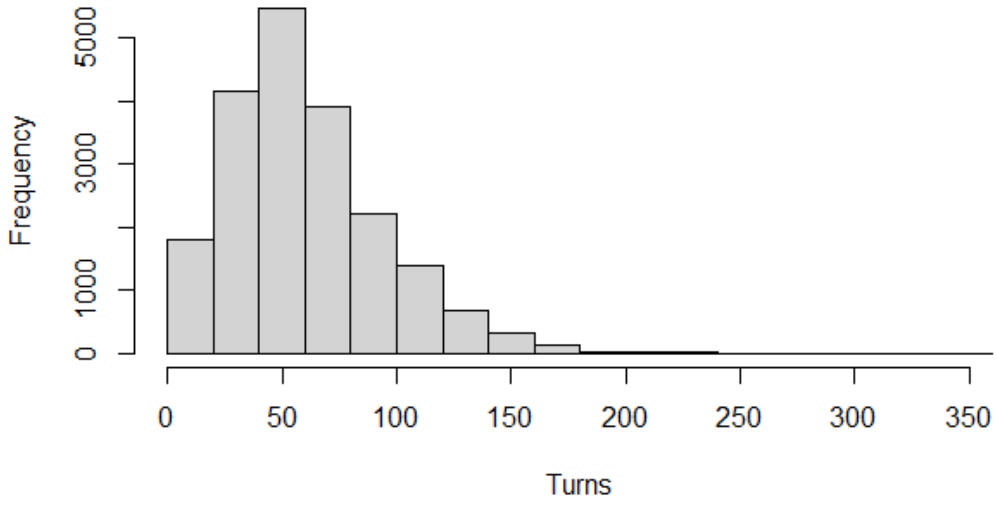
Turns Taken in Chess

Dataset and Guiding Question

For my data project I investigated factors that influence the number of turns played in a game of chess. I based my analysis on dataset of 20,000 LiChess games on Kaggle.com which contained information on the number of turns played in each game, the game's outcome, the time control in which the game was played, the ratings of both players, and whether or not the game was rated. Detailed information was also provided on the move order and opening Eco, which is a Standardized Code of chess openings, but an analysis of the effect of opening choice on the total number of turns in a game is beyond the scope of this analysis.

At the outset of my analysis, I identified six potential predictor variables for the number of turns in a game: Increment per turn in seconds, time per player in minutes, outcome of whether the game was decisive or drawn, whether the game was rated, the average rating of the two participating players, and rating difference between the two players. I analyzed the predictors in that same order, as I hypothesized that the increment value would be the strongest predictor of turn number since games with low or no increment are more likely to end in a sudden blunder or loss on time, and that the format played would be the second strongest predictor for similar reasons. For outcome I expected that drawn games would last more turns overall than decisive ones, and regarding whether the game was rated, I expected rated games to last more turns on average as players stay longer to defend unfavorable positions. Finally, I expected the number of turns would increase as the average rating of players in the game increased but decrease as the rating differential between the players increased.

Figure 1: Turns in Chess



Data Tidying

Creating the variables for my analysis required some data tidying and transformation. The only entry that I used from the initial data entirely unmodified was the rated column, a Boolean variable noting whether or not the game played was formally rated.

Time Control: The games were played using an increment time format, meaning that both players began each game with an equal number of minutes on their clock, and every time each played a move a number of seconds equal to the increment value was added to that players' total time available.

The dataset contained a column called increment_code which contained the time control of each game played as a character variable with format X + Y, where X is the number of minutes given to each player at the start of the game and Y is the amount of seconds that is added to their total time after each move. I split this column into two numeric columns, one containing the time X in minutes, the other containing the increment Y in seconds, and created histograms for both of the new variables, showing first the distribution across all games, then zoomed in on games played with 60 minutes or less and 60 or fewer seconds of increment to get a better view of the distribution of each. Figures 2 and 3 show that the vast majority of games were played in formats with a time control of 10 + 10 or faster.

Figure 2: Time Controls

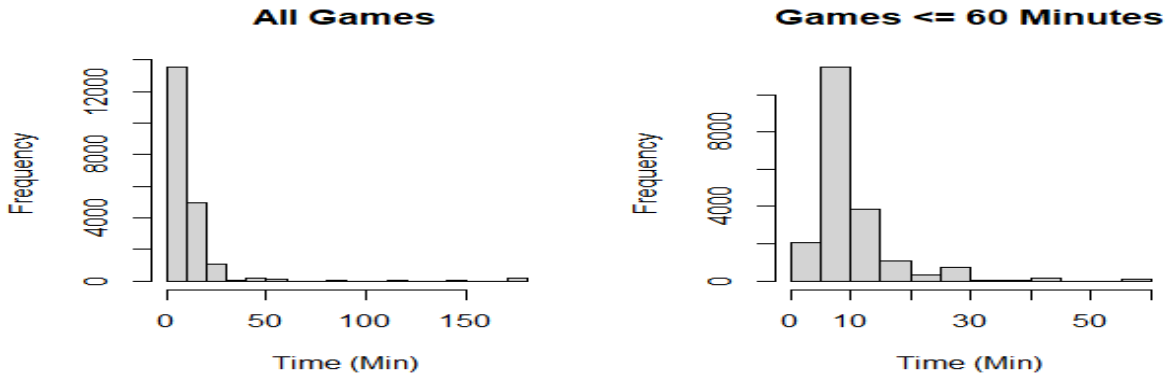
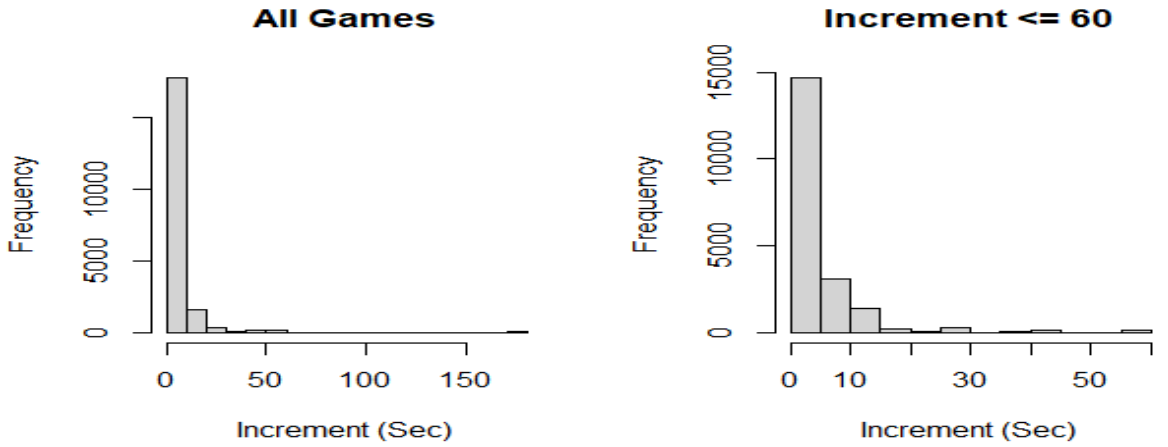
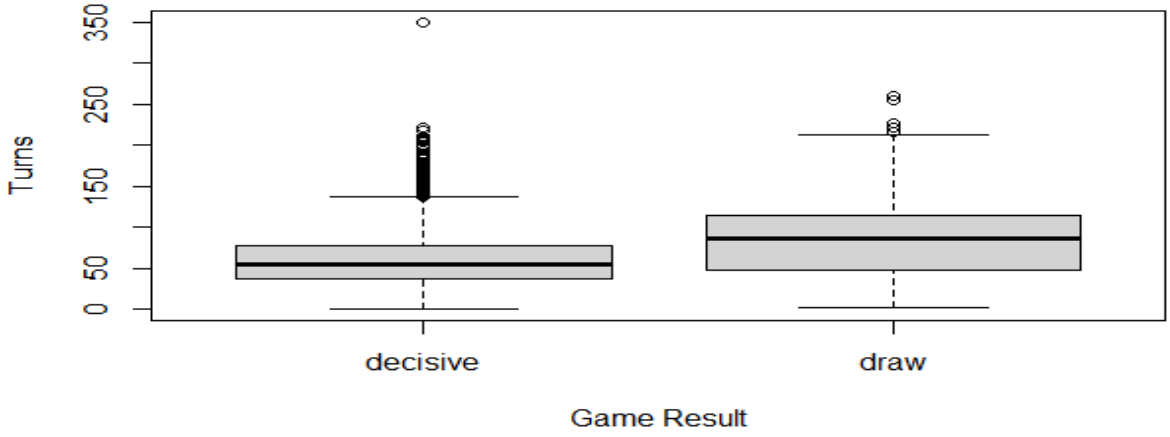


Figure 3: Increment



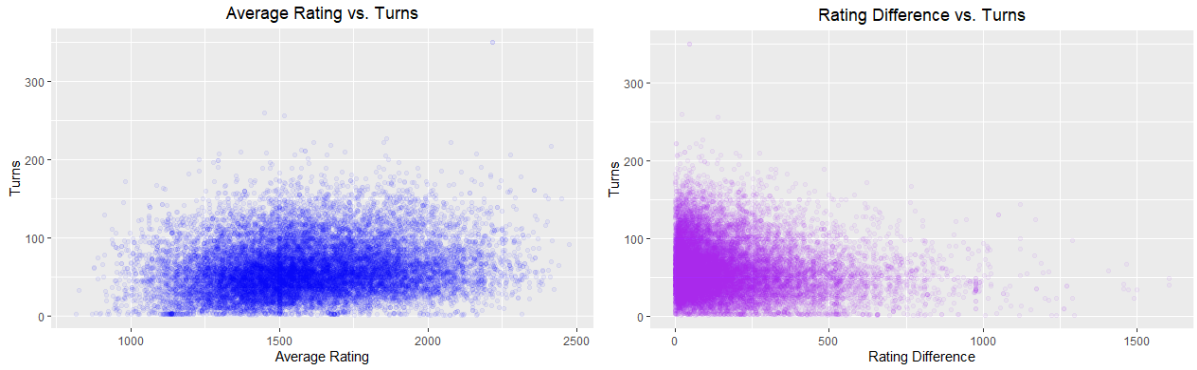
Game Outcome: In the data, the game result was recorded as a variable victory_status, with four possible outcomes – Draw, mate, out of time, and resign. The latter three results are all decisive results of 1-0 or 0-1 (a win for white or a win for black, respectively), so I combined them into a single outcome decisive and created a box blot comparing the distribution of turns for these decisive games compared against games that ended in a draw. As Figure 4 shows, the median and average range of turns played in drawn games is substantially higher than in decisive games. I removed the outlying 350 move decisive game from my analysis to prevent it from potentially biasing the results.

Figure 4: Distribution of Turns by Outcome



The dataset provided the rating of both players in each game, so in order to simplify my analysis of ratings I created two new variables, one for the average rating of the two players in the game, the other for the rating difference between the two players. For the latter variable I calculated the difference in absolute value as what matters the magnitude of the difference between the players, not which player had which color. I created scatterplots comparing both of these variables to the number of turns in a game to look for potential trends in the data. There does not appear to be any obvious pattern between the average rating of the players in a game and the number of turns, but for rating difference the number of turns appears highest when the two players are closely rated to one another and turns downwards as the rating disparity increases.

Figure 5: Average Rating and Rating Difference vs. Turns



Summary Statistics

Variable	Mean	Median	Variance	Standard Dev.	Range
x_1 = increment	5.25015	0	204.21	14.29021	0 – 180
x_2 = time_min	13.82449	10	294.50	17.16	0 – 180
x_3 = outcome	0.9548265	1	0.04313502	0.2076897	0 – 1
x_4 = rated	.8053949	1	0.1567418	0.3959	0 – 1
x_5 = avg_rating	1592.67	1568.5	69228.5	263.11	816.5 – 2475.5
x_6 = rating_diff	173.1041	115	32119.56	179.219	0 – 1605

Covariance Matrix

```
             increment     time_min      outcome         rated
increment   204.2100224  114.7520083 -0.122381674  -0.555804846
time_min    114.7520083  294.4996365 -0.148043419  -0.714236349
outcome      -0.1223817   -0.1480434  0.043135019   0.002427713
rated        -0.5558048   -0.7142363  0.002427713   0.156741774
avg_rating  -97.5721396 -368.6204809 -2.576303891   2.058650769
rating_diff 189.5170126  167.1290183  1.094954989 -15.580437025
turns       -28.6514865  -34.0538003 -1.054593540   1.199924815
               avg_rating  rating_diff        turns
increment     -97.572140   189.517013  -28.651486
time_min     -368.620481   167.129018  -34.053800
outcome        -2.576304     1.094955   -1.054594
rated           2.058651   -15.580437    1.199925
avg_rating  69228.501401  2574.711729 1400.408832
rating_diff  2574.711729 32119.555952 -757.672148
turns        1400.408832  -757.672148 1118.793367
```

Analysis

With my variables selected and prepared in R, I began to work through six models, adding a new predictor variable to each model while keeping the rest in place. For the first model I used = increment as the predictor variable in the equation simple linear regression:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

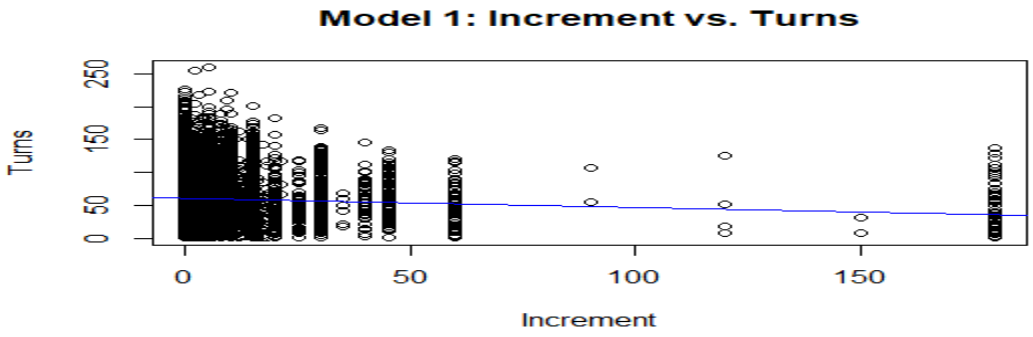
During each subsequent model I added new variables in the order above. Each variable included added a new parameter to the equation for multiple linear regression until I reached the equation below for linear regression using six predictor variables.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6$$

Figure 6: Linear Model Fitted Values

	increment		time_min		outcome (draw = 0)		rated (rated = 1)		avg_rating		Rating Difference		R2	F Significance
	coefficient	*p value*	*coefficient*	*p value*	*coefficient*	*p value*	*coefficient*	*p value*	*coefficient*	*p value*	*coefficient*	*p value*		
model 1	-0.1403	0.00											0.003543	0.000
model 2	-0.0964	0.00	-0.0781	0.00									0.004746	0.000
model 3	-0.1066	0.00	-0.0868	0.00	-25.0486	0.00							0.02883	0.000
model 4	-0.0939	0.00	-0.0739	0.00	-25.3840	0.00	7.3788	0.00					0.0363	0.000
model 5	-0.0991	0.00	-0.0483	0.00	-24.1800	0.00	7.2130	0.00	0.0187	0.00			0.05773	0.000
model 6	-0.0862	0.00	-0.0448	0.00	-23.4300	0.00	5.1510	0.00	0.0196	0.00	-0.0211	0.00	0.06978	0.000

Model 1: I started with a simple analysis of the effect that increment has on the number of turns played in a game. The fitted coefficient is negative and statistically significant with a p-value of 2E-16, but the R-squared value of 0.003543 shows that increment explains very little about the number of turns played in a game.



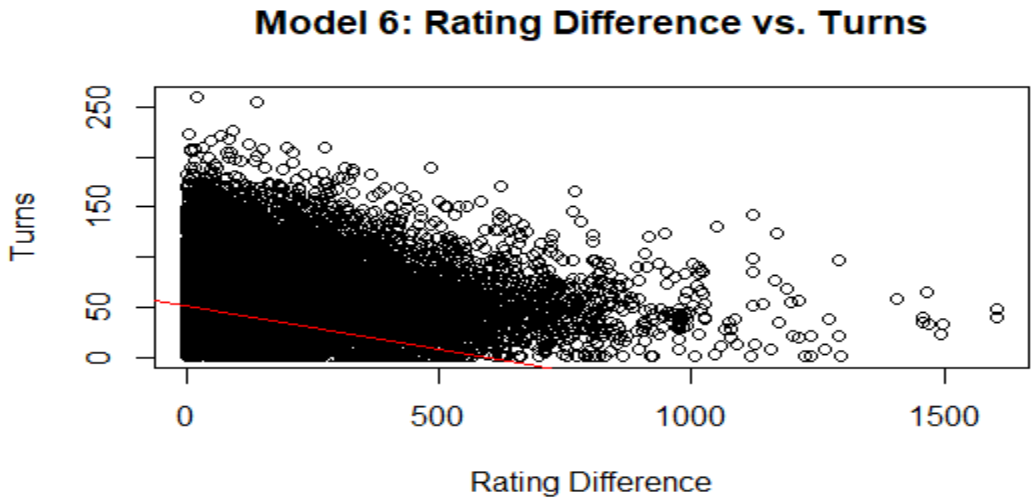
Model 2: In this model I was still investigating only the impact of time control on the number of moves played by adding the time_min variable to the regression model. In both this model and that preceding the fitted parameters for both increment and starting time have p-values indicating statistical significance, but both parameters are very small as is the R-Squared value of 0.004746 showing that the time control of a game has very little impact on the number of moves played.

Model 3: For this model I added a Boolean variable for the outcome of the game, with 0 indicating a drawn game and 1 indicating a decisive game. The fitted coefficient for outcome of -25.05 suggests that there is a strong relationship between a game's result and the number of moves played, and that decisive games end in significantly fewer moves than do drawn games. The inclusion bumped up the R-squared almost an order a magnitude but still indicates that most of the variance in the number of turns played is unexplained by the variables included in this first three models.

Model 4: I added another Boolean variable to the model, this one indicating whether or not the game was played in one of LiChess's rated formats, and the resulting coefficients show that rated games do tend to last significantly more moves than unrated games. Per the slight increase in the R-squared value the variables I've added so far have all improved the model's expected fit to the data, but still does little to explain the response variable from this analysis.

Model 5: In this model I added a variable indicating the average rating of the two players in the game. The fitted model shows that this particular variable has very little impact on turns played with a coefficient of almost 0, which makes me question the improved R-squared value between models here.

Model 6: For the last model of this analysis I included a variable indicating the difference between the ratings of the two players, and found that there is a slight negative trend – as the rating difference increases, the average number of turns played slowly decreases. With all six predictor variables included, this final model had an R-squared value of 0.06978, so the model still leaves a great deal unexplained regarding the estimated number of moves played in a game of chess.



Conclusions and Next Steps

In every model, all fitted coefficients had statistically significant p-values, and new variables added in the later models did not change the trend of any of the variables, meaning that the magnitude and sign of the coefficients did not significantly change as the model was updated. This suggests to me that the model can weakly predict the impact that each variable will have on the expected number of turns played, but the small R-Squared value of even the final model indicates that there is very likely further analysis that may result in a more accurate prediction of turns played in a game. From the covariance matrix I can also see some potential for collinearity between the increment and time_min variables, which is unsurprising since I split a single variable for the full time control into somewhat related two variables for ease of analysis, and there may also be concerns of collinearity between the average ratings and ratings difference variables since a significant difference between the ratings of the two players can only occur if at least one of them is highly rated.

The dataset that I found for this project included only games from the online platform LiChess, so I can think of several ways to find new sources of data that might be helpful in further exploration this question, as well as possible questions for deeper analysis. Regarding data sources, games pulled from another platform such as Chess.com might be helpful in checking for variations in game time controls, formats played, and average turns played. Even more interesting to analyze would be games played in-person over a physical board, especially tournament games, but it is much more difficult to gather significant data sets from the games that occur in such events.

My starting data set also contained full information on the moves played in each game, along with a note of the corresponding Opening ECO (a catalogue of chess openings). The catalogue has a huge amount of variation even within the first few moves so I opted to omit it and any specific analysis of played variations from this project, but it would be very interesting to take a look at the effect that the opening choice has on a game's expected duration in turns, or even deeper questions such as game length when one, both, or neither players castles, or other questions that dig into the nitty-gritty details of gameplay.