Data Sciences Course

(Day X - Session X)

R Simplified

Covers

R basics and startup for data summarization

Understanding R Easily – Getting Data into R

Getting Data into R:

- Several ways – lets have overview of some in day-to-day functions

Entering data

- "c" function – combine or concatenates terms
- Ex: XYZ <- c(10, 15, 12, 18)
- Ex: matrix(c(10, 15, 12, 18), nrow=2, ncol=2)
- Ex: matrix(c(10, 15, 12, 18), nrow=2, ncol=2, byrow = TRUE)
- Ex: array(c(10, 15, 12, 18), dim = c(2, 2))
- Ex: data.entry()

Reading external data

- icities <- read.table("India_Cities.csv",
  header = T,
  sep = ",",
  dec = ".",
  nrows = -1,
  comment.char = "",
  strip.white = T,
  stringsAsFactors = F)
- will read the external file as a data.frame

Reading/Loading data

- load("/dir/icities.RData")
- library(xx); data(xx);
- recollect command;
- try(data(package = "ggplot2"))
- data(cancer, package = "survival")

Generating data

- Ex: a1 <- rnorm(100)
- Ex: a2 <- seq(1, 10, by = 0.01)
- Ex: a3 <- rep(1, 100)
- Ex: a4 <- runif(100)
- Ex: cbind(a1, a2, a3, a4)
- Ex: rbind(a1, a2, a3, a4)

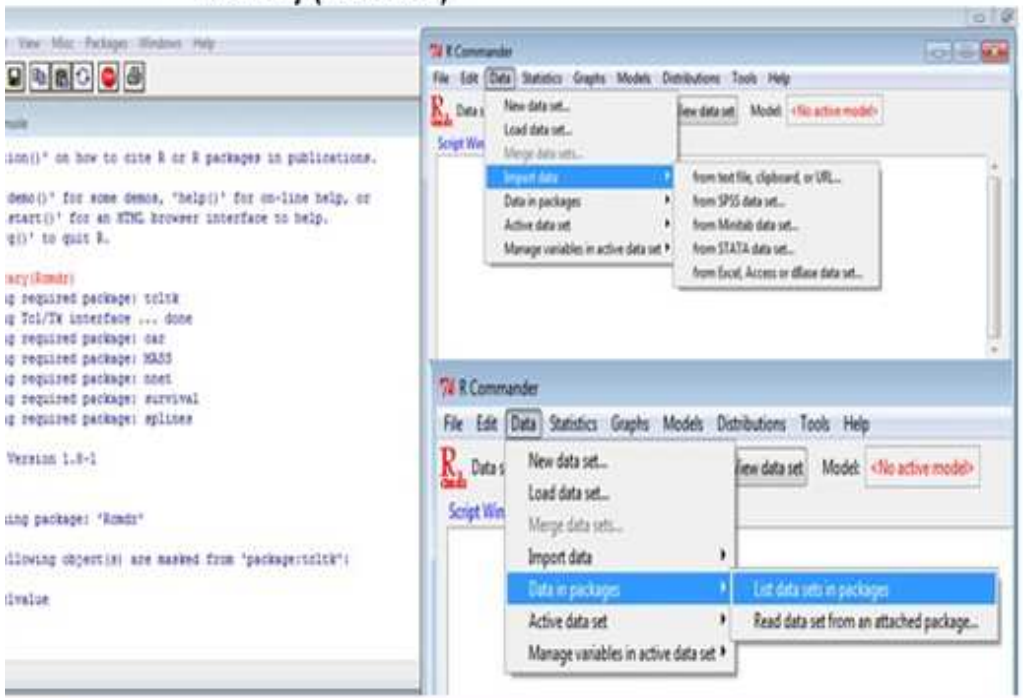
Reading External data

- read.table.url(url) ;
- download.file(url); url.show(url)
- read.csv(f, header=TRUE)
- read.csv2(f, header=TRUE, sep=";" quote="\"", dec=",")
- read.delim(f, header=TRUE, sep="\t", quote="\"", dec=".")
- read.delim2(f, header=TRUE, sep="\t", quote="\"", dec=",")
- read.fwf(file, widths=c(3,5,3), header="FALSE, sep="", as.is=FALSE
- as.is=TRUE; # not to be converted into a factor
- na.strings=c(".", "NA", "", "#") # characters for missing
- save(icities, file = "icities.Rdata")

Understanding R Easily – Using GUI's

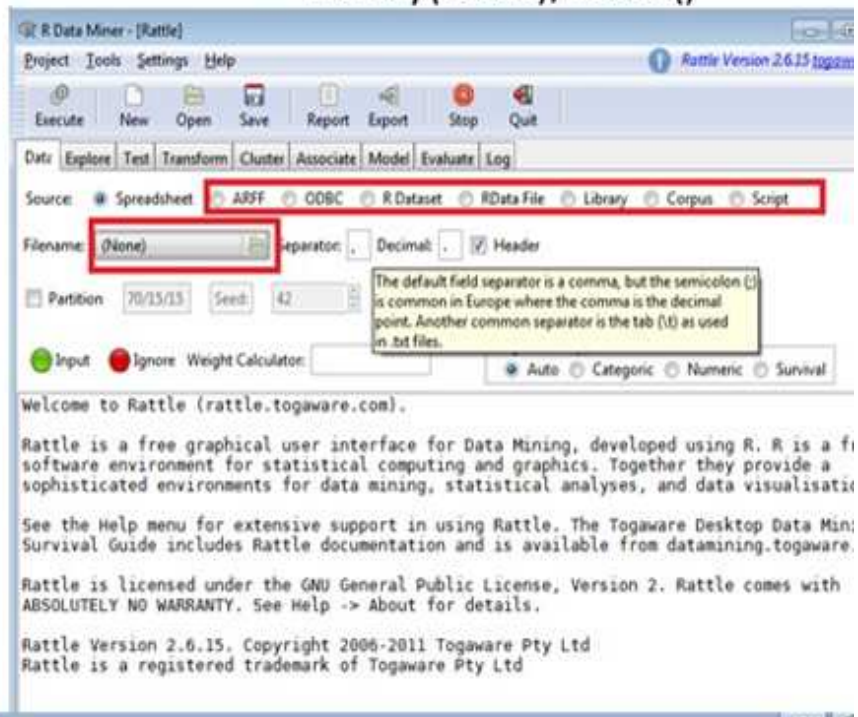
✓ Getting Data in R:

- library(Rcmdr)



✓ Getting Data in R:

- library(rattle); rattle()



✓ Getting XLSX & XLS files in R:

- library(readxl) # damn slow
- test1 <-read_excel("xx:/folder/filename.xslx", sheet = 1, col_names = T)
- library(xlsx)
- test2 <- read.xlsx("datafile.xlsx", sheet=1)

Understanding R Easily – Now lets us know what is in read data

- ✓ **str() # compactly displays the structure of an arbitrary R object**
- ✓ **variable info:**
  - names(df);
  - length(df) # number of elements
  - attributes(df); # more details than "str()"
  - nrow(df); ncol(df)
  - is.data.frame(df)
  - is.character(df$x1)
  - is.numeric(df$x1)
  - is.matrix(df)
  - dim(df)
  - dimnames(df)
- ✓ **Best practice is to have a data summary sheet.**

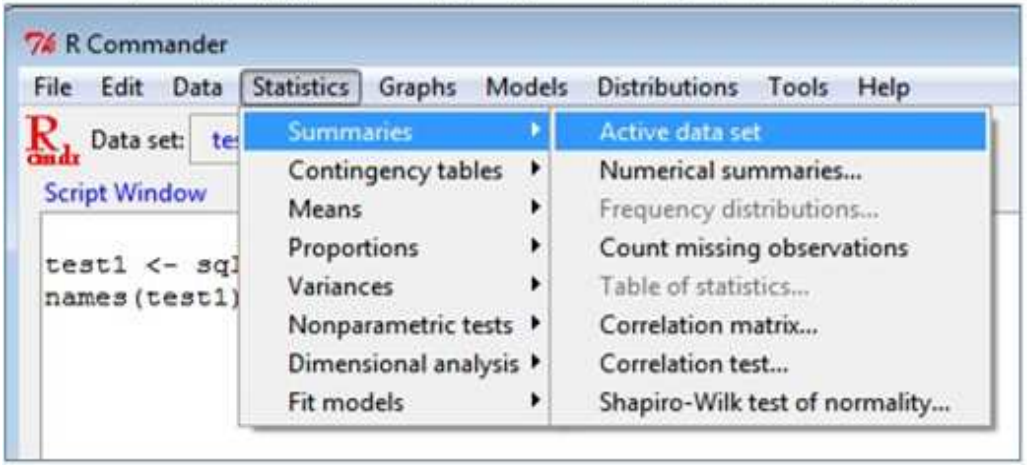
VDM_Data_Summary_File

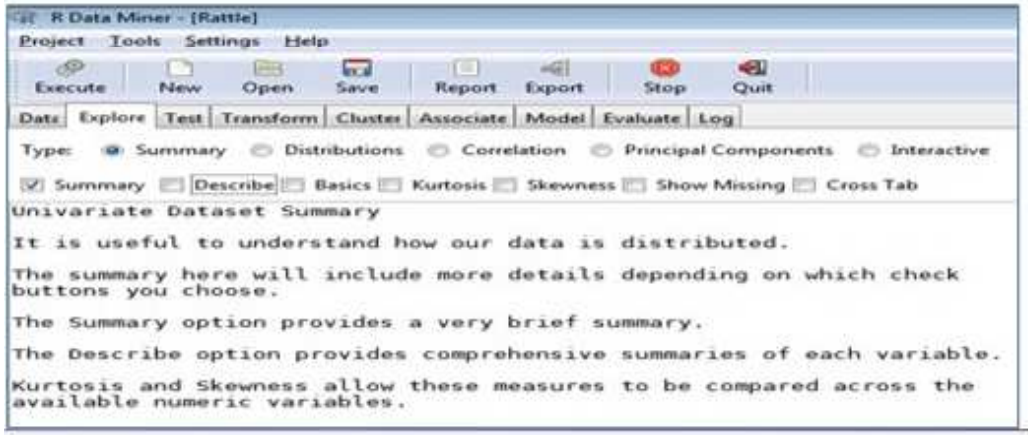
Useful functions and operators for understanding and analyzing the data

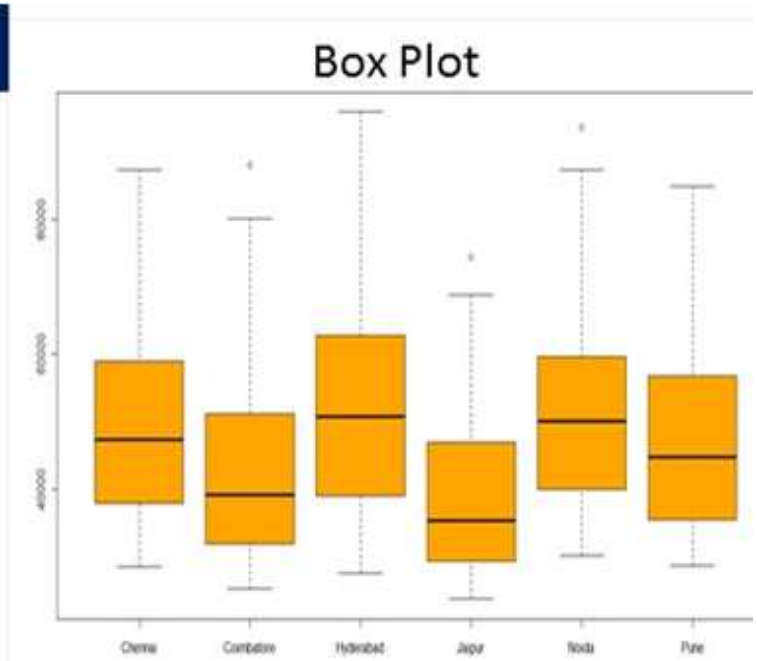
Mathematical & Statistical		Arithmetic		Logical	
log	Natural Logarithm	+	Addition	! X	Logical negation (NOT)
exp	Exponential	-	Subtraction	X & Y	Logical And
sin	SINE	*	Multiplication	X | Y	Logical OR
cos	COSINE	/	Division	xor	indicates elementwise exclusive OR
tan	Tangent	^	Power	Matrices	
sqrt	Square Root	Comparison		t(x)	Transpose of x
min	Returns minimum among numbers	<	Lesser than	diag(x)	Diagonal of x
max	Returns maximum among numbers	>	Greater than	%*%	Matrix Multiplication
sum	Sum of numbers	<=	Lesser than or equal to	solve(x)	Matrix inverse of x
mean	Averages of numbers	>=	Greater than or equal to	rowSums(x)	Sum of rows for a matrix-like object
var	Variance of numbers	==	Equal	colSums(x)	Sum of columns for a matrix-like object
sd	Standard Deviation of numbers	!=	Not equal to	rowMeans(x)	Average of rows for a matrix-like object

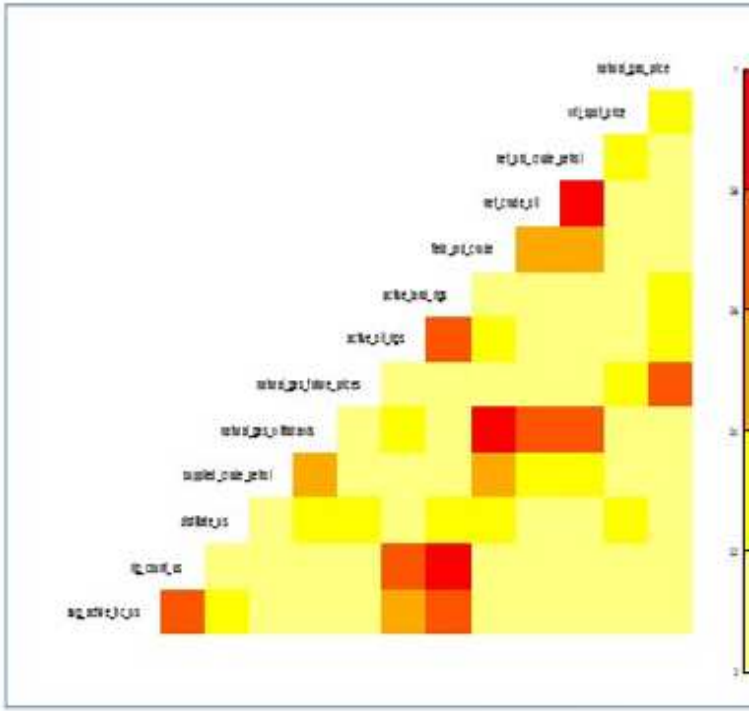
Understanding R Easily – Data Summarization

- ✓ Data summarization/descriptive functions
  - summary(df); fivenum(numeric); cor(df);
  - library(Hmisc); describe(df);
  - mean(); median(); max(); min(); std(); var();
  - library(arm); corrplot(df, col = T);
- ✓ Descriptive plots
  - boxplot(df); dotchart(df); paris(df)
  - hist(df$x1); barplot(df$x1)
  - plot(x, y); plot.ts(x); qqnorm(x); qqplot(x, y)





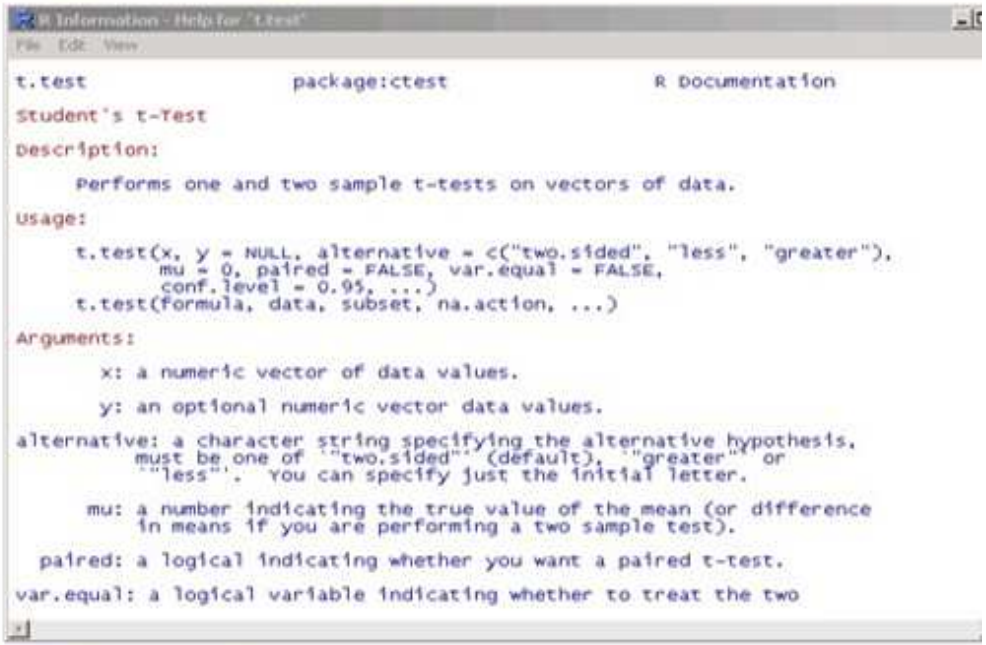


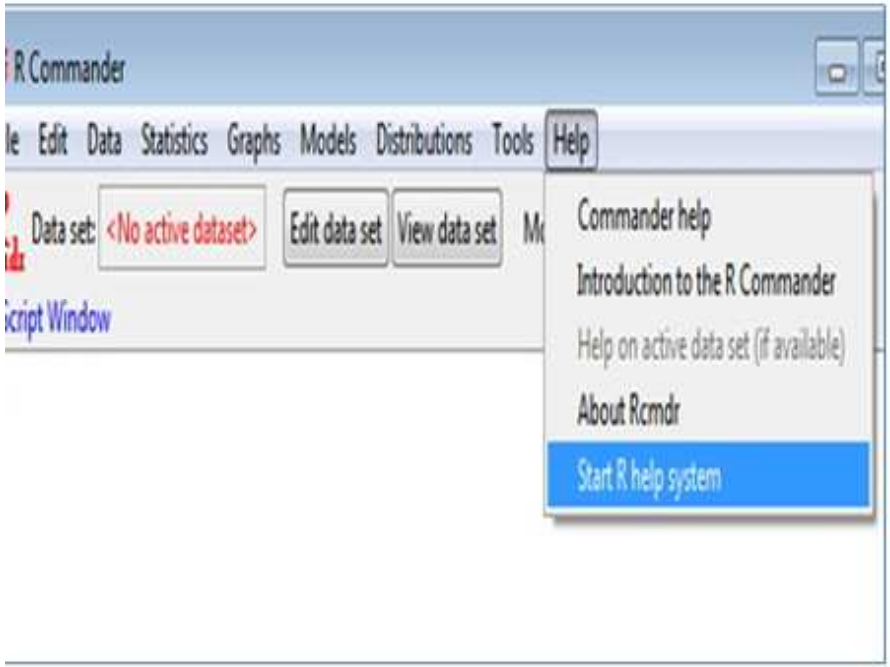


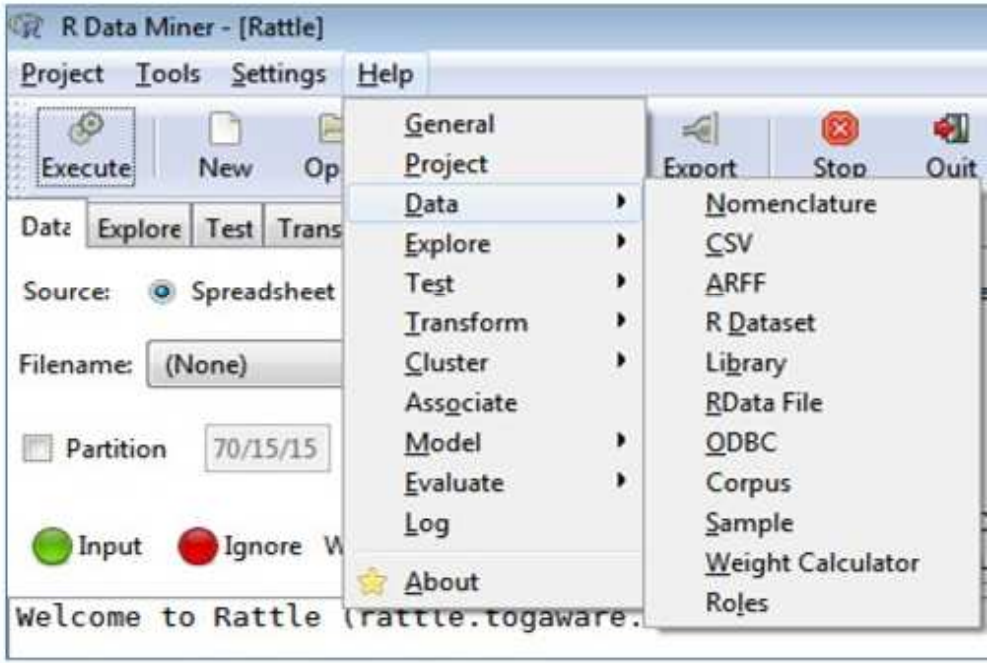
Understanding R Easily – Getting Help in R

Has built-in help pages?

- help(t.test) or ?t.test
- help.start() # starts help page (both in R GUI & Rstudio)
- search() # prints the list of default loaded/attached packages
- help.search("linear models") # searches the help systems and provides all matching string results







Understanding R Easily – Date & Time Formats

✓ Dates & Times

- 'as.Date()' – built-in function – to handle dates.
- 'chron' package has few better options to handle dates, but cannot handle time zones.
- 'POSIXct' and 'POSIXlt' – built-in functions – for handle/control time zones.
- 'Sys.Date()' prints today's date and 'Sys.time()' prints both today's date and current time.
- R also follows – default date of origin – 1970-01-01 - programming languages - it is day zero.
- Except for - POSIXlt class - all dates are stored internally - as the number of days or seconds from reference date.
- Thus, dates will have - a numeric mode - however - class function – explains - format.
- Default format/type of dates – "YYYY-MM-DD" which in R takes the code value as "%Y-%m-%d".
- Coming to POSIXlt class - it stores - date/time values - as a list of components (hour, min, sec, mo etc.) - making them easy for extract time and date components.
- e.g. as.Date('1995-6-16') prints "1995-06-16"
- e.g. as.Date('2/14/2002',format='%m/%d/%Y') prints "2002-02-14"
- e.g. as.Date('April 27, 2009',format='%B %d, %Y') prints "2009-04-27"
- e.g. as.Date('21JUL11',format='%d%b%y') prints "2011-07-21" # better not to use 2-digit year

Time Unit	R-symbol & meaning	Illustration
Day	(%d) Day as a number	01-31
Days of the year	(%j) Days of year as decimal number	001-366
Hours (24)	(%H) Hours as decimal number	00-23
Hours (12)	(%I) Hours as decimal number	01-12
Abbreviated Weekday	(%a) Abbreviated Weekday	Mon
Unabbreviated Weekday	(%A) Unabbreviated Weekday	Monday
Weekday	(%w) Weekday as decimal number	0-6 (0=Sunday)
Week of the year	(%W) Week of the year as decimal	First Monday as day 1 of week 1
Week of the year	(%U) Week of the year as decimal	First Sunday as day 1 of week 1
Month	(%m) Month as a number	01-12
Abbreviated Month	(%b) Abbreviated Month	JUL

Understanding R Easily – Date & Time Formats (contd.)

Time Unit	R-symbol & meaning	Illustration
Unabbreviated Month	(%B) Unabbreviated Month	JULY
Year	(%Y) Four-Digit Year	2011
Year	(%y) Two-Digit Year	11
Minute	(%M) Minute as decimal number	00-59
AM/PM	(%p) AM, PM indicatore in locale (used wih Hours)	AM or PM
Seconds	(%S) Second a decimal number (allows for up-to 2 leap seconds)	00-61
Time Zone	(%z) Abbreviated Time Zone	IST
Locale Options	(%x) Date, Locale Specific Option	Locale Specific
Locale Options	(%X) Time, Locale Specific Option	Locale Specific
Locale Options	(%c) Date and Time, Locale Specific Option	Locale Specific

✓ Don't get confused with date & time classes with data classes and also try 'lubridate' package

Date & Time Class	Data Class
Date or as.Date	ts
POSIXct or as.POSIXct	data.frame
POSIXct or as.POSIXct	vector
numeric	matrix
character	fts
chron	timeSeries
yearmon	mts
yearqtr	its
timeDate	irts
force_tz	zoo

Session open for questions