Champions League Lab

Rajeev Roy and Arthur Watcharananan

2022-10-29

Introduction

The Champions League is an annual soccer competition in which the top teams in Europe fight for the title of best team in Europe. This data set consists of 12 variables to help us answer our question. According to the Champions League Data set which country performed the best and which country performed the worst?

In this lab we will use different variables, such as: Countries, Wins, Loss, Goal Differential, and Participation. These variables will aid to compare different countries and their performance in different categories in an effort to identify the best performing country and the worst performing country.

Definition of Each Variable

- Club: The soccer club that participated in the Champions league
- Country: The country that the soccer club is from. For example, Real Madrid plays in the Spanish league so its country would be ESP.
- Participated: The number of times a club participated in the Champions League.
- Titles: The number of times the team won a Champions League title.
- Played: Number of individual games a team played in the Champions League.
- Wins: Number of times a team won in the Champions League
- Draw: Number of times a team drew in the Champions League
- Loss: Number of times a team lost in the Champions League
- Goals For: Number of goals a team scored in the Champions League
- Goals Against: Number of goals a team conceded in the Champions League
- Points: Total Number of points gained from competing in the Champions League. Win = 3 points, Draw = 1 point, Lose = 0 points.
- Goal Diff: The difference between goals scored and goals conceded. A positive goal difference indicates that a team has scored more goals than has conceded.

Observations & Analysis

Part 1: Background Information

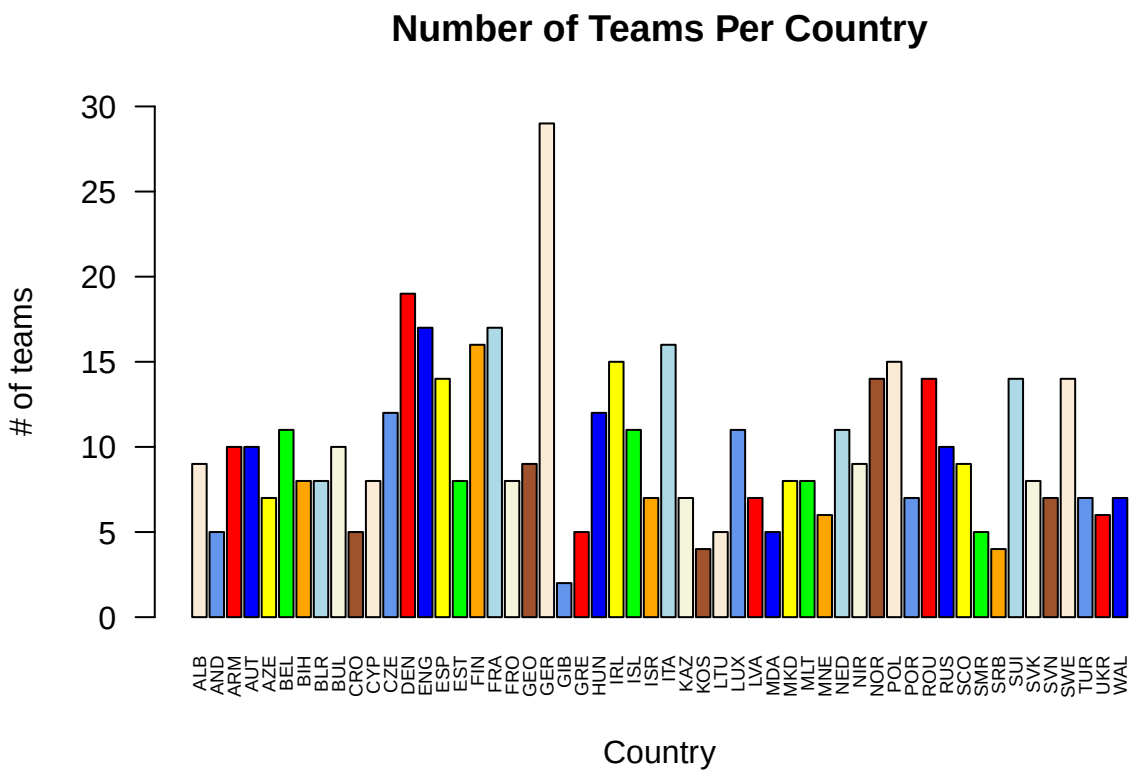
The first step to analyzing the Champions League data to be able to answer our question in to read in our data file...

```
champs <- read.csv("AllTimeRankingByClub.csv")
head(champs, 15)
```

```
##    Position                    Club Country Participated Titles Played Win
## 1         1           Real Madrid CF     ESP           52     14    464 277
## 2         2     FC Bayern M\x9fnchen     GER           38      6    372 221
```

```
## 3         3             FC Barcelona     ESP           32      5    333 195
## 4         4        Manchester United     ENG           30      3    293 160
## 5         5                  Juventus     ITA           36      2    295 152
## 6         6             Liverpool FC     ENG           26      6    240 137
## 7         7                 AC Milan     ITA           29      7    255 126
## 8         8               SL Benfica     POR           41      2    273 120
## 9         9                 FC Porto     POR           36      2    261 117
## 10       10                 AFC Ajax     NED           38      4    241 110
## 11       11           FC Dynamo Kyiv     UKR           38      0    248 101
## 12       12               Chelsea FC     ENG           18      2    191  99
## 13       13               Arsenal FC     ENG           21      0    201 101
## 14       14                Celtic FC     SCO           36      1    216 101
## 15       15 FC Internazionale Milano     ITA           23      3    192  91
##    Draw Loss Goals.For Goals.Against Pts Goal.Diff
## 1    79  108      1021           508 633       513
## 2    75   76       782           367 517       415
## 3    75   63       655           331 465       324
## 4    69   64       533           284 389       249
## 5    70   73       470           288 374       182
## 6    50   53       453           216 324       237
## 7    65   64       422           240 317       182
## 8    64   89       437           319 304       118
## 9    60   84       383           296 294        87
## 10   64   67       385           266 284       119
## 11   54   93       345           308 256        37
## 12   52   40       330           172 250       158
## 13   43   57       332           218 245       114
## 14   37   78       333           255 239        78
## 15   51   50       271           193 233        78
```

```r
barplot(table(champs$Country),
        main = "Number of Teams Per Country",
        xlab = "Country",
        ylab = "# of teams",
        cex.names = 0.6,
        las = 2,
        ylim = c(0,30),
        col = c("antiquewhite", "cornflowerblue", "red", "blue", "yellow", "green",
                "orange", "lightblue", "beige", "sienna")
        )
```



In this barplot, we can see the number of Champions League participants by country. We can clearly see that throughout the history of the Champions League, Germany had the most teams participating, with Denmark coming second. Looking at this graph, we realized that there were way too many countries to analyze, and some of the countries had so little participation that they were not worth analyzing. Because of this, we decided to only keep the top ten countries with the most teams. Now, because of this change a new question has been formulated, which country out of the top 10 participating countries performed the best and which country performed the worst all time in the Champions League?

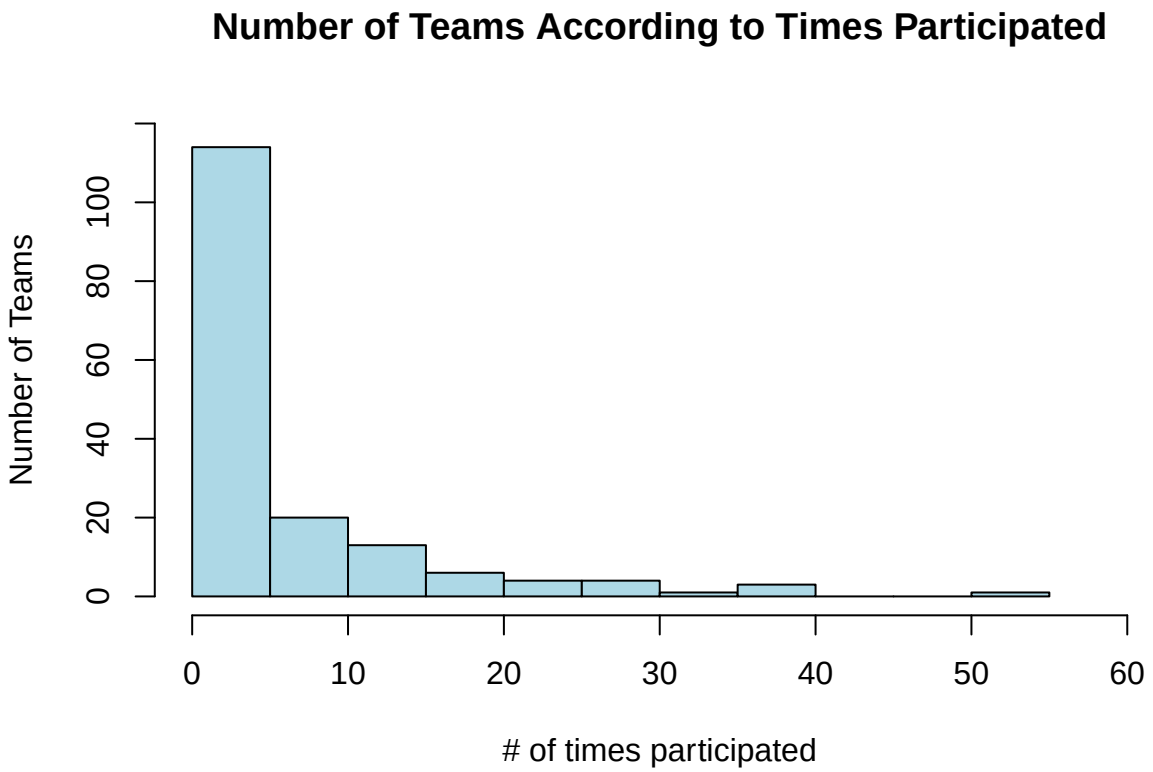
```
champs.top.ten <- champs[(champs$Country=="GER") | (champs$Country=="DEN")|
(champs$Country=="ESP")| (champs$Country=="ITA")| (champs$Country=="FRA")|
(champs$Country=="ENG")| (champs$Country=="POL")| (champs$Country=="SUI")|
(champs$Country=="NOR")| (champs$Country=="NED"), ]

champs.top.ten$Country <- as.factor(champs.top.ten$Country)
```

By running this command, we were able to eliminate the smaller countries, and keep data for the ten countries we thought were interesting to analyze. This is called subsetting. The countries that we kept were Germany, Denmark, Spain, Italy, France, England, Poland, Switzerland, Norway, and the Netherlands. After subsetting the dataset into the vector "champs.top.ten", we changed the variable "Country" into a vector factor because it is a category that isn't specific, and is repeated multiple times.

The first thing we wanted to do was find out how often most teams participated in the champions league. This would help us with our analysis because we would be able to better understand why some of our graphs may have outliers.

```
hist(champs.top.ten$Participated,
     main = "Number of Teams According to Times Participated",
     xlab = "# of times participated",
     ylab = "Number of Teams",
     ylim = c(0,120),
     xlim = c(0,60),
```

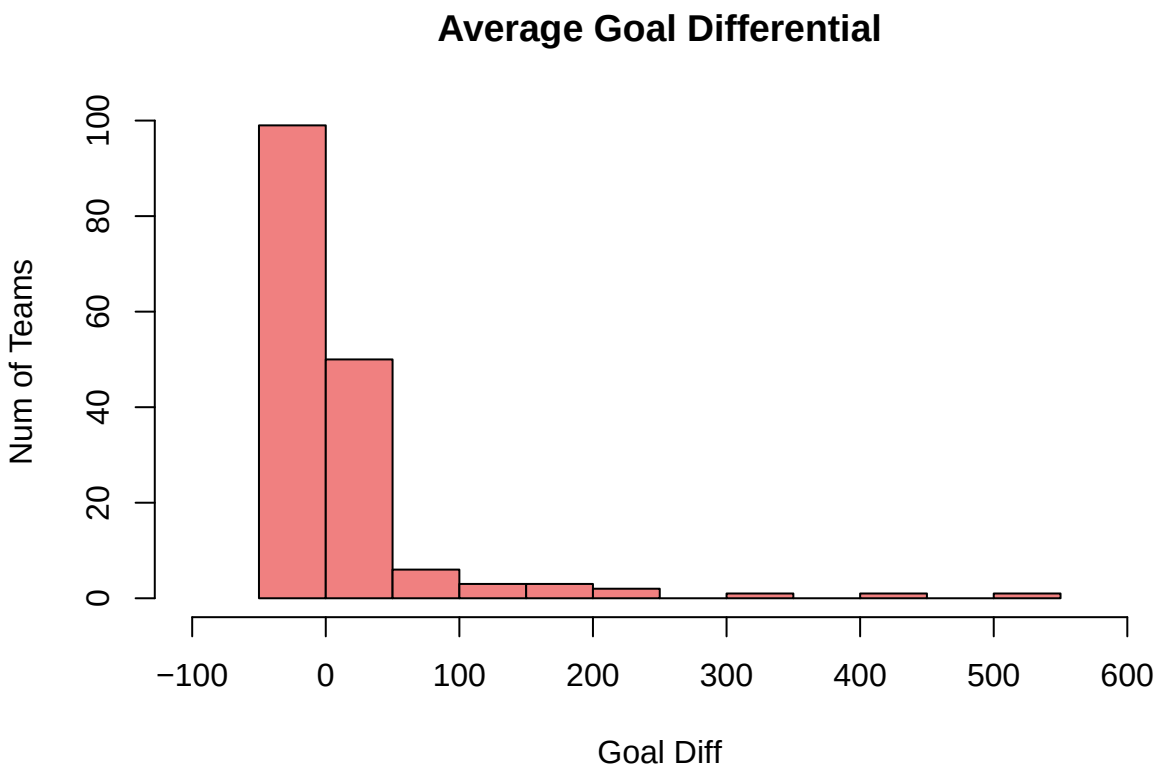
```
     col = "lightblue"
     )
```



With the data of the top 10 countries, we used the command hist() to make a histogram showing how often most teams participated in the champions league. By graphing the histogram, we are able to see where the median lies in the graph, and how skewed the graph is. The graph of Number of Teams According to Times Participated is a unimodal histogram skewed to the right. The median of the data set seems to lie between 1-5 times participated, and there are very noticeable outliers. This indicates that most teams in the Champions League have participated around 1-5 times, and rarely any teams go beyond 10 times. From this we can conclude that no single team would single handedly contribute to a country's performance but rather a country's performance is represented by the performance of all teams within the country.

The second histogram we wanted to examine was the average goal differential for teams in the Champions League.

```
##goal differential histogram by top 15 (average goal differential)
hist(champs.top.ten$Goal.Diff,
     main = "Average Goal Differential",
     xlab = "Goal Diff",
     ylab = "Num of Teams",
     xlim = c(-100, 600),
     col = "lightcoral"
)
```



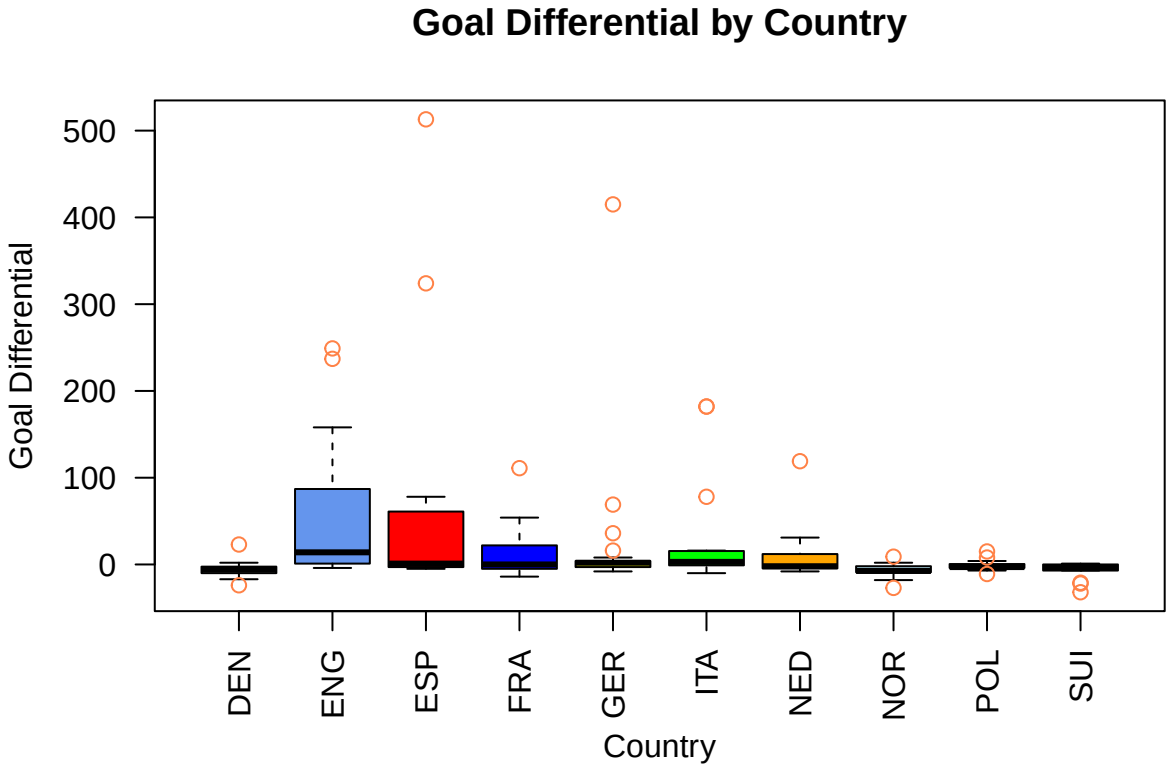
The same principles apply with this histogram. This histogram is a unimodal histogram that is skewed to the right. The median of this histogram seems to lie at around -50 - 0 Goal Differential. This means that most teams that have participated in the Champions League have a negative goal differential, with only a few teams with a positive goal differential.

After finding out how many times teams participated and the average goal differentials, we were able to move on to answer our question: Which country performed the best and which country performed the worst?

Part 2: Who is the Best and Who is the Worst?

The first independent variable we examined was Goal Differential. In theory, Goal Differential is a very accurate measure of how well a team performs in the Champions League because a team that scores more goals than concedes is likely to win more games. For example, if your goal differential is 30, you have scored 30 more goals than have conceded, which probably means you won a lot more games than lost. The graph below shows comparative boxplots on the Goal Differential by country.

```
boxplot(champs.top.ten$Goal.Diff~champs.top.ten$Country,
        main = "Goal Differential by Country",
        ylab = "Goal Differential",
        xlab = "Country",
        col = c("antiquewhite", "cornflowerblue", "red", "blue", "yellow", "green",
                "orange", "lightblue", "beige", "sienna"),
        outcol = "sienna1",
        las = 2)
```

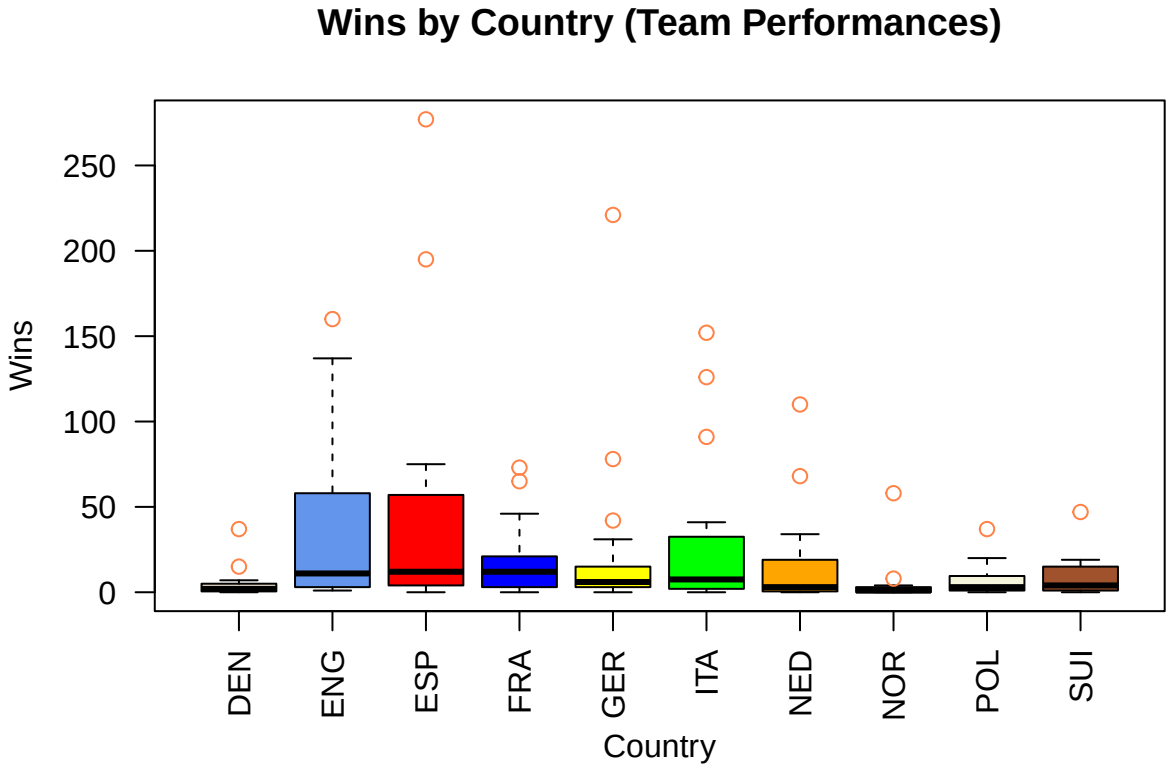


The boxplot above shows the distribution of Goal Differential by Country. Looking at this graph, it is clear that on average, teams from England have the highest Goal Differential. Spain comes in a close second, France third, etc. Although English teams have the highest mean Goal Differential, there are two teams from Spain that have higher Goal Differentials than all teams from England. Furthermore, there is also a team from Germany with a higher Goal Differential than teams from England. Although these outliers are impressive, they have no effect on their country's overall performance.

If we were answering the question "From which country did the team with the best Goal Differential come from," then the clear answer would be Spain; however, we are focusing on ranking the countries as a whole, not a specific club. Consequently, we determined that countries from England have the best overall Goal Differentials. On the other end of the spectrum we see that Norway with a very low lower fence is the worst performing team in terms of Goal differential with a very similar performance to Denmark. This tells us that the bottom 2 teams in terms of Goal Differential are Denmark and Norway.

In order to further examine our question, we decided to use Wins as an independent variable. Obviously, wins is an important variable as teams that win more games are more likely to reach higher stages of the tournament.

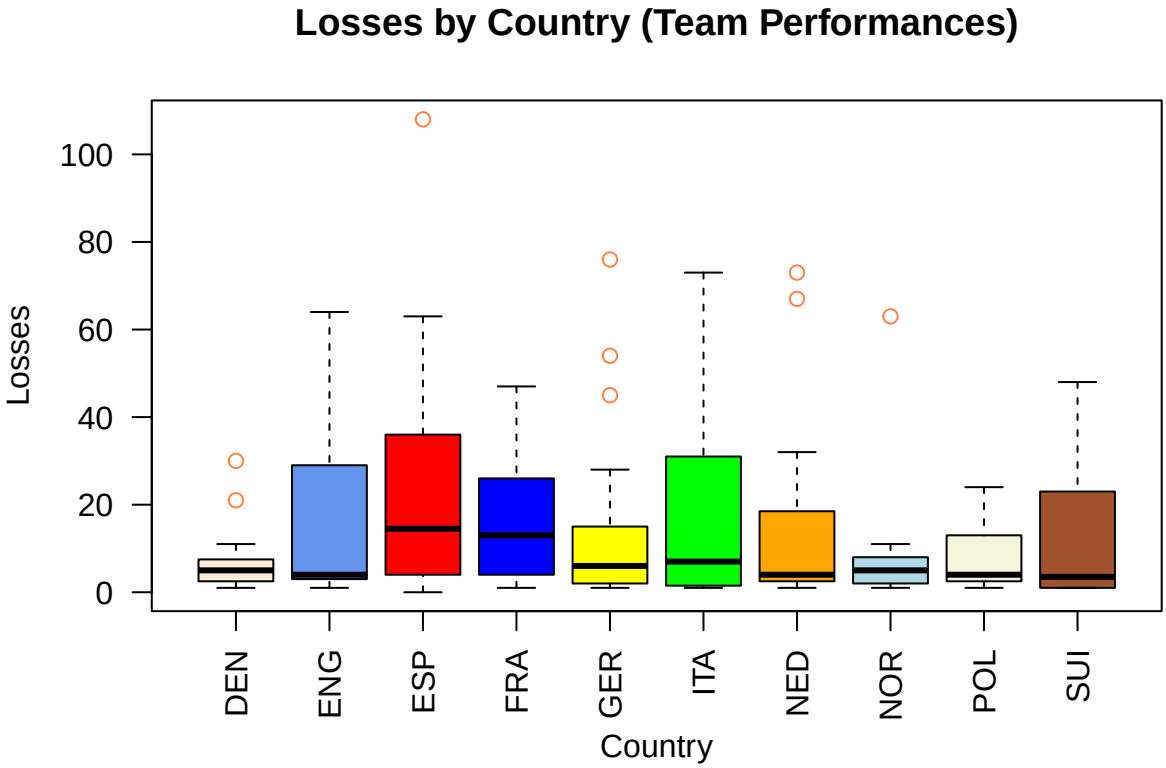
```
boxplot(champs.top.ten$Win~champs.top.ten$Country,
        main = "Wins by Country (Team Performances)",
        ylab = "Wins",
        xlab = "Country",
        col = c("antiquewhite", "cornflowerblue", "red", "blue", "yellow", "green",
                "orange", "lightblue", "beige", "sienna"),
        outcol = "sienna1",
        las = 2)
```



For this variable, we also used comparative boxplots to visualize our data. Looking at the graph above, we can clearly see that England and Spain continue to dominate other countries by having the most wins. Although the interquartile range of both countries are pretty similar, England seems to have a higher upper fence; therefore, we can conclude that England may have more teams with win rate than Spain does. For this boxplot specifically, we determined that teams from England had the most wins, with Spain coming in a close second, Italy third, etc. Additionally, we were able to determine that teams from Norway and Denmark have the least amount of wins with Norway performing the worst.

The last independent variable that we used for our lab research was "Losses." Examining "Losses" will help us determine the best and worst teams because a team that loses more games should theoretically be worse than a team that loses less.

```
boxplot(champs.top.ten$Loss~champs.top.ten$Country,
        main = "Losses by Country (Team Performances)",
        ylab = "Losses",
        xlab = "Country",
        col = c("antiquewhite", "cornflowerblue", "red", "blue", "yellow", "green",
                "orange", "lightblue", "beige", "sienna"),
        outcol = "sienna1",
        las = 2)
```



Looking at the boxplot above, it is evident that teams from Spain, England, Italy, and France had the most losses. Although the interquartile range for England is pretty big, the median is pretty small, meaning that the average number of losses for clubs in England is most likely less than Spain, France, and Italy. This is a major reason why this is a misleading graph. This boxplot also indicates that Denmark and Norway are the countries with the least losses as they have a small interquartile range with the median at below 10 losses. From this we can not extrapolate any useful information without diving deeper into the proportionality of games played to wins, draws, and losses.

Conclusion

Over the course of analyzing variables such as participation, goal differential, wins, and losses we can come to a conclusion that the best performing country in the Champions League of all time is England. England has the best overall performance when it comes to both goal differential as well as wins. While analyzing, we realized that the worst teams in the top 10 participating countries would be from Denmark. Denmark was the second highest participating country but lacked in its performance in wins and goal differential. Although Norway was a close runner up for last place, Denmark had a higher participation count in the competition. This is why we believe they are the worst performing country out of top 10 participating countries in the Champions League.

On a separate side note after analyzing data of all countries we can come up with a rough estimate of the rankings of the 10 countries analyzed in this lab...

1) England
2) Spain
3) France
4) Italy
5) Netherlands
6) Germany

7) Poland
8) Sweden
9) Norway
10) Denmark