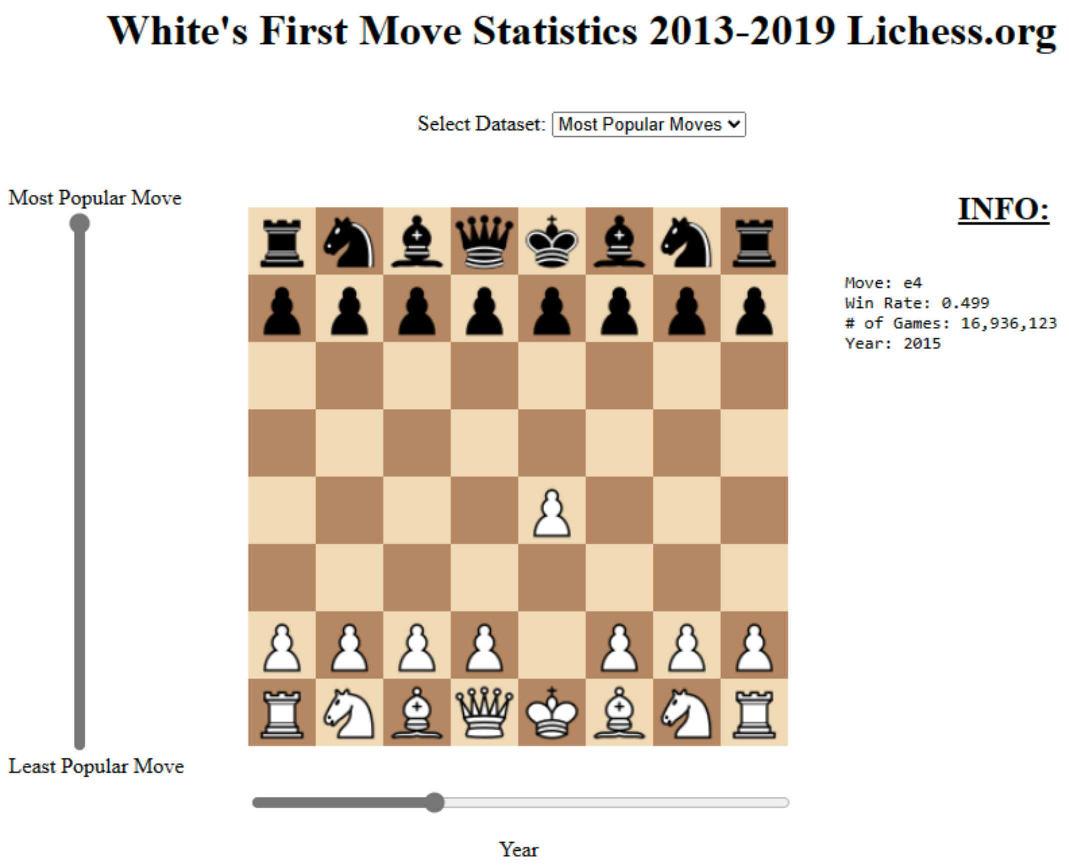
-What is the male to female ratio of chess players?
-What is the average centi-pawn loss per move at different rating levels?
-Where are chess players located?

Average Elo over average age line graph to answer:
-What does the average player’s chess rating look like over the course of their lifetime?
-What is the average age of an individual’s peak rating achievement?
-How long does it take to improve 100 elo rating points given an individual’s current rating?

-Does playing chess exclusively result in a better or worse rating than performing puzzles, and studying games alongside regular play?

Implementation

The goal of this visualization was to show people what the best and most popular first moves are, over a time period using a significant sample size of games. We achieved this with the sliders to change the year, as well as to give them the most to least popular/best moves. When you change either of the sliders, the data that is referenced is changed, and updates the chess board to show you the move that fits your criteria.

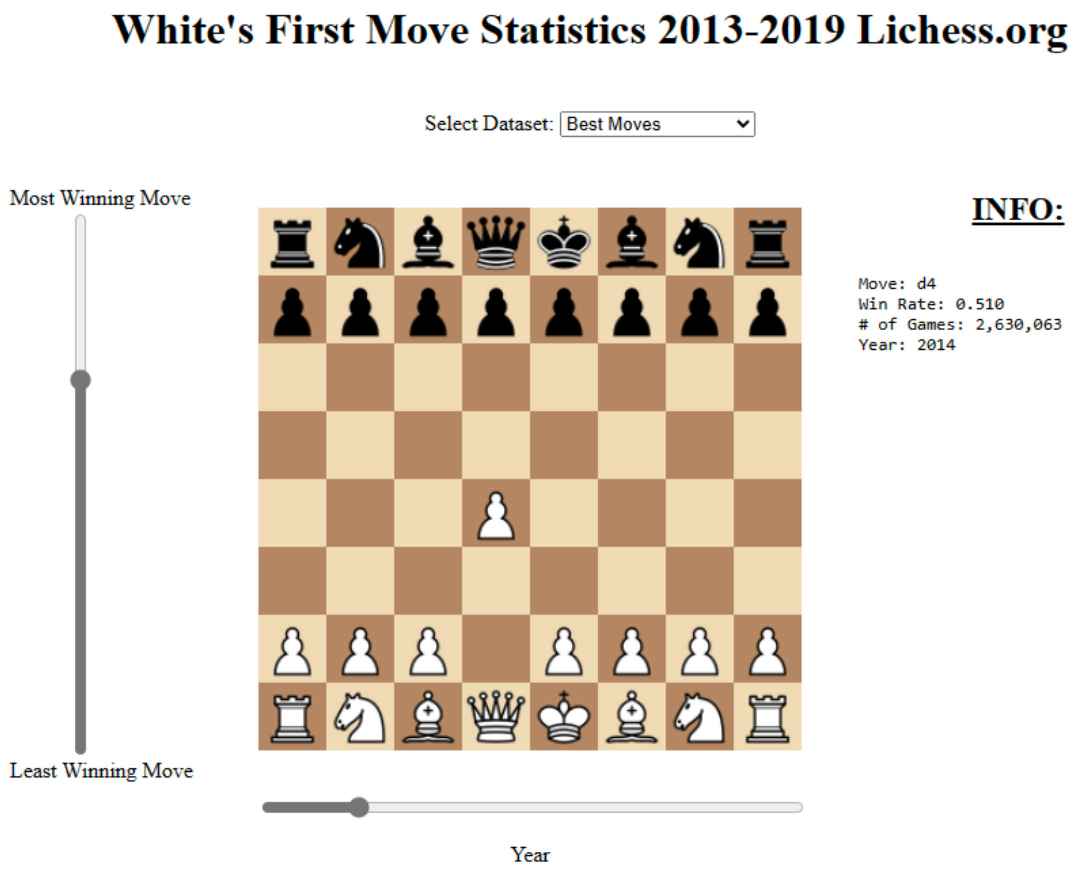


White's First Move Statistics 2013-2019 Lichess.org

Select Dataset: Most Popular Moves

INFO:

```
Move: e4
Win Rate: 0.499
# of Games: 16,936,123
Year: 2015
```



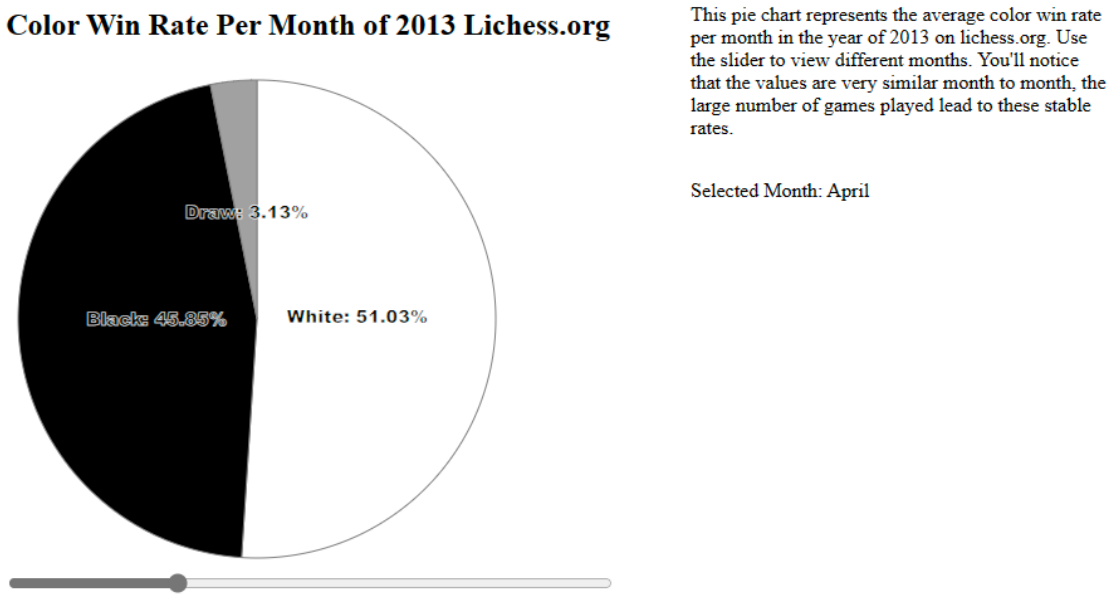
White's First Move Statistics 2013-2019 Lichess.org

Select Dataset: Best Moves

INFO:

```
Move: d4
Win Rate: 0.510
# of Games: 2,630,063
Year: 2014
```

The goal of this visualization was to show black vs white win rates over time. We achieved this with a pie chart, that shows the percentages that white and black win, as well as the percentage of ties in the selected month. We used 2013, as we didn't have time to process the rest of the data that we had. For reference, it took nearly all semester to get through 2013-2019 for the chess visualization, and we had the data up to 2023. Despite the lack of other years data, it still shows a pretty clear cut conclusion, that white wins more than black.

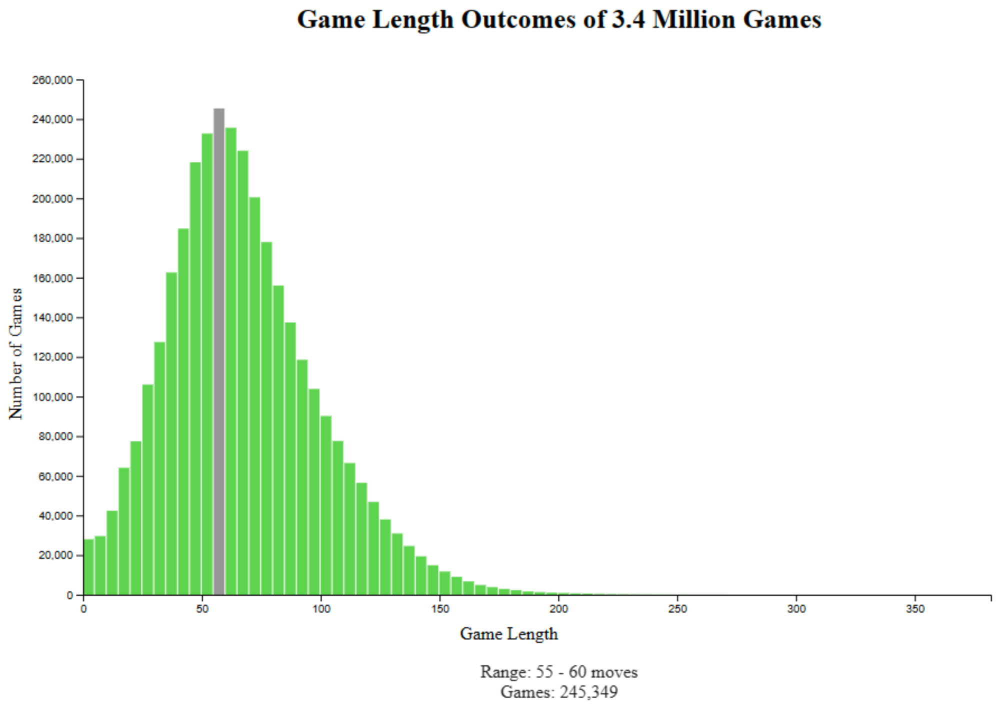


Color Win Rate Per Month of 2013 Lichess.org

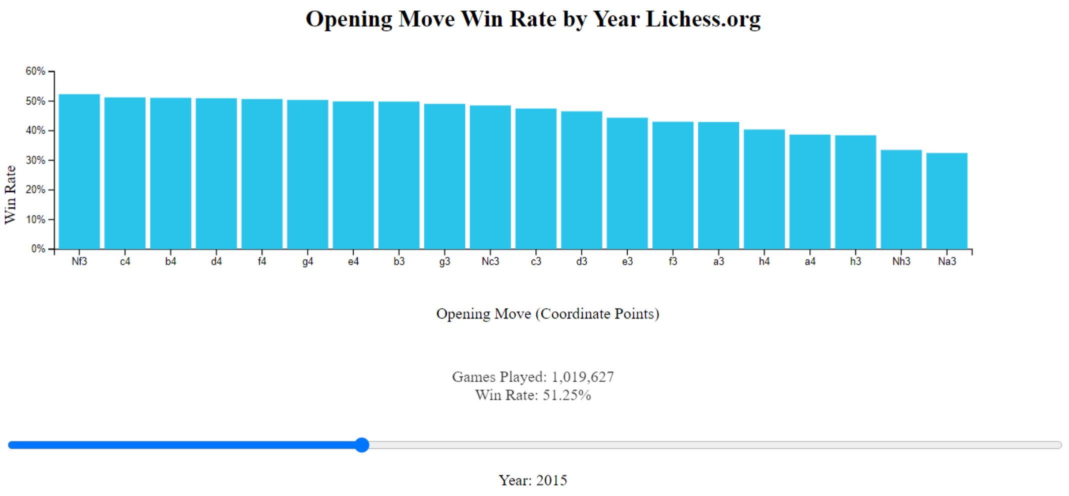
This pie chart represents the average color win rate per month in the year of 2013 on lichess.org. Use the slider to view different months. You'll notice that the values are very similar month to month, the large number of games played lead to these stable rates.

Selected Month: April

The goal of this visualization was to show the distribution of game length. The histogram does a pretty good job at showing the distribution of game lengths, with it being fairly obvious that most games are between 40-80 moves long. We used 3.4 million games as our sample size, as that's enough games to rule out any outliers in the data, but not too many that we have to spend more time processing data that wouldn’t show much else (this processing still took a couple of days).



The goal of this visualization was to display the win rate of different opening moves. This bar chart allows users to click on a bar to display the win rate of the associated opening move. The visualization makes it easy to see what moves tend to perform better than others.



Project Schedule

Meeting at 10am every Saturday
Written standup on Wednesday - couple sentences what you're doing, any blockers etc.

Week 4 - Figure out primary data sources, how we want to structure the data, how we’re going to use the data to build the visualizations
Week 5 - Collect data and build visualizations for **primary** features
Week 6 - Collect data and build visualizations for **primary** features
Week 7 - Collect data and build visualizations for **optional** features
Week 8 - Fall Break
Week 9 - Finish ui visualizations and retrieving data
Week 10 - Begin using collected data on the visualizations
Week 11 - Finish hooking up data to visualizations
Week 12 - Setup the website, put visualizations with real data there
Week 13 - Website should be “done”
Week 14 - Thanksgiving Break
Week 15 - QA

Things we need to do:
Build ui for visualizations - Solon and Ben
Get data from various API’s and downloads, and structure/clean it - Ben and Gavin
Use data to build visualizations we want - Gavin and Ben
Set up basic website to show the visualizations on - Solon and Gavin
Process book - Everyone!

Evaluation

Using our chess board visualization, we are able to look at the most popular and highest win rate starting moves from year to year. We learned that the “best” starting move varied between years. The first 2 years were moving different pawns, while the last years we had all showed moving the knight to F3. We also learned that the “worst” starting move always involved the outside of the board, rather than the inside of the board - like moving a knight to the border of the board, instead of to the middle.

The bar chart also shows a win rate percentage for different starting moves. This is using the same data as the chess board visualization and shows a similar result. However, we like the chess board visualization better, as it shows the result in a much more readable way.

We are also able to look at each color’s win rate data from and game length distribution from games played on lichess in 2013. This was interesting, as it showed white as the clear winner in win rates. If you can choose to play as white instead of black, you should!

These visualizations work well in displaying the previously mentioned data but some of them could benefit from using more data or more recent data. The pie chart and bar chart use a smaller data set from 2013 but they could provide more interesting results if given more games as they could display how data changes during a longer period of time.