



Data Science and Statistics in Research: unlocking the power of your data

Session 1.2: An introduction to R

OUTLINE

Introduction

R and RStudio

Statistical Analyses

Packages

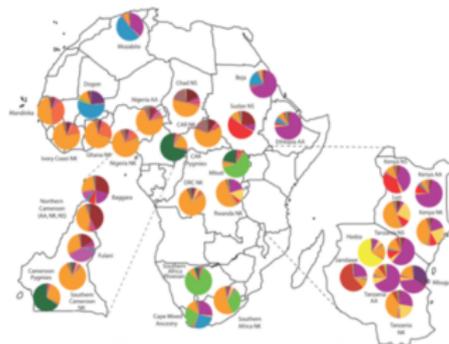
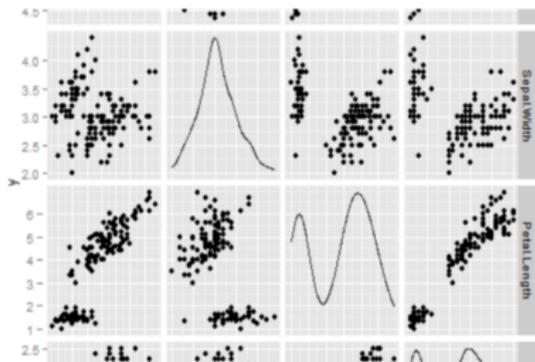
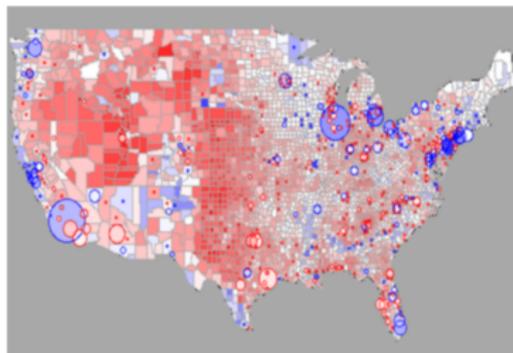
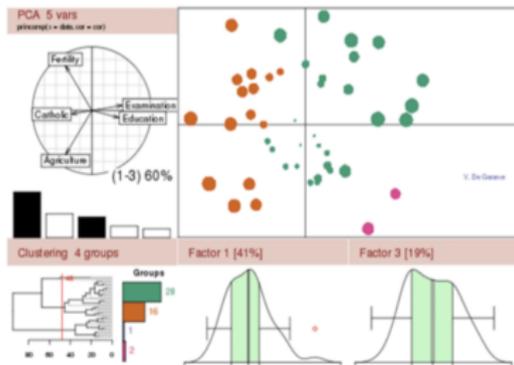
Help

Introduction

STATISTICAL SOFTWARE

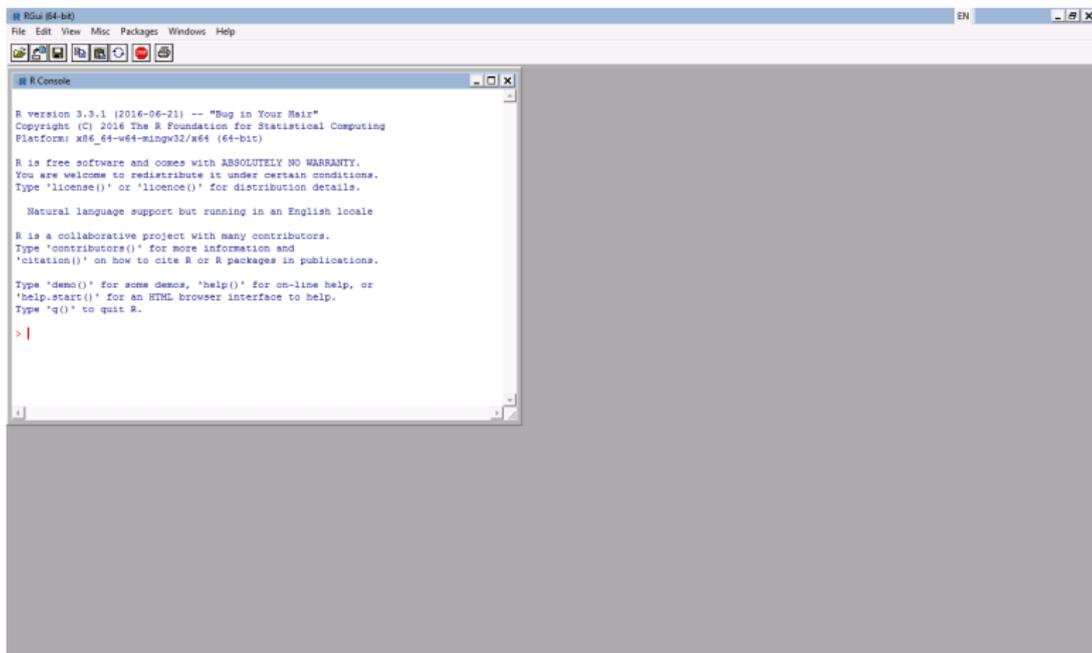
- ▶ Excel
 - ▶ simple descriptive statistics, plots, and regression can be done in the basic installation of Excel
 - ▶ the Analysis Toolpak allows many more methods to be used such as ANOVA and hypothesis tests.
- ▶ SPSS, SAS, Stata
 - ▶ general purpose statistical packages that can perform a very wide variety of analyses
 - ▶ cover everything from initial descriptive analyses to very complex methods.
 - ▶ GUI interfaces: functions found by menus.
- ▶ R
 - ▶ a language and environment for statistical computing and graphics
 - ▶ open source with many many user packages
 - ▶ it's free! (Open-Source)

R GRAPHICS



R and RStudio

R



The screenshot shows the R GUI (64-bit) window. The title bar reads "# RGui (64-bit)" and the menu bar includes "File", "Edit", "View", "Misc", "Packages", "Windows", and "Help". The main window contains an "R Console" pane with the following text:

```
R version 3.3.1 (2016-06-21) -- "Bug In Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

RSTUDIO

RStudio

Project: (None)

Environment History

Global Environment

Data

- dig 6800 obs. of 72 variables
- states.df 15527 obs. of 7 variables

Values

- my.model Large gam (48 elements, 830.9 Kb)
- stateMapEnv "R_MAP_DATA_DIR"

ID	TRTMT	AGE	RACE	SEX	EJF_PER	EJFMETH	CHESTX	BMI
1	1	0	66	1	1	40	2	0.50
2	2	0	77	1	1	32	1	0.56
3	3	0	72	1	2	36	1	0.68
4	4	1	57	1	1	31	1	0.48
5	5	0	74	1	1	35	1	0.53
6	6	0	69	2	2	45	1	0.70
7	7	1	64	1	2	30	1	0.52
8	8	1	60	2	1	39	1	0.40

Displayed 1000 rows of 6800 (5800 omitted)

```

Console ~ /
4 13040 1 468 0 13040 0 13040 0 13040 0
3 746 0 1391 0 1391 0 1391 0 1391 0
4 1157 0 1157 0 1157 0 1157 0 1157 0
5 1550 0 1550 0 1550 0 1550 0 1550 0
6 1620 0 1620 0 1620 1 496 0 1620 0
OCVDDAYS RINF RINFDDAYS OTH OTHDDAYS HOSP HOSPDAYS NHOSP DEATH DEATHDAY
1 1049 0 1438 1 533 1 533 6 0 1438
2 1360 0 1360 1 880 1 468 4 1 1360
3 1391 0 1391 0 1391 1 631 2 0 1391
4 1157 0 1157 0 1157 0 1157 0 0 1157
5 1550 0 1550 1 459 1 191 5 0 1550
6 1620 0 1620 1 966 1 496 5 0 1620
REASON DWHF DWHFDDAYS
1 NA 1 1379
2 1 1 1329
3 NA 1 631
4 NA 0 1157
5 NA 1 191
6 NA 0 1620
> my.model<-gam(DEATH ~ TRTMT + s(DWHFDDAYS), data=dig, family=binomial)
> summary(my.model)$coef
NULL
> plot(my.model)
> View(dig)
>
  
```

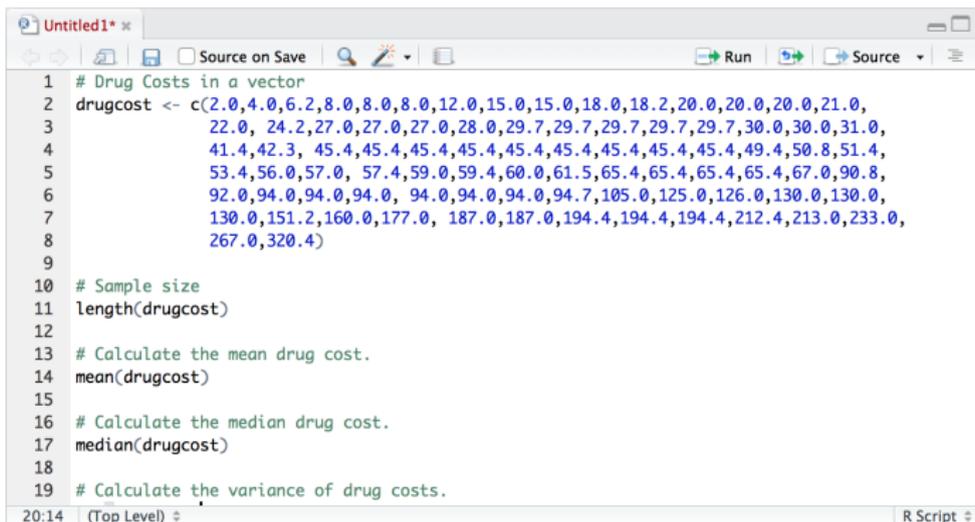
Files Plots Packages Help Viewer

Zoom Export Clear All

RSTUDIO

The editor pane

- ▶ Write, edit and submit R code.



```
1 # Drug Costs in a vector
2 drugcost <- c(2.0,4.0,6.2,8.0,8.0,8.0,12.0,15.0,15.0,18.0,18.2,20.0,20.0,20.0,21.0,
3             22.0, 24.2,27.0,27.0,27.0,28.0,29.7,29.7,29.7,29.7,29.7,30.0,30.0,31.0,
4             41.4,42.3, 45.4,45.4,45.4,45.4,45.4,45.4,45.4,45.4,49.4,50.8,51.4,
5             53.4,56.0,57.0, 57.4,59.0,59.4,60.0,61.5,65.4,65.4,65.4,65.4,67.0,90.8,
6             92.0,94.0,94.0,94.0, 94.0,94.0,94.0,94.7,105.0,125.0,126.0,130.0,130.0,
7             130.0,151.2,160.0,177.0, 187.0,187.0,194.4,194.4,194.4,212.4,213.0,233.0,
8             267.0,320.4)
9
10 # Sample size
11 length(drugcost)
12
13 # Calculate the mean drug cost.
14 mean(drugcost)
15
16 # Calculate the median drug cost.
17 median(drugcost)
18
19 # Calculate the variance of drug costs.
```

20:14 | (Top Level) ⇅ R Script ⇅

RSTUDIO

The editor pane

- ▶ Can also view datasets, similar to Excel in format.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3

Showing 1 to 13 of 32 entries

RSTUDIO

The R console

- ▶ Submitted code and its results appear here.
- ▶ Warnings and errors appear in red.



```
Console ~/1    
> 3+4  
[1] 7  
> 3.4 * 2.5  
[1] 8.5  
> 3.4 . 2.5  
Error: unexpected symbol in "3.4 ."  
>
```

RSTUDIO

The environment and history pane

- ▶ All datasets you are working on appear in the environment window.

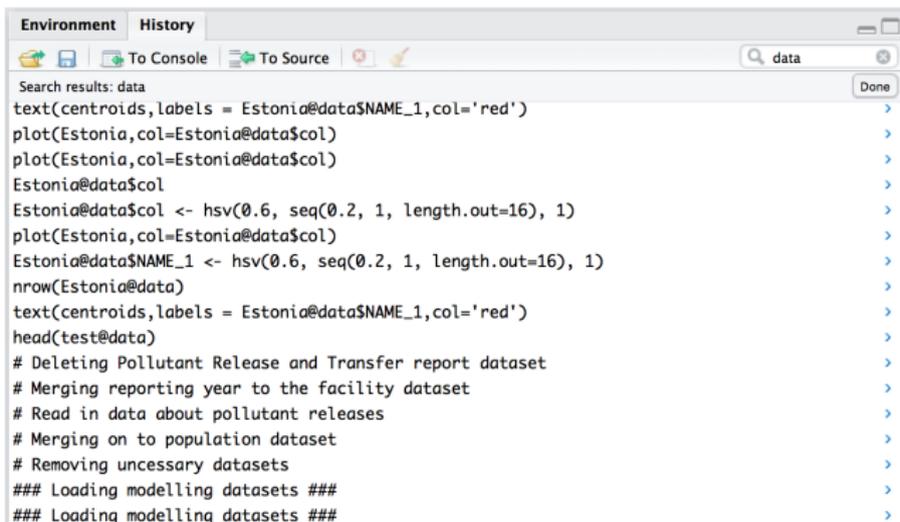
The screenshot shows the RStudio Environment pane. At the top, there are tabs for 'Environment' and 'History'. Below the tabs is a toolbar with icons for 'Import Dataset' and 'List'. The main area is titled 'Global Environment' and contains a search bar. Under the 'Data' section, several objects are listed:

- dat**: 250 obs. of 1 variable
- mtcars**: 32 obs. of 11 variables
 - mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 - cyl : num 6 6 4 6 8 6 8 4 4 6 ...
 - disp: num 160 160 108 258 360 ...
 - hp : num 110 110 93 110 175 105 245 62 95 123 ...
 - drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 - wt : num 2.62 2.88 2.32 3.21 3.44 ...
 - qsec: num 16.5 17 18.6 19.4 17 ...
 - vs : num 0 0 1 1 0 1 0 1 1 1 ...
 - am : num 1 1 1 0 0 0 0 0 0 0 ...
 - gear: num 4 4 4 3 3 3 3 4 4 4 ...
 - carb: num 4 4 1 1 2 1 4 2 2 4 ...
- sleep**: 20 obs. of 3 variables
- tmp**: 32 obs. of 11 variables

RSTUDIO

The environment and history pane

- ▶ You can access all previously run code in the history window.

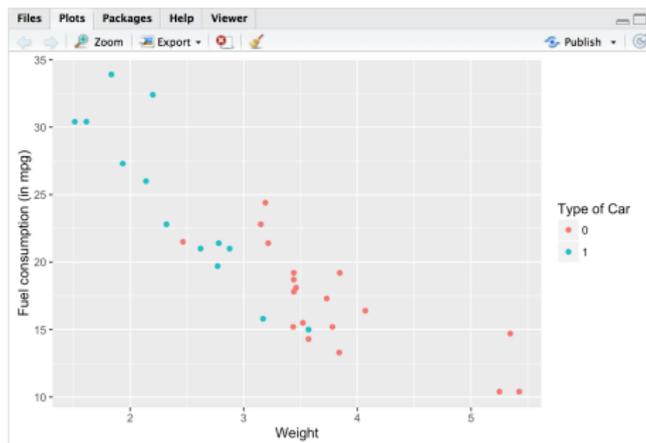


```
Environment History
To Console To Source
Search results: data
text(centroids, labels = Estonia@data$NAME_1, col='red')
plot(Estonia, col=Estonia@data$col)
plot(Estonia, col=Estonia@data$col)
Estonia@data$col
Estonia@data$col <- hsv(0.6, seq(0.2, 1, length.out=16), 1)
plot(Estonia, col=Estonia@data$col)
Estonia@data$NAME_1 <- hsv(0.6, seq(0.2, 1, length.out=16), 1)
nrow(Estonia@data)
text(centroids, labels = Estonia@data$NAME_1, col='red')
head(test@data)
# Deleting Pollutant Release and Transfer report dataset
# Merging reporting year to the facility dataset
# Read in data about pollutant releases
# Merging on to population dataset
# Removing unnecessary datasets
### Loading modelling datasets ###
### Loading modelling datasets ###
```

RSTUDIO

The plot, packages and help window

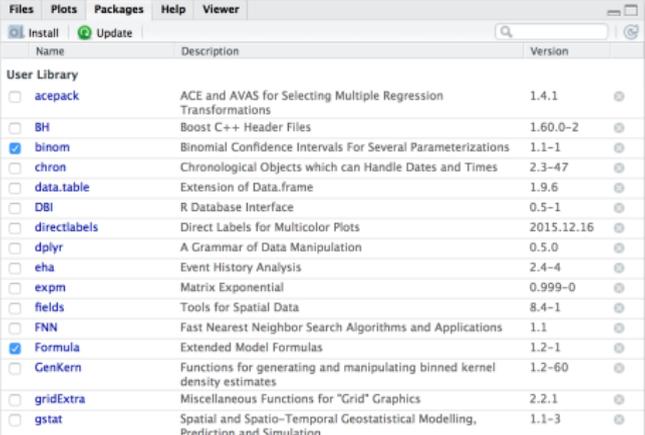
- ▶ Plots will display in the 'Plots' window.



RSTUDIO

The plot, packages and help window

- ▶ Installed packages are listed in the 'Packages' window.



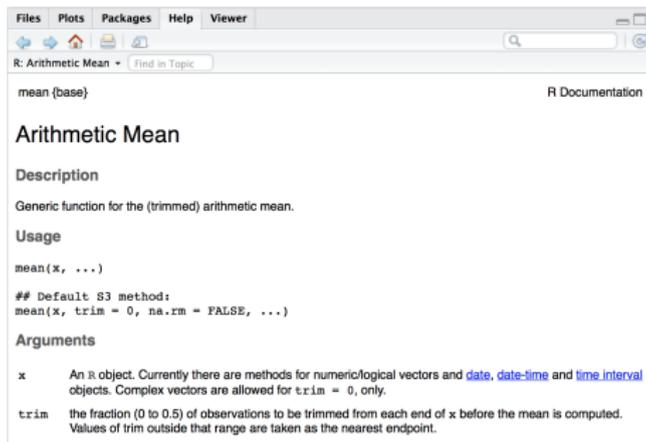
The screenshot shows the 'Packages' window in RStudio. The window has tabs for 'Files', 'Plots', 'Packages', 'Help', and 'Viewer'. The 'Packages' tab is active, showing a list of packages with columns for 'Name', 'Description', and 'Version'. A search bar and an 'Update' button are at the top. The packages listed are:

Name	Description	Version
<input type="checkbox"/> acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1
<input type="checkbox"/> BH	Boost C++ Header Files	1.6.0.0-2
<input checked="" type="checkbox"/> binom	Binomial Confidence Intervals For Several Parameterizations	1.1-1
<input type="checkbox"/> chron	Chronological Objects which can Handle Dates and Times	2.3-47
<input type="checkbox"/> data.table	Extension of Data.frame	1.9.6
<input type="checkbox"/> DBI	R Database Interface	0.5-1
<input type="checkbox"/> directlabels	Direct Labels for Multicolor Plots	2015.12.16
<input type="checkbox"/> dplyr	A Grammar of Data Manipulation	0.5.0
<input type="checkbox"/> eha	Event History Analysis	2.4-4
<input type="checkbox"/> expm	Matrix Exponential	0.999-0
<input type="checkbox"/> fields	Tools for Spatial Data	8.4-1
<input type="checkbox"/> FNN	Fast Nearest Neighbor Search Algorithms and Applications	1.1
<input checked="" type="checkbox"/> Formula	Extended Model Formulas	1.2-1
<input checked="" type="checkbox"/> GenKern	Functions for generating and manipulating binned kernel density estimates	1.2-60
<input type="checkbox"/> gridExtra	Miscellaneous Functions for "Grid" Graphics	2.2.1
<input type="checkbox"/> gstat	Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation	1.1-3

RSTUDIO

The plot, packages and help window

- ▶ Help pages for functions and datasets in the 'Help' window.



The screenshot shows the RStudio Help window for the 'mean' function. The window title is 'R: Arithmetic Mean' and it includes a search bar. The content is organized into sections: 'Description', 'Usage', and 'Arguments'. The 'Description' section states it is a generic function for the (trimmed) arithmetic mean. The 'Usage' section shows the function signature: `mean(x, ...)`. The 'Arguments' section lists `x` as an R object and `trim` as the fraction of observations to be trimmed from each end.

```
mean (base) R Documentation
```

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)
```

Default S3 method:
`mean(x, trim = 0, na.rm = FALSE, ...)`

Arguments

x An R object. Currently there are methods for numeric/logical vectors and [date](#), [date-time](#) and [time interval](#) objects. Complex vectors are allowed for `trim = 0`, only.

trim the fraction (0 to 0.5) of observations to be trimmed from each end of `x` before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.

Statistical Analyses

STATISTICAL ANALYSIS IN R

- ▶ R comes with many statistical tools already installed
 - ▶ descriptive statistics
 - ▶ visualisation
 - ▶ statistical tests
 - ▶ model fitting.

Packages

CAN R DO MORE?

- ▶ The default installation of R has a comprehensive set of tools for statistical analyses.
- ▶ To meet the specific needs of data scientists, many other statistical tools are readily available in the form of packages.
- ▶ Packages are collections of functions and data.
- ▶ "During the last decade, the momentum coming from both academia and industry has lifted R to become the single most important tool for computational statistics, visualisation and data science."

R PACKAGES: EXAMPLES USED IN THIS COURSE

- ▶ ggplot2
- ▶ raster
- ▶ Rmisc
- ▶ mgcv
- ▶ maptools
- ▶ ... many many more!!

OTHER R PACKAGES

- ▶ A list of R Packages can be seen and downloaded from <https://cran.r-project.org>



CRAN

[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R

[R Homepage](#)
[The R Journal](#)

Software

[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation

[Manuals](#)
[FAQs](#)
[Contributed](#)

[A3](#)

[abhyyR](#)

[abc](#)

[ABCanalysis](#)

[abc.data](#)

[abcdeFBA](#)

[ABCoptim](#)

[ABCp2](#)

[ABC.RAP](#)

[abcrf](#)

[abctools](#)

[abd](#)

[abf2](#)

[ABHgenotypeR](#)

[abind](#)

[abn](#)

[abodOutlier](#)

[AbsFilterGSEA](#)

[abundant](#)

Available CRAN Packages By Name

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

Accurate, Adaptable, and Accessible Error Metrics for Predictive Models

Access to Abbyy Optical Character Recognition (OCR) API

Tools for Approximate Bayesian Computation (ABC)

Computed ABC Analysis

Data Only: Tools for Approximate Bayesian Computation (ABC)

ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package

Implementation of Artificial Bee Colony (ABC) Optimization

Approximate Bayesian Computational Model for Estimating P2

Array Based CpG Region Analysis Pipeline

Approximate Bayesian Computation via Random Forests

Tools for ABC Analyses

The Analysis of Biological Data

Load Gap-Free Axon ABF2 Files

Easy Visualization of ABH Genotypes

Combine Multidimensional Arrays

Modelling Multivariate Data with Additive Bayesian Networks

Angle-Based Outlier Detection

Improved False Positive Control of Gene-Permuting GSEA with Absolute Filtering

Abundant regression and high-dimensional principal fitted components

Help

R HELP

- ▶ R-CRAN requires all packages authors to produce manuals detailing the functionality of functions, together with examples of their use.
- ▶ Many forums dedicated to helping people with issues in R
 - ▶ Stack Overflow -
<http://stackoverflow.com/questions/tagged/r>
- ▶ Lots of tutorials available online
 - ▶ Coursera - <https://www.coursera.org/learn/r-programming>
 - ▶ Datacamp - <https://www.datacamp.com>.

Any Questions?