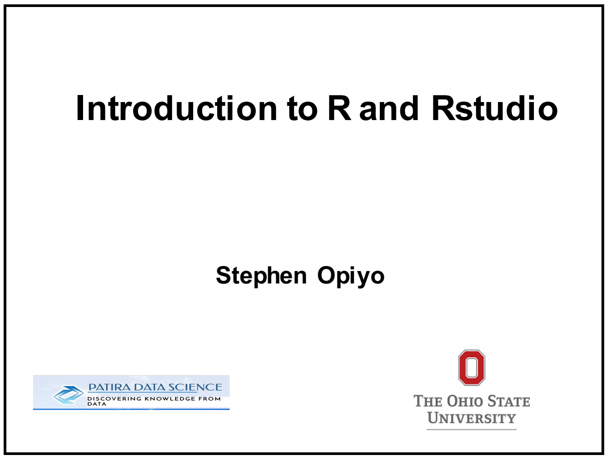
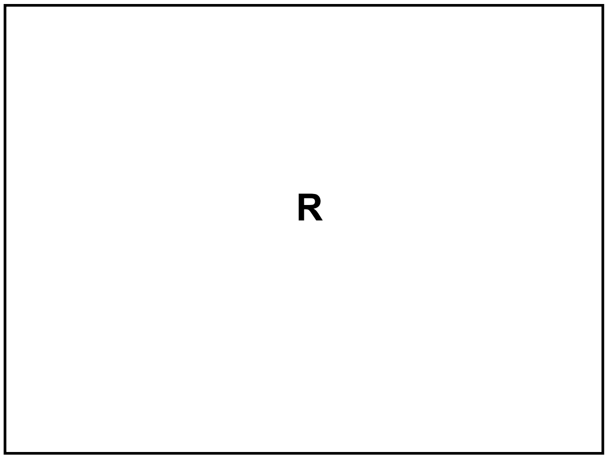


Introduction to R and Rstudio

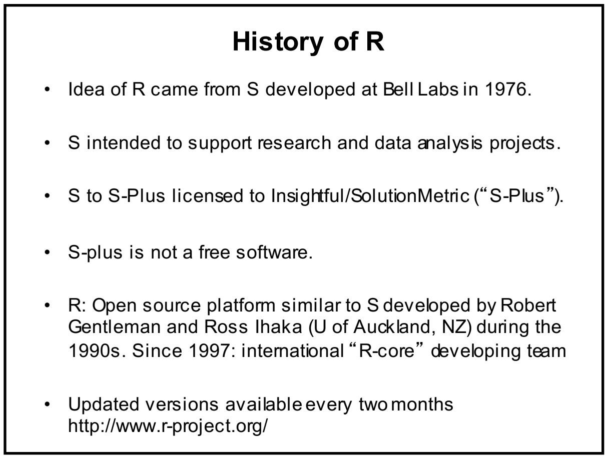
Stephen Opiyo

PATIRA DATA SCIENCE
DISCOVERING KNOWLEDGE FROM DATA

THE OHIO STATE UNIVERSITY



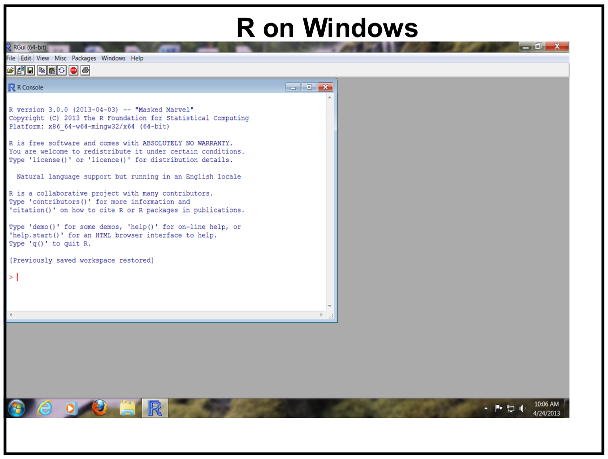
R



History of R

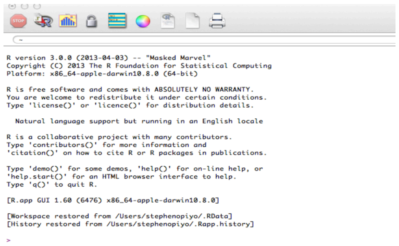
- Idea of R came from S developed at Bell Labs in 1976.
- S intended to support research and data analysis projects.
- S to S-Plus licensed to Insightful/SolutionMetric (“S-Plus”).
- S-plus is not a free software.
- R: Open source platform similar to S developed by Robert Gentleman and Ross Ihaka (U of Auckland, NZ) during the 1990s. Since 1997: international “R-core” developing team
- Updated versions available every two months http://www.r-project.org/

R on Windows



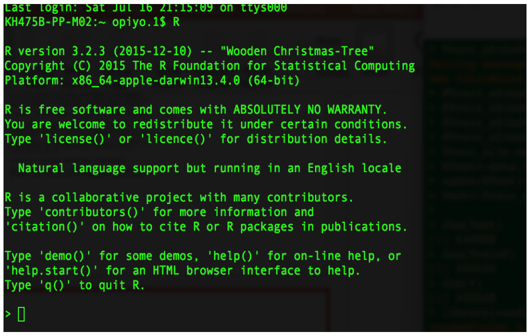
```
R version 3.0.0 (2013-04-03) -- "Masked Marvel"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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.

[Previously saved workspace restored]

> |
```

R on MACs

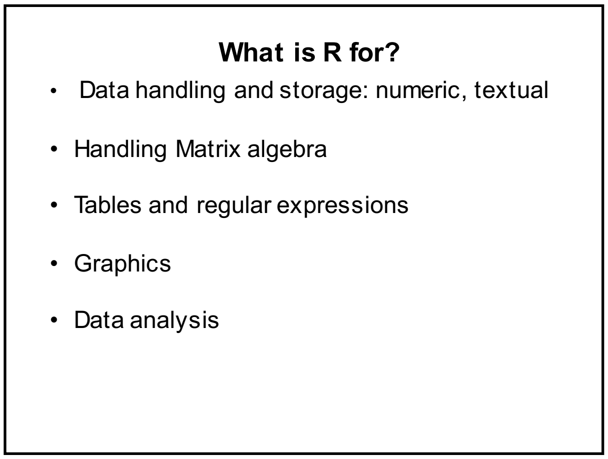


```
R version 3.0.0 (2013-04-03) -- "Masked Marvel"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

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.

[R.app GUI 1.60 (6476) x86_64-apple-darwin10.8.0]

[Workspace restored from /Users/stephenopiyo/.RData]
[History restored from /Users/stephenopiyo/.Rapp.history]

>
```

R on Linux

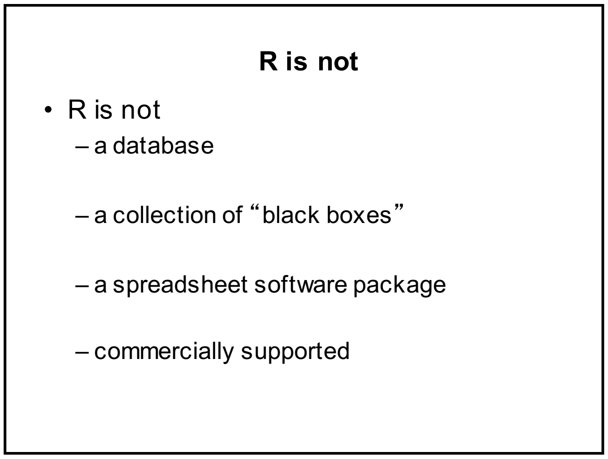


```
Last login: Sat Jul 16 21:15:09 on ttys000
KH475B-PP-M02:~ opiyo.1$ R

R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

> ▯
```



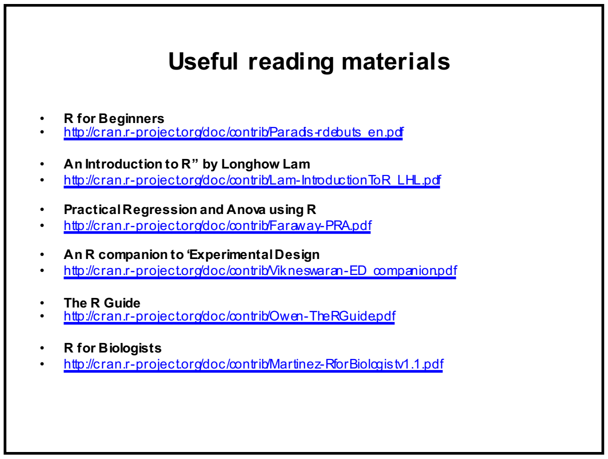
What is R for?

- Data handling and storage: numeric, textual
- Handling Matrix algebra
- Tables and regular expressions
- Graphics
- Data analysis



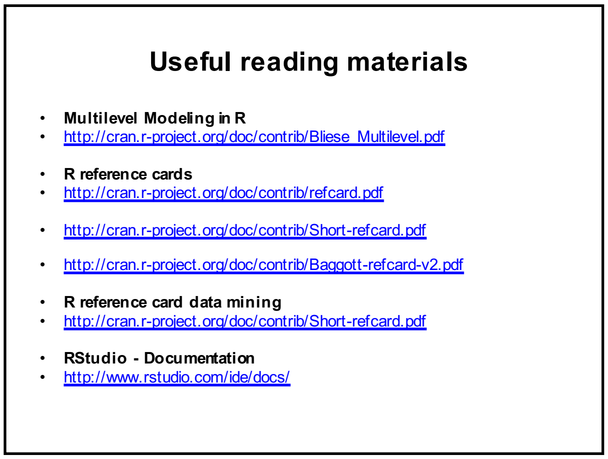
R is not

- R is not
  - a database
  - a collection of "black boxes"
  - a spreadsheet software package
  - commercially supported



Useful reading materials

- **R for Beginners**
- http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf
- **An Introduction to R" by Longhow Lam**
- http://cran.r-project.org/doc/contrib/Lam-IntroductionToR_LHL.pdf
- **Practical Regression and Anova using R**
- http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf
- **An R companion to 'Experimental Design**
- http://cran.r-project.org/doc/contrib/Vikneswaran-ED_companion.pdf
- **The R Guide**
- http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf
- **R for Biologists**
- http://cran.r-project.org/doc/contrib/Martinez-RforBiologistv1.1.pdf



Useful reading materials

- **Multilevel Modeling in R**
- http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf

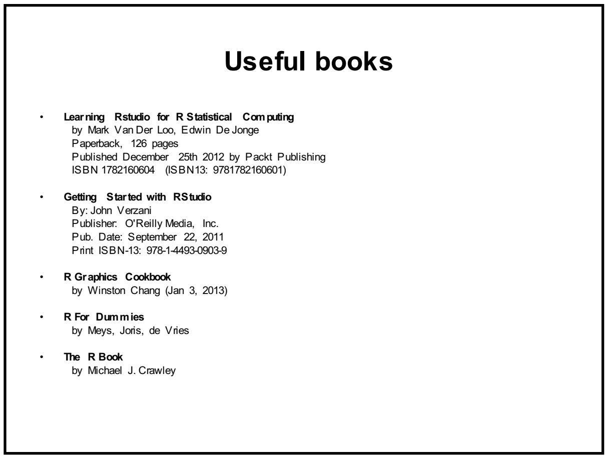
- **R reference cards**
- http://cran.r-project.org/doc/contrib/refcard.pdf

- http://cran.r-project.org/doc/contrib/Short-refcard.pdf

- http://cran.r-project.org/doc/contrib/Baggott-refcard-v2.pdf

- **R reference card data mining**
- http://cran.r-project.org/doc/contrib/Short-refcard.pdf

- **RStudio - Documentation**
- http://www.rstudio.com/ide/docs/



Useful books

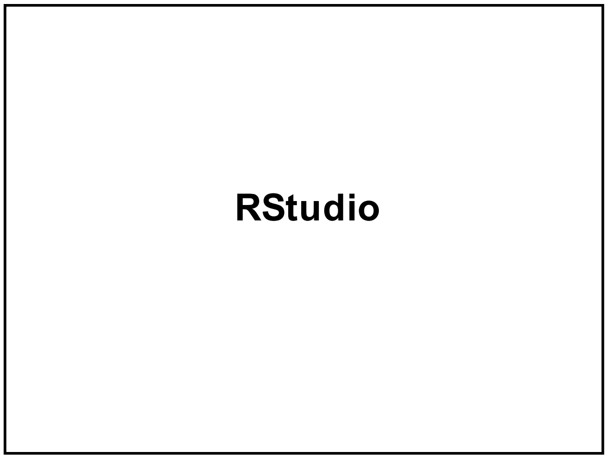
- **Learning Rstudio for R Statistical Computing**
  by Mark Van Der Loo, Edwin De Jonge
  Paperback, 126 pages
  Published December 25th 2012 by Packt Publishing
  ISBN 1782160604 (ISBN13: 9781782160601)

- **Getting Started with RStudio**
  By: John Verzani
  Publisher: O'Reilly Media, Inc.
  Pub. Date: September 22, 2011
  Print ISBN-13: 978-1-4493-0903-9

- **R Graphics Cookbook**
  by Winston Chang (Jan 3, 2013)

- **R For Dummies**
  by Meys, Joris, de Vries

- **The R Book**
  by Michael J. Crawley

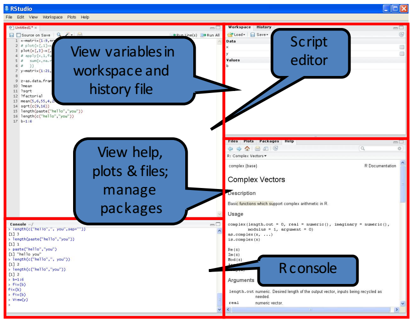


RStudio

RStudio

- RStudio is a free open source integrated development environment for R (http://www.rstudio.com/ide/)
- Free and commercial versions
- RStudio is available in two editions:
  - RStudio Desktop:Run locally as a regular desktop.
  - RStudio Server runs on remote Linux server.
- We will use RStudio in this workshop

Using RStudio


RStudio
File Edit View Workspace Plots Help
Untitled1*
Source on Save
Run Line(s) Run All
Workspace History
Load Save
Data
x
y
Values
b
Files Plots Packages Help
R: Complex Vectors
Console ~/
View variables in workspace and history file
Script editor
View help, plots & files; manage packages
R console

```
x=matrix(1:9,nr
# plot(x[,1]~x
plot(x[,3]~x[,
# apply(x,1,fu
#   sum(x,na.r
#   })
y=matrix(1:21,

z=as.data.fram
?mean
?sqrt
?factorial
mean(5,6,55,4,2
sqrt(c(9,16))
length(paste("hello","you"))
length(c("hello","you"))
b=1:6
```

```
> length(c("hello",", you",sep=""))
[1] 3
> length(paste("hello","you"))
[1] 1
> paste("hello","you")
[1] "hello you"
> length(c("hello",", you"))
[1] 2
> length(c("hello","you"))
[1] 2
> b=1:6
> fix(b)
fix(b)
> fix(b)
> View(y)
>
```

complex {base} R Documentation

Complex Vectors

Description

Basic functions which support complex arithmetic in R.

Usage

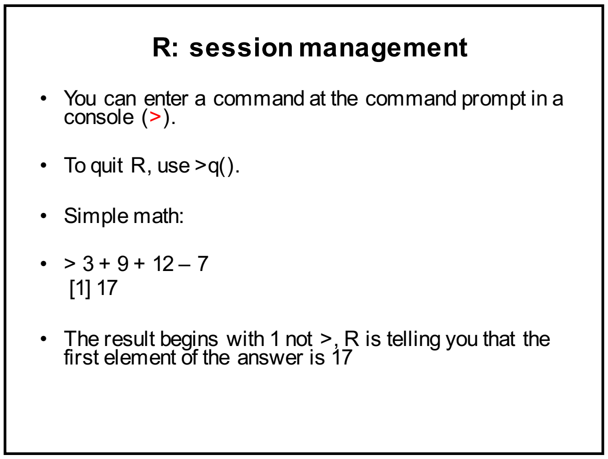
```
complex(length.out = 0, real = numeric(), imaginary = numeric(),
        modulus = 1, argument = 0)
as.complex(x, ...)
is.complex(x)

Re(z)
Im(z)
Mod(z)
```

Arguments

length.out numeric. Desired length of the output vector, inputs being recycled as needed.

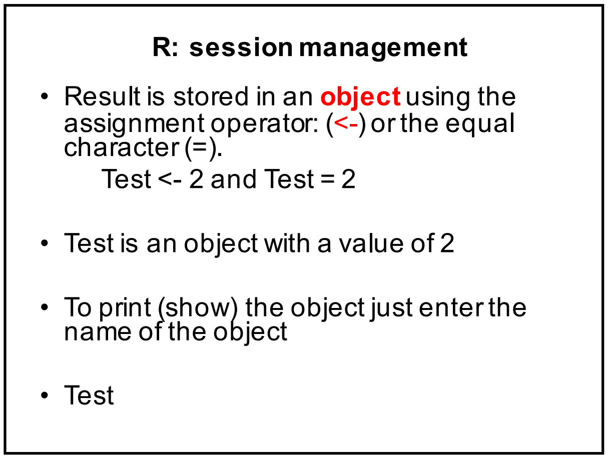
real numeric vector.

R: Session management



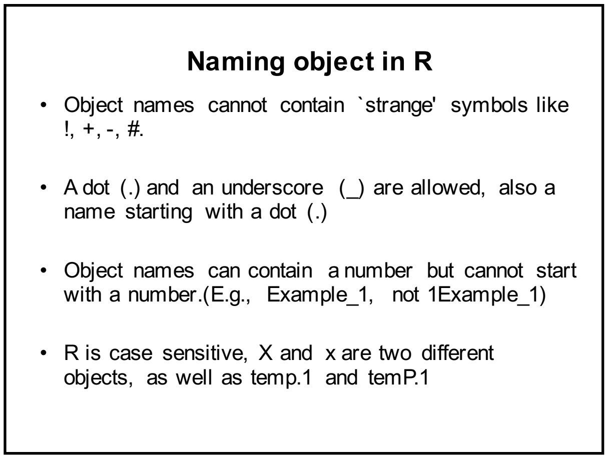
R: session management

- You can enter a command at the command prompt in a console (>).
- To quit R, use >q().
- Simple math:
- > 3 + 9 + 12 – 7
  [1] 17
- The result begins with 1 not >, R is telling you that the first element of the answer is 17



R: session management

- Result is stored in an **object** using the assignment operator: (<-) or the equal character (=).
  Test <- 2 and Test = 2
- Test is an object with a value of 2
- To print (show) the object just enter the name of the object
- Test

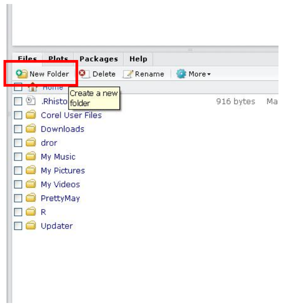


Naming object in R

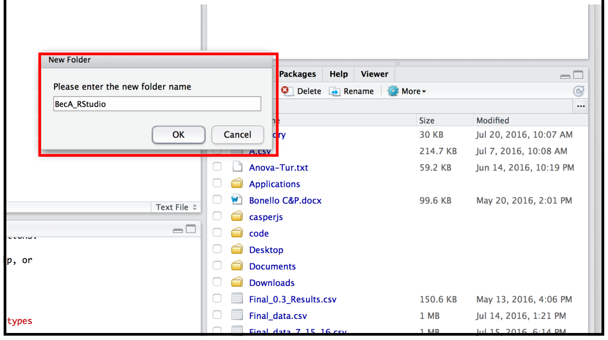
- Object names cannot contain `strange' symbols like !, +, -, #.
- A dot (.) and an underscore (_) are allowed, also a name starting with a dot (.)
- Object names can contain a number but cannot start with a number.(E.g., Example_1, not 1Example_1)
- R is case sensitive, X and x are two different objects, as well as temp.1 and temP.1

Setting up a working directory

Open RStudio and set up a working directory



Open RStudio and set up a working directory



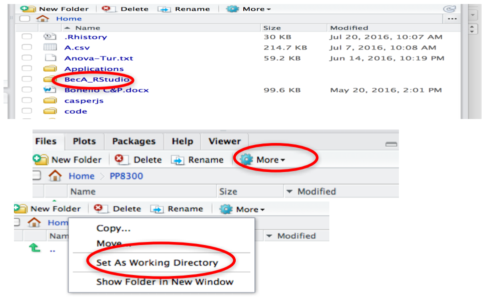
Please enter the new folder name

BecA_RStudio

```
…tions:
p, or
types
```

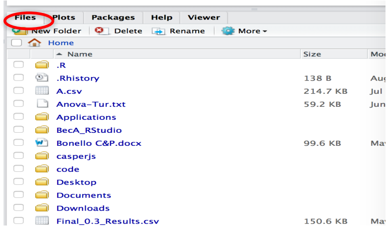
…e	Size	Modified
…ry	30 KB	Jul 20, 2016, 10:07 AM
A.csv	214.7 KB	Jul 7, 2016, 10:08 AM
Anova–Tur.txt	59.2 KB	Jun 14, 2016, 10:19 PM
Applications		
Bonello C&P.docx	99.6 KB	May 20, 2016, 2:01 PM
casperjs		
code		
Desktop		
Documents		
Downloads		
Final_0.3_Results.csv	150.6 KB	May 13, 2016, 4:06 PM
Final_data.csv	1 MB	Jul 14, 2016, 1:21 PM
Final_data_7_15_16.csv	1 MB	Jul 15, 2016, 6:14 PM

Open RStudio and set up a working directory



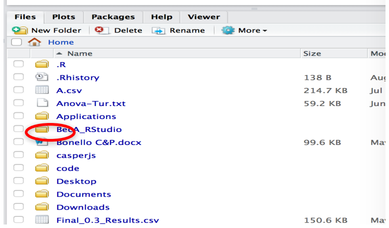
	Name	Size	Modified
☐	.Rhistory	30 KB	Jul 20, 2016, 10:07 AM
☐	A.csv	214.7 KB	Jul 7, 2016, 10:08 AM
☐	Anova–Tur.txt	59.2 KB	Jun 14, 2016, 10:19 PM
☐	Applications		
☐	BecA_RStudio		
☐	Bonello C&P.docx	99.6 KB	May 20, 2016, 2:01 PM
☐	casperjs		
☐	code		

Open RStudio and set BecA_Rstudio as a working directory



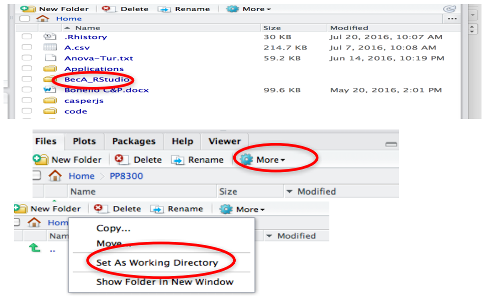
	Name	Size	Mo
☐	.R		
☐	.Rhistory	138 B	Au
☐	A.csv	214.7 KB	Jul
☐	Anova–Tur.txt	59.2 KB	Jun
☐	Applications		
☐	BecA_RStudio		
☐	Bonello C&P.docx	99.6 KB	Ma
☐	casperjs		
☐	code		
☐	Desktop		
☐	Documents		
☐	Downloads		
☐	Final_0.3_Results.csv	150.6 KB	Ma

Open RStudio and set BecA_Rstudio as a working directory



	Name	Size	Mo
☐	.R		
☐	.Rhistory	138 B	Au
☐	A.csv	214.7 KB	Jul
☐	Anova–Tur.txt	59.2 KB	Jun
☐	Applications		
☐	BecA_RStudio		
☐	Bonello C&P.docx	99.6 KB	Ma
☐	casperjs		
☐	code		
☐	Desktop		
☐	Documents		
☐	Downloads		
☐	Final_0.3_Results.csv	150.6 KB	Ma

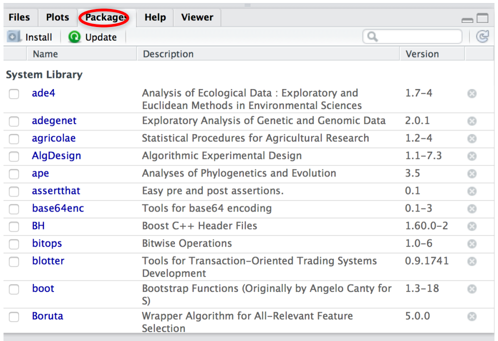
Open RStudio and set up a working directory



Name	Size	Modified
.Rhistory	30 KB	Jul 20, 2016, 10:07 AM
A.csv	214.7 KB	Jul 7, 2016, 10:08 AM
Anova-Tur.txt	59.2 KB	Jun 14, 2016, 10:19 PM
Applications		
BecA_RStudio		
Bonello C&P.docx	99.6 KB	May 20, 2016, 2:01 PM
casperjs		
code		

Installing R package (library) in RStudio

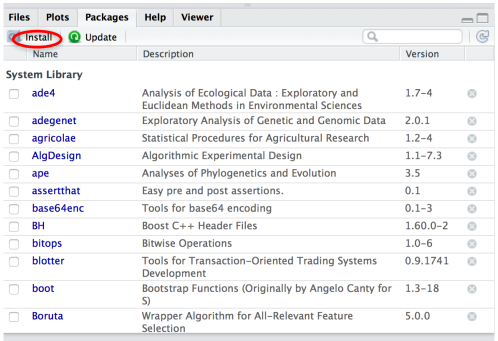
Installing R package in RStudio



System Library

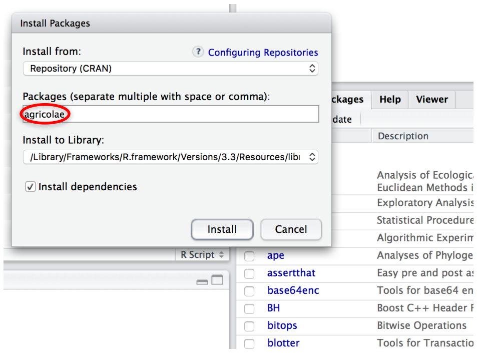
Name	Description	Version
ade4	Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences	1.7-4
adegenet	Exploratory Analysis of Genetic and Genomic Data	2.0.1
agricolae	Statistical Procedures for Agricultural Research	1.2-4
AlgDesign	Algorithmic Experimental Design	1.1-7.3
ape	Analyses of Phylogenetics and Evolution	3.5
assertthat	Easy pre and post assertions.	0.1
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.60.0-2
bitops	Bitwise Operations	1.0-6
blotter	Tools for Transaction-Oriented Trading Systems Development	0.9.1741
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18
Boruta	Wrapper Algorithm for All-Relevant Feature Selection	5.0.0

Installing a package in RStudio



	Name	Description	Version
	System Library		
☐	ade4	Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences	1.7-4
☐	adegenet	Exploratory Analysis of Genetic and Genomic Data	2.0.1
☐	agricolae	Statistical Procedures for Agricultural Research	1.2-4
☐	AlgDesign	Algorithmic Experimental Design	1.1-7.3
☐	ape	Analyses of Phylogenetics and Evolution	3.5
☐	assertthat	Easy pre and post assertions.	0.1
☐	base64enc	Tools for base64 encoding	0.1-3
☐	BH	Boost C++ Header Files	1.60.0-2
☐	bitops	Bitwise Operations	1.0-6
☐	blotter	Tools for Transaction-Oriented Trading Systems Development	0.9.1741
☐	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18
☐	Boruta	Wrapper Algorithm for All-Relevant Feature Selection	5.0.0

Installing a package in RStudio



Install Packages

Install from: Configuring Repositories

Repository (CRAN)

Packages (separate multiple with space or comma):

agricolae

Install to Library:

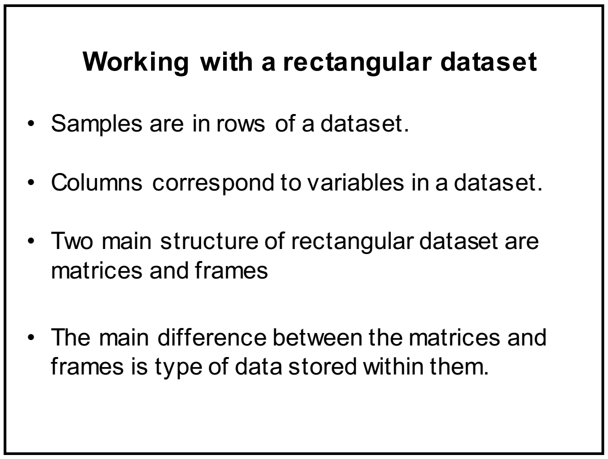
/Library/Frameworks/R.framework/Versions/3.3/Resources/libı

☑ Install dependencies

Install | Cancel

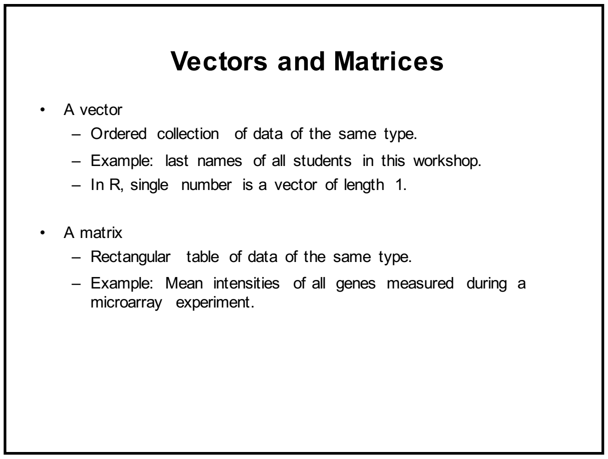
		Description
		Analysis of Ecologic Euclidean Methods i
		Exploratory Analysis
		Statistical Procedure
		Algorithmic Experim
☐	ape	Analyses of Phyloge
☐	assertthat	Easy pre and post a:
☐	base64enc	Tools for base64 en
☐	BH	Boost C++ Header F
☐	bitops	Bitwise Operations
☐	blotter	Tools for Transactio

Basic data types



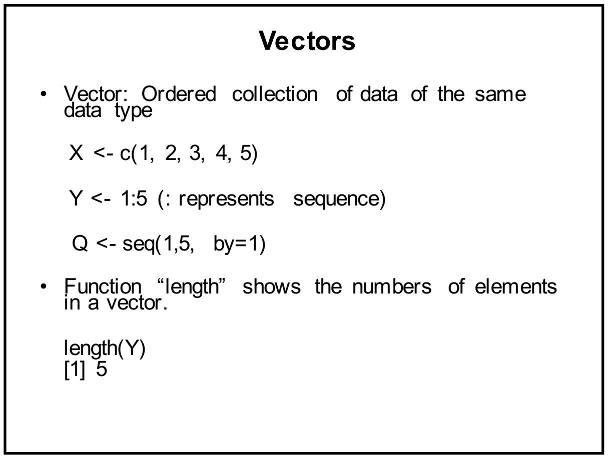
Working with a rectangular dataset

- Samples are in rows of a dataset.
- Columns correspond to variables in a dataset.
- Two main structure of rectangular dataset are matrices and frames
- The main difference between the matrices and frames is type of data stored within them.



Vectors and Matrices

- A vector
  - Ordered collection of data of the same type.
  - Example: last names of all students in this workshop.
  - In R, single number is a vector of length 1.
- A matrix
  - Rectangular table of data of the same type.
  - Example: Mean intensities of all genes measured during a microarray experiment.



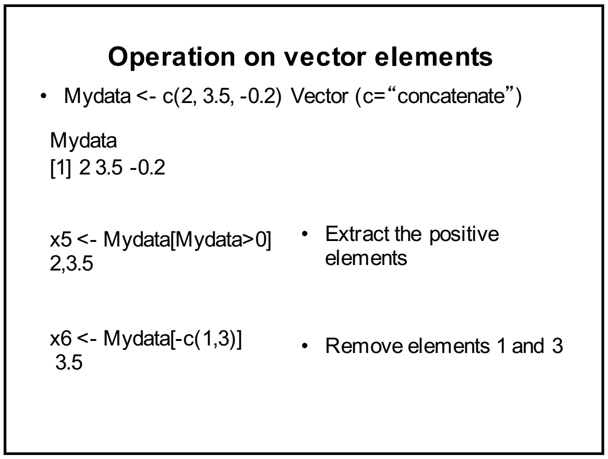
Vectors

- Vector: Ordered collection of data of the same data type

```
 X <- c(1, 2, 3, 4, 5)

 Y <- 1:5 (: represents sequence)

 Q <- seq(1,5, by=1)
```

- Function "length" shows the numbers of elements in a vector.

```
length(Y)
[1] 5
```



Operation on vector elements

- Mydata <- c(2, 3.5, -0.2) Vector (c="concatenate")

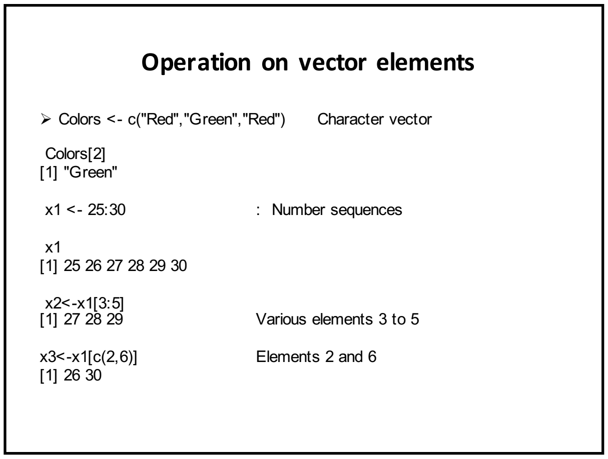
```
Mydata
[1] 2 3.5 -0.2
```

```
x5 <- Mydata[Mydata>0]
2,3.5
```

- Extract the positive elements

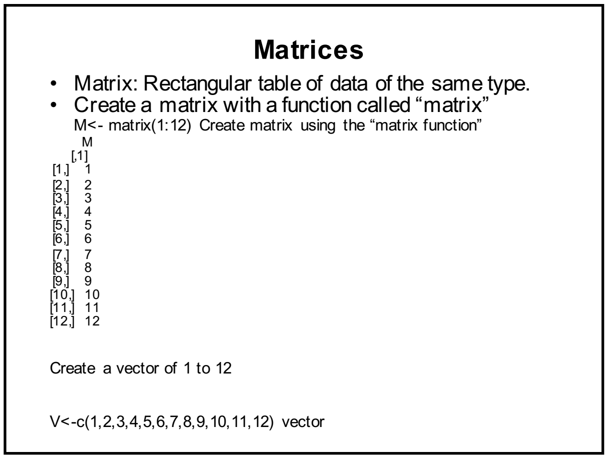
```
x6 <- Mydata[-c(1,3)]
 3.5
```

- Remove elements 1 and 3



Operation on vector elements

```
➢ Colors <- c("Red","Green","Red")       Character vector

 Colors[2]
[1] "Green"

 x1 <- 25:30                      :  Number sequences

 x1
[1] 25 26 27 28 29 30

 x2<-x1[3:5]
[1] 27 28 29                      Various elements 3 to 5

x3<-x1[c(2,6)]                    Elements 2 and 6
[1] 26 30
```



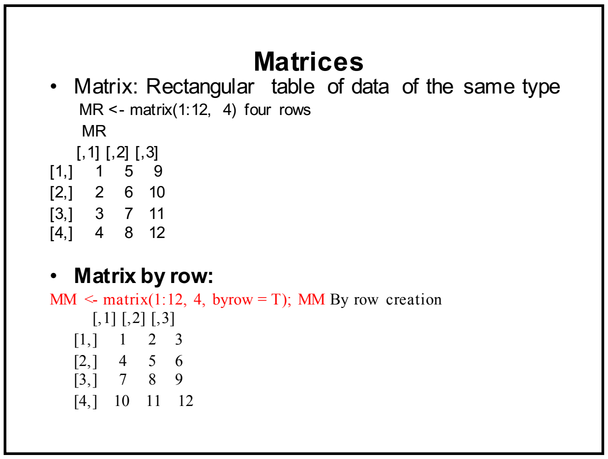
Matrices

- Matrix: Rectangular table of data of the same type.
- Create a matrix with a function called "matrix"

```
   M<- matrix(1:12)  Create matrix  using the "matrix function"
     M
    [,1]
[1,]   1
[2,]   2
[3,]   3
[4,]   4
[5,]   5
[6,]   6
[7,]   7
[8,]   8
[9,]   9
[10,]  10
[11,]  11
[12,]  12
```

Create a vector of 1 to 12

```
V<-c(1,2,3,4,5,6,7,8,9,10,11,12)  vector
```



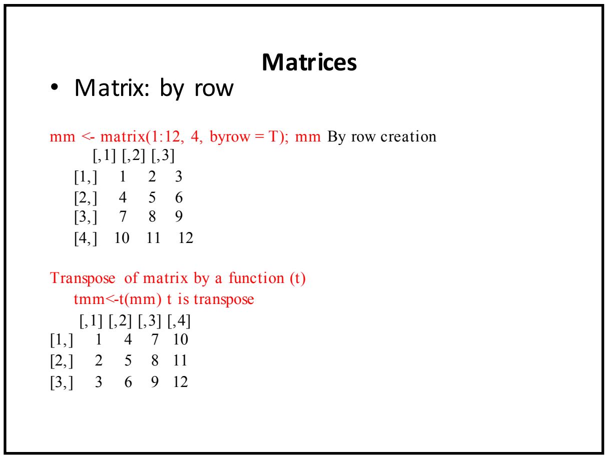
Matrices

- Matrix: Rectangular table of data of the same type

```
    MR <- matrix(1:12, 4) four rows
     MR
    [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
```

- **Matrix by row:**

```
MM <- matrix(1:12, 4, byrow = T); MM By row creation
        [,1] [,2] [,3]
    [1,]    1    2    3
    [2,]    4    5    6
    [3,]    7    8    9
    [4,]   10   11   12
```



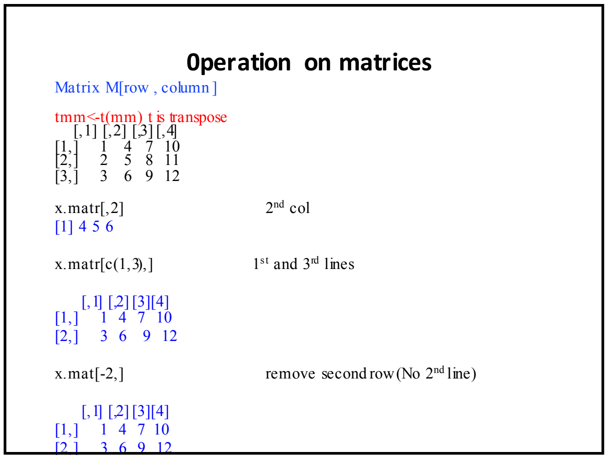
Matrices

- Matrix: by row

```
mm <- matrix(1:12, 4, byrow = T); mm By row creation
        [,1] [,2] [,3]
    [1,]    1    2    3
    [2,]    4    5    6
    [3,]    7    8    9
    [4,]   10   11   12
```

Transpose of matrix by a function (t)

```
    tmm<-t(mm) t is transpose
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```



Operation on matrices

Matrix M[row , column]

```
tmm<-t(mm) t is transpose
    [,1] [,2] [,3] [,4]
[1,]   1    4    7   10
[2,]   2    5    8   11
[3,]   3    6    9   12

x.matr[,2]                    2nd col
[1] 4 5 6

x.matr[c(1,3),]               1st and 3rd lines

     [,1] [,2] [3][4]
[1,]    1   4   7  10
[2,]    3   6   9  12

x.mat[-2,]                    remove second row(No 2nd line)

     [,1] [,2] [3][4]
[1,]    1   4   7  10
[2,]    3   6   9  12
```

Data frame

Data frame

Data frame:

Rectangular table with rows and columns; data within each column has the same type (e.g. number, text, logical), but different columns may have different types.

Example of a data frame with 10 rows and 3 columns

Name	Phone	Country
Stephen	25677643240	Uganda
Richard	19545551234	USA
Monica	2547876655	Kenya
Fred	54113876	Argentina
Jessica	448756509	United Kingdom
Milly	3389756585	France
Norbert	5876454534	Venezuela
Jenifer	23480312345	Nigeria
Jimmy	866586968405	Taiwan
Rose	861069445464	China

Creating a data frame

- # create a data frame from scratch using "data.frame" function
- age <- c(25, 30, 56, 49, 12, 16, 60, 34, 45, 22)
- gender <- c("male", "female", "male","male", "female", "male","male", "female", "male","male")
- weight <- c(160, 110, 220, 100, 65, 120, 179, 134, 165, 153)
- mydata <- data.frame(age,gender,weight)

Importing and exporting data frame in R

Importing Data

- The easiest way to enter data in R is to work with a text file, in which the columns are separated by tabs; or comma-separated values (csv) files .
- Example of importing data are provided below (console).

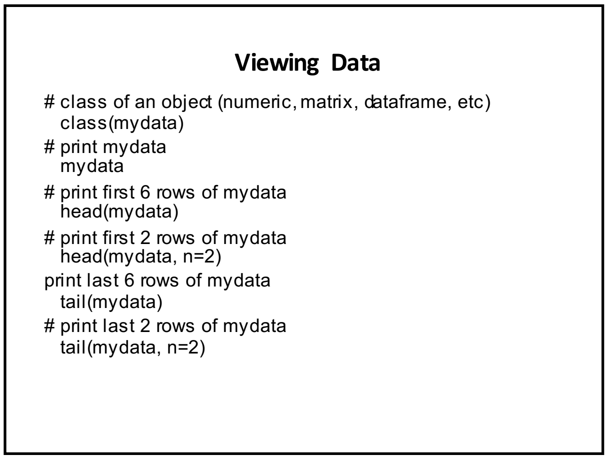
```
mydata  <- read.table("D1_Data_1.csv",    sep=”, ", header=TRUE)
mydata  <- read.csv("D1_Data_1.csv",    header=TRUE)
mydatab<-  read.table("D1_Data_1.txt”,    sep=”,\t", header=TRUE)
mydatab<-  read.delim("D1_Data_1.txt",    header=TRUE)
```

- **Importing data in Rstudio using (Import Dataset) on the Workspace**

Viewing Data

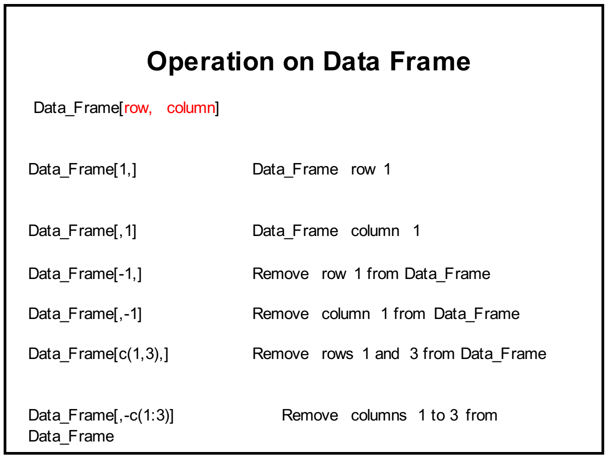
There are a number of functions for listing the contents of an object or dataset.

```
# list the variables in mydata
   names(mydata)

# list the structure of mydata
   str(mydata)

# dimensions of an object
   dim(mydata)
```



Viewing Data

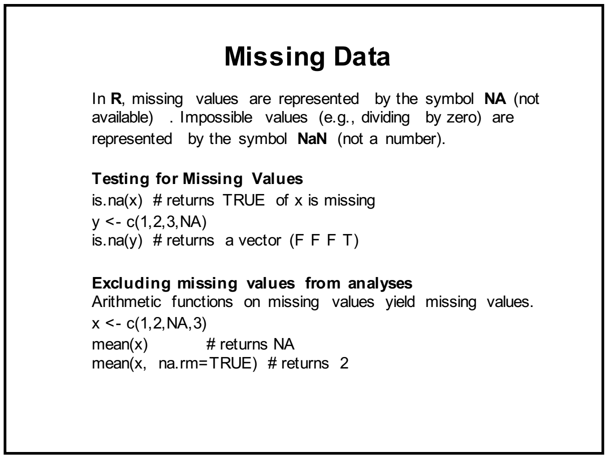
```
# class of an object (numeric, matrix, dataframe, etc)
  class(mydata)
# print mydata
  mydata
# print first 6 rows of mydata
  head(mydata)
# print first 2 rows of mydata
  head(mydata, n=2)
print last 6 rows of mydata
  tail(mydata)
# print last 2 rows of mydata
  tail(mydata, n=2)
```



Operation on Data Frame

Data_Frame[row, column]

Data_Frame[1,]	Data_Frame row 1
Data_Frame[,1]	Data_Frame column 1
Data_Frame[-1,]	Remove row 1 from Data_Frame
Data_Frame[,-1]	Remove column 1 from Data_Frame
Data_Frame[c(1,3),]	Remove rows 1 and 3 from Data_Frame
Data_Frame[,-c(1:3)]	Remove columns 1 to 3 from Data_Frame



Missing Data

In **R**, missing values are represented by the symbol **NA** (not available) . Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number).

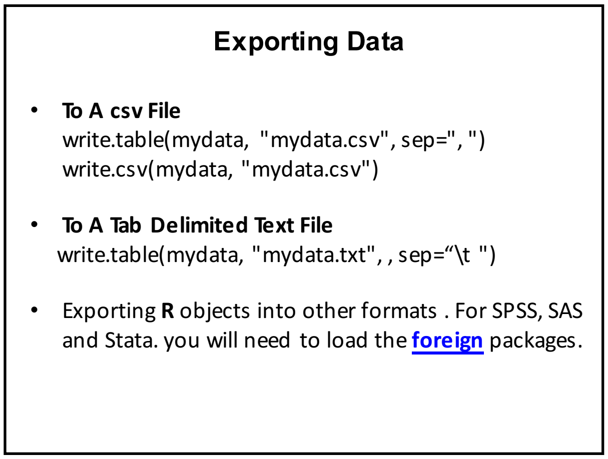
Testing for Missing Values

```
is.na(x) # returns TRUE of x is missing
y <- c(1,2,3,NA)
is.na(y) # returns a vector (F F F T)
```

Excluding missing values from analyses

Arithmetic functions on missing values yield missing values.

```
x <- c(1,2,NA,3)
mean(x)           # returns NA
mean(x, na.rm=TRUE) # returns 2
```



Exporting Data

- **To A csv File**

```
write.table(mydata, "mydata.csv", sep=", ")
write.csv(mydata, "mydata.csv")
```

- **To A Tab Delimited Text File**

```
write.table(mydata, "mydata.txt", , sep="\t ")
```

- Exporting **R** objects into other formats . For SPSS, SAS and Stata. you will need to load the **foreign** packages.

Hands on exercise