Could an Independent Yorkshire Win the World Cup - Data & Scraping

Robert Hickman

2018-06-07

Recently, a Yorkshire national football team appeared in a league of national teams for stateless people. This got me wondering how the historic counties of the UK would do at the world cup. Could any of them compete with full international teams?

This is the complete script for an short article I wrote for CityMetric on the topic. It's split over 6 separate parts and is pretty hefty but contains pretty much everything you need to clone the article.

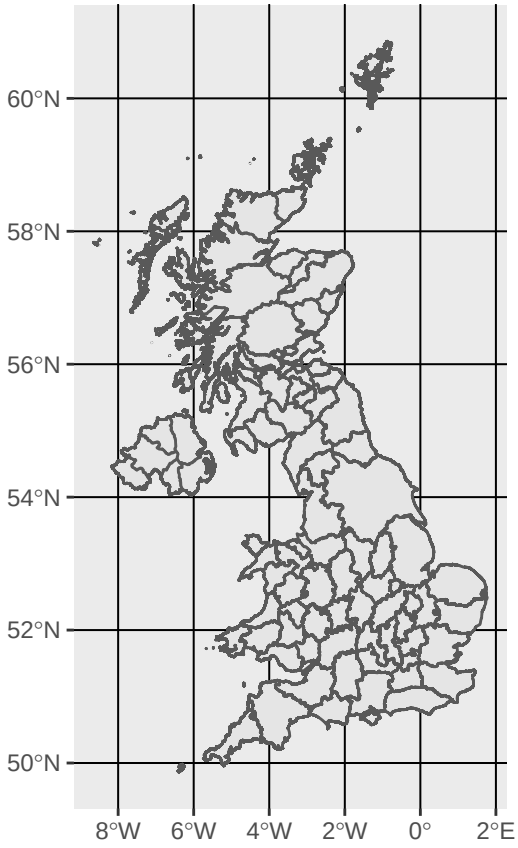
To start, there are 95 historic counts of Great Britain + the 6 counties of Northern Ireland which I included for completeness. These are of a wide variety of sizes and approximate population and demographic, so it's not clear how each would do simply from inspection.

The data for this comes from Boundary Line and the Northern Irish Boundary Database, I've put them together already, but its simple enough to do it in sf.

```
library(dplyr)
library(rvest)
library(data.table)
library(ggplot2)
```

```
p <- ggplot(data = uk_counties) +
  geom_sf()

plot(p)
```



In order to calculate how good each county team would be, I needed a measure of the ability of all of the players they could field. For this I turned to the FIFA18 video game which rates players along a variety of scales.

Scrape Player Data

To get data on every player in the game I wrote a quick scraping function. This finds the links to every player on all 602 pages of https://www.fifaindex.com/players/ and then downloads all the data required on each player.

```
#both steps here take a fair amount of time
#about 10mins and an hour respectively

#get the links to each players page
all_player_links <- unlist(lapply(paste0("https://www.fifaindex.com/players/", 1:602), function(x) {
  player_link <- read_html(x) %>%
    html_nodes("td:nth-child(4) a") %>%
    html_attr("href")
  })) %>%
  paste0("https://www.fifaindex.com/", .)

#big function to scrape every piece of data we could want of each players page
get_player_data <- function(link) {
  #read the players web page
  read <- read_html(link)
```

```
  #basic data
  name <- read %>%html_nodes(".big") %>% html_text()
  club <- read %>% html_nodes(".panel-title a+ a") %>% html_text() %>% .[length(.)]
  if(length(club) == 0) {
    club <- NA
  }
  nationality <- read %>% html_nodes(".subtitle a") %>% html_text()

  #general info on the player
  height <- read %>% html_nodes(".col-lg-5 p:nth-child(1) .pull-right") %>% html_text() %>%
    gsub(" cm", "", .) %>% as.numeric()
  weight <- read %>% html_nodes(".col-lg-5 p:nth-child(2) .pull-right") %>% html_text() %>%
    gsub(" kg", "", .) %>% as.numeric()
  foot <- read %>% html_nodes(".col-lg-5 p:nth-child(3) .pull-right") %>% html_text()
  birthdate <- read %>% html_nodes(".col-lg-5 p:nth-child(4) .pull-right") %>% html_text() %>%
    as.Date("%m/%d/%Y")
  age <- read %>% html_nodes(".col-lg-5 p:nth-child(5) .pull-right") %>% html_text() %>%
    as.numeric()
  main_position <- read %>% html_nodes("body > div.container.main > div:nth-child(3) > div.col-md-8 > di
    .[1] %>% html_attr("title")
  work_rate <- read %>% html_nodes(".col-lg-5 p:nth-child(7) .pull-right") %>% html_text() %>%
    str_split(., " / ") %>% unlist()

  #the players rating for each skill
  ratings <- read %>% html_nodes(".rating") %>% html_text() %>% as.numeric() %>%
    as.matrix() %>% t() %>% as.data.frame()
  names(ratings) <- c("overall","specific","ball_control","dribbling","marking","slide_tackle","stand_ta
                      "aggression","reactions","positioning","interceptions","vision","composure","cross
                      "short_pass","long_pass","acceleration","stamina","strength","balance","sprint_spe
                      "agility","jumping","heading","shot_power","finishing","long_shots","curve","free_
                      "penalties","volleys","gk_positioning","gk_diving","gk_handling","gk_kicking","gk_

  #stick everything into a dataframe to be output
  df <- data.frame(name = name, club = club, nationality = nationality,
                   height = height, weight = weight, foot = foot, birthdate = birthdate, age = age,
                   main_position = main_position, work_rate1 = work_rate[1], work_rate2 = work_rate[2])
    cbind(ratings)
  return(df)
}

#scrape the info on all players
all_players_data <- rbindlist(lapply(all_player_links, get_player_data)) %>%
  setDT() %>%
  #add an id column for each player
  .[, id := 1:.N]
```

Once that's scraped and bound we can take a peek at the data. There's 18k players in total and 48 variables for each so we'll just look at a few for now.

```
#show a selection of the some key info for each player
#the id we gave them, their name, nationality, and their overall ability
head(select(all_players_data, id, name, nationality, overall))
```

```
##    id              name nationality overall
```

```
## 1:  1      Lionel Messi   Argentina      94
## 2:  2 Cristiano Ronaldo    Portugal      94
## 3:  3            Neymar      Brazil      92
## 4:  4       Luis Suárez     Uruguay      92
## 5:  5      Manuel Neuer     Germany      92
## 6:  6            De Gea       Spain      91
```

Over the course of the next posts, we'll use this data to calculate a player's ability in any position on the field. This will then be used to select optimal teams for each nation (or each historic British county). Finally we'll take the average ability of these optimal teams and use them to simulate the World Cup to get the chance each team has to win the tournament.