An Analysis of the 2018/2019 Football World Cups

Billy Hanan
PGD in Data Analytics
x18179797@student.ncirl.ie

December 6, 2019

Abstract

In this paper, StatsBomb event football data from the male and female World Cups of 2018 and 2019 respectively are analysed independently. The raw event data was processed to create statistics/features for each match and the relationships between these features investigated. Logistic models were also created to identify the principal features that determine the match outcome. In both tournaments, the amount of ball possession in the final third and the shots conceded were identified as key predictors. However, ball possession was found to be a far more significant feature in the female matches.

1 Introduction

Football is a leading world sport and the major professional leagues and tournaments attract large viewership figures from around the globe. With so much revenue at stake, top professional clubs now possess an analytics department to examine team performance metrics in an effort to gain an edge over the opposition. In-match data is typically bought from private data companies. One such company is StatsBomb who provide both data and services to clubs, media and gambling organisations [1]. To both develop and publicise the fast growing field of football analytics, StatsBomb makes some of its data publicly available. This free data includes details on all the matches of the 2018 and 2019 Fifa World Cups. In this project we search this data for correlations and attempt to identify match features that are strong predictors of a match outcome.

The data for a single match typically consists of over 3000 rows, each row recording a specific event that took place in the match. Here an event might be a pass, shot, tackle, free kick, etc. The time, location, player and other details related to the event are given in the record. The project tasks were broken down as follows:

- Aggregation of the event data to produce match statistics/features

- Data exploration of the generated match features
- Identification of the key features that are strong predictors of the match outcome

The first task above involved distilling the hundreds of events in a match down into a single record. This match record holds match statistics for each team such as possession percentages, total shots taken and goals scored. A data set was thus generated for all matches in the male 2018 World Cup and another data set generated for the female 2019 World Cup. These two data sets were subsequently investigated independently.

The main objectives of the project were then answer the following questions:

1. Is ball possession important to the match outcome?
2. What other features if any are predictive of the match outcome?

The paper proceeds as follows: In section II we give a detailed description of the raw StatsBomb data and provide details on the generation of the aggregated match records. In section III we briefly discuss the available analytic techniques that may be employed to meet our project objectives. Section IV gives details on the actual techniques employed. Finally, in section V and IV we present the results of our investigations and offer our conclusions respectively.

2 Data Sets

2.1 Raw StatsBomb Event Data

As mentioned in the introduction, we have two data sets in this project that were analysed separately. One involves a data set generated from the StatsBomb event data for the men's 2018 Fifa World Cup. The other data set was similarly generated from StatsBomb event data for the women's 2019 World Cup. For the 2018 World cup, we thus had data covering 64 matches with on average 3,560 event records for each match. Similarly for the 2019 World Cup, we had 52 matches with on average 3,394 event records per match. Below I give a very brief summary of some key fields in an event record. For the full specification, see [2].

The event data comes in Json format and some event types have nested objects containing other fields associated with the event e.g. Outcome of event, end location of a pass etc. There was no need to directly download Json files for this project however. StatsBomb provide a bespoke R package [3] that allows a user to download the data in-memory for any specified competition from their repositories. The package also allows users to more easily extract and manipulate the event data. All the data processing in this project was therefore performed and recorded in an R Jupyter notebook.

Field Name	Description	Sample Values
timestamp	Time (ms) in match of event	00:12:06.293
type	Event type	Dribble, Shot, Pass, Clearance
team	Team associated with event	Brazil
player	Player associated with event	Roberto Firmino
location	Coordinates of event	(60,40)
under_pressure	Action performed while being pressured	True
out	Ball leaves field of play	True,False

2.2 Aggregated Match Data

The above event data is far too rich and detailed for the purposes of this project. Therefore, as previously mentioned, this raw event data was processed to generate statistics/features for each match in our two World Cup data sets. Section IV on methodology might typically be the section where one would usually discuss the details of these aggregated match records, but as these records are informative for the section immediately following, we shall do this here.

Code was written in R that took all the event records for an individual match and aggregated them to produce a single match record. The fields in each such match record are as follows:

- *match_id, match_date*: Unique identifier and date of match
- *team.name* Name of football team
- *Score, Shots, ShotsOnTarget*: Number of goals scored, shots taken and shots on target
- *AvgShotDistance*: Mean distance of shots from the opposition goal
- *XG*: Total expected goals
- *Possession, PctPossession*: Total ball (as measured by number of passes) possession and percentage possession
- *avg.pass.length, avg.pass.length.z1, avg.pass.length.z2, avg.pass.length.z3*: Mean length of passes on entire pitch and in zones 1, 2 and 3
- *length.of.resultant.pass, length.of.resultant.pass.z1, length.of.resultant.pass.z2, length.of.resultant.pass.z3*: Mean length of resultant passes on entire pitch and in zones 1, 2 and 3
- *dirn.of.resultant.pass, dirn.of.resultant.pass.z1, direction.of.resultant.pass.z2, direction.of.resultant.pass.z3*: Mean Direction of resultant passes on entire pitch and in zones 1, 2 and 3
- *z1.passes, z2.passes, z3.passes, z1.passes.pct, z2.passes.pct, z3.passes.pct*: Total number and percentage of passes in zones 1, 2 and 3

- *Result*: Either W (win), L (Lose) or D (Draw)

Note that all the above information relates to the team referenced in the field *team.name*. Similarly, there are a further 27 fields appended to the same match record that give the same information for the opposition team. These fields are prepended with the prefix "Opp_". So for example, we also have fields labelled *Opp_team.name, Opp_Score, Opp_Shots,* etc. In total then, each match record has 56 fields containing information on both teams involved in the match.

As you can see above, the match features that were calculated focus primarily on shot and pass statistics. Note that the football pitch is divided into three equally sized zones with zone 1, 2 and 3 being the defensive, midfield and attacking areas of the pitch. Also note we have references to the "resultant pass" above. To calculate the resultant pass in, for example zone 3, we simply calculate the vector sum of all the teams passes in that zone. The direction of a pass is measured in degrees with zero degrees being in the direction of the opponents goal. Rotating from this direction clockwise (anti-clockwise), the angle takes on positive (negative) values.

Finally, note that *expected goals* is a metric that has recently been popularised in the area of football analytics. The expected goal metric is simply the probability of scoring when a shot is taken at goal. So, for example, the expected goal metric for scoring a penalty, which is determined empirically, is 0.76. Essentially it is a measure of the quality of the goal scoring chance. The field *XG* above is the sum of this metric over the course of the match and thus measures the total quality of all a team's goal attempts.

In summary then, we have two data sets to analyse. We have 64 match records from the 2018 Fifa World Cup and 52 match records for the 2019 World Cup. In the next section, we shall discuss some techniques that might potentially be applied to this data to meet our project goals.

3 Applicable Techniques

Our project goals were outlined in the introduction. To restate them again, they are to use the generated match records to uncover correlations among the match features, investigate the impact of ball possession on the match result and to also identify specific features that are strong predictors of the result. In the following three subsections I briefly discuss the analytical tools that one might use.

3.1 Linear Regression

The Pearson correlation coefficient r can be used to measure the strength of the relationship between two features. When one calculates the correlation, one is effectively performing a linear regression. This technique involves fitting a line to two variables by typically minimising the sum of the squares of the residuals. One typically verifies that the obtained coefficient value(s) is statistically

significant by calculating its p-value. This p-value gives the probability of obtaining the coefficient value by random chance under the null hypothesis that the coefficient is zero.

There are issues with this technique however. Namely, the coefficients and p-values are only accurate if some assumptions hold. For example, the residuals of the fit must be homoskedastic and independent. Most importantly, the relationship between the two variables must be linear. Therefore, if we have two variables that are strongly non-linearly correlated, linear regression (and the correlation coefficient) will not detect this relationship. It is therefore important to perform a visual check on the relationship between the variables, rather than blindly trusting the value of r obtained.

3.2 Classification Techniques

In this project, we attempt to identify the features which are strong predictors of the match result. We do this by performing binary classification. That is, by looking at the match features, we attempt to predict whether the home team wins the match or not. There are numerous classification algorithms in the area of machine learning.

One of the most basic is the k-nearest neighbour algorithm. In this non-parametric method, given some sample feature vector, we classify it by typically choosing the majority class of the k-nearest feature vectors in your data set. One typically requires a large data set for it to be effective however, but if too large, the method can be computationally expensive due to the volume of distance calculations required.

Logistic regression is another classification approach that is a closely related to linear regression. Like linear regression, it involves fitting a curve to the available data to obtain coefficients with p-values. The key difference between the two techniques, other than the form of curve fitted, is that the log likelihood (rather than the sum of residuals squared) is minimised. Like linear regression, it can be applied to small datasets but can lead to erroneous results if the independent variables are strongly correlated (a condition known as multicollinearity).

One might also apply an ensemble method such as random forests for classification purposes. It is a technique that is very robust to overfitting. One can also easily identify the features with the most predictive power. The same cannot be said for support vector machines or neural networks. The former involves transforming the data set using kernel functions and then fitting a hyperplane to separate the classes. In the latter approach, the connection weights in a network of non-linear units are tuned in an iterative fashion to minimise some output error function. Both these approaches suffer from an inherent lack of interpretability. That is, they are essentially black boxes used to churn out predictions with little in the way of explanation.

4 Methodology

The analytical techniques that are suitable for this project are constrained by:

1. the relatively small size of the data sets (64 and 52 match records for the 2018 and 2019 World Cup respectively)

2. the need for the models to be easily interpretable

These constraints effectively rule out nearly all of the methods described in the previous section. As such, in this project we implemented both linear and logistic regression. Note that a significance level of 0.05 was used in all the hypothesis testing performed in this project.

Whereas all the data processing was performed using R, the majority of the analysis was done using Python in a Jupyter notebook. The analysis began with some initial data exploration. For example, any relationships between the match features were investigated by calculating the correlation matrix. A visual check was also performed using the *PerformanceAnalytics* package to check if any obvious non-linear correlations were present (None were found).

In the next section we show some of these plots to demonstrate the most interesting relationships discovered. In some cases, simple linear regressions are performed to measure the statistical significance of these relationships. We pay particular attention to ball possession and its relationship with the other variables. But other correlations are also explored.

Finally, we conclude our analysis by building a logistic regression model for each data set. Here, a target variable y is appended to each match record. A value of $y = 1$ is assigned to the record if the home team (given by the field *team.name*) wins the match. Otherwise the value $y = 0$ is assigned. A logistic model was then fitted using y as our dependent variable and all available match features as our potential independent variables.

The building of the logistic model was done manually by performing forward selection. Briefly, this involved starting with a model with no independent variables ($n = 0$). Multiple logistic models were then fitted using each feature as the independent variable (That is, we have one model of size $n = 1$ for each feature). The feature yielding the lowest p-value was selected for inclusion into our model. We now have a logistic model with $n = 1$ features. The next feature to be added to the model ws selected in a similar fashion. That is, multiple logistic models with $n = 2$ were fitted by selecting each of the remaining features in turn and again the feature with the lowest p-value is chosen. This iterative procedure is repeated until the p-value exceeds the preset criterion $p < 0.05$, at which point we have built our final logistic model.

Note that at each iteration of this procedure, correlations between the features being added to the model were checked. If the correlation exceeded a value of 0.5, the feature was not included in the model. This was to avoid any issues with multicollinearity. Let us proceed to the results and discussion section.

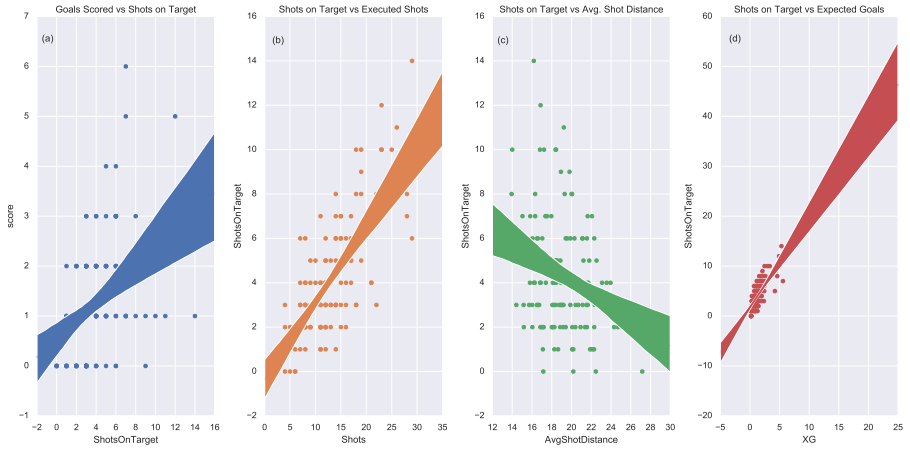

Figure 1: Goal Metric Relationships

5 Results and Discussion

We shall discuss the results of the two datasets separately. In the first subsection we will present the results obtained from the 64 matches in the men's 2018 World Cup before proceeding onto the 52 matches in the 2019 women's World Cup.

5.1 World Cup 2018

From an attacking perspective, the objective is to score as many goals as possible. Therefore we begin by examining the shooting statistics in a match and its relationship to goals scored. The results shown in figure 1 are not surprising. In fig. 1(a), while there is a positive correlation between the number of goals scored and shots on target, it is not a clean linear relationship with a very low R^2. The remaining plots in fig. 1 simply illustrate that to maximise your number of shots on target, and thereby the number of goals scored, you need to take more shots as close to the opponents goal as possible. Fig 1(d) simply demonstrates that the better the quality of chance, the more likely you will hit the target.

In fig. 2, we examine the relationship of percentage ball possession on various other match features. This metric simply measures the fraction of the match that a team is in control of the ball. In football parlance, the team with the greater percentage of ball possession is said to be the dominant team. The interesting aspect of this figure is that as we move from left to right, we see the steady degradation in the statistical significance of the linear coefficient. While in fig. 2(a) we see that clearly having greater ball possession yields more shots at the opposition goal, this greater possession does not necessarily translate into more actual goals scored as shown in fig. 2(d). In fact the p-value for the fitted linear coefficient for fig. 2(d) is 0.607. This implies that with the available data

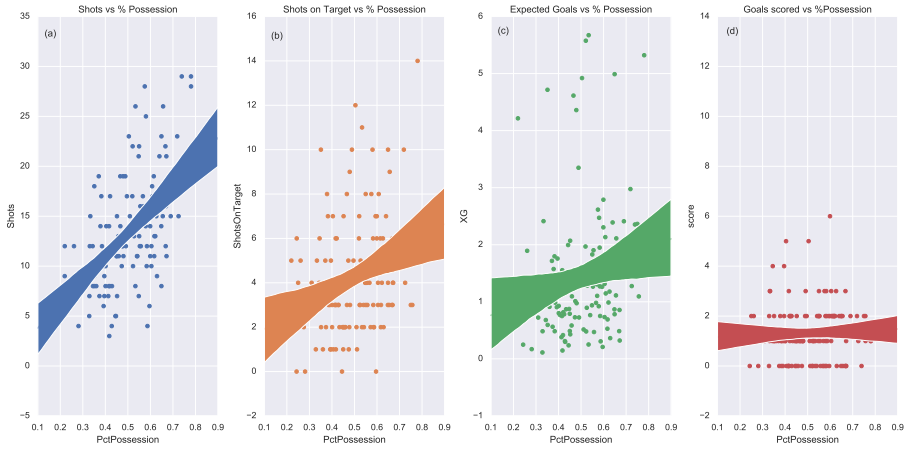

Figure 2: WC2018 Ball Possession

we cannot reject the null hypothesis that ball possession has no influence on goals scored.

This conclusion is further vindicated by performing a simple t-test on two groups of teams. Here we simply look at each of our 64 matches and assign the team with the greater percentage ball possession to group 2 and the opponent to group 1. Group 2 therefore represents the dominant teams in their matches. We then calculate the average number of goals scored by the teams in each group. If ball possession has an impact on goals scored, then we should see a significant difference between the two values. Fig. 3(a) shows a simple box plot distribution of the number of goals scored in group 1 and 2. The red dots denote the mean values of 1.25 and 1.39 for group 1 and group 2 respectively. Though the mean value for group 2 is marginally higher, a simple t-test yields a p-value of 0.49 so that the difference is not statistically significant. Again, we cannot say that greater ball possession leads to a greater number of goals scored.

In fig. 4, we illustrate a simple concept. There is a clear negative correlation between percentage possession and the average pass length. The implication is simple. If a team wishes to dominate the ball, the coach has to implement a shorter passing game.

On a more abstract level, football is about controlling and exploiting space on the pitch. In Fig. 5 we examine this theme by showing the distribution of the average pass length and the mean direction of the passes in the three (defence, midfield and attack) zones of the pitch. Note that the average pass length decreases as one moves up the pitch. This likely implies that passing lanes become congested and space becomes rarer as one approaches the opponents goal. A similar phenomenon is shown in fig 5(b) where we see that the distribution of the mean passing direction actually broadens as the opponents goal is approached. Again as space is at a premium near the goal, the passes

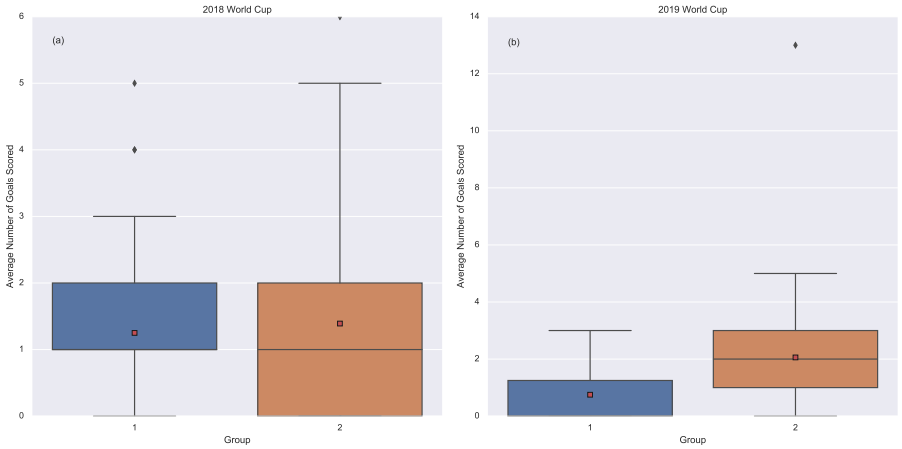

Figure 3: Boxplots of Average Number of Goals per match

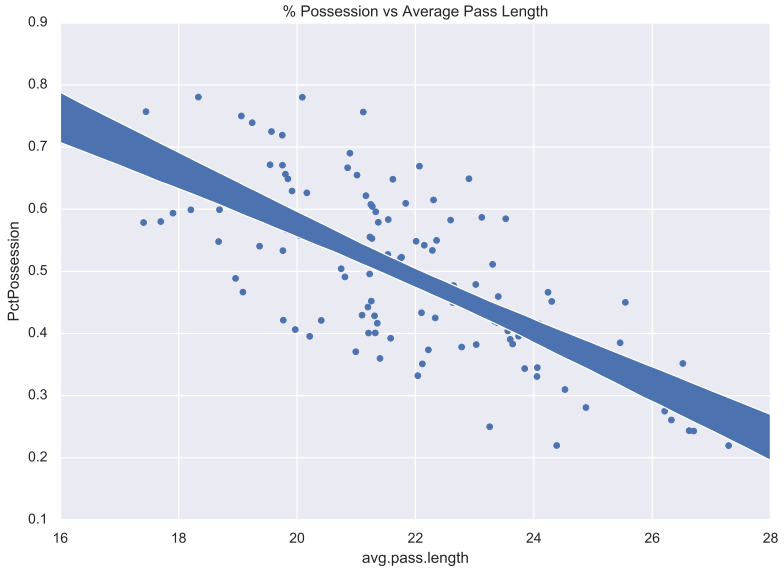

Figure 4: Ball Possession vs. Mean Pass Length

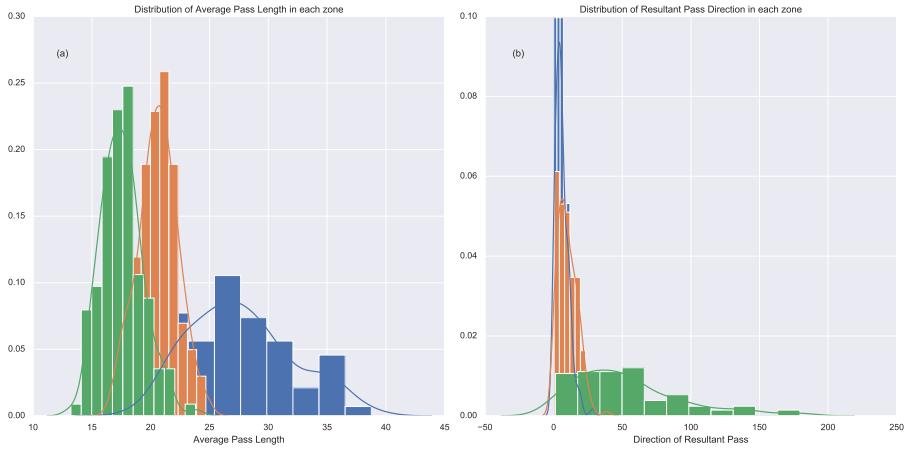

Figure 5: Compression of Space on Pitch

become less vertical and more lateral. That is, the ball is passed sideways more often and even passed backward to be recycled.

In the table below we display the results of the logistic model that was built to identify the most predictive features in determining the match outcome. Recall that the model simply has to predict whether the home team wins or not. The three features *Opp_ShotsOnTarget, Opp_z3.pass.pct* and *length.of.resultant.pass.z3* are all statistically significant and the standardised coefficients (not shown in the table) are -1.1464, 0.7983 and -0.6435 respectively. As such, the most significant feature is the number of shots on target that you allow the opposition. The more shots on target, the smaller your chance of victory. Bizarrely though, the results for *Opp_z3.pass.pct* suggest that as the fraction of time that your opponents spend in your defensive zone with the ball increases, the greater your chance of victory. Finally the feature with the least impact of the three is *length.of.resultant.pass.z3*. Therefore the smaller the length of the resultant pass in the your attacking zone, again the greater your chance of a win. We will discuss these results again in the conclusion later.

Feature Name	**Coefficient Value**	**p-value**
Opp_ShotsOnTarget	-0.6368	0.0011
Opp_z3.pass.pct	12.8069	0.0016
length.of.resultant.pass.z3	-0.4528	0.0373

Note that using a threshold probability value of 0.5, we also found that the in-sample prediction accuracy of this logistic model to be 78.125%. Note that of the 64 World Cup matches, there were 26 victories for the home team. As such, a naive benchmark model where we simply always predict a loss for the home team yields a baseline model accuracy of $(64 - 26)/64 = 59.375\%$. As

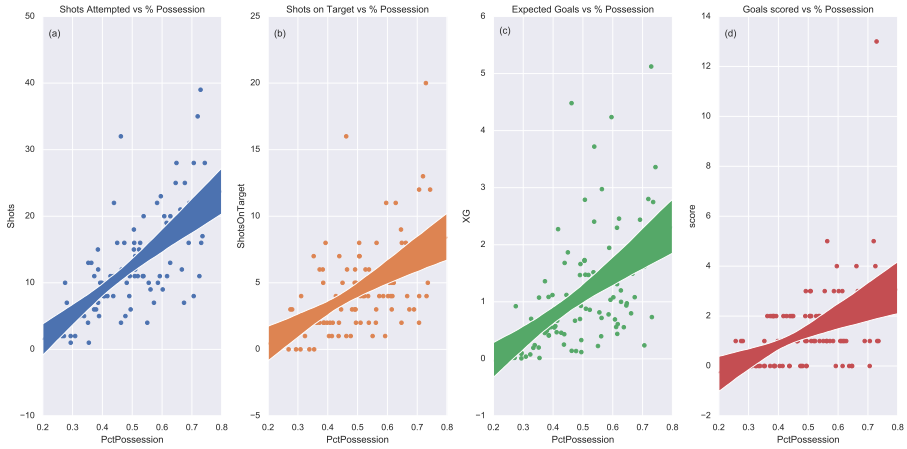

Figure 6: WC2019 Ball Possession

such, the model accuracy significantly beats this benchmark.

5.2 World Cup 2019

The women's 2019 World Cup tournament won by the USA took place over 52 matches. An identical analysis was performed on this dataset of 52 match records as was done with the 2018 World Cup in the previous section.

Performing some initial data exploration on goals scored and shots, we found the same behaviour as that exhibited in Fig. 1. As such, the equivalent plots for the 2019 World Cup are not shown here.

In Fig. 6 we illustrate the relationship between percentage possession and some shot related features. The equivalent diagram from the 2018 World Cup is shown in Fig. 2. Note that there is one marked difference. In Fig. 2, the linear relationship in the figures became weaker as we moved from Fig 2(a) to 2(d). In fact in Fig. 2(d) we concluded that there was no evidence of a relationship between the number of goals scored and percentage possession. This is not the case however in Fig 6(d). Here we find that the linear coefficient has a value of 5.55 and is statistically significant from zero with a p-value of 0.000002. As such we conclude that in the 2019 World Cup, and unlike in the men's 2018 World Cup, ball possession was a significant factor in determining the number of goals scored.

This was corroborated when performing a t-test in the same manner as described in the previous subsection. Here the mean goals scored by teams in the dominant group2 was 2.06 while group1 scored a significantly lower average of 0.75 goals in a match. This is a significant difference and the t-test yielded a p-value of $3.6e-05$. The box plots of the goals scored in each group is shown in Fig. 3(b).

Finally we built a logistic regression model using the 52 sample points in

our dataset. Again, the objective was to build a model to predict whether the home team would win the match or not. The logistic model was again built by manually applying a forward feature selection method while checking for multicollinearity. The coefficients of the fitted regression model are shown in the table below. There are in fact only two statistically significant features in this model, *Opp_ShotsOnTarget* and *z3.passes* which have standardised coefficients (not shown in table) of -0.6091 and 0.0233 respectively. Again we find that as you concede shots on target, your chances of victory are significantly reduced. Though far weaker in effect, your chances of a win increase by maximising the time spent with ball possession in your attacking third.

Feature Name	**Coefficient Value**	**p-value**
Opp_ShotsOnTarget	-0.6091	0.0008
z3.passes	0.0233	0.0010

6 Conclusion

In this final section we will give a concise overview and interpretation of the results in the previous section. Some of the results are not particularly noteworthy and simply provide empirical evidence of what one would intuitively expect. As such we will only comment here on any stand out issues.

As illustrated in Figures 2 and 6, our first major result is that there does appear to be a marked difference in the impact of ball possession on the goals scored in the 2018 and 2019 World Cups. In the former, there was no evidence that holding onto the ball for long periods had any impact on the number of goals scored. In the latter however, ball possession was positively correlated with goals scored. Furthermore, the importance of ball possession in the female World Cup was further corroborated by the presence of the feature *z3.passes* in the logistic regression model.

In fact, the results of the logistic regression model for the 2019 World Cup are simple to interpret. The model features include *Opp_ShotsOnTarget* and *z3.passes*. The latter feature stresses the importance of ball possession to victory while the former feature emphasises the importance of a tight defence that does not allow the opponents to get shots on target.

The logistic regression model built from the men's 2018 World Cup is much more difficult to interpret. Here *Opp_ShotsOnTarget, Opp_z3.pass.pct* and *length.of.resultant.pass.z3* are the key predictors of victory. The most significant is again *Opp_ShotsOnTarget* which has the same simple interpretation as in the 2019 model. However the other two features are particularly tricky to interpret. The feature *Opp_z3.pass.pct* seems to suggest that counter-intuitively, one should allow the opposition to have the ball in your defensive third. That is, it suggests you should defend deep and concede possession to the opposition. The presence of the factor *length.of.resultant.pass.z3* in the model is even more difficult to interpret. In

fact, I will refrain from doing so as it warrants further investigation.

As a final point, as evidenced by the fact that *Opp_ShotsOnTarget* is the most predictive feature in both the 2018 and 2019 logistic models, if there is one clear message to take away from this study, it is that having a tight defence is the most important aspect to implement for a winning football team.

References

[1] https://statsbomb.com/

[2] Full data specification document is available for download from https://github.com/statsbomb/open-data/blob/master/doc/StatsBomb

[3] The R package can be installed by executing *devtools::install_github—("statsbomb/StatsBombR")*