# Homework1

Ryan Shea

9/14/2022

## Homework 1

### Ryan Shea

**I pledge my honor that I have abided by the Stevens Honor system.**

First we will import the necessary packages and data. I chose this chess data set because I have become interested in the game within the past year and would like to investigate some trends in the data set of over 20,000 games.

```
library(ggplot2)
library(gridExtra)


library(readr)
# https://www.kaggle.com/datasets/datasnaek/chess
dataf <- read_csv("games.csv")
```

```
## Rows: 20058 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): id, victory_status, winner, increment_code, white_id, black_id, mov...
## dbl (6): created_at, last_move_at, turns, white_rating, black_rating, openin...
## lgl (1): rated
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
colnames(dataf)
```

```
##  [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"
```

```
relevant_columns <- c("winner", "rated", "turns", "victory_status", "increment_code", "white_rating",
                      "black_rating", "moves", "opening_name", "opening_ply")

df <- dataf[relevant_columns]
head(df, 3)
```

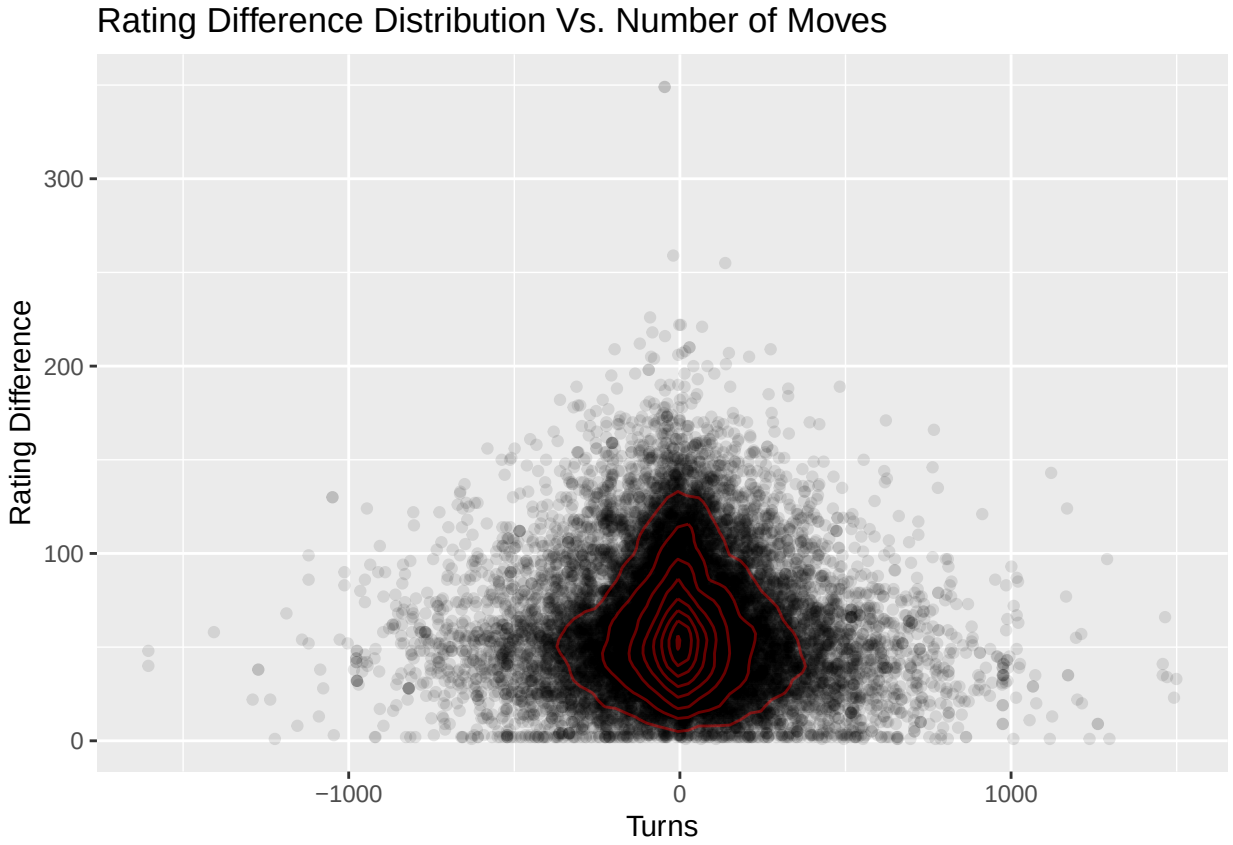
```
## # A tibble: 3 x 10
##   winner rated turns victory_sta~1 incre~2 white~3 black~4 moves openi~5 openi~6
##   <chr>  <lgl> <dbl> <chr>         <chr>     <dbl>   <dbl> <chr> <chr>     <dbl>
## 1 white  FALSE    13 outoftime     15+2       1500    1191 d4 d~ Slav D~       5
## 2 black  TRUE     16 resign        5+10       1322    1261 d4 N~ Nimzow~       4
## 3 white  TRUE     61 mate          5+10       1496    1500 e4 e~ King's~       3
## # ... with abbreviated variable names 1: victory_status, 2: increment_code,
## #   3: white_rating, 4: black_rating, 5: opening_name, 6: opening_ply
```

Now we will do some basic data manipulation in order to create new columns and features in the dataset.

```
df$rating_diff <- df$white_rating - df$black_rating # If positive, white is better
df$average_rating <- (df$white_rating + df$black_rating) / 2
```

The first visualization I will create will be investigating the relationship between the difference in rating and the total number of moves in the game. Because of that I will use a scatter plot with a high level of transparency (as it is a large data set) and layer on a density plot to illustrate it further.

```
ggplot(df, aes(rating_diff, turns)) +
  geom_point(alpha=0.1) +
  geom_density2d(colour = 'red', alpha = 0.4) +
  labs(
    title = "Rating Difference Distribution Vs. Number of Moves",
    x = "Turns",
    y = "Rating Difference")
```

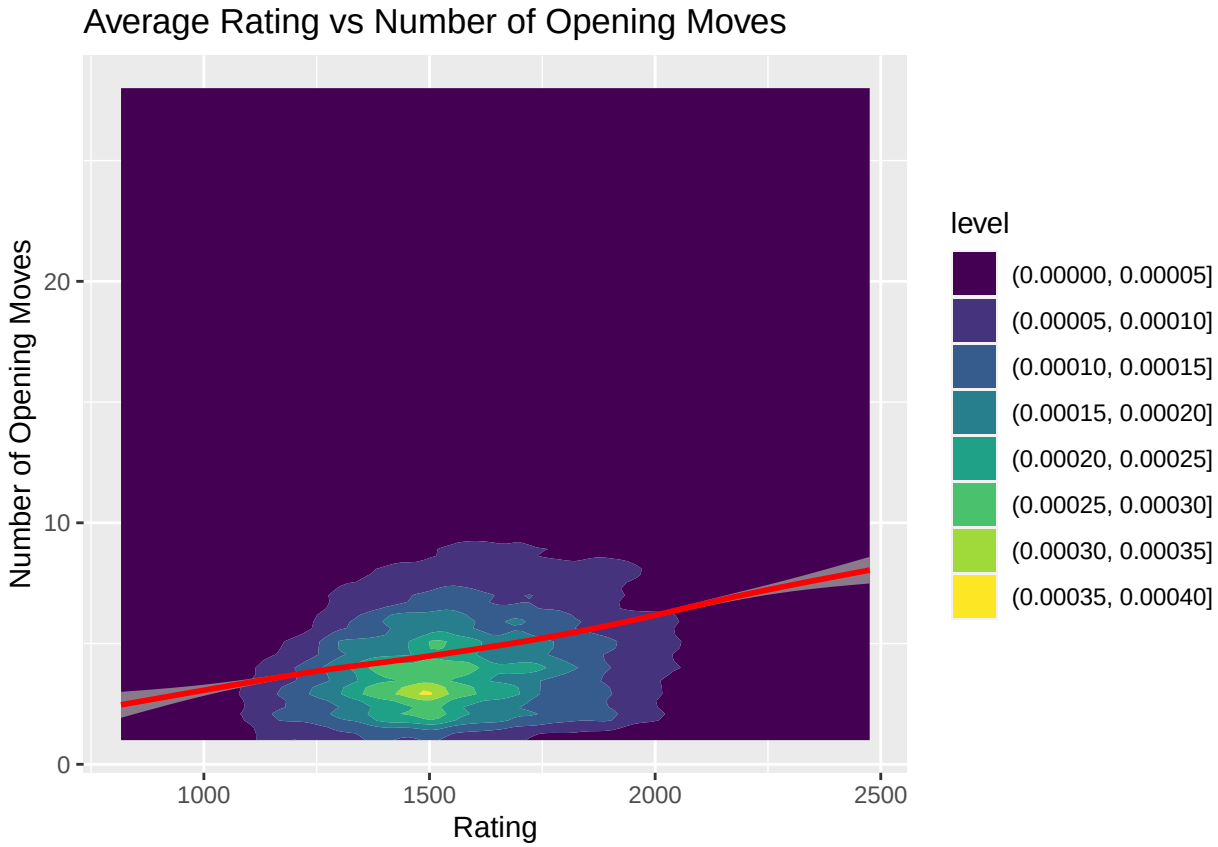



This distribution is showing how as the rating increases, the average number of turns decreases which makes sense. It is more likely that the lower rated player will make more blunders and the higher rated player will capitalize on it quicker than someone their rating would.

For our next visualization we will look at the average rating and the number of moves in the opening. I believe that as the rating increases the number of opening moves will increase as they know more chess theory and following the opening generally will put them at an advantage. The best way to see this is using a filled density chart and to add a trend line on the top.

```
ggplot(df, aes(average_rating, opening_ply)) +
  geom_density2d_filled(h = c(100, 1.6)) +
  labs(
    title = "Average Rating vs Number of Opening Moves",
    x = "Rating",
    y = "Number of Opening Moves"
  ) +
  geom_smooth(
    color = 'red',
    alpha = 0.8,
    method = 'gam'
    )
```

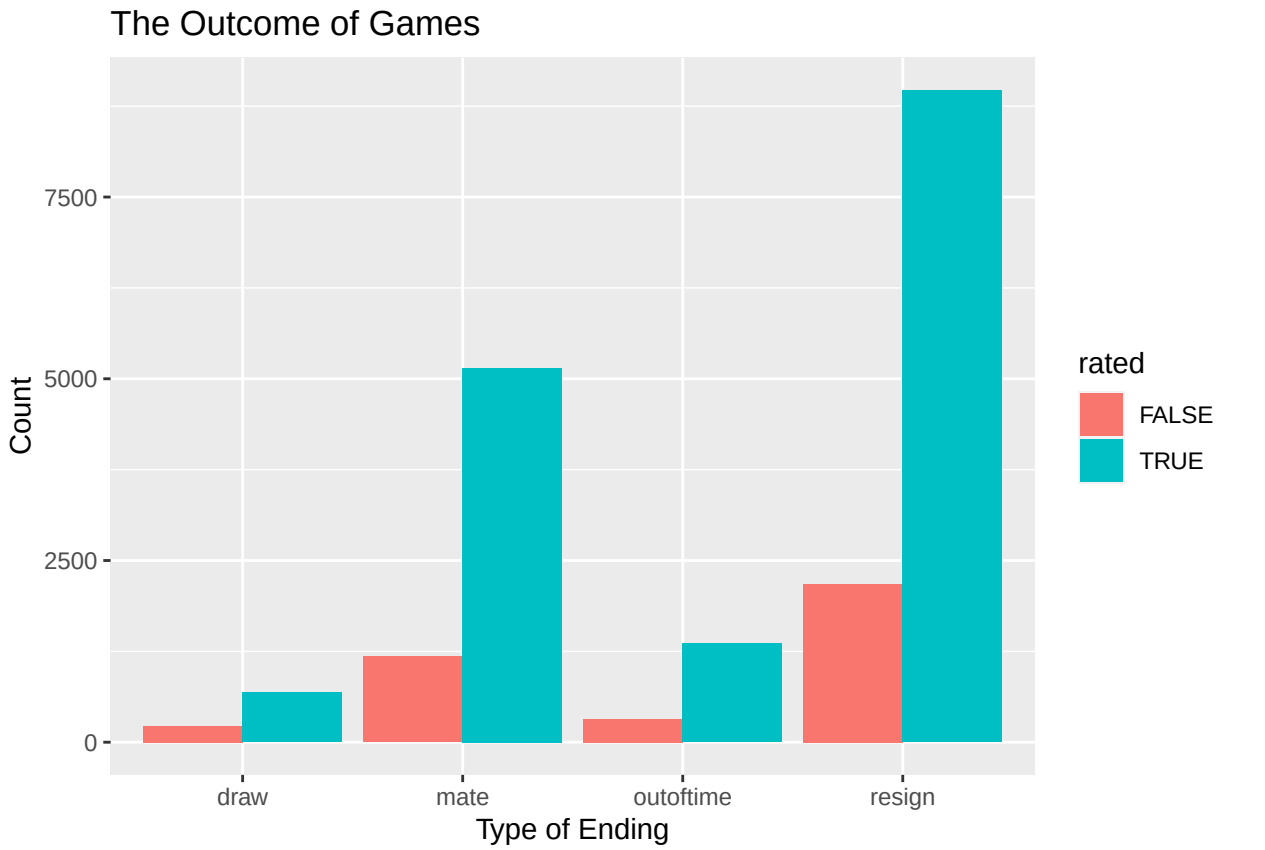
```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```




According to this density chart, you can see a slight increase in the opening moves as the rating increases (up until about 1500). This makes sense as it slightly goes down after that where they might see a mistake most people would miss and might ditch theory in order to capitalize. Even with the drop on density, you can still see there is a strong linear relationship between the rating of the two players and the number of opening moves.

For the last visualization, I will investigate the outcomes of the games and see if whether or not the game being rated has to do with this outcome. I will use a bar chart for this as it will be easy to see the different outcomes (as factors) easily side by side.

```
ggplot(df, aes(victory_status, fill = rated)) +
  geom_bar(position = 'dodge') +
  labs(
    title = "The Outcome of Games",
    x = "Type of Ending",
    y = "Count"
  )
```




The vast majority of games end in resignation whether or not it is rated. You can see that clearly in the bar chart and it seems like whether or not it is rated does not matter in predicting the type of ending.