Cleaning Dataset:

* Replace usernames with real names
* Change ‘time\_control’ column from seconds+seconds to minutes+seconds. Minutes+seconds is standard.
* Use value\_counts() to decide whether to filter rows or remove columns including:
  + Remove non-standard ‘variant’ games
  + Remove the empty ‘ply\_count’ column
  + Remove the ‘round’ column, it’s all question marks
  + Remove ‘game\_id’ column. The ‘site’ column serves as a unique identifier.
  + Remove the ‘file\_name’ column since it’s just named after Magnus’s usernames which will be changed to Magnus anyways.
* There could be a better way to showcase the results of games. I prefer to keep white info on the left side, and black info on the right side, rather than alternating but I’m not 100% sure.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Title | Elo Dif | Elo | Name | Result | Name | Elo | Elo Dif | Title |

* The column names have to be unique so I’ll specify black or white there.
* The winner, winner\_elo, loser, loser\_elo columns are just the white, white\_elo, black, black\_elo columns rearranged. Remove these redundant columns.
* The ‘group\_name\_color’ and ‘group\_name\_result’ shows Magnus’ color and result. I could track his color using another column and change the result column name to ‘Magnus result’ or something similar.
* The ‘date\_played’ and ‘utc\_date’ columns are the exact same. Remove the date\_played column.
* Sort games by utc\_date. Original data is sorted by username.