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.

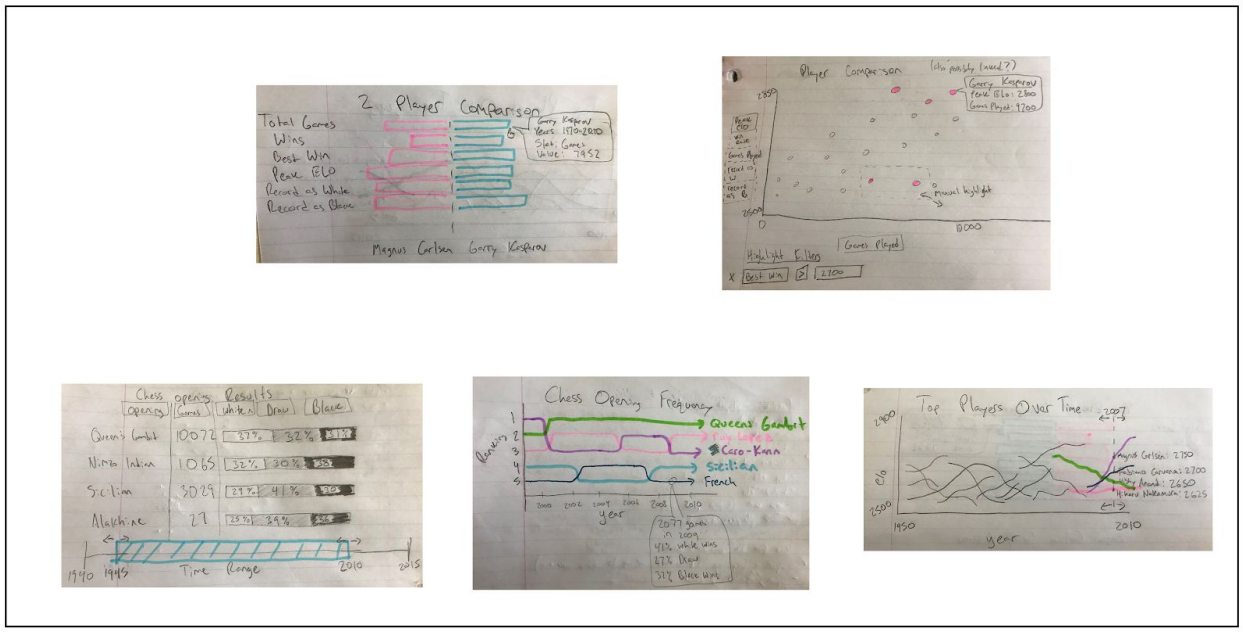
Final Design

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.

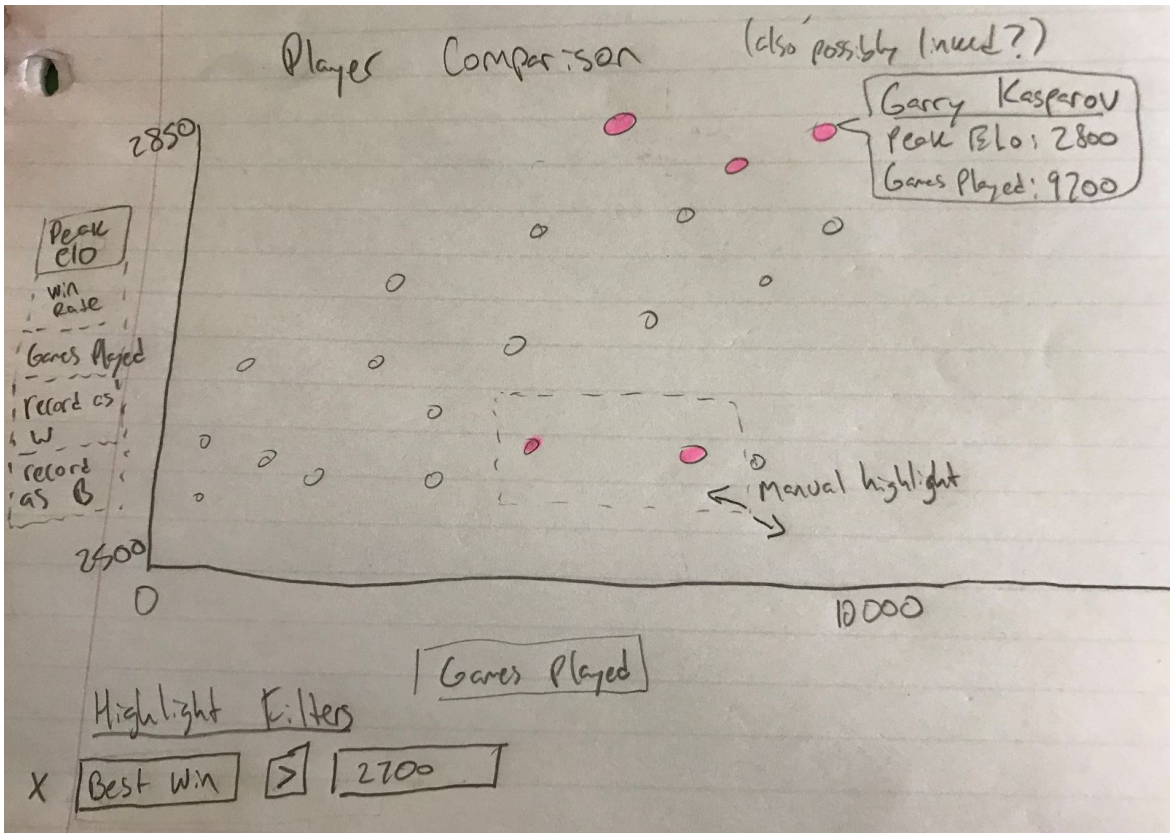
The bottom row focuses both on chess openings and general chess trends over time. The bottom right visualization lists all the played openings and shows the number of times they were played, as well as the win percentages for each player when playing this opening. Each

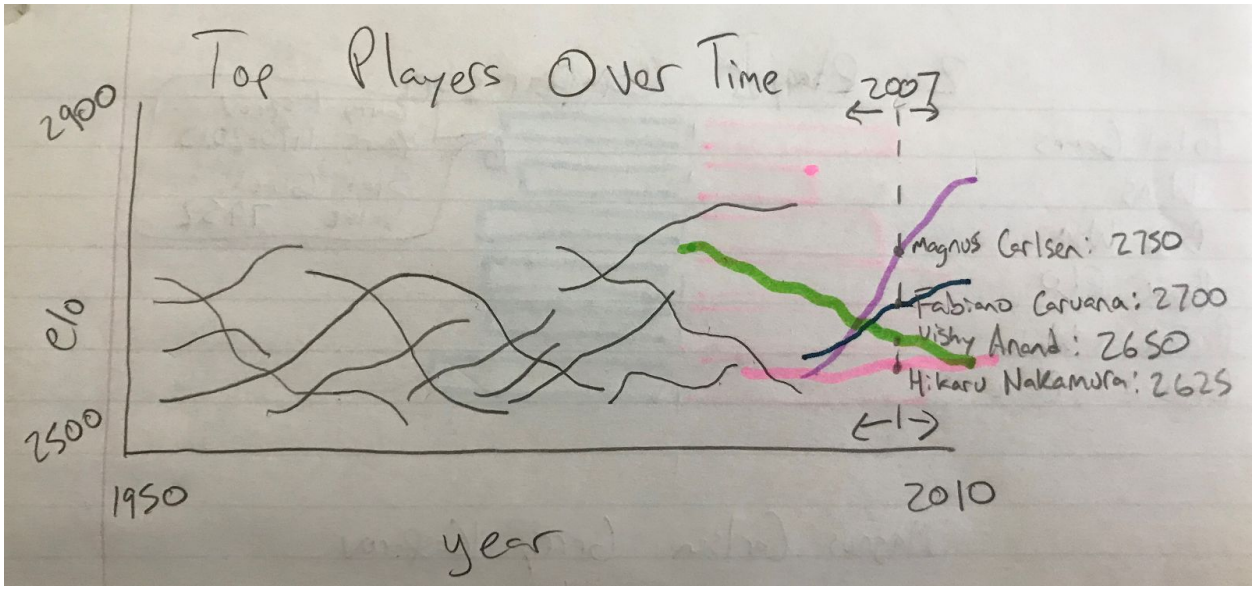
category (games played, black win rate, white win rate, etc) would be sortable, allowing users to quickly see both the most common openings as well as the most effective openings for each color.

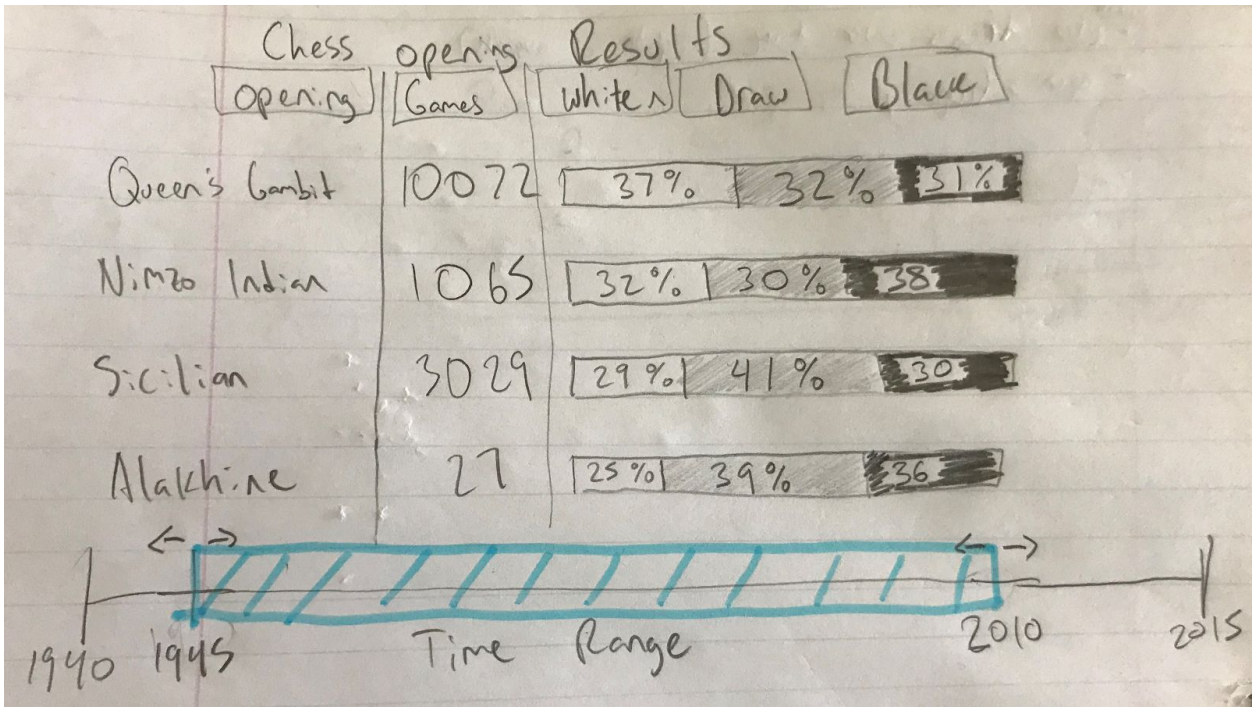
The bottom middle visualization shows the rankings of most popular chess openings over time. This makes it easy to see the trends in chess openings and gives us an idea of which openings have remained effective and which have died out over the years. Additionally, hovering over a point on one of the 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.

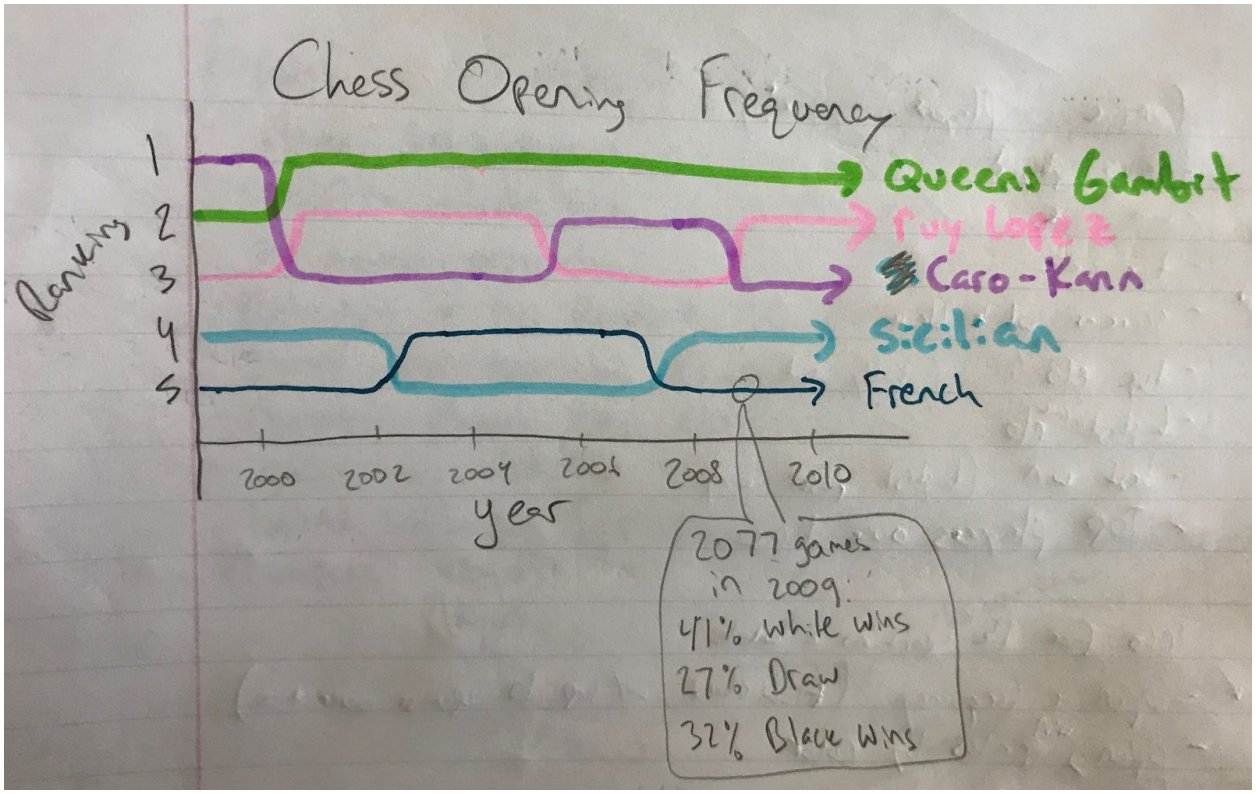
The final visualization at the bottom left shows a plot of the top chess players over time. There is a vertical, dashed line which can be manipulated to indicate the selected year. For the selected year, the top chess players for that year will be highlighted and their names will be displayed. This visualization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


2 Player Comparison
Total Games
Wins
Best Win
Peak ELO
Record as White
Record as Black
Garry Kasparov
Years 1970-2010
Stat: Games
Value: 7952
Magnus Carlsen
Garry Kasparov

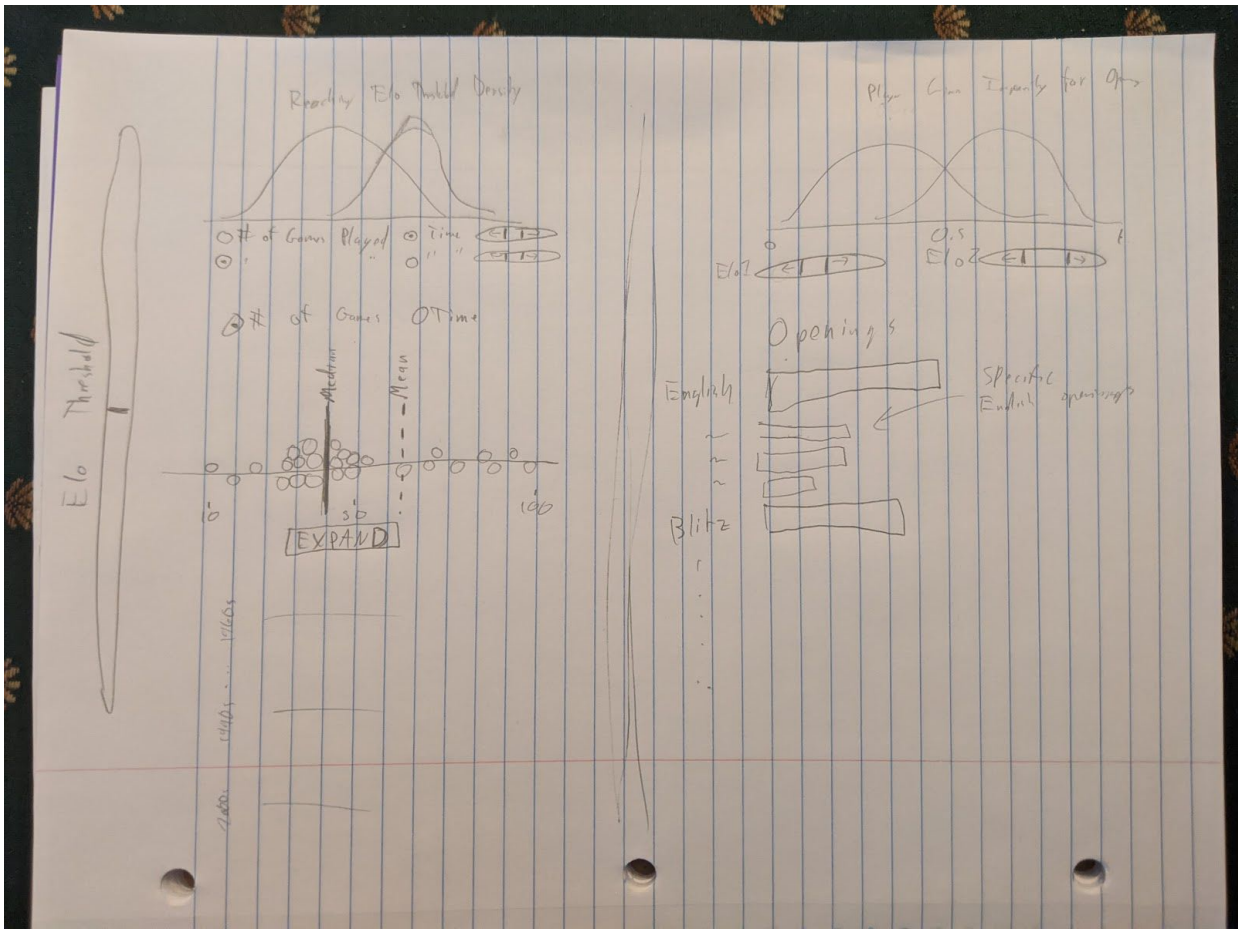

Player Comparison
(also possibly linked?)
Garry Kasparov
Peak ELO: 2800
Games Played: 9200
2850
2500
0
10000
Peak ELO
win rate
Games Played
record as W
record as B
Manual highlight
Games Played
Highlight Filters
X
Best Win
>
2700


Top Players Over Time
2007
2900
elo
2500
1950
2010
year
Magnus Carlsen: 2750
Fabiano Caruana: 2700
Vishy Anand: 2650
Hikaru Nakamura: 2625


Chess openings Results
opening
Games
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
Time Range
2010
2015


Chess Opening Frequency
Ranking
1
2
3
4
5
Queens Gambit
Ruy Lopez
Caro-Kann
Sicilian
French
2000
2002
2004
2006
2008
2010
year
2077 games in 2009:
41% White Wins
27% Draw
32% Black Wins

Prototype II



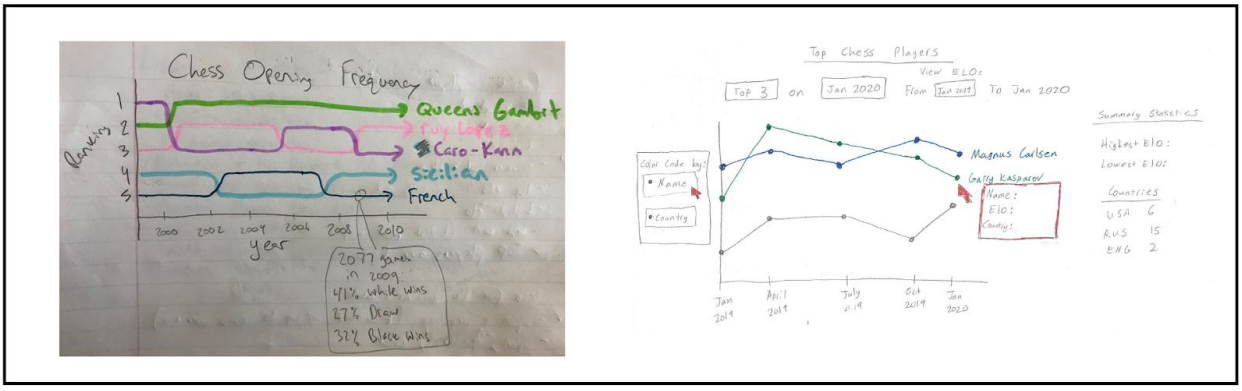
This prototype consists of 4 visualizations. Lets address each of these by their quadrants with quadrant I being located on the top right and quadrant IV being located at the bottom right. The quadrant II and quadrant III visualizations are related through an Elo threshold seen on the far left. The quadrant 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.

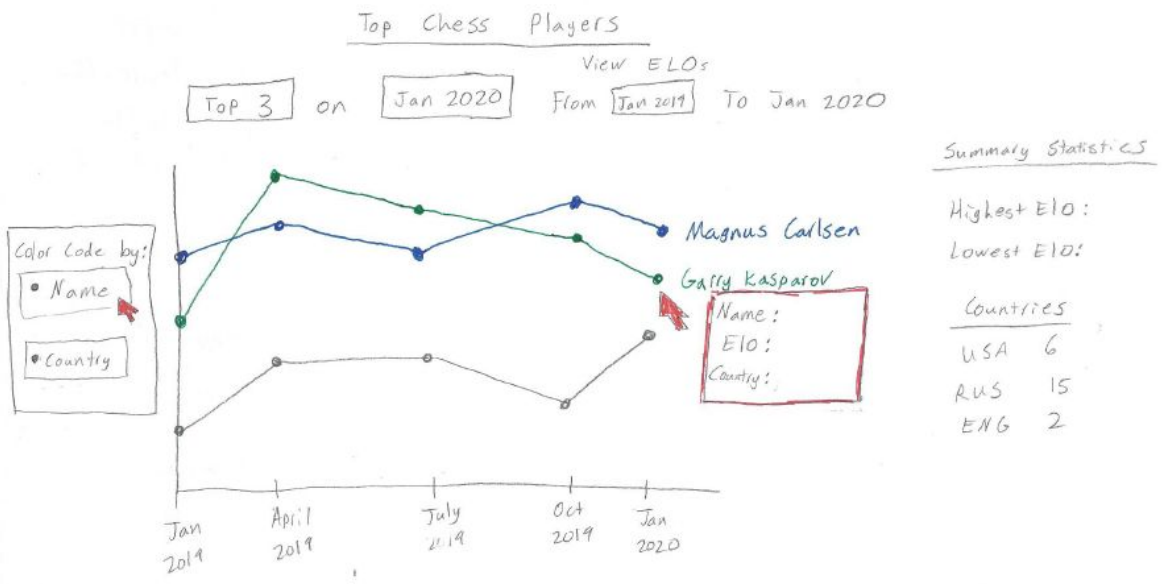
The quadrant III plot shows a beeswarm plot of how many games or how much time it takes to reach the Elo threshold where each circle is a player. The beeswarm plot can then be expanded to see it grouped by the decade of their first game.

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 final design section 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





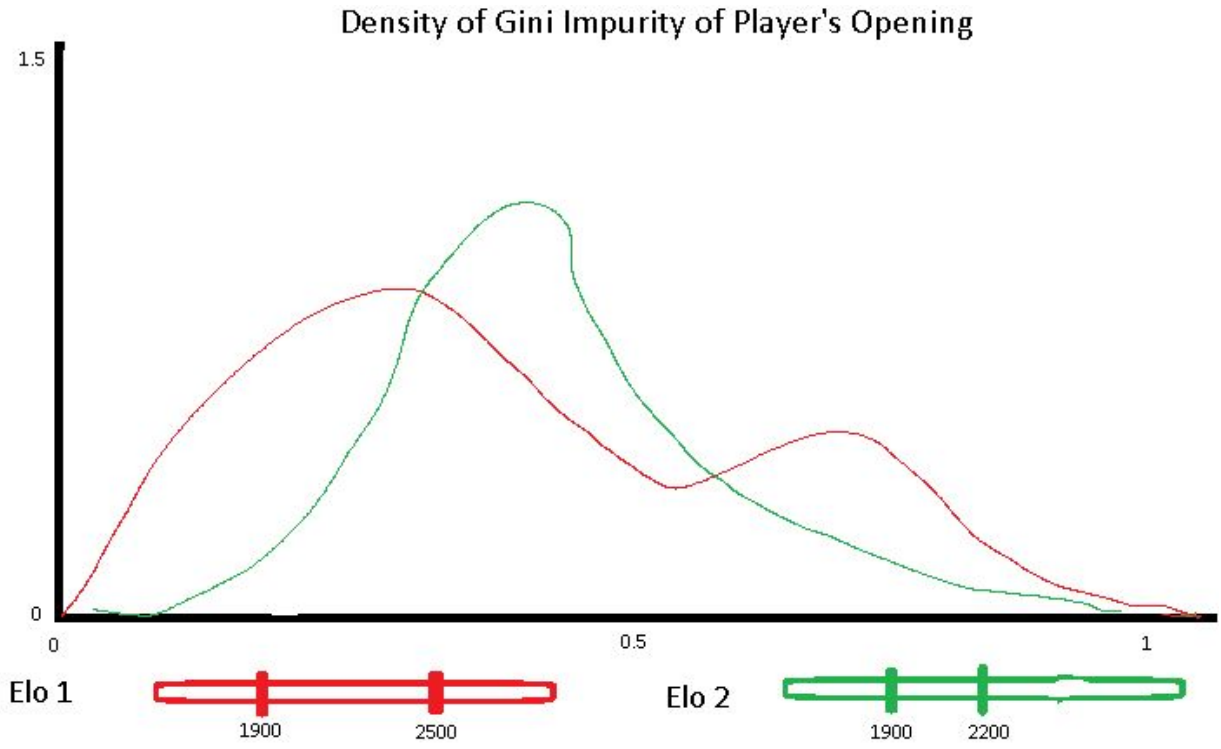
This visualization focused on learning about the top players and openings. This provides a simple UI with information which the general chess player may be interested in. The player could choose how many of the top players to display for a given quarter. (Data source: https://ratings.fide.com/toplist.phtml). 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



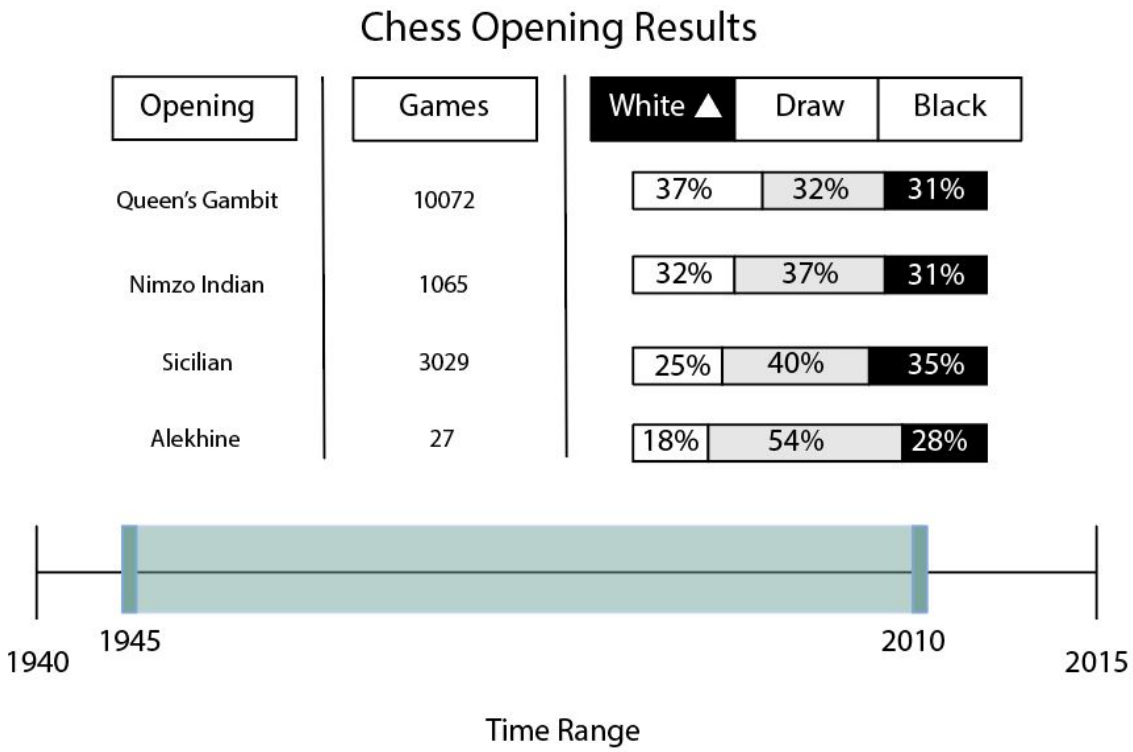
Description:

This at 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, their mood, or for other reasons.

- Must-have features
  - Two plots that show distributions
    - This could be either a density plot or a histogram
  - The ability to control the 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



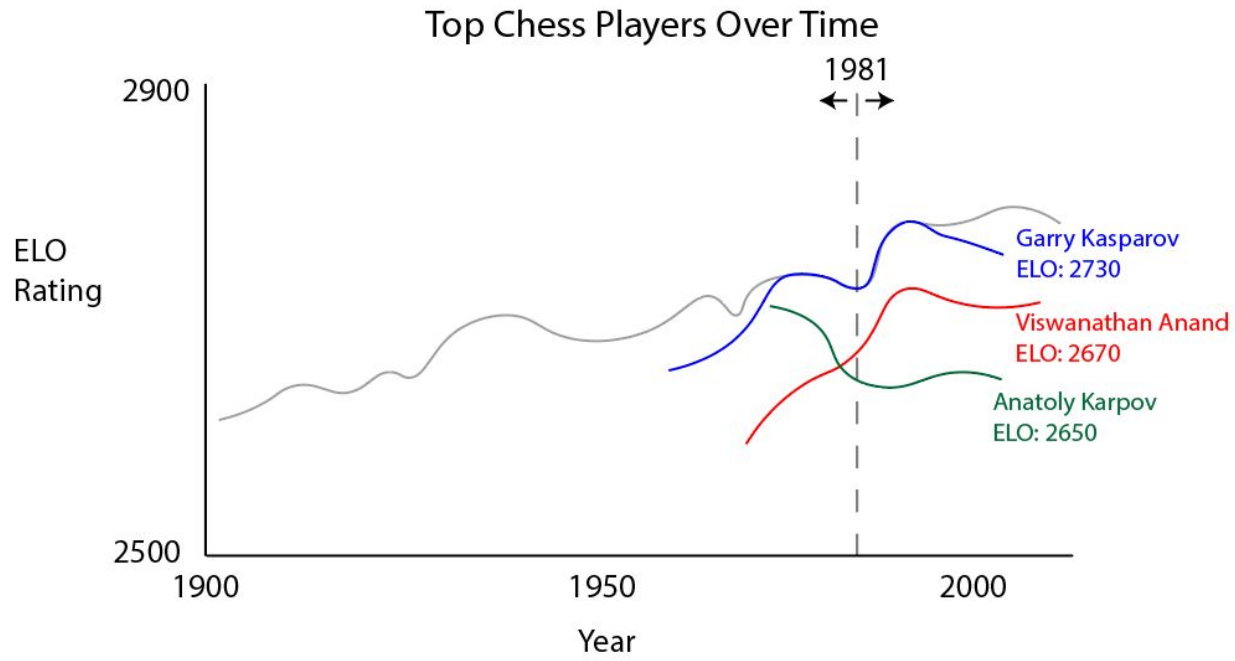
Description:

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

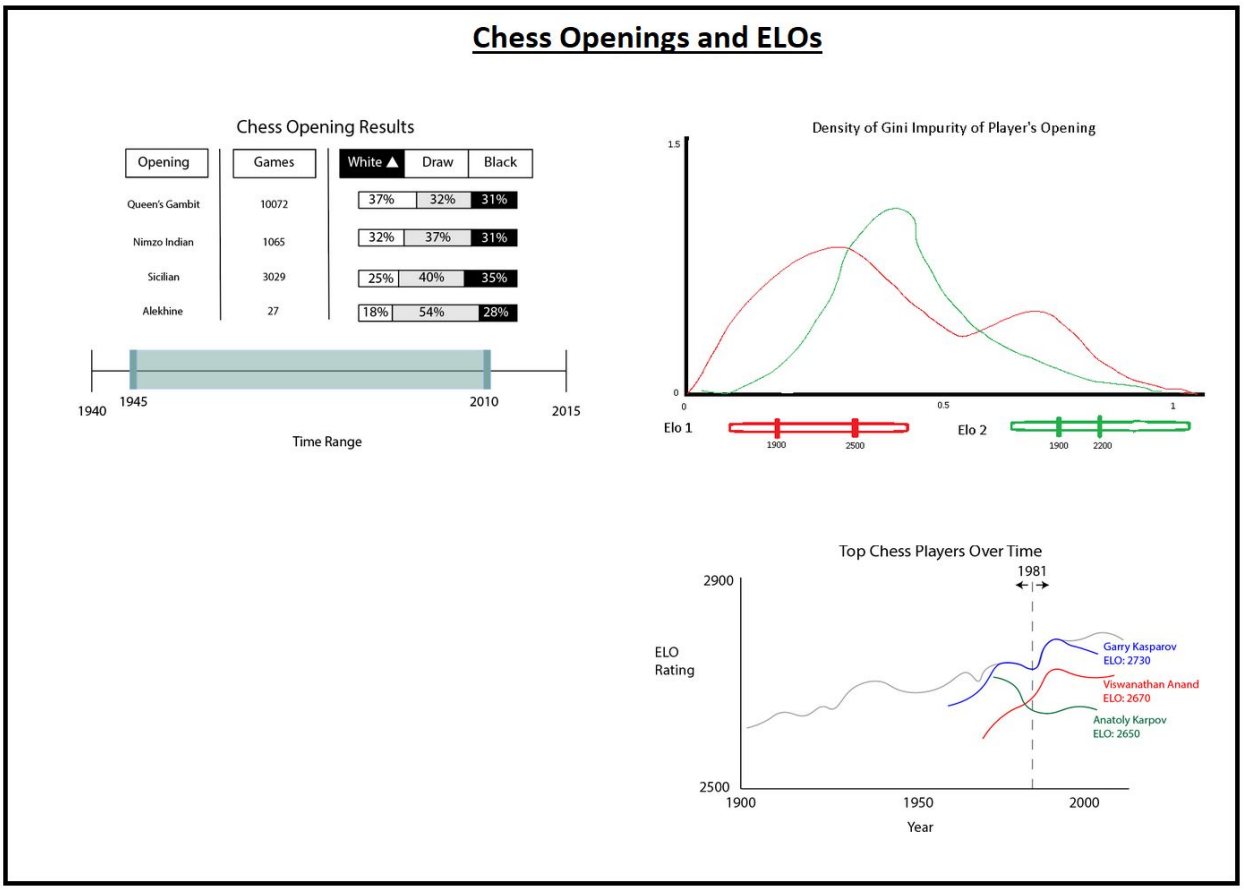
Must-Have features

- 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

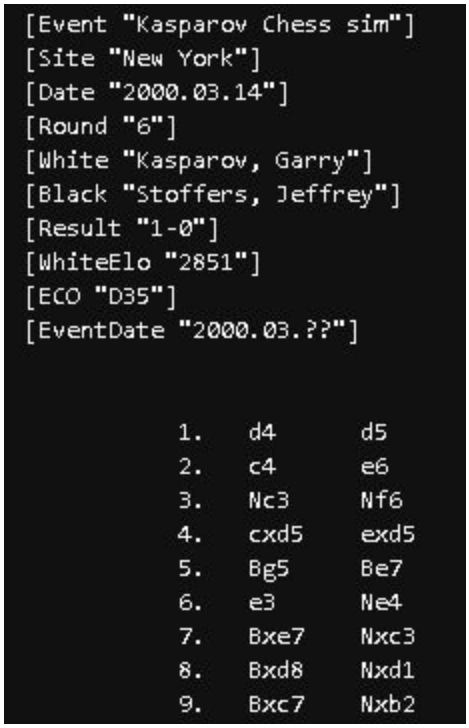
This meeting was brief and it just outlined the few remaining steps that need to be taken on the proposal document before submission.

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 lines and 2.86 GB and created Master.csv that is 3,561,471 tabular 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



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 Goltsman	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

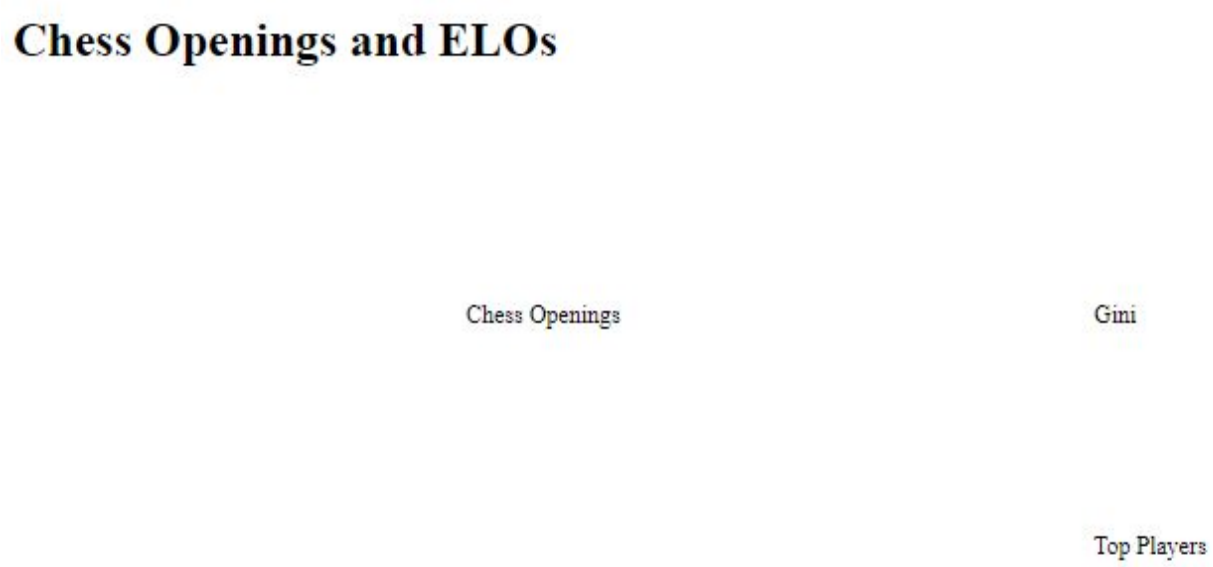
Date: 11/1/20 2:50 PM

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 file is 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:



Added feedback_exercise.pdf :

Date: 11/5/20

We received good student feedback and included this document in our git repository. Below are our notes on the meeting.

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 Student 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. Michaeland 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 (GiniImpurity), 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 ontaking 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”. Wecould 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.

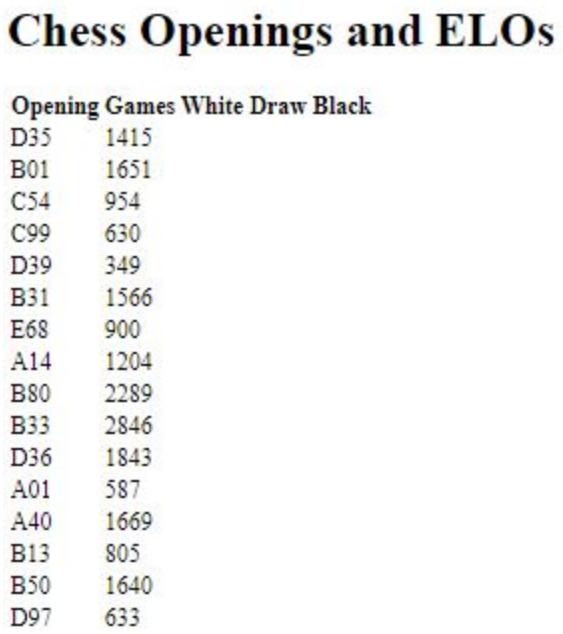
The other major comments that Michael and Jessica gave were related to our visualization layouts. For our chess openings visualization, Michael recommended that we include the selected time range at the top of the visualization in order to make it more clear to the user what data they are selecting. 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 “Games Results by Chess Openings”:

Date: 11/11/20

Added opening and games column data to Chess Openings

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			

Date: 11/11/20

Added game results data to Chess Openings :

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

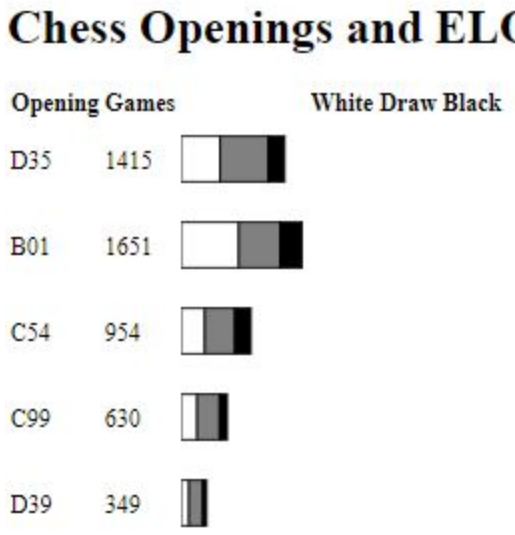
Chess Openings and ELOs

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

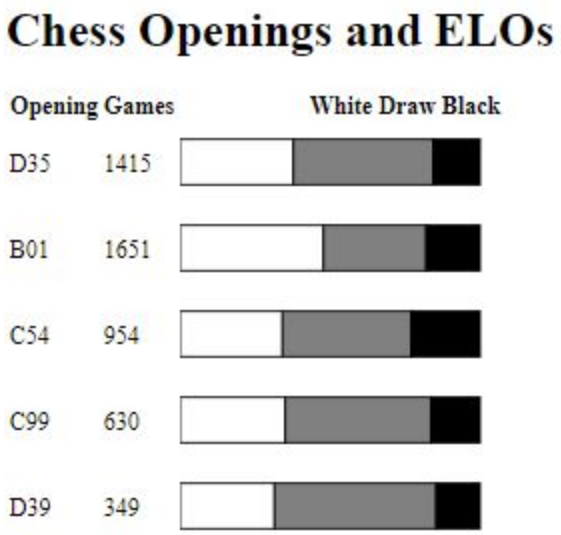
Date: 11/13/20

Added bars and sorting to Chess Openings :

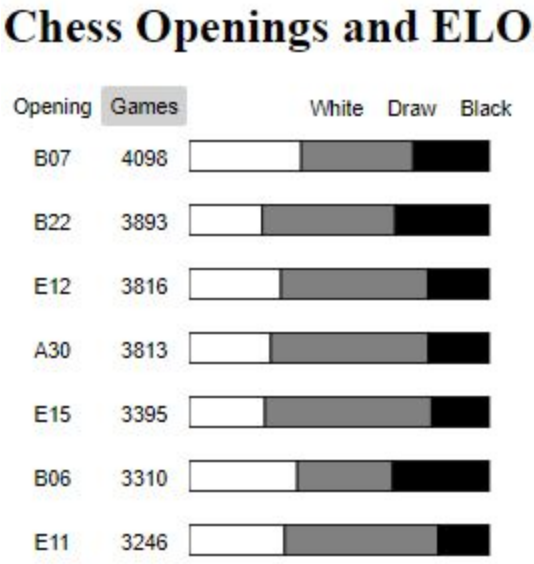
Created basic stacked bar charts which represented the absolute number of games won for each color.



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



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



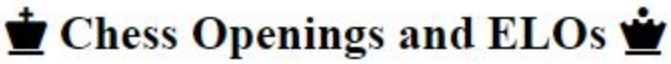
Date: 11/14/20

Improved appearance of html and added percentages to bars :

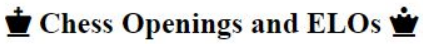
Tried a few different table header appearances.

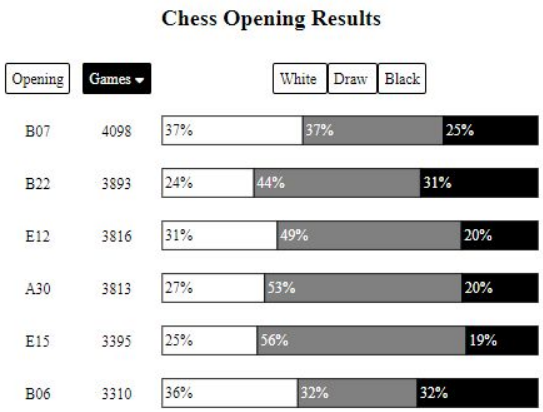
Added the sorting up and down icons to indicate which column is being sorted and if it is ascending or descending.

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.



Added the percentages to the bars.
End appearance:



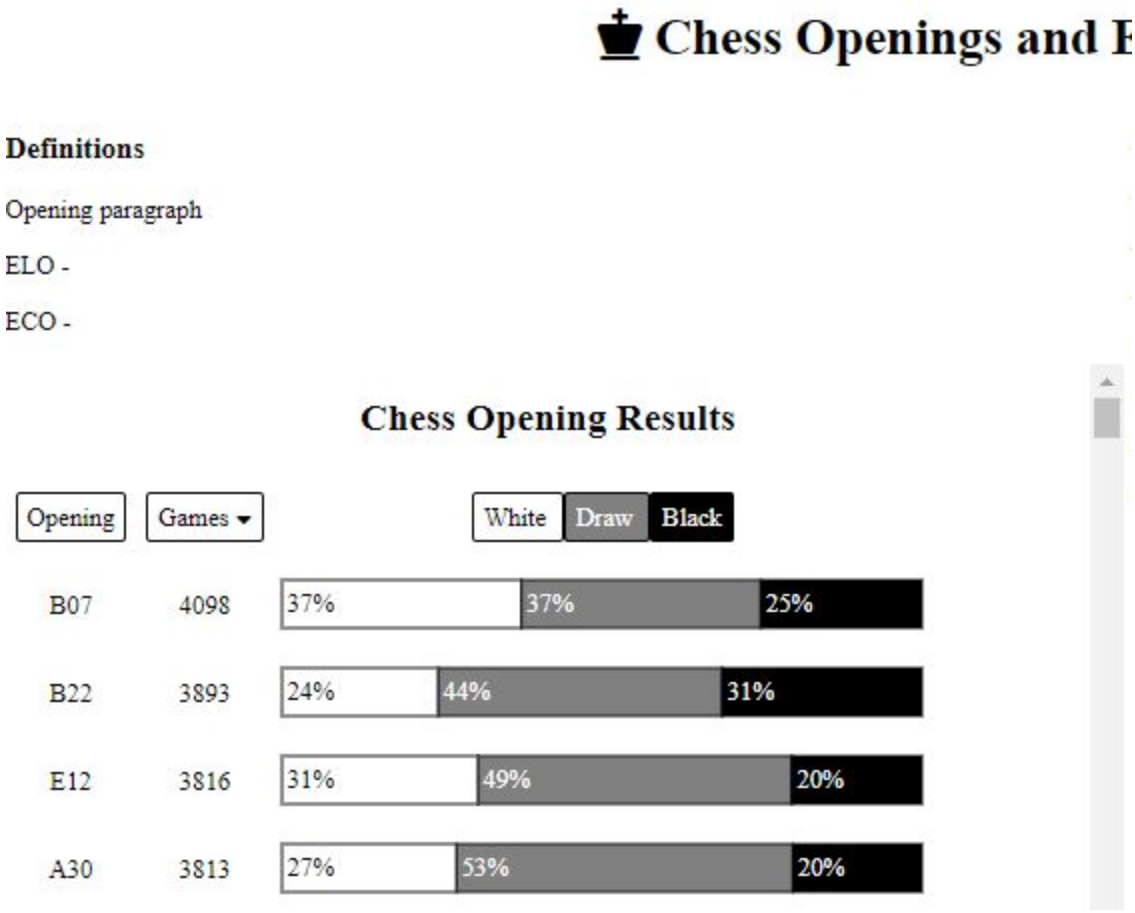


This looks very similar to the design in our ProjectProposal.pdf file. We need to make a csv which maps the ECOs to Opening Names (e.g. B07 = Pirc defence). We also need a slider at the top which filters based upon the selected date range.

Date: 11/28/20

Added definition section and made chess openings scrollable

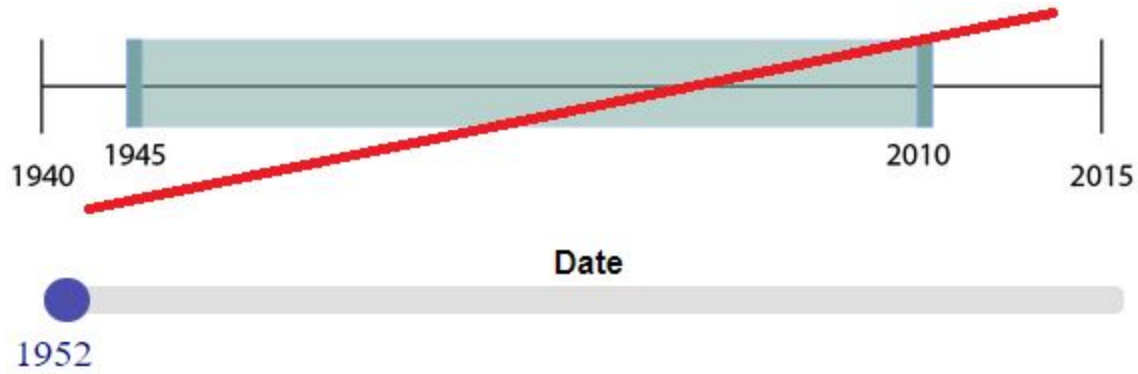
The length of the chess openings table is rather long and the website looks better if it is inside a scrollable section.



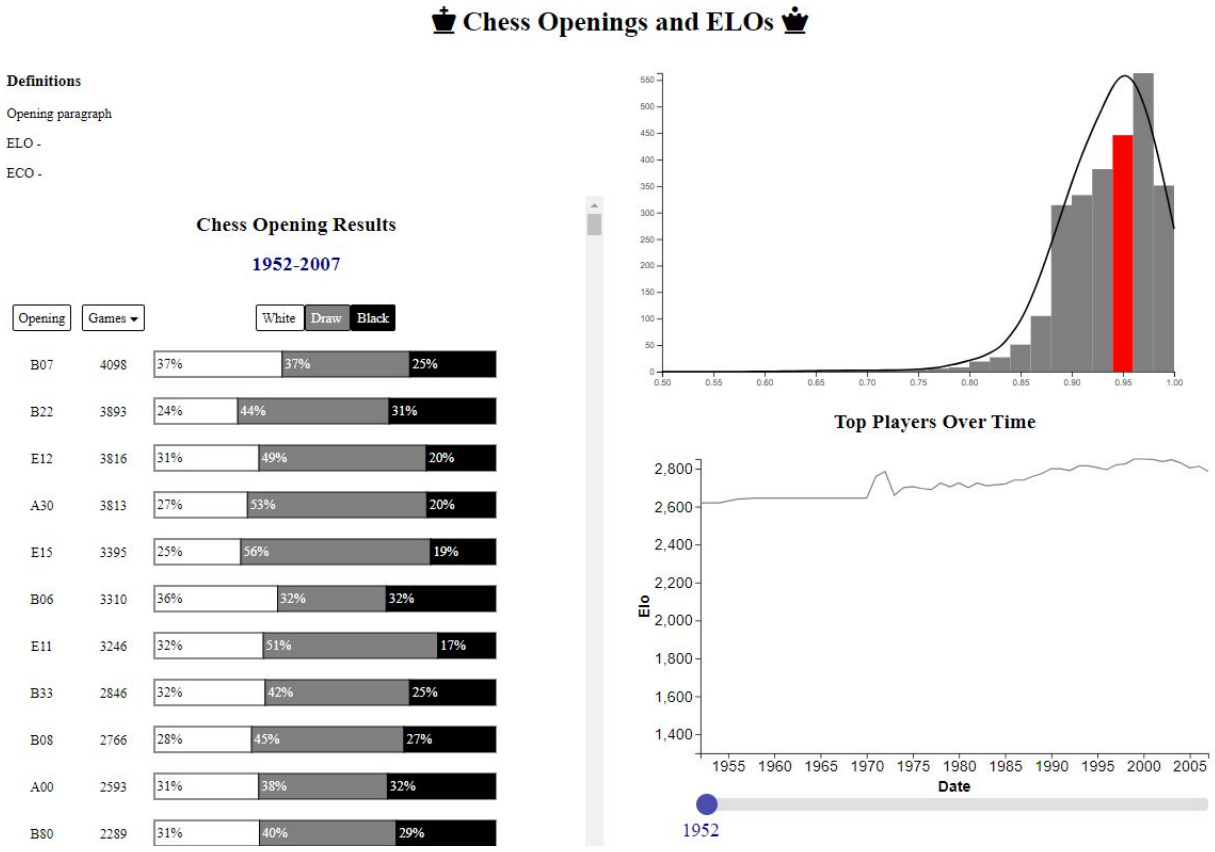
Date: 11/28/20

Deciding on a single node slider

Given the amount of time we had left to complete the project we felt that value obtained from a double node slider did not justify the time required to implement.



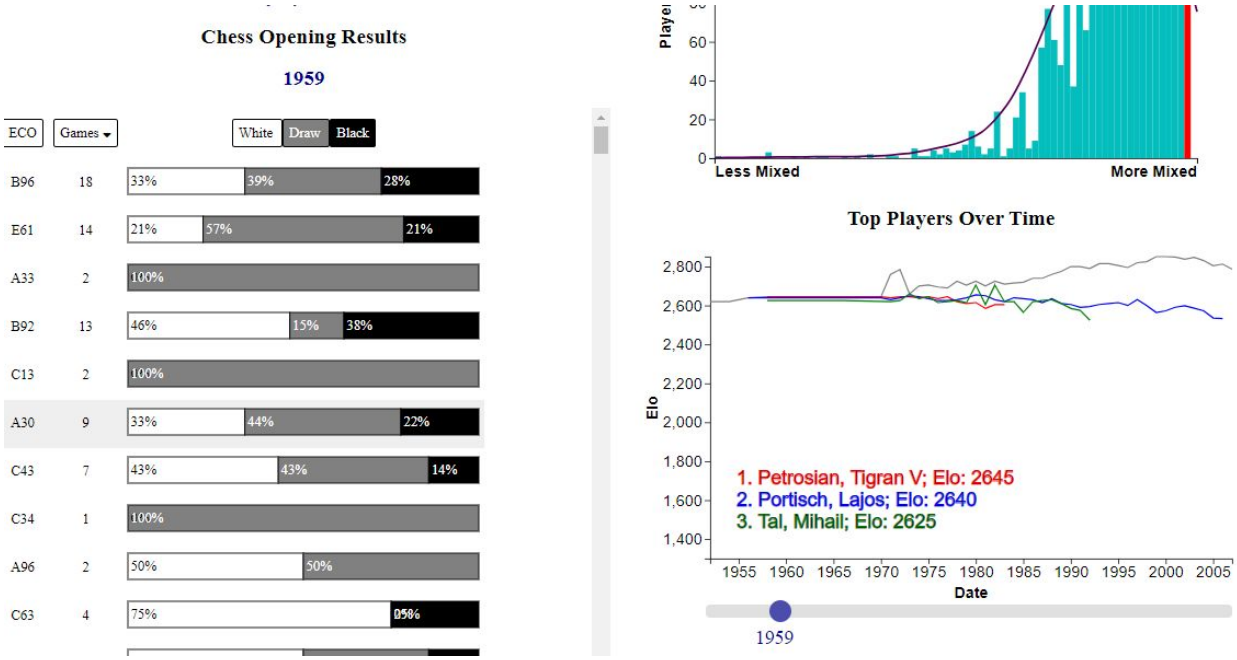
The Gini impurity chart and the chess openings will link to this single year selection.



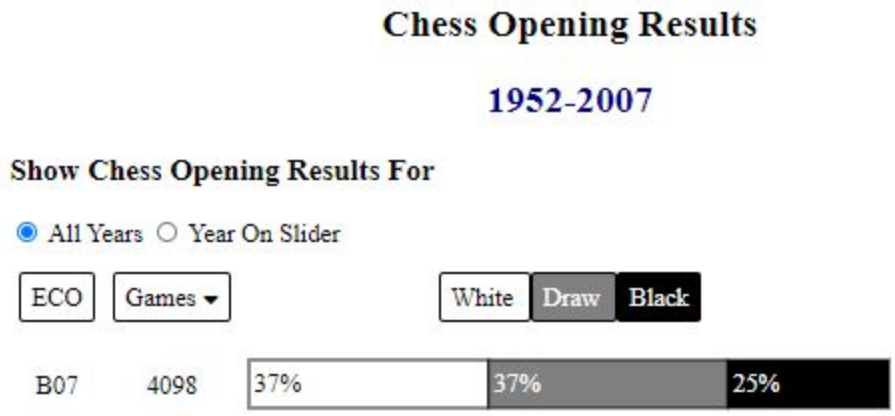
Date: 11/29/20

Chess openings table updates according to slider

The initial load of the chess openings table has data for all of the years, but when the slider moves, the chess openings table shows openings for that year and the title above shows the year. The title color matches the slider color to better show a connection.



The slider scroll speed was significantly slowed down due to the number of rows being drawn in the chess openings table and the data was being calculated at run time for the chess openings. Due to this we decided to add two radio buttons for the chess openings where the user would have to choose to see the openings which corresponded to the slider. Otherwise the default was to show data for all years. We plan on pre-processing this data to improve performance.



Date: 11/30/20

Made placeholder for process book and video in html

Garret added an interactive Python notebook which pre-processed the chess openings data so that it later can be used to improve the chess openings load performance. Essentially there are counts of the number of games, a white wins, draws, and black wins for each ECO and year pair.

♚ Chess Openings and ELOs ♚

Process Book

Insert link to process book here.

Video Explanation

Embed YouTube video here.

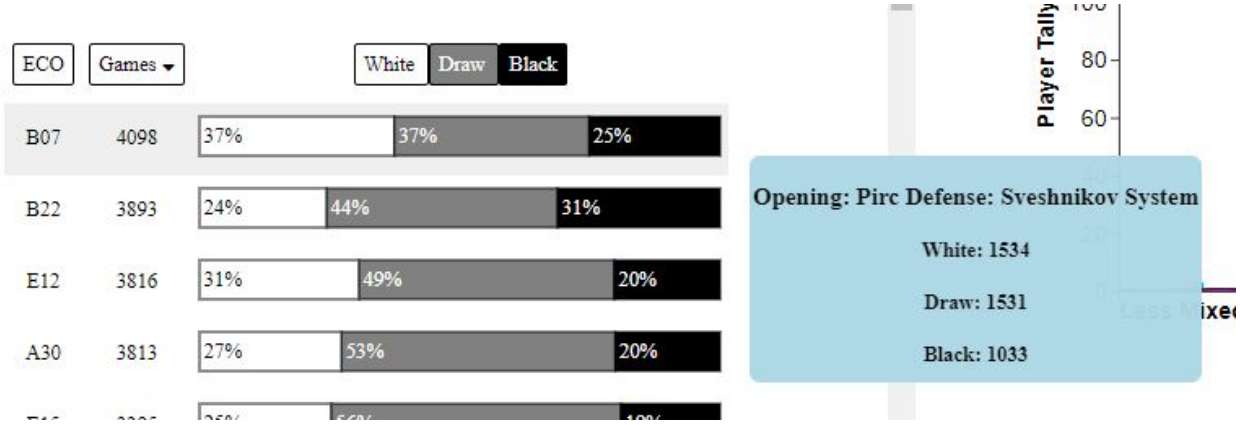
Background Information

- Elo
  - A rating system common to chess where the higher the score the better the player.

Date: 12/1/20

Added data for Elo to opening data and tooltip

Some words



When the user hovers over a row it will display the opening name and the exact count of results for white, draw, and black. It would probably be better to have the row hover color be the same as the color of the tooltip.

Date: 12/2/20

Finalized openings design

We ultimately decided against the stacked bar charts because it would make it harder to compare one row to another visually. This was considered in our first design phase. We also clarified the title and made the row hover color match the tooltip color. We also decided to place the text percentages to the right of the bars rather than inside of the bars. We thought this gave a more uniform feel to the visualization and made the percentages easier to read. Even with pre-processing the openings data, the drawing of the rows tends to still be a speed bottleneck. We kept the default settings for the openings to not be connected to the slider for that reason.

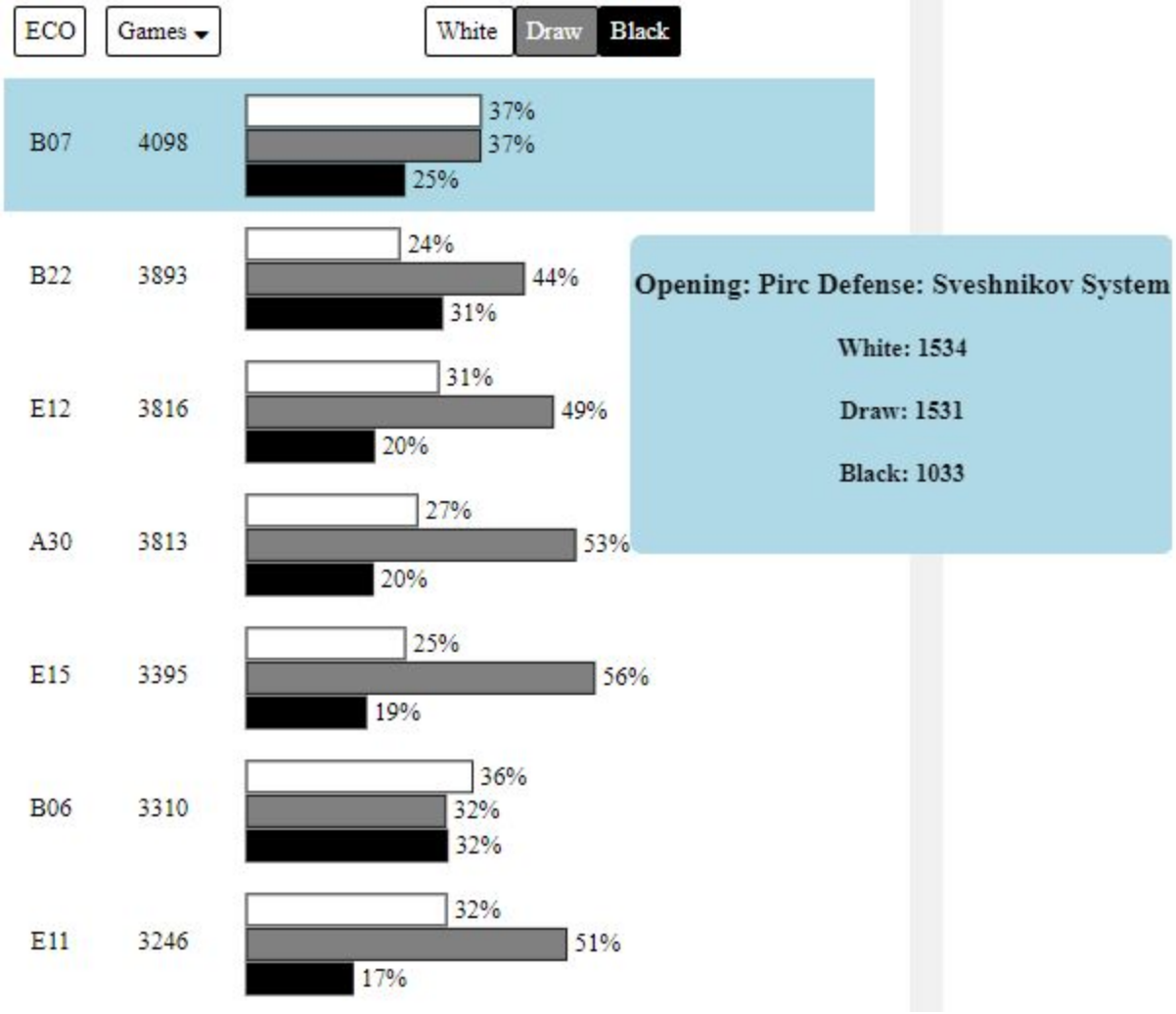
Games Results by Chess Openings

1921-2007

Show Chess Opening Results For

◉ All Years ○ "Top Players Over Time" Year

Click button below to sort



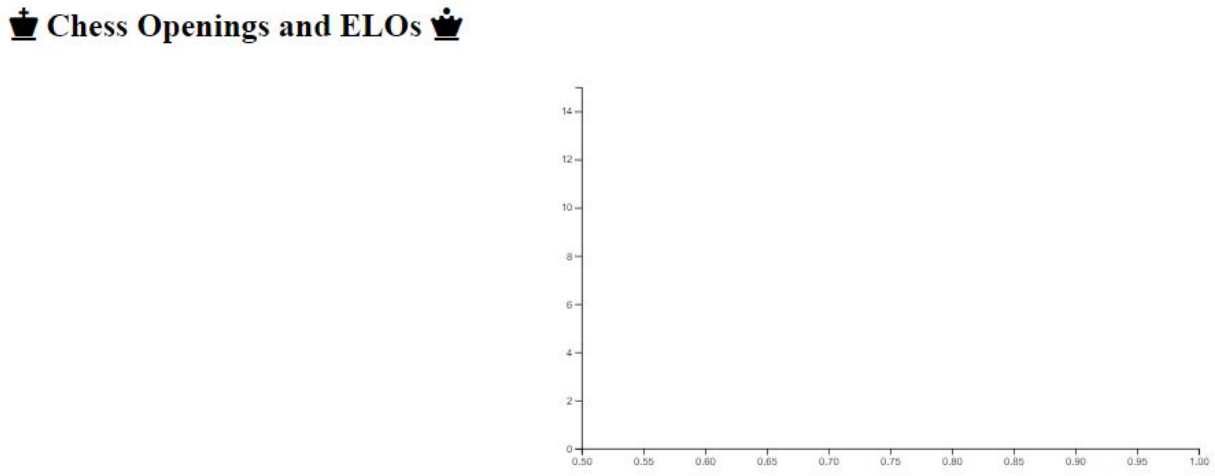
Work on “Diversity of Chess Openings for Top Players”:

Date: 11/14/20

We created GiniImpurity.ipynb to generate GiniImpurity.csv from BestPlayers.csv. This was done by grouping by the white player and then calculating the gini impurity of ECO, the openings. The result, in BestPlayers.csv was a csv with the features white name, Gini impurity, and max Elo. GiniImpurity.csv, like the other csvs, uses semicolons as the delimiter. How to calculate Gini impurity can be seen here.

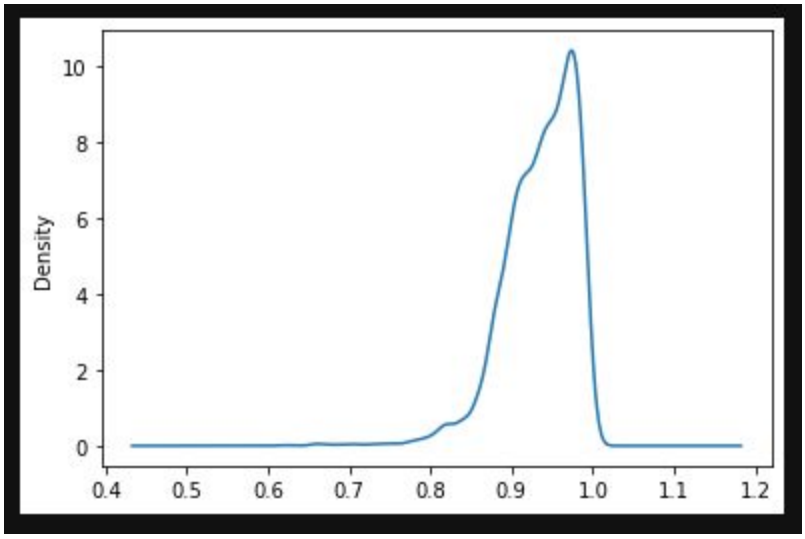
We edited the script such that loadData() returns a key-value object so that all data can be loaded in the same function. I was also able to load GiniImpurity.csv using d3.dsv which allowed us to set the semicolon as the delimiter and freed us from having to edit the csvs.

We created gini_impurity.js with some basic code. Tried to do density plot but failed. I tried to use this online example but nothing was showing. The path was showing when I inspected it but nothing was in the svg. Below is how the setup looked like.



After not being able to get the density plot to work I opened up the same data in python and created a density plot in there using pandas. This was a sanity check to get an idea what exactly the plot *should* look like. This plot is shown below and the code can be seen in GiniImpuritySanityCheck.ipynb. This allowed me to confirm that the used y-axis bounds were reasonable.

Density plot using Python

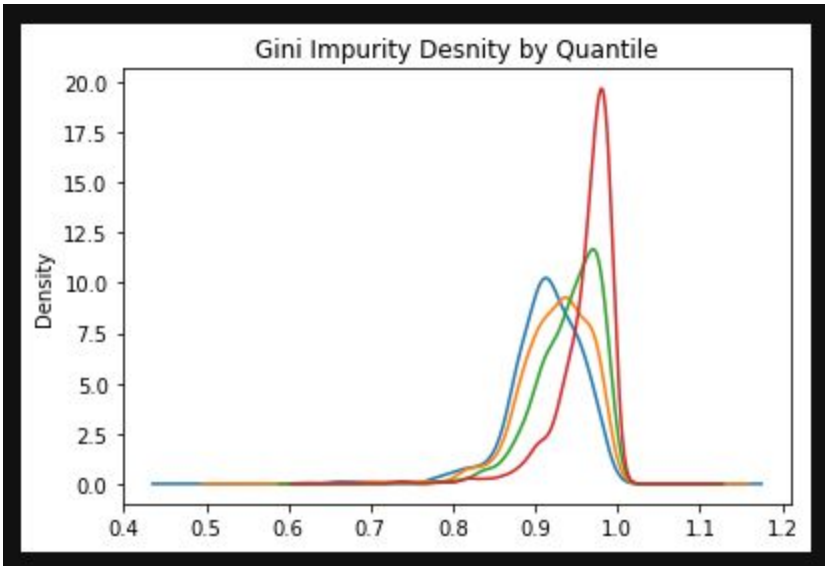


Date: 11/15/20

After looking at the sanity check I went back working on the D3 version. I found that I did have everything correct *except* that I was looking at a range along x from 0 to 1 so my starting bandwidth parameter was off by several orders of magnitude. I was able to play with the parameter and find a good value.

I edited the code to set the y-axis dynamically given the data. After completing this step and the previous step I was able to get the density plot to show.

In GiniImpuritySanityCheck.ipynb I then decided to look at the density plots for each quantile based on max Elo achieved. The quantiles, in order, are blue, orange, green, then red.



Here is when a sinking feeling fell over us. As a reminder, the idea for this visualization was to see if players with higher Elos were specializing in a few openings or playing a wide variety of openings and if that differed from players with lower Elos. If a player had a single game then the max Gini impurity of that person would be 0. If a player had two games then the max Gini impurity of that person would be able to achieve would be 0.5. If a player had three hundred games, and each game's opening was different, then the max Gini impurity of that person would be able to achieve would be 0.997. So the more tracked games in our dataset for a particular player the greater upper bound on their Gini impurity. This means that if there was a massive imbalance of games per player this could be hugely problematic and bias the data.

GamesInDataset was then added to GiniImpurity.csv then in the python sanity check script some more data exploration was done. Sure enough, the greater the Elo the more played games in the dataset. This makes sense in that people are more likely to track the game of a more prestigious player. Below are some results on the games being tracked for each quantile.

Quantile 1	
count	675.000000
mean	26.712593
std	42.210091
min	10.000000
25%	12.000000
50%	15.000000
75%	25.000000
max	569.000000

Quantile 2	
count	699.000000
mean	41.912732
std	76.468598
min	10.000000
25%	14.000000
50%	21.000000
75%	37.000000
max	751.000000

Quantile 3	
count	687.000000
mean	95.836972
std	186.239374
min	10.000000
25%	18.000000
50%	34.000000
75%	72.000000
max	1557.000000

Quantile 4	
count	657.000000
mean	302.207002
std	388.652927
min	10.000000
25%	41.000000
50%	95.000000
75%	494.000000
max	2257.000000

From this we can say that the mean number of games for those in the bottom quintile was 26 while those from the top had a mean number of games of 302. Now the issue becomes is the Gini impurity higher for players with higher score Elos because there are more games and can have a greater mixedness?

This bias essentially destroyed the original intention of the visualization. We thought about if there was a way to normalize it and were unable to. We also tried to use ECO category instead of exact ECO (e.g. A instead of A05) but the same underlying problem remained. We also experimented with putting a threshold of games in the dataset. This already existed at 10 but we tried different values. This limited the size of the total dataset.

Date: 11/16/20

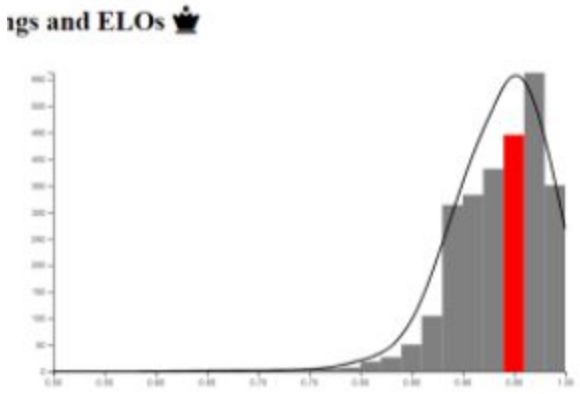
After sleeping on it this visualization was decided to be simplified in order to not lead the user to, potentially, incorrect conclusions. The difference in Gini impurity could be due to higher level players having more games or they could be due to using a wider variety of openings. The new visualization would be a histogram with a density plot overlaid on it of all players.

I also thought of another reason why the best players could have more mixedness in their openings. Maybe the best players have longer careers on average. As the game changes perhaps their games change as well. So for example common openings in 1970 might be uncommon openings in 1990. So maybe the best players specialize but they change their specializations as time goes. Another reason why the original visualization could be misleading.

Date: 11/21/20

The histogram and the density plot of the player's Gini impurity scores was created for all the best players. Below is what it looked like at this time. This histogram was done using the d3 histogram functionality. The density plot was done using a Epanechnikov distribution. The points were then converted to a path and plotted. The max of the domain for both the density and the histogram are generated from the data so they fit nicely over each other. It is important to note that the drawHistogram() takes a player name as parameter and the bin that contains that name is highlighted. This will later be used to provide interactivity between visualizations.

Results at the end of the day on 11/21/20



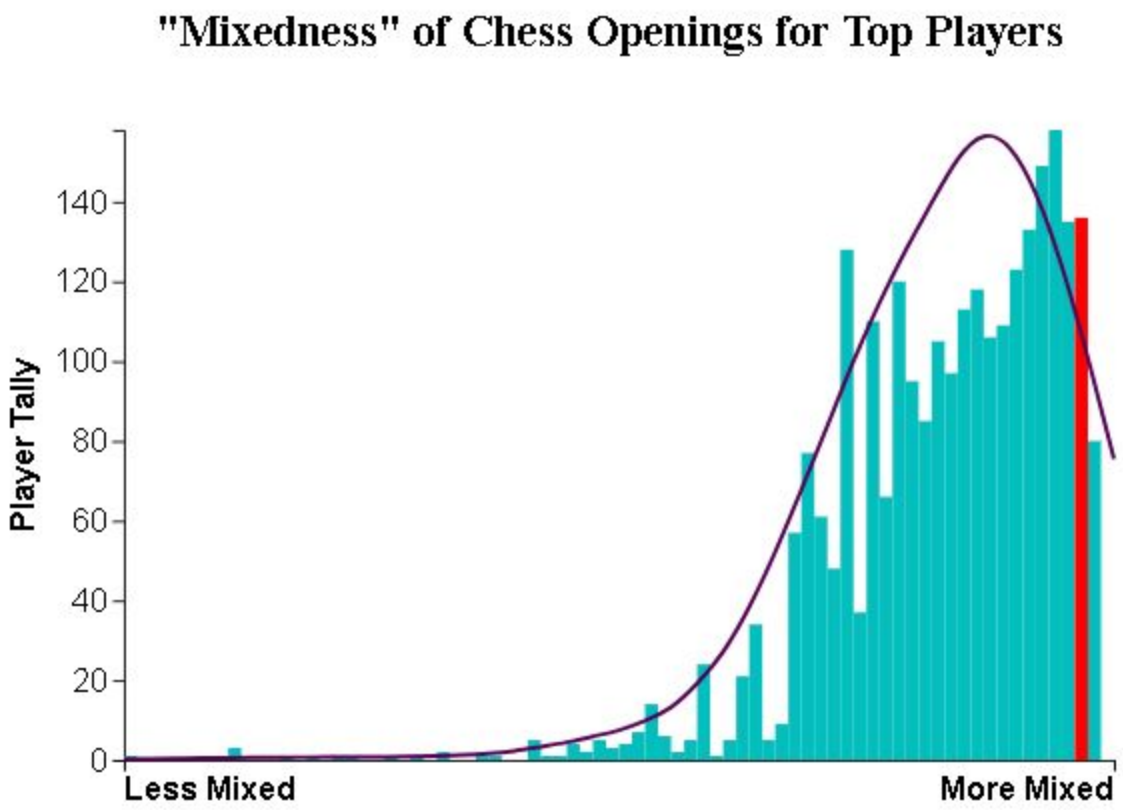
Date: 11/29/20

One piece of advice that we got from the student feedback session and from the TA feedback session is that the Gini impurity concept was that it was confusing. To combat this confusion we changed the x-axis which had decimal values between 0 and 1 to use the simplified “Less/More Mixedness” to make interpretation simpler. We understood that the simplification of this would diminish precision but we think that the increases in clarity, the ability for any user to be able to come to the conclusion that “most good players have a high mix of openings”, was worth the trade-off.

We also implemented some interaction between the “The Top Players Over Time” and the Gini Impurity plot. At any given time there is a number one player on the “The Top Players Over Time” visualization and the Gini impurity portion of the histogram that included that player would be highlighted in red. Most of the number one players were on the far right rectangle. We had to increase the number of bins in the histogram in order to observe any sort of distinction.

We also made some changes to make better use css classes and ids instead of relying directly on changing styling using d3.

Results at the end of the day on 11/29/20



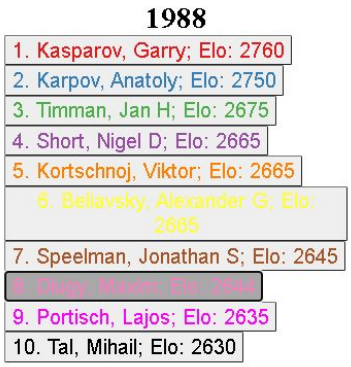
Date: 12/2/20

The word “mixed” in regards to this visualization comes from one persons (Garret T. Cervantez) one word simplification of what the Gini impurity describes when applied to decision trees. On further reflection we thought that the word “diversity” was better than “mixedness” for communication to our users so that was changed.

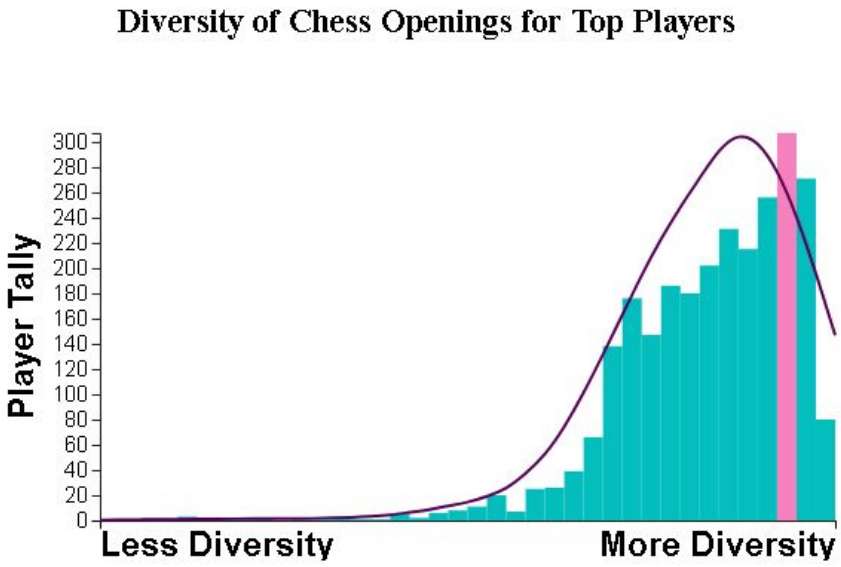
The number of bins was reduced to a more reasonable number which removed these giant changes from bin to bin which likely didn’t mean anything.

We made the highlight have more options. Before the highlight would always be the number one player for a given year. After the changes any of the top players would be selected. An example is shown below. To do this the drawHistogram() function takes an additional parameter of color.

Selection



Highlight of bin containing selection



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.

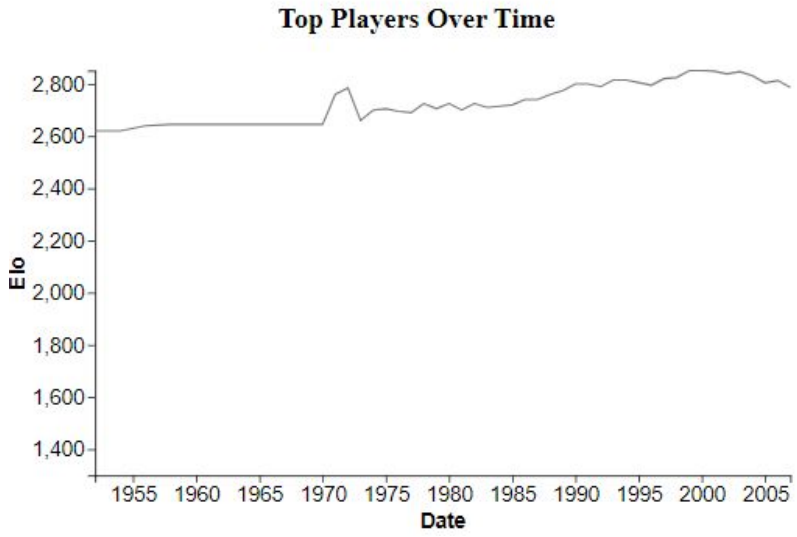
The second dataset I generated is called TopPlayerCareers.csv. A sample of this dataset is shown below.

```
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 Elo for all the players that are in TopPlayersByYear.csv.

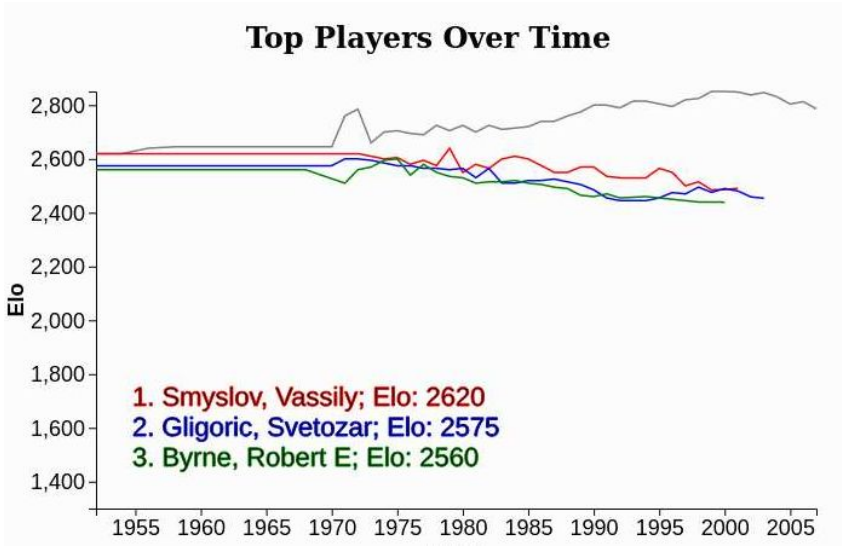
Date: 11/14/20

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.



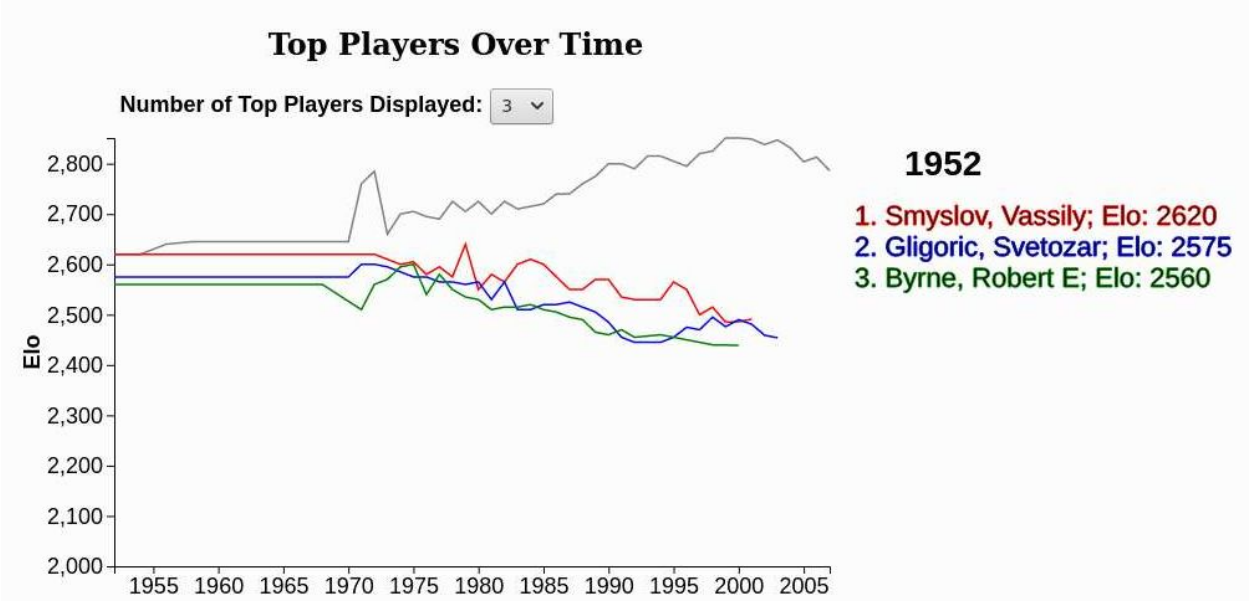
The remaining work to be done for this plot is to display the career paths of the top players for the selected year. This will involve adding a slider and displaying additional paths which have already been created.

Date: 11/29/20



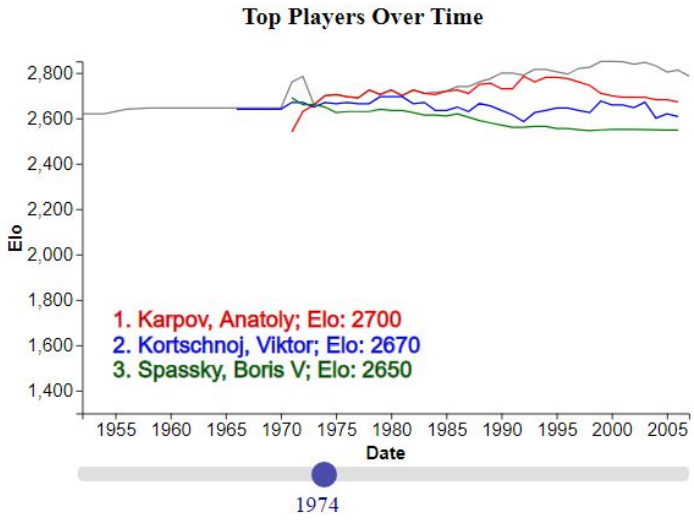
This update to the Top Players plot shows that it now has the functionality to plot the top three players for a given year.

Date: 11/29/20



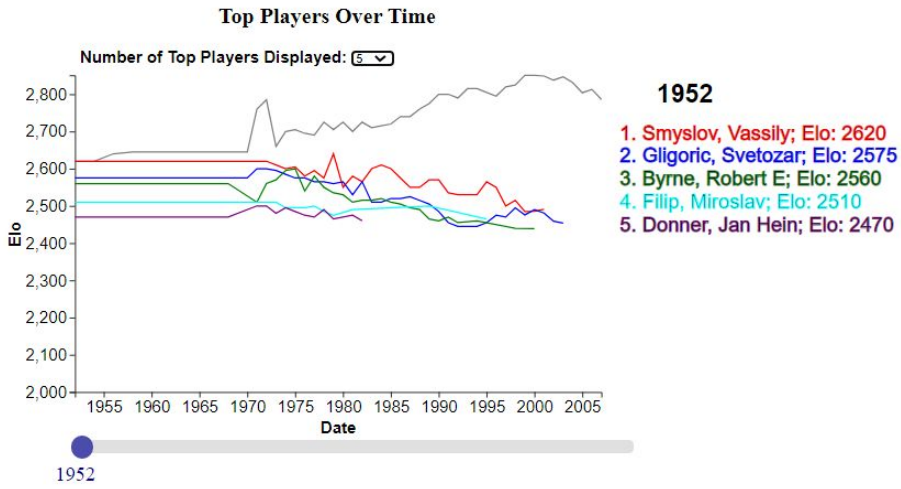
This iteration allows for users to plot more than 3 top players (they can choose from 1, 2, 3, 5, or 10). Additionally, we moved the location of the player names to the right since they were sometimes overlapping the player career paths.

Date: 11/30/20



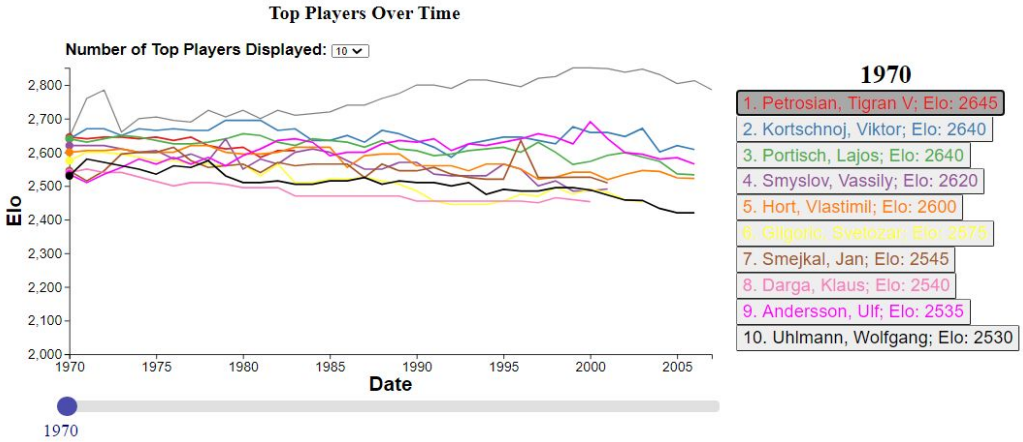
This next iteration has a scrollable bar that allows for users to select the current date that is being selected (it was developed in tandem with the previous iteration so it contains some outdated aspects).

Date: 11/30/20



This next iteration essentially shows the combination of the two previous iterations. It allows users to select the number of top players to display and it includes the scrollable date slider.

Date: 12/1/20



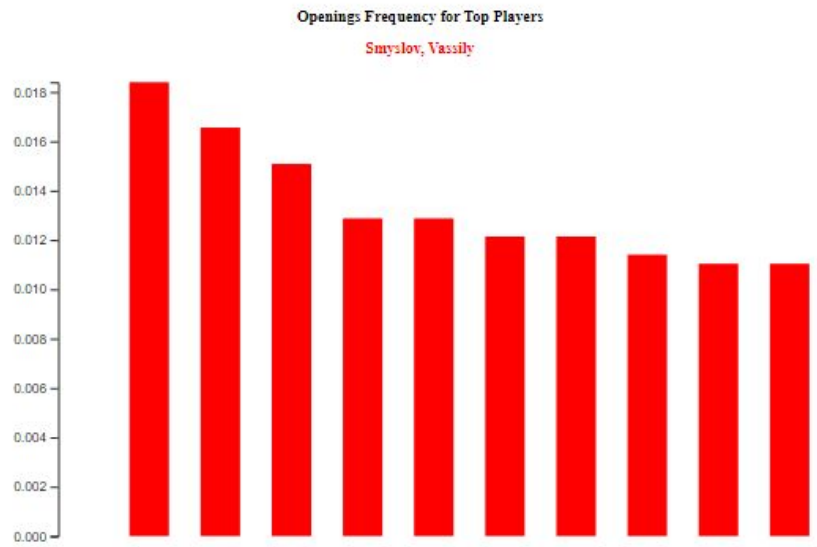
This final iteration shows the addition of buttons when displaying the top players. These buttons make it so that specific players can be selected and displayed in other visualizations. For example, the selected player will have their top openings frequency chart displayed and their bucket within the Gini Impurity histogram highlighted. It does this with a callback function that is passed to the RankingPlot object on creation. This callback function then passes player name, color, and any other appropriate information to the necessary other visualizations any time a new player is selected.

Another thing this final iteration did was limit the range of player dates. We originally provided data all the way back to 1952, but the data before 1970 was very sparse, so we decided to eliminate it in order to emphasize the more meaningful data in the rest of the plot.

Work on “Most Common Openings for a Top Player”:

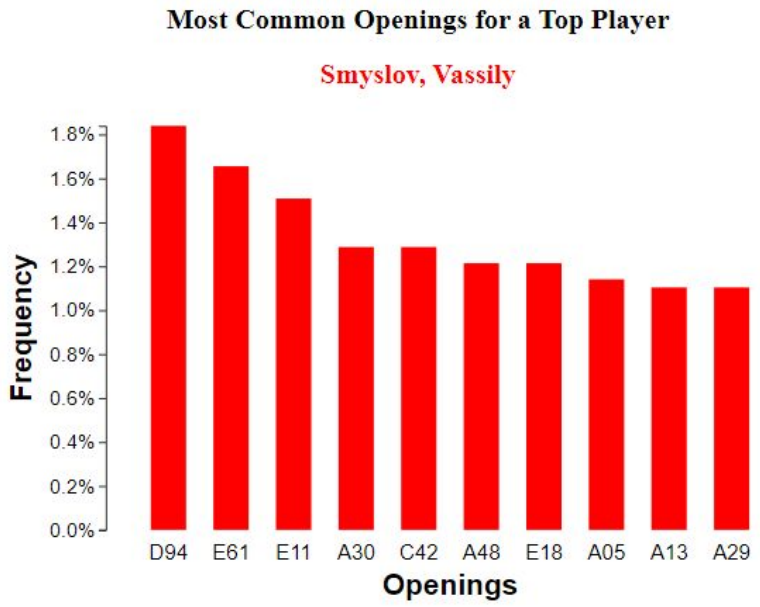
This visualization was a last minute addition to the project, so all of the work was done right before the deadline. The goal behind this visualization was to be able to select users from the rankings plot and show their most used openings. This would allow us to see in even greater detail whether the best players use a wide variety of openings or specialize in a few. Additionally, it would allow for direct comparisons between players.

Date: 11/30/20



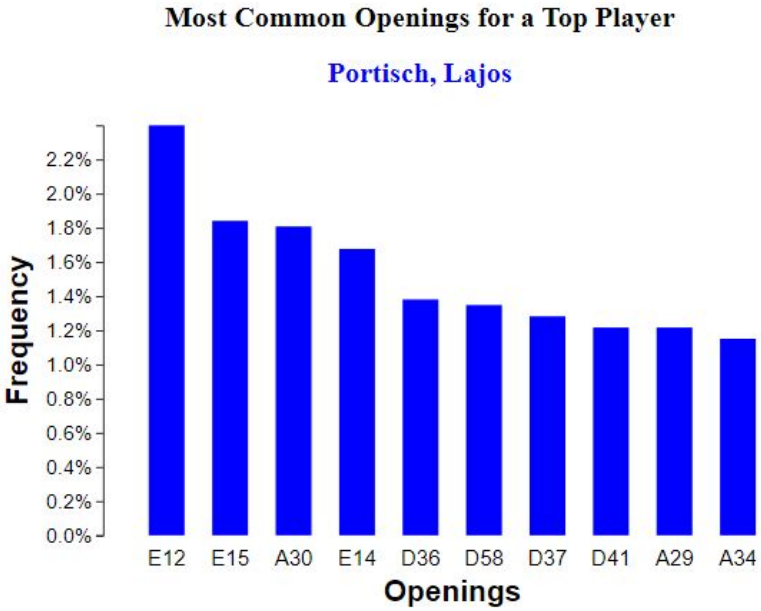
This first iteration simply plotted the Frequency of a player's top 10 openings. There are no axis labels, and it is impossible to tell what the actual openings are. Additionally, this plot is not yet linked to any other plot, so there is no way to set the player being plotted other than manually.

Date: 12/1/20



This second iteration now has nicer axes and labels. Additionally, it labels the individual openings being plotted. It is still not connected to any other visualization, however, and there is no way to see the full opening names.

Date: 12/1/20



This third iteration is now linked to the playerRankings plot and changes color to correspond with the selected player.

Date: 12/1/20

T his final iteration enables hovering over the bars to see additional statistics about the player's use of the opening and the opening's full name. Additionally, this design has a different color since we changed the color scheme for the visualization.

Merged Top Players, Preparing for First GitHub Release:

Date: 11/15/20

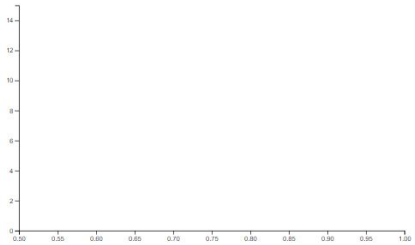
Uncommented Chess Openings Results (The initial loading time for this is a little slow, look to improve this). Merged branches and our repository is ready to release.

Appearance:

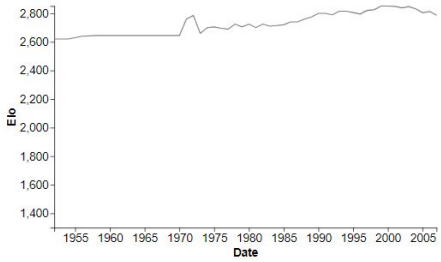
♚ Chess Openings and ELOs ♚

Chess Opening Results

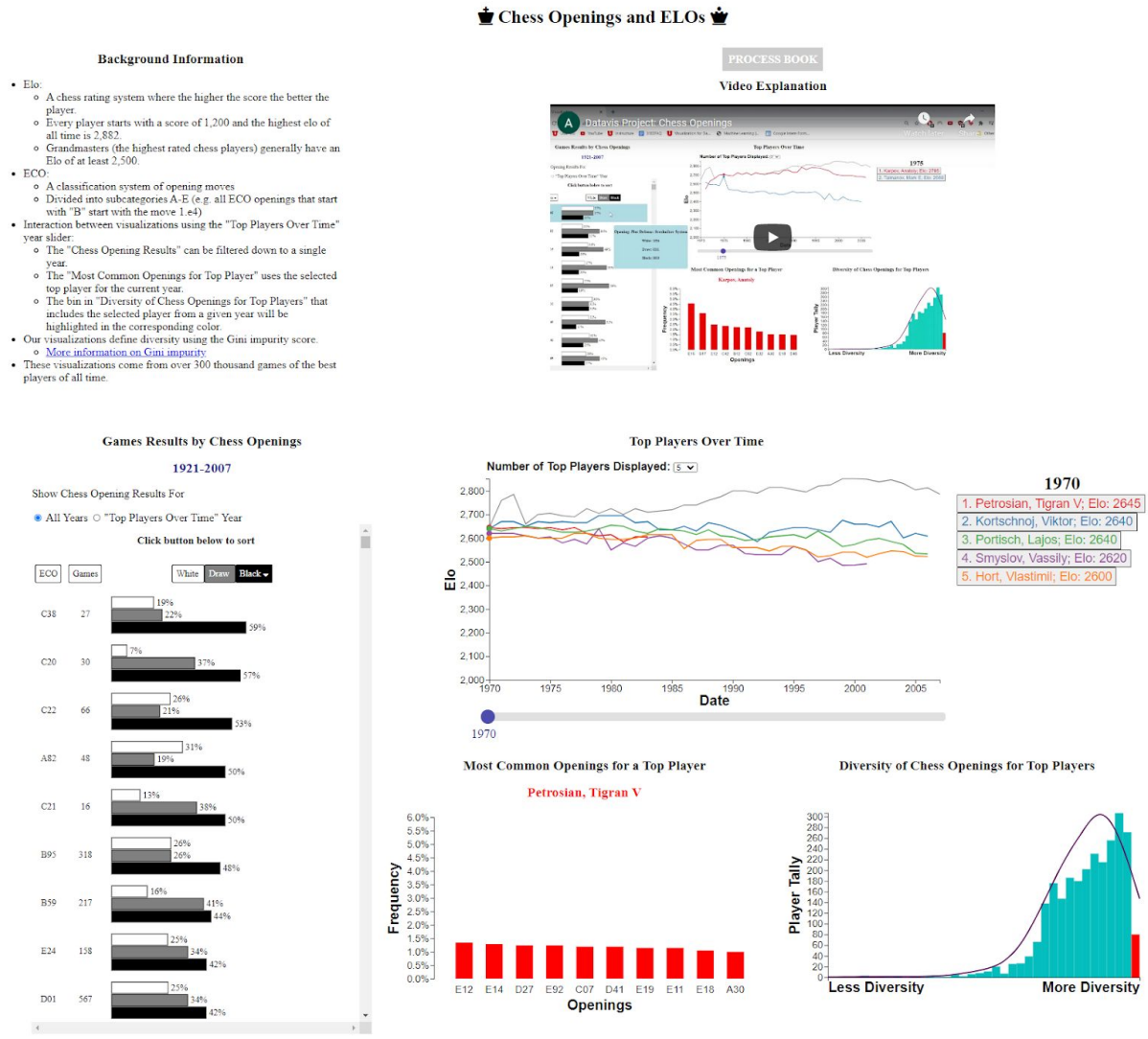
Opening	Games	White	Draw	Black
B07	4098	37%	37%	25%
B22	3893	24%	44%	31%
E12	3816	31%	49%	20%
A30	3813	27%	53%	20%
E15	3395	25%	56%	19%
B06	3310	36%	32%	32%
E11	3246	32%	51%	17%
B33	2846	32%	42%	25%
B08	2766	28%	45%	27%
A00	2593	31%	38%	32%
B80	2289	31%	40%	29%
E94	2262	40%	39%	21%
B42	2209	33%	39%	28%
C42	2181	29%	61%	11%



Top Players Over Time



Final Appearance



Evaluation

Things we learned about chess

We created several interesting visualizations and each led to their own insights. From looking at “Game Results by Chess Openings” one can gain practical information about historically what openings have had the highest success rate. One could imagine a chess aficionado looking for the best openings in order to prioritize which to study and use. There are also questions raised about certain types of openings that may be extremely conservative. For

example the our dataset has the opening D14 used 750 times with draws occurring 86% of the time. Other explanations are possible, but this could indicate that this opening could be a low-risk/low-reward option.

When looking at the “Top Players Over Time” we can see what an outlier Robert “Bobby” Fischer was. The visualization shows what a brief and spectacular career Bobby Fischer had. He only played for a few years and he achieved an Elo score of 2,785 which was the highest score until 1990 when Garry Kasporov achieved a 2,800. Another interesting highlight was seeing how long the career of Garry Kaspoarov was. He was the highest ranked player starting in 1984 and remained the highest rated player (with the exception of 1985 when he was the second highest rated player) until 2006 when was toppled by Veselin Topalov. It is incredible to think about a man staying at the top of his game for over three decades while facing multiple generations of opponents.

The “Most Common Openings for a Top Player” shows how different players use different openings. For example we can see that E15 is a popular opening for many top players. We can also see differences in how often openings are played from player to player. For example, Anatoly Karpov used his most frequent openings a much greater percentage of his games than Garry Kasporov.

In some ways, “Most Common Openings for a Top Player” delivers on the original intention of “Diversity of Chess Openings for Top Players” better. “Diversity of Chess Openings for Top Players” does still show that the top players tend to use a diverse set of openings.

Finally, filtering the chess openings table by year allows us to see which openings are the most popular overall, and whether they align with those used by the best players from that year. It is very interesting to see how much the most used openings vary from year to year. For example, the Modern Defense was the most popular opening from 1970-1973, but by 1976 it wasn't even in the top 10 most popular. Chess is always evolving as players find new openings and counters, so playing novel and modern openings is clearly important.

What could be done to improve our visualizations

One overarching thing that could be done is to change the layout. It is a bit difficult to see all the visualization on the screen at the same time if you don’t zoom out or have a high resolution display. The year slider under “Top Players Over Time” controls all the other visualizations so it should actually probably be located in a more prominent location, perhaps spanning the entire width of the visualization and located above “Games Results by Chess Openings” and “Top Players Over Time”. Our visualization also doesn’t utilize transitions. One place it might be interesting to do that would be to draw the lines in “Top Players Over Time”.

In “Games Results by Chess Openings” it would be nice if the buttons at the top that control ordering were outside the scrollable section. Currently these buttons disappear once the user scrolls down. One other feature that would be nice would be to add a number of games threshold. For example, assume that a user wants to see what the best openings historically are. The user would likely sort by white and see that the 4th best opening for white is C23. On closer examination though the user notices that there are only 5 games. The user might wish to *only* see openings where there are at least 100 games but currently that is not an option. This

visualization is monochromatic to match the monochromatic palette of chess but we wish we could have thought of something to make things more aesthetically appealing.

In “Top Players Over Time” there are players like Bobby Fischer that do not have games every year. The data doesn’t show that very clearly. One way to do this would be to make periods with no games a dashed line instead of a solid line. This visualization also has an option to show up to the top ten players for each year. When the full ten are selected the visualization can be hard to read and, because each rank is color coded, we have a yellow rank which is difficult to see. One option, that would probably make everything look nicer, would be to move away from the white background towards a charcoal, dark theme. Hovering over a line could also highlight the line in some way (e.g. thicker line) which could make things more clear. A great feature in addition to the highlight of a line on hover would be if you could also select a player by clicking on a line.

The “Most Common Openings for a Top Player” is a simple visualization and achieves its aims well. One way to improve this visualization would be the ability to compare multiple players at the same time by putting other color coded rectangles next to the current rectangle. By doing this the user could more easily get an idea about what openings are unique to certain players and which shared among top players. If a way to select players from different eras were possible then one could get an idea about the different openings from different eras.

One change regarding the “Diversity of Chess Openings for Top Players” would have been to use the entropy instead of the Gini impurity. After some cursory research they are similar in some ways but entropy would probably be a better metric. One option that could improve the visualization would be a slider that could control how many bins to use in the histogram. The original question this visualization was trying to answer was are the best players generalists or specialists when it comes to openings? We fell short of answering this question and decided to answer a simpler question, do the best players have a lot of diversity or very little diversity in their opening moves? With enough time we would love to figure out how to answer that first question better. One potential solution that comes to mind would be using bootstrapping in order to get better estimates of the true values of the entropy for every player.