Intelligent data analysis to aid decision making in a commercial environment

Gordon Blunt

Gordon Blunt Analytics Ltd

Statistics: For Businesses, About Businesses
19th February 2013

Outline

1. Background
2. The ‘data environment’
    - Computers
    - Data
    - Data quality
    - Domain knowledge
3. Visualising data
    - Exploratory data analysis
    - Presenting the case
4. Case study
5. Final thoughts
6. References

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

Background

Experience

- Clients who may have little specialist numerate training
- Often with marketing, business or finance degrees
- Typically medium sized companies
- Departments without statistical resource

Background

Experience

- Clients who may have little specialist numerate training
- Often with marketing, business or finance degrees
- Typically medium sized companies
- Departments without statistical resource

I won't be talking about (for example)

- The likes of Google or Microsoft
- Academics in the mathematical sciences

Background

Experience

- Clients who may have little specialist numerate training
- Often with marketing, business or finance degrees
- Typically medium sized companies
- Departments without statistical resource

I won't be talking about (for example)

- The likes of Google or Microsoft
- Academics in the mathematical sciences

We all have clients - might be internal or external

- Our peers
- The Government
- External groups we want to influence

Objective

My main aim is to help everyone take better decisions - based on intelligent use of data - the focus of this talk

Objective

My main aim is to help everyone take better decisions - based on intelligent use of data - the focus of this talk

In other words, to promote the view . . .

‘*It is easy to lie with statistics, but even easier to lie without them*’

Attributed to Frederick Mosteller [Murray 2005]

Not this one . . .

‘*There are three kinds of lies: lies, damned lies, and statistics*’

Benjamin Disraeli (or Mark Twain?)

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

Computers

Computational Statistics, James Gentle

‘*While I think that the PC* sui generis *is the Big Thing, the overall advance in computational power is also important*’ [Gentle 2009]

In a business environment, the former has arguably had more of an impact on the type and frequency of analysis undertaken

Computers

Computational Statistics, James Gentle

‘*While I think that the PC* sui generis *is the Big Thing, the overall advance in computational power is also important*’ [Gentle 2009]

In a business environment, the former has arguably had more of an impact on the type and frequency of analysis undertaken

‘Business desktop’

- Microsoft® dominates, particularly Excel® and PowerPoint®
- ‘*Lets not kid ourselves: the most widely used piece of software for statistics is Excel*’ [Ripley 2002]

Computers

Computational Statistics, James Gentle

'*While I think that the PC* sui generis *is the Big Thing, the overall advance in computational power is also important*' [Gentle 2009]

In a business environment, the former has arguably had more of an impact on the type and frequency of analysis undertaken

'Business desktop'

- Microsoft® dominates, particularly Excel® and PowerPoint®
- '*Lets not kid ourselves: the most widely used piece of software for statistics is Excel*' [Ripley 2002]

Our responsibility as analysts

- Offer help and guidance to those who only use spreadsheets
- Or have little training in statistical graphics

Outline

1. Background

2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge

3. Visualising data
   - Exploratory data analysis
   - Presenting the case

4. Case study

5. Final thoughts

6. References

Nature of the data

Large (ish) data sets

- Tens of millions of cases, thousands of variables
- Problems using common statistical techniques
- Often will not fit ‘standard’ distributions
- Any statistical test likely to prove significant

Nature of the data

Large (ish) data sets

- Tens of millions of cases, thousands of variables
- Problems using common statistical techniques
- Often will not fit ‘standard’ distributions
- Any statistical test likely to prove significant

Scientific data sets can be much larger

- ‘*A year of collisions at a single LHC experiment generates close to 1 million petabytes of raw data*’ [symmetry 2012]
- 1.1×10^{15} gene sequences on the NCBI database - 28,000 times more than 5 years ago [NCBI 2013]

Nature of the data

Opportunistic data?

- Data may be a by-product of operational processes
- Therefore not a properly constructed sample
- Probably contain missing or incorrect fields
- May be difficult to access certain fields or tables
- Fields may have different names over time

It may be difficult to reconcile . . .

- Fields from different tables
- Records from ‘legacy systems’ with more recent data
- Data stored for different purposes
- For example, invoicing vs marketing

Using the data

A lot of data is not necessarily a good thing

‘*In some ways I think that scientists have misled themselves into thinking that if you collect enormous amount of data you are bound to get the right answer. You are not bound to get the right answer unless you are enormously smart.*’

[Efron 2010]

Using the data

A lot of data is not necessarily a good thing

‘*In some ways I think that scientists have misled themselves into thinking that if you collect enormous amount of data you are bound to get the right answer. You are not bound to get the right answer unless you are enormously smart.*’

[Efron 2010]

Practical consequences are important

‘*Politicians must often make decisions with imperfect knowledge, and sometimes those decisions don’t work or have unintended negative consequences*’

[Nutt 2012]

This is relevant to the example later

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

Data quality

Some problems from recent years

- Overdrawn savings accounts - should be impossible
- Consumer offers being used before they were distributed
- Postcodes missing from almost 15% of records
- ‘Unique’ codes being re-used
  - Customers
  - Products
- A mortgage given to a 91 year old
- A personal loan given to a 5 year old
- 12 digit barcodes with last 6 digits zero
- Missing values replaced by plausible, but incorrect, values
- Mortgage balances of £1, £5, £10 . . .

Data cleaning

Data cleaning [Maletic and Marcus 2010]

- Define and determine error types
- Search and identify error instances
- Correct the uncovered errors

Data cleaning

Data cleaning [Maletic and Marcus 2010]

- Define and determine error types
- Search and identify error instances
- Correct the uncovered errors

Data cleaning is harder in practice than in theory

- Cleaning the data can take 80% - 90% of a project’s time
- Automatic fault removal may remove real features
- Distinguishing ‘real’ from systemic patterns may not be trivial
- Need to work closely with the client
- Or the domain expert - may not be the same person

Distrust a data set that has no errors or missing values - someone has probably done something with it

Outline

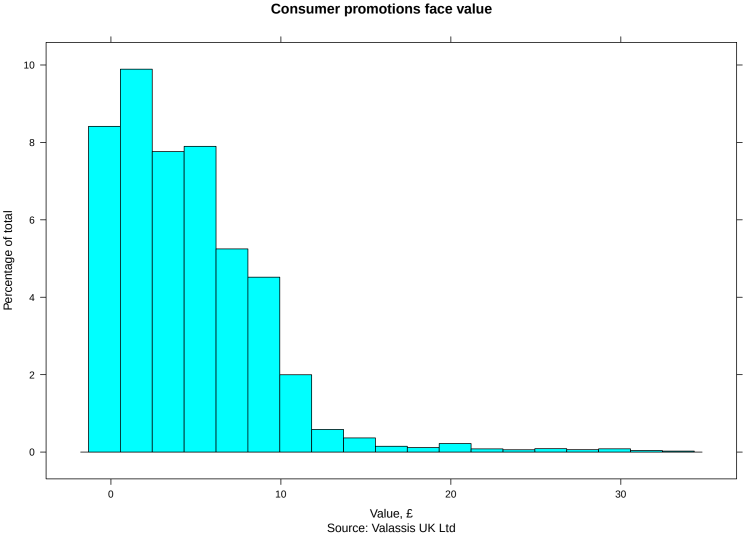
1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

Domain knowledge is essential

Commercial data sets

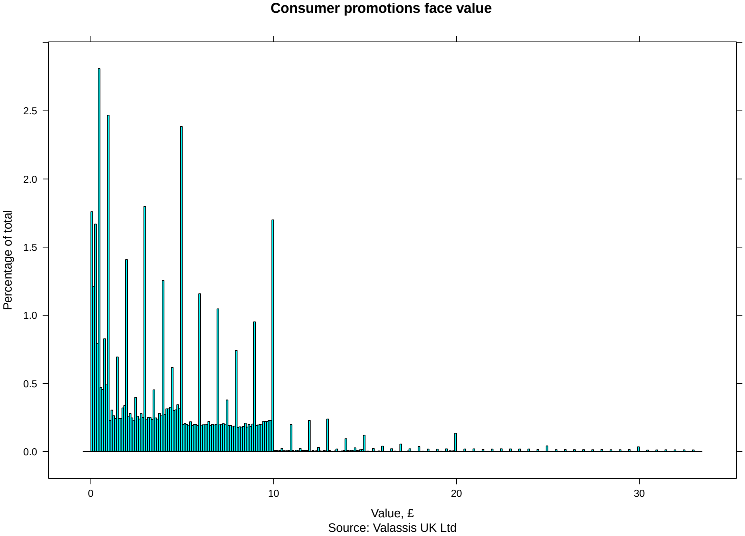
- We may see structures like the following one
- The histogram shows a sample of consumer promotions
- Uses a well known method for calculating the number of bins

Domain knowledge is essential (2)



This histogram uses Sturges' [Sturges 1926] method for calculating the number of breaks

Domain knowledge is essential (3)



Source: Valassis UK Ltd

Scott’s [Scott 1979, Scott 1992], F-D [Freedman and Diaconis 1981] work better, but above is better still

Domain knowledge is essential

Using the client's knowledge

- After discussions with the client, drew the second histogram
- As is clearly seen, there is much more structure apparent
- There appears to be a mixture of distributions, composed of
  - Random values
  - More clearly defined values at whole £ amounts

Domain knowledge is essential

Using the client’s knowledge

- After discussions with the client, drew the second histogram
- As is clearly seen, there is much more structure apparent
- There appears to be a mixture of distributions, composed of
  - Random values
  - More clearly defined values at whole £ amounts
- Some of the spikes were deterministic, some were not
- Some of the other data were deterministic, some were not

Domain knowledge

- Essential to talk to the client or other domain expert
- Throughout the project
- Such discussions will help our understanding

Deterministic or stochastic?

Some problems

- It might not be easy to work out which observations are deterministic and which are not
- The conclusions we draw will be wrong if we decide incorrectly

Deterministic or stochastic?

Some problems

- It might not be easy to work out which observations are deterministic and which are not
- The conclusions we draw will be wrong if we decide incorrectly

By contrast, a stochastic data set

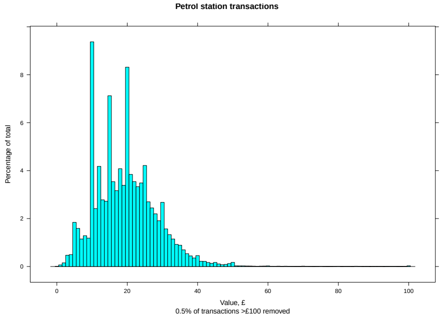
The next two charts show two views of a data set that has some unusual patterns, but which are entirely stochastic

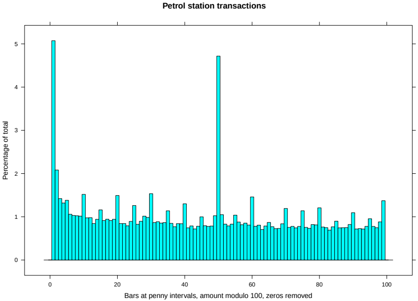
In this case, the patterns are entirely a result of people's behaviour

The data set is a few years old, but I'm pretty certain people still exhibit this type of behaviour

Consumer behaviour

Petrol station transactions on credit cards





Spikes at multiples of £5
and of £6 too . . .

Pence values of transactions
zeros removed

Data first reported in, with more detail, *Prospecting for gems in credit card data* [Hand and Blunt 2001]

Outline

1. Background
2. The ‘data environment’
    - Computers
    - Data
    - Data quality
    - Domain knowledge
3. Visualising data
    - Exploratory data analysis
    - Presenting the case
4. Case study
5. Final thoughts
6. References

The most important part of discussing data with clients

Plot the data

‘*Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency.*’

[Tufte 2001]

‘*Graphics* reveal *data. Indeed graphics can be more precise and revealing than conventional statistical computation.*’

[Tufte 2001]

‘*Visualization is a necessary part of data analysis. Tools matter.*’

[Cleveland 1993]

The most important part of discussing data with clients

Plot the data

‘*Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency.*’

[Tufte 2001]

‘*Graphics* reveal *data. Indeed graphics can be more precise and revealing than conventional statistical computation.*’

[Tufte 2001]

‘*Visualization is a necessary part of data analysis. Tools matter.*’

[Cleveland 1993]

Client facing statistical graphics

- Use the statistical graphic that best reveals structure
- Don’t be restricted to charts that are commonly used in spreadsheets and presentation software

Don’t be afraid of using complex graphics

If the graphic is good our job will be easier

- People can understand ‘non standard’ statistical graphics
- Providing we explain them clearly enough

Don’t be afraid of using complex graphics

If the graphic is good our job will be easier

- People can understand ‘non standard’ statistical graphics
- Providing we explain them clearly enough

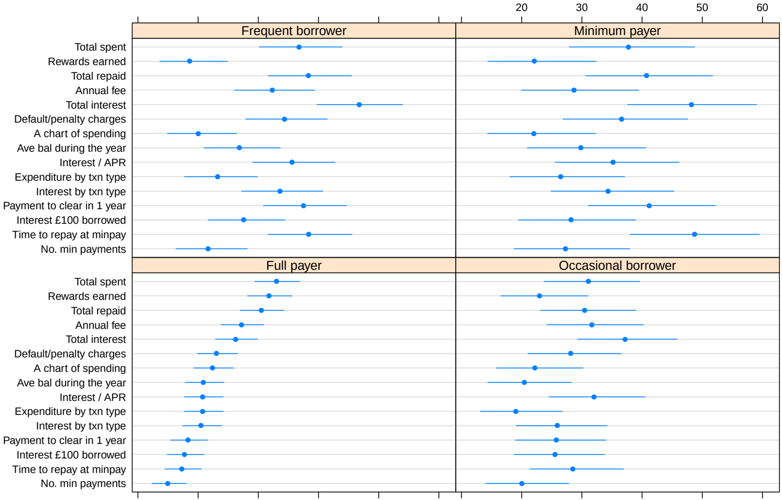
Credit card annual statement research

- Market research survey
- What consumers might value from an annual statement

What to show on an annual statement?

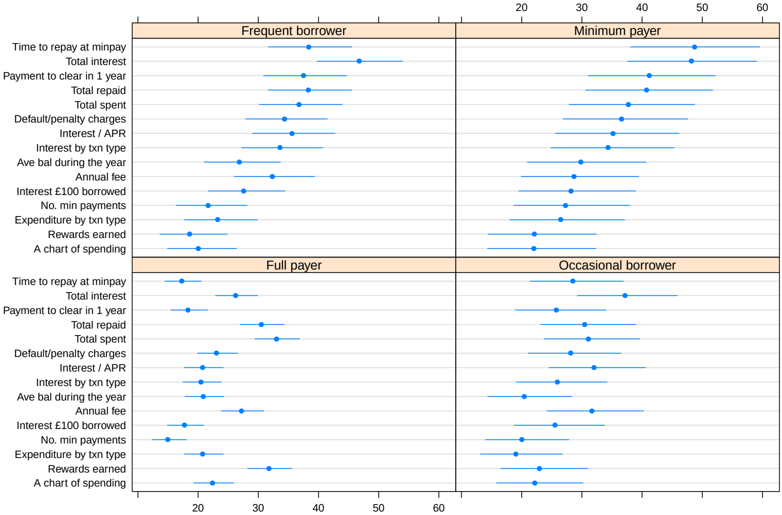
- Different types of card holder may value different features
- How to show this?
- I first saw this idea on a customer satisfaction data set [Clark et al 1999]

Don’t be afraid of using complex graphics (1)



Proportion 'very interested', %
Source: The UK Cards Associaition

Don’t be afraid of using complex graphics (2)



Proportion 'very interested', %
Source: The UK Cards Associaition

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

Exploratory data analysis

> ‘*Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone*’ [Tukey 1977]

EDA is critical

- For every analysis, not just those in business
- Visualising data is an essential first step
- Can also speed up the identification of quality issues

Exploratory data analysis

> ‘*Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone*’ [Tukey 1977]

EDA is critical

- For every analysis, not just those in business
- Visualising data is an essential first step
- Can also speed up the identification of quality issues

There are many benefits to EDA

- Tukey probably didn’t have EDA in mind for data cleaning
- But it’s useful for seeing problems as well as structure
- Very useful for showing patterns to clients

Outline

1 Background

2 The ‘data environment’
- Computers
- Data
- Data quality
- Domain knowledge

3 Visualising data
- Exploratory data analysis
- Presenting the case

4 Case study

5 Final thoughts

6 References

We need to explain results

Use statistical language with caution

- Know your client - whether internal or external
- Use plain English if possible, explain simply and clearly

We need to explain results

Use statistical language with caution

- Know your client - whether internal or external
- Use plain English if possible, explain simply and clearly

Statistical language that might confuse

- Normal
- Error
- Significance
- Variance

We need to explain results

Use statistical language with caution

- Know your client - whether internal or external
- Use plain English if possible, explain simply and clearly

Statistical language that might confuse

- Normal
- Error
- Significance
- Variance

Business imperatives

- Clients often want a quick answer
- In days rather than weeks or months
- Don’t want uncertainty [sic]

Convincing the audience

The case

Whether writing a formal report or giving a presentation, the argument needs to be . . .

- Logical
- Coherent
- Structured

Convincing the audience

The case

Whether writing a formal report or giving a presentation, the argument needs to be . . .

- Logical
- Coherent
- Structured

Remember the audience

- Directors may want the key points on one page
  - Be ready to argue your case
  - Keep it brief
- Analysts will want to see the data and understand the models
  - Always show sources - in an appendix if necessary
  - Be ready to defend the technical aspects
  - Don’t write an academic paper

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

The UK Cards Association

In this example, my client was The UK Cards Association, but the ‘end client’ was the Department for Business Innovation and Skills and the Government

The UK Cards Association

In this example, my client was The UK Cards Association, but the ‘end client’ was the Department for Business Innovation and Skills and the Government

What is it?

- The leading trade association for the card payments industry in the UK
- Its main functions are
  - to facilitate co-operation between industry participants on matters of mutual interest
  - to provide a clear and cohesive industry voice to the outside world
- It provides a forum for its members to come together on non-competitive issues relating to the payments industry

www.theukcardsassociation.org.uk

The Credit and Store Card Consultation

Department for Business, Innovation and Skills (BIS)

- BIS wanted to
  - '*secure a better deal for consumers*'
  - '*give consumers improved control of credit and store card borrowing*'
  - '*ensure that intervention is proportionate, transparent and targeted*'

BIS proposed changes on 5 aspects of credit card operation

The industry's case persuaded BIS that action should be taken on only one of these - and it was the one that consumers saw as unfair

(I worked on the data analysis and presentation)

www.bis.gov.uk/creditconsultation

The response

The response was based entirely on evidence

- Argus Information & Advisory Services LLC
- GfK Financial
- Oxera

The evidence

- Consumer behaviour on 44 million credit card accounts 2008 - 2010
- Market research - qualitative and quantitative
- ‘Market impact assessment’
- Publicly available data - these provided wider context

www.theukcardsassociation.org.uk/responses_to_government/index.asp

The response

Written by a small team of people, working with

- Industry experts
- External experts
- Academics
- Liaison with BIS and other interested parties throughout

The evidence

- The formal response was 230 pages
- The appendices contained 700+ pages

One academic wished his PhD students could write theses as well structured as the Association's response . . .

www.theukcardsassociation.org.uk/responses_to_government/index.asp

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

Final thoughts

1. Visualisation is essential

Final thoughts

1. Visualisation is essential
2. Allow plenty of time for data cleaning

Final thoughts

1. Visualisation is essential
2. Allow plenty of time for data cleaning
3. Talk throughout to the client and other domain experts

Final thoughts

1. Visualisation is essential
2. Allow plenty of time for data cleaning
3. Talk throughout to the client and other domain experts
4. Domain knowledge is essential

Final thoughts

1. Visualisation is essential
2. Allow plenty of time for data cleaning
3. Talk throughout to the client and other domain experts
4. Domain knowledge is essential
5. Sort out the data before doing anything else

Final thoughts

1. Visualisation is essential
2. Allow plenty of time for data cleaning
3. Talk throughout to the client and other domain experts
4. Domain knowledge is essential
5. Sort out the data before doing anything else
6. Structure the case well

Final thoughts

1. Visualisation is essential
2. Allow plenty of time for data cleaning
3. Talk throughout to the client and other domain experts
4. Domain knowledge is essential
5. Sort out the data before doing anything else
6. Structure the case well
7. Know the audience

Final thoughts

1. Visualisation is essential
2. Allow plenty of time for data cleaning
3. Talk throughout to the client and other domain experts
4. Domain knowledge is essential
5. Sort out the data before doing anything else
6. Structure the case well
7. Know the audience

Outline

1. Background
2. The ‘data environment’
   - Computers
   - Data
   - Data quality
   - Domain knowledge
3. Visualising data
   - Exploratory data analysis
   - Presenting the case
4. Case study
5. Final thoughts
6. References

References

Clark LA, Cleveland WS, Denby L and Liu C.
Competitive Profiling Displays: Multivariate Graphs for Customer Satisfaction Survey Data
Marketing Research, **1**, 25-33, 1999.

Cleveland WS.
Visualizing Data
Hobart Press, 1993.

Efron B.
Data Mining and Statistics: What's the connection?
Significance, **7**, 178-181, 2010.

Fermilab/SLAC
Particle physics tames big data
symmetry magazine
http://www.symmetrymagazine.org/article/august-2012/particle-physics-tames-big-data

Freedman D and Diaconis P.
On the histogram as a density estimator: L_2 theory
Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, **57**, 453-476, 1981.

Gentle JE.
Computational Statistics
Springer, 2009.

Hand DJ, Blunt G.
Prospecting for gems in credit card data
IMA Journal of Management Mathematics, **12**:173-200, 2001.

Maletic JI and Marcus A.
A Data Cleansing: A Prelude to Knowledge Discovery
In Maimon O and Rokach L. (Eds.) *Data Mining and Knowledge Discovery Handbook (2nd Ed)*
Springer, 2010.

Murray C.
How to Accuse the Other Guy of Lying with Statistics
Statistical Science, **20**:3, 239-241, 2005.

National Center for Biotechnology Information
Sequence Read Archive database growth
NCBI sequence archive
http://www.ncbi.nlm.nih.gov/Traces/sra/

Nutt D.
Drugs - without the hot air®: Minimising the harms of legal and illegal drugs
UIT, Cambridge, 2012.

References

Ripley BD.
Statistical Methods Need Software: A View of Statistical Computing
Royal Statistical Society Annual Conference, Royal Statistical Society2002.

Scott DW.
On optimal and data-based histograms
Biometrika **66**, 605610, 1979.

Scott DW.
Multivariate Density Estimation. Theory, Practice, and Visualization
Wiley, 1992.

Sturges HA.
Sturges 1926
Journal of the American Statistical Association, **21**, 65-66, 1926.

Tufte ER.
The Visual Display of Quantitative Information
Graphics Press, Connecticut, 2001.

Tukey JW.
Exploratory Data Analysis
Addison Wesley, 1977.