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

Session 1.2:

An introduction to R

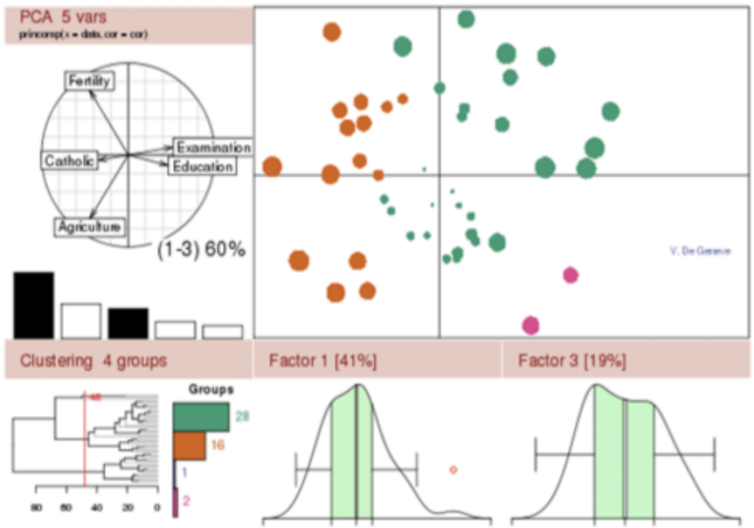
Outline

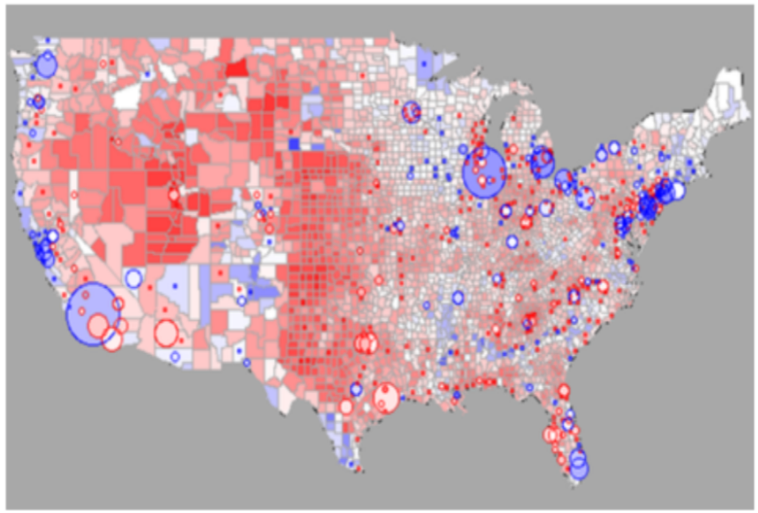
Introduction

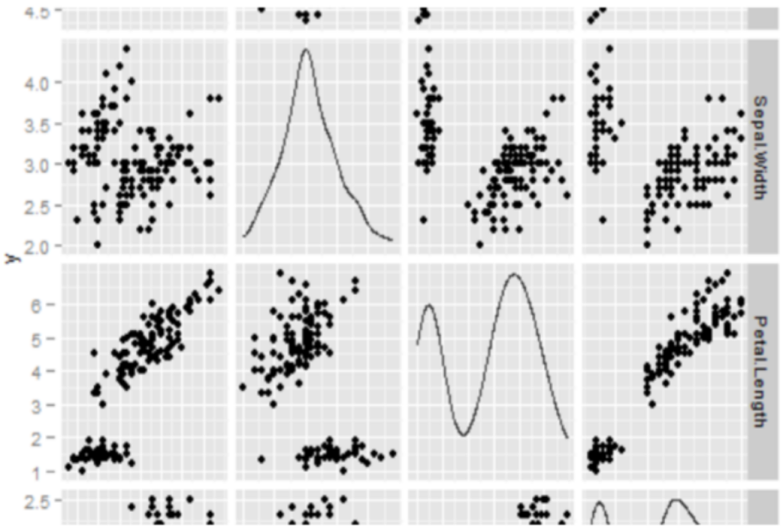
STATISTICAL SOFTWARE

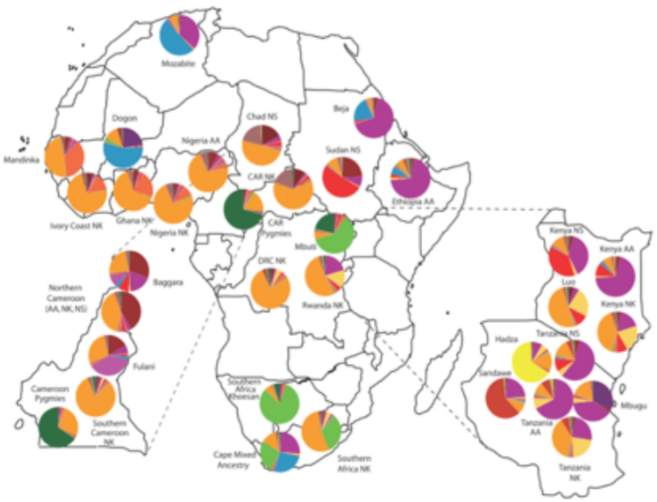
- Excel
  - simple descriptive statistics, plots, and regression can be done in the basic installation of Excel
  - the Analysis Toolpak allows many more methods to be used such as ANOVA and hypothesis tests.
- SPSS, SAS, Stata
  - general purpose statistical packages that can perform a very wide variety 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


PCA 5 vars
Fertility
Catholic
Examination
Education
Agriculture
(1-3) 60%
Clustering 4 groups
Factor 1 [41%]
Factor 3 [19%]
Groups
28
16
1
2

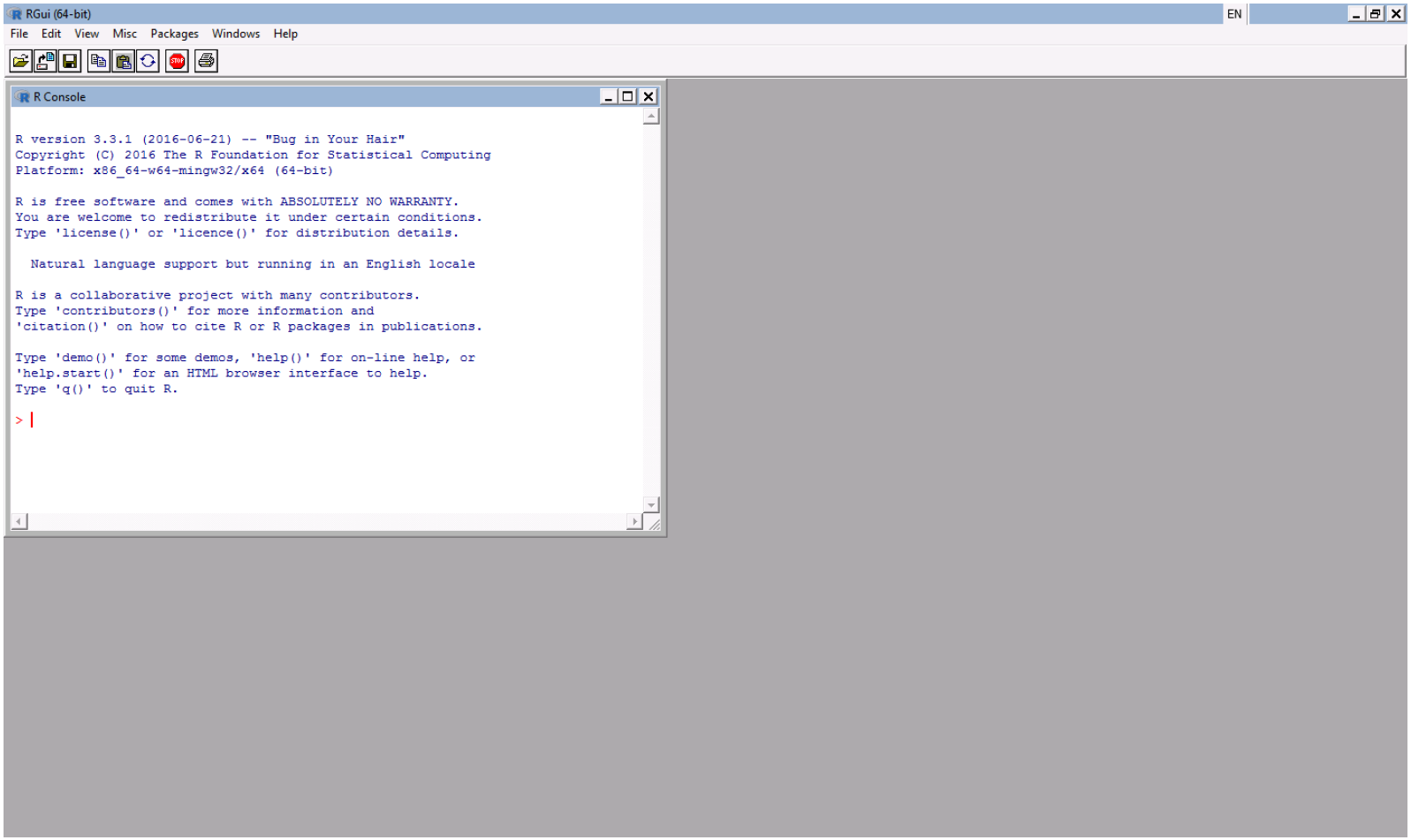



Sepal.Width
Petal.Length
y

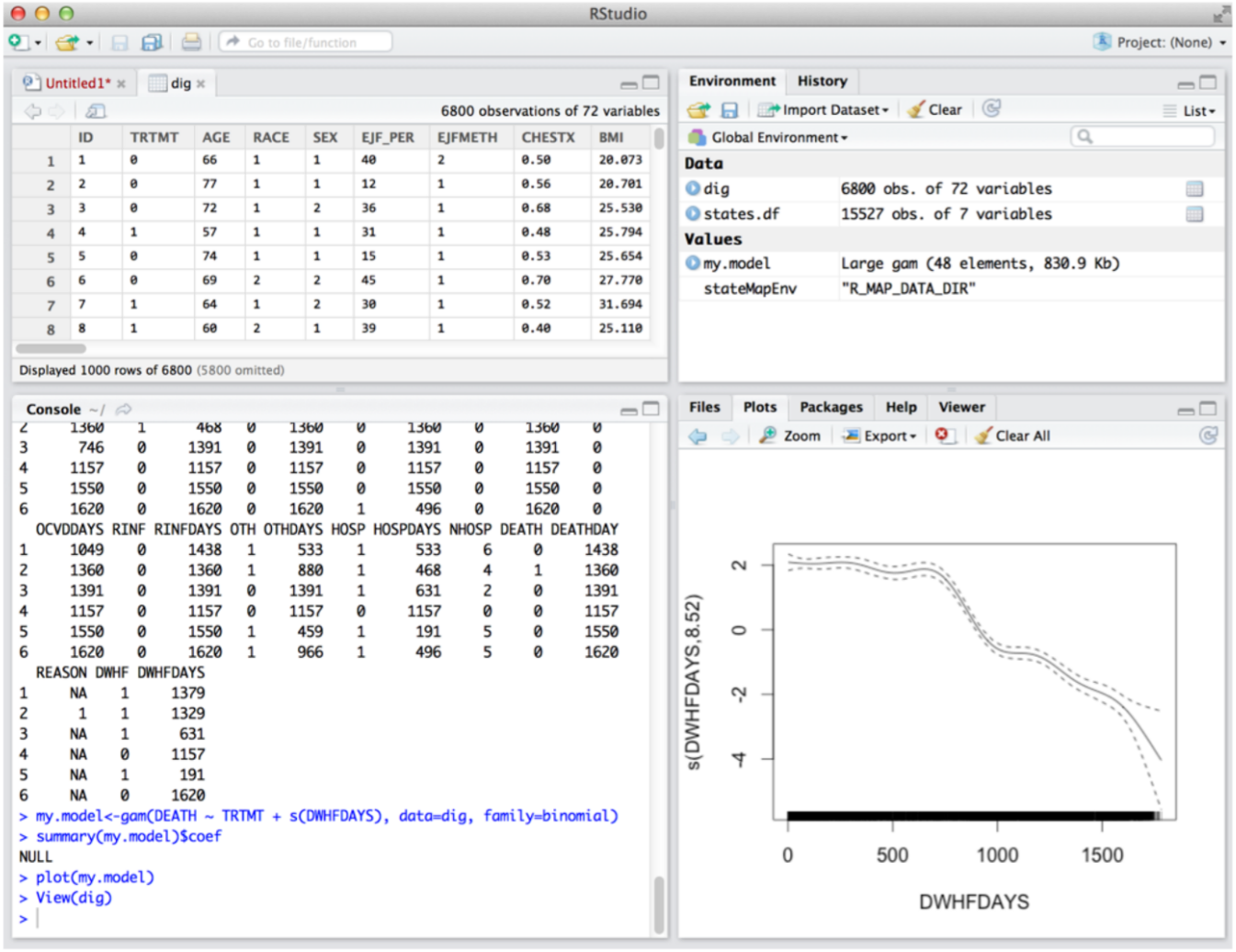


R and RStudio

R


RGui (64-bit)
EN
File Edit View Misc Packages Windows Help
R Console
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.
>

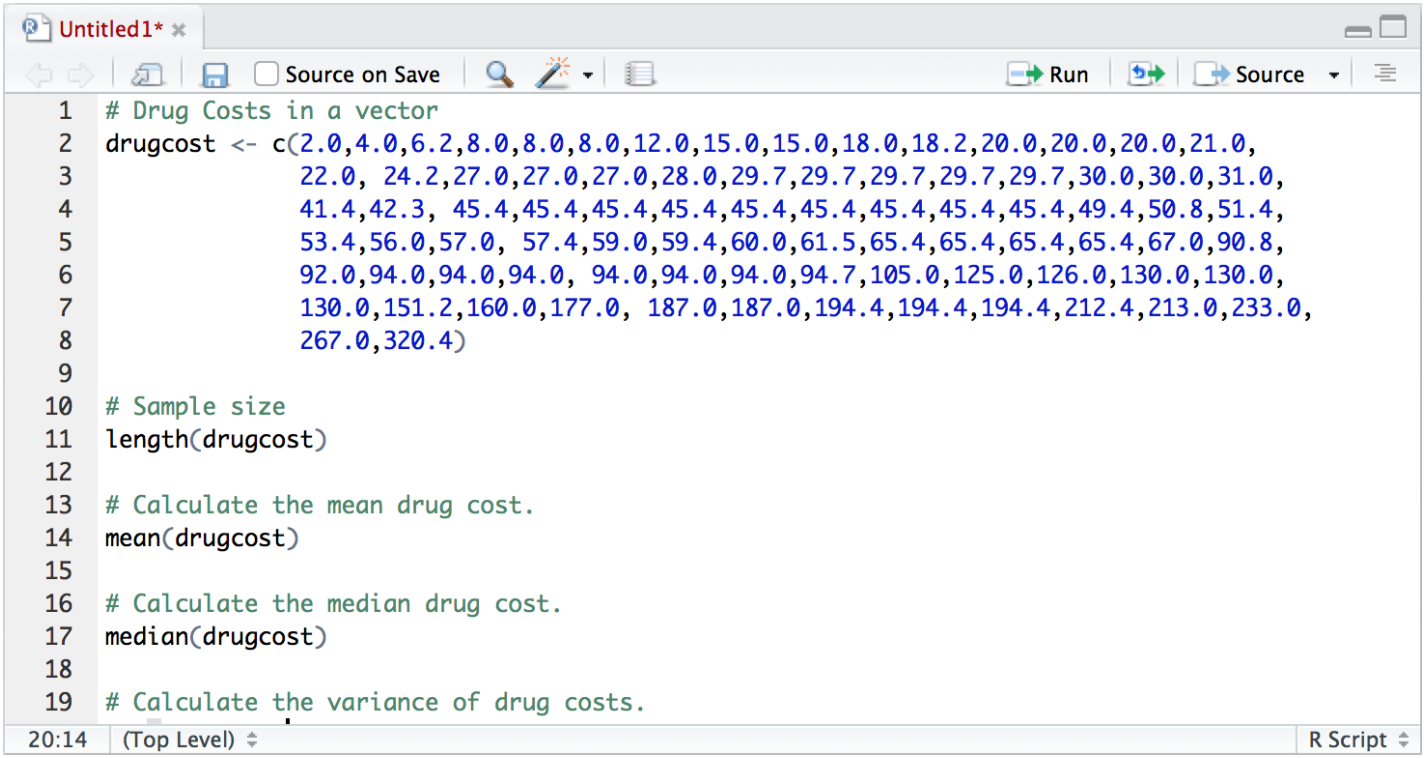
RSTUDIO


RStudio
Untitled1*
dig
6800 observations of 72 variables
Displayed 1000 rows of 6800 (5800 omitted)
Environment
History
Import Dataset
Clear
List
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"
Console
> my.model<-gam(DEATH ~ TRTMT + s(DWHFDAYS), data=dig, family=binomial)
> summary(my.model)$coef
NULL
> plot(my.model)
> View(dig)
Files
Plots
Packages
Help
Viewer
Zoom
Export
Clear All
s(DWHFDAYS,8.52)
DWHFDAYS

RSTUDIO

The editor pane

- Write, edit and submit R code.



RSTUDIO

The editor pane

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

mtcars

Filter

	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

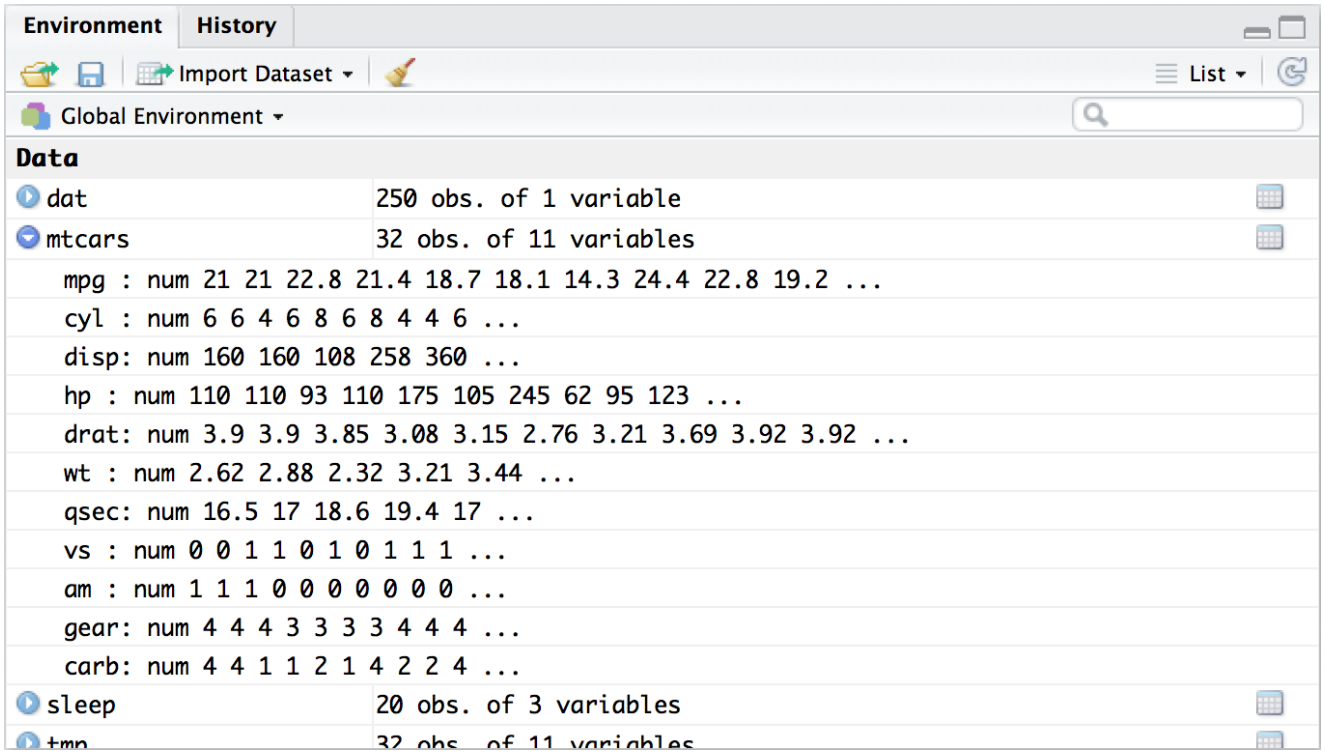
The R console

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

RSTUDIO

The environment and history pane

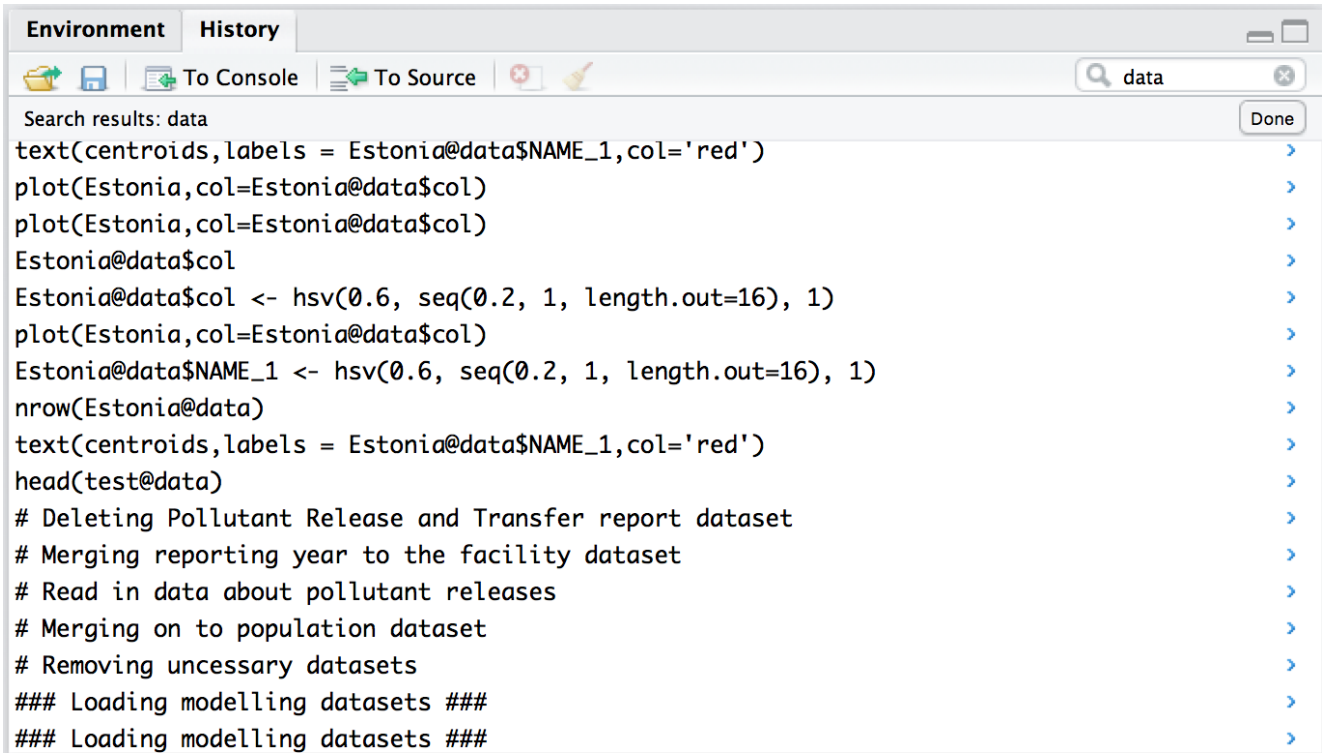
- All datasets you are working on appear in the environment window.



RSTUDIO

The environment and history pane

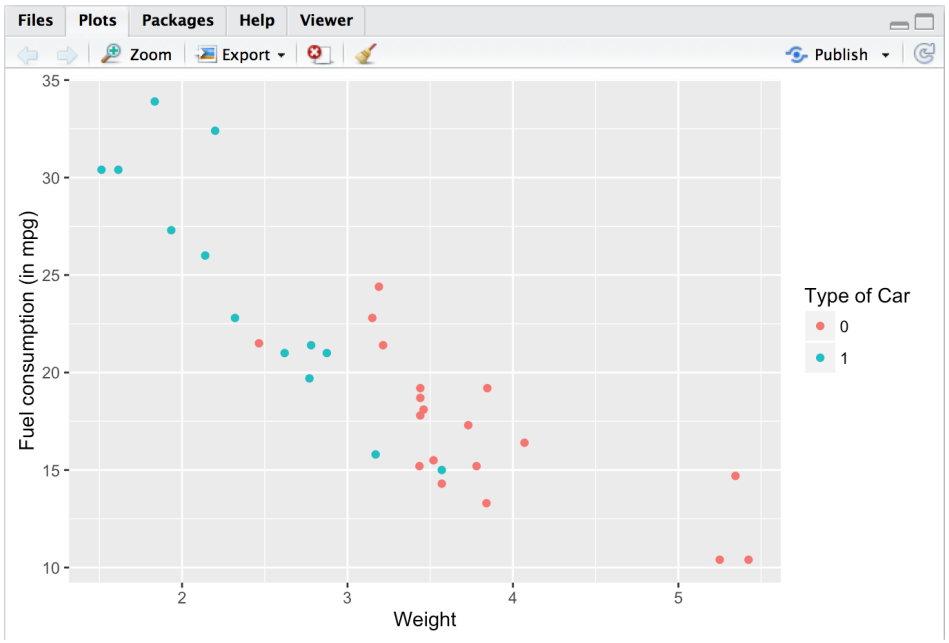
- You can access all previously run code in the history window.



RSTUDIO

The plot, packages and help window

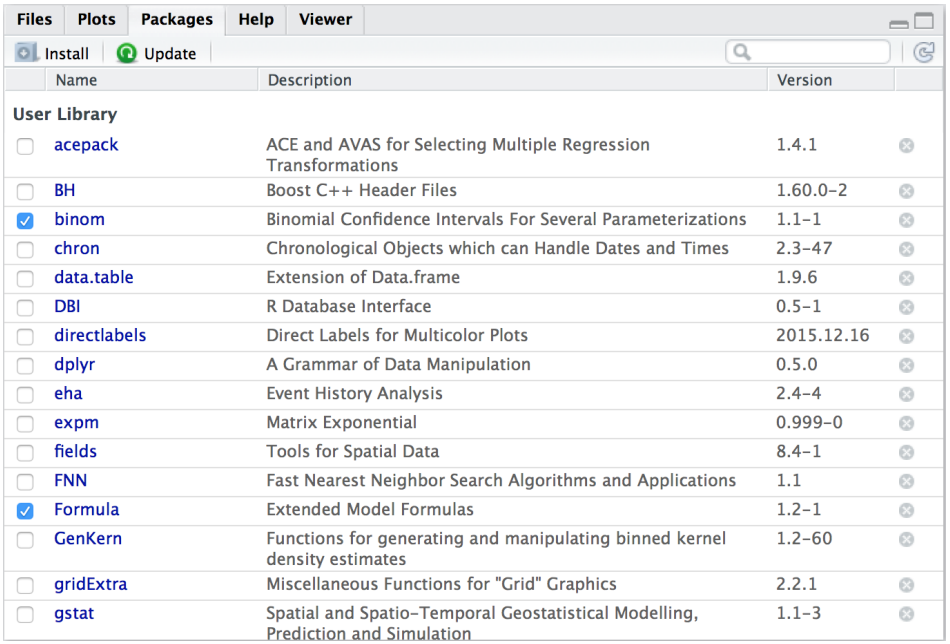
- Plots will display in the 'Plots' window.



RSTUDIO

The plot, packages and help window

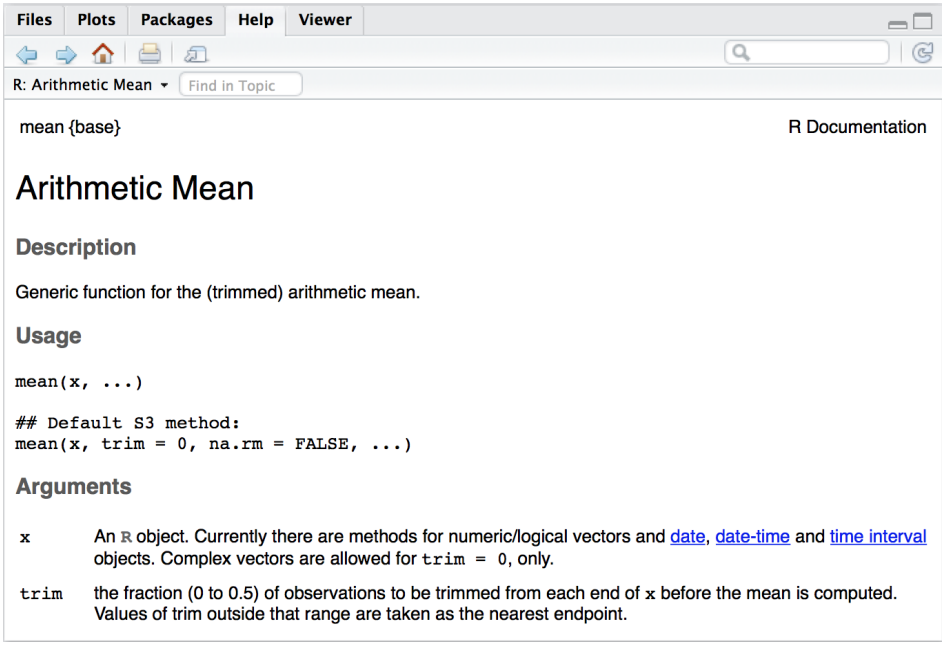
- Installed packages are listed in the 'Packages' window.



RSTUDIO

The plot, packages and help window

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



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?

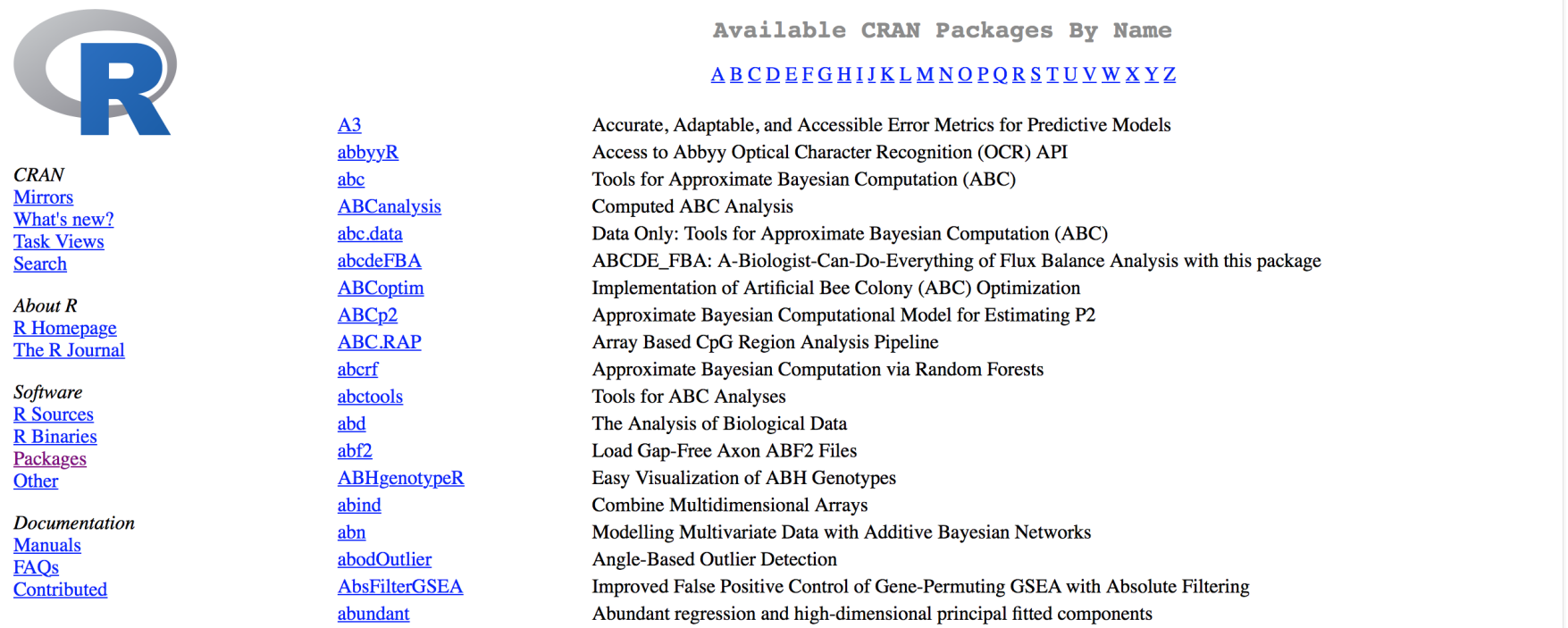
- The default installation of R has a comprehensive set of tools for statistical analyses.
- To meet the specific needs of data scientists, many other statistical tools are readily available in the form of packages.
- Packages are collections of functions and data.
- "During the last decade, the momentum 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

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

A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
ABCanalysis	Computed ABC Analysis
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2	Approximate Bayesian Computational Model for Estimating P2
ABC.RAP	Array Based CpG Region Analysis Pipeline
abcrf	Approximate Bayesian Computation via Random Forests
abctools	Tools for ABC Analyses
abd	The Analysis of Biological Data
abf2	Load Gap-Free Axon ABF2 Files
ABHgenotypeR	Easy Visualization of ABH Genotypes
abind	Combine Multidimensional Arrays
abn	Modelling Multivariate Data with Additive Bayesian Networks
abodOutlier	Angle-Based Outlier Detection
AbsFilterGSEA	Improved False Positive Control of Gene-Permuting GSEA with Absolute Filtering
abundant	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?