# World Cup 2022

Jack Miller
Markus Miller

# Our Goal

- What separates the teams that advance from their group from those who don’t?
- Could we help international managers do their jobs using simple data analysis?
- How can we use the tools learned in this class to determine the factors that help teams succeed at the World Cup?

- Goal: Find two or three statistics that are strong predictors for advancing from a World Cup group.

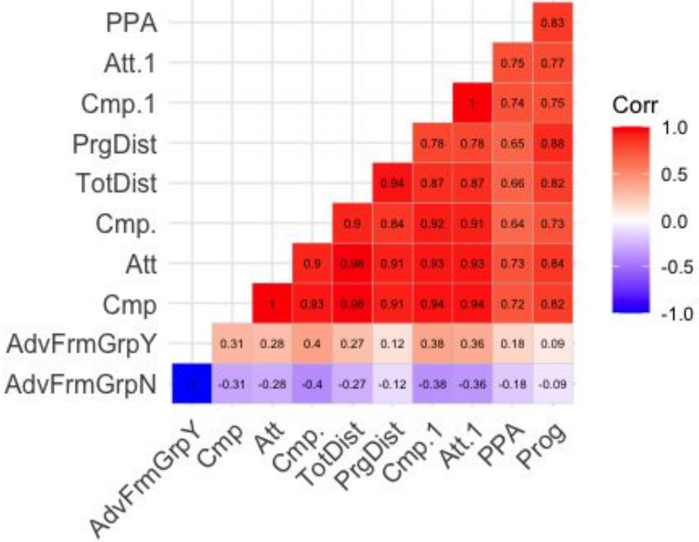
# FBREF - Our Data Source

- FBRef or fbref.com is a central site that contains both team and individual stats for all soccer (or football) competitions around the world. This includes advanced statistics and rate statistics, which we used to analyze the World Cup, since some teams have played more minutes and more games than others.
- Used their World Cup Squad Stat tables: Shooting, Passing, Advanced Goalkeeping, Possession.
- We used “per 90” statistics rather than totals because teams that have advanced from their group have played more matches than those didn’t
    - Per 90 statistics average a statistic over the 90 minute length of a match (e.g. total shots/90 vs. total shots)
- *Note: This project used World Cup data as of December 10, 2022*

SPORTS
REFERENCE

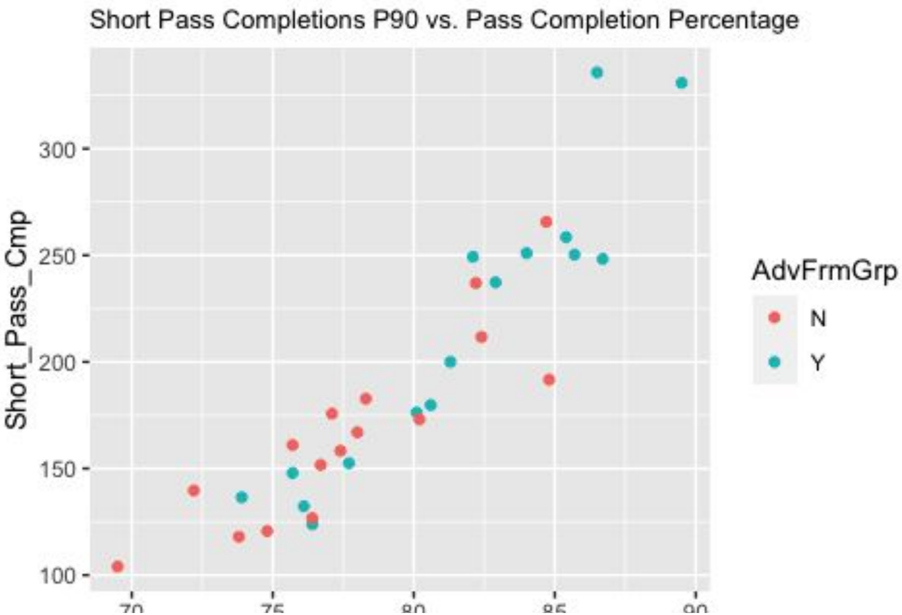
# Initial Exploration

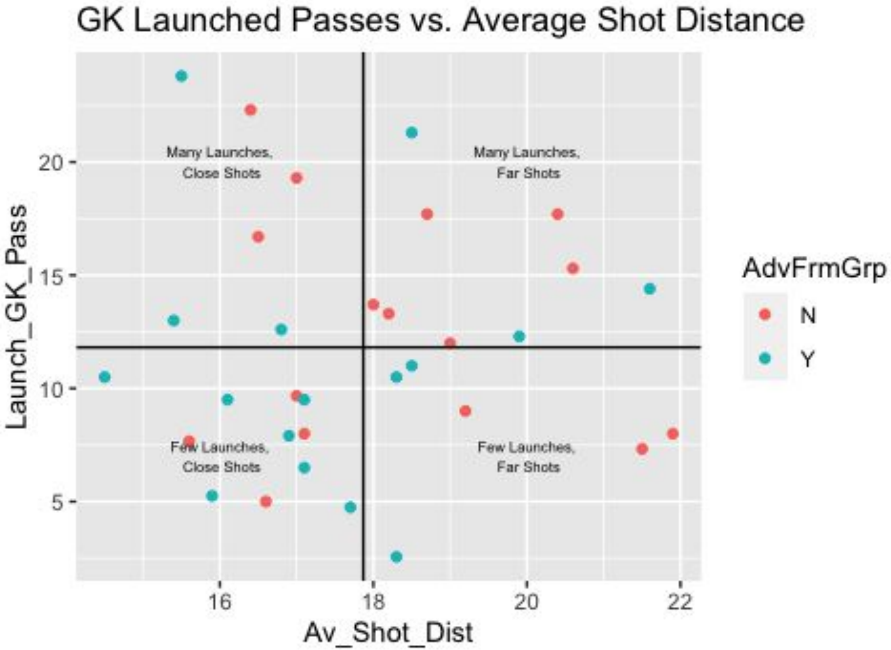
- We browsed FBREF for interesting statistics and data tables.
- Given the large number of statistics available, it was impossible to look at them all. We had to narrow down the list to 5 or 6 interesting statistics per table based on our own soccer knowledge.
- We created a few basic visualizations to aid our process, including the correlation matrix shown in the bottom right. (this one is a selection of passing statistics)
- This was the part of the process that took the longest. It involved searching and sorting through a high number of variables.



# Visualizations

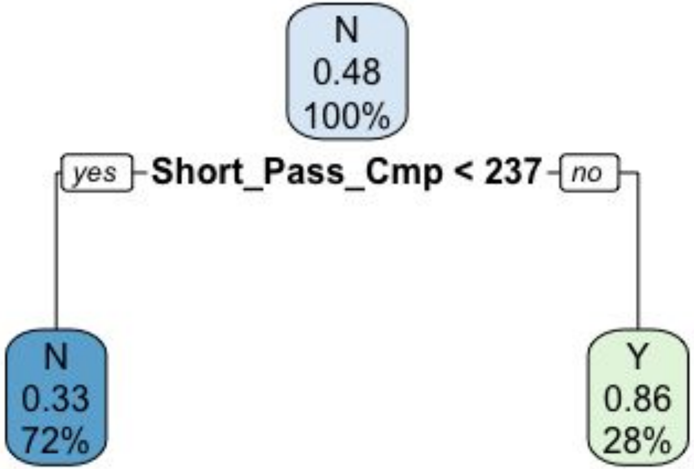
- We made correlation matrices for each of our four data tables, and took one or two stats from each that were most strongly correlated with advancing from the group stage.
- We ended up narrowing down to the following statistics: Pass Completion %, Short Pass Comp, Shot Distance, Total Touches, Att 3rd Touches, Launched Pass Att, PSxG+/-

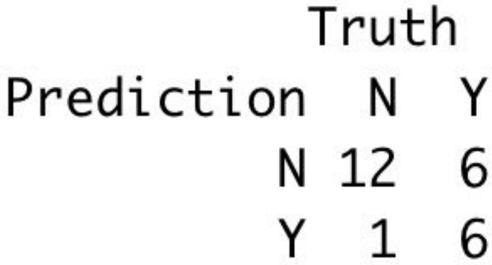




# Decision Tree

- We realized that a decision tree could be perfect for this problem. A decision tree can basically show you the predictive significance of variables, which is exactly what we want.
- And, what's more, it also parameterizes variables, which adds another layer of depth to the solution. So, instead of just saying that more short pass completions predict success, it can give a specific threshold that determines success (e.g. more than 237 short pass completions).
- Our decision tree didn't end up being very sophisticated, as shown to the right.
- However, accuracy was 72% for training data and 71% for testing data.





```
          Truth
Prediction  N  Y
         N 12  6
         Y  1  6
```

# Conclusion

- We dove into this project with a lot to sort through.
- After searching for a while, we're not sure our findings are particularly useful for understanding *how* to advance out of your group, but we do know they help explain the story that has taken place in the 2022 World Cup.
- This may not be directly translatable onto the field, but by sorting through so many different variables and running correlation matrices we were able to see how strongly certain variables interacted with other variables, and how strongly they interacted with the binary variable, "advanced."
- We do have a better picture now of what winning soccer (football) on the international stage looks like, which is important.