We have two datasets, one that contains the general information of the games (around 6M entries) and the other one contains the moves for each game. We split it to be able to work properly since the dataset was too large.

* Data analysis and preparation of dataset 1

PROCESSING

1. Create ID to be able to merge the datasets afterwards
2. Rename AN column to Moves
3. Check if there are nans
4. Create the Moves dataset with ID and Moves to process the text for the embeddings
5. Drop the columns that we do not want (afegir totes les columnes q s’han llevat al codi)
6. Check if there are any nans
7. Data normalization
   1. Normalize all of the event types to the four categories (Bullet, Classical, Blitz and Correspondence)
   2. Drop the \* results because we do not know what they are

DATA ANALYSIS

[**1. General game results 2**](#_3bg8qf3e7wmo)

[**2. termination and who wins more 3**](#_gpsnkaooe98f)

[**3. Difference of elo points in each game 4**](#_pq506u4az3ot)

[**4. Elo correlations (bad) 6**](#_363dkors5o1d)

[**5. Division by elo level 6**](#_qi1q5xoedif7)

[**6. results by elo level 7**](#_9w5e9y6oojx5)

[**7. average game results for each elo level (NOT GOOD) 8**](#_tfgku06db1pd)

[**8. opening effectiveness by color (only used once openings too) 9**](#_f0b2lje9hkdt)

[**9. top 10 openings (frequency) 10**](#_fkjtil518ksz)

[**10. openings used by elo level (frequency) 11**](#_yoto8mrzhh9q)

[**11. opening effectiveness from the most 46 used openings (by color) 12**](#_fiii4y57rqmi)

[**12. openings by elo and effectiveness (winrate) 14**](#_d677lbk1cjeg)

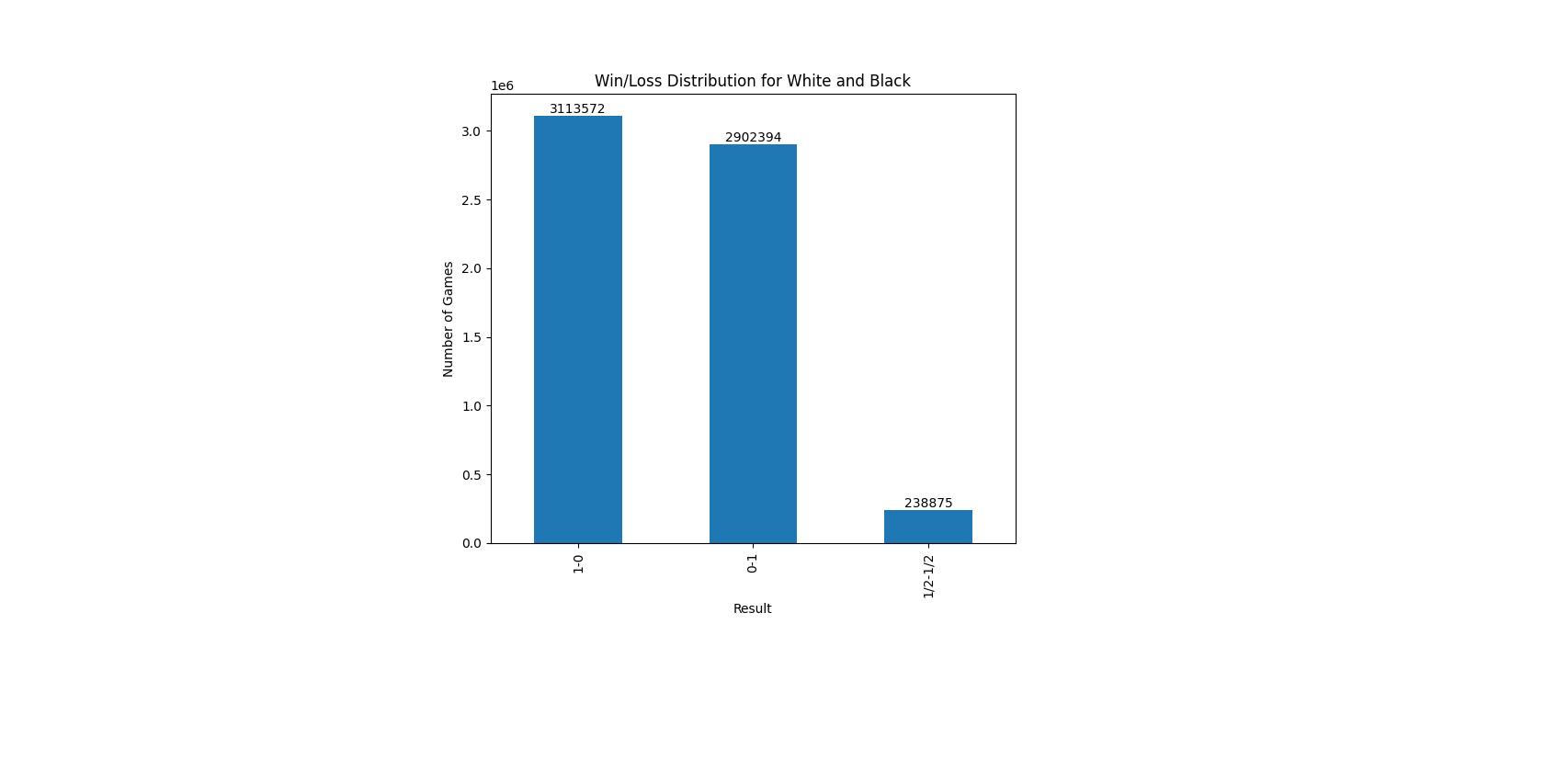
[**13. event types + by elo 15**](#_cgyjq43pyybg)

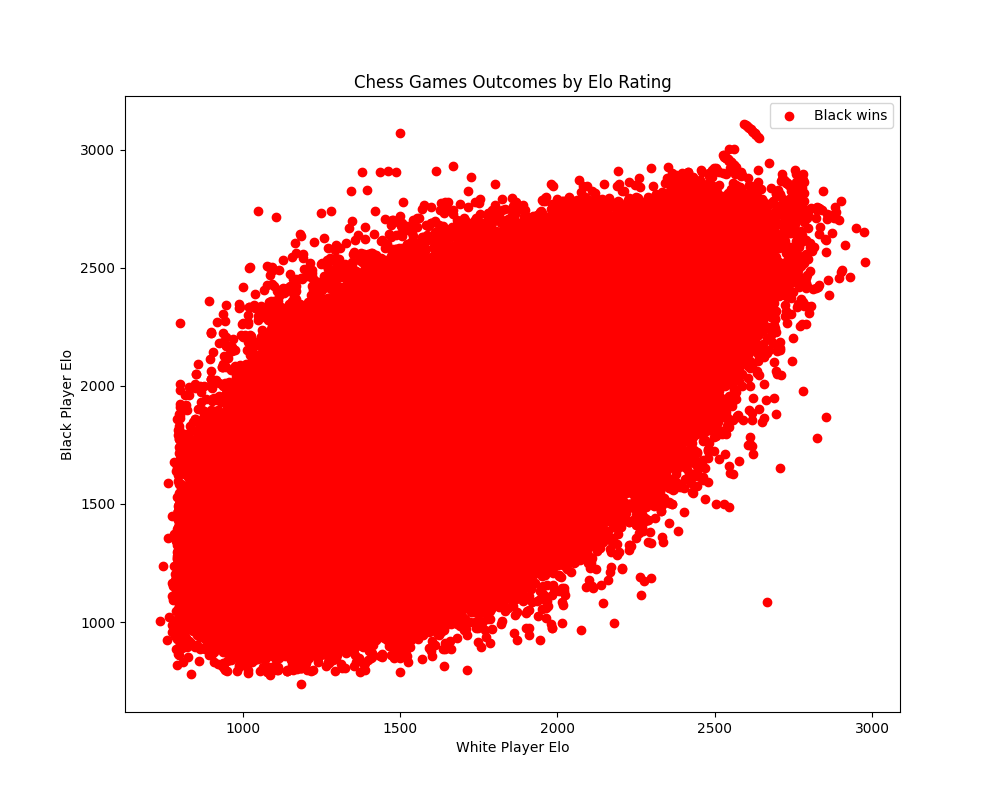
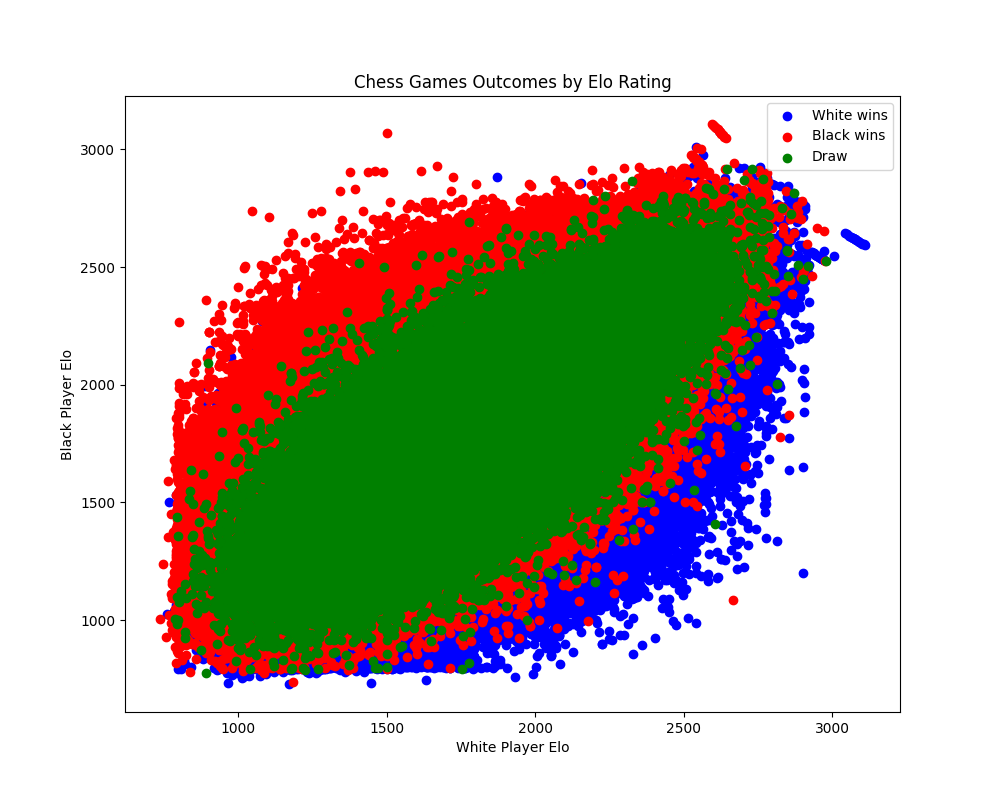
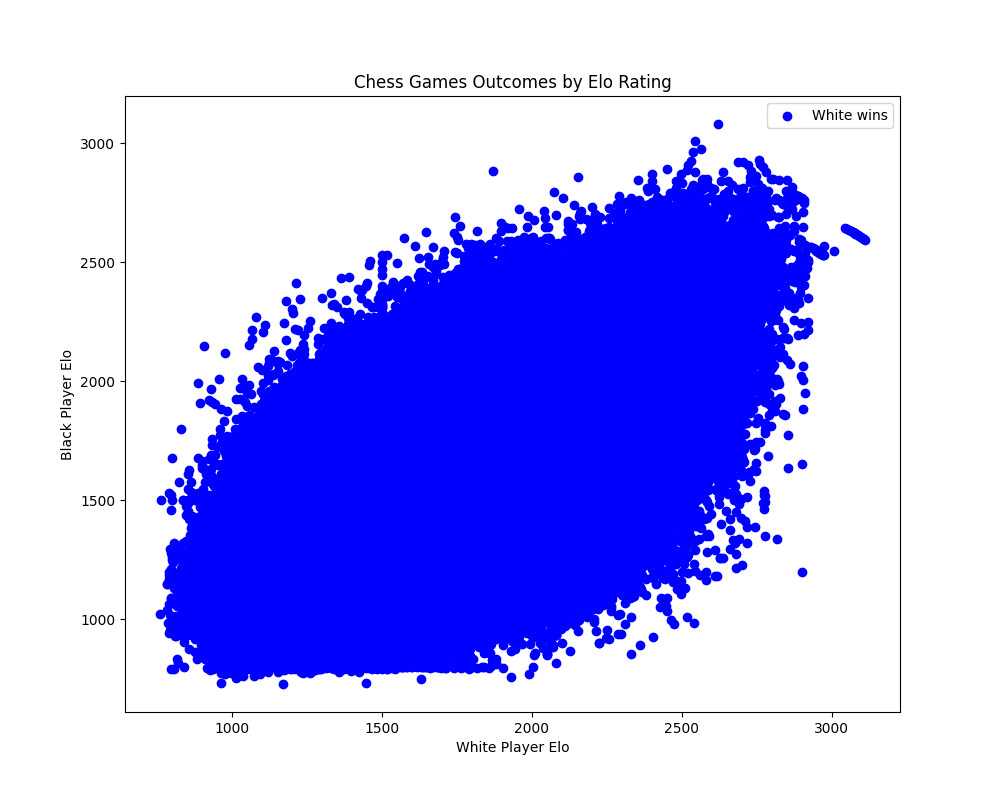
[**14. clustering 16**](#_2jktaixcgi2y)

\*\* all of the opening analysis could have been used if our model relied on openings to predict

### General game results

Whites won more games than black players.





We can see that when black elo is higher, black players win more and when white elo is higher, white players win more (as expected).

### termination and who wins more

### 

* More whites win because of abandonement

print(abandoned\_winloose)

'''

1-0 13281 more whites win because of abandonement

0-1 7

Name: count, dtype

'''

* All infractions have been committed by black players so win players won

'''1-0 128 all that have done infractions are black players'''

* More whites win due to normal wins
* print(normal\_winloose)
* '''1-0 2116730 more whites win
* 0-1 1926100
* 1/2-1/2 187259
* Name: count, dtype'''
* Similar number of games are won by black and white players due to time forfeit
* print(time\_winloose)
* '''1-0 983433
* 0-1 976287
* 1/2-1/2 51616
* Name: count, dtype'''

### Difference of elo points in each game

print(elo\_diff\_stats)

'''

count 6254841.00

mean 147.18

std 139.33

min 0.00

25% 46.00

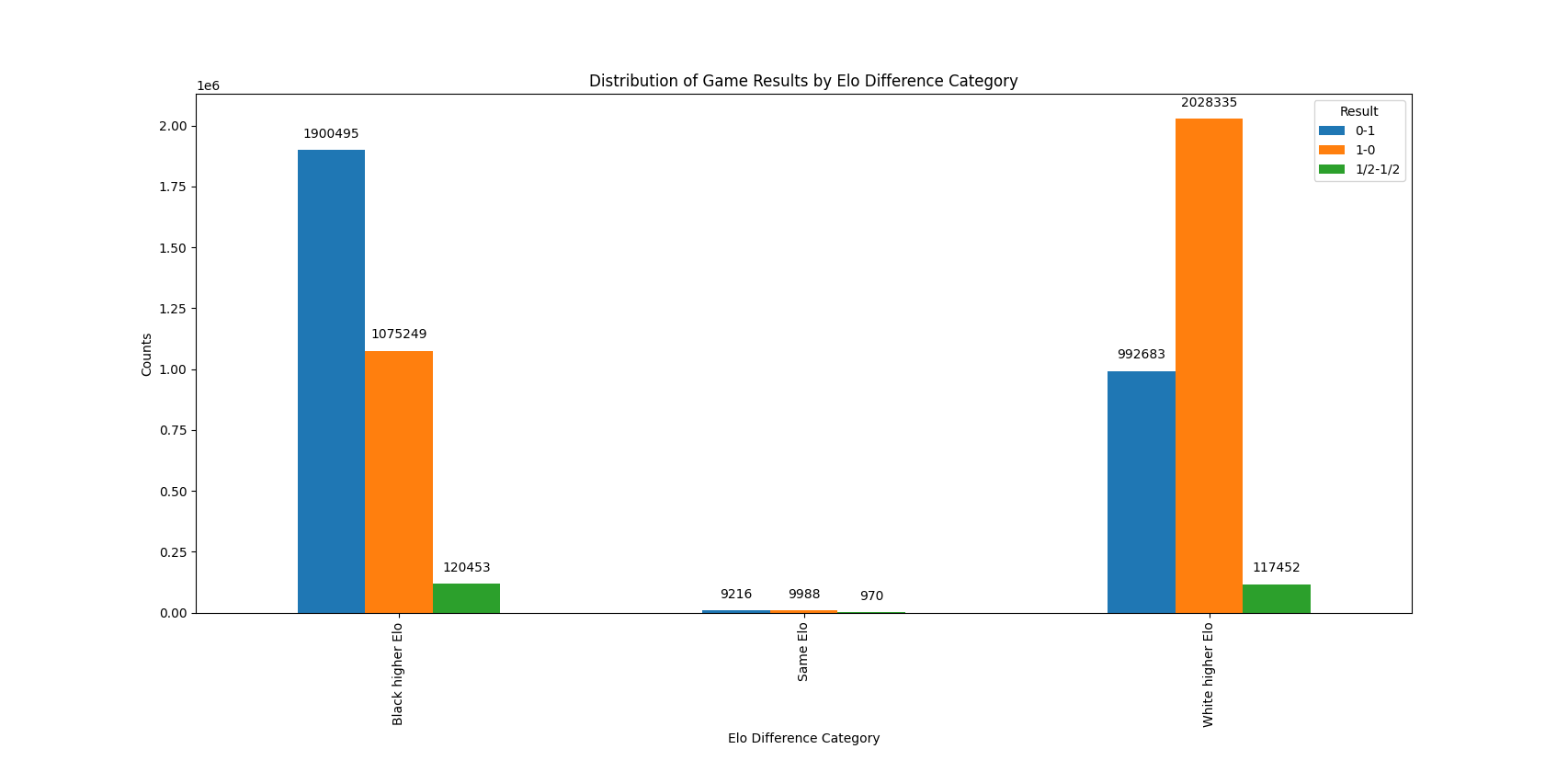
50% 106.00

75% 204.00

max 1702.00 very big difference in elo level

there are games with huge elo gap between players

There are 184550 games that have a level difference above 500 points.

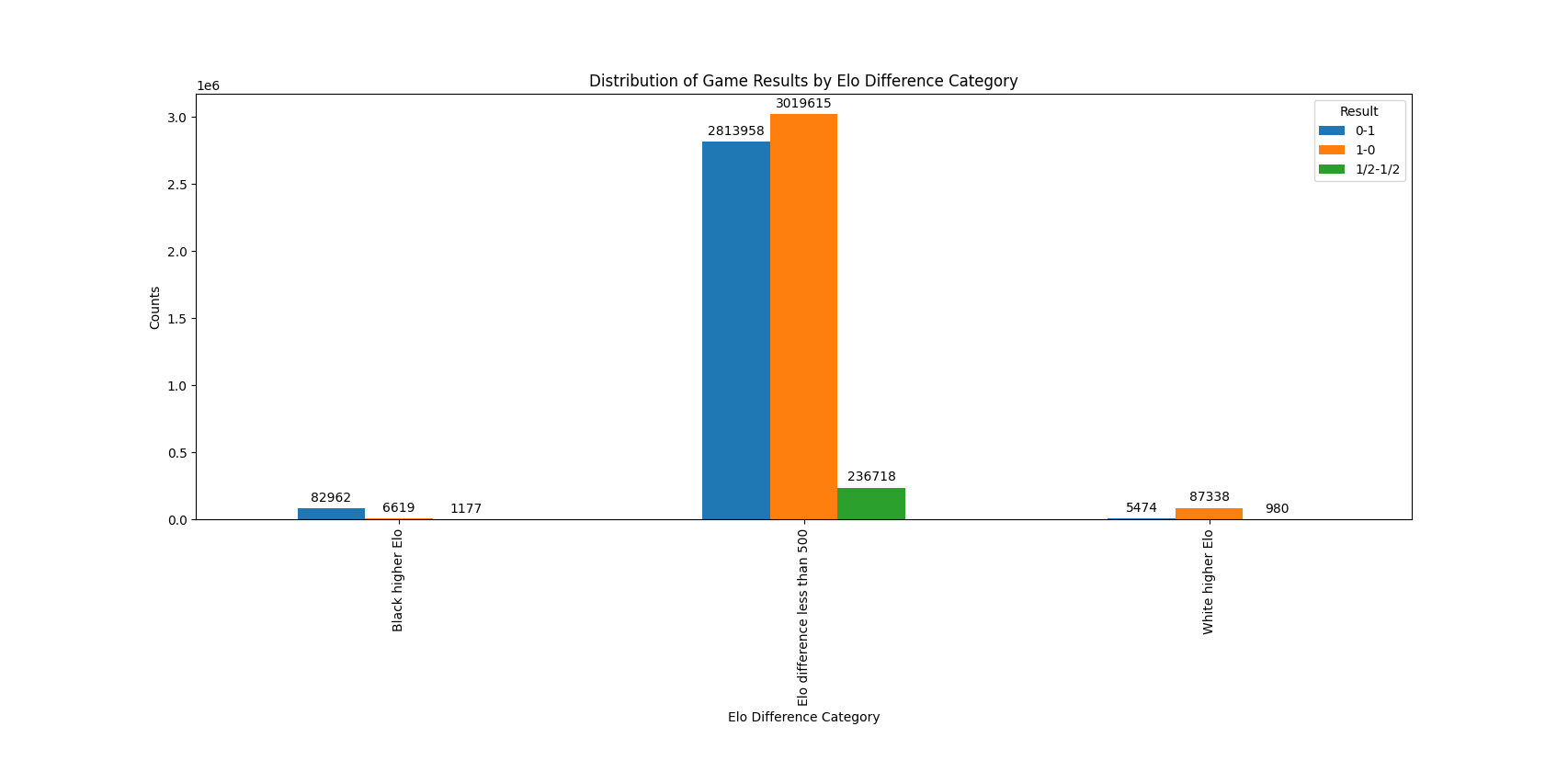


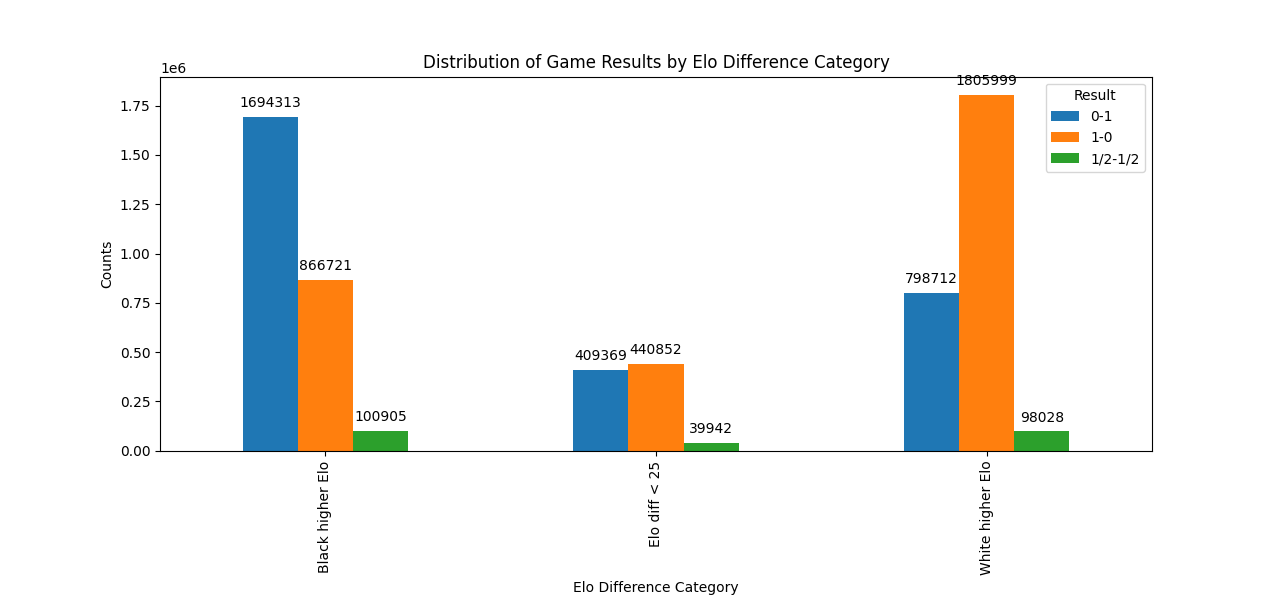
For the games with the same elo, we get similar values of win-loss (central 3 bars)

For the games where the black players are better black players win more although many white players are able to win. Same for games where the white player had an unbalanced Elo. (this plot is showing strictly different ELOs, that is why the same elo is so little).

For the visualization of the games with difference of more than 500 points in elo:

* more whites win.

We do the same with elo difference of 25 because we use this to test the accuracy of our model:



* For games with higher white ELO, whites win more
* For games with higher black ELO, blacks win more
* For games with around the same level (difference less than 25 pts) the white players win more

### Elo correlations (bad)

'''

WhiteElo BlackElo numeric\_result

WhiteElo 1.000000 0.710445 0.140026

BlackElo 0.710445 1.000000 -0.145405

numeric\_result 0.140026 -0.145405 1.000000

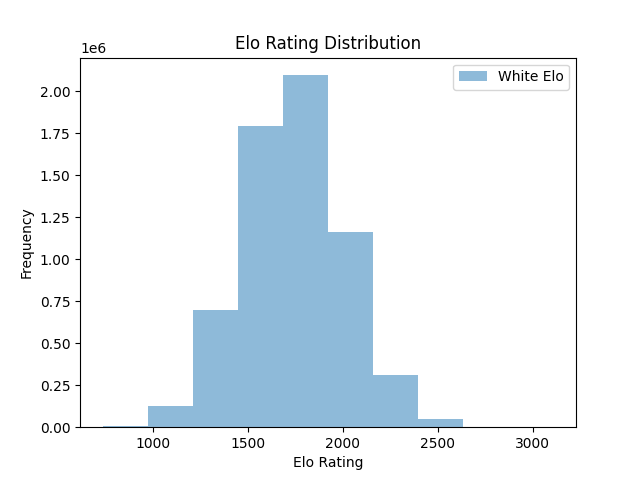
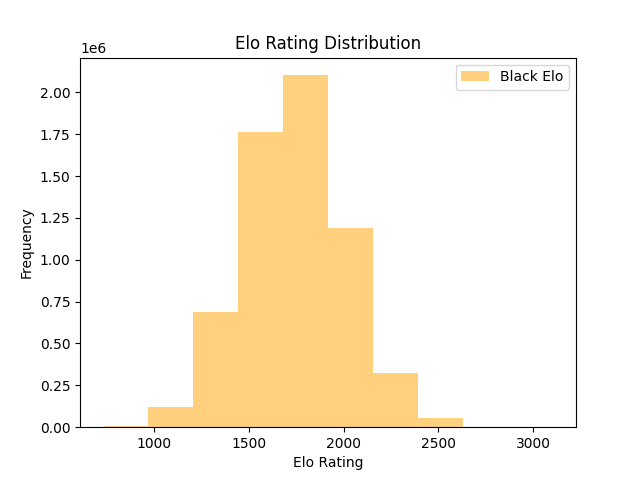
No correlation for the numeric result, maybe that whites win more

'''

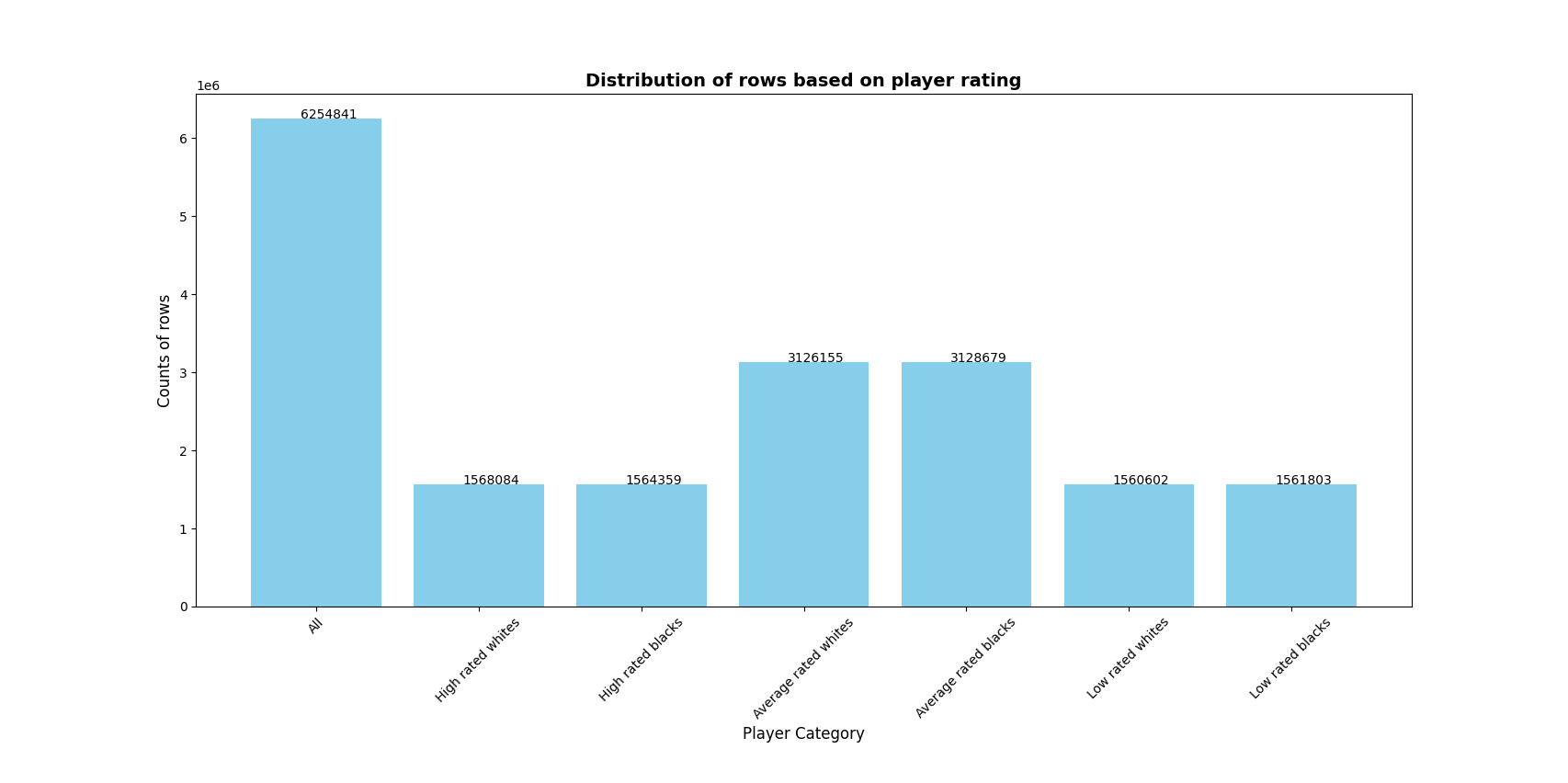
As we have seen, the black and white elo levels are highly related (this means that mainly the players are paired by similar level. The negative correlation between the numeric result and the black elo might indicate that black players win less (no estic segura d’aixó)

### Division by elo level

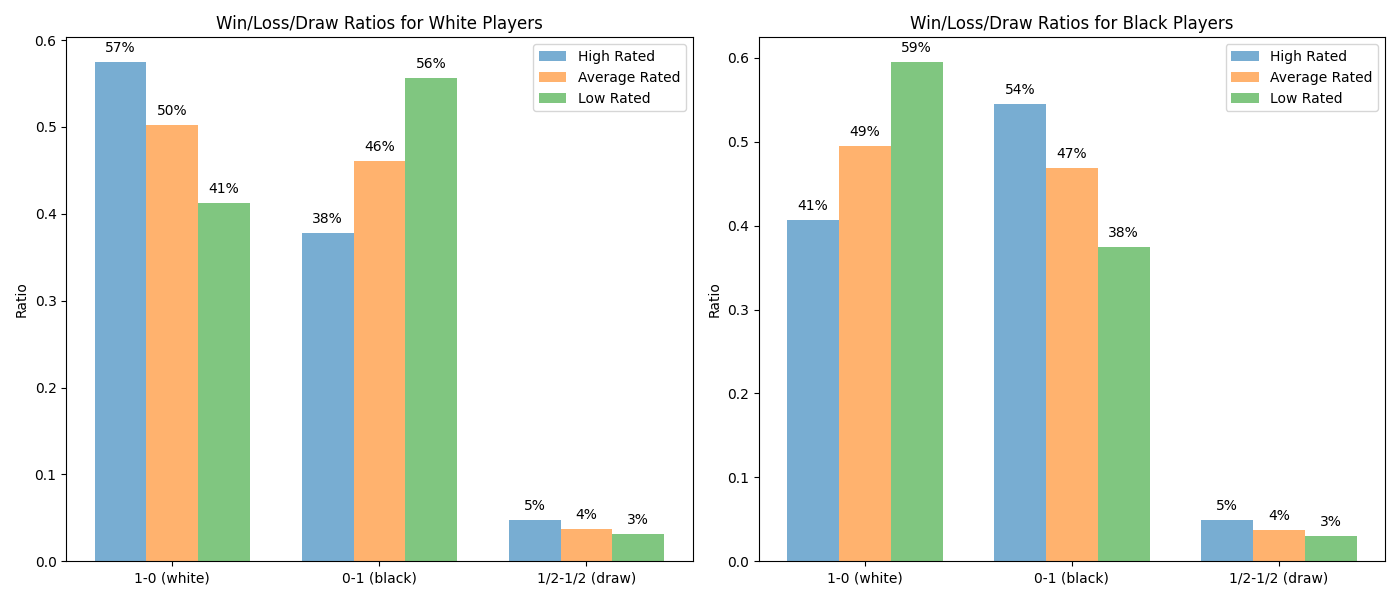
We check how elo levels are distributed for b&w players



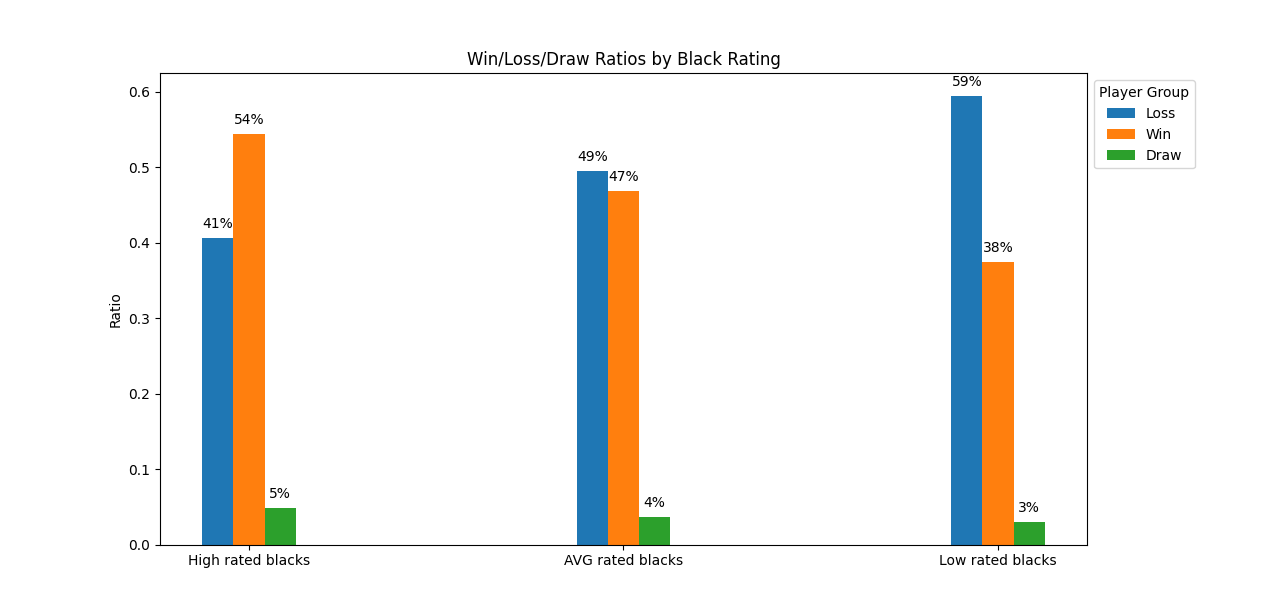
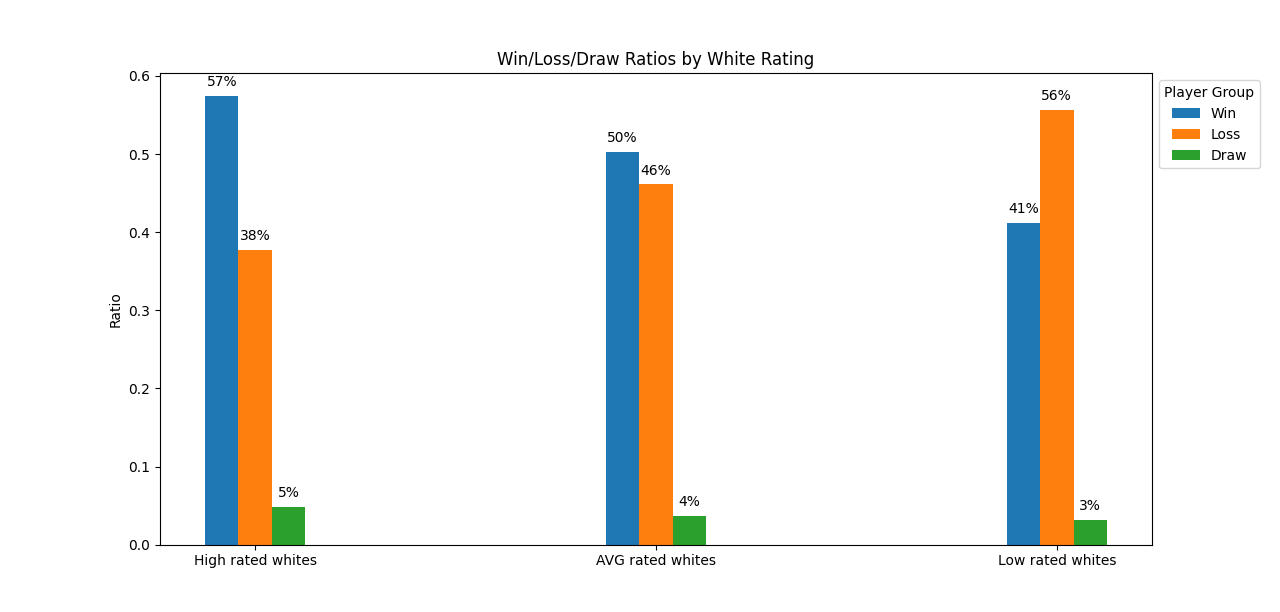
We divide the players by categories, sorting the top 25th percentile and the bottom 25th percentile and we can check that the number of players for each category and color is balanced. We have many average players.



### results by elo level

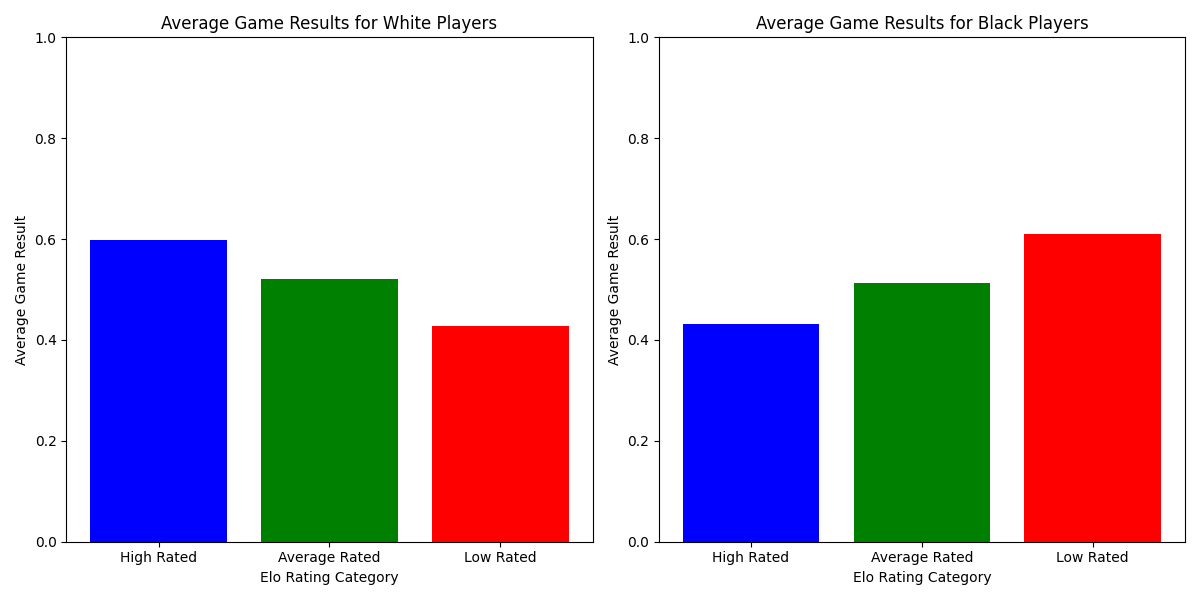


We see that the results are consistent, more than 50% of white and black high rated players win the games (it does not sum 100 because high rated players are also paired with lower level players, the 100 sum is between all of the results of each rating and each color,this is a bit difficult to understand so new plot reorganizing the bars)



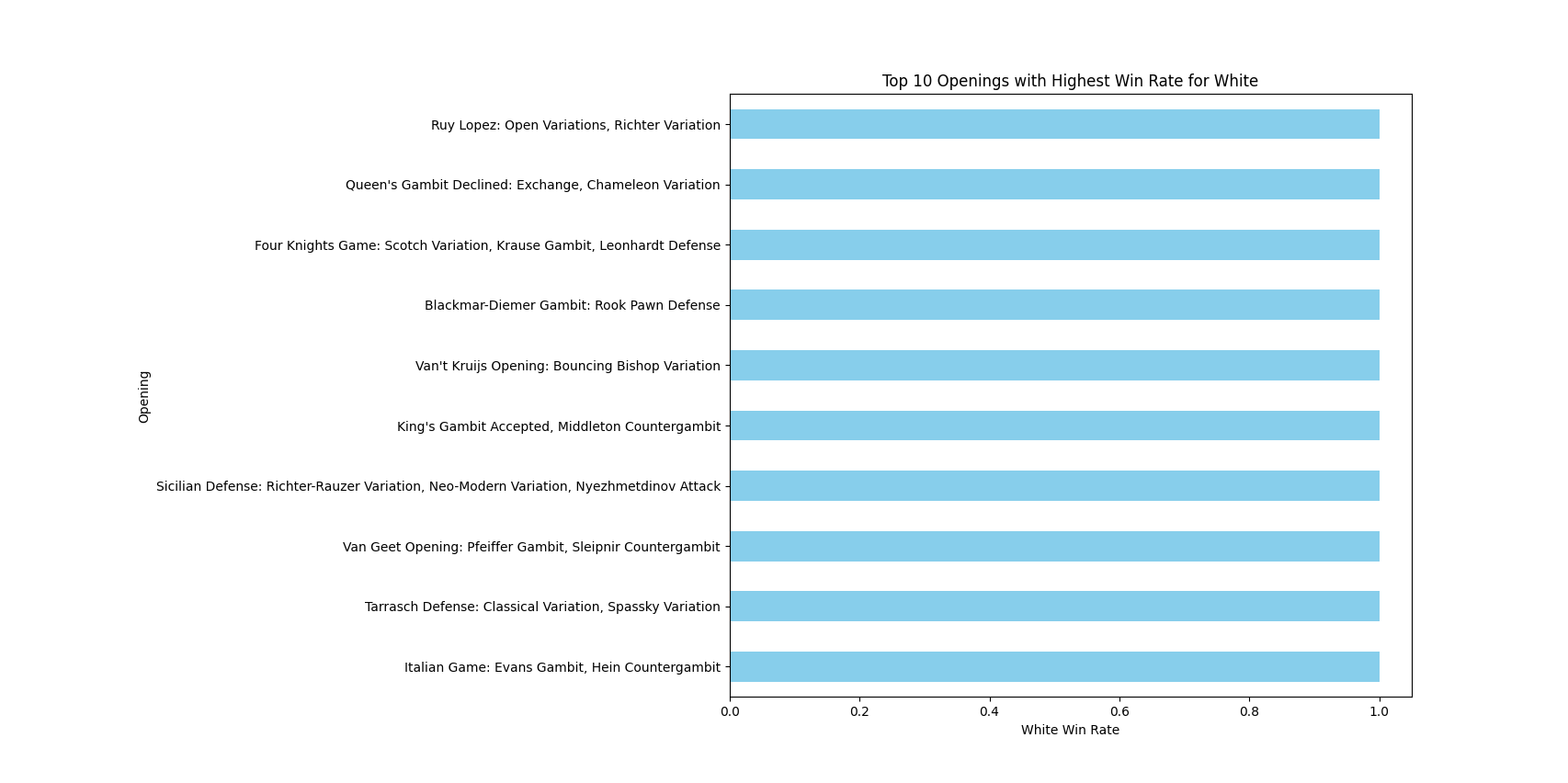
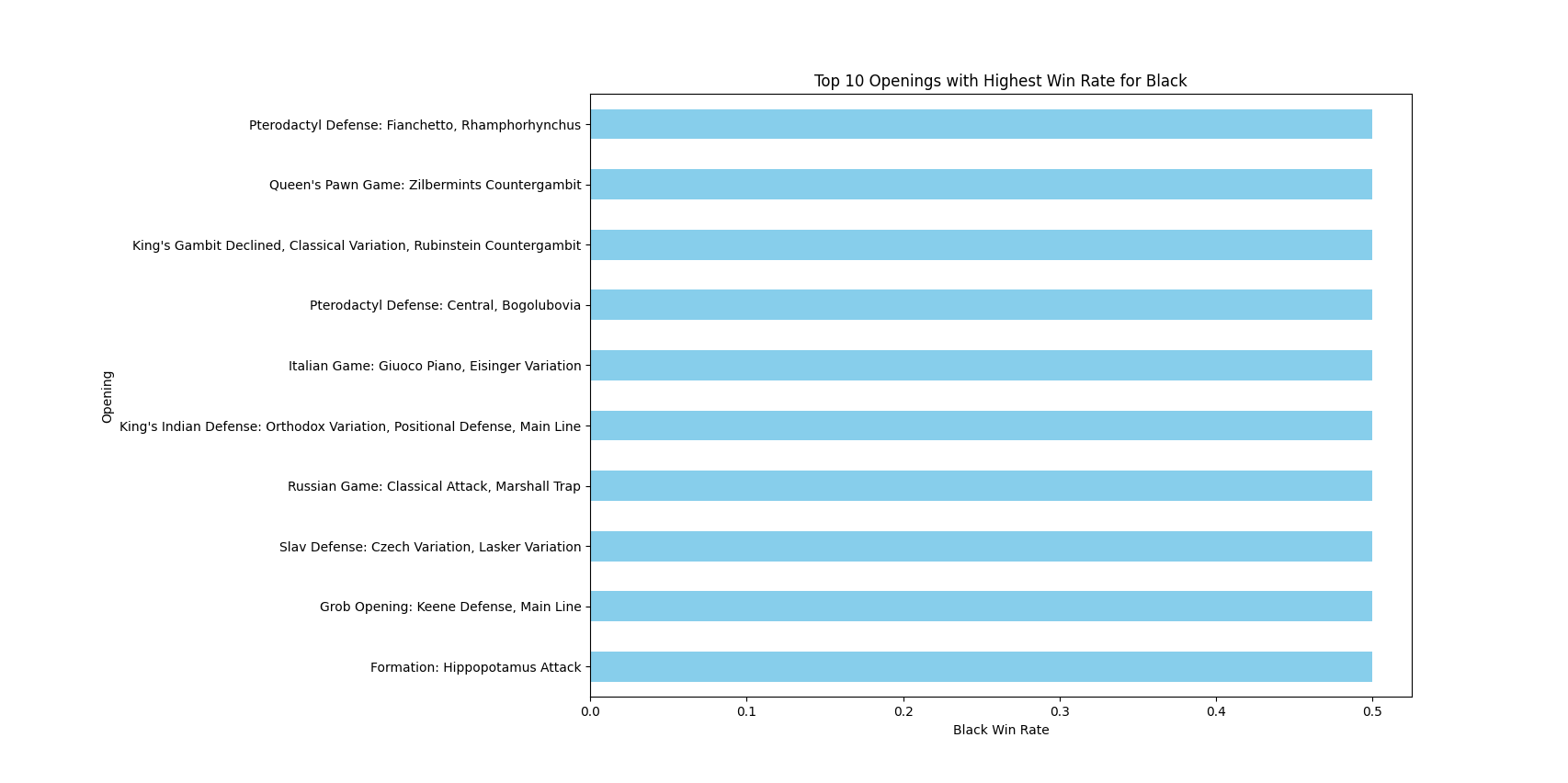
### average game results for each elo level (NOT GOOD)

We miss the wins and the loses, it is a worse version of the plot in the previous section.

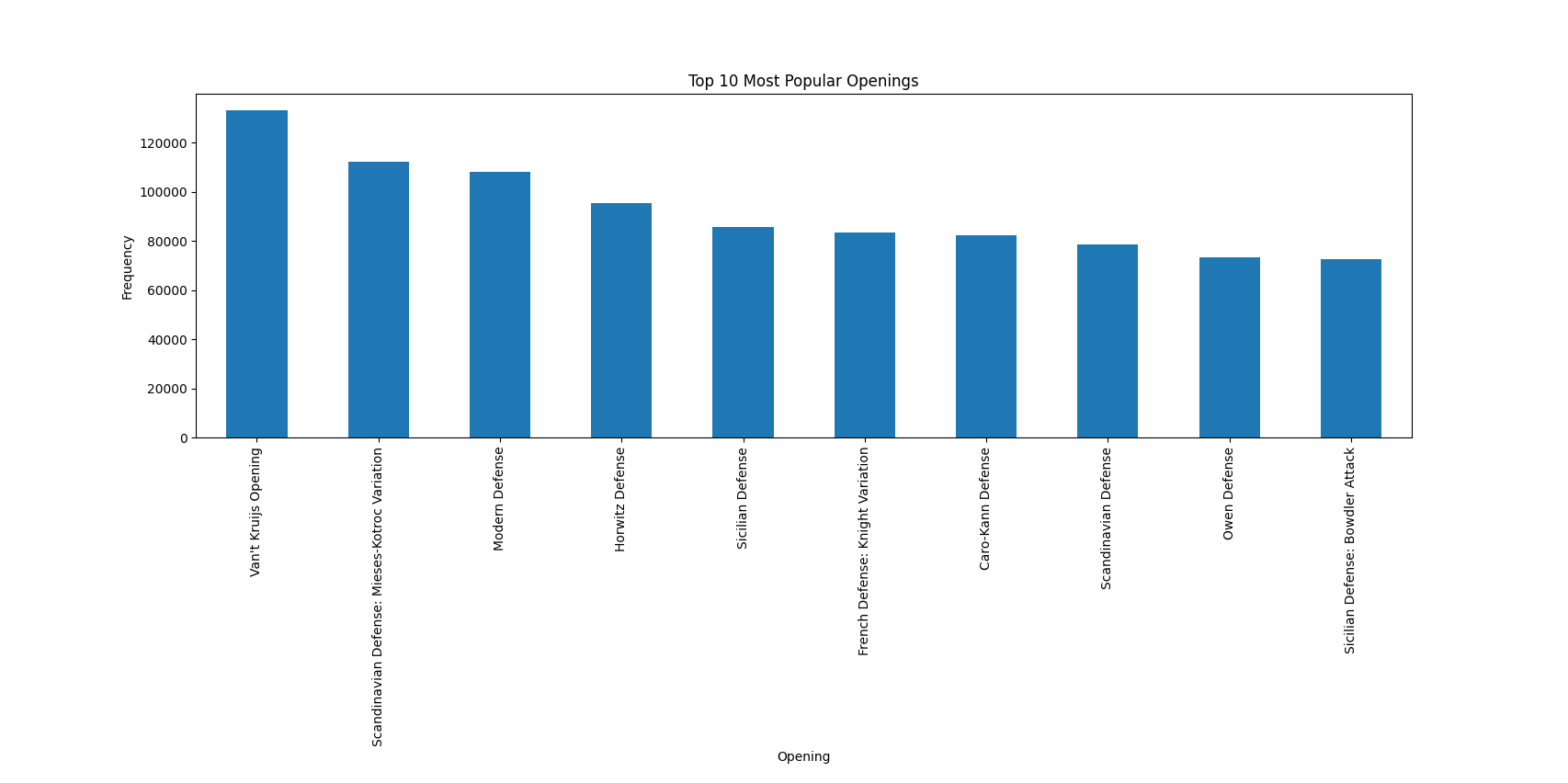


### opening effectiveness by color (only used once openings too)

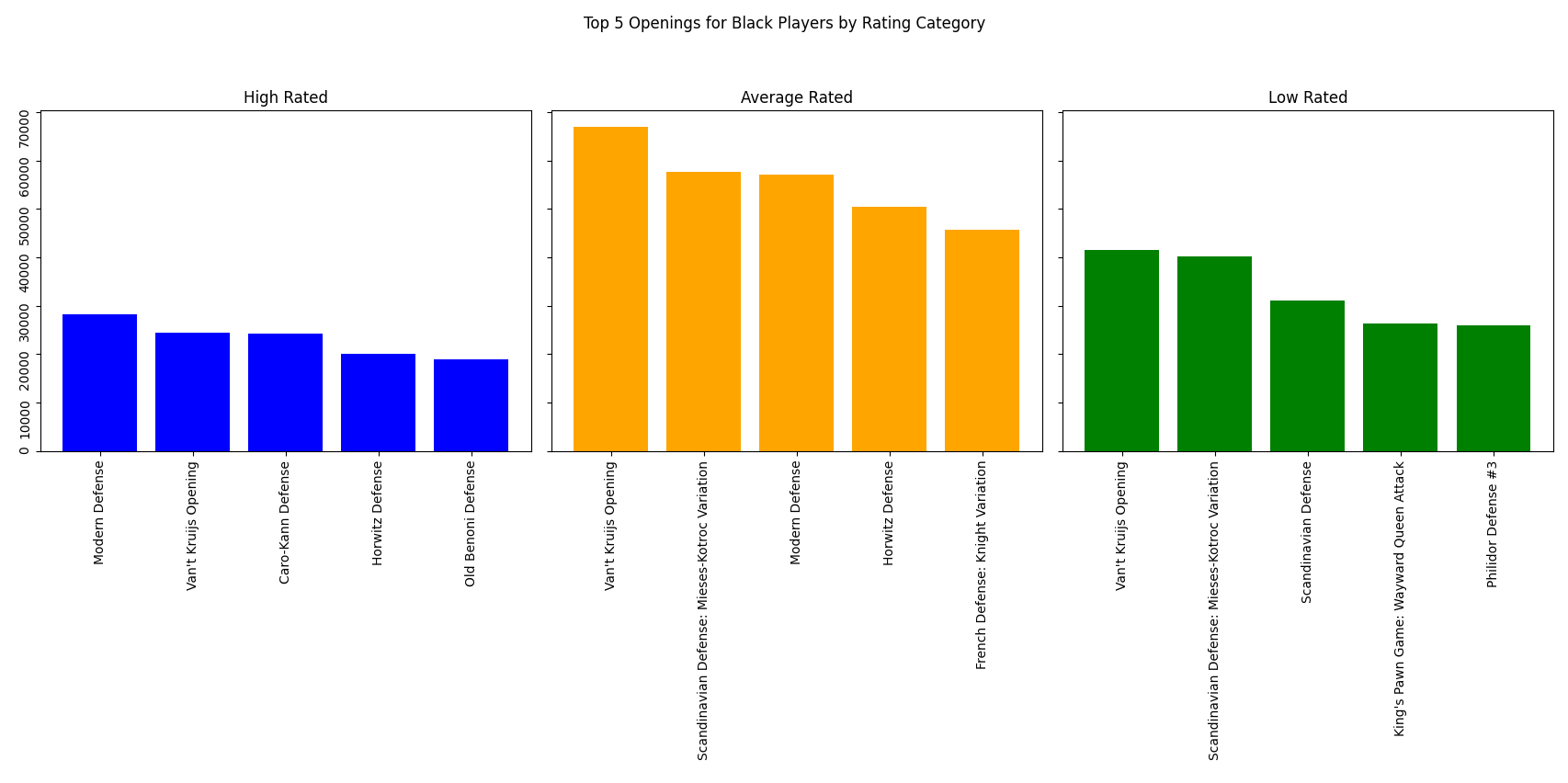
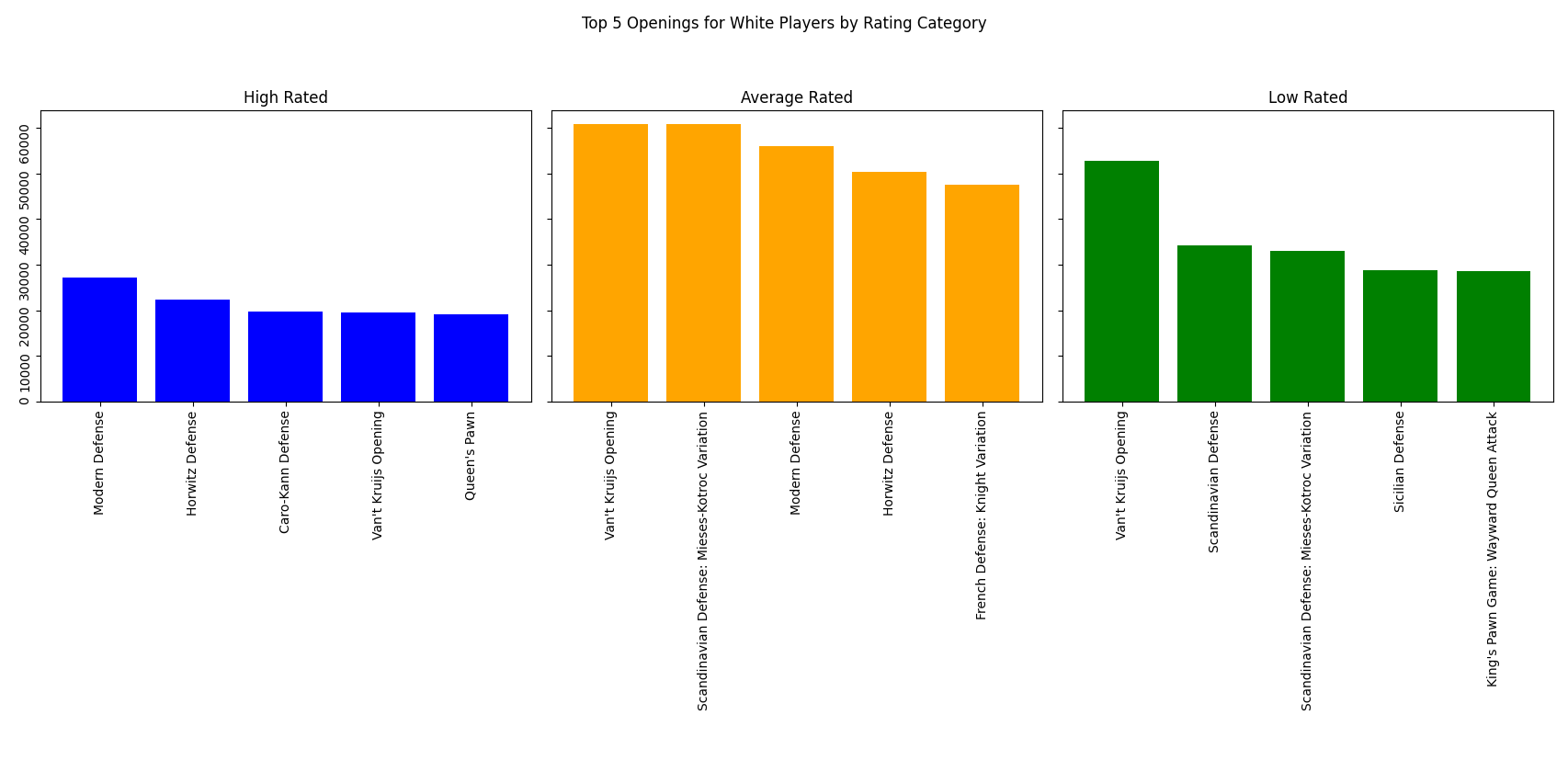
These are lists that have openings that resulted in the win of the player, however, as they have a win rate of 1, this means that these openings were only used once so they are not relevant for our analysis.



### top 10 openings (frequency)

these are the 10 most used openings by all users. this can be used to compare how this list differs from the most frequent and best effectiveness depending on only color or color and ELO category (analyses 11-12)

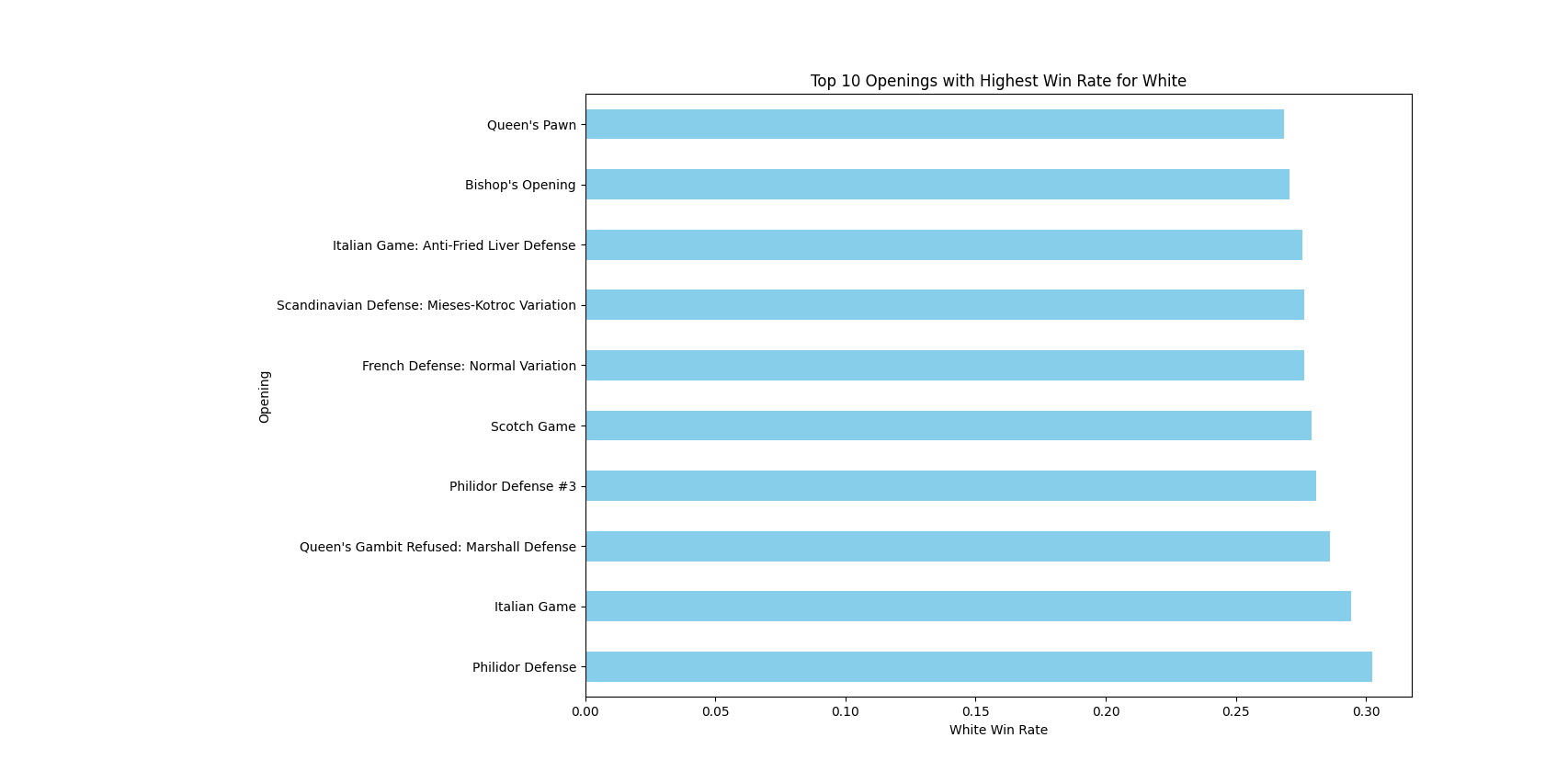
### openings used by elo level (frequency)



this is similar to the plot in section 9 but divided by elo categories, we can compare the most known openings to the most effective openings.

### opening effectiveness from the most 46 used openings (by color)

\*\*marking in yellow the openings that are the same as the most popular openings

'''

Philidor Defense 0.30

Italian Game 0.29

Queen's Gambit Refused: Marshall Defense 0.29

Philidor Defense #3 0.28

Scotch Game 0.28

French Defense: Normal Variation 0.28

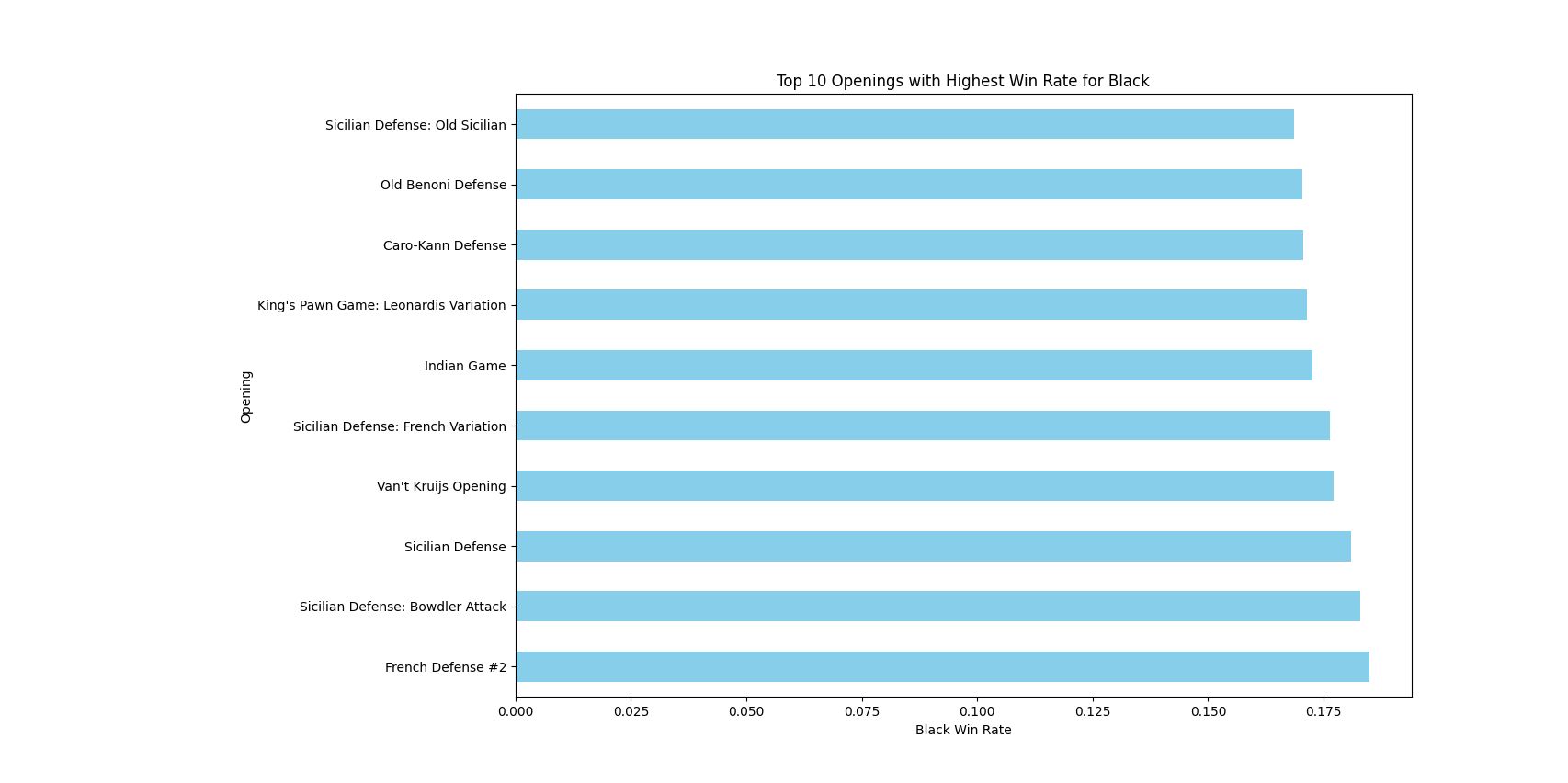
Scandinavian Defense: Mieses-Kotroc Variation 0.28

Italian Game: Anti-Fried Liver Defense 0.28

Bishop's Opening 0.27

Queen's Pawn 0.27

'''



'''

French Defense #2 0.18

Sicilian Defense: Bowdler Attack 0.18

Sicilian Defense 0.18

Van't Kruijs Opening 0.18

Sicilian Defense: French Variation 0.18

Indian Game 0.17

King's Pawn Game: Leonardis Variation 0.17

Caro-Kann Defense 0.17

Old Benoni Defense 0.17

Sicilian Defense: Old Sicilian 0.17

'''

We have taken the list of the 46 most used openings (more than 25000 times used in our dataset). We can see that from the list of the most used, the openings that are giving the best results are not the top 10 most common openings. (Even though we got the most common openings, the ones that give us better results are not in the top 10 list).

### openings by elo and effectiveness (winrate)

We can see that the win rate goes down for lower ranked players no matter the color

'''

Performance of White Players:

----------------------------------------

High Rated White Players:

Result 0-1 1-0 1/2-1/2 WinRate

Opening

Philidor Defense 827 2602 137 0.73

Italian Game: Anti-Fried Liver Defense 772 2326 122 0.72

Italian Game 667 1928 93 0.72

Philidor Defense #3 1589 4391 214 0.71

Scotch Game 1643 4067 238 0.68

----------------------------------------

Average Rated White Players:

Result 0-1 1-0 1/2-1/2 WinRate

Opening

Philidor Defense 4269 8130 501 0.63

Italian Game 4834 8147 461 0.61

Philidor Defense #3 13207 20599 1372 0.59

Italian Game: Anti-Fried Liver Defense 7761 11718 788 0.58

Scotch Game 9966 15199 1144 0.58

----------------------------------------

Low Rated White Players:

Result 0-1 1-0 1/2-1/2 WinRate

Opening

Italian Game 4282 5096 279 0.53

Philidor Defense 3860 4510 362 0.52

Queen's Gambit Refused: Marshall Defense 3559 3619 242 0.49

Scotch Game 8370 8575 690 0.49

Philidor Defense #3 10721 10707 744 0.48

----------------------------------------

Performance of Black Players:

----------------------------------------

High Rated Black Players:

Result 0-1 1-0 1/2-1/2 WinRate

Opening

King's Pawn Game: Wayward Queen Attack 660 201 28 0.74

French Defense #2 2994 1102 167 0.70

Sicilian Defense: Bowdler Attack 8149 3296 426 0.69

King's Pawn Game: Leonardis Variation 2385 1042 132 0.67

Sicilian Defense 11519 5641 844 0.64

----------------------------------------

Average Rated Black Players:

Result 0-1 1-0 1/2-1/2 WinRate

Opening

King's Pawn Game: Wayward Queen Attack 7190 4707 446 0.58

French Defense #2 9394 6405 543 0.57

King's Pawn Game: Leonardis Variation 11881 8792 847 0.55

Sicilian Defense: Bowdler Attack 24012 17996 1535 0.55

Van't Kruijs Opening 36678 28235 2131 0.55

----------------------------------------

Low Rated Black Players:

Result 0-1 1-0 1/2-1/2 WinRate

Opening

Sicilian Defense 10137 11127 691 0.46

French Defense #2 4689 5168 332 0.46

King's Pawn Game: Leonardis Variation 9256 10708 708 0.45

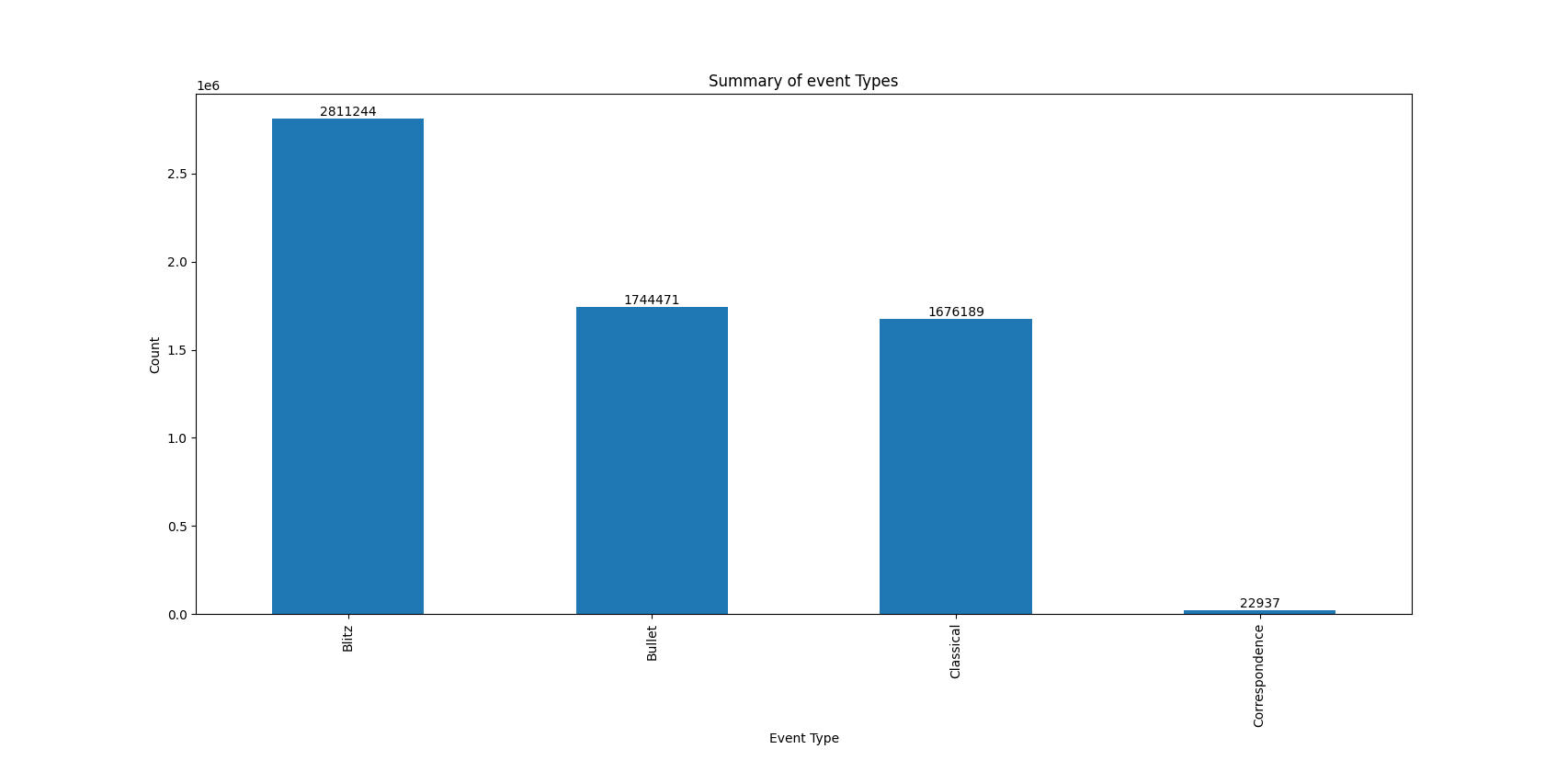
Scandinavian Defense 13912 16176 1063 0.45

Sicilian Defense: Bowdler Attack 7591 8925 514 0.45

----------------------------------------

'''

### event types + by elo



bullet<blitz<classical

(correspondence is longer and different to the regular games, they allow an openings book)

blitz games are the ones most played by all players

Event types of White Players:

----------------------------------------

High Rated White Players:

Event

Blitz 709424

Bullet 555025

Classical 300675

Correspondence 2960

Name: count, dtype: int64

----------------------------------------

Average Rated White Players:

Event

Blitz 1381718

Classical 890740

Bullet 843554

Correspondence 10143

Name: count, dtype: int64

----------------------------------------

Low Rated White Players:

Event

Blitz 720102

Classical 484774

Bullet 345892

Correspondence 9834

Name: count, dtype: int64

----------------------------------------

Event types of Black Players:

----------------------------------------

High Rated Black Players:

Event

Blitz 708692

Bullet 554282

Classical 298350

Correspondence 3035

Name: count, dtype: int64

----------------------------------------

Average Rated Black Players:

Event

Blitz 1384428

Classical 891877

Bullet 842138

Correspondence 10236

Name: count, dtype: int64

----------------------------------------

Low Rated Black Players:

Event

Blitz 1384428

Classical 891877

Bullet 842138

Correspondence 10236

Name: count, dtype: int64

----------------------------------------

'''

### clustering

It is interesting to see how the plots resulting from clustering are the exact opposite of the scatter plots from the beggining

### 