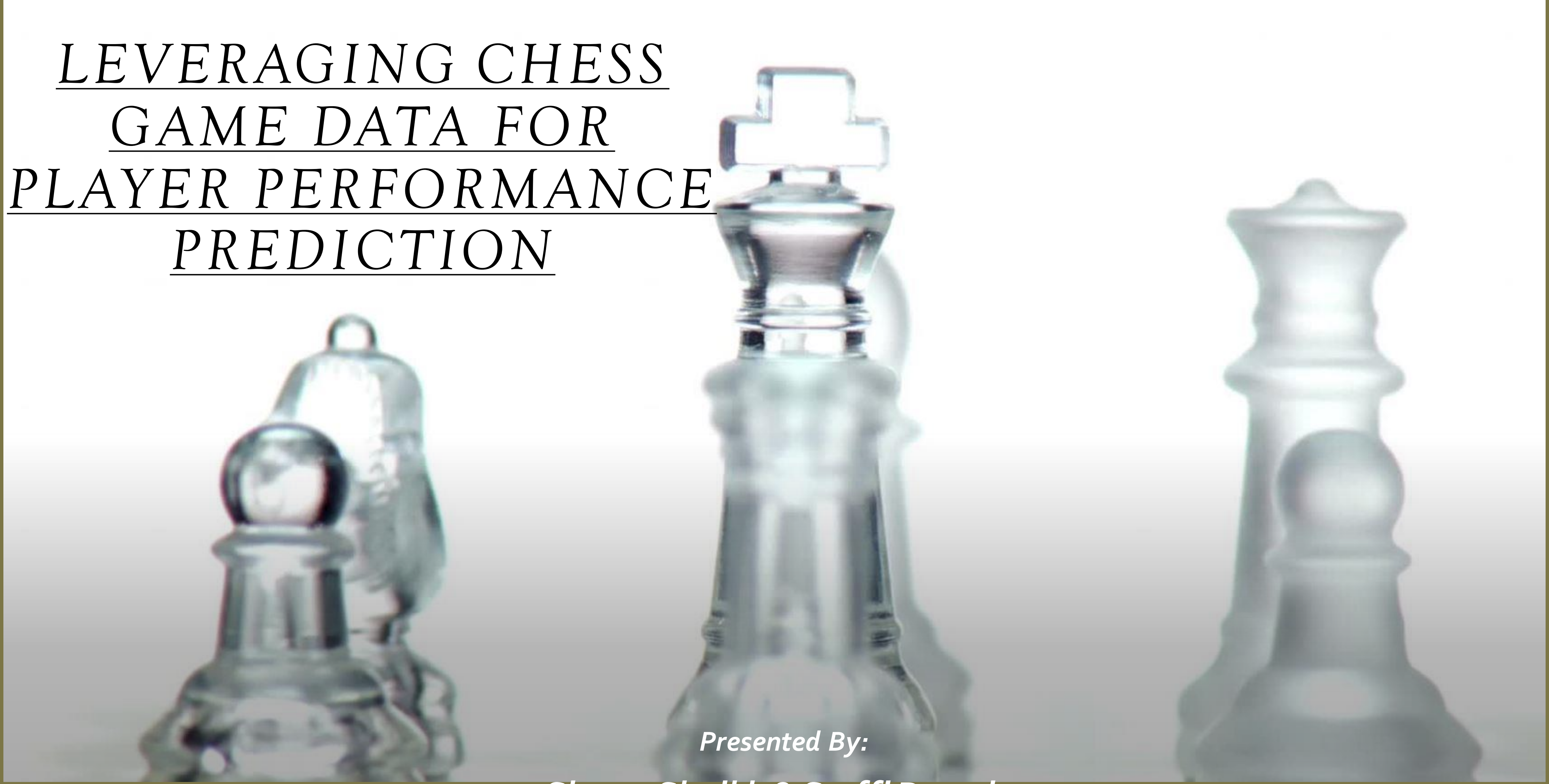
WORCESTER POLYTECHNIC INSTITUTE

DS 502 – Statistical Methods for Data Science



LEVERAGING CHESS GAME DATA FOR PLAYER PERFORMANCE PREDICTION

Presented By:

Sheroz Shaikh & Steffi Dorothy

04/23/2024

Sheroz Shaikh
Ms. Data Science

Steffi Dorothy
Ms. Data Science

TABLE OF CONTENTS

DESCRIPTION OF THE PROBLEM

DESCRIPTION OF THE DATASET

MOTIVATION

EXPLORATORY DATA ANALYSIS

TRAINING AND ASSESSING THE MODELS

RESULTS

CONCLUSION

DESCRIPTION OF THE PROBLEM

Our project aims to analyze a large dataset of chess games from Lichess to predict player performance accurately.

We developed a predictive model based on game attributes and player characteristics to forecast game outcomes.

The goal is to uncover strategic insights by identifying recurring patterns and tactics used by players.

Analyzing moves made by players will help predict the likelihood of winning.

Examining player-specific attributes like skill level and playing style will identify key performance factors.

DESCRIPTION OF THE DATASET

The "Chess Game Dataset (Lichess)" from Kaggle is a credible data set that contains around 20K samples of games from Lichess.org with 16 features.

The dataset includes game outcomes, player ratings, move sequences, time controls, and more, allowing for in-depth exploration of chess gameplay dynamics.

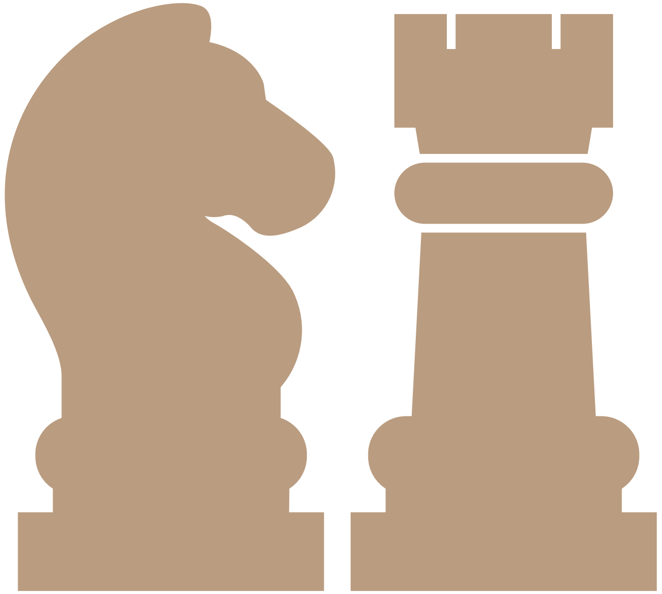
Each row in the dataset represents a specific game, with columns representing attributes such as game identifiers and timestamps.

The dataset is formatted appropriately and licensed under CC0 1.0 Universal Public Domain Dedication.

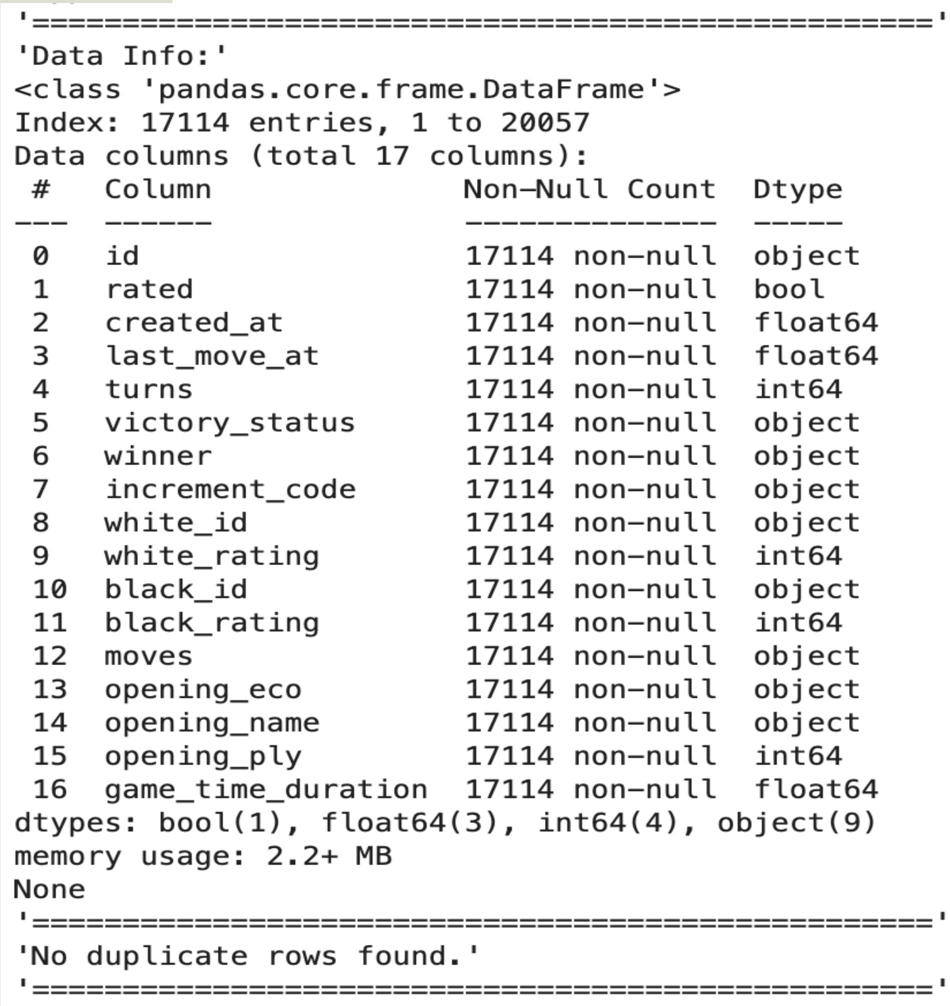
Descriptive statistics, EDA, and machine learning techniques are used to analyze player performance and opening strategies.

MOTIVATION

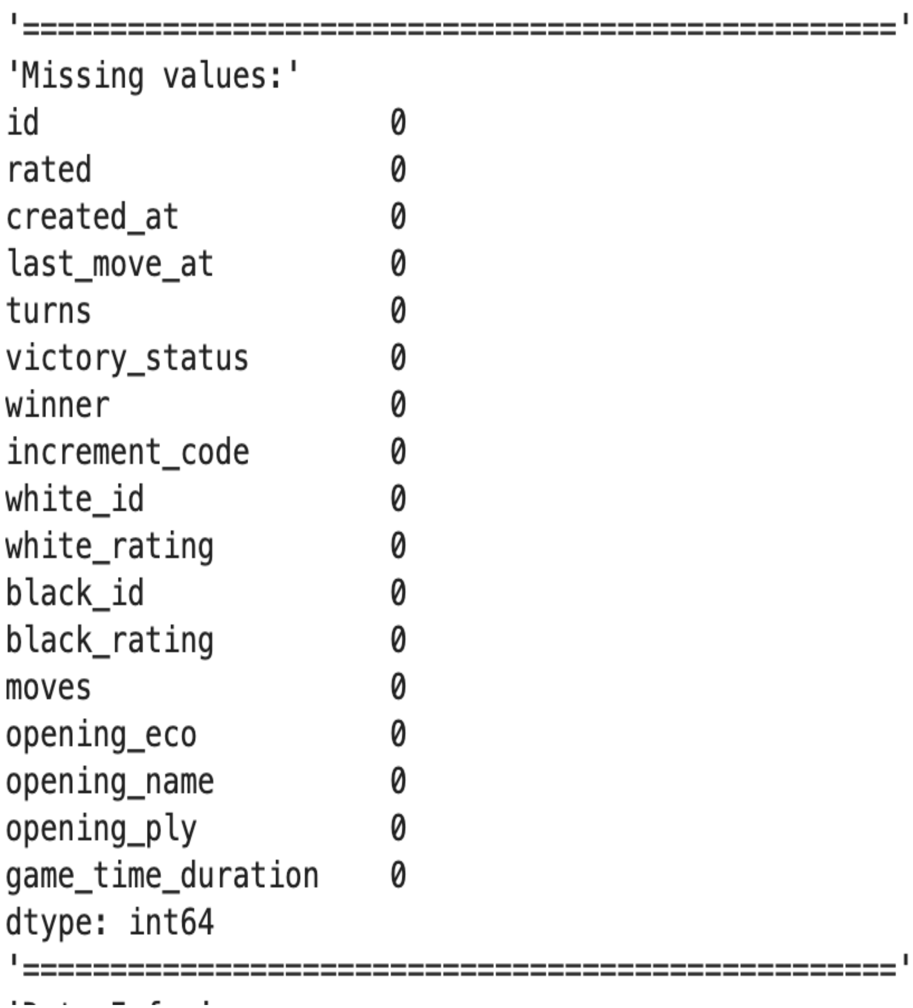
We were curious about the use of data analysis to understand complicated systems like chess.



PREPROCESSING



```
'=================================================='
'Data Info:'
<class 'pandas.core.frame.DataFrame'>
Index: 17114 entries, 1 to 20057
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   id                  17114 non-null  object
 1   rated               17114 non-null  bool
 2   created_at          17114 non-null  float64
 3   last_move_at        17114 non-null  float64
 4   turns               17114 non-null  int64
 5   victory_status      17114 non-null  object
 6   winner              17114 non-null  object
 7   increment_code      17114 non-null  object
 8   white_id            17114 non-null  object
 9   white_rating        17114 non-null  int64
 10  black_id            17114 non-null  object
 11  black_rating        17114 non-null  int64
 12  moves               17114 non-null  object
 13  opening_eco         17114 non-null  object
 14  opening_name        17114 non-null  object
 15  opening_ply         17114 non-null  int64
 16  game_time_duration  17114 non-null  float64
dtypes: bool(1), float64(3), int64(4), object(9)
memory usage: 2.2+ MB
None
'=================================================='
'No duplicate rows found.'
'=================================================='
```



```
'=================================================='
'Missing values:'
id                    0
rated                 0
created_at            0
last_move_at          0
turns                 0
victory_status        0
winner                0
increment_code        0
white_id              0
white_rating          0
black_id              0
black_rating          0
moves                 0
opening_eco           0
opening_name          0
opening_ply           0
game_time_duration    0
dtype: int64
'=================================================='
```

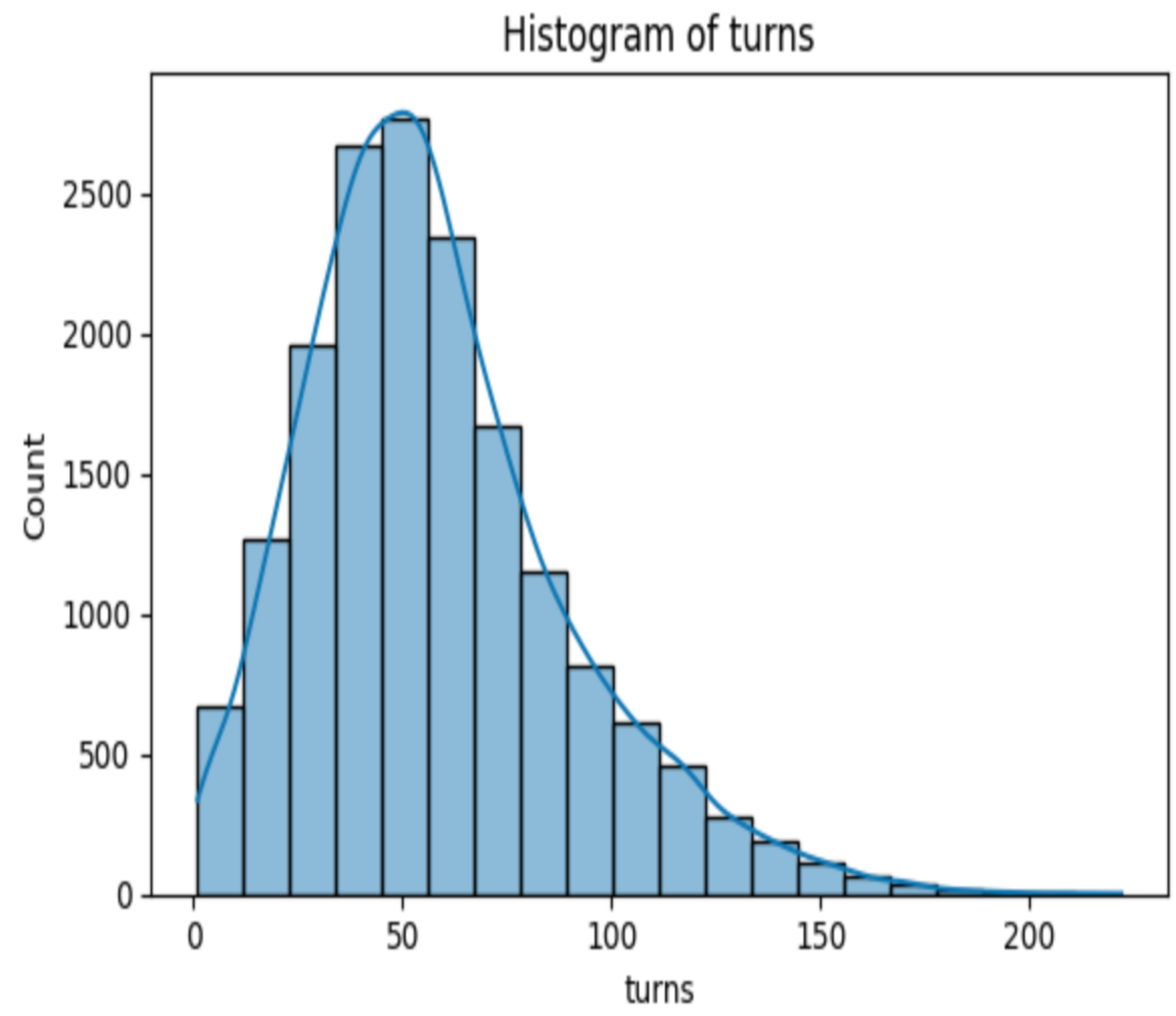
It was found that there were no missing or null values in the dataset, which contained over 20,000 records.

EXPLORATORY DATA ANALYSIS

The progression of moves made by the participants during a game is commonly referred to as "Turns." An analysis of these turns reveals that they follow a Gaussian distribution with a positive skew, which is indicative of a normal data distribution.

This observation can be attributed to the inclusion of various substantial datasets.

Number of turns less than 6:
#284, which makes up 2.0% of all games



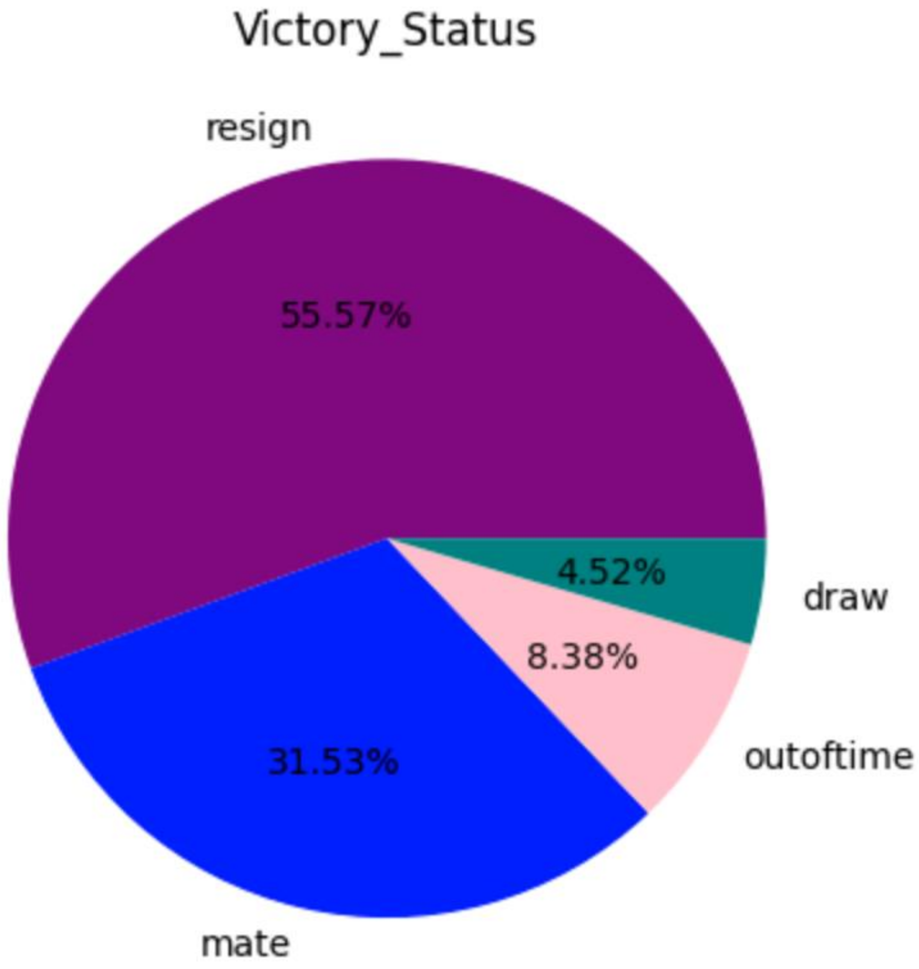
VICTORY_STATUS PIE PLOT

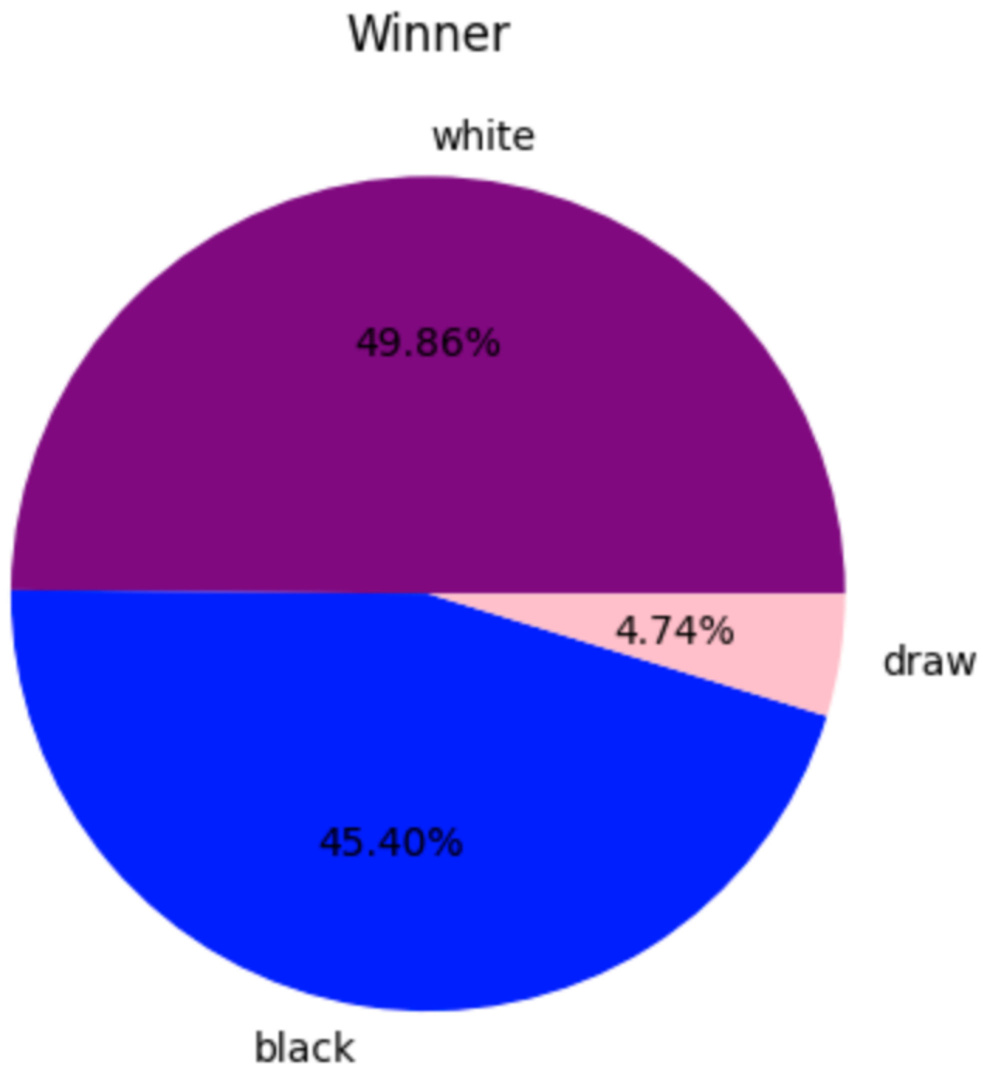
Four primary modalities can determine the outcome of a chess game: expiry of time on the chess clock, voluntary forfeit by resignation, checkmate, or draw.

The "victory_status" attribute in the data provides this information for each recorded match.

An analysis of this attribute has been performed, resulting in the pie plot, which ranks the percentage of each outcome.

The pie plot indicates that the most common method of ending the game was resignation.





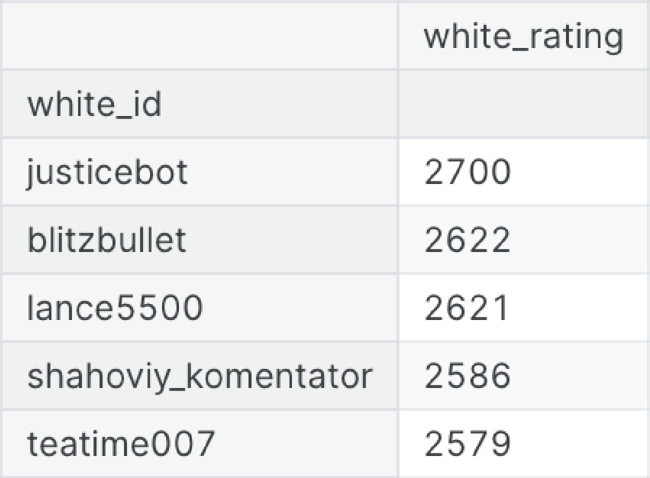
WINNER PIE PLOT

The winner attribute of the data set gives information on the color of a winning player piece.

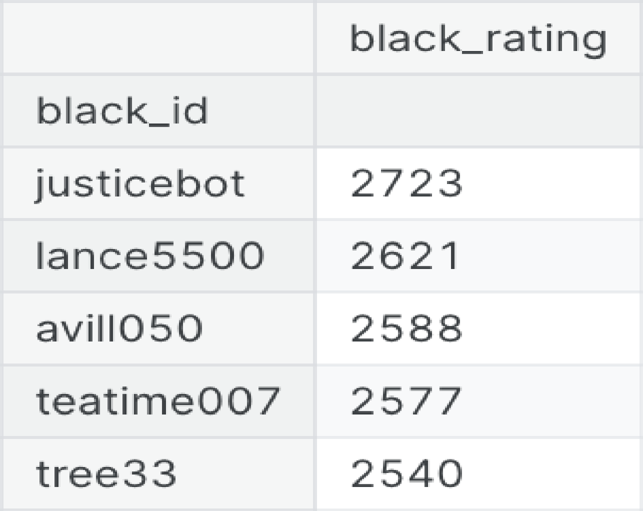
The analysis of the pie chart shows that 49.86% of wins are accounted for by players who use white pieces.

It seems that starting the game with white pieces gives a slight advantage.

HIGHEST RATED PLAYER



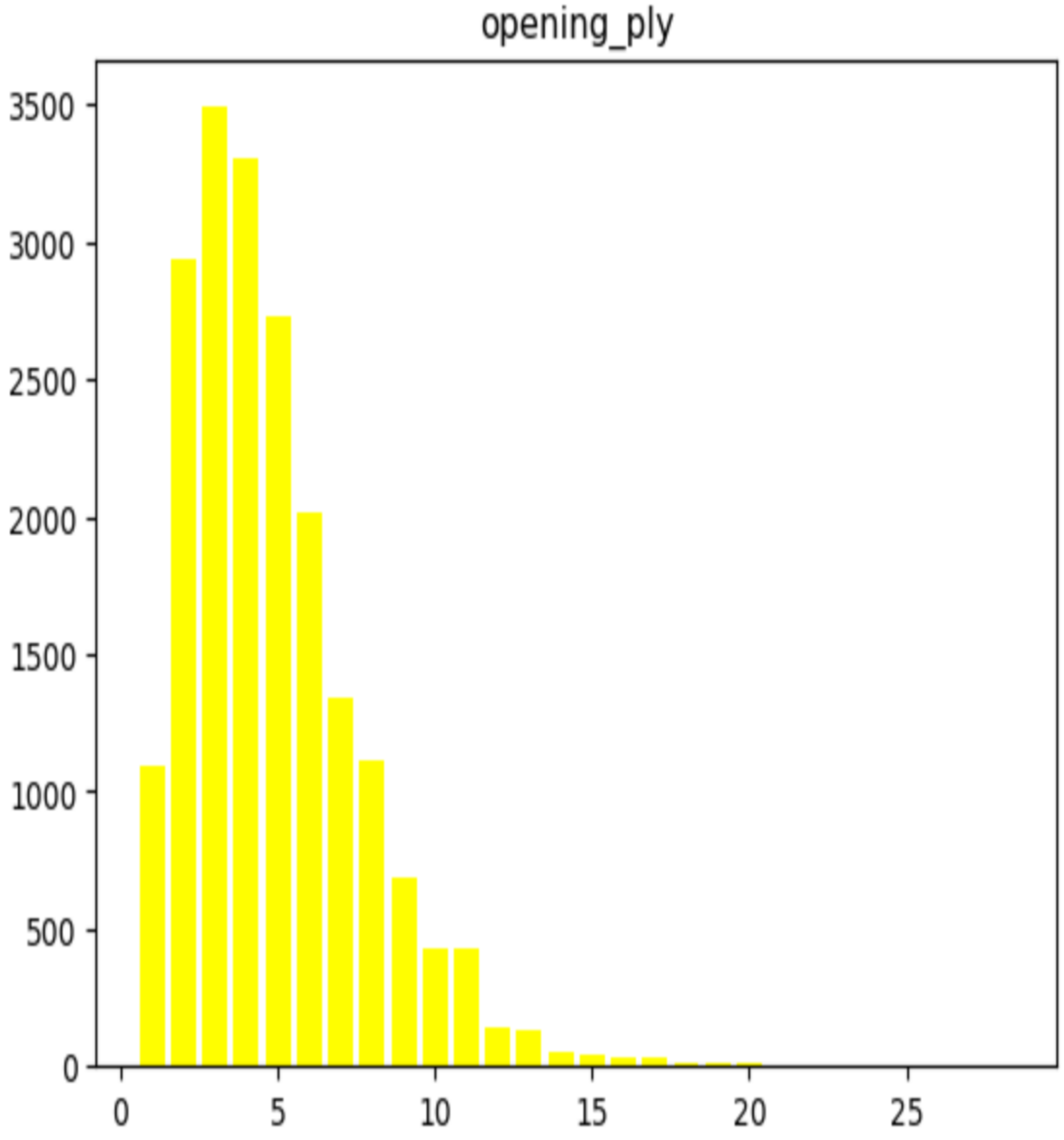
	white_rating
white_id	
justicebot	2700
blitzbullet	2622
lance5500	2621
shahoviy_komentator	2586
teatime007	2579



	black_rating
black_id	
justicebot	2723
lance5500	2621
avill050	2588
teatime007	2577
tree33	2540

- The most highly recorded rating is associated with the player "justicebot," who competes as both white and black and has ratings over 2700 in both.

- The data shows that all the best players had ratings above 2500, suggesting a high level of talent.

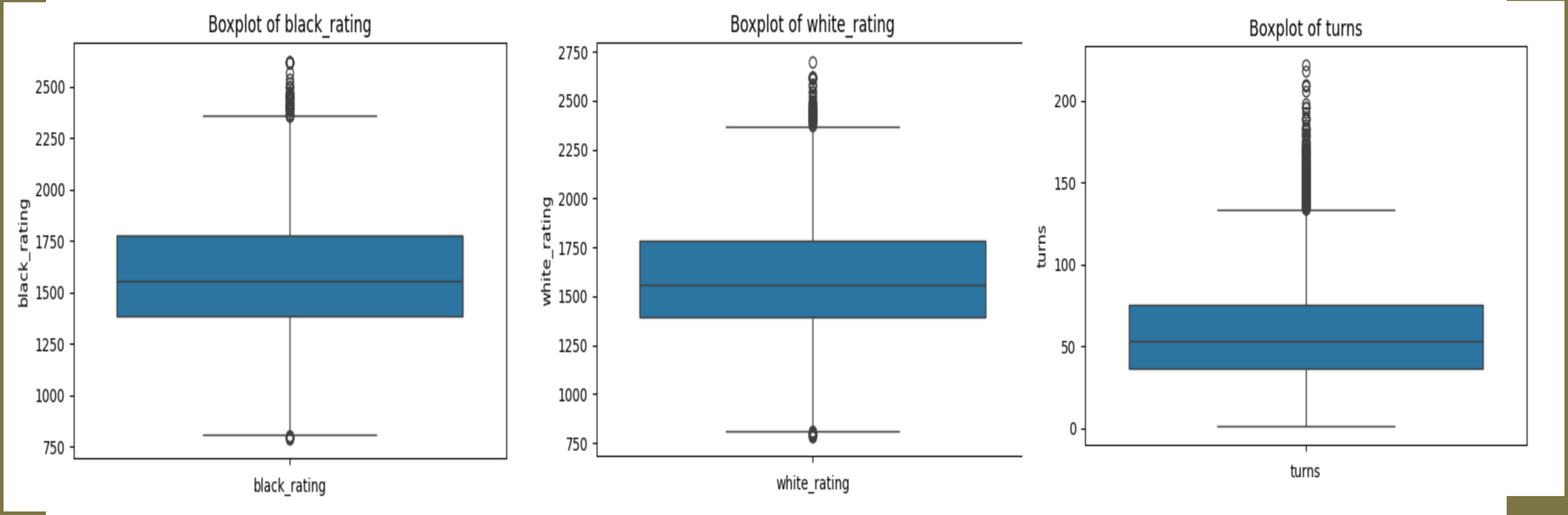


OPENING PLAY

In chess, opening play refers to the two participants performing a half-move or 1 complete move.

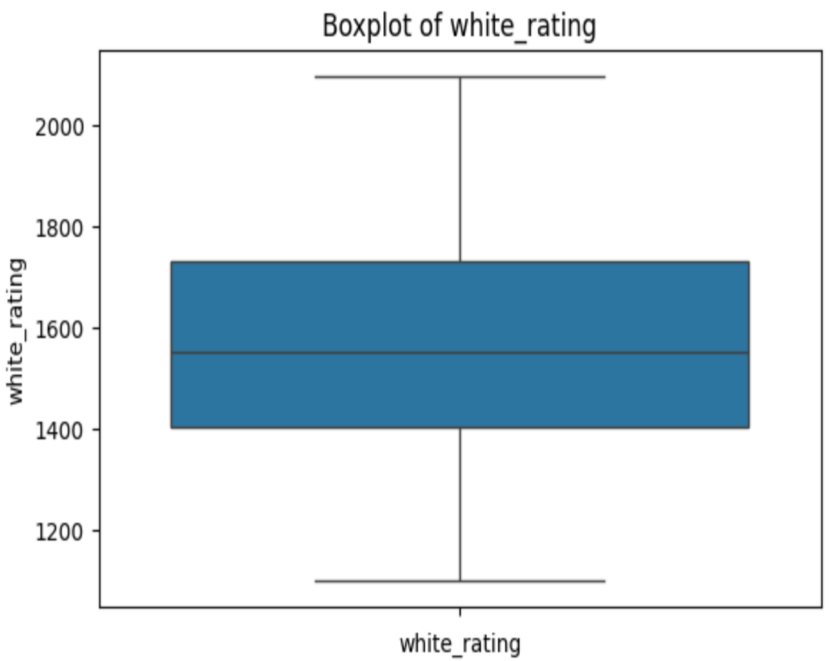
"opening_ply = 3" implies that three complete moves were executed, resulting in six half-moves, combined with White and Black each completing one move.

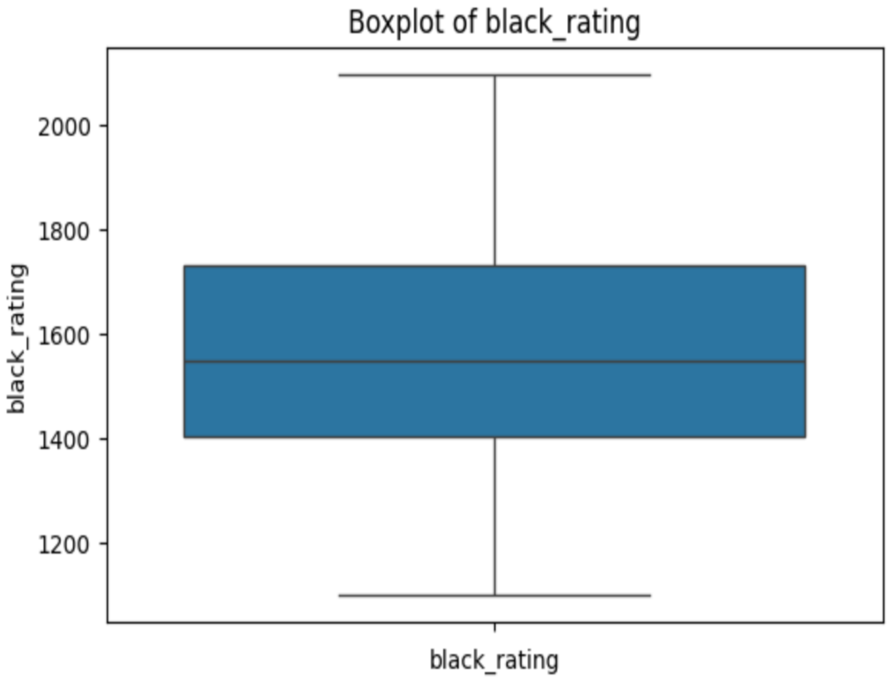
The bar plot demonstrates how the three early movements are the most generally applied.

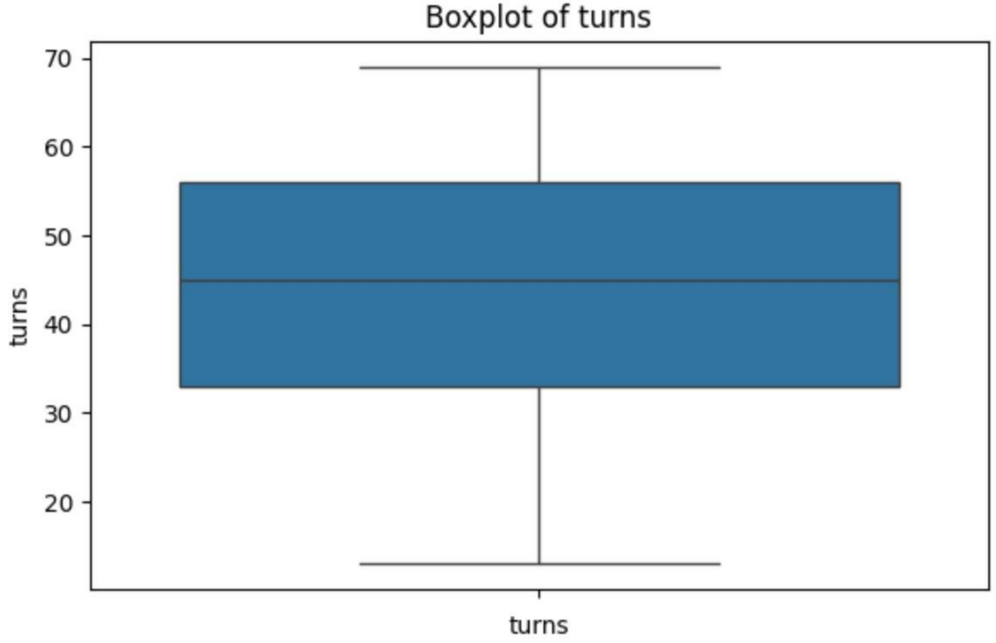


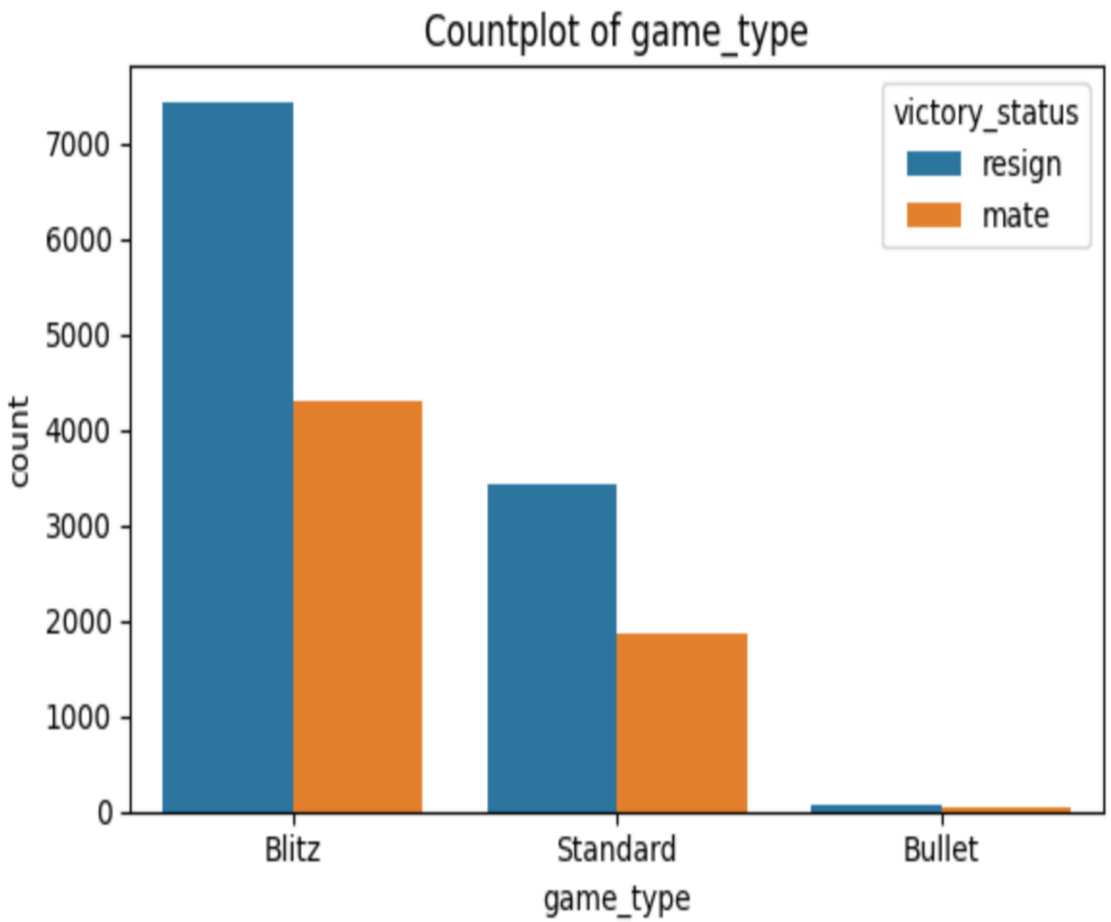
SINCE THERE ARE EXTREME VALUES, WE FILTER DATA USING A FIXED WINDOW.

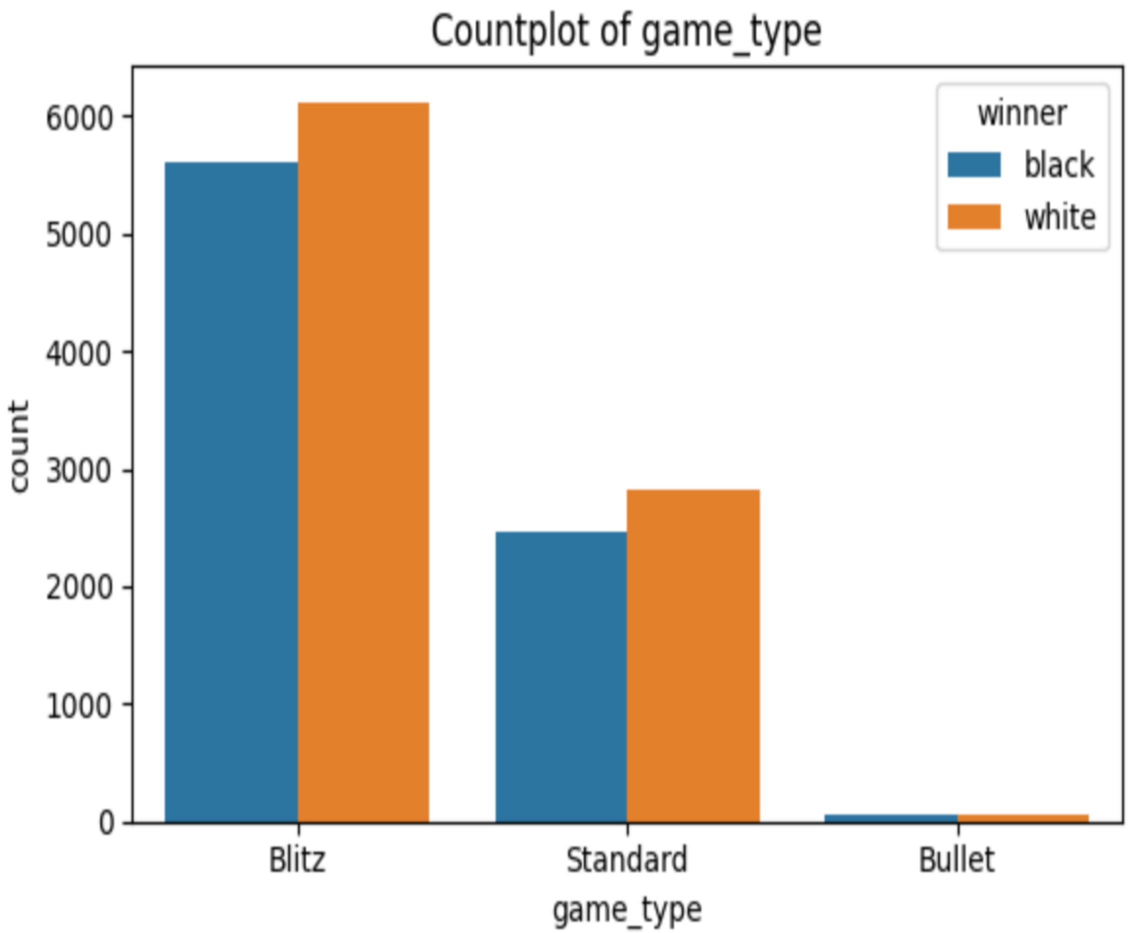
AFTER FILTERING



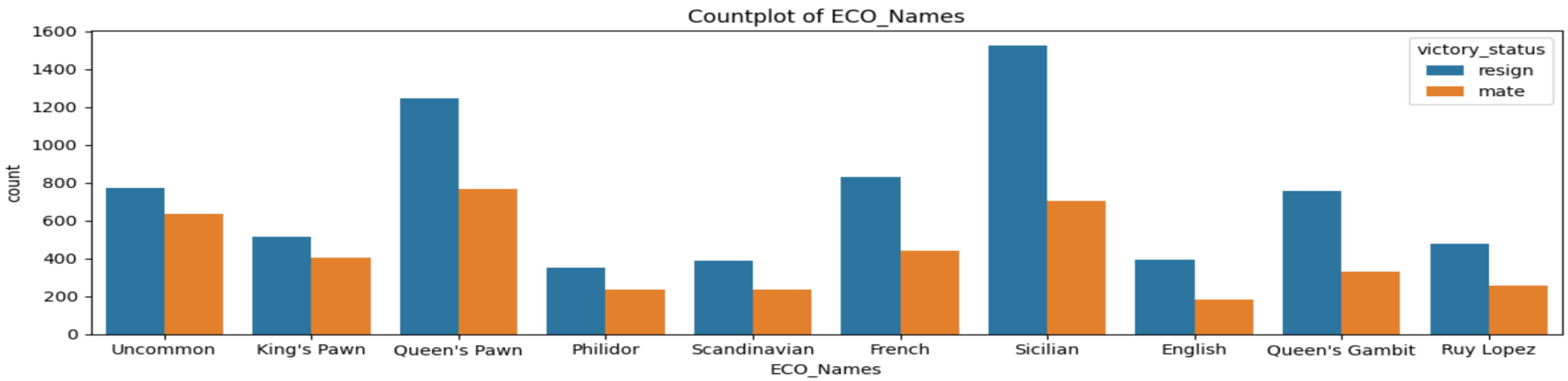




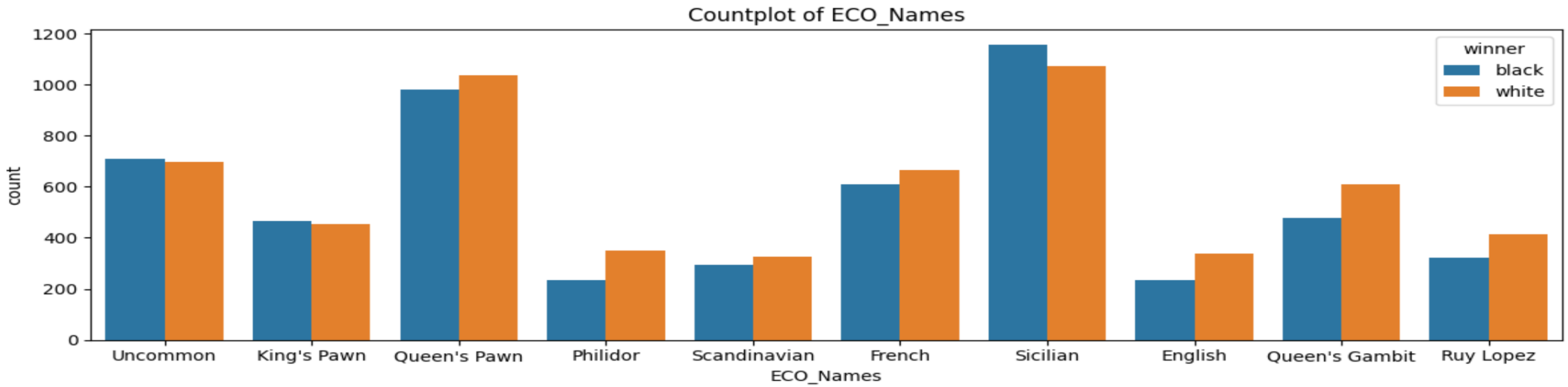




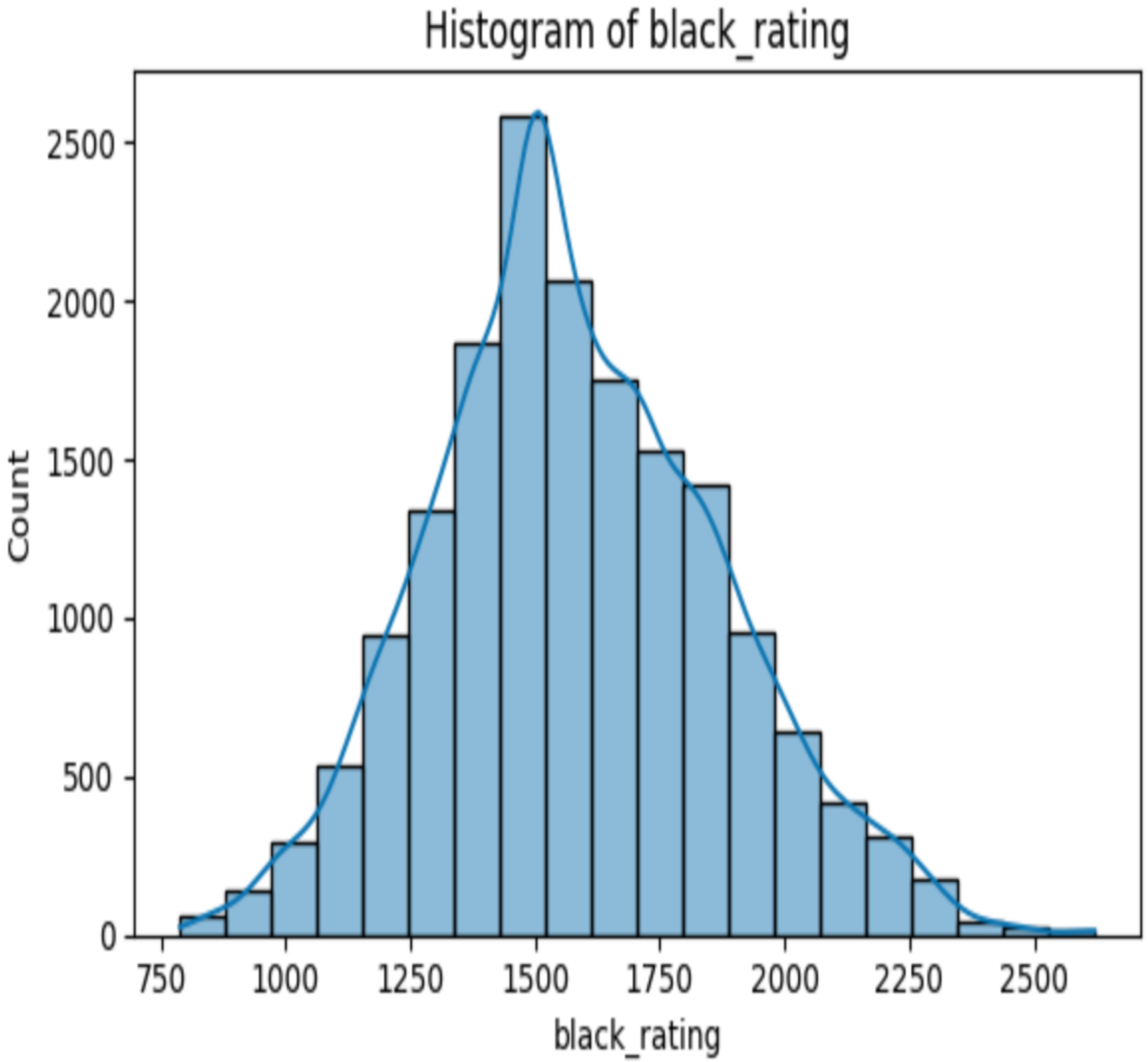
TYPES OF CHESS GAMES

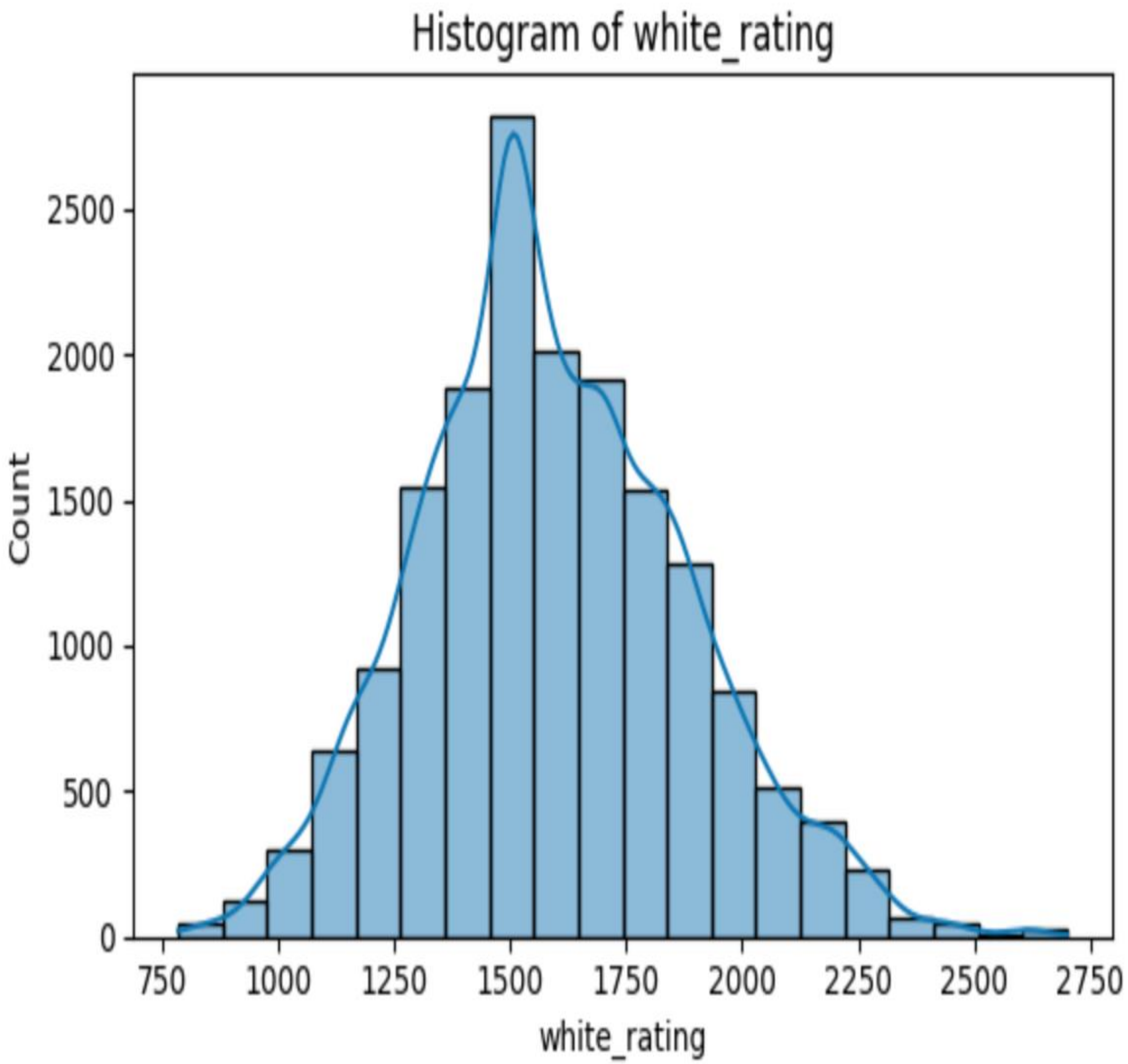


Countplot of ECO_Names

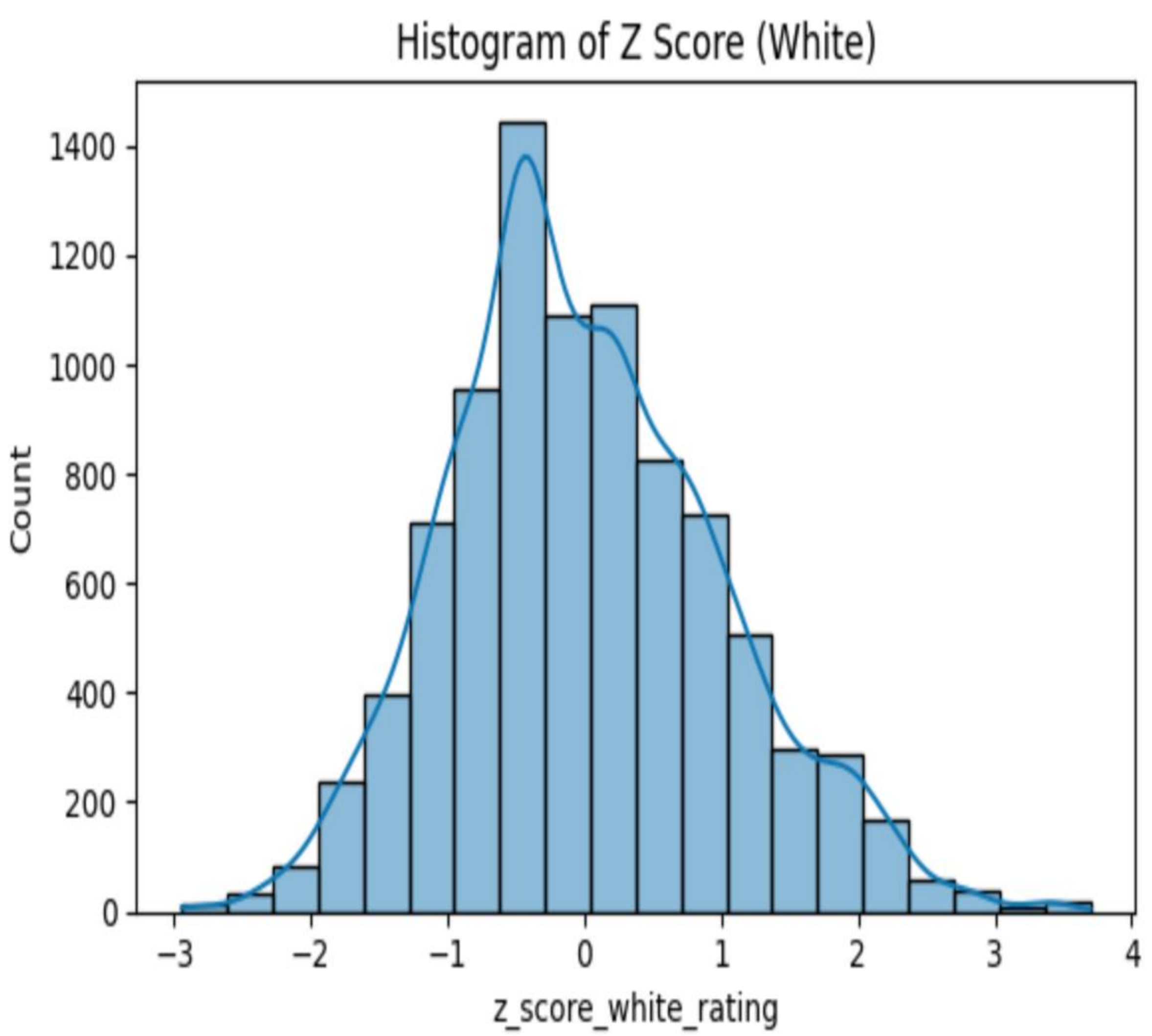


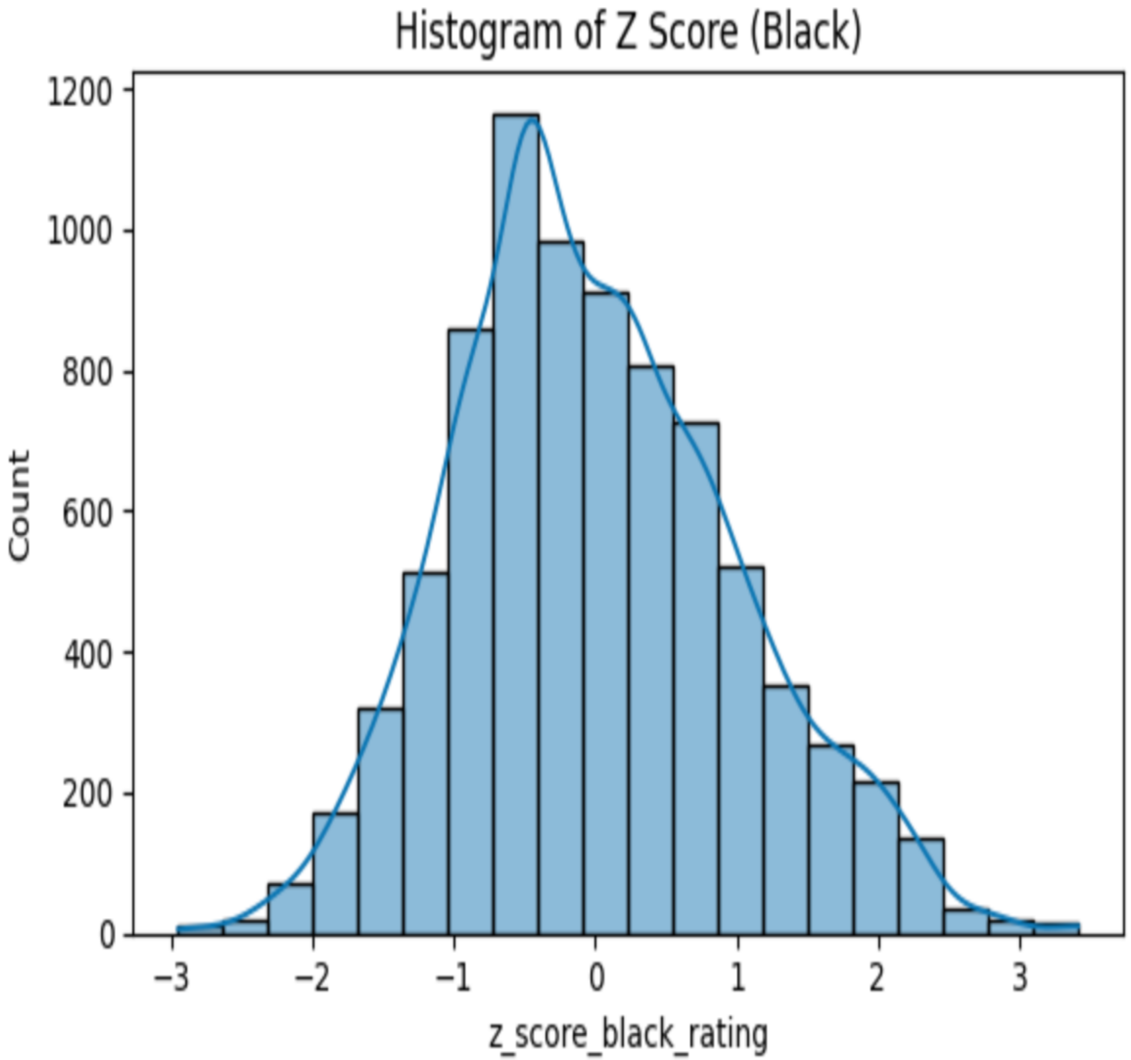
Countplot of ECO_Names





The chess rating is a numerical representation of the player's performance in comparison to others.
It's determined on the basis of a player's performance.





If the score is below -0.689, it's a "Low Rating". If the score is above 0.654, it's a "High Rating". If the score falls between those values, it's a "Mid Rating".

TRAINING AND ASSESSING THE MODELS

Decision trees,

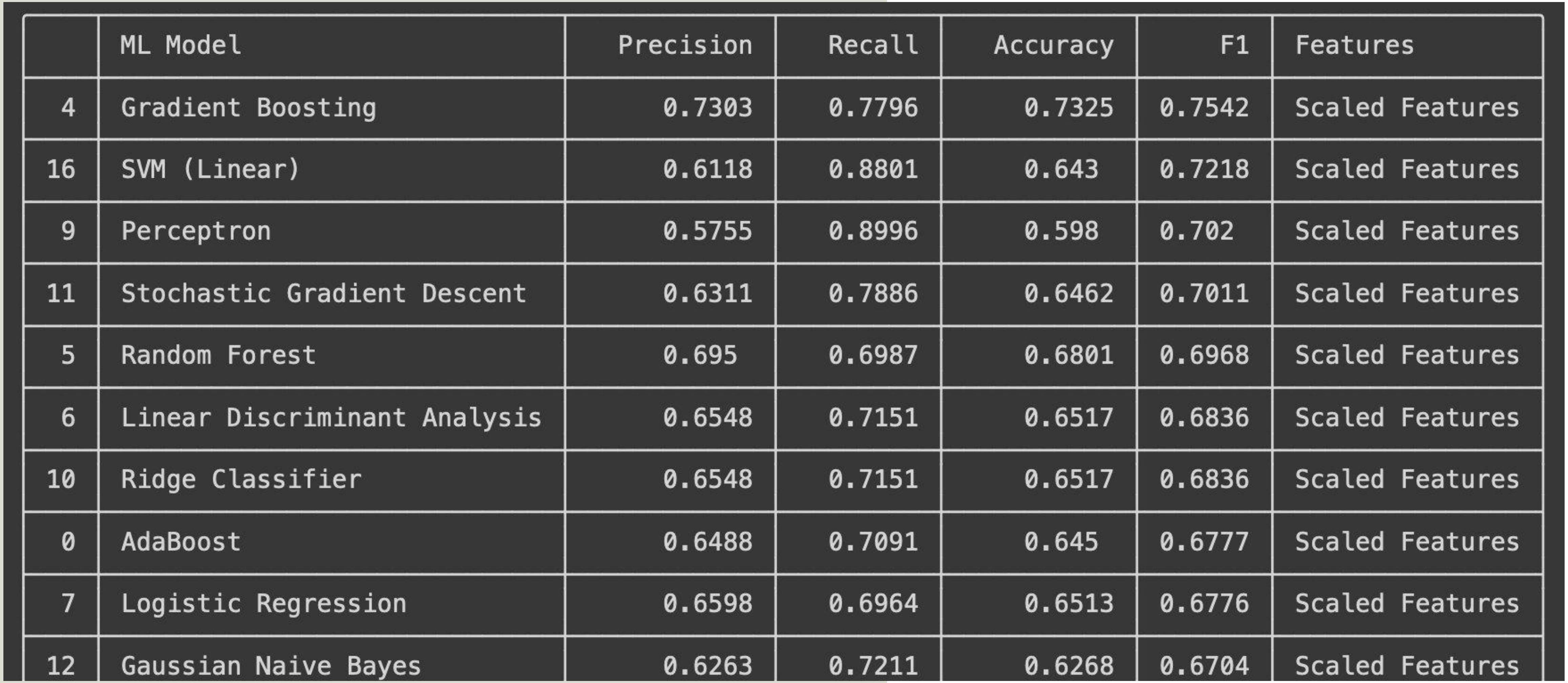
Random forests,

Gradient Boosting

Bagging

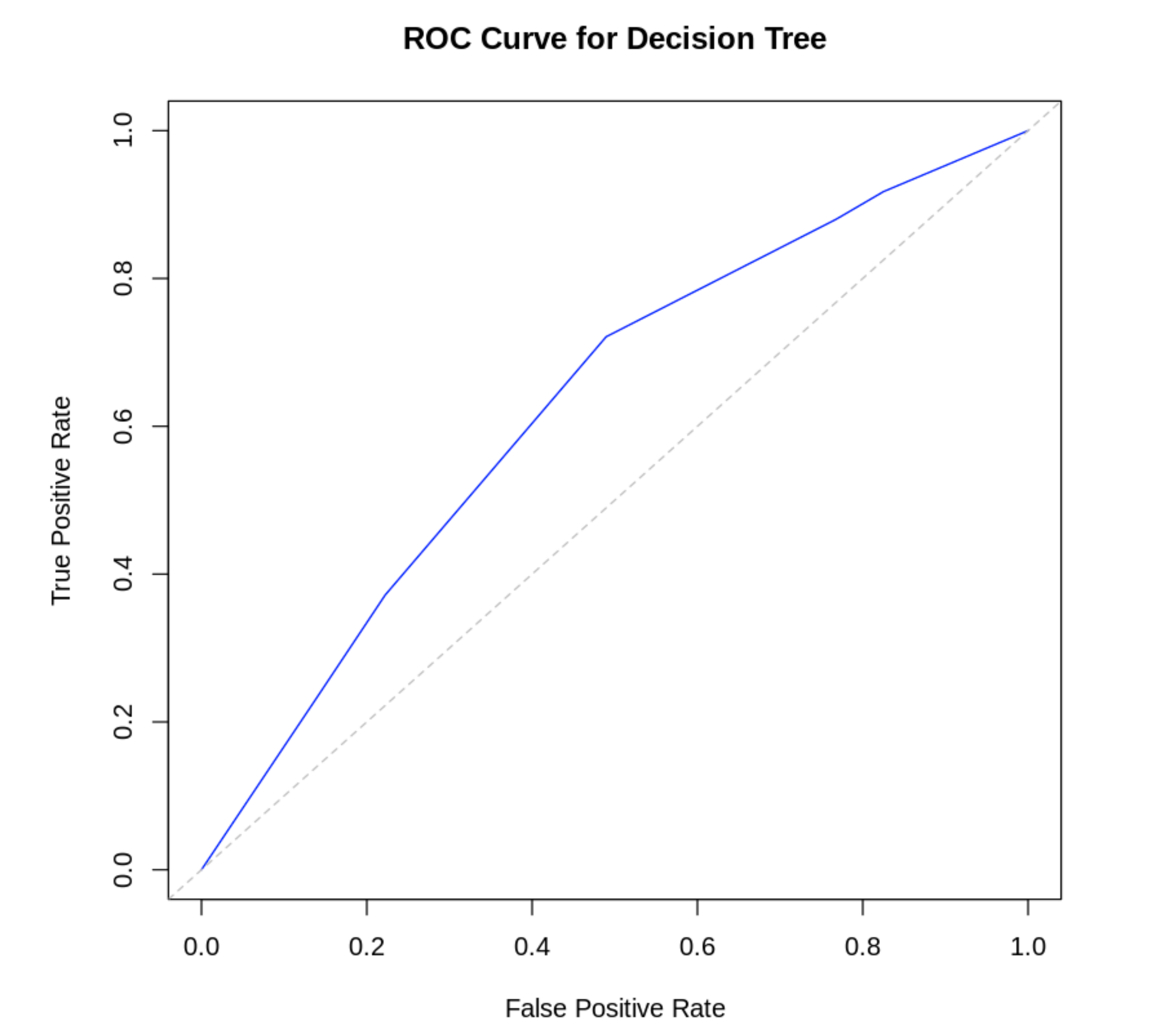
Performance evaluation of models is done using metrics such as accuracy, precision, recall and F1-score.

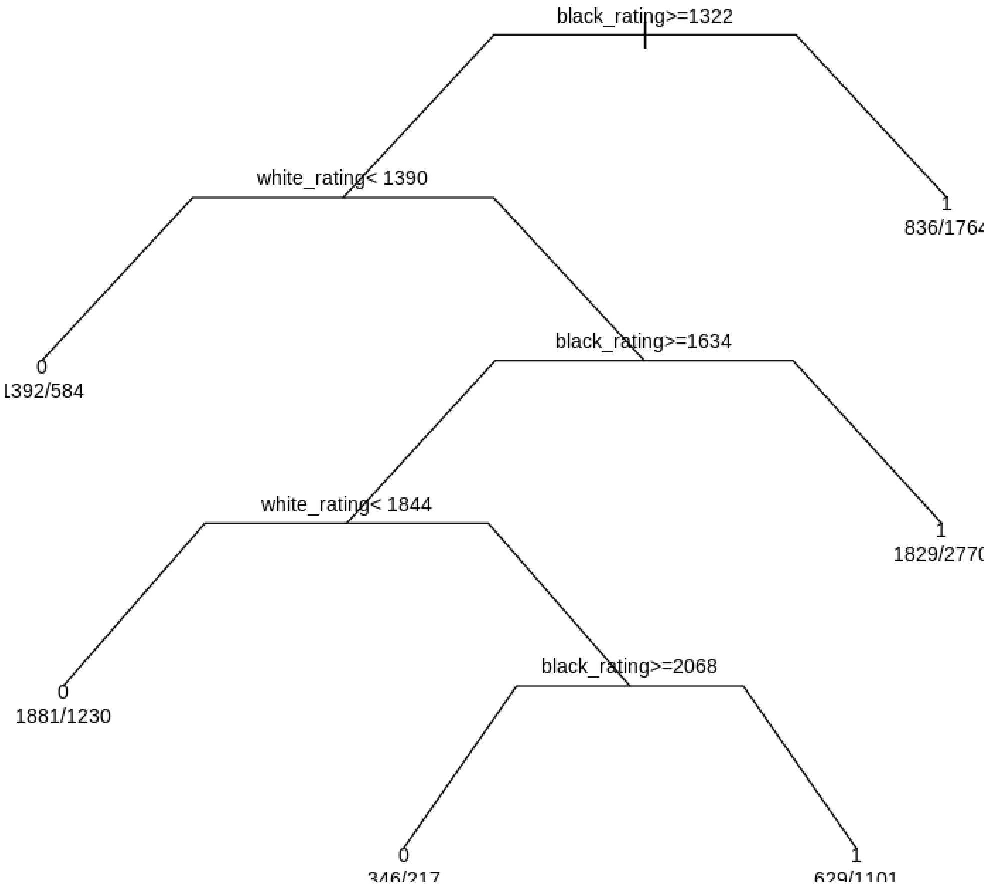
Results

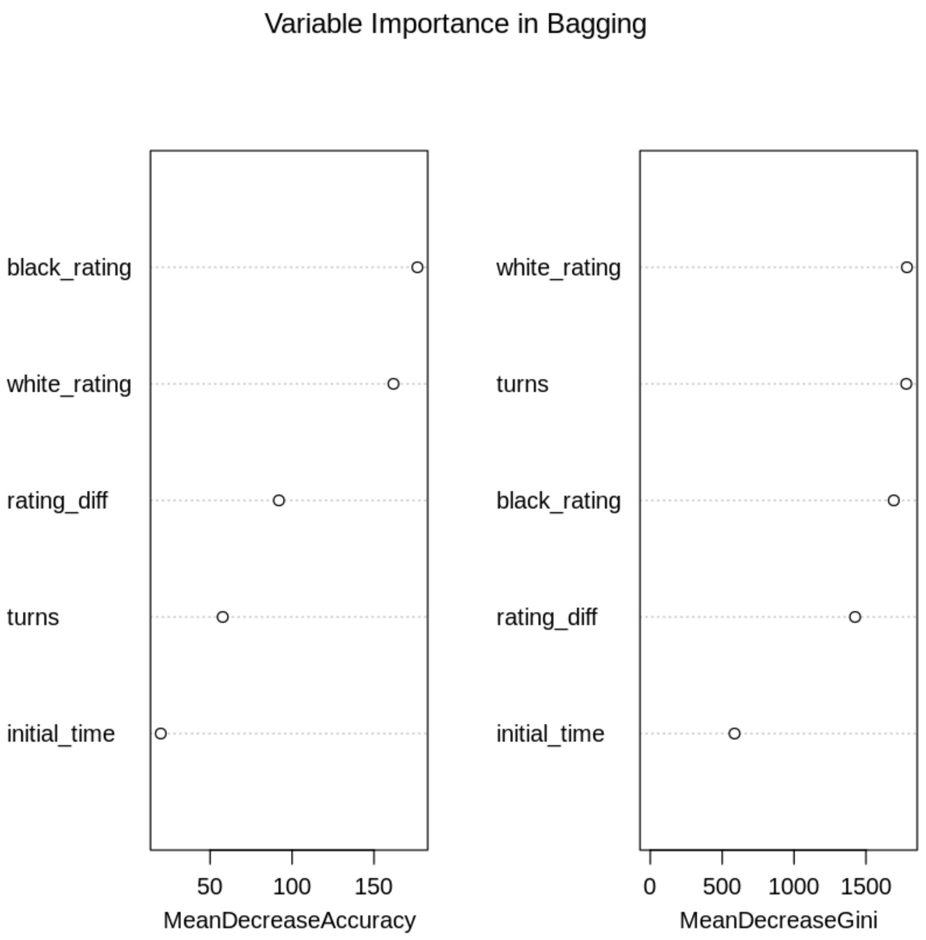


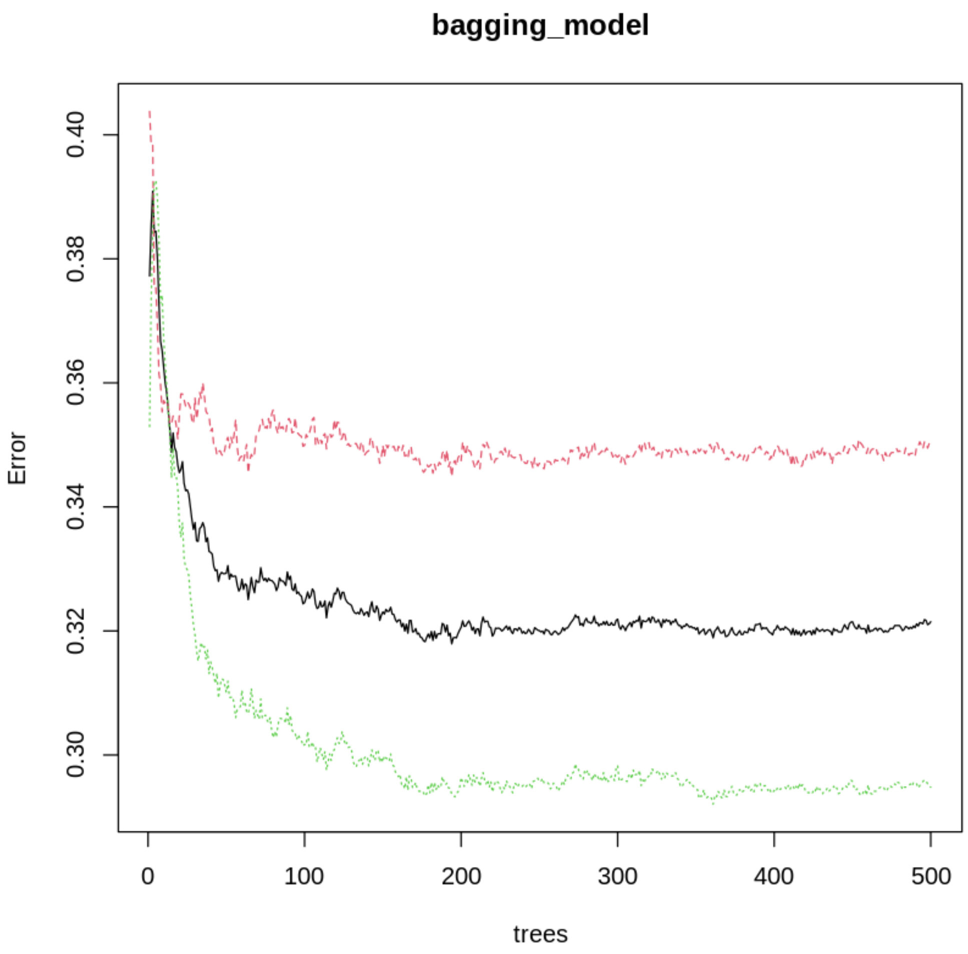
	ML Model	Precision	Recall	Accuracy	F1	Features
4	Gradient Boosting	0.7303	0.7796	0.7325	0.7542	Scaled Features
16	SVM (Linear)	0.6118	0.8801	0.643	0.7218	Scaled Features
9	Perceptron	0.5755	0.8996	0.598	0.702	Scaled Features
11	Stochastic Gradient Descent	0.6311	0.7886	0.6462	0.7011	Scaled Features
5	Random Forest	0.695	0.6987	0.6801	0.6968	Scaled Features
6	Linear Discriminant Analysis	0.6548	0.7151	0.6517	0.6836	Scaled Features
10	Ridge Classifier	0.6548	0.7151	0.6517	0.6836	Scaled Features
0	AdaBoost	0.6488	0.7091	0.645	0.6777	Scaled Features
7	Logistic Regression	0.6598	0.6964	0.6513	0.6776	Scaled Features
12	Gaussian Naive Bayes	0.6263	0.7211	0.6268	0.6704	Scaled Features

TREE BASED MODEL



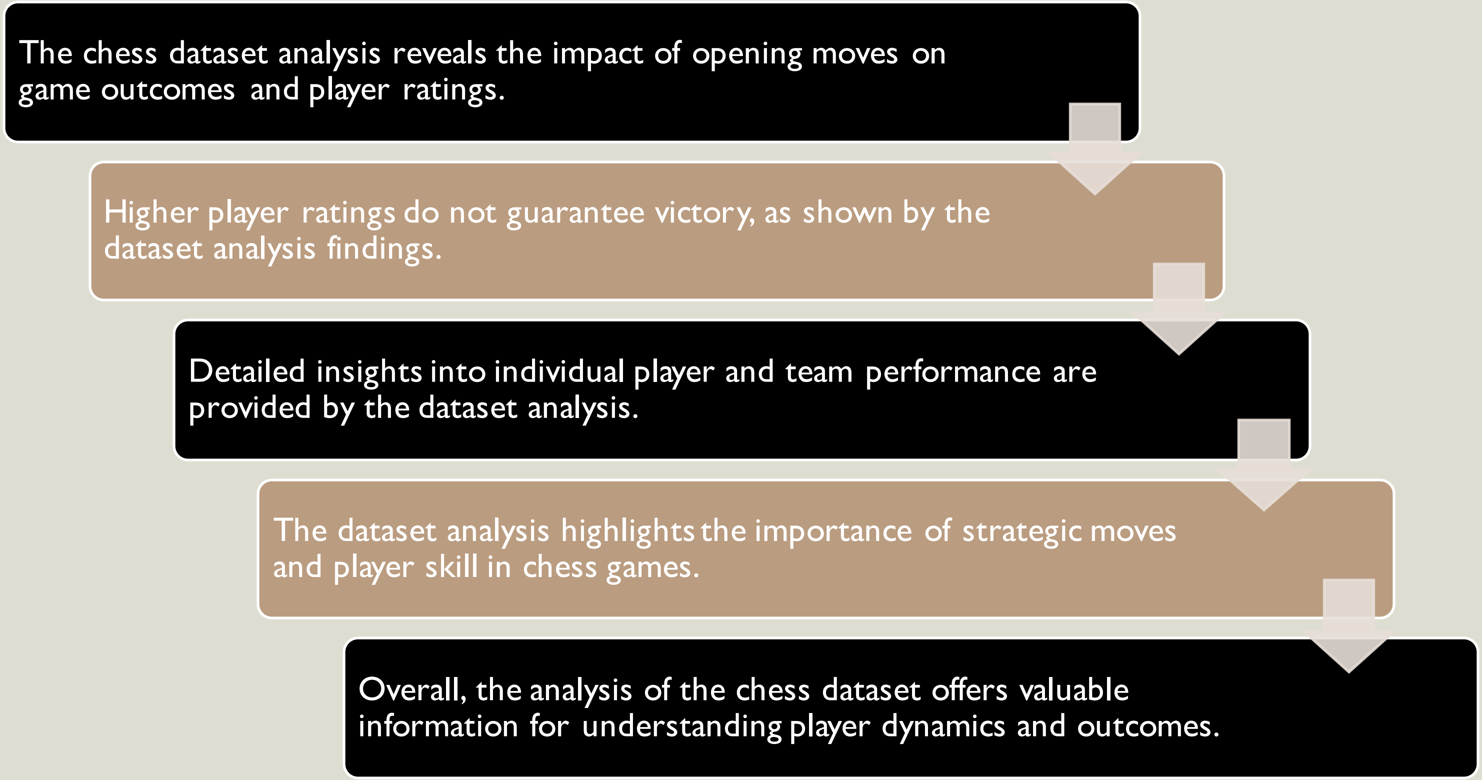






STATISTICAL ANALYSIS IN R

CONCLUSION



The chess dataset analysis reveals the impact of opening moves on game outcomes and player ratings.

Higher player ratings do not guarantee victory, as shown by the dataset analysis findings.

Detailed insights into individual player and team performance are provided by the dataset analysis.

The dataset analysis highlights the importance of strategic moves and player skill in chess games.

Overall, the analysis of the chess dataset offers valuable information for understanding player dynamics and outcomes.

QUESTIONS

- What kind of features could be added in the dataset?
- Which metric should be prioritized Precision or Recall? Why?
- Where do you think this model can fit? Why?

THANK YOU