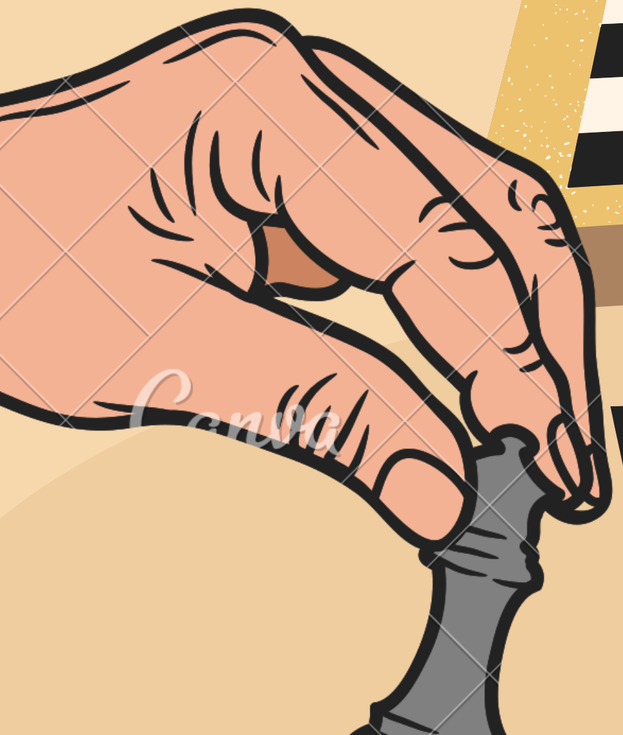
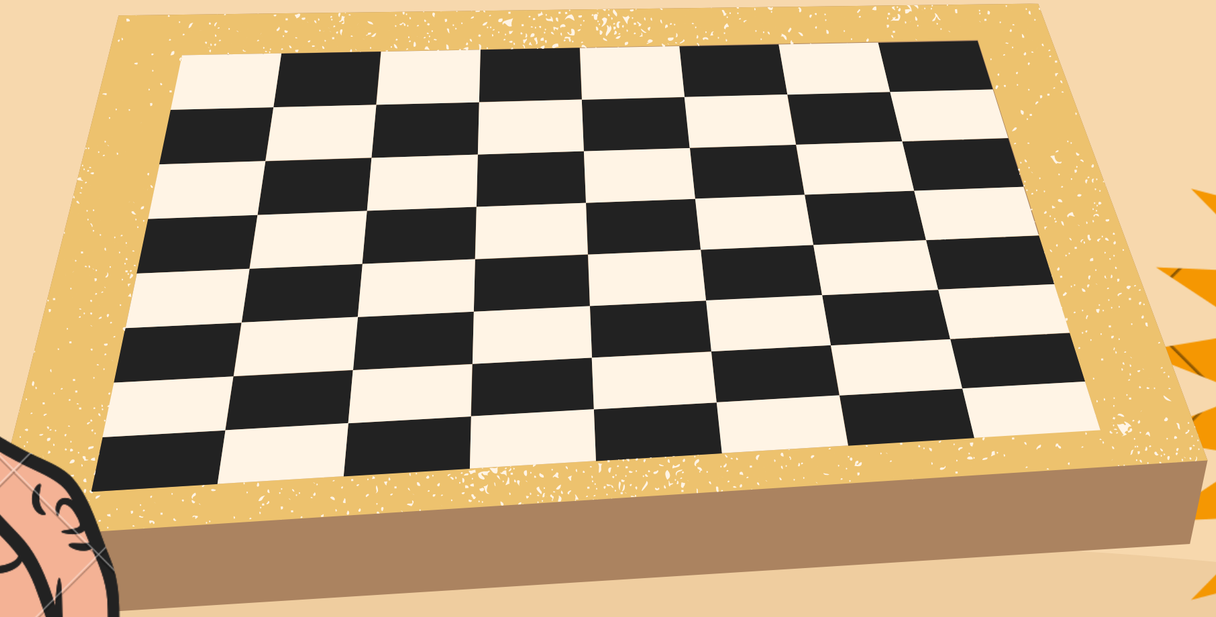
SC1015 MINI PROJECT



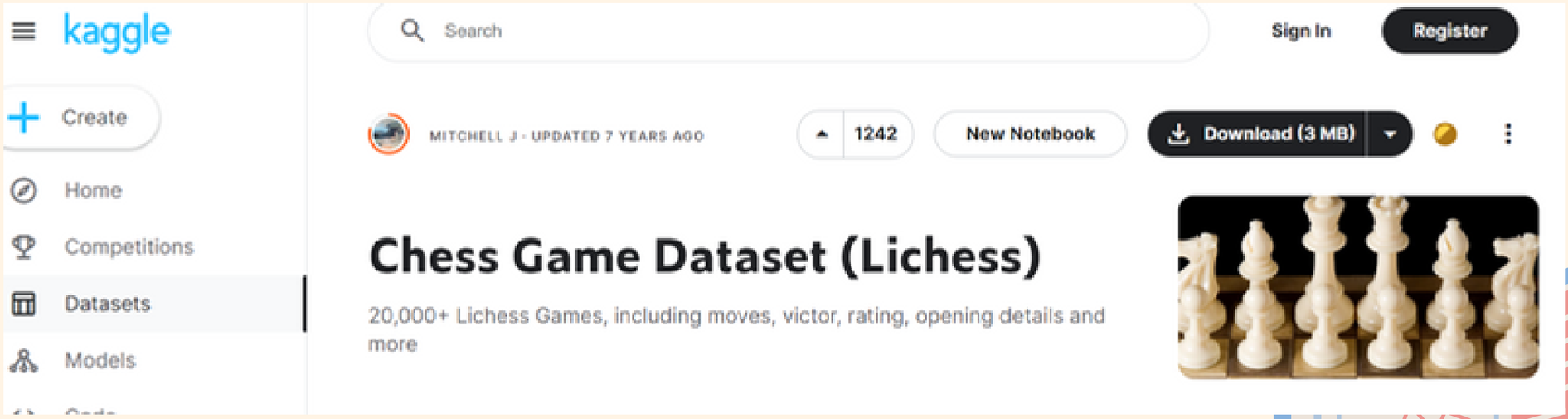


CHESS

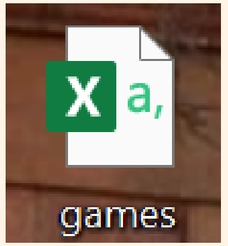
FCSE GROUP 11
LOW WEI JIE
CHUA YOU WANG, NATHAN-HOPE
DOO HONG WEN, EVAN

DATASET USED:

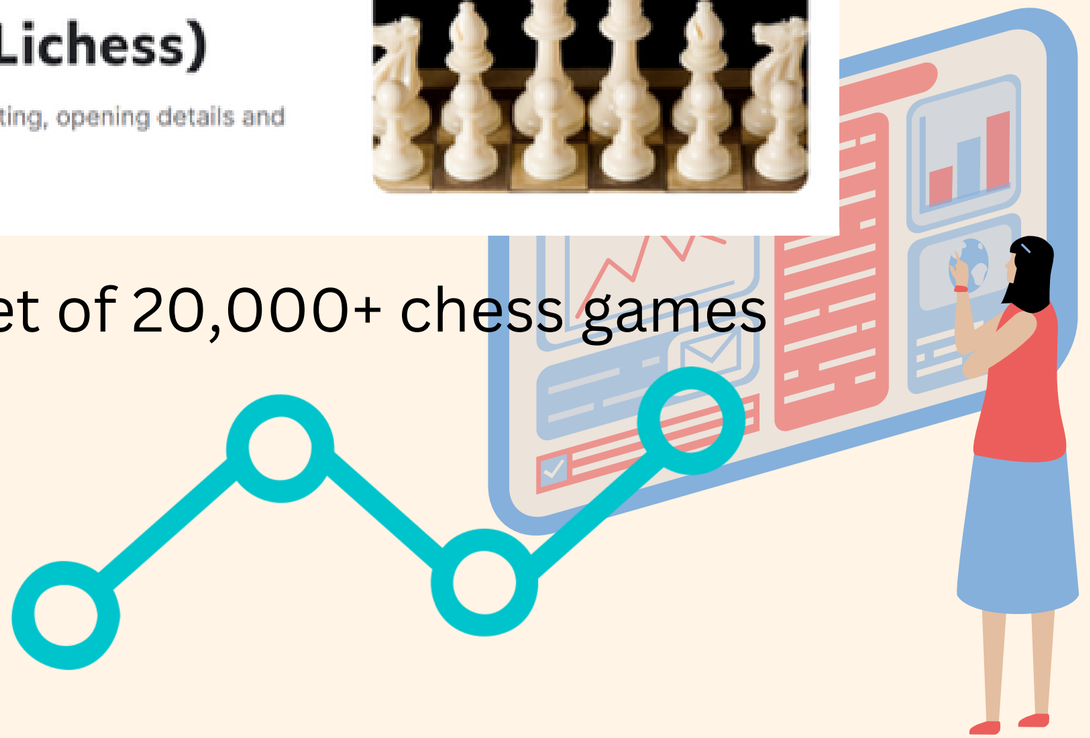
KAGGLE: CHESS GAME DATASET (LICHESS)



Downloaded from kaggle, a dataset of 20,000+ chess games

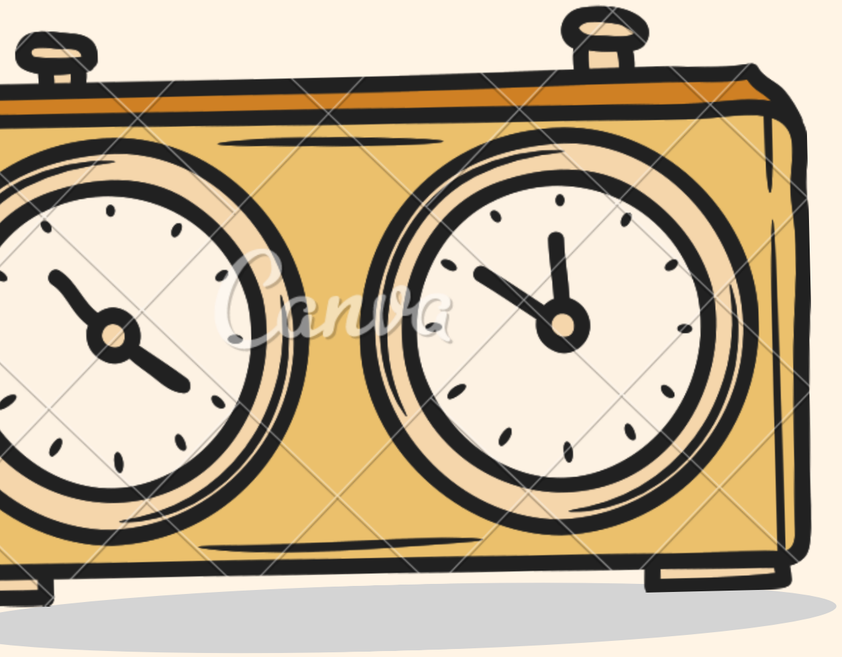


games.csv



EXPLAINING THE COLUMNS

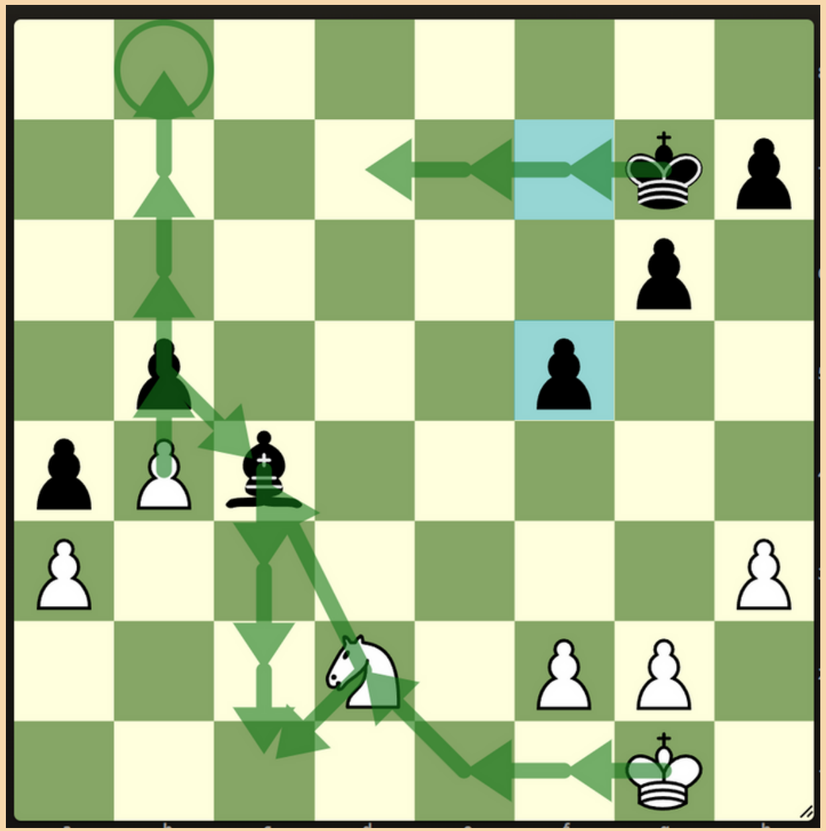
METHOD: DF.INFO()



AOO

```
RangeIndex: 20058 entries, 0 to 20057
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              20058 non-null  object
 1   rated           20058 non-null  bool
 2   created_at      20058 non-null  float64
 3   last_move_at    20058 non-null  float64
 4   turns           20058 non-null  int64
 5   victory_status  20058 non-null  object
 6   winner          20058 non-null  object
 7   increment_code  20058 non-null  object
 8   white_id        20058 non-null  object
 9   white_rating    20058 non-null  int64
 10  black_id        20058 non-null  object
 11  black_rating    20058 non-null  int64
 12  moves           20058 non-null  object
 13  opening_eco     20058 non-null  object
 14  opening_name    20058 non-null  object
 15  opening_ply     20058 non-null  int64
dtypes: bool(1), float64(2), int64(4), object(9)
```

PRACTICAL MOTIVATION

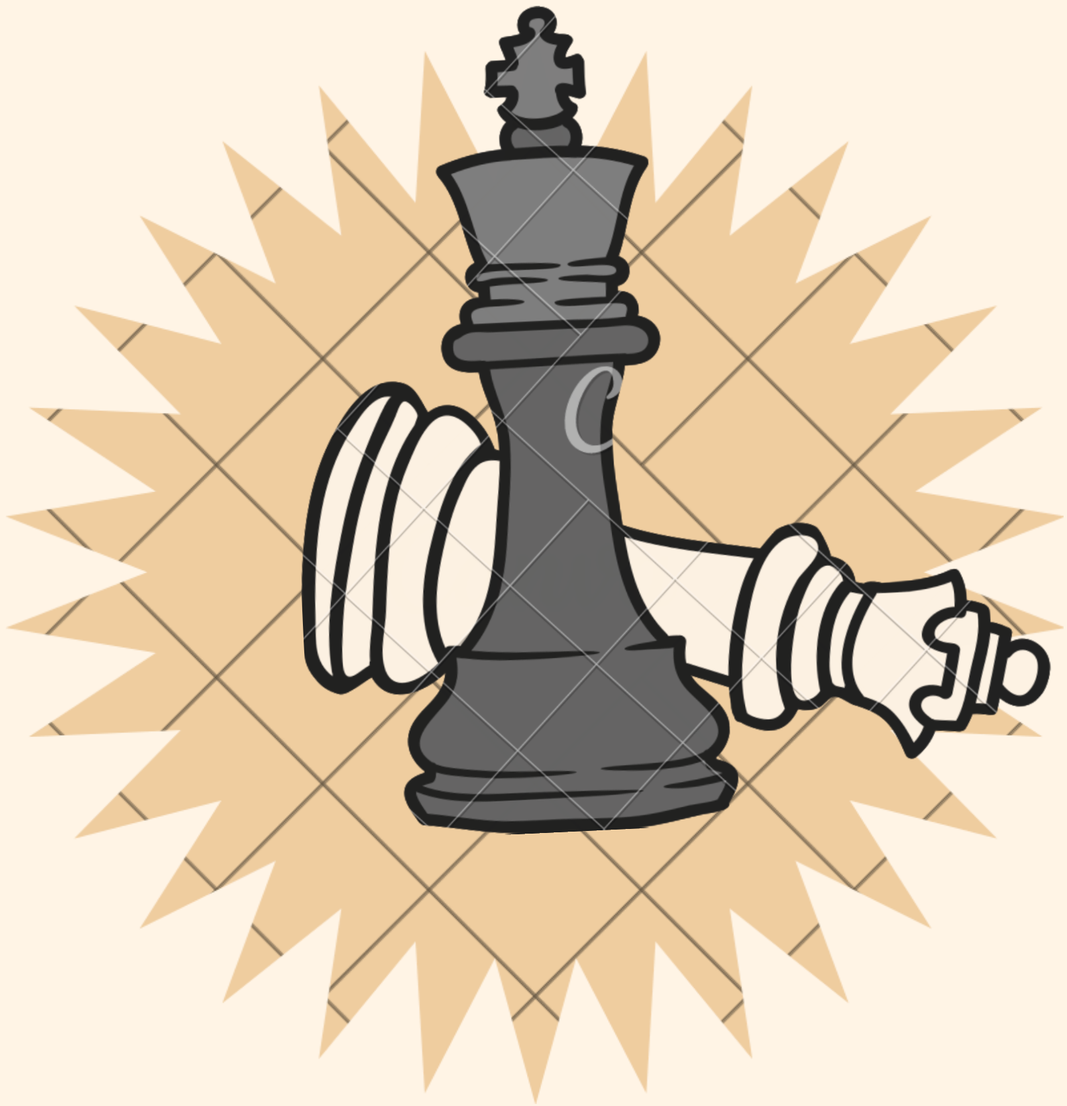


Chess:

- **Highly theoretical**
- **Strategic/ competitive.**
- **Can be studied -> higher chances of winning!**

What we want:

- **To understand where we stand**
- **To see if any opening type we should learn, playing black/white.**

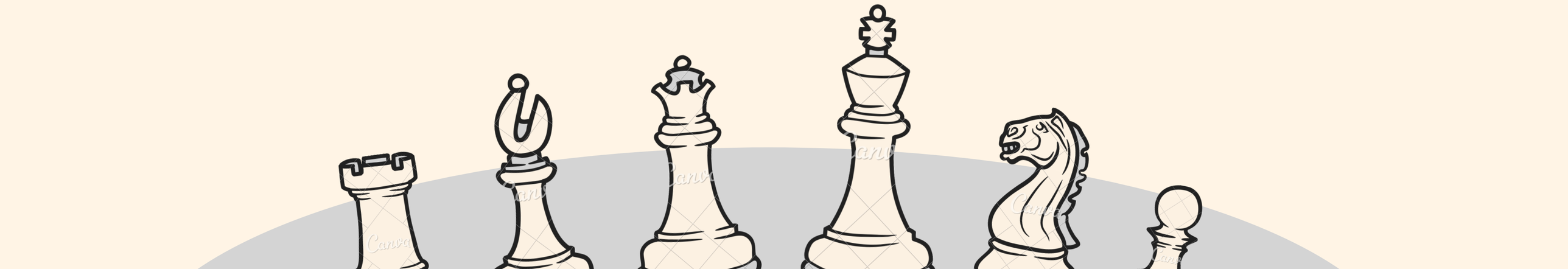


PROBLEM STATEMENT

How can we learn openings to increase our chances of winning?

EXPLORATORY ANALYSIS & DATASET CLEANING

- Preliminary Exploration & Cleaning
- Exploring Player Rating & Number of turns
- Removing games with players under 2000 ELO
- Removing games which end with a draw

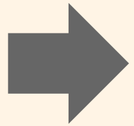


PRELIMINARY EXPLORATION & CLEANING

We first explored the data using basic tools to get an idea of the dataset we are dealing with.

We first renamed the column headers for **better readability**.

increment_code	white_id	white_rating	black_id	black_rating	moves	opening_eco	opening_name	opening_ply
15+2	bourgris	1500	a-00	1191	d4 d5 c4 c6 cxd5 e6 dxe6 fxe6 Nf3 Bb4+ Nc3 Ba5...	D10	Slav Defense: Exchange Variation	5
5+10	a-00	1322	skinnerua	1261	d4 Nc6 e4 e5 f4 f6 dxe5 fxe5 fxe5 Nxe5 Qd4 Nc6...	B00	Nimzowitsch Defense: Kennedy Variation	4
5+10	ischia	1496	a-00	1500	e4 e5 d3 d6 Be3 c6 Be2 b5 Nd2 a5 a4 c5 axb5 Nc...	C20	King's Pawn Game: Leonardis Variation	3

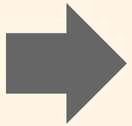


	Created At	Last Move At	Turns	White Rating	Black Rating	Opening Ply
count	2.005800e+04	2.005800e+04	20058.000000	20058.000000	20058.000000	20058.000000
mean	1.483617e+12	1.483618e+12	60.465999	1596.631868	1588.831987	4.816981
std	2.850151e+10	2.850140e+10	33.570585	291.253376	291.036126	2.797152
min	1.376772e+12	1.376772e+12	1.000000	784.000000	789.000000	1.000000
25%	1.477548e+12	1.477548e+12	37.000000	1398.000000	1391.000000	3.000000
50%	1.496010e+12	1.496010e+12	55.000000	1567.000000	1562.000000	4.000000
75%	1.503170e+12	1.503170e+12	79.000000	1793.000000	1784.000000	6.000000
max	1.504493e+12	1.504494e+12	349.000000	2700.000000	2723.000000	28.000000

PRELIMINARY EXPLORATION & CLEANING

We then dropped unnecessary columns such as **“Created At”** and **“Last Move At”** to enable us to focus on data necessary to us.

	Created At	Last Move At	Turns	White Rating	Black Rating	Opening Ply
count	2.005800e+04	2.005800e+04	20058.000000	20058.000000	20058.000000	20058.000000
mean	1.483617e+12	1.483618e+12	60.465999	1596.631868	1588.831987	4.816981
std	2.850151e+10	2.850140e+10	33.570585	291.253376	291.036126	2.797152
min	1.376772e+12	1.376772e+12	1.000000	784.000000	789.000000	1.000000
25%	1.477548e+12	1.477548e+12	37.000000	1398.000000	1391.000000	3.000000
50%	1.496010e+12	1.496010e+12	55.000000	1567.000000	1562.000000	4.000000
75%	1.503170e+12	1.503170e+12	79.000000	1793.000000	1784.000000	6.000000
max	1.504493e+12	1.504494e+12	349.000000	2700.000000	2723.000000	28.000000



	Turns	White Rating	Black Rating	Opening Ply
count	20058.000000	20058.000000	20058.000000	20058.000000
mean	60.465999	1596.631868	1588.831987	4.816981
std	33.570585	291.253376	291.036126	2.797152
min	1.000000	784.000000	789.000000	1.000000
25%	37.000000	1398.000000	1391.000000	3.000000
50%	55.000000	1567.000000	1562.000000	4.000000
75%	79.000000	1793.000000	1784.000000	6.000000
max	349.000000	2700.000000	2723.000000	28.000000

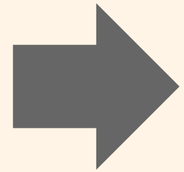
PRELIMINARY EXPLORATION & CLEANING

For columns such as the time control for games, it was in the format **x+y**:

- x= number of minutes allocated to each player for the game
- y= number of seconds added per move

There were many different combinations of time controls. Therefore, we classified them into **“Bullet”, “Blitz”, “Rapid”** and **“Classical”** to **simplify** what we were looking at.

Increment Code
15+2
5+10
5+10
20+0
30+3
...
10+10
10+0
10+0
10+0
10+0

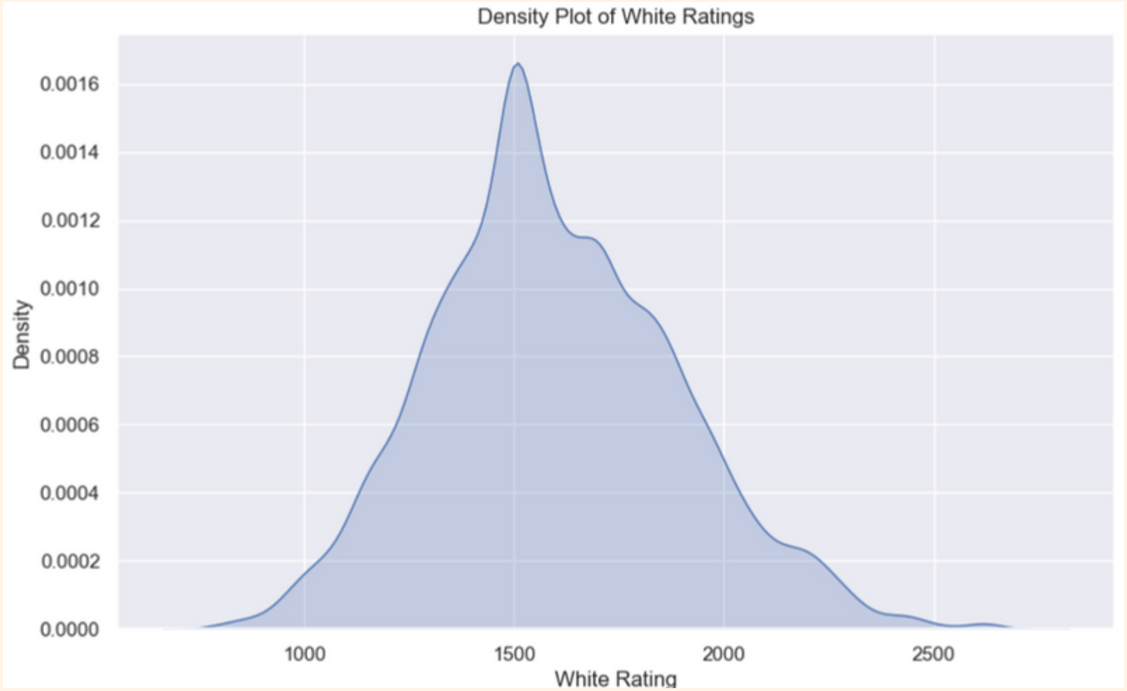


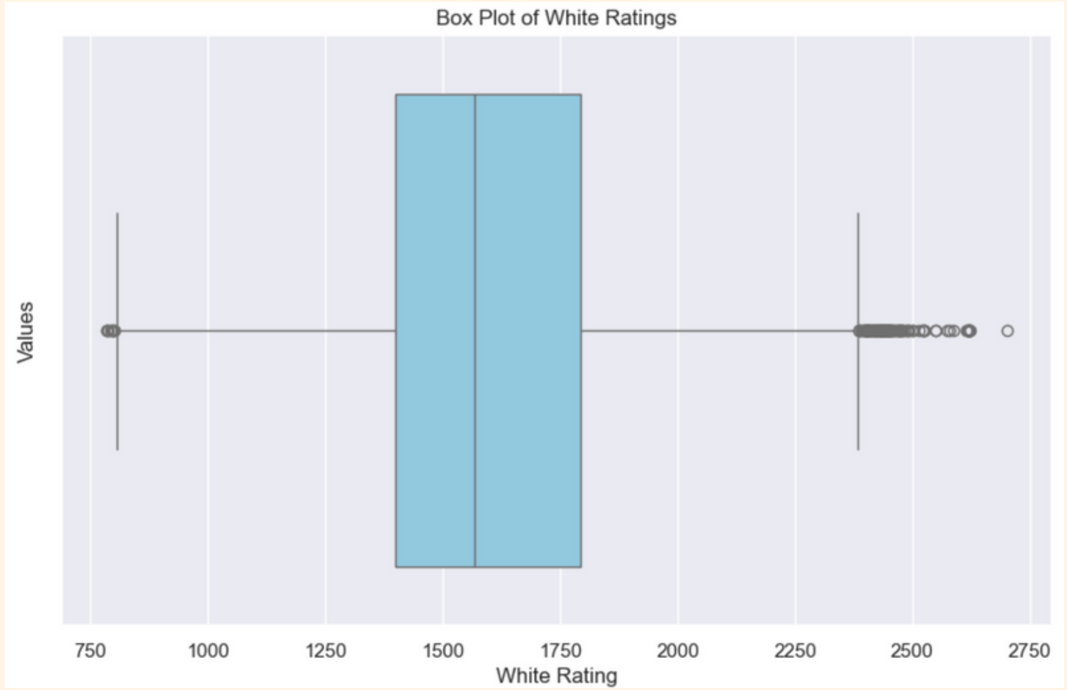
	Time Control
0	Rapid
1	Blitz
2	Blitz
3	Rapid
4	Rapid
...	...
20053	Blitz
20054	Blitz
20055	Blitz
20056	Blitz
20057	Blitz

PLAYER RATING

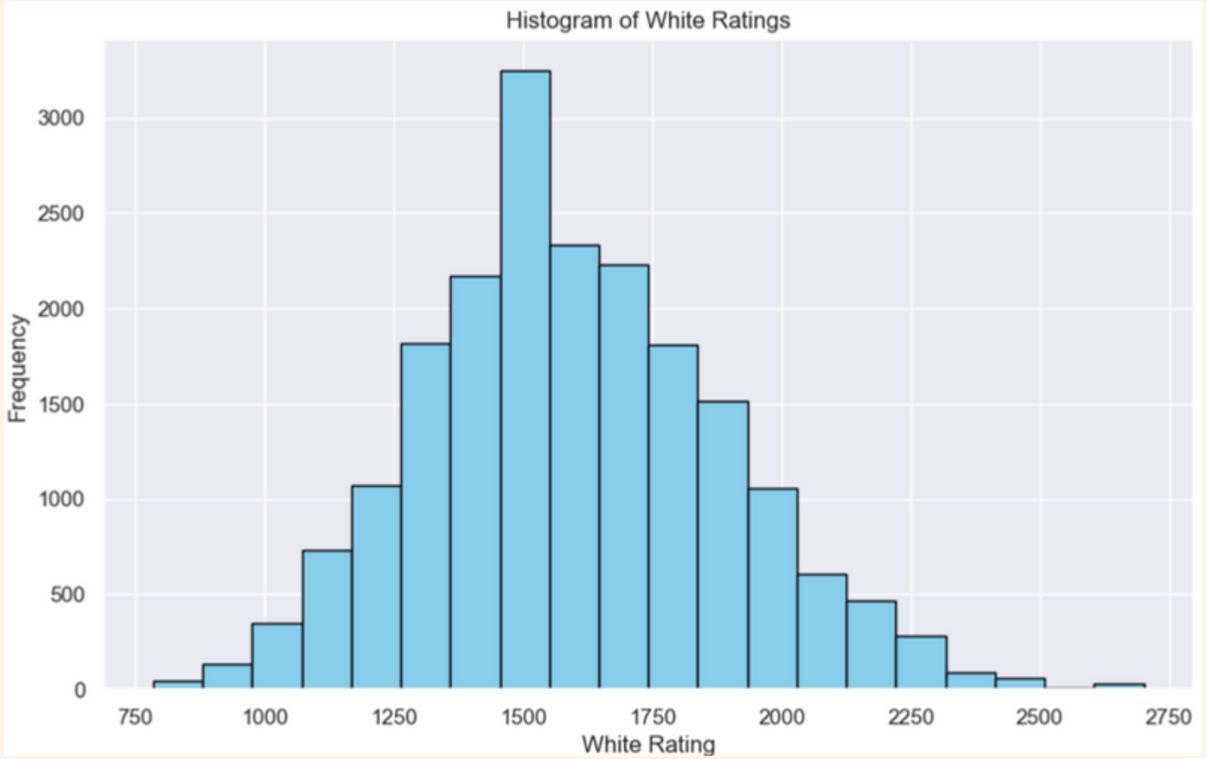
We then plotted the distribution of different variables for a visual representation.

We chose white player ratings as a reference. Here is the density plot and box plot.





PLAYER RATING



We separated the ratings into different bins to plot a histogram as well.

We find out of **20,000+** games:

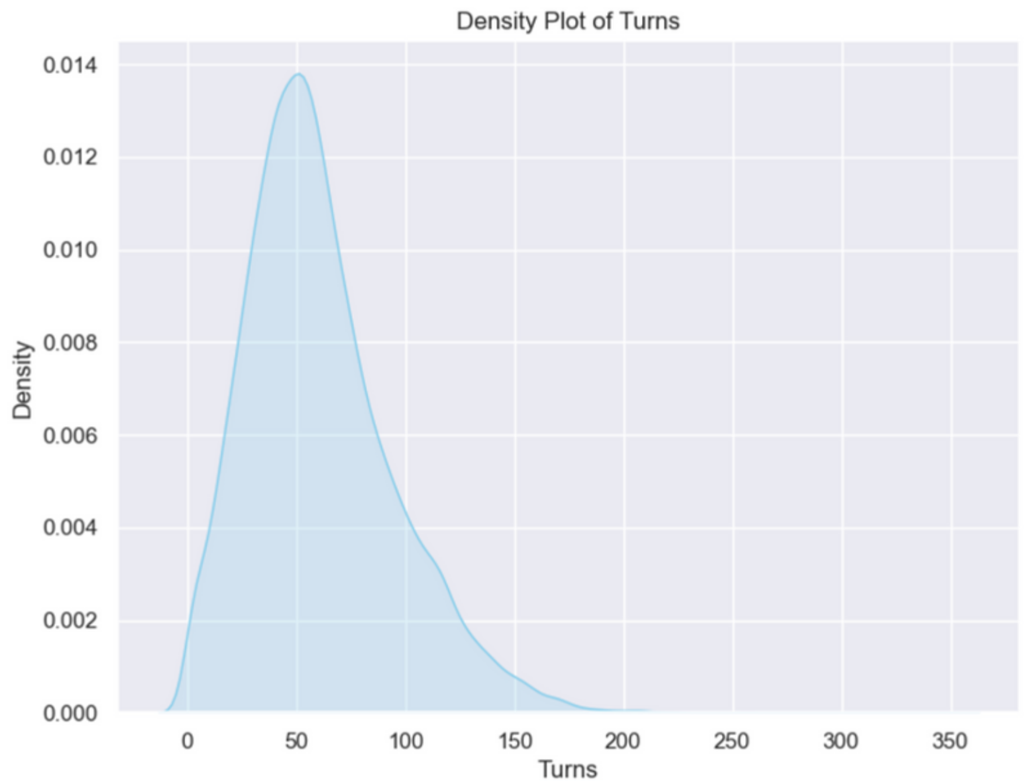
- Median = **1567**
- Skewness = **0.3007**
- Outliers = **135** games

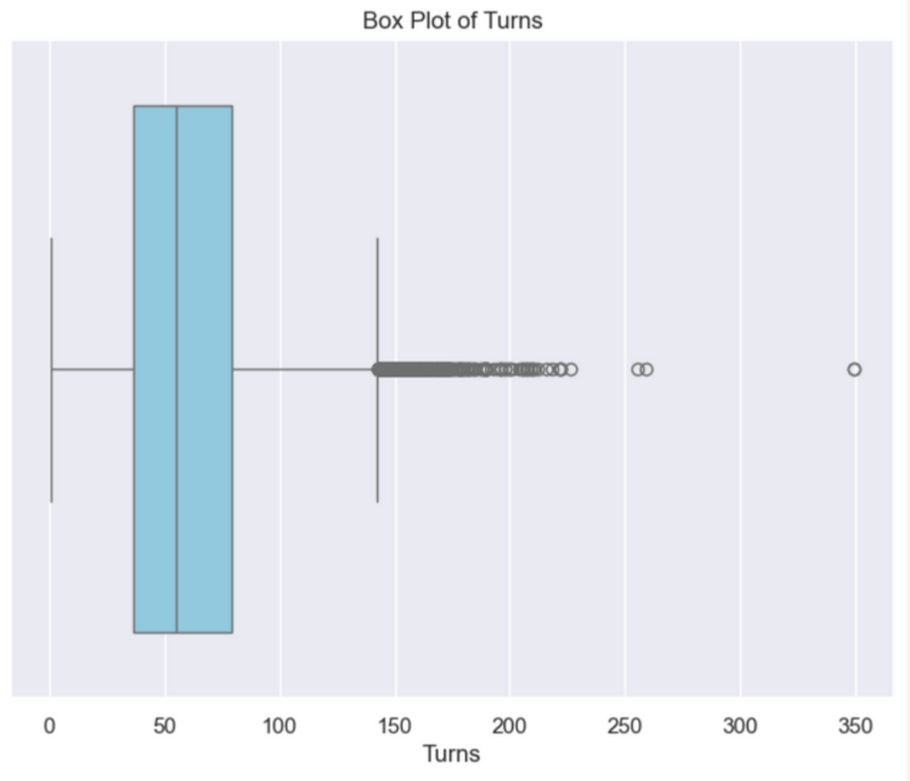
This shows that the distribution is fairly symmetrical.

Density plot shows the largest group of player's ratings is around 1500 ELO.

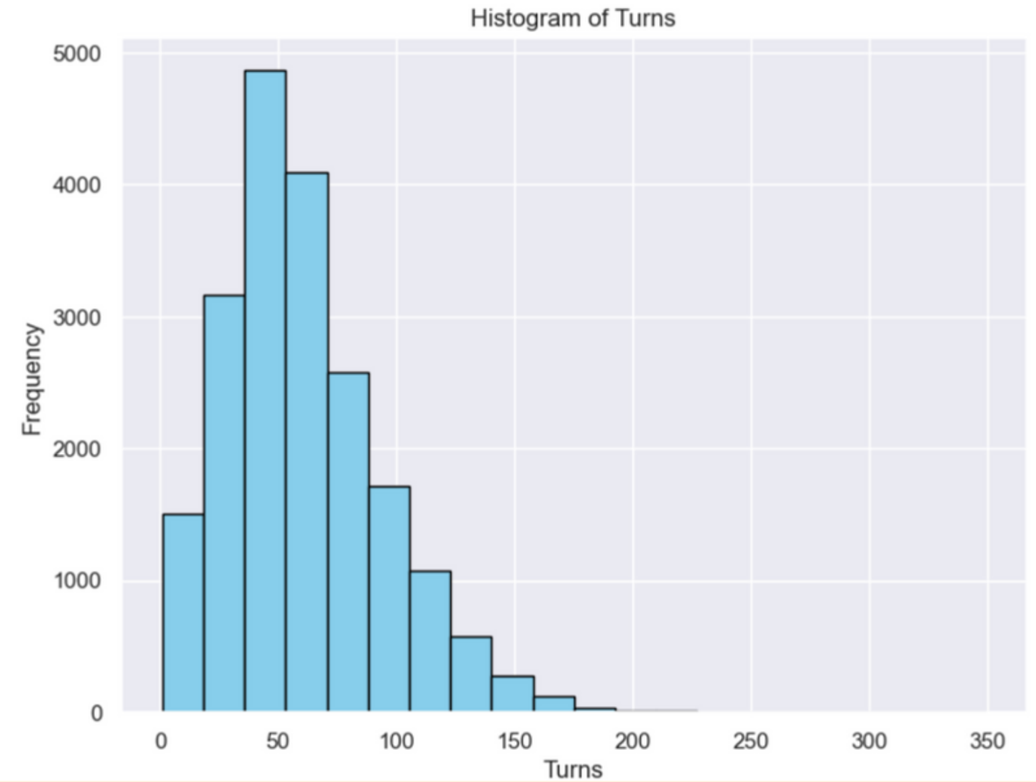
NUMBER OF TURNS

We then plotted the same plots for the number of turns made in the game.





NUMBER OF TURNS



We find out of **20,000+** games:

- Median = **55**
- Skewness = **0.8972**
- Outliers = **428** games

This shows that the distribution is positively skewed.

Density plot tells us that most games end at around 50 turns.

CLEANING THE DATASET

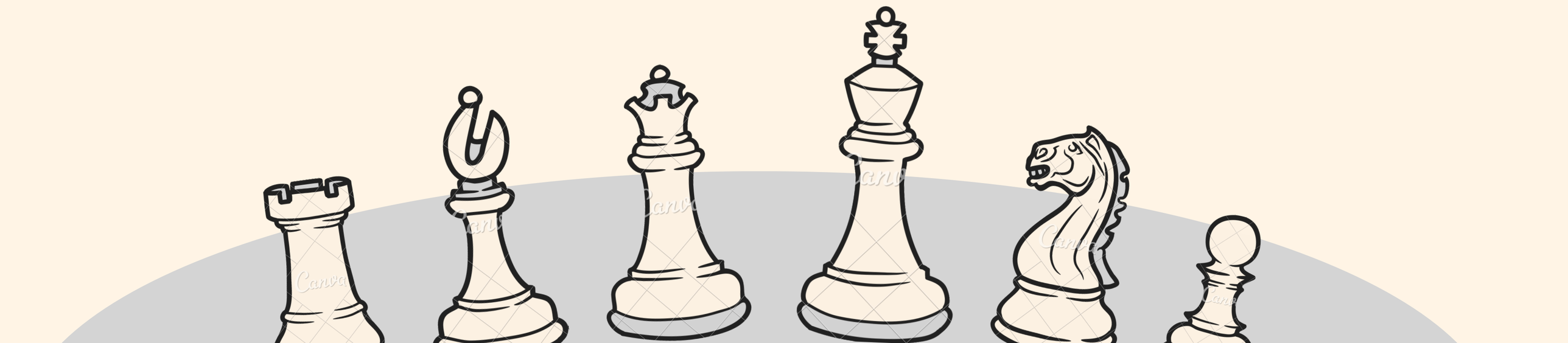
To prepare our data for machine learning, we:

- **Removed** all games where both White and Black Elo rating are **below 2000**.
  - We only consider higher rated players as we want to analyze the effectiveness of opening strategies. Lower rated players may not play strategically.
- **Removed** games which ended in **draw**.
  - We only want to consider how likely a player wins. Therefore, we remove games that end in a draw.

```
Number of people with ratings 2000 and above for both white and black players: 857
Number of rows where the winner is not a draw: 776
```

MACHINE LEARNING

Binary Classification Tree



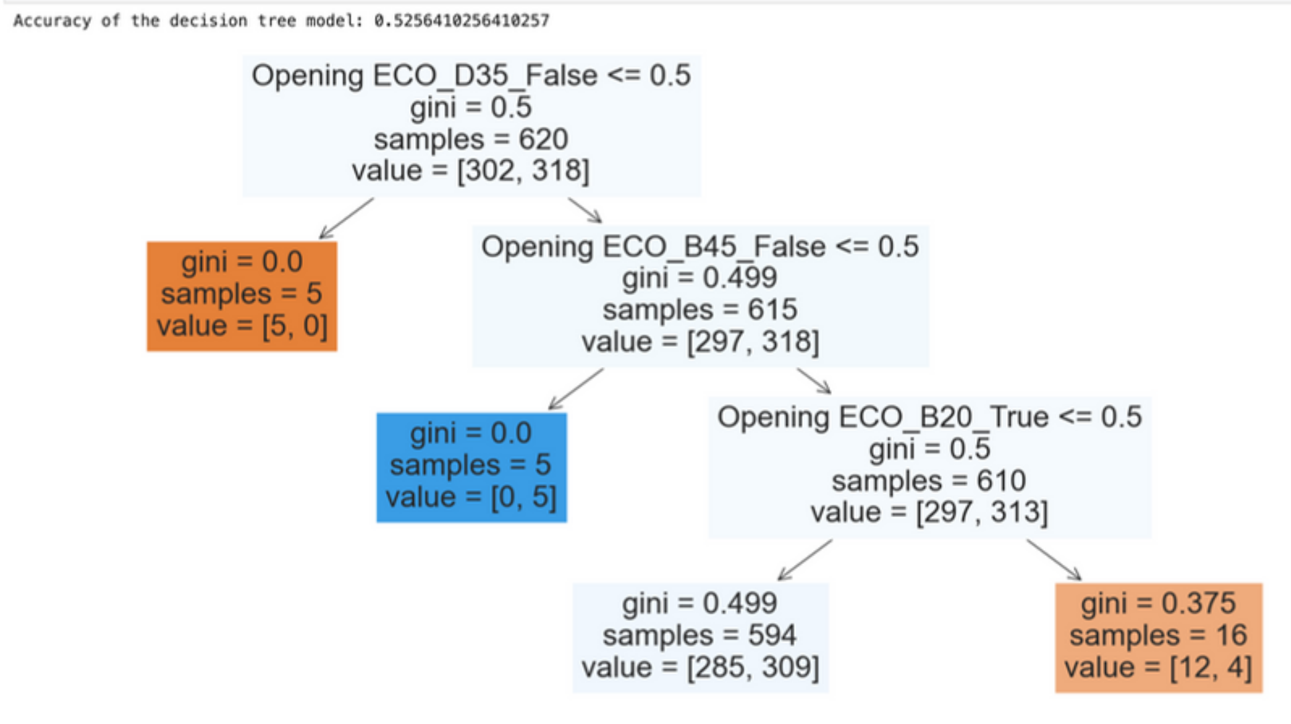
STEPS BEFORE BINARY CLASSIFICATION TREE

1. Convert the column ‘Winner’ to make a boolean variable

2. Do one hot encoding for Opening ECO

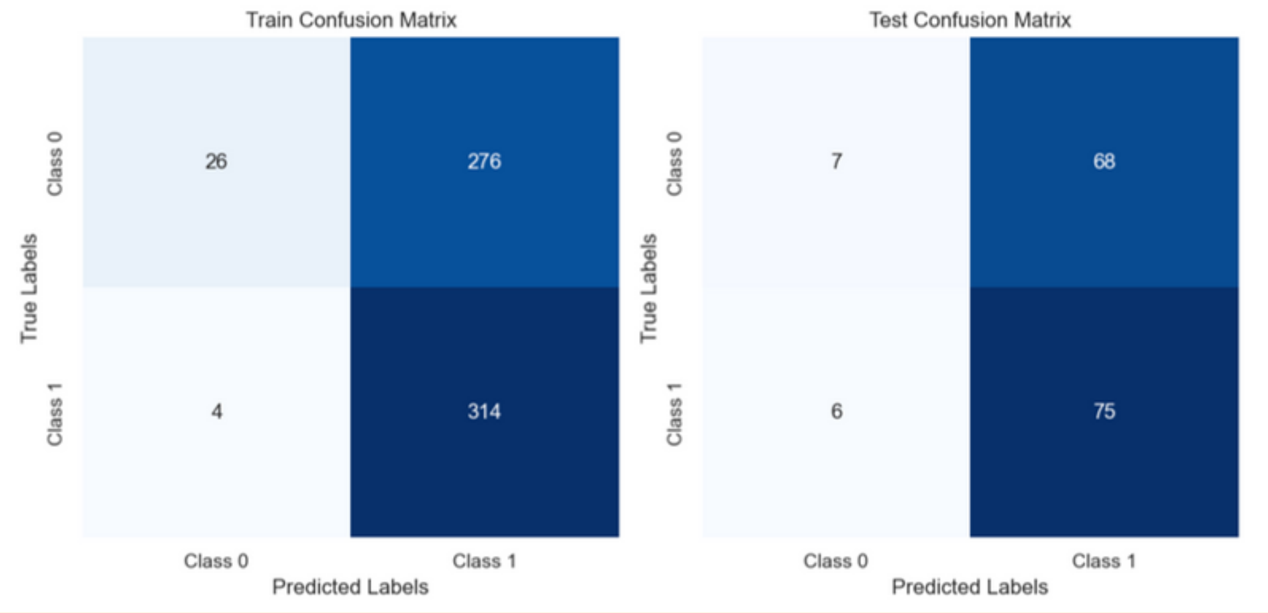
3. Remove non-numeric and boolean, as well as

BINARY CLASSIFICATION TREE



- We can tell different opening types favours different sides (Black or White)
- In the classification tree, values = [a,b] where a is the number of games that black won and b is the number of games white won, in the opening type.

EVALUATING BINARY CLASSIFICATION TREE



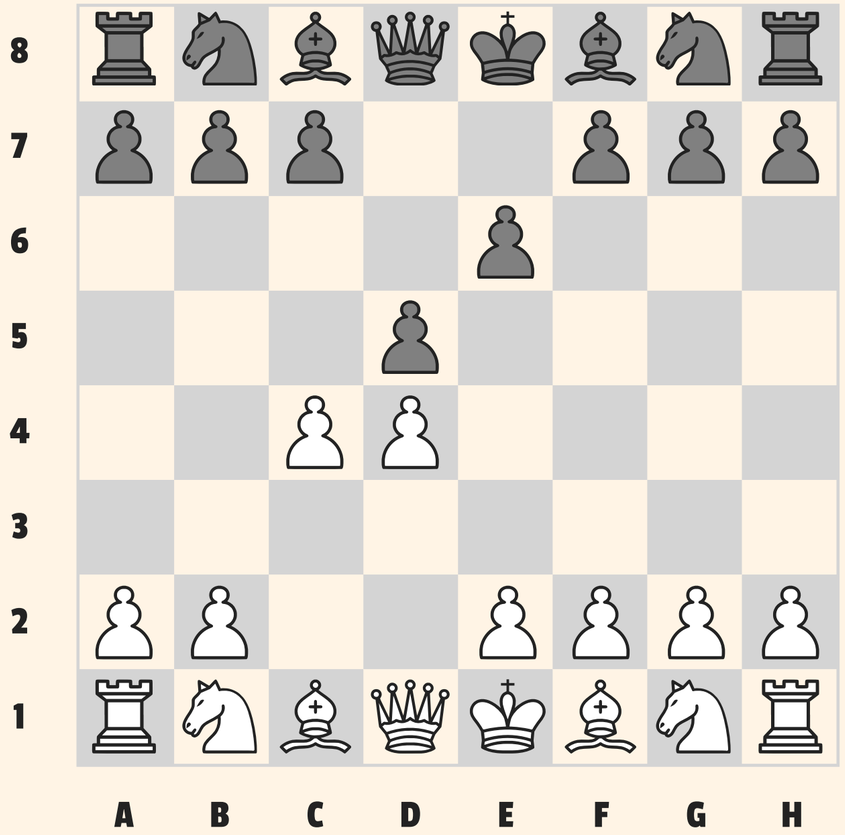

```
Train Set Metrics:
True Positive Rate (Sensitivity): 0.9874213836477987
False Positive Rate: 0.9139072847682119
True Negative Rate (Specificity): 0.08609271523178808
False Negative Rate: 0.012578616352201259

Test Set Metrics:
True Positive Rate (Sensitivity): 0.9259259259259259
False Positive Rate: 0.9066666666666666
True Negative Rate (Specificity): 0.09333333333333334
False Negative Rate: 0.07407407407407407
```

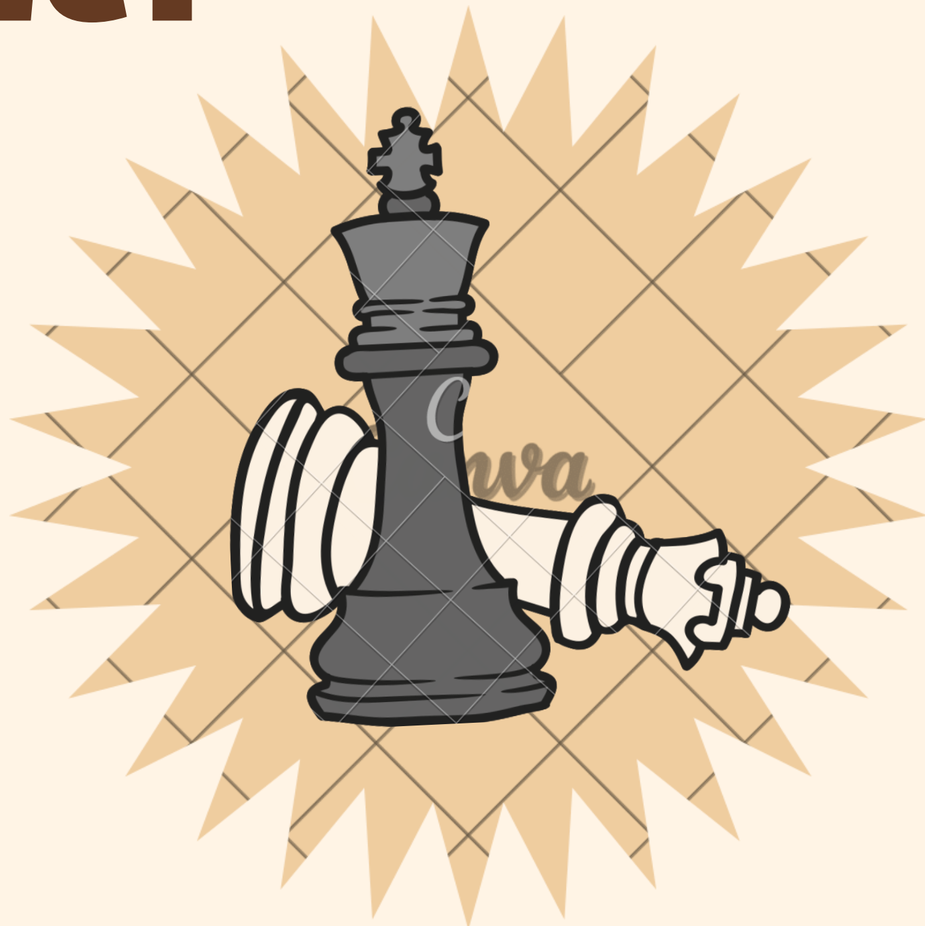
THE COMPARISON

Sale Price - Continuous integer ($), Predictor variable

Central Air - Boolean (Y/N), Response Variable

OUR PROJECT

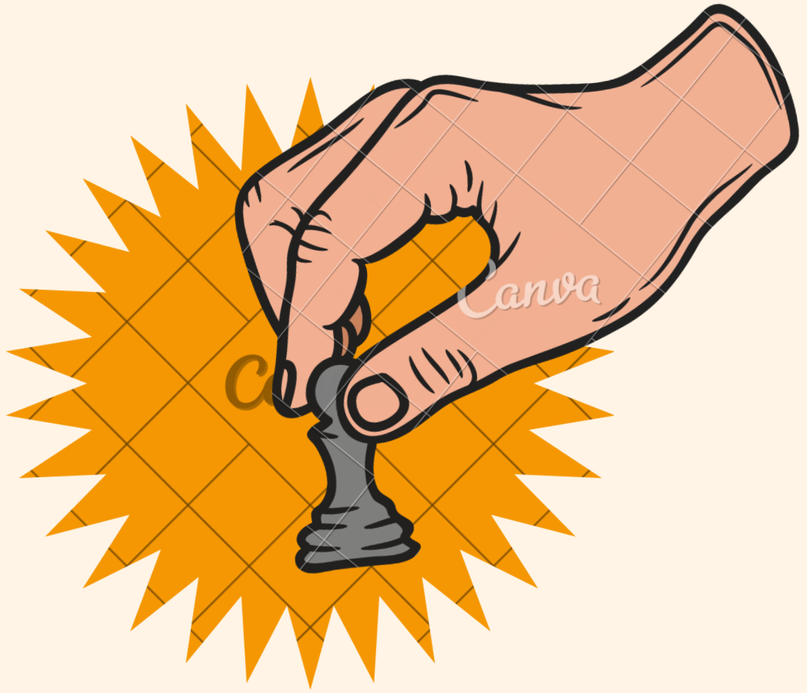




Opening ECO – One hot encoded (222) Boolean (T/F), Predictor variable

Is White Winner – Boolean (Y/N), Response Variable

OUTCOME



It is successful in telling us which opening favours which side.

Test and train matrix:

High true positive rate

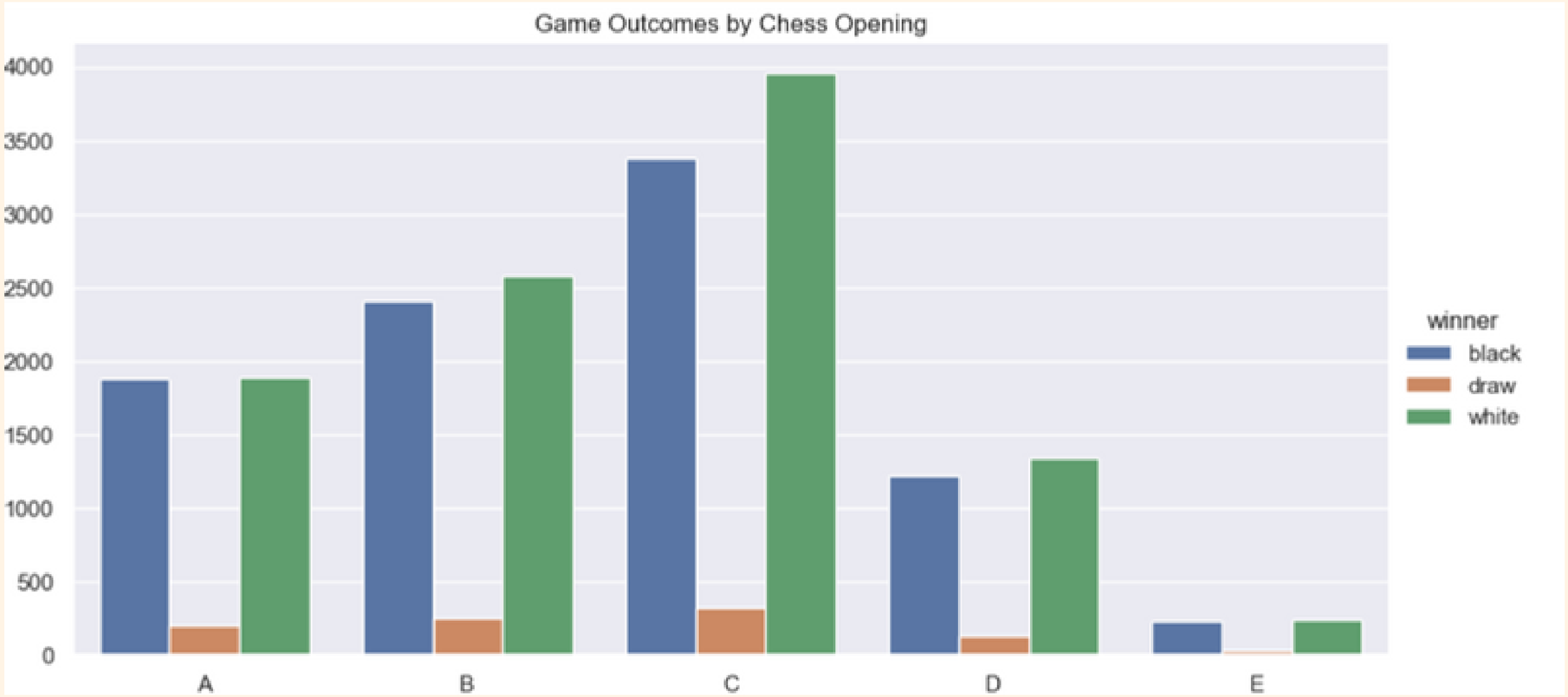
High false positive rate too

Implications:

- **Mostly true, but also a lot of falsely classified as true.**
- **False and true classification = Is White Winner.**
- **Most of the time, the algorithm predicts that white wins.**
- **Half of the time it is correct, half of the time, it predicts wrongly.**

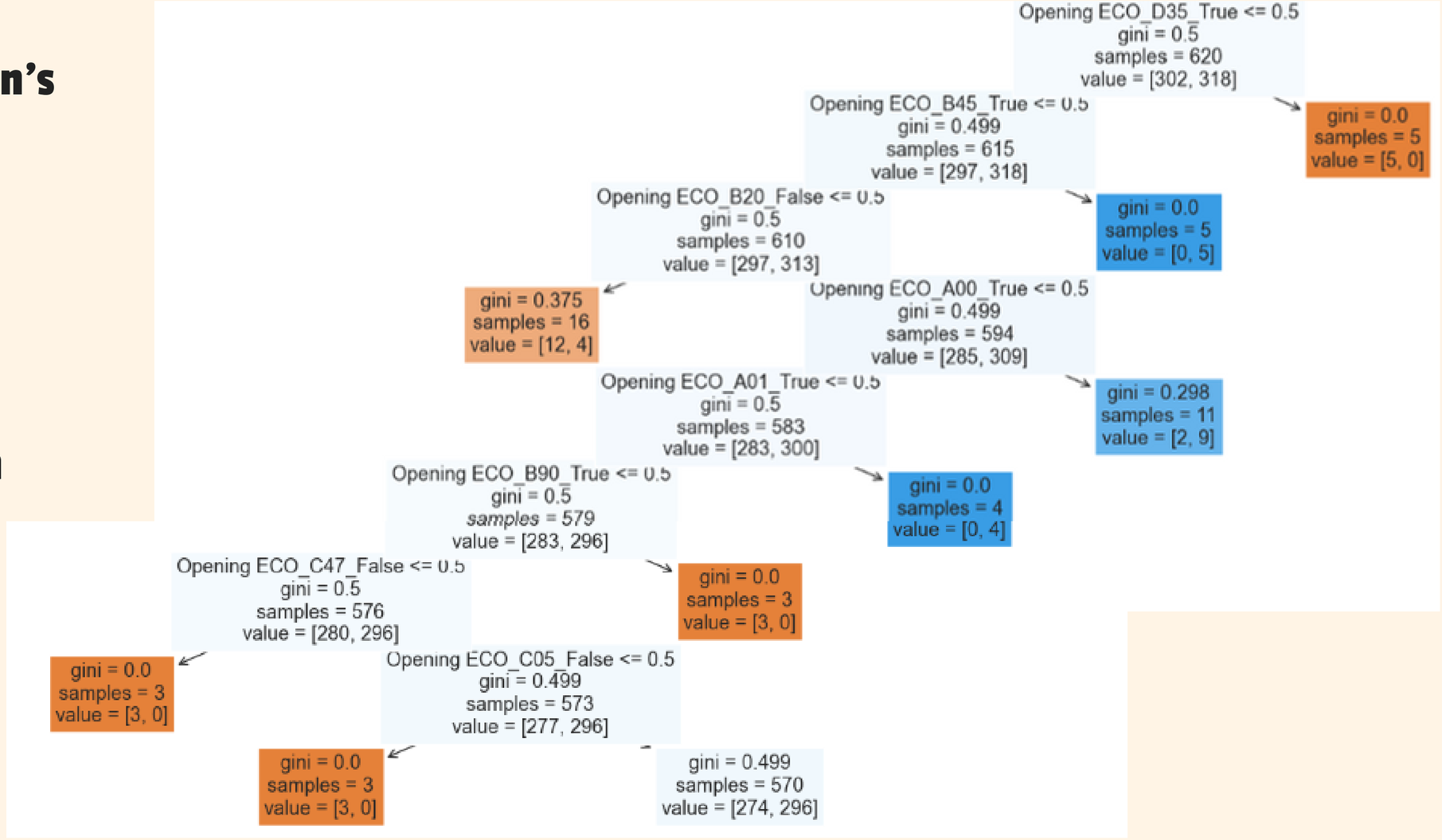
Essentially, white or black, both have equal chances of winning over all openings.

Proof


Game Outcomes by Chess Opening
4000
3500
3000
2500
2000
1500
1000
500
0
A
B
C
D
E
winner
black
draw
white

The binary classification tree

- **Opening ECO D35 = Queen's gambit favours black to win**
- **B45 = Sicilian defense, Taimanov variation favours white to win**
- **More can be found down the binary classification tree**
- **This is if players have studied the opening theory by heart before playing the game.**



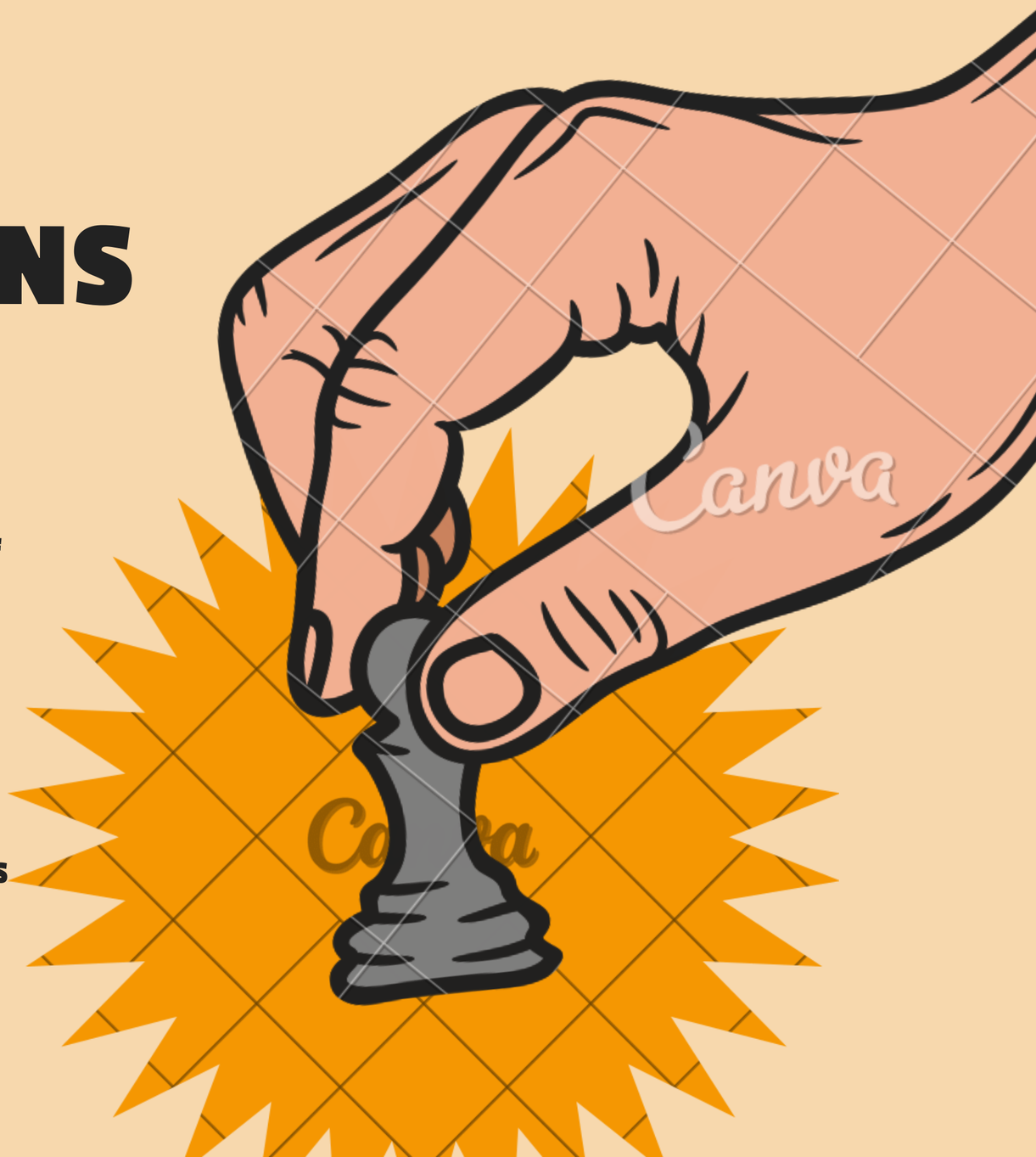
FINAL RECOMMENDATIONS

In reality, we can never get our desired opening in 100% of the games we play.

The sequence of black and white moves both affect the course of the opening.

That is to say, we should learn a variety of openings that branch off an umbrella of openings (The first letter)

But learning a solid foundation for openings can prove to help us win more games, knowing whether it is advantageous for us to play.



THANK YOU FOR LISTENING

LET'S PLAY CHESS!