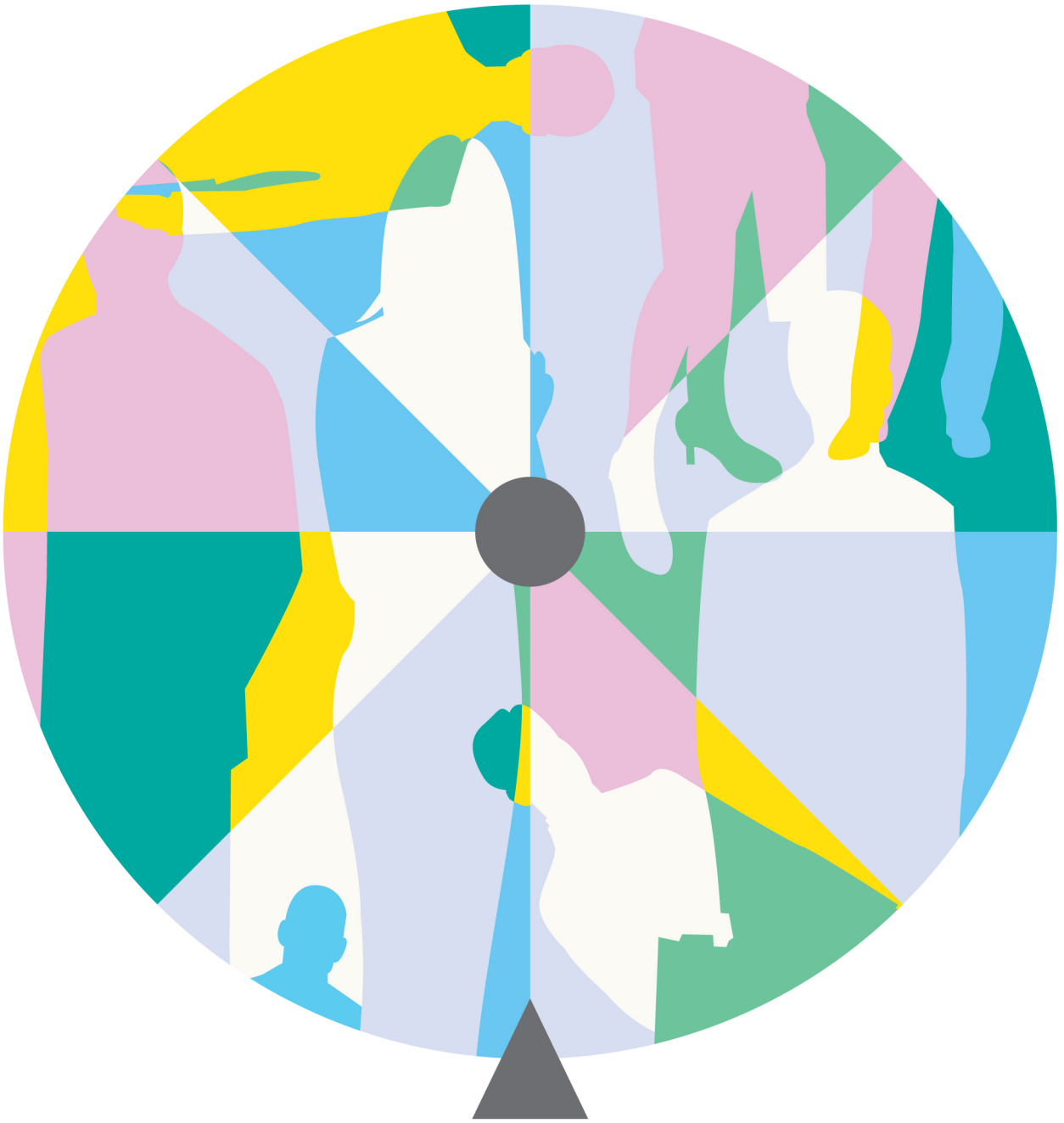


Australia / New Zealand / Edition 7

BUSINESS STATISTICS ABRIDGED

ELIYATHAMBY A SELVANATHAN,
SAROJA SELVANATHAN, GERALD KELLER

Solutions Manual

Solutions Manual

Table of Contents

1 What is statistics?

1.1 a **Population:** The collection of all measurements of interest in a statistical problem; e.g., heights of all Australians.

b **Sample:** Any subset of measurements from a population; e.g., heights of 100 selected Australians.

c **Parameter:** A descriptive measure of the measurements in a population; e.g., average height of all Australians.

d **Statistic:** A descriptive measure of the measurements in a sample; e.g., average height of 100 selected Australians.

e **Statistical inference:** A conclusion about a characteristic of a population, based on the information provided by a sample drawn from the population; e.g., testing whether the average height of all Australian adults is greater than 150cm using the sample information based on the heights of 100 randomly selected Australian adults.

1.2 Descriptive statistics consists of graphical and numerical methods used to describe sets of data, both populations and samples. Inferential statistics consists of a body of methods used for drawing conclusions about characteristics of a population, based on information available in a sample drawn from the population.

1.3 a Views on internet banking of the 12 000 customers

b Views on internet banking of the 300 customers surveyed

c Statistic.

1.4 a The complete production run of light bulbs

b 1000 bulbs selected

c The proportion of the light bulbs that are defective in the whole production run.

d The proportion of bulbs that are defective in the sample of 1000 bulbs selected.

e Parameter

f Statistic

g Because the sample proportion (1%) is much less than the claimed 5%, we can conclude with the confidence that there is evidence to support the claim.

1.5 Select a number of graduates (say 100) from each group (Business group and Arts and Science group) and workout the sample mean salaries of the two groups. Compare the two mean values to see whether there is some support for the claim that in general, average salary of business graduates is greater than the average salary of arts and science graduates.

1.6 **a** Flip the coin 100 times and count the number of heads and tails

b Outcomes of repeated flips

c Outcomes of the 100 flips

d The proportion of heads

e The proportion of heads in the 100 flips.

f If the sample proportion is close to 0.5, we conclude that there is some support for the claim that the coin is a fair coin.

1.7 **a** The coin is not a fair coin.

b The coin may be a fair coin. Need more trials (say, 1000) to confirm this.

c If it is not a fair coin, the answer is no. The number of heads and tails can be anywhere between 0 to 100, out of 100 trials (for example, 60 heads and 40 tails, or 30 heads and 70 tails). If it is a fair coin, one would expect the number of heads and tails to be close to 50, out of 100 trials.

2 Types of data, data collection and sampling

2.1 **Numerical:** **a** Kilometres commuted to work
b Age of students in a statistics class.

Ordinal: **a** Fortnightly Australian family income
i Under $1500
ii $1500–2000
iii $2000–2500
iv $2500 or over.
b Patient's condition: excellent, good, fair or poor.

Nominal: **a** Country of origin of Australians
b Type of car owned.

2.2 **a** Numerical
b Nominal
c Ordinal
d Numerical.

2.3 **a** Nominal
b Nominal
c Numerical
d Ordinal
e Numerical.

2.4 **a** Numerical
b Nominal
c Nominal
d Ordinal
e Numerical
f Ordinal.

2.5 **a** Ordinal
b Numerical
c Nominal
d Numerical
e Numerical.

2.6 **a** Numerical
b Ordinal
c Nominal
d Numerical
e Ordinal
f Nominal.

2.7 **a** Nominal
b Numerical
c Nominal
d Numerical
e Ordinal.

2.8 **a** Numerical
b Ordinal
c Nominal
d Numerical
e Nominal
f Ordinal.

2.9 **a** Numerical
b Numerical
c Nominal
d Ordinal
e Numerical.

2.10 **a** Ordinal
b Ordinal
c Ordinal
d Numerical.

2.11 Primary data are published by the original source. Secondary data are published by someone other than whoever originally collected and published the data. Secondary data sources often summarise much of the original data, resulting in a loss of some information.

2.12 **a** *Australian Bureau of Statistics*; Year Book, Australia (annual); rate of unemployment, population
b *Reserve Bank Bulletin* (monthly); interest rate, exchange rate
c *CIA Fact Book* (annual); electricity consumption, flags of the world
Note: The two specific pieces of information contained in the latest issue of these publications will, of course, vary considerably, unless the instructor is more specific about the information requested.

2.13 In an observational study, there is no attempt to control factors that might influence the variable of interest. In an experimental study, a factor (such as regular use of a fitness centre) is controlled by randomly selecting who is exposed to that factor, thereby reducing the influence of other factors on the variable of interest.

2.14 **a** This is an observational study, because no attempt is made to control factors that might influence cola sales, such as store location or store type.
b Randomly select which stores (both grocery and convenience) receive cola in bottles to reduce the influence of factors like location. Separately analyse the two types of stores in order to reduce the influence of store type.

2.15 **a** Randomly select 4000 people over the age of 50. Compare the proportion of smokers who have lung cancer with the proportion of non-smokers who have lung cancer.

b The study described in part a is observational, because we haven't controlled who smoked.

2.16 **a** A survey can be conducted, for example, by means of a personal interview, a telephone interview, or a self-administered questionnaire.

b A personal interview has a high response rate relative to other survey methods, but is expensive because of the need to hire well-trained interviewers and possibly pay travel-related costs if the survey is conducted over a large geographical area. A personal interview will also probably result in fewer incorrect responses arising from respondents misunderstanding some questions. A telephone interview is less expensive, but will probably result in a lower response rate. A self-administered questionnaire is least expensive, but suffers from lower response rates and accuracy than personal interviews.

2.17 Five important points to consider when designing a questionnaire are as follows:

- The questionnaire should be short.
- Questions should be clearly worded and unambiguous.
- Consider using dichotomous or multiple-choice questions, but take care that respondents needn't make unspecified assumptions before answering the questions.
- Avoid using leading questions.
- When preparing the questions, think about how you intend to tabulate and analyse the responses.

2.18 **a** The sampled population will exclude those who avoid large department stores in favour of smaller shops, as well as those who consider their time too valuable to spend participating in a survey. The sampled population will therefore differ from the target population of all customers who regularly shop at the mall.

b The sampled population will contain a disproportionate number of thick books, because of the manner in which the sample is selected.

c The sampled population consists of those eligible voters who are at home in the afternoon, thereby excluding most of those with full-time jobs (or at school).

2.19 We used Excel to generate 40 three-digit random numbers. Because we will ignore all randomly generated numbers over 800, we can expect to ignore about 20% (or about 8 to 10) of the randomly generated numbers. We will also ignore any duplications. We therefore chose to generate 40 three-digit random numbers, and will use the first 25 unique random numbers less than 801 to select our sample. The 40 numbers generated are shown below, with a stroke through those to be ignored.

6	357	456	449	~~862~~	154	55	412	475	430
~~999~~	~~912~~	60	207	717	651	10	294	327	165
576	~~871~~	~~990~~	354	390	540	~~893~~	181	496	~~870~~
738	~~820~~	32	~~963~~	160	32	231	86	~~970~~	46

2.20 We used Excel to generate 30 three-digit random numbers. Because we will ignore any duplicate numbers generated, we generated 30 three-digit random numbers and will use the first 20 unique random numbers to select our sample. The 30 numbers generated are shown below.

169	470	744	530	554	918
318	858	698	203	383	938
836	116	123	936	539	154
110	630	856	380	145	692
909	269	811	274	553	749

2.21 The operations manager can select stratified random samples where the strata are the four departments. Simple random sampling can be conducted in each department.

2.22 Stratified random sampling is recommended. The strata are the school of business, the faculty of arts, the graduate school and the all the other schools and faculties would be the fourth stratum. The data can be used to acquire information about the entire campus but also compare the four strata.

2.23 A stratified random sampling plan accomplishes the president's goals. The strata are the four areas enabling the statistics practitioner to learn about the entire population but also compare the four areas.

2.24 **a** Sampling error refers to an inaccuracy in a statement about a population that arises because the statement is based only on sample data. We expect this type of error to occur because we are making a statement based on incomplete information. Nonsampling error refers to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

b Nonsampling error is more serious because, unlike sampling error, it cannot be diminished by taking a larger sample.

2.25 Three types of nonsampling errors:

- Error due to incorrect responses
- Nonresponse error, which refers to error introduced when responses are not obtained from some members of the sample. This may result in the sample being unrepresentative of the target population.
- Error due to selection bias, which arises when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample.

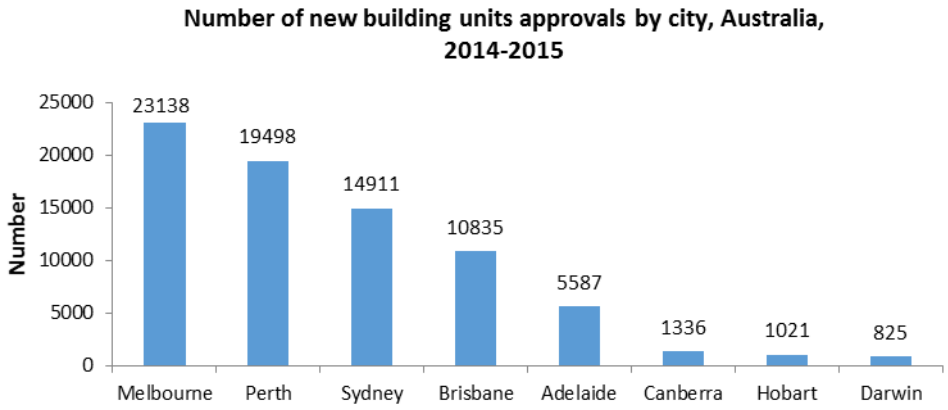
2.26 Yes. A census will probably contain significantly more nonsampling errors than a carefully conducted sample survey.

Part 1

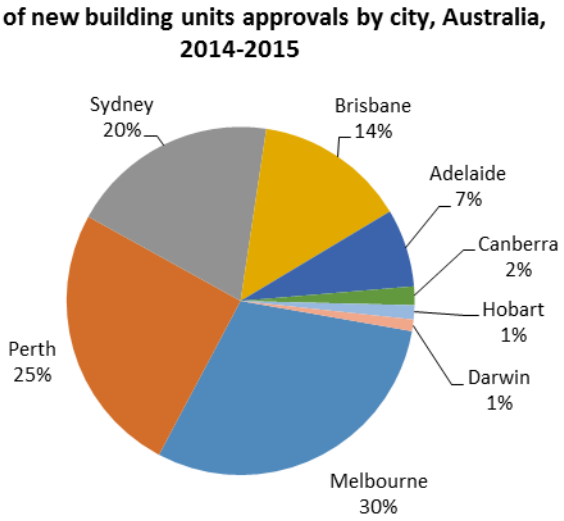
Descriptive measures and probability

3 Graphical descriptive techniques – nominal data

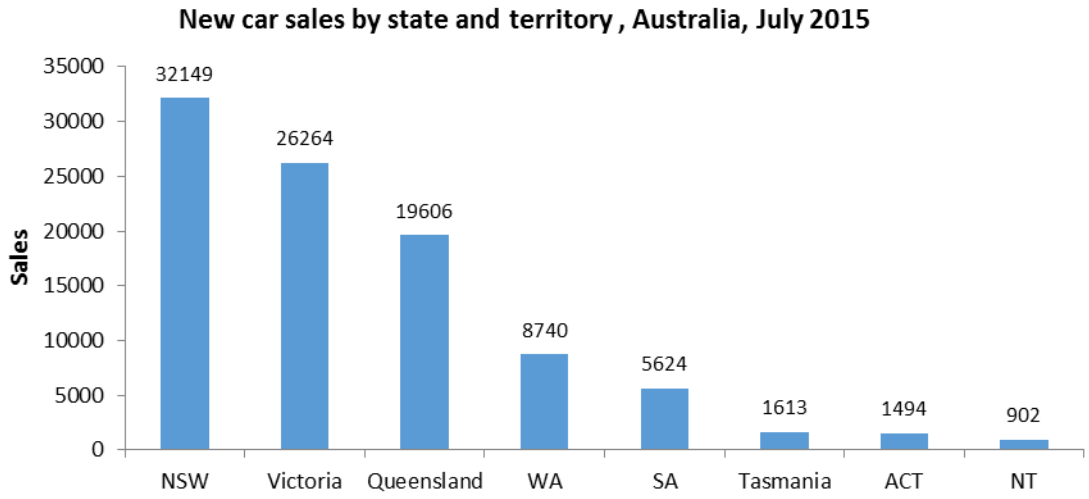
3.1 a



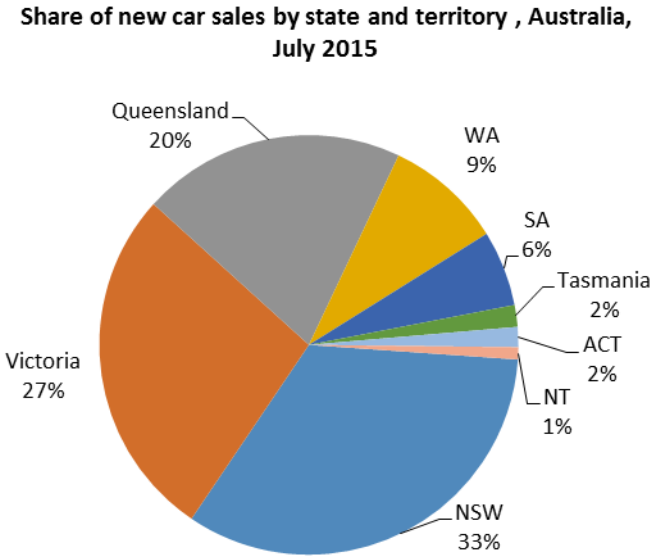
b



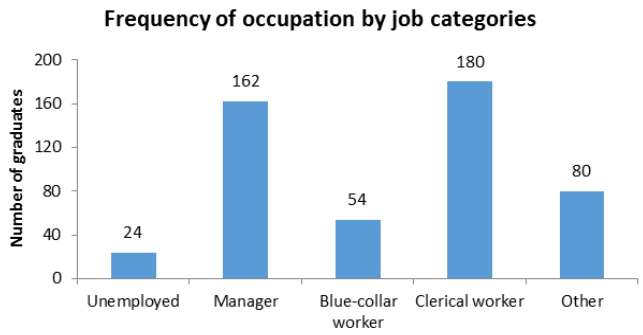
3.2 **a** A bar chart would be appropriate.

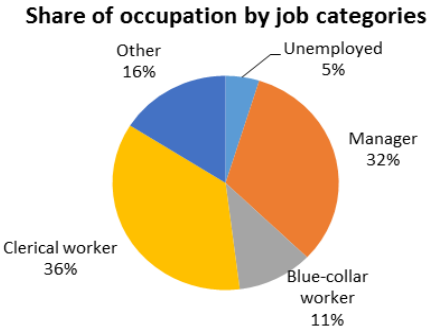


b A pie chart would be appropriate.

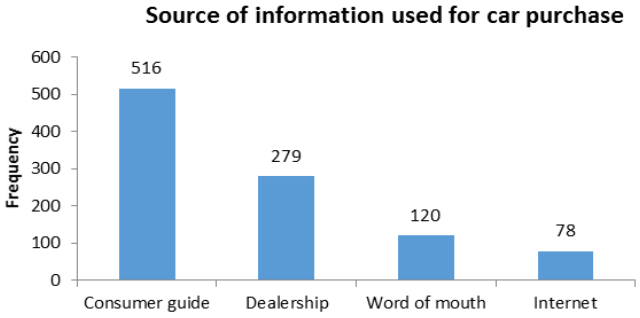


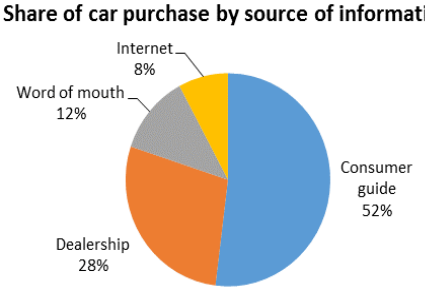
3.3 A pie or bar chart can be used; however, a pie chart is more suitable here.



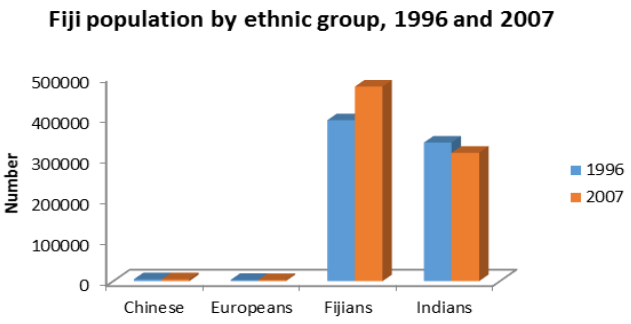


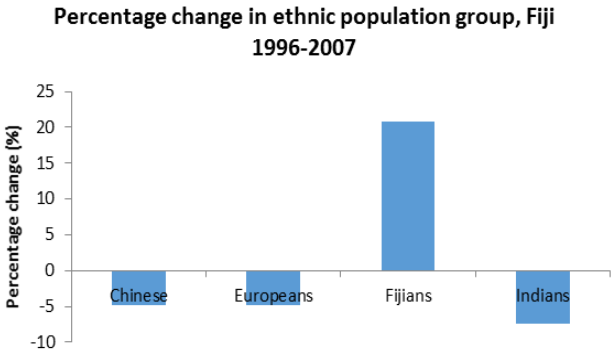
3.4 A pie or bar chart can be used; however, a pie chart is more suitable here.

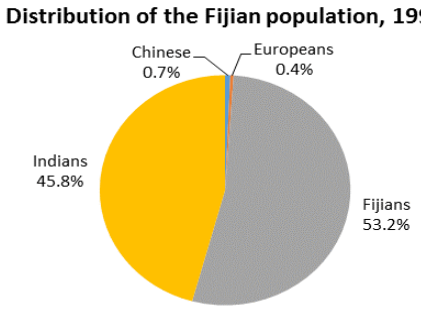


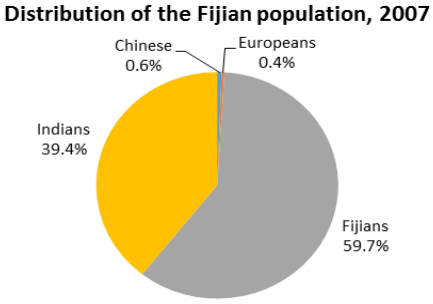


3.5

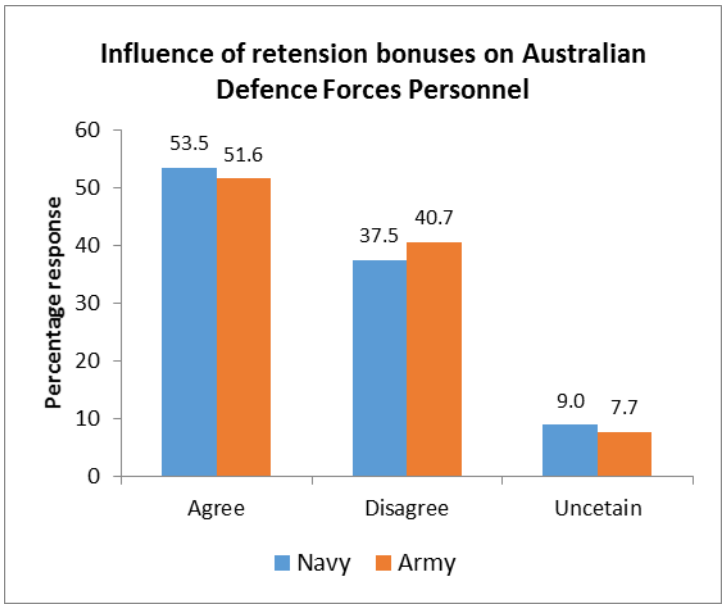


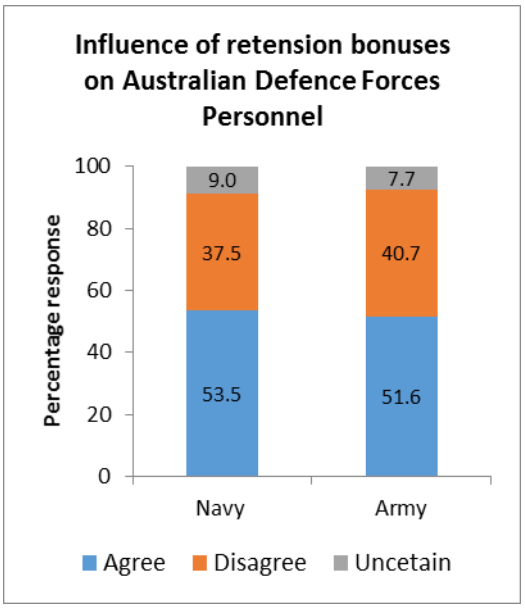




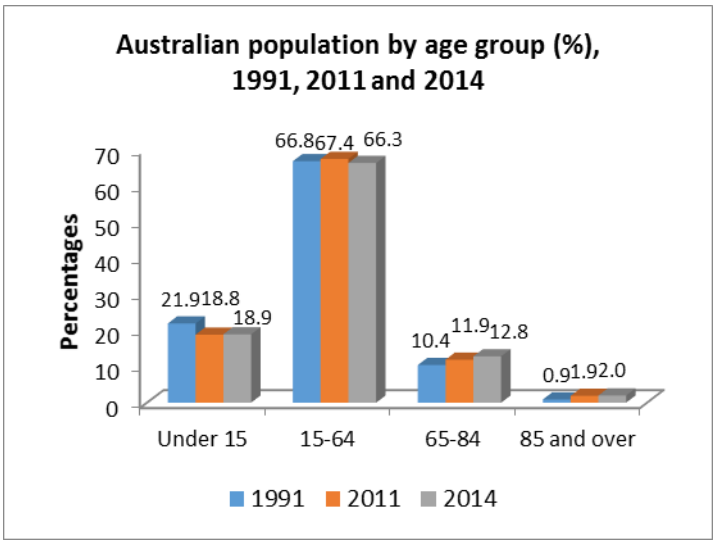


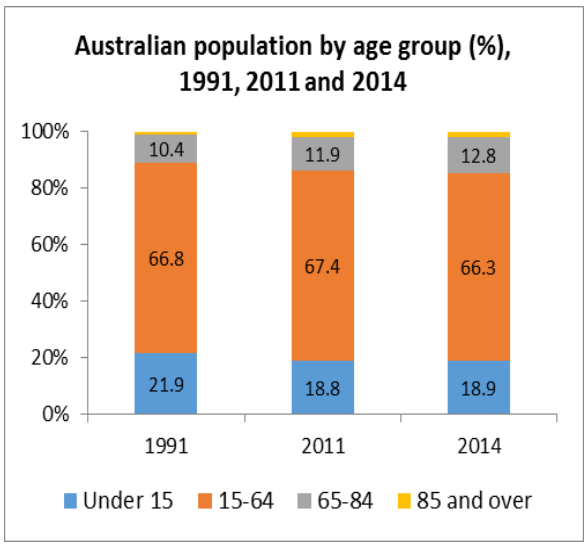
3.6 To compare the responses of the navy and army personnel a bar chart could be used. On the other hand, to compare the responses within the navy and within the army personnel, a component bar chart could be used. A component bar chart can also be used to compare the responses of the navy and army personnel as well.





3.7

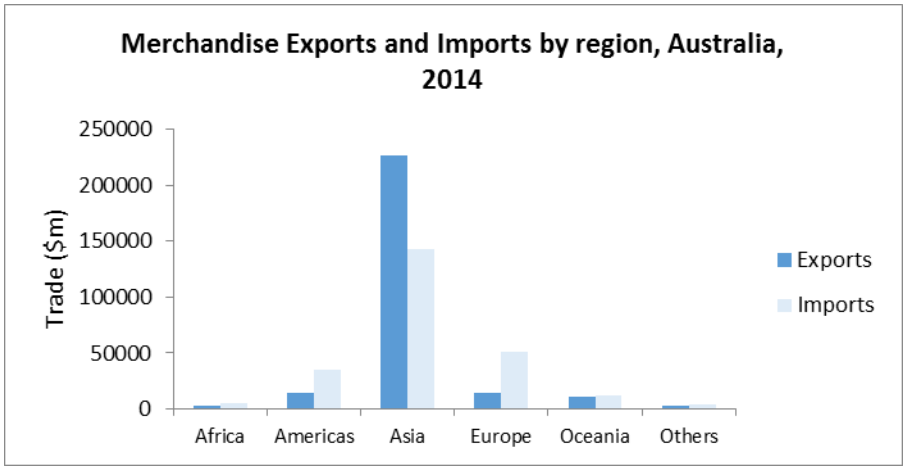




3.8. To compare the exports and imports in a particular year to and from the 6 trading partner regions, bar charts in (i) and (ii) would be useful. To compare the exports during 2012 and 2014, bar chart (iii) would be useful. Similarly, to compare the imports during 2012 and 2014, bar chart (iv) would be useful.

(i)

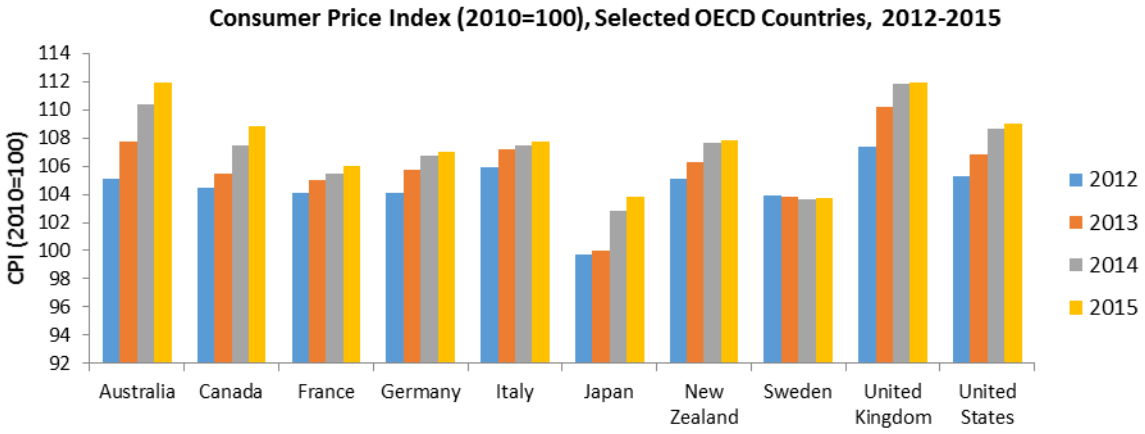
(ii)



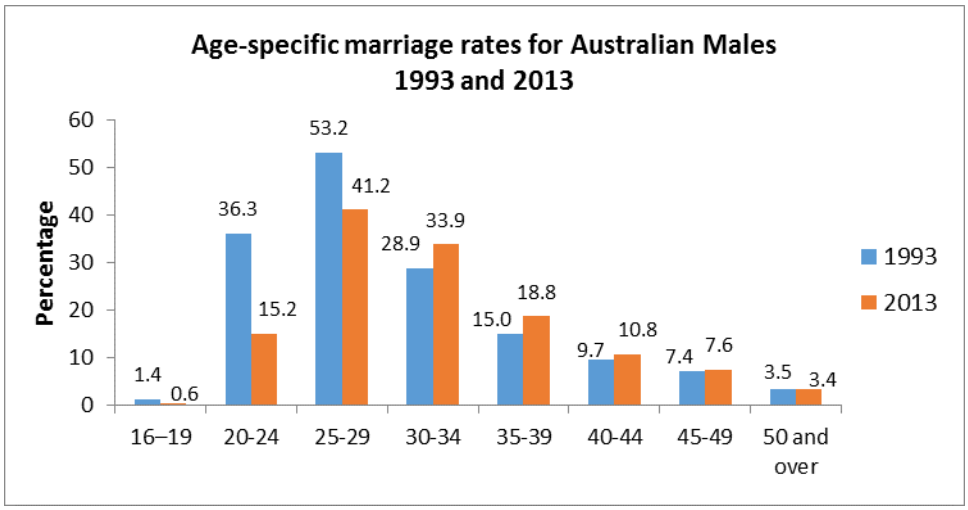
(iii)

(iv)

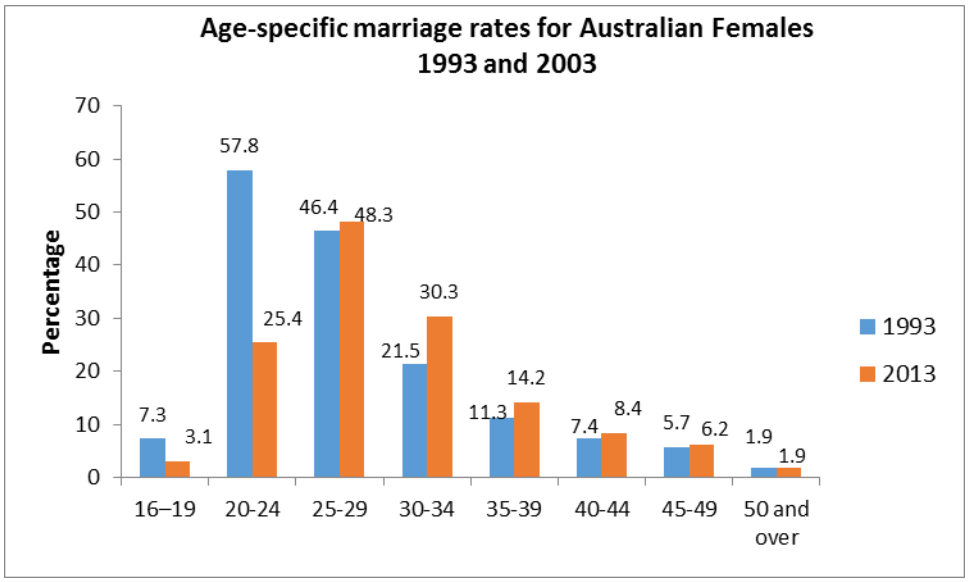
3.9 A bar chart would be appropriate.



3.10 a Appropriate graph to compare the marriage rates for males would be a bar chart using data for Males only.



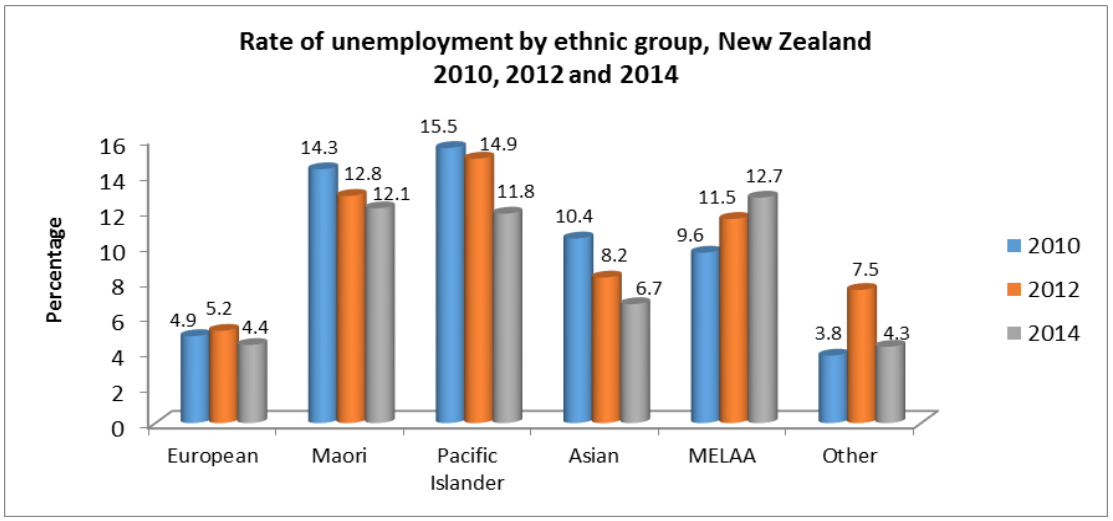
b Appropriate graph to compare the marriage rates for females would be a bar chart using data for Females only.

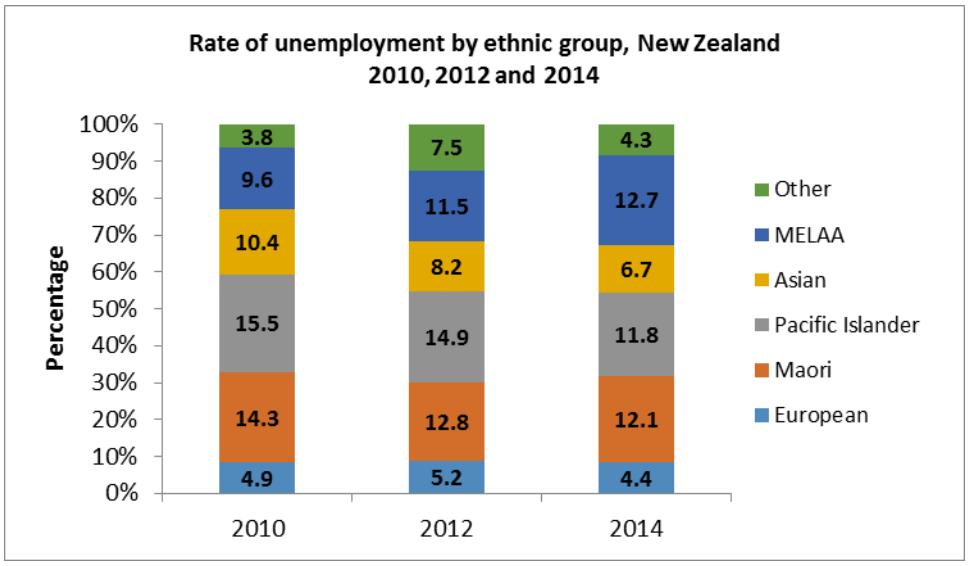


c Among the age groups for males, the marriage rate has fallen for the 16-19, 20-24, and 25-29 age groups, while it has increased for all the age groups 30 years and over between 1993 to 2013. For females, the marriage rate has fallen for the 16-19 and 20-24 age groups, while it has slightly increased for the 25-29 group and increased for all age groups 30 years and over between 1993 and 2003. This shows that the age at marriage has increased between 1993 and 2013 for both males and females. That is, more and more males and females are waiting longer to get married.

d A bar chart is more appropriate as the aim is to directly compare the marriage rates between the years 1993 and 2014, among males and females separately.

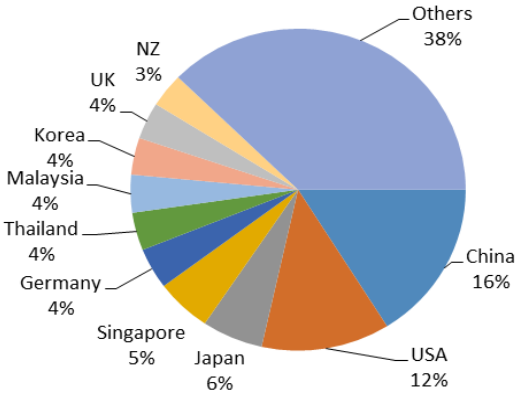
3.11 A bar chart would be appropriate to compare the rate of unemployment during the three years, 2010, 2012 and 2014 for each ethnic group. A component bar chart would be appropriate to compare the level of unemployment among the ethnic groups for each year.



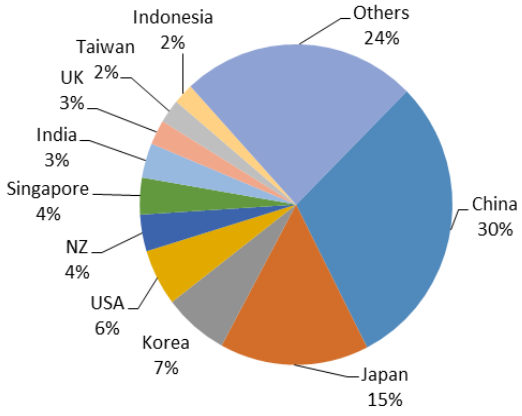


3.12 **a** Since the information is given in percentage shares, either a bar chart or pie chart would be appropriate. As the countries are different, a comparison of imports and exports is not possible.

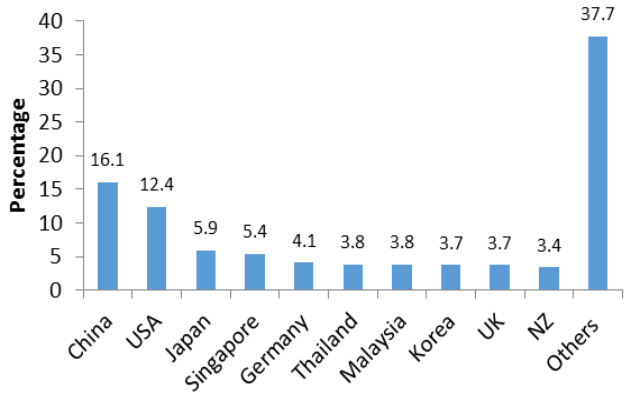
Share of Australian imports by top 10 countries, 2014



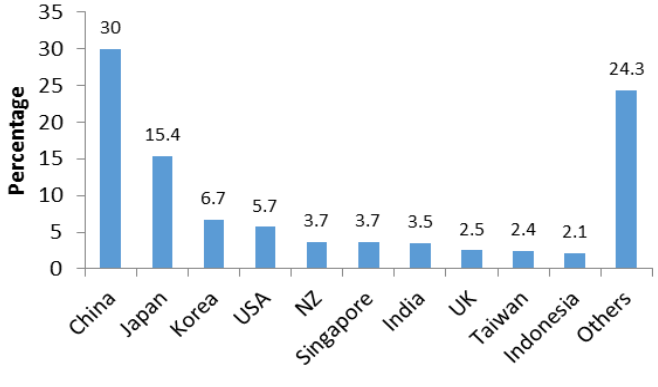
Share of Australian exports by top 10 countries, 2014



Share of Australian imports by top 10 countries, 2014



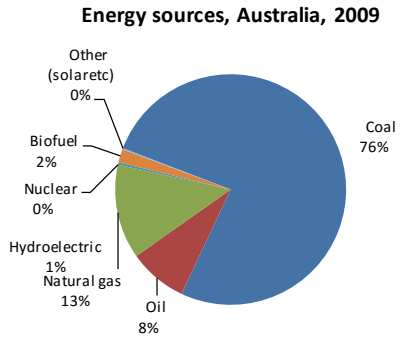
Share of Australian exports by top 10 countries, 2014

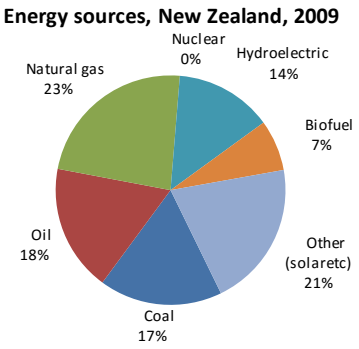


b The bar chart or a pie chart would be useful as the information is already in percentages (shares). However, a pie chart would be more appropriate to easily see the major contributors to Australian exports and imports.

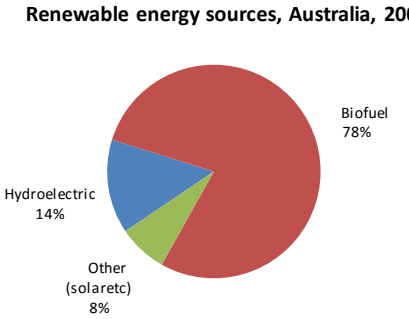
3.13 Pie charts (a) are helpful to compare the share of all forms of energy sources between and within Australia and New Zealand. Pie charts (b) for the non-renewable energy sources for Australia and New Zealand would be appropriate for comparison of the contribution of non-renewable energy sources within and between the two countries. Separate pie charts (c) for the renewable energy sources for Australia and New Zealand would be appropriate for comparison of the contribution of renewable energy sources within and between the two countries.

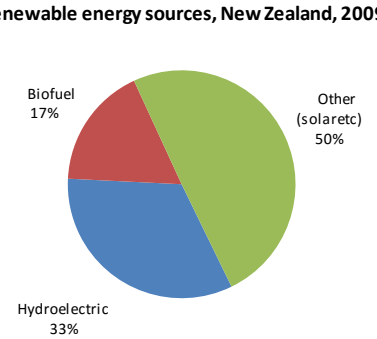
(a)



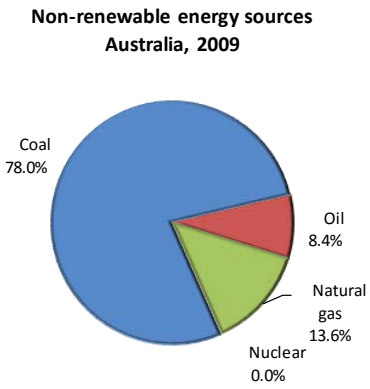


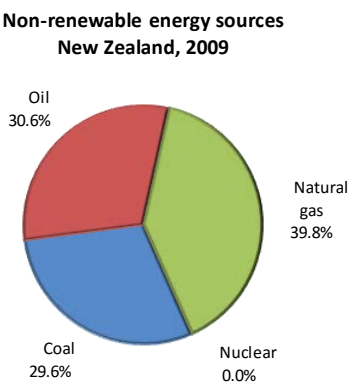
(b)



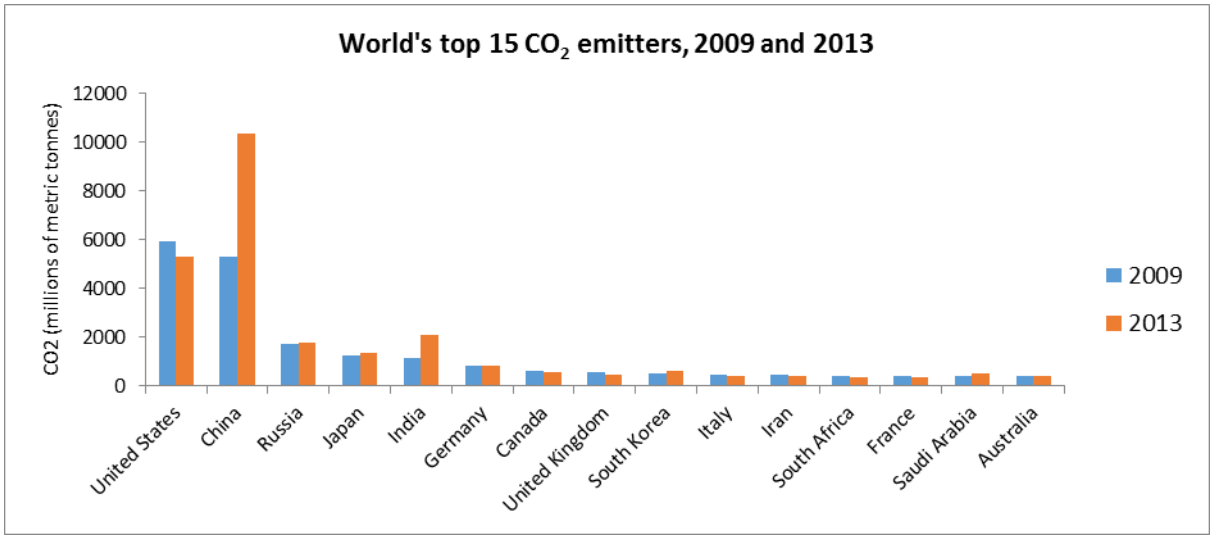


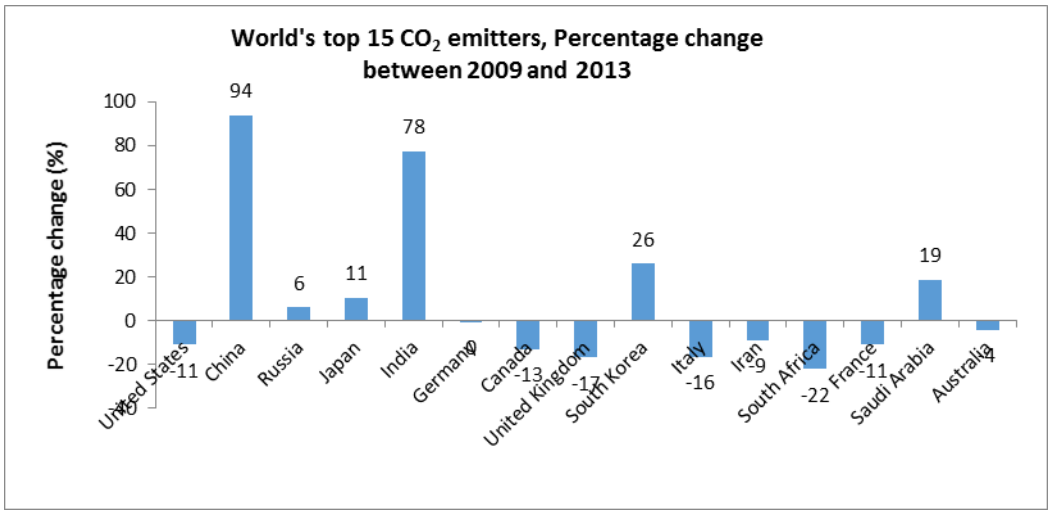
(c)





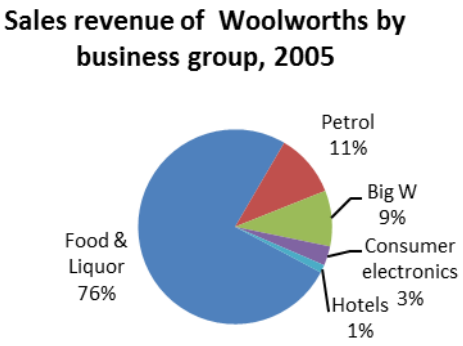
3.14 A bar chart would be appropriate. To see how countries are performing in terms of reducing CO_2 emissions, a bar chart of percentage change in emissions from 2009 to 2013 for the 15 countries would be appropriate.

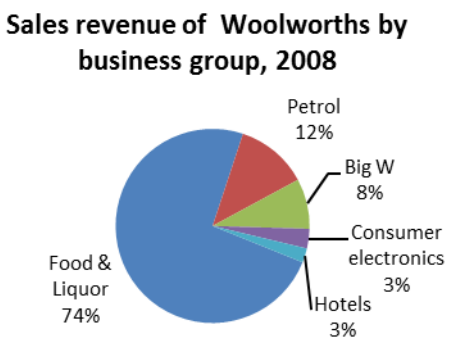


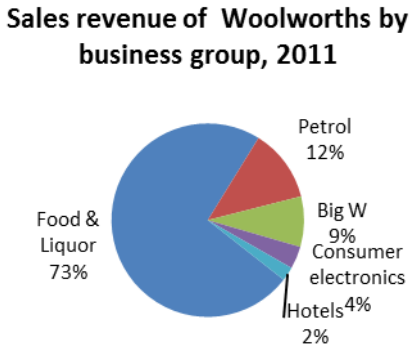


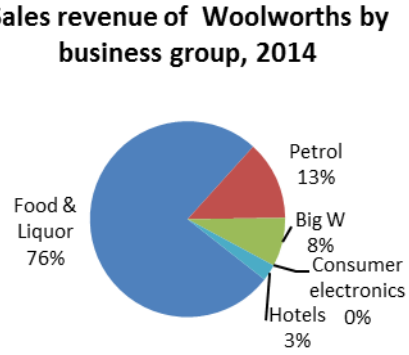
The top CO_2 emitters in the world in 2013 are China followed by the US, India, Russia and Japan in that order. Between 2009 and 2013, China, India, South Korea, Saudi Arabia and Russia have increased their CO2 emissions while US and all other countries including Australia have reduced their CO_2 emissions.

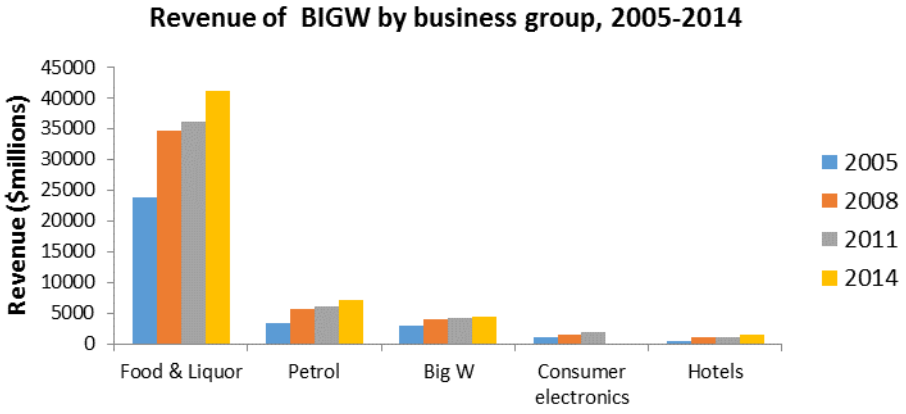
3.15 Pie charts, one for each year, are useful for comparison of the share contribution of revenues from various Woolworths business groups during the four years. A bar chart would be useful to compare the values of each business group during 2005, 2008, 20011 and 2014.



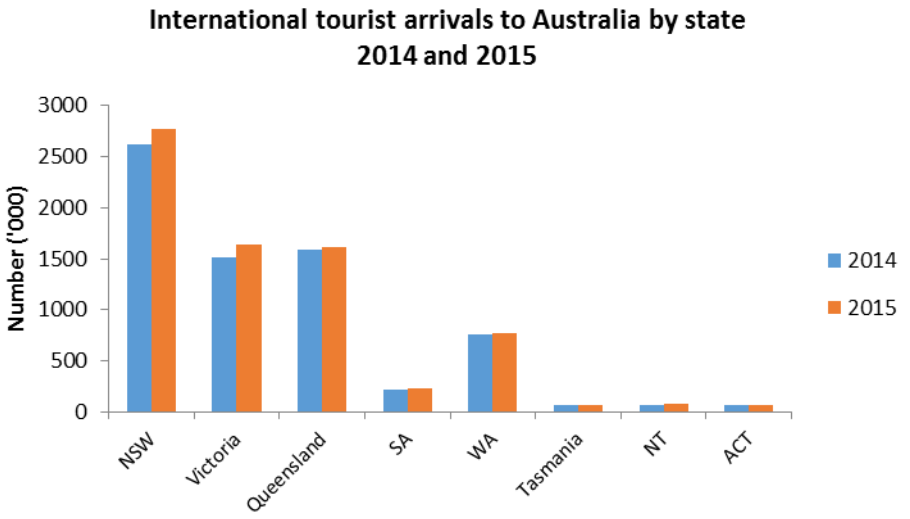




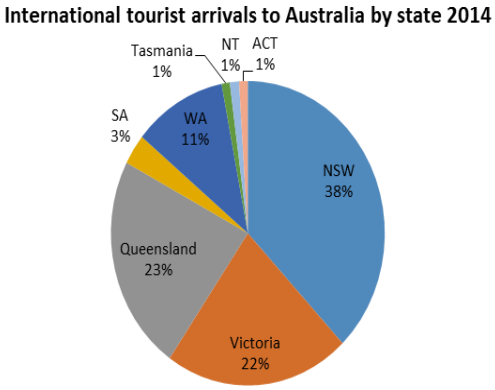


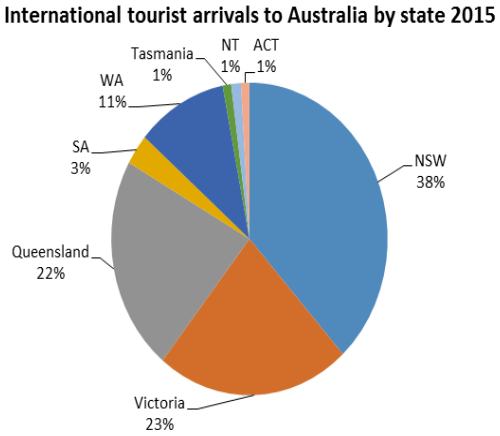


3.16 A bar chart would be appropriate to compare the change in the number of tourist arrivals from 2014 to 2015. The tourist arrivals to NSW and Victoria have increased in 2015 from 2014. Tourist arrivals to other states and territories have remained nearly the same. A pie chart would be useful to analyse the share of tourist arrivals to the different states and territories in Australia. The share of tourist arrivals to the different states and territories have stayed nearly the same in 2014 and 2015.

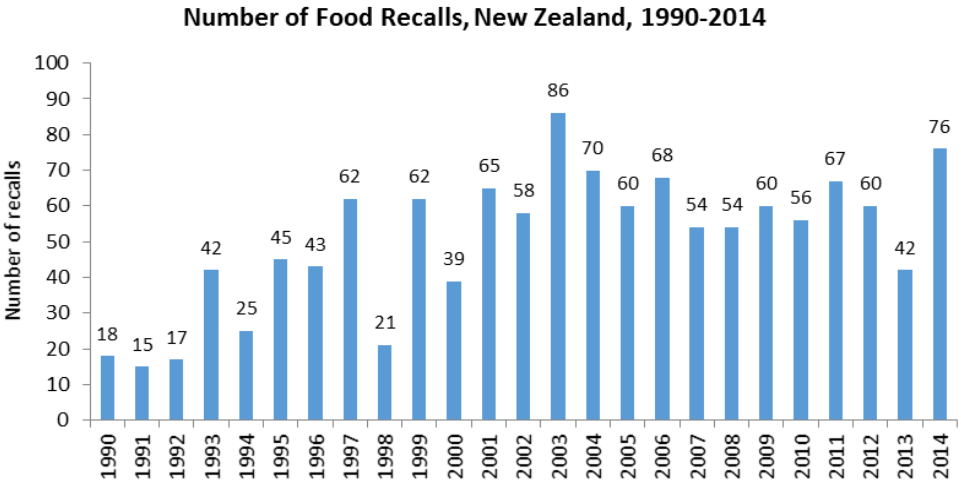


b



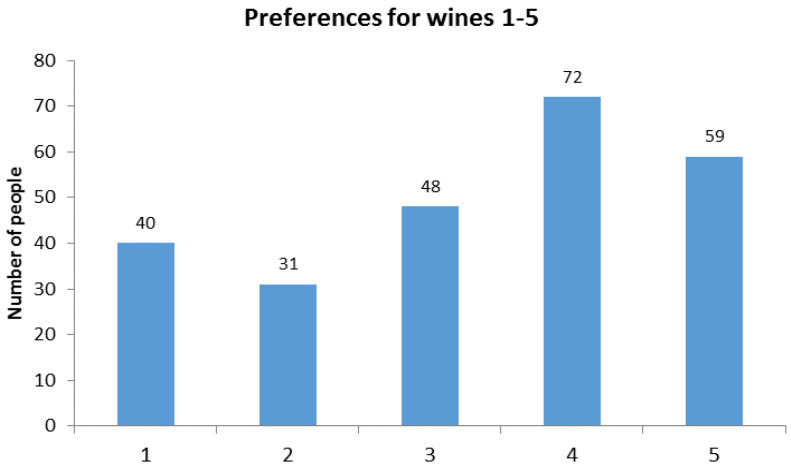


3.17

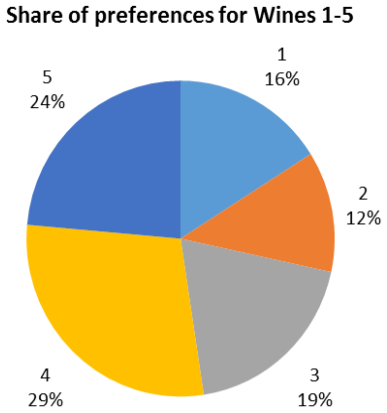


3.18

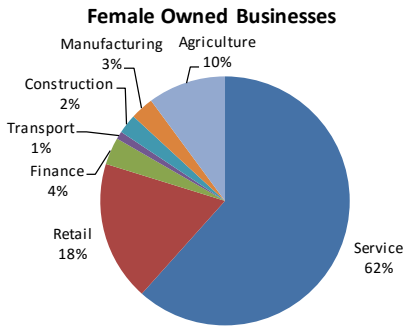
a If the focus is on the actual numbers who prefer each type of wine, then a bar chart would be useful.

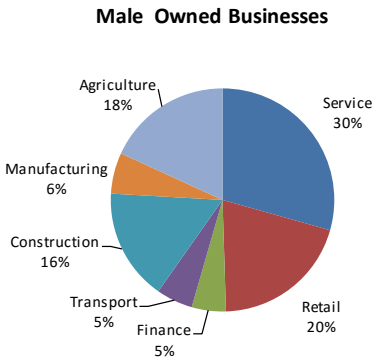


b Pie chart would be useful to show the share of the preferences for each type of wine among all types of wine.



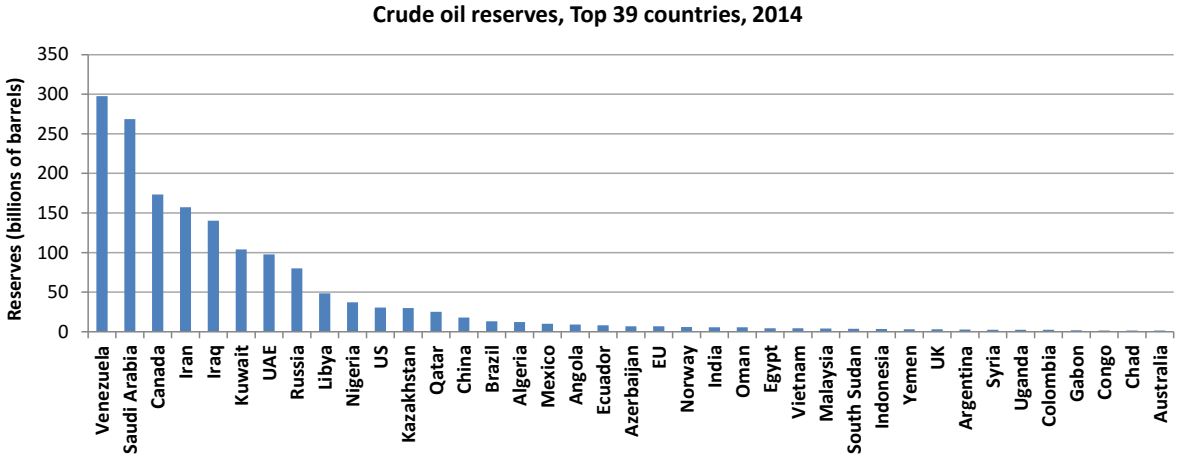
3.19 Pie charts of female-owned and male-owned businesses would provide the share of each business type for comparison.



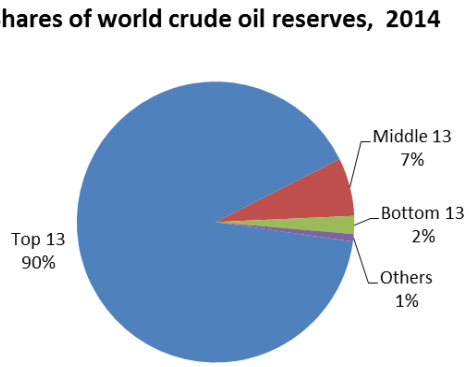


The biggest difference is that 62% of female-owned businesses are in the services sector, compared with 30% of male-owned businesses. Only 3% of female-owned businesses are in the construction sector, compared with 16% for male-owned businesses.

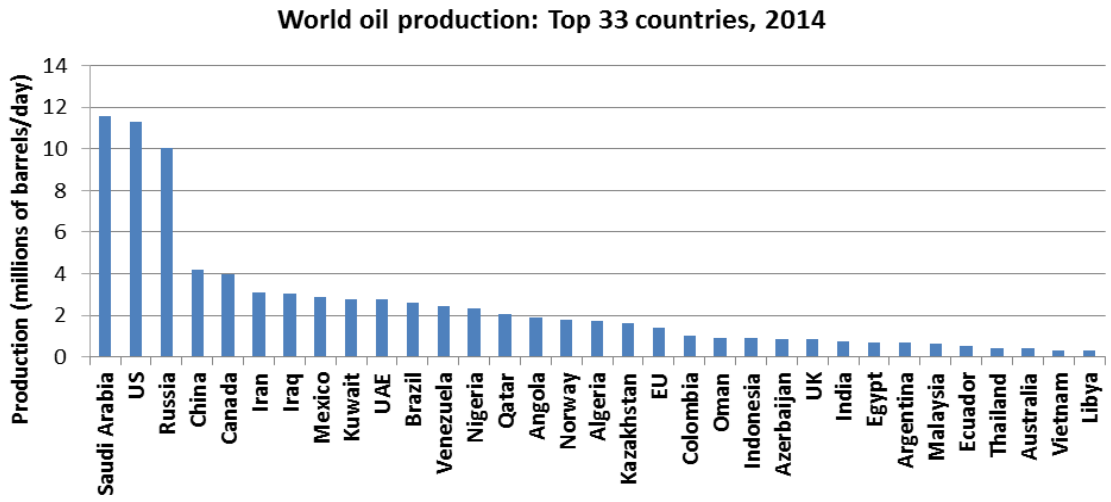
3.20



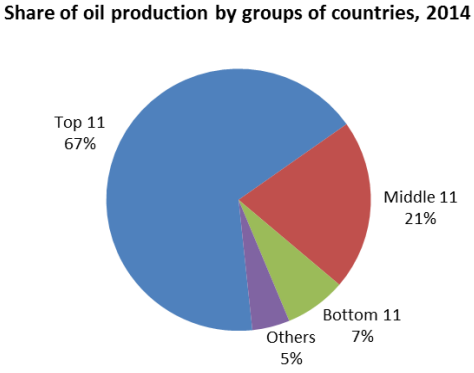
3.21 A pie chart for the four groups of top 13, middle 13, bottom 13, and other countries would emphasise the breakdown of oil reserves among the four groups of countries.



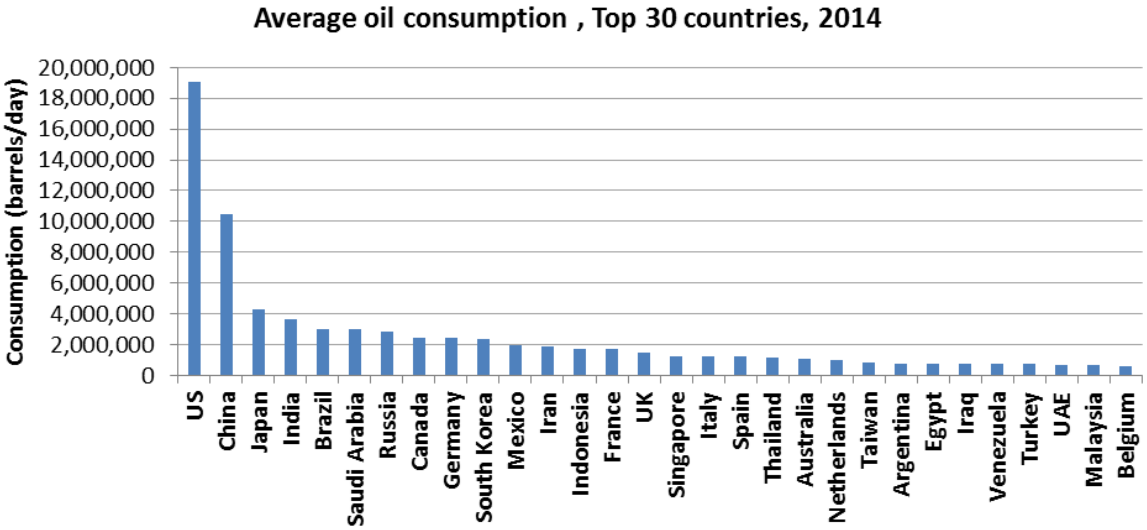
3.22



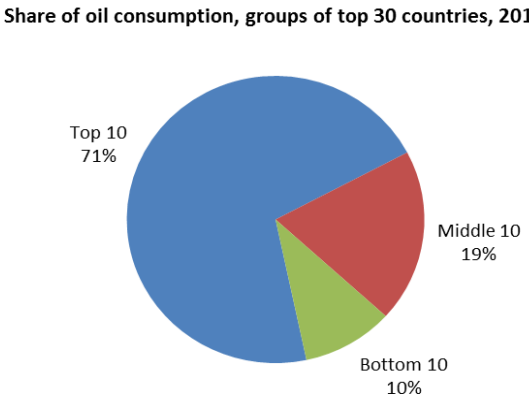
3.23 A pie chart for the four groups, top 11, middle 11, bottom 11, and other countries, would emphasise the breakdown of oil production among the four groups of countries.



3.24



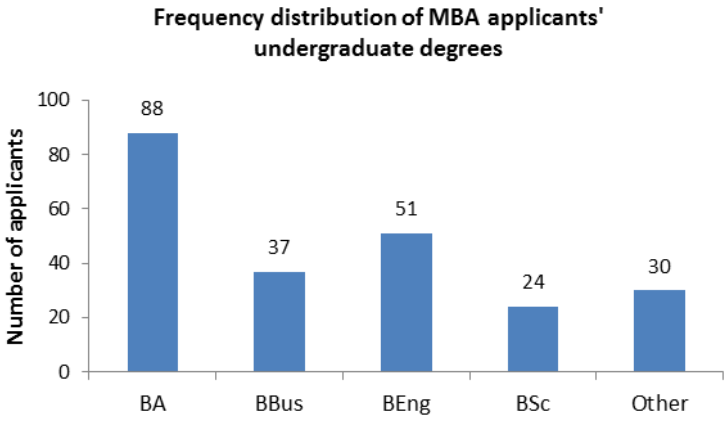
3.25 A pie chart for the three groups of top 10, middle 10 and bottom 10 countries would emphasise the breakdown of oil consumption across the 30 countries.



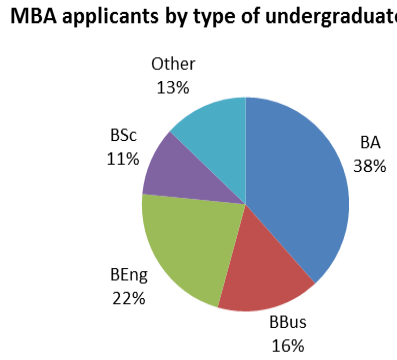
3.26 **a** Frequency distribution of undergraduate degrees of MBA applicants.

Degree	Frequency
BA	88
BBus	37
BEng	51
BSc	24
Other	30
Total	**230**

b

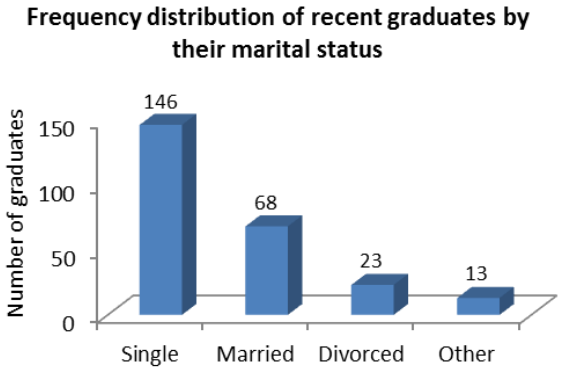


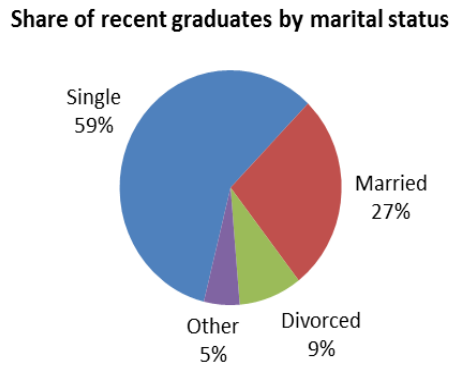
c



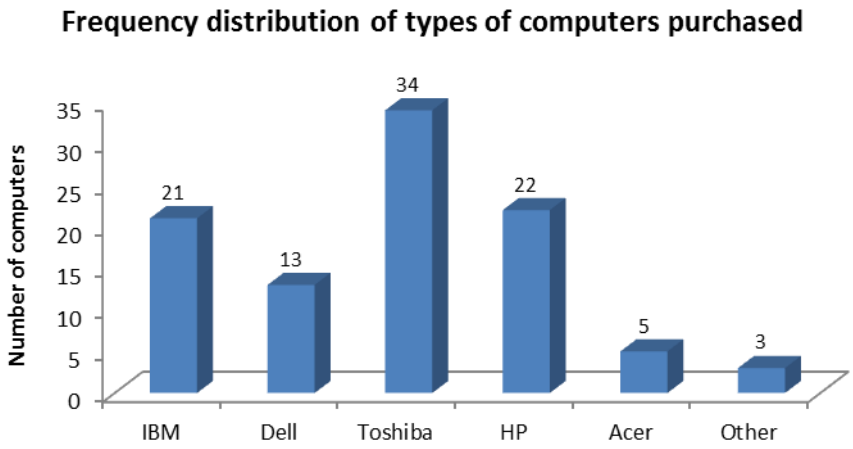
d. The charts in b and c show that a majority of applicants are BA graduates, capturing 88 (38%) of the applicants, followed by BEng 51 (22%), then BBus 37(16%) and BSc 24 (11%).

3.27 A bar or pie chart could be used. The graphs show that a majority of recent graduates are single (59%). About 27% of them are married and 9% are divorced.

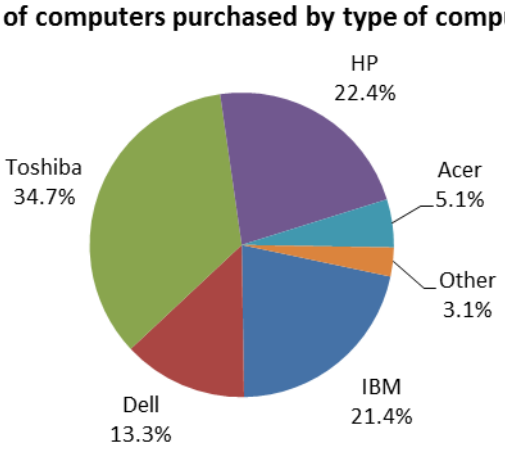




3.28 a A bar chart would be appropriate to depict the frequency distribution.

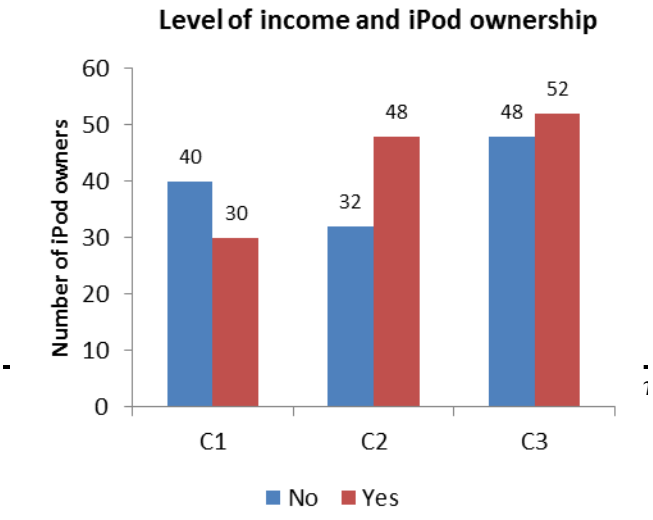


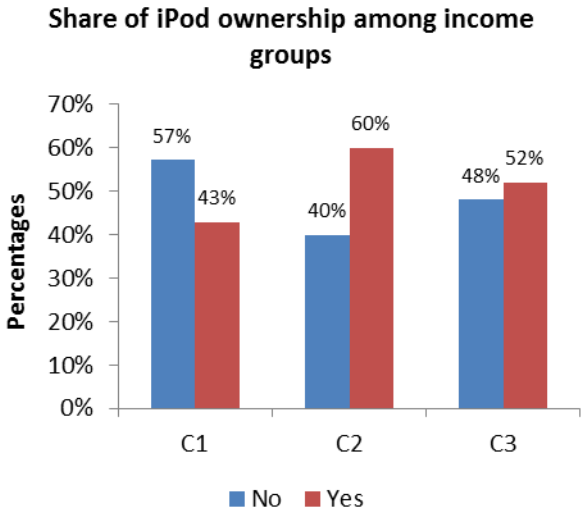
b A pie chart would be appropriate to depict the proportions.



c Based on the sample data provided, the charts show that Toshiba is the most popular brand followed by HP and IBM. Dell, Acer and other brands are the least popular brands among university students when they make a computer purchase.

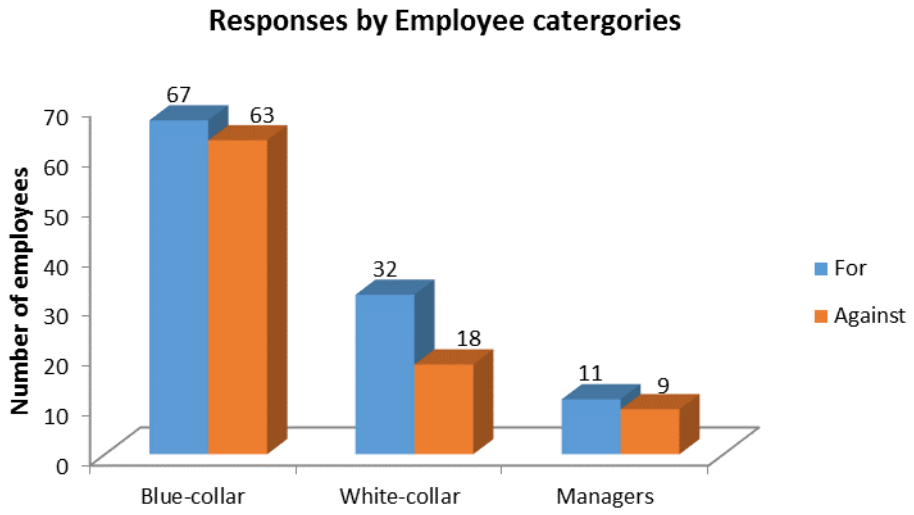
3.29 The frequency distributions are presented as bar charts in frequencies as well as proportions (in percentages).

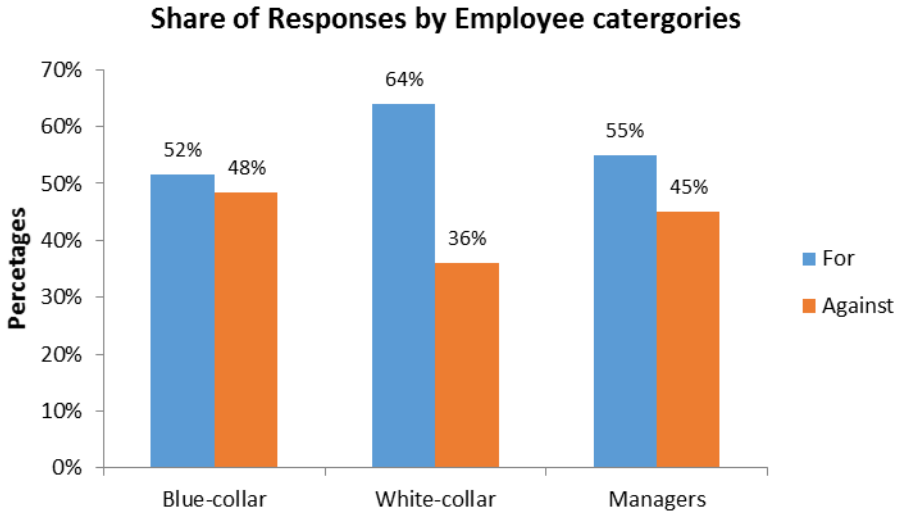




The proportion of iPod ownership is higher among the higher income groups (C2 and C3) compared to the lower income group (C1). If the two variables are unrelated, then the patterns exhibited in the bar charts should approximately be the same. If some relationship exists, then some of the bar charts will differ from other. Since the bar charts for C_1, C_2 and C_3 are all different, there exists some relationship between ownership of iPod and income level.

3.30 A pivot bar chart by type of worker would be more appropriate. Within each employee category, the proportion of 'For' response is greater than that of 'Against' response. However, the proportion of various types of workers 'For' the revision of the scheme is not similar. Similar result is shown for 'Against' as well. Therefore, the responses differ among the 3 groups.

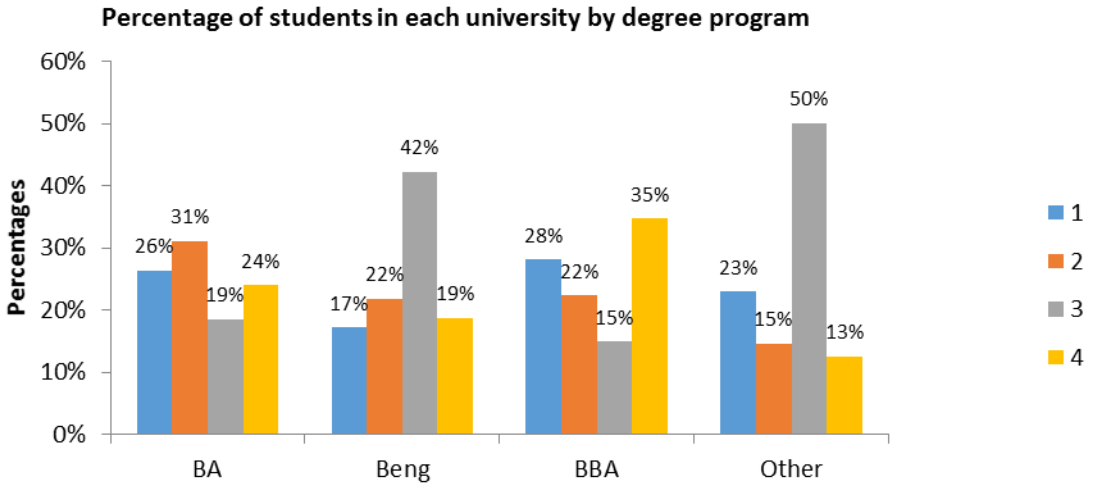




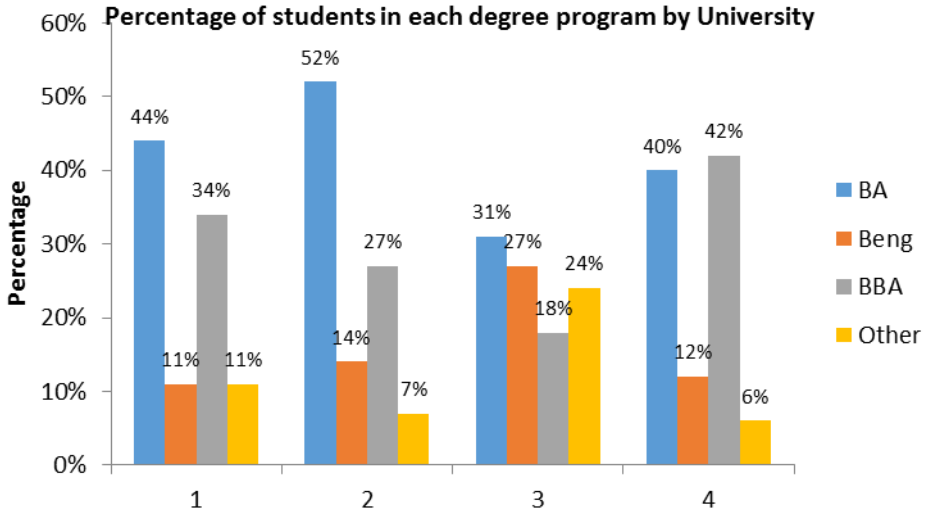
3.31 A pivot table and pivot chart would be appropriate to determine whether the undergraduate degree program and the university each person applied to are related. As can be seen, for each degree program, the proportion of students applying for a particular university is not similar. Similarly, among the universities, the distribution of students applying for a particular degree is not similar. Universities 1 and 2 are similar

and quite dissimilar from universities 3 and 4, which also differ. The two nominal variables appear to be related.

Count of Student	University				
Degree	1	2	3	4	Grand Total
BA	26%	31%	19%	24%	100%
Beng	17%	22%	42%	19%	100%
BBA	28%	22%	15%	35%	100%
Other	23%	15%	50%	13%	100%
Grand Total	25%	25%	25%	25%	100%



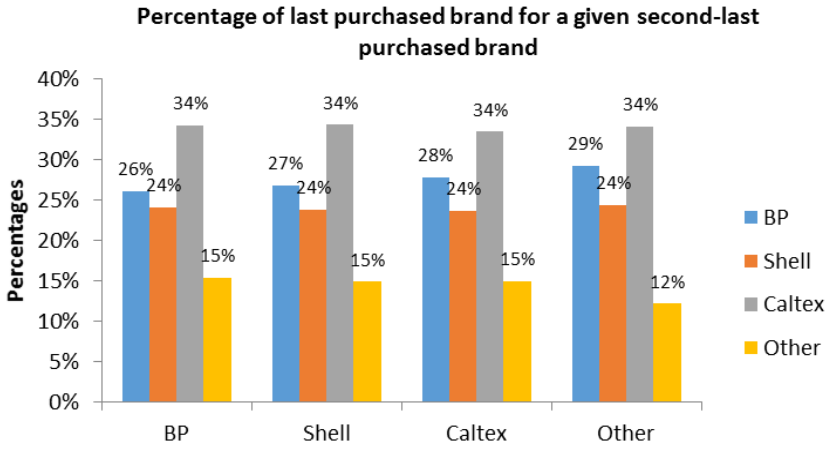
Count of Student	Degree				
University	BA	BEng	BBA	Other	Grand Total
1	44%	11%	34%	11%	100%
2	52%	14%	27%	7%	100%
3	31%	27%	18%	24%	100%
4	40%	12%	42%	6%	100%
Grand Total	42%	16%	30%	12%	100%



3.32 Constructing a pivot table (in percentages) and pivot chart would give the information required to compare and conclude whether there is brand loyalty among car owners.

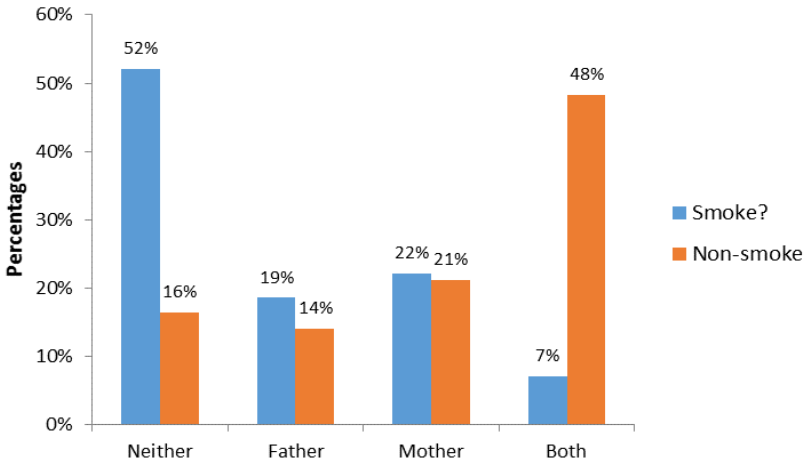
As can be seen, the column proportions are similar; the two nominal variables appear to be unrelated. Similar distributions can be seen from the pivot chart as well, confirming that there does not appear to be any brand loyalty.

Count of Owner	Last				
Second-last	BP	Shell	Caltex	Other	Grand Total
BP	26%	24%	34%	15%	100%
Shell	27%	24%	34%	15%	100%
Caltex	28%	24%	34%	15%	100%
Other	29%	24%	34%	12%	100%
Grand Total	27%	24%	34%	15%	100%



3.33

Count of ID	Smoke?		
Parent	Smoker	Non-smoker	Grand Total
Neither	52%	16%	39%
Father	19%	14%	17%
Mother	22%	21%	22%
Both	7%	48%	23%
Grand Total	100%	100%	100%

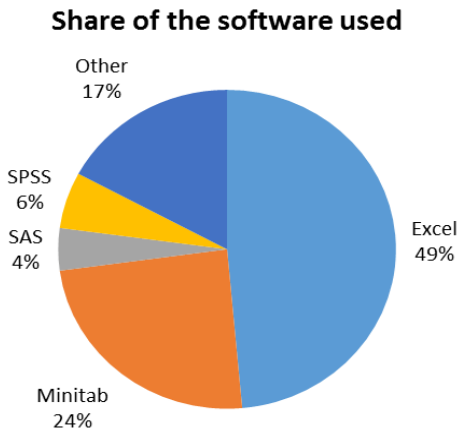


The two variables are related.

3.34 **a** Use Excel to count the number of courses using each software.

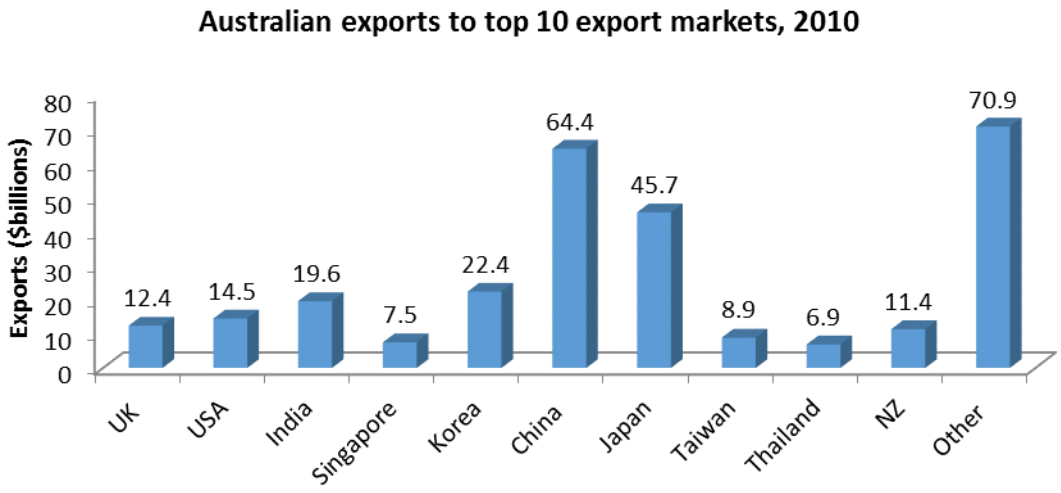
Software	Frequency
Excel	34
Minitab	17
SAS	3
SPSS	4
Other	12

b A pie chart would be appropriate to depict the proportions.

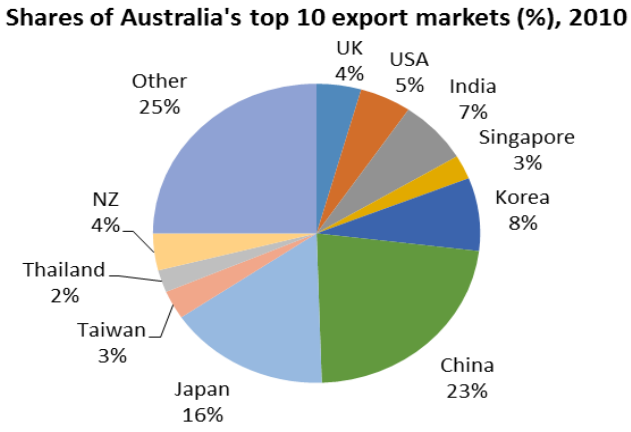


c Excel is the choice of about half the sample, one-quarter have opted for Minitab, and a small fraction chose SAS and SPSS.

3.35 a

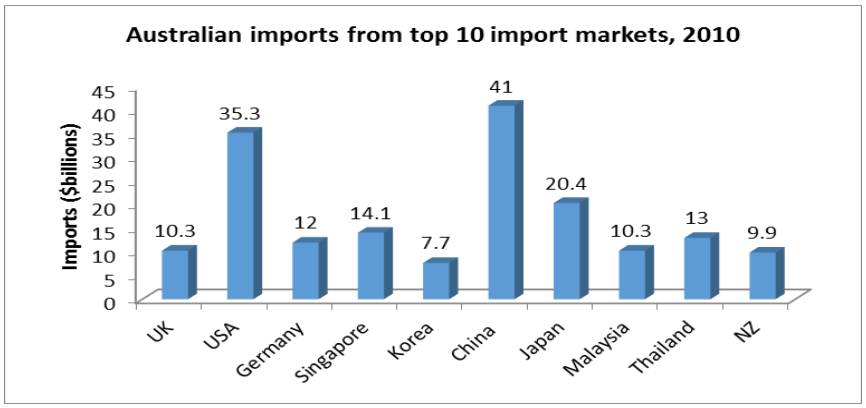


b

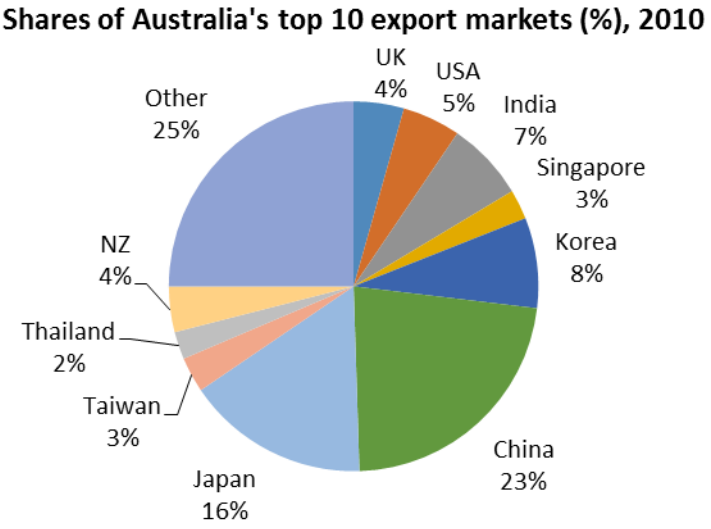


c As can be seen, among the top 10 export countries, China is the largest export market (23%), followed by Japan (16%) and Korea (8%).

3.36 a Depict the amount of imports in a bar chart.

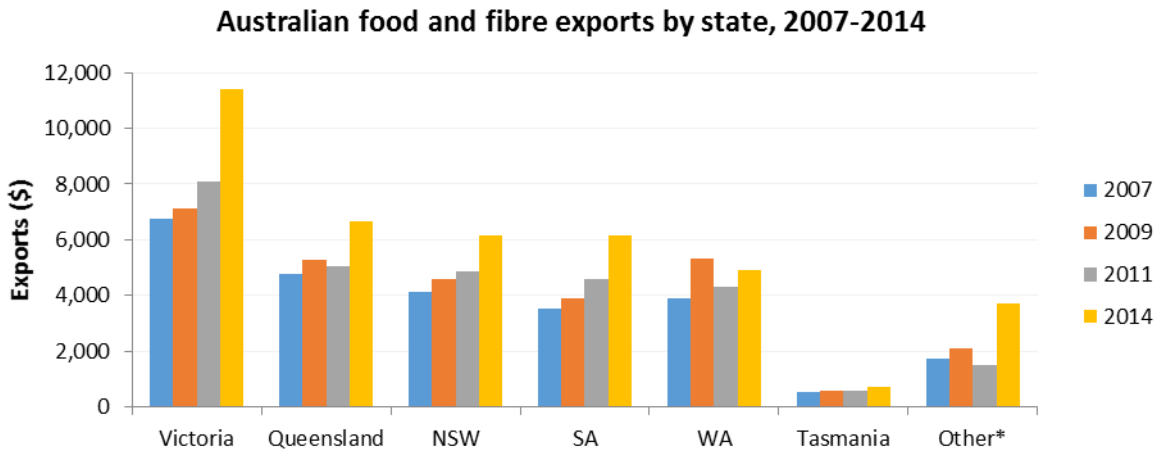


b Depict the share of imports in a pie chart.

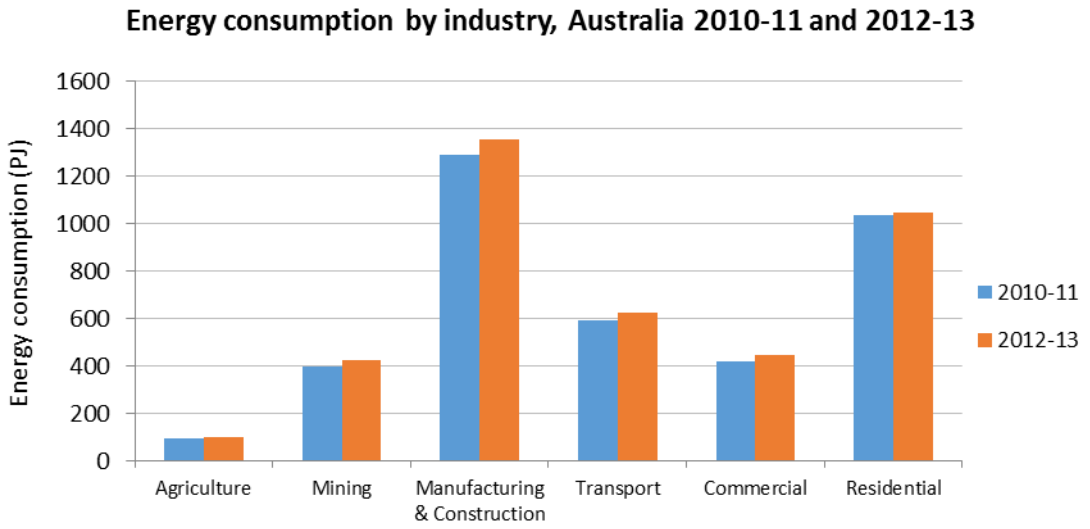


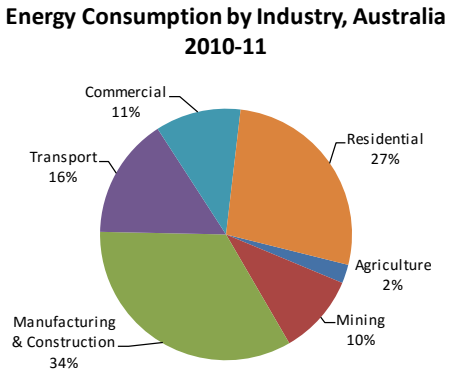
c As can be seen, among the top 10 import markets, China is the largest (15.3%) followed by the US (13.2%) and Japan (7.6%).

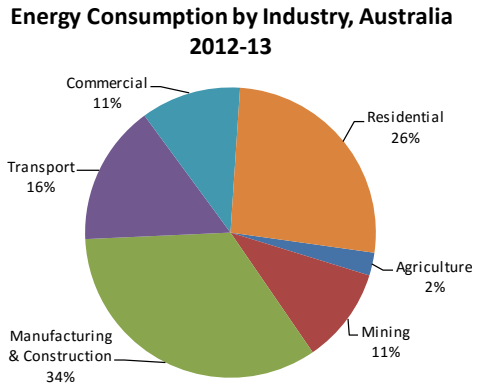
3.37 A combined bar chart by state would be more appropriate. Australian food and fibre exports have generally been on the increase in all states between 2007 and 2014.



3.38 A combined bar chart by industry would be appropriate. Pie charts for the two years would also provide information regarding how the shares of each industry have changed over the two-year period.



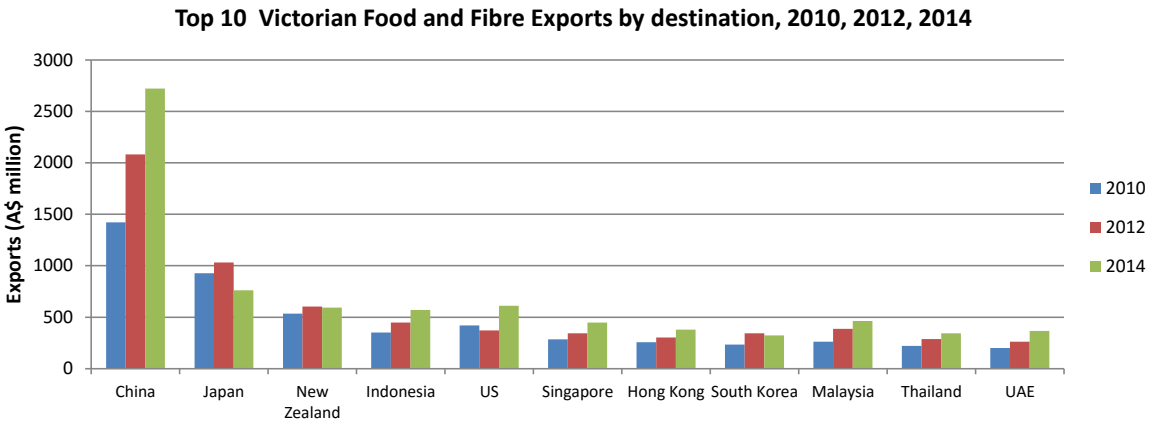


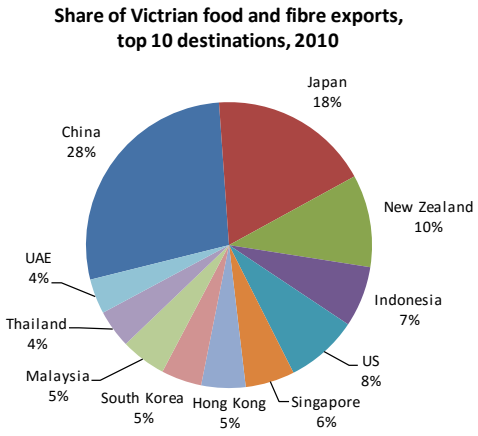


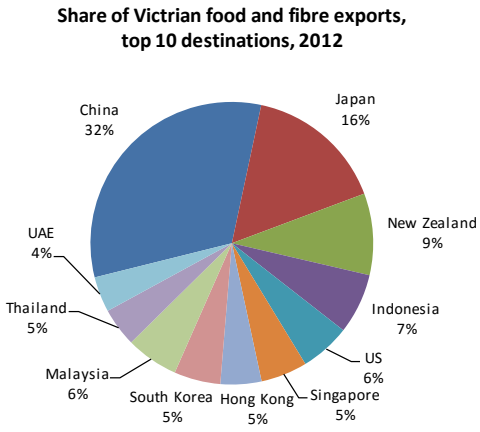
Energy consumption is highest in the manufacturing and construction industry followed by residential, transport, commercial and mining. The agricultural industry

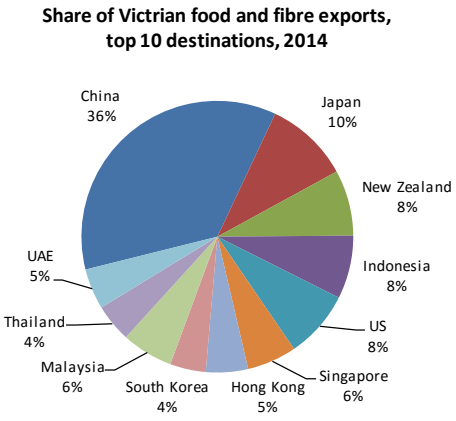
consumes the least. There has been a slight increase in energy consumption in all the industries between 2010-11 and 2012-13. The share of energy consumption in the various industries have not changed between the two year period.

3.39 A combined bar chart by destination would be appropriate. Pie charts for the three years would also provide information regarding how the share of each destination has changed from 2010 to 2014.

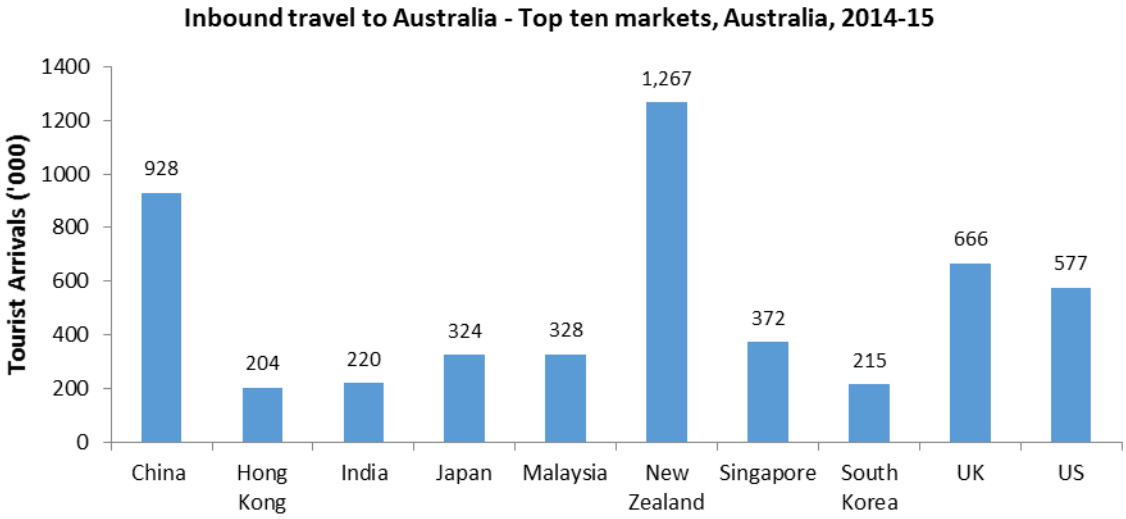






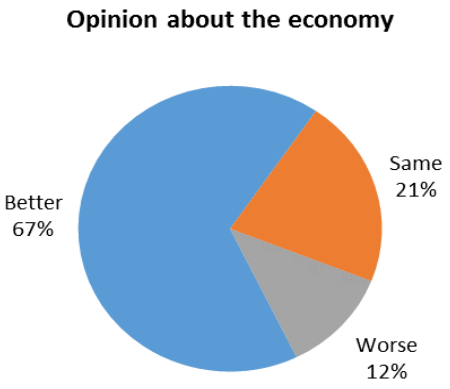


3.40 a A bar chart would be appropriate to compare the arrivals from the top 10 markets.



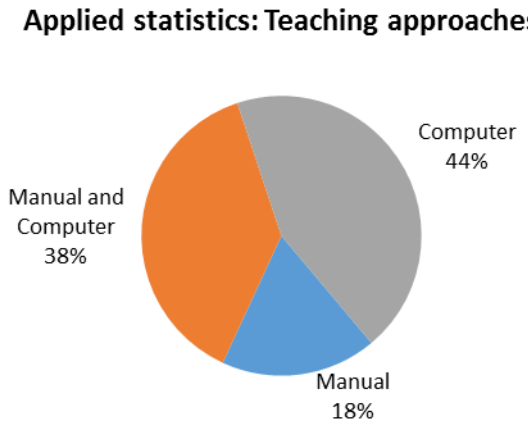
b The graph allows us to easily gauge visually the level of differences in the tourist arrivals from the top 10 countries.

3.41 As the data are nominal survey data, we obtain the frequency for each category and consider the proportions that fall in each category, which can be depicted using a pie chart.



As can be seen, more than two-thirds of those surveyed said that the economy will be better next year, and only 12% said that the economy will be worse.

3.42 As the data are nominal survey data, we obtain the frequency for each category and consider the proportions that fall in each category. A pie chart would be more appropriate.



4 Graphical descriptive techniques – Numerical data

Throughout this chapter,

- **In all frequency distributions, histograms and ogive, the class intervals contain observations greater than their lower limits (except for the first class) and less than or equal to their upper limits.**
- **In all histograms and ogive output using Excel, the upper limits of the class intervals are printed in the centre of the classes.**

4.1 Using Sturge's formula for $n = 250$, the approximate number of classes for a histogram would be

$K = 1 + 3 \log_{10} 250 \approx 8$ to 9 classes

4.2 a Using Sturge's formula for $n = 125$, the approximate number of classes to have in a histogram would be $K = 1 + 3 \log_{10} 125 \approx 7$ to 8 classes

b Approximate class width using $K = 8$ classes can be calculated as

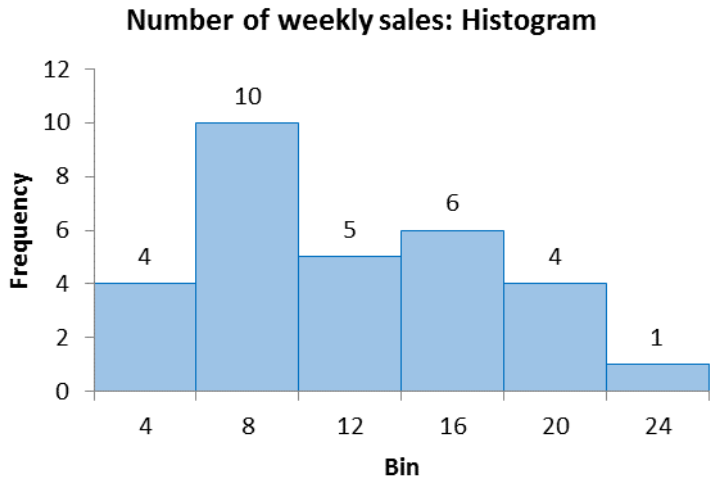
$$d = \frac{(largest - smallest)}{K} = \frac{(188 - 37)}{8} \approx 20$$

Therefore, the class intervals would be

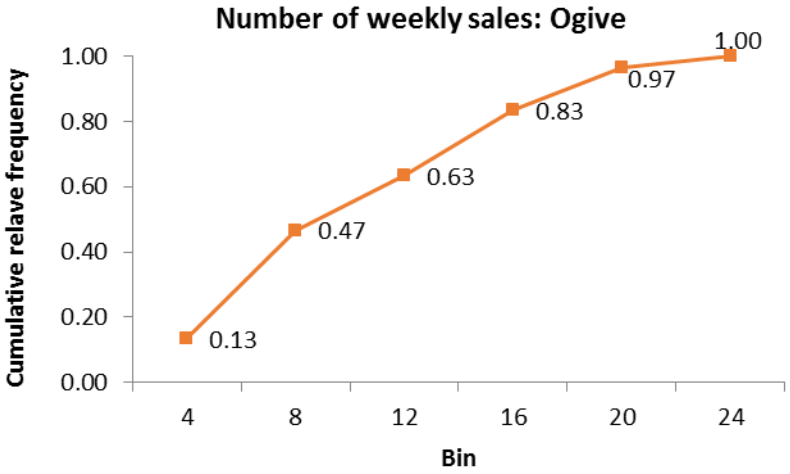
$35 \le X \le 55$, $55 < X \le 75$, $75 < X \le 95$, $95 < X \le 115$,
$115 < X \le 135$, $135 < X \le 155$, $155 < X \le 175$, $175 < X \le 195$

4.3 a

Class interval	Tally	Frequency	Cumulative Frequency	Relative cumulative frequency
0 up to 4	||||	4	4	0.13
4 up to 8	𝍸 𝍸	10	14	0.47
8 up to 12	𝍸	5	19	0.63
12 up to 16	𝍸 |	6	25	0.83
16 up to 20	||||	4	29	0.97
20 up to 24	|	1	30	1.00



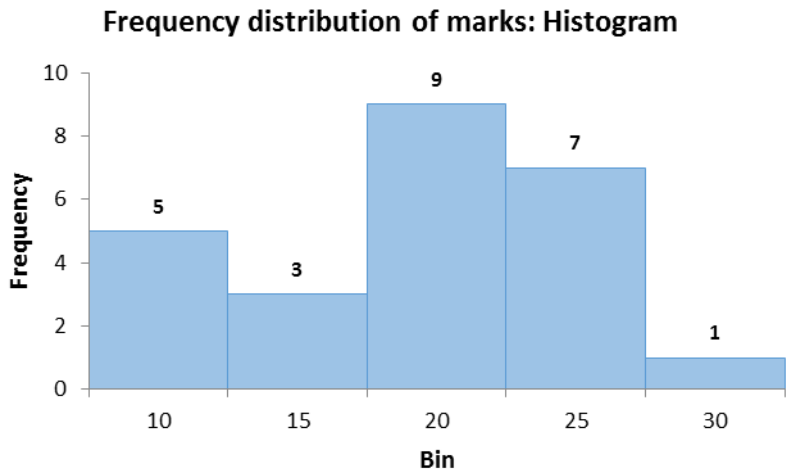
b



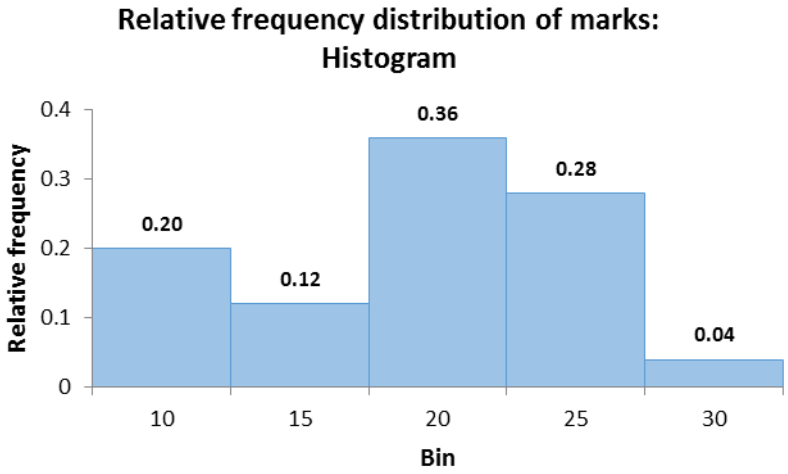
c The distribution is slightly skewed to the right.

4.4 a

Class interval	Bin	Tally	Frequency (f)	Relative frequency
$5 \leq X \leq 10$	10	𝍸	5	0.20
$10 < X \leq 15$	15	|||	3	0.12
$15 < X \leq 20$	20	𝍸 ||||	9	0.36
$20 < X \leq 25$	25	𝍸 ||	7	0.28
$25 < X \leq 30$	30	|	1	0.04



b



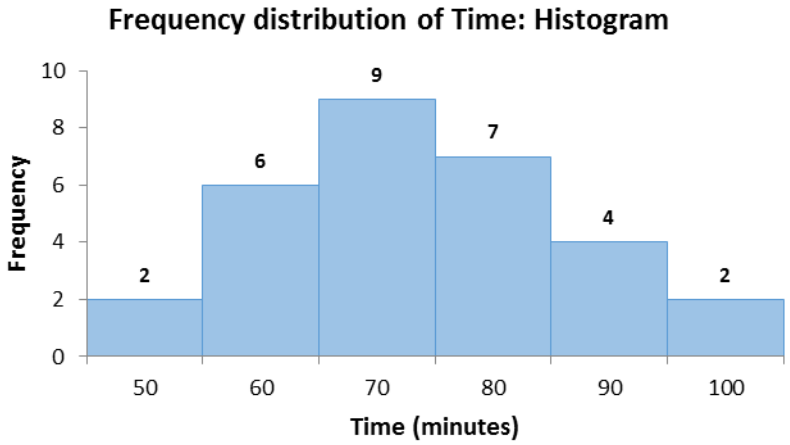
c The area of each rectangular strip is proportional to the relative frequency of that class. As the class widths are all the same the height is represented by the relative frequency of that class.

4.5 a

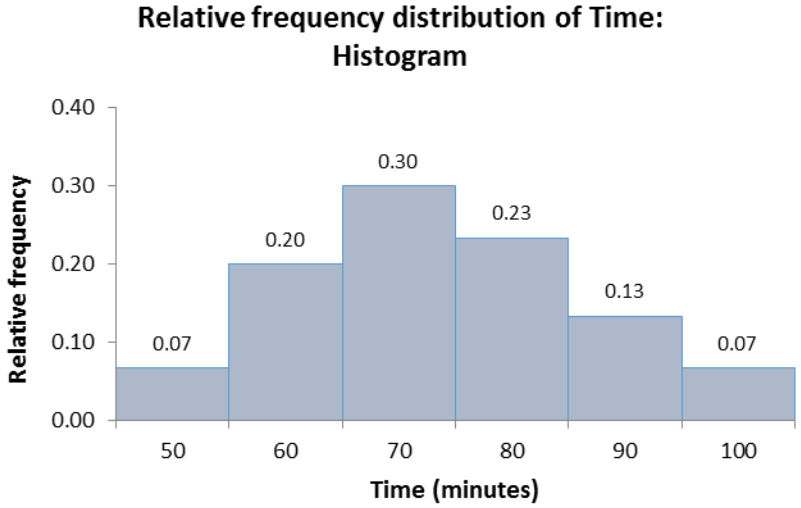
Stem	Leaf
4	58
5	245889
6	11245667
7	0357789
8	02366
9	14

b

Class interval	Bin	Tally	Frequency	Relative Frequency
$40 \le X \le 50$	50	||	2	0.07
$50 < X \le 60$	60	𝍸 |	6	0.20
$60 < X \le 70$	70	𝍸 ||||	9	0.30
$70 < X \le 80$	80	𝍸 ||	7	0.23
$80 < X \le 90$	90	||||	4	0.13
$90 < X \le 100$	100	||	2	0.07
Total			30	1.00



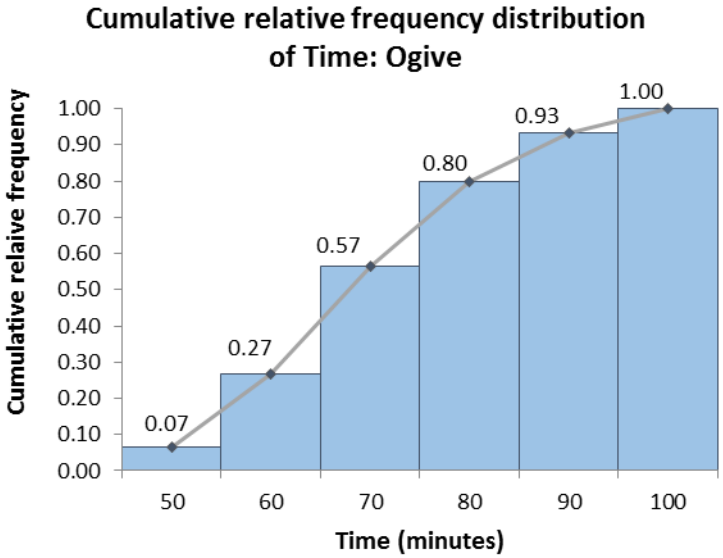
c



d The stem and leaf display and the histograms show a near symmetrical distribution of the time taken to complete the quiz.

e

Class interval	Frequency	Relative frequency	Cumulative relative frequency	Cumulative relative frequency (in %)
40 up to 50	2	0.07	0.07	6.67%
50 up to 60	6	0.20	0.27	26.67%
60 up to 70	9	0.30	0.57	56.67%
70 up to 80	7	0.23	0.80	80.00%
80 up to 90	4	0.13	0.93	93.33%
90 up to 100	2	0.07	1.00	100.00%
Total	30	1.00		



f **(i)** Proportion of the students took less than 70 minutes to complete the quiz = 0.567

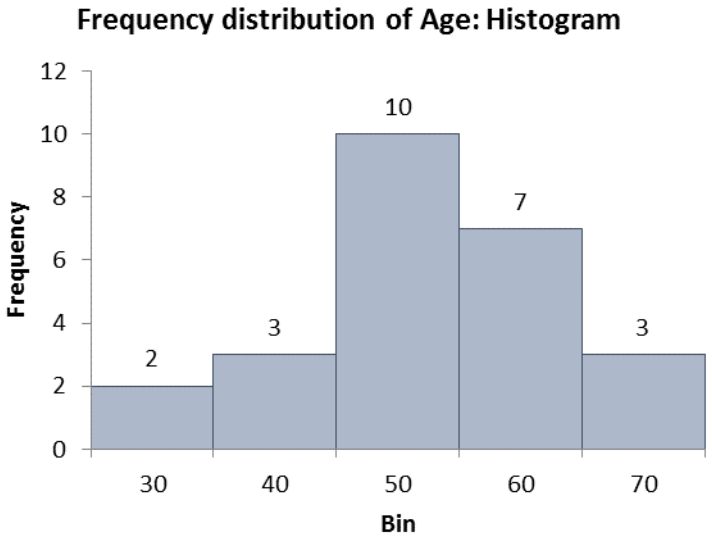
(ii) Proportion of the students took greater than 70 minutes to complete the quiz = 0.433

4.6 a We have chosen to use the ten's digit as the stem and one's digit as the leaf. Although we have recorded the leaves in ascending order, it is not necessary to do so.

Stem	Leaf
2	48
3	268
4	124456779
5	01245789
6	146

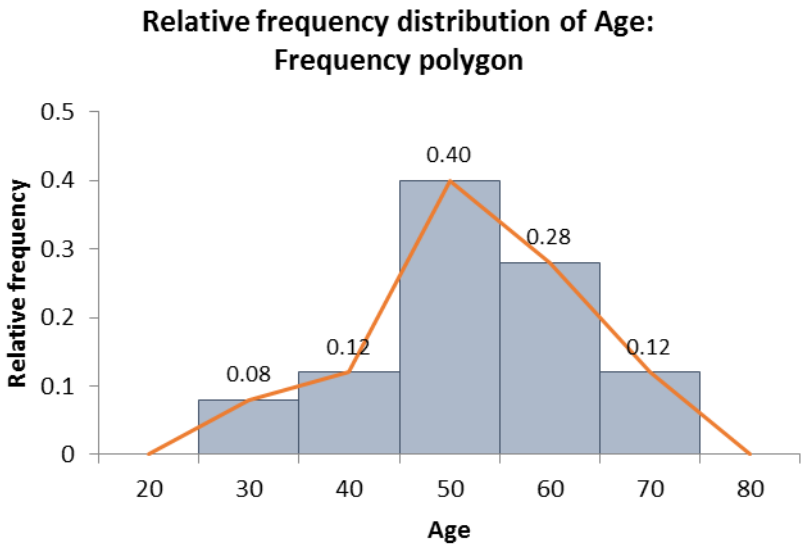
b

Class interval	Tally	Frequency
20 up to 30	||	2
30 up to 40	|||	3
40 up to 50	𝍸 𝍸	10
50 up to 60	𝍸 ||	7
60 up to 70	|||	3
Total		25

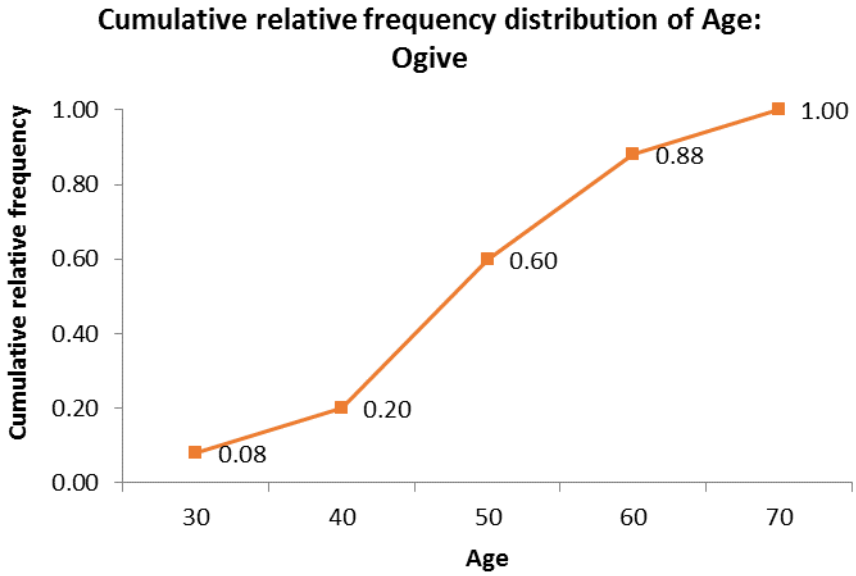


c, d

Class interval	Relative frequency	Cumulative relative frequency
20 up to 30	$\frac{2}{25} = 0.08$	$\frac{2}{25} = 0.08$
30 up to 40	$\frac{3}{25} = 0.12$	$\frac{5}{25} = 0.20$
40 up to 50	$\frac{10}{25} = 0.40$	$\frac{15}{25} = 0.60$
50 up to 60	$\frac{7}{25} = 0.28$	$\frac{22}{25} = 0.88$
60 up to 70	$\frac{3}{25} = 0.12$	$\frac{25}{25} = 1.00$



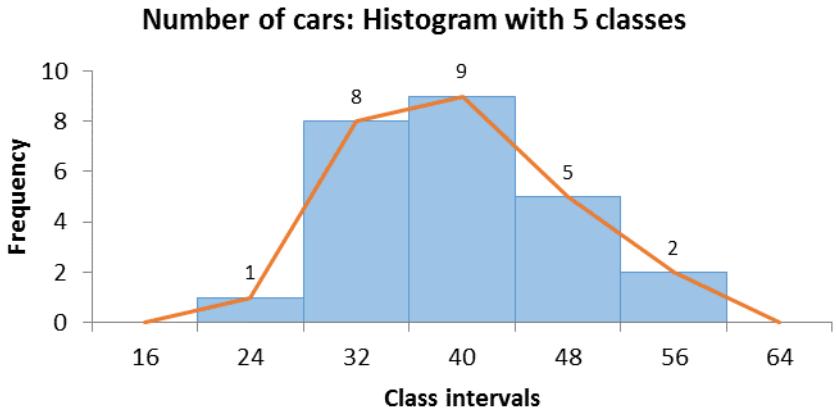
e



f Referring to the relative frequency histogram in part c, the proportion of the total area under the histogram that falls between 20 and 40 is 0.20 (= 0.08 + 0.12).

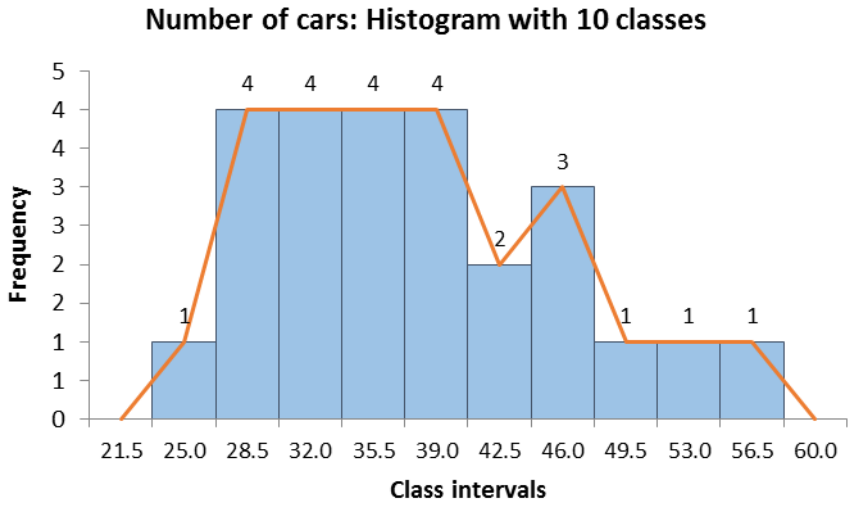
4.7 a & d

Class interval	Bin	Tally	Frequency
$16 \le X \le 24$	24	|	1
$24 < X \le 32$	32	𝍸 |||	8
$32 < X \le 40$	40	𝍸 ||||	9
$40 < X \le 48$	48	𝍸	5
$48 < X \le 56$	56	||	2



b & d

Classes	Bin	Tally	Frequency
$21.5 \le X \le 25.0$	25.0	|	1
$25.0 < X \le 28.5$	28.5	||||	4
$28.5 < X \le 32.0$	32.0	||||	4
$32.0 < X \le 35.5$	35.5	||||	4
$35.5 < X \le 39.0$	39.0	𝍸	5
$39.0 \le X \le 42.5$	42.5	|||	3
$42.5 < X \le 46.0$	46.0	||	2
$46.0 < X \le 49.5$	49.5	|	1
$49.5 < X \le 53.0$	53.0	|	1
$53.0 < X \le 56.5$	56.5		0

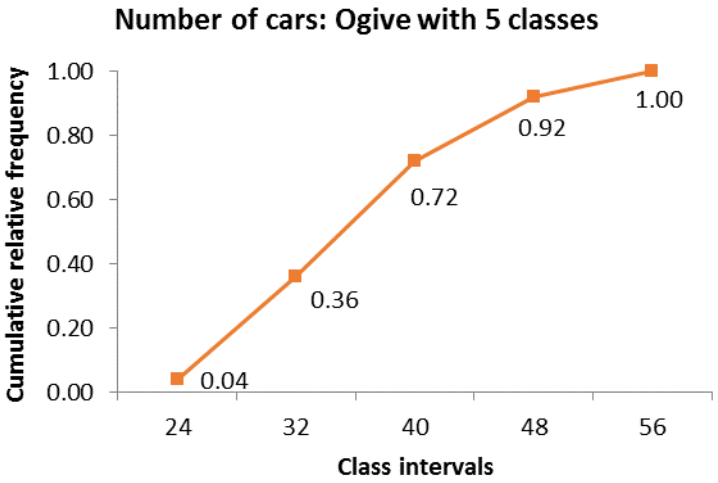


c

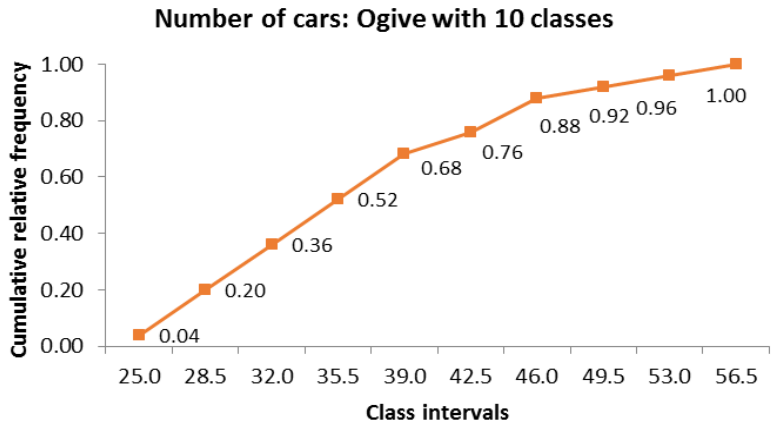
Stem	Leaf
2	4778899
3	1133457788
4	013468
5	06

e

Class interval	Frequency	Cumulative frequency	Relative cumulative frequency	Relative cumulative frequency (in %)
$16 \leq X \leq 24$	1	1	0.04	0.04
$24 < X \leq 32$	8	9	0.36	0.36
$32 < X \leq 40$	9	18	0.72	0.72
$40 < X \leq 48$	5	23	0.92	0.92
$48 < X \leq 56$	2	25	1.00	1.00
Total	25			



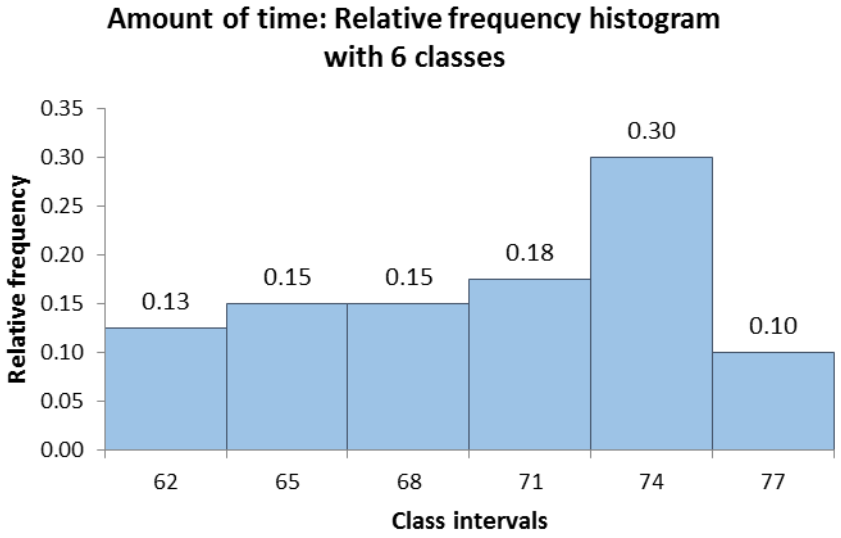
Bin	Frequency	Cumulative frequency	Relative Cumulative frequency
25.0	1	1	0.04
28.5	4	5	0.20
32.0	4	9	0.36
35.5	4	13	0.52
39.0	4	17	0.68
42.5	2	19	0.76
46.0	3	22	0.88
49.5	1	23	0.92
53.0	1	24	0.96
56.5	1	25	1.00
Total	25		



4.8 a Frequency distribution with 6 classes

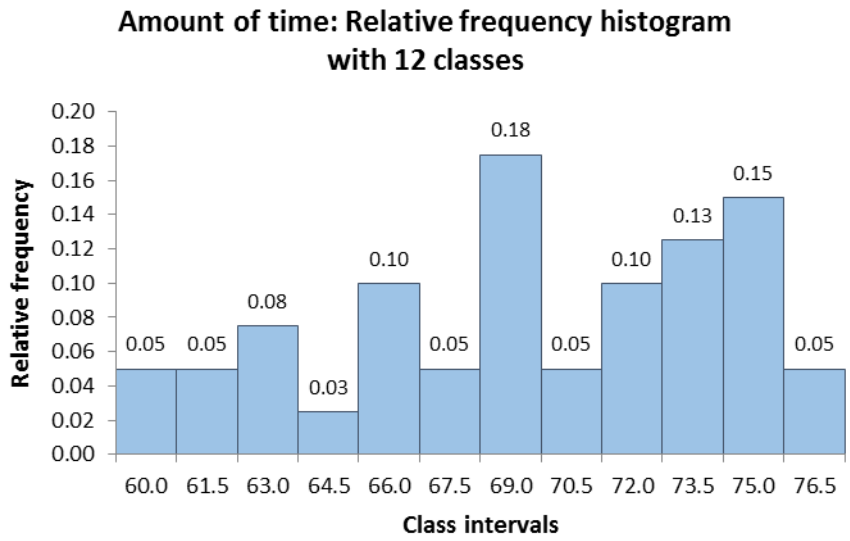
Class	Tally	Frequency	Relative	Cumulative	Cumulative

			frequency	frequency	relative frequency
59 up to 62	𝍸	5	0.125	5	0.125
62 up to 65	𝍸 |	6	0.150	11	0.275
65 up to 68	𝍸 |	6	0.150	17	0.425
68 up to 71	𝍸 ||	7	0.175	24	0.600
71 up to 74	𝍸 𝍸 ||	12	0.300	36	0.900
74 up to 77	||||	4	0.100	40	1.000
Total		40			

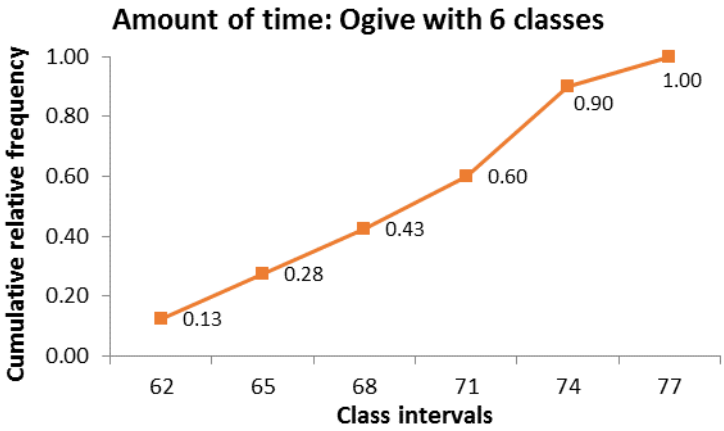


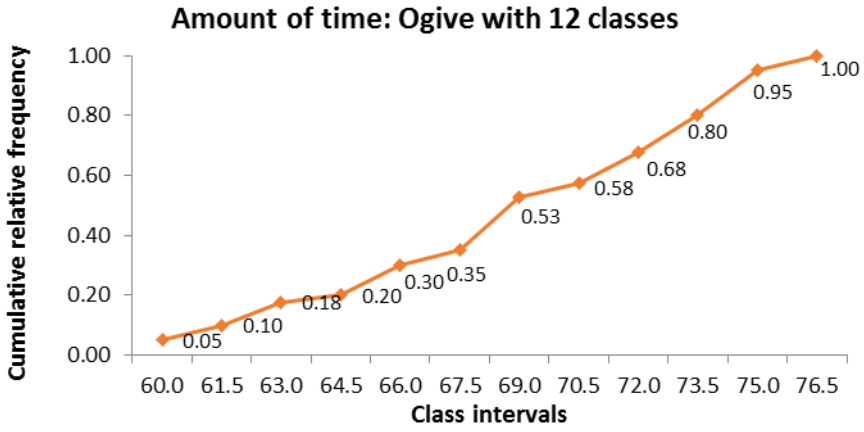
b Frequency distribution with 12 classes

Class	Tally	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
58.5 up to 60.0	||	2	0.05	2	0.05
60.0 up to 61.5	||	2	0.05	4	0.10
61.5 up to 63.0	|||	3	0.08	7	0.18
63.0 up to 64.5	|	1	0.03	8	0.20
64.5 up to 66.0	||||	4	0.10	12	0.30
66.0 up to 67.5	||	2	0.05	14	0.35
67.5 up to 69.0	𝍸 ||	7	0.18	21	0.53
69.0 up to 70.5	||	2	0.05	23	0.58
70.5 up to 72.0	||||	4	0.10	27	0.68
72.0 up to 73.5	𝍸	5	0.13	32	0.80
73.5 up to 75.0	𝍸 |	6	0.15	38	0.95
75.0 up to 76.5	||	2	0.05	40	1.00
Total		40			



c

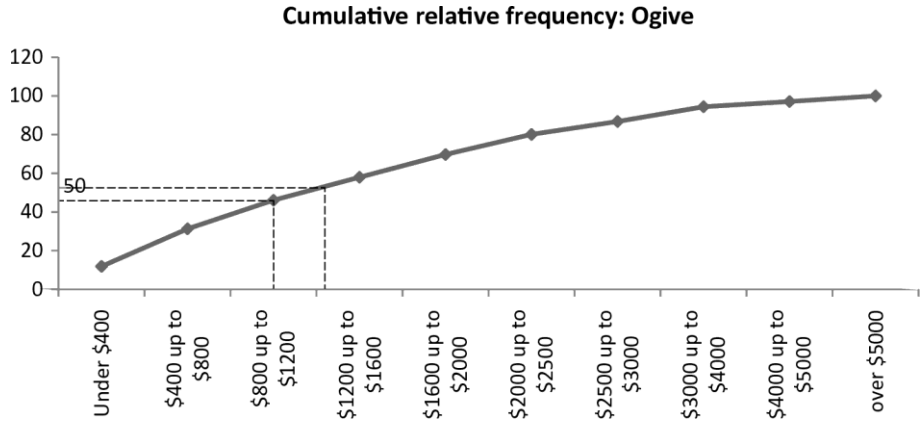




4.9 a

Class intervals	Relative frequency (%)	Cumulative relative frequency (%)
Under $400	11.9	11.9
$400 up to $800	19.4	31.3
$800 up to $1200	14.8	46.1
$1200 up to $1600	12.7	58.8
$1600 up to $2000	10.9	69.7
$2000 up to $2500	10.4	80.1
$2500 up to $3000	6.7	86.8
$3000 up to $4000	7.6	94.4
$4000 up to $5000	2.7	97.1
Over $5000	2.9	100.0

b

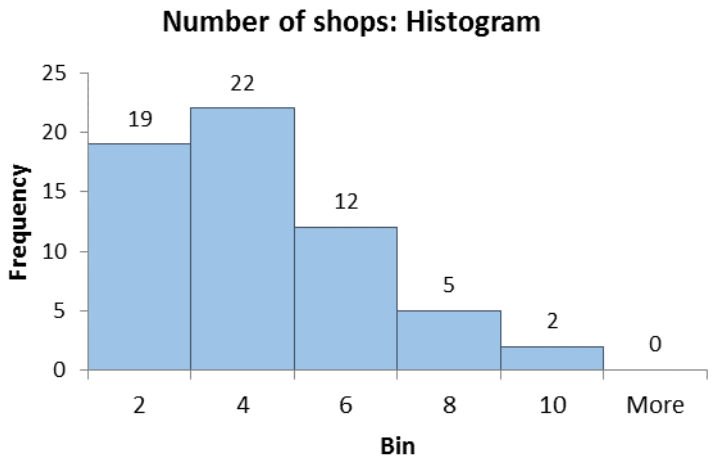


c Annual income = $62400. Weekly income = $1200. Corresponding cumulative relative frequency is 46.1%. Therefore, 46.1% of the annual incomes were less than $62 400 in 2012.

d 50% of the weekly incomes were less than approximately $1250.

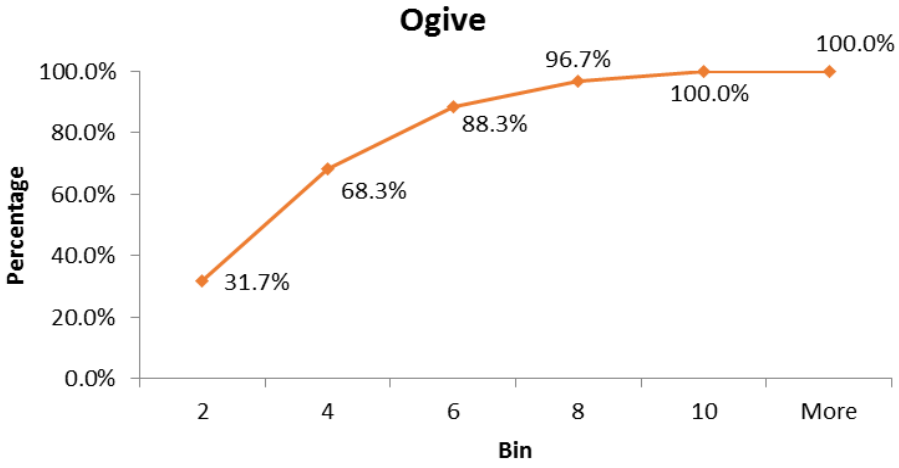
4.10 a

Bin	*Frequency*
2	19
4	22
6	12
8	5
10	2
More	0



b

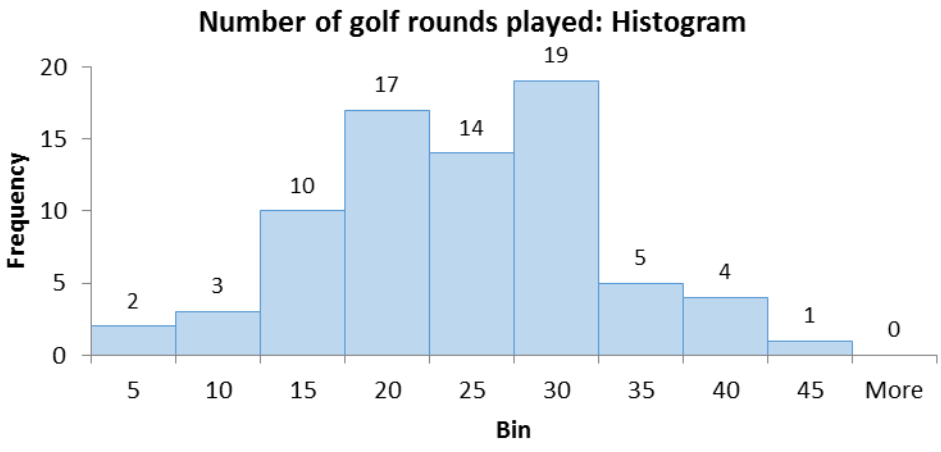
Bin	Frequency	Cumulative %
2	19	31.7%
4	22	68.3%
6	12	88.3%
8	5	96.7%
10	2	100.0%
More	0	100.0%



c The data are skewed to the right, showing that the number of shops entered by most customers is up to 4 shops.

4.11 a

Bin	Frequency
5	2
10	3
15	10



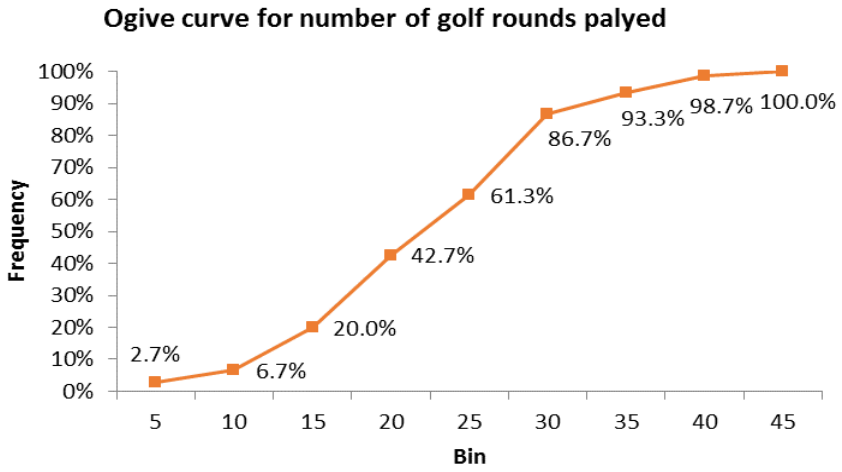
20	17
25	14
30	19
35	5
40	4
45	1
More	0

b

Stem	*Leaf*
0	359
1	00233344455566778888888899
2	000012233344444555666678888899999
3	000001125566668
4	2

c

Bin	*Frequency*	*Cumulative %*
5	2	2.67%
10	3	6.67%
15	10	20.00%
20	17	42.67%
25	14	61.33%
30	19	86.67%
35	5	93.33%
40	4	98.67%
45	1	100.00%



d The number of rounds of golf played has a symmetric distribution, with most players

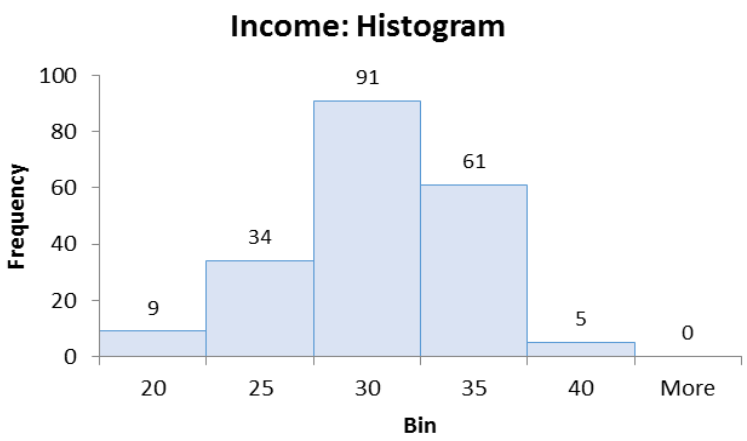
playing on average between 15 and 30 rounds in a year.

4.12 a.

Stem	Leaf
15	36
16	
17	
18	1279
19	169
20	
21	666
22	0003346
23	12223666778
24	24455567889
25	001112344566677799
26	011122334445667788999
27	12445566677799
28	001122333444556689
29	00002222233666778999
30	00111222224444777899
31	00134445555888
32	0011113345666779
33	014579
34	2234479
35	8
36	27
37	3
38	
39	3

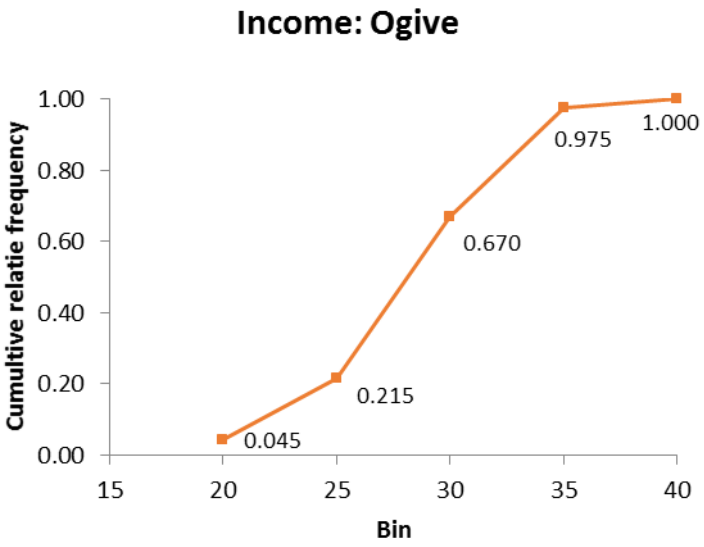
b

Bin	*Frequency*
20	9
25	34
30	91
35	61
40	5
More	0



c The distribution of the annual incomes of the recently-graduated business graduates is approximately bell-shaped, unimodal, with the modal class consisting of incomes between \$25 000 and \$30 000.

d

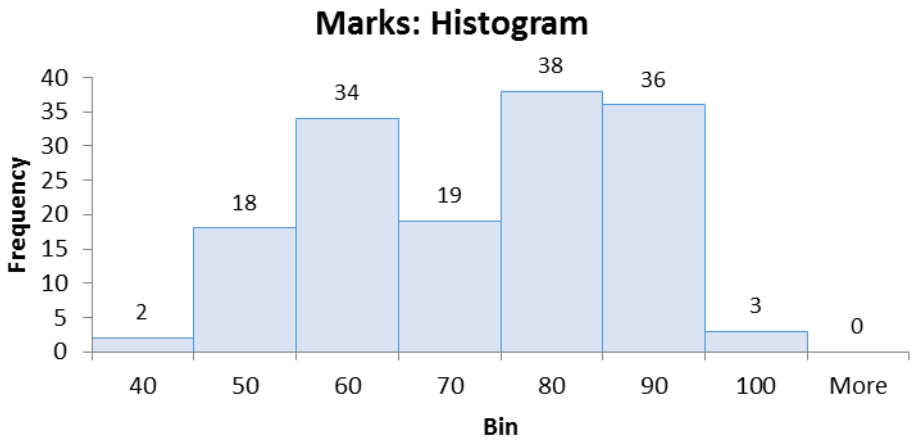


e **i**. The proportion of recently-graduated business graduates who earn less than \$20 000 is 0.045.

ii The proportion who earn more than \$35 000 is 1 – 0.975 = 0.025.

iii The proportion who earn between \$25 000 and \$40 000 is 1.000 – 0.215 = 0.785.

4.13 b

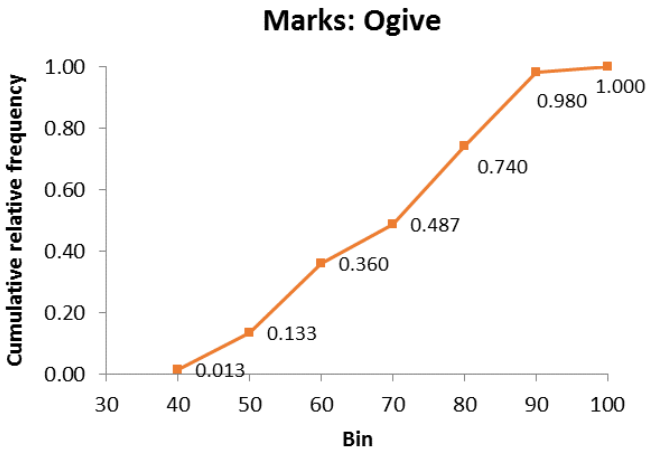


[If you are using Excel, recall that Excel prints the upper limits, bins, of the class intervals in the centre of the classes.]

c The 150 marks range from about 30 up to 100, with two marks below 40 and three marks over 90. The distribution is somewhat bimodal, with approximately 23% of the marks between 50 and 60, and approximately 50% of the marks between 70 and 90.

d

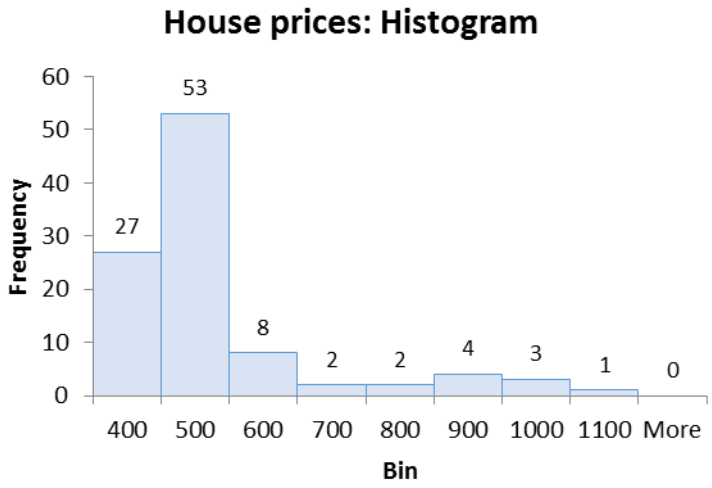
Bin	*Frequency*	*Cumulative %*
40	2	1.33%
50	18	13.33%
60	34	36.00%
70	19	48.67%
80	38	74.00%
90	36	98.00%
100	3	100.00%



e The proportion of marks less than 70 is 0.4867, or 48.67%.

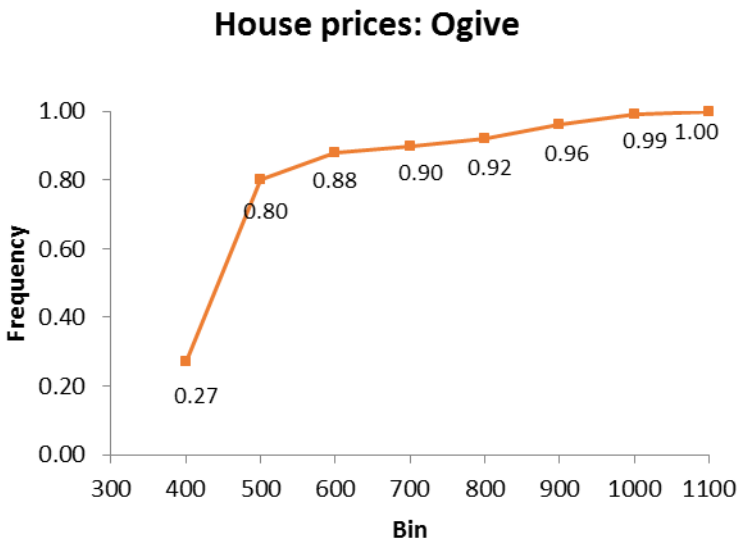
f The proportion of marks less than 75 is $0.6143 = \left(0.4867 + \frac{0.74 - 0.4867}{2}\right)$

4.14 a



The histogram shows that the distribution of the price data is skewed to the right.

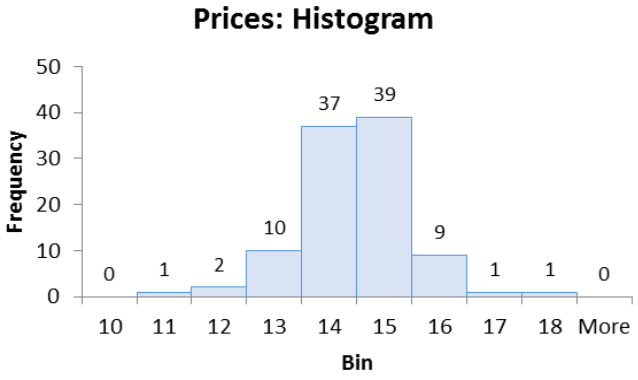
b



c About 27% of the house prices are less than $400 000.

d About 84% of the house prices are less than $550 000.

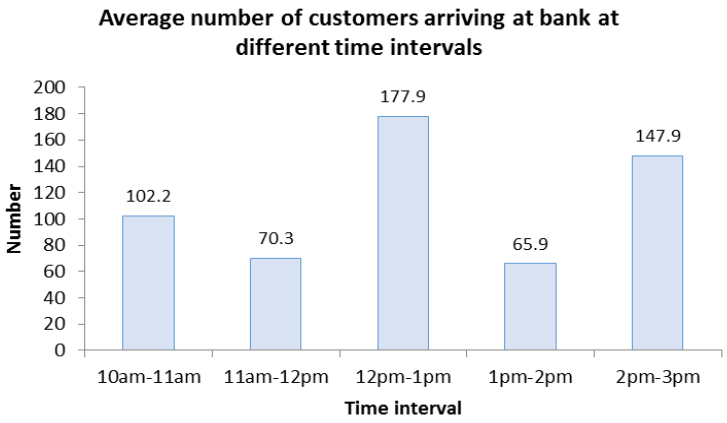
4.15 a

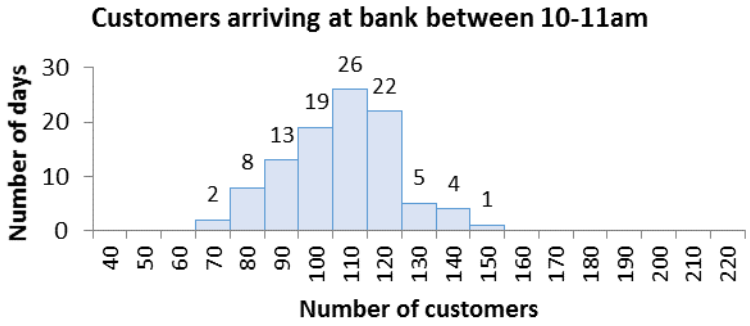


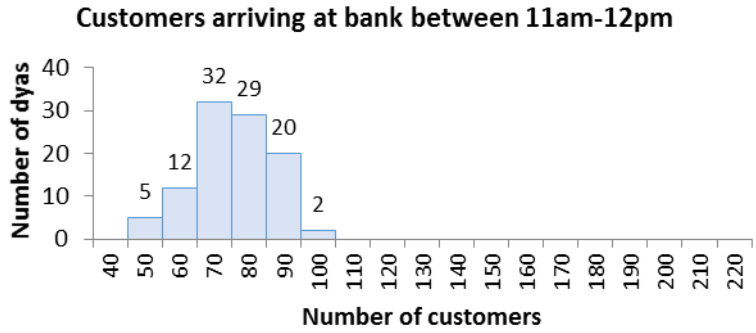
[If using Excel, recall that Excel prints the upper limits of the class intervals in the centre of the classes.]

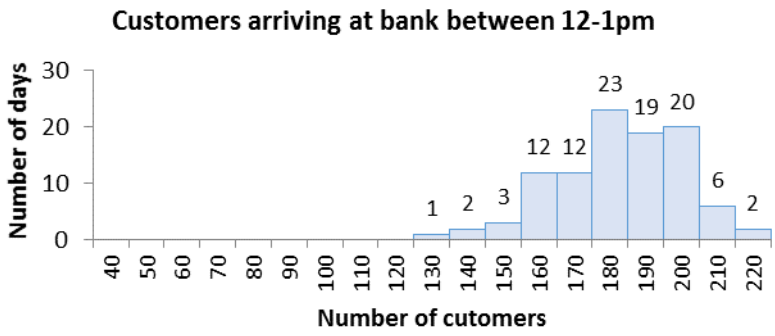
b The distribution of the 100 prices is quite symmetrical, with about 75% of the prices between $13 and $15, and about 95% of the prices between $12 and $16.

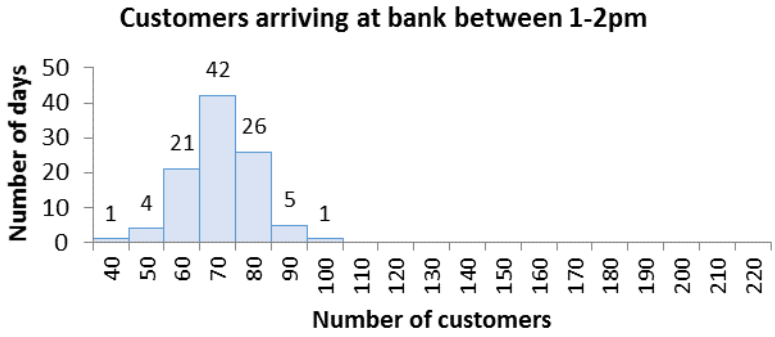
4.16 a

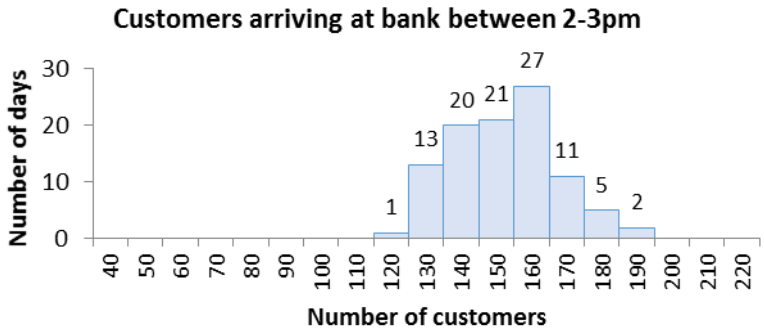












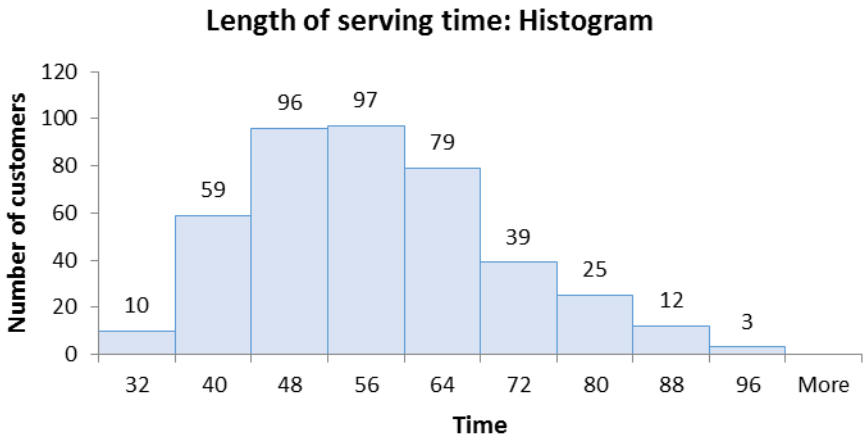
b With the exception of the noon hour (12:00 to 1:00pm), each period's distribution is unimodal and only very roughly bell-shaped. The distribution of the number of customers arriving during the noon hour is more spread out and has three central classes with approximately the same frequencies (each of which is close to being the modal class).

c From the first graph (the bar chart), the noon hour attracts the highest average number of customers, followed by the 2:00 to 3:00pm period. The 1:00 to 2:00pm period has the lowest average number.

d One suggestion would be to use the 11:00 to 12:00pm and 1:00 to 2:00pm periods for tea/lunch breaks for the bank employees, as these are the two periods with the lowest average number of customers.

4.17 **a** This histogram should contain 9 or 10 class intervals. Minimum = 28, Maximum = 92, we chose 9 classes with class width 8.

b



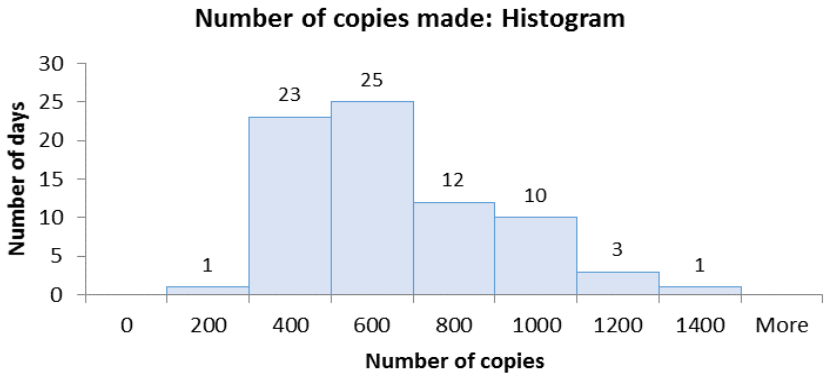
c The histogram is positively skewed.

d This is a unimodal distribution.

e The histogram is not bell-shaped.

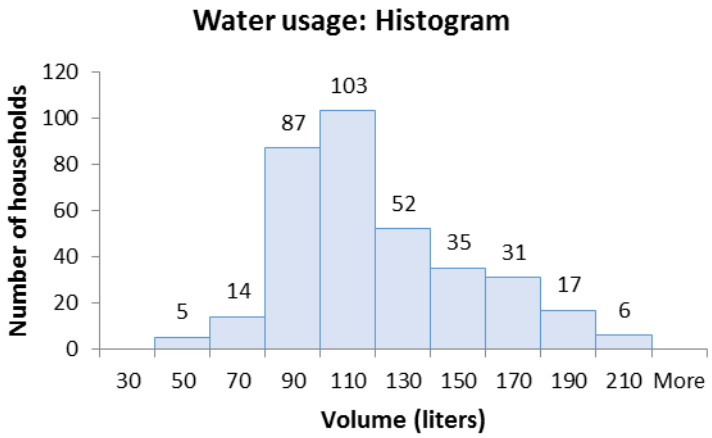
4.18

Bin	*Frequency*
0	0
200	1
400	23
600	25
800	12
1000	10
1200	3
1400	1
More	0



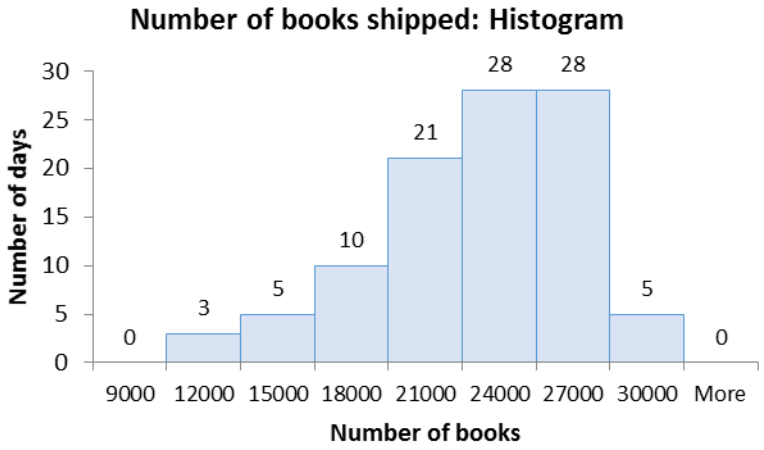
The histogram is unimodal and positively skewed.

4.19



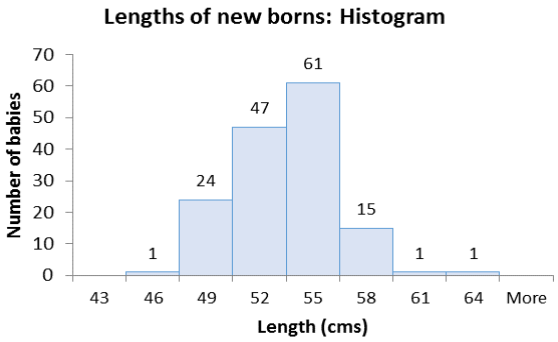
The histogram is positively skewed and unimodal. Most households use between 70 and 130 litres per day. The centre of the distribution appears to be around 70 to 110 litres.

4.20



The histogram of the number of books shipped daily is negatively skewed. It appears that there is a maximum number that the company can ship.

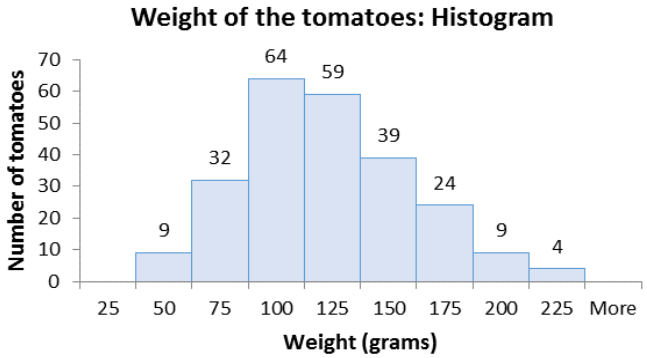
4.21



[If using Excel, recall that Excel prints the upper limits of the class intervals in the centre of the classes.]

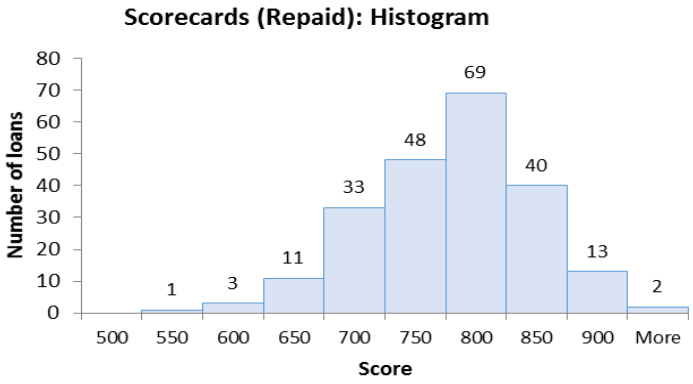
The histogram is unimodal, bell-shaped and roughly symmetric. Most of the lengths lie between 46cms and 55cms.

4.22

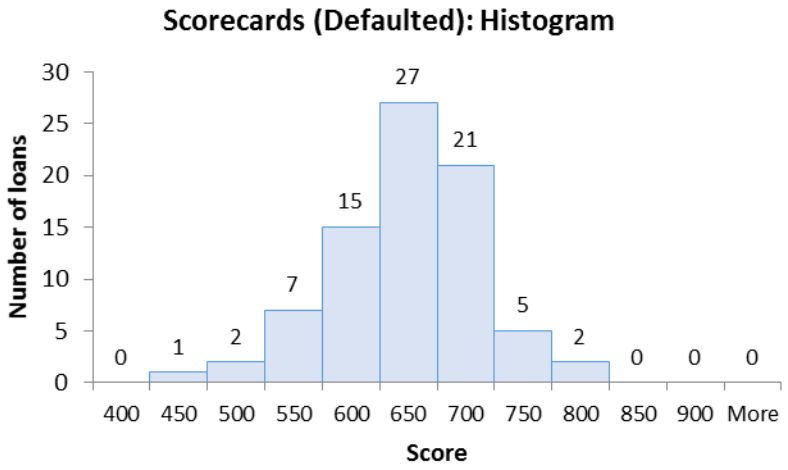


The histogram is unimodal, symmetric and bell-shaped. Most tomatoes weigh between 50 and 175grams with a small fraction weighing less than 50 grams or more than 175 grams.

4.23 a

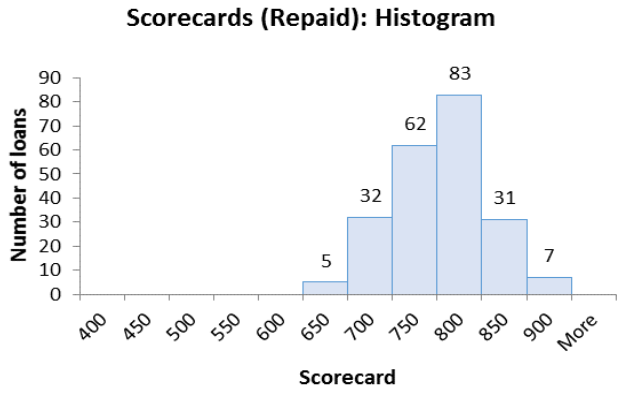


b

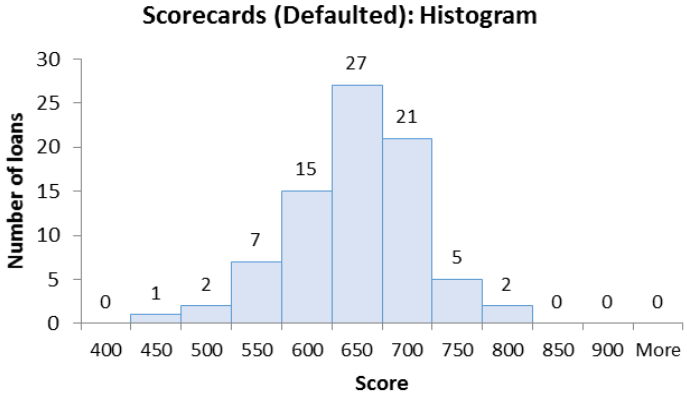


c The scorecards appear to be relatively poor predictors.

4.24 a

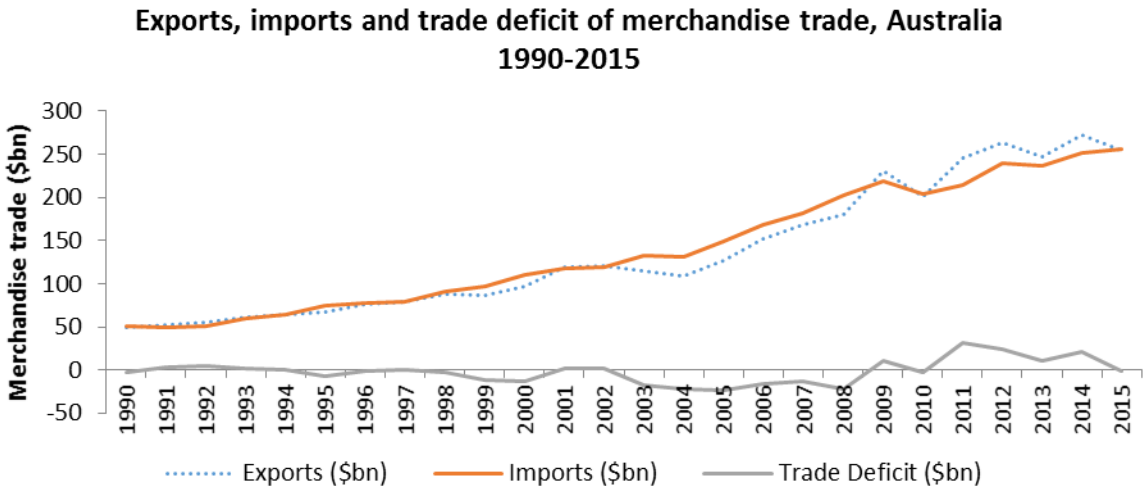


b

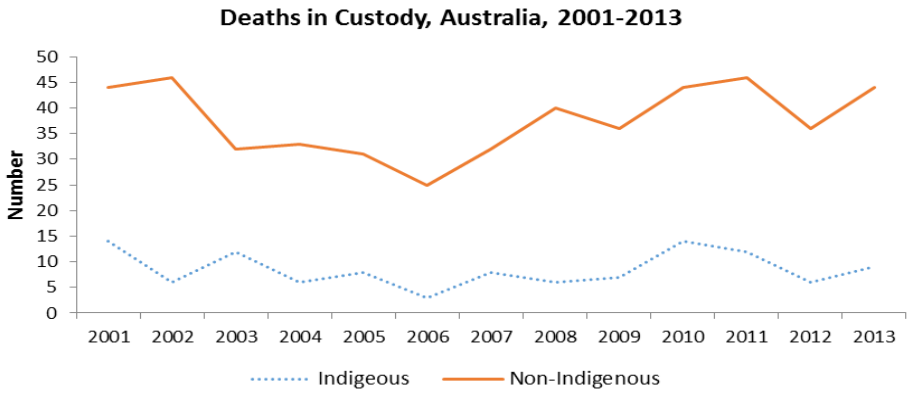


c & **d** This scorecard is a much better predictor.

4.25

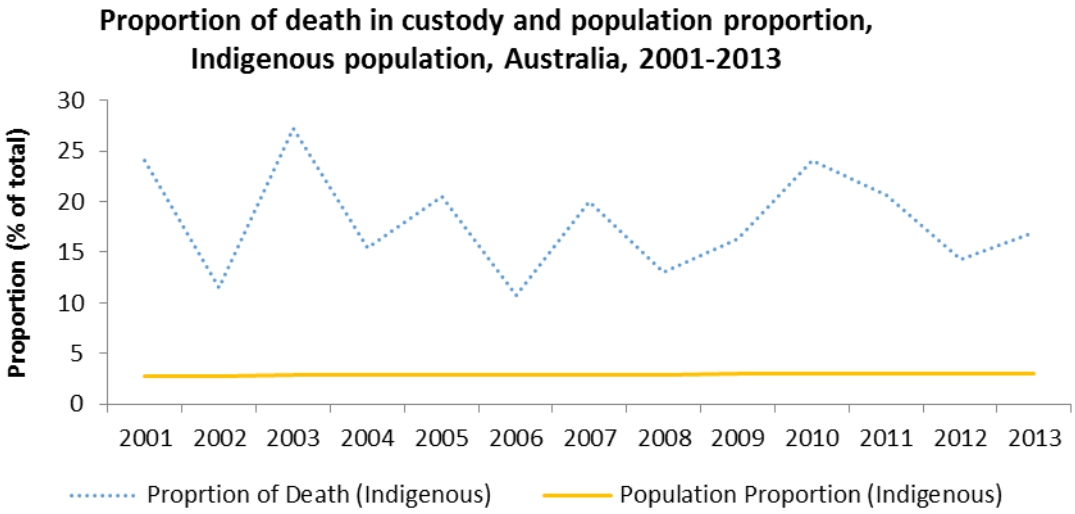


4.26 a



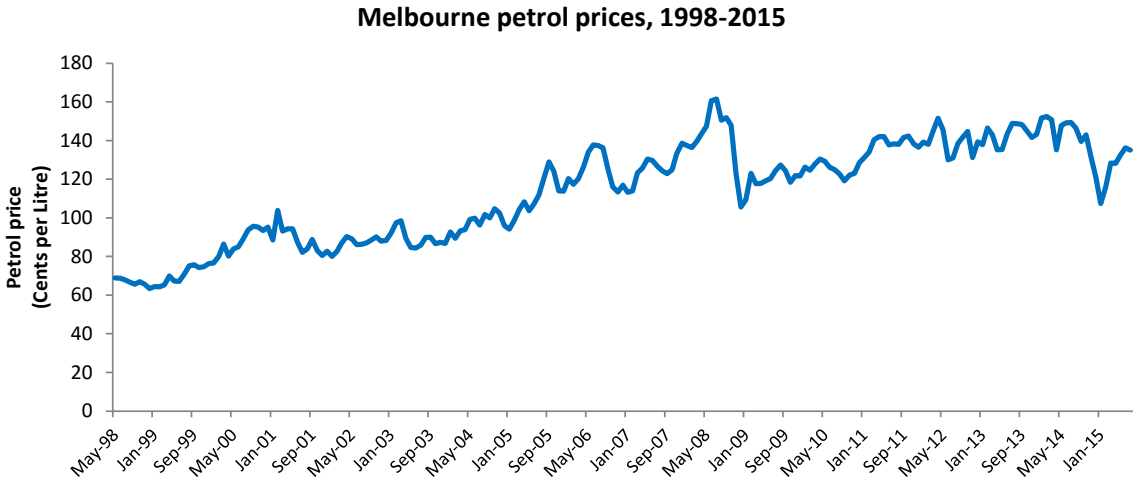
b

Year	Proportion of Death (Indigenous)	Population Proportion (Indigenous)
2001	24.1	2.76
2002	11.5	2.79
2003	27.3	2.83
2004	15.4	2.86
2005	20.5	2.89
2006	10.7	2.92
2007	20.0	2.93
2008	13.0	2.93
2009	16.3	2.94
2010	24.1	2.96
2011	20.7	2.97
2012	14.3	2.98
2013	17.0	3.00



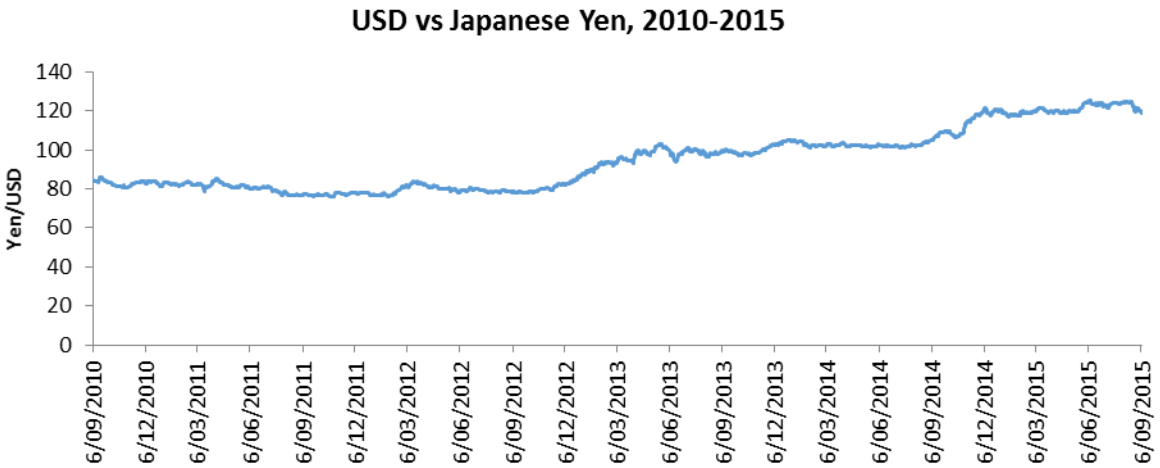
c The proportion of Indigenous Australian population is less than 3%, while the proportion of deaths in custody of Indigenous Australians is more than 10% and sometimes as high as 27%.

4.27



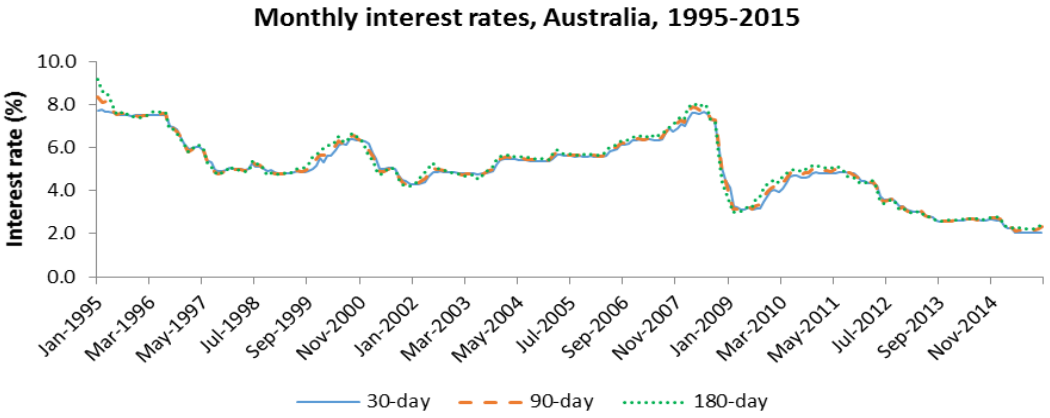
Petrol prices in Melbourne, in general, continued to increase until May 2008 and fell sharply in the latter part of 2008, and then continued to increase again. However, there is a sharp fall in petrol prices in early 2015 but has recovered quickly.

4.28 a

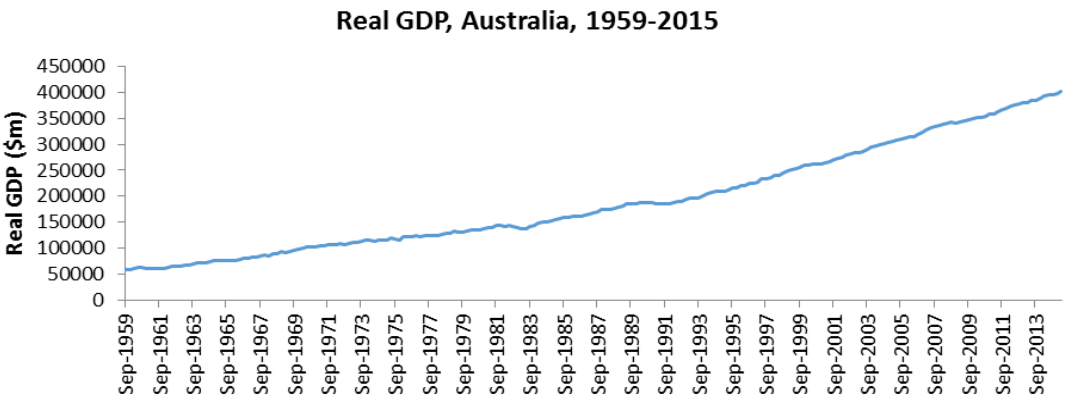


b The line graph shows quite a lot of random movements in the spot rate.

4.29 A line chart would be appropriate. As can be seen, the three interest rate series move similarly throughout the time period.

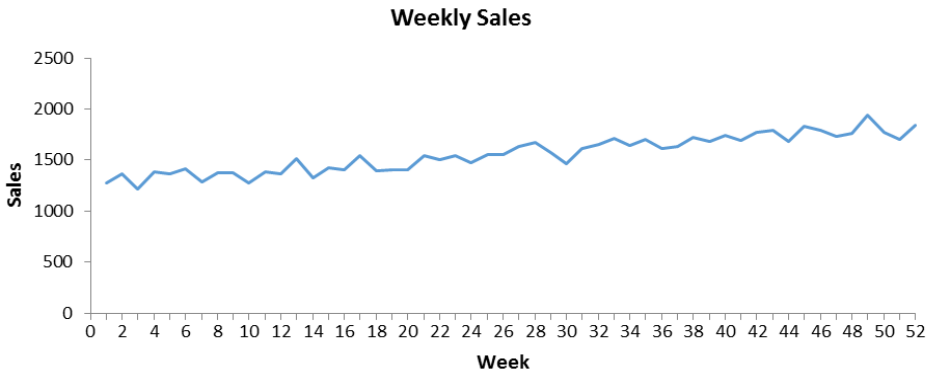


4.30



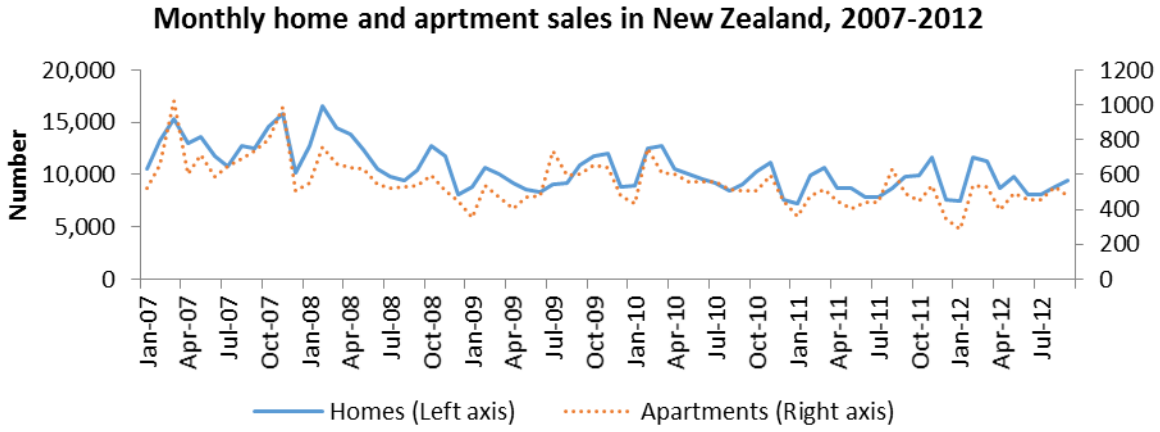
Except for a slight decline in 1983,1991 and 2009, the quarterly GDP series has been steadily increasing throughout 1959 to 2015.

4.31 a



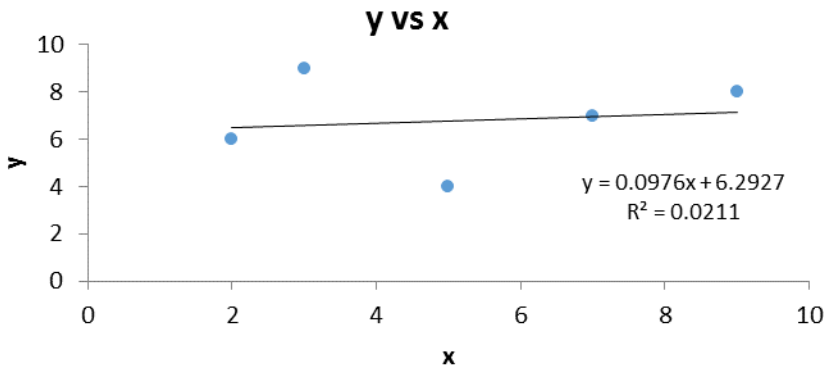
b There has been a gradual weekly increase in sales over the 52 weeks.

4.32

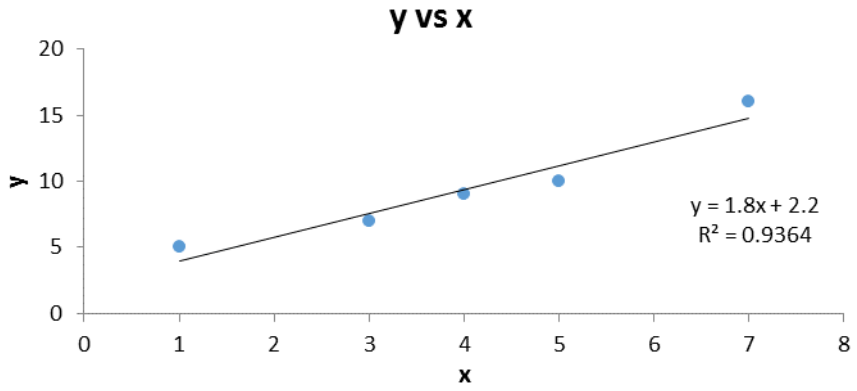


For the two time-series data a line chart would be appropriate. The chart shows that more homes are sold than apartments. The patterns of monthly sale of both types seem to be similar during 2007-2012.

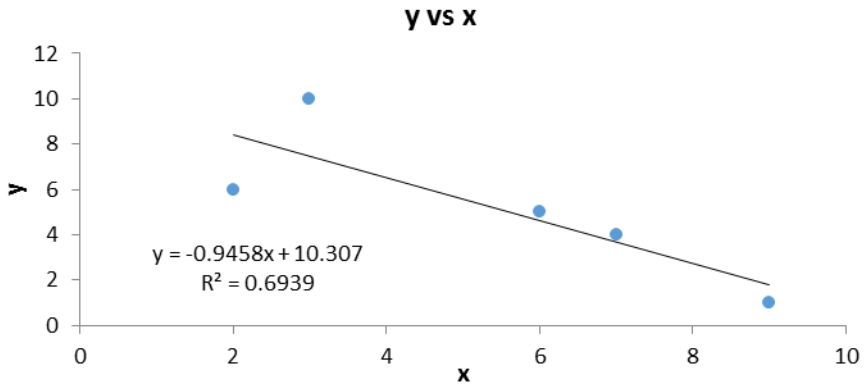
4.33 a



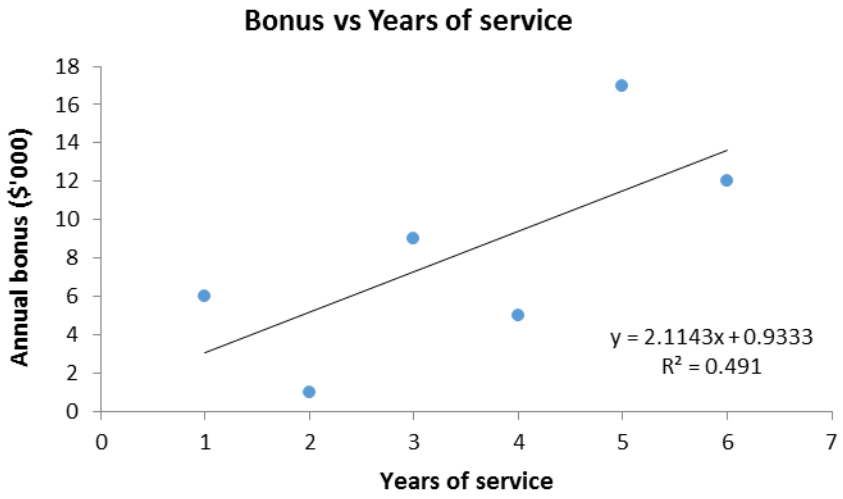
b



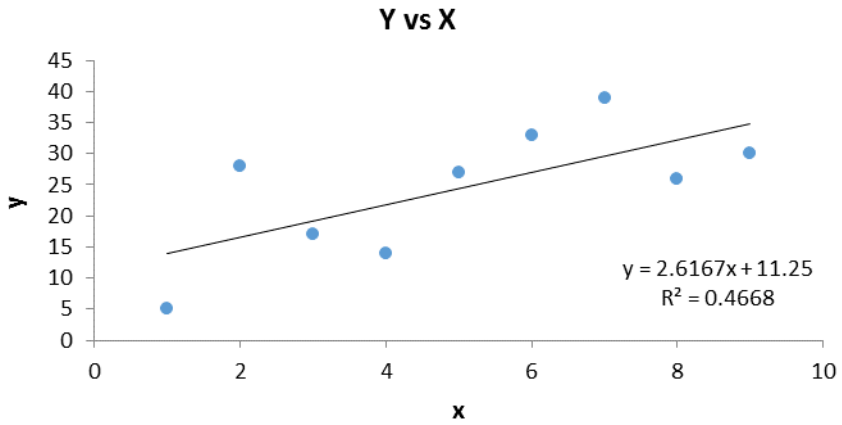
c



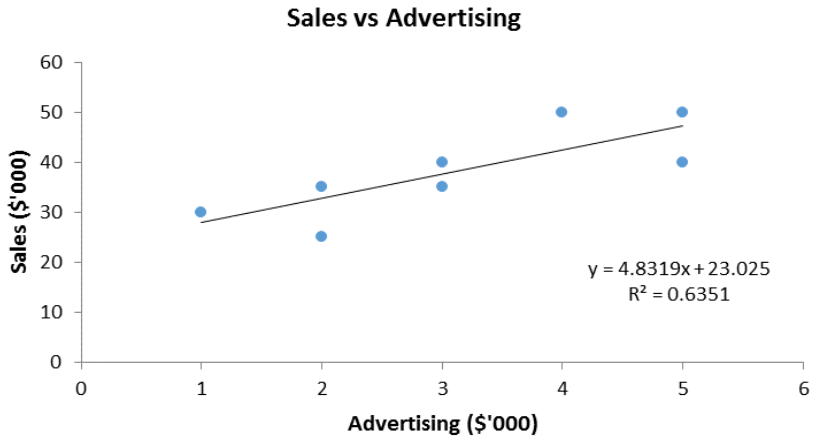
4.34 **a** and **b**



4.35 **a** and **b**

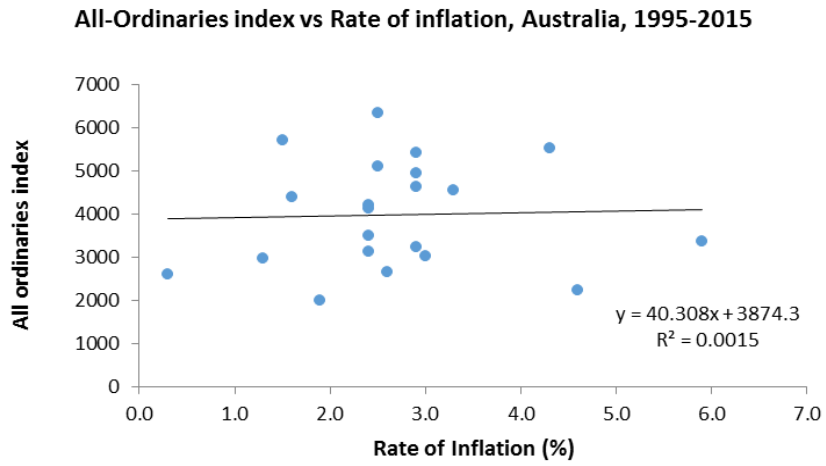


4.36



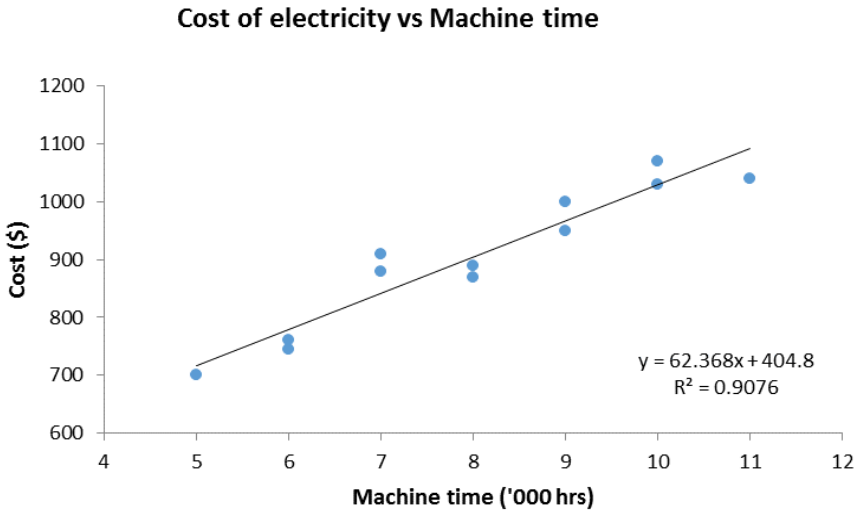
Sales levels seem to have a positive linear relationship with advertising expenditure.

4.37 **a** and **c**



b There is a very weak positive linear relationship between the All Ordinaries index and rate of inflation.

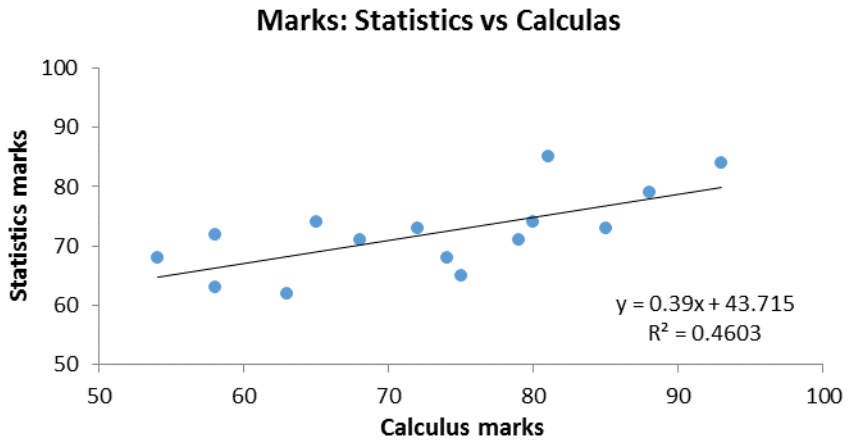
4.38 **a** and **c**



b A positive linear relationship exists between hours of machine usage (x, '000) and – cost of electricity (y, \$).

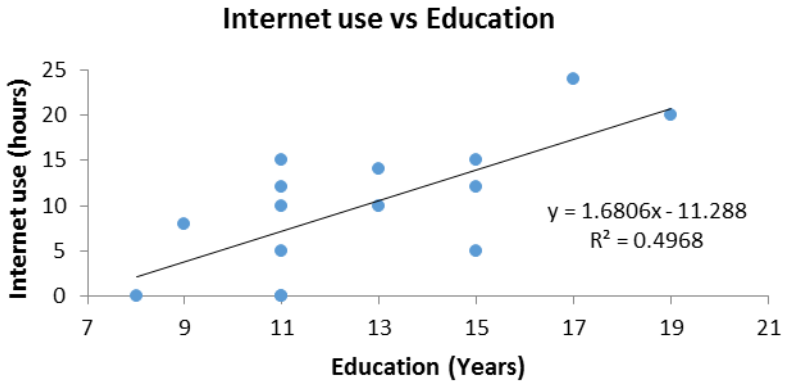
d The least squares line of fit is $\hat{y} = 404.8 + 62.368x$. Fixed cost of electricity is \$404.80 and for every 1000hrs of additional machine time, cost would increase by \$62.37.

4.39 a



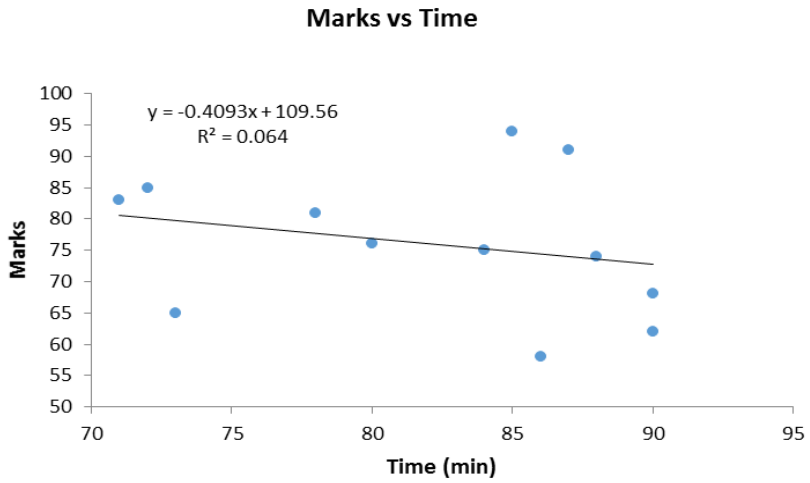
b Higher the calculus marks the higher the statistics marks. The points in the scatter plot are all scattered around an upward sloping straight line. That is, there appears to be a positive linear relationship between the calculus marks and the statistics marks. Considering the R^2, only 46% of the variation in Statistics marks is explained by the calculus marks.

4.40 a



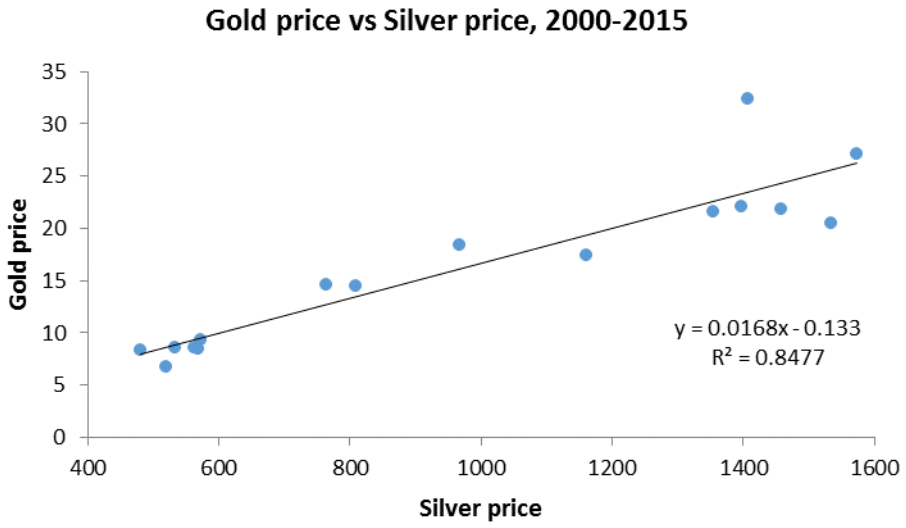
b Higher the level of education the higher the internet-use. The points in the scatter plot are all scattered around an upward sloping straight line. That is, there appears to be a positive linear relationship between the level of education and the hours of internet use. Considering the R^2, only about 50% of the variation in hours spent on the internet is explained by the years of education.

4.41 a



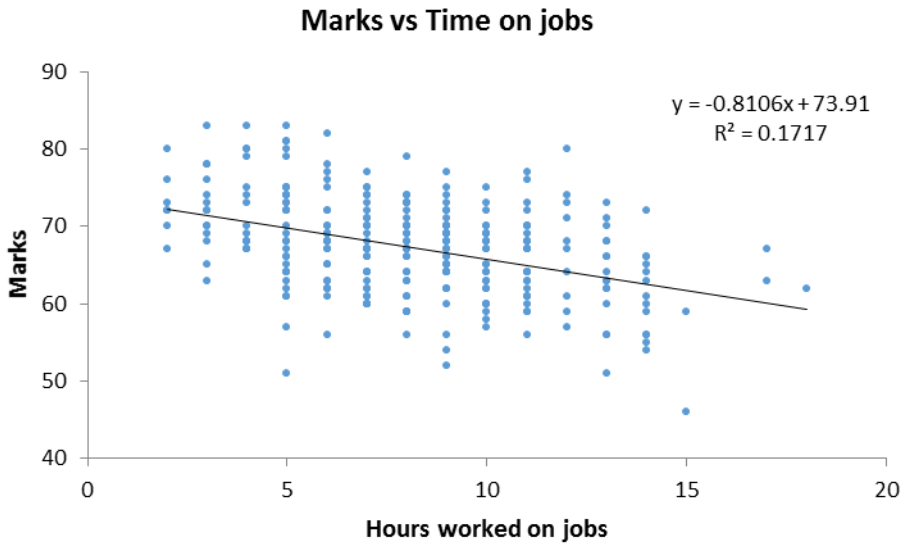
b Higher the time spent on exam the slightly lower the exam mark. The points in the scatter plot are all scattered around a slightly downward sloping straight line. That is, there appears to be a slight negative linear relationship between the amount of time spent and the exam mark. Considering the R^2, only about 6% of the variation in marks is explained by the time spent doing the exam. The relationship is very weak.

4.42 a



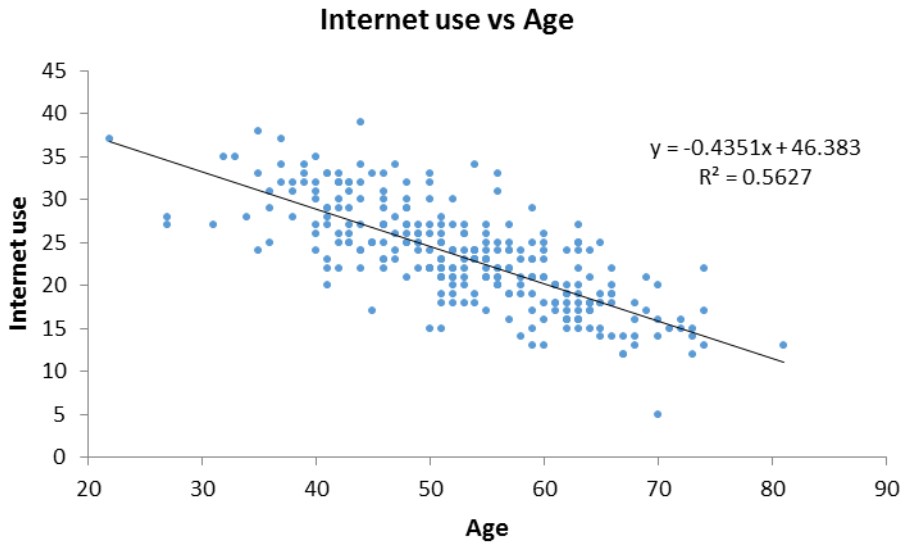
b Overall, it seems that there is a strong positive linear relationship between gold price and silver price.

4.43 a



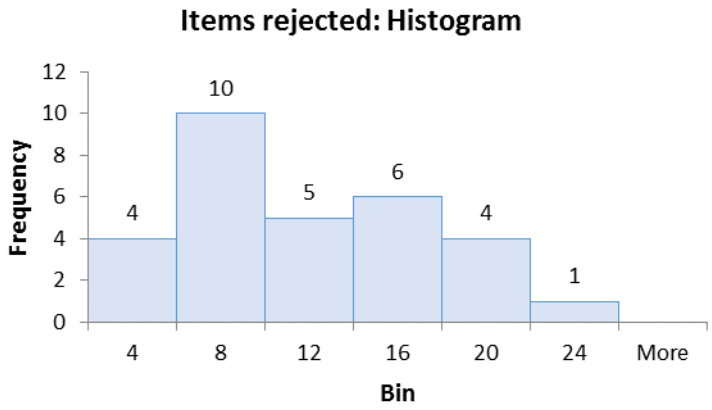
b There is a weak negative linear relationship.

4.44 a A scatter diagram with age on the horizontal *x*-axis and the number of hours of Internet use on the vertical *y*-axis would be appropriate.

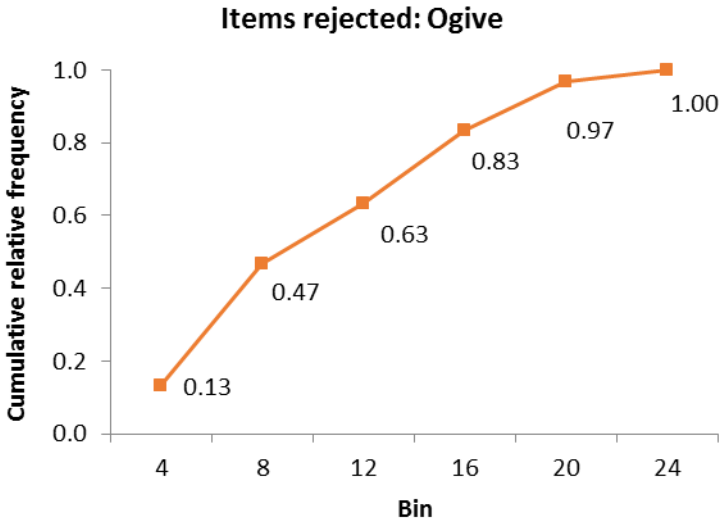


b There appears a moderate negative linear relationship. The older the person, the lesser the Internet use.

4.45 a



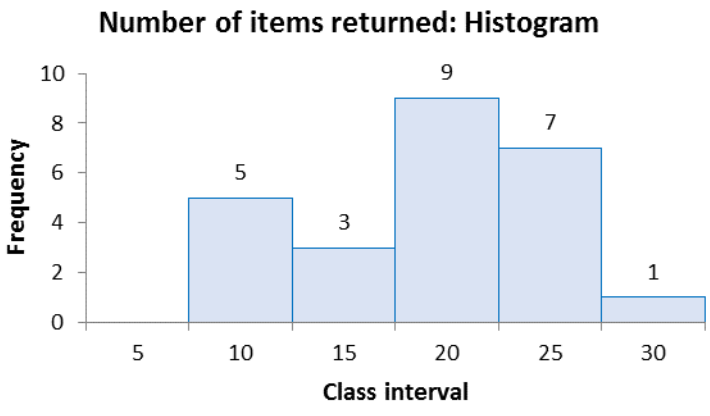
b



c The data are slightly skewed to the right and is bimodal.

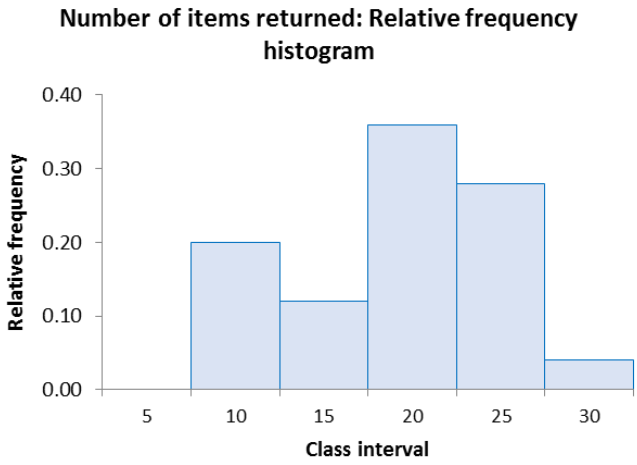
4.46 a

Bin	*Frequency*
5	0
10	5
15	3
20	9
25	7
30	1



b

Bin	*Relative frequency*
5	0.00
10	0.20
15	0.12
20	0.36
25	0.28
30	0.04



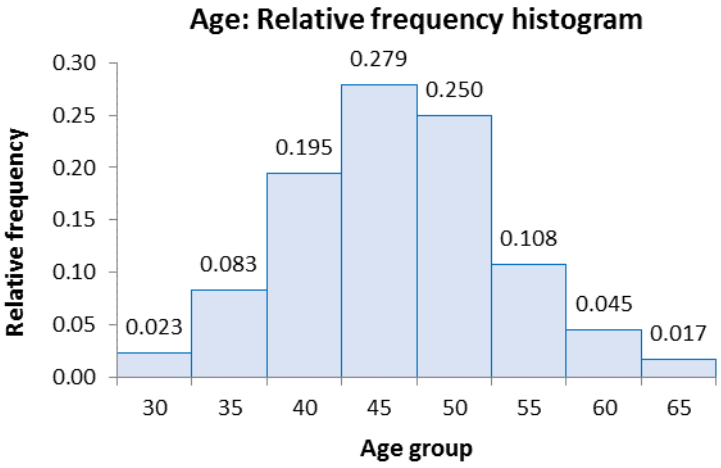
c The two histograms are basically the same except for the vertical axis scale. The relative frequency axis gives the frequencies divided by the total number of items. The area under the frequency histogram is proportional to the area under the relative frequency histogram for each class interval.

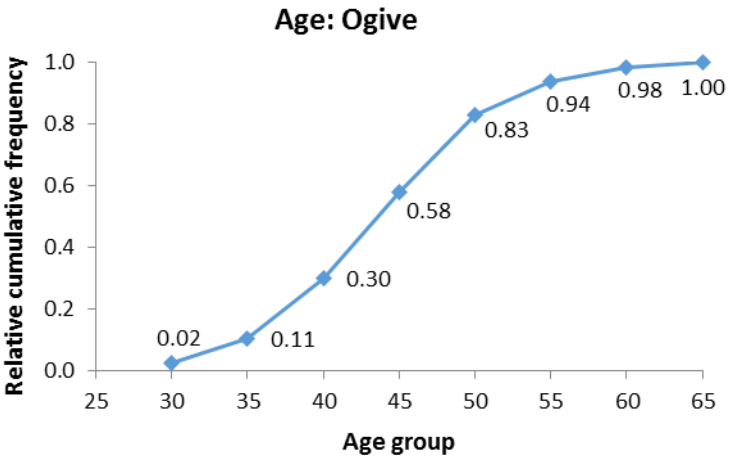
4.47 **a** Line graphs for male and female on the same plot would be appropriate.

b Line graphs are the appropriate plot for time series data. The line graphs on the same plot allows us to compare the male and female unemployment rate and also see the trend and movement of the male and female unemployment rates from 2000-2015.

4.48

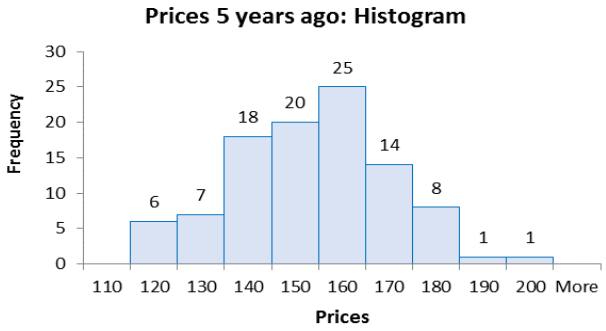
Bin	Relative frequency	Cumulative relative frequency
30	0.023	0.023
35	0.083	0.106
40	0.195	0.301
45	0.279	0.580
50	0.250	0.830
55	0.108	0.938
60	0.045	0.983
65	0.017	1.000

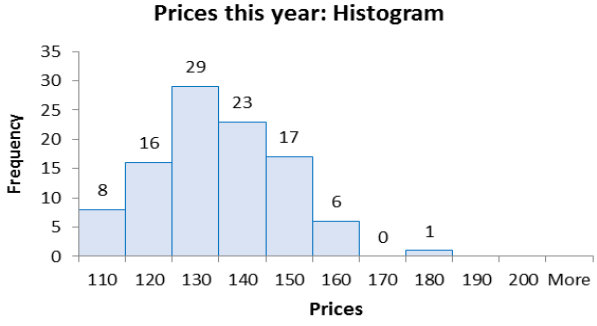




As can be seen from the relative frequency table and the ogive curve, about 42 percent of the academics are now 45 years or older who will retire within the next 20 years. About 2 percent will retire in the next 5 years, 6 percent within the next 10 years, 17 percent within the next 15 years, and 42 percent within the next 20 years. Therefore, careful planning is required to replace these academics as there may be not enough new trained people to be absorbed as academics.

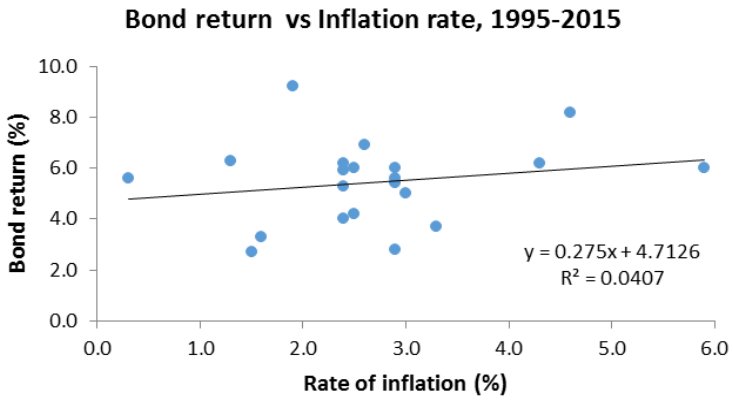
4.49 a





b It appears that prices of houses have fallen and less dispersed this year compared to the house prices 5 years ago.

4.50 a

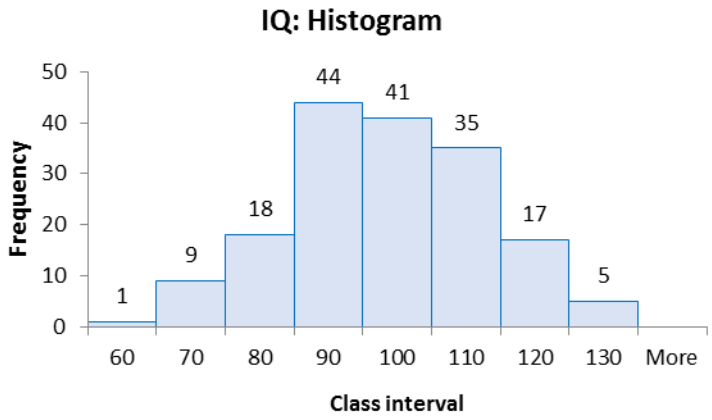


b A positive linear relationship exists between rate of inflation and bond return.

c Yes

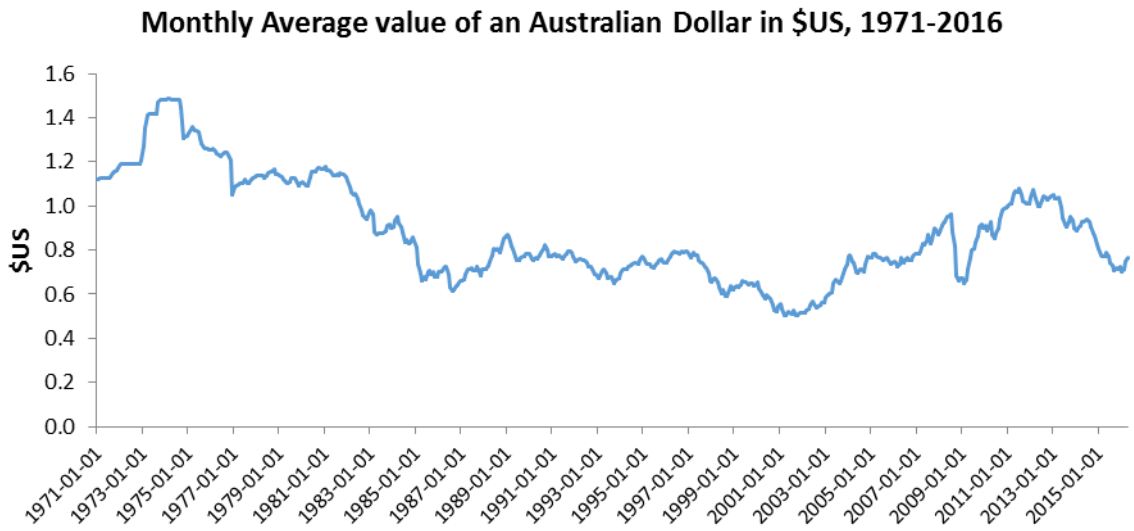
d Bond return = 4.7126 + 0.275 × (Rate of inflation)

4.51

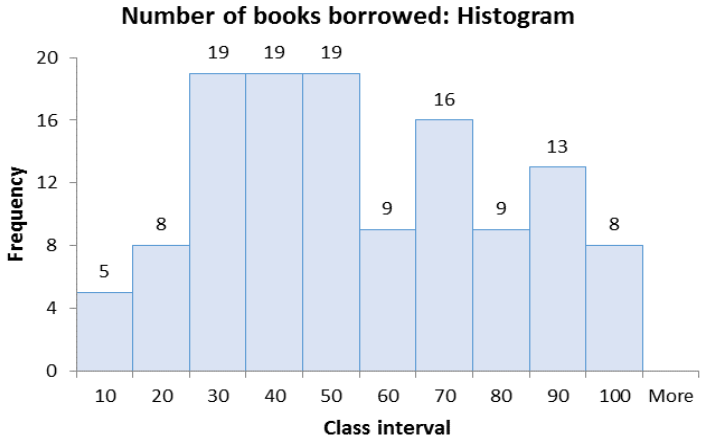


Most of the IQ is somewhere between 80 and 110. The distribution is slightly symmetrical.

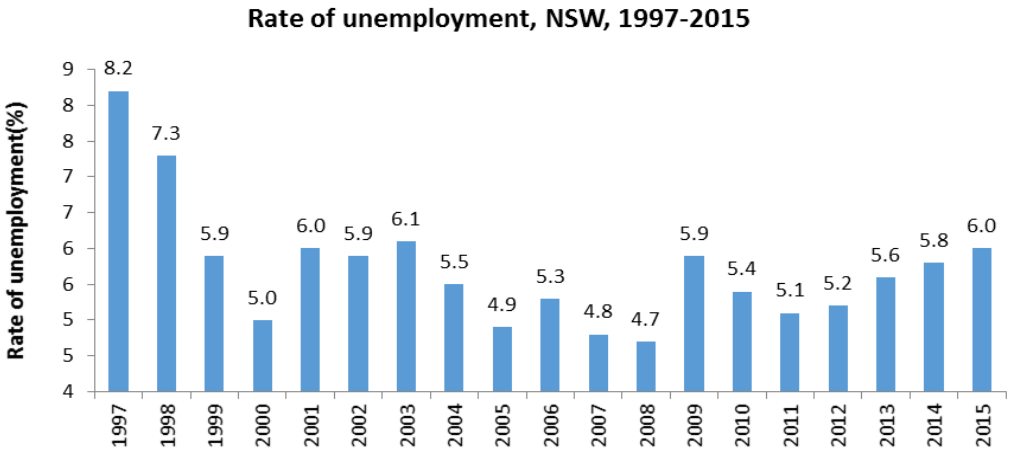
4.52



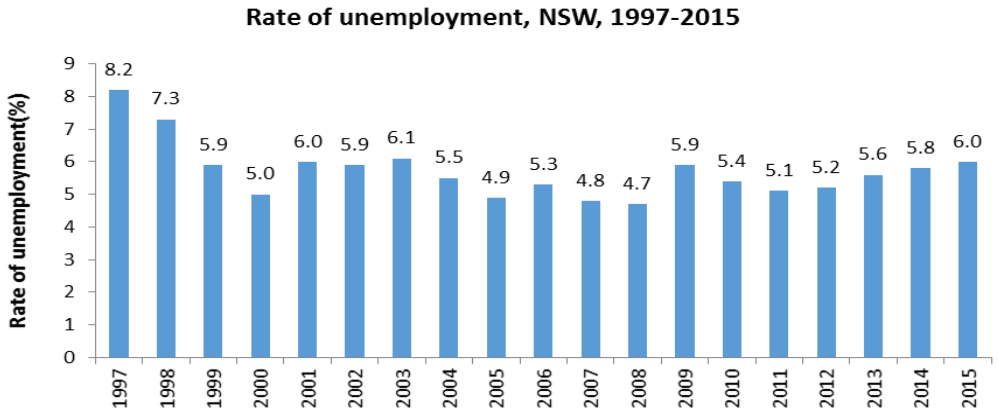
4.53



4.54 a



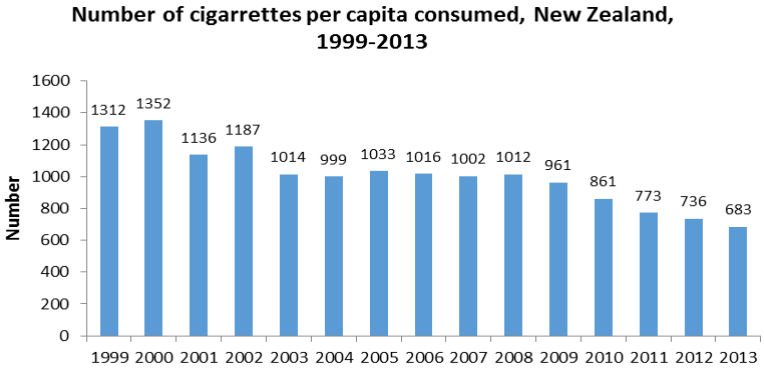
b

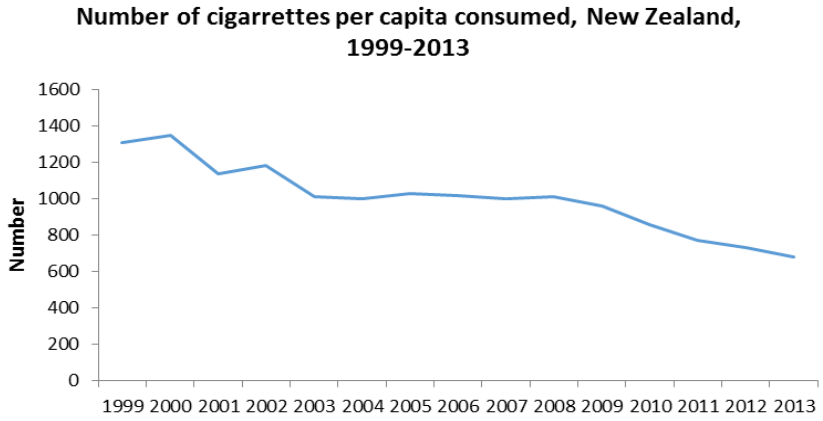


c Even though the data depicted are the same; the first bar chart whose vertical range is only 5%, seems to show the differences in the unemployment rate better than the second bar chart where the vertical range is 9%.

d The bar chart in (a) shows the rate changes better throughout the sample period.

4.55





A time series (line) graph of number of cigarettes per capita against year would be appropriate to show the decline in per capita cigarette consumption.

5 Numerical descriptive measures

5.1 Shift 1 : $\bar{x} = \dfrac{24+17+35+15+19}{5} = \dfrac{110}{5} = 22$

Shift 2 : $\bar{x} = \dfrac{21+13+15+20+18}{5} = \dfrac{87}{5} = 17.4$

5.2 a $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{55+25+15+0+105+45+60+30+35+80+40+5}{12} = \dfrac{495}{12} = 41.25$

Ordered data: 0, 5, 15, 25, 30, 35, 40, 45, 55, 60, 80, 105; Median $= \dfrac{(35+40)}{2} = 37.5$

Mode = N/A

5.3 a

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{5.5+7.2+1.6+22.0+8.7+2.8+5.3+3.4+12.5+18.6+8.3+6.6}{12} = \frac{102.5}{12} = 8.54$$

Ordered data: 1.6, 2.8, 3.4, 5.3, 5.5, 6.6, 7.2, 8.3, 8.7, 12.5, 18.6, 22.0; Median = 6.9
Mode = N/A

b The mean number of kilometres jogged is 8.54. Half the sample jogged more than 6.9 kilometres and the other half jogged less.

5.4 a Mean = $ 96,500, Median = $ 770,000, Mode = $ 770,000

b Mean > Median = Mode. Therefore, the house price is skewed to the right. Median is the best measure to represent the house prices.

5.5 a Mean = (12 + 0 +. . . + 7) / 20 = 7.7 days

Ordered values:

0	2	3	4	4	4	5	6	7	8	8	9	9	10	11	12	12	13	13	14

Median = (10^{th} value + 11^{th} value)/2 = 8

x	0	2	3	4	5	6	7	8	9	10	11	12	13	14
f	1	1	1	3	1	1	1	2	2	1	1	2	2	1

Mode = 4

b Mode < Median, Mode < Mean. The distribution is skewed to the right (or positively skewed).

5.6 a The mean age is

$$\mu = \frac{\sum_{i=1}^{7} x_i}{7} = \frac{19+19+65+20+21+18+20}{7} = \frac{182}{7} = 26$$

To find the median, first arrange the ages in ascending order: 18, 19, 19, 20, 20, 21, 65
The median is the middle age, which is 20. There are two modes: 19 and 20. These are the ages that occur most frequently.

b Suppose that the highest age, 65, is removed from the data set. The new mean age is

$$\mu = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{19+19+20+21+18+20}{6} = 19.5$$

The ages arranged in ascending order are now: 18, 19, 19, 20, 20, 21
The median is 19.5, the mean of the two middle ages. There are still two modes: 19 and 20
Notice that the mean has decreased substantially, the median has decreased only slightly, and the modes haven't changed at all.

5.7 a Mean = 72, median = 72, mode = N/A
b Mean = 82, median = 73, mode = N/A
c The outlier made a significant difference to the mean, but not to the median.

5.8 The mean number of cars owned is

$$\mu = \frac{3(0)+10(1)+4(2)+2(3)+1(4)}{20} = \frac{28}{20} = 1.4$$

The median is the mean of the two middle numbers: $\frac{1+1}{2} = 1$

The mode is 1, the number of cars that occurs most frequently.

5.9 a Mean = 975.5, median = 856. The mean of the 12 highest compensations ($000s) is μ = 975.5. The median is the mean of the two middle compensations: 856.
b Right skewed as mean > median > mode. The mean is considerably larger than the median, indicating that its value is being influenced by a few very large compensations. The distribution is positively skewed.
c Mean = 862.8, median = 856. Notice that the mean has been substantially affected, whereas the median has been only slightly affected.

5.10 a Mean = 23.95
Median = 27.52. Mean < Median. Therefore, the data are left skewed.
b Mean = 29.18, Median = 28.41. Mean has changed due to the 2 outliers. Median is not affected by the 2 outliers.

5.11 a Mean = 66.85
b Median = 72.3
c Mean 64.95, median 72.3. With or without the outliers, median is not affected. But mean

has changed.

5.12 a Before dropping the extreme scores:

$$\overline{x}_A = \frac{6.0 + 7.0 + 7.25 + 7.25 + 7.5 + 7.5 + 7.5}{7} = \frac{50}{7} = 7.14$$

$$\overline{x}_B = \frac{7.0 + 7.0 + 7.0 + 7.25 + 7.5 + 7.5 + 8.5}{7} = \frac{51.75}{7} = 7.39$$

After dropping the extreme scores:

$$\overline{x}_A = \frac{7.0 + 7.25 + 7.25 + 7.5 + 7.5 + 7.5}{6} = \frac{44}{6} = 7.33$$

$$\overline{x}_B = \frac{7.0 + 7.0 + 7.0 + 7.25 + 7.5 + 7.5}{6} = \frac{43.25}{6} = 7.21$$

b Competitor *B* has the highest mean before dropping the extreme scores, while competitor *A* has the highest mean after dropping the extreme scores.

Before dropping the extreme scores:
$\text{Median}_A = 7.25$
$\text{Median}_B = 7.25$
After dropping the extreme scores:
$\text{Median}_A = 7.38$
$\text{Median}_B = 7.13$
The two competitors have similar median score both before and after.

5.13 Mean = 11.19; Median = 11
The mean number of days is 11.19 and half the sample took less than 11 days and half took more than 11 days to pay. The average number of days taken for payment has declined (from 15 to 11.19) with the new colour invoices.

5.14 a Mean = 5.85 minutes
Median = 5 minutes
Mode = 5 minutes

b The 'average' of all the times was 5.85 minutes. Half the times were less than 5 minutes, and half were greater. The times most frequently taken was 5 minutes. A few large observations pulled the mean upward.

5.15 a Mean = 238.02, median = 224.1, mode = 240.

b A bill of $240 was the most common, while half the bills were for less than $224.

5.16 a Mean = $2432.88
Median = $2446.10

b The distribution is reasonably symmetrical. But a few low incomes have pulled the mean below the median, resulting in a distribution slightly skewed to the left.

c Either measure could be used, but the median is better as it is not affected by a few low incomes.

5.17 a Mean = $28016

Median = $28250

b Distribution of incomes may be symmetric as mean and median are reasonably close.

5.18 **a** Mean = $475,910 Median = $435,000

b Mean > Median, the house prices distribution may be skewed to the right.

5.19 **a** Mean = 6.17 metres
Median = 5 metres
Mode = 5 metres

b The mean is unduly influenced by extreme observations, the median doesn't indicate what lengths are most preferred, and the mode doesn't consider any desired lengths other than the one most frequently purchased.

c Use this example to anticipate the usefulness of quartiles and a box plot. Since $Q_1 = 4$, $Q_3 = 6$, and a box plot shows only 8 lengths of 10metres or more, you might decide upon lengths of 4, 5, 6 and 10metres.

5.20 **a** Mean = 31.66 seconds
Median = 32 seconds
Mode = 33 seconds

b All three measures in part (a) are approximately equal. The distribution of times is therefore approximately symmetrical with a single mode of 33. Half the times are less than 32 seconds.

5.21 **a** Mean = 30.53; Median = 31

b The mean training time is 30.53. Half the sample trained for less than 31 hours.

5.22 **a** Mean = $39329; Median = $39461

c Mean < Median, the salaries distribution may be slightly skewed to the left.

5.23 **a** Mean = 32.91; Median = 32; Mode = 32

b The mean speed is 32.91kmph. Half the sample traveled slower than 32kmph and half traveled faster. The mode is 32.

5.24 **a** No. A standard deviation cannot be negative, because it is defined as the positive square root of the variance.

b Yes, a standard deviation is larger than its corresponding variance when the variance is between 0 and 1.

c Yes, when every value of a data set is the same, the variance and standard deviation will be zero.

d Yes. Because CV = SD/Mean. SD is non negative but mean can be negative.

e Yes. Because SD could be zero and hence could CV.

5.25 Range = largest value – smallest value = 24 – 5 = 19

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{5 + 7 + \ldots + 24}{10} = 15$$

$$s^2 = \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{9} = \frac{(5-15)^2 + (7-15)^2 + \ldots + (24-15)^2}{9} = 35.56$$

$$s = \sqrt{35.36} = 5.96$$

$$cv = \frac{s}{\bar{x}} \times 100 = 39.73\%$$

a Mean = 14, variance = 28.75, s = 5.36, cv = 38.3%
b Mean = 17, variance = 35.56, s = 5.96, cv = 35.1%
c Mean = 45, variance = 320.04, s = 17.89, cv = 39.8%

5.26 a $\bar{x} = \frac{14+7+8+11+5}{5} = \frac{45}{5} = 9$

$$s^2 = \frac{(14-9)^2 + (7-9)^2 + (8-9)^2 + (11-9)^2 + (5-9)^2}{4} = 12.5$$

$$s = \sqrt{12.5} = 3.54$$

$$cv = \frac{s}{\bar{x}} \times 100 = 39.3\%$$

b $\bar{x} = 0$

$$s^2 = \frac{(-3)^2 + (-2)^2 + (-1)^2 + (-0)^2 + (1)^2 + (2)^2 + (3)^2}{6} = 4.67$$

$$s = \sqrt{4.67} = 2.16$$

cv not applicable as $\bar{x} = 0$

c $\bar{x} = \frac{4+4+8+8}{4} = \frac{24}{4} = 6$

$$s^2 = \frac{(4-6)^2 + (4-6)^2 + (8-6)^2 + (8-6)^2}{3} = 5.33$$

$$s = \sqrt{5.33} = 2.31$$

$$cv = \frac{s}{\bar{x}} \times 100 = 38.5\%$$

d $\bar{x} = 5$

$s^2 = s = 0$ (by inspection)

$cv = 0$

5.27 **a** Sample 3 has the largest variability, with values ranging from from 16 to 49. Sample 2 has the least variability, with all values close to 30.

b

	s^2	s
Sample 1	46.5	6.82

Sample 2	6.5	2.55
Sample 3	174.5	13.21

c $\Sigma (x_i - \bar{x})$ equals zero for every sample.

5.28 Range = 6 – 0 = 6 hours

$$\bar{x} = \frac{2+5+6+1+4+0+3}{7} = \frac{21}{7} = 3 \text{ hours}$$

$$s^2 = \frac{(2-3)^2 + (5-3)^2 + \ldots + (0-3)^2 + (3-3)^2}{6} = 4.67 \text{ (hours)}^2$$

$$s = \sqrt{4.67} = 2.16 \text{ hours}$$

$$cv = \frac{s}{\bar{x}} \times 100 = 72\%$$

5.29

	Male	Female
Mean	5.9	5.5
Variance	2.85	0.93
SD	1.69	0.96
CV	28.59	17.47

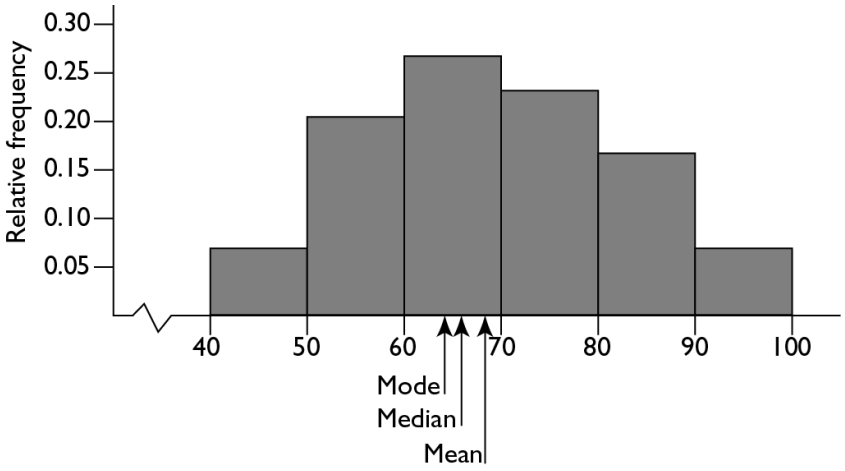
The variability in rate of unemployment across states and territories appears higher for male than female.

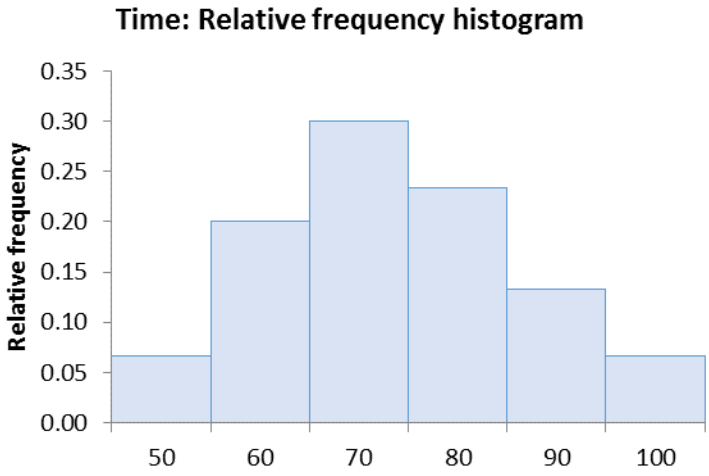
5.30 a $\bar{x} = \frac{75+79+\ldots+59+55}{30} = \frac{2072}{30} = 69.07$

The median is 66.5; the midpoint between the two middle values 66 and 67.

b The modal class of the frequency distribution constructed in Exercise 4.7 is '60 up to 70'. The mode is therefore 65.

c





d $\Sigma x_i^2 = 75^2 + 79^2 + \ldots + 59^2 + 55^2 = 147950$

From part (a), we know that $\Sigma x_i = 2072$

$$s^2 = \frac{1}{n-1}\left(\sum_{i-1}^{10} x_i^2 - \frac{\left(\sum_{i-1}^{30} x_i\right)^2}{n}\right)$$

$$= \frac{1}{29}\left(147\,950 - \frac{(2072)^2}{30}\right) = 167.03$$

5.31 Mean = 6.49, SD = 4.09

5.32 **a** Mean = 1.06, SD = 0.60
b Median = 0.97
c Mean = 1.00, SD = 0.14, median = 0.97

5.33 According to the Empirical Rule:
a Approximately (0.68)(1000) = 680 workers receive wages in (23 400, 27 800) = $\bar{x} \pm s$.
b Approximately (0.95)(1000) = 950 workers receive wages in (21 200, 30 000) = $\bar{x} \pm 2s$.
c Virtually all of the workers receive wages in (19 000,32 200) = $\bar{x} \pm 3s$.

5.34 **a** About 68%
b About 95%
c About 99.7%

5.35 **a** Nothing
b At least 75% lie between 60 and 180.
c At least 88.9% lie between 30 and 210.

5.36 a $\mu = \dfrac{\sum_{i-1}^{15} x_i}{15} = \dfrac{3+5+\ldots+18+14}{15} = 5.67\%$

$$\sigma^2 = \frac{\sum_{i-1}^{15}(x_i-\mu)^2}{15} = \frac{(3-5.67)^2+(5-5.67)^2+\ldots+(18-5.67)^2+(14-5.67)^2}{15} = 277.16$$

$\sigma = \sqrt{277.16} = 16.65\%$

b Range = 70%
Median = 5%

5.37 a $\bar{x} = \dfrac{3+7+4+\ldots+3.5+3}{11} = \dfrac{71}{11} = 6.45\text{meters}$

Arranging the measurements in ascending order, we obtain

2.5, 3, 3, 3, 3.5, 4, 5, 7, 15, 20

The median is 4 metres, which is the middle value.
The mode is 3 metres, since that value occurs most frequently.

Variance: $s^2 = \dfrac{1}{(n-1)}\left[\sum x_i^2 - n\bar{x}^2\right] = \dfrac{1}{10}\left[(3^2+7^2+\ldots+3^3) - 11(6.45)^2\right] = 32.72$

SD: $s = \sqrt{32.72} = 5.72$

b Range = 20 - 2.5 = 17.5

Approximate standard deviation $s \approx \dfrac{\text{Range}}{4} = \dfrac{17.5}{4} = 4.375$

5.38 a $\bar{x} = 2.55$
$s^2 = 0.03945$
$s = 0.20$

b Using the range approximation, we obtain

$s \approx \dfrac{\text{Range}}{4} = \dfrac{2.9-2.1}{4} = \dfrac{0.8}{4} = 0.2$

The approximate value is the same as the actual standard deviation in part (a).

We have assumed that the sample of 20 measurements has a mound-shaped distribution.

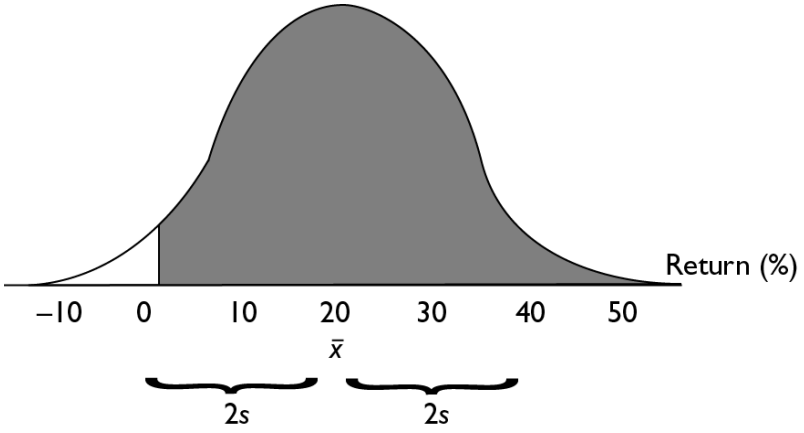
5.39 $\bar{x} = 20\%$ and $s = 10\%$

a Approximately 68% of the stocks had a return in the interval $(10, 30) = \bar{x} \pm s$. Virtually 100% of the stocks had a return in the interval $(-10, 50) = \bar{x} \pm 3s$.

b Approximately 68% of the stocks had a return in the interval $(10, 30) = \bar{x} \pm s$. The

remaining 32% (approximately) of the stocks had a return that was outside this interval.

c Approximately 95% of the stocks had a return in the interval $(0, 40) = \bar{x} \pm 2s$. Therefore, 5% of the stocks had a return outside this interval. Because the distribution is symmetrical, 2.5% of the stocks had a return less than 0, and 2.5% of the stocks had a return greater than 40. Thus, the proportion of stocks that had a positive return was $(1 - 0.025) = 0.975$.



5.40 a

Year	Return
1	13.70 [$= 0.5(12.3) + 0.5(15.1)$]
2	−1.00 [$= 0.5(-2.2) + 0.5(0.2)$]
3	17.15
4	8.25
5	34.20
6	37.60
7	24.80
8	11.45
9	4.40
10	24.45

b $\bar{x}_P = \dfrac{13.70 - 1.0 + 17.15 + \ldots + 24.45}{10} = 17.5\%$

c $s_P^2 = \dfrac{\sum_{i-1}^{10} (x_i - \bar{x}_P)^2}{9}$

$s_P^2 = \dfrac{\sum_{i-1}^{10} (x_i - \bar{x}_P)^2}{9} = \dfrac{(13.7 - 17.5)^2 + (-1.00 - 17.5)^2 + \ldots + (24.45 - 17.5)^2}{9} = 159.45$

$s_P = \sqrt{159.45} = 12.63\%$

d $cv = 72.2\%$

e

Mean return	Variance of returns	Coefficient of variation%

$\bar{x}_A = 20\%$	$s_A^2 = 280.3$	83.72
$\bar{x}_P = 17.5\%$	$s_P^2 = 159.5$	72.16
$\bar{x}_B = 15\%$	$s_B^2 = 99.4$	66.50

Highest average return: Fund A; Lowest risk: Fund B.
[Variance or coefficient of variation of returns will be used as the measure of the riskiness of an investment. Ranking the three investments in order of both decreasing coefficient of variation and decreasing riskiness, we obtain: Fund *A* (highest); Fund *B* (lowest).]

5.41 a, b & c

	Dividend yield		
	Top 1-40	Top 41-80	Top 81-120
Mean	5.66	6.73	7.04
Median	5.09	5.92	5.11
Range	23.40	14.50	22.45
SD	4.11	3.63	5.51
cv	0.73	0.54	0.78

d The average yield is highest among the Top 81-120 companies, but Top 41-80 has higher variability than the yields of Top 1-40 and Top 41-80. Based on median and CV, Top 41-80 is the most stable than the other two groups.

5.42 a & b

	Mean	**Median**	**Range**	**SD**	**cv**
1-year (4-star)	9.87	9.82	5.11	1.82	18.39
3-year (4-star)	5.77	6.17	2.72	1.10	19.10
1-year (3-star)	10.16	10.44	2.42	0.93	9.17
3-year (3-star)	5.69	5.86	1.52	0.63	11.11

c Among the 1-year investments, (i) since the 3-star rated managed funds have a lower cv value (9.17%) than the 4-star rated managed funds (18.39%), the 3-star rated managed funds have a lower risk than the 4-star rated managed funds; and (ii) the 3-star rated managed funds also have higher average performance. Therefore, among the 1-year investments, the 3-star rated funds is preferable.

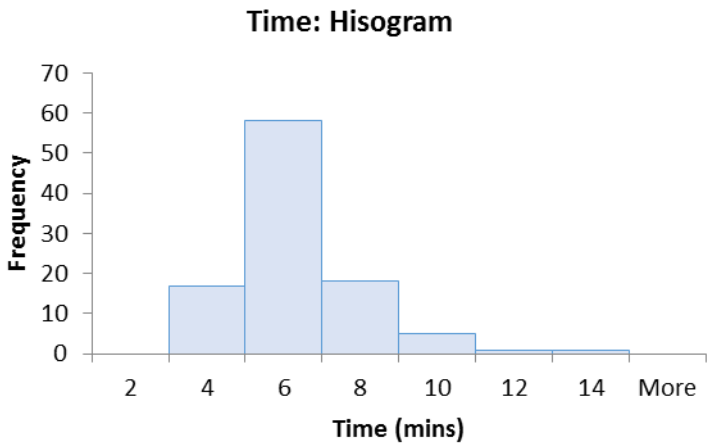
d Among the 4-star rated investments, 1-year has higher (mean) return and lower (cv) risk than 3-year investment. Therefore, 1-year 4-star rated is preferable.

e Among the 3-star rated investments, 1-year also has higher (mean) return and lower (CV) risk. Therefore, 1-year 3-star rated is preferable.

5.43 a Variance = 2.9571(minutes)^2

Standard deviation = 1.7196 minutes

b



5.44 Range = $411
Variance = 9292.41 ($)^2
SD = $96.40
cv = 40.5%

5.45 a

	Mean	**Median**
Morning	52.14	47
Afternoon	52.68	50
Evening	53.28	48

b

	Range	s^2	s
Morning	128	422.72	20.56
Afternoon	109	382.49	19.56
Evening	97	356.85	18.89

c The mean of viewing time is smallest for the morning and largest for the evening**.** The variance of viewing time is smallest for the evening and largest for the morning.

d In the morning, more visitors could be allowed in at a time than in the afternoon or the evening. The fewest could be allowed in at a time in the evening.

5.46 a

	Mean	s
10:00am – 11:00am	102.22	16.07
11:00am – 12:00pm	70.26	10.58
12:00pm – 1:00pm	177.93	18.24
1:00pm – 2:00pm	65.87	9.37
2:00pm – 3:00pm	147.92	14.63

b The noon hour (12:00pm – 1:00pm) is the busiest, followed by the (2:00pm – 3:00pm) and

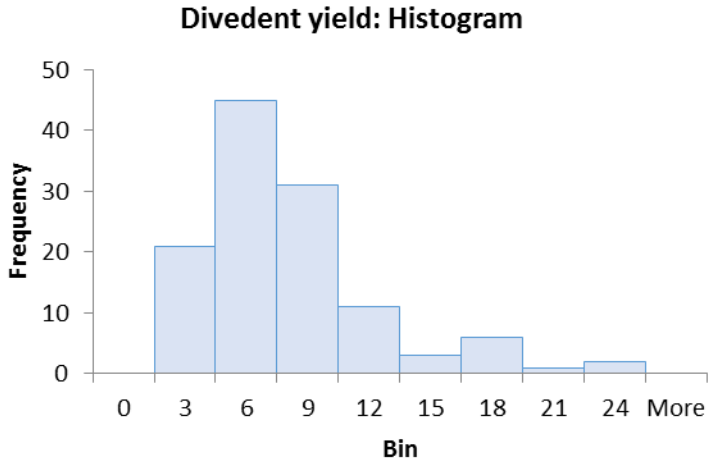
(10:00am – 11:00am) periods. Staff lunch breaks and coffee breaks should be scheduled with this in mind.

5.47 a Standard deviation = 4.490

b Range = 23.5

Approximate standard deviation = $\frac{23.5}{4} = 5.875$

c The actual (4.49) and the approximate (5.88) values of the standard deviation are reasonably close, but not very close. The reason could be the requirement for the approximation that the data be bell-shaped, as can be seen from the histogram of the dividend yields of the top 120 companies below, the data are not symmetric or bell-shaped.



5.48 $s^2 = 40.73$ kph, and $s = 6.38$kph. Based on Chebyshev's theorem, at least 75% of the speeds lie within 12.76 kmph ($2s$) of the mean; at least 88.9% of the speeds lie within 19.15kmph ($3s$) of the mean.

5.49 a

Person	Variance	Standard deviation
1	40.22	6.34
2	14.81	3.85
3	3.63	1.91

b Person 3 is the most consistent.

5.50 $s^2 = 0.0858\text{cm}^2$, and $s = 0.2929$cm. Based on Chebyshev's theorem, at least 75% of the lengths lie within 0.5857cm ($2s$) of the mean; at least 88.9% of the rods will lie within 0.8786cm ($3s$) of the mean.

5.51 $\bar{x} = 175.73$; $s = 62.06$. Based on Chebyshev's theorem, at least 75% of the withdrawals lie between \$51.61 and \$299.86; at least 88.9% of the withdrawals lie between 0 and \$361.92.

5.52 a $s = 15.01$

b In approximately 68% of the hours the number of arrivals falls between 83 (rounded from 83.04) and 113; on approximately 95% of the hours the number of arrivals fall between 68 and 128; on approximately 99.7% of the hours the number of arrivals fall between 53 and 143.

5.53 $n = 15$. Data in ascending order: 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 7, 7, 8, 9, 10

First quartile (Q_1): $L_{25} = (15+1)\frac{25}{100} = (16)(0.25) = 4$; Q_1 = the fourth number = 3.

Second quartile (Q_2): $L_{50} = (15+1)\frac{50}{100} = (16)(0.5) = 8$; Q_2 = the eight number = 5.

Third quartile (Q_3): $L_{75} = (15+1)\frac{75}{100} = (16)(0.75) = 12$; Q_3 = the twelfth number = 7.

5.54 $n = 10$. Data in ascending order: 15, 20, 22, 23, 24, 26, 29, 30, 31, 31

30th percentile P_{30}: $L_{30} = (10+1)\frac{30}{100} = (11)(0.30) = 3.3$; $P_{30} = 22 + 0.3(23 - 22) = 22.3$.

80th percentile P_{80}: $L_{80} = (10+1)\frac{80}{100} = (11)(0.80) = 8.8$; $P_{80} = 30 + 0.3(31 - 30) = 30.8$.

5.55 $n = 10$. Data in ascending order: 39, 43, 51, 52, 60, 61, 64, 71, 73, 88

20th percentile, P_{20}: $L_{20} = (10+1)\frac{20}{100} = (11)(0.20) = 2.2$; $P_{20} = 43 + 0.2(51 - 43) = 44.6$.

40th percentile, P_{40}: $L_{40} = (10+1)\frac{40}{100} = (11)(0.40) = 4.4$; $P_{40} = 52 + 0.4(60 - 52)$ 55.2.

5.56 $n = 13$; 10.0, 10.5, 12.2, 13.9, 13.9, 14.1, 14.7, 14.7, 15.1, 15.3, 15.9, 17.7, 18.5

First quartile (Q_1): $L_{25} = (13+1)\frac{25}{100} = (14)(0.25) = 3.5$; Q_1 = 12.2+0.5(13.9-12.2) = 13.05.

Second quartile (Q_2): $L_{50} = (13+1)\frac{50}{100} = (14)(0.5) = 7$; Q_2 = 14.7.

Third quartile (Q_3): $L_{75} = (13+1)\frac{75}{100} = (14)(0.75) = 10.5$; Q_3 =15.3+0.5(15.9-15.3) = 15.6.

5.57 Third decile, P_{30}: $L_{30} = (15+1)\frac{30}{100} = (16)(0.30) = 4.8$; $P_{30} = 5 + 0.8(7 - 5) = 6.6$.

Sixth decile, P_{60}: $L_{60} = (15+1)\frac{60}{100} = (16)(0.60) = 9.6$; $P_{60} = 17 + 0.6(18 - 17) = 17.6$.

5.58 Interquartile range IQR = Q_3 - Q_1 = 7 – 3 = 4

5.59 Interquartile range IQR = Q_3 - Q_1 = 15.6 – 13.05 = 2.55

5.60 n = 10, Data in ascending order: 2, 5, 6, 8, 9, 10, 11, 14, 18, 21.

$L_{25} = (10+1)\frac{25}{100} = (11)(0.25) = 2.75;$ First quartile Q_1 = 5 + 0.75(6-5) = 5.75.

$L_{75} = (10+1)\frac{75}{100} = (11)(0.75) = 8.25;$ Third quartile Q_3 = 14 + 0.25(18-14) = 15.

Interquartile range IQR = Q_3 - Q_1 = 15 – 5.75 = 9.25

5.61 Arranging the 15 measurements in ascending order, we obtain:

−20, −18, −10, −5, −2, 0, 3, 5, 6, 10, 14, 14, 18, 20, 50

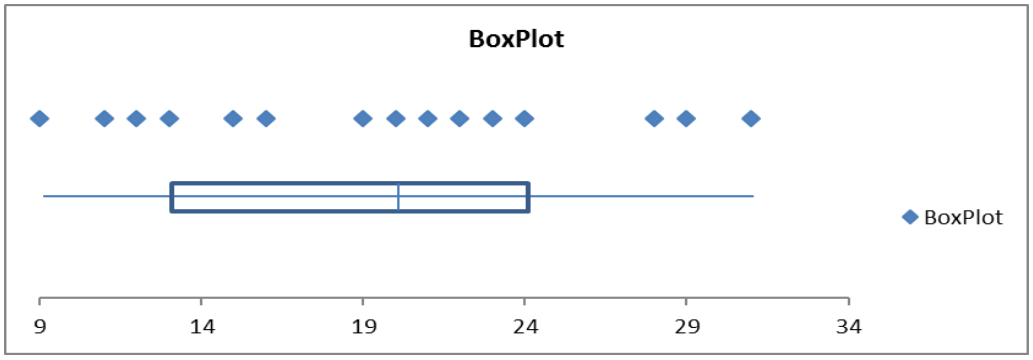
The 20th and 60th percentile:

$$L_{20} = (15+1)\frac{20}{100} = 3.2; \quad P_{20} = -10 + 0.2(5) = -9$$

$$L_{60} = (15+1)\frac{60}{100} = 9.6; \quad P_{60} = 6 + 0.6(10-6) = 8.4$$

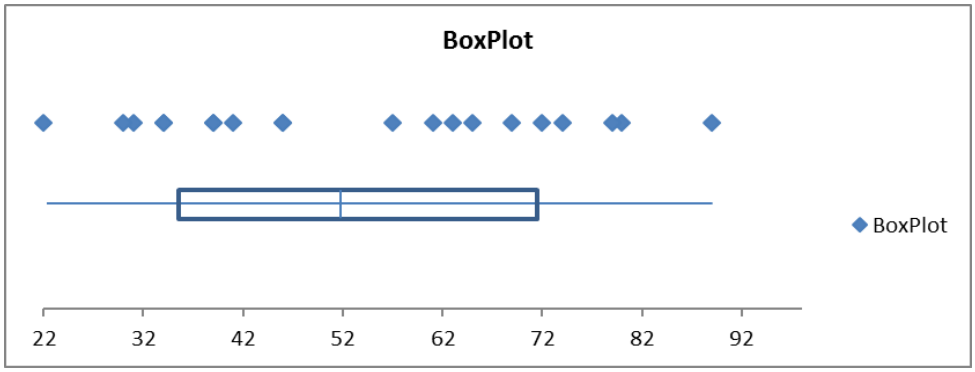
5.62 $S = 9,\ Q_1 = 13,\ Q_2 = 20,\ Q_3 = 24,\ L = 31$

Smallest = 9
Q1 = 13
Median = 20
Q3 = 24
Largest = 31
IQR = 11
Outliers:



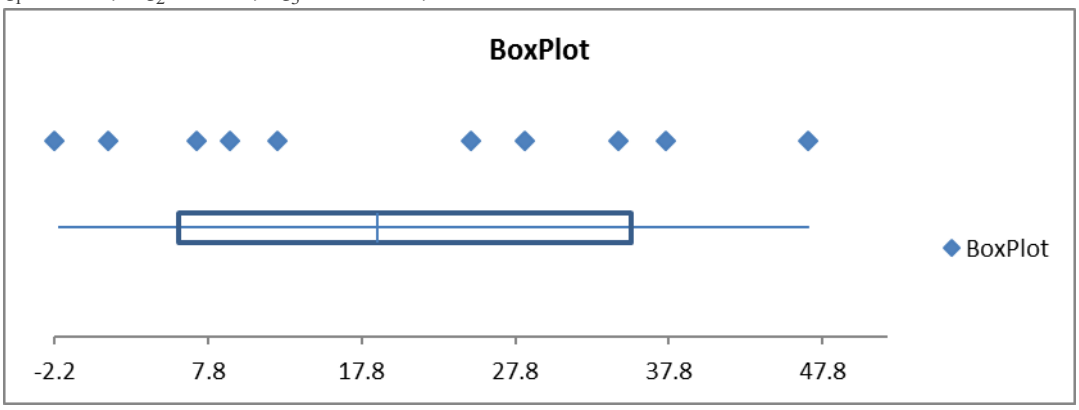
5.63 $S = 22,\ Q_1 = 35.25,\ Q_2 = 51.5,\ Q_3 = 71.25, L = 89$

Smallest = 22
Q1 = 35.25
Median = 51.5
Q3 = 71.25
Largest = 89
IQR = 36
Outliers:



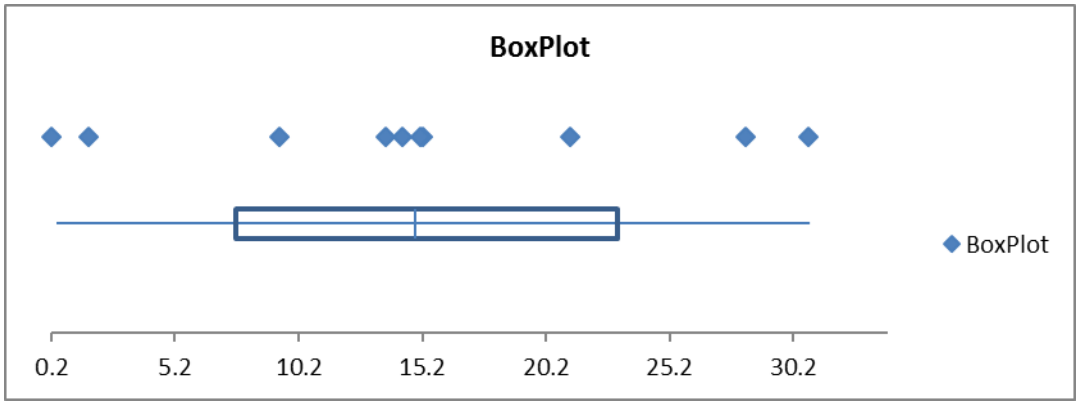
5.64 a S = -2.2, $Q_1 = 5.65,\ Q_2 = 18.6,\ Q_3 = 35.275,\ L = 46.9$

Trust A
Smallest = -2.2
Q1 = 5.65
Median = 18.6
Q3 = 35.275
Largest = 46.9
IQR = 29.625
Outliers:



S = 0.2, $Q_1 = 7.475,\ Q_2 = 14.75,\ Q_3 = 22.975,\ L = 38$

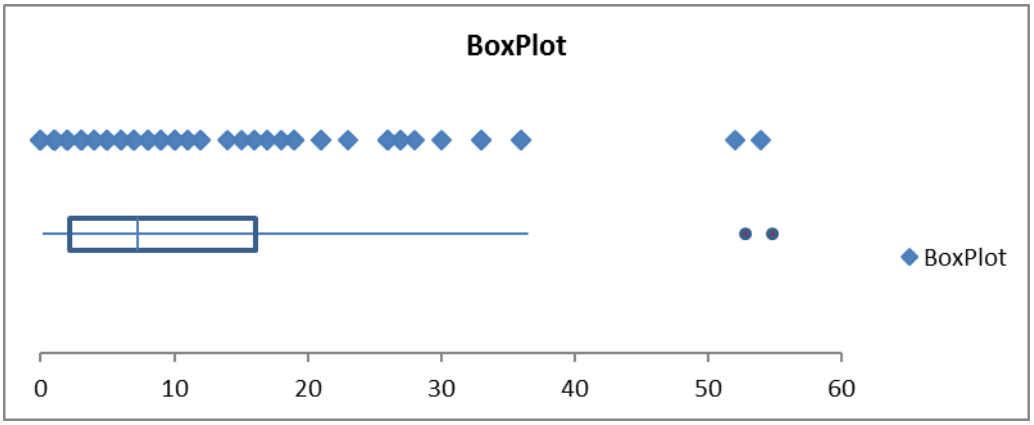
Trust B
Smallest = 0.2
Q1 = 7.475
Median = 14.75
Q3 = 22.975
Largest = 30.8
IQR = 15.5
Outliers:



b The median return for Fund A exceeds the median return for Fund *B*. The returns for Fund *A* are more variable than for Fund *B*, with Fund A having an IQR of 29.6% and a range of 49.1%, compared with an IQR of 15.5% and a range of 30.6% for Fund *B* Neither fund has any outliers.

5.65 a

Smallest = 0
Q1 = 2
Median = 7
Q3 = 15.75
Largest = 54
IQR = 13.75
Outliers: 54, 52,



b $L_{25} = (100+1)\times\frac{25}{100} = 25.25,\ Q_1 = 2$

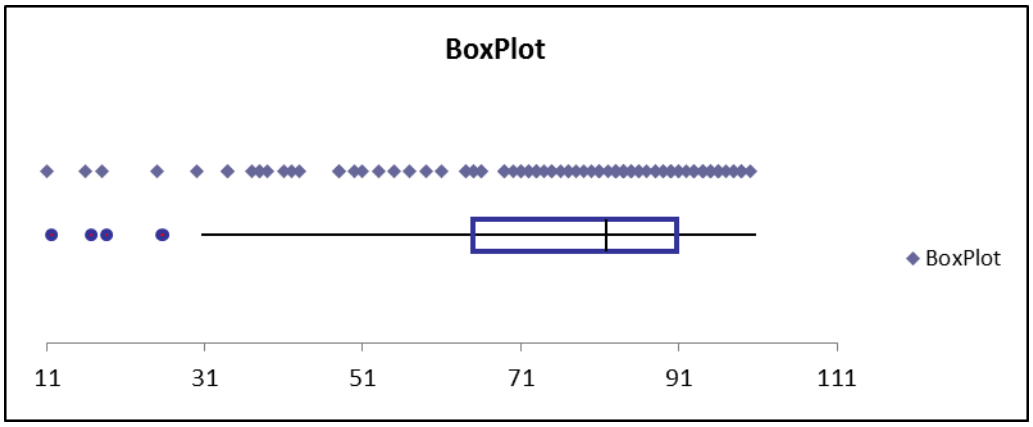
$L_{50} = (100+1)\times\frac{50}{100} = 50.5,\quad Q_2 = 7$

$L_{75} = (100+1)\times\frac{75}{100} = 75.75,\ Q_3 = 15.75$

c Although half the observations are less than 7, they range from 0 to 54. The distribution is highly skewed to the right, with two outliers: 52 and 54.

5.66 a

Smallest = 11
Q1 = 64.25
Median = 81
Q3 = 90
Largest = 100
IQR = 25.75
Outliers: 25, 18, 16, 11,



b $L_{25} = (100+1)\times\frac{25}{100} = 25.25,\ Q_1 = 64.25$

$L_{50} = (100+1)\times\frac{50}{100} = 50.5,\ Q_2 = 81$

$L_{75} = (100+1)\times\frac{75}{100} = 75.75,\ Q_3 = 90$

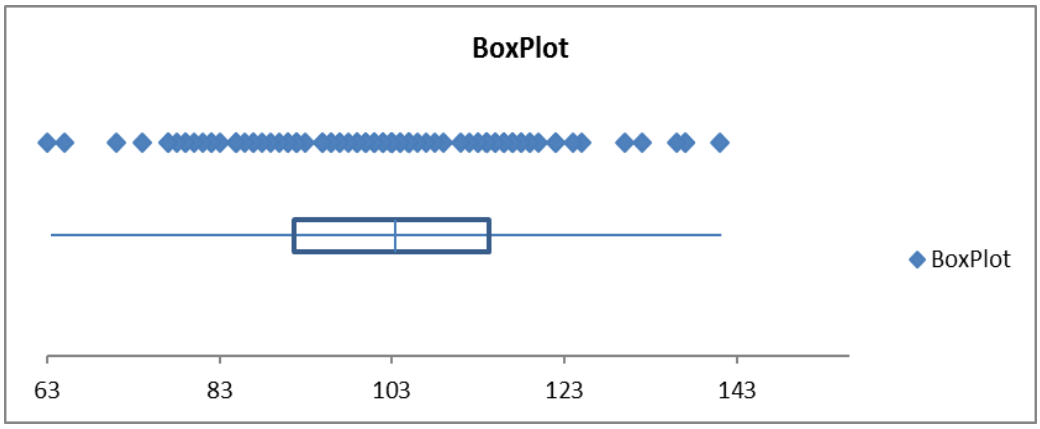
c There are 4 outliers: 11, 16, 18, 25

d Although the marks range from 11 to 100, half of them are over 81. The distribution is

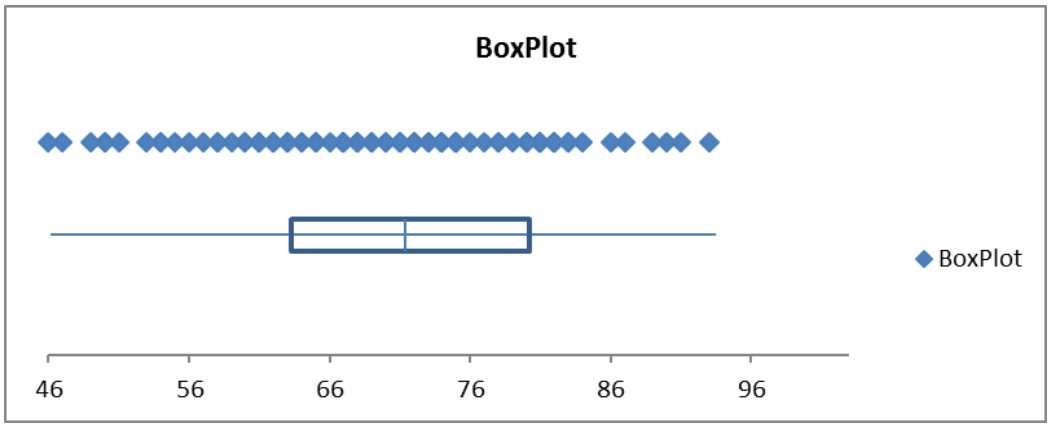
highly skewed to the left. The mean mark of about 74 (computed in Example 3.8) has been pulled much below the median by the four (small) outliers.

5.67 a

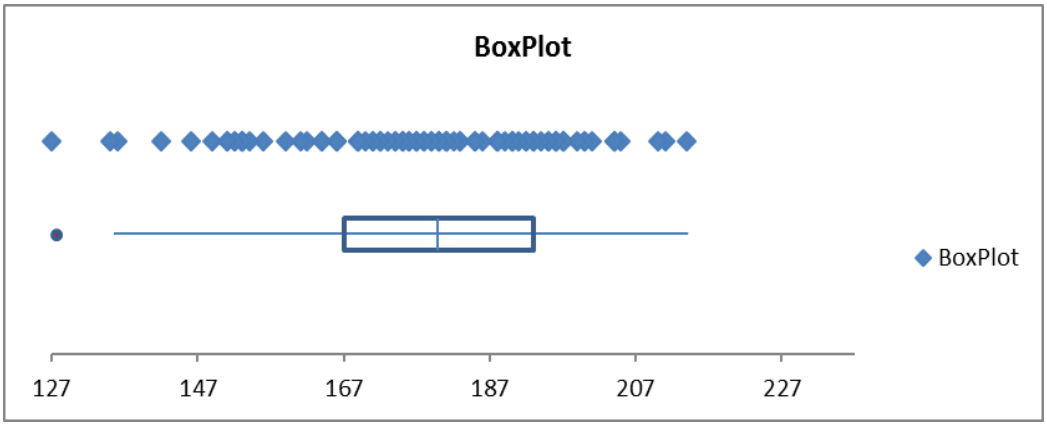
10am-11am
Smallest = 63
Q1 = 91.25
Median = 103
Q3 = 114
Largest = 141
IQR = 22.75
Outliers:



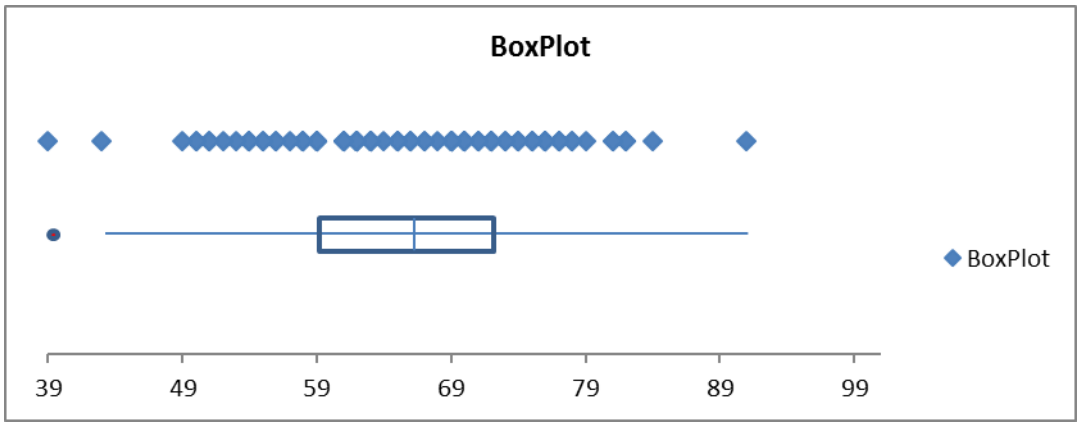
11am-12pm
Smallest = 46
Q1 = 63
Median = 71
Q3 = 79.75
Largest = 93
IQR = 16.75
Outliers:



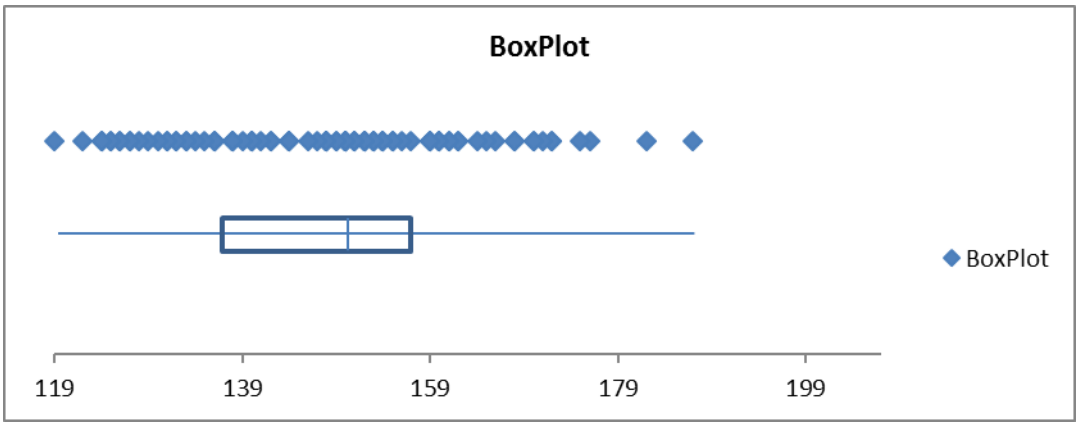
12pm-1pm
Smallest = 127
Q1 = 166.75
Median = 179.5
Q3 = 192.75
Largest = 214
IQR = 26
Outliers: 127,



1pm-2pm
Smallest = 39
Q1 = 59
Median = 66
Q3 = 72
Largest = 91
IQR = 13
Outliers: 39,



2pm-3pm
Smallest = 119
Q1 = 136.5
Median = 150
Q3 = 156.75
Largest = 187
IQR = 20.25
Outliers:



b 10:00 – 11:00 am

The median number of customers was 103, with numbers falling between 91 and 114 about half the time.

11:00 am – 12:00pm

The median number of customers was 71, with numbers falling between 63 and 80 about half the time.

12:00 – 1:00 pm

The median number of customers was 179.5, with numbers falling between 167 and 193 about half the time.

1:00 – 2:00 pm

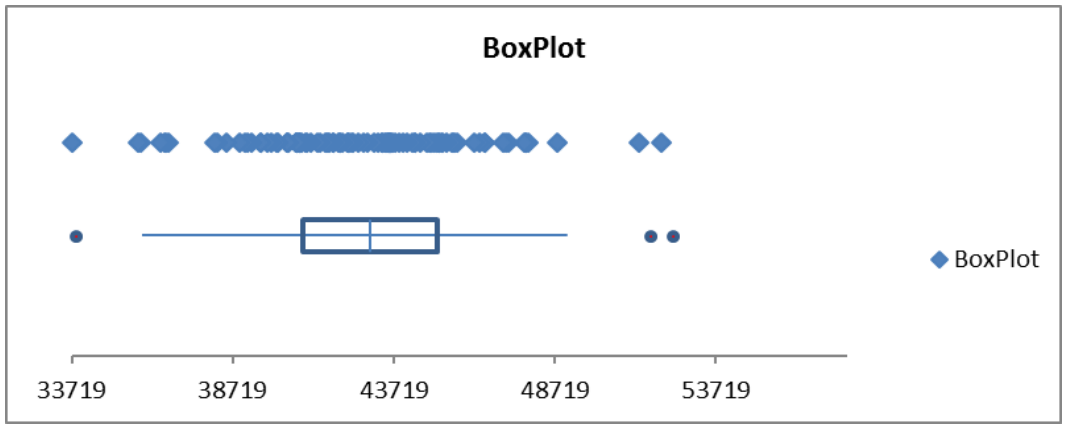
The median number of customers was 66, with numbers falling between 59 and 72 about half the time.

2:00 – 3:00 pm

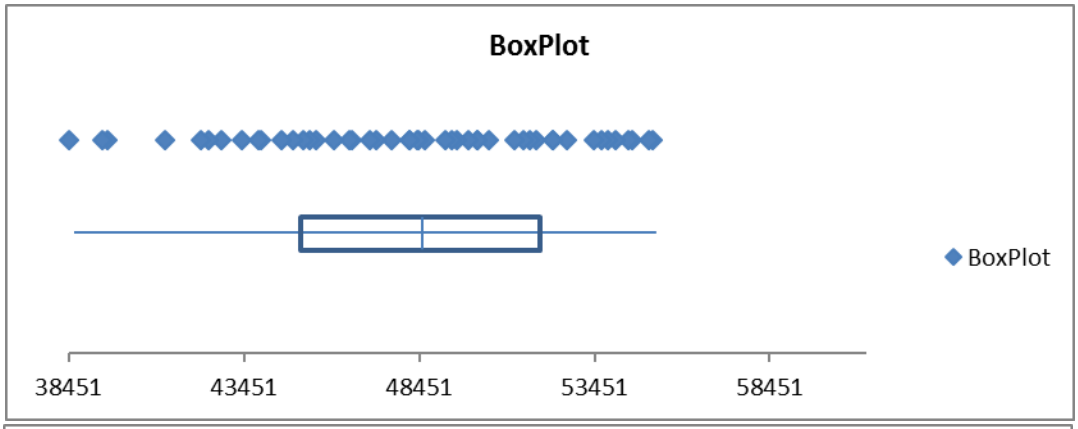
The median number of customers was 150, with numbers falling between 137 and 157 about half the time.

5.68 a

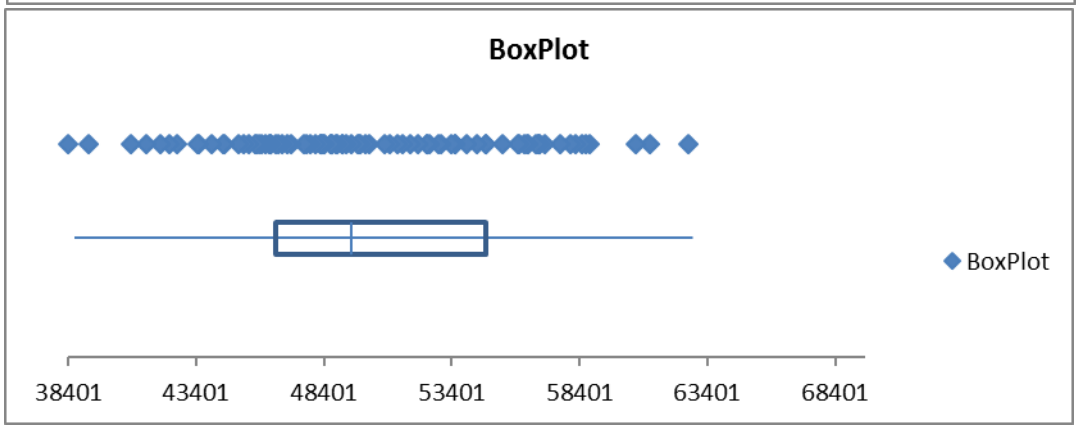
BA
Smallest = 33719
Q1 = 40730
Median = 42765
Q3 = 44835.5
Largest = 52025
IQR = 4105.5
Outliers: 52025, 51345, 33719,



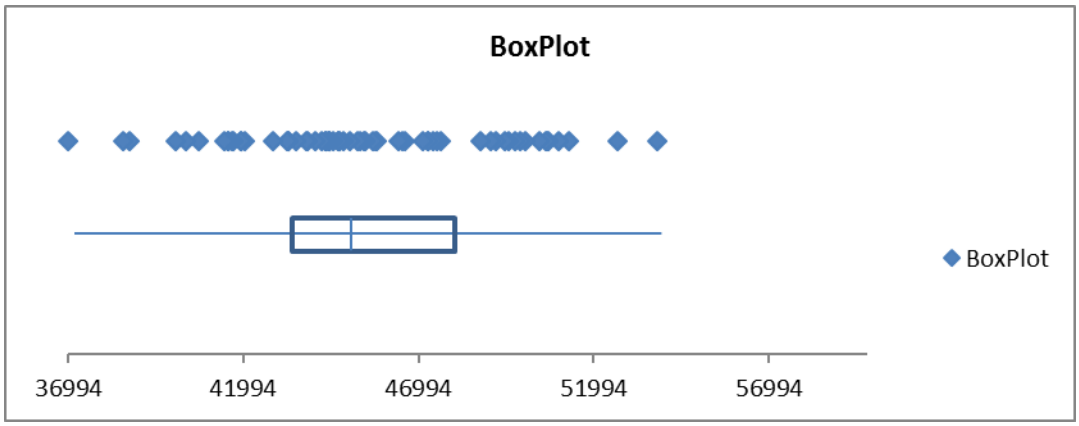
BSc
Smallest = 38451
Q1 = 44927
Median = 48396.5
Q3 = 51745.25
Largest = 55105
IQR = 6818.25
Outliers:



BBA
Smallest = 38401
Q1 = 46316
Median = 49284
Q3 = 54551
Largest = 62639
IQR = 8235
Outliers:

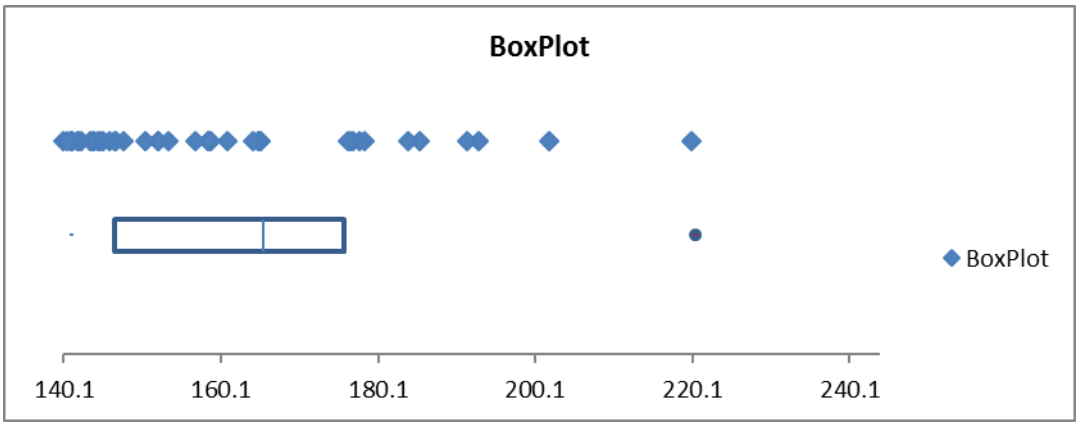


Other
Smallest = 36994
Q1 = 43253.5
Median = 44950.5
Q3 = 47905.25
Largest = 53812
IQR = 4651.75
Outliers:



5.69 a

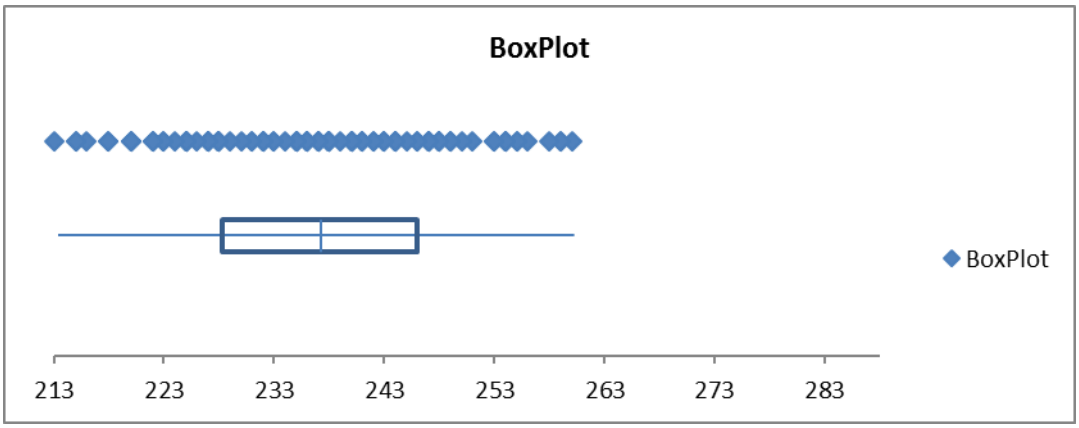
Time
Smallest = 140.1
Q1 = 145.11
Median = 164.17
Q3 = 175.18
Largest = 220.96
IQR = 30.07
Outliers: 220.96,



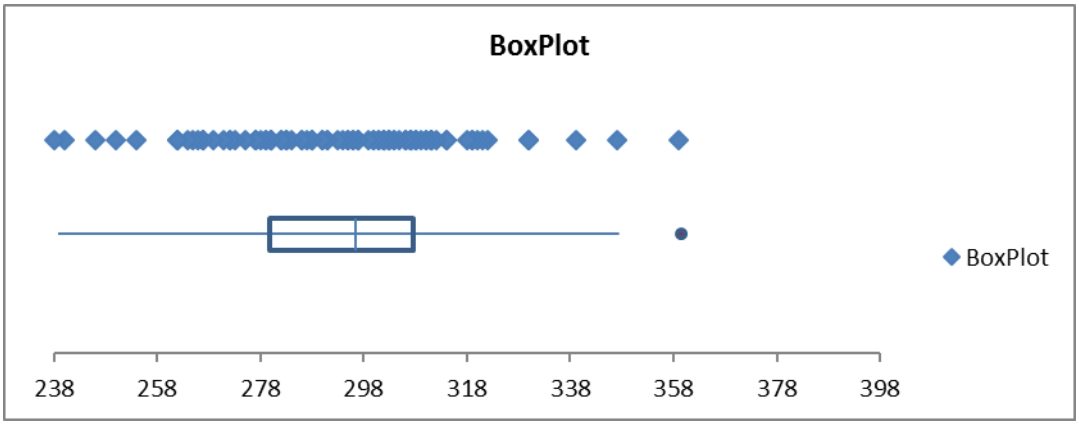
b The quartiles are 145.11, 164.17, and 175.18
c There is one outlier, 220.96.
d The data are positively skewed. One-quarter of the times are below 145.11 and one-quarter are above 175.18. 50% of the running times lie within 30.07mins

5.70 a

Private
Smallest = 213
Q1 = 228
Median = 237
Q3 = 245.75
Largest = 260
IQR = 17.75
Outliers:



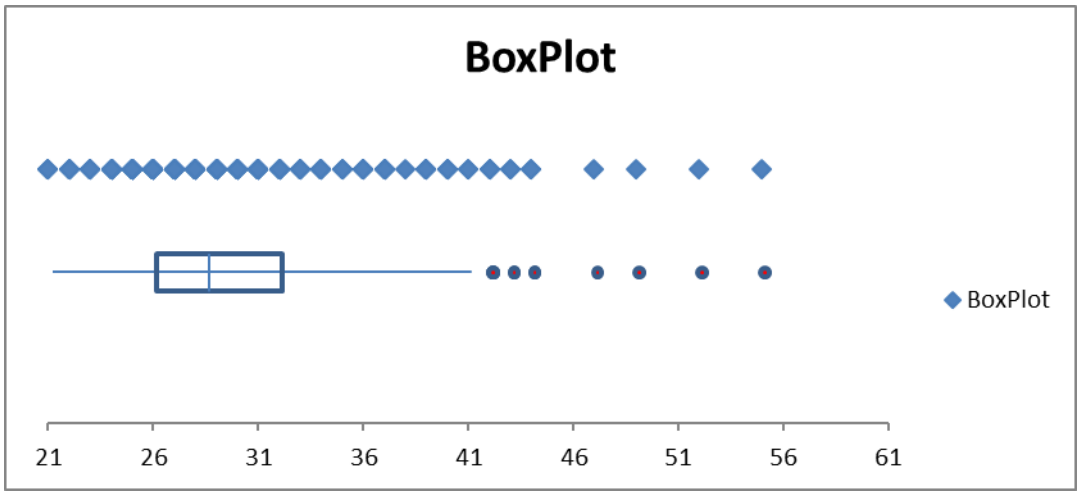
Public
Smallest = 238
Q1 = 279
Median = 296
Q3 = 307
Largest = 359
IQR = 28
Outliers: 359,



b The times for public course is skewed to the right. The amount of time taken to complete rounds on the public course is larger and more variable than those played on private courses.

5.71 a The quartiles are 26, 28.5, and 32

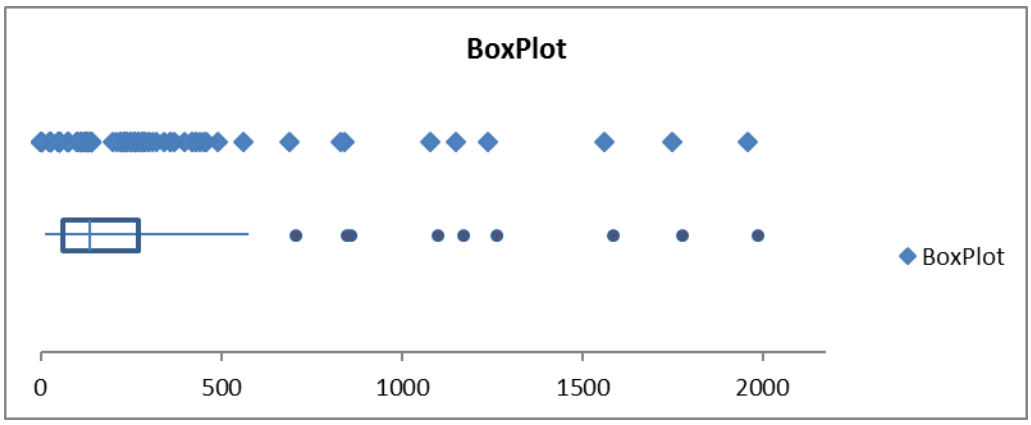
Time
Smallest = 21
Q1 = 26
Median = 28.5
Q3 = 32
Largest = 55
IQR = 6
Outliers: 55, 52, 49, 47, 44, 44, 43, 43, 42, 42, 42,



b The times are positively skewed.

5.72

Amount
Smallest = 0
Q1 = 50
Median = 125
Q3 = 260
Largest = 1960
IQR = 210
Outliers: 1960, 1750, 1560, 1240, 1150, 1080, 840, 830, 690,



The amounts are positively skewed.

5.73 a

Class	f_i	m_i	$f_i m_i$	$f_i m_i^2$
–20 up to –10	8	–15	–120	1 800
–10 up to 0	21	–5	–105	525
0 up to 10	43	5	215	1 075
10 up to 20	48	15	720	10 800
20 up to 30	25	25	625	15 625
30 up to 40	15	35	525	18 375
Total	160		1 860	48 200

$$\bar{x} \approx \frac{\sum_{i=1}^{n} f_i m_i}{160} = \frac{1860}{160} = 11.625$$

$$s^2 \approx \frac{1}{n-1}\left(\sum_{i=1}^{n} f_i m_i^2 - \frac{\left(\sum_{i=1}^{n} f_i m_i \right)^2}{n} \right) = \frac{1}{160-1}\left(48200 - \frac{(1860)^2}{160} \right) = \frac{26577.5}{159} = 167.15$$

$$s \approx \sqrt{167.15} = 12.93$$

b

$$s \cong \frac{\text{Range}}{4} = \frac{40-(-20)}{4} = 15$$

The range approximation of *s* indicates that the group value computed for *s* in part (a) is at least in the ballpark.

5.74 a

Earnings	f_i	m_i	$f_i m_i$	$f_i m_i^2$
8 up to 10	11	9	99	891
l0 up to 12	17	11	187	2 057
12 up to 14	32	13	416	5 408
14 up to 16	27	15	405	6 075
16 up to 18	13	17	221	3 757
Total	100		1 328	18 188

$$\bar{x} \cong \frac{\sum_{i=1}^{s} f_i m_i}{100} = \frac{1328}{100} = \$13.28$$

$$s^2 @ \frac{1}{n-1}\left(\sum_{i=1}^{s} f_i m_i^2 - \frac{\left(\sum_{i=1}^{s} f_i m_i \right)^2}{n} \right) = \frac{1}{99}\left(18\,188 - \frac{(1328)^2}{100} \right)$$

$$= 5.58$$

$$s @ \sqrt{5.58} = \$2.36$$

b We have not computed $\bar{x}$ and *s* using the actual 100 earnings, but have used only the midpoints of the five classes into which the earnings were grouped. We must assume, at the least, that the midpoint of each class approximates the mean of the earnings in that class. The approximations of $\bar{x}$ and *s* obtained from the formulas are only as good as this assumption.

5.75

Fuel consumption (km/litre)	f_i	m_i	$f_i m_i$	$f_i m_i^2$
9 up to 10.5	9	9.75	87.75	855.5625
10.5 up to 12.0	13	11.25	146.25	1 645.3125
12.0 up to 13.5	24	12.75	306.00	3 901.5000
13.5 up to 15.0	38	14.25	541.50	7 716.3750
15.0 up to 16.5	16	15.75	252.00	3 969.0000
Total	100		1 333.50	18 087.7500

$$\bar{x} = \frac{\sum_{i=1}^{s} f_i m_i}{100} = \frac{1333.50}{100} = 13.34 \text{ kms/litre}$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{s} f_i m_i^2 - \frac{\left(\sum_{i=1}^{s} f_i m_i \right)^2}{n} \right)$$

$$= \frac{1}{99}\left(18087.75 - \frac{1333.50)^2}{100} \right) = 3.086$$

$$s = \sqrt{3.086} = 1.76 \text{ kms/litre}$$

5.76 a 1990:

m_i	f_i	$f_i m_i$	$f_i m_i^2$
19	31761	603459	11465721
29	58060	1683740	48828460
39	16170	630630	24594570
49	6371	312179	15296771
59	2865	169035	9973065
69	1314	90666	6255954
84	417	35028	2942352
Sum	116958	3524737	1.19E+08

b 2000:

m_i	f_i	$f_i m_i$	$f_i m_i^2$
19	17454	331626	6300894
29	61111	1772219	51394351

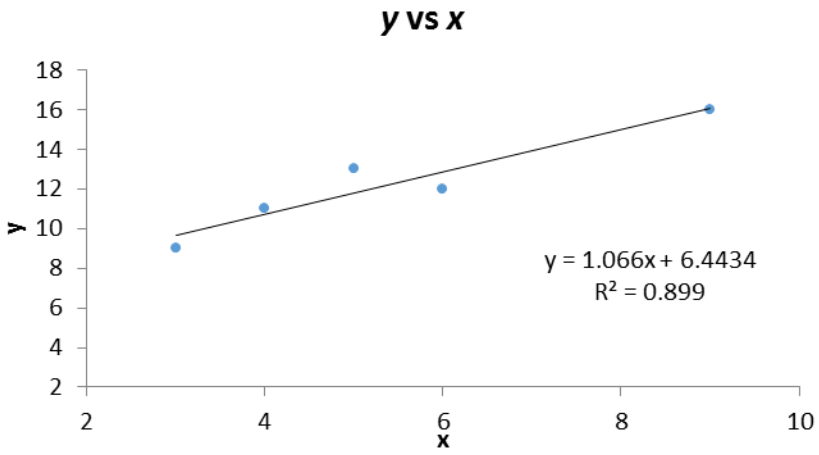
39	20485	798915	31157685
49	9024	442176	21666624
59	3583	211397	12472423
69	1300	89700	6189300
84	472	39648	3330432
Sum	113429	3685681	1.33E+08

c 2010:

m_i	f_i	$f_i m_i$	$f_i m_i^2$
19	14119	268261	5096959
29	64325	1865425	54097325
39	24793	966927	37710153
49	10801	529249	25933201
59	5189	306151	18062909
69	1488	102672	7084368
84	461	38724	3252816
Sum	121176	4077409	1.51E+08

1990	2000	2010
Mean = 30.13678	Mean = 32.49329	Mean = 33.64865
Var = 112.2864	Var = 112.4222	Var = 115.8524
SD = 10.59653	SD = 10.60293	SD = 10.76348

5.77 a



A positive linear relationship between *x* and *y* exists.

b

Obs	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	3	9	−2.4	−3.2	7.68	5.76	10.24
2	6	12	0.6	−0.2	−0.12	0.36	0.04
3	5	13	−0.4	0.8	−0.32	0.16	0.64
4	9	16	3.6	3.8	13.68	12.96	14.44
5	4	11	−1.4	−1.2	1.68	1.96	1.44
Sum	27	61	0	0	22.60	21.20	26.80

$$\bar{x} = \frac{\sum x_i}{n} = \frac{27.0}{5} = 5.4, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{61.0}{5} = 12.2$$

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{21.2}{4} = 5.3, \quad s_x = \sqrt{5.3} = 2.3$$

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{26.8}{4} = 6.7, \quad s_y = \sqrt{6.7} = 2.59$$

$$s_{xy} = \text{cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{22.60}{4} = 5.65$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{5.65}{2.3 \times 2.59} = 0.948$$

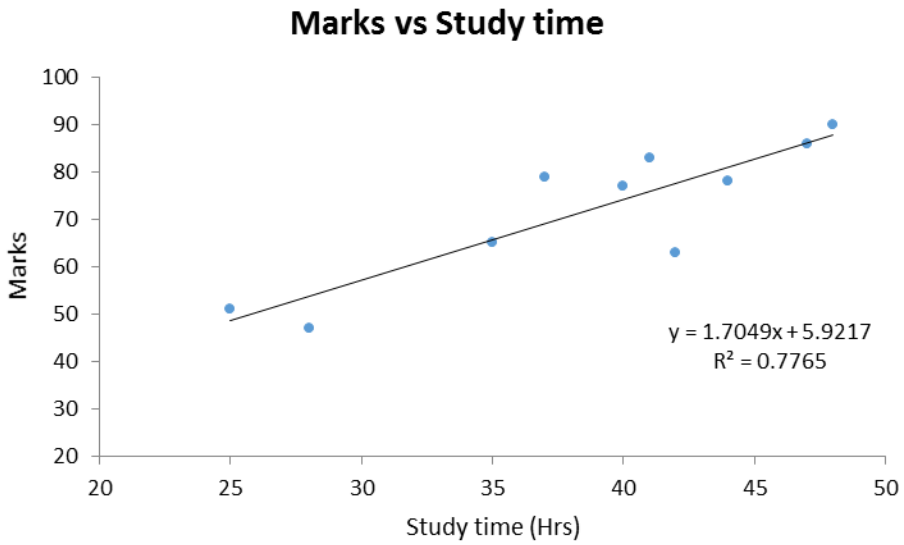
c $\hat{\beta}_1 = \dfrac{s_{xy}}{s_x^2} = \dfrac{5.65}{5.3} = 1.066$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 12.2 - 1.066(5.4) = 6.443$$

$\hat{y} = 6.443 + 1.066x$

Interpretation: When x increases by 1 unit, y would increase by 1.066 units.

5.78 a



b

	x_i	y_i	x^2_i	y^2_i	$x_i y_i$
	40	77	1,600	5,929	3,080
	42	63	1,764	3,969	2,646
	37	79	1,369	6,241	2,923
	47	86	2,209	7,396	4,041
	25	51	625	2,601	1,276
	44	78	1,936	6,084	3,432
	41	83	1,681	6,889	3,403
	48	90	2,304	8,100	4,320
	35	65	1,225	4,225	2,275
	28	47	784	2,209	1,316
Total	387	719	15,497	53,643	28,712

There is a positive linear relationship between marks and time spent studying.

$$\sum_{i=1}^{10} x_i = 387 \quad \sum_{i=1}^{10} y_i = 719 \quad \sum_{i=1}^{10} x_i^2 = 15,497 \quad \sum_{i=1}^{10} y_i^2 = 53,643 \quad \sum_{i=1}^{10} x_i y_i = 28,712$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{10-1}\left[28,712 - \frac{(387)(719)}{10}\right] = 98.52$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{10-1}\left[15,497 - \frac{(387)^2}{10}\right] = 57.79$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{10-1}\left[53,643 - \frac{(719)^2}{10}\right] = 216.32$$

$\text{cov}(x,y) = s_{xy} = 98.52$

c $r = \dfrac{s_{xy}}{s_x s_y} = \dfrac{98.52}{\sqrt{(57.79)(216.32)}} = 0.8811$

d $R^2 = r^2 = 0.8811^2 = 0.7763$

e $\hat{\beta}_1 = \dfrac{s_{xy}}{s_x^2} = \dfrac{98.52}{57.79} = 1.705$

$\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{387}{10} = 38.7$

$\bar{y} = \dfrac{\sum y_i}{n} = \dfrac{719}{10} = 71.9$

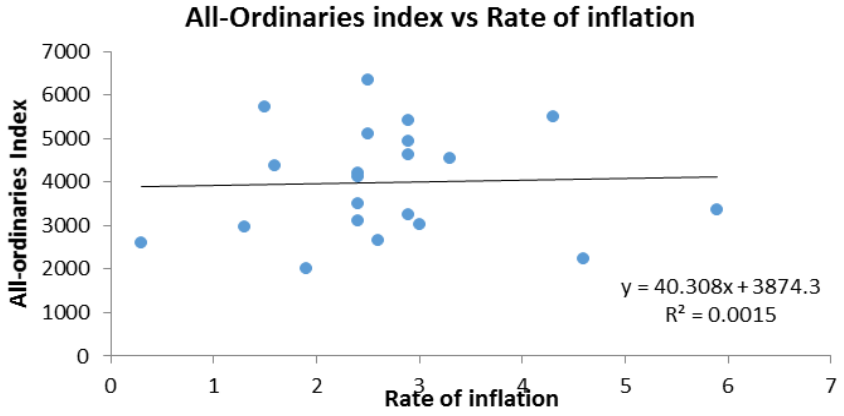
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 71.9 - (1.705)(38.7) = 5.9217$

The least squares line is

$$\hat{y} = 5.9217 + 1.705x$$

f There is a strong positive linear relationship between marks and study time. For each additional hour of study time marks increased on average by 1.705.

5.79



a

$$\sum_{i=1}^{21} x_i = 56.5 \quad \sum_{i=1}^{21} y_i = 83637.4 \quad \sum_{i=1}^{21} x_i^2 = 180.49 \quad \sum_{i=1}^{21} y_i^2 = 363721125 \qquad \sum_{i=1}^{21} x_i y_i = 226172.3$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{21-1}\left[226172.3 - \frac{(56.5)(83637.4)}{21}\right] = 57.394$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{21-1}\left[180.49 - \frac{(56.5)^2}{21}\right] = 1.424$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{21-1}\left[363721125 - \frac{(83637.4)^2}{21}\right] = 1530783.2$$

Cov(x, y) = s_{xy} = 57.39

Corr(x, y) = $r = \dfrac{s_{xy}}{s_x s_y} = \dfrac{57.394}{\sqrt{1.424}\sqrt{1530783.2}} = 0.0389$

b r = 0.0389. There is a very weak positive linear relationship between annual ordinary index and inflation rate.

c $\hat{\beta}_1 = \dfrac{s_{xy}}{s_x^2} = \dfrac{57.394}{1.424} = 40.305$

$\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{56.5}{21} = 2.69.$

$\bar{y} = \dfrac{\sum y_i}{n} = \dfrac{83637.4}{21} = 3982.73$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3982.73 - (40.305)(2.69) = 3874.27$

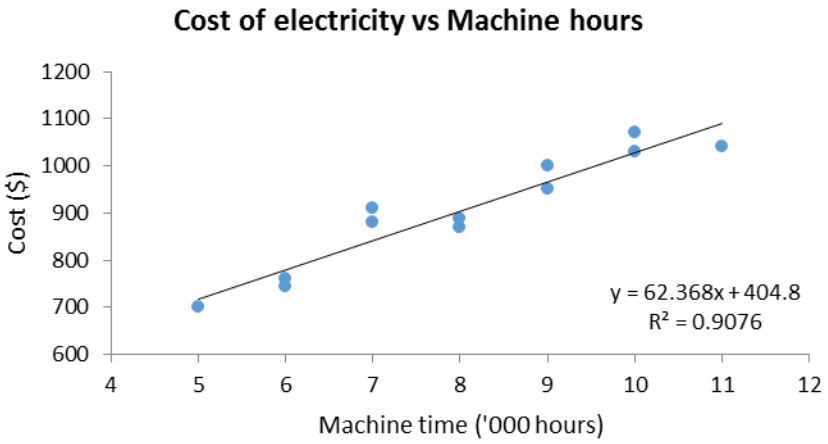
The least squares line is

$$\hat{y} = 3874.27 + 40.305x$$

AO Index = 3874.27 + 40.305 *(Rate of inflation)

d R^2 = 0.0015

5.80 a



The graph shows a positive linear relationship between the cost of electricity and the machine time.

b

	x_i	y_i	x^2_i	y^2_i	$x_i y_i$
	6	760	36	577600	4560
	9	1000	81	1000000	9000
	8	890	64	792100	7120
	7	880	49	774400	6160
	10	1070	100	1144900	10700
	10	1030	100	1060900	10300
	5	700	25	490000	3500
	7	910	49	828100	6370
	6	745	36	555025	4470
	9	950	81	902500	8550
	8	870	64	756900	6960
	11	1040	121	1081600	11440
Total	96	10845	806	9964025	89130

$$\sum_{i=1}^{12} x_i = 96 \qquad \sum_{i=1}^{12} y_i = 10845 \qquad \sum_{i=1}^{12} x_i^2 = 806 \qquad \sum_{i=1}^{12} y_i^2 = 9964025 \qquad \sum_{i=1}^{12} x_i y_i = 89130$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{12-1}\left[89130 - \frac{(96)(10845)}{12}\right] = 215.45$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{12-1}\left[806 - \frac{(96)^2}{12}\right] = 3.455$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{12-1}\left[9964025 - \frac{(10845)^2}{12}\right] = 14805.11$$

Covariance = cov(*X*,*Y*) = s_{xy} = 215.45

Coefficient of correlation = $r = \frac{s_{xy}}{s_x s_y} = \frac{215.45}{\sqrt{3.455}\sqrt{14805.11}} = 0.9526$

c As the coefficient of correlation is positive and close to 1, there is a strong positive linear relationship between hours of machine time (X) and the cost of electrical power (Y).

d

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{215.45}{3.455} = 62.36$$

$$\overline{x} = \frac{\sum x_i}{n} = \frac{96}{12} = 8.$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{10845}{12} = 903.75$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 903.75 - (62.36)(8) = 404.87$$

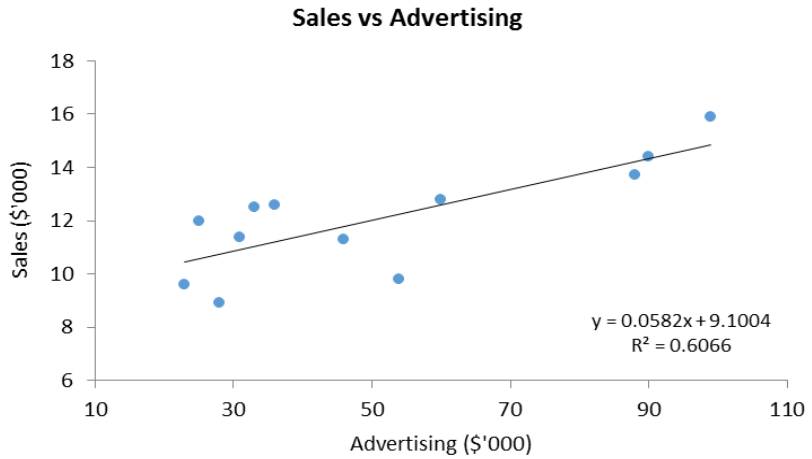
The least squares line is

$$\hat{y} = 404.87 + 62.34x$$

Cost = 404.87 + 62.34*(Machine time)

Fixed cost = $404.87 and variable cost = $62.34/hour

5.81 a



b-d

	x_i	y_i	x^2_i	y^2_i	$x_i y_i$
	23	9.6	529	92.16	220.8
	46	11.3	2116	127.69	519.8
	60	12.8	3600	163.84	768
	54	9.8	2916	96.04	529.2
	28	8.9	784	79.21	249.2
	33	12.5	1089	156.25	412.5
	25	12	625	144	300
	31	11.4	961	129.96	353.4
	36	12.6	1296	158.76	453.6
	88	13.7	7744	187.69	1205.6
	90	14.4	8100	207.36	1296
	99	15.9	9801	252.81	1574.1
Total	613	144.9	39561	1795.77	7882.2

$$\sum_{i=1}^{12} x_i = 613 \quad \sum_{i=1}^{12} y_i = 144.9 \quad \sum_{i=1}^{12} x_i^2 = 39561 \quad \sum_{i=1}^{12} y_i^2 = 1795.77 \quad \sum_{i=1}^{12} x_i y_i = 7882.2$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{12-1}\left[7882.2 - \frac{(613)(144.9)}{12}\right] = 43.66$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{12-1}\left[39561 - \frac{(613)^2}{12}\right] = 749.72$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{12-1}\left[1795.77 - \frac{(144.9)^2}{12}\right] = 4.191$$

Covariance = cov(X,Y) = s_{xy} = 43.66

Coefficient of correlation = $r = \dfrac{s_{xy}}{s_x s_y} = \dfrac{43.66}{\sqrt{749.72}\sqrt{4.191}} = 0.779$

There is a moderately strong positive linear relationship.

e

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{43.66}{749.72} = 0.0582$$

$$\overline{x} = \frac{\sum x_i}{n} = \frac{613}{12} = 51.083.$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{144.9}{12} = 12.075$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 12.075 - (0.0582)(51.083) = 9.102$$

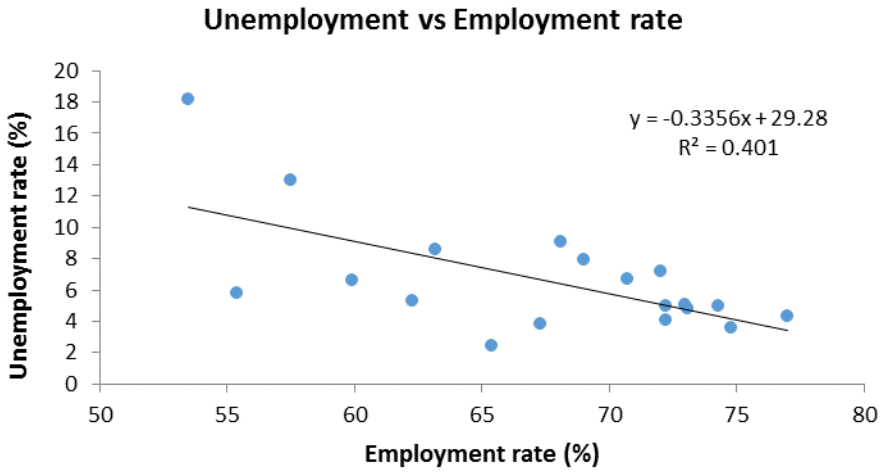
The least squares line is

$$\hat{y} = 9.102 + 0.0582x$$

Sales = 9.102 + 0.0582 × (Advertising)

f $R^2 = 0.606$

5.82

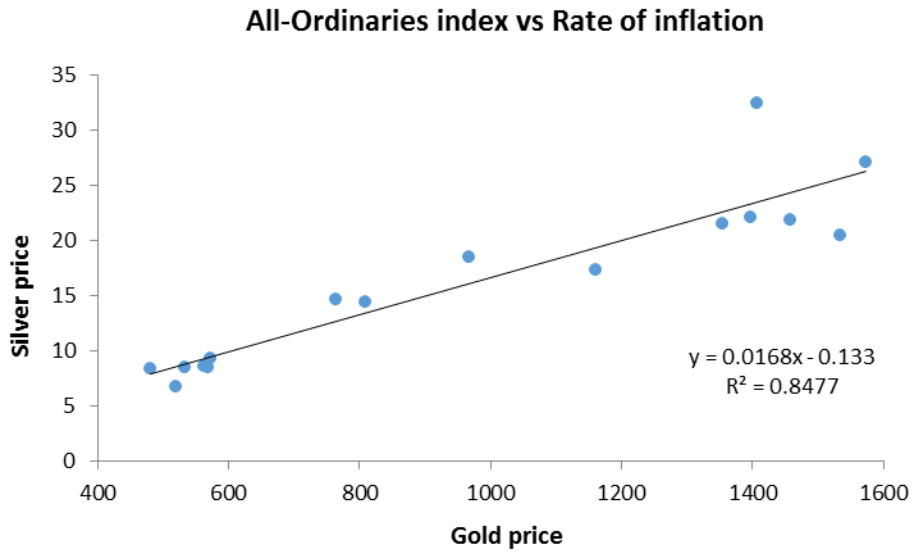


Correlation matrix:

	Employment rate	*Unemployment rate*
Employment rate	1	
Unemployment rate	-0.633246572	1

$R^2 = r^2 = (-0.6332)^2 = 0.4009$; 40.09% of the variation in the employment rate is explained by the variation in the unemployment rate.

5.83



b

	x_i	y_i	x^2_i	y^2_i	$x_i y_i$
	481.68	8.3	232015.6	69.4	4012.4
	533.12	8.5	284216.9	72.9	4552.8
	562.88	8.6	316833.9	73.6	4829.5
	519.18	6.8	269547.9	46.2	3530.4
	568.03	8.5	322658.1	72.1	4822.6
	572.74	9.3	328031.1	86.1	5315.0
	809.31	14.4	654982.7	207.9	11670.3
	764.53	14.6	584506.1	214.0	11185.1
	967.87	18.5	936772.3	340.4	17857.2
	1161.59	17.4	1349291.3	301.4	20165.2
	1457.79	21.8	2125151.7	476.1	31809.0
	1408.2	32.4	1983027.2	1051.7	45667.9
	1573.54	27.1	2476028.1	735.0	42658.7
	1353.87	21.6	1832964.0	464.4	29175.9
	1397.62	22.1	1953341.7	489.7	30929.3
	1534.09	20.5	2353432.1	419.8	31433.5
Total	15666.0	260.4	18002801.0	5120.9	299614.8

$$\sum_{i=1}^{16} x_i = 15666 \qquad \sum_{i=1}^{16} y_i = 260.4 \qquad \sum_{i=1}^{16} x_i^2 = 18002801 \qquad \sum_{i=1}^{16} y_i^2 = 5120.9$$

$$\sum_{i=1}^{16} x_i y_i = 299614.8$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{16-1}\left[299614.8 - \frac{(15666)(260.4)}{16}\right] = 2976.02$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{16-1}\left[18002801 - \frac{(15666)^2}{16}\right] = 177583.3$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{16-1}\left[5120.9 - \frac{(260.4)^2}{16}\right] = 58.84$$

Covariance = cov(X,Y) = s_{xy} = 2976.02

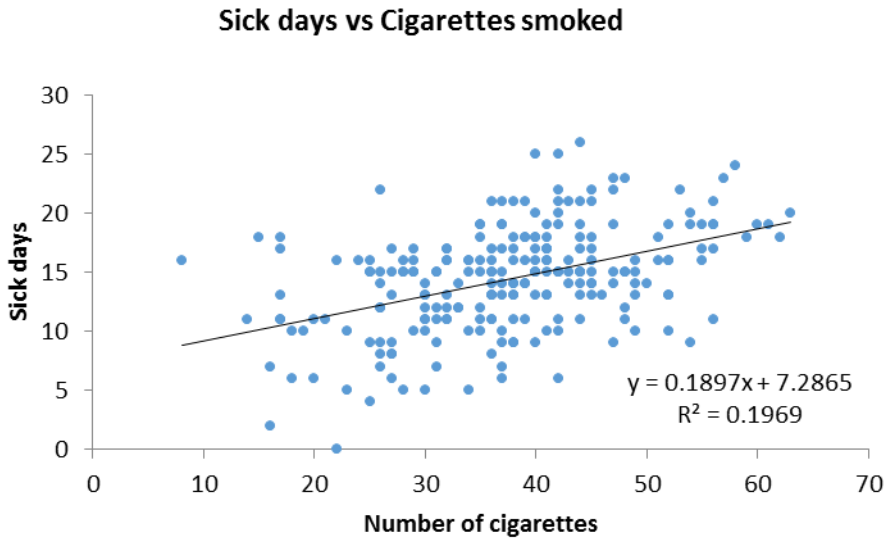
Coefficient of correlation = $r = \dfrac{s_{xy}}{s_x s_y} = \dfrac{2976.02}{\sqrt{177583.3}\sqrt{58.84}} = 0.9207$

a $\text{cov}(R, R_*) = 2976.02$

$r = 0.921$

b There is a positive linear relationship between the monthly spot gold price and the monthly spot silver price.

5.84



a Population covariance matrix (using Excel):

	Cigarettes	Days
Cigarettes	107.81	
Days	20.46	19.72

Sample: $\text{cov}(x, y) = 20.46 \times \left(\frac{231}{230}\right) = 20.55$

$$\text{var } x = s_x^2 = 20.46 \times \left(\frac{231}{230}\right) = 108.28$$

$$\text{var } y = s_y^2 = 19.72\left(\frac{231}{230}\right) = 19.80$$

[Note: Excel calculates population covariance. Hence to work out sample correlation we need to multiply by $(n/n–1)$]

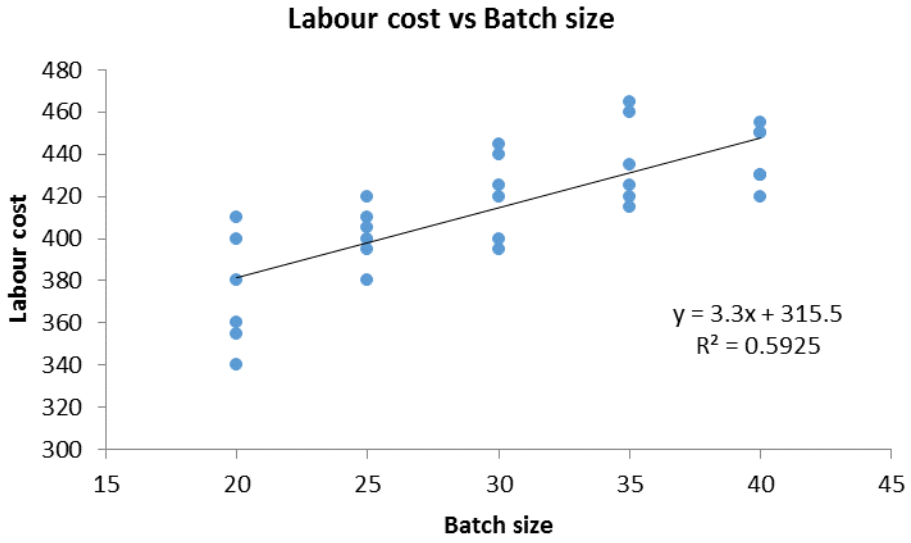
b Correlation matrix (using Excel):

	Cigarettes	Days
Cigarettes	1	
Days	0.4437	1

$r = 0.4437$

c The coefficient of correlation tells us that there is a weak positive linear relationship between smoking and the duration of colds.

5.85



a $r = \sqrt{0.5925} = 0.77$. There exists a strong positive linear relationship between labour cost and batch size.

b $\hat{y}$ = 315.5 + 3.3x; Fixed costs = \$315.50, variable costs = \$3.30

5.86

	A	B
1		*Bone Loss*
2	Mean	35.01
3	Standard Error	0.69
4	Median	36.00
5	Mode	38.00
6	Standard Deviation	7.68
7	Sample Variance	59.04
8	Kurtosis	0.08
9	Skewness	-0.19
10	Range	38.00
11	Minimum	15.00
12	Maximum	53.00
13	Sum	4376.00
14	Count	125.00

a Sample mean $= \bar{x} = 35.01$, Median = 36

b $s = 7.68$

c Half of the bone density losses lie below 36. At least 75% of the observations lie between 19.64 and 50.38, and at least 88.9% of the observations lie between 11.96 and 58.06.

5.87 a $\mu = \dfrac{11+12+\ldots+7+5}{12} = \dfrac{38}{12} = 3.17$

The median is 4.5, the mean of the two middle values: 4 and 5.

$$\sigma^2 = \frac{(11-3.17)^2+(12-3.17)^2+\ldots+(7-3.17)^2+(5-3.17)^2}{12} = 38.81$$

$$s = \sqrt{38.81} = 6.23$$

b Arranging the 12 measurements in ascending order, we obtain:

$-10, -6, -1, 1, 2, 4, 5, 5, 7, 8, 11, 12$

The location of the upper quartile is $L_{75} = (12+1)\dfrac{75}{100} = 9.75.$

Therefore, $P_{75} = 7 + (8-7)0.75 = 7.75.$ Hence the upper quartile is 7.75.

The location of the lower quartile is $L_{25} = (12+1)\dfrac{25}{100} = 3.25.$

Therefore, $P_{25} = -1 + (1-(-1))0.25 = -1 + 0.5 = -0.5.$ Hence the lower quartile is −0.5.

5.88 Arranging the observations in ascending order, we obtain

6, 6, 8, 9, 10, 11, 12, 14, 16, 16, 17, 17, 18, 19, 19, 19, 20, 21, 21, 21, 22, 24, 25, 25, 29

The median (middle) value is 18. There are two modes: 19 and 21.

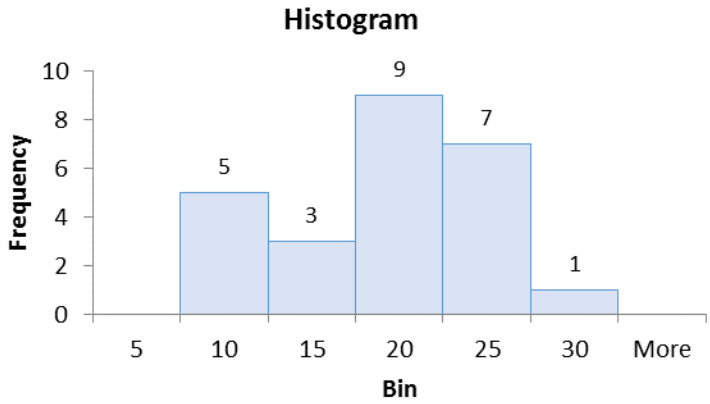
$$\bar{x} = \frac{\sum_{i=1}^{25} x_i}{25} = \frac{21+18+\ldots+29+25}{25} = \frac{425}{25} = 17$$

Using the shortcut formula for the variance,

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{25} x_i^2 - \frac{\left(\sum_{i=1}^{25} x_i\right)^2}{25}\right) = \frac{1}{24}\left(8.149 - \frac{(425)^2}{25}\right) = 38.5$$

$$s = \sqrt{38.5} = 6.20$$

5.89 a $\bar{x} \pm 2s = 17 \pm 2(6.2) = 17 \pm 12.4 = (4.6, 29.4)$



The proportion of the items falling into the interval (4.6, 29.4) is 100%.

b S = 6, L = 29. Therefore, from the histogram in part (a), the distribution of the sample is only very roughly mound-shaped.

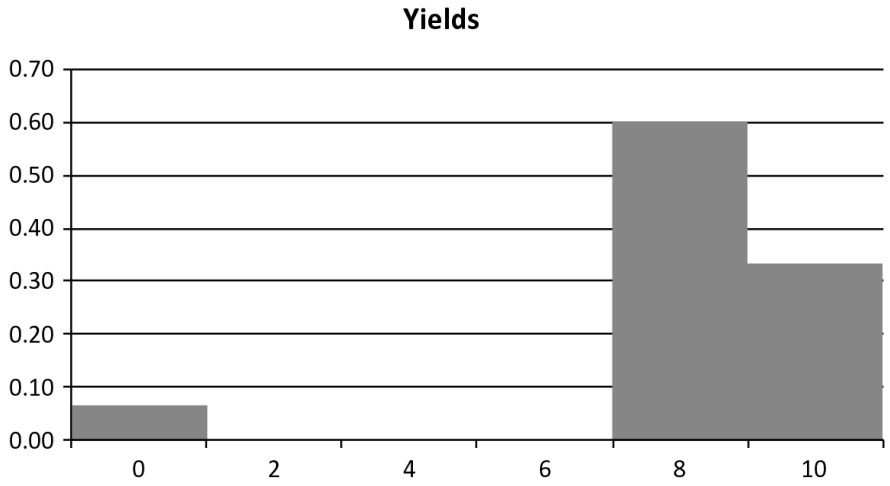
c $\bar{x} \pm s = 17 \pm 6.2 = (10.8, 23.2)$

$\bar{x} \pm 2s = 17 \pm 2(6.2) = 17 \pm 12.4 = (4.6, 29.4)$

Interval	Actual Proportion	Empirical Rule Proportion
[10.8, 23.2]	$\frac{16}{25} = 0.64$	0.68
[4.6, 29.4]	$\frac{25}{25} = 1.00$	0.95

5.90 a Mean = 7.29, SD = 2.14

b



c Median = 7.56

5.91 a We have chosen to use the ten's digit as the stem and the one's digit as the leaf. Although we have recorded the leaves in ascending order, it is not necessary to do so.

Stem	Leaf
2	48
3	268
4	124456779
5	01245789
6	146

b Arranging the 25 measurements in ascending order, we obtain

24, 28, 32, 36, 38, 41, 42, 44, 44, 45, 46, 47, 47, 49, 50, 51, 52, 54, 55, 57, 58, 59, 61, 64, 66

The median age is 47, which is the middle (13^{th}) value.

c $L_{25} = (25+1)\frac{25}{100} = 6.5$; 6^{th} observation = 41 and 7^{th} observation = 42

$P_{25} = 41 + 0.5(42 - 41) = 41.5$

d $L_{75} = 19.5; \quad P_{75} = 55 + 0.5(57 - 55) = 56$

e $L_{80} = (25+1)\frac{80}{100} = 20.8$; 20th observation = 57 and 21st observation = 58

$P_{80} = 57 + 0.8(58 - 57) = 57.8$.

f Yes, the firm has reason to be somewhat concerned about the distribution of ages. Twenty-five percent of its brokers are over 56 and will retire in the next 10 years, requiring the firm to plan for the orderly retirement and replacement of these brokers. Also, only 25% of the brokers are under the age of 42. The firm may wish to increase this