Statistical analysis of the 2023 FIFA Women's World Cup

INSA Toulouse

Date : May, 21st 2024

Authors :
Canouet Eugénie
Laurié Romain
Richaume Julien

Tutors :
Dejean Sébastien
Saint Pierre Phillipe

Contents

1 Abstract

Football is a multi-billion dollar industry, but many questions remain about the intricate predictors of a given team's success. Many have explored the impact of possession such as Collet who studied the impact of possession in 2013. We examined the success of national teams that took part in the 2023 FIFA Women's World Cup, using public data available on StatsBomb. Firstly, we used descriptive statistics to understand the ways teams sustain advantages on other teams. We examined passing, shooting, and advanced metrics such as expected goals (xG) and expected passes (xP). We tried to distinguish tendencies and differences in styles of play between teams, and individual players. We managed to develop our own xG model, therefore finding the most relevant criteria to predict goals, in order to compare it to StatsBomb's model, and build visualizations based on it. We also plotted player's and team's pass completion rate, as well as their goal to shot ratio. Using our model, we managed to explain the success of designated teams, and the failures of others.

2 Introduction

In today's world, sports are at the center of global culture. In order to excel at the best level, players and teams must find solutions, both physical and tactical. Therefore, statistics will play a crucial role in optimizing performance. Previous studies have shed light on various aspects of football analytics. Collet studied the impact of possession in 2013. More recently Liu analyzed the environmental impact in 2021. However, the realm of football remains relatively unexplored in terms of data analysis. Understanding the dynamics of offensive and defensive play is pivotal for teams aiming to excel in competitions.

The research gap lies in the need for a comprehensive analysis of football performance using advanced statistical methods, with a focus on data from platforms like StatsBomb. The impact of certain specific aspects of football analytics, such as shot analysis or passing patterns remains unclear, and a comprehensive understanding of player and team performance is still lacking.

We aimed to address this gap by conducting a detailed analysis of football performance using StatsBomb data. We seeked to identify key performance indicators, assess their impact on match outcomes, and uncover underlying trends and patterns in player and team performance. This report outlines the methodology used for data collection and analysis, presents the findings from the study, and discusses their implications for the future of football analytics.

This report is divided into three parts. In the first section, we conduct an exploratory data analysis to identify certain trends, notably by analyzing shots and goals for each team. Then we seek an optimal statistical model to determine which parameters have the greatest impact on player performance. The last section contains the results of our analysis, including insights into player and team performance derived from StatsBomb data, with graphs examining successful shots and passes.

3 Descriptive data analysis

The package StatsBombR provides the data from 71 national and international competitions from over 3000 matches. For the sake of this project, we narrow our scope down to the most recent competition available : The FIFA Women's World Cup 2023 and its 64 matches. We begin by interpreting the different variables of this large dataset.

The full dataset contains 183 variables for analysis, with the majority having a significant proportion of missing values, as they were used to track very specific patterns of play. As an example, the parameter "goalkeeper.shot_saved_to_post" is attributed "True" only if the goalkeeper saved a shot from going inside the goal, by deflecting it onto a post.

3.1 Analysis of successful shots according to country

First, we look at how the data is stored in the dataset. We do this by plotting the shots of a specific match; we chose Spain-England, which was the final match of the competition.

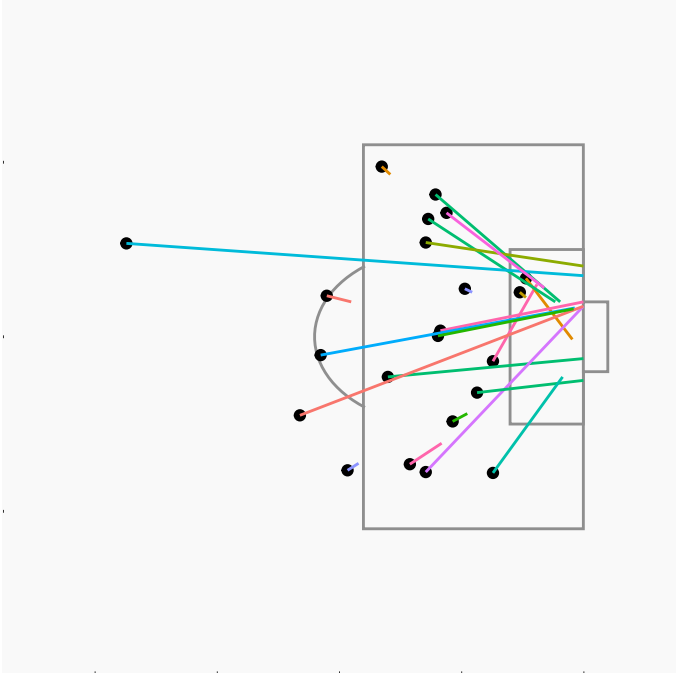


Figure 1: Visualization of the shots of Spain-England on the pitch

This graph shows that regardless of which team scores, the locations of the players are always tracked the same way : for a given team, the home goal is located at x = 0, which corresponds to the left of this graph, and the opposing goal is located at x = 120. This will allow us to directly use the provided variables, without further formatting the data.

Then, we took a look at the number of goals and shots in all matches for each team.
Figure 2 shows a visualization of these results.

From the above figure, we can already see a big disparity in team success. Some teams, like Vietnam and Haiti, didn't even manage to score a goal over the course of the competition, while Sweden and France were

more successful in this aspect. However, simply displaying how much shots and goals a team made provides an incomplete understanding of team effectiveness, and is generally a flawed metric for comparison, as teams that made it further into the competition naturally scored more goals, and had more shots.

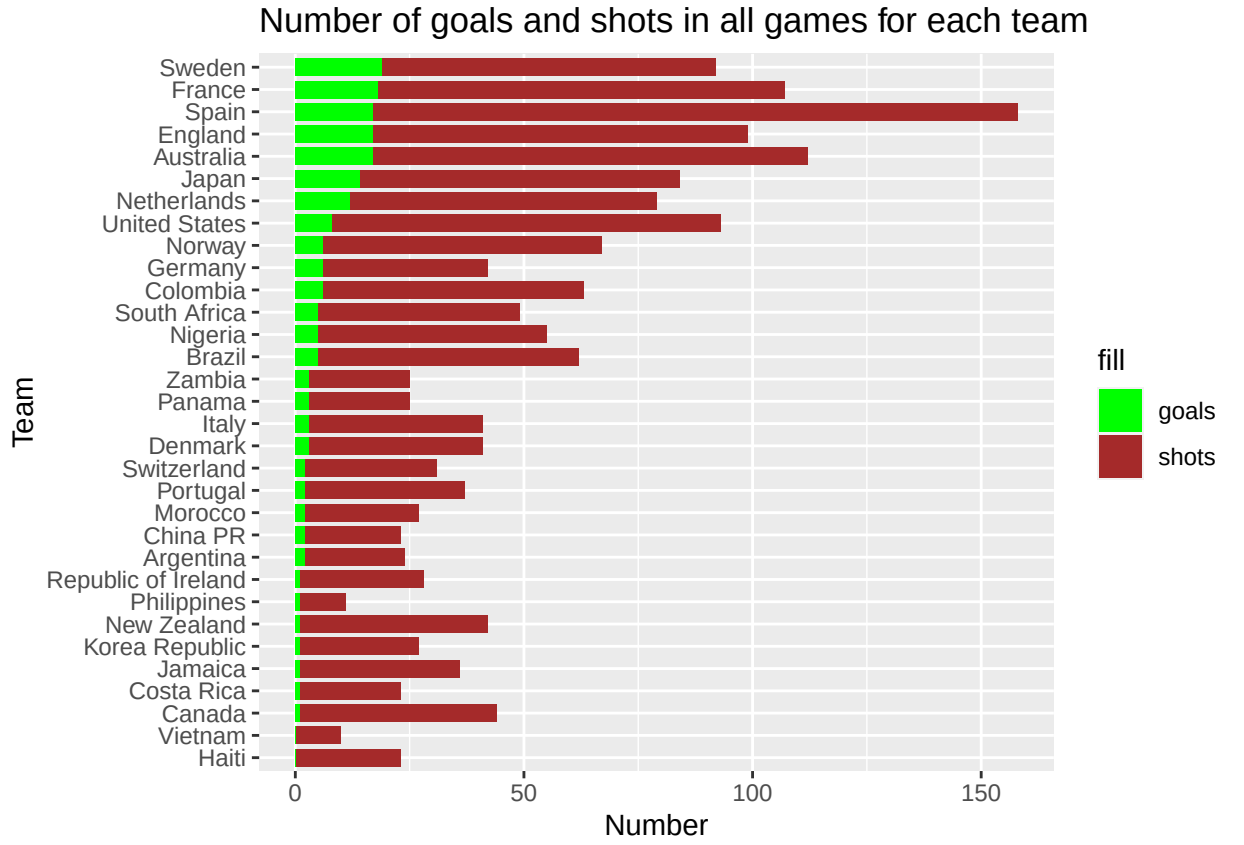


Figure 2: Diagram of the number of goals and shots in all matches for each team

Therefore, we calculated the percentage of shots leading to a goal for every team, and compared them in Figure 3.

From this graph, we are able to identify that Sweden was the most efficient team in scoring by a substantial margin. This is of course very biased, as Sweden finished third in the tournament, only losing one match throughout the entirety of the competition. Keeping that in mind, it is surprising that Spain scores this low on the graph, considering they won the World Cup. It could be that Spain, despite their evident success, was not a very efficient team, or that they had a different playing style than other teams.

Next, we realized the above analyses for singular matches. We chose the four matches played by team France, and created the same graphs for each of their matches.

Figure 4 shows that there were a huge number of goals in two games : Panama-France and Australia-France. The final score of the first match was 3-6 for France, while the latter ended on a draw. However, teams went to penalties, and they scored a total of 13, making this observation flawed.

3.2 Analysis of successful shots according to different variables

We first look at the different types of shots in Figure 5. Four different types of shots were differentiated in the dataset : “Open Play”, “Penalty”, “Free Kick” as well as “Corner.” A shot was deemed as being “Open Play” if it was taken during regular actions of the game. The “Corner” label only applies to a single shot

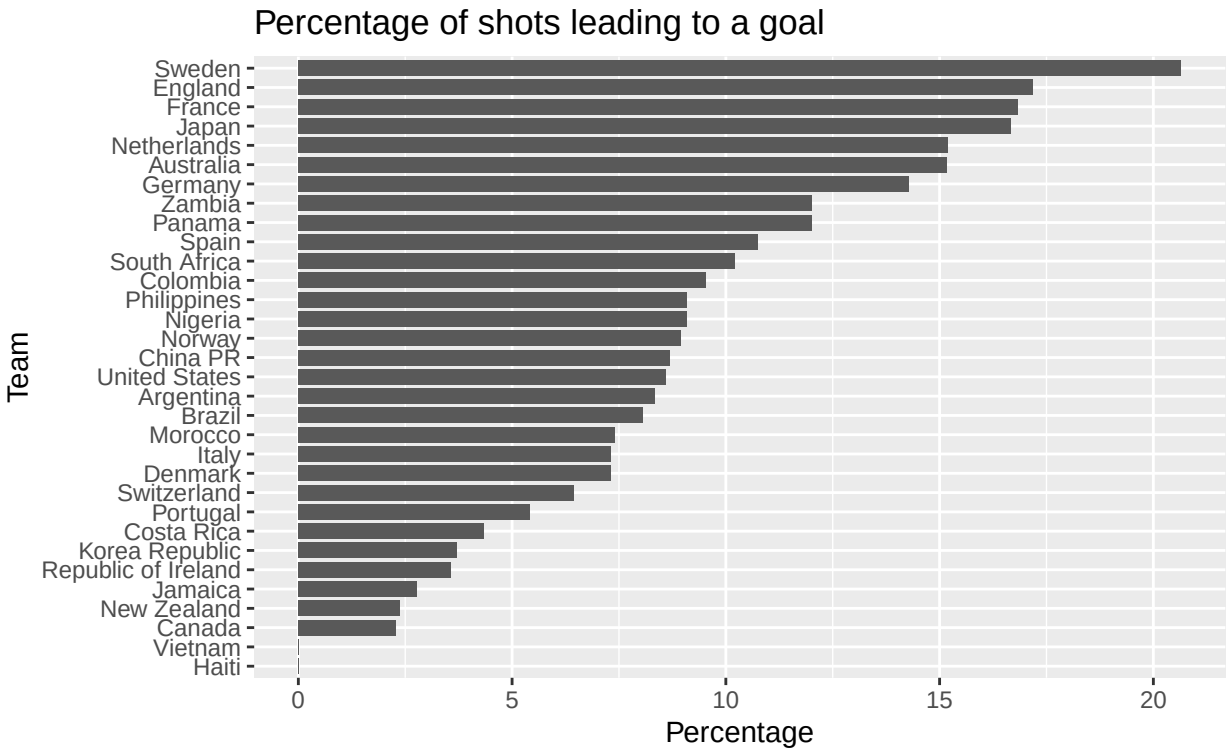


Figure 3: Diagram of the percentage of shots leading to a goal

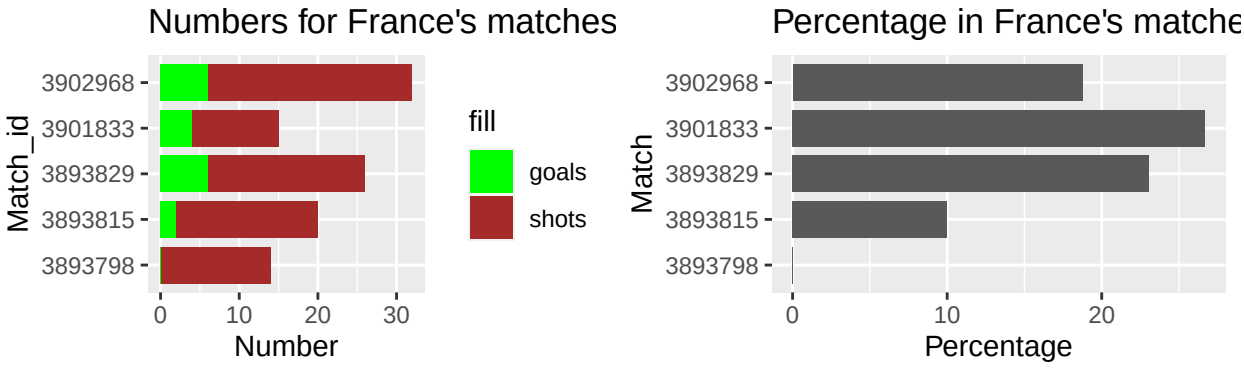


Figure 4: Diagram of the number of shots and goals for each French and percentage

made by Ireland's Katie McCabe that went in. This is something to keep in mind, as it is sure to skew our future models. Other labels are self-explanatory.

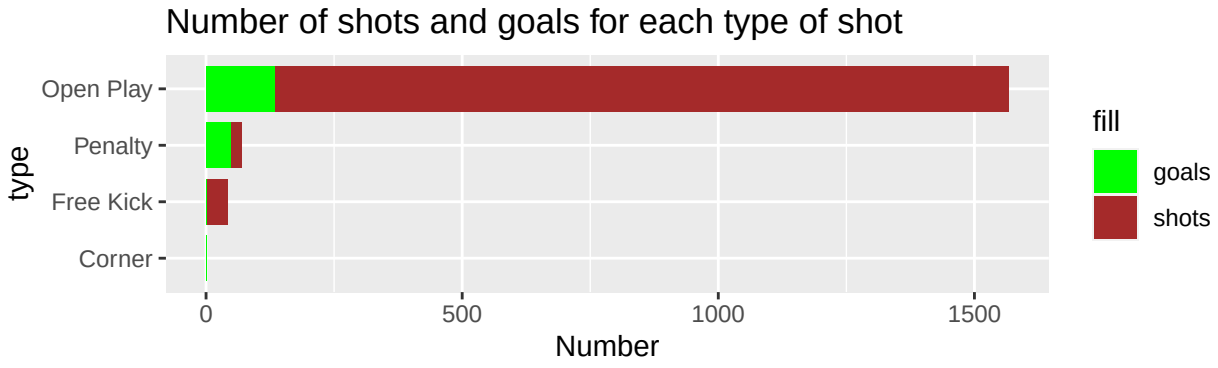


Figure 5: Diagram of the number of shots and goals for each type of shot

Next, we look at the different techniques used by players : how much are being kept a track of in the dataset, how much were each of them used, and which one produced the most goals.

Figure 6 shows that the dataset contains seven types of shots, although only three are consistently being used, that is : "Normal", "Half Volley" and "Volley". Naturally, the "Normal" shot was the most popular, and hence yielded the mos t goals. A "Lob" was very infrequently done, but could prove effective in the right situation : it seems a good proportion of these shots went in.

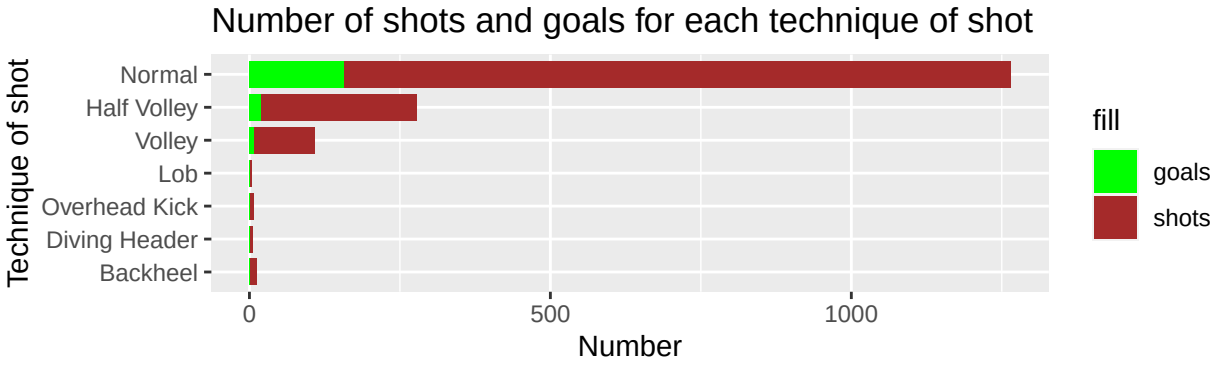


Figure 6: Diagram of the number of shots and goals for each type of shot

Similarly, we visualize in Figure 7 the different body parts used in shooting.

Unsurprisingly, the right foot was most commonly used, and it seems every body part was equally as effective in scoring, apart from the "Other" body zone.

3.3 Analysis of teams and players scoring ratio

In this part we want to describe the number of goals scored by players and teams by using the goal ratio, so the number of goals scored on the number of total shots during the events.

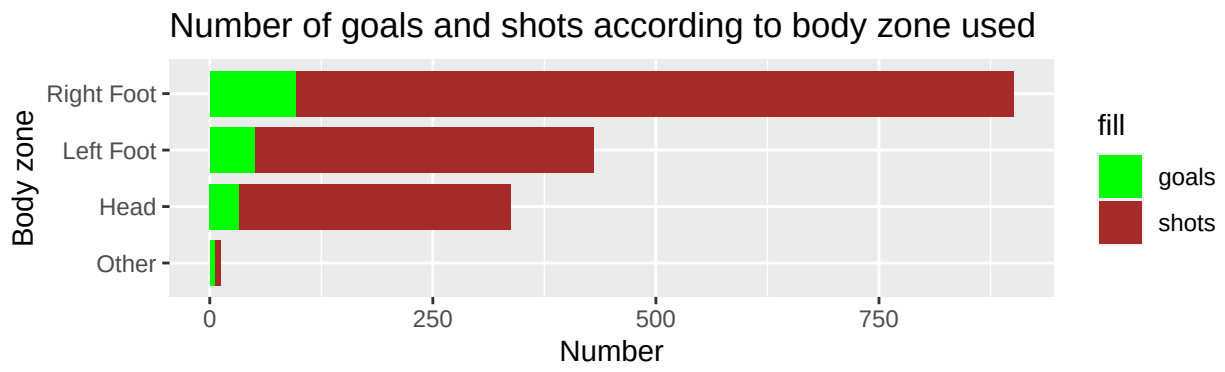


Figure 7: Diagram of the number of goals and shots according to body zone used

3.3.1 Players scoring ratio

Firstly, we look at player scoring ratio, which is calculated by dividing the total amount of goals a player scored, by the number of shots they took throughout the competition. This may lead us to the conclusion that efficiency varies highly from one player to another.

Figure 8 shows a visualization of the goal ratio of players that performed at least 1 shot during the World Cup.

Goal ratio on the number of totals shots for players

Some information on this graph: the color of the dots represented, as indicated in the legend, at what stage of the competition the player's team finished, so the more the dots tend towards purple, the more matches the players played and therefore went further in the competition. The graph shows a clear trend: the more shots players take, the lower their ratio. Even so, the players on the best teams -that is, those who finished in the top positions - often have a higher ratio than those on teams who lost earlier in the tournament. We also notice that the average (represented by a red cross) is quite low compared to the points we see on the graph, intuitively we'd probably have placed it around 10 shots and 0.25 goal ratio. But then we realized that many points, especially those at the bottom of the graph with 0 goals scored, are superimposed, so there are many players who shot without scoring, which is fairly consistent with the match statistics of around ten shots per team for an average of between 1 to 4 goals per match. An important piece of information is missing on this graph. Results show that the more players attempted shots, the worse their ratio of successful shots, but we don't take into account the difficulty of the shots perfomed by the players. Indeed a penalty is intuitively easier to score than an open shot far from the goals and under the pressure of defenders. It's something we already saw in the Figure 4, that goals on penalty occurred more often even though that there are less frequent in games.

So, we ask ourselves base on which criteria we could try to implement the notion of shot's difficulty, and we find that the notion of Expected Goals is what we were looking for and that the StatsBomb dataset provide it with every shots attempt.

The "Expected Goal", often named "xG" is the probability to score given a lot of datas on the shot, for example it can be the shooter and the goal position, the fact that the striker is under pressure or not, with which foot he is attempting this shot and various differents variables. This probability is computed by the xG model of StatsBomb which is probably created on a large amount of data that they were able to collect.

So this notion of xG is well-suited to implement the notion of shots difficulty. We can compute the mean of the xG of every shots that one player attempted to see if the player had easy or hard shots to perform.

This his what the Figure 9 is showing true goal ratio on the StatsBomb xG ratio for every player which attempted at least 1 shots during the World Cup.

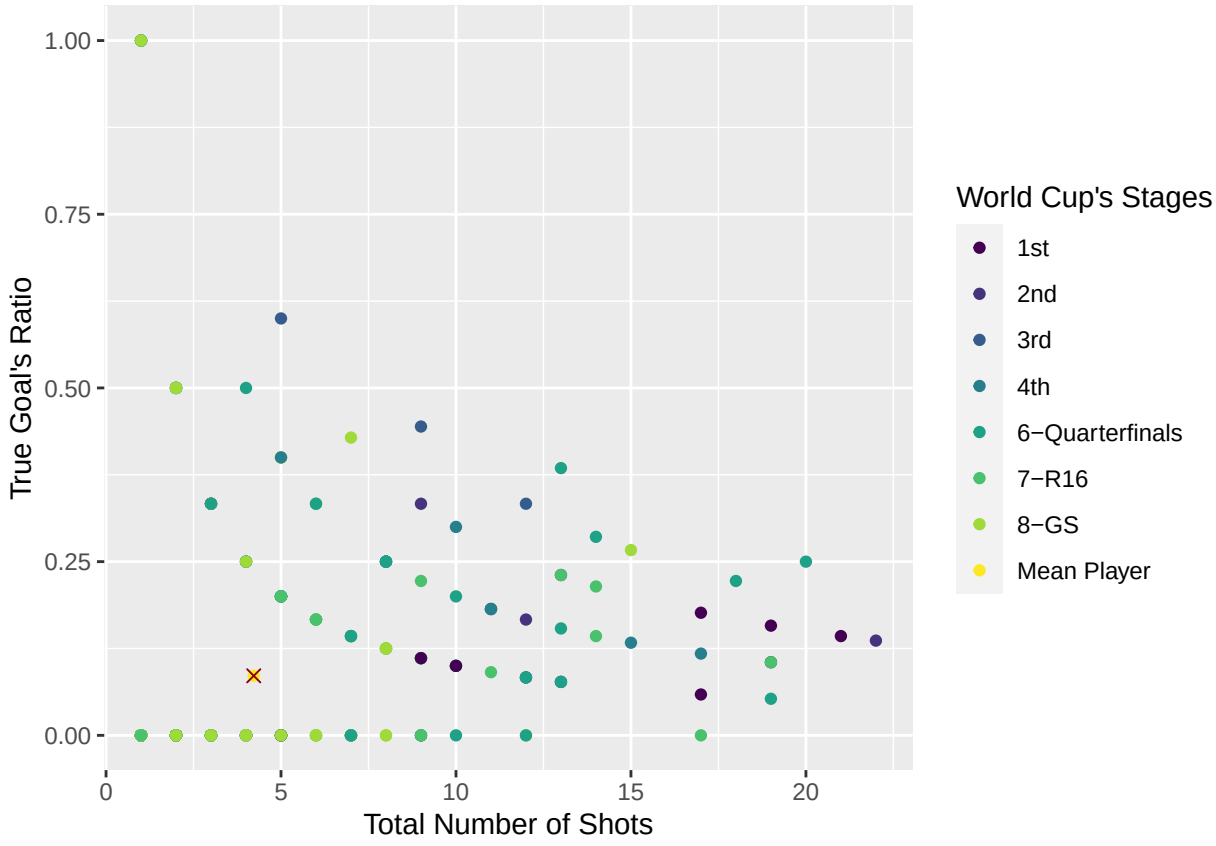


Figure 8: Goal ratio on the number of totals shots for players

Goal ratio on the Expected Goal ratio (StatsBomb model) for players

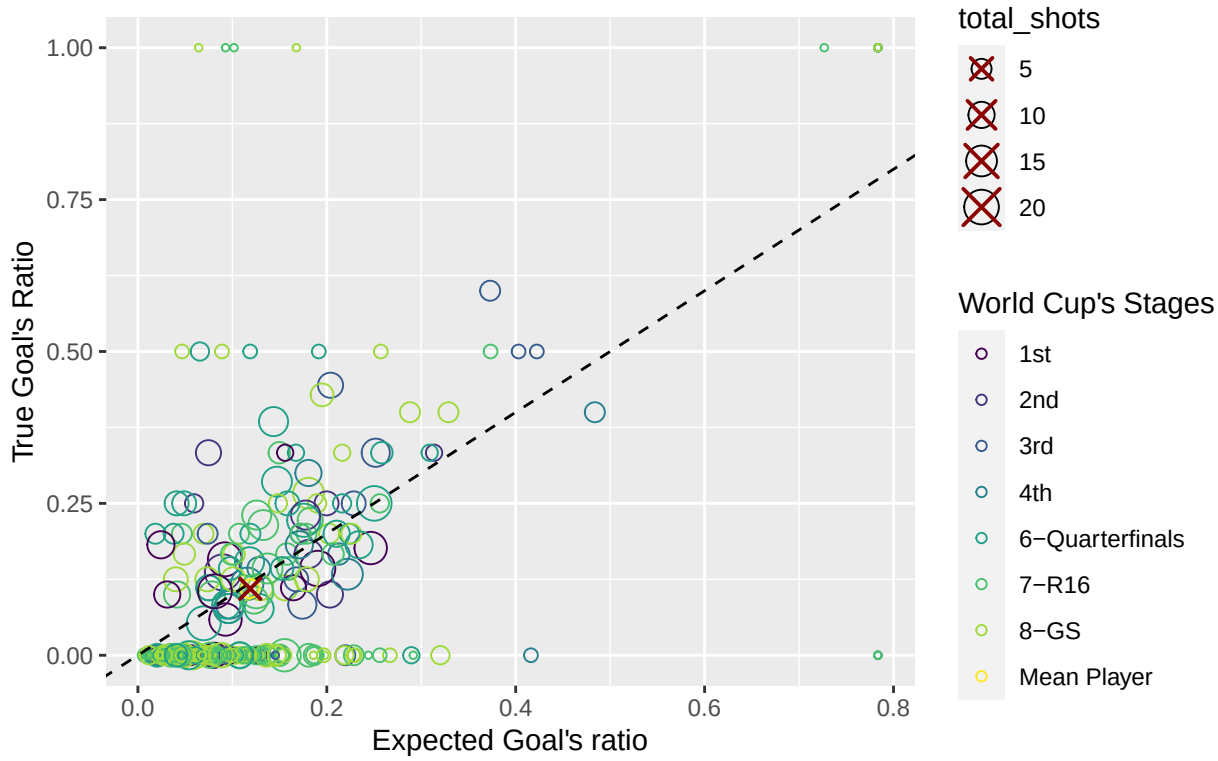


Figure 9: Goal ratio on the Expected Goal ratio (StatsBomb model) for players

This graph compares the number of goals predicted by StatsBomb with the actual goals scored. It shows whether players scored more or fewer goals than the StatsBomb model predicted. We observe the same biases as in the previous graph: players who scored on their only shot appear to be better performers, but this does not guarantee consistent performance, as we cannot assess regularity. This observation also applies to all players represented by the smaller circles.

Since neither axis in this graph displays the number of shots taken, we add this information by modifying the size of the circle representing each player. This adjustment emphasizes the importance of considering the number of matches played and the frequency of shots taken during the game.

Interpreting this graph, we can see from the colors that most players who reached the quarter-finals or beyond are clustered around the dotted line, which symbolizes player performance. This line, represented by the equation y = x, indicates the expected performance of the players. Thus, we can deduce that players from teams that advanced to the quarter-finals and beyond under performed less frequently and over performed less often, as they took more shots, reducing the bias from a single successful difficult shot.

3.3.2 Teams scoring ratio

Secondly, we analyzed teams using the same method we applied to players, which could lead to additional insights. This approach reduces the number of outliers because we are analyzing all team players who took at least one shot during the tournament. For instance, if a midfielder took only one shot and scored, while a striker took ten shots and scored twice, the mid-fielder's bias has less impact on the team's overall performance. Therefore, this team analysis allows us to determine which teams under performed in terms of goals and to investigate whether this correlates with the teams that were eliminated earliest in the competition.

Figure 10 visualizes the goal ratio relative to the total shots taken by teams that made at least one attempt during the World Cup.

Goal ratio on the number of total shots for teams

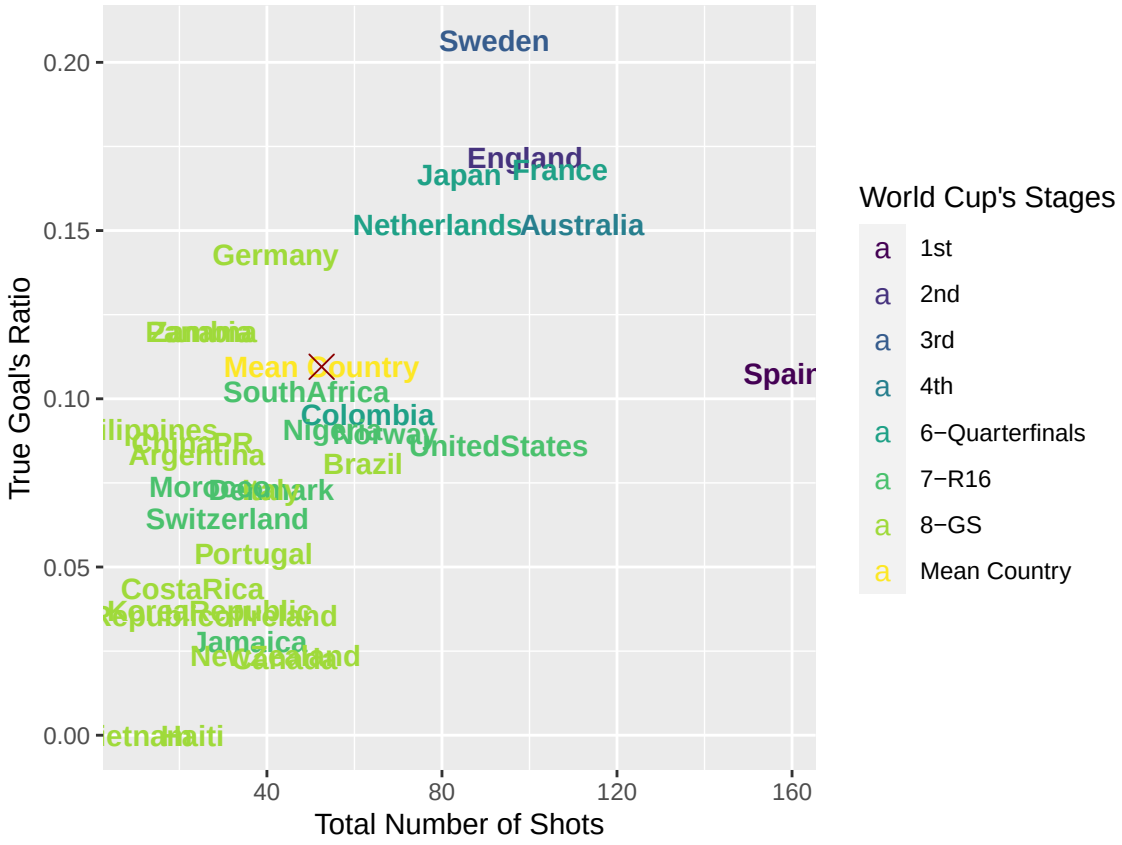


Figure 10: Goal ratio on the number of total shots for teams

In this graph, countries are represented by dots labeled with their names, and the color indicates the stage of the competition they reached. There is a clear separation between the bottom left of the image, which represents countries with few shots and few goals, and the top right, which includes the eight best teams, excluding Colombia.

We observe that Spain, the tournament winner, took many more shots than its rivals but scored less, suggesting strong domination during matches but less effective finishing. It would be interesting to investigate whether the difficulty of the shots explains the lower accuracy or if there is another interpretation.

Other outliers include Sweden, which finished third and had the highest ratio of successful shots, and Vietnam and Haiti, which did not score a single goal during the competition.

Similar to our player analysis, we aim to implement Expected Goal (xG) data to assess and analyze the difficulty of shots taken by teams.

Figure 11 shows the true goal ratio relative to the StatsBomb xG ratio for each team during the World Cup.

Goal ratio on the Expected Goal ratio (StatsBomb model) for teams

In this graph, we can extend our interpretations from the previous graph. Sweden, on the right-hand side of the graph, had the easiest shots on average according to StatsBomb's model and performed as expected, or

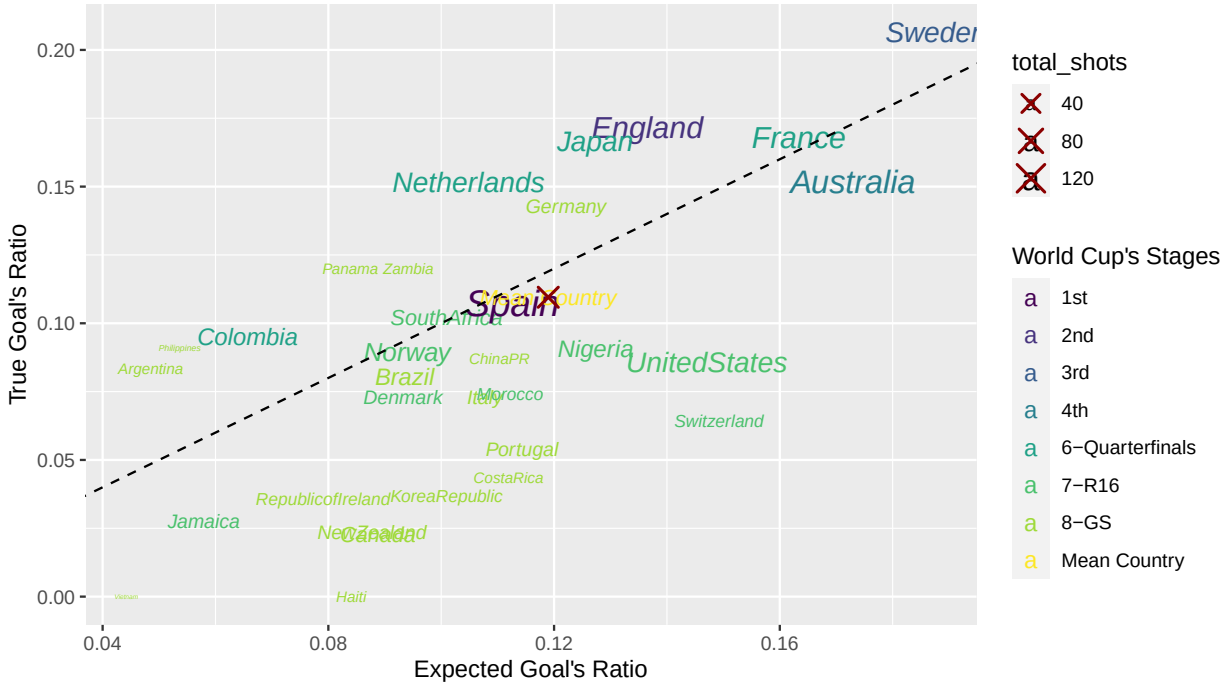


Figure 11: Goal ratio on the Expected Goal ratio (StatsBomb model) for teams

even slightly better, as it lies just above the dotted line that marks expected performance.

For Spain, we see that they had the most difficult shots among all the teams that reached the quarter-finals, except for Colombia. This explains why Spain had a higher number of shots with a relatively low success ratio compared to their main rivals.

It is noticeable that the best teams have higher xGs, indicating that their shots are more likely to succeed on average. This could mean that their forwards are getting into better shooting positions, either due to their superior skill or because the opposing teams' defenders are weaker. This suggests that the defense of the best teams is significantly stronger compared to that of the teams eliminated earlier in the tournament, making it difficult for these weaker teams to score.

Having compared the true goal ratio to the StatsBomb xG ratio, we wanted to understand how an xG model works in detail. Therefore, we decided to create our own xG model.

4 Models analysis

We developed various models to create our own xG model, identifying the most relevant variables for predicting goals.

We run a logistic regression model: we want the output to be 0 or 1 depending on whether the shot turns into a goal.

4.1 First model : body part, technique, type of shot

The first model keeps the variables studied previously : body part, technique, type of shot.

R^2 for the model without interaction is : 0.144105

R^2 for the model with interaction is : 0.1460338 .

4.2 Our target model : the expected goal variable

We now create a model composed of a single variable: the expected goal given in StatsBomb.

Our goal in creating the different models in this section is to find the most accurate model possible, which can have an R^2 close to this model (with only the expected goal as a variable), i.e. an R^2 close to : 0.262161.

4.3 Model 3 : Adding location.x and location.y

Any player on the field is represented as a moving point within a rectangle measuring 80 by 120 meters. The player's horizontal movement, from one goal post to another, is tracked by the variable location.x, while the vertical movement is tracked by location.y. We have now added location.x and location.y to our previously adjusted model.

We test a regression without interaction, and obtain an R^2 of : 0.2054155 .

With interactions, we get an R^2 of : 0.2323348.

In this model, we targeted the main variables to obtain a good model and an R^2 as close to 1 as possible.

4.3.1 Significance of variables ?

We run several tests to see which variables are significant in the model.

```
## Analysis of Deviance Table
##
## Model 1: shot.outcome.name ~ (location.x + location.y + shot.body_part.name +
```

```
##     shot.type.name)^2
## Model 2: shot.outcome.name ~ (shot.body_part.name + shot.technique.name +
##     shot.type.name + location.x + location.y)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1658     917.76
## 2      1637     891.22 21   26.542   0.1865
```

We see that we can remove the technique because $p - value$>0.05 so we can accept the sub-model with a 95% level.

For this sub-model without the technique variable we obtain an R^2 of : 0.2094726

The R^2 is no greater than for model 3 with interactions: this is normal because the R^2 favors models with many variables.
We should look at other variables such as AIC score, which is minimal for model 3 without interactions.

4.3.2 Comparison of norms

We now want to compare model 3 with and without interaction : the closer the 2-norm is to 0, the better the model.

Norm L2 for the model_3 without interaction is equal to : 4.1435027.

The value for the model_3 with interactions is : 4.2588035.

We find the same results as with the AIC criterion. This is consistent with the fact that R^2 favors models with many variables, so it's better to evaluate with AIC. We can conclude that the model 3 without interaction is best.

We do the same to compare model 1 with and without interaction.

The L2 norms are respectively : 4.6785642 and 4.6786731.

Both models are less accurate than the 3rd one.

4.4 Model 4 : adding the under_pressure variable

We now create a new model like the model_3, but adding a variable : under_pressure.

4.4.1 Test for significance of single variables

First, we test the significance of this new variable.

```
##
## Call:
## glm(formula = shot.outcome.name ~ under_pressure, family = binomial(link = "logit"),
##     data = df_model_4)
##
## Coefficients:
```

```
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.94591    0.09486 -20.513   <2e-16 ***
## under_pressureTRUE -0.41957    0.16790  -2.499   0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1160.9  on 1679  degrees of freedom
## Residual deviance: 1154.5  on 1678  degrees of freedom
## AIC: 1158.5
##
## Number of Fisher Scoring iterations: 5
```

We see that $p_{\text{value}} < 0.05$, so we reject H_0 : playing under pressure is significant.

Estimated coefficients are negative, so playing under pressure reduces the probability of scoring.

Testing the model with only the shot.body_part.name variable gives us a p_{value} of : 0.042.
We reject H_0, the technique variable is significant.

We now want to test the model with only the shot.technique.name variable.

```
##
## Call:
## glm(formula = shot.outcome.name ~ shot.technique.name, family = binomial(link = "logit"),
##     data = df_model_4)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.657e+01  6.927e+02  -0.024    0.981
## shot.technique.nameDiving Header  8.145e-08  1.277e+03   0.000    1.000
## shot.technique.nameHalf Volley    1.395e+01  6.927e+02   0.020    0.984
## shot.technique.nameLob            1.547e+01  6.927e+02   0.022    0.982
## shot.technique.nameNormal         1.461e+01  6.927e+02   0.021    0.983
## shot.technique.nameOverhead Kick  8.143e-08  1.141e+03   0.000    1.000
## shot.technique.nameVolley         1.389e+01  6.927e+02   0.020    0.984
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1160.9  on 1679  degrees of freedom
## Residual deviance: 1143.9  on 1673  degrees of freedom
## AIC: 1157.9
##
## Number of Fisher Scoring iterations: 15
```

The reference is backheel : all the other techniques are better, we have a lot of values close to 1, we could do a constant sub-model to see if this variable is significant.

We find a p_{value} of 0.009. We reject H_0, the technique variable is significant.

We are now testing the model with only the shot.type.name variable. We also run a sub-model test.

We find a p_{value} of 0.

The variable shot.type.name is significant, we reject H_0.

We do the same with the variable location.x :

We see a p_{value} of : 0. <0.05 so location.x is highly significant.

We check if the variable location.y is significant as well.

The p_{value} is : 0.915. > 0.05 so location.y is not significant.

4.4.2 Testing the complete model

The model is now tested with all the following variables: shot.body_part.name,shot.technique.name,shot.type.name,location.x

We have an R^2 of 0.2054992 which is good, but it's normal because it's a model with many variables.

We also note a low AIC, which is equal to `954.3696823`.

4.5 Model 5 : adding the position of the goalkeeper

We create the same model as above, but adding the position of the goalkeeper.

4.5.1 Does the position of the goalkeeper in x and y improve our results ?

First we test the model with only the location.x.GK variable.

```
##
## Call:
## glm(formula = shot.outcome.name ~ location.x.GK, family = binomial(link = "logit"),
##     data = df_model_6)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -36.22586    4.49351  -8.062 7.52e-16 ***
## location.x.GK   0.28758    0.03775   7.617 2.59e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1160.9  on 1679  degrees of freedom
## Residual deviance: 1104.7  on 1678  degrees of freedom
## AIC: 1108.7
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] 0.04846669
```

Significant effect of goal position in x because both p_{values} are lower than 0.5.
The AIC value is low, equals to `1108.6754173`.

Then we do the same but with the location.y.GK variable.

We find that the variable for keeper position in y is significant as well. AIC is slightly higher than for position in x, it's equal to `1164.4331267`.

4.5.2 Complete model

For the model with all the preceding variables and without interaction, we find a very low AIC=`950.1303381`.
We can conclude that this model is really good.

4.6 Keeping all significant variables and removing location.y

Since we found that location.y is not significant, we can remove it from the model.

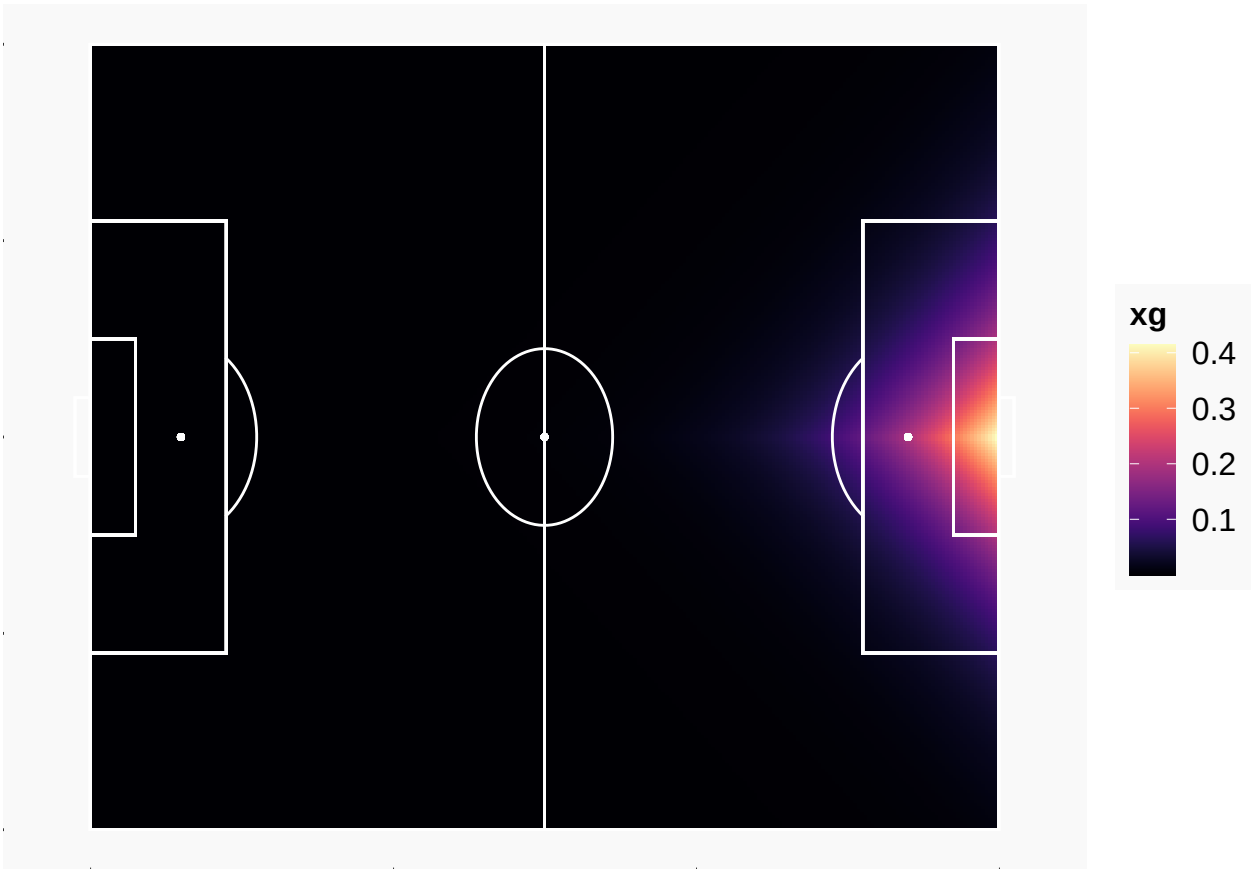
Without this variable, the AIC is even lower, at `866.1274211`.

We can conclude that we have found our best model for now and it's composed of the variables :

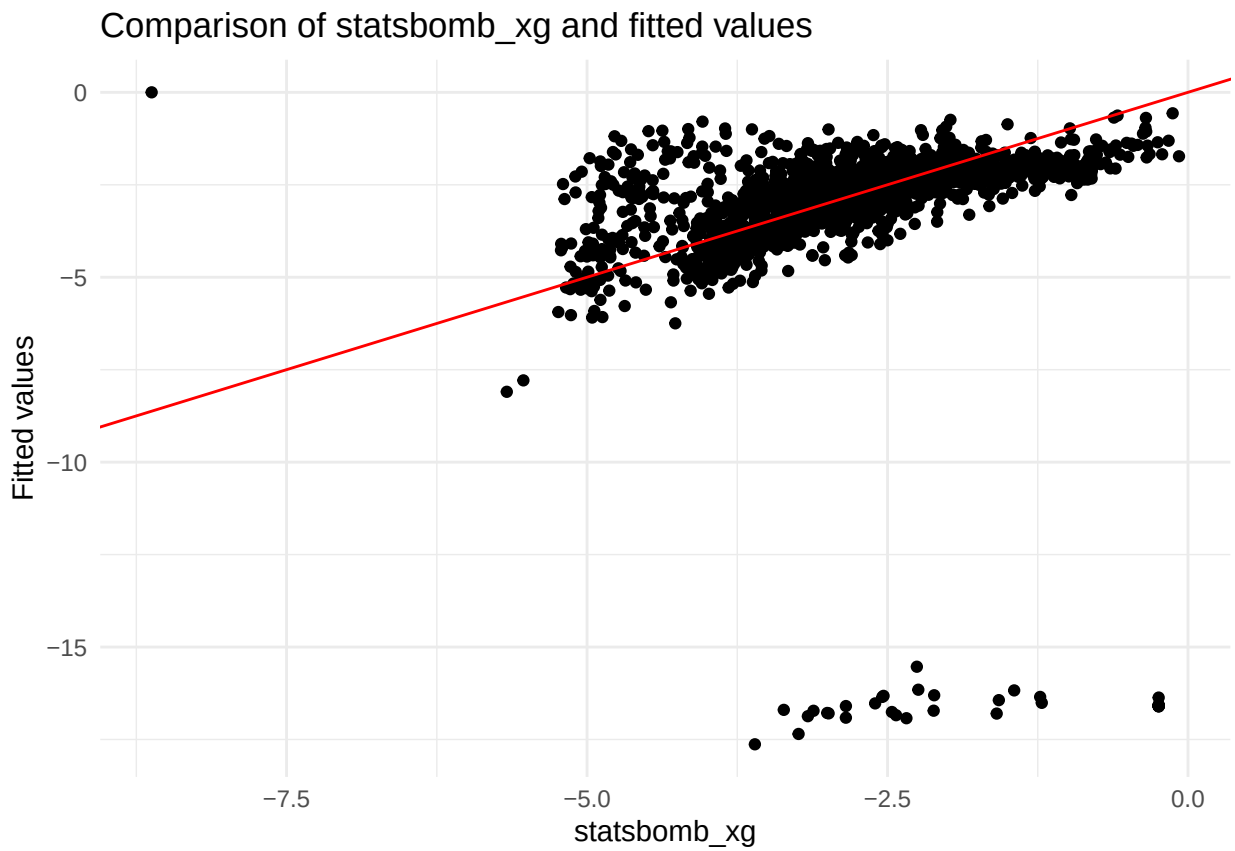
body_part, shot.technique, shot.type, location.x, under_pressure, location.x.GK, location.y.GK.

4.7 Replacing location.y

Since location.y is a very insignificant variable, we now consider offense as symmetrical with respect to the axis passing through the center point of the pitch and the penalty spots. We redefine location.y as the distance to center, with this new variable ranging from 0 to 40 meters. This adjustment should give more meaningful insights, as a high distance to center indicates that a player is very off-center.



The heat map above shows expected results: from a very simple model that only uses location and distance to center, the model predicts better scoring chances for point-blank shots compared to shots taken outside the penalty area. This is easily explained by the fact that our dataset contains few goals made from outside this area.



We have many values close to 0, so we use a log transformation to better observe the residuals.

The data indicate that only one point is being overestimated by our model. This point corresponds to the goal made from a corner kick mentioned previously. Our model predicts an xG of 1, perfectly illustrating the limits of our model. We only worked on this reduced dataset, and since only one corner shot was attempted, our model incorrectly assumes it is the most efficient shot. After transforming the residuals to their logarithmic counterparts, they appear well-adjusted for the most part, with the majority clustering around the desired spot. In the bottom right corner, we find points where our model underestimates the chance of a goal. These seem to be goals taken from far away, and our model consistently underestimates them. This is likely a consequence of our small dataset. Overall, our model gets relatively close to the xG of StatsBomb.

4.8 Finding our best model with the AIC criterion

```
##
## Call:
## glm(formula = shot.outcome.name ~ shot.body_part.name + shot.technique.name +
##     shot.type.name + location.x + location.x.GK, family = binomial(link = "logit"),
##     data = df_model_6)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.26950 2496.42516   0.000  0.99991
## shot.body_part.nameLeft Foot     0.73885    0.28416   2.600  0.00932 **
## shot.body_part.nameOther         1.66474    0.62359   2.670  0.00759 **
## shot.body_part.nameRight Foot    0.61594    0.26028   2.366  0.01796 *
```

```
## shot.technique.nameDiving Header    1.15111 1270.44869   0.001  0.99928
## shot.technique.nameHalf Volley     14.61311  688.68298   0.021  0.98307
## shot.technique.nameLob             16.50566  688.68416   0.024  0.98088
## shot.technique.nameNormal          15.26610  688.68296   0.022  0.98231
## shot.technique.nameOverhead Kick    0.58723 1122.64416   0.001  0.99958
## shot.technique.nameVolley          14.48016  688.68305   0.021  0.98323
## shot.type.nameFree Kick           -17.10201 2399.54497  -0.007  0.99431
## shot.type.nameOpen Play           -16.88807 2399.54474  -0.007  0.99438
## shot.type.namePenalty             -13.31112 2399.54477  -0.006  0.99557
## location.x                          0.13002    0.01748   7.439 1.02e-13 ***
## location.x.GK                      -0.12563    0.04842  -2.594  0.00948 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1160.94  on 1679  degrees of freedom
## Residual deviance:  916.26  on 1665  degrees of freedom
## AIC: 946.26
##
## Number of Fisher Scoring iterations: 15
```

Findings suggest that y-positions are irrelevant for the goalkeeper, as well as for the shooter. However, both x-positions are significant in the best model.

Our best model is composed of 5 variables : The technique of shot, the type of shot, the body part used, and the positions in x for both the player and the opposing goalkeeper.

5 Analysis of the Model performance

Now that we have developed our model composed of these five variables, we have questioned whether this type of model is superior to the StatsBomb xG model. Our primary objective is to assess the performance of our model.

Similar to the comparison conducted in the "Descriptive Analysis" section of our report, we are now comparing the ratio of goals. However, instead of comparing given data, we predict the xG of every shot using our best model and then compare these xG values to those provided by StatsBomb through visual graphics.

Figure 12 presents a visualization of these results.

StatsBomb xG ratio on our xG ratio based on bestmod for players

We observe that when displaying the ratio of the xG we predicted to those from StatsBomb, the results exhibit very little variation, indicating that our model closely aligns with that of StatsBomb! However, for a precise comparison between our model and StatsBomb's, it is essential to display the xG for each shot.

This is depicted in Figure 13 below:

StatsBomb xG on our xG based on bestmod for every shots in the World Cup

The results indicate that our models are comparable for the majority of shots. Nonetheless, we do observe some outliers, which may be elucidated by including additional factors such as the type of shot, whether it is a free kick, an open play shot, a penalty, a corner, and so forth. This is illustrated in Figure 14 below:

Figure 12: StatsBomb xG ratio on our xG ratio based on bestmod for players

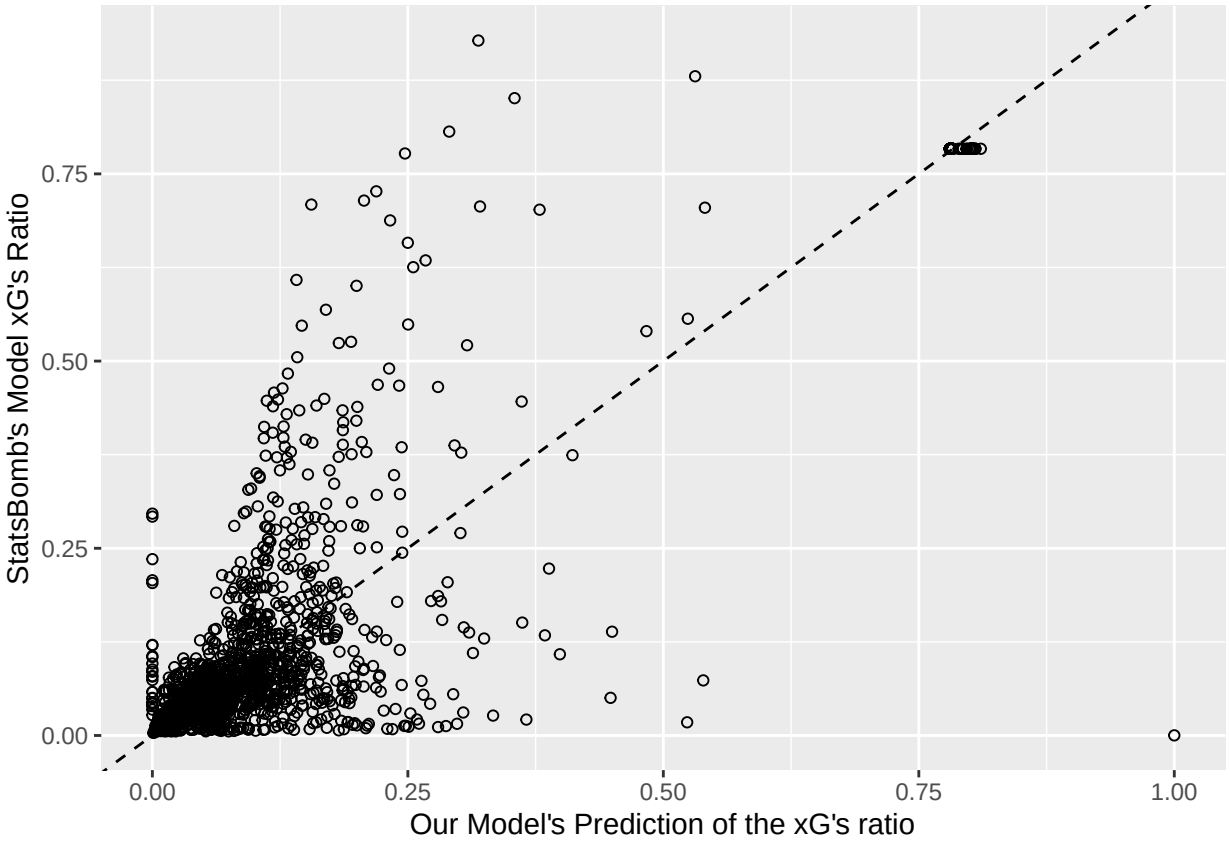


Figure 13: StatsBomb xG on our xG based on bestmod for every shots in the World Cup

StatsBomb xG on our xG based on bestmod and the type of shots for the 100 shots with the higher xG difference

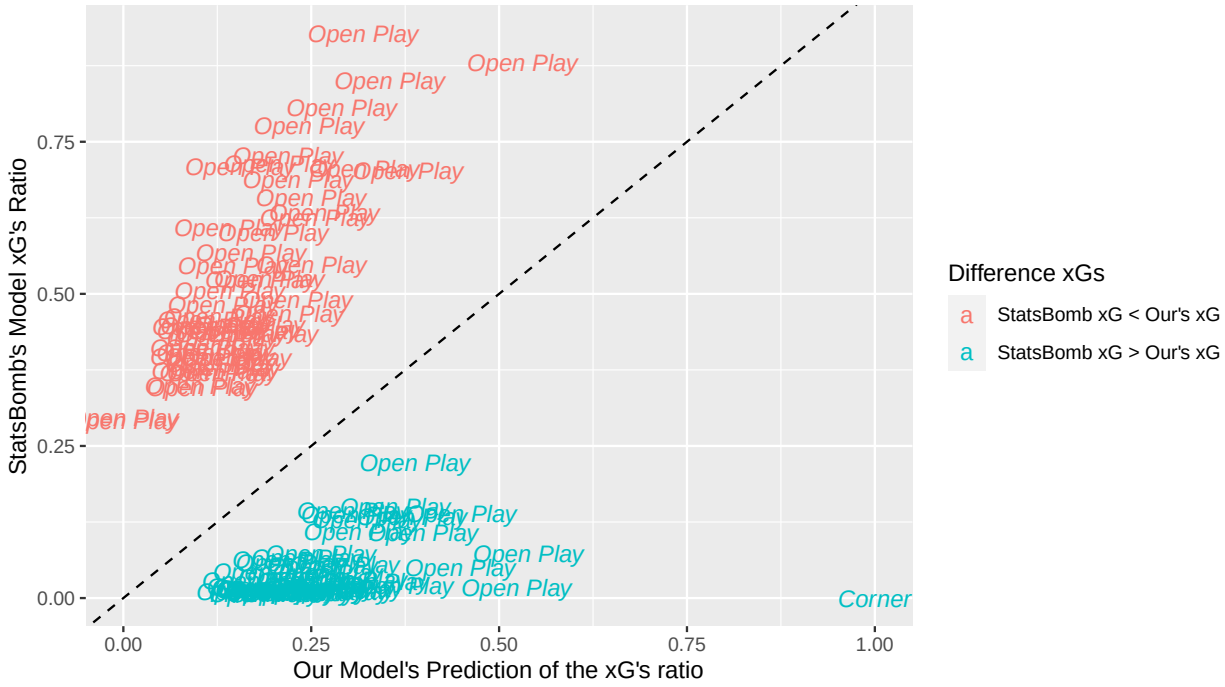


Figure 14: StatsBomb xG on our xG based on bestmod and the type of shots for the 100 shots with the higher xG difference

Our results reveal a prominent bias in the model, particularly evident in the 'Corner' category situated at the bottom right of the graph. In this instance, our model predicts an xG of 1, while StatsBomb predicts an xG of almost 0. During the competition, there was a successful in-swinging corner, leading our model, trained solely on this event, to classify all in-swinging corners as shots with a goal probability exceeding 0.99. Conversely, StatsBomb's model, likely trained on a broader dataset, possesses enough historical data to recognize that this rare type of shot seldom results in a goal, thereby assigning a probability close to 0.

Explaining the other points is more complex. These predominantly consist of open play shots, where numerous other factors come into play. Our model incorporates only five variables: the type of shot (penalty, free kick, open play, etc.), the shot technique (normal, lob, volley, bicycle kick, etc.), the body part used for the shot, and the distance between the player and the goalkeeper. Consequently, we may be overlooking certain information, such as shot angle or whether the player was under pressure. To refine our model and address the idiosyncrasies of this competition, including anomalies like the in-swinging corner, further data from women's football tournaments is required, analyzing more than the current 64 matches.

6 Another Model : Expected Pass

Having scrutinized the Expected Goal (xG) values provided by StatsBomb and formulated our own xG prediction model, we pondered whether StatsBomb offered a comparable metric for passes—an Expected Pass (xP) indicator, forecasting the likelihood of a pass succeeding based on various parameters. To our surprise, such an indicator was not available. Consequently, we undertook the task of implementing one, employing the same methodology utilized for our xG model. We opted to develop this model for passes due to the significant correlation between a football team's performance and goals, which ultimately determine match outcomes. However, upon observing football matches, it becomes apparent that certain teams encounter difficulties in creating scoring opportunities and goal-scoring situations due to challenges in executing passes between lines, through passes, player combinations, technical skills, or dribbles.

We observed that teams advancing to the quarterfinals generally demonstrated superior xG and True Goal Ratios compared to others, albeit with exceptions such as Colombia. This model offers us an avenue to delve into another facet of the intricate sport of football.

For instance, did Colombia excel in play construction compared to their opponents, thereby securing a quarterfinal berth despite similar xG and True Goal Ratios? Furthermore, do we observe a similar trend regarding goal-scoring prowess among top teams? These inquiries motivated the development of this passes analysis model.

6.1 Quick Descriptive Analysis

6.1.1 Player-by-player visualization

True Pass ratio on the Total Number of Pass for every player with 5 pass or more during the WC

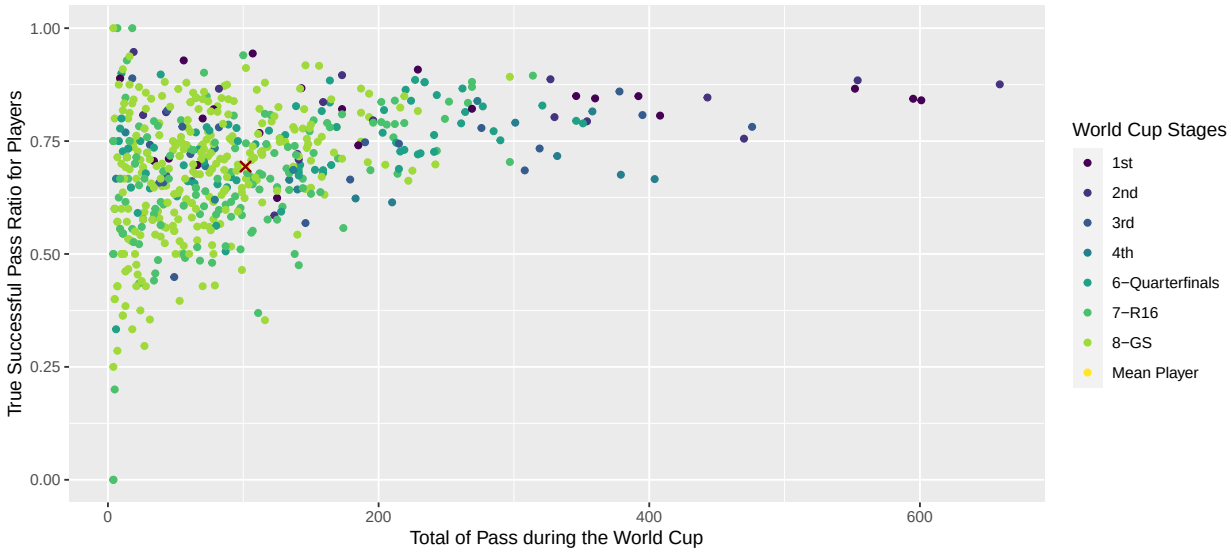


Figure 15: True Pass ratio on the Total Number of Pass for every player with 5 pass or more during the WC

The graph on 15 depicts the ratio of successful passes to the total number of passes. Once again, a noticeable disparity is evident: players from countries that advanced to the quarterfinals occupy the uppermost positions on the graph, boasting a successful pass ratio exceeding 60%, whereas teams eliminated in the round of 16 and group stages display notably lower pass ratios. It is worth mentioning that even within these less successful teams, a considerable number of players exhibit pass ratios comparable to those in the top-performing teams. This suggests potential variations within these countries, likely indicating a

greater degree of diversity in player performance.

6.1.2 Team-by-team visualization

True Pass ratio on the Total Number of Pass for every teams during the WC

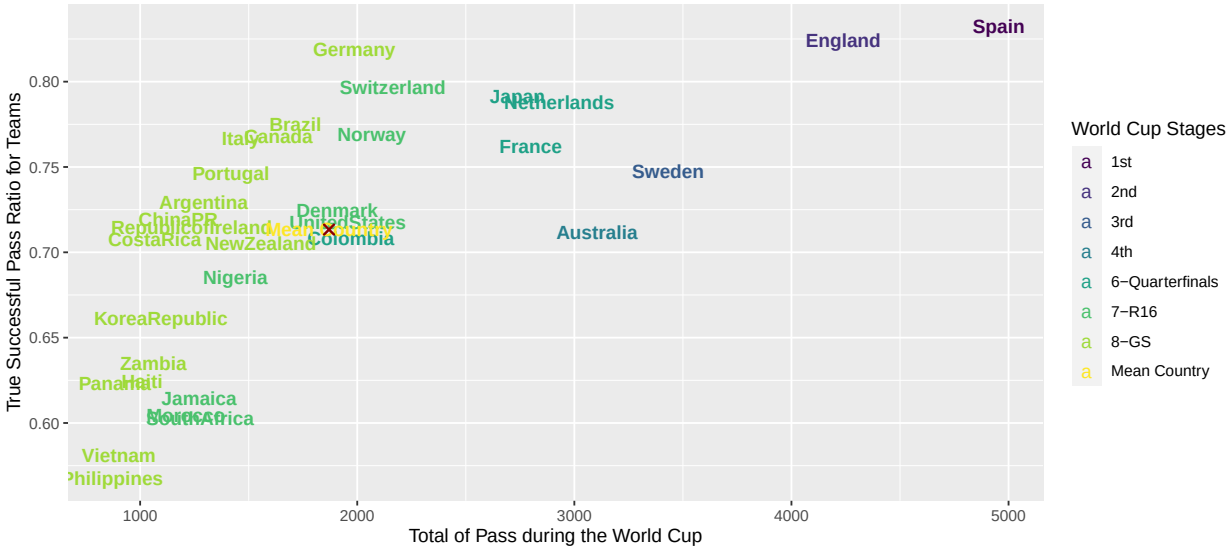


Figure 16: True Pass ratio on the Total Number of Pass for every teams during the WC

We analyzed the statistics by country and observed a consistent trend, with a notable distinction for Spain. Spain made significantly more passes than other teams that played a similar number of matches, such as Sweden, Australia, and England, while also achieving a higher pass ratio. When combined with shooting data, where Spain had many more shots than its competitors, this observation reaffirms that Spain controlled their matches more effectively, with higher possession and more opportunities. This trend likely extends to England, the other finalist team. As for Sweden and Australia, the third and fourth-place teams respectively, we can infer that their passes were riskier, potentially involving more line-breaking passes or through balls. Colombia, once again, is positioned among teams that were eliminated earlier and aligns with the average performance level of the competition.

To provide deeper interpretation and validate our analyses with these visualizations, we will now focus on the Expected Pass (xP) indicator, particularly the difficulty of the passes. However, since StatsBomb did not provide an xP indicator in the dataset, we will not be able to compare our models to some better models with a wider training set.

6.2 Implementing the Pass model we created

After this short descriptive analysis for the passes, we create our model to predict the probability of a success of a pass. We used the same methodology as for the xG model, so we won't detail a lot how we create this model.

This lead us to this model below :

```r
# There seems to be some non-linear dependencies within the model, as our barplot suggested.

dfmod <- dfmod %>% mutate(length_sq = length**2)

mtest <- glm(outcome ~ location.x*distance_to_center*length + angle * location.x + angle * distance_to_
#summary(mtest)

mfinal <- step(mtest, backward=TRUE, k=log(nrow(dfmod)),trace=FALSE)
summary(mfinal)
```

```
##
## Call:
## glm(formula = outcome ~ location.x + distance_to_center + length +
##     under_pressure + length_sq + location.x:distance_to_center +
##     location.x:length + distance_to_center:length + length:under_pressure +
##     under_pressure:length_sq + location.x:distance_to_center:length,
##     data = dfmod)
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.113e+00  1.586e-02  70.149  < 2e-16 ***
## location.x                          -4.966e-03  2.386e-04 -20.812  < 2e-16 ***
## distance_to_center                  -9.711e-03  5.755e-04 -16.875  < 2e-16 ***
## length                              -2.880e-03  7.167e-04  -4.018 5.87e-05 ***
## under_pressure                      -2.791e-01  1.192e-02 -23.410  < 2e-16 ***
## length_sq                           -1.839e-04  7.303e-06 -25.176  < 2e-16 ***
## location.x:distance_to_center        1.573e-04  9.006e-06  17.466  < 2e-16 ***
## location.x:length                    1.435e-04  8.961e-06  16.015  < 2e-16 ***
## distance_to_center:length            2.078e-04  1.979e-05  10.500  < 2e-16 ***
## length:under_pressure                1.727e-02  1.083e-03  15.939  < 2e-16 ***
## under_pressure:length_sq            -2.858e-04  1.962e-05 -14.564  < 2e-16 ***
## location.x:distance_to_center:length -5.619e-06  3.308e-07 -16.984  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1710626)
##
##     Null deviance: 11508  on 59836  degrees of freedom
## Residual deviance: 10234  on 59825  degrees of freedom
## AIC: 64168
##
## Number of Fisher Scoring iterations: 2
```

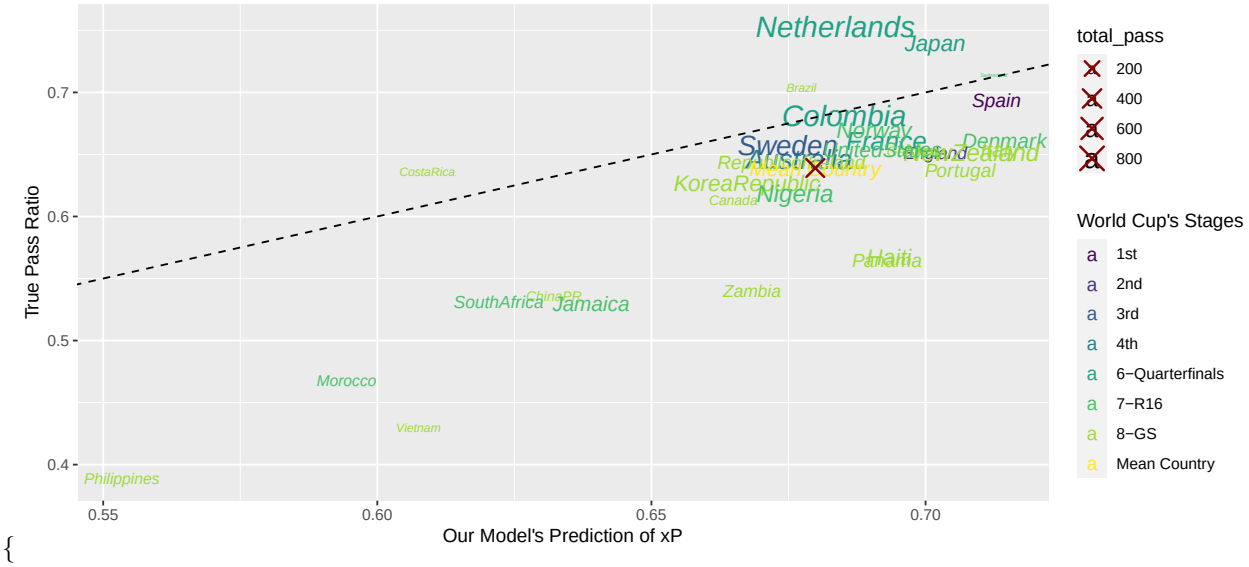
Thus, the model we retain includes the variables: location.x, distance to center, length, under_pressure, length_sq, and several interactions between these variables. We use more variables than in the xG model because we have more data points for passes, given the higher frequency of passes during the competition. This likely reduces bias and enables us to create a more precise model.

As mentioned earlier, we cannot compare our pass model to a similar model by StatsBomb since such a model does not exist. Instead, we decided to extend our analysis by considering the notion of pass difficulty. We divided our dataset into two subsets: one containing the easiest passes and the other containing the most difficult passes. This approach allows us to address the questions we posed about the performance of specific teams.

6.2.1 Comparaison xP to True Pass Ratio - Hardest Passes

Firstly, We look at the 25% hardest pass in term of xP according to our model. This represents all the passes that our model predict to have an probability to be succeeded less than 72%.

\begin{figure}



{

}

\caption{True Pass ratio on our model xP ratio for the 25% hardest pass for every country} \end{figure}

In figure 2 shows a visualization of these results., we observe several interesting points in this graph. Spain, having made more than 5,000 passes, has a significantly lower proportion of difficult passes compared to other teams. On average, Spain's passes were simpler, even considering only the hardest 25% of passes according to our xP model. This suggests that their style of play relied more on possession, involving ball circulation and fewer challenging passes, such as long crosses or through balls during counterattacks, likely due to higher ball possession. Furthermore, consistent with our analysis of Spain, the strongest teams are clustered together, along with some less strong teams, around an xP slightly below 70%. This indicates that their most difficult passes had, on average, a 70% chance of success, which is quite substantial.

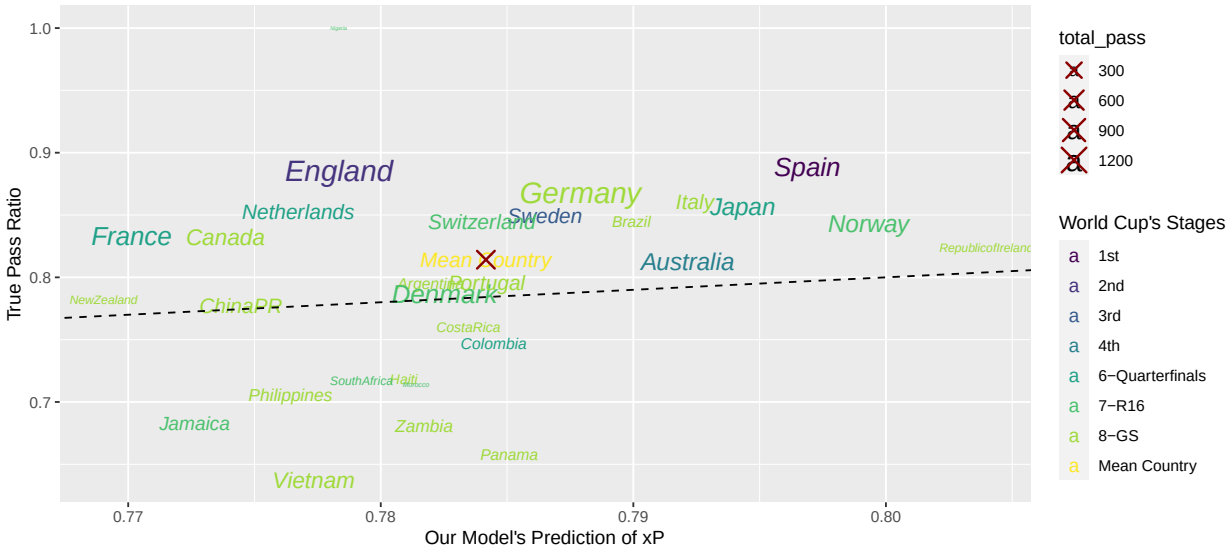
Additionally, we notice three countries—Japan, the Netherlands, and Colombia—within this zone but with a significantly higher number of passes. This indicates that these teams executed a higher proportion of difficult passes compared to other teams. However, they all lie near the equality line between the actual successful pass ratio and the xP ratio. This likely implies that their difficult passes, which are probably more dangerous for the opposing team's defense, were executed relatively successfully. This could explain why these three teams reached the quarterfinals, particularly for Colombia, as previously discussed.

6.2.2 Comparaison xP to True Pass Ratio - Easiest Passes

We know look at the 25% easiest pass in term of xP according to our model. This represents all the passes that our model predict to have an probability to be succeeded greater than 76,27%.

Figure 6.2.2 presents a visualization of these results.
True Pass ratio on our model xP ratio for the 25% easiest pass for every country during the WC

\begin{figure}



{

}

\caption{True Pass ratio on our model xP ratio for the 25% easiest pass for every country during the WC}
\end{figure}

When examining only the easiest 25% of passes according to our model, we immediately notice that the xP display range is very narrow, with a maximum difference of 4% between the xP of two teams, which is relatively small. Additionally, we observe that the majority of teams exceed their expected pass success ratios. It is also evident that the further teams advanced in the competition, the more simple passes they had to make and the more successfully they executed these simple passes. In contrast, other teams, such as Vietnam, missed more simple passes than expected, likely providing their opponents with more frequent scoring opportunities. These situations are likely dangerous, as our model, which accounts for field position, assigns more favorable xP values when a pass is made in the team's own half more frequently.

7 Conclusion

After acquiring a substantial understanding of football indicators, as well as the way StatsBomb keeps track of football data, we were able to distinguish trends in the winning, as well as the losing teams of the 2023 FIFA Women's World Cup. We looked at specific teams, such as France, Spain or Colombia, and managed to explain some of their success. Using basic logistic regression, and with only a very limited dataset, we implemented our own expected goal metric. Through AIC criterion step, we were able to come very close to the xG calculated by StatsBomb. This model contained the following observations, from the moment of the shot : the position of the shooter as well as the opposing goalkeeper, its distance relative to the central axis of the pitch, the type, body part and technique of the shot. Future research could expand on this analysis by incorporating more diverse datasets, including data from other tournaments and leagues, and use more advanced regression techniques. Such techniques could include support vector machine or neural networks, to predict more accurately the probability of a goal.

8 References

[1] Christian Collet (2012), *The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010* Retrieved from https://www.tandfonline.com/doi/full/10.1080/02640414.2012.727455#:~:text=Using%20data%20from%20five%20European%20leagues%2C%20UEFA%20and,team%20quality%20and%20home%20advantage%20were%20accounted%20for.

[2] StatsBomb. (2022), *StatsBombR: R Wrapper for StatsBomb Data.* Retrieved from https://github.com/statsbomb/StatsBombR

[3] StatsBomb. (2022), *StatsBombR: R Wrapper for StatsBomb Data.* Retrieved from https://statsbomb.com/

[4] StatsBombR Specifications. Retrieved from https://github.com/statsbomb/StatsBombR

[5] Working with R: Accessing & Working With StatsBomb Data In R. (2021). Retrieved from https://statsbomb.com/wp-content/uploads/2021/11/Working-with-R.pdf

List of Figures