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.