



Using R as a Research Tool.

NERC E3 DTP Training – 7th November

Dr Susan Johnston, Institute of Evolutionary Biology

Demonstrators: Gergana Daskalova, John Godlee

What is ?

- Environment for statistical computing and graphics.
- Interpreted programming language.
- Powerful research tool.
- **Free and open source** multi-platform software.

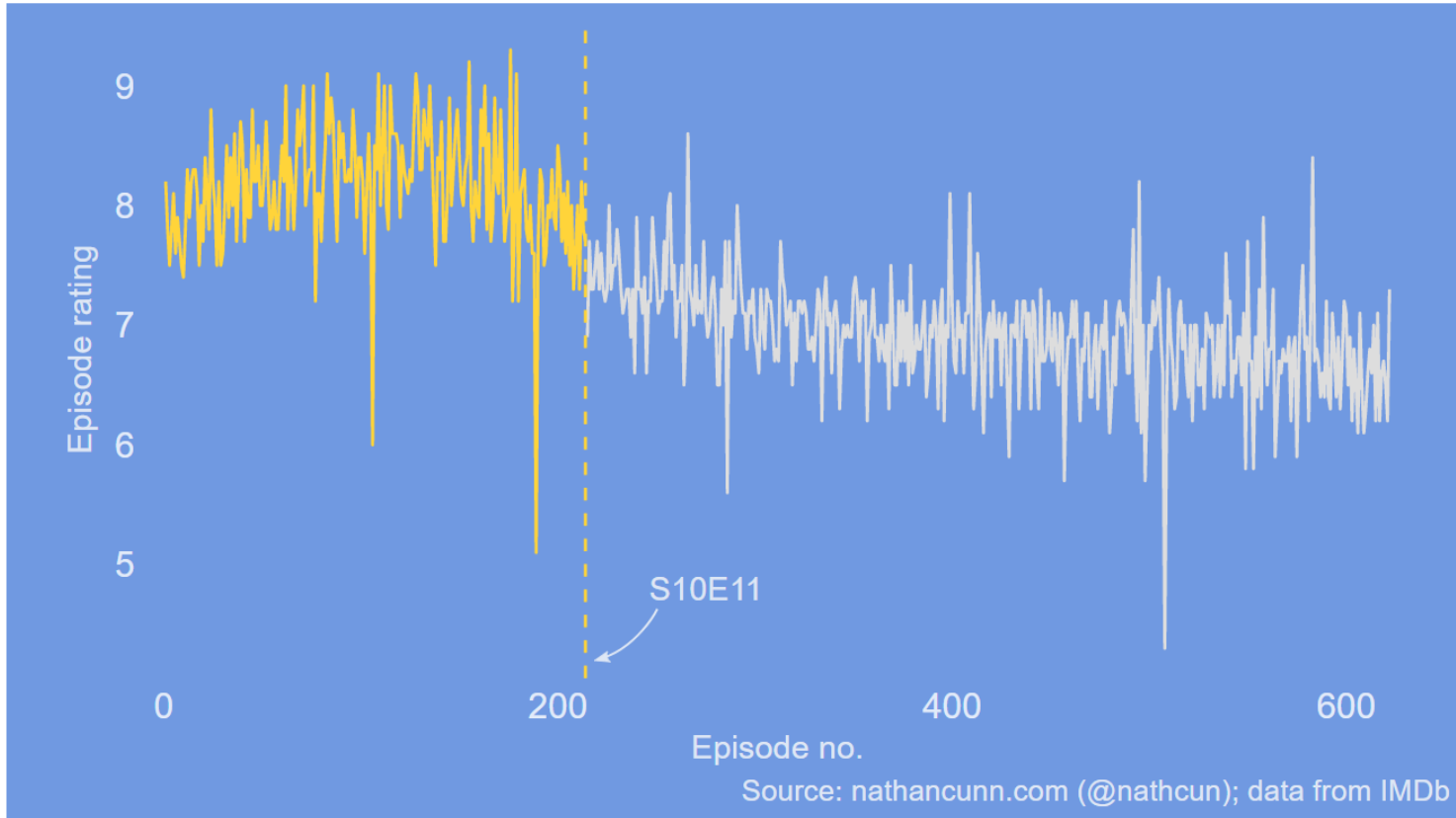
Why use ?

“This is R. There is no if. Only how.”
-- Simon ‘Yoda’ Blomberg, R-help (April 2005)

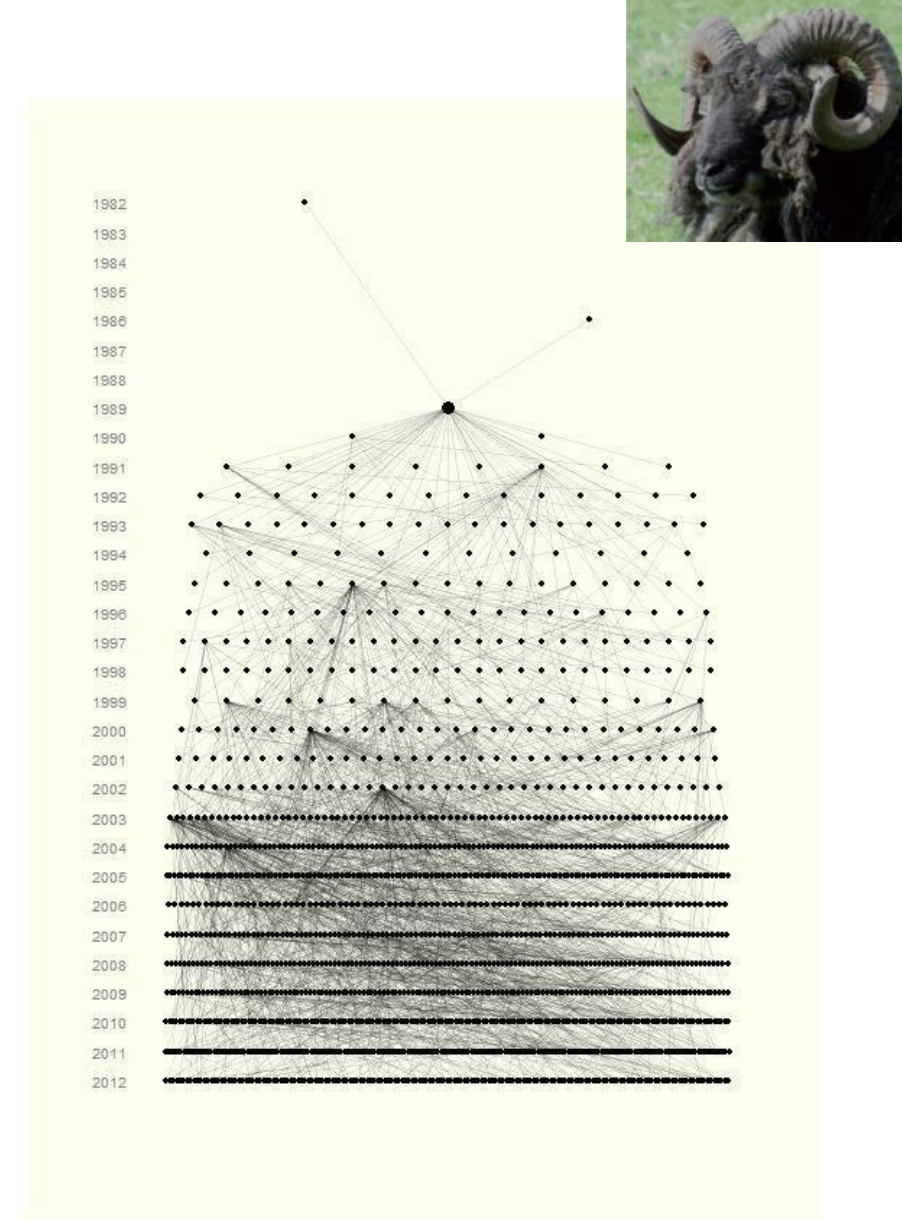
- Statistics.
- Data visualisation.
- Interactive web applications.
- Processing and tidying data.
- Reports and presentations.
- Portable projects.

Data visualisation

e.g. <http://www.r-graph-gallery.com/portfolio/ggplot2-package/>



When did the golden age of The Simpsons end?



**Ancestors and descendants
of a single Soay sheep
called Snowball.**

Report writing

Using R as a Research Tool.

Dr Susan Johnston: Susan.Johnston@ed.ac.uk

Demonstrators: Gergana Dalaskova, John Godlee.
Hat-Tips to Kyle Dexter, The Coding Club and R4all.

November 6, 2017

1 Introduction

1.1 What is R?

R began its life in New Zealand in 1993 as a language and environment for statistical computing and graphics. It is an interpreted programming language, meaning that rather than pointing and clicking, the user types in commands. It is **free** and works across all platforms.

1.2 Why use R?

R Base Graphics: An Idiot's Guide

Comments (-)

Share

Hide Toolbars

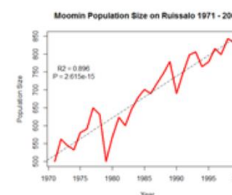
One of the most powerful functions of R is its ability to produce a wide range of graphics to quickly and easily visualise data. Plots can be replicated, modified and even publishable with just a handful of commands.

Making the leap from chiefly graphical programmes, such as Excel and Sigmaplot, may seem tricky. However, with a basic knowledge of R, just investing a few hours could completely revolutionise your data visualisation and workflow. Trust me - it's worth it.

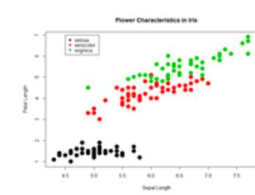
Last year, I presented an informal course on the basics of R Graphics University of Turku. In this blog post, I am providing some of the slides and the full code from that practical, which shows how to build different plot types using the basic (i.e. pre-installed) graphics in R, including:



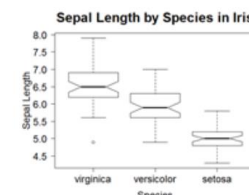
1. Basic Histogram



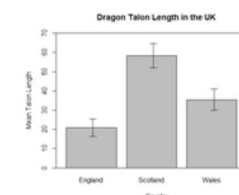
2. Line Graph with Regression



3. Scatterplot with Legend



4. Boxplot with reordered/
formatted axes



5. Boxplot with Error Bars

LaTeX and R Sweave

knitr to HTML

Interactive applications (**shiny**)

<https://scotland.shinyapps.io/babynames/>



Baby names trends in Scotland since 1974

Enter a **name**, select the **gender** and click on '**Apply**' to see how a name's popularity has changed over the years.
App might be slow at busy times. Please be patient.

Name

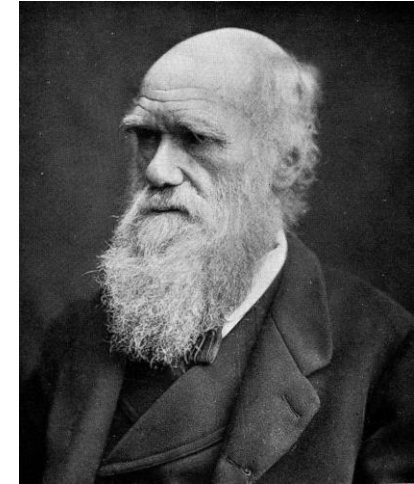
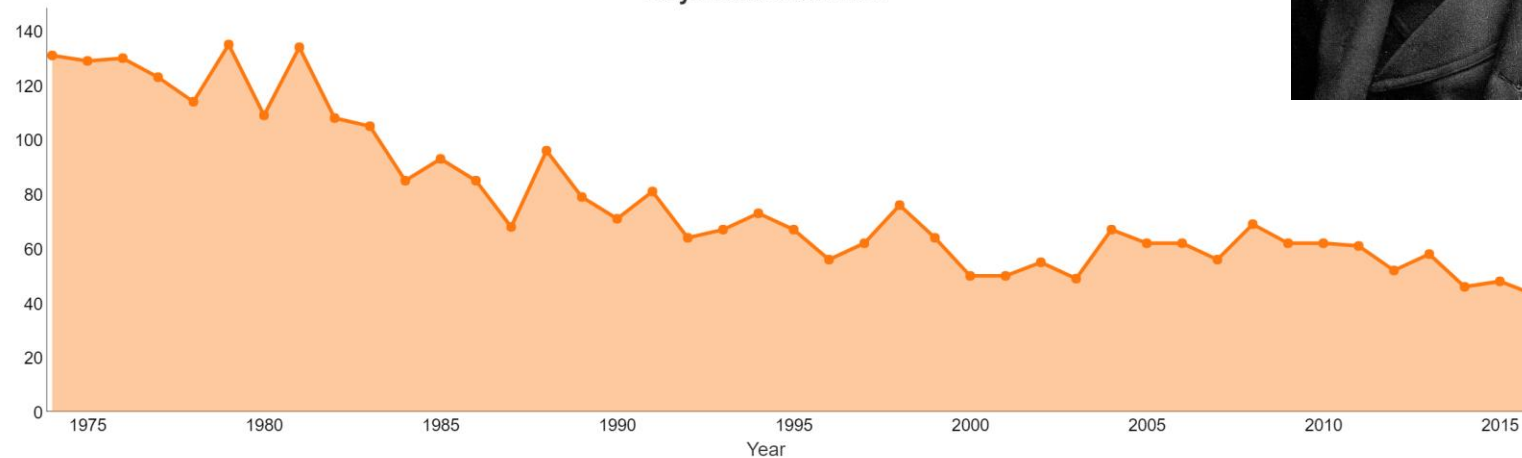
Charles

Gender

Male

Apply

Boys named Charles



How to use this app

Hover over years to highlight individual values
Click and drag to zoom
Double-click to zoom out

Data: [Baby names, Scotland, 1974-2016 \(xlsx\)](#)

Data: [Baby names, Scotland, 1974-2016 \(csv\)](#)

Publications: [Baby names, Scotland, 1974-2016](#)

[National Records of Scotland](#)

© Crown Copyright 2017 - Copyright conditions

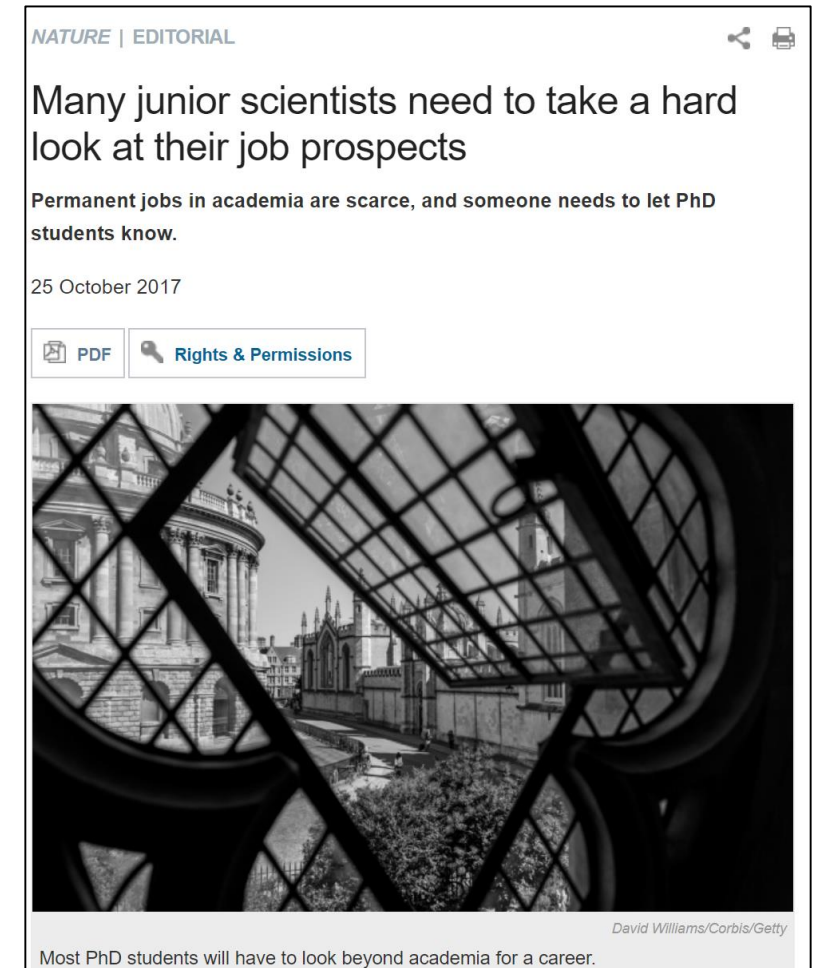
Follow us on Twitter - [@NatRecordsScot](#)

See more [Infographics & Visualisations](#)



Programmers are in demand.

- Massively transferable skill!
- Makes you competitive for postdocs and academic positions.
- Easy path to Python and other languages.
- Research companies, Facebook, Google, Twitter, AirBnB.
- R jobs in Edinburgh at the Scottish Government, RBS, Tesco Bank, Rockstar North, University of Edinburgh, Data Lab, Energy Companies.





facilitates reproducible research.

“Reproducibility is the ability of an entire experiment or study to be reproduced, either by the researcher or by someone else working independently, [and] is one of the main principles of the scientific method.”

-Wikipedia

In the lab:

8/27/08

OPERON-LIKE ORGANIZATION OF THE GAL GENES

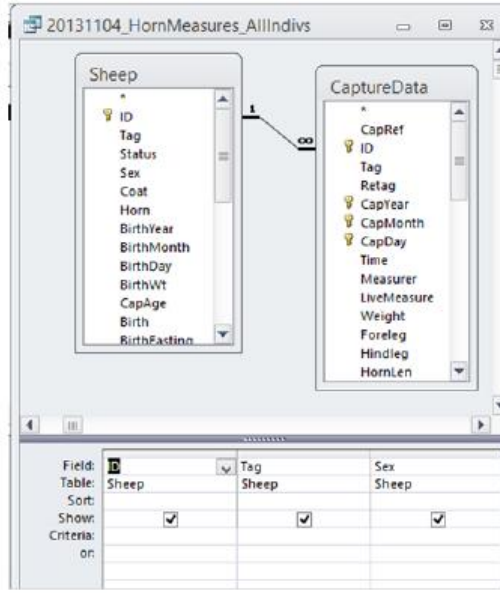
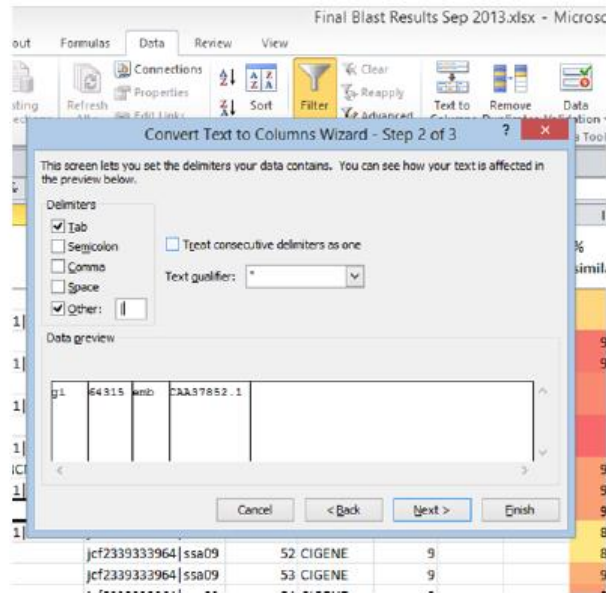
Although eukaryotes lack true operons, there are examples of operon-like gene clusters. Three examples are the galactose utilization genes in *S. cerevisiae* (*GAL1, GAL10, GAL7*), the allantoin degradation genes in *S.c.* (*DAL1, DAL2, DAL3, DAL4, DAL7, DCG1*), and the thiaminol synthesis genes in *Arabidopsis* (*THAS, THAH, THAD*):



Two explanations have been given to account for this organization: ~~genetic linkage~~ and metabolic channeling.

The genetic linkage hypothesis seems to be favored in the literature. It is interesting to note, however, that all three pathways above have intermediates that are toxic to the organism (in red). Here I want to test the hypothesis that the operon-like organization allows for better co-regulation of the genes and helps ~~channel~~ maintain flux through the pathway thus prevent the accumulation of the toxic intermediate →

Many of us are clicking, copying and pasting...



Haggis population density in the Scottish Highlands

S Johnston, University of Edinburgh.

Introduction.

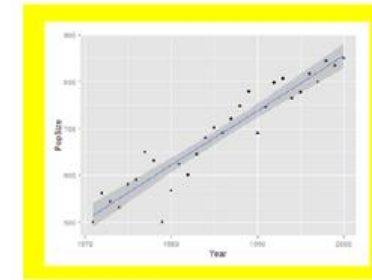


Figure 1: Linear regression of haggis population size and year.

The haggis is a common pest species in the Scottish Highlands. Haggis population densities were recorded annually from 1970 to 2000. We found that the haggis population size increased over this period by 11.67 individuals year⁻¹ ($P < 0.001$, Figure 1).

- Can you repeat all of this again. . .
- . . . and would you get the same results every time?

Worst Case Scenario

Retraction Watch

Archive for the 'not reproducible' Category

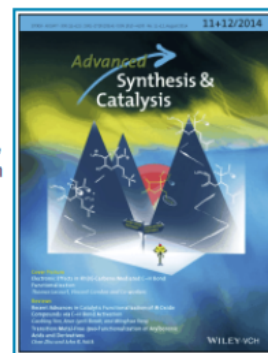
Molecular mixup burns chemistry paper

without comments

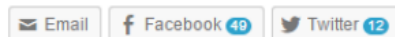
Chemists at Lanzhou University in China did the right thing last month, retracting a [paper](#) in *Advanced Synthesis & Catalysis* because of issues with a reactant that could only be corrected by changing "all the text and quantities."

When the scientists were adding what was labeled Reactant 1 to the mix, they believed it was α -ethoxycarbonyl- α -azido-*N*-phenylacetamides. Unfortunately, what they were actually using was a decomposed version of the molecule, which threw everything off.

Here's the [notice](#) for "*tert*-Butyl Hydroperoxide and Tetrabutylammonium Iodide- Promoted Free Radical Cyclization of α -Azido-*N*-arylamides": [Read the rest of this entry »](#)



Share this:



Written by Cat Ferguson
April 14th, 2015 at 11:30 am

Posted in [Advanced Synthesis and Catalysis](#), [chemistry retractions](#), [china retractions](#), [doing the right thing](#), [freely available](#), [not reproducible](#), [wiley](#)

Two more retractions bring lab break-in biochemist up to eleven

without comments

[Karel Bezouška](#), the Czech biochemist who was caught on hidden camera breaking into a lab fridge to fake results, has [turned it up to eleven](#) with two new retractions.

Both retractions appeared in *Biochemical and Biophysical Research Communications*, one in October 2014 and one in January 2015. His story began two decades ago in 1994, when he published a paper in *Nature* that couldn't be reproduced, and was [eventually retracted in 2013](#).

The best part of the story, of course, is that when his university was attempting to recreate his experiments, Bezouška broke into a lab fridge to tamper with the experiments. Unbeknownst to him, he was caught on hidden camera. [Read the rest of this entry »](#)

Tracking retractions as a window into the scientific process

Subscribe to Blog

Email Address

Subscribe

Pages

[How you can support](#)

[Retraction Watch](#)

[Meet the Retraction Watch staff](#)

[About Adam Marcus](#)

[About Ivan Oransky](#)

[The Center For Scientific Integrity](#)

[Board of Directors](#)

[The Retraction Watch FAQ, including comments policy](#)

[The Retraction Watch Transparency Index](#)

[The Retraction Watch Store](#)

[Upcoming Retraction Watch appearances](#)

[What people are saying about Retraction Watch](#)

Search for:

Search

Search

Recent Posts

[Beleaguered Förster turns down prestigious professorship, citing personal toll](#)

We're on Facebook

Retraction Watch

Scenarios that benefit from reproducibility



- The first researcher who will need to reproduce results is likely to be **YOU**.
- New data becomes available.
- You return to a project after a period of time.
- You give the project to a new student/collaborator.
- A reviewer wants you to change a model parameter.
- When you find an error, but not sure where you went wrong.

Using R as a Research Tool: Overview

- Getting Started: R and the Rstudio Environment.
- Loading data into R.
- Data Management in R with **dplyr**.
- Data visualisation with **ggplot2**.

FREE COURSE

Introduction to R

[Continue Course](#)

 4 hours |  0 Videos |  62 Exercises |  579,784 Participants |  6200 XP | **DOWNLOAD THE APP:**  

Course Description

In this introduction to R, you will master the basics of this beautiful open source language, including factors, lists and data frames. With the knowledge gained in this course, you will be ready to undertake your first very own data analysis. With over 2 million users worldwide R is rapidly becoming the leading programming language in statistics and data science. Every year, the number of R users grows by 40% and an increasing number of organizations are using it in their day-to-day activities. Leverage the power of R by completing this free R online course today!

This course is part of these tracks:

Data Analyst with R

Data Scientist with R

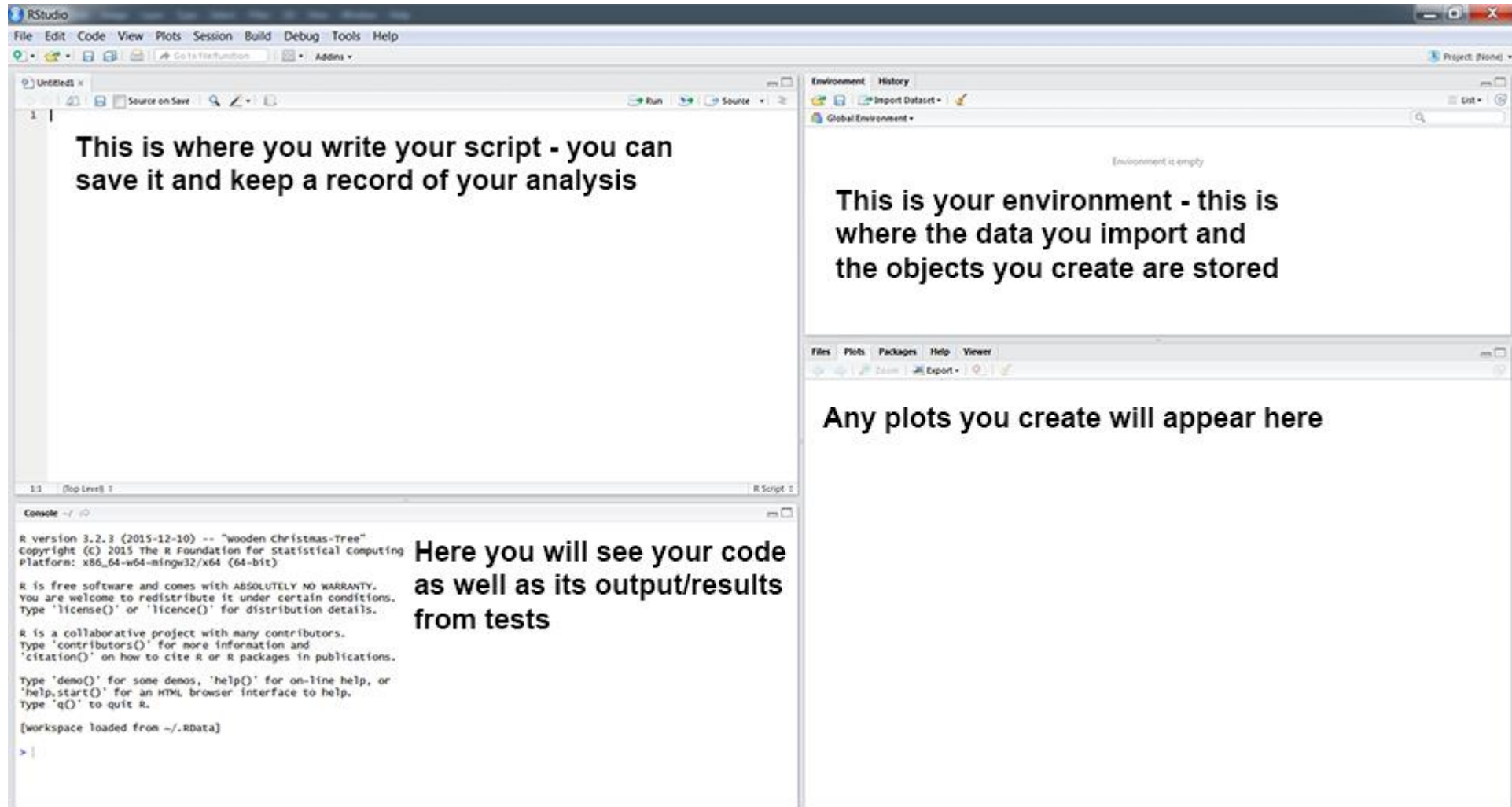
R Programmer

R Programming

Mhairi's Base R cheatsheet...

<http://github.com/rstudio/cheatsheets/raw/master/base-r.pdf>

R and the Rstudio Environment



Finding help.

- In R...

- ? searches for a specific function.
- ?? searches for a specific string.
- Help tab in RStudio

- Online...

- ourcodingclub.github.io
- Stack Overflow
- R Cheatsheets

Loading data into R

Data management in R with base R & `dplyr`

- Summarise data with `summary()`
- Sort data with `arrange()`
- Select columns with `select()`
- Adding columns with `$`
- Select rows with `filter()`

filter()

Operator	Function
<	less than
>	greater than
=<	less than or equal to
=>	greater than or equal to
==	equals
!=	does not equal
%in%	matches

Data visualisation with **ggplot2**

<http://ggplot2.tidyverse.org/reference/>

Base graphics...

<http://rpubs.com/SusanEJohnston/7953>

ggplot2 uses three components to construct a graph.

- Layers: **ggplot()**
 - Data with aesthetic properties (**aes()**)
- Geoms: **geom_...()**
 - Type of plot (line, scatter, box-plot, etc).
- Stats: **stat_...()**
 - Statistical transformations
 - NB. Most geoms have a default stat, so this is not always need.

Using R as a Research Tool: Overview

- Getting Started: R and the Rstudio Environment.
- Loading data into R.
- Data Management in R with **dplyr**.
- Data visualisation with **ggplot2**.