ISQA 521 Final Project

Skye Gilbreth, Jordan Hilton

May 2, 2019

Introduction

We're beginner chess players, and we'd like to improve our game a little. We have a data set of 20,000 chess games played on the website lichess, including a computer-identified opening used in each game. The goal of this project is to analyze this set of games to find out more about what openings Iwe should study, by looking at the performance of each opening using this data.

The set is from this kaggle:

https://www.kaggle.com/datasnaek/chess

Data Load

Let's load the data and take a look at it. We've dropped out some unnecessary columns, including the largest one which is a list of moves for each game.

```
chessdata <- read.csv("data/chessdata.csv")
chessdata <- chessdata[,-c(1,3,4,13,14,16)]
head(chessdata)
```

```
##   rated turns victory_status winner increment_code      white_id
## 1 FALSE    13      outoftime  white           15+2      bourgris
## 2  TRUE    16         resign  black           5+10          a-00
## 3  TRUE    61           mate  white           5+10        ischia
## 4  TRUE    61           mate  white           20+0 daniamurashov
## 5  TRUE    95           mate  white           30+3     nik221107
## 6 FALSE     5           draw   draw           10+0     trelynn17
##   white_rating      black_id black_rating
## 1         1500          a-00         1191
## 2         1322     skinnerua         1261
## 3         1496          a-00         1500
## 4         1439  adivanov2009         1454
## 5         1523  adivanov2009         1469
## 6         1250 franklin14532         1002
##                            opening_name
## 1       Slav Defense: Exchange Variation
## 2 Nimzowitsch Defense: Kennedy Variation
## 3  King's Pawn Game: Leonardis Variation
## 4 Queen's Pawn Game: Zukertort Variation
## 5                      Philidor Defense
## 6   Sicilian Defense: Mongoose Variation
```

Let's go over our available fields quickly. We have:

- rated, whether or not the game was rated
- turns, how many turns were played
- victory_status, how the game ended (by timeout, resignation, draw, or mate)
- winner, the outcome of the game (white wins, black wins, or a draw)

- increment_code, a description of the time control for the game (15+3 means that each player started with 15 minutes on the clock and had a 3 second increment for each move)
- white_id and black_id, the ids of the players
- white_rating and black_rating, the ELO ratings of the players for the category of game they're playing
- opening_name, the computer's description of what opening was played

We're now going to create two additional fields by mutation; one, using regular expressions, consolidates categories of openings (so "Queen's Pawn Game: Han Solo Variant | The Coolest Line" will be grouped in with all the other Queen's Pawn games), and one calculates an average of the two players' ratings. We're also going to use factors to categorize the games by how much time was on the clock, into the categories "bullet", "blitz", and "classical".

```
## fix the names of the openings
chessdata<-mutate(chessdata, shortname=gsub(":.*","",chessdata$opening_name))
chessdata<-transform(chessdata, shortname=gsub("\\|.*","",chessdata$shortname))
## create an average rating for the game
chessdata<-mutate(chessdata, avgrating=(white_rating+black_rating)/2)
## classify the game length
chessdata<-mutate(chessdata, starttime=gsub("\\+.*","",chessdata$increment_code))
chessdata$gametype<-factor(chessdata$starttime)
levels(chessdata$gametype)<-list(Bullet=c("0","1","2"), Blitz=c("3","4","5","6","7","8","9"), Classical=
head(chessdata)
```

```
##   rated turns victory_status winner increment_code      white_id
## 1 FALSE    13      outoftime  white           15+2      bourgris
## 2  TRUE    16         resign  black           5+10          a-00
## 3  TRUE    61           mate  white           5+10        ischia
## 4  TRUE    61           mate  white           20+0 daniamurashov
## 5  TRUE    95           mate  white           30+3     nik221107
## 6 FALSE     5           draw   draw           10+0     trelynn17
##   white_rating      black_id black_rating
## 1         1500          a-00         1191
## 2         1322     skinnerua         1261
## 3         1496          a-00         1500
## 4         1439  adivanov2009         1454
## 5         1523  adivanov2009         1469
## 6         1250 franklin14532         1002
##                            opening_name           shortname avgrating
## 1       Slav Defense: Exchange Variation        Slav Defense    1345.5
## 2 Nimzowitsch Defense: Kennedy Variation Nimzowitsch Defense    1291.5
## 3  King's Pawn Game: Leonardis Variation    King's Pawn Game    1498.0
## 4 Queen's Pawn Game: Zukertort Variation   Queen's Pawn Game    1446.5
## 5                      Philidor Defense    Philidor Defense    1496.0
## 6   Sicilian Defense: Mongoose Variation    Sicilian Defense    1126.0
##   starttime  gametype
## 1        15 Classical
## 2         5     Blitz
## 3         5     Blitz
## 4        20 Classical
## 5        30 Classical
## 6        10 Classical
```

Who's the highest rated player we have?

```
head(sort(chessdata$white_rating, decreasing=TRUE), 1)
```

```
## [1] 2700
```

```
filter(chessdata, white_rating==2700)
```

```
##   rated turns victory_status winner increment_code   white_id white_rating
## 1 False    20         resign  white          30+30 justicebot         2700
##        black_id black_rating                          opening_name
## 1 youredeadmeat         1486 Caro-Kann Defense: Classical Variation
##           shortname avgrating starttime  gametype
## 1 Caro-Kann Defense      2093        30 Classical
```

The answer is "justicebot", long may he reign, playing the Caro-Kann.

Exploratory charts

Let's take a look at what we have in our new "shortnames" field in terms of number of games for each opening:

```
length(table(chessdata$shortname))
```

```
## [1] 180
```

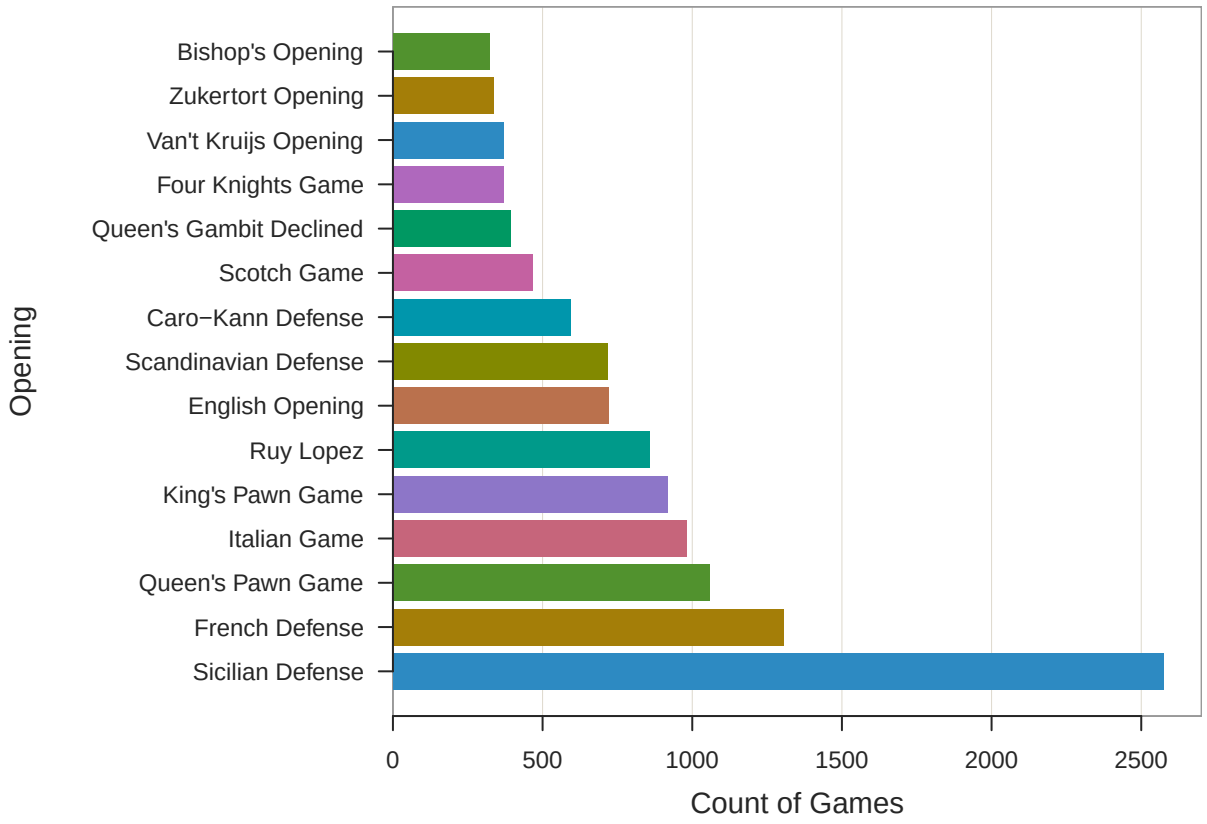
180 is too many openings to neatly chart! Let's filter our table down to the top 15 openings and then take a look at some bar charts.

```
top15openings<-head(names(sort(table(chessdata$shortname), decreasing=TRUE)),15)
filtereddata<-filter(chessdata, shortname %in% top15openings)
```

```
#Export altered data to a CSV file to be used by the Shiny app.
write.csv(filtereddata, "data/filtereddata.csv")
```

Now let's start with a basic bar chart just to see the relative frequency of these top 15 openings.

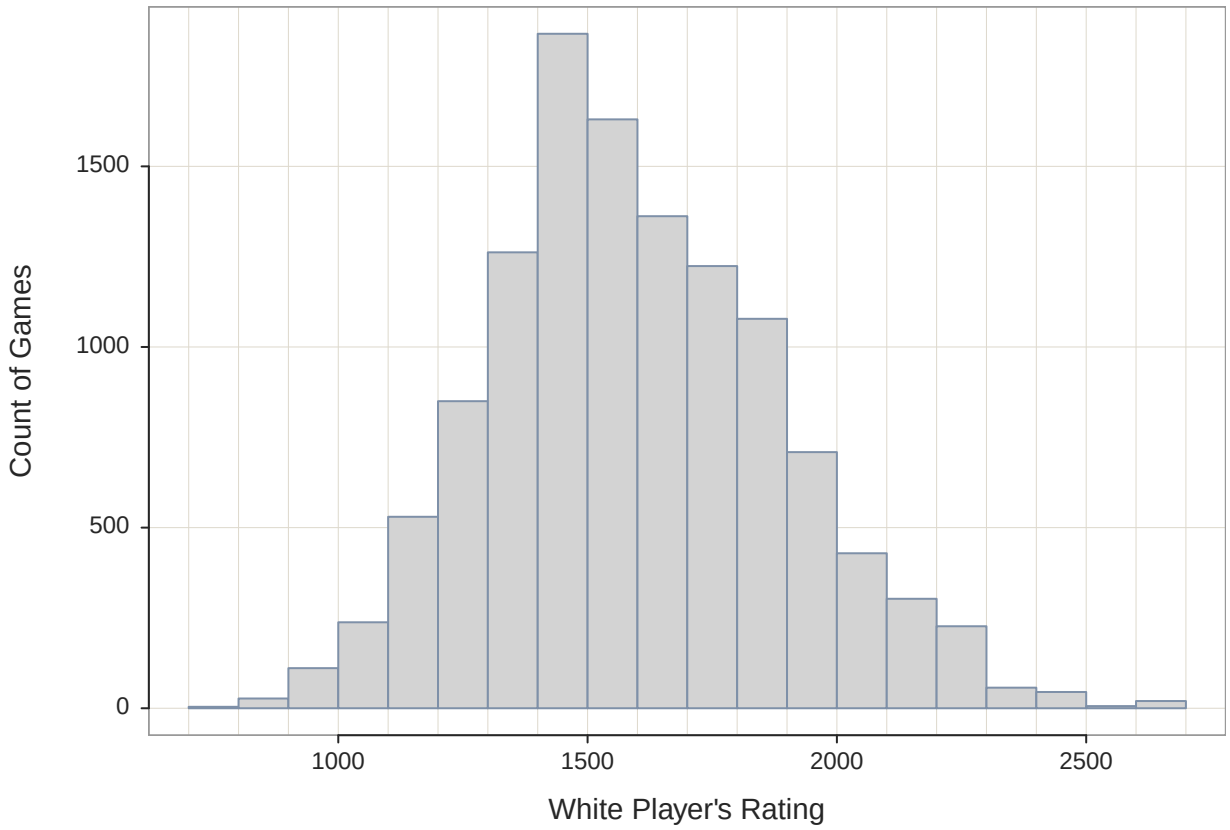
```
BarChart(shortname, data=filtereddata, horiz=TRUE, quiet=TRUE, xlab="Opening", sort="-", ylab="Count of
```



The Sicilian Defense (1. e4 c6) is by far the most popular, with an even distribution after that. We're keeping an eye on the Italian Game (1. e4 e5 2. Nf3 Nc6 3. Bc4) since that's Jordan's favorite opening, and we're pleased to see it in 4th place.

As an excuse to switch to histograms, how about the ratings of the players in this data set?

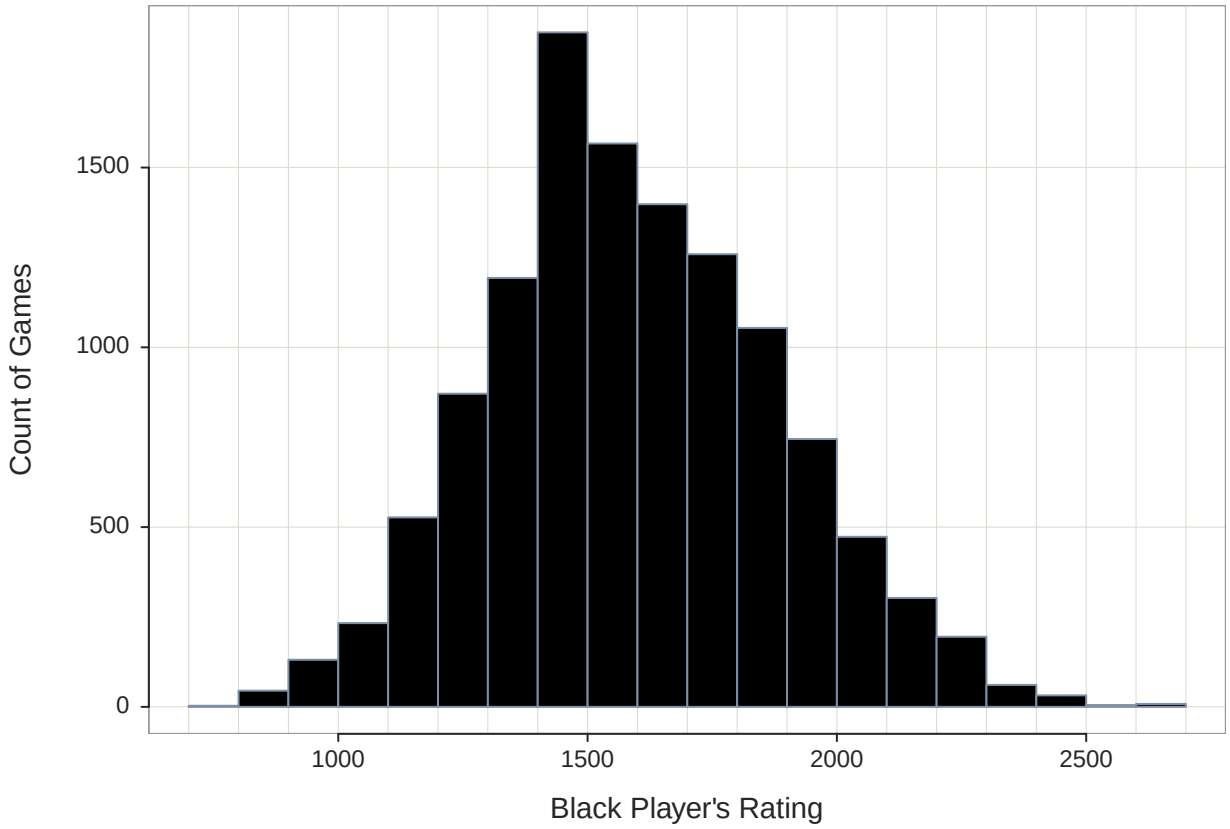
```
Histogram(white_rating, data=filtereddata, fill="lightgray", quiet=TRUE, xlab="White Player's Rating", y
```



This looks like a normal distribution with a slight skew to the left. The default rating is 1500, so there are probably a lot of games by new players at or immediately under 1500. Note the slight uptick at the very top of the range - a buildup of high-rated players.

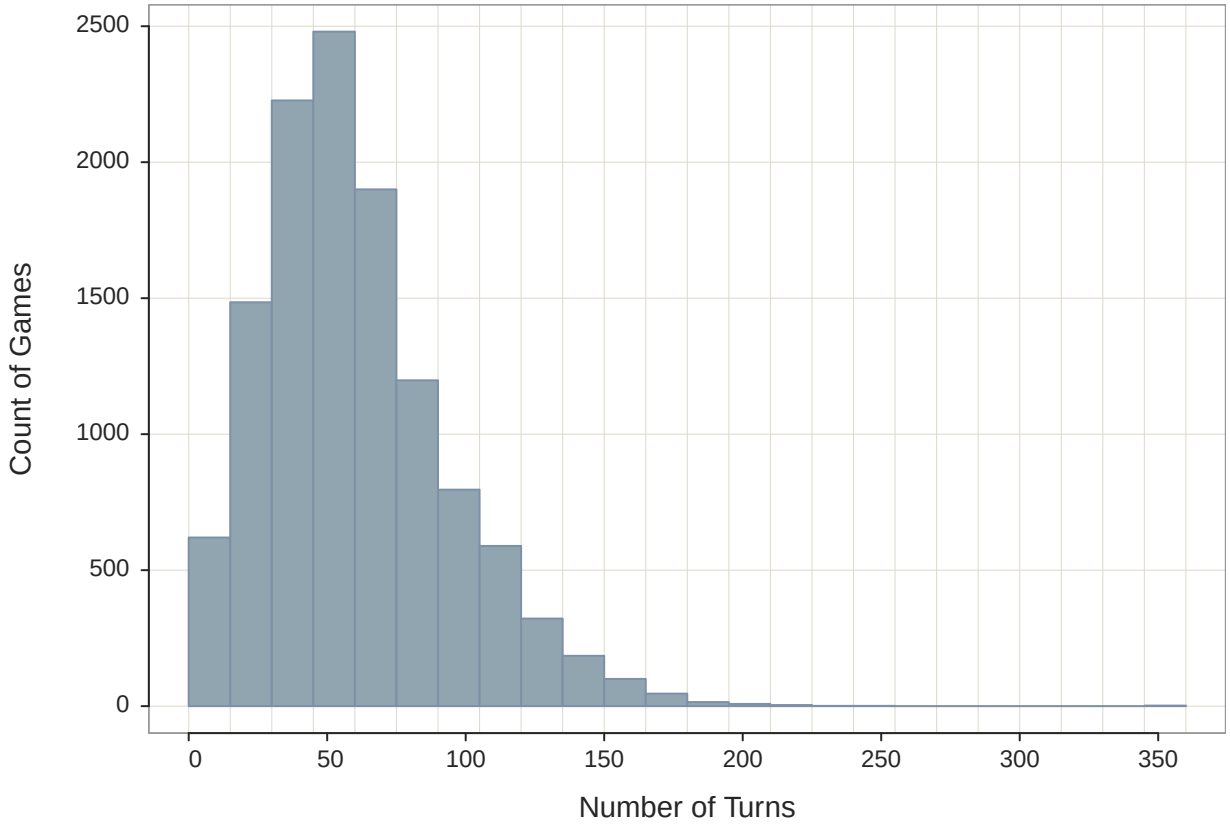
Are the ratings of the black players any different?

```
Histogram(black_rating, data=filtereddata, fill="black", quiet=TRUE, xlab="Black Player's Rating", ylab=
```


Count of Games
1500
1000
500
0
1000
1500
2000
2500
Black Player's Rating

No, the distribution of ratings for the black players are identical; that's as expected. How long is a normal game?

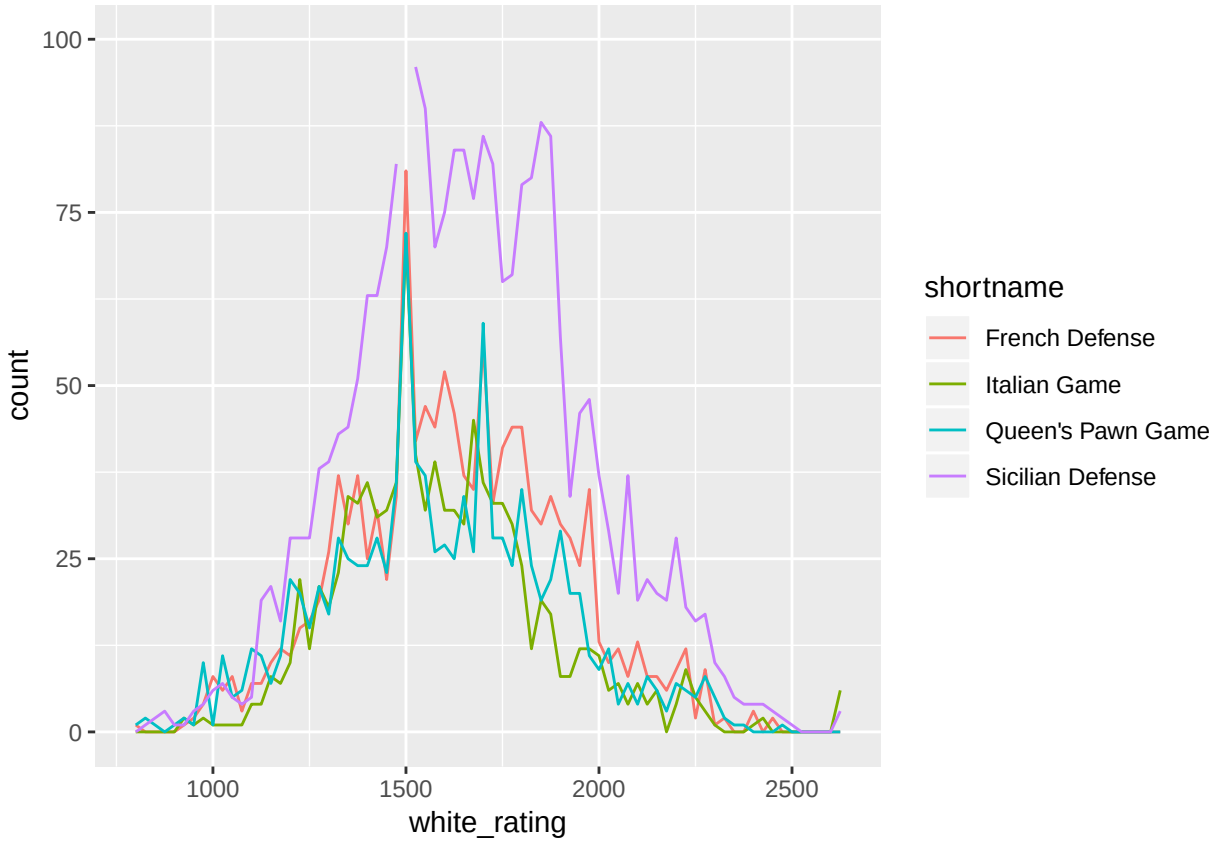
```
Histogram(turns, data=filtereddata, bin.width=15, quiet=TRUE, xlab="Number of Turns", ylab="Count of Ga
```



It looks like a normal game takes around 50 turns, but there's a long tail of games that go all the way out to 350 moves- those are probably complicated endgames with little advantage where somebody is refusing to resign or accept a draw.

Let's take a look at the 4 most frequent openings (couldn't leave out the Italian!). Do any of them get noticeably more frequent in games by higher-rated players, which might indicate that they're just a better opening?

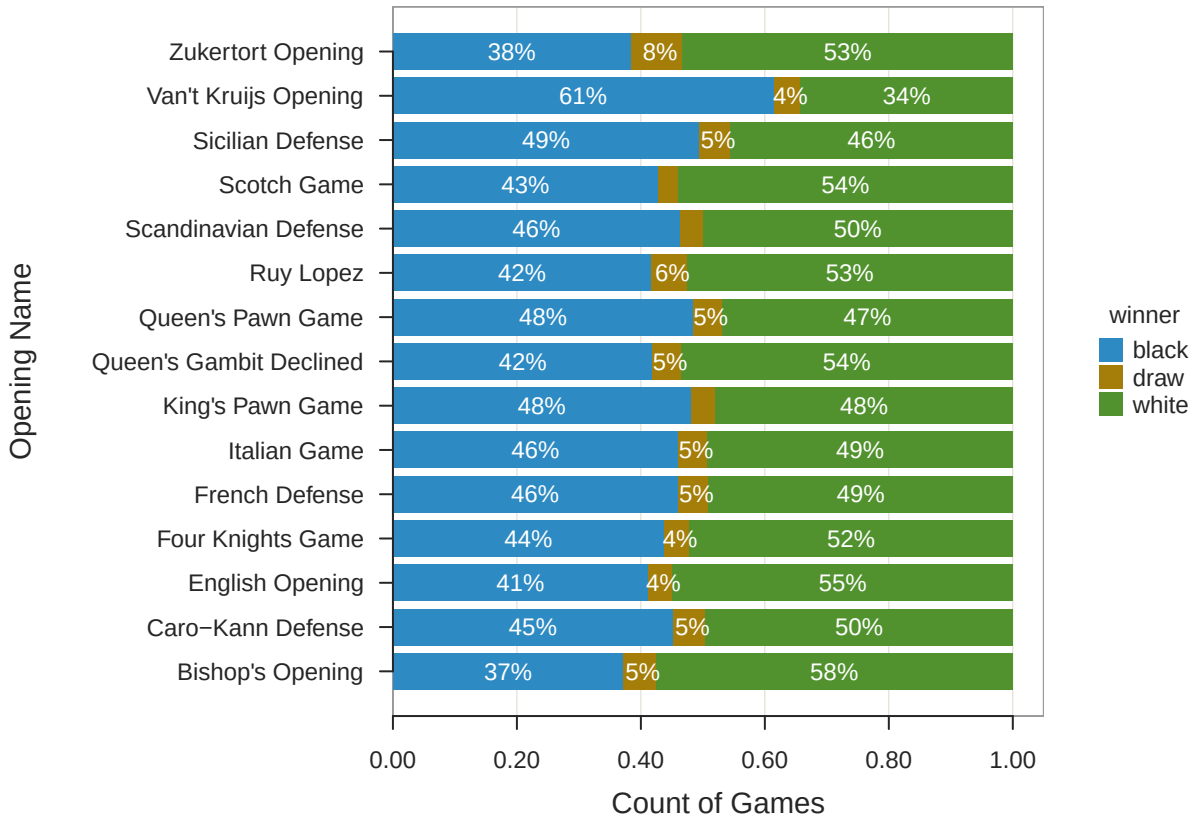
```
top4<-head(names(sort(table(chessdata$shortname), decreasing=TRUE)),4)
filtereddata2<-filter(chessdata, shortname %in% top4)
ggplot(filtereddata2, aes(x=white_rating, color=shortname))+geom_line(stat="bin", binwidth=25)+ylim(0,1
```



The Sicilian Defense is roughly as frequent as other openings at low ratings but quickly gains and retains dominance as rating increases. This probably indicates that it's worth study.

How about the relative win rates of each opening?

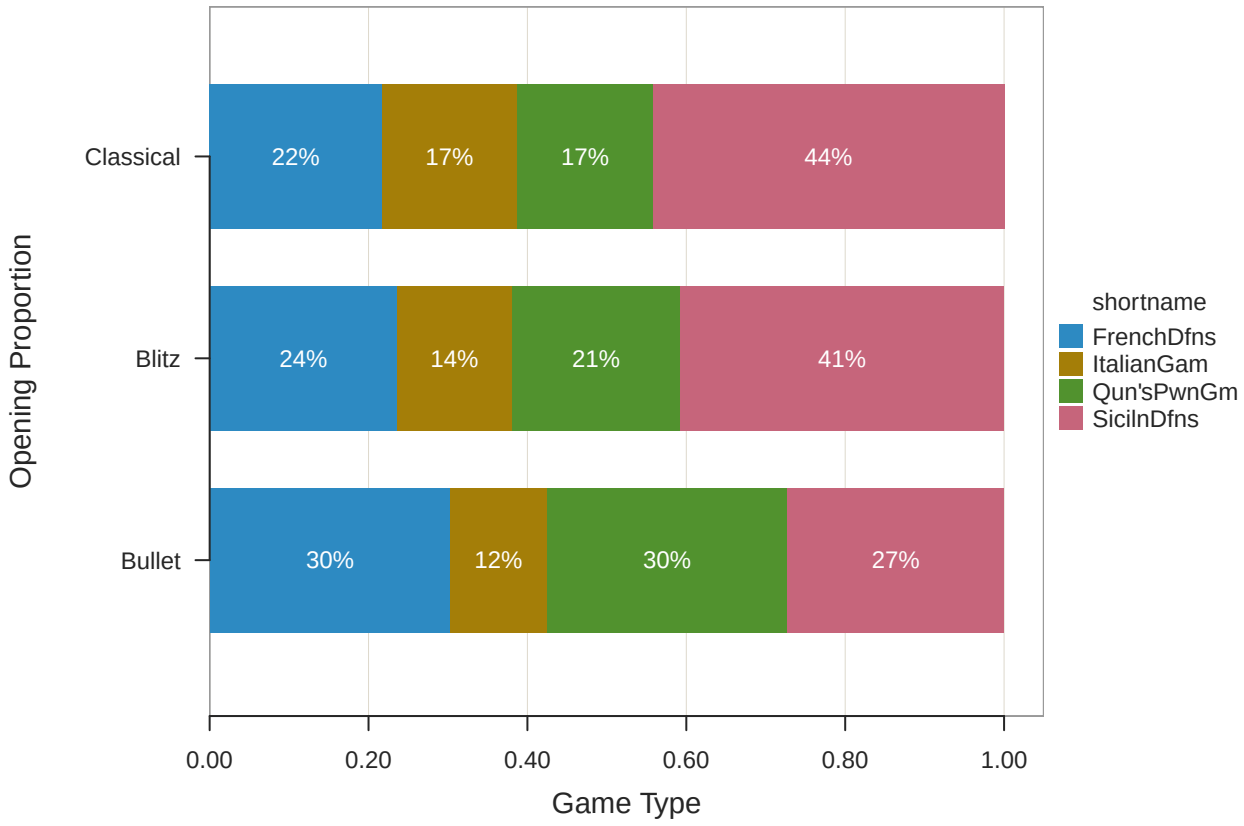
```
BarChart(shortname, by=winner, data=filtereddata, horiz=TRUE, ylab="Count of Games", stat.x="proportion
```



It looks like the Sicilian is very good for black, but the Bishop's opening and the English are good for white.

Lastly, how often do the top 4 openings get played in each time control?

```
BarChart(gametype, by=shortname, data=filtereddata2, horiz=TRUE, ylab="Game Type", stat.x="proportion",
```

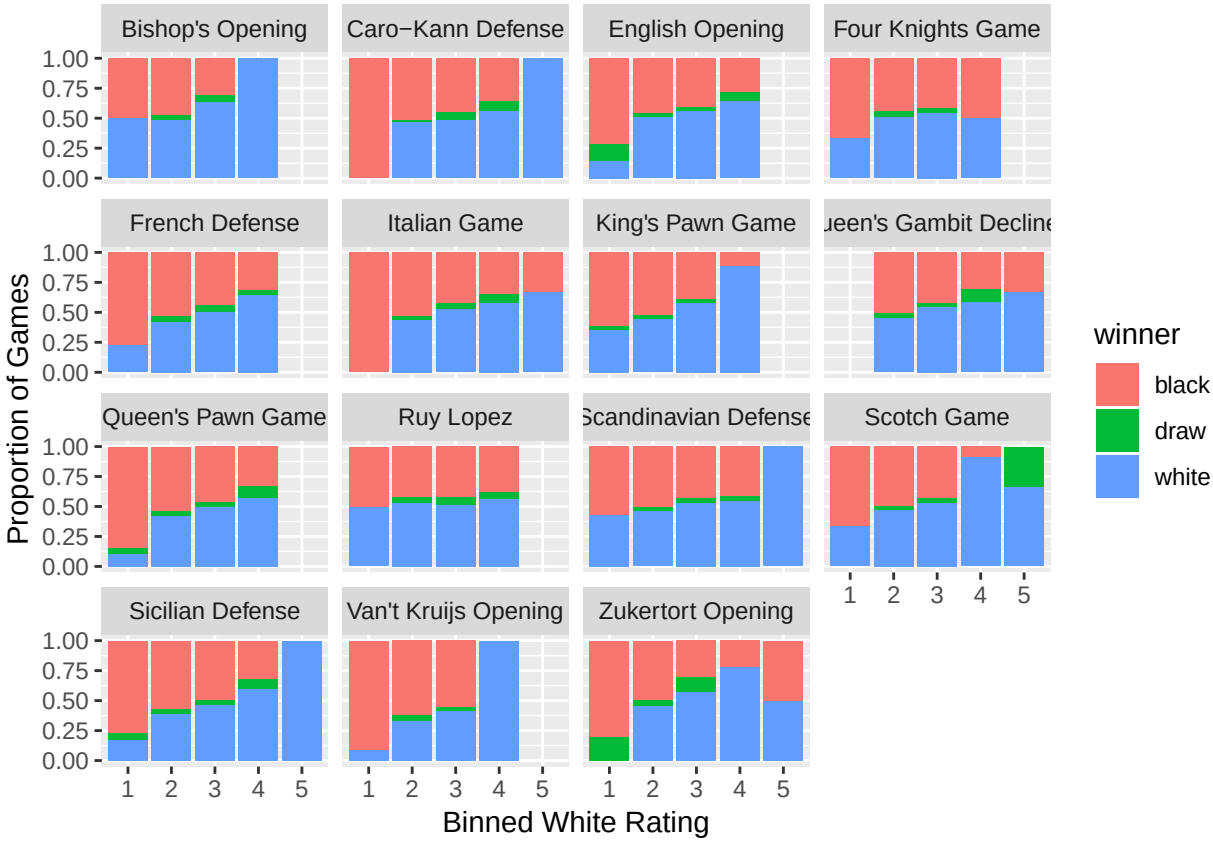


We can see that the French and the Queen's Pawn Game are more popular in short time controls and the Sicilian more popular in longer time controls.

Best Opening for Us

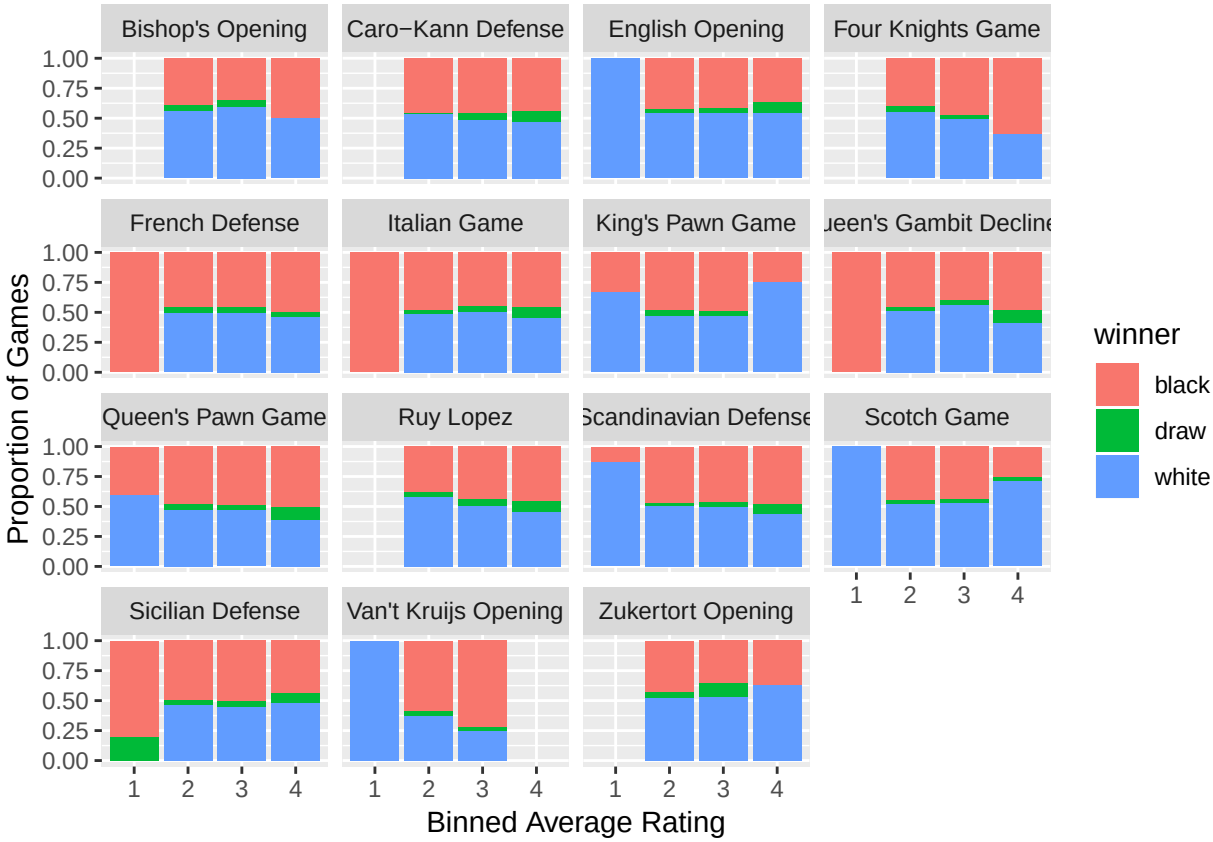
Let's do a trellis plot and look at how each opening's win rate changes by rating.

```
filtereddata$ratingbin<-cut(filtereddata$white_rating, seq(0,3500,500), labels=c("0","1","2","3","4","5
##BarChart(ratingbin, by=winner, by1=shortname, data=filtereddata, horiz=TRUE, ylab="Count of Games", s
ggplot(filtereddata, aes(x = ratingbin, fill=winner)) +
        geom_bar(aes(y = (..count..)/sum(..count..)),position="fill") +
        scale_y_continuous()+facet_wrap(vars(shortname)) +labs(x="Binned White Rating", y="Proportion o
```



Note that each opening seems to improve with rating- this is a methodological error because we've used the white player's rating while ignoring the black player's rating, so with a higher rating just comes a higher win rate! Let's switch to the average rating instead:

```
filtereddata$ratingbin<-cut(filtereddata$avgrating, seq(0,3500,500), labels=c("0","1","2","3","4","5","
##BarChart(ratingbin, by=winner, by1=shortname, data=filtereddata, horiz=TRUE, ylab="Count of Games", s
ggplot(filtereddata, aes(x = ratingbin, fill=winner)) +
        geom_bar(aes(y = (..count..)/sum(..count..)),position="fill") +
        scale_y_continuous()+facet_wrap(vars(shortname)) +labs(x="Binned Average Rating", y="Proportion
```



Here we get to some interesting results. Picking some in particular: the Scotch game and the King's Pawn games are famously sharp, so it doesn't surprise us that they perform well at high and low average rating but poorly in middling ratings. The Four Knights is supposed to be bad, so it makes sense that its performance gets worse at higher ratings. Remember that when looking at overall winrates the Bishop's Opening was supposed to be good for white- since it gets worse at high ratings, we probably shouldn't work with that one.

Shiny Interactive

We did a thing! We built a very cool app with Shiny Interactive, where you can play with the colors and bin width of not one, but two graphs! Not only can you play with these charts, but you can also see the retroactive ggplot2 code developed for your chosen customizations! There are other neat surprises you will find through our app, but we will not spoil the surprise for you. So head on over to our app, by click on the following URL, and have fun!

https://skygil.shinyapps.io/Game_of_Chess_Final/

Conclusion

We should definitely study the Sicilian, since it is by far the most common opening and it performs well at high ratings. In the meantime, though, sticking with the Italian looks like it should work out just fine; it performs decently at all rating levels and doesn't get too drawish at the high rating levels.