Projet_Women_FIFA_WC23_Analysis

2024-05-16

Contents

1 Introduction

In today's world, sports are at the center of global culture. To continue to excel, 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 untapped in terms of data exploration. 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.

2 Descriptive data analysis

We begin by interpreting the elements of the data set.
It is made up of multiple observations with different variables.

2.1 Analysis of successful shots according to country

First, we look at the number of goals and shots in all matches for each team.
Figure 1 shows a visualization of these results.

It would be interesting to make this graph on the average number of goals and shots, as some teams have more games than others, distorting the results a little.

Figure 2 shows a visualization of the percentage of shots leading to a goal.

We now would like to focus on France team.
Figure 3 shows the results.

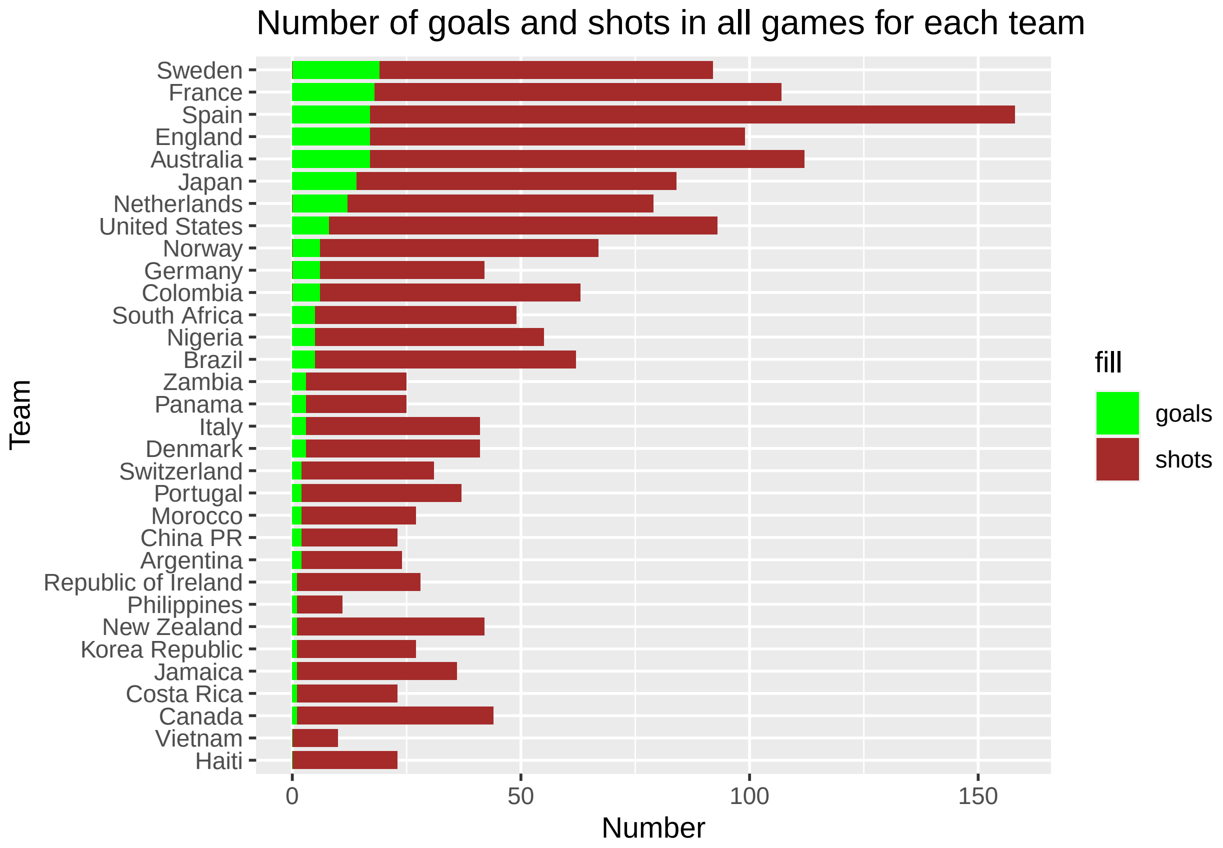


Figure 1: Diagram of the number of goals and shots in all matches for each team

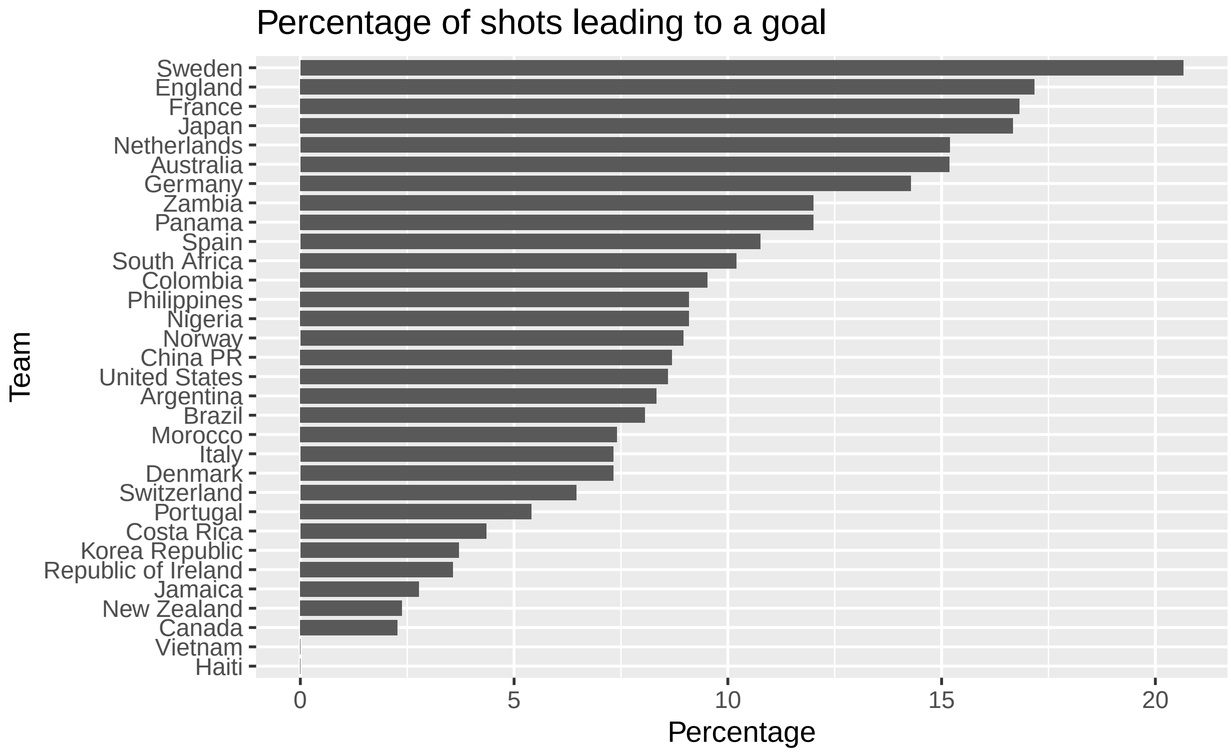


Figure 2: Diagram of the percentage of shots leading to a goal

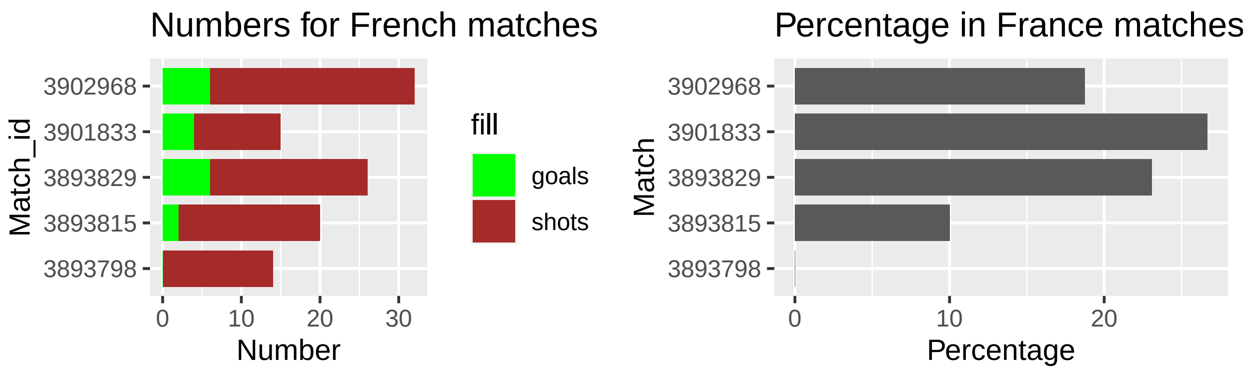


Figure 3: Diagram of the number of shots and goals for each French and percentage

2.2 Analysis of successful shots according to different variables

We first look at the type of shots.

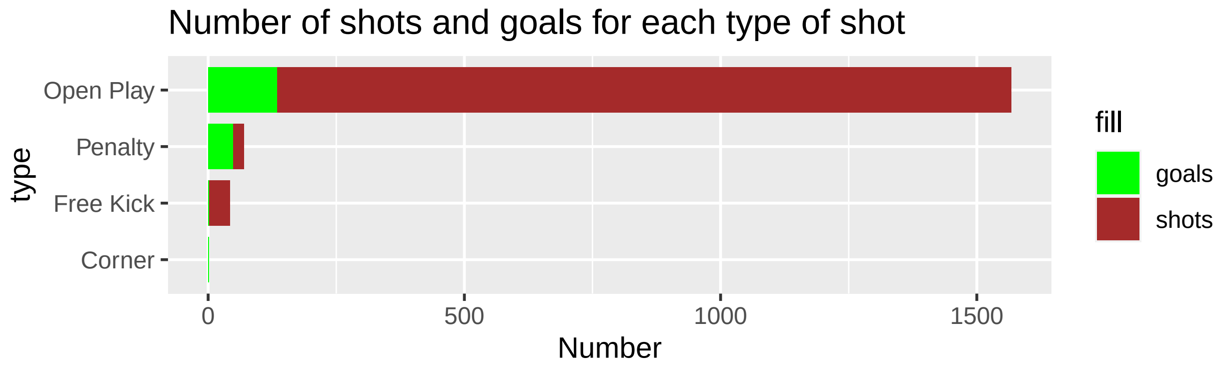


Figure 4: Diagram of the number of shots and goals for each type of shot

We know would like to see if the technique of shot is significant.

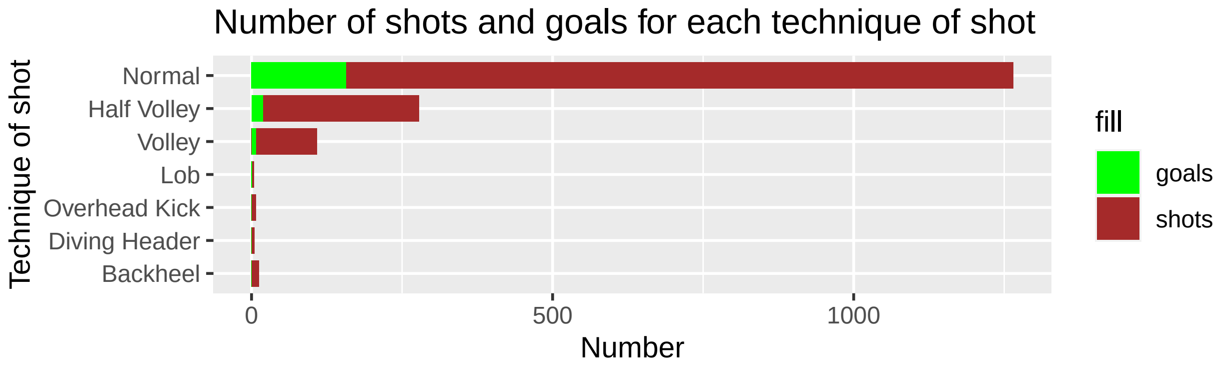
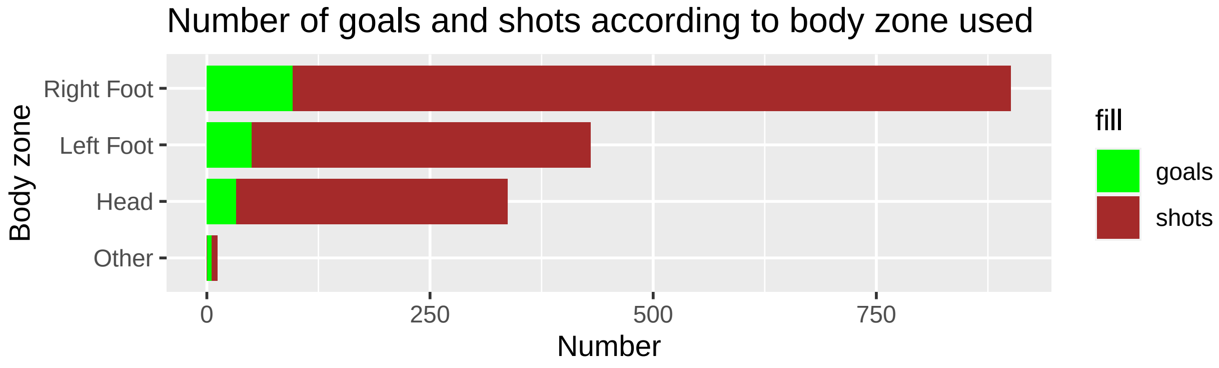


Figure 5: Diagram of the number of shots and goals for each technique of shot

Finally, is the variable body_zone_used relevant ?



3 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.

3.1 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 .

3.2 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.

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

To do that we do the same logistical model as the first one but with the position added : location.x and location.y .

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.

3.3.1 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, which is minimum for model 3 without interactions.

3.3.2 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.

3.4 Model 4 : adding the under_pressure variable

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

3.4.1 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.

3.4.2 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`.

3.5 Model 5 : adding the position of the goalkeeper

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

3.5.1 Does the position of the guard 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)   23.74030    4.90676   4.838 1.31e-06 ***
## location.x.GK -0.22185    0.04174  -5.315 1.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 934.31  on 1620  degrees of freedom
## Residual deviance: 910.15  on 1619  degrees of freedom
##   (59 observations effacées parce que manquantes)
## AIC: 914.15
##
## Number of Fisher Scoring iterations: 5
## [1] 0.02586437
```

Significant effect of goal position in x because both p_{values} are lower than 0.5.
The AIC value is low, equals to `914.1462437`.

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 `937.1556025`.

3.5.2 Complete model

For the model with all the preceding variables and without interaction, we find a very low AIC=`845.3219716`. We can conclude that this model is really good.

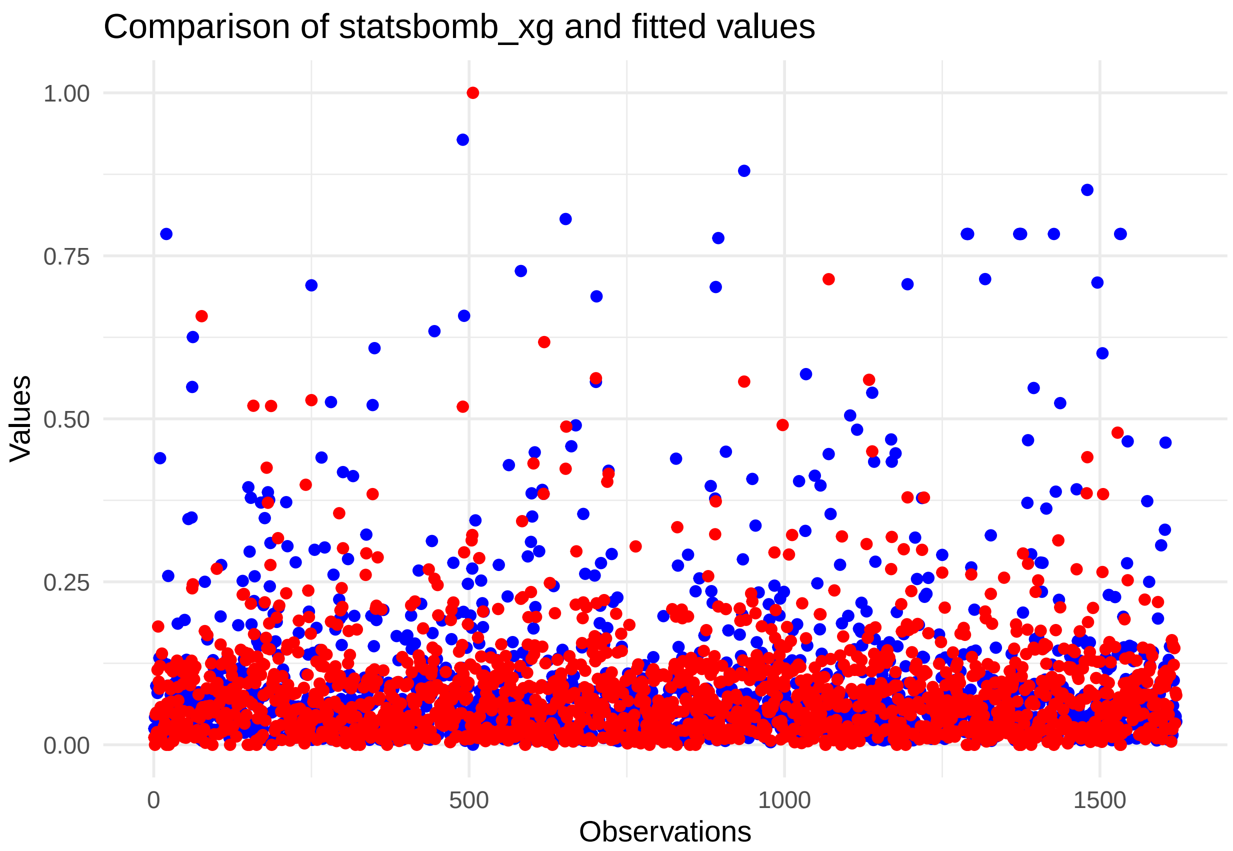
3.6 Last model : keeping all significant variables and removing location.y

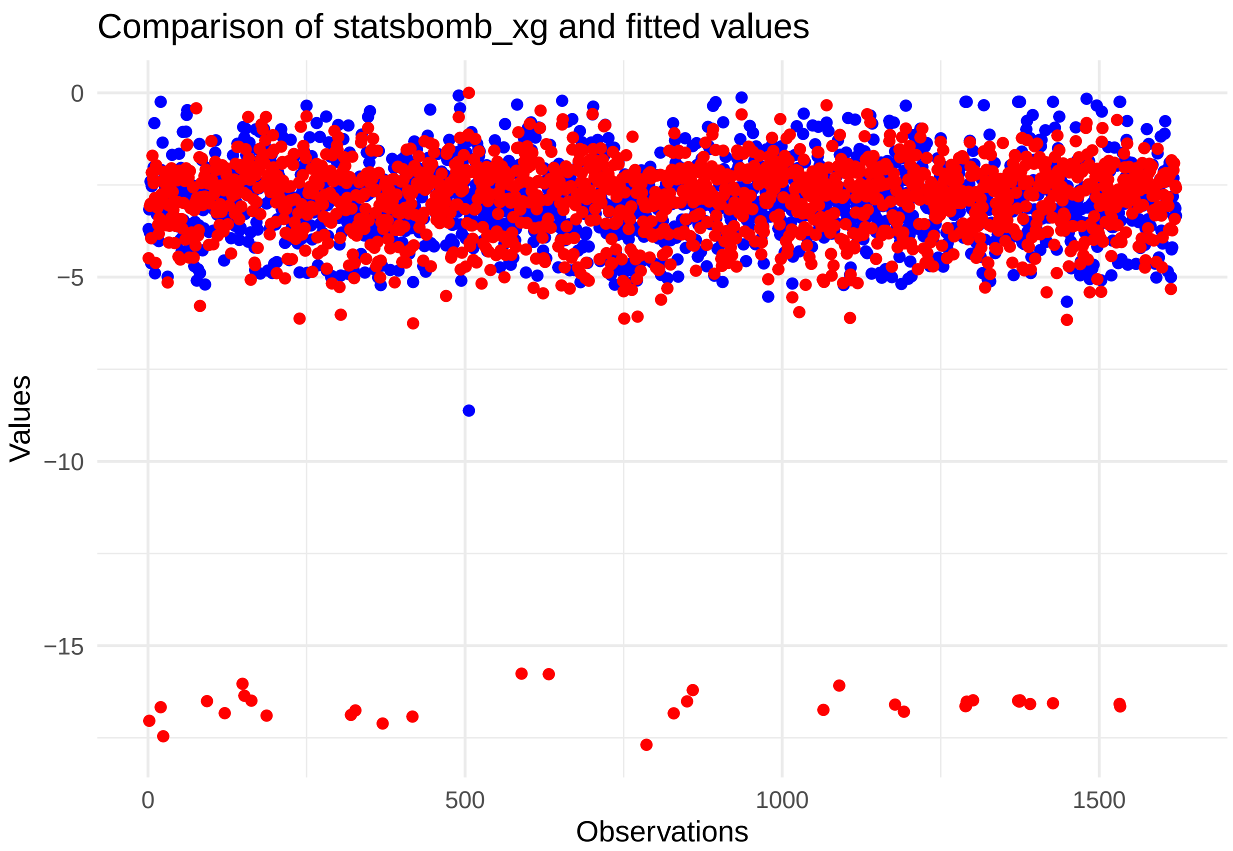
Since we found that location.y is noyt significant, we can remove it from the model.

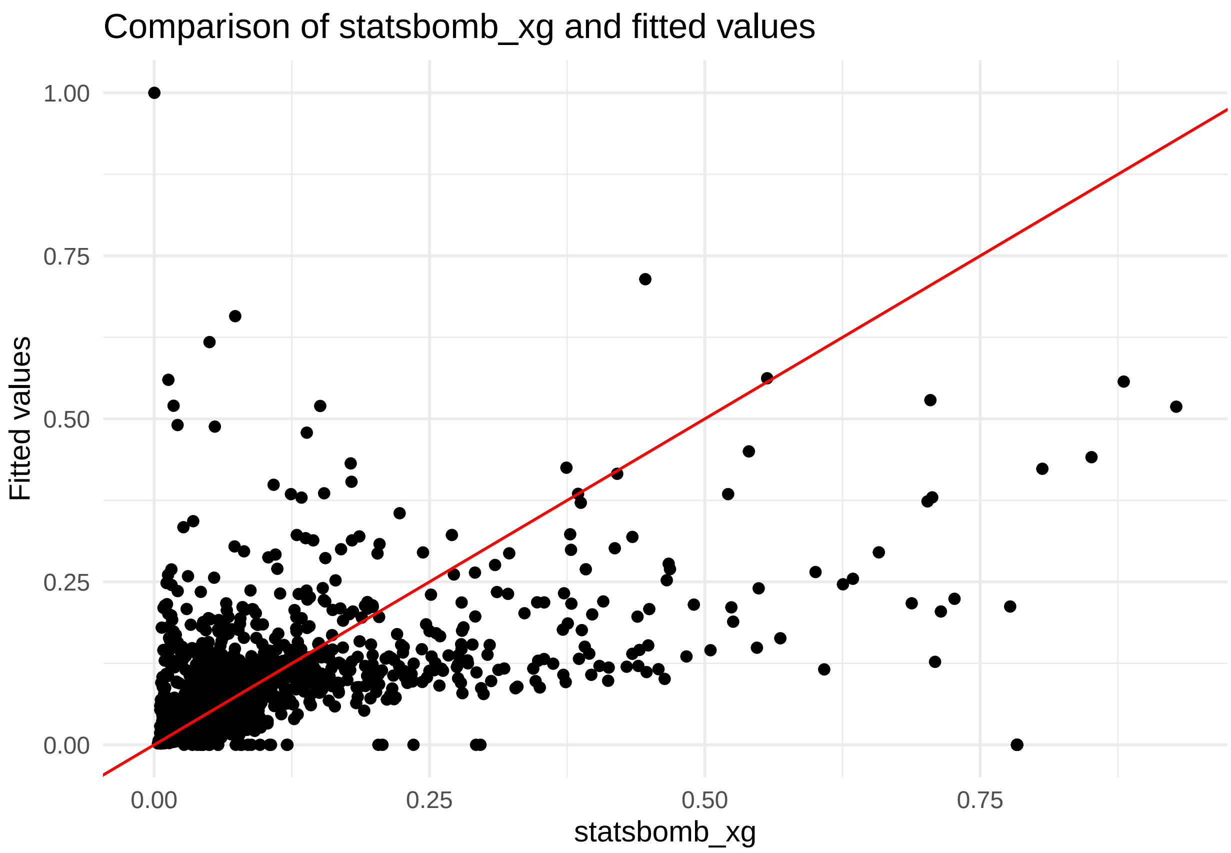
Without this variable, the AIC is even lower, at `844.5352323`.

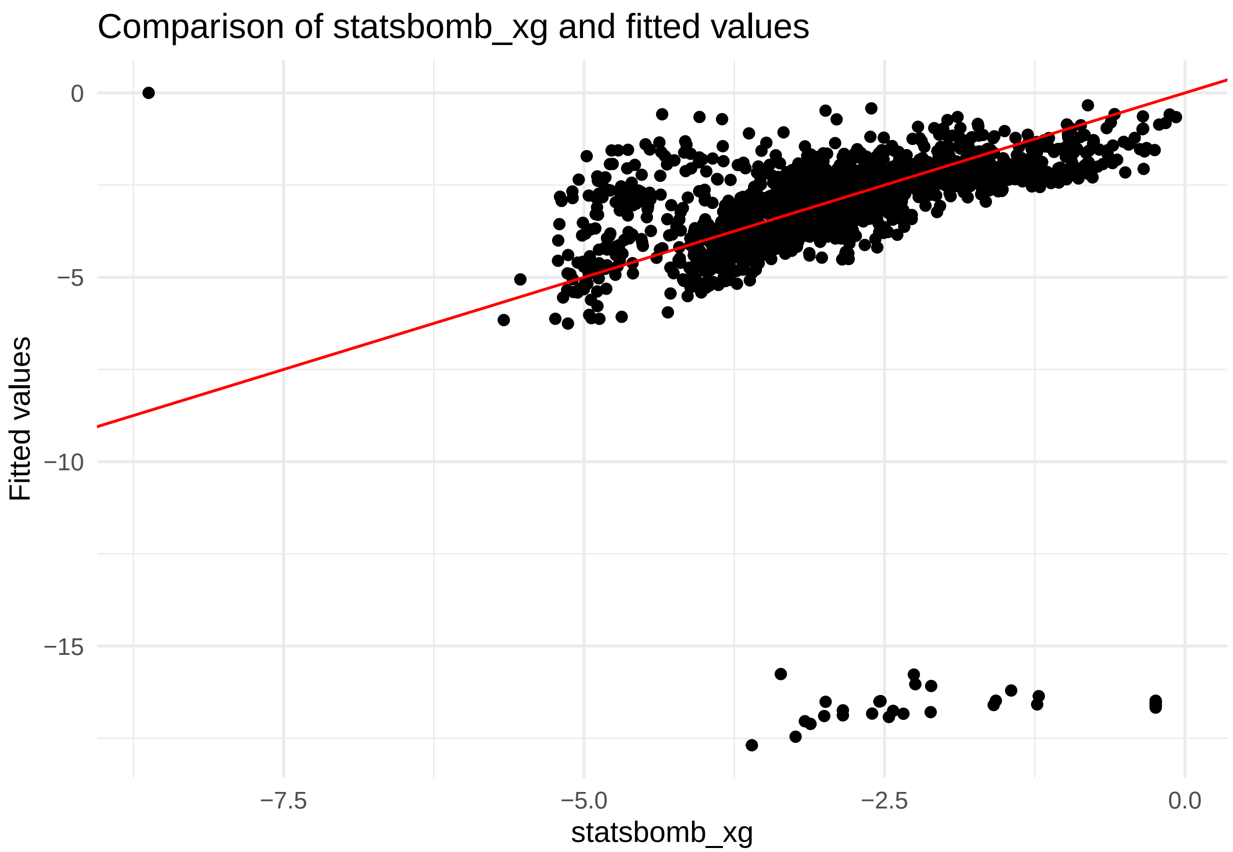
We can conclude that we have found our best model for now and it's composed of the variables :

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


Comparison of statsbomb_xg and fitted values
Values
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
Observations


Comparison of statsbomb_xg and fitted values
0
−5
−10
−15
Values
0
500
1000
1500
Observations


Comparison of statsbomb_xg and fitted values
Fitted values
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
statsbomb_xg



We have a lot of values close to 0, so we do the log to make things clearer.

In log: there's a point (an observation) where we've overestimated the chance of scoring. There are a few points at the bottom right where, on the contrary, we've underestimated the probability, but overall we've got a good prediction based on the Xg of bomb stats.

3.7 Finding our best model with the AIC criterion

```
##
## Call:  glm(formula = shot.outcome.name ~ shot.body_part.name + shot.technique.name +
##     location.x + location.x.GK, family = binomial(link = "logit"),
##     data = df_model_6)
##
## Coefficients:
##                      (Intercept)      shot.body_part.nameLeft Foot
##                          -2.8131                            0.6209
##         shot.body_part.nameOther     shot.body_part.nameRight Foot
##                           1.6863                            0.4945
## shot.technique.nameDiving Header    shot.technique.nameHalf Volley
##                           1.0349                           14.4967
##           shot.technique.nameLob          shot.technique.nameNormal
##                          16.2501                           15.1470
## shot.technique.nameOverhead Kick          shot.technique.nameVolley
##                           0.4921                           14.4811
##                       location.x                      location.x.GK
##                           0.1331                           -0.2488
```

```
##
## Degrees of Freedom: 1620 Total (i.e. Null);  1609 Residual
##   (59 observations effacées parce que manquantes)
## Null Deviance:       934.3
## Residual Deviance: 814.7     AIC: 838.7
```

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

Our best model is composed of 4 variables : The technique of shot, the body part used and the positions in x for the player and the goalkeeper.