Understanding chess with data analysis

Yair Vogel

# Introduction

Playing Chess, every match is unique. The amount of possible Chess moves is huge and grows exponentially as game progresses. That's why playing chess is hard. A player has to think and calculate his next moves constantly, while considering the rival moves, even 7 or 8 moves ahead.

As long as I remember, I loved playing Chess. The time controlled mental challenge charmed me. As a Chess player, you always learn. Every game can be analyzed providing new insights, but there comes a time when even that is not enough.

Here data analysis comes into play. All my past games, as we'll see in this document, were full of new insights. In my opinion data analysis can give unique and special insights about trends that usually evade the micro lens.

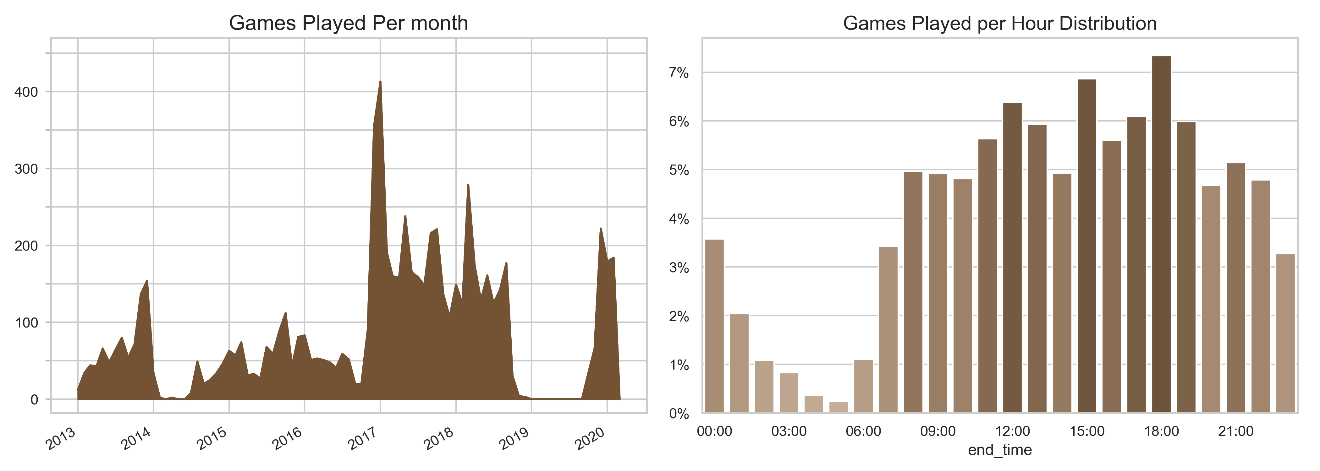
All the chess studies try to find the best play for every player. In this article, I tried to focus not on the 'best' chess moves available, but to focus on *my* games. The idea behind this approach is that in the amateur levels it's hard to find the best move every time, so I wish to conduct a personalized research to improve my game. The measurement in this article will not be the theoretical best moves, but the moves I encounter in my play and that will down-to-earth and relevant conclusions.

With Chess.com beautiful API, I was able to extract very detailed information about all my past games. I was thrilled about this data and immediately went and analyzed it.

I used python with Pandas and NumPy libraries for data extraction and manipulation. I used Matplotlib with Seaborn to visualize the data.

# Overview

Every time I get access to a new dataset, I love playing with it to understand it better. Let's start with understanding when the games were played:



Games Played per month: We can observe very distinct periods across the last 7 years. During 2017 and 2018 I played 4307 games, averaging 178 games per month. That’s roughly 6 games per day, for a period of two years! Later came a huge drop to an absolute 0 games in most of 2019.

Games Played per Hour Distribution: This graph is a lot more balanced with high bars across all waking hours. There is a slight tendency towards afternoon hours, with spikes up to 7% of the games played. Obviously there a very few games played between 01:00 and 06:00 as these are sleep times in general.

# Records

With no specific questions to guide me, I started to investigate my overall record playing chess. The data from a broader scope will be the control group for every question I will be making further in this research.

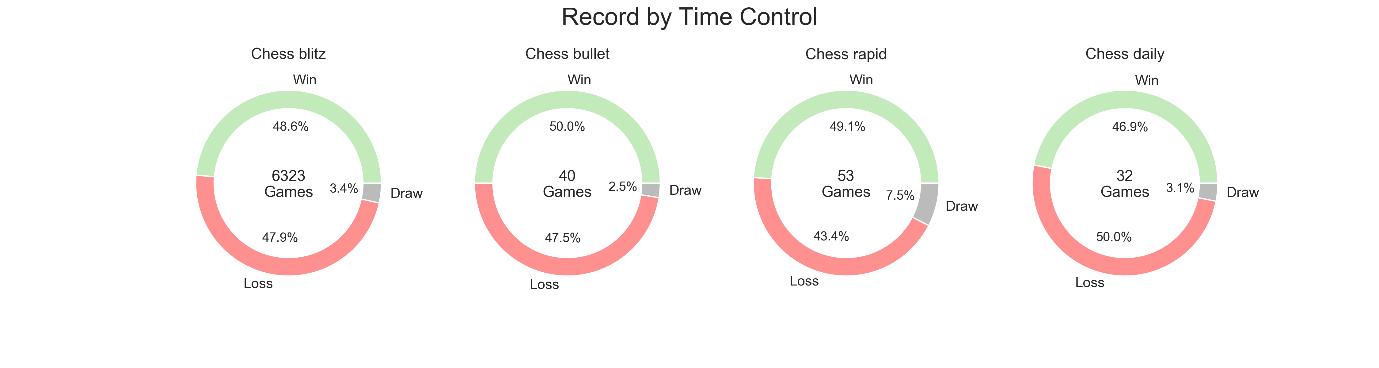
Before showing the next graph, I'd like to explain about *time control*. In chess, both players start with the same time. During the game, the clock will tick only for the player who has to play the next move. The different time control categories are basically the amount of time given for a player. Obviously, the time control impact over the game is huge: while long (classical) games are characterized by complex calculations and heavy thinking, short games will test the players' intuition and feel.

The **Time Controls** are defined as follows:

* **Bullet Games** - games with under 3 minutes per player.
* **Blitz Games** – games with 3 to 14 minutes per player.
* **Rapid games** – games with over 15 minutes per player.
* The **Daily chess** that appears in the graph is a special time control of Chess.com, that gives every player a couple of days per move. A user can manage as many games as he wishes in parallel.

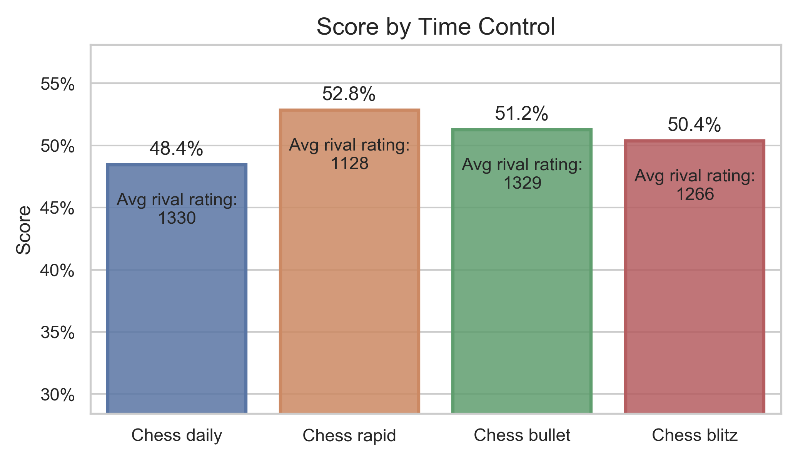
Note that these definitions are the chess.com definitions, while the standard chess league has different definitions, but the main idea is preserved.

Saying this, let's observe this graph:



We can notice a couple of things:

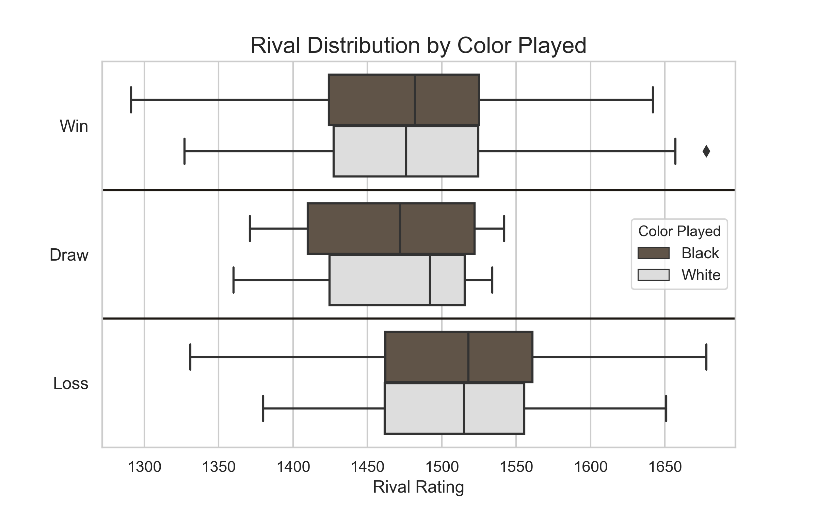
1. Most of my games were blitz, i.e. with 5 or 10 minutes of time control.
2. I have better results playing bullet with 50% wins, but the game sample is not big enough for a definite claim.
3. Putting Daily Chess aside, as it has very different rules, there is a tendency to have fewer draws as time becomes more limited. That's a very interesting, yet a very natural consequence: games with less time tend to end with a defeat because one of the players is out of time and loses automatically, or because of the limited time to calculate the next move, there's a bigger tendency to make crucial mistakes (blunders) that may even finish the game immediately.



**Score** is a metric used in chess tournaments. A chess tournament follows pretty straightforward rules and resemble a soccer league: every player will play every other player, and will accumulate points by winning or drawing: a full point for a win and half a point for a draw. A loss will grant no points. In order to normalize the data for comparison, I looked at all my past games as one big tournament, and generated a rate that is the amount of points I won by the points I could have won.

In a dynamic rating system that gets updated after each game, the score will always tend to 50% because when a player wins a game, its rating will rise and he will face higher-rated players to whom he'll more probably lose. As a result, we have to put the score into the context of my rivals, as it appears on the bars.

We can see that I have the best score playing rapid chess, but my opponents were significantly lower-rated players. On the contrary, we can observe that playing bullet chess I had a better score against stronger players. Perhaps I should be playing more bullet? Interesting indeed.

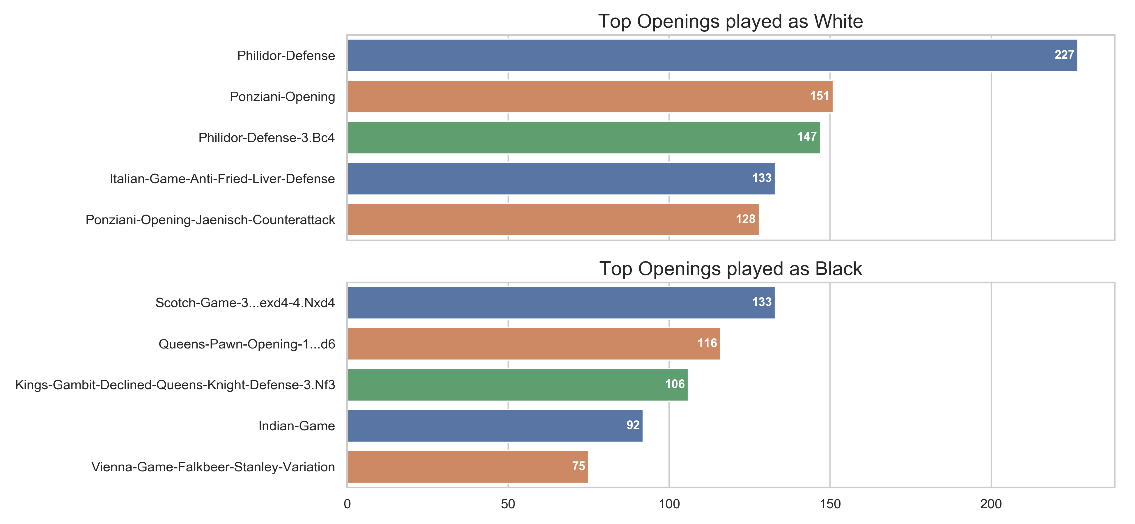


Here's another way to look at my overall record. This time, the data is split into six groups – color played and whether I won the game or not. Plotting a box plot gives an intuition about the expected result of the match given the opponent's rating.   
The first conclusion that can be extracted from this plot is that obviously the opponents to which I lose are better-rated than the opponents I defeat – losing to a median of 1517, and winning to a median of 1479. Another interesting observation is that I draw with weaker players than I win. The conclusion can be that I don't take advantage of playing weaker players and struggle with finishing games decisively.

# Opening Research

After I studied the overall picture that will be my control group – I put as a research question my openings' effectiveness. The game of chess is studied for over 1500 years, and while observing all the moves is practically impossible, a lot of chess experts put their efforts in understanding the first moves white or black should play to gain advantage early on in the game. Usually white openings will be called *attacks* or simply *openings*, while black openings will be called *defenses*. Different continuations of a specific opening will be called *variations* or *lines*. There are openings that include sacrificing a piece to gain strategic advantage, and those are called *gambits*, for either black or white.

Let's dive right in. These are my top openings played:



Warning! Chess terms ahead. If you feel you don't understand the next paragraph feel free to continue on reading.

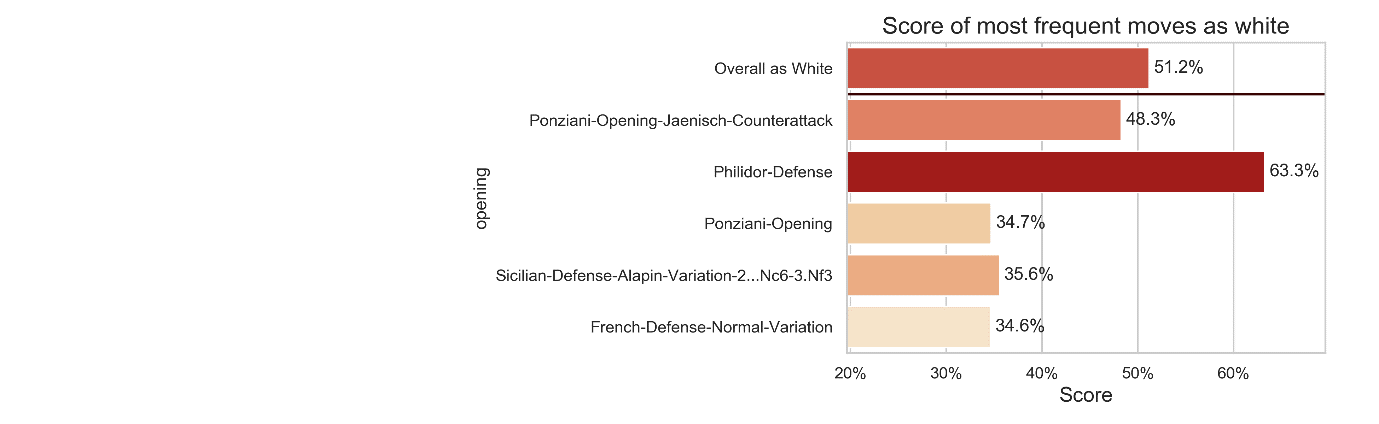
As white, the 4 out of the 5 most frequent openings are [*Philidor Defense*](https://www.chess.com/openings/philidor-defense) and [*Ponziani Opening*](https://www.chess.com/openings/Ponziani-Opening). Those are classic [e4](https://www.chess.com/openings/Kings-Pawn-Opening) openings. When I play 1.e4 I try to get to the *Ponziani Opening* getting a pawn on [c3](https://www.chess.com/openings/Ponziani-Opening) and later establishing a pawn on d4 with castle-king-side and great center control, but most responses I get from black are Philidor Defense (2…d6) that transpose my game into a lesser known variant. Another problematic opening is the [*Jaenisch Counterattack* to the Ponziani](https://www.chess.com/openings/Ponziani-Opening-Jaenisch-Counterattack), a move I fear black plays that usually ruins white structure and attack.

As black there is a more moderate distribution of the openings because as black you usually respond to white moves, so facing more opponents lead to a greater variety of openings. The first opening (*Scotch Game)* is favorable by white, and 3…exd4 is my favorite response as black, leading to a 4.Nxd4 Nge7 which is very uncommon and confuses white many times.

  
the position after move 4 as described above

My favorite response to [1.d4](https://www.chess.com/openings/Queens-Pawn-Opening) is 1…d6 that usually leads to a [Modern defense](https://www.chess.com/openings/Modern-Defense). In this opening black establishes a good defense and gets to harass white development and eventually get an advantage.

After the chess analysis, let's confront my openings with the ultimate challenge – Data Analysis. Let's start by testing My top white openings. The following graph's scope is all my games against >1450 rated players:



I got a score of 51.2% from all the games I played as white. This score will be our control group to whom I'll compare the openings. We can see that 4 out of my 5 most frequent openings I score lower than my average. This means that my favorite opening is just not good enough and loses two thirds of the times approximately.   
I'm an [*e4*](https://www.chess.com/openings/Kings-Pawn-Opening) player. Looking at the graph, it's clear to see that I handle well the Philidor defense, but my line against *2… Nc6* which I respond to with *3. c3* leading for the Ponziani lines is just not good enough.

  
Ponziani Opening

Let's look at my black openings. Like white, this graph takes into consideration games against >1450 rated opponents.



In this graph it looks that most of my frequent openings as black work in my favor – which is logical: As I said above, black usually responds to white's moves. If I as black managed to reach an opening I feel comfortable with, I will be able to score better. Most games I lost as black were the ones were white managed to reach a position me as black didn’t know well and thus committed more mistakes.  
One big exception to point out is the huge difference between the two [*Italian-game*](https://www.chess.com/openings/Italian-Game) lines (4th and 5th in the graph). These two openings are based on the Italian-game, i.e. the first three moves are identical, yet they lead to significantly different results for me. As black, I'm very happy to get the [Knight-attack](https://www.chess.com/openings/Italian-Game-Knight-Attack) as it is (even by looking at the chess.com database) weaker than the [modern bishop](https://www.chess.com/openings/Italian-Game-Two-Knights-Modern-Bishops-Opening) (*4.d3*). according to the chess.com database, the *4.d3* move is the best for white and leaves black with the lowest chances of winning. I have to dive deeper into the Italian game – 4.d3 variation to a better play in this opening.

    
Italian game–Two knights defense Italian game – Knight Attack Italian game – modern bishop opening

## Conclusion

There is a huge amount of information in this field and with it, a lot of hidden observations and conclusions. The goal of this article was to not only get a glimpse of what chess analysis is all about, but to show different approaches to the chess studies. This article was a live example of how one can manage to improve his game by not only studying theory but looking broadly at his own games and learning from mistakes.

Thank you for reading and studying with me this wonderful topic. I've uploaded all the code behind this article – from the constants declared through data fetching and into visualization – it's all open for you to watch and learn.

Please! Let's stay in touch:

Yair Vogel.  
Email: yairvogel@gmail.com   
[LinkedIn](https://www.linkedin.com/in/yair-vogel-517446156/) and [Github](https://github.com/yairvogel)