R for Bioinformatics

Introduction, Programming, Data Analysis and Visualization

Introduction to Data Analysis and R

Gang Chen
chengang@bgitechsolutions.com

November 30, 2013

Outline

1. Data Analysis
2. Data Aanlysis and R
3. Hello R!
4. Development Environment
5. References

Next

1. Data Analysis
2. Data Aanlysis and R
3. Hello R!
4. Development Environment
5. References

Data Analysis

Wikipedia

Analysis of data is a process of **inspecting**, **cleaning**, **transforming**, and **modeling** data with the goal of discovering useful information, suggesting conclusions, and supporting decision making.

Data Analysis

1. Collecting →
2. Cleaning →
3. Transforming →
4. Modeling →
5. Visualizing →
6. Knowledge

Biological Data Analysis

1. Sequencing →
2. QC →
3. Alignment ...→
4. GWAS, EWAS ...→
5. Manhattan Plot, Q-Q plot ...→
6. Paper?

Next

1. Data Analysis
2. Data Aanlysis and R
3. Hello R!
4. Development Environment
5. References

What is R?

R

R is a **free** software environment for statistical computing and graphics.
----R-project.org

Data Analysis

1. Collecting →
2. Cleaning →
3. Transforming →
4. Modeling →
5. Visualizing →
6. Knowledge

R

- **Rcurl...**
- **gsub, unique...**
- **reshape...**
- **e1071...**
- **ggplot2...**
- **knitr...**

Biological Data Analysis and R

Biological Data Analysis

1. Sequencing →
2. QC →
3. Alignment ...→
4. GWAS, EWAS ...→
5. Manhattan Plot, Q-Q plot ...→
6. Paper?

R

- **Rsamtools, Affy...**
- **genomicRanges...**
- **reshape...**
- **e1071...**
- **ggbio...**
- **knitr, shiny...**

History

- Version 0.16, This is the last alpha version developed primarily by Ihaka and Gentleman. Much of the basic functionality from the "White Book" (see S history) was implemented. The mailing lists commenced on April 1, 1997.
- Version 0.49, April 23, 1997, This is the oldest available source release, and compiles on a limited number of Unix-like platforms. CRAN is started on this date, with 3 mirrors that initially hosted 12 packages. Alpha versions of R for Microsoft Windows and Mac OS are made available shortly after this version.
- Version 0.60, December 5, 1997, R becomes an official part of the GNU Project. The code is hosted and maintained on CVS.

History

- Version 1.0.0, February 29, 2000, Considered by its developers stable enough for production use.[28]
- Version 1.4.0, S4 methods are introduced and the first version for Mac OS X is made available soon after.
- Version 2.0.0, October 4, 2004, Introduced lazy loading, which enables fast loading of data with minimal expense of system memory.
- Version 2.1.0, Support for UTF-8 encoding, and the beginnings of internationalization and localization for different languages.
- Version 2.11.0, April 22, 2010, Support for Windows 64 bit systems.
- Version 2.13.0, April 14, 2011, Adding a new compiler function that allows speeding up functions by converting them to byte-code.

History

- Version 2.14.0, October 31, 2011, Added mandatory namespaces for packages. Added a new parallel package.
- Version 2.15.0, March 30, 2012, New load balancing functions. Improved serialization speed for long vectors.
- Version 3.0.0, April 3, 2013, Support for numeric index values 231 and larger on 64 bit systems.

R in China

- 2004, official documents are translated into Chinese
- 2006, some books on R in Bioinformatics
- 2008, the first R conference was hold at Renming University, Beijing.
- 2009 to 2013, China R Conference is hold at Beijing and Shanghai each year.
- 2012, popular R books are translated into Chinese.
- 2013, R in Action ggplot2 R in a nutshell ...are published in China.
- 2013, CUHK-R course is launched.

Applications of R

Applications

- Statistical analysis
- Data Mining
- Life Science
- Business Intelligence
- Data Visualization
- Social Network
- eCommerce
- Integrated Circuit
- Financial
- Media
- Consoluting
- ...

Attendee of R community

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA
cytema
Energy and Environment Science and Technology Campus
CAMPUS OF INTERNATIONAL EXCELLENCE
DPTO. DE MATEMÁTICAS
EIIAB
Escuela de Ingenieros Industriales de Albacete
escuela superior de ingeniería informática
Facultad de Ciencias Económicas y Empresariales
www.uclm.es/ab/fcee
REVOLUTION ANALYTICS
R Studio
Google
ORACLE
MANGOSOLUTIONS
data analysis that delivers
Spotfire
TIBCO Software
Open Analytics
SOLUCIONES INFORMÁTICAS
Tecon

Pros and Cons

Bo Cowgill, Google

``The best thing about R is that it was developed by statisticians. The worst thing about R is that ... it was developed by statisticians.''

Next

1. Data Analysis
2. Data Aanlysis and R
3. Hello R!
    - Hello R!
    - Hello Statistical Analysis!
    - Hello Plot!
4. Development Environment
5. References

Hello R!

```
print("Hello R!")

## [1] "Hello R!"
```

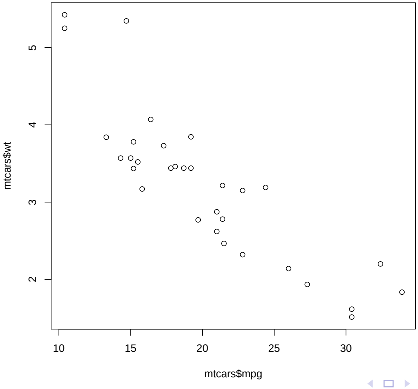
Hello Statistical Analysis

```
data(mtcars)
cor(mtcars$mpg, mtcars$wt)

## [1] -0.8677
```

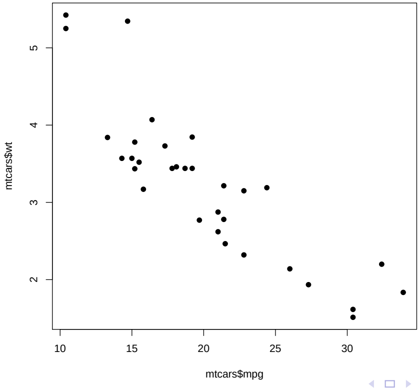
Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt)
```



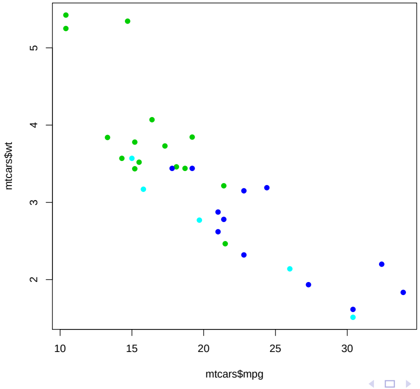
Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19)
```



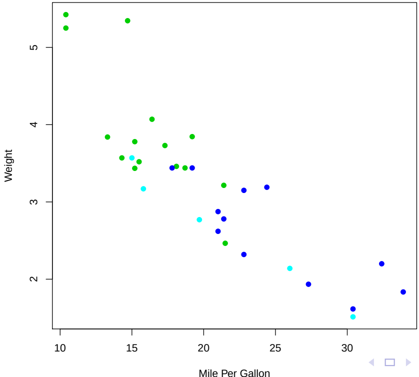
Hello Plot

```r
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear)
```



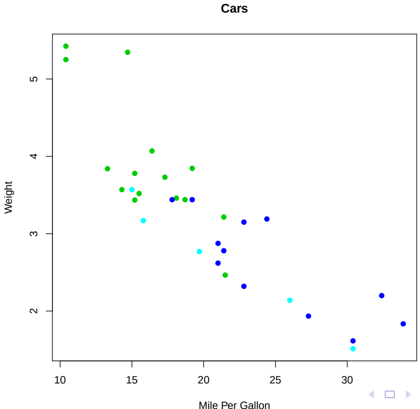
Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mi
     ylab = "Weight")
```



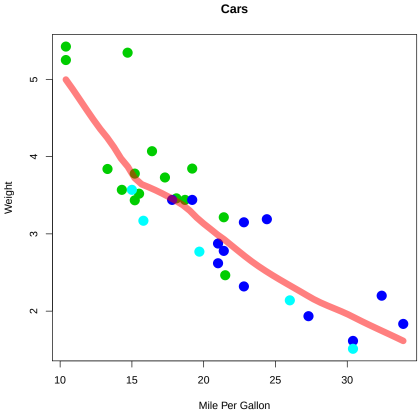
Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mi
    ylab = "Weight", main = "Cars")
```



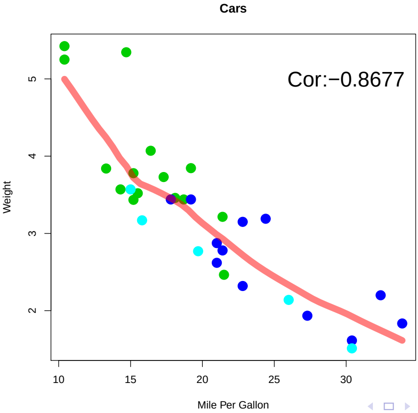
Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mile Per Gallon",
    ylab = "Weight", main = "Cars", cex = 2)
lines(loess.smooth(mtcars$mpg, mtcars$wt), col = rgb(1, 0, 0, 0.5), lwd = 10)
```



Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mile Per Gallon",
    ylab = "Weight", main = "Cars", cex = 2)
lines(loess.smooth(mtcars$mpg, mtcars$wt), col = rgb(1, 0, 0, 0.5), lwd = 10)
text(30, 5, paste("Cor", round(cor(mtcars$mpg, mtcars$wt), 4), sep = ":"), cex = 2)
```



Next

1. Data Analysis
2. Data Aanlysis and R
3. Hello R!
4. Development Environment
   - Obtaining and installing R
   - R in Command Line
   - Editors and IDEs
5. References

Download and Installation

Download

CRAN

Installation

- R: Linux(apt, yum), Mac OS, Windows
- Rtools: Windows
- packages: CRAN, devtools, github, local file

R Commands

R Commands

```
R CMD command args
```

``command'':

- **INSTALL** Install add-on packages
- **REMOVE** Remove add-on packages
- **BATCH** Run R in batch mode

R Command Options

R Command Options

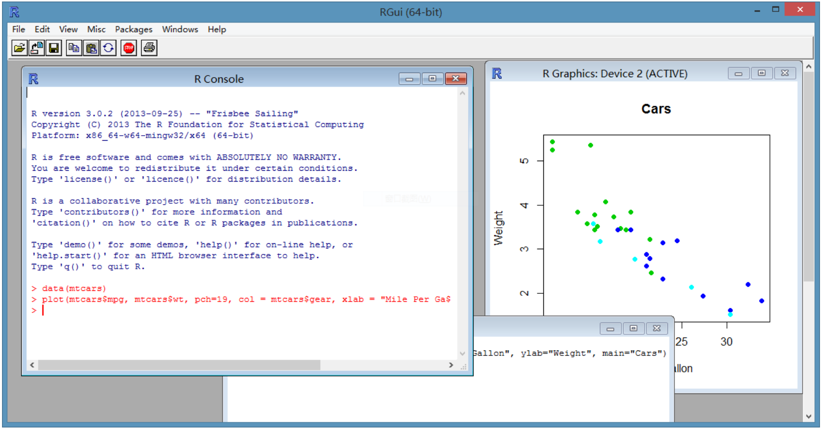
-h, --help	Print usage message and exit
--version	Print version information and exit
--save	Do save workspace at the end of the session
--no-save	Don't save it
--restore	Do restore previously saved objects
--no-restore	Don't restore anything
--vanilla	Combine --no-save, --no-restore, --no-site-file, --no-init-file and --no-environ
-f file, --file=file	Take input from ``file''
-e expression	Use `exression' as input

Editors and IDEs

Editors

- R terminal
- Rgui
- VIM + Vim-R-plugin
- Emacs + ESS
- Notepad++ + NppToR
- ...

R Terminal and Rgui



```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
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.

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.

> data(mtcars)
> plot(mtcars$mpg, mtcars$wt, pch=19, col = mtcars$gear, xlab = "Mile Per Ga$
> 
```

```
allon", ylab="Weight", main="Cars")
```

R Terminal and Rgui

R

- Ctrl + R : run
- Tab: auto complete
- arrow up and down: history

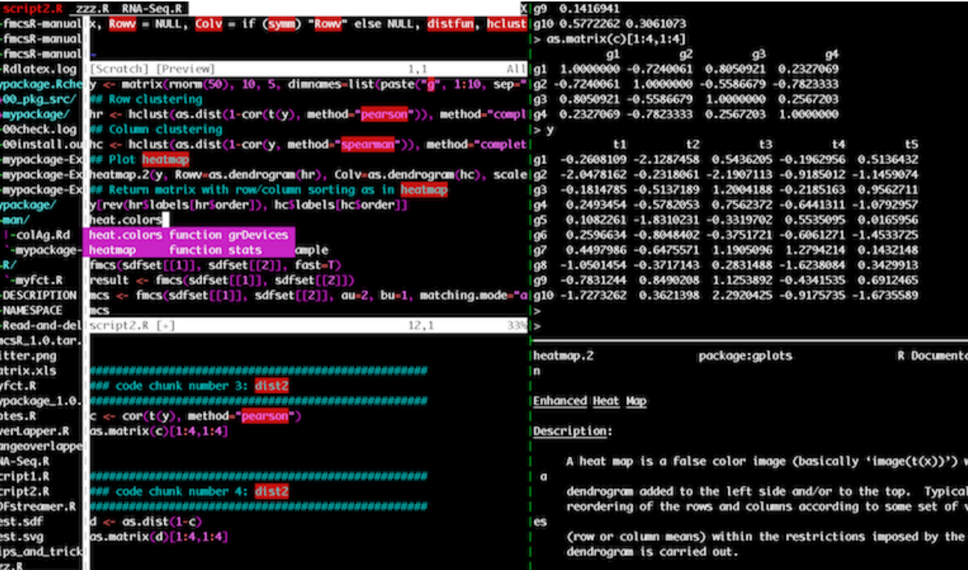
R and Texteditor

- copy and paste
- `source("source.R")`

source

```
sourceDir <- function(path, trace = TRUE, ...) {
    for (nm in list.files(path, pattern = "[.][RrSsQq]$")) {
        if (trace)
            cat(nm, ":")
        source(file.path(path, nm), ...)
        if (trace)
            cat("\n")
    }
}
```

VIM + Vim-R-plugin



```
-fmcsR-manual
-fmcsR-manual
-fmcsR-manual
-Rdlatex.log
ypackage.Rche
-00_pkg_src/
-mypackage/
-00check.log
-00install.ou
-mypackage-Ex
-mypackage-Ex
-mypackage-Ex
ypackage/
-man/
 |-colAg.Rd
 `-mypackage-
-R/
 `-myfct.R
-DESCRIPTION
-NAMESPACE
-Read-and-del
mcsR_1.0.tar.
itter.png
atrix.xls
yfct.R
ypackage_1.0.
otes.R
verLapper.R
angeoverlappe
NA-Seq.R
cript1.R
cript2.R
DFstreamer.R
est.sdf
est.svg
ips_and_trick
zz.R
```

```
script2.R  zzz.R  RNA-Seq.R                                               X
x, Rowv = NULL, Colv = if (symm) "Rowv" else NULL, distfun, hclust

~
[Scratch] [Preview]                                        1,1          All
y <- matrix(rnorm(50), 10, 5, dimnames=list(paste("g", 1:10, sep="
## Row clustering
hr <- hclust(as.dist(1-cor(t(y), method="pearson")), method="compl
## Column clustering
hc <- hclust(as.dist(1-cor(y, method="spearman")), method="complet
## Plot heatmap
heatmap.2(y, Rowv=as.dendrogram(hr), Colv=as.dendrogram(hc), scale
## Return matrix with row/column sorting as in heatmap
y[rev(hr$labels[hr$order]), hc$labels[hc$order]]
heat.colors
heat.colors function grDevices
heatmap     function stats      ample
fmcs(sdfset[[1]], sdfset[[2]], fast=T)
result <- fmcs(sdfset[[1]], sdfset[[2]])
mcs <- fmcs(sdfset[[1]], sdfset[[2]], au=2, bu=1, matching.mode="a
mcs
script2.R [+]                                              12,1         33%


###################################################
### code chunk number 3: dist2
###################################################
c <- cor(t(y), method="pearson")
as.matrix(c)[1:4,1:4]


###################################################
### code chunk number 4: dist2
###################################################
d <- as.dist(1-c)
as.matrix(d)[1:4,1:4]
```

```
g9  0.1416941
g10 0.5772262 0.3061073
> as.matrix(c)[1:4,1:4]
           g1         g2         g3         g4
g1  1.0000000 -0.7240061  0.8050921  0.2327069
g2 -0.7240061  1.0000000 -0.5586679 -0.7823333
g3  0.8050921 -0.5586679  1.0000000  0.2567203
g4  0.2327069 -0.7823333  0.2567203  1.0000000
> y
            t1         t2         t3         t4         t5
g1  -0.2608109 -2.1287458  0.5436205 -0.1962956  0.5136432
g2  -2.0478162 -0.2318061 -2.1907113 -0.9185012 -1.1459074
g3  -0.1814785 -0.5137189  1.2004188 -0.2185163  0.9562711
g4   0.2493454 -0.5782053  0.7562372 -0.6441311 -1.0792957
g5   0.1082261 -1.8310231 -0.3319702  0.5535095  0.0165956
g6   0.2596634 -0.8048402 -0.3751721 -0.6061271 -1.4533725
g7   0.4497986 -0.6475571  1.1905096  1.2794214  0.1432148
g8  -1.0501454 -0.3717143  0.2831488 -1.6238084  0.3429913
g9  -0.7831244  0.8490208  1.1253892 -0.4341535  0.6912465
g10 -1.7273262  0.3621398  2.2920425 -0.9175735 -1.6735589
>
>
```

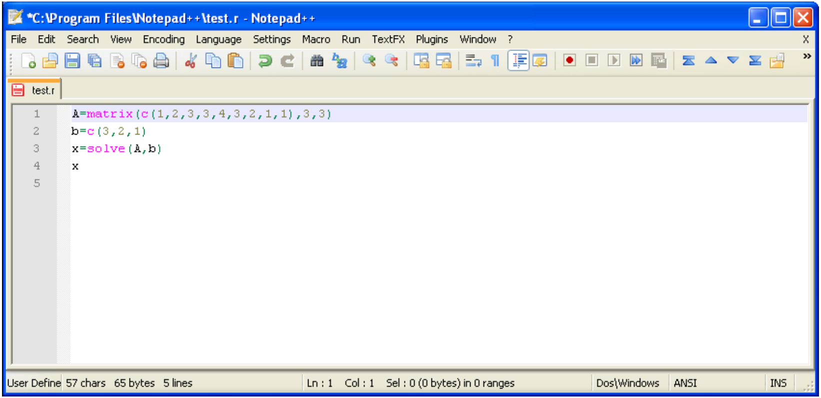
```
heatmap.2                 package:gplots                 R Documenta
n

Enhanced Heat Map

Description:

     A heat map is a false color image (basically 'image(t(x))') w
 a
     dendrogram added to the left side and/or to the top.  Typical
     reordering of the rows and columns according to some set of v
es
     (row or column means) within the restrictions imposed by the
     dendrogram is carried out.
```

Notepad++ + NppToR

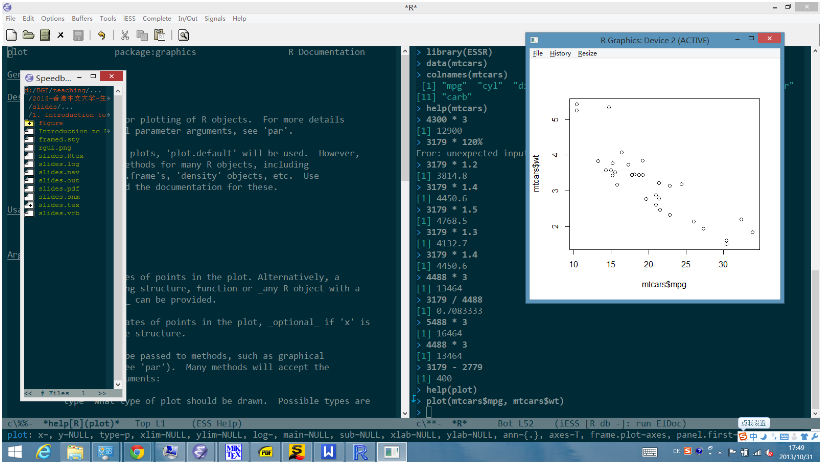


*C:\Program Files\Notepad++\test.r - Notepad++
File Edit Search View Encoding Language Settings Macro Run TextFX Plugins Window ?
test.r

```
A=matrix(c(1,2,3,3,4,3,2,1,1),3,3)
b=c(3,2,1)
x=solve(A,b)
x
```

User Define 57 chars 65 bytes 5 lines Ln : 1 Col : 1 Sel : 0 (0 bytes) in 0 ranges Dos\Windows ANSI INS

Emacs + ESS



```
plot                   package:graphics                   R Documentation

Ger

Des
                      or plotting of R objects.  For more details
                      l parameter arguments, see 'par'.

                       plots, 'plot.default' will be used.  However,
                      ethods for many R objects, including
                      .frame's, 'density' objects, etc.  Use
                      d the documentation for these.

Usa


Arg


                      es of points in the plot. Alternatively, a
                      ng structure, function or _any R object with a
                      _ can be provided.

                      ates of points in the plot, _optional_ if 'x' is
                      e structure.

                      be passed to methods, such as graphical
                      ee 'par').  Many methods will accept the
                      uments:

     type   what type of plot should be drawn.  Possible types are
```

```
d:/BGI/teaching/...
 /2013-香港中文大学-生>
 /slides/...
 /1. Introduction to>
 figure
 Introduction to I>
 framed.sty
 rgui.png
 slides.Rtex
 slides.log
 slides.nav
 slides.out
 slides.pdf
 slides.snm
 slides.tex
 slides.vrb
<<  # Files   1   >>
```

```
> library(ESSR)
> data(mtcars)
> colnames(mtcars)
 [1] "mpg"  "cyl"  "di                                       r"
[11] "carb"
> help(mtcars)
> 4300 * 3
[1] 12900
> 3179 * 120%
Eror: unexpected input
> 3179 * 1.2
[1] 3814.8
> 3179 * 1.4
[1] 4450.6
> 3179 * 1.5
[1] 4768.5
> 3179 * 1.3
[1] 4132.7
> 3179 * 1.4
[1] 4450.6
> 4488 * 3
[1] 13464
> 3179 / 4488
[1] 0.7083333
> 5488 * 3
[1] 16464
> 4488 * 3
[1] 13464
> 3179 - 2779
[1] 400
> help(plot)
> plot(mtcars$mpg, mtcars$wt)
> 
```

```
c\%%-  *help[R](plot)*   Top L1     (ESS Help)          c\**-  *R*      Bot L52    (iESS [R db -]: run ElDoc)
plot: x=, y=NULL, type=p, xlim=NULL, ylim=NULL, log=, main=NULL, sub=NULL, xlab=NULL, ylab=NULL, ann={.}, axes=T, frame.plot=axes, panel.first=
```

Emacs + ESS

What is ESS?

Emacs Speaks Statistics (ESS) is an add-on package for emacs text editors such as **GNU Emacs** and XEmacs. It is designed to support editing of scripts and interaction with various statistical analysis programs such as **R**, S-Plus, SAS, Stata and JAGS.

ESS Website

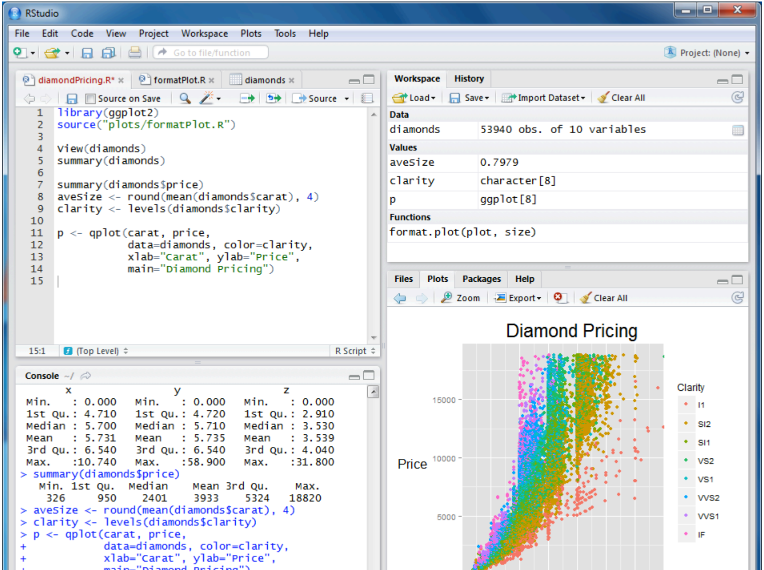
http://ess.r-project.org/

IDEs

IDEs

- RStudio: local and cloud-based
- TinnR
- StatET: eclipse for R
- ...

RStudio



```
library(ggplot2)
source("plots/formatPlot.R")

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")
```

Data	
diamonds	53940 obs. of 10 variables
Values	
aveSize	0.7979
clarity	character[8]
p	ggplot[8]
Functions	
format.plot(plot, size)	

```
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
```

Next

1. Data Analysis
2. Data Aanlysis and R
3. Hello R!
4. Development Environment
5. References

Books

- R in action (also in Chinese)
- Introduction to R (also in Chinese)
- R for beginner (also in Chinese)
- R in a Nutshell (Chinese version is in press)
- The art of R programming (also in Chinese)
- ggplot2. Elegant Graphics for Data Analysis (also in Chinese)

Websites

- R-project and CRAN
- COS.name (Chinese)
- Quick-R
- http://had.co.nz/, Hadley Wickham
- Twitter, github, RForge
- Google

Websites

- R-project and CRAN
- COS.name (Chinese)
- Quick-R
- http://had.co.nz/, Hadley Wickham
- Twitter, github, RForge
- Google Baidu?

Journals

- The R Journal
- Journal of Statistical Software
- BMC Bioinformatics: Software
- Bioinformatics: Application Note