MS-E2112 Project: Online chess

Anni Niskanen

Contents

1 Introduction

Chess is a complex game of strategy, and surely a multitude of variables, both in-game and out-of-game, determines the result of a game. In this project I set out to examine the relation between certain variables that can be recorded from a single game of chess. In particular, I will examine which variables affect the result of the game the most, or whether any such variables can be identified at all. Secondly, the relation between the length of the game and the two players' ratings is of interest to me. Lastly, I seek to answer the question of how often players play with someone of a similar rating, i.e. the relation between the two players' ratings.

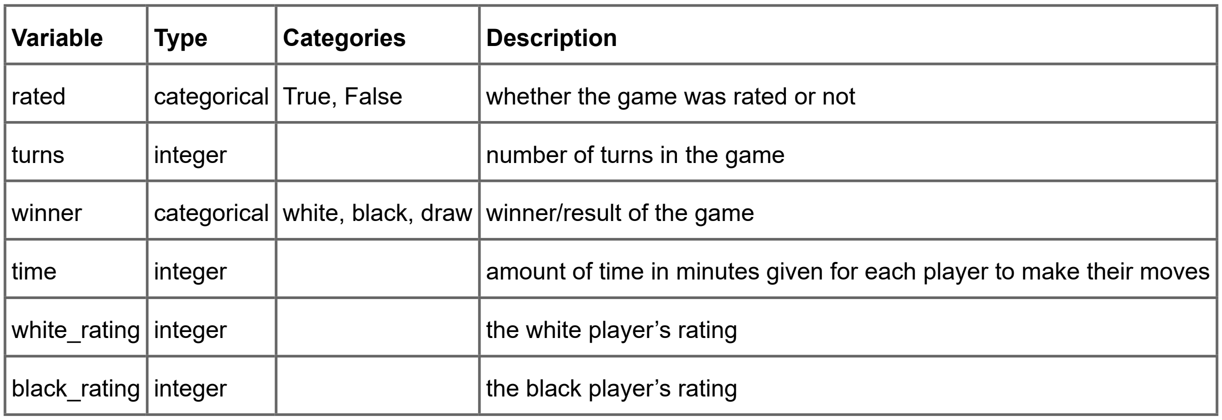
The dataset I utilised in this project was and can be obtained from Kaggle [1], an online site with a variety of datasets. The dataset contains information of over 20,000 online chess games played on the online chess site lichess.org. The variables recorded in the dataset will be introduced in more detail in chapter 2. Because of the nature of my research questions and the categorical nature of some of the variables in the dataset, I plan to apply multiple correspondence analysis (MCA) on the data to conduct my research. MCA will be introduced and conducted in chapter 4.

2 Univariate analysis

2.1 Description of the variables

The chess dataset from [1] contains values of 16 variables from 20,058 online chess games on the site lichess.org. 6 of the 16 available variables were chosen to be analysed. These variables are introduced in table 1 below.

Table 1: The chosen variables, their types, categories and short description of them.



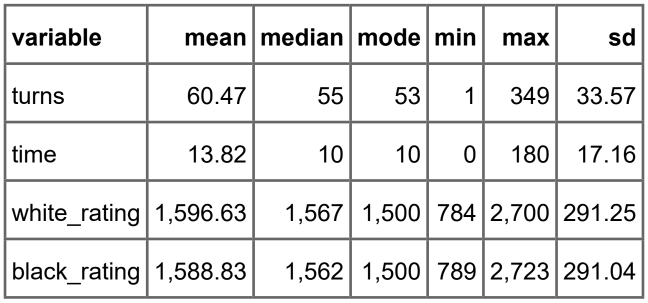
Variable	Type	Categories	Description
rated	categorical	True, False	whether the game was rated or not
turns	integer		number of turns in the game
winner	categorical	white, black, draw	winner/result of the game
time	integer		amount of time in minutes given for each player to make their moves
white_rating	integer		the white player's rating
black_rating	integer		the black player's rating

Note that variable time was obtained by modifying variable increment_code of the original dataset. Variable increment_code contains the amount of time in minutes given for each player to make their moves (A) and the time increment in seconds rewarded for the player for each move (B) in format "A+B". The "+B" part was simply omitted in order to acquire the new variable time.

2.2 Summary statistics

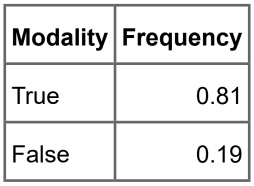
Next I calculated some summary statistics for the 6 variables. Summary statistics for the numerical variables are presented in table 2. Because of their categorical nature, only the relative frequencies of variables rated and winner are presented in tables 3 and 4.

Table 2: Summary statistics for the chosen numerical variables.



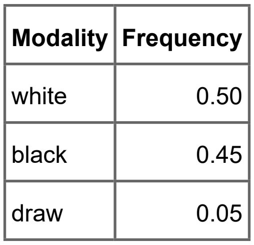
variable	mean	median	mode	min	max	sd
turns	60.47	55	53	1	349	33.57
time	13.82	10	10	0	180	17.16
white_rating	1,596.63	1,567	1,500	784	2,700	291.25
black_rating	1,588.83	1,562	1,500	789	2,723	291.04

Table 3: Relative frequencies for the modalities of the rated variable.



Modality	Frequency
True	0.81
False	0.19

Table 4: Relative frequencies for the modalities of the winner variable.



Modality	Frequency
white	0.50
black	0.45
draw	0.05

2.3 Visualisation and categorisation

Before MCA is applied to the dataset, the numerical variables of the data - namely turns, time, white_rating and black_rating - should be categorised. Histograms of the original numerical data of these variables are plotted in figure 2a. Frequencies of the two already categorical variables, rated and winner, are plotted in figure 1.

MCA is most accurate when there are roughly the same number of modalities for each variable. Obtaining rare modalities should also be avoided when determining the categories. For example, a value higher than 50 for variable time is very rare, so determining a category of time > 50 would be unwise. The variables were categorised based on this knowledge of MCA, the histograms in figure 2a, my own intuition, and lastly some simple trial and error. The categorisation is presented in figure 2b and justified further in appendix A.

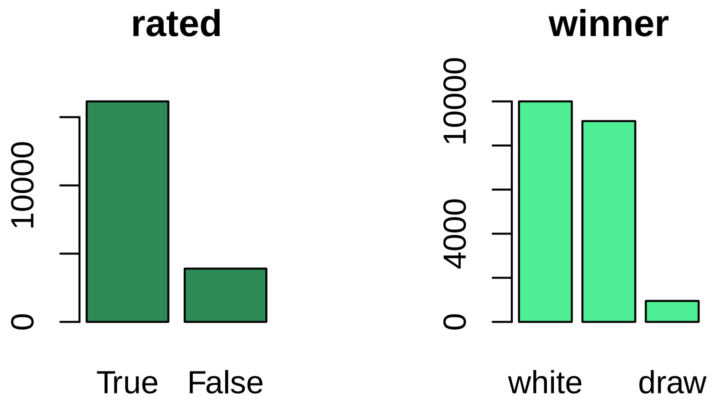


Figure 1: Frequencies of variables rated and winner.

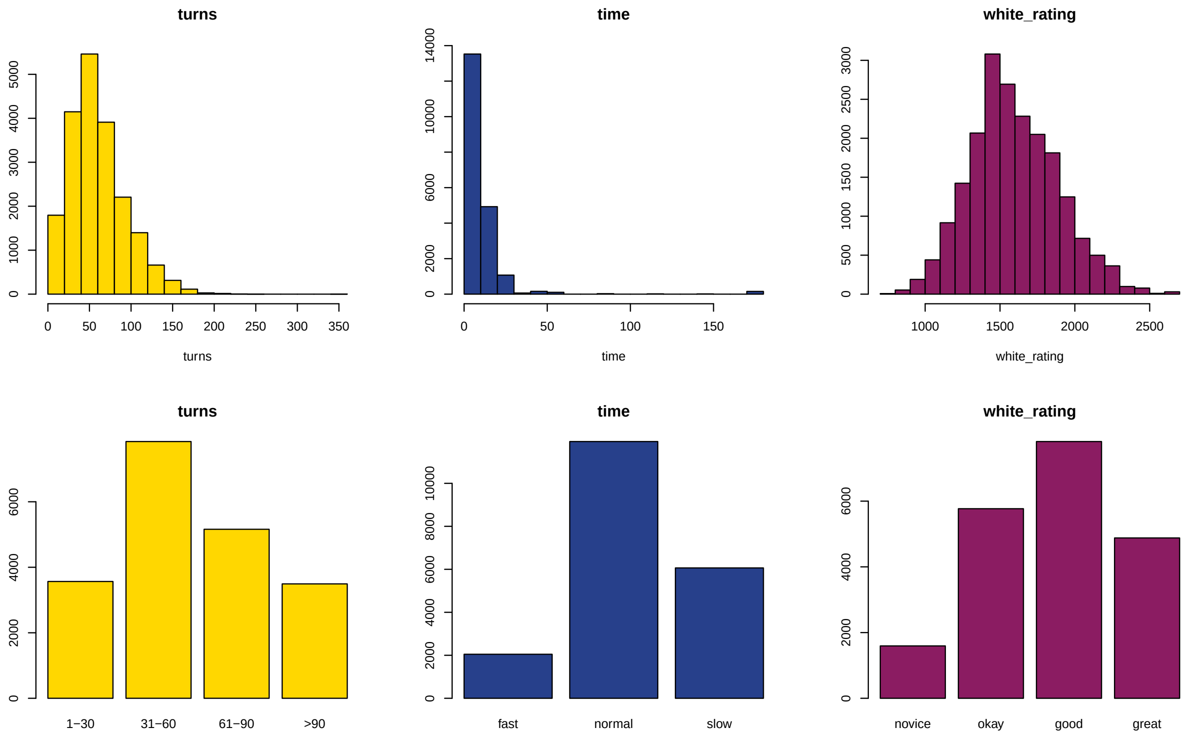


Figure 2: Frequencies of the original and categorised variables turns, time and white_rating. Variable black_rating is omitted because its distribution is very similar to variable white_rating.

3 Bivariate analysis

For categorical variables, it is easiest to model the dependencies between the modalities of the variables. Attraction repulsion indices represent these dependencies. An index higher than 1 indicates attraction and an index below 1 repulsion, and an index close to 1 indicates independence of the two modalities. A heatmap of the attraction repulsion indices of the data is presented in figure 3.

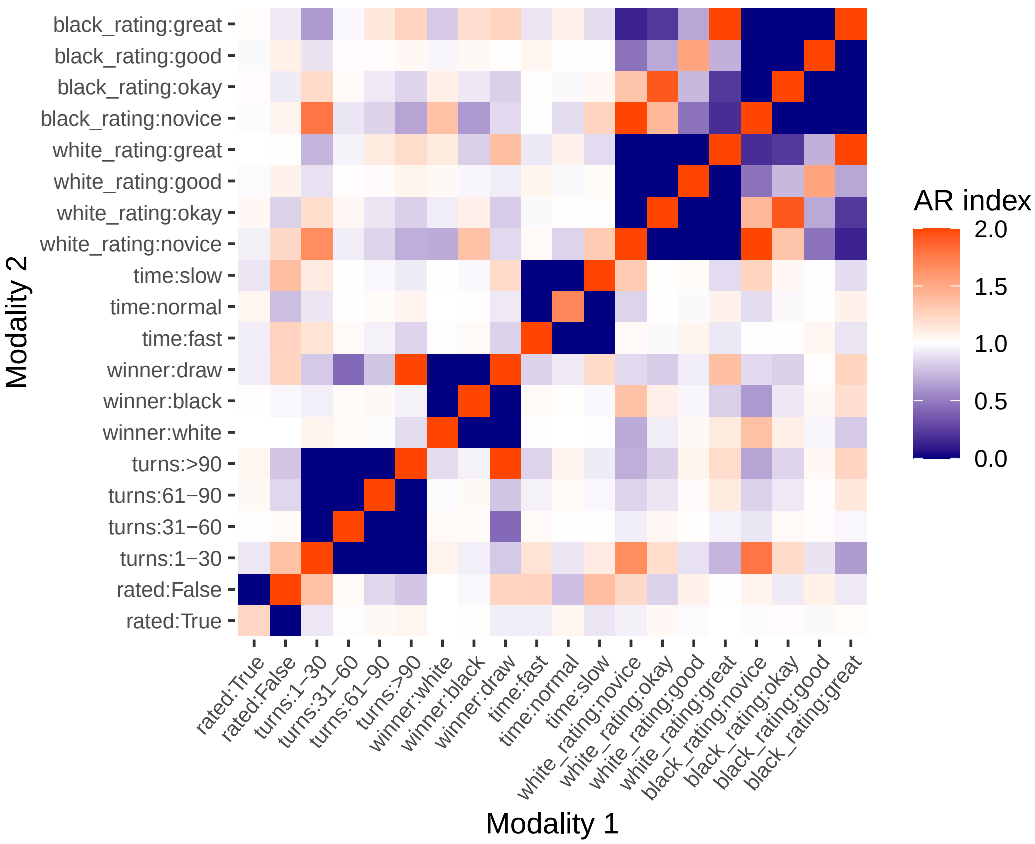


Figure 3: A heatmap of the attraction repulsion indices of the data. Attraction repulsion indices higher than 2.0 were reduced to 2.0 in order to keep the color scale clear and meaningful.

It should be noted that the exceptionally high diagonal values (attraction repulsion indices of the modalities in regard to themselves) in the heatmap are irrelevant, as are the low values between "competing" modalities (such as the True and False modalities of variable rated which logically preclude each other, and therefore their attraction repulsion index is always 0.0).

Quite a lot of interesting dependencies can be seen from the heatmap. Because of the nature of MCA, the analysis on the results of MCA in chapter 4 will present a lot of the observations that could be presented based on the heatmap alone. Therefore I will keep the analysis on the heatmap brief here and only focus on my first research question: which variables affect the result of the game the most. The heatmap shows an attraction between a high rating of a player and that player winning the game, and an even stronger repulsion for a low rating of a player and that player winning the game. A similar phenomenon is seen for the second player: a low rating of the second player attracts the first player winning and vice versa. This dependency, a high rating of a player attracting the modality of that player winning, is quite simple and logical. The dependencies of the draw modality are more interesting. It can be seen that the >90 modality of variable turns is very strongly attracted to the draw modality, while other modalities of variable turns, especially 31-60, are repulsed by it. Moreover, high ratings of both players attract the draw modality, while low ratings repulse it. This indicates that games which have been played for over 90 turns by skilled players are quite likely to end in a draw, while a shorter game by two inexperienced players will likely be one by one of them.

4 Multivariate analysis

4.1 Introduction of the method

Multiple correlation analysis (MCA) aims to produce a graphical display of associations between the modalities of the variables in a lower dimension, without losing too much information provided by the attraction repulsion indices. That is to say, MCA is in a sense a graphical representation of the attraction repulsion indices: it can summarise attractions and repulsions of the modalities in a single graph. This makes MCA ideal for my research questions. Additionally, because MCA can be conducted only for categorical variables, it is suitable for the chess dataset I have chosen for this project.

4.2 Implementation

In this project MCA was implemented in R with the `mjca` function of the `ca` package. The variation of MCA where correspondence analysis (CA) is applied to the complete disjunctive table was implemented by setting the parameter `lambda = "indicator"`.

4.3 Results and their interpretation

Figure 4 summarises the components obtained with MCA. It shows that the first two components together explain only 21.9% of the total variance in the data. Nonetheless, in order to obtain a 2-dimensional graphical representation of our data, only the first two components will be analysed. The scores for the modalities in the first two dimensions are plotted in figures 5, 6 and 7 below. Because there are 20 modalities in total, only subsets including the relevant modalities are plotted for each research question. Relevance is determined by examining figure 3 and choosing only attracting and repelling modalities. In order to check whether the modalities are represented well in these first two dimensions, instead of arrows we plot points representing qualities of representation of the modalities, i.e. how well the modality is represented by the two first components.

Figure 5 examines the relation between the result of the game and player ratings when one of the players wins the game (i.e. the game does not end in a draw). Unfortunately, even though the heatmap in figure 3 showed dependencies between player ratings and the result of the game, those dependencies are not visible here. It is likely that the relation between the result of the game and player ratings is overshadowed by the very strong relation between the player ratings themselves, and therefore MCA cannot create a meaningful graphical representation here.

Figure 6 describes the relation between various modalities considered possibly impactful for the draw modality and the draw modality itself. Inspecting the games which ended in a draw in such a way provides more interesting results than figure 5. As the heatmap indicated, players with high ratings

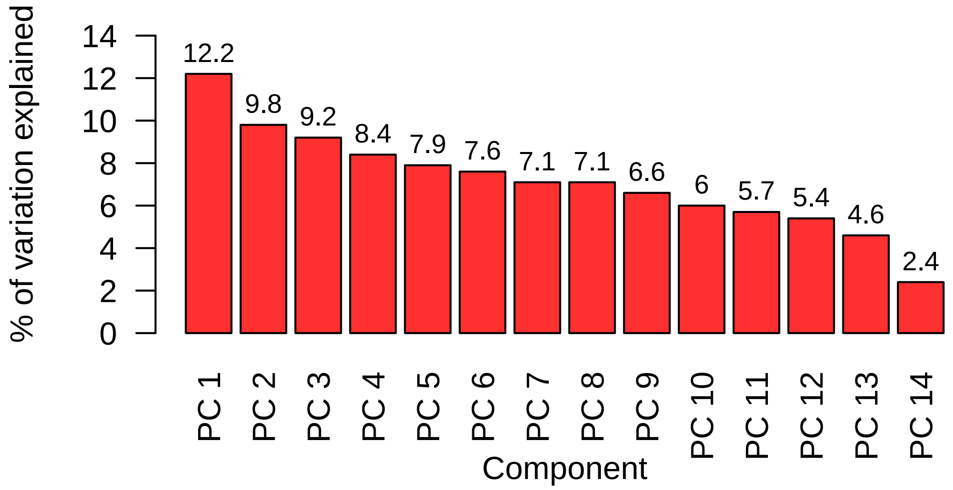


Figure 4: A plot summarising the obtained components and the percentage of variance in the data they explain.

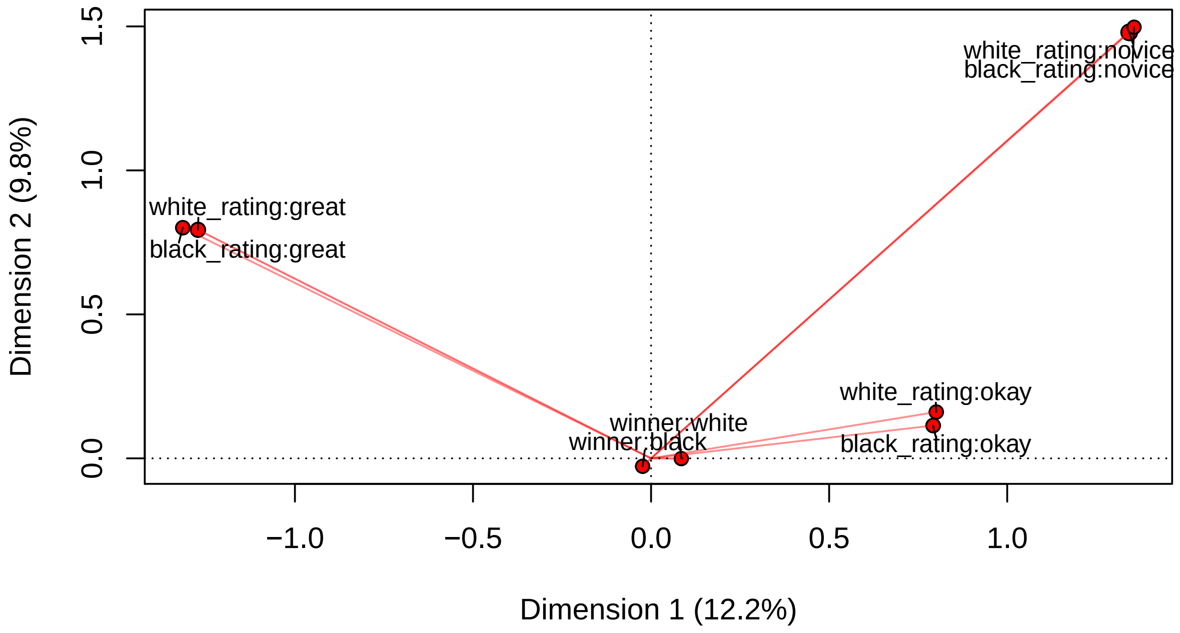


Figure 5: The MCA scores related to the research question ‘Which variables affect the result of the game the most?’, focusing on the games where either the black or white player won.

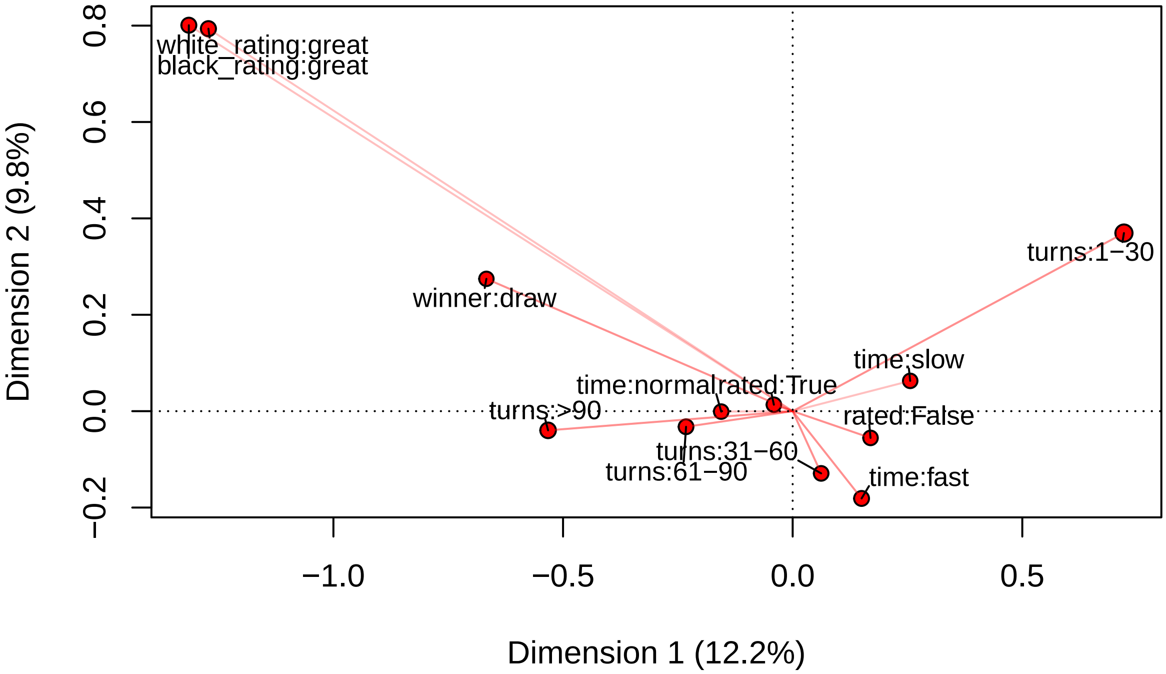


Figure 6: The MCA scores related to the research question ‘Which variables affect the result of the game the most?’, focusing on the games which ended in a draw.

are more likely to play games ending in a draw. It can also be seen that longer games of over 60 turns are more likely to end in a draw, while shorter games repel this modality. Interestingly, the graph completely disagrees with the heatmap on the relation between draws and non-rated games: while the heatmap indicates the two modalities attract each other, according to the graph there is a strong repulsion between them. Lastly, one can see that fast games are less likely to end in draws.

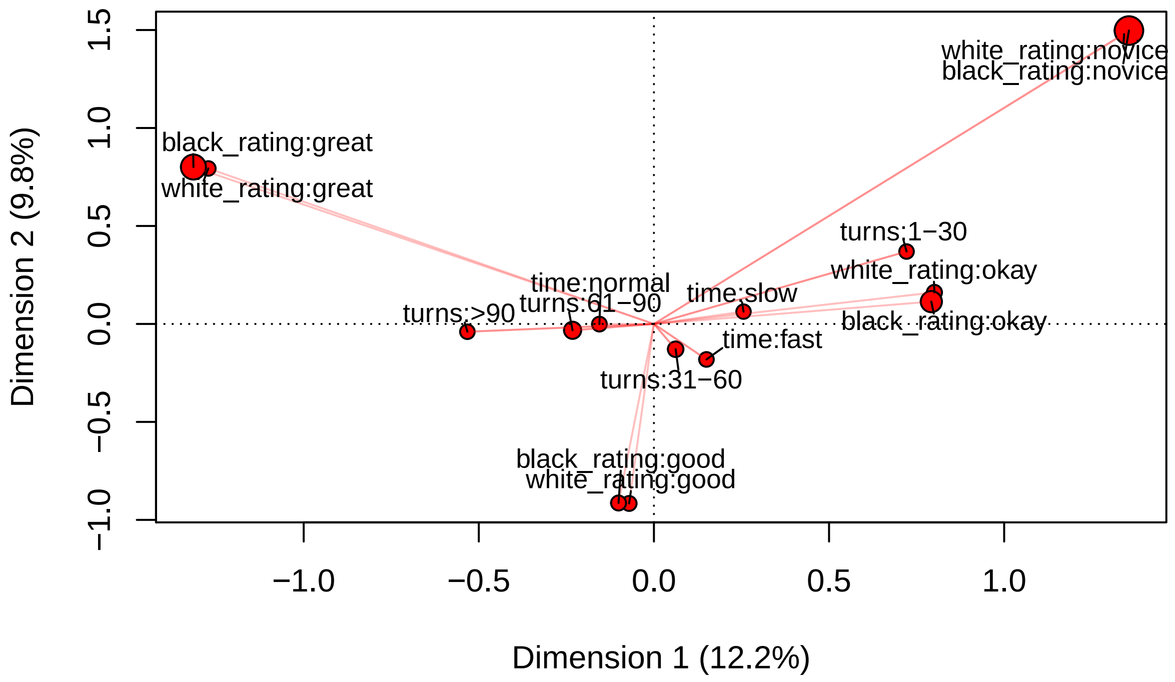


Figure 7: The MCA scores related to the research questions 'What is the relation between the length of the game and the two players' ratings?' and 'How often do players play with someone of a similar rating?'.

Figure 7 aims to help answer the other two research questions. Starting with the second question: The length of a chess game can mean many things. Here it means both the number of turns in the game and the time the players were given to think of their moves in the game. From the graph one can see that players with high ratings seem to be more likely to play longer games with more than 60 turns and normal timing, while novice and okay players will likely play games which end in 1-30 turns and give more time for the player to think.

Figure 7 is also useful for answering our last research question, how often do players play with someone of a similar rating. The simple answer is very, very often. Okay and novice players are somewhat likely to play with each other, but the better the player is, the more likely they are to play with someone of a similar rating.

5 Conclusions

The variables utilised in this analysis were clearly inadequate to explain which conditions lead to either the black or the white player winning the game. As suspected at the start of the project, the matter is too complicated for just a few variables to explain. However, the results do indicate that some of the simple chosen variables can predict if the game will end in a draw. Interesting relations between player ratings and the length of the game, both in terms of number of turns and the time given for the players to think, were also identified, as well as a clear positive correlation between the ratings of the two players. Altogether, skilled players tend to play normal-paced games with quite a lot of turns with each other, and these games tend to end in a draw. Shorter games, more likely to be played by inexperienced players, are slower-paced and end in fewer turns. Players rarely play with someone with a drastically different rating, likely due to the nature of online chess: instead of playing with friends for fun (perhaps more likely leading to a game of two players with drastically different ratings), players will be matched with someone of a similar rating. Novice and okay players seem to be an exception, mixing and playing together to some extent.

Some properties of the utilised chess dataset could have affected the obtained results. First of all, the dataset contains multiple games played by the same players, and therefore the chess games, or samples, are not fully indepedent. Additionally, while the time variable gives some indication to the length of the game, the actual game time was left unexplored in this analysis. Examining the relation between the two - theoretical maximum game time and actual game time - could also be interesting. Furthermore, I believe in online chess a player is more likely to run out of time not because of lack of skill, but by accident, resulting in a loss. Therefore in this dataset a player winning a game cannot automatically be considered a show of that player's skill. Lastly, it should be noted that the results obtained from this analysis should not be generalised to real-life chess.

The conducted analysis with MCA was far from ideal. The first two components analysed explain only around 22% of the variation in the data, and the quality of representation was not too high for most modalities, as can be seen from the points in the graphs. Additionally, the components (not only the first two, but also all the rest!) seemed to mostly focus on the very strong dependency between player ratings, and all other relations between modalities are overshadowed by this. Therefore, the analysis could perhaps be repeated without the variables describing player ratings. Nonetheless, this project was an interesting case study of what happens when very strong dependencies are present in the data, and some interesting dependencies between modalities were still identified - although they should not be accepted blindly.

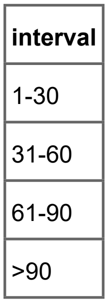
6 References

[1] Kaggle. Chess game dataset (Lichess) [Internet]. 2017. Available from: https://www.kaggle.com/datasnaek/chess

Appendix A: Categorising the variables

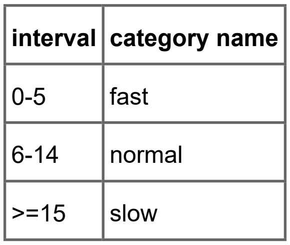
Variables turns, time, white_rating and black_rating were categorised as seen in tables 5, 6 and 7. The new names given to categories of variables time, white_rating and black_rating are presented as well.

Table 5: The intervals utilised to categorise variable turns. As the average number of turns in this data is around 60, it seemed appropriate to use that value as a middle point, with two categories above and two below it.



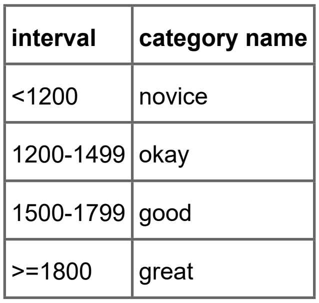
interval
1-30
31-60
61-90
>90

Table 6: The intervals utilised to categorise variable time and the corresponding category names. Note that variable time represents the amount of time in minutes that **one** player is given to make their moves.



interval	category name
0-5	fast
6-14	normal
>=15	slow

Table 7: The intervals utilised to categorise variables white_rating and black_rating and the corresponding category names. Note that on lichess.org, a player's rating starts at 1500 and can then either improve or deteriorate depending on the results of the games they play. Therefore it seemed appropriate to choose 1500 as the middle point, with two categories above and two below it.



interval	category name
<1200	novice
1200-1499	okay
1500-1799	good
>=1800	great