COMP4030

DATA MODELLING AND ANALYSIS

Lecture 6: Data Visualisation and Pre-Processing

LECTURE OUTLINE

1. Summary
2. Lecture Outcomes
3. Data Visualisation
4. Pre-processing steps
    1. Data selection and integration
5. Cleaning data
    1. Missing values
    2. Noise and outliers
6. Transforming data
    1. Aggregation; generalisation
    2. Changing scales; normalising

PREVIOUSLY ON DMA...

DATA MODELLING

System Question

Should I play tennis today?

L2/L3

Model

Input (i.v.)

	V_1	V_2	V_3	V_4
s1	x_{11}	x_{12}	x_{13}	x_{14}
s2	x_{21}	x_{22}	x_{23}	x_{24}

L4/L5/L6

Most models as functions

{S,Q,M}

f(X)=Y

Output (d.v.)

y_1
y_2

Model Input Output

$f(x_{11},x_{12},x_{13},x_{14}) = y_1$

$f(x_{m1},x_{m2},\ldots,x_{mn})=y_m$

PREVIOUSLY IN DMA

DATA VS KNOWLEDGE

Data (N.O.I.R)

Data Exploration (L5/6)

Johannes Kepler

- Essential to understand data
- Determined by type of data

- Large amounts
- No structure
- Easy to obtain
- Can't predict

Data Preprocessing (L6/7)

- Minimal
- Highly structured
- Hard to obtain
- Allows prediction
- Correct, general, useful, comprehensive and new

Tycho Brahe

Data Mining/Learning (L08-L10)

Knowledge

MODELLING DATA

DATA ANALYSIS: DATA TYPES

Provides:	Nominal	Ordinal	Interval	Ratio
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode, Median		✔	✔	✔
The "order" of values is known		✔	✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

- Depending on your model, the same data might belong to different classes.
  - For example: Age, Time

LECTURE OUTCOMES

- At the end of this lecture, you should be able to answer these questions:
  - What is data visualisation?
  - Why is it useful? When is it useful?
  - What are the main graphs we can use to describe data?
  - What pre-processing steps should you take?
  - How and with what purpose do clean data?
  - How and with what purpose do you transform data?

DATA VISUALISATION

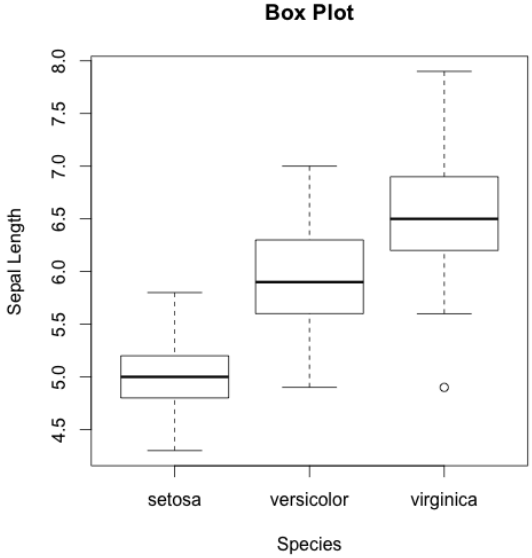
DEFINITION AND PURPOSE

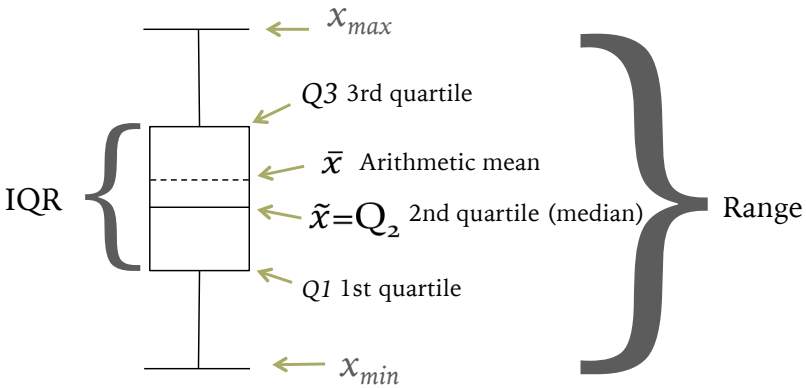
- Visualisation: Any technique for creating images, diagrams or animations **to communicate a message**.
- There are many types (scientific, educational, etc.).
- Data visualisation: creation/study visual representation of data.
- Primary goal: **to communicate information clearly and efficiently** via statistical graphics, plots and information graphics.
- Makes complex data more accessible, understandable and usable.
- There are many types of visualisations you can use. The correct choice will depend on **your data** and **your "message"**.

DATA VISUALISATION

VISUALISING CHARACTERISTIC MEASURES: BOXPLOTS

- Box plots are an effective visualisation technique to depict several characteristic measures



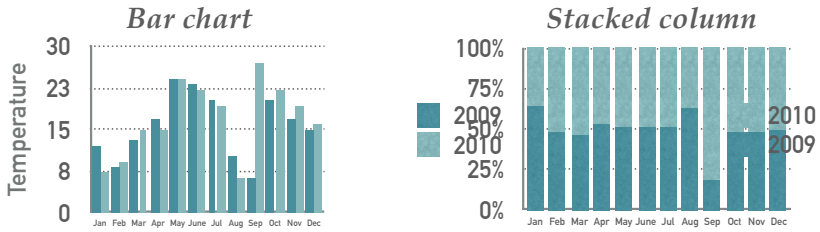


VISUALISING DATA

VISUALISING DISCRETE TRENDS OR VALUES: BARCHARTS

- Bar charts:
  - Wrong impressions of relative proportions can be generated if the y-axis does not start at zero

	Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sep	Oct	Nov	Dec
2009	12	8	13	17	24	23	20	5	6	20	17	15
2010	7	9	15	15	24	22	19	27	27	22	19	16
2011	38	37	36	39	35	20	15	12	13	10	12	37



VISUALISING DATA

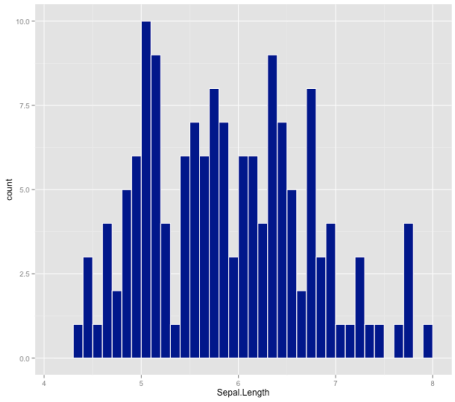
VISUALISING DISTRIBUTIONS: HISTOGRAMS

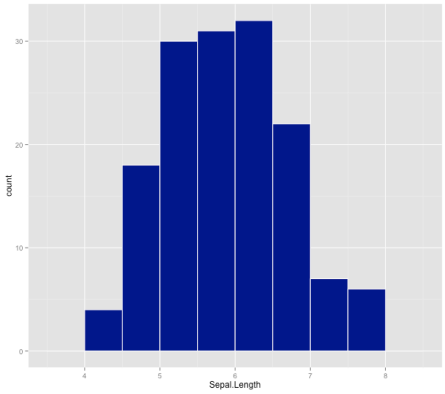
- Histograms
  - Bar char measuring frequency of appearance of values.
  - Representation of the distribution of numerical data.
  - If data is continuous, you might have to create intervals between values.
    - This process is called binning
    - We will see how to bin data manually and using R in this lecture.

VISUALISING DATA

VISUALISING DISTRIBUTIONS: HISTOGRAMS

- Both of these graphs represent the same data. But:
  - One has bins (intervals) of 0.1 between values
  - The other has bins of 0.5
- Granularity is important!



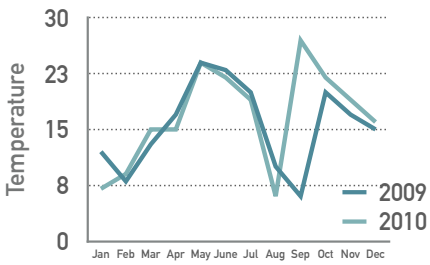


VISUALISING DATA

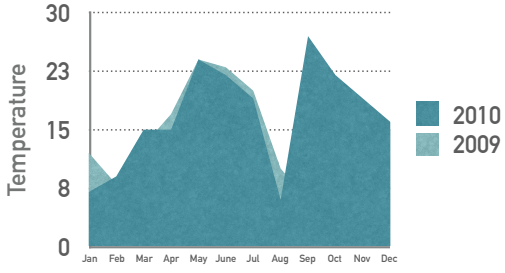
VISUALISING CONTINUOUS DATA: LINE AND AREA CHARTS

	Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sep	Oct	Nov	Dec
2009	12	8	13	17	24	23	20	5	6	20	17	15
2010	7	9	15	15	24	22	19	27	27	22	19	16
2011	38	37	36	39	35	20	15	12	13	10	12	37

Line chart



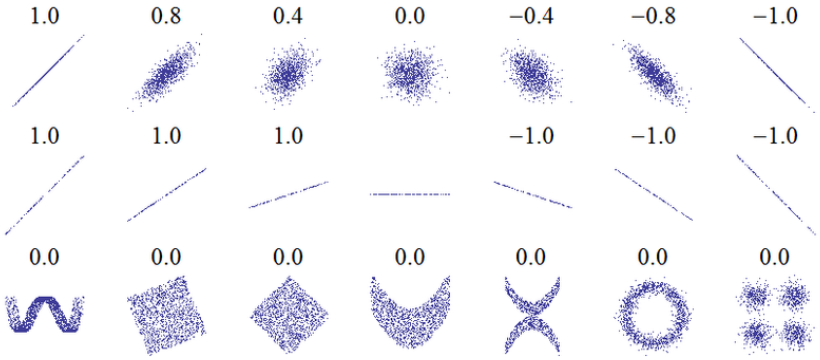
Area chart



DATA VISUALISATION

VISUALISING RELATIONSHIPS BETWEEN VARIABLES: CORRELATION

- Relationships between two variables
  - Visualisation: scatterplot
  - Pearson correlation coefficients for different datasets
- Nonlinear relationships are not captured by this:



The middle correlation coefficient is undefined because the variance along the y-axis is zero.

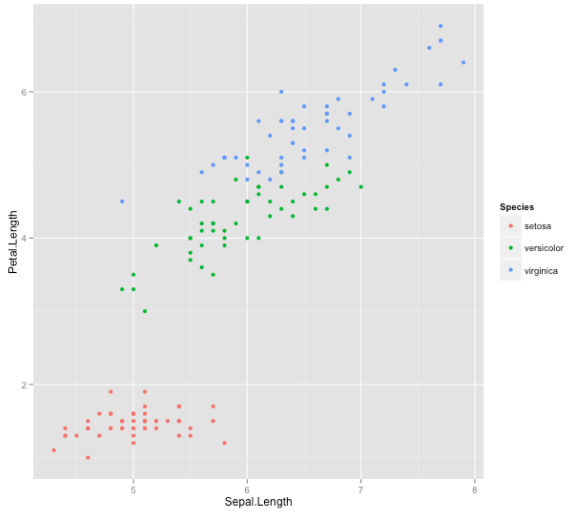
'Correlation': linear

'Association': more general

VISUALISING DATA

VISUALISING DATA RELATIONSHIPS: SCATTERPLOTS

- Useful for revealing relationships between pairs of variables



VISUALISING DATA

VISUALISING RELATIONSHIPS BETWEEN SEVERAL VARIABLES: SCATTERPLOT MATRIX

- But, what if we have 3 or more variables?
  - For example:
    - gender/age/mark
    - food/price/calories
- What visualisation techniques can we use to reveal the relationships between the variables?

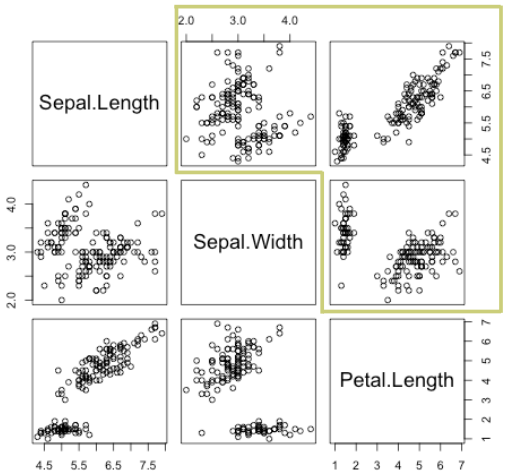
VISUALISING DATA

VISUALISING RELATIONSHIPS BETWEEN SEVERAL VARIABLES: SCATTERPLOT MATRIX

- Scatterplot matrices:
  - depict scatterplots for all possible pairs of distinct variables
  - if we have n variables then there will be $(n\text{-}1)^2$ scatterplots
  - depict the scatterplots in a 2D grid
  - essential property: adjacent graphs must have one axis (i.e. variable) in common
  - the lower triangle of the 2D matrix contains the same scatterplots as the upper triangle, but the axes are flipped.
    - lower triangle is often not included

VISUALISING DATA

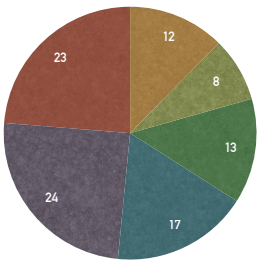
VISUALISING RELATIONSHIPS BETWEEN SEVERAL VARIABLES: SCATTERPLOT MATRIX



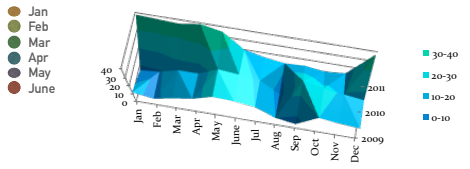
VISUALISING DATA

MISCELLANEOUS REPRESENTATIONS

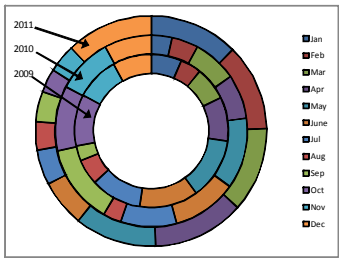
Pie chart (2009)



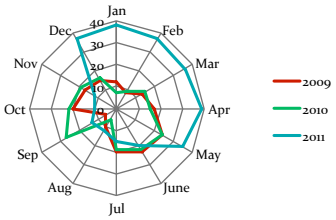
3D surface chart



Doughnut chart



Spider chart

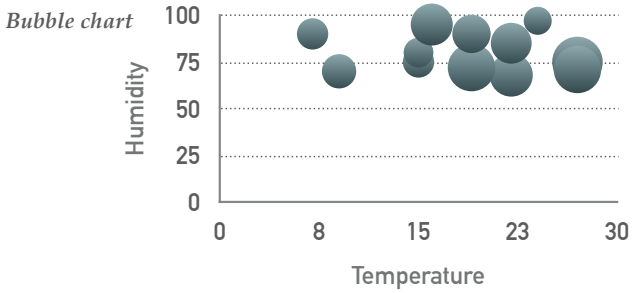


VISUALISING DATA

3D REPRESENTATIONS

- What if each data point needs to represent three possible attributes?
  - Two dimensions can be represented spatially, and the third can be represented in different ways, e.g.
    - colour, size, shape of markers

	Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sep	Oct	Nov	Dec
Temperature	7	9	15	15	24	22	19	27	27	22	19	16
Humidity	90	70	75	80	97	68	72	75	71	85	90	95
Cloudy Days	10	12	10	9	8	19	23	26	23	17	15	18



The number of cloudy days in a month is depicted by the size of the sphere

DATA VISUALISATION

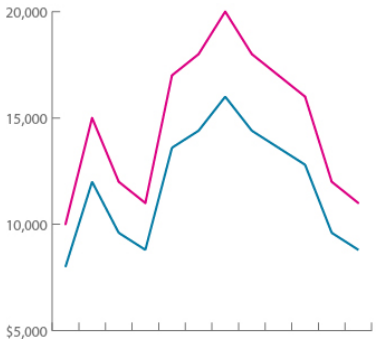
HOW, WHEN AND WHY?

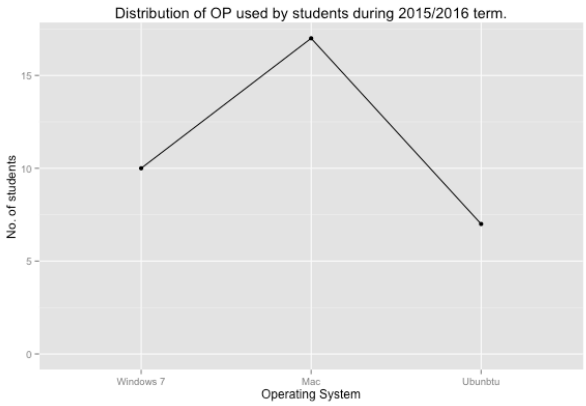
- The biggest problem: to know when, how, and why to use it.
  - When?
  - How?
  - Why?
- If you are not careful, you might end up with graphs that are…

DATA VISUALISATION

IF YOU ARE NOT CAREFUL, YOU MIGHT END UP WITH GRAPHS THAT ARE…

- Wrong:

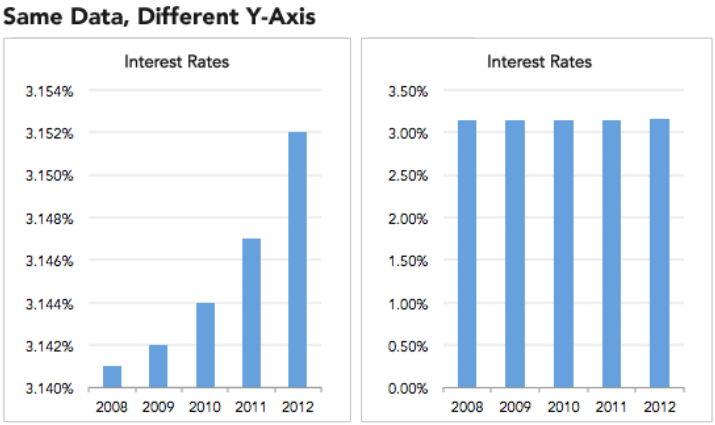




DATA VISUALISATION

IF YOU ARE NOT CAREFUL, YOU MIGHT END UP WITH GRAPHS THAT ARE…

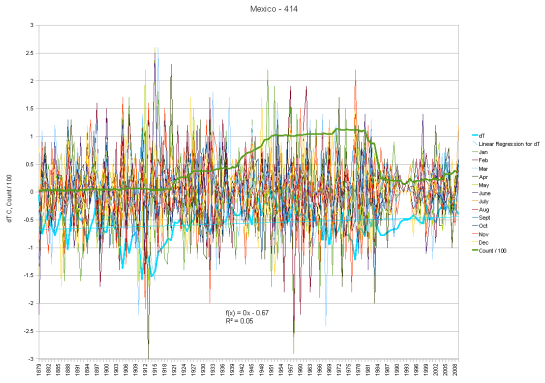
- Misleading:

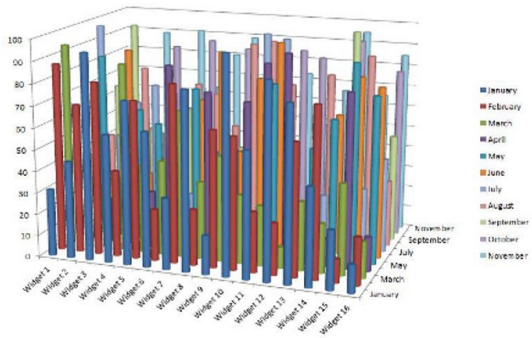


DATA VISUALISATION

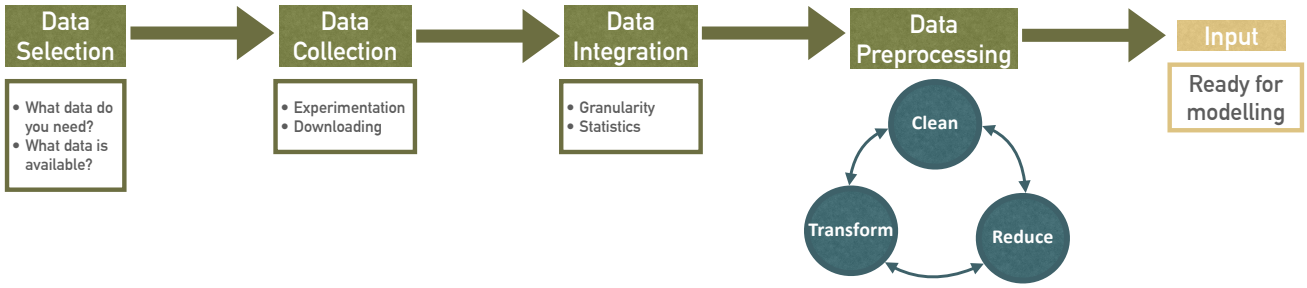
IF YOU ARE NOT CAREFUL, YOU MIGHT END UP WITH GRAPHS THAT ARE…

- Confusing:





DATA PRE-PROCESSING: FROM COLLECTION TO TRANSFORMATION



PRE-PROCESSING STEPS: DATA SELECTION

- Selection can be seen as precursor to data collection
- What data do you need?
- How are you going to obtain it?
  - It includes determination of:
    - data types
    - sources
    - instruments for collecting the data
  - otherwise it involves determining:
    - Relevant data-sets
    - corresponding subsets

PRE-PROCESSING STEPS: DATA SELECTION

- For a model on flu spread in a region, we might select data on:
  - community areas
  - demographics
  - historical flu trends
  - movement patterns (work, home, etc.)
- For a model of 2 species competing for survival in an area:
  - species demographics in that area
  - growth rates of both species in the area
  - death rates of both species in the area

PRE-PROCESSING STEPS: DATA COLLECTION

- Identify source(s) of data:
  - Acquisition: Analyse the area/field. What data is available?
    - Download from the Internet (public benchmarks, specialised datasets)
    - Obtain from experts
  - Experimentation: If the data that you need is not available, can you collect it?
    - Data collection protocol
    - Ethical approval

PRE-PROCESSING STEPS: DATA INTEGRATION

- Once source(s) of data have been selected, you might need to combine them together.
- Heterogeneous data:
  - Data from different sources
    - Creating a dataset of purchases from data from both Amazon and Ebay purchases.
  - Data in different formats
    - Combining image datasets when images from dataset A are PNG and images from dataset B are JPG
- **Objective:** to create a single coherent database

PRE-PROCESSING STEPS: DATA INTEGRATION

- Issues include
  - How to determine if attributes are the same?
    - different attributes may have same names
    - same attributes may have different names
  - Data conflicts
    - different units of measurement
    - granularity issues
    - different contexts (e.g. 'high earner' in USA or UK?)

PRE-PROCESSING STEPS: DATA INTEGRATION

- Before combining any data from different sources:
  - It is vital to ensure that the essential statistical measures relating to any attributes to be combined are the same.
  - Judging 'similarity' requires intelligence and deep knowledge of the problem/data.
  - In general: it is extremely unlikely that they will be identical.

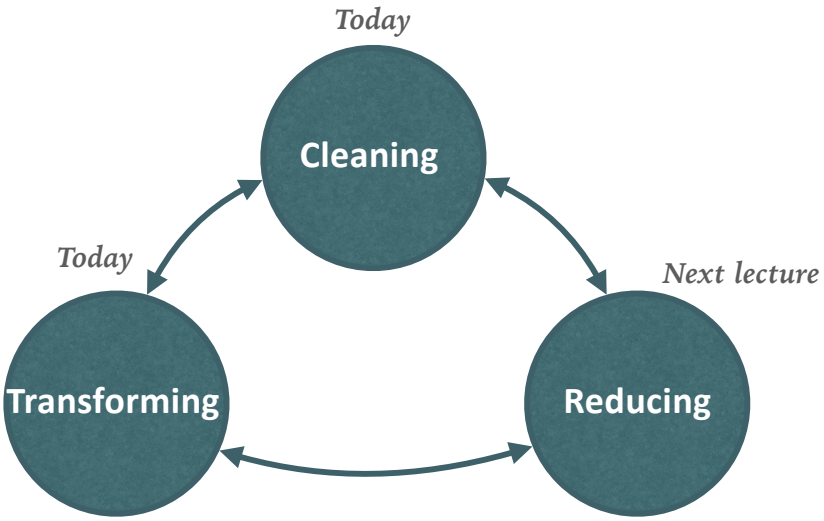
PRE-PROCESSING STEPS: DATA INTEGRATION

- As an absolute minimum, make sure to:
  1. Check similar location:
     - mean,
     - median
     - mode
  2. check similar dispersions:
     - min,
     - max
     - variance / standard deviation
  3. check similar distributions (advanced)

PRE-PROCESSING STEPS: PRACTICAL STEPS

1. Parse the data into columns
   - ➤ delete blank rows and columns
2. Name the attributes
3. Assess presence of target class(es)
   - ➤ assess which attributes are inputs, and which output(s)
4. Perform elementary statistical analysis
   - ➤ determine data types (N, O, I, R)
5. In practice, describing and visualising the data may occur in parallel to 'pre-processing'

PRE-PROCESSING STEPS: THE PRE-PROCESSING CYCLE



CLEANING DATA: MISSING VALUES

- Perhaps, if there are few, we can 'find' the data
  - go back to the original source?
  - ask experts?
- Easiest option is to delete attributes / instances
  - almost all of one attribute missing
    - delete attribute(s) (columns)
  - rare / sporadic missing values, or plenty of data
    - delete instance(s) (rows)

CLEANING DATA: MISSING VALUES

- But be careful!
  - 'missingness' may be informative
    - No driving license number?
    - A missing symptom may be the key to differential diagnosis

CLEANING DATA: INPUTTING MISSING VALUES

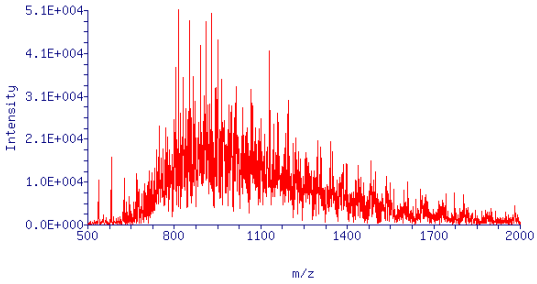
- Inputting: filling-in missing data with values
  - numerical: blank → zero (or e.g. -1)
    - number children: blank may imply zero?
  - categorical: blank → 'unknown'
    - marital status: blank implies 'unknown' or 'don't care'?

CLEANING DATA: INPUTTING MISSING VALUES

- Inputting with specific values
  - mean of attribute
  - mean of subsets of attribute
    - number of children belonging to men and women:
      - for men, replace missing with mean of men
      - for women, replace with mean of women
  - Perhaps use mode rather than mean
- More sophisticated methods:
  - Clustering
  - Regression

CLEANING DATA: NOISY DATA

- Noise: random error in a measured variable
- Some causes:
  - sensor failure
  - data transmission
  - improper data entry
- Smoothing techniques:
  - binning
  - regression
  - clustering
- Detection of anomalies



CLEANING DATA: BINNING

- A form of local smoothing
- Values are corrected by looking at neighbours
- General process:
  - sort values
  - construct a set of bins (or buckets)
  - replace values by other values that are computed relative to each bin/bucket
- Variants:
  - smoothing by bin means
  - smoothing by bin medians
  - smoothing by bin boundaries

CLEANING DATA

BINNING

***Sort data**: 4,8, 15, 21, 21, 24, 25, 28, 34*

Partition into equal-frequency bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smooth by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

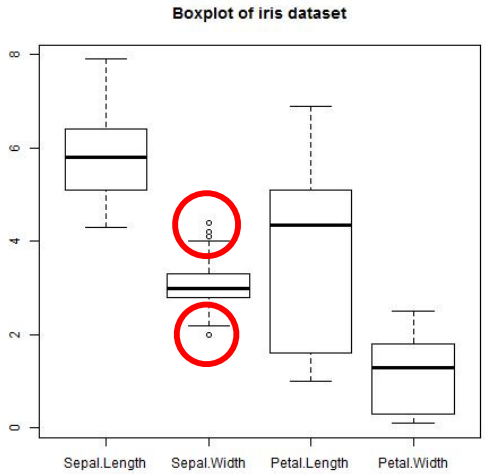
Smooth by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Smoothing by bin medians is the same except that the median of a bin is used instead of the mean

CLEANING DATA

REMOVING OUTLIERS

1. Statistical approach
   - Use visualisation to identify potential outliers
   - Calculate means & standard deviations
   - Define an outlier if value is more than *3* st. devs from the mean?



CLEANING DATA: REMOVING OUTLIERS – EXAMPLE

- Example: Books read by teenagers in one month.
- 13-year olds
- 15-year olds
- 24 instances of samples in each case

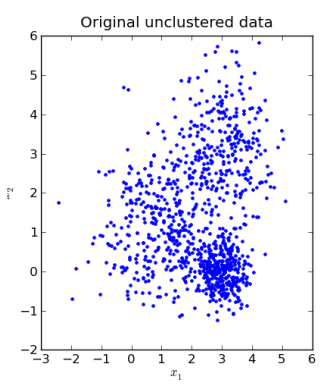
	13 yo	15 yo
1	1	2
2	3	3
3	4	5
4	5	5
5	2	3
6	6	6
7	9	9
8	7	8
9	8	8
10	8	8
11	8	9
12	6	6
13	5	7
14	5	7
15	5	7
16	17	18
17	16	4
18	1	3
19	1	7
20	2	5
21	2	4
22	3	4
23	3	4
24	3	16

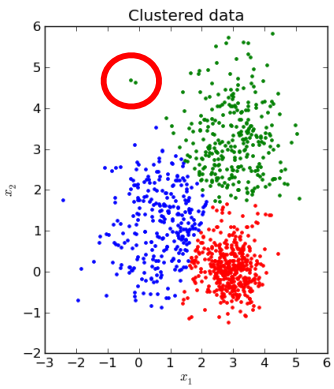
CLEANING DATA: REMOVING OUTLIERS – EXAMPLE

- Visualise data
  - Identify possible outliers
- 13 year olds
  - mean= 4.4
  - std=2.5
  - mean+3*std=11.9
  - mean-3*std= -3.1
  - Samples over 11.9 or under -3.1, can be classified as outliers

CLEANING DATA: REMOVING OUTLIERS

2. Clustering approach
   - Apply clustering algorithm(s)
   - Instance is outlier if 'far' from any / all clusters?





CLEANING DATA: REMOVING OUTLIERS

3. Model building approach
   - Construct multiple alternative predictive models
   - Use models to predict 'training' data
   - Instance is outlier if all models 'fail' to predict

CLEANING DATA: REMOVING OUTLIERS

- Note that all of the methods to detect outliers come *after* initial data exploration
  - Do **not** try to remove outliers before initial data exploration
- There is no guaranteed satisfactory approach
  - Outliers may be incorrect data (e.g. transcription error)
  - Outliers may be critical instances that lead to new knowledge and/or novel insights!

TRANSFORMING DATA: CHANGING DATA TYPES

- There may be a need to transform all data to numeric
  - Easier to deal with (e.g. plotting data)
  - Some machine learning algorithms require numbers
- Booleans
  - False/True → 0 / 1
- Ordinal
  - *N* different ordinal values → integers 1 … *N*
  - Do not mistake them for interval data
- Nominal
  - *N* different nominal values → *N* indicator variables
  - One-hot encoding: Each indicator is 0 (attr is not *X*) or 1 (attr is *X*)
  - Embeddings

TRANSFORMING DATA: CHANGING DATA TYPES

- Example: Digitise data from:

	Original data	Transformation	Transformation II
Number of extremities	4	4	4
Colour (Black, Blue, Brown, ...)	Brown	3	[0,0,1]
Age (Young, Adult, Old)	Young	1	1
Horns	True	1	1

TRANSFORMING DATA: SCALING

- Normalise the range of values
  - many machine learning algorithms require inputs in the ranges of [0,1] or [-1,1]
    - e.g. neural networks
  - give 'equal weighting' to the various attributes
- Normalise the distribution
  - ensure shape of distributions are (roughly) the same
  - in general, requires advanced statistics
    - the validity is questionable without external knowledge
- Normalisation requires great care, since distortions and biases can (will) be inadvertently added

TRANSFORMING DATA: LINEAR TRANSFORMATIONS

- E.g.

$$y_i = \frac{x_i - \min(x_1 \ldots x_n)}{\max(x_1 \ldots x_n) - \min(x_1 \ldots x_n)}$$

y_i *normalized value*
x_i *instance value*

- Or

$$y_i = \frac{2x_i - \min(x_1 \ldots x_n) - \max(x_1 \ldots x_n)}{\max(x_1 \ldots x_n) - \min(x_1 \ldots x_n)}$$

- Obviously, any range is similarly possible

TRANSFORMING DATA: SIGMOID/HYPERBOLIC TRANSFORMS

- E.g.

$$y_i = \frac{1}{1 + e^{-x_i}}$$

- Or

$$y_i = \frac{e^{x_i} - 1}{e^{x_i} + 1}$$

- Use $m(x_i - c)$ rather than x_i to alter slope and intercept

TRANSFORMING DATA: MEAN CENTRING AND STANDARDISATION

- Mean centring
  - subtract mean to move the mean value to zero

$$y_i = x_i - \bar{x}$$

- Standardisation
  - divide by standard deviation to achieve mean of zero and standard deviation of one

$$y_i = \frac{x_i - \bar{x}}{\sigma}$$

- This is a form of distribution normalisation

TRANSFORMING DATA: ATTRIBUTE CONSTRUCTION

- Use combinations of sets of attributes and combining / splitting functions to
  - generate new attributes
  - make data mining easier
- Old attributes
  - awkward to use in standard algorithms
  - limited predictive power
- New attributes
  - easier to use in standard algorithms
  - greater predictive power

TRANSFORMING DATA: ATTRIBUTE CONSTRUCTION – EXAMPLE

- Data mining on credit risk
- Two attributes
  - income
  - expenditure
- Create one new boolean
  - income > expenditure
- Advantages
  - easier to deal with
  - captures important key information

Income	Expenditure	Quality
120	100	Good
50	30	Good
50	70	Bad
200	40	Good
200	210	Bad
...	...	...
160	150	Good

Income	Expenditure	i>e	Quality
120	100	1	Good
50	30	1	Good
50	70	0	Bad
200	40	1	Good
200	210	0	Bad
...	...	...	...
160	150	1	Good

LECTURE OUTCOMES

- At the end of this lecture, you should be able to answer these questions:
  - What is data visualisation?
  - Why is it useful? When is it useful?
  - What are the main graphs we can use to describe data?
  - What pre-processing steps should you take?
  - How and with what purpose do clean data?
  - How and with what purpose do you transform data?

THE END

Questions