Projet Women FIFA WC23 Analysis

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 gain advanced comprehension on the ways teams develop and convert 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 the StatsBomb's one, 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 soccer 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, for a total of 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 data set.

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

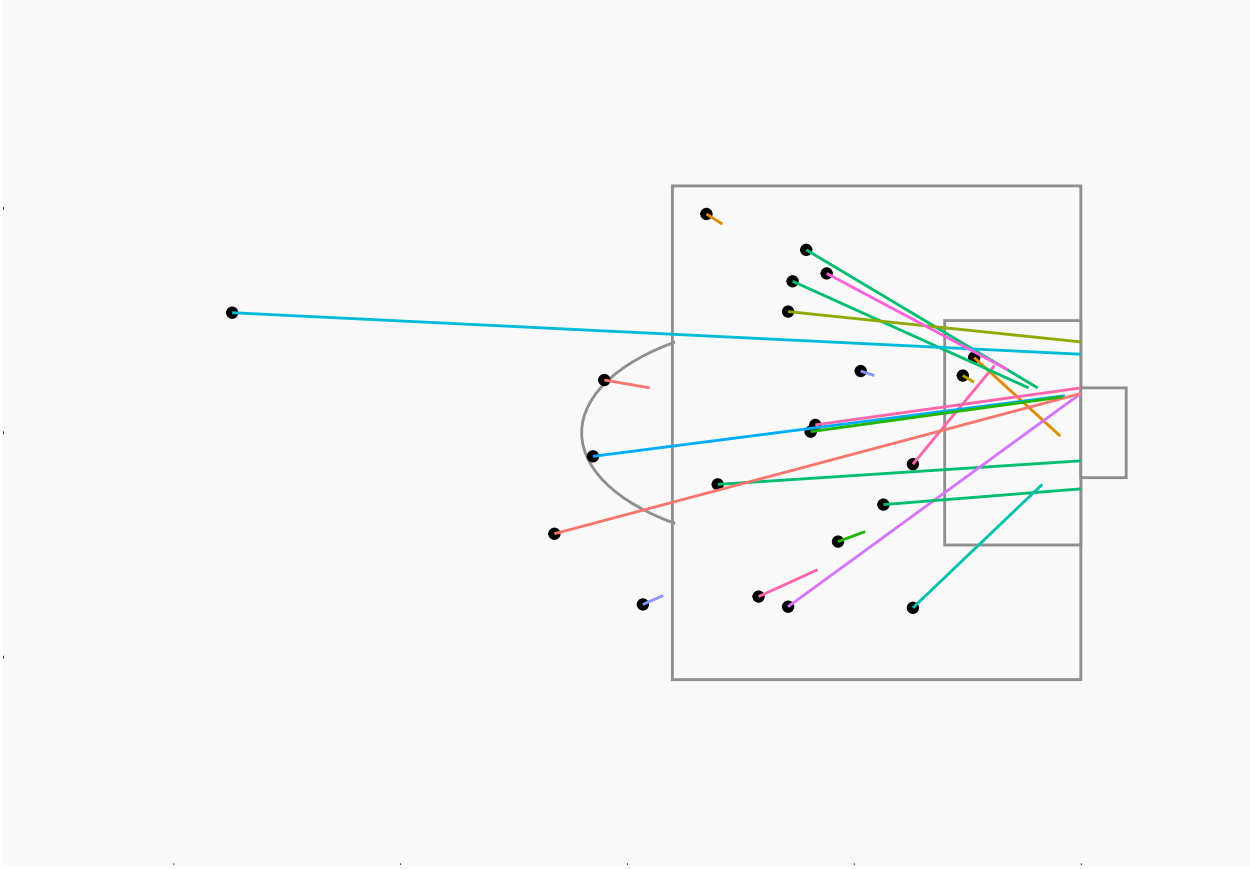


Figure 1: Visualization of the shots of Spain-England on the pitch

We can see 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.

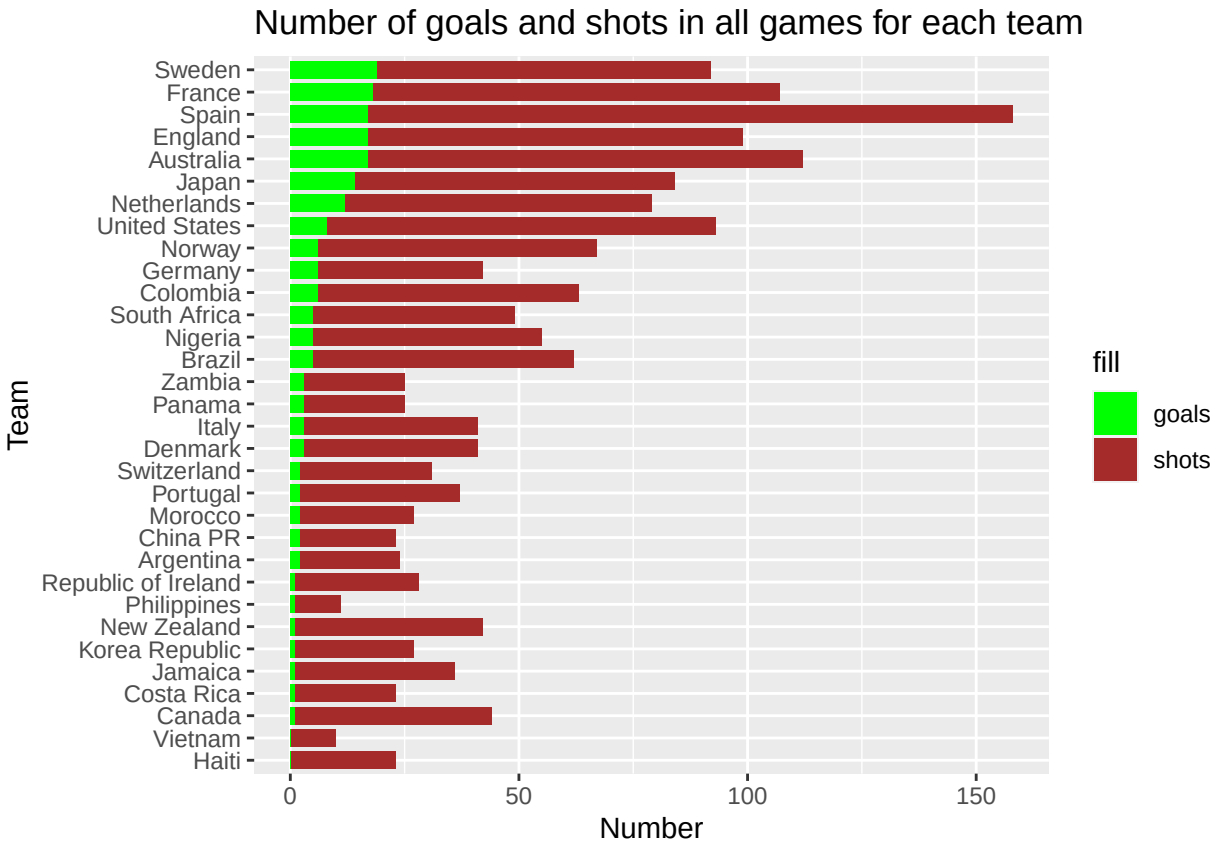


Figure 2: Diagram of the number of goals and shots in all matches for each team

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 over 2.5%. 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 wanted to realize 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 the results.

As we can see, there were a huge number of goals in two games : Panama France and Australia France. The first match ended on a exceptional 3-6 score for France, but the former ended on a draw. However, teams went to penalties, and they scored a total of 13, making this observation flawed.

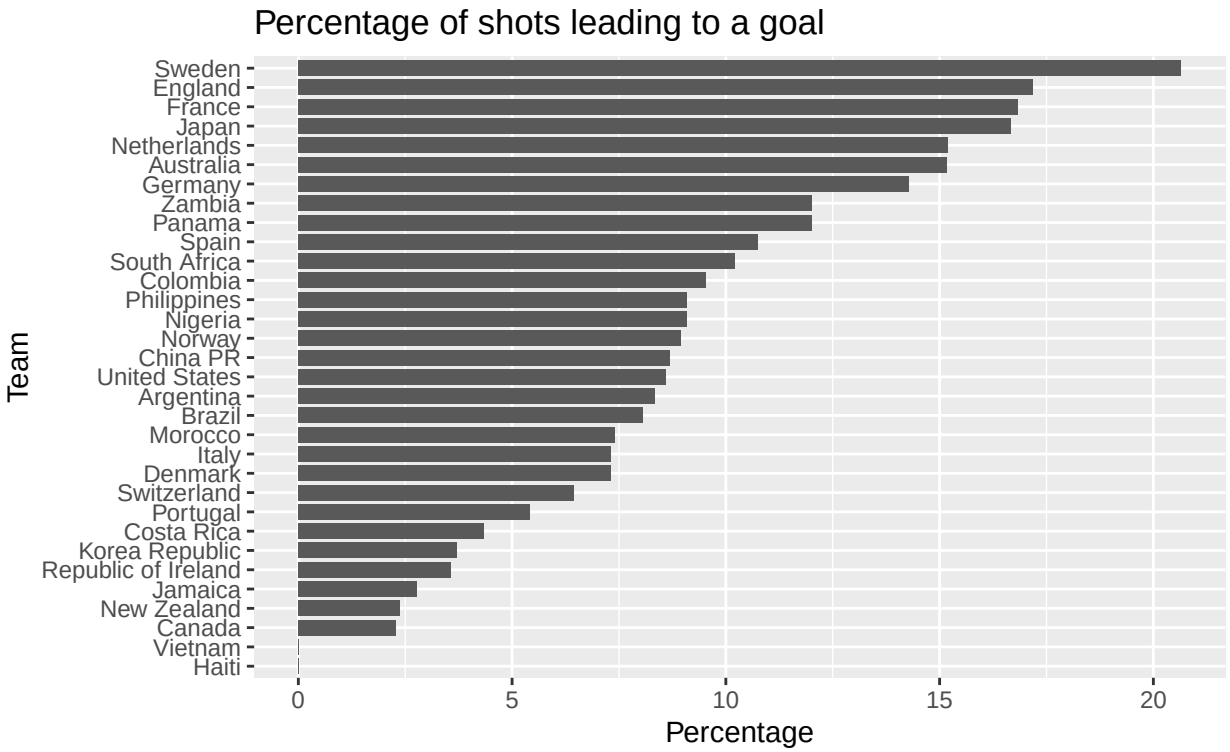


Figure 3: Diagram of the percentage of shots leading to a goal

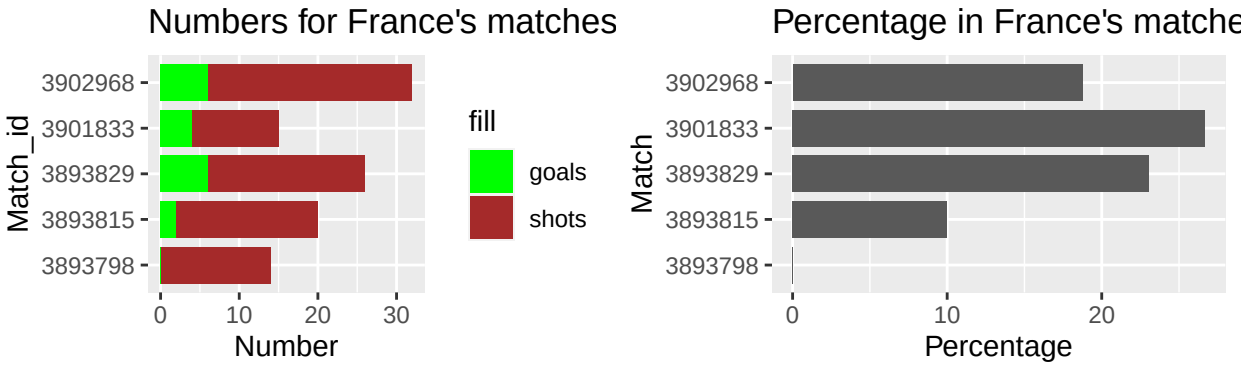


Figure 4: Diagram of the number of shots and goals for each French and percentage

3.2 Analysis of successful shots according to different variables

We first look at the different types of shots in Figure 4. Four different types of shots were differentiated in the data set : “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 made by Ireland’s Katie McCabe, and went in. This is something to keep in mind, as it is sure to skew our future models. Other labels are self-explanatory.

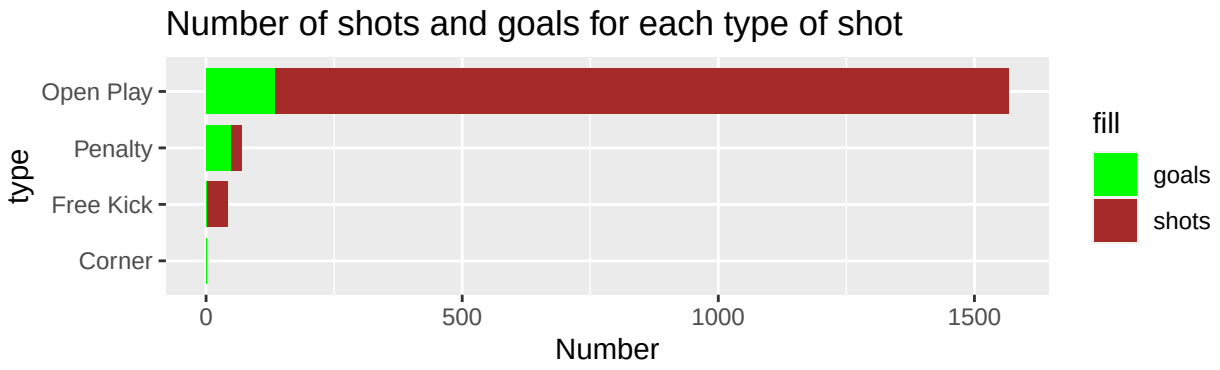


Figure 5: Diagram of the number of shots and goals for each type of shot

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

Figure 5 shows that data set 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 portion of these shots went in.

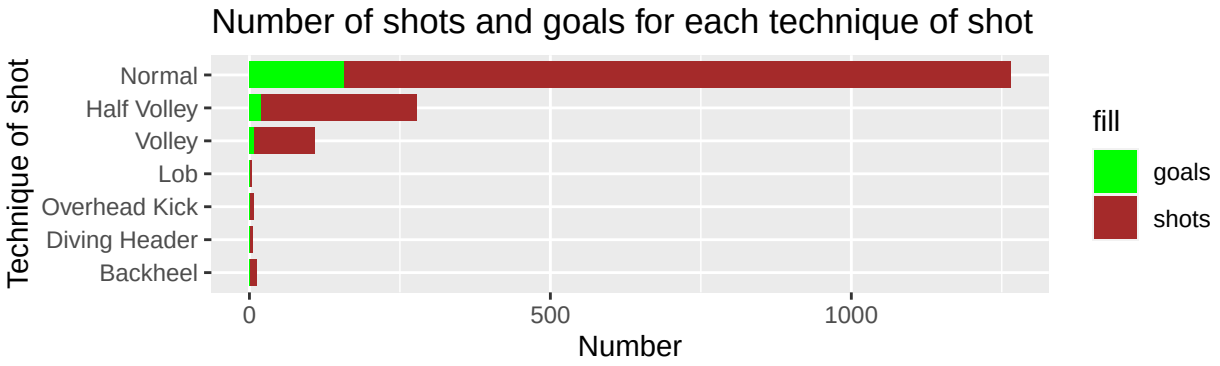


Figure 6: Diagram of the number of shots and goals for each technique of shot

Similarly, we visualize in Figure 6 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.

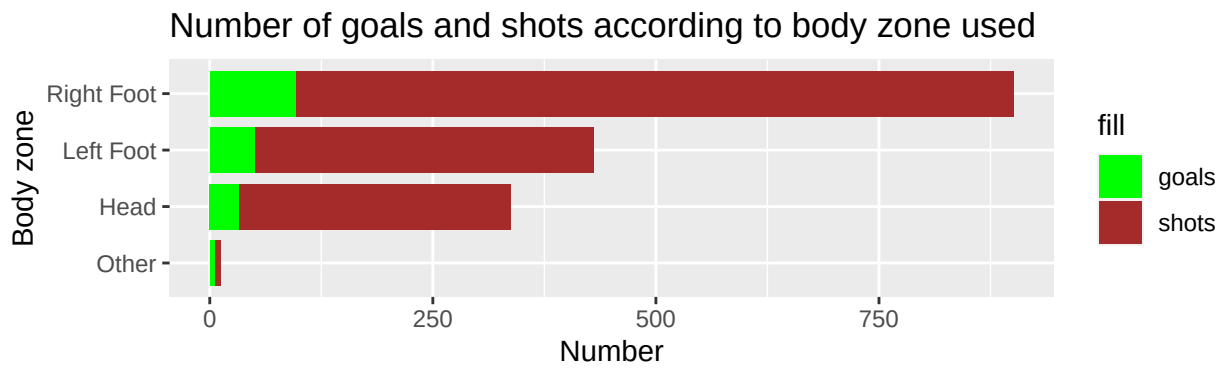


Figure 7: Diagram of the number of goals and shots according to body zone used

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 totals shots during the events.

3.3.1 Players scoring ratio

Firstly, we looked at players, this may lead to find outliers so players which have performed a lot of shots but only few goals, and the opposite players which attempted only few shots but scored on a lot of them. This is a very basic way to start taking an interest in performance of players during this Word Cup.

Figure 8 shows a visualization of the goal ratio on the total shots performed by players which have performed at least 1 shots during the World Cup.

Goal ratio on the number of totals shots for players

Some information on this graph: the color of the dots represents, 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 players have shot, the lower their ratio. Even so, the players on the best teams, those who finished in the best positions, often have a higher ratio than those on teams who lost earlier in the tournament. We also notice that the average (the 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 and 4 goals per match. Now we miss an important information on this graphics. Indeed we can see that the most players attempted shots the less their ratio of successful shots is high, 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 for the goals and under the pressure of defenders. It's something we already saw in the Figure 5, 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 shhoter and the goal position, the fact that the striker is under pressure or not, with

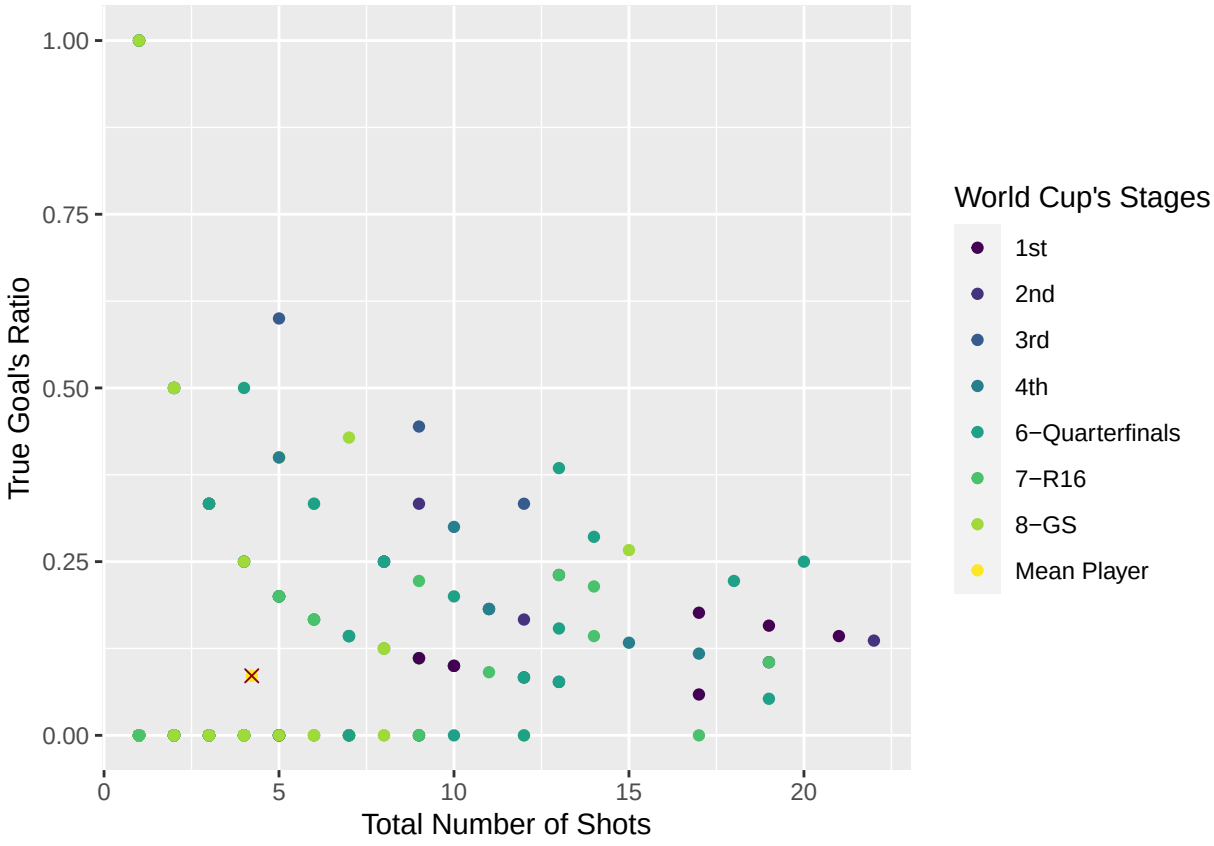


Figure 8: Goal ratio on the number of totals shots for players

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 perfect 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.

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

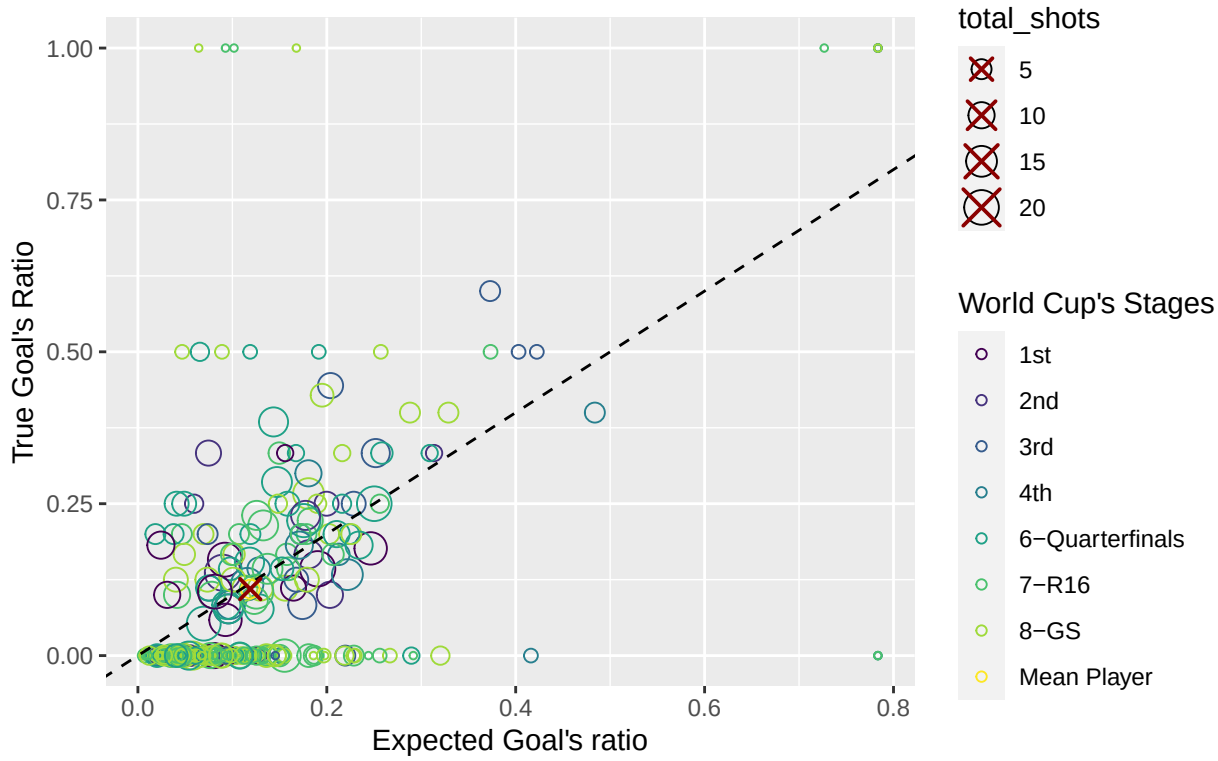


Figure 9: Goal ratio on the Expected Goal ratio (Statsbomb model) for players

This graph compares the number of goals predicted by Statsbomb with the number of goals actually scored. It shows whether players scored more or fewer goals than predicted by the Statsbomb model. We find the same biases as in the graph above: players who scored on their only shot are obviously better performers, but this is no guarantee of performance, as we can't assess the notion of regularity. The same applies to all players represented by the smaller circles. As none of the axes in this graph display the notion of number of shots taken, we are obliged to add this notion by modifying the size of the circle representing the player. This reinforces the caution of interpretation in relation to the number of matches played and how often they had shot during the game.

Now if we interpret this graph, we can see from the colors that the majority of players who have reached the quarter-finals or beyond are still around the dotted line that symbolizes player performance. This line, with the equation y=x, symbolizes the expected performance of the players. Thus, we can deduce that players from teams that made it to the quarter-finals and beyond underperformed less often and even over-performed less, since they took more shots and so their data is less biased by a successful difficult shot.

3.3.2 Teams scoring ratio

Secondly, we decided to look at teams, we are running the same analysis as for the players but this could lead to other analysis. Indeed it's reducing the number of outliers, because we are conduction an analysis on all team players that have shot at least once during the tournament. Indeed if a midfielder player has shot only once and scored and a striker has shot 10 times for only 2 goals this bias of the midfielder has less impact on her team. So, with this team analysis we could say overall which team underperform inter of goals, and try to find if it's correlated to the team who lost earliest during the competition.

Figure 10 shows a visualization of the goal ratio on the total shots performed by teams which have performed at least 1 shots during the World Cup.

Goal ratio on the number of total shots for teams

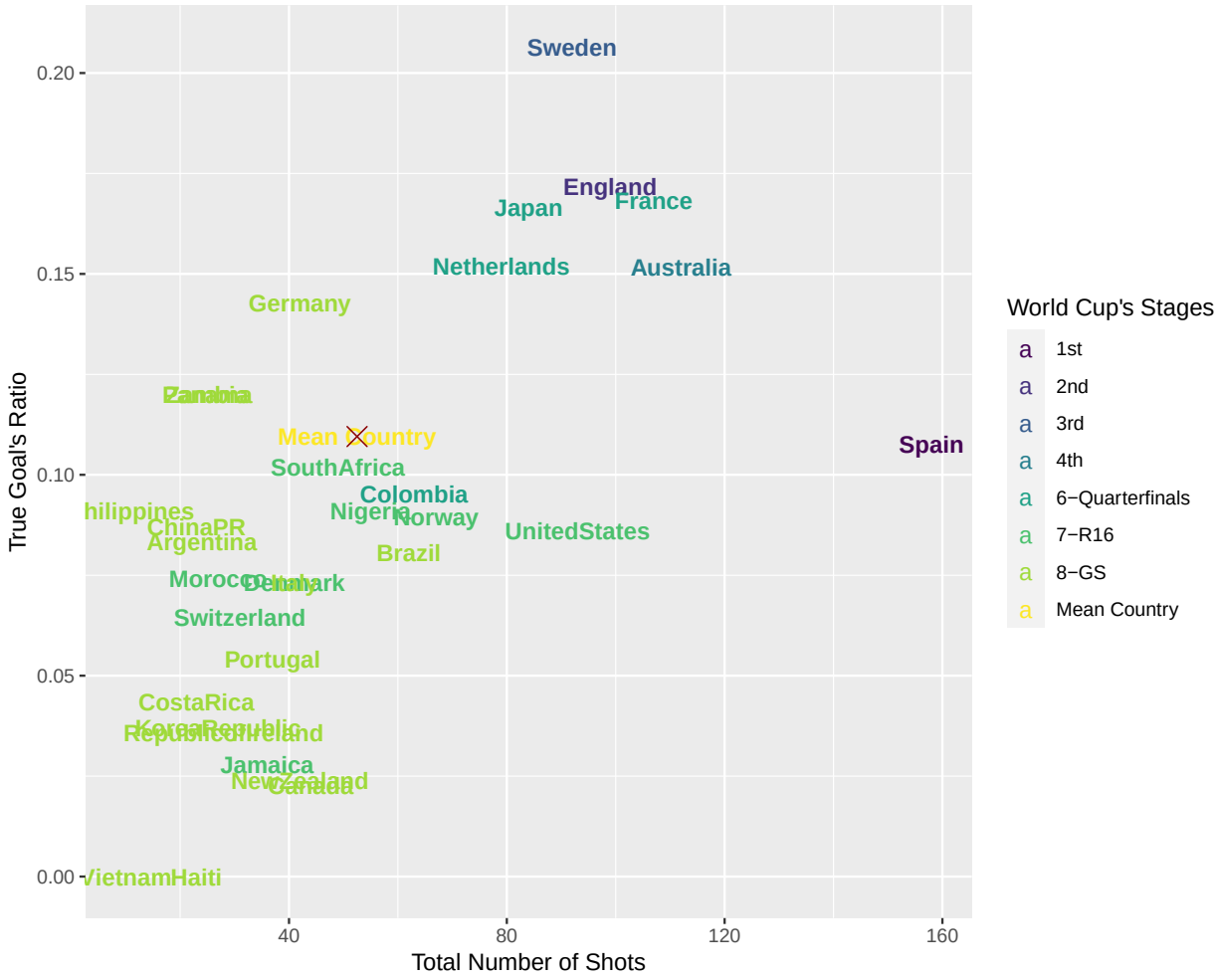


Figure 10: Goal ratio on the number of total shots for teams

In this graph, countries are represented by dots with their name, and the color always indicates the stage of the competition they have reached. There's a clear separation between the bottom left of the image, representing countries with few shots and few goals, with teams that didn't make it beyond the quarter-finals, and the top right with the 8 best teams, excluding Colombia.

We can also see that Spain, winner of the tournament, shot much more than its rivals but scored less, so

this could reflect a strong domination during the matches, but less good finishing. It would be interesting to see if the difficulty of the shots could explain the lower accuracy, or if there's another interpretation.

Other outliers were Sweden, who finished 3rd and had the best ratio of successful shots. And Vietnam and Haiti, who didn't score a single goal during the competition.

As for the players, we want to implement the Expected Goal data, so we can implement and analyze the difficulty of shots performed by teams.

This his what the Figure 11 is showing true goal ratio on the Statsbomb xG ratio for every teams during the World Cup.

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

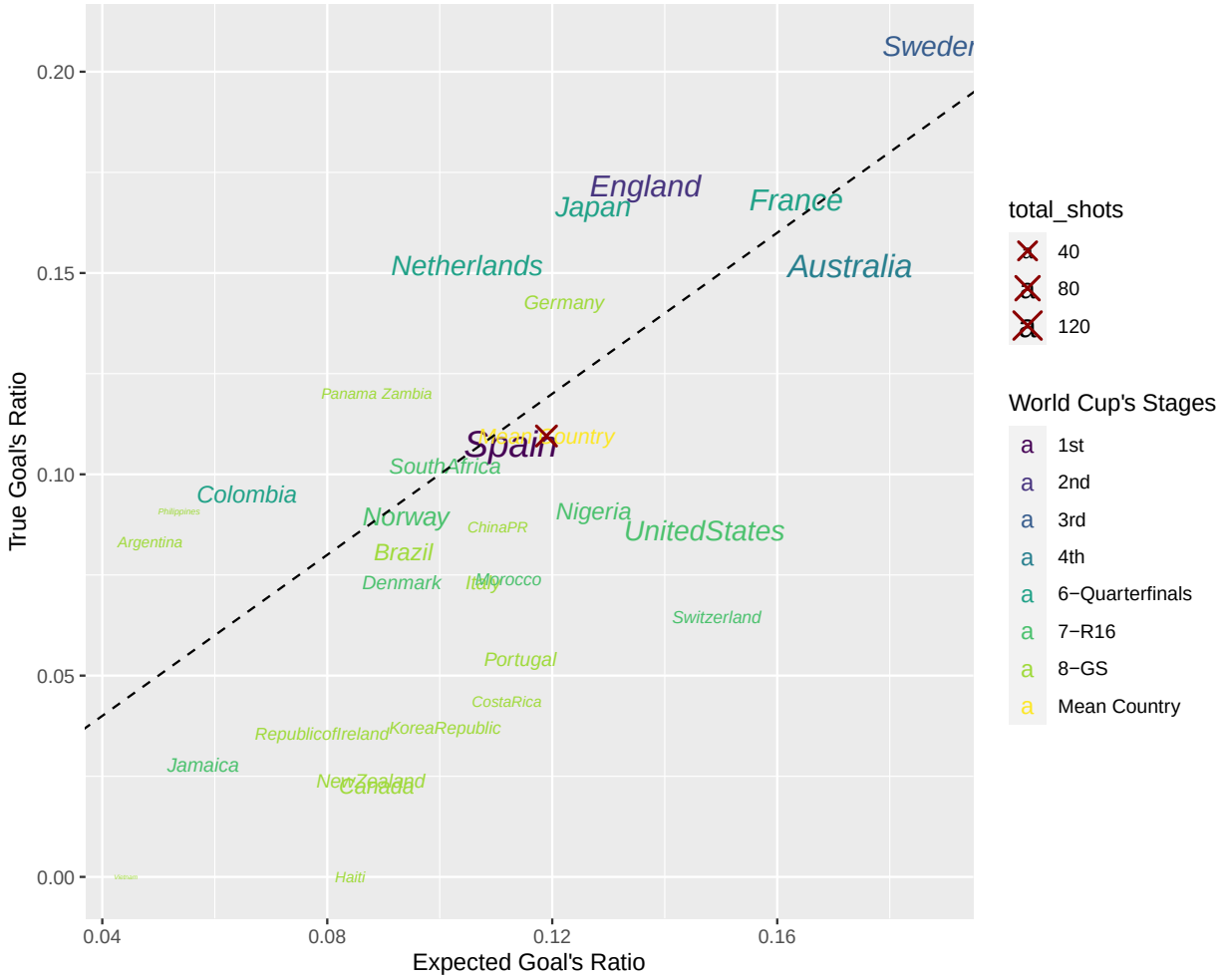


Figure 11: Goal ratio on the Expected Goal ratio (Statsbomb model) for teams

So, on this graph, we can complete our interpretations of the previous graph. Sweden, on the right-hand side of the graph, had the easiest shots on average according to Statsbomb's model, and succeeded as expected, or even a little better, as it lies slightly above the dotted line that marks the performance.

As for Spain, we can see that they had the hardest shots of all the teams that made it through to the quarter-finals, with the exception of Colombia, which explains why they had a greater number of shots with a fairly low ratio compared to their main rivals.

However, it is noticeable that the best teams have higher xGs, which means that their shots are more likely to succeed on average. This could mean that their forwards are putting themselves in better shooting situations, either because they are better or because the other teams' defenders are weaker. Thus, it could mean that the defense of the best teams is very strong compared to that of the teams eliminated earlier in the tournament, so these weaker teams had difficulty scoring. Now that we have compared the true goal ratio to the xG Statsbomb ratio, we were wondering how exactly is one xG model working. So, we decided to create our own xG model.

4 Models analysis

We wanted to create our own xG model. To do that we developed different models, finding the most relevant variables to predict 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 assimilated as a moving point on a rectangle of size 80x120m. Its horizontal movement - that is, going from one goal post to another - is tracked by the variable location.x, while the vertical movement is associated to location.y. We will now add 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.

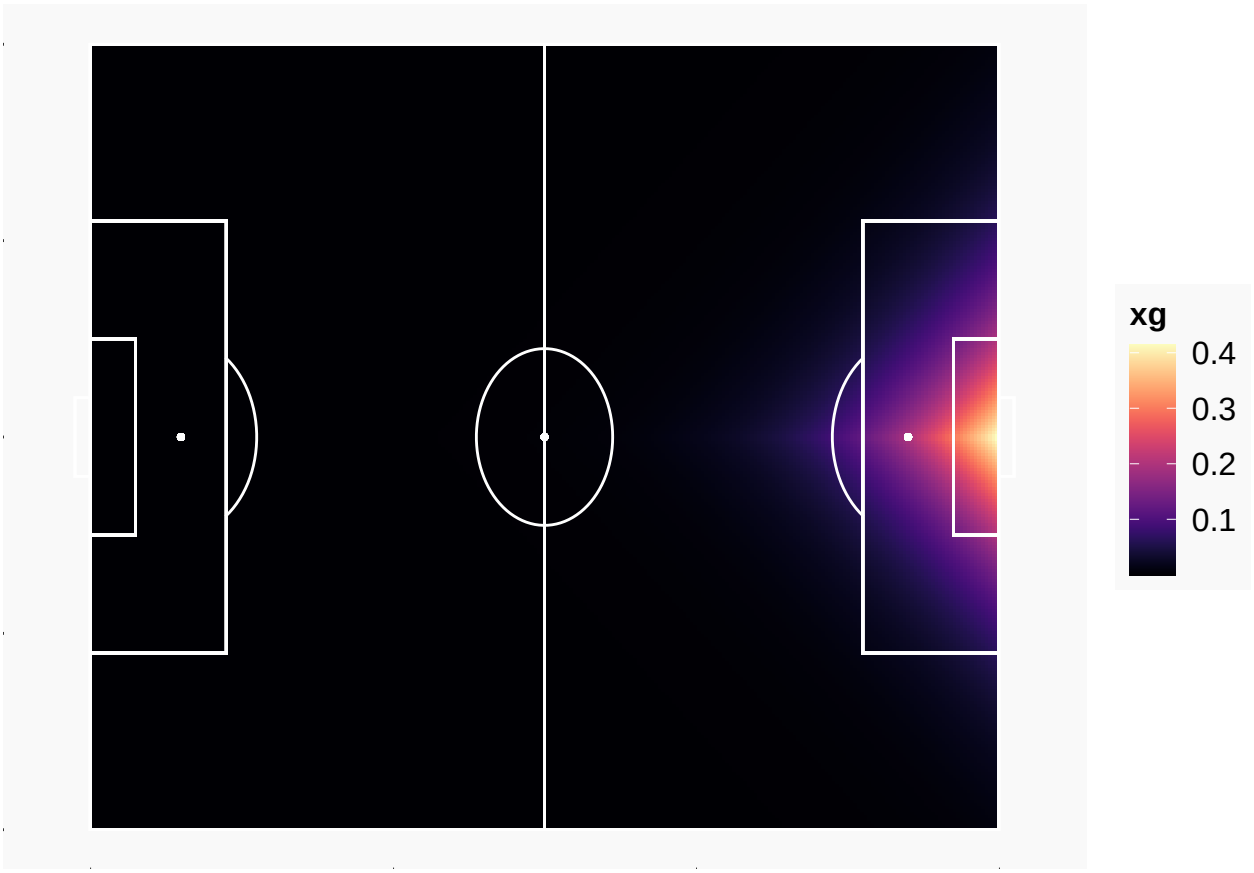
Without this variable, the AIC is even lower, at `866.1274211`.

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

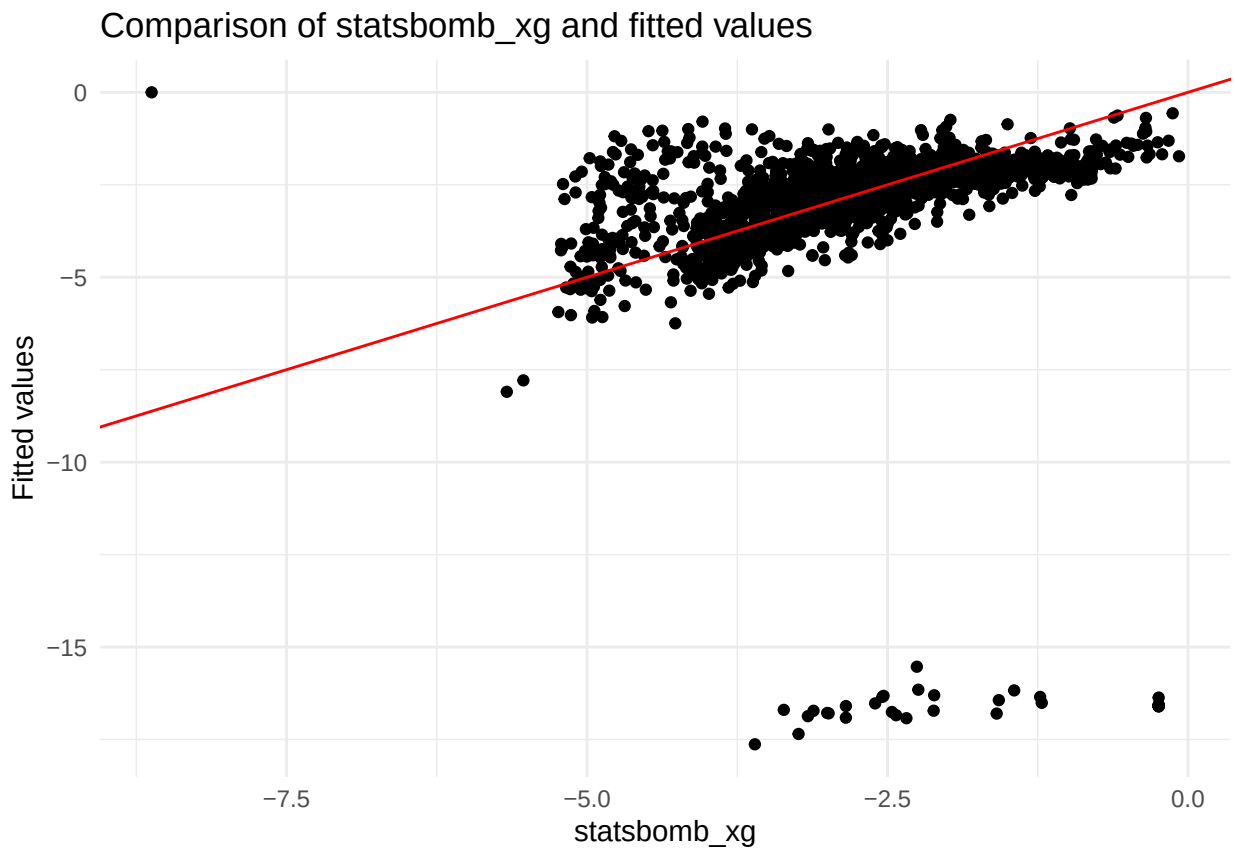
body_part, shot.technique, shot.type, location.x, under_pressure, location.x.GK, location.y.GK.

4.7 Replacing location.y

Seeing as location.y is a very insignificant variable, we now view offense as symmetrical with respect to the axis passing by the center point of the pitch and the penalty spots. We will now refer to location.y as the distance to center, and the new variable will vary from 0 to 40 meters. This hopefully will give more sense to the parameter, as a high distance to center means a player is very off-center.



The above heatmap shows what is to be expected : from a very simple model, only using the location as well as the distance to center, we can see that the model expects a player to have better scoring chances with point-blank shots rather than shots outside the penalty area. This is easily explained by the fact that our data set does not contain many goals made from outside this area.



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

We can see that only one point is being overestimated by our model. It is the goal made from a corner kick mentioned previously. Our model predicts an xG of 1, which perfectly demonstrates the limits of our model. We only worked on this reduced data set, and since only one corner shot was attempted, our model wrongly assumes it is the most efficient shot out there. After transforming the residuals to their logarithmic counterpart, they seem for the most part well-adjusted, with a large majority of them grouping around the desired spot. On the bottom right, we can see some points where our model underestimates the chance of a goal happening. These seem to be goals that were taken from far away, and our model seems to consistently underestimate those. Again, this is most likely a consequence of our small data set. Overall, our model seems to get 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
```

We can see finally that y-positions are useless even for the goalkeeper, only x-positions are significant.

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 the player and the goalkeeper.

5 Analysis of the Model performance

Now that we have our model composed of these 5 variables, we asked ourselves if this type of model is better than the Statsbomb xG one, and overall we wanted to test the perfomance of our model ! So as we previsouly did in the "Descriptive Analysis" part of our report we are conducting a comparison on the ratio of goals. Before, we did it between data we were given, but now we will predict the xG of every shot with our best model and then compare these xG value to the Statsbomb one on graphics.

Figure **??** shows a visualization of these results.

```
## [1] 398
```

```
## [1] 398
```

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

We notice that if we display the ratio of the xG we predicted to those from Statsbomb, the results vary very little, which means our model is quite close to that of Statsbomb! However, to accurately compare our model with Statsbomb's, we should display the xG for each shot.

This is what the 13 below shows:

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

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

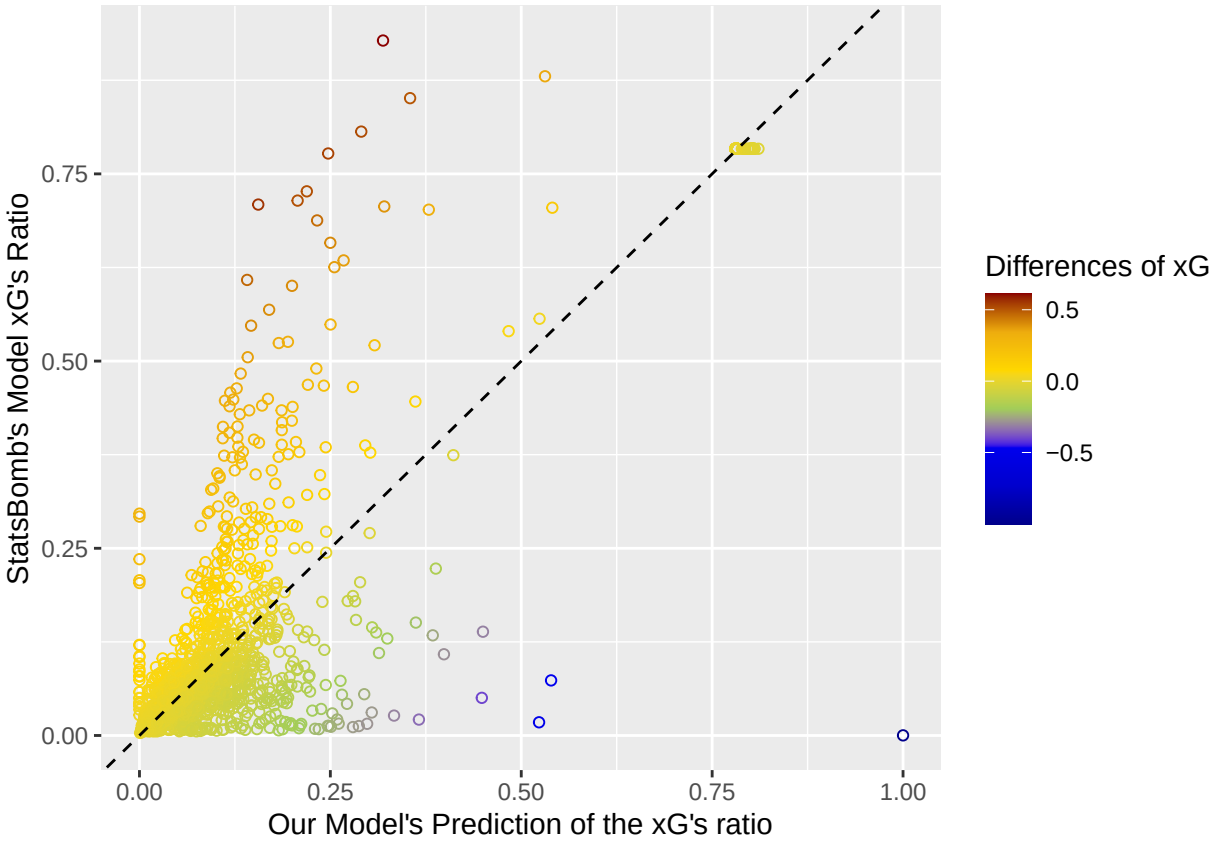


Figure 13: StatsBomb xG on our xG based on bestmod for every shots in the World Cup

We can see that for the vast majority of shots, our models are similar! However, we do notice some outliers, which can be explained by displaying, for example, the type of shot, whether it is a free kick, an open play shot, a penalty, a corner, etc. This is shown in 14 below:

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

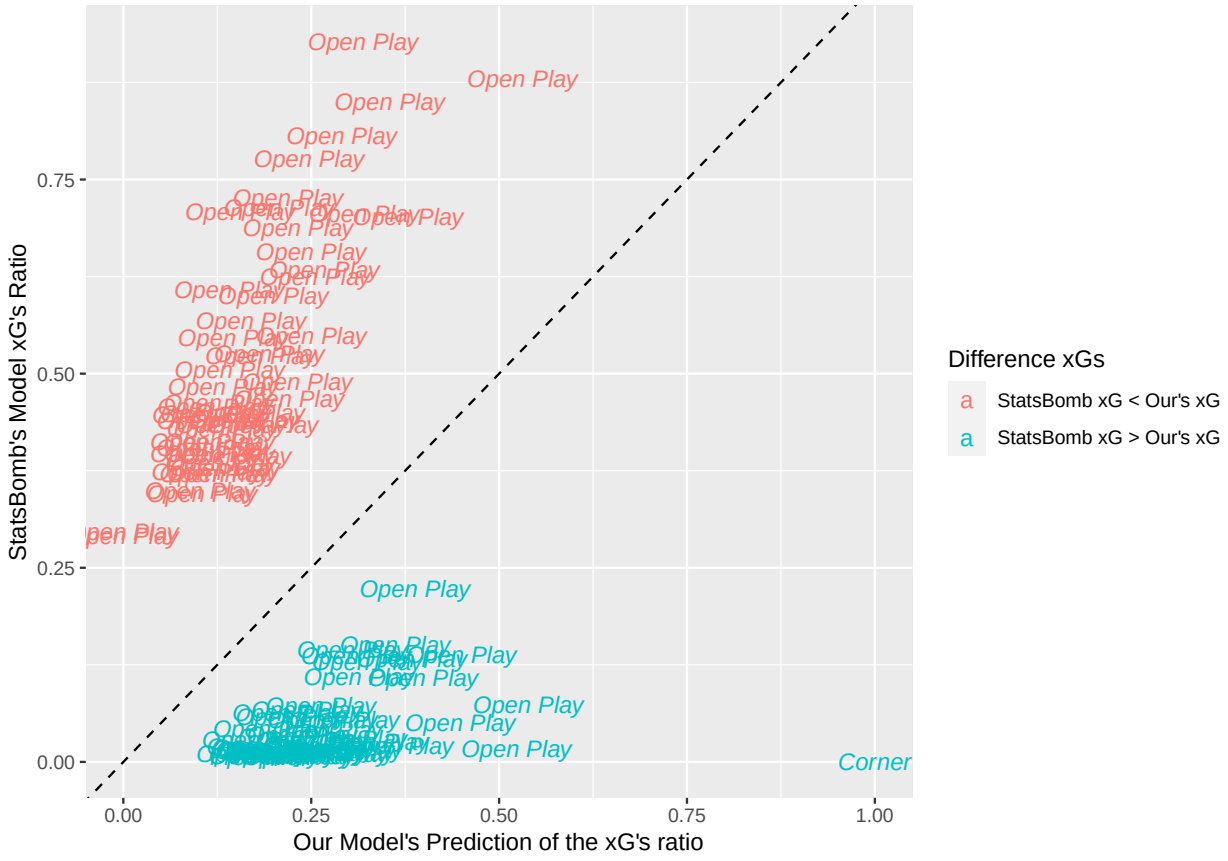


Figure 14: StatsBomb xG on our xG based on bestmod and the type of shots for the 100 shots with the higher xG difference

We can see a clear example of bias in our model: the 'Corner' at the bottom right of the graph, where our model predicts an xG of 1 and Statsbomb predicts an xG of almost 0. In fact, during the competition, there was an in-swinging corner that was attempted and succeeded, so our model, having been trained only on this competition, now classifies all in-swinging corners as shots with a goal probability of more than 0.99. On the other hand, Statsbomb's model, likely trained on many more competitions, has enough historical data to see that this rare type of shot very seldom results in a goal, thus it assigns a probability close to 0.

For the other points, it is more complicated to explain. Indeed, they are almost all open play shots, so many other parameters come into play. Our model only considers five: the type of shot (penalty, free kick, open play, etc.), the shot technique (normal, lob, volley, bicycle kick, etc.), the part of the body used for the shot, and the distance to the goal of the player and the goalkeeper. Therefore, we are probably missing some information, such as the shot angle or whether the player was under pressure, or we need more data from women's football tournaments to refine our model and smooth out the peculiarities of this competition and the impact of anomalies, like the in-swinging corner, by analyzing more than 64 matches.

6 Another Model : Expected Pass

Having analyzed the Expected Goal (xG) values provided by StatsBomb and developed our own xG prediction model, we questioned whether StatsBomb offered a similar indicator for passes—an Expected Pass (xPass) metric, which predicts the probability of a pass being successful based on various parameters. Surprisingly, they do not have such an indicator. Consequently, we decided to implement one using the same methodology employed for our xG model. We chose to create this model for passes because a football team's performance is strongly correlated with goals, as goals ultimately determine the outcome of matches. However, when observing football matches, it is evident that some teams struggle to create scoring opportunities and goal situations due to their inability to execute passes between the lines, through passes, player combinations, technical skills, or dribbles. We have noted that teams reaching the quarterfinals generally exhibited better xG and True Goal Ratios compared to others, with some exceptions like Colombia. This model allows us to analyze another dimension of the complex sport of football.

For instance, did Colombia excel in the area of play construction compared to their opponents, thereby enabling them to reach the quarterfinals despite having similar xG and True Goal Ratios? Moreover, do we observe the same trend regarding goal-scoring capabilities in the top teams? These questions prompted us to develop this model for analyzing passes.

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

This graph illustrates the ratio of successful passes to the total number of passes. Once again, a striking difference is observed: players from countries that reached the quarterfinals are positioned at the top of the graph, with a successful pass ratio exceeding 60%, whereas teams eliminated in the round of 16 and group stages exhibit considerably lower pass ratios. It is noteworthy that within these less successful teams, there are still a significant number of players with pass ratios as high as those in the top-performing teams. This suggests potential disparities within these countries, likely indicating a higher level of heterogeneity 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

We examined the statistics by country and observed a consistent trend with a notable distinction for Spain, which 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. Combining this with shooting data, where Spain had many more shots than its competitors, it reaffirms that Spain controlled their matches more effectively, with higher possession and more opportunities. This observation likely extends to England, the other finalist team. For Sweden and Australia, the third and fourth place teams respectively, we can infer that their passes were more risky, 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 (xPass) indicator, particularly the difficulty of the passes. However, because Statsbomb did not provide a xPass indicator inside the data base, we will not be able to compare our models to the ones used in practice.

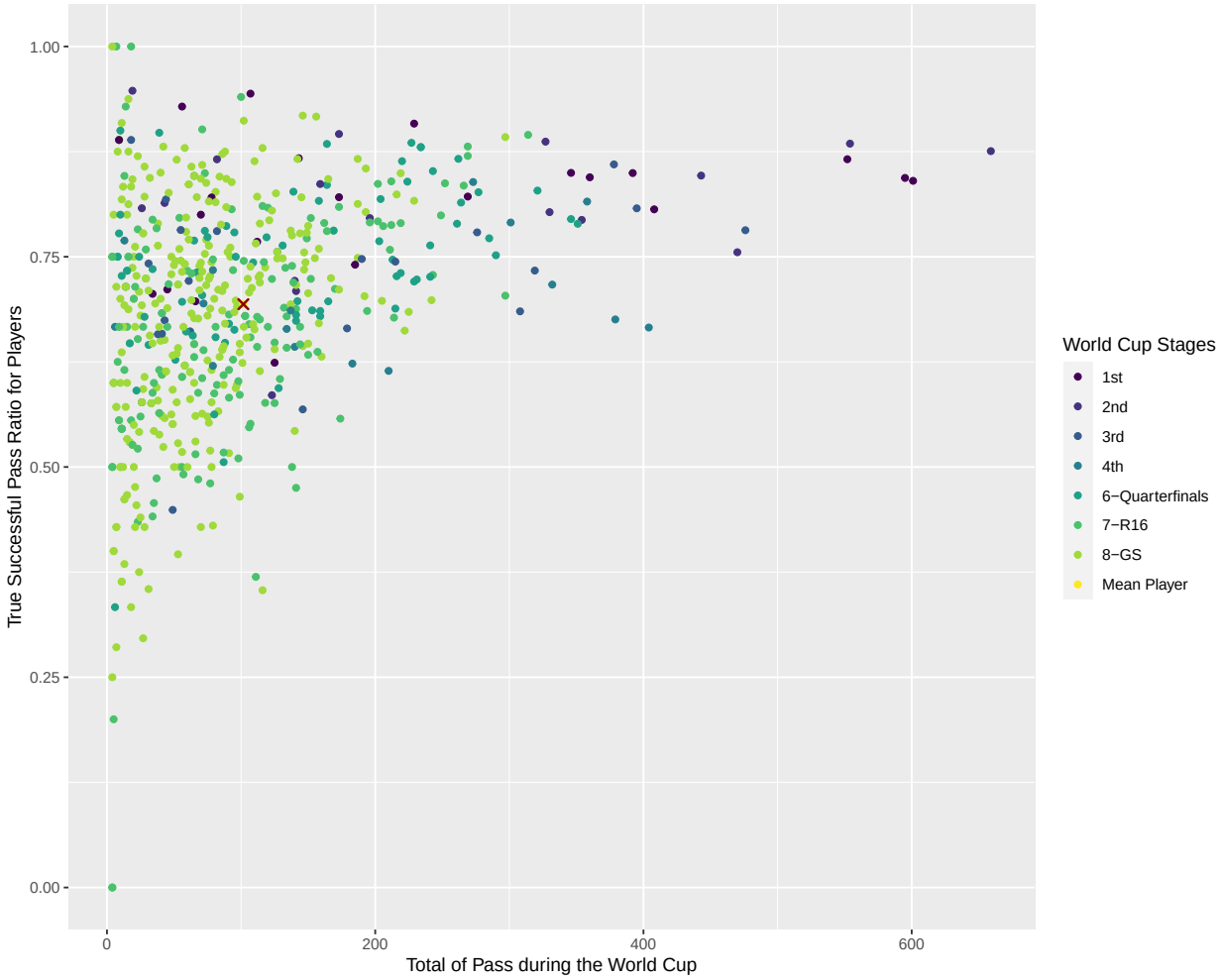


Figure 15: True Pass ratio on the Total Number of Pass for every player with 5 pass or more during the WC

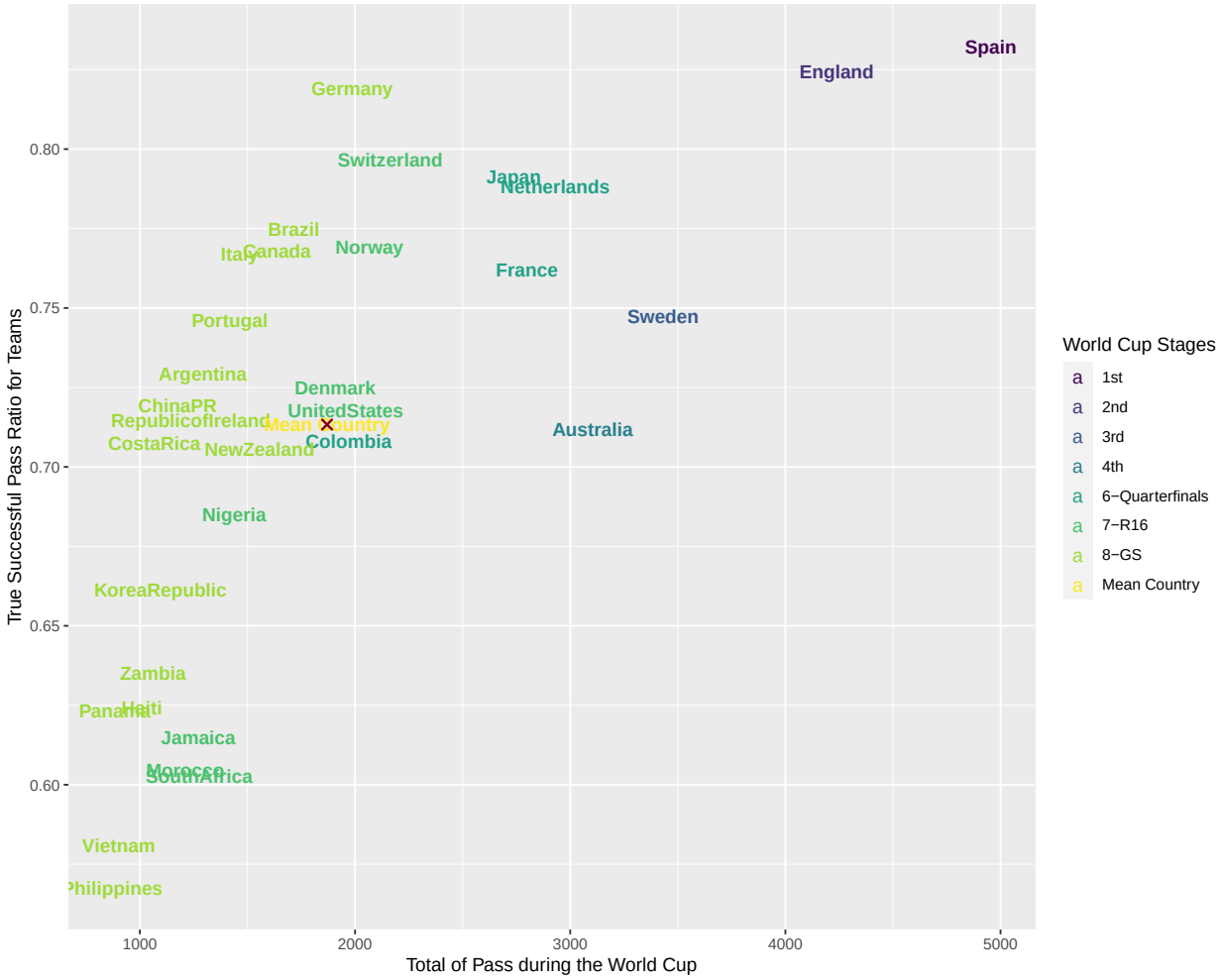


Figure 16: True Pass ratio on the Total Number of Pass for every teams during the WC

6.2 Implementing the Pass model we created

6.2.1 Comparaison xP to True Pass Ratio visualization

6.2.1.1 For hard shots

6.2.1.2 For easy shots

6.2.1.3 For all shots

###Comparaison xP to True Pass Ratio visualization - Pays par Pays

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 or Colombia, and managed to explain some of their success. Using basic logistic regression, and with only a very limited data set, we implemented our own expected goal metric. Through many variable selection steps, we were able to come very close to the xG calculated by StatsBomb. 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 as 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