“Exploring Online Chess Games Using a Data Science Based Approach”

Yadvender Singh
50466664
yadvende@buffalo.edu
Computer Science and Engineering
University at Buffalo
Buffalo, New York, USA

Shawna Saha
50468278
shawnasa@buffalo.edu
Computer Science and Engineering
University at Buffalo
Buffalo, New York, USA

I. Introduction

Since time immemorial Chess has been a prevalent game that has captivated the minds of a large audience. The highly esoteric nature of this game, along with its simple rules but overly complex and intricate plays make it an ideal sport for information extraction via data analysis. Data Science can be applied to chess game analysis to build player profiles, build fraud detection system in online chess games, recommender system for chess training and practice etc.

II. Problem Statement

Though the game of Chess has been played for millennia, there is a lack of comprehensive understanding of the dynamics of the game. Determining the optimum set of moves to achieve victory is a significant struggle since there is no one right strategy to play the game. The game's arcane nature has made it a challenge even for the brightest brains.

With the development of technology acquiring data on chess games have become relatively feasible. By analyzing the enormous databases of chess games, data science approaches can help in spotting patterns and trends, create models that may forecast upcoming moves and results.

Thus, this project aims to employ data science techniques in the analysis of a large dataset of online chess games to gain insights into the patterns, trends and strategies used by different chess players at different levels of the game.

The primary questions we are trying to find answers for through this project are:

1. What are the most common openings used by successful chess players and how does it affect their game?
2. Is it possible to predict the outcome of a game based on the opening move chosen by a player?
3. What sequence of moves have the largest probability to win?
4. How does a player’s rating affect the choice of opening move and the outcome of the game?
5. How do players of similar ratings perform when played against each other?
6. Does a white player have an advantage in how the game unfolds due to the rule of the game that states – “White moves first”?

Answering these questions will help in attaining the following objectives:

1. By analyzing the data to better understand a player’s performance such as – number of wins, number of losses, preferred first move etc. can help in coaching and suggest improvements to the game.
2. Insights into the most common opening strategies, number of turns, sequence of moves, the most successful tactics etc. can help in creating better training materials.

III. Data Sources

The dataset used for the implementation of the project are detailed below:
Name of the dataset: Chess Game Dataset (Lichess)
Source: **www.kaggle.com**
Link: **https://www.kaggle.com/datasets/datasnaek/chess?datasetId=2321&searchQuery=eda**

Description: This is a dataset containing data collected from 20,000 online chess games from the site Litchess.org. The dataset has 16 features in total. They are as detailed below:

1. ID: Game id.
2. RATED: If a player is rated or not.
3. CREATED AT: Start time of the game.
4. LAST MOVE AT: End time of the game.
5. TURNS: Number of turns in the game.
6. VICTORY STATUS: The outcome of the game.
7. WINNER: Which player won.
8. INCREMENT CODE: The amount of time added to the clock after each move made by a player.
9. WHITE ID: Id of the white player.
10. WHITE RATING: Rating of the white player.
11. BLACK ID: Id of the black player.
12. BLACK RATING: Rating of the white player.
13. MOVES: Set of moves made by the players in chess notation.
14. OPENING ECO: Standardized codes for the chess opening names.
15. OPENING NAME: The chess opening used by the player.
16. OPENING PLY: Number of moves in the opening phase.

IV. DATA CLEANING / PROCESSING

The detailed process of cleaning the dataset is described as below:

1. **Column Names Formatting**: Total number of rows in the original dataset is 20,058. At first, the column names were formatted for better visual representation.

```
Formatted column names with associated datatype are as folows :
ID                 object
RATED                bool
CREATED AT        float64
LAST MOVE AT      float64
TURNS               int64
VICTORY STATUS     object
WINNER             object
INCREMENT CODE     object
WHITE ID           object
WHITE RATING        int64
BLACK ID           object
BLACK RATING        int64
MOVES              object
OPENING ECO        object
OPENING NAME       object
OPENING PLY         int64
dtype: object
```

2. **Changing column datatype:** "CREATED AT" and "LAST MOVE AT" features were converted from object to datetime format.

```
ID                         object
RATED                        bool
CREATED AT        datetime64[ns]
LAST MOVE AT      datetime64[ns]
TURNS                      int64
VICTORY STATUS            object
WINNER                    object
INCREMENT CODE            object
WHITE ID                  object
WHITE RATING               int64
BLACK ID                  object
BLACK RATING               int64
MOVES                     object
OPENING ECO               object
OPENING NAME              object
OPENING PLY                int64
dtype: object
```

3. **Adding a column:** “MATCH DURATION” feature was added to store the duration of each Chess match.

   MATCH DURATION = LAST MOVE AT - CREATED AT

```
ID                          object
RATED                         bool
CREATED AT         datetime64[ns]
LAST MOVE AT       datetime64[ns]
TURNS                       int64
VICTORY STATUS             object
WINNER                     object
INCREMENT CODE             object
WHITE ID                   object
WHITE RATING                int64
BLACK ID                   object
BLACK RATING                int64
MOVES                      object
OPENING ECO                object
OPENING NAME               object
OPENING PLY                 int64
MATCH DURATION            float64
dtype: object
```

4. **Selecting features:** The features of interest were extracted from the dataframe. The list of features selected are:

```
RATED                 bool
TURNS                int64
VICTORY STATUS      object
WINNER              object
INCREMENT CODE      object
WHITE ID            object
WHITE RATING         int64
BLACK ID            object
BLACK RATING         int64
MOVES               object
OPENING NAME        object
OPENING PLY          int64
MATCH DURATION     float64
dtype: object
```

5. **Dropped ‘seconds’ information from “INCREMENT CODE”:** The column increment code contained data of form “10+2” where the part after ‘+’ sign represented seconds. For time increment we are only interested in the minutes hence formatted the column by splitting the data and keeping only the minutes information. Then datatype of the column was changed from object to Integer.

```
print(game_data['INCREMENT CODE'].head(5))
```

```
0    15
1     5
2     5
3    20
4    30
Name: INCREMENT CODE, dtype: int64
```

6. **Checking for null values:** There were no null values in the code, hence null removal operation was not required.

7. **Removing duplicates:** There was duplicate data in the dataset, so the duplicate columns were dropped. Data count after removal of duplicates was 19,629 records.

8. **Dropping games shorter than 2 moves:** The shortest possible checkmate in Chess is “Fool’s mate” where the black player wins after two moves. Any number of turns less than 2 is erroneous and insignificant. Hence data containing turns less than 2 were dropped. Data count after this operation was 19,345.

9. **Converting column to list:** Each entry in the MOVES column was converted to list for easy access in future.
10. **Extracting Opening Categories:** OPENING NAMES contained the main category of opening names along with the subcategories mentioned. Hence, the main category of opening names was extracted and saved in a new column "OPENING CATEGORY" for better analysis.

```
Sicilian Defense       2508
French Defense         1269
Queen's Pawn Game      1017
Italian Game            953
King's Pawn Game        869
                       ...
Alekhine Defense #3       1
Doery Defense             1
Petrov's Defense #5       1
Global Opening            1
Pterodactyl Defense       1
Name: OPENING CATEGORY, Length: 180, dtype: int64
```

11. **Adding "MEAN RATING" column:** MEAN RATING column was added to the dataset.

    MEAN RATING = (WHITE RATING + BLACK RATING) / 2
12. **Diving the game into levels:** The game was divided into 4 levels namely- Beginner, Intermediate, Advanced and Expert based on the rating distribution. This information was stored in a new column "GAME LEVEL".

```
Beginner : 800 to 1225
Intermediate : 1225 to 1650
Advanced : 1650 to 2075
Expert : 2075 to 2500
```

13. **Removing columns having 0 game duration:** On exploratory data analysis it was found that 25% of the MATCH DURATION data had 0 values which was erroneous. Hence only those columns were retained where MATCH DURATION was greater than 0. Data count after this operation was 10,885 records.
14. **Segregating games into variants:** The data was segregated into four variants based on the increment time: Bullet, Blitz, Rapid and Classical and the information was stored in a new column "GAME VARIANT".
15. **Removing least used openings:** On exploratory data analysis it was observed that 25% of the openings were used only once. Hence openings that were used less than 20 times were removed. The count of data after this operation was 7,672 records.

V. EXPLORATORY DATA ANALYSIS

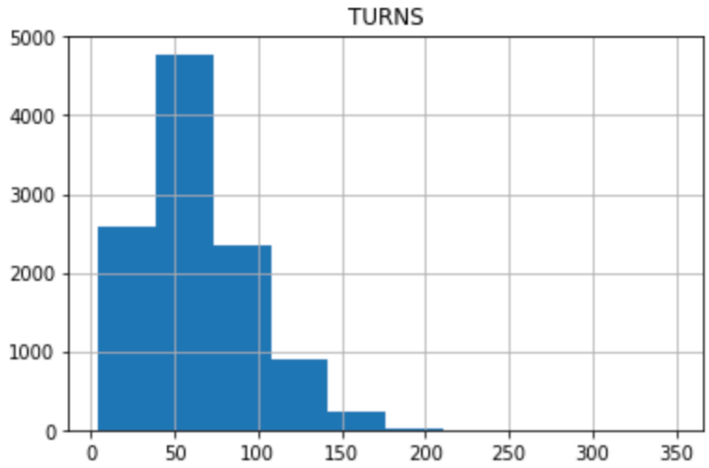
The process of exploratory data analysis is described as below:

1. **Data distribution analysis:** The details of data distribution were analyzed to understand how the statistical representation of the data. Through this process the mean, standard deviation, data count for each column etc. were accessed. This gave a better idea about the maximum and minimum value present in each column. By studying the below chart, it is evident that 25% of the data in MATCH DURATION column had 0 values which is erroneous. These records were then removed.

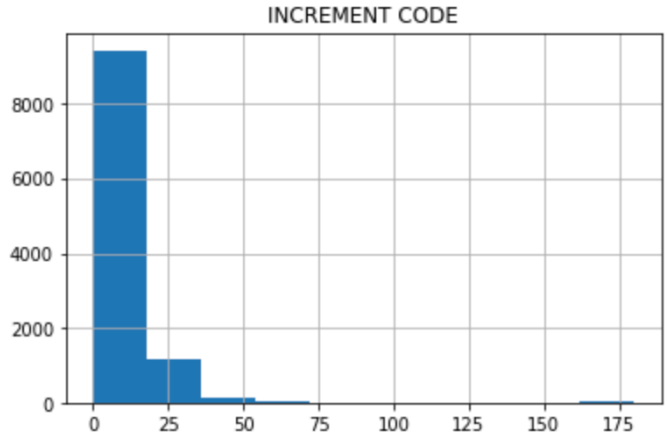
	TURNS	INCREMENT CODE	WHITE RATING	BLACK RATING	OPENING PLY	MATCH DURATION	MEAN RATING
count	19345.000000	19345.000000	19345.000000	19345.000000	19345.000000	19345.000000	19345.000000
mean	61.313363	13.580408	1598.331300	1590.755286	4.847713	14.495559	1594.543293
std	32.991158	15.960817	287.969503	288.395886	2.788960	80.429841	260.699355
min	4.000000	0.000000	784.000000	789.000000	1.000000	0.000000	827.000000
25%	38.000000	10.000000	1402.000000	1396.000000	3.000000	0.000000	1410.500000
50%	56.000000	10.000000	1568.000000	1563.000000	4.000000	3.970067	1570.500000
75%	79.000000	15.000000	1792.000000	1785.000000	6.000000	13.230350	1771.500000
max	349.000000	180.000000	2700.000000	2621.000000	28.000000	10097.411683	2475.500000

2. **Plotting Histograms for numerical features:** Plotting histograms corresponding to TURNS, INCREMENT CODE, WHITE RATING, BLACK RATING, OPEN PLY, and MEAN RATING were plotted.

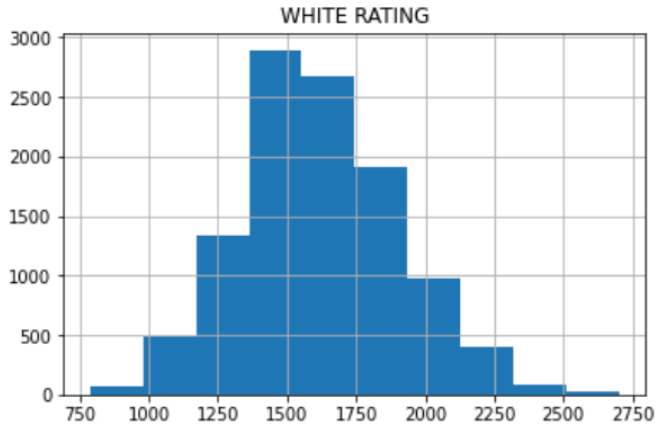
   TURNS: The **x-axis** represents the range of turns divided into equally spaced bins and the **y-axis** represents the number of observation or the frequency of the turns that fall into each bin.



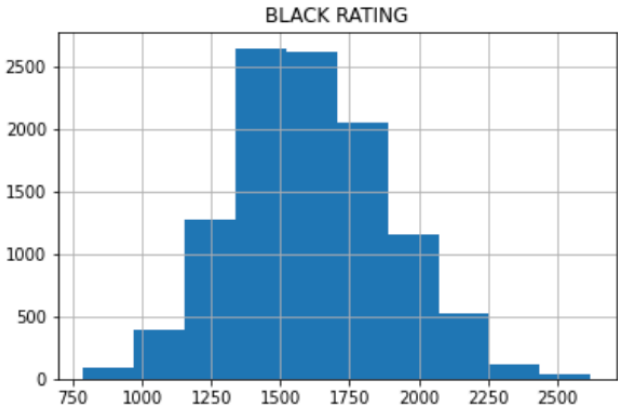
   INCREMENT CODE: The **x-axis** represents the range of increment code divided into equally spaced bins and the **y-axis** represents the number of observation or the frequency of the increment code that fall into each bin.



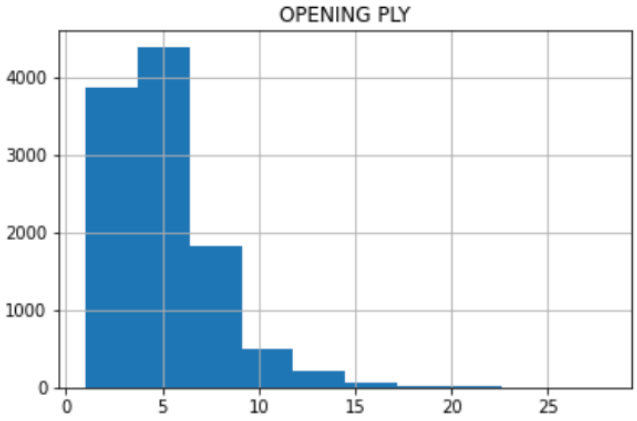
WHITE RATING: The **x-axis** represents the range of white rating divided into equally spaced bins and the **y-axis** represents the number of observation or the frequency of the white rating that fall into each bin.



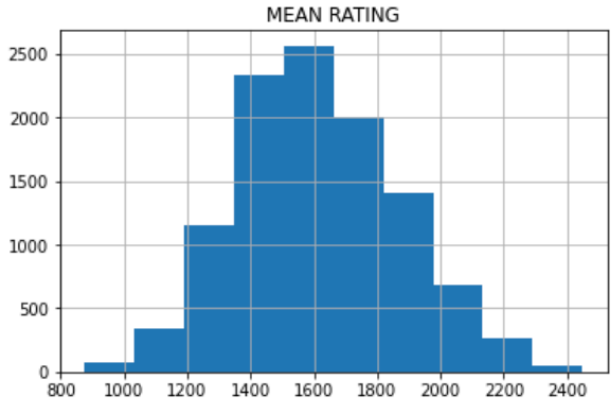
BLACK RATING: The **x-axis** represents the range of black rating divided into equally spaced bins and the **y-axis** represents the number of observation or the frequency of the black rating that fall into each bin.



OPEN PLY: The **x-axis** represents the range of open ply divided into equally spaced bins and the **y-axis** represents the number of observation or the frequency of the open ply that fall into each bin.

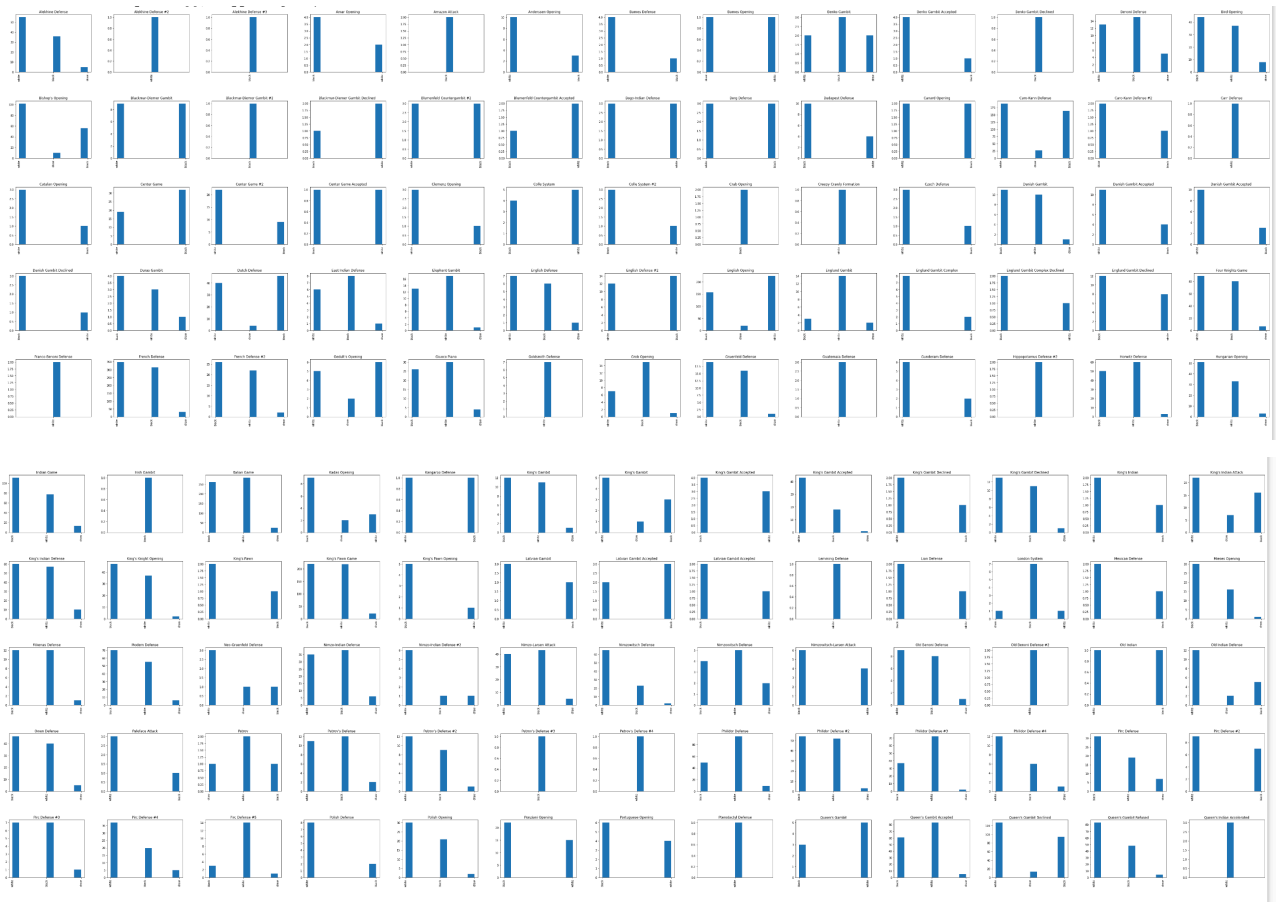


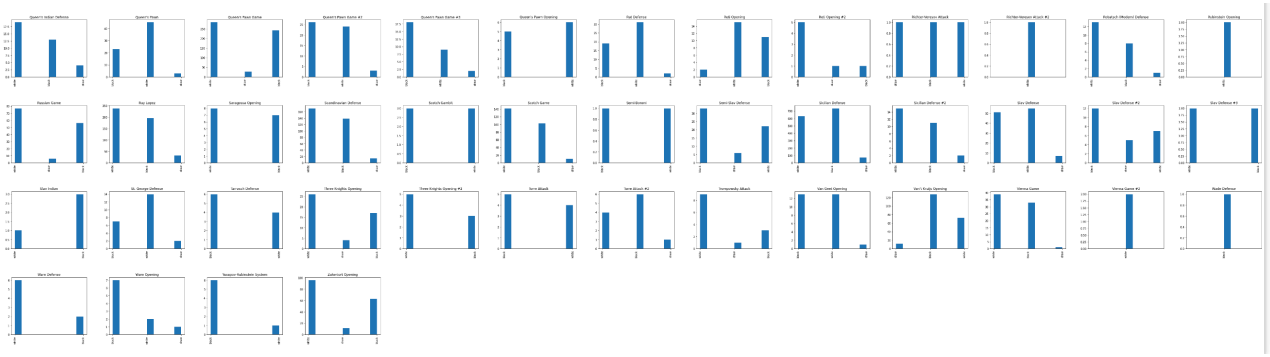
MEAN RATING: The **x-axis** represents the range of mean rating divided into equally spaced bins and the **y-axis** represents the number of observation or the frequency of the mean rating that fall into each bin.



3. **Plotting histogram to visualize the winner against each opening game:** For each opening category this visual representation provided information about the count of white and black players that won. The **x-axis** represents the winner bin and **y-axis** represents the count for each winner bin. For example, for Alekhine Defense count of white winnings is more than count of black winnings.

   Owing to the different types of opening categories, the below attached pictures of the graphs may not be as clear. Please refer the code for better clarity plots.





4. **Opening name distribution analysis:** On exploring the distribution of the OPENING NAME feature, it was observed that the min and 25% of the openings were used only once.

```
count    1261.000000
mean        8.632038
std        18.329767
min         1.000000
25%         1.000000
50%         3.000000
75%         8.000000
max       212.000000
Name: OPENING NAME, dtype: float64
```

To extract the openings that were used only once, the value counts corresponding to each opening was analyzed.

```
Sicilian Defense                1439
French Defense                   694
Italian Game                     568
Queen's Pawn Game                551
Ruy Lopez                        467
                                ...
Irish Gambit                       1
Creepy Crawly Formation            1
Alekhine Defense #2                1
Blackmar-Diemer Gambit #2          1
Petrov's Defense #4                1
Name: OPENING CATEGORY, Length: 173, dtype: int64
```

Openings that were used only once do not have significant contribution in the prediction of which player will win or the best set of moves that may contribute to a positive outcome in the game. Since the standard distribution was 18, openings that were used less than 20 times were removed from the dataset.

5. **Least and most used openings:**

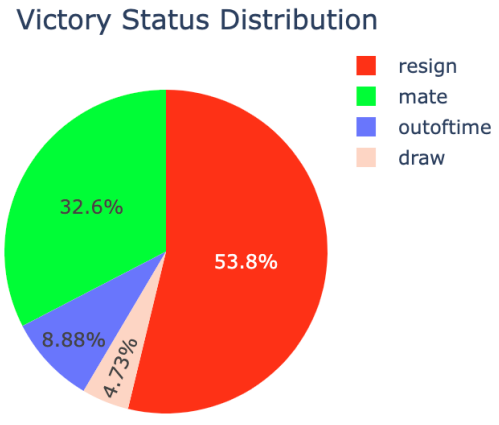
On further analyzing the openings, it was observed that preferred openings by Chess players are as below:

```
Van't Kruijs Opening                              212
Sicilian Defense                                  192
Sicilian Defense: Bowdler Attack                  174
Scotch Game                                       147
Queen's Pawn Game: Mason Attack                   136
French Defense: Knight Variation                  133
Caro-Kann Defense                                 127
Scandinavian Defense: Mieses-Kotroc Variation     119
Indian Game                                       114
Queen's Pawn Game: Chigorin Variation             114
Name: OPENING NAME, dtype: int64
```

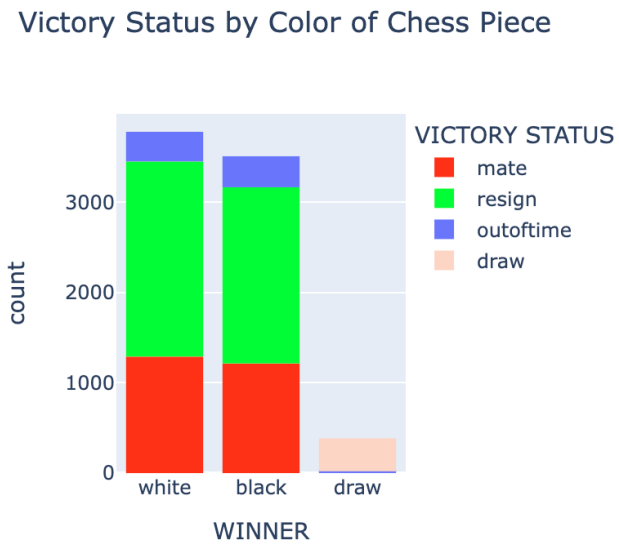
List of openings that are not usually preferred and very few white players choose to start the game with are as below:

```
English Opening: Agincourt Defense                        21
Ruy Lopez: Morphy Defense | Caro Variation                21
Pirc Defense: Classical Variation                         21
French Defense: Queen's Knight                            21
Italian Game: Classical Variation | Giuoco Pianissimo     21
Four Knights Game: Spanish Variation                      21
Nimzo-Indian Defense: Ragozin Variation                   21
Semi-Slav Defense                                         20
Caro-Kann Defense: Advance Variation | Short Variation    20
Italian Game: Giuoco Pianissimo                           20
Name: OPENING NAME, dtype: int64
```

6. **Outcome analysis:** On plotting the victory status of the chess games, it was found that most of the games end with a player resigning.



7. **Winner analysis:** By analyzing the plot it can be inferred that whites win more than black.

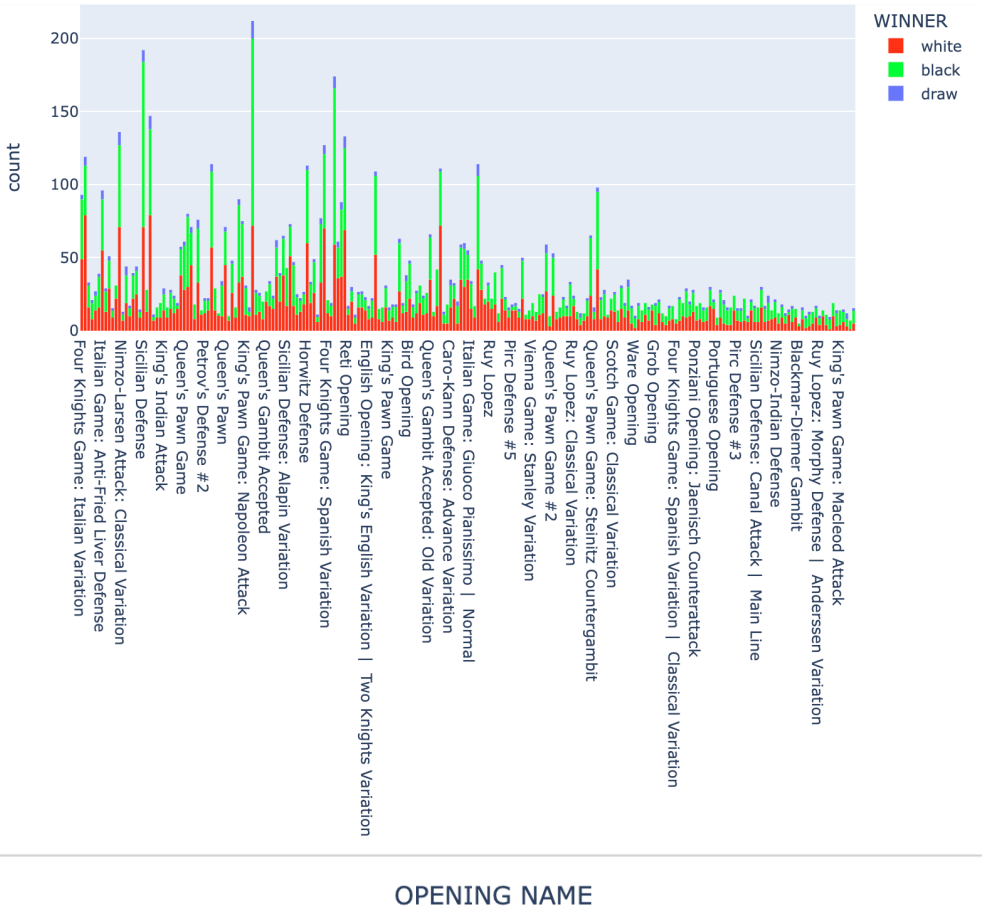


8. **Analyzing winner by opening name:** By analyzing the below graph it can be inferred that 'Van't Kruji's Opening' is the opening that won the most games.

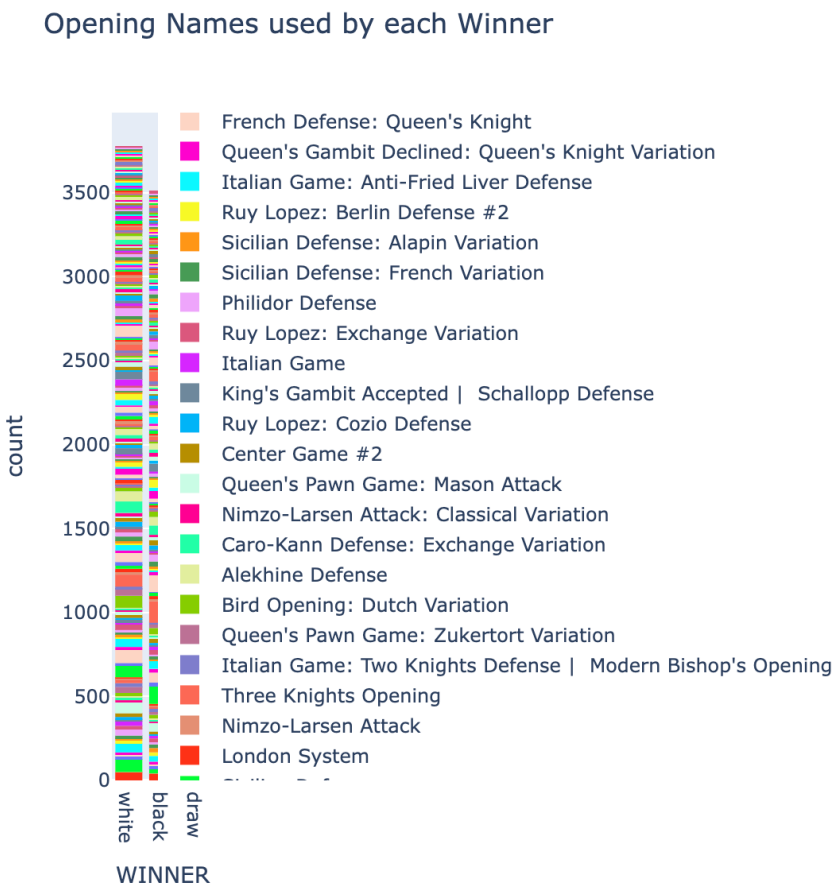
   Black Winner = 128

   White Winner = 72

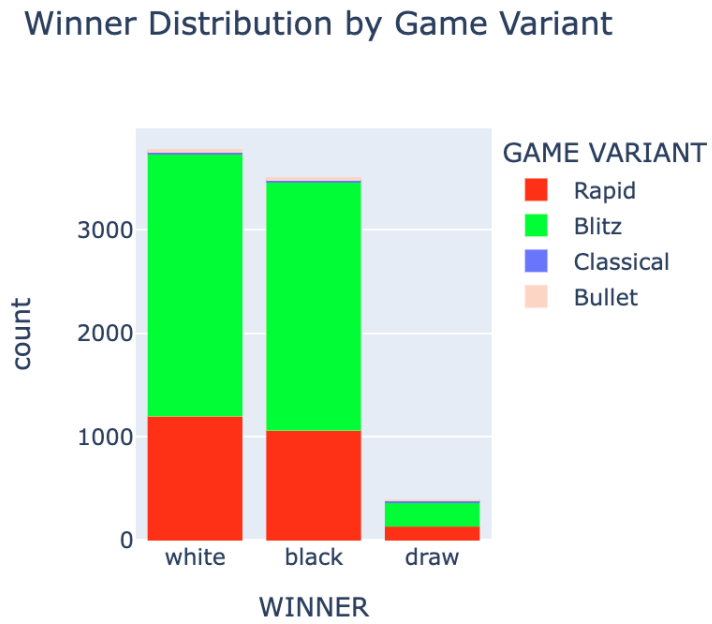
   Total wins by using this opening = 200



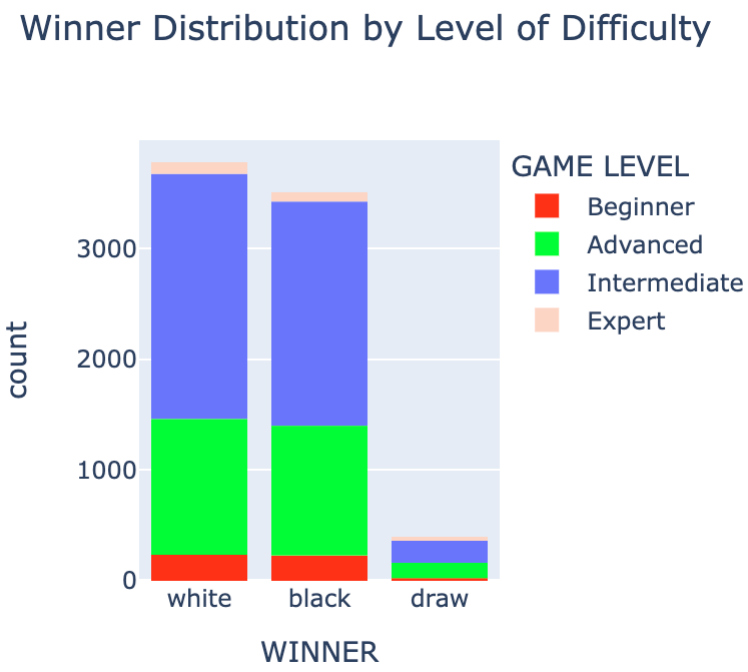
9. **Analyzing opening names by winner count:** By analyzing the plot it can be observed that whites won most using the 'Scandinavian Defense: Mieses-Kotroc Variation' opening (count = 79) whereas blacks won most by using the 'Van't Kruji's Opening' (count = 128).



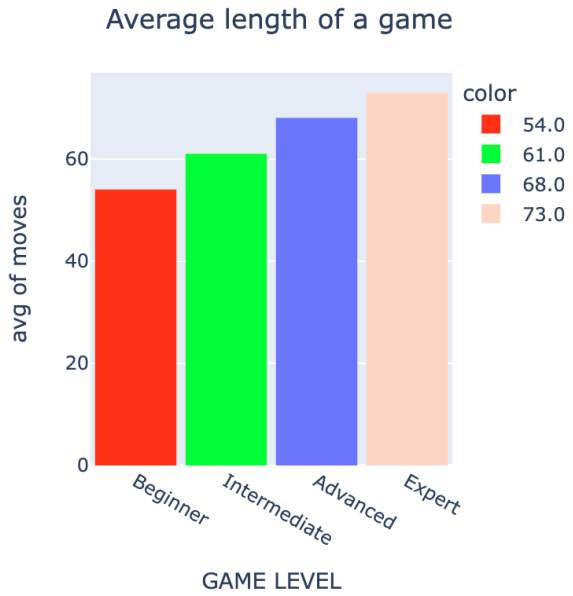
10. **Winner distribution by game variant:** It can be observed from the below plot that whites won mostly on Blitz (count = 2527) while blacks won on Bullet (count = 21).



11. **Winner distribution by game difficulty:** From the plot it can be inferred that whites won on every level of difficulty.



12. **Average length of the game by difficulty:** It can be inferred from the below plot that with the increase of difficulty level, the average length of the match also increases.



13. **Moves analysis:**

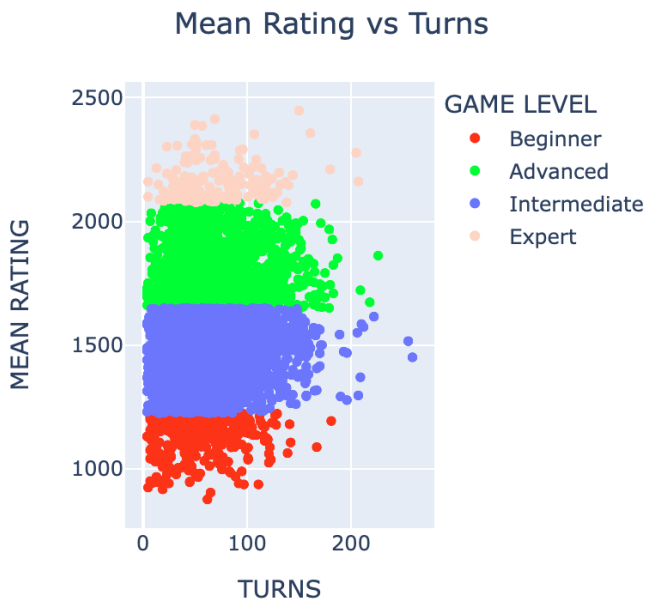
Top 5 set of move winners used are shown below:

```
f4 e6 g4 Qh4#
6
e4 e5 Qh5 g6 Qxe5+ Be7 Qxh8
3
e4 e5 Bc4 Na6 Qf3 b6 Qxf7#
2
e4 b6 f4 Bb7 d3 Nf6 Be3 e6 Nd2 g6 c4 Bg7 e5 Nh5 g3 Na6 Ngf3 c5 g4 Nxf4
Bxf4 f6 Qe2 fxe5 Bxe5 Bxf3 Nxf3 Bxe5 Qxe5 Rf8 Bg2 Nb4 O-O Nxd3 Qc3 Nf4 Ne5
Ne2+ Kh1 Nxc3 Rxf8+ Kxf8 Bxa8 Qxa8+ Kg1 Ne2+ Kf2 Qe4 Nxd7+ Ke7 Nb8 Qxc4
Nc6+ Kf8 Rf1 Nf4 Ke1 Qe2#    2
e4 e5 d3 Bc5 c4 Qf6 Ne2 Qxf2+ Kd2 Bb4+ Kc2 a5 Kb3 Qc5 Bg5 Nc6 a4 Nd4+ Ka2
f6 Nbc3 b5 Qb3 bxa4 Nxd4 exd4 Qxa4 dxc3 Kb3 Bb7 Bf4 Bc6 Bxc7 Bxa4+ Ka2
Qxc7 g4 Qf4 Rd1 Qf2 Bh3 Qxb2#
2
Name: MOVES, dtype: int64
```

First Moves used mostly by players are shown below:

```
e4    4164
d4    1321
e3     218
Nf     131
c4     108
f4      86
g3      66
d3      36
b4      33
Nc       8
Name: MOVES, dtype: int64
```

14. **Mean Rating vs Turns:** It can be observed from the plot below that with the increase of mean rating, the number of turns used in the match also increases.



ACKNOWLEDGMENT

We would like to take advantage of this opportunity to express our gratitude to Dr. Eric Mikida and Dr. Shamsad Parvin, our lecturers and Smarana Shrikant Pankati, our assigned project mentor for their insightful comments, patience, and time.

REFERENCES

1. https://bookdown.org/pq2142/finalproj/

2. https://community.ibm.com/community/user/ai-datascience/blogs/moloy-de1/2020/08/27/points-to-ponder

3. https://towardsdatascience.com/analysing-lichess-games-with-r-c4f8b0bc512c

4. Demo code provided in class for cleaning and EDA operation.