When this blog moved from bioinformatics to data science I ran a Twitter poll to ask whether I should start afresh at a new site or continue here. “Continue here”, you said.

So let’s test the tolerance of the long-time audience and celebrate the start of the 2019 season as we venture into the world of – [Australian football](https://en.wikipedia.org/wiki/Australian_rules_football) (AFL) statistics!

I’ve been hooked on the wonderful sport of AFL since attending my first game, the ANZAC Day match between the Sydney Swans and Melbourne in 2003, and have hardly missed a Swans home game since. However, I don’t think you need to be a sports fanatic – I certainly am not – to appreciate that sport is a rich source of data on which you can practice your R, statistics and data science skills. A large part of data science is figuring out what makes an interesting question, then querying the data to get the answer. Sport of course is full of trivia questions: the first, the last, the highest, the longest; and so provides many opportunities to devise questions and find answers. Sports fans also tend to hold strong opinions and make bold statements – not always backed up with evidence – which can be fun to engage with, armed with a little data.

As an example we’ll use [this list of predictions for the 2019 season](https://www.theroar.com.au/2019/03/21/my-23-crazy-fearless-predictions-for-the-2019-afl-season/) which tells us that:

Carlton will score 100 points  
A gentle one off the bat. The Blues sub-ton streak stands at 55 games, making it one of the longest in league history.

Let’s look at some ways to visualise how long it’s been since a team scored 100+ points. For years the go-to site for AFL data has been the wonderful [AFL Tables](https://afltables.com/afl/afl_index.html). The HTML and text files at this site are relatively easy to scrape into a dataframe using *rvest*. However, recent years have seen the development of another data source to which we bow down in awe and gratitude: the fitzRoy R package.

The results of every game since 1897 are stored in match\_results and look like this:

# A tibble: 15,407 x 16

Game Date Round Home.Team Home.Goals Home.Behinds Home.Points Away.Team Away.Goals

1 1 1897-05-08 R1 Fitzroy 6 13 49 Carlton 2

2 2 1897-05-08 R1 Collingw~ 5 11 41 St Kilda 2

3 3 1897-05-08 R1 Geelong 3 6 24 Essendon 7

4 4 1897-05-08 R1 Sydney 3 9 27 Melbourne 6

5 5 1897-05-15 R2 Sydney 6 4 40 Carlton 5

To explain some basics of the game: a score in AFL can be a goal (between the two big posts) for 6 points, or a behind (the ball hits a big post, goes between big and small post, or is taken through the posts by a defender) for 1 point. So the total points = (6 x goals) + behinds.

Any match statistic can be viewed from the perspective of either the home or away team. For our example we don’t care whether teams were home or away – we just want their total score. So we can simplify the results like this:

# packages for this post

library(fitzRoy)

library(tidyverse)

library(lubridate)

library(ggrepel)

match\_results %>%

select(Date, Team = Home.Team, Points = Home.Points) %>%

bind\_rows(select(match\_results, Date, Team = Away.Team, Points = Away.Points))

Result:

# A tibble: 30,814 x 3

Date Team Points

1 1897-05-08 Fitzroy 49

2 1897-05-08 Collingwood 41

3 1897-05-08 Geelong 24

4 1897-05-08 Sydney 27

5 1897-05-15 Sydney 40

So: how long since your team scored 100+ points? We can plot how long in days with a couple of simple filters:

match\_results %>%

select(Date, Team = Home.Team, Points = Home.Points) %>%

bind\_rows(select(match\_results, Date, Team = Away.Team, Points = Away.Points)) %>%

filter(year(Date) > 1999,

Points > 99) %>%

mutate(Days = as.numeric(Sys.Date() - Date)) %>%

group\_by(Team) %>%

filter(Date == max(Date)) %>%

ggplot(aes(Date, Points)) +

geom\_point() +

geom\_text\_repel(aes(label = paste(Team, Days)),

size = 3,

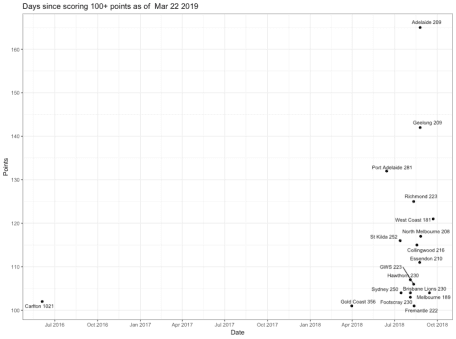
force = 2) +

scale\_x\_date(date\_breaks = "3 months",

date\_labels = "%b %Y") +

scale\_y\_continuous(breaks = seq(100, 180, 10)) +

labs(title = paste("Days since scoring 100+ points as of ", format(Sys.Date(), "%b %d %Y")))



Days since AFL teams scored 100+ points

It has indeed been a long time for Carlton. Every other team scored 100+ points in at least one game during the 2018 season.

How unusual is this time between 100+ scores, for Carlton or any other club? Let’s filter for the maximum days between 100+ scores:

match\_results %>%

select(Date, Team = Home.Team, Points = Home.Points) %>%

bind\_rows(select(match\_results, Date, Team = Away.Team, Points = Away.Points)) %>%

filter(Points > 99) %>%

group\_by(Team) %>%

arrange(Date) %>%

mutate(Days = as.numeric(Date - lag(Date))) %>%

filter(Days == max(Days, na.rm = TRUE)) %>%

arrange(desc(Days))

Result:

# A tibble: 19 x 4

# Groups: Team [19]

Date Team Points Days

1 1914-06-08 Fitzroy 108 2844

2 1910-07-02 Geelong 100 2506

3 1921-09-17 Melbourne 115 2324

4 1956-07-07 St Kilda 129 1862

5 1919-06-28 Essendon 113 1799

6 1968-06-29 Footscray 104 1512

7 1964-05-02 North Melbourne 121 1365

8 1918-07-20 Sydney 101 1134

9 1933-06-10 Hawthorn 105 1134

10 1901-08-24 Collingwood 143 1092

11 1924-08-23 Richmond 121 1064

12 1918-06-15 Carlton 111 1029

13 2012-08-11 Gold Coast 109 462

14 2011-05-07 Brisbane Lions 116 364

15 2018-03-31 Fremantle 106 328

16 2014-06-14 GWS 125 315

17 2001-05-12 Adelaide 103 302

18 2000-04-02 Port Adelaide 149 301

19 2011-04-02 West Coast 116 259

It’s certainly the longest period of the modern era. It’s also likely that Carlton will break their previous record drought of 1029 days which ended in June 1918. Fitzroy hold the record, with 2844 days between 100+ scores. It might seem unlikely that a team could go almost 8 years without scoring 100+ but we can filter the data to show that it is true:

match\_results %>%

select(Date, Team = Home.Team, Points = Home.Points) %>%

bind\_rows(select(match\_results, Date, Team = Away.Team, Points = Away.Points)) %>%

filter(Team == "Fitzroy",

between(Date, as.Date("1900-01-01"), as.Date("1914-12-31"))) %>%

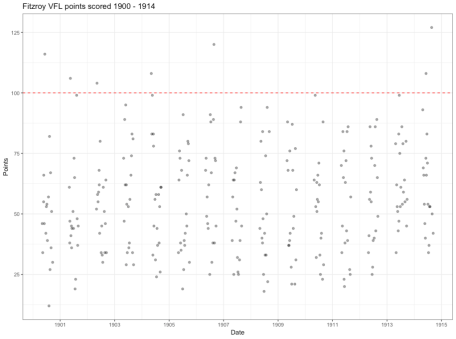
ggplot(aes(Date, Points)) +

geom\_point(alpha = 0.4) +

geom\_hline(yintercept = 100, color = "red") +

scale\_x\_date(date\_breaks = "2 years", date\_labels = "%Y") +

labs(title = "Fitzroy points scored in VFL matches 1900 - 1914")



Fitzroy scores 1900-1914

“Days since” is maybe not the best measure. The game is not played year-round and players who score goals don’t play every game.

A better measure, as in the linked article, could be “games since scoring 100+ points”. There are undoubtedly more elegant solutions to this question than mine, but here it is:

* Group the data by team and arrange by ascending date of game
* Create a new variable is100 with value 1 (Points >= 100) or 0 (Points < 100)
* Create a second variable is100cs, the cumulative sum of is100
* Filter for rows where is100cs = its maximum value (the most recent value)
* That number of rows is the games since (and including) the most recent 100+ score

match\_results %>%

select(Date, Team = Home.Team, Points = Home.Points) %>%

bind\_rows(select(match\_results, Date, Team = Away.Team, Points = Away.Points)) %>%

group\_by(Team) %>%

arrange(Date) %>%

mutate(is100 = ifelse(Points > 99, 1, 0),

is100cs = cumsum(is100)) %>%

filter(is100cs == max(is100cs)) %>%

summarise(n = n()) %>%

arrange(desc(n))

Result:

# A tibble: 20 x 2

Team n

1 University 126

2 Carlton 56

3 Gold Coast 21

4 Port Adelaide 11

5 Sydney 8

6 St Kilda 7

7 Collingwood 6

8 Hawthorn 6

9 GWS 5

10 Richmond 5

11 Brisbane Lions 4

12 Footscray 4

13 Fitzroy 3

14 Fremantle 3

15 Geelong 2

16 Melbourne 2

17 West Coast 2

18 Adelaide 1

19 Essendon 1

20 North Melbourne 1

Two teams in this list no longer play in the AFL: Fitzroy, which merged with Brisbane in 1996 and University. The latter played 7 seasons from 1908-1914 and in fact, never scored 100+ points. Once again this leaves Carlton at the top of the current 100+ drought club.

One last question: is Carlton’s current 55 games without scoring 100+ their longest ever? How about the longest ever of any team?

To find that we can group on team and our is100cs variable and again, count the rows and filter for the maximum count.

match\_results %>%

select(Date, Team = Home.Team, Points = Home.Points) %>%

bind\_rows(select(match\_results, Date, Team = Away.Team, Points = Away.Points)) %>%

group\_by(Team) %>%

arrange(Date) %>%

mutate(Games = row\_number(),

is100 = ifelse(Points > 99, 1, 0),

is100cs = cumsum(is100)) %>%

group\_by(Team, is100cs) %>%

summarise(n = n()) %>%

filter(n == max(n)) %>%

arrange(desc(n))

Result:

Team is100cs n

1 St Kilda 0 172

2 Fitzroy 5 140

3 Carlton 0 133

4 University 0 126

5 Sydney 0 122

6 Geelong 6 116

7 Melbourne 5 116

8 Essendon 6 85

9 Footscray 120 79

10 Hawthorn 3 60

11 North Melbourne 84 60

12 Collingwood 1 54

13 Richmond 15 48

14 Gold Coast 2 35

15 Brisbane Lions 233 21

16 Fremantle 147 17

17 GWS 0 17

18 Port Adelaide 16 17

19 Adelaide 39 13

20 West Coast 225 12

21 West Coast 303 12

This show us that Saint Kilda did not score more than 100 in their first game, nor in the 171 games that followed. Fitzroy’s aforementioned longest drought began after their fifth game, while Carlton did not score 100+ in their first 133 games.

At the other end of the scale, West Coast have gone at most only 11 games without scoring 100+, following the 225th and 303rd occasions on which they did so.

We can get some sense of how often each team scored 100+ in the following chart:

match\_results %>%

select(Date, Team = Home.Team, Points = Home.Points) %>%

bind\_rows(select(match\_results, Date, Team = Away.Team, Points = Away.Points)) %>%

group\_by(Team) %>%

arrange(Date) %>%

mutate(Games = row\_number(),

is100 = ifelse(Points > 99, 1, 0),

is100cs = cumsum(is100)) %>%

ggplot(aes(Games, is100cs)) +

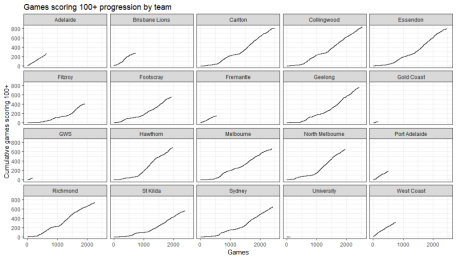
geom\_line() +

facet\_wrap(~Team) +

scale\_x\_continuous(breaks = seq(0, 3000, 1000)) +

labs(y = "Cumulative games scoring 100+",

title = "Games scoring 100+ progression by team")



Progression of 100+ scores by team

**Summary**  
In summary: tidy data in a nice package + tidyverse tools = much easier to slice, dice, query and aggregate in order to answer those burning AFL trivia questions.

Sports data science – give it a go! And if this has left you curious about AFL, I leave you with my own subjective assessment of its finest day in the last 10 years.