Analysis on Football data

Presented by - Asim Chitre, Faizan Ali Saiyad, Mrunal Ghadge, Shabeeha Ahmed, Siddhesh Unhavane

INTRODUCTION

- Football also called soccer is a popular sport.
- Our project is focused on finding important statistical factors which affect the match outcomes.
- We have a dataset consisting of data from English Premier League.
- We aim to find out insights which can help and contribute to real world application.

LITERATURE REVIEW

- **Brillinger -Modelling some Norwegian soccer data (2006) -** 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, which is same as one of our research questions.

- **Croucher - Using Statistics to Predict Scores in English Premier League Soccer (2004) -** The author 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. In our project, we have tried to see number of shots on target, home and away factors have a relation to winning or losing.

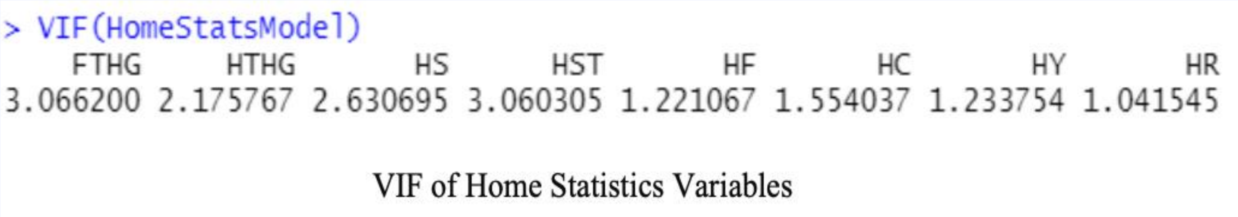
- **Oberstone- Differentiating the Top English Premier League Football Clubs from the Rest of the Pack (2009) -** The author Oberstone organized football actions into categories like passing, defending, crossing and attempts. He segregated the teams into top, bottom and middle tier which was analyzed using ANOVA and multiple regression models, which is a different from our approach of dividing everything on the basis of home and away teams.

RESEARCH QUESTIONS

S.No	Hypotheses	Methodologies
1.	Do home statistics affect the match outcome?	Variance Inflation Factor (to check multi collinearity), Linear Multiple Regression
2.	Do away statistics affect the match outcome?	Variance Inflation Factor (to check multi collinearity), Linear Multiple Regression
3.	How do match stats help shape the final results?	Spearman’s Correlation Plot
4.	Does half time results affect the final game results?	Bar graph
5.	Which betting company provides a fair share chance of winning?	Linear Regression
6.	Decision Tree for winning or losing	Decision Tree

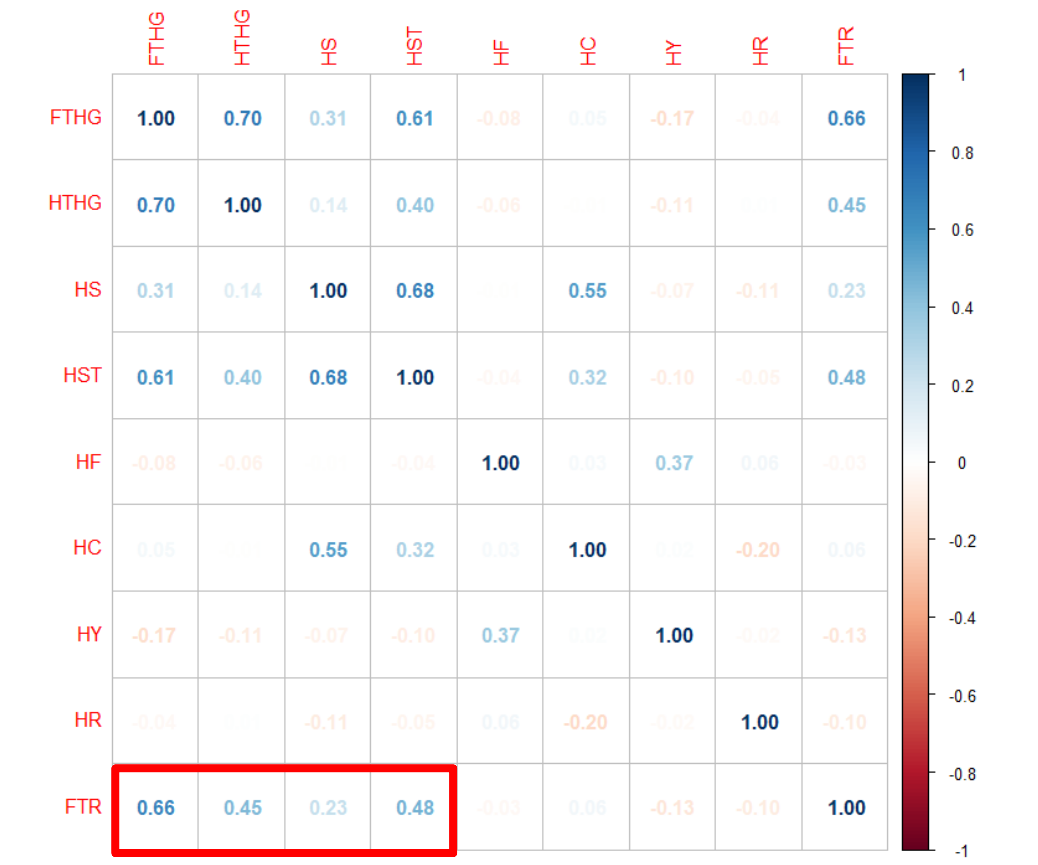
Do home statistics affect the match outcome?

Methodologies: We first isolate the home statistics followed by applying Variance Inflation Factors (VIF) to check the multi collinearity. VIF <5, thus there is no multi-collinearity. Hence, we apply Linear Regression Model.



```
> VIF(HomeStatsModel)
    FTHG     HTHG       HS      HST       HF       HC       HY       HR 
3.066200 2.175767 2.630695 3.060305 1.221067 1.554037 1.233754 1.041545 
```

VIF of Home Statistics Variables

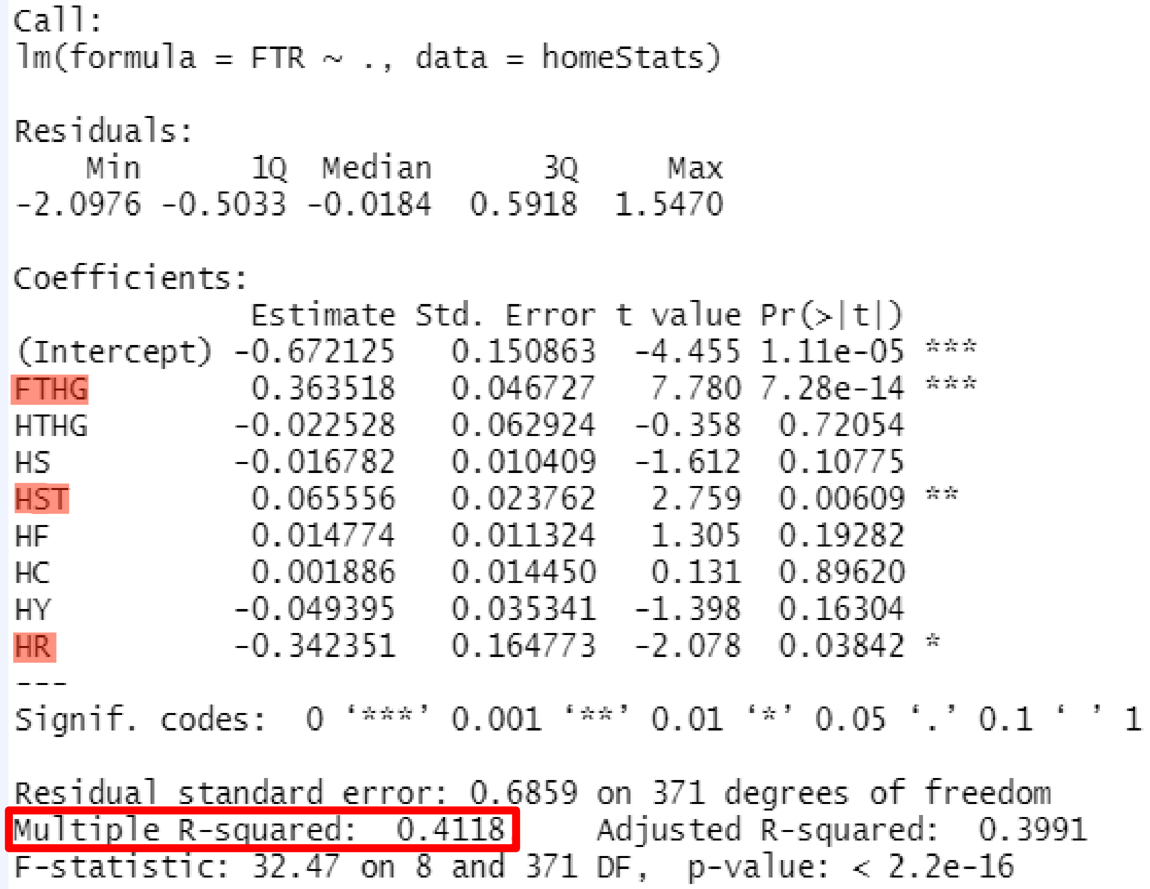


	FTHG	HTHG	HS	HST	HF	HC	HY	HR	FTR
FTHG	1.00	0.70	0.31	0.61	-0.08	0.05	-0.17	-0.04	0.66
HTHG	0.70	1.00	0.14	0.40	-0.06	0.01	-0.11	0.01	0.45
HS	0.31	0.14	1.00	0.68	-0.01	0.55	-0.07	-0.11	0.23
HST	0.61	0.40	0.68	1.00	-0.04	0.32	-0.10	-0.05	0.48
HF	-0.08	-0.06	-0.01	-0.04	1.00	0.03	0.37	0.06	-0.03
HC	0.05	0.01	0.55	0.32	0.03	1.00	0.02	-0.20	0.06
HY	-0.17	-0.11	-0.07	-0.10	0.37	0.02	1.00	-0.02	-0.13
HR	-0.04	0.01	-0.11	-0.05	0.06	-0.20	-0.02	1.00	-0.10
FTR	0.66	0.45	0.23	0.48	-0.03	0.06	-0.13	-0.10	1.00

Do home statistics affect the match outcome?

Conclusions: Out of all home statistics, FTHG (Full Time Home Goals), the HST (Home Shots on Target), and HR (Home Red card)statistics significantly affect the results.

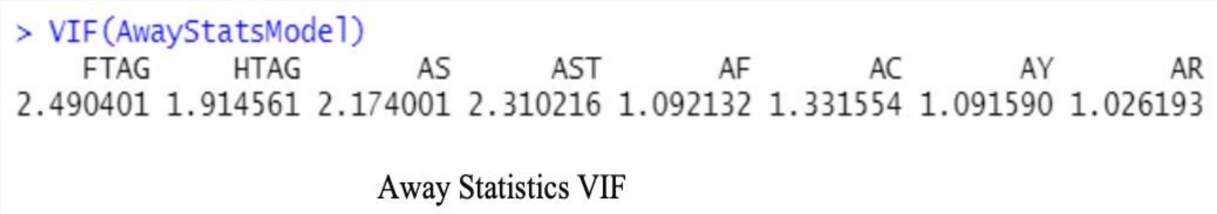
Applications: This analysis will help the teams strategize better as, if the home team score more goals, take more shots on target, and carefully play the match without getting a red card, the chances of home team winning look very promising.



```
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
```

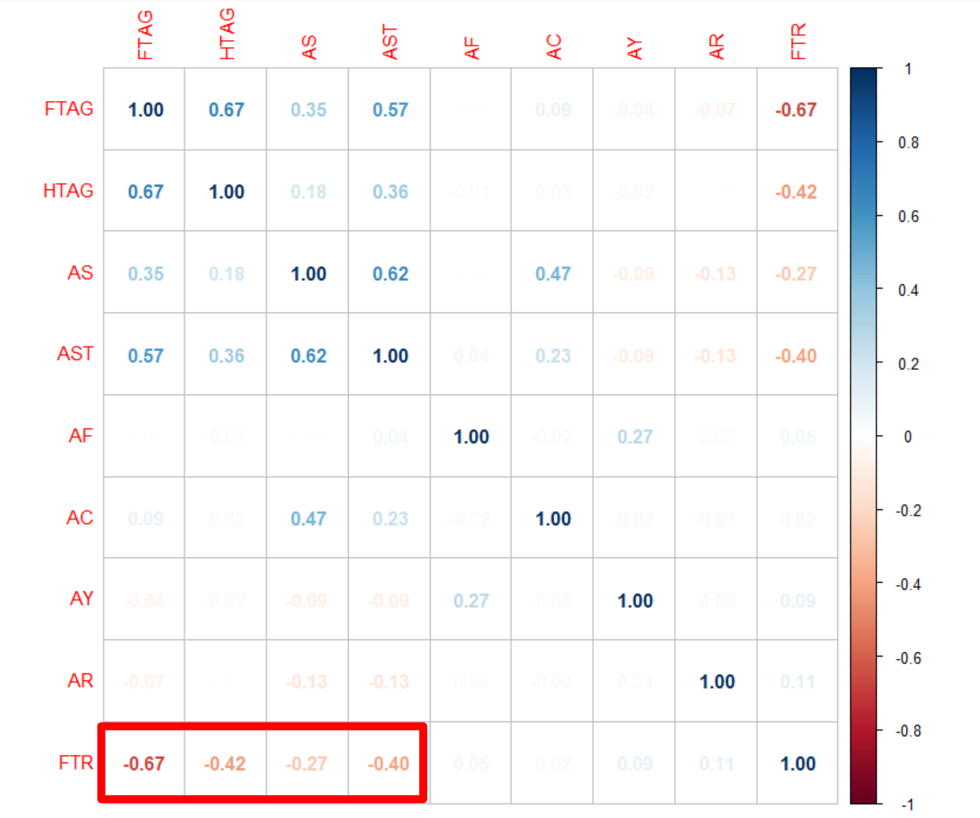
Do away statistics affect the match outcome?

Methodologies: We first isolate the home statistics followed by applying Variance Inflation Factors (VIF) to check the multi collinearity. VIF <5, thus there is no multi-collinearity . Hence, we apply Linear Regression Model.



```
> VIF(AwayStatsModel)
    FTAG     HTAG       AS      AST       AF       AC       AY       AR 
2.490401 1.914561 2.174001 2.310216 1.092132 1.331554 1.091590 1.026193 
```

Away Statistics VIF

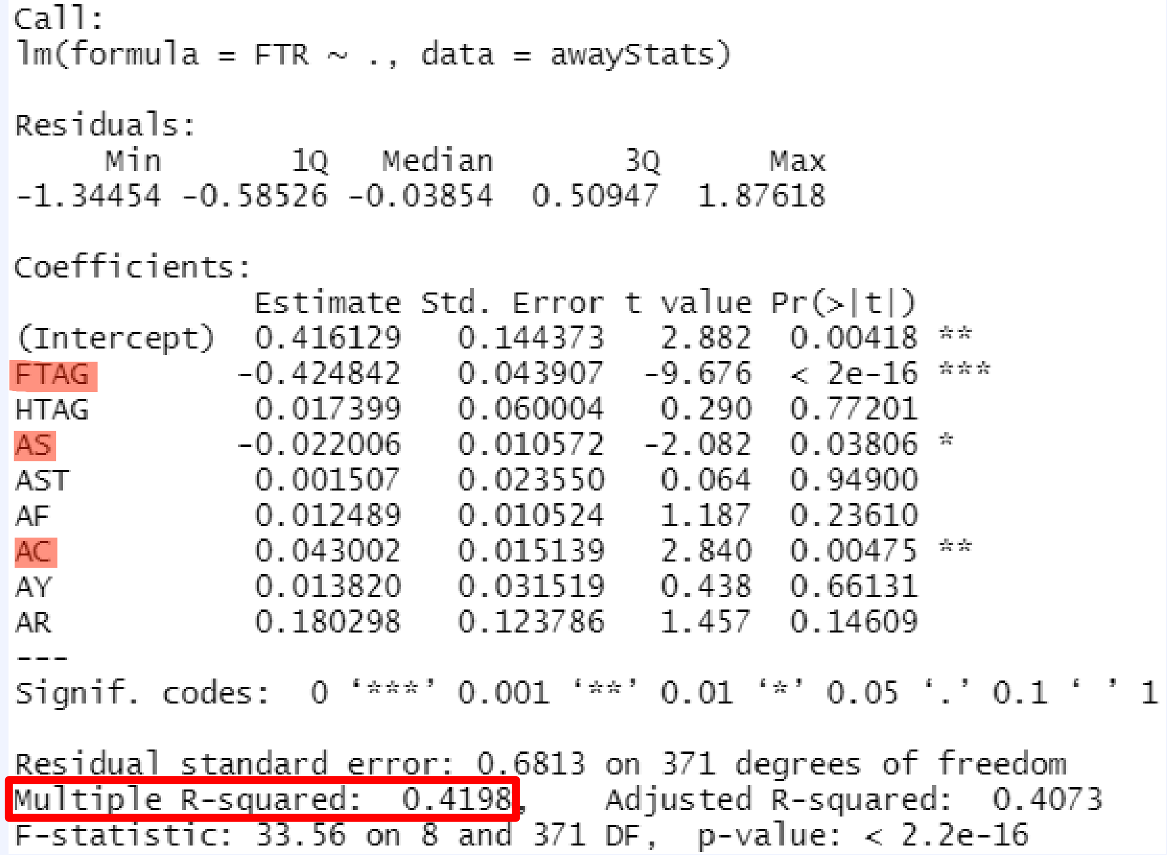


	FTAG	HTAG	AS	AST	AF	AC	AY	AR	FTR
FTAG	1.00	0.67	0.35	0.57	0.01	0.09	-0.04	-0.07	-0.67
HTAG	0.67	1.00	0.18	0.36	-0.01	0.03	-0.02	0.00	-0.42
AS	0.35	0.18	1.00	0.62	0.01	0.47	-0.09	-0.13	-0.27
AST	0.57	0.36	0.62	1.00	0.04	0.23	-0.09	-0.13	-0.40
AF	0.01	-0.01	0.01	0.04	1.00	-0.02	0.27	0.02	0.05
AC	0.09	0.03	0.47	0.23	-0.02	1.00	-0.03	-0.03	0.02
AY	-0.04	-0.02	-0.09	-0.09	0.27	-0.03	1.00	0.03	0.09
AR	-0.07	0.00	-0.13	-0.13	0.02	-0.03	0.03	1.00	0.11
FTR	-0.67	-0.42	-0.27	-0.40	0.05	0.02	0.09	0.11	1.00

Do away statistics affect the match outcome?

Conclusions: 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.

Applications: The away team can strategize their game plan as to take more shots and get maximum number of corners to pressurize the home team. This is again very true in real life soccer scenarios.

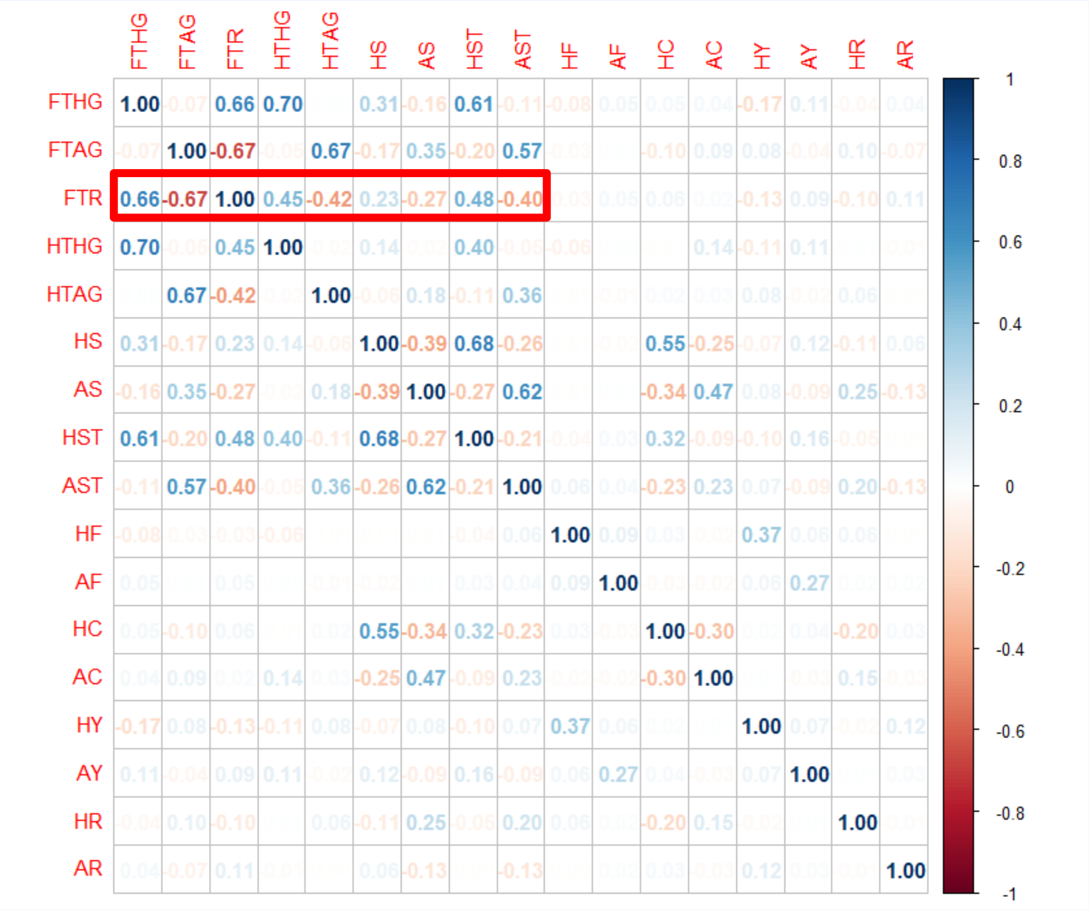


```
Call:
lm(formula = FTR ~ ., data = awayStats)

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
```

How do match stats help shape the final results?

Methodologies: 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.

Conclusions: As we can see from the correlation plot, a significant number of home and away statistics is related to the full-time result. There are no highly correlated variables throughout the dataset, and this is established through the above correlation plot



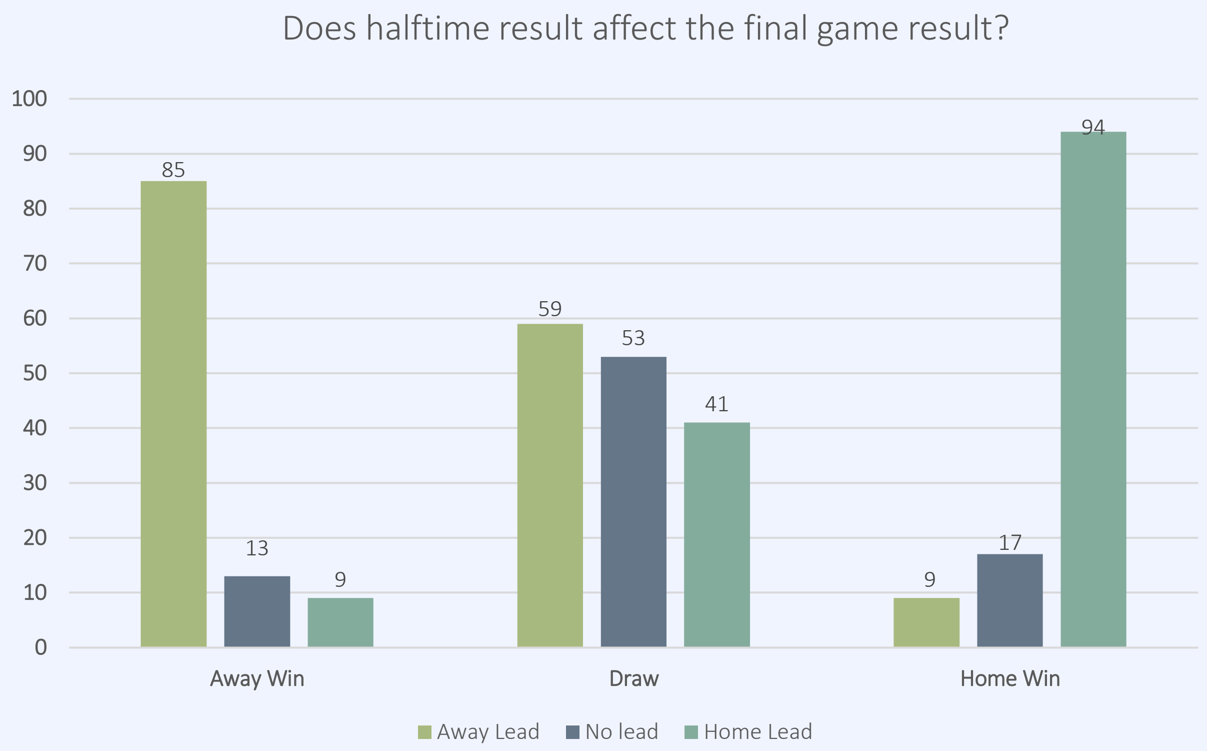
	FTHG	FTAG	FTR	HTHG	HTAG	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
FTHG	1.00	-0.07	0.66	0.70	[illegible]	0.31	-0.16	0.61	-0.11	-0.08	0.05	0.05	0.04	-0.17	0.11	-0.04	0.04
FTAG	-0.07	1.00	-0.67	-0.05	0.67	-0.17	0.35	-0.20	0.57	-0.03	[illegible]	-0.10	0.09	0.08	-0.04	0.10	-0.07
FTR	0.66	-0.67	1.00	0.45	-0.42	0.23	-0.27	0.48	-0.40	0.03	0.05	0.06	0.02	-0.13	0.09	-0.10	0.11
HTHG	0.70	-0.05	0.45	1.00	-0.02	0.14	-0.02	0.40	-0.05	-0.06	[illegible]	-0.01	0.14	-0.11	0.11	[illegible]	-0.01
HTAG	[illegible]	0.67	-0.42	-0.02	1.00	-0.06	0.18	-0.11	0.36	[illegible]	-0.01	0.02	0.03	0.08	-0.02	0.06	[illegible]
HS	0.31	-0.17	0.23	0.14	-0.06	1.00	-0.39	0.68	-0.26	[illegible]	-0.02	0.55	-0.25	-0.07	0.12	-0.11	0.06
AS	-0.16	0.35	-0.27	-0.02	0.18	-0.39	1.00	-0.27	0.62	[illegible]	[illegible]	-0.34	0.47	0.08	-0.09	0.25	-0.13
HST	0.61	-0.20	0.48	0.40	-0.11	0.68	-0.27	1.00	-0.21	-0.04	0.03	0.32	-0.09	-0.10	0.16	-0.05	[illegible]
AST	-0.11	0.57	-0.40	-0.05	0.36	-0.26	0.62	-0.21	1.00	0.06	0.04	-0.23	0.23	0.07	-0.09	0.20	-0.13
HF	-0.08	-0.03	-0.03	-0.06	[illegible]	[illegible]	[illegible]	-0.04	0.06	1.00	0.09	0.03	-0.02	0.37	0.06	0.06	[illegible]
AF	0.05	[illegible]	0.05	[illegible]	-0.01	-0.02	[illegible]	0.03	0.04	0.09	1.00	-0.03	-0.02	0.06	0.27	0.02	0.02
HC	0.05	-0.10	0.06	[illegible]	0.02	0.55	-0.34	0.32	-0.23	0.03	-0.03	1.00	-0.30	0.02	0.04	-0.20	0.03
AC	0.04	0.09	0.02	0.14	0.03	-0.25	0.47	-0.09	0.23	-0.02	-0.02	-0.30	1.00	[illegible]	-0.03	0.15	-0.03
HY	-0.17	0.08	-0.13	-0.11	0.08	-0.07	0.08	-0.10	0.07	0.37	0.06	0.02	[illegible]	1.00	0.07	-0.02	0.12
AY	0.11	-0.04	0.09	0.11	-0.02	0.12	-0.09	0.16	-0.09	0.06	0.27	0.04	-0.03	0.07	1.00	[illegible]	0.03
HR	-0.04	0.10	-0.10	[illegible]	0.06	-0.11	0.25	-0.05	0.20	0.06	0.02	-0.20	0.15	-0.02	[illegible]	1.00	-0.01
AR	0.04	-0.07	0.11	-0.01	[illegible]	0.06	-0.13	[illegible]	-0.13	[illegible]	0.02	0.03	-0.03	0.12	0.03	-0.01	1.00

Does half time results affect the final game results?

Methodologies: A simple bar graph made by selecting the number of goals scored by home and away teams, we calculate the instances of home team winning/ losing/ drawing the match, given that they were ahead in the first half.

Conclusions: It is evident from the graph that the team, which is leading in the half time, wins the game.

Applications: This result helps the team management and team coaches to tell their players to press more in the first half, that is attack more in the first 45 minutes of the game.

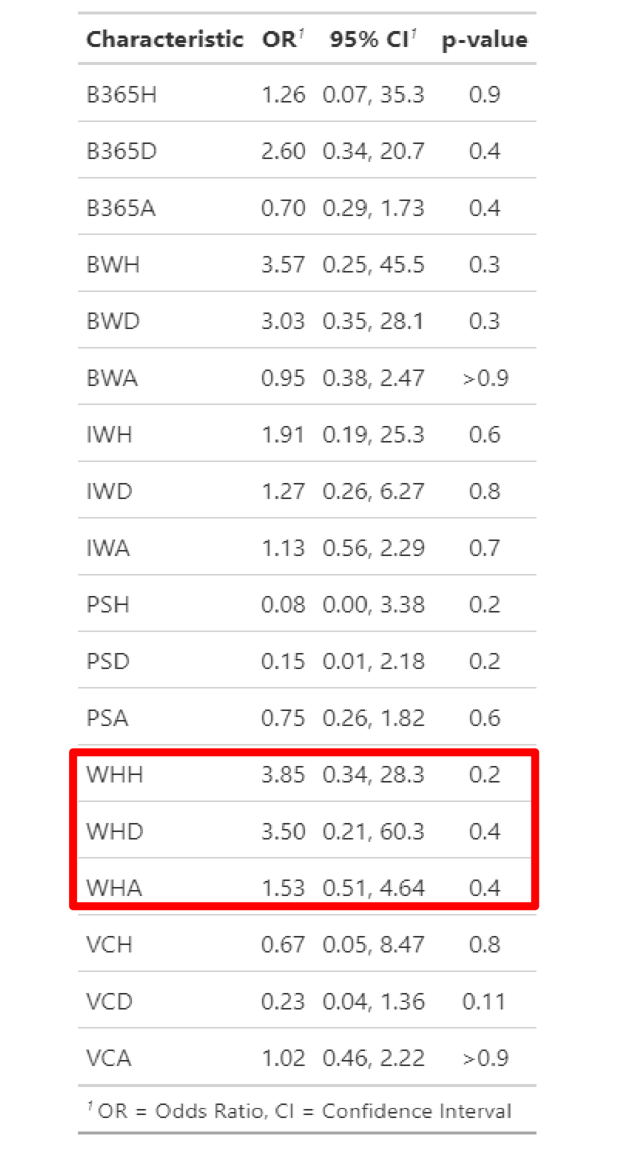


Which betting company provides a fair share chance of winning?

Methodologies: Applied 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.

Conclusions: William Hill has the 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 which means they all are significant.

Applications: We can suggest the potential betters to Bet on website as it provides a greater chance for the customers to gain profit.



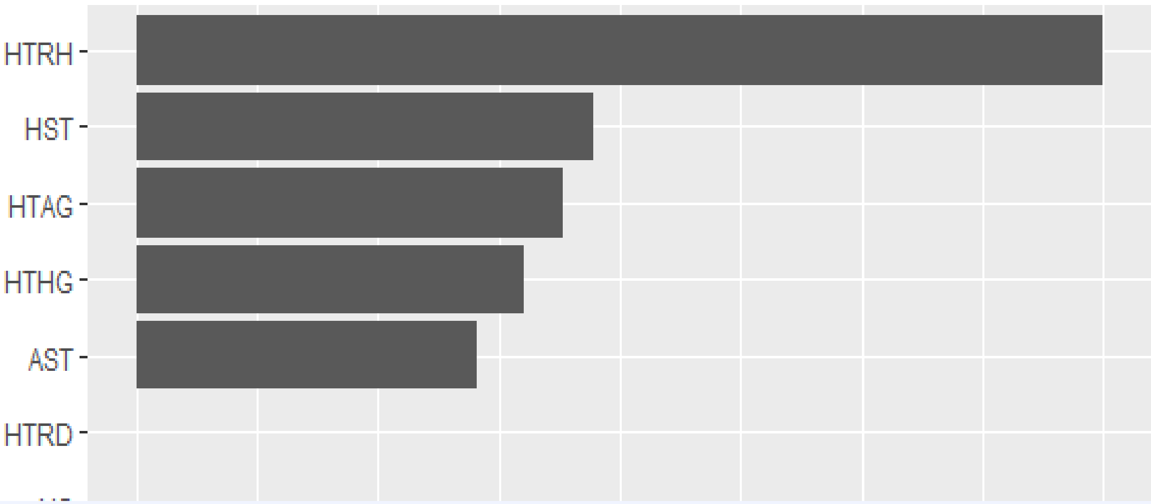
Characteristic	OR[1]	95% CI[1]	p-value
B365H	1.26	0.07, 35.3	0.9
B365D	2.60	0.34, 20.7	0.4
B365A	0.70	0.29, 1.73	0.4
BWH	3.57	0.25, 45.5	0.3
BWD	3.03	0.35, 28.1	0.3
BWA	0.95	0.38, 2.47	>0.9
IWH	1.91	0.19, 25.3	0.6
IWD	1.27	0.26, 6.27	0.8
IWA	1.13	0.56, 2.29	0.7
PSH	0.08	0.00, 3.38	0.2
PSD	0.15	0.01, 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

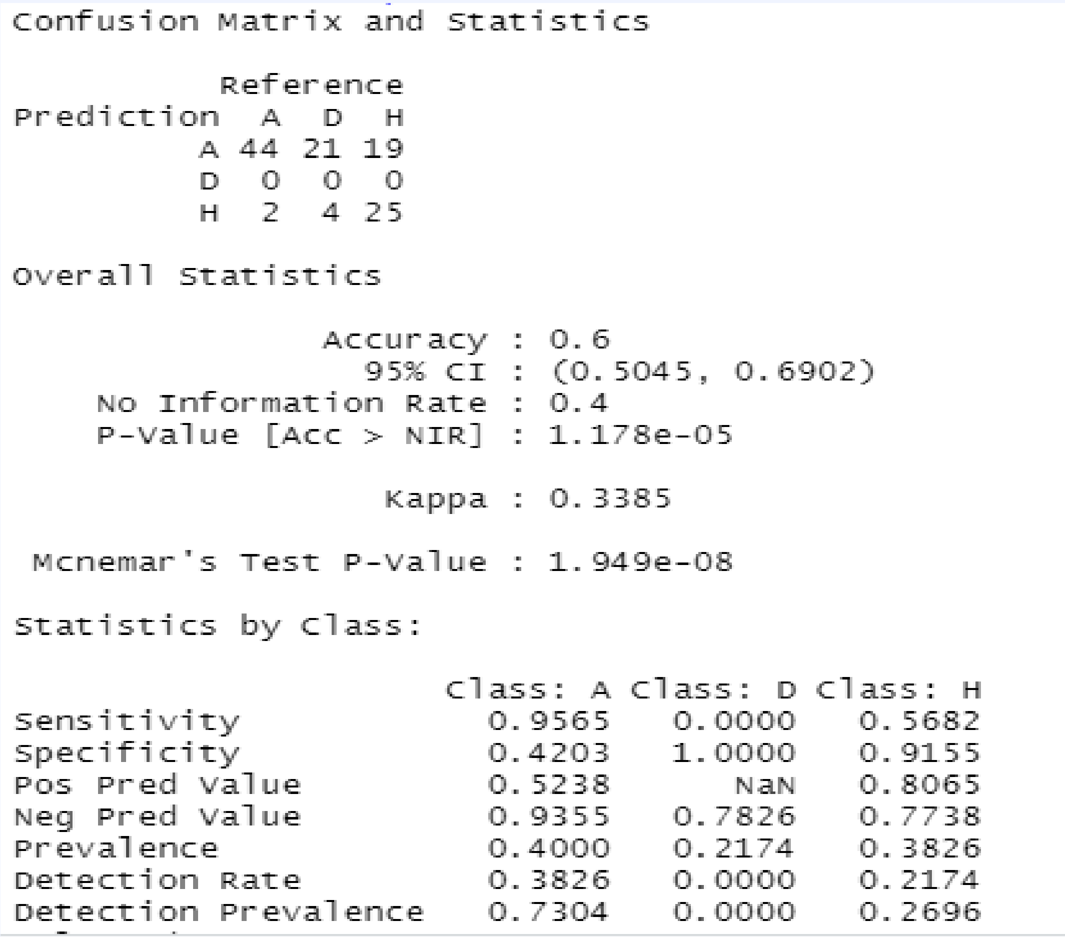
[1] OR = Odds Ratio, CI = Confidence Interval

Decision Tree for winning or losing:

Methodologies:

A decision tree with 60% accuracy giving the likelihood of an event occurring. Apart from full time home and away goals, what other factors are significant is shown in the graph.





```
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
```

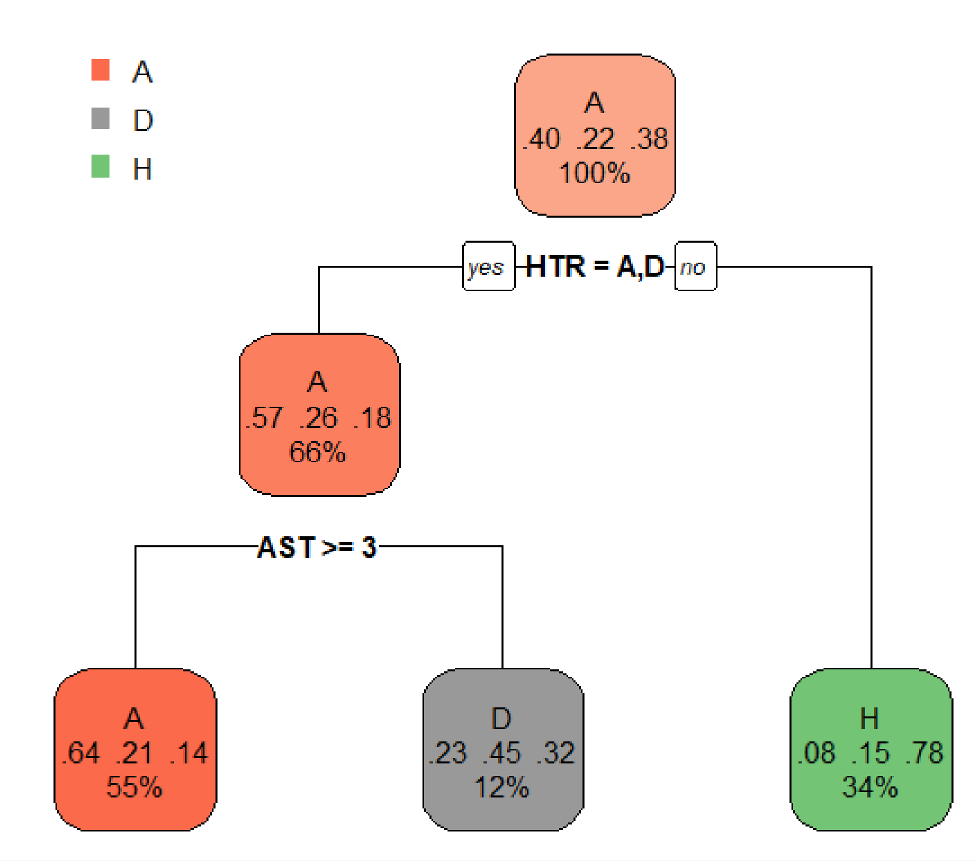
Decision Tree for winning or losing:

Conclusions:

- 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 it is not the case, then Home team has 34% chances of winning.
- 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.

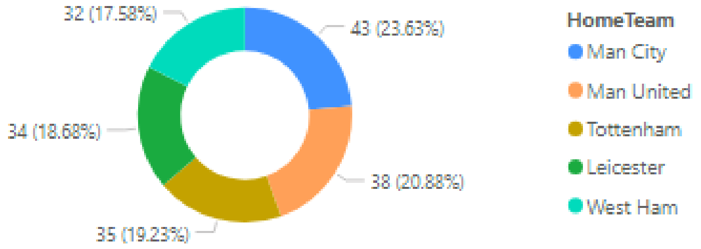
Applications:

To perform better in the match, away team should focus on scoring in the first half of the match. Later, in order to convert shots on target to goals, take at least 3 shots on the target.

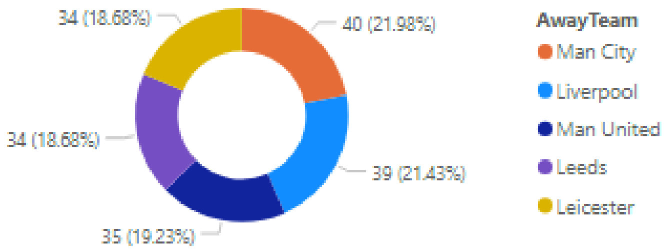


Performace by Top 5 Teams

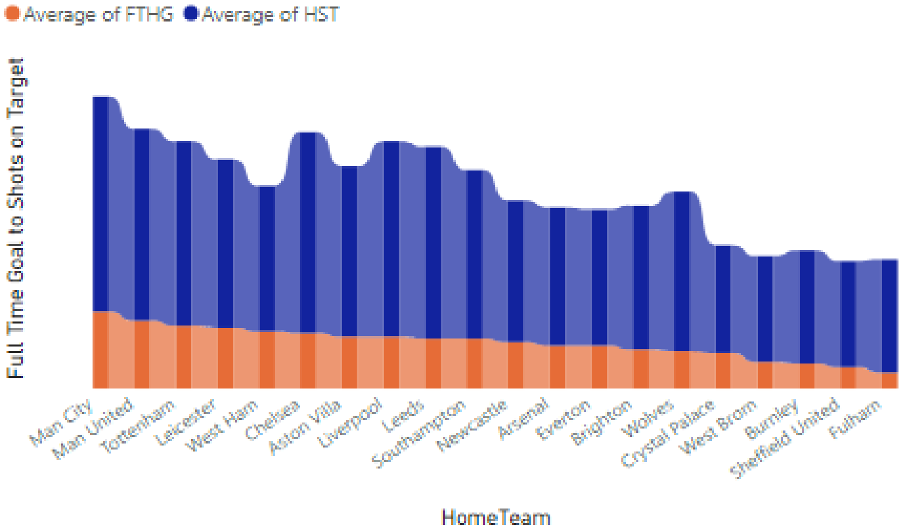
Full Time Home Goals



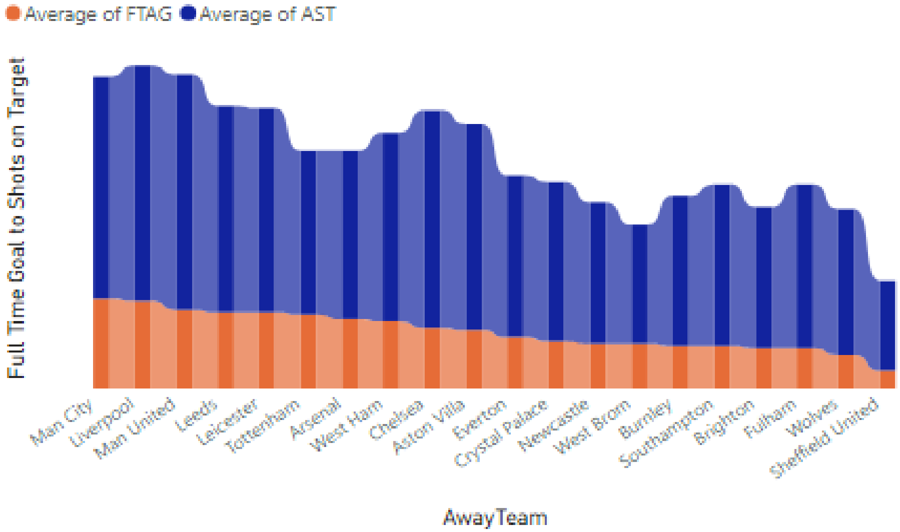
Full Time Away Goals



Conversion of Shots on Target to Goals (Home Team)



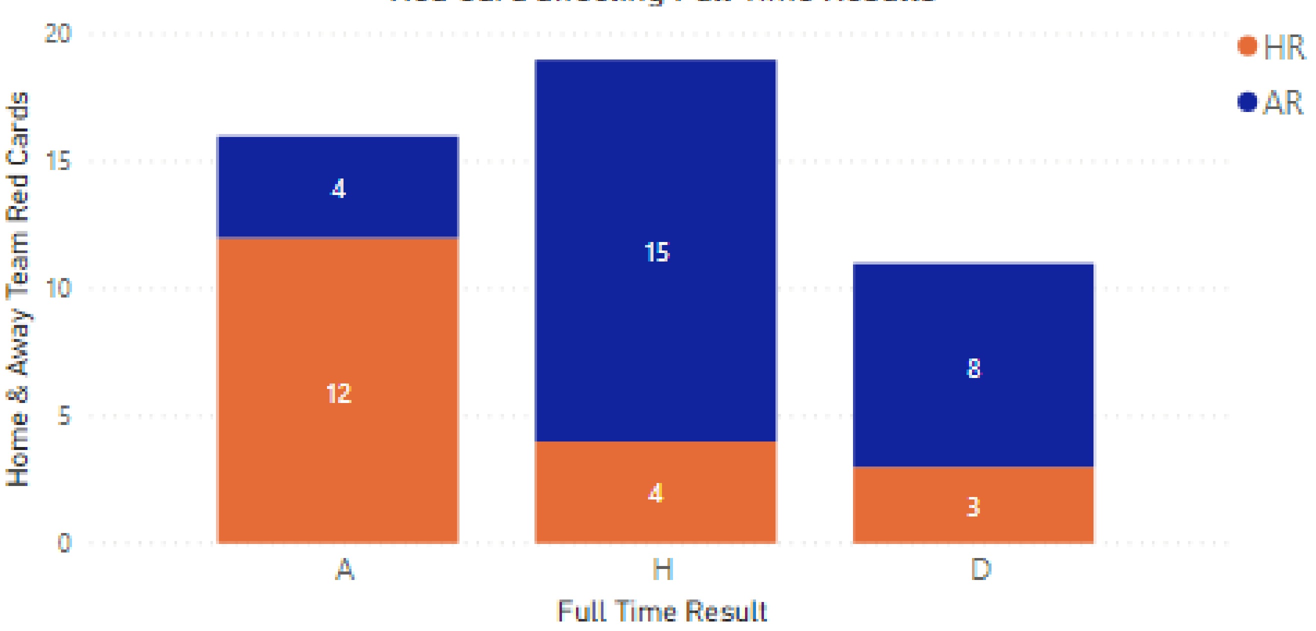
Conversion of Shots on Target to Goals (Away Team)



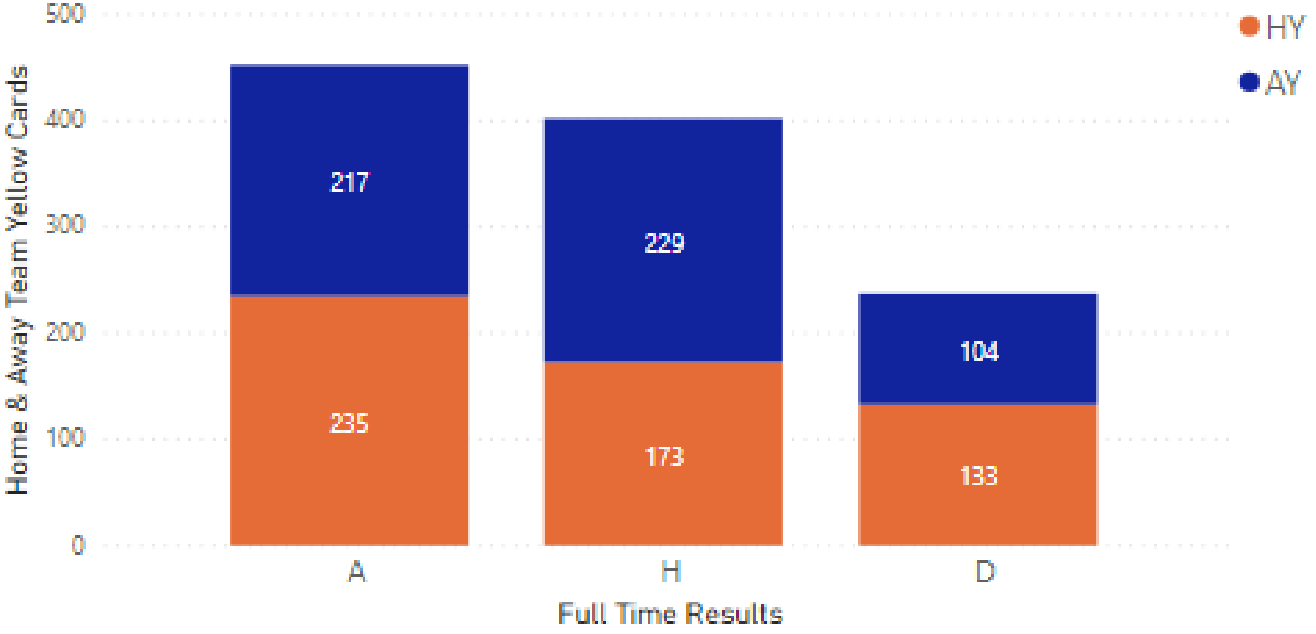
Key take aways:

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.



Red Card affecting Full Time Results



Yellow Card affecting Full Time Results

Key take aways:

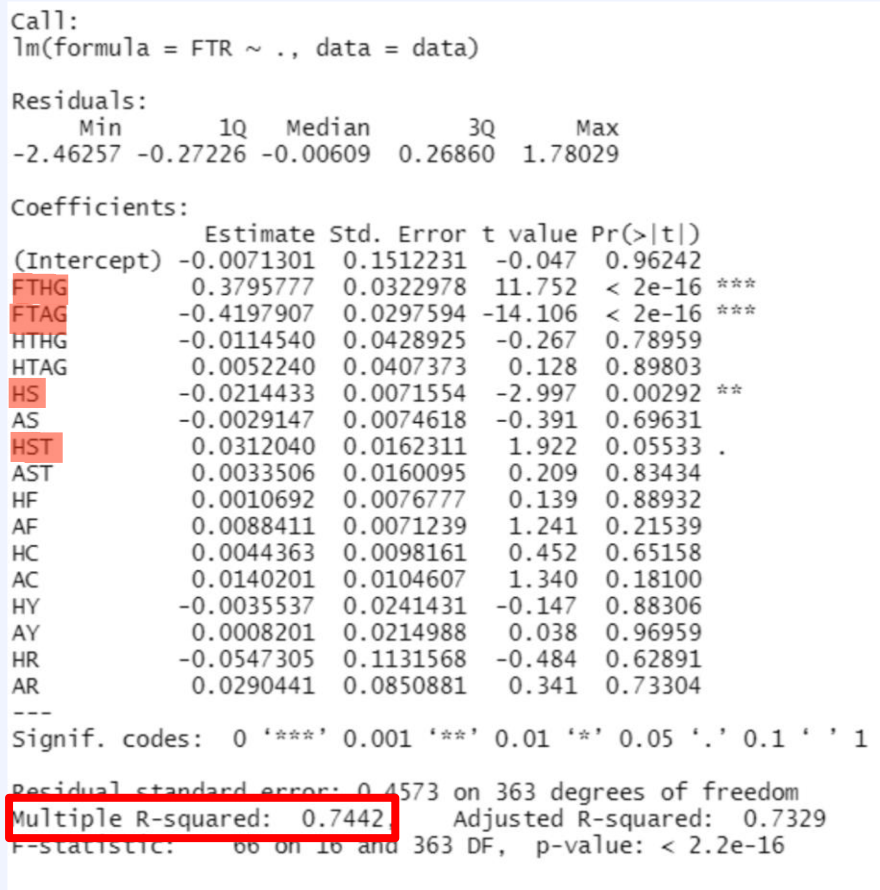
1. If away team gets red card more than away team, away team is likely to win. The same is applicable for yellow cards.

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.

CONCLUSION

- After applying LR model with Home and Away statistics, we can see that both the models gave same statistical results and can only explained 42% of the match outcome.
- Combined statistics of both Home and Away statistics can explain about 74% of the match outcome.
- This is the evidence that no match can is single handedly affected by Home or Away statistics.
- We conclude that there are handful of home and away statistics that affect the Full Time Result. But both the set of factors must be considered simultaneously to better calculate the influence of the match statistics on the Full Time Result.



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

Residuals:
     Min       1Q   Median       3Q      Max 
-2.46257 -0.27226 -0.00609  0.26860  1.78029 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0071301  0.1512231  -0.047  0.96242    
FTHG         0.3795777  0.0322978  11.752  < 2e-16 ***
FTAG        -0.4197907  0.0297594 -14.106  < 2e-16 ***
HTHG        -0.0114540  0.0428925  -0.267  0.78959    
HTAG         0.0052240  0.0407373   0.128  0.89803    
HS          -0.0214433  0.0071554  -2.997  0.00292 ** 
AS          -0.0029147  0.0074618  -0.391  0.69631    
HST          0.0312040  0.0162311   1.922  0.05533 .  
AST          0.0033506  0.0160095   0.209  0.83434    
HF           0.0010692  0.0076777   0.139  0.88932    
AF           0.0088411  0.0071239   1.241  0.21539    
HC           0.0044363  0.0098161   0.452  0.65158    
AC           0.0140201  0.0104607   1.340  0.18100    
HY          -0.0035537  0.0241431  -0.147  0.88306    
AY           0.0008201  0.0214988   0.038  0.96959    
HR          -0.0547305  0.1131568  -0.484  0.62891    
AR           0.0290441  0.0850881   0.341  0.73304    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4573 on 363 degrees of freedom
Multiple R-squared:  0.7442,	Adjusted R-squared:  0.7329 
F-statistic:    66 on 16 and 363 DF,  p-value: < 2.2e-16
```

All Statistics Model Summary

LIMITATIONS

- Dataset is very 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.

FUTURE SCOPE

- We could approach teams playing football and understand from them about various factors which influence them while playing a match.

- We can add additional data about the crowd like their attendance, their cheering and noise levels.

- We can incorporate data for multiple seasons and multiple leagues for our research.

REFERENCES

- 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.

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

THANK YOU