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/datasets/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[1] *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[2] code *as opening_eco*
- Opening name[3] *as opening_name*
- Opening ply[4] *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.

Data Visualization

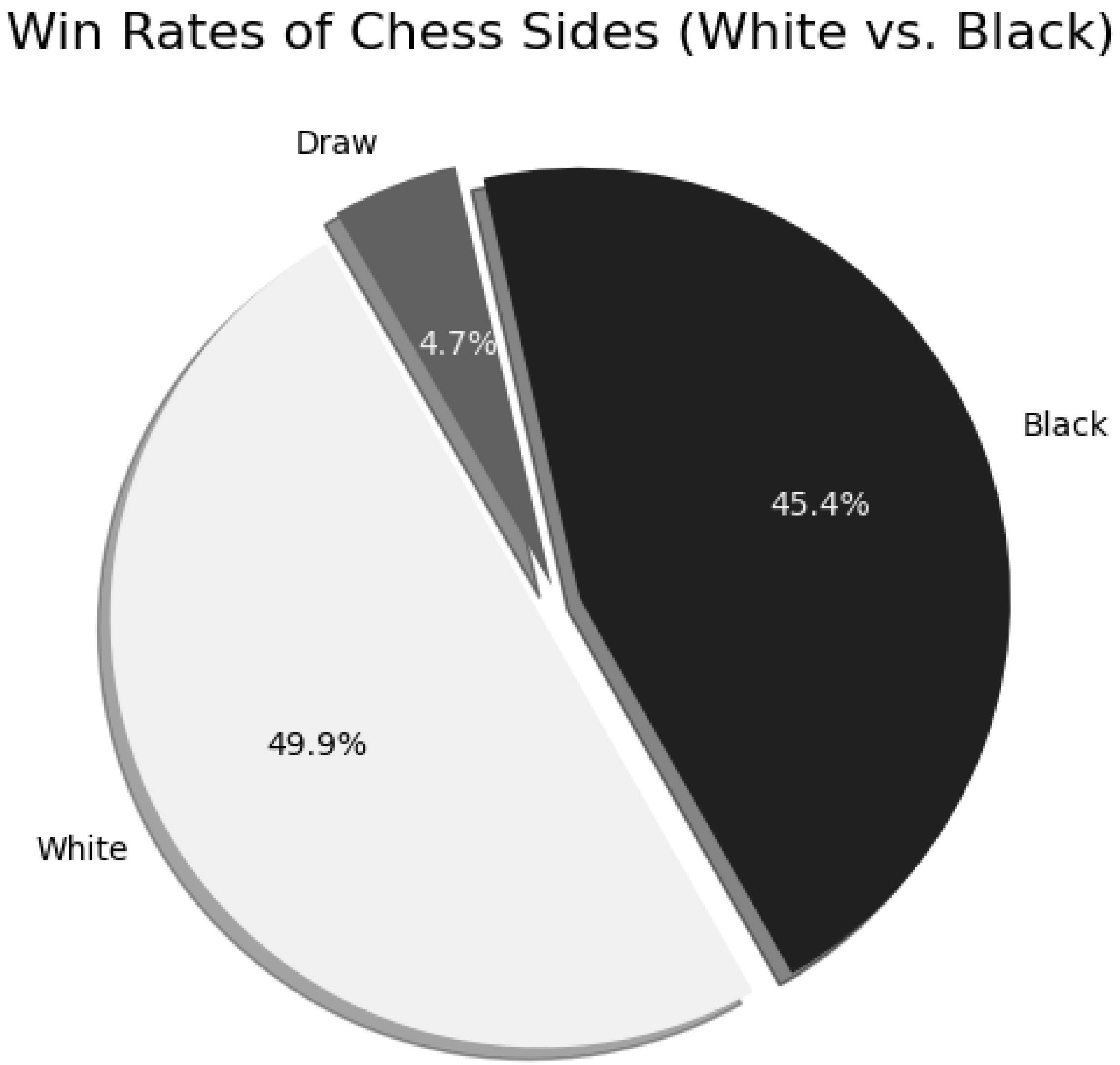


Figure 1 Win Rates of Chess Sides

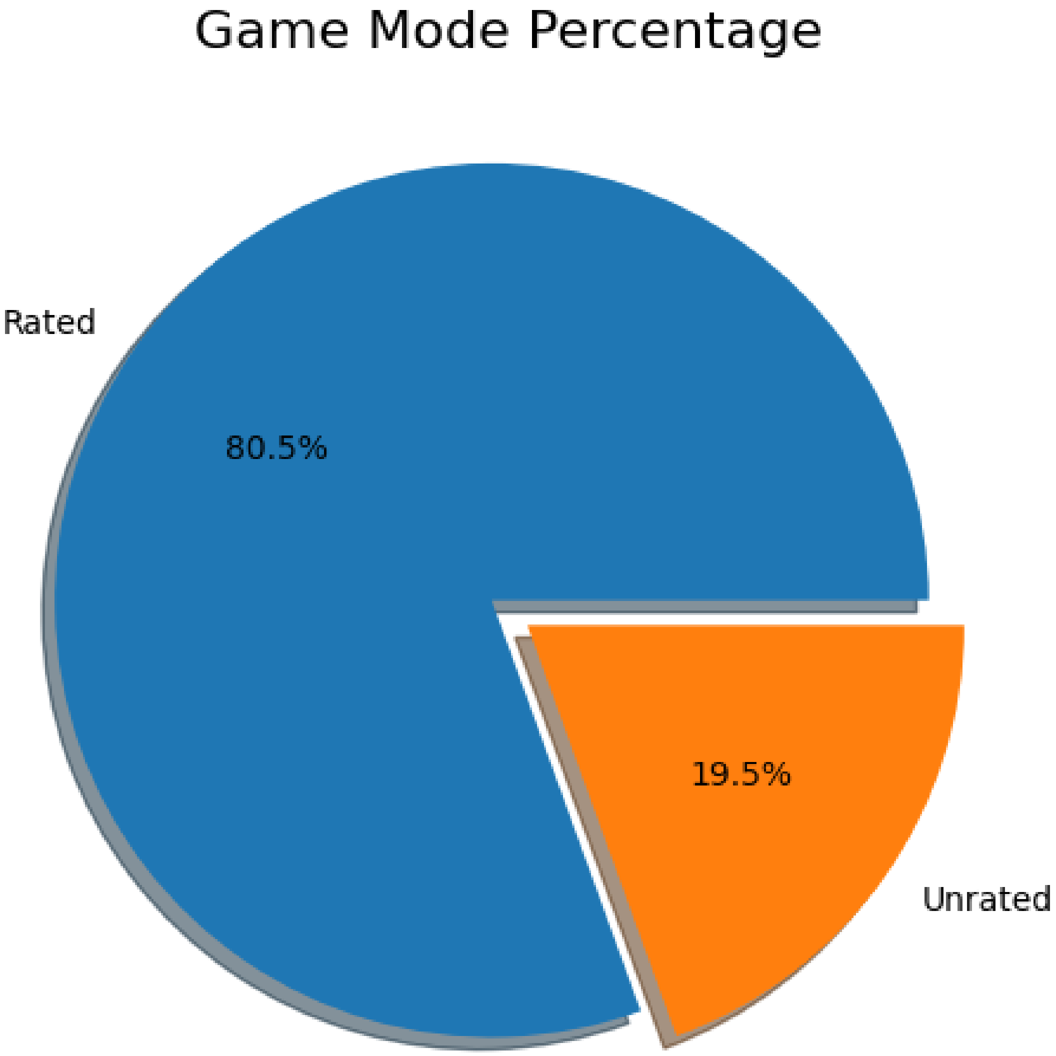


Figure 2 Game Mode Percentage

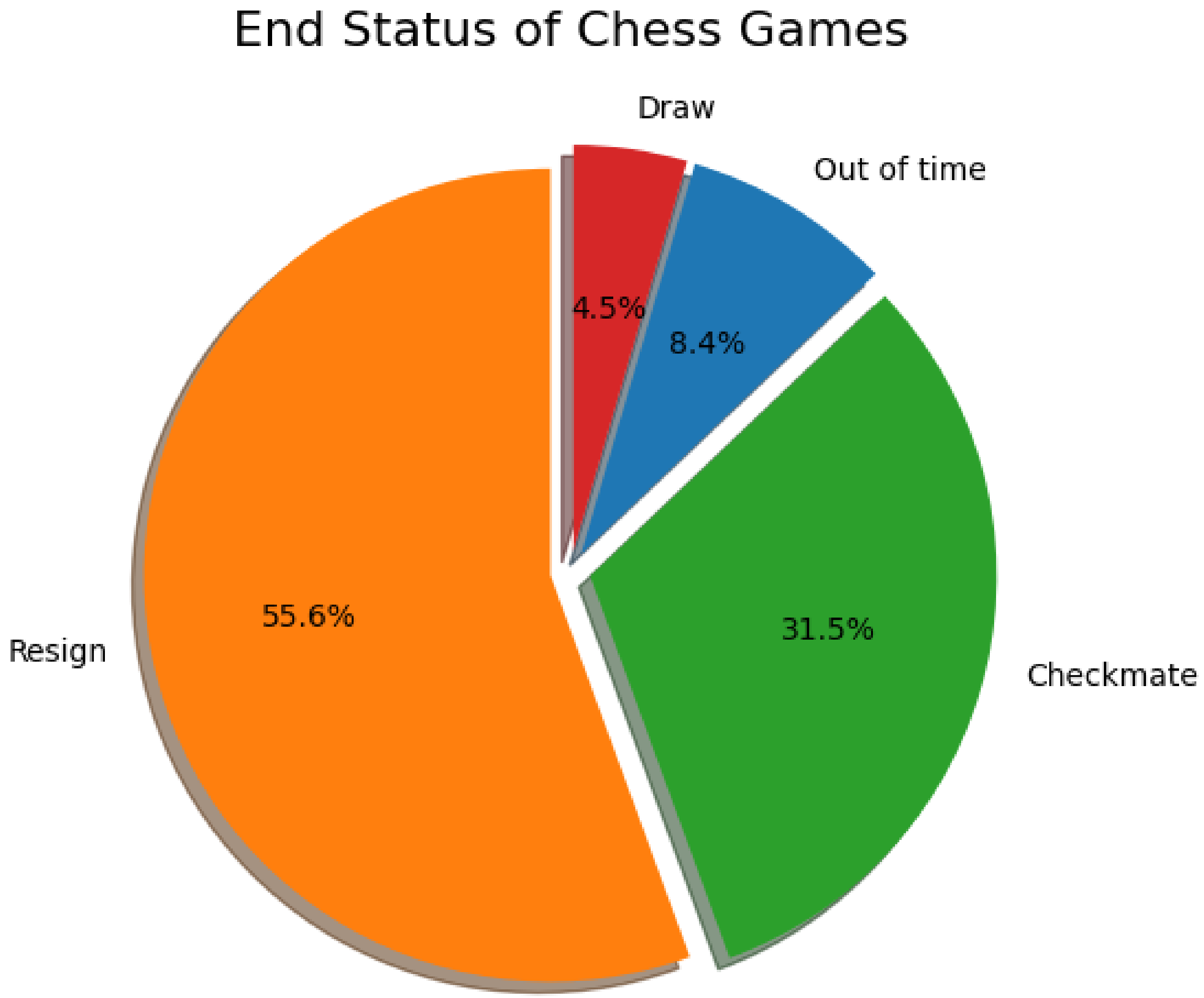


Figure 3 End Status of Chess Games

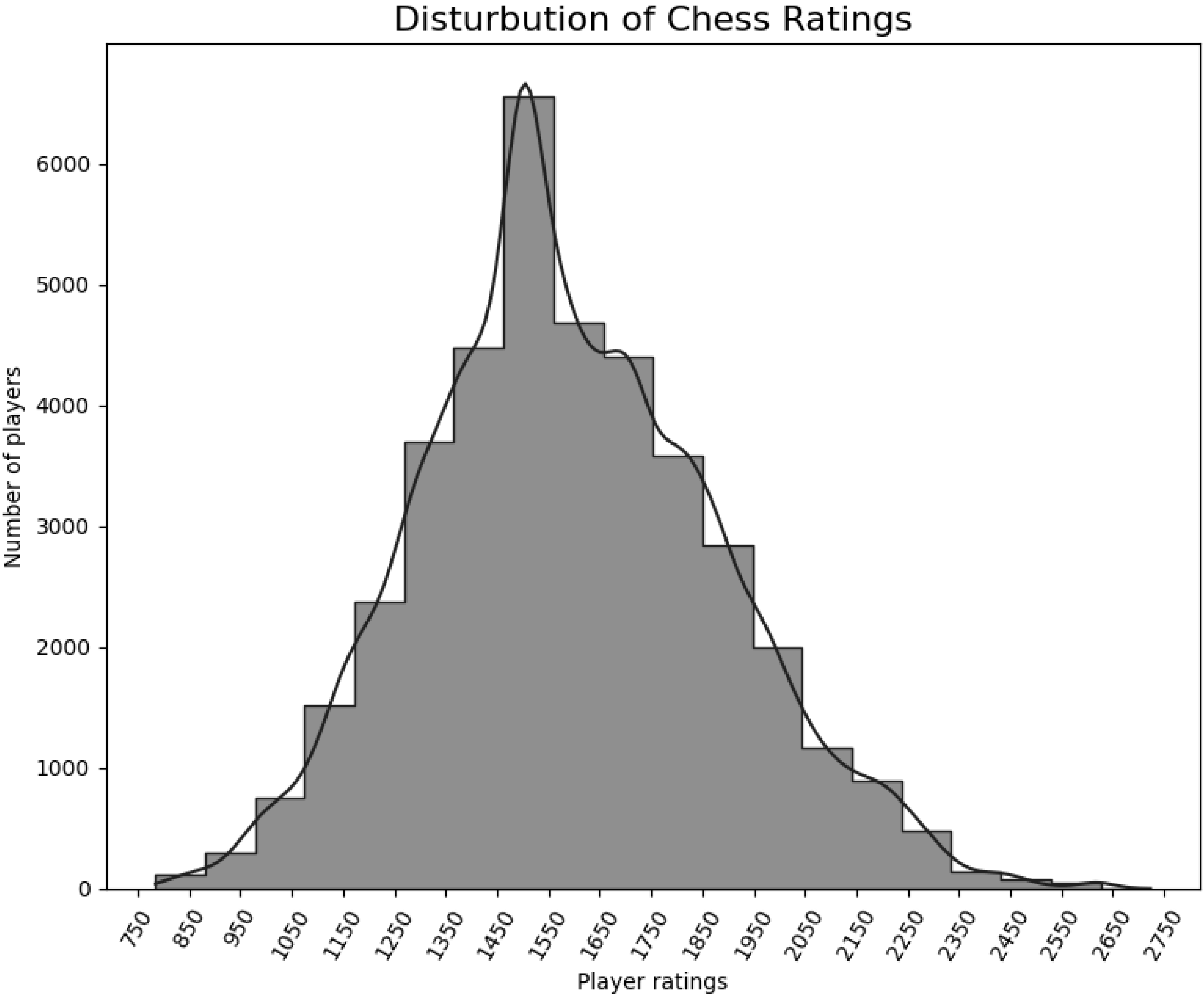


Figure 4 Disturbution of Chess Ratings

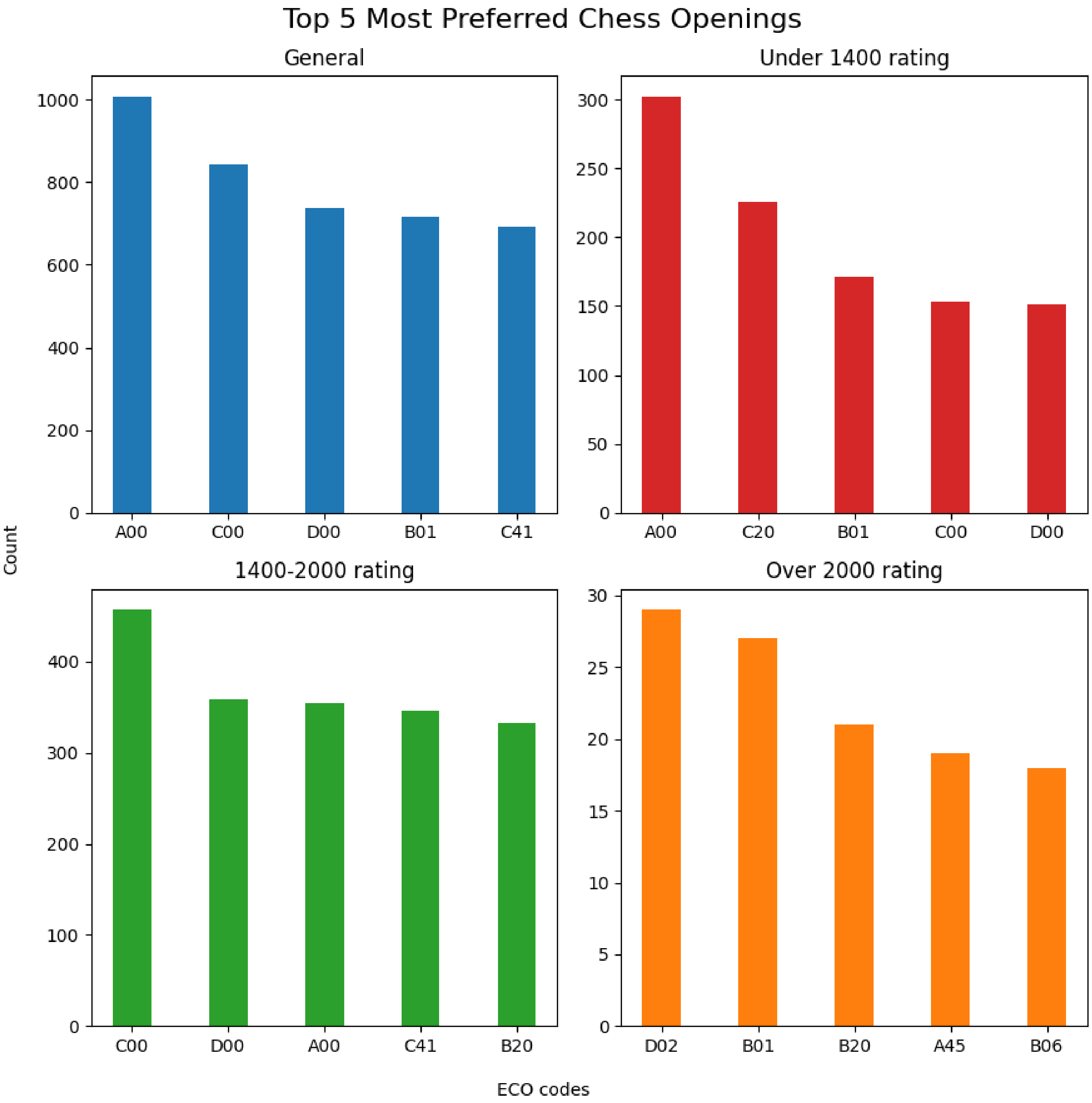


Figure 5 Top 5 Most Preferred Chess Openings

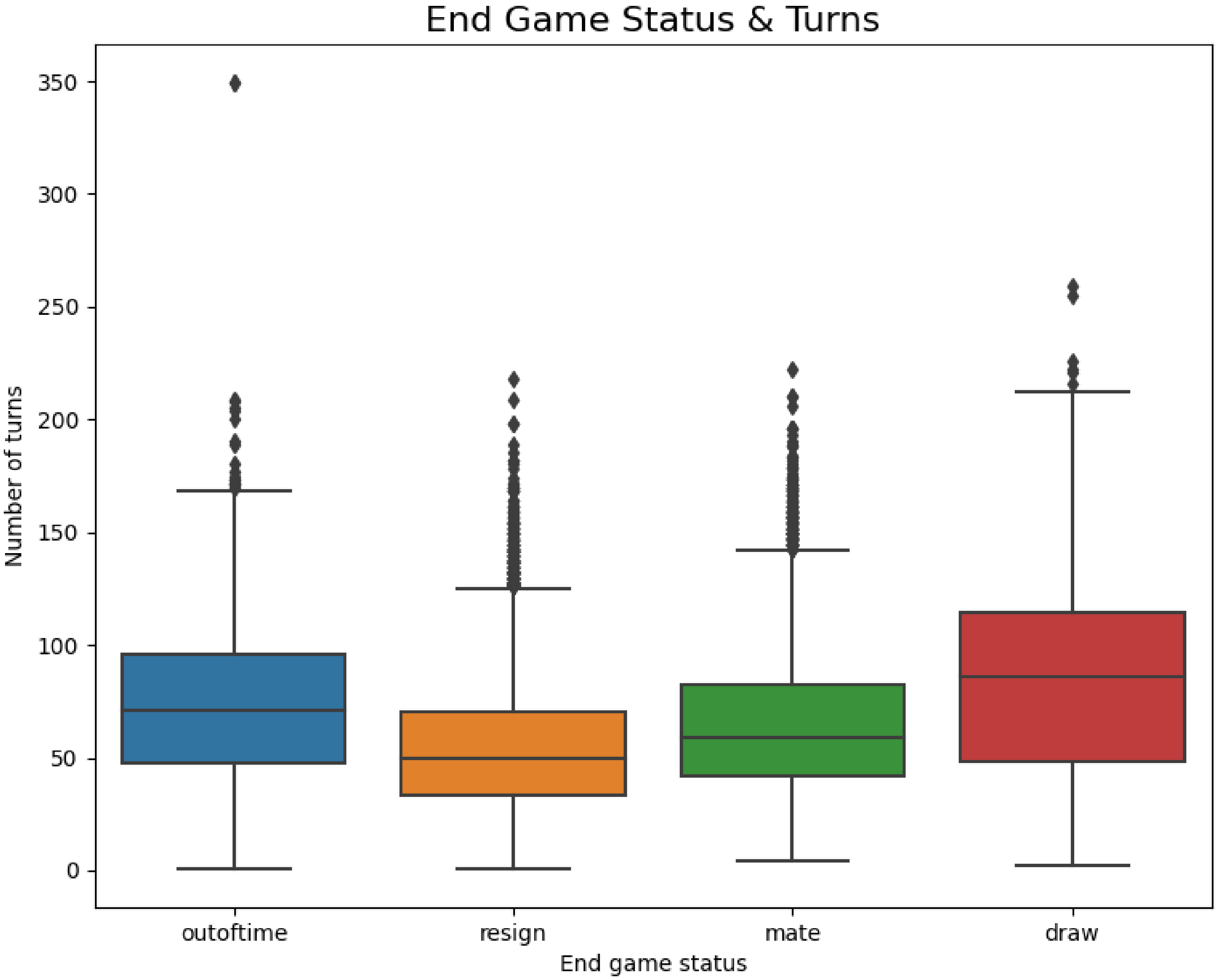


Figure 6 End Game Status & Turns

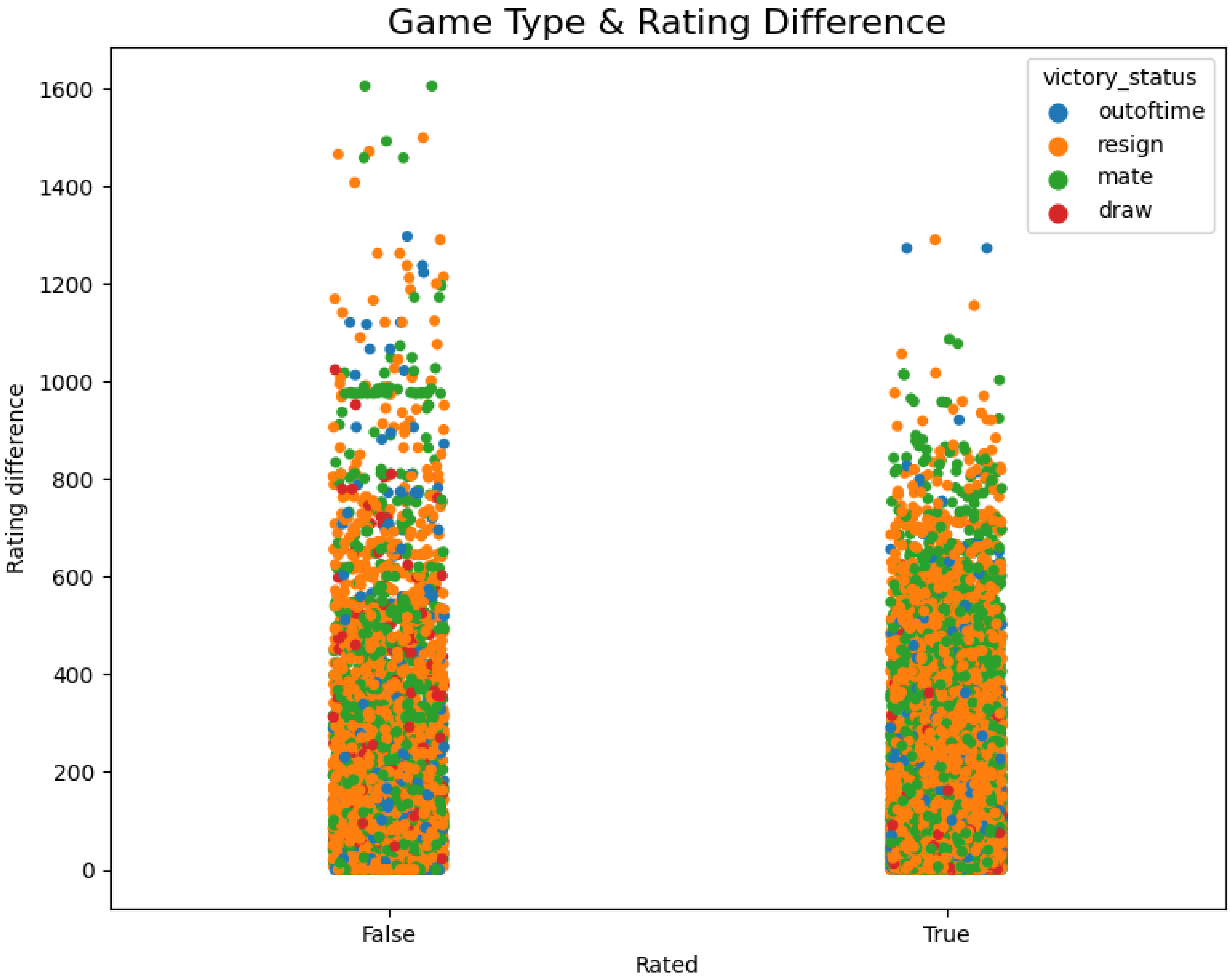


Figure 7 Game Type & Rating Difference

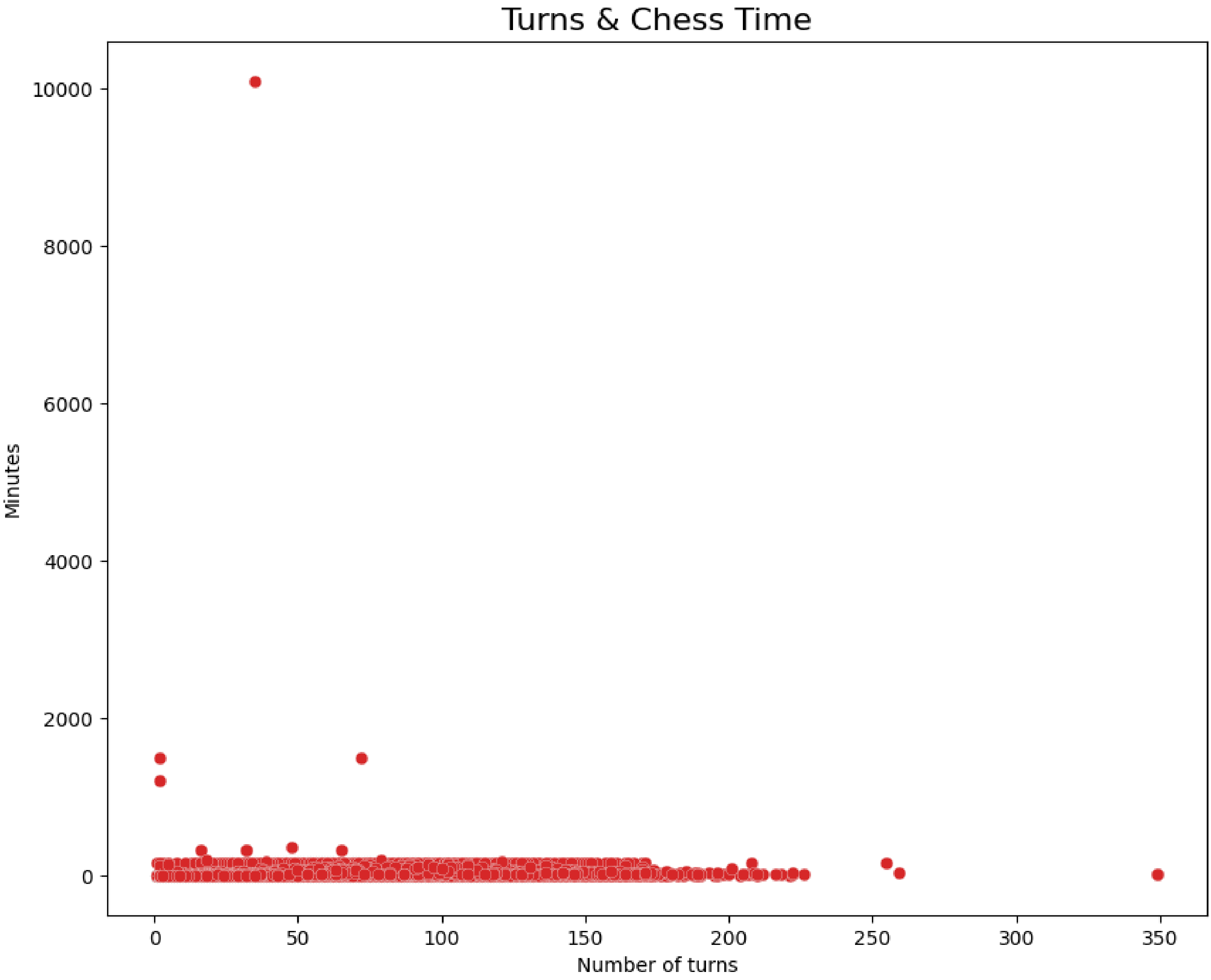


Figure 8 Turns & Chess Time

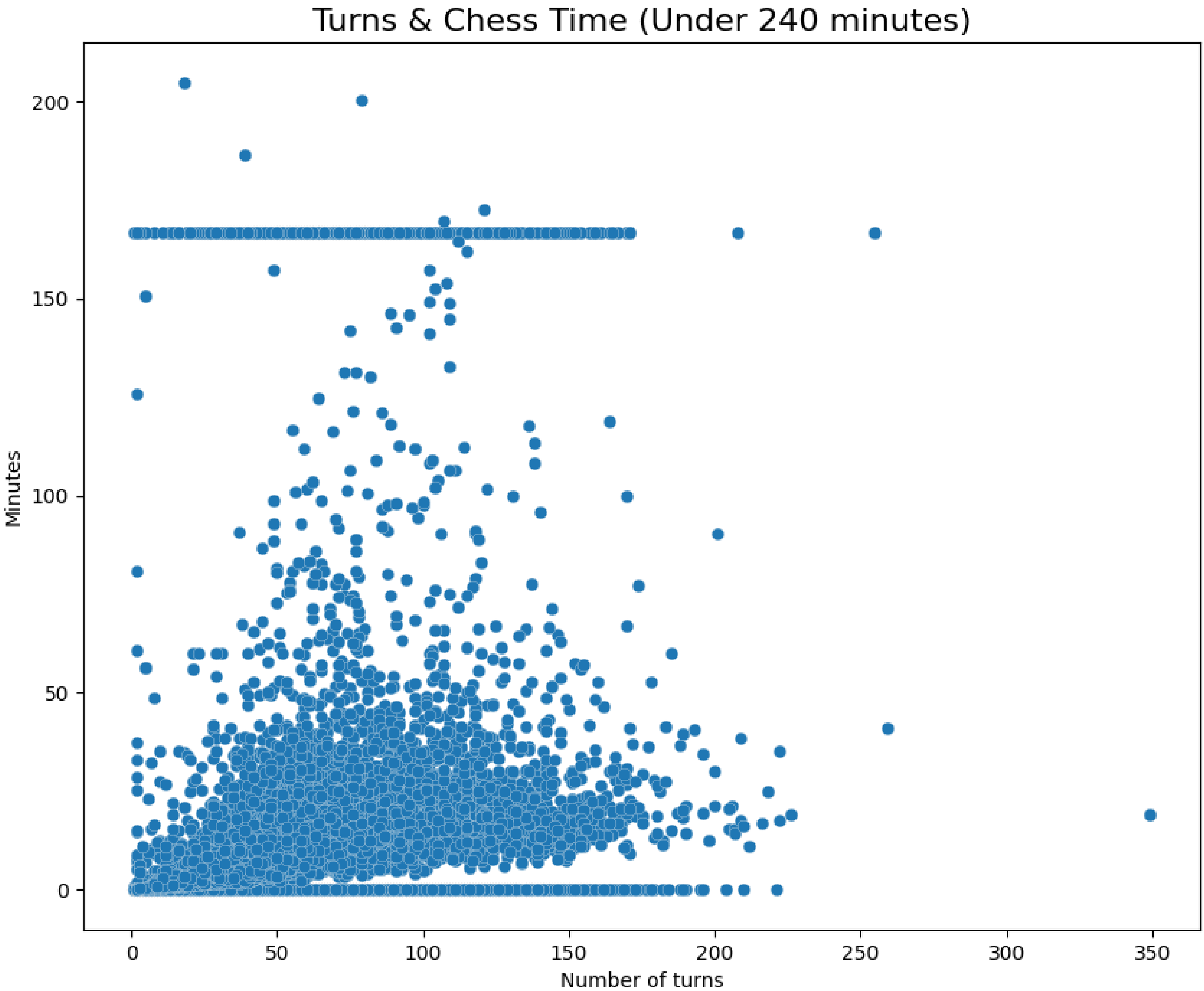


Figure 9 Turns & Chess Time (Under 240 minutes)

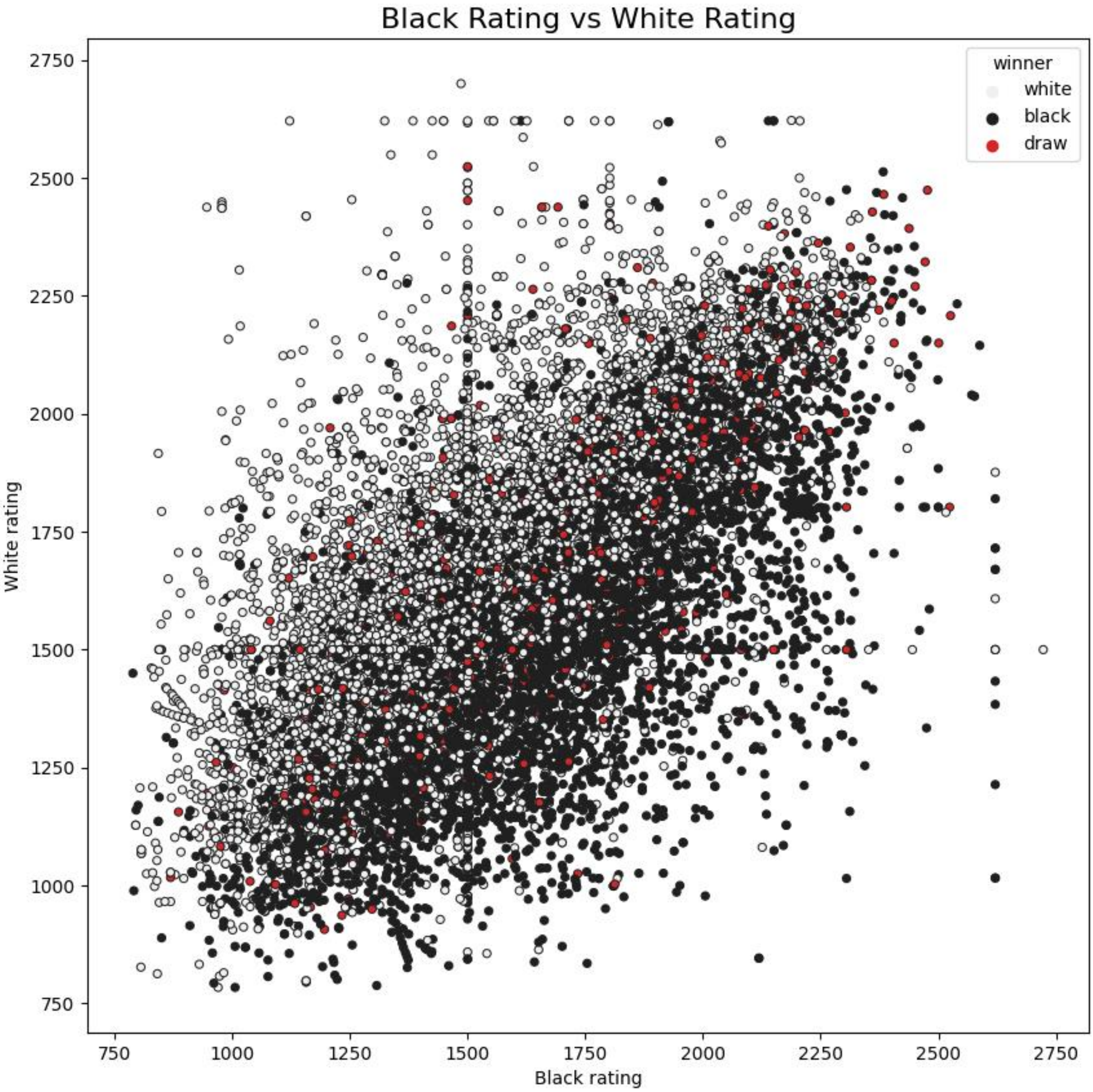


Figure 10 Black Rating vs White Rating

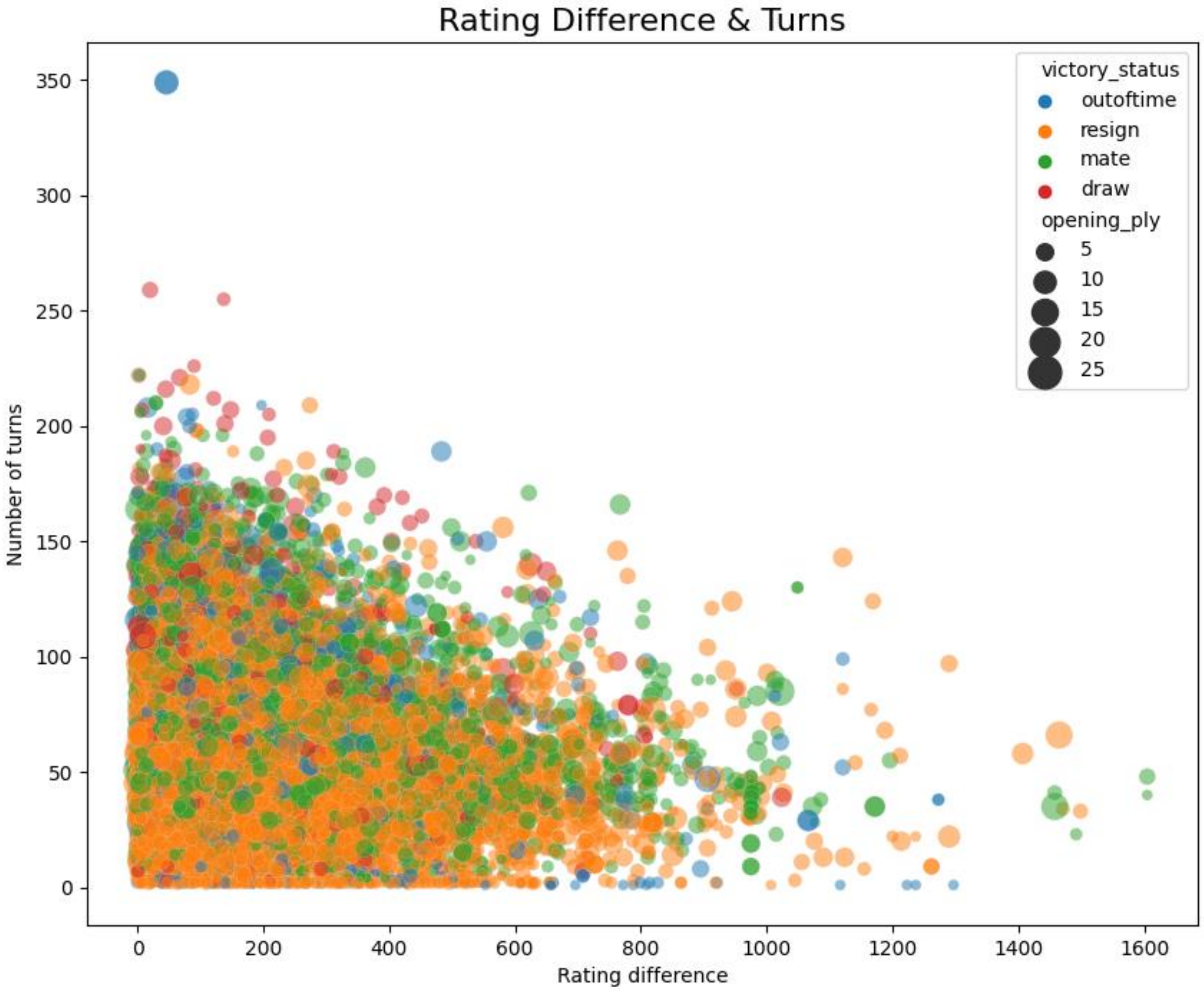


Figure 11 Rating Difference & Turns

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 "*incremenet_code", "moves", "white_id"* and *"black_id"* have been discarded. "o*pening_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 training set and 33% were splitted as 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

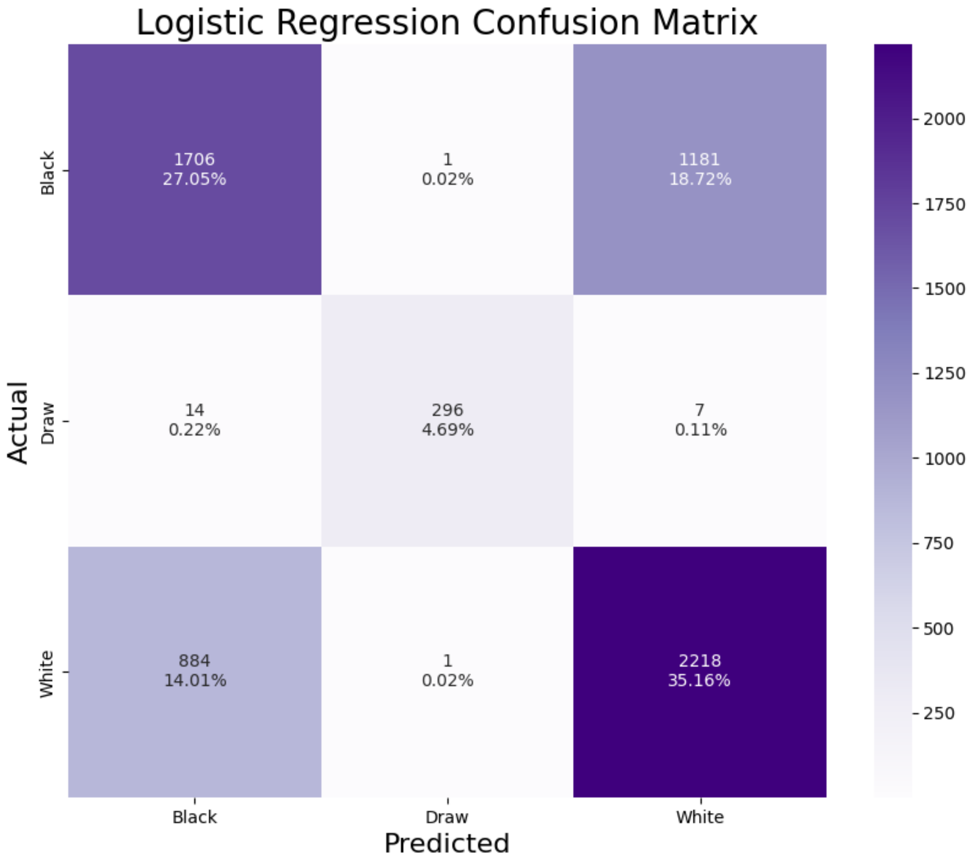
Classification model was preferred due to predicting categorical output from numeric inputs in the problem. Logistic regression, K-nearest neighbor (K-NN), support vector machine (SVM), Naïve Bayes, decision tree and random forest methods were used in modelling phase.

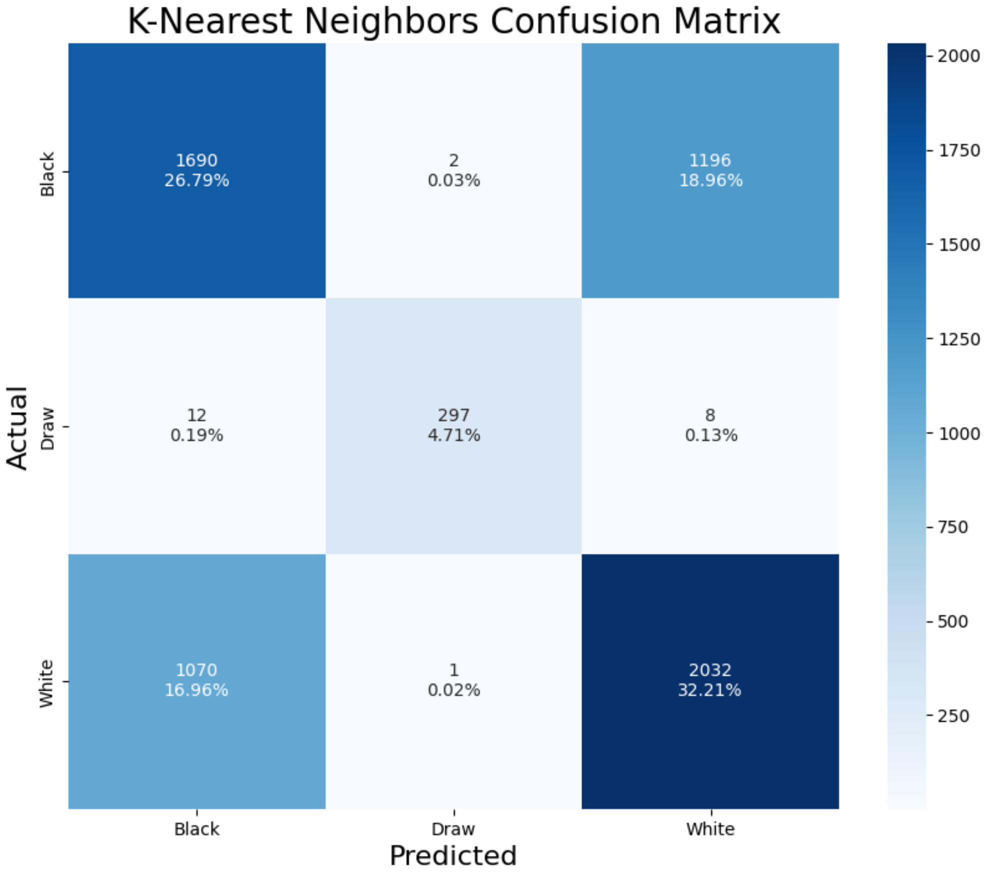
Model Evaluation

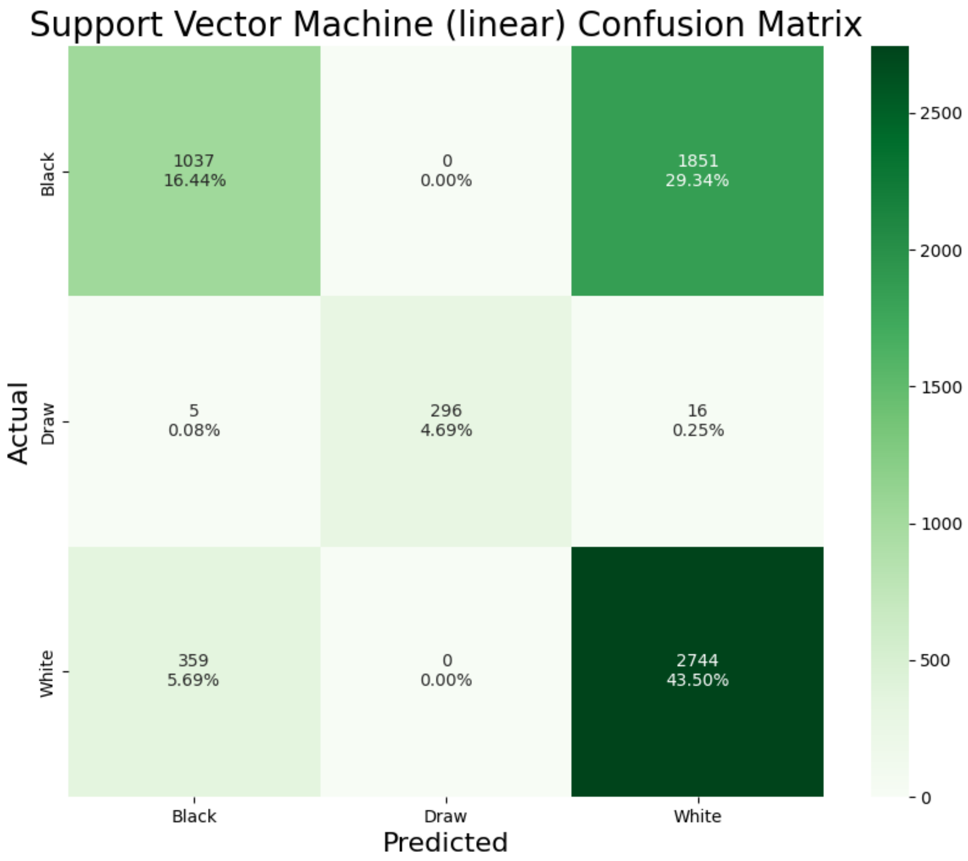
The problem is counted as multiclass classification problem due to having three classes as “black”, “draw” and “white”. Therefore, 3 by 3 confusion matrix is used in evaluation phase.

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.

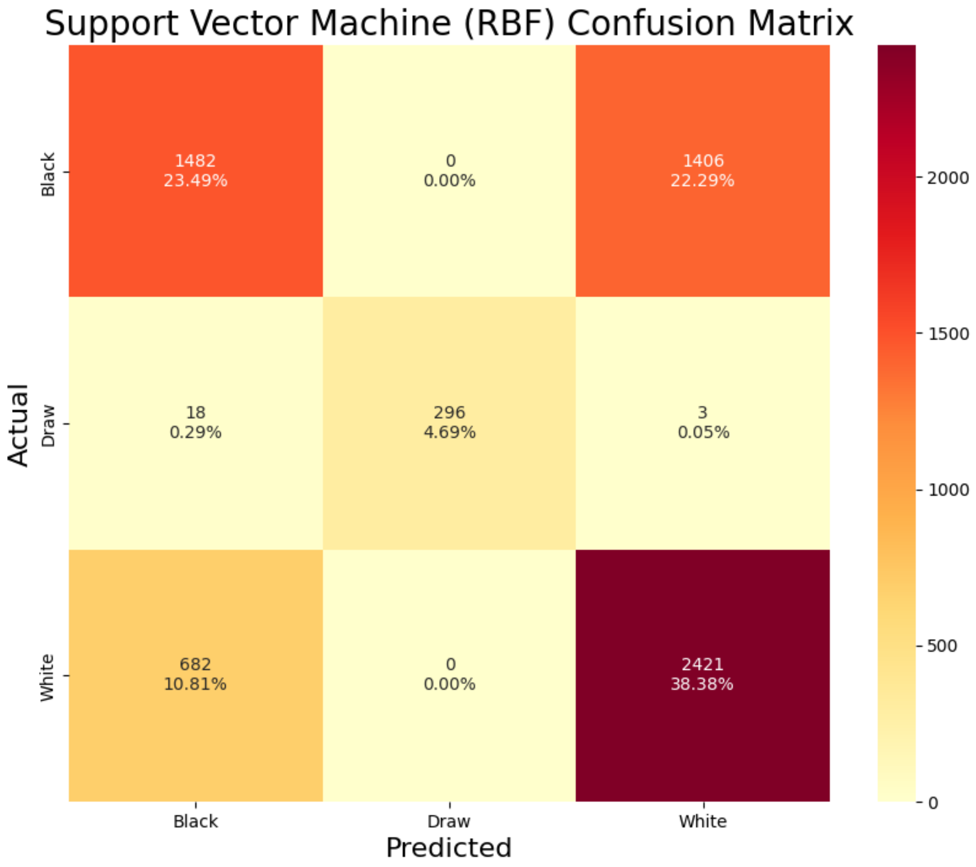

Logistic Regression Confusion Matrix
Actual
Black
Draw
White
Predicted
Black
Draw
White
1706
27.05%
1
0.02%
1181
18.72%
14
0.22%
296
4.69%
7
0.11%
884
14.01%
1
0.02%
2218
35.16%
2000
1750
1500
1250
1000
750
500
250


K-Nearest Neighbors Confusion Matrix
Actual
Black
Draw
White
Predicted
Black
Draw
White
1690
26.79%
2
0.03%
1196
18.96%
12
0.19%
297
4.71%
8
0.13%
1070
16.96%
1
0.02%
2032
32.21%
2000
1750
1500
1250
1000
750
500
250



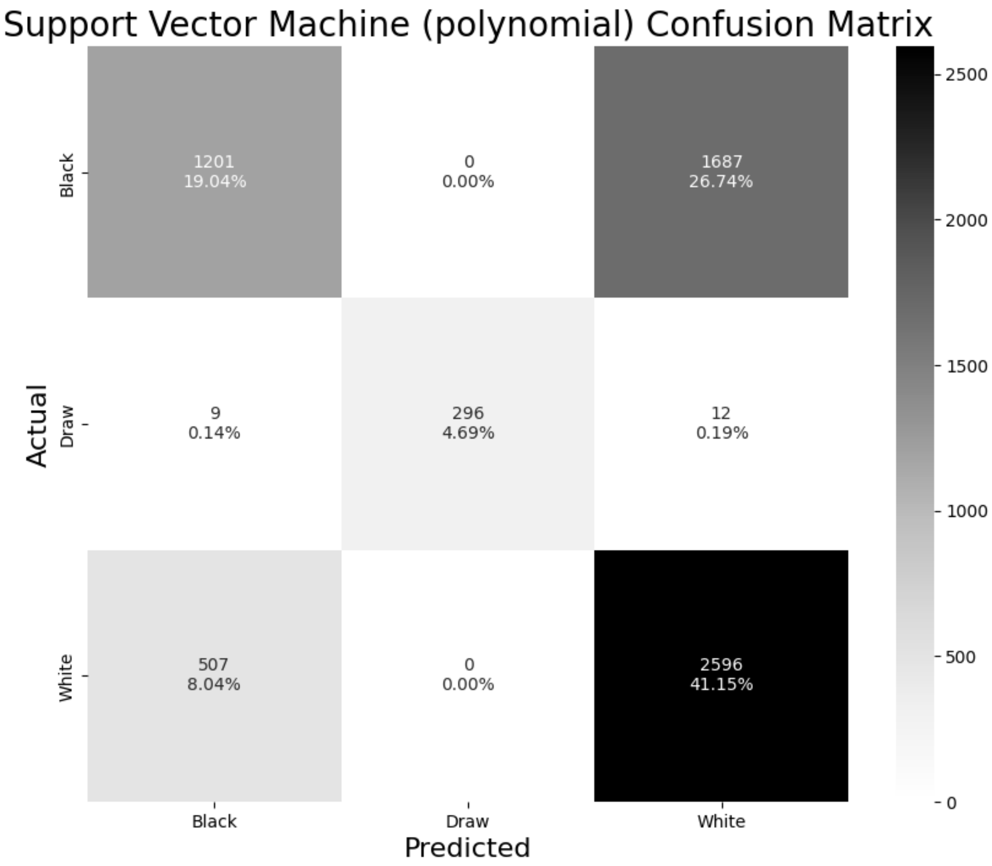
Support Vector Machine (linear) Confusion Matrix

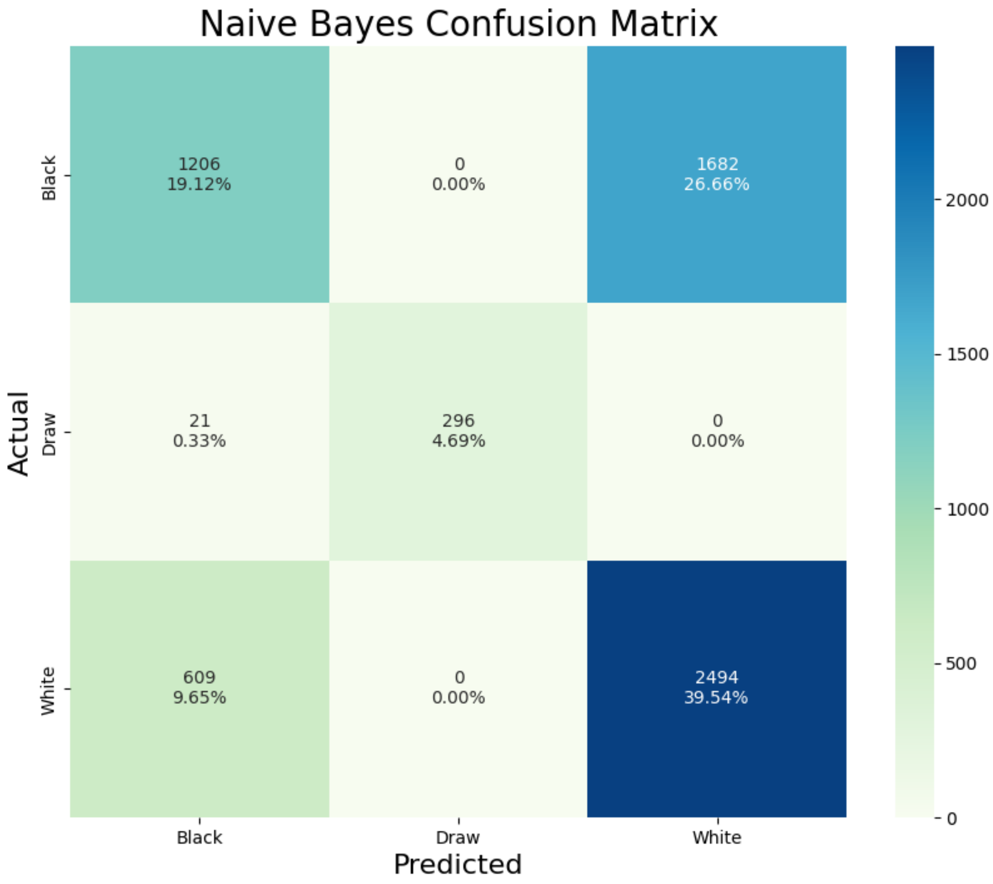
Actual \ Predicted	Black	Draw	White
Black	1037 16.44%	0 0.00%	1851 29.34%
Draw	5 0.08%	296 4.69%	16 0.25%
White	359 5.69%	0 0.00%	2744 43.50%

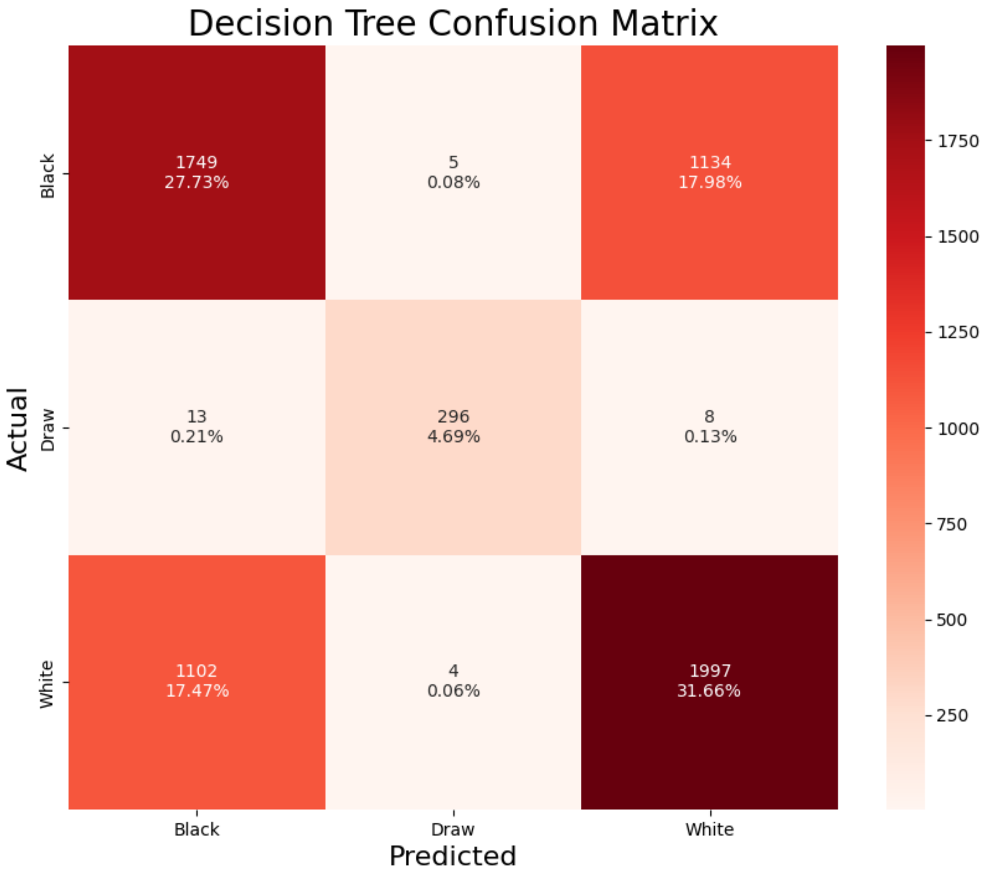


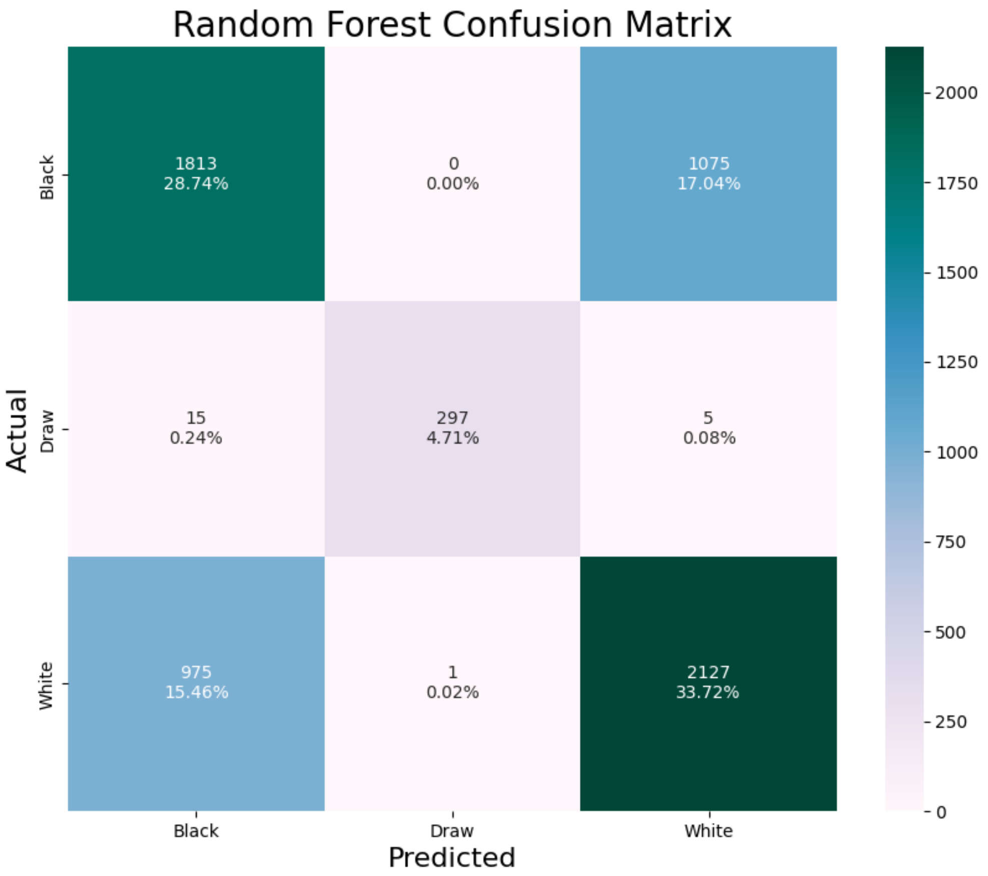
Support Vector Machine (RBF) Confusion Matrix

Actual \ Predicted	Black	Draw	White
Black	1482 23.49%	0 0.00%	1406 22.29%
Draw	18 0.29%	296 4.69%	3 0.05%
White	682 10.81%	0 0.00%	2421 38.38%


Support Vector Machine (polynomial) Confusion Matrix
Actual
Black
Draw
White
Predicted
Black
Draw
White
1201
19.04%
0
0.00%
1687
26.74%
9
0.14%
296
4.69%
12
0.19%
507
8.04%
0
0.00%
2596
41.15%
2500
2000
1500
1000
500
0


Naive Bayes Confusion Matrix
Actual
Black
Draw
White
Predicted
Black
Draw
White
1206
19.12%
0
0.00%
1682
26.66%
21
0.33%
296
4.69%
0
0.00%
609
9.65%
0
0.00%
2494
39.54%
2000
1500
1000
500
0


Decision Tree Confusion Matrix
Actual
Black
Draw
White
Predicted
Black
Draw
White
1749
27.73%
5
0.08%
1134
17.98%
13
0.21%
296
4.69%
8
0.13%
1102
17.47%
4
0.06%
1997
31.66%
1750
1500
1250
1000
750
500
250


Random Forest Confusion Matrix
Actual
Black
Draw
White
Predicted
Black
Draw
White
1813
28.74%
0
0.00%
1075
17.04%
15
0.24%
297
4.71%
5
0.08%
975
15.46%
1
0.02%
2127
33.72%
2000
1750
1500
1250
1000
750
500
250
0

In the evaluation process of the classification methods, four evalutaion 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	Sensivity	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