2012

The "R" Statistics library: Research Applications

David Allen
Edith Cowan University

Abhay Singh
Edith Cowan University

The R Statistics Library: Research Applications

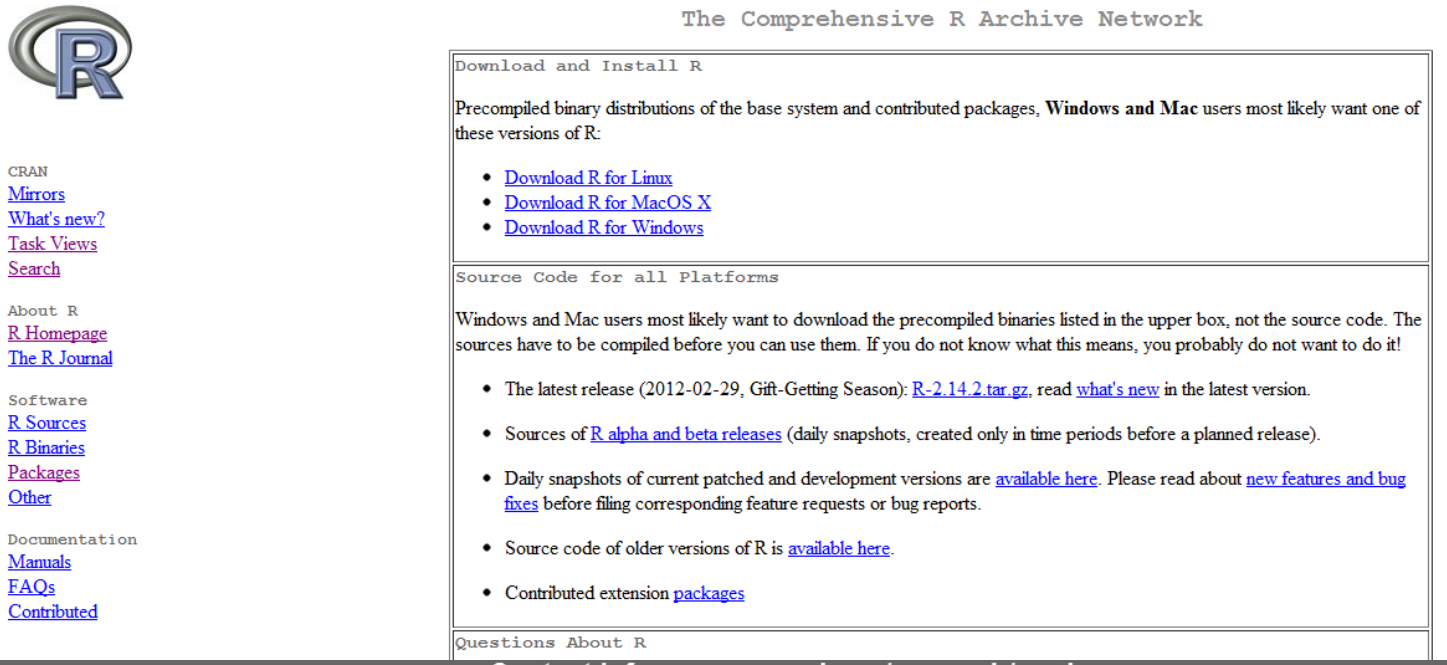
David E Allen & Abhay K Singh

School of Accounting Finance & Economics
Edith Cowan University

Use of The R Library

Use of the R library See http://cran.r-project.org/
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

Figure: R Website



R

- R is the Rolls Royce of free open source programs for statistical analysis .
- The syntax of the R language is very similar to that of the S language which underlies S-PLUS. In fact, you can think of R as 'GNU S'. The main differences that users familiar with S will notice are (1) objects are not saved as separate files in a directory but are stored internally; and (2) the packages available in R are not the same as the libraries available in S-PLUS.
- Rob Hyndman Professor of statistics at Monash University suggests the following pros and cons of using R for statistical and econometric analysis.

Advantages of R

Pros:

- Free. Students can have copies at home.
- Portable. Once students invest in learning this program, they can take it with them and install it again wherever they may end up working.
- Versatile. The software exists for more platforms than virtually any existing commercial program.
- General. A very large number of statistical/econometric tools are available, so the software could be used for many (maybe all) subjects
- Cutting-edge. It includes the very latest methods.
- Programmable. It is easy for students to program new methods or develop modifications of existing methods.
- Matrix language. The R language handles vectors and matrices directly (as do Gauss, Matlab and Ox). This makes programming much simpler for students and reinforces the matrix notation used in class.

Pros and cons of R

- Object-oriented language. Students may take a little time to adjust to the object-oriented way of thinking, but it simplifies things greatly. For example, the plot function and the summary function can be used on all types of data and fitted models.
- Great graphics.
- Relatively fast.

Cons:

- Command-driven. Although the command line, which is similar to that of the bash shell, is extremely powerful and easy to use, some students accustomed to drag-and-drop menu programs may find R awkward to use.
- Missing functions. Some desirable functions have not yet been written. Of course, it is easy to add such functions yourself.
- Inconsistent syntax. Command syntax is not always consistent between packages which do similar things.

R Packages

R functionality is based around the concept of “packages". A package is a collection of functions to carry out certain tasks (rather like Gauss modules or Matlab toolboxes). For example, the nls package does nonlinear regression, the ts package contains a variety of time-series functions, and so on. The base packages are automatically available with a default installation. Contributed packages, on the other hand, need to be installed individually. One can obtain these packages by following the download link on the R home page.

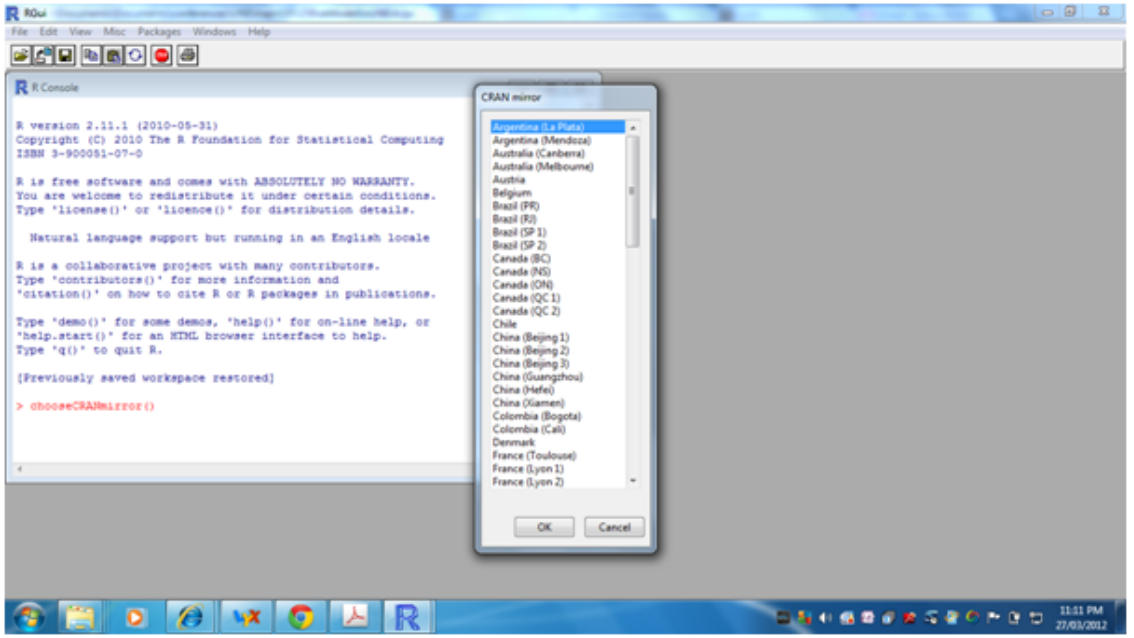
R Packages

By its very nature, R is a dynamic, evolving computing environment, and packages are continuing to be written at a rapid rate. Very often, the capabilities of two or more packages overlap. For example, both the tseries and ts packages provide procedures for fitting ARMA models.
R is a powerful and well-written open-source statistical software package.

Cran mirrors

R can be obtained from a variety of CRAN mirror sites located around the globe. There are 3 here in Australia.

Figure: Cran Mirrors



Increasing Interest in R (Tsay University of Chicago)

Ruey S. Tsay Teaching Page

This page has teaching and book materials.

Books:

1. A Course in Time Series Analysis (ed. Pena, Tiao, & Tsay)
   Wiley, 2001, ISBN: 0-471-36164-X

2. Analysis of Financial Time Series , Wiley, 2002
   ISBN: 0-471-41544-8

3. Analysis of Financial Time Series, Second Edition, Wiley, 2005
   ISBN: 0-471-69074-0

4. Analysis of Financial Time Series, Third Edition, Wiley, 2010.
   ISBN: 0-470-41435-9; 10-digits: 978-0470414354

5. An Introduction to Analysis of Financial Data with R, Wiley 2012

Courses:

Autumn Quarter 2008:

1) Business 41910: Time Series Analysis for Forecasting and Model Building

Tsay University of Chicago

HOME / MATHEMATICS & STATISTICS / APPLIED PROBABILITY & STATISTICS / STATISTICS FOR FINANCE, BUSINESS & ECONOMICS /

An Introduction to Analysis of Financial Data with R

Wiley's Series in PROBABILITY AND STATISTICS

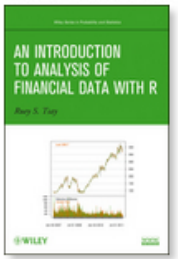
RELATED SUBJECTS

- Statistics for Social Sciences
- Survival Analysis
- Biometrics
- Econometric & Statistical Methods
- General Applied Probability & Statistics
- Environmental Statistics & Environmetrics
- Applied Probability & Statistics Models

RELATED TITLES

More From This Series

Stochastic Geometry and its Applications 3e
by Professor Dietrich Stoyan, Dr Wilfrid S. Kendall, Dr Sung Nok Chiu, Joseph Mecke



Larger Image

An Introduction to Analysis of Financial Data with R
Ruey S. Tsay

ISBN: 978-0-470-89081-3

Hardcover
420 pages
November 2012

AUD $155.00 / NZD $180.00 Add to Cart

This price is valid for Australia. Change location to view local pricing and availability.

Evaluation Copy

Instructors may request an evaluation copy for this title.

Permissions

To reuse content from this title
Request Permissions

Description | Table of Contents

This book provides a systematic and mathematically accessible introduction to financial econometric models and their applications in modeling and predicting financial time series data. It emphasizes empirical financial data and focuses on real-world examples. Following this approach, readers will master key aspects of financial time series, including volatility modeling, neural network applications, market microstructure, and high-frequency financial data. S-Plus commands and illustrations are used extensively throughout the book in order to highlight accurate interpretations and graphical representations of financial data. Exercises are included in order to provide readers with more opportunities to put the models and methods into everyday practice. The tools provided in the text aid readers in developing a deeper understanding of financial markets through firsthand experience in working with financial data, most importantly without needless computation.

International Conference on the use of R

R/Finance 2012: Applied Finance with R

May 11 & 12, Chicago, IL, USA

home agenda register travel committee 2009 Conference 2010 Conference 2011 Conference

> home(2012)

The fourth annual R/Finance conference for applied finance using R, the premier free software system for statistical computation and graphics, will be held this spring in Chicago, IL, USA on **Friday May 11 and Saturday May 12, 2012**. The two-day conference will cover portfolio management, time series analysis, advanced risk tools, high-performance computing, econometrics and more. All will be discussed within the context of using **R** as a primary tool for financial risk management, analysis and trading.

We invite you to submit complete papers or one-page abstracts (in txt or pdf format) for consideration. Academic and practitioner proposals related to R are encouraged. We welcome submissions for full talks, abbreviated "lightning talks", and for a limited number of (longer) pre-conference seminar sessions.

Presenters are strongly encouraged to provide working R code to accompany the presentation/paper. Data sets should also be made public for the purposes of reproducibility (though we realize this may be limited due to contracts with data vendors). Preference may be given to presenters who have released R packages.

Travel and accommodation grants may be available for selected presenters at the discretion of the committee. In addition, the conference will award prizes for best papers. To be eligible for a best paper award, a submission must be a full paper. Extended abstracts, even if a full paper by conference time, are not eligible for a best paper award.

The submission deadline was January 31, 2012.

The **draft agenda** is now available, as is **information about registration**.

The 2012 conference will build upon the success of the three previous events. We expect invited keynote lectures by:

- Blair Hull,
- Paul Gilbert,
- Rob McCulloch and
- Simon Urbanek.

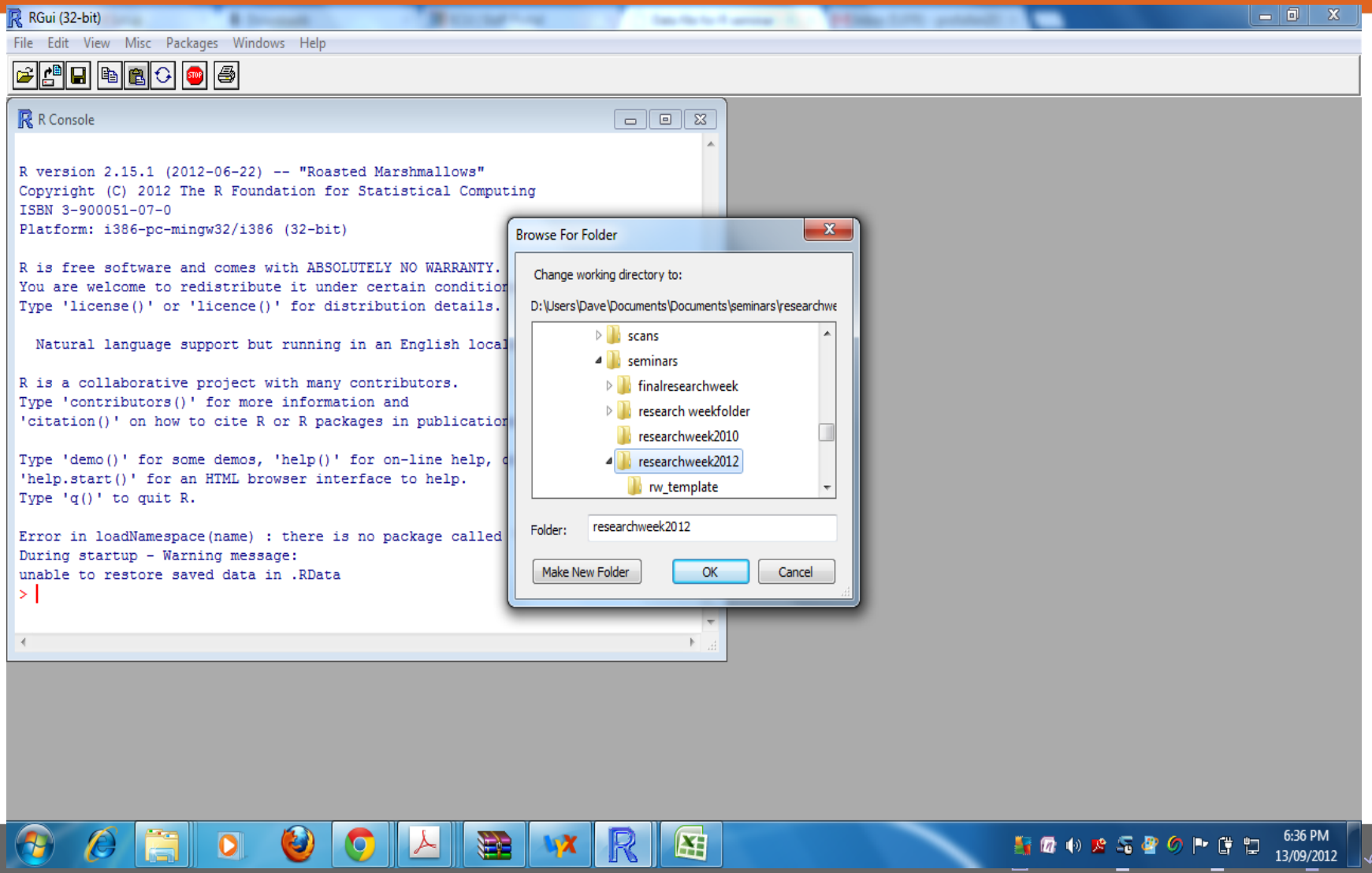
plus presentations of contributed papers, short "lightning-style" presentations, and optional pre-conference tutorials. It also presents a unique opportunity to meet fellow R users and developers, and a thus provides a chance to discuss the future of R in Finance.

The inaugural 2009 conference featured keynotes by Patrick Burns, Robert Grossman, David Kane, Roger Koenker, David Ruppert, Diethelm Wuertz, and Eric Zivot, as well as a number of excellent presentations. The 2010 conference followed up with keynotes by Bernhard Pfaff, Ralph Vince, Marc Wildi, and Achim Zeileis. Last year's conference featured keynotes by Meb Faber, Stefano Iacus, John Bollinger and Louis Kates.

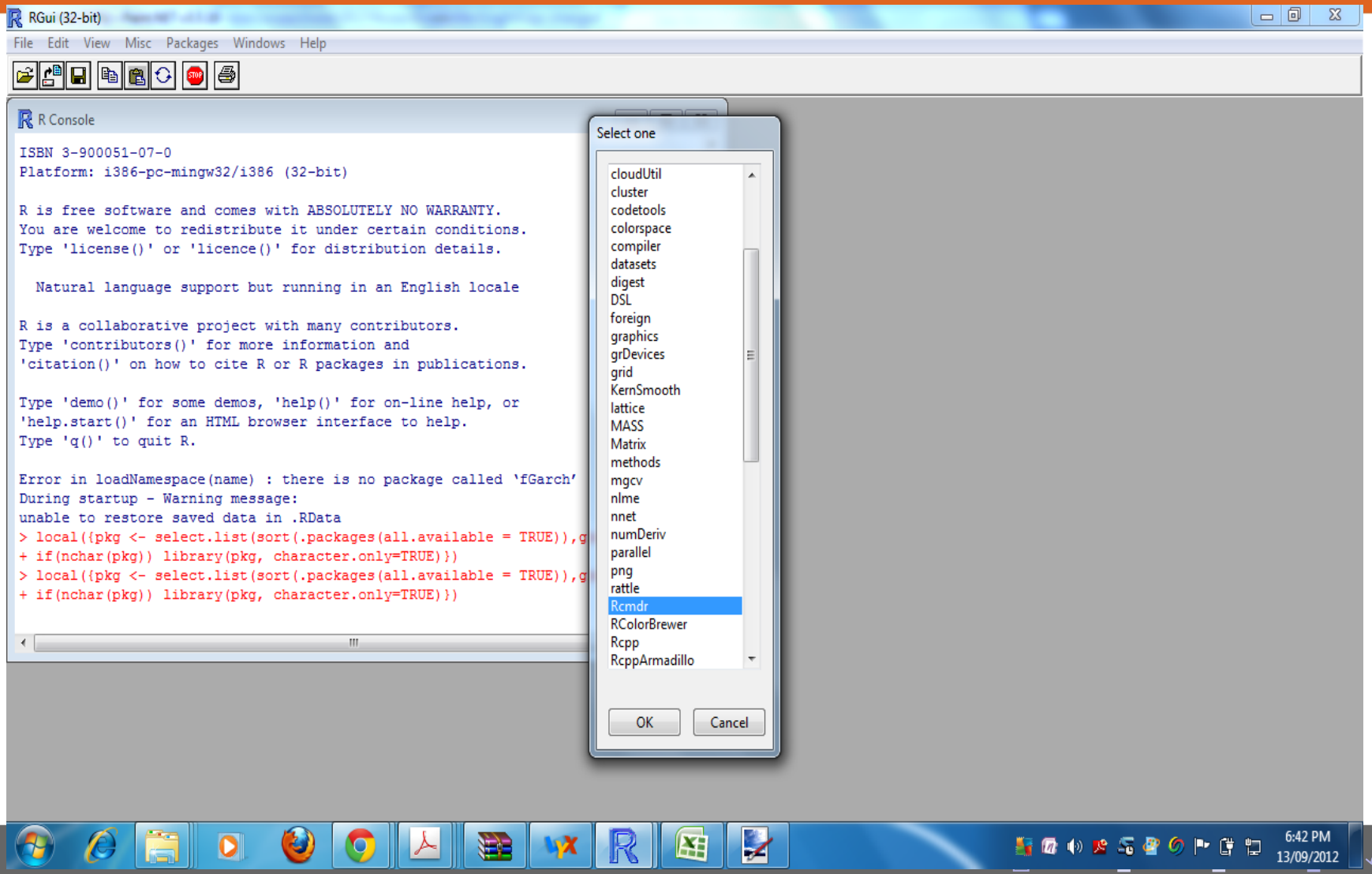
Complete programs of the previous conferences, along with downloadable presentation slides, are available via the links above and below.

The R/Finance 2012 conference is again organized by a local group of R package authors and community contributors, and hosted by the International Center for Futures and Derivatives [ICFD] at the University of Illinois at Chicago. Limited sponsorship opportunities are available.

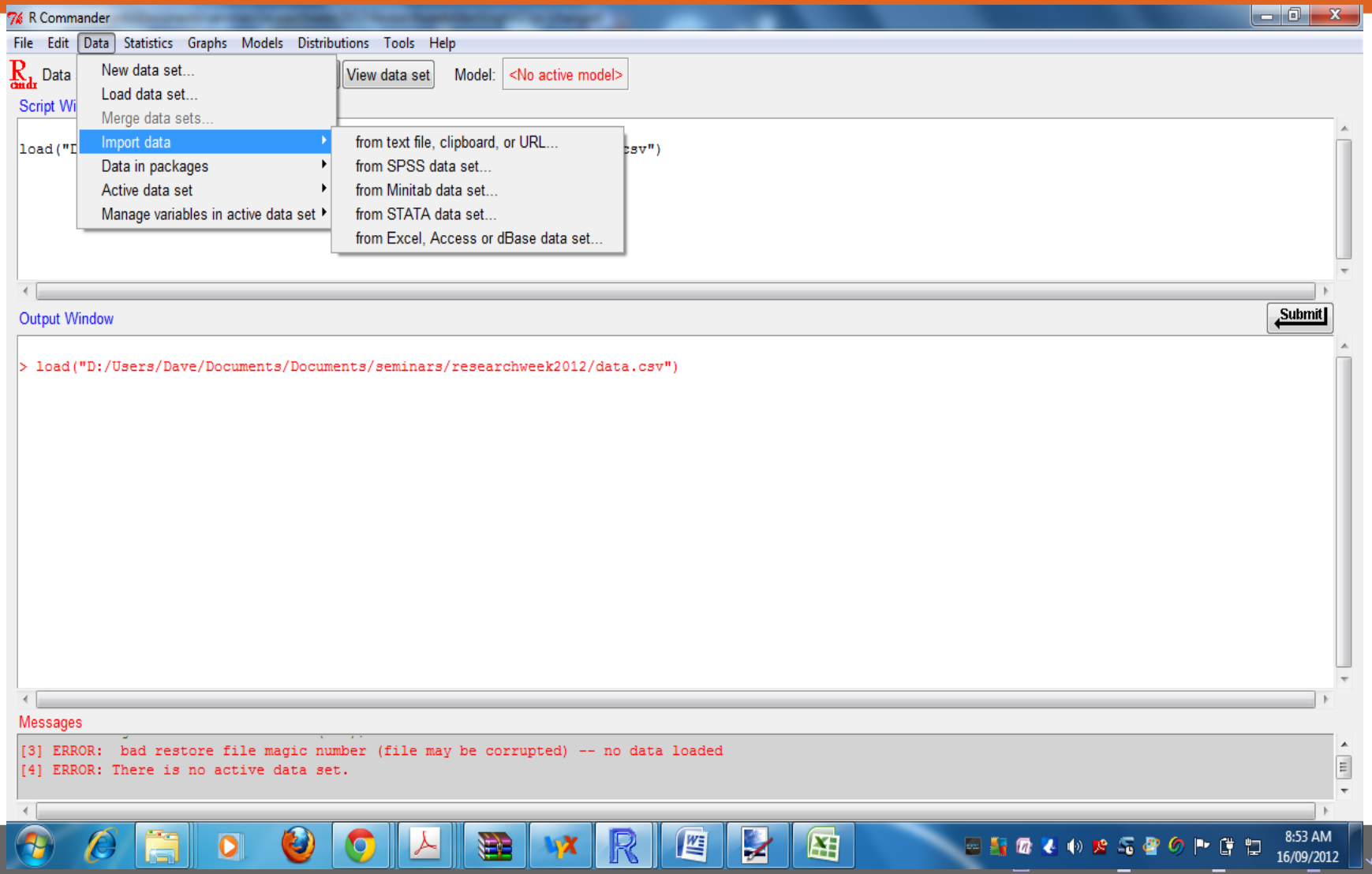
Using R: Start up and set directory



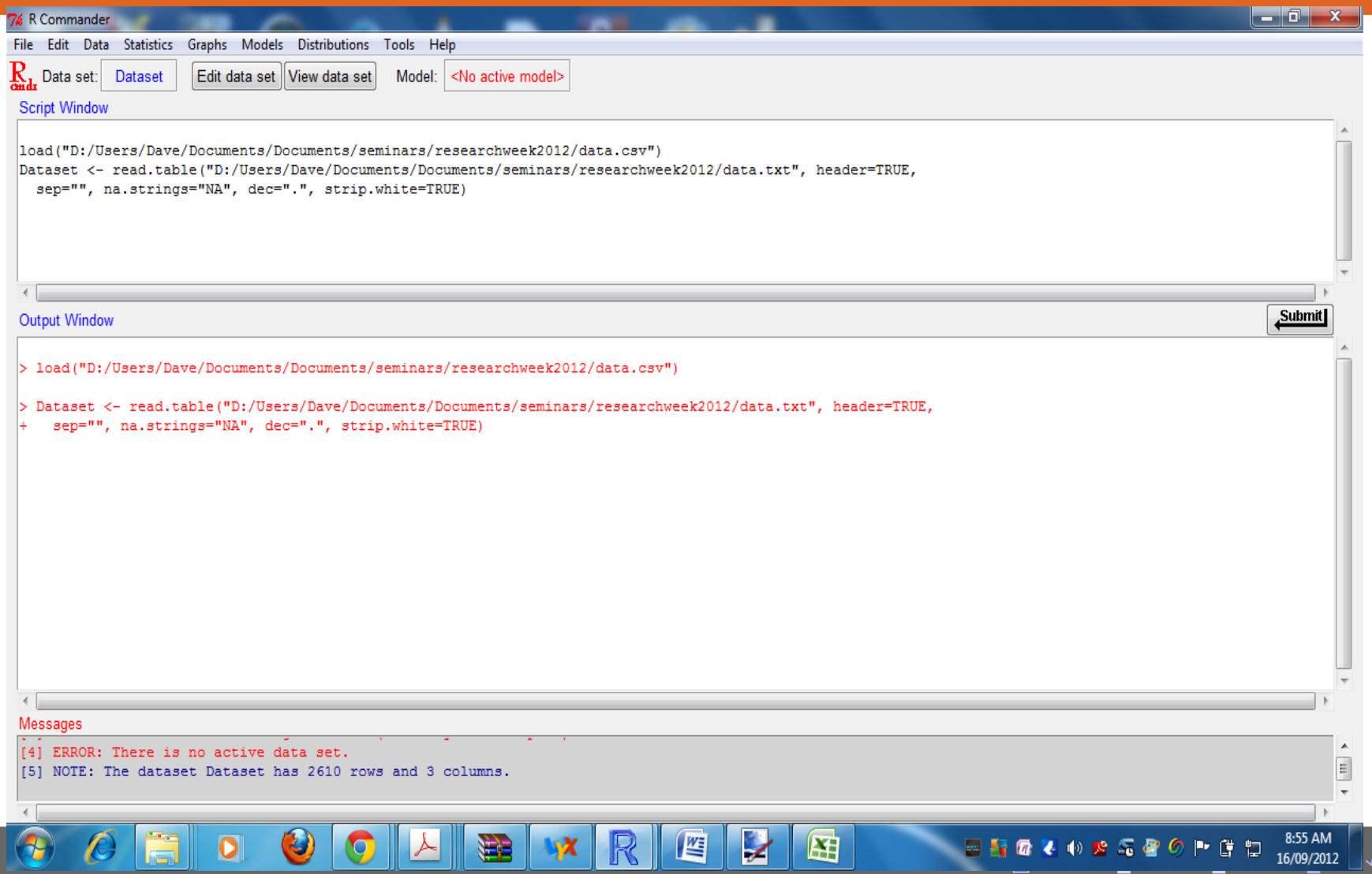
Load required packages



RCommander data import



RCommander data import code



Rcommander Summary statistics

R Commander

File Edit Data Statistics Graphs Models Distributions Tools Help

Data set: Dataset | Edit data set | View data set | Model: <No active model>

Script Window

```
load("D:/Users/Dave/Documents/Documents/seminars/researchweek2012/data.csv")
Dataset <- read.table("D:/Users/Dave/Documents/Documents/seminars/researchweek2012/data.txt", header=TRUE,
  sep="", na.strings="NA", dec=".", strip.white=TRUE)
summary(Dataset)
```

Output Window — Submit

```
> load("D:/Users/Dave/Documents/Documents/seminars/researchweek2012/data.csv")

> Dataset <- read.table("D:/Users/Dave/Documents/Documents/seminars/researchweek2012/data.txt", header=TRUE,
+   sep="", na.strings="NA", dec=".", strip.white=TRUE)

> summary(Dataset)
         Date          SP200           Rio
 2002/09/05:   1   Min.   :2700   Min.   : 22.22
 2002/09/06:   1   1st Qu.:3707   1st Qu.: 32.33
 2002/09/09:   1   Median :4397   Median : 58.76
 2002/09/10:   1   Mean   :4430   Mean   : 57.04
 2002/09/11:   1   3rd Qu.:4929   3rd Qu.: 72.91
 2002/09/12:   1   Max.   :6829   Max.   :123.12
 (Other)   :2604
```

Messages

```
[4] ERROR: There is no active data set.
[5] NOTE: The dataset Dataset has 2610 rows and 3 columns.
```

8:57 AM
16/09/2012

Descriptive Statistics

- Easy way to calculate common descriptive statistics using Performance Analytics package.
- Data has two variables with last 10 year's prices of S&P/ASX-200 and Rio Tinto.

Code

```
#Load the Library
library(PerformanceAnalytics)
# Reading Data from the CSV file
d1=read.csv("data.csv",strip.white=TRUE)
#Keeping the data in the workspace
attach(d1)
#Descriptive statistics using Performance Analystics package's function table.Stats
t1=table.Stats(d1$SP200)
t2=table.Stats(d1$Rio)
#Combining the two tables using Column Bind (cbind)
T=cbind(t1,t2)
# Name the columns
colnames(T)=c("SP200","Rio")
```

Descriptive Statistics

	SP200	Rio
Observations	2610.00	2610.00
NAs	0.00	0.00
Minimum	2700.40	22.22
Quartile 1	3706.82	32.33
Median	4396.70	58.76
Arithmetic Mean	4429.70	57.04
Geometric Mean	4333.04	51.91
Quartile 3	4928.98	72.91
Maximum	6828.70	123.12
SE Mean	18.18	0.47
LCL Mean (0.95)	4394.04	56.12
UCL Mean (0.95)	4465.35	57.95
Variance	862725.71	567.01
Stdev	928.83	23.81
Skewness	0.35	0.31
Kurtosis	-0.39	-0.71

Descriptive Statistics

Convert prices to percentage log returns

```
#Prices to returns
n <- nrow(d1)
sp200.ret=100*(log(d1$SP200[2:n]) - log(d1$SP200[1:(n - 1)]))
rio.ret= 100*(log(d1$Rio[2:n]) - log(d1$Rio[1:(n - 1)]))
#Create new dataset with returns
d2=data.frame(d1$Date[2:n],sp200.ret,rio.ret) colnames(d2)=c("date","sp200","rio")
attach(d2)
```

Plotting Histograms

```
par(mfrow=c(1,2)) #graphical parameters
#histograms using function of PerformanceAnalytics package
chart.Histogram(sp200,breaks=50,method=c("add.rug"),main="S&P/ASX-200")
chart.Histogram(rio,breaks=50,method=c("add.rug"),main="Rio Tinto")
```

Histograms

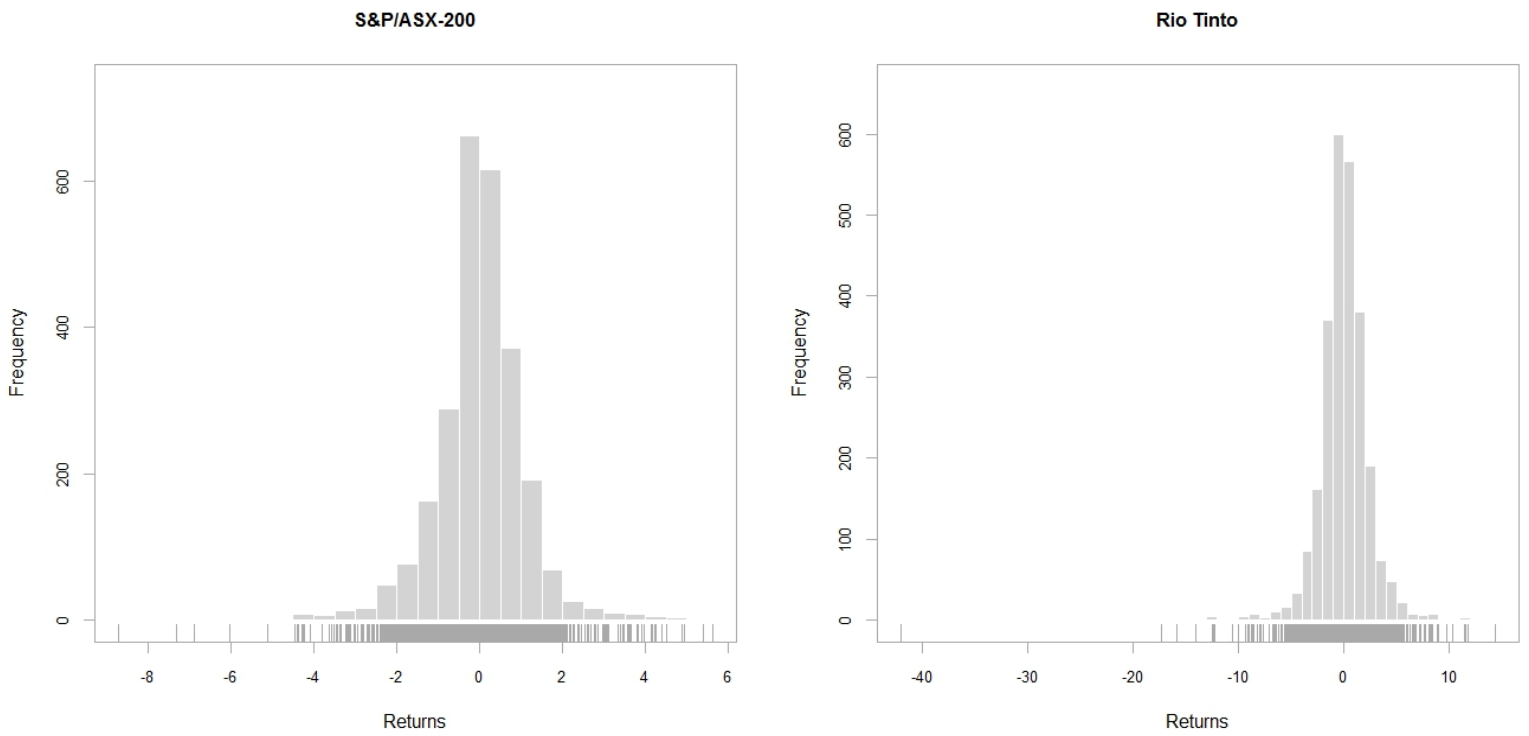


Figure: Histograms

Linear Regression

- CAPM model (Assuming zero risk free rate).

$$S_R = \alpha + \beta M_R + e$$

- Simple linear regression approach

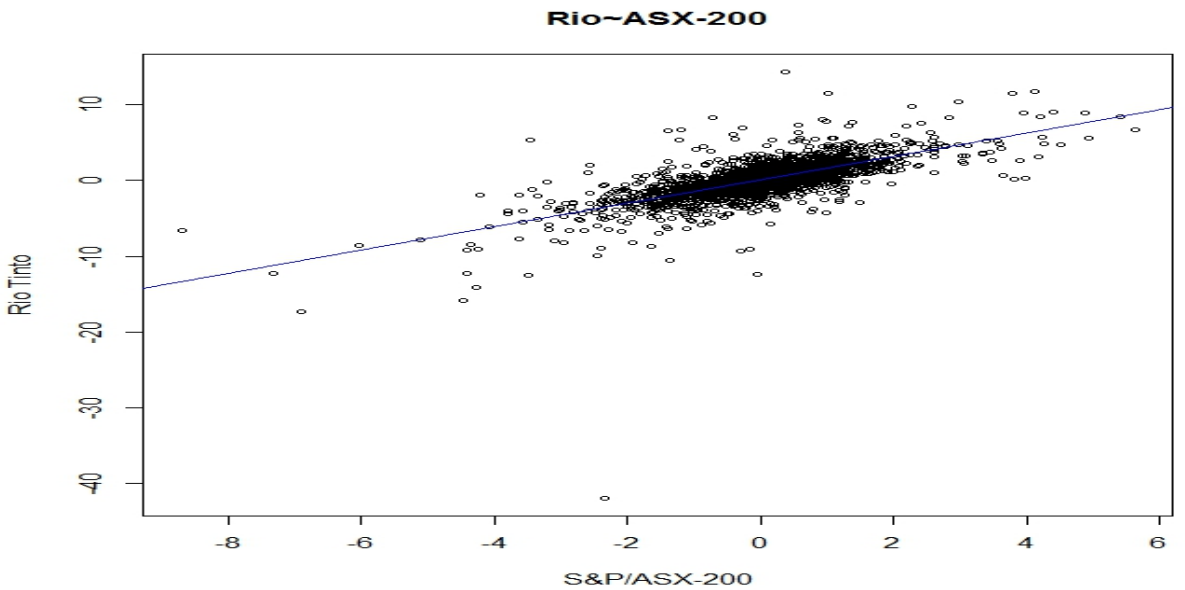
```
#linear regression using lm command
fitlr=lm(rio~1+sp200)
#print summary
s1=summary(fitlr)
print(s1)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0077	0.0343	0.22	0.8224
d2$sp200	1.5409	0.0313	49.28	0.0000

Linear Regression

- Regression Plot

```
#simple regression plot
plot(sp200,rio,xlab="S&P/ASX-200", ylab="Rio Tinto",main="Rio~ASX-200",cex=0.75)
abline(coef(fitlr),col=4)
```

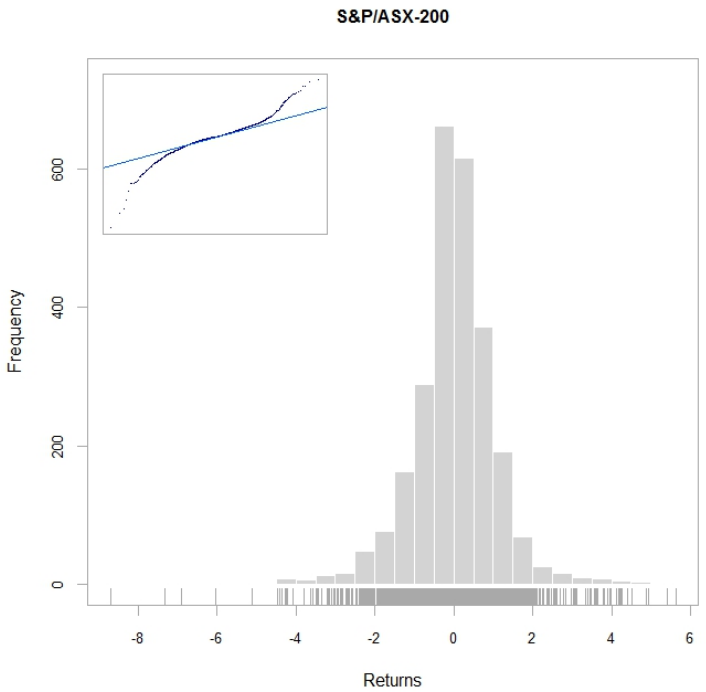


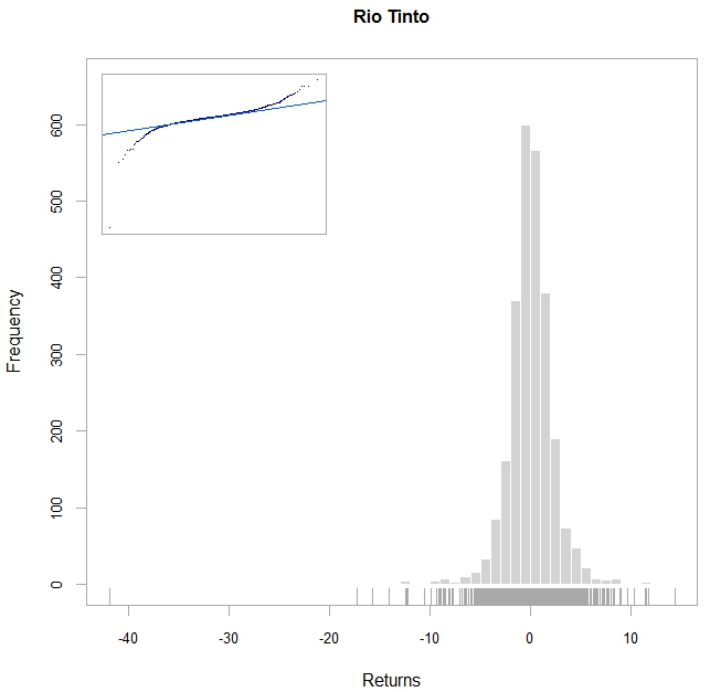
Quantile Regression

- Financial return distribution is not usually normal (against the most popular assumption)
- How to visualize this?
  - Q-Q Plot

```
#histograms with Q-Q plots using function of PerformanceAnalytics package
chart.Histogram(sp200,breaks=50,method=c("add.rug","add.qqplot"),main="S&P/ASX-200") dev.new()
chart.Histogram(rio,breaks=50,method=c("add.rug","add.qqplot"),main="Rio Tinto")
```

Quantile Regression





Quantile Regression

- Package : Quantreg
  - Very powerful R package to model Quantile Regression with support for Linear and Non-Linear Quantile Regression models.

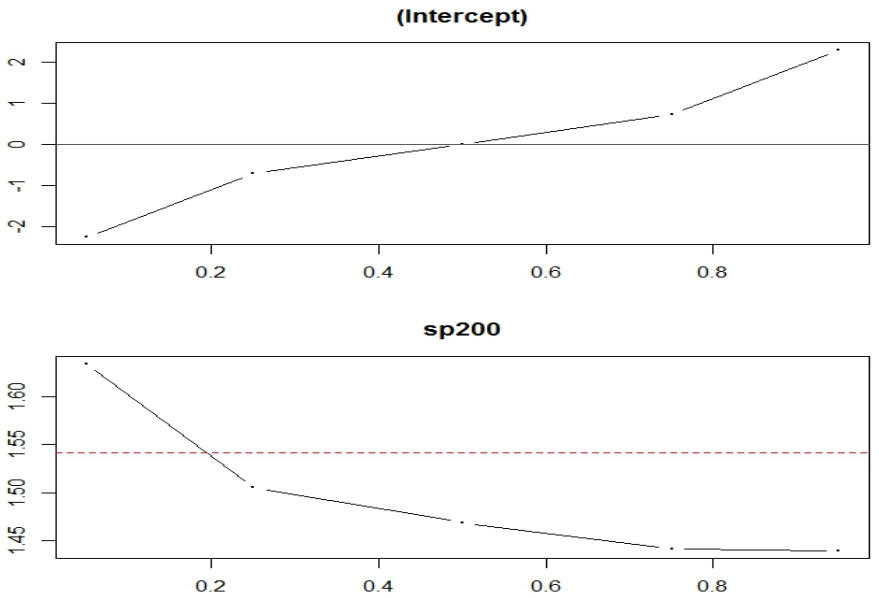
```
#load package
library(quantreg)
#quantile regression fit
taus=c(0.05,0.25,0.5,0.75,0.95)
fitrq=rq(rio~sp200,tau=taus)
s2=summary(fitrq,se="nid")
```

Quantiles	(Intercept)	sp200
0.05	−2.256	1.634
	(0.094)	(0.086)
0.25	−0.703	1.505
	(0.035)	(0.019)
0.50	0.000	1.469
	(0.020)	(0.017)
0.75	0.730	1.442
	(0.034)	(0.026)
0.95	2.291	1.439
	(0.114)	(0.089)

Quantile Regression

Plotting Quantile Regression Coefficients.

```
#plot QR estimates of alpha and beta
plot(fitrq)
```



Quantile Regression

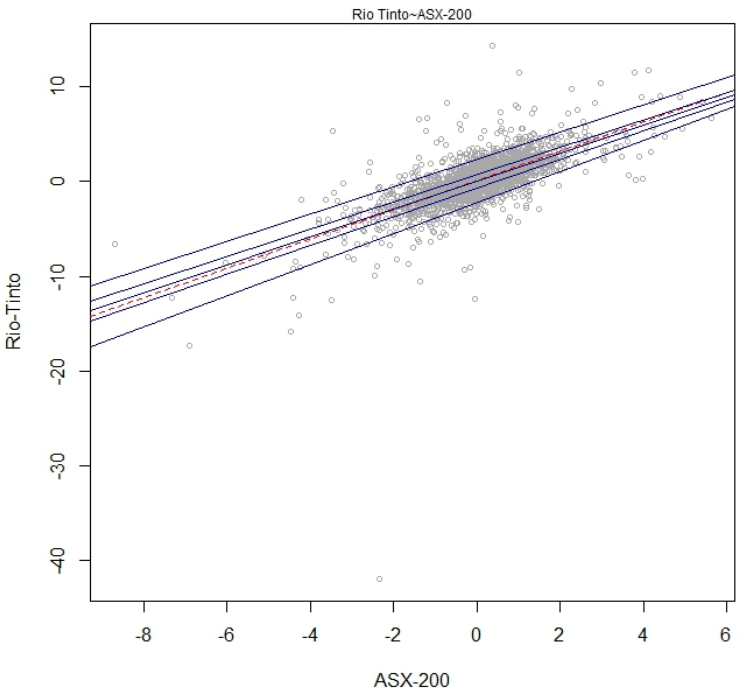
Plotting Regression Slopes with a loop

```
plot(sp200,rio,type="n",ylab="Rio-Tinto",
xlab="ASX-200")
points(sp200,rio,cex=.75,col="darkgray")
abline(lm(rio~sp200),lty=2,col="red")
for( i in 1:length(taus)){
abline(rq(rio~sp200,tau=taus[i]),col="darkblue",lty=1)
}
mtext("Rio Tinto~ASX-200",side=3,cex=0.75)
```

Quantile Regression

Plotting Regression Slopes with a loop

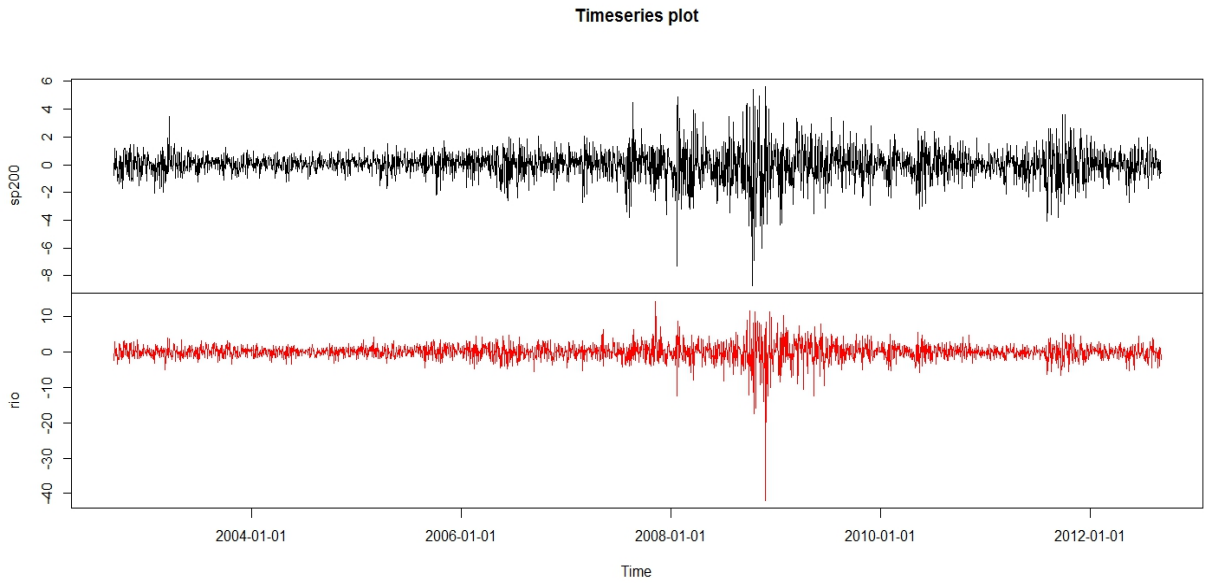
```
plot(sp200,rio,type="n",ylab="Rio-Tinto",
xlab="ASX-200")
points(sp200,rio,cex=.75,col="darkgray")
abline(lm(rio~sp200),lty=2,col="red")
for( i in 1:length(taus)){
abline(rq(rio~sp200,tau=taus[i]),col="darkblue",lty=1)
}
mtext("Rio Tinto~ASX-200",side=3,cex=0.75)
```



Time Series

Converting data to time series data.

```
library(timeSeries)
#note that the date should be in R recognizable format (e.g., YYYY-MM-DD) or other pre-processing is
required see help(strptime)
d3=as.timeSeries(d2)
#plotting
plot(d3, main="Timeseries Plot")
```



Time Series-Basic GARCH

Garch modelling using fGarch Package.

```
#basic garch fit using fGarch library(fGarch)
fitg=garchFit(~garch(1,1),d3[,1])
summary(fitg) #prints the summary on the R console
#interactive plots for the garch fit

plot (fitg)
```

The code used till here along with data.csv can be downloaded from my personal homepage
https://sites.google.com/site/drabhayksingh/codes

Text Mining Using R and Twitter

- Useful webpage to get started `https://sites.google.com/site/miningtwitter/home`
- Useful R Packages
  - twitteR (twitter client for R) `http://cran.r-project.org/web/packages/twitteR/vignettes/twitteR.pdf`
  - XML (for parsing XML and HTML documents) `http://www.omegahat.org/RSXML/`
  - tm (for text mining) `http://tm.r-forge.r-project.org/index.html`
  - ggplot2 (for cool and elegant graphics) `http://had.co.nz/ggplot2/`
  - stringr (makes string functions simpler and easier to use)`https://github.com/hadley/stringr/blob/master/README.md`
  - igraph (for graphs and networks) `http://igraph.wikidot.com/`
  - RColorBrewer (for nice color palettes) `http://colorbrewer2.org/`
  - wordcloud (for wordclouds aka tag clouds)
  - sentiment (for sentiment analysis) `http://cran.r-project.org/web/packages/sentiment/index.html`
  - CRAN Task View: Natural Language Processing `http://cran.r-project.org/web/views/NaturalLanguageProcessing.html`

Carbon Tax-A Sentiment Analysis

A simple sentiment analysis of Twitter tweets related to carbon tax. Code obtained from `https://sites.google.com/site/miningtwitter/home`.

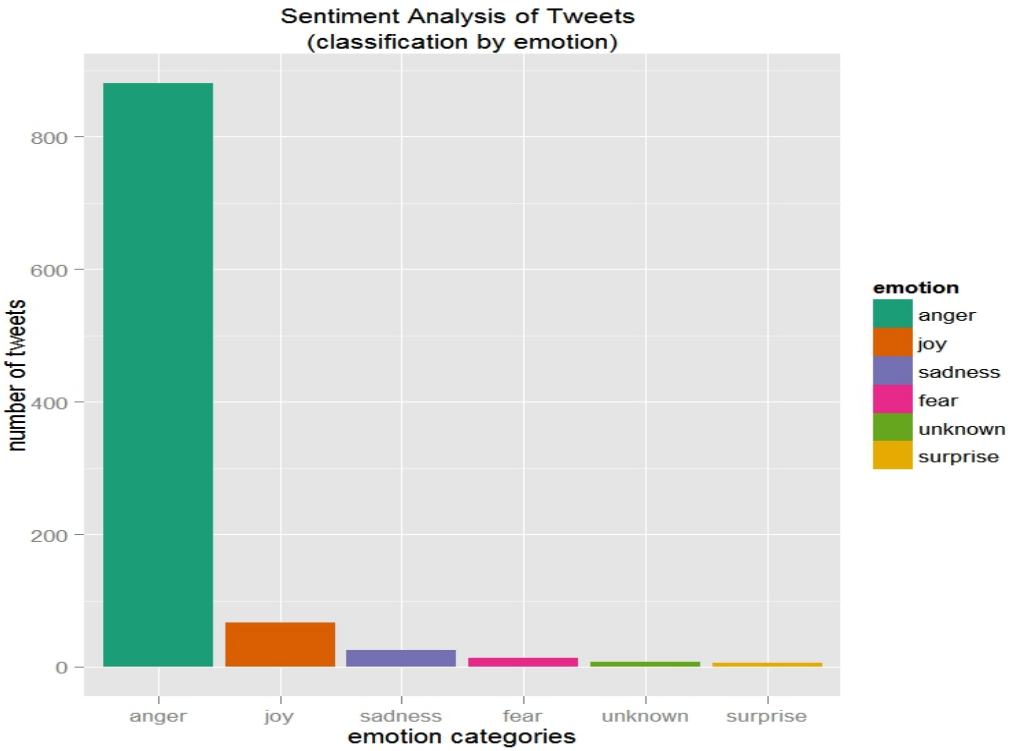


Figure: Number of tweets based on sentiments

Carbon Tax-A Sentiment Analysis

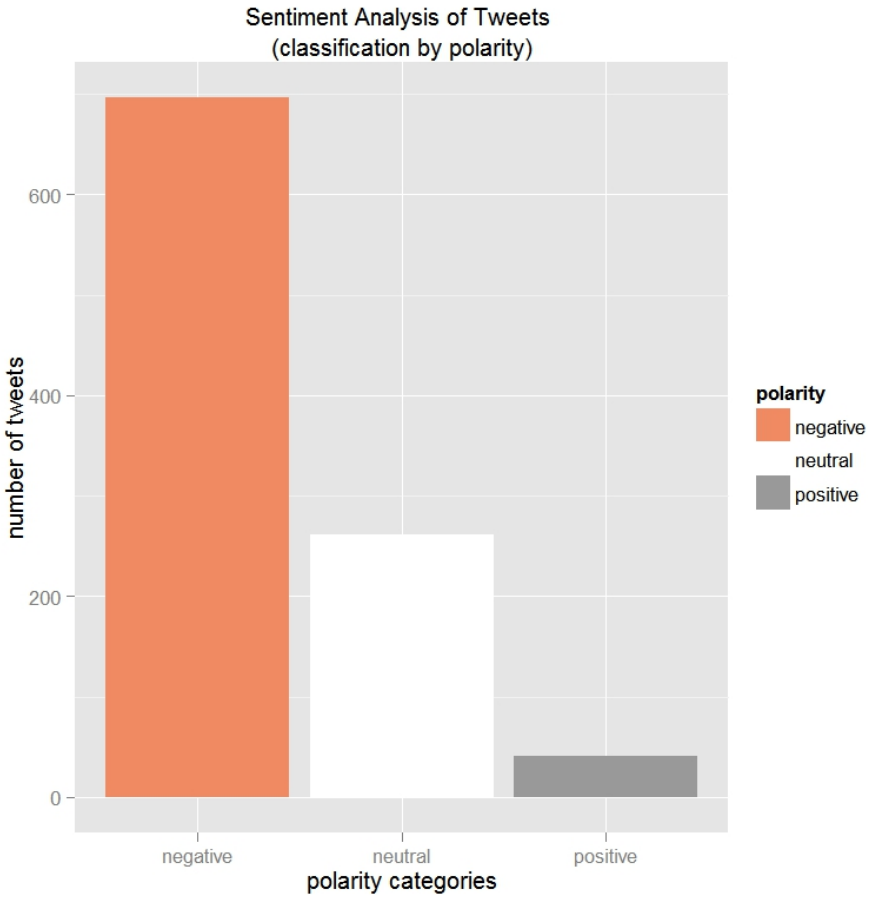


Figure: Classification of Tweets by polarity

Carbon Tax-A Sentiment Analysis

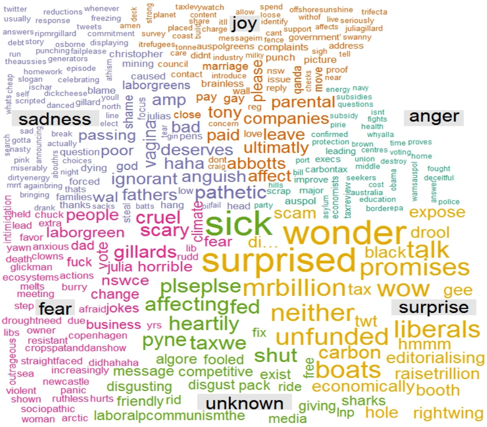


Figure: Comparison Word Cloud

Carbon Tax-A Sentiment Analysis

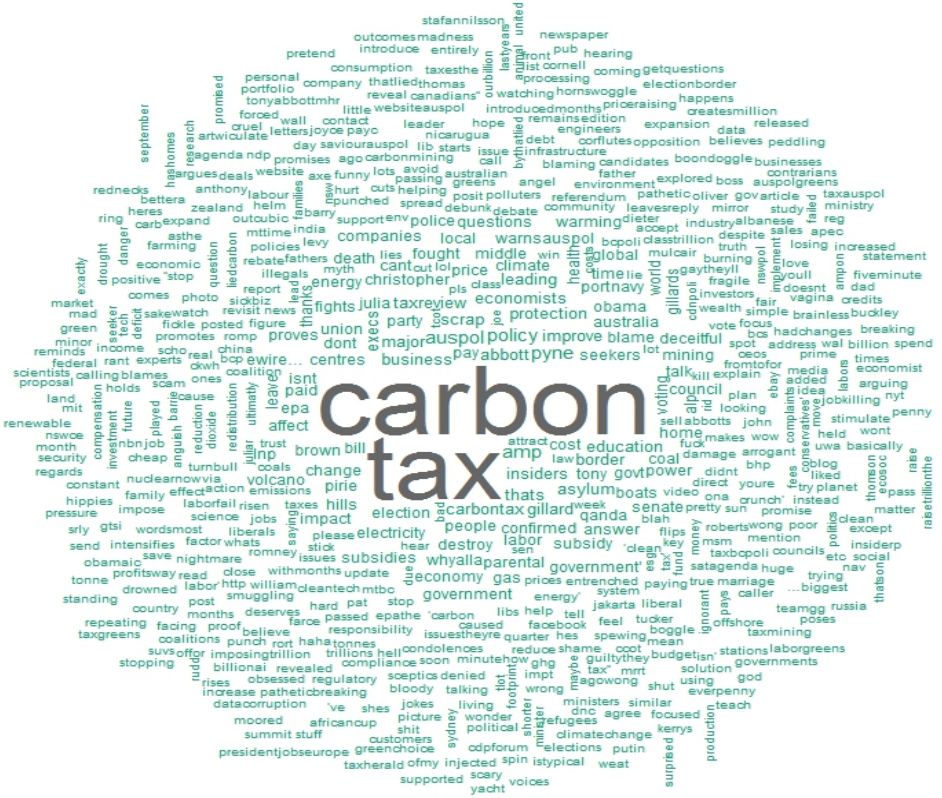


Figure: A simple wordcloud

Thank You