CS306
Data Analysis and Visualization

Course Instructor: Prof. Pankaj Kumar
Teaching Assistant: Mr Umang Patel

Project Report

Mihir Desai	201801033
Hemang Nakarani	201801158
Bhargav Dave	201801402

Contents

1 Introduction

A chess game has a variety of data points available in it, especially one which is being played on a professional level or in an online platform. Online platforms are a good way to analyze professional chess as they provide a sort of dummy environment. LiChess is one of the major platforms for chess at the moment, we have extracted a dataset of games played in a small time frame on the platform (in the entire month of September 2020 to be precise), which contains thousands of games played over that time interval.

Chess is a game which has a myriad of data including win-loss, white/black ELO rating, rating changes, time taken per game, time taken per move, openings, win rates of openings, accuracy of moves, types of victories and so on. Hence a dataset with thousands of games contains a huge amount of data.

What we do in our analysis is we take a set of hypothesis which maybe common knowledge or intuitive regarding chess games and use various methods apply Data Analysis and Visualization techniques on this data to validate or dismiss these hypothesis.

We try to extract valuable inferences and test various hypotheses. Among many, some of our hypothesis and problems tackled may include hypothesis testing about whether or ratings of players follow Gaussian distributions, finding correlations between ratings of the player with the white pieces with the one with the black pieces, finding correlation between the number of blunders and/or mistakes/inferior moves and the average rating of the game, clustering in order to identify how lower tier, middle tier and higher tier players approach games and play a particular opening and so on, alongside an analysis of correlations between the various statistics.

The methods we use include KS-Test, ANOVA, K-Means Clustering, Pearson Correlation, Linear Regression and Gaussian Curve Analysis.

2 About the Dataset and regarding Data Cleaning

2.1 Information

Lichess.org is a 100% free website where people can play chess. The games played on the website are collected in monthly extracts that are stored in here: `https://database.lichess.org/`. The Dataset(1) used is made from the extracts of the above mentioned site for the month of September 2020.

It contains the data of the players, their ratings, number of turns taken, total moves, the black and white blunders, mistakes, results, the openings etc, This dataset has around 900MB of data and has 40 columns and over 8 lacs rows, which makes it a very good dataset to conduct

a study on considering the studies we intend to cover.

2.2 Cleaning

There were 3739909 rows in the original dataset, with some of the values having NaN value. As the dataset is already huge, the rows having a single NaN value have been removed. After removing those, 3716475 rows were left, i.e. 23434 rows were removed.

3 Distribution Analysis

3.1 Distribution Results for ELO

- We start off by analyzing the distribution of White and Black ELO ratings. Our hypothesis here is that both of these distributions should be follow a normal distribution
- The reason for the same is that because these are essentially measures of the skill of a chess player, it should follow a Gaussian Curve and should hence be a normal distribution

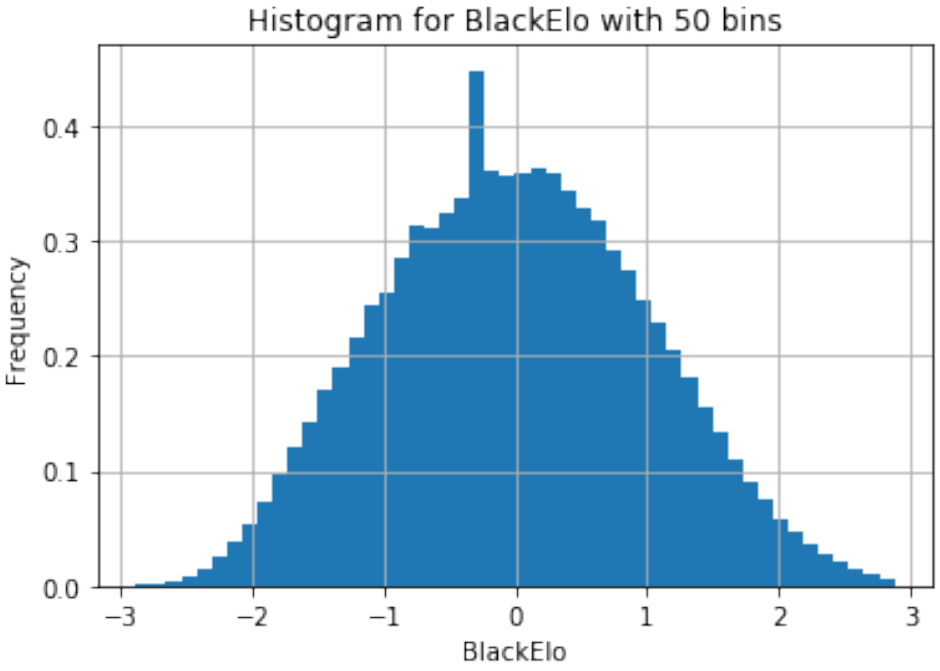

Figure 1: Histogram of Black Players' Rating for all games

As can be seen in figures 1 and 19, from a visual standpoint the data seems to follow a gaussian distribution, however when we plot the box plot of the same along with the two quartiles, we can see that the distribution is not exactly Gaussian. (Note that the spike we can see at 1500 ELO is due to the fact that whenever a player starts out on a website the starting rating is 1500 and hence due to all the new players, the rating 1500 appears an overwhelmingly large number of times)

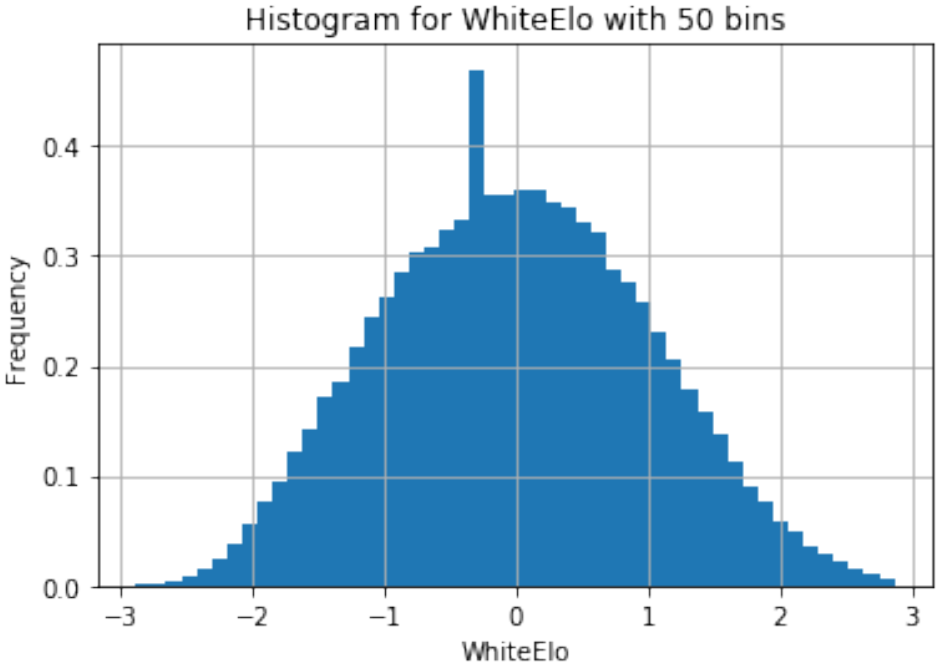

Figure 2: Histogram of White Players' Rating for all games

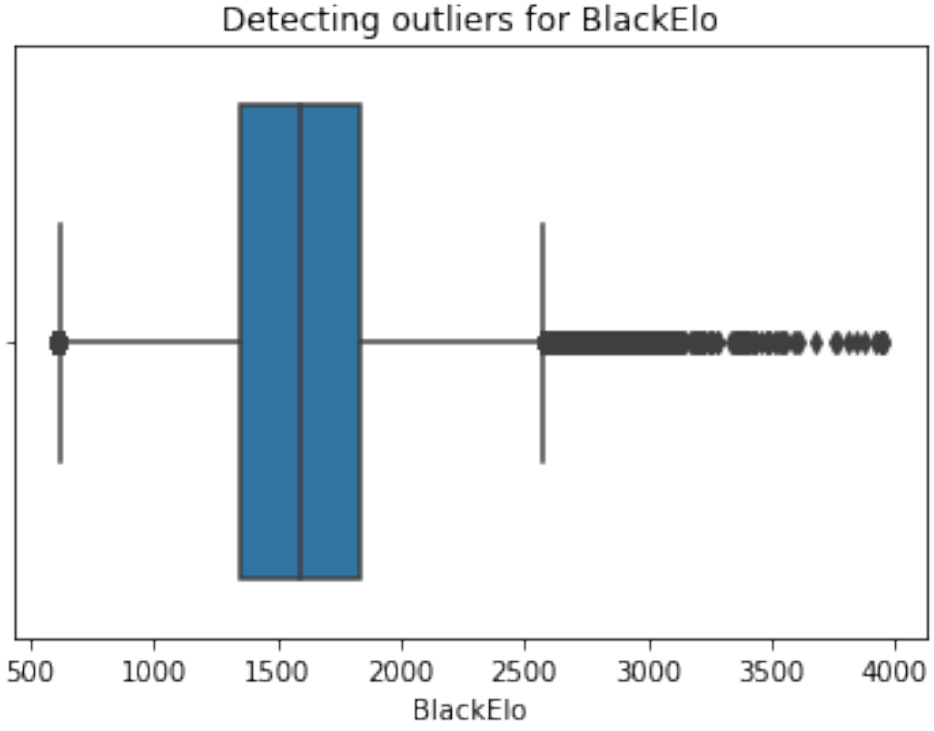

Figure 3: Box Plot of White Players' Rating for all games

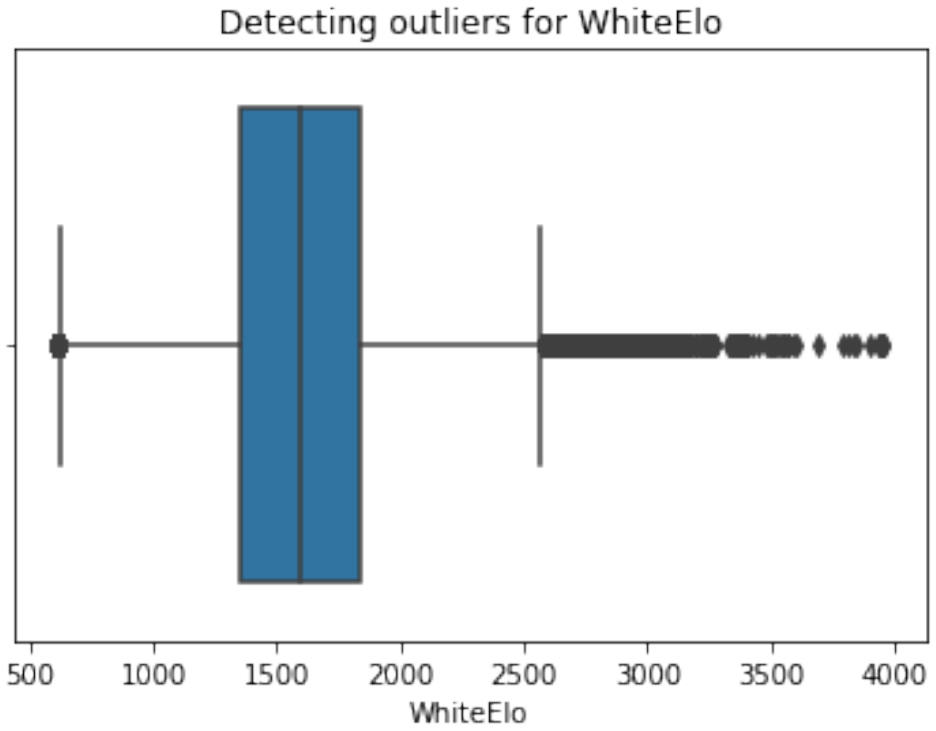

Figure 4: Box Plot of White Players' Rating for all games

As we can clearly seen the observations with rating above the 3^{rd} quartile are large and more spread out as compared to the observations below the 1^{st} quartile. Hence we can see that the distribution is not exactly Gaussian which we can determine by carrying out the KS-Test in the following section

3.2 Kolmogorov-Smirnov Test

KS Test is a factor used to determine the similarity between a given distribution and another known distribution. Here we use it to determine the similarities between the given distributions of the ELO rating and the normal distribution as shown in the following figure:

The results of our KS-test are:

- The D-value for the KS Test for the Black Elo is 0.01734.
- The D-value for the KS Test for the White Elo is 0.01955.
- For the value of $\alpha = 0.2$, the null hypothesis for both of the ratings, Black Elo and White Elo are rejected as,

  $$D_{critical} = c(\alpha)\sqrt{\frac{m+n}{mn}}$$

  and for $\alpha = 0.2$, $c(\alpha) = 1.073$, and here $m = n$ and the critical value is 0.0007 which is less than both of the D-value of the KS-test performed for both Elo ratings.

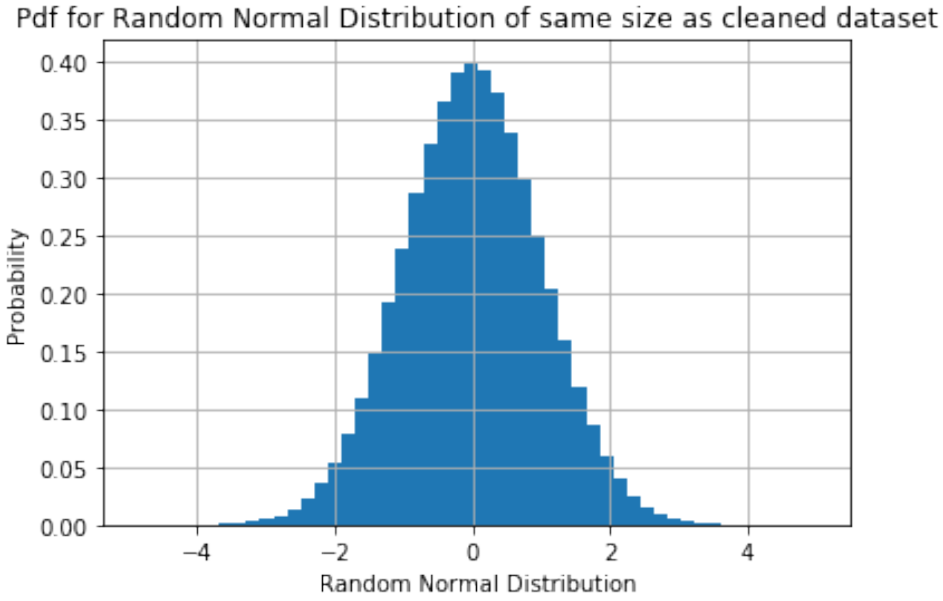

Figure 5: Normal Distribution

3.3 Analysis of Variance

Here we perform ANOVA between the Elo Ratings of Black and White Players after removing the outliers, and considering them to be kind of showing the same kind of behaviour. So the Null Hypothesis in this case would be that both the distributions of Black and White Elo Rating are taken from the same population.

Source	Sum of Squares	Degress of Freedom	Mean Square	F-value
Between samples	1.089911×10^6	1	1.089911×10^6	9.558899
Within samples	8.436037×10^{11}	7398696	1.140206×10^5	

4 Correlation Analysis and Linear Regression

In this section we use Pearson Correlation in order to determine whether we can draw any inferences or see some sort of a relationship/correlation between various parameters of a game. We also plot the linear correlation between the two in order to show a general trend in the data.

The overall correlation can be given by:

We can see clearly from this figure that various interesting relationships can be identified using the amount of correlations between two features. For example, we can see that if the correlation between white/black inferior moves and game flips is 0.71, which implies that the more the inferior moves the players make the higher the chances of the game flips. The same can also be said for the correlation between the white inferior moves and the black inferior moves, which is also close to 0.7 implying a relationship which shows a positive linear relationship between the two and hence we can infer that in a game where one side makes more inferior moves, the other side also can potentially make more inferior moves, which is an interesting observations

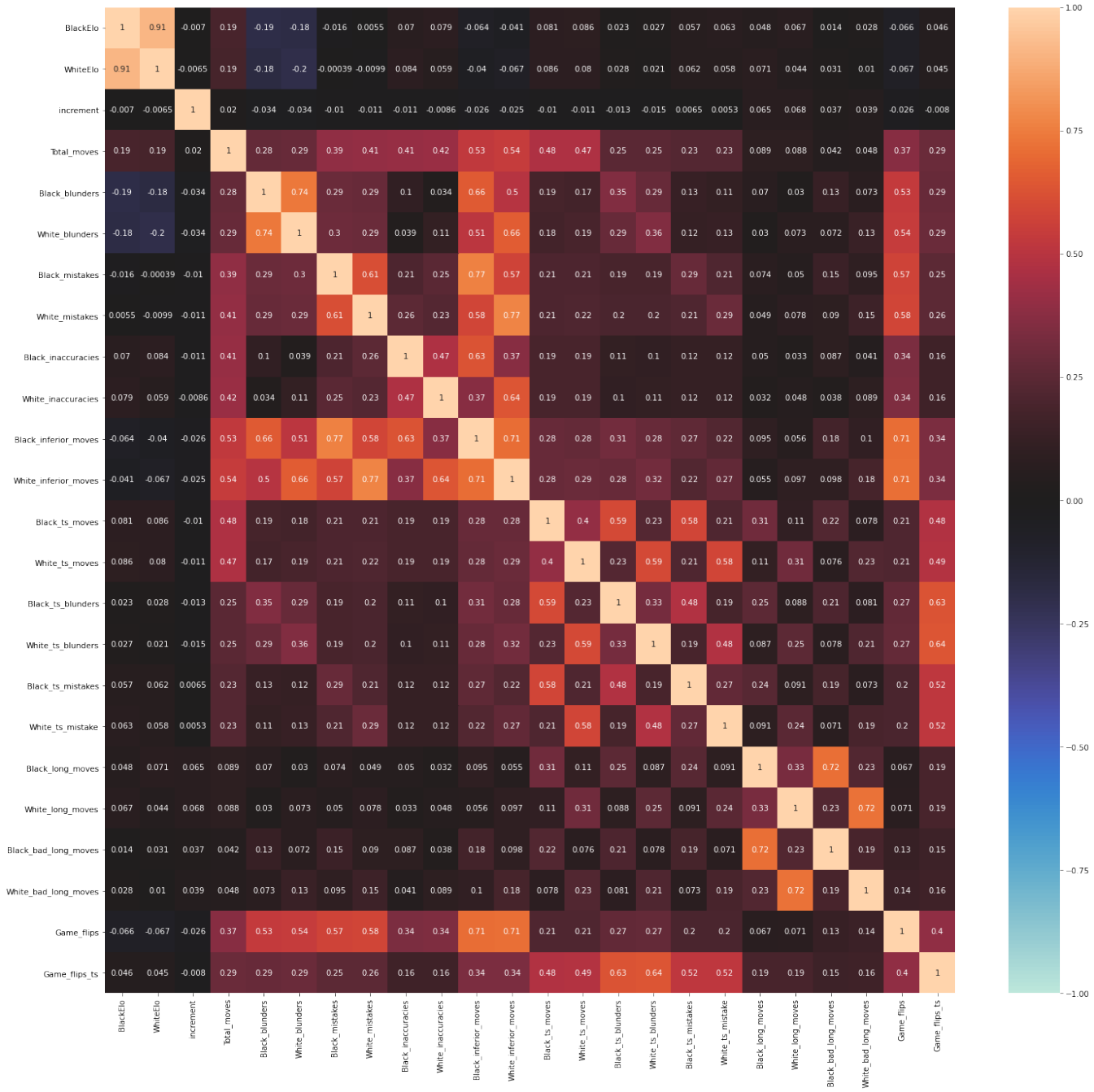

Figure 6: Correlation between the Various Features of the Data

considering the psychology of the game. Similarly we can also see some positive correlations between the number of Black Blunder and Black Time Scramble moves and hence we can say that the more the moves a player makes in a time scramble the higher the number of blunders which is an intuitive conclusion. However the correlation is not as high as one would expect and the relationship is fuzzy linear, this time scrambles do not imply higher number of blunders in all cases but have a fuzzy linear relationship. We can point out some other such other infer-

ences which come as a result of the correlation table. We list some of it in the following sections.

4.1 Correlation between Black Rating vs White Rating

Here we check our hypothesis that there should be high correlation between the ratings of the player with the white pieces and the one with the black pieces in a particular game. This can be attested to the good matching algorithm of the website which for the sake of fairness would match highly rated players with other similarly high rates players and vice verse. The plot of the same is given by:

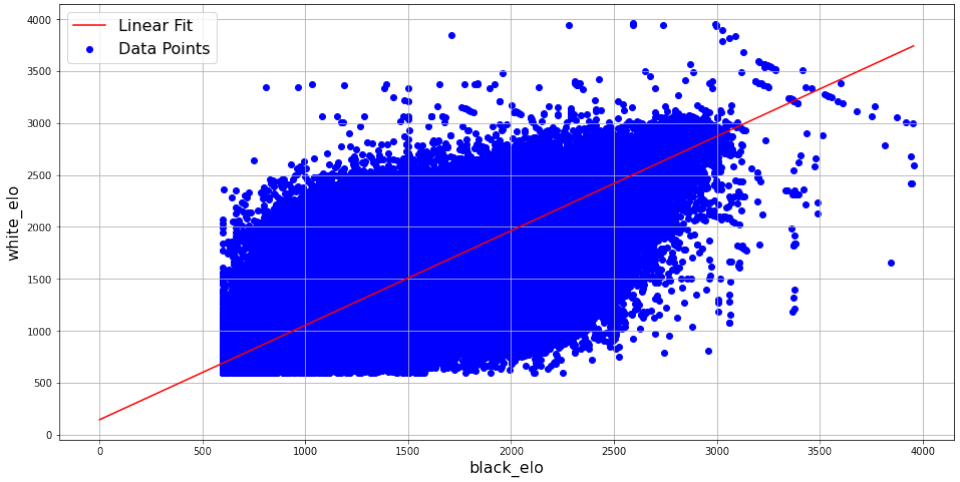

Figure 7: Plot of Black ELO vs White ELo and Correlation

The correlation between the two is almost 0.91 which implies a strong positive linear relationship and hence we can conclude that in a game, if the ELO rating of the white player is high, the same would be the case or the player with the black pieces

4.2 Rating vs Inferior Moves in Rapid Games

It was our hypothesis that in Rapid games, as the rating of a player increased, the player would make less and less inferior moves, this is because the higher rated the player, the better they are at the game and hence they would make lesser and lesser moves which a re inaccurate. Thus according to this hypothesis the correlation between Rating and inferior moves must be negative, as rating goes up the number of inferior moves go down, however, we can clearly see that that is not true with the following figure:

Hence with the above figure we can say that on an average, a player makes the same number of mistakes and inaccuracies and blunders in a game regardless of their rating and that the rating does not affect the number of inferior moves since the correlation between the two is -0.0083, which is very less showing no relation

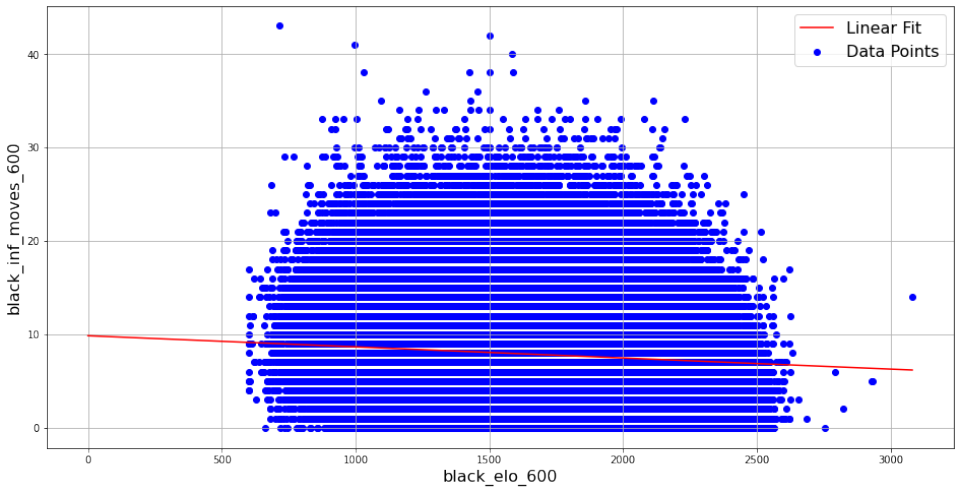

Figure 8: ELO vs Blunders for a rapid game with time control 600+0

4.3 Total Moves vs Inferior Moves

Next we test the hypothesis that the longer a game goes the higher the chances of a player starting to make inaccuracies and inferior moves because of the longer game being stressful, the human accuracy not being so high and lastly because of the sheer number of moves being made. The plot for the same is as follows:

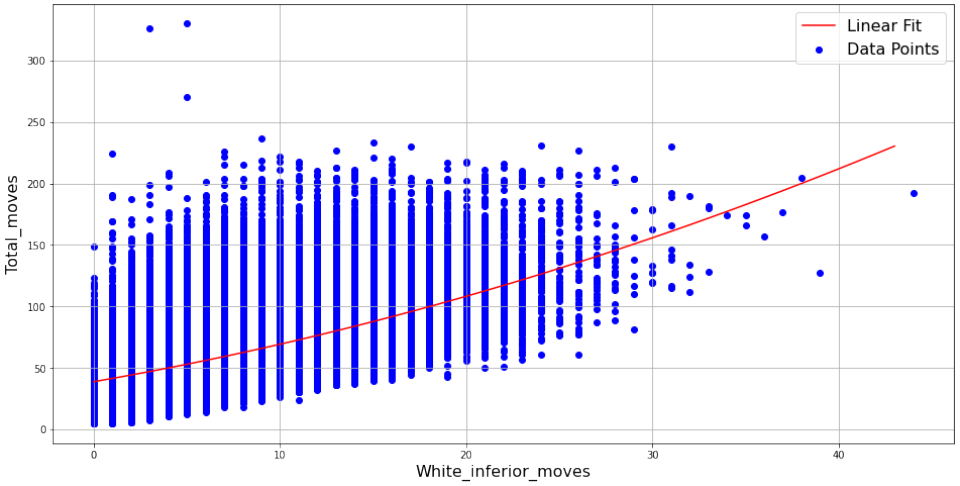

Figure 9: Total Moves vs Inferior Moves

Here the Pearson Correlation Coefficient was found out to be 0.556, which implies a fuzzy linear relationship hence we can say that our hypothesis is not exactly false and more often than not, the game with the higher number of total moves will have a higher number of inferior moves

4.4 ELO Rating vs Game types

This is not a correlation analysis but an analysis of whether or now we can say by looking at the rating of the game which type of game it was. The following plot shows the average rating of a game vs the type of game. It can be clearly seen that correspondence games are played between lower rated players whilst high rated players usually prefer playing Bullet Chess. We can see that the longer the time control the lower the ratings of the player who play it on average

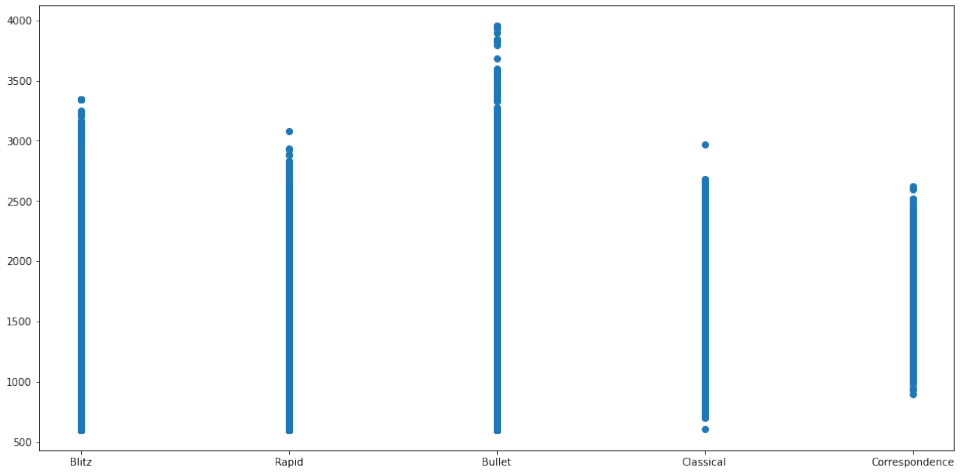

Figure 10: ELO vs Game type

5 Clustering Analysis

In this section, we do a more qualitative data analysis in order to determine the answers to a few questions:

- What are the most played openings?
- What are the openings played when white wins?
- What are the openings played when black wins?
- Given a set of openings, can we use K-means clustering to see which section of players (low rated, high rated and average rating) play it the most?
- What are the length of the games which the aforementioned openings are played?

5.1 Visualization of Openings

The following Bar Charts show the most popular openings for overall games, for games when white won and for games when black won. As we can see in Figure 11, Queen's Pawn Game leads the most popular openings as it is also considered the 2^{nd} most popular opening after 1. e4 and the King's Pawn openings have various names. Next, we can see in Figure 12, the most popular opening when White Won were also very similar to the openings in Figure 11 this is

because White usually dictates the openings being played. Lastly Figure 13 shows the openings when Black Won the most and we can see that Sicilian reigns supreme as is expected.

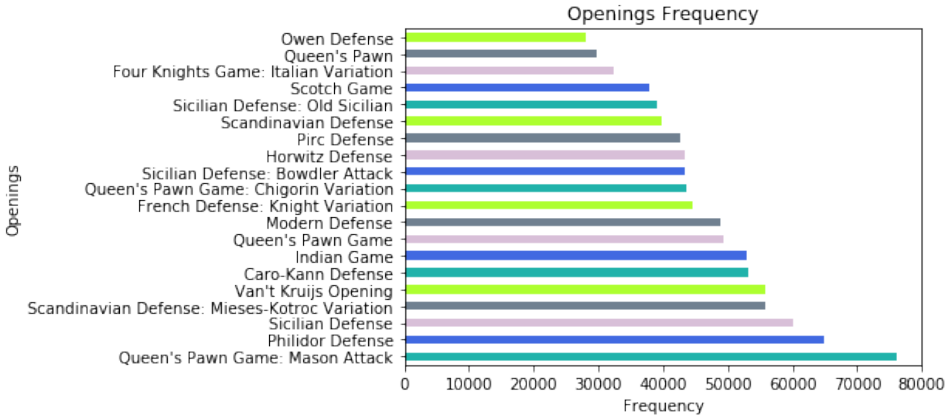

Figure 11: Most popular openings

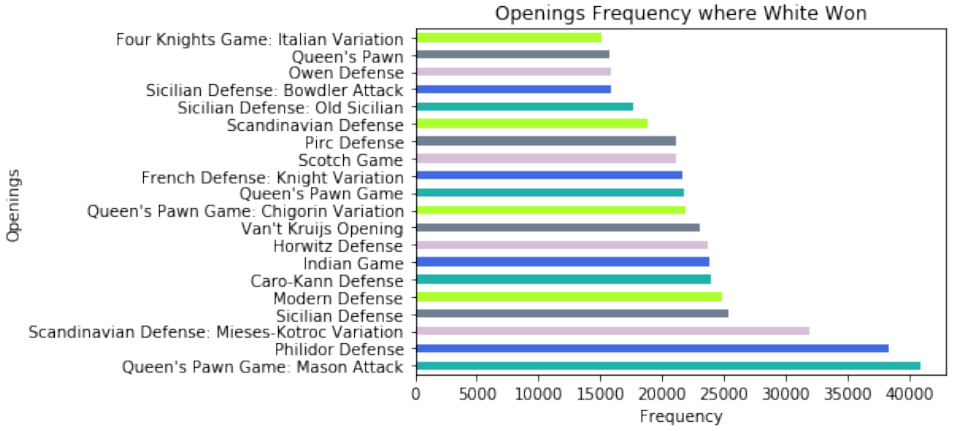

Figure 12: Openings played when White won

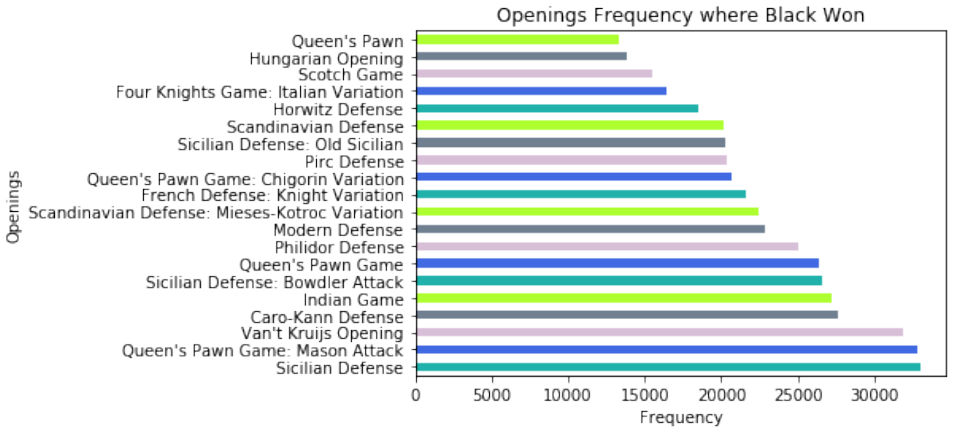

Figure 13: Openings played when Black won

5.2 K means clustering and related analysis

Here we perform clustering on various openings which are popular and obscure to see the number of moves and the category of players which play it often. The analysis according to the cluster is as follows:

- A03: Bird's opening is historically considered an irregular chess opening, however it has seen surge in popularity in modern times as we can see that a lot of players play it lately, however we can clearly see that a lot of those games have less moves and hence the opening is not sound, plus it is seen to be played at mid to high levels of play
- B03: Alekhine's defense is an opening which is preferred at higher levels and we can clearly see this happens seeing the cluster
- C35: King's Gambit is one of the rarest openings in chess and is something that is only played soundly at higher levels, all of this can be seen in the clusters, with the most games being played at the higher levels and virtually no games at lower levels
- C54: Giuoco Piano is one of the most basic openings which is very popular amongst low to mid level players which can also be seen in the clusters
- D06:Queen's Gambit is one of the most popular openings played along all the levels which can be seen in the clusters shown in the figure

The following figure shows the elbow plot from which we can see that k=3 can be a decent value for number of clusters.

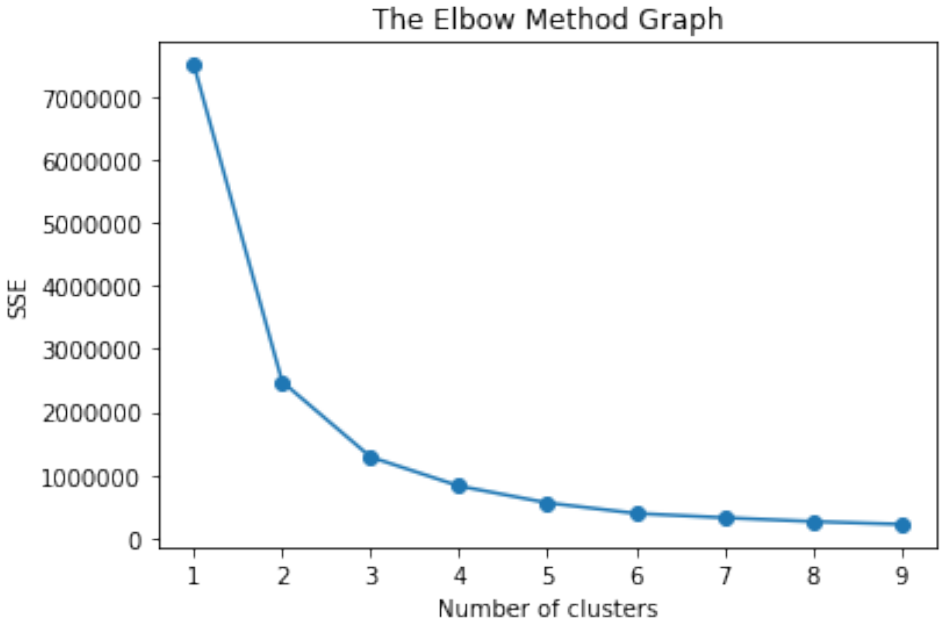

Figure 14: Elbow Plot for A03 (is almost the same for all openings)

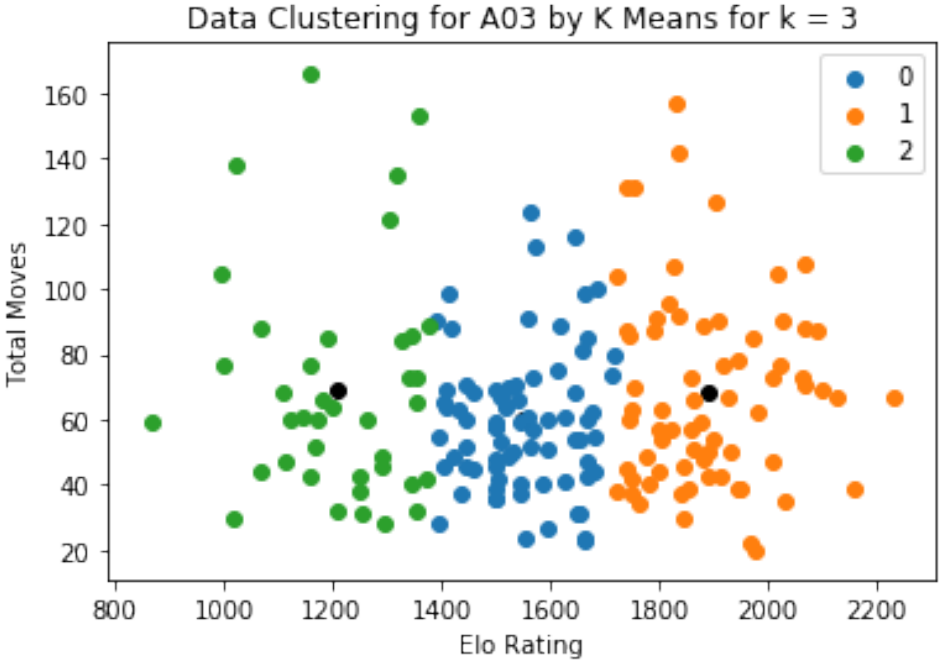

Figure 15: Clustering for A03

Figure 16: Clustering for B03

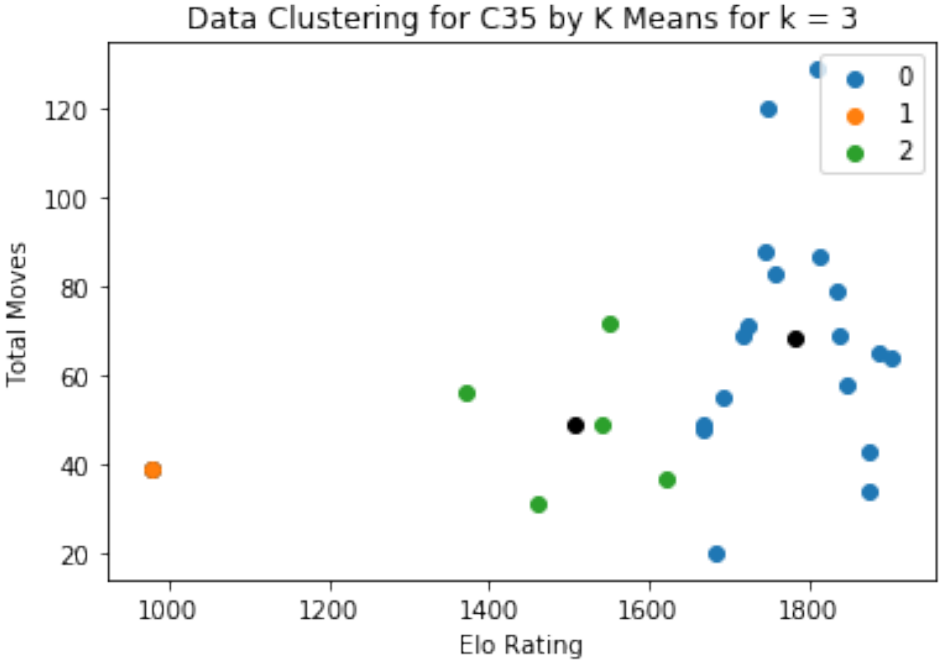

Figure 17: Clustering for C35

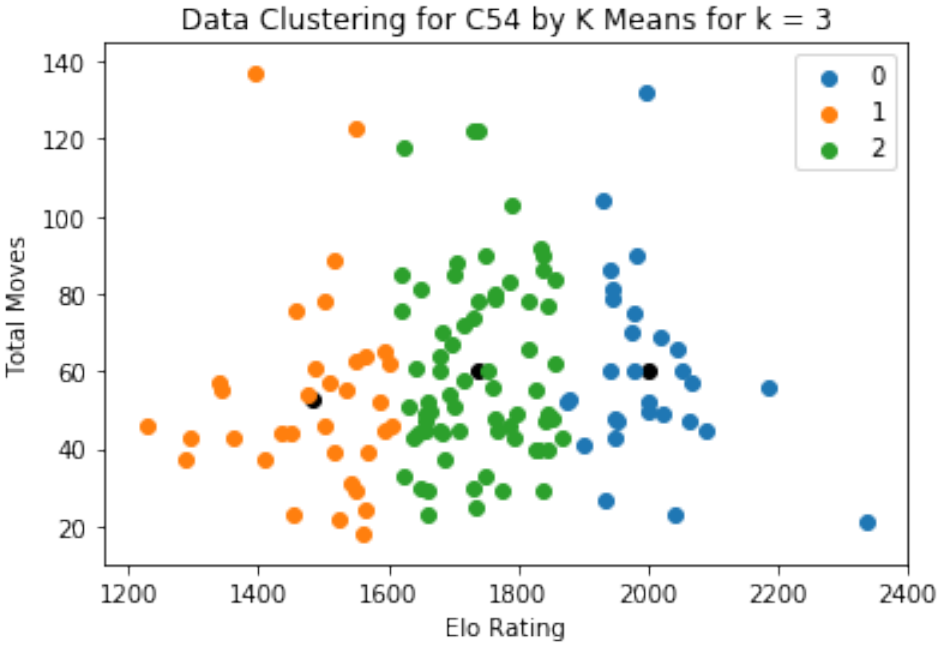

Figure 18: Clustering for C54

Figure 19: Clustering for D06

References

[1] *https://www.kaggle.com/noobiedatascientist/lichess-september-2020-data?select=Sept_20_analysis.csv*

[2] *https://www.365chess.com/eco.php*

[3] Our GitHub repository *https://github.com/desaimihir12/lichess-analysis*

[4] *https://en.wikipedia.org/wiki/Queen%27s_Pawn_Game*

[5] *https://en.wikipedia.org/wiki/Sicilian_Defence*