**Introduction**  
**- Why was the project undertaken?**

We chose to analyze Chess Data because it seemed very fitting for machine learning. Also, because that chess provided interesting data to dissect from game results to individual move types. We wanted to see if we can organize and examine the data to see if we can make meaningful predictions on the outcome of games. It was interesting to see if throughout the entirety of the dataset, which moves were predictable.

**- What was the research question, the tested hypothesis, or the purpose of the research?**

The purpose of the research was that We wanted to analyze a data set of chess games and use the information to make predictions about who will win a game of chess.

**Selection of Data**

**- What is the source of the dataset? Characteristics of data?**

The source of the dataset is hosted from Internet Archive. The full data set is hosted here, <https://archive.org/download/chess-data>, and the csv file is here, <https://ia902501.us.archive.org/23/items/chess-data/Chess_Data.txt>

**- Any munging or feature engineering?**

The data set is given with many columns that we do not need, data in the wrong format, and many records that are not useful. To make things easier we upload the data from the local machine and then generate a PKL file of the prepared data frame.

As you can see from the drop statements, we have dropped some of the records that are not needed and some columns that were just left blank. From that and additional statements that were used in munging the code, we are finally able to have a dataset that has been prepared and formatted to our needs.

**Methods**

**- What materials/APIs/tools were used or who was included in answering the research question?**

In answering the research question, we imported a plotting library called matplotlib, while doing this all on Jupyter notebook which helps turning code into a report. We also hosted a compressed version of the data on GitHub.

**Results**

**- What answer was found to the research question; what did the study find? Was the tested hypothesis true? Any visualizations**?

Some of the questions we had to analyze this dataset were: “What were the most common first 8 moves for white and black each year?”, “For each year in the data set, which pieces were most left in play at the endgame?”, and “How many pieces are usually left when the game ends?”. Our main prediction was to see if White will win based on the predictors in our model.

From some of the hypothesis’ questions we found that the white chess pieces had a higher ELO rating than the black pieces. Therefore, after further analysis, we saw that the advantage was to white pieces because the white pieces get to move first. To answer our first hypothesis, we found that Black’s opening moves were more diverse compared to White’s and that each player had established a response to their opponents move.

We found out that as the game went on, there were less unique moves to do. After analyzing throughout the years as time passed, we found the number of unique moves towards the end kept on increasing. After comparing the 70s and the 90s, we concluded that the diversity of the endgame meta has greatly increased, which in a way shows our evolution.

**Discussion**

**- What might the answer imply and why does it matter? How does it fit in with what other researchers have found? What are the perspectives for future research? Survey about the tools investigated for this assignment.**

From all this we can see that White’s ELO rating is higher than Black’s which shows that the White pieces do win a bit more than black pieces. This proves that our predictors were correct. Our results showed a 68%-win advantage for White, while others showed 53%-win rate for White pieces, which is about 15% better than baseline.

**Summary**

**Most important findings.**

For the machine learning part, we found out that Classification was better suited for our prediction because we wanted to put games in 2 buckets, those that white would win or lose. We set the predictors and scaled the data; we built a classifier for the precisions and printed the first 10 and compared those to the first ten values which showed us the accuracy.

<https://youtu.be/la1dwOYPHys> -- Project Video