Notebook 1: Establishing Data

1. Introduction

This notebook introduces the chess openings dataset, describes its source, and provides an overview of its features. This forms the foundation for further exploratory data analysis (EDA) and modeling.

---

2. Data Acquisition

Source

The dataset was downloaded from Kaggle

- **Link:** https://www.kaggle.com/datasets/alexandrelemercier/all-chess-openings
- **Format:** CSV

How to Get the Data

1. Navigate to the dataset link.
2. Click **Download.**
3. Save the CSV file as `chess_openings.csv` in your working directory.

---

3. Who Produced the Data

The dataset was produced by chess enthusiasts and/or platform users who tracked nearly 3.5 million games from over 1800 openings over time. It contains aggregated statistics for various openings including:

- Opening name
- Number of games played
- Player performance metrics
- Move sequences

It is intended for analysis of opening effectiveness and popularity.

In [2]:
```
import pandas as pd
df = pd.read_csv("chess_openings.csv")
df.head()
```

Out[2]:

	Unnamed: 0	Opening	Colour	Num Games	ECO	Last Played	Perf Rating	Avg Player	Player Win %	Draw %	...
0	0	Alekhine Defense, Balogh Variation	white	692	B03	2018-06-22	2247	2225	40.8	24.3	...
1	1	Alekhine Defense, Brooklyn Variation	black	228	B02	2018-06-27	2145	2193	29.8	22.4	...
2	2	Alekhine Defense, Exchange Variation	white	6485	B03	2018-07-06	2244	2194	40.8	27.7	...
3	3	Alekhine Defense, Four Pawns Attack	white	881	B03	2018-06-20	2187	2130	39.7	23.2	...
4	4	Alekhine Defense, Four Pawns Attack, Fianchett...	black	259	B03	2018-05-20	2122	2178	37.8	21.2	...

5 rows × 26 columns

Column Name	Description
Opening	Name of the chess opening (e.g., Sicilian Defense)
Colour	Player color (white or black)
Num Games	Number of games recorded for this opening and color
ECO	ECO (Encyclopedia of Chess Openings) code
Last Played	Date the opening was last played in this dataset
Perf Rating	Performance rating of the player using this opening
Avg Player	Average rating of opponents
Player Win %	Win percentage for the player using this opening
Draw %	Percentage of games ending in a draw
Opponent Win %	Win percentage for the opponent

Column Name	Description
Moves	Sequence of moves in standard algebraic notation
moves_list	Moves as a Python list for analysis
move1w, move1b ...	Individual moves by turn (white/black)
White_win%	Overall white win percentage for this opening
Black_win%	Overall black win percentage
White_odds	Odds of white winning (numerical)
White_Wins	Number of games white won
Black_Wins	Number of games black won

In [4]:
```
# Basic info
print(df.info())

# Summary statistics
print(df.describe())

# Check for missing values
print(df.isnull().sum())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1884 entries, 0 to 1883
Data columns (total 26 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Unnamed: 0      1884 non-null   int64  
 1   Opening         1884 non-null   object 
 2   Colour          1884 non-null   object 
 3   Num Games       1884 non-null   int64  
 4   ECO             1884 non-null   object 
 5   Last Played     1884 non-null   object 
 6   Perf Rating     1884 non-null   int64  
 7   Avg Player      1884 non-null   int64  
 8   Player Win %    1884 non-null   float64
 9   Draw %          1884 non-null   float64
 10  Opponent Win %  1884 non-null   float64
 11  Moves           1884 non-null   object 
 12  moves_list      1884 non-null   object 
 13  move1w          1884 non-null   object 
 14  move1b          1869 non-null   object 
 15  move2w          1814 non-null   object 
 16  move2b          1744 non-null   object 
 17  move3w          1628 non-null   object 
 18  move3b          1501 non-null   object 
 19  move4w          1340 non-null   object 
 20  move4b          1186 non-null   object 
 21  White_win%      1884 non-null   float64
 22  Black_win%      1884 non-null   float64
 23  White_odds      1884 non-null   float64
 24  White_Wins      1884 non-null   float64
 25  Black_Wins      1884 non-null   float64
dtypes: float64(8), int64(4), object(14)
memory usage: 382.8+ KB
None
        Unnamed: 0     Num Games  Perf Rating   Avg Player  Player Win %  \
count  1884.000000   1884.000000  1884.000000  1884.000000   1884.000000   
mean    941.500000   1846.019108  2235.945860  2236.531847     35.159395   
std     544.008272   2739.103462   135.260392   127.723711      9.077139   
min       0.000000    100.000000  1583.000000  1577.000000      7.500000   
25%     470.750000    314.750000  2157.000000  2166.000000     28.900000   
50%     941.500000    788.500000  2252.500000  2255.000000     35.100000   
75%    1412.250000   2225.000000  2329.000000  2326.000000     41.125000   
max    1883.000000  22482.000000  2536.000000  2492.000000     77.600000   

            Draw %  Opponent Win %   White_win%   Black_win%   White_odds  \
count  1884.000000     1884.000000  1884.000000  1884.000000  1884.000000   
mean     29.914066       34.928715    39.745701    30.342410     1.448725   
std       8.043043        9.180450     7.671108     7.976305     0.673991   
min       4.000000        6.700000    13.600000     6.700000     0.308642   
25%      24.500000       28.900000    34.800000    25.100000     1.037277   
50%      29.400000       34.650000    39.100000    29.900000     1.325008   
75%      34.625000       40.525000    44.000000    35.000000     1.674116   
max      68.500000       77.500000    77.600000    64.800000     9.810127   

        White_Wins   Black_Wins  
count  1884.000000  1884.000000  
```

```
mean    708.835970   557.051955
std    1037.027669   866.788831
min      21.000000     8.946000
25%     124.740500    91.026250
50%     310.274000   230.952000
75%     824.243000   651.262750
max    8295.858000  8700.534000
Unnamed: 0          0
Opening             0
Colour              0
Num Games           0
ECO                 0
Last Played         0
Perf Rating         0
Avg Player          0
Player Win %        0
Draw %              0
Opponent Win %      0
Moves               0
moves_list          0
move1w              0
move1b             15
move2w             70
move2b            140
move3w            256
move3b            383
move4w            544
move4b            698
White_win%          0
Black_win%          0
White_odds          0
White_Wins          0
Black_Wins          0
dtype: int64
```

In [ ]: