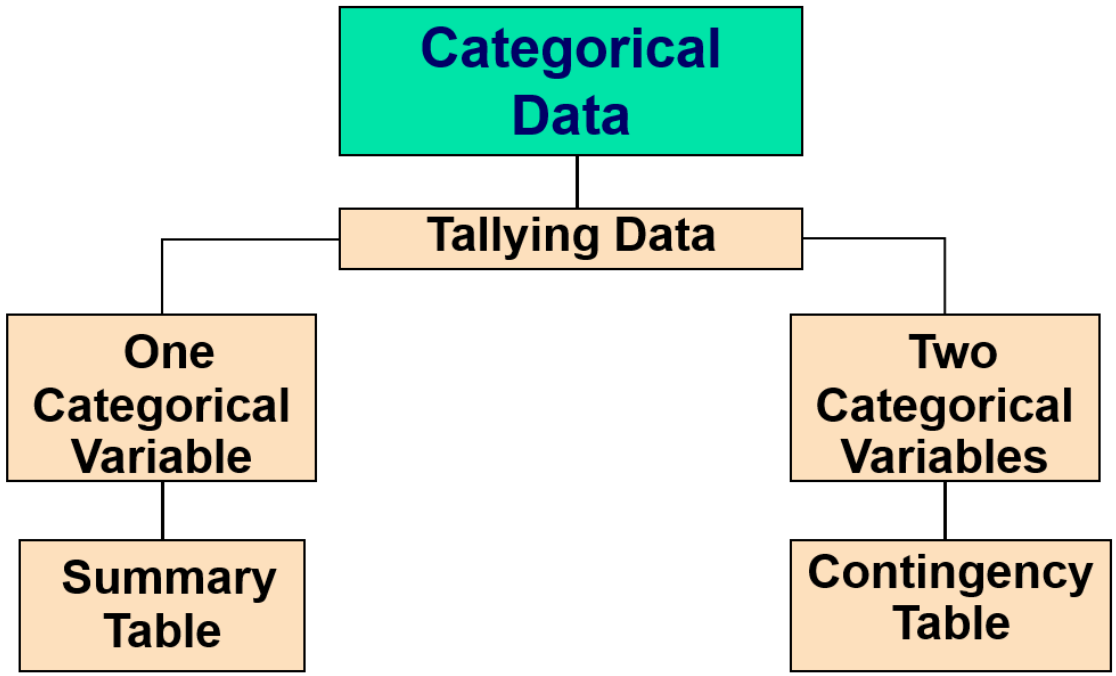
Topic 2: Organising and Presenting Data

Dr. Lei Pan

School of Accounting, Economics and Finance
Curtin University

Categorical data are organised by utilising tables



Organising and Visualising One Categorical Variable (1 of 2)

- Summary table
  - summarises categorical or numerical data
  - gives the frequency, proportion or percentage of data values in each category or class

Summary Table

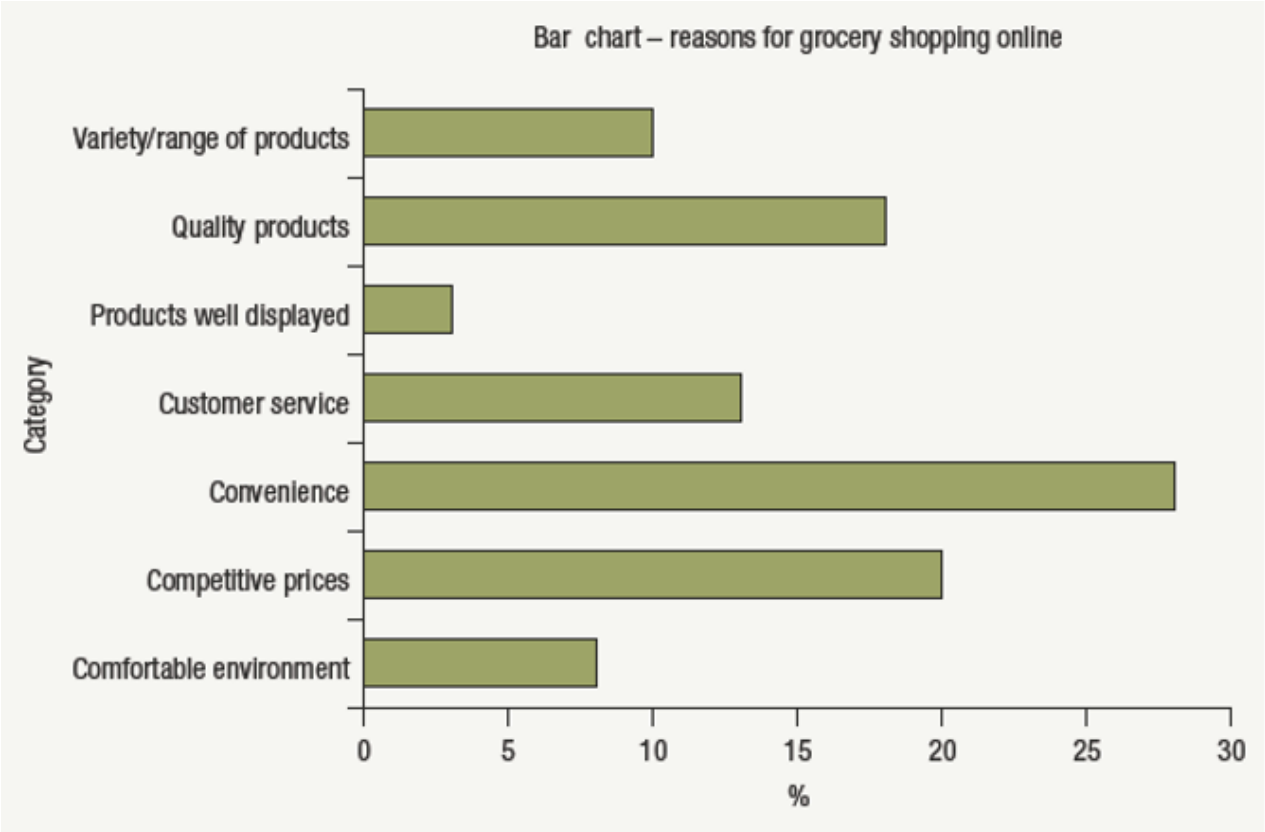
Reason	Percentage
Comfortable environment	8
Competitive prices	20
Convenience	28
Customer service	13
Products well displayed	3
Quality products	18
Variety/range of products	10
	100

Example of a summary table, which shows reasons for grocery shopping online

Organising and Visualising One Categorical Variable (2 of 2)

- Bar and pie charts
  - Bar charts and pie charts are often used for qualitative data (categories or nominal scale).
  - The length of bar, or size of pie slice, shows the frequency or percentage for each category.
  - Bar charts are preferred for comparing categories.
  - Pie charts are preferred for observing portion of the total which lies in a particular category (e.g., market share).

Bar Chart



Microsoft Excel bar chart of the reasons for grocery shopping online

Pie Chart

Microsoft Excel pie chart of the reasons for grocery shopping online

Organising Two or More Categorical Variables

- Contingency table
  - Used to study patterns that may exist between the responses of two or more categorical variables.
  - Cross tabulates or tallies jointly the responses of the categorical variables.
  - For two variables the tallies for one variable are located in the rows and the tallies for the second variable are located in the columns.

Contingency Table

- A random sample of 400 invoices is drawn.
- Each invoice is categorized as a small, medium, or large amount.
- Each invoice is also examined to identify if there are any errors.
- This data are then organised in the contingency table.

Example: Contingency table

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

Contingency table showing frequency of invoices categorised by size and the presence of errors

Contingency Table Based on Percentage of Overall Total

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

42.50% = 170 / 400
25.00% = 100 / 400
16.25% = 65 / 400

	No Errors	Errors	Total
Small Amount	42.50%	5.00%	47.50%
Medium Amount	25.00%	10.00%	35.00%
Large Amount	16.25%	1.25%	17.50%
Total	83.75%	16.25%	100.0%

83.75% of sampled invoices have no errors and 47.50% of sampled invoices are for small amounts.

Contingency Table Based on Percentage of Row Totals

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

89.47% = 170 / 190
71.43% = 100 / 140
92.86% = 65 / 70

	No Errors	Errors	Total
Small Amount	89.47%	10.53%	100.0%
Medium Amount	71.43%	28.57%	100.0%
Large Amount	92.86%	7.14%	100.0%
Total	83.75%	16.25%	100.0%

Medium invoices have a larger chance (28.57%) of having errors than small (10.53%) or large (7.14%) invoices.

Curtin University

Contingency Table Based on Percentage of Column Totals

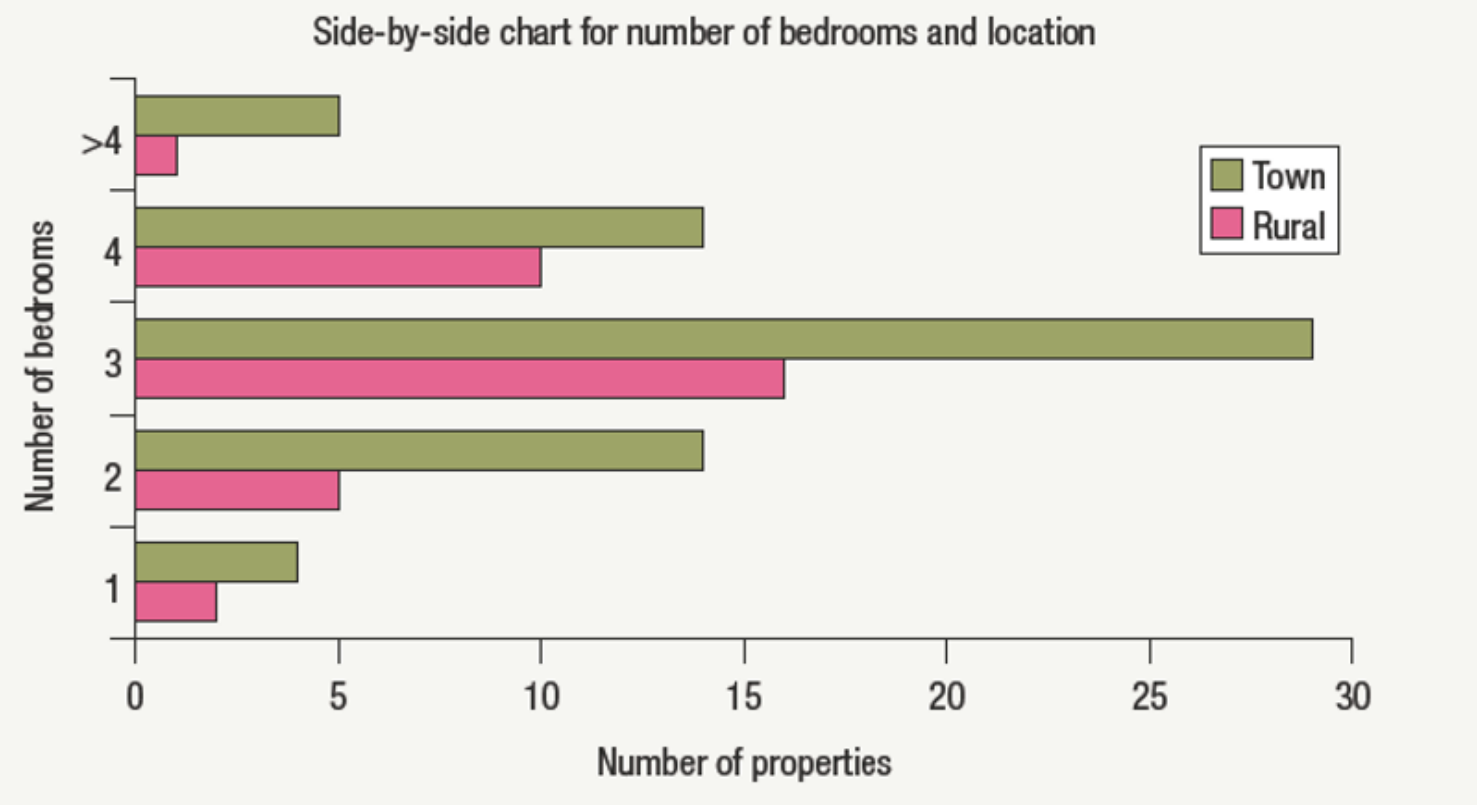
	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

50.75% = 170 / 335
30.77% = 20 / 65

	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%

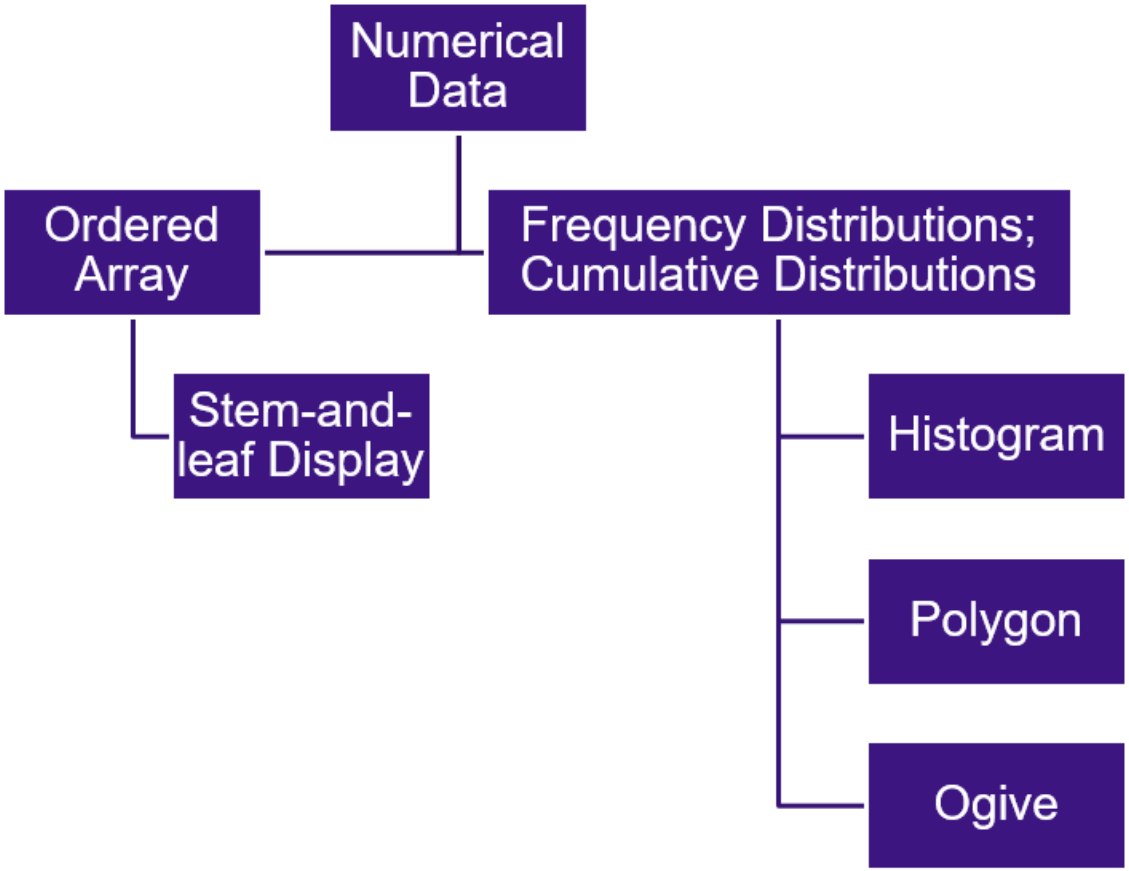
There is a 61.54% chance that invoices with errors are of medium size.

Side-by-Side Charts



A side-by-side chart is a graphical representation of a cross-classification table

Organising Numerical Data



Ordered Array

- A **sequence of data** in **rank** order.
- Shows range (minimum to maximum).
  - e.g. 24, 26, 24, 21, 27, 27, 30, 41, 32, 38
    becomes 21, 24, 24, 26, 27, 27, 30, 32, 38, 41
- Provides some signals about variability within the range and may help identify outliers.
- If the data set is large or if the data is highly variable, the ordered array is less useful.

Stem-and-Leaf Display

- A quick and simple way to see distribution details in a data set.
- **Method**: Separate the sorted data series into groups (the **stem**) and the values within each group (the **leaves**).

Example

Data in an ordered array: 21, 24, 26, 27, 27, 30, 32, 38, 41

- 21 is shown as →
- 38 is shown as →
- 41 is shown as →

Stem	Leaf
2	1
3	8
4	1

Example (Cont'd)

Completed stem-and-leaf display

Stem	Leaves
2	1 4 6 7 7
3	0 2 8
4	1

You Do It

Sam got his friends to do a long jump and got these results:

2.3, 2.5, 2.5, 2.7, 2.8 3.2, 3.6, 3.6, 4.5, 5.0

Construct a stem-and-leaf plot for the given data set.

Frequency Distribution (1 of 2)

- A summary table in which data are arranged into numerically ordered classes or intervals.
- The number of observations in each ordered class or interval becomes the corresponding frequency of that class or interval.
- Why use a frequency distribution?
  - Summarise numerical data
  - Condense raw data into a more useful form
  - Create a quick visual interpretation of the data and a first inspection of the shape of the data

Frequency Distribution (2 of 2)

- You must give attention to
  - selecting the appropriate **number** of class groupings for the table
  - determining a suitable **width** of a class grouping
  - establishing the **boundaries** of each class grouping to avoid overlapping.
- The number of classes depends on the number of values in the data.
  - In general, a frequency distribution should have **at least 5 but no more than 15 classes**.
- To determine the width of a class interval, you divide the range (Highest value - Lowest value) of the data by the number of class groupings desired.

Example

Price of main meal ($)	City frequency	Suburban frequency
$10 but less than $15	1	0
$15 but less than $12	0	0
$20 but less than $25	2	4
$25 but less than $30	3	13
$30 but less than $35	7	13
$35 but less than $40	14	12
$40 but less than $45	8	4
$45 but less than $50	5	1
$50 but less than $55	8	2
$55 but less than $60	1	1
$60 but less than $65	1	0
Total	**50**	**50**

A frequency distribution of the price per main meal for 50 city restaurants and 50 suburban restaurants

Frequency Distributions: Some Tips

- **Different class boundaries** may provide different pictures for the same data (especially for smaller data sets).
- Shifts in **data concentration** may show up when different class boundaries are chosen.
- As the size of the data set increase, the impact of alterations in the selection of class boundaries is greatly reduced.
- When comparing two or more groups with different sample sizes, you must use either a relative frequency or a percentage distribution.

Visualising Numerical Data: Histogram

- A graph of the data in a frequency distribution is called a **histogram**.
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
- The vertical axis is either **frequency**, **relative frequency**, or **percentage**.
- Bars of the appropriate heights are used to represent the frequencies (number of observations) within each class or the relative frequencies (percentage) of that class.

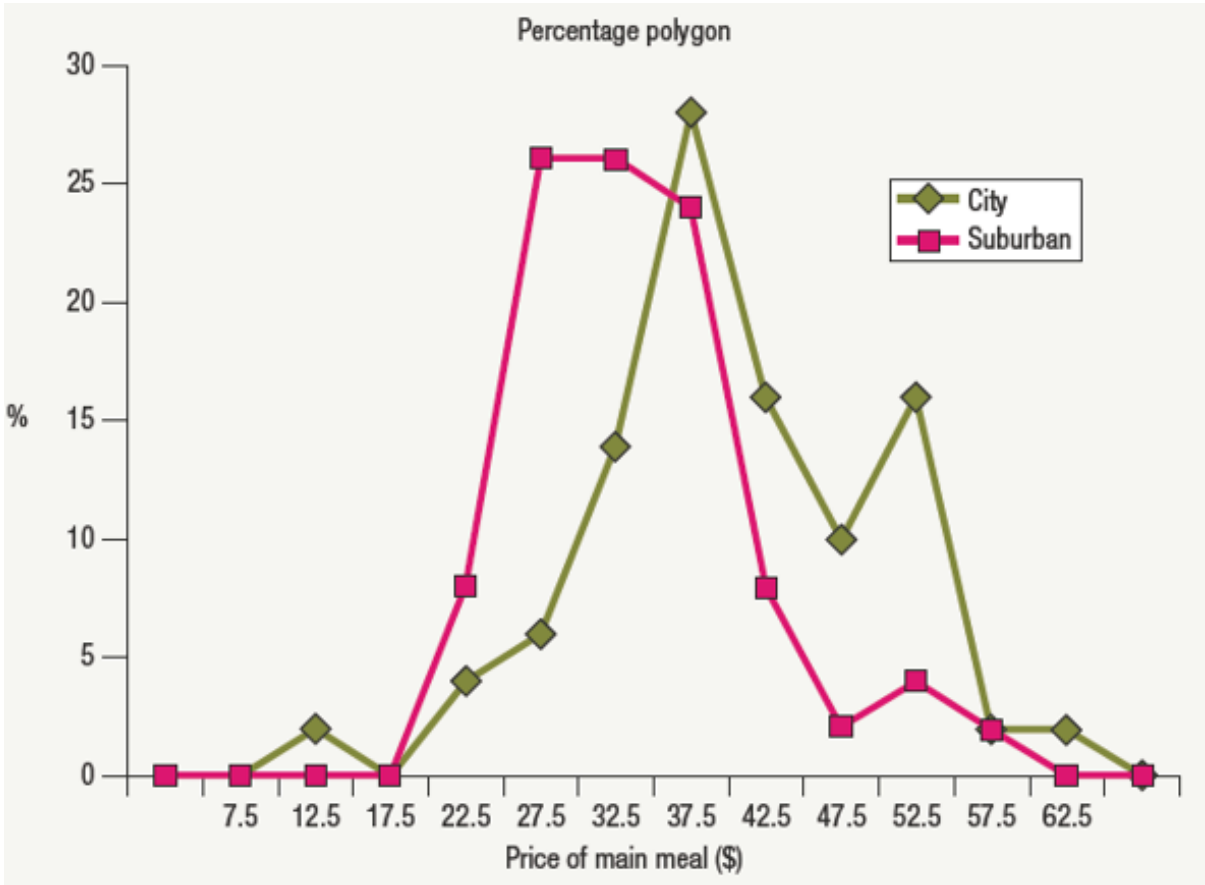
Example

An Excel histogram of the price of main meals at city restaurants

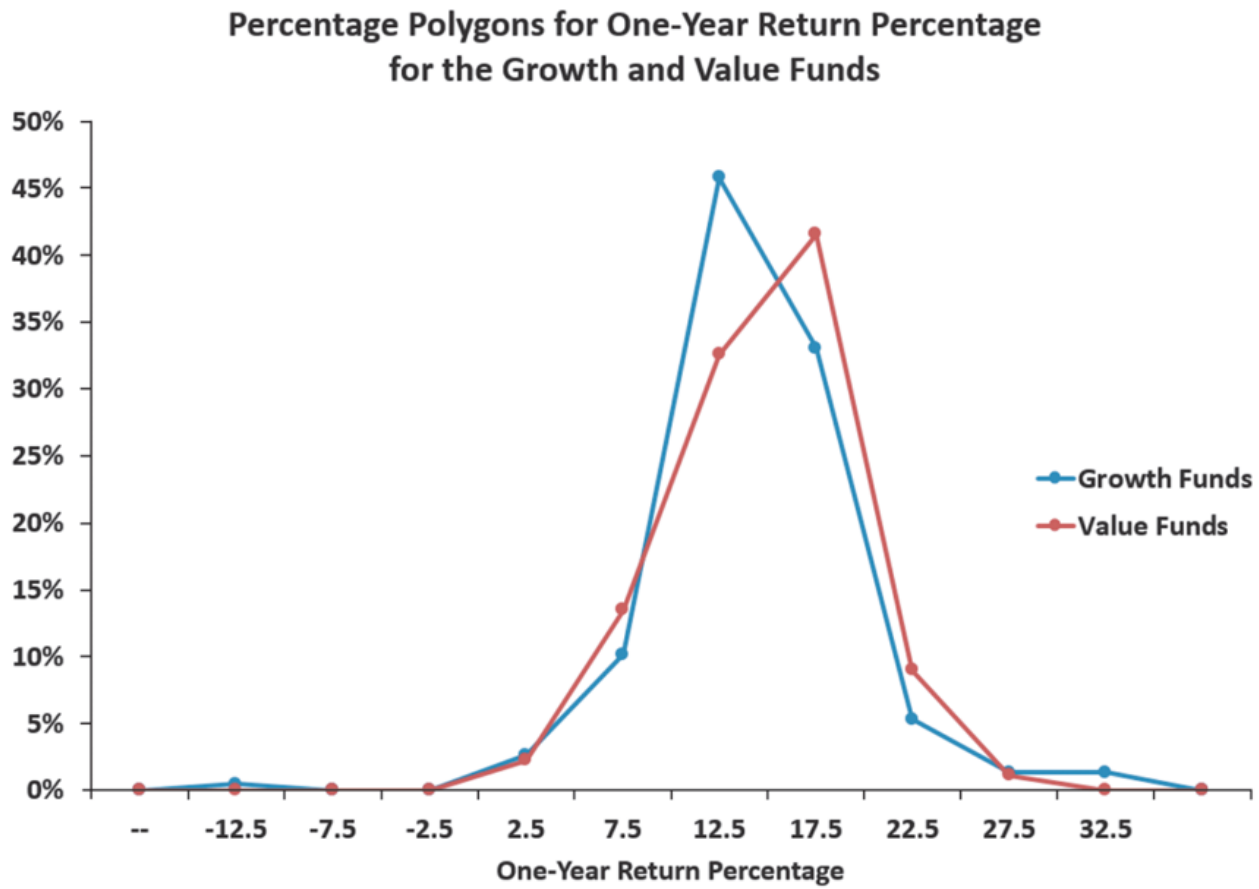
Visualising Numerical Data: Polygon

- **A percentage polygon** is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages.
- The **cumulative percentage polygon**, or **ogive**, displays the variable of interest along the X axis, and the cumulative percentages along the Y axis.
- Useful when there are two or more groups to compare.

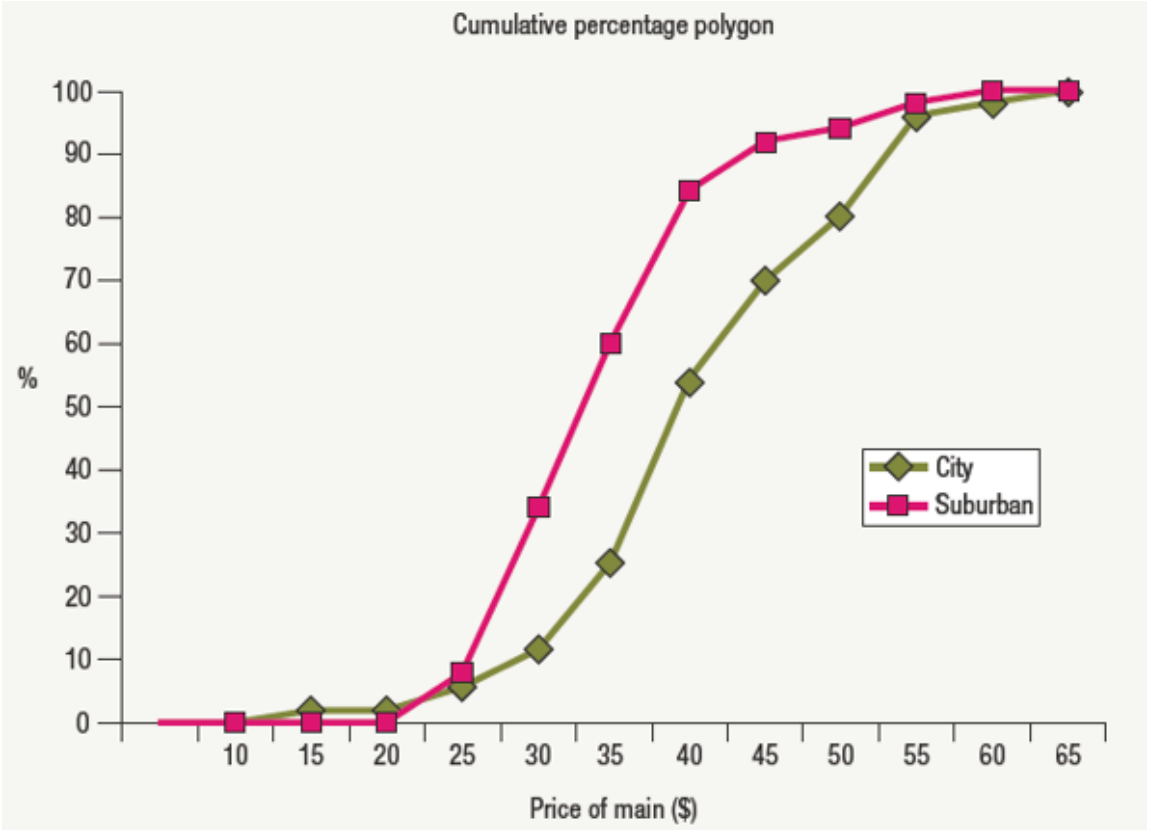
Example: Percentage Polygon (1 of 2)



Example: Percentage Polygon (2 of 2)



Example: Cumulative Percentage Polygon



Visualising Two Numerical Variables: Scatter Diagrams

- A **scatter diagram** is a graphical representation of the relationship between two numerical variable.
- Plotted points represent the given values of the independent variable and corresponding dependent variable.
- In a scatter diagram:
  - One variable is measured on the vertical axis (Y).
  - The other variable is measured on the horizontal axis (X).

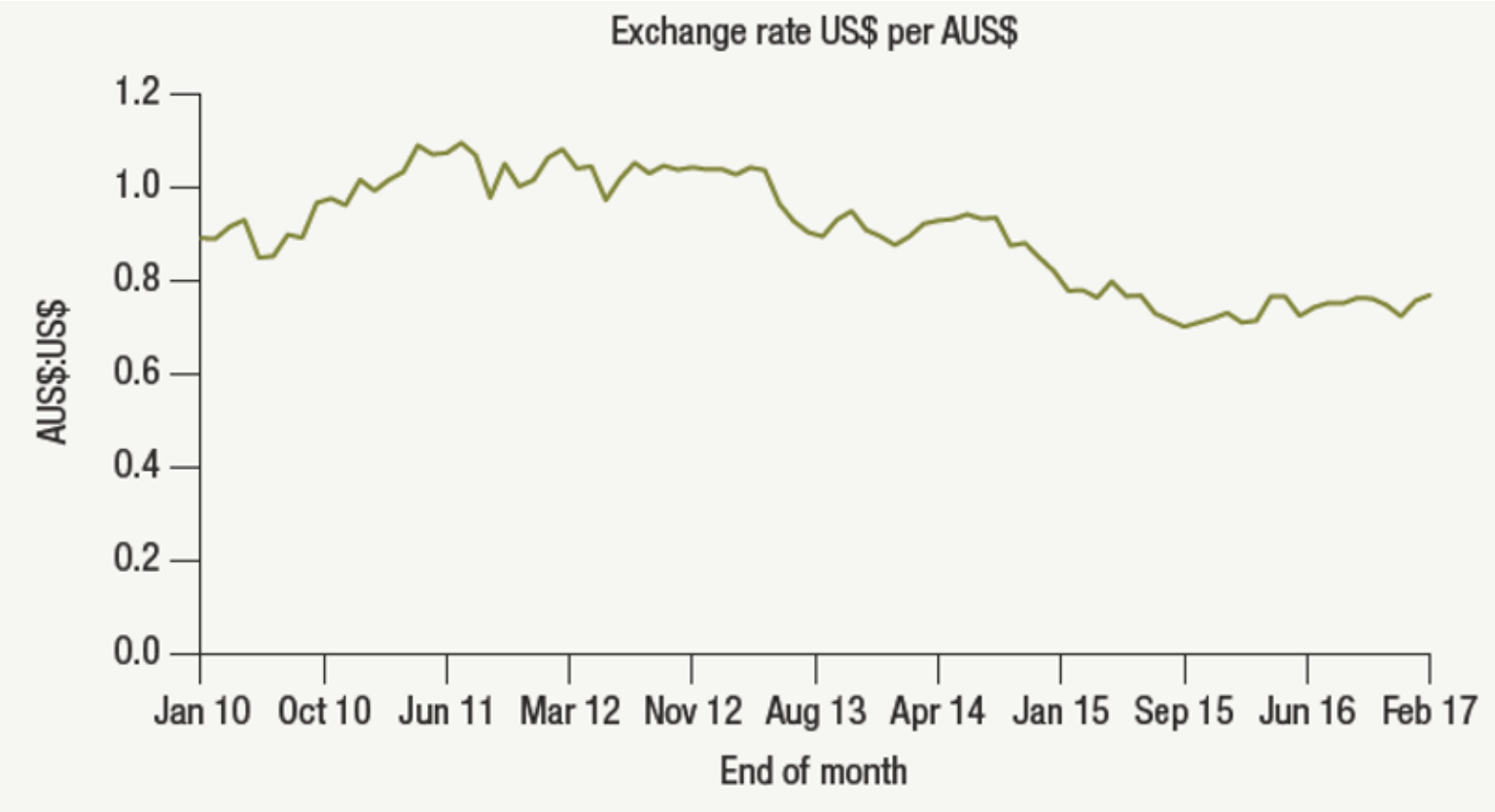
Example

Microsoft Excel scatter diagram for number of bedrooms and asking price

Visualising Two Numerical Variables: Time Series Plots

- A **time-series plot** is a graphical representation of the value of a numerical variable over time.
- In a time-series plot:
  - One variable is measured on the vertical axis (Y).
  - The time period is measured on the horizontal axis (X).

Example

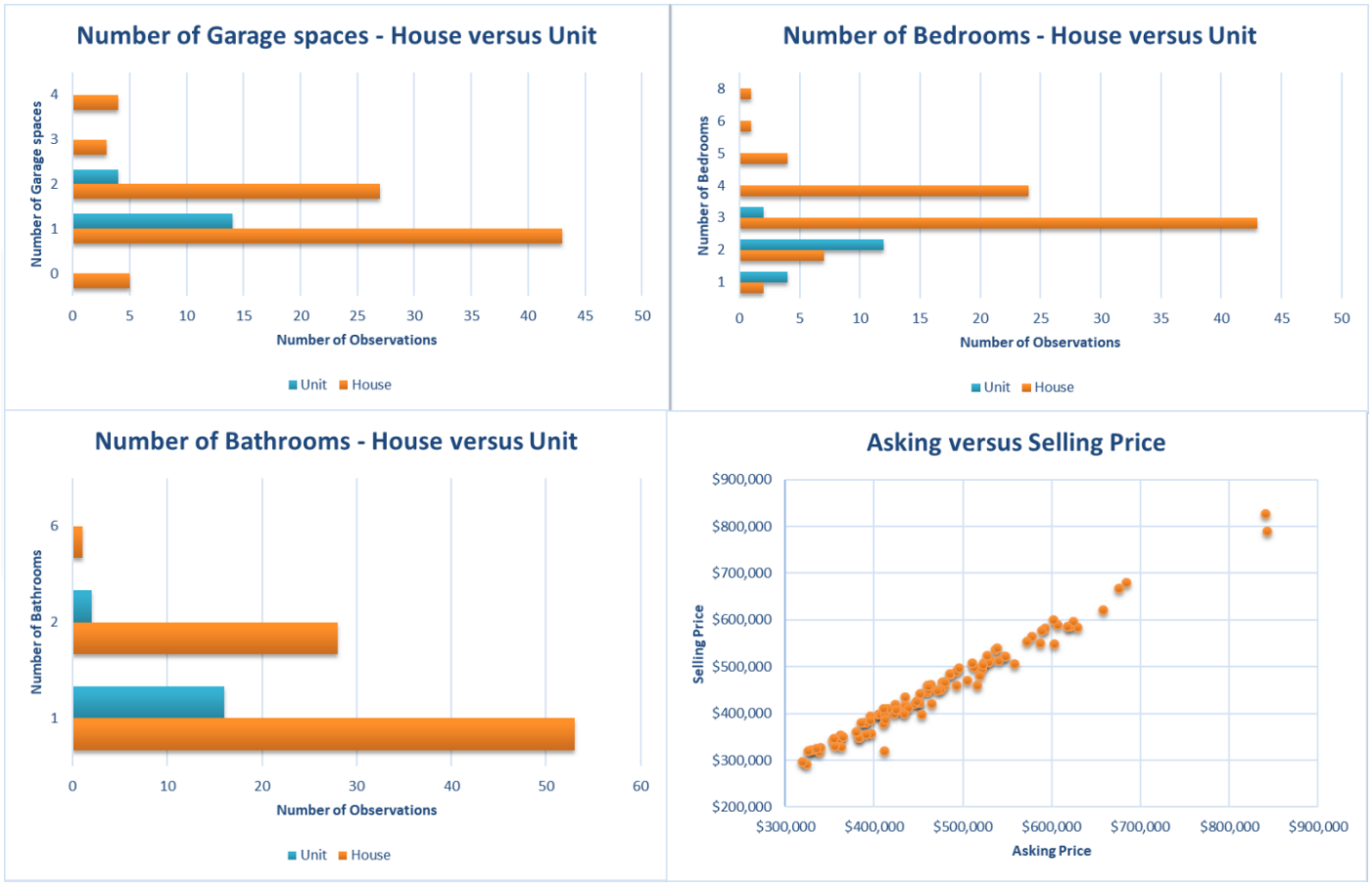


Microsoft Excel time-series plot of exchange rates: Australian dollar against US dollar 2010 to 2017

Business Analytics Applications - Descriptive Analytics

- **Business analytics** are skills, technologies and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.
- They include **descriptive** analytics, **predictive** analytics and **prescriptive** analytics.
- Descriptive analytics methods includes:
  - Dashboards
  - Sparklines (not covered in this unit)
  - Guages (not covered in this unit)
  - Bullet Graph (not covered in this unit)
  - Treemaps (not covered in this unit)

Example - Dashboard



Misusing Graphs and Ehtical Issues

- Do not distort the data
  - Frequency/quantity should be proportional to the area/volume.
- Avoid unnecessary adornments
  - No **chartjunk**
- Use a scale for each axis on a two-dimensional graph
  - Should be properly scaled along each axis.
  - All axes should be labelled.
  - Vertical axis should begin at zero unless there is justification for truncation.
  - Graph should contain a title.
  - Use the simplest graph for a given set of data.