Proposed Analytics Report

Rishi Garg

2022-10-24

Abstract

This report pertains to the project I'm working on as a part of one of my master's courses, "Visualization". This is an initial proposal report laying out the introduction of the dataset being used, various possible plots, graphs, comparisons, and inferences from the data. This shall be used as a brief about the project until the final project and the final report is ready.

Introduction

This project is about developing an analytics dashboard for a specific dataset. I'll use R and R-shiny throughout this project and the dashboard shall contain various graphs and plots and possibly an interactive graph as well. I'll use R-Studio to deploy the app on the web.

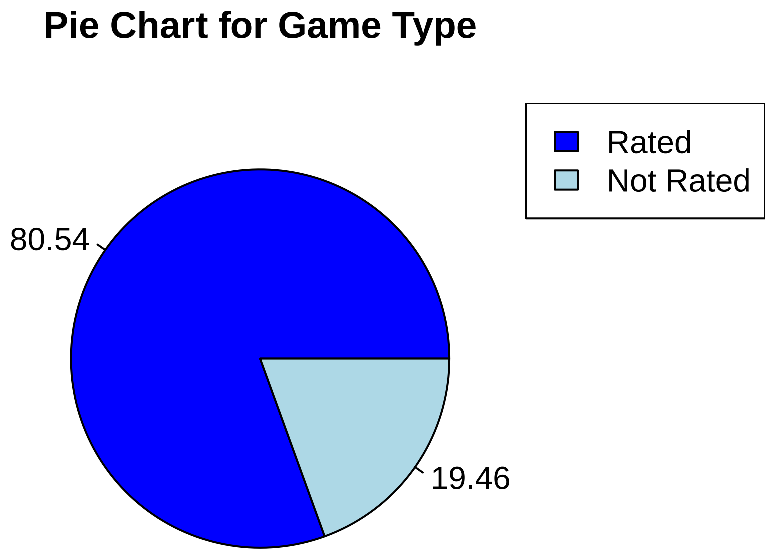
Dataset Description

The dataset I'm using for this project is about **Online Chess Games**. It contains data about game moves, victor, players' ratings, opening details, the game is rated or not and much more from more than 20,000 games played online on Lichess.

Our objective shall be to verify, using various plots, graphs, comparisons, and inferences from the data, whether the games and ratings on Lichess are rational or not.

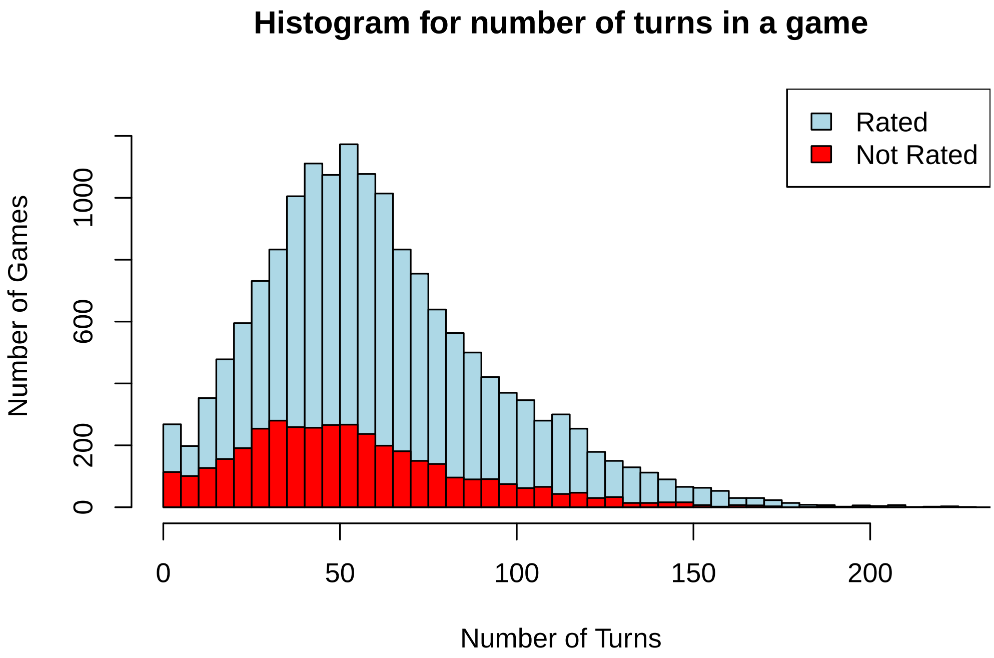
Analysis

We start by analyzing and comparing the number of rated and non-rated games played on lichess as per the dataset used. Consider the below pie chart for the same:

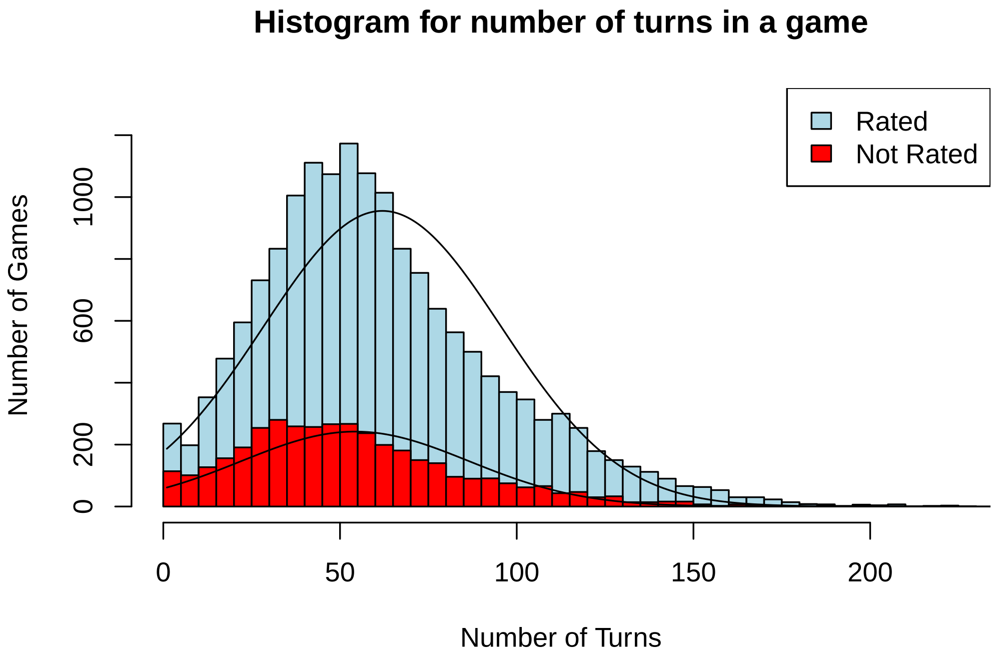


It can be seen that the total proportion of rated games is 80.54% which is much larger than that of non rated games (accounting to just 19.46%). And, thus, it can be concluded that the number of rated games played is much larger than the non rated games played.

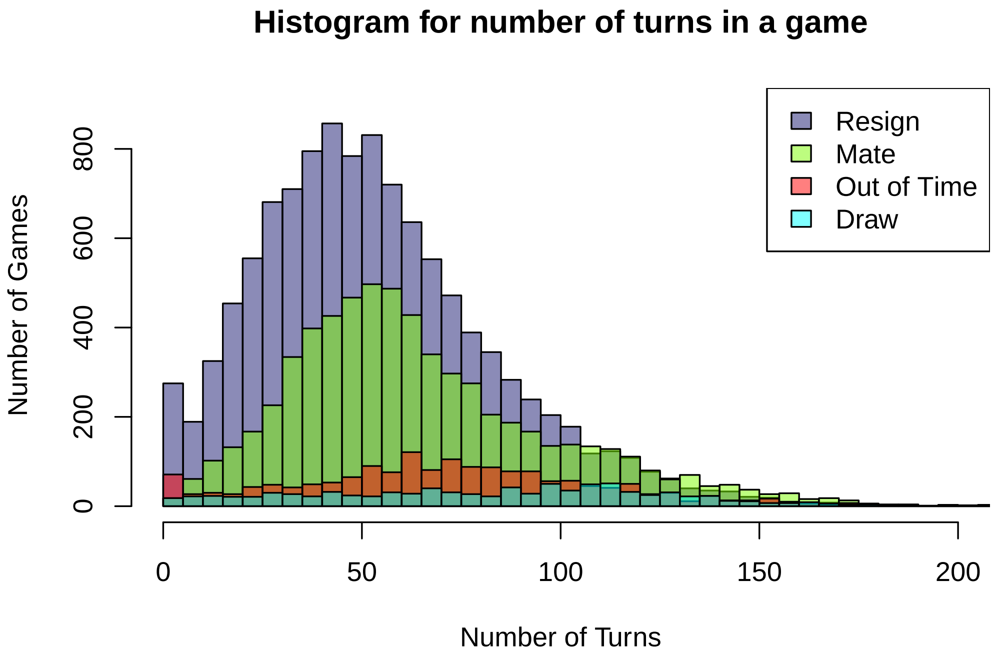
From the given dataset, considering the columns `turns` and `rated`, below is the histogram representing the number of turns it took in a game to finish along with the comparison for both `rated` as well as `non rated` games.



It can be visualized that the histograms/distributions seem like a normal distribution. To verify the same, I also draw a normal curve on the same graph to verify whether the two follows a normal distribution or not. The graph can be visualized as follows:

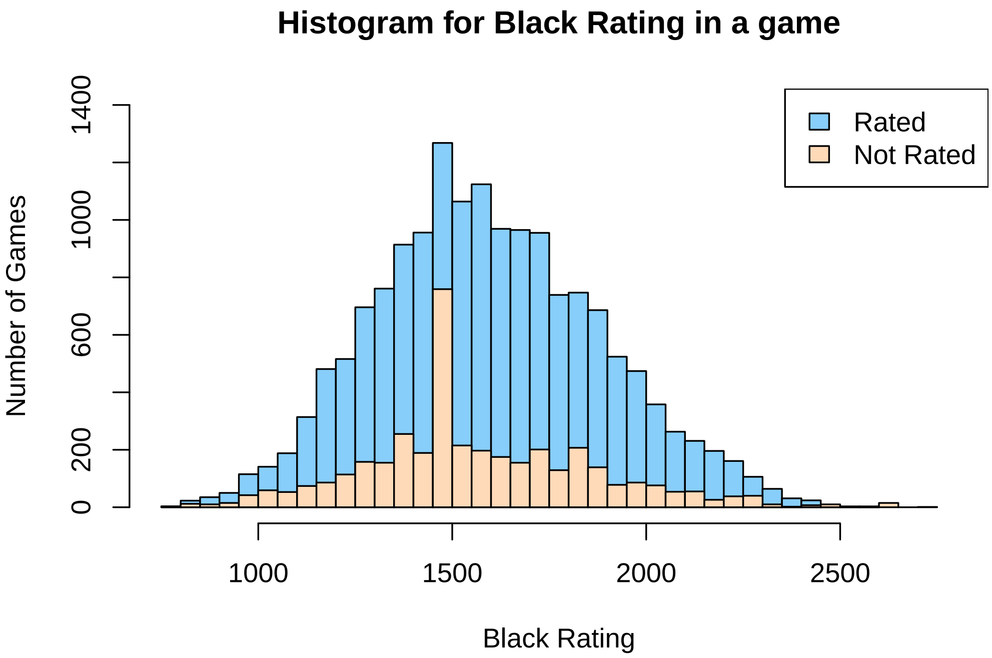


Below we have the histogram representing the number of turns it took in a game to reach the final result factored by the result status being *Resign*, *Mate*, *Out of Time*, and *Draw*.



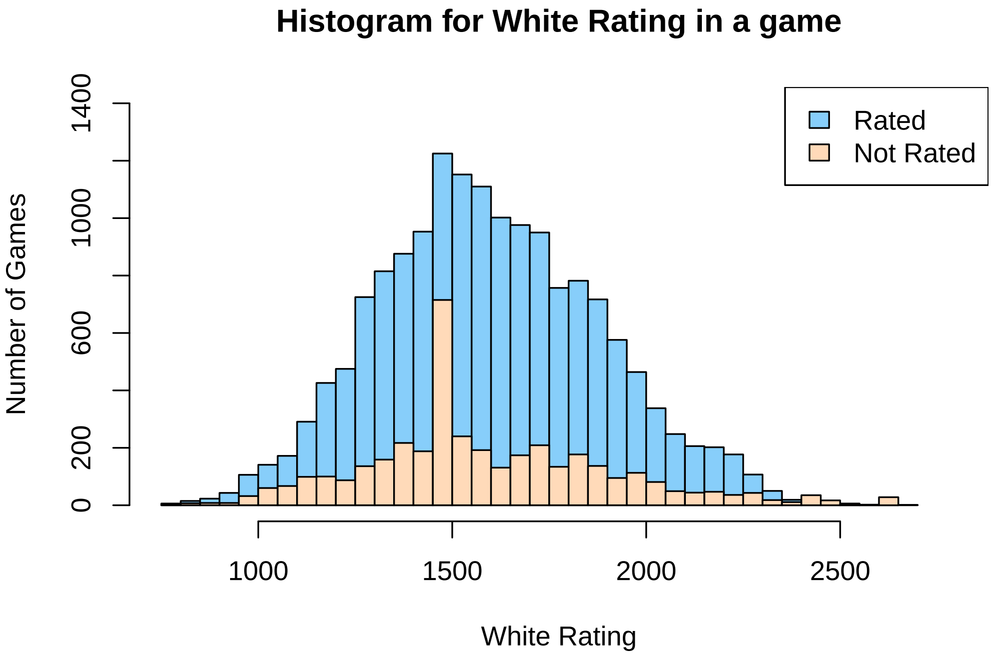
It can be seen from the above plot that for a fixed number of turns, the number of games in which one of the players resigned is much larger than those where the game resulted in a win of one of the players.

Consider the below histogram for the Black Ratings in all the games played. The plot is factored by the type of the game, it being *Rated* or *Non Rated*.



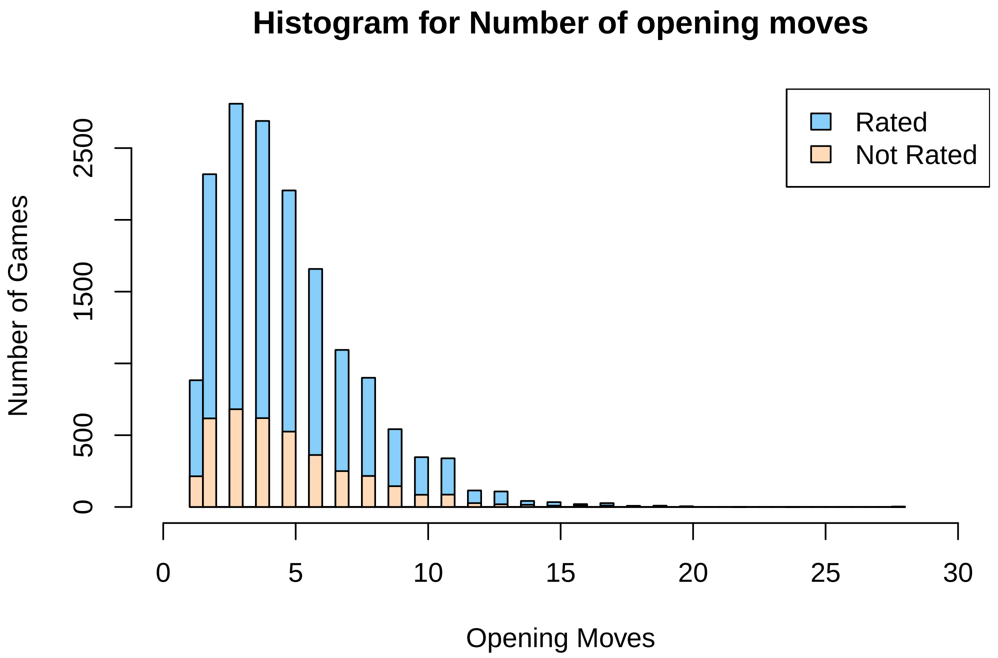
It can be seen from the above graph that the number of rated games played are more than the non rated games for almost every Black rating. Also that the number of games played are the maximum when the black rating is between 1500 and 1600 and it decreases as the rating goes downward or upward.

Consider the below histogram representing the White ratings factored by the type of the game.



The same analysis as the one for the Black ratings can be done for the White ratings. The two plots seem to be similar for making the simple comparisons.

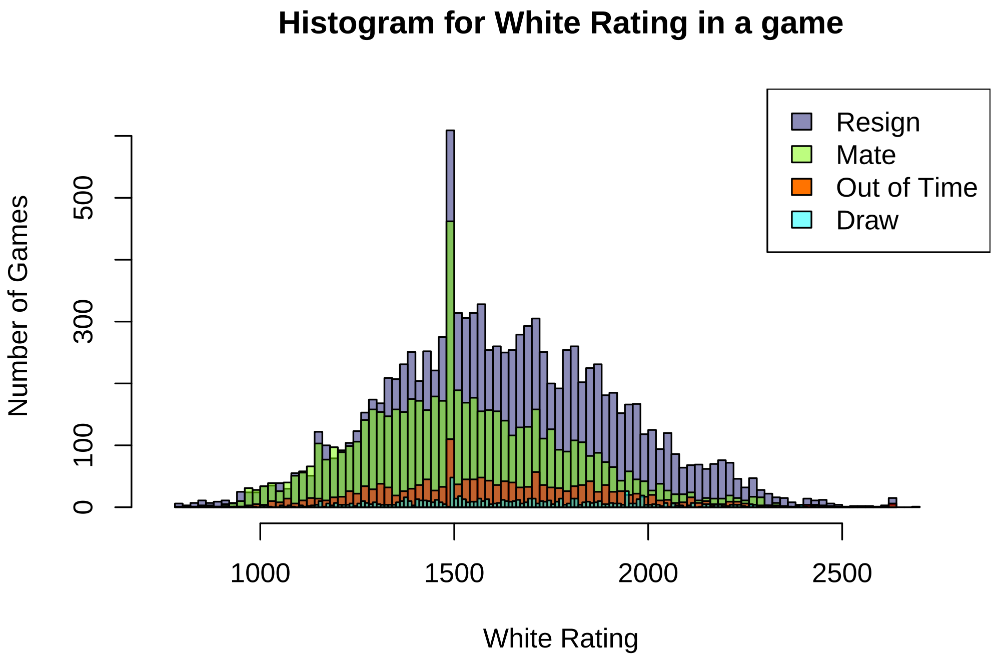
For any chess game, it is most important to use a sound opening strategy to increase the winning chances. Each opening strategy has a different number of moves. Consider the below histogram representing the number of opening moves in a game factored by the type of the game i.e., *Rated* or *Not Rated*.



From the above graph, it can be seen that the maximum number of games used opening strategy with 3 opening moves irrespective of the type of the game being *Rated* or *Not Rated* followed by 4 and 2 opening moves respectively.

It can also be seen from the graph that there are almost 0 games each that used more than 20 opening moves.

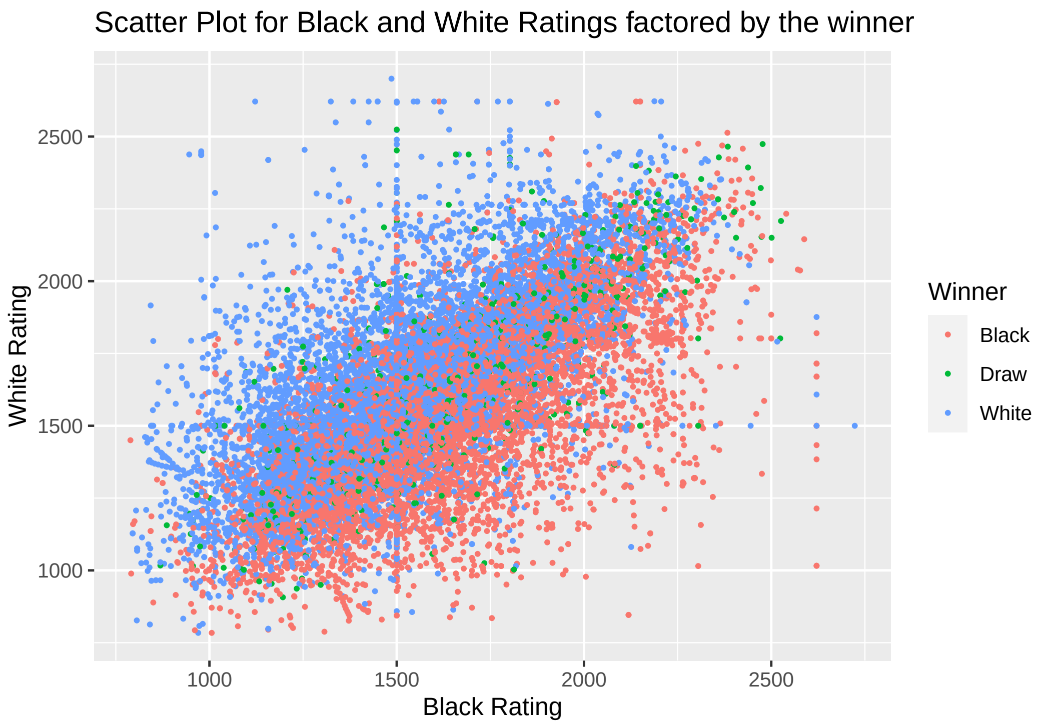
After discussing various plots and comparisons, let's also see the effect of white ratings on the final result status. Here is the histogram representing the White Ratings factored by the Victory Status:



It can be interpreted from the above graph that the maximum number of games that resulted in a "Resign" were those where White Rating was more than 1500 and the majority of those resulted in *Mate* had White Ratings less than or around 1500.

It can also be seen that the number of games with final status as *Out of Time* or *Draw* are much less than those with *Resign* or *Mate* as their final status. However, it is not very clear from the above graph to define a specific range of values of White Ratings for the *Out of Time* and *Draw* status as they are not concentrated in any specific small range and rather distributed for all values of White Ratings.

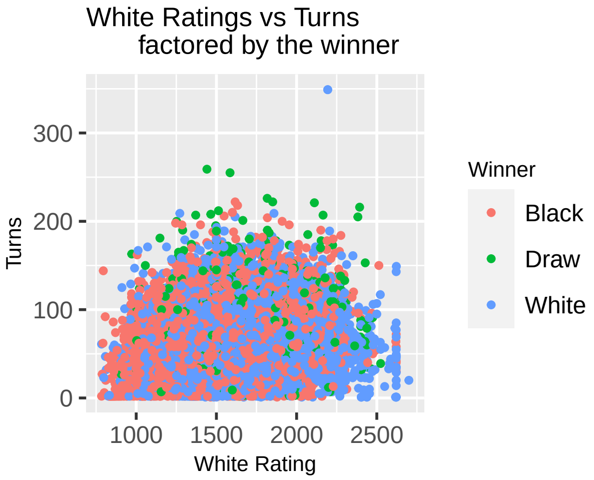
Consider the below scatter plot for the Black Ratings against the White Ratings. The plot is factored by the winner in the game.

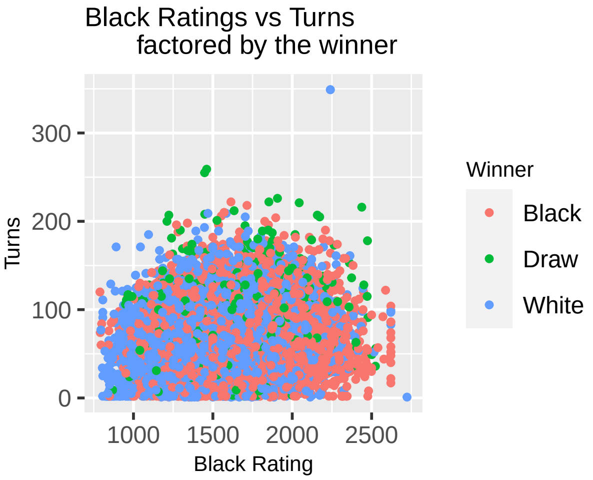


From the above scatter plot, it can be seen that the black and white ratings have a strongly positive linear correlation which implies that the matches/games played on lichess are rational and that the players of similar skills are matched for a game and thus, it can be concluded that the players with huge difference in their ratings are usually not matched for a game.

In the above plot, it can be seen that the majority of points below the line $y = x$ are red indicating that the black player won while the majority of points above the $y = x$ line are blue indicating that the white player won. Thus, it can be concluded that the player with higher rating than the opponent wins in the majority of the cases and that, the games and ratings are rational.

Now, we have the following two scatter plots representing the number of turns it took in a game to reach its final result ('win' or 'draw') against the Black Ratings and against the White Ratings. The $x-$axis represents the Ratings while the $y-$axis represents the number of turns it took in the game. The plots are factored by the winner of the game.


White Ratings vs Turns
factored by the winner


Black Ratings vs Turns
factored by the winner

From the above two plots, it can be easily interpreted that the majority of the games that resulted in a *Draw* took more 100 turns to reach the final status.

Also, there are only 3 out of 20,000 games that took more than 250 turns to reach the final status and only 1 game out of those that took more than 300 turns.

From the two plots, it can be seen that as the rating goes beyond 1500, the respective player wins more games than the other and for rating of a player below 1250, the other player wins more often which indicates that the games are indeed rational.

Summary/Conclusion

From all the plots and graphs discussed above, it can be concluded that the matching algorithm on lichess that matches the two players is good enough to ensure that the players matched are of equivalent skills and that no player has an undue advantage.

Also, that the number of rated games played is more than the number of non-rated games.

It can also be summarized that the number of rated games are more than non rated games when filtered by almost any parameter. When compared the Black ratings, White Ratings, Number of opening moves, we saw that the rated games are much larger than the non rated games. And, thus, it would not be wrong to conclude that the players prefer playing rated games than non rated games irrespective of their ratings or any other game mode/factor.

The final R dashboard shall have all these graphs along with more graphs and plots to make detailed analysis and comparisons among various variables in the dataset. It shall also have an interactive graph.