Final Project: Chess Game Dataset

Isabella Sturm

2/28/2021

Data Importing & Cleaning Steps

Data importing and cleaning steps are explained in the text and in the Github exercises. (Tell me why you are doing the data cleaning activities that you perform). Follow a logical process. Write a paragraph that involves data cleaning requirement such as variable selection, outlier detection and missing data.

One of the first steps of data cleaning is variable selection. Considering the chess game dataset, I am interested in data that affects who will win a chess game such as how long a game lasted (Created_At, Last_Move_At), how many turns it took (Turns), What the game status was (victory_status) and who won (winner), and the player rating (white_rating, black_rating). Other potentially interesting variables are opening_ply (the number of moves in the opening phase), as well as opening_eco (the opening move) and moves (character string representing the game play); though the last two may be hard to analyze in the scope of this project. Next is outlier detection. In this dataset, I wouldn't expect there to be much room for outliers. The data is pretty simple and straightforward being that it is given a chess game between two players, likely of similar ranking, what was the outcome of the game. While there aren't outliers to detect, it is important to know whether or not there is missing data which can be done by calling is.na however this dataset also does not have any missing data because it was collected from online games.

1. Load the data into RStudio
2. Removed columns that are not needed: rated, increment_code, white_id, black_id. Who the player is doesn't really matter; we care more about their rating only.
3. Victory_status is categorical (outoftime, resign, mate, draw) so this will be turned into a factor
4. Winner needs to be numeric because it is out dependent variable. 1=white won, 0=black won, -1=draw
5. Created_At and Last_Move_At can be used to determine how long the game will be. It currently represented as an integer so they will need to be changed to dates. Also, a new column will be created "Game_Length" for how long the game lasted
6. The rest of the data is already in the format we would want them.

```
# Load libraries
library(ggplot2)

# Load the data
setwd('/Users/isabellasturm/Documents/DataScience/DSC520Winter2020/ChessFinal/')
full_chess_df <- read.csv('games.csv')

# Clean the data
# Remove unneeded columns
chess_df <- subset(full_chess_df, select=c(id, created_at, last_move_at, turns,
                                           victory_status, winner, white_rating,
                                           black_rating, moves, opening_eco,
                                           opening_name, opening_ply))
# Set factors: victory_status
chess_df$victory_status <- as.factor(chess_df$victory_status)
# Set winner to numeric: white = 1, black = 0, draw = -1
```

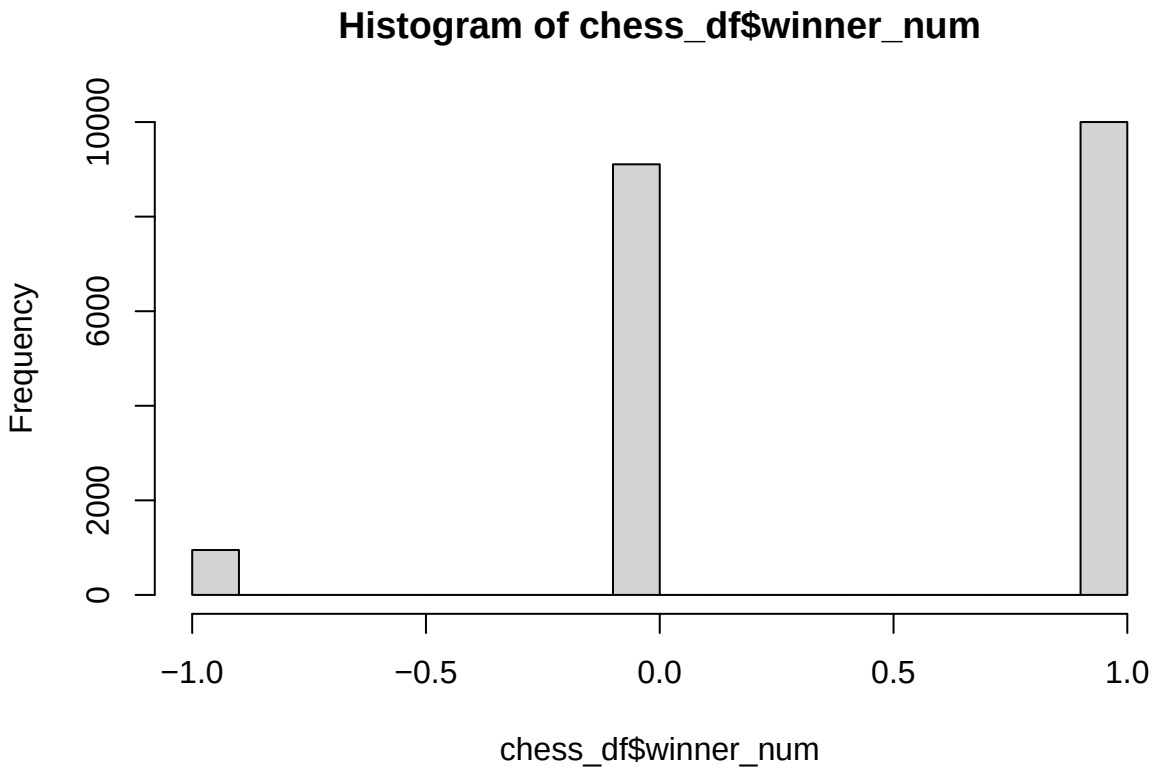
```r
chess_df$winner_num[chess_df$winner == "white"] <- 1
chess_df$winner_num[chess_df$winner == 'black'] <- 0
chess_df$winner_num[chess_df$winner == 'draw'] <- -1
chess_df$winner_num <- as.numeric(chess_df$winner_num)
# Set dates: created_at, last_move_at
chess_df$created_at <- as.Date.numeric(chess_df$created_at, origin='1970-01-01')
chess_df$last_move_at<- as.Date.numeric(chess_df$last_move_at, origin = '1970-01-01')
# Create new column: game_length, rating_diff
chess_df$game_length <- difftime(chess_df$last_move_at, chess_df$created_at, unit="secs")
chess_df$rating_diff <- chess_df$white_rating - chess_df$black_rating # negative if black is rated high
```

Data Summary

```r
# Get a summary of the data
summary(chess_df)
```

```
##       id              created_at              last_move_at
##  Length:20058       Min.   :-5877641-06-23   Min.   :-5877641-06-23
##  Class :character   1st Qu.:-5877641-06-23   1st Qu.:-5877641-06-23
##  Mode  :character   Median :-5877641-06-23   Median :-5877641-06-23
##                     Mean   :-5877641-06-23   Mean   :-5877641-06-23
##                     3rd Qu.:-5877641-06-23   3rd Qu.:-5877641-06-23
##                     Max.   :-5877641-06-23   Max.   :-5877641-06-23
##      turns          victory_status     winner           white_rating
##  Min.   :  1.00   draw     :  906   Length:20058       Min.   : 784
##  1st Qu.: 37.00   mate     : 6325   Class :character   1st Qu.:1398
##  Median : 55.00   outoftime: 1680   Mode  :character   Median :1567
##  Mean   : 60.47   resign   :11147                      Mean   :1597
##  3rd Qu.: 79.00                                       3rd Qu.:1793
##  Max.   :349.00                                       Max.   :2700
##   black_rating     moves           opening_eco        opening_name
##  Min.   : 789   Length:20058       Length:20058       Length:20058
##  1st Qu.:1391   Class :character   Class :character   Class :character
##  Median :1562   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1589
##  3rd Qu.:1784
##  Max.   :2723
##   opening_ply       winner_num      game_length        rating_diff
##  Min.   : 1.000   Min.   :-1.0000   Length:20058       Min.   :-1605.0
##  1st Qu.: 3.000   1st Qu.: 0.0000   Class :difftime    1st Qu.: -108.0
##  Median : 4.000   Median : 0.0000   Mode  :numeric     Median :    3.0
##  Mean   : 4.817   Mean   : 0.4512                      Mean   :    7.8
##  3rd Qu.: 6.000   3rd Qu.: 1.0000                      3rd Qu.:  122.0
##  Max.   :28.000   Max.   : 1.0000                      Max.   : 1499.0
```

```r
# Plot histogram to show frequency of winner of the chess game
hist(chess_df$winner_num)
```



Need to Learn: Data Importing & Cleaning

The dataset I chose seems to be pretty straightforward and not require much work to get cleaned and in the format needed to answer my research questions. The one thing that I think I will need to look into more is how to get the dates into a date format and get the time difference between two dates. I looked a little into this, but I am still getting dates that don't make sense like -5877641-06-23.

Uncovering New Information

The chess dataset currently is nothing more than data – who won, how long/how may moves did it take, who were the players, what moves were made taken directly with no manipulation from an online game. To uncover new information in the dataset, I will aggregate the data to see how often either play wins with what kind of outcome (victory_status) and what the player rating differences were. Also, I will look for correlations and significant variables that affect whether or not one player is more likely to win over the other. For example, is white always more likely to win? Or is white more likely to win only if they are of a higher player rating or only if they can finish the game under x number of moves in x amount of time.

Looking at the Data

Summary statistics are a good place to start for looking at the data. They give a good idea of where to start and where you can look into the data more (I explain this a little bit more in summarizing data).

Plotting the data in various plots (histograms, scatterplots, bar graphs) are great for visualizing the data and getting insights that help answer the questions at hand.

Lastly, looking into correlation of the variables and the outcome are really the best ways for me to answer the questions I have about the chess dataset because I am most interested in factors that affect winning a game of chess (rating, playing black/white, number of moves).

Plan for Expanding Data

I do not plan on slicing and dicing the dataset as I want to compare all the datapoints as one and I don't have another data set to expand upon the current dataset. However, there are a couple new variables I will create, such as rating_diff and game_length to make it easier to compare the ratings of two players and how long it took for a winner to be announced.

Summarizing Data

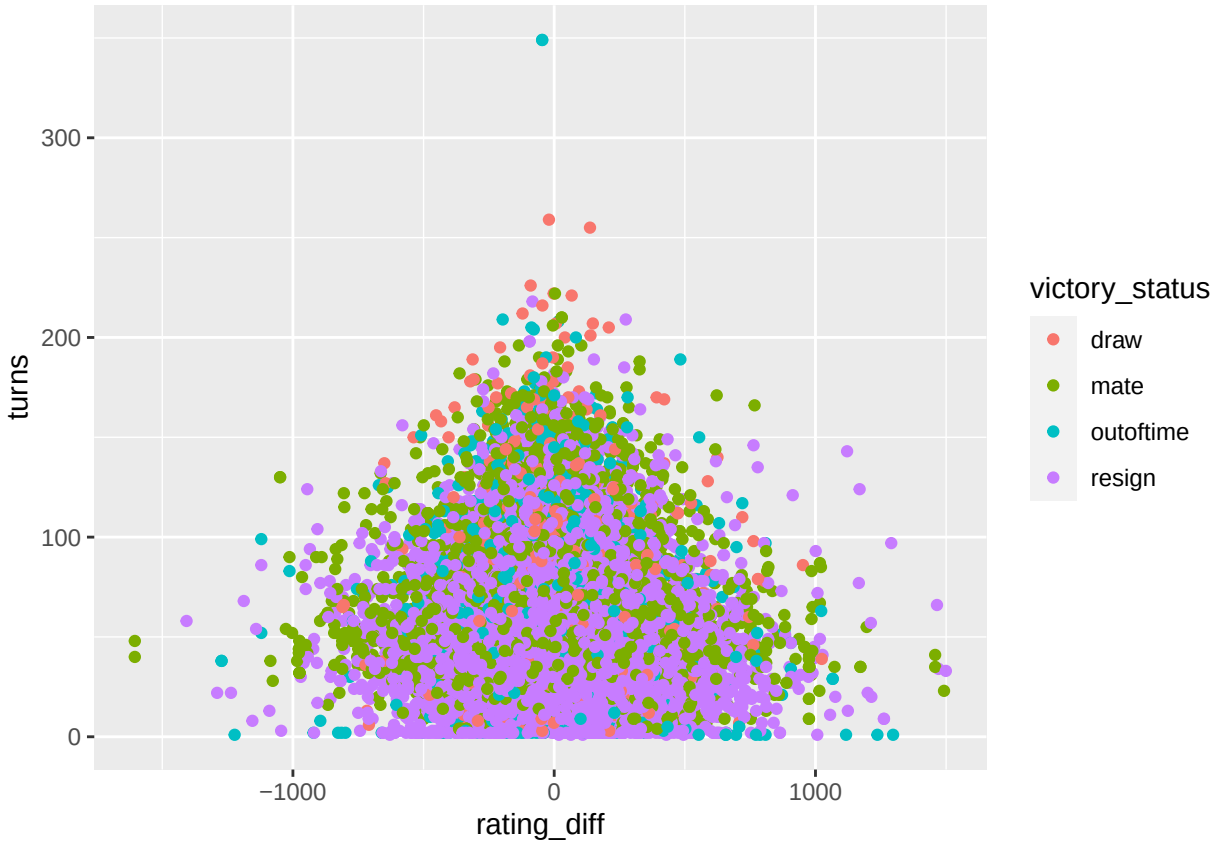
The summary() function is useful for summarizing the data to answer my questions about the data. For instance, looking at the victory_status, mate and resign are two outcomes that where the winner is undisputable and that makes up the majority of the outcomes. Looking at the rating_diff, it is possible that the games have a mismatch in player ratings (the largest difference being 1605 points!); this is interesting and would want to look more into how big of a factor rating plays in winning the match.

A histogram of the winners shows that white wins most often though black wins a lot too, so this means it should look more into the data to see what factors impact who will win.

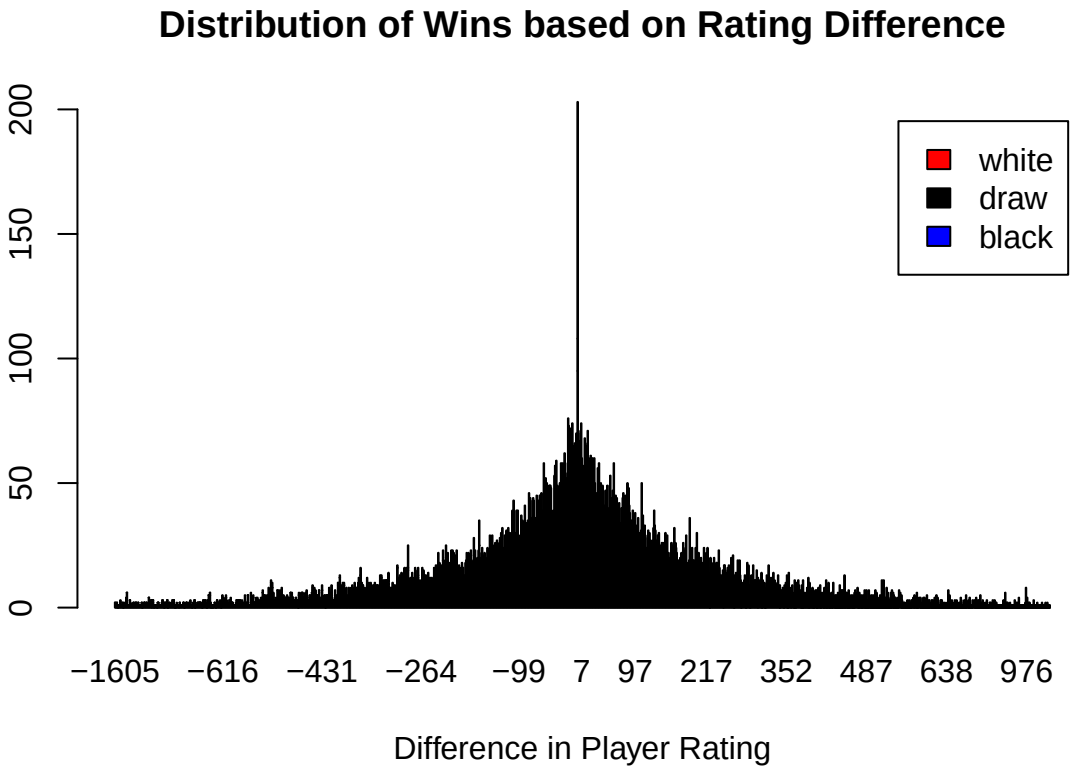
Plots & Tables

Here are a couple plots about the data. There are some tweaks I still need to do/figure out to get it to display properly and more I would like to add. Scatterplots and histograms/barplots are useful visualizations for the data. But most of the analysis will be better explained through correlation analysis.

```
# Scatterplot of player rating difference vs number of moves, color
ggplot(data=chess_df, aes(x=rating_diff, y=turns, color=victory_status)) +
        geom_point()
```



```r
# Barplot of Winner based on Difference in Player Rating
win_counts = table(chess_df$winner, chess_df$rating_diff)
barplot(win_counts, main="Distribution of Wins based on Rating Difference",
        xlab="Difference in Player Rating", col=c("blue", "black", "red"),
        legend=rownames(win_counts))
```



Need to Learn: Answering Questions

I need to learn more about some of the plots I want to use that we haven't gone over in the course, like barplots. I want to make sure my plots are displaying how I want/need. Other than that, it is just finessing the analysis techniques we've learned this term.

Plan for Machine Learning

For the scope of this project, I do not plan on incorporating any machine learning techniques as I believe it goes beyond what is necessary to answer my research questions. I think a next step or a future project on this dataset could be to use machine learning to analyze the moves to determine which moves will result in a win for either player, or even just if there is a certain opening or an opening phase that could results in a win for either player. For example, if the opening move is the Alekhine Defense, are you more likely to win if you complete the opening phase in 2 moves instead of 6 moves or if you open at B02 vs B03?