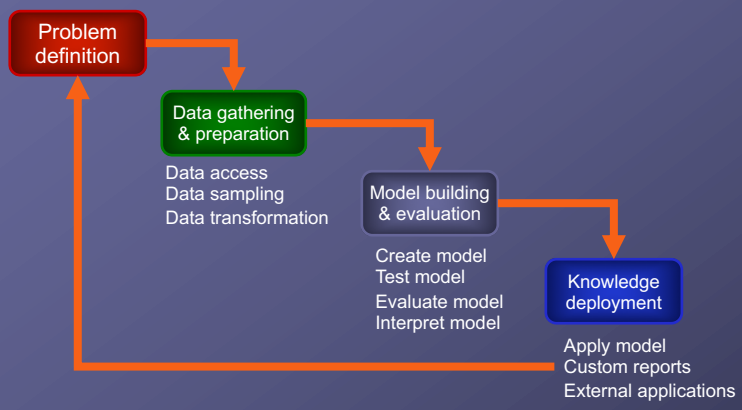
Data Pre-processing

Descriptive Data Summarisation

Learning Outcomes

- Why pre-processing the data?
- Different pre-processing tasks
- Descriptive data summarisation
  - Central Tendency
  - Data Dispersion

Data Mining Process



Why Data Pre-processing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

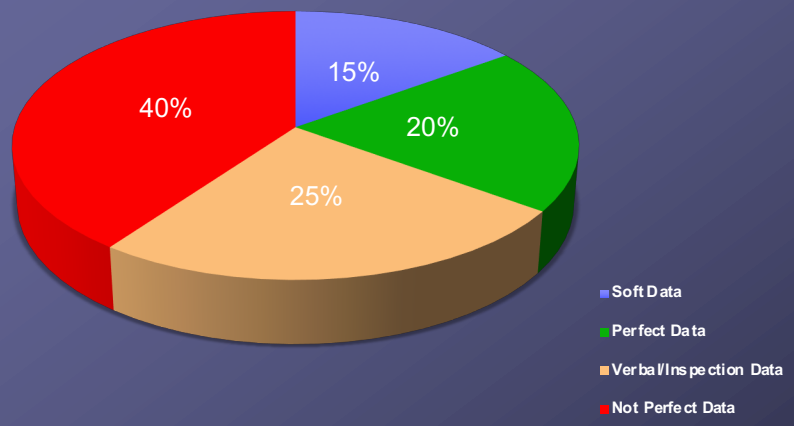
Data Quality

- Perfect data
  - Data is valid, complete, and reliable. No data extrapolation is needed
- Not Perfect data
  - Data with NO serious flaws, but needs some pre-processing
- Verbal/Inspection data
  - Data with serious gaps → requires additional documentation and verification prior to its inclusion in the DM process
- Soft data
  - Data relied on the memories of experienced personnel of the participating facility
  - The most difficult to summarise

Examples

- Not Perfect data
  - The data recorded in a dimension which is not important
- Verbal/Inspection data
  - Wrong or non-recorded values of an airplane flight parameters
- Soft data
  - Memories of an experienced analyst who dealt with the same problem before

Data Quality



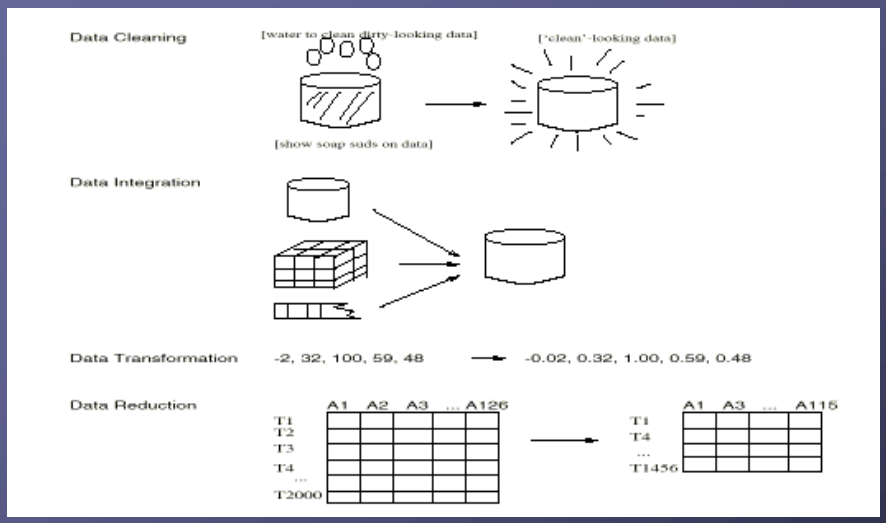
Analysis Plan

- Purpose of the analysis
  - Identify population groups and domains of interests
  - Example
    - Population groups: customers, personnel, etc.
    - Domain of interest: sales, profit, stock, products, etc.
- Audience for the analysis
  - Agencies, companies, directors, communities, etc.
- Data availability and data quality
  - Choices about which data to include
  - Etc.

Major Tasks in Data Pre-processing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalisation and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretisation
  - Part of data reduction but with particular importance, especially for numerical data

Forms of data pre-processing



Data Summarisation

- Descriptive Data Summarisation
  - Identify typical properties of the data
  - Highlight which data values should be treated as noise or outliers
- Descriptive Statistics
  - Understand the distribution of the data
  - Central Tendency: mean, median, midrange
  - Data Dispersion: quartiles, inter-quartile range (IQR), variance

Central Tendency

- Arithmetic Mean
  - Effective numerical measure of the centre
  - Let $x_1, x_2, \ldots, x_N$ a set of observations

$$\overline{X} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- Drawbacks
  - Sensitivity to extreme values (outliers)
  - Trimmed mean: obtained after removing the extremes

Central Tendency (2)

- Median
  - Used for skewed (asymmetric data)
  - Let $\{x_1, x_2, \ldots, x_N\}$ a set of ordered observations
  - The median is the middle value if N is odd and is the average of the two middle values if N is even
  - Example: consider the set of values (1, 2, 3, 4, 5, 90}. Calculate the mean and the median
    - The mean: 17.5
    - The trimmed mean with p=40% is: 3.5
    - The median: 3.5

$$median(X) = \begin{cases} x_{r+1} & \text{if } N = 2r+1 \\ \frac{1}{2}(x_r + x_{r+1}) & \text{if } N = 2r \end{cases}$$

Central Tendency (3)

- Mode
  - Indicates the value that occurs most frequently in the set
  - Example: {140 (0.33), 160 (0.27), 130 (0.22), 170 (0.18)
    - The mode is: 140
- Midrange
  - Range(X) = max(X) - min(X)
  - Midrange(X) = [max(X) + min(X)]/2

Data Dispersion

- Dispersion of the Data
  - The degree to which numerical data tend to spread
  - The most common measures are
    - Range
    - Five-number summary
    - Inter-quartile range
    - Standard deviation
- Standard Deviation
  - Measures spread about the mean
  - Can only be used when the mean is chosen as the measure of the centre
  - Let $X=\{x_1, x_2, \ldots, x_N\}$

$$std(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Data Dispersion (2)

- Percentiles
  - Let X be a set of ordered observations
  - k^{th} percentile of X is x_i such that k% of X is below x_i
- Quartile = 25th percentile
  - 1st quartile (Q_1), 2nd quartile (Q_2), 3rd quartile (Q_3)
  - Inter-quartile range (IQR): $IQR = Q_3 - Q_1$
- Five-number summary
  - Includes information about end-points
  - = {min(X), Q_1, median, Q_3, max(X)}

Examples: Q1 & Q3

Let X = {70, 65, 54, 56, 57, 80, 71, 46, 55, 63, 62, 53, 68, 76, 58, 54}.

Calculate Q1 and Q3

1) Sorting X.
{46, 53, 54, 54, 55, 56, 57, 58, 62, 63, 65, 68, 70, 71, 76, 80}

2) Calculate positions
Q1 position: 16 × 25% = 4,
Q3 position: 16 × (100 − 25)% = 12

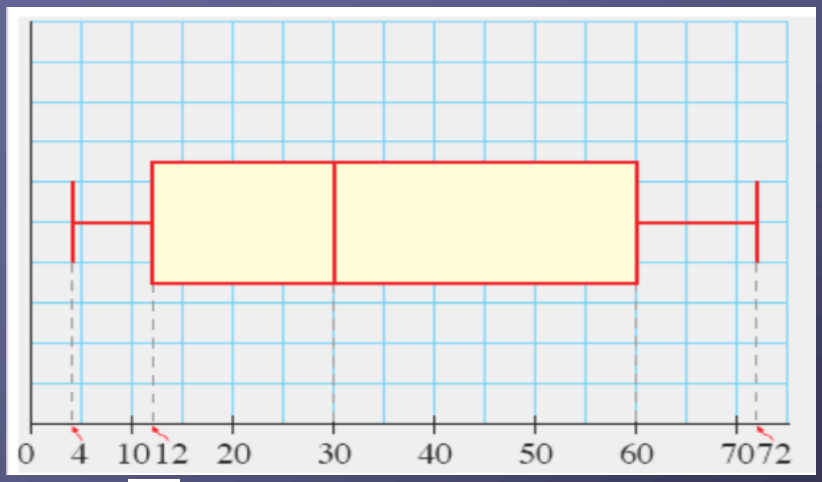
3) Find the values by the positions
(The average of the 4th and 5th values, n=16)
Q1 = (54 + 55) ÷ 2 = 54.5
Q3=?

Examples of Quartiles

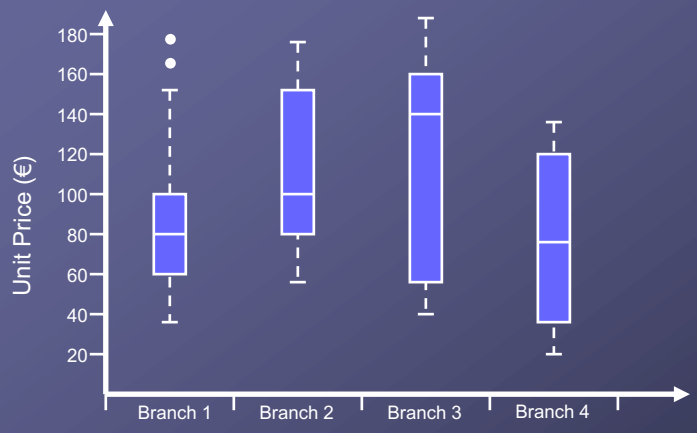
N	values	Median	Q1	Q3
13	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	7	4	10
14	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	7.5	4	11
15	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	8	4	12
16	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16	8.5	4.5	12.5

Examples: boxplot

X = {30, 30, 50, 52, 4, 6, 10, 12, 23, 25, 60, 67, 70, 72}



Visualising a Distribution



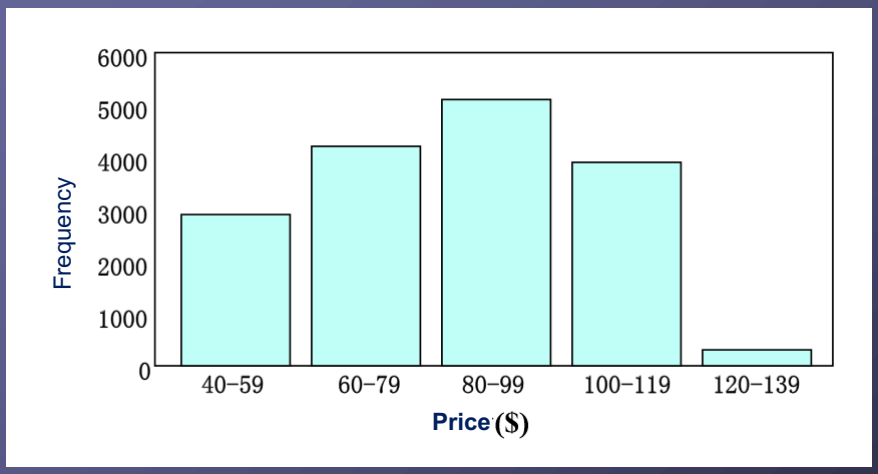
Boxplot

Visualising a Distribution

- Histograms
  - Frequency histograms
  - A graphical method for summarising the distribution of a given attribute
- Quantile plot
  - Simple way to have a 1st look at a univariate data distribution
  - Allows us to compare different distributions based on their quantiles
- Quantile-Quantile Plot (q-q plot)
  - Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
  - Is a powerful visualisation tool
- Scatter Plot
  - Each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane
  - Most effective graphical methods for determining if there appears to be relationship, patterns, or trends between two numerical attributes

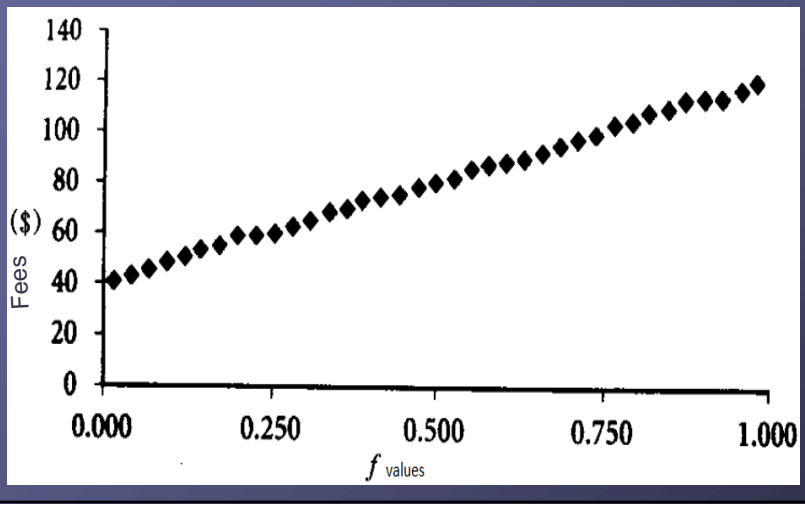
Visualising a Distribution (cont.)

- Histograms



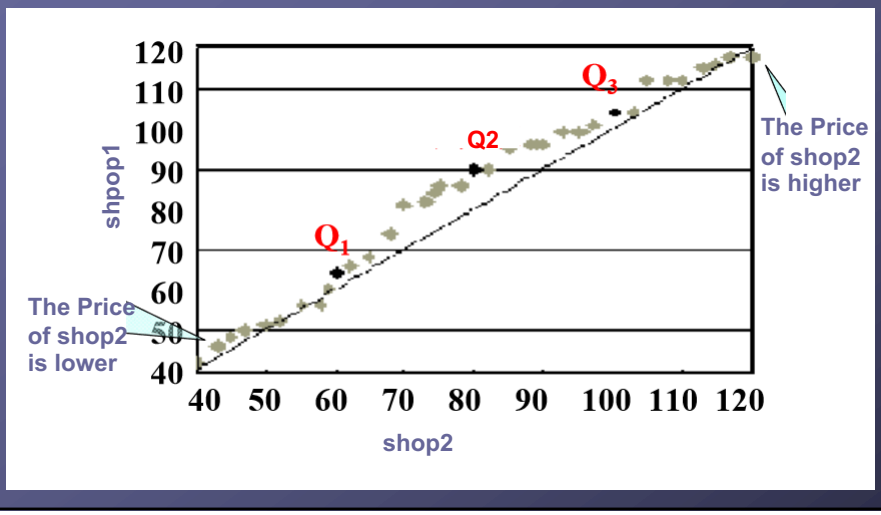
Visualising a Distribution (cont.)

- Quantile plot



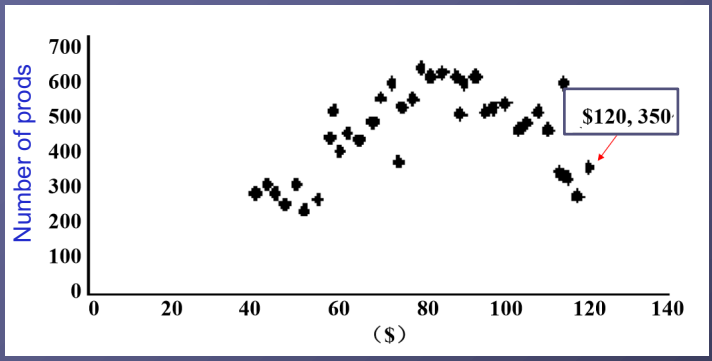
Visualising a Distribution (cont.)

- Quantile-Quantile Plot (q-q plot)



Visualising a Distribution (cont.)

- Scatter Plot



Home Work

Suppose a hospital tested the **age** and **body fat** data for 18 randomly selected adults with the following results.

1. Calculate the mean, median, and standard deviation of *age and %fat*.
2. Draw the boxplots for age and %fat.
3. Draw a scatter plot and q-q plot based on these two variables.
4. Normalise the two variables based on *z-score normalisation*.
5. Calculate the correlation coefficient (Pearson's product moment coefficient).
6. Are these two variables positively or negatively correlated?

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31	4	25.9	27.4	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7