#### Using R for data analysis

#### Daniel Müllensiefen Goldsmiths, University of London

August 18, 2009

イロト イヨト イヨト イヨト

æ

Daniel Müllensiefen Goldsmiths, University of London Using R for data analysis

#### Introduction

What's it good for? R and its competitors Core characteristics History

Analysing data: The iris data example Getting data in Summarising data

回 と く ヨ と く ヨ と

æ

What's it good for? R and its competitors Core characteristics History

・ロト ・回ト ・ヨト ・ヨト

æ

# R is good for

Flexible Data Analysis (programmable)

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

æ

- Flexible Data Analysis (programmable)
- Using different analysis techniques

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

æ

- Flexible Data Analysis (programmable)
- Using different analysis techniques
- Data Visualisation

What's it good for? R and its competitors Core characteristics History

- 4 回 2 - 4 回 2 - 4 回 2 - 4

æ

- Flexible Data Analysis (programmable)
- Using different analysis techniques
- Data Visualisation
- Numeric Accuracy

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

- Flexible Data Analysis (programmable)
- Using different analysis techniques
- Data Visualisation
- Numeric Accuracy
- Rapid prototyping of analysis / process models

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

- Flexible Data Analysis (programmable)
- Using different analysis techniques
- Data Visualisation
- Numeric Accuracy
- Rapid prototyping of analysis / process models
- Pre-processing data from different sources
  - ▶ textfiles (.txt) and binary files (e.g. SPSS .sav, Excel)

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

- Flexible Data Analysis (programmable)
- Using different analysis techniques
- Data Visualisation
- Numeric Accuracy
- Rapid prototyping of analysis / process models
- Pre-processing data from different sources
  - textfiles (.txt) and binary files (e.g. SPSS .sav, Excel)
  - Audio files

What's it good for? R and its competitors Core characteristics History

- 4 回 2 - 4 □ 2 - 4 □

- Flexible Data Analysis (programmable)
- Using different analysis techniques
- Data Visualisation
- Numeric Accuracy
- Rapid prototyping of analysis / process models
- Pre-processing data from different sources
  - textfiles (.txt) and binary files (e.g. SPSS .sav, Excel)
  - Audio files
  - databases

What's it good for? R and its competitors Core characteristics History

- 4 回 2 - 4 □ 2 - 4 □

- Flexible Data Analysis (programmable)
- Using different analysis techniques
- Data Visualisation
- Numeric Accuracy
- Rapid prototyping of analysis / process models
- Pre-processing data from different sources
  - textfiles (.txt) and binary files (e.g. SPSS .sav, Excel)
  - Audio files
  - databases
  - texts (linguistic data)

What's it good for? R and its competitors Core characteristics History

・ロト ・回 ト ・ヨト ・ヨト

æ

## R is considered less good for

Graphical User Interfaces

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

æ

## R is considered less good for

- Graphical User Interfaces
- Internet programming

What's it good for? R and its competitors Core characteristics History

- - 4 回 ト - 4 回 ト

# R is considered less good for

- Graphical User Interfaces
- Internet programming
- Low-level programming

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

æ

# R is considered less good for

- Graphical User Interfaces
- Internet programming
- Low-level programming

► ...

What's it good for? R and its competitors Core characteristics History

・ロト ・ 日 ・ ・ ヨ ・ ・ ヨ ・ ・

æ

#### R compares to

Matlab (open source, community driven, not commercial)

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

#### R compares to

- Matlab (open source, community driven, not commercial)
- SPSS, SAS, Stata (programming language, not program)

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

#### R compares to

- Matlab (open source, community driven, not commercial)
- SPSS, SAS, Stata (programming language, not program)
- Weka (driven by community, not individuals)

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

### R compares to

- Matlab (open source, community driven, not commercial)
- SPSS, SAS, Stata (programming language, not program)
- Weka (driven by community, not individuals)
- SciPy and other software libraries (entire language specialised for data analysis)

What's it good for? R and its competitors Core characteristics History

< □ > < □ > < □ > < □ > < □ > .

æ

### Pros and Cons

Huge community support

Daniel Müllensiefen Goldsmiths, University of London Using R for data analysis

What's it good for? R and its competitors Core characteristics History

イロン イヨン イヨン イヨン

æ

### Pros and Cons

- Huge community support
- Cross-plattform and command-line based

What's it good for? R and its competitors Core characteristics History

イロン イヨン イヨン イヨン

### Pros and Cons

- Huge community support
- Cross-plattform and command-line based
- Interactive: interpreted not complied

What's it good for? R and its competitors Core characteristics History

- 4 回 2 - 4 回 2 - 4 回 2 - 4

## Pros and Cons

- Huge community support
- Cross-plattform and command-line based
- Interactive: interpreted not complied
- Mainly functional

What's it good for? R and its competitors Core characteristics History

イロン イヨン イヨン イヨン

æ

### How R came about

1976: John Chambers releases 1st version of S: Language for statistics, stochastic simulation and data visualisation

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

æ

- 1976: John Chambers releases 1st version of S: Language for statistics, stochastic simulation and data visualisation
- ▶ 1995: Ross Ihaka and Robert Gentleman release R as GPL

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

- 1976: John Chambers releases 1st version of S: Language for statistics, stochastic simulation and data visualisation
- ▶ 1995: Ross Ihaka and Robert Gentleman release R as GPL
- ▶ 1998: Comprehensive R Archive Network (CRAN) founded

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

- 1976: John Chambers releases 1st version of S: Language for statistics, stochastic simulation and data visualisation
- ▶ 1995: Ross Ihaka and Robert Gentleman release R as GPL
- ▶ 1998: Comprehensive R Archive Network (CRAN) founded
- 2001: R News published for 1st time

What's it good for? R and its competitors Core characteristics History

<ロ> <同> <同> < 同> < 同> < 同><<

- 1976: John Chambers releases 1st version of S: Language for statistics, stochastic simulation and data visualisation
- ▶ 1995: Ross Ihaka and Robert Gentleman release R as GPL
- ▶ 1998: Comprehensive R Archive Network (CRAN) founded
- 2001: R News published for 1st time
- > 2004: 1st *useR!* conference

What's it good for? R and its competitors Core characteristics History

イロト イヨト イヨト イヨト

- 1976: John Chambers releases 1st version of S: Language for statistics, stochastic simulation and data visualisation
- ▶ 1995: Ross Ihaka and Robert Gentleman release R as GPL
- ▶ 1998: Comprehensive R Archive Network (CRAN) founded
- ▶ 2001: R News published for 1st time
- > 2004: 1st *useR*! conference
- ▶ 2009: More than 1000 packages available on CRAN

Getting data in Summarising data

æ

#### Basic data in and out

Start R

Getting data in Summarising data

(4回) (1日) (日)

æ

### Basic data in and out

- Start R
- Save file from

 $\label{eq:http://www.doc.gold.ac.uk/~mas03dm/teaching/r/iris.data.txt to R's working directory (using getwd()$ 

Getting data in Summarising data

・ロン ・回と ・ヨン ・ヨン

2

### Basic data in and out

- Start R
- Save file from http://www.doc.gold.ac.uk/~mas03dm/teaching/r/iris.data.txt to R's working directory (using getwd()
- Get data into R using the read.table command (useful operations help(read.table) and assignment operator "
  —")

Getting data in Summarising data

### Basic data in and out

#### Start R

- Save file from http://www.doc.gold.ac.uk/~mas03dm/teaching/r/iris.data.txt to R's working directory (using getwd()
- Get data into R using the read.table command (useful operations help(read.table) and assignment operator "
  —")
- Change the species label of the 3rd observations to your own first name (using the indexing function [ , ]), save this dataset (using write.table())

イロン イヨン イヨン イヨン

Getting data in Summarising data

### Basic data in and out

#### Start R

- Save file from http://www.doc.gold.ac.uk/~mas03dm/teaching/r/iris.data.txt to R's working directory (using getwd()
- Get data into R using the read.table command (useful operations help(read.table) and assignment operator "
  —")
- Change the species label of the 3rd observations to your own first name (using the indexing function [ , ]), save this dataset (using write.table())
- Remove the altered dataset (using rm()) and get the original dataset in again

イロン イヨン イヨン イヨン

Getting data in Summarising data

・ロン ・回 と ・ ヨ と ・ ヨ と

æ

#### Data summary and plots

Summarise dataset (summary(), str())

Getting data in Summarising data

- 4 回 2 - 4 □ 2 - 4 □

æ

- Summarise dataset (summary(), str())
- Plot 1st column vs 2nd column (plot())

Getting data in Summarising data

・ 同 ト ・ ヨ ト ・ ヨ ト

- Summarise dataset (summary(), str())
- Plot 1st column vs 2nd column (plot())
- Attach dataset to search path (attach())

Getting data in Summarising data

(4月) (日) (日)

- Summarise dataset (summary(), str())
- Plot 1st column vs 2nd column (plot())
- Attach dataset to search path (attach())
- Plot Species vs Petal.Width and give graph a title and axes names

Getting data in Summarising data

- 4 周 ト - 4 日 ト - 4 日 ト

- Summarise dataset (summary(), str())
- Plot 1st column vs 2nd column (plot())
- Attach dataset to search path (attach())
- Plot Species vs Petal.Width and give graph a title and axes names
- Plot histogram of Petal.Length (hist()

Getting data in Summarising data

御 ト く ヨ ト

- Summarise dataset (summary(), str())
- Plot 1st column vs 2nd column (plot())
- Attach dataset to search path (attach())
- Plot Species vs Petal.Width and give graph a title and axes names
- Plot histogram of Petal.Length (hist()
- Plot scattergram of full dataset (plot(dataset,col=Species))

Getting data in Summarising data

- 4 周 ト - 4 日 ト - 4 日 ト

- Summarise dataset (summary(), str())
- Plot 1st column vs 2nd column (plot())
- Attach dataset to search path (attach())
- Plot Species vs Petal.Width and give graph a title and axes names
- Plot histogram of Petal.Length (hist()
- Plot scattergram of full dataset (plot(dataset,col=Species))
- Add non-parametric smoother (plot(dataset, col=Species, panel=panel.smooth))

Getting data in Summarising data

イロト イヨト イヨト イヨト

æ

### More plots and a function

Do boxplot(Petal.Length Species,notch=TRUE). What are the notches?

Getting data in Summarising data

イロト イヨト イヨト イヨト

æ

### More plots and a function

- Do boxplot(Petal.Length Species,notch=TRUE). What are the notches?
- Set the graphical device to be split into a 2x2 panel: op ← par(mfrow = c(2,2)