Basic Info. The project title, your names, e-mail addresses, UIDs, a link to the project repository.

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

Background and Motivation. Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

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.

Project Objectives. Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

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. From where and how are you collecting your data? If appropriate, provide a link to your data sources.

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. Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?

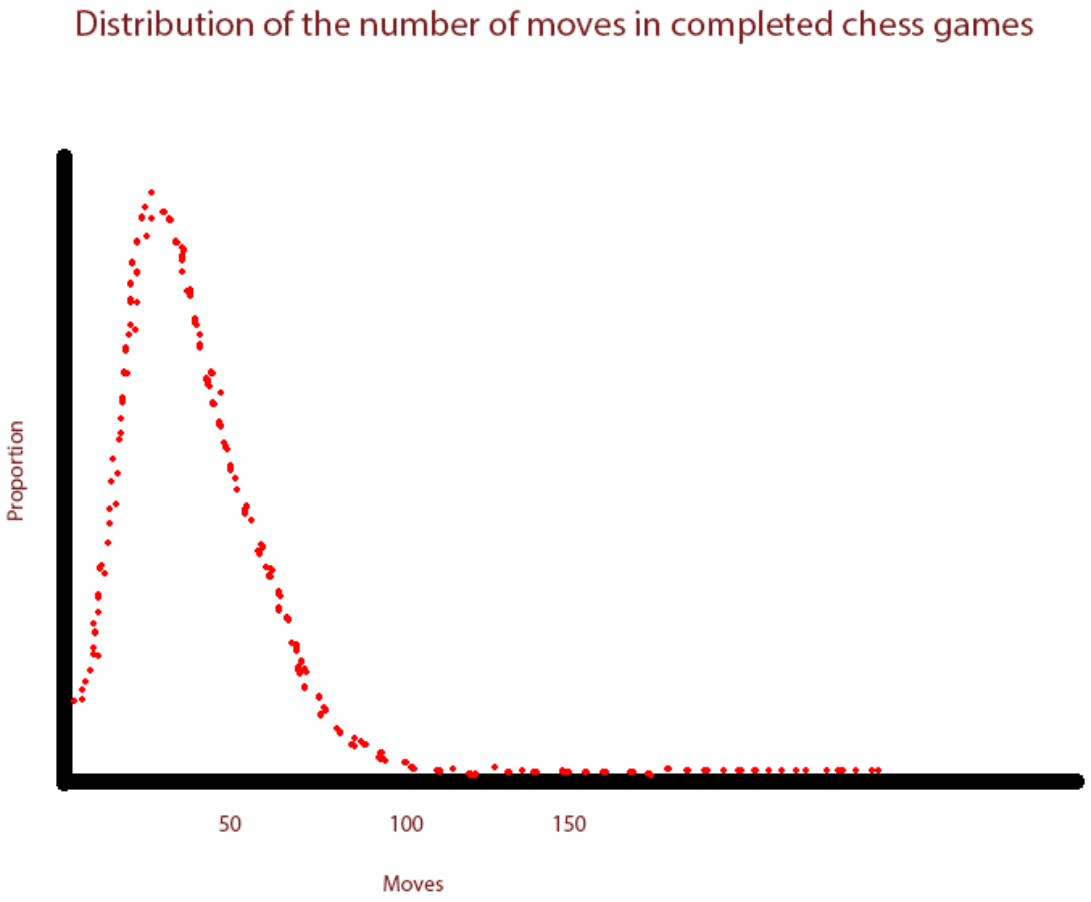
We are going to have to do a substantial amount of data cleanup, as chess.com's main way to get stats is going directly through player ids. Then we would want to aggregate those stats, and display them in a nice way to users.

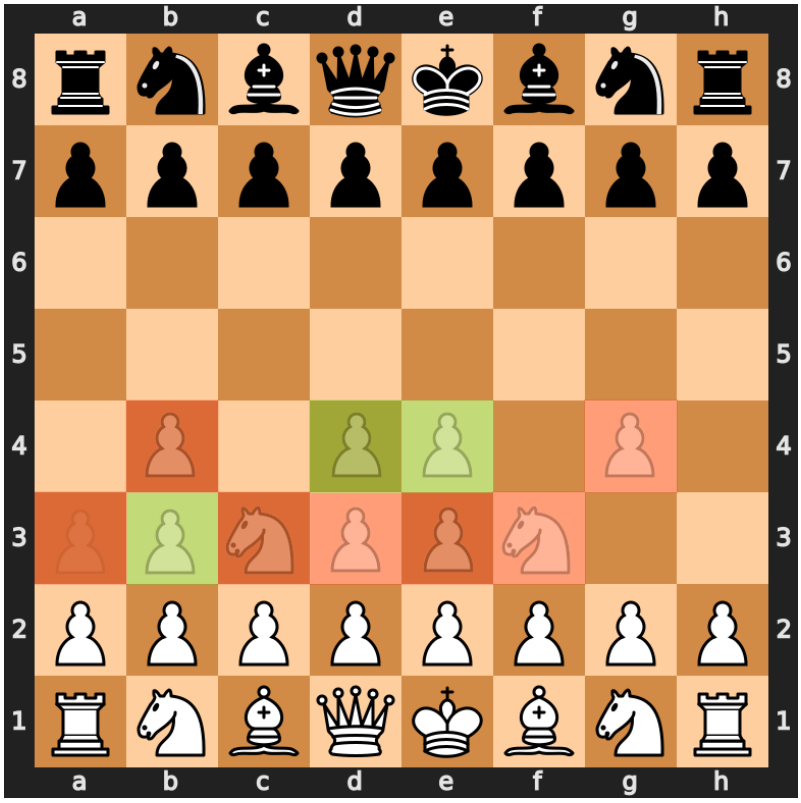
Lichess has a ton of games (almost 5 billion) available to download. There would be a lot of data processing required to get the data/stats we want out of it. We can split up the games based on rating to give us rating specific data that we can use.

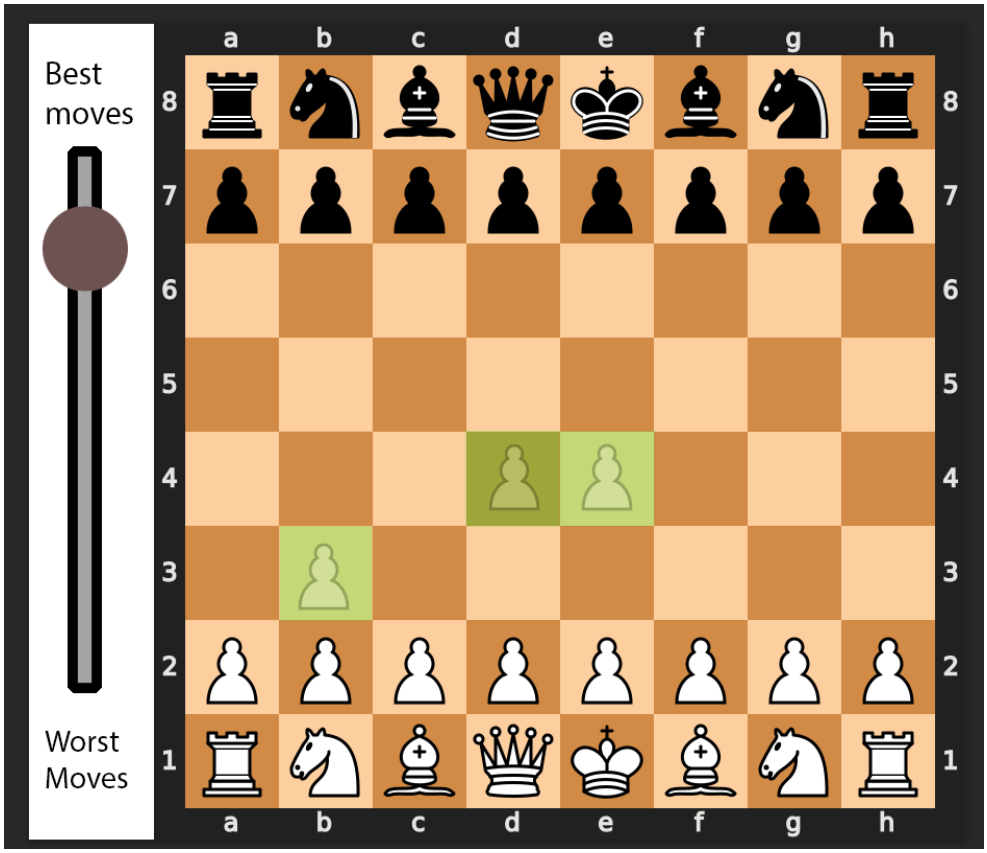
Visualization Design. How will you display your data? Provide some general ideas that you have for the visualization design. Develop three alternative prototype designs for your visualization. Create one final design that incorporates the best of your three designs. Describe

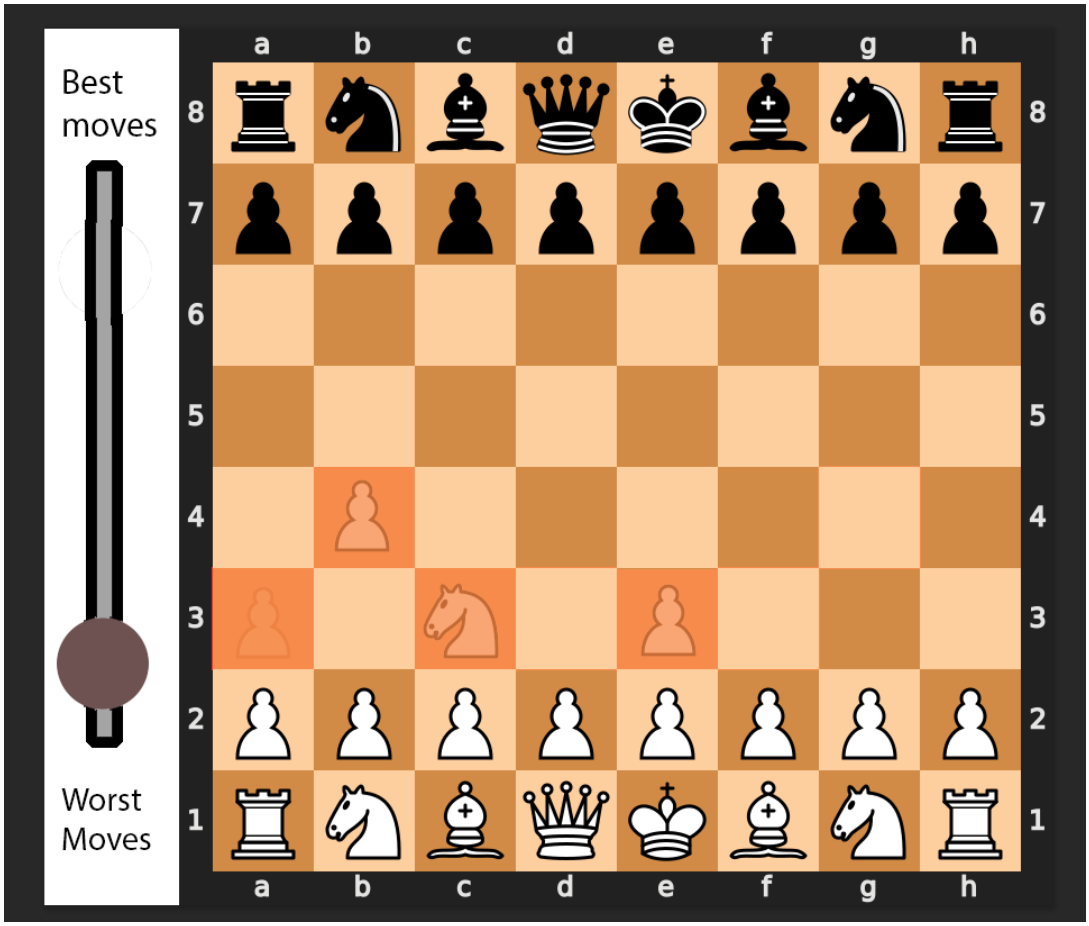
your designs and justify your choices of visual encodings. We recommend you use the [Five Design Sheet Methodology].

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




a b c d e f g h
8 7 6 5 4 3 2 1


Best moves
Worst Moves
a b c d e f g h
8 7 6 5 4 3 2 1

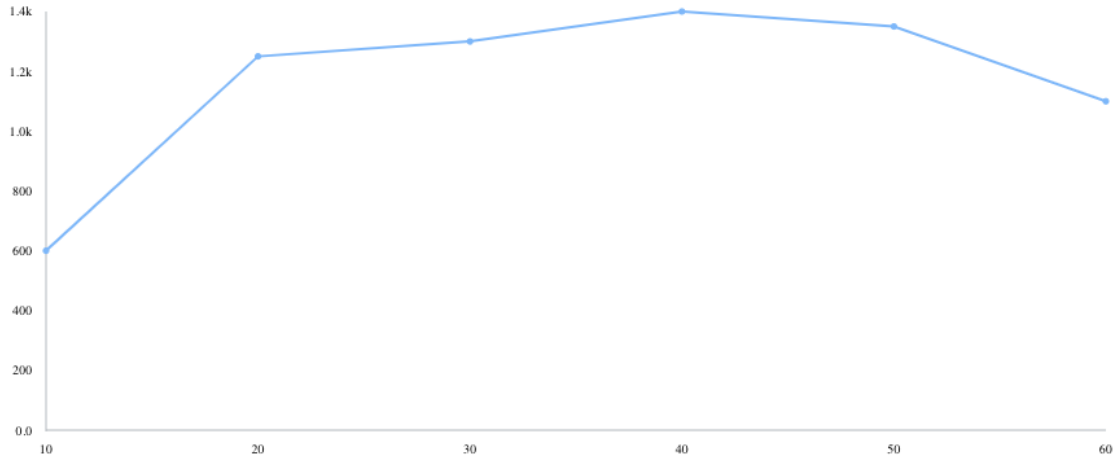


Some visualizations we have might be a heatmap on a chess board, to display “the best openings”, or other interesting stats about chess moves. This can be shown as the opening moves having a red to green tint based on how likely they are to result in a win. We could do multiple boards, perhaps one highlighting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

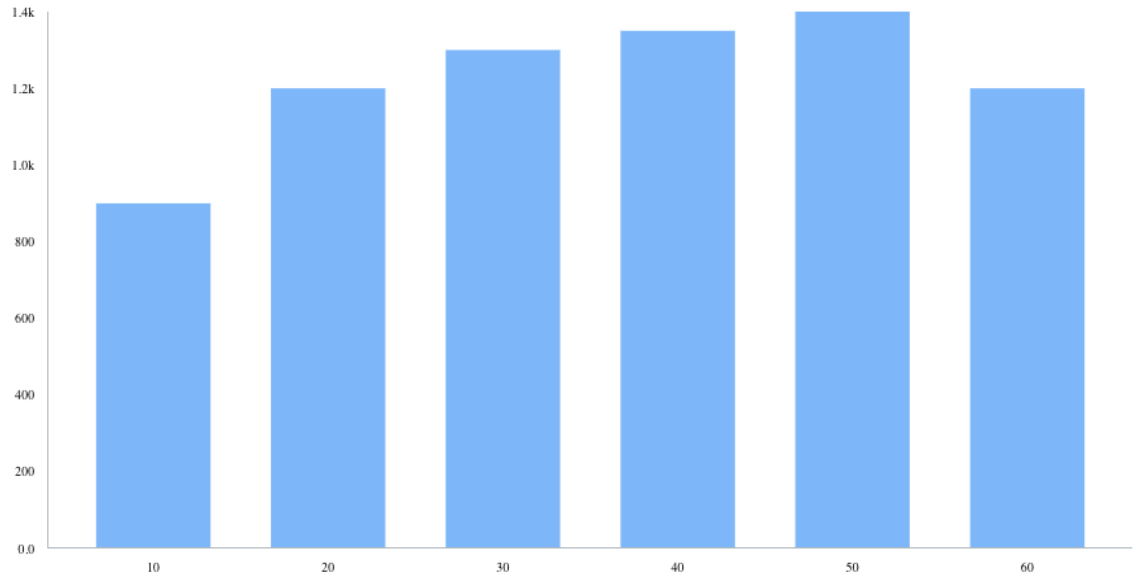
We can add a “Filter by ranking” picker to the side of the chess board visualization, that would allow the user to see what the best opening moves are by ranking.

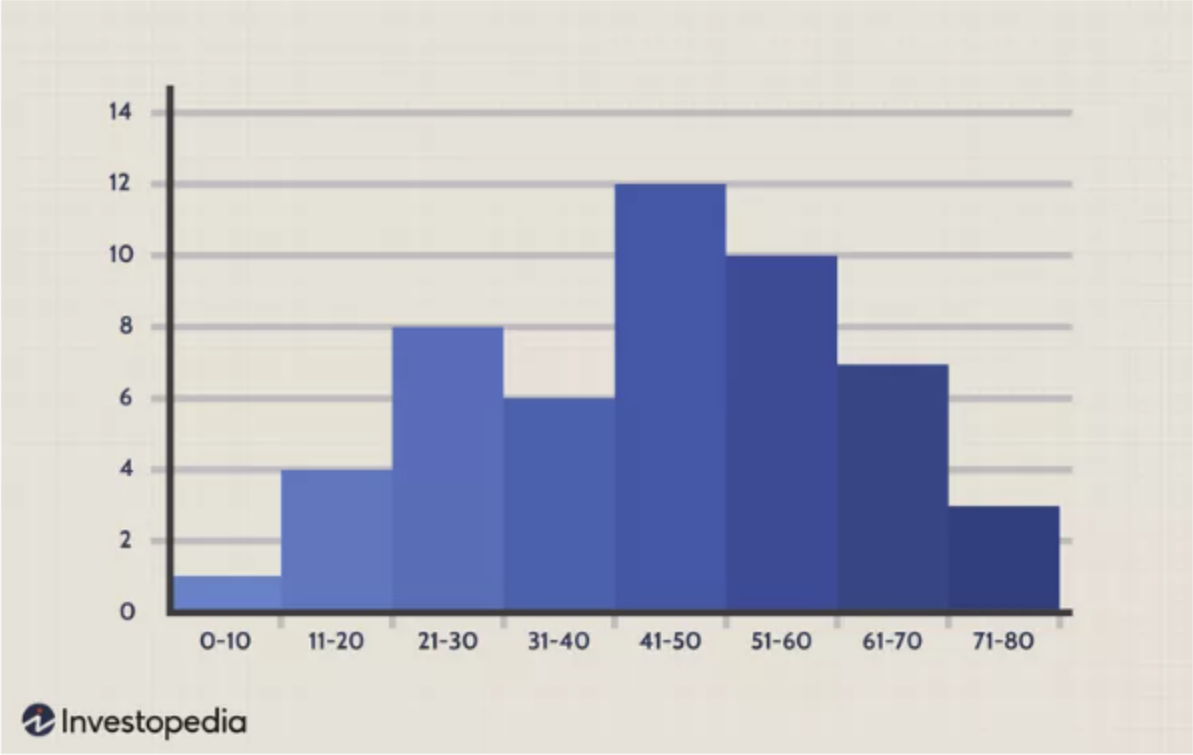
We would also want to use a generic bar/line chart for some of these stats, to visualize the distribution of turn counts, and average rating over the average ages.

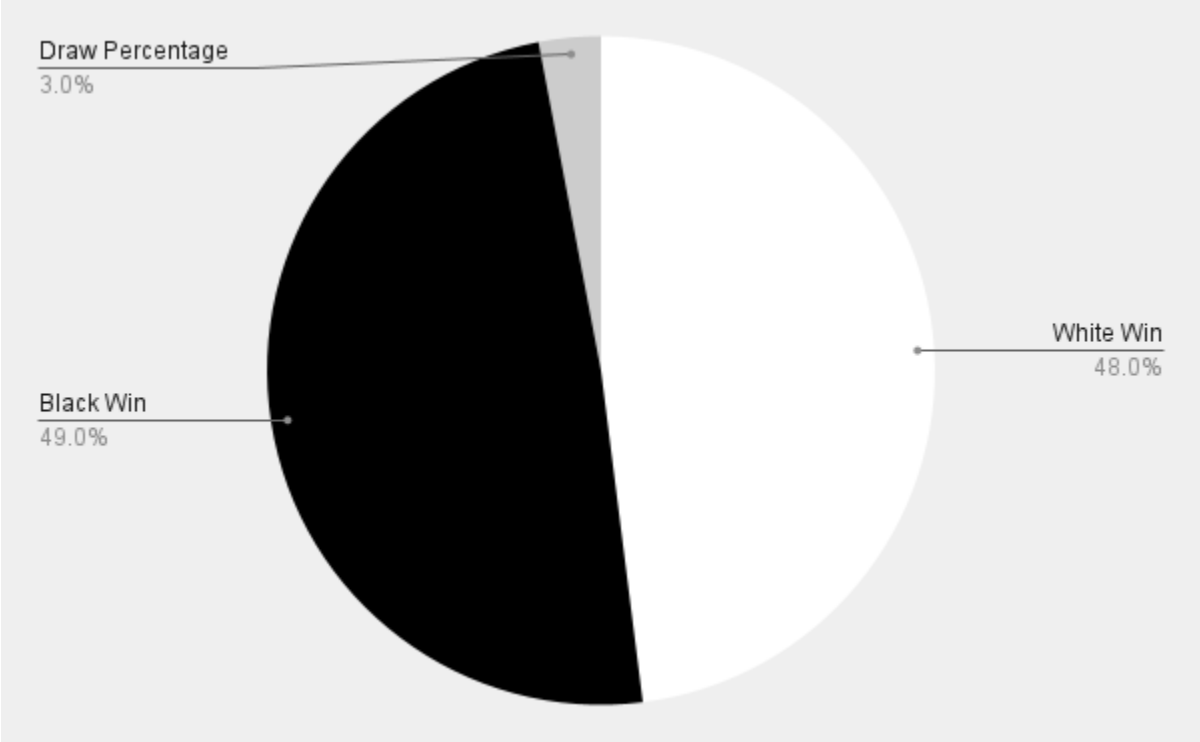
Player Elo by Ages



Average Player Elo by Ages







Must-Have Features. List the features without which you would consider your project to be a failure.

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?

Visualizations that answer the following questions
-Does playing chess exclusively result in a better or worse rating than performing puzzles, and studying games alongside regular play?
-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:
https://chess.stackexchange.com/questions/2506/what-is-the-average-length-of-a-game-of-chess#:~:text=The%20average%20length%20is%2079,due%20to%20people%20resigning%20early
.

Optional Features. List the features which you consider to be nice to have, but not critical.

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?

Project Schedule. Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

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!