UNIVERSITÀ
DI TRENTO

Introduction to the R language

Biological Networks – Mod 1
Academic year 2021/2022
Mario Lauria

Credits: Arin Basu, Tyler Perrachione

TOPICS

What is R

How to obtain and install R

How to read and export data

How to do basic statistical analyses

Data analysis packages in R

WHAT IS R

- Software for Statistical Data Analysis
- Based on S
- Programming Environment
- Interpreted Language
- Data Storage, Analysis, Graphing
- Free and Open Source Software

STRENGTHS AND WEAKNESSES

- Strengths
  - Free and Open Source
  - Strong User Community
  - Highly extensible, flexible
  - Implementation of high end statistical methods
  - Flexible graphics and intelligent defaults
- Weakness
  - Steep learning curve
  - Slow for large datasets

OBTAINING R

- Current Version: R-4.1.2
- Comprehensive R Archive Network:

  http://cran.r-project.org
- Binary and source codes
- Windows executables
- Compiled RPMs for Linux
- Can be obtained on a CD

INSTALLING R

- Binary (Windows/Linux): One step process
  - exe, rpm (Red Hat/Mandrake), apt-get (Debian)
- Linux, from sources:

```
$ tar -zxvf "filename.tar.gz"
$ cd filename
$ ./configure
$ make
$ make check
$ make install
```

STARTING R

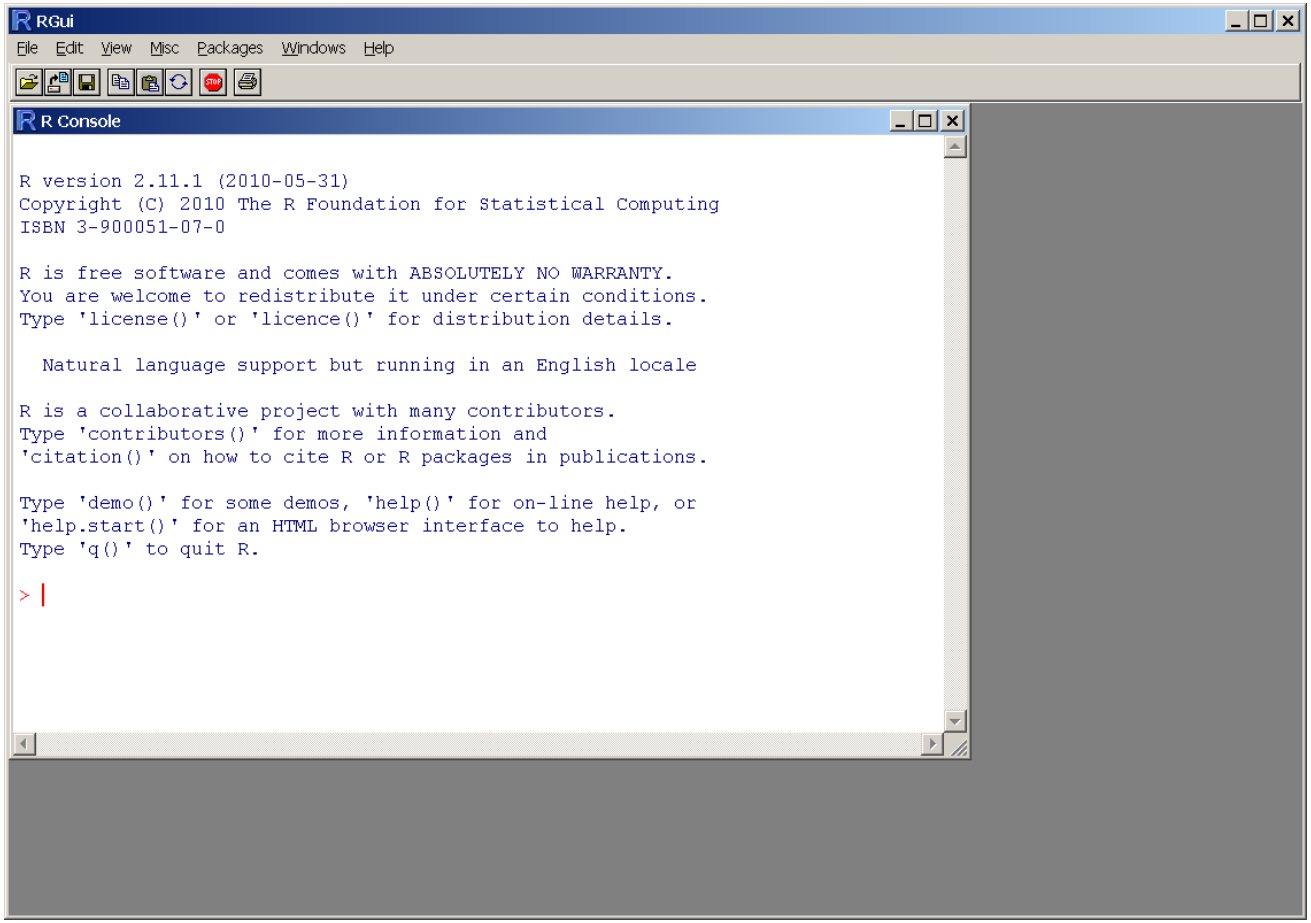
Windows, Double-click on Desktop Icon

```
$ R
```

Linux, type R at command prompt

RUNNING AND INTERACTING WITH R

Reference implementation of R includes a minimal programming environment – a console and no editor



RUNNING R

- An R program is created using an editor
  - Notepad++, WinEdt, Tinn-R: Windows
  - Xemacs, ESS (Emacs speaks Statistics)
- To run a program in the console:
  - source(“filename.R”)
  - Outputs can be diverted by using sink(“filename.Rout”)

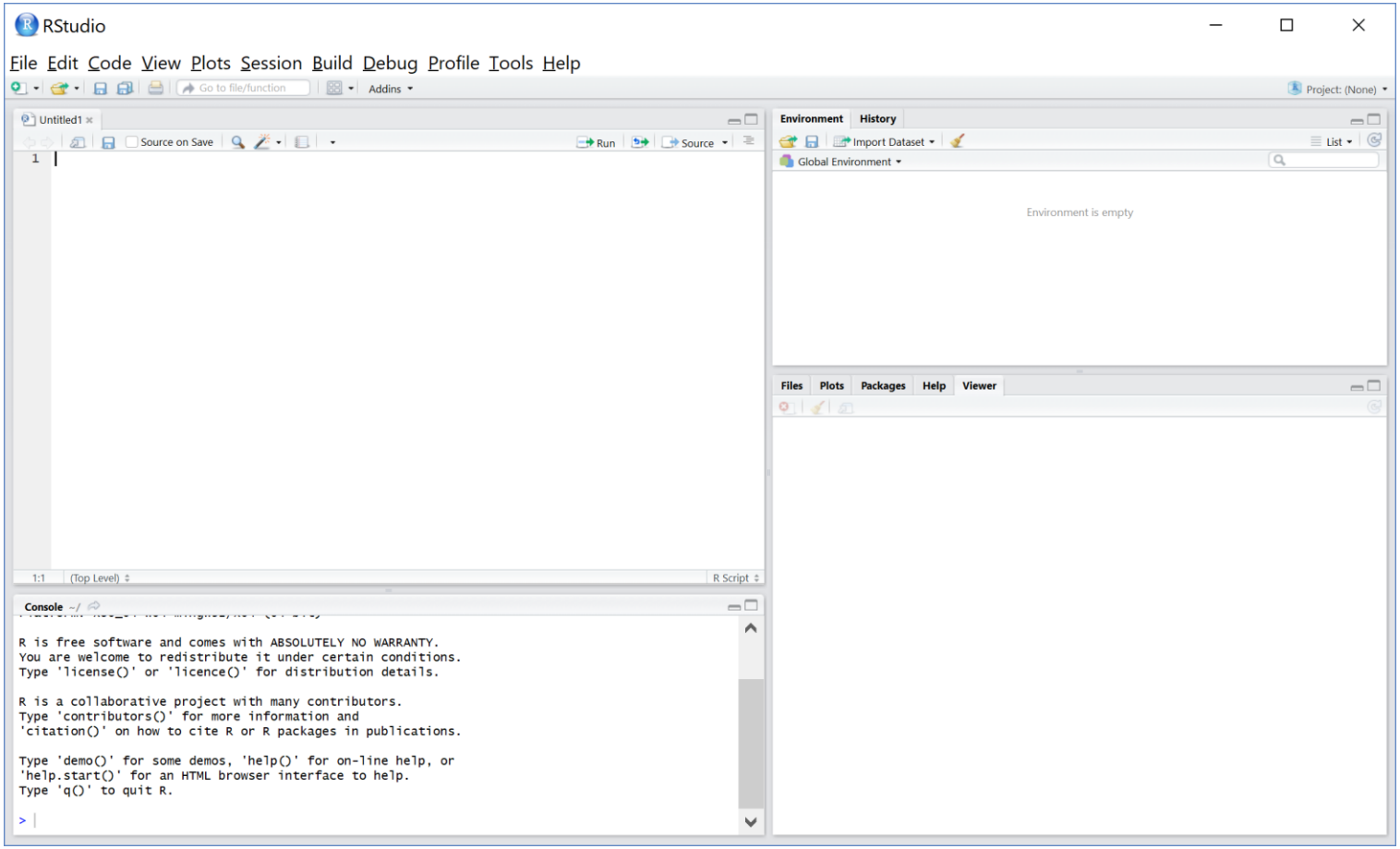
RUNNING R

- What makes R so useful is the ease of installation of additional libraries
  - Libraries are called ‘packages’ in R parlance
  - Packages add functionality to the basic R installation
  - Hundreds packages are available in open online repositories
- Several IDEs are available for R to facilitate code development
  - a very popular one is Rstudio - available at http://www.rstudio.com

```
#
# basic code example
#

# install the XLConnect package
install.packages("XLConnect")
library("XLConnect")

# see what libraries are installed
library()

# change working directory
setwd("C:/Users/Mario/Documents")
getwd()

# load internal toy data set
data(iris)

# let’s inspect the iris data set
?iris         # same as: help(iris)
class(iris)
dim(iris)
head(iris)
tail(iris)

q()
```

RUNNING R WITH A IDE

- IDE: Integrated Development Environment
- RStudio: an open source IDE for R



GETTING HELP IN R

- From built-in documentation:
  - ?WhatIWantToKnow
  - help("WhatIWantToKnow")
  - help.search("WhatIWantToKnow")
  - help.start()
  - getAnywhere("WhatIWantToKnow")
  - example("WhatIWantToKnow")
- Reference manual: "An Introduction to R"
  - https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- Many active mailing lists and blogs
  - typically found by using the right keywords in Google

DATA STRUCTURES

- Supports virtually any type of data
- Numbers, characters, logicals (TRUE/ FALSE)
- Arrays of virtually unlimited sizes and number of dimensions
- Simplest data structures: Vectors and Matrices
- Lists: Can Contain mixed type variables
- Data Frame: Rectangular Data Set

SYNOPSIS OF OPERATORS

Operator	Usually means	In Formula means
+ or -	add or subtract	add or remove terms
*	multiplication	main effect and interactions
/	division	main effect and nesting
:	sequence	interaction only
^	exponentiation	limiting interaction depths
%in%	no specific	nesting only

EXPRESSIONS IN R

Math:

```
> 1 + 1
[1] 2

> 1 + 1 * 7
[1] 8

> (1 + 1) * 7
[1] 14
```

Variables:

```
> x <- 1
> x
[1] 1
> y = 2
> y
[1] 2
> 3 -> z
> z
[1] 3
> (x + y) * z
[1] 9
```

EXPRESSIONS IN R

Arrays:

```
> x <- c(0,1,2,3,4)
> x
[1] 0 1 2 3 4

> y <- 1:5
> y
[1] 1 2 3 4 5

> z <- 1:50
> z
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
[16] 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
[31] 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
[46] 46 47 48 49 50
```

EXPRESSIONS IN R

Math on arrays:

```
> x <- c(0,1,2,3,4)
> y <- 1:5
> z <- 1:50
> x + y
[1] 1 3 5 7 9
> x * y
[1]  0  2  6 12 20
> x * z
 [1]   0   2   6  12  20   0   7  16  27  40   0
[12]  12  26  42  60   0  17  36  57  80   0  22
[23]  46  72 100   0  27  56  87 120   0  32  66
[34] 102 140   0  37  76 117 160   0  42  86 132
[45] 180   0  47  96 147 200
```

DATA STRUCTURES IN R

	Linear	Rectangular
All Same Type	VECTORS	MATRIX
Mixed	LIST	DATA FRAME

VECTORS AND MATRICES

```
# The c() function can be used to create vectors
> x <- c(0,1,2,3,4)
> x
[1] 0 1 2 3 4

# Use the vector() function to create empty vectors
> y <- vector("numeric", length = 10)
> y
[1] 0 0 0 0 0 0 0 0 0 0

> z <- 1:25
> z
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
[16] 16 17 18 19 20 21 22 23 24 25

# There is a matrix() function too
> m <- matrix(0, 2,3)
> m
     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    0    0
>
```

ACCESSING MATRIX AND VECTOR ELEMENTS

- Subscripts are an essential tool
  - x[1] identifies first element in vector x
  - x[-1] identifies all elements in vector x except the first
  - x[2:4] identifies elements from the 2nd to the 4th in vector x
  - x[c(2,4,5)] identifies specific elements in vector x
  - x[x > 3] identifies all the elements grater than 3 in vector x
  - y[2,3] identifies the element in row 2, column 3 in matrix y
  - y[1,] identifies first row in matrix y
  - y[,1] identifies first column in matrix y
  - y[1:5,] gives first 5 rows of data

- To inspect a data object:
  - edit(<mydataobject>)

THE INFAMOUS NA

- R uses NA to represent data that is "not available"
- The function **`is.na()`** tests for NA
- NA's are often an annoyance, see here an example of how to deal with them

```
> x <- c(1,2,NA,4,NA,5)
> x
[1]  1  2 NA  4 NA  5

> bad <- is.na(x)
> bad
[1] FALSE FALSE  TRUE FALSE  TRUE FALSE

> !bad
[1]  TRUE  TRUE FALSE  TRUE FALSE  TRUE

> x[!bad]
[1] 1 2 4 5
```

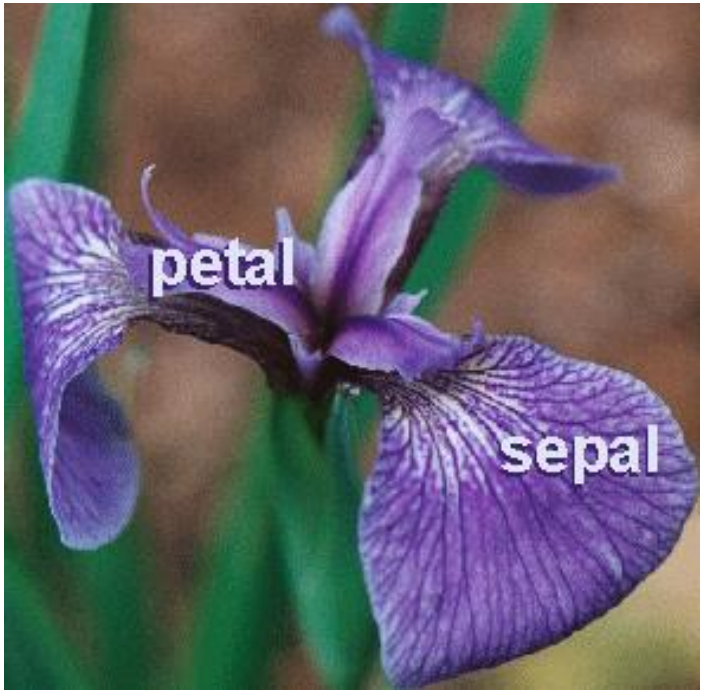
DATA FRAMES

- Data frames are R's fundamental data structures
- They are used to store data in a table format
- Unlike matrices, data frames can store different type of data
- Data frames are usually created by calling the `read.table()` or `read.csv()` functions
- They can be converted to a matrix by calling `data.matrix()`

BUILT-IN DATA SETS

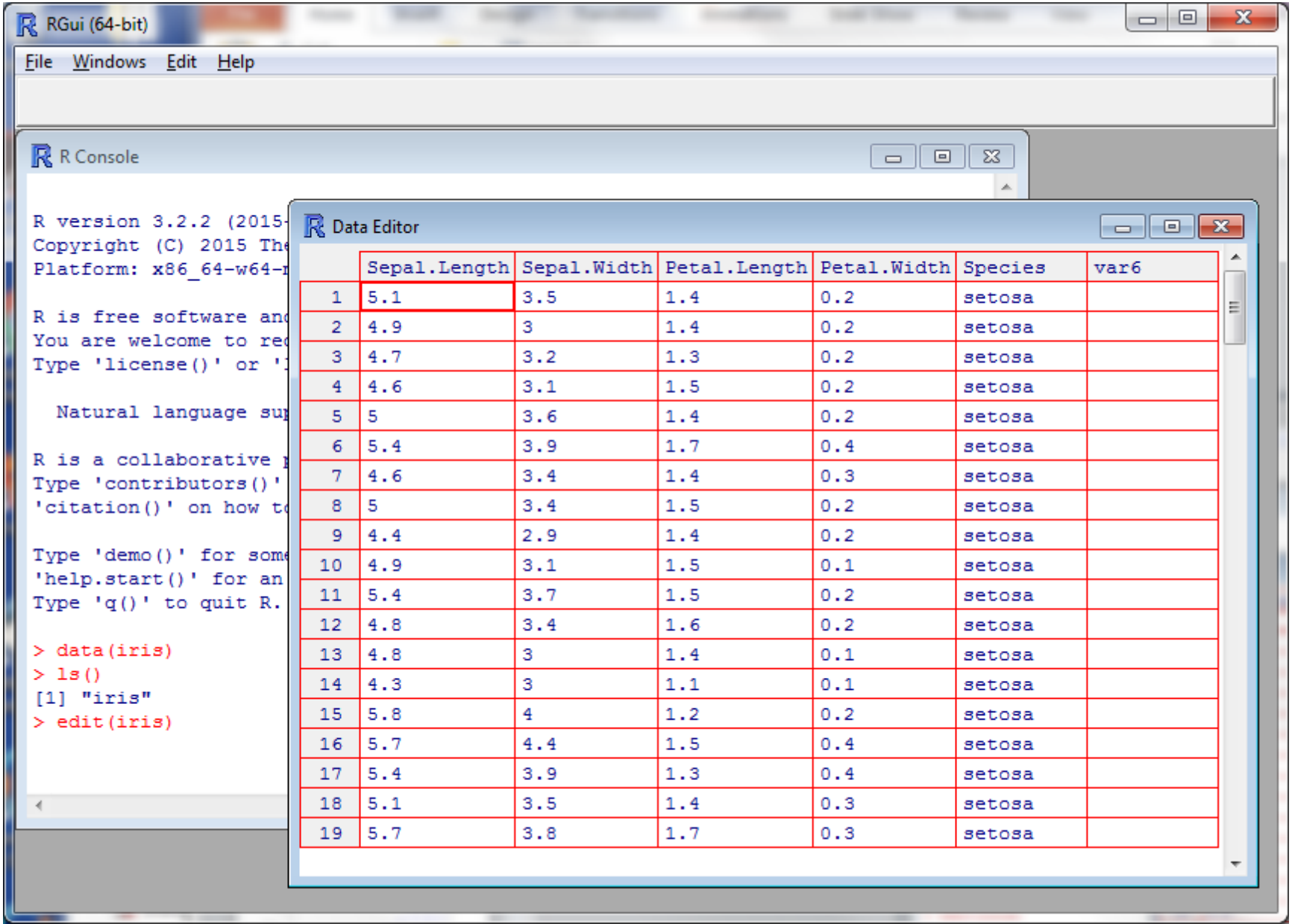
- You can practice the use of data frames by importing some of the built-in data sets available in R
- `data()` lists the built-in data sets
- `Iris` is a nice one

Iris virginica



Iris setosa

IRIS DATA SET



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	var6
1	5.1	3.5	1.4	0.2	setosa	
2	4.9	3	1.4	0.2	setosa	
3	4.7	3.2	1.3	0.2	setosa	
4	4.6	3.1	1.5	0.2	setosa	
5	5	3.6	1.4	0.2	setosa	
6	5.4	3.9	1.7	0.4	setosa	
7	4.6	3.4	1.4	0.3	setosa	
8	5	3.4	1.5	0.2	setosa	
9	4.4	2.9	1.4	0.2	setosa	
10	4.9	3.1	1.5	0.1	setosa	
11	5.4	3.7	1.5	0.2	setosa	
12	4.8	3.4	1.6	0.2	setosa	
13	4.8	3	1.4	0.1	setosa	
14	4.3	3	1.1	0.1	setosa	
15	5.8	4	1.2	0.2	setosa	
16	5.7	4.4	1.5	0.4	setosa	
17	5.4	3.9	1.3	0.4	setosa	
18	5.1	3.5	1.4	0.3	setosa	
19	5.7	3.8	1.7	0.3	setosa	

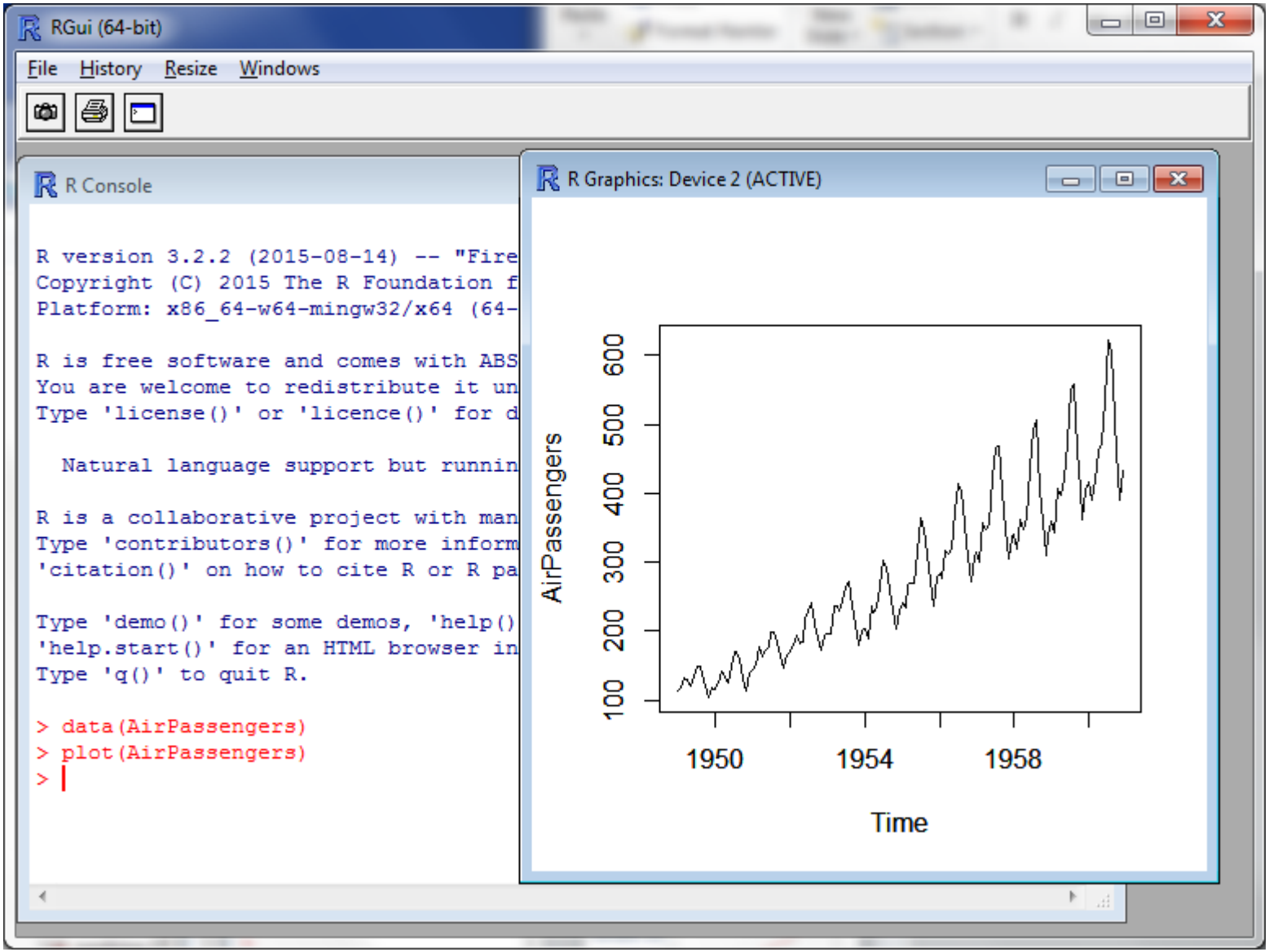
SUBSETTING DATA

- Using subset function
  - subset() will subset the dataframe
- Subscripting from data frames
  - myframe[,1] gives first column of myframe
- Using logical expressions
  - myframe[myframe[,1] < 5,] gets all rows that contain a value less than 5 in the first column
- Use the $ sign for lists and data frames
  - myframe$age gets age variable of myframe
  - attach(dataframe) -> make visible in the R environment the row/column names of the dataframe

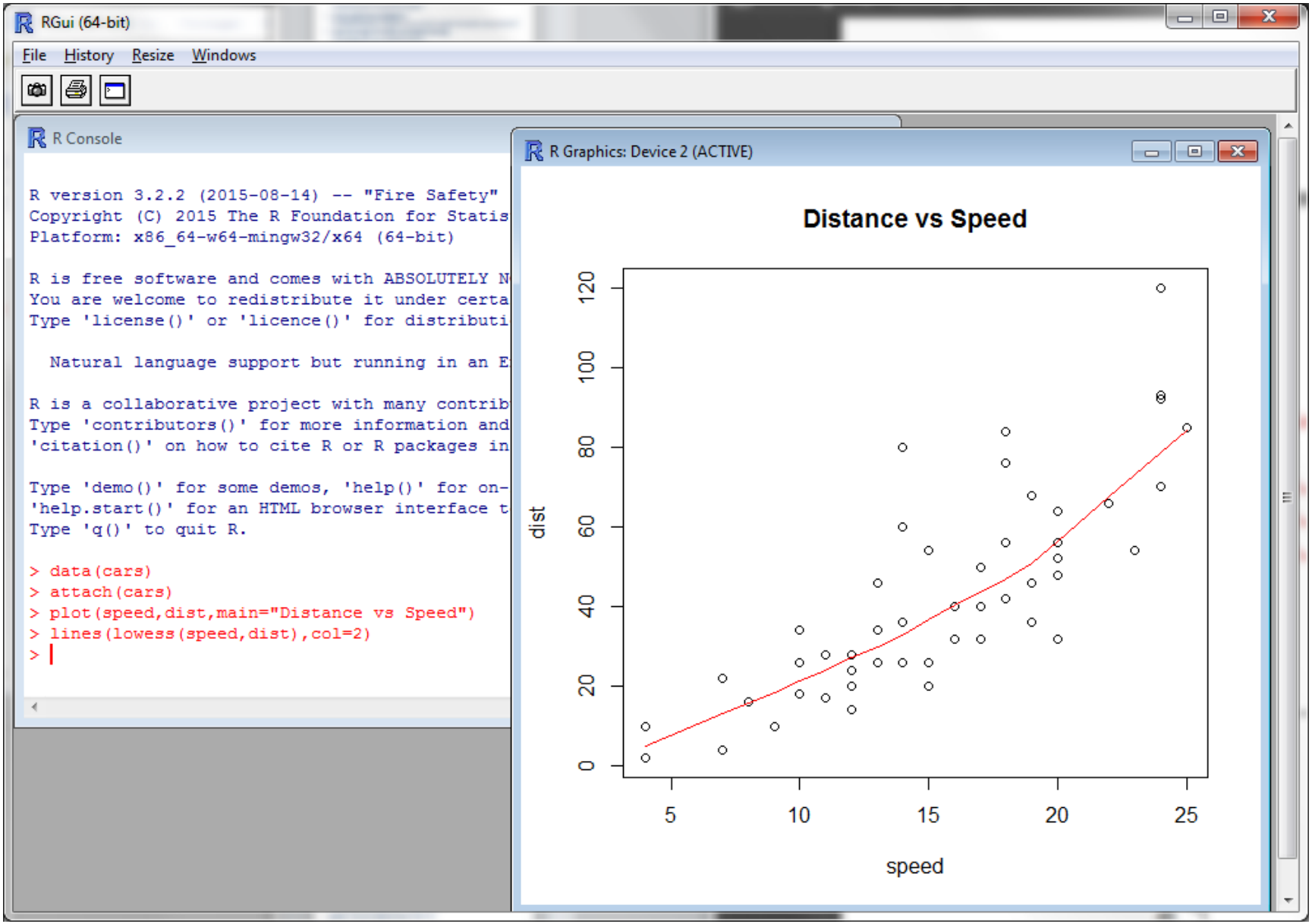
GRAPHICS

- Plot an object, like: plot(num.vec)
  - here plots against index numbers
- Plot sends to graphic devices
  - can specify which graphic device you want
  - postscript, gif, jpeg, etc…
- Two types of plotting commands
  - high level: graphs drawn with one call (R does all the work, good results most of the time)
  - Low Level: add additional information to existing graph (i.e. a graph title, a smoothing curve, etc.) to improve quality of the graph

HIGH LEVEL: GENERATED WITH PLOT()



LOW LEVEL: SCATTERGRAM WITH LOWESS



READING DATA INTO R

- R not well suited for data preprocessing
- Preprocess data elsewhere (SPSS, etc…)
- Easiest form of data to input: text file
- Spreadsheet like data:
  - Small/medium size: use read.table()
  - Large data: use scan()
- Use specialized packages to read from other systems:
  - Use the library “foreign” to read files from other statistical tools: library(“foreign”)
    - Can import from SAS, SPSS, Epi Info, can export to STATA format
  - Use the library “XLConnect” to read Excel files: library(“XLConnect”)

MORE BASIC R

```
#
# code example showing how to read/write data
#


# load internal toy data set
data(iris)

#create own copy
my.copy.of.iris <- iris

# how to extract columns from a data.frame
my.iris <-
data.frame(SepLen=iris$Sepal.Length,
           SepWid=iris$Sepal.Width,
           Species=iris$Species)

# quicker ways of extracting/removing cols
my.iris <- iris[,c(1,2,5)]
my.iris <- iris[,-c(3,4)]

# how to save/load a data set (R format)
save(my.copy.of.iris, file="mydata.RData")
mydata <- load(file="mydata.RData")
```

```
# how to save/load a data set (TAB sep. format)
write.table(iris, "myfile.dat", quote = FALSE,
              sep = "\t")
my.data <- read.table("myfile.dat",header = TRUE,
                        sep = "\t", row.names = 1 )


# how to save/load a data set (CSV format)
write.csv(iris, "myfile.dat")
my.data <- read.csv("myfile.dat", row.names = 1)

# sometimes data is available in Excel format
# many packages can be used for this purpose
# install the XLCoonect package
install.packages("XLConnect")
library("XLConnect")

wb1 <- loadWorkbook("TP-Function.xlsx")
f1 <- readWorksheet(wbf1, sheet="F1_48")


q()
```

R PROGRAMMING STYLE

- Functions & Operators typically work on entire vectors
  - You are encouraged to write your own functions and use them
- Codes separated by newlines
  - ";" not necessary, sometimes useful to improve code clarity
- You are highly encouraged to use a lot of explicative comments to make your code easier to maintain!
  - Anything following a "#" is a comment and ignored by R

STATISTICAL FUNCTIONS IN R

- Descriptive Statistics
- Statistical Modeling
  - Regressions: Linear and Logistic
  - Probit, Tobit Models
  - Time Series
- Multivariate Functions
- A large and diverse mix of built-in packages and contributed packages

DESCRIPTIVE STATISTICS

- R has functions for all common statistics
- summary() gives lowest, mean, median, first, third quartiles, highest for numeric variables
- stem() gives stem-leaf plots
- table() gives tabulation of categorical variables

STATISTICAL MODELING

- Over 400 functions
  - lm, glm, aov, ts
- Numerous libraries & packages
  - survival, coxph, tree (recursive trees), nls, …
- Distinction between factors and regressors
  - factors: categorical, regressors: continuous
  - you must specify factors unless they are obvious to R
  - dummy variables for factors created automatically
- Use of data.frame makes life easy

HOW TO MODEL

Specify your model like this:

- $y \sim x_i + c_i$, where
- y = outcome variable, x_i = main explanatory variables, c_i = covariates, + = add terms
- Operators have special meanings

+ = add terms, : = interactions, / = nesting, so on…

Modeling -- object oriented

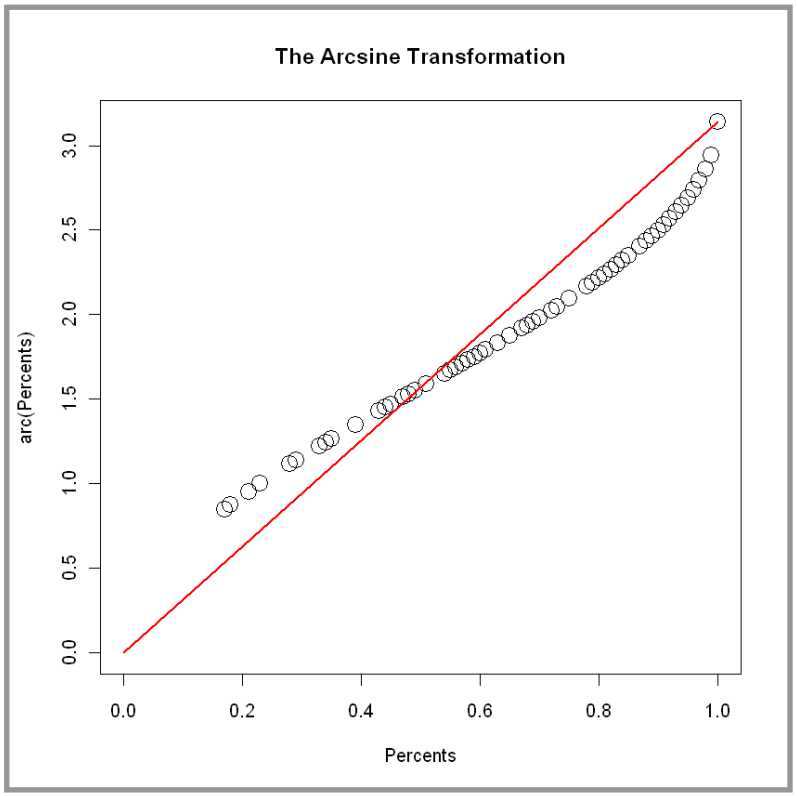
- each modeling procedure produces objects
- classes and functions for each object

MODELING EXAMPLE: REGRESSION

```
carReg <- lm(speed~dist, data=cars)
# carReg becomes an object
# to get summary of this regression, we type
summary(carReg)

# to get only coefficients, we type
coef(carReg), or carReg$coef
# don't want intercept? add 0, so
carReg <- lm(speed~0+dist, data=cars)
```

Functions:

```
> arc <- function(x) 2*asin(sqrt(x))
> arc(0.5)
[1] 1.570796
> x <- c(0,1,2,3,4)
> x <- x / 10
> arc(x)
[1] 0.0000000 0.6435011 0.9272952
[4] 1.1592795 1.3694384
```

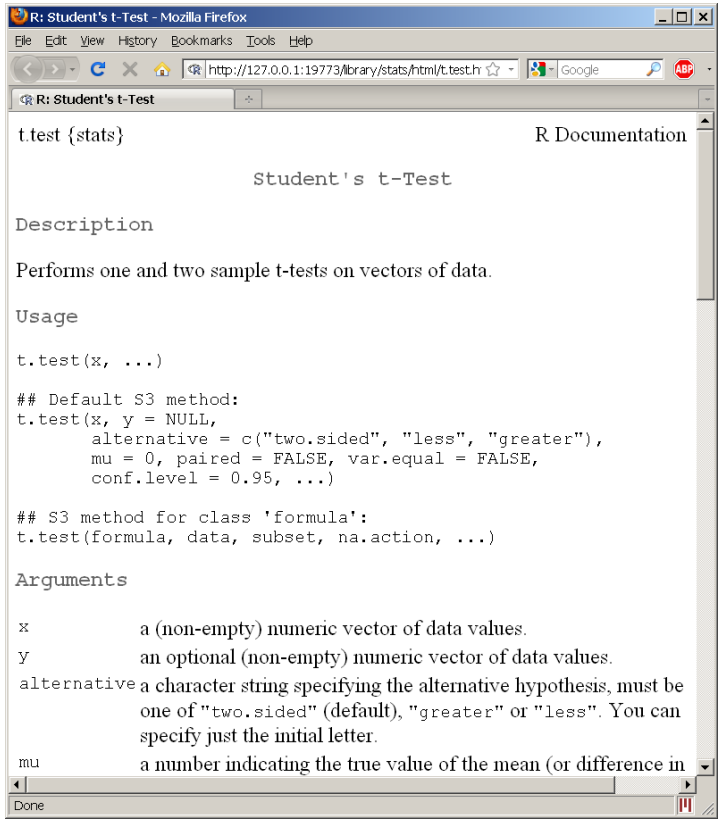


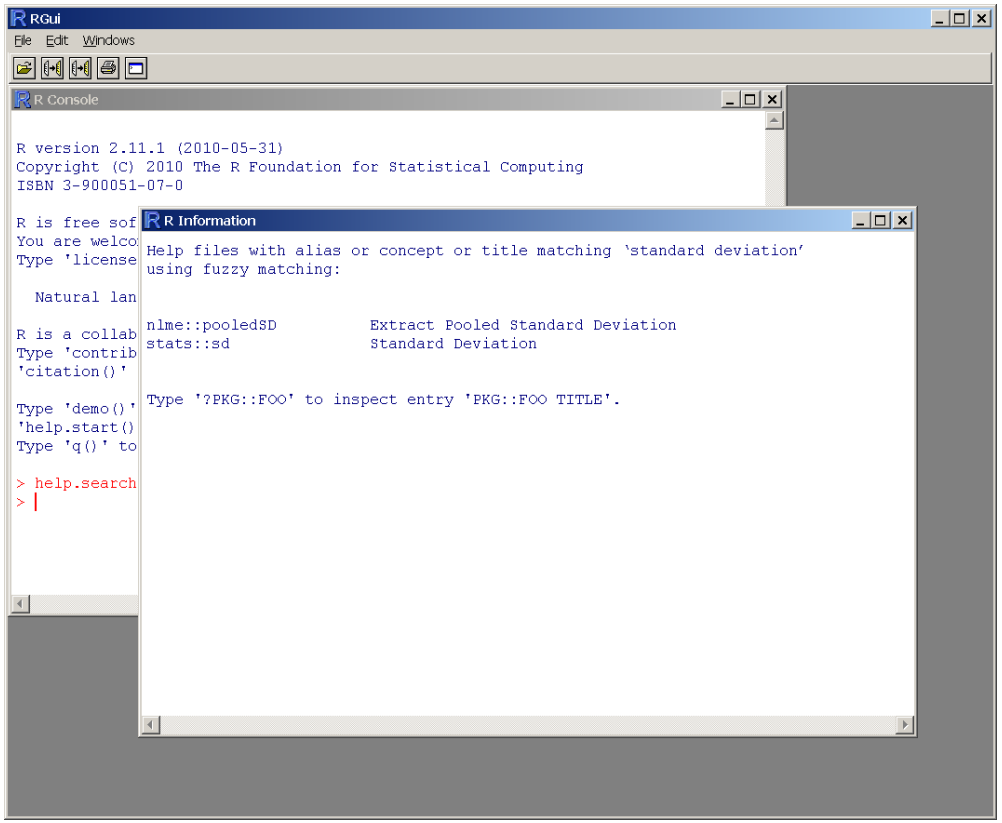

```
> plot(arc(Percents)~Percents,
+ pch=21,cex=2,xlim=c(0,1),ylim=c(0,pi),
+ main="The Arcsine Transformation")
> lines(c(0,1),c(0,pi),col="red",lwd=2)
```

INSTALLING, RUNNING, AND INTERACTING WITH R

Getting help:

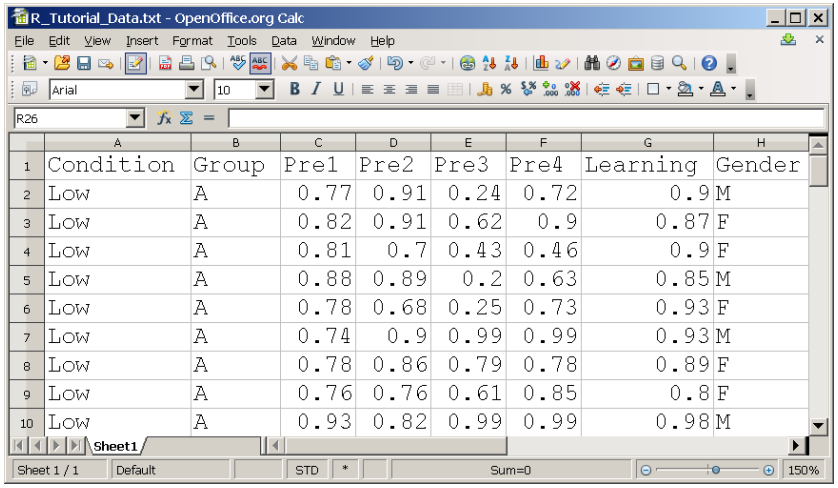
```
> help(t.test)
> help.search("standard deviation")
```



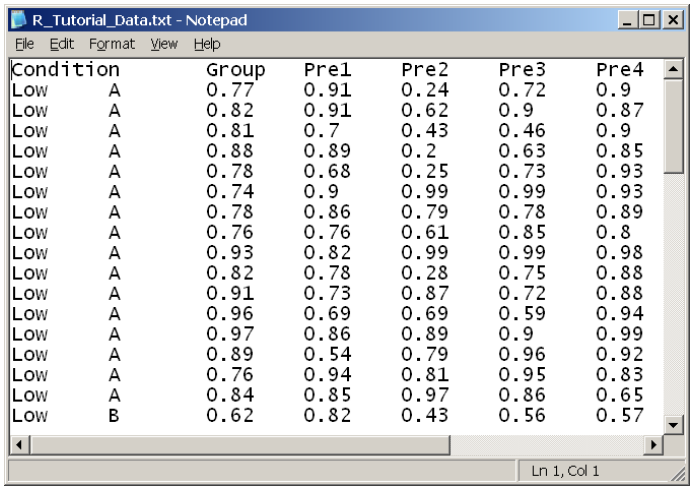


Reading data from files:

```
> myData <- read.table("R_Tutorial_Data.txt",
+ header=TRUE, sep="\t")
> myData
   Condition Group Pre1 Pre2 Pre3 Pre4 Learning
1        Low     A 0.77 0.91 0.24 0.72     0.90
2        Low     A 0.82 0.91 0.62 0.90     0.87
3        Low     A 0.81 0.70 0.43 0.46     0.90
. . .
61      High     B 0.44 0.41 0.84 0.82     0.29
62      High     B 0.48 0.56 0.83 0.85     0.48
63      High     B 0.61 0.82 0.88 0.95     0.28
```



	A	B	C	D	E	F	G	H
1	Condition	Group	Pre1	Pre2	Pre3	Pre4	Learning	Gender
2	Low	A	0.77	0.91	0.24	0.72	0.9	M
3	Low	A	0.82	0.91	0.62	0.9	0.87	F
4	Low	A	0.81	0.7	0.43	0.46	0.9	F
5	Low	A	0.88	0.89	0.2	0.63	0.85	M
6	Low	A	0.78	0.68	0.25	0.73	0.93	F
7	Low	A	0.74	0.9	0.99	0.99	0.93	M
8	Low	A	0.78	0.86	0.79	0.78	0.89	F
9	Low	A	0.76	0.76	0.61	0.85	0.8	F
10	Low	A	0.93	0.82	0.99	0.99	0.98	M

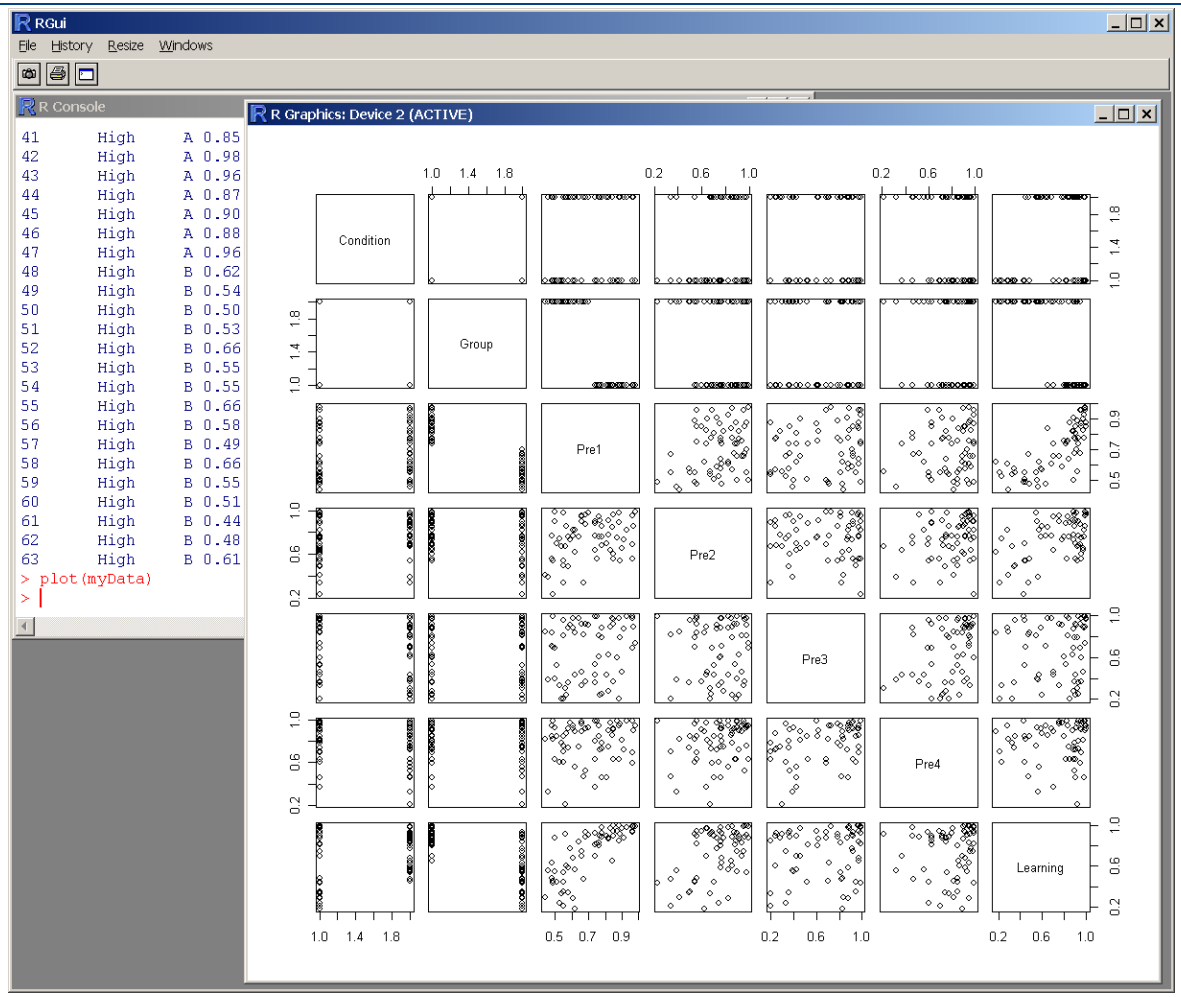


Condition	Group	Pre1	Pre2	Pre3	Pre4
Low	A	0.77	0.91	0.24	0.72	0.9
Low	A	0.82	0.91	0.62	0.9	0.87
Low	A	0.81	0.7	0.43	0.46	0.9
Low	A	0.88	0.89	0.2	0.63	0.85
Low	A	0.78	0.68	0.25	0.73	0.93
Low	A	0.74	0.9	0.99	0.99	0.93
Low	A	0.78	0.86	0.79	0.78	0.89
Low	A	0.76	0.76	0.61	0.85	0.8
Low	A	0.93	0.82	0.99	0.99	0.98
Low	A	0.82	0.78	0.28	0.75	0.88
Low	A	0.91	0.73	0.87	0.72	0.88
Low	A	0.96	0.69	0.69	0.59	0.94
Low	A	0.97	0.86	0.89	0.9	0.99
Low	A	0.89	0.54	0.79	0.96	0.92
Low	A	0.76	0.94	0.81	0.95	0.83
Low	A	0.84	0.85	0.97	0.86	0.65
Low	B	0.62	0.82	0.43	0.56	0.57

INSTALLING, RUNNING, AND INTERACTING WITH R

Examining datasets:

```
> plot(myData)
```

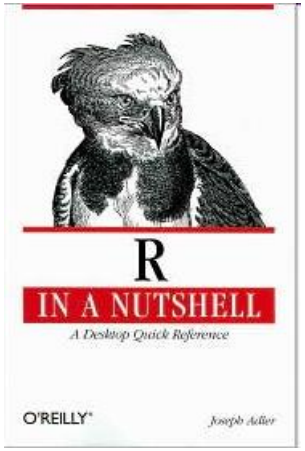


HOW TO FIND R HELP AND RESOURCES ON THE INTERNET

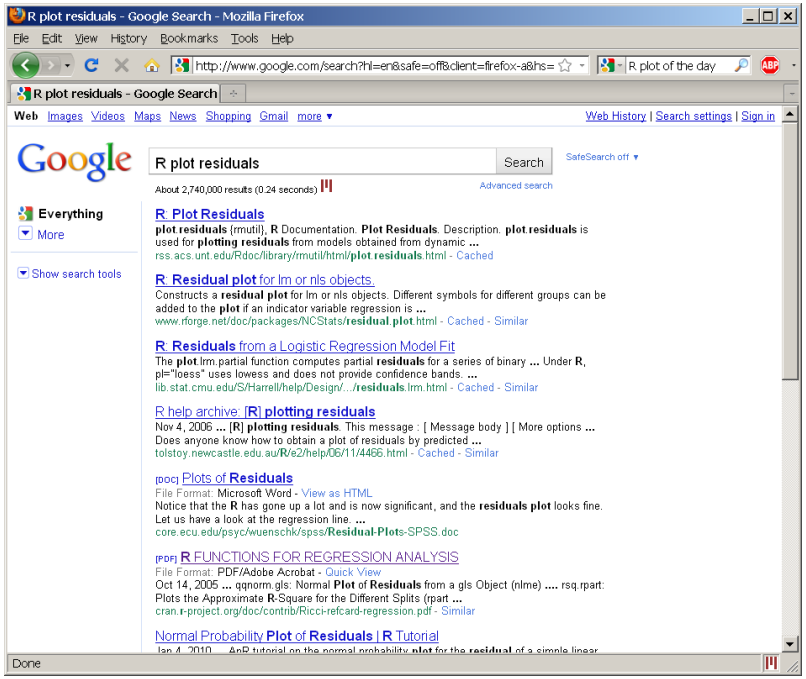
R wiki:
http://rwiki.sciviews.org/doku.php

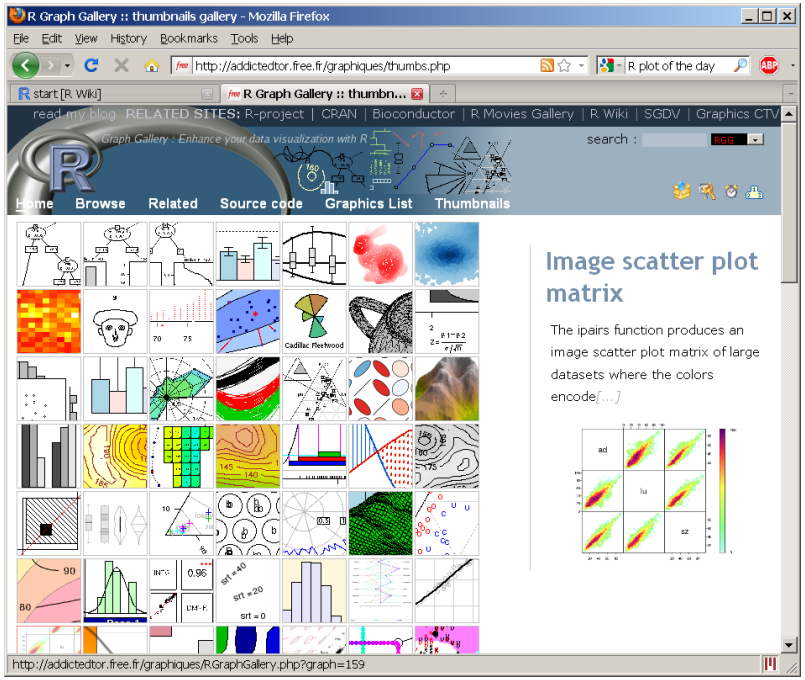
R graph gallery:
http://addictedtor.free.fr/graphiques/thumbs.php

Kickstarting R:
http://cran.r-project.org/doc/contrib/Lemon-kickstart/



ISBN: 9780596801700





FOR MORE RESOURCES, CHECK OUT…

- R home page
  - http://www.r-project.org
- R discussion group
  - http://www.stat.math.ethz.ch/mailman/listinfo/r-help
- Cheat sheets:
  - https://www.rstudio.com/resources/cheatsheets/
- Search Google for R and Statistics
  - Typically the most useful hits are those from the www.stackoverflow.com blog