Group S Final Report

Sarah Fatihi, Sophie Diop, James Street

2024-05-01

Abstract

We want to study International Federation of Association Football (FIFA) Stats in comparison to their in game valuation. Our goal is to predict in game market value from the FIFA video game stats. We looked at correlations between FIFA video game predictors including pace, shooting, passing, dribbling, defense, and physicality and in-game transfer market value. We found that pace, shooting, passing, and physicality were significant quantitative predictors of transfer market value. The models accounted for a lot of the variability in transfer market value, although log transformations were used to make the data more linear.

Background and Meaning

Originally, we attempted to try and analyse in-game statistics in relation to real world transfer prices, but found that due to the complexity of the transfer market (cyclical economics, club relationships, regulations) the in-game variables held very little significance. In the FIFA game however, a player's transfer market value holds immense significance for the gamer, reflecting their monetary worth in the in-game transfer market. This value is determined by the developers of the game, and are meant to reflect a players worth based on their skill level and statistics. Our project seeks to explore the relationship between FIFA video game stats and in-game transfer market value. Specifically, we aim to assess the accuracy of FIFA game stats in predicting in-game transfer market value and identify the best and worst predictors of a player's market worth. FIFA game stats, which use attributes such as pace, shooting, passing, dribbling, defense, and physicality, serve as key indicators of a player's ability and performance in the virtual realm of FIFA video games. However, how far these statistics influence their in-game value set by developers is unclear. To achieve our objective, we employed a multiple linear regression model to test the null hypothesis that there is no correlation between FIFA video game stats and transfer market value. By analyzing the relationship between these variables, we sought to determine the extent to which FIFA video game stats accurately predict a player's in-game market worth. This approach allowed us to assess how accurate FIFA video game stats are on predicting transfer market value and identify any significant predictors.

Methods

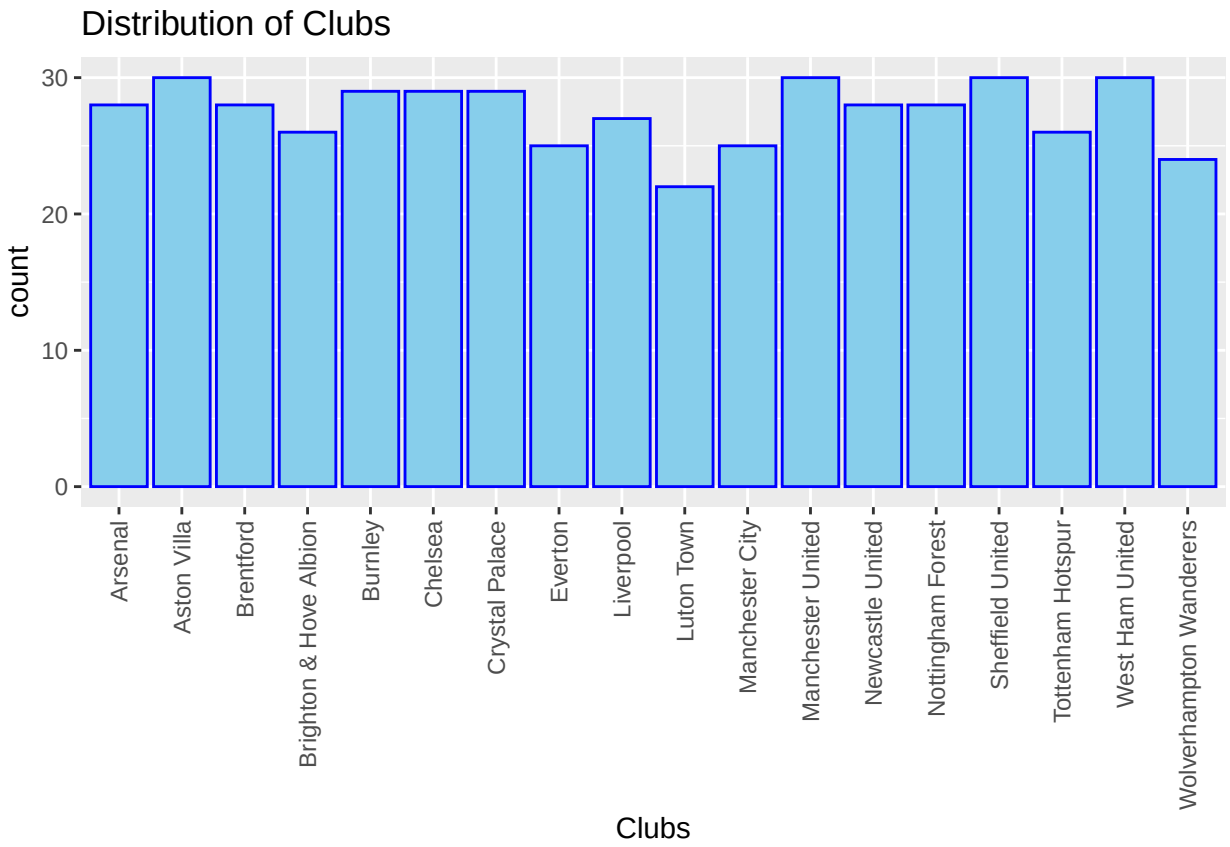
Data The FIFA Football Players dataset, https://www.kaggle.com/datasets/rehandl23/fifa-24-player-stats-dataset, is a comprehensive collection of information about football (soccer) players from around the world. This dataset offers a wealth of attributes related to each player, making it a valuable resource for various analyses and insights into the realm of football, both for gaming enthusiasts and real-world sports enthusiasts. The data is based on the game FIFA 24 which contains information on soccer players from over 19,000 fully licensed players, 700 teams, and 30 leagues in the year 2023-2024. We are planning on slimming down this data to only include players in the premier league. The Premier League is the top professional football (soccer) league in England. It consists of 20 teams, and it's widely regarded as one of the most competitive and popular football leagues in the world.

Variables: We have both categorical and quantitative variables as our explanatory variables. These are purchasing club (categorical), in-game Passing (quantitative), in-game Shooting (quantitative), in-game Dribbling (quantitative), in-game Defending (quantitative), in-game Physicality (quantitative), in-game Pace (quantitative). All of our quantitative in-game stats are on a scale of 0-100 determined by the people working at FIFA. Below is a list of the explanatory variables we intend on using: Club - This predictor is the club that the soccer player plays on. The premiere league contains 20 different clubs. These include, Aresenal, Aston Villa, Bournemouth, Brentford, Brighton, Burnley, Chelsea, Crystal Palace, Everton, Fulham, Liverpool, Luton Town, Man. City, Manchester Utd, Newcastle, Nottingham, Sheffield Utd, Tottenham, West Ham, and Wolves. Short Passing - This is a measure at how good the soccer player is at passing in real time performance on a scale of 0-100 determined by the people working at FIFA. Shooting - This is a measure at how good the soccer player's shot is in real time performance on a scale of 0-100 determined by the people working at FIFA. Dribbling - This is a measure at how good the soccer player is at dribbling in real time performance on a scale of 0-100 determined by the people working at FIFA. Defending - This is a measure at how good the soccer player is at Defending in real time performance on a scale of 0-100 determined by the people working at FIFA. Physicality - This is a measure at how physical the soccer player is in real time performance on a scale of 0-100 determined by the people working at FIFA. Pace - This is a measure at how fast the soccer player is in real time performance on a scale of 0-100 determined by the people working at FIFA.

Statistical Methods: We checked conditions for each model we created using a Fitted vs. Residuals plot to check for linearity and equal variance. We also used the QQ plot to check for normality. Our original data did not meet conditions so we used the log transformation of transfer value to see if this would improve the linearity of our data and meet conditions. We found that this worked and the Fitted vs Residuals plot as well as the QQ plot looked much better and met conditions. Since all players from the Premier League are included in the data set, there isn't a process of random selection involved. Instead, our analysis is focused on the population of players within the Premier League, and we are making inferences specifically about that population rather than generalizing to a larger population or making causal claims. To conduct bivariate analysis for the quantitative variables, we created scatterplots of them as a predictor of in-game transfer value. For the bivariate analysis for our categorical variables, we created box plots with each variable as a predictor of in-game transfer value. Additionally, we looked at the variation inflation factor (VIF) for each predictor variable to check for multicollinearity. We ended up with one final model which was a multiple linear regression model. We tested for significance of each model as a predictor of household income by using F-statistic and associated p-values. Based on this information, we determined which predictor variables were significant in each model. We also used the best subsets method to evaluate different combinations of predictor variables to determine the. Most effective set for predicting in-game transfer value. Through this, we landed on one final model that we found was best at predicting in-game transfer value.

Results

Univariate Descriptive Stats and Figures



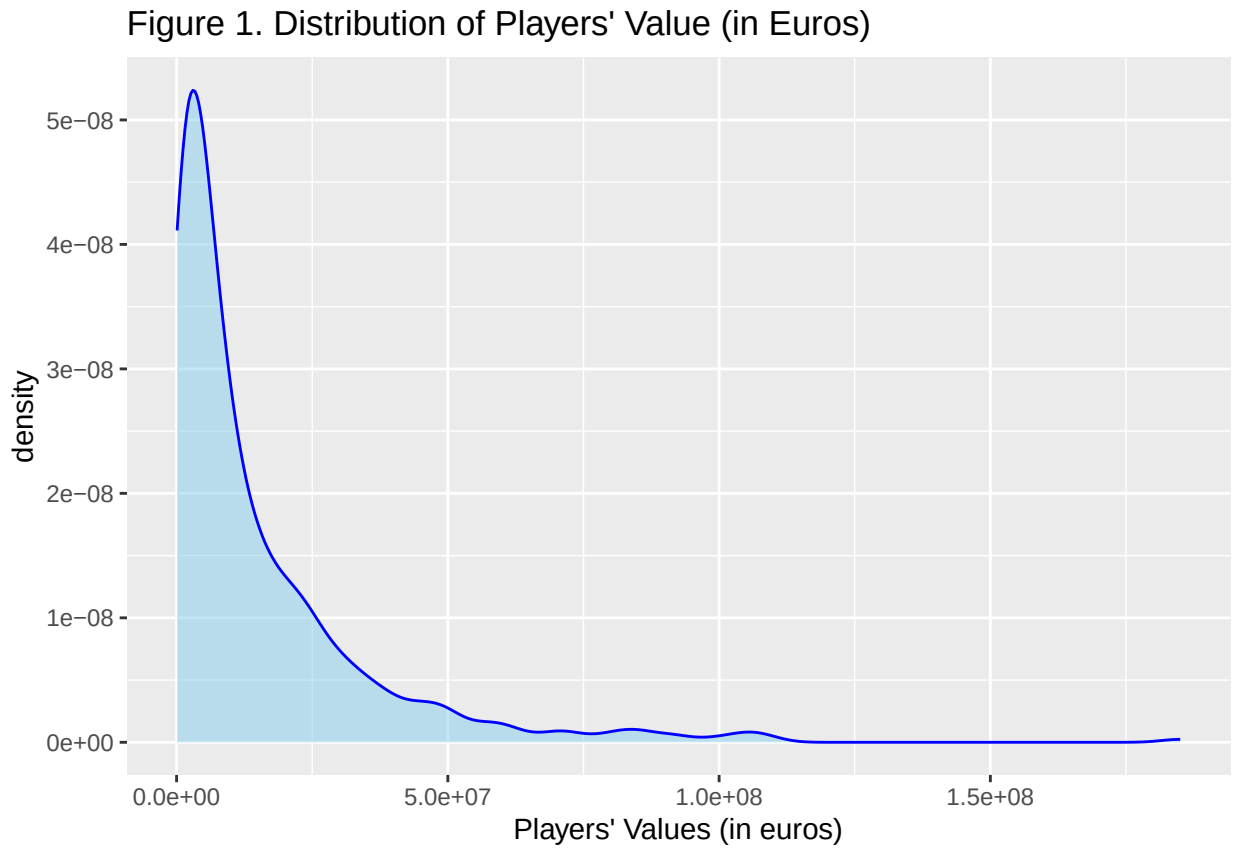


Figure 1. The density plot for players' value, our response variable, is unimodal and has a strong right skew. The median of value is €7,750,000 and the IQR is €17,500,000. Financial data typically has a strong skew, which is consistent with our distribution.

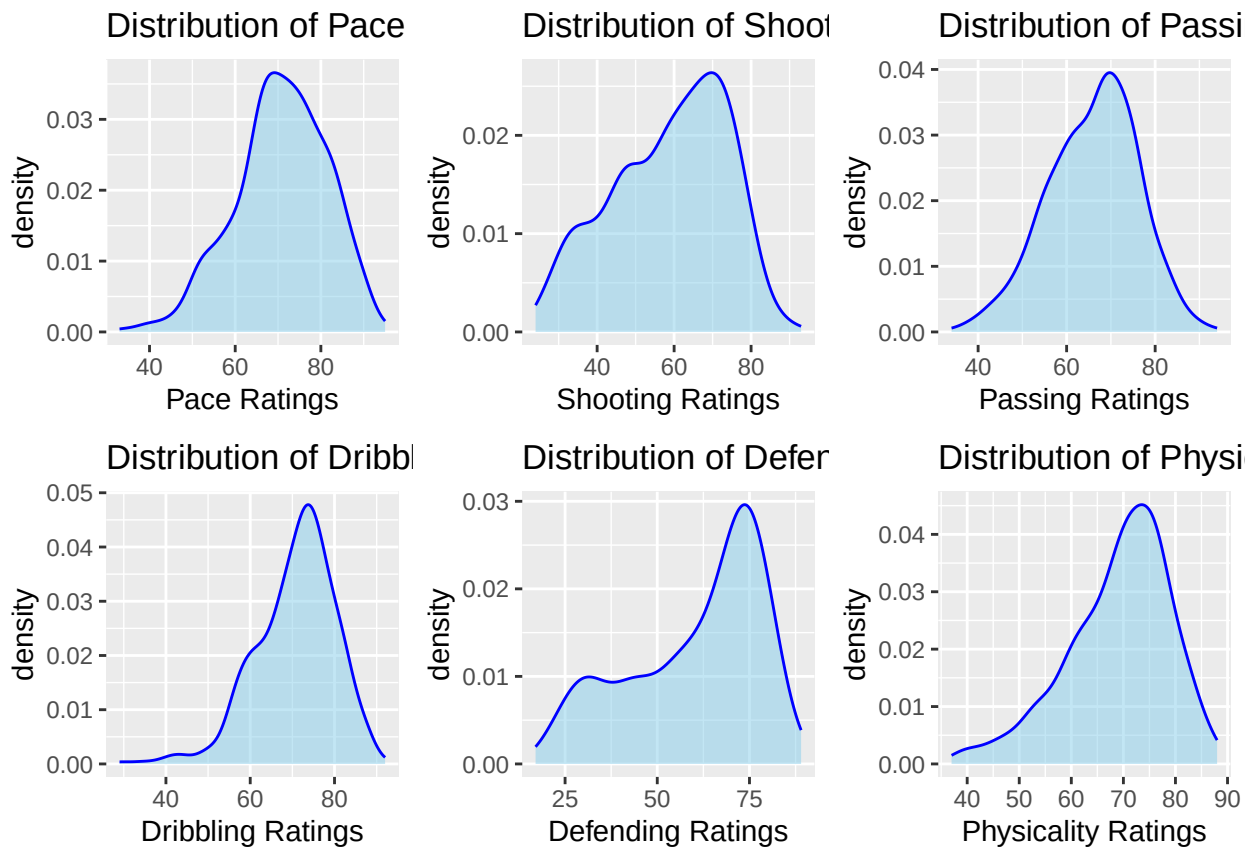


Figure 2. For our univariate analysis we ran density plots to see the distribution of each of of our quantitative explanatory variables.

Pace: The ratings of players' pace is normally distributed, so we use the mean of 70.75 as a measure of center, and the standard deviation of 10.73 as a measure of spread.

Shooting: The ratings of players' shooting is normally distributed, so we use the mean of 58.7206 as a measure of center, and the standard deviation of 14.87 as a measure of spread.

Passing: The ratings of players' passing is normally distributed, so we use the mean of 65.92 as a measure of center, and the standard deviation of 10.16 as a measure of spread. Dribbling: The ratings of players' dribbling is normally distributed, so we use the mean of 70.77 as a measure of center, and the standard deviation of 9.42 as a measure of spread.

Defending: The ratings of players' defending is left-skewed with a potential second peak, so we use the median of 66 as a measure of center, and the IQR of 29 (first quartile = 46, third quartile = 75) as a measure of spread.

Physicality: The ratings of players' physicality is normally distributed, so we use the mean of 69 as a measure of center, and the standard deviation of 9.89 as a measure of spread.

Bivaraite Analysis

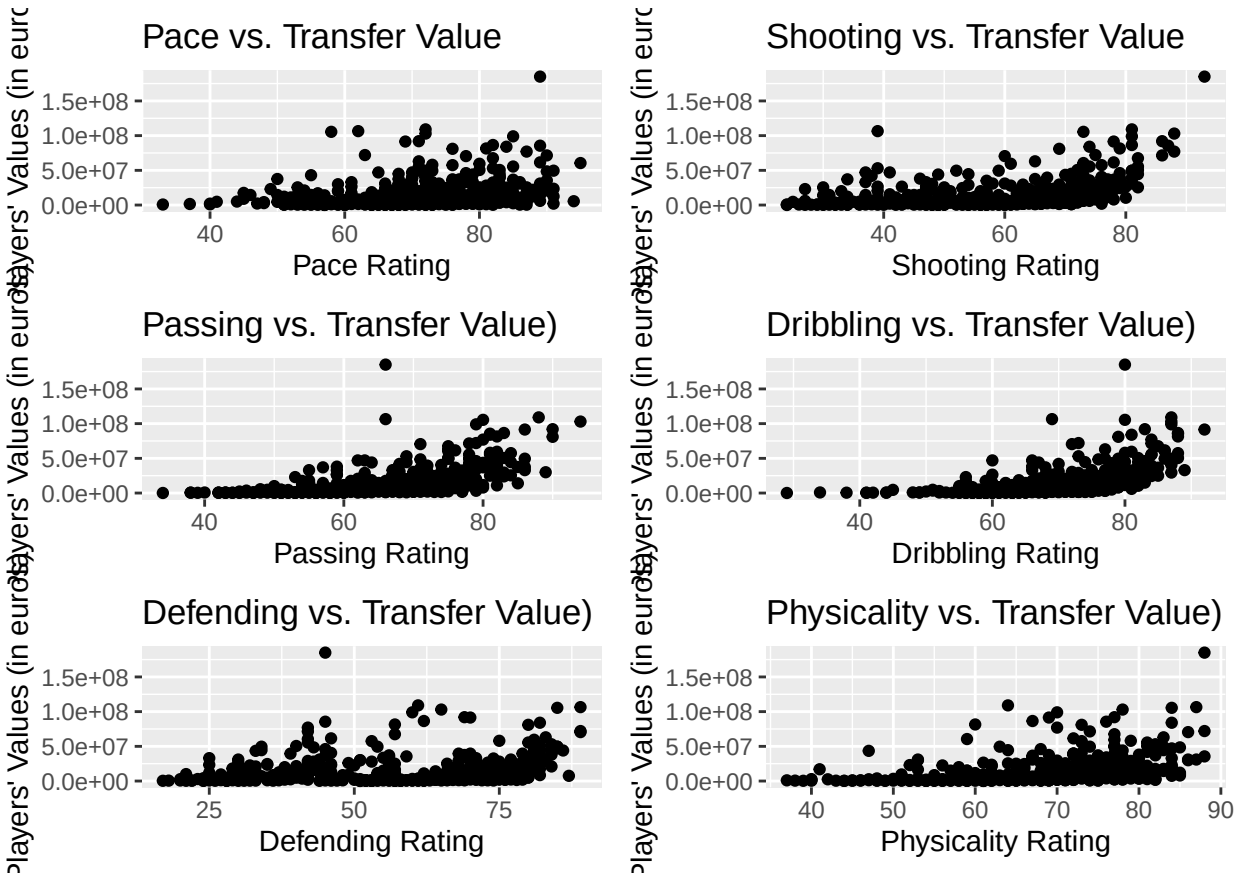


Figure 3. To conduct bivariate analysis for the quantitative we made scatterplots of them as predictors of market value. We saw that there seems to be positive relationships between all the variables and transfer value. In other words, as a rating for a variable goes up, so does the transfer value. However, we can also see that none of these relationships seem to be linear and instead seem to be curved. This suggests that we need to do a transformation is needed to satisfy our linearity condition

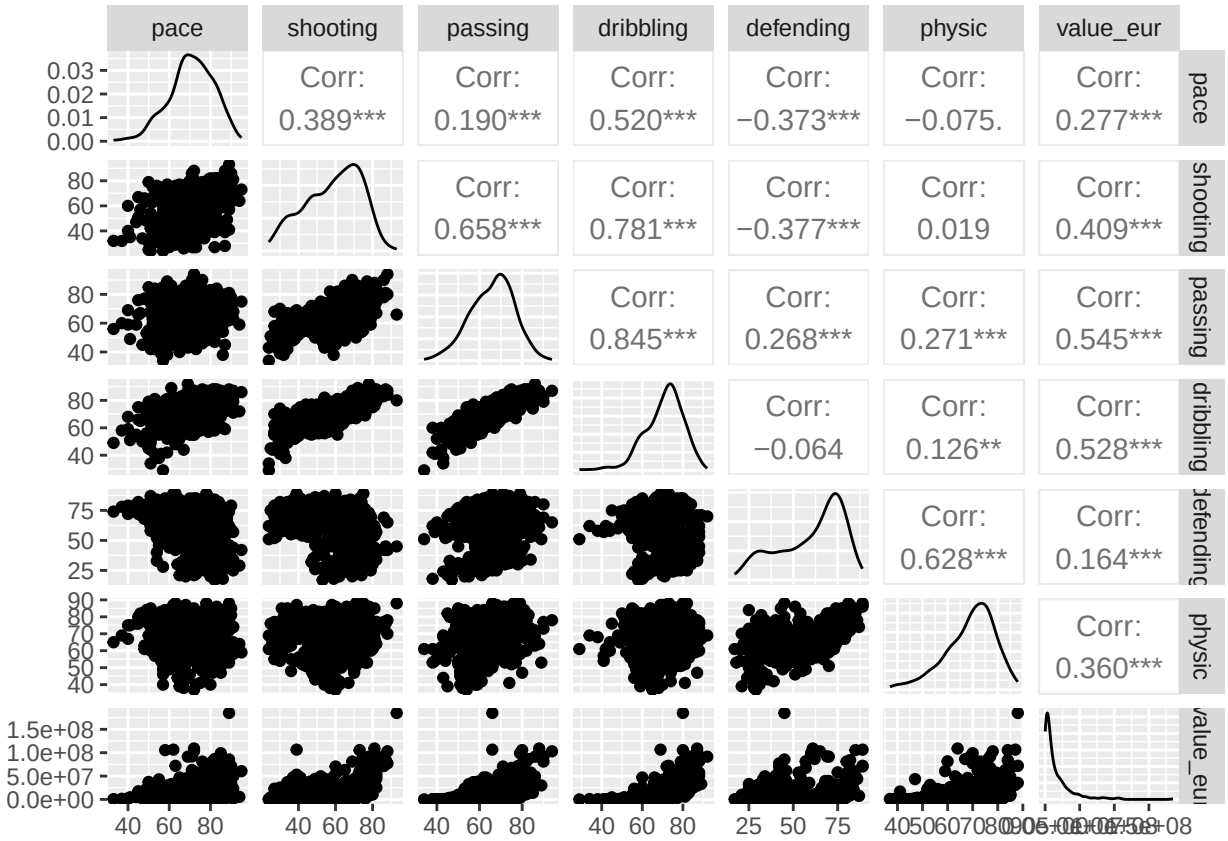
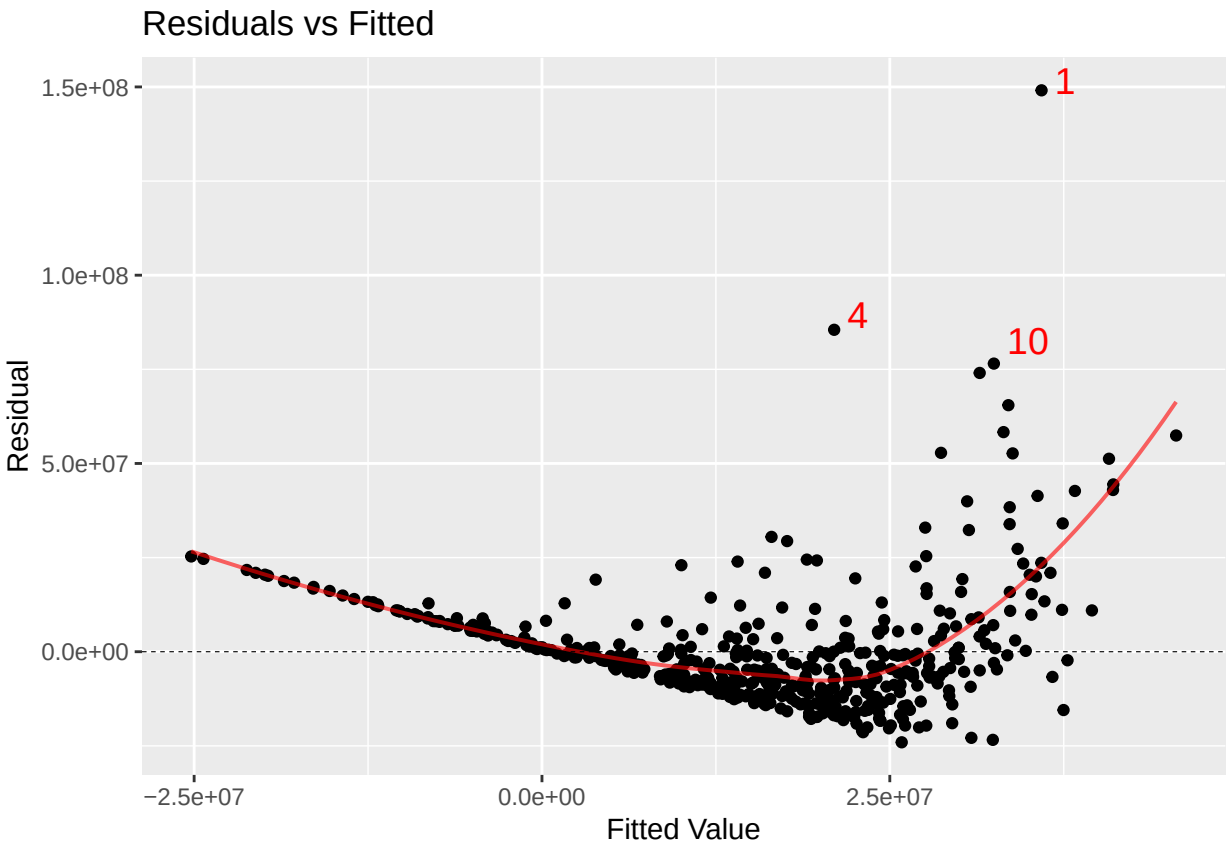
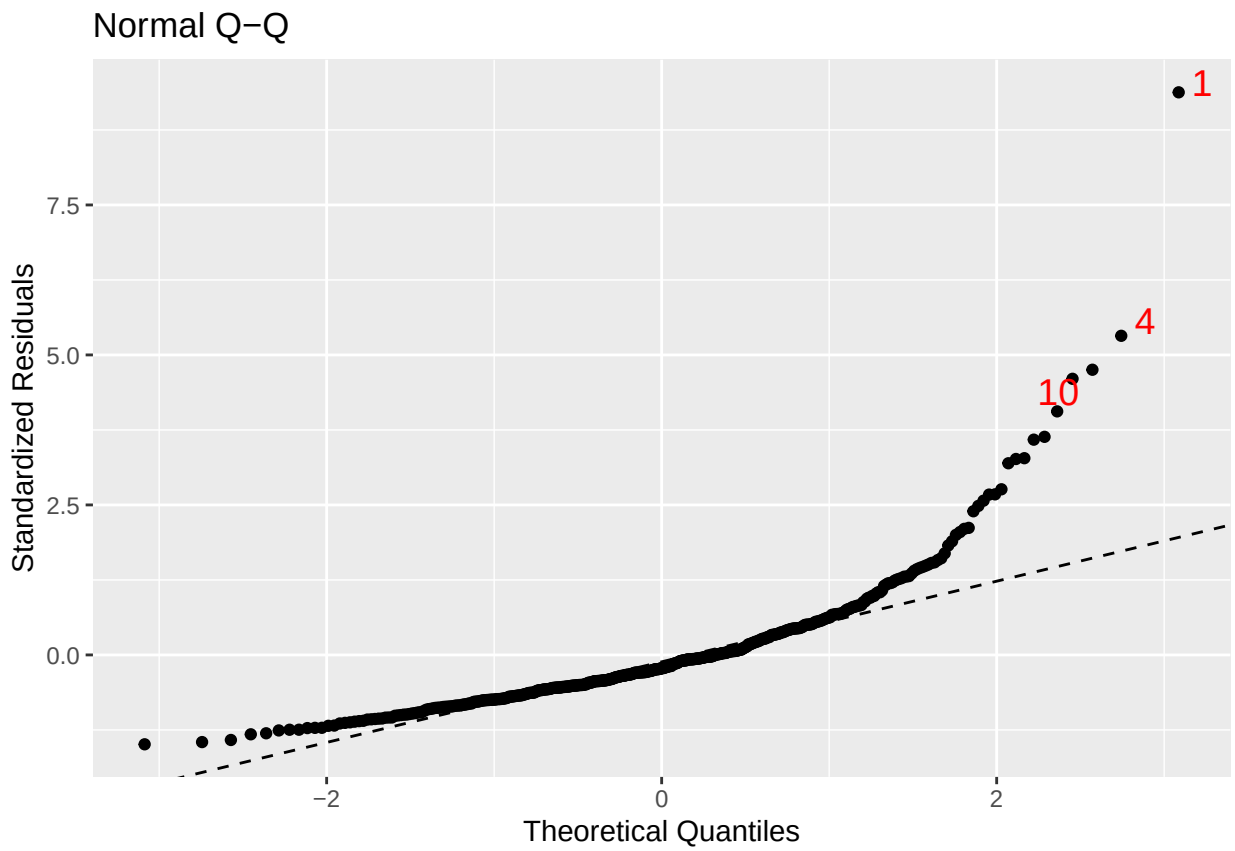


Figure 4. This ggpairs model shows that some of our variables within our dataset are highly correlated with each other. This is especially true for dribbling-passing, dribbling-shooting, passing-shooting and defending-physic. This is concerning and we would need to look into running a multicollinearity test to decide which of our variables are over the threshold of multicollinearity.

Initial model

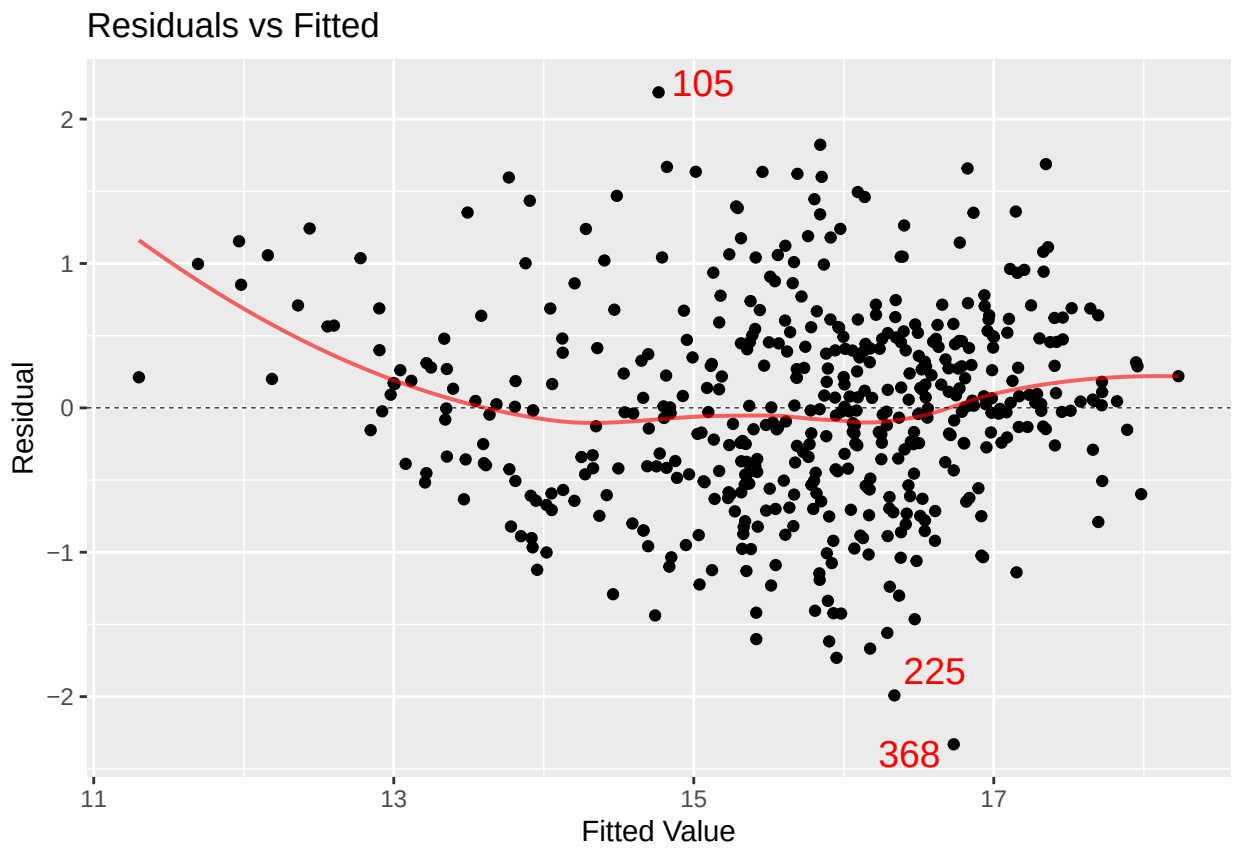
```
## `geom_smooth()` using formula = 'y ~ x'
```

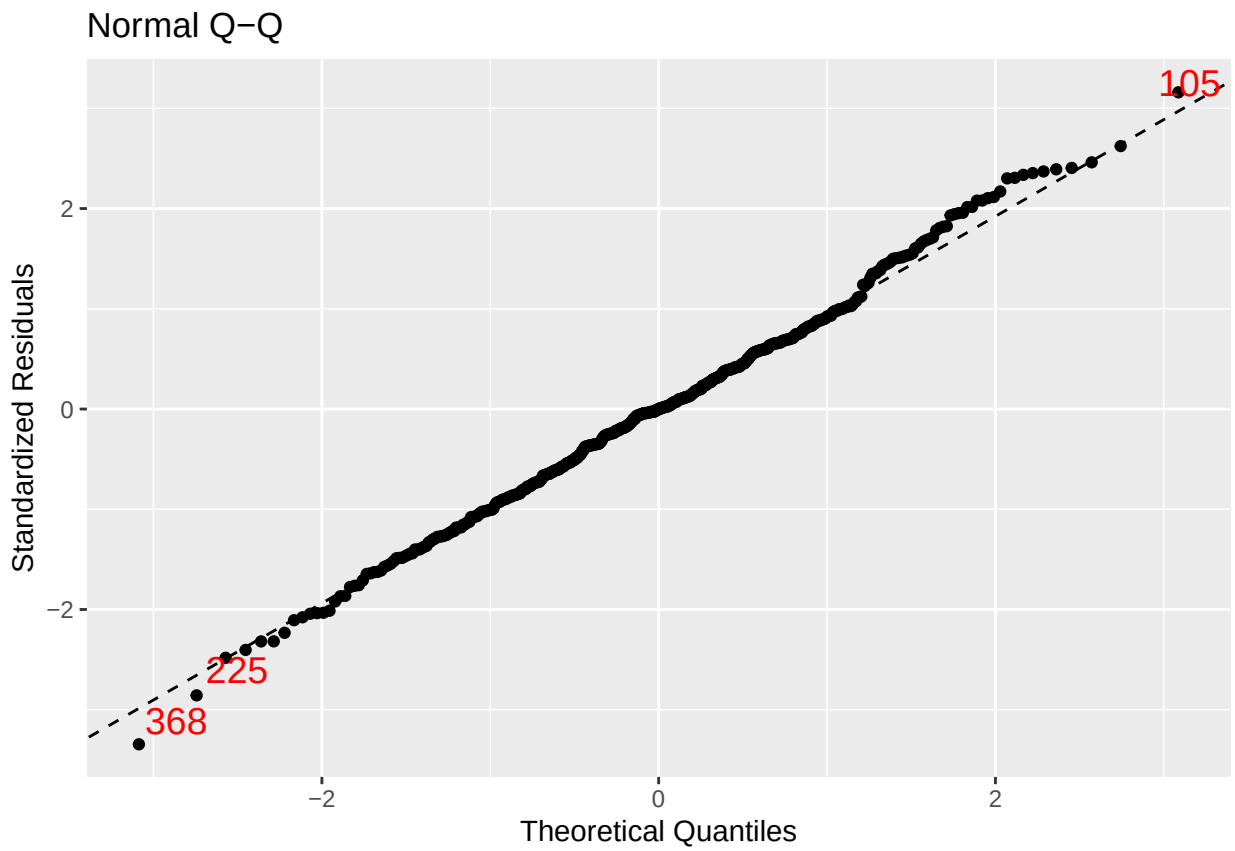

Residuals vs Fitted
1.5e+08
1.0e+08
5.0e+07
0.0e+00
Residual
1
4
10
−2.5e+07
0.0e+00
2.5e+07
Fitted Value



REWRITE: We begin by looking at a kitchen sink model to get a good sense of all of our variables. There are some major issues with the kitchen sink model. The conditions for multiple linear regression are clearly not met. The residuals vs fitted plot is heteroskedastic and has a nonlinear/nonrandom pattern (which violates _ condition). The QQ plot testing normality is far less concerning as most points fall on or near our reference line, however we see the right tail strays away which is a problem. The look of our residuals vs. fitted plot confirms our suspicion that we may need to transform the data. Looking at our summary of the model there are many variables that are currently insignificant which we want to address. Some variables (name them) also have VIF's above five which raises issues of multicollinearity that we also need to address. We will try and fix this issues in our next model beginning with a transformation and then carefully selecting variables to slim down the model and hopefully raise our R^2 value.

```
## `geom_smooth()` using formula = 'y ~ x'
```

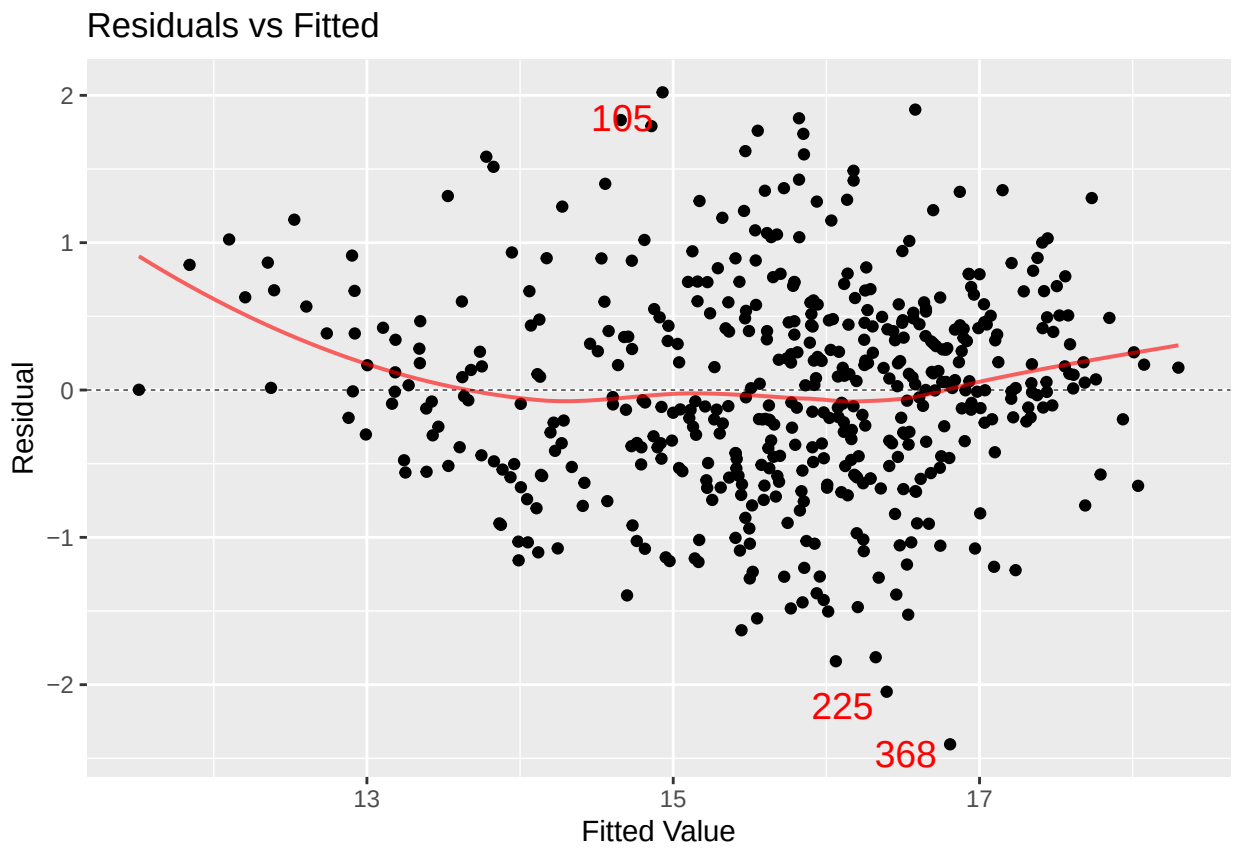

Residuals vs Fitted
105
225
368
2
1
0
−1
−2
Residual
11
13
15
17
Fitted Value

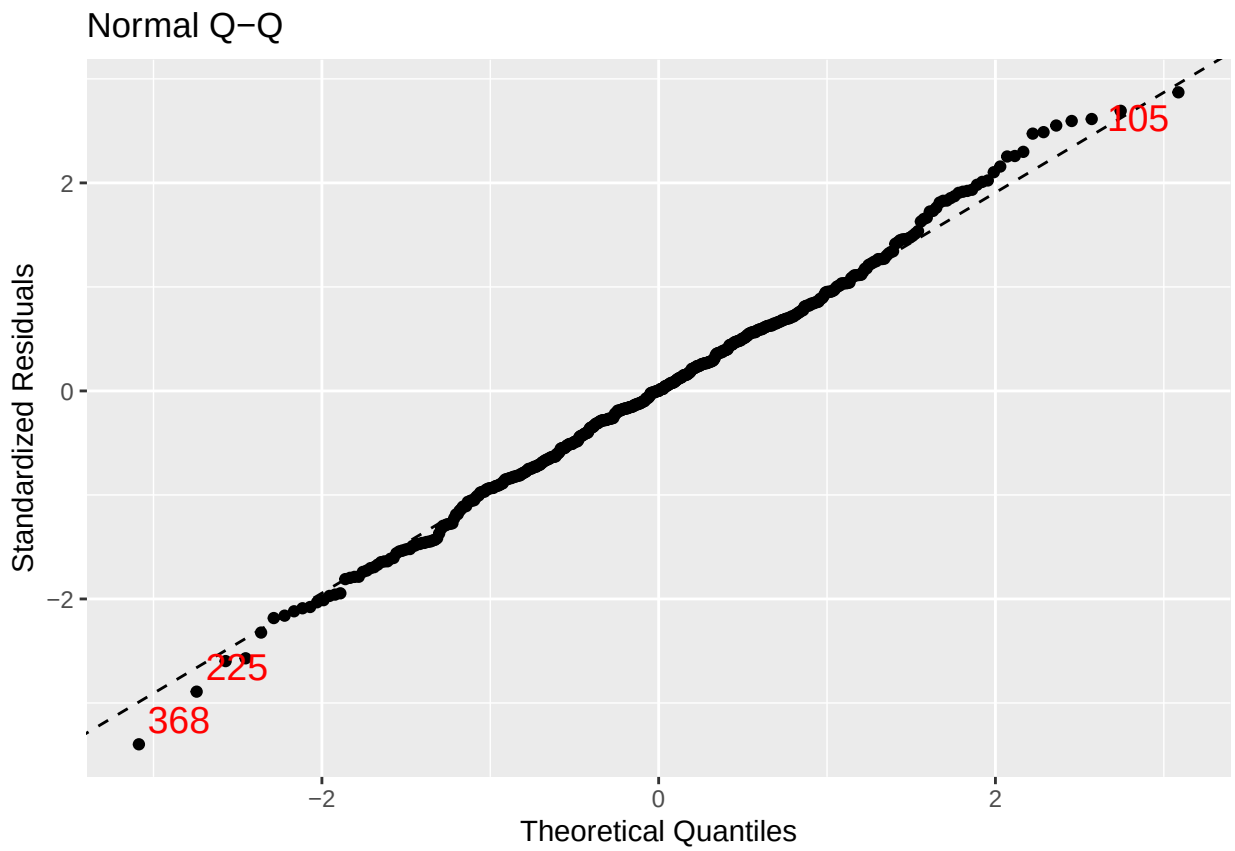


REWRITE: We found the log transformation works best for our data. We immediately see and improvement in conditions. The residuals vs fitted plot still shows some heteroskadisticity (this just seems to be the pattern of the data. There aren't any majorly concerning points. Overall there is a band of randomly distributed points and these plots don't raise enough concern so we proceed with the log transformation. The QQ plot also looks really good. The tails do veer away from the reference line but not nearly as drastically as before. With this transformation we also see a much higher R^2 (state value and compare to past). We also see many more of our variables are now significant. A lot of them have VIFS greater than 5 so we will use a best subsets _____ (I don't remember what if this is a test or not) to narrow down our varaibles and see if there if we can reduce multicollinearity and improve R^2.

REWRITE: Looking at the best subsets output it looks as though 4 variables will give us the best model. Adding more than five variables doesn't significantly improve the R^2 or the Mallows CP which have values of ___ and ___. Additionally 4 variables is not overly complex (in terms of numbers of variables). So we continue with a model that contains (state variables).

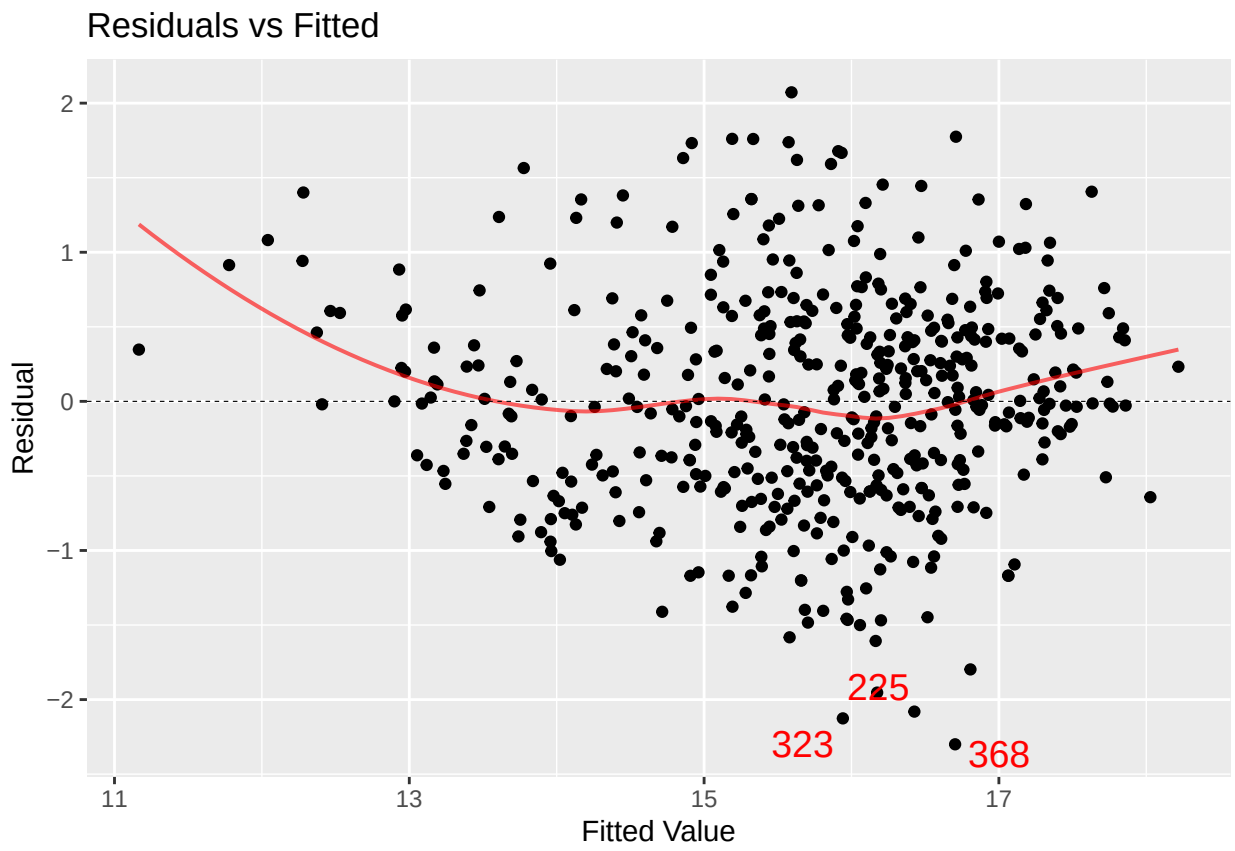
```
## `geom_smooth()` using formula = 'y ~ x'
```

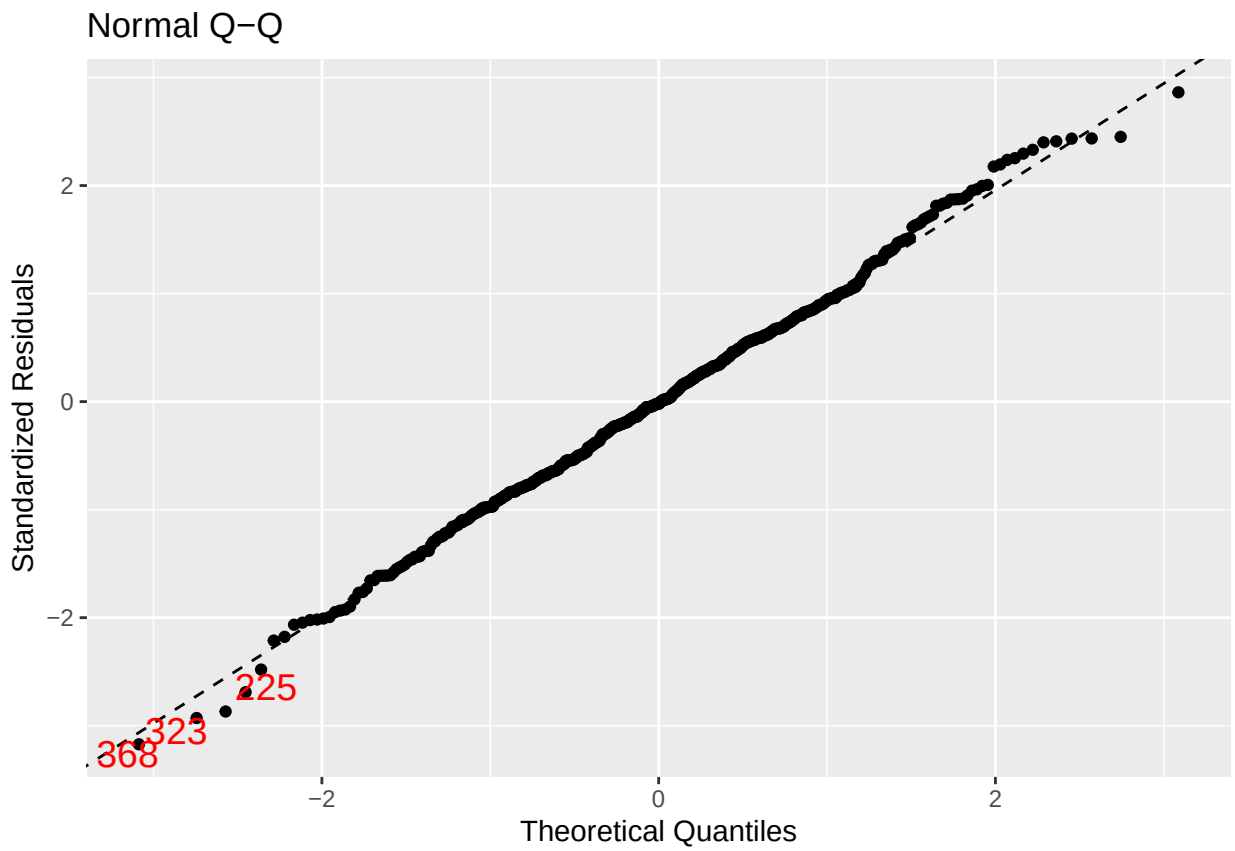

Residuals vs Fitted
105
225
368
2
1
0
−1
−2
Residual
13
15
17
Fitted Value



REWRITE: The model looks good. All variables are significant. Conditions look good as well. Still have high collinearity (passing and dribbling). We reference back to best subsets and remove pace.

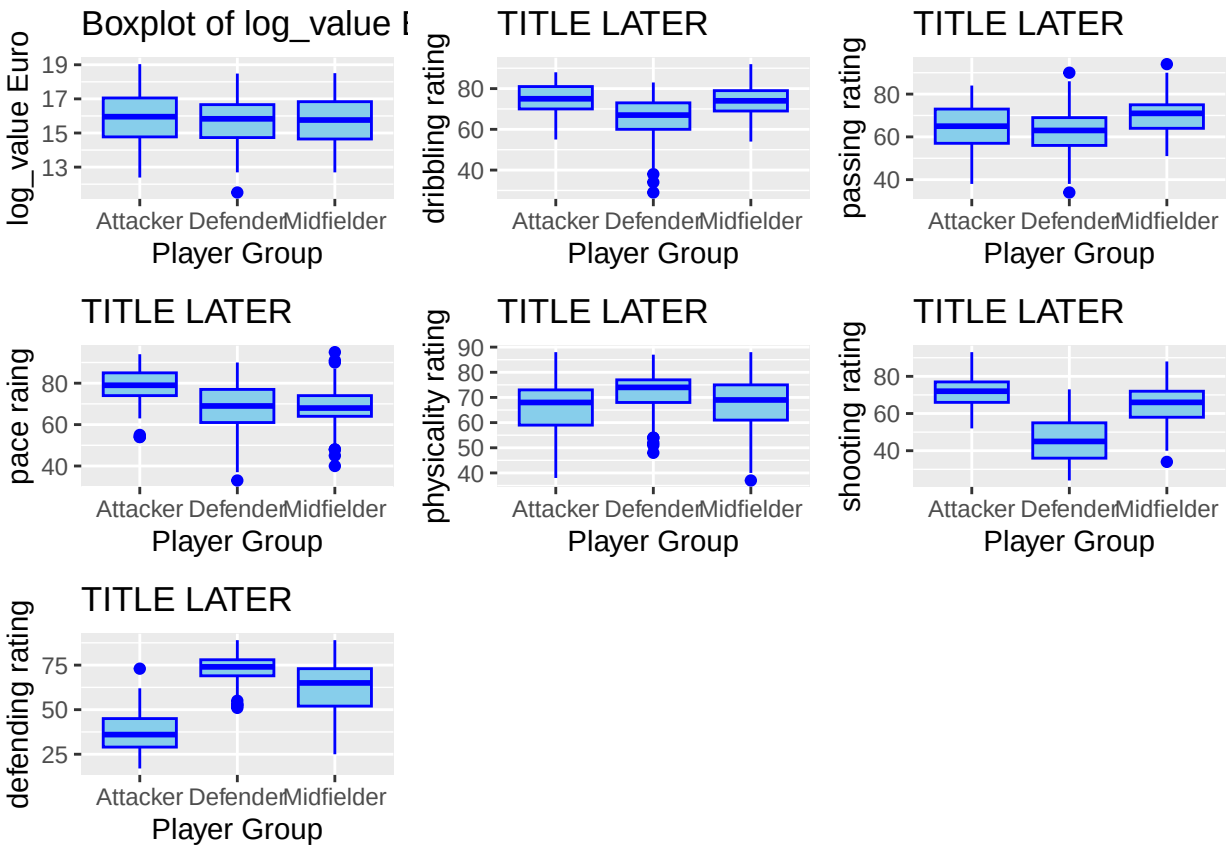
```
## `geom_smooth()` using formula = 'y ~ x'
```


Residuals vs Fitted
Residual
Fitted Value
2
1
0
−1
−2
11
13
15
17
225
323
368

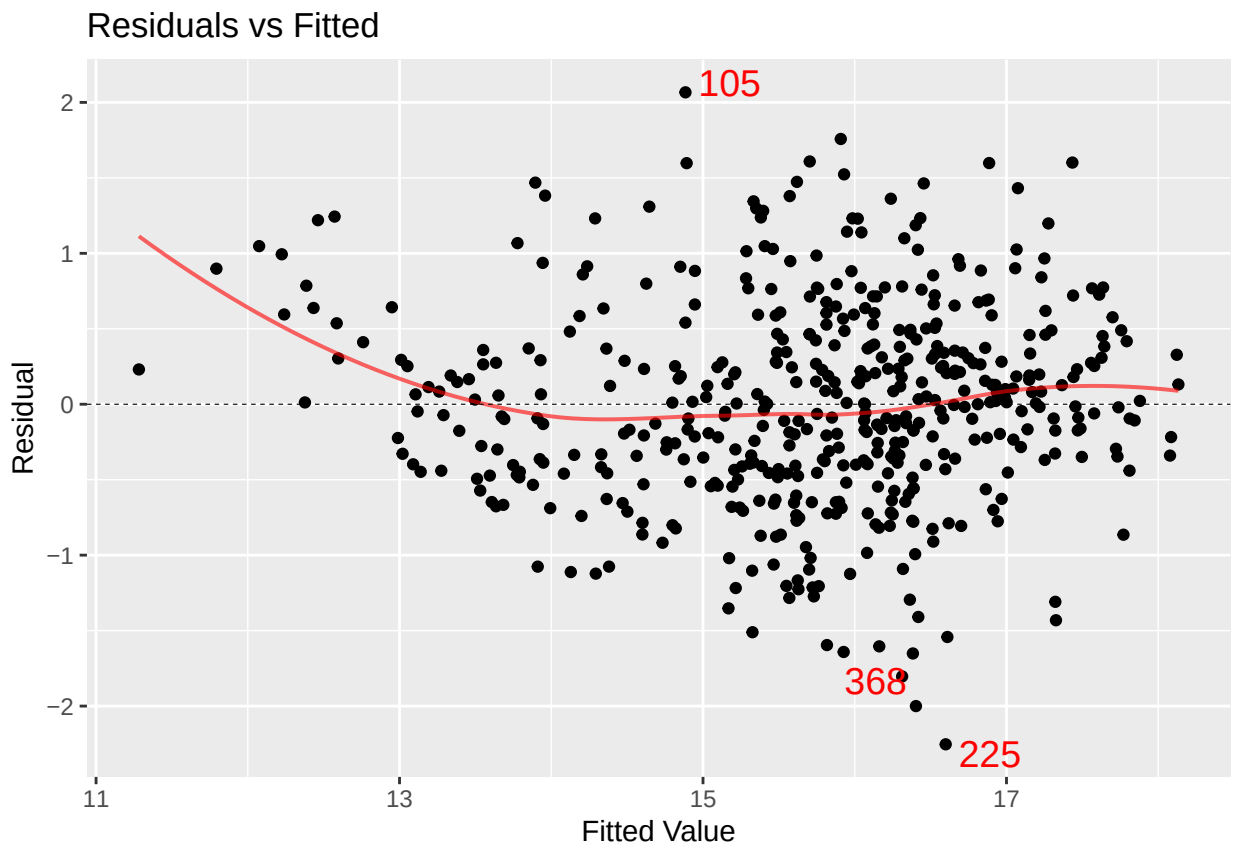


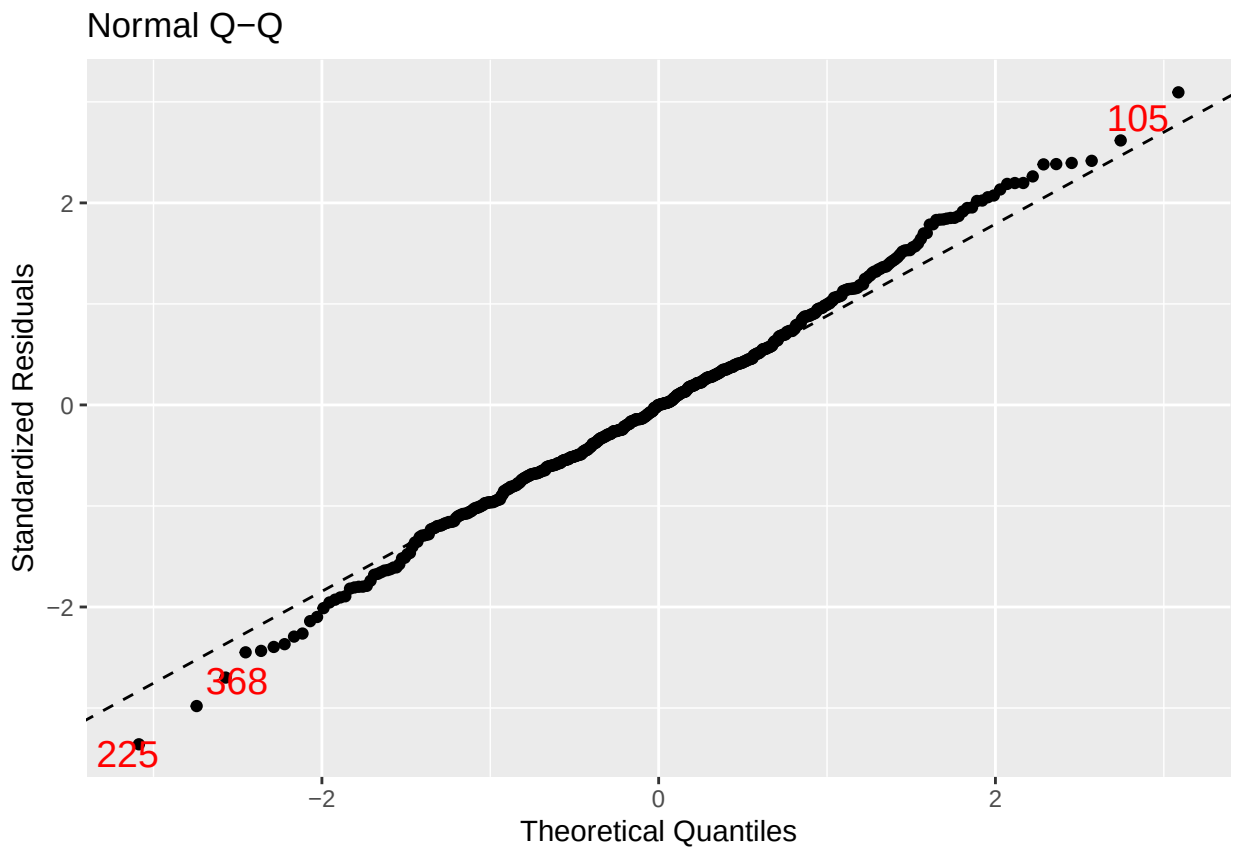
NEED TO EXPLAIN FINDINGS OF THIS MODEL.

REWRITE: We realized that different players have different skills based on position despite having similar valuations. We figured including a variable relating to position could aid in making the model more accurate.



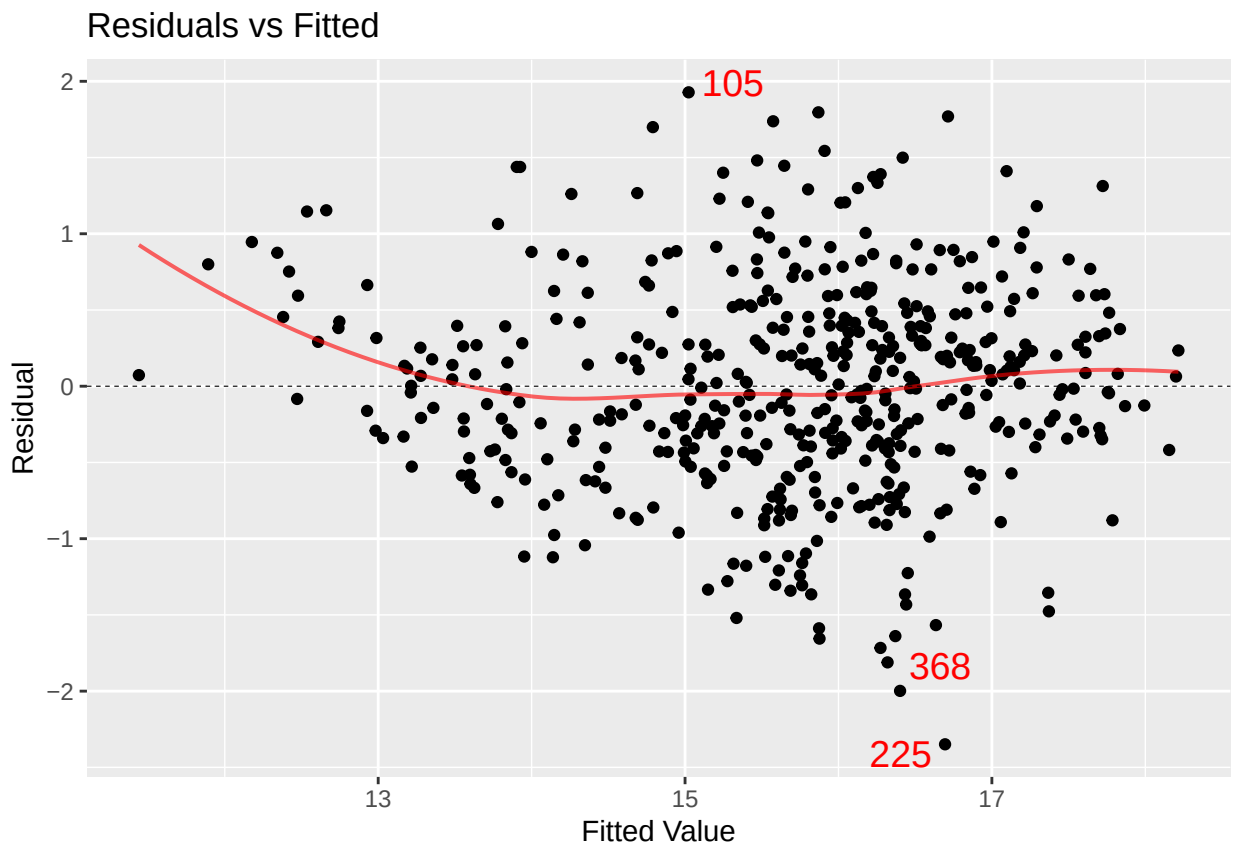

```
## `geom_smooth()` using formula = 'y ~ x'
```

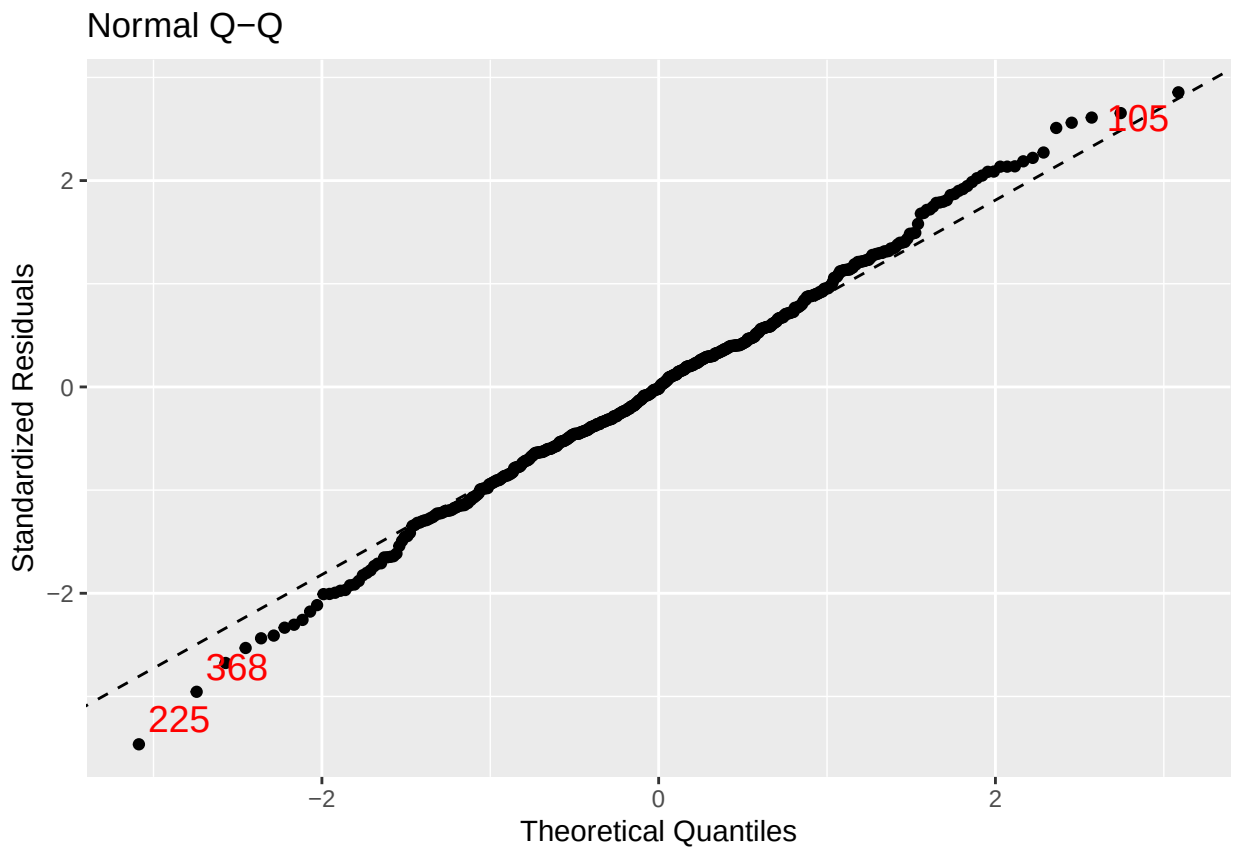

Residuals vs Fitted
105
368
225
2
1
0
−1
−2
Residual
11
13
15
17
Fitted Value



REWRITE: Bringing back all varaibles to this new kitchen sink we see a lot of similar results to our previous kitchen sink model. The plots for conditions are decent. There are issues of VIFs> 5 and some insignificant variables. We'll use best subsets again to simplify the model. Looking at best subsets a five variable model seems to be the best with R^2 and cp of _. mallows cp increases once we get past 6 variables and a five variable model won't be too complex. We'll build a model using (state variables).

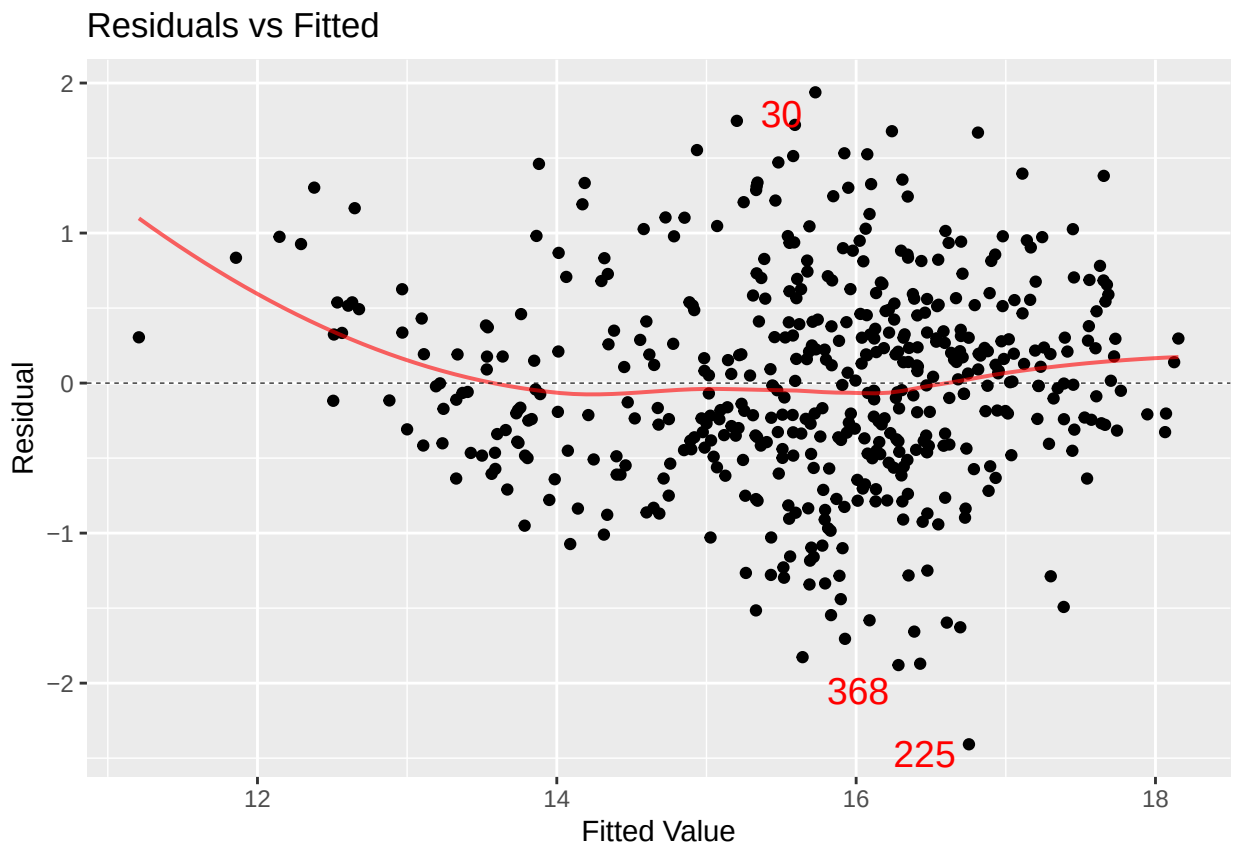
```
## `geom_smooth()` using formula = 'y ~ x'
```

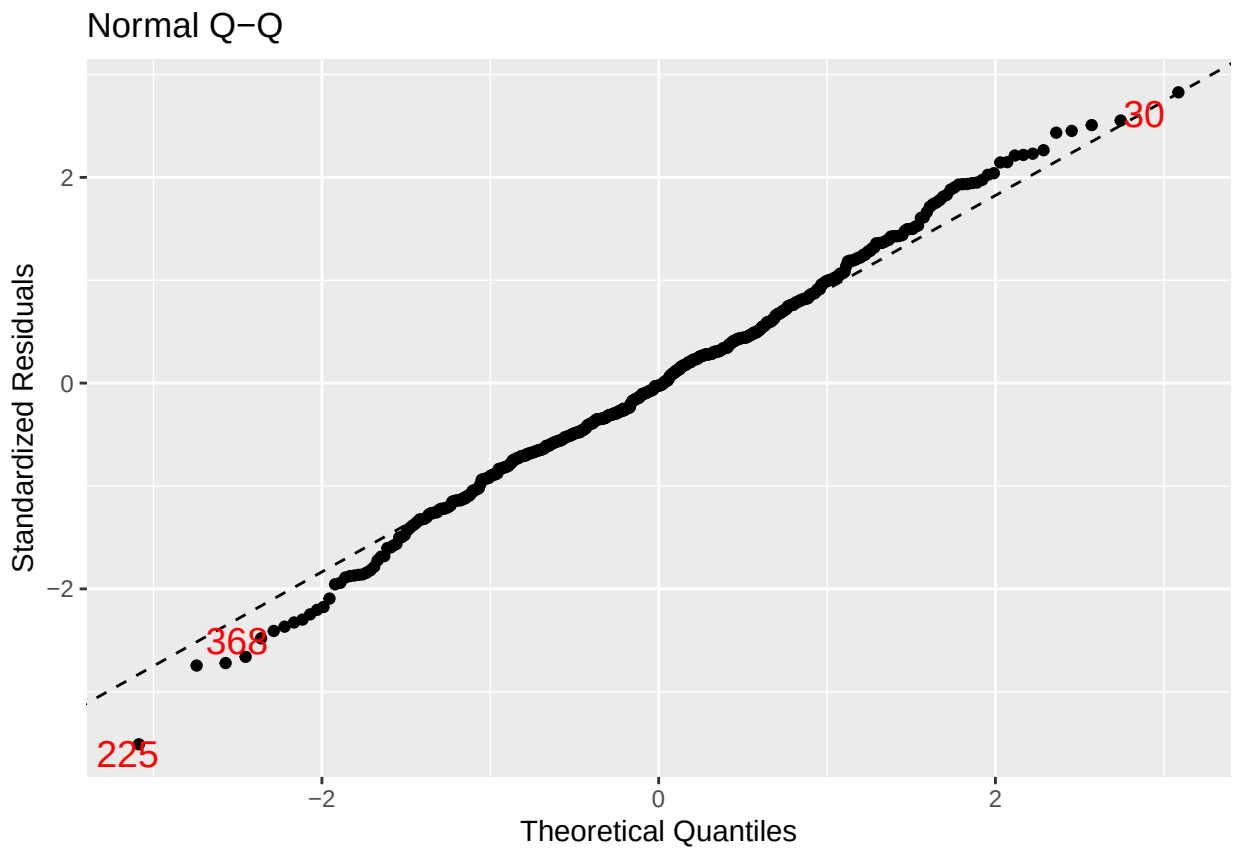

Residuals vs Fitted
105
368
225
2
1
0
−1
−2
Residual
13
15
17
Fitted Value



REWRITE: Conditions look good and all variables significant but still have high VIFs so we'll refer to kitchen sink model and get rid of a variable (state which).

```
## `geom_smooth()` using formula = 'y ~ x'
```


Residuals vs Fitted
Residual
Fitted Value
2
1
0
−1
−2
12
14
16
18
30
368
225



Interpretations ### Multiple Linear Regression: Using the best subsets regression technique, the best model we found for predicting logged transfer value was using variables passing, physical, dribbling and is_midfielder. In this model we had three significant quantitative variables (passing, physical and dribbling) and one binary variable (is_midfielder). After checking for issues with multicollinearity, we found that none of our variables reached the threshold of vif > 5 and therefore there was no cause for concern from a collinearity perspective. All these variables were significant with $P > 0.05$. Analyzing our regression output, we can see that a higher passing, physical, dribbling and pace rating increases a player's logged in-game transfer value. However, we can also see that if the player is a midfielder, it decreases the logged in-game transfer value. Combining all of our significant predictors selected, we account for 76.5% of the variance within in-game transfer value with a residual standard error of 0.689.

Conclusion