Predicting the Premier League

By: Jake Haggard

Table of Context

I. Visualization

II. Models

III. Insights and Recommendations

I. Visualization

Exploring the six visualizations that were made for the project showed the characteristics of the data and how it can be best used in producing the model. The visualization consists of relationships between seasons, game results, home team aggressions, shots on target from away and home, and the average fouls called by the referee. Through these relationships the discovery of shots on target at home and away

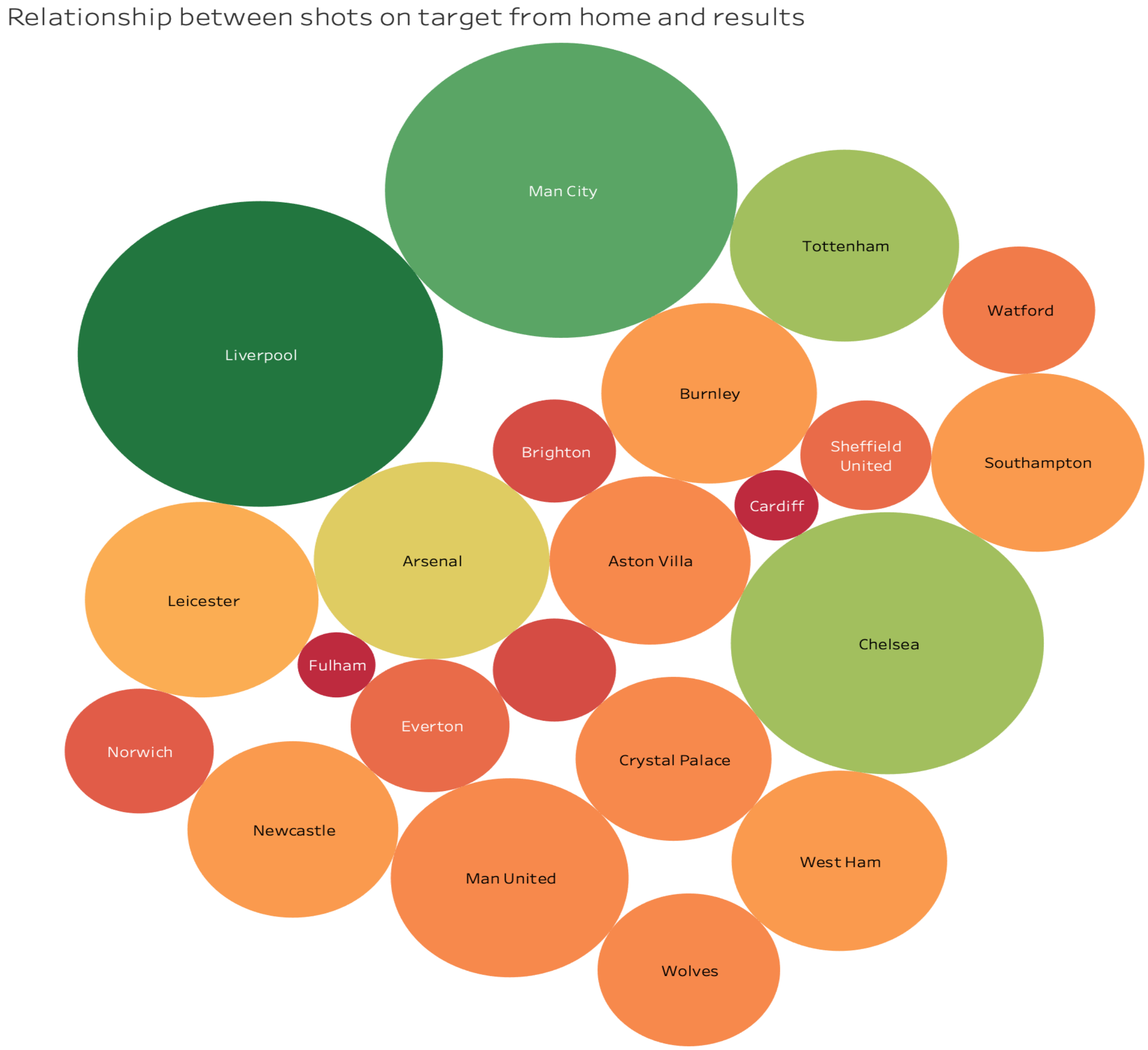

Figure 1

have a major effect on the result of the game. In figure 1.1, it is possible to understand that the team with the most shot on target wins most of their games. In figure 1.1 as well one can say ‘of course, bigger teams will always have more shots on target and more likely to win’ but with figure 1.2, shots on target when away from home, mostly underdogs have the advantage. This shows that there is a correlation between bigger teams performing better at home than away and vice versa, smaller teams performing better when they are away from home. To also to introduce a factor that shows the likelihood

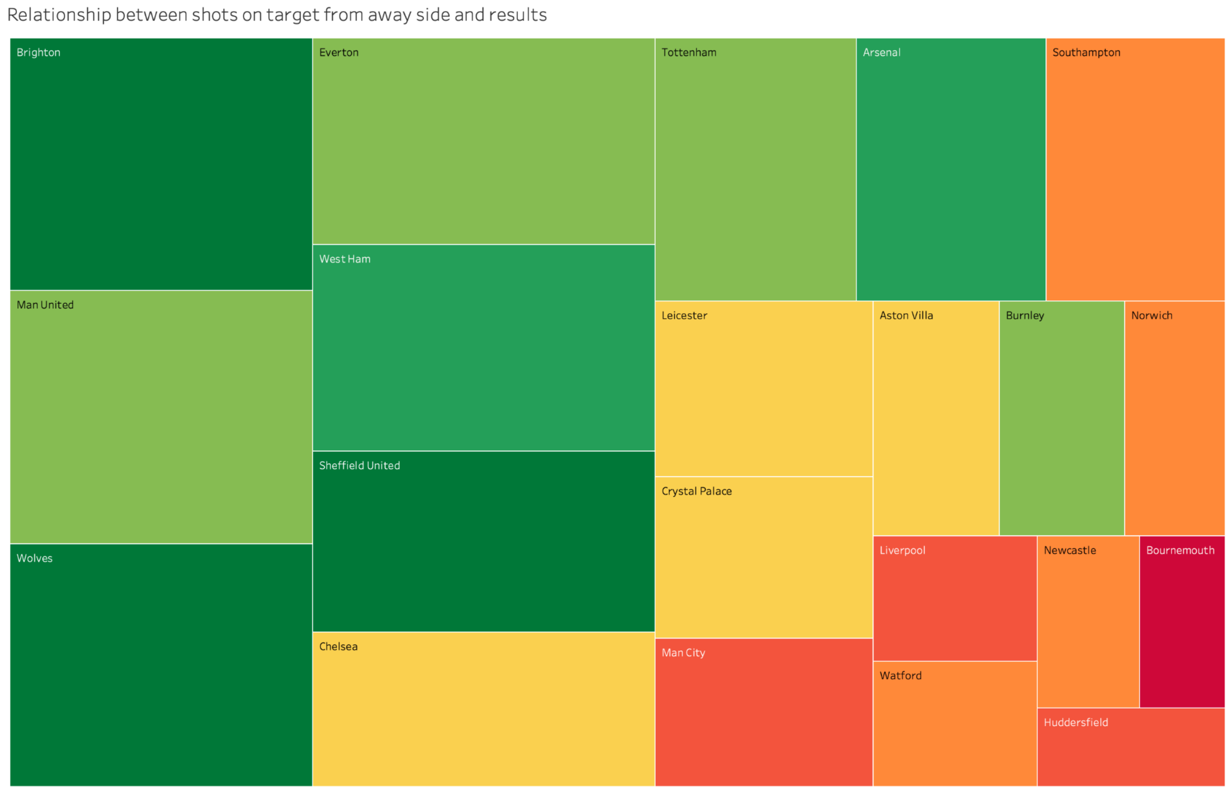

Figure 1.2

of fouls called per game by referees have a factor on the result, figure 1.3 also shows the average of fouls called per game by certain referees, to see if it has a factor on the result.

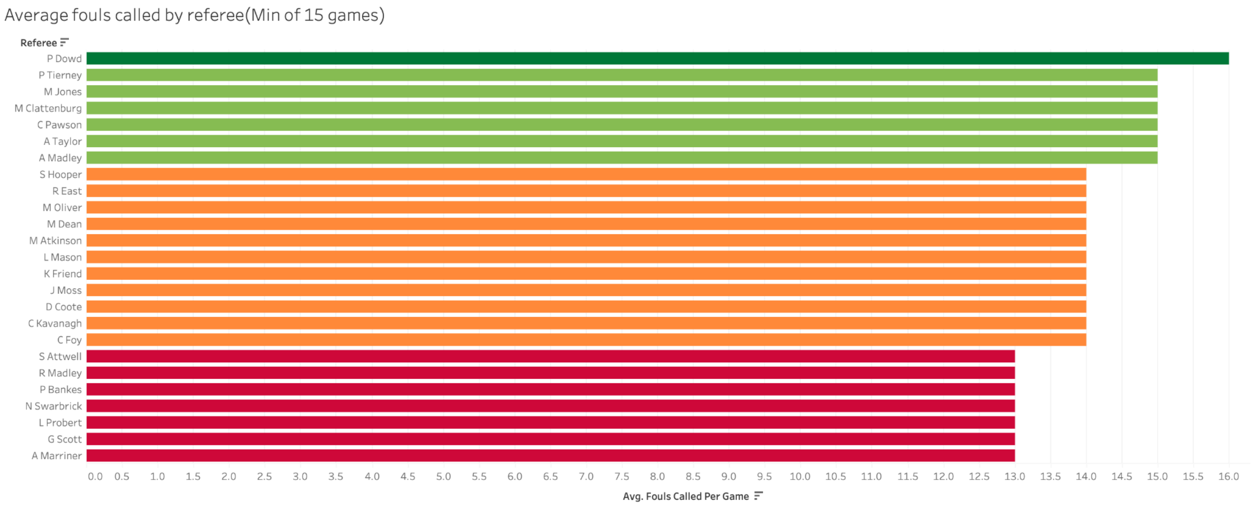

Figure 1.3

To do this the exploration of games won at home and away and its relationship with more/less frequent foul calling officators have on the results through figure 1.4, 1.5, 1.6, and 1.7. As it shows that no matter who's officiating the game it does not show any impact or associations on the result of the game.

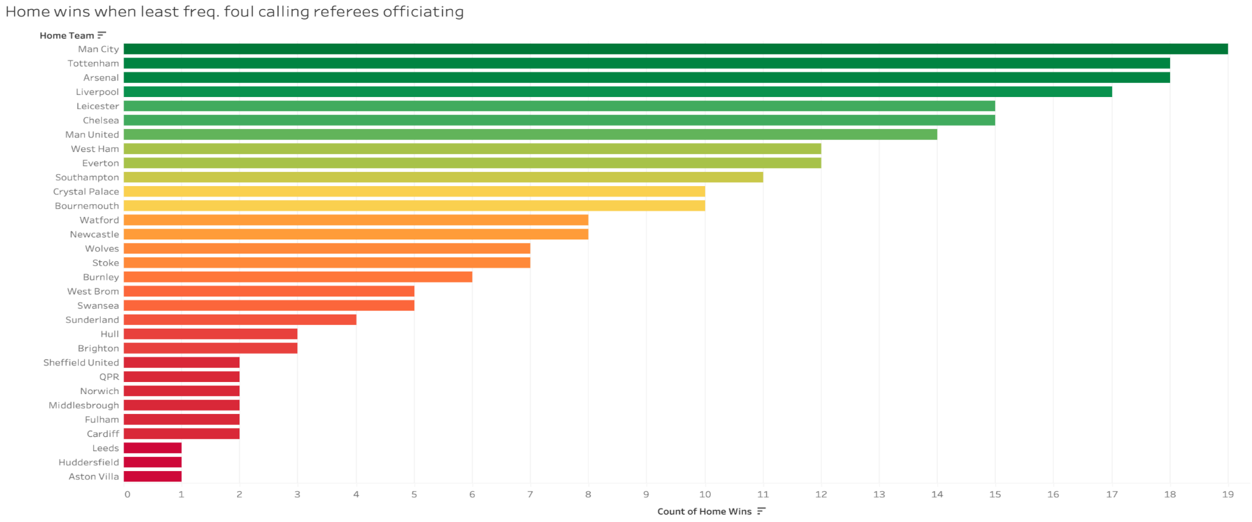

Figure 1.4

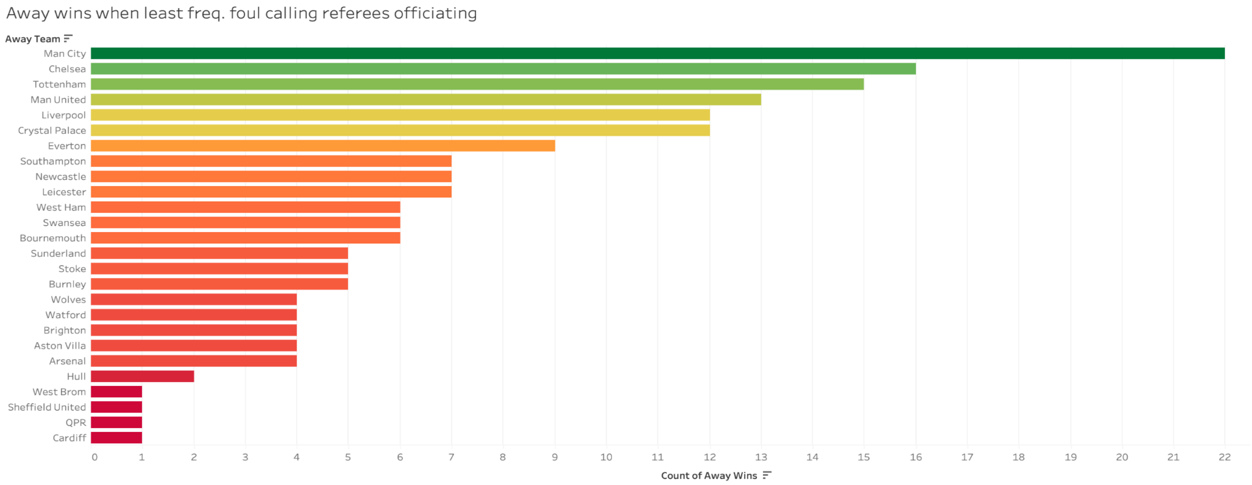

Figure 1.5

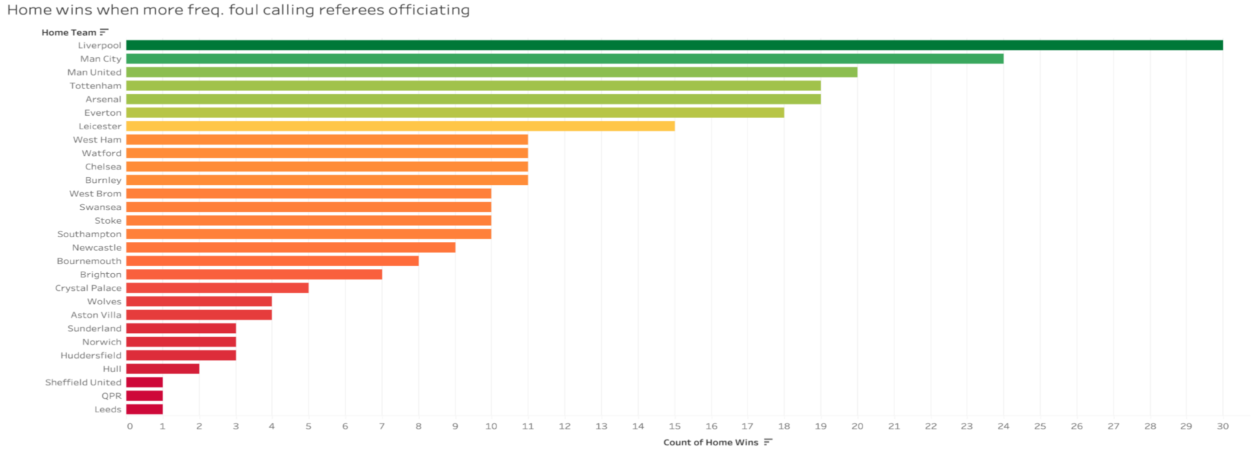

Figure 1.6

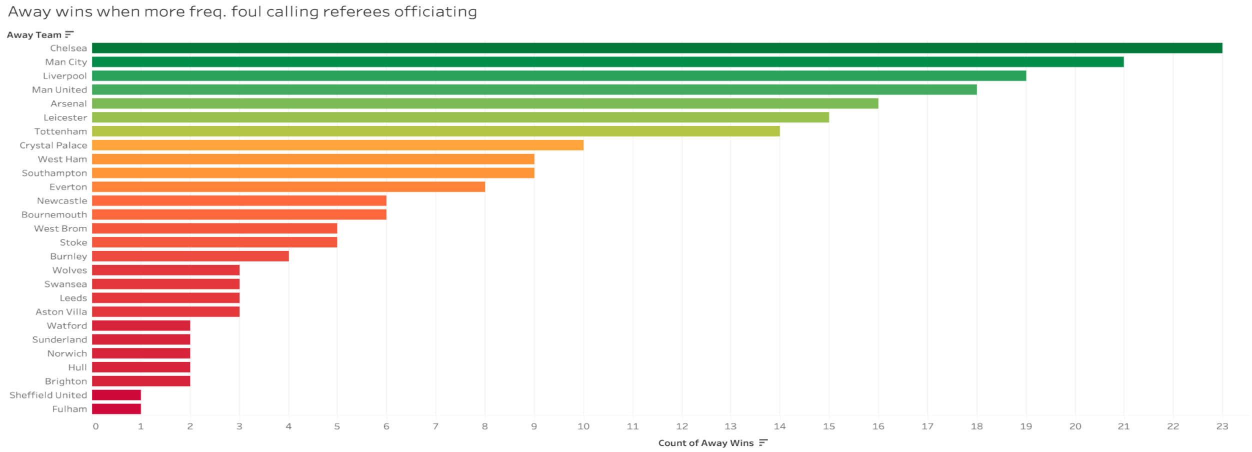

Figure 1.7

To examine the home performance of each team I made a visual that showcased the amount of wins at home throughout the course of the year. A team's performance can be influenced by transfers, injuries, and manager changes throughout the season. I chose wins at home because winning at home is something that is desired across all sports, the best teams win consistently at home in front of their supporters. This visual was filtered by each season of the year and the size of the circle is determined by the number of home wins for that respective team. This visual helped I gather the insight that a team's performance can alter over the course of a year, you can see that Manchester United got less wins as the year went on in the visuals below. The four visuals below are all the same, just filtered with respect to each season; the order is fall (1.8), winter (1.9), spring (1.10), and summer (1.11).

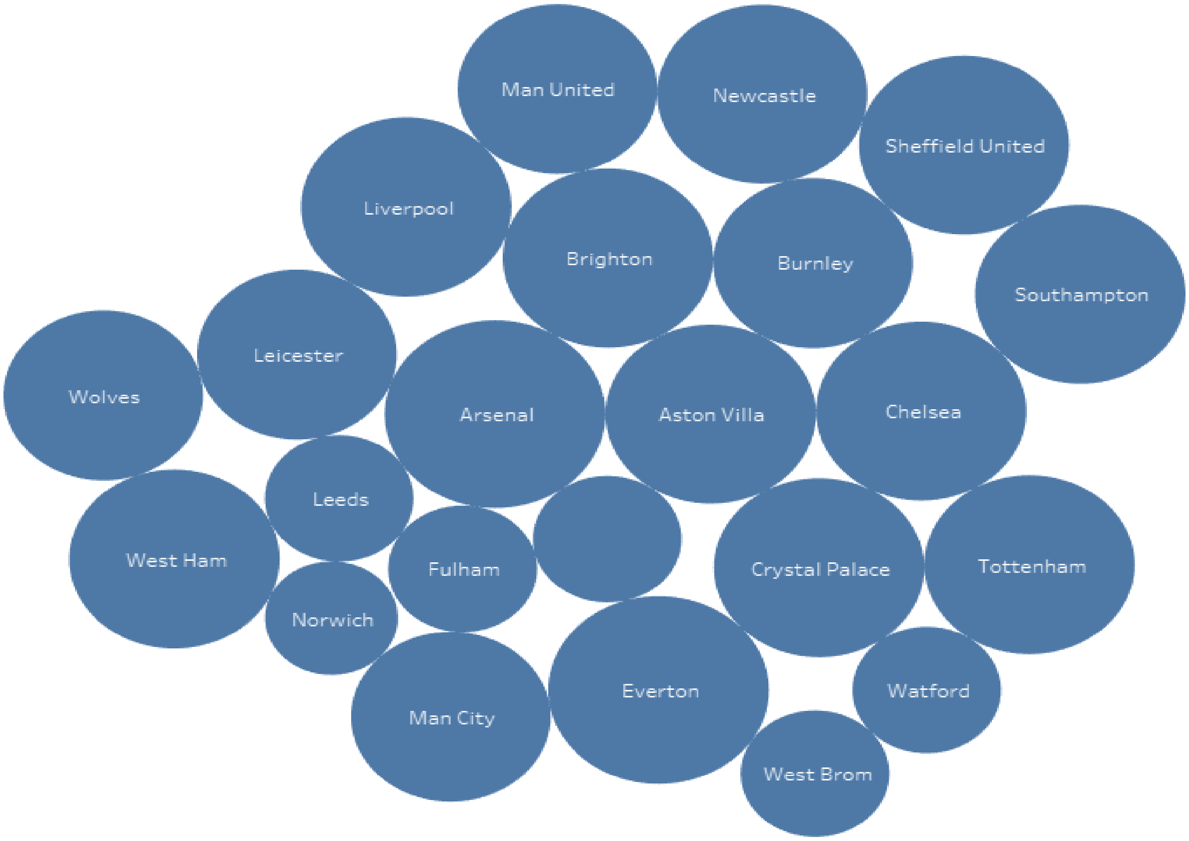

Figure 1.8

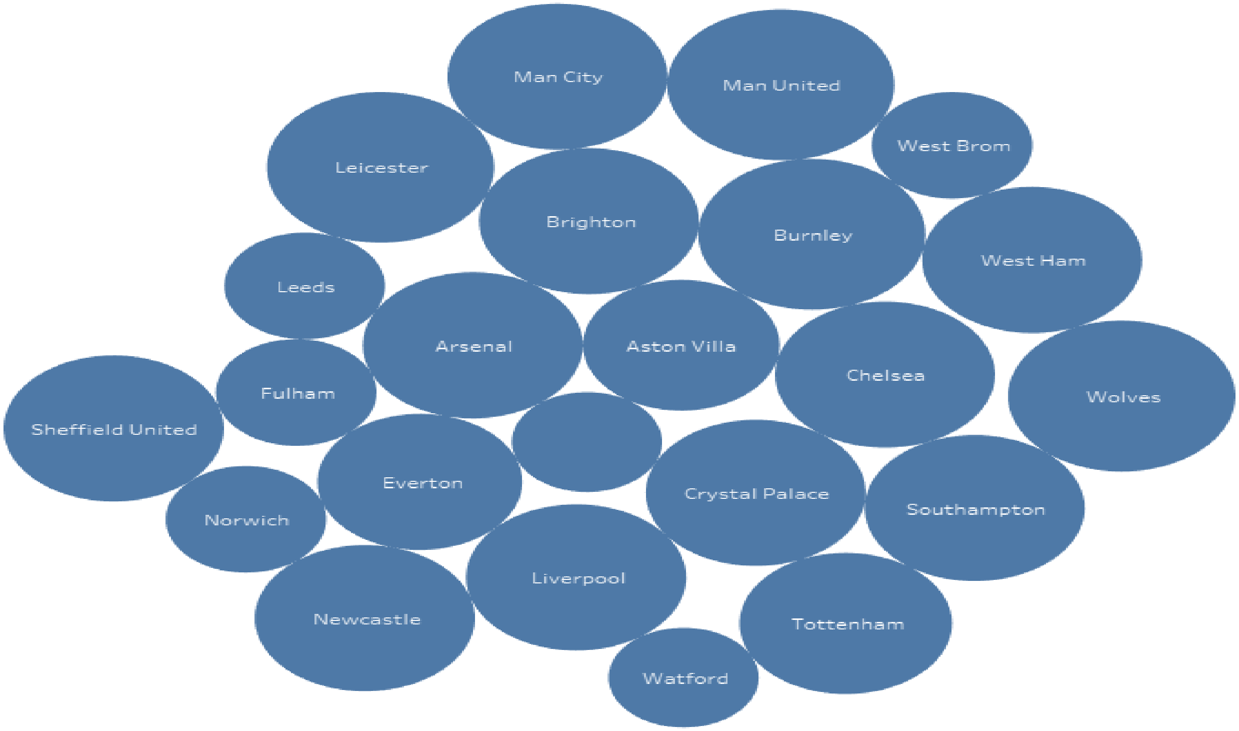

Figure 1.9

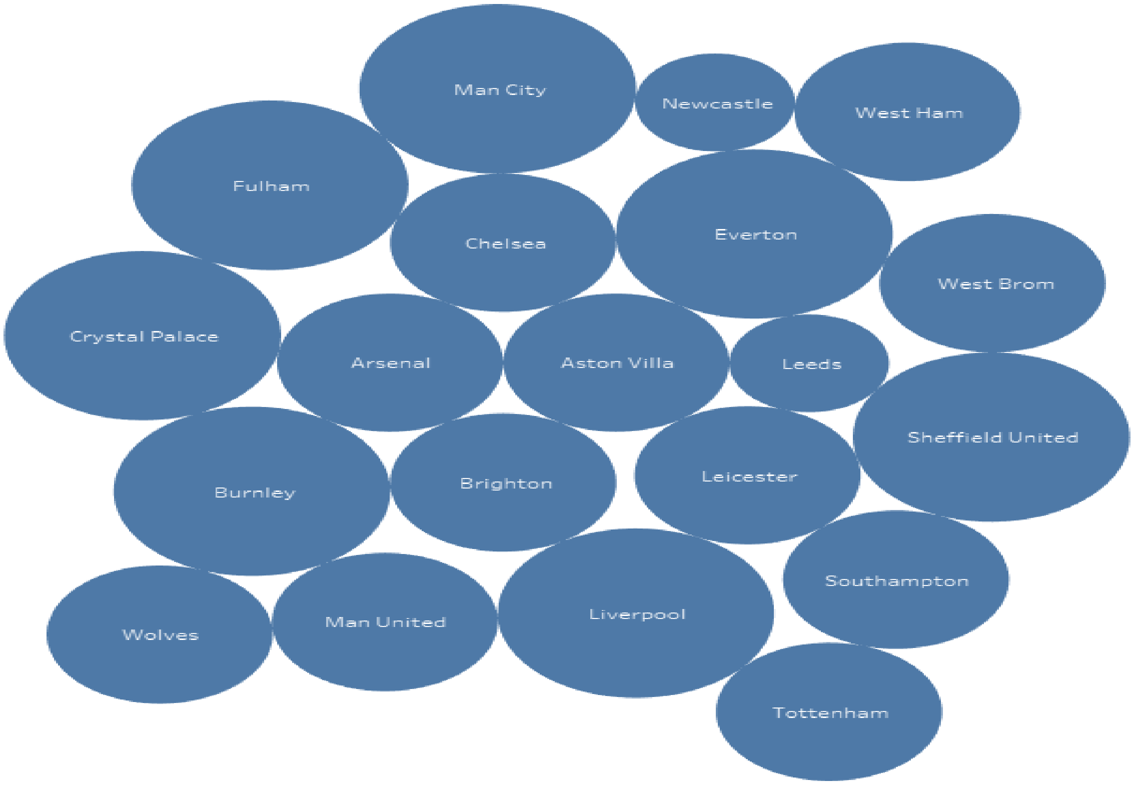

Figure 1.10

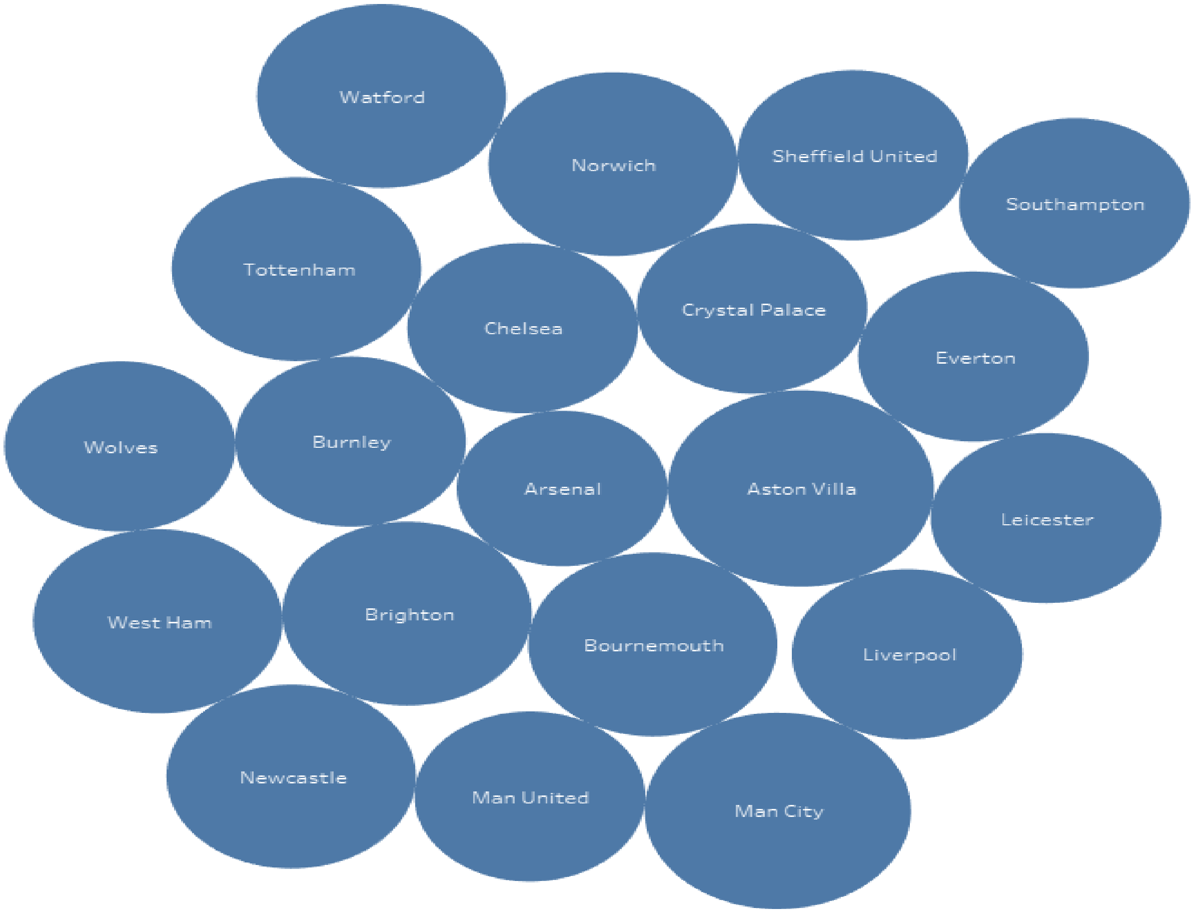

Figure 1.11

After investigating the performance of teams across the four seasons of a year, I wanted to examine the performance of a team during the course of a day. Essentially I wanted to figure out if the time of a day impacted a team's performance as well as the current season. In sports teams play at various times and at times it can seem to make an impact. If a team that normally plays later in the day ends up playing in the afternoon they can start off as sluggish as they aren't used to playing at that time. One major takeaway was teams tend to win less at home during the summer in the afternoon, this could be because they are playing earlier in the day and as well as playing in the heat however, there could also be more reasons that I was unaware of. I have the visual below filtered by each time period within a day; afternoon (1.12), mid-day (1.13), and late-day (1.14).

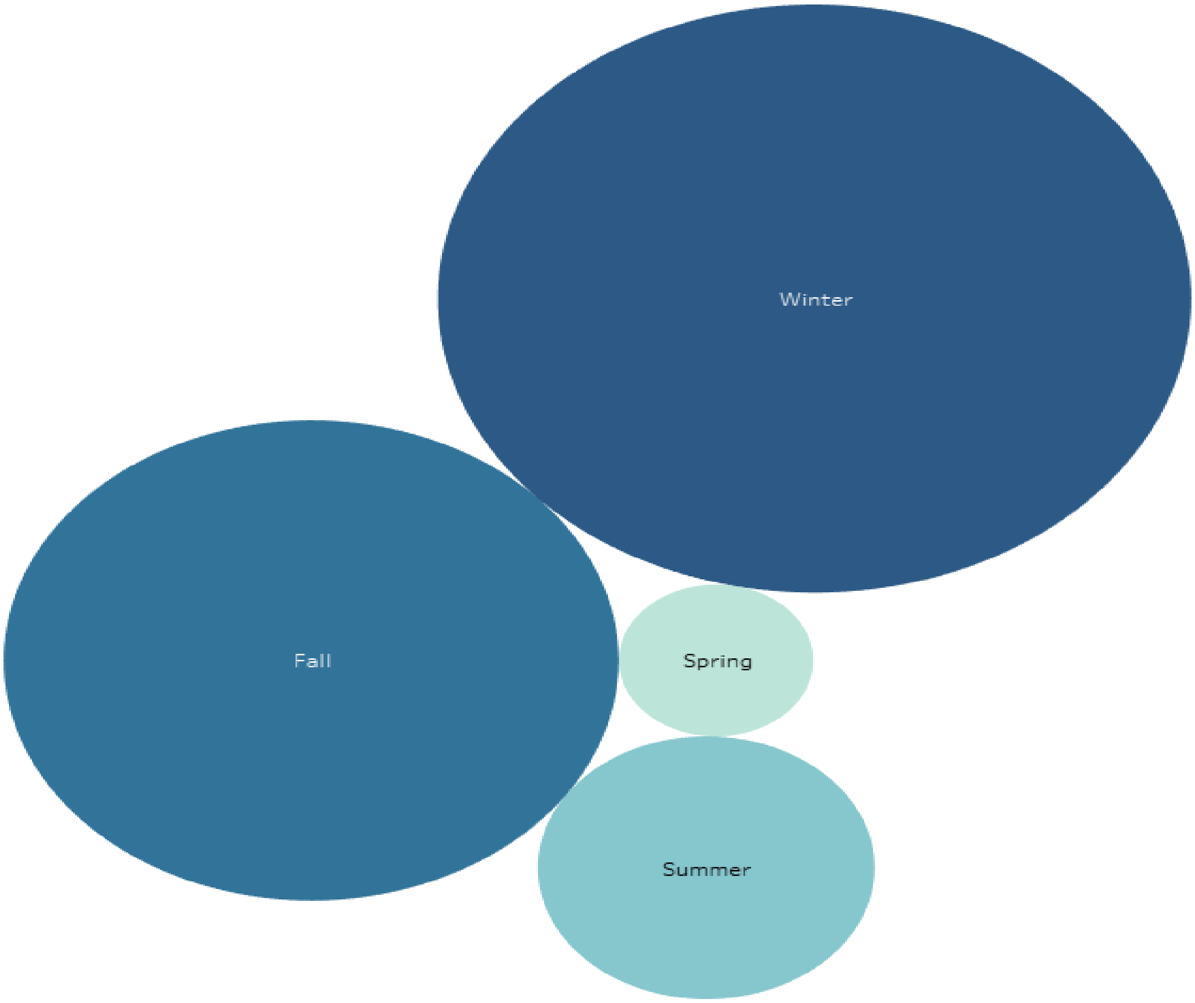

Figure 1.12

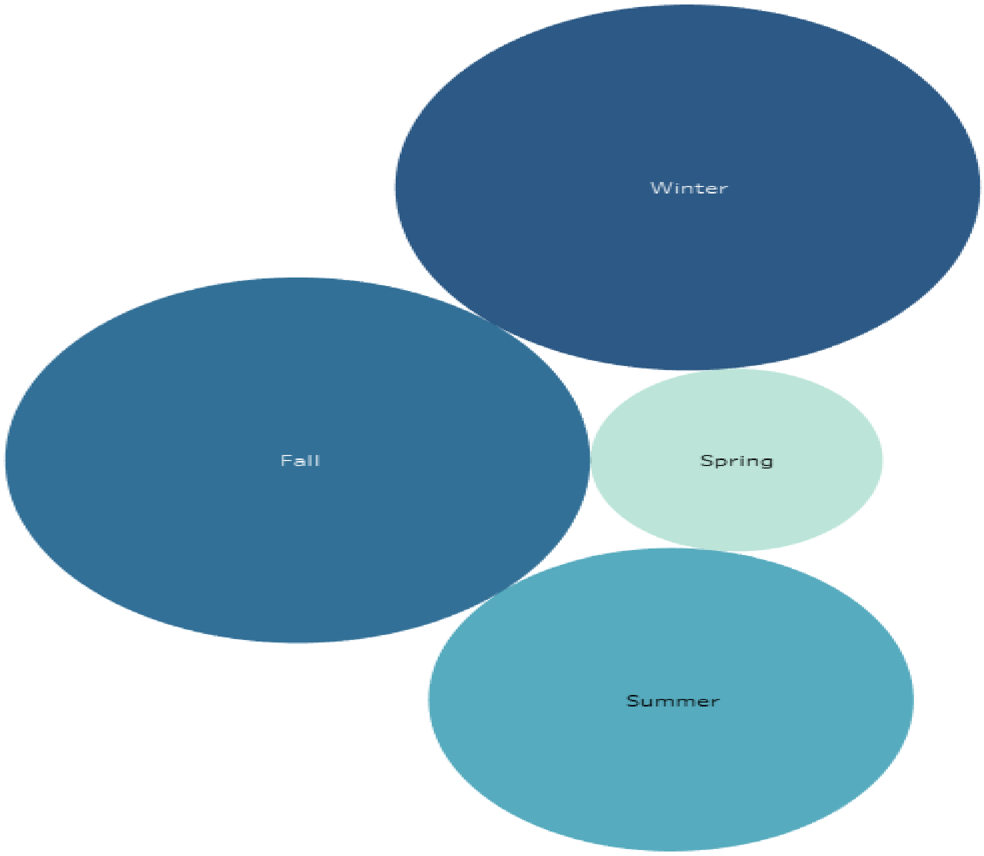

Figure 1.13

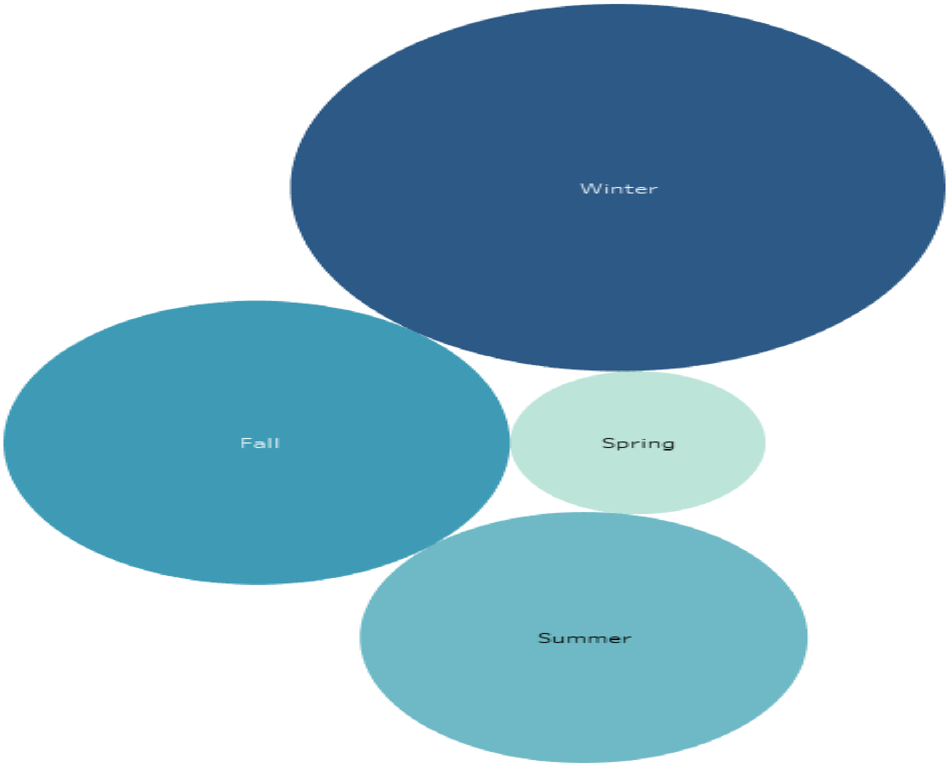

Figure 1.14

Our final visual showcases how aggressive teams play to win at home. To measure aggression of the home-team I used the amount of fouls called against the home team to represent defensive/physical aggression and total number of shots for the home team to represent offensive aggression. I also filtered the visual with respect to each individual referee, this was because referees have a tendency to call fouls differently (i.e. some referees are more lenient than others whereas others can be more strict). This is also filtered for each home team so I could look into how a cluster of teams fared or how an individual team fared. From this visual I gained the insight that the better clubs such as Chelsea, Arsenal, Manchester City, Manchester United, Liverpool, and Tottenham all win the most of the matches at home and get more shots at home but they also tend to foul the most at home games in which they win when compared to the number of fouls these teams commit when they draw or lose at home. This shows the better clubs play aggressively to win at home. The visual below is filtered to contain all of the referees and the clubs Chelsea, Arsenal, Manchester City, Manchester United, Liverpool, and Tottenham.

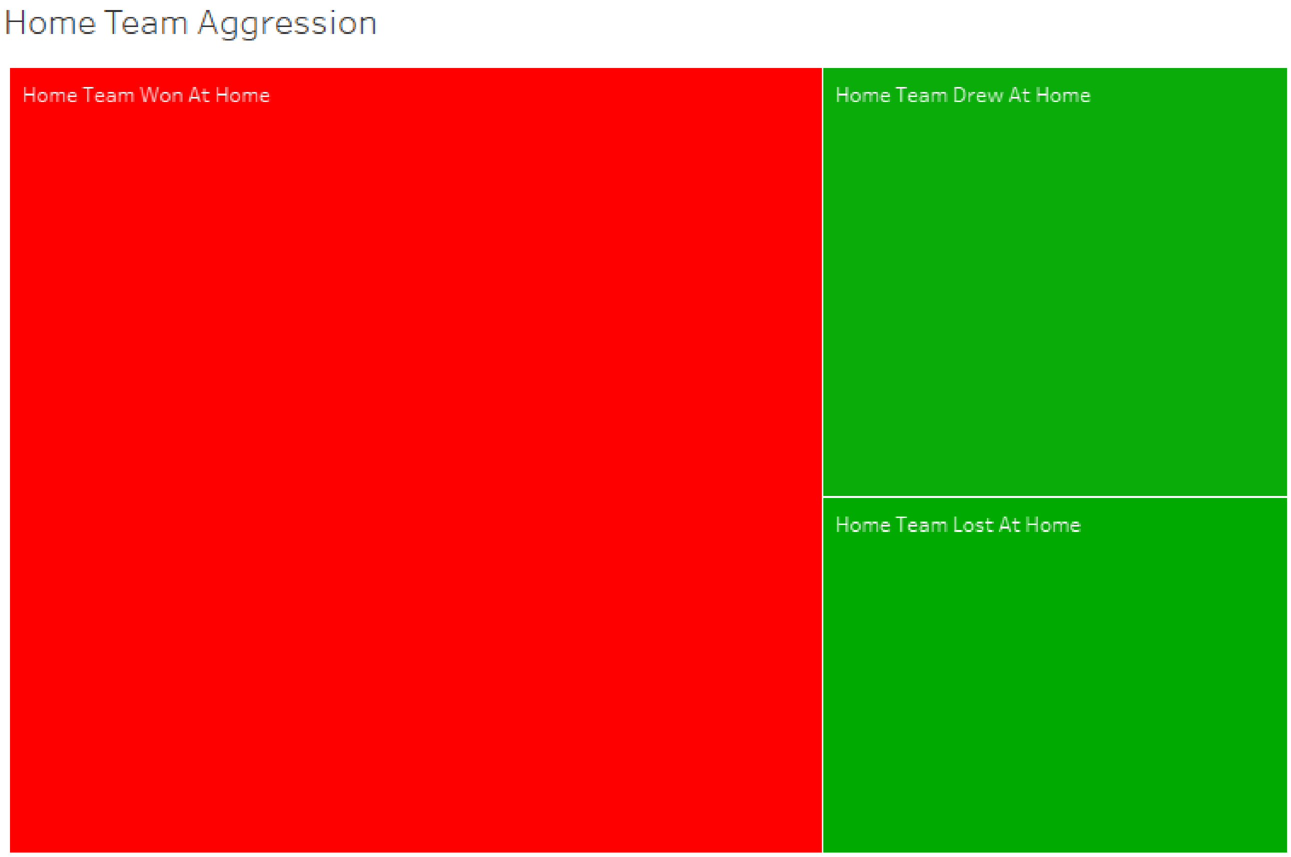

Figure 1.15

II. Models

With the insights from the visualization, I have created multiple models that consist of Logistic Regression, Decision Tree, K-Nearest Neighbors, Stochastic Gradient Descent, Naive Bayes, Neural Network (Multi Layer Perceptron), Ensemble Method, and Random Forest. Using the library of SK-Learn to build these models allowed for simple and efficient predictive data analysis. For this project the elected models to use were the Random Forest, Neural Network, Ensemble Method, Logistic Regression models since they performed the best out of all the models that were tested out. For the Random Forest model it was made to have ten estimators, a maximum of ten features, a maximum depth of five, and a minimum sample split of three. The Logistic Regression model was made using a OneVsRest multi-class classifier, a penalty of “L2”, and a tolerance of 1e-3. I wanted to use a OneVsRest classifier because it was less computationally expensive than doing a OnveVsOne classification for all possible scenarios as it required less classifiers to be made for this task. For the K-NN model, it was made to look at fifty neighbors and has its weights set to ‘distance ’. The Neural Network was constructed with the ‘adam’ solver, a ‘tanh’ activation function, a limit of 250 iterations, and a hidden layer size of (25, 15). The Ensemble Method model was made using all four of the previously stated models. After fitting all of the models, the models had their accuracy evaluated. Logistic Regression scored 61.842%, Random Forest scored 63.157%, K-Nearest Neighbors scored 58.552%,Neural Network scored 63.157%, and Ensemble Method scored 62.5%. The appropriate variables for these models are:

- Season Encoding - The four season, such as Fall, Winter, Spring, and Summer

- Time - Split into three categories; Early Day, Mid Day, and Late Day
- Home Team Encoding - ID for the home team
- Away Team Encoding - ID for the away team
- Referee Encoding - ID for the game’s referee
- YearOfSeason Encoding - The years of the season
- Fouls Called Per Game - Average fouls called per game by referee
- HS - Home team’s total shots
- AS - Away team’s total shots
- HST - Home Team Shots on Target
- AST - Away Team Shots on Target
- HC - Home Team Corners
- AC - Away Team Corners
- HY - Home Team Yellow Cards
- AY - Away Team Yellow Cards
- HR - Home Team Red Cards
- AR - Away Team Red Cards

To justify the choice of variables for the model is through the insights of the visualization. The understanding of shots and aggressiveness of the teams played a big part in the selection of the variables. It shows that teams that tend to be more aggressive win more games and understand this relationship improves the models.

Figure 2.1

In figure 2.1, this shows the normalized confusion matrix of the Random Forest. True positive prediction of home and away wins average 81.5% of accuracy. This helps the understanding that the model can accurately predict home win and away wins 4 out of 5 times, but the only limitation is that is fails to accurately predict draws as such in figure 2.2, the normalized confusion matrix of the decision tree, both the logistic regression and decision tree do not well when trying to predict draws.

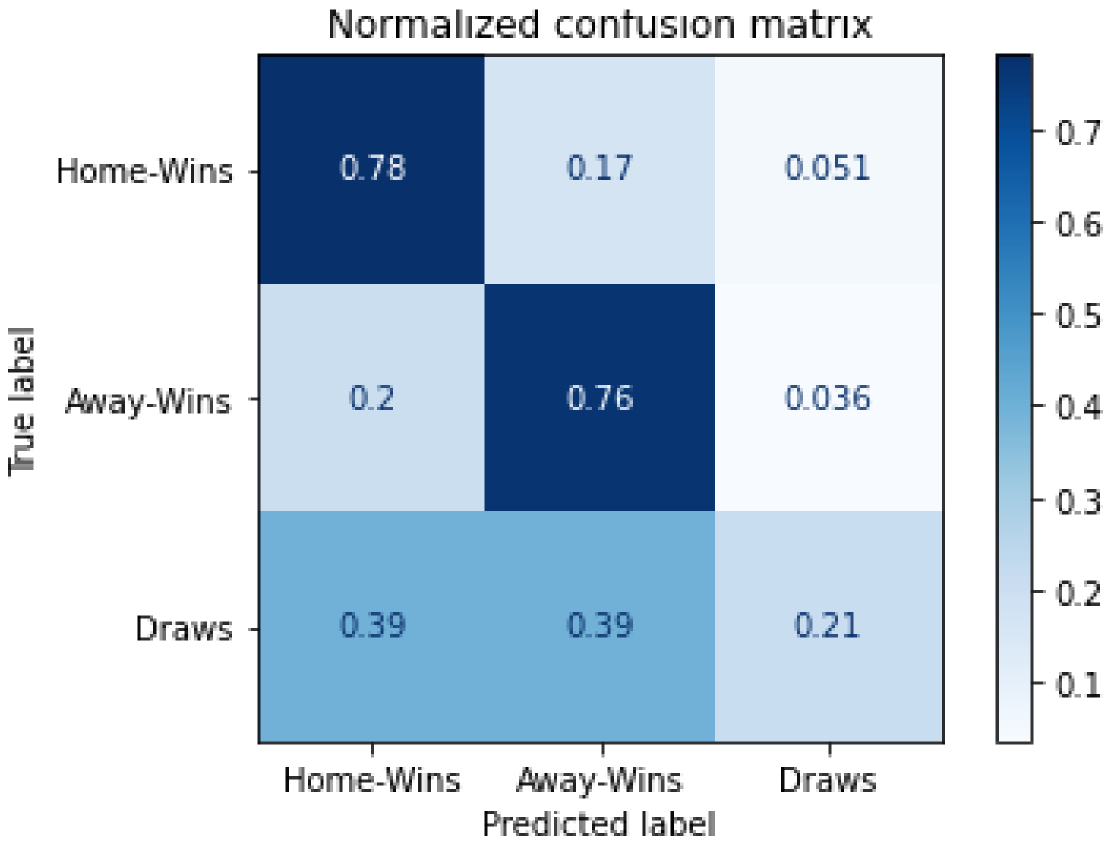

Figure 2.2

Other metrics I used to evaluate our models were the ROC-Curve. This also helped me get a graph to represent how well our model was able to predict the proper classes. When I compared the AUC values for both the Random Forest (fig 2.3) and Logistic

Regression (fig 2.4) models I can see that the Logistic Regression AUC value is less than the one for Random Forest. This played a role in selecting the Random Forest model as our best performing model. Another factor I considered when choosing a final model was the Precision, Recall, and F1-Score metrics, the Random Forest model had the favorable numbers, which can be seen in the table below.

	Precision	Recall	F1-Score
Random Forest	0.64214	0.57211	0.53166
Logistic Regression	0.61230	0.55769	0.51076

The final factor was the accuracy of each model, which Random Forest performed better in as discussed earlier. The one metric that favored the Logistic Regression model was the Adjusted R^2 value, the Random Forest model had a score of -0.96449 and the Logistic Regression model had a score o-0.85972.

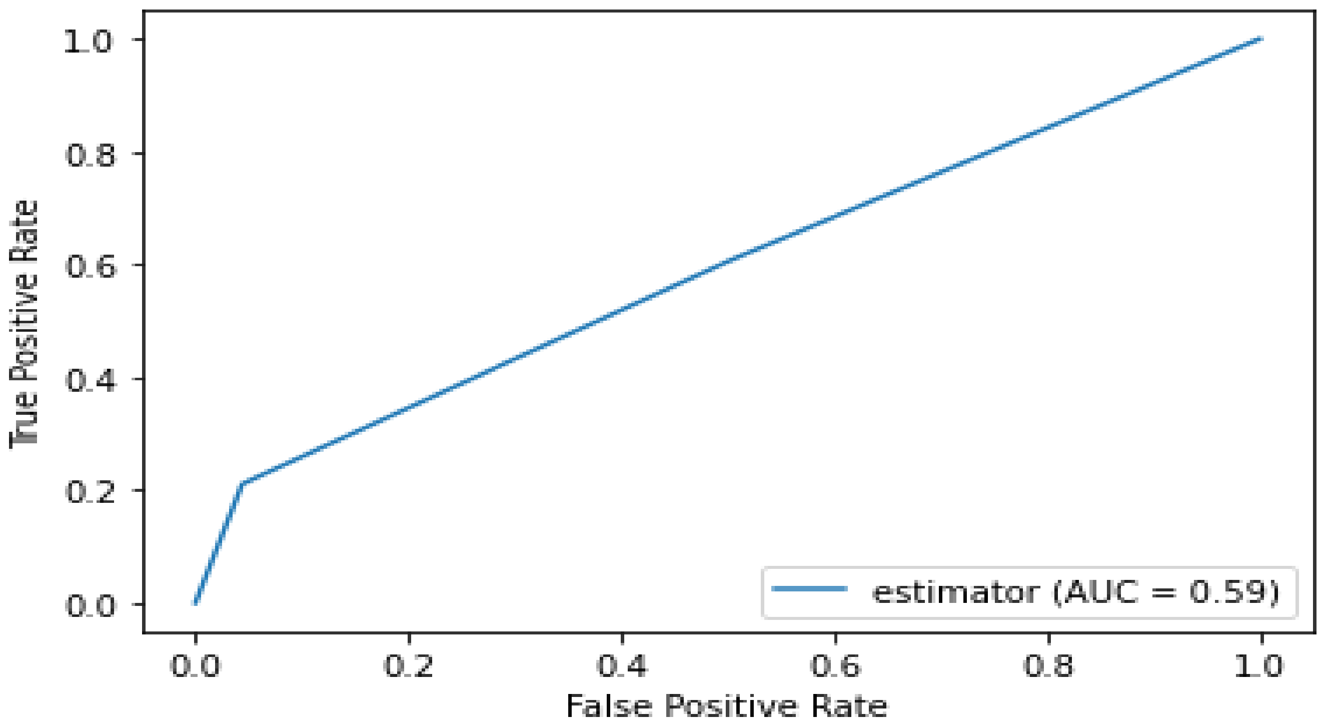

Figure 2.3

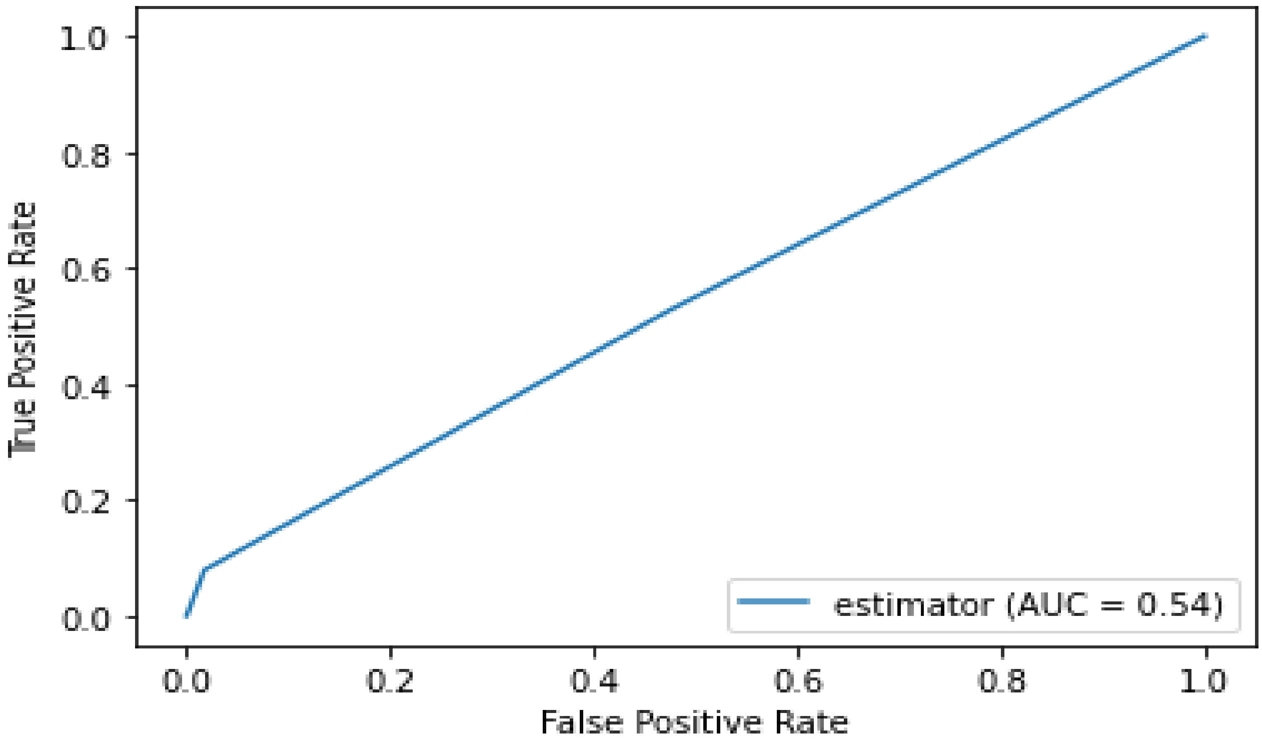

Figure 2.4

III. Insights and Recommendations

Insights that were recovered from the analysis explains that through the model it was able to predict wins with an accuracy of 81.5%. An outstanding number only to be dampened by the accuracy of predicting draws. Thorough analysis the recommendations to be issued are:

- Only wins are more reliant than when the model predicts draws.
- Further research to introduce more variables to help the accuracy of the model predictions (eg: variables about players, team's formations, and managers).
- Find a dataset that has more samples with a time variable, the models with the time variable performed better than the models without the time variable.