

## DATA ANALYSIS OF THE CHESS GAMES

### Purpose of the Research

In this study, it is aimed to determine the potential winning strategies of chess game by considering game attributes such as duration of the game, openings, type of game (rated or regular) and chess sides (black and white). Determination process includes over 20.000 online match histories of players at different levels of chess.

### Dataset: Chess Game Dataset

🔗 <https://www.kaggle.com/datasnaek/chess>

This dataset includes 20.000+ chess matches which are obtained from Lichess.org and it has 16 different columns as follows:

- Game ID *as id*
- Rated (T/F) *as rated*
- Start time *as created\_at*
- End time *as last\_move\_at*
- Number of turns *as turns*
- End game status *as victory\_status*
- Winner *as winner*
- Time increment<sup>1</sup> *as increment\_code*
- White player ID *as white\_id*
- White player Rating *as white\_rating*
- Black player ID *as black\_id*
- Black player Rating *as black\_rating*
- All moves in standard chess notation *as moves*
- Opening ECO<sup>2</sup> code *as opening\_eco*
- Opening name<sup>3</sup> *as opening\_name*
- Opening ply<sup>4</sup> *as opening\_ply*

1 - Time increment in a chess game is an amount of time added to the clock after each move is made.

2 - The Encyclopaedia of Chess Openings (ECO) is a classification system for the chess opening moves.

3 - Opening name is the title that is corresponding to the relevant ECO code.

4 - Ply is a half move and each turn consists 2 plies in a chess game. Opening ply refers to number of half moves that played in opening phase.

## Win Rates of Chess Sides (White vs. Black)

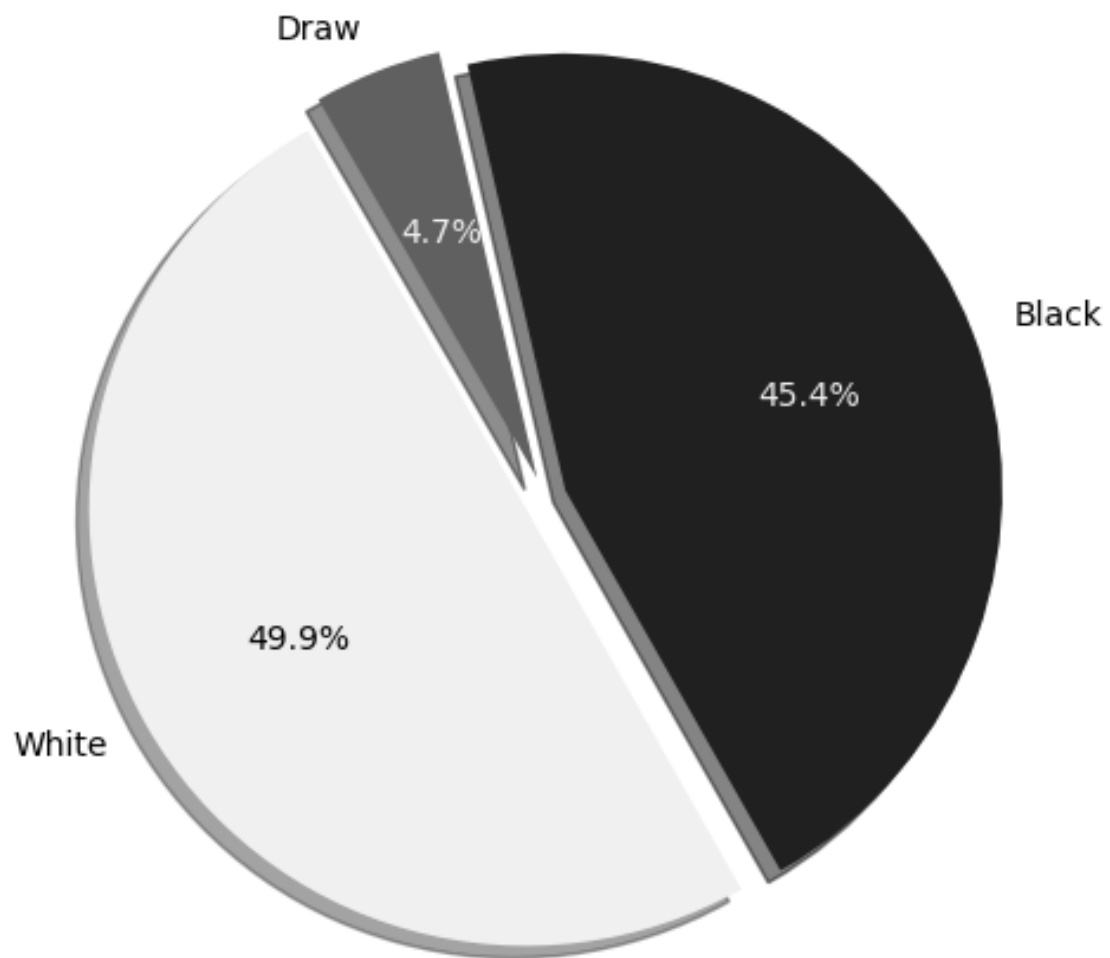
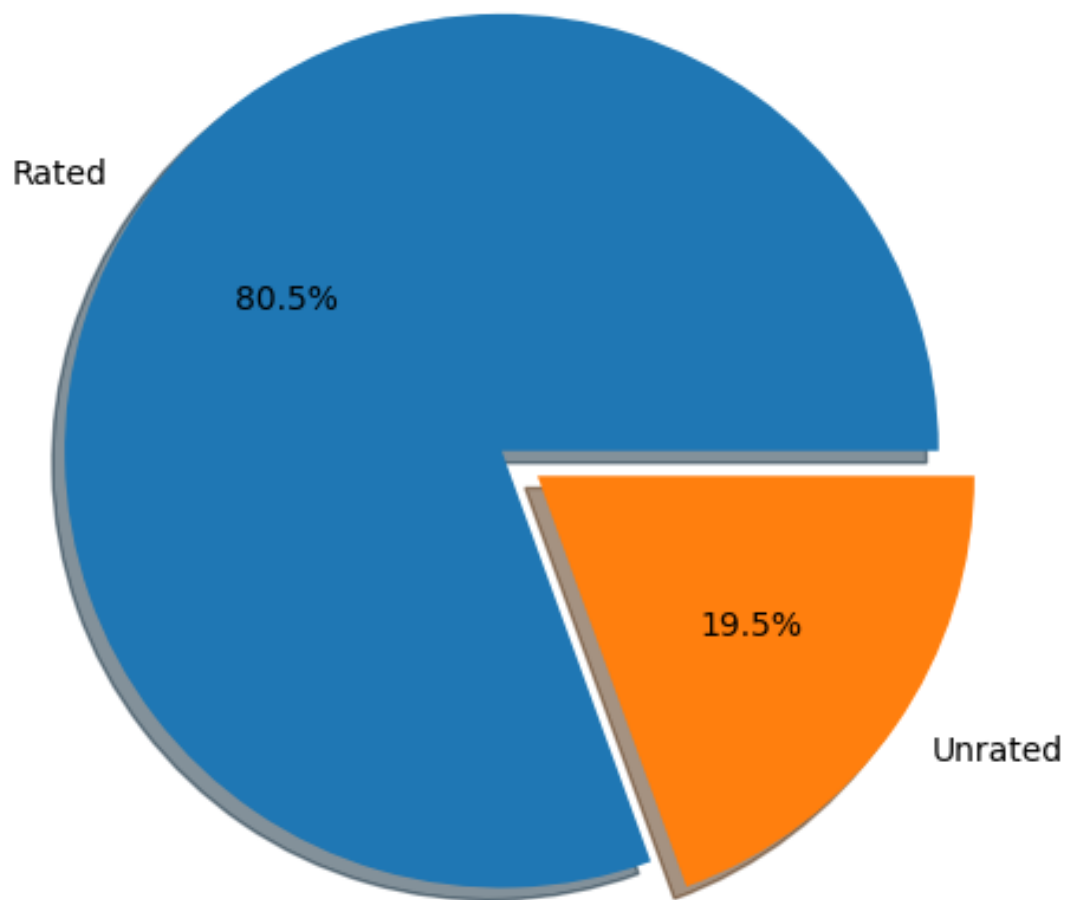


Figure 1 Win Rates of Chess Sides

## Game Mode Percentage



*Figure 2 Game Mode Percentage*

## End Status of Chess Games

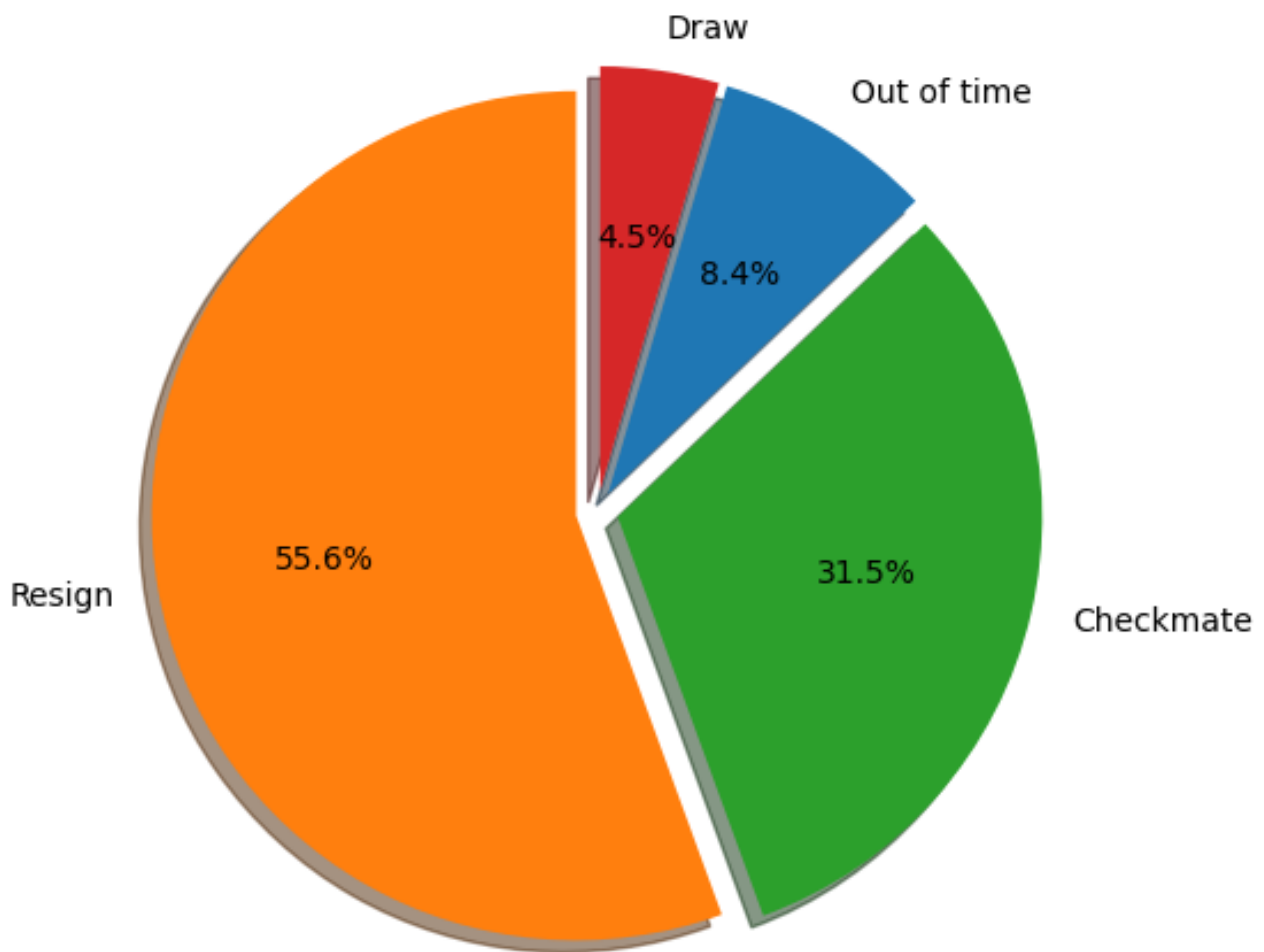


Figure 3 End Status of Chess Games

Disturbution of Chess Ratings

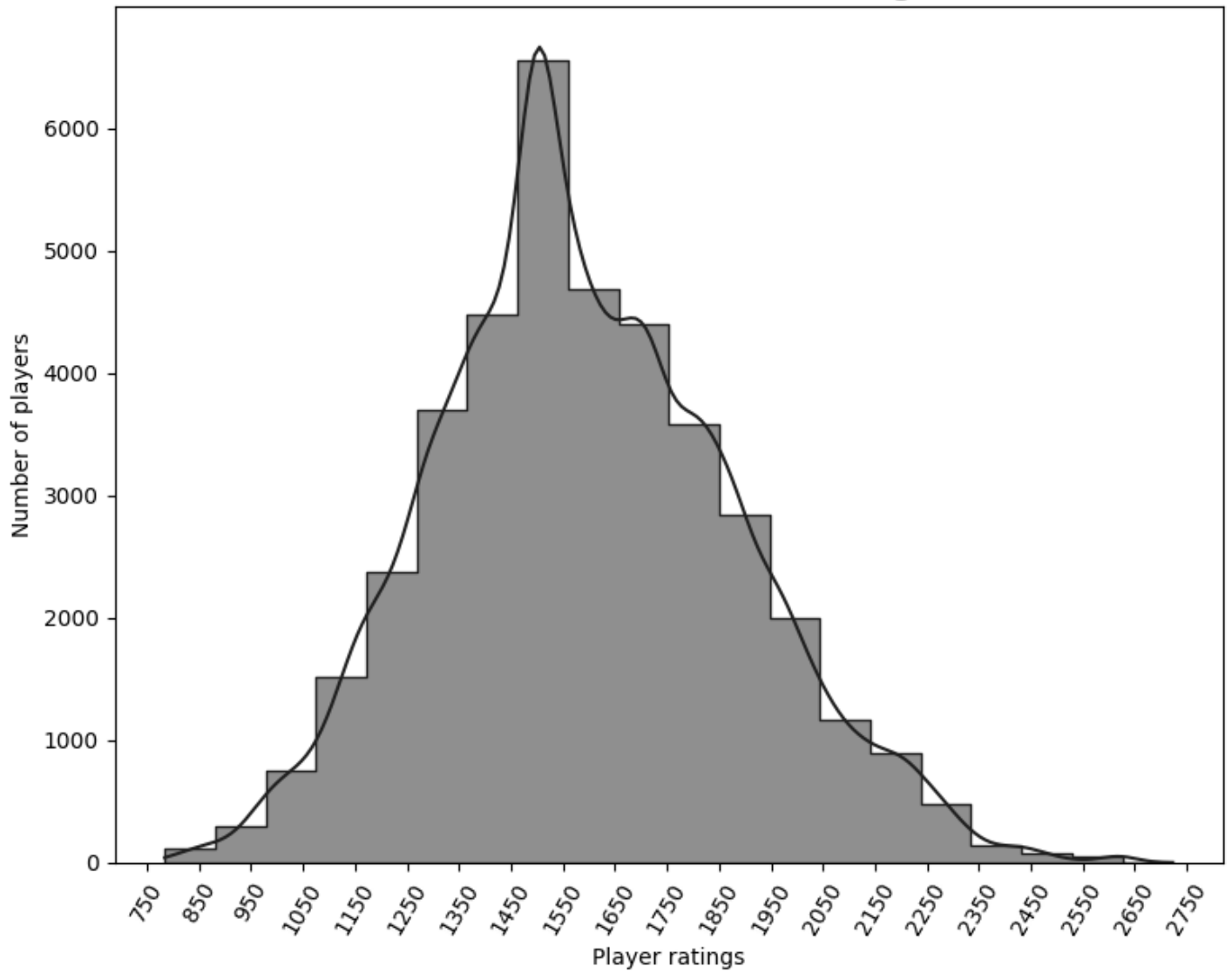


Figure 4 Disturbution of Chess Ratings

## Top 5 Most Preferred Chess Openings

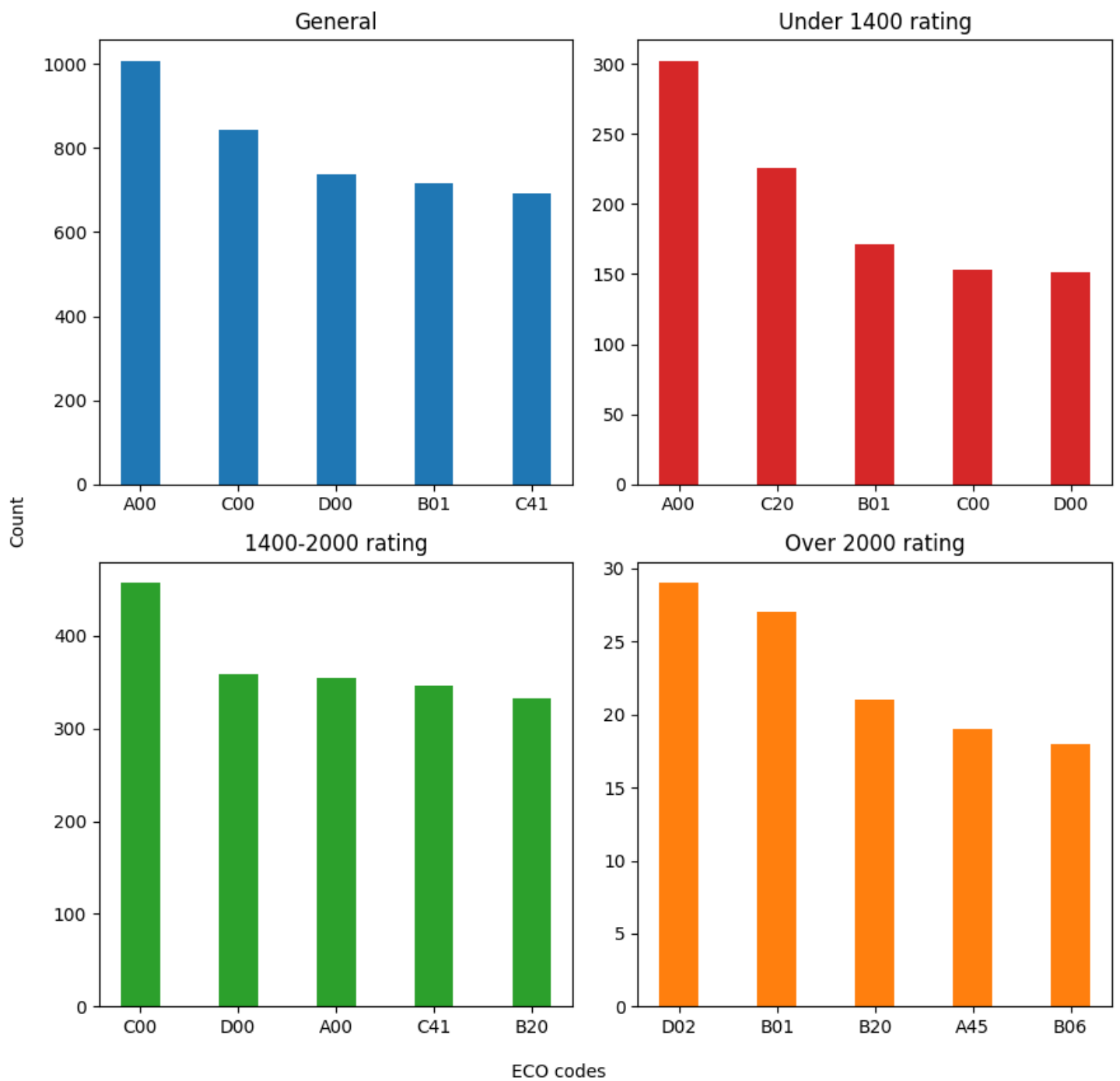


Figure 5 Top 5 Most Preferred Chess Openings

End Game Status & Turns

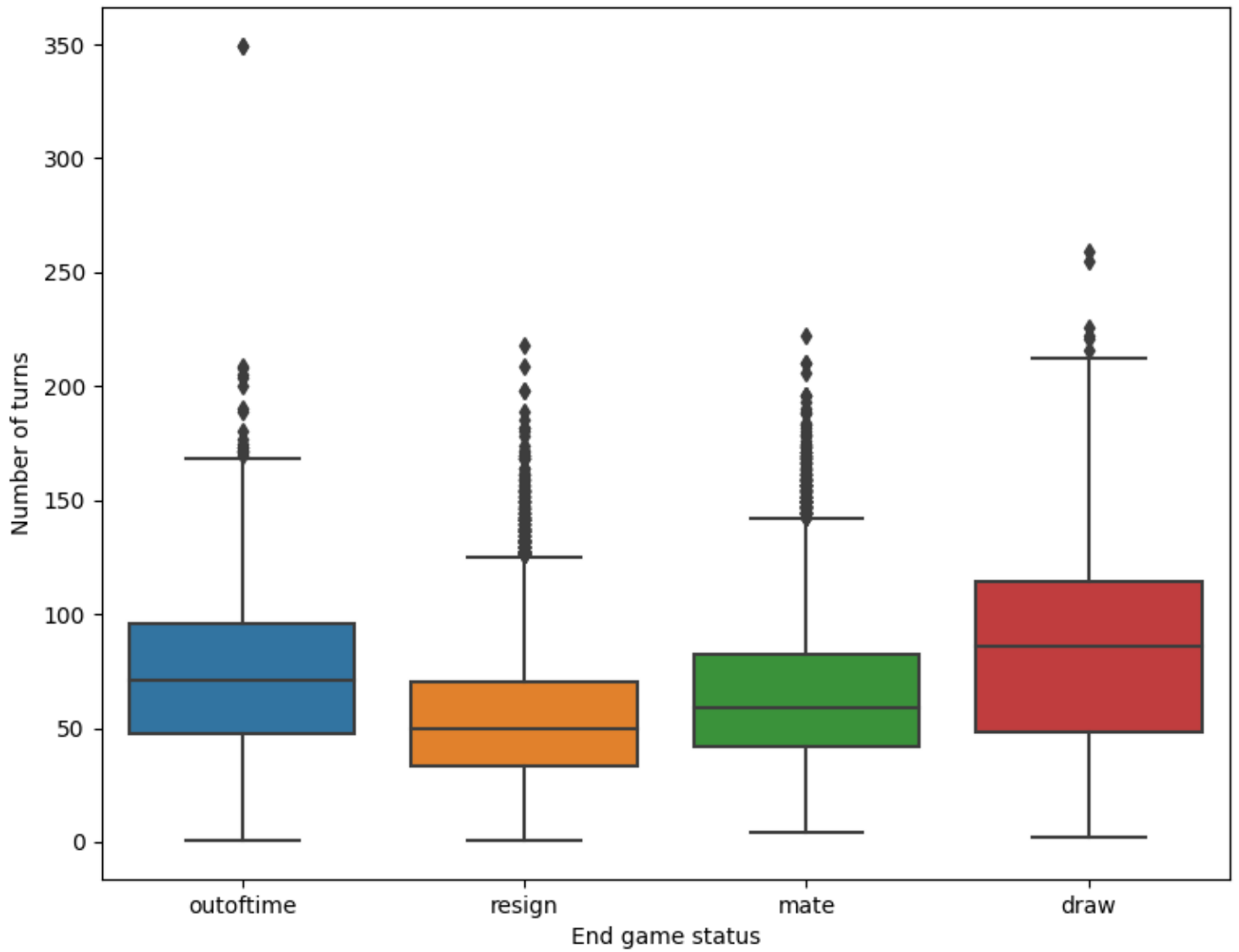


Figure 6 End Game Status & Turns

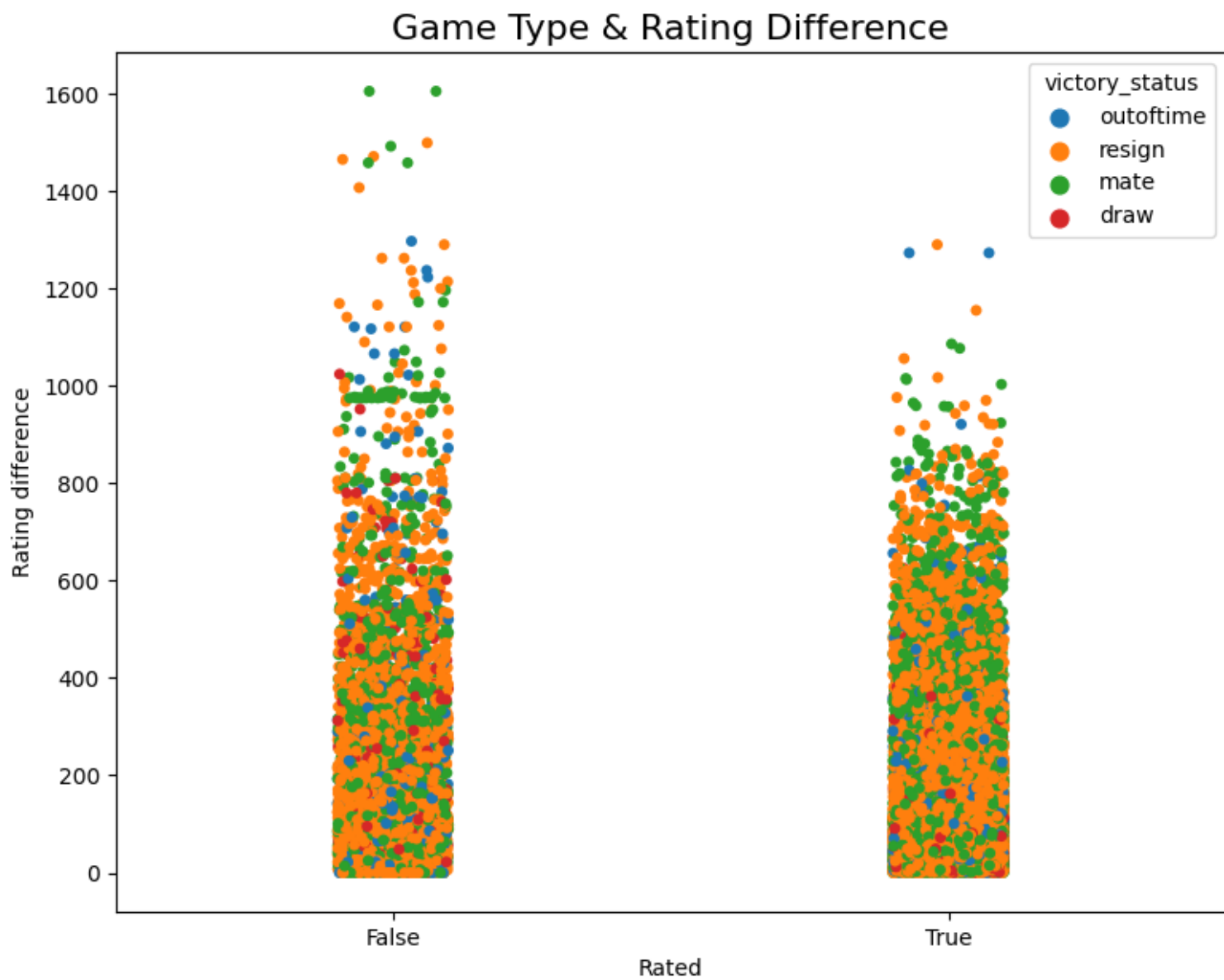


Figure 7 Game Type & Rating Difference



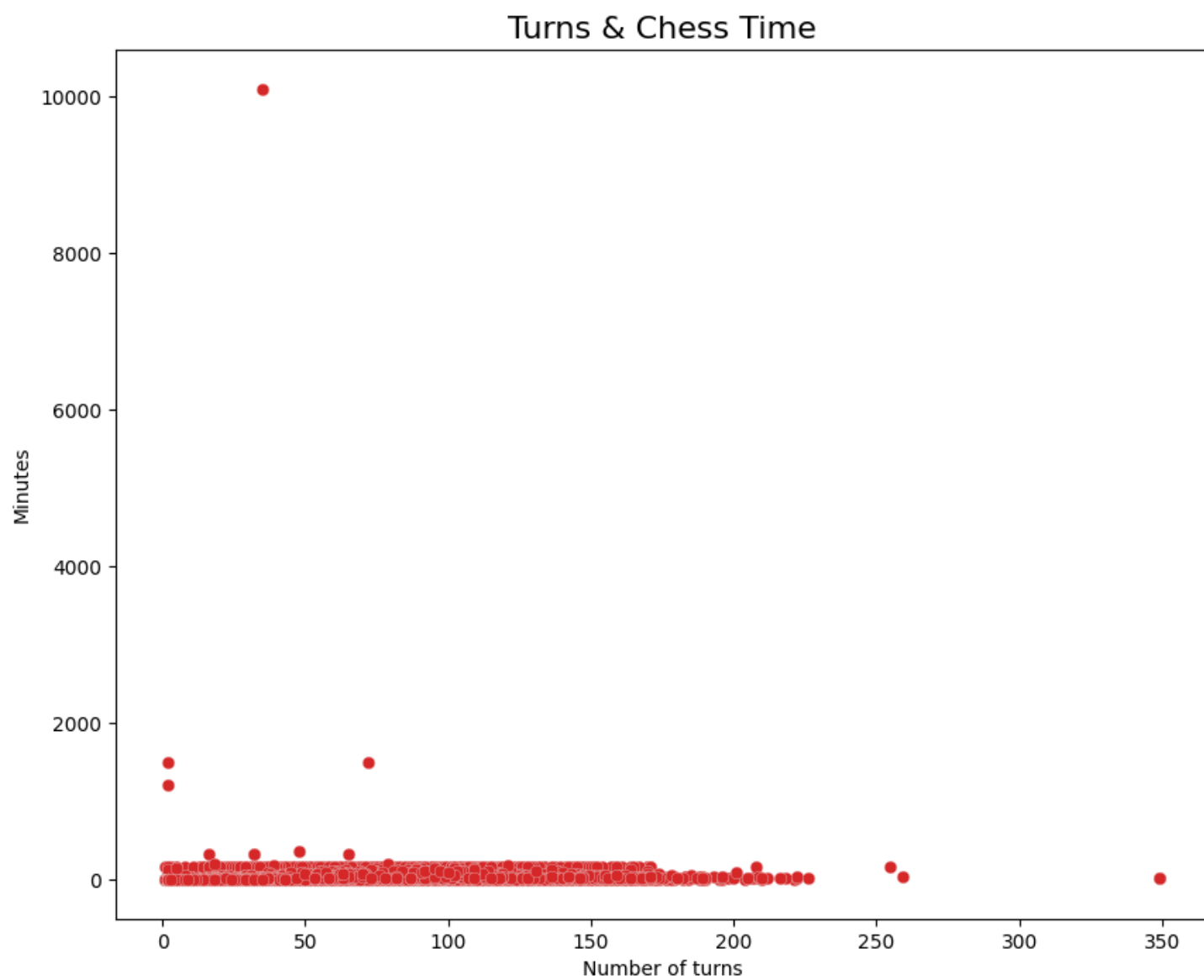


Figure 8 Turns & Chess Time

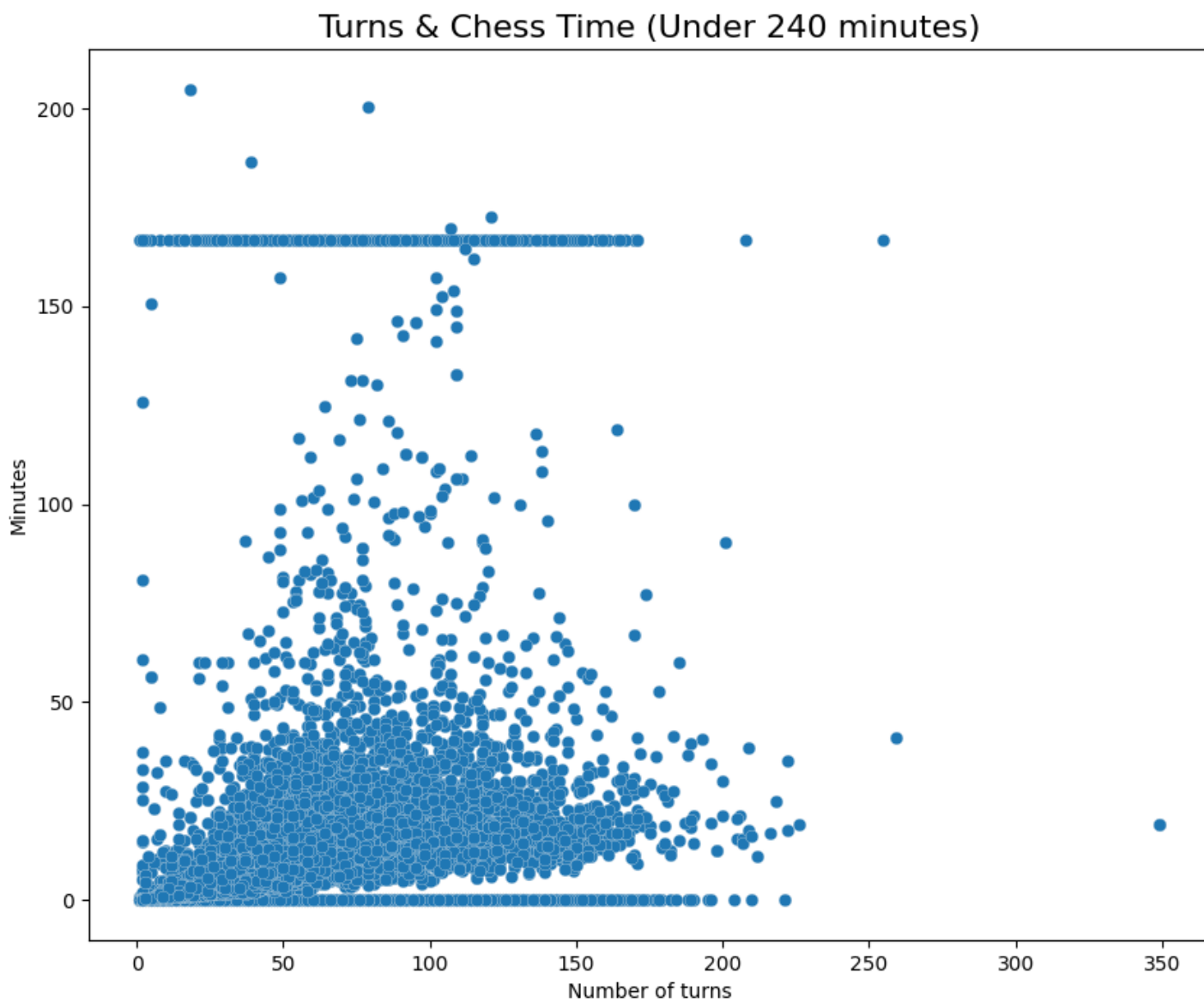


Figure 9 Turns & Chess Time (Under 240 minutes)

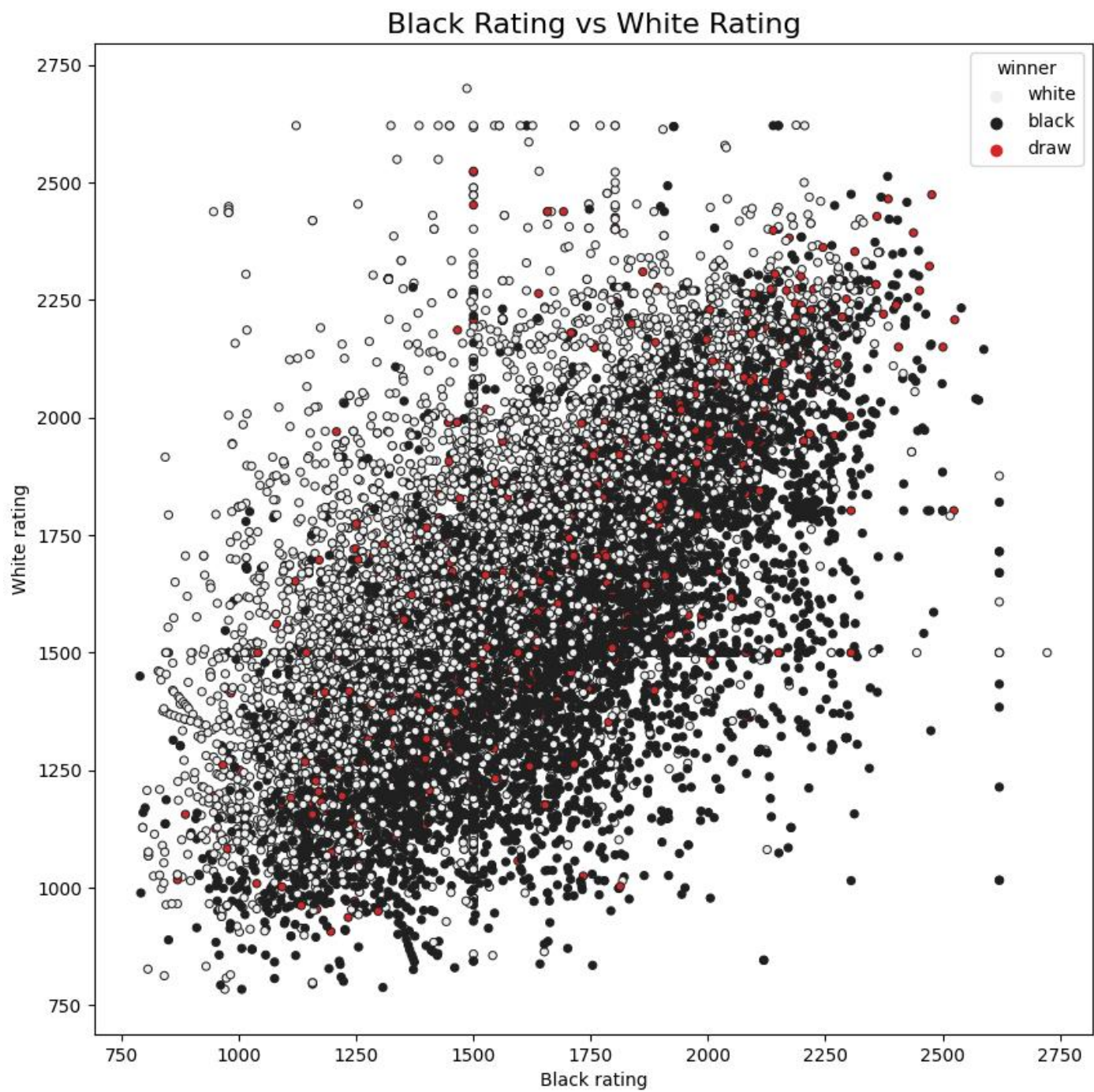


Figure 10 Black Rating vs White Rating

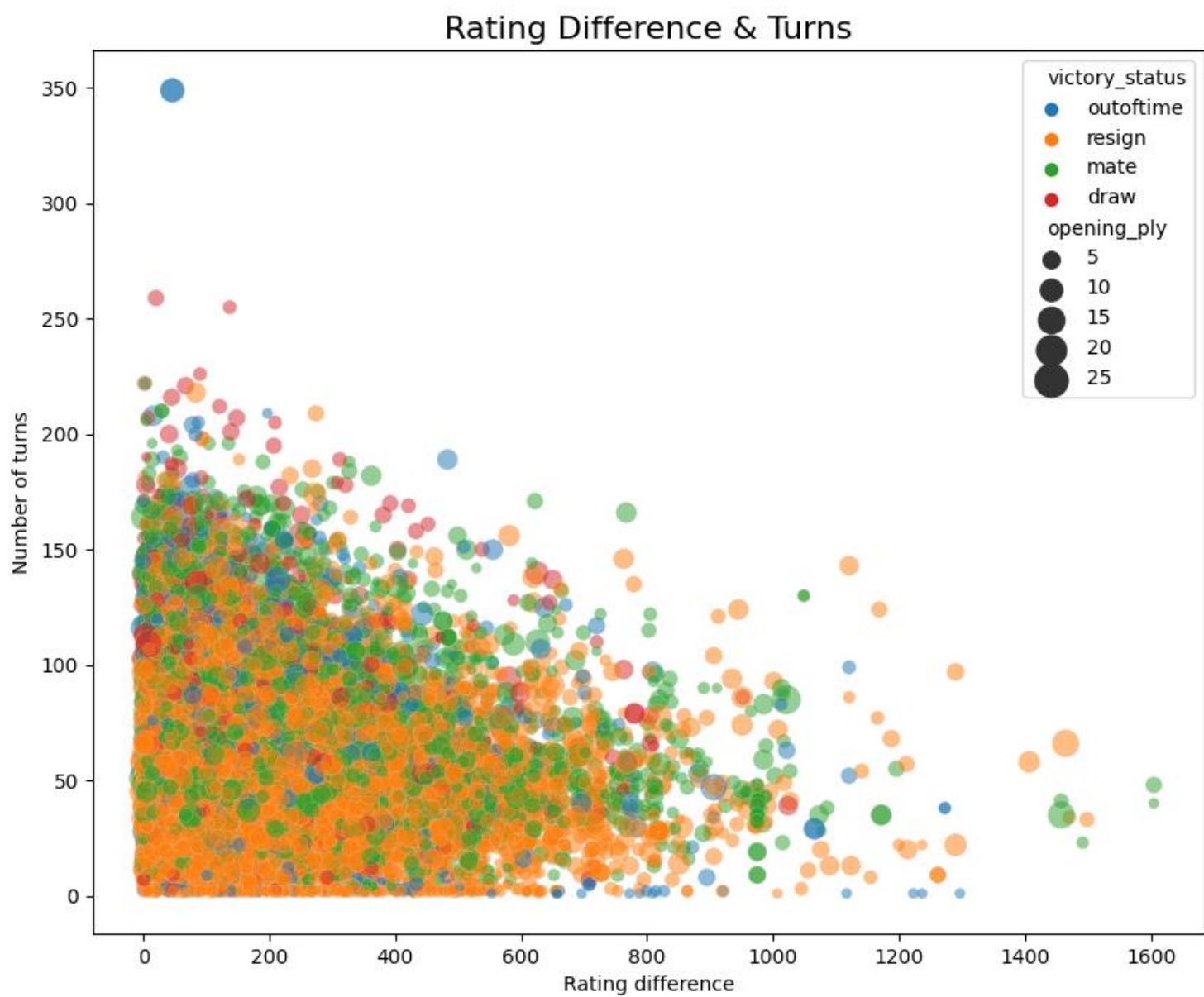


Figure 11 Rating Difference & Turns

## Data Preprocessing

### a) Data Formatting

*“created\_at”* and *“last\_move\_at”* columns were combined into single column *“chess\_time”* and removed from the dataset. *“chess\_time”* data were converted from UNIX time format to minutes.

*“rating\_difference”* column was created from the absolute value of the difference of *“white\_rating”* and *“black\_rating”*.

In the analysis process, redundant features like *“increment\_code”*, *“moves”*, *“white\_id”* and *“black\_id”* have been discarded. *“opening\_name”* feature was eliminated as *“opening\_eco”* has a similar use.

9282 out of 20058 *“chess\_time”* data, approximately 46% of all data, were accumulated at 0 and 166.67 points, which can be observed in Figure 8 & 9. After minutes to seconds conversion, new points were detected as 0 and 10000 seconds which are inconsistent values for chess time. It has been determined that *“chess\_time”* data did not reflect the truth when the other features were compared with time values. Thus, *“chess\_time”* column also removed from the dataset due to corrupted UNIX time data.

### b) Data Cleaning

No null data have been encountered during the analysis process. Therefore, null data imputation was not needed.

Identical rows were detected via comparing *“id”* features and eliminated from the dataset. Approximately 4.71% of all data, 945 rows, were discarded and number of data samples decreased from 20058 to 19113. After duplicate value elimination, the dataframe was reindexed with suitable numbers.

### c) Encoding

In the encoding process of categorical features, one-hot encoding method and label encoder method has been applied. *“rated”*, *“opening\_eco”* and *“victory\_status”* features were encoded as follows:

Label Encoder:

- Column *“rated”* = False: 0, True: 1 (binomial values). Records saved in Column *“rated\_or\_not”*.
- Column *“opening\_eco”* encoded with label encoder. Records saved in Column *“eco”*.

One-hot Encoder:

- Column *“victory\_status”* = *“draw”*, *“mate”*, *“outoftime”*, *“resign”*

| victory_status | vic_outoftime | vic_resign | vic_mate | vic_draw |
|----------------|---------------|------------|----------|----------|
| outoftime      | 1             | 0          | 0        | 0        |
| resign         | 0             | 1          | 0        | 0        |
| mate           | 0             | 0          | 1        | 0        |
| draw           | 0             | 0          | 0        | 1        |

Eventually, original columns *“rated”* and *“victory\_status”* were dropped from the dataframe for preventing from dummy variables.

## Feature Selection – Extraction

In this study, a categorical variable “winner” was tried to be estimated from numeric variables such as encoded variables in previous section. Therefore, Analysis of Variance (ANOVA) F-value method is used in feature selection.

Selected features are as follows:

- “turns”: Number of turns
- “white\_rating”: White player rating
- “black\_rating”: Black player rating
- “rating\_difference”: Rating difference between chess sides
- “vic\_draw”: End game status is draw or not
- “vic\_mate”: End game status is checkmate or not
- “vic\_outoftime”: End game status is out of time or not
- “vic\_resign”: End game status is resign or not

Extracted features are as follows:

- “opening\_ply”: Number of plies in opening phase
- “rated\_or\_not”: Game is rated or not
- “eco”: Encoded Opening ECO number

## Dataset Splitting

The eight values which were selected in feature selection determined as input (x) variables and “winner” value are determined as output (y) variable. 67% of data were splitted as a training set and 33% were splitted as a test set.

## Data Scaling

The numerical differences between features affect the modelling phase of the problem. Thus, standardization has been applied as a scaling to prevent possible errors.

## Model Training – Classification

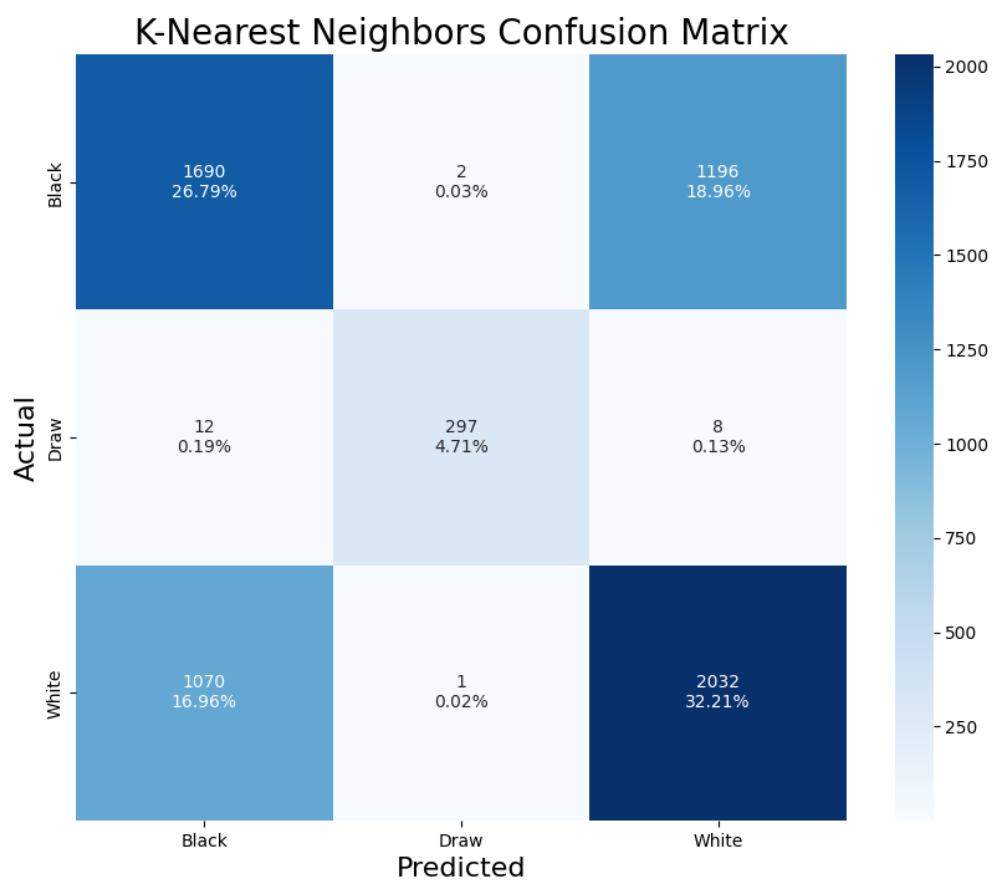
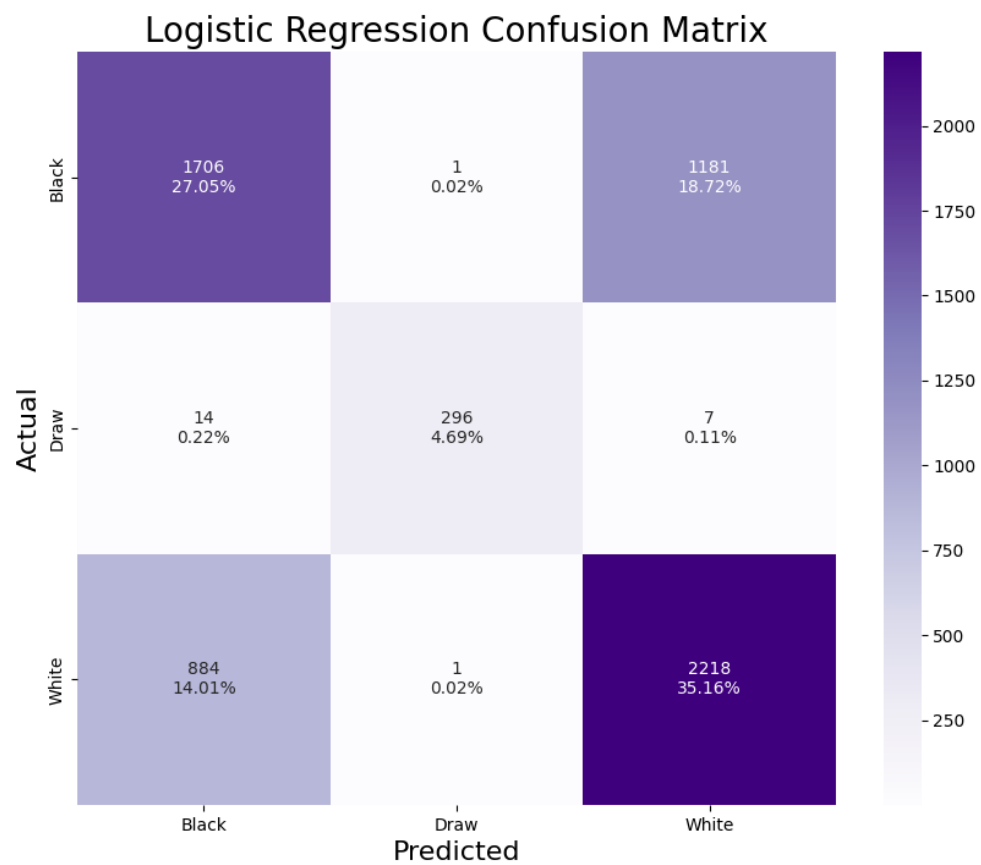
Classification model was preferred due to predicting categorical output from numeric inputs in the problem. Logistic regression, K-nearest neighbor (K-NN), support vector machine (SVM), Naïve Bayes, decision tree and random forest methods were used in the modelling phase.

## Model Evaluation

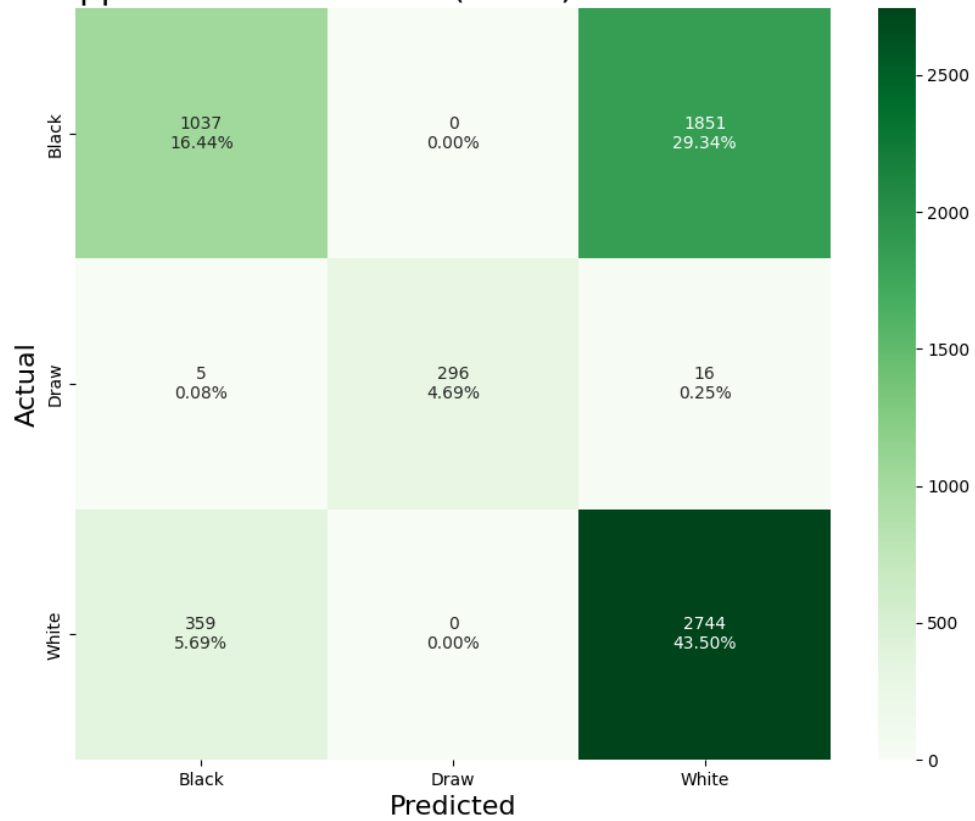
The problem is counted as multiclass classification problem due to having three classes as “black”, “draw” and “white”. Therefore, 3 by 3 confusion matrix is used in evaluation phase.

In confusion matrix, there are four crucial concepts as follows:

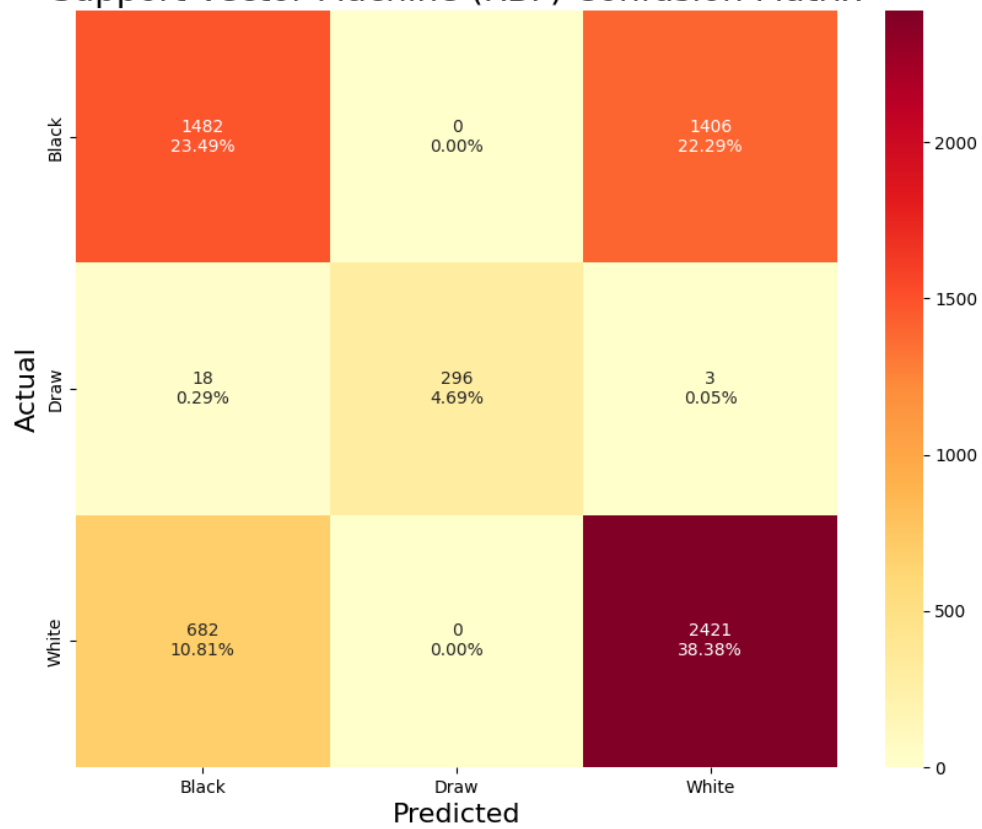
- True Positive (TP): The label belongs to the class and it is predicted as positive.
- True Negative (TN): The label does not belong to the class and it is predicted as negative.
- False Positive (FP): The label does not belong to the class and it predicted as positive.
- False Negative (FN): The label belongs to the class and it predicted as negative.



Support Vector Machine (linear) Confusion Matrix

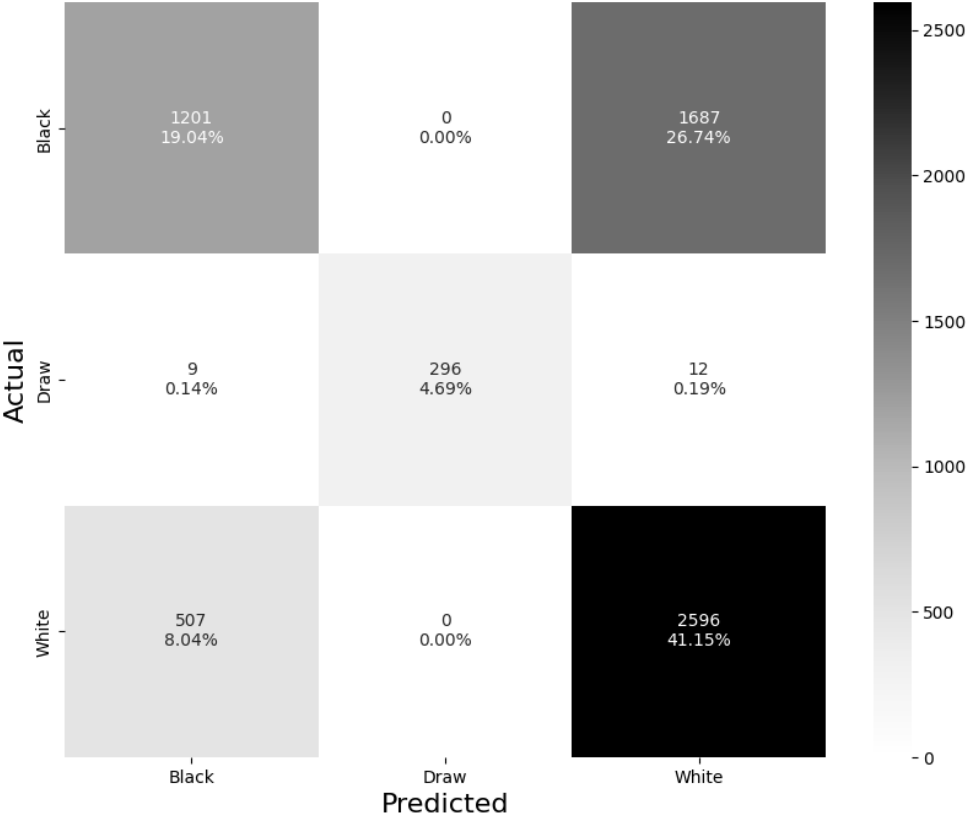


Support Vector Machine (RBF) Confusion Matrix

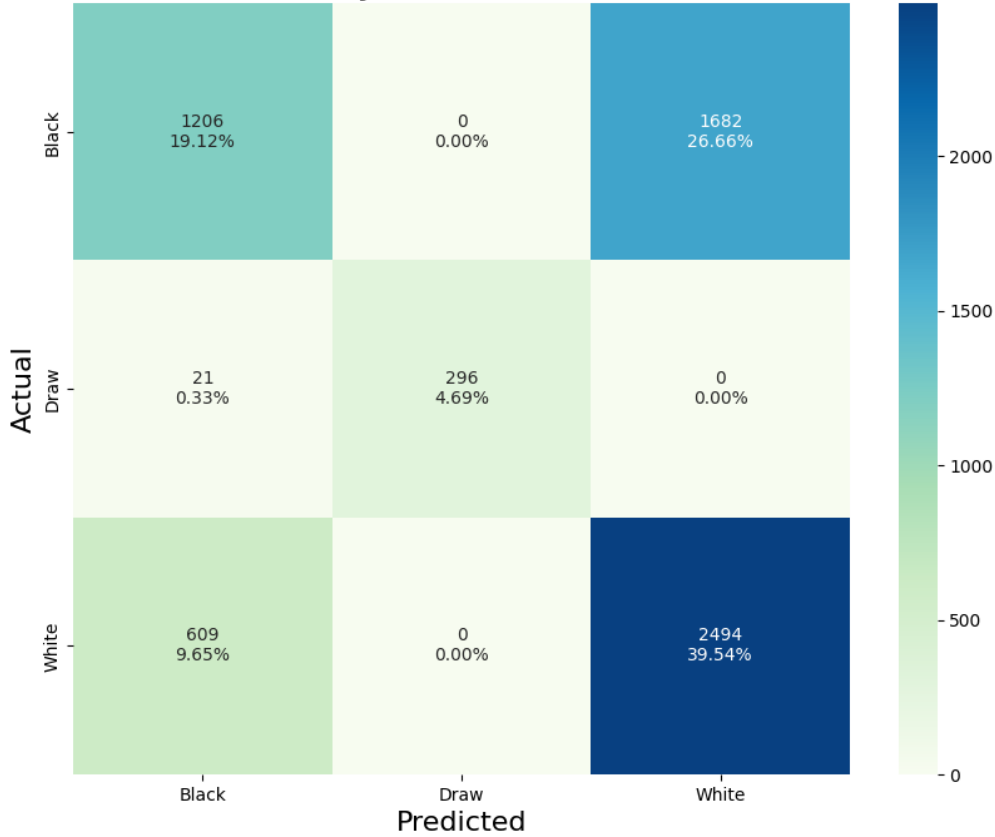


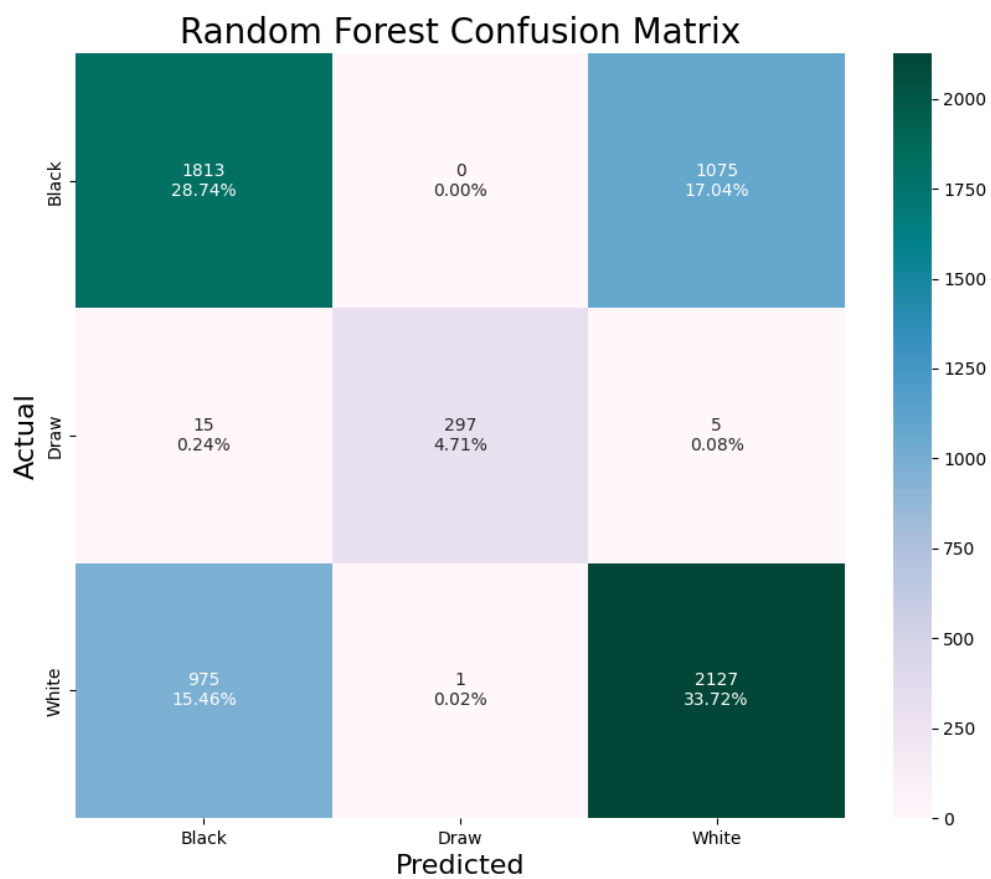
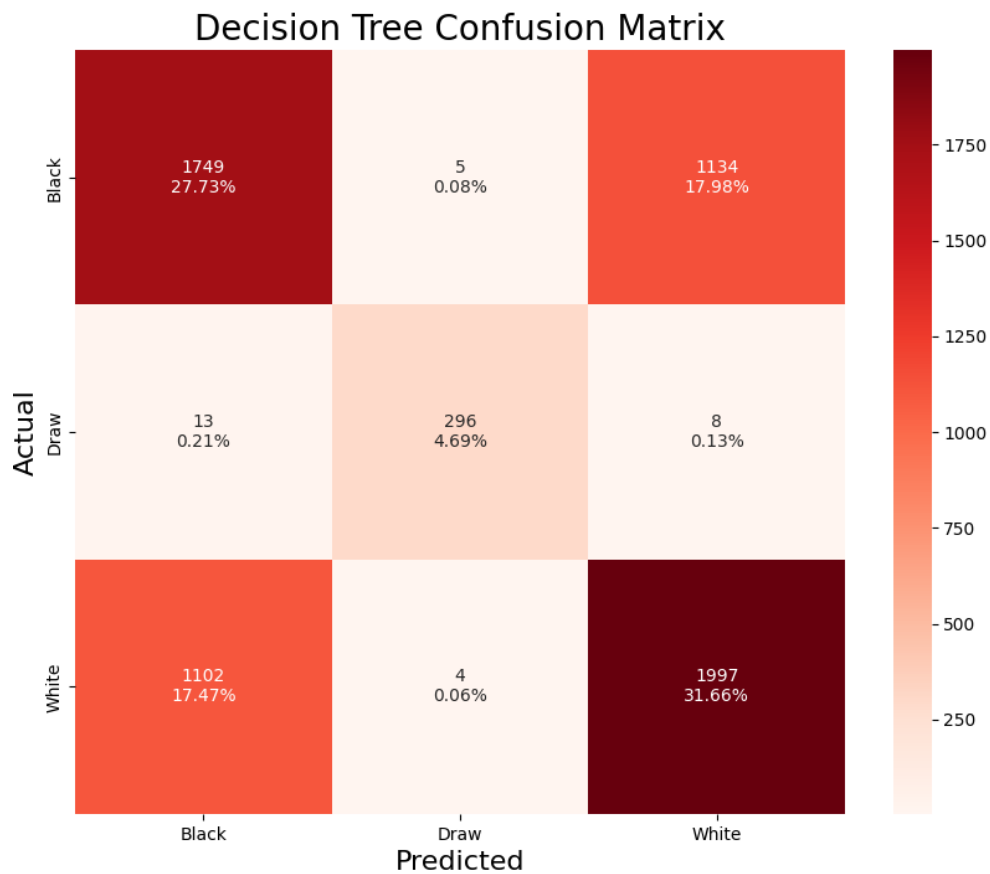


Support Vector Machine (polynomial) Confusion Matrix



Naive Bayes Confusion Matrix





In the evaluation process of the classification methods, four evaluation metrics were used: accuracy, precision, sensitivity and f1 score. These metrics can be calculated as follows:

- Accuracy =  $(TP + TN) / (TP + TN + FP + FN)$
- Precision =  $TP / (TP + FP)$
- Sensitivity =  $TP / (TP + FN)$
- F1 Score =  $2 * Precision * Sensitivity / (Precision + Sensitivity)$

### Evaluation Report

| Method              | Accuracy | Precision | Sensitivity | F1 Score |
|---------------------|----------|-----------|-------------|----------|
| Logistic Regression | 0.66899  | 0.76655   | 0.74642     | 0.75513  |
| K-NN                | 0.63713  | 0.74253   | 0.72565     | 0.73367  |
| SVM (linear)        | 0.64632  | 0.77843   | 0.72571     | 0.72025  |
| SVM (RBF)           | 0.66566  | 0.77044   | 0.74237     | 0.74959  |
| SVM (polynomial)    | 0.64886  | 0.76797   | 0.72874     | 0.72972  |
| Naïve Bayes         | 0.63348  | 0.75136   | 0.71836     | 0.72053  |
| Decision Tree       | 0.63871  | 0.73870   | 0.72724     | 0.73285  |
| Random Forest       | 0.67422  | 0.76962   | 0.75181     | 0.76032  |