1. Topic:

Chess opening choice and success for low rated players

Chess Grandmasters Hikaru Nakamura and Ben Finegold both independently claimed that players under a rating of 2000elo should not spend their time studying opening theory, and that for these low-rated players openings don't matter.

As a low-rated chess player myself I am suspicious of this claim.

It is my hypothesis that low rated players who study openings do better than players who do not.

I believe that this topic is challenging enough to allow for meaningful analysis that will let our team show off the data-science skills that we have developed so far, yet focused enough that we will be able to manage the task in the allocated time without having to slave over piles of dirty data that may never come up clean, and without getting ourselves tangled up in abstruse philosophical discussions that prove irresolvable.

2. Planning the analysis

2.1 Data selection

Chess is a game that has always been close to the hearts of mathematicians, and computer scientists, as such there is an abundance of data surrounding this game dating back centuries. Selecting an appropriate data source and deciding on an appropriate sample size to answer our hypothesis will in itself be a challenge.

Before the advent of the internet and openly available digital databases, chess data mostly consisted of recordings and analyses of the games of the very best players in the world.

Today, with platforms like Chess.com and Lichess, recording and analysing millions of games a day, and billions a year, there is more data available than we could ever hope to analyse.

Chess.com is a commercial platform but does seem to have an API that would allow us to access statistical information. This may be quite limited and we may not get all the information we need. For instance we may not be able to query readily see the openings that players played.

From a quick scout about it seems like documentation is also limited.

Lichess on the other hand is open-source and has detailed documentation on how to use their API so I am confident if we wanted to we could build a dataset that would help us answer our question by using Lichess data.

As chess is a popular subject for data science projects already, there are many kaggle projects that use chess datasets already, so if making our own dataset is too challenging or outside of the scope of this project I am sure we could find appropriate data from kaggle or similar data science learning platforms.

2.2 Data Preparation

The data will need likely need to be cleaned before use although I believe this would be a matter of transforming CSV data into a data frame, removing certain columns, filtering out players that we believe do not count as low-rated, and perhaps splitting our data across multiple data frames for easier comparisons.

Given that chess data is collected digitally I don't think we will have to deal with missing values and garbled text.

2.3 Analysis approach

We would need to determine the most common to the least common chess openings and split our data this way this will require the creation of summary statistics.

We will then need to look at average win, draw, and loss rates for each opening, we might then want to look at players who primarily use obscure openings and compare their win rates with those who use more common openings.

We might look at how the average ratings of both sets of players changes over time. Do players who use obscure openings tend to increase in rating more quickly than those who use common openings.

3 Choosing Visualisation Techniques

To help us understand the statistics that we gather as the project proceeds it will be necessary to create a variety of visualisations such as: bar charts, line graphs, scatter plots, violin plots, probability distributions.

These plots should be made in a simple uncluttered style, with well chosen titles, and axes labels. Flashy features such as 3D effects on bar charts should be avoided as this can often make the data more difficult to interpret and this could potentially be used to mislead.

4. Planning for Communication

Different methods of presentation should be considered.

A traditional report written with tex/latex would be my preferred choice. This way we could all share our formatting choices and easily incorporate any 2D graphics we create and any mathematical formulae that we use.

However we should also consider the possibility of creating a jupyter notebook. This would seem like a logical approach as we all got used to working within them in semester one.

Jupyter notebooks are a very good way of presenting information to students as you can include a certain amount of interactivity

We could also consider building a dashboard with something like dash by plotly.

This is a good way to present data particularly in a business setting.

It allows for a great deal of interactivity and is a good choice when you desire to present graphs that you would like your audience to be able to interact with. However dash is not so great at presenting text heavy work.

5. Reflection

In the process of coming up with this plan I struggled most with choosing a topic that had the right scope.

I first considered looking at wine reviews but found lots of garbled text based data and not more to do with it.

I then considered looking at growing inequality in the UK. There I found a quagmire of philosophical conundrums and disparate scattered data sets to difficult to comprehend and wrangle into anything manageable.

This question of whether a subset of chess players who play obscure openings do better than their peers appealed to me due to the quality of the data and the narrowness of the scope.

I’m not sure what challenges we will face during the project if we choose to go with my idea. I don’t find any of this easy so I tend to try to deal with most challenges the same way I deal with elephants: one bite at a time.