IS – 507

Final Project Rough Draft

- Asim Chitre, Faizan Ali 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, which we have obtained from https://www.football-data.co.uk/.

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 final 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.

The dataset that we took was very small and covered just one season of the Premier League. The dataset was clean with no null values. To answer the first research question, does half time results affect the final game results, we did some exploratory data analysis and finally made a bar chart to understand the results. It is evident from the graph that the team, which is leading in the half time, wins the game.

For the second and third 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 the 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 final 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 significant number of home and away statistics is related to the full-time result.

We have several limitations of the dataset, which are- 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. We can conclude that home and away factors play a very important 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.

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 very 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 are the most significant to determine the full-time results. We have used Variance Inflation Factor to check multicollinearity, 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 also 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:

The dataset used in our project is consists of statistics from the English Premier League, season 2020. The dataset consists of data like- goals scored by home and away team in half time, full time, number of shots on target, number of red and yellow cards received by the home and away team, data on betting odds, etc. We have collected the data from https://www.football-data.co.uk/.

The data can give us good insights about the home and away factors influencing the match statistics and also tell us which factors are more significant than the others so that different strategies can be formed to be able to optimize the team's performance in 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 model 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 and 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 an important factor in Soccer and will use regression models and hypothesis testing to analyze the data.

We will be using various 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:

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

For the first research question i.e., does the half time score affect the final results, the game of soccer is played in two halves, with 45 minutes in each half. Selecting the number of goals scored by home and away teams, as well as the result, we create another data frame. We create 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. Similar 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.

Coming to our next research question of finding whether home statistics affect the final results, the match is being played either in home stadium or away stadium. But whenever a match is being played at home stadium, it is a general notion that the odds are in favor of the home team. The home fan's termed as the 'twelfth man' as they add to the team performance. Teams perform better when they are playing in front of their home crowd.

For answering whether the home statistics turn the odds in favor of the home team, we first separate the home statistics variables from all statistics. This isolates the home statistics and gives us better idea of effects of home statistics on full time result. Then we check Variance Inflation Factor (VIF) to further check the multi collinearity. Ideally, the VIF should be less than 5 to ensure no multi collinearity but below 10 is acceptable as VIF between 5 and 10 means that there is mild multi collinearity. In our case all the variables have VIF less than 5 meaning there exists no multi collinearity between the variables.

```
> 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

Since there is no multi collinearity, we can 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.

For the third research question i.e., finding whether away statistics affect the final results, here we look at the away statistics of the match being played at the opponent's home stadium. We are

trying 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.

It is very interesting to note that the away statistics affect the result in negative way. Which is true as better the away team plays, less are the chances of home team winning. This relationship can be clearly established from the above correlation diagram. To add to this, FTAG (Full Time Away Goals), HTAG (Half Time Away Goals), AS (Away Shots), and AST (Away Shots on Target) are the factors that heavily influence the match outcome. These variables have the ability to alter the match results and can easily tip the match outcome in away team's favor. Further the away red card has a positive correlation here which is very similar to the home statistic correlation. This is because if any of the team, be it home or away loses an on-field player due to red card, then the other team will heavily benefit from this and may have a chance to win the match.

The correlation plot also shows that there are no variables that are highly correlated. Hence, there is no need to eliminate any variable. We now look at VIF (Variance Inflation Factor).

```
> 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

The VIF is below 5 and well below 10. This shows that there is no multi collinearity within the dataset and gives a green signal that the variables are good to be used as predictor variables for our Linear Regression Model. Hence, we move on with the available variables and model a Linear Regression with target variable as FTR (Full Time Result).

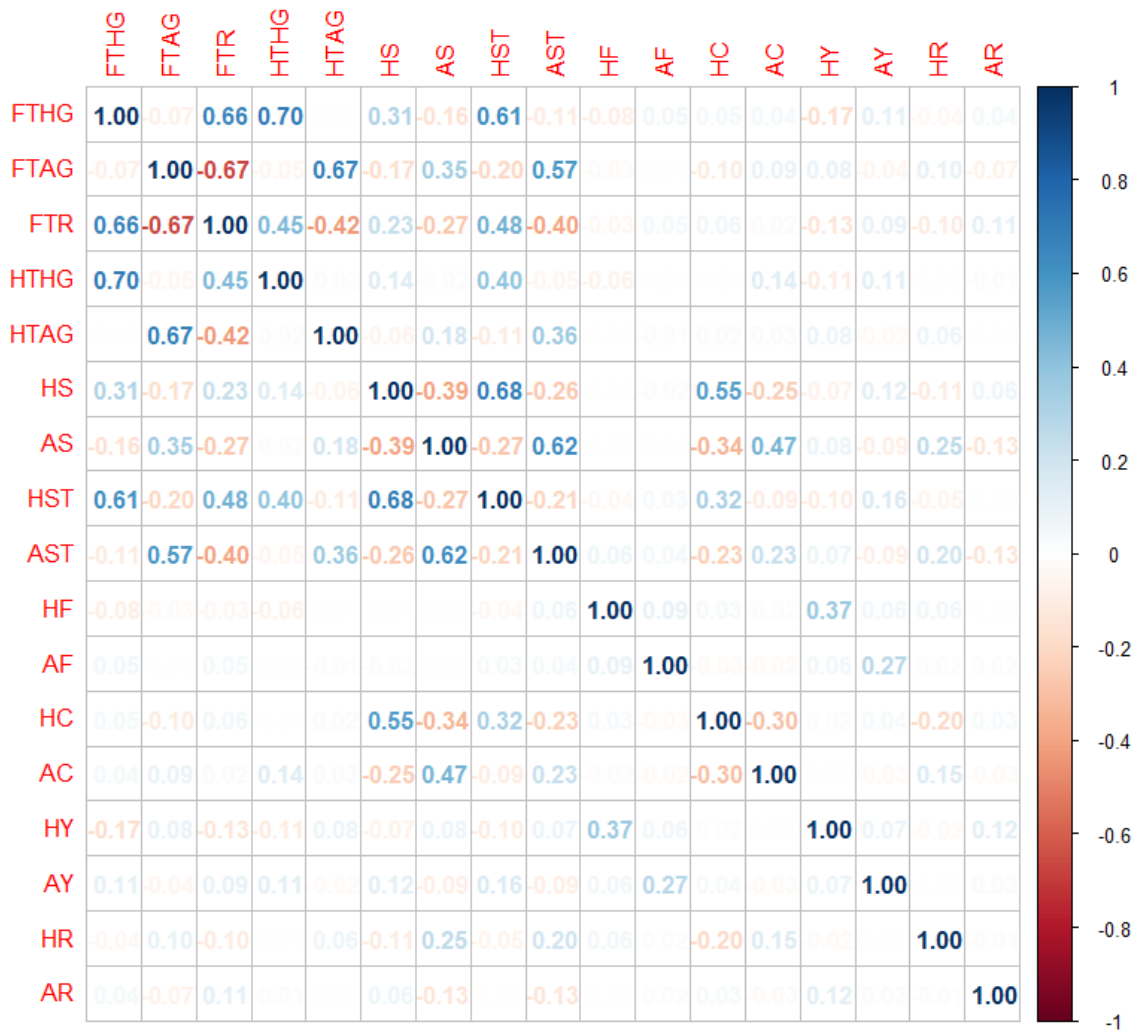
Moving forward with the next research question i.e., how do match statistics help shape the final results, match statistics are an important part of the data collected by football teams and the

authorities. These statistics can be used to effectively establish relationship between various kind of outcomes. 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.

For our last research question i.e., which betting company provides a fair share chance of winning, betting companies rely on customers for their income. A common notion of betting is that the "house" should earn money and for that same reason, the betting companies try to keep the probability of customers profiting from the best much lower. 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.

Discussion and Results:

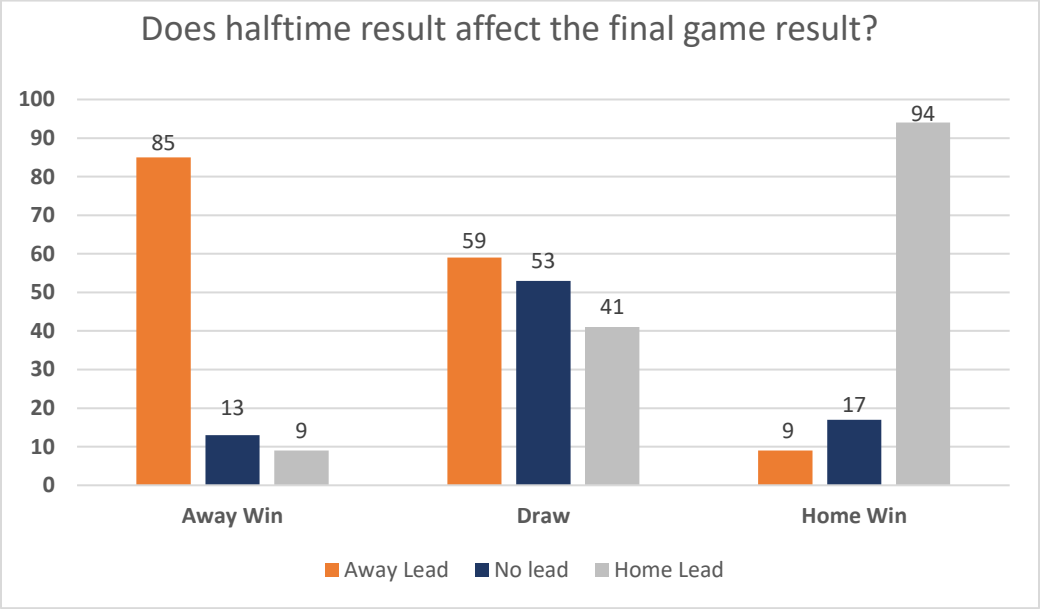
In this section, we will discuss about all the mentioned hypotheses and their results.



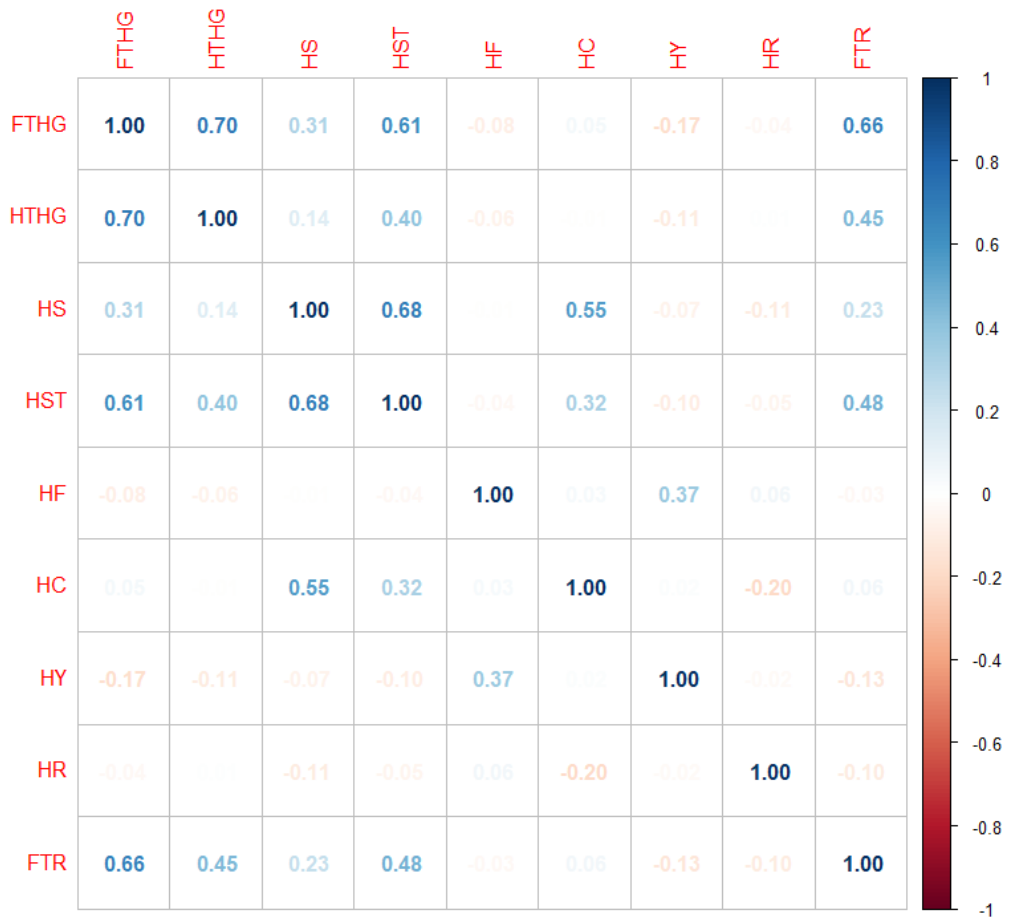
All Statistics Correlation

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. 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.

For the first research question, the below graph shows the number of matches that each possibility happens. 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.



Next, 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.



Home Statistics Correlation

In this case, 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 clearly 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 prove with data that just taking shots towards goal does not pressurize the away team, rather taking shots on target does help the result to tip in the favor of home team. This analysis is very 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 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.

Just from the coefficient we can establish very significant relationship. The model summary for home statistics dives deeper and explains more about the effect of home statistics on full time result.

```
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
```

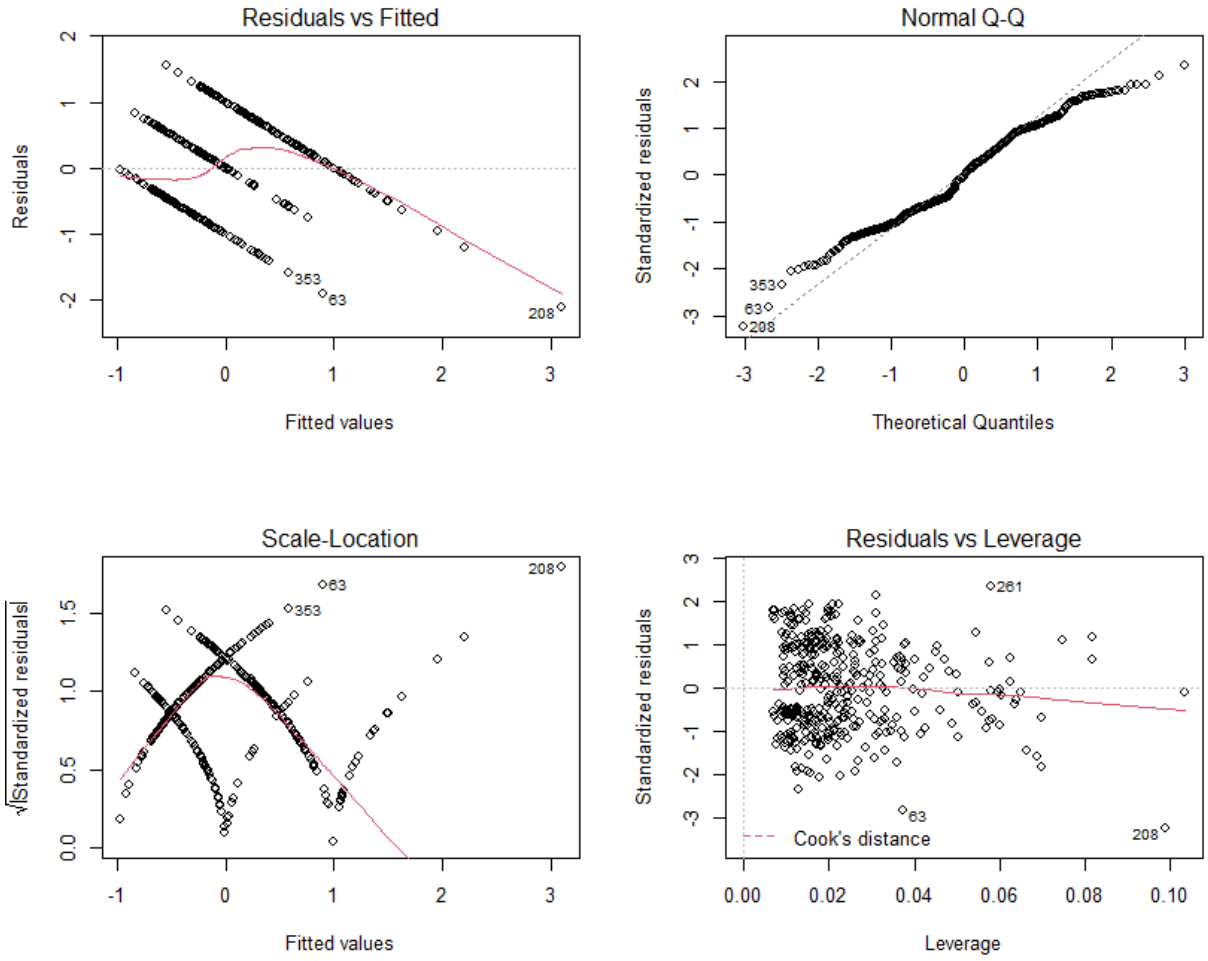
Model Summary for Linear Regression Model of Home Statistics

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 effectively 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). In conclusion, out of all home statistics, the above mentioned three statistics significantly affect the results.

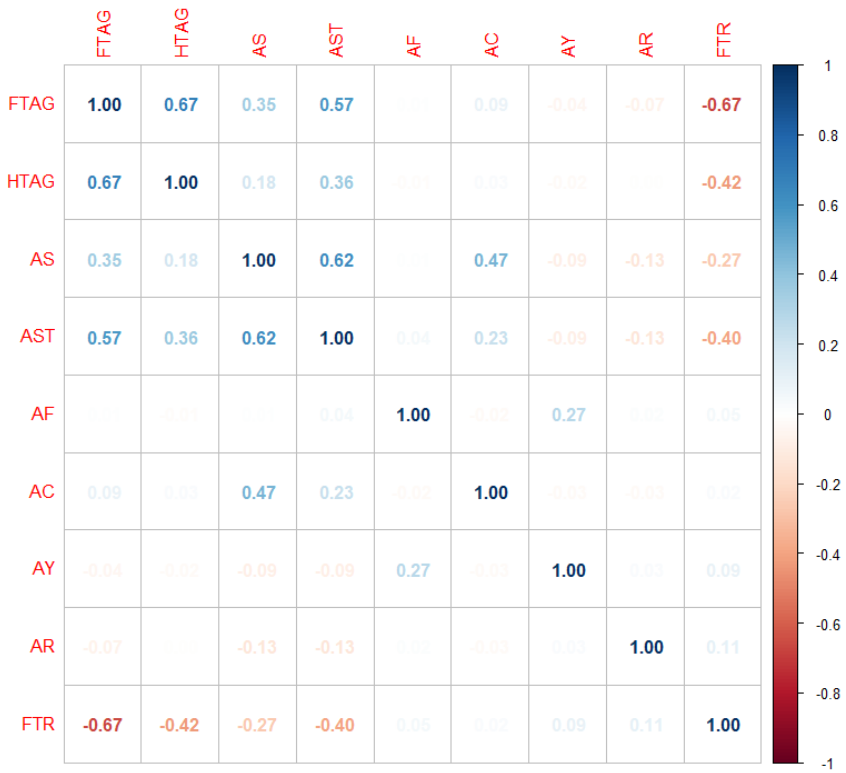
Moving on to diagnostic plots for the Linear Regression Model. The diagnostic plots help us understand the Linear Regression Model in much deeper sense. The diagnostic plot talks about various metrics like overfitting, residual, normality etc., and can be easily interpreted by only seeing the graphs.

The first plot that we see is the Residual vs Fitted plot. The main goal of this plot is show if there is a linear relationship between our predictor variable and target variable. In this case, it is very clear from the first plot that there exists a linear relation to some extent but overall, there is no linear relationship. Next, we look at the second plot which Normal Q-Q plot. This 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 third plot is the Scale-Location plot. To put it in simple term, this plot shows how the residuals are spread or in statistical term, Homoscedasticity. Although, the residuals are not spread randomly, i.e., they do follow a pattern, but the line is not totally horizontal. This indicated the presence of random spread of residuals within our model. The last plot is Residuals vs Leverage. This plot serves the purpose of identifying influential outliers within our total spread of the residual values. In this case, we see that all the outliers are within Cook’s distance and hence there is no presence of influential outliers.



Home Statistics Diagnostic Plot

Now we move on with our third research question, The coefficients just reiterate on the conclusions that we drew from the correlation plot.



Away Statistics Correlation

The FTAG (Full Time Away Goals) and AS (Away Shots) work against the home team as they negatively affect the outcome and push it in away team's favor. The AR (Away Red cards) work in favor of the home team and hence have a positive coefficient.

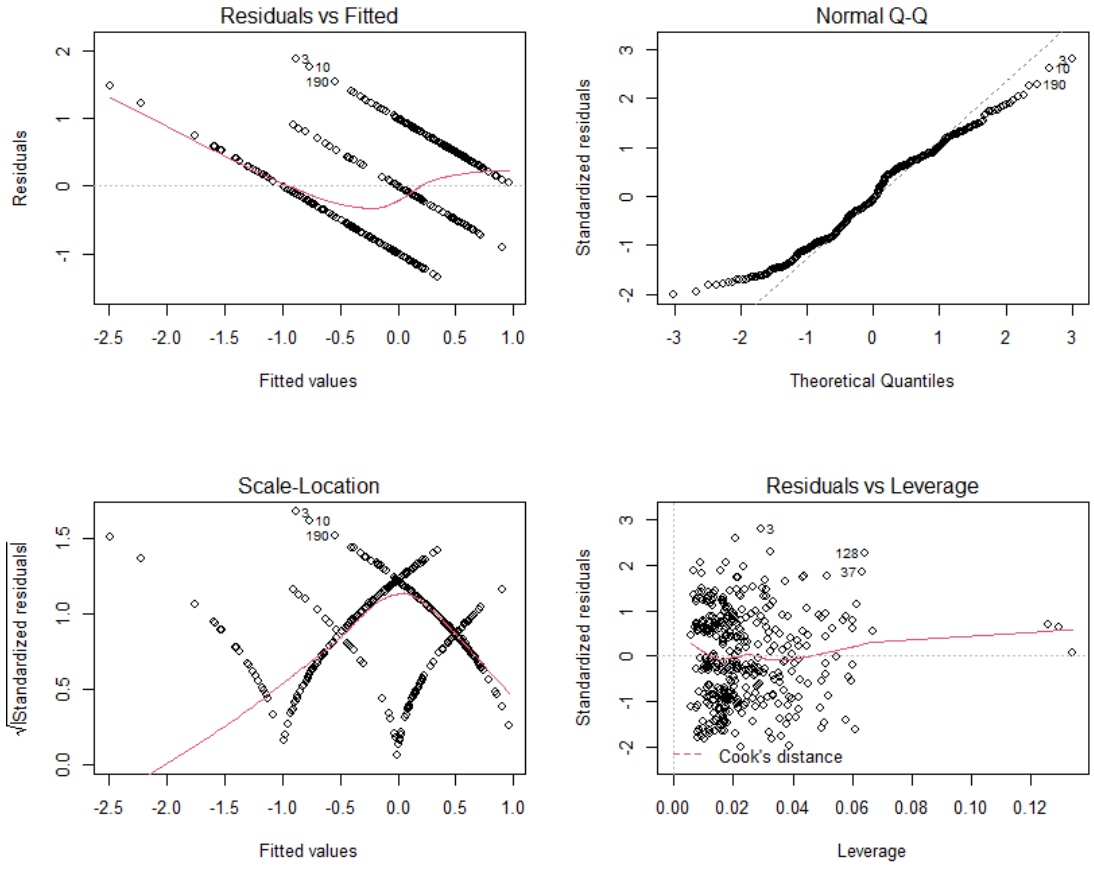
```
Call:
lm(formula = FTR ~ ., data = awayStats)

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
```

Model Summary for Linear Regression Model of Away Statistics

Looking at the summary of Linear Regression Model with Away Statistics as the predictor variable and FTR (Full Time Result), we see very similar summary statistics. The Multiple R-squared is roughly 42%. That means about 42% of the outcome is explained by our model. The residual error is also less. But here we see three significant variables. When we look closely, 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 effect the match outcome. Although the numbers of this model are very similar to the numbers shown by home statistic model, the explaining variables are different. In case of away statistics, away corners play

an important role. This single outcome can be very much influential in turning around the results in the favor of away team. 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.

To explain the model better, we now look at the diagnostic plot of the regression model. This will specifically talk about the model and not about the research questions.



Away Statistics Diagnostic Plot

From the first plot, i.e., the Residuals vs Fitted plot, we conclude 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 second plot, i.e., 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 third plot, i.e., 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 fourth plot, i.e., Residuals vs Leverage shows that there are no influential outliers. We need to check whether an outlier is influential or not before we eliminate it. In this case, there are couple of outliers but since they lie within the Cook's distance, they are not that influential. Hence, no influential outliers exist in our residuals.

Furthermore, our next research question i.e., which betting company provides a fair share chance of winning, for keeping a biased decision in respect to winnings, we will keep 50% confidence interval. This will ensure that betting companies as well as the betters 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 that WH which is William Hill, has been affected by the final results as it has 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 answers our research question.

Characteristic	OR[†]	95% CI[†]	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

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

Conclusion:

The conclusions are very sensible when we talk practically about soccer as a sport. Hypothetically, if home statistics heavily influence the match outcome, then almost 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 and square 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.

```
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

Here, we can see that combined statistics of both home and away teams are able to explain about 74% of the match outcome. This is a significant jump from 42%. This is evidence that no match is single handedly affected by home or away statistics. Rather 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 score less goals. This will ensure the home team victory. But as we concluded above, there are some factors that can help the either of the team perform better. Like for 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 in no one can ensure that either of the team wins.

Hence, we conclude that there are handful of home and away statistics that affect the Full Time Result. But in reality, both the set of factors have to 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.