BRM5012 R practical

BRM5012 R practical

2018-08-27

What is R?

A programming language for data science and statistics.

Open source: free to use and build on.

Many contributors and packages.

What is R?

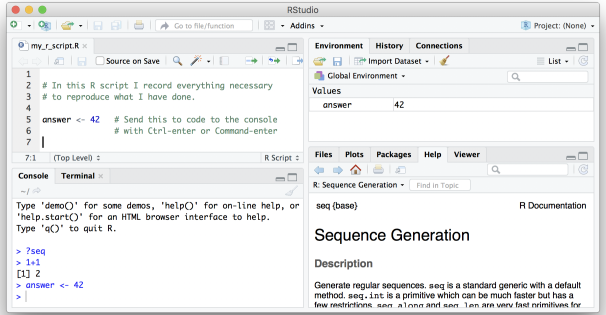
A programming language for data science and statistics.
Open source: free to use and build on.
Many contributors and packages.

R and RStudio

R is a programming language.

- Can be used directly from the "command line".

RStudio is an environment for using R, either locally or over the web.



We're using RStudio over the web today.

Can install on your own laptop. Need to install both R and RStudio.

Data science skills are generic

- Load data
- Tidy
- Explore, visualize
- Transform
- Summarize, model, statistical testing
- Report

See book “R for Data Science” by Garrett Grolemund and Hadley Wickham.

Data science skills are generic
- Load data
- Tidy
- Explore, visualize
- Transform
- Summarize, model, statistical testing
- Report

See book “R for Data Science” by Garrett Grolemund and Hadley Wickham.

Analysis of almost any data can be viewed as being made of these steps.

We can learn to recognize these steps. Something is wrong if one of them is missing.

These are interconnected steps, eg exploration informs transformation and testing.

R makes data science reproducible and automated

An "R script" is text that

- describes exactly what the computer should do.
- is a record of exactly what was done.

Can be turned into "functions" that become the building block of future analysis.

R makes data science reproducible and automated

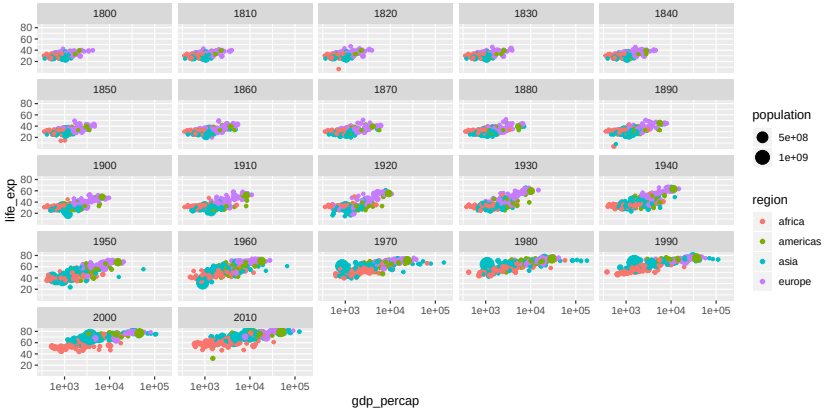
An "R script" is text that
- describes exactly what the computer should do.
- is a record of exactly what was done.

Can be turned into "functions" that become the building block of future analysis.

Excel or a statistics program may be easier to use, but can be hard to describe steps taken.

R makes data science reproducible and automated

```
ggplot(gap_geo,
    aes(x=gdp_percap, y=life_exp,
        size=population, color=region)) +
    geom_point() + facet_wrap(~ year) + scale_x_log10()
```



This code will make sense to you by the end of this practical.

The point here is that it is text, rather than a series of steps "click on this", "click on that", ...

Tidyverse

A modern collection of R packages that work well together, mostly written by Hadley Wickham.

Key packages:

- `dplyr` for manipulating tabular data
- `ggplot2` for visualization

Tidyverse

A modern collection of R packages that work well together, mostly written by Hadley Wickham.

Key packages:

- dplyr for manipulating tabular data
- ggplot2 for visualization

R and predecessor S have 4 decades of history.

Today we're going to be looking at two particular flavours of R: the "tidyverse" and "bioconductor".

The Tidyverse is a recent development that makes R a lot more accessable for casual users.

Bioconductor

Bioconductor is a repository of R packages for working with biological data.

- Special data types and file formats.
- Need to deal with large quantity of data.
- $p \gg n$, high throughput experiments often produce detailed information about a small number of biological samples, requiring special statistical methods.

Bioconductor

Bioconductor is a repository of R packages for working with biological data.

- Special data types and file formats.
- Need to deal with large quantity of data.
- $p \gg n$, high throughput experiments often produce detailed information about a small number of biological samples, requiring special statistical methods.

Aims for these practicals

- Hands on experience using R.
- Enough about tidyverse for it to be immediately useful.
- Know enough to know the next question to ask, and where to look for the answer.

Not aiming to cover

- Statistical modelling and testing (`lm`, etc).
- Bioconductor in depth.

Developing fluency in R will take further reading and practice.

Aims for these practicals
- Hands on experience using R.
- Enough about tidyverse for it to be immediately useful.
- Know enough to know the next question to ask, and where to look for the answer.

Not aiming to cover
- Statistical modelling and testing (lm, etc).
- Bioconductor in depth.

Developing fluency in R will take further reading and practice.

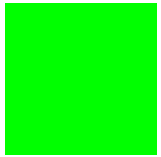
We consider the tidyverse flavour of R an essential skill, something any lab scientist should know.

Bioconductor requires more work to master, is a more specialized skill, and has sub-specialities.

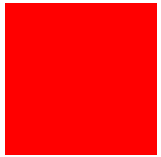
Workshop format

- Follow presenter in your own RStudio session, with variations. Try things out!
- Short challenges to apply what you are learning.
- Etherpad to share challenge solutions. Alternative place to ask questions.

Sticky notes



All good
Challenge completed



Something is broken
It doesn't work
Something doesn't make sense

(or raise hand/call out)

Workshop format
- Follow presenter in your own RStudio session, with variations. Try things out!
- Short challenges to apply what you are learning.
- Etherpad to share challenge solutions. Alternative place to ask questions.

Sticky notes

All good
Challenge completed

Something is broken
It doesn't work
Something doesn't make sense

(or raise hand/call out)

We're here to teach, ask for help early.

R cares a lot about missing punctuation. Easy to miss a quote or a bracket and get R into a confused state.

Digressions are good.