THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2002
Campus: City

STATISTICS

Information Visualisation

(Time allowed: THREE hours)

NOTE: Attempt all FIVE questions.
All questions are worth equal marks.
You should allot equal time for answering each question.

1. The following set of data shows how the intake of all Swedish upper secondary school students in 1991 can be broken down by gender and area of study.

Intake into three-year courses at upper secondary school, autumn, 1991.

	No. entrants	*Percentage women*
Theoretical subjects:		
Humanities	4104	85
Social sciences	12048	71
Science	8520	52
Economics	12431	56
Engineering	10930	19
Vocational Subjects:		
Construction	465	1
Electrical engineering	815	4
Vehicles and transport	1126	4
Business studies	1200	62
Industrial engineering	1752	3
Agriculture	1289	50
Nursing (care)	3427	87
Catering	513	58
Nursing	772	92

(The data are taken from the book *Graphing Statistics & Data*, written by a group of Swedish Government statisticians.)

The book shows a number of different ways of presenting the data, including the following plot:

Intake by Sex into Upper Secondary School Three–Year Courses 1991
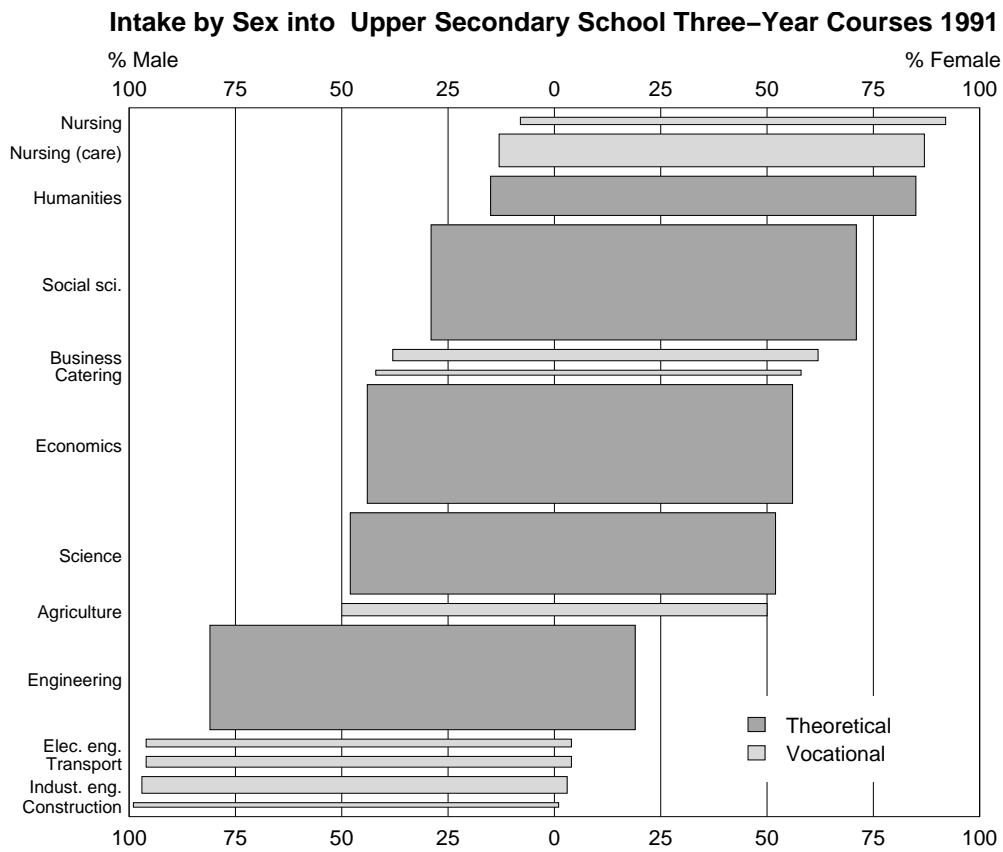
(a) Briefly explain and interpret this graph.

(b) Describe the encodings used to represent the data in this graph.

(c) What are the good and bad features of this graph?

(d) Suggest a better way of presenting this information, and explain why your method is better.

(e) Explain (in detail) how your type of graph could be produced in R.

2. (a) For humans, colour is an inherently three-dimensional phenomenon. Describe why this is.

   (b) The representation of colour is different at different stages of the visual pathway (from the eye to the visual cortex). Explain the representations used at **three** different stages of the visual pathway.

   (c) Describe **three** different colour spaces, and indicate how **one** of the spaces can be used when choosing colours for graphs.

   (d) Describe **three** ways of using colour which should be avoided when producing graphs.

3. A question on graphical encoding in simple graphs.

(a) Using what you know about perception, explain which of stacked bar charts or or side-by-side bar charts is usually a better way of showing data.

(b) When might it be useful to use the the other kind of bar plot.

Explain how to create a mosaic plot from these values (using pen and paper). Note: There are two possible mosaics — choose the one you feel is most appropriate.

(c) There have been claims that there is racial prejudice in the the way that the death penalty is applied in the United States. The following table shows the results of 332 murder cases where the death penalty could have been applied.

Victim's Race	Defendant's Race	Death Penalty Yes	Death Penalty No
White	White	19	132
	Black	11	52
Black	White	0	9
	Black	6	97

A mosaic plot for these data is shown below. Write a short description of what this plot shows.
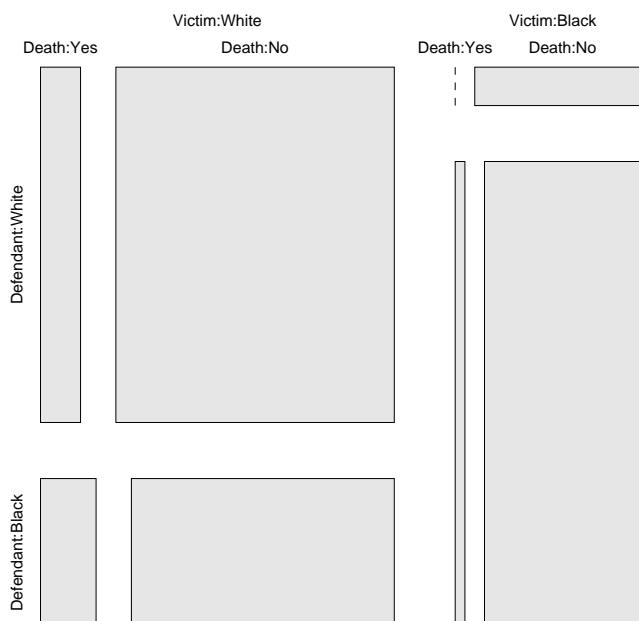
4. (a) Histograms and density plots both provide a graphical representation of the distribution of a set of values. Explain why density plots provide a better representation than histograms.

    (b) Describe precisely what it is that a density plot shows.

    (c) Describe how to construct a quantile-quantile plot comparing two samples of numerical values.

    (d) Female Yranigami beetles range from 1cm to 1.5cm in length. Male Yranigami beetles have a similar distribution of lengths, but are smaller, ranging from 0.6cm to 1.1cm in length. Sketch the appearance you would expect for a quantile-quantile plot made to compare the lengths of 100 male and 100 female Yranigami beetles.

    (e) In contrast to the Yranigami, the males of Puedam beetle range from 2 cm to 4 cm in length and are typically twice the size of the female Puedam beetle. Sketch the appearance you would expect for a quantile-quantile plot made to compare the lengths of 100 male and 100 female Puedam beetles.

5. (a) To investigate the relationship between a dependent variable *z* and independent variables *u* and *v* the following R statements are used to create the plot below.

```
> v.intervals = equal.count(v, number = 6, overlap = 0.5)
> xyplot(z ~ u | v.intervals)
```

Explain what these statements do and explain in detail the resulting graph. Describe the nature of the relationship between *z*, *u* and *v*.
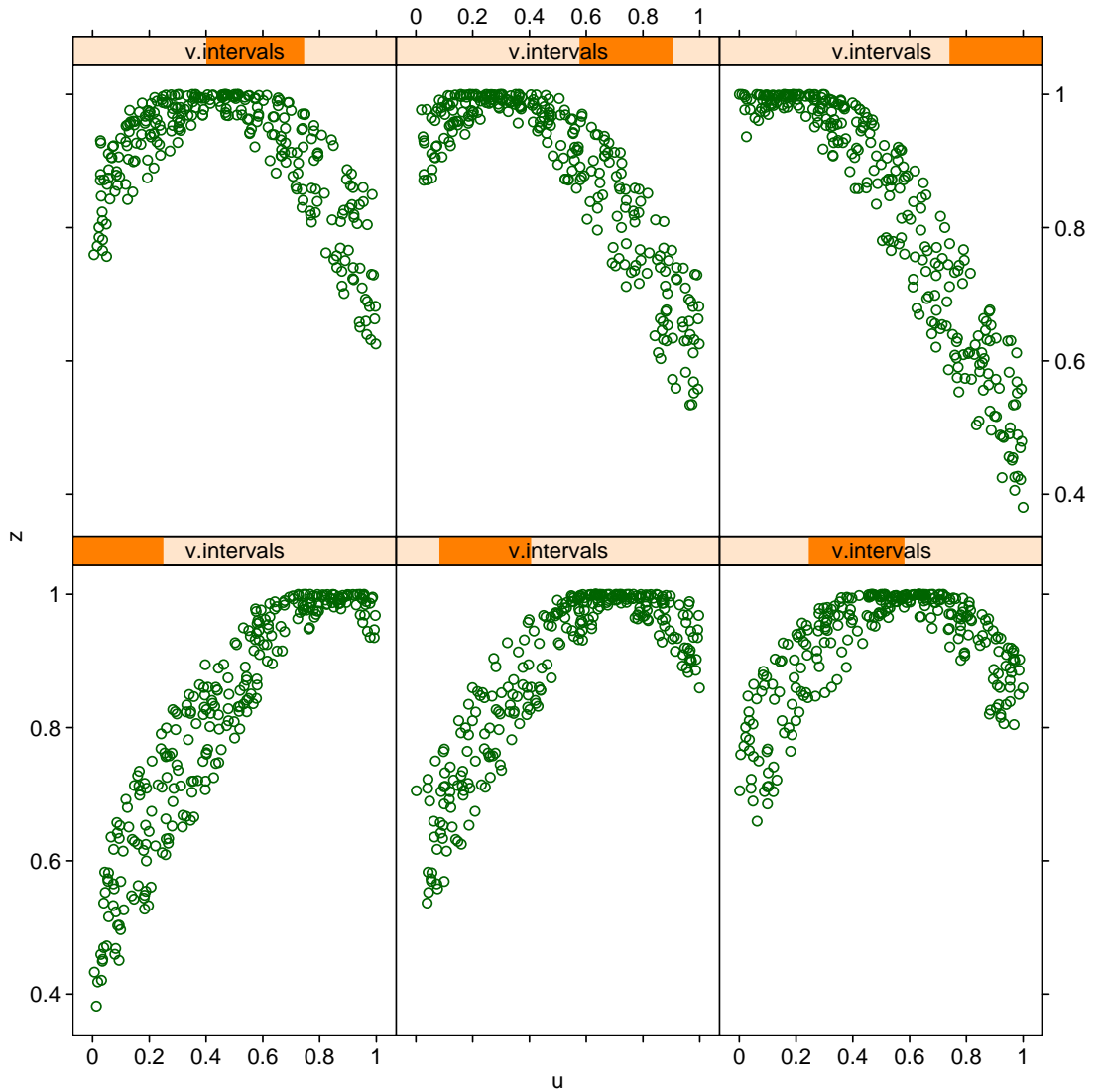
(b) The figure below shows a graphical display produced from the time series of monthly deaths from lung diseases (bronchitis, emphysema and asthma) in the UK, 1974-1979.

Explain what the graph shows and indicate (with a sketch) how it should be labelled.