Data Analytics with HPC

Data Cleaning

Reusing this material

Contents of Lecture 2

- Lecture Aim:
  - To introduce the CRISP-DM data analytics process.This process provides a framework in which to describe data cleaning.
- Lecture Contents:
  - CRISP-DM
    - Business Understanding
    - Data Understanding
    - Data Preparation
  - Data Cleaning Techniques

CRISP-DM DATA ANALYTICS PROCESS

epcc

THE UNIVERSITY OF EDINBURGH

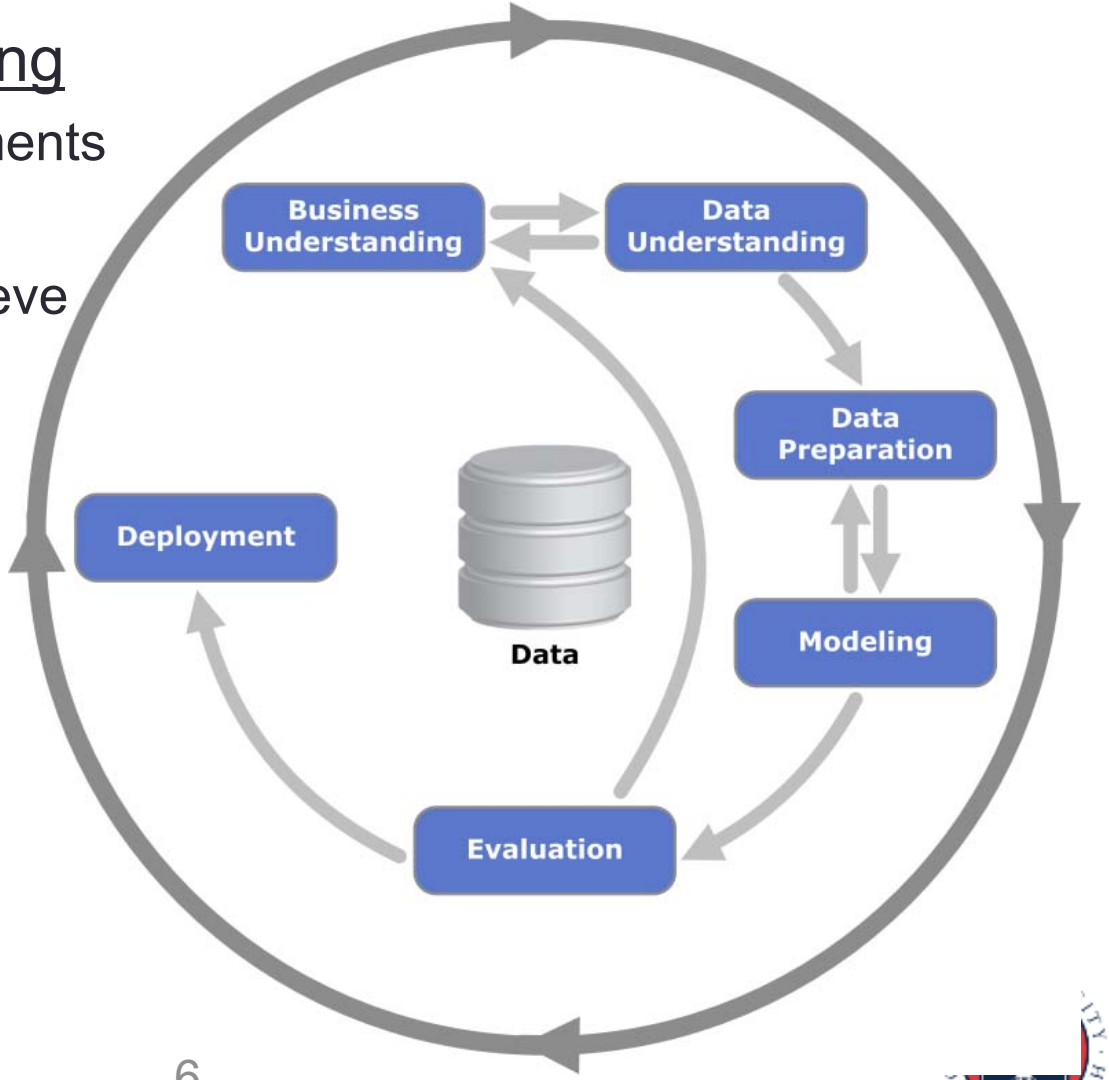
CRISP-DM

- Cross Industry Standard Process for Data Mining
  - C. Shearer, “The CRISP-DM model: the new blueprint for data mining”, Journal of Data Warehousing, Vol. 5 (4), 2000

Diagram from Wikipedia, see
http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

CRISP-DM Business Understanding

- Business Understanding
  - Objectives and requirements
  - Define the problem
  - Preliminary plan to achieve objectives



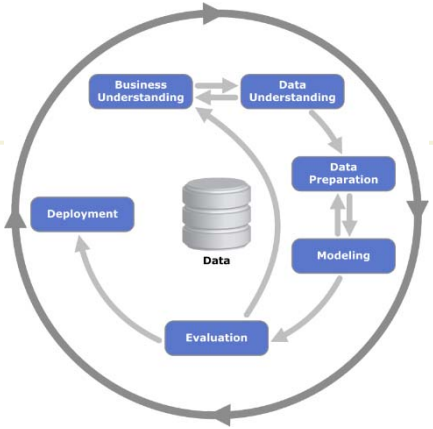
CRISP-DM Data Understanding

- Data Understanding
  - Initial data collection
  - Data familiarisation
  - Identify data quality issues
  - First data insights

CRISP-DM Data Preparation

- Data Preparation
  - All activities to construct from the initial raw data the dataset(s) to feed into modelling tool(s)
  - Potentially many iterations
  - Tasks include
    - Table, record, attribute selection
    - Data Cleaning
    - Data restructuring and transformations

CRISP-DM BUSINESS UNDERSTANDING



epcc

THE UNIVERSITY OF EDINBURGH

Business Understanding

- Determine the Analytics Goal
- Formulate the question(s) to be answered.

Why?

Determine the Analytics Goal – Why?

- Without it, far more difficult to know if you have found anything useful.
- Helps define the data required and any preparation required for that data.
- Helps define the resources (people, their skills, machines, the techniques to apply) required too.
- Easy for the question to be forgotten about in the rush to collect and analyse the data
- Gives a focus – eg business relevance

Examples

- *"What customer behaviour characteristics correlate with mortgage early redemption?"*
  - Customer churn to identify early redeemers
  - Need to define what an early redeemer is
- *"Are product features correlated with mortgage early redemption?"*
- *"What is the effect of reliability on customer satisfaction?"*
  - The effect of roadworks on bus customers
- *"What are the characteristics of customers whose accounts are suspended within 6 months?"*
  - Mobile Telecommunications

Requirements/Objectives

- These state what you are looking for
  - Usually given by the Domain Expert
  - Very, very important
  - Define the direction of the study
  - Business questions help define the data to acquire

- These are a good starting point
  - Even in speculative studies
  - Easier to find if you know what you are looking for

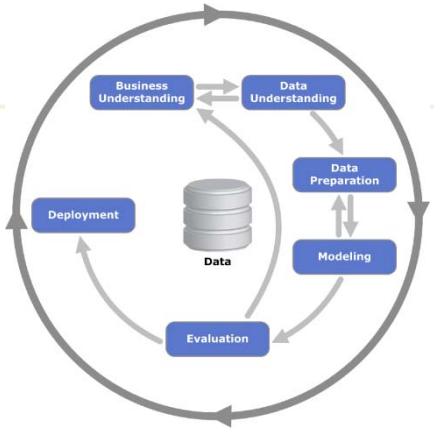
Document Requirements/Objectives

- Always document the requirements/objectives.

Why?

- To retain focus throughout the lifetime of the study
  - It's easy to forget
  - So if requirements/objectives change, document them
- The entire analytics team (data analysis, computing, domain) need to be able to review the requirements document. Why ?
  - So the data analysis experts can say if data can support the requirements, the tools, techniques to use
  - So the computational experts can say if the hardware/software can cope

CRISP-DM DATA UNDERSTANDING



epcc

THE UNIVERSITY OF EDINBURGH

Data Understanding

- Initial data collection
- Data familiarisation
- Identify data quality issues
- First data insights

Initial Data Collection

- **Is the data available to meet the analytic goals?**
- Published Data
  - Existing (publicly) available data: May need to buy it or perhaps it can be obtained for free
- Experimental Data
  - Data collected from an experiment specifically designed to test the hypothesis under investigation
  - Data collected from an experiment **NOT** specifically designed to test the hypothesis under investigation
- Operational Data
  - Data collected for some other purpose
    - eg. mobile phone records, web site logs, ….
  - "Big Data": when volumes of such data are immense

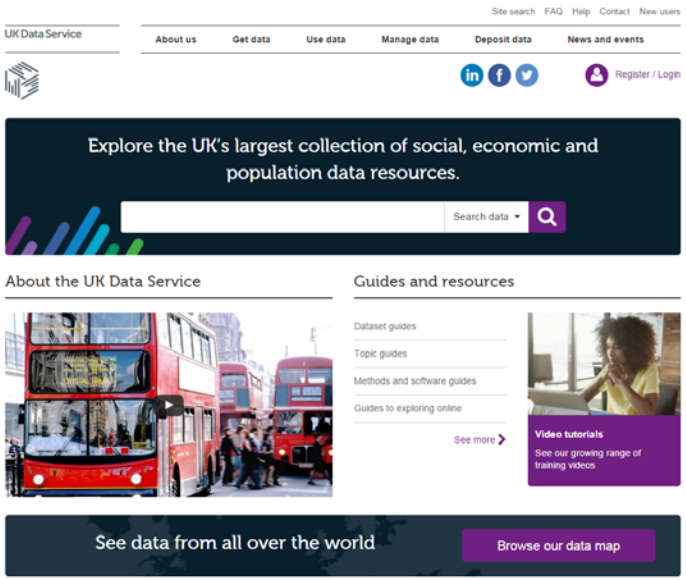
Data Source Categories

- **So there are essentially two categories**
- Data collected from an experiment specifically designed to test the hypothesis under investigation
  - Some references for experimental design if you want to find out more. Note this will NOT be covered in this course
    - Chapters 2 and 3 of Jason W. Osborne. 2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org.ezproxy.is.ed.ac.uk/10.4135/9781452269948
    - Chapter 8 of Elements of Statistics Daly, F.; Hand, D. J.; Jones, M.C; Lunn, A. D. and McConway, K. J. (1995). Elements of Statistics. Wokingham: Addison-Wesley Publishing Company. Usually available from http://www.open.edu/openlearnworks/mod/oucontent/view.php?id=18263&section=1.2.1
- Data from elsewhere

Data Sources

Data from elsewhere – scientific, governmental, …

- Public data repositories, e.g.
  - http://ukdataservice.ac.uk/
  - http://data.gov.uk/
- Institutional Repositories e.g.
  - http://datashare.is.ed.ac.uk/



- Discipline repositories e.g.
  - Sanger Institute (http://www.sanger.ac.uk/science/data)
  - UCI (See http://mlr.cs.umass.edu/ml/)
- APIs e.g.
  - Twitter (https://dev.twitter.com/docs/api/streaming)
  - ….

Data Sources (2)

Data from elsewhere – business/commercial/operational

- Transaction data
  - eg. Mobile phone records, mortgages, …
- Sensor data
  - eg. Fitbits, …
- Web logs

Data Sources (3) - Data Fusion

- When no single data set has all the data you need
- Involves fusing (merging) data from different sources on a common field
- Requires an appropriate common field (key)

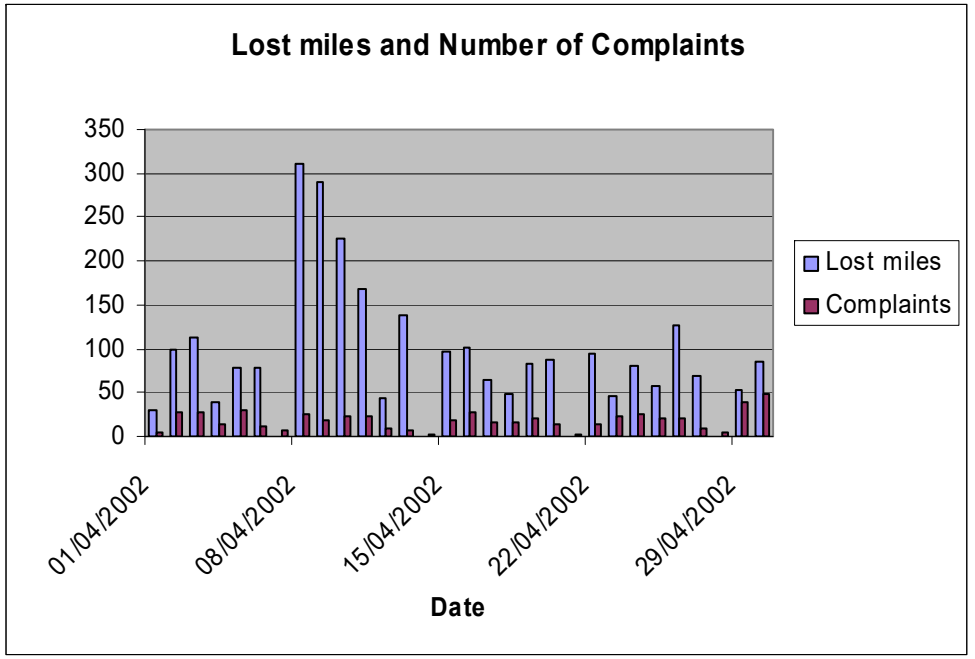


Figure 5: Lost miles and Number of Compliants per day

Data Fusion Example - FirstDIG

- Data sources were
  - Customer contact correspondence
  - Vehicle Mileage
  - Ticket Revenue
  - Schedule Adherence
- Systems at various company sites, on differing platforms in different databases (eg. SQL, ODBC, COBOL,..)
- Objectives/Business Questions
  - the effect of lost mileage on revenue, where lost mileage is due to activities such as road works and breakdowns;
  - the effect of lost mileage on the number of complaints;
  - the effect of reliability on revenue;
  - the effect of reliability on complaints received.

Pitfalls of Data Fusion

- With data fusion – always have to be careful about using data that was not collected specifically for your analysis.
- For example,
  - there could some bias in the data so always have to be wary.
  - the fields used to fuse the data need to be sensible and have the same meanings across the different datasets.
  - Timestamp issues eg. time zones
  - ….

Pitfalls of Data Fusion (cont.)

- Public Perception: Care.data
  - *"supposed to link all NHS data about all patients together into one giant database"*
  - *"The care.data project was promoted in two ways: we will use your data for lifesaving research, and we will give it to the private sector for commercial exploitation, creating billions for the UK economy. … public support public research, but are nervous about commercial exploitation of their health data."*

    http://www.theguardian.com/society/2014/feb/21/nhs-plan-share-medical-data-save-lives
  - See also http://www.wired.co.uk/news/archive/2014-02/07/a-simple-guide-to-care-data
- Cancelled: July 2016
  - http://www.computerweekly.com/news/450299728/Caldicott-review-recommends-eight-point-consent-model-for-patient-data-sharing

Data Relevance

- May not need any data at all
  - Do the objectives/ business questions require all the proposed data ?
  - Do the objectives/ business questions really need data for them to be answered?
- May not be able to answer objectives/ business questions with available data
- If suitable extract cannot be identified then analytics team must consider
  - Stopping the study, or
  - Changing objectives/ business questions

Other Resources

Other than data, what else do you need ?

Other resources

- Do you need domain expertise to understand the data ?
  - Do you need business knowledge to understand the processes by which the data has been collected and previously manipulated?
  - Eg. certain non-obvious fields used to indicate an early redeemer
- How much staff effort is needed for the project or is available?
  - Do your goals match the amount of effort that is available?
- Do you have the statistical knowledge ?
- Do you have the computational resources ?
- What software do you need ?

Other resources (cont.)

- Do you have the storage to handle the data volumes ?
- Do you need an analytic sand box where you can safely and efficiently manipulate the data?
- Do you have/need a data warehouse ?

Metadata

- If getting data from elsewhere – need Metadata
  - Do you have a codebook or dataset description ?
    - Eg. http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.names
    - http://archive.ics.uci.edu/ml/datasets/Abalone
  - Do you have sufficient expertise to understand the data fields and structure?
    - What's the Data Model of the data source?
      - RDBMS, COBOL, CSV, flat file, multiple formats, …
    - If existing documentation not available eg. legacy systems then document it
- Review the data model to identify an appropriate extract
  - Will the extract support the analytic goals?

Metadata – processes

- Important to understand the processing the data has undergone
  - Sometimes not available in written form or at all
    - Person left company
    - Data acquired through company acquisitions
  - If not written down, need to document as much as possible
- Involve relevant data processing staff from data source organisation
  - Ideally involve throughout study
  - Help identify false results - artefacts
- Analytics results MUST always be reviewed with knowledge about processing the data has undergone
- May not be able to answer business questions
  - Study stops?
  - Change business questions?

Data Transfer & Storage

- Storage
  - Where will the data be put? A Data warehouse?
- Data volume
  - may be too big for easy transfer/transportation (networking, snail mail)
- Data format
  - may be held in a non-useful format, eg. EBCDIC to ASCII

ETL vs. ELT Data

- ETL = Extract Transform Load
- ELT = Extract Load Transform

ETL

- ETL – Extract Transform Load
  - Traditional approach
  - Subset of operational data extracted and transformed prior to its loading into data warehouse.
  - Extract and transformations applied are based on the analytic goals.
  - Potentially reduced turnaround time for analytics goals
  - Data warehouse is smaller
- BUT
  - Transforming data before load means possible loss of important information eg. outliers.
    - In fraud detection, it's the outliers that are important.

ELT

- ELT – Extract Load Transform
  - exact copy of the operational data loaded into data warehouse without reference to analytic goals
  - future proofs data warehouse since as analytic goals change, data warehouse is still usable
  - Very large data ware house
- BUT
  - Potentially longer turnaround times for analytic goals

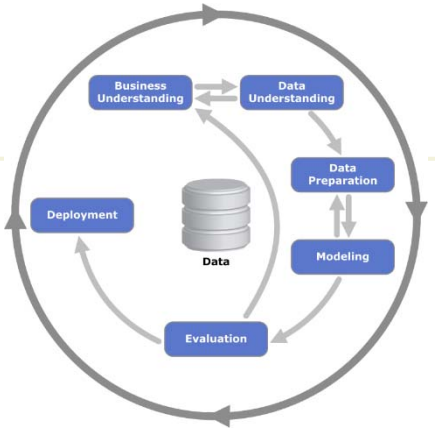
Important Considerations

- Things to consider when transferring data
- Privacy, legal, anonymisation
  - Does any of data need to be stored in a secure environment?
    - Analytic sandboxes
  - Will anonymisation affect analytic goals?
- Obtaining commercial data within a project
  - Always a delay in getting the data
    - things sometimes get hidden etc.
  - Need to have somebody who understands the domain…

Other Resources

- Domain expertise to understand the data ?
  - Business knowledge to understand how data has been collected and manipulated.
- Technical Skills
  - Computing
  - Data Analysis
- Computational resources
  - hardware
  - software
  - storage

CRISP-DM DATA PREPARATION



epcc

THE UNIVERSITY OF EDINBURGH

Data Preparation

- All activities to construct from the initial raw data the dataset(s) to feed into modelling tool(s)
- Potentially many iterations
- Tasks include
  - Table, record, attribute selection
  - Data Cleaning
  - Data restructuring and transformations

“It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003).”

Hadley Wickham, Journal of Statistical Software, August 2014, Volume 59, Issue 10.

“Literally hundreds of practicing data miners and statistical modelers, most of them working at major corporations supporting extensive analytical projects, have reported that they spend 80% of their effort in manipulating the data so that they can analyze it!”

Dan Steinberg 2013, ““How Much Time Needs to be Spent Preparing Data for Analysis?”
http://1.salford-systems.com/blog/bid/299181/How-Much-Time-Needs-to-be-Spent-Preparing-Data-for-Analysis

"Most statistical theory focuses on data modeling, prediction and statistical inference while it is usually assumed that data are in the correct state for data analysis. In practice, a data analyst spends much if not most of (their) time on preparing the data before doing any statistical operation. It is very rare that the raw data one works with are in the correct format, are without errors, are complete and have all the correct labels and codes that are needed for analysis."

Edwin de Jonge and Mark van der Loo,

"An introduction to data cleaning with R"

Statistics Netherlands Discussion Paper, 2013, 13

(Available from http://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf)

WHY SO MUCH TIME?

Data Quality

- Data quality
  - crucial to successfully achieving analytic goals
- Data must be
  - Consistent
  - Reliable
  - Appropriate
  - ….
- Need to examine structure and content to
  - Fix missing fields
  - Fix/remove incorrect values
  - Remove unnecessary information
- May also need to
  - Select relevant data; generate new fields

Formats

- Is the data in a format you can load into your preferred visualisation/analytic/modelling tool?
  - Binary
  - Character encoding (e.g. EBCDIC, ASCII, UTF-8, …)
  - CSV – comma separated values
  - FWF – fixed width format
  - Tab-delimited
  - Is there a header row?
  - Is it unstructured data ? E.g. free text or non-uniformly formatted i.e. different formats on different lines

Anomalies

- Remove anomalous (incongruous) values, for example
  - 99 year bank loan but bank only has 30 year loan products
  - amount of loan equals zero
  - A person's age may be negative
  - A person's age may be very large
  - An under-age person may have a driving license
  - …

    But…
  - **BEWARE**: sometimes those are the analytic goal e.g. fraud detection

Artefacts

- Remove artefacts a.k.a data leakage
  - i.e. proxies for the feature of interest in an analytic goal
  - eg. early redemption –a person who pays off a bank loan early
    - a code in a particular field is used to signify an early redeemer is entered after the person has paid off the loan early
    - If trying to predict early redeemers then this field and code has to be removed.
  - Amazon Case Study: Big Spenders
    - Kaggle competition to predict high spending customers using transaction data
    - The winning model was “free shipping = TRUE”
    - But customers only get free shipping if they are big spenders in the first place.
    - The model is therefore useless – do not know the value of free shipping until the customer has made the purchase.

Confounders

- Need to deal with confounding variables
  - Similar'ish to data leakage
- Usually described as - a relationship between two variables that is actually attributable to the relationship both these variables have with a third variable.
  - For example: shoe size and reading ability
    - See http://www.ma.utexas.edu/users/mks/statmistakes/causality.html
- Correlation ≠ Causation

Confounding example

"a highly publicized study pronounced that left-handed people did not live as long as right handed people (Coren and Halpern, 1991). In one part of the study, the researchers had sent letters to next of kin for a random sample of recently deceased individuals, asking which hand the deceased had used for writing, drawing, and throwing a ball. They found that the average age of death for those who had been left-handed was 66, while for those who had been right-handed it was 75.

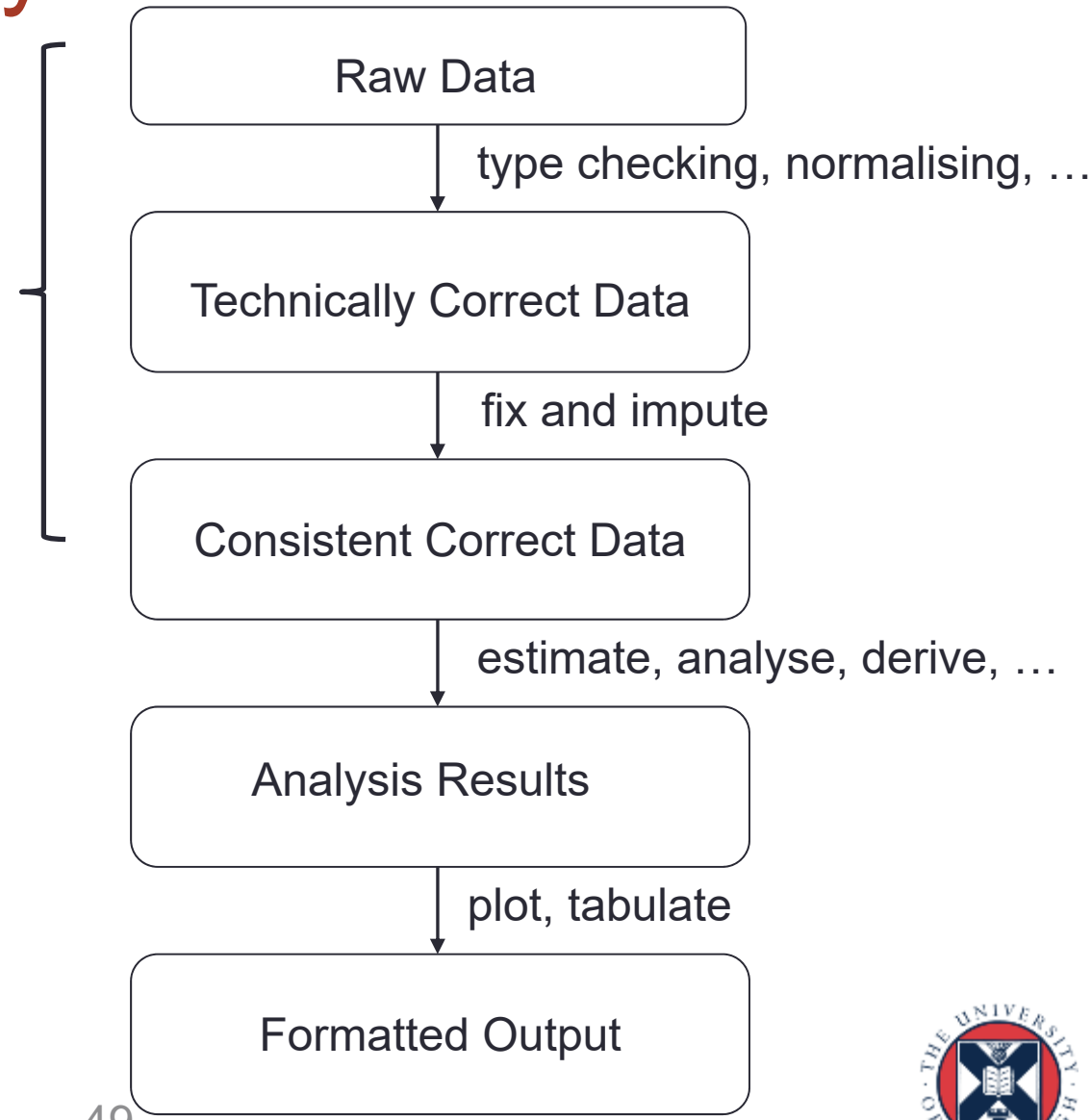
What the researchers failed to take into account was that in the early part of the 20th century many children were forced to write with their right hands, even if their natural inclination was to be left-handed. Therefore, people who died in their 70s and 80s during the time of this study were more likely to be right-handed than those who died in their 50s and 60s. The confounding factor of how long ago one learned to write was not taken into account. A better study would be a prospective one, following current left- and right-handers to see which group survived longer, but you can imagine the practical difficulties of conducting such a study. The participants could well outlive the researchers."

"J.Utts, R Heckard, "Statistical Ideas and Methods", Thomson Brooks/Cole, 2006

DATA CLEANING TECHNIQUES

Statistical Analysis Value Chain

- A way to visualise the data flow when cleaning and preparing data
  - From Statistics Netherlands Discussion Paper, 2013, 13



Raw Data to Technically Correct Data

A data set is a collection of data that describes attribute values (variables) of a number of real-world objects (units).

With data that are technically correct, we (have) *a data set where each value:*

1. can be directly recognized as belonging to a certain variable;

2. is stored in a data type that represents the value domain of the real-world variable.

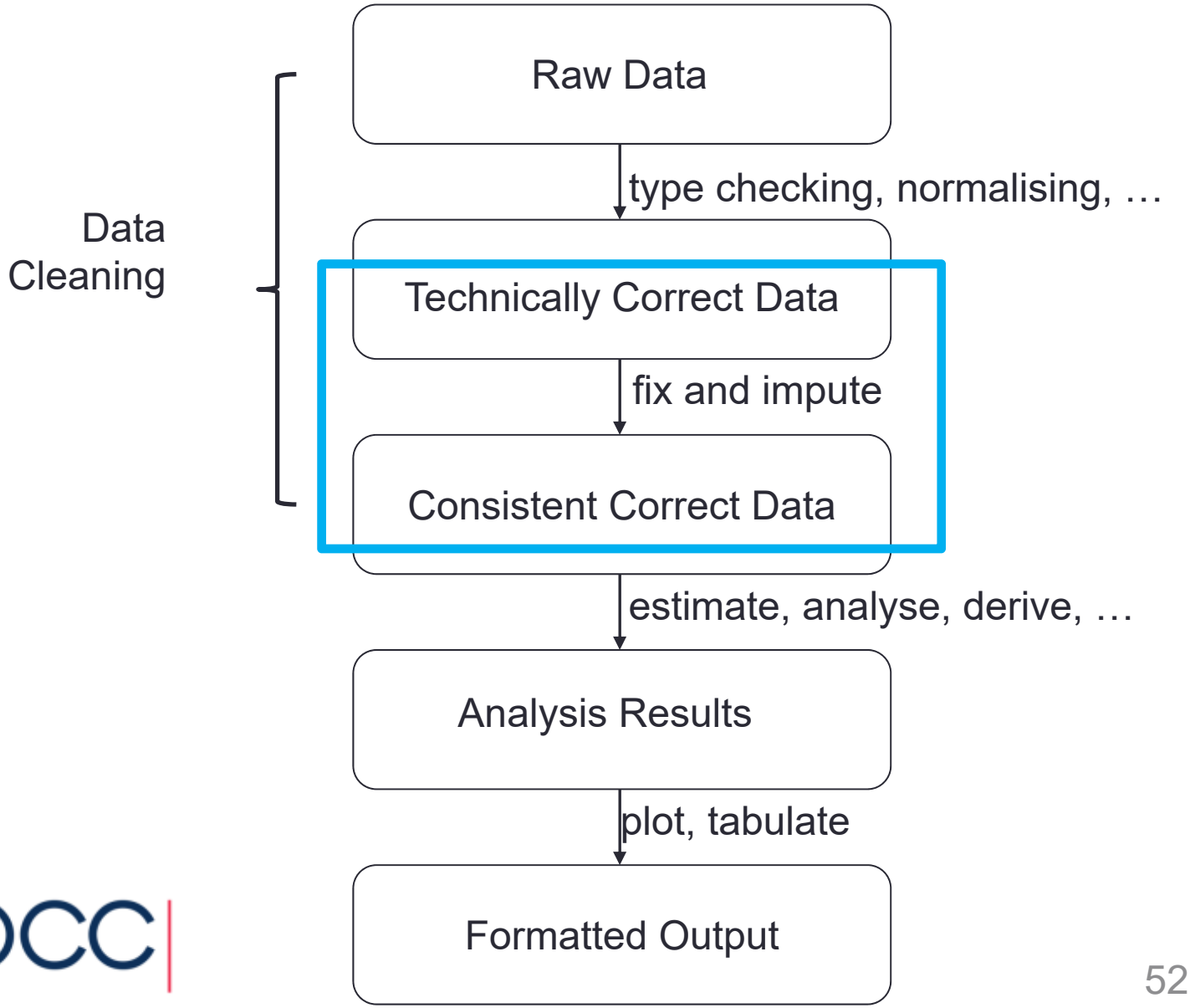
In other words, for each unit, a text variable should be stored as text, a numeric variable as a number, and so on, and all this in a format that is consistent across the data set.

Edwin de Jonge and Mark van der Loo,
"An introduction to data cleaning with R"
Statistics Netherlands Discussion Paper, 2013, 13
(Available from http://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

Raw to Technically Correct Techniques

- Format: Binary, CSV, FWF, XML, …
- Character Encoding: EBCDIC, ASCII, UTF-8, …
- Type conversion: also known as coercion
- String normalisation: eg space removal
- Approximate string matching: two types
  - Extraction of substrings
  - Use string distance to define the difference
- Categorical variables: Recoding
- Date conversion
- Transforming non-uniformly formatted (unstructured) data
  - selecting lines that contain data
  - splitting out required data fields
  - standardise rows with same number/type of fields

Statistical Analysis Value Chain



Technically Correct to Consistent Data

- Consistent Data are:
  - technically correct data that are fit for statistical analysis.
  - In which missing values, special values, (obvious) errors and outliers are either removed, corrected or imputed.
  - consistent with constraints based on real-world knowledge about the subject that the data describe
- Consistent Data have:
  - In-record consistency
    - i.e. no contradictory information is stored in a single record
  - Cross-record consistency
    - i.e. statistical summaries of different variables don't conflict with each other
- Cross-dataset consistency
  - the dataset being analyzed is consistent with other datasets pertaining to the same subject matter

Process towards Consistency

Comprises three steps:

1. **Detection**:
   - establish which constraints are violated eg. age variable is constrained to non-negative values
2. **Selection**:
   - of field (s) causing inconsistency. Sometimes can be difficult eg. marital status of a child. Is the age field wrong? Or is the marital status field wrong? Are both fields wrong?
3. **Correction**:
   - of the fields deemed erroneous
     - For example, is the record removed ?

ERROR DETECTION

Missing Values

- Missing values
  - Need to be dealt with prior to analysis
  - In R a missing value is represented by NA (Not Available)
- Do not confuse a missing value with a default or category
  - Some numerical software may silently impute a value leading to erroneous results
  - Common mistake:
    - if a category variable has a value “unknown”, do not confuse this with NA or a missing value. For example, place of birth can have a category “unknown”, but this is not a missing value. NA means there is no information to determine if place of birth is known or not
- Once missing values are detected you need to decide what to do about them
  - For example in R, na.rm option

Special Values

- Most programming and scripting languages have exceptions to normal values in a type. For example, in R

  NA – Not Available. a placeholder for a missing value

  NULL - the empty set from mathematics. NULL is special since it has no class (its class is NULL) and has length 0 so it does not take up any space.

  Inf – Infinity

  NaN – Not a Number. generally the result of a calculation of which the result is unknown, but it is surely not a number

- Calculations involving special values often result in special values
  - Therefore it is desirable to handle special values prior to analysis

Outliers

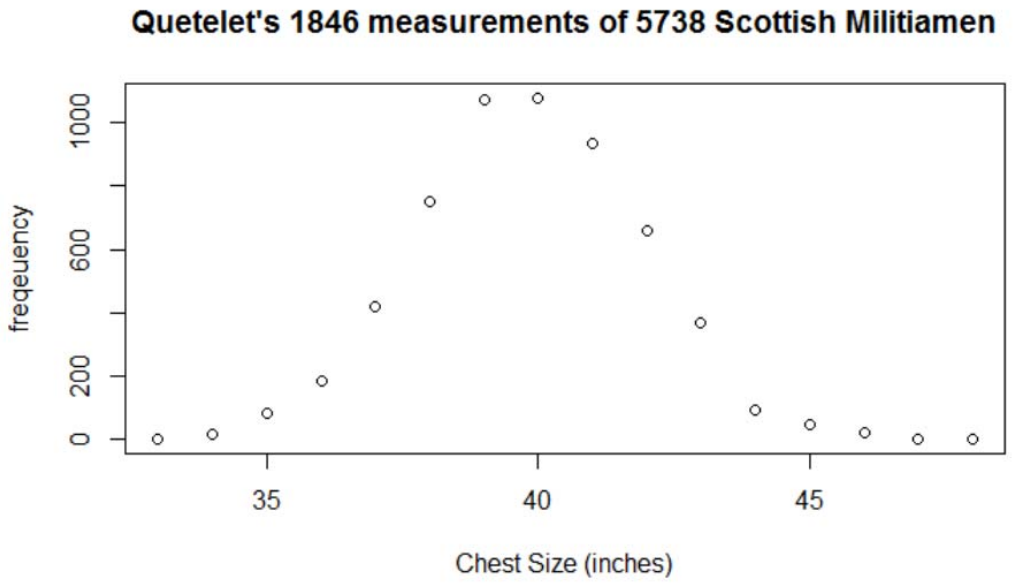
- Definition
  - an outlier in a data set is an observation (or set of observations) which appear to be inconsistent with that set of data. Barnett & Lewis, Outliers in statistical data. Wiley, New York, NY, 3rd edition, 1994
- Note: Outliers do not equal errors.
  - They should be detected, but not necessarily removed. Their inclusion in the analysis is a statistical decision.

Outliers

- For more or less unimodal and symmetrically distributed data, boxplots (also known asTukey's box-and-whisker plots) for outlier detection is often appropriate.
  - an observation is an outlier when it is larger than the so-called whiskers of the set of observations
  - The upper whisker is computed by adding 1.5 times the interquartile range to the third quartile and rounding to the nearest lower observation. (the factor 1.5 is arbitrary and can altered)
  - The lower whisker is computed likewise
  - Interquartile range is difference between 1st and 3rd quartile

Outliers: Unimodal, symmetric

- A data set is unimodal when there is a single mode.



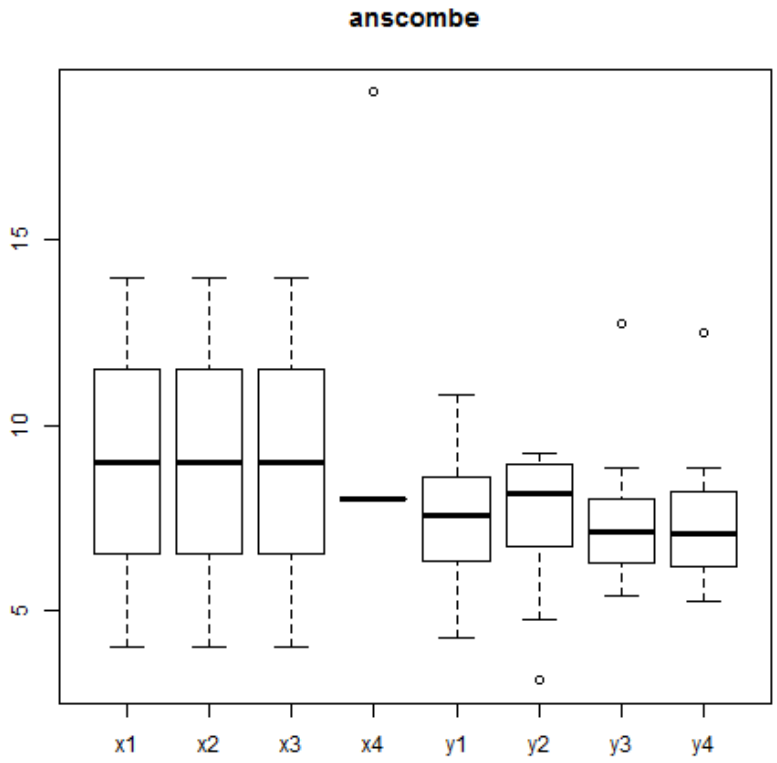
- Symmetric distribution – left side is a mirror of right side.
- Symmetric distributions can multimodal
- Skewed : a lack of symmetry

Outlier Example

Box and whisker plot (a.k.a. boxplot)

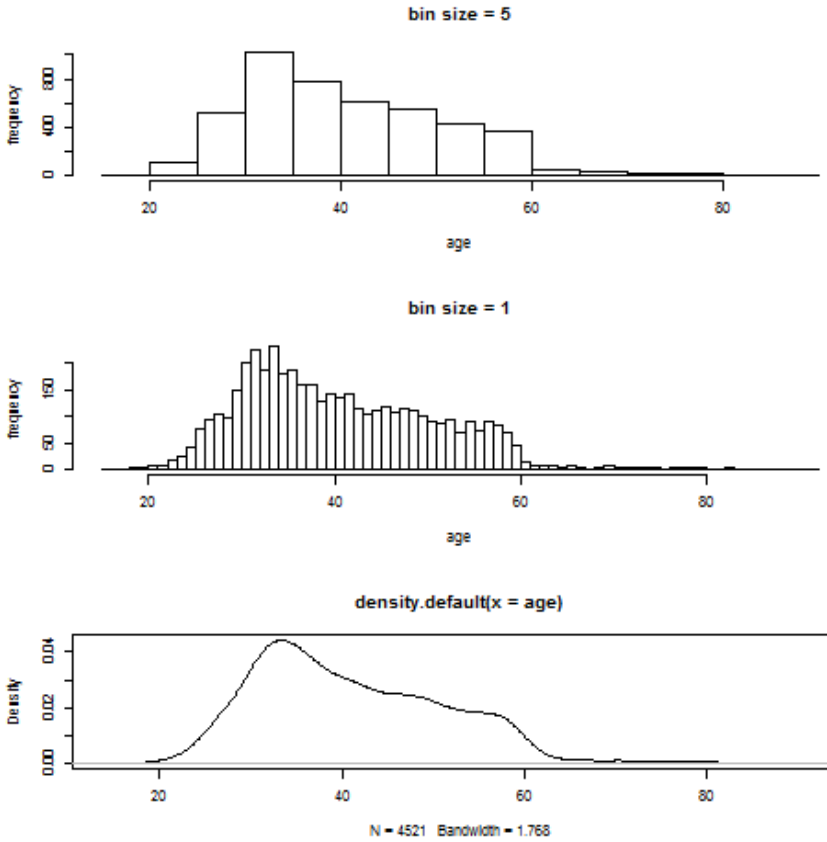
- box indicates the interquartile range and the median
- the whiskers are represented at the ends of the box-and-whisker plots
- outliers are indicated as separate points above or below the whiskers.

Fails when data is skewed, categorical



Outliers: Distribution of continuous variable

- Histograms
  - Note: shape of distribution is affected by bin size



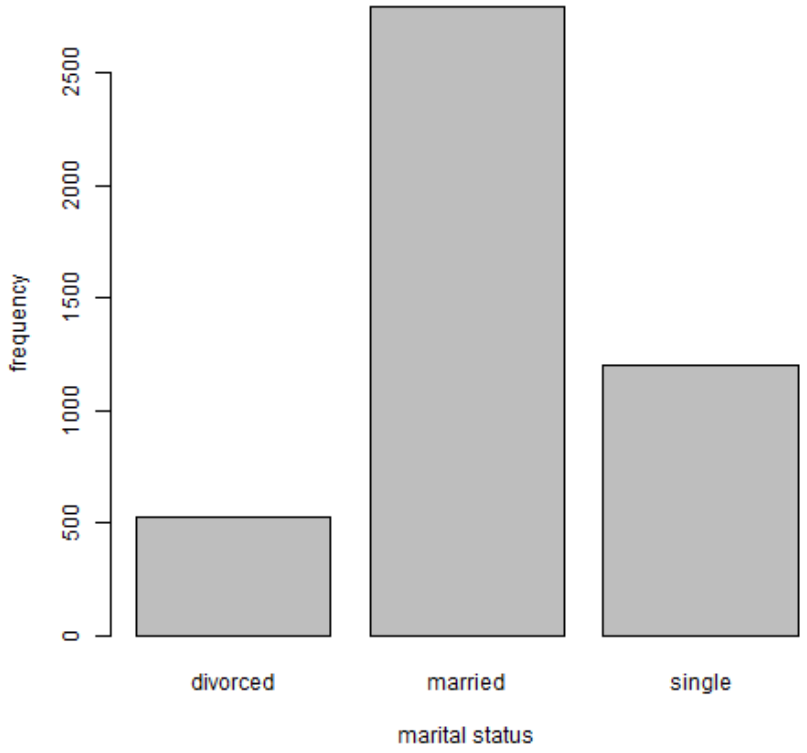
- Density plots
  - Area under curve =1
  - Smoothed curve
  - Be wary of missing fine structure ie. compare with bin size =1

epcc

THE UNIVERSITY OF EDINBURGH

Distribution of a single categorical variable

- Frequency of values in the variable (field)
- Looking for
  - Skew
  - Distribution
  - Outliers
- Where's widowed?
  - See the data definition
  - http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
  - "note: 'divorced' means divorced or widowed"



Obvious Inconsistencies

- Occurs when a record contains a value or combination of values that cannot correspond to a real-world situation
  - Eg. under-age driving license holder
- Some are easy to check
  - Non-negative
- But complexity increases as number of variables involved increases
  - Typically need to define set of rules and apply them (eg. R editrules package)

EXPLORATORY DATA ANALYSIS (EDA)

epcc

THE UNIVERSITY OF EDINBURGH

Exploratory Data Analysis (EDA)

- *"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…"*
- *"EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect"*
- *"…there is a large collection of statistical tools that we generally refer to as graphical techniques. These include: scatter plots, histograms, probability plots, residual plots, box plots, block plots."*
- *"The EDA approach relies heavily on these and similar graphical techniques. Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use."*

NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook/, Oct'2014.

ANSCOMBE'S QUARTET

Why summaries are not enough?

- *"If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data."*

- *"Quantitative statistics are not wrong per se, but they are incomplete."*
  - NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, Oct'2014

Anscombe's Quartet

- Four datasets with the same mean, variance, correlation, regression line, ….

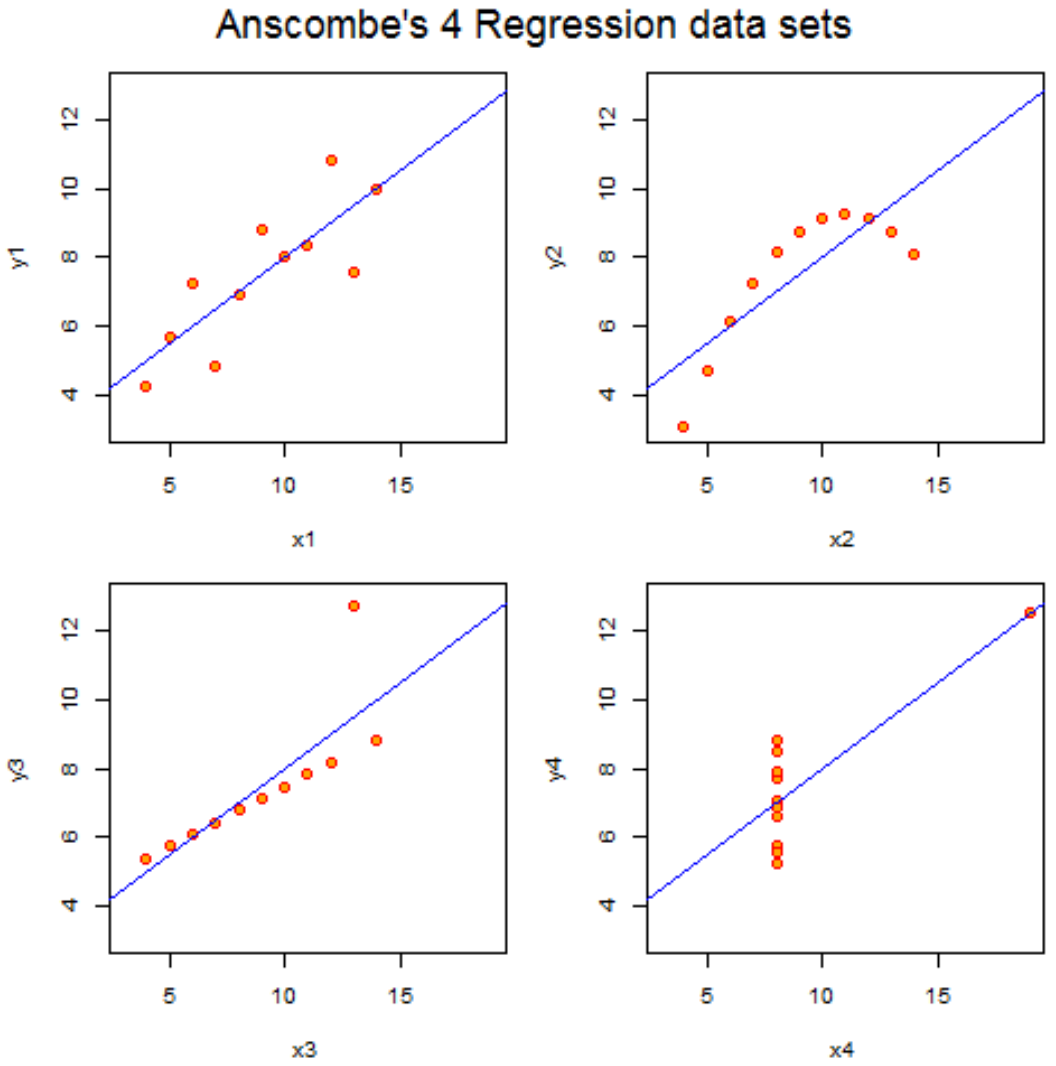
- Anscombe, Francis J. (1973) "Graphs in statistical analysis", American Statistician, 27, 17–21

Anscombe's Quartet

For all 4 datasets

- Number of (x,y) pairs = 11
- Mean of X = 9
- Standard deviation of X = 3.32
- Mean of Y = 7.5
- Standard deviation of Y = 2.03
- Slope of regression line = 0.5
- Intercept of regression = 3
- Correlation coefficient = 0.82

Anscombe's Quartet Plots



Anscombe's Quartet Conclusions

- *"Conclusions from the scatter plots are:*
  - *- data set 1 is clearly linear with some scatter.*
  - *- data set 2 is clearly quadratic.*
  - *- data set 3 clearly has an outlier.*
  - *- data set 4 is obviously the victim of a poor experimental design with a single point far removed from the bulk of the data 'wagging the dog'."*

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, Oct'2014

Statistics Refresher (i)

Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample Median

If a data set $x_1, x_2, \cdots, x_n$ is re-ordered as $x_{(i)}$, $i = 1, 2, \ldots, n$, where

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \quad (1)$$

then the median is defined by

$$m = x_{\left(\frac{1}{2}(n+1)\right)}$$

Statistics Refresher (2)

Sample Quartiles

If a data set $x_1, x_2, \cdots, x_n$ is re-ordered as $x_{(i)}, i = 1, 2, \ \ldots, n$, where

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} (1)$$

then the lower sample quartile (a.k.a. 1st quartile) is defined by

$$q_L = x_{(\frac{1}{4}(n+1))}$$

and the upper sample quartile (a.k.a 3rd quartile) is defined by

$$q_U = x_{(\frac{3}{4}(n+1))}$$

Statistics refresher (3)

Sample Standard Deviation

A measure of dispersion in a sample $x_1, x_2, \cdots, x_n$
with sample mean as $x_{(i)}$ is given by the sample standard deviation s,
where s is obtained by averaging the squared residuals, and taking the square root of that average

$$s = \sqrt{\frac{r_1^2 + r_2^2 + \cdots r_n^2}{n-1}}$$

$$= \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Regression

See Linear Regression Lecture later in the course

Statistics refresher (4)

Pearson correlation product-moment coefficient

The Pearson correlation coefficient r from bivariate data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, where the means of the x-values and the y-values are $\bar{x}$ and $\bar{y}$ and their standard deviations are s_X and s_Y is calculated

as follows $$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X}\right) \left(\frac{y_i - \bar{y}}{s_Y}\right)$$

When $r = +1$, x and y have an exact straight line relation with +ve slope
When $r = -1$, x and y have an exact straight line relation with -ve slope
When $r = 0$, the two variables are unrelated

Using EDA: A guide

- Typically, as well statistical summaries, want to

1. Plot the distribution of each variable (column/field)
   - Look for outliers, unexpected values, unexpected distributions
   - Decide whether our or not the records associated with anomalous values need to be removed from the study. Perhaps such records need a separate investigation since they may highlight an issue with the business process. (Eg. the field may be predominantly one value when a range was expected)
2. Examine the relationship (correlation) between pairs of variables (Especially, those that relate directly to a feature(s) of the analytic goals)
   - Helps with identifying artefacts, data leakage, proxies
3. Analyse each variable over time (eg. is there a sequence of values that a variable might have)

After Cleaning

- After cleaning, is there enough appropriate data to meet the analytic goals ?
  - Data imputation

Extremely important side benefit of cleaning and preparation

- Helps develop a better overall view of the data

HPC in Data CLEANING & PREPARATION

HPC IN DATA CLEANING & PREPARATION

Is HPC needed?

- Dependency on
  - Volume – how much data needs to be cleaned
    - Operations on elements are easily parallelised
  - Complex – how complex is the cleaning task
    - Does it involve bringing together a variety of data sources
  - Frequency
    - Number of times cleaning operations will be repeated
- Trade-offs
  - Programming overhead
  - Elegance of computation, 80:20
  - ELT vs ETL,
  - …

WRAP-UP

- Questions?

epcc

THE UNIVERSITY OF EDINBURGH