Kyle Sherman

CS-2610

3/28/2022

Research Project: Data Analysis Plan

**Does white have a first move advantage?**

To answer this question, I will need to use the winner, white\_rating, black\_rating variables. It is important to note that the ratio between the white and black rating will likely play a factor in the outcome of the game as it would be hard for a player of low skill to defeat a skilled player regardless of any first move bias. To counter this, I will filter out any games where the ELO difference is outside of a +- 250 points boundary. I will also need to filter the victory type to only include games which result in a mate or a draw; if a game ends as a result of time running out or a resignation, then there is a good chance that the result could be impacted from external factors and could skew the outcome. I think it would also be beneficial to group the white team win/loss frequency by ELO rating as it will allow me to get a better insight into whether the skill level of the player will change the respective player’s ability to utilize any start bias that may exist. To answer this question, I will be looking at the win/loss ratio of the white team grouped by ELO rating.

**What openings have the highest and lowest average win rate grouped by player ELO rating?**

This question is going to require the most work to solve. There are many factors which could have an impact on the outcome, so it will be important to consider any outliers. In order to ensure the integrity of the findings, I will need to filter out any of the non-rated games because players will typically play unrated when they want to try new openings that they don’t know or they want to mess around, which would have an impact on the legitimacy of the outcome. Additionally, I will want to filter out any games that don’t end in a mate or a draw. Lastly, I will filter out any games where the ELO difference is outside of a +- 250 points boundary. Anything outside of this range is going to lead to significantly unpredictable variability in prediction accuracy. Because there are many different openings, it would likely be best to select only the top 5 most played openings for each ELO group. The least played openings are going to have such a small ratio of the total game pool that the impact of a single game will be too significant. The win rate for each opening selected will be the win/loss ratio for the respective opening.

This is all preliminary and is subject to change as I understand more about the dataset and how to address my specific questions.

This section is very preliminary and is subject to change.