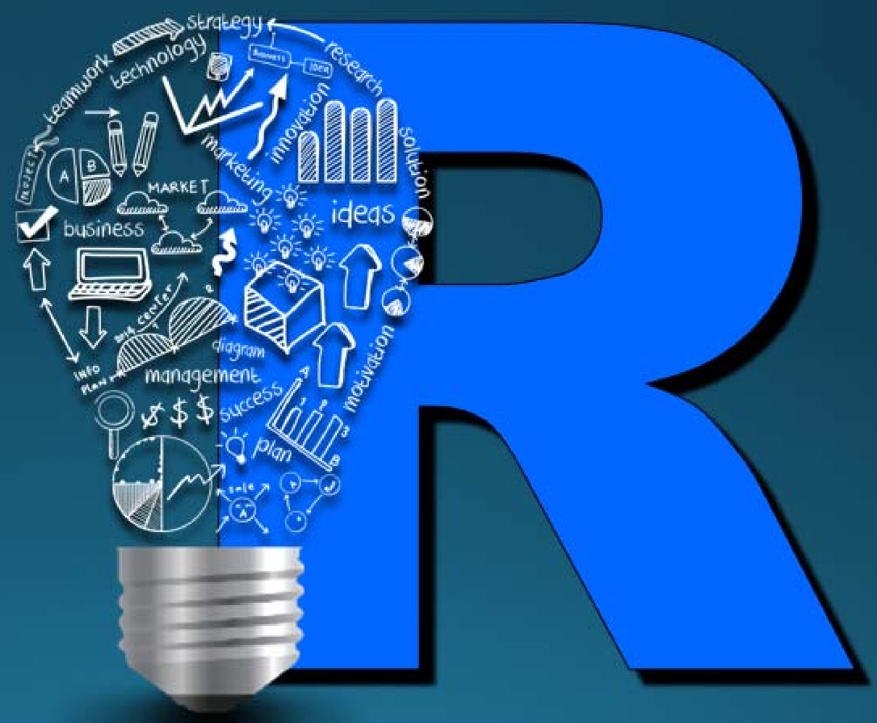
C00-1

探索式資料分析簡介 (EDA)



吳漢銘
國立政治大學 統計學系

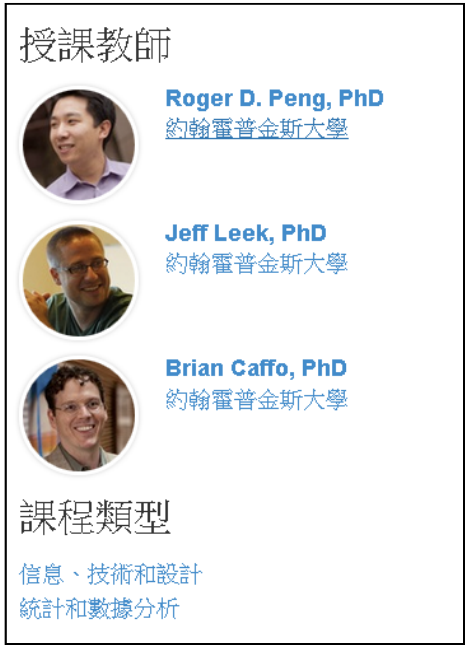
https://hmwu.idv.tw

主要參考書目

https://www.coursera.org/course/exdata

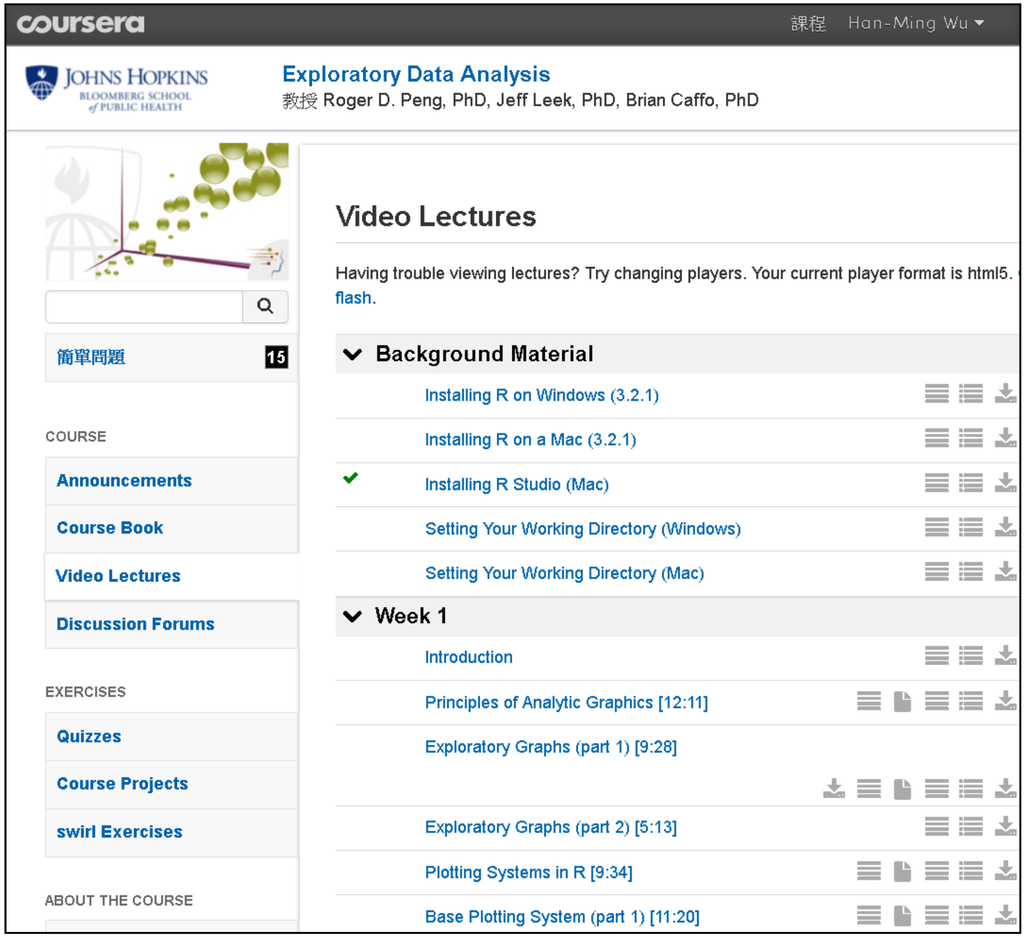
EDA with R: Course Content

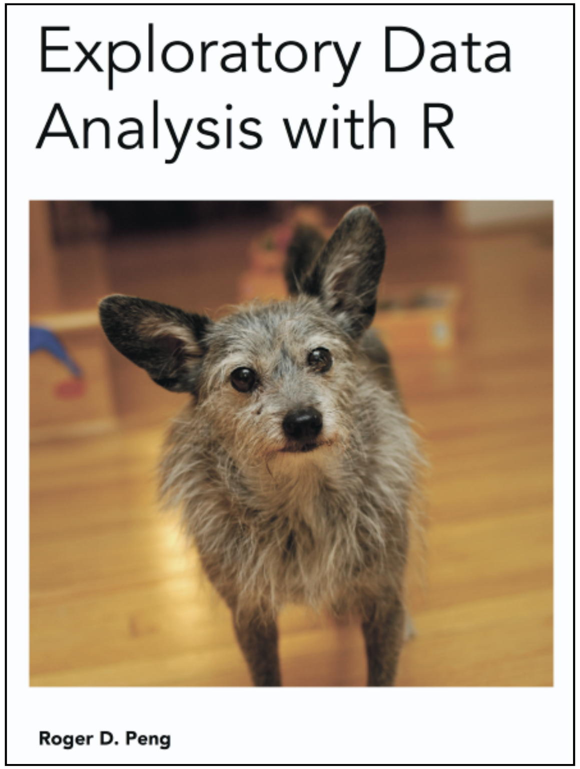
- Making exploratory graphs
- Principles of analytic graphics
- Plotting systems and graphics devices in R
- The base, lattice, and ggplot2 plotting systems in R
- Clustering methods
- Dimension reduction techniques

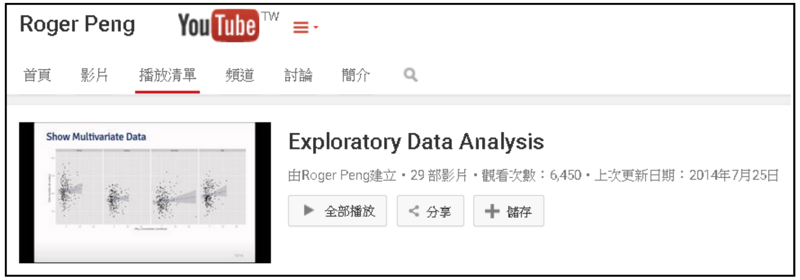


主要參考書目

https://class.coursera.org/exdata-a030/lecture







https://www.youtube.com/playlist?list=PLjTlxb-wKvXPhZ7tQwlROtFjorSj9tUyZ

UDACITY

UDACITY Nanodegree Catalog Sign In Sign Up

Data Analysis with R

Visually Analyze and Summarize Data Sets

Intermediate

Approx. 2 months

Assumes 6hr/wk (work at your own pace)

Built by facebook

Join 41,351 Students

Course Summary

Exploratory data analysis is an approach for summarizing and visualizing the important characteristics of a data set. Promoted by John Tukey, exploratory data analysis focuses on exploring data to understand the data's underlying structure and variables, to develop intuition about the data set, to consider how that data set came into existence, and to decide how it can be investigated with more formal statistical methods.

Start Free Course

Start free course

Free

You get

- Instructor videos
- Learn by doing exercises and view project instructions

Syllabus

Lesson 1: What is EDA? (1 hour)

We'll start by learn about what exploratory data analysis (EDA) is and v
amazing instructors for the course and find out about the course struct

Lesson 2: R Basics (3 hours)

EDA, which comes before formal hypothesis testing and modeling, mak
and summarize data sets. R will be our tool for generating those visuals
lesson, we will install RStudio and packages, learn the layout and basic
basic R scripts, and inspect data sets.

Lesson 3: Explore One Variable (4 hours)

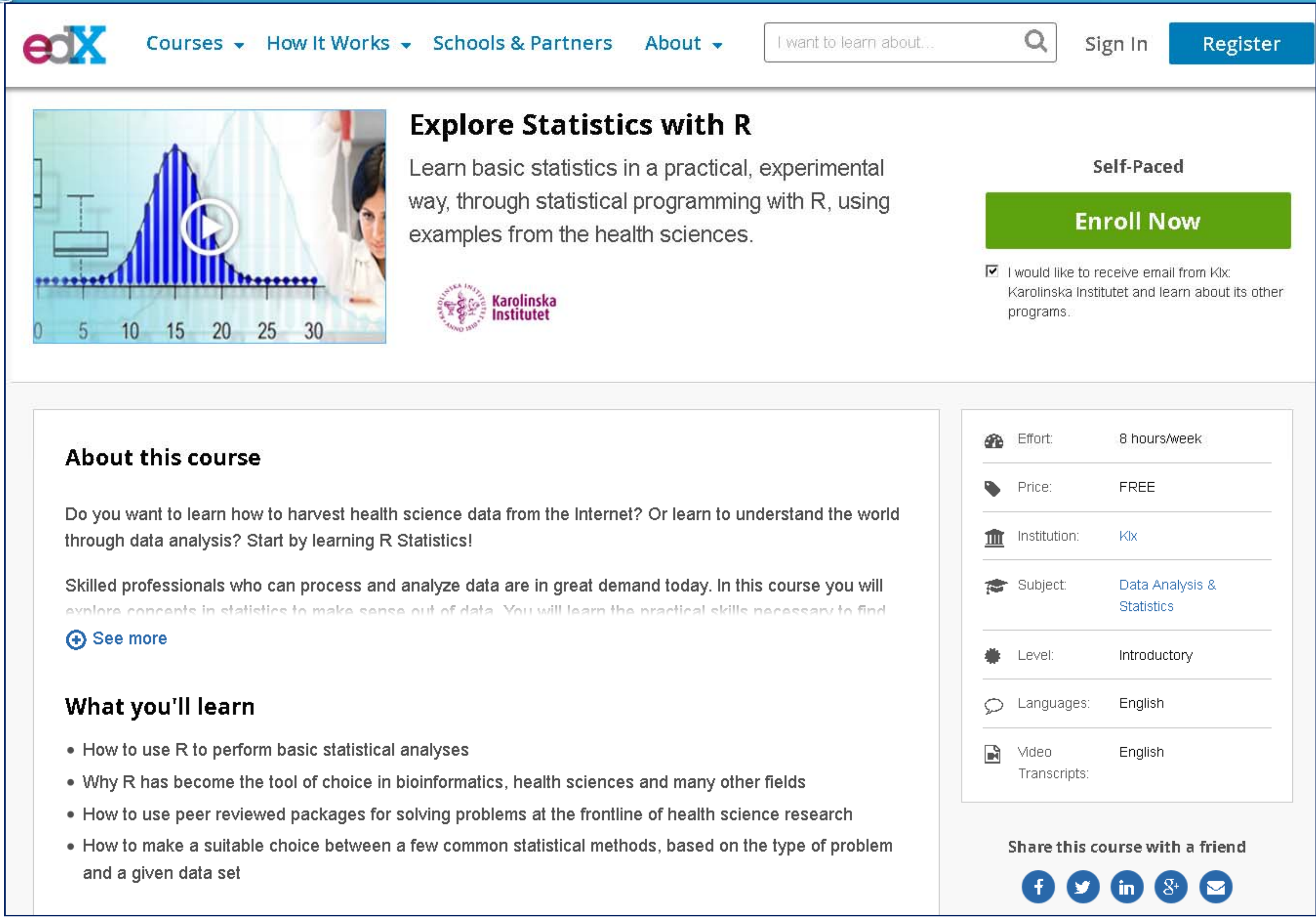
Lesson 4: Explore Two Variables (4 hours)

Lesson 5: Explore Many Variables (4 hours)

Lesson 6: Diamonds and Price Predictions (2 hours)

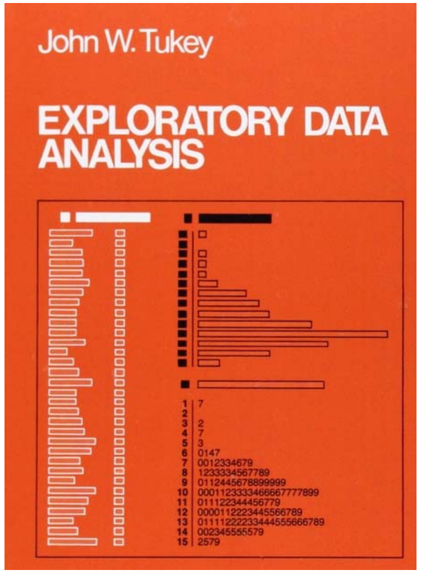
Final Project (10+ hours)

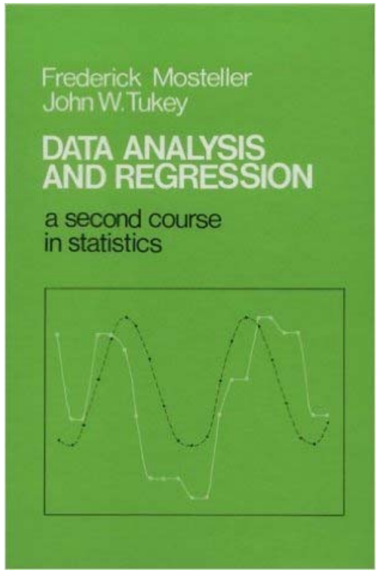
edX

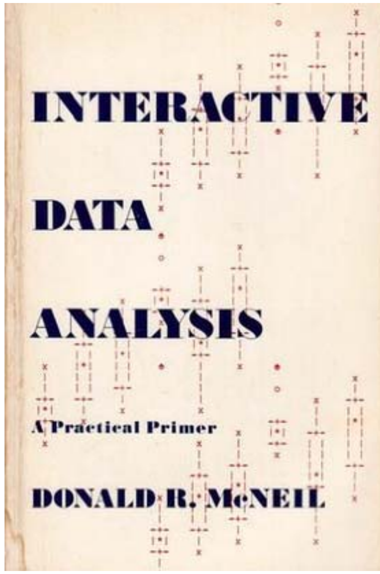


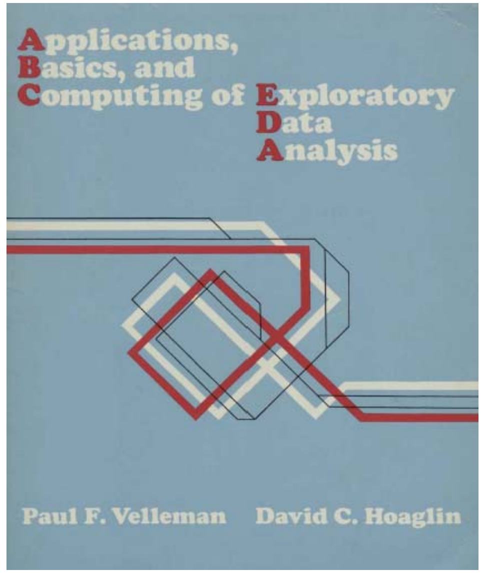
註: 三大MOOC 巨頭：Coursera 、Udacity 、edX 比較
http://www.owstartup.com/2014/05/30/coursera-edx-udacity-review/

EDA經典書籍



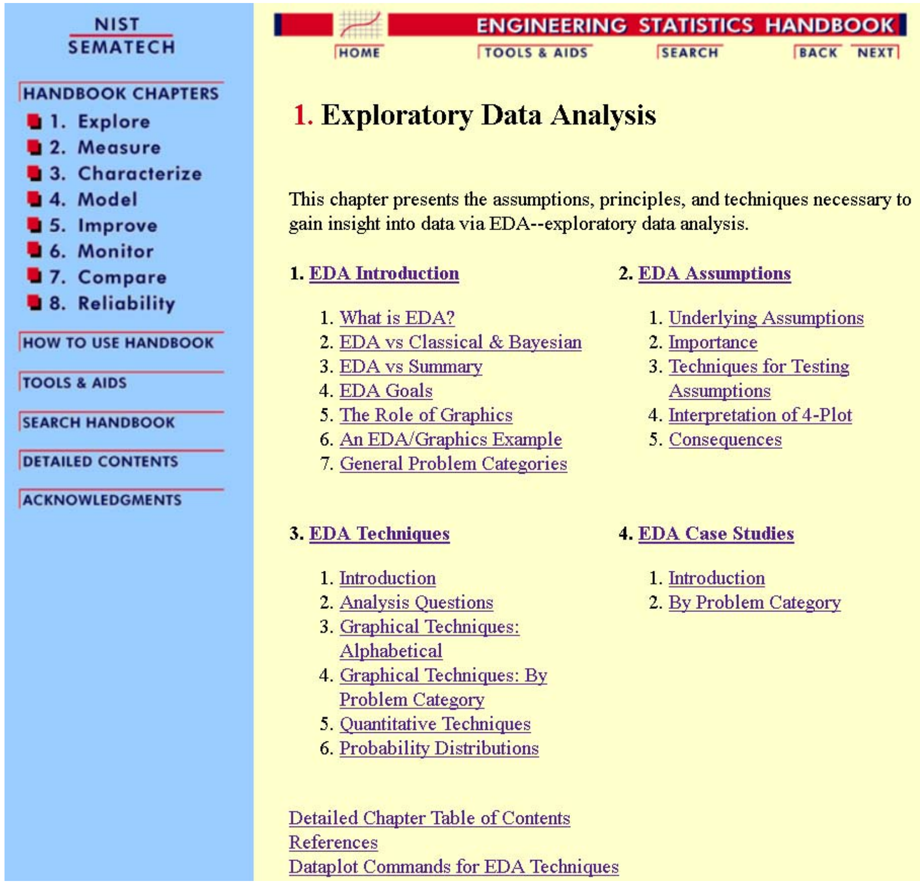






- The seminal work in EDA is *Exploratory Data Analysis*, Tukey, (1977). Over the years it has benefitted from other noteworthy publications such as *Data Analysis and Regression*, Mosteller and Tukey (1977), *Interactive Data Analysis*, Donald (1977), *The ABC's of EDA*, Velleman and Hoaglin (1981) and has gained a large following as "the" way to analyze a data set.

參考文獻

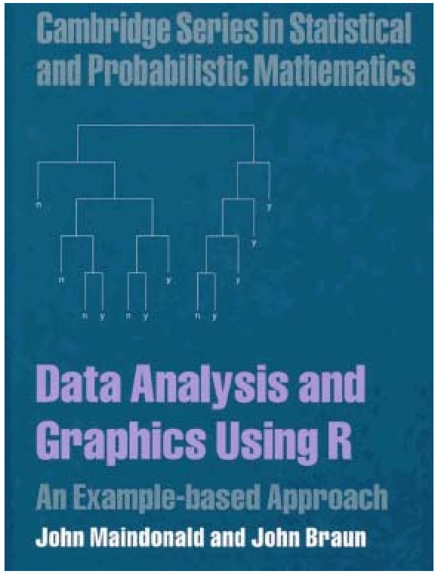


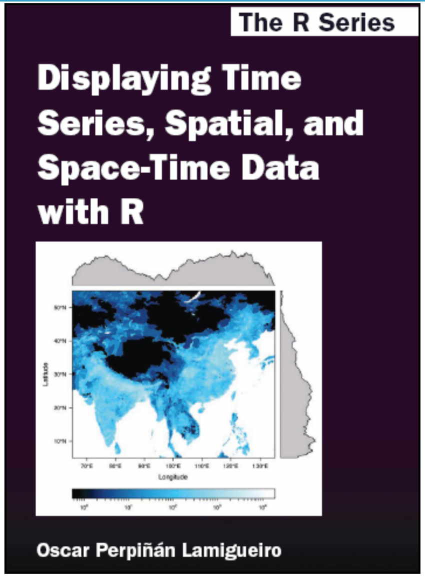
http://www.itl.nist.gov/div898/handbook/index.htm

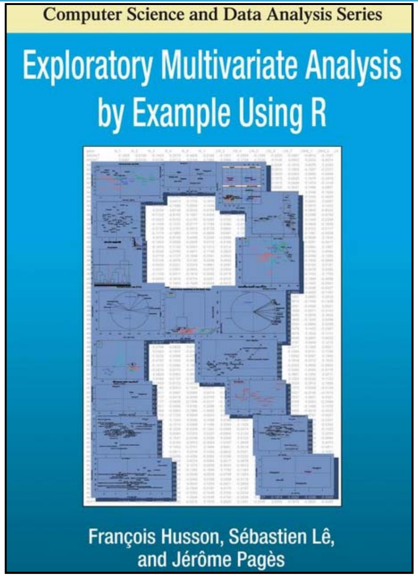
Selected References For EDA

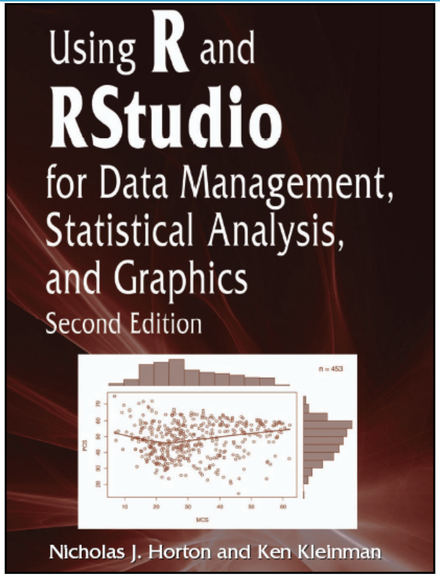
- Anscombe, F. and Tukey, J. W. **(1963)**, The Examination and Analysis of Residuals, Technometrics, pp. 141-160.
- Box, G. E. P. and Cox, D. R. **(1964),** An Analysis of Transformations, Journal of the Royal Statistical Society, pp. 211-243, discussion pp. 244-252.
- Wilk, M. B. and Gnanadesikan, R. **(1968)**, Probability Plotting Methods for the Analysis of Data, Biometrika, 5(5), pp. 1-19.
- Anscombe, F. **(1973)**, Graphs in Statistical Analysis, The American Statistician, pp. 195-199.
- Filliben, J. J. **(1975)**, The Probability Plot Correlation Coefficient Test for Normality, Technometrics, pp. 111-117.
- McNeil, Donald **(1977)**, Interactive Data Analysis, John Wiley and Sons.
- Tukey, John **(1977)**, Exploratory Data Analysis, Addison-Wesley.
- Velleman, Paul and Hoaglin, David **(1981)**, The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis, Duxbury.
- Wainer, Howard **(1981)**, Visual Revelations, Copernicus.
- Tufte, Edward **(1983)**, **The Visual Display of Quantitative Information**, Graphics Press.
- Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, **(1983)**, Graphical Methods for Data Analysis, Wadsworth.
- Cleveland, William **(1985)**, **Elements of Graphing Data**, Wadsworth.
- du Toit, Steyn, and Stumpf **(1986)**, Graphical Exploratory Data Analysis, Springer-Verlag.
- Cleveland, William and Marylyn McGill, Editors **(1988)**, Dynamic Graphics for Statistics, Wadsworth.
- Cleveland, William **(1993)**, **Visualizing Data**, Hobart Press.
- Barnett and Lewis **(1994)**, Outliers in Statistical Data, 3rd. Ed., John Wiley and Sons.
- Harris, Robert L. **(1996)**, Information Graphics, Management Graphics.

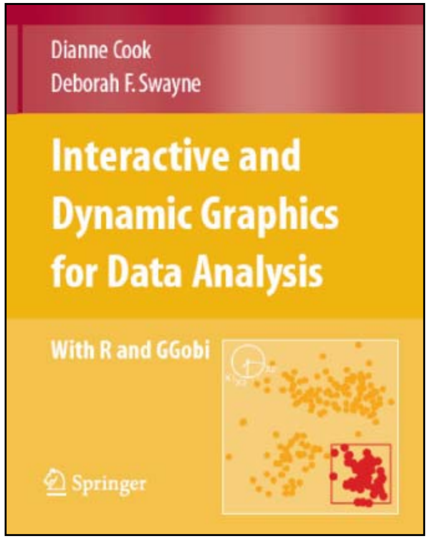
參考書目 (1/2)

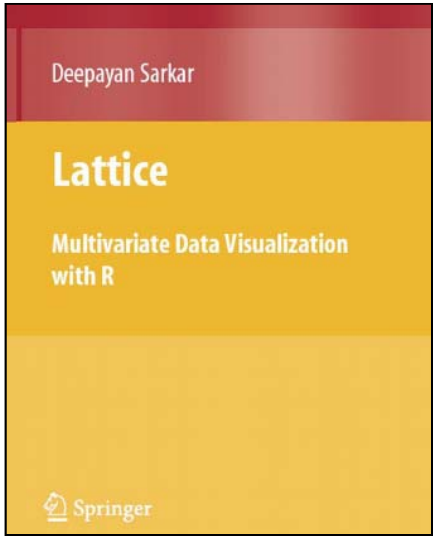

Cambridge Series in Statistical and Probabilistic Mathematics
Data Analysis and Graphics Using R
An Example-based Approach
John Maindonald and John Braun

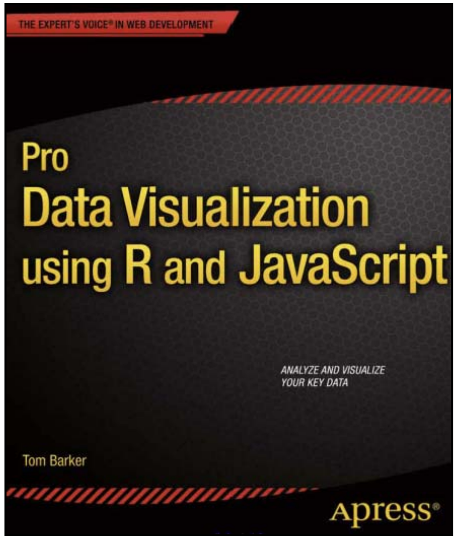

The R Series
Displaying Time Series, Spatial, and Space-Time Data with R
Oscar Perpiñán Lamigueiro

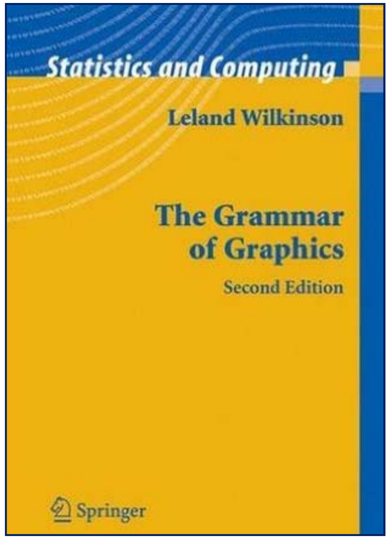

Computer Science and Data Analysis Series
Exploratory Multivariate Analysis by Example Using R
François Husson, Sébastien Lê, and Jérôme Pagès


Using R and RStudio for Data Management, Statistical Analysis, and Graphics
Second Edition
Nicholas J. Horton and Ken Kleinman

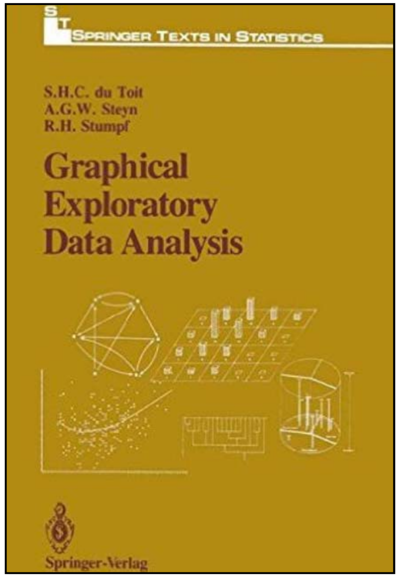

Dianne Cook
Deborah F. Swayne
Interactive and Dynamic Graphics for Data Analysis
With R and GGobi
Springer

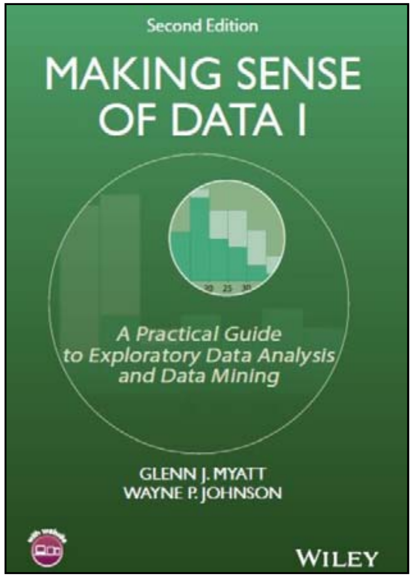

Deepayan Sarkar
Lattice
Multivariate Data Visualization with R
Springer

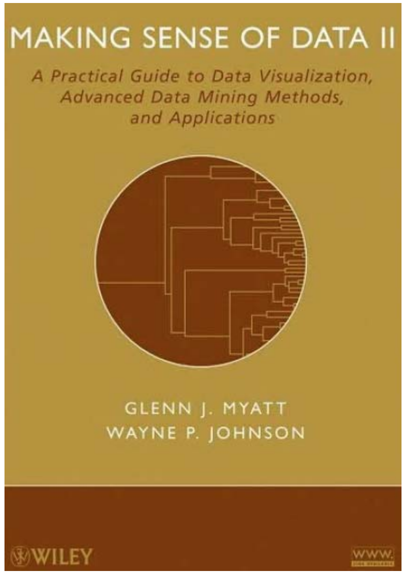

THE EXPERT'S VOICE® IN WEB DEVELOPMENT
Pro
Data Visualization using R and JavaScript
ANALYZE AND VISUALIZE YOUR KEY DATA
Tom Barker
Apress®

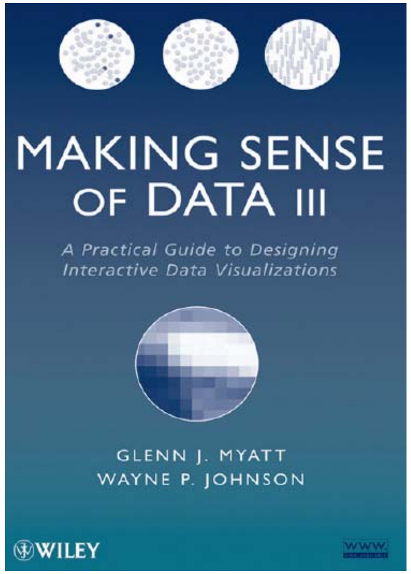

Statistics and Computing
Leland Wilkinson
The Grammar of Graphics
Second Edition
Springer

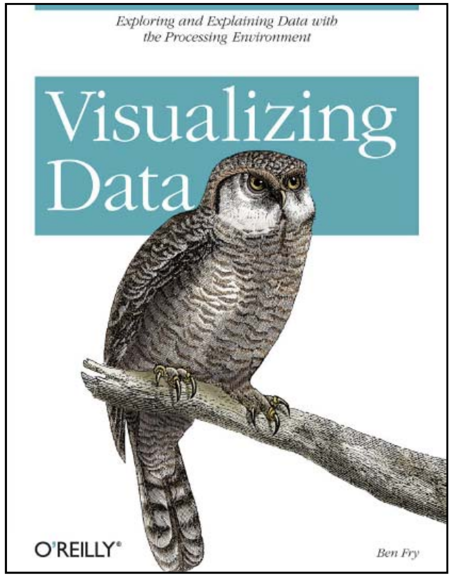
參考書目 (2/2)

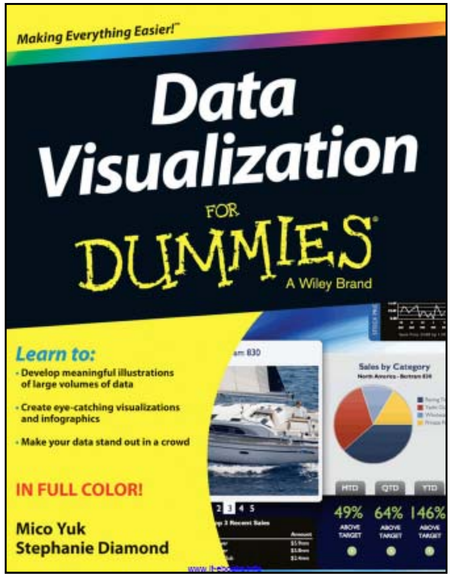

Springer Texts in Statistics
S.H.C. du Toit
A.G.W. Steyn
R.H. Stumpf
Graphical Exploratory Data Analysis
Springer-Verlag

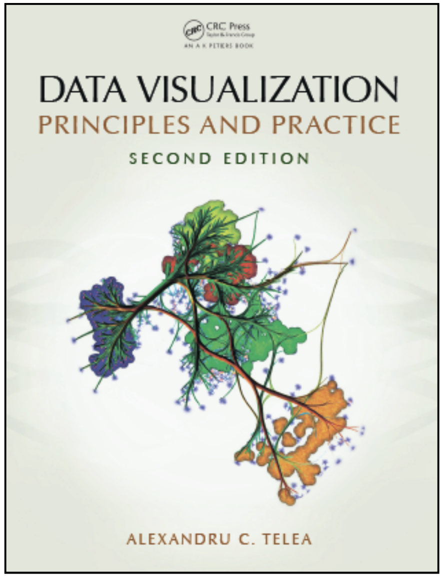

Second Edition
MAKING SENSE OF DATA I
A Practical Guide to Exploratory Data Analysis and Data Mining
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY

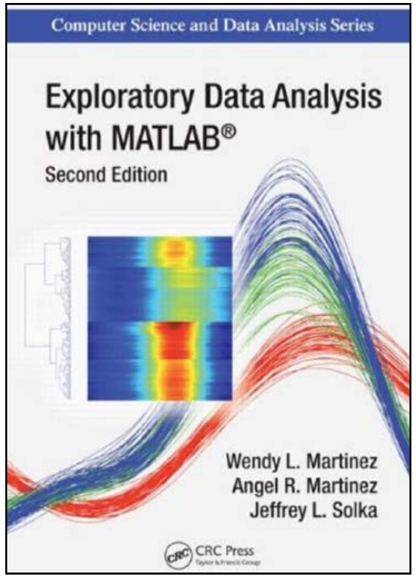

MAKING SENSE OF DATA II
A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY


MAKING SENSE OF DATA III
A Practical Guide to Designing Interactive Data Visualizations
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY


Exploring and Explaining Data with the Processing Environment
Visualizing Data
O'REILLY
Ben Fry


Making Everything Easier!
Data Visualization FOR DUMMIES
A Wiley Brand
Learn to:
Develop meaningful illustrations of large volumes of data
Create eye-catching visualizations and infographics
Make your data stand out in a crowd
IN FULL COLOR!
Mico Yuk
Stephanie Diamond


CRC Press
DATA VISUALIZATION
PRINCIPLES AND PRACTICE
SECOND EDITION
ALEXANDRU C. TELEA

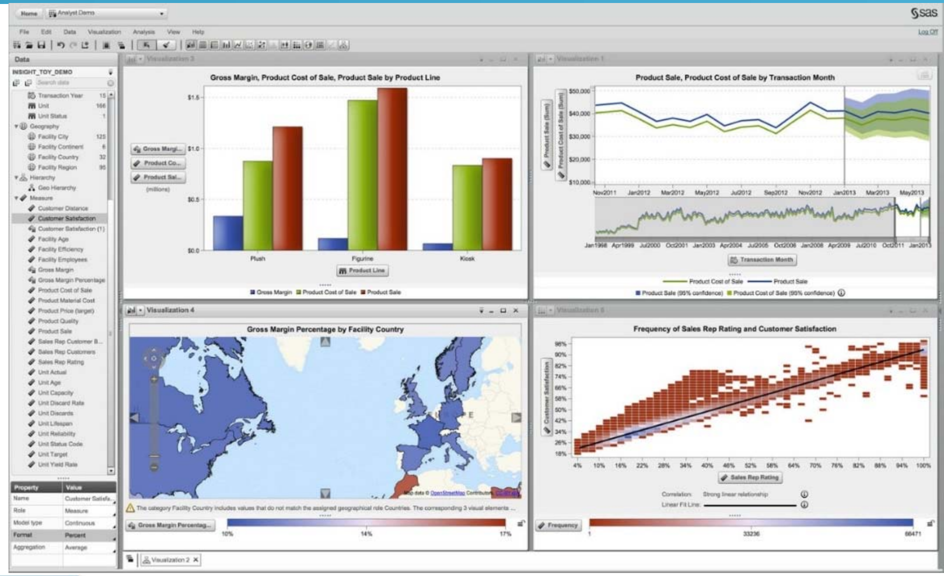

Computer Science and Data Analysis Series
Exploratory Data Analysis with MATLAB®
Second Edition
Wendy L. Martinez
Angel R. Martinez
Jeffrey L. Solka
CRC Press

EDA Software (1/3)

THE POWER TO KNOW.

SAS Visual Analytics

優越的資料視覺化介面、簡單的分析操作
帶您發現無限可能的新世界



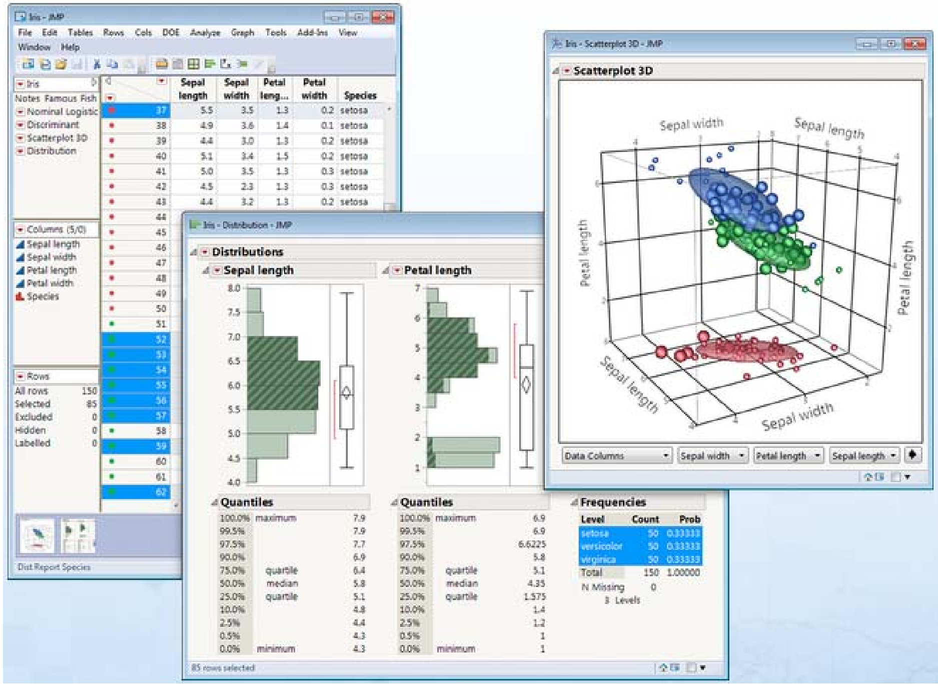
SAS Visual Statistics

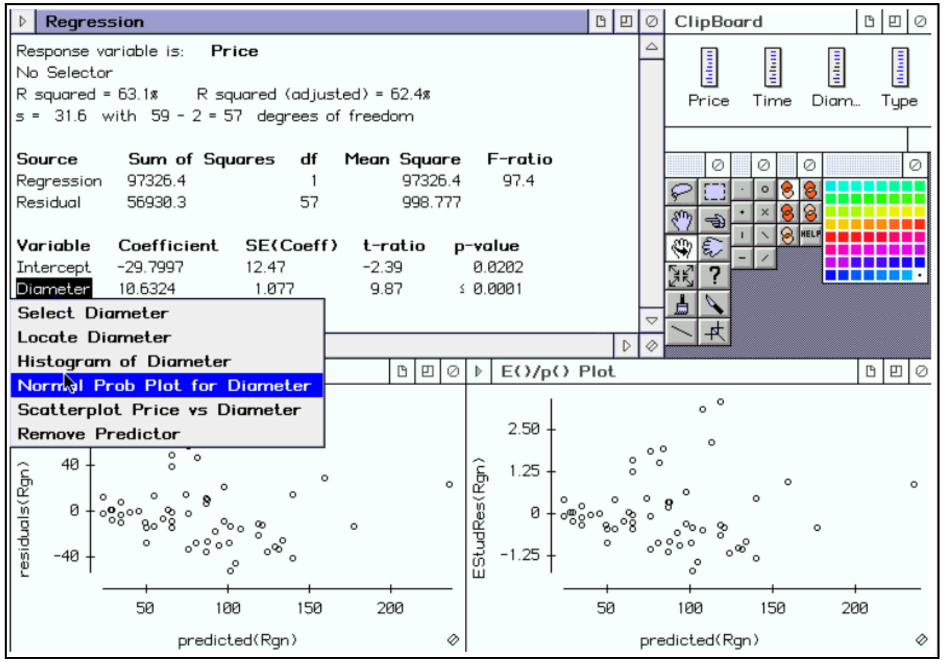
以互動方式建立、執行及評估高效能分析模型，
立即獲得結果

EDA Software (2/3)

jmp Statistical Discovery.™ From SAS.

Source: http://www.jmp.com/en_us/software/jmp.html





DataDesk® 7

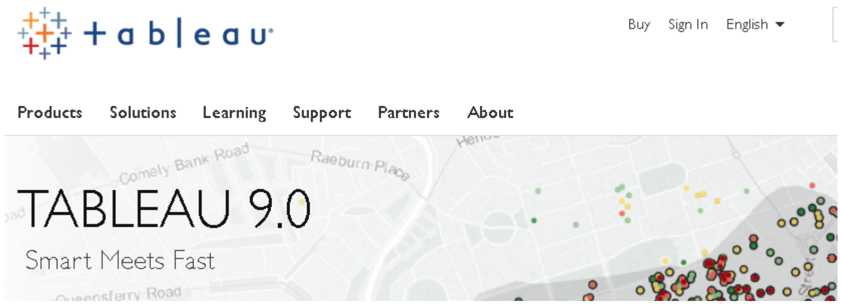
Source:
http://www.datadesk.com/products/data_analysis/datadesk/

Others: Fathom (Keypress), Data Explorer

EDA Software (3/3)

http://www.ggobi.org/

The current version of GGobi is 2.1.10a, released 12 March 2010, and updated 10 June 2012 for 64 bit.



http://www.tableau.com/new-features/9.0

Others:

- Fathom (Keypress)
- Data Explorer
- http://factominer.free.fr/
- FactoMineR is an R package dedicated to multivariate Exploratory Data Analysis.

The Best Data Visualization Tools for Big Data

The Best Data Visualization Tools for Big Data, http://www.scriptiny.com/2013/09/best-data-visualization-tools-big-data/

- **Tableau**, Birst, QlikView, SAP BusinessObjects

2015 Best TopTen Data Visualization Software, http://data-visualization-software-review.toptenreviews.com/

- Advizor, Birst, Datawatch, Pentaho Software, SAP Lumira, **SAS Visual Analytics**, **Tableau**, Targit, TIBCO Spotfire, ZingChart

The 14 best data visualization tools, http://thenextweb.com/dd/2015/04/21/the-14-best-data-visualization-tools/

- D3.js, FusionCharts, Chart.js, Google Charts, Highcharts, Leaflet, dygraphs, Datawrapper, **Tableau**, Raw, Timeline JS, Infogram, Plotly, ChartBlocks

The 37 best tools for data visualization, http://www.creativebloq.com/design-tools/data-visualization-712402

- Dygraphs, ZingChart, InstantAtlas, Timeline, Exhibit, Modest Maps, Leaflet, WolframAlpha, Visual.ly, Visualize Free, Better World Flux, FusionCharts, jqPlot, Dipity, Many Eyes, D3.js, JavaScript InfoVis Toolkit, jpGraph, Highcharts, Google Charts, **Excel**, CSV/JSON, Crossfilter, Tangle, Polymaps, OpenLayers, Kartograph, CartoDB, Processing, NodeBox, R, Weka, Gephi, iCharts , Flot, Raphaël, jQuery Visualize

39 Data Visualization Tools for Big Data, https://blog.profitbricks.com/39-data-visualization-tools-for-big-data/

- Polymaps, NodeBox, Flot, Processing, Processingjs.org, Tangle, D3.js, FF Chartwell, Google Maps, **SAS Visual Analytics**, Raphael, Inkscape, Leaflet, Crossfilter, OpenLayers, Kartograph, Microsoft Excel, Modest Maps, CartoDB, Google Charts, Gephi, Flare, Envision.js, Miso, **The R Project**, **Tableau Public**, Timeline JS, Quadrigram, Prefuse, Many Eyes, Cytoscape, NetworkX, Arbor.js, iCharts, Databoard, Q Research Software, Dapresy, Visualize Free, Jolicharts

Top Data Visualization Software Products, http://www.capterra.com/data-visualization-software/

John Tukey (1915~2000): 統計學界的畢卡索

John W. Tukey is the first statisticians to provide a detailed description of exploratory data analysis (EDA).

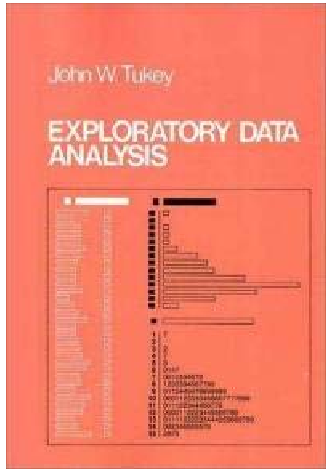
生平

- 布朗大學**化學**學士及碩士。
- 1939年: 普林斯頓大學**數學**博士。(數理統計)
- 二次大戰加入火砲控制研究室，以及後來加入**AT&T**貝爾實驗室(**創立統計組**)，接觸統計上的實際問題。

「對正確的問題有個近似的答案，
勝過對錯的問題有精確的答案。」

對後世的貢獻

- 發明快速傅立葉轉換(FFT)。
- 創造bit (位元)及 software(軟體) 。
- 探索性的資料分析 (Exploratory Data Analysis, EDA, 1977)



Source: http://www.unige.ch/ses/sococ/cl/bib/eda/tukey.html

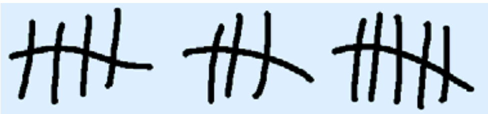
「統計應該是科學，而非數學！」

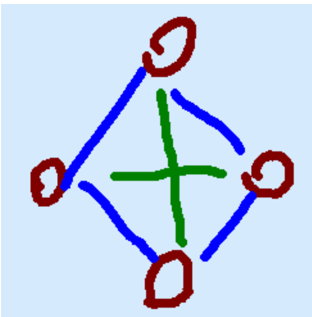
Stanford Linear Accelerator (1973)

他曾挑戰當時主流的數理統計學家，堅持 data analysis 是統計分析中不可忽視的步驟，數學的假設需要 data 加以驗證才可行。 Tukey 說過統計應該是科學，而非數學！

數學思維 vs 統計思維

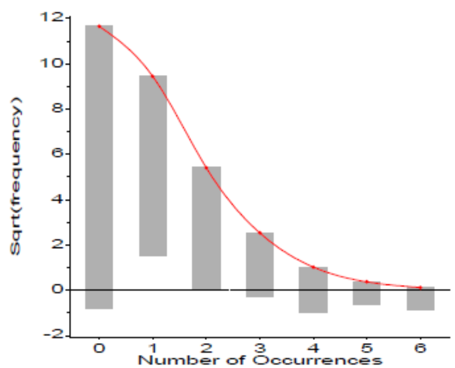
証明在哪裏? vs 數據在哪裏?



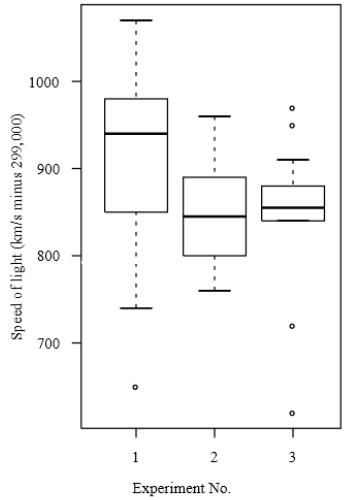


Stem and Leaf Plot

```
42 | 0
44 | 0000
46 | 000000
48 | 00000000000
50 | 0000000000000000000
52 | 00000
54 | 0000000000000
56 | 00000000000000
58 | 0000000000
60 | 000000000000
62 | 0000000000000
64 | 000000000000
66 | 0000000000
68 | 0000000
70 | 00
72 | 0000
74 | 0
76 | 00000
78 | 0
```



Box-and-whisker plot



What is EDA? (1/2)

- Exploratory Data Analysis (EDA) is an **approach/philosophy** for data analysis that employs a variety of techniques (mostly **graphical**) to
  - maximize **insight** into a data set;
  - uncover underlying **structure**;
  - extract important variables;
  - detect **outliers** and anomalies (detection of mistakes);
  - test underlying **assumptions**;
  - develop parsimonious **models** (preliminary selection of appropriate models);
  - determine **optimal** factor settings;
  - determine **relationships** among the explanatory variables; and
  - assess the direction and rough size of relationships between explanatory and outcome variables.

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

What is EDA? (2/3)

- **Goal: get a general sense of the data**
  - means, medians, quantiles, histograms, boxplots
  - You should always look at every variable - you will learn something!
- **Data-driven (model-free)**
- **Think interactive and visual**
  - Humans are the best pattern recognizers
  - You can use more than 2 dimensions!
    - x, y, z, space, color, time....
- **Especially useful in early stages of data mining**
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g. log(x))
- Bottom line: it is always well worth looking at your data!

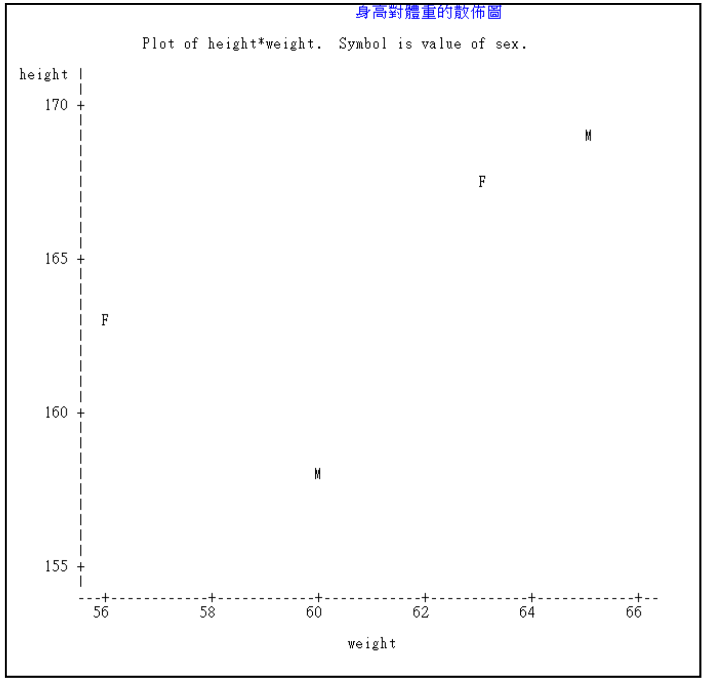
Source: http://www2.research.att.com/~volinsky/DataMining/Columbia2011/Slides/Topic2-EDAViz.ppt

What is EDA? (3/3)

- Defined EDA as **detective** work - **numerical** detective work - or **counting** detective work - or **graphical** detective work.
- Any method of looking at data that does not include **formal statistical modeling and inference** falls under the term EDA.
- Before 1970, computers were not widely available, the data sets tended to be somewhat small. Nowadays, EDA engaged in **computationally intensive methods** for pattern discovery and statistical visualization.
- The philosophy of EDA is the same - that those engaged in it are **data detectives**.

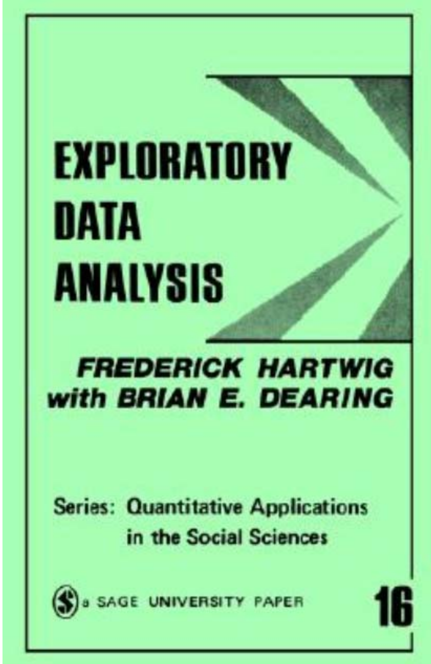
```
Stemleaf:GNPAgr(20) %GNP for Agriculture
 Groups defined by Continents
 Legend:  0|0 stands for 0.00;   5|5 for 55.00
     Asia                  Africa                Europe
    0|0111112234          |34                   |1123333334444
    0|55778               |55567                |55666788
    1|01                  |1224444              |134
    1|6899                |566                  |667
    2|123                 |111112               |03
    2|5567                |577                  |
    3|24                  |011344               |3
    3|9                   |567                  |
    4|12                  |3444                 |
    4|                    |5555677              |
    5|11                  |01123                |
    5|                    |5                    |

 hi  |(* 4)               |GNEQ                 |
```



What Do They Say About EDA? (1/4)

- **Hartwig and Dearing (1979)** specify two principles for EDA: **skepticism** and **openness**. This might involve
  - visualization of the data to look for anomalies or patterns,
  - the use of resistant (or robust) statistics to summarize the data,
  - openness to the transformation of the data to gain better insights, and
  - the generation of models.



- **Chatfield (1985):**
  - EDA emphasis on starting with the **noninferential** approach in data analysis.
  - the need for looking at how the data were collected, what are the objectives of the analysis.

• Chatfield, C. 1985. The initial examination of data, Journal of the Royal Statistical Society, A, 148:214-253

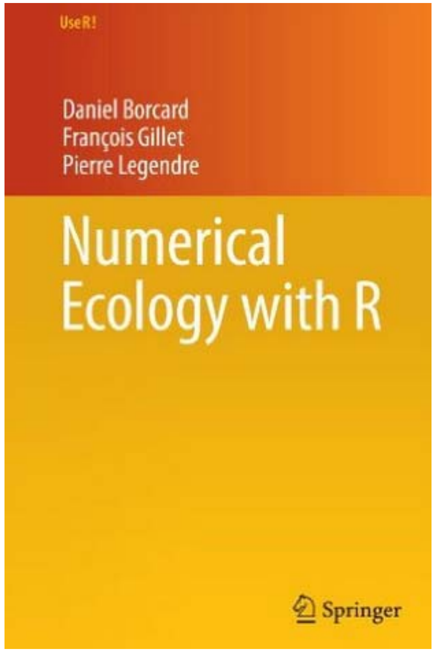
What Do They Say About EDA? (2/4)

- **Hoaglin (1982):** EDA encompasses four themes:
  - Resistant: data analysis methods where an arbitrary change in a data point or small subset of the data yields a small change in the result.
    - robustness
  - Residuals: what we have left over after a summary or fitted model has been subtracted out. (residual = data - fit.)
    - Residuals should be looked at for lack of fit, heteroscedasticity (nonconstant variance), nonadditivity, and other interesting characteristics of the data.
  - Re-expression: the transformation of the data to some other scale that might make the variance constant, might yield symmetric residuals, could linearize the data, or add some other effect.
    - The goal of reexpression for EDA is to facilitate the search for structure, patterns, or other information.
  - Display: visualization techniques for EDA.
    - Often the only way to discover patterns, structure, or to generate hypotheses is by visual transformations of the data.

- Hoaglin, D. C. 1982. Exploratory data analysis, in Encyclopedia of Statistical Sciences, Volume 2, Kotz, S. and N. L. Johnson, eds., New York: John Wiley & Sons.

What Do They Say About EDA? (3/4)

- **Daniel Borcard, Francois Gillet, Pierre Legendre (2011):**
  - A first exploratory look at the data can tell much about them.
  - Information about simple parameters and distributions of variables is important to consider in order to choose more advanced analyses correctly.
  - EDA is often neglected by people who are eager to jump to more sophisticated analyses. It should have an important place.



What Do They Say About EDA? (4/4)

- **Howard J. Seltman (2015)**, Experimental Design and Analysis.
  - EDA need not be restricted to techniques you have seen before; sometimes you need to **invent a new way** of looking at your data.
  - You should always perform appropriate EDA before further analysis of your data.
  - Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables.
  - EDA is not an exact science, it is a very **important art**!

Philosophy of EDA

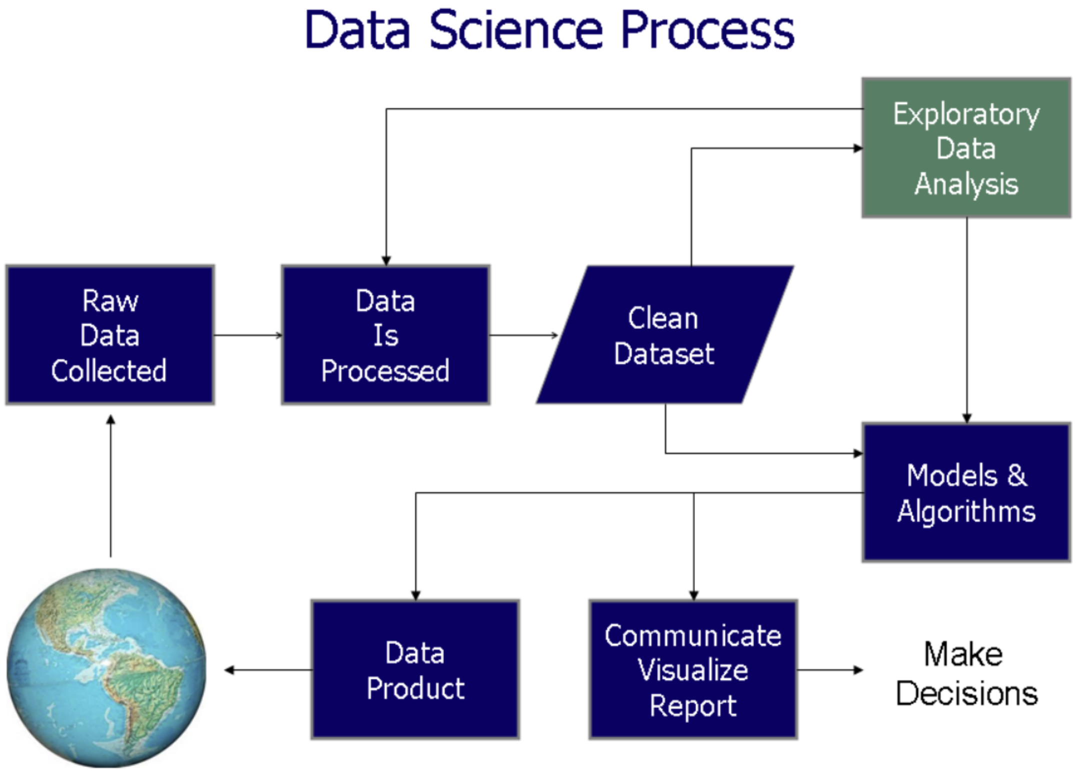
- The EDA approach is precisely that--an approach--not a set of techniques, but an **attitude/philosophy** about how a data analysis should be carried out.

- EDA is not identical to statistical graphics (two terms are used almost interchangeably.)
  - Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect.
  - EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the **data itself** to reveal its underlying structure and model.
  - EDA is a philosophy as to how we **dissect a data set**; what we look for; how we look; and how we interpret.

Source: http://www.itl.nist.gov/div898/handbook/eda/

Data Analysis Procedures

- Statistics and data analysis procedures can broadly be split into two parts:
  - **Graphical techniques** include scatter plots, histograms, probability plots, residual plots, box plots, block plots.
  - **Quantitative techniques** are the set of statistical procedures that yield numeric or tabular output: hypothesis testing, analysis of variance, point estimates and confidence intervals, least squares regression (classical analysis).



Source: https://en.wikipedia.org/wiki/Exploratory_data_analysis

EDA Techniques

- The main role of EDA is to open-mindedly explore:
  - **Plotting the raw data** (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.
  - **Plotting simple statistics** such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
  - **Positioning such plots** so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

- The graphical tools are the **shortest path** to gaining insight into a data set in terms of
  - testing assumptions, model selection, model validation, estimator selection, relationship identification, factor effect determination, outlier detection.

Four Types of EDA

- The four types of EDA:
  - **Non-graphical methods** generally involve calculation of summary statistics, while **graphical methods** obviously summarize the data in a diagrammatic or pictorial way.
  - **Univariate methods** look at one variable (data column) at a time, while **multivariate methods** look at two or more variables at a time to explore relationships.
- Perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.
- Each of the categories of EDA have further divisions based on the role (**outcome or explanatory**) and type (**categorical or quantitative**) of the variable(s) being examined.

Source: google images

Why EDA?

- Much of the quality of scientific work is determined by the quality of the hypotheses and models used by the researcher. Can data analysis help suggest hypotheses?
  - Data analysis tools are typically used for Hypothesis testing and Parameter estimation.
  - Graphics tools are typically used for **presentation**.
- Quantitative statistics are **incomplete**:
  - The numeric summaries focus on a particular aspect of the data (e.g., location, intercept, slope, degree of relatedness, etc.) by judiciously reducing the data to a few numbers.
  - Doing so also filters the data, necessarily omitting and screening out other sometimes **crucial information**. (misleading at worst)

The Objectives of EDA (1/3)

- The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:
  - a good-fitting, parsimonious model,
  - a list of outliers,
  - a sense of robustness of conclusions,
  - estimates for parameters,
  - uncertainties for those estimates,
  - a ranked list of important factors,
  - conclusions as to whether individual factors are statistically significant, and
  - optimal setting.

The Objectives of EDA (2/3)

- The objectives of EDA
  - Suggest hypotheses about the causes of observed phenomena.
  - Assess assumptions on which statistical inference will be based.
  - Support the selection of appropriate statistical tools and techniques.
  - Provide a basis for further data collection through surveys or experiments.

- Many EDA techniques have been adopted into **data mining**, as well as into big data analytics. They are also being taught to young students as a way to introduce them to **statistical thinking**.

Source: https://en.wikipedia.org/wiki/Exploratory_data_analysis

The Objectives of EDA (3/3)

- **Insight into the Data**
  - Insight implies detecting and uncovering underlying structure in the data.
  - Graphics are **irreplaceable**--there are no quantitative analogues that will give the same insight as well-chosen graphics.

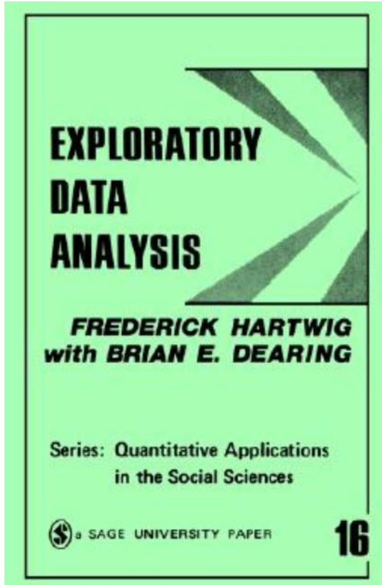
- **"Feel" for the data**
  - The "feel" for the data comes almost exclusively from the application of various graphical techniques.
  - It is not enough for the analyst to know what is in the data; the analyst also must know what is not in the data.
  - The only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data.

EDA v.s. CDA (1/2)

- Confirmatory data analysis (CDA): data analysis that is mostly concerned with statistical hypothesis testing, confidence intervals, estimation, etc.

- EDA and CDA should be used in a **complementary way**:
  - The analyst explores the data looking for patterns and structure that leads to hypotheses and models.

- **Hartwig and Dearing (1979)**:
  - CDA: the one that answers questions such as "Do the data confirm hypothesis XYZ?"
  - EDA: tends to ask "What can the data tell me about relationship XYZ?"



EDA v.s. CDA (2/2)

- **Tukey (1980)** presents a typical straight-line methodology for CDA:
  1. State the questions to be investigated.
  2. Design an experiment to address the questions.
  3. Collect data according to the designed experiment.
  4. Perform a statistical analysis of the data.
  5. Produce an answer.

- To incorporate EDA, Tukey revises the first two steps as follows:
  1. Start with some idea.
  2. Iterate between asking a question and creating a design.

Tukey, J.W. (1962) The future of data analysis, Annals of Mathematical Statistics 33(1), pp. 1-67.
Tukey, J.W. (1980, page 24), We need both exploratory and confirmatory, The American Statistician, 34(1) pp. 23-25

EDA vs. Summary Analysis

- Summary
  - A summary analysis is simply a **numeric reduction** (summary table, e.g., mean and sd) of a historical data set. Its focus is in the past.
  - Summary statistics are passive and historical.

- Exploratory
  - EDA desires to gain insight into the engineering/scientific process behind the data.
  - In an attempt to "**understand**" the process and improve it in the future, EDA uses the data as a "window" to peer into the heart of the process that generated the data.
  - EDA is active and futuristic.

Three Analysis Techniques

- Three popular data analysis approaches, classical, exploratory, Bayesian, are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions.

- **Classical analysis:**

  Problem → Data → Model → Analysis → Conclusions

- **EDA:**

  Problem → Data → Analysis → Model → Conclusions

- **Bayesian:**

  Problem → Data → Model → Prior Distribution → Analysis → Conclusions

Classical, EDA, Bayesian

Method of Dealing with Underlying Model

- Classical analysis:
  - the data collection is followed by the **imposition of a model** (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.
- EDA:
  - the data collection is followed immediately by analysis with a goal of **inferring what model** would be appropriate.
- Bayesian analysis:
  - the analyst attempts to incorporate scientific/engineering **knowledge/expertise** into the analysis by imposing a data-independent distribution on the parameters of the selected model.
  - the analysis consists of combining both the **prior distribution** on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

- In the real world, data analysts freely mix elements of all of the above three approaches (and other approaches).

Exploratory vs. Classical: Model

- EDA vs classical, differ at (1) Models, (2) Focus, (3) Techniques, (4) Rigor, (5) Data Treatment, and (6) Assumptions.

- **Classical**
  - The classical approach imposes models (both **deterministic** and **probabilistic**) on the data.
  - Deterministic models include, for example, regression models and analysis of variance (ANOVA) models.
  - The most common probabilistic model assumes that the **errors** about the deterministic model are normally distributed--this assumption affects the validity of the ANOVA F tests.

- **Exploratory**
  - EDA does not impose deterministic or probabilistic models on the data.
  - EDA allows the data to suggest admissible models that best fit the data.

Exploratory vs Classical: Focus and Techniques

- Classical
  - the focus is on the model--**estimating parameters** of the model and generating **predicted values** from the model.
  - Classical techniques are generally **quantitative** in nature.
  - e.g., ANOVA, t tests, chi-squared tests, and F tests.

- Exploratory
  - the focus is on the data--its **structure**, outliers, and models suggested by the data.
  - EDA techniques are generally **graphical**.
  - e.g., scatter plots, character plots, box plots, histograms, bihistograms, probability plots, residual plots, and mean plots.

Exploratory vs Classical: Rigor

- Classical
  - Classical techniques serve as the **probabilistic foundation** of science and engineering;
  - the most important characteristic of classical techniques is that they are **rigorous**, **formal**, and "**objective**".

- Exploratory
  - EDA techniques do not share in that rigor or formality.
  - EDA techniques make up for that lack of rigor by being very **suggestive**, **indicative**, and insightful about what the appropriate model should be.
  - EDA techniques are **subjective** and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions.

Exploratory vs Classical: Data Treatment

- Classical
  - Classical estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers ("**estimates**").
  - These few numbers focus on **important characteristics** (location, variation, etc.) of the population.
  - Concentrating on these few characteristics can filter out other characteristics (skewness, tail length, autocorrelation, etc.) of the same population.
  - In this sense there is a loss of information due to this **"filtering" process**.

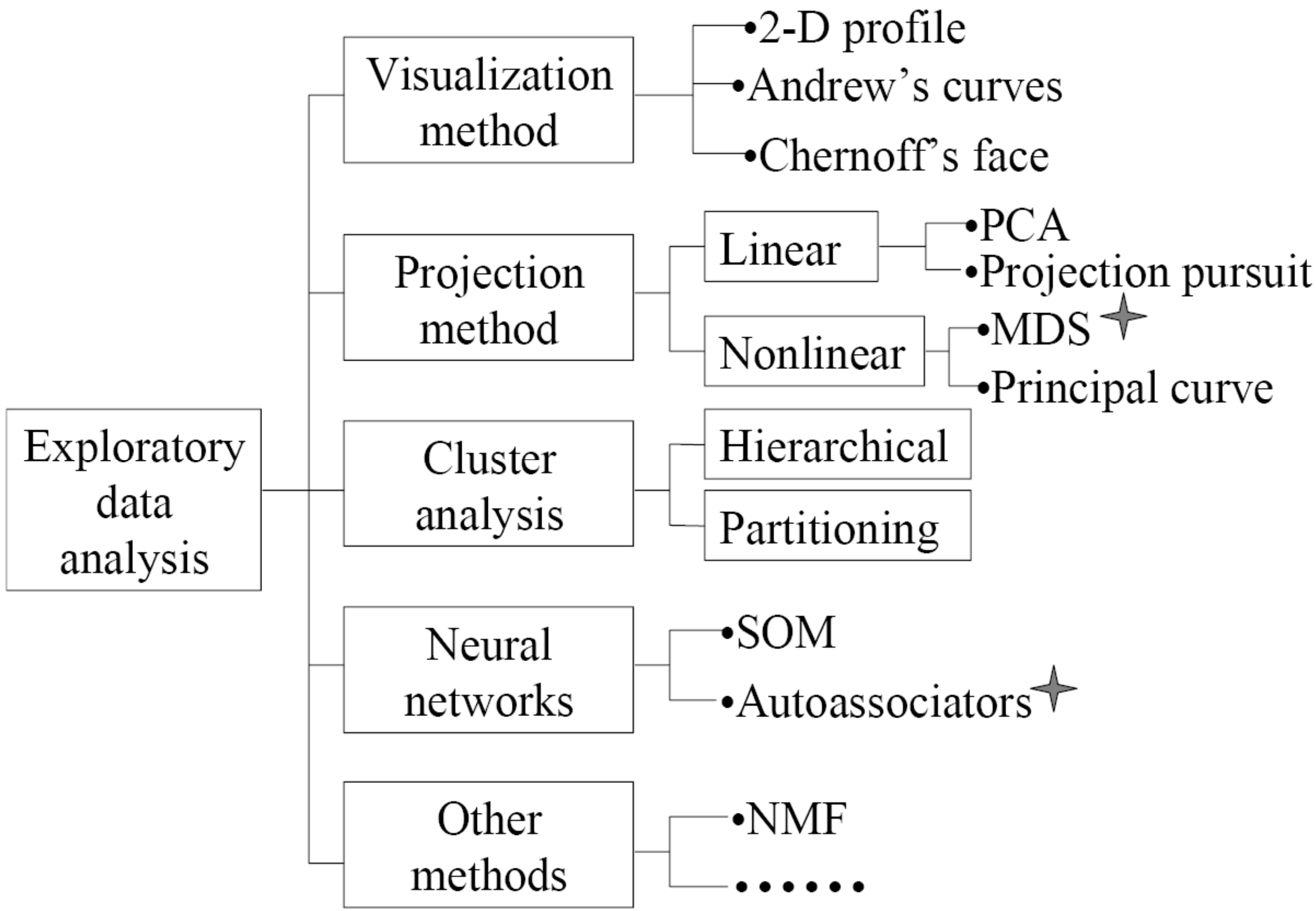
- Exploratory
  - EDA often makes use of (and shows) all of the available data.
  - In this sense there is no corresponding loss of information.

Exploratory vs Classical: Assumptions

- Classical
  - Classical tests depend on underlying **assumptions** (e.g., normality), and hence the validity of the test conclusions becomes dependent on the validity of the underlying assumptions.
  - The exact underlying assumptions may be **unknown** to the analyst, or if known, untested.
  - Thus the validity of the scientific conclusions becomes intrinsically linked to the validity of the underlying assumptions.
  - In practice, if such assumptions are unknown or untested, the validity of the scientific conclusions becomes suspect.

- Exploratory
  - Many EDA techniques make little or no assumptions--they present and show the data--all of the data.

Classification System for EDA techniques



Source: http://bioinformatics.sdstate.edu/users/gex/index/indexfiles/ch1.pdf

General Problem Categories (1/4)

- UNIVARIATE
  - Data: A single column of numbers, Y.
  - Model: $y = constant + error$
  - Output:
    - A number (the estimated constant in the model).
    - An estimate of uncertainty for the constant.
    - An estimate of the distribution for the error.
  - Techniques: 4-Plot (run sequence plot, lag plot, histogram, normal probability plot.), Probability Plot, PPCC Plot (Probability Plot Correlation Coefficient Plot)

- CONTROL
  - Data: A single column of numbers, Y.
  - Model: $y = constant + error$
  - Output: A "yes" or "no" to the question "Is the system out of control?".
  - Techniques: Control Charts.

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda17.htm

General Problem Categories (2/4)

- COMPARATIVE
  - Data: A single response variable and k independent variables $(Y, X1, X2, \ldots, Xk)$, primary focus is on one (the primary factor) of these independent variables.
  - Model: $y = f(x1, x2, \ldots, xk) + error$
  - Output: A "yes" or "no" to the question "Is the primary factor significant?".
  - Techniques: Block Plot, Scatter Plot, Box Plot
- SCREENING
  - Data: A single response variable and k independent variables $(Y, X1, X2, \ldots, Xk)$.
  - Model: $y = f(x1, x2, \ldots, xk) + error$
  - Output:
    - A ranked list (from most important to least important) of factors.
    - Best settings for the factors.
    - A good model/prediction equation relating Y to the factors.
  - Techniques: Block Plot, Probability Plot, Bihistogram

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda17.htm

General Problem Categories (3/4)

- OPTIMIZATION
  - Data: A single response variable and k independent variables *(Y, X1, X2, ... , Xk)*.
  - Model: $y = f(x1, x2, ..., xk) + error$
  - Output: Best settings for the factor variables.
  - Techniques: Block Plot, Least Squares Fitting, Contour Plot

- REGRESSION
  - Data: A single response variable and k independent variables *(Y, X1, X2, ... , Xk)*. The independent variables can be continuous.
  - Model: $y = f(x1, x2, ..., xk) + error$
  - Output: A good model/prediction equation relating Y to the factors.
  - Techniques: Least Squares Fitting, Scatter Plot, 6-Plot

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda17.htm

General Problem Categories (4/4)

- TIME SERIES
  - Data: A column of time dependent numbers, Y. In addition, time is an independent variable. The time variable can be either explicit or implied. If the data are not equi-spaced, the time variable should be explicitly provided.
  - Model: $yt = f(t) + error$,
    The model can be either a time domain based or frequency domain based.
  - Output: A good model/prediction equation relating Y to previous values of Y.
  - Techniques: Autocorrelation Plot, Spectrum, Complex Demodulation Amplitude Plot, Complex Demodulation Phase Plot, ARIMA Models.

- MULTIVARIATE
  - Data: k factor variables $(X1, X2, \ldots , Xk)$.
  - Model: The model is not explicit.
  - Output: Identify underlying correlation structure in the data.
  - Techniques: Star Plot, Scatter Plot Matrix, Conditioning Plot, Profile Plot, Principal Components, Clustering, Discrimination/Classification

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda17.htm

Why Data Visualization?

- It is not about "infographics", the beautiful, heavily customized products of expert graphic designers.
- Data visualization can provide clear understanding of patterns in data, detect hidden structures in data, condense information.
- **Anscombe's quartet** comprises four datasets. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
- Four datasets have nearly identical simple statistical properties, yet appear very different when graphed.

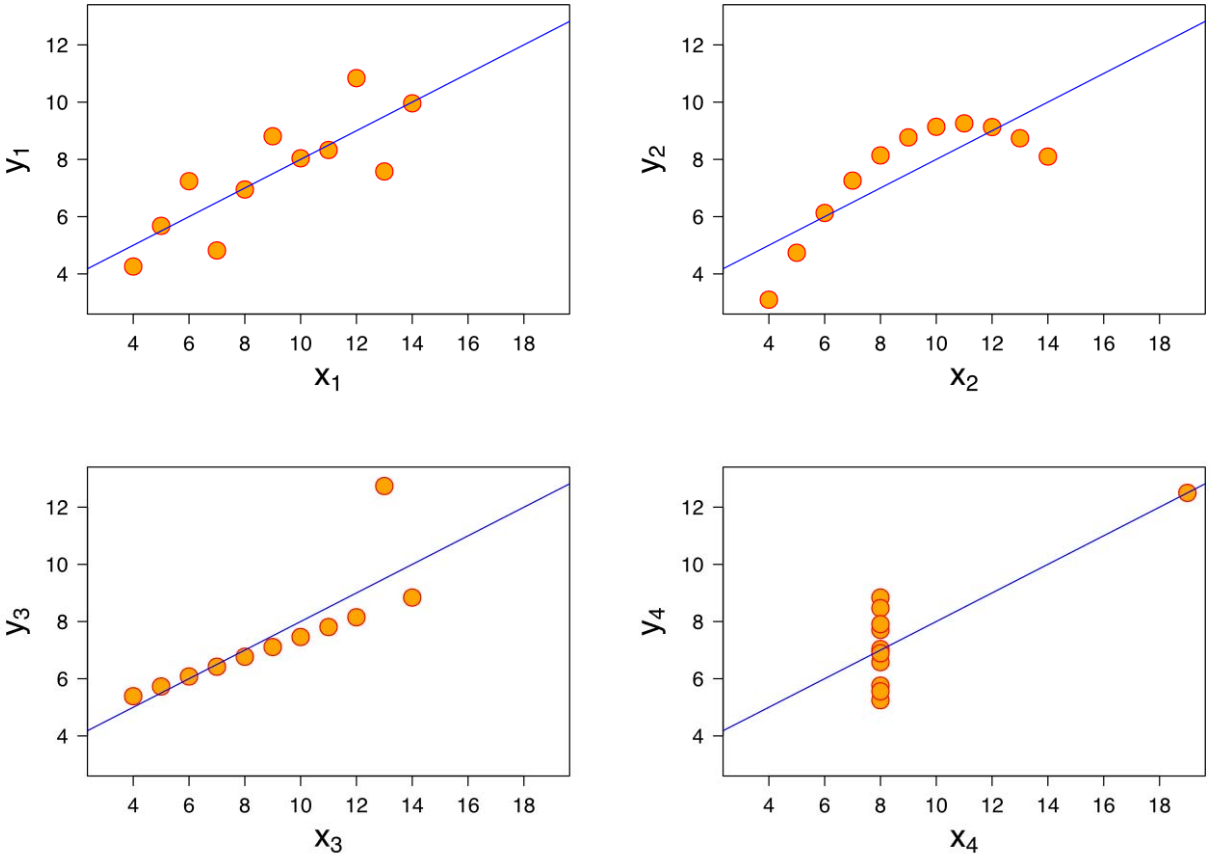
	I		II		III		IV	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

https://en.wikipedia.org/wiki/Anscombe's_quartet

http://ryanwomack.com/IASSIST/DataViz/

Anscombe's Quartet

- **Mean of x** in each case: 9 (exact)
- **Sample variance of x** in each case: 11 (exact)
- **Mean of y** in each case: 7.50 (to 2 decimal places)
- **Sample variance of y** in each case: 4.122 or 4.127 (to 3 decimal places)
- **Correlation** between x and y in each case: 0.816 (to 3 decimal places)
- **Linear regression line** in each case: y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)



EDA and Visualization

- **Data visualization** is the presentation of data in a pictorial or graphical format.
  - Any effort to help people understand the significance of data by placing it in a visual context.
  - Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

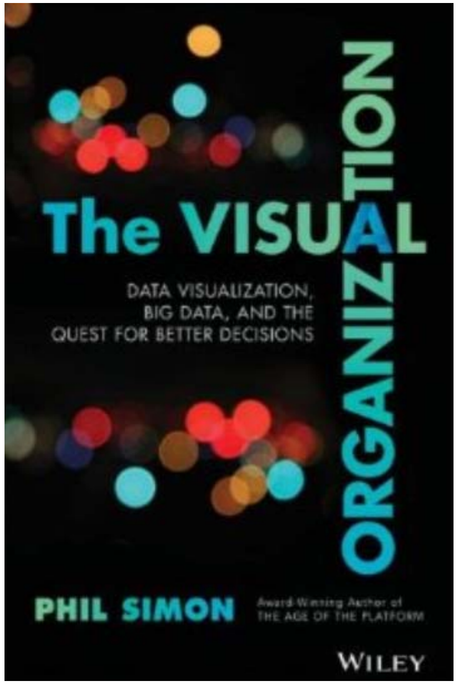
- **Get to know your data**: distributions (symmetric, normal, skewed), data quality problems, outliers, correlations and inter-relationships, subsets of interest, suggest functional relationships.

- **Visualizing data**: One variable, Two variables, More than two variables, Other types of data, Dimension reduction.

- **Interactive data visualization**: using computers and mobile devices to drill down into charts and graphs for more details, and interactively (and immediately) changing what data you see and how it is processed.

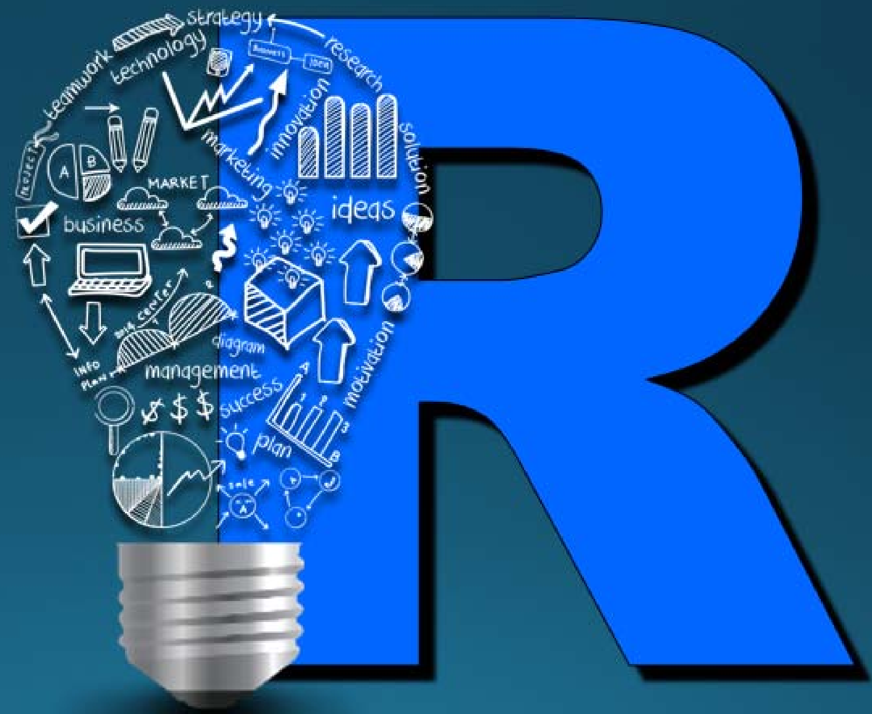
Some DataViz Sites

- Data Visualization Is The Future - Here's Why
http://www.forbes.com/sites/dorieclark/2014/03/10/data-visualization-is-the-future-heres-why/
- Phil Simon, 2014, The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions, Wiley. ISBN: 9781118794388 | 1118794389



- Information Aesthetics: http://infosthetics.com/
- Chart Porn: http://chartporn.org/
- Eagereyes: https://eagereyes.org/
- We Love Datavis: http://datavis.tumblr.com
- A New Generation Tool For (big) Data Visualization: http://www.stratio.com/datavis/kbase/
- Visualizing.org: http://www.visualizing.org/explore
- VizWiz: http://vizwiz.blogspot.ca/
- US Census Data Visualization Gallery: http://www.census.gov/dataviz/

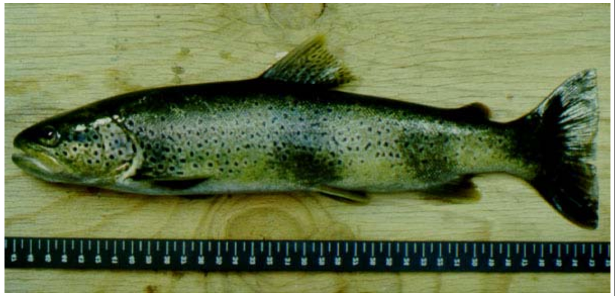
探索式資料分析
兩個EDA的例子



吳漢銘
國立政治大學 統計學系

https://hmwu.idv.tw

Example 1: The Doubs Fish Data



- **Fish communities** were good biological indicators of these water bodies: Verneaux (1973) (Verneaux et al. 2003) proposed to use **fish species** to characterize ecological zones along European rivers and streams. (River Doubs, 杜河)

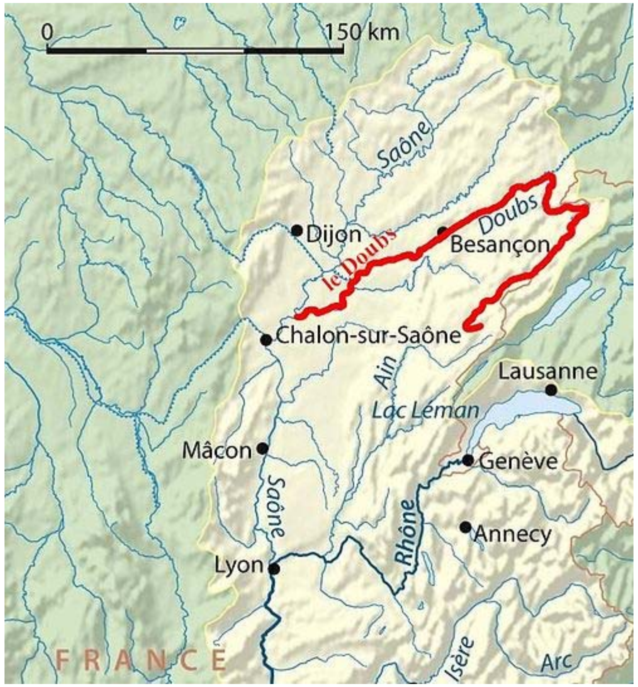
- Verneaux proposed a **typology** in four zones, and he named each one after a characteristic species:
  - the **trout (鱒魚，鮭鱒魚) zone** (from the brown trout Salmo trutta fario),
  - the **grayling (鱒魚) zone** (from Thymallus),
  - the **barbell (鲃, 有觸鬚的魚) zone** (from Barbus) and
  - the **bream (歐鯿, 鯉科淡水魚) zone** (from the common bream Abramis brama).

- The two upper zones are considered as the “**Salmonid (鮭魚) region**" and the two lowermost ones constitute the “**Cyprinid (鯉科之魚) region**”.

D. Borcard et al., Numerical Ecology with R, Use R, DOI 10.1007/978-1-4419-7976-6_2, © Springer Science+Business Media, LLC 2011

Image Source:
http://www.qub.ac.uk/bb-old/prodohl/TroutConcert/images/gallery/c_lagiader-me07-18-trout.jpg
http://www.bamboorods.ch/guiding/bilder/grayling2.jpg
https://en.wikipedia.org/wiki/Barbus_barbus#/media/File:Barbel.jpg
http://www.ultimateangling.co.za/index.php?topic=15775.0

River Doubs Map



背景知識、問題、資料收集方式、變數資訊、參與人角色、資料處理、探索(分析)方法、資料/過程/結果呈現。

Source: https://en.wikipedia.org/wiki/Doubs_%28river%29

The Doubs Fish Data: 檔案

- The Doubs data set have been collected at **30 sites** along the Doubs River (near the France–Switzerland border in the Jura Mountains.)

- The corresponding ecological conditions, with much variation among rivers, range from relatively pristine, well oxygenated and oligotrophic (湖泊沼地等水草植物不多、營養不足的) to eutrophic (營養正常的) and oxygen-deprived (貧困的) waters.

- **DoubsSpe**: contains coded abundances (豐富充足) of **27 fish species**.
- **DoubsEnv**: contains **11 environmental variables** related to the hydrology, geomorphology and chemistry of the river.
- **DoubsSpa**: contains the **geographical coordinates** (Cartesian, X and Y) of the sites.

DoubsSpe.csv

```
,   CHA, TRU, VAI, LOC, OMB, BLA, HO'
1,  0,   3,   0,   0,   0,   0,   0,
2,  0,   5,   4,   3,   0,   0,   0,
3,  0,   5,   5,   5,   0,   0,   0,
4,  0,   4,   5,   5,   0,   0,   0,
5,  0,   2,   3,   2,   0,   0,   0,
6,  0,   3,   4,   5,   0,   0,   0,
7,  0,   5,   4,   5,   0,   0,   0,
8,  0,   0,   0,   0,   0,   0,   0,
9,  0,   0,   1,   3,   0,   0,   0,
10, 0,   1,   4,   4,   0,   0,   0,
11, 1,   3,   4,   1,   1,   0,   0,
```

DoubsEnv.csv

```
,   das,   alt, pen, deb,  pH,  dur,
1,  0.3,   934, 48,  0.84, 7.9, 45,
2,  2.2,   932, 3,   1,    8,   40,
3,  10.2,  914, 3.7, 1.8,  8.3, 52,
4,  18.5,  854, 3.2, 2.53, 8,   72,
5,  21.5,  849, 2.3, 2.64, 8.1, 84,
6,  32.4,  846, 3.2, 2.86, 7.9, 60,
7,  36.8,  841, 6.6, 4,    8.1, 88,
8,  49.1,  792, 2.5, 1.3,  8.1, 94,
9,  70.5,  752, 1.2, 4.8,  8,   90,
10, 99,    617, 9.9, 10,   7.7, 82,
11, 123.4, 483, 4.1, 19.9, 8.1, 96,
```

DoubsSpa.csv

```
,   x,   y
1,  88,  7
2,  94,  14
3,  102, 18
4,  100, 28
5,  106, 39
6,  112, 51
7,  114, 61
8,  110, 76
9,  136, 100
10, 168, 112
11, 186, 130
12, 205, 145
```

The Doubs Fish Data: 前置處理

- Working with the environmental data available in the R package **ade4** (version 1.4-14), we corrected a mistake in the das variable and restored the variables to their original units (Table 1.1.)

- Verneaux used a semi-quantitative, species-specific, abundance scale (0–5) so that comparisons between species abundances make sense. (However, species-specific codes cannot be understood as unbiased estimates of the true abundances (number or density of individuals) or biomasses at the sites.)

Table 1.1 Environmental variables of the Doubs data set used in this book and their units

Variable	Code	Units
Distance from source	das	km
Altitude	alt	m a.s.l.
Slope	pen	‰
Mean minimum discharge	deb	$m^3 s^{-1}$
pH of water	pH	–
Calcium concentration (hardness)	dur	$mg L^{-1}$
Phosphate concentration	pho	$mg L^{-1}$
Nitrate concentration	nit	$mg L^{-1}$
Ammonium concentration	amm	$mg L^{-1}$
Dissolved oxygen	oxy	$mg L^{-1}$
Biological oxygen demand	dbo	$mg L^{-1}$

Data Extraction: Read Data

- 每一檔案之大小、資料維度、關聯。
- (報告中)列出每一變數之
  - 名稱、所代表意義。
  - 型態(連續、類別、順序、時間等等)、單位
  - 編碼、範圍(五數摘要)、遺失值比例(分佈)。
- 若是類別變數，則列出每一類別之次數分佈。交叉次數表。

```
> # Load the required package, vegan: Community Ecology Package
> library(vegan)

> # Load additionnal functions
> # (files must be in the working directory)
> source("panelutils.R")

> # Import the data from CSV files
> # Species (community) data frame (fish abundances)
> spe <- read.csv("DoubsSpe.csv", row.names=1)
> # Environmental data frame
> env <- read.csv("DoubsEnv.csv", row.names=1)
> # Spatial data frame
> spa <- read.csv("DoubsSpa.csv", row.names=1)
```

```
> library(ade4)
> data(doubs)
> ?doubs
```

Source: Borcard D., Gillet F. & Legendre P. Numerical Ecology with R, Springer, 2011

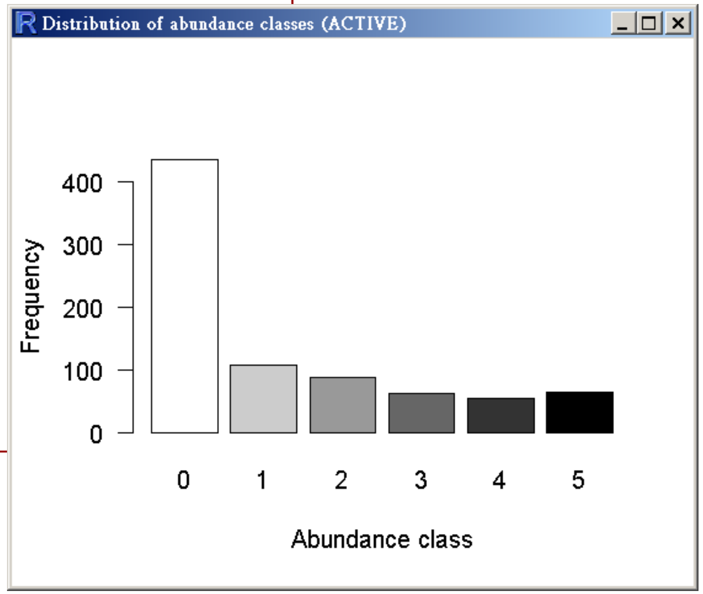
Species Data: First Contact
Basic functions

```
> spe   # Display the whole data frame in the console
   CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT
1    0   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
...
> spe[1:5,1:10]     # Display only 5 lines and 10 columns
  CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE
1   0   3   0   0   0   0   0   0   0   0
...
> head(spe)     # Display only the first few lines
  CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR
1   0   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
...
> nrow(spe)     # Number of rows (sites)
[1] 30
> ncol(spe)     # Number of columns (species)
[1] 27
> dim(spe)     # Dimensions of the data frame (rows, columns)
[1] 30 27
> colnames(spe)     # Column labels (descriptors = species)
 [1] "CHA" "TRU" "VAI" "LOC" "OMB" "BLA" "HOT" "TOX" "VAN" "CHE" "BAR" "SPI"
...
> rownames(spe)     # Row labels (objects = sites)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
...
> summary(spe)     # Descriptive statistics for columns
      CHA            TRU            VAI              LOC              OMB
 Min.   :0.00   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.00
 1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.00
 Median :0.00   Median :1.00   Median :3.000   Median :2.000   Median :0.00
 Mean   :0.50   Mean   :1.90   Mean   :2.267   Mean   :2.433   Mean   :0.50
 3rd Qu.:0.75   3rd Qu.:3.75   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:0.75
 Max.   :3.00   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :4.00
  ...
```

Overall Distribution of Abundances (Dominance Codes)

Compare median and mean abundances. Are most distributions symmetrical?

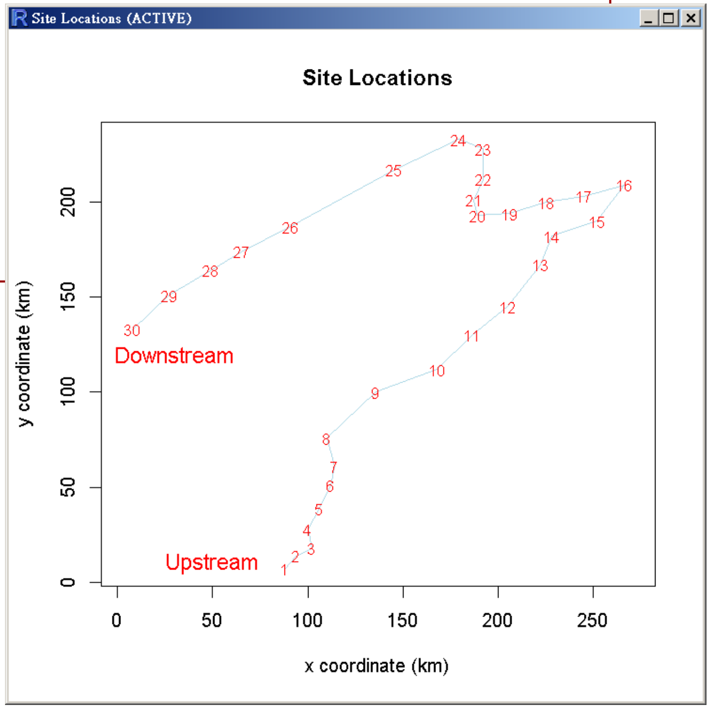
```
> # Minimum and maximum of abundance values in the whole data set
> range(spe)
[1] 0 5
> # Count cases for each abundance class
> (ab <- table(unlist(spe)))
  0   1   2   3   4   5
435 108  87  62  54  64
> # Create a graphic window with title
> windows(title="Distribution of abundance classes")
>
> # Barplot of the distribution, all species confounded
> barplot(ab, las=1, xlab="Abundance class",
+  ylab="Frequency", col=gray(5:0/5))
> # Number of absences
> sum(spe==0)
[1] 435
> # Proportion of zeros in the community data set
> sum(spe==0)/(nrow(spe)*ncol(spe))
[1] 0.537037
```



How do you interpret the high frequency of zeros (absences) in the data frame?

Species Data: A Closer Look
Map of the Locations of the Sites

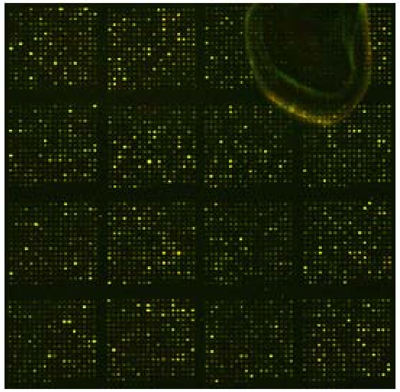
```
> windows(title="Site Locations")
> # Create an empty frame (proportional axes 1:1, with titles)
> # Geographic coordinates x and y from the spa data frame
> plot(spa, asp=1, type="n", main="Site Locations",
+ xlab="x coordinate (km)", ylab="y coordinate (km)")
> # Add a blue line connecting the sites (Doubs river)
> lines(spa, col="light blue")
> # Add site labels
> text(spa, row.names(spa), cex=0.8, col="red")
> # Add text blocks
> text(50, 10, "Upstream", cex=1.2, col="red")
> text(30, 120, "Downstream", cex=1.2, col="red")
```

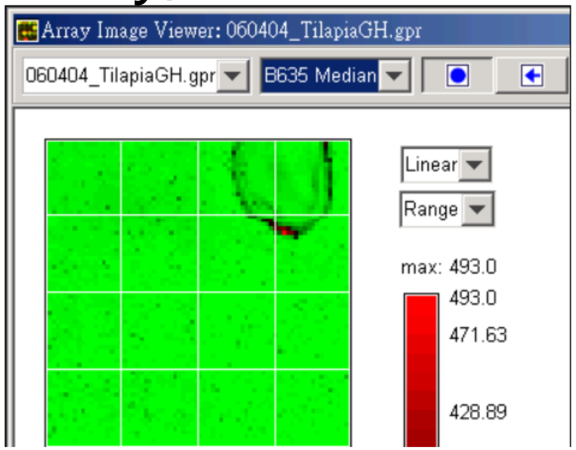


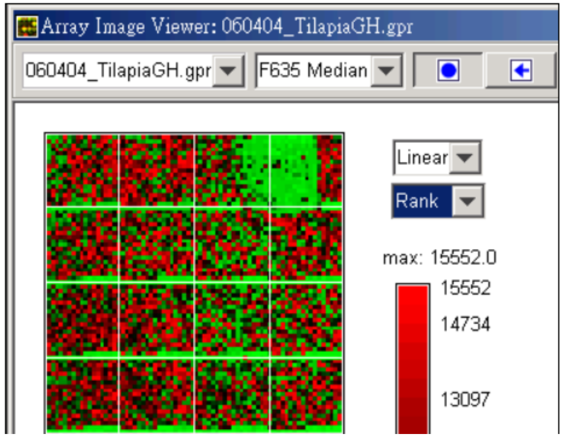
The river looks more real, but where are the fish?

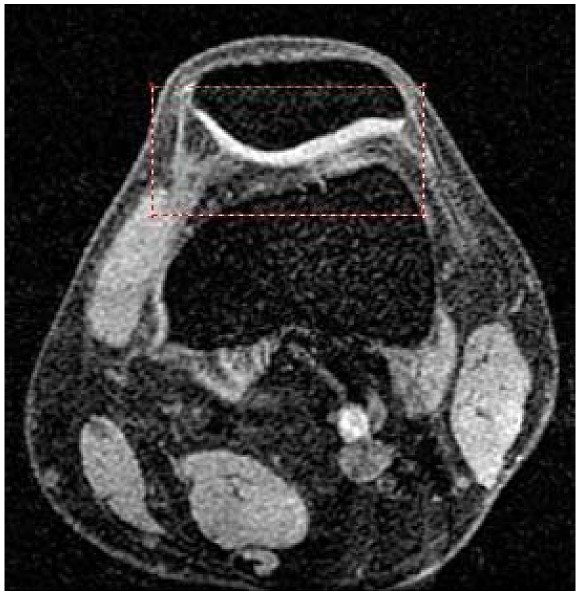
註: 重建 Reconstruction

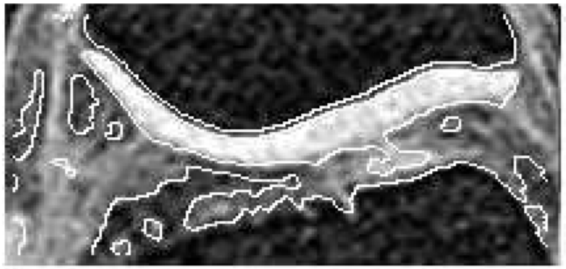
生物晶片 (Microarray)

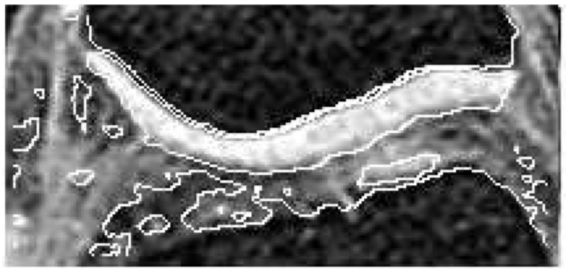




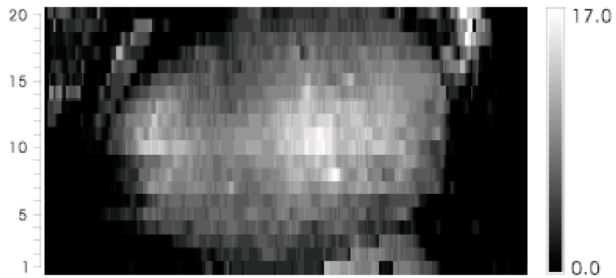








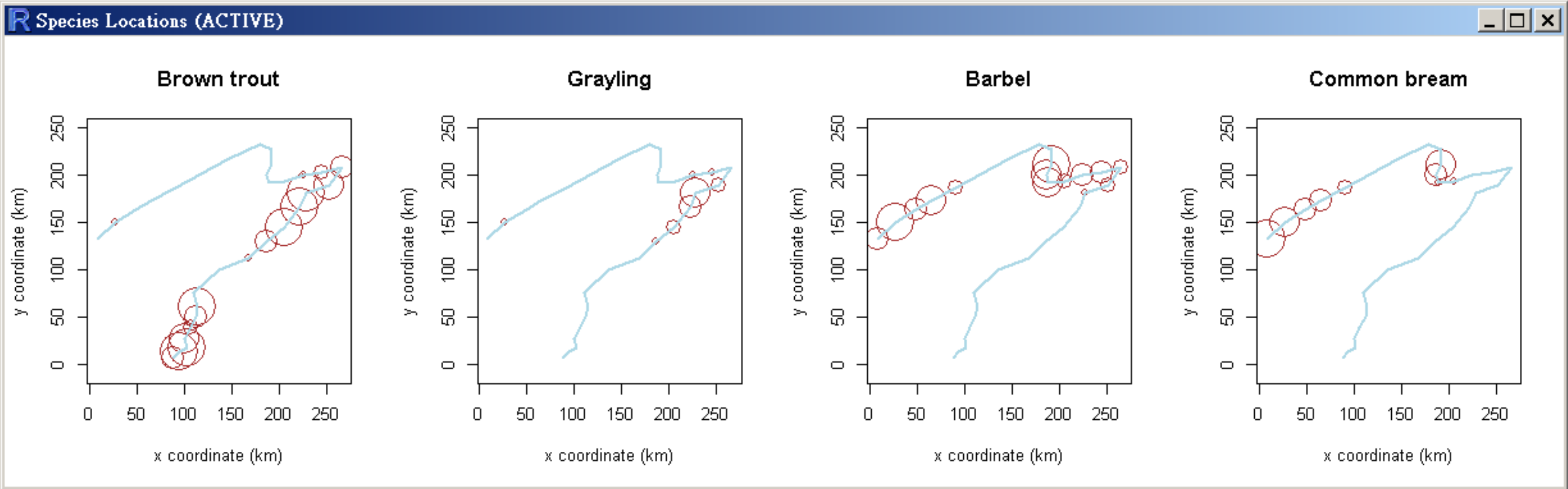
醫學影像 (fMRI)



Maps of Some Fish Species

```
> # New graphic window (size 9x9 inches)
> windows(title="Species Locations", 9, 9)
> par(mfrow=c(1,4))
> # Plot four species
> xl <- "x coordinate (km)",
> yl <- "y coordinate (km)"
> plot(spa, asp=1, col="brown", cex=spe$TRU, main="Brown trout", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$OMB, main="Grayling", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$BAR, main="Barbel", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$BCO, main="Common bream", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
```

From these graphs you should understand why these four species were chose as ecological indicators.



Bubble maps of the abundance of four fish species

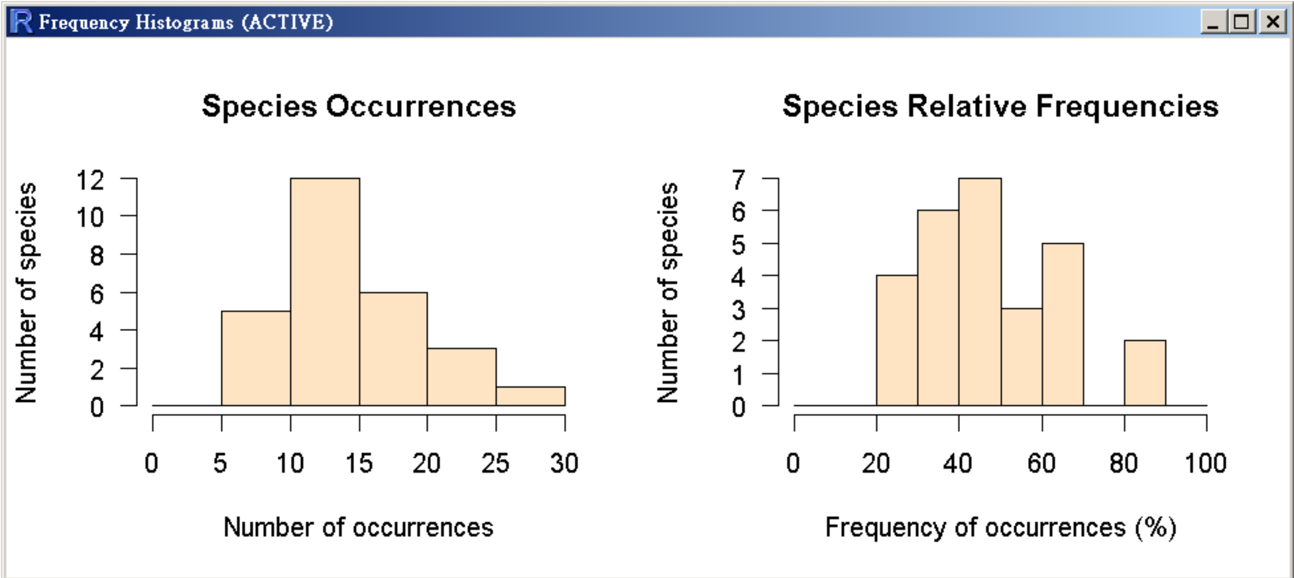
Compare Species: Number of Occurrences

At how many sites does each species occur? Calculate the relative frequencies of species (proportion of the number of sites) and plot histograms.

```
> # Compute the number of sites where each species is present
> # To sum by columns, the second argument of apply(), MARGIN, is set to 2
> spe.pres <- apply(spe > 0, 2, sum)
> # Sort the results in increasing order
> sort(spe.pres)
PCH CHA OMB BLA BCO BBO TOX BOU ROT ANG HOT SPI CAR GRE PSO BAR ABL PER TRU TAN
  7   8   8   8   9  10  11  11  11  11  12  12  12  12  13  14  14  15  17  17
VAN BRO GAR VAI GOU LOC CHE
 18  18  18  20  20  24  25
> # Compute percentage frequencies
> spe.relf <- 100*spe.pres/nrow(spe)
> # Round the sorted output to 1 digit
> round(sort(spe.relf), 1)
 PCH  CHA  OMB  BLA  BCO  BBO  TOX  BOU  ROT  ANG  HOT  SPI  CAR  GRE  PSO  BAR
23.3 26.7 26.7 26.7 30.0 33.3 36.7 36.7 36.7 36.7 40.0 40.0 40.0 40.0 43.3 46.7
 ABL  PER  TRU  TAN  VAN  BRO  GAR  VAI  GOU  LOC  CHE
46.7 50.0 56.7 56.7 60.0 60.0 60.0 66.7 66.7 80.0 83.3
```

Compare Species: Number of Occurrences

```
> # Plot the histograms
> windows(title="Frequency Histograms",8,5)
> # Divide the window horizontally
> par(mfrow=c(1,2))
> hist(spe.pres, main="Species Occurrences", right=FALSE, las=1,
+ xlab="Number of occurrences", ylab="Number of species",
+ breaks=seq(0,30,by=5), col="bisque")
> hist(spe.relf, main="Species Relative Frequencies", right=FALSE,
+ las=1, xlab="Frequency of occurrences (%)", ylab="Number of species",
+ breaks=seq(0, 100, by=10), col="bisque")
```



Compare Sites: Species Richness

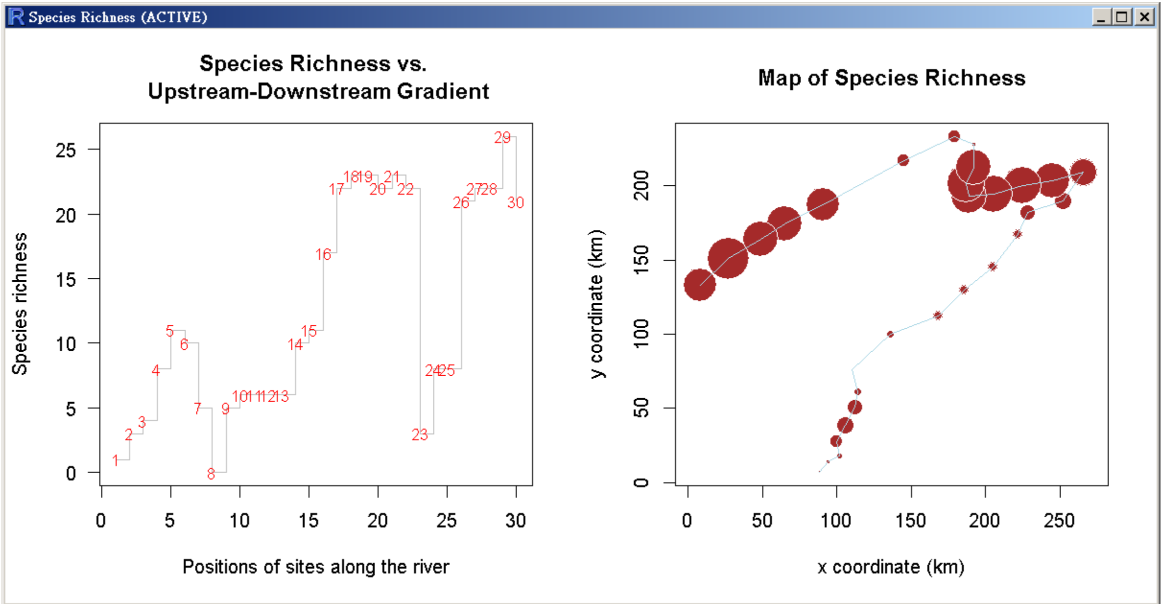
Now that we have seen at how many sites each species is present, we may want to know how many species are present at each site (species richness).

```
> # Compute the number of species at each site
> # To sum by rows, the second argument of apply(), MARGIN, is set to 1
> sit.pres <- apply(spe > 0, 1, sum)
> # Sort the results in increasing order
> sort(sit.pres)
 8  1  2 23  3  7  9 10 11 12 13  4 24 25  6 14  5 15 16 26 30 17 20 22 27 28 18 19
 0  1  3  3  4  5  5  6  6  6  6  8  8  8 10 10 11 11 17 21 21 22 22 22 22 22 23 23
21 29
23 26
```

Compare Sites: Species Richness

```
> windows(title="Species Richness", 10, 5)
> par(mfrow=c(1,2))
> # Plot species richness vs. position of the sites along the river
> plot(sit.pres,type="s", las=1, col="gray",
+ main="Species Richness vs. \n Upstream-Downstream Gradient",
+ xlab="Positions of sites along the river", ylab="Species richness")
> text(sit.pres, row.names(spe), cex=.8, col="red")
> # Use geographic coordinates to plot a bubble map
> plot(spa, asp=1, main="Map of Species Richness", pch=21, col="white",
+ bg="brown", cex=5*sit.pres/max(sit.pres), xlab="x coordinate (km)",
+ ylab="y coordinate (km)")
> lines(spa, col="light blue")
```

Can you identify richness hot spots along the river?



Compute Alpha Diversity Indices of the Fish Communities

Finally, one can easily compute classical diversity indices from the data. Let us do it with the function **diversity()** of the **vegan** package.

diversity {vegan} R Documentation

Ecological Diversity Indices and Rarefaction Species Richness

Description

Shannon, Simpson, and Fisher diversity indices and rarefied species richness for community ecologists.

Usage

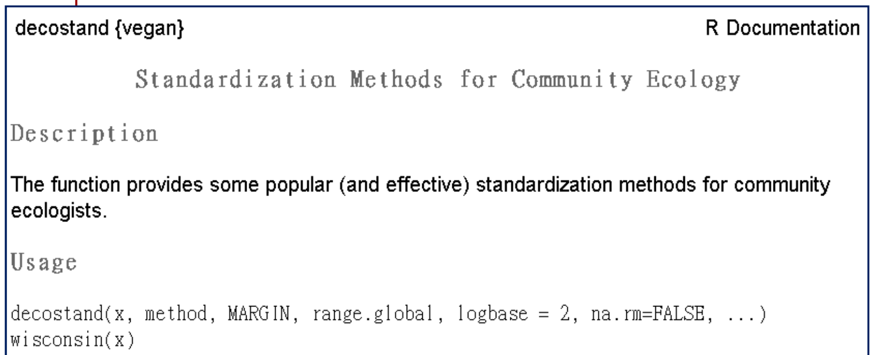
```
diversity(x, index = "shannon", MARGIN = 1, base = exp(1))
```

```
> # Get help on the diversity() function
> ?diversity
>
> N0 <- rowSums(spe > 0)         # Species richness
> H <- diversity(spe)            # Shannon entropy
> N1 <- exp(H)                   # Shannon diversity (number of abundant species)
> N2 <- diversity(spe, "inv")    # Simpson diversity (number of dominant species)
> J <- H/log(N0)                 # Pielou evenness
> E10 <- N1/N0                   # Shannon evenness (Hill's ratio)
> E20 <- N2/N0                   # Simpson evenness (Hill's ratio)
> (div <- data.frame(N0, H, N1, N2, E10, E20, J))
   N0        H        N1       N2       E10       E20         J
1   1 0.000000  1.000000 1.000000 1.0000000 1.0000000       NaN
2   3 1.077556  2.937493 2.880000 0.9791642 0.9600000 0.9808340
3   4 1.263741  3.538634 3.368421 0.8846584 0.8421053 0.9115962
4   8 1.882039  6.566883 5.727273 0.8208604 0.7159091 0.9050696
5  11 2.329070 10.268387 9.633333 0.9334897 0.8757576 0.9712976
6  10 2.108294  8.234184 7.000000 0.8234184 0.7000000 0.9156205
...
```

Transformation and Standardization of the Species Data

- The **decostand()** function of the **vegan** package provides many options for common standardization of ecological data.
- In this function, standardization, as contrasted with simple transformation (such as square root, log or presence–absence), means that the values are not transformed individually but relative to other values in the data table.
- Standardization can be done relative to sites (site profiles), species (species profiles), or both (double profiles), depending on the focus of the analysis.

```
> # Get help on the decostand() function
> ?decostand
> ## Simple transformations
> # Partial view of the raw data (abundance codes)
> spe[1:5, 2:4]
  TRU VAI LOC
1   3   0   0
...
> # Transform abundances to presence-absence (1-0)
> spe.pa <- decostand(spe, method="pa")
> spe.pa[1:5, 2:4]
  TRU VAI LOC
1   1   0   0
...
```



decostand {vegan} R Documentation

Standardization Methods for Community Ecology

Description

The function provides some popular (and effective) standardization methods for community ecologists.

Usage

```
decostand(x, method, MARGIN, range.global, logbase = 2, na.rm=FALSE, ...)
wisconsin(x)
```

Transformation and Standardization of the Species Data

```
> Species profiles: 2 methods: presence-absence or abundance data
> ## Species profiles: standardization by column
> # Scale abundances by dividing them by the maximum value for each species
> # Note: MARGIN=2 (column, default value) for this method
> spe.scal <- decostand(spe, "max")
> spe.scal[1:5,2:4]
  TRU VAI LOC
1 0.6 0.0 0.0
...
> # Display the maximum by column
> apply(spe.scal, 2, max)
CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR TAN
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
BCO PCH GRE GAR BBO ABL ANG
  1   1   1   1   1   1   1
> # Scale abundances by dividing them by the species totals
> # (relative abundance by species)
> # Note: MARGIN=2 for this method
> spe.relsp <- decostand(spe, "total", MARGIN=2)
> spe.relsp[1:5,2:4]
         TRU        VAI        LOC
1 0.05263158 0.00000000 0.00000000
...
> # Display the sum by column
> apply(spe.relsp, 2, sum)
CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR TAN BCO
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
PCH GRE GAR BBO ABL ANG
  1   1   1   1   1   1
```

Did the scaling work properly? Keep an eye on the results by a plot or by the use of summary statistics

Scale Abundances by Dividing Them by the Site Totals

```
> ## Site profiles: 3 methods; presence-absence or abundance data
> ## standardization by row
> # Scale abundances by dividing them by the site totals
> # (relative abundance, or relative frequencies, per site)
> # (relative abundance by site)
> # Note: MARGIN=1 (default value) for this method
> spe.rel <- decostand(spe, "total")
> spe.rel[1:5,2:4]
         TRU        VAI        LOC
1 1.00000000 0.00000000 0.00000000
...
> # Display the sum of row vectors to determine if the scaling worked properly
> apply(spe.rel, 1, sum)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1 
29 30 
 1  1 
> # Give a length of 1 to each row vector (Euclidean norm)
> spe.norm <- decostand(spe, "normalize")
> spe.norm[1:5,2:4]
        TRU       VAI       LOC
1 1.0000000 0.0000000 0.0000000
...
> # Verify the norm of row vectors
> norm <- function(x) sqrt(x%*%x)
> apply(spe.norm, 1, norm)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1 
29 30 
 1  1 
```

The chord transformation: the Euclidean distance function applied to chord-transformed data produces a chord distance matrix. Useful before PCA and K-means.

Compute Relative Frequencies by Rows (Site Profiles)

- The Hellinger transformation can be also be obtained by applying the chord transformation to square-root-transformed species data.

```
> # Compute relative frequencies by rows (site profiles), then square root
> # Compute square root of relative abundances by site
> spe.hel <- decostand(spe, "hellinger")
> spe.hel[1:5,2:4]
        TRU       VAI       LOC
1 1.0000000 0.0000000 0.0000000
2 0.6454972 0.5773503 0.5000000
3 0.5590170 0.5590170 0.5590170
4 0.4364358 0.4879500 0.4879500
5 0.2425356 0.2970443 0.2425356
> # Check the norm of row vectors
> apply(spe.hel, 1, norm)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1 
29 30 
 1  1 
```

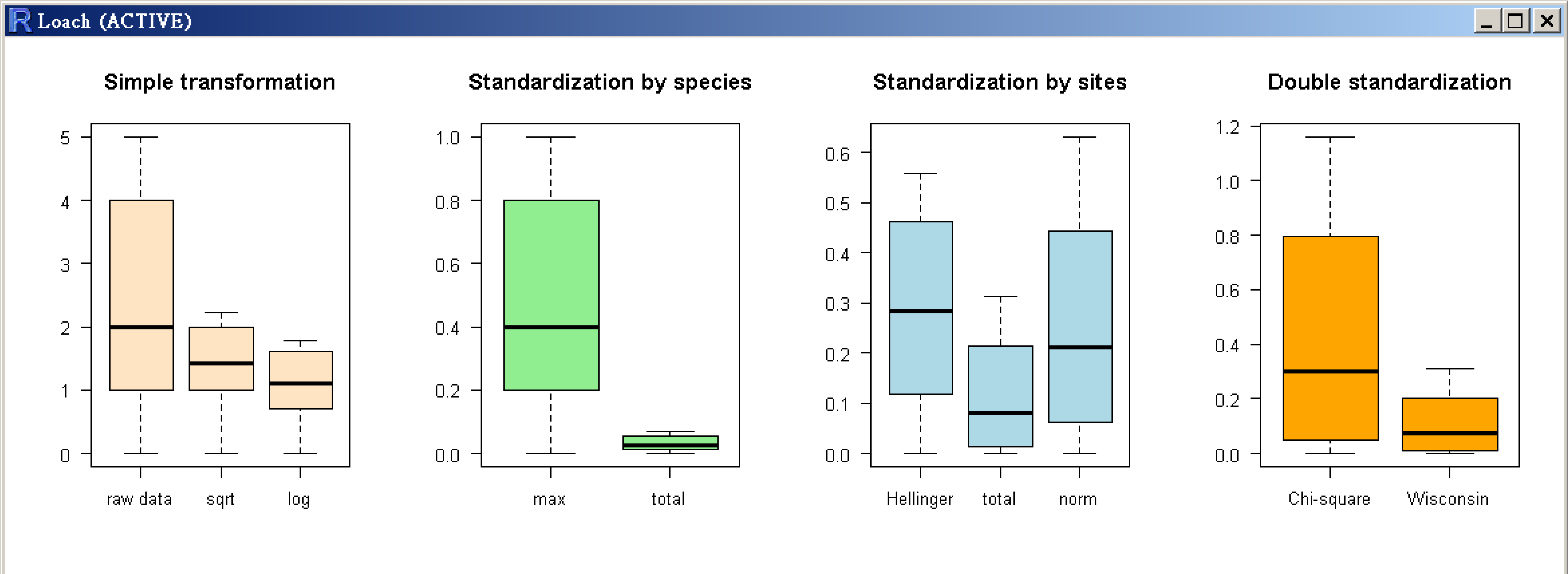
http://artax.karlin.mff.cuni.cz/r-help/library/analogue/html/tran.html

Standardization by Both Columns and Rows

```
> # Chi-square transformation
> spe.chi <- decostand(spe, "chi.square")
> spe.chi[1:5,2:4]
        TRU       VAI       LOC
1 4.1969078 0.0000000 0.0000000
2 1.7487116 1.2808290 0.9271402
3 1.3115337 1.2007772 1.1589253
4 0.7994110 0.9148778 0.8829907
5 0.2468769 0.3390430 0.2181506
> # Check what happened to site 8 where no species was found
> spe.chi[7:9,]
  CHA      TRU       VAI       LOC OMB BLA HOT TOX      VAN       CHE BAR SPI GOU BRO
7   0 1.311534 0.9606217 1.1589253   0   0   0   0 0.302004 0.2646384   0   0   0   0
8   0 0.000000 0.0000000 0.0000000   0   0   0   0 0.000000 0.0000000   0   0   0   0
9   0 0.000000 0.2744634 0.7946916   0   0   0   0 0.000000 1.5122194   0   0   0   0
  PER BOU PSO ROT CAR       TAN BCO PCH GRE      GAR BBO ABL ANG
7   0   0   0   0   0 0.0000000   0   0   0 0.000000   0   0   0
8   0   0   0   0   0 0.0000000   0   0   0 0.000000   0   0   0
9   0   0   0   0   0 0.3373903   0   0   0 1.140587   0   0   0
> # Wisconsin standardization
> # Abundances are first ranged by species maxima and then by site totals
> spe.wis <- wisconsin(spe)
> spe.wis[1:5,2:4]
         TRU        VAI        LOC
1 1.00000000 0.00000000 0.00000000
2 0.41666667 0.33333333 0.25000000
3 0.31250000 0.31250000 0.31250000
4 0.19047619 0.23809524 0.23809524
5 0.05882353 0.08823529 0.05882353
```

Boxplots of Transformed Abundances of a Common Species (Stone Loach)

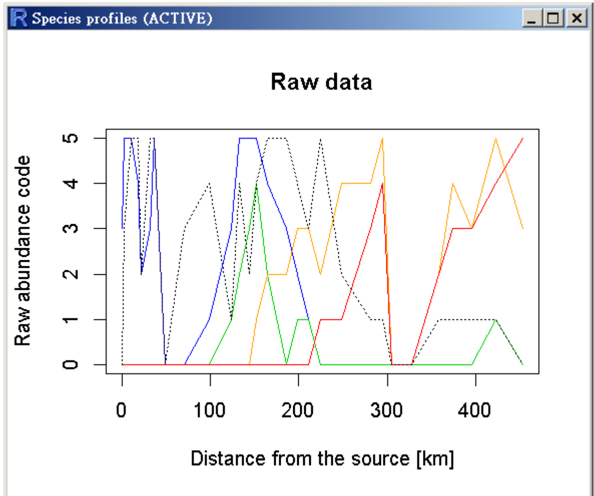
```
> windows(title="Loach")
> par(mfrow=c(1,4))
> boxplot(spe$LOC, sqrt(spe$LOC), log1p(spe$LOC), las=1, main="Simple transformation",
+ names=c("raw data", "sqrt", "log"), col="bisque")
> boxplot(spe.scal$LOC, spe.relsp$LOC, las=1, main="Standardization by species",
+ names=c("max", "total"), col="lightgreen")
> boxplot(spe.hel$LOC, spe.rel$LOC, spe.norm$LOC, las=1, main="Standardization by sites",
+ names=c("Hellinger", "total", "norm"), col="lightblue")
> boxplot(spe.chi$LOC, spe.wis$LOC, las=1, main="Double standardization",
+ names=c("Chi-square", "Wisconsin"), col="orange")
```

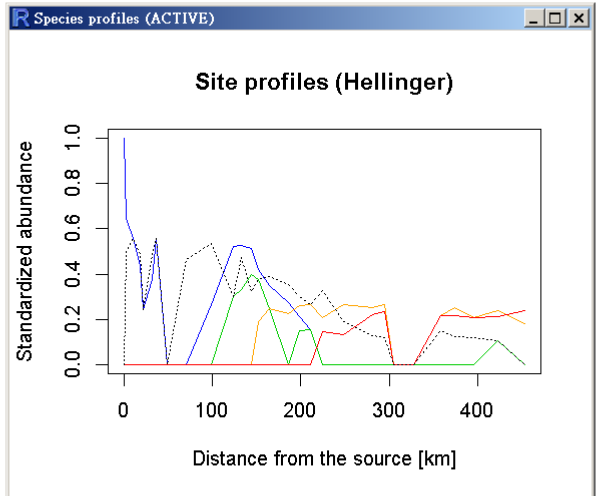


Boxplots of transformed abundances of a common species, Nemacheilus barbatulus (stone loach)

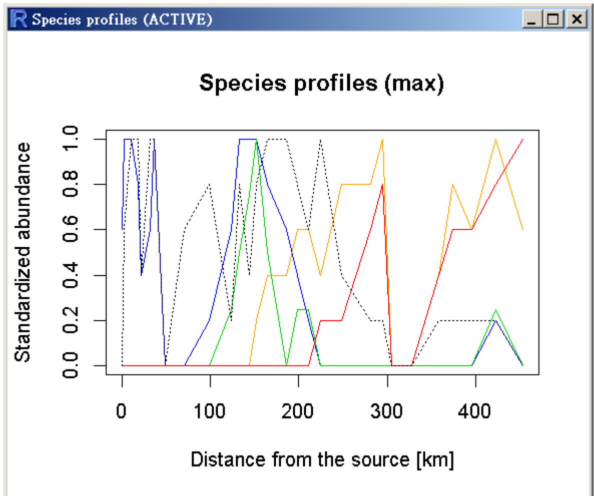
Plot Profiles Along the Upstream-Downstream Gradient

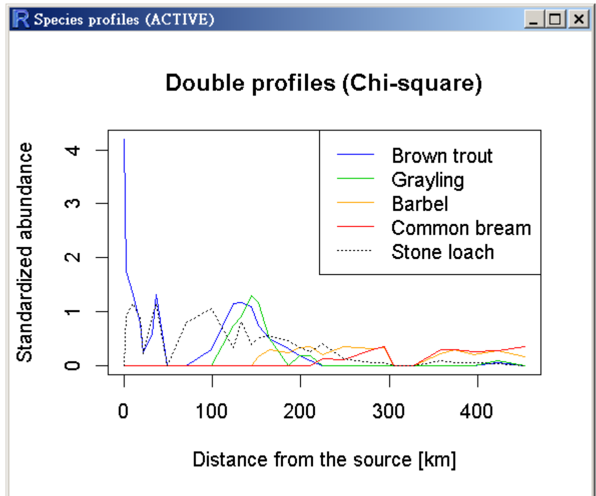
Another way to compare the effects of transformations on species profiles is to plot them along the river course.





Compare the profiles and explain the differences.





Plot Profiles Along the Upstream-Downstream Gradient

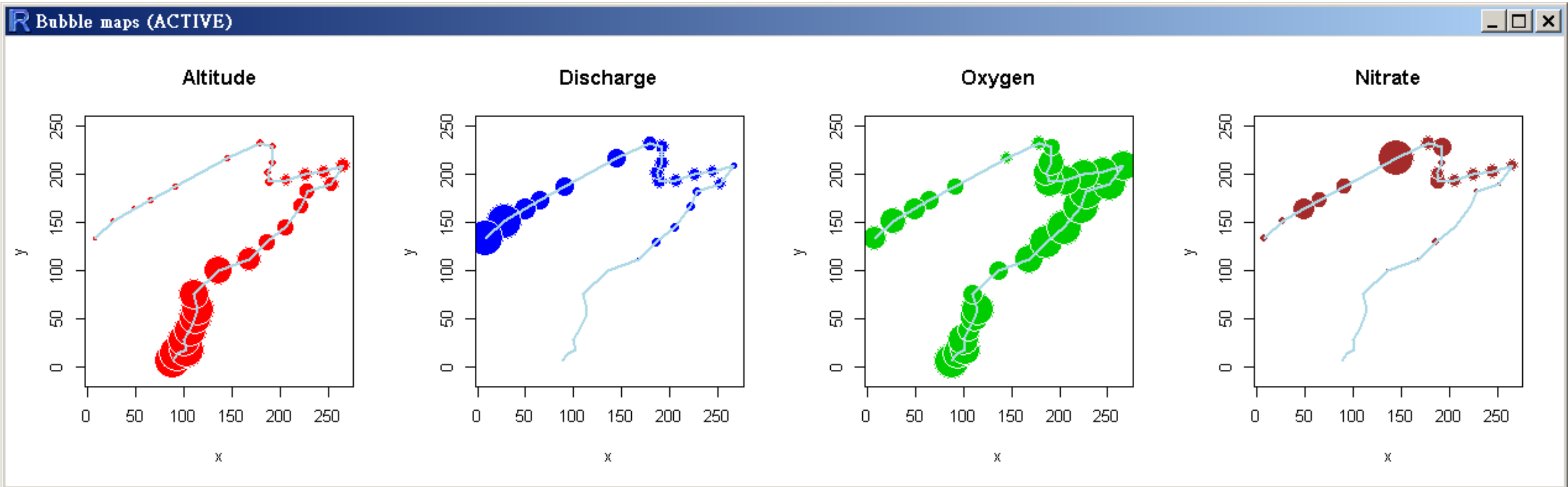
```
> windows(title="Species profiles", 9, 9)
> plot(env$das, spe$TRU, type="l", col=4, main="Raw data",
+ xlab="Distance from the source [km]", ylab="Raw abundance code")
> lines(env$das, spe$OMB, col=3); lines(env$das, spe$BAR, col="orange")
> lines(env$das, spe$BCO, col=2); lines(env$das, spe$LOC, col=1, lty="dotted")
>
> plot(env$das, spe.scal$TRU, type="l", col=4, main="Species profiles (max)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.scal$OMB, col=3); lines(env$das, spe.scal$BAR, col="orange")
> lines(env$das, spe.scal$BCO, col=2); lines(env$das, spe.scal$LOC, col=1, lty="dotted")

> plot(env$das, spe.hel$TRU, type="l", col=4, main="Site profiles (Hellinger)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.hel$OMB, col=3); lines(env$das, spe.hel$BAR, col="orange")
> lines(env$das, spe.hel$BCO, col=2); lines(env$das, spe.hel$LOC, col=1, lty="dotted")
>
> plot(env$das, spe.chi$TRU, type="l", col=4, main="Double profiles (Chi-square)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.chi$OMB, col=3); lines(env$das, spe.chi$BAR, col="orange")
> lines(env$das, spe.chi$BCO, col=2); lines(env$das, spe.chi$LOC, col=1, lty="dotted")
> legend("topright", c("Brown trout", "Grayling", "Barbel", "Common bream", "Stone loach"),
+ col=c(4,3,"orange",2,1), lty=c(rep(1,4),3))
```

Bubble Maps of Some Environmental Variables

```
> windows(title="Bubble maps", 9, 9)
> par(mfrow=c(1,4))
> plot(spa, asp=1, main="Altitude", pch=21, col="white",
+ bg="red", cex=5*env$alt/max(env$alt), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Discharge", pch=21, col="white",
+ bg="blue", cex=5*env$deb/max(env$deb), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Oxygen", pch=21, col="white",
+ bg="green3", cex=5*env$oxy/max(env$oxy), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Nitrate", pch=21, col="white",
+ bg="brown", cex=5*env$nit/max(env$nit), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
```

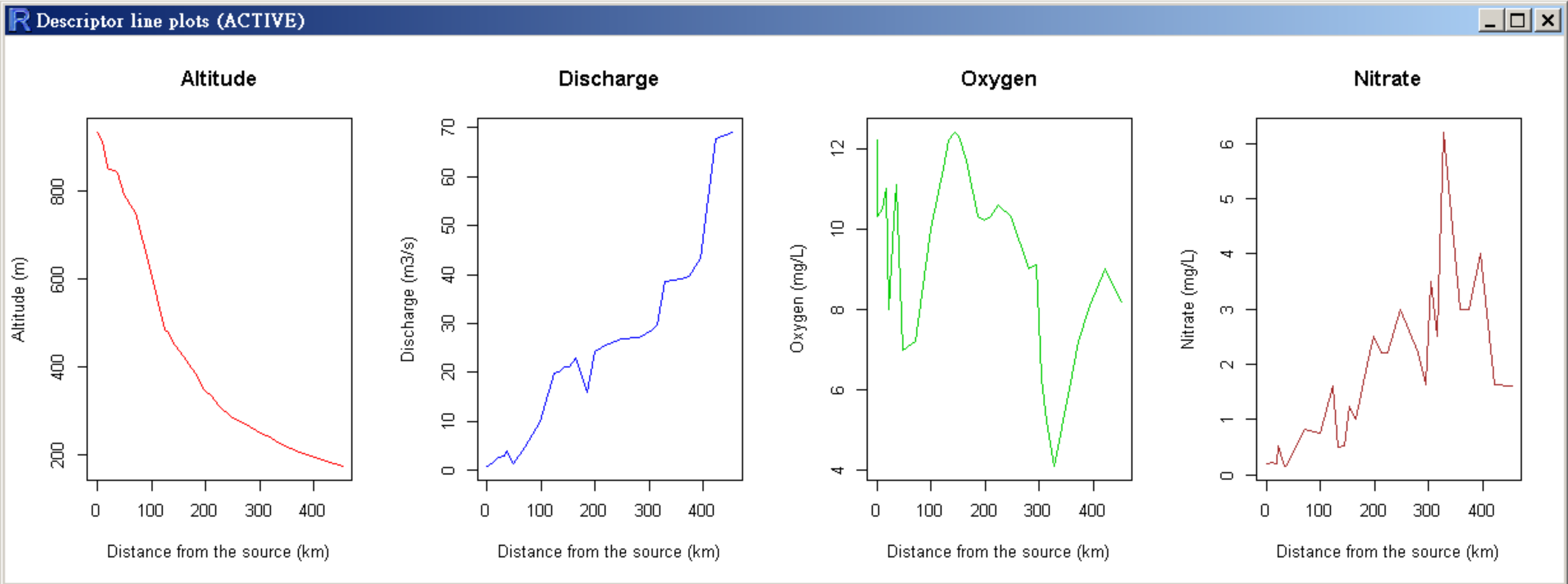
Apply the basic functions to **env**. While examining the **summary()**, note how the variables differ from the species data in values and spatial distributions. Draw maps of some of the environmental variables.



Which ones of these maps display an upstream-downstream gradient? How could you explain the spatial patterns of the other variables?

Examine the Variation of Some Descriptors Along the Stream: Line Plots

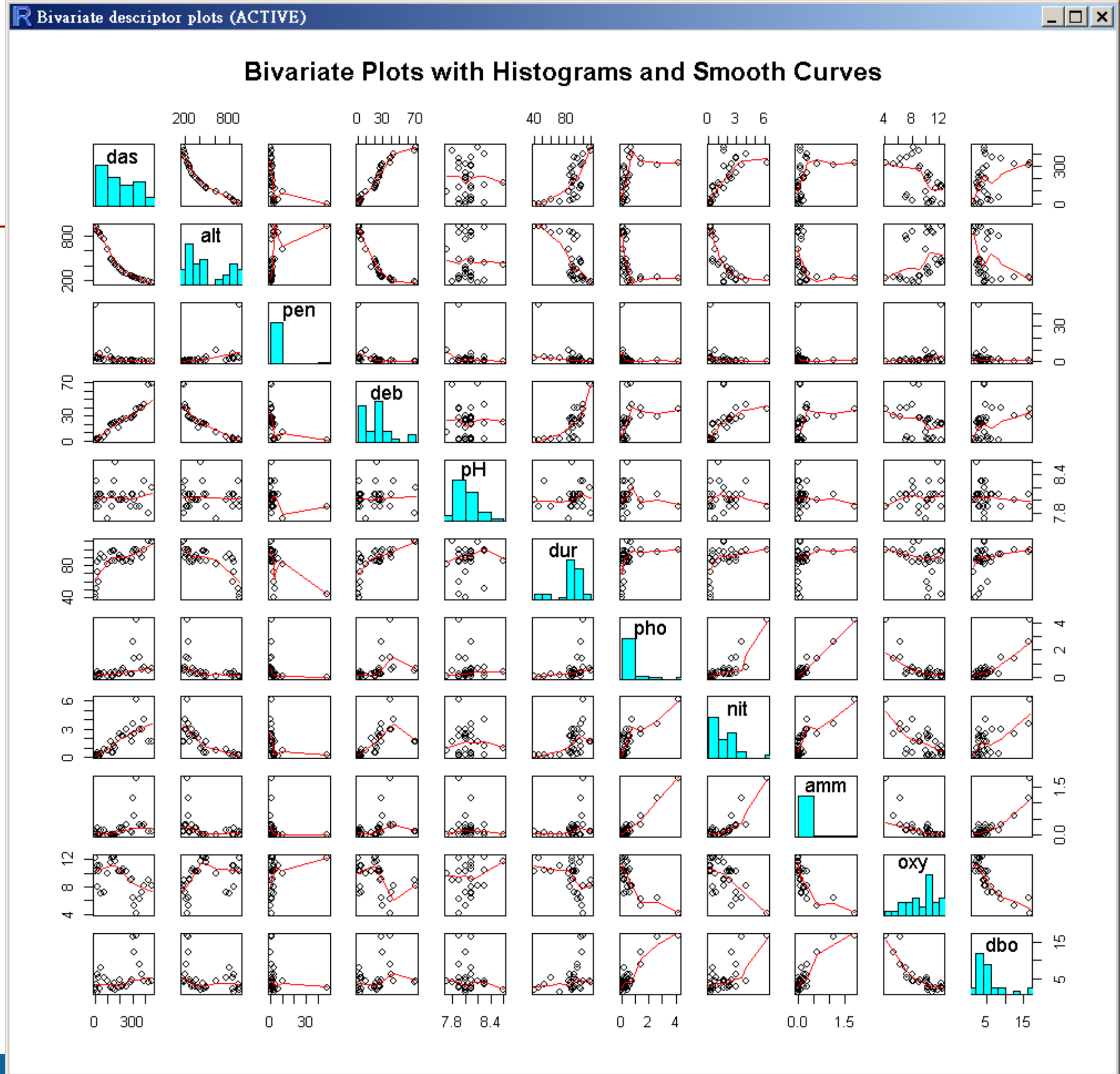
```
> windows(title="Descriptor line plots")
> par(mfrow=c(1,4))
> plot(env$das, env$alt, type="l", xlab="Distance from the source (km)",
+ ylab="Altitude (m)", col="red", main="Altitude")
> plot(env$das, env$deb, type="l", xlab="Distance from the source (km)",
+ ylab="Discharge (m3/s)", col="blue", main="Discharge")
> plot(env$das, env$oxy, type="l", xlab="Distance from the source (km)",
+ ylab="Oxygen (mg/L)", col="green3", main="Oxygen")
> plot(env$das, env$nit, type="l", xlab="Distance from the source (km)",
+ ylab="Nitrate (mg/L)", col="brown", main="Nitrate")
```



Note the scaleings.

Scatter Plots for All Pairs of Environmental Variables

```
> windows(title="Bivariate descriptor plots")
> source("panelutils.R")
> op <- par(mfrow=c(1,1), pty="s")
> pairs(env, panel=panel.smooth,
diag.panel=panel.hist,
main="Bivariate Plots with
Histograms and Smooth Curves")
> par(op)
```

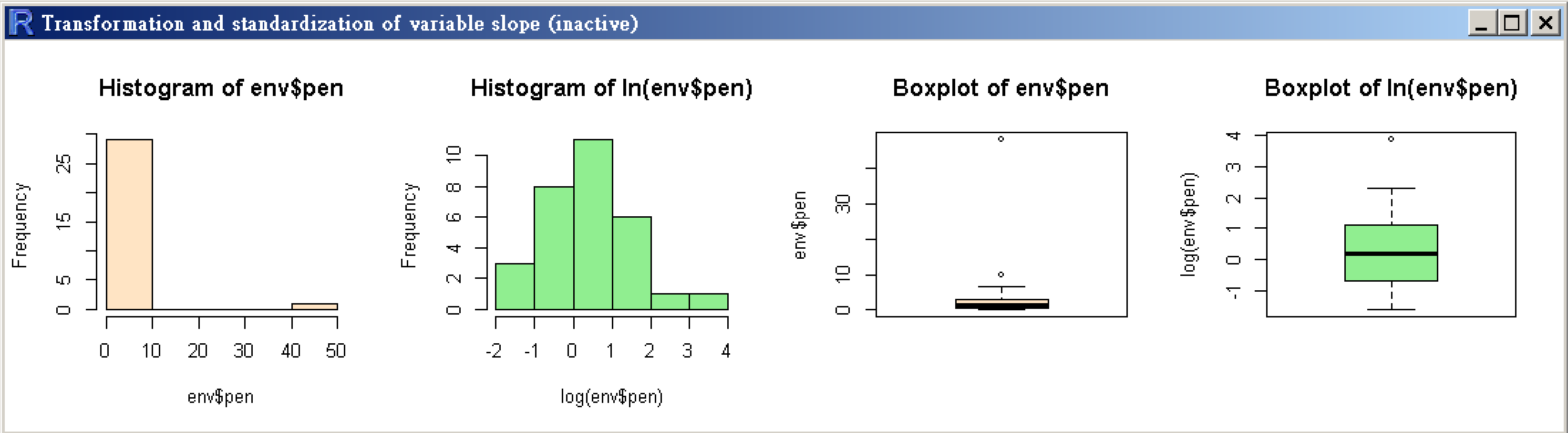


Do many variables seem normally distributed?
Do many scatter plots show linear or at least monotonic relationships?

Simple Transformation of An Environmental Variable

- Simple transformations, such as the log transformation, can be used to improve the distributions of some variables (make it closer to the normal distribution).
- Because environmental variables are dimensionally heterogeneous (expressed in different units and scales), many statistical analyses require their standardization to zero mean and unit variance. These centred and scaled variables are called z-scores.

```
> range(env$pen)
[1]  0.2 48.0
> # Log-transformation of the slope variable (y = ln(x))
> # Compare histograms and boxplots of raw and transformed values
> windows(title="Transformation and standardization of variable slope")
> par(mfrow=c(1,4))
> hist(env$pen, col="bisque", right=FALSE)
> hist(log(env$pen), col="light green", right=F, main="Histogram of ln(env$pen)")
> boxplot(env$pen, col="bisque", main="Boxplot of env$pen", ylab="env$pen")
> boxplot(log(env$pen), col="light green", main="Boxplot of ln(env$pen)",
+ ylab="log(env$pen)")
```



Standardization of All Environmental Variables

```
> # Center and scale = standardize variables (z-scores)
> env.z <- decostand(env, "standardize")
> apply(env.z, 2, mean)   # means = 0
          das           alt           pen           deb            pH           dur 
 1.000429e-16  1.814232e-18 -1.659010e-17  1.233099e-17 -4.096709e-15  3.348595e-16 
          pho           nit           amm           oxy           dbo 
 1.327063e-17 -8.925898e-17 -4.289646e-17 -2.886092e-16  7.656545e-17 
> apply(env.z, 2, sd)   # standard deviations = 1
das alt pen deb  pH dur pho nit amm oxy dbo 
  1   1   1   1   1   1   1   1   1   1   1 
>
> # Same standardization using the scale() function (which returns a matrix)
> env.z <- as.data.frame(scale(env))
> env.z
           das          alt         pen         deb         pH          dur
1  -1.34949526  1.667360909  5.14106053 -1.18004457 -0.8635475 -2.436958124
2  -1.33585215  1.659991358 -0.05737533 -1.17120570 -0.2878492 -2.733425049
...
```

小結 & 想想看

- The EDA tools allow researchers to obtain a general impression of their data.
- Information about simple parameters and distributions of variables is important to consider in order to choose more advanced analyses correctly.
- Graphical representations may help generate hypotheses about the processes acting behind the scene.

try heatmap!

- EDA is often neglected by people who are eager to jump to more sophisticated analyses.
- **想想看: Doubs Fish Data經過這一連串的資料探索，還有哪一些有趣的問題可以提出?**

Example 2: Hourly Ozone Data

Source: Roger D. Peng, (2015), *Exploratory Data Analysis with R*, Coursera.

Exploratory Data Analysis Checklist

0) Prepare your data
1) Formulate your question
2) Read in your data
3) Check the packaging
4) Run `str()`
5) Look at the top and the bottom of your data
6) Check your "n"s
7) Validate with at least one external data source
8) Try the easy solution first
9) Challenge your solution
10) Follow up

Together with graphics!

0. Prepare Your Data (1/3)

- **Dataset:** an air pollution (hourly ozone levels) dataset from the U.S. Environmental Protection Agency (EPA) for the year 2014.

  http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

- U.S. EPA on hourly ozone measurements in the entire U.S. for the year 2014. The data are available from the EPA's Air Quality System web page.

- The dataset is a comma-separated value (CSV) file, where each row of the file contains one hourly measurement of ozone at some location in the country.

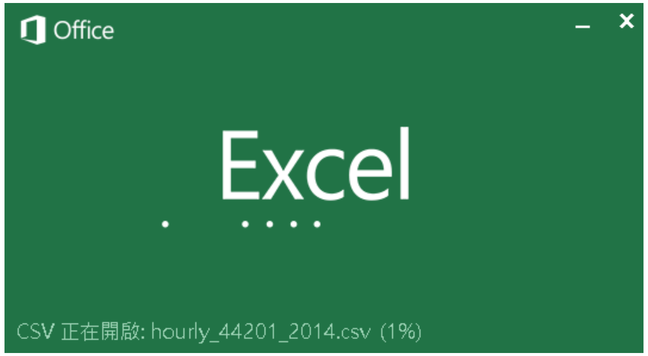
0. Prepare Your Data (2/3)

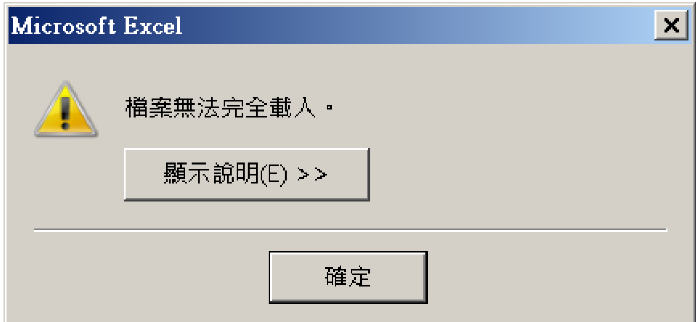
Hourly Data

Criteria Gases

Year	Ozone (44201)	SO2 (42401)	CO (42101)	NO2 (42602)
2015	hourly_44201_2015.zip 1,575,854 Rows 11,553 KB As of 2015-06-20	hourly_42401_2015.zip 813,370 Rows 5,510 KB As of 2015-06-20	hourly_42101_2015.zip 468,398 Rows 3,415 KB As of 2015-06-20	hourly_42602_2015.zip 677,518 Rows 5,303 KB As of 2015-06-20
2014	hourly_44201_2014.zip 8,967,571 Rows 66,326 KB As of 2015-06-20	hourly_42401_2014.zip 3,724,805 Rows 24,719 KB As of 2015-06-20	hourly_42101_2014.zip 2,457,531 Rows 16,890 KB As of 2015-06-20	hourly_42602_2014.zip 3,382,360 Rows 25,630 KB As of 2015-06-20

dataset:
hourly_44201_2014.zip (64.7M)
hourly_44201_2014.csv (1.89G)





	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Parameter N	Date Local	Time Local	Date GMT	Time GMT	Sample Mea	Units of Me	MDL	Uncertainty	Qualifier	Method Typ	Method Coc	Method Nar	State Name	County Nan	Date of Last Change		
1048572	Ozone	2014/1/22	19:00	2014/1/23	03:00	0.002	Parts per mi	0.005			FEM	87	INSTRUMI	California	Merced	2014/8/4		
1048573	Ozone	2014/1/22	21:00	2014/1/23	05:00	0.002	Parts per mi	0.005			FEM	87	INSTRUMI	California	Merced	2014/8/4		
1048574	Ozone	2014/1/22	22:00	2014/1/23	06:00	0.012	Parts per mi	0.005			FEM	87	INSTRUMI	California	Merced	2014/8/4		
1048575	Ozone	2014/1/22	23:00	2014/1/23	07:00	0.013	Parts per mi	0.005			FEM	87	INSTRUMI	California	Merced	2014/8/4		
1048576	Ozone	2014/1/23	00:00	2014/1/23	08:00	0.014	Parts per mi	0.005			FEM	87	INSTRUMI	California	Merced	2014/8/4		

0. Prepare Your Data (3/3)

```
hourly_44201_2014.csv
1 "State Code","County Code","Site Num","Parameter Code","POC","Latitude","Longitude","Datum","Parameter Name","Date Local","Time Local","Date GMT"
2 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","01:00","2014-03-01","07:00",0.047,"Parts per million",0.005,"","",
3 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","02:00","2014-03-01","08:00",0.047,"Parts per million",0.005,"","",
4 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","03:00","2014-03-01","09:00",0.043,"Parts per million",0.005,"","",
5 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","04:00","2014-03-01","10:00",0.038,"Parts per million",0.005,"","",
6 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","05:00","2014-03-01","11:00",0.035,"Parts per million",0.005,"","",
7 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","06:00","2014-03-01","12:00",0.035,"Parts per million",0.005,"","",
8 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","07:00","2014-03-01","13:00",0.034,"Parts per million",0.005,"","",
9 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","08:00","2014-03-01","14:00",0.037,"Parts per million",0.005,"","",
10 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","09:00","2014-03-01","15:00",0.044,"Parts per million",0.005,"","",
11 "01","003","0010","44201",1,30.498001,-87.881412,"NAD83","Ozone","2014-03-01","10:00","2014-03-01","16:00",0.046,"Parts per million",0.005,"","",

8967568 "80","006","0007","44201",1,31.7122,-106.3953,"NAD83","Ozone","2014-08-31","19:00","2014-09-01","02:00",0.019,"Parts per million",0.005,"","","FE
8967569 "80","006","0007","44201",1,31.7122,-106.3953,"NAD83","Ozone","2014-08-31","20:00","2014-09-01","03:00",0.021,"Parts per million",0.005,"","","FE
8967570 "80","006","0007","44201",1,31.7122,-106.3953,"NAD83","Ozone","2014-08-31","21:00","2014-09-01","04:00",0.024,"Parts per million",0.005,"","","FE
8967571 "80","006","0007","44201",1,31.7122,-106.3953,"NAD83","Ozone","2014-08-31","22:00","2014-09-01","05:00",0.002,"Parts per million",0.005,"","","FE
8967572 "80","006","0007","44201",1,31.7122,-106.3953,"NAD83","Ozone","2014-08-31","23:00","2014-09-01","06:00",0.002,"Parts per million",0.005,"","","FE
8967573
1.89 GB (2,034,887,869 字節)，8,967,573 行。                TeX  行 8967573, 欄 1
```

There are 34 variables with 8967571 observations:

"State Code","County Code","Site Num","Parameter Code", "POC", "Latitude","Longitude","Datum","Parameter Name","Date Local", "Time Local","Date GMT","Time GMT","Sample Measurement","Units of Measure", "MDL","Uncertainty","Qualifier","Method Type","Method Code", "Method Name","State Name","County Name","Date of Last Change"

註: 如何呈現這些變數的內容及資訊?

1. Formulate Your Question

- A general question:
  - Are air pollution levels higher on the east coast than on the west coast?
- A more specific question:
  - Are hourly ozone levels on average higher in New York City than they are in Los Angeles?
- Figure out what is the question you're really interested in, and narrow it down to be as specific as possible.

2. Read in Your Data (1/3)

- Sometimes the data need to be cleaned up.
- You can read in a subset by specifying a value for the **n_max** argument to **read_csv()** that is greater than 0.

```
# getwd(), setwd(), list.files()
# The readr package is a nice package for reading in flat files very fast.
> library(readr)
警告訊息：
package ‘readr’ was built under R version 3.1.3

#  If col_types is not specified and read_csv() will try to figure it out.
> ozone <- read_csv("data/hourly_44201_2014.csv")
|                        | 0%    2 MB
...
|========================| 100% 1940 MB
警告訊息：
44153 problems parsing 'data/hourly_44201_2014.csv'. See problems(...) for
more details.
```

比較一下:

```
read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

```
read_csv(file, col_names = TRUE, col_types = NULL, na = "NA", skip = 0,
  n_max = -1, progress = interactive())
```

2. Read in Your Data (2/2)

```
> head(problems(ozone))
   row col        expected actual
1 6019  18 T/F/TRUE/FALSE      2
2 6020  18 T/F/TRUE/FALSE      2
3 6021  18 T/F/TRUE/FALSE      2
4 6022  18 T/F/TRUE/FALSE      2
5 6023  18 T/F/TRUE/FALSE      2
6 6024  18 T/F/TRUE/FALSE      2

>  #Rewrite the names of the columns to remove any spaces.
> names(ozone)
 [1] "State Code"          "County Code"         "Site Num"            "Parameter Code"
 [5] "POC"                 "Latitude"            "Longitude"           "Datum"
 [9] "Parameter Name"      "Date Local"          "Time Local"          "Date GMT"
[13] "Time GMT"            "Sample Measurement"  "Units of Measure"    "MDL"
[17] "Uncertainty"         "Qualifier"           "Method Type"         "Method Code"
[21] "Method Name"         "State Name"          "County Name"         "Date of Last Change"
> (names(ozone) <- make.names(names(ozone)))
 [1] "State.Code"          "County.Code"         "Site.Num"            "Parameter.Code"
 [5] "POC"                 "Latitude"            "Longitude"           "Datum"
 [9] "Parameter.Name"      "Date.Local"          "Time.Local"          "Date.GMT"
[13] "Time.GMT"            "Sample.Measurement"  "Units.of.Measure"    "MDL"
[17] "Uncertainty"         "Qualifier"           "Method.Type"         "Method.Code"
[21] "Method.Name"         "State.Name"          "County.Name"         "Date.of.Last.Change"
```

3. Check the Packaging

- check the number of rows and columns.

```
> nrow(ozone)
[1] 7147884
> ncol(ozone)
[1] 23
```

- check the original text file to see if the number of columns printed out (23) here matches the number of columns you see in the original file.

```
> memory.size(max = FALSE) # 目前使用的記憶體量
[1] 2613.55
> memory.size(max = TRUE) # 從作業系統可得到的最大量記憶體
[1] 2953.06
> memory.limit(size = NA) # 列出目前記憶體的限制
[1] 16343
> memory.limit(size = 2048) # 設定新的記憶體限制為 2048MB
[1] 16343
警告訊息：
In memory.limit(size = 2048) : 無法減少記憶體限制：已忽略
> print(object.size(ozone), units = "Mb")
1607.9 Mb
```

4. Run `str()` (1/2)

```
> str(ozone)
Classes 'tbl_df', 'tbl' and 'data.frame':    8967571 obs. of  24 variables:
 $ State Code         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ County Code        : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Site Num           : int  10 10 10 10 10 10 10 10 10 10 ...
 $ Parameter Code     : int  44201 44201 44201 44201 44201 44201 44201 44201 44201 44201 ...
 $ POC                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Latitude           : num  30.5 30.5 30.5 30.5 30.5 ...
 $ Longitude          : num  -87.9 -87.9 -87.9 -87.9 -87.9 ...
 $ Datum              : chr  "NAD83" "NAD83" "NAD83" "NAD83" ...
 $ Parameter Name     : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
 $ Date Local         : Date, format: "2014-03-01" "2014-03-01" ...
 $ Time Local         : chr  "01:00" "02:00" "03:00" "04:00" ...
 $ Date GMT           : Date, format: "2014-03-01" "2014-03-01" ...
 $ Time GMT           : chr  "07:00" "08:00" "09:00" "10:00" ...
 $ Sample Measurement : num  0.047 0.047 0.043 0.038 0.035 0.035 0.034 0.037 0.044 0.046 ...
 $ Units of Measure   : chr  "Parts per million" "Parts per million" "Parts per million" "Parts per
million" ...
 $ MDL                : num  0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005 ...
 $ Uncertainty        : logi  NA NA NA NA NA NA ...
 $ Qualifier          : logi  NA NA NA NA NA NA ...
 $ Method Type        : chr  "FEM" "FEM" "FEM" "FEM" ...
 $ Method Code        : int  47 47 47 47 47 47 47 47 47 47 ...
 $ Method Name        : chr  "INSTRUMENTAL - ULTRA VIOLET" "INSTRUMENTAL - ULTRA VIOLET" "INSTRUMENTAL
- ULTRA VIOLET" "INSTRUMENTAL - ULTRA VIOLET" ...
 $ State Name         : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ County Name        : chr  "Baldwin" "Baldwin" "Baldwin" "Baldwin" ...
 $ Date of Last Change: Date, format: "2014-06-30" "2014-06-30" ...
 - attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame':    44153 obs. of  4 variables:
  ..$ row     : int  6019 6020 6021 6022 6023 6024 6025 6363 6364 6365 ...
  ..$ col     : int  18 18 18 18 18 18 18 18 18 18 ...
  ..$ expected: chr  "T/F/TRUE/FALSE" "T/F/TRUE/FALSE" "T/F/TRUE/FALSE" "T/F/TRUE/FALSE" ...
  ..$ actual  : chr  "2" "2" "2" "2" ...
```

4. Run str() (2/2)

```
> remove(ozone)
> ozone <- read_csv("data/hourly_44201_2014.csv", col_types = "ccccinnccccccncnncccccccc")
|================================| 100% 1940 MB
> names(ozone) <- make.names(names(ozone))
> str(ozone )
Classes 'tbl_df' , 'tbl' and 'data.frame':      8967571 obs. of  24 variables:
 $ State Code         : chr  "01" "01" "01" "01" ...
 $ County Code        : chr  "003" "003" "003" "003" ...
 $ Site Num           : chr  "0010" "0010" "0010" "0010" ...
 $ Parameter Code     : chr  "44201" "44201" "44201" "44201" ...
 $ POC                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Latitude           : num  30.5 30.5 30.5 30.5 30.5 ...
 $ Longitude          : num  -87.9 -87.9 -87.9 -87.9 -87.9 ...
 $ Datum              : chr  "NAD83" "NAD83" "NAD83" "NAD83" ...
 $ Parameter Name     : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
 $ Date Local         : chr  "2014-03-01" "2014-03-01" "2014-03-01" "2014-03-01" ...
 $ Time Local         : chr  "01:00" "02:00" "03:00" "04:00" ...
 $ Date GMT           : chr  "2014-03-01" "2014-03-01" "2014-03-01" "2014-03-01" ...
 $ Time GMT           : chr  "07:00" "08:00" "09:00" "10:00" ...
 $ Sample Measurement : num  0.047 0.047 0.043 0.038 0.035 0.035 0.034 0.037 0.044 0.046 ...
 $ Units of Measure   : chr  "Parts per million" "Parts per million" "Parts per million" "Parts per
million" ...
 $ MDL                : num  0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005 ...
 $ Uncertainty        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Qualifier          : chr  "" "" "" "" ...
 $ Method Type        : chr  "FEM" "FEM" "FEM" "FEM" ...
 $ Method Code        : chr  "047" "047" "047" "047" ...
 $ Method Name        : chr  "INSTRUMENTAL - ULTRA VIOLET" "INSTRUMENTAL - ULTRA VIOLET" "INSTRUMENTAL
- ULTRA VIOLET" "INSTRUMENTAL - ULTRA VIOLET" ...
 $ State Name         : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ County Name        : chr  "Baldwin" "Baldwin" "Baldwin" "Baldwin" ...
 $ Date of Last Change: chr  "2014-06-30" "2014-06-30" "2014-06-30" "2014-06-30" ...
```

c: character
n: numeric
i: integer

5. Look at the Top (head) and the Bottom (tail) of Your Data

- Make sure to check all the columns and verify that all of the data in each column looks the way it's supposed to look.

```
> head(ozone) #tail(ozone)
  State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude Datum Parameter.Name
1         01         003     0010          44201   1   30.498 -87.88141 NAD83          Ozone
2         01         003     0010          44201   1   30.498 -87.88141 NAD83          Ozone
3         01         003     0010          44201   1   30.498 -87.88141 NAD83          Ozone
4         01         003     0010          44201   1   30.498 -87.88141 NAD83          Ozone
5         01         003     0010          44201   1   30.498 -87.88141 NAD83          Ozone
6         01         003     0010          44201   1   30.498 -87.88141 NAD83          Ozone
  Date.Local Time.Local   Date.GMT Time.GMT Sample.Measurement  Units.of.Measure   MDL
1 2014-03-01      01:00 2014-03-01    07:00              0.047 Parts per million 0.005
2 2014-03-01      02:00 2014-03-01    08:00              0.047 Parts per million 0.005
3 2014-03-01      03:00 2014-03-01    09:00              0.043 Parts per million 0.005
4 2014-03-01      04:00 2014-03-01    10:00              0.038 Parts per million 0.005
5 2014-03-01      05:00 2014-03-01    11:00              0.035 Parts per million 0.005
6 2014-03-01      06:00 2014-03-01    12:00              0.035 Parts per million 0.005
  Uncertainty Qualifier Method.Type Method.Code                Method.Name State.Name
1          NA                   FEM         047 INSTRUMENTAL - ULTRA VIOLET    Alabama
2          NA                   FEM         047 INSTRUMENTAL - ULTRA VIOLET    Alabama
3          NA                   FEM         047 INSTRUMENTAL - ULTRA VIOLET    Alabama
4          NA                   FEM         047 INSTRUMENTAL - ULTRA VIOLET    Alabama
5          NA                   FEM         047 INSTRUMENTAL - ULTRA VIOLET    Alabama
6          NA                   FEM         047 INSTRUMENTAL - ULTRA VIOLET    Alabama
  County.Name Date.of.Last.Change
1     Baldwin          2014-06-30
2     Baldwin          2014-06-30
3     Baldwin          2014-06-30
4     Baldwin          2014-06-30
5     Baldwin          2014-06-30
6     Baldwin          2014-06-30
```

6. Check Your "n"s (1/3)

- Check the dataset to make sure that you have data on all subjects.
- Use the fact that the dataset purportedly contains hourly data for the entire country. These will be our two landmarks for comparison.
- The hourly ozone data comes from monitors across the country. The monitors should be monitoring continuously during the day, so all hours should be represented.
- We can take a look at the `Time.Local` variable to see what time measurements are recorded as being taken.

6. Check Your "n"s (2/3)

- Almost all measurements in the dataset are recorded as being taken on the hour, some are taken at slightly different times.
- Such a small number of readings are taken at these off times that we might not want to care.
- But it does seem a bit odd, so it might be worth a quick check.

```
> table(ozone$Time.Local)

 00:00  00:01  01:00  01:02  02:00  02:03  03:00
288698      2 290871      2 283709      2 282951
 03:04  04:00  04:05  05:00  05:06  06:00  06:07
     2 288963      2 302696      2 302356      2
 07:00  07:08  08:00  08:09  09:00  09:10  10:00
300950      2 298566      2 297154      2 297132
 10:11  11:00  11:12  12:00  12:13  13:00  13:14
     2 298125      2 298297      2 299997      2
 14:00  14:15  15:00  15:16  16:00  16:17  17:00
301410      2 302636      2 303387      2 303806
 17:18  18:00  18:19  19:00  19:20  20:00  20:21
     2 303795      2 304268      2 304268      2
 21:00  21:22  22:00  22:23  23:00  23:24
303551      2 295701      2 294549      2
```

```
> table(ozone$Time.Local)

 00:00  01:00  02:00  03:00  04:00  05:00  06:00  07:00  08:00
365278 366282 355919 349843 353867 380124 379771 378238 375373
 09:00  10:00  11:00  12:00  13:00  14:00  15:00  16:00  17:00
373472 373225 374424 374742 376722 378593 380236 381336 381889
 18:00  19:00  20:00  21:00  22:00  23:00
381806 382354 382329 381144 371224 369380
```

6. Check Your "n"s (3/3)

- Since EPA monitors pollution across the country, there should be a good representation of states. Perhaps we should see exactly how many states are represented in this dataset.
- There are 52 states in the dataset, but only 50 states in the U.S.!
- Now we can see that Washington, D.C. (District of Columbia) and Puerto Rico are the "extra" states included in the dataset. Since they are clearly part of the U.S. (but not official states of the union) that all seems okay)

```
> select(ozone, State.Name) %>% unique %>% nrow
[1] 53
> unique(ozone$State.Name)
 [1] "Alabama"        "Alaska"         "Arizona"          "Arkansas"             "California"
 [6] "Colorado"       "Connecticut"    "Delaware"         "District Of Columbia" "Florida"
[11] "Georgia"        "Hawaii"         "Idaho"            "Illinois"             "Indiana"
[16] "Iowa"           "Kansas"         "Kentucky"         "Louisiana"            "Maine"
[21] "Maryland"       "Massachusetts"  "Michigan"         "Minnesota"            "Mississippi"
[26] "Missouri"       "Montana"        "Nebraska"         "Nevada"               "New Hampshire"
[31] "New Jersey"     "New Mexico"     "New York"         "North Carolina"       "North Dakota"
[36] "Ohio"           "Oklahoma"       "Oregon"           "Pennsylvania"         "Rhode Island"
[41] "South Carolina" "South Dakota"   "Tennessee"        "Texas"                "Utah"
[46] "Vermont"        "Virginia"       "Washington"       "West Virginia"        "Wisconsin"
[51] "Wyoming"        "Puerto Rico"    "Country Of Mexico"
```

7. Validate with at Least One External Data Source (1/3)

- Making sure your data matches something outside of the dataset is very important.
- External validation can often be as simple as checking your data against a single number.
- In the U.S. we have **national ambient air quality standards**, and for ozone, the current standard(*) set in 2008 is that the "annual fourth-highest daily maximum 8-hr concentration, averaged over 3 years" should not exceed **0.075** parts per million (ppm).
- The 8-hour average concentration should not be too much higher than 0.075 ppm (it can be higher because of the way the standard is worded).

NOTE: 背景知識!

http://www.epa.gov/ttn/naaqs/standards/ozone/s_o3_history.html

7. Validate with at Least One External Data Source (2/3)

- The hourly measurements of ozone

```
> summary(ozone$Sample.Measurement)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.01900 0.03000 0.03011 0.04100 0.24100
```

- From the summary we can see that the maximum hourly concentration is quite high (0.241 ppm) (0.349 ppm) but that in general, the bulk of the distribution is far below 0.075.

- We can get a bit more detail on the distribution by looking at deciles of the data.

```
> quantile(ozone$Sample.Measurement, seq(0, 1, 0.1))
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
0.000 0.009 0.016 0.022 0.026 0.030 0.034 0.038 0.043 0.050 0.241
```

7. Validate with at Least One External Data Source (3/3)

- Knowing that the national standard for ozone is something like **0.075**, we can see from the data that
  - The data are at least of the right order of magnitude (i.e. the units are correct)
  - The range of the distribution is roughly what we' d expect, given the regulation around ambient pollution levels.
  - Some hourly levels (less than 10%) are above 0.075 but this may be reasonable given the wording of the standard and the averaging involved.

8. Try the Easy Solution First (1/4)

- The original question: which counties in the United States have the highest levels of ambient ozone pollution?
- We need a list of counties that are ordered from highest to lowest with respect to their levels of ozone.
  - levels of ozone: take the average across the entire year for each county and then rank counties according to this metric.

```
> ranking <- group_by(ozone, State.Name, County.Name) %>%
+            summarize(ozone = mean(Sample.Measurement)) %>%
+            as.data.frame %>%
+            arrange(desc(ozone))
>
> head(ranking, 10) #the top 10 counties
       State.Name County.Name      ozone
1      California    Mariposa 0.04849027
2      California      Nevada 0.04821713
3         Wyoming      Albany 0.04738065
4      California        Inyo 0.04469113
5            Utah    San Juan 0.04457553
6      California   El Dorado 0.04363664
7          Nevada  White Pine 0.04344640
8  North Carolina      Yancey 0.04337582
9  North Carolina     Jackson 0.04314067
10       Colorado    Gunnison 0.04302312
```

To identify each county we will use a combination of the **State.Name** and the **County.Name** variables.

It seems interesting that all of these counties are in the western U.S., with 4 of them in California alone.

8. Try the Easy Solution First (2/4)

- How many observations there are for the highest level counties, Mariposa County, California in the dataset.

```
> filter(ozone, State.Name == "California" & County.Name == "Mariposa") %>% nrow
[1] 12130
```

- Always be checking. Does that number of observations sound right? Well, there's 24 hours in a day and 365 days per, which gives us *8760*.
- Sometimes the counties use alternate methods of measurement during the year so there may be "extra" measurements.

8. Try the Easy Solution First (3/4)

- We can take a look at how ozone varies through the year in this county by looking at monthly averages.

```
> # convert the date variable into a Date class.
> ozone <- mutate(ozone, Date.Local = as.Date(Date.Local))
> # split the data by month to look at the average hourly levels.
> filter(ozone, State.Name == "California" & County.Name == "Mariposa") %>%
       mutate(month = factor(months(Date.Local), levels = month.name)) %>%
       group_by(month) %>%
       summarize(ozone = mean(Sample.Measurement))
```

- Ozone appears to be higher in the summer months and lower in the winter months.
- There are two months missing (November and December) from the data. It's probably worth investigating a bit later on.

8. Try the Easy Solution First (4/4)

```
> # look at one of the lowest level counties, Caddo County, Oklahoma
> tail(ranking, 5)
     State.Name          County.Name       ozone
787    Oklahoma                Caddo 0.017435731
788 Puerto Rico               Juncos 0.013466699
789      Alaska Fairbanks North Star 0.013419708
790 Puerto Rico              Bayamon 0.009246600
791 Puerto Rico               Catano 0.005014176
> filter(ozone, State.Name == "Oklahoma" & County.Name == "Caddo") %>% nrow
[1] 7562
> filter(ozone, State.Name == "Oklahoma" & County.Name == "Caddo") %>%
+        mutate(month = factor(months(Date.Local), levels = month.name)) %>%
+        group_by(month) %>%
+        summarize(ozone = mean(Sample.Measurement))
Source: local data frame [1 x 2]

  month      ozone
1    NA 0.01743573
```

- Here we can see that the levels of ozone are much lower in this county and that also three months are missing (October, November, and December).
- Given the seasonal nature of ozone, it’ s possible that the levels of ozone are so low in those months that it’ s not even worth measuring.
- In fact some of the monthly averages are below the typical method detection limit of the measurement technology, meaning that those values are highly uncertain and likely not distinguishable from zero.

9. Challenge Your Solution

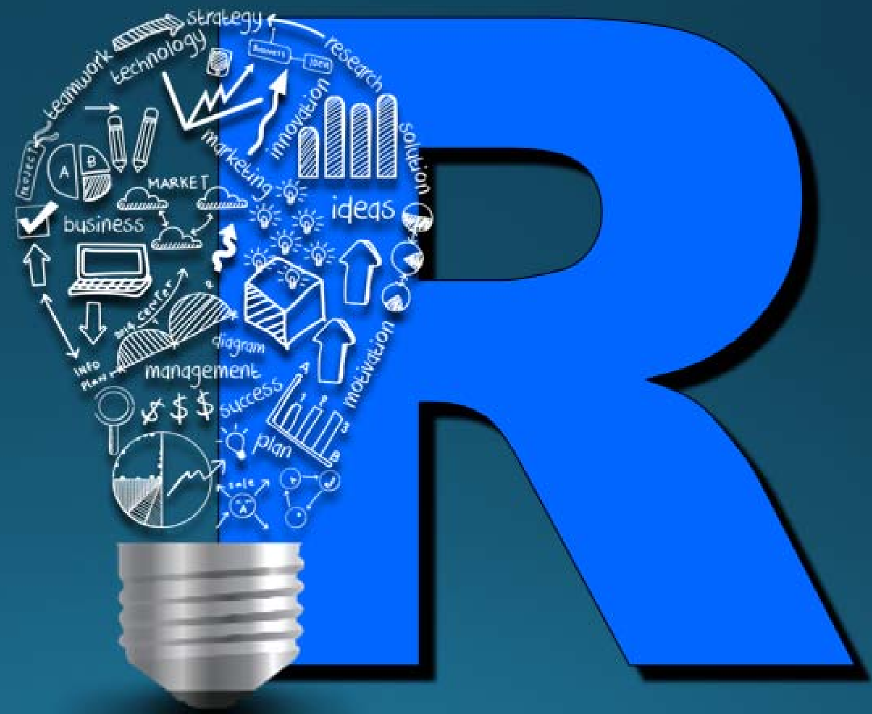
- You should always be thinking of ways to challenge the results, especially if those results **comport with your prior expectation**.
- 幾個問題:
  - Some counties do not have measurements every month. Is this a problem?
  - Would it affect our ranking of counties if we had those measurements?
  - How stable are the rankings from year to year?
  - We could get a sense of the stability of the rankings (use bootstrap samples to validate.) by shuffling the data around a bit to see if anything changes.
    - The ozone data are different randomly from year to year, but generally follow similar patterns across the country. So the shuffling process could approximate the data changing from one year to the next. It could give us a sense of how stable the rankings are.

10. Follow Up Questions

- Do you have the right data?
  - Sometime the dataset is not really appropriate for the question.
- Do you need other data?
  - e.g., whether the county rankings were stable across years?
  - We addressed this by resampling the data once to see if the rankings changed, but the better way to do this would be to get the data for previous years and re-do the rankings.
- Do you have the right question?
  - e.g., which counties were in violation of the national ambient air quality standard?
  - However, this is a much more complicated calculation to do, requiring data from at least 3 previous years.
- The goal of EDA is to get you thinking about your data and reasoning about your question. We can **refine our question** or **collect new data**, all in an iterative process to get at the truth.

探索式資料分析

EDA Assumptions



吳漢銘
國立政治大學 統計學系

https://hmwu.idv.tw

EDA Assumptions (1/2)

- There are **four assumptions** that underlie all measurement processes: the data from the process at hand "behave like":
  - random drawings;
  - from a fixed distribution;
  - with the distribution having fixed location; and
  - with the distribution having fixed variation.

- The general model for Univariate (Single Response Variable):

  response = deterministic component + random component

  becomes

  response = constant + error

http://www.itl.nist.gov/div898/handbook/eda/section2/eda2.htm

EDA Assumptions (2/2)

response = constant + error

- Assumptions for Univariate Model : the "fixed location" is simply the *unknown constant*.

- The process at hand to be operating under constant conditions that produce **a single column of data** with the properties that
  - the data are uncorrelated with one another;
  - the deterministic component consists of only a constant;
  - the random component has a fixed distribution; and
  - the random component has fixed variation.

Extrapolation to a Function of Many Variables

- The univariate model can be extended to the more general case: the deterministic component is a function of **many variables**.

http://www.itl.nist.gov/div898/handbook/eda/section2/eda2.htm

Underlying Assumptions

Residuals Will Behave According to Univariate Assumptions

- Regardless of how many factors there are, how complicated the function is, if we choose a good model, then the differences (**residuals**) between the raw **response data** and the **predicted values** from the **fitted model** should themselves behave like a univariate process.

- The residuals from this univariate process fit will behave like:
  - random drawings;
  - from a fixed distribution;
  - with fixed location; and
  - with fixed variation.

- **Validation of Model:** if the residuals violate one or more of the above univariate assumptions, then the chosen fitted model is inadequate.

「假設」的重要性

(1) Predictability and Statistical Control

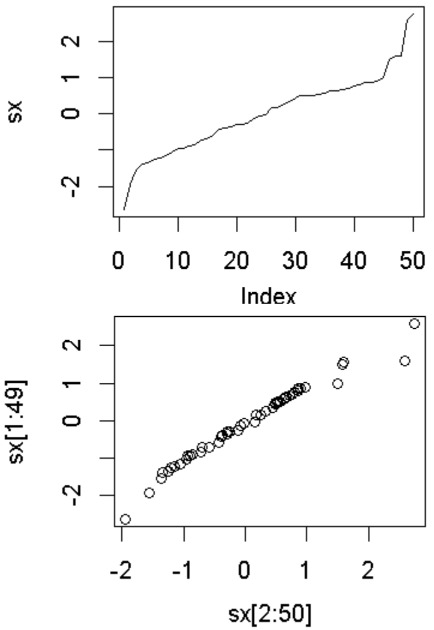
- If the four underlying assumptions hold, then we have achieved probabilistic predictability--the ability to make probability statements not only about the process in the **past**, but also about the process in the **future**.
- Such processes are said to be "in statistical control".

(2) Validity of Engineering Conclusions

- If the four assumptions are valid, then the process is amenable to the generation of valid scientific and engineering conclusions.
- If the four assumptions are not valid, then the process is unpredictable, and out of control. Such process leads to engineering conclusions that are not valid, and which are not repeatable in the laboratory.

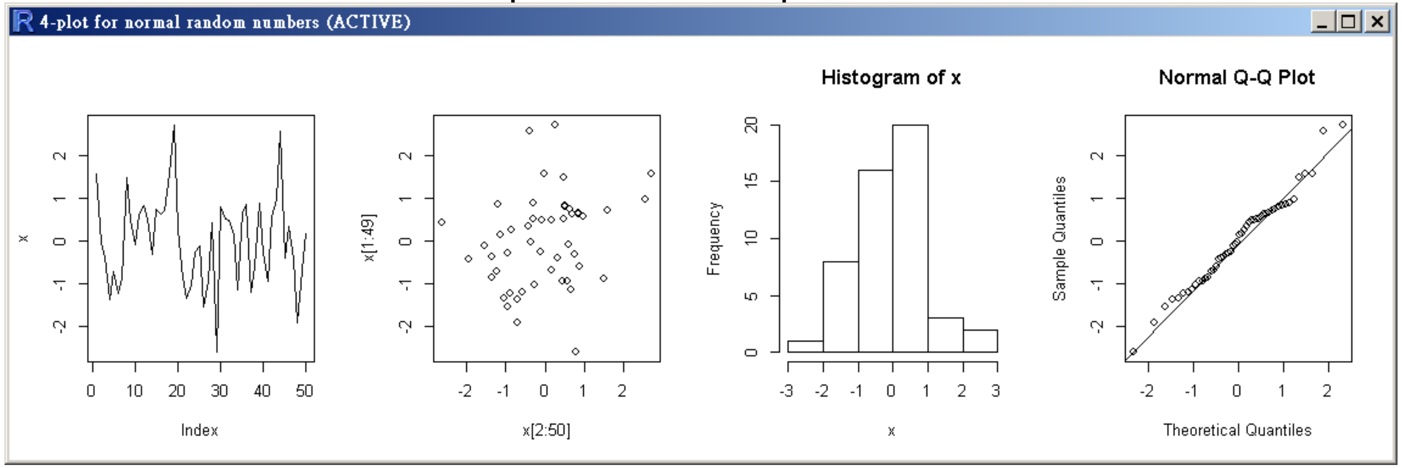
Techniques for Testing Assumptions (1/2)

- Four Techniques (4-plot) to Test Underlying Assumptions
  - run sequence plot (Y_i vs i)
  - lag plot (Y_i vs Y_{i-1})
  - histogram (counts vs subgroups of Y)
  - normal probability plot (ordered Y vs theoretical ordered Y)



This 4-plot reveals a process that has fixed location, fixed variation, is random, apparently has a fixed approximately normal distribution, and has no outliers.

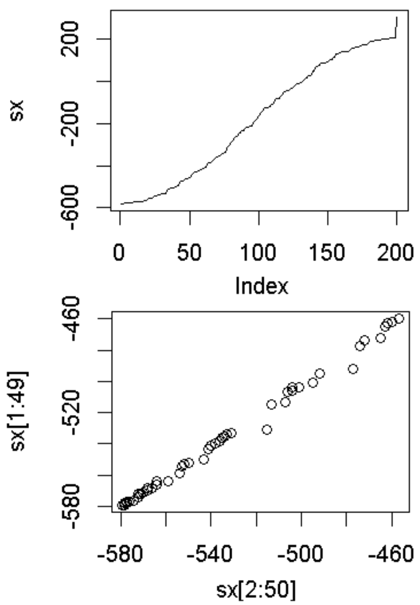
Sample Plot: Assumptions Hold



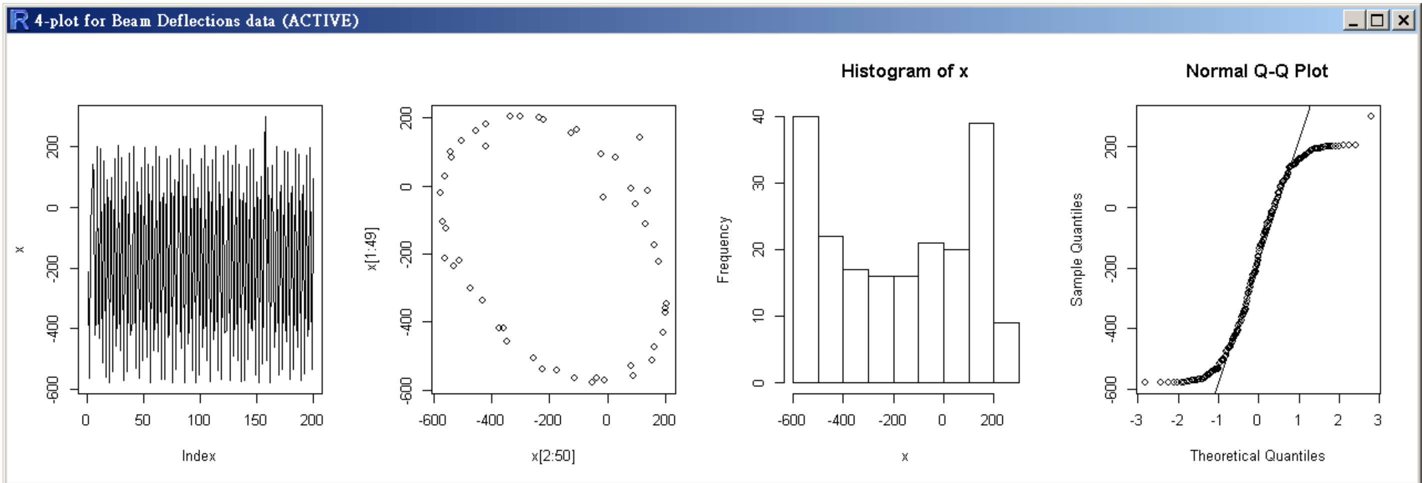
Techniques for Testing Assumptions (2/2)

- If one or more of the four underlying assumptions do not hold, then it will show up in the various plots

- This 4-plot reveals a process that has fixed location, fixed variation, is non-random (oscillatory), has a non-normal, U-shaped distribution, and has several outliers.



Sample Plot: Assumptions Do Not Hold

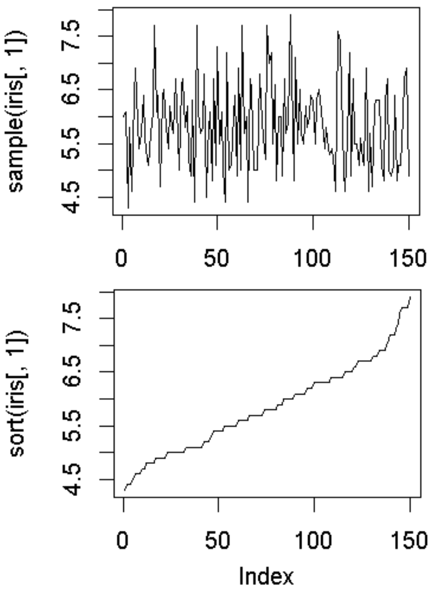


Interpretation of 4-Plot

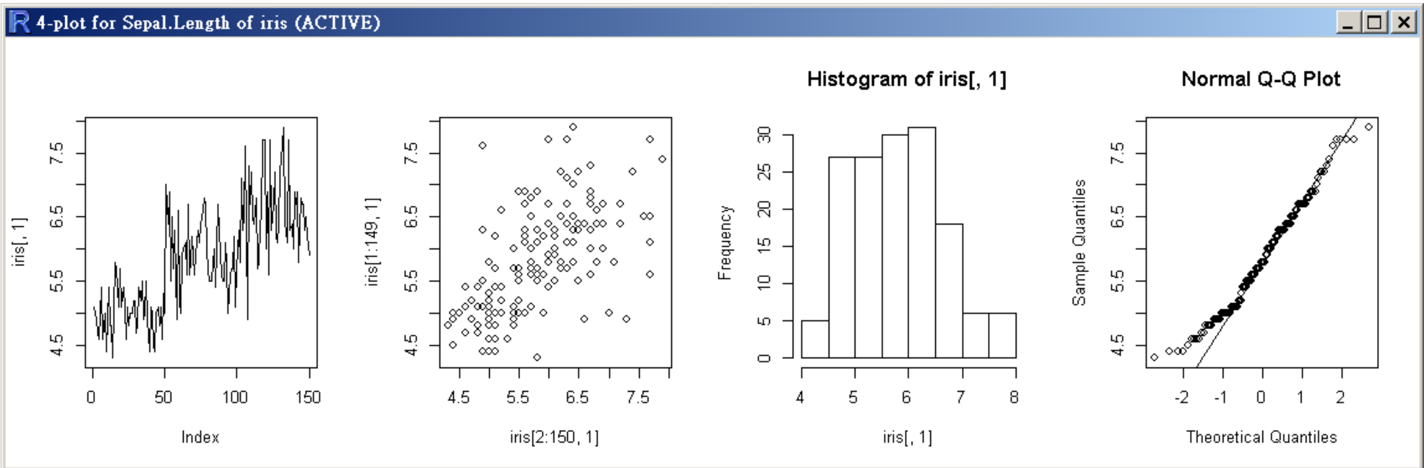
Case: Flat and Equi-Banded, Random, Bell-Shaped, and Linear

- Fixed Location holds: the run sequence plot will be flat and non-drifting.
- Fixed Variation holds : the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.
- Randomness holds: the lag plot will be structureless and random.
- Fixed (normal) Distribution holds: the histogram will be bell-shaped, and the normal probability plot will be linear.

```
plot(x, type="l")
plot(x[2:n], x[1:(n-1)])
hist(x)
qqnorm(x);qqline(x)
```



4-plot for the Sepal Length of Iris Data:



Plots Utilized to Test the Assumptions

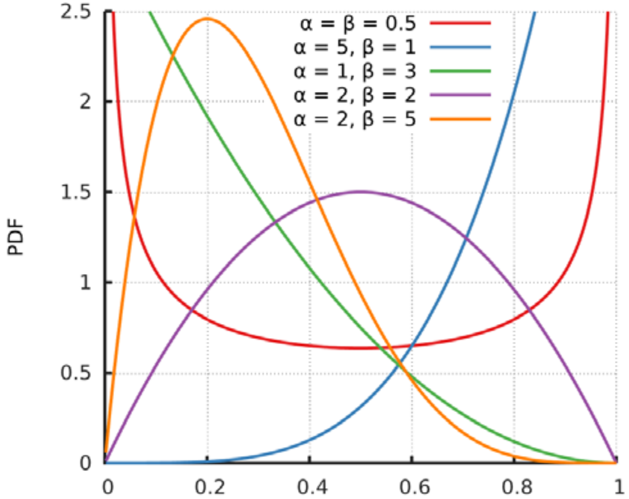
- Run Sequence Plot:
  - if the run sequence plot is flat and non-drifting, the fixed-location assumption holds.
  - If the run sequence plot has a vertical spread that is about the same over the entire plot, then the fixed-variation assumption holds.

- Lag Plot: if the lag plot is structureless, then the randomness assumption holds.

- Histogram: if the histogram is bell-shaped, the underlying distribution is symmetric and perhaps approximately normal.

- Normal Probability Plot: if the normal probability plot is linear, the underlying distribution is approximately normal.

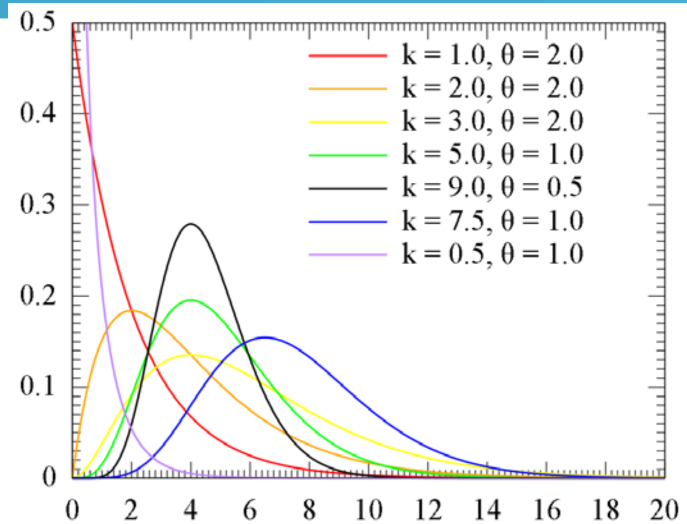
Some Probability Density Functions



Beta distribution

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

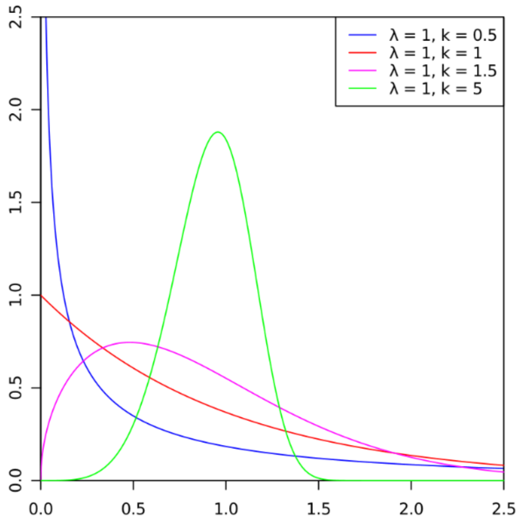
$\alpha > 0$ shape
$\beta > 0$ shape



Gamma distribution

$$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

$k > 0$ shape
$\theta > 0$ scale



Weibull distribution

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$k > 0$ shape
$\lambda > 0$ scale

If the quantity X is a "time-to-failure", the Weibull distribution gives a distribution for which the failure rate is proportional to a power of time.

https://en.wikipedia.org/wiki/

Consequences of Non-Fixed Location Parameter

- The usual estimate of location is the mean

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

- If the run sequence plot does not support the assumption of **fixed location**, then
  - The location may be drifting.
  - The single location estimate may be meaningless (if the process is drifting).
  - The choice of location estimator (e.g., the sample mean) may be sub-optimal.
  - The usual formula for the uncertainty of the mean:

$$s(\bar{Y}) = \frac{1}{\sqrt{N(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

  may be invalid and the numerical value optimistically small.
  - The location estimate may be poor.
  - The location estimate may be biased.

http://www.itl.nist.gov/div898/handbook/eda/section2/eda252.htm

Consequences of Non-Fixed Variation Parameter

- The usual estimate of variation is the standard deviation

$$s_Y = \frac{1}{\sqrt{(N-1)}} \sqrt{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}$$

- If the run sequence plot does not support the assumption of **fixed variation**, then
  - The variation may be drifting.
  - The single variation estimate may be meaningless (if the process variation is drifting).
  - The variation estimate may be poor.
  - The variation estimate may be biased.

Consequences of Non-Randomness

- The randomness assumption is the most critical but the least tested.

- If the randomness assumption does not hold, then
  - All of the usual statistical tests are invalid.
  - The calculated uncertainties for commonly used statistics become meaningless.
  - The calculated minimal sample size required for a pre-specified tolerance becomes meaningless.
  - The simple model: y = constant + error becomes invalid.
  - The parameter estimates become suspect and non-supportable.

Non-Randomness Due to Autocorrelation

- Autocorrelation is the correlation between Y_t and Y_{t-k}, where k is an integer that defines the lag for the autocorrelation.

- Autocorrelation is a time dependent non-randomness. This means that the value of the current point is highly dependent on the previous point if $k = 1$ (or k points ago if k is not 1).

- Autocorrelation is typically detected via an autocorrelation plot or a lag plot.

- If the data are not random due to autocorrelation, then
  - Adjacent data values may be related.
  - There may not be n independent snapshots of the phenomenon under study.
  - There may be undetected "junk"-outliers.
  - There may be undetected "information-rich"-outliers.

Consequences Related to Distributional Assumptions

- The mean (average) is routinely used to estimate the "middle" of a distribution. It is not so well known that the variability and the noisiness of the mean as a location estimator are intrinsically linked with the underlying distribution of the data.

- For any given distribution, the estimator with **minimum variability/noisiness** is an optimal choice.

- This optimal choice may be, for example, the **median**, the **midrange**, the **midmean**, the mean, or something else.

- The *implication* of this is to "estimate" the distribution first, and then--based on the distribution--choose the optimal estimator.

Other Consequences Related to Distributional Assumptions (1/2)

- Distribution
  - The distribution may be changing.
  - The single distribution estimate may be meaningless (if the process distribution is changing).
  - The distribution may be markedly non-normal.
  - The distribution may be unknown.
  - The true probability distribution for the error may remain unknown.

Other Consequences Related to Distributional Assumptions (2/2)

- Model
  - The model may be changing.
  - The single model estimate may be meaningless.
  - The model "Y = constant + error" may be invalid.
  - If the default model is insufficient, information about a better model may remain undetected.
  - A poor deterministic model may be fit.
  - Information about an improved model may go undetected.
- Process
  - The process may be out-of-control.
  - The process may be unpredictable.
  - The process may be un-modelable.

R package: DataExplorer

DataExplorer 0.8.1 | Reference | Articles | Changelog

DataExplorer

Background

Exploratory Data Analysis (EDA) is the initial and an important phase of data analysis/predictive modeling. During this process, analysts/modelers will have a first look of the data, and thus generate relevant hypotheses and decide next steps. However, the EDA process could be a hassle at times. This R package aims to automate most of data handling and visualization, so that users could focus on studying the data and extracting insights.

Installation

The package can be installed directly from CRAN.

```
install.packages("DataExplorer")
```

However, the latest stable version (if any) could be found on GitHub, and installed using `devtools` package.

```
if (!require(devtools)) install.packages("devtools")
devtools::install_github("boxuancui/DataExplorer")
```

If you would like to install the latest development version, you may install the develop branch.

```
if (!require(devtools)) install.packages("devtools")
devtools::install_github("boxuancui/DataExplorer", ref = "develop")
```

validate: Data Validation Infrastructure

validate: Data Validation Infrastructure
https://cran.r-project.org/web/packages/validate/index.html

The Data Validation Cookbook
Mark P.J. van der Loo
2020-12-08 | validate version 1.0.1
https://data-cleaning.github.io/validate/