ISQA 521 Final Project

Jordan Hilton

May 2, 2019

Introduction

I'm a beginner chess player, and I'd like to improve my game a little. I 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 I 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. I've dropped out some unnecessary columns, including the largest one which is a list of moves for each game.

```r
chessdata <- read.csv("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

```
chessdata<-mutate(chessdata, shortname=gsub(":.*","",chessdata$opening_name))
chessdata<-transform(chessdata, shortname=gsub("\\|.*","",chessdata$shortname))
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
## 1       Slav Defense: Exchange Variation        Slav Defense
## 2 Nimzowitsch Defense: Kennedy Variation Nimzowitsch Defense
## 3  King's Pawn Game: Leonardis Variation    King's Pawn Game
## 4 Queen's Pawn Game: Zukertort Variation   Queen's Pawn Game
## 5                       Philidor Defense    Philidor Defense
## 6   Sicilian Defense: Mongoose Variation    Sicilian Defense
```

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))
```

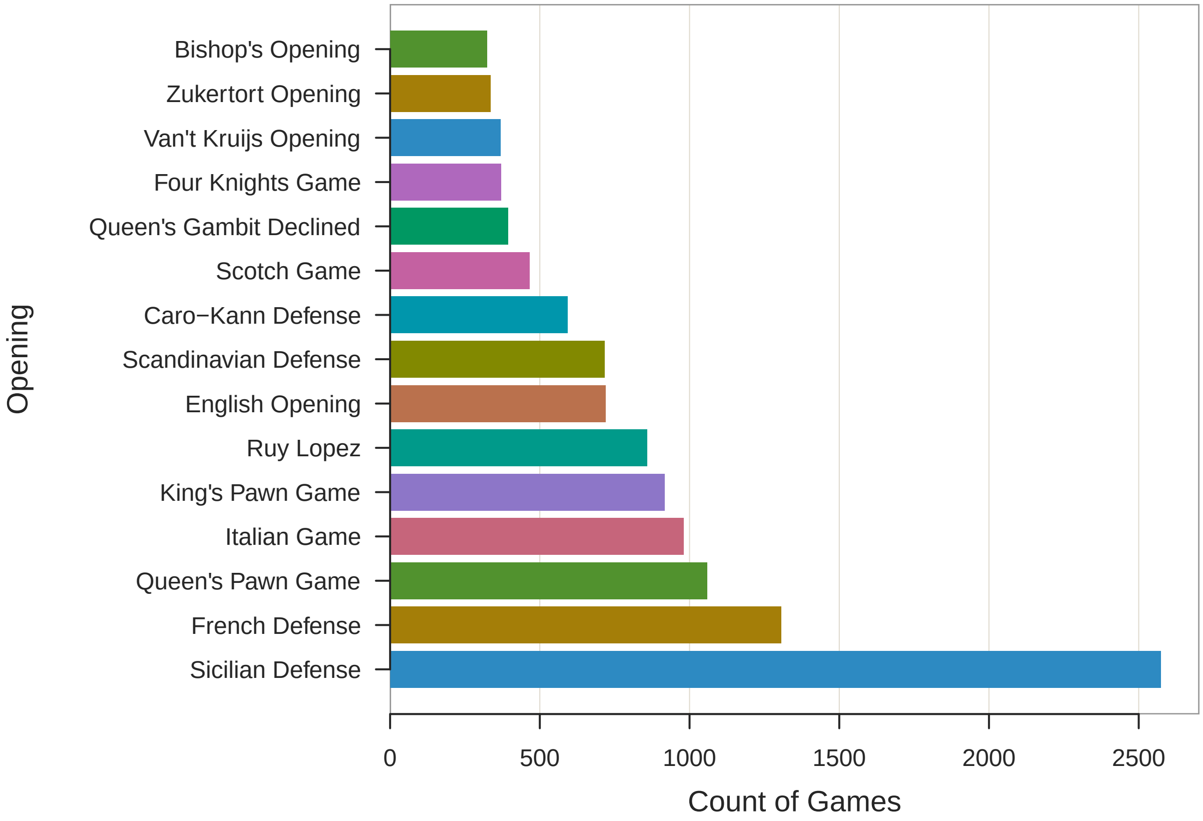
```
## [1] 180
```

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)
```

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

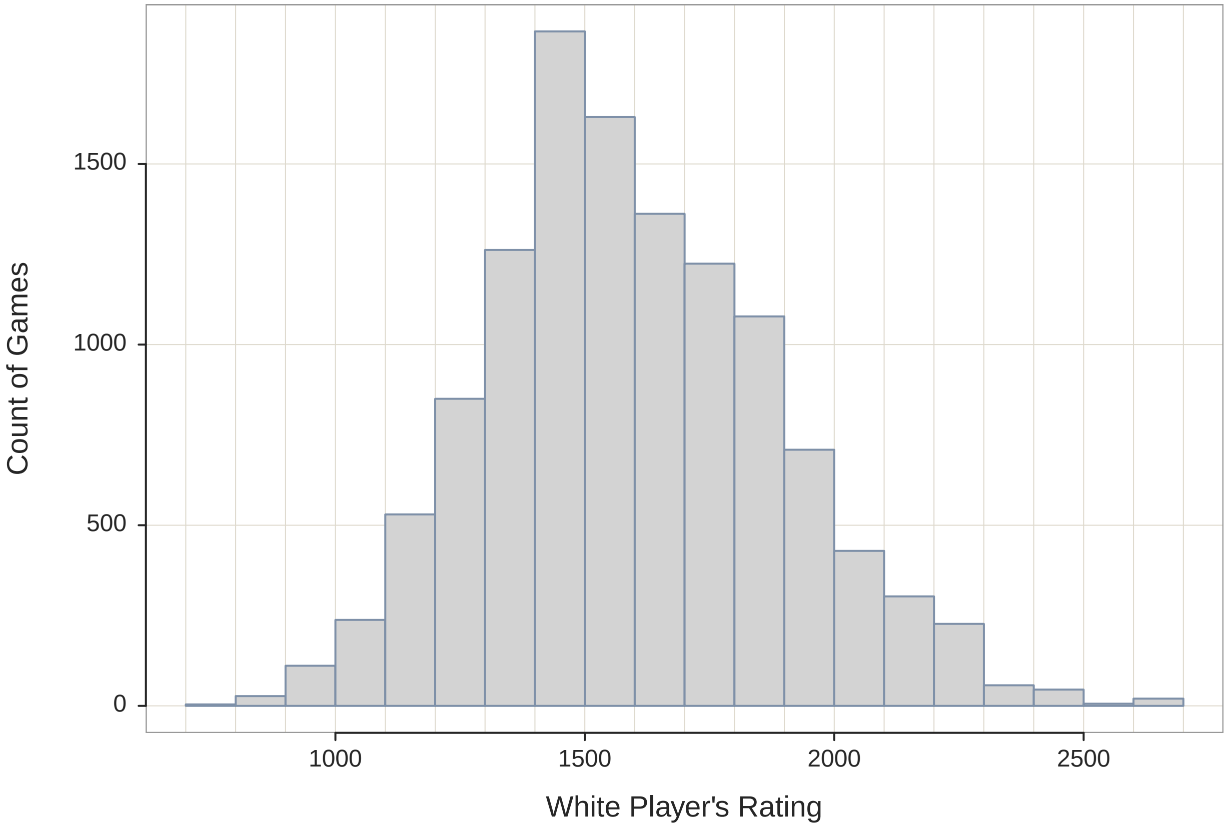
```
BarChart(shortname, data=filtereddata, horiz=TRUE, quiet=TRUE, xlab="Opening", sort="-", ylab="Count of
```



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 my favorite opening, and I'm 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?

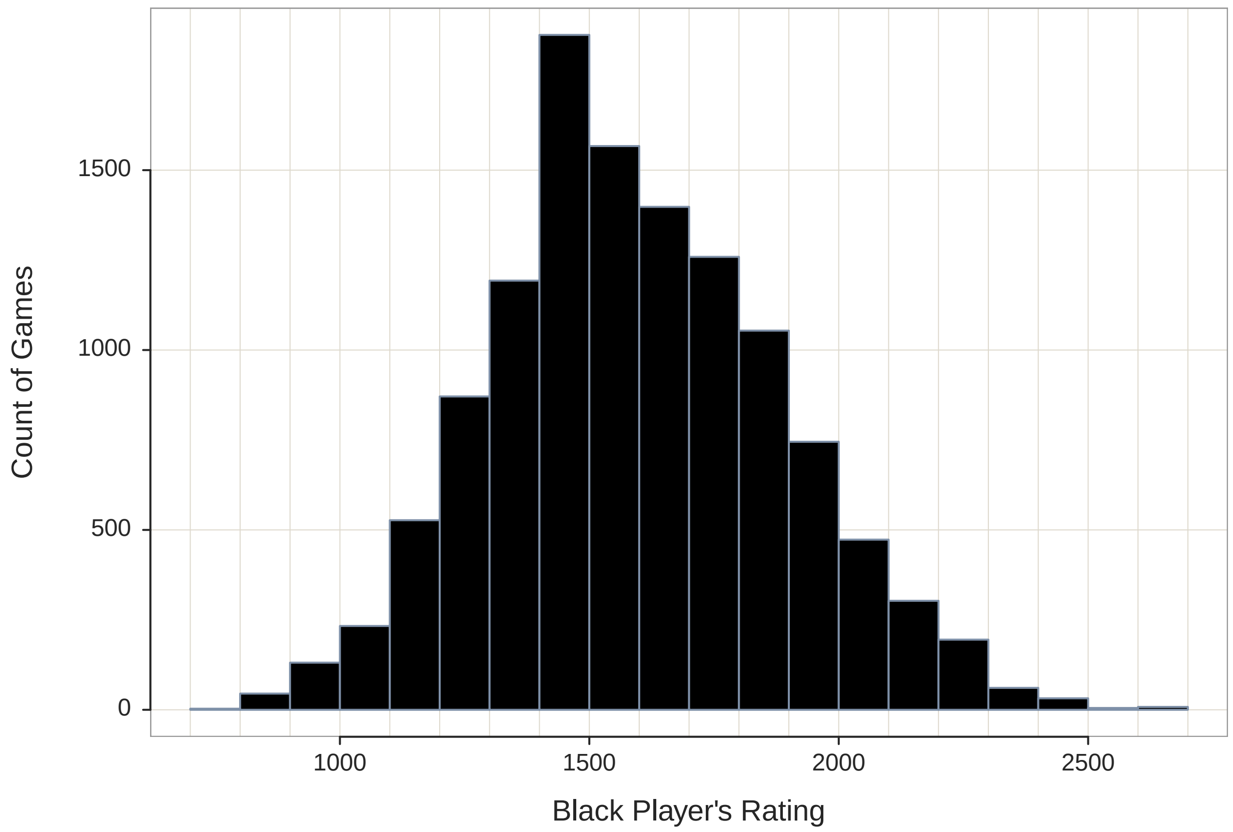
```
Histogram(white_rating, data=filtereddata, fill="lightgray", quiet=TRUE, xlab="White Player's Rating", 
```



This looks like a normal distribution with a slight skew to the left. I'd explain this by saying your 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.

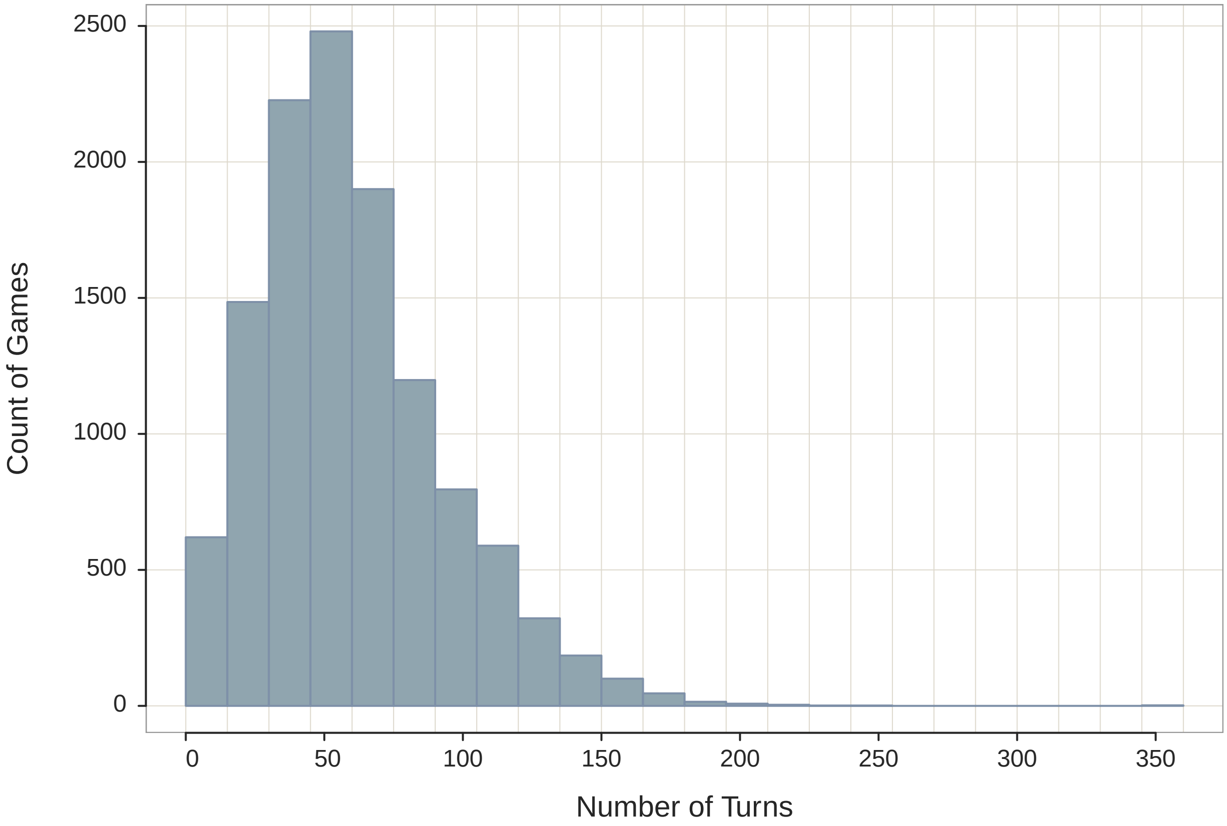
Are the ratings of the black players any different?

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

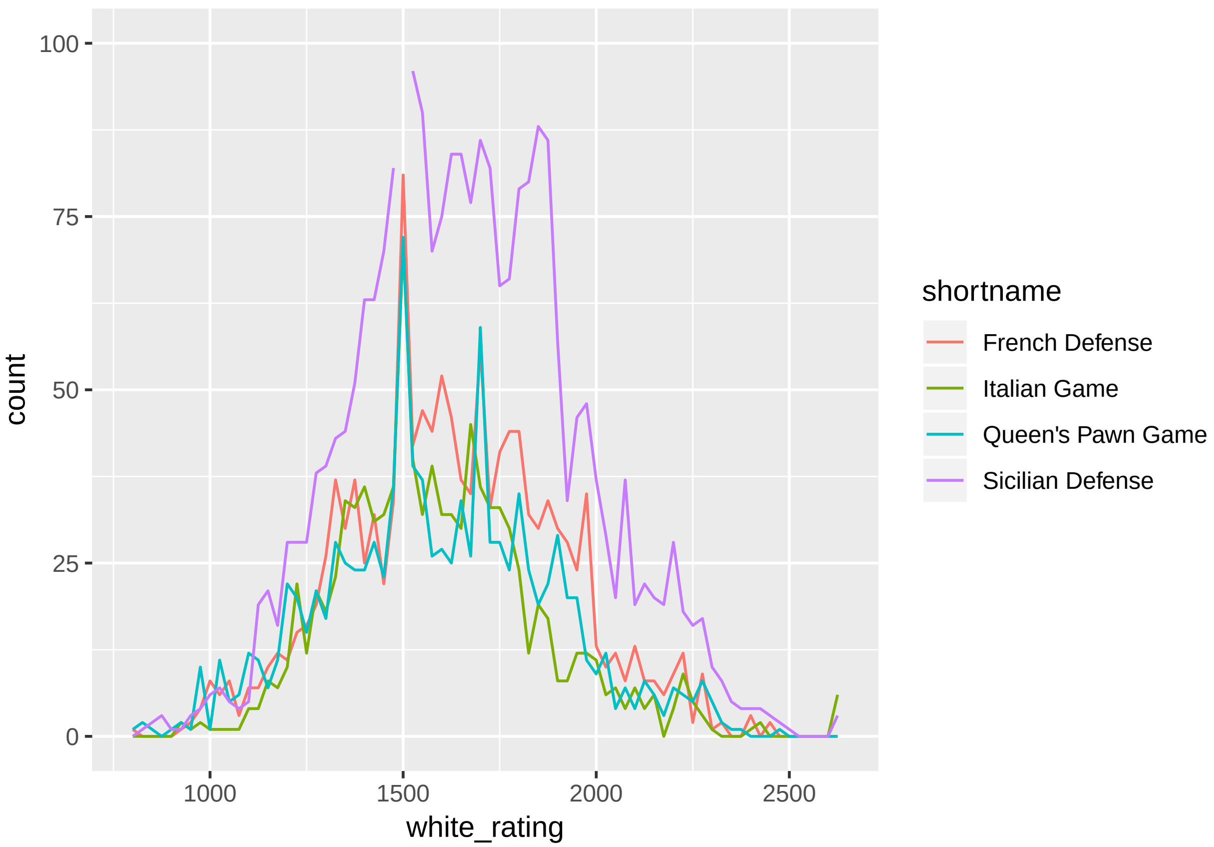


No, the distribution of ratings for the black players are identical; that's as expected. How long is a normal game?

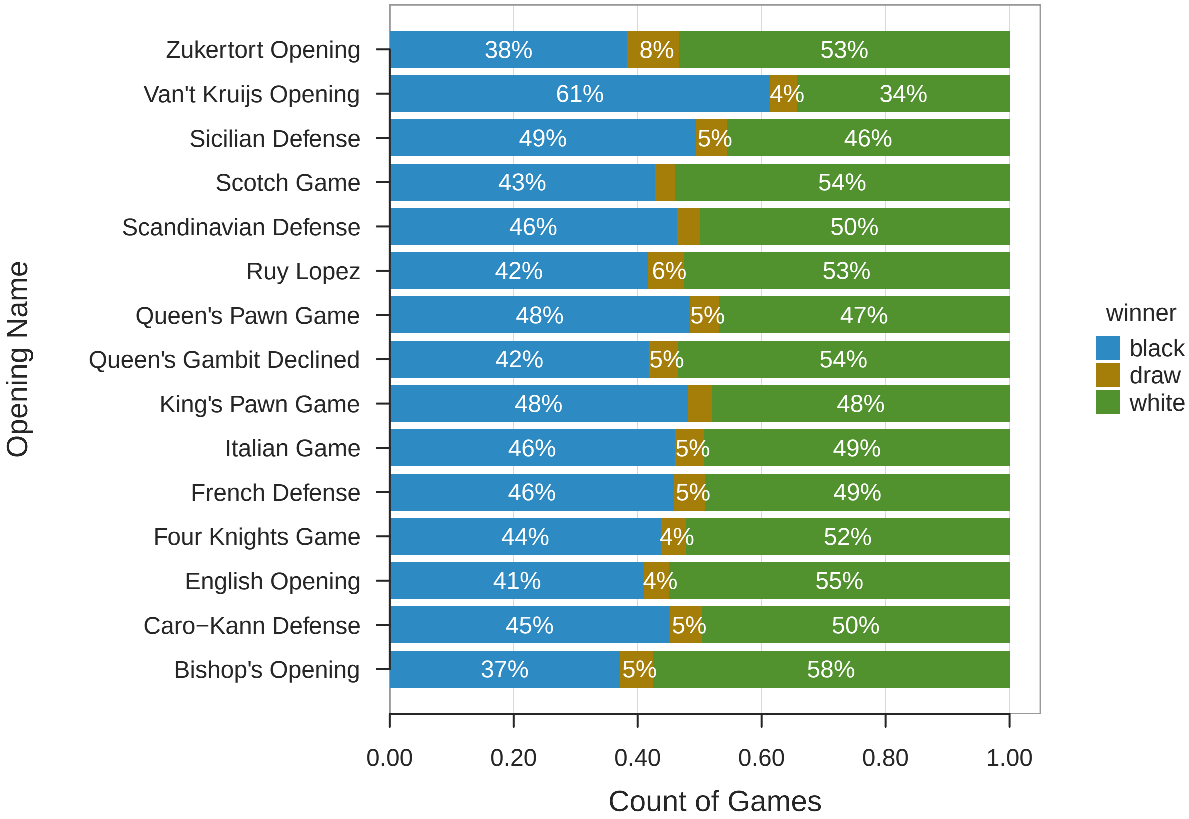
```
Histogram(turns, data=filtereddata, bin.width=15, quiet=TRUE, xlab="Number of Turns", ylab="Count of Gar
```



```r
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(0,10
```



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



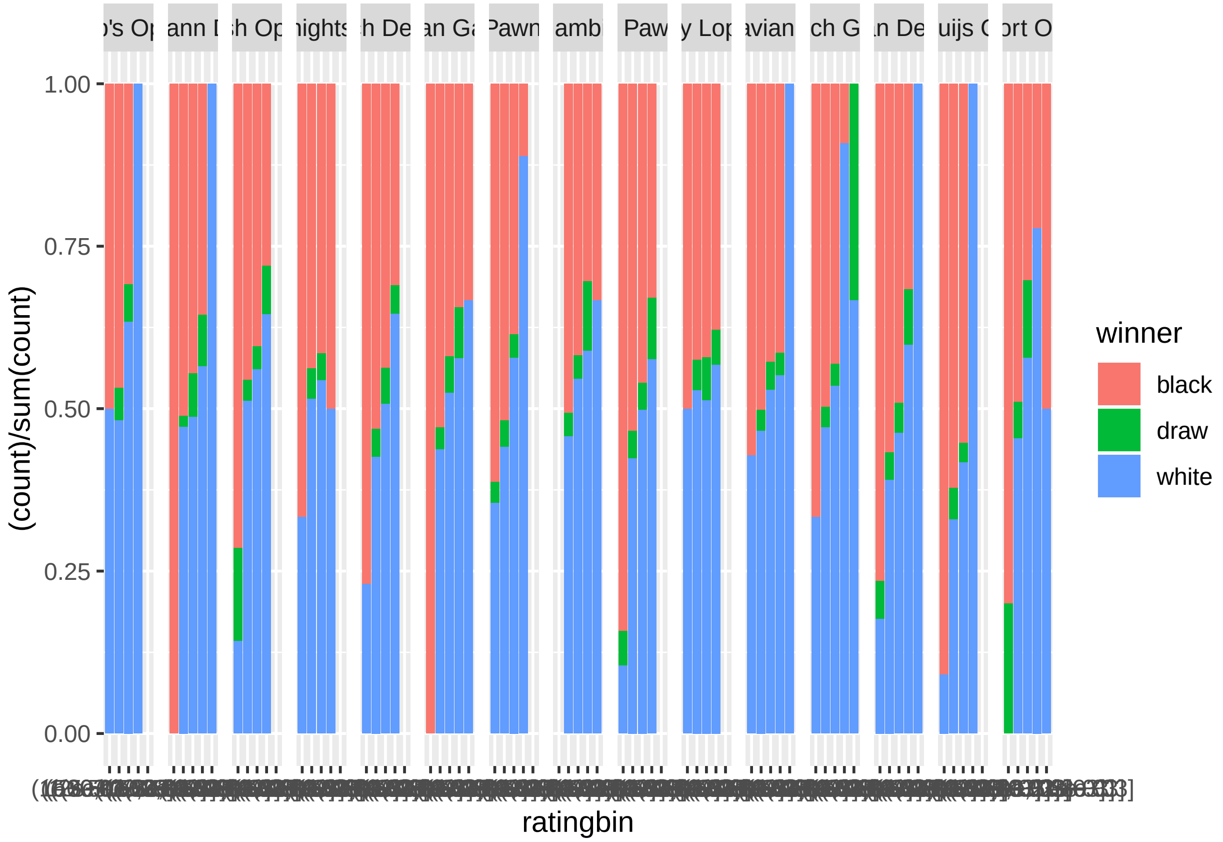
—most common opening by rating?

— draw rate by opening and rating

Best Opening for Me

—highest winrate by rating

```
filtereddata$ratingbin<-cut(filtereddata$white_rating, seq(0,3500,500))
##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_grid(cols=vars(shortname)) ## version 3.0.0
```



—highest winrate by time control

—linechart of performance of top 3 openings by rating

Shiny Interactive

– highest winrate by binned rating as a shiny