Proj 9

Elizabeth Bell,Chris Bryan, Meera Haridasa

December 7, 2017

Getting our data from the web:

UNIX CODE:

#wget -A zip -m -p -E -k -K -np https://www.pgnmentor.com/files.html  
  
#unzip '\*.zip'  
  
#$ cat \*.pgn > combined.pgn  
  
#//Remove lines without brackets:  
  
 #sed -i '/\[/!d' combined.pgn  
  
#//For games that do not contain an ECO code, place an "NA" Eco variable  
  
#cat combined.pgn | awk '{ if(count == 9) { if(substr($0,1,4) == "[ECO") {print $0; count = 0;} else { print "[ECO \"\"]"; print $0; count = 1;}} else {print $0; count++;}}' > combinedNEW.txt  
  
#//Isolate lines with brackets, place in new file  
#$ sed -i '/\[/d' combined2.pgn  
  
#//Delete blank space. Put all in one line. Count number of fields  
#cat combined2.pgn | sed 's/.$//' | awk -F\. 'BEGIN{prevblank = 1}{if (NF == 0) {print $0; prevblank = 1;} else if(prevblank ==1){prevblank = 0; printf $0;} else if (prevblank == 0) {prevblank = 0; printf ""; printf $0;}} END {printf "/n"}' | awk -F\. '{print NF}' | sort | uniq -c | sort -n > numMoves.txt

IN R:

pgn <- read.table("~/290/gp09g2/scratch/www.pgnmentor.com/openings/combinedNEW.txt", quote="", sep="\n", stringsAsFactors=FALSE)  
# get column names  
colnms <- sub("\\[(\\w+).+", "\\1", pgn[1:10,1])  
pgn.df <- data.frame(matrix(sub("\\[\\w+ \\\"(.+)\\\"\\]", "\\1", pgn[,1]),  
 byrow=TRUE, ncol=10))  
names(pgn.df) <- c("Event", "Site", "Date", "Round", "White", "Black", "Result", "WhiteElo", "BlackElo","ECO")

QUESTIONS

1. In chess, each opening sequence of moves has a name and a corresponding code, known as an ECO code. Out of these opening sequences, which ones will lead to the highest win percentage for white? How about for black?

WHITE:

factors <- levels(pgn.df$ECO)  
pgnWINS <- pgn.df[pgn.df$Result == "1-0", ]  
  
winsExtract <- function(name)  
{  
 win <- length(grep(name, pgnWINS$ECO))  
 total <- length(grep(name, pgn.df$ECO))  
 names <- name;  
 winpct <- win / total  
   
}  
winpct <- sapply(factors, winsExtract)  
winpct <- sort(winpct, decreasing = TRUE)  
  
head(winpct)

## D37/22 C62 E08 C70 C40 A74   
## 1.0000000 0.5689714 0.5264741 0.5220126 0.5166517 0.5134100

mean(winpct)

## [1] 0.3871814

The 5 highest win percentage openings for white are C62(Ruy Lopez, Old Steinitz Defense), E08 (Catalan, Closed), C70(variant of Ruy Lopez), C40(King's Knight Opening), and A74(Benoni Classical, 9.a6) White's average win percentage is .3871814.

BLACK:

pgnBLACK <- pgn.df[pgn.df$Result == "0-1", ]  
  
winsBlackExtract <- function(name)  
{  
 winB <- length(grep(name, pgnBLACK$ECO))  
 totalB <- length(grep(name, pgn.df$ECO))  
 namesB <- name;  
 winpctB <- winB / totalB  
   
}  
winpctB <- sapply(factors, winsBlackExtract)  
winpctB <- sort(winpctB, decreasing = TRUE)  
  
head(winpctB)

## C38 B59 B95 E24 C20 B50   
## 0.5185185 0.4880531 0.4550847 0.4479638 0.4375000 0.4316354

mean(winpctB)

## [1] 0.2863016

The 5 highest win percentage openings for black are C38(King's Gambit Accepted), B59 (Sicilian Boleslavsky Variation), B95 (Sicilian Najdorf), E24 (Nimzo Indian Samisch), C20(King's Pawn Game), and B50 (Sicilian Variation) Black's average win percentage is .286302.

1. In chess elo, the "expected score" is defined as the expected win percentage plus expected draw percentage / 2. By following the elo formula, a player rated 100 higher than their oppenent should have an expected score of .64, while a player rated 200 higher should have an expected score of .76. Does this match up with the data?

WhiteHigher <- pgn.df[(as.numeric(pgn.df$WhiteElo) - as.numeric(pgn.df$BlackElo)) == 100, ]  
length(grep("1-0", WhiteHigher$Result))/ 4144 + (length(grep("1/2-1/2", WhiteHigher$Result)) / 4144) / 2

## [1] 0.6691602

#White's expected score is .6691602   
  
White200Higher <- pgn.df[(as.numeric(pgn.df$WhiteElo) - as.numeric(pgn.df$BlackElo)) == 200, ]  
length(grep("1-0", White200Higher$Result))/ 1988 + (length(grep("1/2-1/2", White200Higher$Result)) / 1988) / 2

## [1] 0.7811871

#White's expected score is .7811871  
  
BlackHigher <- pgn.df[(as.numeric(pgn.df$BlackElo) - as.numeric(pgn.df$WhiteElo)) == 100, ]  
length(grep("0-1", BlackHigher$Result))/ 2494 + (length(grep("1/2-1/2", BlackHigher$Result)) / 2494) / 2

## [1] 0.5771852

#Black's expected score is .5771852  
  
Black200Higher <- pgn.df[(as.numeric(pgn.df$BlackElo) - as.numeric(pgn.df$WhiteElo)) == 200, ]  
length(grep("0-1", Black200Higher$Result))/ 2149 + (length(grep("1/2-1/2", Black200Higher$Result)) / 2194) / 2

## [1] 0.6779066

#Black's expected score is .677906

White's expected scores are higher than the formula suggests, while black's are lower. This indicates that a higher rated player with white pieces will have a larger advantage than a higher rated player with black pieces. The simple explanation for white having a higher expected score is that the player with the white pieces gets to go first and that is an advantage.

1. Best player in the world's opening moves win percentages?

Magnus Carlten is the best player in the world and these are his win percentages based on the ECOs in the data (the opening moves). In chess there is wins, losses, and ties and so in the data set I made wins 1 point, losses 0 points, and ties .5 points. That is why some of the percentages are close to .5 because in chess ties are very common. Also some win percentages may be 1 or 0 and that may not be because it is a really bad or good opening move but rather that he has only done that move 1 or a few times and thus it wouldn't be a good average.

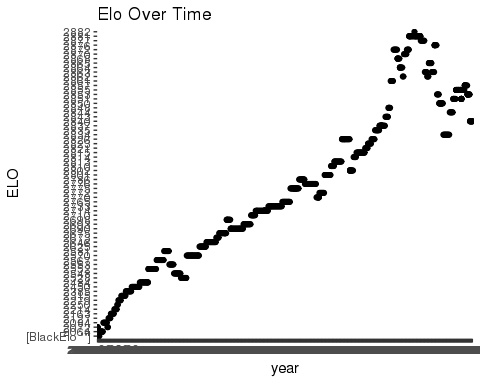
MCWHITE.df <- subset(pgn.df, pgn.df$White == "Carlsen,M" | pgn.df$White == "Carlsen,Magnus")  
MCWHITE.df$WL[MCWHITE.df$Result == "0-1"] <- 0   
MCWHITE.df$WL[MCWHITE.df$Result == "1-0"] <- 1  
MCWHITE.df$WL[MCWHITE.df$Result == "1/2-1/2"] <- .5  
  
MCBLACK.df <- subset(pgn.df, pgn.df$Black == "Carlsen,M" | pgn.df$Black == "Carlsen,Magnus")  
MCBLACK.df$WL[MCBLACK.df$Result == "0-1"] <- 1   
MCBLACK.df$WL[MCBLACK.df$Result == "1-0"] <- 0  
MCBLACK.df$WL[MCBLACK.df$Result == "1/2-1/2"] <- .5  
  
MC.df <- rbind(MCWHITE.df, MCBLACK.df)  
sort(tapply(MC.df$WL, MC.df$ECO, mean, na.rm = TRUE))

## A13 A62 B37 B59 B73 C17 C26   
## 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000   
## C61 C71 C96 D05 D20 D48 D89   
## 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000   
## E14 E29 E30 A33 A80 B87 C19   
## 0.0000000 0.0000000 0.0000000 0.1666667 0.1666667 0.2500000 0.2500000   
## C24 D07 D76 E05 C86 B65 B97   
## 0.2500000 0.2500000 0.2500000 0.2500000 0.3000000 0.3333333 0.3333333   
## C28 C43 C92 E24 B22 B46 D82   
## 0.3333333 0.3333333 0.3333333 0.3333333 0.3750000 0.3750000 0.3750000   
## E92 B20 B04 C63 A20 A45 B35   
## 0.4000000 0.4166667 0.4230769 0.4285714 0.4444444 0.4444444 0.4615385   
## E32 A09 A14 A15 A21 A25 A27   
## 0.4750000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## A30 A31 A34 A38 A39 A61 B02   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## B03 B05 B14 B25 B36 B48 B54   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## B56 B60 B61 B77 C25 C33 C39   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## C42 C56 C80 C81 C89 C97 D11   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## D13 D14 D27 D41 D46 D55 D61   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## D63 D77 D78 D83 D88 D90 D93   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## E04 E09 E16 E18 E33 E34 E35   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## E39 E43 E47 E48 E50 E61 E63   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000   
## E65 E68 E76 E90 E98 D37 E06   
## 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5156250 0.5294118   
## B45 C83 D17 B33 B12 B76 E15   
## 0.5500000 0.5500000 0.5500000 0.5571429 0.5588235 0.5714286 0.5724638   
## A07 B51 B90 B23 C02 C41 D16   
## 0.5750000 0.5769231 0.5769231 0.5833333 0.5833333 0.5833333 0.5833333   
## E21 E17 C67 B19 D44 E97 D38   
## 0.5882353 0.5909091 0.5980392 0.6000000 0.6000000 0.6000000 0.6060606   
## A29 C91 D47 C84 A11 A36 A59   
## 0.6111111 0.6111111 0.6111111 0.6206897 0.6250000 0.6250000 0.6250000   
## B10 C08 C54 C60 C68 D85 D86   
## 0.6250000 0.6250000 0.6250000 0.6250000 0.6250000 0.6250000 0.6250000   
## D91 B31 B42 A17 C45 C90 D87   
## 0.6250000 0.6333333 0.6333333 0.6428571 0.6428571 0.6428571 0.6428571   
## B30 B92 D45 C65 A57 B09 B38   
## 0.6500000 0.6562500 0.6562500 0.6595745 0.6666667 0.6666667 0.6666667   
## B78 B84 C09 C49 C55 C69 D52   
## 0.6666667 0.6666667 0.6666667 0.6666667 0.6666667 0.6666667 0.6666667   
## E00 E13 E25 E54 E55 C77 B52   
## 0.6666667 0.6666667 0.6666667 0.6666667 0.6666667 0.6764706 0.6818182   
## C88 A28 D97 E12 A10 B43 B67   
## 0.6818182 0.6875000 0.6875000 0.6956522 0.7000000 0.7000000 0.7000000   
## E37 B53 C50 E20 B32 E94 C53   
## 0.7000000 0.7142857 0.7187500 0.7187500 0.7222222 0.7222222 0.7272727   
## D15 [ECO ""] A16 A22 A32 A37 A46   
## 0.7380952 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000   
## A90 B01 B08 B11 B17 B18 B41   
## 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000   
## B47 B96 C00 C11 C16 C18 D12   
## 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000   
## D19 D22 D36 D43 D80 D81 E36   
## 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000 0.7500000   
## E52 E62 C95 A48 A01 A04 A58   
## 0.7500000 0.7500000 0.7592593 0.7692308 0.8000000 0.8000000 0.8000000   
## B66 B91 E11 B07 C36 C76 D31   
## 0.8000000 0.8000000 0.8000000 0.8333333 0.8333333 0.8333333 0.8333333   
## E46 A56 B06 C48 D35 A00 A03   
## 0.8333333 0.8750000 0.8846154 0.9000000 0.9285714 1.0000000 1.0000000   
## A06 A08 A35 A42 A47 A67 A70   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000   
## A87 A88 B00 B21 B26 B28 B29   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000   
## B34 B40 B50 B70 B72 B74 B79   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000   
## B80 B82 B95 C01 C03 C04 C05   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000   
## C07 C10 C15 C44 C47 C57 C66   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000   
## C78 C85 C98 D00 D18 D34 D42   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000   
## D75 D94 D99 E01 E02 E27 E38   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000   
## E42 E59 E60 E72 E75 E81   
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000

1. Best player's elo over time?

Magnus Carleton's elo over time is shown on the graph. He has been on the upward climb for years. He peaked in 2014 and has been sliding a bit ever since. This is because he is the top chess player and the only place to go from the top is down, even if he ties his rating goes down. It does not matter if you graph the white elo or the black elo; the graphs look exactly the same.

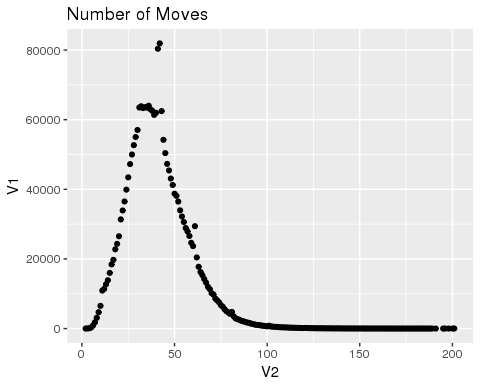
library("ggplot2")  
  
ggplot((MCBLACK.df), aes(x=Date, y=BlackElo)) + geom\_point() + ggtitle("Elo Over Time") + xlab("year") + ylab("ELO")



#ggplot((MCWHITE.df), aes(x=Date, y=WhiteElo)) + geom\_point() + ggtitle("Elo Over Time") + xlab("year") + ylab("ELO")

1. What is the distribution for the number of moves in a chess game? What is the mean and median number of moves?

numMoves <- read.table("~/290/gp09g2/scratch/www.pgnmentor.com/openings/numMoves.txt", quote="", sep="", stringsAsFactors=FALSE)  
library("ggplot2")  
ggplot((numMoves), aes(x=V2, y=V1)) + geom\_point() + ggtitle("Number of Moves")



d2 <- rep(numMoves$V2, numMoves$V1) ## expands the data by frequency of score  
  
multi.fun <- function(x) {  
 c(mean = mean(x), median = median(x), var = var(x), sd = sd(x))  
}  
  
multi.fun(d2)

## mean median var sd   
## 40.83511 39.00000 261.41170 16.16823

#mean median var and sd.

This distribution is likely NOT normal, because it has a slight right skew. The reason there some games that are super low moves like close to zero it is most likely because the person resigned the game. The average game of chess is about 40 moves.

1. What is the most popular site for the chess games? What about most popular event? Do these two have any connection?

tail(sort(table(pgn.df$Site)) , 20)

##   
## Berlin GER Prague CZE Copenhagen   
## 8771 8892 9332   
## Copenhagen DEN Helsingor DEN HUN   
## 9364 9436 9684   
## Khanty-Mansiysk RUS Reykjavik ISL Pardubice CZE   
## 11098 11456 11723   
## Warsaw POL Istanbul TUR Dresden GER   
## 11836 12199 12992   
## Buenos Aires Moscow RUS Budapest   
## 14102 20368 21910   
## Moscow St Petersburg RUS Budapest HUN   
## 24545 24855 28875   
## ? Germany   
## 40694 60484

tail(sort(table(pgn.df$Event)) , 20)

##   
## Thessaloniki ol (Men) 37th Olympiad New York op   
## 3007 3027 3156   
## Masters SVK-ch Hoogovens   
## 3159 3251 3336   
## URS-chT CSR-ch ARG-ch   
## 3396 3442 3500   
## It JUG-chT HUN-ch   
## 3511 3786 3873   
## Cappelle op JUG-ch URS-ch sf   
## 3885 4587 4891   
## Wch U20 BCF-ch Open A   
## 5395 7047 9483   
## Politiken Cup Open   
## 11544 33138

Yes, Politiken Cup is in Copenhagen DEN is one example. This data is not in the best format, like Germany is just by itself but then there is two different Copenhagen Sites. So it would be possible to go through and make this more accurate if it was really important to you and you hoped to learn something significant from this data.