Derek Sullivan

Capstone Project 1-Chess Dataset

January 11, 2021

Problem Statement

The world of chess is data rich and has been the target of many different studies over the course of history. The amount of variation and possibilities in the game has inspired data scientists and machine learning specialists to study chess on very advanced levels and even create programs that are nearly impossible to defeat.

In this study I plan on analyzing a dataset featuring 20,000 games and dividing them by skill level (determined by player rating, a well established feature of chess matches for decades) and seeing how different skilled players approach the game. I intend to suss out what higher level players tend to do and compare that to those in the middle and lower end of the skill spectrum. I will analyze how many moves they stay on the opening move patterns, what the most common chess openings (sequence of moves) are, and how many moves it takes to win.

Goals

1. **Determine what makes great players so good:** Discovering the patterns and tactics of highly skilled players will help anyone with intentions to become a great player.
2. **See how lower ranking people behave:** We can assume that lower level players have worse habits than higher level players. This will make the differences between the top skill group and find pitfalls to be avoided by any player.

Specifications

We will be using the Chess dataset available on Kaggle.com and writing code in Jupyter Notebooks.

Milestones

1. Preparing Data

   The dataset includes 17 variables for each of the 20000 games. These need to be prepared for analysis. There will be some removed and some created and it will be important to make sure the datatypes are appropriate for analysis.

2. EDA

   Exploratory data analysis will be applied to the problem using tools taught from the various datacamp resources provided in this course. Data Visualization using matplotlib.pyplot and seaborn will be used to determine patterns in the data, as well as statistical analysis tools provided by pandas and numpy.

3. Communicating Findings

   The final milestone of the project will be communicating the findings of the project and discussing their significance. Differentiating between the practices of low skill and high skill players will be an interesting effect to observe and discuss.