

MODULE ONE: PRESENTING AND DESCRIBING INFORMATION

TOPIC 2: VISUALISING DATA

"A PICTURE IS WORTH A THOUSAND WORDS"



+ Learning Objectives

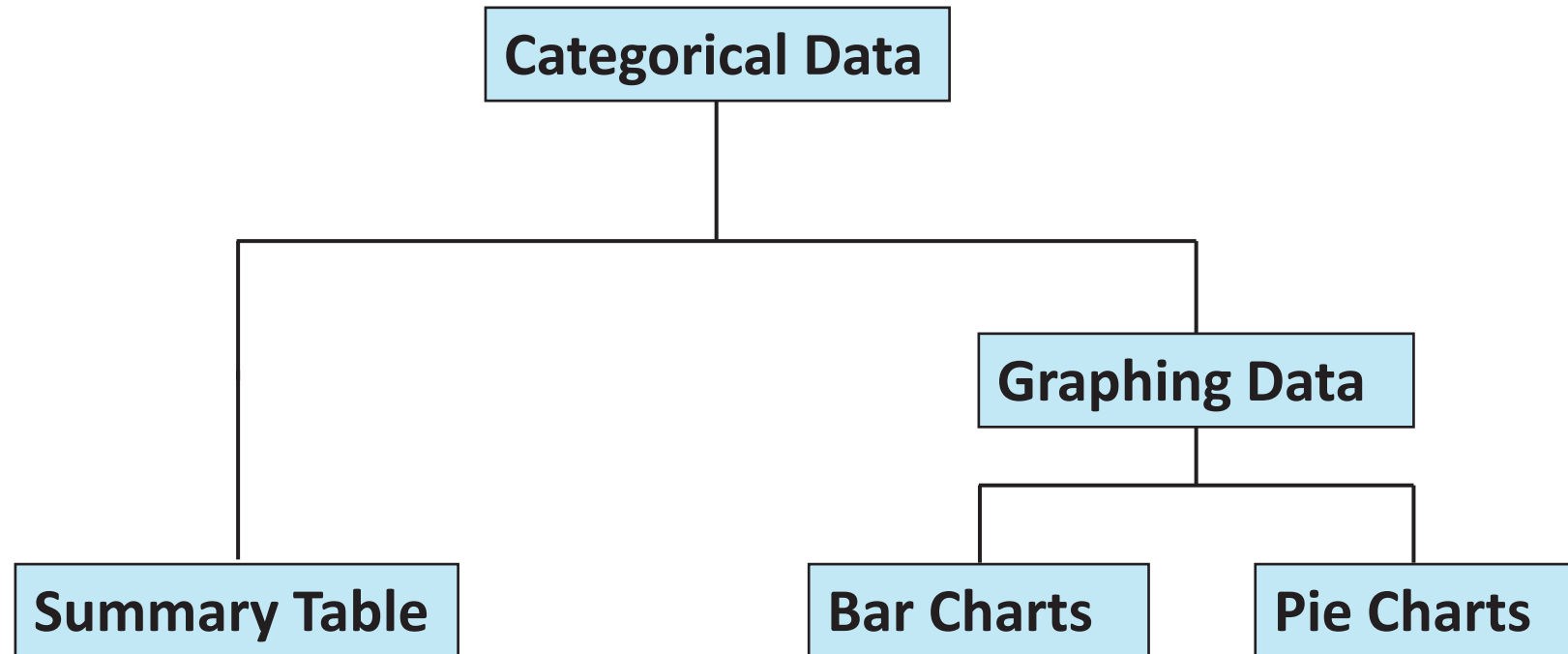
2

At the completion of this topic, you should be able to:

- describe the distribution of a single categorical variable using tables and charts
- describe the distribution of a single numerical variable using tables and graphs
- correctly present data in graphs

+2.1 Tables and Charts for Categorical Data

3



+Summary Tables

4

Table 2.2A

A frequency and percentage summary table for the location of 100 recent property sales

Location	Number (frequency) of properties	Percentage of properties
Rural	34	34.0
Town	66	66.0
Total	100	100.0

Table 2.2B

A frequency and percentage summary table for type of 100 recent property sales

Type	Number of properties	Percentage of properties
House	82	82.0
Unit	18	18.0
Total	100	100.0

+ Bar (Or Column) and Pie Charts

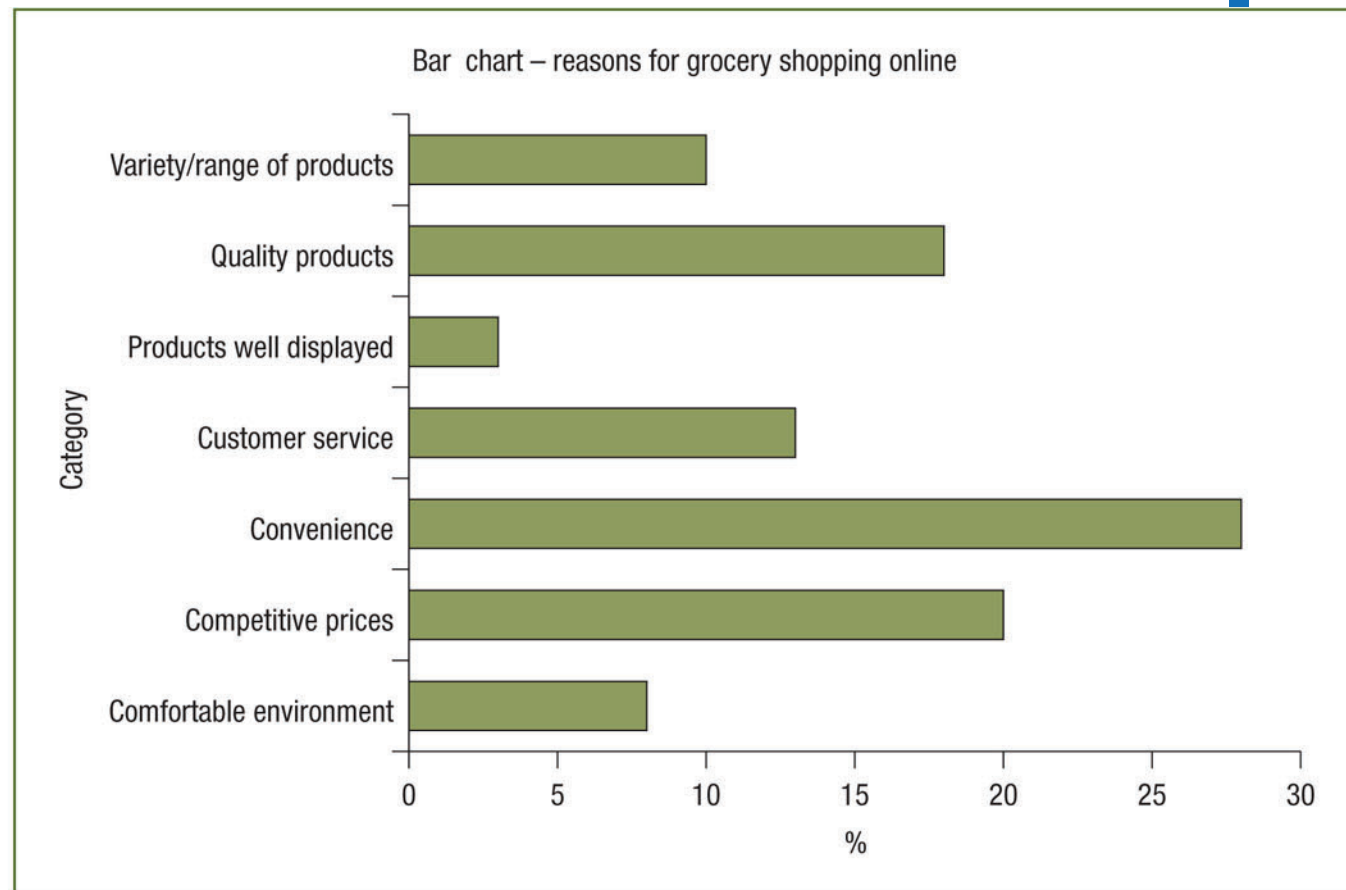
- Bar charts and pie charts are often used for qualitative data (categories or nominal scale)
- The length of bar, or size of pie slice, shows the frequency or percentage for each category
- Bar charts are preferred for comparing categories
- Pie charts are preferred for observing portion of the total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

+Bar Charts

6

Figure 2.1

Microsoft Excel bar chart
of the reasons for grocery
shopping online



Microsoft® product screen shots are reprinted with permission from Microsoft Corporation.

+Bar Charts

7

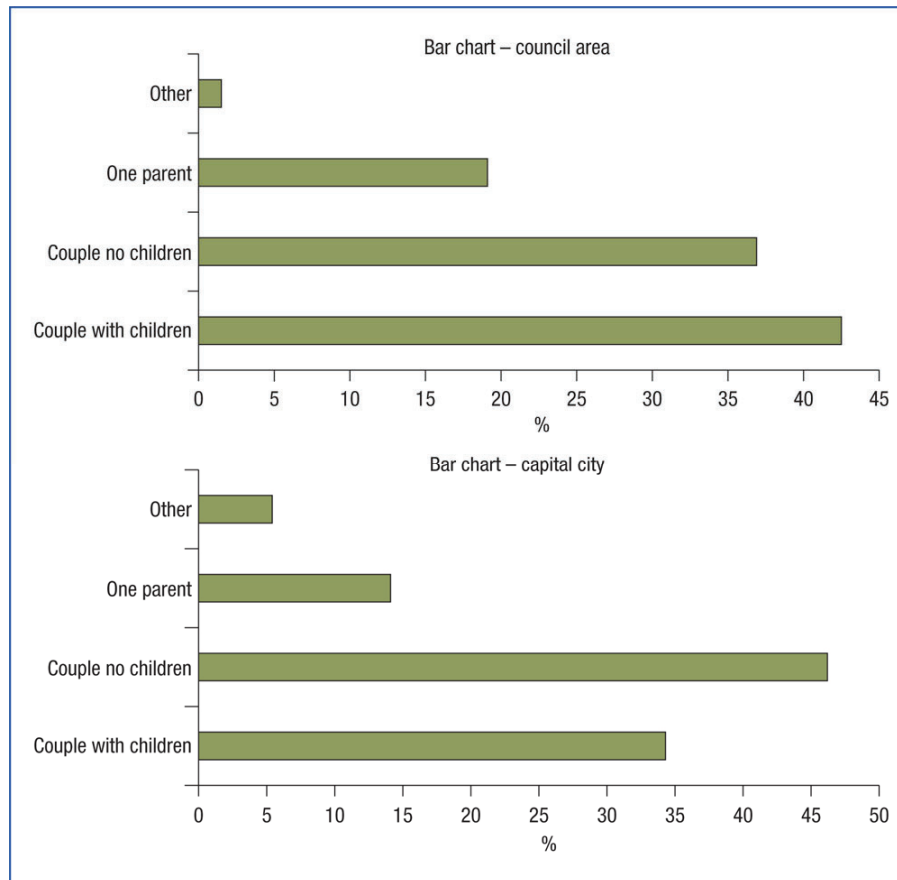


Figure 2.2
Microsoft Excel bar chart
for family type

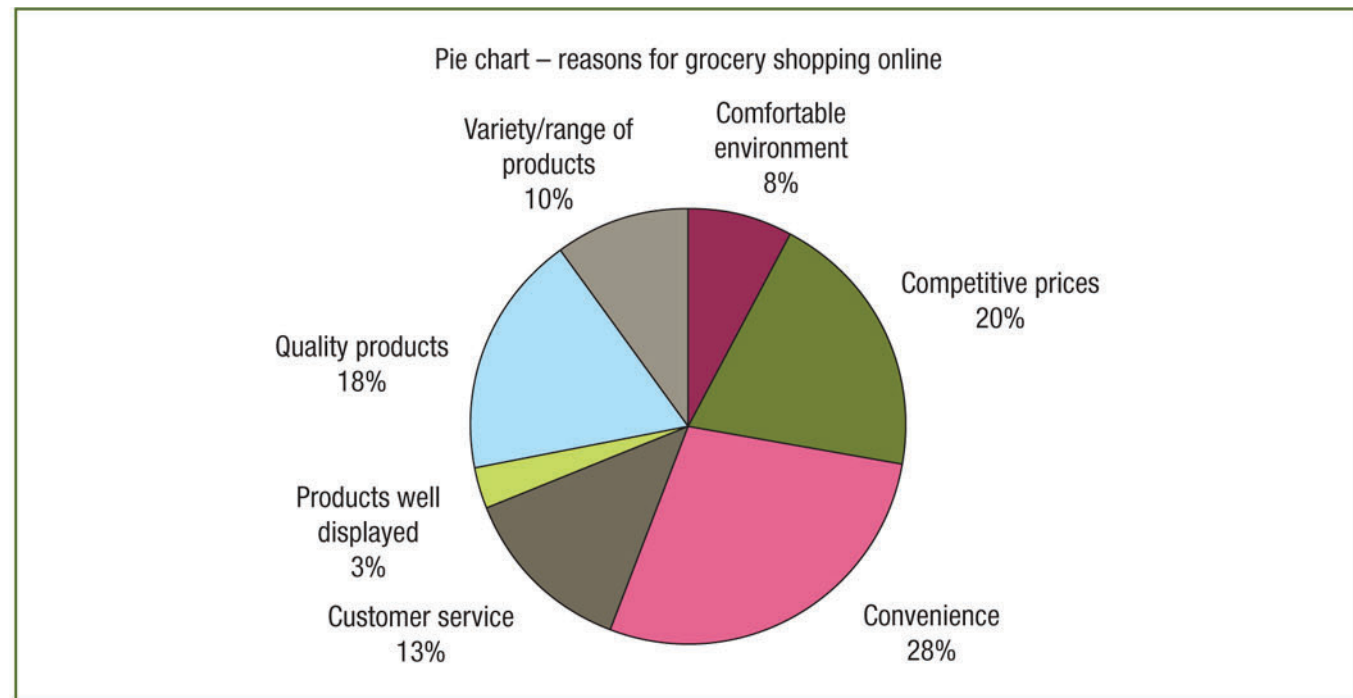
Microsoft® product screen shots are reprinted with permission from Microsoft Corporation.

+Pie Charts

8

Figure 2.3

Microsoft Excel pie chart
of the reasons for grocery
shopping online

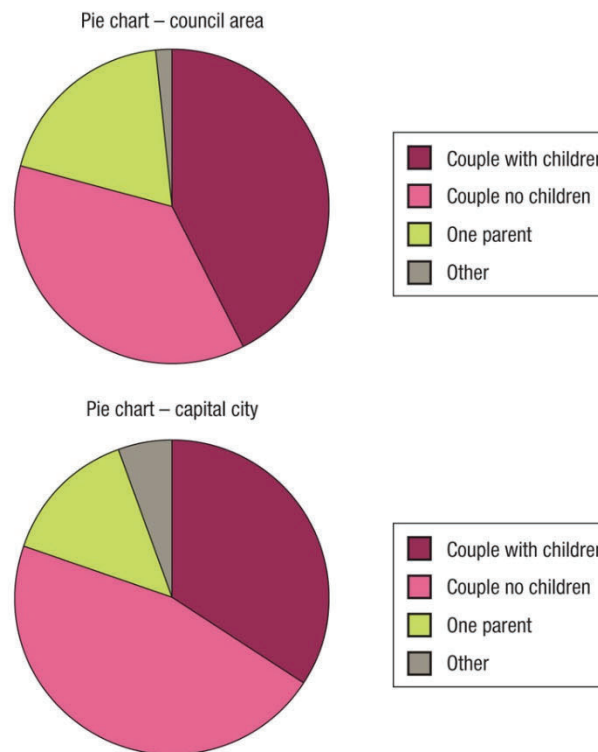


Microsoft® product screen shots are reprinted with permission from Microsoft Corporation.

+Pie Charts

Figure 2.4

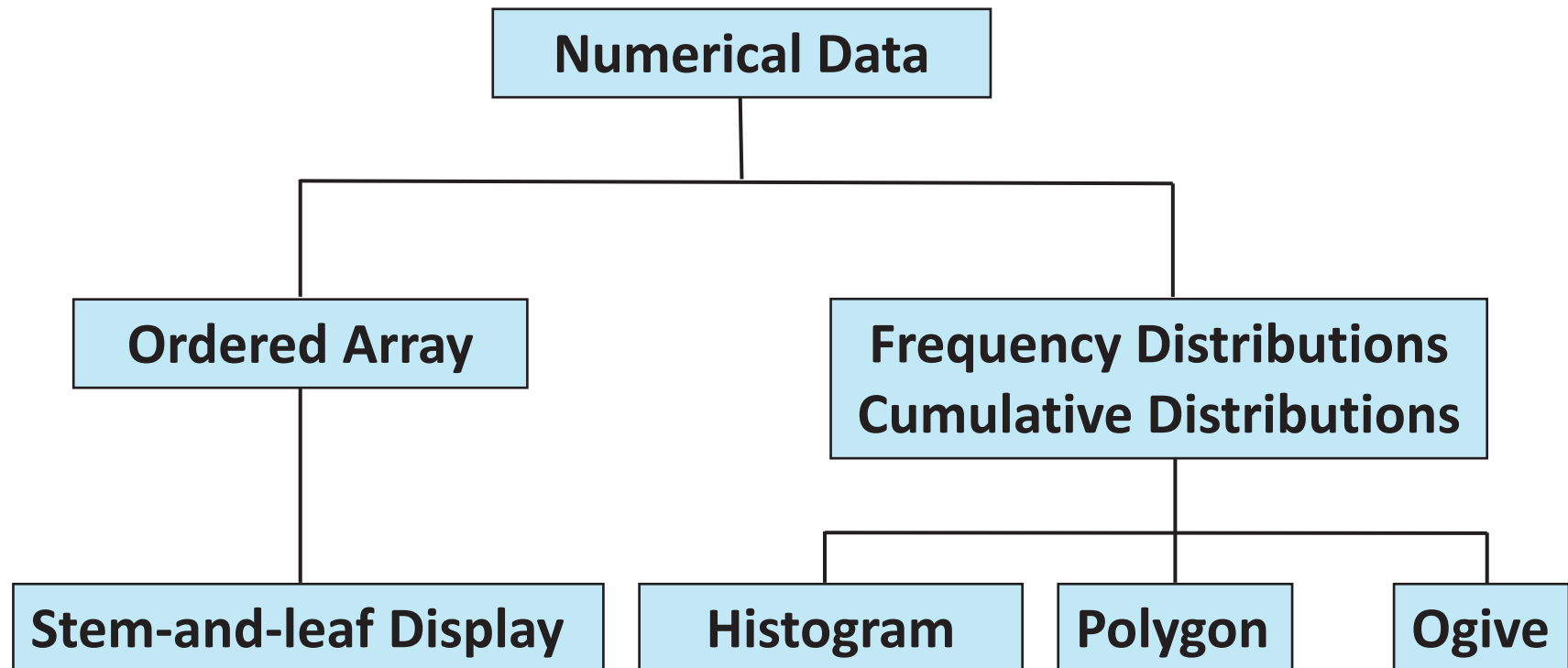
Microsoft Excel pie chart for family type



Microsoft® product screen shots are reprinted with permission from Microsoft Corporation.

+2.2 Organising Numerical data

10



+Ordered Arrays

11

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

+Ordered Arrays

Table 2.3 Price per main meal at 50 city restaurants and 50 suburban restaurants

City									
50	38	43	56	51	36	25	33	41	44
34	39	49	37	40	50	50	35	22	45
44	38	14	44	51	27	44	39	50	35
31	34	48	48	30	42	26	35	32	63
36	38	53	23	39	45	37	31	39	53
Suburban									
37	37	29	38	37	38	39	29	36	38
44	27	24	34	44	23	30	32	25	29
43	31	26	34	23	41	32	30	28	33
26	51	26	48	39	55	24	38	31	30
51	30	27	38	26	28	33	38	32	25

City									
14	22	23	25	26	27	30	31	31	32
33	34	34	35	35	35	36	36	37	37
38	38	38	39	39	39	39	40	41	42
43	44	44	44	44	45	45	48	48	49
50	50	50	50	51	51	53	53	56	63
Suburban									
23	23	24	24	25	25	26	26	26	26
27	27	28	28	29	29	29	30	30	30
30	31	31	32	32	32	33	33	34	34
36	37	37	37	38	38	38	38	38	38
39	39	41	43	44	44	48	51	51	55

Table 2.4 Ordered array of price per main meal at 50 city restaurants and 50 suburban restaurants

+Stem-and-Leaf Displays

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**)

An example: Data in an ordered array (21), 24, 27, 30, 32, (38), (41)

2	1	4	7
3	0	2	8
4	1		

21 is shown as →

38 is shown as →

41 is shown as →

Stem	Leaf
2	1
3	8
4	1

+Stem-and-Leaf Displays (cont)

14

Figure 2.5

PhStat2 stem-and-leaf display for festival expenditure by interstate visitors

Festival expenditure by interstate visitors

Stem unit: \$100

Leaf unit: \$10

2		2 7 8
3		1 2 3 5 9 9 9
4		0 2 3 3 5 5 6 7 8 8 9
5		1 2 5 5 6 8 9
6		0 0 0 3 3 3 4 6 6 8 9
7		3 5 6 7 7 8 9
8		0 6 7
9		1 1 4
10		4

+Frequency Distributions

15

What is a frequency distribution?

- A frequency distribution is 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?

- It is a way to summarise numerical data
- It condenses the raw data into a more useful form
- It allows for a quick visual interpretation of the data and first inspection of the shape of the data

+Frequency Distributions (cont)

Class Intervals and Class Boundaries

- Each data value belongs to one - and only one - class
- Each class grouping has the same width
- Determine the width of each interval by:

$$\text{Width of Interval} \cong \frac{\text{Range}}{\text{Number of desired class groupings}}$$

- Usually at least 5 - but no more than 15 - groupings
- Class boundaries must be mutually exclusive
- Classes must be collectively exhaustive
- Round up the interval width to get desirable endpoints

+Frequency Distributions (cont)

17

.....
Table 2.5 Frequency
distribution of the price per
main meal for 50 city
restaurants and 50 suburban
restaurants

Price of main meal (\$)	City frequency	Suburban frequency
\$10 but less than \$15	1	0
\$15 but less than \$20	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	<u>1</u>	<u>0</u>
Total	50	50

+Relative Frequency Distributions and Percentage Distributions

Price of main meal (\$)	City		Suburban	
	Relative frequency	Percentage	Relative frequency	Percentage
\$10 but less than \$15	0.02	2.0	0.00	0.0
\$15 but less than \$20	0.00	0.0	0.00	0.0
\$20 but less than \$25	0.04	4.0	0.08	8.0
\$25 but less than \$30	0.06	6.0	0.26	26.0
\$30 but less than \$35	0.14	14.0	0.26	26.0
\$35 but less than \$40	0.28	28.0	0.24	24.0
\$40 but less than \$45	0.16	16.0	0.08	8.0
\$45 but less than \$50	0.10	10.0	0.02	2.0
\$50 but less than \$55	0.16	16.0	0.04	4.0
\$55 but less than \$60	0.02	2.0	0.02	2.0
\$60 but less than \$65	0.02	2.0	0.00	0.0
Total	1.00	100.0	1.00	100.0

Table 2.7

Relative frequency distribution and percentage distribution of the price of main meals at city and suburban restaurants

Relative Frequency = Frequency/Total

% Frequency = (Frequency/Total) * 100%

+Histograms

19

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

+Histograms (cont)

20

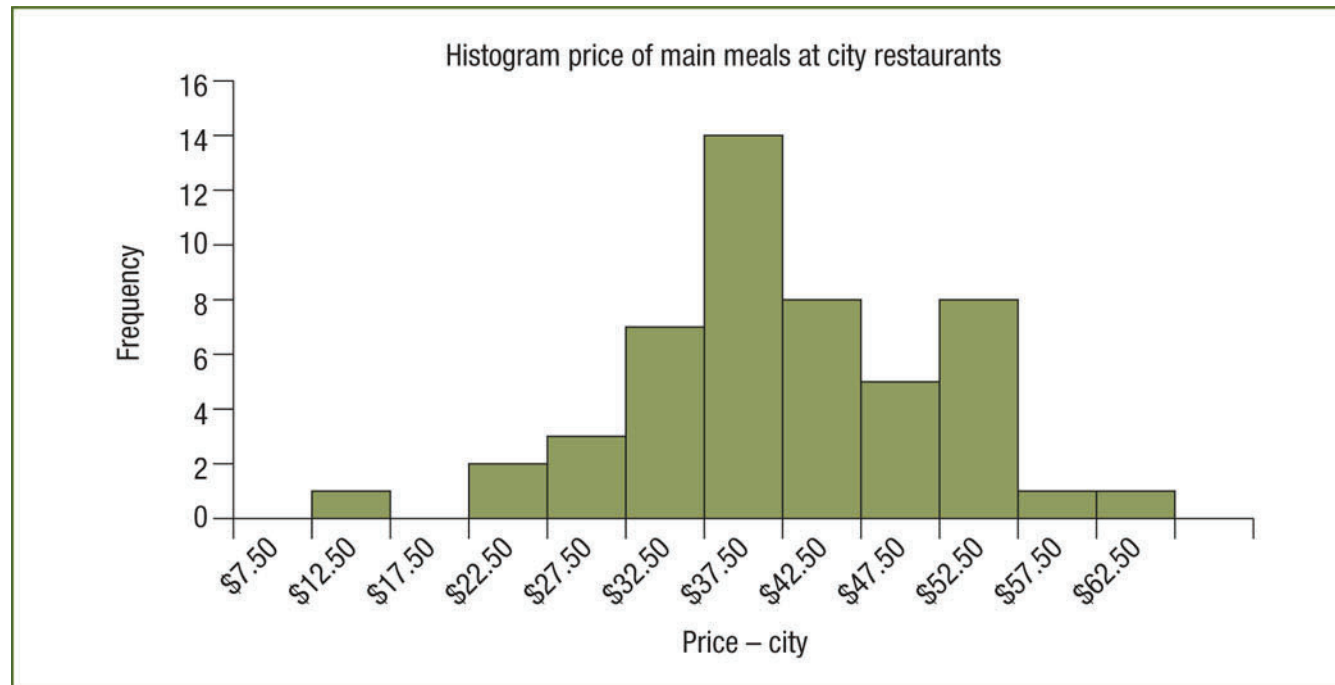


Figure 2.6

Excel histogram of the price of main meals at city restaurants

Microsoft® product screen shots are reprinted with permission from Microsoft Corporation.

+Polygons

21

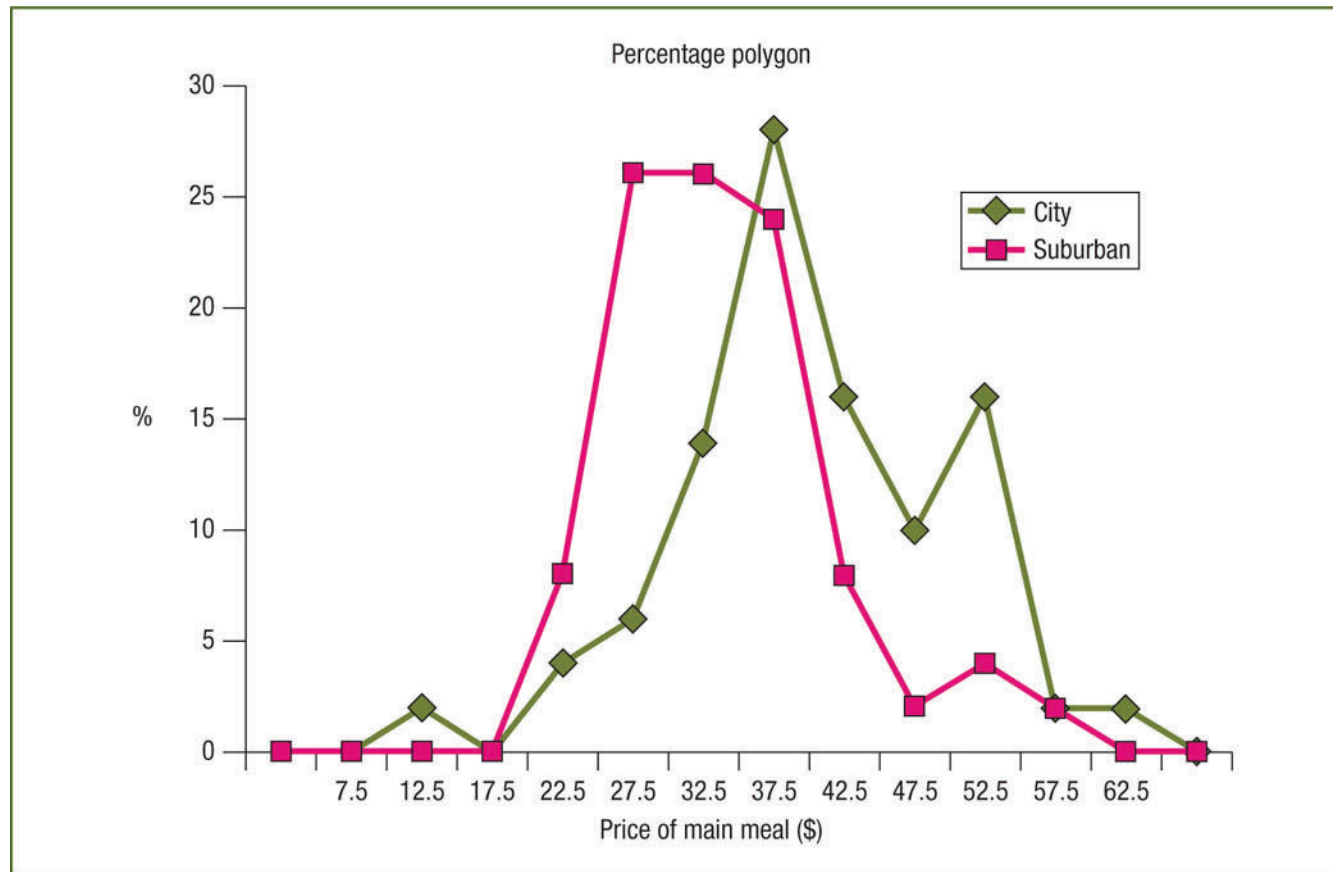


Figure 2.8

Percentage polygons for the price of main meals in city and suburban restaurants

+Cumulative Distributions

Table 2.9 Cumulative percentage distributions of

urban

Price of main meal (\$)	City		Suburban	
	Relative frequency	Percentage	Relative frequency	Percentage
\$10 but less than \$15	0.02	2.0	0.00	0.0
\$15 but less than \$20	0.00	0.0	0.00	0.0
\$20 but less than \$25	0.04	4.0	0.08	8.0
\$25 but less than \$30	0.06	6.0	0.26	26.0
\$30 but less than \$35	0.14	14.0	0.26	26.0
\$35 but less than \$40	0.28	28.0	0.24	24.0
\$40 but less than \$45	0.16	16.0	0.08	8.0
\$45 but less than \$50	0.10	10.0	0.02	2.0
\$50 but less than \$55	0.16	16.0	0.04	4.0
\$55 but less than \$60	0.02	2.0	0.02	2.0
\$60 but less than \$65	0.02	2.0	0.00	0.0
Total	1.00	100.0	1.00	100.0

Price (\$)	City percentage of restaurants less than indicated value	Suburban percentage of restaurants less than indicated value
\$10	0	0
\$15	2	0
\$20	2	0
\$25	6	8
\$30	12	34
\$35	26	60
\$40	54	84
\$45	70	92
\$50	80	94
\$55	96	98
\$60	98	100
\$65	100	100

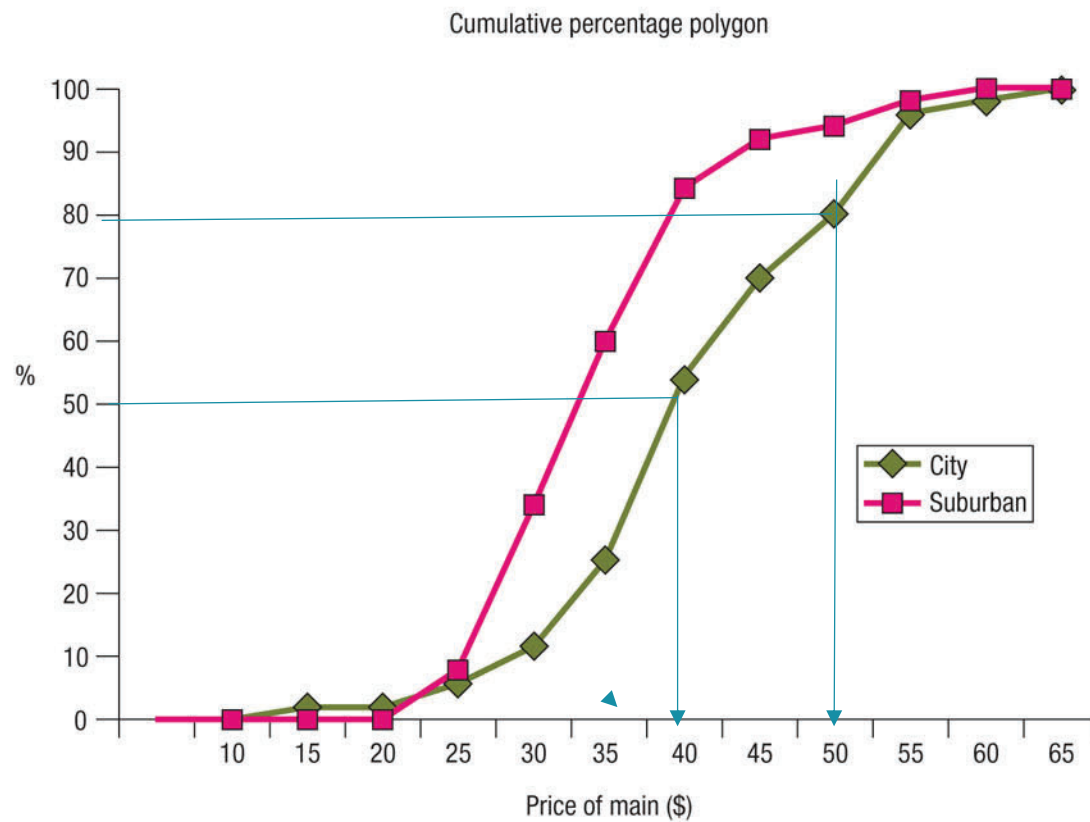
Cumulative Frequency = Number of observations that lie below or equal to a particular value/interval in a given data set

+ Cumulative Percentage Polygons (Ogives)

23

Figure 2.10

Cumulative percentage polygons of the cost of main meals at city and suburban restaurants



+Roadmap for Selecting Tables and Charts

Table 2.16

Roadmap for selecting tables and charts

Type of analysis	Type of data	
	Numerical	Categorical
Tabulating, organising and graphically presenting the values of a variable	Ordered array, stem-and-leaf display, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution, histogram, polygon, cumulative percentage polygon (Sections 2.2 and 2.3)	Summary table, bar chart, pie chart (Section 2.1)
Graphically presenting the relationship between two variables	Scatter diagram, time-series plot (Section 2.5)	Contingency table, side-by-side bar chart (Section 2.4)

+2.6 Misusing Graphs and Ethical Issues

Do not distort the data

- Frequency/quantity should be proportional to the area/volume

Avoid unnecessary adornments

- No 'chart junk'

Use a scale for each axis on a two-dimensional graph

- Should be properly scaled along each axis
- All axes should be labelled

+2.6 Misusing Graphs and Ethical Issues (cont)

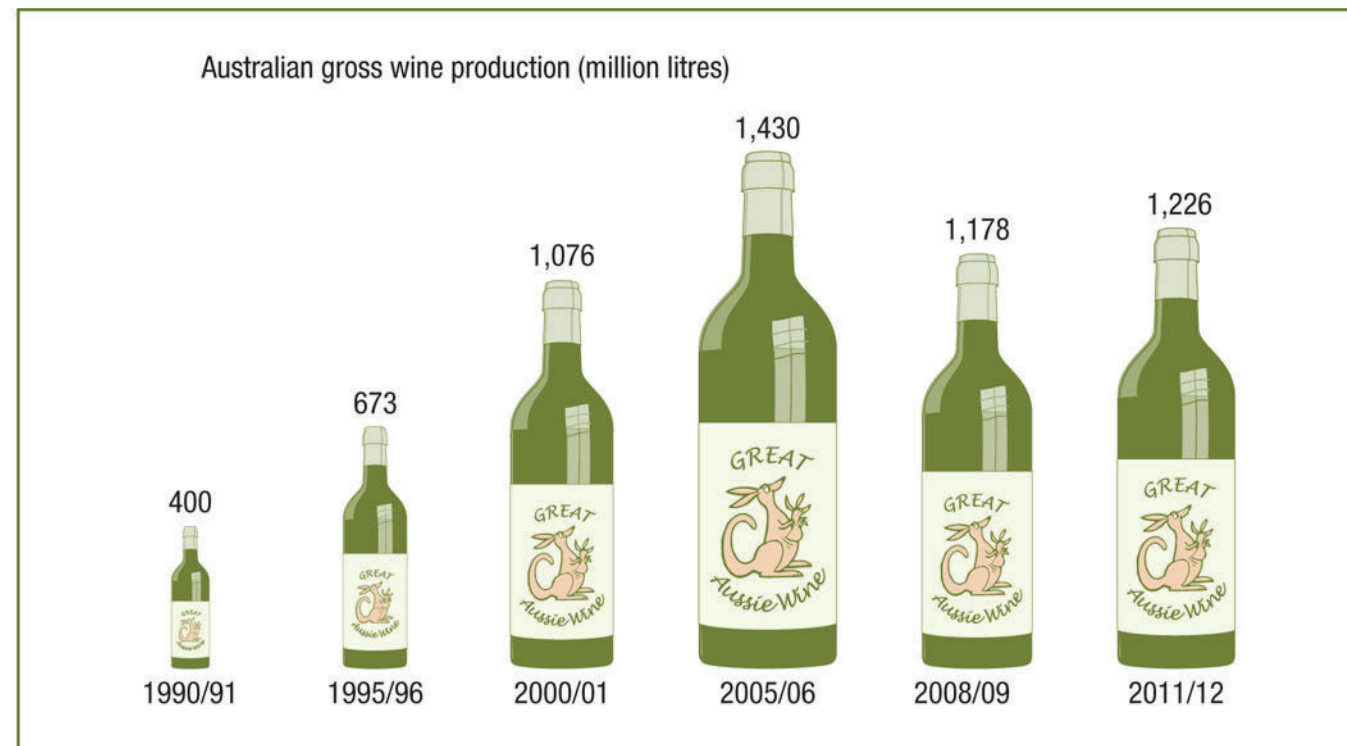
- The vertical axis scale should begin at zero unless there is justification for truncation (which must be clearly labelled and explained to the reader)
- The graph should contain a title
- Use the simplest graph for a given set of data

+2.6 Misusing Graphs and Ethical Issues (cont)

Figure 2.16

Misleading display of
Australian wine production

Source: Data obtained from
'Australian Gross Wine
Production – pdf format',
Wine Australia Corporation
<[www.wineaustralia.com/
australia](http://www.wineaustralia.com/australia)> accessed
December 2013.

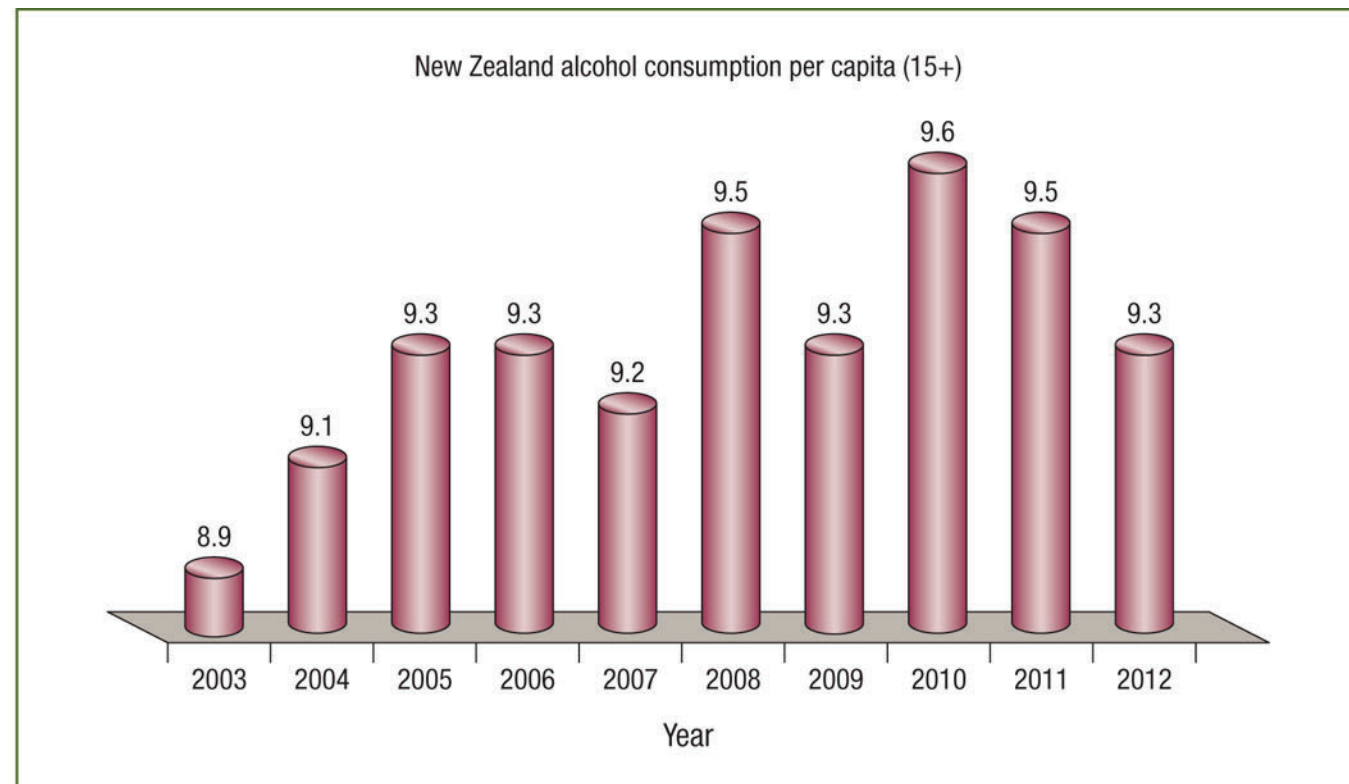


+2.6 Misusing Graphs and Ethical Issues (cont)

Figure 2.17

Misleading display of New Zealand alcohol consumption

Source: Data from OECD (2011 and 2013), 'Alcohol consumption', Health: Key Tables from OECD, No. 24. doi: 10.1787/alcoholcons-table-2013-2-en and 10.1787/alcoholcons-table-2011-1-en, accessed December 2013.



+Ethical Concerns

29

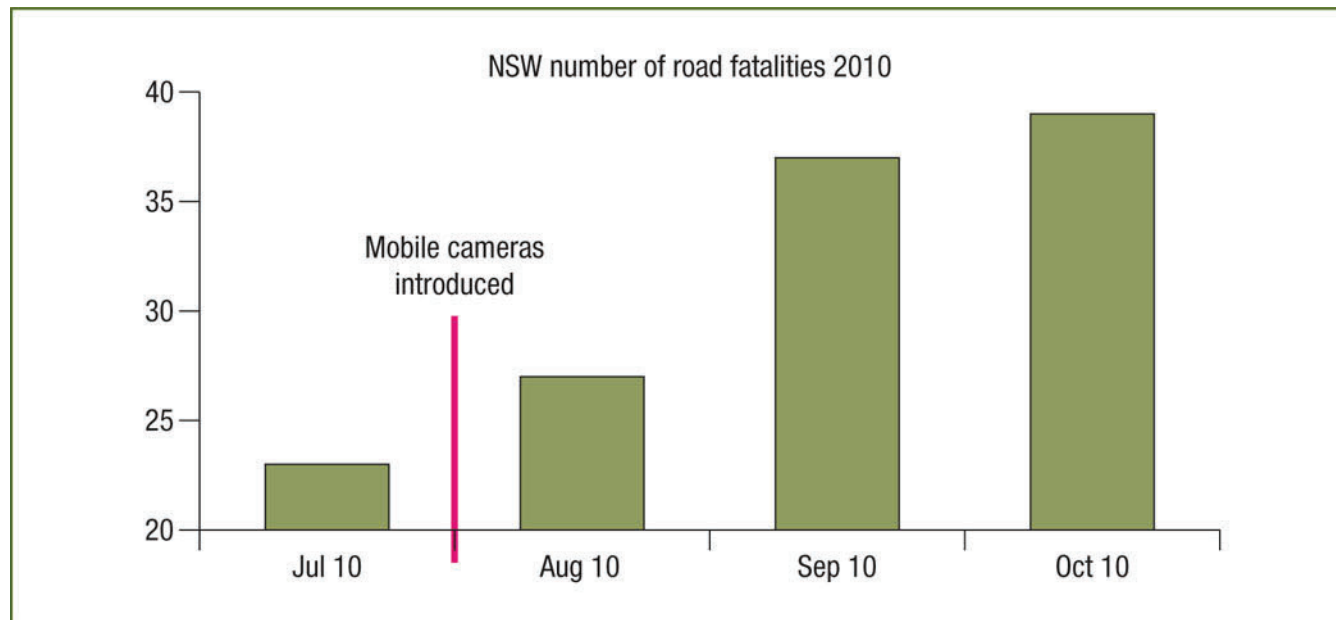


Figure 2.18a

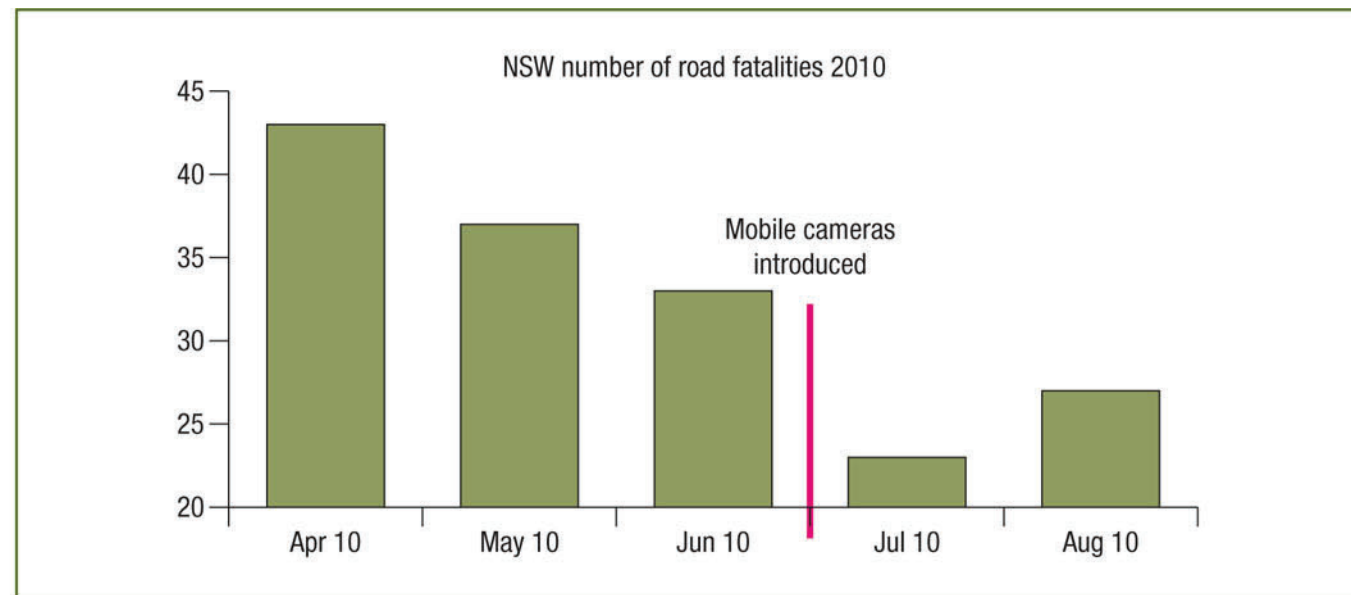
NSW road fatalities 2010

+Ethical Concerns

30

Figure 2.18b

NSW road fatalities 2010



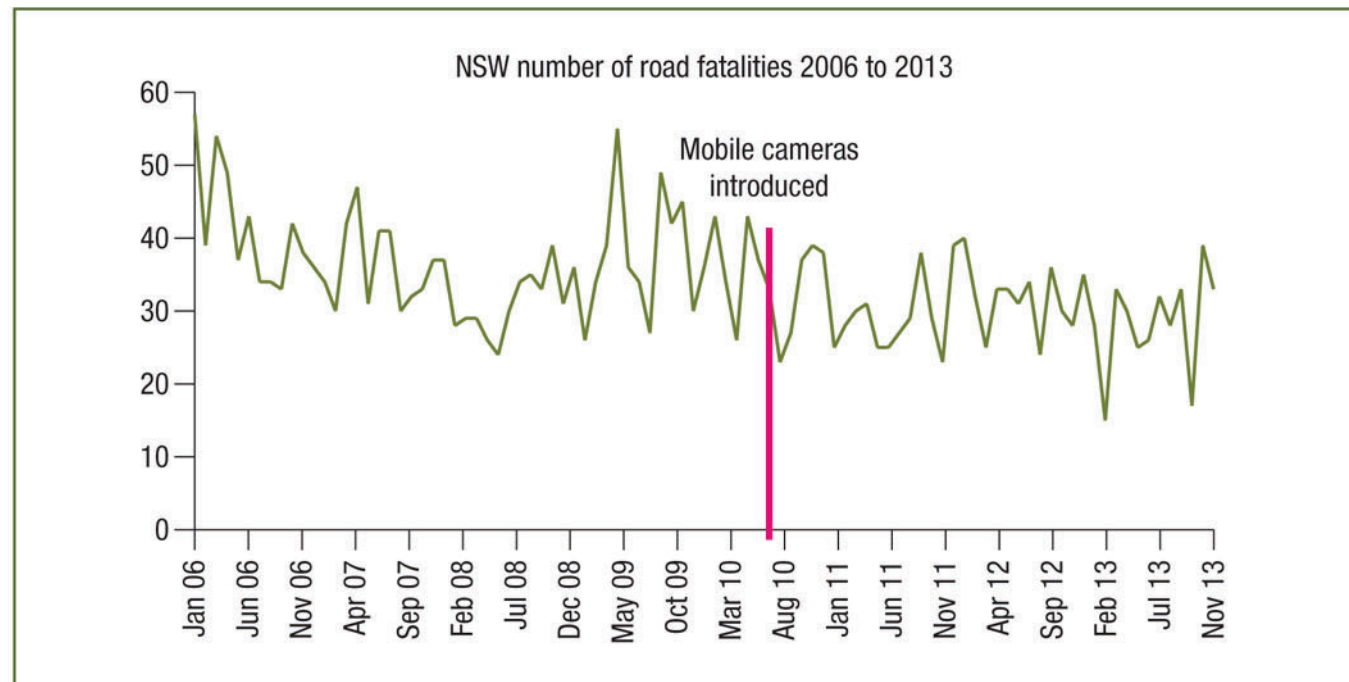
+Ethical Concerns

31

Figure 2.18c

NSW road fatalities 2006 to 2013

Source: Data in Figures 2.18(a)–(c) obtained from Australian Road Deaths Database, <www.bitre.gov.au/statistics/safety/fatal_road_crash_database.aspx>, accessed 27 December 2013.



+Misleading graphs in real life



Image: Media Matters

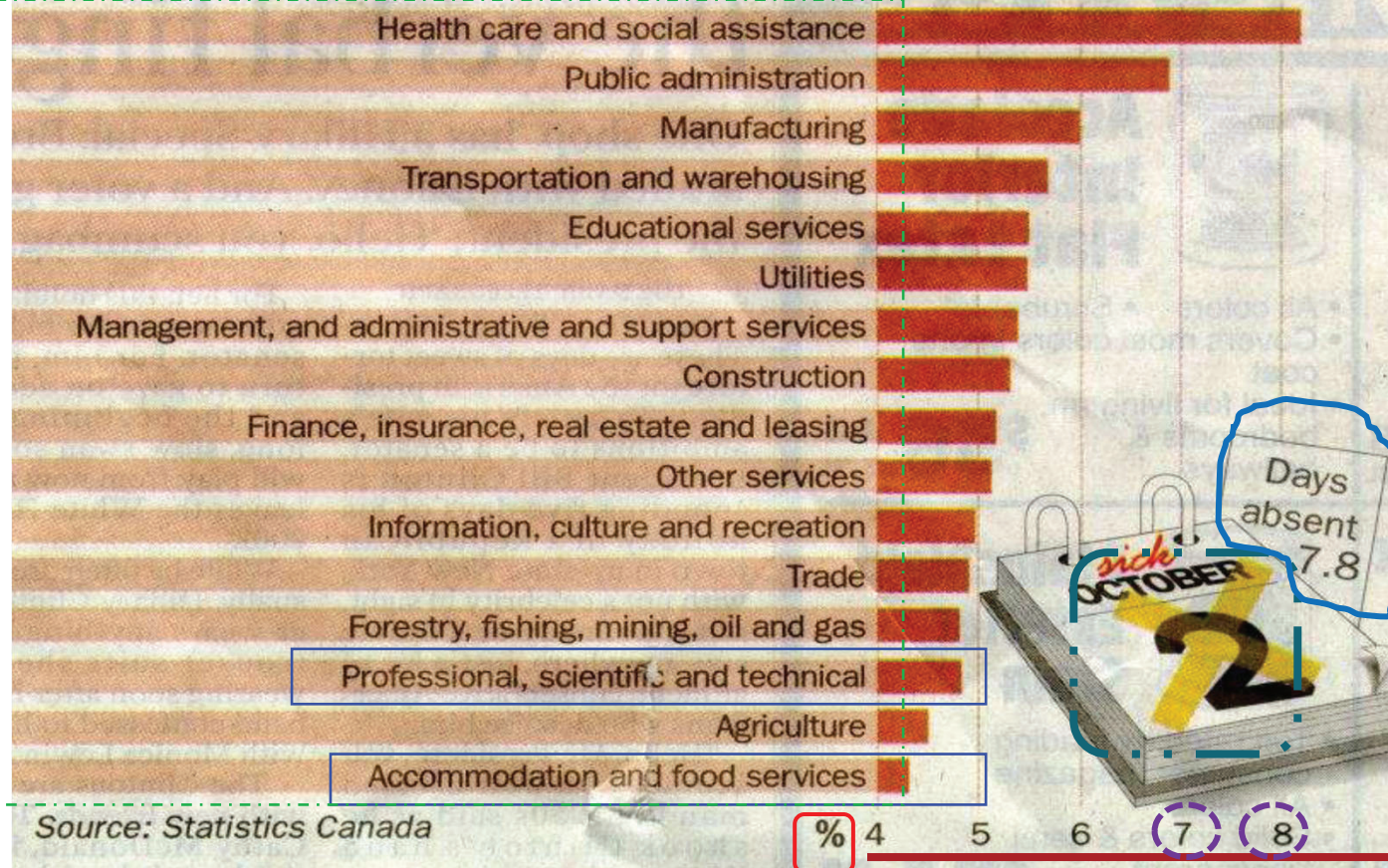
*"Fox's 108.6 million figure for the number of 'people on welfare' comes from a Census Bureau's account...of participation in means-tested programs, which include 'anyone residing in a household in which one or more people received benefits' in the fourth quarter of 2011, thus **including individuals who did not themselves receive government benefits**. On the other hand, the 'people with a full time job' figure Fox used included only individuals who worked, not individuals residing in a household where at least one person works."*

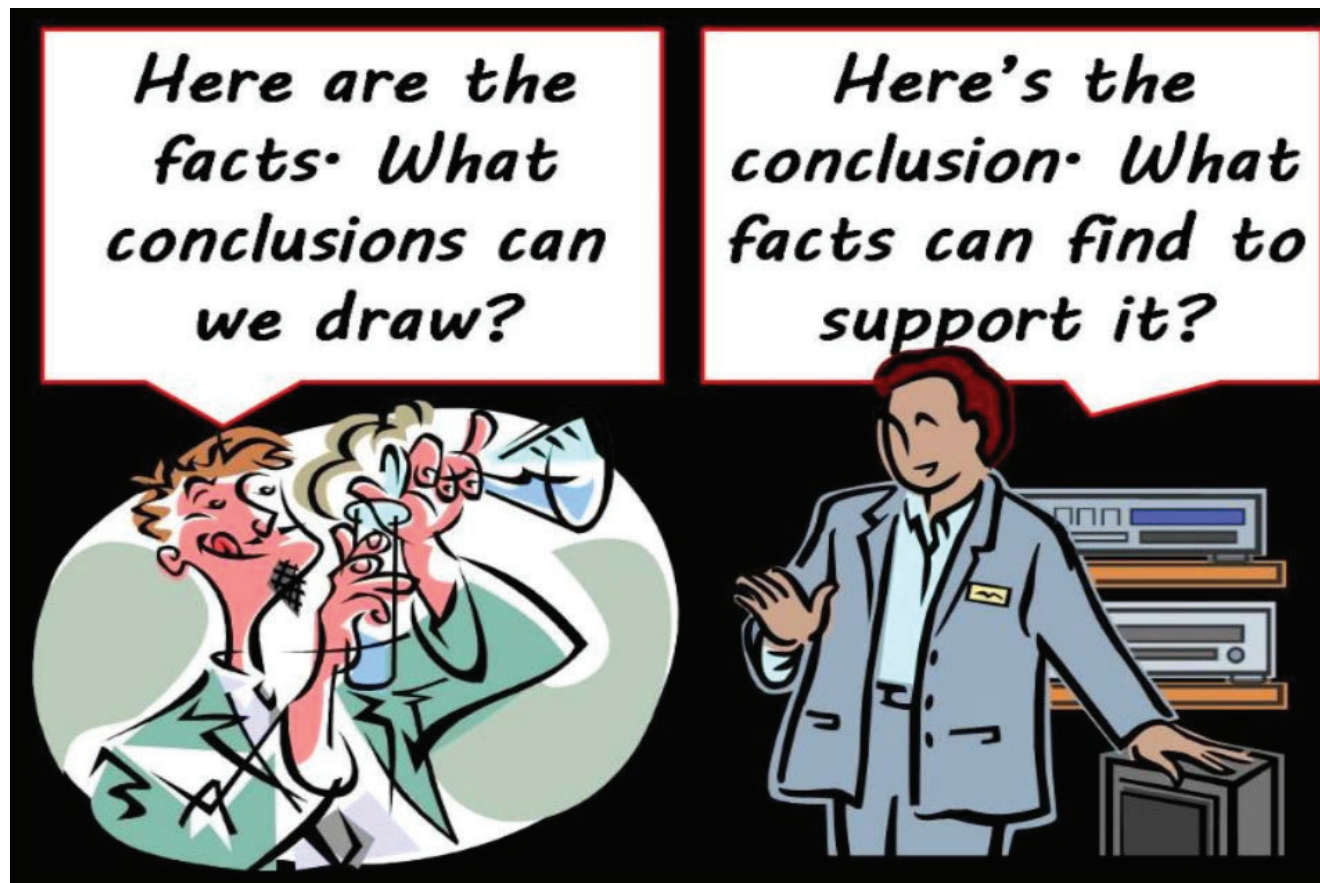
+

Absenteeism rates up

Workdays missed by full-time employees for personal reasons rose from an average of 7.4 in 1997 to 7.8 in 1998.

Workdays missed by full-time employees for personal reasons - 1998





Source: Unknown



+ TTD Week Two

By the end of the week 2 make sure you ...

- Can construct 'basic' visualisations that are appropriate for a particular data set
- Are able to critique/analyse visualisations drawn by others
- Have completed the relevant questions from the textbook
- Read the suggested sections from Chapters 2 of the textbook.
- Summarise the key terms introduced this week
- Read over the assignment one details