**Codes R Us Final Project Proposal**

Date: 4/2/2022

**Research Topic:** Model goal scoring probabilities and win probabilities using soccer match data from the 2017/2018 season for Europe’s top five leagues (English Premiere League, Spanish La Liga, German Bundesliga, Italian Serie A, and French Ligue Un), the 2016 European Championship, and the 2018 World Cup.

**Dataset:** The dataset contains all matches played in Europe’s top five leagues during the 2017/2018 season (380 matches per league for a total of 1900 matches, producing 643,149 match events), the 2016 European Championship (51 matches, producing 78,139 match events), and the 2018 World Cup (64 matches, 101,758 match events). It was obtained from the work posted by [Luca Pappalardo on Figshare](https://figshare.com/collections/Soccer_match_event_dataset/4415000).

In this dataset, each event happening in a match is expressed as a particular row. Each match may have hundreds of different events, which include, but are not limited to, Shots, Key Passes, Assists, and Goals. One key feature of the dataset is that it allows us to see the relative (X, Y) coordinates of the field from which an event begins and ends. It also tells, through different variables, if an event resulted in a successful goal or not. (DAN)

**Smart Questions:**

1. ~~Which match variables result in the highest probability of a shot on goal being successful and can we successfully use them to predict a shot being successful or not?~~ Given the variables that describe the (X, Y) coordinates of the beginning and end of an event, which configuration of the variables result in the best fitted model that explains the probability of a Goal being successful?

1. Based on match summary variables (i.e., percentage of possession of the ball, number of shots, number of shots on target, being home team vs being visitor team, etc.), can we predict who the winner of a particular match was?

**Github Repo:** [rmesina /T2\_Codes-R-Us](https://github.com/rmesina/T2_Codes-R-Us)