Projet_Women_FIFA_WC23_Analysis

2024-04-27

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.

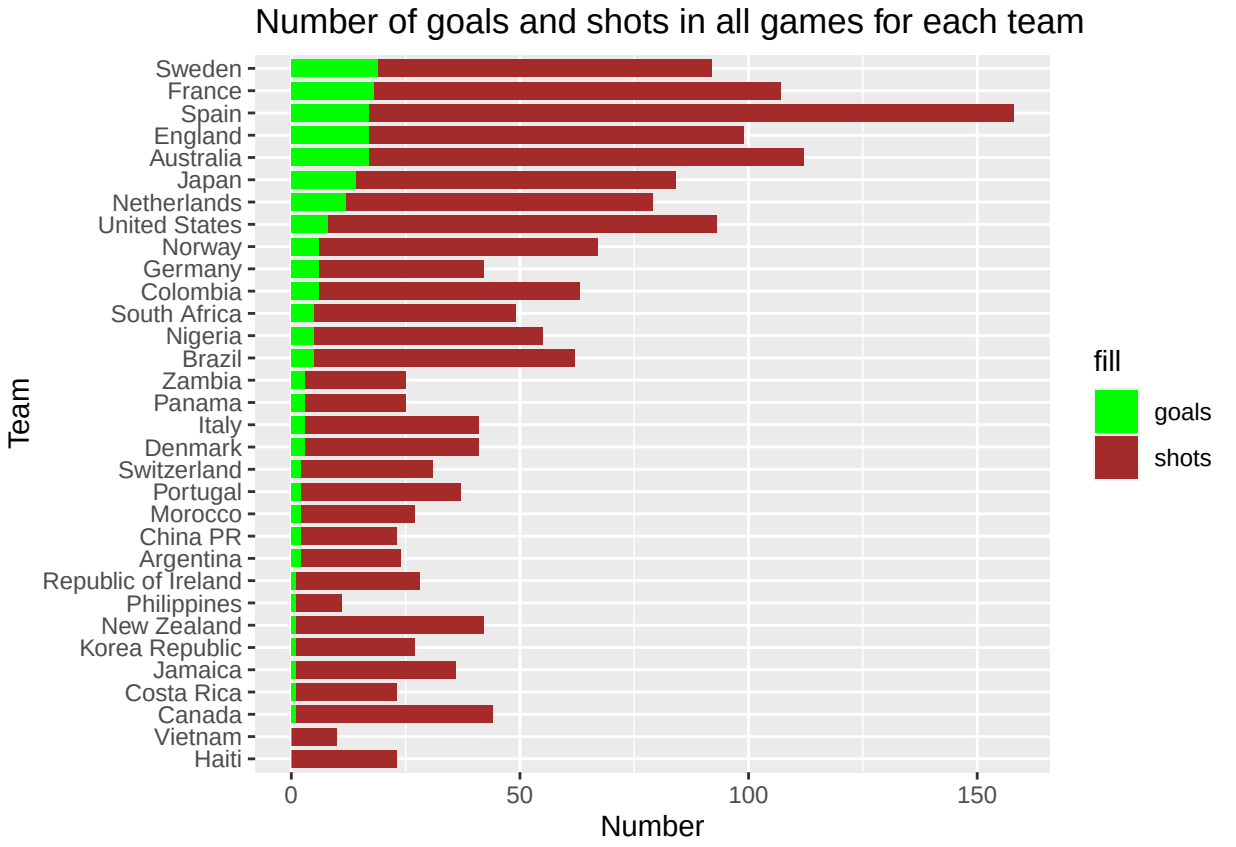


Figure 1: Diagram of the number of goals and shots in all matches for each team

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.

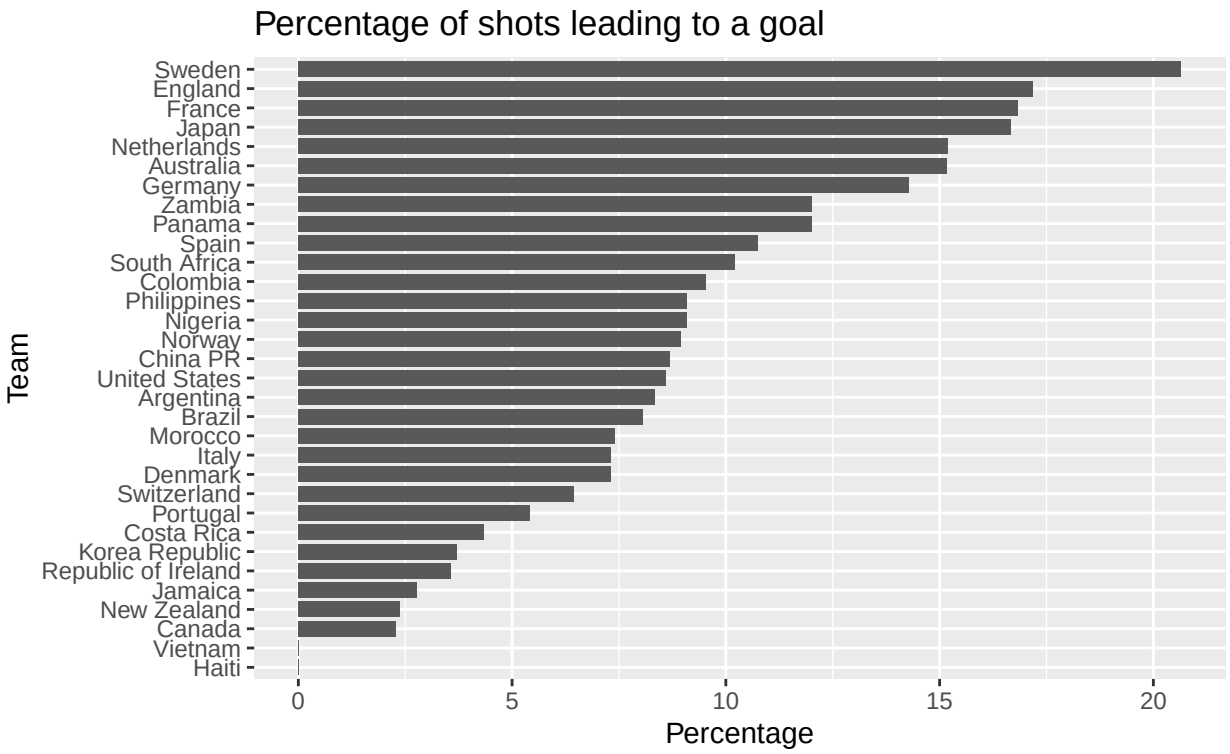


Figure 2: Diagram of the percentage of shots leading to a goal

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

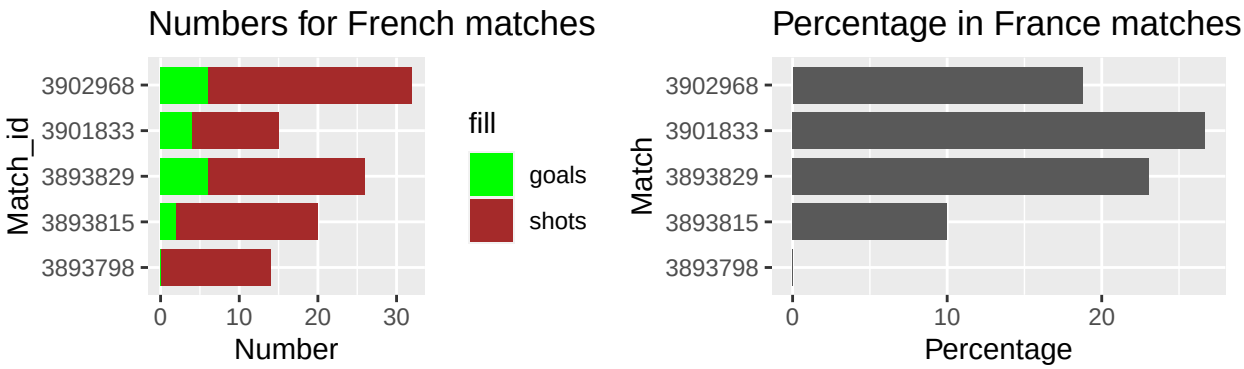


Figure 3: Diagram of the number of shots and goals for each French and percentage

2.2 Analysis of successful shots according to different variables

We first look at the type of shots.

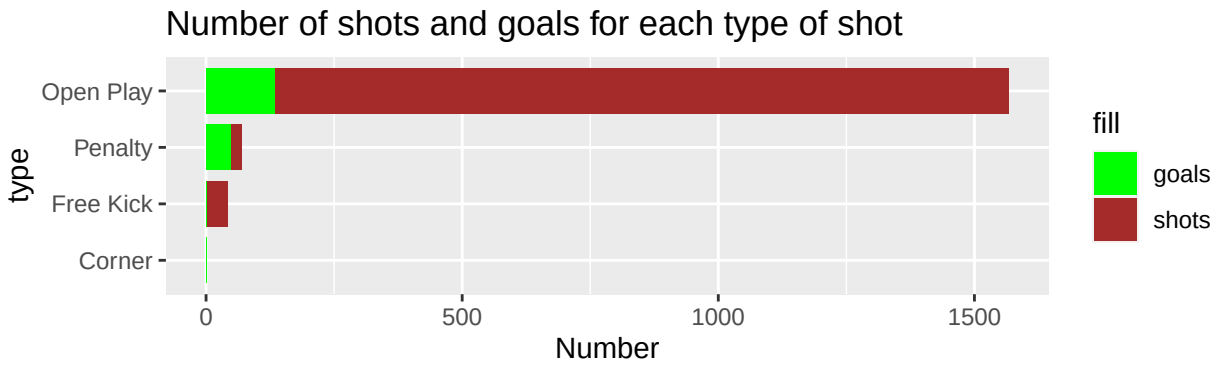


Figure 4: Diagram of the number of shots and goals for each type of shot

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

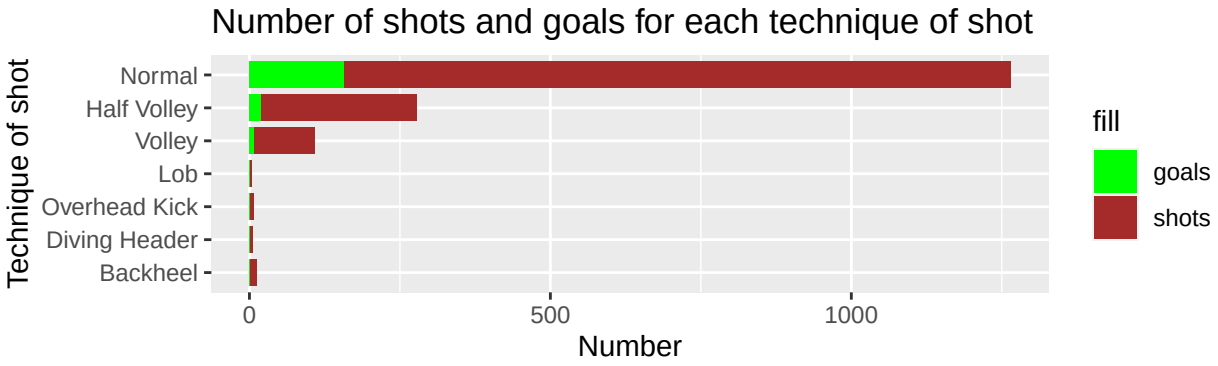
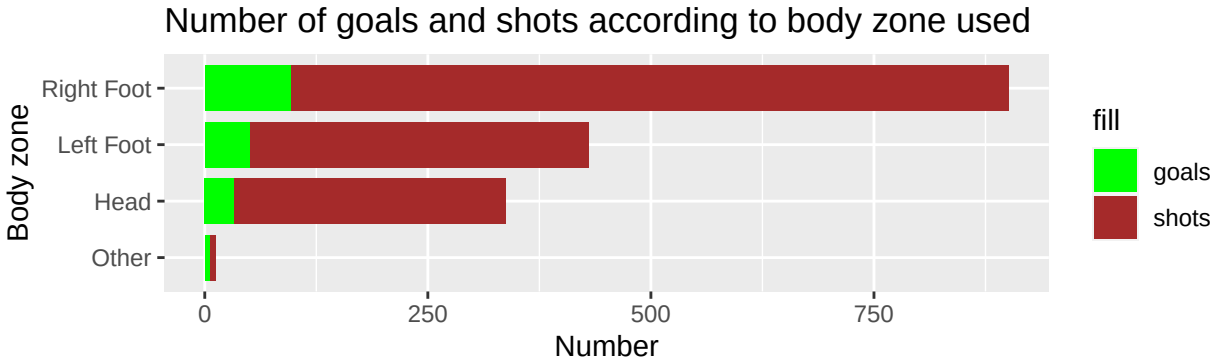


Figure 5: Diagram of the number of shots and goals for each technique of shot

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 :

```
## [1] 0.144105
```

R^2 for the model with interaction is :

```
## [1] 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 :

```
## [1] 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 :

```
## [1] 0.2054155
```

With interaction, and get an R^2 of :

```
## [1] 0.2323348
```

With interactions, 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 :

```
## [1] 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.