CHESS BEST OPENING INDEX

Data analysis and visualization - CA1 Specification Index Generation and Visualization

Alexandre Desbos

Table of Contents

Theorical framework: ***2***

Methodology ***2***

Data selection and preparation **2**

Data Distribution 3

Data Aggregation and Modifications 5

Data Cleaning 6

Data Normalization 6

Multivariate Analysis **7**

Effectiveness Indicator 7

Complexity Indicator 9

Popularity Indicator 10

Improvement Indicator 10

Overall Index 11

Findings ***13***

The Results **13**

Link to other Indicators **14**

Theorical framework:

In chess, the selection of an opening often plays a pivotal role in determining a player's success. To address this, I have developed the "Best Chess Opening Composite Index", designed to quantify the multifaceted nature of chess openings. This index integrates data across three concept: effectiveness, popularity, and complexity, offering a comprehensive resource for players at all skill levels to take well-informed decisions about their opening strategies. By adjusting the weighting of indicators, I also generate 2 supplement indexes, one for beginner and one for experiment players to allow a choice of opening levels.

The data underpinning this index come from an extensive database (https://www.kaggle.com/datasets/alexandrelemercier/all-chess-openings) that encompasses a vast array of recorded games. This diverse dataset ensures that the index is robust and reflective of strategies employed across the entire spectrum of the chess-playing community. The variables integrated into the index include quantifiable measures such as win and draw percentages, frequency of opening utilization, and number of moves of an opening.

By using this data, the index provides a nuanced view of the strategic value of different openings. It serves as a tool for strategic preparation and decision-making, enabling players to choose openings that not only align with their personal style and strengths but also enhance their chances of winning or securing a draw.

Methodology

All code available: https://github.com/alexandredesbos/DAV-CA1

Data selection and preparation

The dataset contains a variety of variables, so the first step was to analyse them to select the variables I can use to build my sub-indicators (complexity, popularity, improvement, effectiveness) , which will then enable me to create my index.

Data select:

- **Number of game** : The total number of games played with this opening.
- **Perf Rating**: The average performance rating of players who have played this opening.
- **Player Rating**: The overall average rating of players in the dataset.
- **Player Win %:** The win rate for players using the opening.

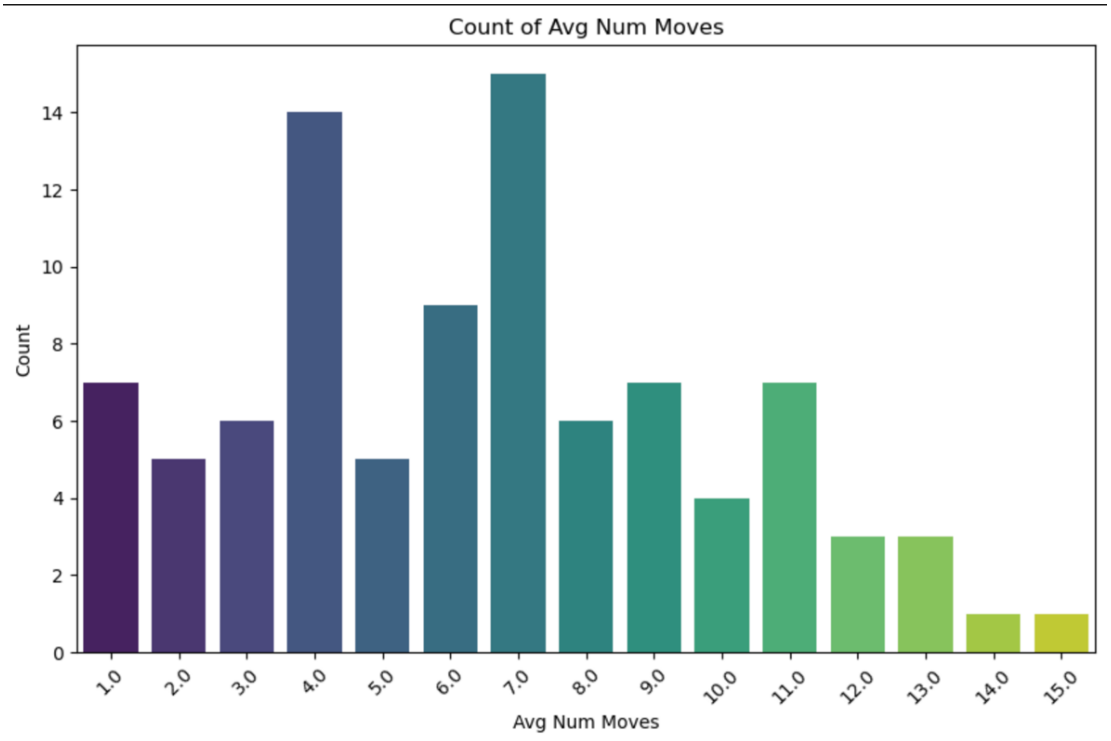
- **Draw %**: The percentage of games that ended in a draw.
- **Opponent Win %:** The win rate against players using the opening.
- **Moves List**: A comprehensive list of all moves made in the opening sequence.

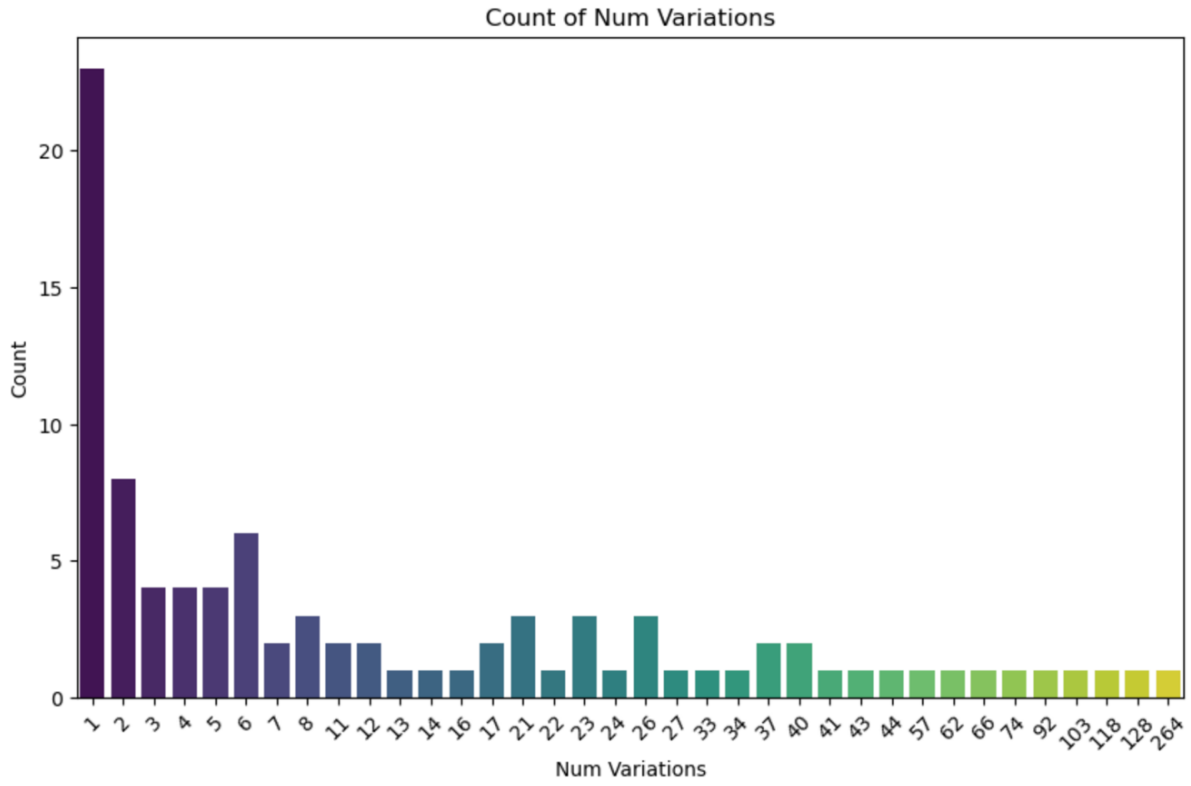
Data create using the dataset:

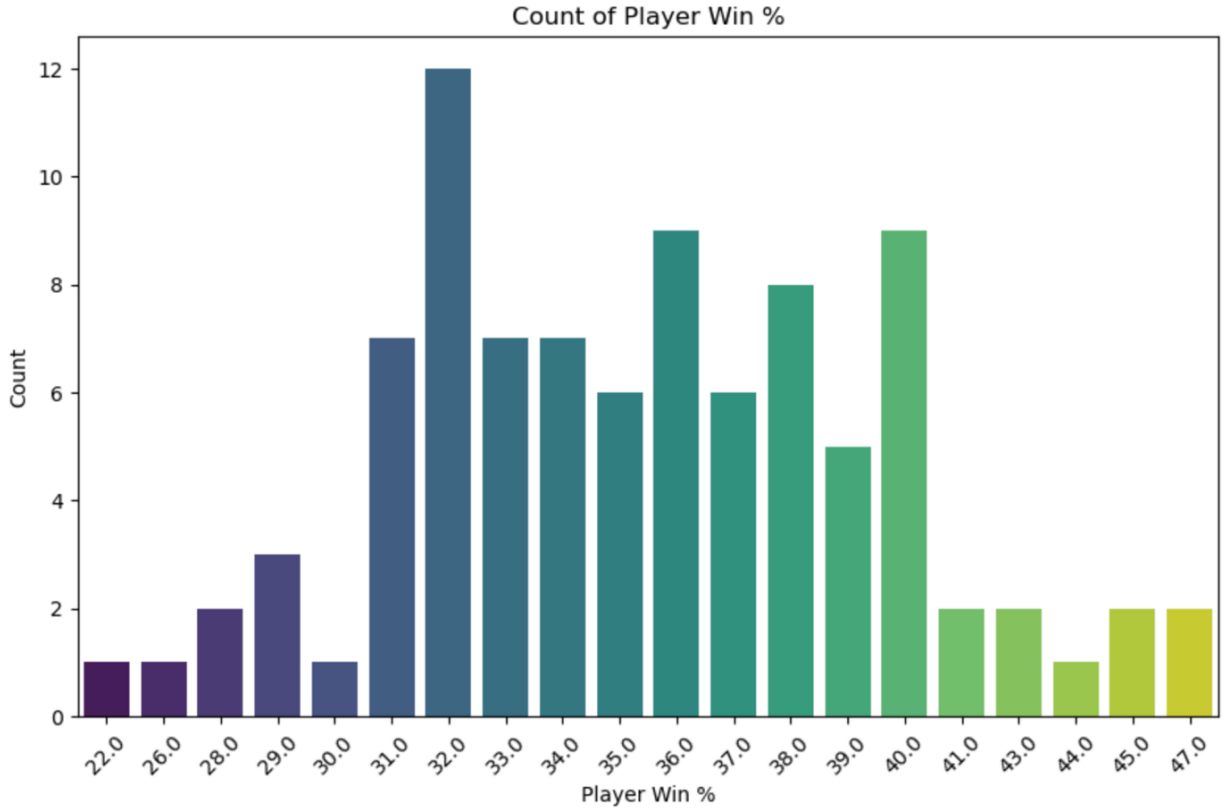
- **Number of variations:** The number of variations for each opening
- **Delta Perf:** Difference between player Rating and his performance rating

Data Distribution

Plot of variable distributions to assess the modifications and aggregations to be made.




Count of Num Variations
Count
Num Variations
0
5
10
15
20
1
2
3
4
5
6
7
8
11
12
13
14
16
17
21
22
23
24
26
27
33
34
37
40
41
43
44
57
62
66
74
92
103
118
128
264


Count of Player Win %
Count
Player Win %
0
2
4
6
8
10
12
22.0
26.0
28.0
29.0
30.0
31.0
32.0
33.0
34.0
35.0
36.0
37.0
38.0
39.0
40.0
41.0
43.0
44.0
45.0
47.0

Data Aggregation and Modifications

In order to have more relevant data, I had to make aggregation and modification of variables.

First of all, I change the variables "*moves list*" to "*number of moves*" by changing it to an int of the length of the list.
Then, in the dataset I'm using, the openings are divided for each variation of it, so I had to aggregate the data by opening. For each variation of an opening I added up the number of games and averaged the other variables (perf rating, player rating, player win%, draw%, opponent win%, number of moves).
This also allowed me to create a new variable, 'Number of variations', by adding up the number of variations for each opening, which I'll be able to use for the Complexity sub-index.
Finally, I have calculated and add a Delta Perf variable by calculating the difference between the average rating and the performance rating. This variable is useful for determining whether a player can get a better score for a game with a specific opening than for all his games.

Before aggregation and modifications:

	Opening	Num Games	Perf Rating	Avg Player	Player Win %	Draw %	Opponent Win %	moves_list
0	Alekhine Defense, Balogh Variation	692	2247	2225	40.8	24.3	35.0	['1.e4', 'Nf6', '2.e5', 'Nd5', '3.d4', 'd6', '4.Bc4']
1	Alekhine Defense, Brooklyn Variation	228	2145	2193	29.8	22.4	47.8	['1.e4', 'Nf6', '2.e5', 'Ng8']
2	Alekhine Defense, Exchange Variation	6485	2244	2194	40.8	27.7	31.5	['1.e4', 'Nf6', '2.e5', 'Nd5', '3.d4', 'd6', '4.c4', 'Nb6', '5.exd6']

After aggregation and modifications:

	Opening Name	Num Games	Perf Rating	Avg Player	Player Win %	Draw %	Opponent Win %	Avg Num Moves	Num Variations	DeltaPerf
0	**Alekhine Defense**	34710	2207.925925925930	2208.44	36.133	26.78	37.08	7.62	27	-0.51
1	**Anderssen Opening**	1308	2124.0	2126.0	35.7	25.6	38.7	1.0	1	-2.0
2	**Benko Gambit**	24543	2245.058823529 4100	2229.29	40.13	25.17	34.68	10.58	17	15.76

Data Cleaning

I ensured the consistency of the data by cleaning it up. This process involved:

1. **Removing Duplicates**: No duplicates were found in the dataset

```
Number of duplicate: 0
```

2. **Handling Missing Values**: I didn't need to impute any data because there were no missing values.

```
Opening Name      0
Num Games         0
Perf Rating       0
Avg Player        0
Player Win %      0
Draw %            0
Opponent Win %    0
Avg Num Moves     0
Num Variations    0
DeltaPerf         0
dtype: int64
```

Data Normalization

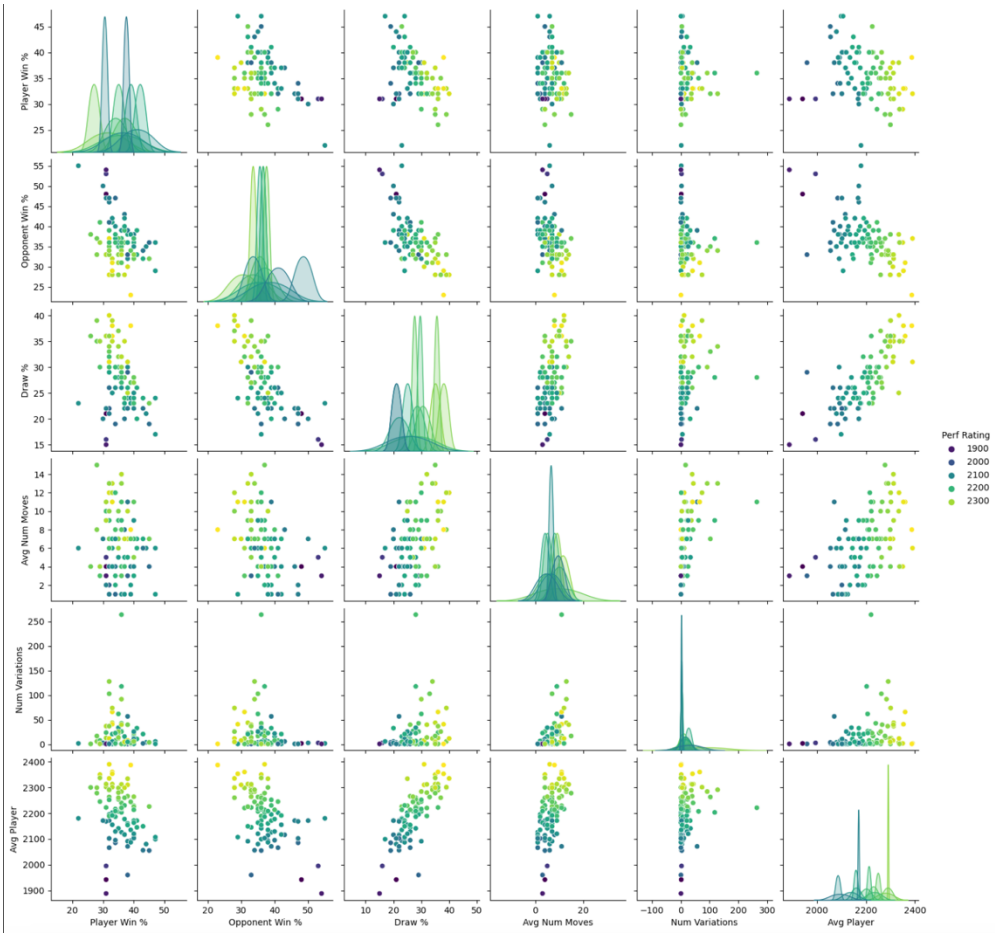
To ensure that each variable contributed equally to the final index without bias from different scales or units, I applied the Min-Max normalization technique. This method was used for each variable and enabled them to be scaled within a range of 0 to 1.

Before applying the normalisation on the “Number of games”, I applied a Log function to make it more symmetric because it was skewed data that gave me not relevant results.

Data after normalization:

Opening Name	Num Games	Perf Rating	Avg Player	Player Win %	Draw %	Opponent Win %	Avg Num Moves	Num Variations	DeltaPerf	Log Num Games
Alekhine Defense	34710	0.67	0.63	0.55	0.45	0.56	0.52	0.90	0.55	0.65
Andersse n Opening	1308	0.51	0.470	0.54	0.41	0.51	1.0	1.0	0.54	0.28
Benko Gambit	24543	0.75	0.68	0.71	0.39	0.63	0.31	0.94	0.65	0.61

These standardised data form the basis of my multivariate analysis, making it possible to construct a reliable and significant composite index of chess openings and to easily choose the contribution of variables with weightings.



Multivariate Analysis

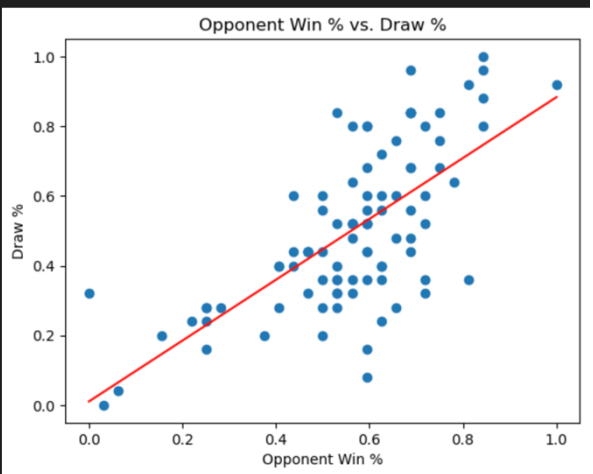
I made a multivariate analysis for each of my sub-indicators and then on my overall index to make sure my indexes are consistent.

Effectiveness Indicator

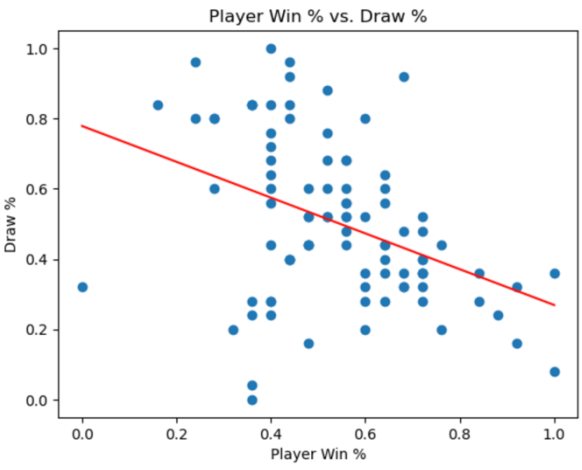
- Player Win %
- Draw %
- Opponent Win %

=>Direct outcomes when the opening is used.

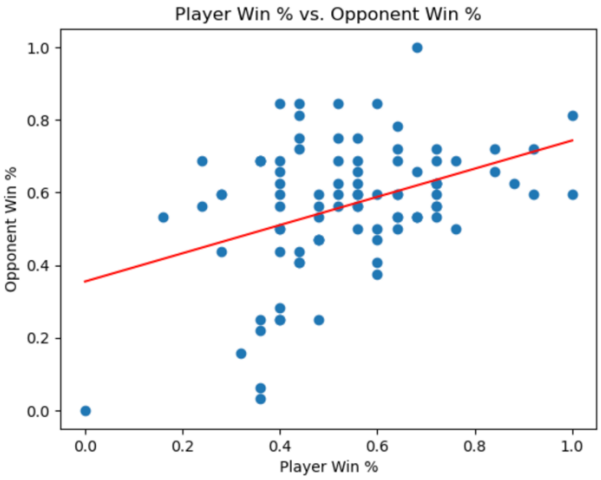
```
PearsonRResult(statistic=0.6810766000663936, pvalue=5.817675071124944e-14)
```




```
PearsonRResult(statistic=-0.4016984580204248, pvalue=6.582121431480686e-05)
```




```
PearsonRResult(statistic=0.3924268427527624, pvalue=0.00010003094742932919)
```




```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Player Win %   R-squared:                       0.989
Model:                            OLS   Adj. R-squared:                  0.988
Method:                 Least Squares   F-statistic:                     3944.
Date:                Sun, 05 May 2024   Prob (F-statistic):           2.27e-88
Time:                        20:01:46   Log-Likelihood:                 235.85
No. Observations:                  93   AIC:                            -465.7
Df Residuals:                      90   BIC:                            -458.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3267      0.007     48.755      0.000       0.313       0.340
Draw %        -0.9841      0.012    -81.606      0.000      -1.008      -0.960
Opponent Win %     1.2555      0.015     81.245      0.000       1.225       1.286
==============================================================================
Omnibus:                        4.126   Durbin-Watson:                   2.207
Prob(Omnibus):                  0.127   Jarque-Bera (JB):                4.666
Skew:                           0.145   Prob(JB):                       0.0970
Kurtosis:                       4.058   Cond. No.                         11.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
0.019475344222886482
```

We can see that the model have a good R-squared and adjust R-square, the standar errors is very low (good).

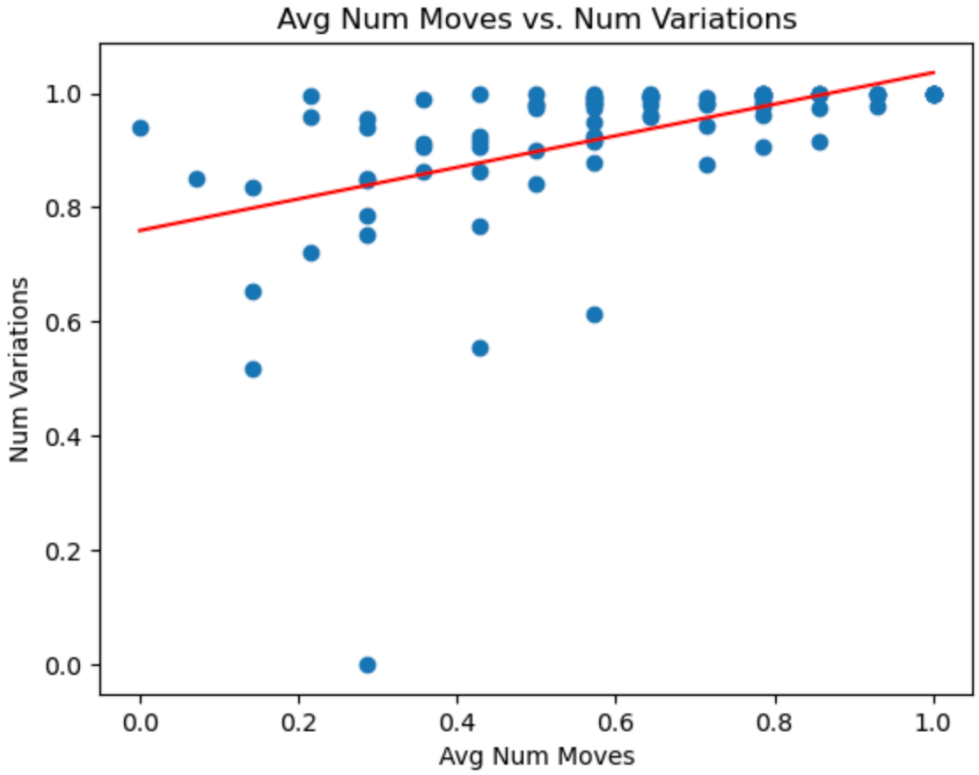
Complexity Indicator

To build the complexity indicator, I wanted to use two variables, so I performed a scatterplot analysis to spot and avoid potential multicollinearity.

Variables for complexity indicator:

- Number of moves: The length and complexity of the opening moves can indicate strategic depth.
- Number of variations: The number of possible move sequences can reflect the complexity of the opening.

Results of the scatterplot analysis and correlation:



PearsonRResult(statistic=0.5471878958641806, p-value=2.8813028407624773e-08)

We can see the person correlation coefficient value is 0.54, that represent a moderate relationship between the two variables and the p-value is way smaller than 0.05 so it's very significant.
With this analysis, I decided to keep both of my variables because a correlation of 0.55 is generally not high enough to cause concerns on about multicollinearity and I believe both

variables contribute unique information to the index and reflect different aspects of the complexity of chess opening.

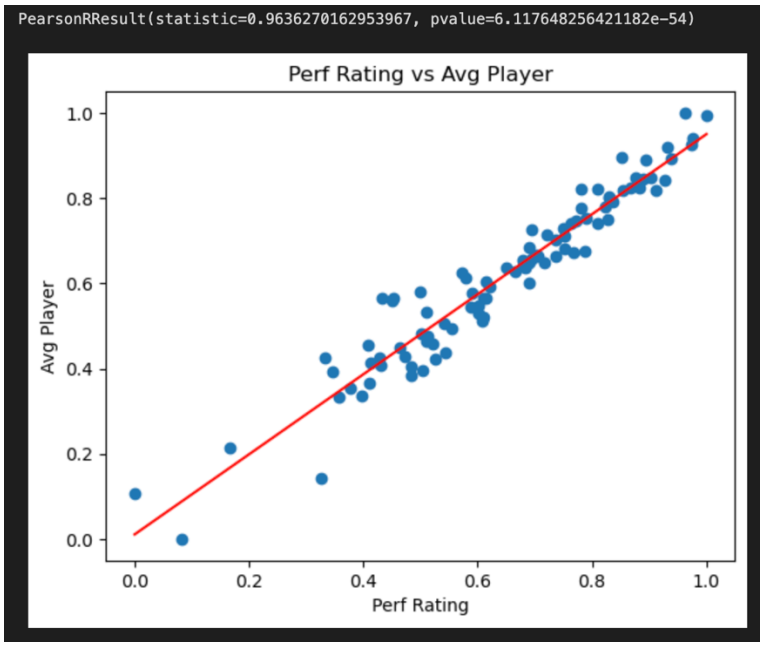
Popularity Indicator

The popularity indicator uses only the number of games played, so no analysis has been carried out for this indicator.

- Num Games: Represent the number of times the opening is used.

Improvement Indicator

The improvement indicator also uses only on variable, the Delta perf to represent if in average, playing an opening can have a better performance rating than player rating. It's better to use Delta perf than perf rating and Avg rating because they are highly correlated.

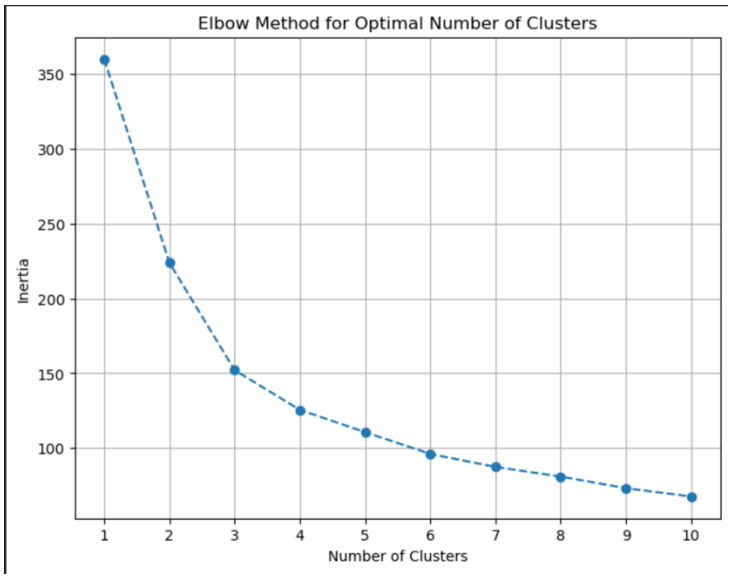


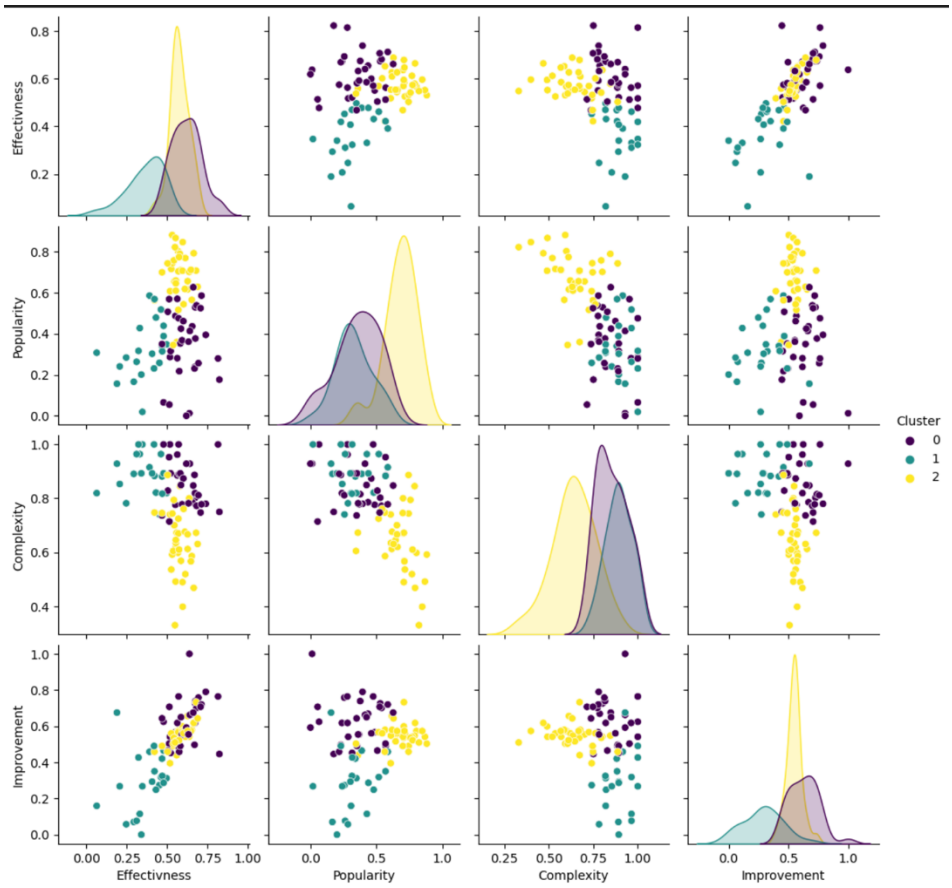
The Avg player and Perf Rating are highly correlated (0.9636) and this is very significant (pvalue < 0.05)

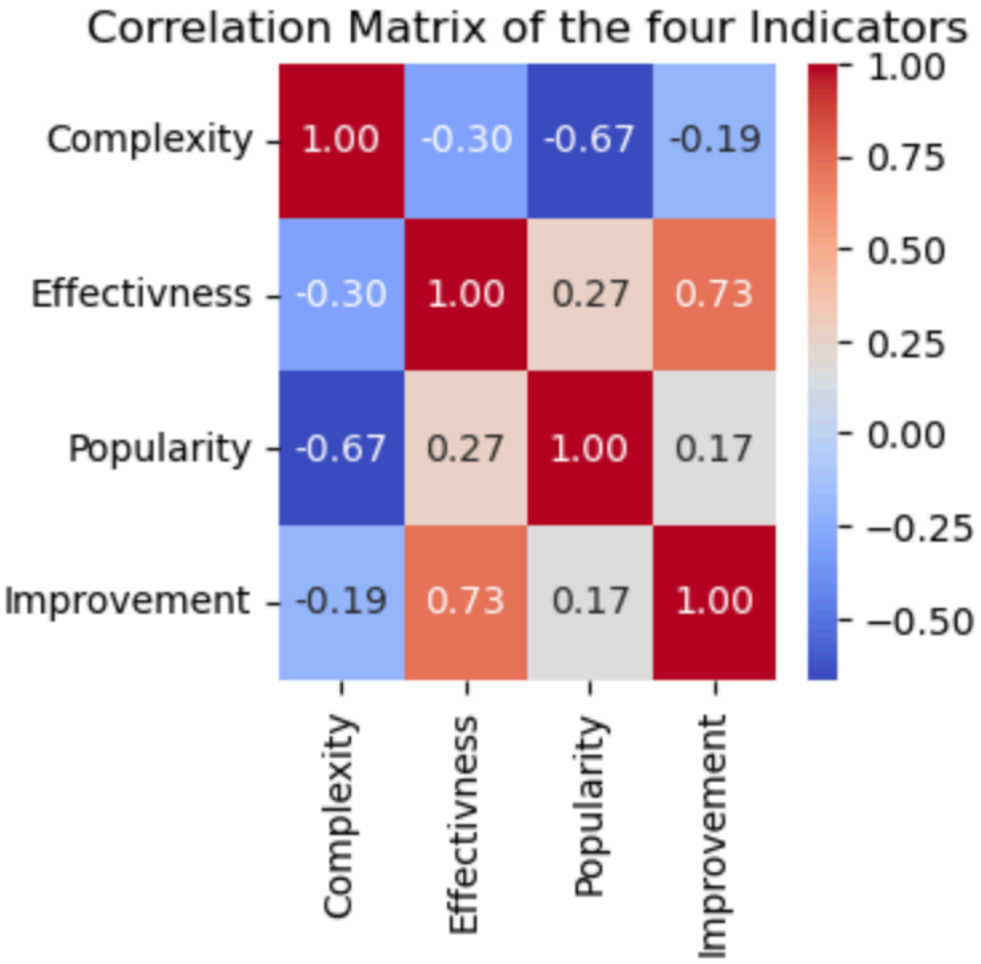
- Delta perf : Difference between player Rating and his performance rating

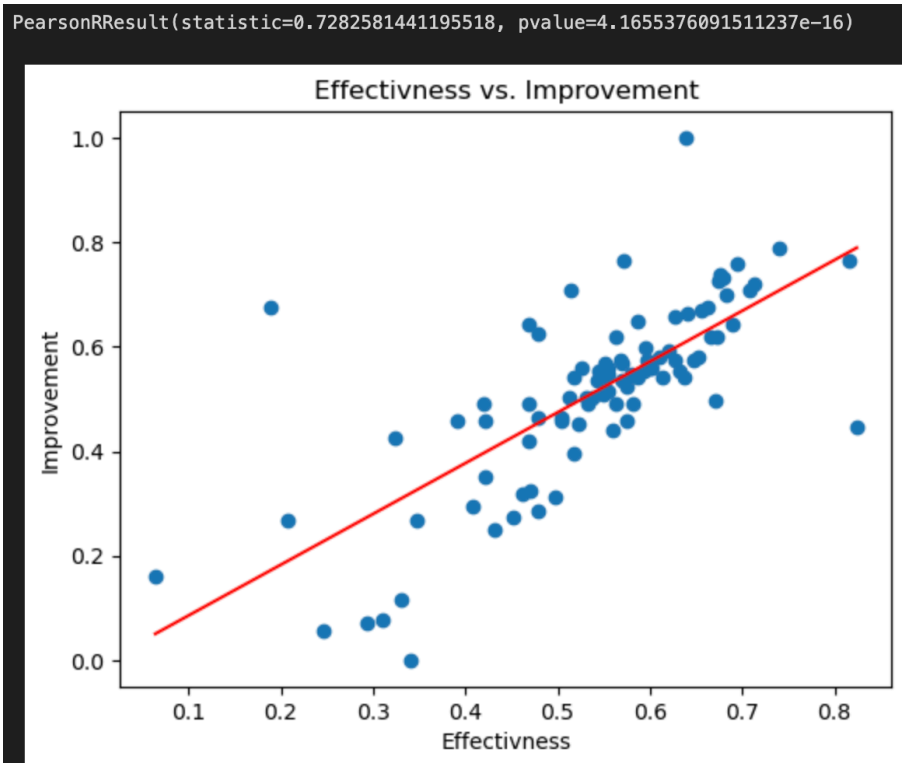
Overall Index

Cluster Analysis






Correlation Matrix of the four Indicators
Complexity
Effectivness
Popularity
Improvement
1.00
-0.30
-0.67
-0.19
-0.30
1.00
0.27
0.73
-0.67
0.27
1.00
0.17
-0.19
0.73
0.17
1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Complexity
Effectivness
Popularity
Improvement


PearsonRResult(statistic=0.7282581441195518, pvalue=4.1655376091511237e-16)
Effectivness vs. Improvement
Improvement
Effectivness
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

The correaltion Matrix of the indicators show that Effectiveness and Improvement look highly correlated and this is confirm by the scatterplot, the p-value is < 0.05 so the corelation is very significant.

To avoid multicollinearity, I remove the Improvement index.

Findings

The Results

The complete results from the index, including the overall result of 10 best opening as well as the individual rankings within the 3 categories. (complete result in the file final_data.csv)

Opening Name	Overall Index		Opening Name	Popularity
Indian Game	0.722		**English Opening**	0.882
Queen Pawn Opening	0.710		**French Defense**	0.867
Trompowsky Attack	0.707		**King's Indian Defense**	0.847
King's Indian Attack	0.705		**Spanish Game**	0.822
Torre Attack	0.694		**Caro-Kann Defense**	0.803
Zukertort Opening	0.690		**Queen's Gambit Declined**	0.793
Pirc Defense	0.686		**Indian Game**	0.790
Queen Pawn Game	0.682		**Slav Defense**	0.774
King's Pawn Opening	0.671		**Nimzo-Indian Defense**	0.769
Bishop's Opening	0.669		**Queen Pawn Game**	0.758

Opening Name	Effectiveness		Opening Name	Complexity
Indian Game Defense	0.824		**Clemenz Opening**	1.0
Queen Pawn Opening	0.815		**Van't Kruijs Opening**	1.0
Lion Defense	0.739		**Kadas Opening**	1.0
Torre Attack	0.713		**King's Pawn Opening**	1.0
Vienna Game	0.707		**Saragossa Opening**	1.0
Latvian Gambit	0.694		**Queen Pawn Opening**	1.0
Russian Game	0.6891		**Anderssen Opening**	1.0
Old Indian Defense	0.682		**Owen Defense**	0.964
Pirc Defense	0.679		**Horwitz Defense**	0.964
Center Game	0.675		**Polish Defense**	0.964

I also create 2 more index, based on Effectiveness, complexity and popularity but with different weight, one is for the Experienced player so it can be complex opening, the second is for beginner so it has less complexity

Opening Name	Overall Advanced		Opening Name	Overall Beginner
Queen's Gambit Declined	0.690		Queen Pawn Opening	0.819
Pirc Defense	0.690		King's Pawn Opening	0.809
Indian Game	0.690		Nimzo-Larsen Attack	0.788
King's Indian Attack	0.679		Zukertort Opening	0.781
English Opening	0.677		Indian Game	0.781
King's Indian Defense	0.677		Horwitz Defense	0.774
French Defense	0.672		Nimzowitsch-Larsen Attack	0.768
Torre Attack	0.668		Trompowsky Attack	0.766
Trompowsky Attack	0.666		Anderssen Opening	0.760
Nimzo-Indian Defense	0.666		Bird Opening	0.757

Link to other Indicators

I didn't find any index similar to mine, I add bellow some ranking of chess opening but they are make only using one indicator.

Here is a ranking of the chess opening by popularity, we see that the 3 first are also in my ranking respectively 9, 2,1.

- [1] 1. e4, *King's Pawn*, 1362K games --> [4], [8]
- [2] 1. d4, *Queen's Pawn*, 1048K --> [3], [9]
- [3] 1. d4 Nf6, *Indian Game*, 622K --> [6]
- [4] 1. e4 c5, *Sicilian Defense*, 606K --> [5]
- [5] 1. e4 c5 2. Nf3, *Open Sicilian*, 483K --> [14]
- [6] 1. d4 Nf6 2. c4, *Indian Game: Main Line*, 429K --> [13]
- [7] 1. Nf3, *Réti Opening*, 299K games
- [8] 1. e4 e5, *Open Game*, 296K --> [10]
- [9] 1. d4 d5, *Closed Game*, 263K games --> [15]
- [10] 1. e4 e5 2. Nf3, *Open Game: Main Line*, 262K --> [11]
- [11] 1. e4 e5 2. Nf3 Nc6, *Open Game: Main Line*, 226K games
- [12] 1. c4, *English Opening*, 211K games
- [13] 1. d4 Nf6 2. c4 e6, *East Indian Defense*, 209K games
- [14] 1. e4 c5 2. Nf3 d6, *Modern Sicilian*, 195K games
- [15] 1. d4 d5 2. c4, *Queen's Gambit*, 191K games

Here we can see a ranking of the chess opening but he is only based on the white win% and draw%. We can see that this ranking is not corresponding to my ranking based on the effectiveness.

White's Best 10 Openings

(Based on White Win % + (Draw % x 0.5))

Rank	Opening	White Win %	Draw %	Points per 100 games
1	**Queen's Gambit**	**40**	**36**	**58**
2	Blackmar Diemer Gambit	49	16	57
3	Ruy Lopez	40	33	56.5
4	Bishop's Opening	41	30	56
5	Benko Opening	38	35	55.5
=5	Reti Opening	37	37	55.5
=5	Vienna Game	41	29	55.5
8	Centre Game	44	22	55
=8	English Opening	38	34	55
=8	Scotch Game	40	30	55