English Premier League - What Makes a winning team?

Anthony Meyers, Danny Moncada, & Frankie Cancino

Introduction

Predicting the winner of a single soccer match is notoriously difficult. Weather conditions, the location of the game, individual skill, the performance of the referee, and now with the advent of Video Assistant Referee (VAR), performance of technology can all affect the outcome of the match. Predicting the winner of the English Premier League is even more difficult, because you have to account for all of these factors over the course of a thirty eight game season. Many unknowns can affect a single result; injuries, player form, transfers, player suspensions, even the galvanizing effect of losing a manager in the middle of season can transform a team from a contender to out of the title race over the span of a weekend.

With this in mind, one of the biggest challenges facing any prospective advertiser is who to "bet" on prior to the start of a new league season, and determine what factors influence a team's final standing in the league table, especially the "top four". Why are the teams in the top four so important? This grants these teams automatic entry into the group stages of the UEFA Champions League, the annual club football competition organized by the Union of European Football Association. This is the most prestigious club competition in European football, and the UEFA Champions League final is the most watched annual sporting event worldwide. If a team qualifies for the group stage, they will earn €15,250,000, and an additional €2,700,000 for each additional victory!

Additionally, sponsorship deals are constantly being updated for the top performing teams, as the exposure that these teams have in the USA and across the globe is increasing every year. For example, NBC Sports reported in May 2018 that the average television audience for a Premier League match increased from 447,000 in the 2016-2017 season to 449,000 viewers for the current season.

Once you've identified the top teams of the league, the next step is determining what matches would be the best for an advertiser to generate interest in their product and market themselves well; matches between the two best teams will draw a lot of interested football fans! Being able to understand the different variables will allow for effective deployment of marketing dollars.

In order to conduct a thorough analysis, we had to look at what data exists, and luckily for us, football provides a wealth of information for each match. This includes things

like goals scored, yellow/red cards received, number of saves per match, number of shots taken by each time. With all of that information available, we determined that a multiple linear regression model would be the best way to investigate what factors influence a football team's final position at the end of a campaign. Once we've determined a good model, we can then use that to draw some conclusions and then iterate over that model. Iterations would include taking a look at new data, re-fitting models based off of this new data, choosing different k-values to enhance our feature selection, and experimenting with different methods such as logistic regression.

Our data set comes from a report that contains results for a time period of the last 10 seasons of the English Premier League, including the current season; this gives us a sample of over 3000 matches to look at. The data contained few outliers and little to no missing data; with a plethora of data and no abnormalities in the variables being used as predictors, this made it an ideal data set. When determining appropriate predictors, we looked at a variety of factors including goals for and against, total shots and shots on target, fouls for and against, and yellow/red cards for and against.

Analysis

In order to formulate an appropriate model, we decided the best approach would be to aggregate the results over the course of a season, rather than look at it game by game. Over the course of a game, the only real variables that determine the outcome of a game are goals for and goals against, which is to say, a team will win 100% of the time if they simply outscore their opposition - no need to interpret anything here! In order to be able to extrapolate the results and build an accurate model, we decided to take our data set and group it by team, and then look at team by team results.

When we started looking at different models*, we determined that the best method would be to start with the most complex model and then slowly whittle it down until we felt comfortable it would be cohesive and allow us to tell the appropriate story. Have all the predictors in one model was simply too much to allow for us to gleam any meaningful information from it. We took that initial model and determined that goals and shots were probably the most important factors in determining the outcome of a match. However, we noticed that there was a heavy degree of collinearity (found in the appendix) between goals and shots; that is, there is a direct relationship between the two factors. Intuitively, this makes sense; it is extremely difficult to score a goal without taking a shot!

* This table demonstrates the different approaches we took to come up with our final model

model	k	r-squared	s	p-value
All predictors	22	0.8581877	2.325087	2.20E-16
Goals and Shots	8	0.8510219	2.28347	2.20E-16
Goals	4	0.8492233	2.270806	2.20E-16
Net Goals	1	0.8442622	2.288332	2.20E-16

From there, we determined that net goals was probably the most accurate predictor, as it had very similar summary statistics to the previous models. However, it did not provide a lot of actionable feedback on how to rank different teams; naturally a team that had more goals scored than goals conceded were likely to win their matches and thus climb up to the top of the table. For example, a team could be involved in high scoring affairs, winning their games 5-4, which did not really help us determine what makes an effective team. After further analysis, we decided that measuring goals scored at home and away matches, and goals allowed at home and away matches would provide us with the most accurate picture*:

* This output demonstrates that 85% of the varia tion in the final standing of a team at the end of a league season is explained by that team's total goals for and goals against, split by home and away matches

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.87444    1.42615  -5.521 1.20e-07 ***
Home_GF      0.15763    0.02390   6.596 4.82e-10 ***
Home_GA     -0.25643    0.03165  -8.101 9.01e-14 ***
Away_GF      0.20166    0.02851   7.072 3.51e-11 ***
Away_GA     -0.20474    0.02838  -7.214 1.58e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.271 on 175 degrees of freedom
Multiple R-squared:  0.8492,    Adjusted R-squared:  0.8458
F-statistic: 246.4 on 4 and 175 DF,  p-value: < 2.2e-16
```

We use the 90% confidence interval to define the range in which the true population effect of the predictors will lie.

				90% Confidence Interval	
Predictor	**Description**	**β**	**Estimate**	**5%**	**95%**
Home_GF	Goals for at home stadium	β1	0.15763	0.11811	0.19714
Home_GA	Goals against at home stadium	β2	-0.25643	-0.30877	-0.20408
Away_GF	Goals for at away stadiums	β3	0.20166	0.15450	0.24881
Away_GA	Goals against at away stadiums	β4	-0.20474	-0.25167	-0.15781

Conclusion

The estimates from the linear model provide predicted value changes in the final standing of a team based on goals for and goals against as the home and away team. These predictors can be utilized to efficiently deploy marketing capital to the team segments, like the cities and their fan bases to maximize earned revenue or to maximize

broadcasting capital to *potential* high profile matches. Based on the final model using goals scored and against from our sample, we determined that an away goal is 27.9% more valuable than a home goal to a higher finish. Likewise, conceding a goal at home is 25.2% more damaging than conceding a goal away.

Therefore, our recommendation is to consider heavier marketing towards teams that score at high rates as the *visiting team* and that do not concede goals *at home*. With further research and data, predictors of the goals for and goals against predictors could supplement the analysis and allow for prediction and forecasting capabilities. Further evaluation of market research can compliment analytical approaches for counter cases that fall out of the scope of our analysis results.

References

Data Set:

support@datahub.io . (2018, October). *English Premier League (football).* Retrieved from https://datahub.io/sports-data/english-premier-league

NBC Sports Coverage:

NBC Sports. (2018, May 17).*Premier League Viewership Highlights.* Retrieved from http://nbcsportsgrouppressbox.com/2018/05/17/record-39-3-million-americans-tuned-into-nbc-sports-coverage-of-2017-18-premier-league-season-on-the-networks-of-nbcuniversal/

Champions League Prize Money:

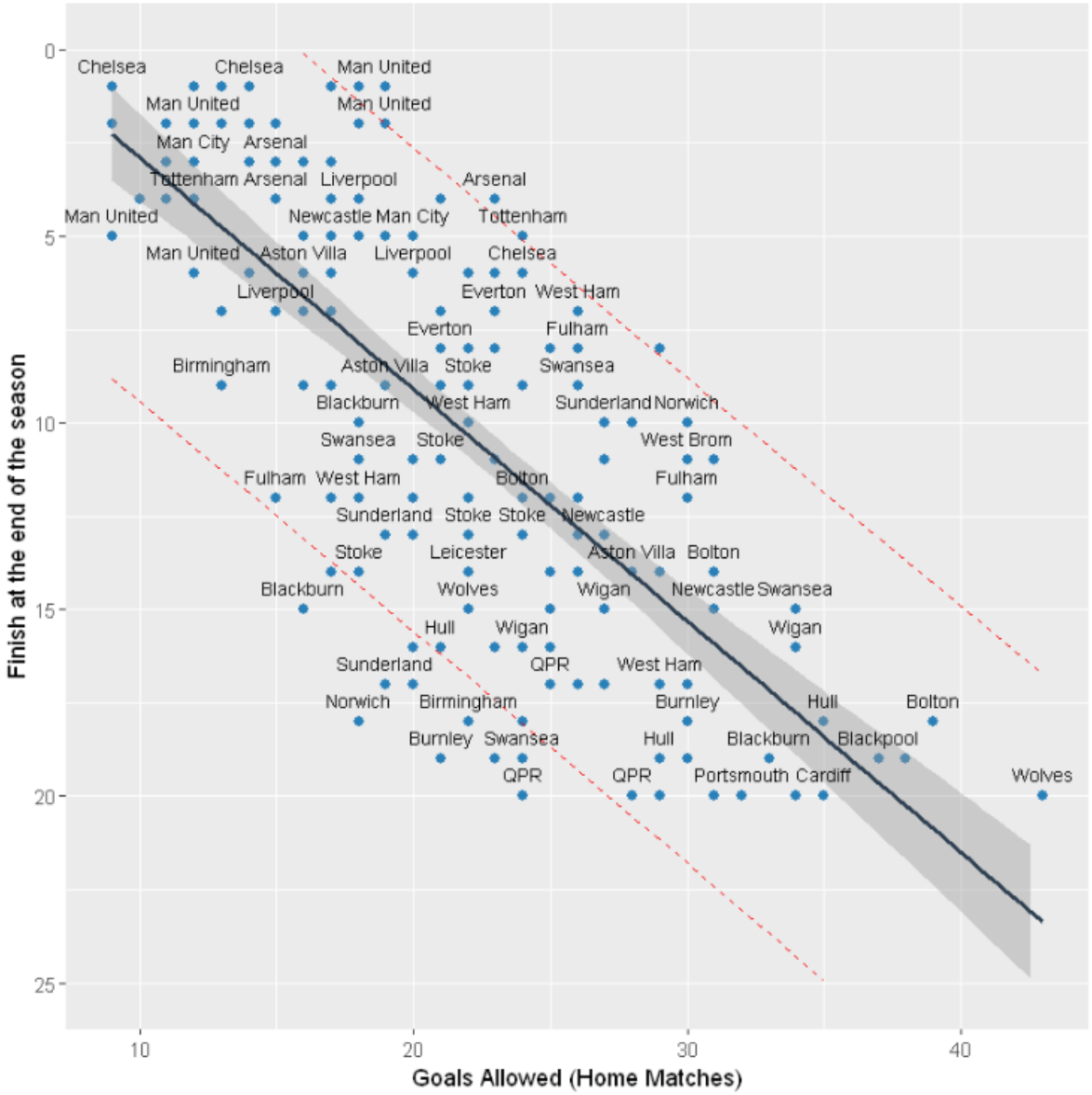
UEFA. (2018, June 12).*How clubs' 2018/19 UEFA Champions League revenue will be shared.* https://www.uefa.com/uefachampionsleague/news/newsid=2562033.html#/2562033

Champions League Final Viewership:

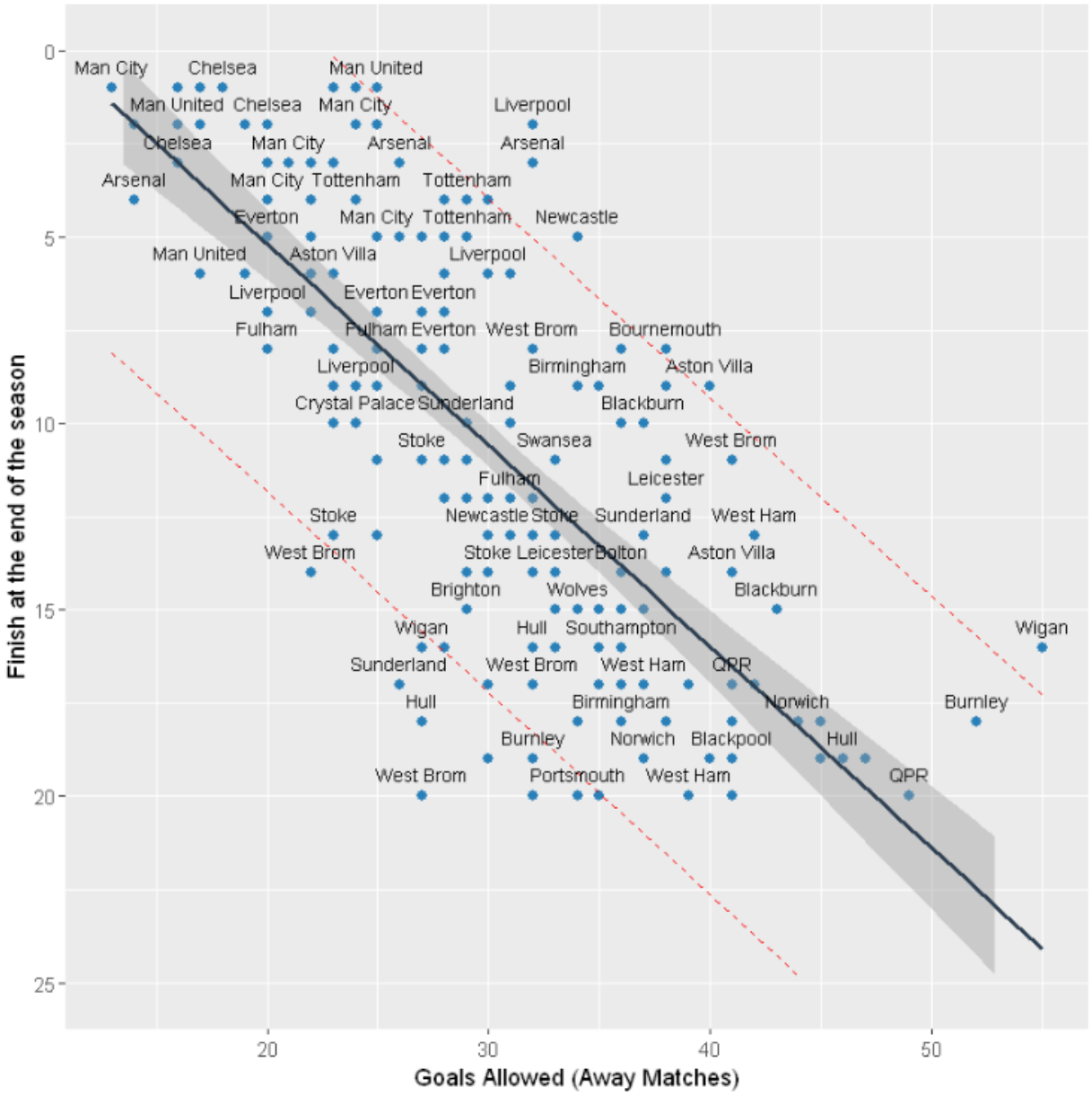
UEFA. (2018, May 23).*Wembley final proves global pulling power.* Retrieved from https://www.uefa.com/uefachampionsleague/news/newsid=1957523.html

Appendix

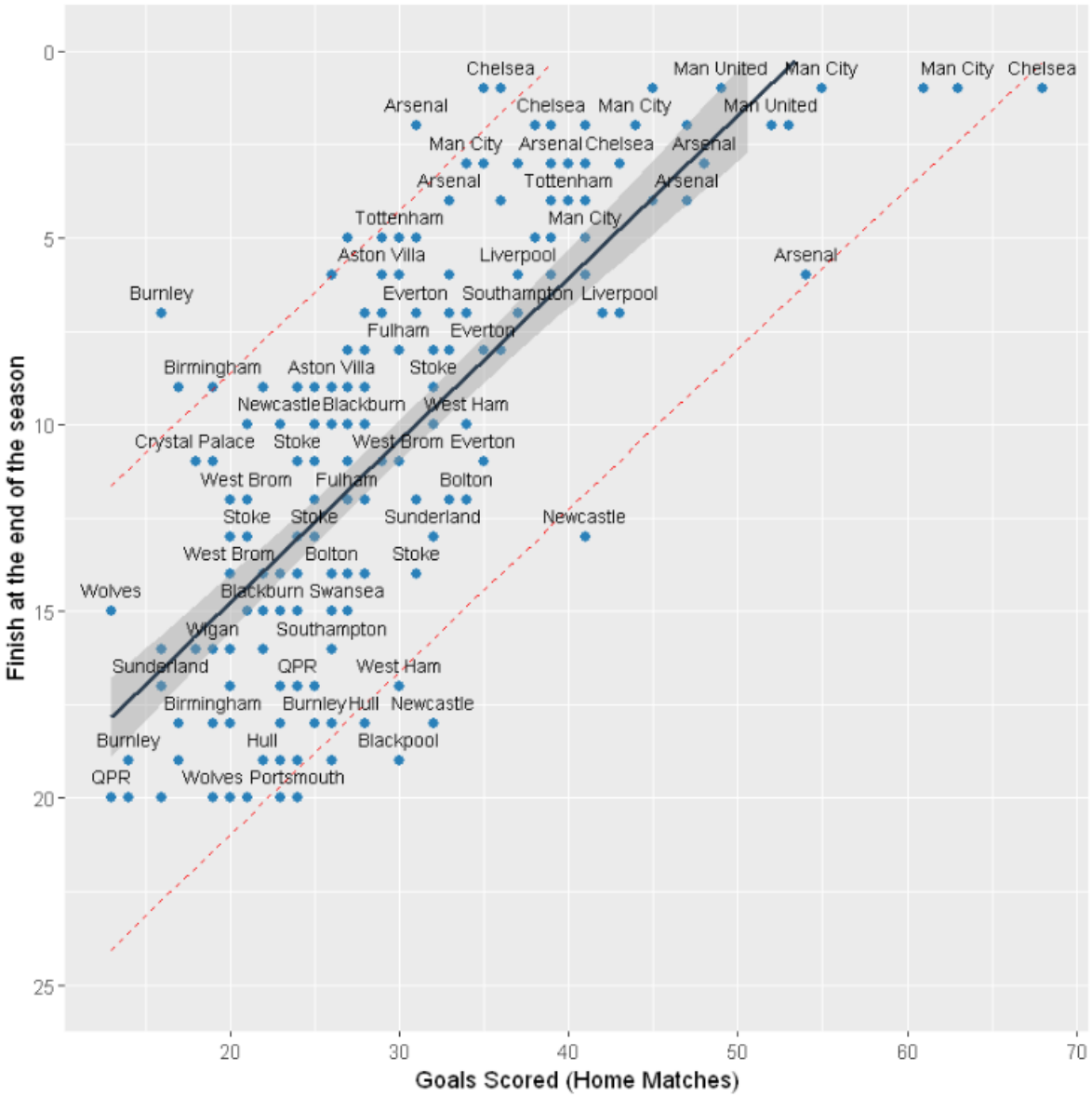
Linear regression model of goals conceded at home matches and the finishing position at the end of the season:



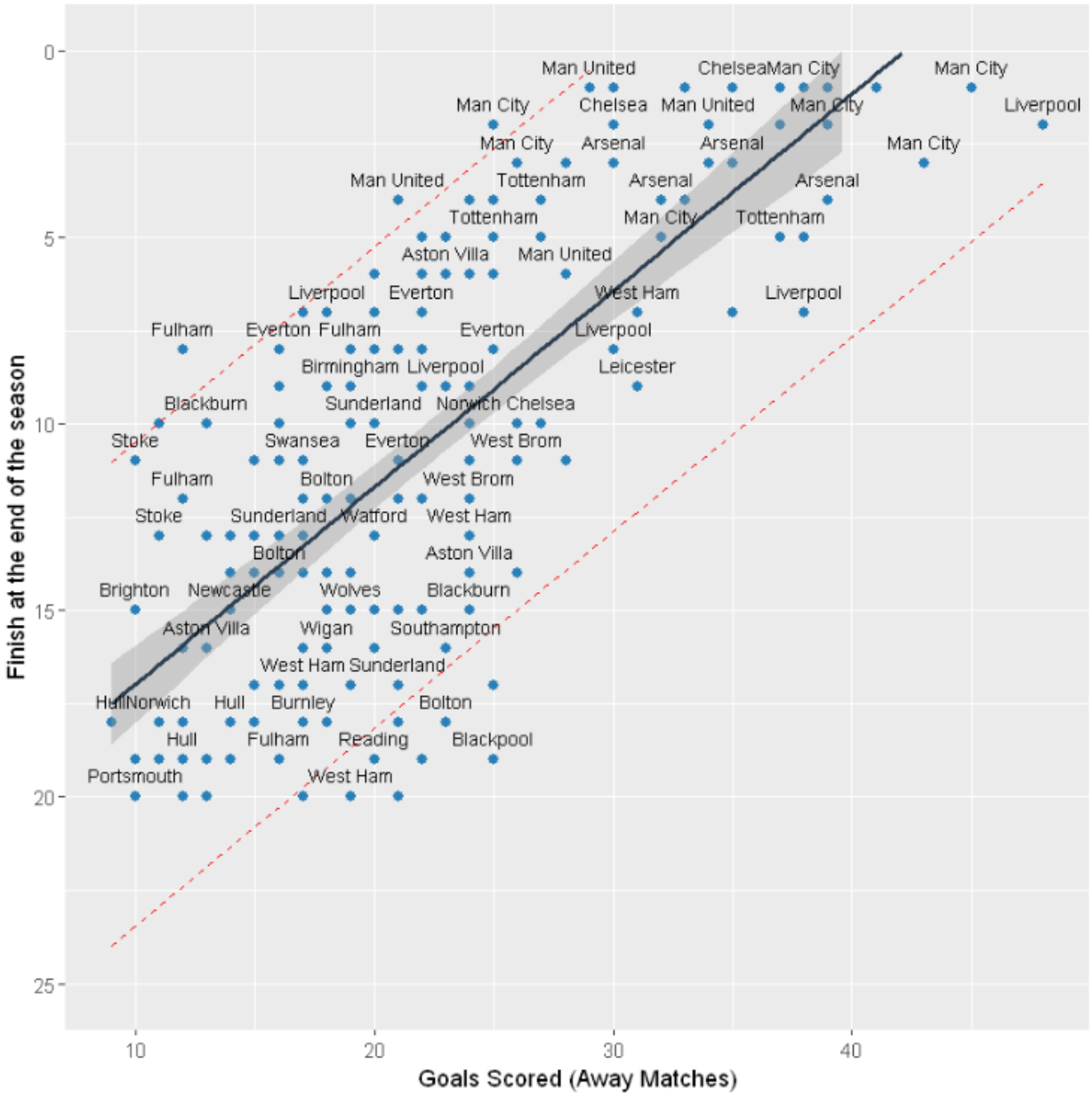
Linear regression model of goals conceded at away matches and the finishing position at the end of the season:



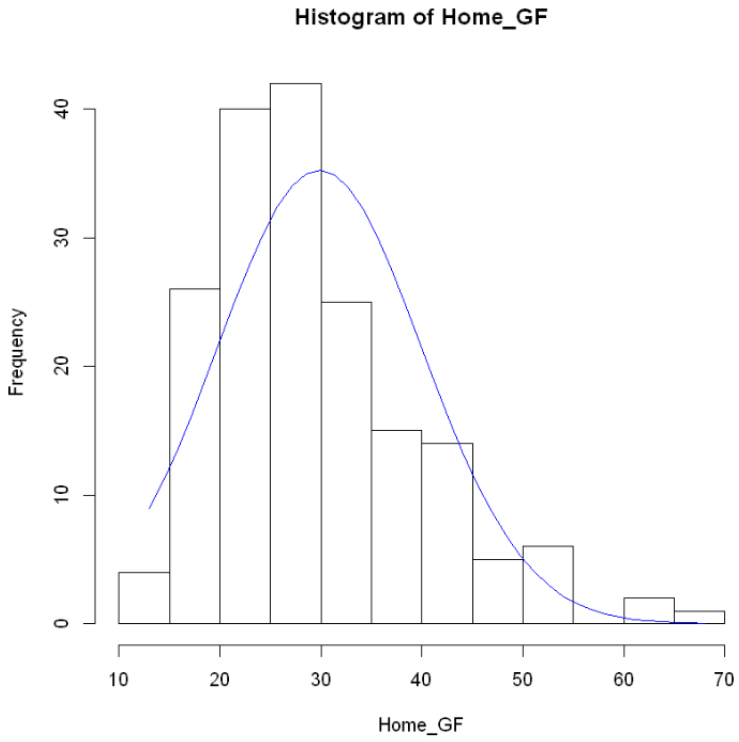
Observing linear regression model of goals scored at home and the total number of points at the end of the season:

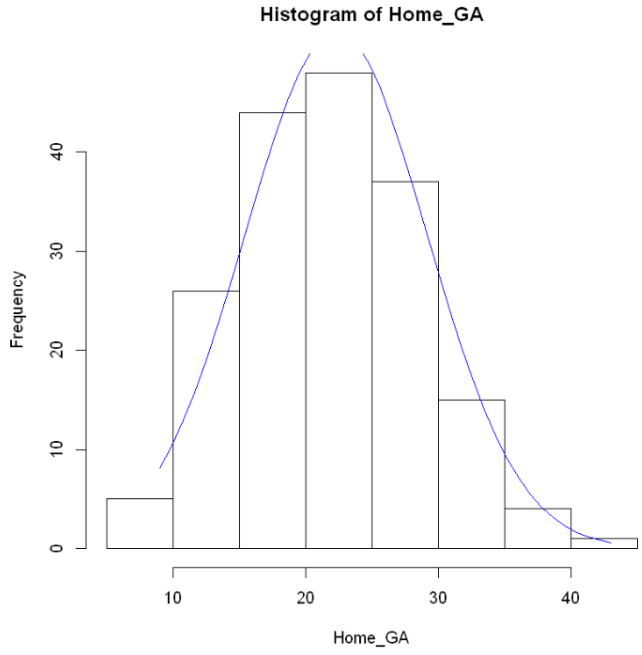


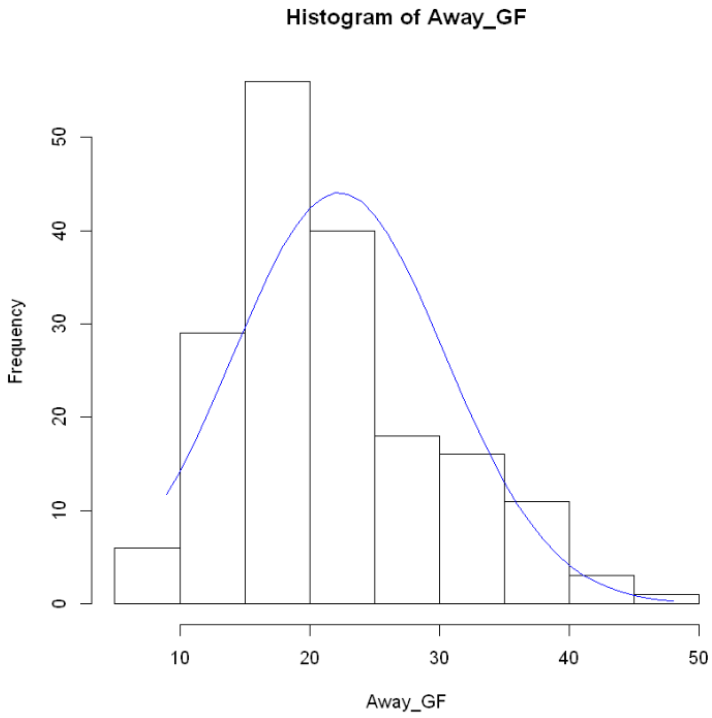
Linear regression model of goals scored at away matches and the total number of points at the end of the season:

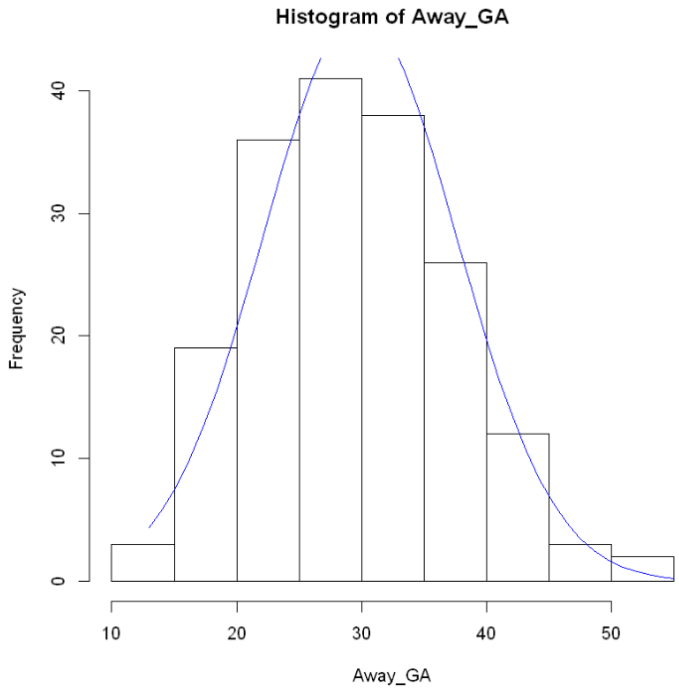


Checking normalcy for selected predictors:

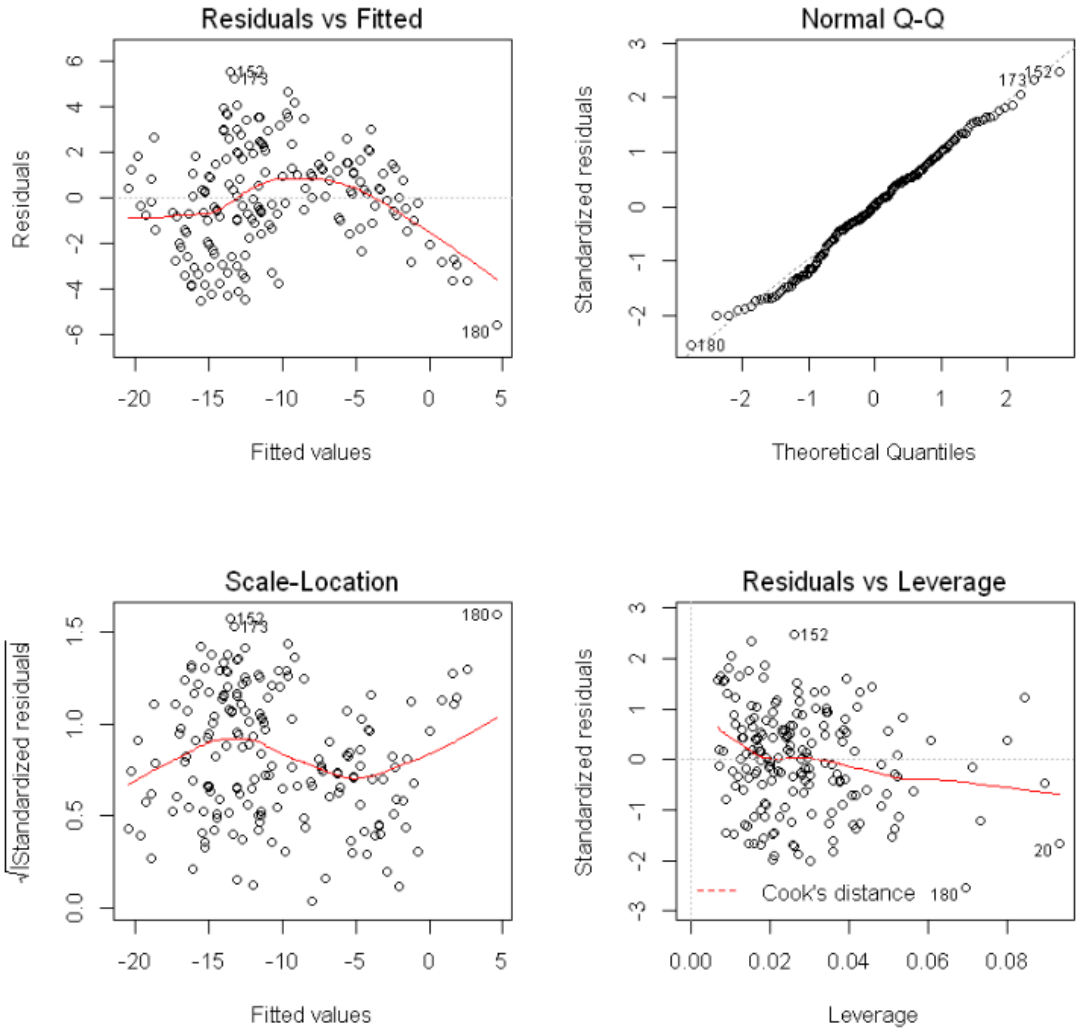

Histogram of Home_GF
Frequency
0
10
20
30
40
10
20
30
40
50
60
70
Home_GF


Histogram of Home_GA
Frequency
0
10
20
30
40
10
20
30
40
Home_GA





Confirming normalcy for our selected model and the factors:



Anova table analysis to det ermine how much of the variance of a team's finish position is contributed by our predictors :

```
Analysis of Variance Table

Response: -Finish
           Df Sum Sq Mean Sq F value    Pr(>F)
Home_GF     1 3520.7  3520.7 682.764 < 2.2e-16 ***
Home_GA     1 1006.4  1006.4 195.176 < 2.2e-16 ***
Away_GF     1  287.1   287.1  55.675 3.840e-12 ***
Away_GA     1  268.4   268.4  52.042 1.579e-11 ***
Residuals 175  902.4     5.2
```

Testing our assumption regarding correlation between Shots and Goals Scored from Home and Away matchest:

	Home_GF	Home_ShotsFor	Home_GA	Home_ShotsAgainst
Home_GF	1.0000000	0.7131591	-0.5009005	-0.6062895
Home_ShotsFor	0.7131591	1.0000000	-0.4793260	-0.6354719
Home_GA	-0.5009005	-0.4793260	1.0000000	0.6268720
Home_ShotsAgainst	-0.6062895	-0.6354719	0.6268720	1.0000000

	Away_GF	Away_ShotsFor	Away_GA	Away_ShotsAgainst
Away_GF	1.0000000	0.7675868	-0.4525418	-0.5853232
Away_ShotsFor	0.7675868	1.0000000	-0.5485788	-0.6990074
Away_GA	-0.4525418	-0.5485788	1.0000000	0.5977395
Away_ShotsAgainst	-0.5853232	-0.6990074	0.5977395	1.0000000

Testing our assumption regarding correlation between Home and Away Goals Scored/Against:

```
            Home_GF    Away_GF    Home_GA    Away_GA
Home_GF   1.0000000  0.6512784 -0.5009005 -0.5497240
Away_GF   0.6512784  1.0000000 -0.5015553 -0.4525418
Home_GA -0.5009005 -0.5015553  1.0000000  0.5524449
Away_GA -0.5497240 -0.4525418  0.5524449  1.0000000
```