Introduction to R

Dr. Peter Morgan

MorganPH@Cardiff.ac.uk

Office 029-20875727 Mobile 0780-9677213

Pros and Cons of R

Advantages

- Very compact - A complete analysis often done by a single function
- Hugely versatile and powerful – thousands of packages available over and above the base package
- Freedom from restrictions of pull down menus
- Graphics are of high quality and capacity
- Excellent colour-handling
- Creates 'objects' which can be used to bridge between different analyses
- Written by practising statisticians and data analysts so default options often all you need
- Useful for many different types of data – numbers, text, visual, audio, mixed, etc.
- Updated very regularly – so any bugs are cured quickly
- Free
- Generic functions 'do their best' to provide an answer
- A huge and enthusiastic world-wide user community
- Very well supported by documentation and searchable user group lists and blogs
- Built-in and copy/pasteable examples which can be modified for own use

Disadvantages

- Takes a little bit of getting used to if you use SPSS, Minitab, SAS, etc. from their menu-driven options
- If you make a mistake, R will often do what it did before as a default (look out for error messages)
- Error messages can be cryptic but web search on an error message will most often find an explanation
- Some technical documentation can be obtuse but user group help messages will often provide an answer
- Very addictive

Reading in Data

- Data can be read in as a table from
  - text files, txt (tab delimited), csv (comma separated value)
  - directly from data sets in websites
  - via the clipboard or entered via the keyboard
- Data in text form can be read from
  - webpages (webscraping)
  - text files
  - PDFs
- Sound and image data can also be read in and processed

A typical R analysis

Put in an object

Read data

Date table has a top row of column names

```
> wages=read.table(file="data/Wages.txt", header=TRUE)
```

Random sample of 534 Wages from the 1985 US Current Population Survey

Generic function

File location

```
> summary(wages)
```

```
   Education          South          Sex          Experience          Union          Wage            Age       
 Min.   : 2.00   Not_South:378   Female:245   Min.   : 0.00   Non_Union:438   Min.   : 1.000   Min.   :18.00  
 1st Qu.:12.00   South    :156   Male  :289   1st Qu.: 8.00   Union    : 96   1st Qu.: 5.250   1st Qu.:28.00  
 Median :12.00                                Median :15.00                   Median : 7.780   Median :35.00  
 Mean   :13.02                                Mean   :17.82                   Mean   : 9.024   Mean   :36.83  
 3rd Qu.:15.00                                3rd Qu.:26.00                   3rd Qu.:11.250   3rd Qu.:44.00  
 Max.   :18.00                                Max.   :55.00                   Max.   :44.500   Max.   :64.00  
      Race                Occupation                Sector          Married     
 Hispanic: 27   Clerical     : 97   Construction : 24   Married  :350  
 Other   : 67   Management   : 55   Manufacturing: 99   Unmarried:184  
 White   :440   Other        :156   Other        :411                  
                Professional:105                                       
                Sales        : 38                                      
                Service      : 83                                      
```

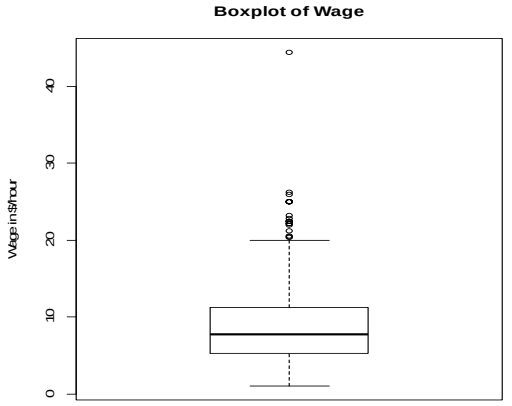
Five number summary + mean

$ selects a column

Text summarized as a table

```
> median(wages$Wage)
[1] 7.78
```

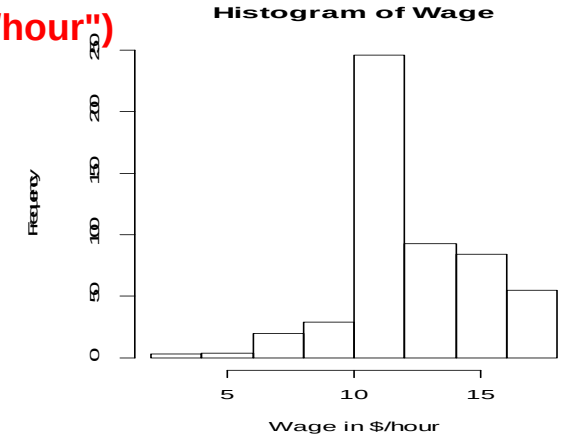
```
> boxplot(wages$Wage,main="Boxplot of Wage", ylab="Wage in $/hour")
```




```
> hist(wages$Education, main="Histogram of Wage", xlab="Wage in $/hour")
```

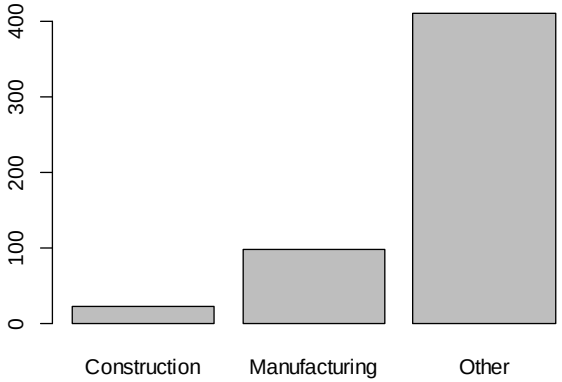
main = "title"

xlab = " x-label "



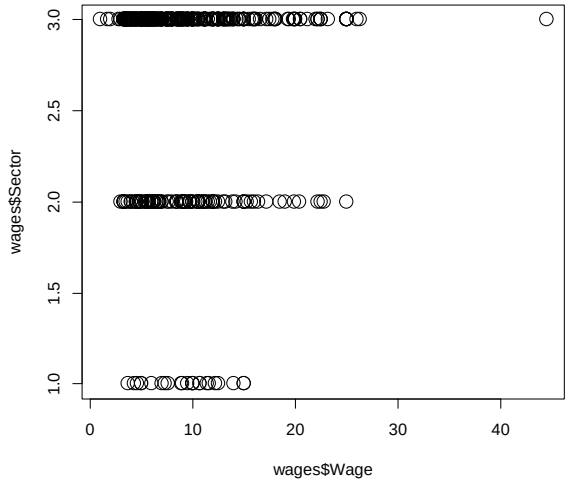
Flexibility of a generic function such as plot()

plot(wages$Sector)



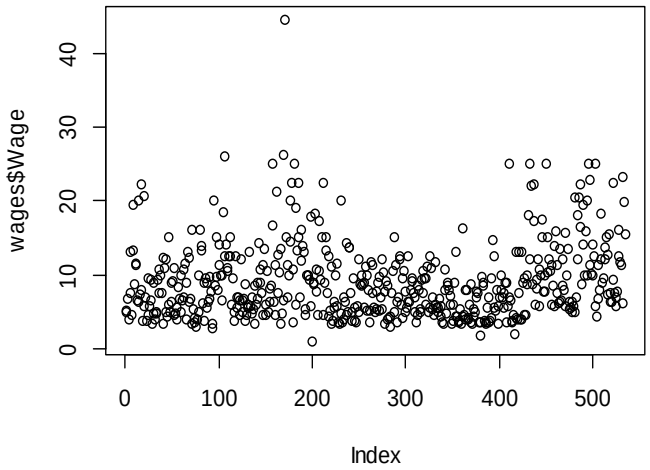
text

plot(wages$Wage,wages$Sector)



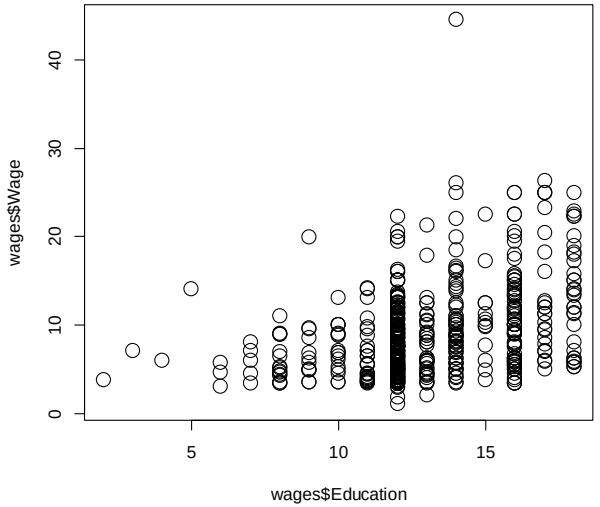
text
+
numeric

plot(wages$Wage)



numeric

plot(wages$Education,wages$Wage)



numeric
+
numeric

Further analysis

```
> tab1=table(wages$Union,wages$Married)
> tab1
            Married Unmarried
  Non_Union     278       160
  Union          72        24

> chisq.test(tab1)

        Pearson's Chi-squared test with Yates' continuity correction
data:  tab1
X-squared = 4.1384, df = 1, p-value = 0.04192

> plot(tab1)

> plot(tab1, main="Mosaic Plot of Union & Married", cex=1.5)
```

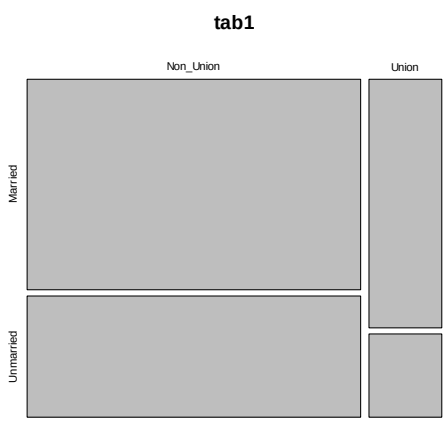
Crosstab of two factors

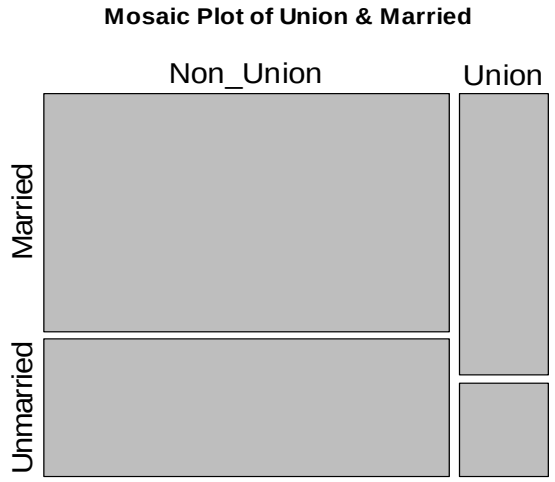
Output the table object

Use the object for a test

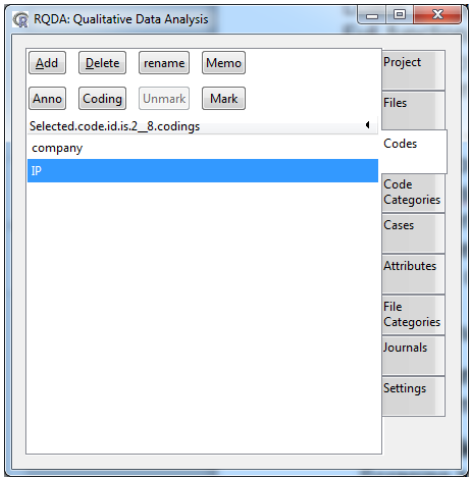
plot() is generic
'tries to do its best for you'

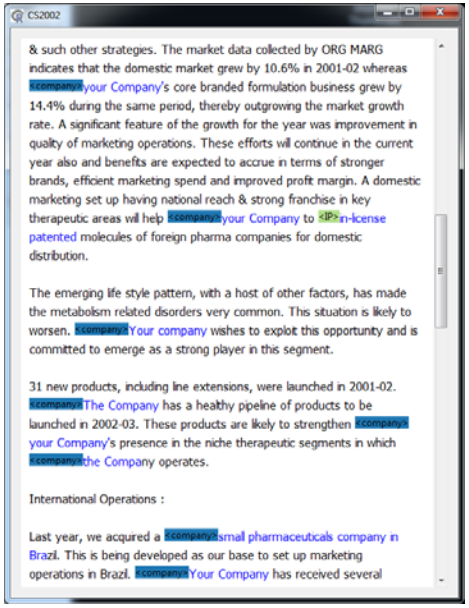
Tweak the default option





A Few Qualitative & Text Based Analysis Packages





'RQDA' - R Qualitative Data Analysis

(CAQDAS – Computer Aided Qualitative Data Analysis Software)

- Loads through R and runs on its own GUI (Graphic User Interface)
- Full functionality of R available alongside it
- Input transcriptions/texts and code them
- Visualization of codes and code categories

'tm' package - Text Mining package

- Forming corpora – many language bases available
- Preprocessing text – stopword and whitespace removal, etc.
- Forming term document matrices
- Count-based evaluation
- Document classification and clustering
- Synonyms
- Content analysis

'NLP' – Natural Language Processing

- Latent semantic analysis – scoring documents

'twitteR' package to interface with Twitter

- Scraping tweets from Twitter

'sentiment' – sentiment analysis

- Assessing sentiment – positivity/negativity and emotion classification

A Few Quantitative Packages

Base package does a huge range of

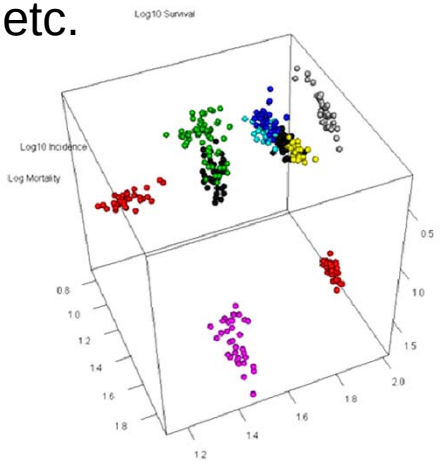
summarizing – tables, summary measures,
statistical testing – parametric (t-tests, Chi-squared, ANOVA, etc.)
– nonparametric (Mann-Whitney, Spearman, etc.)
model fitting – regression, logit, etc.
data manipulation – subsetting and merging tables, reshaping, etc.
plotting – histograms, barcharts, scatterplots, conditioning plots, etc.

'psych' package – factor analysis, item and scale analysis, etc.

'cluster' package - does a wide variety of cluster analysis methods
- dendrograms, cluster membership, etc.

'ggplot2' – high quality 2-D graphics

'rgl' – 3-D interactive graphics – use mouse cursor to rotate plots in 3-D

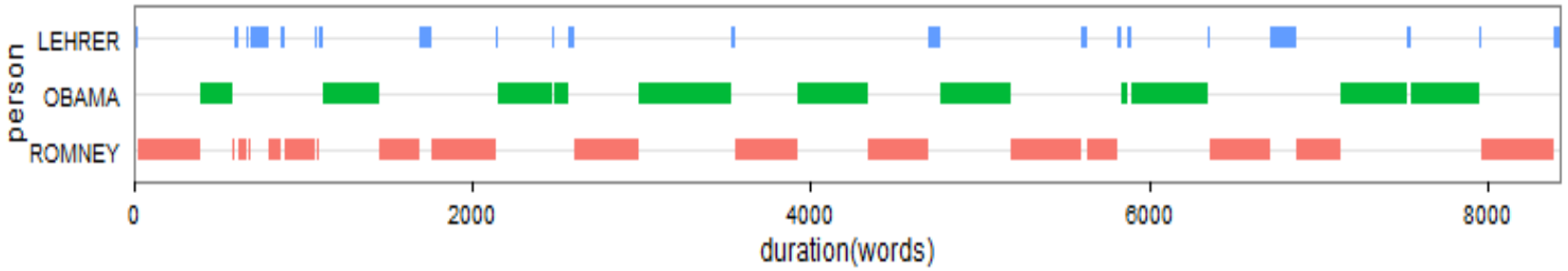


A Qualitative-Quantitative Bridge?

‘qdap’ – a Quantitative Discourse Analysis Package by Tyler Rinker

The example below is a graphic from the package author’s webpages and shows an analysis of one of the Obama-Romney presidential debates with additional analysis of word categories and a language formality index.

http://trinkerrstuff.wordpress.com/2012/10/04/presidential-debates-with-qdap-beta/



The plot above uses the ‘ggplot2’ graphics package mentioned on the previous slide.

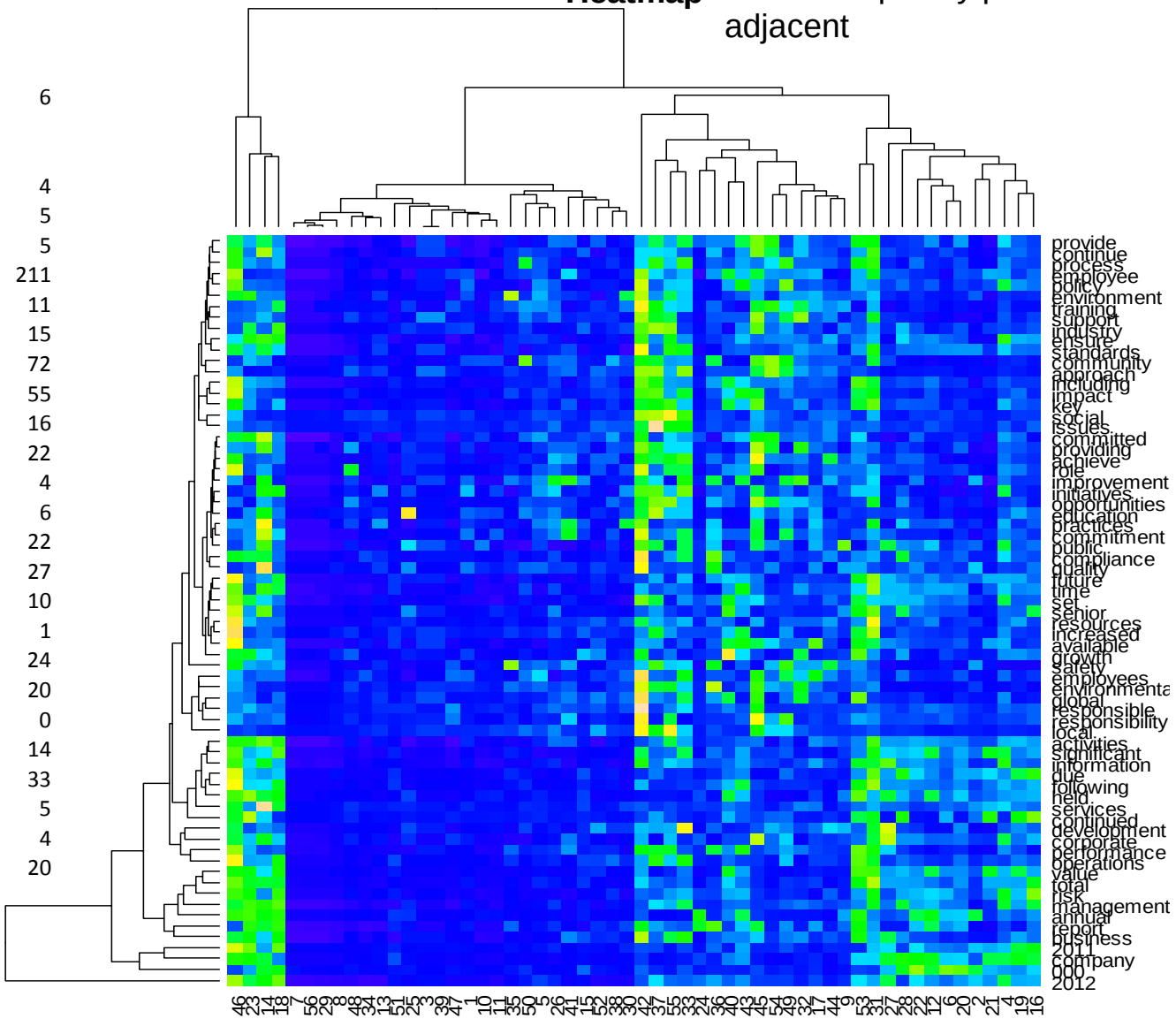
Content analysis + Regression

56 UK & Malaysian Company Corporate Social Responsibility Reports
Read into a corpus, preprocessed then tokenized and 1 – 2 word terms identified

Extract from Term Document Matrix
Showing frequency of terms

	Docs					
Terms	1	2	3	4	5	6
commitment	13	2	4	15	9	4
committed	5	5	6	12	12	5
community	21	6	2	11	18	5
company	13	94	13	234	11	211
compliance	5	6	5	33	5	11
continue	1	5	14	18	13	15
continued	4	44	2	11	12	72
corporate	24	36	6	117	14	55
development	12	9	23	49	15	16
due	2	45	2	39	1	22
education	10	2	2	9	9	4
employee	18	16	11	28	19	6
employees	45	18	14	47	82	22
ensure	7	8	8	69	20	27
environment	20	22	9	16	35	10
environmental	18	21	19	6	36	1
following	2	15	3	45	3	24
future	8	22	11	36	6	20
global	3	4	21	37	8	0
growth	13	5	5	28	1	14
held	2	23	1	42	11	33
impact	3	8	3	17	16	5
improvement	8	12	6	9	10	4
including	6	6	15	32	16	20

Heatmap



Heatmap sorts rows (columns) so that terms (firms) with similar frequency profiles are adjacent

Data collected by Ms. Jingju LU as part of her MSc Dissertation

Content analysis + **Regression**

For UK companies, how does frequency of **“human rights”** depend on **R**eturn **O**n **A**ssets, **M**arket **Cap**italization, **Liability** and **Current Liability**?

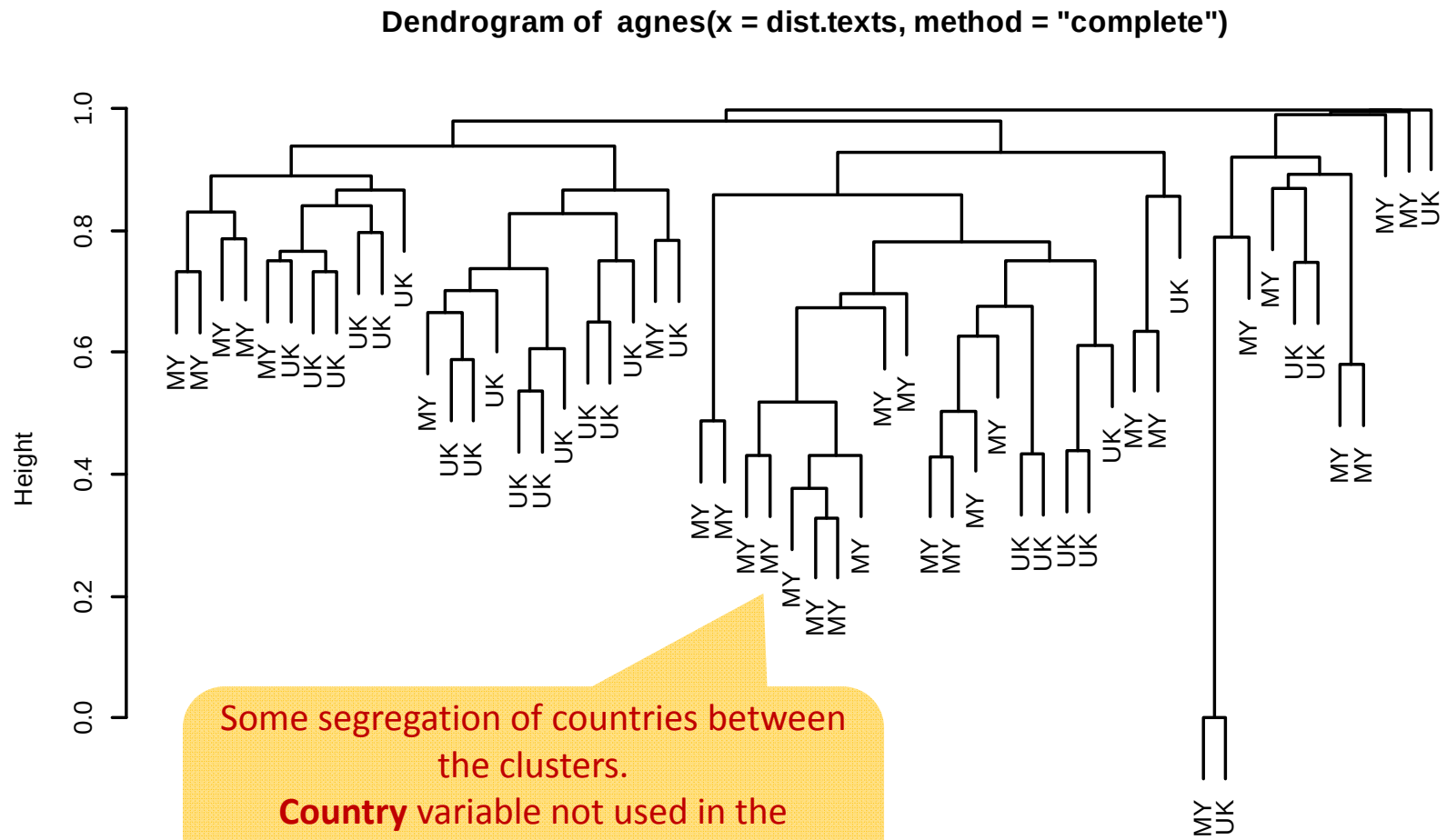
```
Zero hurdle model coefficients (binomial with logit link):
                Estimate     Pr(>|z|)
(Intercept)     -10.5525     <2e-16 ***
ROA              0.1477       0.017 *

Count model coefficients (truncated poisson with log link):
                Estimate   Pr(>|z|)
(Intercept)     -3.86       0.00078 ***
ROA              0.07705   1.05e-11 ***
log(MCAP)       -0.75385   7.86e-10 ***
log(CurrLiable) -1.28982   2.47e-11 ***
log(Liable)      2.02799    < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This tells us that the probability of the mere mention of “human rights” is increased by ROA

This tells us that, once human rights is mentioned, the frequency of its occurrence is **increased**/**decreased** as shown

Clustering

Complete linkage clustering of the Corporate Social Responsibility reports for companies in Malaysia and UK by pattern of co-occurrence of terms between documents



Some segregation of countries between the clusters.
Country variable not used in the clustering and, hence, this gives some external validation of the clusters

Getting Help and Ideas

Quick R

http://www.statmethods.net/

The Comprehensive R Archive Network

http://cran.r-project.org/ downloads and 'approved' packages

Journal of Statistical Software

http://www.jstatsoft.org/search Not all about R but many feature articles on R packages

R-Bloggers

http://www.r-bloggers.com/ excellent source of ideas and examples

R-help mailing list

https://stat.ethz.ch/mailman/listinfo/r-help

ROC Curve for mention of “human rights”