Chess.com Player Ratings and Data

Luke Clement and Nathan Zeigler

Assignment Description

In this project you are going to use the skills that you've learned about regression on a dataset of your own. You may choose any dataset that you wish as long as it is not one that we've already discussed in the course. You may want to consult me about your choice of dataset, just to make sure it is suitable.

After making a suitable dataset choice, you need to complete the following steps:

- Narrative: You need to formulate a question in which you can address using your chosen techniques. This is the overall goal of your analysis.
- You need to perform proper pre-processing and cleaning of the data before your analysis begins. Depending on your data, this step may be fairly short or quite lengthy.
- You need to have a substantial exploratory data analysis (EDA) section. This section should include summaries, graphs (univariate, bivariate, and possibly multivariate), and other techniques from DS 1 to describe your data. You should also investigate possible interactions between variables. Your EDA should show a progression of understanding about the data and your research question.
- You need to choose at least two regression techniques (most likely a multiple linear regression model and a penalized regression method) to use in your analysis. You should explain your modeling choices and how they were informed by your EDA.
- You need to address the assumptions of each method with graphical and/or numeric evidence.
- You need to use cross-validation or a related method to compare the two or more methods.
- You need to come to your final answer using an iterative process that you show throughout your project.
- You need to discuss the shortcomings of your modeling approach. Also, if appropriate, you discuss improvements that could be made.
- You need to discuss how the model approach/output works toward answering the question.

- You need to discuss your major takeaways from the project. This part is meant to be a reflection on what you learned about the data and your increase in knowledge about data science during the process of the project.

Place Work Below

Narrative: I think a quick introduction to the data is important to understanding what our project is all about. We decided that we wanted to do a project on titled chess players and their associated stats on chess.com. Games are broken into 3 major categories: bullet, blitz, and rapid. In bullet games both players receive 1-2 minutes for the duration of the game. In blitz they receive 3-5 minutes, and in rapid they receive 10-60 minutes. We wanted to look at how well the bullet, blitz, rapid, and FIDE ratings of titled players correlated with other factors on chess.com. (This also includes averaging the three online ratings)

name - Acts like an index (actual usernames were removed for privacy reasons) title - Awarded to players for performance at high rated chess tournaments and based on their FIDE rating. The titles are awarded based on the following ratings. Grandmaster (GM) - 2500+ International Master (IM) - 2400+ FIDE Master - (FM) 2300+ Candidate Master - (CM) 2200+ Woman Grandmaster - (WGM) 2300 + Woman International Master - (WIM) 2200+ Woman FIDE Master (WFM) 2100+ Women Candidate Master (WCM) 2000+ Note that women can earn all titles and it is personal preference whether or not they use GM vs WGM, etc

country - abbreviation of country they represent joined - When they joined chess.com last online - When they were last online followers - Number of people that follow them on chess.com league - Another form of ranking chess.com uses. From worst to best: Legend, Champion, Elite, Crystal, Silver, Bronze, Stone, Wood bullet_last - Bullet rating (as of taking this data) bullet_best - Best Bullet rating bullet_wins - Bullet wins bullet_losses - Bullet Losses bullet_draws - Bullet draws blitz_last - Blitz (as of taking this data) blitz_best - Best Blitz Rating blitz_wins - Blitz Wins blitz_losses - Blitz Losses blitz_draws - Blitz draws rapid_last - Rapid rating (as of taking this data) rapid_best - Best Rapid rating rapid_wins - Rapid wins rapid_losses - Rapid losses rapid_draws - Rapid draws online_rating - (average of three ratings) fide - FIDE rating (FIDE is the international chess governing association) tactics_rating - Tactics rating puzzle_attempts - Number of puzzles attempted for best puzzle rush puzzle_best - Number of correct puzzles for puzzle rush on best attempt

```
library("jsonlite")
library("httr")
library("tidyverse")
library("tidymodels")
library("rvest")
library("XML")
library("ggplot2")
```

```
library("dplyr")
library("caret")
```

Get csv data

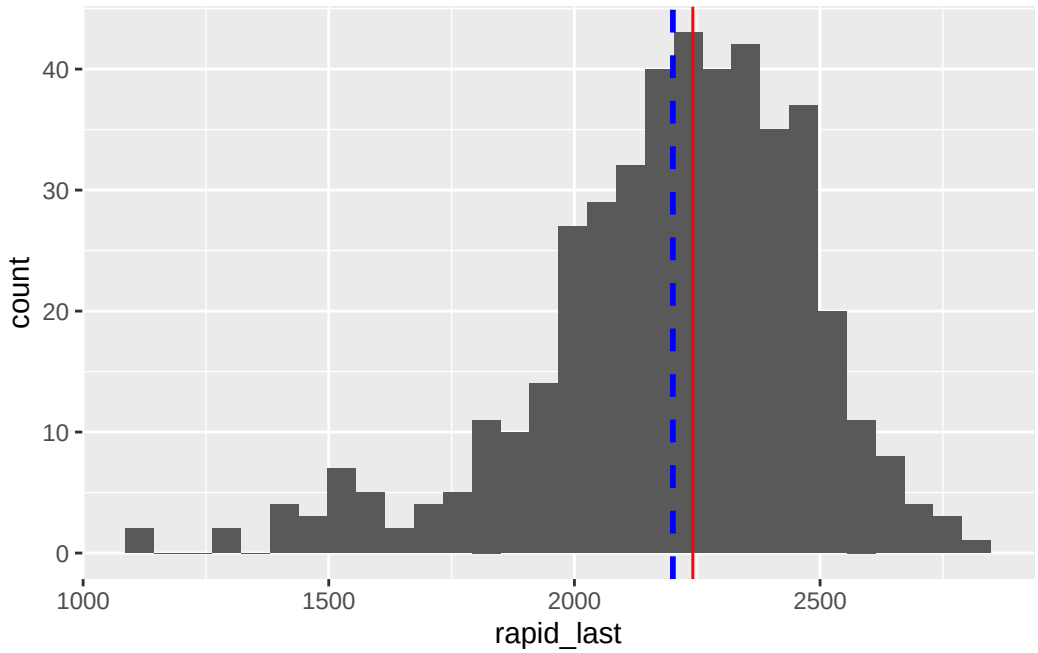
```
players <- read_csv("player_data.csv")

players$rapid_total <- players$rapid_draw + players$rapid_win + players$rapid_loss
players$blitz_total <- players$blitz_draw + players$blitz_win + players$blitz_loss
players$bullet_total <- players$bullet_draw + players$bullet_win + players$bullet_loss

players <- players %>%
  mutate(online_rating = (rapid_last+blitz_last+bullet_last)/3,
         total_games = (rapid_total + blitz_total + bullet_total),
         wins = (rapid_win + bullet_win + blitz_win),
         draws = (rapid_draw + bullet_draw + blitz_draw),
         losses = (rapid_loss + bullet_loss + blitz_loss)) %>%
  filter(fide > 1000 & fide < 2900)

players$country <- as.numeric(as_factor(players$country))
players$title <-  as.numeric(as_factor(players$title))

rap_total_mean <- mean(players$blitz_last)
```

EDA: Looking at the distribution of ratings from our sample. Note that for the simplicity of our EDA, some graphs will be discussed but not shown because it would be excessive to include all of them. These graphs will still be present just commented out.

First: Rapid Ratings

```
#look at rating distribution graphically
rap_mean <- mean(players$rapid_last)

rap_median <- median(players$rapid_last)
players %>%
  ggplot(aes(x = rapid_last)) +
  geom_histogram() +

  #adding line for mean
```

```
  geom_vline(aes
             (xintercept= rap_mean),
            color="blue", linetype="dashed", size=1) +

  #adding line for median
  geom_vline(xintercept = median(players$rapid_last),
             col = "red",show.legend = true
            )
```




```
rap_mean
```

```
[1] 2200.442
```

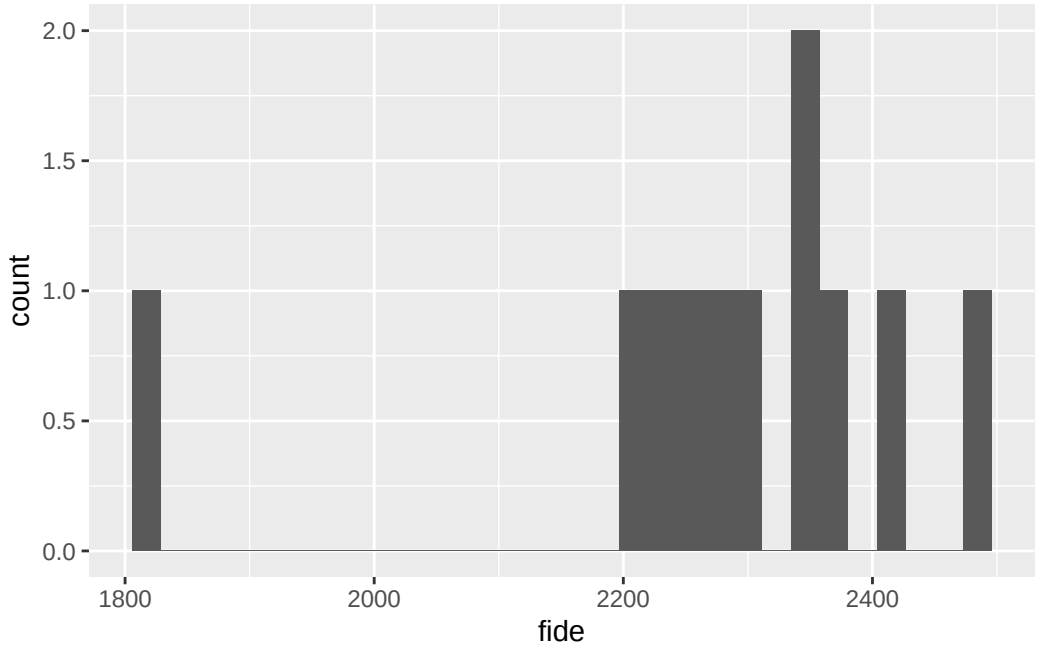
```
rap_median
```

```
[1] 2241
```

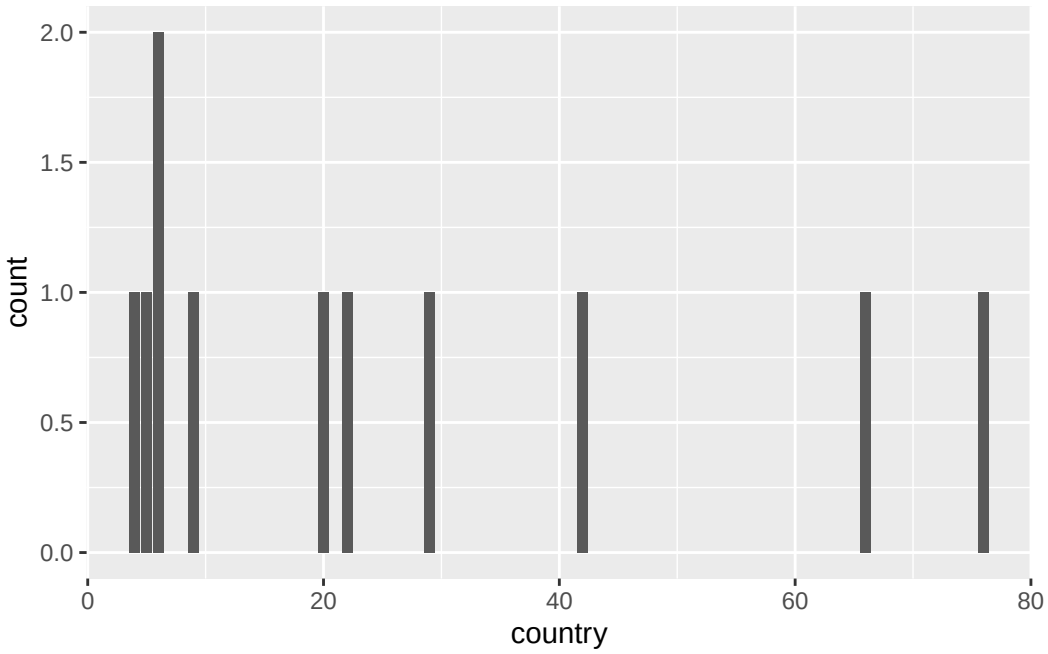
Rapid ratings are skewed to the left. We see about a 100 point gap between the mean and median with the median being higher. There are potential outliers with titled players rated below 1500, which is really surprising.

Let's look at the players with the lowest rapid_ratings

```
low_rap_players <- players %>%
  arrange(rapid_last) %>%
  filter(rapid_last < 1500)

#Look at the title distribution of the players with the lowest ratings

# low_rap_players %>%
#   #ordering from highest frequency to lowest frequency
#   ggplot(aes( x = reorder(title, -table(title)[title]))) +
#               geom_bar() + labs(x = "chess title")

#look at the distribution of players with the lowest rapid ratings by fide
low_rap_players %>%
  ggplot(aes(x = fide)) + geom_histogram()
```




```
#look at country distribution
low_rap_players %>%
  ggplot( aes( x = country)) + geom_bar()
```



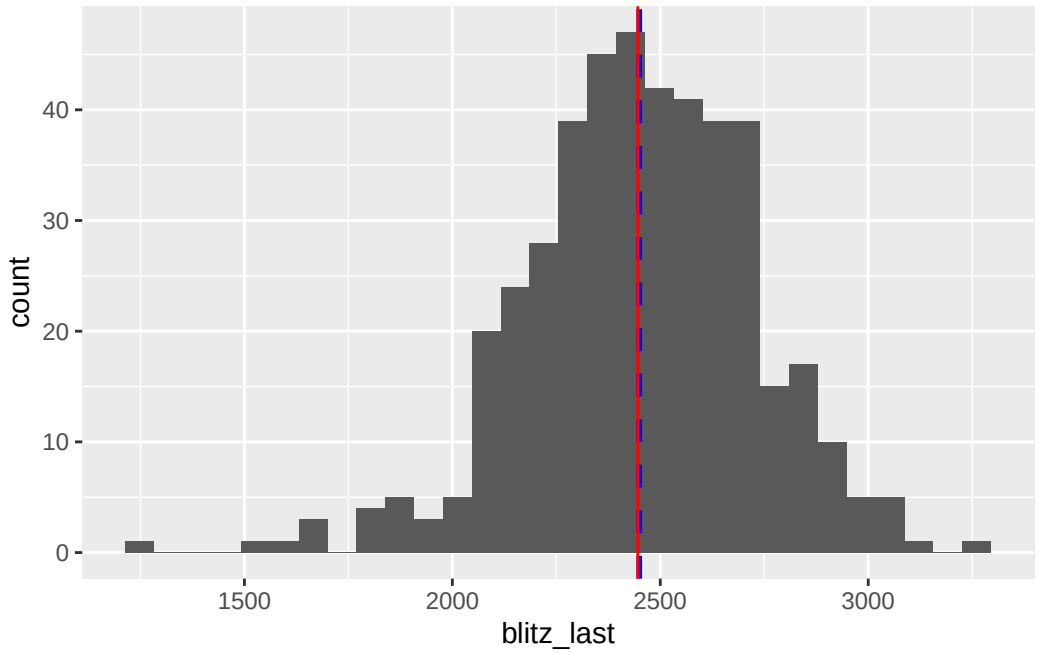
The listed FIDE ratings (the official international ratings of players) looks like it's much higher than their listed Rapid Ratings. This is a bit weird. Generally speaking, the conventional wisdom is that your FIDE rating should be more than 200 points higher than your Chess.com rapid rating.

If we look at the lowest rapid ratings by country, there isn't really anything notable. US and Russia probably have a count higher than 1 because they have more chess players in that country, but it's hard to know using just this data

Now let's do the same thing for blitz and bullet rating!

```
#Blitz first
bltz_mean <- mean(players$blitz_last)

bltz_median <- median(players$blitz_last)

players %>%
  ggplot(aes(x = blitz_last)) +
  geom_histogram() +

  #adding line for mean
  geom_vline(aes
             (xintercept= bltz_mean),
            color="blue", linetype="dashed", size=1) +
```

```
  #adding line for median
  geom_vline(xintercept = bltz_median,
             col = "red"
            )
```




```
bltz_mean
```

```
[1] 2449.367
```

```
bltz_median
```

```
[1] 2446
```

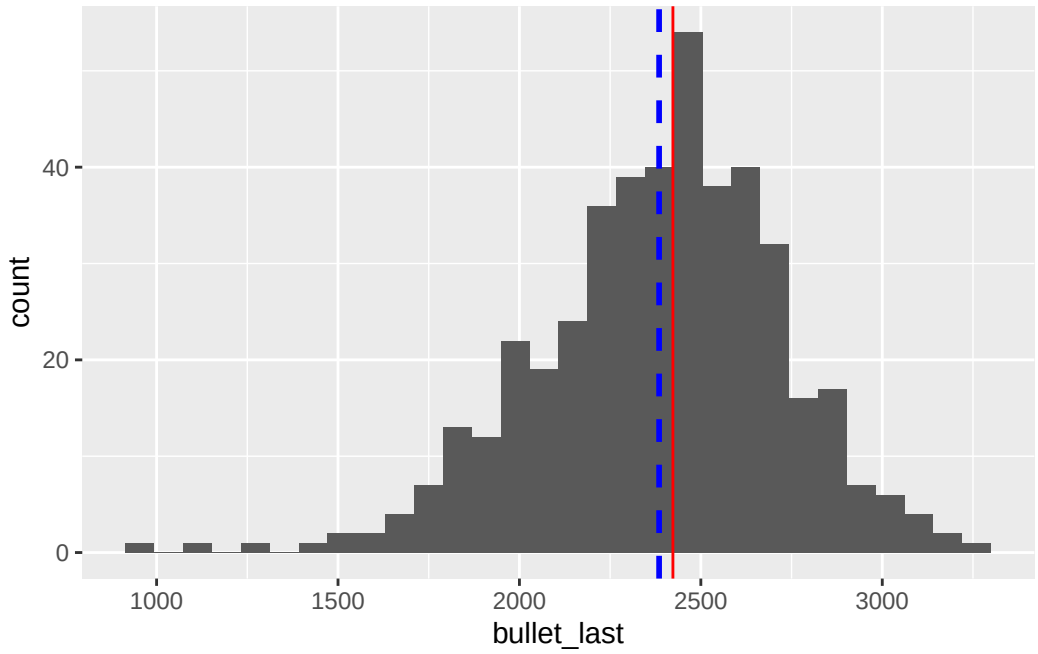
```
#basically the same number!

#Now let's look at the 20 lowest rated players
# low_blitz_players <- players %>%
#   arrange(blitz_last) %>%
```

```
#   head(20)

#Look at the title distribution of the players with the lowest ratings

# low_blitz_players %>%
#   #ordering from highest frequency to lowest frequency
#   ggplot(aes( x = reorder(title, -table(title)[title]))) + geom_bar()
```

Big difference doing the same thing with blitz ratings. The blitz ratings look pretty normally distributed and we see that no grandmasters are within the lowest 20 blitz rated players and the frequency of titled player occurs from the lowest titles to the highest titles (WCM being the lowest and WIM). The titles of players may be a good predictor for rating.

Now for bullet!

```
bull_mean <- mean(players$bullet_last)

bull_median <- median(players$bullet_last)

players %>%
  ggplot(aes(x = bullet_last)) +
  geom_histogram() +

  #adding line for mean
  geom_vline(aes
             (xintercept= bull_mean),
            color="blue", linetype="dashed", size=1) +

  #adding line for median
  geom_vline(xintercept = bull_median,
             col = "red"
            )
```




```
bull_mean
```

```
[1] 2384.934
```

```
bull_median
```

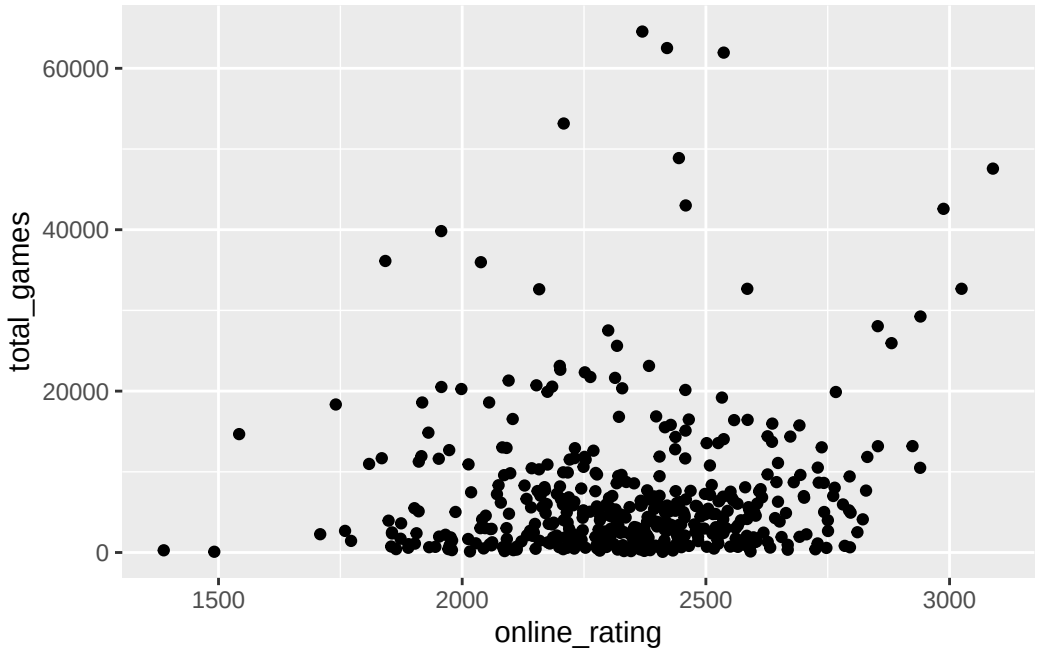
```
[1] 2423
```

We see pretty similar results to blitz ratings. A bit of a bigger gap between the mean and median, but it's bullet so we expect more volatility and unexpectedness because of how fast the time control is. It's good to know that bullet ratings seem to behave similarly to rapid and blitz ratings.

Is there any correlation between number of games played and rating?

```
# players %>%
#   ggplot(aes(x = rapid_last, y = rapid_total)) + geom_point()
#
# players %>%
#   ggplot(aes(x = blitz_last, y = blitz_total)) + geom_point()
#
```

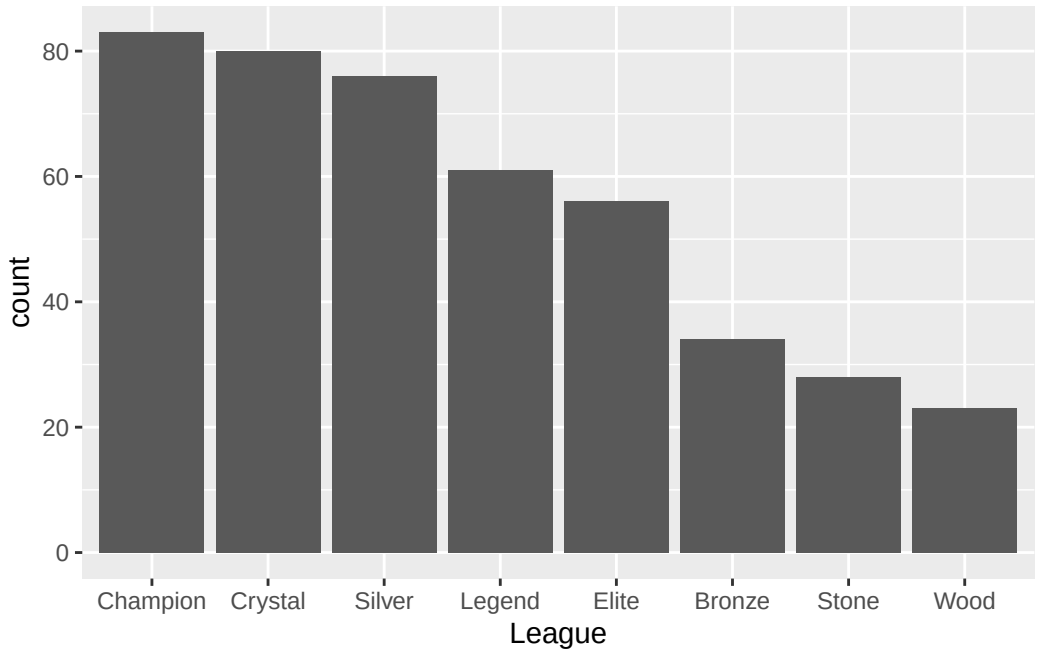
```
 # players %>%
 #   ggplot(aes(x = bullet_last, y = bullet_total)) + geom_point()

players %>%
   ggplot(aes(x = online_rating, y = total_games)) + geom_point()
```



Visually, there seems to be a weak correlation between the number of games played in that time control and the associated rating. Also, adding up total games compared to average rating produces the same result. It will be interesting to see how significant the number of games played will affect predictions of ratings.

"League" is a variable in the data set that is an additional ranking method on chess.com. Players can gain points to move up a league by winning games. It encourages people to play more, but is not related to rating. Let's take a look into this variable! (the order, from highest league to lowest league goes from Legend, Champion, Elite, Crystal, Silver, Bronze, Stone, Wood)

```
#What's the league distribution for our sample of titled players?
players %>%
  ggplot(aes( x = reorder(league, -table(league)[league]))) +
  geom_bar() + labs(x="League")
```



Looks like the leagues distribution is pretty spread out with the middle and upper middle leagues being the most common.

Let's see if there's any correlation in the league and the ratings of the players in the league

```
library(magrittr)

players %>%
   group_by(league) %>%
   summarize(Rapid = median(rapid_last), Blitz = median(blitz_last),
             Bullet = median(bullet_last)) %>%
   arrange(-Rapid)
```

```
# A tibble: 8 x 4
  league   Rapid Blitz Bullet
  <chr>    <dbl> <dbl>  <dbl>
1 Champion 2309  2441   2436
2 Legend   2302  2442   2378
3 Wood     2253  2434   2488
4 Elite    2246  2509   2449
5 Crystal  2208. 2474.  2450
6 Stone    2202. 2400   2324
7 Bronze   2196  2423   2412.
```

```
8 Silver   2132. 2427   2336.
```

```
players %>%
  group_by(league) %>%
  summarize(Rapid = mean(rapid_last), Blitz = mean(blitz_last),
            Bullet = mean(bullet_last)) %>%
  arrange(-Rapid)
```

```
# A tibble: 8 x 4
  league   Rapid Blitz Bullet
  <chr>    <dbl> <dbl>  <dbl>
1 Legend   2313. 2494.  2414.
2 Champion 2268. 2465.  2426.
3 Wood     2213. 2399   2392.
4 Bronze   2182. 2413.  2359.
5 Elite    2173. 2473.  2388.
6 Crystal  2168. 2442.  2398.
7 Stone    2142. 2409.  2314.
8 Silver   2118. 2434.  2336.
```

There doesn't appear to be any strong correlation between the league someone is in and their online ratings. However, as expected, the top leagues, legend and champion are on top, but unexpectedly the lowest league is third. League may not be a good predictor

```
players %>%
  group_by(country) %>%
  count()%>%
  arrange(-n)
```

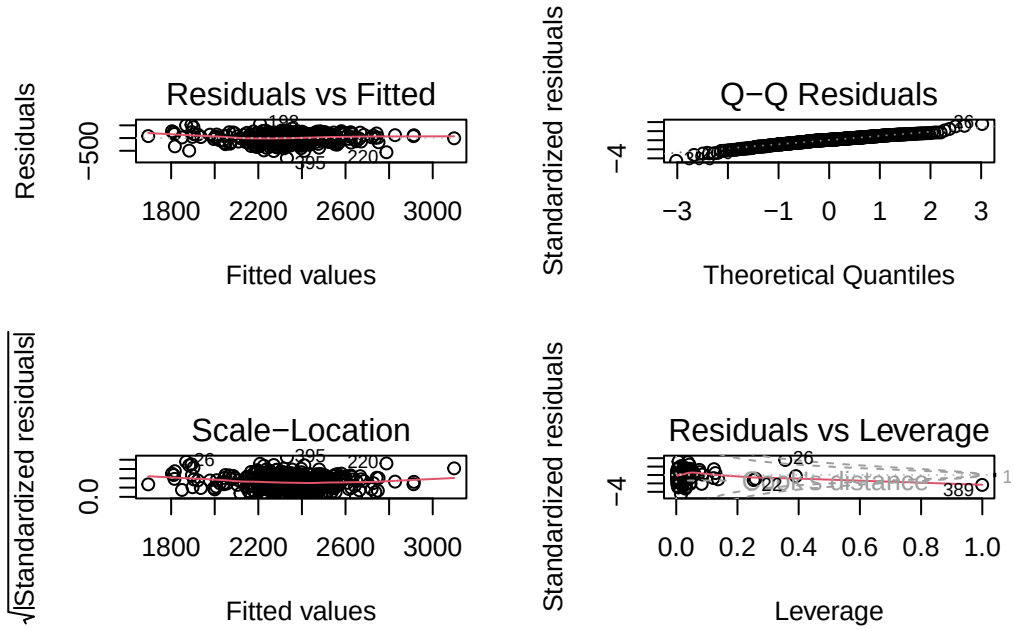
```
# A tibble: 95 x 2
# Groups:   country [95]
   country     n
     <dbl> <int>
 1       5    83
 2       4    24
 3       6    21
 4      12    18
 5      22    18
 6       3    12
 7      21    12
```

```
 8      29    12
 9      20    11
10      37    11
# i 85 more rows
```

We see that US has the most amount of players in our data set. This is followed by India and Russia. This is expected because Chess.com is a US-based country and has lots of US players, while India and Russia are known as big chess countries. Ultimately, nothing surprising about this tibble.

Consolidating the different online ratings into one online rating and doing analysis from there. We are going to first use a linear analysis.

```
#Let's do a multivariate linear regression model.

#split data into testing and training
split <- initial_split(players, prop = .9)

train <- training(split)
test <- testing(split)

players_model <- lm(online_rating ~ fide + followers + tactics_rating + country +
                    total_games + wins + draws + losses , data = train)

par(mfrow = c(2, 2))
plot(players_model)
```




```
summary(players_model)
```

```
Call:
lm(formula = online_rating ~ fide + followers + tactics_rating +
    country + total_games + wins + draws + losses, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-788.92 -101.32   15.19  111.04  553.39

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.427e+02  1.183e+02   2.052   0.0409 *
fide            8.092e-01  5.125e-02  15.790  < 2e-16 ***
followers      -1.367e-04  2.012e-04  -0.679   0.4974
tactics_rating  7.840e-02  1.441e-02   5.440 9.45e-08 ***
country        -2.092e-03  3.692e-01  -0.006   0.9955
total_games    -1.649e-02  7.329e-03  -2.249   0.0250 *
wins            1.965e-02  1.247e-02   1.575   0.1160
draws           1.226e-01  3.118e-02   3.933 9.94e-05 ***
losses                 NA         NA      NA       NA
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 174.6 on 388 degrees of freedom
Multiple R-squared:  0.5283,    Adjusted R-squared:  0.5198
F-statistic: 62.09 on 7 and 388 DF,  p-value: < 2.2e-16
```

```r
predictions <- predict(players_model, test)
data.frame( R2 = R2(predictions, test $ online_rating),
            RMSE = RMSE(predictions, test $ online_rating))
```

```
        R2     RMSE
1 0.6088011 174.5512
```

```r
players$loss
```

```
NULL
```

The assumptions for the linear graph appear to be sufficient. According to this linear regression model, FIDE ratings, tactics ratings, and number of draws are the strongest predictors of ratings. As expected, country isn't a strong predictor. Conceptually, it makes sense that more draws happen as ratings rise because if both players play perfectly, the result is always a draw. FIDE and Tactics ratings also make sense because they are similar rating systems that measure strength. Ultimately, nothing surprising here.

Now let's build a regression decision tree to compare our results

```r
#set model
tree_model <- decision_tree(
  mode = "regression",
  cost_complexity = tune(),
  tree_depth = tune()
) %>%
  set_engine("rpart")

#set recipe
tree_recipe <- recipe(online_rating ~ fide + followers + tactics_rating +
                        country + total_games + wins + draws + losses, train)
```

```r
#make samples
tree_samples <- vfold_cv(train)

#create grid
tree_grid <- grid_regular(
  cost_complexity(),
  tree_depth(),
  levels = 5
)

#make workflow
online_workflow <- workflow() %>%
  add_recipe(tree_recipe) %>%
  add_model(tree_model)

#do the tuning
tree_res <- online_workflow %>%
  tune_grid(
    resamples = tree_samples,
    grid = tree_grid
  )

#find best collection
online_best_tree <- tree_res %>%
  select_best("rmse")

#finalize workflow
online_final_wf <- online_workflow %>%
  finalize_workflow(online_best_tree)

#metrics of tuned model
online_final_wf %>%
  last_fit(split) %>%
  collect_metrics()
```

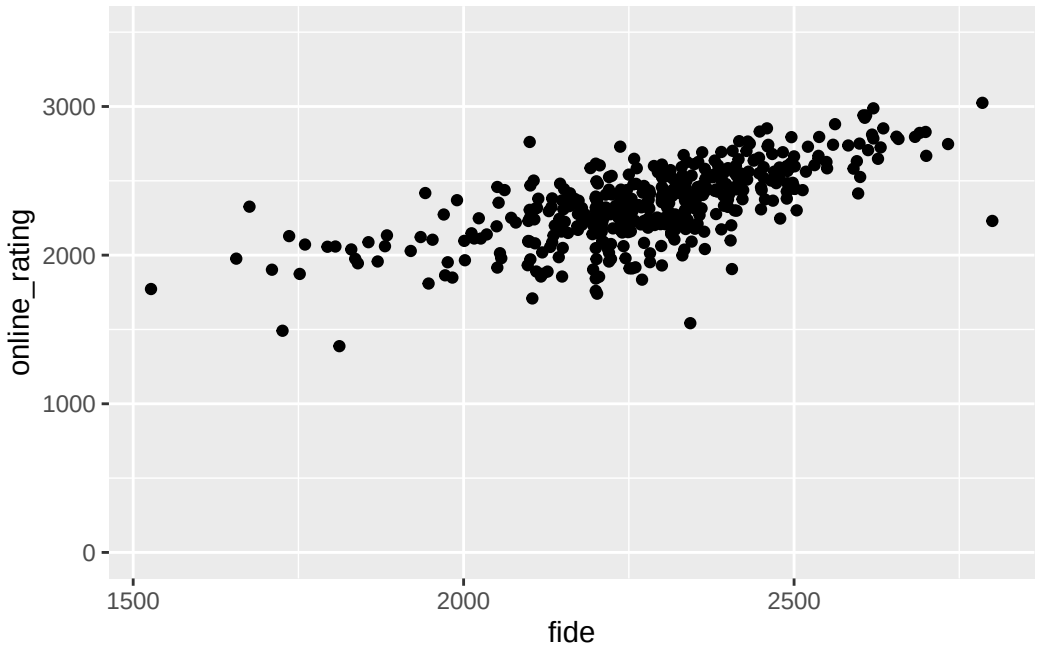
```
# A tibble: 2 x 4
  .metric .estimator .estimate .config
  <chr>   <chr>          <dbl> <chr>
1 rmse    standard     198.    Preprocessor1_Model1
2 rsq     standard       0.498 Preprocessor1_Model1
```

From the regression decision tree, we see slightly better results than the linear regression model

with a lower rmse and a slightly higher R-squared.

Now we are going to shift gears a bit and look at how factors correlate with FIDE rating. We are going to use a slightly more slimmed down version of the data.

```
player <- players %>%
  filter(followers < 100000) %>%
        select(c("title","fide","country","joined","last_online","followers",
                 "league","online_rating","wins","losses","draws","tactics_rating",
                 "puzzle_best"))

# player %>%
#   ggplot(aes(x=fide,y=bullet_last)) + geom_point() + ylim(max=c(0,3500))
#
# player %>%
#   ggplot(aes(x=fide,y=blitz_last)) + geom_point() + ylim(max=c(0,3500))
#
# player %>%
#   ggplot(aes(x=fide,y=rapid_last)) + geom_point() + ylim(max=c(0,3500))

player %>%
   ggplot(aes(x=fide,y=online_rating)) + geom_point() + ylim(max=c(0,3500))
```

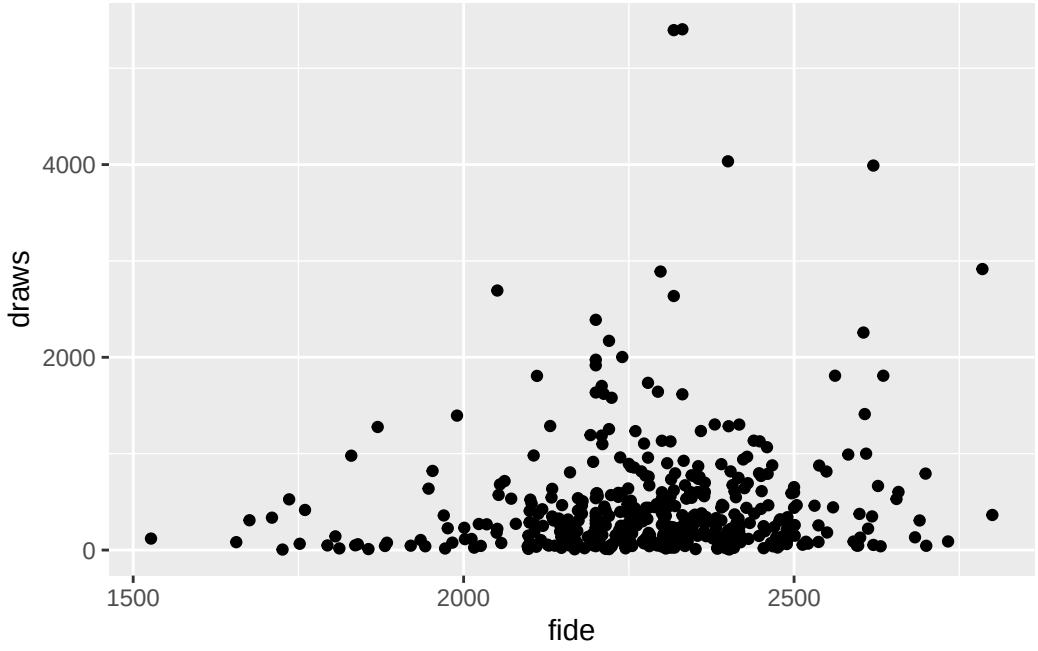


This portion of the exploratory analysis shows that bullet, blitz, and rapid ratings seem to be

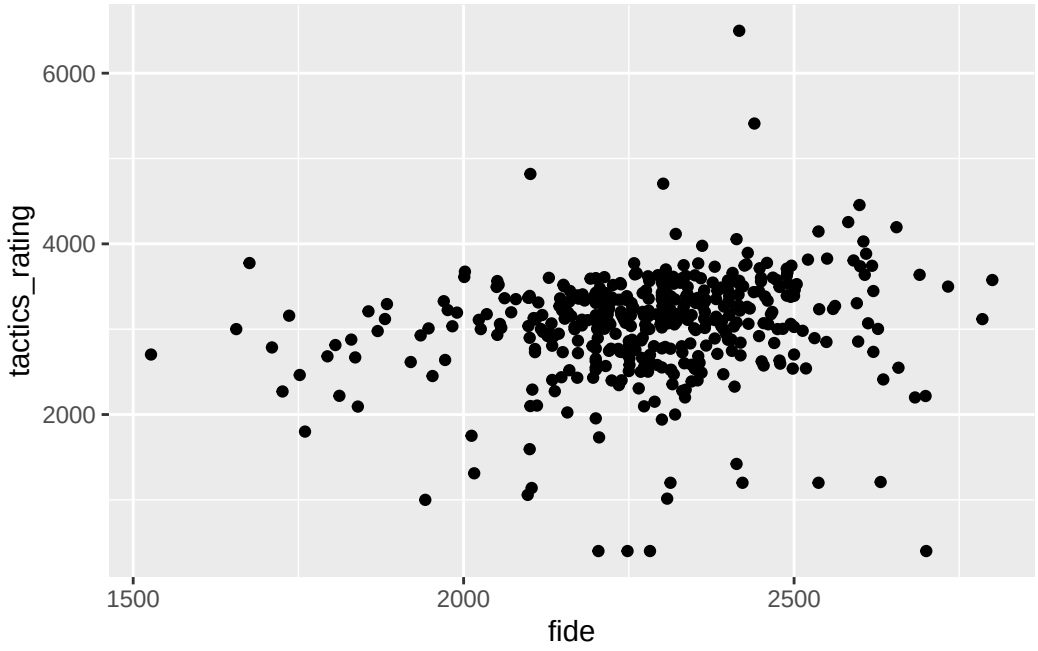
linearly correlated in a slightly positive direction. The average graph also has less variation than the other 3 as we would expect. This hints at a possibility they will correlate fairly well with FIDE rating.

Other factors we will look at include country, followers, league, wins, losses, draws, tactics_rating, and puzzle_best. I will again graph all of them in my own analysis but will only show the most important graphs for simplicity of this project.

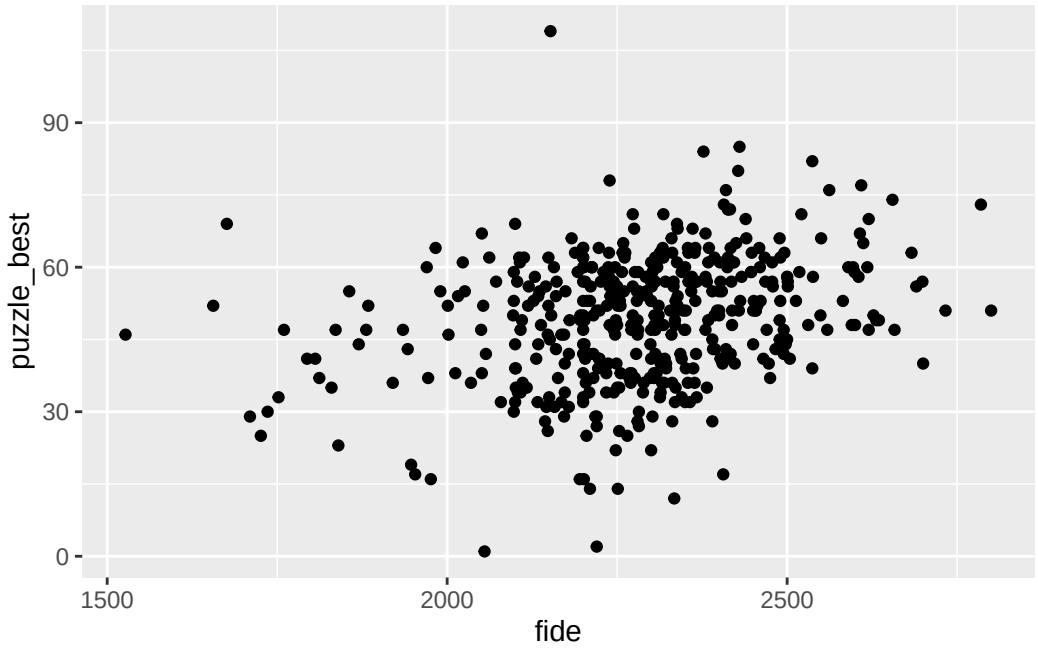
```
# followers, country, league, wins, losses
# player %>%
#   group_by(country) %>%
#   summarise(median = median(fide), mean = mean(fide)) %>% arrange(desc(median))
# player %>%
#   group_by(league) %>%
#   summarise(median = median(fide), mean = mean(fide)) %>% arrange(desc(median))
# player %>%
#    ggplot(aes(x=fide,y=wins)) + geom_point()
# player %>%
#    ggplot(aes(x=fide,y=losses)) + geom_point()
# player %>%
#    ggplot(aes(x=fide,y=followers)) + geom_point() + ylim(max=c(0,20000))

# draws, tactics_rating, puzzles_best
player %>%
   ggplot(aes(x=fide,y=draws)) + geom_point()
```




```
player %>%
   ggplot(aes(x=fide,y=tactics_rating)) + geom_point()
```

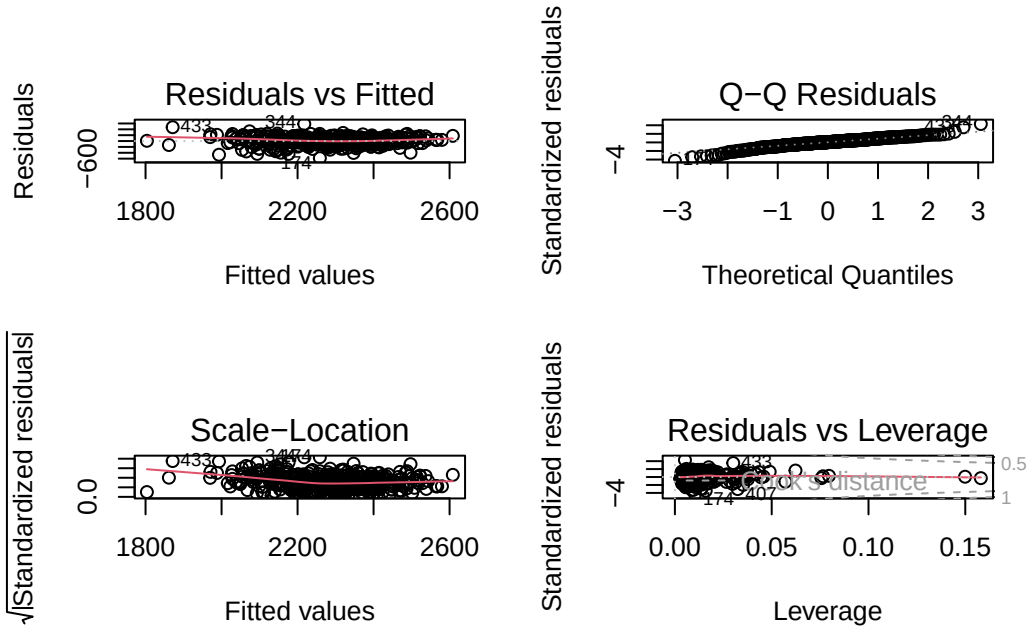


```
player %>%
   ggplot(aes(x=fide,y=puzzle_best)) + geom_point()
```



Country, league, wins, and losses appear to be too scattered to be useful for our models. Number of followers is vaguely correlated. The number of draws, tactics_ratings, and puzzle_best seemed to be somewhat correlated, but those are the best three so those are the ones we are going to use along with online_rating to predict FIDE rating.

```
model <- lm(fide ~ online_rating + draws + tactics_rating + puzzle_best, data = player)
par(mfrow = c(2, 2))
plot(model)
```




```
summary(model)
```

```
Call:
lm(formula = fide ~ online_rating + draws + tactics_rating +
    puzzle_best, data = player)

Residuals:
    Min      1Q  Median      3Q     Max
-582.68  -70.38    8.35   80.13  582.29

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.123e+03  6.327e+01  17.748   <2e-16 ***
online_rating   5.179e-01  3.011e-02  17.202   <2e-16 ***
draws          -6.647e-03  1.057e-02  -0.629    0.530
tactics_rating -1.055e-02  1.136e-02  -0.929    0.353
puzzle_best    -3.890e-01  5.704e-01  -0.682    0.496
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136 on 435 degrees of freedom
```

```
Multiple R-squared:  0.4624,    Adjusted R-squared:  0.4574
F-statistic: 93.53 on 4 and 435 DF,  p-value: < 2.2e-16
```

```
# model2 <- lm(fide ~ online_rating + draws + tactics_rating +
#              puzzle_best + followers, data = player_online)
# plot(model)
# summary(model)
```

The assumptions for a linear model appear to be sufficient. All of the assumption graphs don't appear to produce any anomalies or patterns. Our data produced an R-squared of .45 which is okay, but we were expecting to get a model with an R-squared closer to .8 but unfortunately that wasn't the case. I did not graph it, but adding in other factors particularity followers did not increase the R-squared value.

I decided to use a decision tree model to cross validate the data. If they produce a similar r-squared, then that would lend support to the linear model's conclusion because the r-squared while not the best metric for comparison implies that there isn't much room for improvement. If it differs greatly, then more analysis is required. We will predict FIDE rating using all variables and one using just the ones we used in linear regression.

```
decision_tree_test <- function(data,testVar) {

data <- data %>% mutate(testVar = testVar)
data_split <- initial_split(data)
train_data <- training(data_split)
test <- testing(data_split)

d_recipe <- recipe(testVar ~ ., train_data)
d_samples <- vfold_cv(train_data)

# set model
d_model <- decision_tree(mode = "regression",
                         cost_complexity = tune(),
                         tree_depth = tune()) %>%
    set_engine("rpart")

# workflow
d_wf <- workflow() %>%
    add_recipe(d_recipe) %>%
    add_model(d_model)

# create grid
```

```r
  d_tree_grid <- grid_regular(cost_complexity(),
                              tree_depth(),
                              levels = 5)

  # does the tuning (takes some time to run)
  d_tree_res <- d_wf %>%
      tune_grid(
          resamples = d_samples,
          grid = d_tree_grid
      )

  # find the best collection of cost_complexity and tree_depth
  d_best_tree <- d_tree_res %>%
      select_best("rmse")

  # finalize the workflow
  d_final_wf <- d_wf %>%
      finalize_workflow(d_best_tree)

  # metrics of the tuned model
  d_final_wf %>%
      last_fit(data_split) %>%
      collect_metrics()

  # predictions of tuned model
  pred <- d_final_wf %>%
      last_fit(data_split) %>%
      collect_predictions()

  # metrics
  pred %>% metrics(testVar, .pred)
  }

  player_fide <- player %>% select(c("fide","online_rating","draws",
                                     "tactics_rating","puzzle_best"))

  decision_tree_test(player %>% select(-c("fide")),player$fide)
```

```
# A tibble: 3 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
```

```
1 rmse    standard     125.
2 rsq     standard       0.551
3 mae     standard      90.8
```

```
decision_tree_test(player_fide %>% select(-c("fide")),player$fide)
```

```
# A tibble: 3 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard     135.
2 rsq     standard       0.410
3 mae     standard     108.
```

Conclusion and Analysis:

No assumptions are needed for a decision tree so that is covered.

The results produced an rmse around 100-150 and an R-squared around .45 for both models. This means additional factors did not improve the model. These results are close to identical to the linear regression results. This indicates that FIDE rating is somewhat correlated with the factors on chess.com, particularly online_rating, draws, tactics_rating, and puzzle_best. This is a bit unexpected as were predicting the R-square to be close to .8.

The reason for this less than optimal outcome could be the result of different time controls. FIDE games use much longer time controls and so faster games online may result in diverse outcomes. Another reason may be the environment. Playing online is a much different environment than playing over the board which could heavily influence outcomes. Other factors such as wins/draws/losses, tactics, puzzles, etc weren't strongly correlated either. Again, maybe those players didn't invest as much effort into playing a lot of games online or didn't try tactics or puzzles as much relative to their FIDE rating. It simply could be the strong FIDE players are investing greatly varying amounts of time to online chess subsequently producing varying results.

When we looked at Online rating in relation to other factors, we got similar results. So to answer the quesion: "How well do bullet, blitz, rapid, and FIDE ratings of titled players correlated with other factors and each other on chess.com." the answer is somewhat based on our models.

Improvements and limits of the model: It is important to keep in mind some factors like number of draws shouldn't necessarily correlate with online or FIDE rating, so some of the variables themselves shouldn't be expected to have an strong correlation. This is one of the limits of the data we obtained. The models could be improved if we had access to more chess.com data such as chess openings that each players uses, accuracy percentage, and more.

Those might be better factors to look at since they would shed more light on a player's abilities which could produce more useful models. Another thing is if we looked at data in terms of relative percentages such as win percentage, then maybe we would would have seen different outcomes.

Reflection:

(Nathan):

I found the variability of ratings of titled players to be really interesting. It was a lot more spread out than I anticipated. I was also surprsied to see that the number of games wasn't very significant in determining rating. My thought was that the more games you play the more likely you have a higher rating, but that wasn't the case.

Thinking about data science as a whole, this project helped me internalize what EDA means and what the process is like. I feel like I had pretty good intuition about the data set because I'm so involved in the chess world, but I can definitely understand how valuable it and necessary it would be if you're exploring a data set that about a topic that is more foreign to you, and it was already valuable already knowing a lot about the subject.

(Luke):

Overall, I would say this was a very good learning project. One of the major components was web scrapping the data as this was the most work intensive part. It taught me a lot about data collection. The actual analysis was also pretty interesting. We thought that if players did well in FIDE rating tournaments, then they would likely excel in the online chess environment. This is not necessarily the case and even factors on chess.com varied with online rating. Regardless, I think the experiment was successful considering both the linear and decision tree models supported each other. Maybe in the future we can expand upon it.