R Programming Language

Josip Šaban

30.11.2012

Outline

- What is analytics?
- What is R and how people use it?
  - Advantages and limitations
  - Support and resources
- Ways to run R
  - Enterprise versions
- Demo

What is analytics?

What is analytics?

- Statistics +
- Machine Learning +
- Artificial Intelligence…
- Used to explain or predict information
- Business analytics software: $10.5 billion in 2010
- + Service + Hardware = $50 billion
- Growing 13.5%/year

Business inteligence

Gartner BI Marke Share Data - 2010

2009 Rank	2010 Rank	Chg	BI Vendors	2008	2009	2010	Share 2009	Share 2010	Growth 2009	Growth 2010
1	1		SAP	2,105	2,066	2,413	22%	23%	-1.8%	16.8%
2	2		Oracle	1,285	1,350	1,646	15%	16%	5.1%	21.9%
3	3		SAS Institute	1,287	1,325	1,387	14%	13%	3.0%	4.7%
4	4		IBM	997	1,136	1,222	12%	12%	14.0%	7.6%
5	5		Microsoft	681	739	914	8%	9%	8.5%	23.6%
6	6		MicroStrategy	280	295	338	3%	3%	5.4%	14.4%
7	7		Fico	302	277	288	3%	3%	-8.3%	4.1%
9	8	+1	Qliktech	104	141	205	2%	2%	36.0%	45.2%
10	9	+1	Infor Global Solutions	147	139	151	2%	1%	-5.3%	8.4%
8	10	-2	Information Builders	185	156	147	2%	1%	-15.9%	-6.0%
11	11		Actuate	117	113	115	1%	1%	-3.5%	1.8%
13	12	+1	TIBCO	65	65	80	1%	1%	0.2%	22.7%
12	13	-1	Minitab	76	72	67	1%	1%	-5.0%	-7.2%
14	14		Accelrys	47	48	49	1%	0%	2.3%	3.1%
23	15	+8	Tableau	13	18	38	0%	0%	36.9%	113.5%
			Other Vendors	1,249	1,338	1,463	14%	14%	7.1%	9.4%
			BI Total	**8,939**	**9,278**	**10,522**	**100%**	**100%**	**3.8%**	**13.4%**

What is R and how people use it?

What is R?

- Open source statistical language
  - And no, it is not about the money!
- De facto standard for statistical research
- Grew out of Bell Labs' S (1976, 1988)
- Licensed by *AT&T/Lucent* to *Insightful Corp*. Product name: *S-plus*.

What is R?

- Language + package + environment for graphics and data analysis
- Free and open source
- Created by Ross Ihaka & Robert Gentleman 1996 & extended by many more
- An implementation of the S language by John Chambers and others
- R has roughly (year 2012) 5000 add-ons and about 100,000 functions

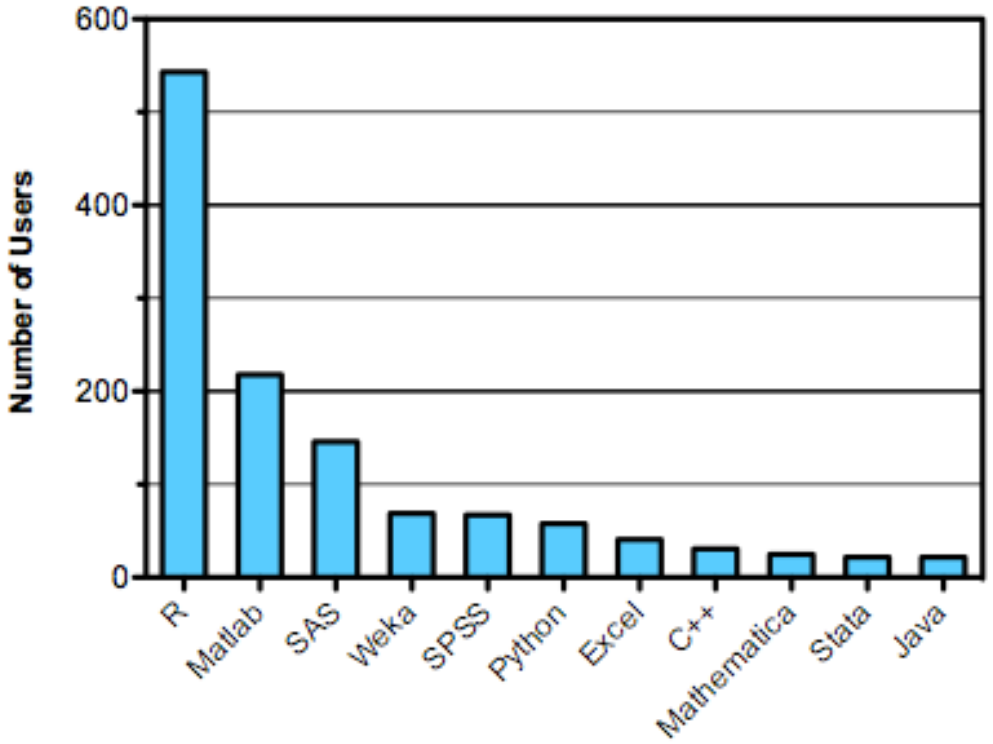
“Open source”... that just means I don’t have to pay for it, right?

- „Free” is NEVER an advantage
  - Provides full access to algorithms and their implementation
  - Gives you the ability to fix bugs and extend software
  - Provides a forum allowing researchers to explore and expand the methods used to analyze data
  - Is the product of 1000s of leading experts in the fields they know best - it is CUTTING EDGE

“Open source”... that just means I don’t have to pay for it, right?

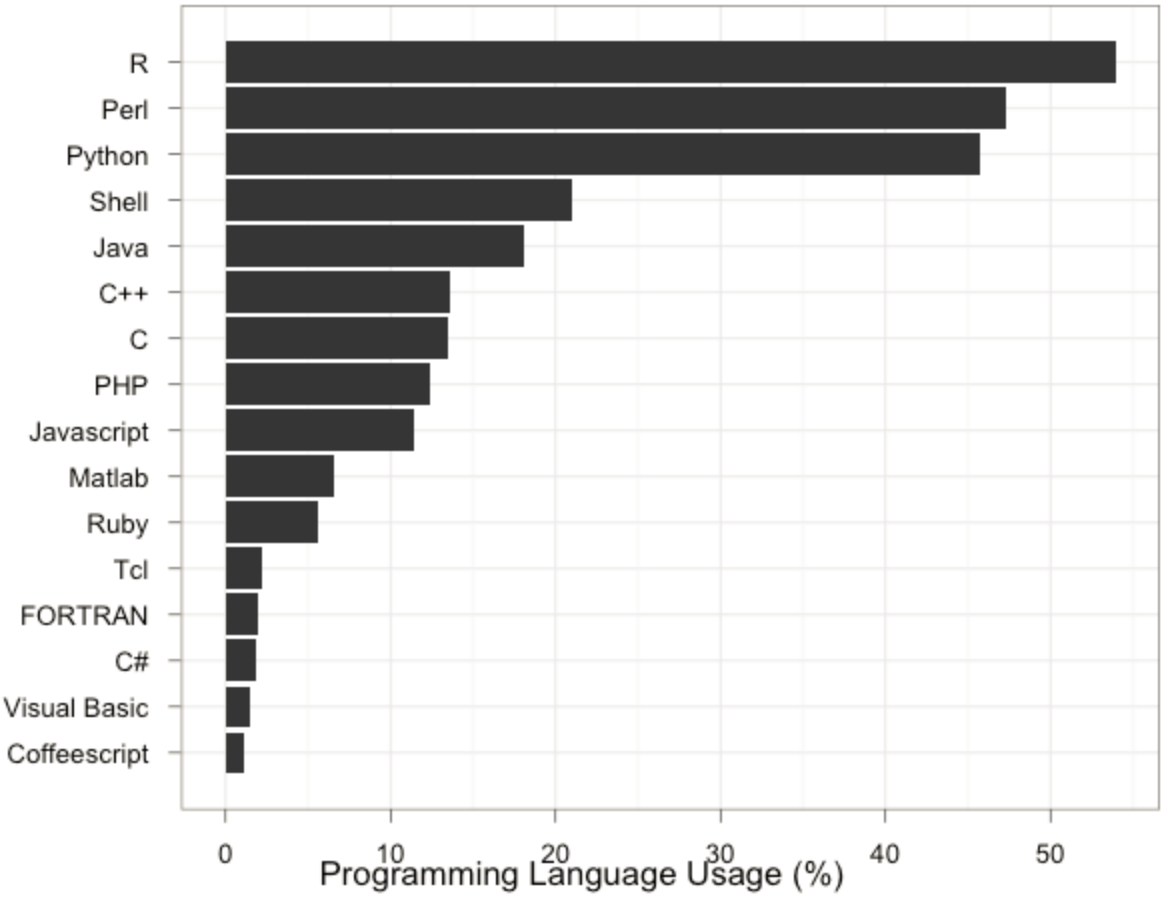
- Ensures that scientists around the world - and not just ones in rich countries - are the co-owners to the software tools needed to carry out research
- Promotes reproducible research by providing open and accessible tools
- Most of R is written in… R! This makes it quite easy to see what functions are actually doing

R in data analysis



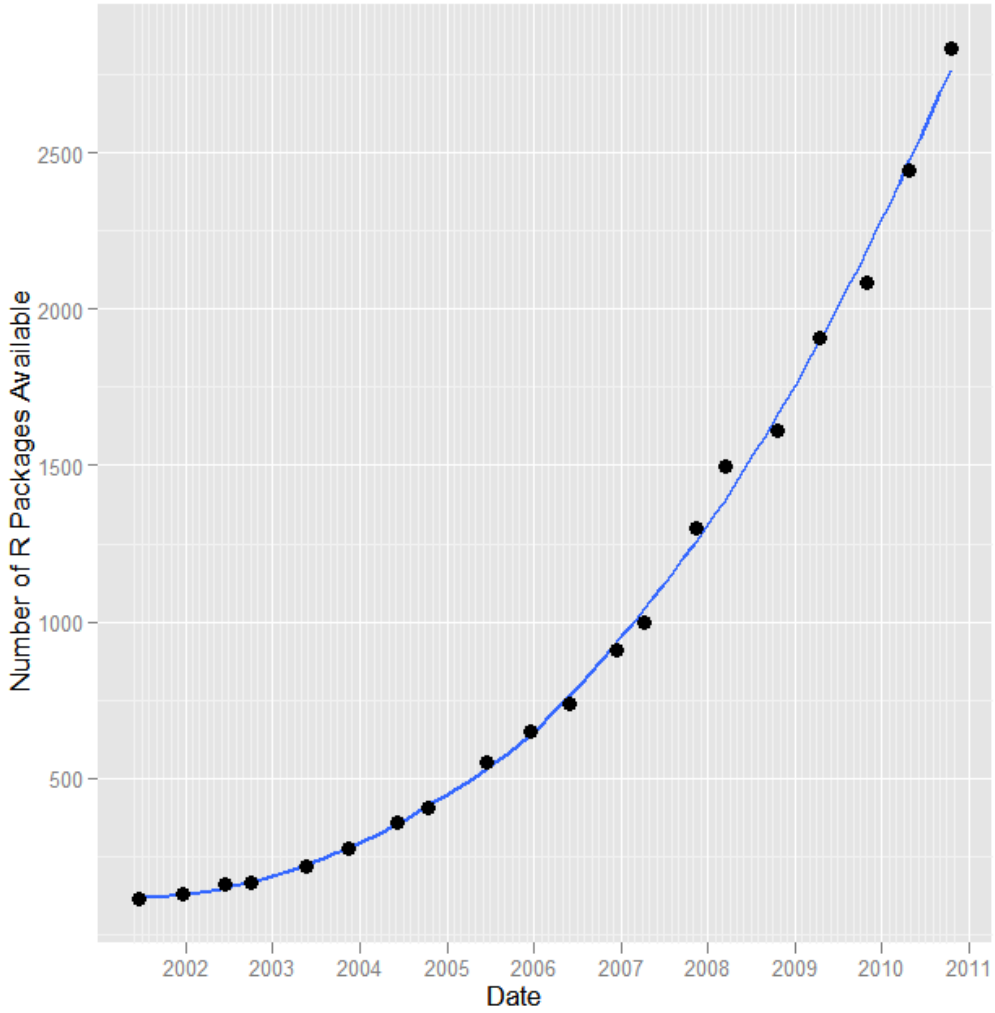
- Languages used in Kaggle.com data analysis competition 2011
  Source: http://r4stats.com/popularity

R in bioinformatics (2012)



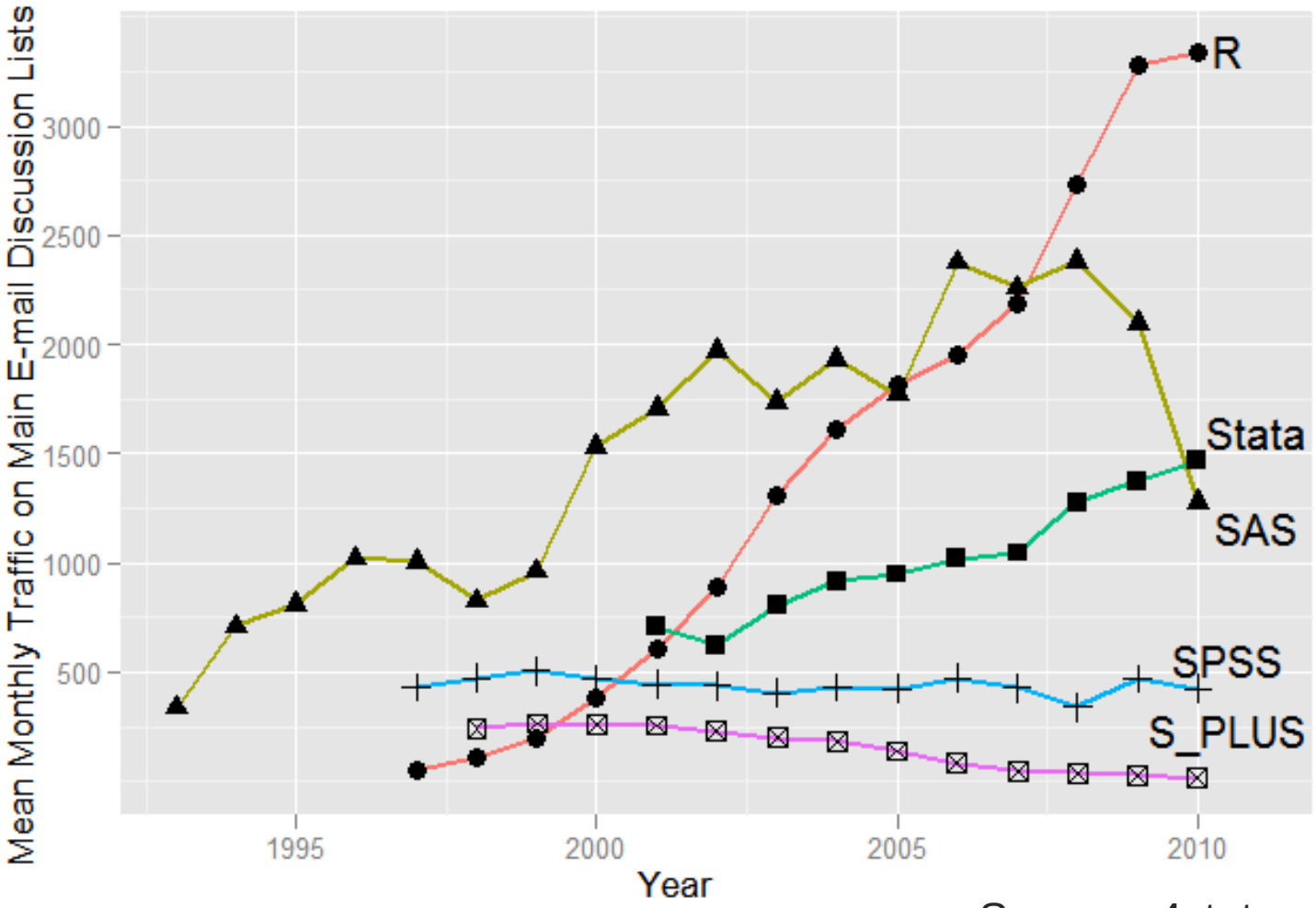
- http://bioinfsurvey.org/analysis/programming_languages/

Growth in R addon packages



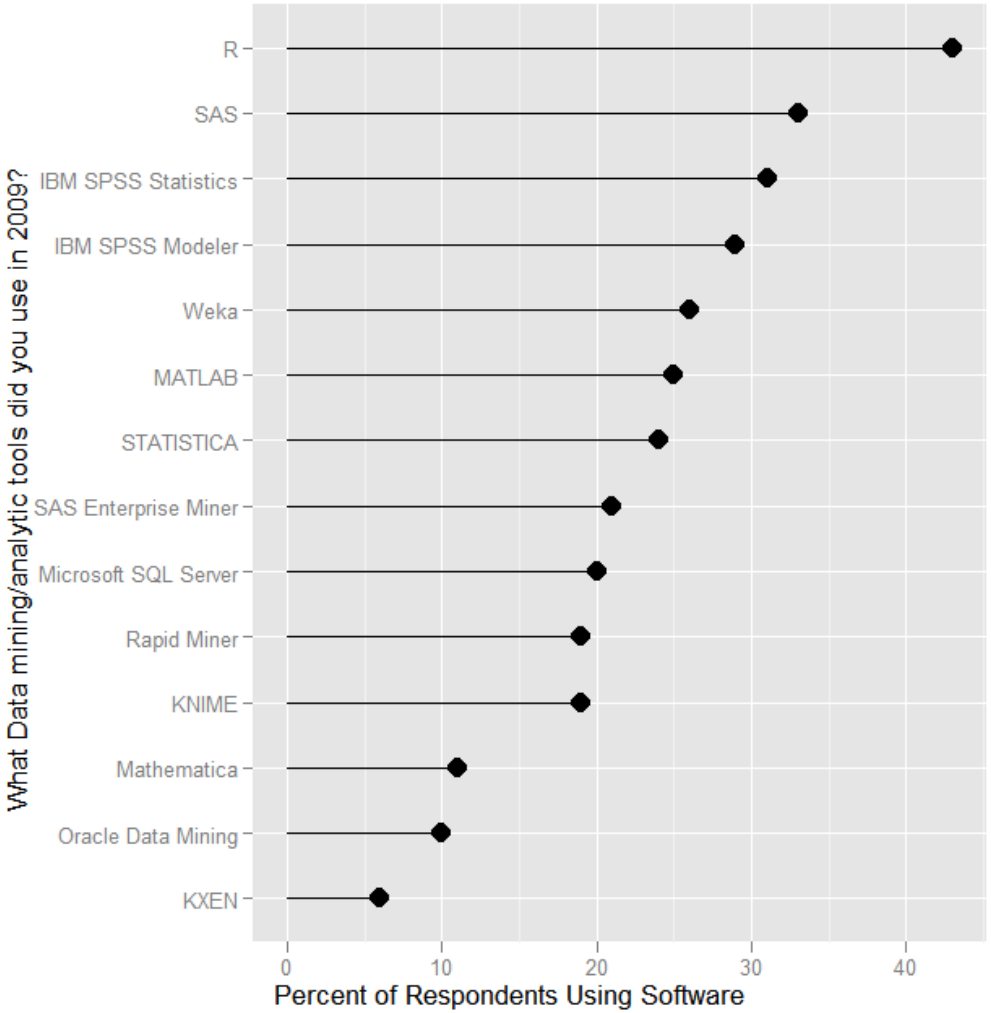
Source: r4stats.com/popularity

Growth in Internet Discussion



Source: r4stats.com/popularity

Rexer Analytics Poll on “tools”

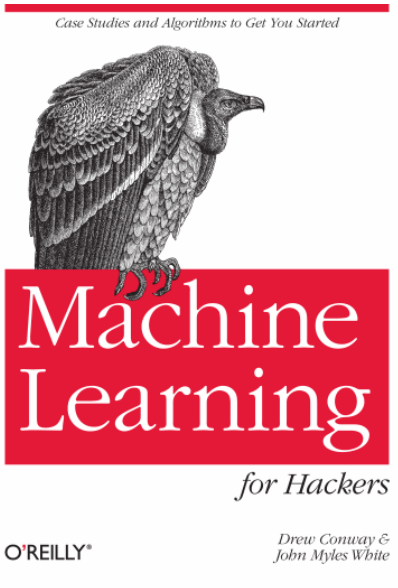


KDnuggests.com Poll on “languages”

R (257)	45%
SQL (184)	32%
Python (140)	25%
Java (139)	24%
SAS (121)	21%
MATLAB (83)	15%
C/C++ (73)	13%
Unix shell/awk/gawk/sed (59)	10%
Perl (45)	7.9%
Hadoop/Pig/Hive (35)	6.1%
Lisp (4)	0.7%
Other (70)	12.0%
None (7)	1.2%

Drew Conway/John Myles White

- "… R has a unique and somewhat prickly syntax and tends to have a steeper learning curve than other languages."



What about speed?

- Maybe 100x slower than C++, though it varies greatly.

What about tool support?

- Limited compared to, for example, first release of Visual Studio (1995).

Domain-specific language (DSL)

- To understand a DSL, start with D, not L.
- The alternative to R isn't Python or C#, it's SAS.
- People love their DSL, and will use it outside of its domain.

- "The best thing about R is that it was written by statisticians. The worst thing about R ..."
  Bo Cowgill, Google

What are statisticians like?

- Different priorities than software developers
- Different priorities than mathematicians
- Learn bits of R in parallel with statistics

What is a statistical DSL?

- Statistical functions easily accessible
- Convenient manipulation of tables
- Vector operations
- Smooth handling of missing data
- Patterns for common tasks

Advantages of R

- A powerful programming language
- Designed for **interactive** data analysis
- Easier to program than, e.g., SAS
- Open source, interpreted, portable
- Succinct notation for querying and filtering
- Succinct notation for linear regression
- Commands are written in that language

Advantages of R

- Commands are visible and changeable
- Commands you write that are on equal footing
- Output that easily becomes input
- Models easily applied to new data
- Legions of developers, thousands of add-ons
- Internet archives make add-ons easy to find

R’s Limitations

- Must find R and its add-ons yourself
- Documentation is sparse & complex
- Graphical user interfaces not as polished
- Language is somewhat harder to learn
- Most R functions hold data in main memory

What about accuracy?

- *Base R plus Recommended Packages* like:
  - Base SAS, SAS/STAT, SAS/GRAPH, SAS/IML Studio
  - SPSS Stat. Base, SPSS Stat. Advanced, Regression
- Tested via extensive validation programs
- But add-on packages written by…
  - Professor who invented the method?
  - A student interpreting the method?

What about tech-support?

- Email support is free, quick, 24-hours: https://stat.ethz.ch/mailman/listinfo/r-help
- You may get “too much” help
- Phone support available commercially e.g. Revolution Analytics
- Use http://www.rseek.org/ instead of Google

Add-on packages

- Alphabetical list at: http://cran.r-project.org
- Use any search engine: “neural network” + “R package”
- Crantastic.org

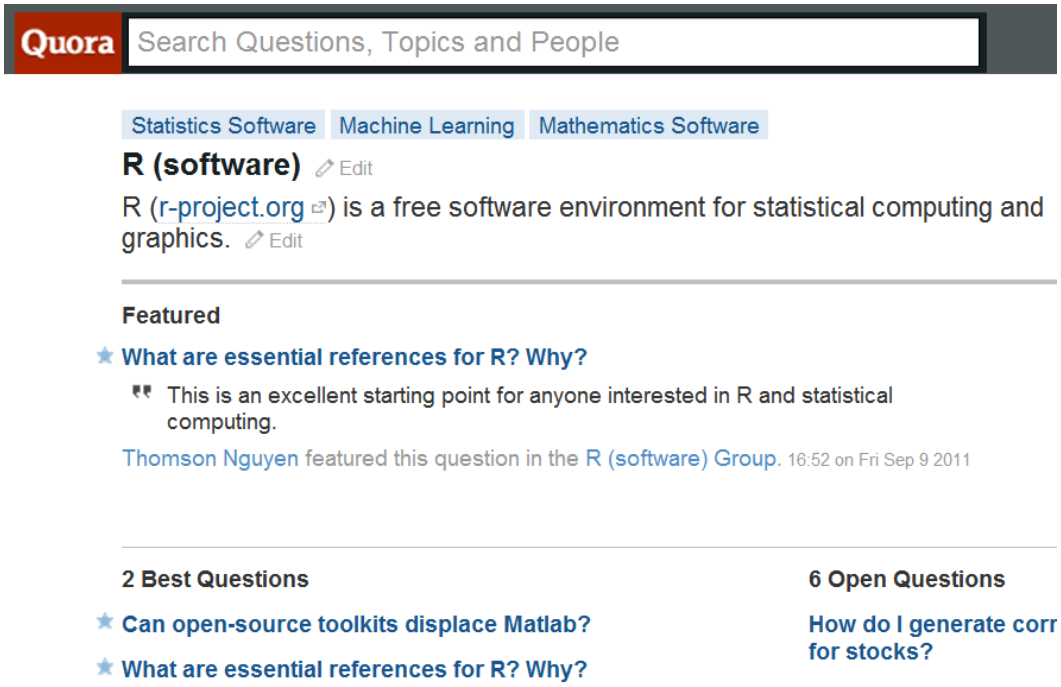
Add-on packages

- http://r4stats.com/add-ons

Topic	SAS Product	SPSS Product	R Package(::Function)
Advanced Models	SAS/STAT	IBM SPSS Advanced Statistics	R, MASS, many others
Association Analysis	Enterprise Miner	IBM SPSS Association	arules, arulesNBMiner, arulesSequences
Basics	Base SAS	IBM SPSS Statistics Base	R
Bootstrapping	SAS/STAT	IBM SPSS Bootstrapping	BootCL, BootPR, boot, bootRes, BootStepAIC, bootspecdens, bootstrap, FRB, gPdtest, meboot, multtest, pvclust, rqmcmb2, scaleboot, simpleboot
Classification Analysis	Enterprise Miner	IBM SPSS Classification	rattle, see the neural networks and trees entries in this table.
Conjoint Analysis	SAS/STAT: PROC TRANSREG	IBM SPSS Conjoint	homals, psychoR, bayesm
Correspondence Analysis	SAS/STAT: PROC CORRESP	IBM SPSS Categories	ade4, cocorresp, FactoMineR, homals, made4, MASS, psychoR, PTAk, vegan
Custom Tables	Base SAS, PROC REPORT, PROC SQL, PROC TABULATE, Enterprise Reporter	IBM SPSS Custom Tables	aggregate, Epi::stat.table, reshape, rreport, tapply, xtable
Data Access	SAS/ACCESS	SPSS Data Access Pack	DBI, foreign, gdata::read.xls, Hmisc::sas.get, sasxport.get, RODBC, sas7bdat, WriteXLS, xlsReadWrite
Data Collection	SAS/FSP	IBM SPSS Data Collection Family	RSQLite, and the other open source programs MySQL or PostgreSQL are popular among R users for this purpose.
Data Mining	Enterprise Miner	IBM SPSS Modeler (formerly Clementine)	arules, FactoMineR, Rattle, Red-R, RWeka link to Weka, various functions
Data Mining, In-database Processing	SAS In-Database Initiative with Teradata	IBM SPSS Modeler	PL/R for PostgreSQL, RODM for Oracle

Relevant sites

- quora.com/R-software
- stackoverflow.com/questions/tagged/r

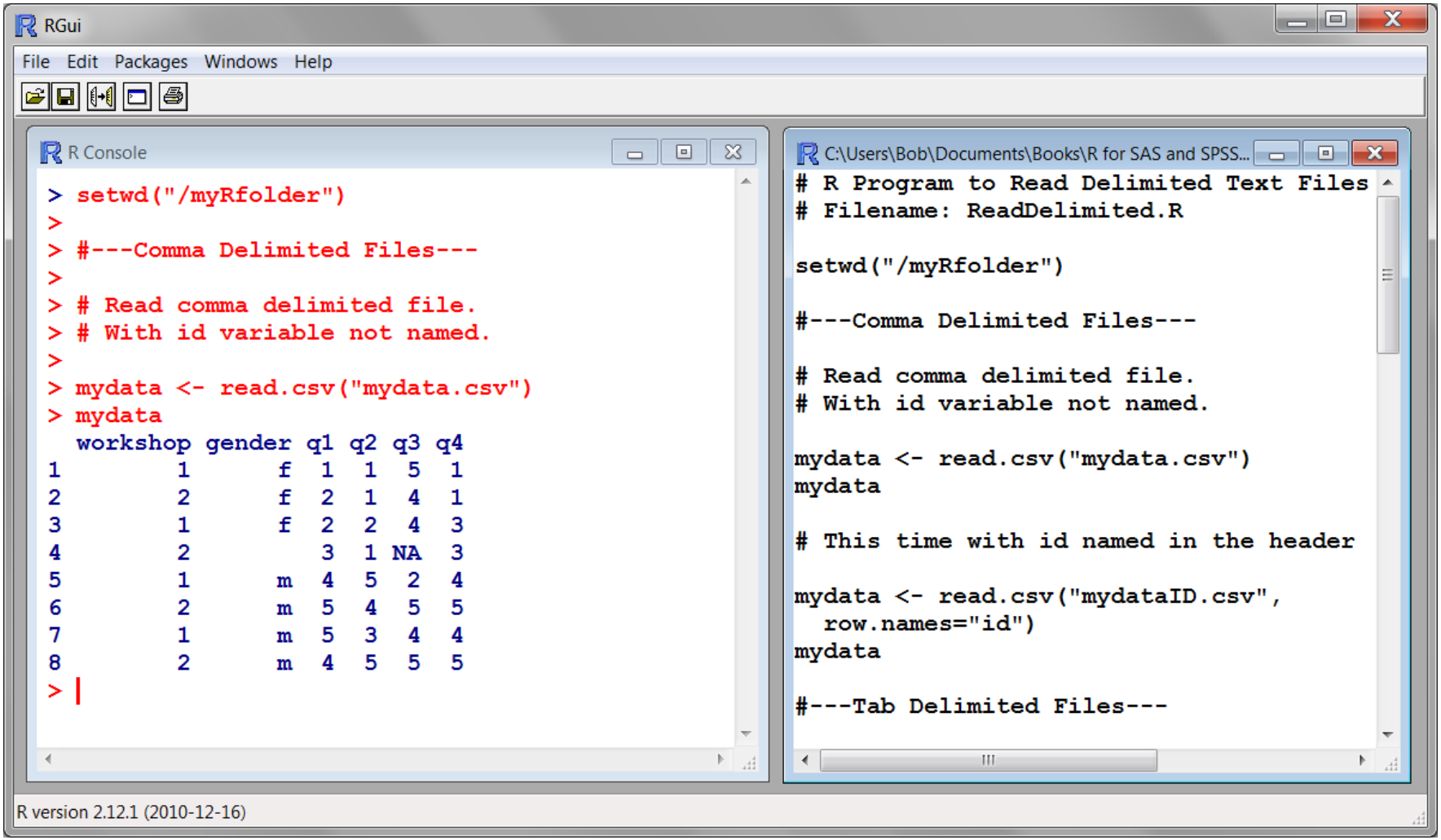


Relevant sites

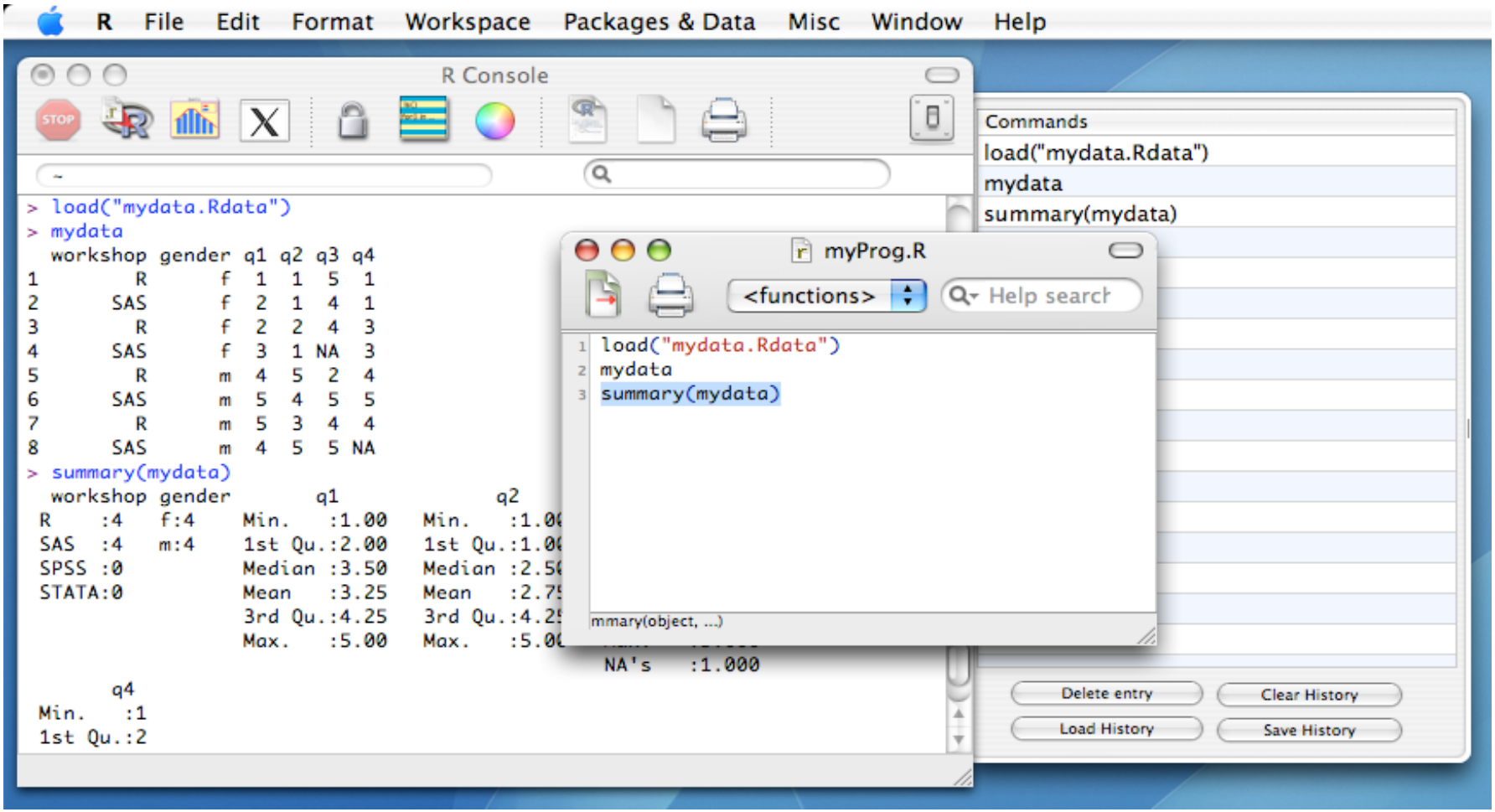
- R wiki:
  - http://rwiki.sciviews.org/doku.php
- R graph gallery:
  - http://addictedtor.free.fr/graphiques/thumbs.php
- Kickstarting R:
  - http://cran.r-project.org/doc/contrib/Lemon-kickstart/

Ways to run R

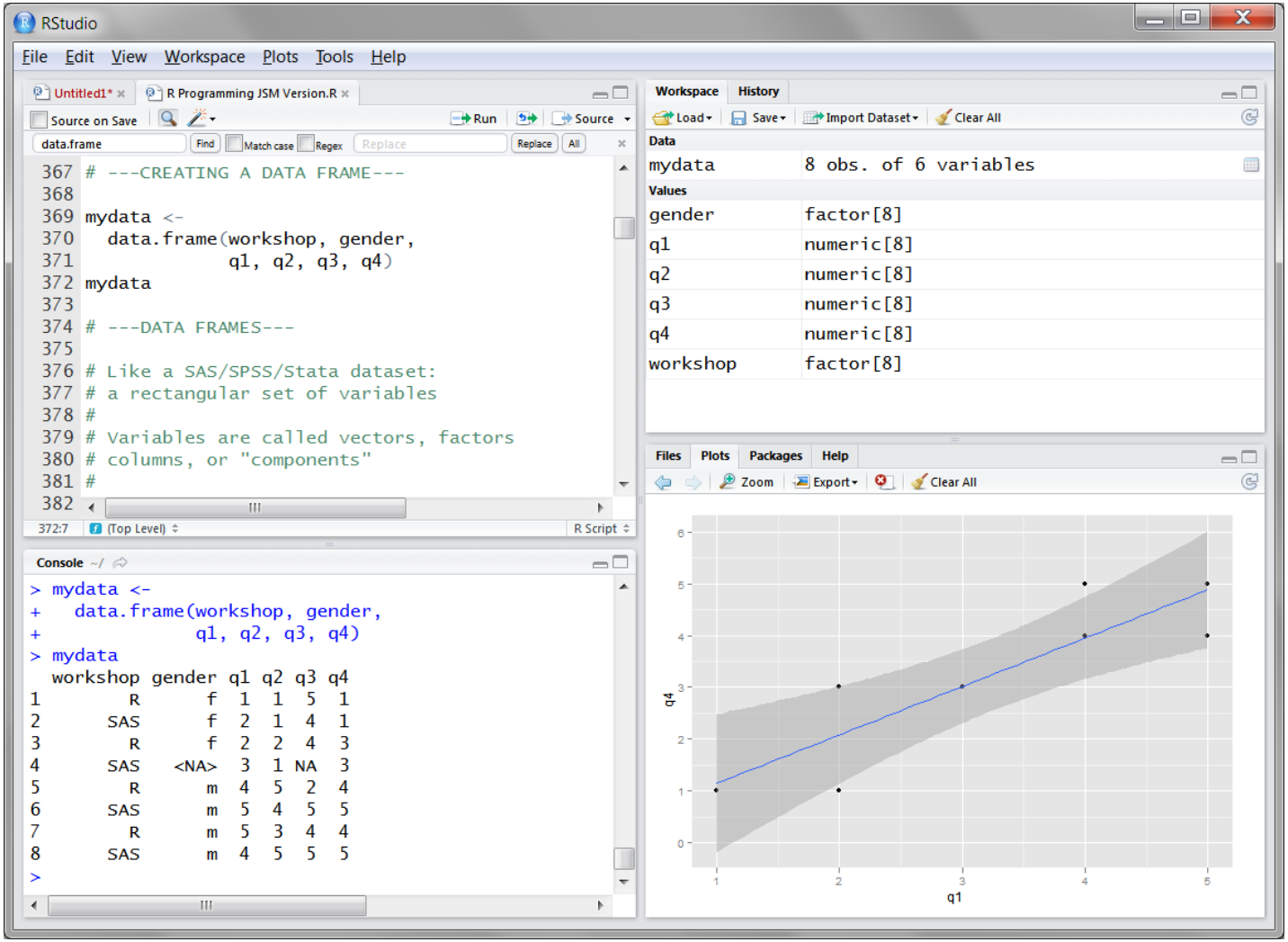
Standard Windows Interface



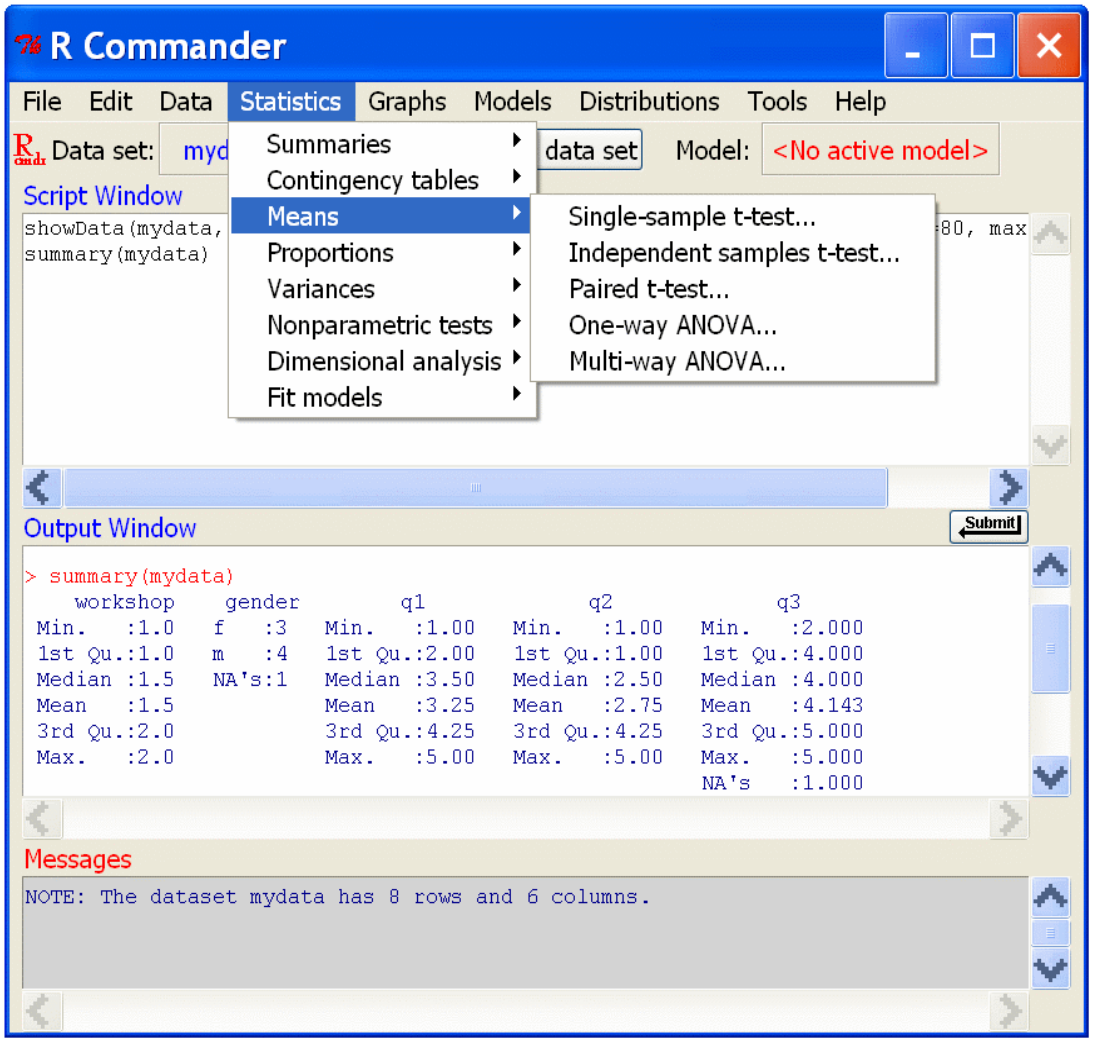
Standard MacIntosh interface



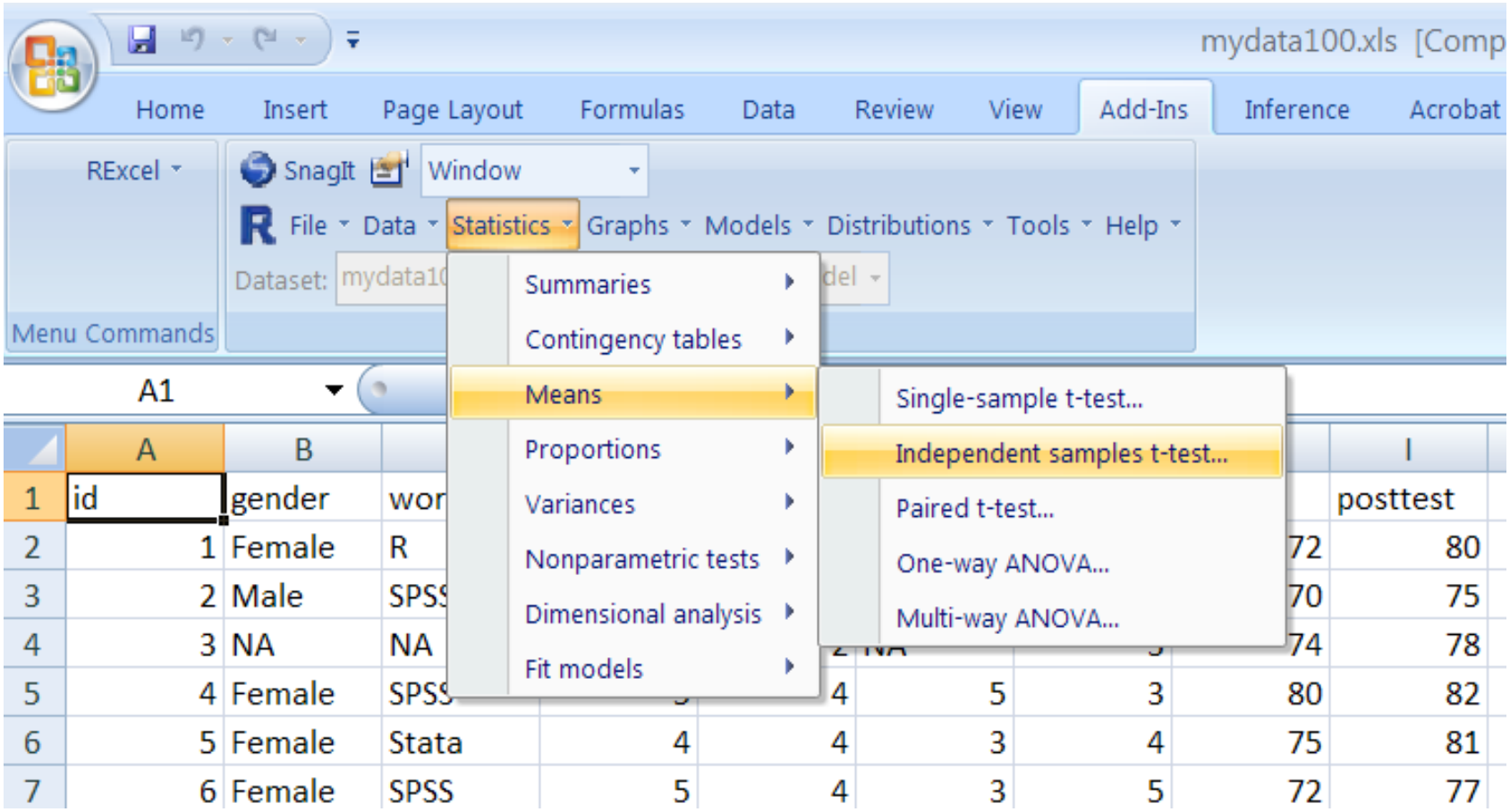
RStudio (http://RStudio.org)



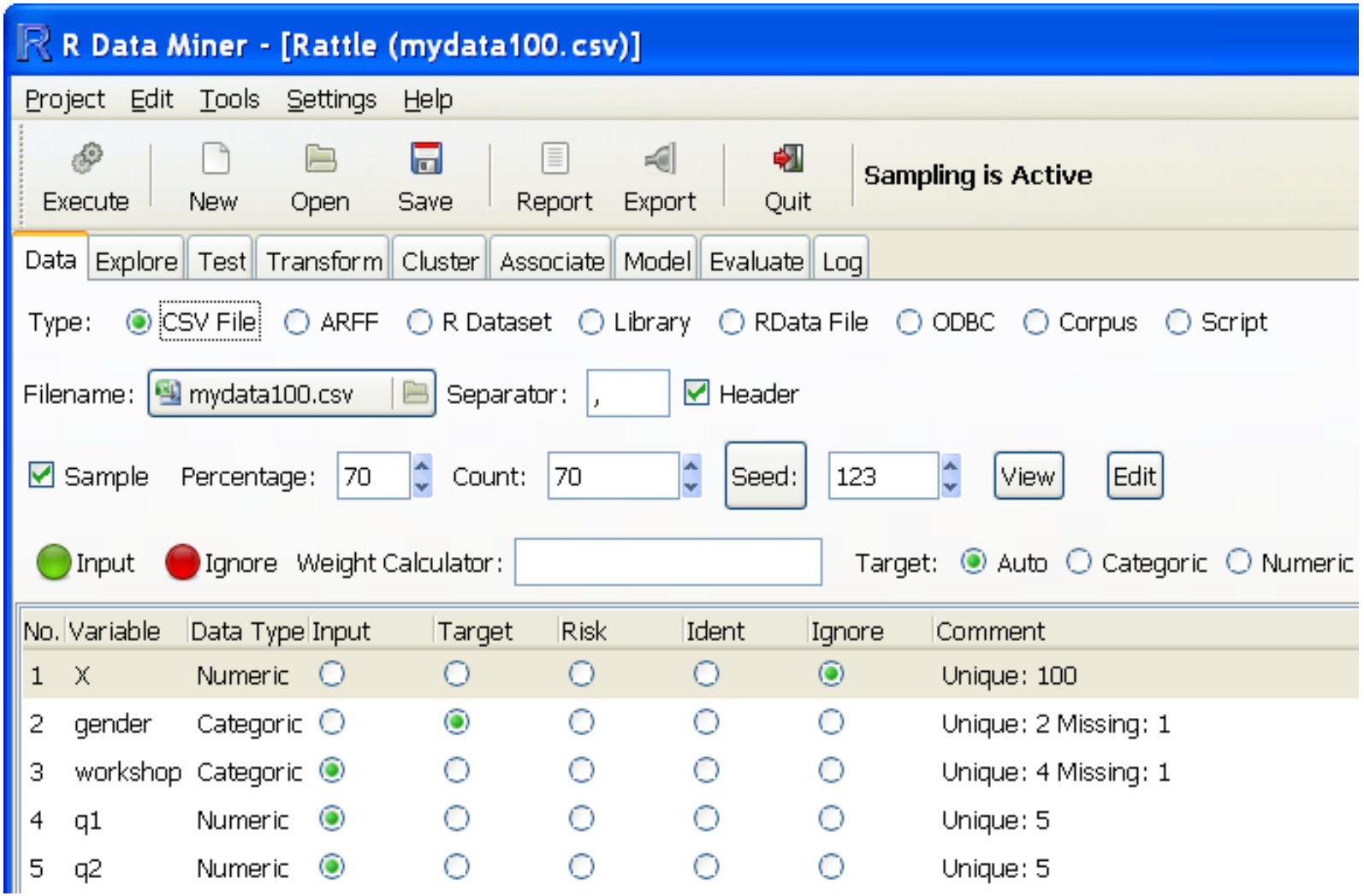
R Commander


R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set: myd
data set
Model: <No active model>
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Single-sample t-test...
Independent samples t-test...
Paired t-test...
One-way ANOVA...
Multi-way ANOVA...
Script Window
showData(mydata,
summary(mydata)
Output Window
Submit
> summary(mydata)
workshop gender q1 q2 q3
Min. :1.0 f :3 Min. :1.00 Min. :1.00 Min. :2.000
1st Qu.:1.0 m :4 1st Qu.:2.00 1st Qu.:1.00 1st Qu.:4.000
Median :1.5 NA's:1 Median :3.50 Median :2.50 Median :4.000
Mean :1.5 Mean :3.25 Mean :2.75 Mean :4.143
3rd Qu.:2.0 3rd Qu.:4.25 3rd Qu.:4.25 3rd Qu.:5.000
Max. :2.0 Max. :5.00 Max. :5.00 Max. :5.000
NA's :1.000
Messages
NOTE: The dataset mydata has 8 rows and 6 columns.

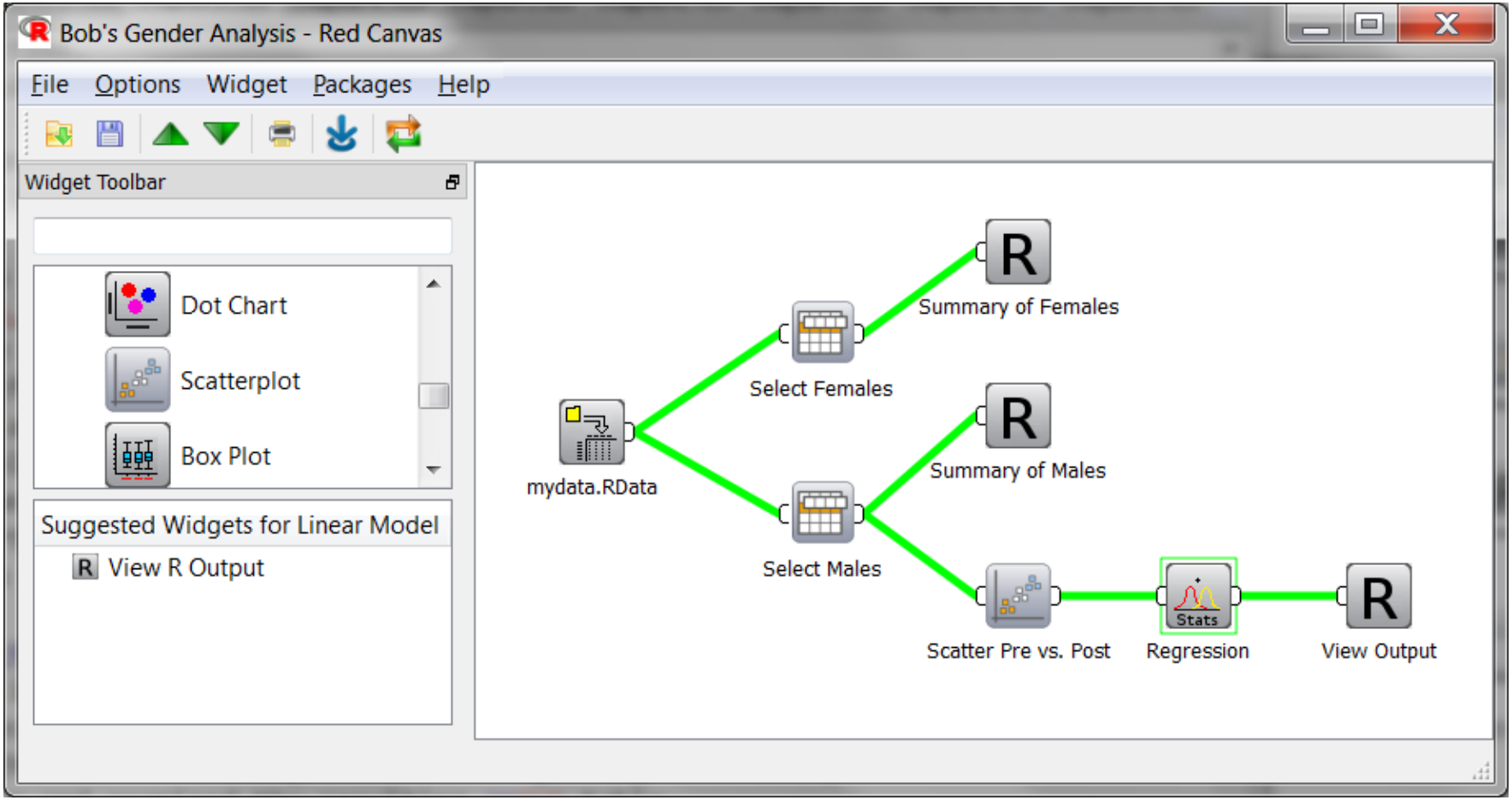
Running R from Excel



Rattle: R Analytical Tool To Learn Easily



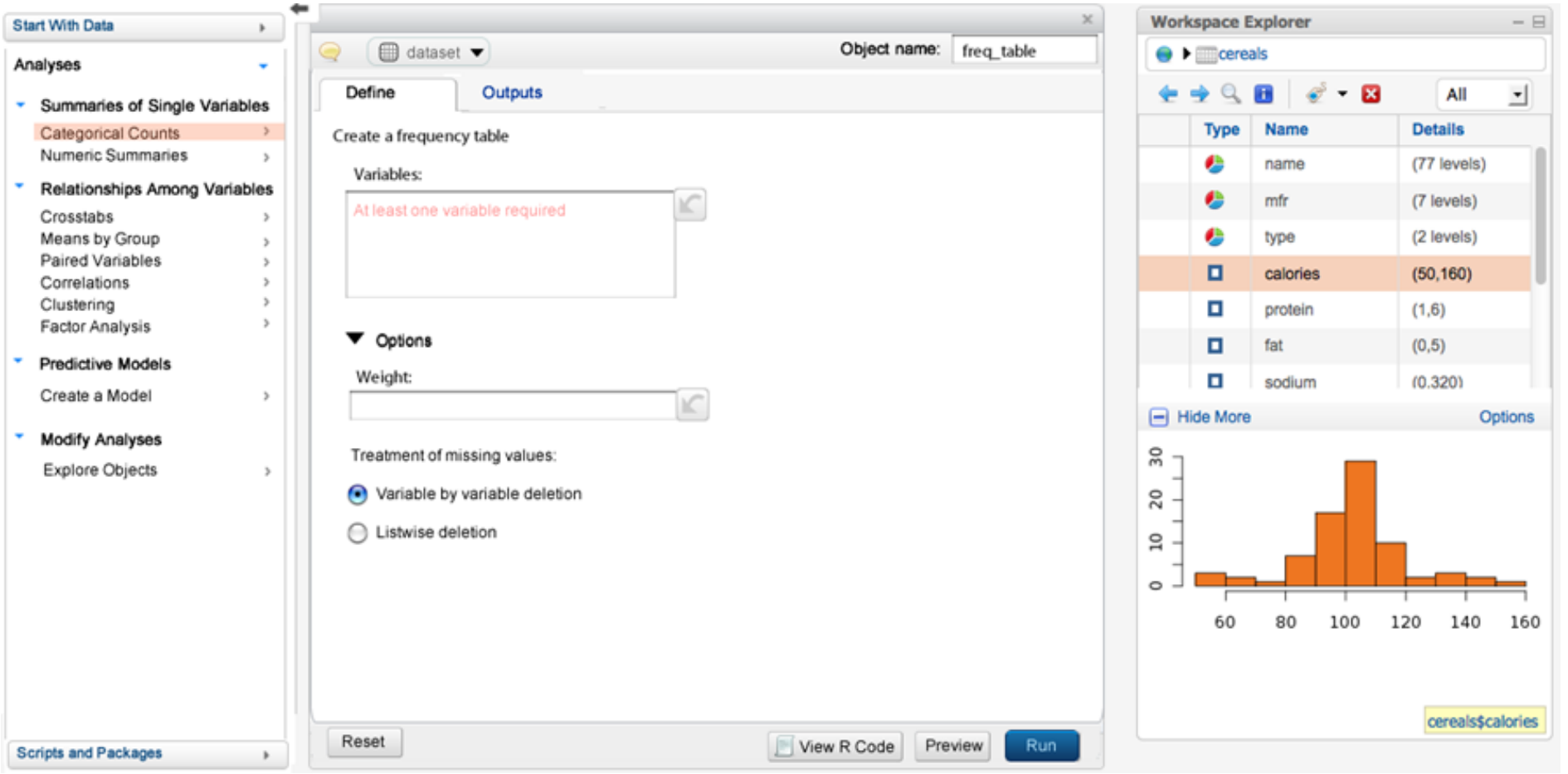
Red-R (http://www.red-r.org/)



Revolution R Enterprise

- From Revolution Analytics
- Company run by SPSS founders Norman Nie and “Tex” Hull
- Recompiled for speed using optimized compilers
- Supports multi-core processors
- ParallelR for clusters
- Includes R Productivity Environment

Revolution Analytics User Interface



Code comparison

Task	R	SAS	SPSS

Analysis of Variance	`myModel <- aov(posttest ~ workshop, data = mydata100)`
	`summary(myModel)`
	`pairwise.t.test(posttest, workshop)`
	`TukeyHSD(myModel, "workshop")`
	`plot(TukeyHSD(myModel, "workshop"))`
		`PROC GLM;`
		`CLASS workshop;`
		`MODEL posttest = workshop;`
		`MEANS workshop / TUKEY;`
			`UNIANOVA posttest BY workshop`
			`/POSTHOC = workshop ( TUKEY )`
			`/PRINT = ETASQ HOMOGENEITY`
			`/DESIGN = workshop.`

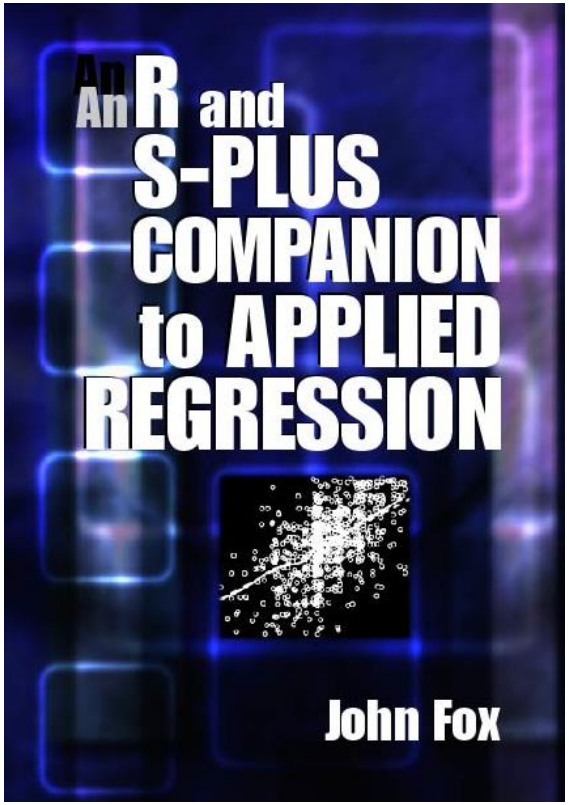
Correlate, Pearson	`cor( mydata[3:6], method = "pearson", use = "pairwise")`
	`cor.test(mydata$q1, mydata$q2, use = "pairwise")`
	
	`library("Rcmdr")`
	`rcorr.adjust( mydata[3:6] )`
		`PROC CORR; VAR q1-q4;`
		`RUN;`
			`CORRELATIONS`
			`/VARIABLES=q1 TO q4.`

Correlate, Spearman	`cor( mydata[3:6], method = "spearman", use = "pairwise")`
	`cor.test(mydata$q1, mydata$q2, use = "pairwise")`
	
	`library("Rcmdr")`
	`rcorr.adjust( mydata[3:6] )`
		`PROC CORR SPEARMAN;`
		`VAR q1-q4;`
			`NONPAR CORR`
			`/VARIABLES=q1 to q4`
			`/PRINT=SPEARMAN.`

Crosstabulation & Chi-squared	`myWG <- table(workshop, gender)`
	`chisq.test(myWG)`
	
	`library("gmodels")`
	`CrossTable(workshop, gender, chisq = TRUE, format = "SAS")`
		`PROC FREQ;`
		`TABLES workshop*gender /CHISQ;`
			`CROSSTABS`
			`/TABLES=workshop BY gender`
			`/FORMAT= AVALUE TABLES`
			`/STATISTIC=CHISQ`
			`/CELLS= COUNT ROW`
			`/COUNT ROUND CELL.`

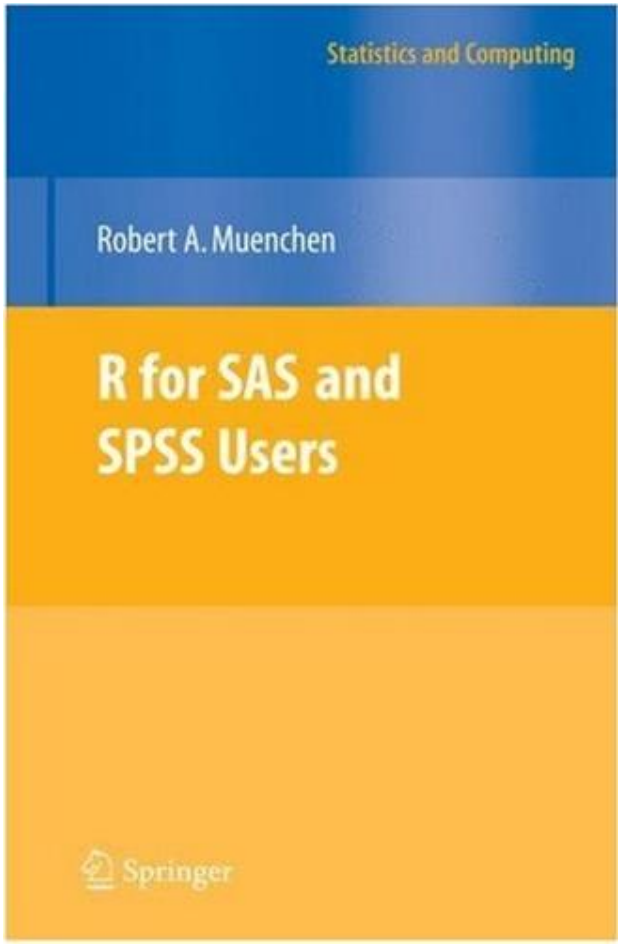
1st recommended book

- An R and S-PLUS Companion to Applied Regression: An excellent overview of R, not just regression in R. Highly recommended.



2nd recommended book

- R for SAS and SPSS Users: This book is geared to people who already know SAS or SPSS and want to learn R. If that describes you, you might consider buying this book.



A few more books

- *R for Stata Users*, Muenchen & Hilbe
- *R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*, Heiberger & Neuwirth
- Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Williams

Demo

Design of the R system

- The R system is divided into 2 conceptual parts:
  - The “base” R system
  - Everything else
- R functionality is divided into a number of packages

Design of the R system

- The “base” R system contains, among other things, the **base** package which is required to run R and contains the most fundamental functions
- The other packages contained in the “base” system include utils, stats, datasets, graphics, grDevices, grid, methods, tools, parallel, compiler, splines, tcltk, stats4

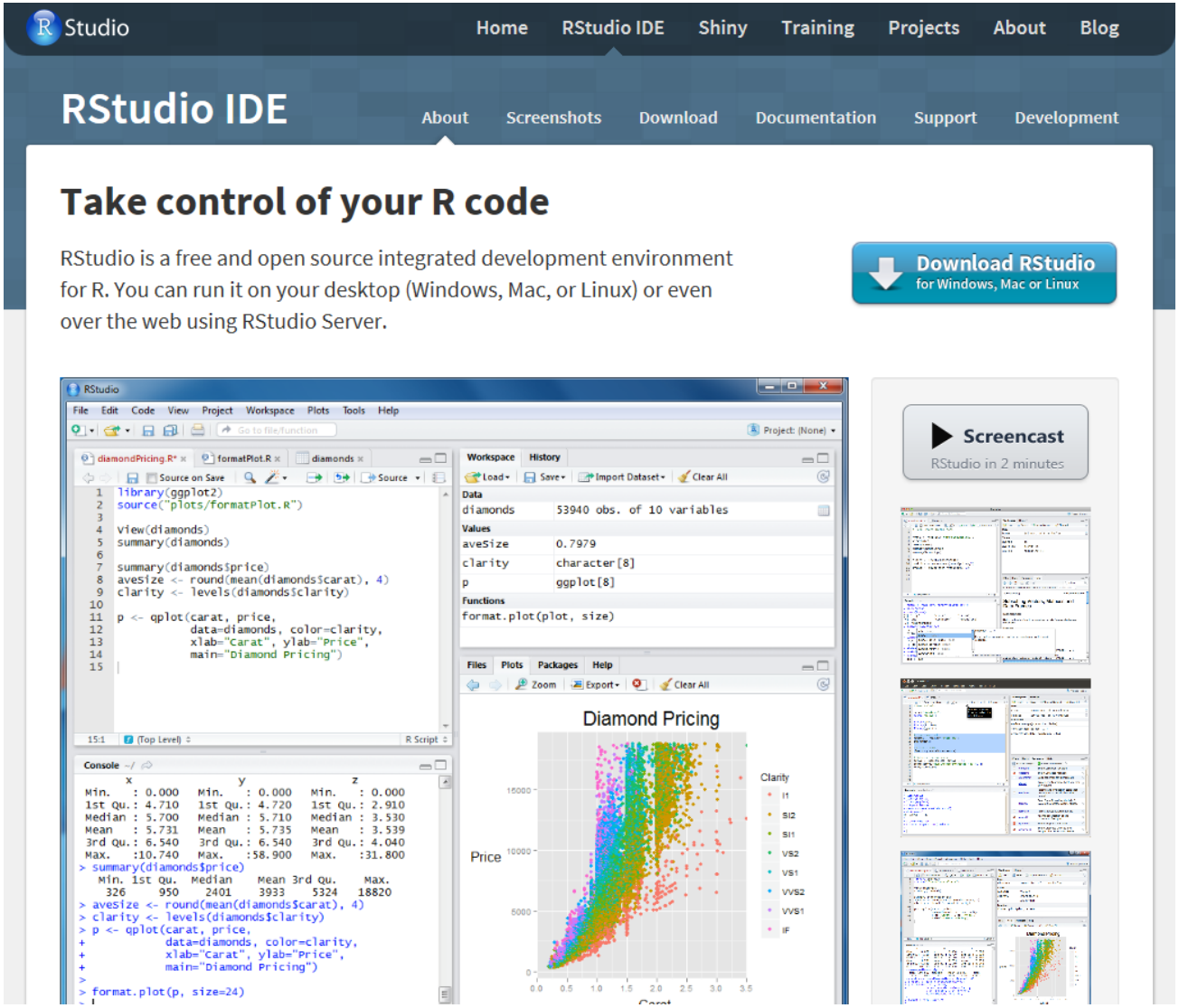
Design of the R system

- And there are many other packages available
- There are also many packages associated with the Bioconductor project (http://bioconductor.org)
- People often make packages available on their personal websites; there is no reliable way to keep track of how many packages are available in this fashion

Demo requirements

- Software
  - RStudio (http://www.rstudio.com/ide/)
- Source files
  - http://web.mit.edu/tkp/www/R/R_Tutorial_Data.txt
  - http://web.mit.edu/tkp/www/R/R_Tutorial_Inputs.txt

RStudio


Studio
Home
RStudio IDE
Shiny
Training
Projects
About
Blog
RStudio IDE
About
Screenshots
Download
Documentation
Support
Development
Take control of your R code
RStudio is a free and open source integrated development environment for R. You can run it on your desktop (Windows, Mac, or Linux) or even over the web using RStudio Server.
Download RStudio
for Windows, Mac or Linux
Screencast
RStudio in 2 minutes
Diamond Pricing
Price
Clarity

Demo

- Entering data
  - Math, Variables, Arrays
  - Math on arrays
  - Functions
- Getting help
- Reading data from files
- Selecting subsets of data

Demo

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
```

Demo

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
```

Demo

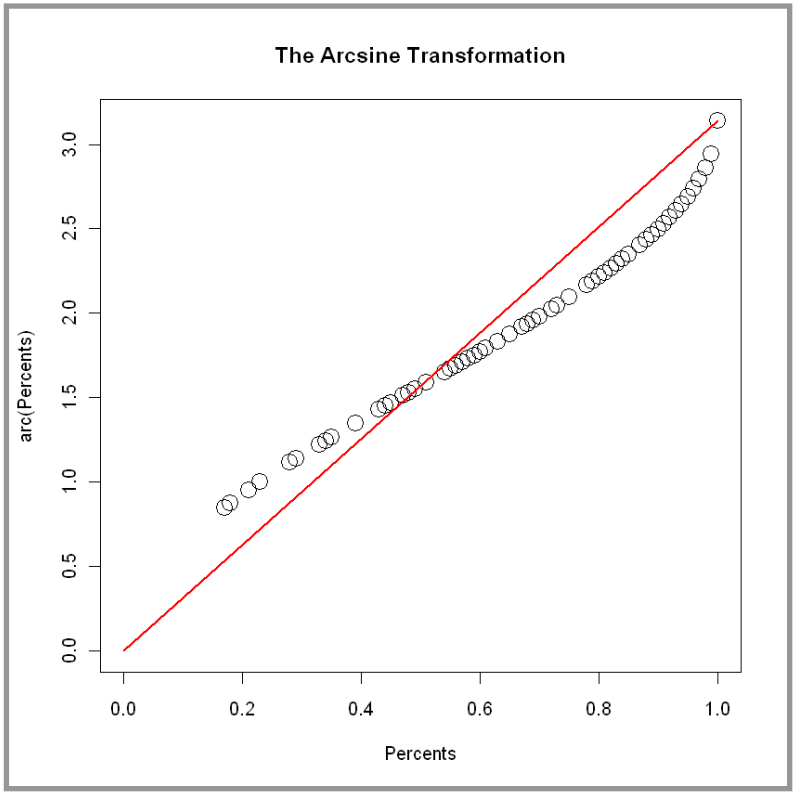
Math on arrays:

```
> 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
```

Demo

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
```

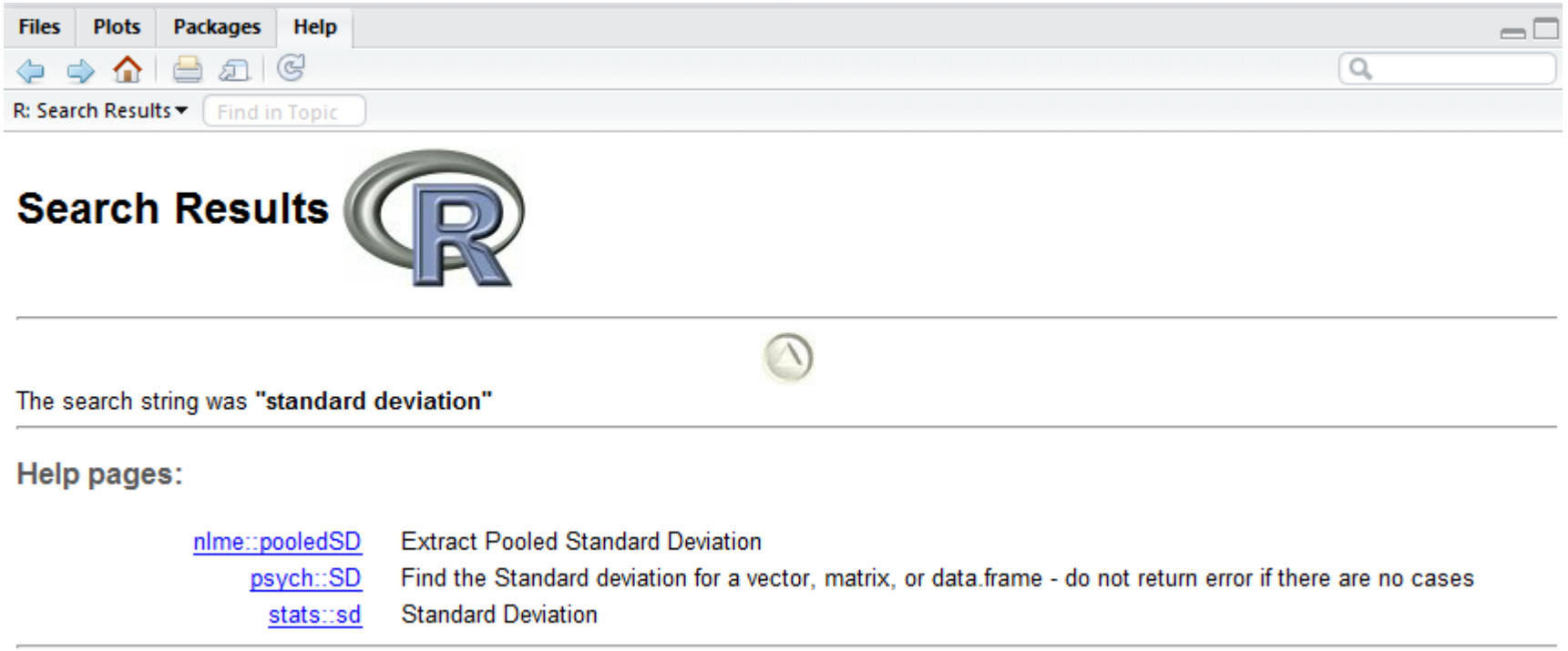



```
> plot(arc(Percents)~Percents,
+ pch=21,cex=2,xlim=c(0,1),ylim=c(0,pi),
+ main="The Arcsine Transformation")
> lines(c(0,1),c(0,pi),col="red",lwd=2)
```

Demo

Getting help:

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



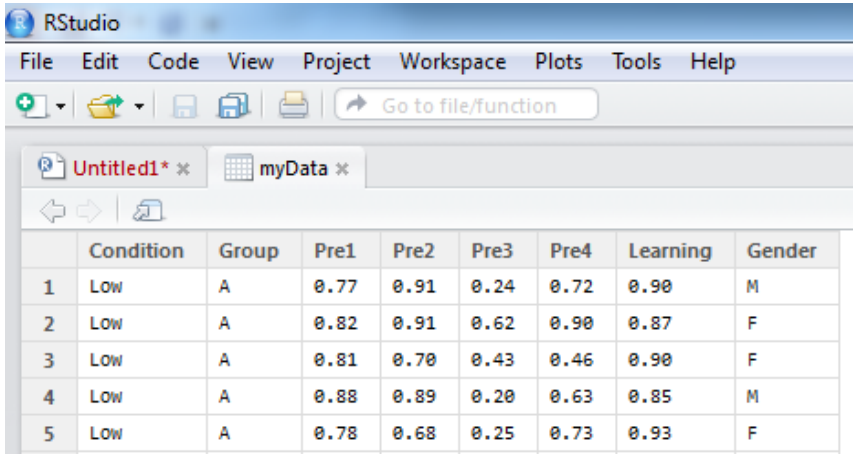
Demo

- Example experiment:
  - Subjects learning to perform a new task:
  - Two groups of subjects
    - (“A” and “B”; high and low aptitude learners)
  - Two types of training paradigm
    - (“High variability” and “Low variability”)
  - Four pre-training assessment tests
- Example data in “R_Tutorial_Data.txt”

Demo

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
```

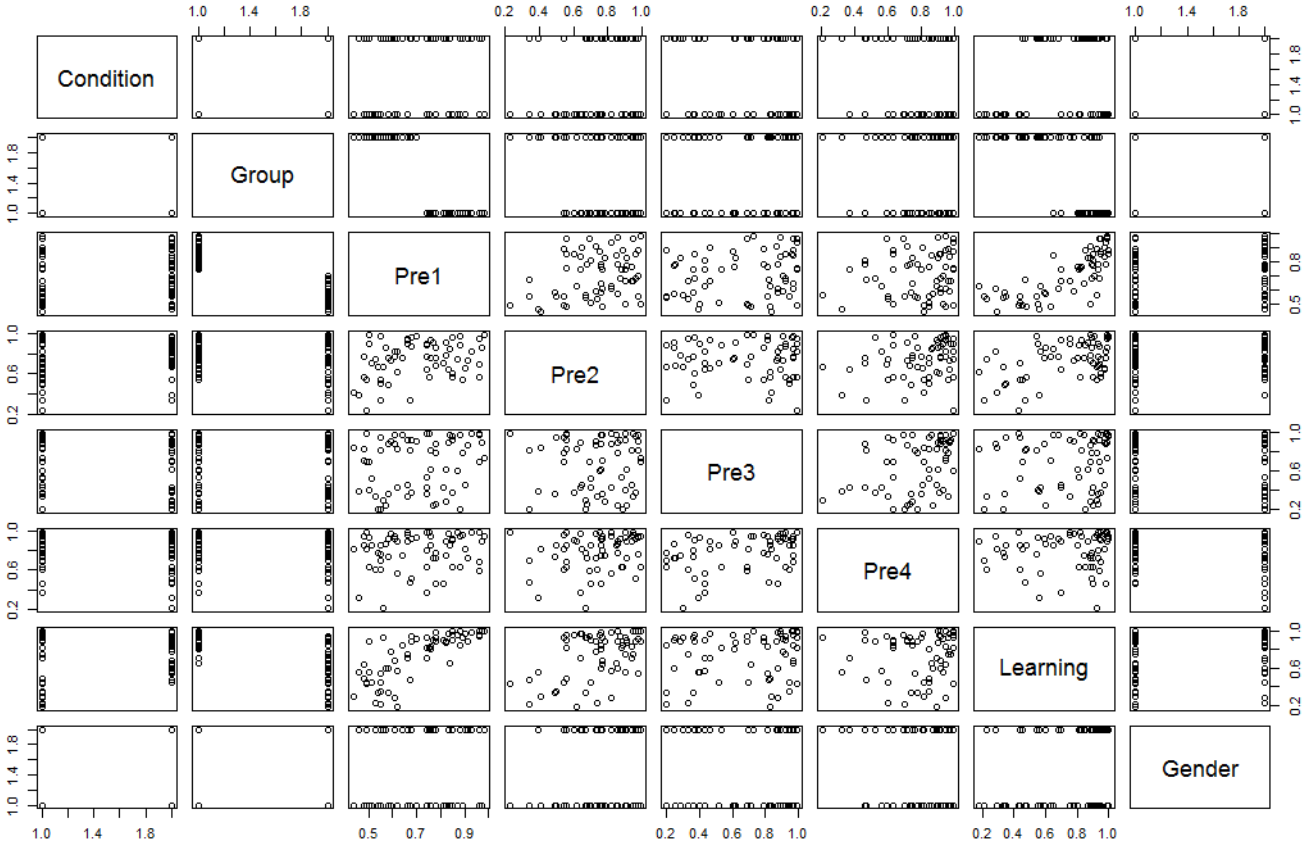


	Condition	Group	Pre1	Pre2	Pre3	Pre4	Learning	Gender
1	Low	A	0.77	0.91	0.24	0.72	0.90	M
2	Low	A	0.82	0.91	0.62	0.90	0.87	F
3	Low	A	0.81	0.70	0.43	0.46	0.90	F
4	Low	A	0.88	0.89	0.20	0.63	0.85	M
5	Low	A	0.78	0.68	0.25	0.73	0.93	F

Demo

Examining datasets:

```
> plot(myData)
```



Demo

Selecting subsets of data:

```
> myData$Learning
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.57 0.55
[19] 0.94 0.68 0.89 0.60 0.63 0.84 0.92 0.56 0.78
[28] 0.54 0.47 0.45 0.59 0.91 0.98 0.82 0.93 0.81
[37] 0.97 0.95 0.70 1.00 0.90 0.99 0.95 0.95 0.97
[46] 1.00 0.99 0.18 0.33 0.88 0.23 0.75 0.21 0.35
[55] 0.70 0.34 0.43 0.75 0.44 0.44 0.29 0.48 0.28
> myData$Learning[myData$Group=="A"]
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.98 0.82
[19] 0.93 0.81 0.97 0.95 0.70 1.00 0.90 0.99 0.95
[28] 0.95 0.97 1.00 0.99
```

Demo

Selecting subsets of data:

```
> myData$Learning
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.57 0.55
[19] 0.94 0.68 0.89 0.60 0.63 0.84 0.92 0.56 0.78
[28] 0.54 0.47 0.45 0.59 0.91 0.98 0.82 0.93 0.81
[37] 0.97 0.95 0.70 1.00 0.90 0.99 0.95 0.95 0.97
[46] 1.00 0.99 0.18 0.33 0.88 0.23 0.75 0.21 0.35
[55] 0.70 0.34 0.43 0.75 0.44 0.44 0.29 0.48 0.28
> attach(myData)
> Learning
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.57 0.55
[19] 0.94 0.68 0.89 0.60 0.63 0.84 0.92 0.56 0.78
[28] 0.54 0.47 0.45 0.59 0.91 0.98 0.82 0.93 0.81
[37] 0.97 0.95 0.70 1.00 0.90 0.99 0.95 0.95 0.97
[46] 1.00 0.99 0.18 0.33 0.88 0.23 0.75 0.21 0.35
[55] 0.70 0.34 0.43 0.75 0.44 0.44 0.29 0.48 0.28
```

Demo

Selecting subsets of data:

```
> Learning[Group=="A"]
 [1] 0.90 0.87 0.90 0.85 0.93 0.93 0.89 0.80 0.98
[10] 0.88 0.88 0.94 0.99 0.92 0.83 0.65 0.98 0.82
[19] 0.93 0.81 0.97 0.95 0.70 1.00 0.90 0.99 0.95
[28] 0.95 0.97 1.00 0.99
> Learning[Group!="A"]
 [1] 0.57 0.55 0.94 0.68 0.89 0.60 0.63 0.84 0.92
[10] 0.56 0.78 0.54 0.47 0.45 0.59 0.91 0.18 0.33
[19] 0.88 0.23 0.75 0.21 0.35 0.70 0.34 0.43 0.75
[28] 0.44 0.44 0.29 0.48 0.28
> Condition[Group=="B"&Learning<0.5]
 [1] Low  Low  High High High High High High High
[10] High High High High High
Levels: High Low
```

Demo - Statistics and Data Analysis

- Parametric Tests
  - Independent sample t-tests
  - Paired sample t-tests
  - One sample t-tests
  - Correlation
- Nonparametric tests
  - Shapiro-Wilks test for normality
  - Wilcoxon signed-rank test (Mann-Whitney U)
  - Chi square test
- Linear Models and ANOVA

Basic parametric inferential statistics

Independent sample t-tests:

```
> t.test(Pre2[Group=="A"],
+ Pre2[Group=="B"],
+ paired=FALSE)

        Welch Two Sample t-test

data:  Learning[Group == "A"] and Learning[Group == "B"]
t = 1.6117, df = 53.275, p-value = 0.1129
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0179193  0.1645725
sample estimates:
mean of x mean of y
0.7764516 0.7031250
```

http://www.wellesley.edu/Psychology/Psych205/indepttest.html

Basic parametric inferential statistics

Independent sample t-tests:

```
> t.test(Pre2[Group=="A"],
+ Pre2[Group=="B"],
+ paired=FALSE,
+ var.equal=TRUE)

        Welch Two Sample t-test

data:  Learning[Group == "A"] and Learning[Group == "B"]
t = 1.601, df = 61, p-value = 0.1145
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0179193  0.1645725
sample estimates:
mean of x mean of y
0.7764516 0.7031250
```

Basic parametric inferential statistics

Independent sample t-tests:

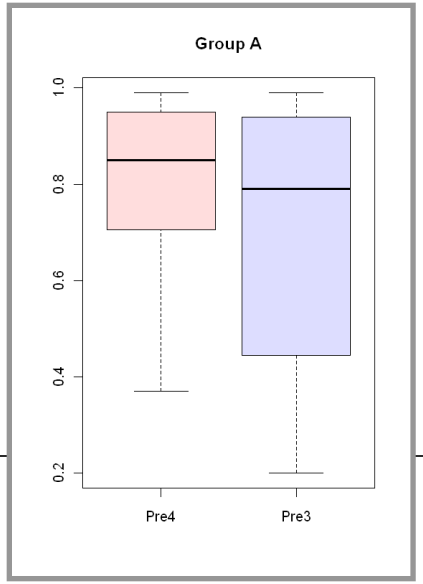
```
> t.test(Pre2[Group=="A"],
+ Pre2[Group=="B"],
+ paired=FALSE,
+ var.equal=TRUE,
+ alternative="greater")

        Welch Two Sample t-test

data:  Learning[Group == "A"] and Learning[Group == "B"]
t = 1.601, df = 61, p-value = 0.5727
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.003169388          Inf
sample estimates:
mean of x mean of y
0.7764516 0.7031250
```

Basic parametric inferential statistics

Paired sample t-test:

```
> t.test(Pre4[Group=="A"],
+ Pre3[Group=="A"],
+ paired=TRUE)
        Paired t-test
data:  Pre4[Group == "A"] and Pre3[Group == "A"]
t = 2.4054, df = 30, p-value = 0.02253
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01641059 0.20100876
sample estimates:
mean of the differences
              0.1087097
> boxplot(Pre4[Group=="A"],
+ Pre3[Group=="A"],
+ col=c("#ffdddd","#ddddff"),
+ names=c("Pre4","Pre3"),main="Group A")
```

http://www.wellesley.edu/Psychology/Psych205/pairttest.html

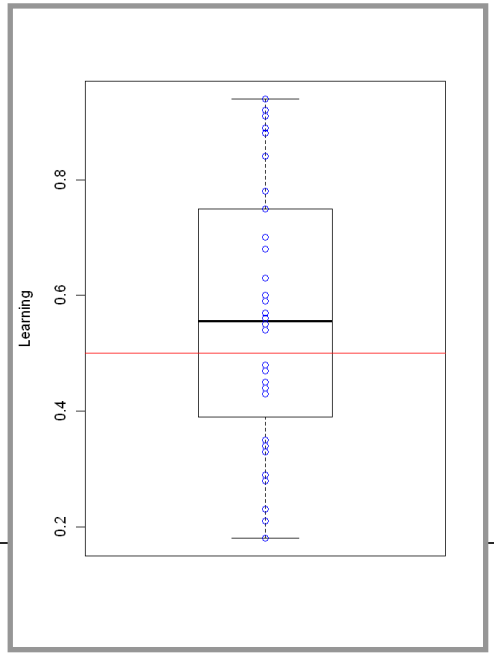


Basic parametric inferential statistics

One sample t-test:

```
> t.test(Learning[Group=="B"], mu=0.5, alternative="greater")

        One Sample t-test

data:  Learning[Group == "B"]
t = 1.5595, df = 31, p-value = 0.06452
alternative hypothesis: true mean is greater than 0.5
95 percent confidence interval:
 0.4945469       Inf
sample estimates:
mean of x
   0.5625
> boxplot(Learning[Group=="B"],
+ names="Group B", ylab="Learning")
> lines(c(0,2), c(0.5, 0.5), col="red")
> points(c(rep(1,length(Learning[Group=="B"]))),
+ Learning[Group=="B"], pch=21, col="blue")
```

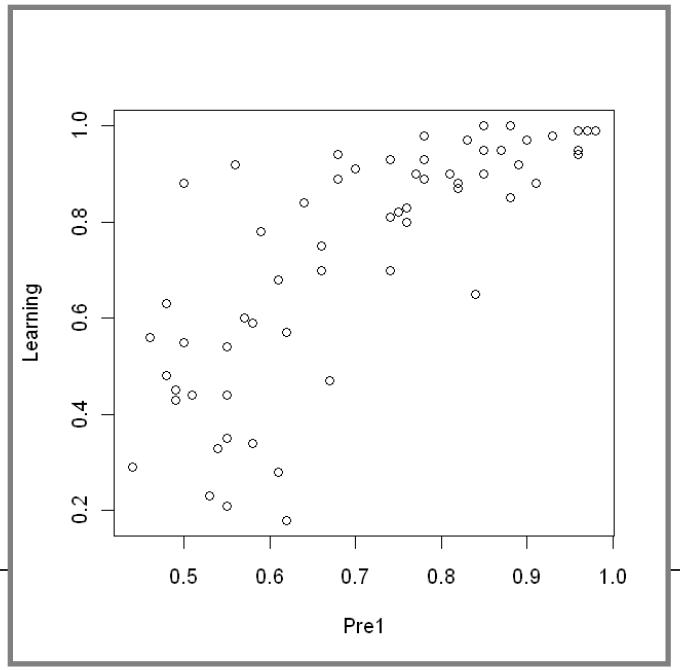
http://www.wellesley.edu/Psychology/Psych205/onettest.html



Basic parametric inferential statistics

Correlation:

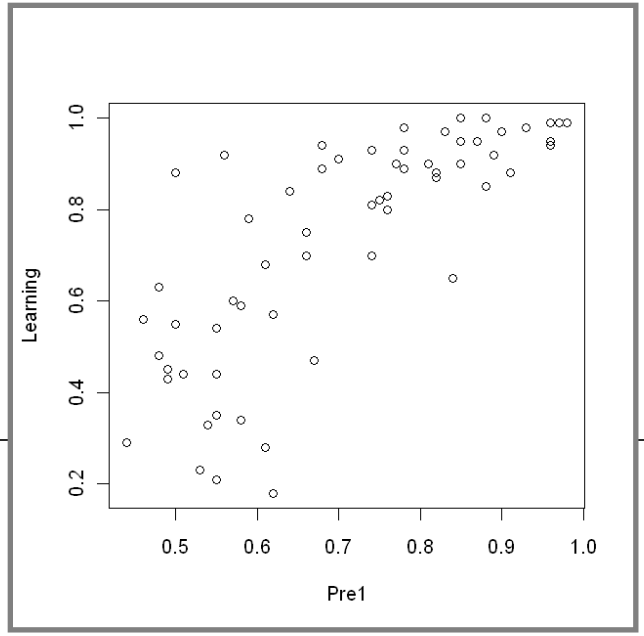
```
> cor.test(Pre1,Learning,method="pearson")
        Pearson's product-moment correlation
data:  Pre1 and Learning
t = 9.2461, df = 61, p-value = 3.275e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6366698 0.8506815
sample estimates:
      cor
0.7639292
> plot(Pre1,Learning)
```



Basic parametric inferential statistics

Correlation (fancier plot example):

```
> cor.test(Pre1,Learning,method="pearson")
        Pearson's product-moment correlation
data:  Pre1 and Learning
t = 9.2461, df = 61, p-value = 3.275e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6366698 0.8506815
sample estimates:
      cor
0.7639292
> plot(Pre1,Learning)
```



Statistics and data analysis

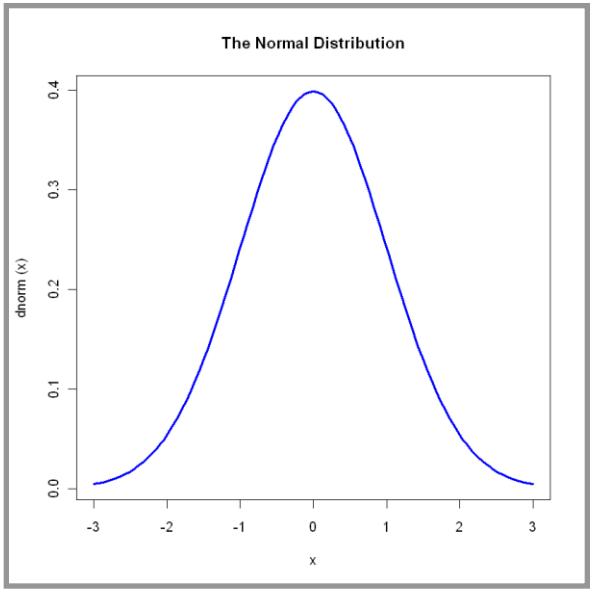
Are my data normally distributed?

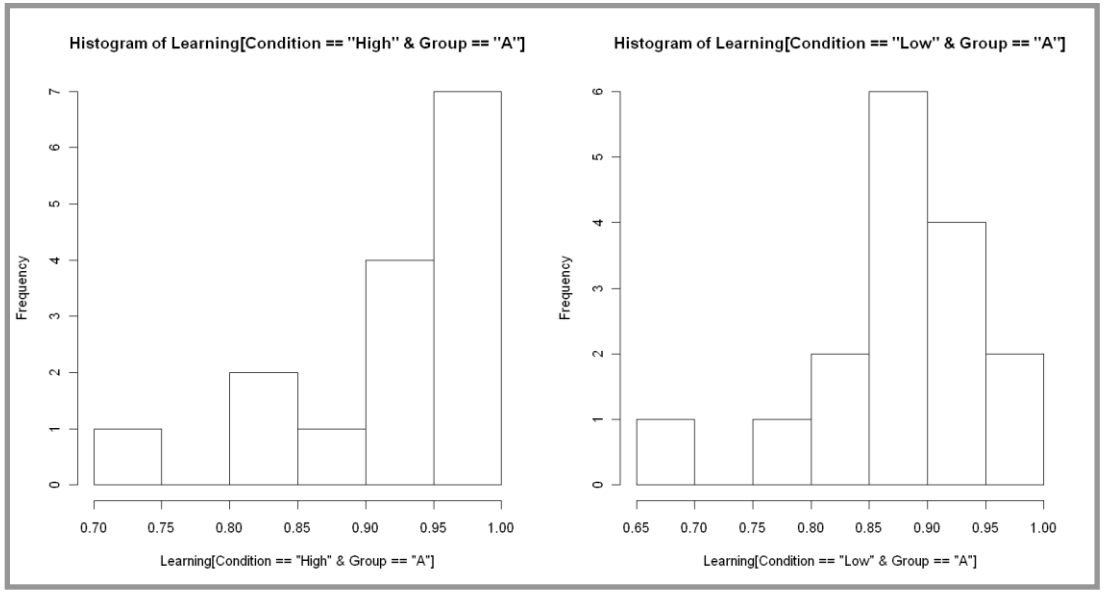
```
> t.test(Learning[Condition=="High"&Group=="A"],
+ Learning[Condition=="Low"&Group=="A"])

        Welch Two Sample t-test

data:  Learning[Condition == "High" & Group == "A"] and
    Learning[Condition == "Low" & Group == "A"]
t = 1.457, df = 28.422, p-value = 0.1561
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01764821  0.10481488
sample estimates:
mean of x mean of y
0.9273333 0.8837500
```

Statistics and data analysis

Are my data normally distributed?

```
> plot(dnorm,-3,3,col="blue",lwd=3,main="The Normal Distribution")
> par(mfrow=c(1,2))
> hist(Learning[Condition=="High"&Group=="A"])
> hist(Learning[Condition=="Low"&Group=="A"])
```





Statistics and data analysis

Are my data normally distributed?

```
> shapiro.test(Learning[Condition=="High"&Group=="A"])

        Shapiro-Wilk normality test

data:  Learning[Condition == "High" & Group == "A"]
W = 0.7858, p-value = 0.002431

> shapiro.test(Learning[Condition=="Low"&Group=="A"])

        Shapiro-Wilk normality test

data:  Learning[Condition == "Low" & Group == "A"]
W = 0.8689, p-value = 0.02614
```

http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

Basic non-parametric inferential statistics

Wilcoxon signed-rank/Mann-Whitney U tests:

```
> wilcox.test(Learning[Condition=="High"&Group=="A"],
+ Learning[Condition=="Low"&Group=="A"],
+ exact=FALSE,
+ paired=FALSE)

        Wilcoxon rank sum test with continuity correction

data:  Learning[Condition == "High" & Group == "A"] and
    Learning[Condition == "Low" & Group == "A"]
W = 173.5, p-value = 0.03580
alternative hypothesis: true location shift is not equal to 0
```

http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U

Basic non-parametric inferential statistics

Chi-squared tests:

```
> x <- matrix(c(
+ length(Learning[Group=="A"&Condition=="High"&Gender=="F"]),
+ length(Learning[Group=="A"&Condition=="Low"&Gender=="F"]),
+ length(Learning[Group=="B"&Condition=="High"&Gender=="F"]),
+ length(Learning[Group=="B"&Condition=="Low"&Gender=="F"])),
+ ncol=2)
> x
     [,1] [,2]
[1,]    4   12
[2,]   10    7
> chisq.test(x)

        Pearson's Chi-squared test with Yates' continuity correction

data:  x
X-squared = 2.5999, df = 1, p-value = 0.1069
```

http://en.wikipedia.org/wiki/Chi-squared_distribution

Linear models and ANOVA

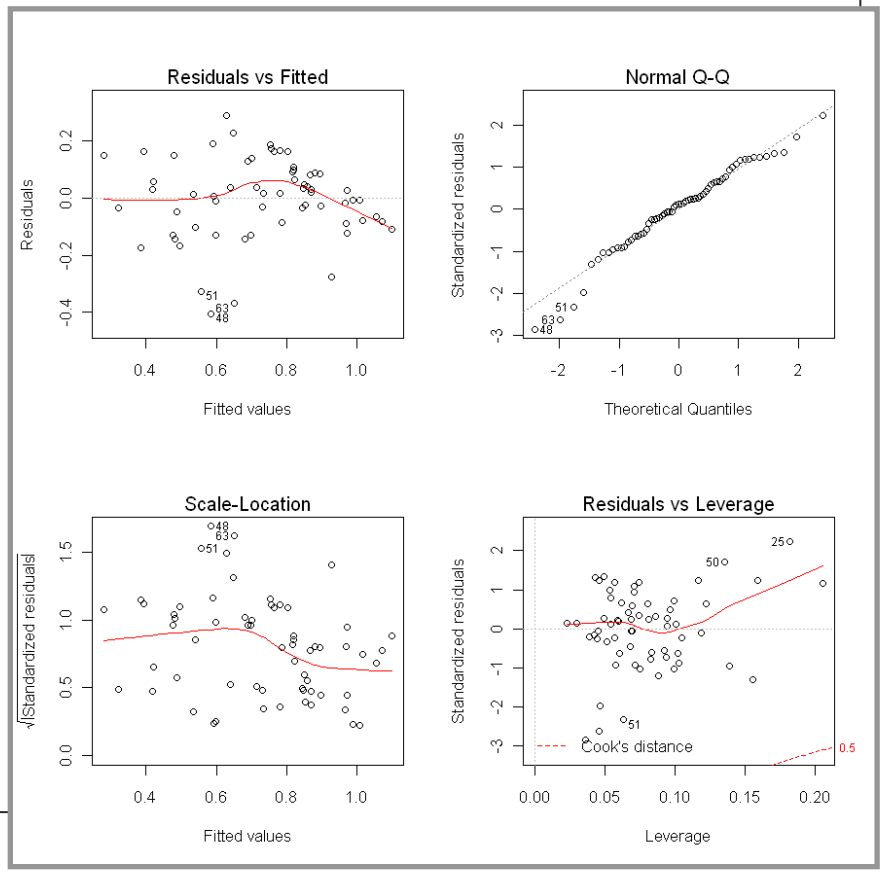
Linear models:

```
> myModel <- lm(Learning ~ Pre1 + Pre2 + Pre3 + Pre4)
> par(mfrow=c(2,2))
> plot(myModel)
```

http://online.stat.psu.edu/online/development/stat501/05model_check/03model_check_rvf.html

http://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

http://www.jerrydallal.com/LHSP/summary.htm



Linear models and ANOVA

Linear models:

```
> summary(myModel)

Call:
lm(formula = Learning ~ Pre1 + Pre2 + Pre3 + Pre4)
Residuals:
     Min       1Q   Median       3Q      Max
-0.40518 -0.08460  0.01707  0.09170  0.29074
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.22037    0.11536  -1.910 0.061055 .
Pre1         1.05299    0.12636   8.333 1.70e-11 ***
Pre2         0.41298    0.10926   3.780 0.000373 ***
Pre3         0.07339    0.07653   0.959 0.341541
Pre4        -0.18457    0.11318  -1.631 0.108369
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1447 on 58 degrees of freedom
Multiple R-squared: 0.6677,     Adjusted R-squared: 0.6448
F-statistic: 29.14 on 4 and 58 DF,  p-value: 2.710e-13
```

Linear models and ANOVA

Linear models:

```
> step(myModel, direction="backward")
Start:  AIC=-238.8
Learning ~ Pre1 + Pre2 + Pre3 + Pre4

       Df Sum of Sq    RSS     AIC
- Pre3  1   0.01925 1.2332 -239.81
<none>              1.2140 -238.80
- Pre4  1   0.05566 1.2696 -237.98
- Pre2  1   0.29902 1.5130 -226.93
- Pre1  1   1.45347 2.6675 -191.21

Step:  AIC=-239.81
Learning ~ Pre1 + Pre2 + Pre4

       Df Sum of Sq    RSS     AIC
- Pre4  1   0.03810 1.2713 -239.89
<none>              1.2332 -239.81
- Pre2  1   0.28225 1.5155 -228.83
- Pre1  1   1.54780 2.7810 -190.58

                ...
```

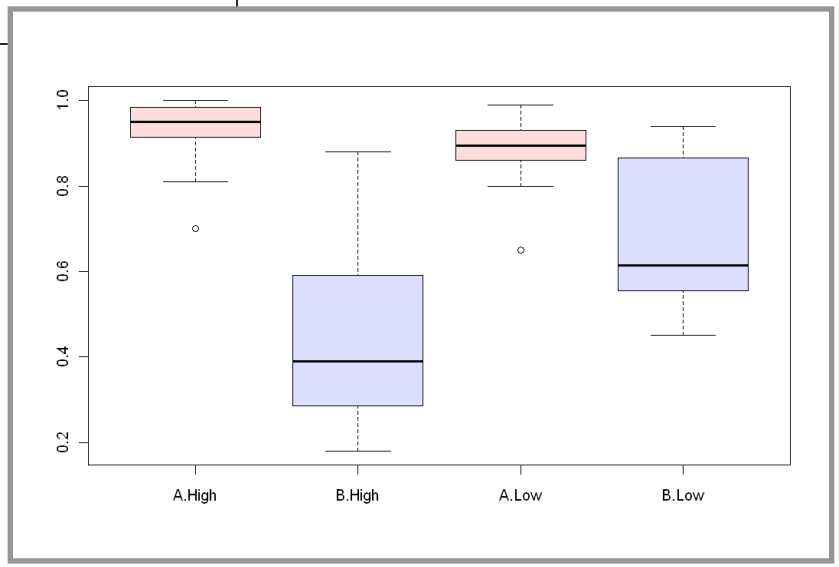
```
...

Step:  AIC=-239.89
Learning ~ Pre1 + Pre2

       Df Sum of Sq    RSS     AIC
<none>              1.2713 -239.89
- Pre2  1   0.24997 1.5213 -230.59
- Pre1  1   1.52516 2.7965 -192.23

Call:
lm(formula = Learning ~ Pre1 + Pre2)

Coefficients:
(Intercept)         Pre1         Pre2
    -0.2864       1.0629       0.3627
```

Linear models and ANOVA

ANOVA:

```
> myANOVA <- aov(Learning~Group*Condition)
> summary(myANOVA)
                Df Sum Sq Mean Sq F value    Pr(>F)
Group            1 1.8454 1.84537 81.7106 9.822e-13 ***
Condition        1 0.1591 0.15910  7.0448 0.0102017 *
Group:Condition  1 0.3164 0.31640 14.0100 0.0004144 ***
Residuals       59 1.3325 0.02258
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> boxplot(Learning~Group*Condition,col=c("#ffdddd","#ddddff"))
```

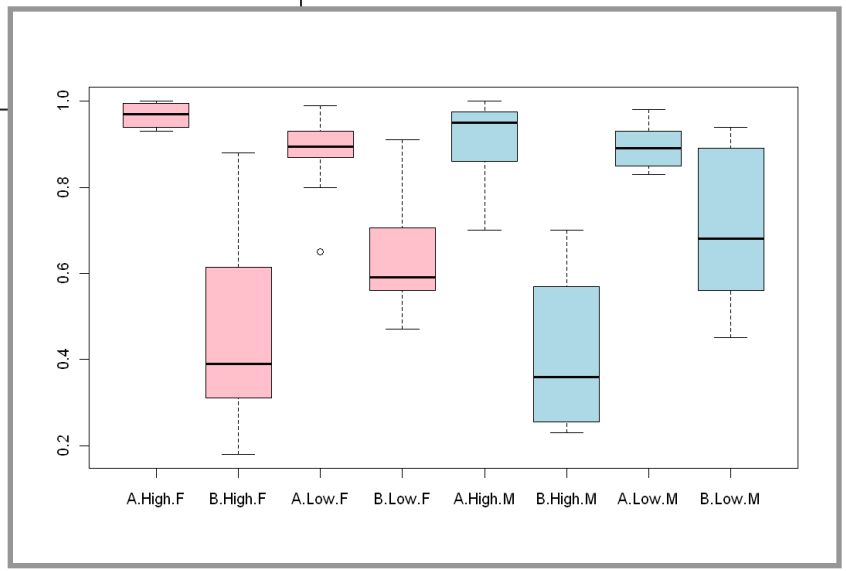
http://en.wikipedia.org/wiki/Analysis_of_variance



Linear models and ANOVA

ANOVA:

```
> myANOVA2 <- aov(Learning~Group*Condition+Gender)
> summary(myANOVA2)
                Df  Sum Sq Mean Sq F value    Pr(>F)
Group            1 1.84537 1.84537 80.3440 1.523e-12 ***
Condition        1 0.15910 0.15910  6.9270  0.010861 *
Gender           1 0.04292 0.04292  1.8688  0.176886
Group:Condition  1 0.27378 0.27378 11.9201  0.001043 **
Residuals       58 1.33216 0.02297
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> boxplot(Learning~Group*Condition+Gender,
+ col=c(rep("pink",4),rep("light blue",4)))
```



The End

- josipsaban@gmail.com