R for statistics: first steps

CAR meeting, Raleigh, Feb 26 2011

Peter Aldhous,
San Francisco Bureau Chief

peter@peteraldhous.com

A note of caution before using any stats package: Beware running with scissors!

You can bamboozle yourself, and your readers, by misusing statistics.

Make sure you understand the methods you are using, in particular the assumptions that must be met for them to be valid (e.g. normal distribution for many common tests).

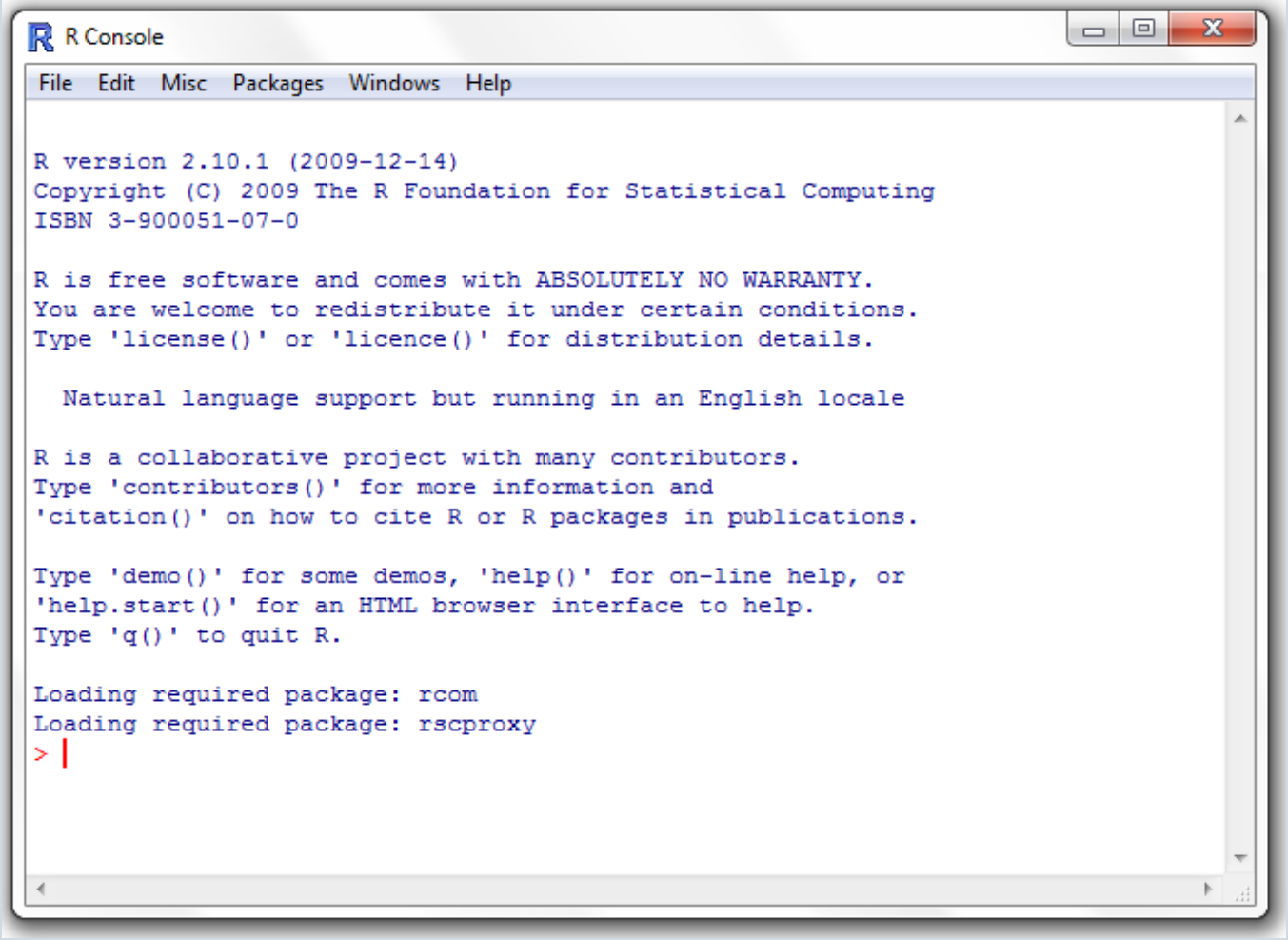
So before rushing in, consult:

- **IRE tipsheets** (e.g. Donald & LaFleur, #2752; Donald & Hacker, #2731)
- **Statistical textbooks** (e.g. http://www.statsoft.com/textbook/ is free online)
- **Experts who can provide a reality check on your analysis!**

Getting started:

Download R, instructions at: http://www.r-project.org/

Start the program:

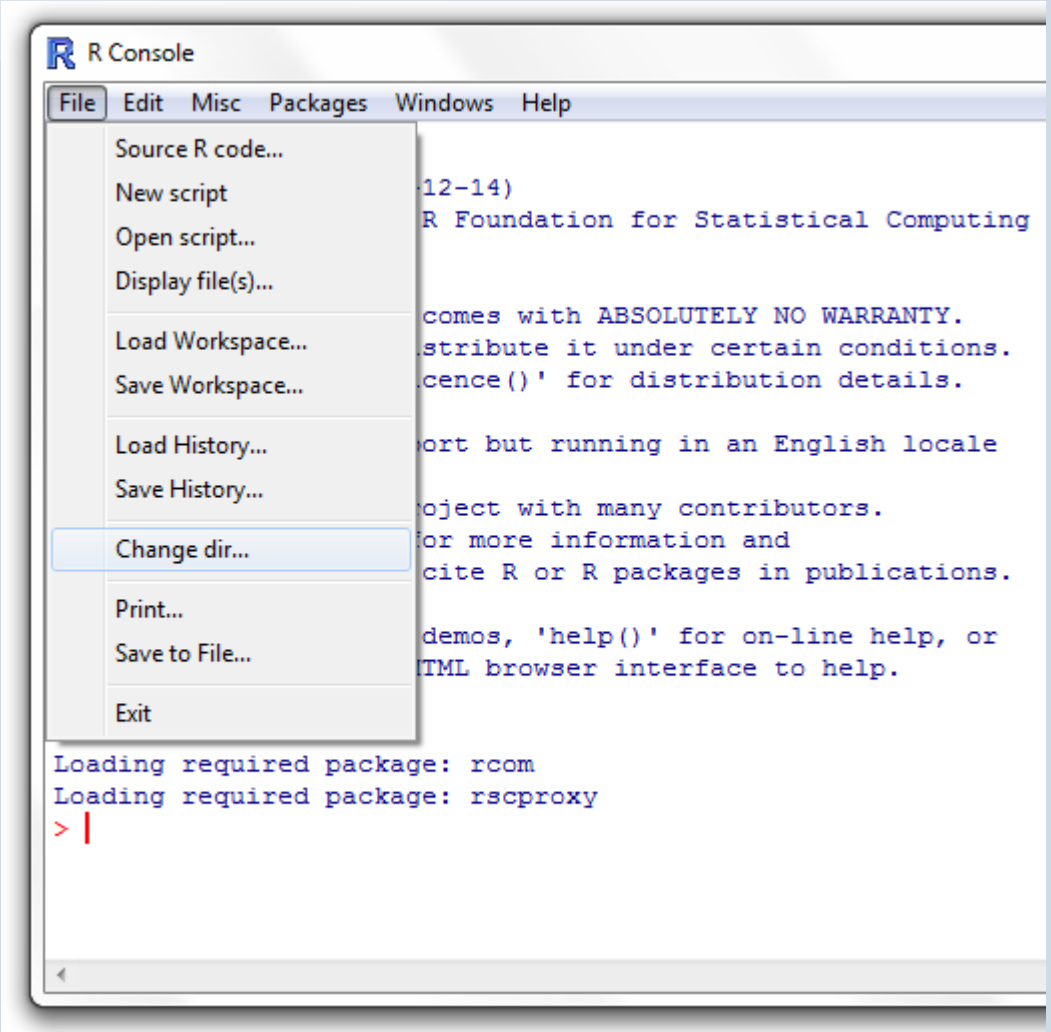


Prepare the data:

1) Save as a csv file

	A	B	C	D	E
1	Age	Salary	Sector		
2	74	8	Financial services		
3	70	2.13	Manufacturing		
4	69	7.5	Retail		
5	69	4.82	Manufacturing		
6	63	1.49	Manufacturing		
7	62	3.96	Manufacturing		
8	62	2.34	Retail		
9	61	5.43	Manufacturing		
10	61	2.54	Financial services		
11	61	1.98	Retail		
12	60	7.26	Manufacturing		
13	59	2.96	Manufacturing		
14	58	2.17	Retail		
15	58	0.21	Retail		
16	57	11.03	Financial services		
17	57	3.17	Financial services		
18	56	8.02	Financial services		
19	56	3.5	Manufacturing		
20	56	2.82	Retail		
21	56	2.04	Manufacturing		
22	55	7.36	Manufacturing		
23	55	4.98	Retail		
24	55	4.24	Financial services		
25	53	4.06	Retail		

2) Point R at the data



First steps in the R command line:

Load and examine the data:

```
> CEOs <- read.csv("CEOs.csv", header=T)
> ls()
[1] "CEOs"
> str(CEOs)
'data.frame':   59 obs. of  3 variables:
 $ Age   : int  74 70 69 69 63 62 62 61 61 61 ...
 $ Salary: num  8 2.13 7.5 4.82 1.49 3.96 2.34 5.43 2.54 1.98 ...
 $ Sector: Factor w/ 3 levels "Financial services",..: 1 2 3 2 2 2 3 2 1 3 ...
```

View a basic summary:

```
> summary(CEOs)
      Age             Salary                Sector
 Min.   :32.00   Min.   : 0.210   Financial services:15
 1st Qu.:45.50   1st Qu.: 2.500   Manufacturing     :20
 Median :50.00   Median : 3.500   Retail            :24
 Mean   :51.54   Mean   : 4.042
 3rd Qu.:57.00   3rd Qu.: 5.395
 Max.   :74.00   Max.   :11.030
```

What is the mean age and salary for CEOs in each sector?

```
> tapply(CEOs$Age, CEOs$Sector, mean)
Financial services      Manufacturing             Retail
          51.06667           54.30000           49.54167
> tapply(CEOs$Salary, CEOs$Sector, mean)
Financial services      Manufacturing             Retail
          5.192667           4.236000           3.160417
```

An alternative, computing mean and standard deviation in one go:

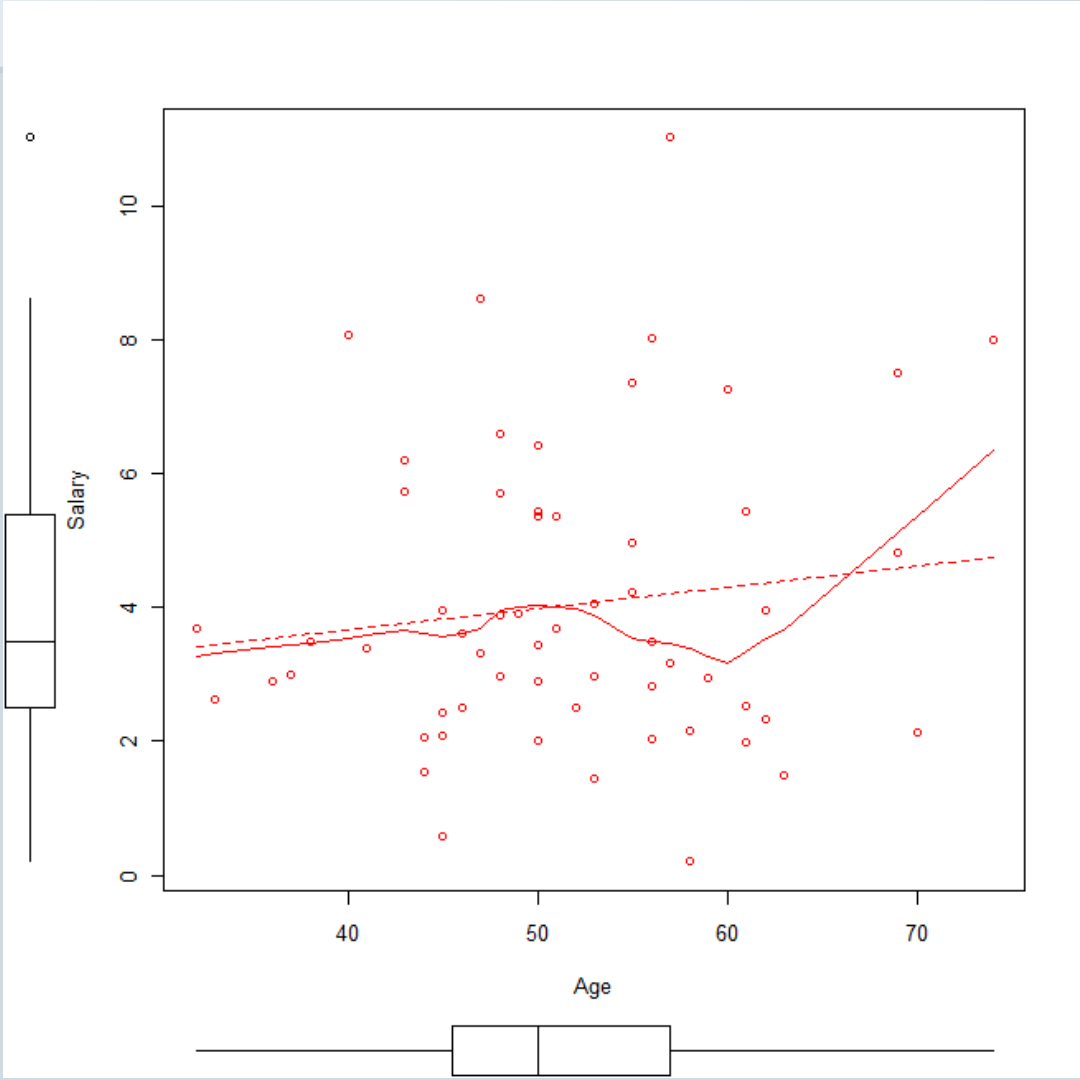
```
> numSummary(CEOs$Salary, groups=CEOs$Sector, statistics=c("mean", "sd"))
                        mean       sd  n
Financial services  5.192667 2.361245 15
Manufacturing       4.236000 2.157409 20
Retail              3.160417 1.821692 24
```

Is there a correlation between age and salary?

```
> cor.test(CEOs$Age, CEOs$Salary, method="pearson")

        Pearson's product-moment correlation

data:  CEOs$Age and CEOs$Salary
t = 0.971, df = 57, p-value = 0.3357
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1328669  0.3715027
sample estimates:
      cor
0.1275554
```

Draw a graph to explore the correlation analysis:

```
> scatterplot(Salary ~ Age, data=CEOs)
```



Does mean CEO age differ significantly across sectors?

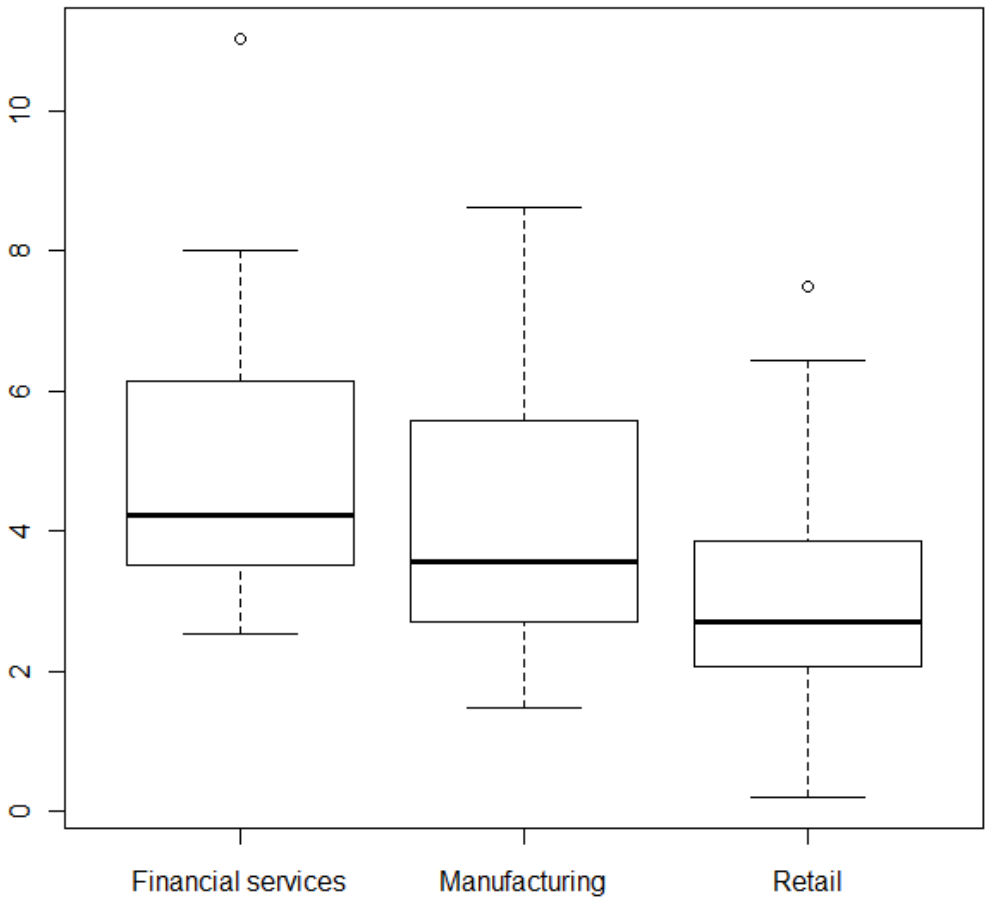
```
> AnovaAge <- aov(Age ~ Sector, data=CEOs)
> summary(AnovaAge)
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
Sector        2   251.6  125.776   1.5917  0.2126
Residuals    56  4425.1   79.019
```

Does mean CEO salary differ significantly across sectors?

```
> AnovaSalary <- aov(Salary ~ Sector, data=CEOs)
> summary(AnovaSalary)
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
Sector        2  39.266  19.633   4.5279  0.01504 *
Residuals    56 242.817   4.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Draw a graph to explore the distribution of salary by sector:

```
> boxplot(Salary ~ Sector, data=CEOs)
```

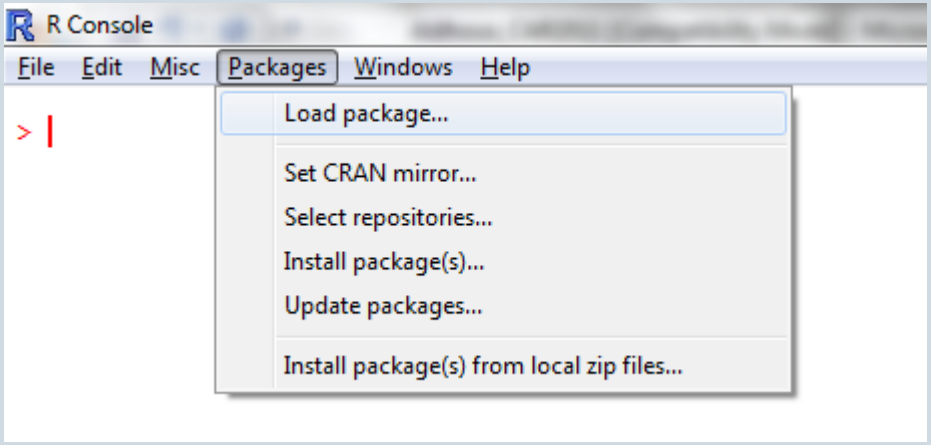


Moving beyond the basic functions: R packages

The R community has written 2800+ extensions to R's basic statistical and graphical functions, available from the CRAN repository:
http://cran.r-project.org/web/packages/

Follow the links to find a PDF Reference Manual for each package.

To see what packages you have available:



If you can't find the package you need, select "Install package(s)" and download from the nearest mirror of the CRAN repository

Some examples of R packages:

- **ggplot2**: for sophisticated statistical graphics

- **survival**: for analysis of data where the dependent variable is the time that elapses to a particular event. Used in medical research to see if drugs prolong life; you might use these methods to analyze the time taken to process permit applications for different groups of people

- **rgdal**: for geospatial data analysis; allows you import shapefiles and other forms of geodata and perform GIS-type analyses

Back to our analysis of CEO salary by sector:

Previous summary:

```
> AnovaSalary <- aov(Salary ~ Sector, data=CEOs)
> summary(AnovaSalary)
             Df  Sum Sq Mean Sq F value  Pr(>F)
Sector        2  39.266  19.633  4.5279 0.01504 *
Residuals    56 242.817   4.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Load multcomp package using "Load package" or:

```
> require(multcomp)
Loading required package: multcomp
Loading required package: mvtnorm
Loading required package: survival
Loading required package: splines
```

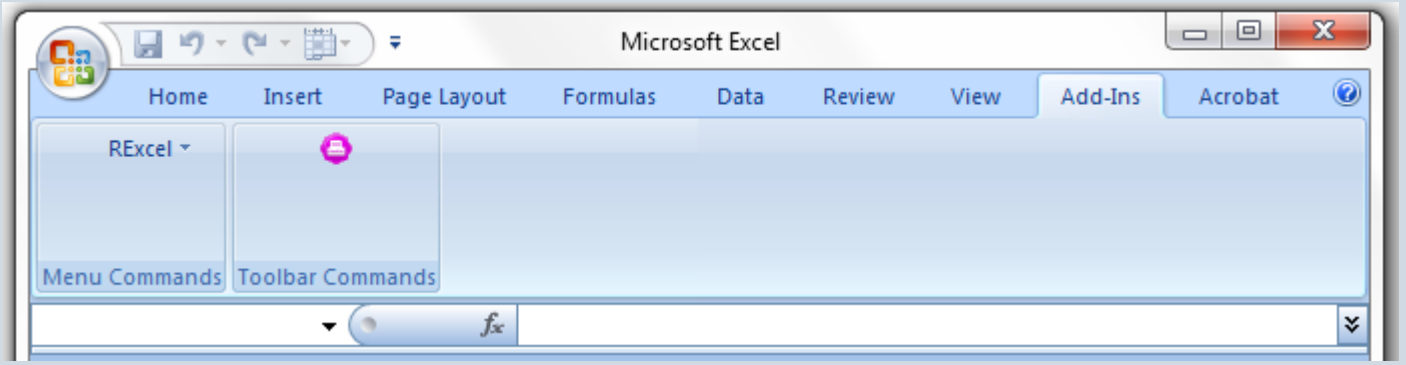
Perform multiple comparisons:

```
> Comparisons <- glht(AnovaSalary, linfct = mcp(Sector = "Tukey"))
> summary(Comparisons)

     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Salary ~ Sector, data = CEOs)

Linear Hypotheses:
                                          Estimate Std. Error t value Pr(>|t|)
Manufacturing - Financial services == 0  -0.9567    0.7112  -1.345  0.3759
Retail - Financial services == 0         -2.0322    0.6854  -2.965  0.0121 *
Retail - Manufacturing == 0              -1.0756    0.6305  -1.706  0.2114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

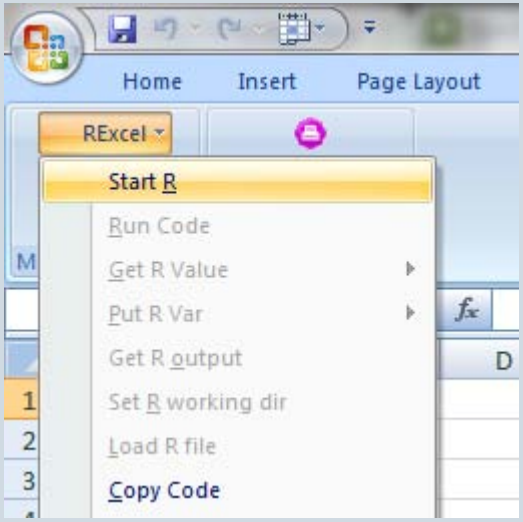
If the command line isn't for you:
R through Excel

Download RExcel, instructions at: http://rcom.univie.ac.at/download.html

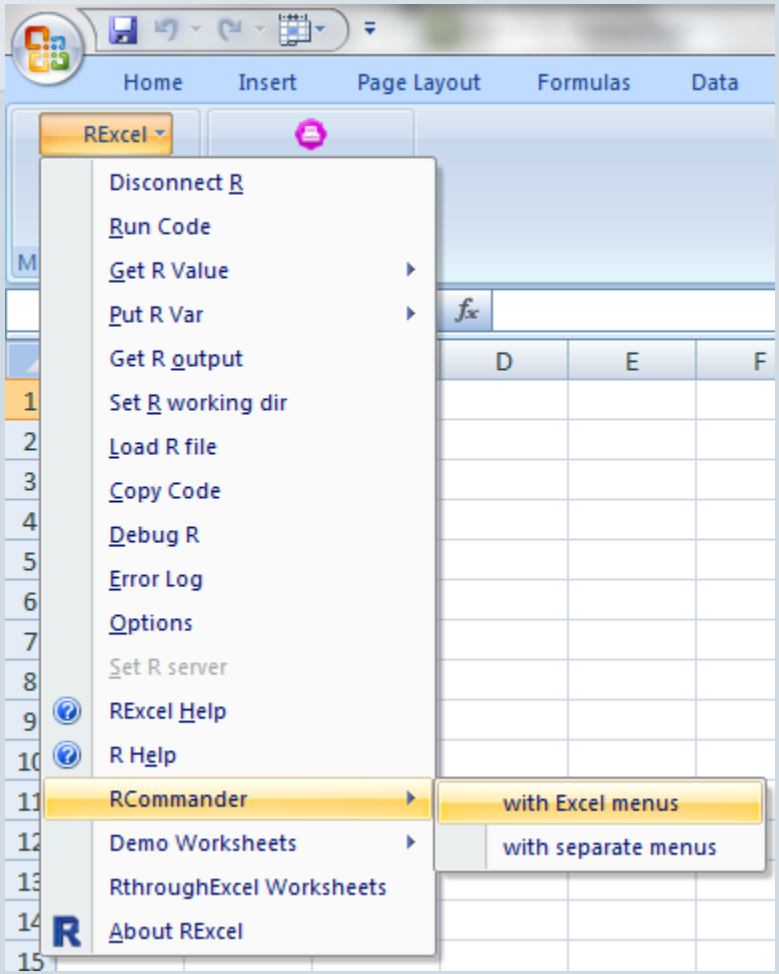
Getting started: 1)



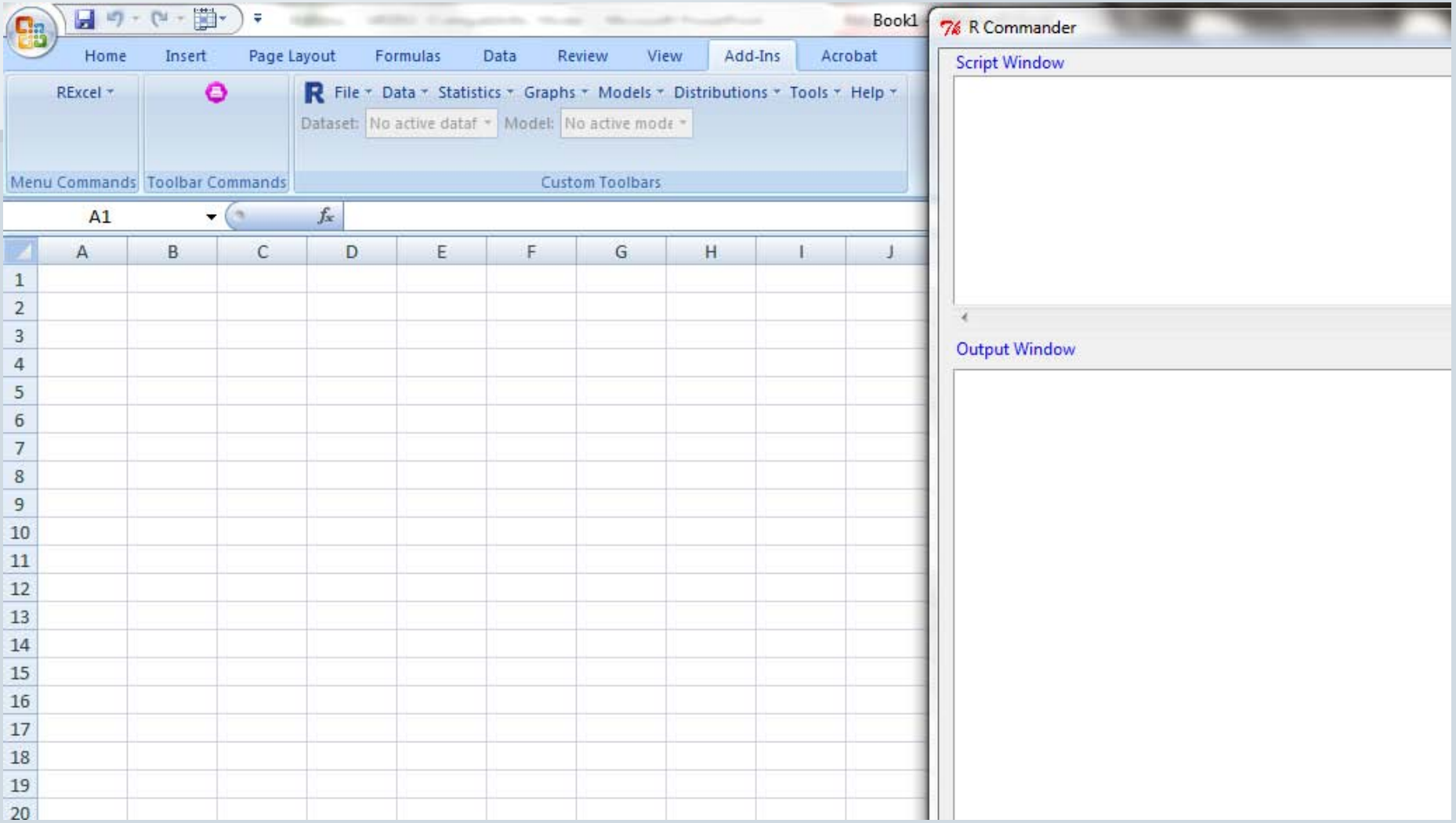
2)



Getting started: 3)

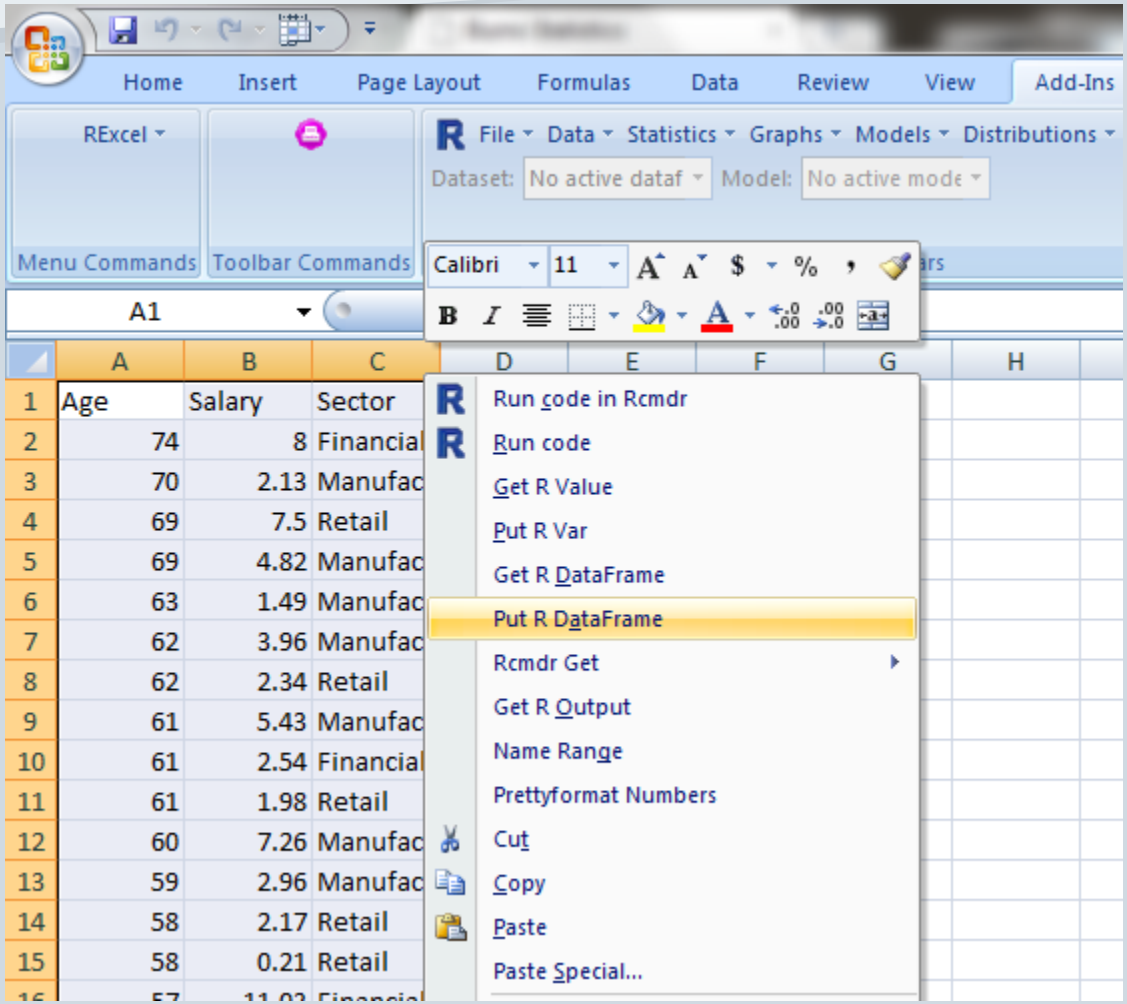


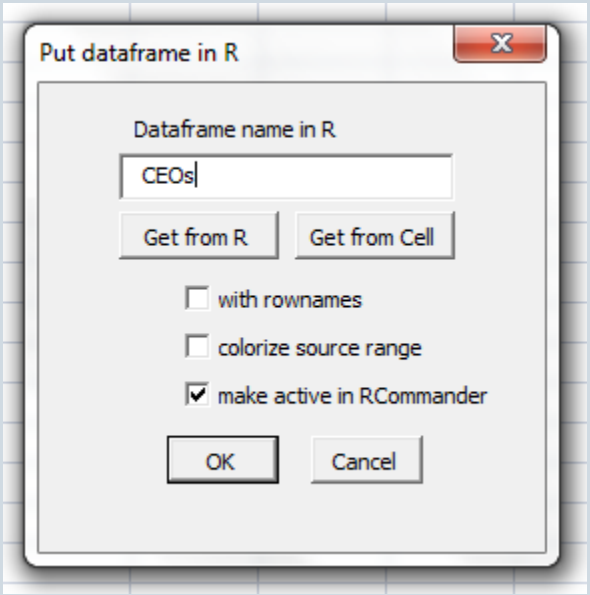
Now we're ready to go:



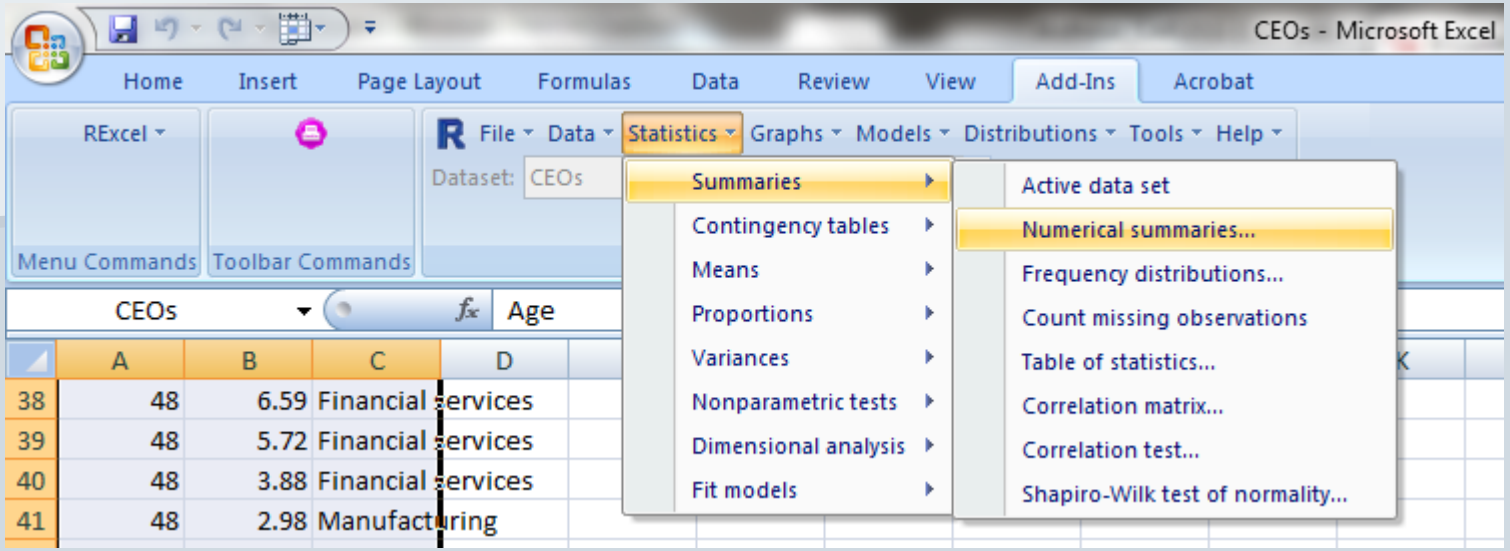
Load the data into R:

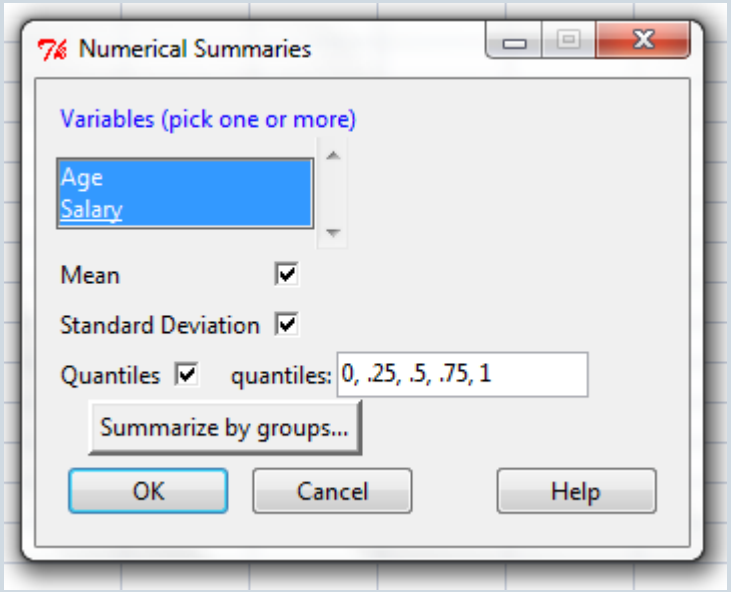
Highlight the data range, right click and select “Put R DataFrame”

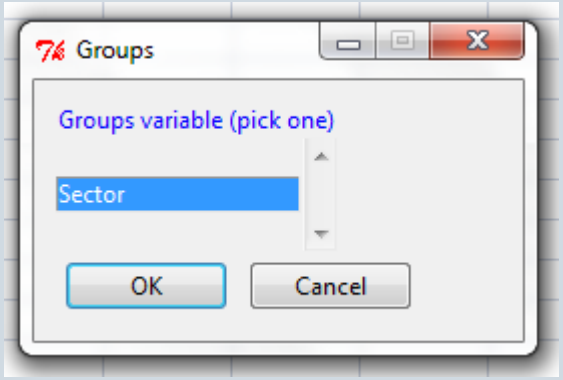




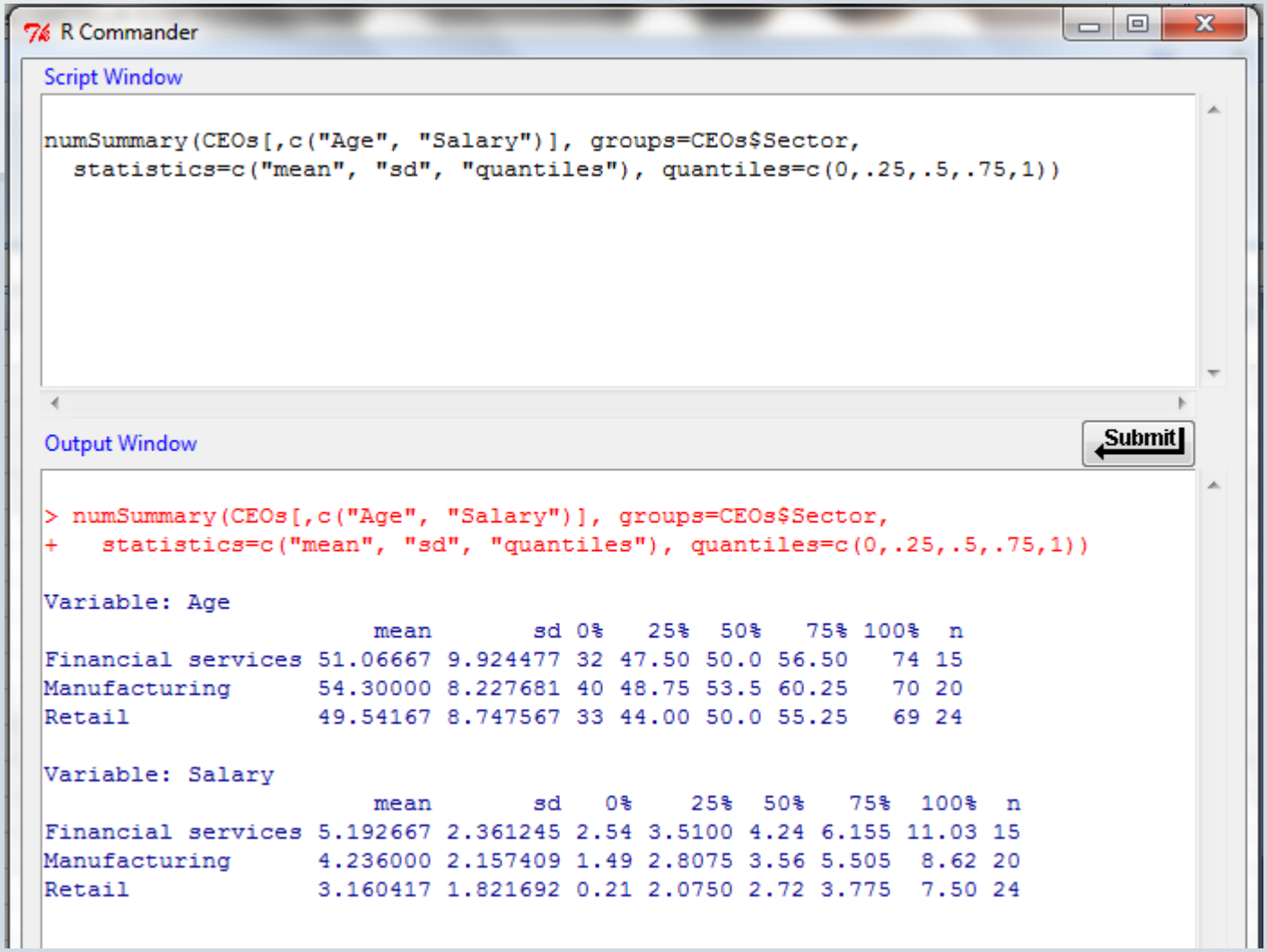
Compute summary statistics:



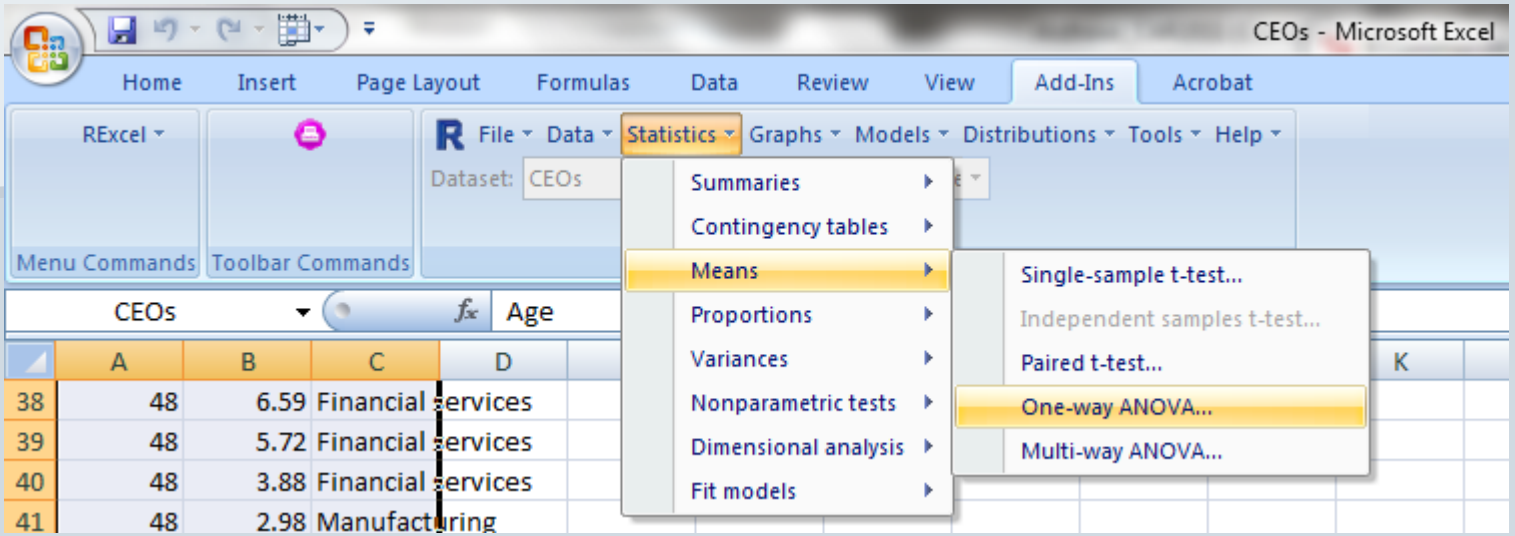


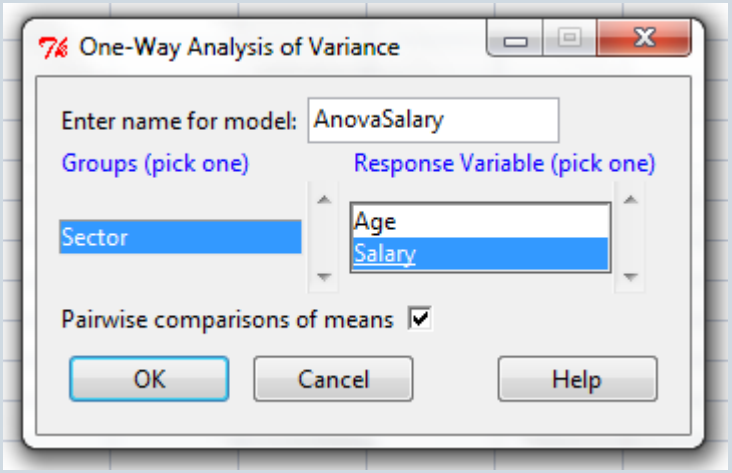


Output in R Commander:

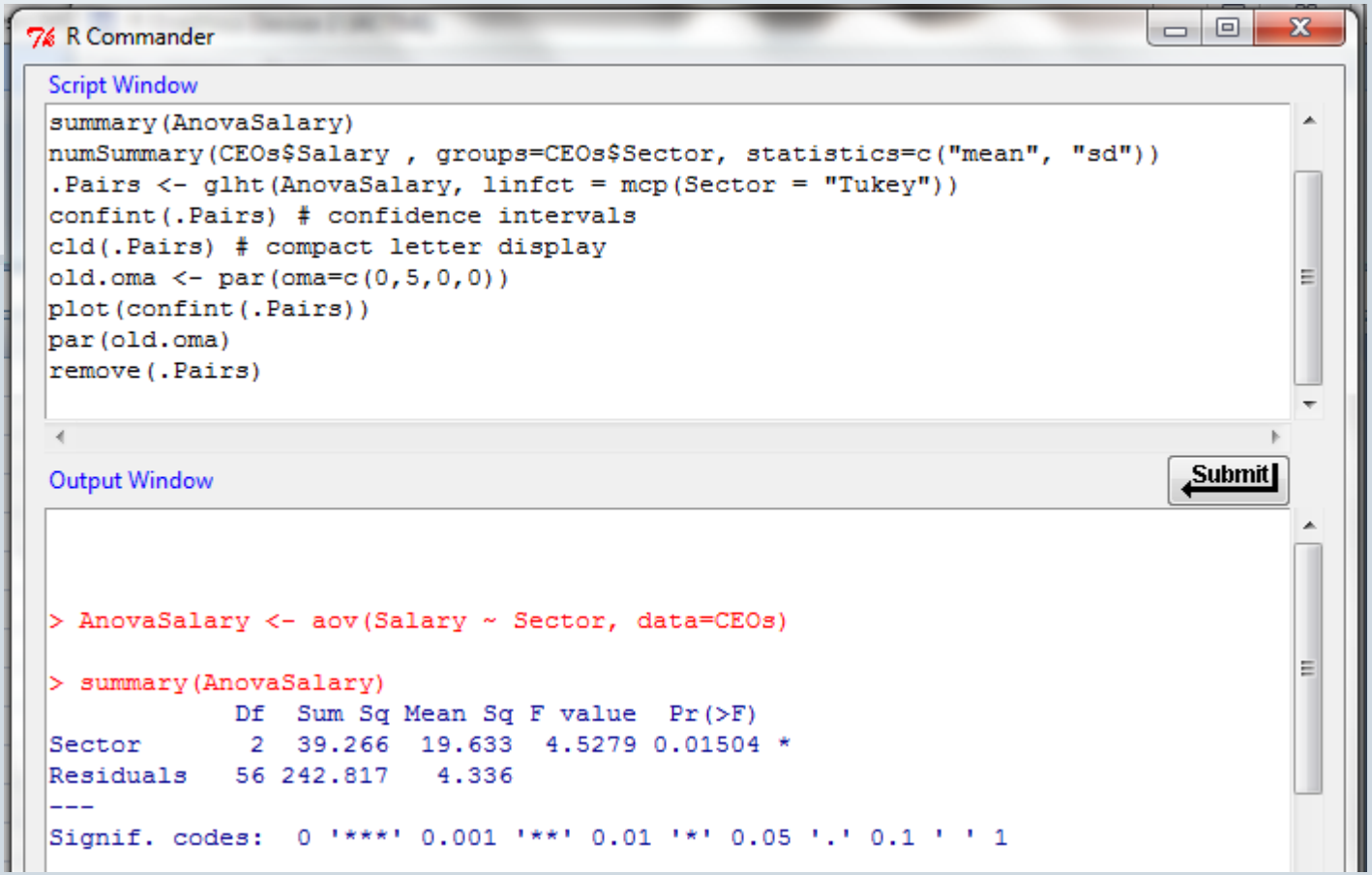


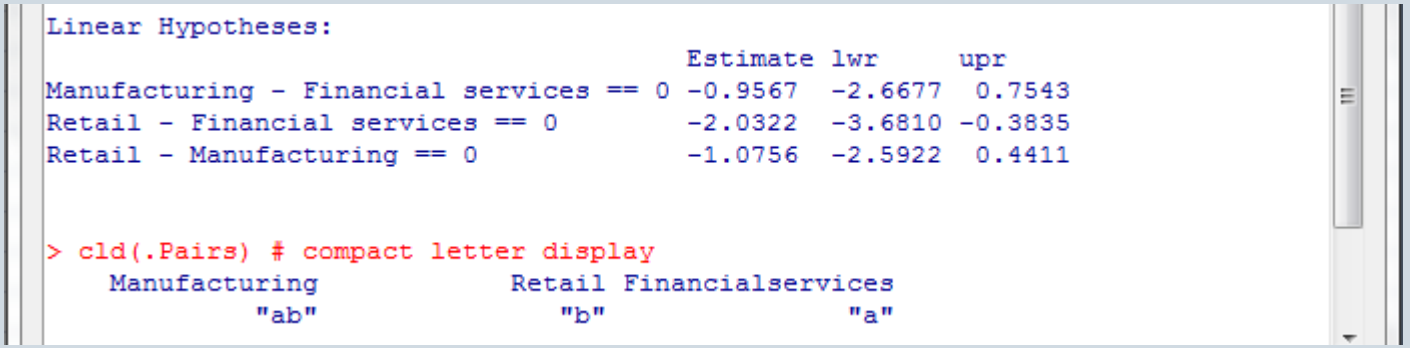
ANOVA of CEO salary by sector:





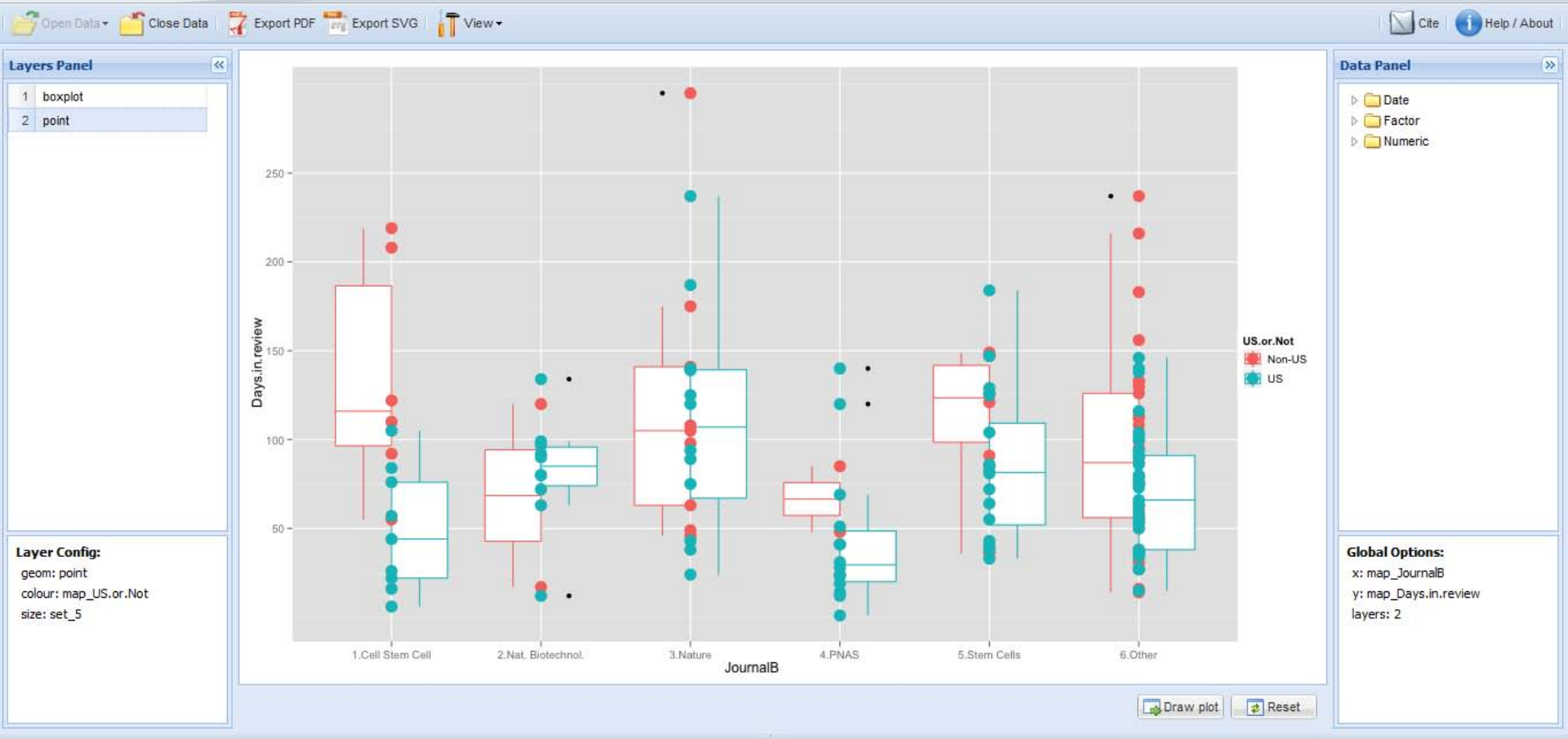
Output in R Commander:





Statistical graphics with R made easy:

Web application for ggplot2 package: http://www.yeroon.net/ggplot2/



R tutorials:

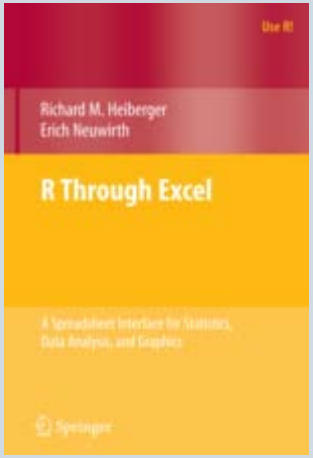
We've only scratched the surface with these examples. To learn how to perform other common statistical analyses in R (e.g. linear and logistic regression; t-tests; tests for normality; non-parametric methods), here are some resources:

- **Quick-R**: intended for people switching from SAS/SPSS/Stata, but easy to follow even if you haven't used these packages

- **Rtips**: comprehensive, but not quite so easy to navigate

- **Burns Statistics**: put together by a specialist in quantitative finance; tutorials on the menu at left

- **R Through Excel**: well-illustrated guide to using the add-in



R for statistics: first steps

CAR meeting, Raleigh, Feb 26 2011

Slides: http://www.peteraldhous.com/CAR/Aldhous_CAR2011_RforStats.pdf

Simulated data used here: http://www.peteraldhous.com/Data/CEOs.csv

Peter Aldhous,
San Francisco Bureau Chief