R

Greater Manchester Public Health Intelligence Network Session
22 March 2019

Dr Henry Partridge
Trafford Data Lab Manager

Hello

- Studied history, philosophy and crime science
- Data analyst in public sector
- Used R since 2013

Why use R?

ben goldacre ✓
@bengoldacre

Following

I think choice of tool is an important debate in science. I'm always concerned by anyone using point and click for research. Scripts best for reproducibility, transparency, external sense-checking, good audit, and code re-use. And (tenuously…) arguably makes u think more clearly

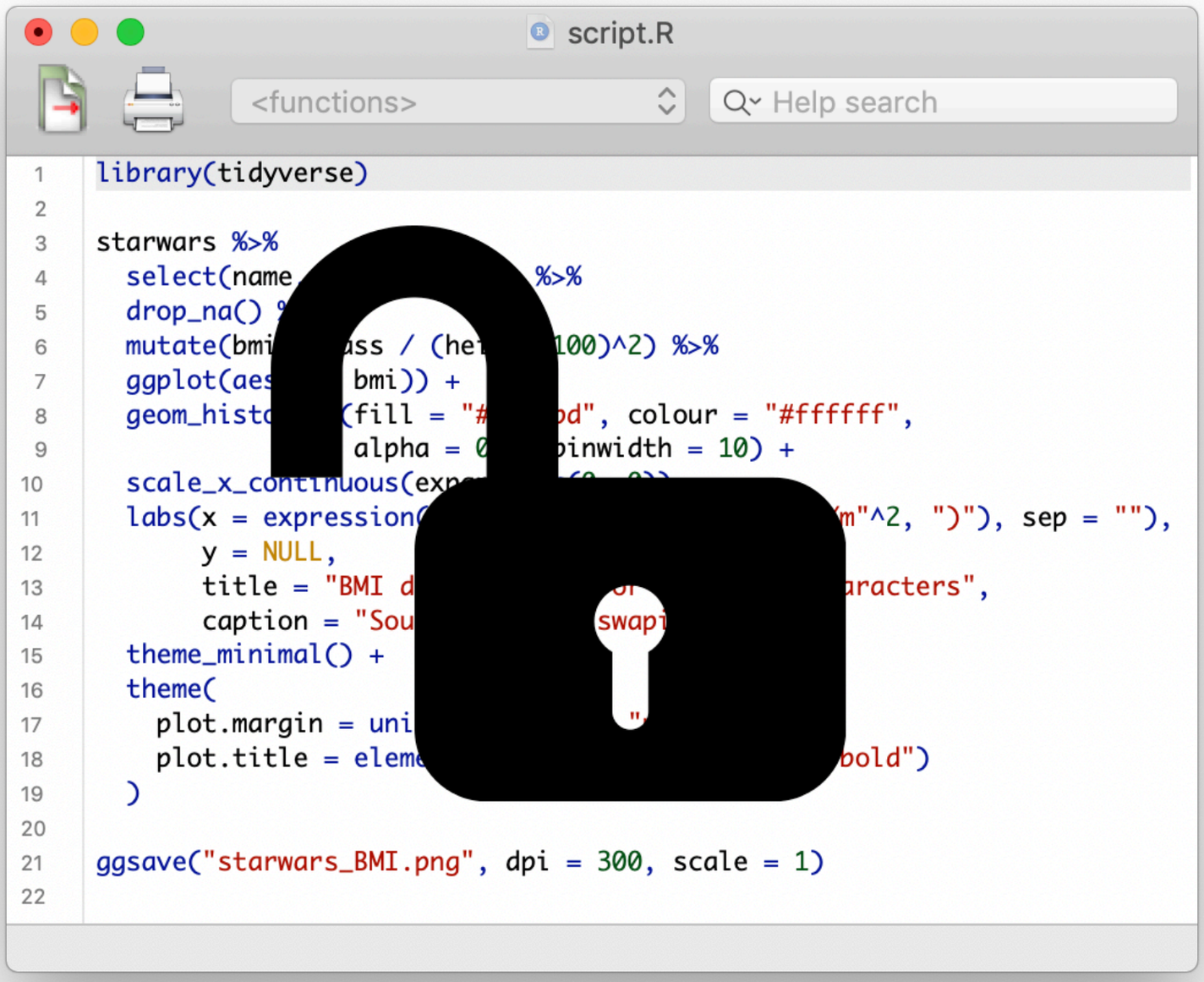
> **Andy Field** @ProfAndyField
> Replying to @LaurentWada @VictorKovalets and 10 others
>
> People get way too snobby about software, it's just a tool to do a job. If you're happy with SPSS use it, if you prefer R use it, love Excel? No problem. Use what works for you (and your collaborators).

1:23 PM - 21 Nov 2017

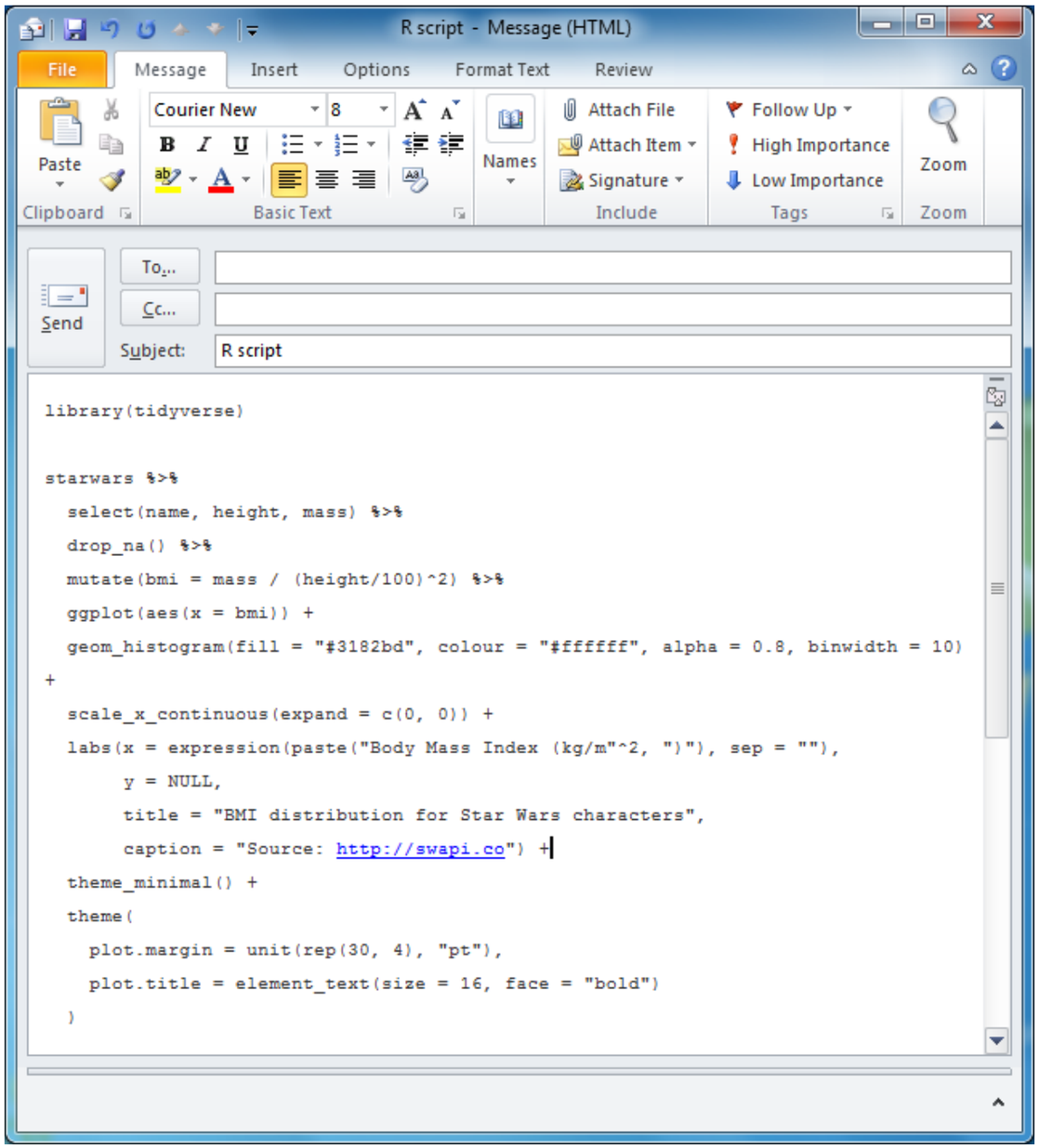
39 Retweets **108** Likes

19 39 108

Open


script.R
<functions>
Help search
1 library(tidyverse)
2
3 starwars %>%
4 select(name, %>%
5 drop_na() %
6 mutate(bmi ass / (hei 100)^2) %>%
7 ggplot(aes bmi)) +
8 geom_histo (fill = "# od", colour = "#ffffff",
9 alpha = 0 inwidth = 10) +
10 scale_x_continuous(exp
11 labs(x = expression(m"^2, ")"), sep = ""),
12 y = NULL,
13 title = "BMI d aracters",
14 caption = "Sou swapi
15 theme_minimal() +
16 theme(
17 plot.margin = uni
18 plot.title = eleme bold")
19)
20
21 ggsave("starwars_BMI.png", dpi = 300, scale = 1)
22

Shareable


R script - Message (HTML)
File
Message
Insert
Options
Format Text
Review
Courier New
8
Paste
Clipboard
Basic Text
Names
Attach File
Attach Item
Signature
Include
Follow Up
High Importance
Low Importance
Tags
Zoom
Zoom
Send
To...
Cc...
Subject:
R script
library(tidyverse)
starwars %>%
select(name, height, mass) %>%
drop_na() %>%
mutate(bmi = mass / (height/100)^2) %>%
ggplot(aes(x = bmi)) +
geom_histogram(fill = "#3182bd", colour = "#ffffff", alpha = 0.8, binwidth = 10)
+
scale_x_continuous(expand = c(0, 0)) +
labs(x = expression(paste("Body Mass Index (kg/m"^2, ")"), sep = ""),
y = NULL,
title = "BMI distribution for Star Wars characters",
caption = "Source: http://swapi.co") +
theme_minimal() +
theme(
plot.margin = unit(rep(30, 4), "pt"),
plot.title = element_text(size = 16, face = "bold")
)

Human readable

human_readable.txt

```
load the tidyverse R package

load the starwars dataframe
  select name, height and mass variables
  remove rows with missing values
  calculate the body mass index of each character
  call a ggplot object with bmi on the x-axis
  encode the data in a histogram
  remove padding around zero
  create an x-axis title
       leave the y-axis title blank
       add a plot title
       add a plot caption
add a minimalistic theme

    add padding to the plot
    increase the size of the title


save the plot as a PNG image in high resolution
```

script.R

```
library(tidyverse)

starwars %>%
  select(name, height, mass) %>%
  drop_na() %>%
  mutate(bmi = mass / (height/100)^2) %>%
  ggplot(aes(x = bmi)) +
  geom_histogram(fill = "#3182bd", colour = "#ffffff", alpha = 0.8, binwidth = 10) +
  scale_x_continuous(expand = c(0, 0)) +
  labs(x = expression(paste("Body Mass Index (kg/m"^2, ")"), sep = ""),
       y = NULL,
       title = "BMI distribution for Star Wars characters",
       caption = "Source: http://swapi.co") +
  theme_minimal() +
  theme(
    plot.margin = unit(rep(30, 4), "pt"),
    plot.title = element_text(size = 16, face = "bold")
  )

ggsave("starwars_BMI.png", dpi = 300, scale = 1)
```

Diffable

Showing **1 changed file** with **1 addition** and **1 deletion**.

Unified | Split

2 2019-03-22_GMPHIN/R/script.R — View file

```
@@ -2,7 +2,7 @@ library(tidyverse)
```

```
2
3   starwars %>%
4     select(name, height, mass) %>%
5   - filter(!is.na(height), !is.na(mass)) %>%
6     mutate(bmi = mass / (height/100)^2) %>%
7     ggplot(aes(x = bmi)) +
8     geom_histogram(fill = "#3182bd", colour = "#ffffff",
```

```
2
3   starwars %>%
4     select(name, height, mass) %>%
5   + drop_na() %>%
6     mutate(bmi = mass / (height/100)^2) %>%
7     ggplot(aes(x = bmi)) +
8     geom_histogram(fill = "#3182bd", colour = "#ffffff",
```

Community driven

R consortium

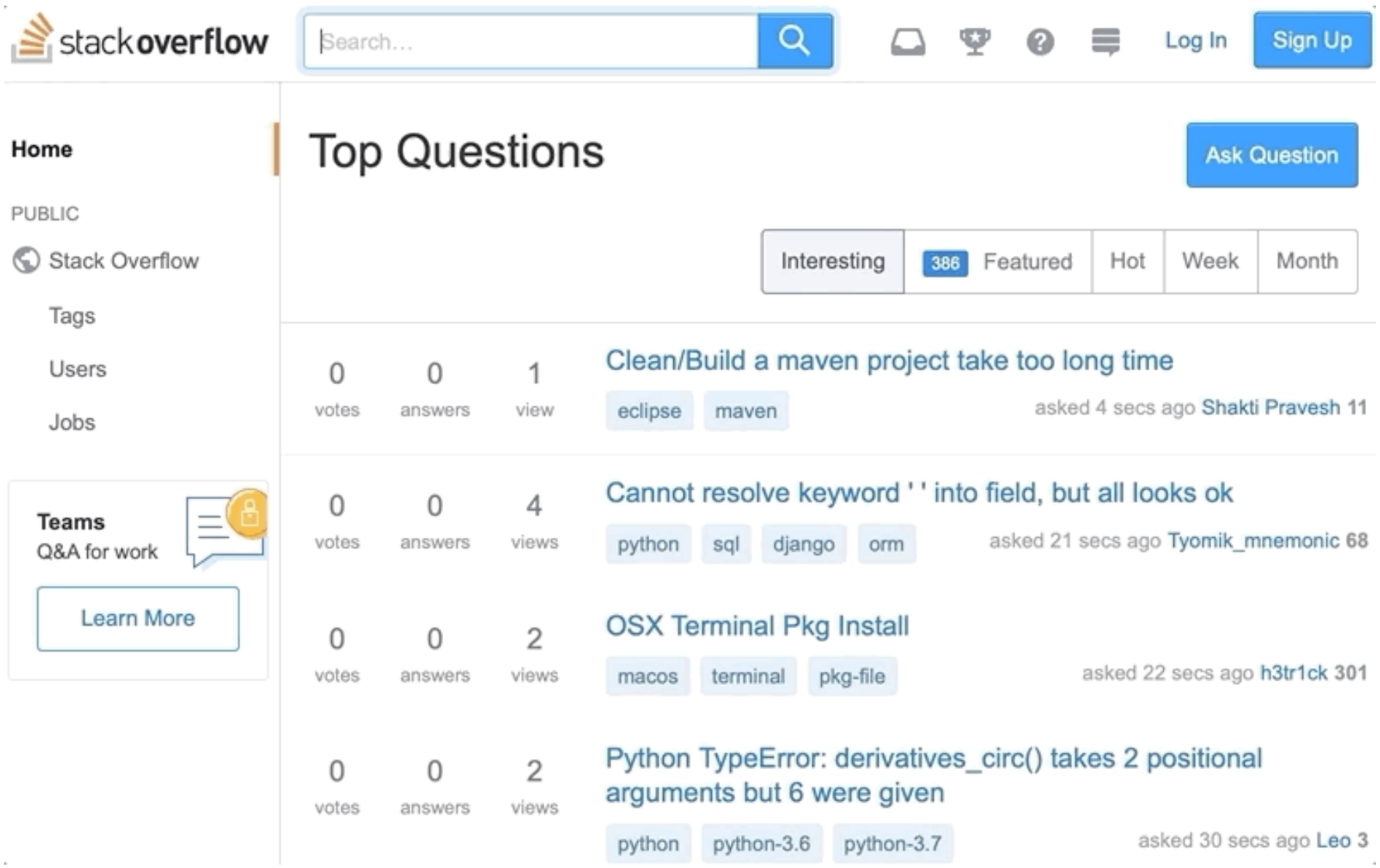
R Ladies

NHS-R COMMUNITY

Manchester R

SUPPORTING THE R COMMUNITY

Supportive

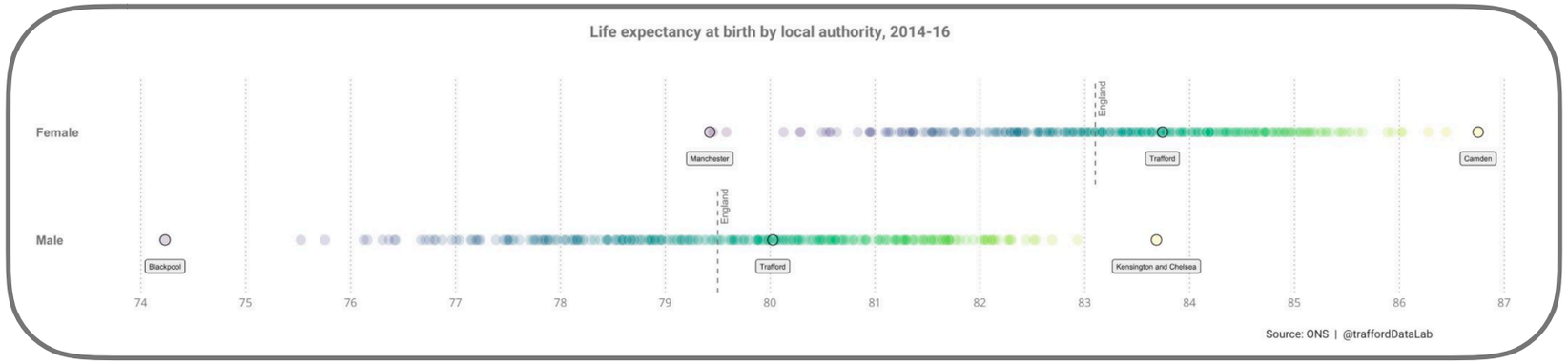

stackoverflow
Search...
Log In
Sign Up
Home
PUBLIC
Stack Overflow
Tags
Users
Jobs
Top Questions
Ask Question
Interesting
386 Featured
Hot
Week
Month
0 votes
0 answers
1 view
Clean/Build a maven project take too long time
eclipse
maven
asked 4 secs ago Shakti Pravesh 11
Teams
Q&A for work
Learn More
0 votes
0 answers
4 views
Cannot resolve keyword ' ' into field, but all looks ok
python
sql
django
orm
asked 21 secs ago Tyomik_mnemonic 68
0 votes
0 answers
2 views
OSX Terminal Pkg Install
macos
terminal
pkg-file
asked 22 secs ago h3tr1ck 301
0 votes
0 answers
2 views
Python TypeError: derivatives_circ() takes 2 positional arguments but 6 were given
python
python-3.6
python-3.7
asked 30 secs ago Leo 3

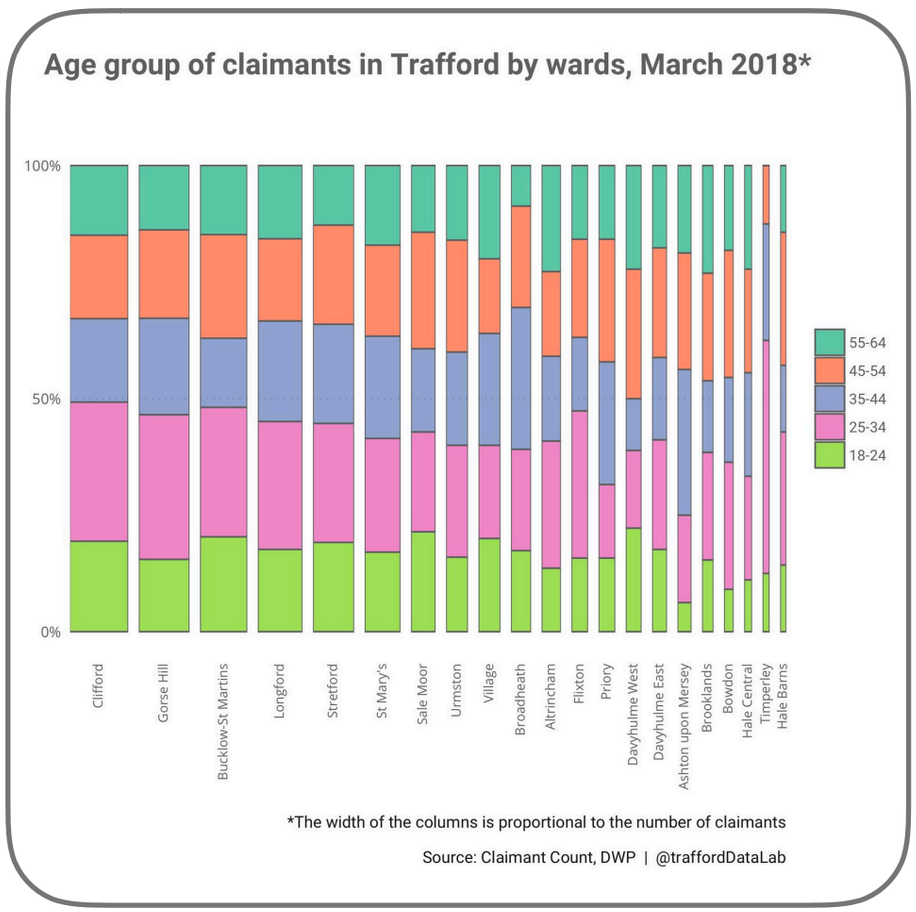
Cutting edge analytics

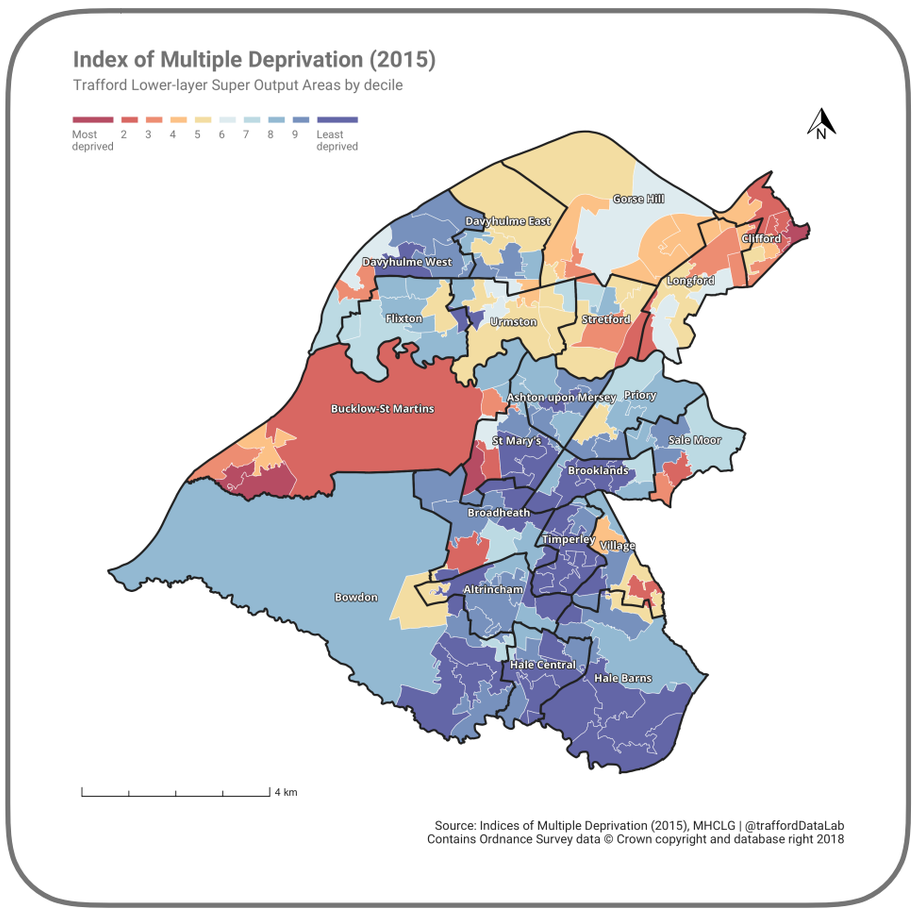
```r
# How many @traffordDataLab followers are bots? #

library(tidyverse) ; library(rtweet) ; library(tweetbotornot)

# retrieve followers of @OpenGovInt
followers <- get_followers("traffordDataLab", n = "all")
followers_info <- lookup_users(followers$user_id) %>%
  select(screen_name, name, followers = followers_count, following = friends_count)

# how many followers are bots?
bot <- tweetbotornot(followers_info$screen_name[1:50], fast = FALSE) %>%
  arrange(prob_bot)

# arrange by probability estimates
bot[order(bot$prob_bot), ]
```

```r
# plot probability estimates
bot %>%
  select(screen_name, prob_bot) %>%
  arrange(prob_bot) %>%
  ggplot()  +
  geom_col(aes(x = reorder(screen_name, -prob_bot), y = prob_bot), fill = "#E44690") +
  scale_y_continuous(limits = c(0, 1), expand = c(0,0)) +
  coord_flip() +
  labs(title = "Probability of @traffordDataLab followers being bots",
       x = NULL, y = NULL) +
  theme_minimal() +
  theme(plot.margin=unit(c(1,1,1,1),"cm"),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.text.y = element_text(hjust = 0))

ggsave("bot_or_not.png", dpi = 300, scale = 1)
```

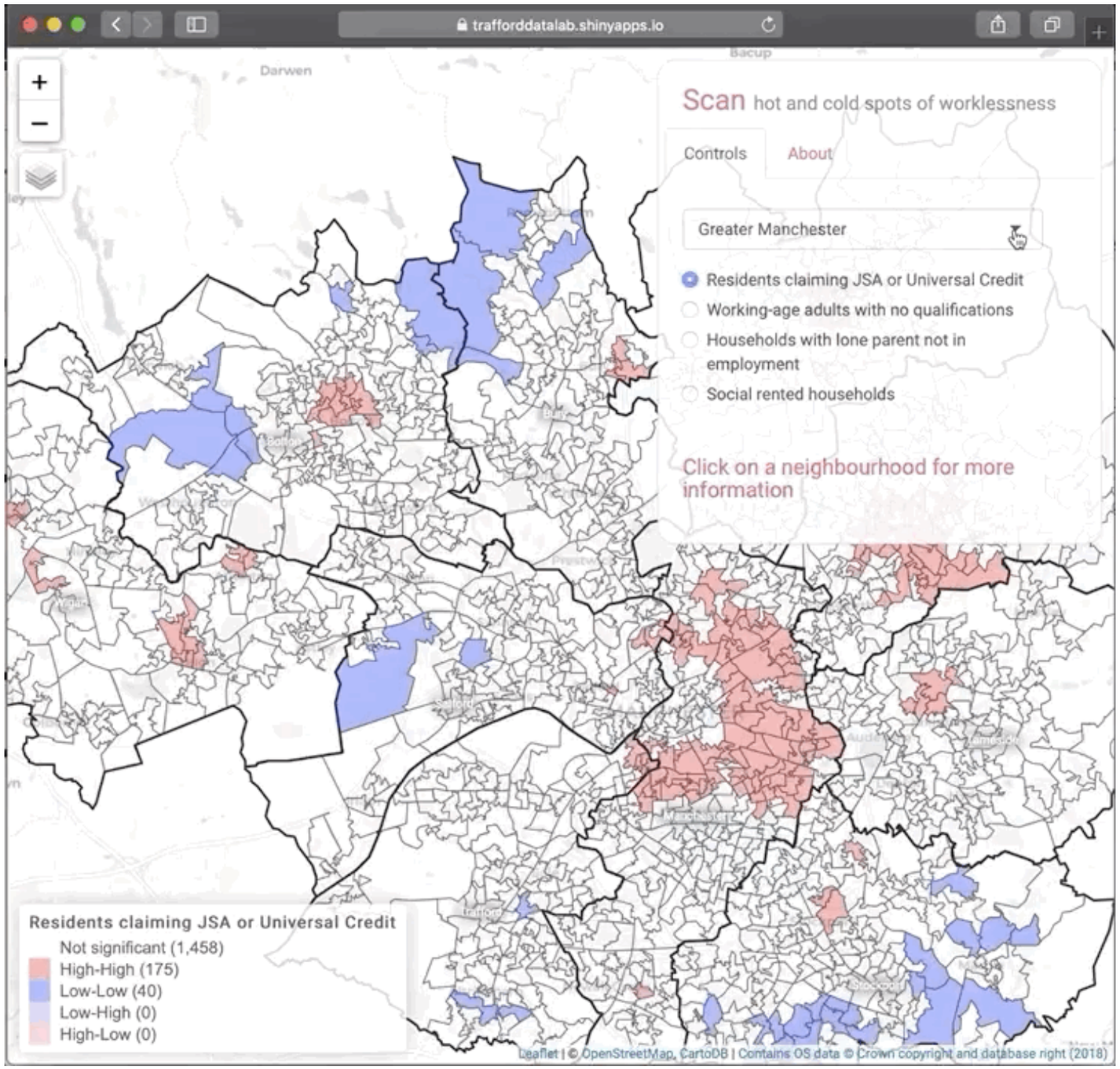
Graphics


Life expectancy at birth by local authority, 2014-16
Female
Male
Manchester
Trafford
Camden
Blackpool
Kensington and Chelsea
England
74
75
76
77
78
79
80
81
82
83
84
85
86
87
Source: ONS | @traffordDataLab


Age group of claimants in Trafford by wards, March 2018*
100%
50%
0%
55-64
45-54
35-44
25-34
18-24
Clifford
Gorse Hill
Bucklow-St Martins
Longford
Stretford
St Mary's
Sale Moor
Urmston
Village
Broadheath
Altrincham
Flixton
Priory
Davyhulme West
Davyhulme East
Ashton upon Mersey
Brooklands
Bowdon
Hale Central
Timperley
Hale Barns
*The width of the columns is proportional to the number of claimants
Source: Claimant Count, DWP | @traffordDataLab


Index of Multiple Deprivation (2015)
Trafford Lower-layer Super Output Areas by decile
Most deprived
2
3
4
5
6
7
8
9
Least deprived
N
Davyhulme East
Gorse Hill
Clifford
Davyhulme West
Longford
Flixton
Urmston
Stretford
Bucklow-St Martins
Ashton upon Mersey
Priory
St Mary's
Sale Moor
Brooklands
Broadheath
Timperley
Village
Bowdon
Altrincham
Hale Central
Hale Barns
4 km
Source: Indices of Multiple Deprivation (2015), MHCLG | @traffordDataLab
Contains Ordnance Survey data © Crown copyright and database right 2018

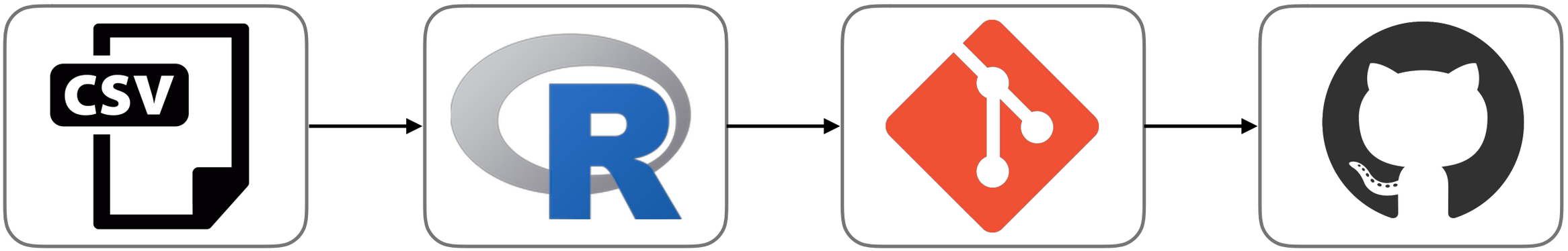
Interactive web apps



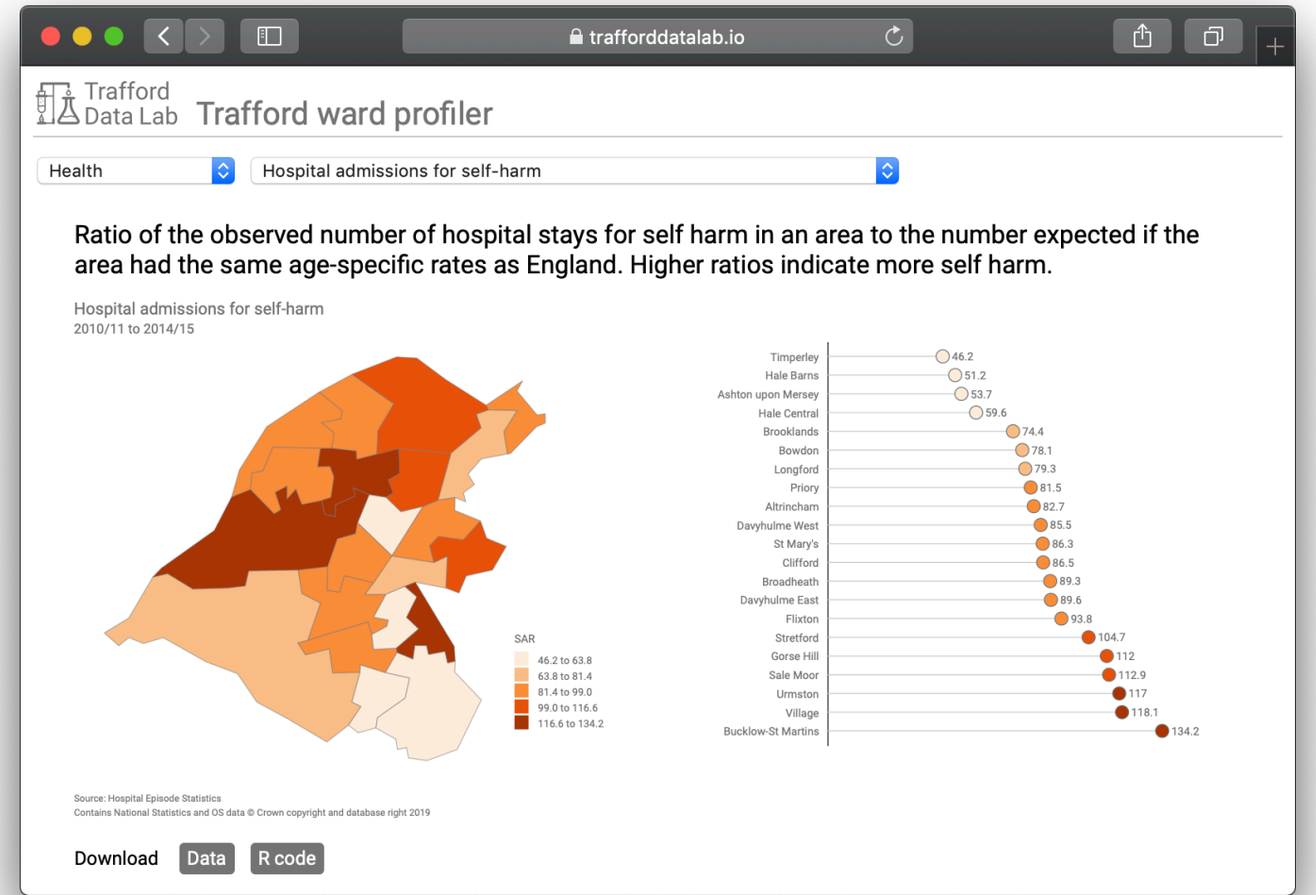
How the Lab uses R

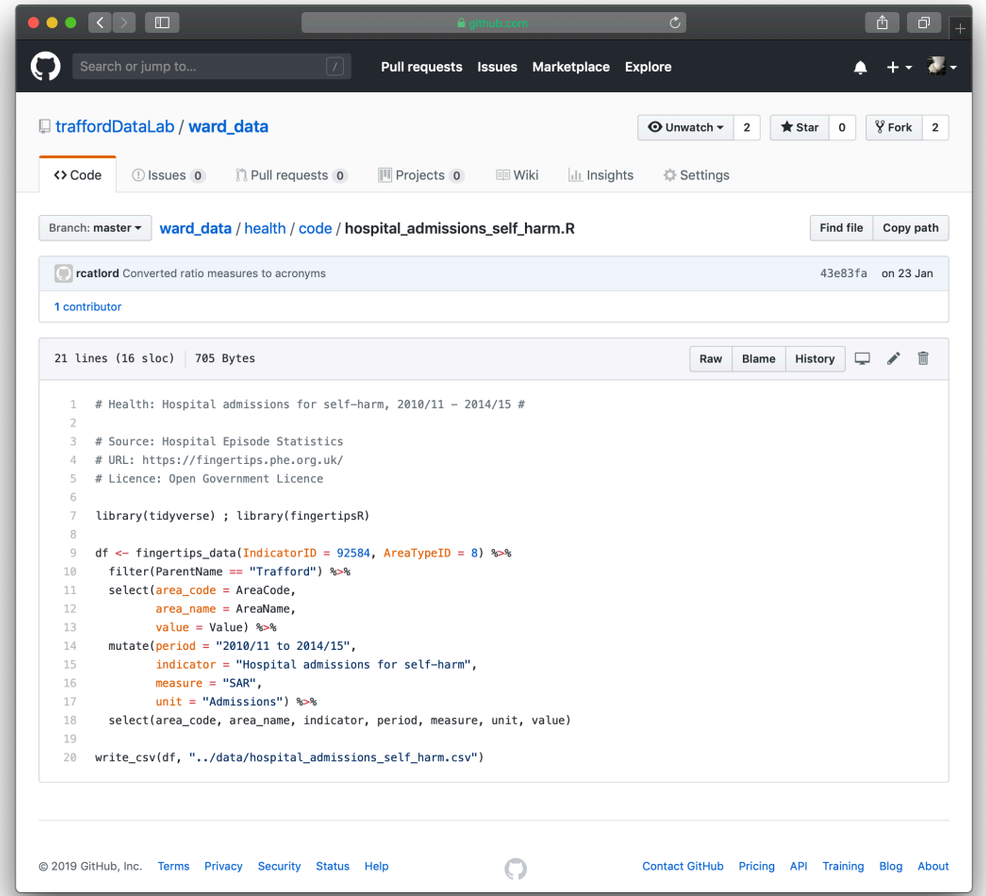
Reproducible workflow

Open data → **Programming** → **Version control** → **Sharing**



An example


trafforddatalab.io
Trafford Data Lab
Trafford ward profiler
Health
Hospital admissions for self-harm
Ratio of the observed number of hospital stays for self harm in an area to the number expected if the area had the same age-specific rates as England. Higher ratios indicate more self harm.
Hospital admissions for self-harm
2010/11 to 2014/15
SAR
46.2 to 63.8
63.8 to 81.4
81.4 to 99.0
99.0 to 116.6
116.6 to 134.2
Timperley 46.2
Hale Barns 51.2
Ashton upon Mersey 53.7
Hale Central 59.6
Brooklands 74.4
Bowdon 78.1
Longford 79.3
Priory 81.5
Altrincham 82.7
Davyhulme West 85.5
St Mary's 86.3
Clifford 86.5
Broadheath 89.3
Davyhulme East 89.6
Flixton 93.8
Stretford 104.7
Gorse Hill 112
Sale Moor 112.9
Urmston 117
Village 118.1
Bucklow-St Martins 134.2
Source: Hospital Episode Statistics
Contains National Statistics and OS data © Crown copyright and database right 2019
Download
Data
R code

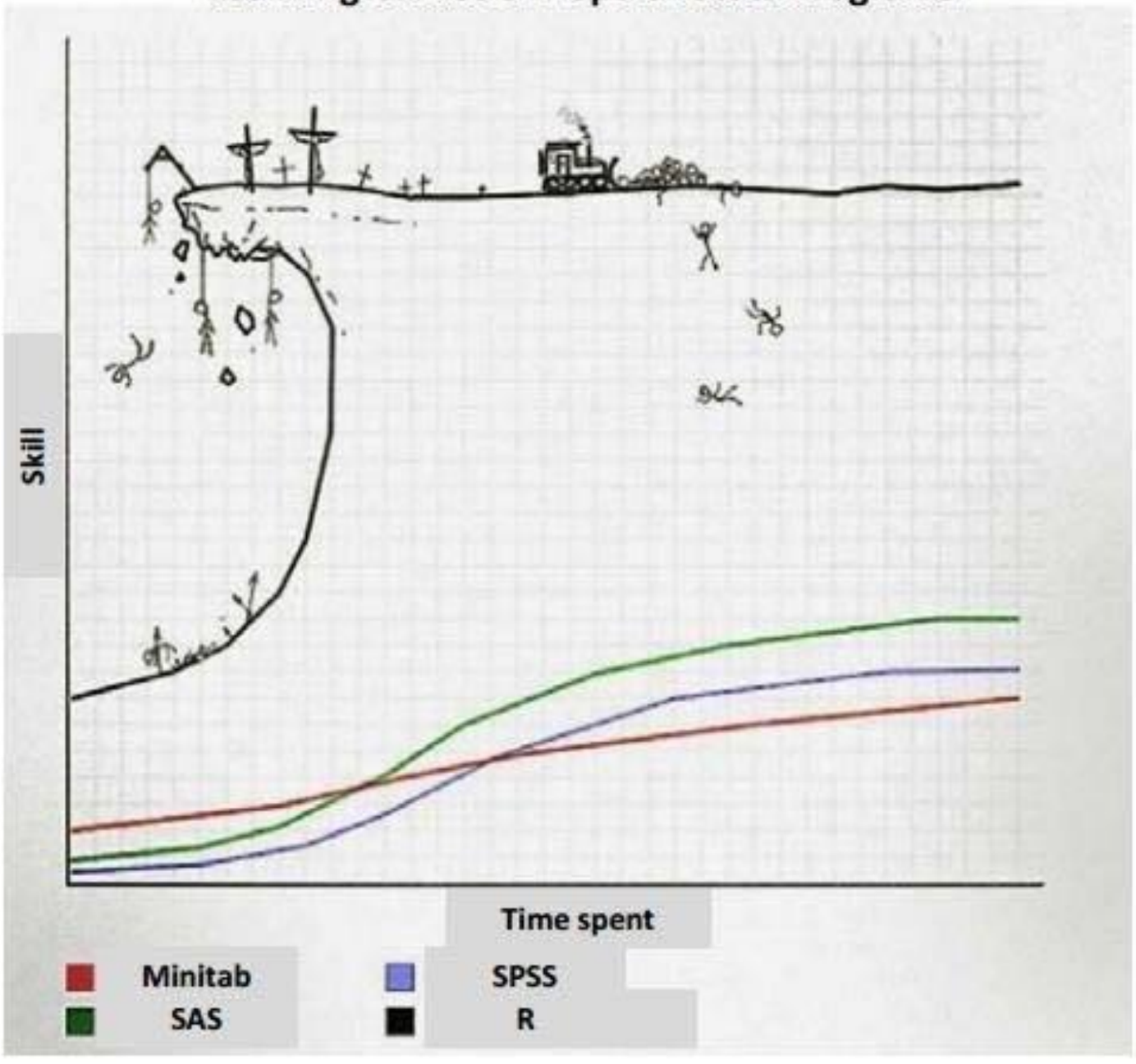

Search or jump to...
Pull requests
Issues
Marketplace
Explore
traffordDataLab / ward_data
Unwatch 2
Star 0
Fork 2
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Insights
Settings
Branch: master
ward_data / health / code / hospital_admissions_self_harm.R
Find file
Copy path
rcatlord Converted ratio measures to acronyms
43e83fa on 23 Jan
1 contributor
21 lines (16 sloc) 705 Bytes
Raw
Blame
History
# Health: Hospital admissions for self-harm, 2010/11 - 2014/15 #
# Source: Hospital Episode Statistics
# URL: https://fingertips.phe.org.uk/
# Licence: Open Government Licence
library(tidyverse) ; library(fingertipsR)
df <- fingertips_data(IndicatorID = 92584, AreaTypeID = 8) %>%
filter(ParentName == "Trafford") %>%
select(area_code = AreaCode,
area_name = AreaName,
value = Value) %>%
mutate(period = "2010/11 to 2014/15",
indicator = "Hospital admissions for self-harm",
measure = "SAR",
unit = "Admissions") %>%
select(area_code, area_name, indicator, period, measure, unit, value)
write_csv(df, "../data/hospital_admissions_self_harm.csv")
© 2019 GitHub, Inc.
Terms
Privacy
Security
Status
Help
Contact GitHub
Pricing
API
Training
Blog
About

Scalability

```r
# Health: Hospital admissions for self-harm, 2010/11 - 2014/15 #

# Source: Hospital Episode Statistics
# URL: https://fingertips.phe.org.uk/
# Licence: Open Government Licence

library(tidyverse) ; library(fingertipsR)

df <- fingertips_data(IndicatorID = 92584, AreaTypeID = 8) %>%
  filter(ParentName == "Trafford") %>%
  select(area_code = AreaCode,
         area_name = AreaName,
         value = Value) %>%
  mutate(period = "2010/11 to 2014/15",
         indicator = "Hospital admissions for self-harm",
         measure = "SAR",
         unit = "Admissions") %>%
  select(area_code, area_name, indicator, period, measure, unit, value)

write_csv(df, "../data/hospital_admissions_self_harm.csv")
```

Learning R

Learning Curves of Popular Stats Programs



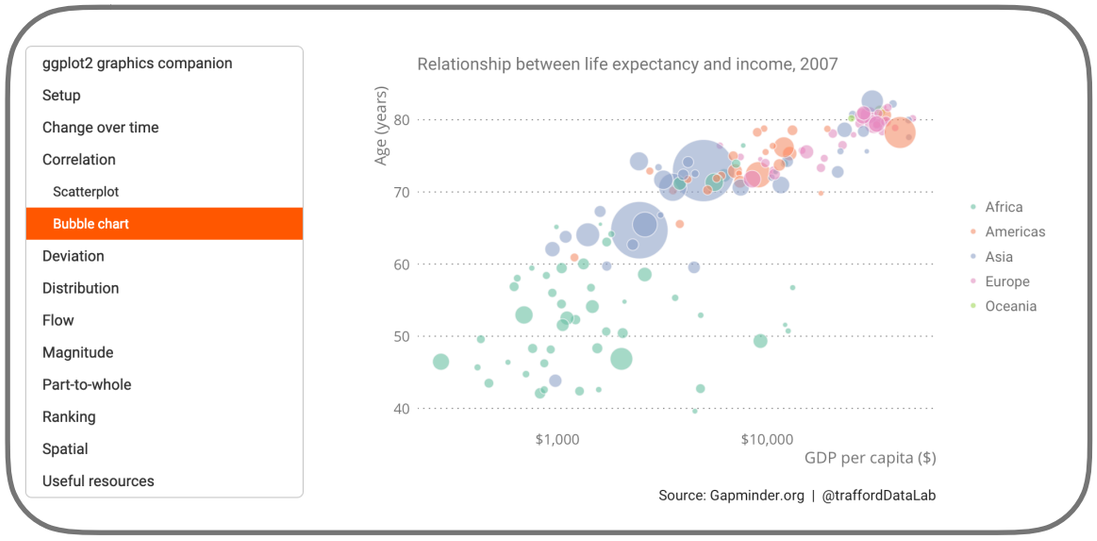
Getting started

- Reach for R rather than Excel
- Follow @hadleywickham, @JennyBryan, @dataandme, @NHSrCommunity and monitor #rstats tweets on Twitter
- Find answers or seek help on stackoverflow and RStudio Community
- Get a GitHub profile and commit your code
- Create a blog using R and post some tutorials
- Set up an R User Group with your colleagues

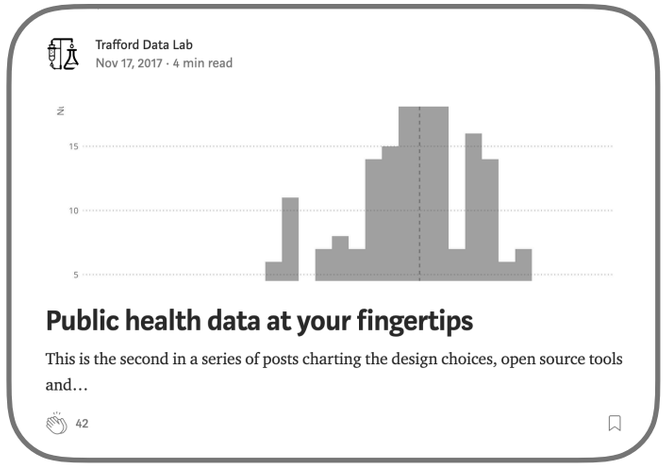
Lab resources

GitHub

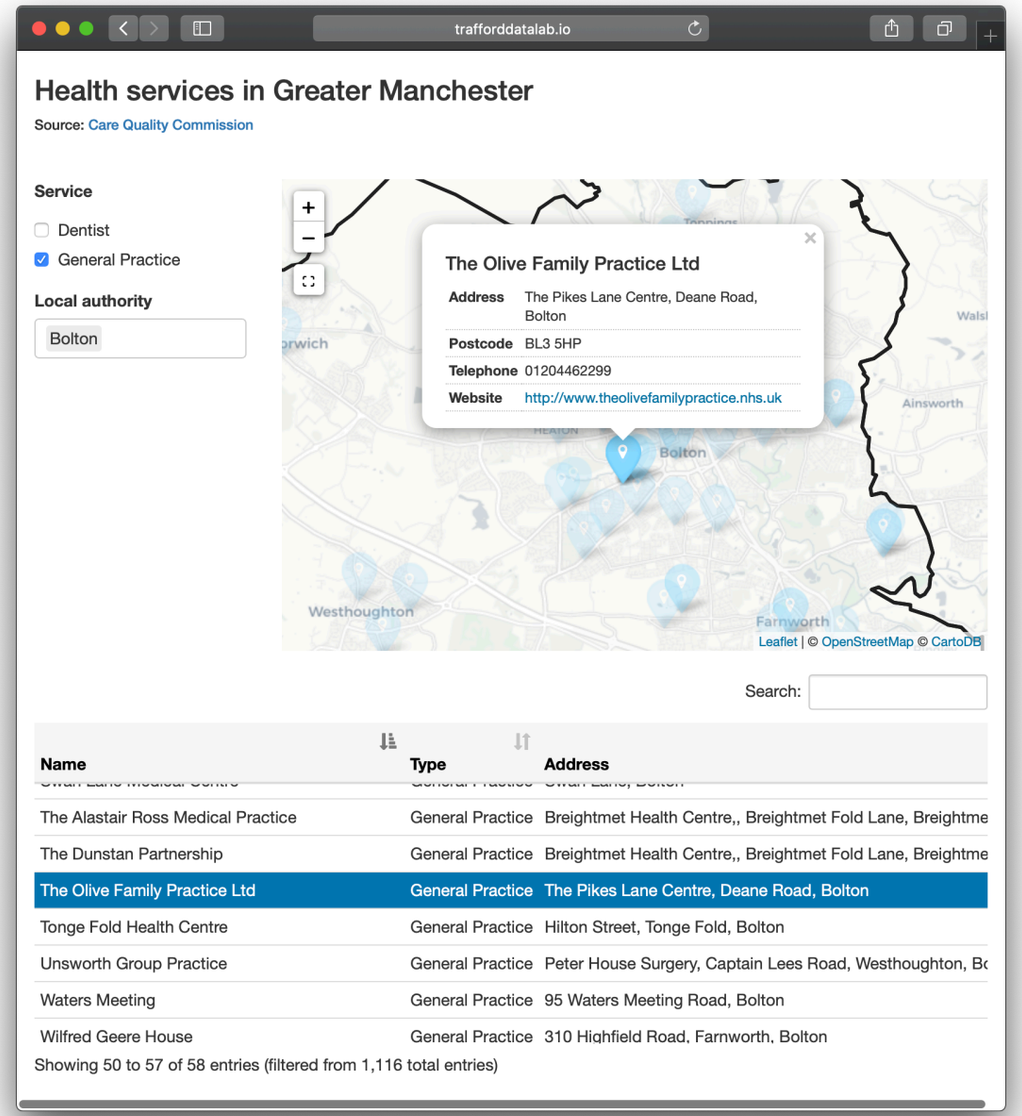
ggplot2 graphics companion



tutorials



Have a play



</> https://github.com/traffordDataLab/talks/tree/master/2019-03-22_GMPHIN/play

Questions?