Wrap Up

My project was to analyze the Magnus Carlsen Lichess Games Dataset. This dataset contains each game Magnus has played on Lichess, each move he and his opponent has made, and a username to real name table. I wanted to figure out which variables affect his win rate the most. The variables I considered were initially color (black or white), time control (how long each player gets to complete a game), opening (first few moves of the game), and strength of opponents (Lichess rating). As I worked on my project, I decided to include year as a variable and changed opening to just the first move. I made this change because the opening classification system would be hard to explain and there are at least 1,300 different openings. Also, instead of win rate, I looked at his fractional score. Fractional score is equal to number of wins + number of ties * 0.5 divided by total number of games. I think it's a better metric than win rate because chess has a lot of draws and draws affect ratings. Chess tournaments use fractional scores as well.

I used Python and the NumPy, pandas, Matplotlib and Seaborn libraries to analyze this dataset. The first thing I did was clean the raw dataset using before analyzing it. I mainly used pandas methods including df.drop(), df.rename(), df.replace(), df.loc(), df.merge(), df.sort_values(), df.reset_index(), and df.apply() to clean the dataset. I haven't used df.replace(), df.merge(), and df.reset_index() before starting this project. I only used df.replace() because the dataset had a csv that matched Lichess usernames to real names. I learned how to convert that csv into a dictionary using zip() and then I used df.replace(dict) to convert almost all of the usernames in the games csv to real names. While doing that, I also learned about the encoding parameter for pd.read_csv() which dealt with foreign characters. I wanted to include first moves in my analysis and every move that Magnus has made on Lichess was in another csv. After getting all the first moves by Magnus from the moves csv, I used df.merge() to merge the first

moves with the games csv. I also used df.merge() when creating a new column but in hindsight, I think using df.apply() would have been better for that. I ended up using df.reset_index() to add empty rows to dataframes created from df.loc(). This allowed me to df.merge() based on index. In addition, I learned how to rearrange columns using df[[<column names>]].

After cleaning the dataset, I mainly used matplotlib to create bar charts so I can visualize the data. I have previously done data visualization using matplotlib so that was not new to me. However, I needed to do some linear regression analysis and create scatter plots. I have not done this before using Python. I ended up using df.plot.scatter() and plt.scatter() to create scatter plots from dataframes and lists respectively. I used df.corr() and np.corrcoef() to get correlation coefficients. I also used np.polyfit() to get a line of best fit to use in my linear regression analysis. To plot the residuals from my main scatter plot, I used sns.residplot(). I wanted to annotate my graphs as always, but I could not figure out how to get the residual values from the plot itself. I ended up calculating the residuals using np.arrays and annotated my residual plot accordingly.

I think I successfully answered my initial question with my analysis of this dataset. Fractional score and opponent Lichess rating have a correlation coefficient of -0.95. This means that there is a really strong, negative association between fractional score and opponent rating. This finding makes a lot of sense since a person would win a lot less against stronger opponents and win a lot more against weaker opponents. Since this association is so strong, it's hard to make conclusions based on just the fractional scores of other variables. For example, Magnus had a 0.731 fractional score in 2018 and a fractional score of 0.707 in 2020. Does this mean that Magnus played worse in 2020 or he fought stronger opponents in 2020? It's hard to tell with just a bar chart that graphs the fractional scores of variables like year, color, first move, and time

control. To deal with this issue, I tried to use a metric called linear performance rating. It is a simple way of getting someone's rating using the average rating of their opponents and their fractional score against those opponents. These two factors vary a lot and ideally, the linear performance rating will reduce that variance. The equation is (average rating of opponents) + 800 * (fractional score) – 400. Unfortunately, there are issues with this equation that I did not realize until after I graphed the linear performance rating of each variable. The linear performance rating equation makes the assumption that the fractional score will approximately change from 1 to 0 when the average opponent rating increases by 800. This is a fair assumption, but it only makes sense when someone fights against both stronger and weaker players in that rating range of 800. For example, let's say I have a fractional score of 0.39 against 1800 rated players and a fractional score of 0.87 against 1400 rated players. This means my linear performance rating ranges from 1696 to 1712. This range is a lot smaller compared to my opponent rating range of 400. In this case, I could successfully use linear performance rating as a metric to judge other variables like year and color without average opponent rating determining anything. However, a person that only fights against stronger opponents will almost never have a fractional score above 0.5. A person that only fights against weaker opponents will almost never have a fractional score below 0.5. This person is Magnus Carlsen since he is the best player in the world. Since Magnus's fractional score ranges from 0.64 to 0.84 depending on the variables I have analyzed, the linear performance rating ends up being heavily positively correlated with average opponent rating. To prove this, I created a scatter plot to show the correlation between linear performance rating and average opponent rating. The correlation coefficient between those two ratings is 0.997. This ultimately means that there is no point in using linear performance rating as a way to isolate average opponent rating from other variables. With this knowledge, I

ended up creating a scatter plot for fractional score vs average opponent rating to use as my main analysis to draw conclusions from. I plotted 22 points based on variables such as year (2018, 2019, 2020, 2021), color (Black, White), opponent rating range (<2700, 2700s, 2800s, 2900s, 3000s, 3100s), time control (1 minute+0 second games or not), first move as black (d5, Nf6, c5, g6), first move as white (e4, d4, c4, Nf3) and another point based on the dataset in its entirety. Then I created a residual plot to clearly see which variables led to a higher than expected fractional score. According to this plot, Magnus plays better as white than as black. This makes a lot of sense since white moves first and in general, white wins more games than black. Magnus's best first move as white is d4 and Magnus's best first move as black is g6. There is a huge 5 point difference between g6 and the 2nd best move as black, c5. I am not sure why this is the case, but this significant difference also exists in Lichess's game database. For example, when white plays e4, black wins 26% of the time with c5, and 30% of the time with g6. In addition, the plot shows that Magnus's best year was 2019 and his worst year was 2018. This could be because Magnus started playing on Lichess from December 2017. In 2019, maybe he had a skill gap against most other players, and they slowly caught up in 2020 and 2021. I cannot be 100% sure. To add on, Magnus plays a lot worse when he's not playing 1 minute+0 second time control. The 'not 1+0' time control has the lowest residual at -0.046. Magnus plays better than expected against opponents rated less than 2700 and 2900-2999 rated opponents. It makes sense that Magnus plays better than expected against opponents rated less than 2700 because at some point, the relationship between fractional score and opponent rating stops being linear. For example, Magnus would probably have at least a 0.99 fractional score against opponents rated between 0 and 2000. I don't know how to explain why he plays better than expected against specifically 2900-2999 rated opponents. He plays worse than expected against opponents rated

more than 3100. This makes sense for a similar reason to why he plays better than expected against really low rated opponents. The correlation coefficient for the fractional scores and average opponent ratings from each variable value is -0.703. R^2 is 0.494. This means that about half of the variance of fractional score shown in the scatter plot can be explained by average opponent rating. That means that the other half of the variance is due to other factors including time control, color, first move, and year. In addition, Magnus's mood, time of day, specific opponents can also influence this model.