Final Part 2

Taylor Anderson

8/4/2019

### a.

##### Data importing and cleaning steps are explained in the text and in the DataCamp exercises.(Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.

First, I checked the data to see if there were any N/A values that I needed to worry about. Luckily, the data did not contain any.

apply(games, 2, function(x) any (is.na(x)))

## id rated created\_at last\_move\_at turns   
## FALSE FALSE FALSE FALSE FALSE   
## victory\_status winner increment\_code white\_id white\_rating   
## FALSE FALSE FALSE FALSE FALSE   
## black\_id black\_rating moves opening\_eco opening\_name   
## FALSE FALSE FALSE FALSE FALSE   
## opening\_ply   
## FALSE

Next, I looked through the columns to make sure the variables made sense.

* id
  + While there are 20,058 games in this data set there are 19,113 unique ids. This is a little concerning because there is likely duplicated data in this dataset. I removed these duplicates with the dplyr package.

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.1

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

games <- distinct(games)

* rated
  + The “rated” column showed that there were 4 factors when really it should only be a boolean. I found that the column had TRUE, FALSE, True, and False. So I created a new column called NewRated that converted them all to TRUE and FALSE.

games$NewRated <- ifelse(games$rated==FALSE|games$rated=="False", FALSE, TRUE)

### b.

##### With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

str(games)

## 'data.frame': 19629 obs. of 17 variables:  
## $ id : Factor w/ 19113 levels "005lW0Xz","009mKOEz",..: 15433 10012 10832 9930 3074 11034 13555 14110 5539 3341 ...  
## $ rated : Factor w/ 4 levels "False","FALSE",..: 2 4 4 4 4 2 4 2 4 4 ...  
## $ 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: Factor w/ 4 levels "draw","mate",..: 3 4 2 2 2 1 4 4 4 2 ...  
## $ winner : Factor w/ 3 levels "black","draw",..: 3 1 3 3 3 2 3 1 1 3 ...  
## $ increment\_code: Factor w/ 400 levels "0+12","0+13",..: 111 299 299 188 248 20 20 115 103 20 ...  
## $ white\_id : Factor w/ 9438 levels "--jim--","-l-\_jedi\_knight\_-l-",..: 1288 17 3824 2018 6058 8569 1468 2023 2526 2023 ...  
## $ white\_rating : int 1500 1322 1496 1439 1523 1250 1520 1413 1439 1381 ...  
## $ black\_id : Factor w/ 9331 levels "-0olo0-","-l-\_jedi\_knight\_-l-",..: 12 7749 12 135 135 3066 2041 7867 2041 5691 ...  
## $ black\_rating : int 1191 1261 1500 1454 1469 1002 1423 2108 1392 1209 ...  
## $ moves : Factor w/ 18920 levels "a3 b6 d4 c5 Nf3 Bb7 e3 cxd4 exd4 d6 c4 d5 Nc3 dxc4 Bxc4 Bxf3 Qxf3 Nh6 Bxh6 Nd7",..: 1823 4372 10772 3199 12201 8481 2951 17373 10374 9595 ...  
## $ opening\_eco : Factor w/ 365 levels "A00","A01","A02",..: 249 72 172 242 193 97 240 72 202 73 ...  
## $ opening\_name : Factor w/ 1477 levels "Alekhine Defense",..: 1388 752 651 993 784 1293 92 755 504 1142 ...  
## $ opening\_ply : int 5 4 3 3 5 4 10 5 6 4 ...  
## $ NewRated : logi FALSE TRUE TRUE TRUE TRUE FALSE ...

### c.

##### What do you not know how to do right now that you need to learn to import and cleanup your dataset?

There are a few variables that I am not sure how to use very well right now. For example; moves, opening\_eco, opening\_name, and opening\_ply. I may need to do additional work to make these variables useful to me.

### d.

##### Discuss how you plan to uncover new information in the data that is not self-evident.

My first plan is to try to predict wins and losses based on the ratings of the two players. Then to improve that prediction I will test how much some other variables play into this. The variables that come to mind are the color they are playing and the increment. I would also like to see if there are any changes to the style of game depending on rating. Are there certain openings that are popular at different ratings? Does the average number of turns change? Are there more draws in higher or lower rated games? etc.

### e.

##### What are different ways you could look at this data to answer the questions you want to answer?

I think that logistic and linear regression are going to be the main two ways to answer the questions that I have. There may be some k-nearest applications that I could use as well but I haven’t thought that through quite as much yet.

### f.

##### Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

I had some concern about games that are rated (meaning that they impact your rating) vs. games that are not rated. I was worried that people may not play their best in unrated games or they may try to play in a way that doesn’t reflect their ability (blind-folded, trick games, etc.) I am going to analize this and see if there is a major difference. I am hoping I can keep them in because unrated games are a large part of the data set.

I have already made a new variable in the data cleaning process called NewRated but that wasn’t for analytical reasons. One new variable I am sure that I will make is to break up the increments. Increments are in the format X+Y where X is the amount of minutes in the game and Y is the number of seconds you get after each move. For example, a common increment is 15+10 where each player starts with 15 minutes and gets 10 seconds added to the timer after each move. There are 400 different increments in this data set which is too many to work with. I plan on breaking these into 3 to 5 groups ranging from short games to long games. I may do a similar thing with openings. There are 1,477 different openings in the data set but I may break them into groups for the first 1 or two moves so it is more manageable.

### g.

##### How could you summarize your data to answer key questions?

Like I have mentioned in other question, by using plots, linear regression, and possibly knn models I will be able to see what kinds of variables will impact the results of a game of chess. After finding which variables have the largest impact on the game, I will create training and testing data sets to see if I can accuratly predict the outcome of games which would answer my question.

### h.

##### What types of plots and tables will help you to illustrate the findings to your questions?

I will use histograms to summarize some of the different variables in the data set. Scatter plots will be used to show correlation along with linear models. I have seen a few clever uses of heat maps for chess data that I would like to try as well.

### i.

##### Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I will be creating a knn model to see how it compares to a linear model. I am not sure how well it will work with the data set so I am excited to learn more about the methods advantages and limitations.