CS439 Final Project Proposal: "An Insight into Chess.com's Playerbase"
Nicholas Myrick
nmyrick@purdue.edu, 0033294190

Who?
Nicholas Myrick.
This will be a solo project.

What?
Dataset(s):

Chess.com Titled Tuesday Dataset:
https://huggingface.co/datasets/kirillgoltsman/titled-tuesday-chess-games
Preprocessing:
- Remove rows that do not have player accuracy, player Glicko score, and/or tournament date.

Chess.com random user game dataset: https://www.kaggle.com/datasets/lukasimek/3m-chess-com-games
Preprocessing:
- Remove rows that do not have player accuracy, player Glicko score, and/or game date.
- Remove duplicate games

Grandmaster (GM) games on Chess.com:
https://www.kaggle.com/datasets/grgoirepelletier/6-2m-chess-games
Preprocessing:
- Remove rows that do not have player accuracy, player Glicko score, and/or game date.
- Remove duplicate games

All GM Chess Games on Chess.com
https://www.kaggle.com/datasets/dimitrioskourtikakis/gm-games-chesscom

General preprocessing for GM games:
- Combine datasets to get a more expansive look on GM games and accuracy.

General preprocessing for all games:
- If a dataset is too big, find a way to randomly remove games per elo bracket to make the dataset smaller.

Why?
The goal of this project is to give insight into Chess.com's player scores and rating, as well as the trend in performance in its main paid tournament "Titled Tuesday".

For the general player base, I would like to visualize:
1) A comparison of player Glicko rating vs player accuracy
2) The statistical winning chances when paired against a differently rated opponent

3) The difference in the accuracy to Glicko rating depending on the time control.

I think these visualizations will give a good insight into Chess.com's implementation of the Glicko rating system to show what a player can reasonably expect of their own statistics and performance against other opponents.

Recently, there has been speculation that GMs have been cheating in Chess.com's titled Tuesday event to win its top prize.

To provide some insight into if these claims have any basis, I would also like to visualize:
1) The trend in accuracy/skill of the titled Tuesday event from 2022-2023, along with its top players.
2) How those top players compare in accuracy from their normal games to their Titled Tuesday games

Finally, to give insight into Chess.com's top players, I would like to visualize:
1) The trend in Glicko rating in GMs over the course of the years
2) Allow search of GM name to see their individual and Titled Tuesday stats

How?

Visualization solutions for each category:

General Player Visualization
1) A comparison of player Glicko rating vs player accuracy
    a. This can be a bar chart where the x-axis is categorized by groups of Glicko score (500-600, 600-700, 700-800, etc), and the y-axis is the player accuracy from 0 to 100%.
    b. We can a selection of which time controls we would like to display (checkboxes beneath the graph), and have multiple bars per Glicko group, one for each time control.
2) The statistical winning chances when paired against a differently rated opponent
    a. This can also be a bar chart where the x-axis is categorized in groupings of difference in player Glicko scores (something like opponent is 20-50pts lower, 50-70, 70-90, etc), and the y-axis will be winning chance from 0 to 100%.
3) I want the option to select the date range of these visualizations, so one can see how these accuracies and scores have changed depending on the year.
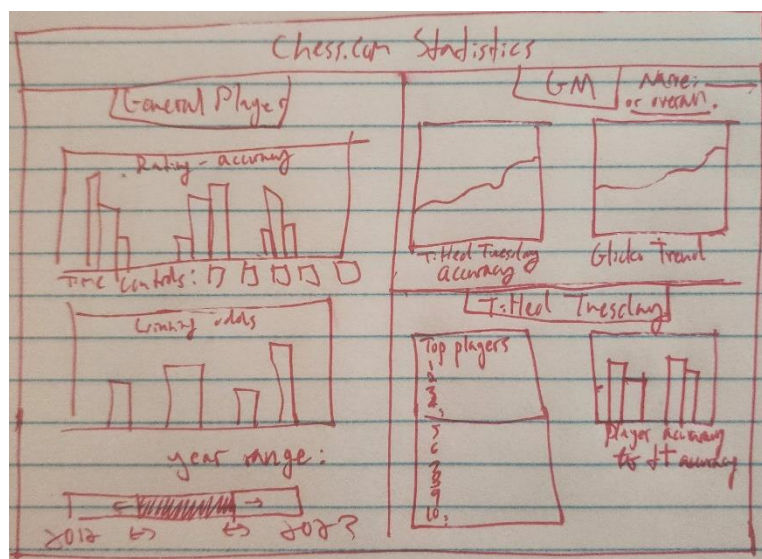
GM Visualization
1) Trend line of average Titled Tuesday accuracy along with normal accuracy
2) Trend line of Glicko score
3) Option to select an individual's statistics for these trends, or overall. If the GM did not participate in Titled Tuesday, then there would be no Titled Tuesday accuracy line.

Titled Tuesday Visualization

1) I want an interactive list of the top n players, showing their accuracy and number of wins, and clicking on it updates the **GM Visualization** to that player.
2) There should also be a bar chart visualization of the top players, where it has two bars for each player, one of their general accuracy and one of their Titled Tuesday accuracy.
3) There should be a bar where the user can choose the range of dates/years.

UI Rough Draft:

Libraries: Matplotlib for the graphs, pandas and numpy for data processing, and PyQt6 for the GUI framework.

Timeline
Milestones:
- Milestone 1: Finish all necessary preprocessing, by end of next week
- Milestone 2: Finish general player visualizations from the 3 million dataset
    1) Finish this by the first report
- Milestone 3: Finish GM visualizations week after
- Milestone 4: Finish Titled Tuesday visualizations for the final week
- Allows 1 week of extra time if something takes longer than expected

Dependencies between pieces of the work:
    The titled Tuesday will depend on the GM visualization in order to switch it to the clicked-on player.

Plan B
- The order of things removed due to time constraints will be:
  1) GM name search and click feature for Titled Tuesday
  2) Date/year range selections of Titled Tuesday statistics