What Best Predicts Success of a Club in UEFA Champions League?

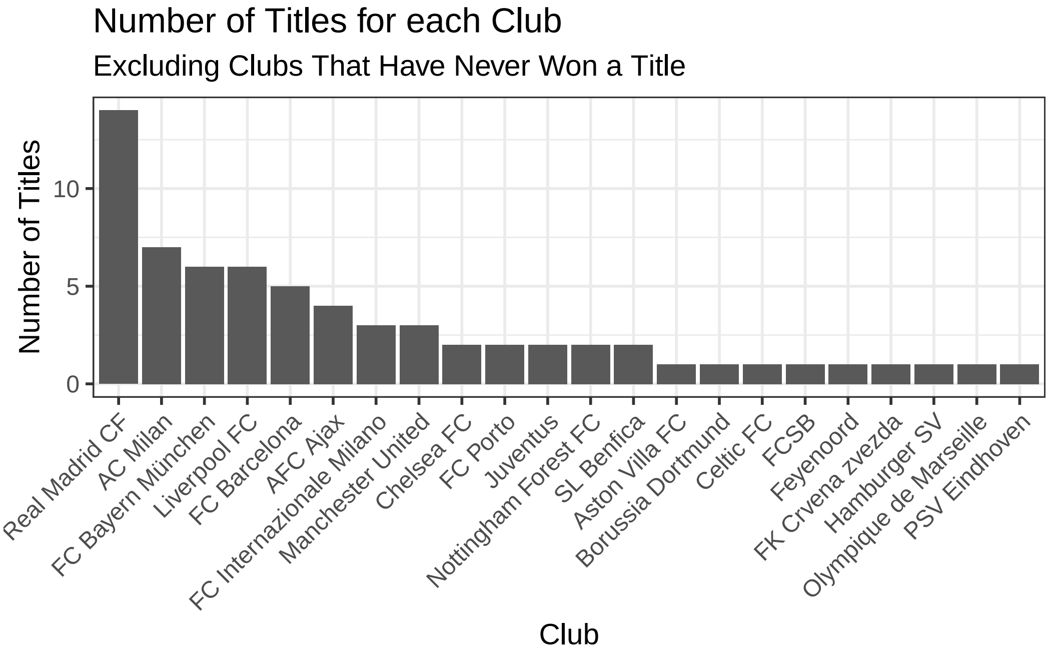
Cole Walker, Madison Griffin

Introduction and Data

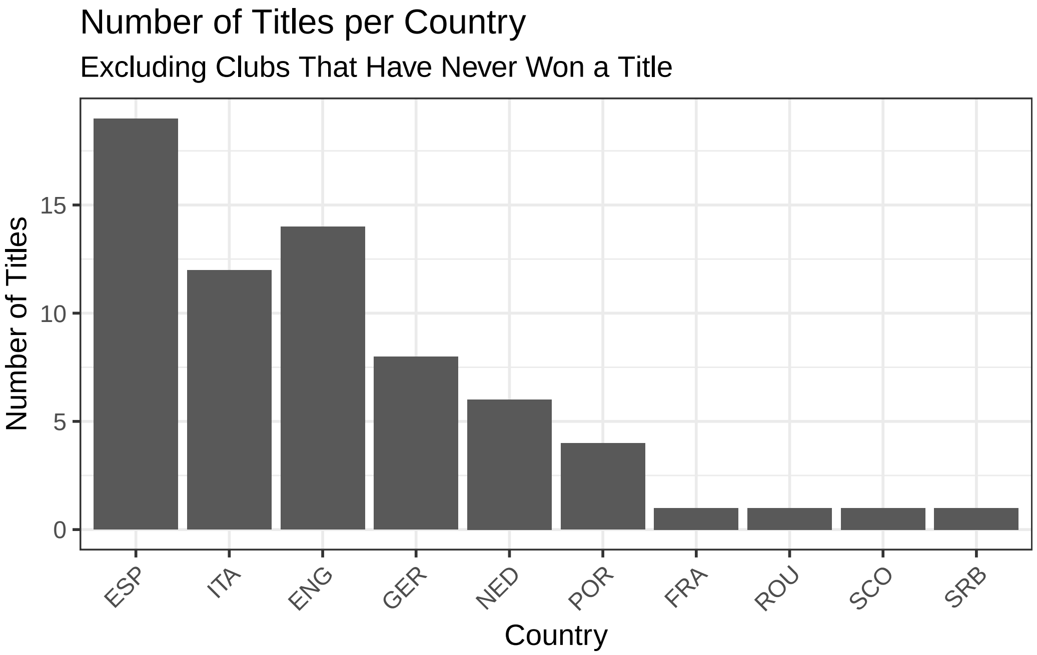
Data Cleaning

EDA

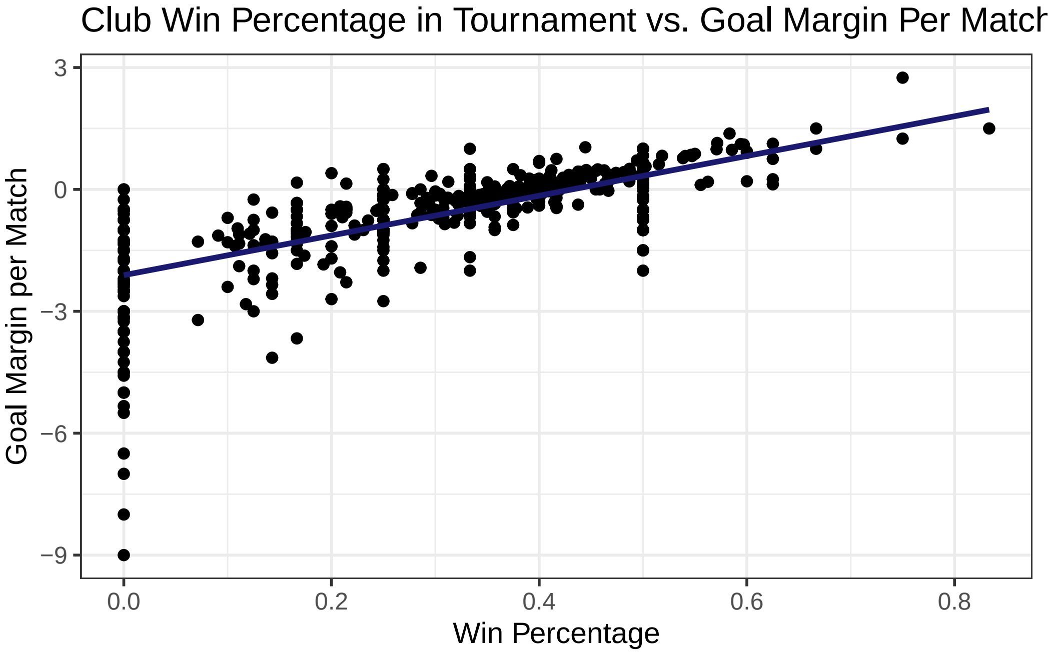
Plot 1:



Plot 2:


Number of Titles per Country
Excluding Clubs That Have Never Won a Title
Number of Titles
0
5
10
15
ESP
ITA
ENG
GER
NED
POR
FRA
ROU
SCO
SRB
Country

Plot 3:


Club Win Percentage in Tournament vs. Goal Margin Per Match
Goal Margin per Match
3
0
−3
−6
−9
0.0
0.2
0.4
0.6
0.8
Win Percentage

Methods

Since our outcome variable is numeric and continuous, we knew our model was either a linear regression or a linear mixed effects model. We hypothesized that the best predictors to determine the success of a club in the UEFA Champions league were win percentage, goals scored per match, goals scored against per match, goal margin per match, and whether a club belonged to a top five country league (England, France, Italy, Spain, or Germany).

To select the most effective variables we conducted five different variable selection processes: all subset, stepwise (forward, backward, and both), and LASSO.

The variables selected in forward selection, both directions selection, and LASSO were win percentage, goal margin per match, and top five league. The variables selected in backward selection were win percentage, goals scored per match, goals scored against per match, and top five league. The variable selected for all subset selection using Mallo's CP was only win percentage.

Comparing RMSE after variable selection

To compare each of these models, we compared their RMSE.

RMSE All Subset = 0.1642128

RMSE Best Backward = 0.1348656

RMSE Best Both, Forward, and Lasso (because they chose the same variables) = 0.1348667

Since the model from backwards selection had the lowest RMSE, we decided to use those variables to assess linear regression assumptions and conditions (win percentage, goals scored per match, goals against per match, and top five league).

We hypothesized that there could be a violation of independence because clubs in the same country, especially countries that are in the top 5 league, could have access to more money, better facilities, coaches, and training regimens, which could violate their independence from each other. Because of this, we tested the linear model assumptions and conditions using the variables chosen by backward selection in two models: 1) a linear regression, and 2) a linear mixed effects model with a random intercept for top five leagues.

Checking Assumptions

Model 1: Linear Regression

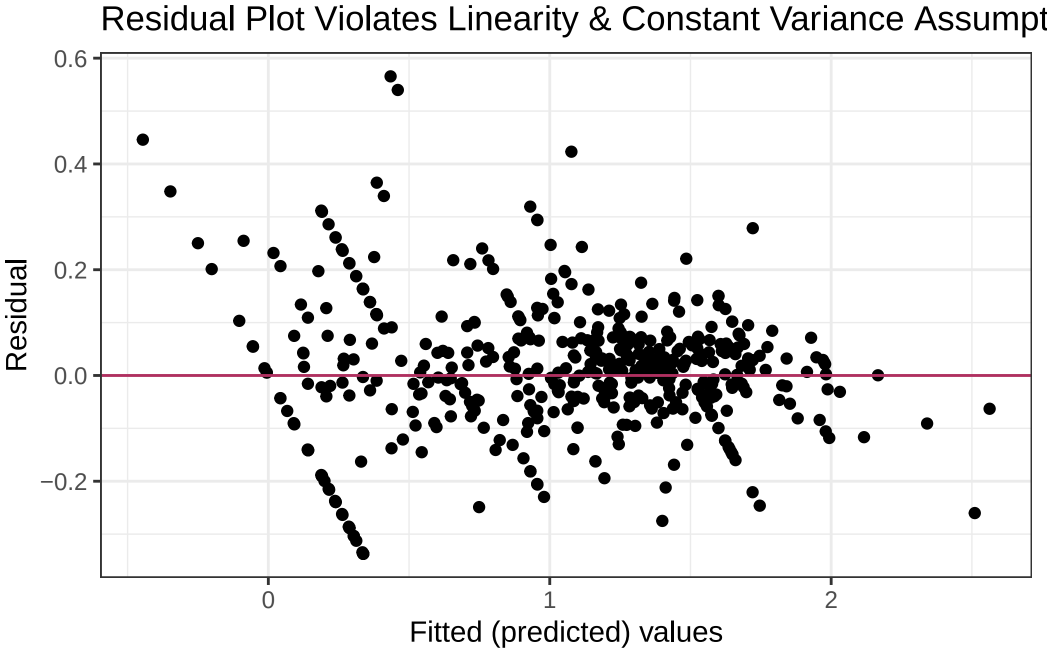
Outcome:

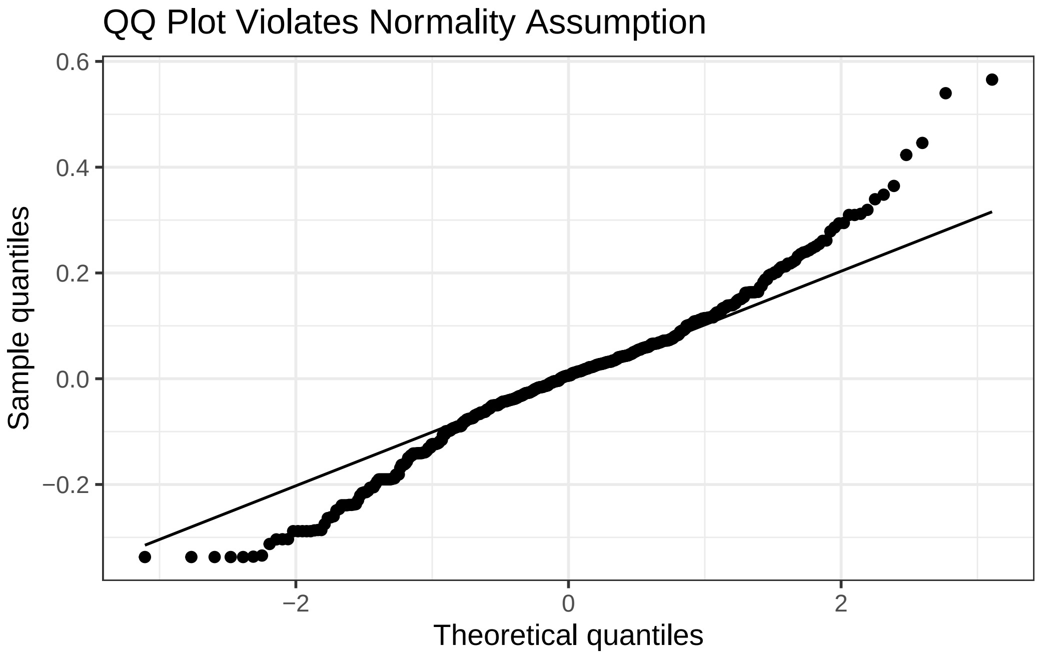
- points per match

Predictors:

- win percentage
- goal scored per match
- goals against per match
- top five league

For the linear regression, the residual plot (shown below) violates both linearity and constant variance. The model starts to underpredict more on the right side of the graph, and there are three diagonal patterns across the residual plot. The Q-Q plot (shown below) also deviates from the line in the bottom left and upper right, thus violating normality.





Model 2: Linear Mixed Effects Model

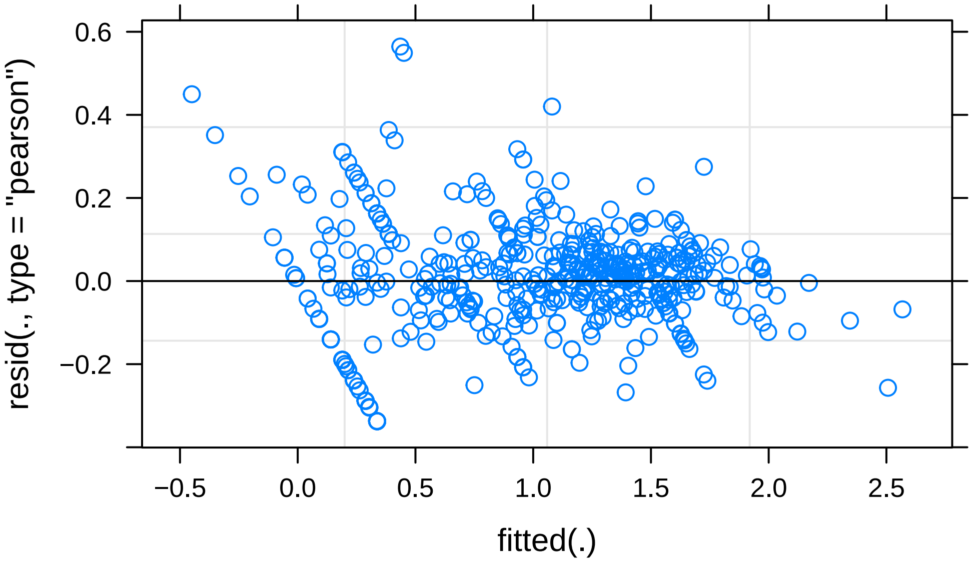
Outcome:

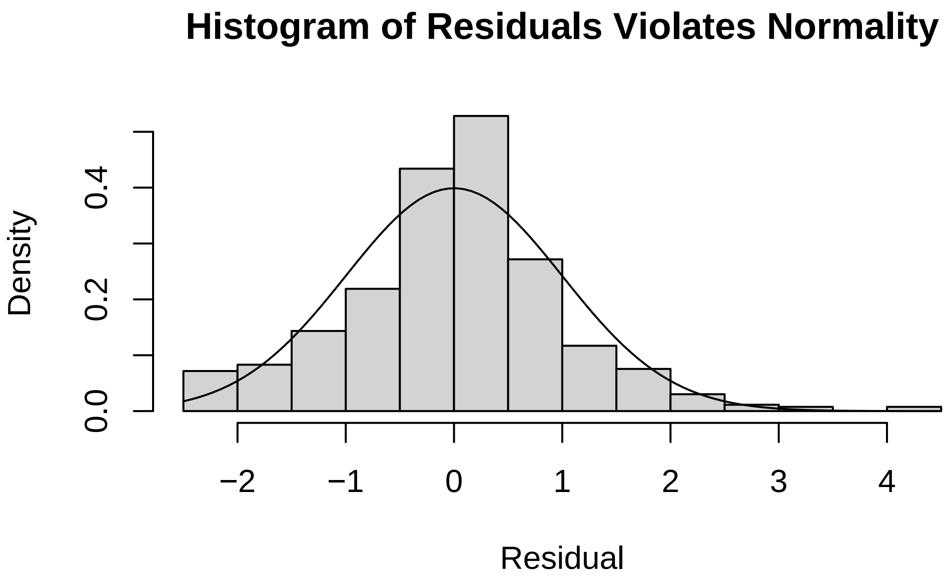
- points per match

Predictors:

- win percentage
- goal scored per match
- goals against per match
- random intercept for top five league

The residual plot (shown below) for the linear mixed effects model also has three diagonal patterns. The residual plot is also clumped, and begins to underpredict on the right side of the plot. The histogram of residuals (shown below) also violates normality, as the bins are relatively large and the bars in the middle fall outside of the normal curve.





Both models violated the assumptions, however, with the potential violation of independence because of the variable top five league, we will choose the linear mixed model as our final model.

Results

FINAL MODEL:

$y_{ij} = (\gamma_{00} + \mu_{0j}) + \gamma_1 WinPercentage_{ij} + \gamma_2 TopFiveLeague_{ij} + \gamma_3 GoalsPerMatch_{ij} + \gamma_4 GoalsAgainstperMatch_{ij} + \epsilon_{ij}$

where

y_{ij} = points per match

γ_1: wins percentage

γ_2: top five league, 1 = top five

γ_3: goals per match

γ_4: goals against per match

```
Linear mixed model fit by REML ['lmerMod']
Formula:
pointspermatch ~ 1 + winpercentage + goalspermatch + goalsagainstpermatch +
    (1 | topfiveleague)
   Data: soccer

REML criterion at convergence: -589.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4908 -0.5166  0.0546  0.5113  4.1647

Random effects:
 Groups        Name        Variance  Std.Dev.
 topfiveleague (Intercept) 0.0001943 0.01394
 Residual                  0.0183620 0.13551
Number of obs: 530, groups:  topfiveleague, 2

Fixed effects:
                      Estimate Std. Error t value
(Intercept)           0.443476   0.023240  19.082
winpercentage         2.383025   0.049148  48.487
goalspermatch         0.097878   0.014635   6.688
goalsagainstpermatch -0.098400   0.006629 -14.843

Correlation of Fixed Effects:
            (Intr) wnprcn glsprm
```

```
winpercentg -0.342
goalsprmtch -0.237 -0.687
glsgnstprmt -0.676  0.436 -0.134
```

The coefficient for win percentage (fixed effect) is 2.383. This means that at a given league (top five or not top five), every additional one-unit increase in a given club's win percentage, their predicted points per match is expected to increase by 2.383 points per match, while controlling for other variables in our model. The coefficient for goals against per match is -0.0984. This means that at a given league (top five or not top five), for every additional goal against per match, points per match is expected to decrease by 0.0984, while controlling for other variables in our model.

Though the output does not provide p-values, we will refer to t values to interpret significance of our fixed effects. Both win percentage (t value = 19.082) and goals against per match (t value = -14.843) have high t values, indicating their significance. Goals scored per match have a small t values (t value = 6.688), showing it was not as significant as the other predictors.

In summary, win percentage, a club's defensive ability, a club's offensive ability, and whether a club is from a country in the top five leagues were found to be the best predictors of UEFA Champions League success, however win percentage and a club's defensive ability were the most significant.

Conclusion