

R-reproducible workflows

1-day workshop

Morning practical session



Brendan Palmer,

Statistics & Data Analysis Unit,

Clinical Research Facility - Cork

A: Tibbles



- The tidyverse equivalent of `data.frames`

4 main points of difference:

1. Printing in the console
2. Subsetting (The use of a placeholder ("`.`")
3. Interacting with older code
4. Tibbles don't change the input

- Open the script `01_pm_tibbles.R`

A: readr and more



- fast way to read rectangular data (like csv, tsv)
- `read_csv()`: comma separated (CSV) files
- `read_tsv()`: tab separated files
- `read_delim()`: general delimited files



- readxl supports both the legacy .xls format and the modern xml-based .xlsx format
- Need to load explicitly



- `read_sas()`: SAS files
- `read_sav()`: SPSS files
- `read_dta()`: Stata files
- Also need to load explicitly
- Open the script 02_pm_readr.R

[Advanced Search](#)

IRELAND'S **OPEN DATA** PORTAL

Promoting innovation and transparency through the publication of Irish Public Sector data in open, free and reusable formats.

[Explore Datasets](#)

5481

Datasets

102

Publishers



Agriculture,
Fisheries,
Forestry &
Food



Arts, Culture
and Heritage



Justice, Legal
System, and
Public Safety



Economy and
Finance



Education and
Sport



Energy



Environment



Government
and Public
Sector



Health



Housing and
Zoning



Population and
Society



Science and
Technology



Regions and
Cities



Transport

Data structures

#vectors:

These come in two forms

- A: Atomic vectors contain exactly one type of data

```
all_numbers      <- c(1, 2, 0.5, -0.5, 3.4)
```

```
all_characters    <- c("One", "too", "3")
```

```
all_logical       <- c(TRUE, FALSE)
```

- B: Lists allow combinations of different types of data

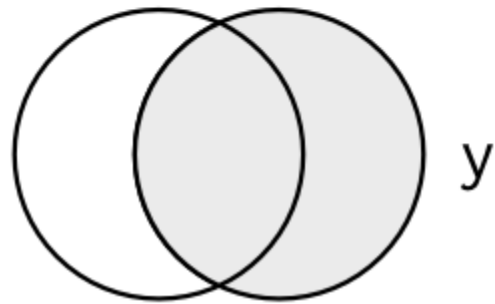
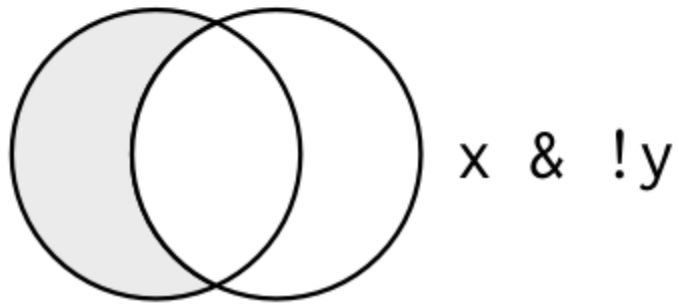
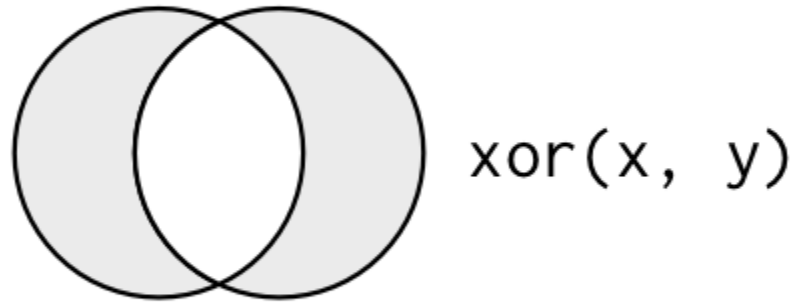
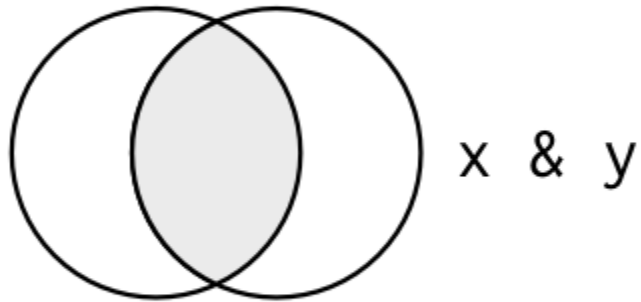
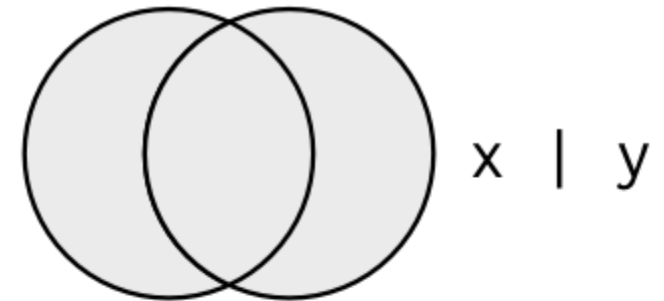
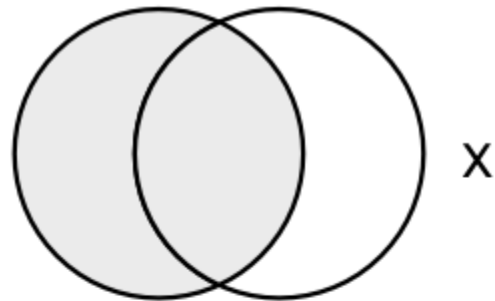
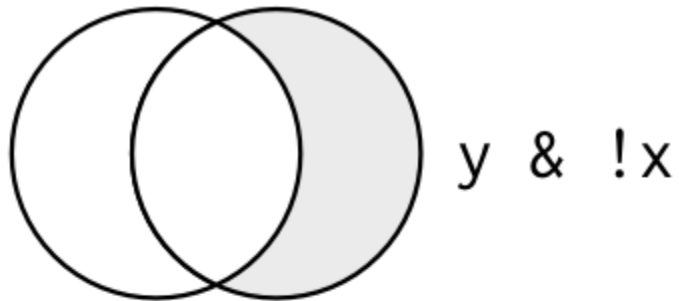
```
this_is_a_list    <- list(1, TRUE, "Three")
```

- If you try to create a vector with more than one data type, then it will undergo coercion to the least common denominator

- The coercion rule goes:

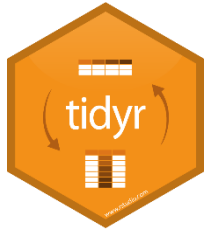
logical -> integer -> numeric -> complex -> character

Logical operators and conditional subsetting



- $\&$ -> AND
- \mid -> OR (inclusive)
- $!$ -> NOT
- $==$ -> EQUAL (identity)
- $!=$ -> NOT EQUAL

A: tidyr



- The tidyverse works with tidy data
 1. Each variable a column
 2. Each observation a row
 3. Each value is a cell



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II.

<http://www.jstatsoft.org/>

Main functions:

- `gather()`
- `separate()`

- Open the script `03_pm_tidyr.R`

Tidy Data

Hadley Wickham
RStudio

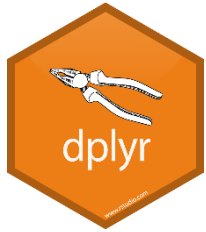
Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

DOI: [10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)

B: dplyr for data transformation



- Solves the most common data manipulation challenges

Main functions:

- `select()`
- `filter()`
- `mutate()`
- `group_by()`
- `summarise()`
- and many many more
- Open the script `04_pm_dplyr.R`

Time for some hands on application

- Open the script `05_pm_practise.R`

C: ggplot2



- Data visualisation based on "[The Grammar of Graphics](#)"

`ggplot(data = <DATA>) +`

`<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>)) +`

`linear model +`

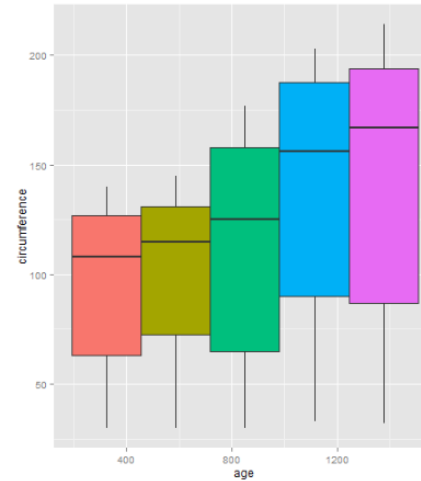
`axes formatting +`

`legend formatting +`

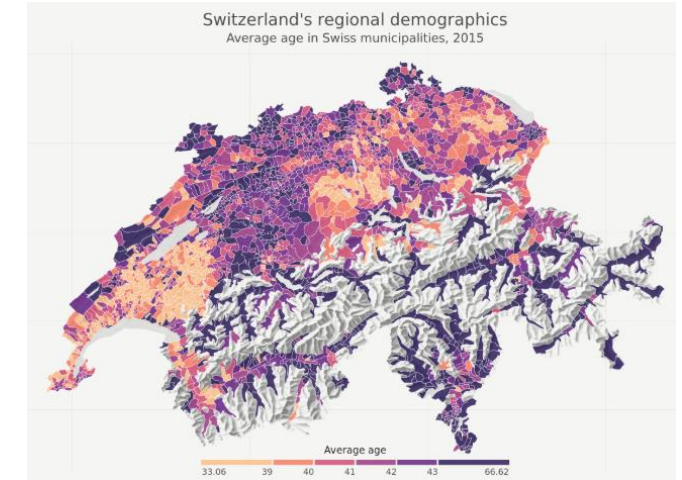
`title + etc. etc.`

C: ggplot2

- Very versatile
- Allows you to go from

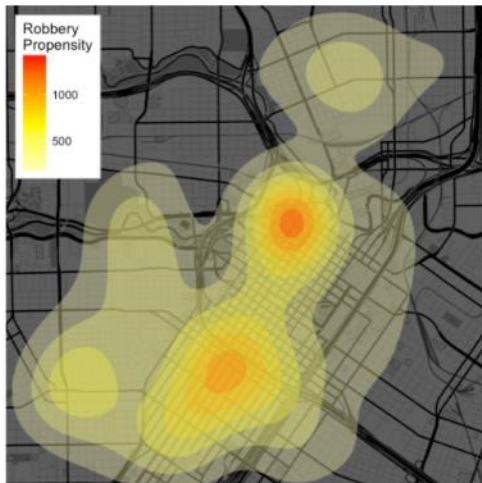


to

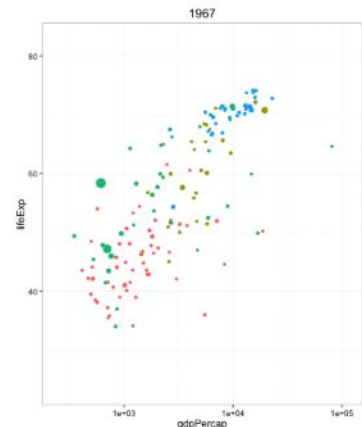


- Lots of add-on packages

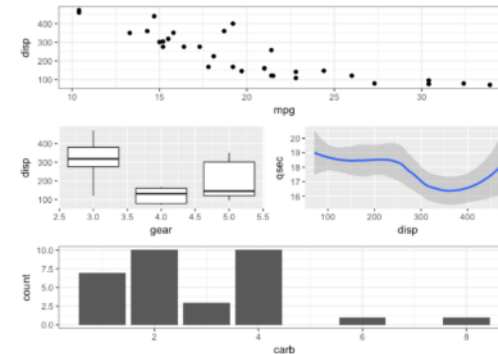
ggmap



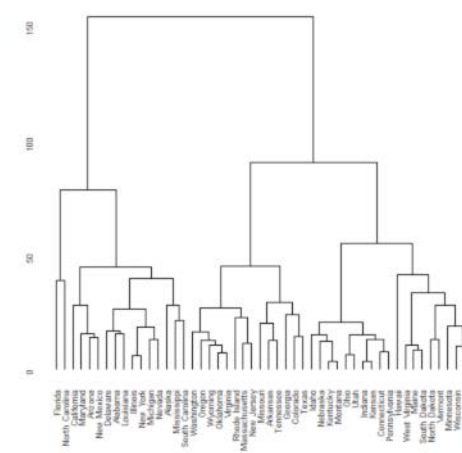
gganimate



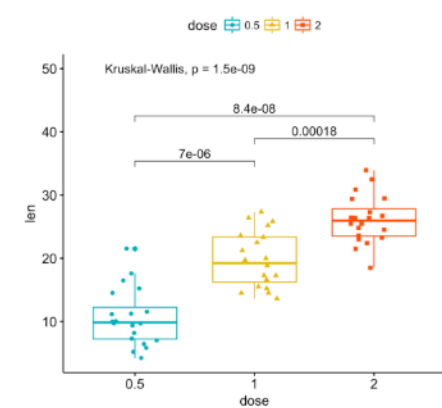
patchwork



ggdendro



ggpubr

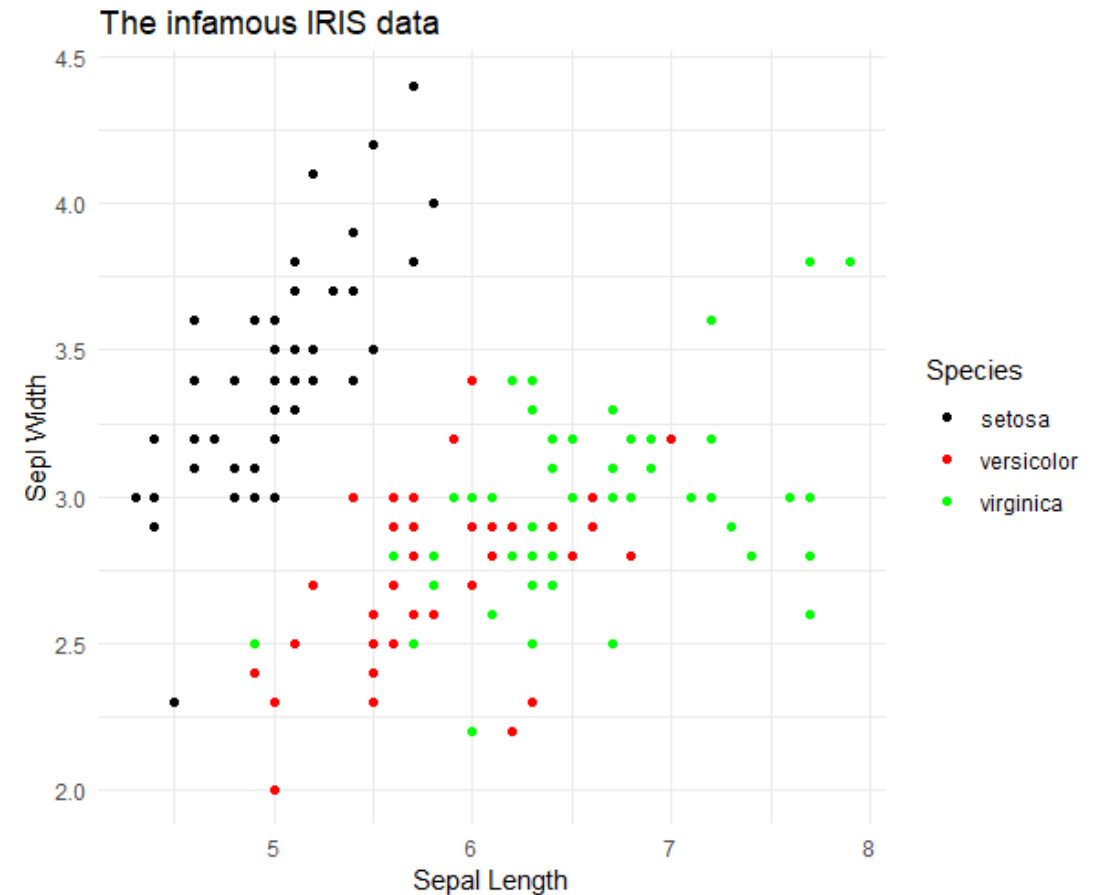
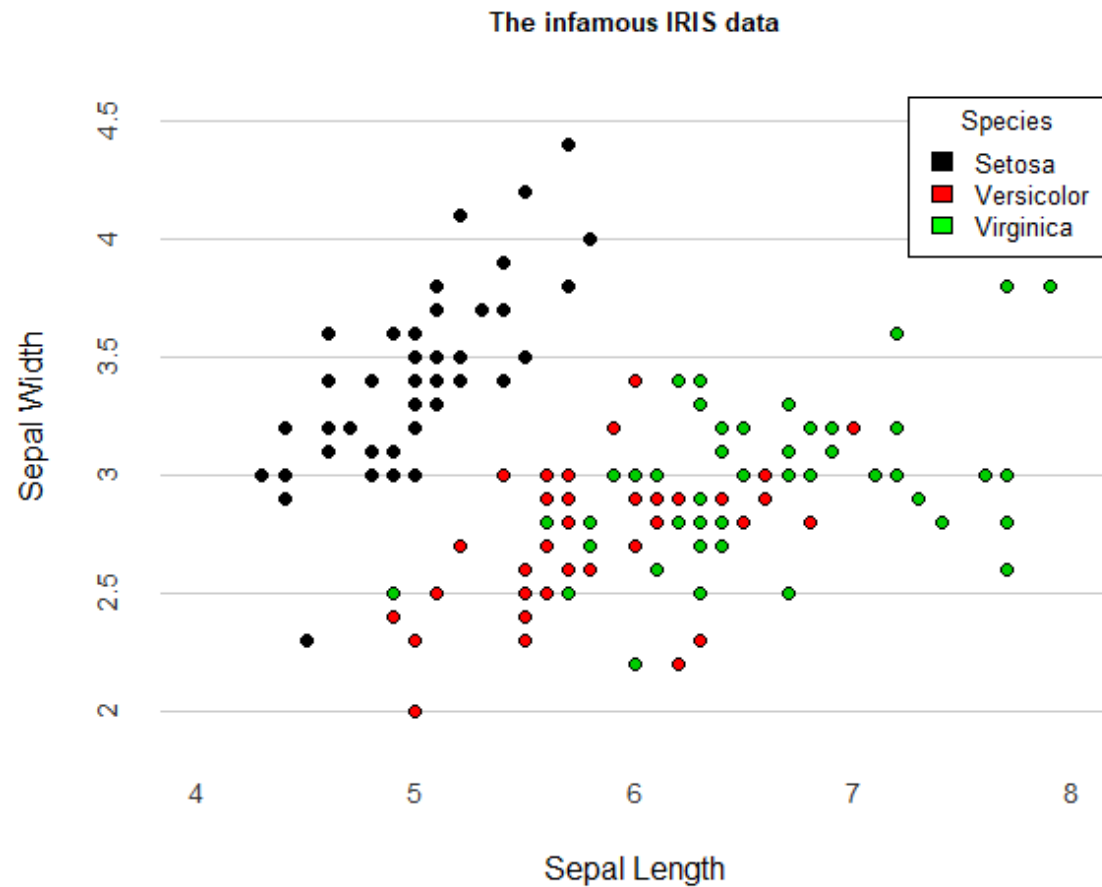


C: Plotting using base R graphics vs ggplot2

```
7 # Here's an example using the graphics packages that comes with base R
8 plot(iris$Sepal.Length, iris$Sepal.Width,
9      bg = iris$Species, # Fill colour
10     pch = 21, # Shape: circles that can be filled
11     xlab = "Sepal Length", ylab = "Sepal Width", # Labels
12     axes = FALSE, # Don't plot the axes
13     frame.plot = FALSE, # Remove the frame
14     xlim = c(4, 8), ylim = c(2, 4.5), # Limits
15     panel.first = abline(h = seq(2, 4.5, 0.5), col = "grey80"))
16
17 at = pretty(iris$Sepal.Length)
18 mtext(side = 1, text = at, at = at,
19       col = "grey20", line = 1, cex = 0.9)
20
21 at = pretty(iris$Sepal.Width)
22 mtext(side = 2, text = at, at = at, col = "grey20", line = 1, cex = 0.9)
23
24 legend("topright", legend = c("Setosa", "Versicolor", "Virginica"),
25       title = "Species", fill = c("black", "red", "green"), cex = 0.8)
26
27 title("The infamous IRIS data",
28       cex.main = 0.8, font.main = 2, col.main = "black")
29
```

```
30 # Now let's view the ggplot2 version
31 # library(tidyverse)
32 ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, colour = Species)) +
33   geom_point() +
34   scale_colour_manual(values = c("black", "red", "green")) +
35   theme_minimal() +
36   labs(title = "The infamous IRIS data",
37        x = "Sepal Length",
38        y = "Sepal Width")
39
```

C: Plotting using base R graphics vs ggplot2



- Open the script `06_pm_graphics_example.R` to see for yourself

C: Whistle-stop tour of ggplot2

Main features:

1. The data
2. The geoms
3. The mappings (x, y, colour, shape etc.)
4. Legends
5. Labels
6. Themes

and many many more

- Open the script 07_pm_ggplot2.R
- Open the script 08_pm_practise_plots.R

Where to next?

[Products](#)[Resources](#)[Pricing](#)[About Us](#)[Blogs](#)

rstudio::conf 2019



RStudio Connect Past, present, and future - Jeff Allen

RStudio Connect is a publishing platform that helps to operationalize the data science work you're doing. We'll review the current state and unveil some exciting new features.

[WATCH VIDEO >](#)

Shiny in production: Principles, practices, and tools - Joe Cheng

Shiny is a web framework for R. As such, Shiny has always faced

[WATCH VIDEO >](#)

Welcome and RStudio Vision - Tareef Kawaf

[WATCH VIDEO >](#)

Working with names and expressions in your tidy eval code - Lionel Henry

In practice there are two main flavors of tidy eval functions:

[WATCH VIDEO >](#)

Why TensorFlow eager execution matters - Sigrid Keydana

In current deep learning with Keras and TensorFlow, when

[WATCH VIDEO >](#)

Visualizing uncertainty with hypothetical outcomes plots - Claus Wilke

[WATCH VIDEO >](#)

vctrs: Tools for making size and type consistent functions - Hadley Wickham

[WATCH VIDEO >](#)

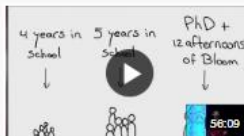
Using R, the Tidyverse, H2O, and Shiny to reduce employee attrition - Matt Dancho

An organization that loses 200



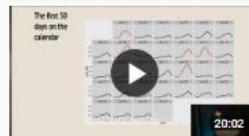
pagedown: Creating beautiful PDFs with R. Markdown and CSS - Yihui Xie

Traditional way to create PDF



Explicit Direct Instruction in Programming Education - Felienne

In education, there is and has always been debate about how to



Melt the clock Tidy time series analysis - Earo Wang

Time series can be frustrating to work with, particularly when



The unreasonable effectiveness of public work - David Robinson

In this talk, I'll lay out the reasons that blogging, open source

WELCOME!

[Follow @rbloggers](#) 66.2k

Here you will find daily news and tutorials about R, contributed by over 750 bloggers.

There are many ways to follow us -
By e-mail:

Your e-mail here

Subscribe

51562 readers
BY FEEDBURNER

On Facebook:



R blog...
75K likes

rstudio::conf 2019 Workshop materials now available

February 5, 2019

By Mine Çetinkaya-Rundel

[Like 16](#) [Share](#) [Share](#)

(This article was first published on [RStudio Blog](#), and kindly contributed to R-bloggers)

[f](#) Share

[T](#) Tweet

rstudio::conf 2019 featured 15 workshops on tidyverse, Shiny, R Markdown, modeling and machine learning, deep learning, big data, and what they forgot to teach you about working with R. Some of the new workshops for this year touched on topics like putting Shiny applications into production at scale and R

SEARCH R-BLOGGERS

RECENT POPULAR POSTS

Learning Data Science: Modelling Basics
Using the uniform sum distribution to introduce probability

MOST VISITED ARTICLES OF THE WEEK

1. How to write the first for loop in R
2. How GPL makes me leave R for Python :-(
3. How to Make a Histogram with Basic R
4. 5 Ways to Subset a Data Frame in R
5. Using apply, sapply, lapply in R