Projet_Women_FIFA_WC23_Analysis

2024-05-15

Contents

1 Introduction

Blalblablabla

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.

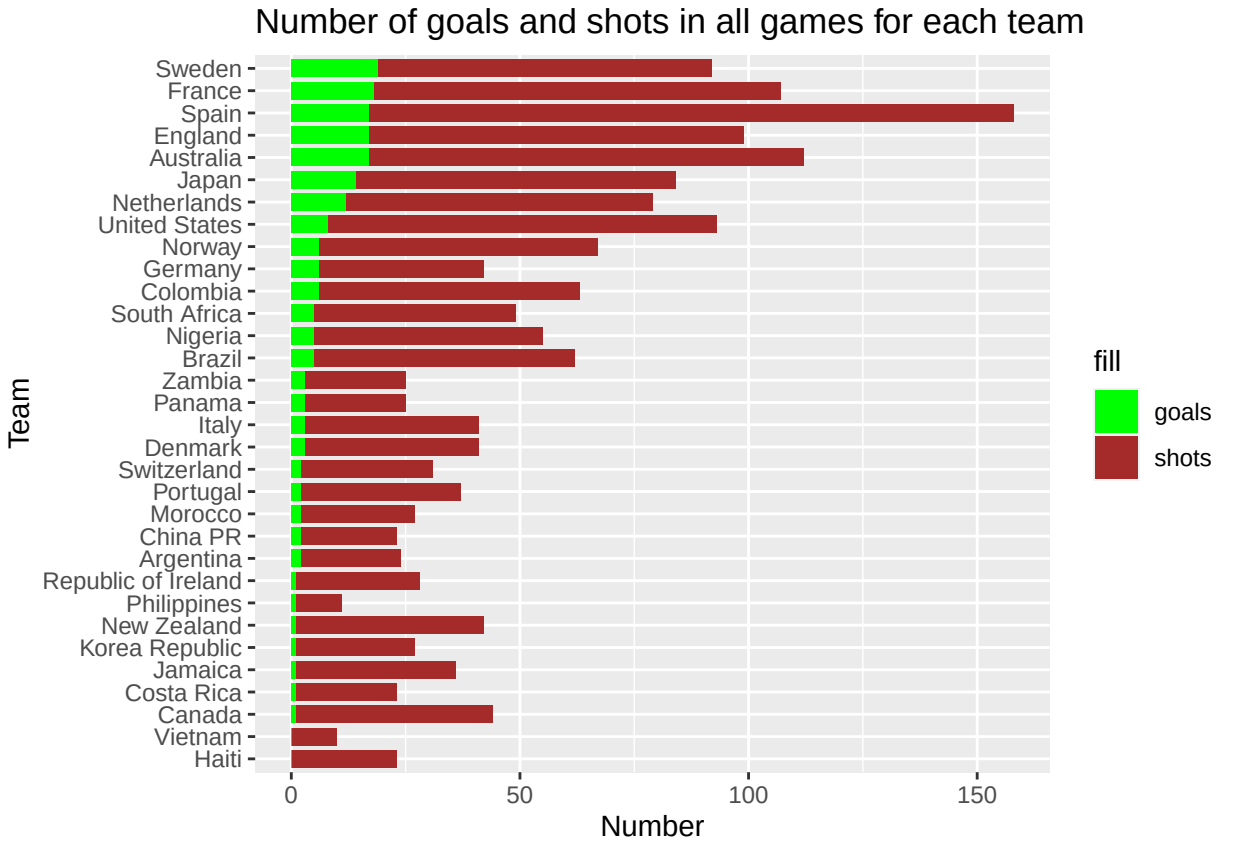


Figure 1: Diagram of the number of goals and shots in all matches for each team

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.

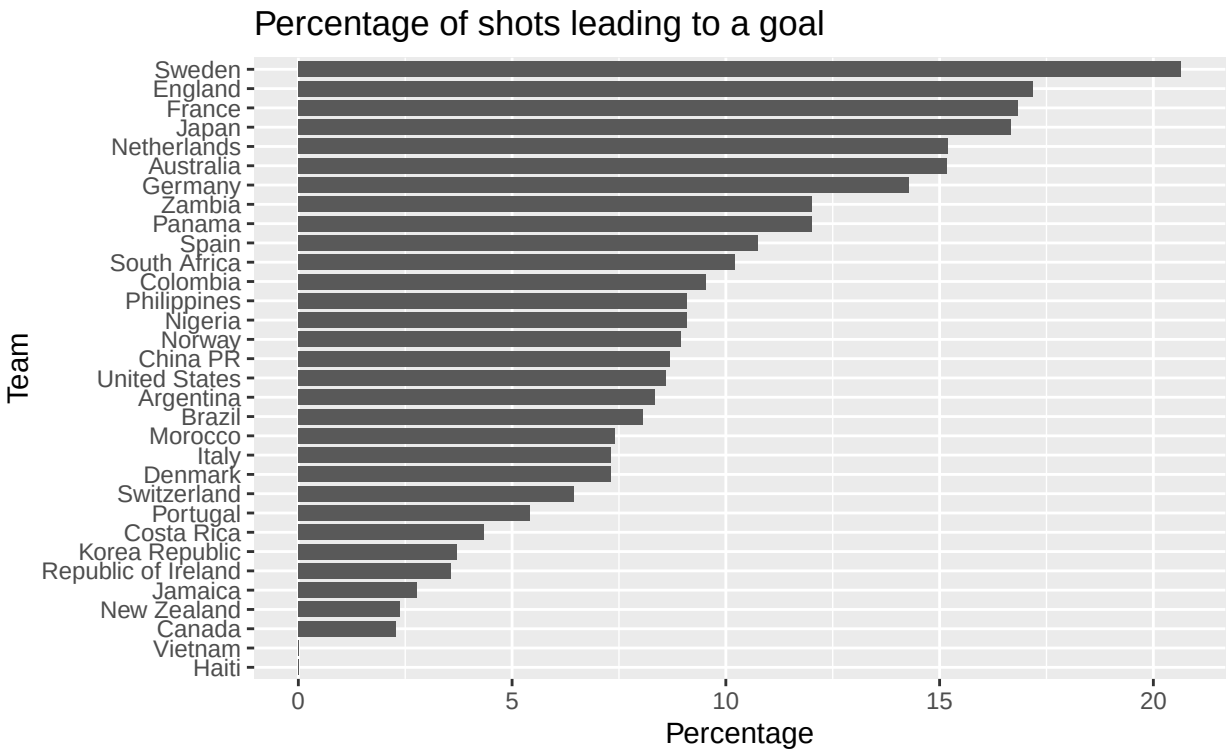


Figure 2: Diagram of the percentage of shots leading to a goal

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

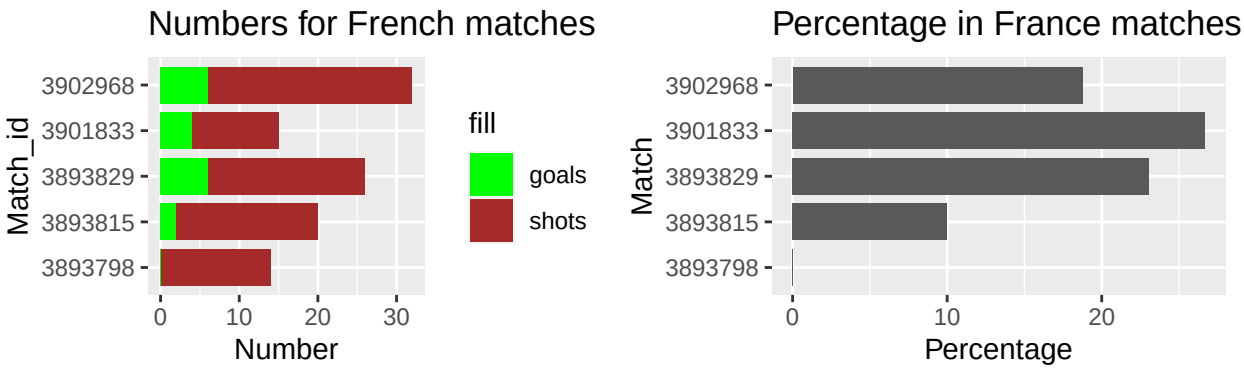


Figure 3: Diagram of the number of shots and goals for each French and percentage

2.2 Analysis of successful shots according to different variables

We first look at the type of shots.

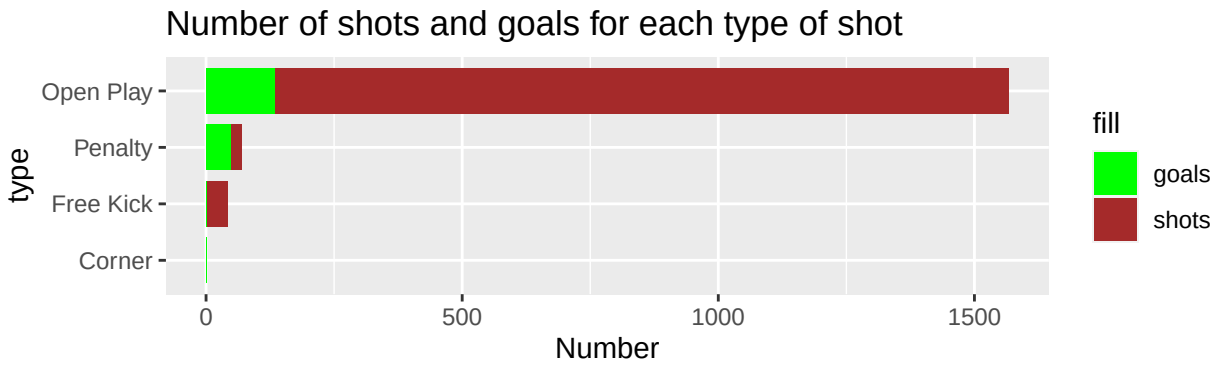


Figure 4: Diagram of the number of shots and goals for each type of shot

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

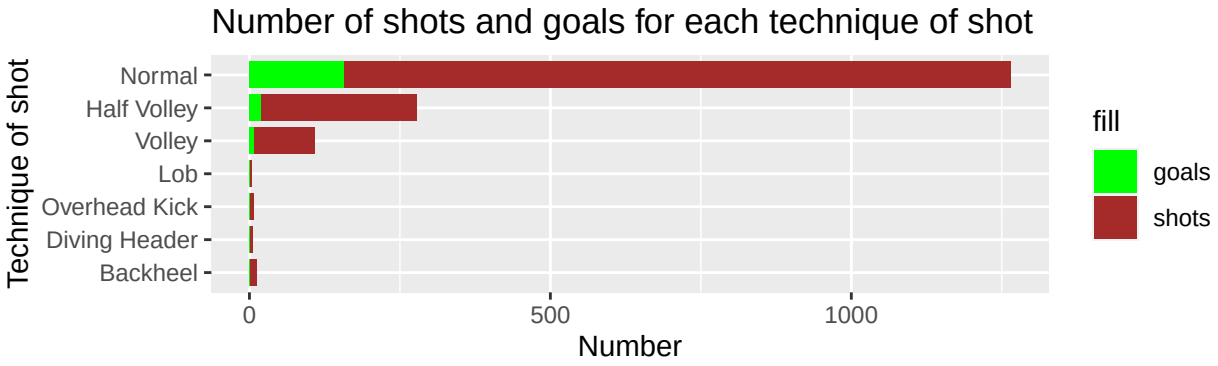
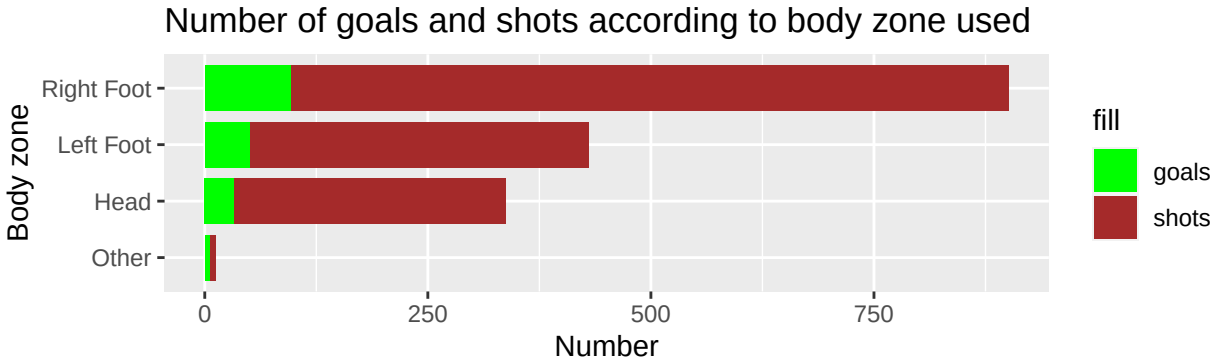


Figure 5: Diagram of the number of shots and goals for each technique of shot

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.

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 We make a model with the given expected goal as variable.

We would therefore like to find a model with an R^2 value close to this model, i.e. an R^2 close to : 0.262161

We do the same logistical model 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.

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 model 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.

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. 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.

We now create a new model like the model_3, but adding a variable : under_pressure. 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 : $8.6349383 \times 10^{-34}, 0.4171638, 0.0021353, 0.6584544$

The probability of marking the head is lower than for other parts of the body.

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 (tallonade): 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 type.name is significant, we reject H_0.

We do the same with the variable location.x :

We see a p_{value} of : $8.6349383 \times 10^{-34}, 0.4171638, 0.0021353, 0.6584544$ <0.05 so location.x is highly significant.

#———————————————————————

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`.

We create the same model as above, but adding the position of the goalkeeper. First we test the model with only the location.x.GK variable.

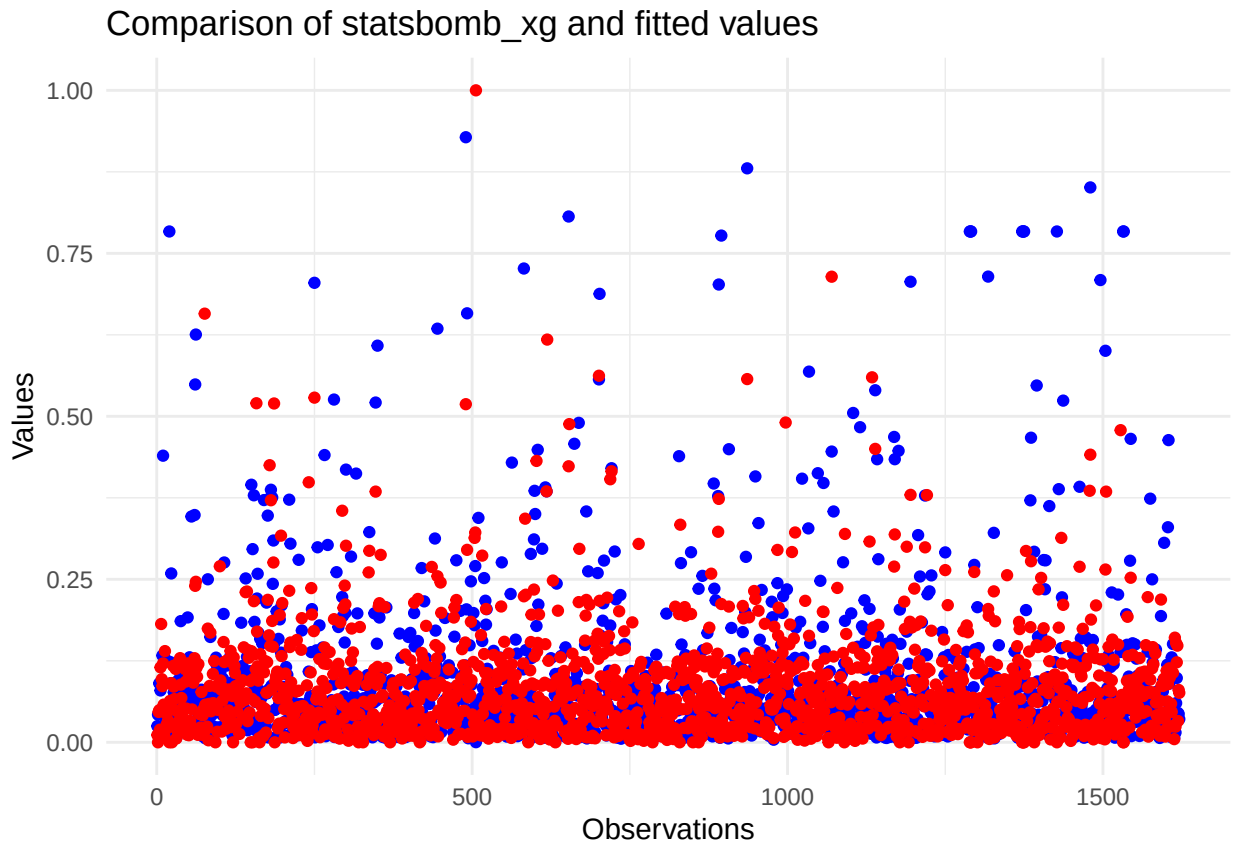
Significant effect of goal position in x. The AIC value is low, equals to `914.1462437`.

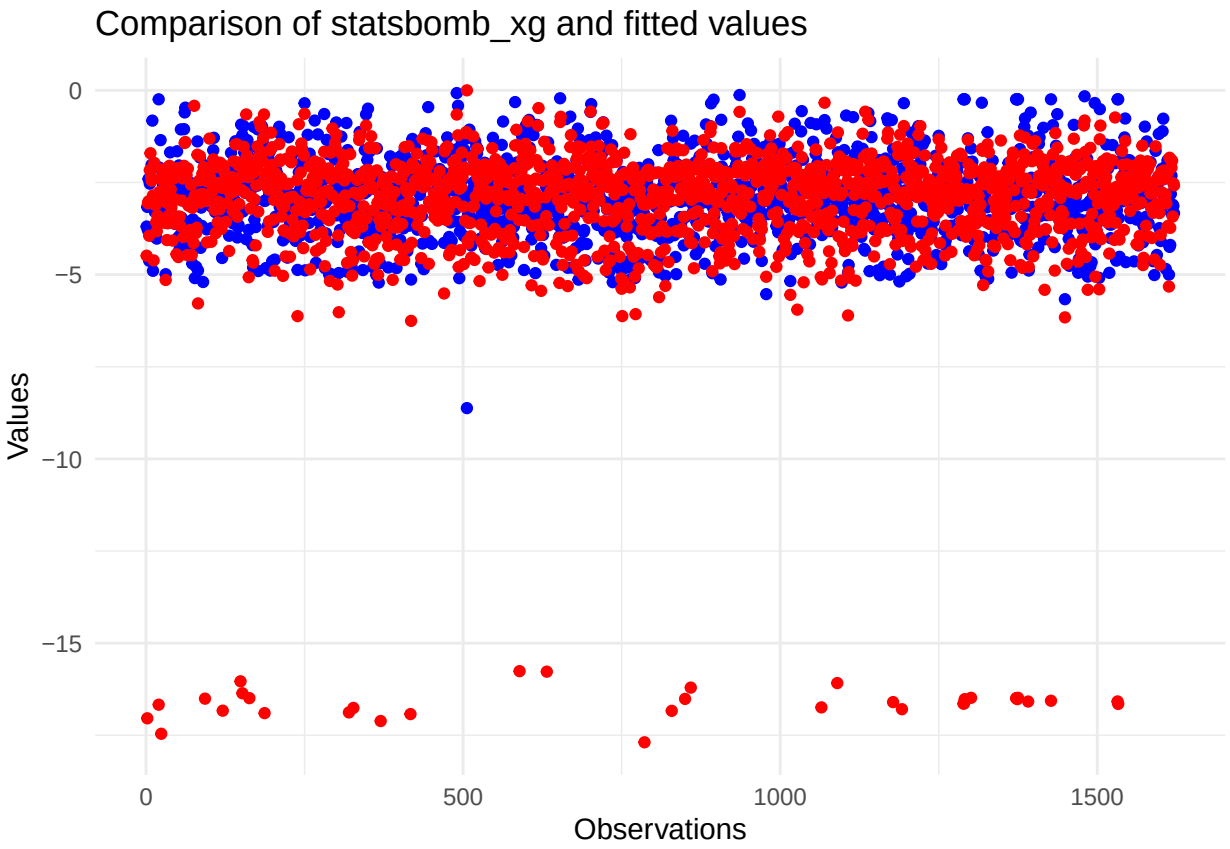
Then we do the same but with the location.y.GK variable.

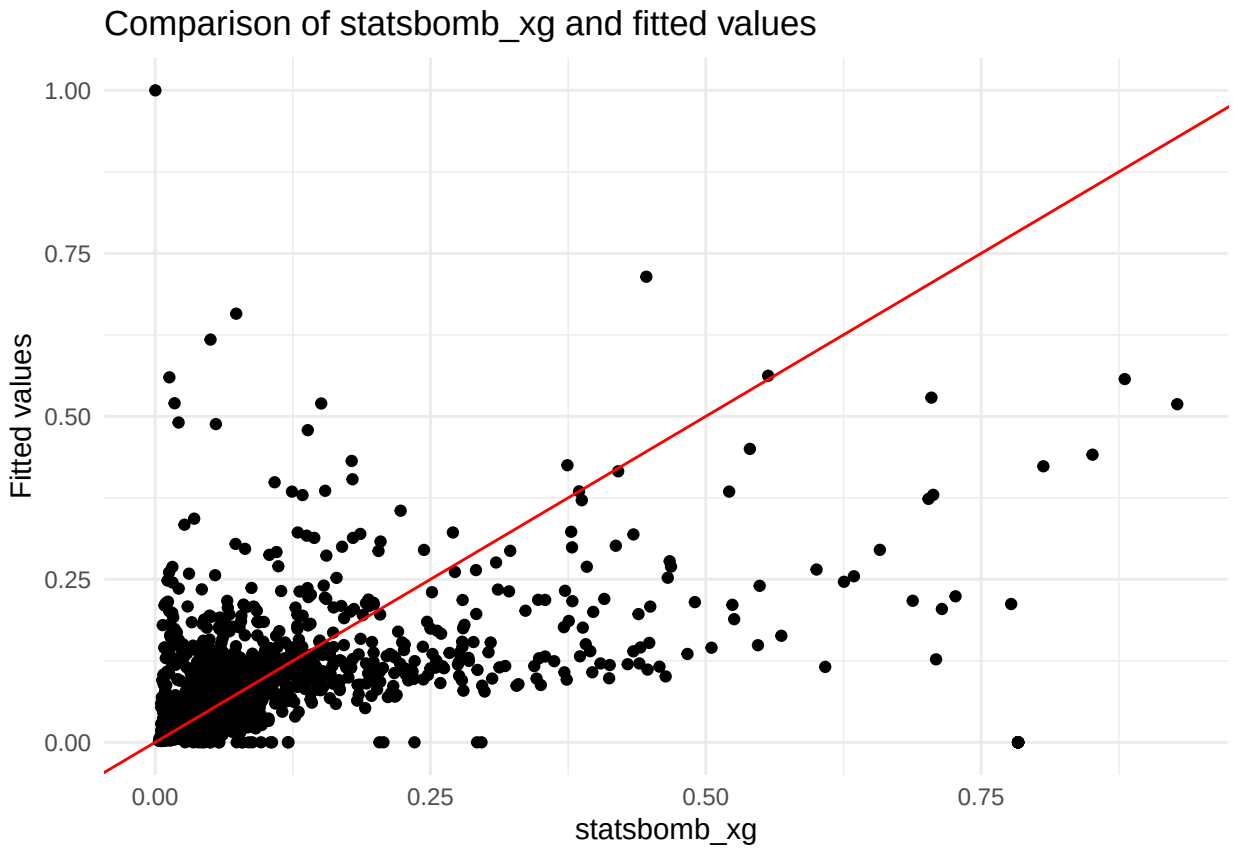
Variable for keeper position in y significant. AIC is slightly higher than for position in x, it's equal to `937.1556025`.

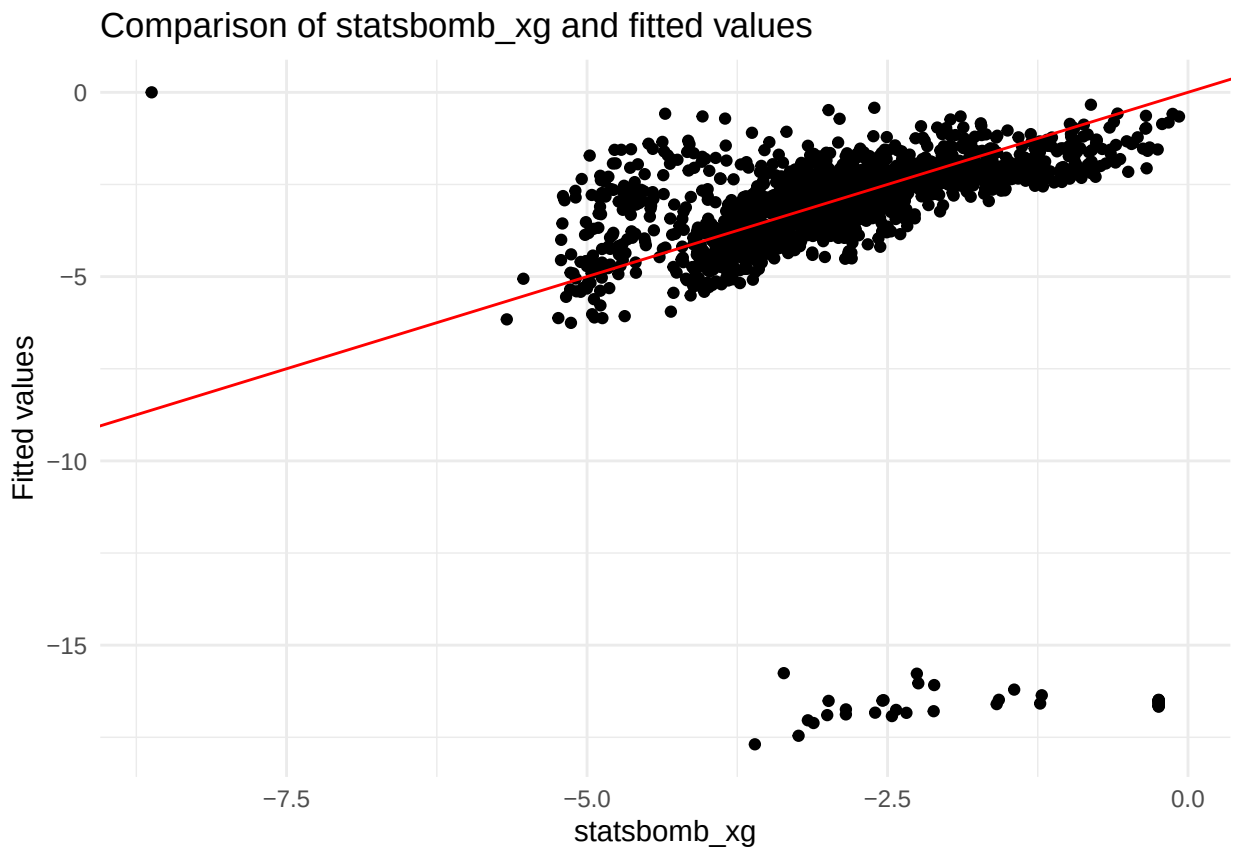
Very low AIC=`845.3219716`. We can conclude that this model is really good.

We can therefore remove the variable location.y The AIC is even lower, at `844.5352323`.
-> Best model for now


Comparison of statsbomb_xg and fitted values
1.00
0.75
0.50
0.25
0.00
Values
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.

```
##
## 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
```

You can see that y-positions are useless, only x-positions are significant, as well as the body part and the technique used. -> Best model