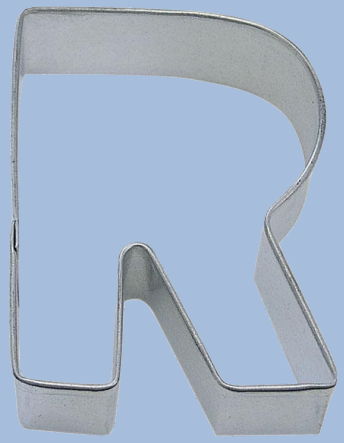
A Reproducible Workflow



using R and GitHub

Henry Partridge | BRITE event | 3 July 2019

Who am I?

I'm the manager of the Trafford Data Lab. I have an academic background in German philosophy and crime science. I've previously worked at TfL and MMU. I've been a cheerleader for #rstats since 2013.

What is reproducibility?

reproducibility /ˌriːprəˌdjuːsəˈbɪlɪti/ *noun* to obtain the same results using the method and data of the original study

which is different from ...

replication /rɛplɪˈkeɪʃ(ə)n/ *noun* to obtain the same results using the method of the original study and independently collected data

Why is reproducibility important?

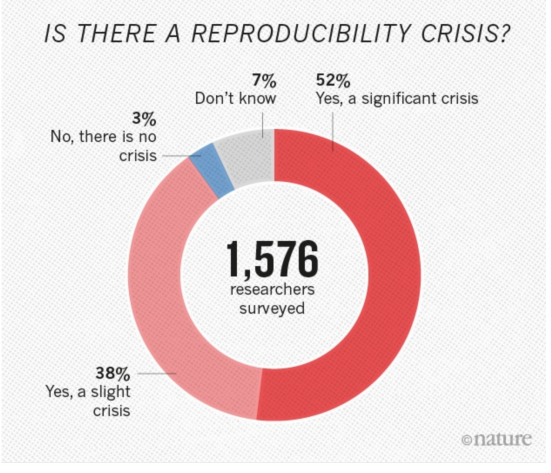
> non-reproducible single occurrences are of no significance to science

Karl Popper, The Logic of Scientific Discovery

Why is reproducibility important?

- allows checking and double checking by yourself and others
- enables rigorous peer review
- gives confidence in results

"Reproducibility crisis"



Source: nature.com

100 experimental and correlational studies in psychology were repeated with larger sample sizes. 97% of the original studies had statistically significant results but only 36% of the replications did. The replication effects were on average half the magnitude of the mean effect size of the original effects.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716, DOI:10.1126/science.aac4716

Open science initiatives

- Registered Reports
- study pre-registrations
- many-lab replication projects e.g ManyLabs
- sharing data
- open access publishing

> Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226-1227, DOI:10.1126/science.1213847

Practical steps for a reproducible workflow

Organised projects

- make your project folder self-contained
- quarantine your raw data

```
└── project
    ├── data
    │   ├── raw          # read-only pre-processed datasets
    │   └── processed    # intermediate datasets
    ├── R                # R scripts
    ├── outputs          # tables, charts
    ├── README.md        # project description
    ├── LICENCE.txt
    └── .gitignore
```

Readable code

- avoid absolute paths
- adopt a consistent style
- comment your code
- write functions

```
# this an absolute path
df <- read.csv("/Users/henrypartridge/Documents/project/data/foo.csv", string

# this is a relative path
df <- read.csv("data/foo.csv", stringsAsFactors = FALSE)
```

Literate programming

- avoid word processing software like MS Word
- combine code with human-readable plain text in R Markdown

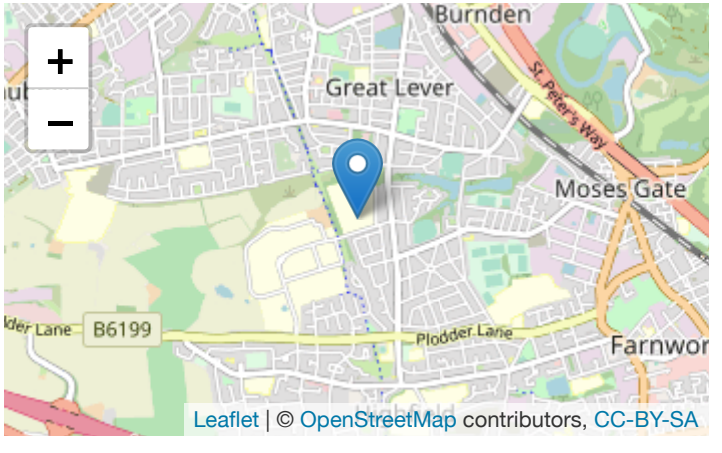
R Markdown

````
Bolton Science and Technology Centre
is located on Minerva Road.

```{r out.width = '100%',
fig.height = 3, echo = FALSE}
leaflet() %>%
  addTiles() %>%
  addMarkers(-2.424208, 53.554980,
             popup = "Bolton Science
             and Technology Centre")
```
````

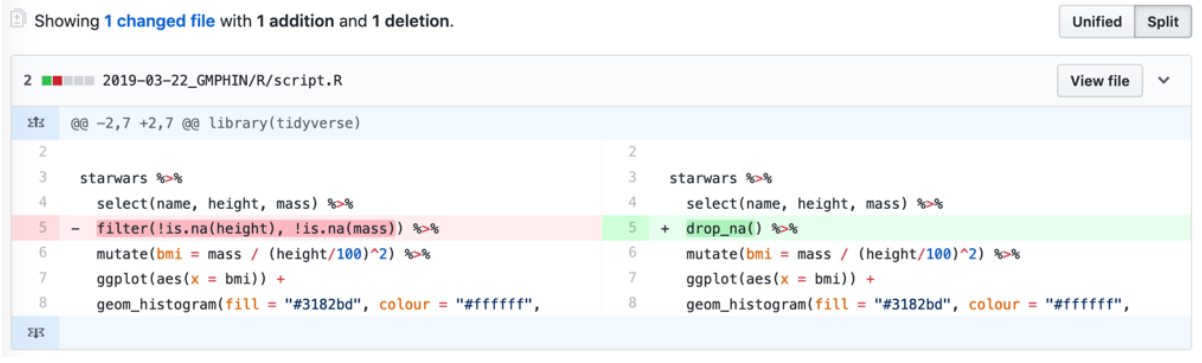
HTML output

Bolton Science and Technology Centre is located on Minerva Road.



Version control

- tracks changes to code and plain text files without need for version v0.1 etc.
- timestamps your work
- encourages collaboration
- integrates with RStudio
- remote copies of local projects can be stored on GitHub which also provides issue tracking, wikis and website hosting

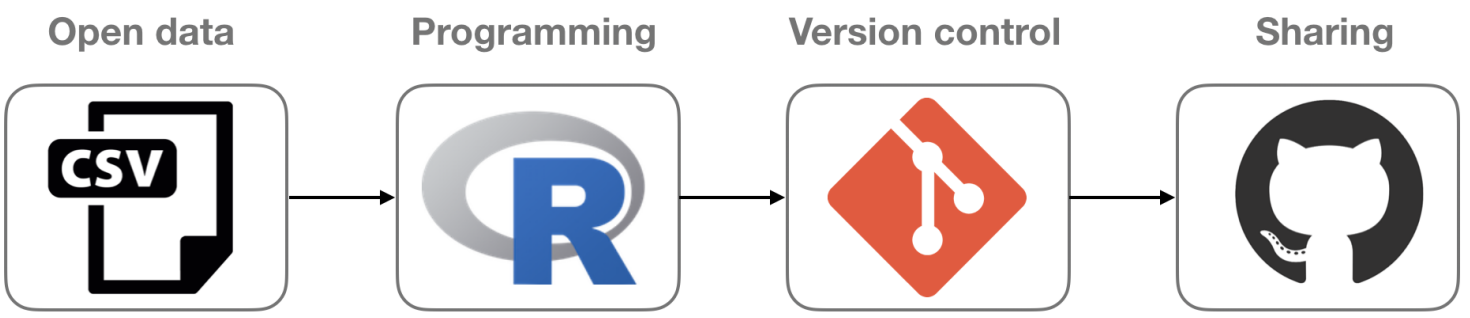


Licensing

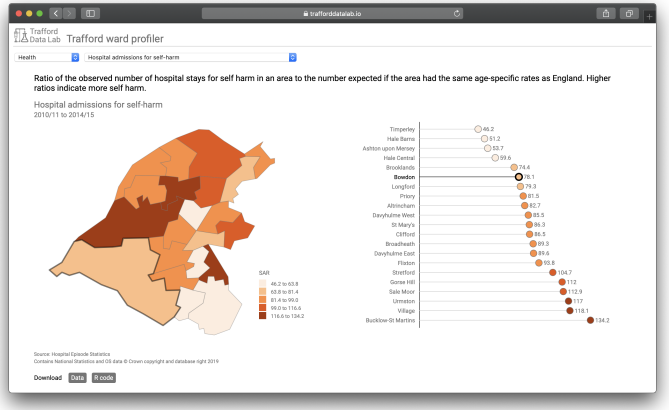
Give people permission to use your data and code:

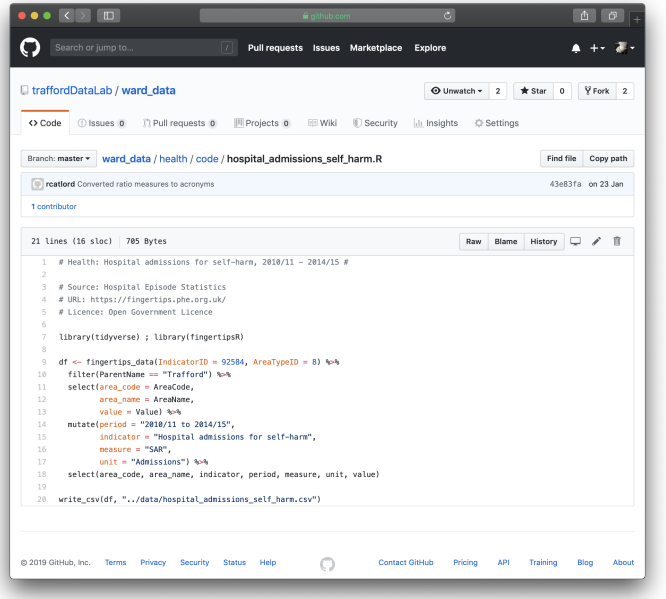
- Open Government Licence 3.0 for government published data
- CC-BY (Creative Commons Attribution) for media and text
- MIT licence for code

The Lab's workflow

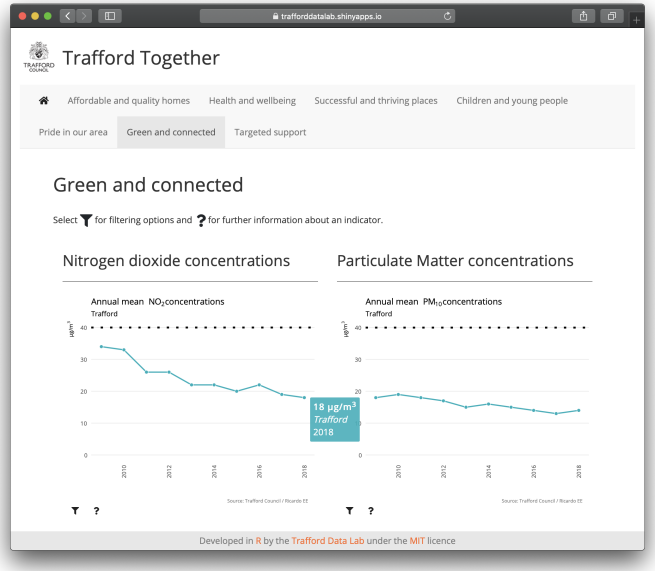

Open data
Programming
Version control
Sharing
CSV
R

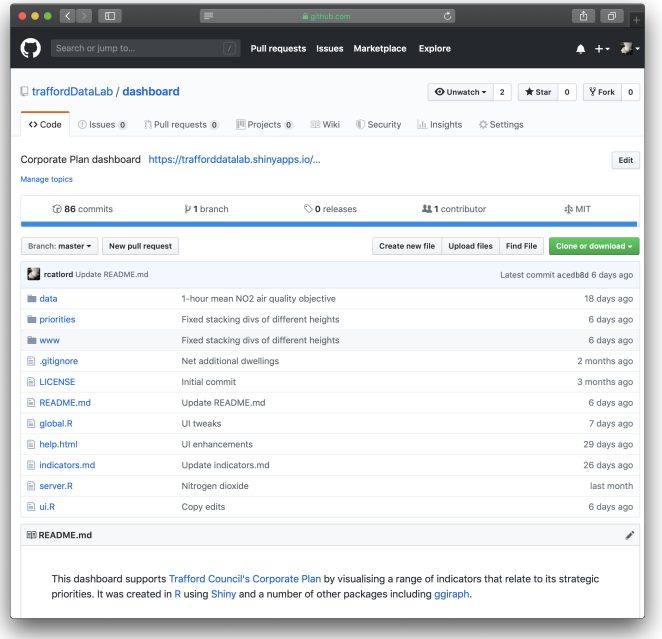
Example #1


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
Download
Data
R code

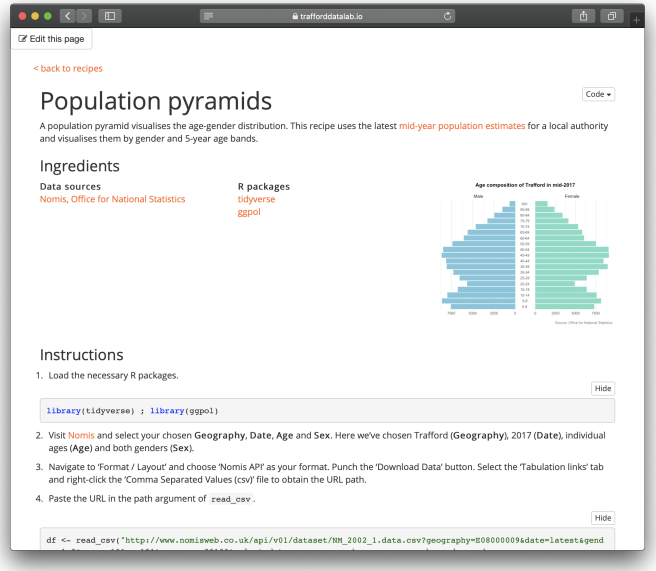

Search or jump to...
Pull requests
Issues
Marketplace
Explore
traffordDataLab / ward_data
Unwatch
Star
Fork
Code
Issues
Pull requests
Projects
Wiki
Security
Insights
Settings
Branch: master
ward_data / health / code / hospital_admissions_self_harm.R
Find file
Copy path
ratlord Converted ratio measures to acronyms
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

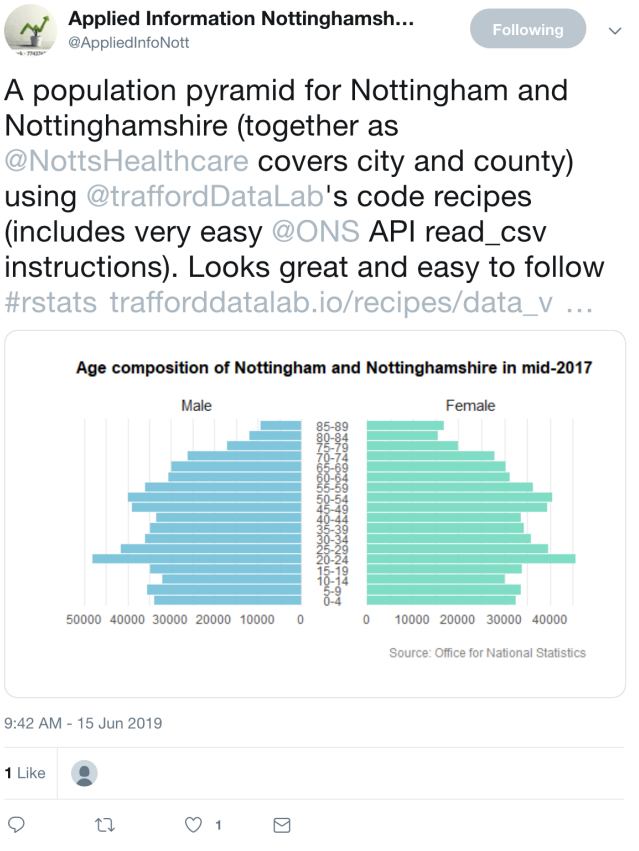
Example #2


Trafford Together
Affordable and quality homes
Health and wellbeing
Successful and thriving places
Children and young people
Pride in our area
Green and connected
Targeted support
Green and connected
Select for filtering options and for further information about an indicator.
Nitrogen dioxide concentrations
Particulate Matter concentrations
Annual mean NO2concentrations
Trafford
Annual mean PM10concentrations
Trafford
18 µg/m3
Trafford
2018
Developed in R by the Trafford Data Lab under the MIT licence


Pull requests
Issues
Marketplace
Explore
traffordDataLab / dashboard
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Security
Insights
Settings
Corporate Plan dashboard https://trafforddatalab.shinyapps.io/...
Edit
Manage topics
86 commits
1 branch
0 releases
1 contributor
MIT
Branch: master
New pull request
Create new file
Upload files
Find File
Clone or download
rcatlord Update README.md
Latest commit acedb8d 6 days ago
data 1-hour mean NO2 air quality objective 18 days ago
priorities Fixed stacking divs of different heights 6 days ago
www Fixed stacking divs of different heights 6 days ago
.gitignore Net additional dwellings 2 months ago
LICENSE Initial commit 3 months ago
README.md Update README.md 6 days ago
global.R UI tweaks 7 days ago
help.html UI enhancements 29 days ago
indicators.md Update indicators.md 26 days ago
server.R Nitrogen dioxide last month
ui.R Copy edits 6 days ago
README.md
This dashboard supports Trafford Council's Corporate Plan by visualising a range of indicators that relate to its strategic priorities. It was created in R using Shiny and a number of other packages including ggiraph.

Example #3


trafforddatalab.io
Edit this page
< back to recipes
Population pyramids
Code
A population pyramid visualises the age-gender distribution. This recipe uses the latest mid-year population estimates for a local authority and visualises them by gender and 5-year age bands.
Ingredients
Data sources
Nomis, Office for National Statistics
R packages
tidyverse
ggpol
Instructions
1. Load the necessary R packages.
Hide
library(tidyverse) ; library(ggpol)
2. Visit Nomis and select your chosen Geography, Date, Age and Sex. Here we've chosen Trafford (Geography), 2017 (Date), individual ages (Age) and both genders (Sex).
3. Navigate to 'Format / Layout' and choose 'Nomis API' as your format. Punch the 'Download Data' button. Select the 'Tabulation links' tab and right-click the 'Comma Separated Values (csv)' file to obtain the URL path.
4. Paste the URL in the path argument of read_csv .
Hide
df <- read_csv("http://www.nomisweb.co.uk/api/v01/dataset/NM_2002_1.data.csv?geography=E08000009&date=latest&gend


Applied Information Nottinghamsh...
@AppliedInfoNott
Following
A population pyramid for Nottingham and Nottinghamshire (together as @NottsHealthcare covers city and county) using @traffordDataLab's code recipes (includes very easy @ONS API read_csv instructions). Looks great and easy to follow #rstats trafforddatalab.io/recipes/data_v …
Age composition of Nottingham and Nottinghamshire in mid-2017
Male
Female
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
50000 40000 30000 20000 10000 0
0 10000 20000 30000 40000
Source: Office for National Statistics
9:42 AM - 15 Jun 2019
1 Like
1

take-home message

GCSE
Mathematics
Specification (8300/3F)

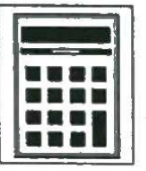
F

Paper 3 Foundation tier

Date | Morning | 1 hour 30 minutes

Materials

For this paper you must have:
- a calculator
- mathematical instruments.



Instructions
- Use black ink or black ball-point pen. Draw diagrams in pencil.
- Fill in the boxes at the bottom of this page.
- Answer **all** questions.
- You must answer the questions in the spaces provided. Do not write outside the box around each page or on blank pages.
- Do all rough work in this book.
- In all calculations, show clearly how you work out your answer.

Source: AQA

Useful resources

- Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10, DOI:10.1080/00031305.2017.1375989
- Bryan, J. (2018). Happy Git and GitHub for the useR
- Bryan, J. (2017) Project-oriented workflow
- rOpenSci, Reproducibility in Science

thank you

@trafforddatalab
@trafforddatalab
trafforddatalab.io

Slides created with **remark.js** and the R package **xaringan**