

**TASK**

**Exploratory Data Analysis on the Chess Dataset from Lichess.org**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

**DATA CLEANING**

**As we can observe, the data set is 20,000 rows and 16 columns. Due to the scale of this data set, it is important to keep in mind that visualizations and patterns may need to be presented in a way that is easy to interpret and understand. Fortunately, we can see that the data set is complete with no missing values, and all data types have been verified as accurate.**

**To clean the data I have used an array of techniques to allow it easier for me to display data in graphs. Some of these techniques include Dropping information not needed, renaming columns, converting Unix time to date and time. I have then changed datatypes of some columns which would notallow me to visualise certain data how I want.**

**Upon examination of the data set's columns, I find that all of them are relevant to the analysis. However, I have identified that 'created\_at' and 'last\_move\_at' columns are in Unix time format which may be difficult to interpret. To rectify this, I will convert these columns into a human-readable format by creating new columns and storing the new values. Once the conversion is complete, the original columns 'created\_at' & 'last\_move\_at' will be removed. Additionally, 'id' and 'increment\_code' columns are not essential to the analysis and will be removed as well.**

**MISSING DATA**

**There was no missing data, all data was visualised using a missingno matrix to visualise if there was any data missing along with checking if there were any nulll values.**

**DATA STORIES AND VISUALISATIONS**

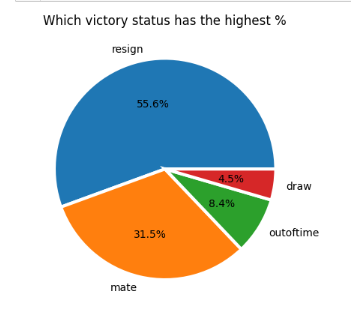
**Objective: To uncover the correlation between the strategies employed by chess players and their performance in the game.**

**Data Set: The chess data set comprises a wealth of information on games played by a diverse set of players, including player ratings, openings used, and game outcomes (win, loss, or draw).**

**Methods: Through a comprehensive analysis, I will employ a variety of techniques to uncover key insights and relationships between a player's approach to the game and their performance. Specifically, I will delve into the correlation between a player's opening move and their success rate, in order to determine the significance of the initial move in determining the outcome of the game.**

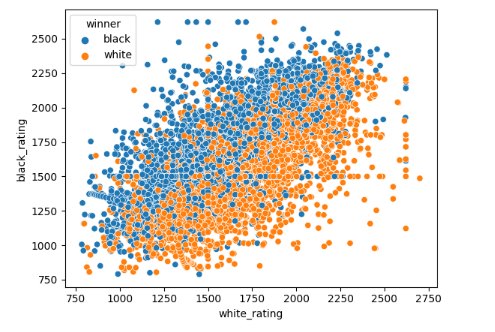
**Expected Outcomes: The analysis is expected to reveal a strong correlation between the strategies employed by players and their performance in the game. Additionally, it is anticipated that the analysis will indicate that the opening move is a crucial factor in determining the success of a player's performance.**

# **Graph displaying which status had the highest %**



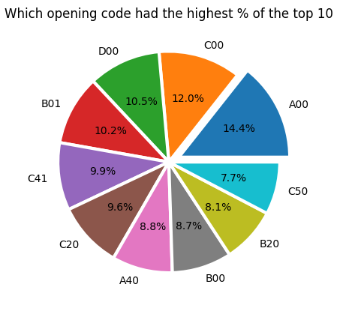
**As we can see from the pie chart resign had the biggest %. Is there a reason for this? Are the quality of players higher so opponents quit a lot more?**

# **Relation between the game ending in a resign and the opponets rating**



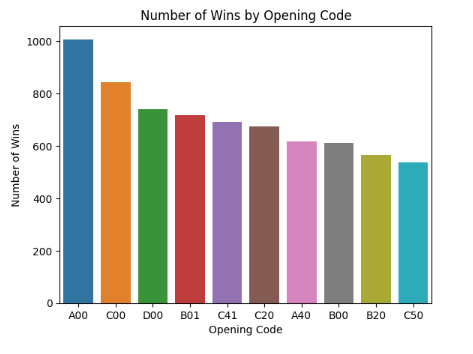
**Based on the scatter plot analysis, there does not appear to be a strong correlation between a player's high rating and their opponent resigning. It is worth noting that chess can be a game that can change rapidly with one poor move, which may contribute to the high rate of resignations, whether due to frustration or the perception that continuing the game would be futile and a waste of time.**

# **Number of occurrences of the opening code showing only the top 10**



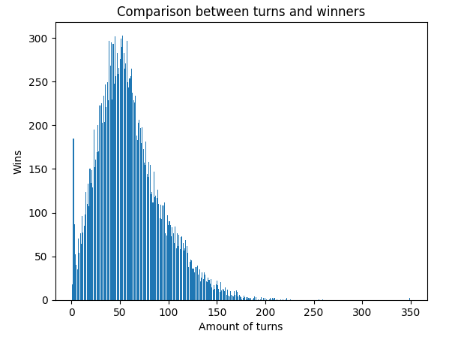
**In order to make the pie chart more visually appealing, I chose to only display the top 10 most common openings used. From the data, it appears that the top 4 openings have relatively similar percentages. To gain a deeper understanding of this, further analysis is needed to compare more data. This could potentially provide insight into the best starting position for a chess game. It would be interesting to compare more data and see if this trend holds true.**

# **Number of Wins by Opening Code**



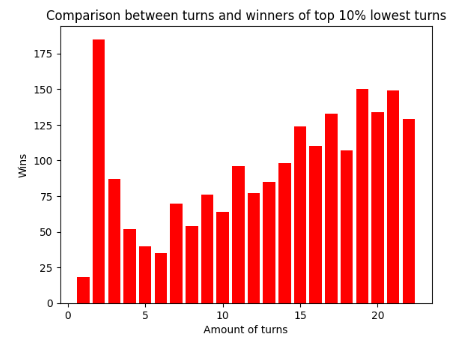
**The two charts clearly indicate that the top 10 openings used have the highest win rate in order. This observation may explain why these openings are used so frequently. It is worth considering that the high usage of these openings may be a contributing factor to their high win rate, or it could be that they are simply the most effective starting strategies to secure a win**

# **Comparison between opening turns and winners**



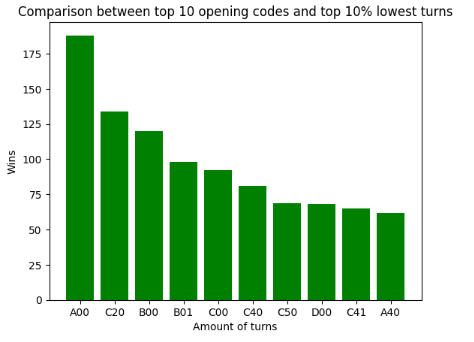
**From the comparison, it is evident that the majority of wins occur around 35-70 turns. Additionally, there is a significant spike at the very beginning of the game. This raises the question of whether there is a correlation between the number of turns taken and the likelihood of winning. To gain further insight, it would be beneficial to examine the gameplay of the top 10% of players with the lowest amount of turns. These players are likely among the most skilled and studying their strategies and tactics could yield valuable information on how to improve one's own gameplay.**

# **Compare data of the top 10% lowest turns**



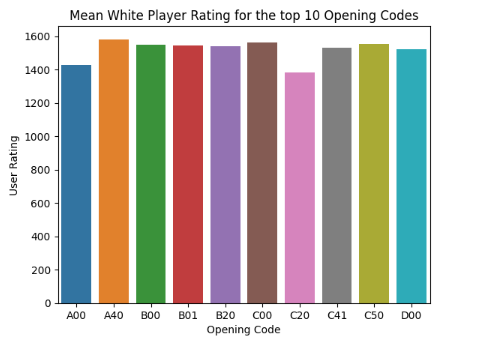
**Based on the data, it appears that the second turn has the highest number of wins. From my knowledge of chess, this move is commonly referred to as the "fool's mate," which is typically executed by an inexperienced player. It is important to note that this move is often a result of a mistake made by the opponent, and thus, it may not be a reliable strategy to win in a competitive setting. Therefore, it may be wise to disregard this play in the analysis and focus on more sound strategies.**

# **Comparison between top 10% of quickest games and their first move**



**This comparison is cross-referencing the top 10 opening codes with the top 10% of games with the lowest number of turns. From the data, it appears that A00 is once again at the top, which was also the most common opening code associated with wins. Based on this information, it is safe to conclude that A00 is the best starting position to play from. I will look further into what the top rated players use**

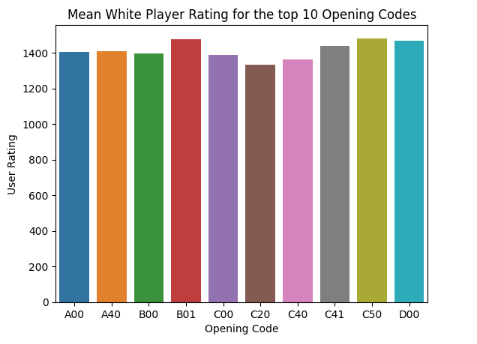
# **The Mean of White Player Rating for the top 10 Opening Codes**



**The graph illustrates the average rating of white players for the top 10 opening codes in chess. In order to effectively present the data, the mean was utilized given the large amount of information. As we can observe, the A00 opening code is not utilized by the highest-rated players, with A40 taking its place. This is intriguing as A40 typically ranks lower in other charts and A00 has the highest win rate.**

**Additionally, charts for black player ratings will also be created for further analysis.**

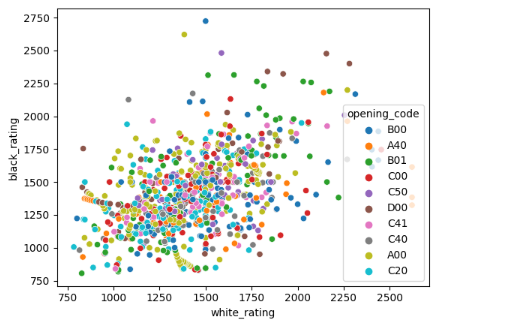
# **The Mean of black Player Rating for the top 10 Opening Codes**



**The graph displays the average rating of black players for the top 10 opening codes in chess. As the data indicates, the A00 opening code is not favored among the highest-rated players, with B20 taking its place. This is noteworthy as B20 is typically not as well represented or ranked lower in previous charts and A00 has the highest win rate.**

**Despite A00 not reaching the top spot, it's still within the top 10 of the highest rated players. This is still positive data to consider.**

# **Scatterplot of The Mean of Player Rating for The Top 10 Opening Codes**



**To provide a clearer understanding of the comparison, I thought it would be beneficial to present an alternate representation. At first glance, it appears that A00 is the most frequently used, however, it seems to be primarily utilized by players with average to below average ratings.**

**The results of extensive data analysis confirm that the initial move plays a significant role in determining the outcome of a game. The A00 move, which is the most commonly used among top players, has been found to have the highest win rate. Furthermore, the top players consistently utilize the top 10 moves, and their superior performance serves as evidence of the correlation between move choice and game outcome.**

**THIS REPORT WAS WRITTEN BY : Connor Walker**

