Final Report

Aiden Santoro and Daniel Berry

2025-04-17

Final Report

Introduction and Data

```
# A tibble: 6 x 22
    Pos Team              M     W     D     L    GF    GA   Dif Points   GPM  GAPM
  <dbl> <chr>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
1     1 Liverpool ~    19    18     1     0    48    14    34     55  2.53  0.74
2     2 Leicester ~    19    12     3     4    41    18    23     39  2.16  0.95
3     3 Manchester~    19    12     2     5    52    23    29     38  2.74  1.21
4     4 Chelsea FC     19    10     2     7    33    27     6     32  1.74  1.42
5     5 Wolverhamp~    19     7     9     3    29    24     5     30  1.53  1.26
6     6 Tottenham ~    19     8     5     6    34    27     7     29  1.79  1.42
# i 10 more variables: Final_Points <dbl>, Season <chr>, `#` <dbl>,
#   Win_Percentage <dbl>, Loss_Percentage <dbl>, Draw_Percentage <dbl>,
#   Goals_by_Win <dbl>, GCBW <dbl>, ADR <dbl>, team_category <chr>
```

Research Question

Can a team's performance metrics after the first half of the English Premier League season reliably predict their final standings?

Background and Motivation

The English Premier League is one of the most competitive soccer leagues globally, with performance in the first half of the season often indicative of final standings. By understanding these relationships this report can help teams, analysts, and fans make mid-season predictions and strategic adjustments to ensure their desired outcome. There have been many similar studies that state early seasons performance metrics are strong predictors to end of season standings, but the relationships can vary a lot per league and season. By looking at multiple seasons for the English Premier League this report aims to build a more comprehensive understanding of performance indicators on seasonal outcomes that can be generalized for each year.

Data Description

Source: https://www.worldfootball.net/schedule/eng-premier-league-2019-2020-spieltag/19/

Collection Method: We manually copied the data into an excel file where we performed feature engineering in order to create new variables that would better the performance of our models in the future. From there after cleaning and formatting the data table we saved it as a csv and uploaded it into RStudio.

- **Variables**:
  - **Outcome Variable**: `Final_Points` (total points at the end of the season).
  - **Explanatory Variables** (All from first half of respective season):
    - Please view the code book to view each variable and its description.

Data Wrangling:

Initial Problems:

We came across an initial problem of finding a multiple data sets that would allow us to move further with our research questions. More specifically:

1. We will be going back many years so data sets change from year to year and what they offer will also change. In more recent years there are more in-depth tables describing the league while in years like 2015 they aren't as descriptive.

    1. **Solution:** We are going to bring the years being used for the training data to more present seasons. Starting from the 2019-2020 season until the 2023-2024 will give us a good amount of data in order to accurately predict the final standings.

2. We originally said 15 games but it makes more sense to just do halfway through the season which is 19 games and this data seems to be more readily available.

    1. **Solution:** We were able to find data tables online that give information halfway through the season for each year we are going to use.

Data Cleaning and Variable Creation

For these specific data sets there was no missing information but there were a few things we had to add and change. First off, to pull these data sets into RStudio we copied them into excel, saved them as a csv file, and then uploaded them to our qmd file.

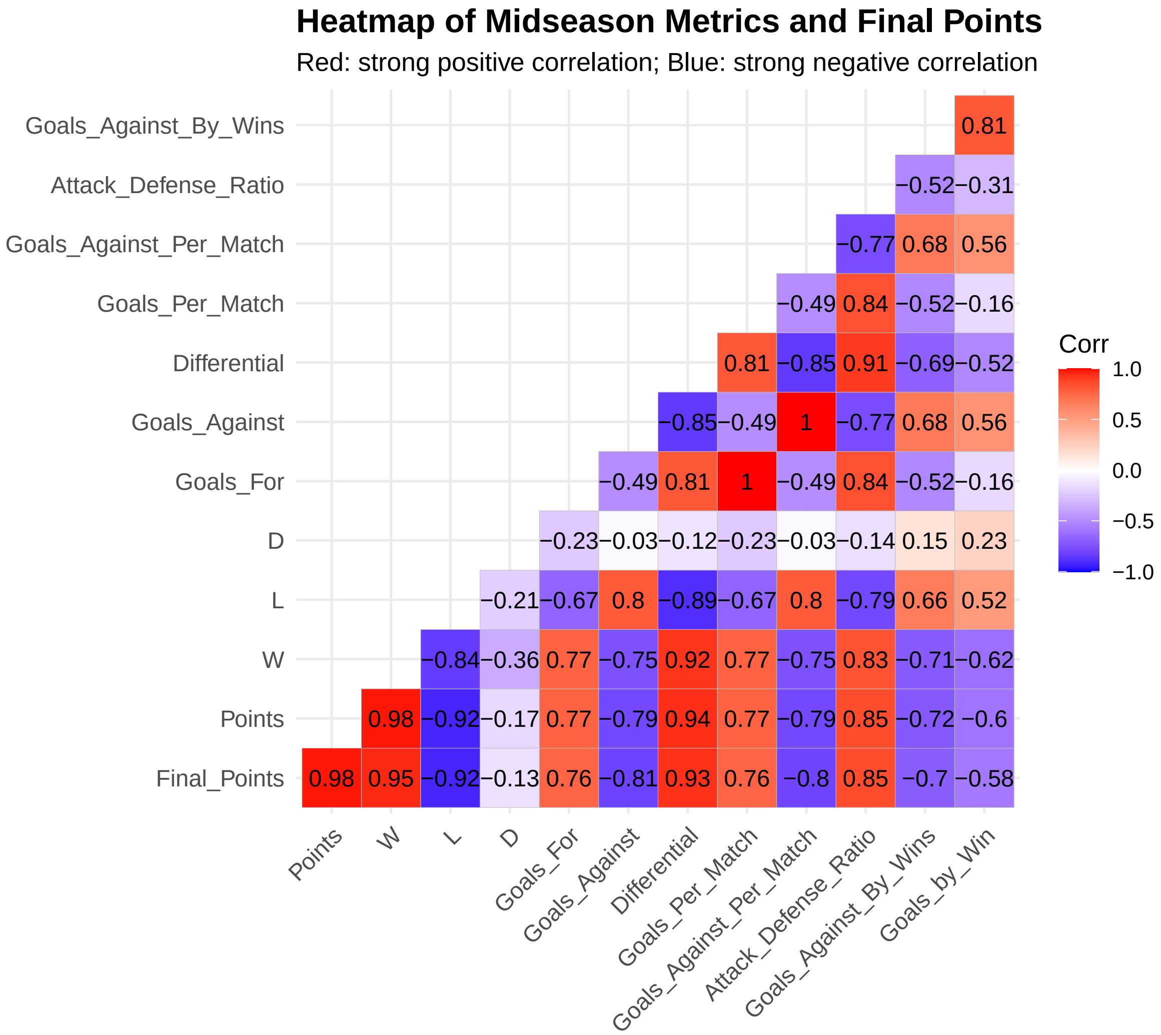
1. Getting rid of the logo column. There was a column in the data set that just consisted of the logo for each team so we deleted that.
2. Changed Column names to something more readable. For example, Goal differentials was taken down in this format originally: 48:22, meaning a team scored 48 and let in 22 goals. We changed it to two different columns GF and GA that has the first number in GF and the second number in GA. Then in a column name Diff we have the differential of those two.
3. We added a few numerical columns to make our data more specific. These include the GF, GA, GPM, and GAPM, Win%, Loss%, Goals_by_Win, GCBW, and ADR.
4. Also we will add a categorical column to categorize teams into either: Aggressive, Dominant, Under performers, and Defensive, based on their GF and GA.
5. Finally we added a Final_Points column that has each teams point total after the season is done. This will be useful in comparing our predictions to the actual final outcomes.

Methodology

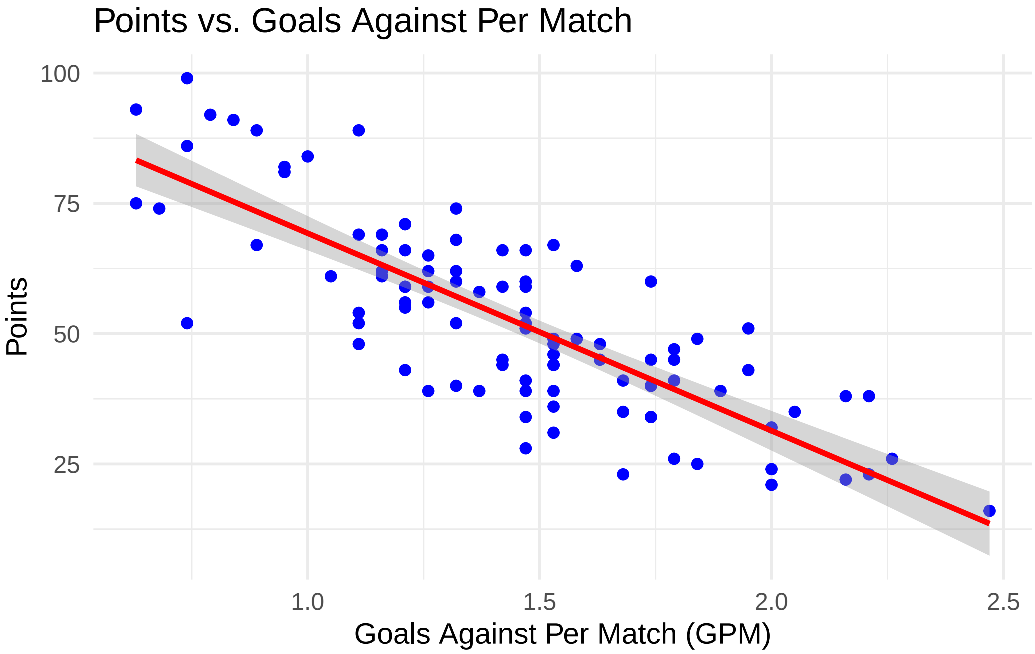
Approach

- **Exploratory Data Analysis (EDA)**:
  - Numerous insights can be drawn from our brief analysis. Firstly, the heat map displayed how a teams performance in the first half of the season is quite similar to that of the second half. This can be inferred as many different variables all had a high correlation with the end of season points. Next, the distribution of team performance is quite variable. Both shown through the summary statistics table and the win percentage distribution graph, the club's performance data were quite spread out with few of them remaining close to the average. So, in this league there can be pure dominance and utter failure rather than most teams simply performing at the average. Finally defensive teams perform a little better than aggressive teams as conceding less goals seems to be more important that scoring goals.
- **Statistical Modeling**: Two Models to quantify the relationship between explanatory variables and `Final_Points`.
- **Validation**: Cross-validation to assess model performance.

Key Visualizations (Reuse from EDA)



The heat map visualization displays the correlation between all major performance variables and the outcome variable, Final Points It clearly shows that goal differential (Dif), mid season points, and goals against (GA) are highly correlated with final standings. For example, Dif shows a correlation of approximately 0.93 with Final_Points indicating that teams with better goal differentials midway through the season tend to finish with more points. Similarly, GA shows a strong negative correlation, suggesting that strong defensive performance early on is a key driver of final outcomes. This analysis confirms that first-half metrics are not only relevant but highly predictive of season-end performance, providing support for building regression models around these features.



The scatter plot comparing Goals Against Per Match (GAPM) with Final_Points shows a clear negative relationship. As GAPM increases—meaning teams are conceding more goals per match—their final point totals tend to decrease. This trend is reinforced by the red linear regression line, which slopes downward, suggesting a strong negative association between defensive performance and season success. This visualization supports the conclusion that solid defensive metrics midway through the season are strong indicators of how a team will perform overall. As such, GAPM is a valuable explanatory variable to include in predictive models for final standings.

Based on the heat map, we selected variables for our linear regression model that showed the strongest correlations with Final_Points. Goal Differential (Diff) had the highest positive correlation, making it a clear choice. Goals Against (GA) showed a strong negative correlation, reinforcing the importance of defensive performance. Additionally, Win_Percentage captured match-level success and also demonstrated a strong positive relationship with final outcomes. These variables were chosen not only for their statistical relationship but also for their interpretability in a soccer performance context.

Model 1: Simple Linear Regression

We created this simple linear regression model using only goal differential (Diff) to serve as a baseline. It allows us to evaluate the individual predictive strength of one key variable and compare it to more complex models with multiple predictors. The goal is to assess how much additional accuracy is gained by incorporating other performance metrics like Goals Against and Win Percentage.

```
Call:
lm(formula = Final_Points ~ Differential, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.7428  -4.5627  -0.6634   4.1808  12.5676 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  52.06270    0.76481   68.07   <2e-16 ***
Differential  1.09242    0.05021   21.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.664 on 74 degrees of freedom
Multiple R-squared:  0.8648,	Adjusted R-squared:  0.863 
F-statistic: 473.4 on 1 and 74 DF,  p-value: < 2.2e-16
```

Model 2: Multiple Linear Regression

```
Call:
lm(formula = Final_Points ~ Differential + Goals_Against + Win_Percentage, 
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.7128  -2.4906   0.1933   2.8244   8.6254 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     37.3884     4.3500   8.595 1.19e-12 ***
Differential     0.3114     0.1087   2.866  0.00545 ** 
Goals_Against   -0.3108     0.1378  -2.255  0.02715 *  
Win_Percentage  60.3391     7.2390   8.335 3.63e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.803 on 72 degrees of freedom
Multiple R-squared:  0.9317,    Adjusted R-squared:  0.9288
F-statistic: 327.3 on 3 and 72 DF,  p-value: < 2.2e-16

R² Score: 0.92

RMSE: 5.79

Linear Regression

100 samples
  3 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 89, 92, 90, 90, 88, 90, ...
Resampling results:

  RMSE      Rsquared   MAE
  5.096393  0.9364352  3.948435

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Justification for Methods

We used multiple linear regression to predict a team's final point total because the outcome variable, Final_Points, is continuous and numerical. Our exploratory analysis showed strong linear relationships between key mid season performance metrics—such as Goal Differential, Goals Against, and Win Percentage—and final standings. Multiple linear regression allowed us to model the combined effect of these variables while maintaining interpretability, enabling us to quantify how each factor contributes to a team's overall success. This method also served as a strong baseline for future comparisons with more complex models.

Along with this, we created a simple linear model before it so we could compare how adding more variables can increase the model's predictive power. In this case, when comparing the two, the multiple regression model clearly outperforms the single-variable model. While the simple model using only goal differential (Diff) explained about 86.5% of the variance in final point totals, the multiple regression model—which also included Goals Against (GA) and Win_Percentage—explained over 93% of the variance and achieved a lower RMSE. This demonstrates that incorporating multiple relevant performance metrics provides a more accurate and reliable prediction of team success, highlighting the value of a multivariate approach.

Model 3: LASSO Regression for Feature Importance

Given that we saw through previous analysis that a lot of these variables are quite successful in determining end of season performance, we wanted to understand which are the best few. In the code chunk below we are setting up a LASSO regression model. LASSO was selected as it will drive less important coefficients to zero. So the model will automatically select the most predictive features. Also it is good to note that in Premier League seasons that data points are not independent, as performance in one season can influence later ones. So by randomly splitting the data we risk data leakage as future information can influence the model. A GroupKFold is used on the seasons variable to ensure all samples from that season stay together. Using GroupKFold in the LASSO regression was important to avoid data leakage across seasons. Also, since we used ElasticNetCV the model also maintained the linear assumptions inherent to LASSO regression. Finally I want to note that we ran this LASSO model before with the inclusion of the Points column and it dominated the feature importance, downweighting a lot of the other features present. So for our actual analysis we excluded it to make insights about other important variables that can provide insights about a club's performance.

```
[-38.60175165  -6.80858178 -18.96284821 -13.09833573 -14.93933311]
```

Average MSE for the models

```
-18.48217009804261
```

Since we are predicting total season points, this implies that the average for our LASSO model's predictions are off by about 4 points per season.

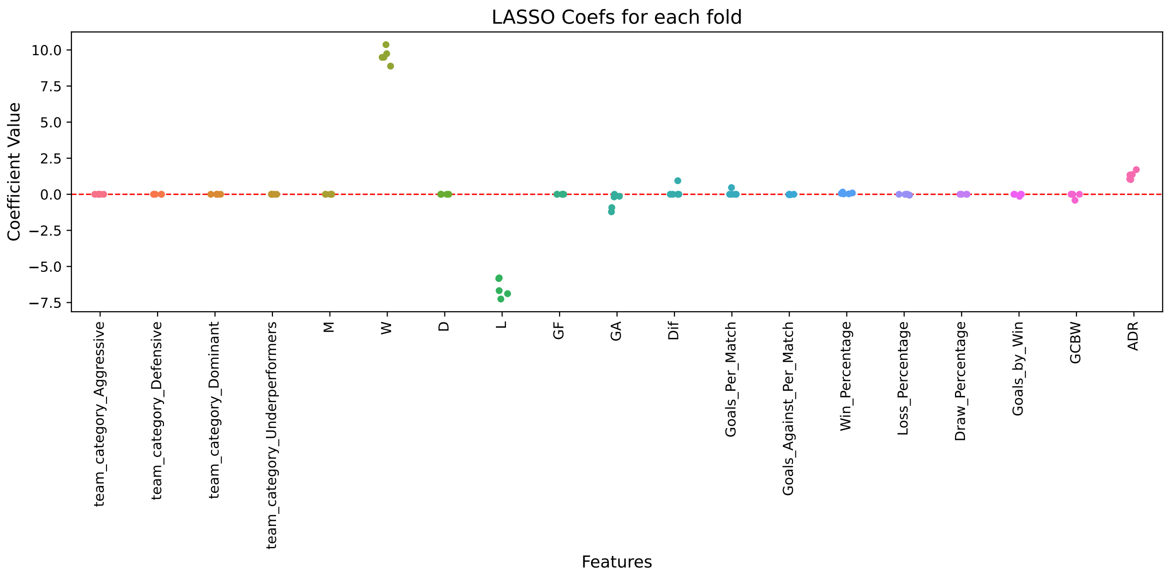
Now we will build a DF to specifically show the importance of each of these features.

```
   team_category_Aggressive  team_category_Defensive  team_category_Dominant  \
0                       0.0                     -0.0                    0.0   
1                      -0.0                     -0.0                    0.0   
2                       0.0                     -0.0                    0.0   
3                       0.0                     -0.0                    0.0   
4                       0.0                     -0.0                    0.0   

   team_category_Underperformers    M          W    D         L   GF  \
0                           -0.0  0.0  10.366332 -0.0 -5.836522  0.0   
1                            0.0  0.0   9.491258 -0.0 -6.675295  0.0   
2                            0.0  0.0   8.879876 -0.0 -6.886656  0.0   
3                            0.0  0.0   9.486415 -0.0 -7.256328  0.0   
```

```
4                            0.0  0.0   9.739292 -0.0 -5.791711  0.0

         GA       Dif  Goals_Per_Match  Goals_Against_Per_Match  \
0 -0.183662  0.000000         0.000000                -0.027626   
1 -0.921992  0.000000         0.000000                -0.022466   
2 -1.214911  0.000000         0.000000                -0.000000   
3 -0.000000  0.000000         0.464698                -0.000000   
4 -0.136345  0.939521         0.000000                -0.000000   

   Win_Percentage  Loss_Percentage  Draw_Percentage  Goals_by_Win      GCBW  \
0        0.091590        -0.004523             -0.0     -0.140931 -0.000000   
1        0.055664        -0.000116             -0.0     -0.000000 -0.000000   
2        0.029353        -0.062109             -0.0      0.000000  0.000000   
3        0.154849        -0.000165             -0.0     -0.000000 -0.414414   
4        0.028538        -0.003833             -0.0     -0.000000 -0.000000   

        ADR  
0  1.372067  
1  1.339772  
2  1.018465  
3  1.058514  
4  1.707975  
```

As seen from above a bunch of the less important features get zero'd out overtime. So for example Draw_percentage is observed as a significantly less predictive feature than ADR.

```
([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18], [Text(0, 0, 'team_catego
```



From this strip plot it becomes clear that Wins and Losses do the best job at predicting the end of season points. However it is not quite clear what is the next best.

We will now find the top 10 features

```
W                  9.592635
L                  6.489302
ADR                1.299359
GA                 0.491382
Dif                0.187904
Goals_Per_Match    0.092940
GCBW               0.082883
Win_Percentage     0.071999
Goals_by_Win       0.028186
Loss_Percentage    0.014149
dtype: float64
```

Though first half of the season points were by far the best predictors, stats like Wins, Losses, ADR, and Goals Against are also quite predictive as well.

Now lets look at which features were zero'd out the least

```
ADR                5
W                  5
Loss_Percentage    5
Win_Percentage     5
```

```
L                    5
dtype: int64
```

Results

Statistical Models

- **Model 1**: Simple linear regression of `Final_Points` on `Diff`.
    - Equation: `Final_Points = a + b * Diff`.
    - Interpretation: For every unit increase in `Diff`, `Final_Points` increases by `b`.
- **Model 2**: Multiple regression including `GF`, `GA`, and `Win_Percentage`.
    - Assess which variables are most predictive.
- **Model 3**: LASSO regression
    - Assess which features are the best/worst at predicting Final_Points.

Key Findings of Multiple Linear Regression Model

To evaluate the performance of our multiple linear regression model predicting Final_Points, we used two complementary validation techniques: a traditional train/test split and 10-fold cross-validation.

We first applied a 75/25 train/test split, where the model was trained on 75% of the data and tested on the remaining 25%. This allowed us to assess how well the model performs on a single, unseen subset of data. From this approach, we achieved an R^2 score of 0.92 and an RMSE of 5.79, indicating a strong fit but still subject to the randomness of one specific split.

To obtain a more robust and generalizable assessment, we also performed 10-fold cross-validation. This method divides the dataset into 10 parts, training the model on 9 and testing on the 1 remaining fold, repeating the process 10 times. The performance metrics are then averaged across all folds. This approach yielded an even stronger result, with an R^2 of 0.936 and an RMSE of 5.10.

Using both techniques allowed us to compare the model's performance on a specific split versus its average performance across multiple random splits. The cross-validation results suggest that the model performs consistently well across different subsets of the data and is unlikely to be overfitting to a particular sample. This enhances our confidence in the model's reliability and predictive strength when applied to new data.

The following is the multiple regression formula:

$$\mathbf{Final\_Points = 37.39 + 0.31 \times Diff - 0.31 \times GA + 60.34 \times Win\_Percentage}$$

We applied this formula to a team that had the following stats halfway through the season:

- Goal Differential (Diff) = 6
- Goals Against (GA) = 27
- Win Percentage (Win_Percentage) = 0.526 (i.e., won 10 of 19 matches)

Plugging these into the formula we predict that the team will finish around 62.61 points which is only 3.4 points off there actual final points which was 66. This example reinforces the model's strength in using early-season metrics to make accurate and interpretable end-of-season predictions.

Key Findings from the LASSO Model

The LASSO model identified the most significant predictors of Final League standings, while also filtering out the less impactful variables. The strongest predictor, by no surprise, was Wins and Losses. These predictors confirmed that a team's performance during the season is the most reliable factor for how they finish. Other factors like ADR, and GA highlighted the importance that not letting up goals is more important than scoring them. Interestingly team categories and raw counts like Goals Scored were not as relevant in the LASSO model. The model achieved a MSE of 18.5, which implies the predictions were off by 4 points per season.

Discussion

This project set out to answer the question: Can a team's midseason performance metrics reliably predict their final point total in the English Premier League? Based on our statistical analysis and multiple linear regression model, the answer is yes — midseason indicators such as goal differential, goals against, and win percentage showed strong predictive power. Our final model achieved an R^2 of 0.936 with a root mean squared error of 5.10, meaning it could predict final point totals with impressive accuracy using only a few simple inputs.

From a methodological standpoint, we selected variables based on exploratory visualizations and correlation analysis. The stepwise comparison between the simple and multiple linear regression models demonstrated that including more meaningful variables significantly improved predictive accuracy. In terms of validation, we used both a 75/25 train-test split and 10-fold cross-validation, allowing us to compare performance under different scenarios. Cross-validation provided a more reliable assessment of model generalizability and gave us greater confidence in the robustness of our findings.

Then for the LASSO model the findings follow suit. While current performance is the best predictor, defensive weakness significantly influences outcomes as well. Honing in on minimizing how many goals are scored on your club could prove to be helpful in boosting end of season points. Additionally while LASSO's feature selection simplifies the model, excluded variables like Win Percentage can still be a great predictor in the multiple regression.

Both modelling approaches show that mid-season performance strongly predicts final standings. The multiple regression quantifies how metrics like win percent drive success, while LASSO isolated the most impactful factors (such as showing defensive strength is crucial for maximizing points). This dual perspective can lead to actionable insight that tracking core performance provides reliable forecasts, but defensive improvements can produce rewarding outcomes.

Limitations

- **Data Scope**: Limited to 5 seasons; may not capture long-term trends or anomalies.
- **Generalization**: Findings may not apply to other leagues with different competitive dynamics.
- **Future Years:** the model may not generalize perfectly to future seasons due to external factors.

Future Work

- Include more seasons or leagues to validate findings.
- Explore more advanced models (e.g., machine learning) for non-linear relationships.
- Investigate the impact of external factors (e.g., injuries, managerial changes).

Different Approaches

A different approach we could have taken is using time-series modeling. Then we would treat each team's performance over the season as a temporal sequence rather than a single mid season snapshot. For example, we could have used a dynamic regression model to incorporate lagged variables which are past values of predictors to better capture the trends and momentum over time.

- For our premier league project we could incorporate this to see the trajectory of performance rather than just season-to-date averages or totals at mid season.

- This could help capture momentum and trends that static mid season averages miss and can improve prediction for teams with unusual mid season trajectories. Moreover, it will make the model more adaptive to real-world developments like slumps or injuries. However, this approach would require more granular data and a more complex modeling framework.

Overall, a dynamic regression would allow you to model not just how well a team has performed, but how they are trending. This is often crucial in forecasting final outcomes in sports.

What We Would Do Differently

If we were to start this project over, one key change we would make is collecting more granular match-level data rather than relying solely on mid season summaries. This would allow us to explore trends in performance over time, such as streaks, form fluctuations, and the impact of specific events like injuries or managerial changes. With this data, we could implement dynamic regression or time-series models that better capture momentum and provide more nuanced predictions. Additionally, we would focus earlier on standardizing our feature engineering process across seasons to ensure full consistency in the variables used. This would have saved time during data cleaning and improved model generalizability. Lastly, we would invest more time in visualizing model residuals and diagnostic plots to evaluate model assumptions more thoroughly and potentially uncover patterns or outliers that merit further investigation.

Answers to Comments

One surprising insight we encountered was how consistently strong defensive performance—especially low Goals Against—outperformed offensive metrics like Goals Scored in predicting final points. While we expected overall performance to matter, the relative weight of defensive strength was more pronounced than anticipated. This finding led us to reevaluate common assumptions that high-scoring teams dominate, highlighting instead that preventing goals may be a more reliable strategy for long-term success. We also appreciate your suggestion about using prior seasons' data for future prediction and agree it would be a compelling extension to test year-over-year consistency and deeper trends.