

Process Book: Chess Openings and ELOs

Alex Hamrick, Stephen Harman, Garret Cervantez

Background and Motivation

Alex: I grew up playing chess from a very young age, and I placed second in the Idaho State Chess Championship at the age of 7. After that, I quit playing and following chess until 2018 when my interest in chess was revived by the World Chess Championship. I watched all of the games between Magnus Carlsen and Fabiano Caruana, and I began playing and studying the game again.

As I have begun to re-learn about chess, I've found so many new openings, strategies, and tactics that I never knew as a kid, and I have often wondered how to best prioritize my time learning the game. As a (now) average skilled player, studying and memorizing uncommon openings has allowed me to play above my skill level by catching my opponents off guard with moves they have never seen before. This led me to question whether such strategies would continue to work at the highest level, and so I wanted to do a project to learn more about chess openings, their success rates at the highest level, and how effective becoming a specialist at a given opening would be. I also wanted to do a project that would allow us to analyze the effects of technology on chess players, as youtube and online chess engines have been instrumental in my chess revival, and I imagine similar tools have also helped those playing at the highest level.

Ultimately, I proposed a chess related project due to my own background and interest in chess, but I think our proposed project will also interest those who lack any real background with chess.

Stephen: I also grew up playing chess, but casually with friends and family. I am an intermediate player who will still play from time to time. I really enjoy chess and thought that it would be a fun project to work on.

Project Questions and Purpose

The purpose of this project is to learn more about the history of chess, understand how chess play and players have evolved in the modern era, and derive information regarding any correlations between openings, player rating, and win rate.

The first major question our project hopes to answer is: do skilled players develop expertise in one or two openings, or do they have a broader control of a large quantity of openings? Is it more effective for players to invest large amounts of time in becoming experts at a few openings, or is their time better spent studying other things? For example, Magnus Carlsen is known for playing the Sicilian Defense, but does he really play this opening a disproportionate amount compared to other players of his caliber, and if so, does this niche expertise wind up ultimately helping or hurting him?

The next question our project hopes to answer is: what are the general trends for various common openings? Do certain openings minimize the possibility of a draw between the two

players? Which openings favor black or white? If you are in a must win situation, which openings should you consider?

Our project also hopes to discover how chess players (their habits and elo) have changed over time. As chess has gotten more and more popular, we expect that the best players will reflect an overall upward trajectory in elo over time. Additionally, we want to be able to analyze how the rise of computers and chess AI have affected chess players. Now that machines surpass humans at chess, have players been able to effectively utilize new technologies to improve even further? Or has such a discovery led to an increase in drawn games as players are more able to effectively prepare for their opponents?

There are two big advantages that would come from answering the above questions. The first applies to the chess community. Understanding which openings are most effective and how specialization affects players will give players more insight as to what they should spend their time studying and could be a good indicator as to how important opening theory really is. Especially since modern openings rely on heavy memorization of hundreds or thousands of possible lines, this could indicate whether the best players specialize in just a few lines or whether they need to study every line equally.

The second benefit can be applied more generally. Understanding how play has evolved over time allows us to speculate as to how things like chess AI and technology have affected other games as well. If we find a sharp increase in skill among all high rated players in the modern era, it is likely because of the increased accessibility of high quality chess resources via the internet. Alternatively, if we see that only the best players improve, then we can hypothesize that the easy access to information allows the most skilled players to maximize their skill gap compared to other players. Insights like these can be applied more generally to other games, sports, and activities and could inspire related research in other areas.

Finally, this visualization will hopefully provide people with the ability to answer their own questions about chess and its history while also encouraging them to discover questions of their own!

Finalizing Project Ideas

Date: 10/22/20

We all met today to discuss the different types of insights that we hoped to gain from our established chess dataset. We planned to each draw up a prototype for the project design so that we could combine our ideas into one final design. In general, we were interested in opening success, opening frequency, player ratings over time, draw rate over time, and player comparisons. Our plan is to reconvene on Tuesday and create the final project design.

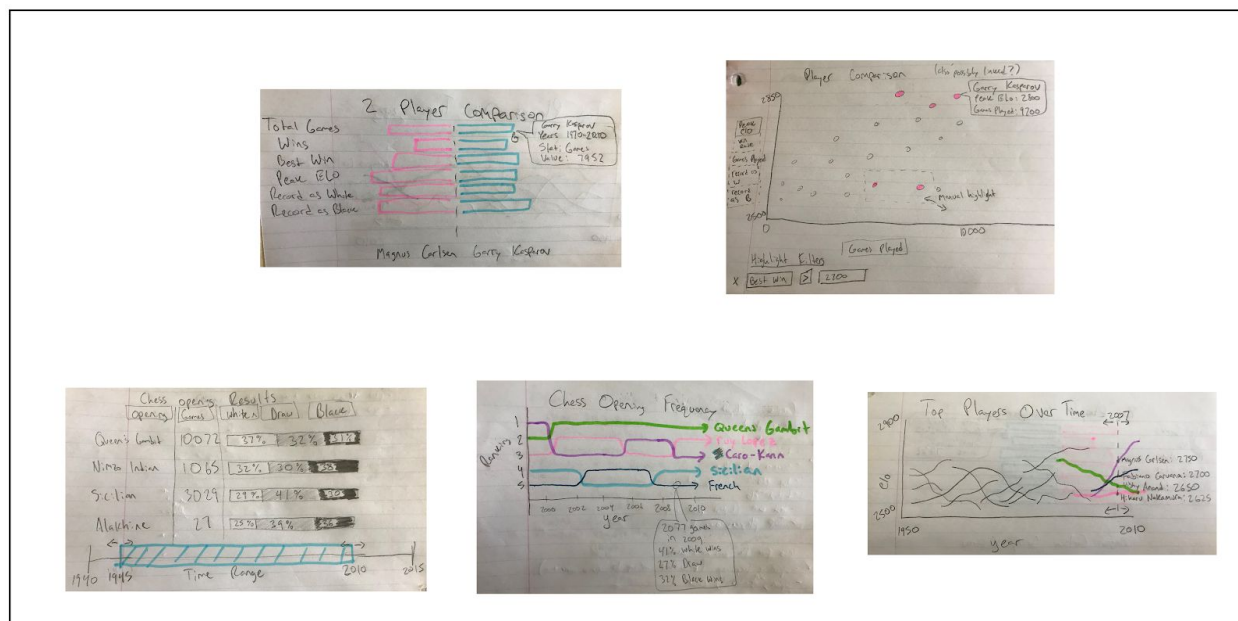
Final Design

Date: 10/27/20

Today we each showed our project prototypes and decided on our favorite visualizations. We ultimately decided that our final design should include the gini impurity density plot, the player ranking over time graph (which we slightly modified to display career data for players that were a top player in the selected year), and the chess opening data table. We will meet again on Friday to add the finishing touches to our proposal document. The prototype and final designs are shown below.

Prototypes

Prototype I



A better view of each visual can be seen below. Let's discuss each visualization individually first. We will go left to right, and top to bottom.

The top row is focused on specific players and their rankings based on different stats. The top left image shows a visualization in which users can select two different players on which we have data and compare them based on a variety of factors (wins, games played, etc). This would be a nice way to display specialized information about individuals.

The top right image, on the other hand, displays all the players from a given time period and plots them based on customizable attributes (again, using wins, games played, max elo, etc). This scatterplot would allow highlighting of players who meet a certain criteria and would show more information about the individual when hovering over their data point.

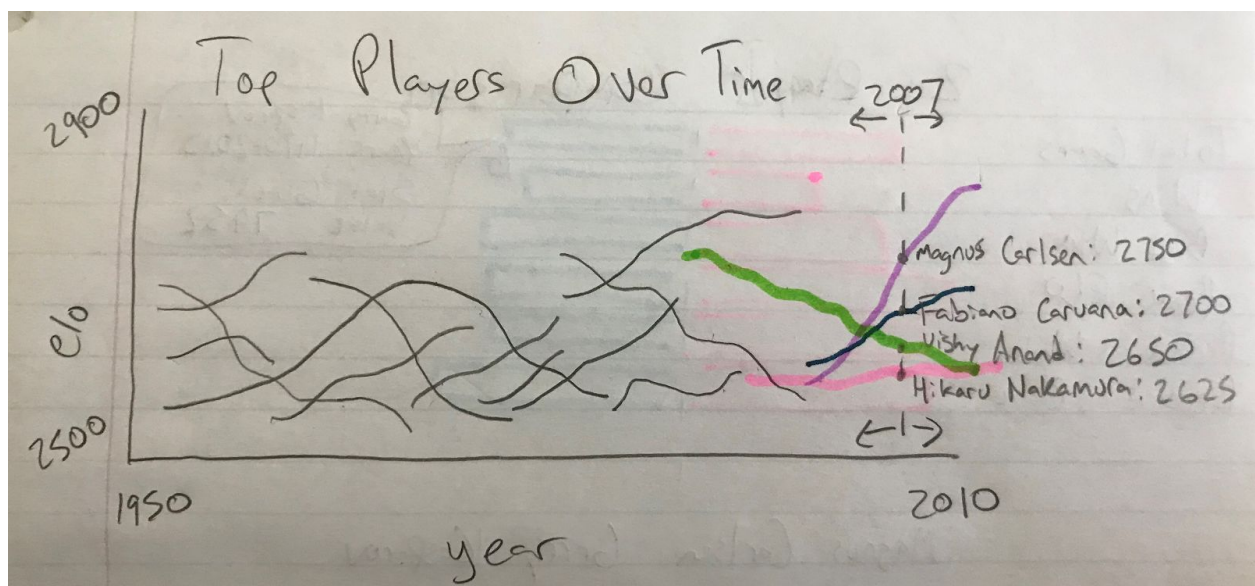
The bottom row focuses both on chess openings and general chess trends over time. The bottom right visualization lists all the played openings and shows the number of times they were played, as well as the win percentages for each player when playing this opening. Each

category (games played, black win rate, white win rate, etc) would be sortable, allowing users to quickly see both the most common openings as well as the most effective openings for each color.

The bottom middle visualization shows the rankings of most popular chess openings over time. This makes it easy to see the trends in chess openings and gives us an idea of which openings have remained effective and which have died out over the years. Additionally, hovering over a point on one of the lines would provide additional info about the games which played that opening during that year, allowing users to dive deeper into the opening trends than just their frequency.

The final visualization at the bottom left shows a plot of the top chess players over time. There is a vertical, dashed line which can be manipulated to indicate the selected year. For the selected year, the top chess players for that year will be highlighted and their names will be displayed. This visualization allows users to see both the top players and player elo trends over time.





Chess opening	opening Games	Results		
		White	Draw	Black
Queen's Gambit	10072	37%	32%	31%
Nimzo Indian	1065	32%	30%	38%
Sicilian	3029	29%	41%	30%
Alakhine	27	25%	39%	36%

← →

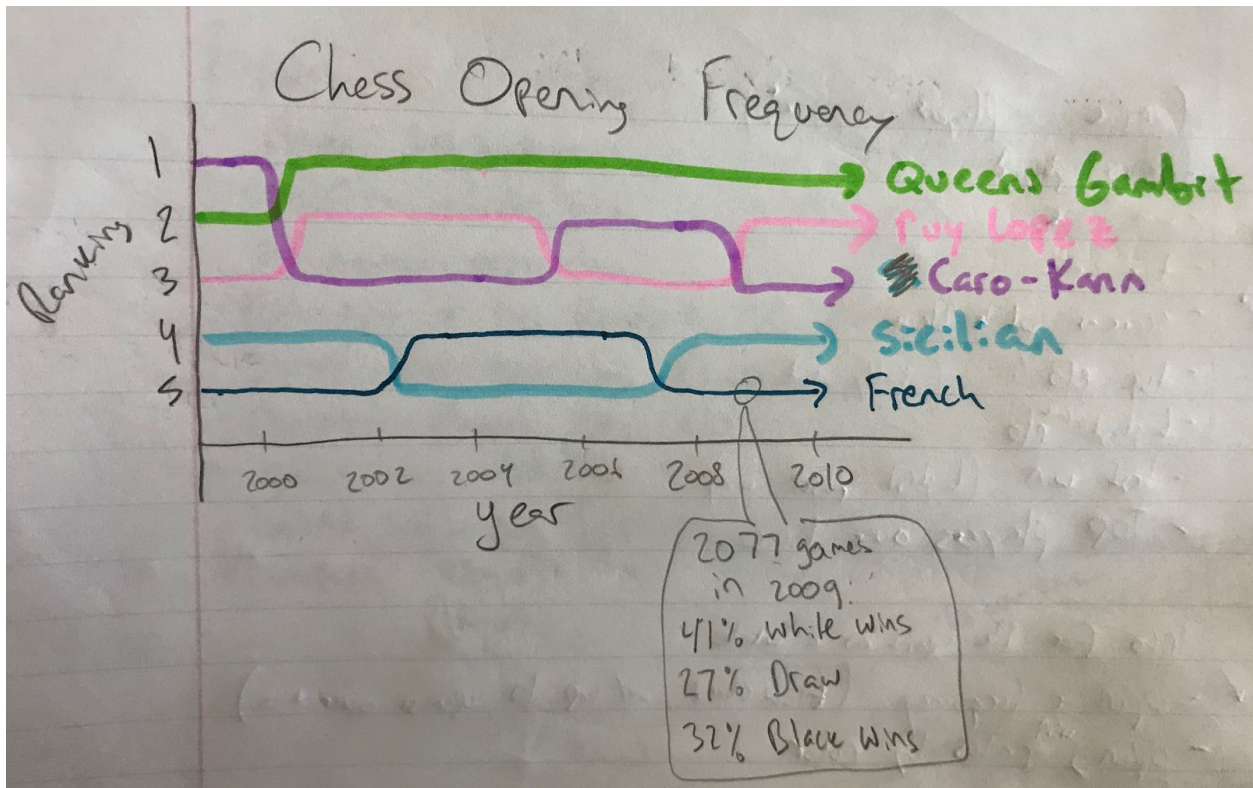
1940 1945

←

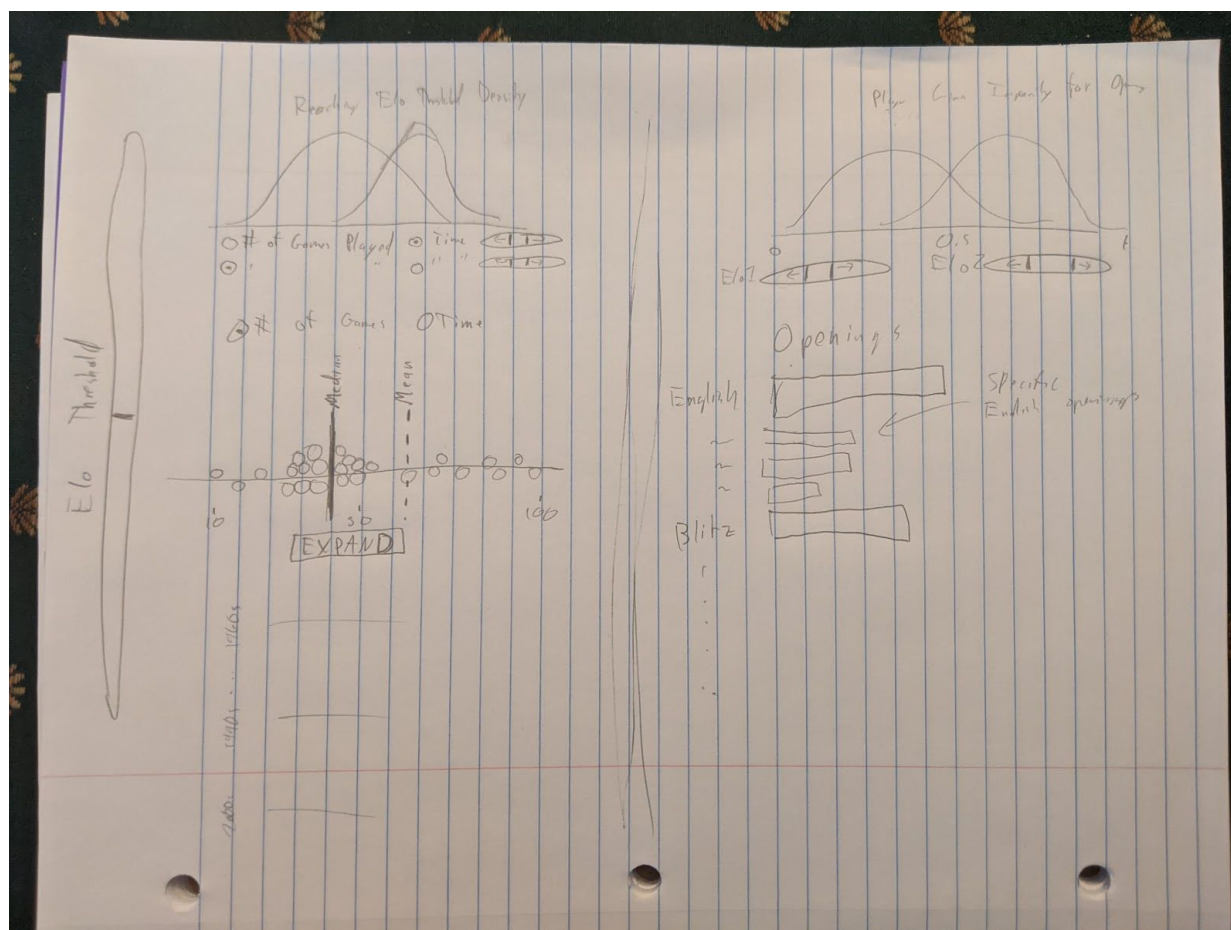
→

2010 2015

Time Range



Prototype II



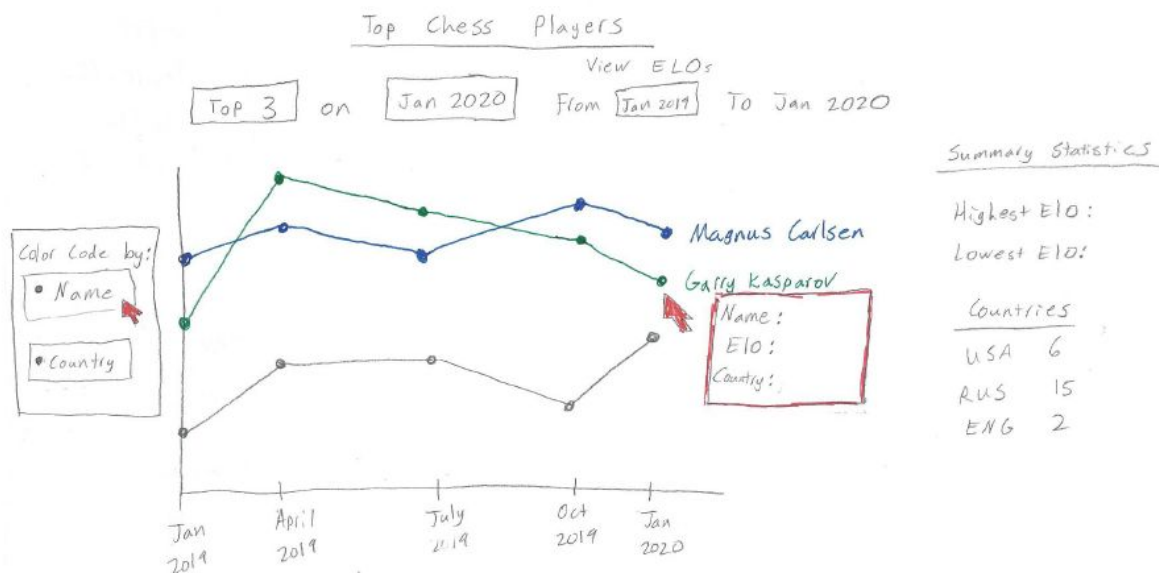
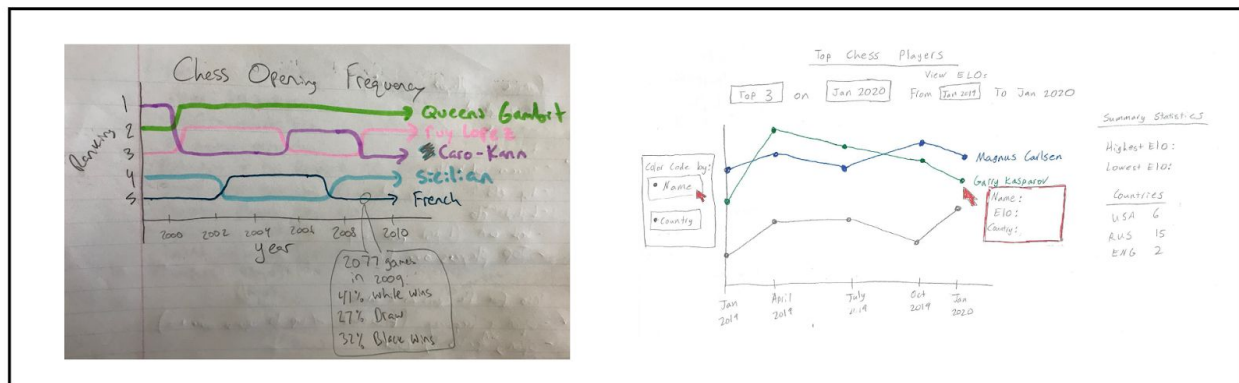
This prototype consists of 4 visualizations. Lets address each of these by their quadrants with quadrant I being located on the top right and quadrant IV being located at the bottom right. The quadrant II and quadrant III visualizations are related through an elo threshold seen on the far left. The quadrant II visualization is a density plot of either the number of games or the time it took to reach a certain elo threshold. The time and the dates are options for each density plot. There is also a filter for each density plot that limits the years you want to look at e.g. only look at the distribution from 1980 to 1990.

The quadrant III plot shows a beeswarm plot of how many games or how much time it takes to reach the elo threshold where each circle is a player. The beeswarm plot can then be expanded to see it grouped by the decade of their first game.

The quadrant I visualizes a density plot of the gini impurity of each player's opening moves. In the most simplified terms, this shows if they used many different openings or few openings. More is done with this in the deep dive below.

The quadrant IV visualization is a bar chart that shows the number of opening moves given an opening move category (e.g. The English) then by clicking on the move a bar graph of each specific move would be shown.(e.g. Troeger defence).

Prototype III



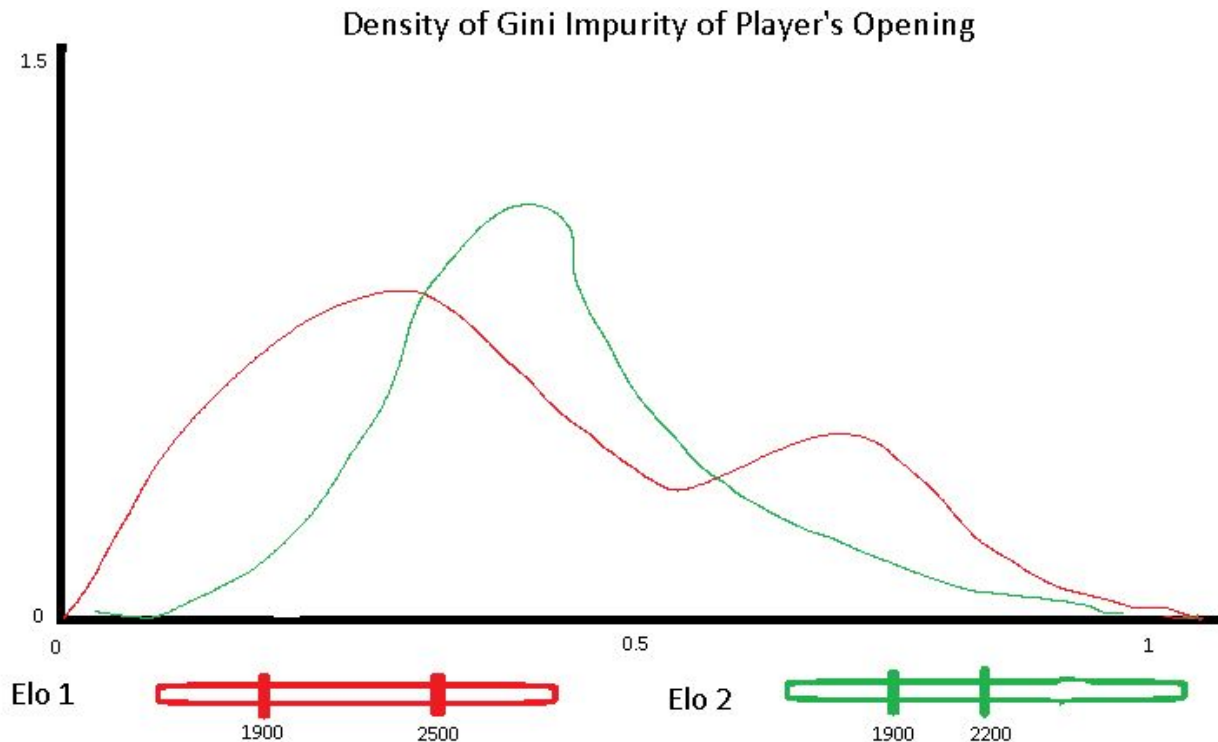
This visualization focused on learning about the top players and openings. This provides a simple UI with information which the general chess player may be interested in. The player could choose how many of the top players to display for a given quarter. (Data source: <https://ratings.fide.com/toplist.phtml>). Then the range of dates can be chosen to see those players' ELO temporal progression. The top chess moves could be linked to the user's chosen dates and then insight could be gained from exploring what openings were popular at the times leading up to the players' top ranking.

Summary statistics would be a nice addition to the right since it wouldn't clutter up the UI and it gives the opportunity to give a general summary not easily determined from the visualizations. The user could choose to have the lines color-encoded by name or by country. This would add an easy extra dimension to consider and explore alongside the ELO progression. On mouse-hover, the line could pop-out and display the player's name, exact ELO,

and country. We decided against this because we thought having a different dataset for different visualizations could be misleading to the user.

Final Designs

Density of Gini Impurity of Player's Openings



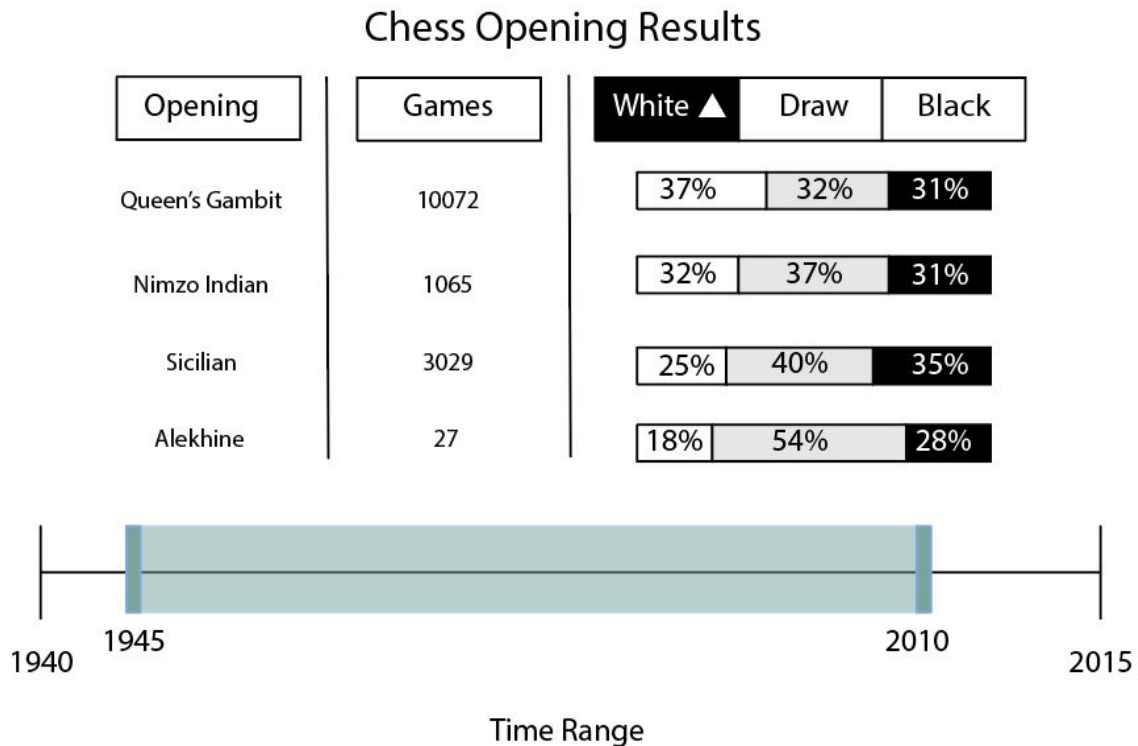
Description:

This is a density plot of the gini impurity of all players where their max elo was within a certain range. In oversimplified terms, one can think of the gini impurity as a measure of “mixedness” and by looking at the distribution of the opening moves it could suggest interesting trends. For example maybe the great players are specialists, i.e. they get really good at a specific opening, or maybe great players are generalists, i.e. they use a wide variety of openings depending on their opponent, or other related things.

- Must-have features
 - Two plots that show distributions
 - This could be either a density plot or a histogram
 - The ability to control the range of their max elo for the distribution plots
 - Ideally this would be a slider with an upper and lower bound but could also be a “less than elo x” slider and “greater than elo x” slider.
- Optional Features
 - Kernel density options
 - Density plot

- Instead of histogram
- Double (lower and upper bound slider)
 - instead of single lower or upper bound

Chess Opening Results



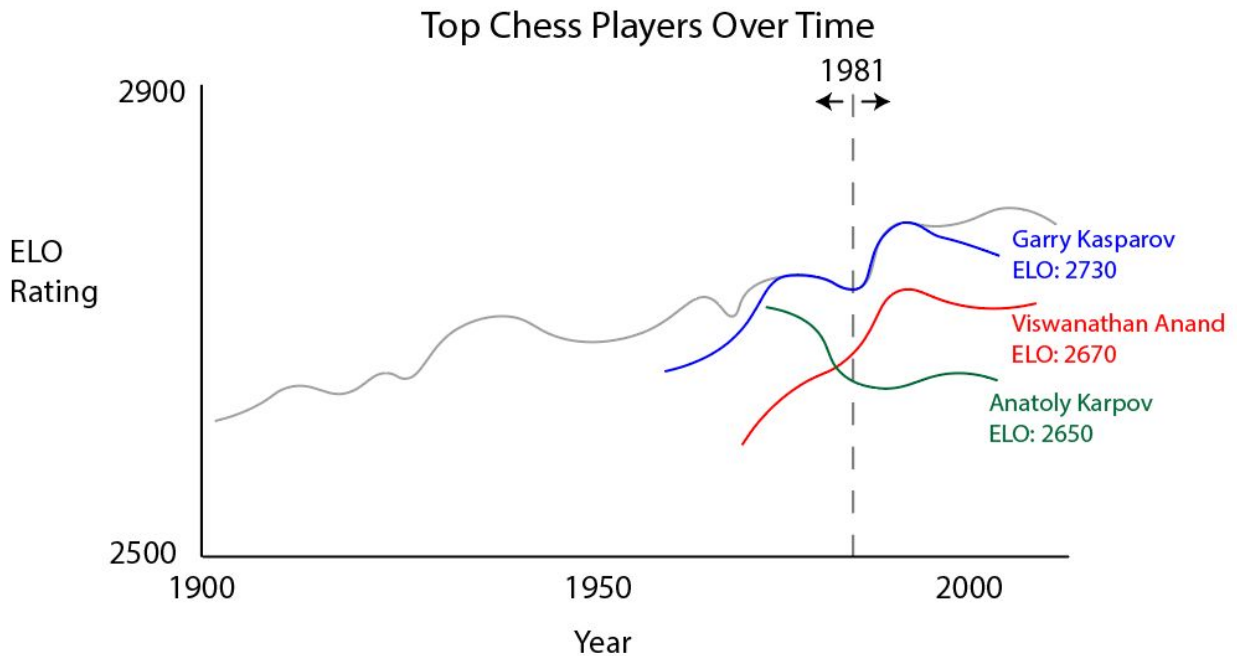
Description:

This table will show the number of games and the game result for each chess opening in the user chosen time range. The table will be sortable based upon any of its columns. This will allow the user to see how frequently the opening occurred in the games alongside the result of the games.

- Must-have features
 - Openings
 - Games
 - White, Draw, Black
 - Sorting
 - Choose Time Range
- Optional Features

- Limiting the number of openings displayed to a custom value

Top Chess Players Over Time



Description:

This chart displays both the overall top chess ratings from every year that our data covers, as well as the career ELOs of the three highest rated players from that year. The dotted vertical line can be manipulated to change the current active year. This visualization allows users to clearly see the overall ELO trend over time, while also allowing them to get a sense for the best players over time and the various career trendlines that exist among top-tier players.

- Must-have features
 - Max elo per year line
 - Ability to manipulate current year
 - Player elo trend lines (depending on the year)
 - Labels for player name and elo in the selected year
- Optional Features
 - Toggleable number of best players (doesn't always have to be just 3)
 - Additional filters on the players considered in the visualization

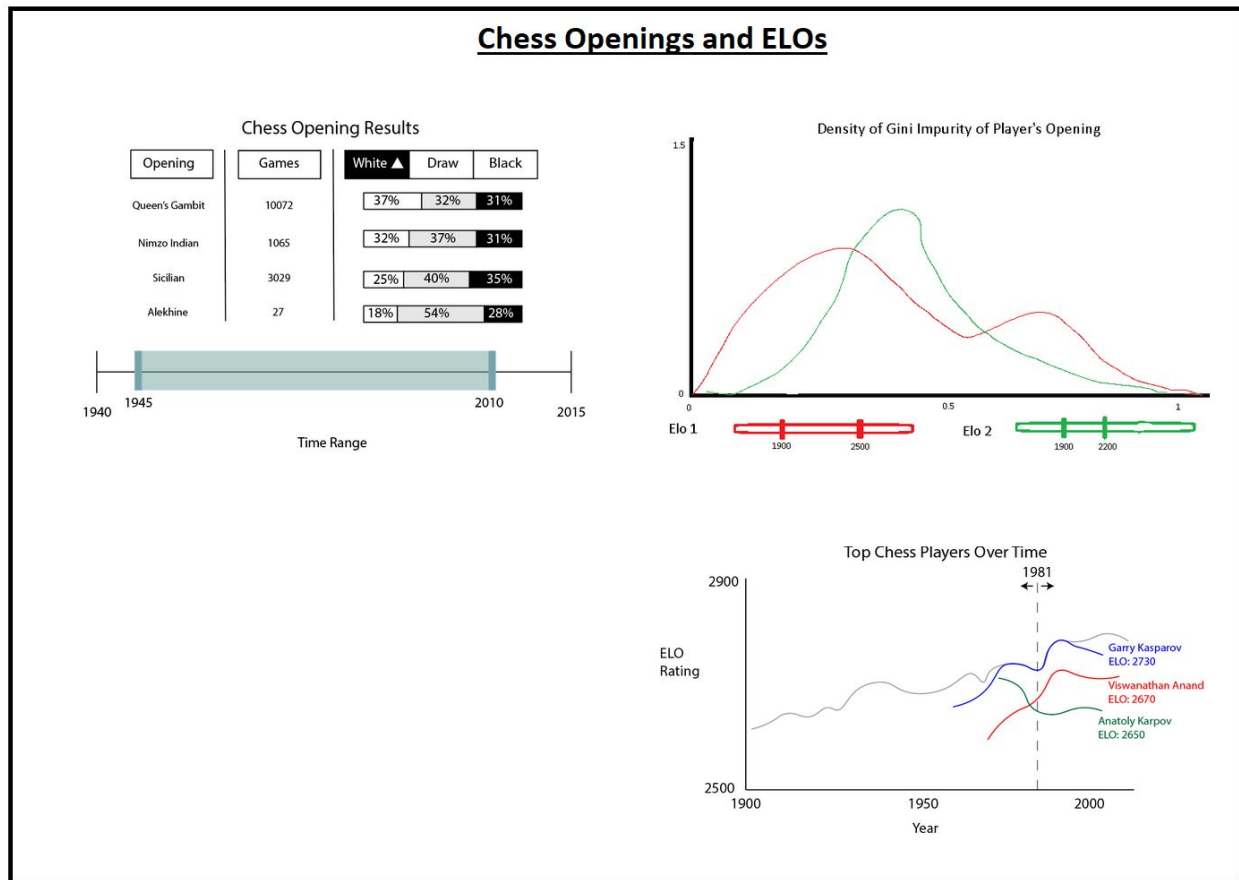
Must-Have features

- We must have the following visualizations as described above:
 - Chess Opening Results
 - Density of Gini Impurity of Player's Opening
 - Top Chess Players Over time

Optional Features

- Dark-mode-esque color theme
- Density plot of peak ELO vs how many games it took to reach that ELO

Final Design Layout



We decided on combining these three visualizations into our final design because we felt that these three provided the most interesting data to explore. The left column would contain all of the chess openings for the chosen time range and the right half would contain The Gini Impurity and Top Chess Players visualizations. These three visualizations provide the user a great

opportunity to analyze periods of time in chess and make connections between openings, gini impurity density, and the top chess players over time.

Proposal Document Meeting

Date: 10/30/20

This meeting was brief and it just outlined the few remaining steps that need to be taken on the proposal document before submission.

Data Cleaning: Turning pgn data into tabular data

Date: 10/31/20 8:53 AM

We took the all.pgn file which is 187,768,942 and 2.86 GB and created Master.csv that is 3,561,471 rows and 220 MB. The PGN was parsed by iterating over each line and then updating a ChessGame object property based on the PGN tag. Once the game result was reached this was an indication that a new chess game was about to start and the previous ChessGame was added to an array. Once the entire PGN was parsed then we iterated over the ChessGame array and wrote each line to a csv. The names in the PGN have a comma (e.g. "Kasparov, Garry") so we delineated using semicolons. Below are pictures of what the PGN looked like and what the tabular data looked like after it was parsed. Work can be seen in "MasterCsv.ipynb".

all.pgn

```
[Event "Kasparov Chess sim"]
[Site "New York"]
[Date "2000.03.14"]
[Round "6"]
[White "Kasparov, Garry"]
[Black "Stoffers, Jeffrey"]
[Result "1-0"]
[WhiteElo "2851"]
[ECO "D35"]
[EventDate "2000.03.??"]
```

1.	d4	d5
2.	c4	e6
3.	Nc3	Nf6
4.	cxd5	exd5
5.	Bg5	Be7
6.	e3	Ne4
7.	Bxe7	Nxc3
8.	Bxd8	Nxd1
9.	Bxc7	Nxb2

Master.csv

Date	WhiteName	WhiteElo	BlackName	BlackElo	Result	ECO
2000.03.??	Kasparov, Garry	2851	Stoffers, Jeffrey	None	1-0	D35
2000.03.??	Kasparov, Garry	2851	Tomasso, Santiago	None	1-0	B01
1999.11.20	Kasparov, Garry	2851	Teixeira, Rafael Goltzman	None	1-0	C54
1999.11.20	Kasparov, Garry	2851	Quintino, Luis Felipe Pires	None	1-0	B01
2000.02.09	Kasparov, Garry	2851	Piket, Jeroen	2633	1/2-1/2	C99

Data Cleaning: Filtering Master.csv based on maximum elo

Date: 11/1/20 2:50 PM

The Master.csv needed to be filtered down based on elo. To do this we only wanted to look at players that were the top 100 players of any given decade. We defined the best players of a decade by their max elo during that time period. This filtering was done by first creating a new dataset where both the white and black players were being looked at. Then from here we grouped by decade, then grouped by player, then took the max elo for each player. We then created a set of the best players' names derived from the above process. Then we filtered out any games where either black or white was not in our set of best players. There were records where the date was missing and these were also filtered out. This took the number of records from 3,561,471 to 344,953 and the size from 220 MB to 23.5 MB. This table has the same schema as the Master.csv table. This table is also included under the Data folder. Work can be seen in "BestPlayers.ipynb".

Basic code structure:

Date: 11/1/20

Added the very basics for the code structure including index.html, script.js (empty), and style.css. Id tags for each section were included in the index.html file. Also added an additional file BestPlayersReformat.csv which had a comma delimiter instead of semicolon (This was later removed and d3.dsv(';',...) was used) Appearance:

Chess Openings and ELOs

Chess Openings

Gini

Top Players

Added feedback_exercise.pdf :

Date: 11/5/20

We received good feedback and included this document in our git repository.

Notes

- Explain chess terms. **Michael**
- Who is the audience? **Michael Jessica**
 - “Pick your audience and stick with it.”
 - This applies specifically to the Gini Impurity visualization which is complicated and, in some ways, is very niche.
- Link the slider on the chess openings chart and the top players for the year sliders. **Michael**
- For the chess opening table, move the slider above the table. **Michael**
- On the chess openings bar chart show the current years selected on the slider at the top of the visualization. **Michael**
 - For example show “1940 - 2010” below chess openings on a new line.
- The top three chess players could be hard to read because they are too close. To solve this we could have a legend where the top player name and elo changes but the colors remain the same based on their position. **Michael**

Analyzing Feedback

One of the big points of feedback that we got was related to the accessibility of our visualizations. Since they are all focused on chess, and we discuss things like openings, ratings(elo), and players, it would be very easy to isolate users who are not familiar with chess. Michael and Jessica rightfully pointed out that it would be useful for us to have a small blurb or section that describes the chess terms that we use and possibly provides some context for the project as a whole. Along with that, Jessica mentioned that we should decide who our intended audience is, especially since one of our visualizations uses a rather complicated metric (Gini Impurity), and we will have to really explain this in a simple way if we want to avoid alienating less technical users. I think our goal is to make our visualizations very accessible, so we plan on taking this feedback and incorporating descriptive labels and comments in our visualization that will make clear what each visualization is measuring and any chess terms that get used. For the visualization using Gini Impurity our current plan is to simplify the naming of the visualization and then have a tooltip that explains what is technically being done. So for example the visualization might be renamed something to the effect of “Mixedness* of Player Openings” and then have text that explains what is going on when you hover over the word “mixedness”. We could also simplify the x-axis instead of going from 0 to 1 it goes from “Specialized Openings” to “Generalized Openings”.

We liked their storytelling aspect and considered an additional optional feature of including storytelling in our visualizations as well. We could explore the data and see if there are any particularly interesting observations. We could include something like when was the first time a computer beat the best chess player. Perhaps even include computer chess elo trends for storytelling.

The other major comments that Michael and Jessica gave were related to our visualization layouts. For our chess openings visualization, Michael recommended that we include the selected time range at the top of the visualization in order to make it more clear to the user what data they are selecting. This would also be important if we are listing a large number of openings, since users may have to scroll or look far down the list to see the selected years otherwise. Additionally, Michael recommended that we move the whole time range scale above the data table, and this would put a larger emphasis on time range aspect, which I think is good. We don't want the time range to be some additional feature, but rather we want people to discover how opening frequency and success rates have changed over time, and putting the time range scale at the very top of the visualization will likely encourage users to use it.

Work on Chess Openings Table :

Added opening and games column data to Chess Openings :

Date: 11/11/20

Did some very basic data processing in ChessOpenings.js and displayed the data in the first two columns.

Chess Openings and ELOs

Opening	Games	White	Draw	Black
D35	1415			
B01	1651			
C54	954			
C99	630			
D39	349			
B31	1566			
E68	900			
A14	1204			
B80	2289			
B33	2846			
D36	1843			
A01	587			
A40	1669			
B13	805			
B50	1640			
D97	633			

Added game results data to Chess Openings :

Date: 11/11/20

Calculated the wins for each opening and displayed a simple text output in the third column.

Chess Openings and ELOs






Opening	Games	White	Draw	Black
D35	1415	white:535	draw:659	black:221
B01	1651	white:790	draw:565	black:296
C54	954	white:326	draw:409	black:219
C99	630	white:221	draw:306	black:103
D39	349	white:110	draw:188	black:51
B31	1566	white:660	draw:606	black:300
E68	900	white:340	draw:352	black:208
A14	1204	white:362	draw:622	black:220
B80	2289	white:705	draw:917	black:667
B33	2846	white:924	draw:1204	black:718
D36	1843	white:708	draw:876	black:259
A01	587	white:208	draw:177	black:202
A40	1669	white:608	draw:645	black:416
B13	805	white:233	draw:394	black:178

Added bars and sorting to Chess Openings :

Date: 11/13/20

Created basic stacked bar charts which represented the absolute number of games won for each color.

Chess Openings and ELOs

Opening Games		White	Draw	Black
D35	1415			
B01	1651			
C54	954			
C99	630			
D39	349			

Changed the absolute numbers to ratios of the number of games won for each color.

Chess Openings and ELOs

Opening	Games	White	Draw	Black
D35	1415	<div></div>	<div></div>	<div></div>
B01	1651	<div></div>	<div></div>	<div></div>
C54	954	<div></div>	<div></div>	<div></div>
C99	630	<div></div>	<div></div>	<div></div>
D39	349	<div></div>	<div></div>	<div></div>

Each of the icons Opening, Games, White, Draw, and Black will toggle the sorting based upon the selection.

Chess Openings and ELOs

Opening	Games	White	Draw	Black
B07	4098	<div></div>	<div></div>	<div></div>
B22	3893	<div></div>	<div></div>	<div></div>
E12	3816	<div></div>	<div></div>	<div></div>
A30	3813	<div></div>	<div></div>	<div></div>
E15	3395	<div></div>	<div></div>	<div></div>
B06	3310	<div></div>	<div></div>	<div></div>
E11	3246	<div></div>	<div></div>	<div></div>

Improved appearance of html and added percentages to bars :

Date: 11/14/20

Tried a few different table header appearances.

Opening	Games	White	Draw	Black
B07	4098	<div></div>	<div></div>	<div></div>

Added the sorting up and down icons to indicate which column is being sorted and if it is ascending or descending.

Opening	Games ▼	White	Draw	Black
B07	4098			

Used the same API as the sorting down and up arrows to get the king and queen icons in the title. These can be removed later, but they look kinda fun for now.

♔ Chess Openings and ELOs ♚

Added the percentages to the bars.
End appearance:

♔ Chess Openings and ELOs ♚

Chess Opening Results				
Opening	Games ▼	White	Draw	Black
B07	4098	37%	37%	25%
B22	3893	24%	44%	31%
E12	3816	31%	49%	20%
A30	3813	27%	53%	20%
E15	3395	25%	56%	19%
B06	3310	36%	32%	32%

This looks very similar to the design in our ProjectProposal.pdf file. We need to make a csv which maps the ECOs to Opening Names (e.g. B07 = Pirc defence). We also need a slider at the top which filters based upon the selected date range.

Work on Gini Impurity :

Date: 11/14/20

Added some scripts to generate GinImpurity.csv from BestPlayers.csv

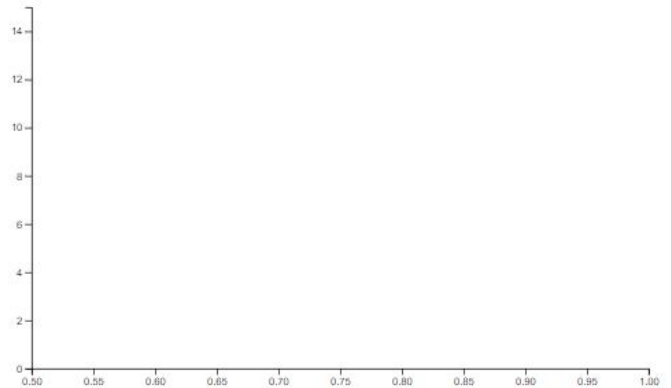
Edited the script such that loadData() returns a key-value object so that all data can be loaded in the same function.

Created gini_impurity.js

Tried to do density plot but failed. Added sanity check in python to see what the result should look like roughly. I also added some setup for the Gini Impurity plot.

Appearance:

♔ Chess Openings and ELOs ♚



Work on Top Players Over Time :

Date: 11/13/20

First, I added a new .ipynb file that would reformat and clean the data from BestPlayers.csv in a way that makes it easier to utilize in the Top Players Over Time graph. The first new csv that I generated is called TopPlayersByYear.csv, and a sample of this data is shown below.

```

Date;PlayerName;Elo
1952;Smyslov, Vassily;2620
1952;Gligoric, Svetozar;2575
1952;Byrne, Robert E;2560
1952;Filip, Miroslav;2510
1952;Donner, Jan Hein;2470
1952;Rossetto, Hector;2465
1952;Bisguier, Arthur Bernard;2430
1952;Cobo Arteaga, Eldis;2420
1952;Pedersen, Eigil;2370
1952;Enevoldsen, Jens;2350
1954;Smyslov, Vassily;2620
1954;Gligoric, Svetozar;2575
1954;Darga, Klaus;2540
1954;Matanovic, Aleksandar;2515
1954;Filip, Miroslav;2510
1954;Donner, Jan Hein;2470
1954;Rossetto, Hector;2465
1954;Robatsch, Karl;2460
1954;Yanofsky, Daniel Abraham;2460
1954;Bobotsov, Milko G;2455
1956;Portisch, Lajos;2640

```

This CSV contains the top ten players (by peak ELO rating) for each of the years in which there we have game data. So, for 1952, there are 10 associated players and peak ELO ratings. This dataset makes it easy to locate the top players during any given year.

The second dataset I generated is called TopPlayerCareers.csv. A sample of this dataset is shown below.

```

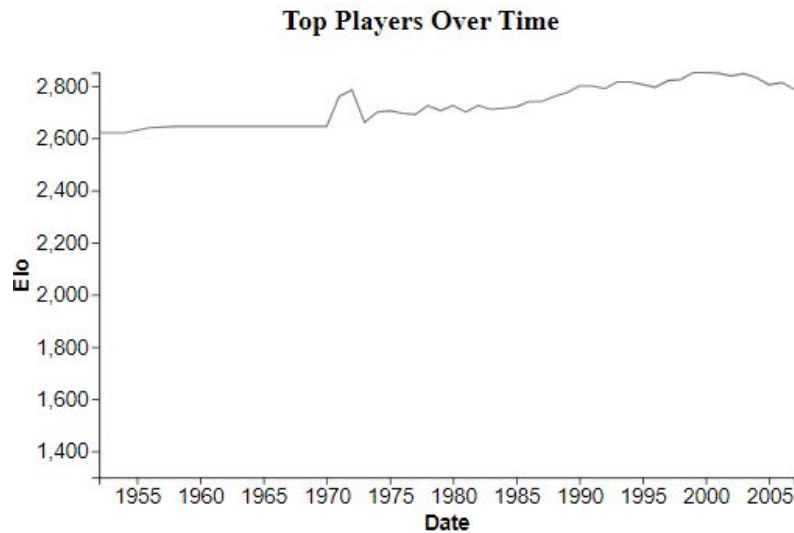
Date;PlayerName;Elo
1985;Adams, Michael;2360
1986;Adams, Michael;2295
1987;Adams, Michael;2360
1988;Adams, Michael;2460
1989;Adams, Michael;2510
1990;Adams, Michael;2590
1991;Adams, Michael;2615
1992;Adams, Michael;2620
1993;Adams, Michael;2630
1994;Adams, Michael;2675
1995;Adams, Michael;2660
1996;Adams, Michael;2685
1997;Adams, Michael;2680
1998;Adams, Michael;2716
1999;Adams, Michael;2716
2000;Adams, Michael;2755
2001;Adams, Michael;2750
2002;Adams, Michael;2752
2003;Adams, Michael;2734
2004;Adams, Michael;2740

```

This dataset shows the peak elo ratings for each player in every year that they were active, but it only contains players who were a top 10 player in at least one year. This means, this dataset contains the career elos for all the players that are in TopPlayersByYear.csv.

Date: 11/14/20

I created the general plot for the Top Players, with appropriately labelled and marked axes. I also added the background line which demonstrates the top elo score for the given years. An image of the plot is shown below.



The remaining work to be done for this plot is to display the career paths of the top players for the selected year. This will involve adding a slider and displaying additional paths which have already been created.

Merged Top Players, Preparing for First GitHub Release:

Date: 11/15/20

Uncommented Chess Openings Results (The initial loading time for this is a little slow, look to improve this). Merged branches and our repository is ready to release.

Appearance:

♔ Chess Openings and ELOs ♚

Chess Opening Results				
Opening	Games	White Draw Black		
B07	4098	37%	37%	25%
B22	3893	24%	44%	31%
E12	3816	31%	49%	20%
A30	3813	27%	33%	20%
E15	3395	25%	56%	19%
B06	3310	36%	32%	32%
E11	3246	32%	31%	17%
B33	2846	32%	42%	25%
B08	2766	28%	45%	27%
A00	2593	31%	38%	32%
B80	2289	31%	40%	29%
E94	2262	40%	30%	21%
B42	2209	33%	39%	28%
C42	2181	29%	61%	11%

