Final_Project – Chess Dataset

Kevin Hou

12/12/2021

```
chess<-read.csv(file = 'games.csv')
head(chess)
```

```
##         id rated  created_at last_move_at turns victory_status winner
## 1 TZJHLljE FALSE 1.50421e+12  1.50421e+12    13      outoftime  white
## 2 l1NXvwaE  TRUE 1.50413e+12  1.50413e+12    16         resign  black
## 3 mIICvQHh  TRUE 1.50413e+12  1.50413e+12    61           mate  white
## 4 kWKvrqYL  TRUE 1.50411e+12  1.50411e+12    61           mate  white
## 5 9tXo1AUZ  TRUE 1.50403e+12  1.50403e+12    95           mate  white
## 6 MsoDV9wj FALSE 1.50424e+12  1.50424e+12     5           draw   draw
##   increment_code      white_id white_rating      black_id black_rating
## 1           15+2      bourgris         1500          a-00         1191
## 2           5+10          a-00         1322     skinnerua         1261
## 3           5+10        ischia         1496          a-00         1500
## 4           20+0 daniamurashov         1439  adivanov2009         1454
## 5           30+3     nik221107         1523  adivanov2009         1469
## 6           10+0     trelynn17         1250 franklin14532         1002
##
## 1
## 2
## 3
## 4
## 5 e4 e5 Nf3 d6 d4 Nc6 d5 Nb4 a3 Na6 Nc3 Be7 b4 Nf6 Bg5 O-O b5 Nc5 Bxf6 Bxf6 Bd3 Qd7 O-O Nxd3 Qxd3 c6
## 6
##   opening_eco                            opening_name opening_ply
## 1         D10       Slav Defense: Exchange Variation           5
## 2         B00 Nimzowitsch Defense: Kennedy Variation           4
## 3         C20  King's Pawn Game: Leonardis Variation           3
## 4         D02 Queen's Pawn Game: Zukertort Variation           3
## 5         C41                      Philidor Defense           5
## 6         B27   Sicilian Defense: Mongoose Variation           4
```

```
names(chess)
```

```
##  [1] "id"             "rated"          "created_at"     "last_move_at"
##  [5] "turns"          "victory_status" "winner"         "increment_code"
##  [9] "white_id"       "white_rating"   "black_id"       "black_rating"
## [13] "moves"          "opening_eco"    "opening_name"   "opening_ply"
```

```
str(chess)
```

```
## 'data.frame':    20058 obs. of  16 variables:
##  $ id            : chr  "TZJHLljE" "l1NXvwaE" "mIICvQHh" "kWKvrqYL" ...
##  $ rated         : chr  "FALSE" "TRUE" "TRUE" "TRUE" ...
```

```
##  $ created_at    : num  1.5e+12 1.5e+12 1.5e+12 1.5e+12 1.5e+12 ...
##  $ last_move_at  : num  1.5e+12 1.5e+12 1.5e+12 1.5e+12 1.5e+12 ...
##  $ turns         : int  13 16 61 61 95 5 33 9 66 119 ...
##  $ victory_status: chr  "outoftime" "resign" "mate" "mate" ...
##  $ winner        : chr  "white" "black" "white" "white" ...
##  $ increment_code: chr  "15+2" "5+10" "5+10" "20+0" ...
##  $ white_id      : chr  "bourgris" "a-00" "ischia" "daniamurashov" ...
##  $ white_rating  : int  1500 1322 1496 1439 1523 1250 1520 1413 1439 1381 ...
##  $ black_id      : chr  "a-00" "skinnerua" "a-00" "adivanov2009" ...
##  $ black_rating  : int  1191 1261 1500 1454 1469 1002 1423 2108 1392 1209 ...
##  $ moves         : chr  "d4 d5 c4 c6 cxd5 e6 dxe6 fxe6 Nf3 Bb4+ Nc3 Ba5 Bf4" "d4 Nc6 e4 e5 f4 f6 dxe
##  $ opening_eco   : chr  "D10" "B00" "C20" "D02" ...
##  $ opening_name  : chr  "Slav Defense: Exchange Variation" "Nimzowitsch Defense: Kennedy Variation" "
##  $ opening_ply   : int  5 4 3 3 5 4 10 5 6 4 ...
```

Dataset Explain

This dataset contains all games on lichess both rated and non-rated games. Also, it haves every game who wins who losses and drew. Plus, it also have every move in the game. There are 20058 observations and 17 variables inside of this dataset.

Why this dataset is interest to me

I start playing chess when I was 12 years old, which means I have played more than 9 years of chess. During this 9 years stretch, I have win more than 50 trophies and medals. It is one of my hobbies. Back in high school, I have started a chess club, which gave me another 2 hours to play chess. Also when I play chess, my brain will become more focus, which helps me easier to think what move should I make. I'm once a candidate master back in my country. My rating is around 2100, which is around 92 percentile on lichess.com.

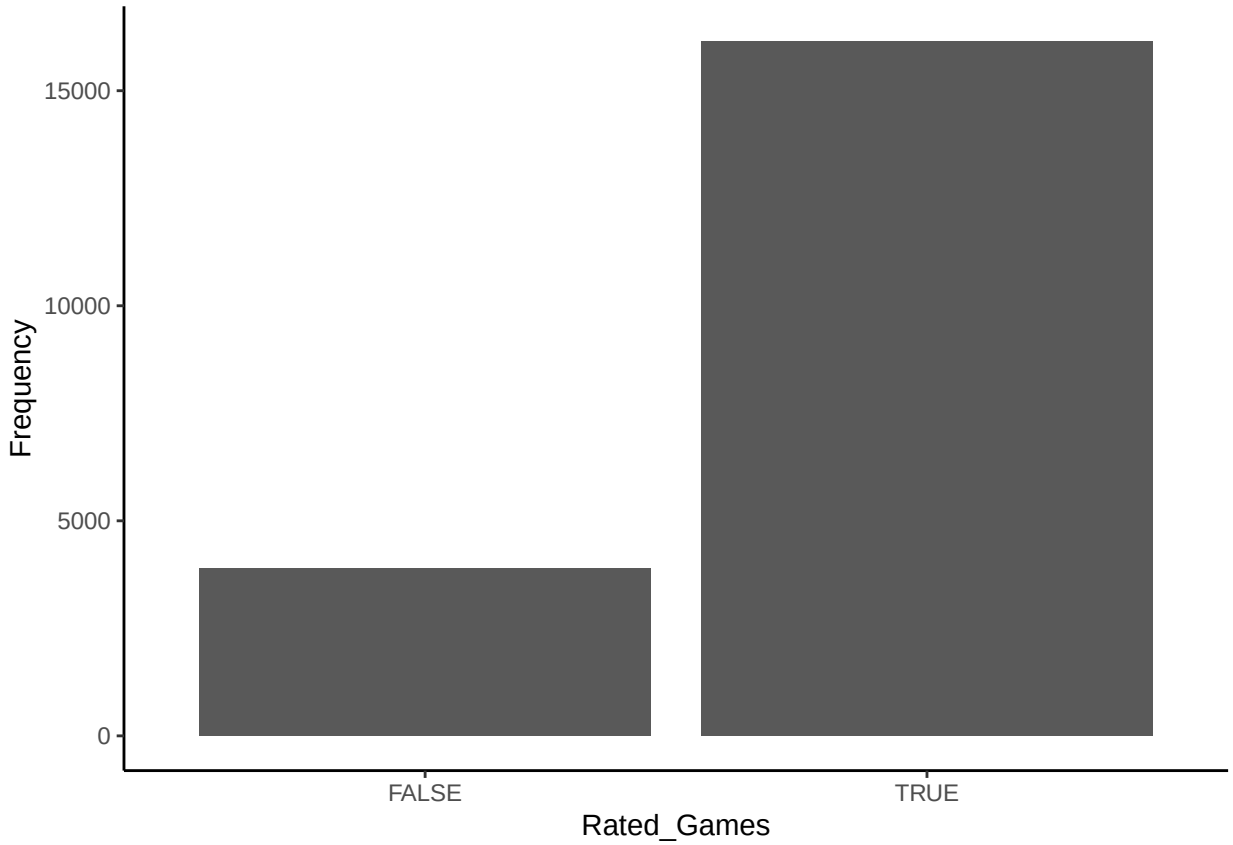
GM Hikaru

GM Hikaru once said that "I fear not the man who has practiced 10,000 openings once, but I fear the man who has practiced one opening 10,000 times" GM Hikaru once the world 2nd in chess ranking. ## State The Problem I will focus on Which opening has the highest winning percentage for both white and black.

Step 1

I want to see what is the percentage of games that are rated and not rated

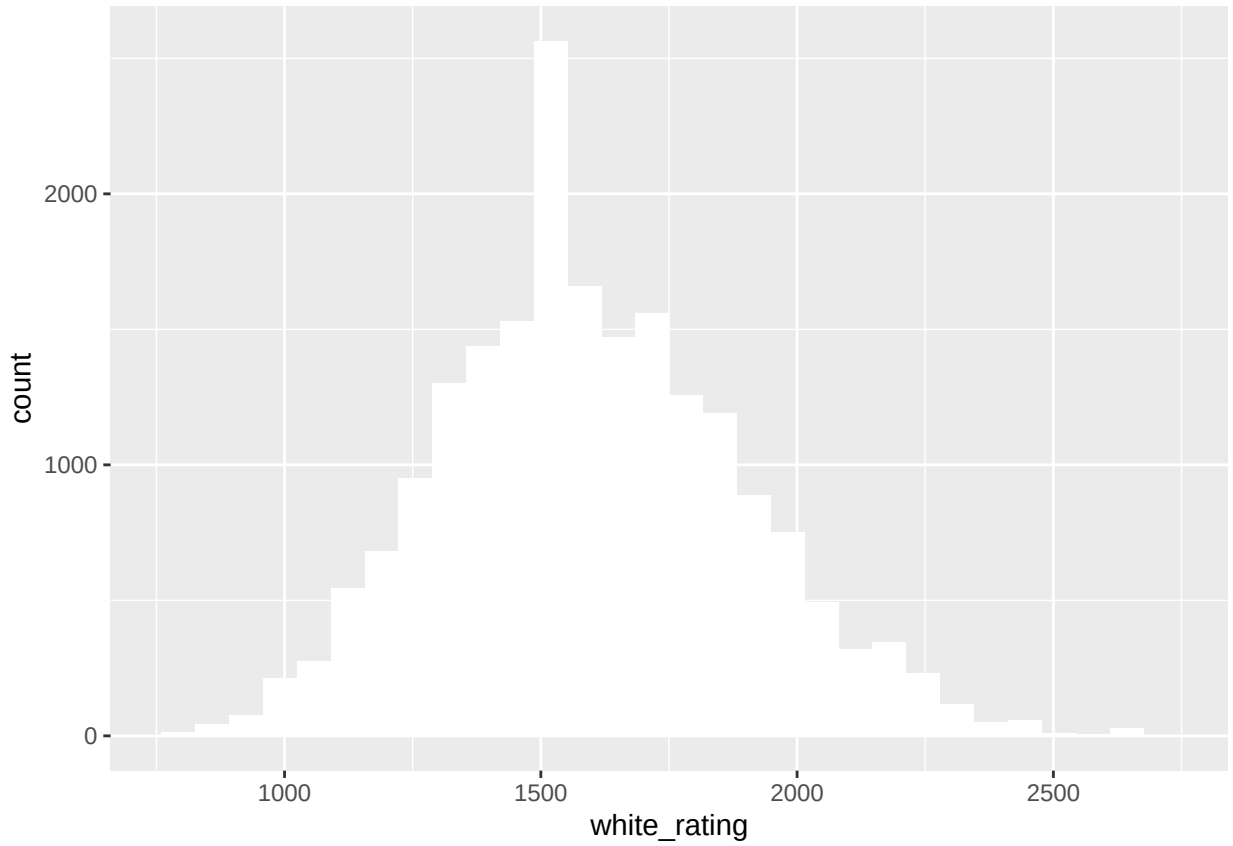
```
ggplot(chess,aes(x=toupper(rated)))+geom_bar()+xlab(label =
"Rated_Games")+ylab(label = "Frequency")+theme_classic()
```



Step 2 I want to find out the average rating of both black and white. As you see the graph below, 1500 rating has the most people for both black and white. As the rating goes up or down, the amount of people drop significantly, which means that most people are around 1500 rating range.

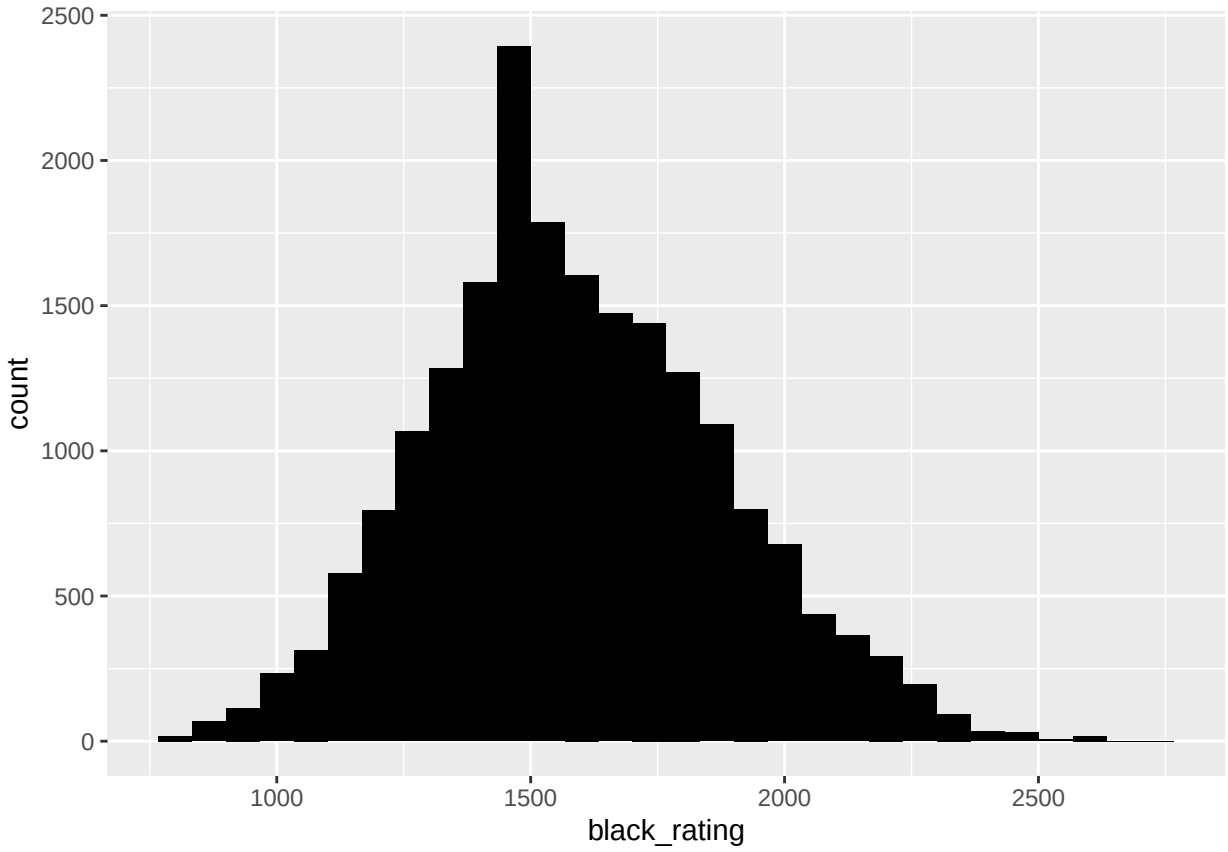
```
ggplot(data = chess)+geom_histogram(aes(x=white_rating),fill = "white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = chess)+geom_histogram(aes(x=black_rating),fill = "black")
```

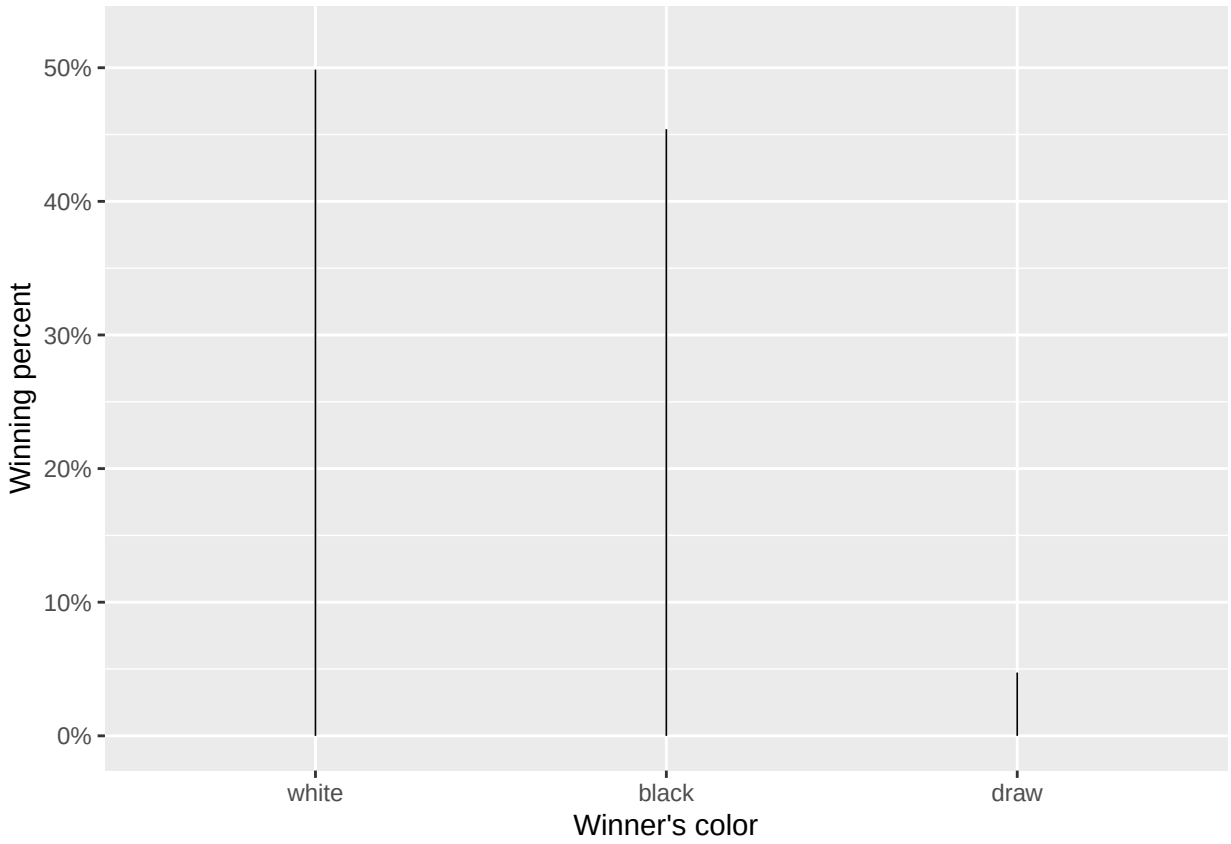
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Step 3

I want to check the winning percentage of white, black and draw. As you see the graph below, the winning percentage of white is 50% and for black is around 45% and there are 5% of draw. The reason why white has the highest winning percentage because white can move first and can choose it's own favorite opening. Therefore, it will has the highest winning percentage.

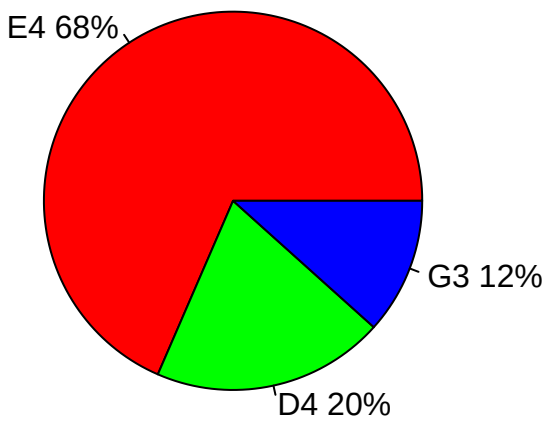
```
chess %>%
  group_by(winner) %>%
  summarise(count = n(), .groups = 'drop') %>%
  mutate(count = count/sum(count)) %>%
  ggplot(., aes(reorder(winner, -count), count))+
  geom_segment(aes(xend = winner, yend = 0), size = 0.3, colour = "black")+
  scale_y_continuous(labels = scales::percent_format(accuracy = 1), limits = c(0,0.52))+
  labs(x = "Winner's color", y = "Winning percent")
```



Step 4

In this step, I want to check what is the top 10 opening moves. 7 of them the first move is E4, which is 68% of the top ten opening move. D4 comes in the second at 20%. And G3 is around 13%. As shown in the pie chart below.

```
slices <- c(2341, 679,398)
lbls <- c("E4", "D4", "G3")
pct <- round(slices/3418*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Opening move")
```



Step 5

I want to find out what is the top 10 opening being played on lichess through database.

```
opening<-filter(summarise(group_by(chess,opening_name), count=length(opening_name)),count>200)
opening
```

```
## # A tibble: 10 x 2
##    opening_name                                  count
##    <chr>                                         <int>
##  1 French Defense: Knight Variation                271
##  2 Horwitz Defense                                 209
##  3 Queen's Pawn Game: Chigorin Variation           229
##  4 Queen's Pawn Game: Mason Attack                 232
##  5 Scandinavian Defense                            223
##  6 Scandinavian Defense: Mieses-Kotroc Variation   259
##  7 Scotch Game                                     271
##  8 Sicilian Defense                                358
##  9 Sicilian Defense: Bowdler Attack                296
## 10 Van't Kruijs Opening                            368
```

Step 6

I calculated how many games are win by each side so that later on I will be able to find the highest winning percentage for both colors.

```
chess%>%
  group_by(winner)
```

```
## # A tibble: 20,058 x 16
## # Groups:   winner [3]
##    id       rated    created_at  last_move_at turns victory_status winner
##    <chr>    <chr>         <dbl>         <dbl> <int> <chr>          <chr>
##  1 TZJHLljE FALSE 1504210000000 1504210000000    13 outoftime      white
##  2 l1NXvwaE TRUE  1504130000000 1504130000000    16 resign         black
##  3 mIICvQHh TRUE  1504130000000 1504130000000    61 mate           white
##  4 kWKvrqYL TRUE  1504110000000 1504110000000    61 mate           white
```

```
##  5 9tXo1AUZ TRUE  1504030000000 1504030000000    95 mate           white
##  6 MsoDV9wj FALSE 1504240000000 1504240000000     5 draw           draw
##  7 qwU9rasv TRUE  1504230000000 1504230000000    33 resign         white
##  8 RVNON3VK FALSE 1503680000000 1503680000000     9 resign         black
##  9 dwF3DJHO TRUE  1503510000000 1503510000000    66 resign         black
## 10 afoMwnLg TRUE  1503440000000 1503440000000   119 mate           white
## # ... with 20,048 more rows, and 9 more variables: increment_code <chr>,
## #   white_id <chr>, white_rating <int>, black_id <chr>, black_rating <int>,
## #   moves <chr>, opening_eco <chr>, opening_name <chr>, opening_ply <int>
```

```
white_game<-0
black_game<-0
draw_game<-0
for(game in chess$winner){
  if(game=="white"){
    white_game<-white_game+1
  }else if(game=="black"){
    black_game<-black_game+1
  }else(draw_game<-draw_game+1)
}
paste("There were",white_game,"white wins and",draw_game,"draws and",black_game,"black wins")
```

```
## [1] "There were 10001 white wins and 950 draws and 9107 black wins"
```

Step 7

I want to find out the top 10 highest opening winning percentage for white color. I divide it by 10001 because in step 6 I found out that there are 10001 games win by white.

```
white<-chess%>%
  filter(winner=='white')%>%
  group_by(opening_name)%>%
  count()%>%
  mutate(percent=n/10001)%>%
  arrange(desc(percent))
white
```

```
## # A tibble: 1,181 x 3
## # Groups:   opening_name [1,181]
##    opening_name                                    n percent
##    <chr>                                       <int>   <dbl>
##  1 Scandinavian Defense: Mieses-Kotroc Variation   164  0.0164
##  2 Sicilian Defense                                149  0.0149
##  3 Scotch Game                                     145  0.0145
##  4 French Defense: Knight Variation                135  0.0135
##  5 Philidor Defense #3                             127  0.0127
##  6 Van't Kruijs Opening                            126  0.0126
##  7 Sicilian Defense: Bowdler Attack                119  0.0119
##  8 Queen's Pawn Game: Mason Attack                 116  0.0116
##  9 Queen's Pawn Game: Chigorin Variation           112  0.0112
## 10 Horwitz Defense                                 110  0.0110
## # ... with 1,171 more rows
```

Step 8

I want to find out the top 10 highest opening winning percentage for black color. I divide it by 9107 because in step 6 I found out that there are 9107 games win by black.

```
black<-chess%>%
  filter(winner=='black')%>%
  group_by(opening_name)%>%
  count()%>%
  mutate(percent=n/9107)%>%
  arrange(desc(percent))
black
```

```
## # A tibble: 1,145 x 3
## # Groups:   opening_name [1,145]
##    opening_name                              n percent
##    <chr>                                 <int>   <dbl>
##  1 Van't Kruijs Opening                    226  0.0248
##  2 Sicilian Defense                        194  0.0213
##  3 Sicilian Defense: Bowdler Attack        164  0.0180
##  4 Scandinavian Defense                    123  0.0135
##  5 French Defense: Knight Variation        121  0.0133
##  6 Scotch Game                             115  0.0126
##  7 Queen's Pawn Game: Chigorin Variation   109  0.0120
##  8 Queen's Pawn Game: Mason Attack         103  0.0113
##  9 Indian Game                             100  0.0110
## 10 Philidor Defense #2                      96  0.0105
## # ... with 1,135 more rows
```

Step 9

I want to find out the top 10 highest opening drawing percentage. I divide it by 950 because in step 6 I found out that there are 950 draw games.

```
draw<-chess%>%
  filter(winner=='draw')%>%
  group_by(opening_name)%>%
  count()%>%
  mutate(percent=n/950)%>%
  arrange(desc(percent))
draw
```

```
## # A tibble: 413 x 3
## # Groups:   opening_name [413]
##    opening_name                             n percent
##    <chr>                                <int>   <dbl>
##  1 Van't Kruijs Opening                    16  0.0168
##  2 French Defense: Knight Variation        15  0.0158
##  3 Sicilian Defense                        15  0.0158
##  4 Queen's Pawn Game: Mason Attack         13  0.0137
##  5 Sicilian Defense: Bowdler Attack        13  0.0137
##  6 Indian Game                             12  0.0126
##  7 Italian Game                            12  0.0126
##  8 Scotch Game                             11  0.0116
##  9 Italian Game: Anti-Fried Liver Defense  10  0.0105
## 10 Scandinavian Defense                    10  0.0105
```

```
## # ... with 403 more rows
```

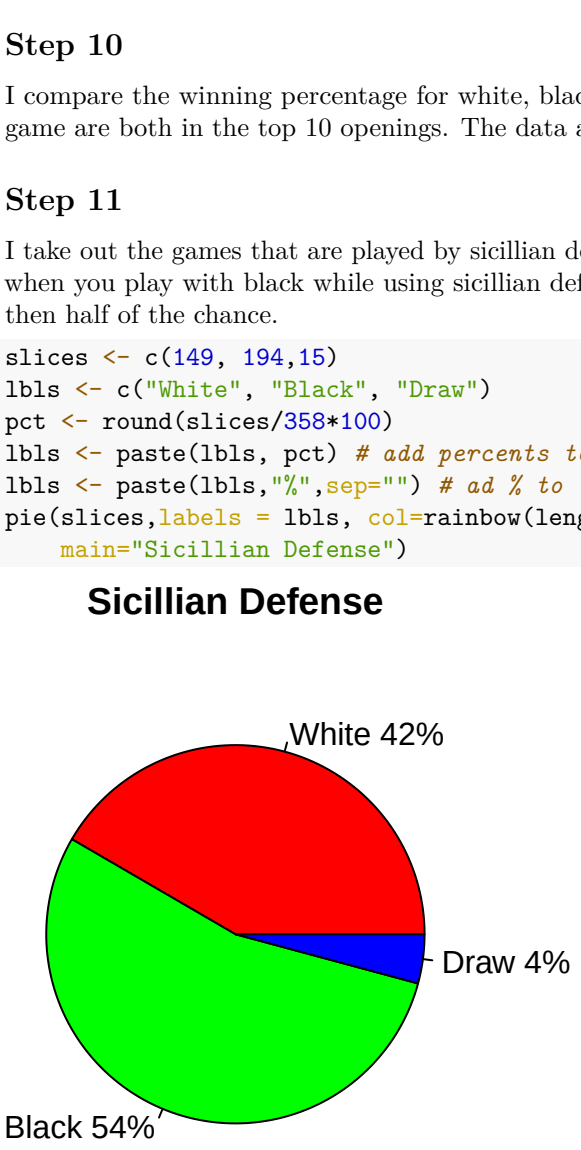
Step 10

I compare the winning percentage for white, black, and draw. I found out that Scotch game and Sicillian game are both in the top 10 openings. The data are all in step 7 through 9.

Step 11

I take out the games that are played by sicillian defense and make this pie chart. As shown in this pie chart, when you play with black while using sicillian defense you have a 54% of winning the game, which is more then half of the chance.

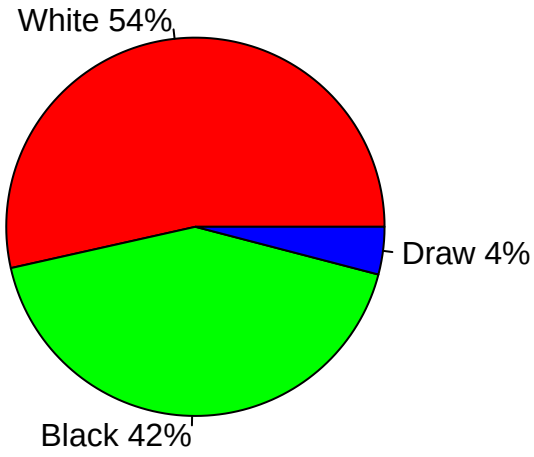
```
slices <- c(149, 194,15)
lbls <- c("White", "Black", "Draw")
pct <- round(slices/358*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Sicillian Defense")
```



Step 12

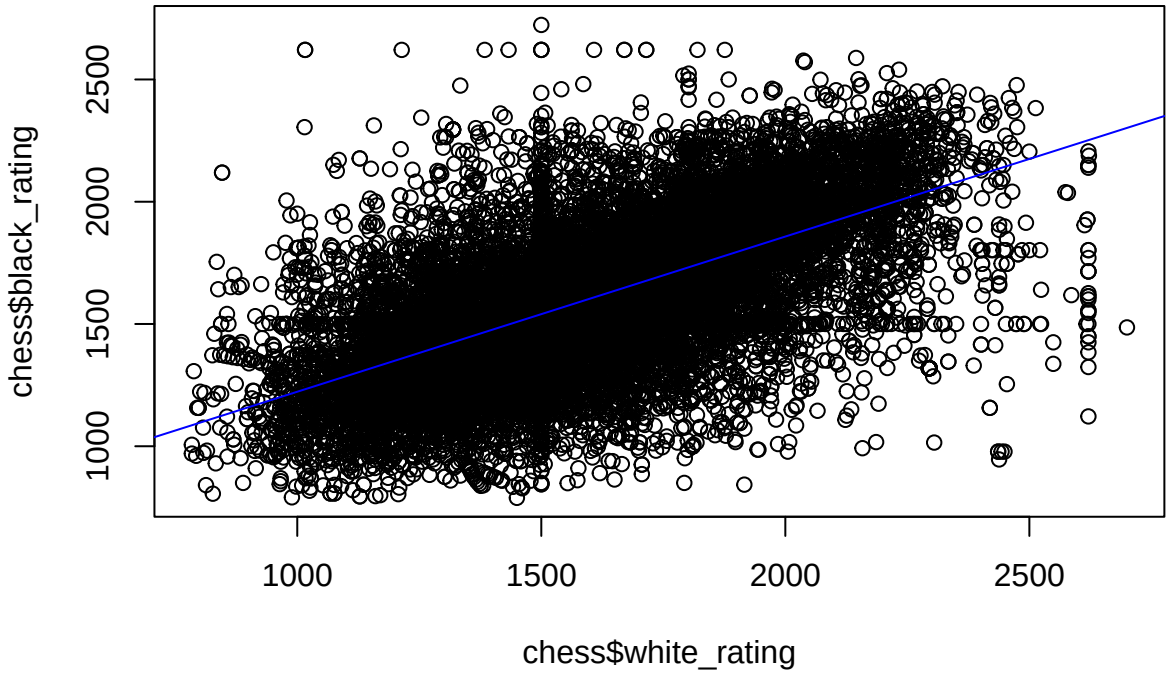
I take out the games that are played by scotch game and make this pie chart. As shown in this pie chart, when you play with white while using scotch game you have a 54% of winning the game, which is more then half of the chance.

```
slices <- c(145, 115,11)
lbls <- c("White", "Black", "Draw")
pct <- round(slices/271*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Scotch Game")
```



Linear Regression Comparing white rating to black rating.

```
plot(chess$white_rating,chess$black_rating)
chess.regression <-lm(white_rating~black_rating,data = chess)
abline(chess.regression, col ="blue")
```



##

Challenges The main challenges that I faced are that the data I tried to analyze are all mainly words, so I need to keep on using filter in order to compare the data. It is hard to made the linear regression out of the data that I have chose. But I managed to make it at last. Because I couldn't make linear regression in the beginning, I decided to make pie and bar chart.

Why I didn't choose Van't kruijs Opening

The reason why I didn't choose Van't Kruijs Opening because this opening is start from E3, which is not even in the top opening first moves despite being one of the top opening. The reason why it is one of the top opening is because there are not that many opening that start with E3 move. So if e3 is played, there is a

very high chance that it will be Van't Krujs Opening.

Conclusion

Base on this dataset and what GM Hikaru has said. I will play white using scotch game and black using sicillian defense. Even though I never played these two openings before, I am going to start learning these openings. If I has taken this class earler, I might have a higher rating if I know that these two are the top openings for white and black. From now on when I play chess, I will focus on scotch game and sicillian defense.