What is best associated with how successful a team has performed in the UEFA Champions League?

Cole Walker, Madison Griffin

Introduction and Data

The UEFA Champions League (UCL) is one of the most entertaining sports tournaments in the world. The tournament consists of 32 clubs separated into eight groups, and clubs around Europe's top-division leagues qualify based upon where they finish in the final standings from their previous domestic league season. Each league is allocated a certain number of qualifying spots based on how well their clubs have performed in European continental competitions over the previous five years (Snowman Sports Media, 2020). Because of this, clubs from Europe's most talented leagues in England, Germany, Spain, Italy, and France typically have 3-4 representatives compete per year, whereas clubs from smaller, less-talented leagues in countries like Austria, Belgium, and Serbia typically only have 1-2 representatives.

Winning a UCL title remains one of the most impressive feats in all of sports given that each club in the tournament is either a domestic league champion or at least one of the other top-four finishers, so the competition is especially stiff. In our analysis, we wanted to identify potential statistical predictors and how they are associated with a club's historical success in the UCL. We asked the question, what relationships exist between each of these predictors and the response variable that we chose to best represent success in the tournament? We found a dataset created by Bashar Naji titled "AllTimeRankingsByClub" which compiled official data from UEFA's website and highlighted tournament statistics from every club that has competed in the UCL from its 1955 founding up until the group stage of the 2021/22 season (Naji, 2022). The dataset incorporated variables including the name of every club that has competed in the UCL up to the 2021/2022 season, the country they belong to, the number of tournaments they have participated in, the total number of matches played, the total number of wins, losses, and draws, and the total number of goals scored, conceded, and overall goal differential throughout their history participating in the tournament.

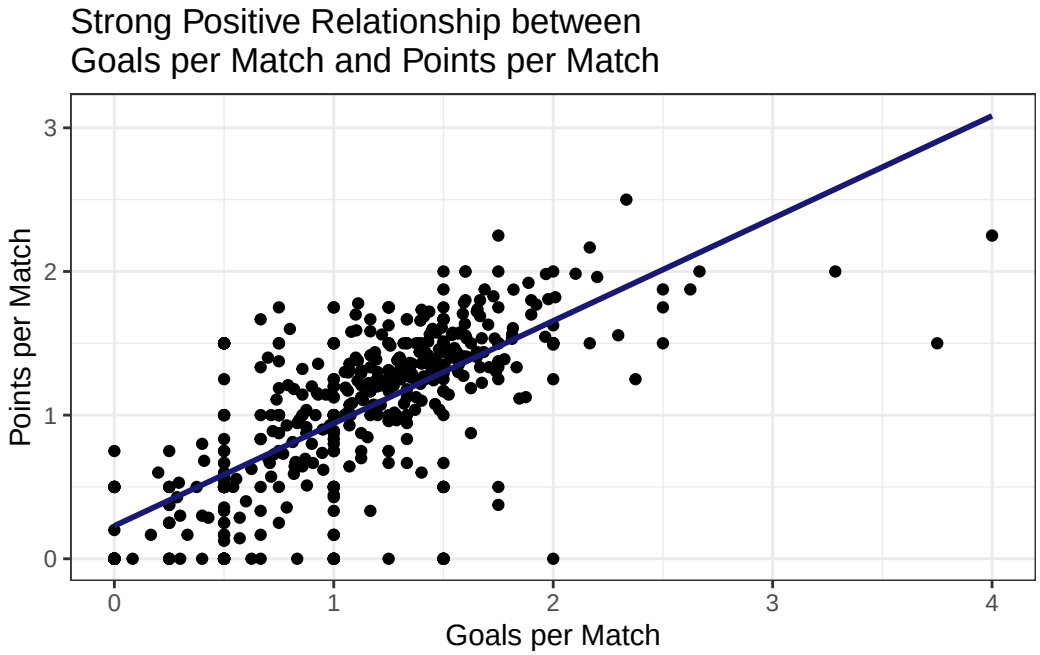
Data Cleaning

Given that the dataset examines more than 60 years of history and that some clubs such as juggernauts Real Madrid and Bayern Munich have far more wins and higher goal scoring statistics than clubs with far fewer appearances in the tournament, we decided to undergo a process of data cleaning to best match this reality. First, we mutated the wins statistic to a win-percentage ratio (wins/total matches played) to create a more generalized point of comparison between the clubs. We then mutated the three goal-related statistics to per-match versions with the same intention. In addition, we mutated the club's country league categorical variable into a binary categorical variable that differentiated whether the club came from a top-five league (England, Germany, Spain, Italy, and France) or not to more clearly discern the competitive environments that clubs competed in before qualifying. Lastly, we calculated the total number of points per UCL guidelines (3 points for a win and 1 point for a draw) that each club has accumulated throughout their UCL history and created new "points" and "points per match played" variables to obtain an additional measure of success.

EDA

We will introduce our final model and the best predictors that it included in our Results section. However, this section explains our rationale for choosing the initial list of variables that we could include in our model and for determining which variable would work best as our response variable and which ones would better serve as predictors. Given that only 22 clubs have ever won the UCL and the other 508 clubs included in the dataset would have a zero value for their titles, we decided to go against the typical practice of using number of titles as our response variable. Instead, we determined that the mutated "points per match played" variable would provide a more accurate and representative determinant of success because it takes into account clubs' overall quality of performance throughout all stages of the tournament. We decided to utilize win percentage, the three per-match goal statistics, as well as the binary "topfiveleague" variable as the potential predictors. Win percentage was a relatively easy inclusion given how essential securing wins are for boosting point totals. We included the three per-match goal variables because we wanted to examine the relationship between clubs' offensive prowess, defensive brilliance, and overall balance were and their success in the UCL. Lastly, we chose the binary "topfiveleague" variable as a predictor because we wanted to examine the relationship between the strength of a club's domestic league and their success in the tournament.

Attached below is a visualization showcasing the strong positive relationship between a given club's goals scored per match and its average points accrued per match. This visualization represents one example of the linear relationships that we found between our chosen predictors and response.



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

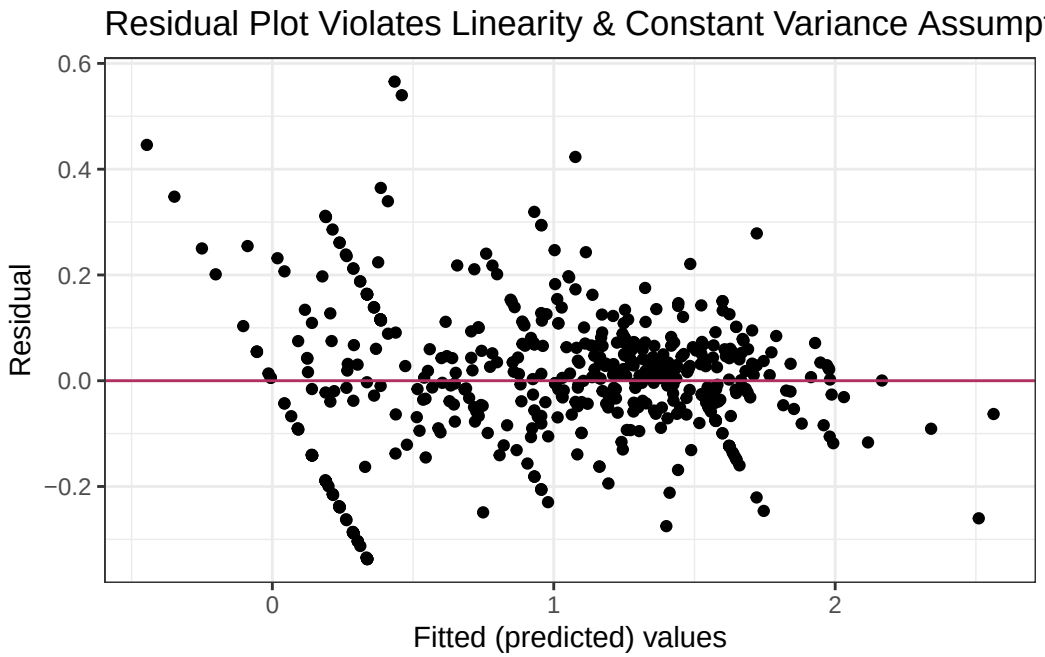
Outcome:

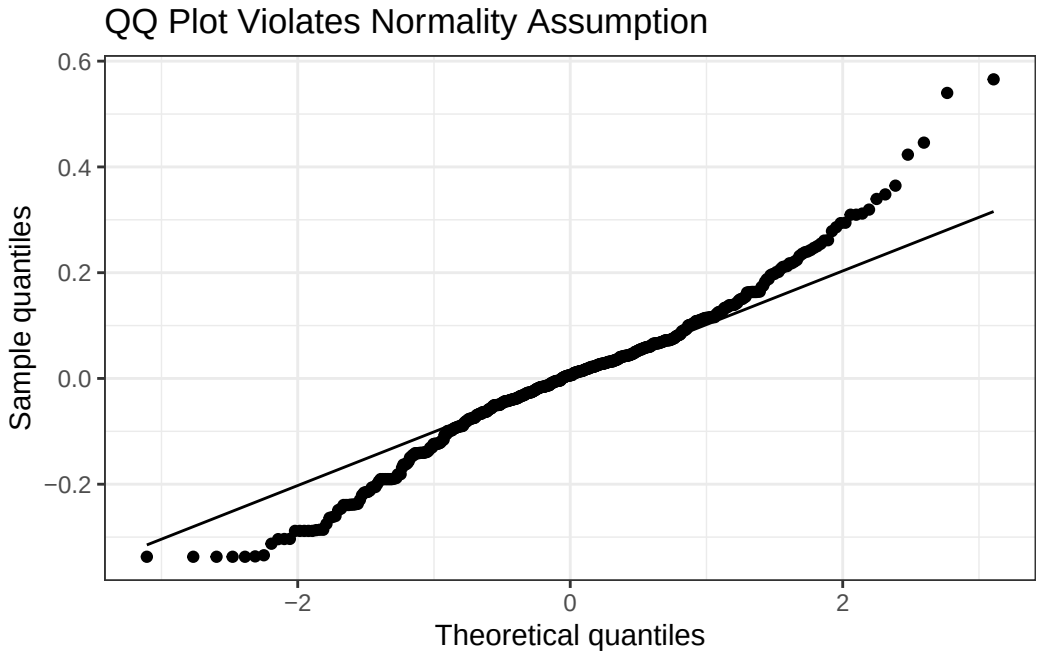
- points per match

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

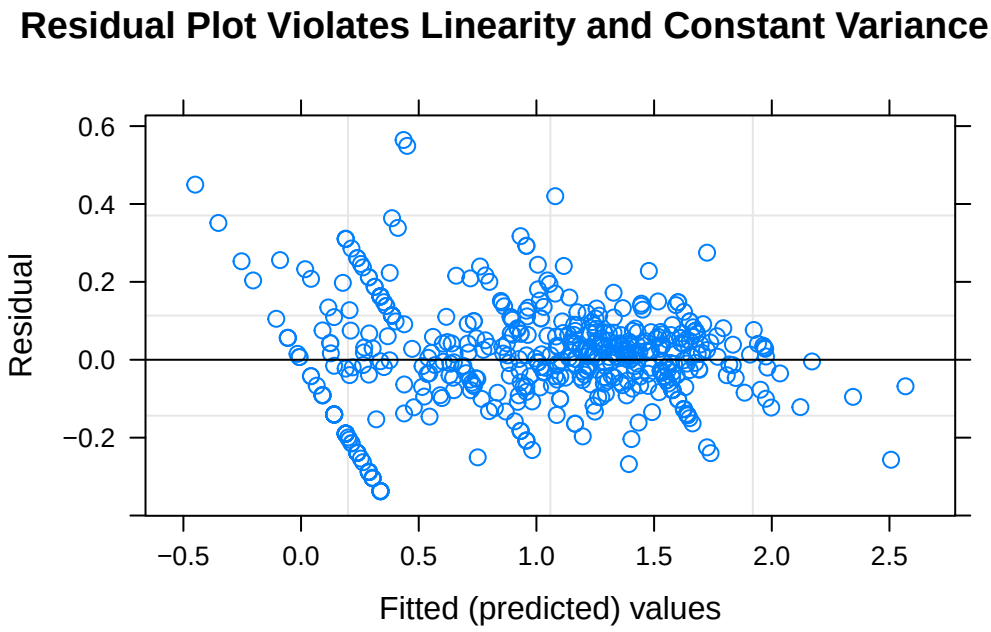
Outcome:

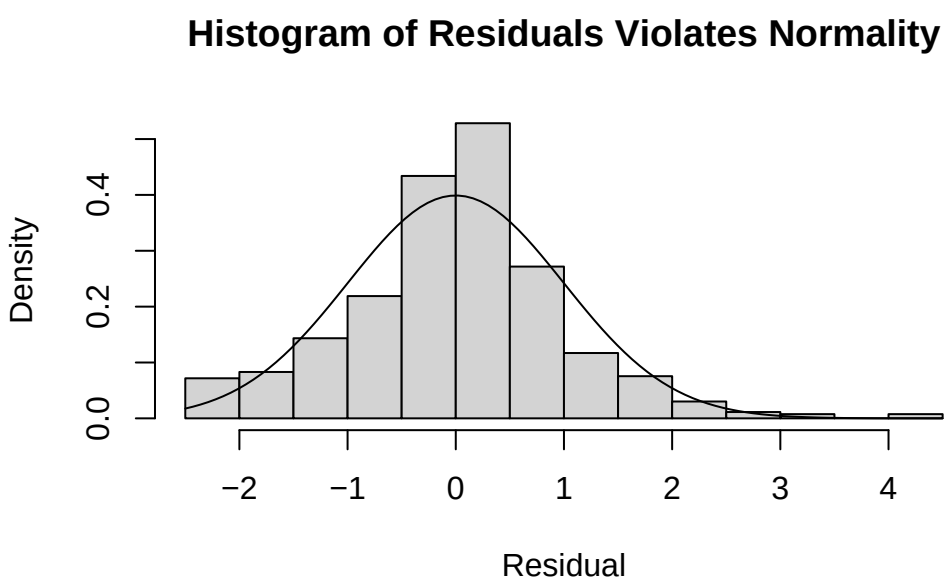
- points per match

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

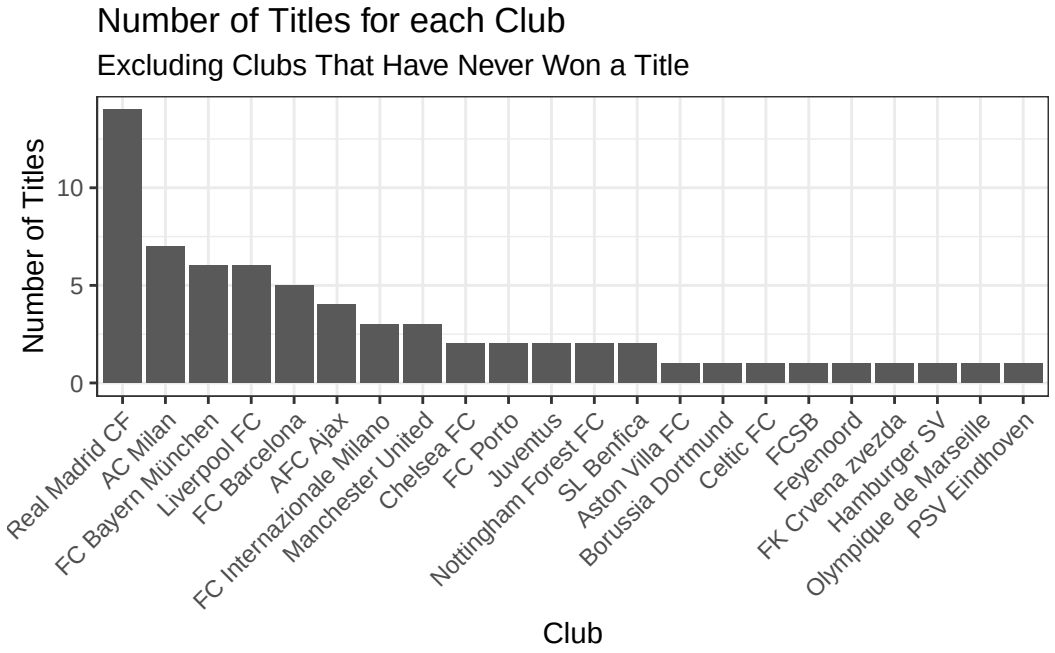
In summary, our analysis allowed us to examine the positive and negative relationships between potential statistical predictors associated with a given club's historical success in the UCL. Our model found that each of the three fixed predictors chosen by our variable selection methods had a linear relationship with our chosen response variable that best represented success in the tournament. In statistical language, regardless of the club's domestic league, our model predicted that there was a positive relationship between both a club's win percentage and their predicted points per match as well as between their goals per match and their predicted points per match, while adjusting for the other variables in the model. Our model also predicted that, regardless of the club's domestic league, there was a negative relationship between a club's goals against per match stat and their predicted points per match, while adjusting for the other variables in the model. These findings match real-world expectations given that there is a three-point gain for winning UCL matches and that maximizing goals scored and minimizing goals allowed attributes to winning and thus gaining points. By randomizing the effect of a club belonging to top five league or non-top-five league, we were able to address potential violations of independence between clubs given that clubs from the same league or tiered leagues may have similar access to resources, training equipment, and quality of coaching personnel.

Unfortunately, some clear limitations still exist in our data and two especially stand out. First, our original dataset did not provide any additional information on which specific years each team has competed in the UCL, so we were unable to determine whether or not their statistics incorporated data from the pre-1992 tournament rebranding and subsequent gradual rule/qualification changes, which could have affected the rigors of the tournament. Second, as was referenced in the introduction, we utilized points per game as our response variable that best attributed success in the UCL rather than titles won to better incorporate all our observations even though the latter is almost always seen as the best determinant of success in any sports tournament.

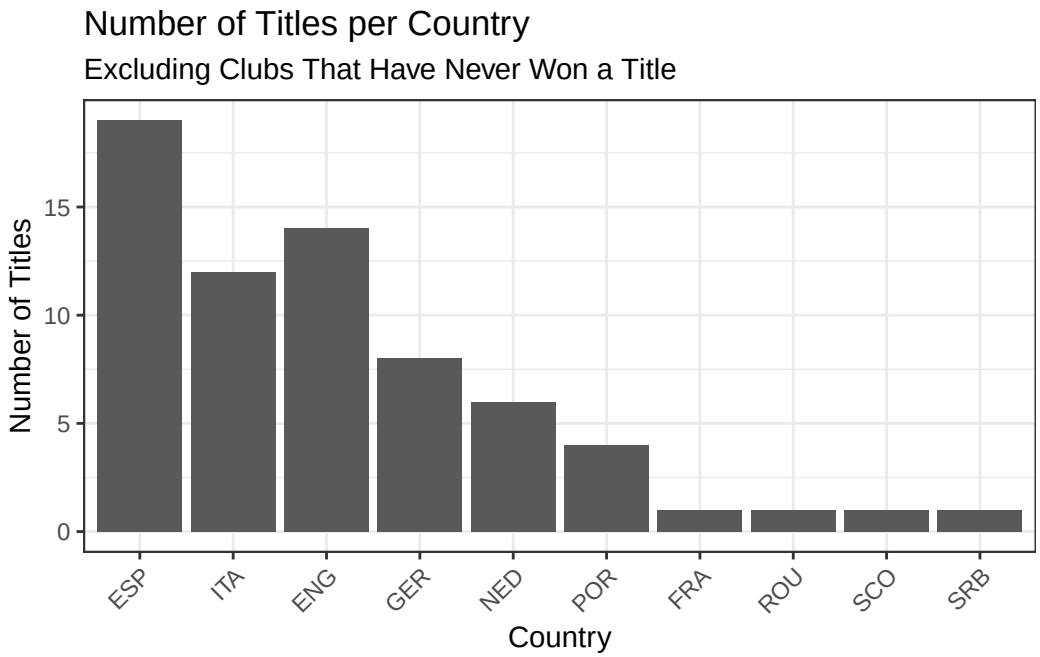
Future studies could be designed to address some of these limitations. First, more rigorous research could be conducted to incorporate each club's participating seasons in the UCL into the dataset to account for potential rule changes over time. Second, a future dataset could incorporate a new variable that averages how long. Second, perhaps a study only incorporating the clubs who have won the UCL could examine the relationship between the same predictors we chose and a new response variable called "title success" (number of titles divided by number of tournaments participated in). This analysis could enable researchers to better understand how factors such as offensive prowess, defensive stoutness, and the ability to secure wins each associate with a team's title success.

Appendix

The plot below shows the number of titles per club, excluding clubs that have never won a title. As shown here, this would severely cut down our dataset to only 22 observations, which is why we decided to not define number of titles as "success".



The plot below shows the number of titles per country, excluding clubs that have never won a title. This shows that Spain, England, Germany, Italy, and the Netherlands have clubs that have won the most UCL titles. Spain, England, Germany, Italy, and France are all considered the most talented leagues, showing why we created a variable to separate the top five leagues from other countries.


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