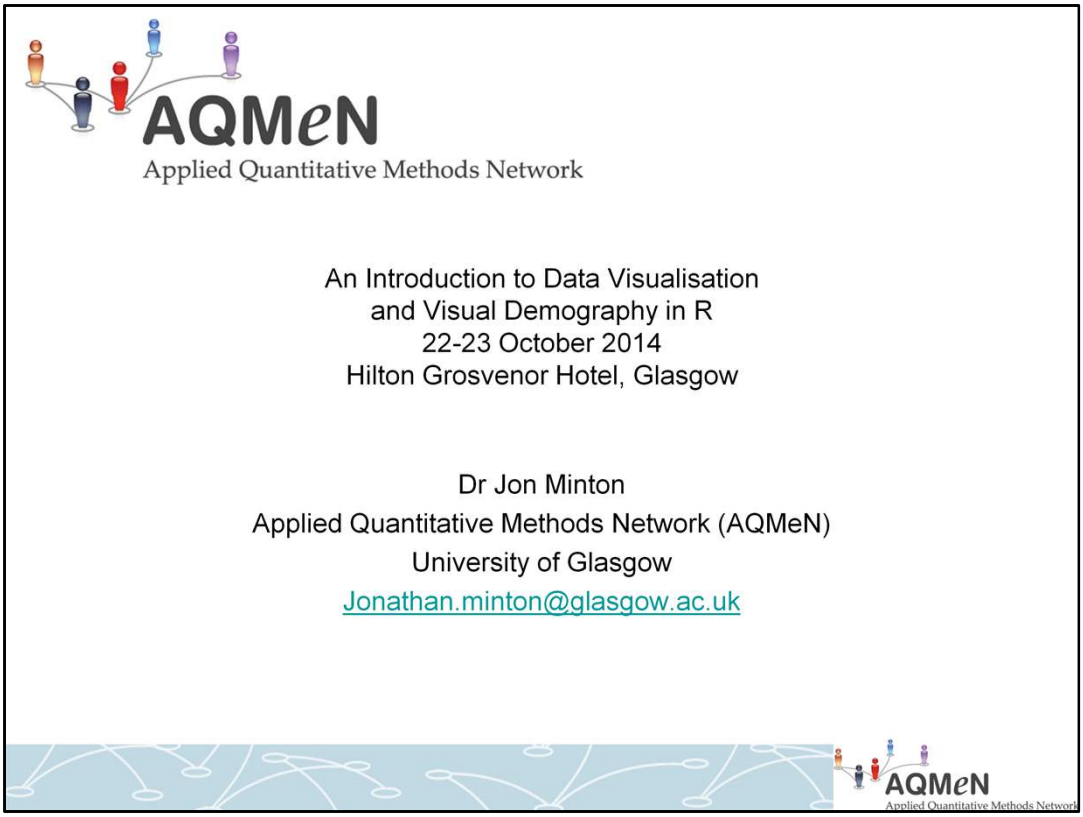
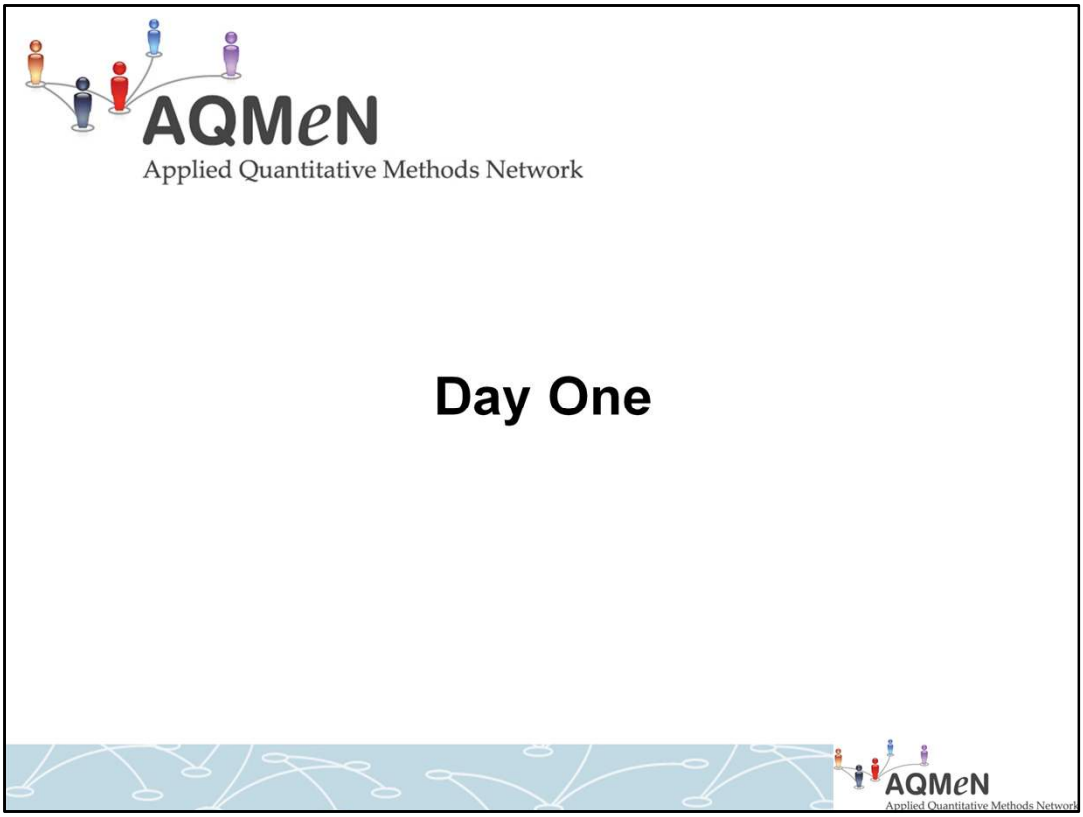
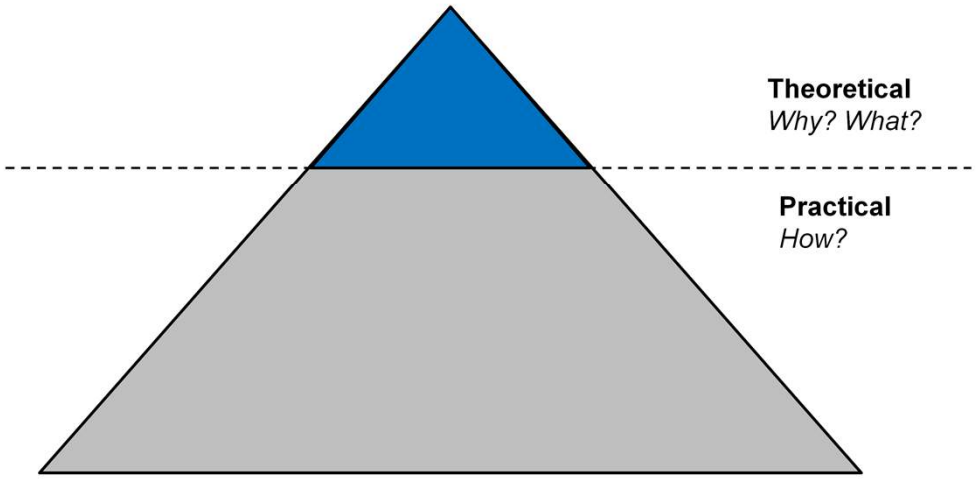

AQMeN
Applied Quantitative Methods Network
An Introduction to Data Visualisation
and Visual Demography in R
22-23 October 2014
Hilton Grosvenor Hotel, Glasgow
Dr Jon Minton
Applied Quantitative Methods Network (AQMeN)
University of Glasgow
Jonathan.minton@glasgow.ac.uk
AQMeN
Applied Quantitative Methods Network


AQMeN
Applied Quantitative Methods Network
Day One
AQMeN
Applied Quantitative Methods Network

Structure of this workshop

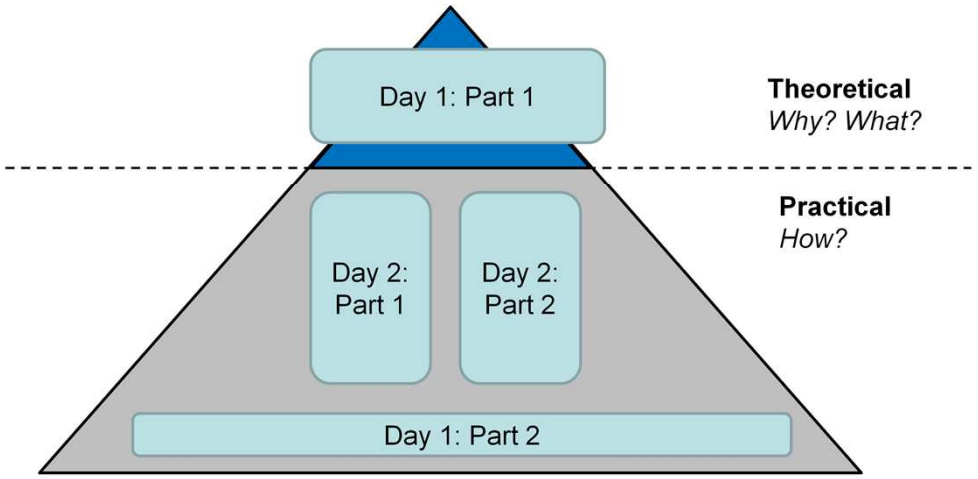
- Two Days; two distinct parts per day
- Day One:
  - Part One: Theory of Data Visualisation
  - Part Two: Introduction to R
    - R Installation Clinic
- Day Two:
  - Part One: ggplot2
  - Part Two: Visual Demography
    - *Scope to work more on ggplot2 if preferred*

Guiding Principles and Structure of this workshop



AQMeN
Applied Quantitative Methods Network

Guiding Principles and Structure of this workshop



AQMeN
Applied Quantitative Methods Network

Approach for this workshop

- INTERRUPTIONS are fine
- Talking concentrated in the first half of the first day

- Approach to the practical's:
  - Organic: if it's useful and there's time, keep going
  - Frequent feedback: check in with the group every 10 or so minutes
  - Nobody left behind
  - Learning as teaching; teaching as learning

Introductions

- Who are you?
- Where are you from and what do you do?
- Why are you interested in data visualisation?
- What do you understand by the term ‘data visualisation’?
- What do you want to achieve by the end of the workshop?
- What do want to achieve regarding data visualisation in the longer term?

Cutting to the Chase: Main Arguments

- Information visualisation *is not* data visualisation. (But data visualisation is a type of information visualisation)
- Data visualisation has a grammar and comes in layers
- We can't have too much information, but we can have too much data; data visualisation (and statistics) is about making data more informative
- Good data visualisations are good matches
- You need to wear Three Hats (and will spend most of your time wearing a Hard Hat)

(Mis)information is Beautiful

AQMeN
Applied Quantitative Methods Network

(Mis)information is Beautiful

- Tim Harford
  - Economist
  - Presenter of Radio 4's More or Less
  - Argues we should beware of data visualisation
  - Data visualisation dazzles, rather than informs

http://view6.workcast.net/?pak=8615439470431294#

AQMeN
Applied Quantitative Methods Network

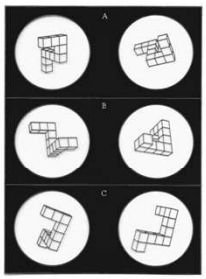
My position

- I agree (mostly)
- But you're using the wrong term.
- You're talking mainly about 'infographics' rather than data visualisations
- Infographics are NOT data visualisations
  - Infographics are information visualisations
  - Data visualisations are information visualisations
  - But (most) infographics are not data visualisations
- So, what makes a data visualisation a data visualisation rather than some other kind of information visualisation?

AQMeN
Applied Quantitative Methods Network

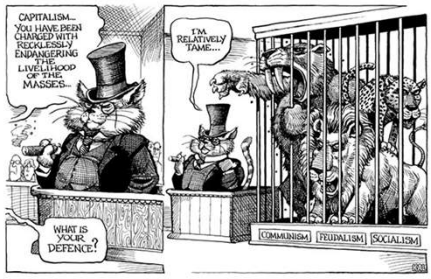
Thought is Visual

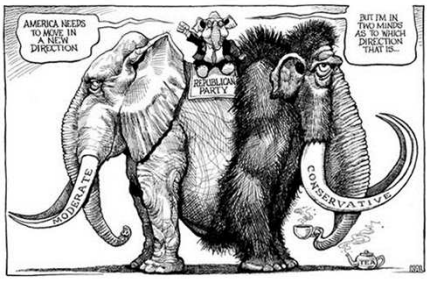
Stephen Kosslyn



George Lakoff

Examples



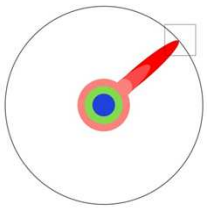


AQMeN
Applied Quantitative Methods Network

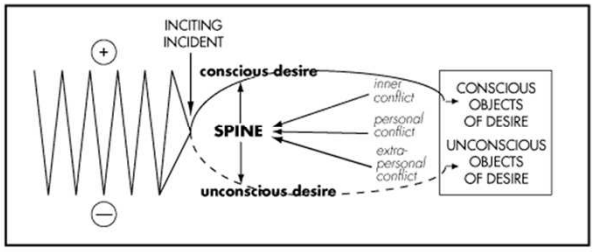
Examples

NOUN

NOUN VERBING



Robert McKee "THE QUEST"



AQMeN
Applied Quantitative Methods Network

Time is spatial...

Then → *Now*

...but is it cultural?

Then → Now

The quick brown fox jumps over the lazy dog

Now ← Then

أدنا ملزمون، بسبب جنسيتك، باخطار الممثليين القنصليين لدولتك هنا فى الولايات المتحدة الامريكية بأنه قد تم القاء القبض عليك او احتجازك. بعد اخطار المسؤولين القنصليين لدولتك، قد يقوموا بالاتصال بك هاتفياً او بزيارتك. انت غير ملزم بقبول مساعدتهم، ولكنهم قد يستطيعوا، من بين اشياء اخرى، مساعدتك فى الحصول على محام، والإتصال بعائلتك. وزيارتك فى مكان احتجازك. سنقوم باخطار الممثليين القنصليين لدولتك بأسرع ما يمكن.

Now

↑

Then

AQMeN
Applied Quantitative Methods Network

Statistics is...

- Data
- A summary of the data
- A model used to summarise the data
- Inferences, predictions, projections based on the model used to summarise the data

Statistics is...

(mostly)

DATA

reduction

Statistical Inference and Data Visualisation

- Complementary Roles

- Statistical Inference *lowers the hurdle*

- Data Visualisation *lets people jump higher*

AQMeN
Applied Quantitative Methods Network

Data Visualisation is…

- … you tell me

Let's (briefly) explore the history of data visualisation

History of Data Visualisation

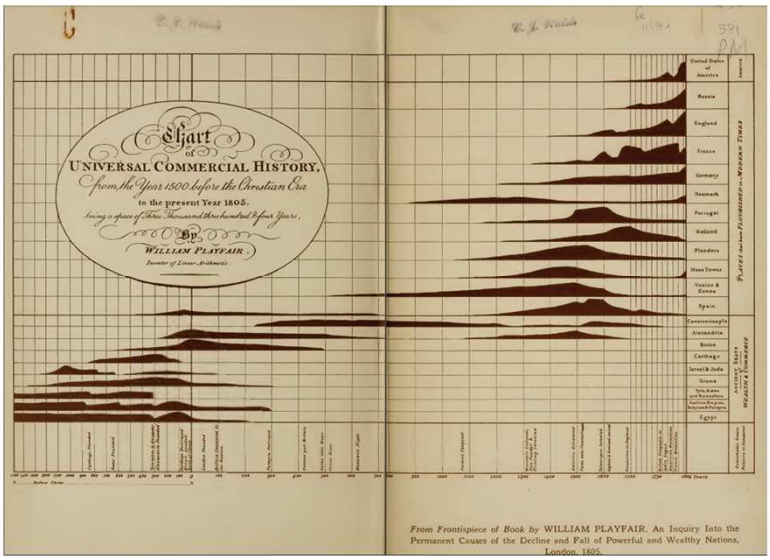
- The Milestones Project
  - 1644: First data graph (Michael Langren)
  - 1701: First contour map (Edmond Halley)
  - 1752: Developments in contour maps (Phillippe Buache)
  - 1753: Annotated timelines (Jacques Barbeu-Dubourg)
  - 1765: Annotated biographical timelines (Joseph Priestley)
  - 1786: Bar charts, and line graphs (**William Playfair**)
    - Let's explore Playfair's visualisations in more detail...

William Playfair

Born 1759
Died 1823

- Architect
- Mathematician
- Economist

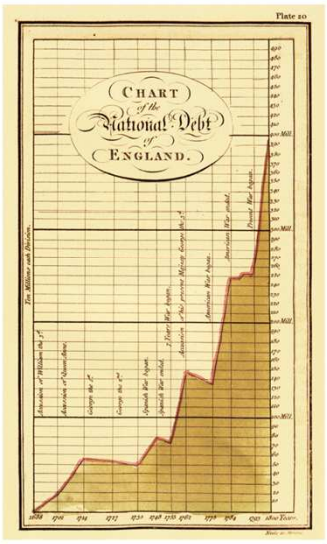
William Playfair



From Frontispiece of Book by WILLIAM PLAYFAIR. An Inquiry Into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations, London. 1805.

AQMeN
Applied Quantitative Methods Network

William Playfair



Note the use of

proportions

Visualisations as rhetoric

AQMeN
Applied Quantitative Methods Network

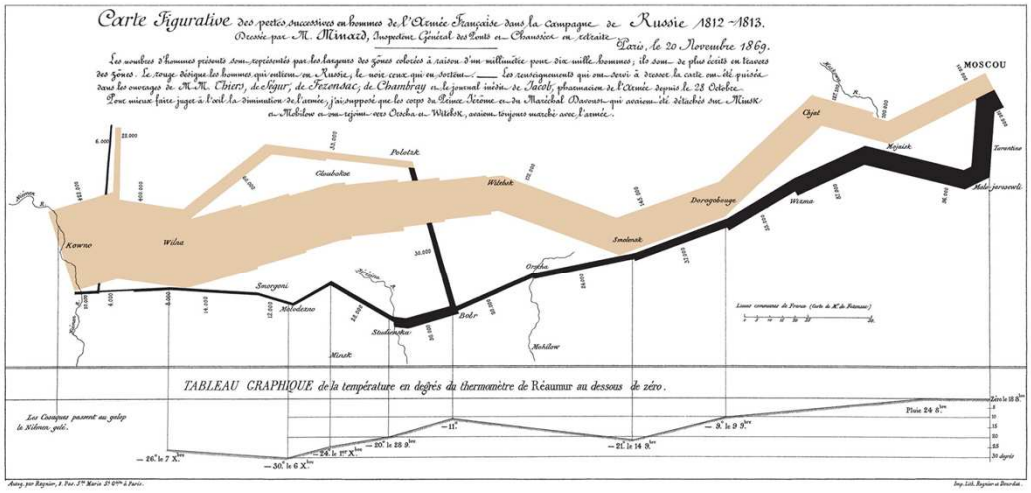
Charles Minard

Born 1781
Died 1870

- Civil Engineer until 1851
- Retired* from 1852

** This is when he produced most of his visualisations*

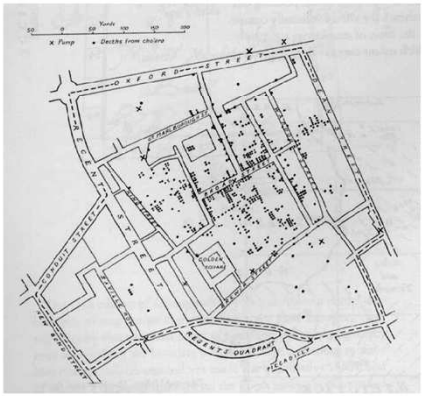
Charles Minard



"...the best statistical graphic ever drawn"? – Edward Tufte

History of Data Visualisation

- The Milestones Project (Continued)
  - 1855: Dot maps to identify origin of a cholera outbreak in London (**John Snow**)

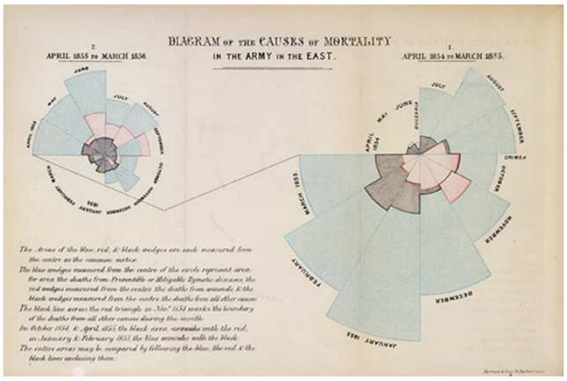


Dots save lives

AQMeN
Applied Quantitative Methods Network

History of Data Visualisation

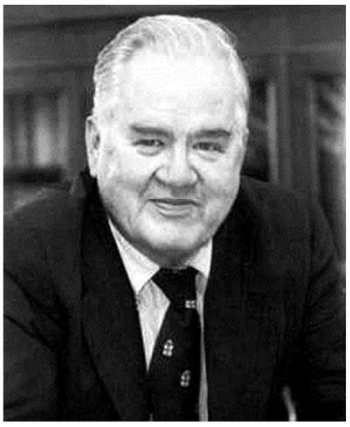
- The Milestones Project (Continued)
  - 1857: Polar area charts AKA coxcombs (**Florence Nightingale**)



Wedges save lives

AQMeN
Applied Quantitative Methods Network

John Tukey



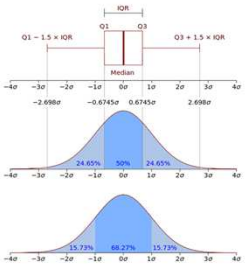
Born 1915
Died 2000

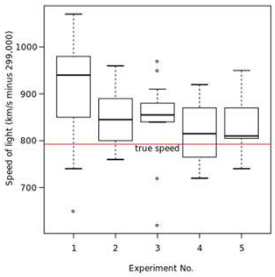
- Mathematician
- (Electronic Engineer)

Contributions to data visualisation:

- Box plots
- 'Exploratory Data Analysis'

John Tukey



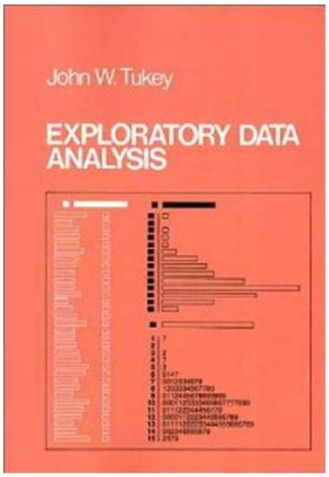


Box plots

- Visual representation of a number of key statistics used to summarise a sample of observations

AQMeN
Applied Quantitative Methods Network

John Tukey



Exploratory Data Analysis

Crudely:

- A series of (mainly visual) methods for trying to be less dependent on summary statistics and the assumptions used to construct them.

AQMeN
Applied Quantitative Methods Network

Why use visualisation for Exploratory Data Analysis?

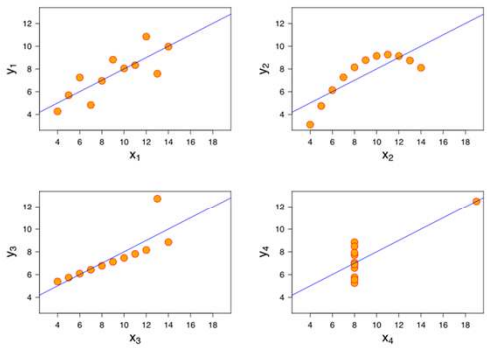
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The Anscombe Quartet (1973)

AQMeN
Applied Quantitative Methods Network

Why use visualisation for Exploratory Data Analysis?

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

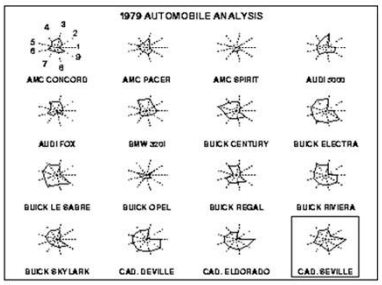


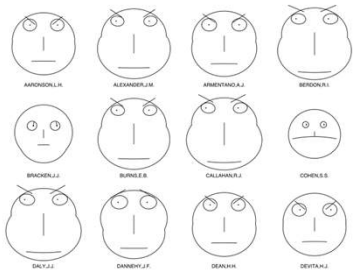
The Anscombe Quartet (1973)

AQMeN
Applied Quantitative Methods Network

History of Data Visualisation

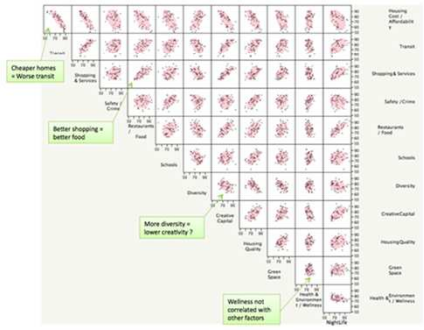
- The Milestones Project (Continued)
  - 1971: Irregular polygons AKA Star Plots (Friedmen & Siegel)
  - 1973: Cartoon faces (Herman Chernoff)

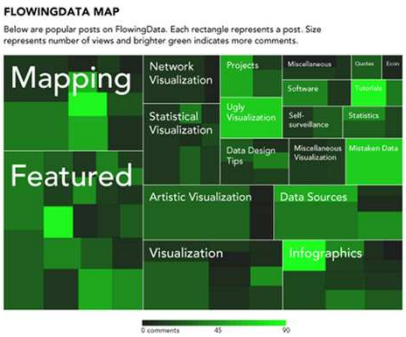




History of Data Visualisation

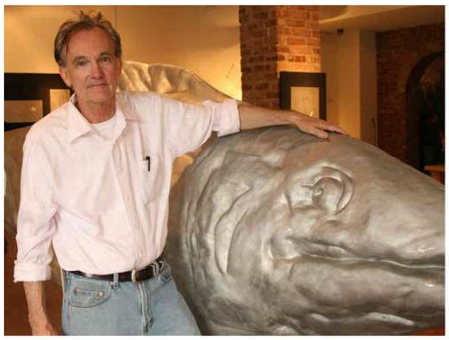
- The Milestones Project (Continued)
  - 1975: Scatterplot matrices – tables and graphs (John Hartigan)
  - 1981: Mosaic plots (Hartigan & Kleiner)





AQMeN
Applied Quantitative Methods Network

Edward Tufte



Born 1942

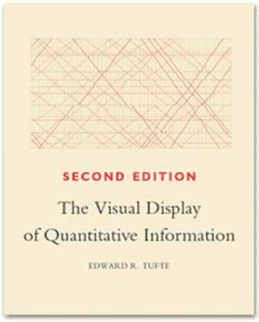
- Statistician
- Sculptor
- Advisor to Obama

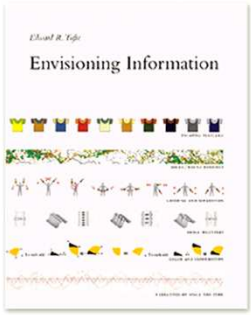
Contributions to data visualisation

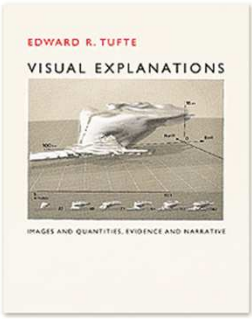
- Conceptual Tools
  - *Data Ink Ratios*
  - *Lie Factors*
  - *Chart Junk*
- Visualisations
  - *Small multiples*
  - *Sparklines*

AQMeN
Applied Quantitative Methods Network

Books by Edward Tufte







AQMeN
Applied Quantitative Methods Network

Data Visualisation is…

- You tell me. (Again)
- Nudging questions:
  - What is not a data visualisation?
  - Data visualisation is, or data visualisations are?

Data Visualisation is…

1. …the development and consistent application of rules for mapping between data and graphics.

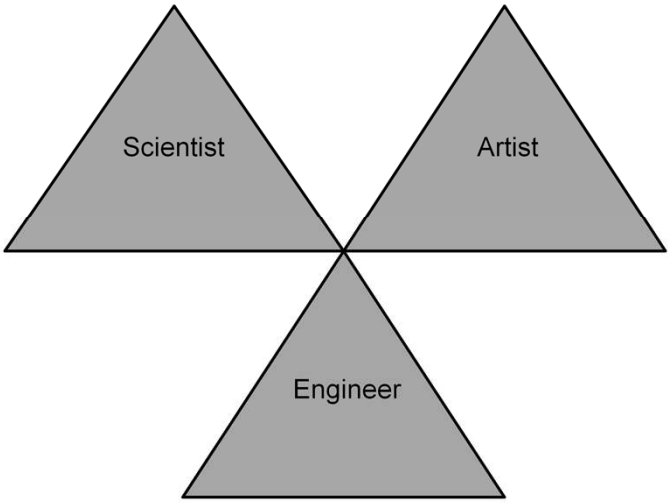
2. (equivalently) ... using quantitative (statistical) data as *instructions* for the production of images.

- A *necessary* condition for something to be a data visualisation; but not a *sufficient* condition for something to be a good data visualisation.

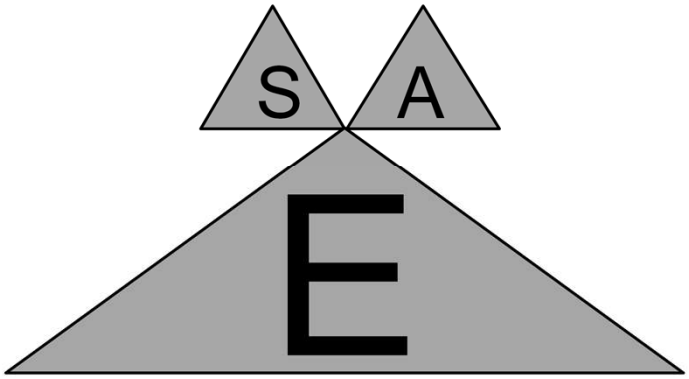
The Three Hats

- Three different skill sets
- Three different mind sets
- What are they?

The Three Hats

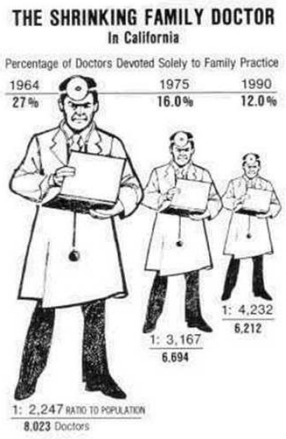


The Engineer must have broad shoulders



AQMeN
Applied Quantitative Methods Network

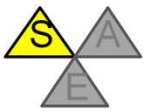
Are your visualisations...lying?



What's the 'lie factor'?
- Edward Tufte

"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented"

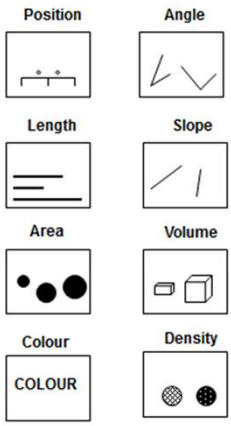
Tufte 1991

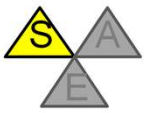


AQMeN
Applied Quantitative Methods Network

Are your visualisations...Accurately interpretable?

	Quantitative	Ordinal	Nominal
more	Position	Position	Position
↓	Length	Density	Colour Hue
↓	Angle	Colour Saturation	Texture
↓	Slope	Colour Hue	Connection
↓	Area	Texture	Containment
accu-	Volume	Connection	Density
rate	Density	Containment	Colour Saturation
↓	Colour saturation	Length	Shape
↓	Colour Hue	Angle	Length
↓	Texture	Slope	Angle
↓	Connection	Area	Slope
↓	Containment	Volume	Area
less	Shape	Shape	Volume

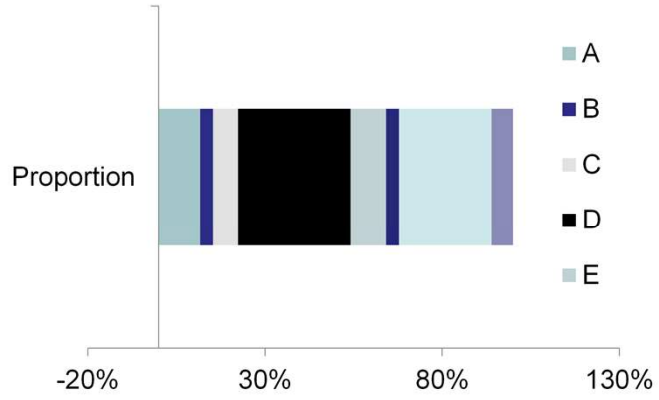




Cleveland W & McGill R (1985), 'Graphical Perception and Graphical Methods for Analysing Scientific Data', *Science* (229) 4716: 828-833

AQMeN
Applied Quantitative Methods Network

Are your visualisations...Sensible?

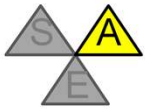


AQMeN
Applied Quantitative Methods Network

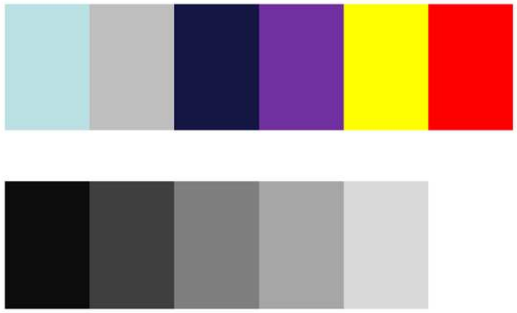
Visualisations and colour

Colour and accessibility

Why is Facebook blue?

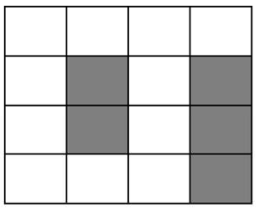


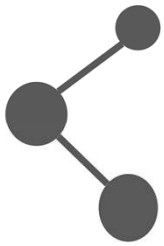
Colour and appropriateness to variables



AQMeN
Applied Quantitative Methods Network

Circles and Squares: A faultline?





AQMeN
Applied Quantitative Methods Network

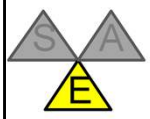
Data Management

- Very important
- Useful things to know
  - Importing from plain text
  - Importing from web tables
  - Exporting
  - Cleaning
  - Formatting
  - Rearranging Data

Tidy Data

- Database theory

- Can we distinguish parts of a table into:
  - Keys
    - Something that uniquely identifies an observation
  - Variables
    - Something that we measure when we make observations
  - Values
    - The result of measuring a specific variable in a specific observation



Making Data 'Tidy'

- What are the keys, variables, and values?

Surname	First name	Month of trial	Height	Condition	Weight
Smith	Mary	1	78	Healthy	45
Smith	John	2	62	Sick	52
Jones	William	1		Healthy	
Jones	William	2	86		84
Smith	Mary	3	78	Sick	43

Making Data 'Tidy'

Variables

Surname	First name	Month of trial	Height	Condition	Weight
Smith	Mary	1	78	Healthy	45
Smith	John	2	62	Sick	52
Jones	William	1		Healthy	
Jones	William	2	86		84
Smith	Mary	3	78	Sick	43

Values

Keys

AQMeN
Applied Quantitative Methods Network

Keys: Unique identifiers of observations

Surname	First name	Month of trial	Key
Smith	Mary	1	Smith_Mary_1
Smith	John	2	Smith_John_2
Jones	William	1	Jones_William_1
Jones	William	2	Jones_William_2
Smith	Mary	3	Smith_Mary_3

Keys: Unique identifiers of observations

Surname	First name	Month of trial	Key
Smith	Mary	1	Smith_Mary_1
Smith	John	2	Smith_John_2
Jones	William	1	Jones_William_1
Jones	William	2	Jones_William_2
Smith	Mary	3	Smith_Mary_3

Long format data

- 'Clay': Something that is easy to reform into many alternative structures

Key	Variable	Value
Smith_Mary_1	Height	78
Smith_Mary_1	Condition	Healthy
Smith_Mary_1	Weight	45
Smith_John_2	Height	62
Smith_John_2	Condition	Sick
Smith_John_2	Weight	52
Jones_William_1	Condition	Healthy
Jones_William_2	Height	86
Jones_William_2	Weight	84
Smith_Mary_3	Height	78
Smith_Mary_3	Condition	Sick
Smith_Mary_3	Weight	43

- A good format for data storage and archiving

‘Casting’ Data

Person	Month 1	2	3	4	5	6	7	8	9	10
A	78	79	75	69	64	65	62	59	59	62
B	56	59	62	61	55	58	63	67	69	62
C	92	99	105	101	104	111	114	105	97	99
D	48	49	49	56	59	55	63	66	61	52
E	66	70	64	62	68	63	64	64	62	67
F	99	98	87	88	86	85	86	82	79	85
G	154	160	161	154	148	149	142	139	137	133

Adding Summary Margins

Person	1	2	3	4	5	6	7	8	9	10
A	100	101	96	88	82	83	79	76	76	79
B	100	105	111	109	98	104	113	120	123	111
C	100	108	114	110	113	121	124	114	105	108
D	100	102	102	117	123	115	131	138	127	108
E	100	106	97	94	103	95	97	97	94	102
F	100	99	88	89	87	86	87	83	80	86
G	100	104	105	100	96	97	92	90	89	86
Mean	100	104	102	101	100	100	103	102	99	97

A Question

- What is a good data visualisation?

The Answer

- It depends
- Good data visualisations don't have to be:
  - Quick
  - Simple
  - Easy to understand
  - Original
  - Eye-catching
- What matters is how well matched the visualisation is with the audience

Two types of audience

- Internal/Specialist
  - Data Visualisation as Exploratory Data Analysis
  - A recursive, two-way process
- External/Generalist
  - 'Narrative' Data Analysis
  - A one-way process

The Complexity Challenge

Social Research Association

Unidimensional ← → *Multidimensional*

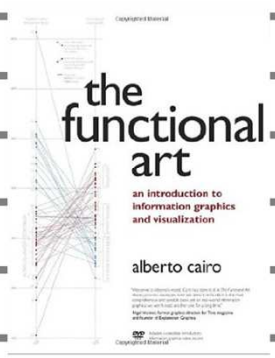
Light ← → *Dense*

Familiar ← → *Original*

Figurative ← → *Abstract*

Decorative ← → *Functional*

Redundancy ← → *Novelty*



AQMeN
Applied Quantitative Methods Network

Data Visualisation is...

... using quantitative (statistical) data as *instructions* for the production of images

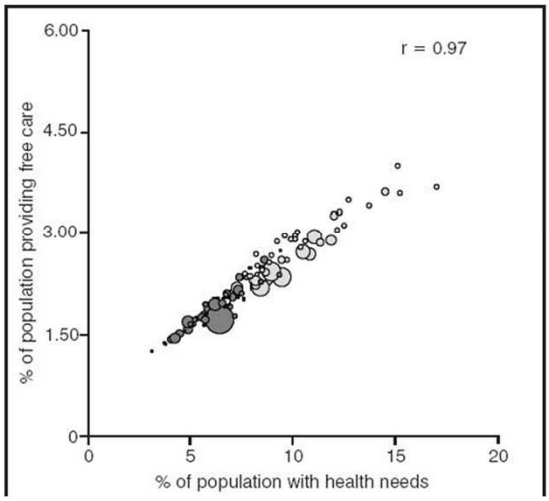
QUESTIONS

- What data?
- How are the data converted into instructions?
- Who/what is being instructed?
- What image?

As an example...

How are data being converted into instructions here?

How many *dimensions* does this visualisation have?



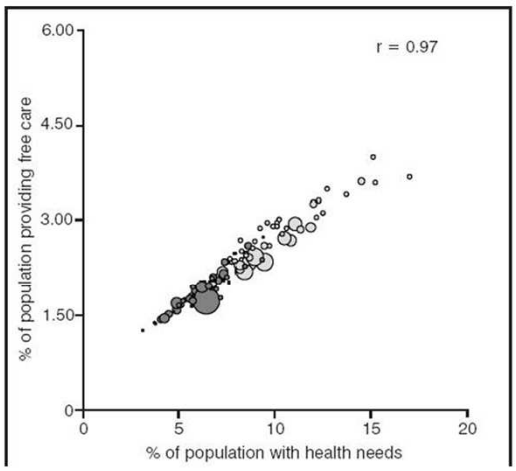
AQMeN
Applied Quantitative Methods Network

As an example...

How are data being converted into instructions here?

How many *dimensions* does this visualisation have?

Each circle is a county, unitary, or former metropolitan authority drawn in proportion to its population in 2001. Circles are shaded light if they lie west or north of the counties of Gloucestershire, Warwickshire, Leicestershire, and Lincolnshire (the Severn–Wash divide). Circles are drawn with area in proportion to total population (with the largest circle representing London). Each circle is positioned on the x-axis according to the proportion of the population who live there who have both poor health and limiting long-term illness, and on the y-axis according to the proportion of the population who live there who provide 50 hours' or more per week unpaid care, which includes: looking after, giving help or support to family members, friends, neighbours or others, because of long-term physical or mental ill-health or disability or problems relating to old age. On each of the graphs only the y-axis alters. The x-axis and the size and number of each circle remains constant.



Shaw M & Dorling D (2004) "Who cares in England and Wales? The Positive Care Law: cross-sectional study" *British Journal of General Practice,* 54 (509): 899-903

AQMeN
Applied Quantitative Methods Network

Data Visualisation as Box Wiring

Data Variable		Graphical Aesthetic
% of population providing free care		Position along horizontal axis
% of population with health needs		Position along vertical axis
Areal unit population		Size of bubble
Geographical location (North or south)		Colour of bubble

Data Visualisation as Box Wiring

Data Variable		Graphical Aesthetic
% of population providing free care		Position along horizontal axis
% of population with health needs		Position along vertical axis
Areal unit population		Size of bubble
Geographical location (North or south)		Colour of bubble

Double-Encoding for Emphasis

Data Variable		Graphical Aesthetic
% of population providing free care		Position along horizontal axis
% of population with health needs		Position along vertical axis
Areal unit population		Size of bubble
Geographical location (North or south)		Colour of bubble

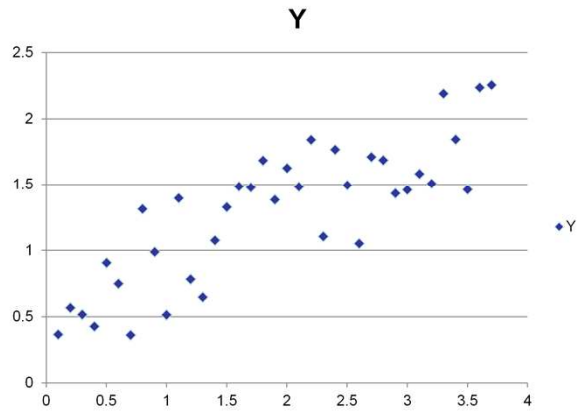
An Exercise (Time permitting)

- Find a data visualisation and ‘wire the boxes’
- Data Variables on the left hand side
- Graphical Aesthetics on the right hand side
- Draw ‘wires’ connecting them
- Time: 10-15 minutes?

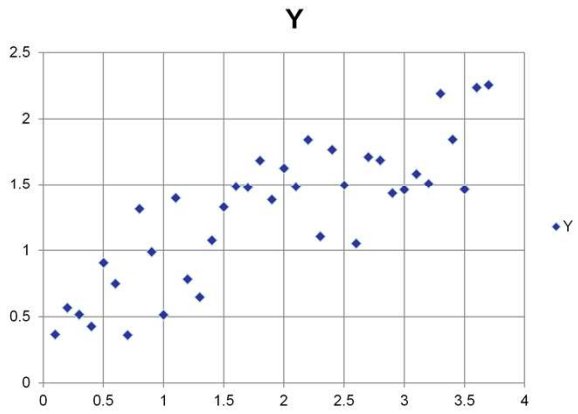
Layers of Data Visualisation

- The Guide Layer
  - Scales
  - Coordinate Systems
  - Legends
- The Data Layer
  - Transformations
  - Aesthetics
  - Geometrics
- The Annotation Layer
  - Everything else

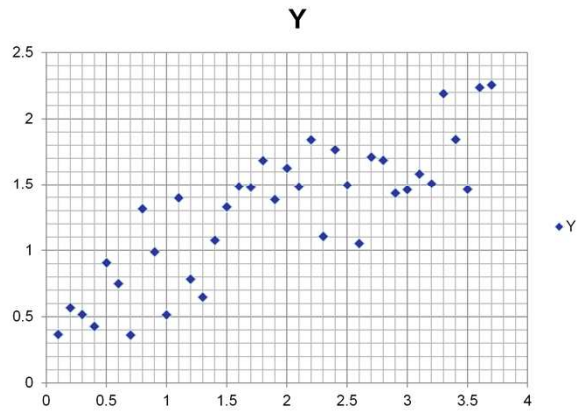
The Guide Layer: Excel Default



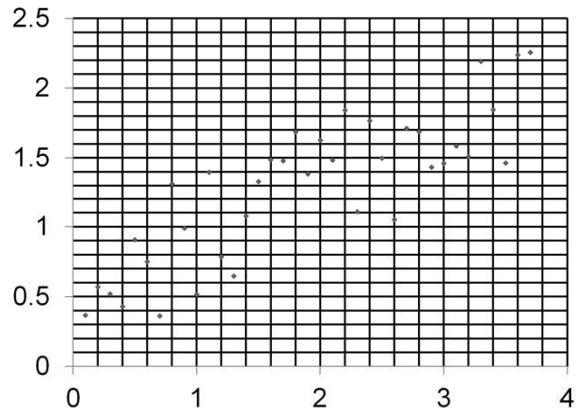
The Guide Layer: Major Gridlines



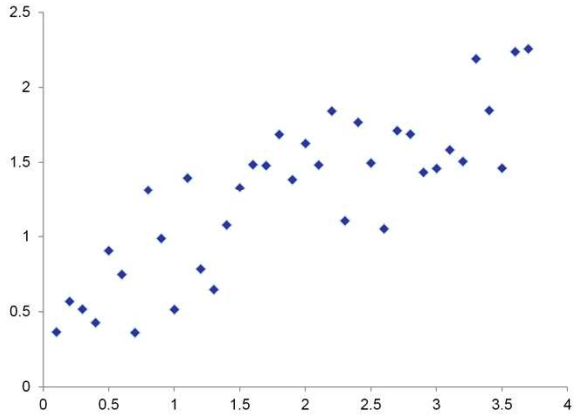
The Guide Layer: Major and Minor Gridlines



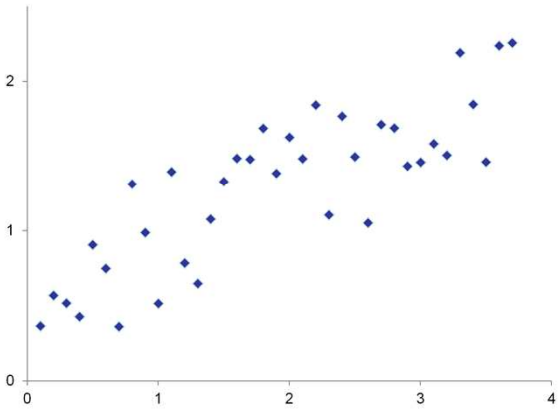
The Guide Layer: Major and Minor Gridlines



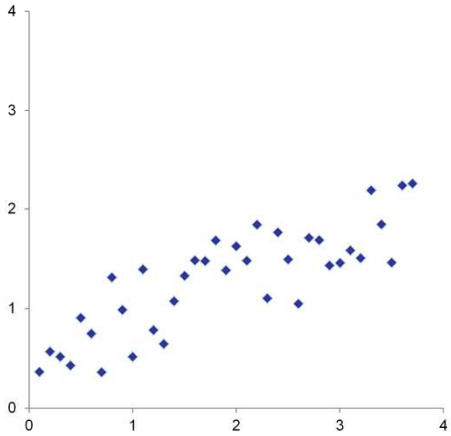
The Guide Layer: Gridlines Removed



The Guide Layer: Unit interval tickmarks

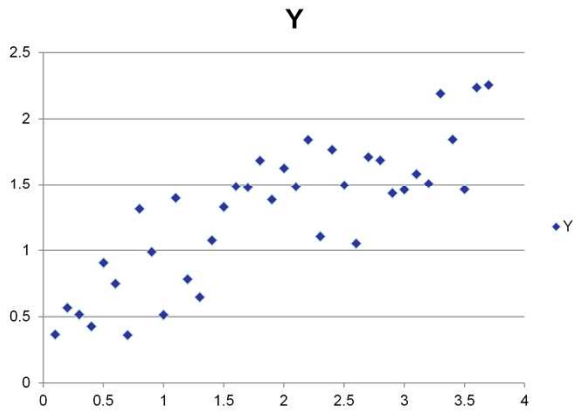


The Guide Layer: 1:1 Ratio

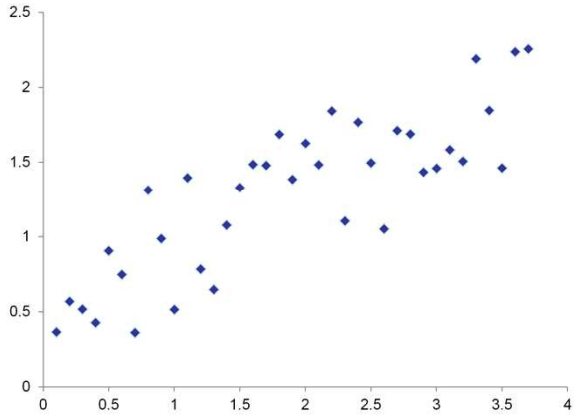


AQMeN
Applied Quantitative Methods Network

The Support Layer: Excel Default

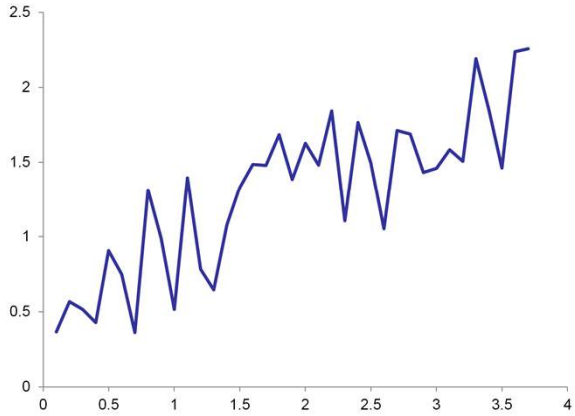


The Geom: Points



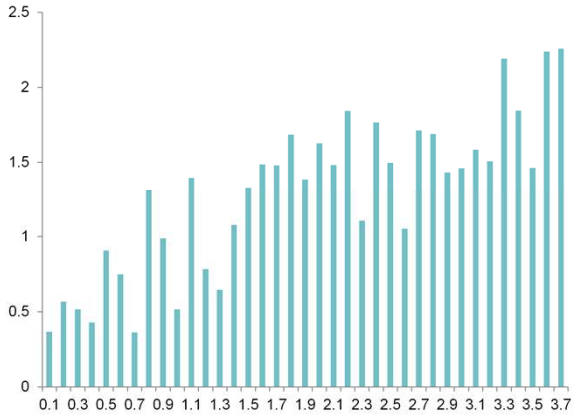
AQMeN
Applied Quantitative Methods Network

The Geom: Lines



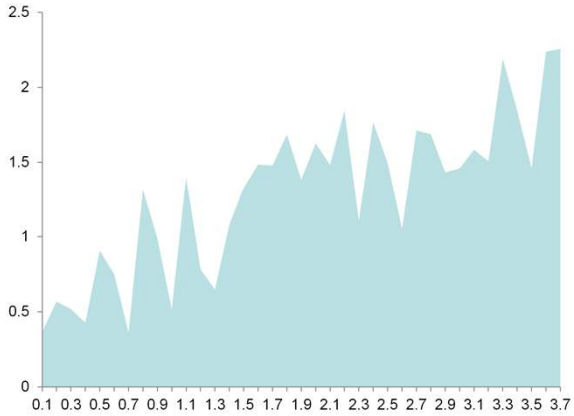
AQMeN
Applied Quantitative Methods Network

The Geom: Bars



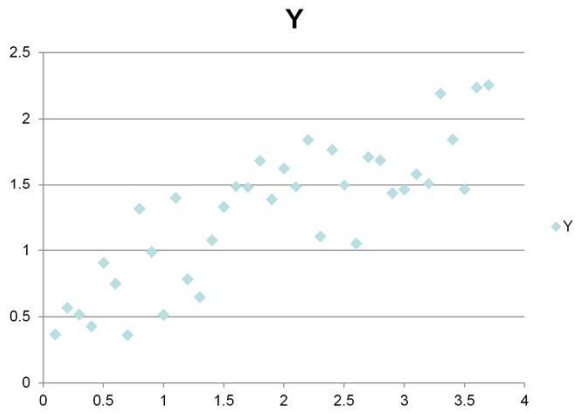
AQMeN
Applied Quantitative Methods Network

The Geom: Area

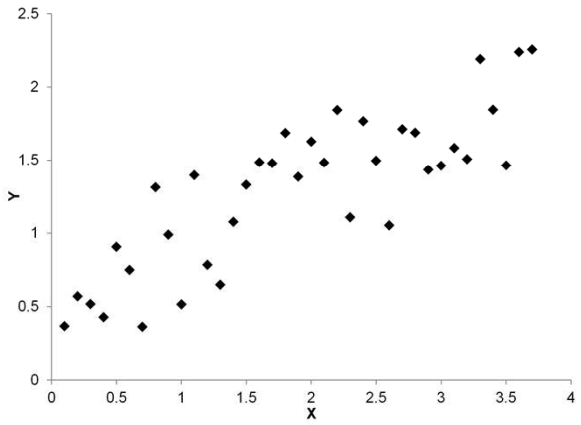


AQMeN
Applied Quantitative Methods Network

The Annotation Layer



The Annotation Layer



The Annotation Layer

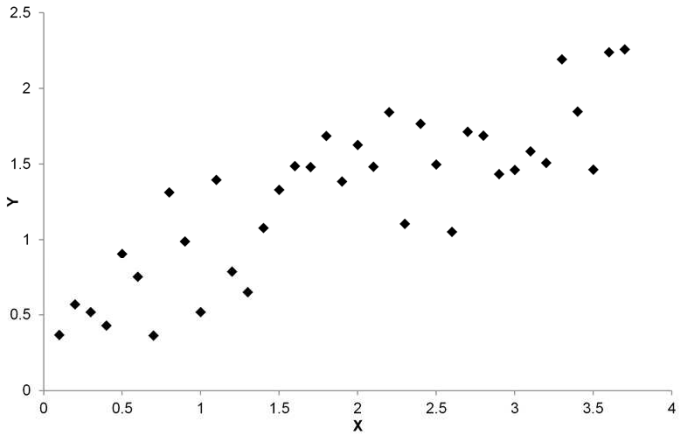


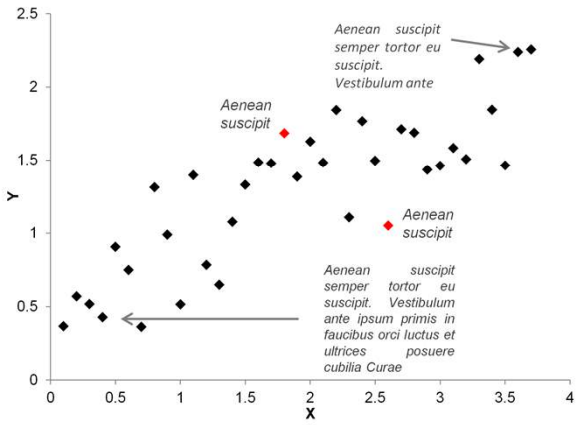
Fig XXX: A plot of y against x. Y represents... X represents... Note the ...Sed consequat dolor a purus molestie aliquet. Maecenas dignissim, urna nec condimentum egestas, lectus tellus interdum enim, vel viverra felis arcu eget odio. Nam placerat faucibus sem ut ornare. Etiam et nulla quam. Suspendisse dolor diam, feugiat sit amet erat sed, accumsan hendrerit erat. Aenean dictum vulputate eros, sed adipiscing erat pulvinar a. Fusce mattis convallis metus, quis te.

Data source:...

AQMeN
Applied Quantitative Methods Network

The Annotation Layer

Sed at dui at urna ullamcorper cursus



Data source:...

AQMeN
Applied Quantitative Methods Network

The rest of the workshop

- This afternoon
  - An introduction to R
  - R installation workshop
- Tomorrow Morning
  - A brief introduction to ggplot2
  - Extended practical on ggplot2
- Tomorrow Afternoon
  - A talk on visual demography
  - A practical: either on visual demography; or continuing to work on ggplot2
  - Summing up

Introduction to R

- Please go to the following url:
  - https://www.codeschool.com/courses/try-r
- This course does not require installation of R. Instead, it simulates R in a web browser. However, it produces a simulation of R that is much friendlier and more helpful than the real thing!
- The course has a number of levels. To complete each level a number of challenges have to be completed. Completing each challenge involves learning about and mastering something new about R.

Introduction to R continued…

- The added value I will be providing, compared to the course alone, is you! Developing a basic familiarity with R is essential for day two. Some of you already have this, some don't. However I hope everyone will gain from this exercise: the newcomers by learning something from the first time; and people with more experience by having to revise these skills, and having to explain to newcomers how and why R works the way it does.

- Every few minutes I will be asking for progress checks. I will ask people who are furthest ahead to help out the people who are furthest behind. When everyone in the class has finished a particular level, I will ask everyone to stop so we can discuss what we have just learned.

- The guiding principle of this session is **nobody left behind**.

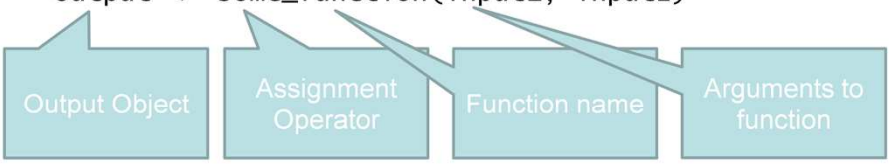
What have we learned?

- Level 1: Using R
  - Discussion here
- Level 2: Vectors
  - Discussion here
- Level 3: Matrices
  - Discussion here
- Level 4: Summary Statistics
  - Discussion here
- Level 5: Factors
  - Discussion here
- Level 6: Data Frames
  - Discussion here
- Level 7: Real-World Data
  - Discussion here

General ideas/Principles

- R is about functions and objects
- Most functions look a bit like this:

```
output <- some_function(input1, input2)
```



- Functions take objects, do things to them, and make other objects. Functions also have 'side effects'. Printing, either to the console, or to a graphics device, is classed as a 'side effect'. (So, 'side effects' are good, as long as they're understood.)

General Ideas/Principles

- There are a large number of R object types, but almost all R objects are vectors, with elements that can be individually accessed. For example, a single value is conceived of by R as a vector of length 1; a 2x2 matrix is conceived of as a vector of length 4 with some additional attributes.
- One of the most common R object types is the dataframe. Dataframes hold data in a rectangular format, with each row representing a different case, and each column representing a different variable. Each variable can be of a different class: numeric, logical, factor, and so on. Dataframes should be very familiar to quantitative social scientists who work with SPSS, Stata or Excel.

General Ideas/Principles continued…

- One of the most important classes of object is the list. A list is an object that can contain all other kinds of object, including other this. This means it can be used to manage data with a complex non-rectangular data. (For example, spatial data).
- Technically, a dataframe is 'non-ragged' list. This means it is thought of a list of length k, where k is the number of variables, and where each of the elements of the list is a vector of length n, the number of cases or observations. Each of the list elements can, however, be of a different type.

Packages

- R has grown so fast because it's open-source. Anyone could (in principle) create a new collection of functions for a particular purpose, and so save other people the effort of creating these functions themselves.
- These collections of functions are known as 'packages'. Thousands of packages are available on CRAN.
- To install a package using the command line, type install.packages("the_package_name"). A package only has to be installed once.
- Once a package has been installed on your machine, it has to be loaded onto the current R session. You can do this by typing either library("the_package_name") or require("the_package_name"). (There are subtle differences between the library() and require() functions. I would recommend using the require() function.)

Packages continued…

- Within RStudio, a list of packages that can be downloaded, along with descriptions, can be found in the 'packages' tab in the bottom right quadrant of the screen. Clicking on the checkbox next to the package name will install or load it for you.

- In fact, there are so many packages that there is a package for helping to manage packages, called 'task views'. A task view is a collection of packages that an expert in a particular subject area has curated and managed for others in that subject area to use. For example, there are specific task views for social sciences, finance, Bayesian inference, spatial statistics, and much more. By installing a particular task view, dozens of packages useful for your particular research field will be fetched and installed for you. For more information, please see http://cran.r-project.org/web/views/

Getting help

- It is possible to get limited and highly structured descriptions of functions in R packages that have been loaded by typing either help("function_name") or ?function_name (The ? Symbol is a shorthand for 'help').
- The problem with the information presented by the help files is you may need a lot of help to understand it. (They tend to be very technical, and not to present easy-to-understand examples).
- Another option is the apropos function, which searches for all functions containing the argument name.
- For more bespoke and specific queries, I would recommend searching through the R users mailing list, and through the website stackoverflow. If you cannot find a solution, I would recommend asking colleagues, or posting on stackoverflow, but only if you have read the guidelines and exhausted other possibilities.

Friendly websites

- Quick-R
  - http://www.statmethods.net/
  - Still my favourite website for R help
  - Examples are generally very clear
- Coursera, Data Science Course
  - https://www.coursera.org/specialization/jhudatascience/1
  - If I had the time and resources to, I'd go on this, and recommend you do too!

Friendly books

- Kabacoff, Robert (2011), **R in Action: Data Analysis and Graphics with R.** (2nd Edn.) *Manning Publications.*
  - The book associated with the Quick-R website.
- Field, Miles, Field (2012), **Discovering Statistics Using R**, *Sage Publications*
  - The R version of Andy Field's introduction to statistics book.
- Fox & Weisberg (2011), **An R Companion to Applied Regression** (2nd Edn.) *Sage Publications*
  - Associated with the R package 'car', which contains a number of useful functions, especially for data coding.
- Matloff, N (2011), **The Art of R Programming: A Tour of Statistical Software Design**. *No Starch Press*
  - Much easier to understand and more useful than it may first appear.

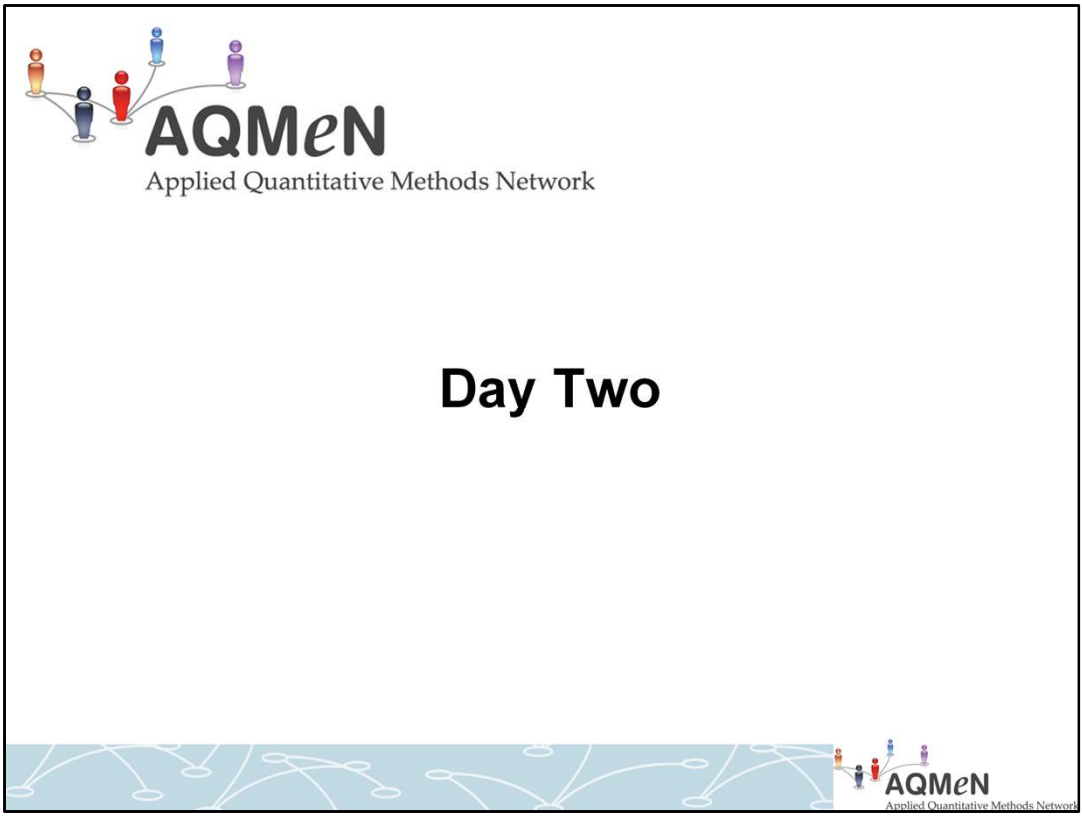
Other AQMEN Courses

- See AQMEN Website
  - Courses run by experts including Dr Colin Gillespie & Prof Lindsay Paterson, amongst others
  - Keep checking out the website for more training opportunities

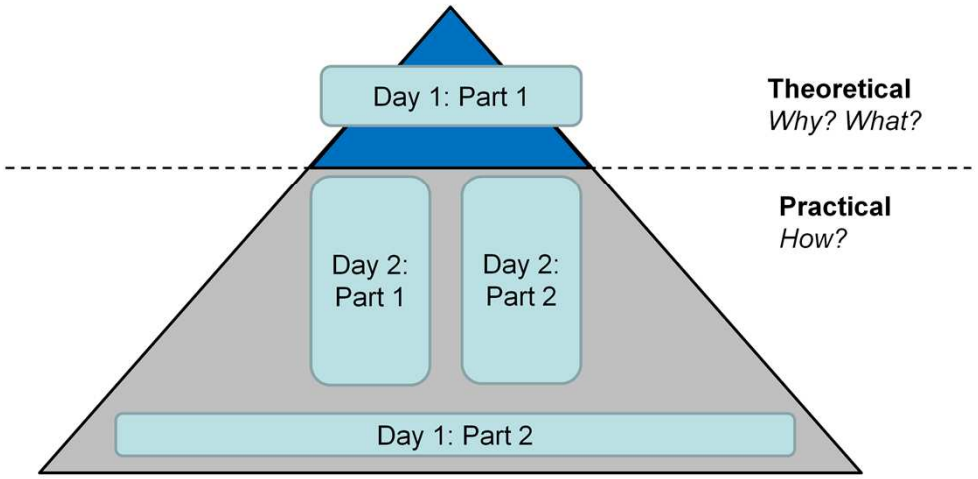
In summary…

- Today was a day of two halves
  - The first half of the day was all theoretical
  - The second half was all about the practicalities of using R
- Tomorrow, the theoretical foundations from this morning will be linked with the practical foundations from this afternoon.
- Next
  - an R/GitHub 'clinic'
  - And then… relax!

See you tomorrow!


AQMeN
Applied Quantitative Methods Network
Day Two
AQMeN
Applied Quantitative Methods Network

Guiding Principles and Structure of this workshop



AQMeN
Applied Quantitative Methods Network

Covered today

- Morning
  - An introduction to ggplot2
  - Extended practical on using ggplot2
- Afternoon
  - A talk on demographic visualisation
  - **A choice**: A) continue with the morning's practical; or B) do some practical exercises on demographic visualisation

But first…

- Any comments/ queries/suggestions about the material covered yesterday?
  - [Make notes here]

An introduction to ggplot2

- The gg in ggplot2 stands for the 'grammar of graphics'
- The basic principle of the grammar of graphics was covered yesterday, within the definition of data visualisation
- Important key concepts:
  - Layers
  - Mapping rules
  - Coordinate systems
  - Aesthetics
  - Geometrics

What is the grammar of graphics?

- A book by the statistician Leland Wilkinson
- Provides a language/framework for thinking about and describing data visualisations
- Wilkinson has worked mainly with SPSS, and with colleagues at SPSS developed the program nVizn, which applies the 'grammar of graphics' approach
  - http://www.cs.uic.edu/~wilkinson/Publications/ibm.pdf
  - http://www.cs.uic.edu/~wilkinson/nViZn/nvizn.html
- However, because of the increasing popularity of R, Wilkinson's grammar of graphics principles have had more influence mediated through Hadley Wickham, the creator of the ggplot2 package

Hadley Wickham & ggplot2

- One of the problems with R comes out of one of its advantages:
  - Open source means many packages;
  - But many packages means many contributors;
  - Many contributors means many different standards and approaches
  - Inconsistency, redundancy, reinvention

Hadley Wickham & ggplot2 continued…

- Hadley Wickham
  - Has developed a wide range of packages for R for helping to work with and explore data
  - These grew out of his PhD thesis.
  - http://www.cs.uic.edu/~wilkinson/nViZn/nvizn.html
  - Perhaps because of this packages by Wickham tend to have much more of a sense of cohesiveness than most packages. They each work well with each other and have complementary roles in the process of helping researchers gain insight from data.
  - ggplot2 is the main package Wickham has developed for the visual exploration of data. It is based on Wilkinson's grammar of graphics 'philosophy', but involves a number of different features and uses some slightly different terminology.

Using ggplot2

- Uses a 'Lego' philosophy…
- Data visualisations in ggplot2 are built up, piece by piece, using the '+' symbol.
  - (Technically, this is an example of 'operator overloading')

- At a minimum, the user needs to specify
  - The dataset
  - The **aes**thetics: mapping from variables to graphical elements)
  - The **geom**etrics: the associations between graphical elements

- Additionally, the user can specify:
  - Data transformations
  - Coordinate systems
  - Support layer elements
  - Facets

Description in Hadley's words

- A plot is made up of multiple layers
- A layer consists of:
  - data;
  - a set of mappings between variables and aesthetics;
  - A geometric object; and
  - A statistical transformation

Source: https://www.youtube.com/watch?v=RHu5vgBZ1yQ

Facets

- Facets are a very important and useful way of arranging data
- Effectively they specify how to arrange multiple figures in a tabular format
- Similar in principle to trellis/lattice plots (Cleveland) or small multiples (Tufte)

- To use:
  g2 <- g1 + facet_grid([arguments])
  g2 <- g1 + facet_wrap([arguments])

- The first argument defines the arrangement of the facets
  - See http://docs.ggplot2.org/0.9.3.1/facet_grid.html and the differences between:
    p + facet_grid(. ~ cyl)
    p + facet_grid(cyl ~ .)

Ggplot2: use within R

- Two main approaches
  - qplot (quickplot): this allows people to be less specific about elements of the graphic, by making (usually) intelligent assumptions about the graphical elements that work best given the data. It uses similar arguments, and *looks* similar, to the basic plot function used in R's basic graphics package.
  - ggplot: this is the main function used in ggplot2 to create a foundation for building your data visualisation on top of. At the minimum, the ggplot function just requires one object: the dataset to be used. All other instructions required can then be added to this first object separately.

Something else to note

- If you assign a composite of a ggplot object to another object, you won't print it on the graphical device.
- Printing is a 'side effect' of a function. This is caused in two ways:
  - Use the print function with the ggplot object as the argument, e.g. type
    print(g3)
  - Type the ggplot object without assigning it anywhere, e.g. type
    g3

The practical: Play with ggplot2

- The practical: *In groups of (ideally) three, choose a dataset, and use ggplot2 to explore it.*
- Discussion:
  - If ggplot2 is Lego, then this practical should be **play**, not work.
  - I won't tell you what to build, or how to build it
- However… some suggestions:
  - Use a dataset that you are actively interested in if possible.
  - SAVE THE CODE as you generate it
  - Be scientific in your play: look at changing only one thing at a time
  - Have designated roles within your team: work to your strengths
  - Consider drawing out the mapping rules on pen and paper first

ggplot2: The stairs or the escalator?

- **The Stairs**: If you want to use 'pure' ggplot2, do so.
- However, if you are new to both R, and to ggplot2, and you are not too comfortable with command line interfaces, this might be too many hurdles at once. Luckily there's GUI alternative that allows you to start to understand the core principles of ggplot2, without having to also work out how to specify it through code:
- **The Escalator**: http://rweb.stat.ucla.edu/ggplot2/
- If you're using the Escalator, please *open the code panel* so you'll know how to use the stairs later.

Getting data

- Hopefully, some data kindly contributed by fellow delegates should be available on the GitHub repository.
  - https://github.com/JonMinton/AQMEN_Data_Vis_Workshop
- Otherwise, data are available within many R packages, including ggplot2; and the datasets package
  - Typing data() to show all datasets available on your computer
  - Installing more packages will increase the list.
  - The package argument in data() allows you to specify only those datasets in a package
  - To move a specific dataset into current workspace, type the dataset name as the only argument in the data() function
  - If you're going to use the web based GUI, you will need to export the data from R proper, and import the data to the GUI. Just ask me, and I will help. (And if you've been helped, help others!)

Suggested phases

- Phase One:
  - Form a team; name the team; decide on roles; decide on approach; decide on data.
- Phase Two:
  - Start to understand the format and dimension of the data; decide on the mapping rules (variables and aesthetics); geometrics (lines, points, bars etc.) and so on; sketch out what this might look like.

Suggested phases continued…

- Phase Three:
  - Research how this mapping can be achieved in ggplot2 (n.b. not everything can be); search for 'recipes' and modify to taste; write up code with comments so you can understand and explain the development process.
- Phase Four:
  - Explore how to modify the cosmetic features of ggplot graphics, how to save in different formats; explore the effect of using different geoms on the same variables, or different variables on the same variables; incorporate ggplot2 images in a report or powerpoint presentation.

Phase One

- Team Names - [fill in here]

- Progress, challenges, achievements, ideas

Phase One

- Team Names - [fill in here]

- Progress, challenges, achievements, ideas

Phase Two

- Team Names - [fill in here]

- Progress, challenges, achievements, ideas

Phase Three

- Team Names - [fill in here]

- Progress, challenges, achievements, ideas

Phase Four

- Team Names - [fill in here]

- Progress, challenges, achievements, ideas

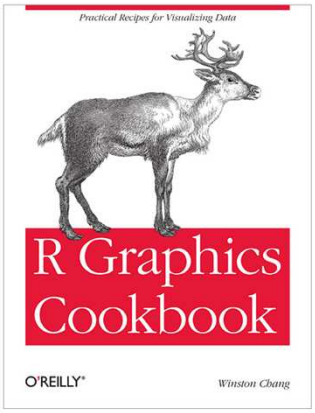
Summary

- What have we learned?

Next steps

- Practice, Practice, Practice
- A very strong (but occasionally unfriendly) online community

- This book:
  - Chang, W (2012) **R graphics Cookbook**, O'Reilly
  - Recommended by Wickham in preference to his own book



AQMeN
Applied Quantitative Methods Network

After lunch

- A talk on demographic visualisation
- A choice:
  - Either continue practicing ggplot2
  - Or start to learn how I produced some of the visualisations in the talk. [More niche?]

Demographic Scars & Demographic Futures

Jonathan Minton
University of Glasgow

Presentation Structure

1. Origins of the research
2. Co-plots
3. Dead Parrots & Demography
4. Thoughts about statistics and data visualisation
5. Contour plots
6. Practical considerations
7. A live demonstration (If there's time, and I'm feeling lucky…)
8. The future

1) Origins of the research

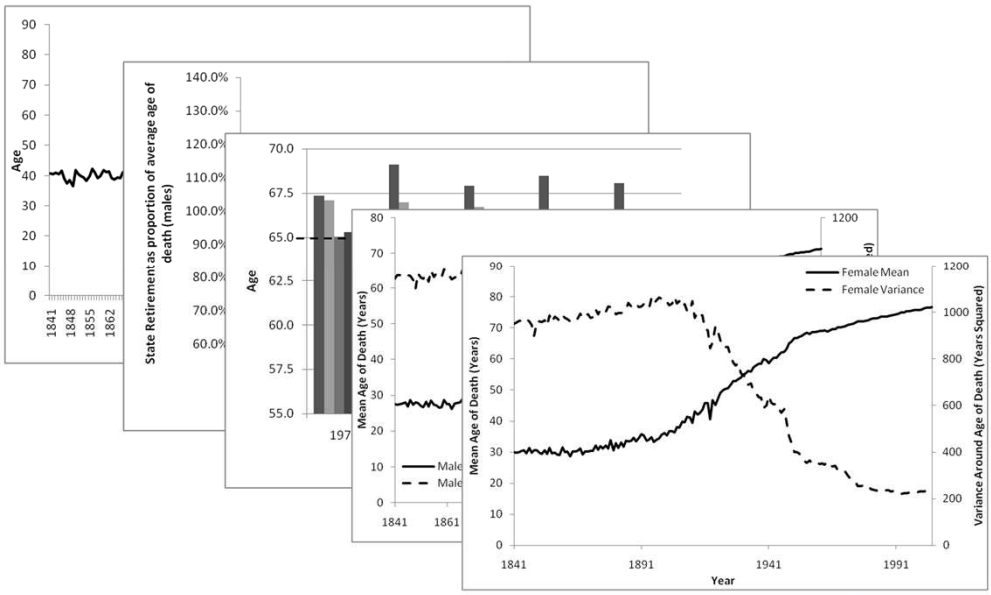
"The visualisations developed from a PhD in Sociology & Human Geography at the University of York, and a friendly argument with his former PhD supervisor, Danny Dorling, about the interpretation of two lines on a graph."

http://blog.oup.com/2013/09/demographic-landscape-good-news/

1) Origins of the research
Reasons for data visualisations of demographic data

- Substantive
  - To get a better sense of how pension liabilities have 'crept up' on richer nations over time due to increased longevity
  - To see how life has become less 'risky' over time
- Methodological
  - To avoid over-reliance on summary statistics
  - To perform rich, nuanced exploratory data analysis of a lot of data

1) Origins of the research

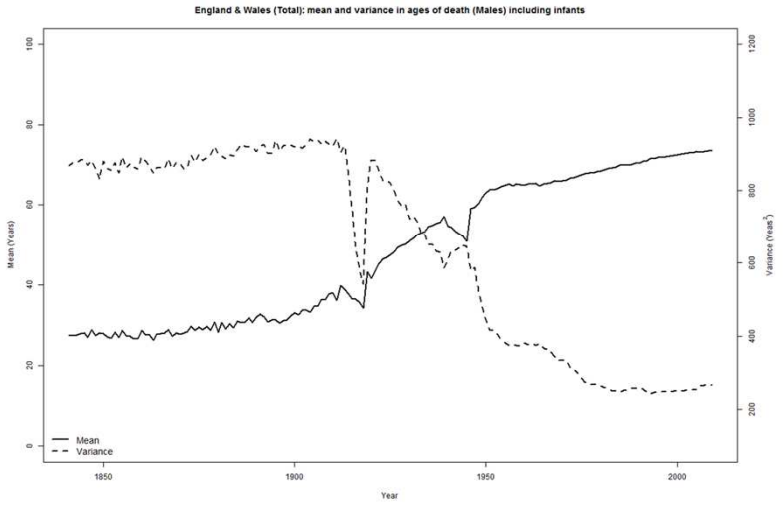


AQMeN
Applied Quantitative Methods Network

2) Coplots

AQMeN
Applied Quantitative Methods Network

2) Coplots

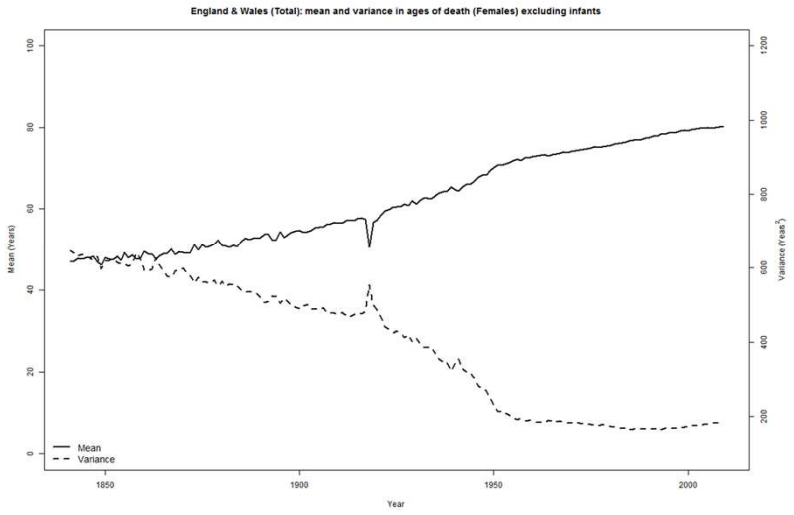


AQMeN
Applied Quantitative Methods Network

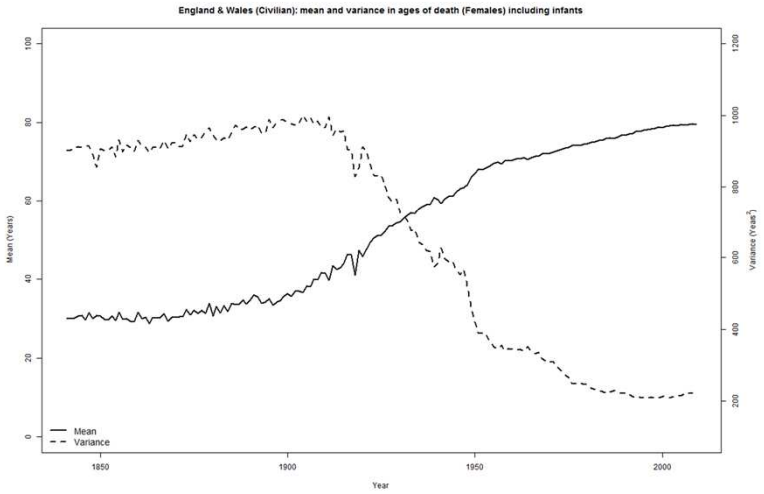
The argument

- Variance is not life riskiness, it's infant mortality.
- Variance may reduce entirely due to reduced infant mortality, without any reduction in mortality risk at other ages.

2) Conditional co-plots



2) Conditional Coplots



2) Further thoughts

- The co-plots are dependent on the choice of threshold.
- The dangers:
  - Over-reduction
  - Devil-'s-in-the-detail Dependence
- Time for a re-think.

3) Dead Parrots

3) Demography: The three mortality effects

- Age effects
  - p(death) = f1(age)
- Period effects
  - p(death) = f2(year)
- Cohort effects
  - p(death) = f3(age*year)
- But it's not this simple: each effect is non-static over time

3) Demography: Areas of substantive importance

- Micro-problems
  - Parental investment
  - Investment in retirement
- Macro-problems
  - Labour market & dependency ratios
  - Pensions
  - Healthcare

4) Statistics & Data Visualisation

- My perspective
  - The main benefit of statistical inference is (honest) data reduction
  - Data reduction is needed because people can't cope with masses of information in making decisions
  - BUT data do not need to be reduced as much if it is presented effectively
- Demographic Data Visualisation
  - 'Thousands of Data at a Glance' (Vaupel et al 1987)

4) The Lexis Surface

- Lexis Surfaces
  - x axis: age
  - y axis: year
  - z axis: value associated with {age, year} configuration
- Z could be
  - Population count
  - Death count
  - Morbidity count/rate (But do you trust the morbidity data?)
  - 'Economic' variable such as median income (But do you trust the economic data?)
- In my case:
  - Crude death rate : Death Count/Population Count

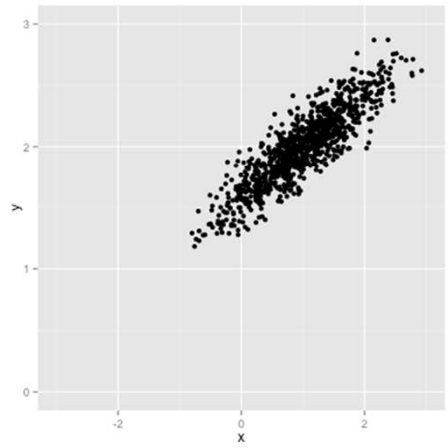
4) The Data Source

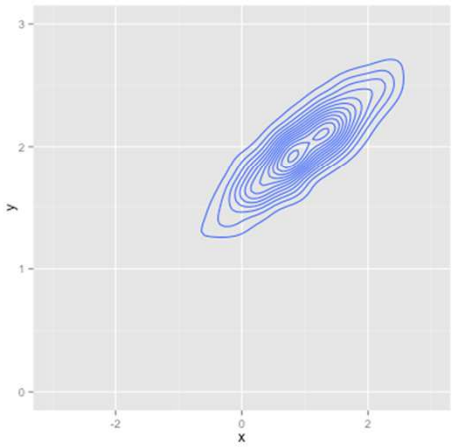
- Human Mortality Database

- 48 datasets from 37 countries, some going back to the 18th century

AQMeN
Applied Quantitative Methods Network

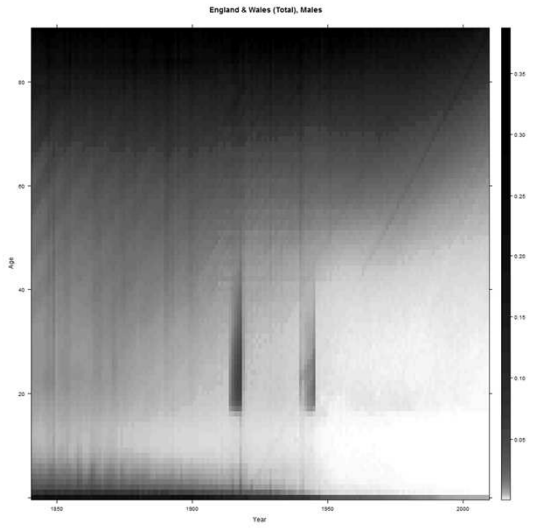
4) The Method: Contour plots



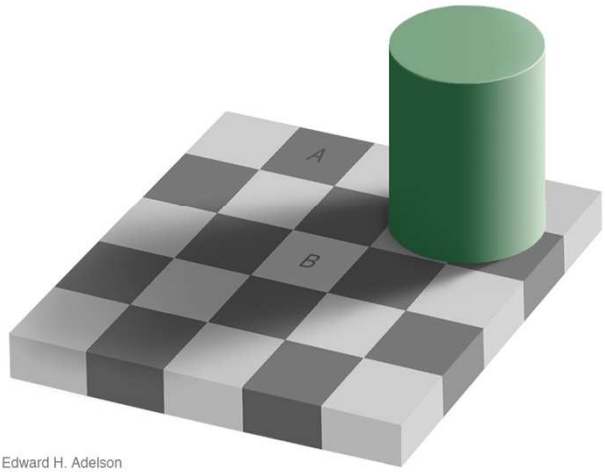


AQMeN
Applied Quantitative Methods Network

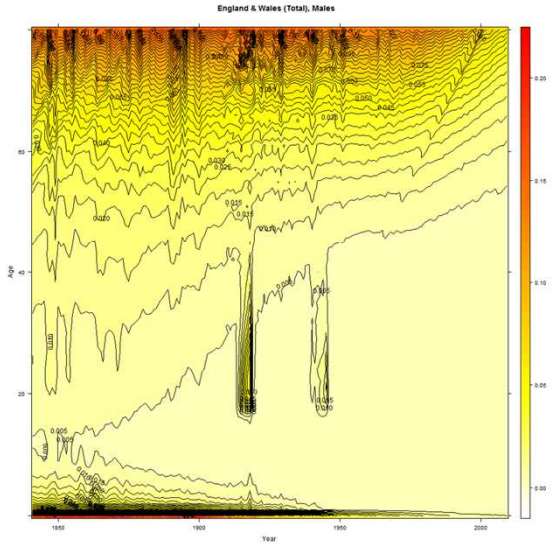
4) Question: Why not a bitmap?



4) Answer: Because of this.

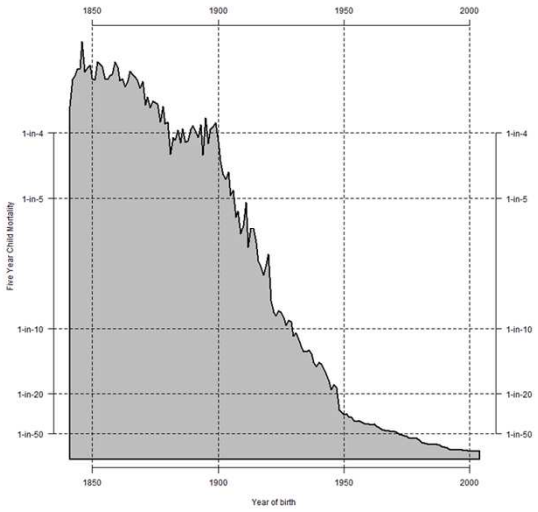


Edward H. Adelson



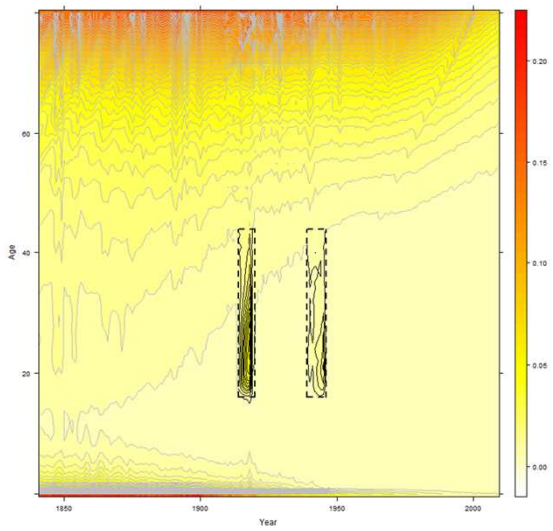
AQMeN
Applied Quantitative Methods Network

Age



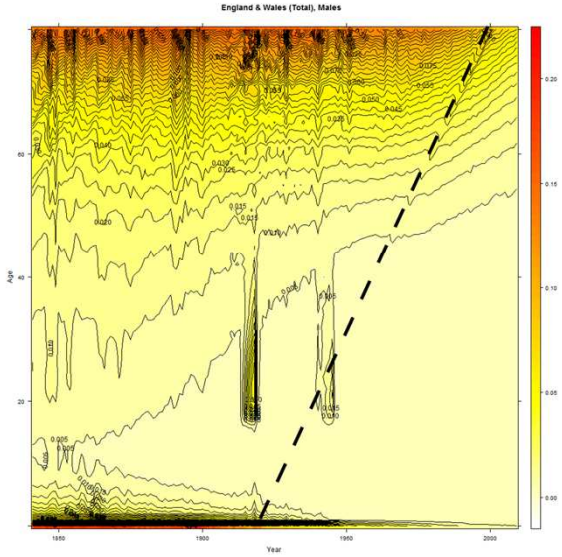
AQMeN
Applied Quantitative Methods Network

Period



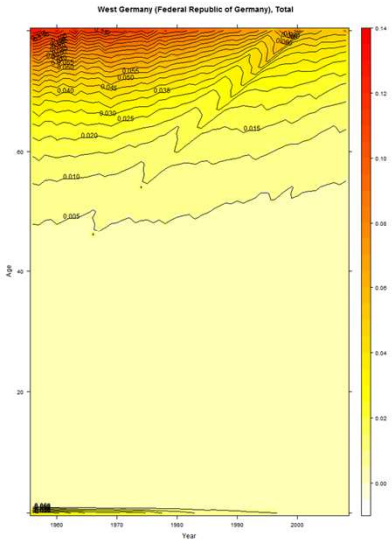
AQMeN
Applied Quantitative Methods Network

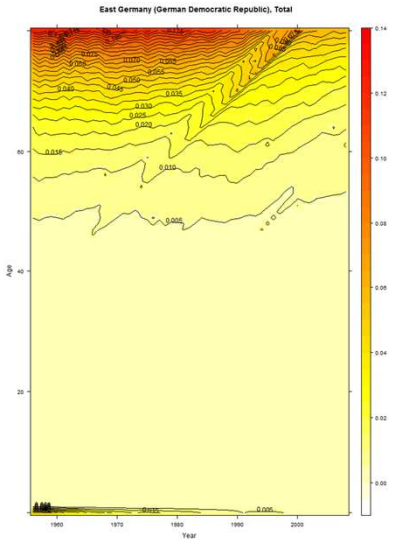
Cohort



AQMeN
Applied Quantitative Methods Network

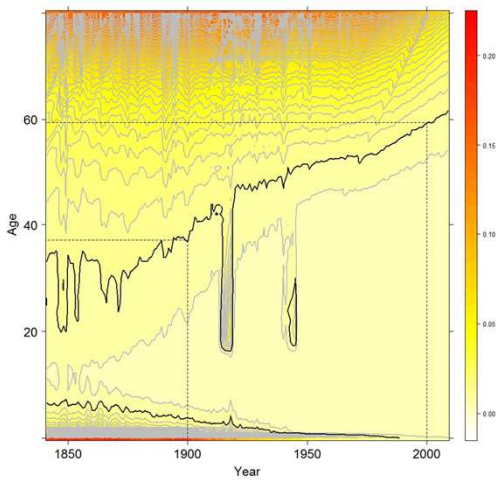
Natural Experiments





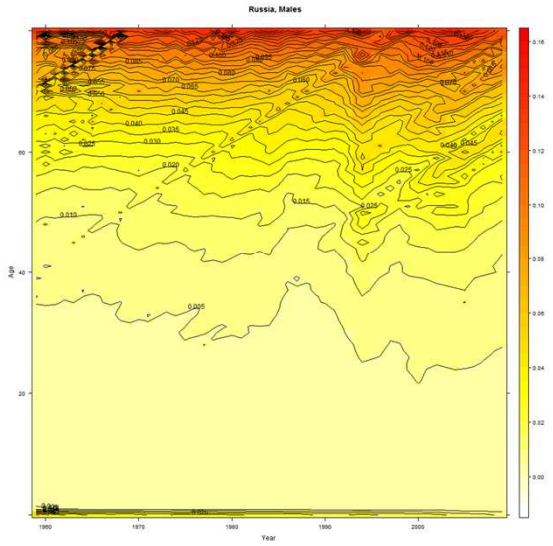
AQMeN
Applied Quantitative Methods Network

Trends: Following a Contour



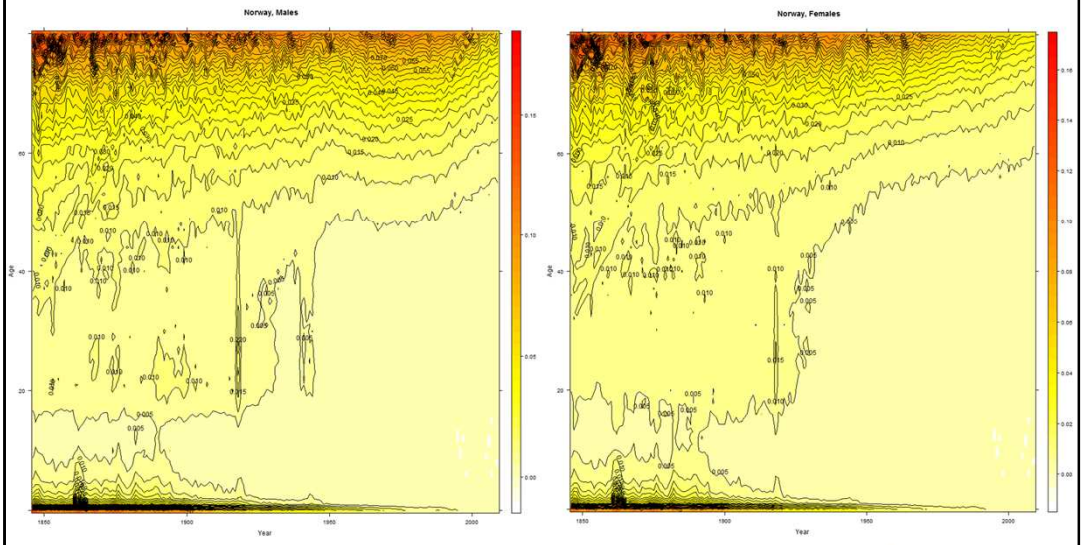
AQMeN
Applied Quantitative Methods Network

Male Mortality & The Collapse of Communism



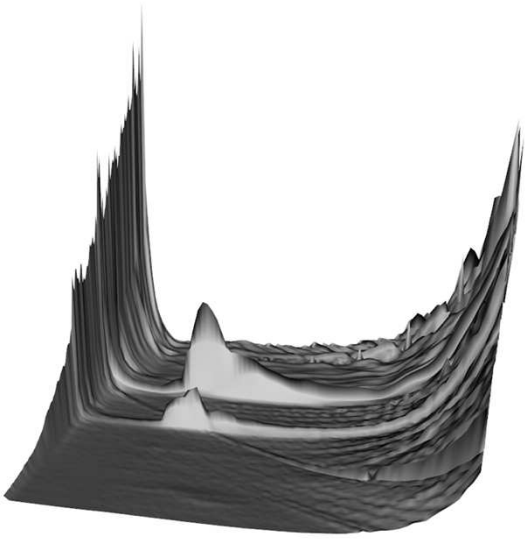
AQMeN
Applied Quantitative Methods Network

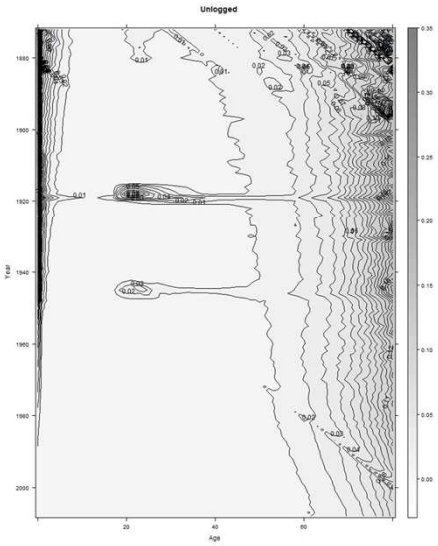
The Long Arc of History



AQMeN
Applied Quantitative Methods Network

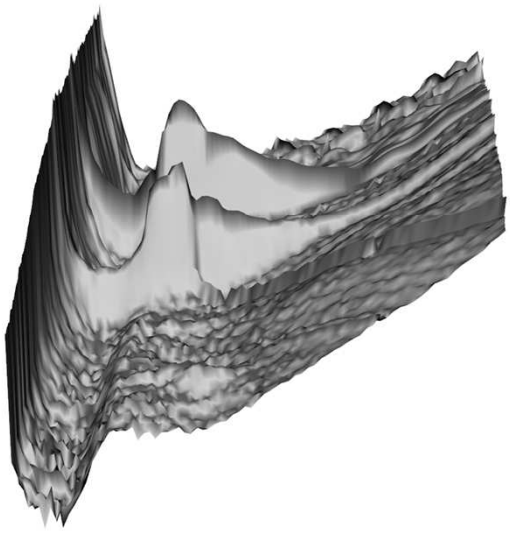
Italian Bathtubs: Unlogged

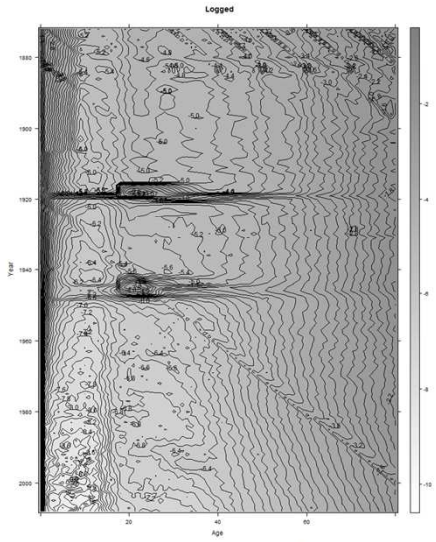




AQMeN
Applied Quantitative Methods Network

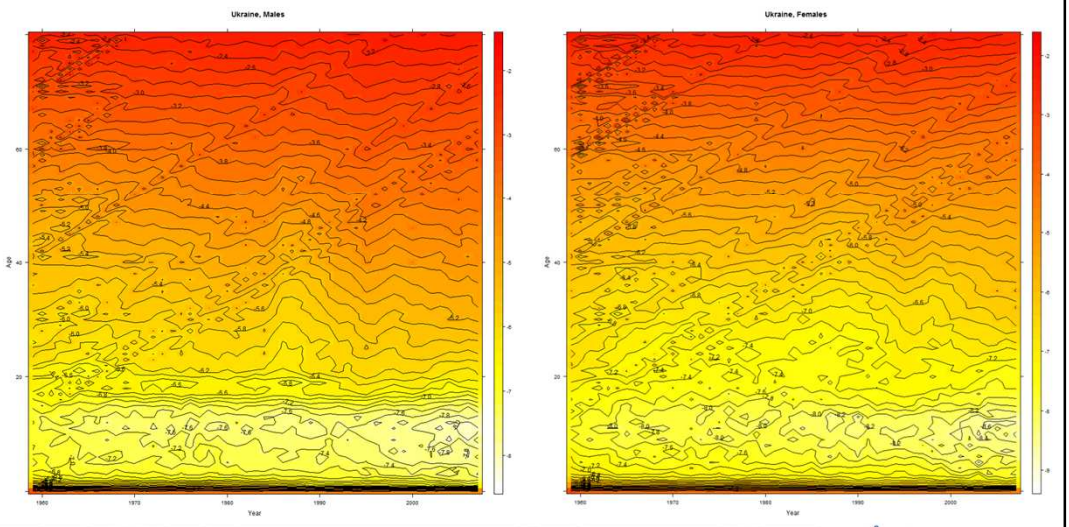
Italian Bathtubs: Logged



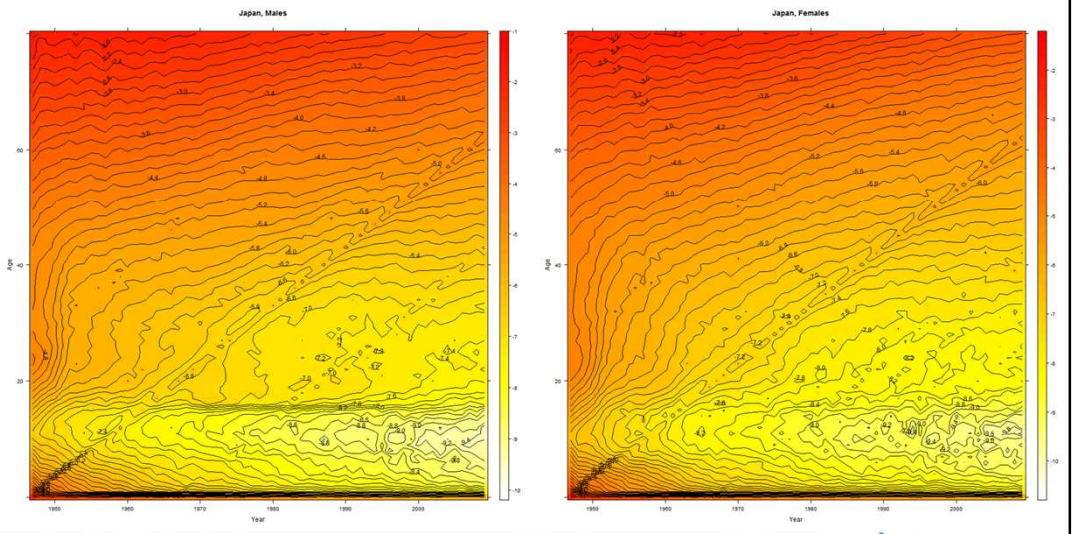


AQMeN
Applied Quantitative Methods Network

More Recent Data: Switching to the Logarithmic Lens

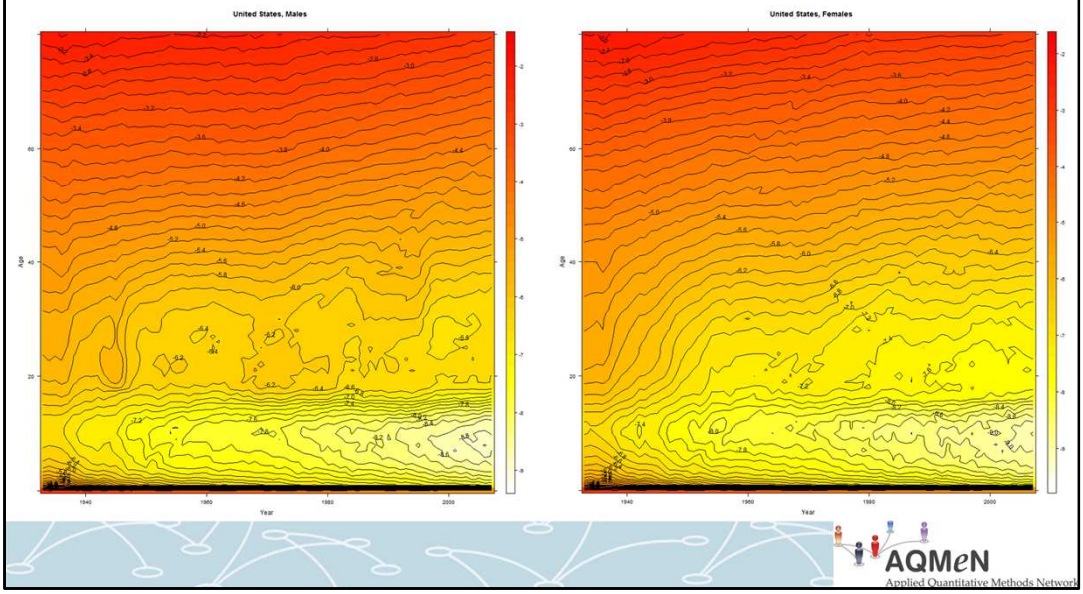


More Recent Data: Switching to the Logarithmic Lens



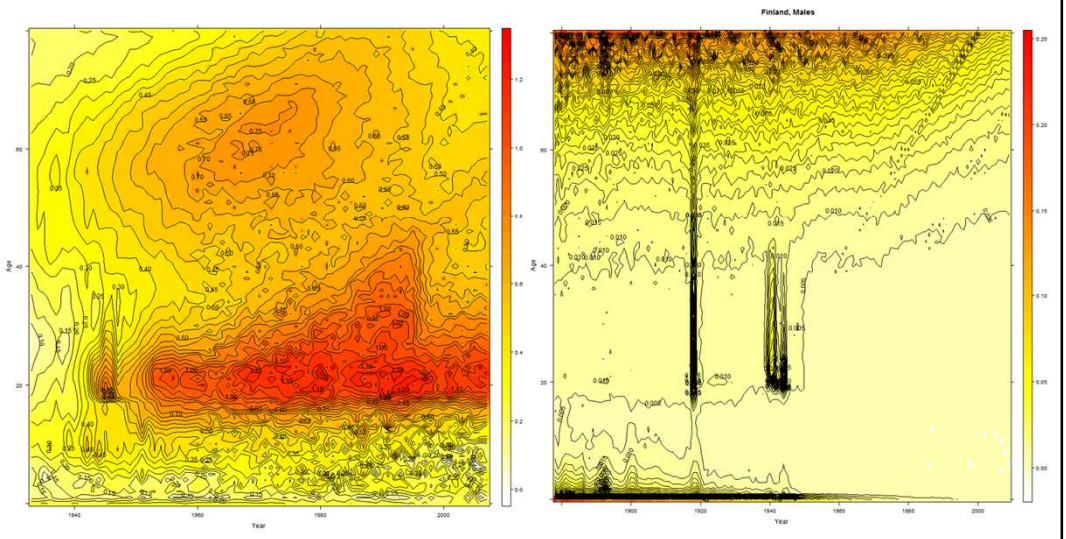
AQMeN
Applied Quantitative Methods Network

USA: Male & Female Log Mortality



AQMeN
Applied Quantitative Methods Network

Unexpected Patterns



AQMeN
Applied Quantitative Methods Network

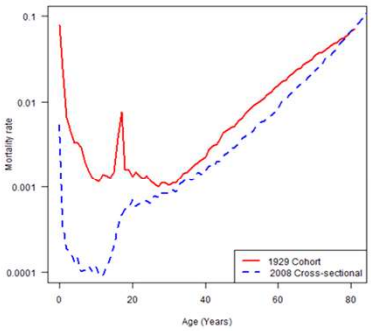
Using shaded contour plots to estimate bathtub curves

These figures show how 'bathtub curves', mortality rates as a function of age, can be read from contour maps.

The **red** and **blue** lines in figure a indicate two planes which cut the landscape at right angles to both age and year. The corresponding **red** and **blue** lines in figure b effectively shows what the cross-sections formed by 'cutting' these surfaces along the planes would look like, i.e. they are tomographs, as commonly used in medical imaging.

The bathtub curve public health researchers really want to be able to estimate is that associated with the **thin, dashed purple line** in figure a. The corresponding tomograph produced would be the bathtub curve which will be experienced by a new cohort, in this those case born in 2008.

Although we do not have the data to produce the bathtub, we can estimate it by extrapolating the contours produced from the data we do have. This is likely to produce better estimates for the 2008 cohort than the alternatives shown here.

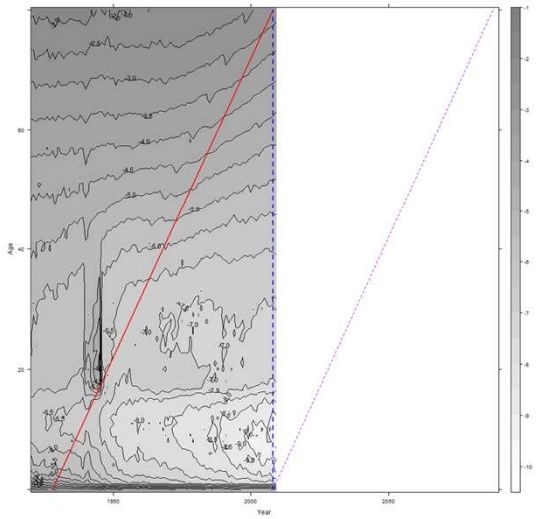


b) Mortality as a function of age for the two colour lines added to figure a

AQMeN
Applied Quantitative Methods Network

Using shaded contour plots to estimate bathtub curves 2

a) Contour plot of mortality surface for Males, England & Wales with two coloured lines indicating the 1929 cohort (red), and a synthetic cohort (cross section) based on age-related mortality in 2008 (dashed blue line).



AQMeN
Applied Quantitative Methods Network

Current Publications

- Minton, Vanderbloemen, Dorling (2013)
  - 'Visualising Europe's Demographic Scars with Coplots and Contour Plots', *International Journal of Epidemiology*
  - http://www.ncbi.nlm.nih.gov/pubmed/24062300
- Minton (2014a)
  - 'Logs, Lifelines and Lie Factors', *Environment & Planning A*
  - http://www.envplan.com/abstract.cgi?id=a130208g
- Minton (2014b)
  - 'Real Geographies and Virtual Landscapes', *Spatial & Spatiotemporal Epidemiology*
  - http://www.sciencedirect.com/science/article/pii/S1877584514000173
- Minton (2014c)
  - 'If Europe were a country…", *Environment & Planning A*
  - Accepted
- Minton (2014d???)
  - 'Hunting Demographic Ghosts', *Environment & Planning A*
  - Under review

Further Research Avenues

- Visualisation Approach
  - Graphics engines
  - 3D printing
- Modelling
  - Formalising estimation of counterfactuals
  - Comparison between nations
- Sources of Data
  - Population Statistics
  - World Health Organisation
  - Human Fertility Database

But… hundreds of visualisations are already available.

Why not develop and explore your own?

RunMyCode & GitHub

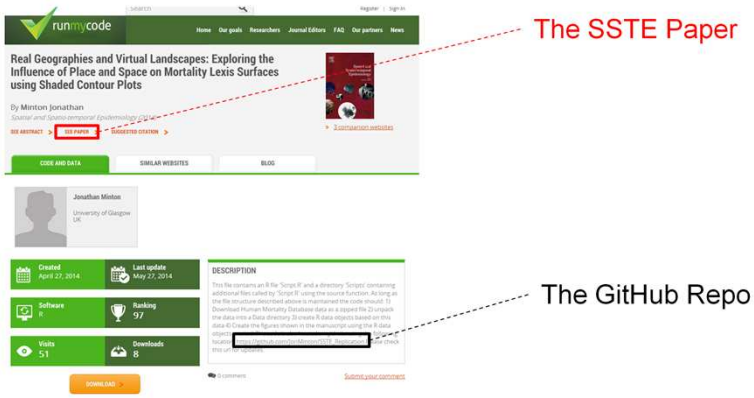
- https://www.sheffield.ac.uk/scharr/sections/heds/ije

- http://www.runmycode.org/companion/view/631

AQMeN
Applied Quantitative Methods Network

RunMyCode & GitHub

- https://www.sheffield.ac.uk/scharr/sections/heds/ije



- http://www.runmycode.org/companion/view/631

AQMeN
Applied Quantitative Methods Network

The code itself

1. Fetch HMD Zipfiles
2. Produce death rate data for each country
3. Automate producing and labelling of coplots and Lexis plots

Structure of Code

```
rm(list=ls())

require(RCurl)
require(repmis)
require(httr)
require(digest)
require(devtools)
require(lattice)
require(latticeExtra)
require(downloader)
require(xlsx)
require(rgl)
require(tcltk)

Run_3D_Vis = TRUE
OlderData=TRUE
Replicate_Figures = FALSE

source("Scripts/Functions.R")
source("Scripts/Manage_Prerequisites.R")

if (Replicate_Figures){
  source("Scripts/Make_Figures.R")
}
```

Clear workspace

Load packages
(Will need to use install.packages first time)

Flags
(Change to TRUE or FALSE)

Unconditional scripts

Unconditional scripts

Working with R List Objects

- Lists: Magical containers

```
> names(DeathRates)
 [1] "AUS"      "AUT"      "BEL"      "BGR"      "BLR"      "CAN"      "CHE"      "CHL"      "CZE"      "DEUTE"
[11] "DEUTFRG"  "DEUTGDR"  "DEUTNP"   "DEUTW"    "DNK"      "ESP"      "EST"      "FIN"      "FRACNP"   "FRATNP"
[21] "GBR_NIR"  "GBR_NP"   "GBR_SCO"  "GBRCENW"  "GBRTENW"  "HUN"      "IRL"      "ISL"      "ISR"      "ITA"
[31] "JPN"      "LTU"      "LUX"      "LVA"      "NLD"      "NOR"      "NZL_MA"   "NZL_NM"   "NZL_NP"   "POL"
[41] "PRT"      "RUS"      "SVK"      "SVN"      "SWE"      "TWN"      "UKR"      "USA"
> head(DeathRates[["AUT"]])
  Year Age      Female        Male       Total
1 1947   0 0.088839243 0.110771668 0.100119146
2 1947   1 0.006173176 0.008010499 0.007106740
3 1947   2 0.004161685 0.004608881 0.004388334
4 1947   3 0.003300049 0.003324345 0.003312387
5 1947   4 0.002267236 0.002600526 0.002436772
6 1947   5 0.001669668 0.001873756 0.001773553
> head(DeathRates[["GBRTENW"]])
  Year Age     Female       Male      Total
1 1841   0 0.13348342 0.16673412 0.15022053
2 1841   1 0.06091973 0.06495909 0.06292776
3 1841   2 0.03681716 0.03741289 0.03711287
4 1841   3 0.02514631 0.02629801 0.02571779
5 1841   4 0.01827503 0.01881921 0.01854612
6 1841   5 0.01373185 0.01396886 0.01385050
```

Producing the contour plots

- Lattice package; contourplot function

```
this.ds <- DeathRates[["NOR"]]

this.ds <- subset(this.ds, Age < 81)

tiff("Figures/Figure07.tiff", height=1000, width=1000)

print(contourplot(
  Male ~ Year * Age,
  data=this.ds,
  region=T,
  col.regions=rev(heat.colors(200)),
  cuts=50,
  main="Norway, Males")
)
dev.off()
```

Other possible formulae

Male ~ Year * Age
Female ~ Year * Age
Total ~ Year * Age

Female ~ Age * Year

log(Male) ~ Year * Age

I(log(Male) – log(Female)) ~ Year * Age

Structure of Code: An Experimental Addition

```
rm(list=ls())

require(RCurl)
require(repmis)
require(httr)
require(digest)
require(devtools)
require(lattice)
require(latticeExtra)
require(downloader)
require(xlsx)
require(rgl)
require(tcltk)

Run_3D_Vis = TRUE
OlderData=TRUE
Replicate_Figures = FALSE

source("Scripts/Functions.R")
source("Scripts/Manage_Prerequisites.R")

if (Replicate_Figures){
  source("Scripts/Make_Figures.R")
}

if (Run_3D_Vis){
  tmp <- Make_3D_Plot.UI(DeathRates)
}
```

Clear workspace

Load packages
(Will need to use install.packages first time)

Flags
(Change to TRUE or FALSE)

Unconditional scripts

Unconditional scripts

WARNING: Experimental

AQMeN
Applied Quantitative Methods Network

Final Practical

- **Option A**
  - Go to the RunMyCode link provided earlier, and run my code; explore the code and how the data and figures are generated; create your own variations of these image
- **Option B**
  - Continue working with and exploring ggplot2; either by starting afresh with a new dataset and project, or by further developing and refining the morning's project
- *I won't be offended if most people go for Option B. This is your decision about how best to use your time.*

Notes on the afternoon's practical

- [Make notes here]

Summing up

- A workshop of two days
- Two days of two halves
- Yesterday: AM Notes-
- Yesterday: PM Notes-
- Today: AM Notes-
- Today: PM Notes-

Summing up continued…

- Ideas for next steps
  - Graphical theory
  - R
  - Ggplot2
  - Lattice
  - Collaborative working: any partnerships formed today?
- Any final suggestions/Ideas? Notes-

Finally…

- Thank you for attending! I hope you found the workshop both informative and enjoyable.

- Feel free to contact me:
  - Jonathan.minton@glasgow.ac.uk