5. **Occupational mobility**
   According to the R help page, the *Yamaguchi87* dataset in **vcdExtra** has become a classic for models comparing two-way mobility tables.
   (a) How do the distributions of occupations of the sons in the three countries compare?
   (b) How do the distributions of the sons' and fathers' occupations in the UK compare?
   (c) Are you surprised by the results or are they what you would have expected?

6. **Whisky**
   The package **bayesm** includes the dataset *Scotch*, which reports which brands of whisky 2218 respondents consumed in the previous year.
   (a) Draw a barchart of the number of respondents per brand. What ordering of the brands do you think is best?
   (b) There are 20 named brands and a further category `Other.brands`. That entails drawing a lot of bars. If you decided to plot only the biggest brands individually and group the rest all together in the 'Other' group, what cut-off would you use for defining a big brand?
   (c) Another version of the dataset called *whiskey* is given in the package **flexmix**. It is made up of two data frames, *whiskey* with the basic data, and *whiskey_brands* with information on whether the whiskeys are blends or single malts. How would you incorporate this information in your graphics, by using colour, by using a different ordering, or by drawing two graphics rather than one?
   (d) Which of the spellings, 'whisky' or 'whiskey', is more appropriate for this dataset?

7. **Choice of school**
   The dataset *GSOEP9402* in the package **AER** provides data on 675 14-year-old children in Germany. The data come from the German Socio-Economic Panel for the years 1994 to 2002.
   (a) Which variables are nominal, ordinal, or discrete?
   (b) Draw barcharts for the variables. Are any similar in form, and what explanations would you suggest for these similarities?
   (c) The variable `meducation` refers to the mother's educational level in years. Would you describe it as ordinal or discrete, and how should it be displayed?
   (d) Summarise briefly the main information shown by your graphics.

8. **Olive oils from Italy**
   The olive oils dataset is well known and can be found in several packages, for instance as *olives* in **extracat**. The original source for the data is the paper [Forina et al., 1983].
   (a) Draw a scatterplot matrix of the eight continuous variables. Which of the fatty acids are strongly positively associated and which strongly negatively associated?
   (b) Are there outliers or other features worth mentioning?

9. **Boston housing**
   The Boston dataset was introduced in Chapter 3.
   (a) Draw a splom of all the continuous variables (i.e., all except the variable `chas`). Which variables are positively associated with `medv`, the median home value?
   (b) Several of the scatterplots involving the variable `crim`, the per capita crime rate, have an unusual form, where higher values of `crim` only occur for one particular value of the other variable. How would you explain this?
   (c) There are many different scatterplot forms in the display. Pick out five and describe how you would interpret them.

10. **Hertzsprung-Russell**
    The Hertzsprung-Russell diagram is a famous scatterplot of the relationship between the absolute magnitudes of stars and their effective temperatures and is over one hundred years old. Although examples of the plot can be found all over the place, it is surprisingly difficult to find the data underlying them. There is a dataset of 47 cases, *starsCYG*, in the package **robustbase**, but that is really too small. The dataset *HRstars* with 6220 stars in package **GDAdata** is from the Yale Trigonometric Parallax Dataset and was downloaded from [Mihos, 2005].
    (a) Plot Y against X. How does your plot differ from the plots you find on the web, for instance from a Google search for images of the Hertzsprung-Russell diagram?
    (b) The plots seem to use different numbers of stars. Are some more likely to be used than others?
    (c) You can colour and annotate your plot using techniques described in Chapter 13. What would you suggest?

11. **Intermission**
    The painting *La Grande Jatte* by Georges Seurat hangs in the *Art Insitute of Chicago*, a classic of putting dots together to form an overall impression. Do you think the artist intended his painting only to be viewed from a distance?

2. **Pottery**
   The package **HSAUR2** includes a dataset on the chemical composition of Romano-British pottery, *pottery*, with 45 cases and 10 variables.
   (a) Draw a pcp of the nine composition variables. What features can you see?
   (b) Make a new variable with the cases with low values on `MgO`. How are these cases different from the rest on the other variables?
   (c) Colour your original pcp using the site information, `kiln`. Which kilns can be easily distinguished from the others using which variables?

3. **Olive oils**
   The olive oils dataset was introduced in Exercise 8 of Chapter 5.
   (a) Draw a default parallel coordinate plot and describe the various features you can see.
   (b) Draw the same plot and additionally colour the oils by the region they come from. What additional information can you find?
   (c) Discuss which features of the dataset are easier to see with a pcp and which are easier to see with a scatterplot matrix.

4. **Cars**
   The dataset *Cars93* was introduced in §5.4. Draw a pcp of the nine variables `Price`, `MPG.city`, `MPG.highway`, `Horsepower`, `RPM`, `Length`, `Width`, `Turn.circle`, and `Weight`.
   (a) What conclusions would you draw from your plot?
   (b) What plot would you draw to compare US cars with non-US cars on these variables? What does the plot tell you about the differences between US cars and the others?
   (c) Is a pcp with `unimax` scaling informative? Try colouring it by the factor variable `Cylinder` to gain additional insight.

5. **Bodyfat**
   The dataset *bodyfat* is available in the **MMST** package. It provides estimates of the percentage of body fat of 252 men, determined by underwater weighing, and body circumference measurements. The dataset is used as a multiple regression example to see if body fat percentage can be predicted using the other measurements. Draw a parallel coordinate plot for the dataset.
   (a) Are there any outliers? What can you say about them?
   (b) Can you deduce anything about the height variable?
   (c) What can you say about the relationship between the first two variables, *density* and *bodyfat*?
   (d) Do you think the ordering of the variables is sensible? What alternative orderings might be informative?

6. **Exam marks**
   In the package **SMPracticals**, there is a dataset *mathmarks* with the marks out of 100 in five subjects for 88 students. The dataset is fairly old, first appearing in the statistical literature in [Mardia et al., 1979] and it was used in an example at the end of §5.6. It is interesting to note that all students had marks in all subjects. Possibly students who missed an exam were excluded.

   (a) Explore the dataset using pcp's. What information can you uncover and which pcp would you use to present your results to others?
   (b) Apparently the first two exams (`mechanics` and `vectors`) were closed book, while the other three were open book. Draw a pcp with boxplots to see if there is evidence that the students got lower marks on closed-book exams. Is it useful to superimpose the polygonal lines (possibly using alpha-blending) or not?

7. **Wine**
   The *wine* dataset can be found in the packages **gclus**, **MMST**, **pgmm**, and **rattle**. They took the data from the UCI Machine Learning Repository [Bache and Lichman, 2013]. The original source is an Italian software package [Forina et al., 1988]. The version in **pgmm** has about twice as many variables as the others, and the version in **MMST** includes the names of the three classes of wine, rather than the numeric coding that the other versions use.

   (a) Use pcp's to investigate how well the variables separate these classes.
   (b) Are there any outliers?
   (c) Is there evidence of subgroups within the classes?

8. **Boston housing**
   Carry out a cluster analysis of the *Boston* data using Ward's methods (`method=ward.D2`) on standardised variables and choose the four cluster solution. You could present your results in a number of ways:

   (a) with a single pcp of the variables with the case profiles coloured by cluster;
   (b) with several pcp's, one for each set of cluster cases (at the time of writing this will not work with *ggparcoord* if all the cases in a cluster have the same value on a variable when a default or `uniminmax` scaling is used);
   (c) with several pcp's, one for each set of cluster cases and with the remaining cases plotted in the background.

   What are the advantages and disadvantages of the three alternatives? What plot or group of plots would you choose for displaying your clustering results?

9. **Intermission**
   Jackson Pollock's *Convergence* is in the *Albright-Knox Art Gallery* in Buffalo, New York. What can you see in this picture?