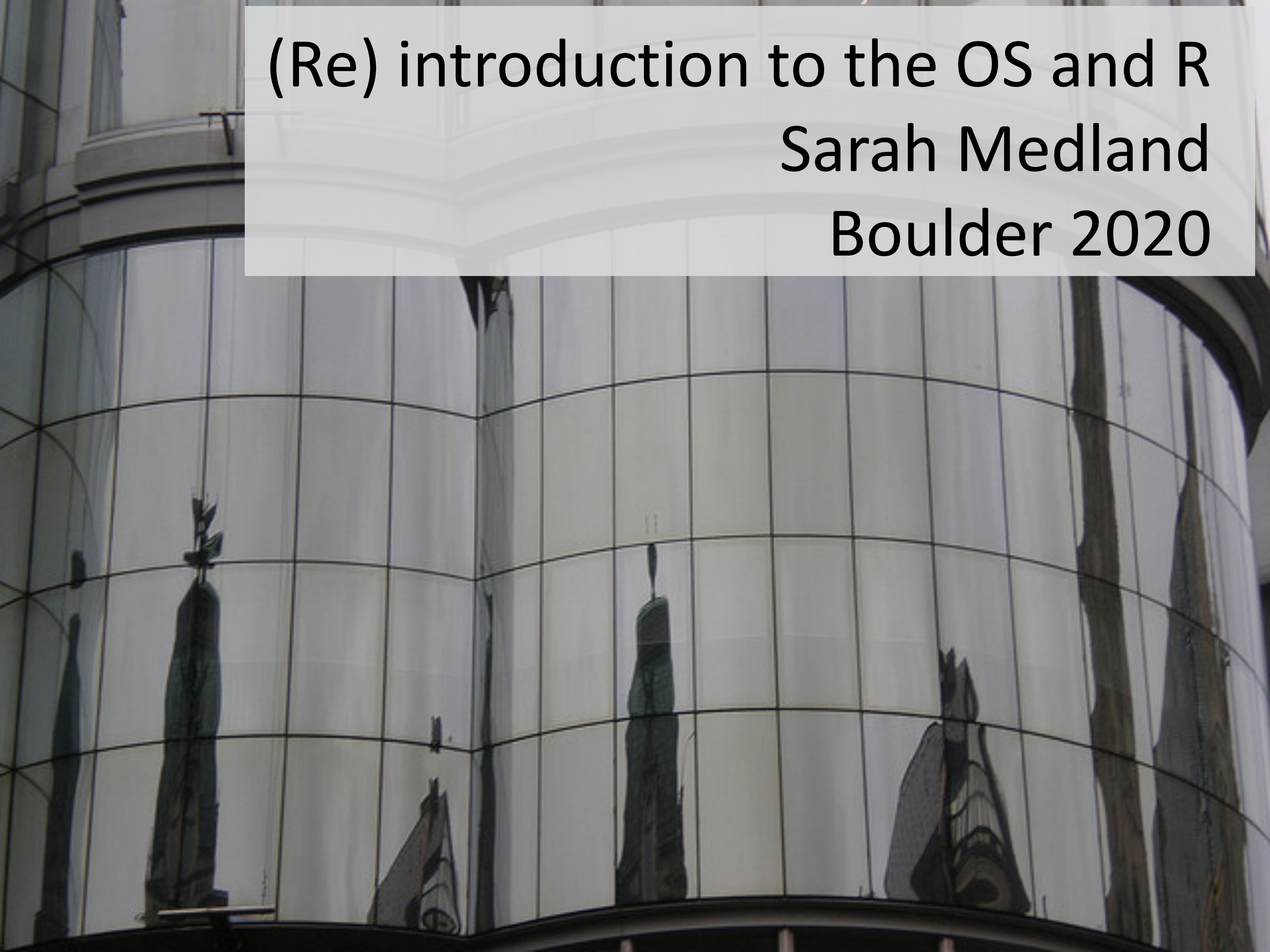
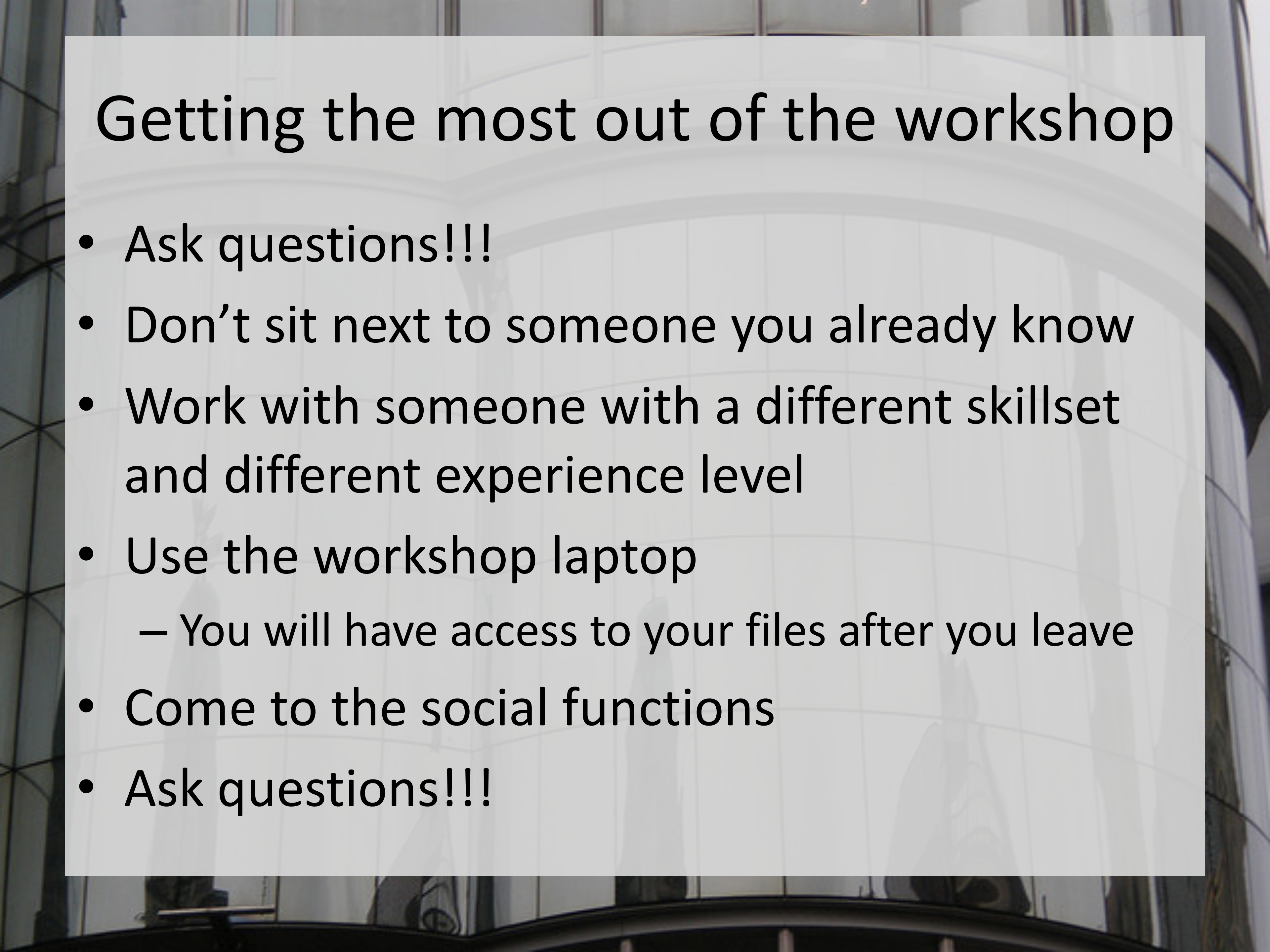


(Re) introduction to the OS and R

Sarah Medland

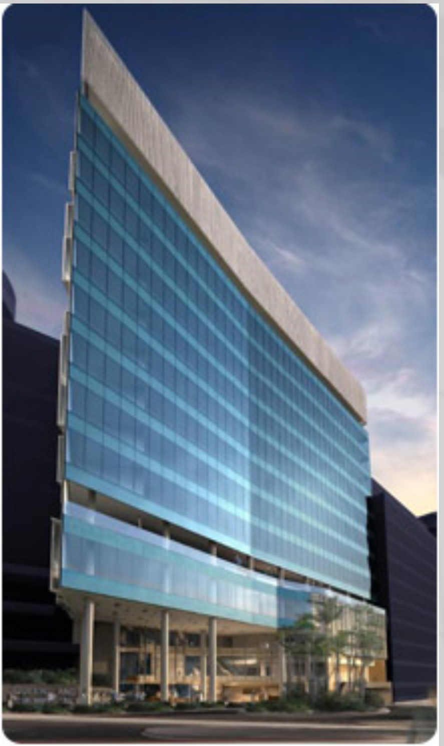
Boulder 2020

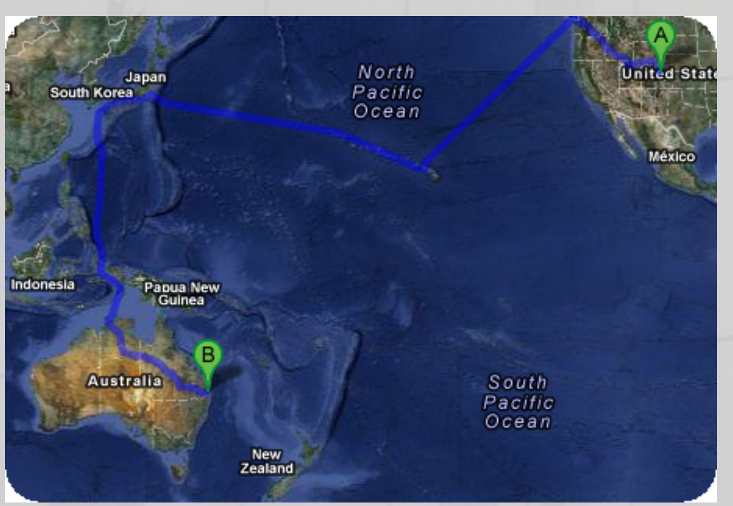


Getting the most out of the workshop

- Ask questions!!!
- Don’t sit next to someone you already know
- Work with someone with a different skillset and different experience level
- Use the workshop laptop
  - You will have access to your files after you leave
- Come to the social functions
- Ask questions!!!

I work in Brisbane at QIMR





Sarah Medland

not

獼猴桃

猕猴桃

Morning sessions

- Optional
  - Feel free to wander in and out/check email etc
- Topics
  - Shift in response to feedback
  - Tomorrow: Plink and Genetic relatedness
  - Wednesday: GWAS and LDscore
  - Thursday: Polygenic Risk Scores
  - Friday: Modelling challenges and solutions

Superfast intro to Linux

This year's OS

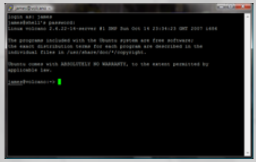
- Debian (linux)
  - Free
  - Many free software packages available
    - Open office
    - R
    - PSPP
    - Terminal
- Based on Unix
  - long and venerable history
  - http://en.wikipedia.org/wiki/Unix

Close but not the same…

- Most basic shortcuts will work
  - crtl+C for copy crtl+V for paste etc
- Supports folder based navigation
- /\ BIG PROBLEM is \ vs /
- You will have used some version of unix previously

Learning the Mac OS X Terminal

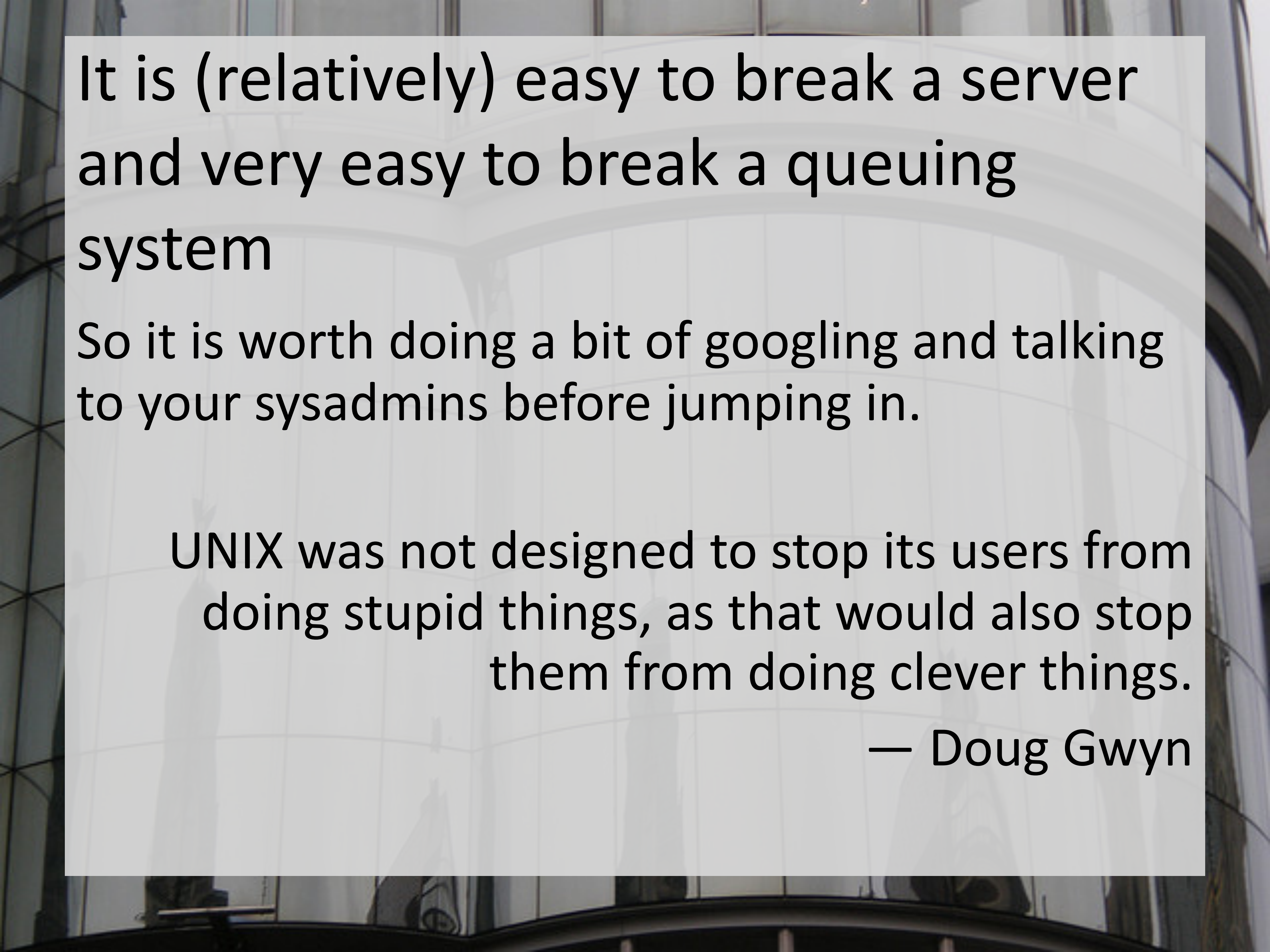
MAC



```
login as: james
james@shell's password:
Linux volcano 2.6.22-14-server #1 SMP Sun Oct 14 23:34:23 GMT 2007 i686

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

james@volcano:~>
```

File hygiene is very important

- Files are stored in Unix format not DOS or Mac
  - Changes the line ending characters
  - Use dos2unix, unix2dos, mac2unix, unix2mac to change formats
  - Can use the file command to check format
- Unix systems are case sensitive!
- NO SPACES in your file/directory names!!
- Wildcards ie dos2unix *.dat



It is (relatively) easy to break a server and very easy to break a queuing system

So it is worth doing a bit of googling and talking to your sysadmins before jumping in.

> UNIX was not designed to stop its users from doing stupid things, as that would also stop them from doing clever things.
>
> — Doug Gwyn

Superfast intro to R

What is it?

- R is an interpreted computer language.
  - System commands can be called from within R

- R is used for data manipulation, statistics, and graphics. It is made up of:
  - operators (+ - <- * %*% …) for calculations on arrays & matrices
  - large, coherent, integrated collection of functions
  - facilities for making unlimited types of publication quality graphics
  - user written functions & sets of functions (packages); 800+ contributed packages so far & growing

Advantages

- Fast** and free.
- State of the art: Statistical researchers provide their methods as R packages. SPSS and SAS are years behind R!
- great graphics.
- Active user community
- Excellent for simulation, programming, computer intensive analyses, etc.
- Forces you to *think* about your analysis.
- Interfaces with database storage software (SQL)

Disadvantages

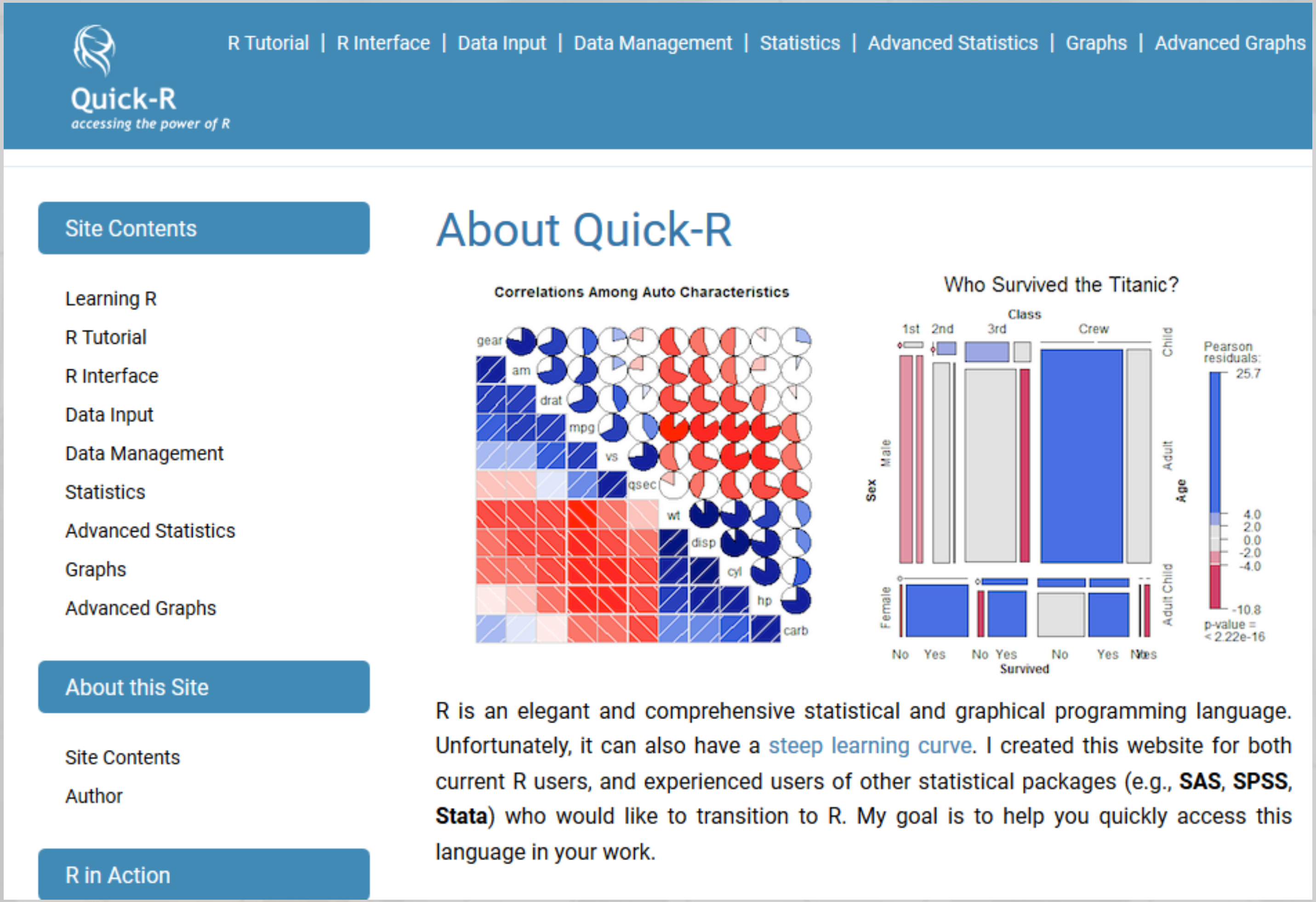
- Not user friendly @ start - steep learning curve, minimal GUI.
- No commercial support; figuring out correct methods or how to use a function on your own can be frustrating.
- Easy to make mistakes and not know.
- Working with large datasets is limited by RAM!!!
- Data prep & cleaning can be messier & more mistake prone in R vs. SPSS or SAS
- Hostility on the R listserve

Learning R....

R-help listserve….

- *Once you appreciate that you have seriously misread the page, things will become a lot clearer. (2005)*
- *You will need to do your homework a lot more carefully, as it seems you don't have enough knowledge to recognise the errors you are making.* (2007)
- *Well, don't try to use a Makefile as you do not know what you are doing. (2013)*
- *It is user lack-of-understanding: there is no error here. (2013)*

Quick R



Quick-R
accessing the power of R

R Tutorial | R Interface | Data Input | Data Management | Statistics | Advanced Statistics | Graphs | Advanced Graphs

Site Contents

- Learning R
- R Tutorial
- R Interface
- Data Input
- Data Management
- Statistics
- Advanced Statistics
- Graphs
- Advanced Graphs

About this Site

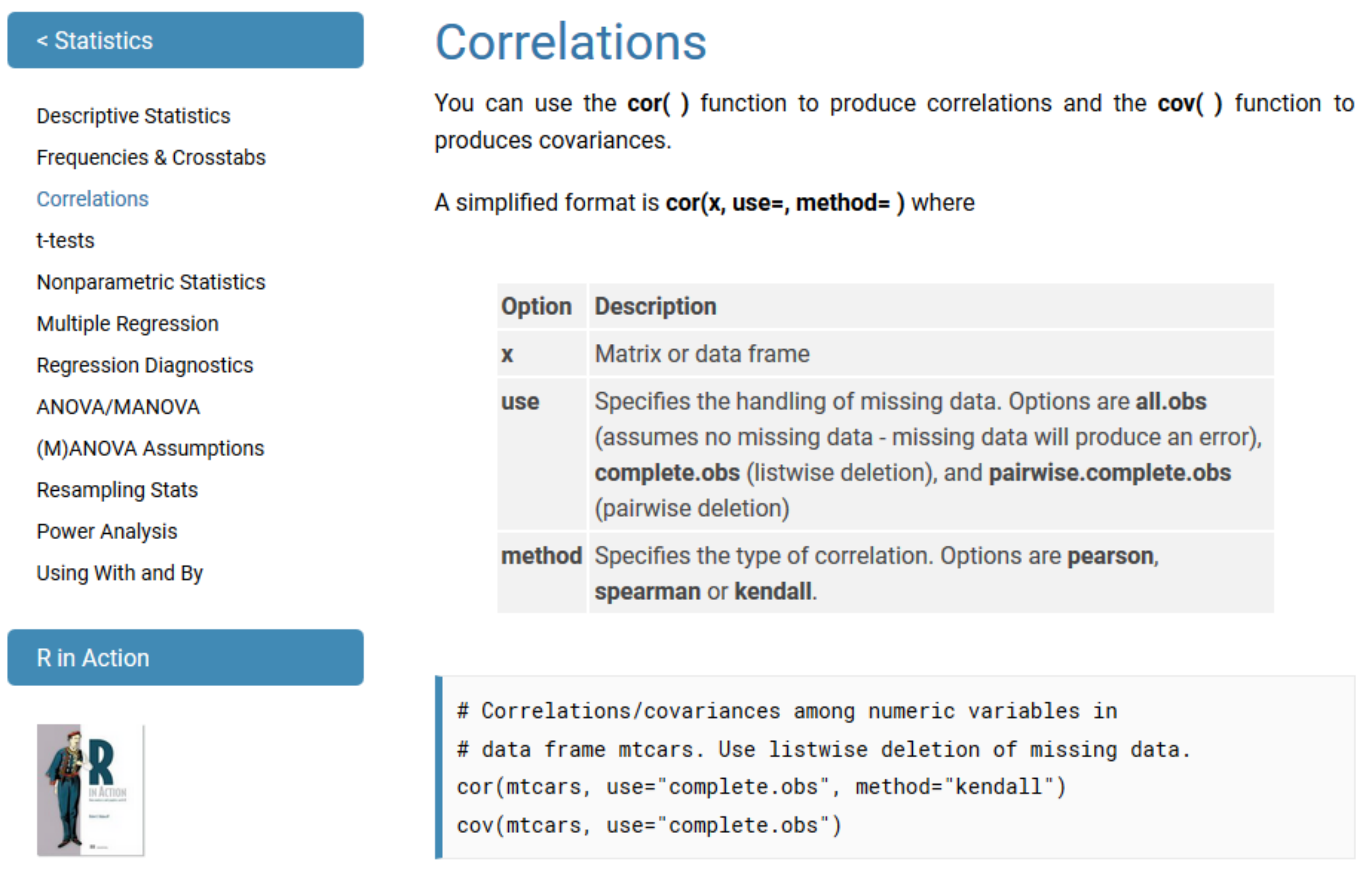
- Site Contents
- Author

R in Action

About Quick-R

R is an elegant and comprehensive statistical and graphical programming language. Unfortunately, it can also have a steep learning curve. I created this website for both current R users, and experienced users of other statistical packages (e.g., **SAS**, **SPSS**, **Stata**) who would like to transition to R. My goal is to help you quickly access this language in your work.

Quick R



Correlations

You can use the **cor( )** function to produce correlations and the **cov( )** function to produces covariances.

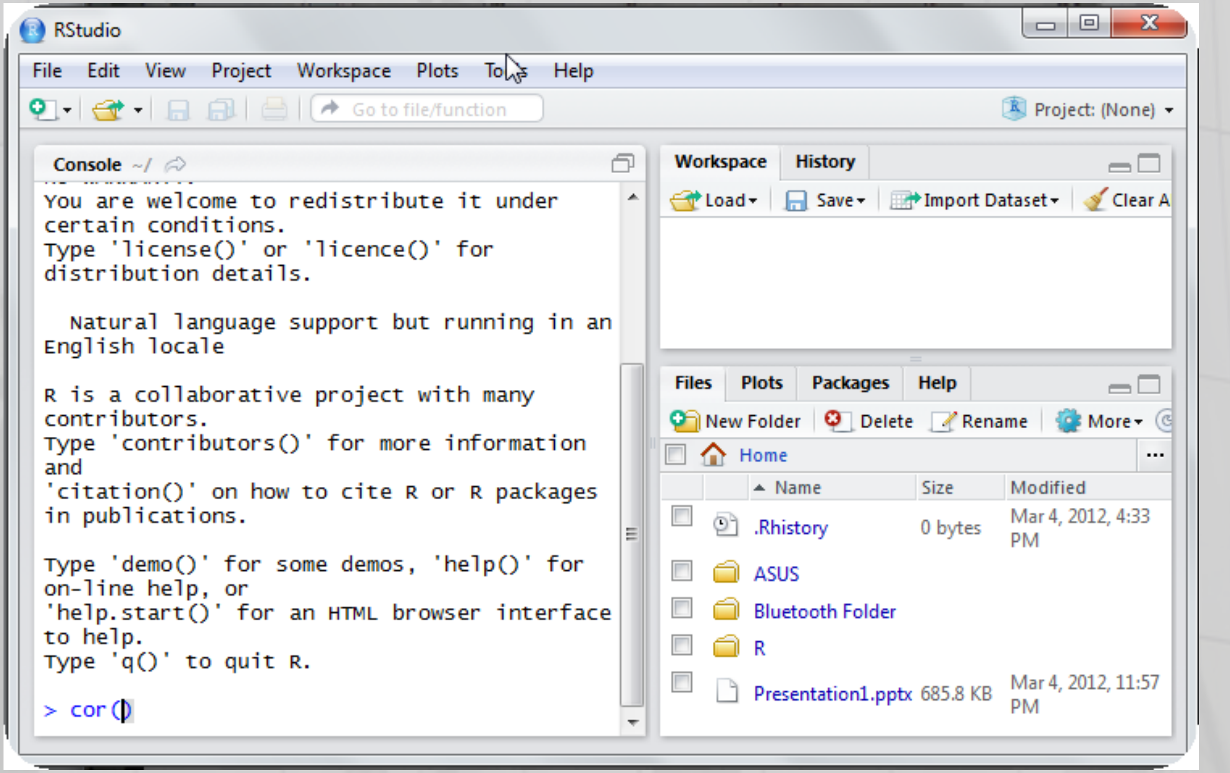
A simplified format is **cor(x, use=, method=)** where

Option	Description
x	Matrix or data frame
use	Specifies the handling of missing data. Options are **all.obs** (assumes no missing data - missing data will produce an error), **complete.obs** (listwise deletion), and **pairwise.complete.obs** (pairwise deletion)
method	Specifies the type of correlation. Options are **pearson**, **spearman** or **kendall**.

```
# Correlations/covariances among numeric variables in
# data frame mtcars. Use listwise deletion of missing data.
cor(mtcars, use="complete.obs", method="kendall")
cov(mtcars, use="complete.obs")
```

Using R this week

- R-studio http://rstudio.org/

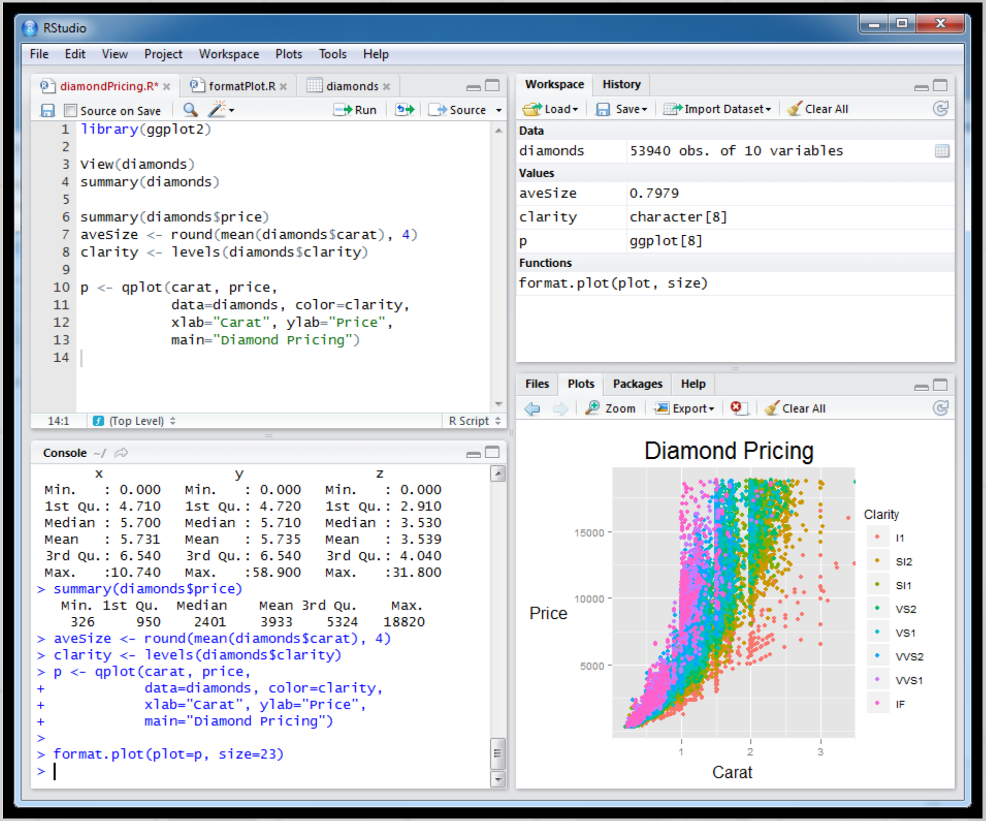


```
You are welcome to redistribute it under
certain conditions.
Type 'license()' or 'licence()' for
distribution details.

  Natural language support but running in an
English locale

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.

> cor()
```

Name	Size	Modified
.Rhistory	0 bytes	Mar 4, 2012, 4:33 PM
ASUS		
Bluetooth Folder		
R		
Presentation1.pptx	685.8 KB	Mar 4, 2012, 11:57 PM

Setting this up at home

- Install R first
- Install R studio
- Install packages



```r
library(ggplot2)

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")

```

```
       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")
>
> format.plot(plot=p, size=23)
> 
```

Data

diamonds	53940 obs. of 10 variables

Values

aveSize	0.7979
clarity	character[8]
p	ggplot[8]

Functions

format.plot(plot, size)

Start up R via R studio

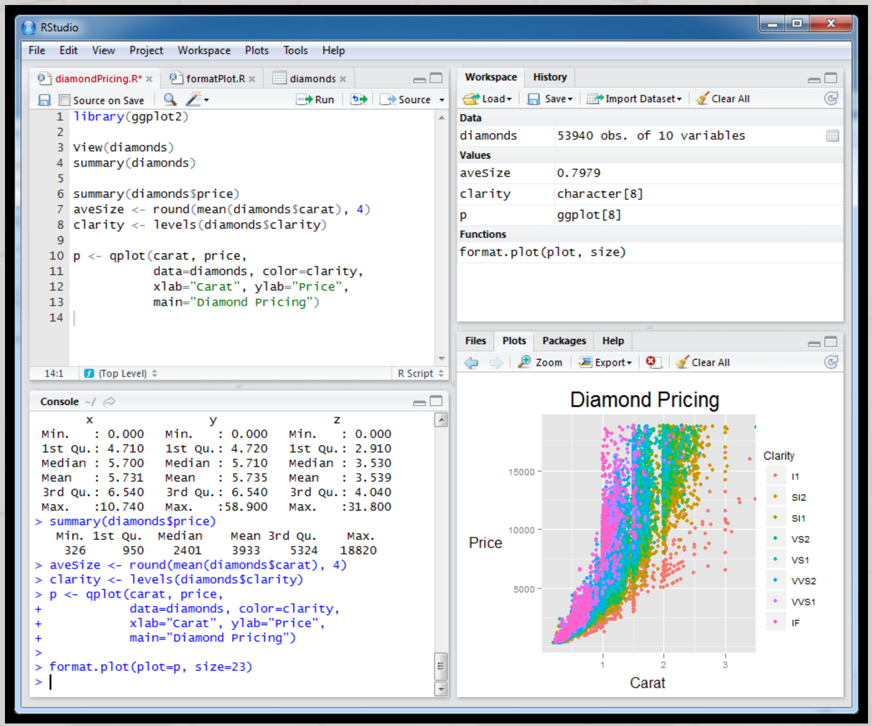
4 windows:

Syntax – can be opened in regular txt file - saved

Terminal – output & temporary input - usually unsaved

Data manager – details of data sets and variables

Plots etc

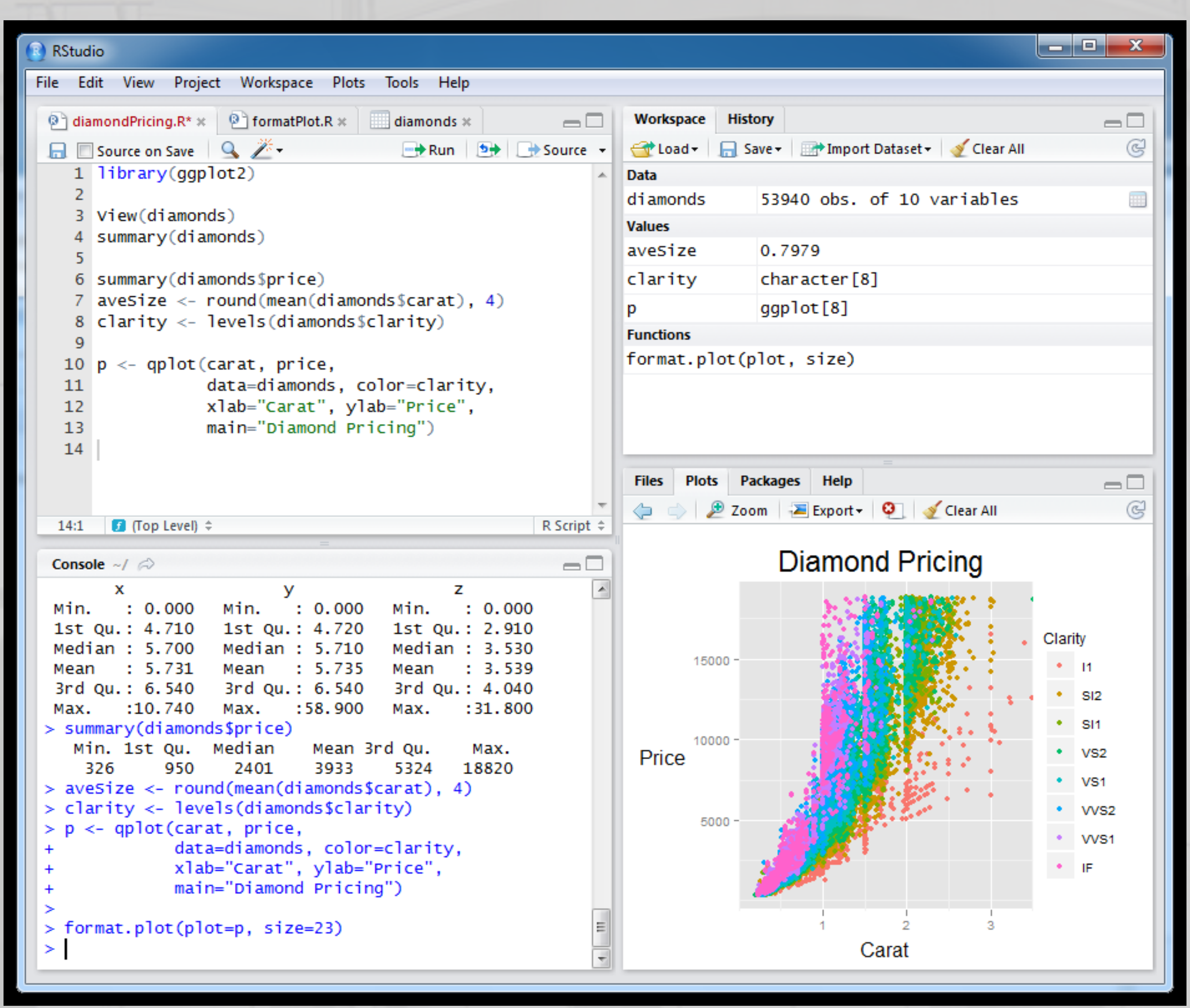


```
library(ggplot2)

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")

```

```
      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")
> 
> format.plot(plot=p, size=23)
> 
```

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

R sessions are *interactive*



```
library(ggplot2)

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")

```

```
       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")
> 
> format.plot(plot=p, size=23)
> 
```

Data

diamonds	53940 obs. of 10 variables

Values

aveSize	0.7979
clarity	character[8]
p	ggplot[8]

Functions

format.plot(plot, size)

Final Words of Warning

“Using R is a bit akin to smoking. The beginning is difficult, one may get headaches and even gag the first few times. But in the long run, it becomes pleasurable and even addictive. Yet, deep down, for those willing to be honest, there is something not fully healthy in it.” --Francois Pinard

GETTING STARTED

How to use help in R?

- R has a help system built in.
- If you know which function you want help with simply use ?_________ or help(_______) with the function in the blank.
  - ?hist.
  - help(hist)
- If you don’t know which function to use, then use help.search(“_________”).
  - help.search(“histogram”).

Importing Data

First make sure your data is in an easy to read format such as space, tab or CSV

Use code:

```
D <- read.table(“ozbmi2.txt”,header=TRUE)
D <-read.table(“ozbmi2.txt”,na.strings=“-
  99”,header=TRUE)
D <- read.table(“ozbmi2.csv”, sep=“,”
header=TRUE)
D <- read.csv(“ozbmi2.csv”, header=TRUE)
```

Exporting Data

Tab delimited

```
write.table(D, “newdata.txt”,sep=“\t”)
```

csv

```
write.csv(D, “newdata.csv”)
```

Other options include writing to xls, spss, sas and other formats

Checking data

#list the variables in D

```
names(D)
```

# dimensions of D

```
dim(D)
```

# print the first 10 rows of D

```
head(D, n=10)
```

#referring to variables in D

#format is Object$variable

```
head(D$age, n=10)
```

Basic Manipulation

#You can make new variables within an existing object

```
D$newage<- D$age*100
```

#Or overwrite a variable

```
D$age<- D$age*100
```

#Or recode a variable

```
#D$catage <- ifelse(D$age > 30,
c("older"), c("younger"))
```

Describing data

```
#Mean and variance
  mean(D$age, na.rm =TRUE)
  var(D$age , na.rm =TRUE)
```

Describing data

A bit more info

```
summary(D$age)
summary(D$age[which(D$agecat==1)])
```

What about a categorical variable

```
table(D$agecat)
table(D$agecat,D$zyg)
```

Some basic analysis

Correlations anyone?

```
cor(D$wt1,D$bmi1, use="complete")
cor(D$ht1,D$bmi1, use="complete")
```

Basic plots

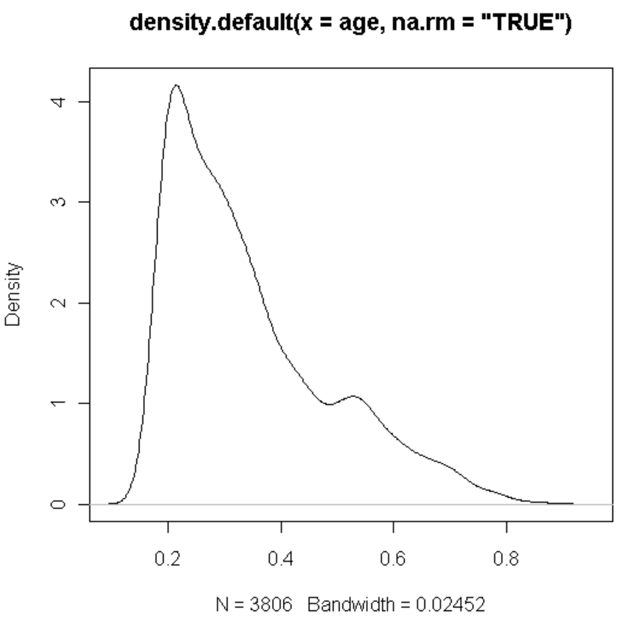
Histogram

```
#basic
  hist(D$age)
#basic
  hist(D$age, breaks=12, col=‘red’)
# Add labels
  hist(D$age, breaks=12, col='red', xlab='age in
  years',main='Histogram of age‘)
```

Looking at your data…

#Kernal density plot

```
density(D$age, na.rm = "TRUE")
# returns the density data
```



Looking at your data…

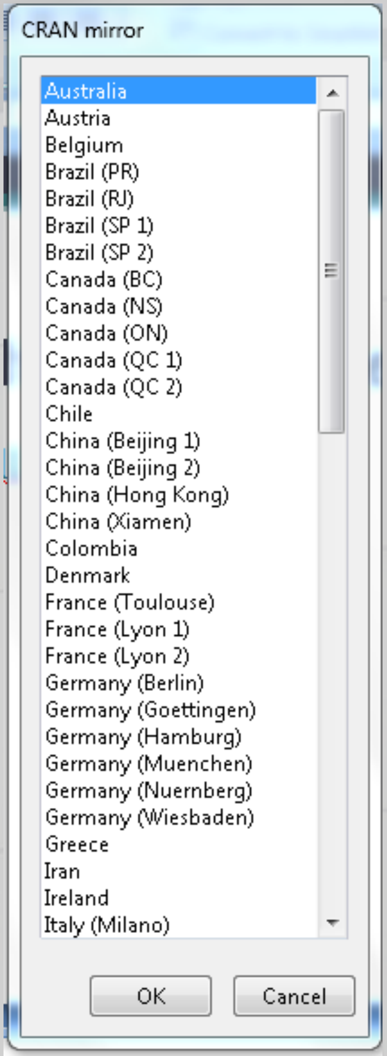
```
#Kernal density plot by zyg?
library(sm)
attach(D)
# create value labels
zyg.f <- factor(zyg, levels= seq(1,5), labels = c("MZF", "MZM", "DZF", "DZM", "DZOS"))

# plot densities
sm.density.compare(age, zyg, xlab="Years")
title(main="Years by ZYG")

# add legend
colfill<-c(2:(2+length(levels(zyg.f))))
legend(.8,3, levels(zyg.f), fill=colfill)
```

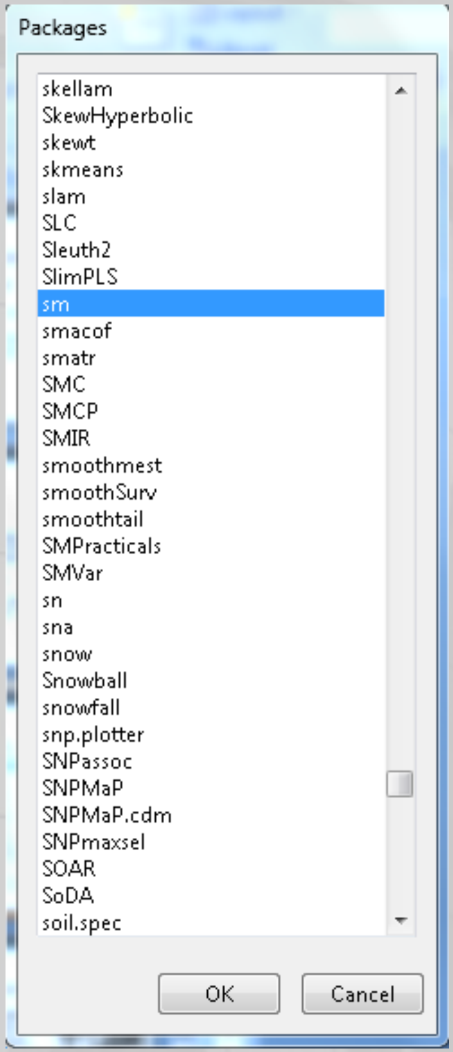
Huh what?

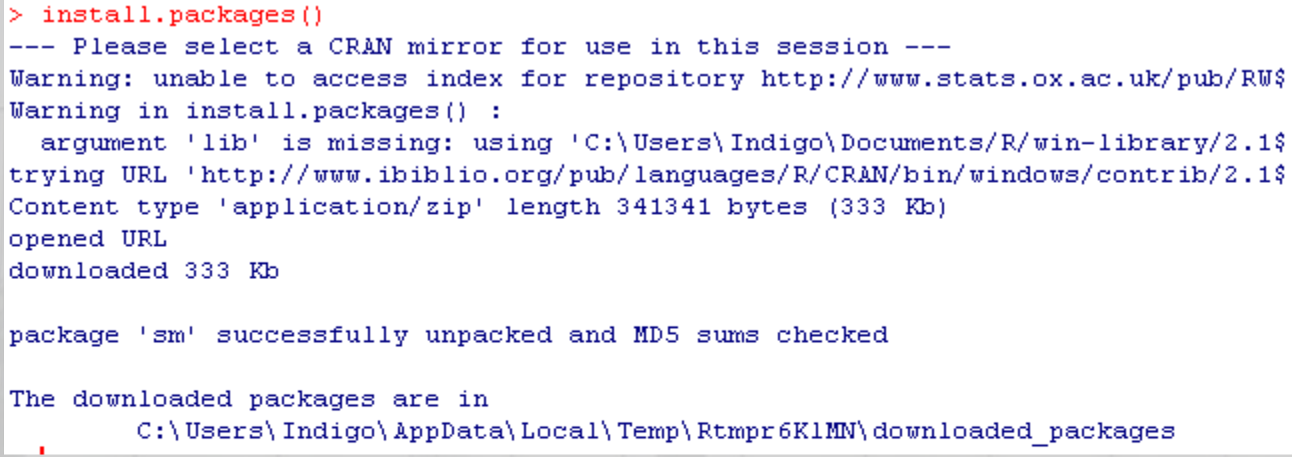
```
> library(sm)
Error in library(sm) : there is no package called 'sm'
> sm.density.compare(age, zyg, xlab="Years")
Error: could not find function "sm.density.compare"
```

Adding a package…

install.packages()





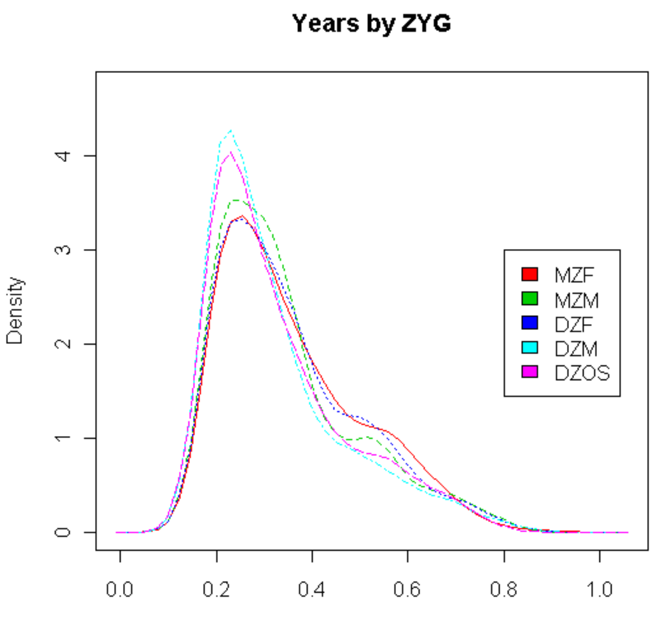


```
> install.packages()
--- Please select a CRAN mirror for use in this session ---
Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RW$
Warning in install.packages() :
  argument 'lib' is missing: using 'C:\Users\Indigo\Documents/R/win-library/2.1$
trying URL 'http://www.ibiblio.org/pub/languages/R/CRAN/bin/windows/contrib/2.1$
Content type 'application/zip' length 341341 bytes (333 Kb)
opened URL
downloaded 333 Kb

package 'sm' successfully unpacked and MD5 sums checked

The downloaded packages are in
        C:\Users\Indigo\AppData\Local\Temp\Rtmpr6K1MN\downloaded_packages
```

Looking at your data...

#Kernal density plot by zyg?

```
library(sm)
attach(D)
# create value labels
zyg.f <- factor(zyg, levels= seq(1,5),
  labels = c("MZF", "MZM", "DZF", "DZM", "DZO

# plot densities
sm.density.compare(age, zyg, xlab="Years”)
title(main="Years by ZYG")

# add legend
colfill<-c(2:(2+length(levels(zyg.f))))
legend(.8,3, levels(zyg.f), fill=colfill)
```


Years by ZYG
Density
🟥 MZF
🟩 MZM
🟦 DZF
🟦 DZM
🟪 DZOS