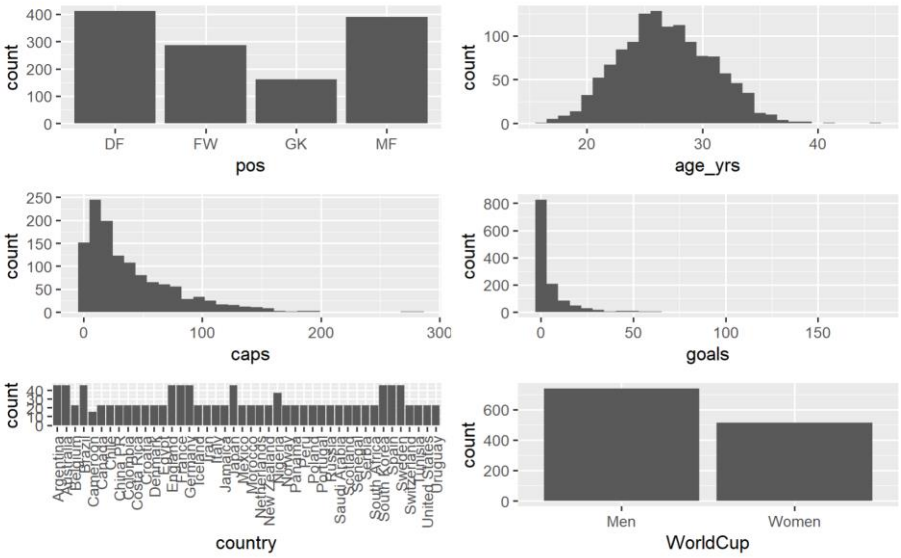
What affects the number of goals a FIFA player makes?

Introduction

Soccer is one of the most popular games around the world and strategic gameplay requires understanding the factors that improve the chances of a goal. In this study, we will carefully explore, assess, and evaluate the FIFA World Cup (2018) and the FIFA Womens World Cup (2019) data. We will consider the number of goals as the response. Among the predictors, age_yrs and caps are continuous variables while pos, country, and WorldCup are categorical variables.

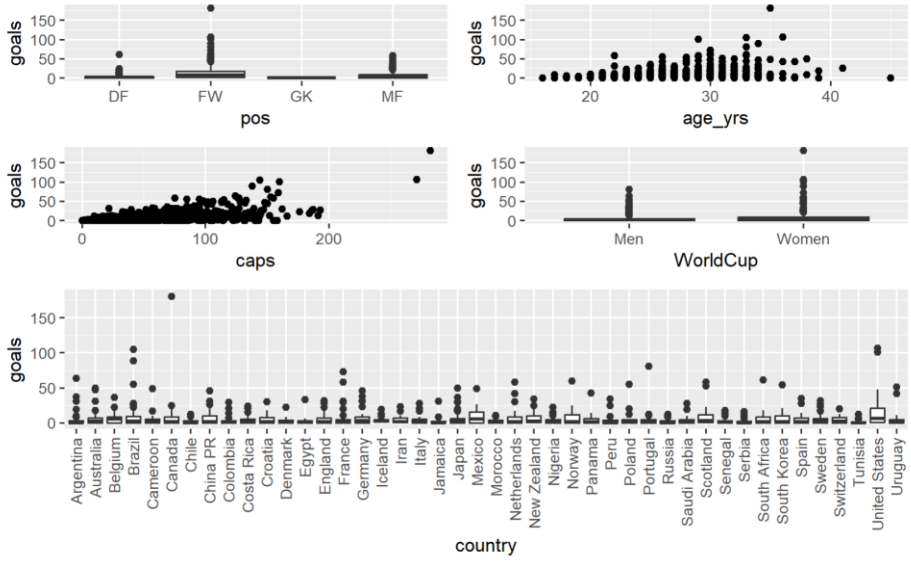
Data exploration

We first exclude all the players where number of goals is missing. Next, we look at the *distribution of factors* (**Fig 1**) that might be influencing the number of goals scored by a player. We see that all the levels of the *categorical data* seem to be well represented. For the *continuous predictors*, age seems to have a normal looking distribution. Caps and goals demonstrate large skewness. This could cause issues with the normality assumption when fitting a linear model.



Next, we use scatter plots and box and whisker plots (**Fig. 2**) to get a rough idea regarding the influence of various predictors on the number of goals. We observe the following:

1) Number of goals by FW players is higher than DF and MF players and is also higher than GK as one would expect
2) Number of goals seems to increase with age due to increase in experience gained up till age 37 after which number of goals reduces perhaps due to reduced physical strength.
3) Number of goals increases as the caps increases (especially at caps>100) as more the international appearances mean more experience.
4) The total number of goals by women are higher than men.
5) There are some countries such as Cananda, USA, and Brazil, where the highest number of goals is higher than other countries.



Model fitting

We consider the following interaction terms:

1) We consider interactions between the age and pos because older players may have more goals in their lifetime, unless they played as a goalkeeper. In other words, the effect of age on goals may depend on the position at which a player plays.
2) Interaction between the predictors Age and WorldCup is considered because the sex of the player can affect the number of goals the player can make at a certain age, older women may have lower bone densities post childbirth affecting their physical ability to score goals.
3) The predictors Age and country may interact as every country has different nutritional standards; this can affect the physical fitness of players as they age.
4) The predictors caps and pos may interact as a player with many international appearances may have a higher number of goals subject to their playing position. A goalkeeper and a forward may both have played 50 tournaments, but the forward would have more goals.
5) The predictors caps and country are also considered to interact as certain countries have had more resources for training players, hence, a player from country "A" and a player from country "B" with the same number of international appearances may have different number of goals.

The predictors WorldCup and caps are not considered to interact, as the sex of the player should not influence how their number of international appearances affect the number of goals they score.

Model diagnostics

After model fitting, we found departures from assumptions of constant variance and normality. We also found 28 bad high leverage points, 9 outliers and 1 influential point in the model, which were probably due to the non-constant variance and non-normality issues. We used a box-cox transformation to fix the issues in the model.

Model transformation

For box-cox all the responses should be strictly positive. As some players have 0 goals, we update the goals random variable as "goals+1" and fit a new model. The new model was passed as an input to the box-cox function yielding an optimal lambda of -0.12, which we round to -0.1. We use the box-cox based transformed response to fit a new model.

After transformation, we see that the variance and normality assumptions are satisfied. We also see that the correlation between continuous predictors is not very high hence there is no issue of collinearity in the design matrix. We find 3 bad high leverage points corresponding to players - Christine Sinclair (captain), Carli Lloyd (co-captain), and Cristiano Ronaldo (captain). All these players are captains and are well known to be extraordinarily talented players. Therefore, having them as leverage points makes sense. We also observe no influential points.

Model selection

We then use the sequential ANOVA test for model selection:

```
Analysis of Variance Table

Response: goals_plus1_trans
                  Df Sum Sq Mean Sq   F value    Pr(>F)
age_yrs            1  87.97   87.97  312.8792 < 2.2e-16 ***
caps               1 460.79  460.79 1638.7925 < 2.2e-16 ***
pos                3 266.82   88.94  316.3096 < 2.2e-16 ***
WorldCup           1   0.02    0.02    0.0541 0.8161725
country           43  21.11    0.49    1.7460 0.0023016 **
age_yrs:pos        3  23.44    7.81   27.7930 < 2.2e-16 ***
age_yrs:WorldCup   1   3.10    3.10   11.0406 0.0009206 ***
age_yrs:country   43  24.24    0.56    2.0050 0.0001597 ***
caps:pos           3  33.36   11.12   39.5471 < 2.2e-16 ***
caps:country      43  27.90    0.65    2.3074  5.12e-06 ***
Residuals       1105 310.70    0.28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Predictors with a p-value lower than 0.05 are considered significant. We see that all the interaction terms are significant, so we cannot drop any of the variables. On checking if the categorical terms are significant, we see that pos and country are significant since the p-value < 0.05. Although WorldCup has a p-value greater than 0.05, we know that all the interaction terms consisting of WorldCup are significant. Therefore, we cannot drop WorldCup according to the hierachichal rule. Both the continuous variables are significant, as age_yrs and caps have a p-value less than 0.05.

We now use the model to draw inferences regarding the influence of different predictors on the number of goals.

1) In terms of differences between men and women soccer players when it comes to scoring goals, we find that the interaction term for age and sex (i.e., worldcup) is statistically significant. However, the coefficient for worldcup alone is not significant. Hence, the sex of the player only changes the effect age has on the number of goals.
2) We find that the position, as well as the interaction of position with number of international appearances and age are significant. Hence, there is an effect of position on number of goals.
3) We can also determine the effect of age on the number of goals by looking at the coefficients of the trained model; note that the box-cox transformation we applied is a monotonic transformation, and hence, a positive coefficient still means an increase in the response. We see that the beta for age is positive, and hence, age generally increases the number of goals a player can make. However, interactions can also positively and negatively influence the effect age has on goals as seen from the sequential ANOVA results.

In conclusion, our results show that a player’s playing position, age, experience in terms of international appearances, nationality, and sex all influence the number of goals they score. In general, an increase in age and an increase in experience increases the number of goals a player may score, which may be valuable information for stakeholders. However, several factors such as the sex, playing position, and nationality impact the effect of age and experience on a player's scoring ability.