The Knowledge 4th August 2018

Robert Hickman

2018-08-04

The Guardian publish a weekly set of questions and answers on a variety of football minutiae at The Knowledge. Forutnately, some of these are extremely tractable using R, so I thought I'd have a go at working through the archives to see if I can shed light on any of the questions.

```
library(rvest)
library(dplyr)
library(magrittr)
library(data.table)
library(zoo)
library(ggplot2)
library(rvest)
library(stringr)

#jalapic/engsoccerdata
library(engsoccerdata)
```

We Ain't Going To The Town..

'This season, Tranmere Rovers return to contest League Two alongside eight teams with the suffix Town, including six successive fixtures against these clubs over the New Year. What is the record for successive fixtures versus clubs with the same (or no) prefix or suffix?'

For this question I decided to ignore prefixes as the dataset I'm using doesn't have any that could be matches between teams except the 'West' in West Ham and West Bromwich Albion. That dataset is the excellent engsoccerdata from James Curley found at his github here and on CRAN.

```
#take all of the english soccer data in the package and bind it together
england_data <- bind_rows(
    select(engsoccerdata::england,
           .data$home, .data$visitor, date = .data$Date),
    select(engsoccerdata::englandplayoffs,
           .data$home, .data$visitor, date = .data$Date),
    select(engsoccerdata::england1939,
           .data$home, .data$visitor, date = .data$Date)) %>%
  setDT() %>%
  #convert the date to date class
  .[, date := as.Date(date)]

#get a list of each unique team in the dataset
all_teams <- unique(c(as.character(england_data$home),
                      as.character(england_data$visitor)))

#melt the dataset by each teams matches
find_chains <- rbindlist(lapply(all_teams, function(team) {
  england_data %>%
    .[home == team | visitor == team] %>%
    .[, matching_team := team]
```

```
  })) %>%
  .[home == matching_team, other := visitor] %>%
  .[visitor == matching_team, other := home] %>%
  .[, c("date", "matching_team", "other")] %>%
  #get the suffixes and prefixes of the other team
  .[, other_prefix := gsub(" .*", "", other)] %>%
  .[, other_suffix := gsub(".* ", "", other)] %>%
  #arrange by team and date
  .[order(matching_team, date)] %>%
  #convert to an id
  .[, suffix_id := as.numeric(as.factor(other_suffix))] %>%
  #if playing consecutively against the same suffix id (ignoring prefixes for now) put in same 'chain'
  .[, match := suffix_id - lead(suffix_id), by = "matching_team"] %>%
  .[match == 0 & lead(match) != 0, chain_id := 1:.N] %>%
  .[match == 0] %>%
  .[, chain_id := na.locf(chain_id, fromLast = TRUE)] %>%
  .[, chain_length := .N, by = chain_id] %>%
  #take only chains at least as long as Tranmere's run (6)
  .[chain_length > 5] %>%
  .[order(chain_length)] %>%
  .[, c("date", "matching_team", "other", "chain_length")]

#print the chains of equal length to Tranmere's run
print(find_chains)
```

```
##           date  matching_team              other chain_length
##  1: 1950-12-30   Chesterfield      Leicester City            6
##  2: 1951-01-13   Chesterfield     Manchester City            6
##  3: 1951-01-20   Chesterfield       Coventry City            6
##  4: 1951-02-03   Chesterfield        Cardiff City            6
##  5: 1951-02-17   Chesterfield     Birmingham City            6
##  6: 1951-02-24   Chesterfield        Swansea City            6
##  7: 2009-03-21 Leicester City   Colchester United            6
##  8: 2009-03-28 Leicester City Peterborough United            6
##  9: 2009-04-04 Leicester City     Carlisle United            6
## 10: 2009-04-11 Leicester City     Hereford United            6
## 11: 2009-04-13 Leicester City        Leeds United            6
## 12: 2009-04-18 Leicester City     Southend United            6
## 13: 1921-05-02         Fulham           Hull City            7
## 14: 1921-05-07         Fulham           Hull City            7
## 15: 1921-08-27         Fulham       Coventry City            7
## 16: 1921-08-29         Fulham      Leicester City            7
## 17: 1921-09-03         Fulham       Coventry City            7
## 18: 1921-09-05         Fulham      Leicester City            7
## 19: 1921-09-10         Fulham           Hull City            7
## 20: 1920-04-17  Leyton Orient     Birmingham City            7
## 21: 1920-04-24  Leyton Orient     Birmingham City            7
## 22: 1920-04-26  Leyton Orient      Leicester City            7
## 23: 1920-05-01  Leyton Orient      Leicester City            7
## 24: 1920-08-28  Leyton Orient      Leicester City            7
## 25: 1920-08-30  Leyton Orient        Cardiff City            7
## 26: 1920-09-04  Leyton Orient      Leicester City            7
## 27: 1920-10-09   Notts County          Stoke City            7
## 28: 1920-10-16   Notts County          Stoke City            7
```

```
## 29: 1920-10-23   Notts County        Cardiff City            7
## 30: 1920-10-30   Notts County        Cardiff City            7
## 31: 1920-11-06   Notts County       Coventry City            7
## 32: 1920-11-13   Notts County       Coventry City            7
## 33: 1920-11-20   Notts County      Leicester City            7
##           date  matching_team               other chain_length
```

so In fact an identical length chain on matching suffixes has occured twice, with Chesterfield playing a range of cities at the start of 1951 in League Two, and much more recently, Leicester playing 6 different Uniteds in a row at the tail end of the 2008/2009 season. This is also the season that saw them recover from being relegated from the Chmapionship and start moving towards winning the title in 2015-2016 season.

Some longer chains involving cities happened in the 1920-1921 seasons in the Second Division, but it seems like the scheduling worked differently then and teams played back to back more, so doesn't really count.

Having originally misread the question, I also wanted to find out the longest chain of a team playing teams that matched *their own* suffix. We can do this using a similar method

```
matching_fixtures <- england_data %>%
  #get only matches between teams with matching prefix/suffixes
  .[, home_suffix := gsub(".* ", "", home)] %>%
  .[, away_suffix := gsub(".* ", "", visitor)] %>%
  .[home_suffix == away_suffix, match := home_suffix] %>%
  .[!is.na(match)] %>%
  #remove matches where teams from the same city play each other
  .[!match %in% c("Bradford", "Bristol", "Burton", "Manchester", "Sheffield")]

#get all the teams that have played teams with matching suffixes
matching_teams <- unique(c(as.character(matching_fixtures$home),
                           as.character(matching_fixtures$visitor)))

#elongate the data and look for chains
find_chains <- rbindlist(lapply(matching_teams, function(team) {
  england_data %>%
    .[home == team | visitor == team] %>%
    .[order(date)] %>%
    .[, matching_team := team]
  })) %>%
  .[home == matching_team, other := visitor] %>%
  .[visitor == matching_team, other := home] %>%
  #id matches and remove matches not involving teams with identical suffixes
  .[, match_id := 1:.N, by = matching_team] %>%
  .[!is.na(match)] %>%
  #find chains of identical suffixed matches
  .[, chain := match_id - lag(match_id)] %>%
  .[chain == 1 & lag(chain) != 1, chain_id := 1:.N] %>%
  .[chain == 1] %>%
  .[, chain_id := na.locf(chain_id)] %>%
    .[, chain_length := .N, by = chain_id] %>%
  #take only chains at least as long as Tranmere's run (6)
  .[chain_length > 4] %>%
  .[order(chain_length)] %>%
  .[, c("date", "matching_team", "other", "chain_length")]

print(find_chains)
```

```
##           date   matching_team             other chain_length
##  1: 1919-12-13      Stoke City   Birmingham City            5
##  2: 1919-12-20      Stoke City    Leicester City            5
##  3: 1919-12-25      Stoke City     Coventry City            5
##  4: 1919-12-26      Stoke City     Coventry City            5
##  5: 1919-12-27      Stoke City    Leicester City            5
##  6: 1919-09-01       Hull City        Stoke City            5
##  7: 1919-09-06       Hull City   Birmingham City            5
##  8: 1919-09-08       Hull City        Stoke City            5
##  9: 1919-09-13       Hull City        Leeds City            5
## 10: 1919-09-20       Hull City        Leeds City            5
## 11: 1988-09-24 Carlisle United  Rotherham United            5
## 12: 1988-09-30 Carlisle United  Cambridge United            5
## 13: 1988-10-04 Carlisle United Colchester United            5
## 14: 1988-10-08 Carlisle United   Hereford United            5
## 15: 1988-10-15 Carlisle United    Torquay United            5
```

So the record for that is only slightly shorter! with Stoke and Hull City playing a range of cities in the 1919-1920 season (but see above for scheduling differences) and Carlisle United playing 5 other different Uniteds in a row in the old Fourth Division.

Answer

The record is 7 matches set by Notts County, Leyton Orient, and Fulham in 1920/1921 playing 7 teams with the suffix 'city' in a row. The Leyton Orient and Fulham chains stretch over the end of one season and into the next, so only Notts County really satisifies the question. However, the scheduling in these years involved a lot of back to back matches and so is cheating a bit.

More recently Chesterfield played 6 different teams with the suffix 'city' in a row in 1950/1951, and Leceister played 6 different 'united's in a row in their promotion season from League One in 2008/2009.

Even more bizarre, Carlisle United played 5 other different United's at the start of the 1988/1989 old Fourth Division season.

Youth Of The Nation

"If Lucas Hernández was born a year and a half later, his age would be a lower than his shirt number (21). Have any World Cup winners achieved this?" muses Edward Gibson.

The easiest way to check this is just to scrape all of the squads off of the wiki pages for the World Cups. I only did from 1954 onwards as before this the squad no and birthdate data is a bit patchy.

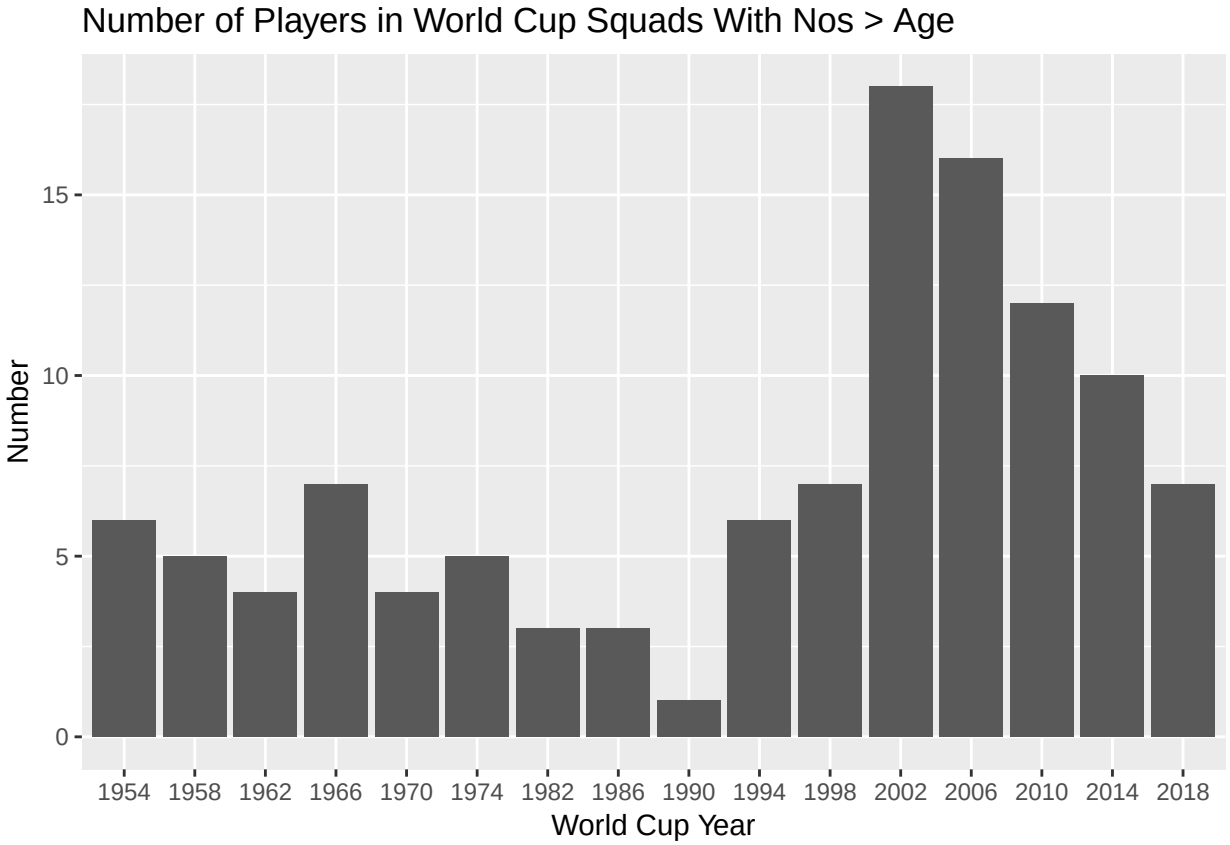
```
#links to the world cup squads pages
wiki_cup_squads <- sprintf("https://en.wikipedia.org/wiki/%s_FIFA_World_Cup_squads",
                           seq(1954, 2018, by = 4))

#scrape all the player data we need
world_cup_squads <- rbindlist(lapply(wiki_cup_squads[1:17], function(link) {
  year <- gsub(".*\\/wiki\\/", "", gsub("_FIFA_World.*", "", link))
  read <- read_html(link)

  sq_no <- read %>%
    html_nodes(".plainrowheaders td:nth-child(1)") %>%
    html_text() %>%
```

```
    as.numeric()
  sq_names <- read %>%
    html_nodes(".plainrowheaders a:nth-child(1)") %>%
    html_text() %>%
    .[. != ""] %>%
    .[!grepl("^\\[", .)] %>%
    .[. != "Unattached"] %>%
    .[!grepl("captain", .)]
  sq_dobs <- read %>%
    html_nodes(".plainrowheaders td:nth-child(4)") %>%
    html_text() %>%
    str_extract(., "[0-9]{4}-[0-9]{2}-[0-9]{2}") %>%
    as.Date()
  countries <- read %>% html_nodes("h3 .mw-headline") %>%
    html_text() %>%
    trimws()

  if(year > 2006) countries <- countries[1:32]

  squad_data <- data.frame(name = sq_names,
                           no = sq_no,
                           dob = sq_dobs,
                           year= year) %>%
    setDT() %>%
    .[!grepl("Nery Pumpido", name)] %>%
    .[no == 1, country := countries] %>%
    .[, country := na.locf(country)] %>%
    .[, c("name", "no", "dob", "year", "country")]
}))

#find all world cup squad players with shirt numbers greater than their age in years
young_players <- world_cup_squads %>%
  .[, age := as.numeric(difftime(as.Date(paste0(year, "-07-01")), dob)) / 365] %>%
  .[age < no]

print(young_players)
```

```
##                        name no        dob year    country      age
##   1:   Aleksandar Petakovic 22 1932-08-06 1954 Yugoslavia 21.91507
##   2:         Ranulfo Cortés 22 1934-07-09 1954     Mexico 19.99178
##   3:             Coskun Tas 22 1935-04-23 1954     Turkey 19.20274
##   4:           Omar Méndez 20 1934-08-07 1954    Uruguay 19.91233
##   5:          Johnny Haynes 21 1934-10-17 1954    England 19.71781
##  ---
## 110: Trent Alexander-Arnold 22 1998-10-07 2018    England 19.74521
## 111:    José Luis Rodríguez 21 1998-06-19 2018     Panama 20.04658
## 112:       Dávinson Sánchez 23 1996-06-12 2018   Colombia 22.06575
## 113:         Dawid Kownacki 23 1997-03-14 2018     Poland 21.31233
## 114:           Moussa Wagué 22 1998-10-04 2018    Senegal 19.75342
```

Overall 114 players are found. England actually have the most players with shirt numbers higher than their age with 9: Haynes, Hooper, Owen, Ferdinand, Carson, Walcott, Barkeley, Shaw, Alexander-Arnold. Surprisingly, most of these young English callups are pretty recent.

```
p <- ggplot(data = young_players, aes(year)) +
  geom_bar() +
  ggtitle("Number of Players in World Cup Squads With Nos > Age") +
  xlab("World Cup Year") +
  ylab("Number")

print(p)
```



It seems that the real high point for this was the turn of the century with young players being given a shot at the tail end of squads, which is returning to pre-1998 levels by 2018.

The data on these squad players is then merged with the data on the winning teams to find those who played for nations who went on to win the world cup.

```
wc_winners <- data.frame(winner = c("West Germany","Brazil","Brazil","England",
                                    "Brazil","West Germany","Argentina","Italy",
                                    "Argentina","West Germany","Brazil","France",
                                    "Brazil","Italy","Spain","Germany","France"),
                         year = seq(1954, 2018, 4))

#merge data with winners and find matches
young_players %<>% .[, year := as.numeric(as.character(year))] %>%
  .[, country := gsub("(^\\s+)|(\\s+$)", "", country)] %>%
  merge(., wc_winners, by = "year") %>%
  .[winner == country]

#kaka only one to have played as per https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_winners#By_ye
```

```
print(young_players)
```

```
##    year    name no        dob country      age winner
## 1: 1970    Leão 22 1949-07-11  Brazil 20.98630 Brazil
## 2: 1994 Ronaldo 20 1976-09-22  Brazil 17.78356 Brazil
## 3: 2002    Kaká 23 1982-04-22  Brazil 20.20548 Brazil
```

So only the great Émerson Leão, Ronaldo and Kaka satisfy the question. However, of these only Kaka played any part during the tournament, which only amounted to 25 minutes vs Costa Rica.

Which players *could* have satisfied this if they had a larger squad number?

```
youngest_players <- world_cup_squads %>%
  .[, age := as.numeric(difftime(as.Date(paste0(year, "-07-01")), dob)) / 365] %>%
  .[age < 23] %>%
  .[, country := gsub("(^\\s+)|(\\s+$)", "", country)] %>%
  .[, year := as.numeric(as.character(year))] %>%
  merge(., wc_winners, by = "year") %>%
  .[winner == country] %>%
  .[, dob := NULL]

#gives 53 potential results with world cup winners under the age of 23
print(youngest_players)
```

```
##     year                 name no      country      age       winner
##  1: 1954          Horst Eckel  6 West Germany 22.40822 West Germany
##  2: 1954     Ulrich Biesinger 18 West Germany 20.91507 West Germany
##  3: 1958                Pelé 10       Brazil 17.69863       Brazil
##  4: 1958              Moacir 13       Brazil 22.13425       Brazil
##  5: 1958             Orlando 15       Brazil 22.79452       Brazil
##  6: 1958             Mazzola 18       Brazil 19.95068       Brazil
##  7: 1962            Coutinho  9       Brazil 19.06849       Brazil
##  8: 1962                Pelé 10       Brazil 21.70137       Brazil
##  9: 1962             Jurandir 14       Brazil 21.64658       Brazil
## 10: 1962            Mengálvio 17       Brazil 22.55342       Brazil
## 11: 1962        Jair da Costa 18       Brazil 21.99178       Brazil
## 12: 1966            Alan Ball  7      England 21.15068      England
## 13: 1966        Martin Peters 16      England 22.66027      England
## 14: 1966        Norman Hunter 18      England 22.68767      England
## 15: 1970            Clodoaldo  5       Brazil 20.77534       Brazil
## 16: 1970        Marco Antônio  6       Brazil 19.41096       Brazil
## 17: 1970          Paulo Cézar 18       Brazil 21.05479       Brazil
## 18: 1970                Edu 19       Brazil 20.91507       Brazil
## 19: 1970             Zé Maria 21       Brazil 21.13425       Brazil
## 20: 1970                Leão 22       Brazil 20.98630       Brazil
## 21: 1974        Paul Breitner  3 West Germany 22.83562 West Germany
## 22: 1974           Uli Hoeneß 14 West Germany 22.50137 West Germany
## 23: 1974        Rainer Bonhof 16 West Germany 22.27123 West Germany
## 24: 1978    Alberto Tarantini 20    Argentina 22.59178    Argentina
## 25: 1978 José Daniel Valencia 21    Argentina 22.75890    Argentina
## 26: 1982        Franco Baresi  2        Italy 22.16164        Italy
## 27: 1982     Giuseppe Bergomi  3        Italy 18.53699        Italy
## 28: 1982      Daniele Massaro 17        Italy 21.12055        Italy
## 29: 1986       Claudio Borghi  4    Argentina 21.76986    Argentina
## 30: 1986           Luis Islas 15    Argentina 20.53699    Argentina
```

```
## 31: 1990       Andreas Möller 17 West Germany 22.84384 West Germany
## 32: 1994              Ronaldo 20       Brazil 17.78356       Brazil
## 33: 1998       Patrick Vieira  4       France 22.03562       France
## 34: 1998        Thierry Henry 12       France 20.88493       France
## 35: 1998      David Trezeguet 20       France 20.72329       France
## 36: 2002           Ronaldinho 11       Brazil 22.29315       Brazil
## 37: 2002                Kaká 23       Brazil 20.20548       Brazil
## 38: 2006     Daniele De Rossi  4        Italy 22.95342        Italy
## 39: 2010            Juan Mata 13        Spain 22.18904        Spain
## 40: 2010      Sergio Busquets 16        Spain 21.97260        Spain
## 41: 2010                Pedro 18        Spain 22.94247        Spain
## 42: 2010        Javi Martínez 20        Spain 21.84110        Spain
## 43: 2014      Matthias Ginter  3      Germany 20.46027      Germany
## 44: 2014       Julian Draxler 14      Germany 20.79178      Germany
## 45: 2014            Erik Durm 15      Germany 22.15068      Germany
## 46: 2014          Mario Götze 19      Germany 22.09041      Germany
## 47: 2014     Shkodran Mustafi 21      Germany 22.21918      Germany
## 48: 2018      Benjamin Pavard  2       France 22.27397       France
## 49: 2018     Presnel Kimpembe  3       France 22.89863       France
## 50: 2018         Thomas Lemar  8       France 22.64932       France
## 51: 2018        Kylian Mbappé 10       France 19.54247       France
## 52: 2018      Ousmane Dembélé 11       France 21.14247       France
## 53: 2018      Lucas Hernández 21       France 22.39178       France
##     year                 name no      country      age       winner
```

```
#most of these young players actually played at their world cups and many appeared in finals
youngest_players_appeared <- youngest_players[c(1, 3:6, 8, 12:13, 15:18, 21:23, 24:25, 27, 29, 31, 33:35

#find nearest matches
youngest_players_appeared %<>% .[, diff := age - no]
```

The closest other players to make it are David Trezeguet (1998, 21.7years no 20), Shkodran Mustafi (2014, 22.2years, no 21) and then Lucas Hernandez (22.4years, no 21). Hernandez is the closest one to actually play in the World Cup final. Alberto Tarantini is his closest competition at 22.6 years old and wearing shirt number 20 in the 1978 final.

Answer

Yes, three winners have appeared in World Cups with an age less than their shirt number. All Brazilians: Émerson Leão in 1970, Ronaldo in 1994, and Kaka in 2002. However only Kaka actually played (for 25 minutes vs. Costa Rica) in the finals.

Other close calls are David Trezeguet (21.7, no 20 in 1998) and Shkodran Mustafi (22.2, no 21 in 2014).

Hernandez *is* the closest to acheiving this having played in the final itself, with only Alberto Tarantini (22.5, no 20 in 1978) and Mario Goetze (22.1, no 19 in 2014) in close competition.