

# **Determining Factors Contributing to Point Totals across the Top Four European Soccer Leagues**

By Will Drury and Kabir Wagle

Stat 324 Section 72

3/11/2021

## II. Introduction:

For our project, we will be investigating the point totals for soccer teams across the top 4 leagues in Europe (20 teams each). In soccer, there is always much debate about which metrics define a team's success. Teams have won titles in the past by being defensively astute, refusing to concede goals, and winning games 1-0. Other sides will focus on possession, controlling the ball in such a way that the other team doesn't even have a chance to score. Most commonly, teams will go all-out attack and look to outscore the opponent. We will be seeking to find out which metric is the best determinant of a team's success.

We got all our data from WhoScored, which describes themselves as a “Provider of football data through a web platform created to have an abundance of football statistics and analysis to help bettors, fans and club staff.” In soccer, there is an idea of the “Top 5” leagues which are the English, Spanish, Italian, French, and German leagues. However, the German league only has 18 teams, so we used the other 4 leagues which all have 20 teams. Given that the current season is ongoing, we moved to collect data from the 2019-2020 season. However, upon realizing that the French league did not finish their season due to Covid-19 we will be using data from the 2018-2019 season.

As we touched on before, our **observational units** will be teams in the English(Premier League), Spanish(La Liga), Italian(Serie A), and French(Ligue 1) top divisions who played 38 games in the 2018-2019 season. Whoscored has numerous team statistics for these leagues listed on their website. We singled out Goals Scored, Goals Conceded, Shots-per game, Possession Percentage, and Pass Completion Percentage as our **Quantitative Explanatory Variables**. We felt that these would be strong indicators of a team's success. Our **Categorical Explanatory Variable** will be the league that the team plays in as we are using data from four different leagues.

In soccer, you get 3 points for winning a game, 1 point for tying, and 0 for losing. At the end of the season, the team with the most points wins the league. We will be using the Point Total at the end of the season as our **Response Variable** as that is the truest indicator of how well the team did that year.

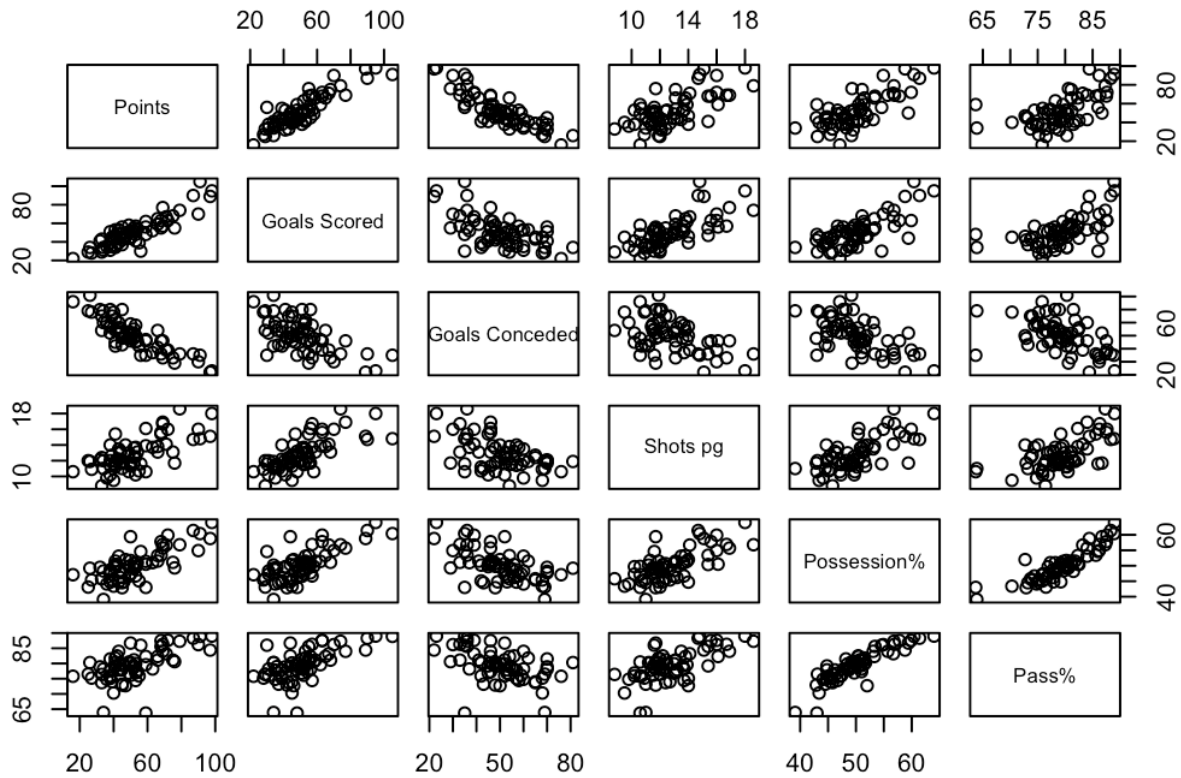
### **III. Split the Data:**

We first need to split up our data into two different data sets: training data and test data. The training data is a random sample of 80% of our data. Since we have 80 teams in the dataset, our training data consists of 64 teams. The remaining 16 teams will be set aside as test data. We will use the test data to validate our model later on. The model validation process will allow us to see how well our model fits data outside of our sample. In R, we are using seed number 333 for the `set.seed` function. This ensures that the R software will pick the same sample every time we run our code. We will then split the data up into two dataframes called `traindata` and `testdata`, and perform all of our model analysis using `traindata`.

#### IV. Data Visualization:

Part A:

Scatterplot Matrix:



### Correlation Matrix:

	Points	Goals.Scored	Goals.Conceded	Shots.pg	Possession.	Pass.
Points	1.0000000	0.8803463	-0.8575851	0.6899313	0.7523717	0.6501526
Goals.Scored	0.8803463	1.0000000	-0.6010423	0.7119272	0.7360814	0.6285562
Goals.Conceded	-0.8575851	-0.6010423	1.0000000	-0.5281754	-0.6163045	-0.5447906
Shots.pg	0.6899313	0.7119272	-0.5281754	1.0000000	0.6853557	0.6204469
Possession.	0.7523717	0.7360814	-0.6163045	0.6853557	1.0000000	0.8812252
Pass.	0.6501526	0.6285562	-0.5447906	0.6204469	0.8812252	1.0000000

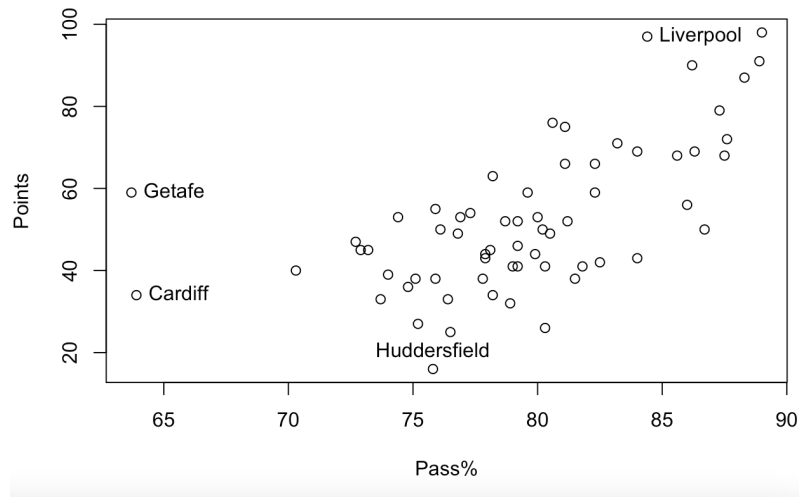
We first investigated the relationships between the explanatory variables and the response variable. Goals scored and goals conceded appeared to be most strongly correlated with number of points. Goals scored exhibited a strong positive linear relationship with points. Goals conceded had a strong negative linear relationship with points, with some slight curvature to the data. We identified that a power transformation decreasing the x power might be needed to improve the linearity of the relationship.

Our three other explanatory variables, Shots pg, Possession%, and Pass%, all appeared to show a moderately strong positive linear relationship with points. However, these plots show slightly more variability and outliers than the goals scored and goals conceded plots, and also show slightly smaller correlations within the correlation matrix.

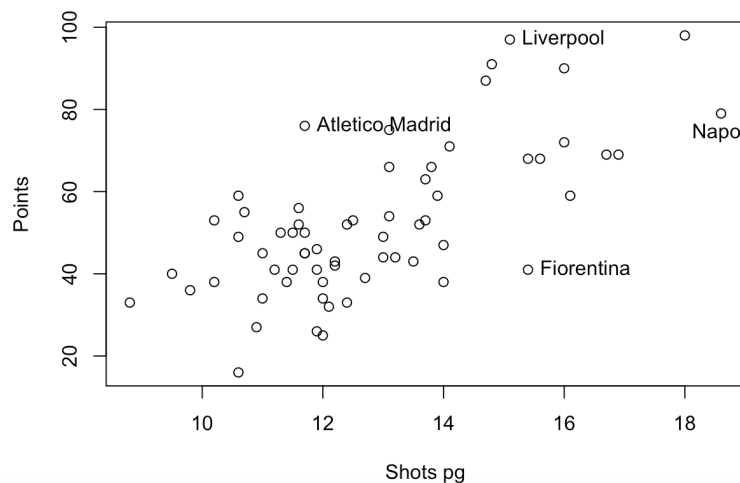
We subsequently investigated the relationships between each of the explanatory variables. Pass% and Possession% showed a strong positive linear relationship. Shots pg and goals scored exhibited a moderately strong positive linear relationship. Possession % and goals scored showed a moderate positive linear relationship. Possession % and shots per game also showed a moderate positive linear relationship. Goals conceded and shots per game were the least correlated explanatory variables.

The associations behave generally as expected. There is a positive relationship between all of the explanatory variables and points and a negative relationship between goals conceded and points. This makes sense intuitively because more goals conceded would indicate a higher chance of losing games.

### Unusual observations:



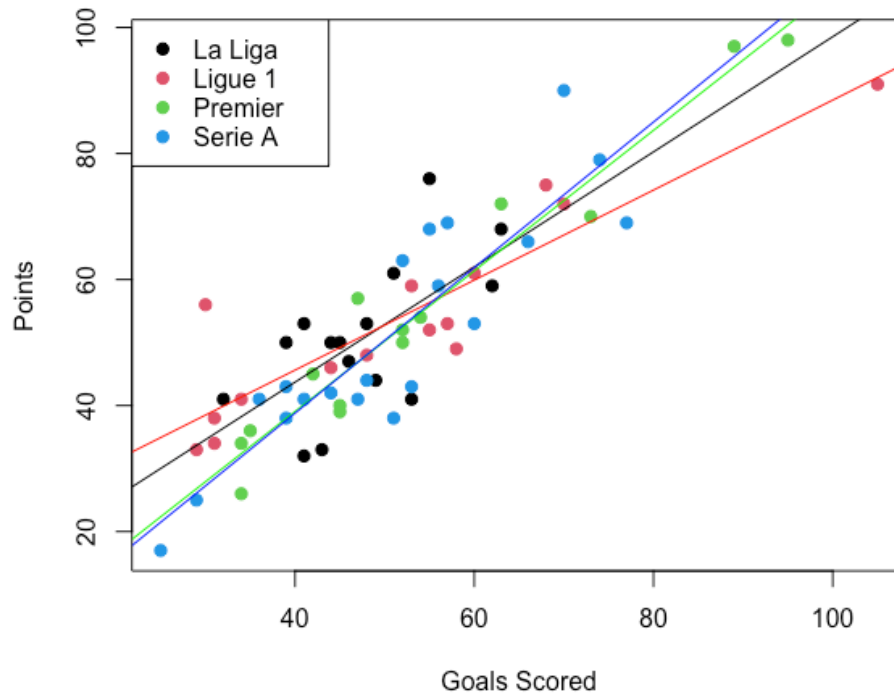
The graph of Pass% vs Points exhibited some unusual outliers. Specifically, Getafe had a very high point total of 59 given the extremely low 63.7% pass percentage relative to the other teams. Also, Huddersfield showed an abnormally small point score of 16 given a 75.8% pass percentage.



The graph of shots pg vs. points yielded a couple outliers. Napoli had a relatively low point score of 79 given their extremely high 18.6 shots per game. Fiorentina had a very low point score of 41 given their comparatively high 15.4 shots per game.

Part B:

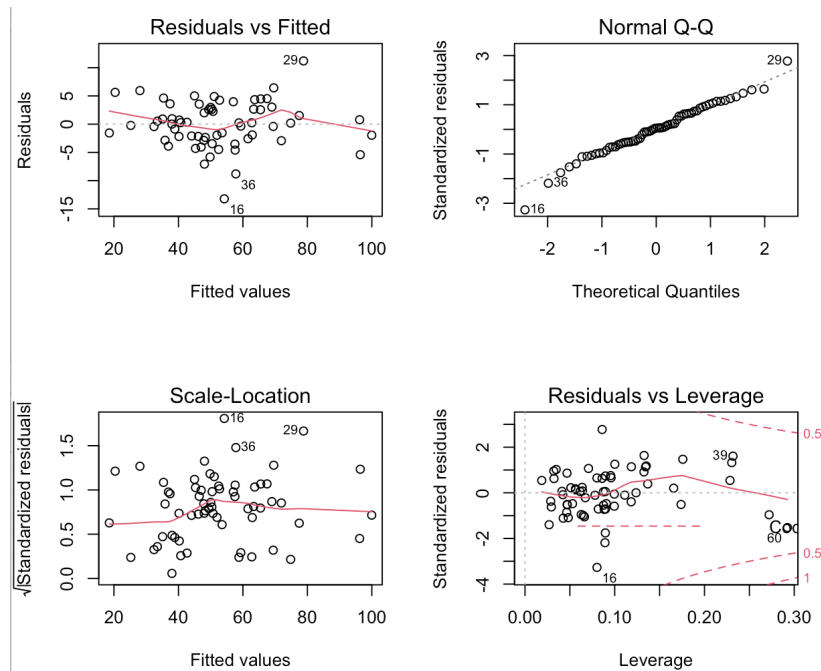
**Interaction plot between points (response) vs. goals scored (quantitative) and league (categorical)**



In order to identify any interaction between goals scored and league, we plotted the graph of points vs. goals scored and color-coded the points to identify each team. We then ran a model of points vs. goals scored, league, and league \* goals scored (interaction term). Fitting the model gave us separate lines for each league with both different intercepts and different slopes. To say there is an interaction between goals scored and league means that the effect of goals scored is significantly different depending on the league. We suspected the interaction might be significant given that each line had a noticeably different slope.

## V. Variable Pre-Processing:

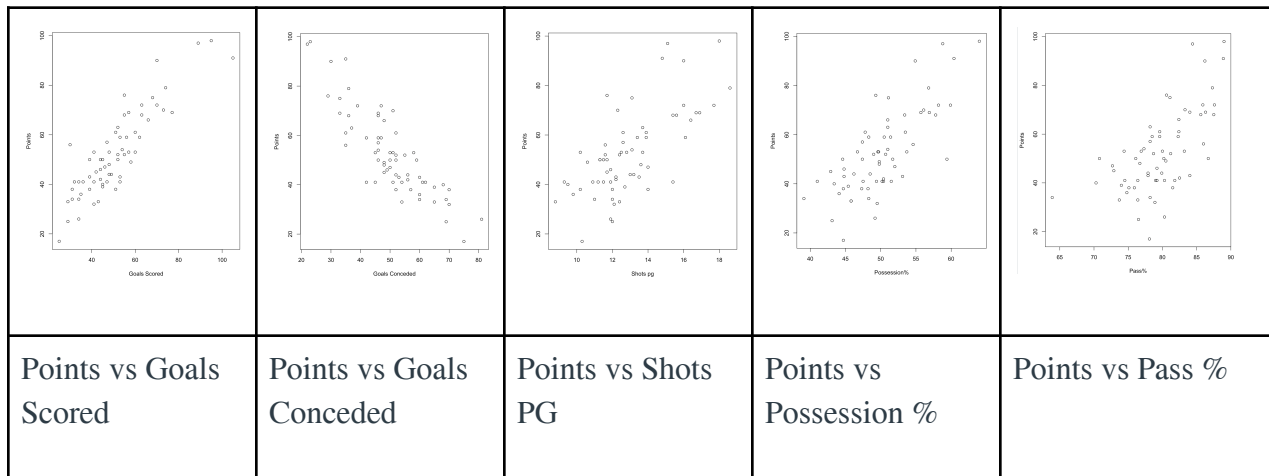
Residual plots (full model):



In order to assess whether transformations are needed, we looked at the initial residual plots of the model with all predictor variables included. The residuals vs. fitted plot appeared to show a slight issue with linearity. The equal variance assumption appeared to be met. The normal Q-Q plot showed slight deviation from the line but generally looked to satisfy the normality assumption. Given that the only issue seemed to be with linearity, we decided to look at transformations of the predictor variables.



## Scatterplots:



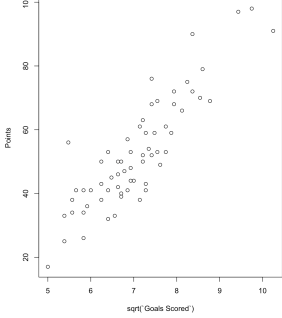
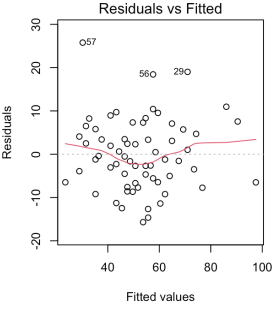
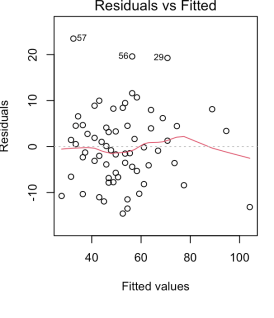
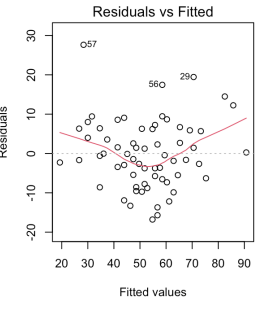
Looking at the scatterplots, most of the explanatory variables show a moderate to strong linear association with point totals. Goals Scored and Goals Conceded seem to have a very strong relationship. Shots PG and Possession % don't have as strong of a relationship, but the scatter plot seems to be more linear than anything else. Pass % on the other hand has a slight linear curve to it, prompting a linear transformation. While the relationship between Goals conceded and goals scored looked very strong, it also showed slight curvature, prompting us to investigate those residual plots.

As mentioned before, the graphs of Shots PG and Pass % seem to have more unusual observations than the other two graphs. Specifically, Fiorentina, Napoli, and Atletico for Shots PG and Getafe, Cardiff, and Huddersfield for Pass%.

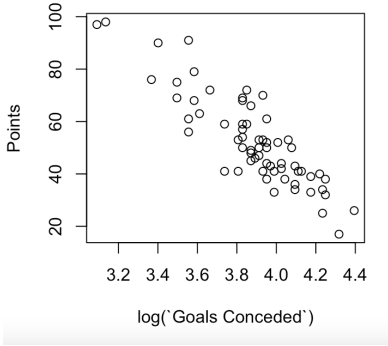
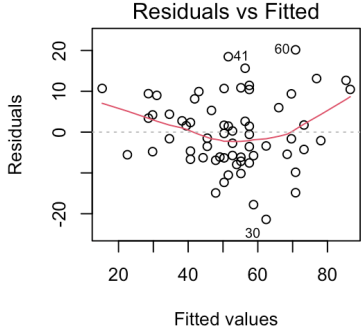
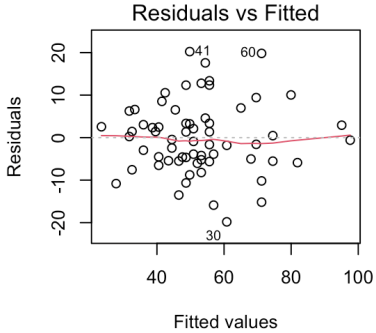
Explanatory Variable	Standard Deviation
Goals Scored	8.223
Goals Conceded	8.916
Shots PG	12.55
Possession %	11.42
Pass %	13.17

Additionally, Goals Scored and Goals Conceded have a much smaller Standard deviation than the other three explanatory variables.

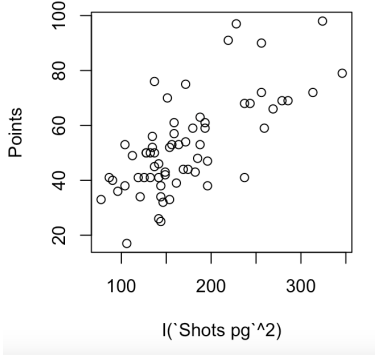
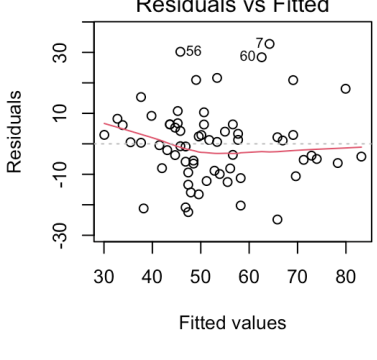
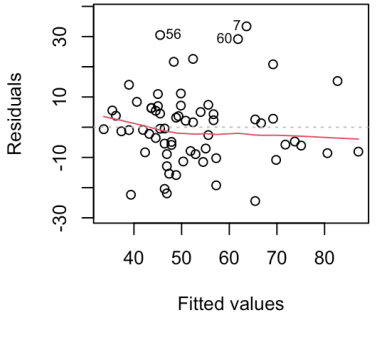
## Points vs Goals Scored Transformation:

Scatter Plot of Points vs sqrt(Goals Scored)	Residual plot of Points vs Goals Scored	Residual plot of Points vs sqrt(Goals Scored)	Residual plot of Points vs log(Goals Scored)
			
<p>After reducing the power of x based on the curvature, the scatter plot of Points vs sqrt(Goals Scored) appears to be more linear than the original.</p>	<p>Upon plotting the residual plot of Points vs Goals Scored, we could see issues with linearity, and decided to reduce the power given how well the transformation worked in the scatterplot.</p>	<p>The residual plot of Points vs sqrt(Goals Scored) looks to have less curvature, prompting us to conclude that this model is more linear than the original one.</p>	<p>To test if sqrt(Goals Scored) was the best reduction of x transformation we also plotted a residual plot of Points vs log(Goals Scored). As you can see, there are even more issues with linearity than the original plot now.</p>

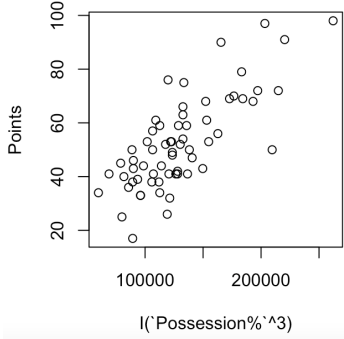
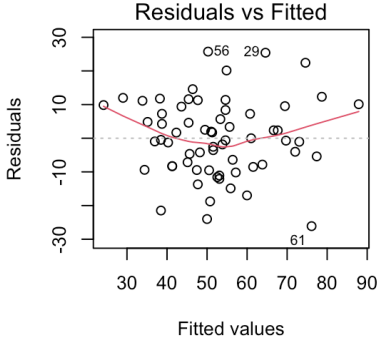
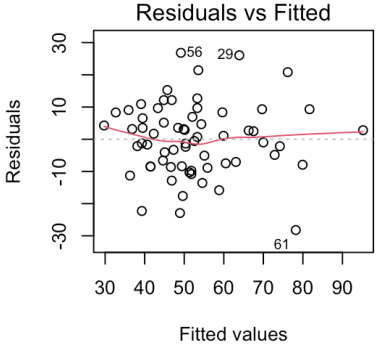
# Points vs Goals Conceded Transformation:

Scatter Plot of Points vs log(Goals Conceded)	Residual plot of Points vs Goals Conceded	Residual plot of Points vs log(Goals Conceded)
		
<p>There appeared to be a slight curvature in the Points vs goals conceded plot, after decreasing the power of x to log(goals conceded) the graph now looks more linear.</p>	<p>Upon plotting the residual plot of Points vs Goals Conceded, we could see issues with linearity, and decided to reduce the power given how well the transformation worked in the scatterplot.</p>	<p>Plotting a residual plot of Points vs log(Goals Conceded) confirmed this theory as there is significantly less curvature, prompting us to conclude that this model is more linear than the original one.</p>

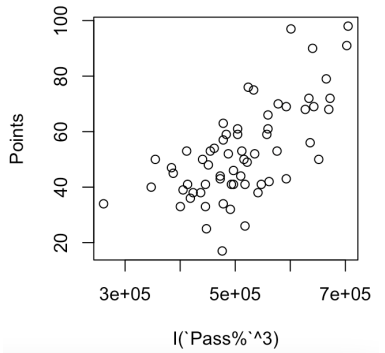
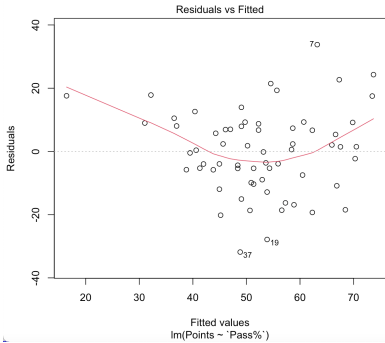
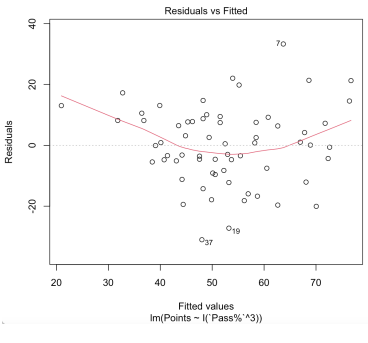
Points vs Shots PG Transformation:

Scatter Plot of Points vs. (ShotsPG) <sup>2</sup>	Residual plot of Points vs. ShotsPG	Residual plot of Points vs. (ShotsPG) <sup>2</sup>
		
<p>After increasing the power of Shots PG to 2, the scatterplot now looks more linear.</p>	<p>After plotting the residual plot of Points vs Shots PG, we could see issues with linearity.</p>	<p>Plotting a residual plot of Points vs (Shots PG)<sup>2</sup> appeared to improve linearity compared to the original residual plot of Points vs Goals Scored.</p>

# Points vs Possession% Transformation:

Scatter Plot of Points vs. (Possession%)^3	Residual plot of Points vs. Possession%	Residual plot of Points vs. (Possession%)^3
		
<p>After increasing the power of possession% to 3, the scatterplot now looks more linear.</p>	<p>After plotting the residual plot of Points vs Possession%, we could see issues with linearity.</p>	<p>Plotting a residual plot of Points vs (Possession%)^3 appeared to improve linearity compared to the original residual plot of Points vs Possession%</p>

## Points vs Pass% vs. Transformation:

Scatter Plot of Points vs (Pass% <sup>3</sup> )	Residual plot of Points vs Pass%	Residual plot of Points vs Points vs (Pass% <sup>3</sup> )
		
<p>To address the curvature in the Points vs Pass% graph, we decided to apply a power transformation to pass %. Raising the power to 3 appeared to improve the linearity.</p>	<p>Upon plotting the residual plot of Points vs Pass%, we could see issues with linearity, and decided to reduce the power given how well the transformation worked in the scatterplot.</p>	<p>The residual plot of Points vs (Pass%<sup>3</sup>) looks to have less curvature, prompting us to conclude that this model is more linear than the original one.</p>

## Current Full Model

Points = sqrt(Goals Scored)+ log(Goals Conceded) + (ShotsPG)^2 + (Possession%)^3 + (Pass%^3)

Original Full Model	Transformed Full Model																																																																																										
<p>Call: lm(formula = Points ~ `Goals Scored` + `Goals Conceded` + `Shots pg` `Possession%` + `Pass%`)</p> <p>Residuals:</p> <table><tr><th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr><tr><td>-13.2239</td><td>-2.4015</td><td>0.2082</td><td>2.7300</td><td>11.1936</td></tr></table> <p>Coefficients:</p> <table><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(&gt; t )</th></tr><tr><td>(Intercept)</td><td>52.81616</td><td>12.20515</td><td>4.327</td><td>6.03e-05 ***</td></tr><tr><td>`Goals Scored`</td><td>0.58892</td><td>0.05627</td><td>10.466</td><td>5.59e-15 ***</td></tr><tr><td>`Goals Conceded`</td><td>-0.70876</td><td>0.05768</td><td>-12.288</td><td>&lt; 2e-16 ***</td></tr><tr><td>`Shots pg`</td><td>0.12847</td><td>0.36991</td><td>0.347</td><td>0.730</td></tr><tr><td>`Possession%`</td><td>0.14807</td><td>0.26306</td><td>0.563</td><td>0.576</td></tr><tr><td>`Pass%`</td><td>-0.04822</td><td>0.23087</td><td>-0.209</td><td>0.835</td></tr></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 4.218 on 58 degrees of freedom Multiple R-squared: 0.9446, Adjusted R-squared: 0.9398 F-statistic: 197.8 on 5 and 58 DF, p-value: &lt; 2.2e-16</p>	Min	1Q	Median	3Q	Max	-13.2239	-2.4015	0.2082	2.7300	11.1936		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	52.81616	12.20515	4.327	6.03e-05 ***	`Goals Scored`	0.58892	0.05627	10.466	5.59e-15 ***	`Goals Conceded`	-0.70876	0.05768	-12.288	< 2e-16 ***	`Shots pg`	0.12847	0.36991	0.347	0.730	`Possession%`	0.14807	0.26306	0.563	0.576	`Pass%`	-0.04822	0.23087	-0.209	0.835	<p>Call: lm(formula = Points ~ sqrt(`Goals Scored`) + log(`Goals Conceded`) + I(`Shots pg`^2) + I(`Possession%`^3) + I(`Pass%`^3))</p> <p>Residuals:</p> <table><tr><th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr><tr><td>-12.6201</td><td>-2.3783</td><td>-0.0259</td><td>2.8710</td><td>9.3487</td></tr></table> <p>Coefficients:</p> <table><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(&gt; t )</th></tr><tr><td>(Intercept)</td><td>1.226e+02</td><td>1.344e+01</td><td>9.123</td><td>8.31e-13 ***</td></tr><tr><td>sqrt(`Goals Scored`)</td><td>8.228e+00</td><td>7.989e-01</td><td>10.300</td><td>1.03e-14 ***</td></tr><tr><td>log(`Goals Conceded`)</td><td>-3.387e+01</td><td>2.610e+00</td><td>-12.980</td><td>&lt; 2e-16 ***</td></tr><tr><td>I(`Shots pg`^2)</td><td>3.504e-03</td><td>1.312e-02</td><td>0.267</td><td>0.790</td></tr><tr><td>I(`Possession%`^3)</td><td>2.094e-05</td><td>3.133e-05</td><td>0.669</td><td>0.506</td></tr><tr><td>I(`Pass%`^3)</td><td>9.690e-08</td><td>1.195e-05</td><td>0.008</td><td>0.994</td></tr></table> <p>--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 4.039 on 58 degrees of freedom Multiple R-squared: 0.9492, Adjusted R-squared: 0.9448 F-statistic: 216.8 on 5 and 58 DF, p-value: &lt; 2.2e-16</p>	Min	1Q	Median	3Q	Max	-12.6201	-2.3783	-0.0259	2.8710	9.3487		Estimate	Std. Error	t value	Pr(> t )	(Intercept)	1.226e+02	1.344e+01	9.123	8.31e-13 ***	sqrt(`Goals Scored`)	8.228e+00	7.989e-01	10.300	1.03e-14 ***	log(`Goals Conceded`)	-3.387e+01	2.610e+00	-12.980	< 2e-16 ***	I(`Shots pg`^2)	3.504e-03	1.312e-02	0.267	0.790	I(`Possession%`^3)	2.094e-05	3.133e-05	0.669	0.506	I(`Pass%`^3)	9.690e-08	1.195e-05	0.008	0.994
Min	1Q	Median	3Q	Max																																																																																							
-13.2239	-2.4015	0.2082	2.7300	11.1936																																																																																							
	Estimate	Std. Error	t value	Pr(> t )																																																																																							
(Intercept)	52.81616	12.20515	4.327	6.03e-05 ***																																																																																							
`Goals Scored`	0.58892	0.05627	10.466	5.59e-15 ***																																																																																							
`Goals Conceded`	-0.70876	0.05768	-12.288	< 2e-16 ***																																																																																							
`Shots pg`	0.12847	0.36991	0.347	0.730																																																																																							
`Possession%`	0.14807	0.26306	0.563	0.576																																																																																							
`Pass%`	-0.04822	0.23087	-0.209	0.835																																																																																							
Min	1Q	Median	3Q	Max																																																																																							
-12.6201	-2.3783	-0.0259	2.8710	9.3487																																																																																							
	Estimate	Std. Error	t value	Pr(> t )																																																																																							
(Intercept)	1.226e+02	1.344e+01	9.123	8.31e-13 ***																																																																																							
sqrt(`Goals Scored`)	8.228e+00	7.989e-01	10.300	1.03e-14 ***																																																																																							
log(`Goals Conceded`)	-3.387e+01	2.610e+00	-12.980	< 2e-16 ***																																																																																							
I(`Shots pg`^2)	3.504e-03	1.312e-02	0.267	0.790																																																																																							
I(`Possession%`^3)	2.094e-05	3.133e-05	0.669	0.506																																																																																							
I(`Pass%`^3)	9.690e-08	1.195e-05	0.008	0.994																																																																																							
	<p>In the transformed full model, the residual standard error has decreased and the R^2 has increased.</p>																																																																																										

## Proposing a Model:

We decided to perform the best subsets algorithm on the full transformed model of:

$\text{Points} \sim \sqrt{\text{Goals Scored}} + \log(\text{Goals Conceded}) + (\text{ShotsPG})^2 + (\text{Possession\%})^3 + (\text{Pass\%})^3$

The categorical model of “League” was omitted from the best subsets because it is the only categorical variable in our data set and needed to be included in our model anyways.

```
> bs
      sqrt..Goals.Scored.. log..Goals.Conceded.. I..Shots.pg..2. I..Possession...3.
1 ( 1 ) *
1 ( 2 ) *
2 ( 1 ) * *
2 ( 2 ) * *
3 ( 1 ) * * *
3 ( 2 ) * *
4 ( 1 ) * * *
4 ( 2 ) * * *
5 ( 1 ) * * *
      I..Pass...3. Pars      Cp      AIC      BIC      Rsq      Rsq.adj      s
1 ( 1 )      2 203.360883 273.2275 269.2275 0.7693958 0.7656763 8.324592
1 ( 2 )      2 211.345040 275.1389 271.1389 0.7624047 0.7585725 8.449836
2 ( 1 )      3  1.433691 179.9530 173.9530 0.9479586 0.9462524 3.986893
2 ( 2 )      3 129.390307 253.4463 247.4463 0.8359172 0.8305374 7.079325
3 ( 1 )      4  2.072941 180.4707 172.4707 0.9491501 0.9466076 3.973694
3 ( 2 )      *  4  2.567773 181.0137 173.0137 0.9487168 0.9461527 3.990588
4 ( 1 )      5  4.000066 182.3903 172.3903 0.9492139 0.9457708 4.004713
4 ( 2 )      *  5  4.071310 182.4689 172.4689 0.9491516 0.9457042 4.007172
5 ( 1 )      *  6  6.000000 184.3902 172.3902 0.9492140 0.9448359 4.039087
```

Based on the subsets output, the best model seems to be 3(1) as it has:

1. A CP close to the # of Pars
2. The second smallest CP
3. The 2nd smallest AIC
4. A small BIC
5. A large  $R^2$
6. The largest  $R^2$  adjusted
7. The smallest s

\*We also considered model 2(1) because it had similar statistics, but decided to go with model 3(1) because it had a slightly better  $R^2$  adj and slightly smaller BIC.

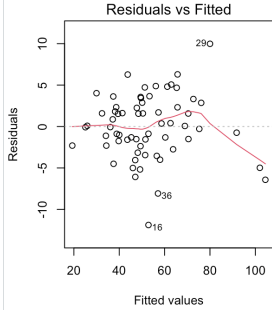
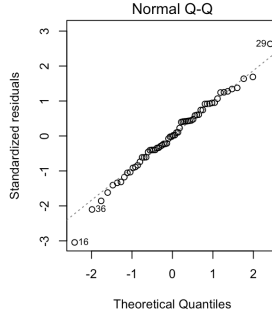
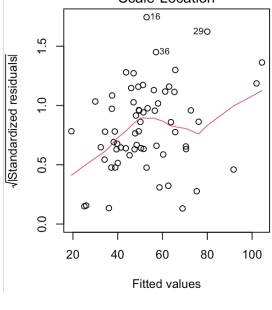
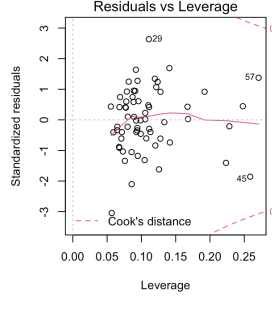
This means that the best model for this data set will be:

**$\text{Points} \sim \sqrt{\text{Goals Scored}} + \log(\text{Goals Conceded}) + (\text{Possession\%})^3 + \text{League}$**



## VI. Residual Analysis:

Residual plots:

Residuals vs. Fitted	Normal Q-Q	Scale-Location	Residuals vs. Leverage
			
<p>The equal variance assumption looked good. There were a couple negative outlying points to the right of the residuals plot but we tried all the transformations we could. The linearity looked mostly good after these transformations.</p>	<p>There is very little deviation from the line, meaning that the normality assumption appears to be met.</p>	<p>The scale-location plot showed equal variance and linearity assumption mostly satisfied. There was a slight pattern shown by the trendline but it was much improved from before the transformations.</p>	<p>The residuals vs. leverage plot showed a few outliers in the x and y direction but no points appeared to cross the Cook's distance line indicating influence. Overall it looked pretty good.</p>

## VII. Fit a Linear Model :

Model = Points ~ sqrt(Goals Scored)+ log(Goals Conceded) + (Possession %)^3 + League

Equation:

**Points-hat = 124.1 + 8.197 \* sqrt(goals scored) - 34.17 \* log(goals conceded) + .00002381 \* (Possession%)^3 + -.1074 \* (League = Ligue 1) + 1.336 \* (League = Premier) - .3838 \* (League = Serie A)**

The implications of our model was that points increases with goals scored, decreases with goals conceded, and increases with possession%. All three of these interpretations made sense contextually as goals win matches and possession increases your chances of scoring goals. The league coefficients indicated that there is a slight increase or decrease in points between each league. This also made sense contextually given that each league is composed of completely different teams. The intercept of 124.1 did not make much sense contextually because it would indicate a point total of 124.1 for a team that has 0 in all of its statistics. However, we are trying to fit a model for teams that played 38 games in a season, and no team who played 38 games would feasibly have those stats, so we don't need to pay much attention to the intercept.

The overall behavior of the model was very good. The coefficients all made sense given the context of the model. The model explained a very large portion of the variability in the response. The R^2 value was .9506, indicating that the percentage of variation in number of points explained by adding all 4 predictors to the model is 95.06%. The s value was 4.02 indicating the "typical" error using the model was 4.02 points. Given that the mean number of points for a given team in our sample was 52.57 points, the s value of 4.02 appeared very small, further confirming the usefulness of the model.

Intercept:

For a team in the La Liga league with 0 goals scored, 0 goals conceded, and 0 possession%, the predicted number of points is 124 (\*extrapolation because we are only looking at teams that played a full 38 games).

Quantitative variable:

The slope coefficient for  $B(\log(\text{Goals Conceded}))$  was - 34.17. If we multiply the number of goals conceded by  $e$  (the base of the natural log), we predict a 34.17 decrease in the predicted number of points while holding goals scored, possession%, and league constant.

Categorical variable:

The slope coefficient for  $B(\text{League} = \text{Premier})$  was 1.336 using La Liga as the reference group. For a fixed number of goals scored, goals conceded, and possession%, we predict the mean number of points for Premier League teams to be 1.336 points higher than La Liga.

Multicollinearity:

Variable	VIF
Sqrt(goals scored)	2.466
Log(goals conceded)	1.856
(Possession%)^3	2.477
League	1.057

The VIF values were all well under the threshold of 5 or 10, indicating no issues with multicollinearity.

## VIII. Statistical Inference:

### 1. Overall Model

Ho:  $\beta(\text{sqrt}(\text{Goals Scored})) = \beta(\log(\text{Goals Conceded})) = \beta(\text{Possession\%}^3) = \beta(\text{League})$

Ha: At least one of the  $\beta_i$ 's  $\neq 0$

Test Statistic: 182.6, DoF: 6 and 57, p-value:  $< .001$

At a significance level of .05, with such a small p-value, we reject the null hypothesis. This means that we have enough evidence to conclude that  $\text{sqrt}(\text{Goals Scored})$ ,  $\log(\text{Goals Conceded})$ ,  $\text{Possession\%}^3$ , and League are all statistically significant predictors of Points for teams in the English, Spanish, Italian, and French top divisions in the 2018-2019 season.

### 2. Partial F-Test

Analysis of Variance Table

Model 1: Points ~ sqrt(`Goals Scored`)

Model 2: Points ~ sqrt(`Goals Scored`) + log(`Goals Conceded`) + I(`Possession%^3`) +  
League

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	62	4296.5				
2	57	921.2	5	3375.4	41.772	< 2.2e-16 ***

---

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

After conducting a partial F-test comparing our model to a model with only one explanatory variable, at a significance level of .05, with such a small p-value, we reject the null hypothesis. This means that we have enough evidence to conclude that including  $\log(\text{Goals Conceded})$ ,  $\text{Possession\%}^3$ , and League is a statistically significant improvement on the model.

### 3. Interaction

We performed the following hypothesis test using ANOVA type I between a model including the interaction terms between league and goals scored and the model without:

Ho:  $\beta(\text{Ligue1} * \text{Goals Scored}) = \beta(\text{Premier} * \text{Goals Scored}) = \beta(\text{SerieA} * \text{Goals Scored}) = 0$

Ha: At least one of the  $\beta_i$ 's  $\neq 0$

Df: 3 and 56 DF, F-statistic: 3.275, P-value = .02759

Conclusion: At a .05 significance level, there is strong evidence that the difference in slopes for each league is significant. Thus we conclude the effect of goals scored on points is significantly different depending on the league. This confirms our prediction from our interaction graph.

### 4. CI's

The 2003-2004 Arsenal team, nicknamed "The Invincibles" went the entire season without losing a game and are widely regarded as one of the greatest teams in history. We wanted to see how many points they would've finished with in the 18/19 season based on our model. The Invincibles had 73 Goals Scored and 26 Goals conceded that year.

	fit	lwr	upr
1	83.68208	81.36242	86.00174

	fit	lwr	upr
1	83.68208	75.08669	92.27747

We are 95% confident that the mean number of points for a team with 73 goals scored and 26 goals conceded is between 81.36 and 86 points.

We are 95% confident that the number of points for a team with 73 goals scored and 26 goals conceded will fall between 75.09 and 92.28 points.

In reality, the Invincibles totalled 90 points that season, which falls outside of the confidence interval, but within the prediction interval. One reason why this point total is higher than the midpoint could be because the Invincibles won a lot of games with the scoreline 1-0. This means that they didn't have as many Goal Scored, but won games anyways.

### Confidence Intervals for the Model Coefficients:

a. (Intercept) 9.793456e+01 1.502320e+02

We are 95% confident that the mean point total for a team with 0 Goals Scored, Goals Conceded, and Possession % will be between 97.93 and 150.2 for teams in the English, Spanish, Italian, and French top divisions in the 2018-2019 season.

b. sqrt(Goals\_scored) 6.707873e+00 9.686520e+00

We are 95% confident that after adjusting for Goals Conceded, Possession %, and League, for every one point increase in Point total, there will be an expected increase in Goals Scored between  $\sqrt{6.71}$  and  $\sqrt{9.69}$  for teams in the English, Spanish, Italian, and French top divisions in the 2018-2019 season.

c. log(Goals\_conceded) -3.940167e+01 -2.893030e+01

We are 95% confident that after adjusting for Goals Scored, Possession %, and League, for every one point increase in Point total, there will be an expected decrease in Goals Conceded between  $\log(28.93)$  and  $\log(39.4)$  for teams in the English, Spanish, Italian, and French top divisions in the 2018-2019 season.

d. I(Possession%<sup>3</sup>) -1.553287e-05 6.314660e-05

We are 95% confident that after adjusting for Goals Scored, Goals Conceded, and League, for every one point increase in Point total, there will be an expected change in Possession% between an  $-(.0000155^3)$  decrease and an  $(.000063^3)$  increase for teams in the English, Spanish, Italian, and French top divisions in the 2018-2019 season.

e. LeagueLigue 1 -3.059297e+00 2.844531e+00  
LeaguePremier -1.713682e+00 4.385141e+00  
LeagueSerie A -3.158949e+00 2.391283e+00

We are 95% confident that the mean difference in points is between -3.059 lower and 2.844 higher for Ligue 1 compared to La Liga in the 2018-2019 season.

We are 95% confident that the mean difference in points is between -1.714 lower and 4.385 higher for the Premier league compared to La Liga in the 2018-2019 season.

We are 95% confident that the mean difference in points is between -3.159 lower and 2.39 higher for Serie A compared to La Liga in the 2018-2019 season.

## IX. Model validation

```
> fit = lm(Points ~ sqrt(`Goals Scored`)+ log(`Goals Conceded`) + I(`Possession%`^3) + League, data=train
data)
> predicted=predict(model, testdata)
> actual= testdata$Points
> MSPE=mean((predicted-actual)^2)
> MSPE
[1] 9.794974
```

After applying the model to the test data, we calculated the MSPE to be 9.795.

```
> MSE1 = (anova(fit))$'Mean Sq'[length((anova(fit))$'Mean Sq')];MSE1
[1] 16.16075
```

We found the MSE of our model to be 16.161

The fact that the MSPE is much smaller (and closer to 0) than the MSE means that the predictive ability of the model is even better for our test data. This validates the predictive capability of the model.

## X. Conclusion

After analyzing the data, we learned that the model with goals scored, goals conceded, possession%, and league was very efficient at predicting the number of points. The capability of our model can be seen in the extremely large  $R^2$  value of .9506. The goals scored and goals conceded predictors appeared to be explaining a large portion of the variability in points, while possession% and league added in additional predictive capability to the model. For these reasons, we found our model to be significant. After applying the model to the test data, the validity beyond our sample was confirmed.

Our model showed that points increases with goals scored, decreases with goals conceded, and increases with possession%. None of these relationships were perfectly linear, but we were able to apply transformations in order to fit a valid model. Our model also showed that the number of points differs across leagues, and the interaction term between goals scored and league proved to be significant. Thus, we concluded that the effect of goals scored was significantly different across the leagues. Given more time, we would like to continue to explore the relationship between league and the other predictors.

Our model also had some weaknesses. According to our best subsets analysis, the model with only goals scored and goals conceded had fairly similar  $C_p$ , AIC, BIC, etc. We chose the model including possession% because it explained slightly extra variability, but the significance of adding this term to the model is debatable because a simpler model can be better.

Our model also showed some weakness in the residual plots. There appeared to still be some pattern in the residuals vs. fitted plot after we had tried all possible transformations.

We also had a relatively limited amount of test data (only 16 teams), so we weren't fully confident in the model validation.

Given more time to work on the project, we would probably look into more variables including conditions (dry vs. wet vs. snow), fouls per game, dribbles per game, crosses per game, and shot location. We feel that these additional variables could reveal more insights into predicting a team's success during the season.



