Online Chess Games Analysis

Marcus Kamen

2023-04-11

In this paper, I will be analyzing the "Online Chess Games" data set from Kaggle.com. The link to this data set can be found at https://www.kaggle.com/datasets/mysarahmadbhat/online-chess-games?resource=download. This data set analyzes thousands of online chess matches, tracking moves, openings, ratings, as well as who won and how. In this analysis, I will explore what distinguishes two players of different ratings other than just their rating. Essentially, what can be used to predict the rating of a player? This information may be helpful to someone who wants to improve at chess. If I can determine what a lower rated player lacks, then I can find out what they can improve on. As a start, I hypothesize that lower rated players more often lose early in the game, initiate and lose with unusual openings, and win with white, all more than higher rated players. For my analysis, I'm going to first graph the average rating against other variables and see the trends. I will also explore any other relationships having to do with player behavior that may seem relevant to the analysis. Then, I am going to try to use regression to determine which variables have a significant effect on rating. From there, I can make conclusions, using knowledge of how the game of chess works, to determine where lower rated players are lacking compared to higher rated players.

To start my analysis, I will filter the data. Game_id, white_id, and black_id will not be very helpful in the analysis since I do not care about the username of the players. I will only take into account games that were rated. Also, games that lasted less than four moves will be filtered out. If a game lasts one move and then a player resigns, that is not very good information to use in an analysis of a player's actual skill. I will also filter out the information about which moves were played. Although this information could be interesting, it is beyond the scope of this analysis (and possibly my knowledge of chess) to determine whether any specific moves were good or bad (As a side note, interestingly, there were only 18920 unique games with about 20100 data points, so about 1200 games were repeated by different players). I am going to add the variable average_rating because it will help keep my rating analysis in terms of one variable.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)

chess <- read.csv('chess_games.csv')
chess <- chess %>%
  filter(turns >= 4) %>%
  select(-game_id) %>%
```

```
  select(-white_id) %>%
  select(-black_id) %>%
  select(-moves) %>%
  mutate(average_rating = (white_rating + black_rating)/2)

head(chess)
```

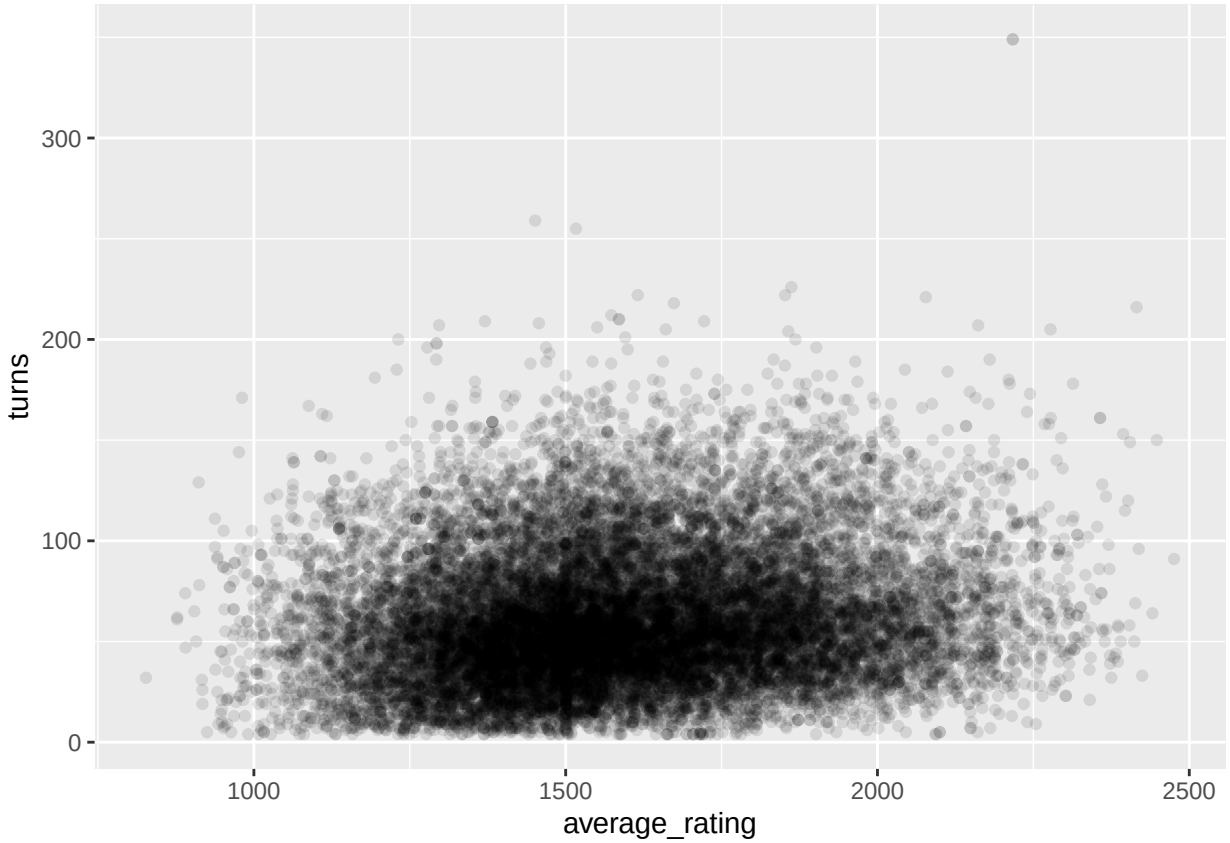
```
##   rated turns victory_status winner time_increment white_rating black_rating
## 1 FALSE    13    Out of Time  White           15+2         1500         1191
## 2  TRUE    16         Resign  Black           5+10         1322         1261
## 3  TRUE    61           Mate  White           5+10         1496         1500
## 4  TRUE    61           Mate  White           20+0         1439         1454
## 5  TRUE    95           Mate  White           30+3         1523         1469
## 6 FALSE     5           Draw   Draw           10+0         1250         1002
##   opening_code opening_moves                      opening_fullname
## 1          D10             5       Slav Defense: Exchange Variation
## 2          B00             4 Nimzowitsch Defense: Kennedy Variation
## 3          C20             3  King's Pawn Game: Leonardis Variation
## 4          D02             3 Queen's Pawn Game: Zukertort Variation
## 5          C41             5                      Philidor Defense
## 6          B27             4   Sicilian Defense: Mongoose Variation
##     opening_shortname opening_response   opening_variation average_rating
## 1        Slav Defense                  Exchange Variation         1345.5
## 2 Nimzowitsch Defense                   Kennedy Variation         1291.5
## 3    King's Pawn Game                 Leonardis Variation         1498.0
## 4   Queen's Pawn Game                 Zukertort Variation         1446.5
## 5    Philidor Defense                                             1496.0
## 6    Sicilian Defense                  Mongoose Variation         1126.0
```

Above is a sample of a few data points.

Now, I will do a preliminary analysis of the data. If I want to analyze what differently rated players do, I should start by graphing average player rating against other variables that make sense to see if there is a significant relationship.

First, let's plot turns against average rating below.

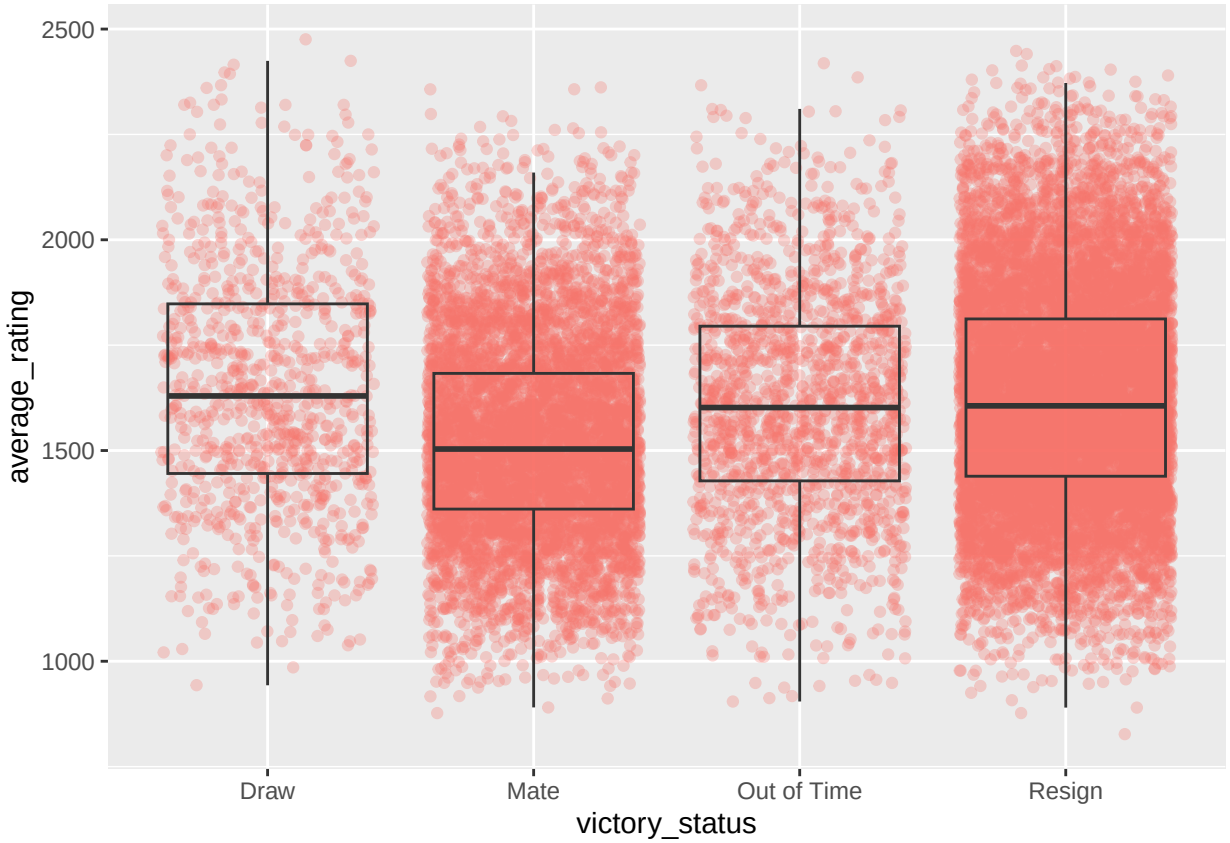
```
ggplot(chess, aes(average_rating, turns)) + geom_point(alpha = 0.1)
```



This data suggests a slight increase in game length for higher rated players, but it does not seem to be too significant. This may be because higher rated players are less likely to make huge, immediately game losing mistakes. Instead, they will make smaller mistakes that will build up over the game and eventually cause them to lose much later. It should be noted that this relationship is probably stronger than it seems in this graph because, as shown later, higher rated players are more likely to resign, so their games end earlier than they could more often.

Let's plot average rating against victory status below.

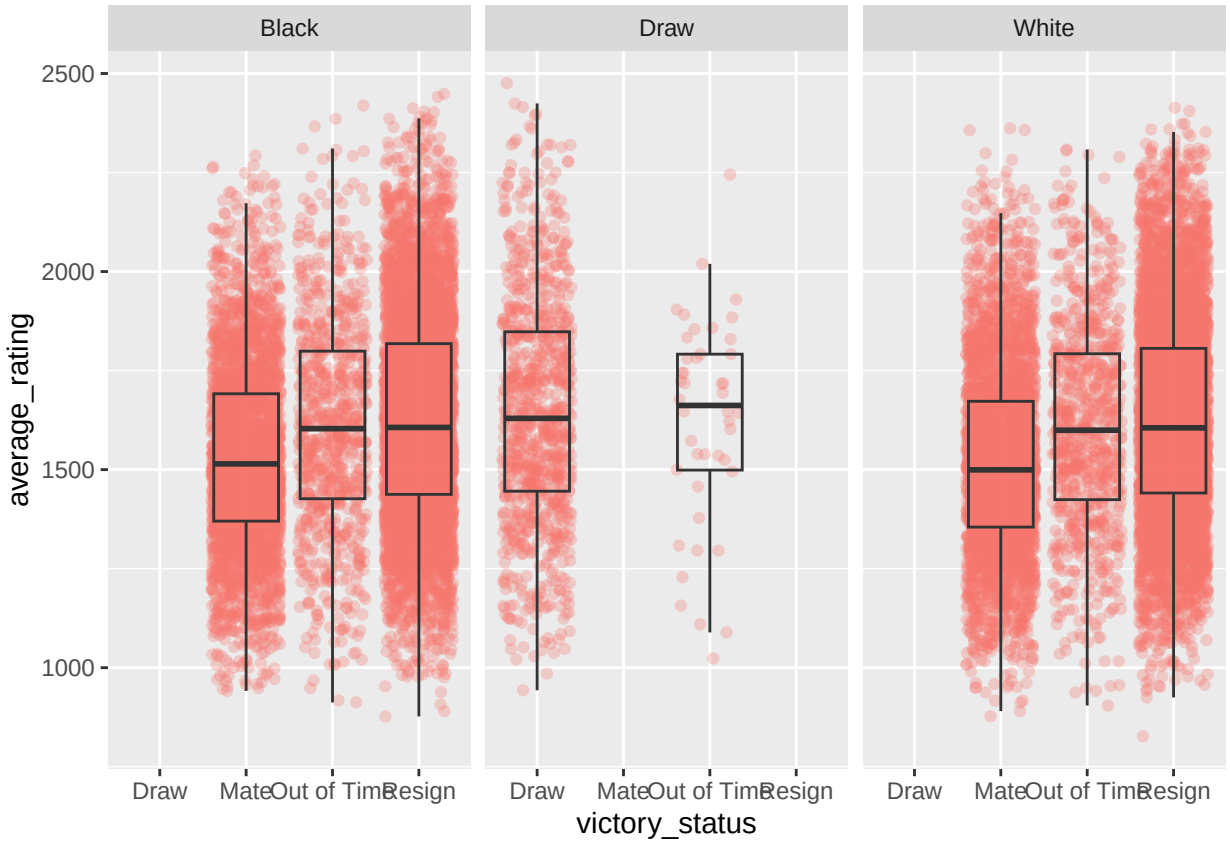
```
ggplot(chess, aes(victory_status, average_rating)) + geom_jitter(alpha=.3, aes(color = 'maroon')) +
  geom_boxplot(alpha=0) + guides(color = 'none')
```



This data suggests that lower rated players are less likely to resign. From my knowledge of chess, this could mean that lower rated players are less likely to give up early. But also, this could mean that lower rated players are less likely to see that they are in a losing position, and so continue playing a game even though they are about to be checkmated or are about to lose significant material. Also, lower rated players are not as likely to draw. This may be because a lower rated player is more likely to make a big mistake that immediately ends the game.

Let's do the same analysis, but wrap by who won the game. This can be seen below.

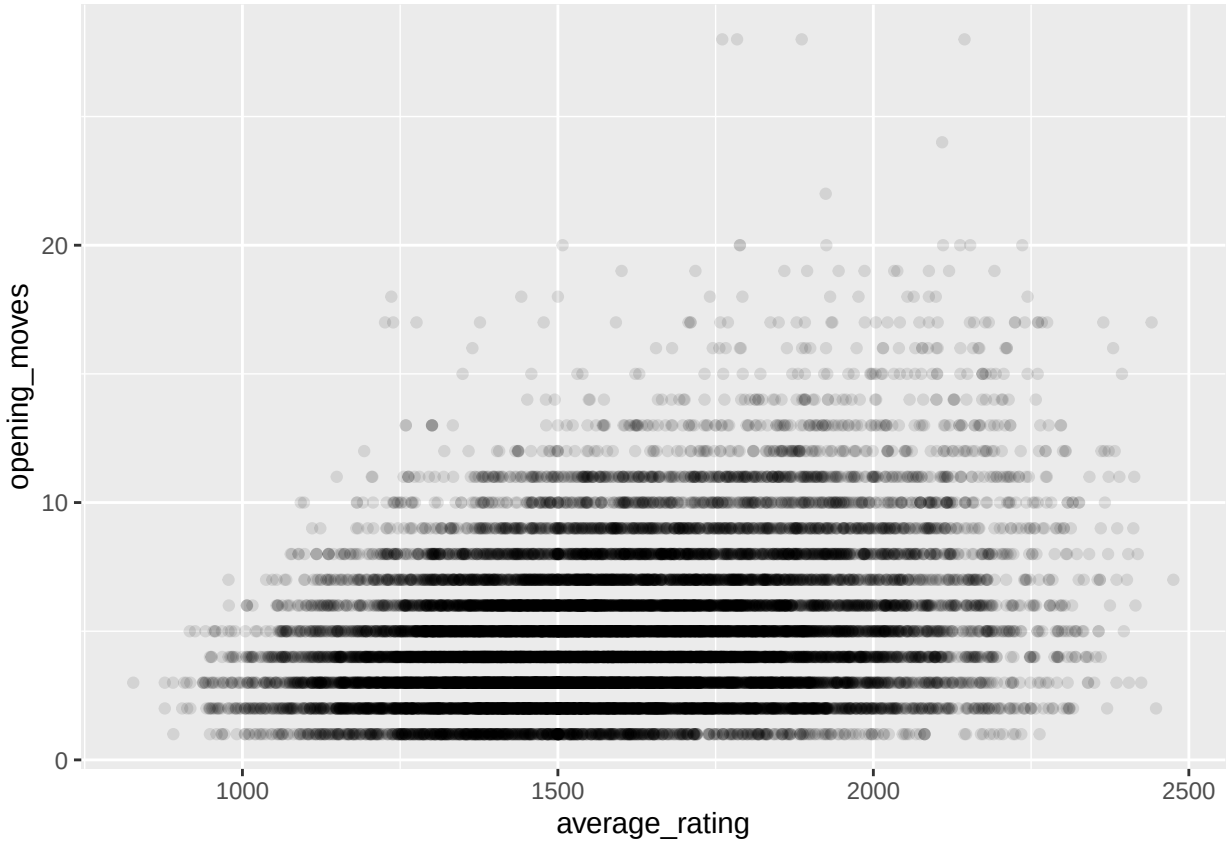
```
ggplot(chess, aes(victory_status, average_rating)) + geom_jitter(alpha=.3, aes(color = 'maroon')) +
  geom_boxplot(alpha=0) + guides(color = 'none') + facet_wrap(~winner)
```



As can be seen, there doesn't seem to be much of a difference in rating between players that win, resign, draw, or run out of time with white vs. black. It is known in chess that white has a slight advantage over black. From this data, it does not seem like this advantage is any more pronounced in higher or lower rated players.

Let's plot opening moves against average rating below.

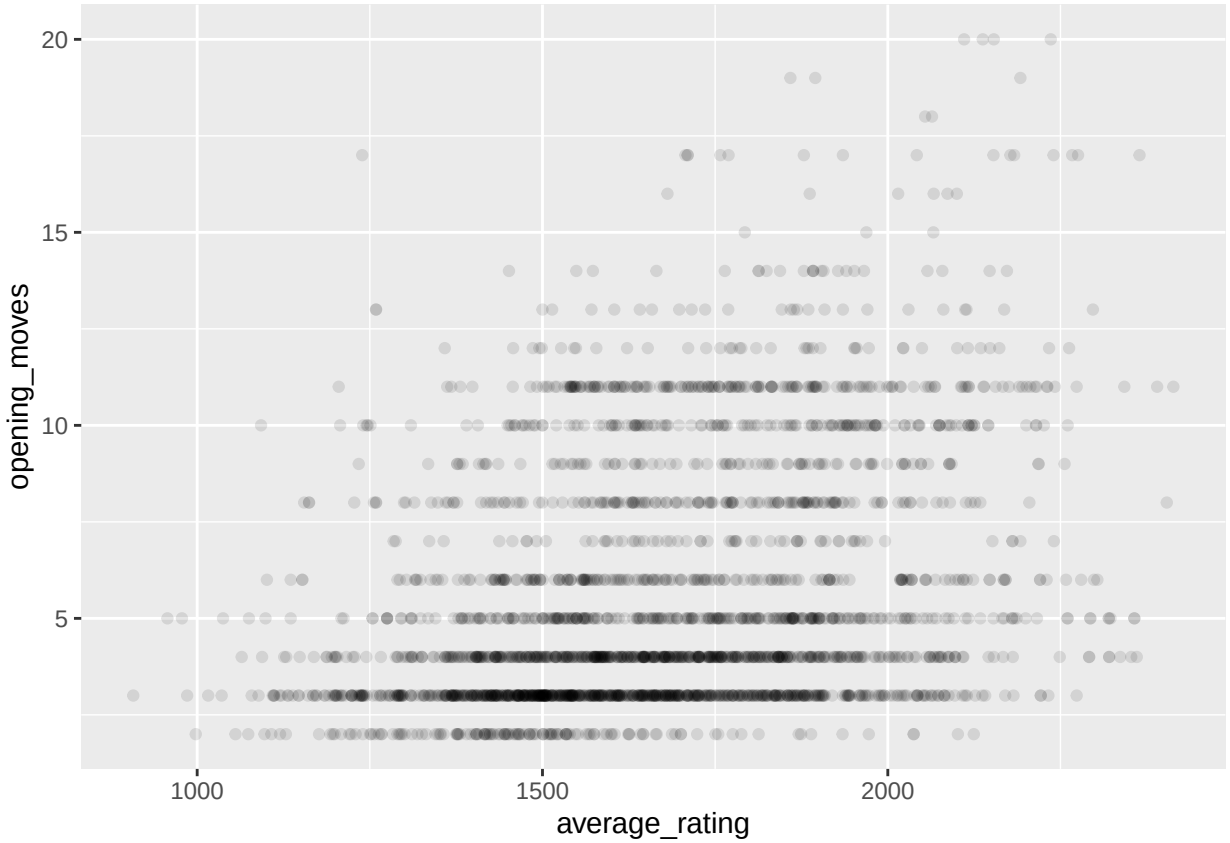
```
ggplot(chess, aes(average_rating, opening_moves)) + geom_point(alpha = 0.1)
```



This data seems to suggest that higher rated players tend to play more moves in the opening. This might suggest that they have memorized openings to use in a game, or, because they are better at finding the best move, they can find opening moves more easily.

Let's do the same analysis, but for the most common openings. Since there are many, many possible openings to choose from, I will stick to the Sicilian Defense, the most common opening in this data.

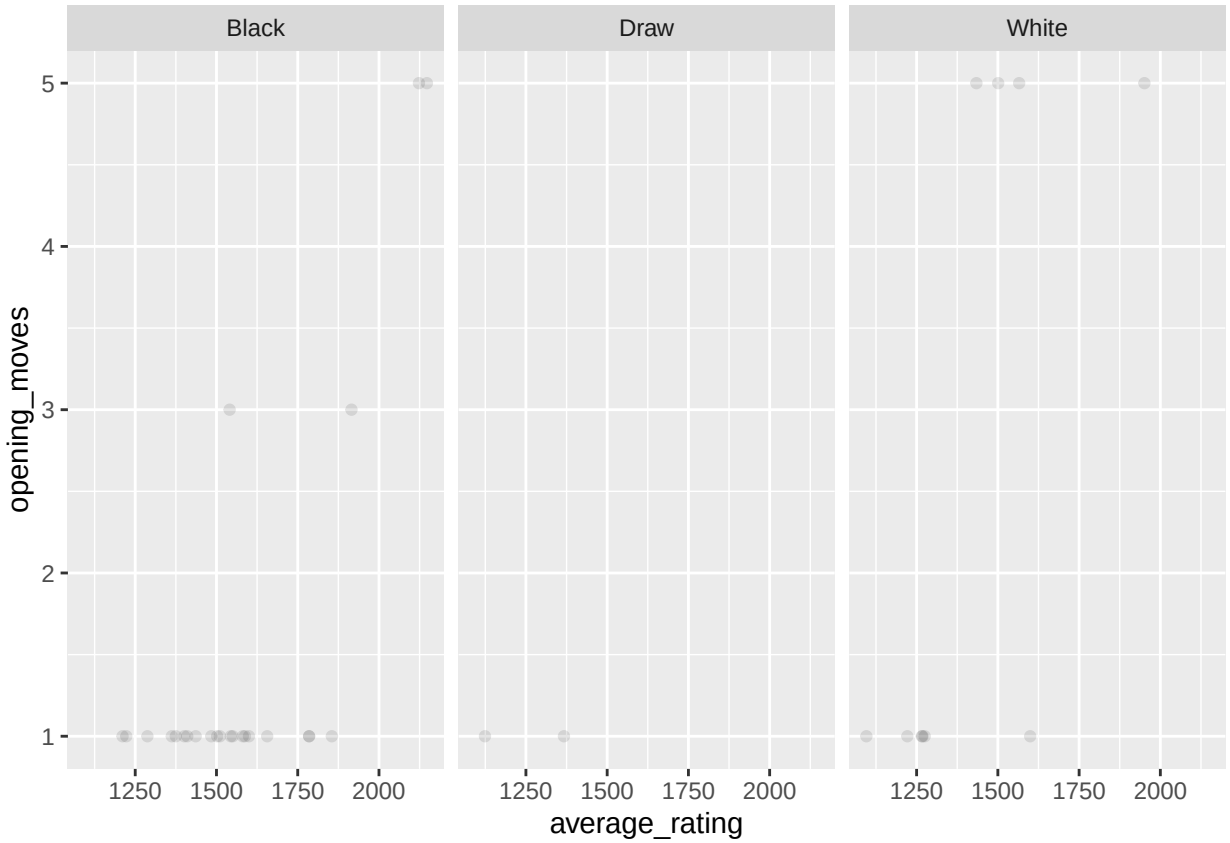
```
commonOpenings <- chess %>%
  filter(opening_shortname == "Sicilian Defense")

ggplot(commonOpenings, aes(average_rating, opening_moves)) + geom_point(alpha = 0.1)
```



From this analysis, it can again be seen that higher rated players spend more time playing defined opening moves. In fact, in the Sicilian Defense, it seems that higher rated players play more moves in opening comparatively to an even larger extent than the average opening. This may be because, if a lower rated player enters the Sicilian, it may just be because they know it exists. However, a higher rated player will know and play the many defined opening moves that the Sicilian has.

Let's do the same analysis, but this time, for uncommon openings. Let's also wrap by who won. Again, since there are many openings to choose from, I'm going to stick to an opening that is objectively bad if computer analysis is used, but players may play anyway. For this, I will use the Grob Opening.

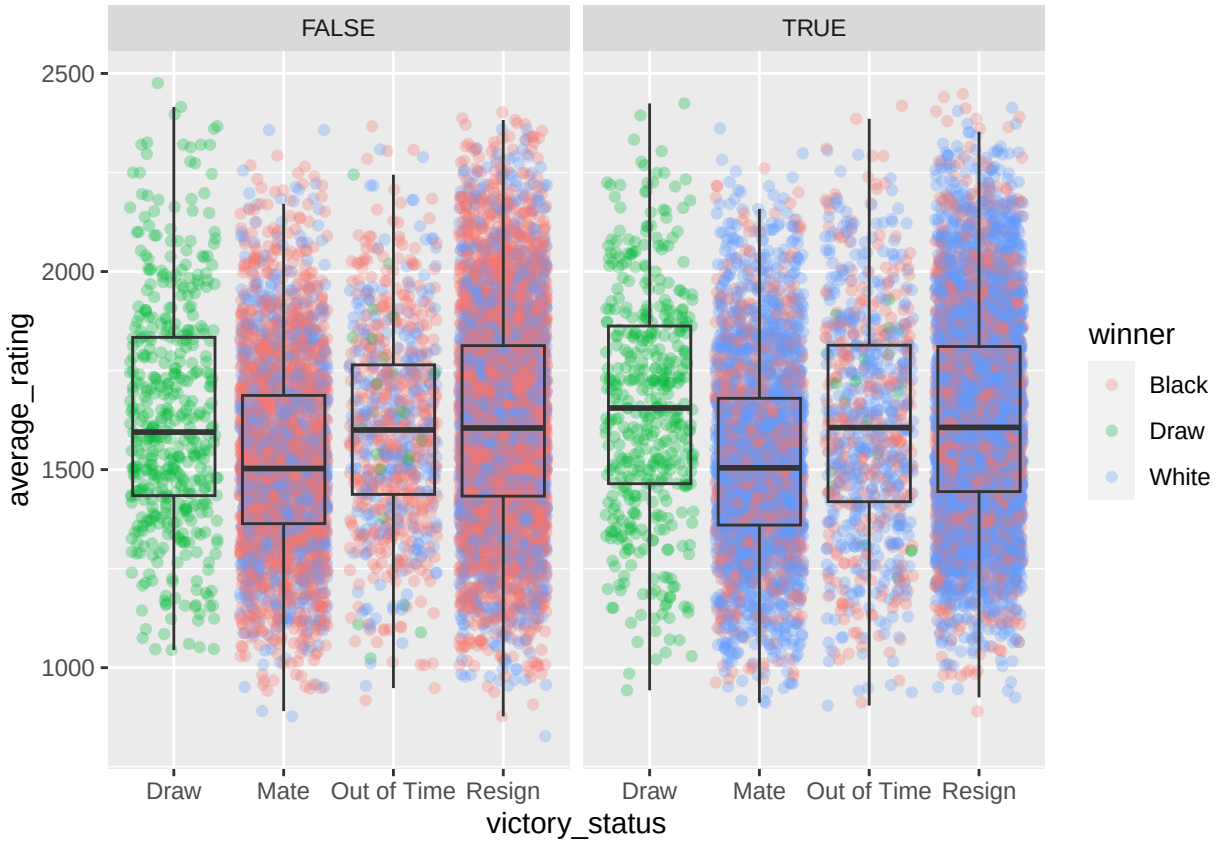
```
uncommonOpenings <- chess %>%
  filter(opening_shortname == "Grob Opening")

ggplot(uncommonOpenings, aes(average_rating, opening_moves)) + geom_point(alpha = 0.1) +
  facet_wrap(~winner)
```



In can be seen from this data that lower rated players are more likely to enter this opening, less likely to play the (albeit little) theory of the opening, and more likely to win with black than with white. Because this opening is instigated by white, this means that playing into this opening can be an extremely risky move for that player if they are too lowly rated.

Let's plot average rating against victory status, except let's also wrap by whether white was higher rating, coloring by who won.

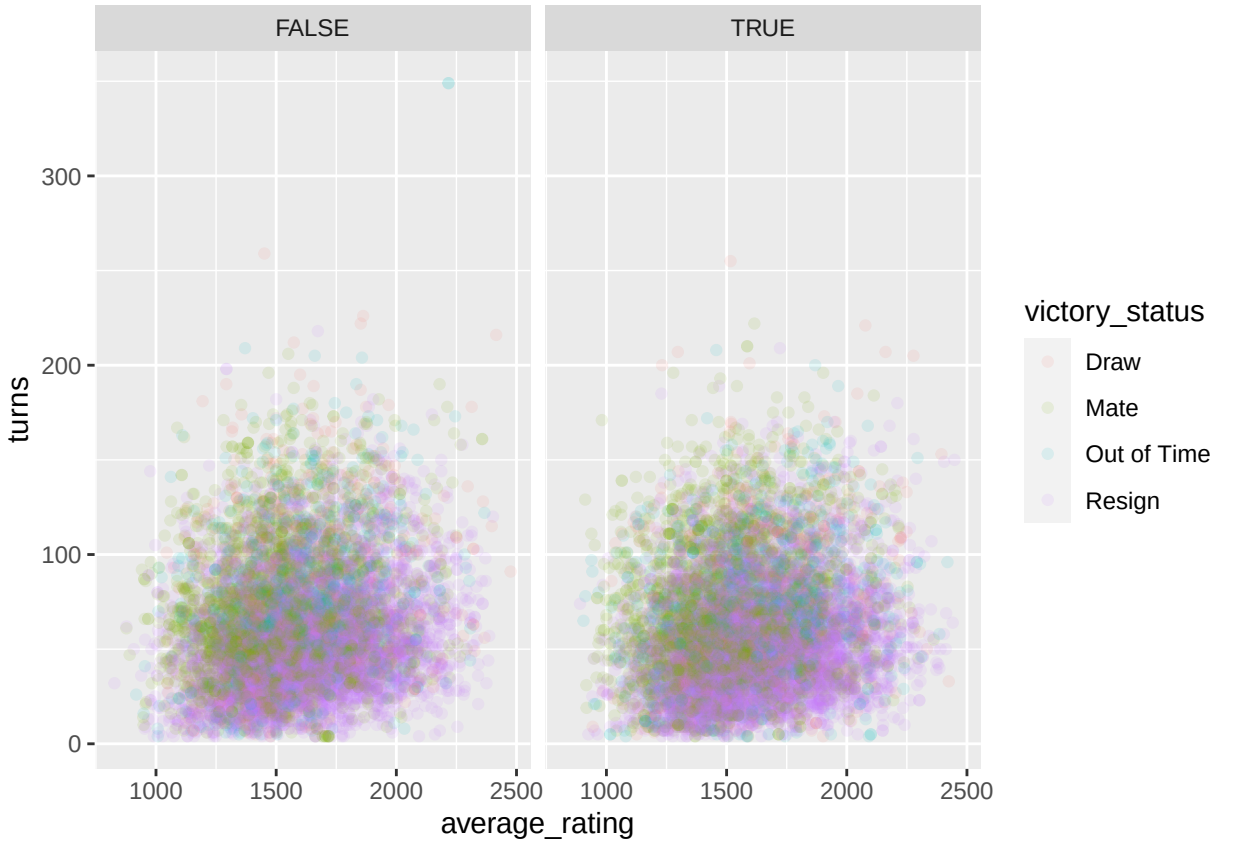
```
whiteHigher <- chess %>%
  mutate(whiteHigherRated = (white_rating > average_rating))

ggplot(whiteHigher, aes(victory_status, average_rating)) +
  geom_jitter(alpha=.3, aes(color = winner)) + geom_boxplot(alpha=0) + facet_wrap(~whiteHigherRated)
```



Unsurprisingly, the player with the higher rating is much more likely to win. More interestingly, a draw with a higher rated white indicates a much higher average rating than a draw with a lower rated white. Honestly, I'm not too sure why this would be the case, but it is interesting to note nonetheless.

Let's plot turns against average rating again, but this time, wrap by whether white was higher rated, also coloring by victory status.

```
ggplot(whiteHigher, aes(average_rating, turns)) +
  geom_point(alpha = 0.1, aes(color = victory_status)) + facet_wrap(~whiteHigherRated)
```



As can be seen, resigns happen most often with higher rated players in shorter games. Which player is higher rated doesn't seem to have much of an effect on this. This may be because higher rated player resign more often anyway. Also, if a player makes a huge mistake early, they may just resign right away. But, if they make a small mistake late, they may drag the game out.

Now that we have investigated some interesting aspects of this data, let's try to create a linear regression model to predict player rating. We will use turns and opening moves as our variables since those are numerical values that can go into a mathematical formula. Below are the coefficients of the regression and the confidence intervals for these coefficients. We will attempt to use multiple variables with polynomial regression to start.

```
model <- lm(average_rating ~ poly(turns, 3) + poly(opening_moves, 3), data=chess)
sm <- summary(model)
print(sm$coefficients)
```

```
##                              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                1595.0118   1.756774 907.920743 0.000000e+00
## poly(turns, 3)1            5216.7071 247.134100  21.108811 7.928598e-98
## poly(turns, 3)2           -1543.3203 247.048547  -6.247033 4.268676e-10
## poly(turns, 3)3            1398.4742 247.343367   5.653979 1.589400e-08
## poly(opening_moves, 3)1 10482.0994 247.477549  42.355759 0.000000e+00
## poly(opening_moves, 3)2   534.5661 247.017855   2.164079 3.047023e-02
## poly(opening_moves, 3)3 -1267.3147 247.030425  -5.130197 2.921761e-07
```

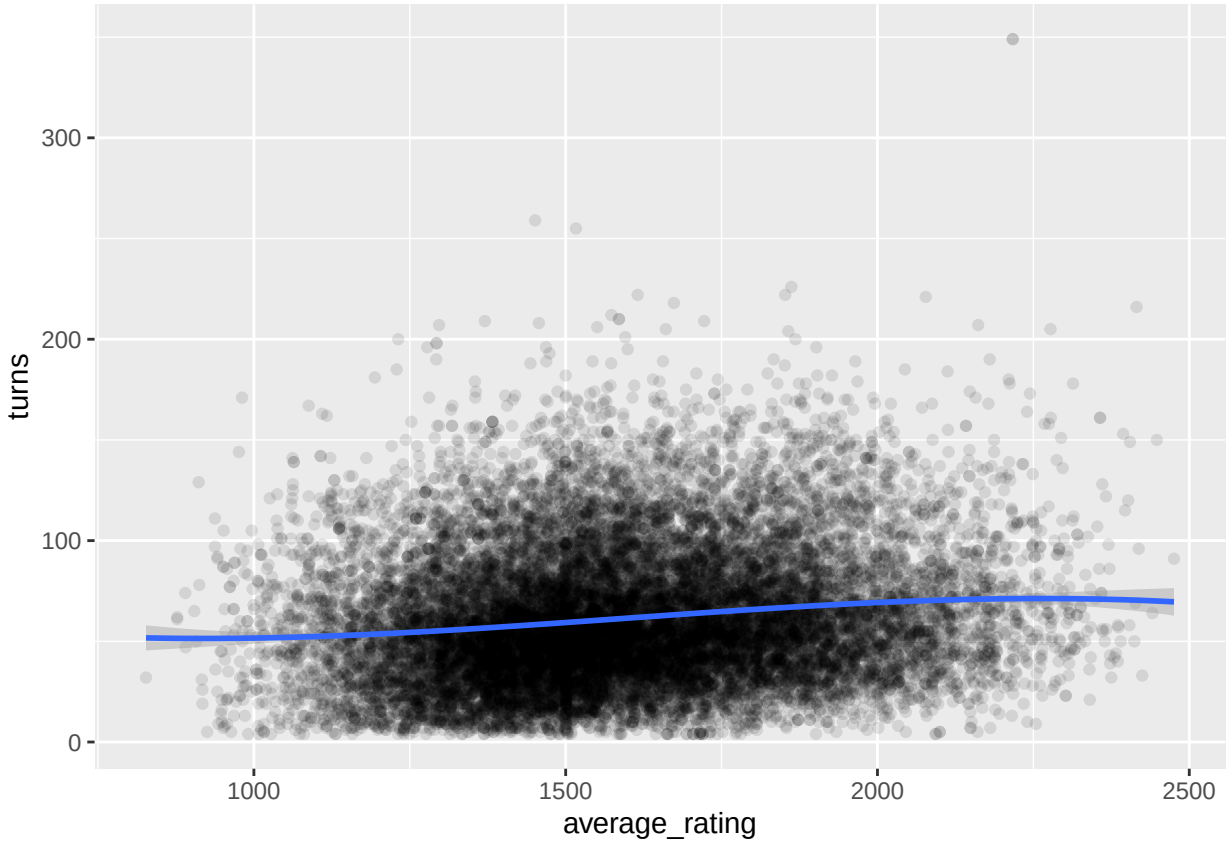
```
confint(model, level=0.95)
```

```
##                                 2.5 %     97.5 %
## (Intercept)                  1591.56836  1598.4552
## poly(turns, 3)1              4732.30352  5701.1107
## poly(turns, 3)2             -2027.55624 -1059.0844
## poly(turns, 3)3               913.66042  1883.2880
## poly(opening_moves, 3)1      9997.02256 10967.1762
## poly(opening_moves, 3)2        50.39039  1018.7419
## poly(opening_moves, 3)3     -1751.51512  -783.1143
```

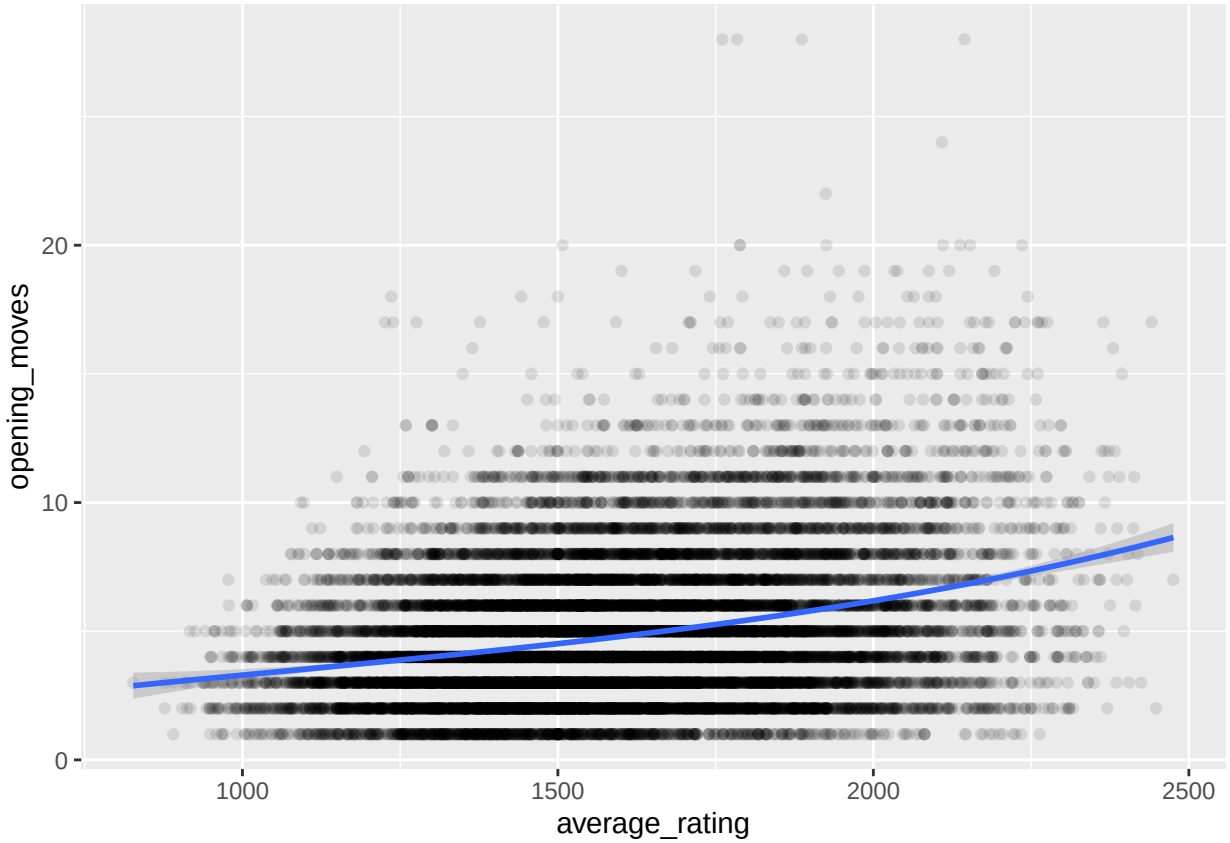
This model seems to show that the higher rated terms are decently statistically significant, at least enough to be used in the analysis. The 95% confidence interval for all predictors does not include 0, and the p-values for the predictors are very small. At the very least, these coefficients and confidence intervals clearly show that turns and opening moves affect average rating.

Let's plot each variable's linear model individually.

```
ggplot(chess, aes(x = average_rating, y = turns)) + geom_point(alpha = 0.1) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 3))
```




```
ggplot(chess, aes(x = average_rating, y = opening_moves)) + geom_point(alpha = 0.1) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 3))
```



Again, these graphs clearly show that both turns and opening moves increase with higher rated players, with opening moves having the more significant effect. Both the regression and the confidence intervals can be seen in the graphs. The graphs have a very small confidence interval for the regression.

Below is the R squared and adjusted R squared for this regression.

```
cat("R Squared: ", sm$r.squared)
```

```
## R Squared:  0.1099012
```

```
cat("Adjusted R Squared: ", sm$adj.r.squared)
```

```
## Adjusted R Squared:  0.1096309
```

These R squared values are close to 0. Therefore, although turns and opening moves have a statistically significant effect on average rating, there is still a very large amount of variance in these variables. While any player can easily have a long opening and a long game, higher rated players are more likely to.

In conclusion, I was correct that lower rated players lose more often early in the game. Also, lower rated players play shorter openings while initiate and lose with unusual openings more often. Lower rated players are less likely to resign and draw, while being more likely to finish the game with a checkmate. This data stays the same whether white or black wins. Higher rated players especially play common openings well, while lower rated players can easily trap themselves if they play an uncommon opening. I was incorrect to say hypothesize that lower rated players have a higher winning percentage with white.

If a lower rated player wants to improve, I would suggest learning one easy opening very well that can be reached consistently in games. Also, I would suggest learning to not make one-move mistakes, rather

considering what an opponent will do to counter, helping to extend a game as long as possible. Lower rated players should consider draws as a more possible outcome to the game. I would suggest a lower rated player actually keep their tendency to drag a game out to checkmate as it can allow an opponent to make a mistake along the way.

Although these suggestion may seem somewhat obvious to someone familiar with chess, it is interesting to see how these suggestion are backed up by evidence from this data set.