Introduction to R

BIOINF 525 Lab2-1

W17

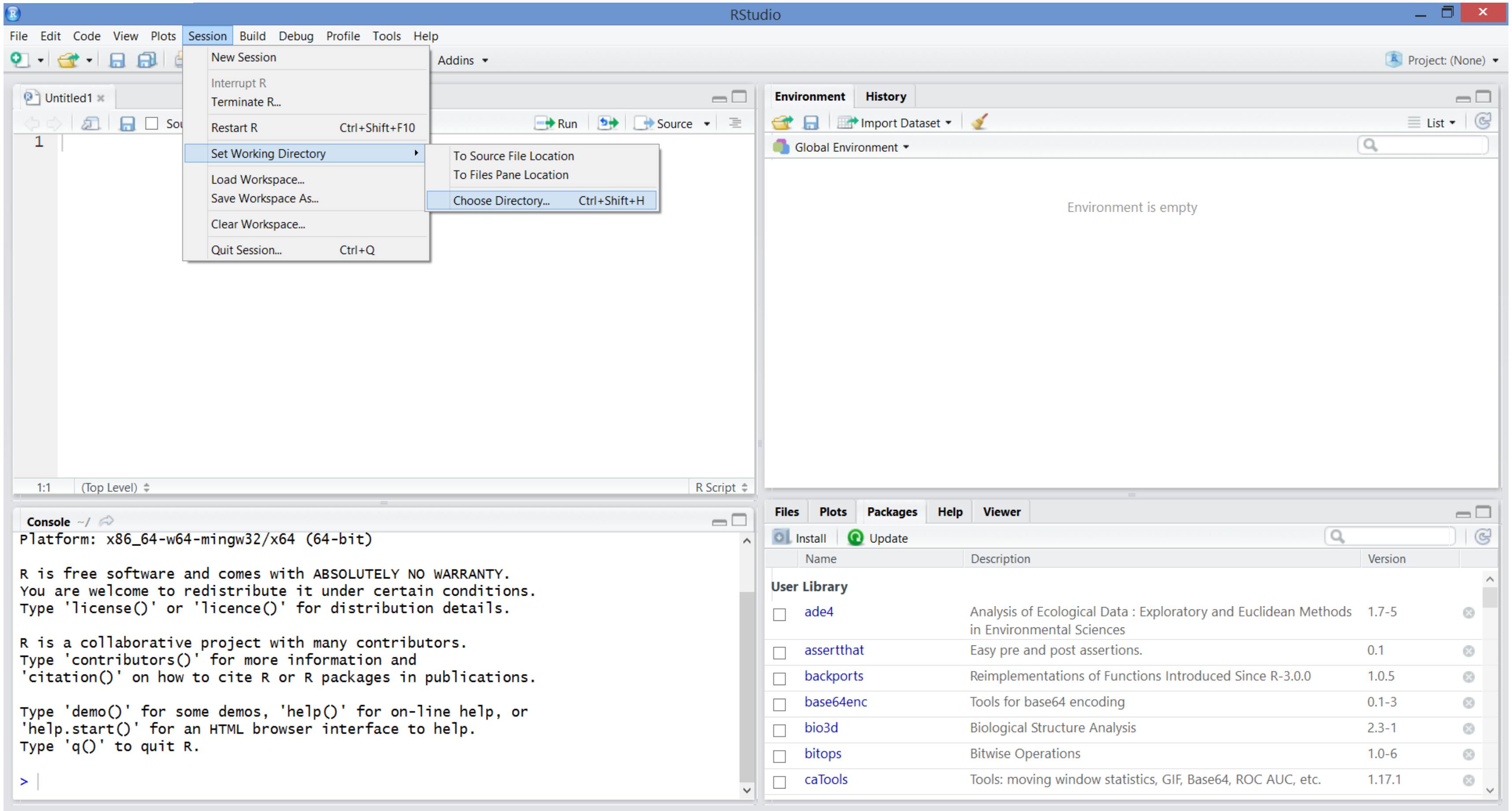
What is R?

- Powerful programming language designed for statistical analysis and graphics.
- R is very popular in the field of bioinformatics.
- Available for Mac, Windows and Unix.
- Free but commercial quality.

- www.r-project.org
- www.rstudio.com

Preparation

- Open Rstudio. Click on **Session > Set Working Directory > Choose Directory**

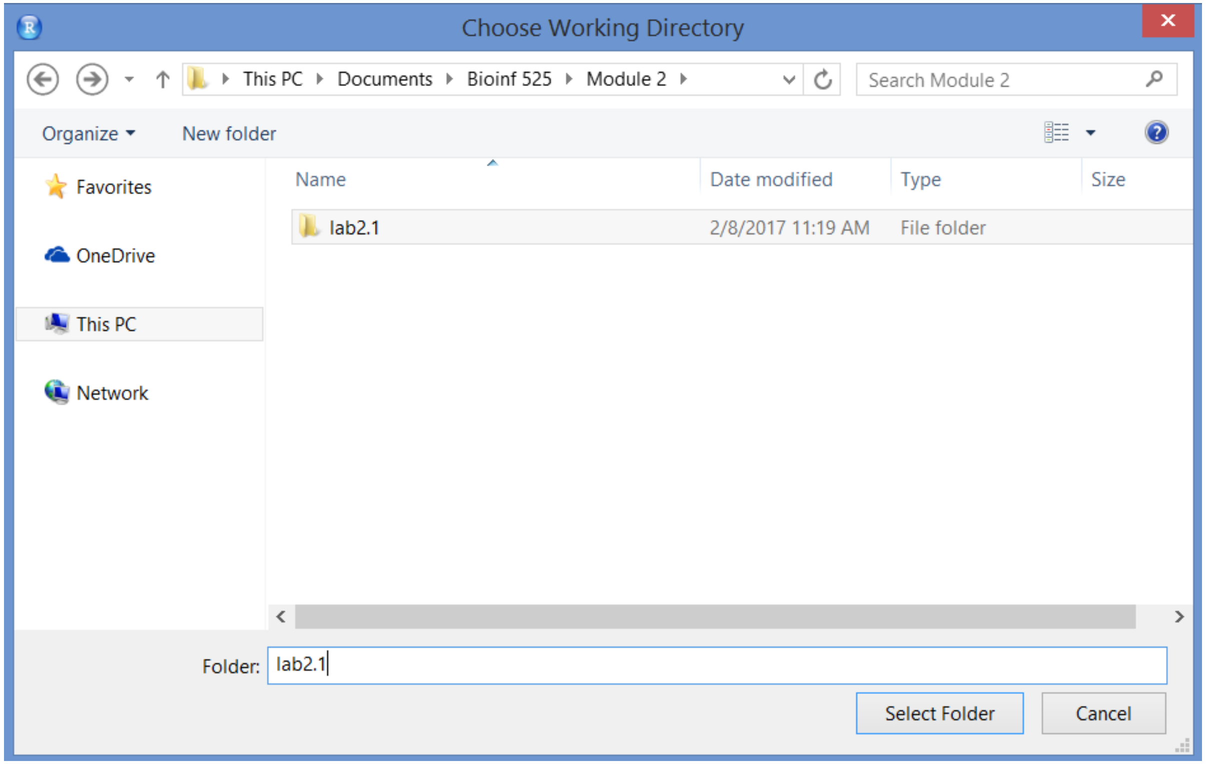


```
Platform: x86_64-w64-mingw32/x64 (64-bit)

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.

>
```

	Name	Description	Version
	User Library		
□	ade4	Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences	1.7-5
□	assertthat	Easy pre and post assertions.	0.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.0.5
□	base64enc	Tools for base64 encoding	0.1-3
□	bio3d	Biological Structure Analysis	2.3-1
□	bitops	Bitwise Operations	1.0-6
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1

Preparation

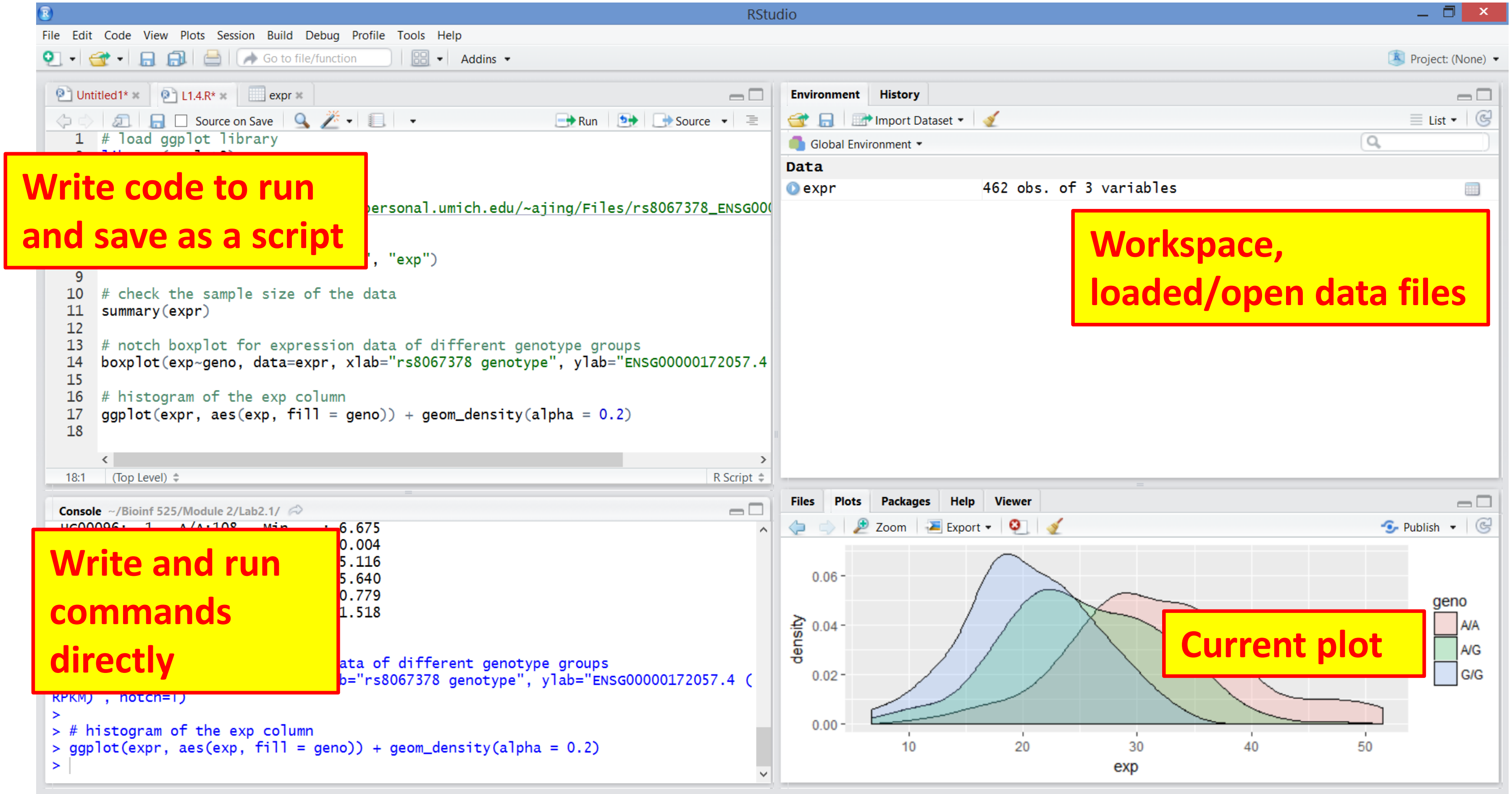
- Open Rstudio. Click on **Session > Set Working Directory > Choose Directory**
- Make and enter a lab2.1 folder



Name	Date modified	Type	Size
lab2.1	2/8/2017 11:19 AM	File folder	

Folder: lab2.1

RStudio



```
# load ggplot library
ersonal.umich.edu/~ajing/Files/rs8067378_ENSG00(
', "exp")

# check the sample size of the data
summary(expr)

# notch boxplot for expression data of different genotype groups
boxplot(exp~geno, data=expr, xlab="rs8067378 genotype", ylab="ENSG00000172057.4

# histogram of the exp column
ggplot(expr, aes(exp, fill = geno)) + geom_density(alpha = 0.2)
```

```
6.675
0.004
5.116
5.640
0.779
1.518

ata of different genotype groups
b="rs8067378 genotype", ylab="ENSG00000172057.4 (
RPKM)", notch=T)
>
> # histogram of the exp column
> ggplot(expr, aes(exp, fill = geno)) + geom_density(alpha = 0.2)
>
```

Data

expr	462 obs. of 3 variables

Write code to run and save as a script

Workspace, loaded/open data files

Write and run commands directly

Current plot

As a calculator

```
> # This is a comment (begins with hash #)
> 10+10
> 10*10
> 10**10
> 10^10

> # Follows order of operations:
> 10+10/2
> (10+10)/2

> # Comparisons
> # logical values TRUE or FALSE
> 10 == 5
> 10 != 5
> 10 > 5
> 10 <= 5
```

Saving variables

```
> a <- 10+10

OR

> a = 10+10

OR

> 10+10 -> a

> a
[1] 20
```

What happens if you type A instead of a?

Vectors

```
> b <- c(1,2,3,4,5)

> b2 <- 1:5

> d <- seq(10,50,by=10)

> e <- rep(0,10)
```

What do b, b2, d, and e look like?

Getting help

```
> # if you know the function

> help(mean)
```

OR

```
> ?mean
```

```
> # if you don’t know the function

> help.search(“variance”)
```

OR

```
> ??variance
```

More Vectors

```
> # logical vectors
> d > 30


> # You can extract part of a vector
> d[1:2]

> d[d>30]

> # Everything except for a selection
> d[-(2:4)]
```

More Vectors

```
> # character/string vectors

> words=c(‘a’, ‘second’, ‘asdf’)

> words[1]

> words[2]
```

Matrices

```
> cbind(b,d)
```

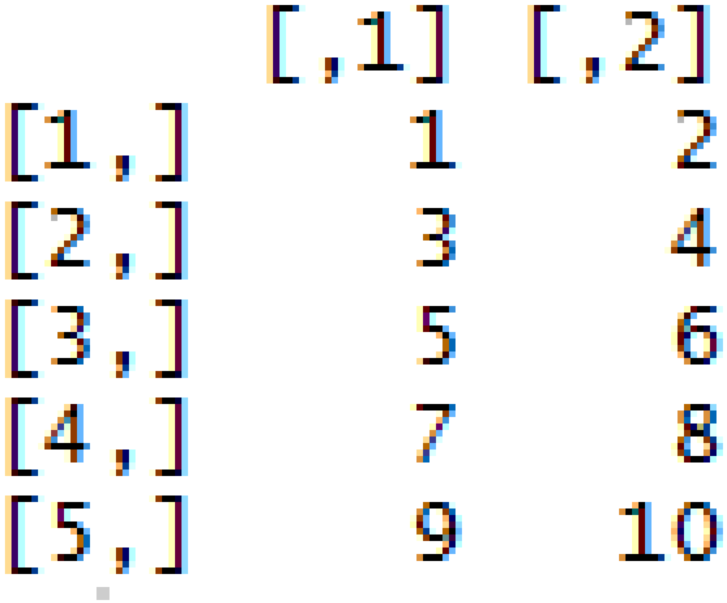
How would you concatenate rows instead of columns?

```
> mat1=matrix(1:10, nrow=5, ncol=2,
byrow=TRUE)
> mat2=matrix(1:10, nrow=5, ncol=2,
byrow=FALSE)
```

What’s the difference between these two?

More Matrices

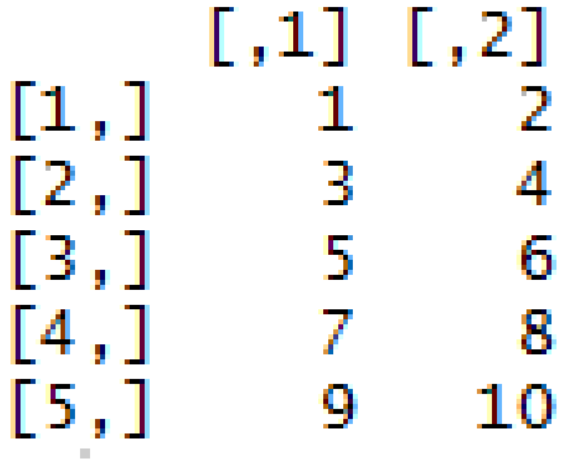
```
> # Just like vectors, we can select part of a matrix

> # What do each of these do?

> mat1[1,2]

> mat1[2:3,1]

> mat1[,1]

> mat1[2,]

> mat1[mat1>3]
```



```
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
[4,]    7    8
[5,]    9   10
```

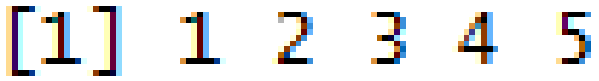
Vector and Matrix Math

```
> mat1
> t(mat1)
> mat1+mat1
> mat1 + 1
> mat1 * 4
```



```
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
[4,]    7    8
[5,]    9   10
```

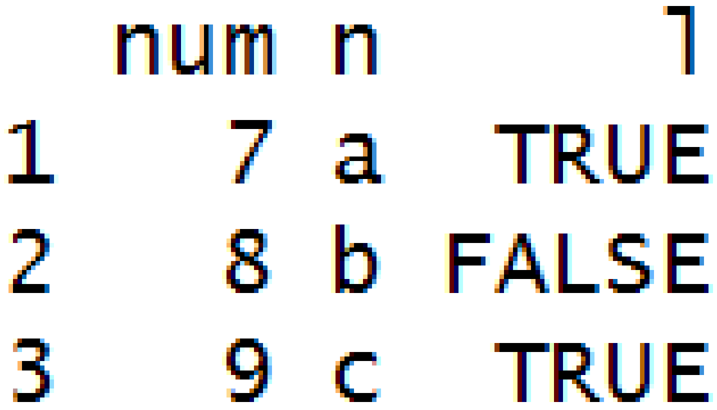
```
> b*b
> mat1*mat1
```



```
[1] 1 2 3 4 5
```

data.frames

```
> # A data frame is a list of vectors of the same length

> num = c(7,8,9)
> n = c("a","b","c")
> l = c(TRUE,FALSE,TRUE)
> df = data.frame(num,n,l)

> df[2,]
> df[,"n"]
```



```
  num n     l
1   7 a  TRUE
2   8 b FALSE
3   9 c  TRUE
```

Generating normal random numbers

```
> rand=rnorm(10,mean=10,sd=4)

> mean(rand)
> sd(rand)
```

Do these match our inputs?

```
> min(rand)
> max(rand)
```

What are some other distributions?

How could we sample from those instead?

Other Distributions

```
> runif(n, min, max)
> rpois(n, lambda)
> rbinom(n, size, prob)

> # Many more
> ?Distributions
```

Opening a file

What if we want to work with outside data?

read.csv, read.delim, read.table

```
> a = read.csv(filename,header,sep)
```

The filename is a string indicating which file to open (including folders, etc.)

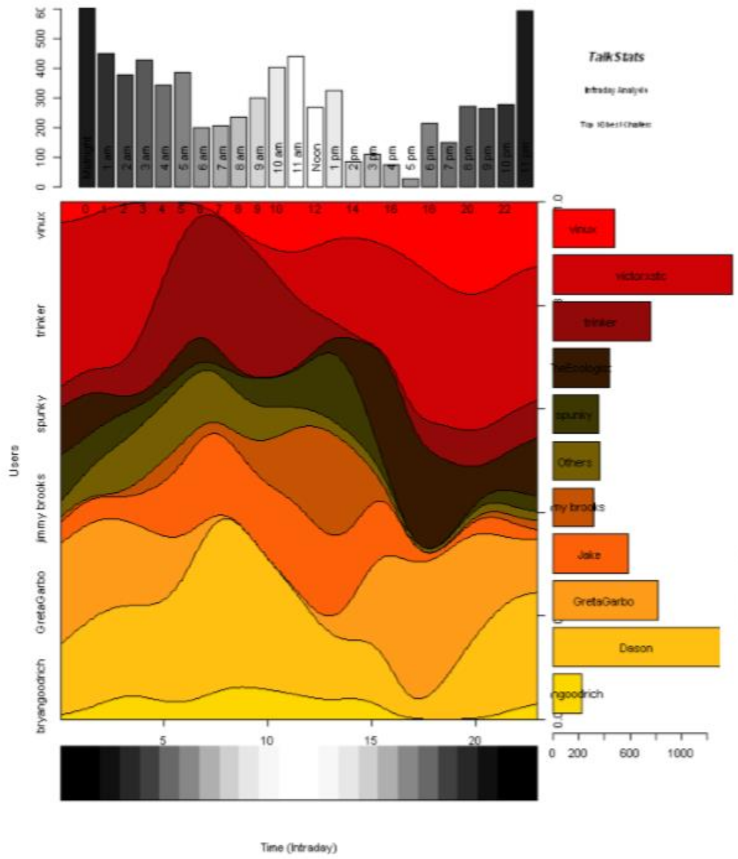
Example filename: “~/Downloads/TROPHY.csv”

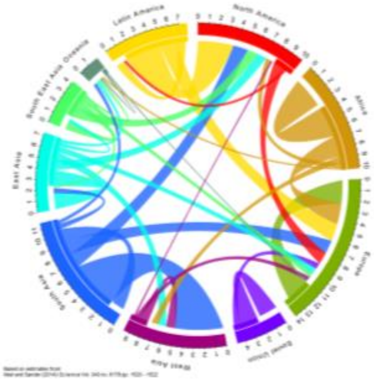
For read.csv(), what do header and sep do?

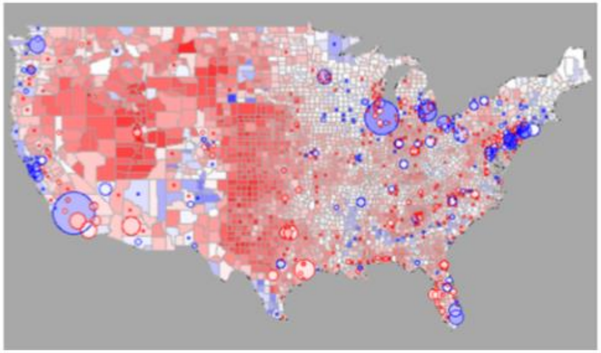
Plotting

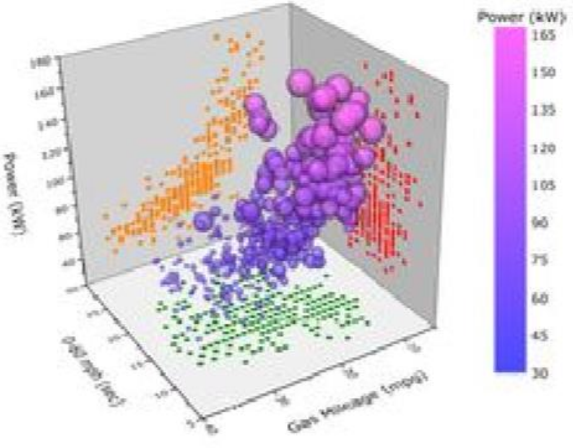
```
> # Create 2 vectors of 1000 random
> # numbers from a normal distribution
> # with mean of 0 and sd of 10

> rand1=rnorm(1000,mean=0,sd=10)
> rand2=rnorm(1000,mean=0,sd=10)
> plot(rand1,rand2)
> hist(rand1)
> boxplot(rand1)
```

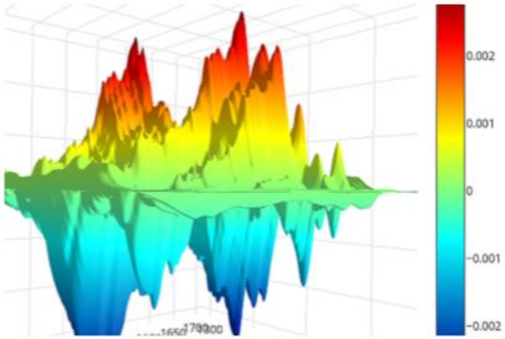
Virtually endless plotting capabilities….











Tutorials and references

- http://cran.r-project.org/doc/manuals/R-intro.html
- http://www.statmethods.net/
- http://bioinformatics.knowledgeblog.org/2011/06/21/using-r-a-guide-for-complete-beginners/
- http://www.cyclismo.org/tutorial/R/
- Many, many more on the internet.