Analysis of Football Data

By

Asim Chitre, Faizan Saiyad, Mrunal Ghadge, Shabeeha Ahmed, Siddhesh Unhavane

Executive Summary:

Football, also known as soccer, is one of the most popular sports played in the world with a lot of popularity in Europe. There are 5 major leagues played in Europe, namely- the Premier League in England, the Bundesliga in Germany, La Liga in Spain, Serie A in Italy, and Ligue 1 in France. In our project, we have studied data of season 2020 of the English Premier League.

The project aims at answering research questions like- do home statistics affect the match outcome, do away statistics affect the match outcome, how do match stats help shape the results, which betting company provides a fair share chance of winning. We have also implemented a decision tree for winning or losing and a Microsoft Power BI dashboard which shows data regarding the top 5 teams.

For the research question - do home and away statistics affect the match outcome, we first isolate the home statistics followed by applying Variance Inflation Factors (VIF) to check the multi collinearity. As VIF <5 in both cases, thus there is no multi-collinearity. Hence, we apply the Linear Regression Model and see that the home statistics, FTHG (Full Time Home Goals), the HST (Home Shots on Target), and HR (Home Red card) statistics significantly affect the results. In case of away statistics, FTAG (Full Time Away Goals), AS (Away Shots), and (Away Corners) are the three variables that affect the match outcome. The fourth research question is how do match stats help shape the outcome, we use Spearman’s correlation plot because the data is not normal,

and Spearman’s is a non-parametric correlation test, from which we can clearly find out that a substantial number of home and away statistics is related to the full-time result.

We have several limitations of the dataset, which are- Dataset is exceedingly small and is further divided into training and testing data. There is data of just one season, which may be influenced by many factors. Data does not have many factors which are associated with home/ away team influence like- crowd attendance, crowd noise level, etc. We can conclude that home and away factors play a particularly significant role in football and are related to full-time results. In home statistics, FTHG (Full Time Home Goals), the HST (Home Shots on Target), and HR (Home Red card) statistics significantly affect the results. In case of away statistics, FTAG (Full Time Away Goals), AS (Away Shots), and (Away Corners) are the three variables that affect the match outcome.

Our research can attempt to help the teams and players to gain valuable insights about the opponent about their playing styles and when they try to score their most goals in what period of the match. Similarly, this can also help for the teams to increase their chances of winning the game by seeing where they lack in the match and try to rectify their mistakes. Apart from on the field results, this research is also useful for the people who are involved in legal betting, as this would help them to understand which betting companies provide the maximum returns and would help them earn significant profits per match.

Abstract:

Football/ soccer is one of the most popular sports in the world with billions of fans around the world. Events like the Premier League have remarkably high stakes and there needs to be a lot of strategizing before playing the game. The main motive of our research is to analyze some of the factors like home and away advantages, goals scored in halftime to strategize better. Our research focuses on finding out if home and away factors affect the team and which factors determine the full-time results. We have used Variance Inflation Factor, Linear Multiple Regression, Spearman's Correlation, and decision tree to answer our research questions. After analyzing, we find that home and away stats affect the game play significantly and it is clearly seen that the team which is leading in the halftime wins the game. Thus, we can make strategies like pressing more in the first half, that is to attack more in the first 45 minutes of the game to have more chances of winning.

Introduction:

Soccer is the world's most popular sport and is a national sport in many European and Latin American countries. About 140 countries play soccer religiously. There is a huge craze of "Club Football" in this globalization phase. The top teams play in different leagues in England, Spain, Germany, France, and Italy. In addition to that, the world cup is played every 4 years and it is one of the most watched events in the world. The dataset used in our project is consists of statistics from the English Premier League, season 2020. We used this data to check whether these factors affect the final score and chances of winning the match. We also found that this dataset had betting odds of various companies and decided to check which company the people should use to gain maximum profits. We believe our results would help teams and players to shape their playing styles and customize tactics to have increased chances of winning the match.

Literature Review:

Oberstone (2009) organized the football actions into 5 broad categories; Attempts, Passing, Defending, Crossing and discipline and differentiated the teams based on the same. There were three separate classes for the teams: the top four, the bottom four and the rest 12 middle teams. He used Analysis of Variance (ANOVA) and multiple regression models to analyze the data.

However, Croucher used Poisson distribution and multiple regression and used only the goals scored by the teams to test whether any team finished significantly higher or lower. This paper also had a separate section dedicated to the drawn matches, used conditional probability, and expected Poisson distribution. Similarly, Brillinger (2006) also used Poisson distribution as a measure to calculate the home and away team's effect on the match in which he used the data of Norwegian Football League. Based on the research done by various papers, we plan to take some concepts from these papers, and we will be doing our own research. We will be proving that the home advantage still is a crucial factor in Soccer and will use regression models and hypothesis testing to analyze the data.

We will be using numerous factors to determine the difference between what the table should have looked like and what the table finally was at the end of the season. This includes ratio of shots to shots on targets, number of goals and the number of fouls committed by the team. At the end, we also will determine whether playing at home is an added advantage to the team. From this analysis, we would be implementing a detailed analysis that would contain regression models and descriptive statistics and would be creating visualizations for the same.

Methods:

Our first research question attempts to answer whether home statistics affect the match outcome. For answering this question, we first separate the home statistics variables from all statistics. This isolates the home statistics and gives us a better idea of the effects of home statistics on full time result. Then we check Variance Inflation Factor (VIF) to further check the multi collinearity. We proceed with Linear Multiple Regression analysis for home statistics with target variable as full-time results. First, we get the coefficients for establishing the relation between feature variables and target variables, which in this case is home statistic variables and full-time results.

The second research question tried to establish a relationship between away statistics and the match outcome. We tried to establish a relationship between the away statistics and match outcome in such a way as to give fair chance of opponent to win the match. We move on with the available variables and model a Linear Regression with target variable as FTR (Full Time Result).

The next research question states whether match stats help shape the results. To check this, we used the concept of correlation and made a visualization using Spearman’s correlation plot which resulted in a correlation table having values ranging from –1 to 1. The match statistics that we are going to discuss about include Full Time Home Goals, Fulltime Away Goals, Full Time Results, Half Time Home Goals, Half Time Away Goals, Home Shots, Away Shots, Home Shots on Target, Away Shots on Target, Home Fouls, Away Fouls, Home Corners, Away Corners, Home Yellow cards, Away Yellow cards, home red cards, and Away Red cards.

Our next research question was whether half time results affect the final game results. We selected the number of goals scored by home and away teams, as well as the result and we created another data frame. We also created a new column which will have the data of which team is in the lead. Using this processed data, we calculate the instances of the home team winning the match, drawing the match, or losing the match, given that they were ahead in the first half. Comparable results were shown for away teams and well. Finally, we also saw the games which ended in a draw given that either the half-time score was 0-0 or the home or away team was in the lead.

The next research question answers the question as to which betting company provides a fair share chance of winning. To answer this research question, we first take the top six betting companies and apply a linear regression model, with the target variable predicting the outcome. The target variable is categorical whereas the betting odds are continuous variables. We then compute the odds ratio and p value of this model.

Last research question focuses on chances of away team winning, home team winning, or draw based on various factors. To answer this research question, we determined the most significant factors based on their level of significance. We have taken the top five significant variables and decided tree predicting the possibility of each event happening. Then we created a decision tree. To find how the match outcome is dependent on various factors, we plotted a Variable Importance Plot (Figure 10) indicating the level of significance of each variable. The top five significant variables will contribute the most in determining the outcome of the match.

Discussion and Results:

We can see from the correlation plot (Figure 1), a considerable number of home and away statistics are related to the full-time result. There are no highly correlated variables throughout the dataset, and this is established through the above correlation plot. The method used for the correlation plot is Spearman's because the data is not normal, and Spearman's is a non-parametric correlation test.

Figure 2 shows the analysis of full-time results given the half-time results. It is evident from the graph that the team, which is leading in the half time, wins the game. We see that 85 matches were won by the away team when they were leading, and they lost just 9 out of a possible 107 matches. The home team has won 94 matches from a possible 120 matches. However, there are just 9 instances when they come from behind and won the match. Draw at the halftime prefers more to the away team rather than home team.

For the second research question, a general notion about coefficient is that positive means that it affects the target variable positively and vice-a-versa. In figure 3, it is true since FTHG (Full Time Home Goals) has the maximum coefficient which is true as more full-time goals tips the result in the favor of the home team. Whereas HR (Home Red cards) means a loss of player. This is a disadvantage for the home team and hence has a negative effect on the end results as it tips the result in favor of the opponent or away team. Further, it is also interesting to note that just taking shots towards goal does not pressurize the away team, taking shots on target does help the result to tip in the favor of home team. This analysis is especially important as the results from this analysis can help the home team strategize their plan to win a game keeping the home statistics under control as home statistics are heavily influenced by the home team. Thus, if the home team scores more goals, takes more shots on target, and carefully plays the match without getting a red card, the chances of the home team winning look very promising.

The model summary for home statistics (Figure 4) dives deeper and explains more about the effect of home statistics on full time result. The most important metric that we see here is the multiple R-squared value which is 0.4118. This means about 41% of the outcome is explained by home statistics. The hypothesis that we assumed in the conclusion is proved by this model as the FTHG (Full Time Home Goals), HST (Home Shots on Target), and HR (Home Red card) have relationship with our target variable i.e., FTR (Full Time Results). The diagnostic plots in Figure 5 help us understand the Linear Regression Model in a much deeper sense. The diagnostic plot talks about various metrics like overfitting, residual, normality etc., and can be easily interpreted by only seeing the graphs.

The Residual vs Fitted shows if there is a linear relationship between our predictor variable and target variable. It is clear from the first plot that there exists a linear relation to some extent but overall, there is no linear relationship. The Normal Q-Q plot shows whether the residuals are normally distributed or not. In this case, the residuals are normally distributed with some exceptional cases that can be clearly seen near the origin of the plot. The Scale-Location plot shows how the residuals are spread or in statistical terms, Homoscedasticity. Although the residuals are not spread randomly, i.e., they do follow a pattern, but the line is not horizontal. This indicated the presence of random spread of residuals within our model. The Residuals vs Leverage plot serves the purpose of identifying influential outliers within our total spread of residual values. We see there is no presence of influential outliers.

In context to the third research question, the coefficients just reiterate on the conclusions that we drew from the correlation plot in Figure 6. The FTAG (Full Time Away Goals) and AS

(Away Shots) work against the home team and push the outcome in away team's favor. The AR (Away Red cards) favors the home team and hence have a positive coefficient.

From the summary of Linear Regression Model in Figure 7, we see similar summary statistics. The Multiple R-squared is 42% meaning 42% of the outcome is explained by our model. The residual error is also less. But here we see three significant variables. We can see that significance codes are being marked by '*' (asterisk) beside the predictor variables. The p-value can also be used to find out the significant variables. In case of away statistics, FTAG (Full Time Away Goals), AS (Away Shots), and (Away Corners) are the three variables that affect the match outcome. In case of away statistics, away corners play a significant role. This single outcome can be influential in turning around the results in favor of the away team. The away team can strategize their game plan to take more shots and get maximum number of corners to pressurize the home team. This is again true in real life soccer scenarios.

Looking at the diagnostic plot in Figure 8 of the regression model especially the first plot, i.e., the Residuals vs Fitted plot, we observe that there is a partial linear relationship between predictor variables and target variables. But overall, the model does not show a linear relationship. From the Normal Q-Q plot, we can see that the residuals are normal with a handful of exceptional outliers that are in the top left corner of the plot. Overall, the residuals are normally distributed. From the Scale-Location plot, we see that the spread follows a pattern, but it is still random. This proves the existence of Homoscedasticity within our model, as the spread is random and is difficult to determine. The Residuals vs Leverage shows that there are no influential outliers.

Our next research question i.e., which betting company provides a fair chance of winning, we keep a biased decision with respect to winnings. We keep 50% confidence interval. This will ensure that betting companies as well as the better have a fair chance of winning. We need to first find that which companies are getting affected by the results, as well as those companies which have p values less than 0.5. Each company gives odds of home team winning, away team winning as well as a draw. We find out from Figure 9 that WH, which is William Hill, has been affected by the results as it has a higher odds ratio of 3.85,3.50 and 1.53 in all the possible scenarios. In addition to that, the p-values are 0.2,0.4,0.4. Hence, they are all significant and answer our research question.

From the summary of the decision tree in Figure 11, we get an accuracy of 60%, sensitivity as 95.6% and 43% specificity i.e., 95.6% of positive outcomes and 43% of negative outcomes is determined by the model accurately. We observe from Figure 12 that if in half time, Away team is in the lead or both the teams have performed equally, the chances of Away team winning or the match being draw is 66%. If the number of shots on target by the away team is greater than or equal to 3, away team has 55% chances of winning and 12% chances of the match being draw. If in the half time, home team is in the lead, they have 34% chances of winning.

We created a dashboard in Power BI to show the performance of top five teams in the league and how match statistics are affected by the fouls i.e., number of red and yellow cards shown to the team. As far as the performance of top five teams is considered, Manchester City has performed the best on both home and away ground (Figure 13). We found that as the number of shots on target decreases, number of full-time goals also decrease. The number of fouls can also affect the match outcome. If home team gets red card more than away team, away team is more likely to win. If away team gets red card more than home team, home team is more likely to win (Figure 14). The same is applicable for yellow cards. However, interpretations cannot be made for the match being draw based on the number of fouls.

Conclusion:

The conclusions are very sensible when we talk about soccer as a sport. If home statistics heavily influence the match outcome, then every home match should have been won just by focusing on home statistics, which is not the case. Similarly, if the match results are heavily influenced by away statistics, then every away team can avoid the face of defeat by focusing on the statistics that heavily affect the match outcome. But again, that is not what we see in real life scenario. A match is fair when we talk only about home statistics and away statistics. A model that consists of both the home and away statistics is much more efficient in determining the match outcome.

We observed that no match is single handedly affected by home or away statistics. Home and away statistics both affect the match outcome. The home team must perform better, play high intensity, and cautious football to win the match. If the away team succumbs to this pressure, eventually the away statistics will take a hit as they will take less shots and eventually scoreless goals. This will ensure the home team victory. But as we concluded above, there are some factors that can help either of the teams perform better. Like for the home team, they can push their team to take more shots on target rather than just wasting opportunities and playing carefully without committing too many fouls. On the other hand, the away team can focus more on pressurizing the home team by taking more shots and winning more corners. This will slightly improve their odds of winning, but no one can ensure that either of the teams wins.

Hence, we conclude that there are a handful of home and away statistics that affect the Full Time Result. But the set of factors must be considered simultaneously to better calculate the influence of the match statistics on the Full Time Result.

References:

1. Baboota, R., & Kaur, H. (2018, March 28). *Predictive analysis and modelling football results using Machine Learning Approach for English Premier League*. International Journal of Forecasting. Retrieved October 25, 2021, from https://www.sciencedirect.com/science/article/pii/S0169207018300116.

2. Brillinger, David. (2006). *Modelling some Norwegian soccer data*. 10.1142/9789812708298_0001.

3. Croucher J.S. (2004). *Using Statistics to Predict Scores in English Premier League Soccer*. 2004, In: Butenko S., Gil-Lafuente J., Pardalos P.M. (eds) Economics, Management and Optimization in Sports. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24734-0_4.

4. Eggels, H., Elk, R. van, & Pechenizkiy, M. (2019, May 13). *Explaining soccer match outcomes with Goal Scoring Opportunities Predictive Analytics*. Eindhoven University of Technology research portal. Retrieved October 25, 2021, from https://research.tue.nl/en/publications/explaining-soccer-match-outcomes-with-goal-scoring-opportunities-.

5. Oberstone, Joel. *Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success*. 2009, https://core.ac.uk/download/pdf/216977938.pdf.

Figures:

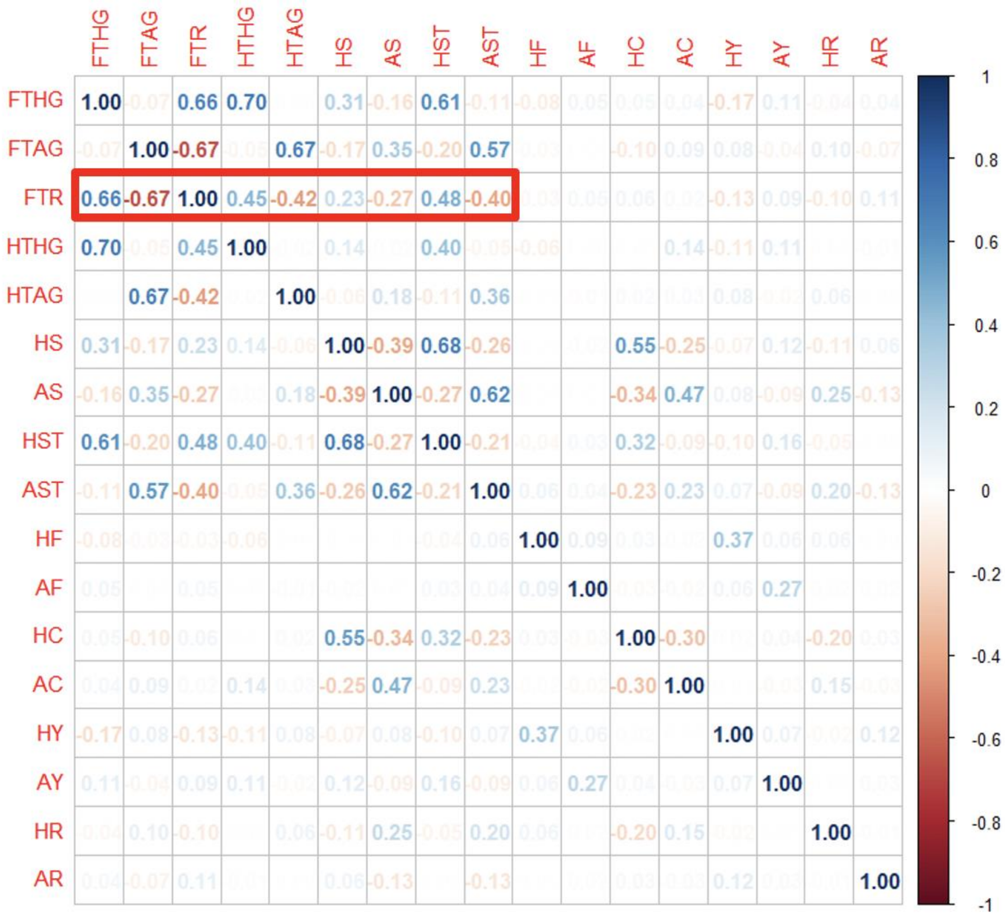


Figure 1: Correlation plot for all statistics.

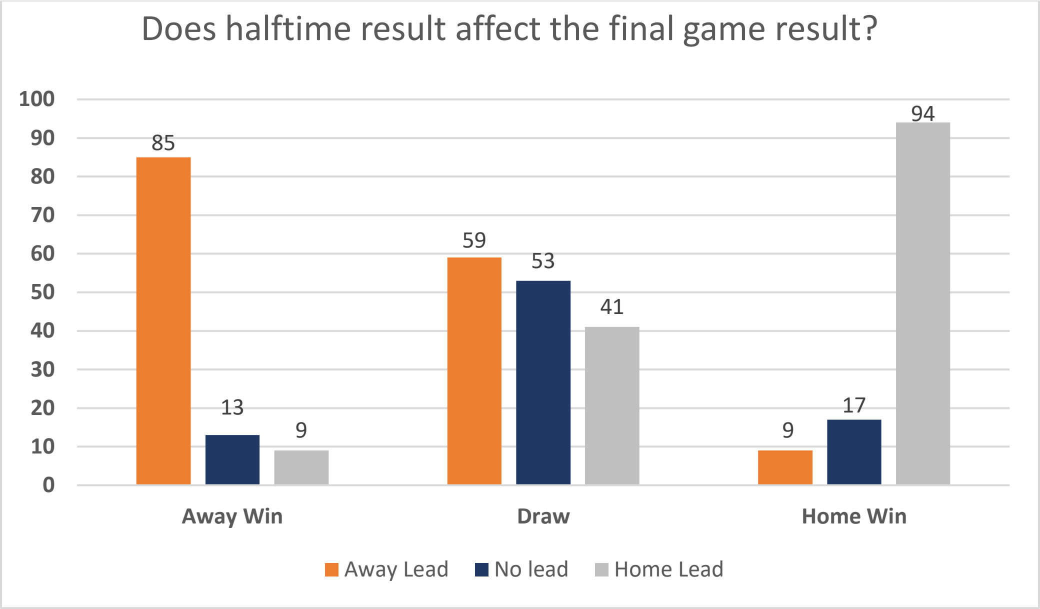


Figure 2: Analysis of half-time scores and full-time results.

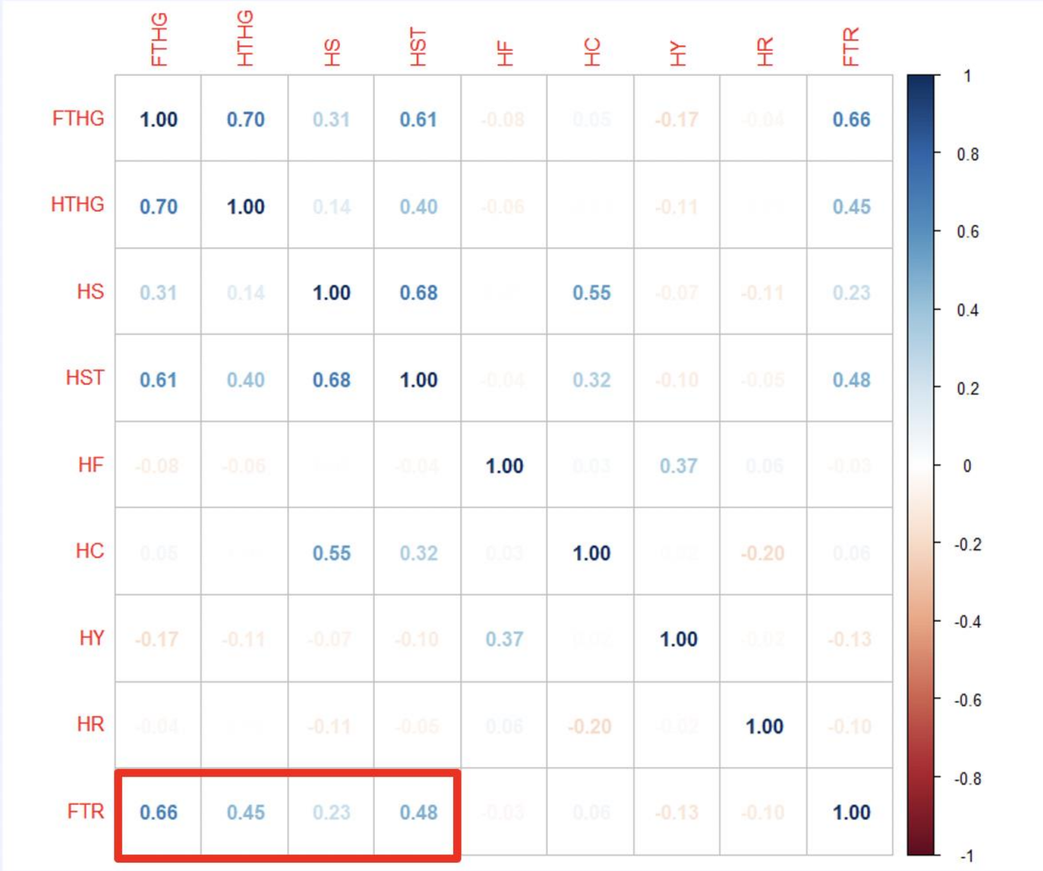
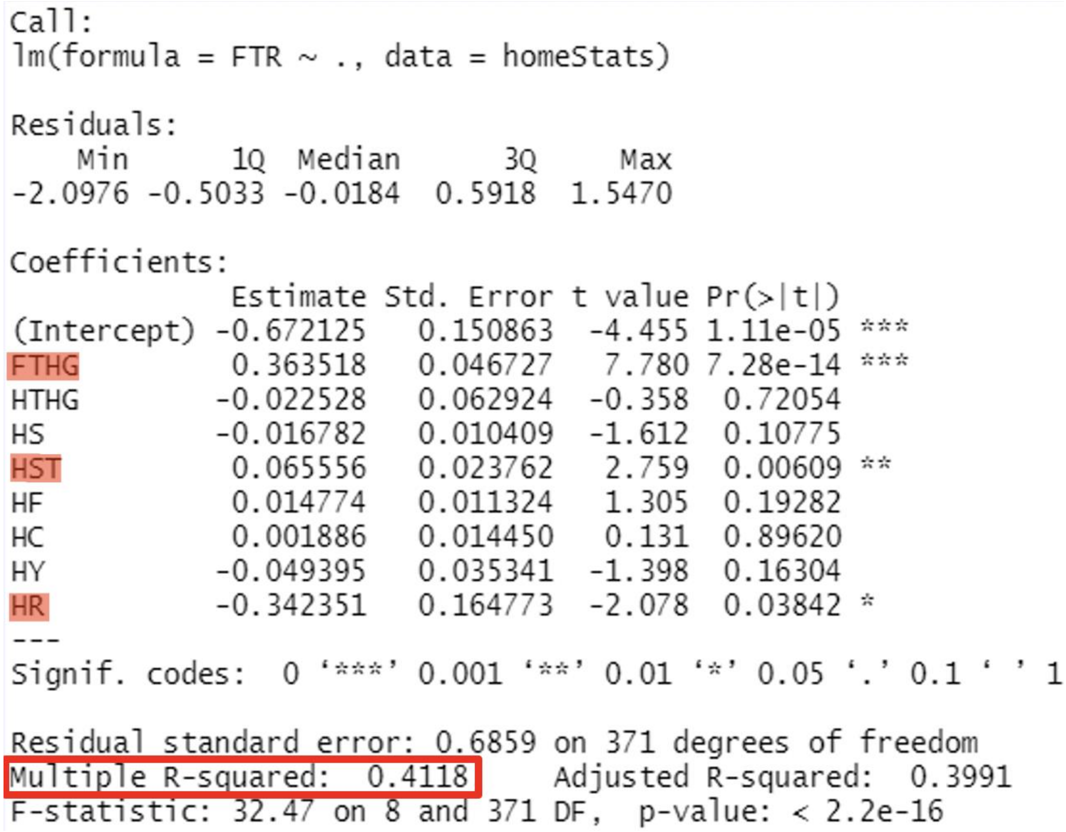


Figure 3: Correlation plot for home statistics.



```
Call:
lm(formula = FTR ~ ., data = homeStats)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0976 -0.5033 -0.0184  0.5918  1.5470 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.672125   0.150863  -4.455 1.11e-05 ***
FTHG         0.363518   0.046727   7.780 7.28e-14 ***
HTHG        -0.022528   0.062924  -0.358  0.72054    
HS          -0.016782   0.010409  -1.612  0.10775    
HST          0.065556   0.023762   2.759  0.00609 ** 
HF           0.014774   0.011324   1.305  0.19282    
HC           0.001886   0.014450   0.131  0.89620    
HY          -0.049395   0.035341  -1.398  0.16304    
HR          -0.342351   0.164773  -2.078  0.03842 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6859 on 371 degrees of freedom
Multiple R-squared:  0.4118     Adjusted R-squared:  0.3991 
F-statistic: 32.47 on 8 and 371 DF,  p-value: < 2.2e-16
```

Figure 4: Model Summary for Linear Regression Model of Home Statistics.

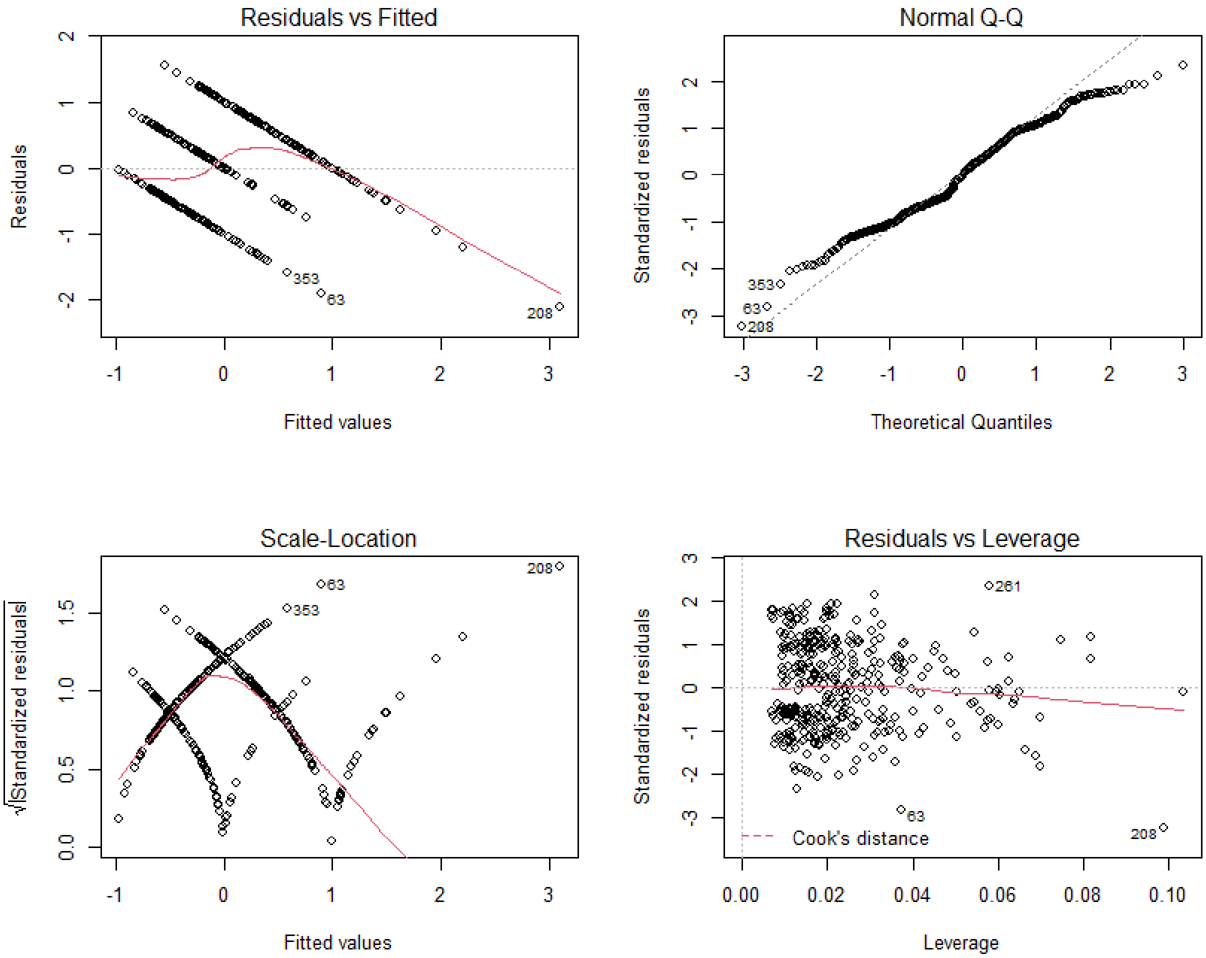


Figure 5: Home Statistics Diagnostic Plot.

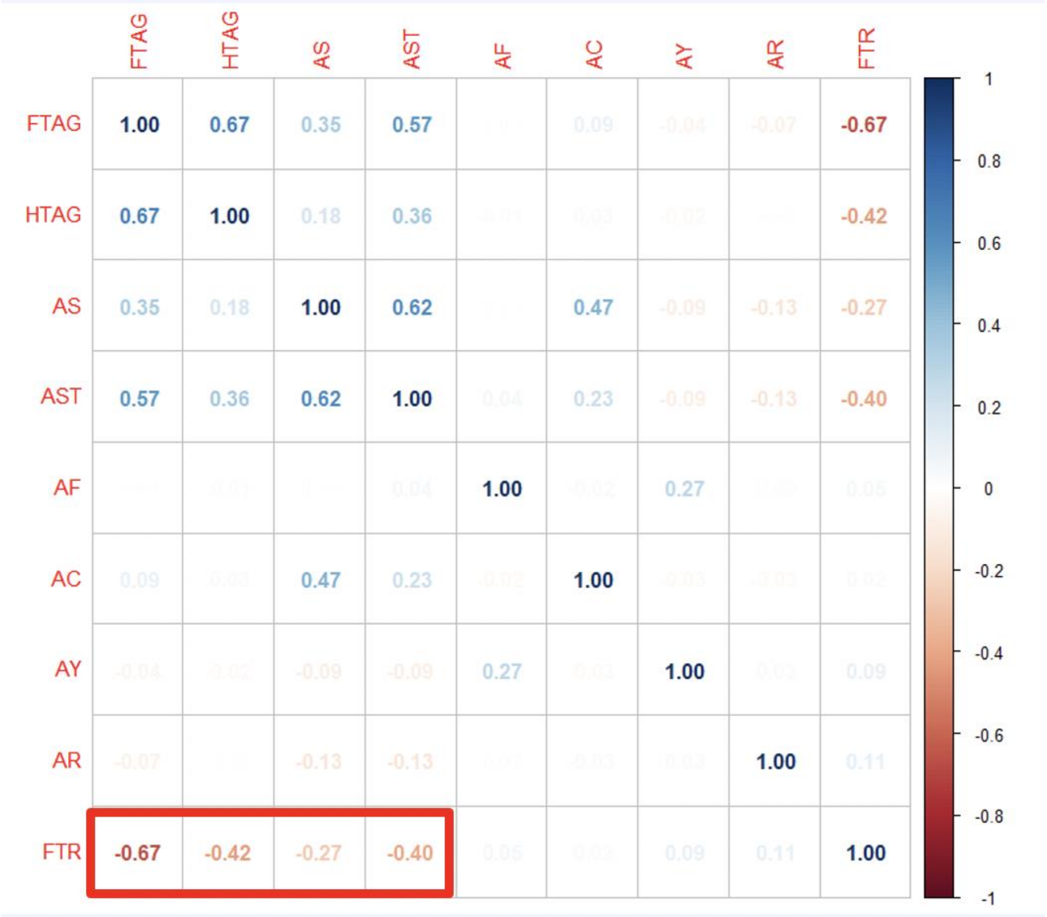
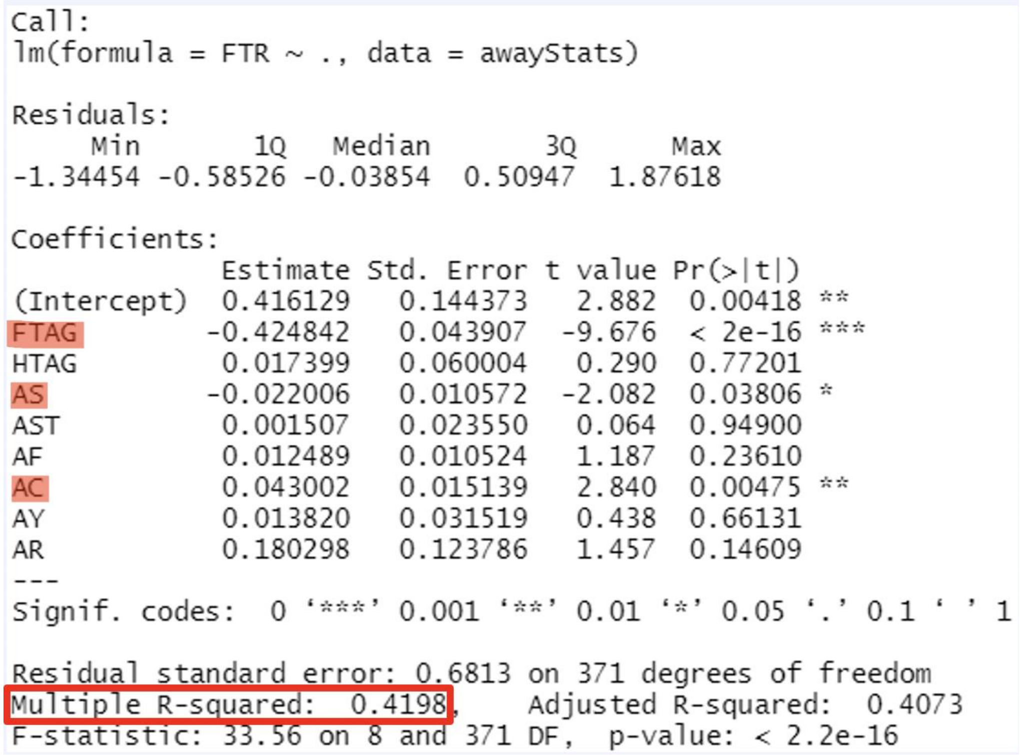


Figure 6: Correlation plot for away statistics.



```
Call:
lm(formula = FTR ~ ., data = awayStats)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.34454 -0.58526 -0.03854  0.50947  1.87618 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.416129   0.144373   2.882  0.00418 ** 
FTAG        -0.424842   0.043907  -9.676  < 2e-16 ***
HTAG         0.017399   0.060004   0.290  0.77201    
AS          -0.022006   0.010572  -2.082  0.03806 *  
AST          0.001507   0.023550   0.064  0.94900    
AF           0.012489   0.010524   1.187  0.23610    
AC           0.043002   0.015139   2.840  0.00475 ** 
AY           0.013820   0.031519   0.438  0.66131    
AR           0.180298   0.123786   1.457  0.14609    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6813 on 371 degrees of freedom
Multiple R-squared:  0.4198,	Adjusted R-squared:  0.4073 
F-statistic: 33.56 on 8 and 371 DF,  p-value: < 2.2e-16
```

Figure 7: Model Summary for Linear Regression Model of Away Statistics,

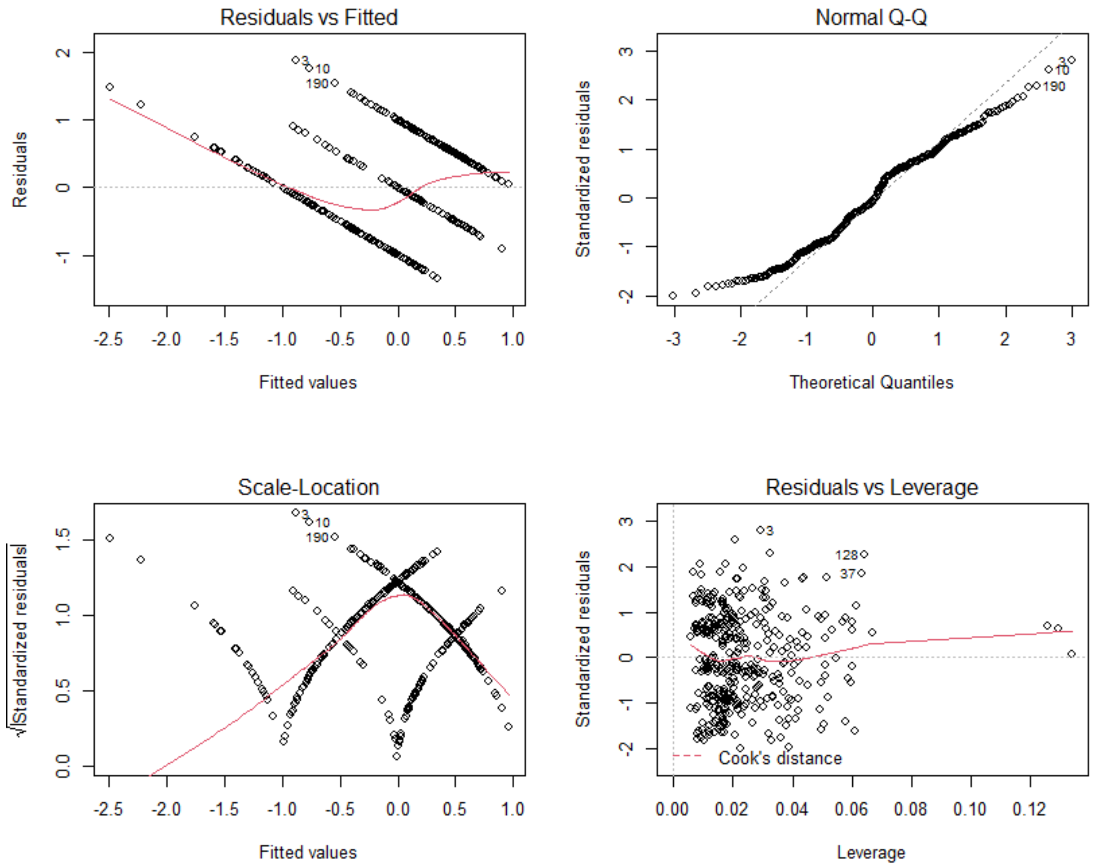
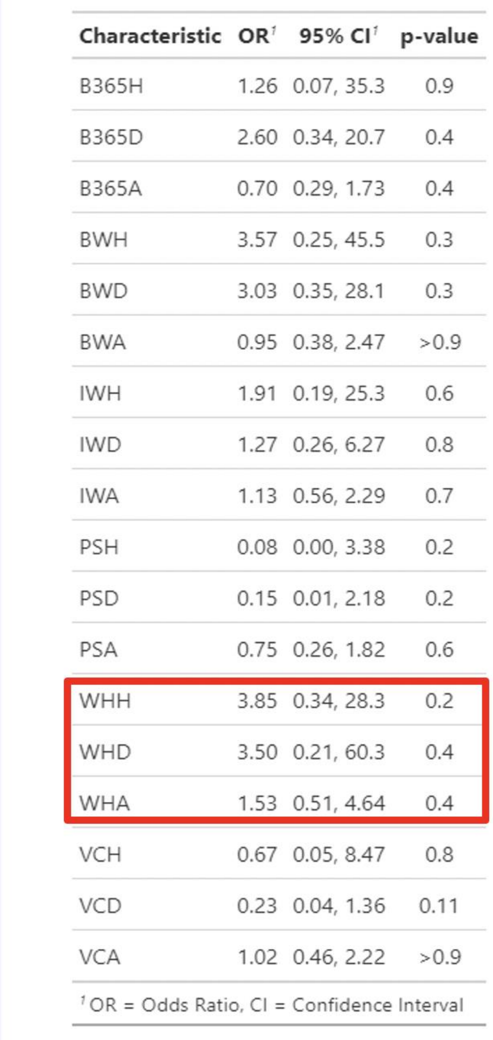


Figure 8: Away Statistics Diagnostic Plot.



Characteristic	OR[1]	95% CI[1]	p-value
B365H	1.26	0.07, 35.3	0.9
B365D	2.60	0.34, 20.7	0.4
B365A	0.70	0.29, 1.73	0.4
BWH	3.57	0.25, 45.5	0.3
BWD	3.03	0.35, 28.1	0.3
BWA	0.95	0.38, 2.47	>0.9
IWH	1.91	0.19, 25.3	0.6
IWD	1.27	0.26, 6.27	0.8
IWA	1.13	0.56, 2.29	0.7
PSH	0.08	0.00, 3.38	0.2
PSD	0.15	0.01, 2.18	0.2
PSA	0.75	0.26, 1.82	0.6
WHH	3.85	0.34, 28.3	0.2
WHD	3.50	0.21, 60.3	0.4
WHA	1.53	0.51, 4.64	0.4
VCH	0.67	0.05, 8.47	0.8
VCD	0.23	0.04, 1.36	0.11
VCA	1.02	0.46, 2.22	>0.9

[1] OR = Odds Ratio, CI = Confidence Interval

Figure 9: Analysis of betting companies.

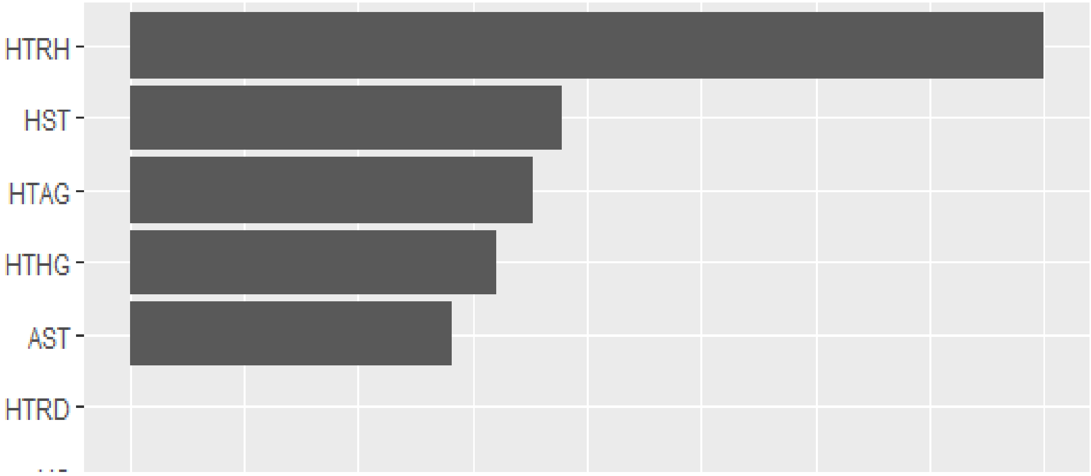
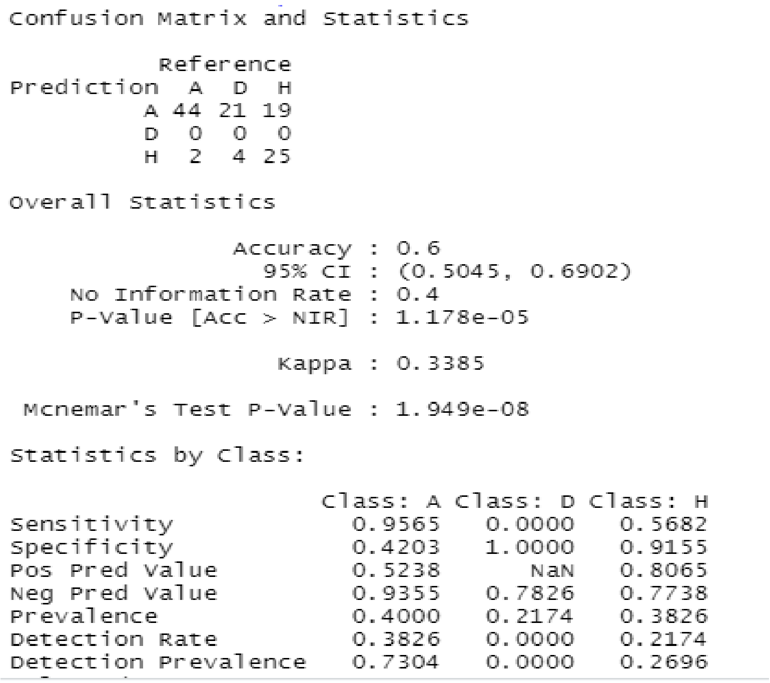


Figure 10: Variable Importance Plot



```
Confusion Matrix and Statistics

          Reference
Prediction  A  D  H
         A 44 21 19
         D  0  0  0
         H  2  4 25

Overall Statistics

               Accuracy : 0.6
                 95% CI : (0.5045, 0.6902)
    No Information Rate : 0.4
    P-Value [Acc > NIR] : 1.178e-05

                  Kappa : 0.3385

 Mcnemar's Test P-Value : 1.949e-08

Statistics by Class:

                     Class: A Class: D Class: H
Sensitivity            0.9565   0.0000   0.5682
Specificity            0.4203   1.0000   0.9155
Pos Pred Value         0.5238      NaN   0.8065
Neg Pred Value         0.9355   0.7826   0.7738
Prevalence             0.4000   0.2174   0.3826
Detection Rate         0.3826   0.0000   0.2174
Detection Prevalence   0.7304   0.0000   0.2696
```

Figure 11: Decision tree summary

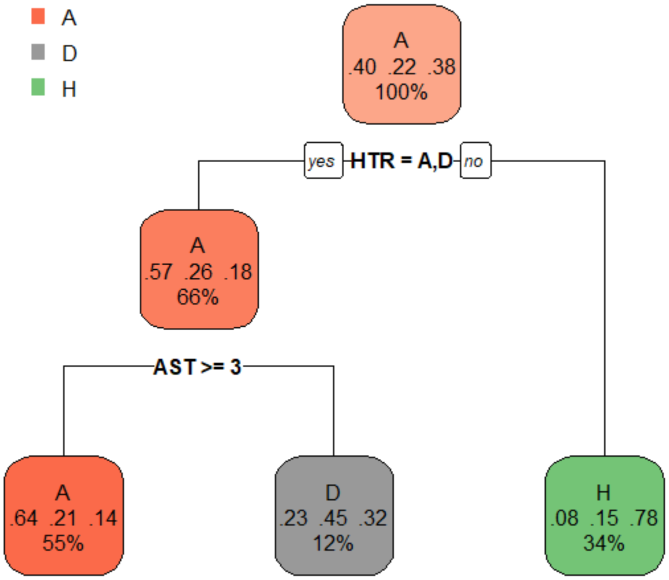
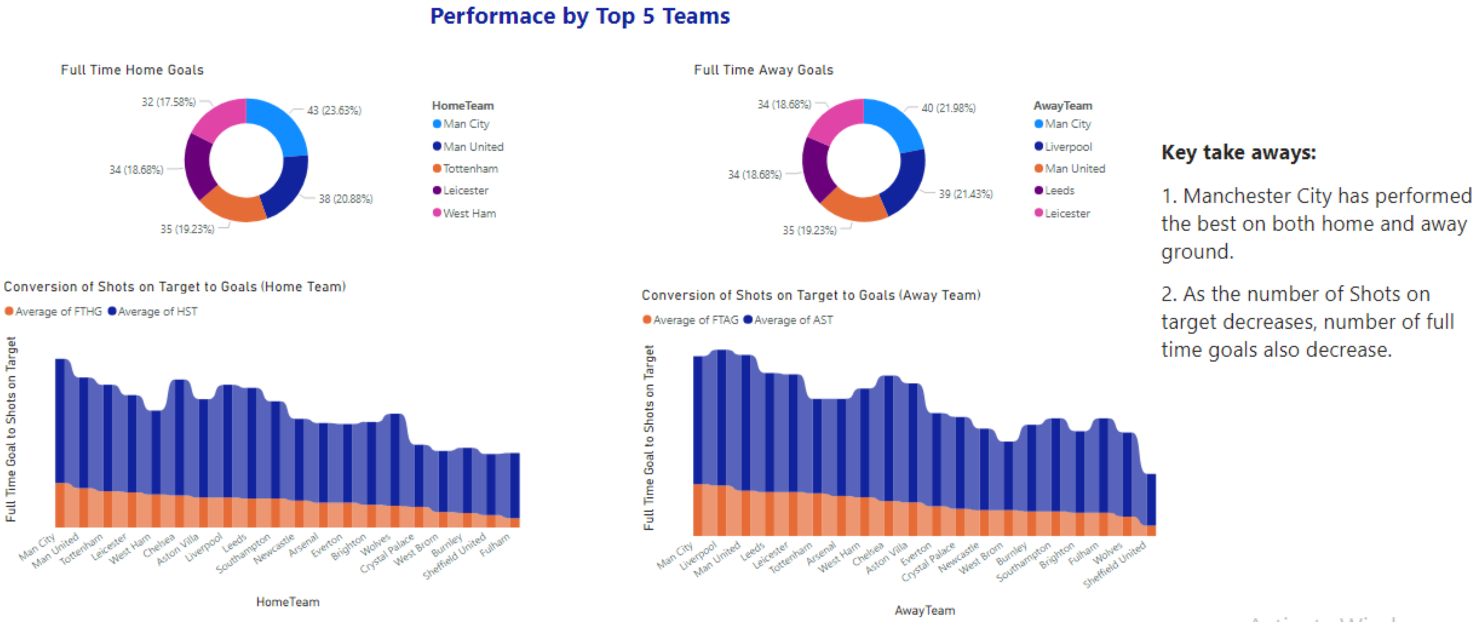


Figure 12: Decision tree

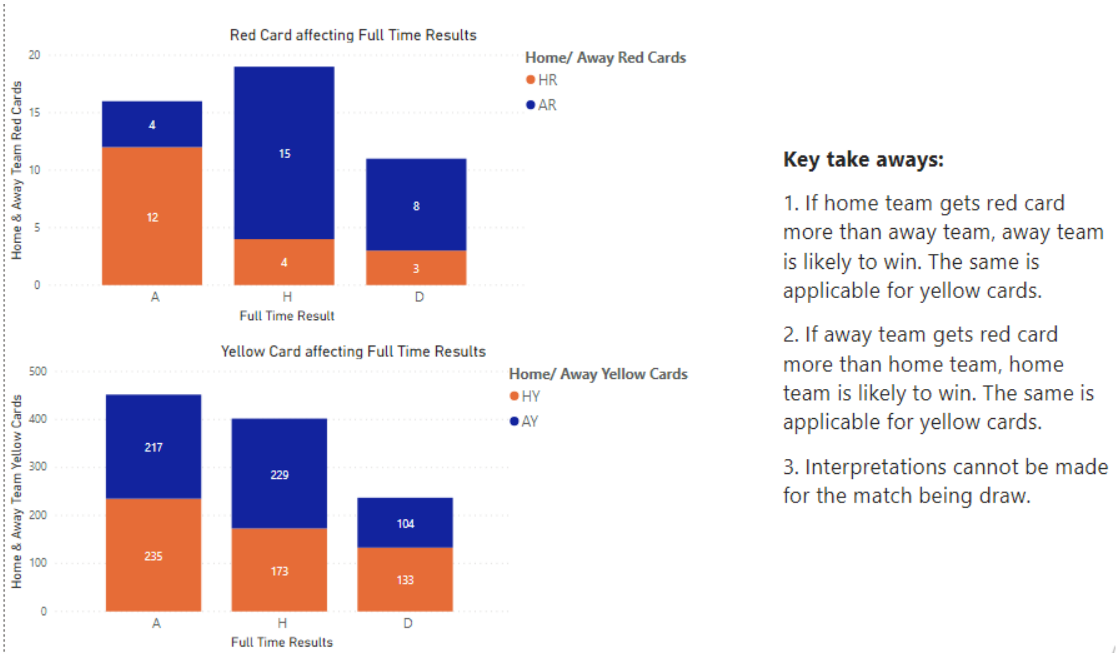


Key take aways:

1. Manchester City has performed the best on both home and away ground.

2. As the number of Shots on target decreases, number of full time goals also decrease.

Figure 13: Performance of top 5 teams



Key take aways:

1. If home team gets red card more than away team, away team is likely to win. The same is applicable for yellow cards.

2. If away team gets red card more than home team, home team is likely to win. The same is applicable for yellow cards.

3. Interpretations cannot be made for the match being draw.

Figure 14: Analysis of match outcome based on Red & Yellow cards