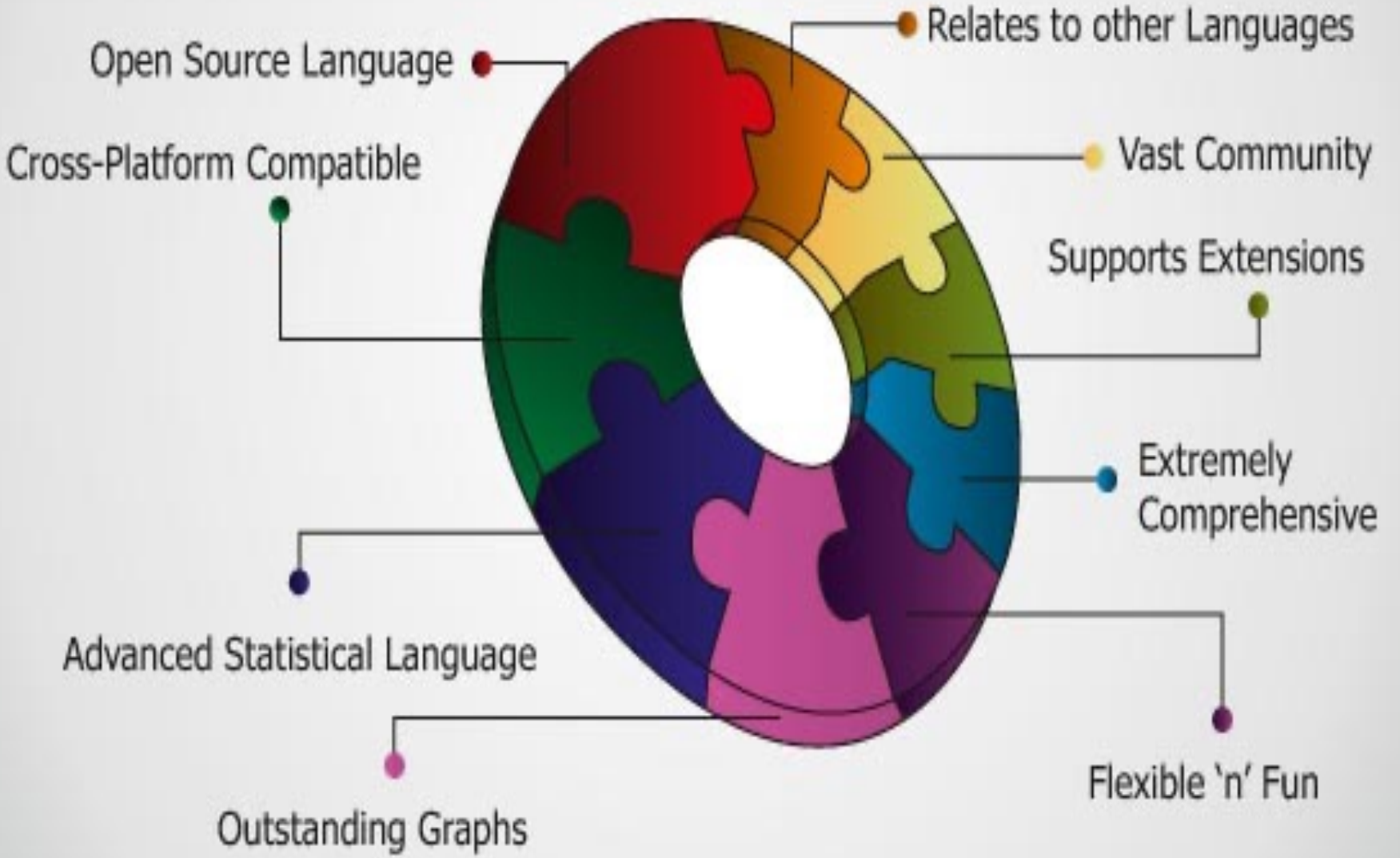
Getting Started with R

What Will We Learn

- R background and environment
- Create, modify data in R
- Know your data using R functions
- Import and Export data
- Use help in R
- A glance at R Studio

Why Learn R?



DATA SCIENCE INSTITUTE

What is R?

- A language and environment for statistical computing and graphics.
- Wide variety of statistical & graphical techniques built in mainly used in education and as a research tool.
- Free and Open Source software.
- Compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

History of R

- The first version of R was developed by Robert Gentleman and Ross Ihaka, at the University of Auckland around 1997, for teaching S+.
- It is now under active development by a group of statisticians called 'the R core team', with a home page at www.r-project.org.
- R 1.0.0 was released in February 2000.
- R is very well accepted globally because of two of its important attributes viz. the ability to transform and to evolve.

Users of R

- 2 million Users and thousands of developers
- Academicians and Researchers
- Banks like ANZ, Bank of America
- Regulators like FDA (Food and Drug Administration)
- Social Media giants like Facebook and Twitter
- Google, Mozilla, New York Times, Thomas Cook, Uber.....

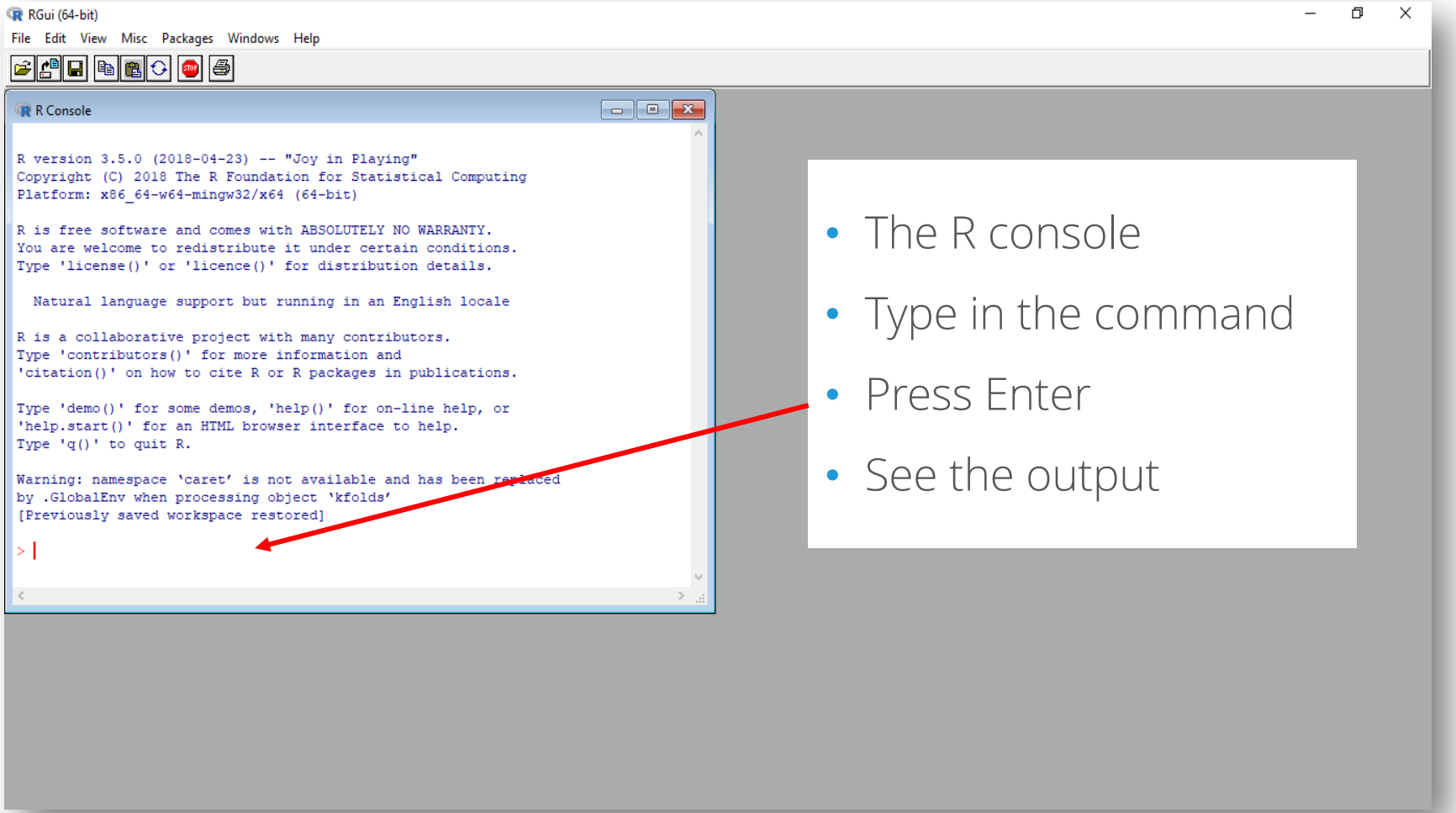
R Environment

- Most functionality through built in functions. Basic and advanced functions available by default.
- R is built using packages. There are approximately 12000 packages.
- All datasets created in the session remain in Memory.
- Output can be used as input to other functions.
- R commands are Case Sensitive (mean is not same as Mean).

R-CRAN

- The Comprehensive R Archive Network.
- A network of global web servers storing identical, up-to-date, versions of code and documentation for R.
- Use the CRAN mirror nearest to you to download R setup at a faster speed.
- http://cran.r-project.org/

R- User Interface



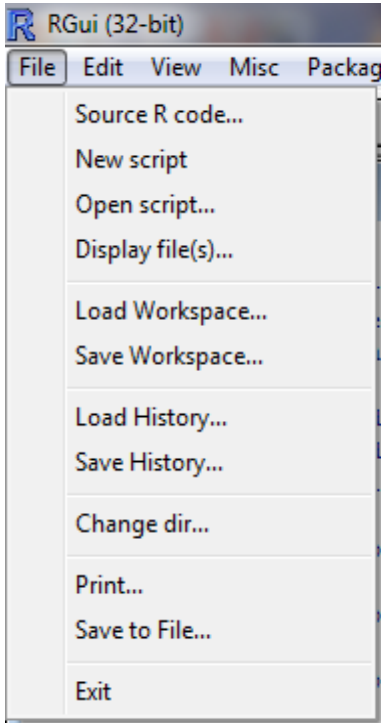
DATA SCIENCE INSTITUTE

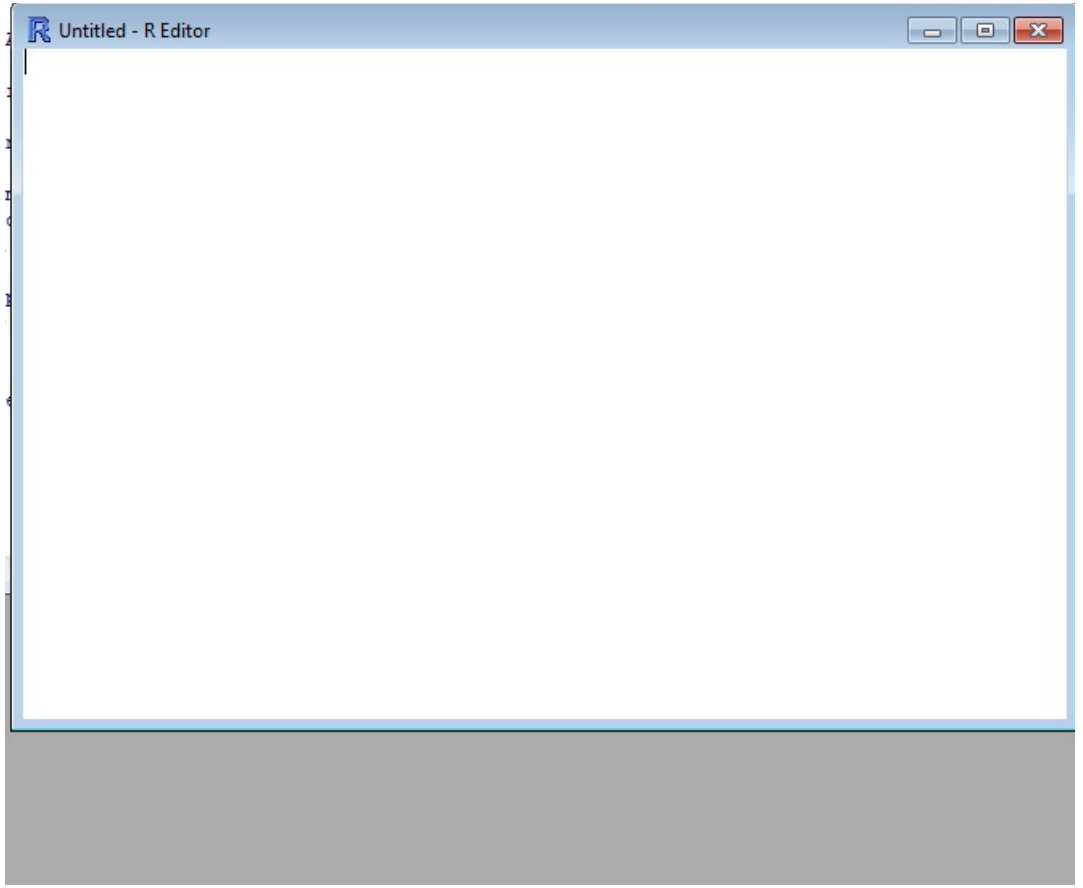
R Package

- A Package is a collection of R functions with comprehensive documents.
- A Package includes: R functions, Data Example, Help Files, Namespace and Description.
- Many packages are installed by default when R is installed on the computer.
- The function of R could be extended by loading R packages.

Create and Save Your Script

- Open a new script, write commands, execute using F5 key. Save the file at a desired folder. Helps to save the script for a later use.





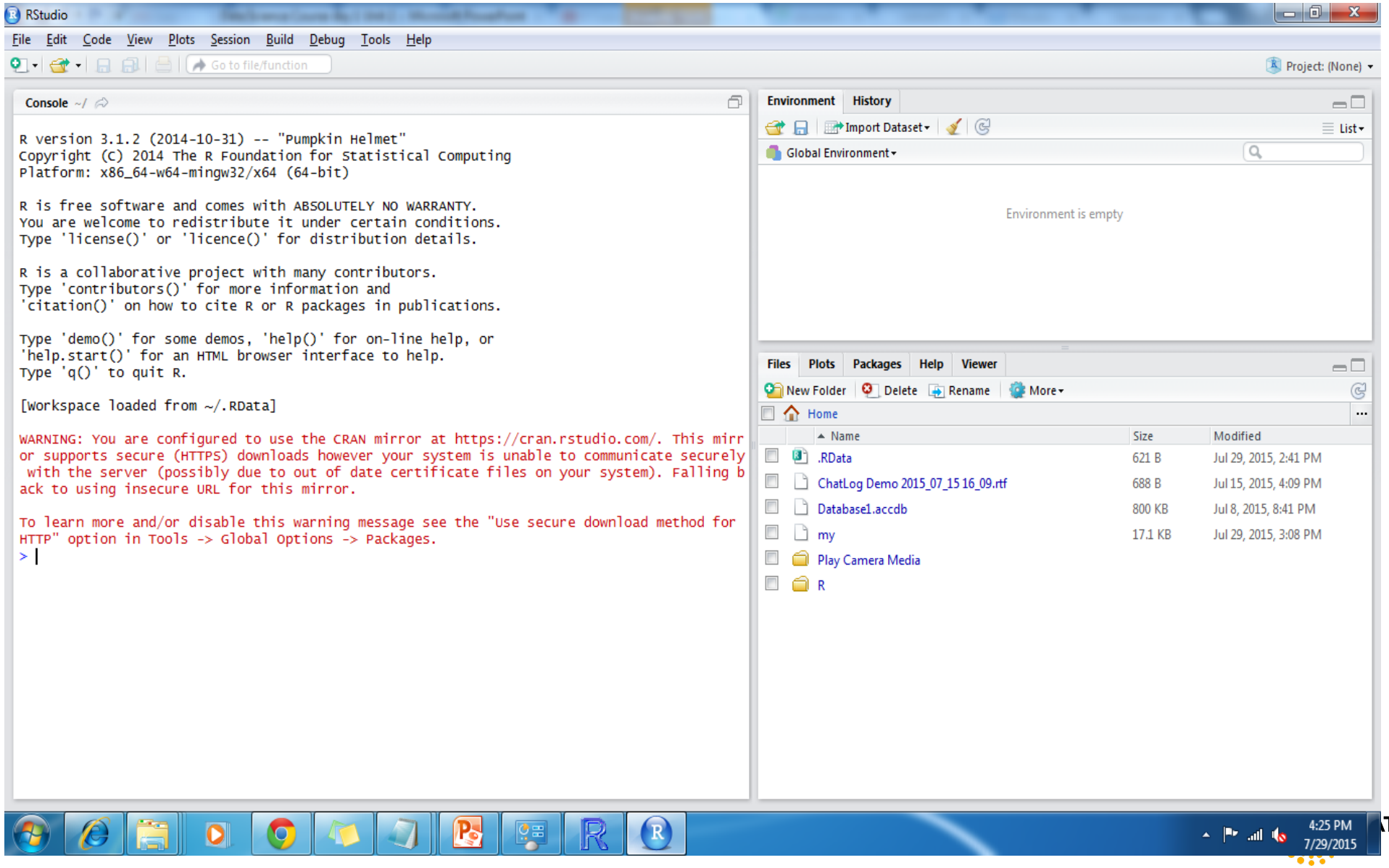
DATA SCIENCE INSTITUTE

R Studio

- RStudio is a free and open source integrated development environment (IDE) for R.
- User friendly.
- Features like code completion.
- Has an organized layout and several extra options.
- The usual Rstudio screen has four windows:

- Console
- Workspace and history
- Files, plots, packages and help
- The R script(s) and data view.

R Studio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
Console ~/
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 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.
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.
[Workspace loaded from ~/.RData]
WARNING: You are configured to use the CRAN mirror at https://cran.rstudio.com/. This mirr
or supports secure (HTTPS) downloads however your system is unable to communicate securely
with the server (possibly due to out of date certificate files on your system). Falling b
ack to using insecure URL for this mirror.
To learn more and/or disable this warning message see the "Use secure download method for
HTTP" option in Tools -> Global Options -> Packages.
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Name Size Modified
.RData 621 B Jul 29, 2015, 2:41 PM
ChatLog Demo 2015_07_15 16_09.rtf 688 B Jul 15, 2015, 4:09 PM
Database1.accdb 800 KB Jul 8, 2015, 8:41 PM
my 17.1 KB Jul 29, 2015, 3:08 PM
Play Camera Media
R
4:25 PM
7/29/2015

Create Your Data Set

- x<-c(12,23,45)
- y<-c(13,21,6) Create vectors x, y, z on R Console
- z<-c("a","b","c")
- data1<-data.frame(x,y,z) Combine them in a table
- data1 Type data name for output

- #Note that R is Case-Sensitive

	x	y	z
1	12	13	a
2	23	21	b
3	45	6	c

DATA SCIENCE INSTITUTE

Try Some Basic Functions

```
min(data1$x)
[1] 12
```

Use $ sign after data set name to use specific columns

```
max(data1$y)
[1] 21
```

```
length(data1$x)
[1] 3
```

```
mean(data1$y)
[1] 13.33333
```

```
levels(data1$z)
[1] "a" "b" "c"
```

Gives a list of unique categories in the data

DATA SCIENCE INSTITUTE

Deriving & Removing Variables

Add new variables using mathematical operators

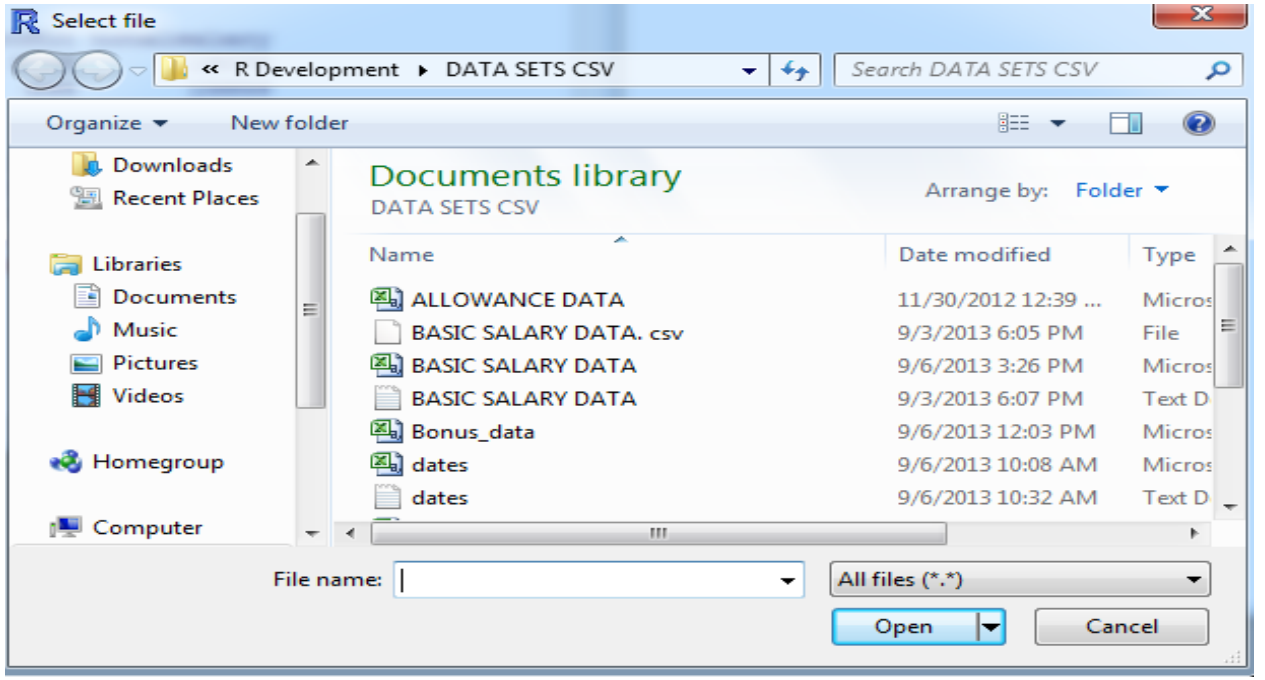
- data1$sum <- data1$x+data1$y
- > data1
- x y z sum
- 1 12 13 a 25
- 2 23 21 b 44
- 3 45 16 c 61

- Delete an unwanted column from your table by using the following method
- data1$x <- NULL

Export Your Updated Table

- write.csv(data1,file.choose())

Select the path, name your file with .csv extension, click open and Yes to export your table



DATA SCIENCE INSTITUTE

Data Object Types

<table>
<tr><td>• Vector</td><td>• 1 dimension</td><td rowspan="3">• All elements have the same data types

• Numeric
• Character
• Logic
• Factor</td></tr>
<tr><td>• Matrix</td><td>• 2 dimensions</td></tr>
<tr><td>• Array</td><td>• 2 or more
• dimensions</td></tr>
<tr><td>• Data frame</td><td>• 2 dimensions</td><td>• Table-like data object allowing different data types for different columns</td></tr>
<tr><td>• List</td><td colspan="2">• Collection of data objects, each element of a list is a data object</td></tr>
</table>

R Workspace

- Objects that you create during an R session are held in memory, the collection of objects that you currently have is called the workspace.
- This workspace is not saved on disk unless you tell R to do so.
- This means that your objects are lost when you close R and not save the objects, or worse when R or your system crashes on you during a session.

R Workspace

- If you have saved a workspace image and you start R the next time, it will restore the workspace
- So all your previously saved objects are available again
- You can also explicitly load a saved workspace i.e., it could be the workspace image of someone else.
- Go the 'File' menu and select 'Load workspace'

R Workspace

- #Display all previous commands
- history()
- #Display last 25 commands

- history(max.show=25)
- #Save your command history to a file. Default is ".Rhistory“

- savehistory(file="myfile")
- #Myfile is saved in My Documents

- #Recall your command history. Default is ".Rhistory”
- loadhistory(file="myfile")

In-Memory Computing

- Storage of information in the main RAM of dedicated servers rather than in complicated relational databases operating on comparatively slow disk drives

- Helps business customers, including retailers, banks and utilities, to quickly detect patterns, analyze massive data volumes on the fly, and perform their operations quickly

Import .csv Data File

- basic_salary<
- read.table("C:/Users/Dell/Desktop/BASIC_SALARY.csv",

  header=TRUE, sep=",")

- Command requires the file path separated by a /
- header=TRUE if there are column labels
- For CSV data file, read.table can be replaced by read.csv. In this case sep="," is not required.

- Instead of specifying path, file.choose() can be used and path can be selected interactively.

- basic_salary<-read.csv(file.choose(),header=T)

Check Your Data

- Use head function to get an idea about how your data looks like
- Note that it displays first 6 rows by default.

- > head(basic_salary)

- First_Name Last_Name Grade Location ba
- 1 Mahesh Joshi GR1 DELHI 17990
- 2 Rajesh Kolte GR1 DELHI 19250
- 3 Neha Rao GR1 DELHI 19235
- 4 Priya Jain GR1 DELHI 23280
- 5 Sneha Joshi GR1 DELHI 20660
- 6 Mahesh Rane GR1 DELHI 23160

Check Your Data.

- Using tail function; Display the last 5 rows

- >tail(basic_salary,5)

- First_Name Last_Name Grade Location ba
- 37 Archa Narvekar GR2 MUMBAI 10940
- 38 Shiva Jathar GR2 MUMBAI 12860
- 39 Anu Bhutala GR2 MUMBAI 13650
- 40 Nita Punjabi GR2 MUMBAI 14050
- 41 Ketan Kharkar GR2 MUMBAI 13140

Check Your Data..

```
dim(basic_salary)
[1] 41  5

> str(basic_salary)
'data.frame':   41 obs. of  5 variables:
 $ First_Name: Factor w/ 37 levels "Ajit","Ameet",..: 19 29 23..
 $ Last_Name : Factor w/ 38 levels "Arora","Bhide",..: 10 13...
 $ Grade     : Factor w/ 2 levels "GR1","GR2": 1 1 1 1 1 1 1...
 $ Location  : Factor w/ 2 levels "DELHI","MUMBAI": 1 1 1 1...
 $ ba        : int  17990 19250 19235 23280 20660 23160...

> names(basic_salary)
[1] "First_Name" "Last_Name"  "Grade"      "Location"   "ba"
```

DATA SCIENCE INSTITUTE

Summary

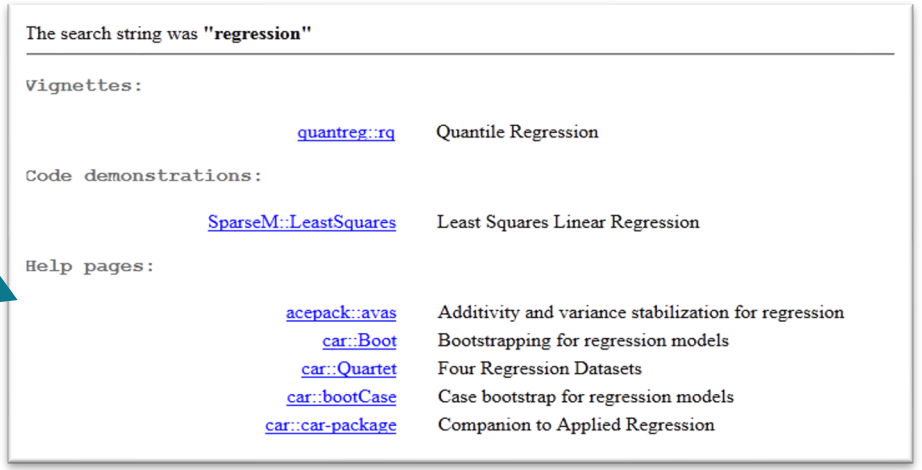
- Summarise all variables using summary function
- Summary function may show different output for "chr" variables.

- > summary(basic_salary)

- First_Name Last_Name Grade Location ba
- Kavita : 2 Joshi : 2 GR1:24 DELHI :19 Min. :10940
- Mahesh : 2 Shah : 2 GR2:17 MUMBAI:22 1st Qu.:13780
- Nishi : 2 Singh : 2 Median :16000
- Priya : 2 Arora : 1 Mean :17062
- Ajit : 1 Bhide : 1 3rd Qu.:19250
- Ameet : 1 Bhutala: 1 Max. :29080
- (Other):31 (Other):32

Help in R

- If you know the topic but not the exact function
- help.search("topic")
- > help.search("regression")


The search string was **"regression"**

Vignettes:

quantreg::rq	Quantile Regression

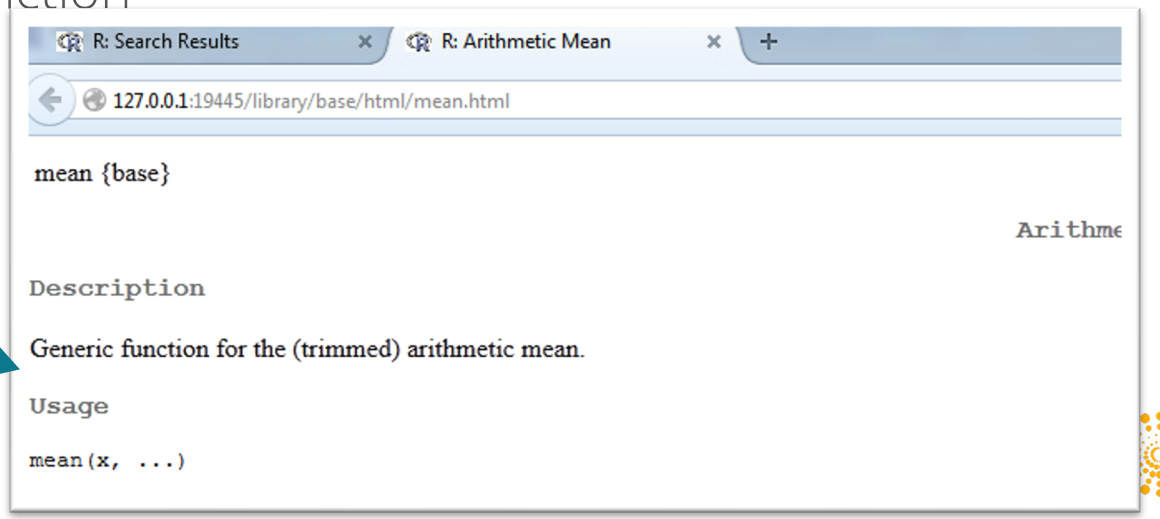
Code demonstrations:

SparseM::LeastSquares	Least Squares Linear Regression

Help pages:

acepack::avas	Additivity and variance stabilization for regression
car::Boot	Bootstrapping for regression models
car::Quartet	Four Regression Datasets
car::bootCase	Case bootstrap for regression models
car::car-package	Companion to Applied Regression

- If you know the exact function
- help(function name) OR
- ?functionname
- > ?mean



DATA SCIENCE INSTITUTE

A Quick Recap

- A Quick Recap

- data.frame
- min
- max
- mean
- levels
- edit
- write.csv
- history

- help.search
- help
- read.table
- head
- tail
- dim
- str
- names
- summary

DATA SCIENCE INSTITUTE

THANK YOU!!

DATA SCIENCE
INSTITUTE