Summary of DSC 530 Final

The statistical question I wanted to ask is, is there a correlation between the rating difference of the two players, and the number of turns until the end of the game. I wanted to ask this question because it seemed to me like if one player was much better than the worse player would lose quickly. Games with similarly skilled players would possibly go on longer.

My EDA showed me right away that there was a lot of bias in my variable for the rating difference between players. I got the data from lichess.org and they match similar players against each other. That meant that many values were near zero for the difference between players. This ended up making it more difficult to find a correlation. If I could do it all over again, I would try harder to find a dataset with more random assignments of ratings for games. The problem is that all chess websites match similar players and so would tournaments. The only real random possibility are more casual games for fun which brings in its own bias. Using non-normal data is usually ok but in this case I think it was so much bias that it impacted my findings. I would also be interested in seeing number of turns based on age of players. There is a hypothesis that younger players will play more aggressive and short games while older players play longer, more defensive games.

Overall, I was happy with the analysis that I did. I think a lot of times in data science it may feel like the goal to make an incredible discovery but a lot of the times I am sure it works at that discoveries aren’t so impressive. Honest reporting of the facts to the best of your knowledge should be the real goal in data science.