Analytics Programming
Lecture 01

Nick Tothill

Outline

1. Structure of this unit
2. Introduction to Data Science
3. Introduction to R

Topics

- Data handling
- The basic of R
- Data types
- Data manipulations in R
- R programming
- Simulation using R
- Inputs and outputs
- Graphics
- SQL
- R markdown

Unit Structure

- 12 lectures (12 hours)
- 11 Practicals (22 hours)
- 5 Quizzes (30 min each)
- 1 assignment (3 weeks)
- Final computer Test (1 hour)
- Reference book: "**The Art of R Programming: A Tour of Statistical Software Design**", Norman Matloff, No Starch Press 2011 (*available online in our library!*)
- There are *many* other R books available and many free online resources. Use them!

Data are everywhere[1]

- Sales (supermarkets, front end shops, outlets,. . .)
- Manufacturing (cars, consumable electronics, . . .)
- Web services (Google, Facebook, Twitter, . . .)
- Health (clinics, hospitals, . . .)
- Sciences (Environment, medicine, . . .)
- . . .

[1] 'data' is a plural noun. Write 'data are. . . ', 'data were. . . ', etc.

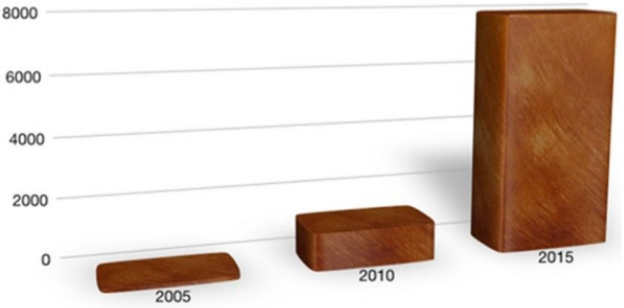
Data generators

- Paper and pencil
- Computer terminals (EFTPOS, ATM, . . .)
- Personal electronics (mobiles, wearable electronics, . . .)
- . . .

Data volume

- B, KB, MB, GB, TB, PB, EB, ZB, YB . . .
- The volume of data is growing at very fast pace

A Decade of Digital Universe Growth: Storage in Exabytes



Source: IDC's Digital Universe Study, sponsored by EMC, June 2011

- Big data: when the volume is large

How to use data?

- Is data useful?
- What data is telling us?
- How to get information from data?
- How to utilise the information?
- Yes, sure.
- Data Science is the solution to these questions.

How to approach the end goal of using data?

- Data collection
- Data storage
- Data manipulation
- Data analysis
- Data visualisation
- . . .

Data collection and storage

- Spreadsheets
- Databases (DB)
- Very large databases (VLDB)
- Data warehouses
- Data cloud
- . . .

Data manipulation

- Select
- Insert
- Update
- Delete

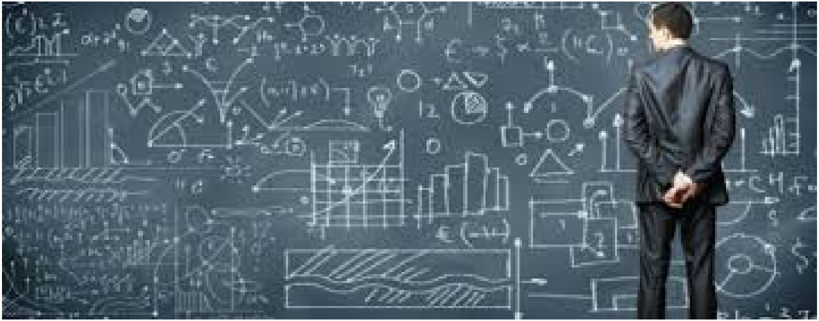
For the purpose of

- cleaning
- transferring
- exploring

data for further analysis

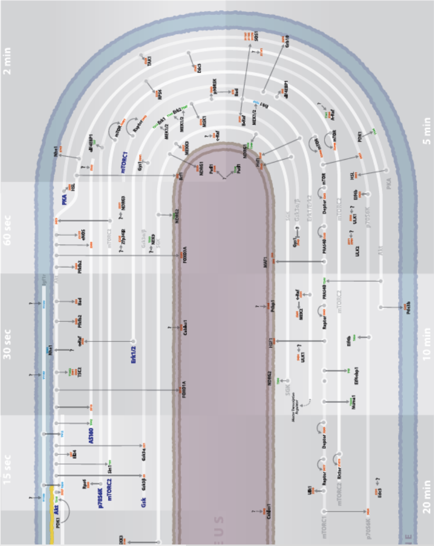
Data analysis

- Statistics
- Machine Learning
- Data mining



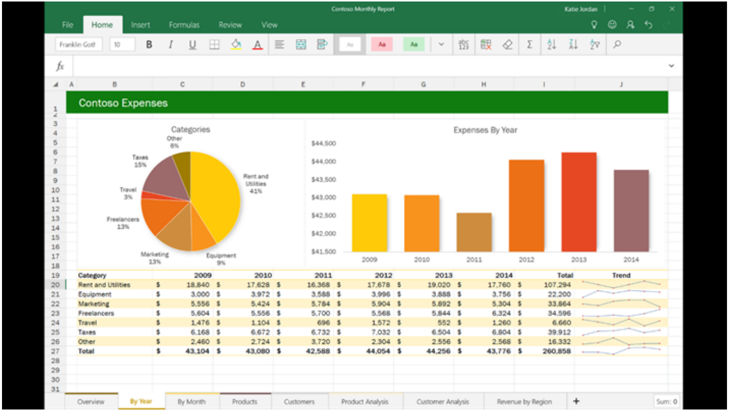
Data visualisation

- Plots
- Graphs
- Others (right: the visualisation of phosphorylation time-series data on insulin response)



Excel

- The starting point for most of applications
- Easy to use (hopefully)
- Some data manipulation and analysis functionalities

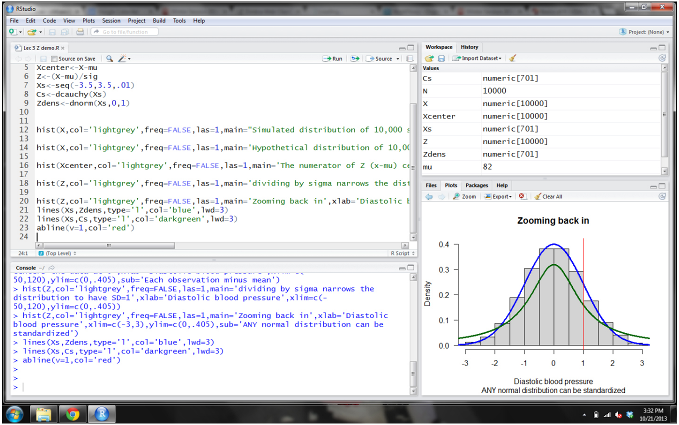


Contoso Expenses

Category	2009	2010	2011	2012	2013	2014	Total	Trend
Rent and Utilities	$ 18,840	$ 17,628	$ 16,368	$ 17,678	$ 19,020	$ 17,760	$ 107,294	
Equipment	$ 3,000	$ 3,972	$ 3,588	$ 3,996	$ 3,888	$ 3,756	$ 22,200	
Marketing	$ 5,556	$ 5,424	$ 5,784	$ 5,904	$ 5,892	$ 5,304	$ 33,864	
Freelancers	$ 5,604	$ 5,556	$ 5,700	$ 5,568	$ 5,844	$ 6,324	$ 34,596	
Travel	$ 1,476	$ 1,104	$ 696	$ 1,572	$ 552	$ 1,260	$ 6,660	
Taxes	$ 6,168	$ 6,672	$ 6,732	$ 7,032	$ 6,504	$ 6,804	$ 39,912	
Other	$ 2,460	$ 2,724	$ 3,720	$ 2,304	$ 2,556	$ 2,568	$ 16,332	
Total	**$ 43,104**	**$ 43,080**	**$ 42,588**	**$ 44,054**	**$ 44,256**	**$ 43,776**	**$ 260,858**	

About R

- A quick introduction to R/RStudio



```r
Xcenter<-X-mu
Z<-(X-mu)/sig
Xs<-seq(-3.5,3.5,.01)
Cs<-dcauchy(Xs)
Zdens<-dnorm(Xs,0,1)


hist(X,col='lightgrey',freq=FALSE,las=1,main="Simulated distribution of 10,000 s

hist(X,col='lightgrey',freq=FALSE,las=1,main='Hypothetical distribution of 10,00

hist(Xcenter,col='lightgrey',freq=FALSE,las=1,main='The numerator of Z (x-mu) ce

hist(Z,col='lightgrey',freq=FALSE,las=1,main='dividing by sigma narrows the dist

hist(Z,col='lightgrey',freq=FALSE,las=1,main='Zooming back in',xlab='Diastolic b
lines(Xs,Zdens,type='l',col='blue',lwd=3)
lines(Xs,Cs,type='l',col='darkgreen',lwd=3)
abline(v=1,col='red')

```

```
50,120),ylim=c(0,.405),sub='Each observation minus mean')
> hist(Z,col='lightgrey',freq=FALSE,las=1,main='dividing by sigma narrows the
distribution to have SD=1',xlab='Diastolic blood pressure',xlim=c(-
50,120),ylim=c(0,.405))
> hist(Z,col='lightgrey',freq=FALSE,las=1,main='Zooming back in',xlab='Diastolic
blood pressure',xlim=c(-3,3),ylim=c(0,.405),sub='ANY normal distribution can be
standardized')
> lines(Xs,Zdens,type='l',col='blue',lwd=3)
> lines(Xs,Cs,type='l',col='darkgreen',lwd=3)
> abline(v=1,col='red')
>
>
> 
```

Cs	numeric[701]
N	10000
X	numeric[10000]
Xcenter	numeric[10000]
Xs	numeric[701]
Z	numeric[10000]
Zdens	numeric[701]
mu	82

What is R?

R is a software environment for statistical computing and graphics. It runs on just about any platform (even for iPad!) and is *completely free* (in the GNU sense).

It is used extensively by academic statisticians for research and teaching and is gaining ground in business.

R is inspired by S by AT&T, which later became S-Plus as a commercial statistical computing software.

Features of R and R programming language

R is original command line based, now several GUI (graphical user interface) are available such as RStudio.

R is an interpreted language while the famous C programming language is an imperative (procedural) language (must be compiled to run). R code can be run in either interactive mode or batch mode.

R has powerful statistics and mathematics functions (will see later) and efficient data handling functionalities.

R programming has similar to C syntax, object-oriented programming (OOP), functional programming capabilities.

R Extensions

It has >**10000** extension packages available!

Pros Its free and open source. It has most methods for most things mostly before any other package. It has the best graphics. It extendable.

Cons It has a steep learning curve. No GUI by default. Poor (but improving) memory management; difficulty with very large data set (but improving as well).

R Resources

- `http://www.r-project.org` — Main R website.
- CRAN — `http://cran.csiro.au` — Comprehensive R Archive Network — base software and add-on packages.
- RStudio — `http://www.rstudio.com` — is a powerful IDE for R
- R Commander — `install.package(Rcmdr)` — is a partial GUI interface to R — requires TclTk.
- R Graph Gallery — `http://gallery.r-enthusiasts.com/` — loads of pretty pictures.
- SAS to R wiki — `http://kenkleinman.net/sasrwiki` — shows how to convert between SAS and R code.
- `http://cran.csiro.au/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf` — "A (very) short Introduction to R"

R Reference Books (*available online in WSU library!*)

- "The Art of R Programming: A Tour of Statistical Software Design", Norman Matloff, No Starch Press 2011
- “Beginning R: The Statistical Programming Language”, Mark Gardener, Wrox 2013.
- “R Object-oriented Programming”,Kelly Black, Packt Publishing 2014.
- “Introductory Statistics with R”, Peter Dalgaard, Springer 2008.

Getting started with R!

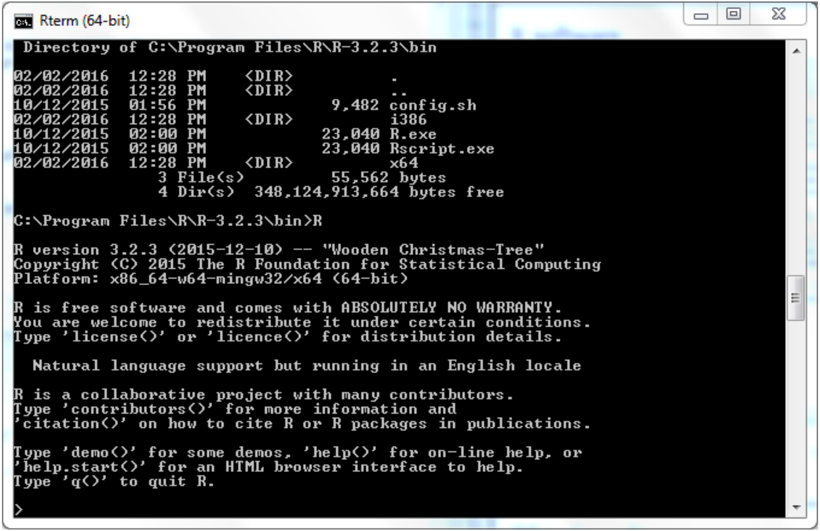
Set up the working environment - R and GUI installation:

1. Go to Main R website `http://www.r-project.org` and download the correct version of R for your system (Windows, Linux, Mac)
2. Install the R software
3. Go to RStudio `http://www.rstudio.com` and install correct version of RStuio for your system
4. Run R or RStudio

Installation order

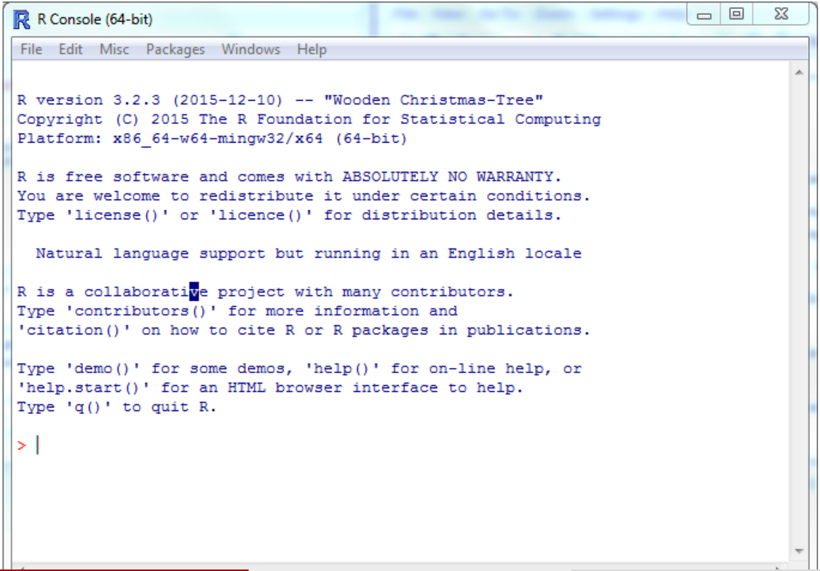
The order of installation is not very important now but it is better to install R first and then RStudio.

R software



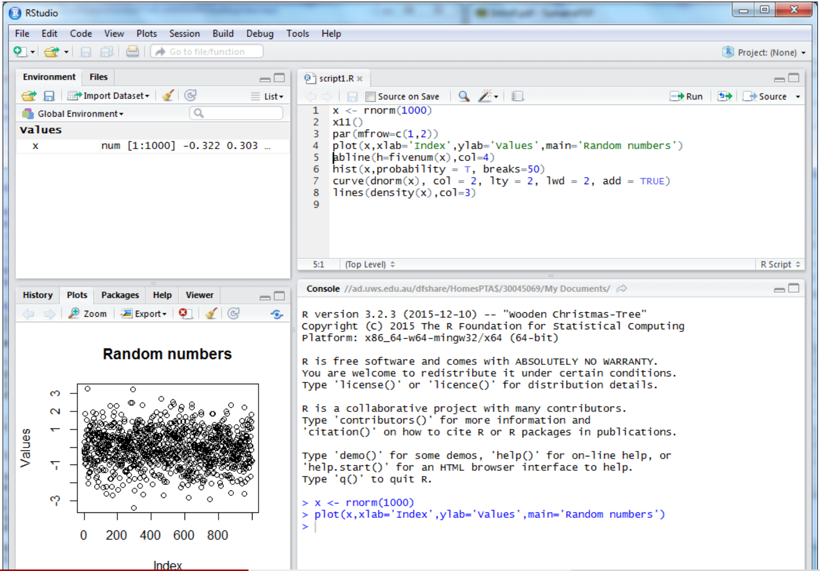
```
 Directory of C:\Program Files\R\R-3.2.3\bin

02/02/2016  12:28 PM    <DIR>          .
02/02/2016  12:28 PM    <DIR>          ..
10/12/2015  01:56 PM             9,482 config.sh
02/02/2016  12:28 PM    <DIR>          i386
10/12/2015  02:00 PM            23,040 R.exe
10/12/2015  02:00 PM            23,040 Rscript.exe
02/02/2016  12:28 PM    <DIR>          x64
               3 File(s)         55,562 bytes
               4 Dir(s)  348,124,913,664 bytes free

C:\Program Files\R\R-3.2.3\bin>R

R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 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.

  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.

>
```

R software



```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 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.

  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.

> 
```

RStudio software



```
Values
  x                num [1:1000] -0.322 0.303 ...
```

```
x <- rnorm(1000)
x11()
par(mfrow=c(1,2))
plot(x,xlab='Index',ylab='Values',main='Random numbers')
abline(h=fivenum(x),col=4)
hist(x,probability = T, breaks=50)
curve(dnorm(x), col = 2, lty = 2, lwd = 2, add = TRUE)
lines(density(x),col=3)
```

```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 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.

> x <- rnorm(1000)
> plot(x,xlab='Index',ylab='Values',main='Random numbers')
> 
```

R Commands

R can be used as a basic calculator.

```
> 1+1
[1] 2
> sqrt(2)
[1] 1.414214
> 2^5
[1] 32
```

R Commands

It can store things as named objects

```
> x <- 1
> print(x)
[1] 1
```

R Commands

It understands vectors and matrices

```
> x <- c(1,2)
> m <- matrix(c(1,2,3,4), ncol=2, byrow=TRUE)
> print(m)
     [,1] [,2]
[1,]    1    2
[2,]    3    4
> m %*% x
     [,1]
[1,]    5
[2,]   11
```

R Commands

It has functions, and you can write them

```
> x <- sqrt(2)
> sqr <- function(x) x^2
> sqr(2)
[1] 4
```

Data in R is stored in data.frames

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> dim(iris)
[1] 150   5
```

Some columns are numeric others are factors

```
> sapply(iris, class)
Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
   "numeric"   "numeric"    "numeric"   "numeric"  "factor"
```

Data can be read from text files (`read.csv` and `read.table`) and various formats using the `foreign` package.

Basic Statistics

```
> x <- rnorm(1000)
> mean(x)
 [1] 0.01115976
> var(x)   ### sd(x)
 [1] 0.9491709
> fivenum(x)
 [1] -3.12386993 -0.65184275 0.02283834 0.62827772 3.0558
```

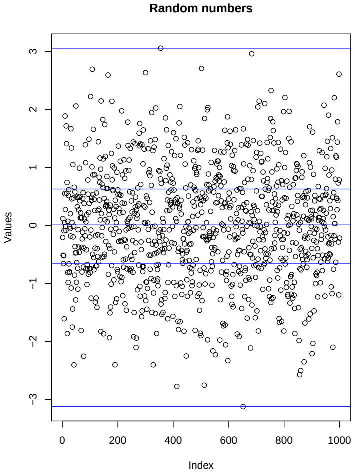
Basic Statistics

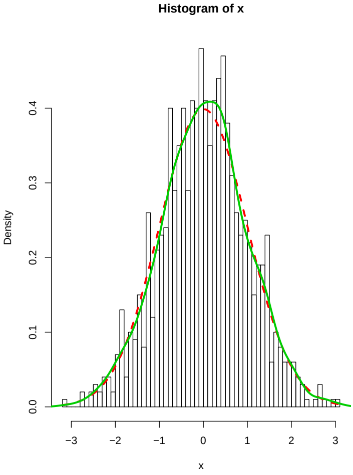
```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

Basic Statistics

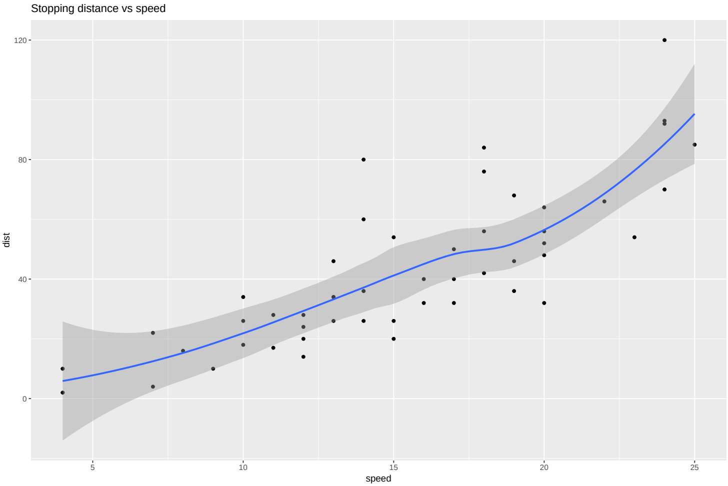
```
> t.test(x)

      One Sample t-test

data: x
t = 0.36223, df = 999, p-value = 0.7173
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.04929727 0.07161678
sample estimates:
 mean of x
0.01115976
```

R plotting


Random numbers
Values
3
2
1
0
−1
−2
−3
0
200
400
600
800
1000
Index


Histogram of x
Density
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3
x

R plotting


Stopping distance vs speed
dist
speed