# Data Science Project

**Chinmay Kulkarni** (PES1UG19CS127), **Anudeep CVS** (PES1UG19CS070),
**Debaditya Ray** (PES1UG19CS133), **Brundha P** (PES1UG19CS116)
Semester 3, Section B

## I. **ABSTRACT**

Lots of information is contained within a single chess game, let alone a full dataset of multiple games. It is primarily a game of patterns, and data science is all about detecting patterns in data, which is why chess has been one of the most invested in areas of AI in the past. In this project we proceed to analyze the various variables present in a chess game and the unknown relations and interpretations which exist between them.

## II. **INTRODUCTION**

Being students of data science, we first perform some exploratory data analysis on the multiple chess games by describing the various variables present in the dataset. We clean the dataset of missing data by examining and imputing and remove all unwanted observations not required. That is followed by classifying the variables into different categories and visualising the values.

It is then normalised using z-statistics for easy manipulation and comparison of data followed by hypothesis testing. In hypothesis testing we make several possible predictions and check whether they are supported or rejected by using p-value tests. Finally, we finish with correlation by graphing various variables and checking for their shapes and obtaining inferences

## III. **DATASET**

The original dataset has been taken from https://www.kaggle.com/datasnaek/chess. It contains information of about 20000 games of chess collected from a selection of users on the site Lichess.org. It consists of 16 columns initially and over 20000 rows. Here is the list of the columns.

- Id – Unique id generated for every chess game
- Rated – Whether the game was a rated game where if a player wins or loses it improves or decreases the profile rank.
- Start time (Omitted) – The time at which the game started
- End time (Omitted) – The time at which the game ended
- Game time – The duration of the game obtained by subtracting start time from end time
- Turns – Number of turns played in the game

- Victory_status – The result of the game whether it was checkmate, draw, the players gave up or time ran out
- Winner – The winner of the game
- Increment_code – The increment code of the game. In chess increment code is of the format X+Y where X is time in minutes for a player to make the first move and Y is the increment in seconds added to each move's time if a move is played within Y seconds
- White_id – The player id of white
- White_rating – Rank or rating of white's player
- Black_id – The player id of black
- Black_rating – Rank or rating of black's player
- Moves – List of moves played in the game in standard chess notation
- Opening_eco – Standardised code given for every game's opening move
- Opening_name – Name of the opening move
- Opening_ply – Number of moves in the opening phase

After modifications, the data considered for study consists of 15 columns and 9367 games with 450 introduced null values. Categorical variables include victory_status and winner while rated, turns, white_rating, black_rating, opening_ply come under discrete variables and continuous variable consists of game time.

## IV. **DATA CLEANING**

We manually filter out the games whose game time is 0 since the users must have lost access to internet and those games hardly have any data. Likewise, we remove the rows with huge erroneous game_times which are not supposedly generated by the system.

Data cleaning is required since later there will be errors while inferencing something from the data due to incomplete value and unwanted outliers and the graphs obtained will not be accurate. Thus, we proceed to clean the data. For cleaning the dataset, we handle the missing values by different ways.

- For game ids we generate a unique string combination for every missing game id.
- Since rated is a binary variable we replace half of the missing values with 1 and other half with 0's
- White_id's and black id's being unique strings we assign a prefix of Player_Whie or Black with the row number of the cell
- For values of increment code, the mode of the existing values is given for every missing value
- Black_rating and white_rating values are filled with the mean value
- Game time value is filled with median value
- For missing opening names, we check if opening ply moves matches with other rows whose opening name is present. For the unique variation moves we check all possible combinations by comparing every subset from beginning of the moves to the nth or opening ply move. From the list we just manually check in an opening move book for the match.
- Opening_ply values are obtained by finding the already present opening ply values with similar opening name records. For unique opening names we just manually check the opening ply up

- Values of turns are obtained by counting number of moves in moves column
- Winner values are filled with checkmate if the last character in moves column ends with ++ or #
- For moves, victory_status and opening_eco there is no standardised way to fill them so we just leave them as it is

The outlier records are removed if the values of the variables game_time, opening
ply, black_rating, white_rating and turns are outside 1.5*IQR times from the mean of the data values.

## V. EXPLORATORY DATA ANALYSIS

By visualizing the data, we can make several inferences.

- By plotting the rated variable, we see that a greater number of games are rated which implies that the players on Lichess are more of a competitive nature than casual game play as they are concerned with their profile rank. A lot of players resign or give up and do not have the patience to play the game till the last move.
- Black_rating and white_rating are moderately positively correlated.

Similarly several other inferences can be drawn from the graphs.
We then proceed to normalize the variables for data consistency and preventing anomalies. Normalizing the data sets all numeric variables to a common scale without changing their differences. Also the ranges in numeric form are extremely huge and not efficient for data manipulation. It is done by replacing z scores for every value of a numeric variable.
From the graphs, we can see that the data has been normalized.

## VI. HYPOTHESIS TESTING

We conduct several hypotheses but two of them are

**Q:** *Is the winner of the games most of the time white at a 90% level?*

**A:** Since the dataset is a sample dataset from a population of games played, this is a hypothesis test for population proportion.
Where p0= 0.3750 (data taken from https://en.wikipedia.org/wiki/First-move_advantage_in_chess#Winning_percentages)

H0 is that white is not the winner of most games p0 <= 0. 3750
H1 would be that white is the winner of most games p0 > 0. 3750

Where n is the number of total games in the sample.
The output which the code produces is Reject H0.
Therefore we can conclude that H0 is false and that white wins the game most of the time.

**Q:** Can we prove that the most common opening move is not the Open Game (or Double King's Pawn Opening)/ ["e4", "e5"] at a 95% level?

**A:** H0 is that most games begin with the opening e4, e5 that is p0 >= 0.5
H1 would be that most games do not begin with the opening e4, e5 that is p0 < 0.5

The output which the code produces is Reject H0.

Therefore we can conclude that H0 is false and that the most common opening move is not the Double King's Pawn Opening.

## VII. **RESULTS AND DISCUSSION**

From doing this project we obtained a lot of information.

1. Players on Lichess are more of a competitive nature than casual game play as they are concerned with their profile rank.
2. Lot of players resign or give up and do not have the patience to play the game till the last move.
3. Black_rating and white_rating are moderately positively correlated.
4. White wins the game most of the time.
5. The most common opening move is not the Double King's Pawn Opening.
6. The mean white rating is greater than the mean black rating
7. Most games ending in a checkmate is plausible.
8. Most games ending in the range of 40-70 moves is plausible
9. Most of the games are rated.

By using only some simple statistics, we have obtained so many results for such a small dataset, this shows how informative can chess get along with the power of statistics.