A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 2

Data Analysis Using Graphical Displays: Malignant Melanoma in the US and Chinese Health and Family Life

2.1 Introduction

2.2 Initial Data Analysis

2.3 Analysis Using R

2.3.1 Malignant Melanoma

We might begin to examine the malignant melanoma data in Table **??** by constructing a histogram or boxplot for *all* the mortality rates in Figure 2.1. The `plot`, `hist` and `boxplot` functions have already been introduced in Chapter 1 and we want to produce a plot where both techniques are applied at once. The `layout` function organizes two independent plots on one plotting device, for example on top of each other. Using this relatively simple technique (more advanced methods will be introduced later) we have to make sure that the x-axis is the same in both graphs. This can be done by computing a plausible range of the data, later to be specified in a plot via the `xlim` argument:

```
R> xr <- range(USmelanoma$mortality) * c(0.9, 1.1)
R> xr
[1]  77.4 251.9
```

Now, plotting both the histogram and the boxplot requires setting up the plotting device with equal space for two independent plots on top of each other. Calling the `layout` function on a matrix with two cells in two rows, containing the numbers one and two, leads to such a partitioning. The `boxplot` function is called first on the mortality data and then the `hist` function, where the range of the x-axis in both plots is defined by $(77.4, 251.9)$. One tiny problem to solve is the size of the margins; their defaults are too large for such a plot. As with many other graphical parameters, one can adjust their value for a specific plot using function `par`. The R code and the resulting display are given in Figure 2.1.

Both the histogram and the boxplot in Figure 2.1 indicate a certain skewness of the mortality distribution. Looking at the characteristics of all the mortality rates is a useful beginning but for these data we might be more interested in comparing mortality rates for ocean and non-ocean states. So we

```
R> layout(matrix(1:2, nrow = 2))
R> par(mar = par("mar") * c(0.8, 1, 1, 1))
R> boxplot(USmelanoma$mortality, ylim = xr, horizontal = TRUE,
+          xlab = "Mortality")
R> hist(USmelanoma$mortality, xlim = xr, xlab = "", main = "",
+       axes = FALSE, ylab = "")
R> axis(1)
```

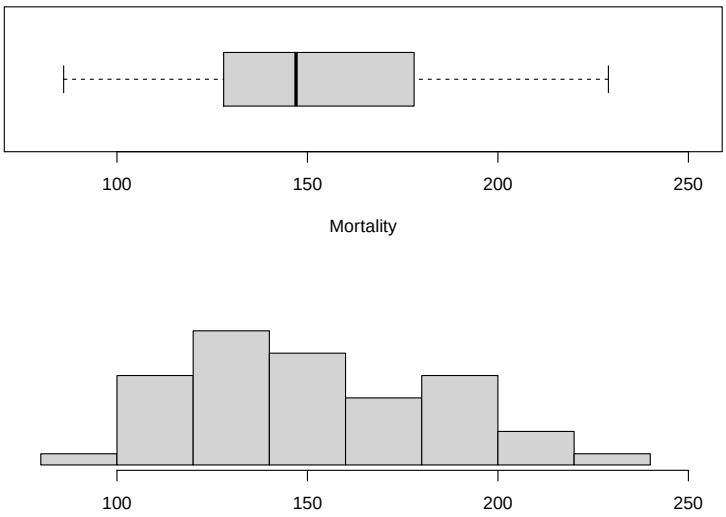

Figure 2.1 Histogram (top) and boxplot (bottom) of malignant melanoma mortality rates.

might construct two histograms or two boxplots. Such a *parallel boxplot*, visualizing the conditional distribution of a numeric variable in groups as given by a categorical variable, are easily computed using the `boxplot` function. The continuous response variable and the categorical independent variable are specified via a *formula* as described in Chapter 1. Figure 2.2 shows such parallel boxplots, as by default produced the `plot` function for such data, for the mortality in ocean and non-ocean states and leads to the impression that the mortality is increased in east or west coast states compared to the rest of the country.

Histograms are generally used for two purposes: counting and displaying the distribution of a variable; according to Wilkinson (1992), 'they are effective for neither'. Histograms can often be misleading for displaying distributions because of their dependence on the number of classes chosen. An alternative

```
R> plot(mortality ~ ocean, data = USmelanoma,
+       xlab = "Contiguity to an ocean", ylab = "Mortality")
```

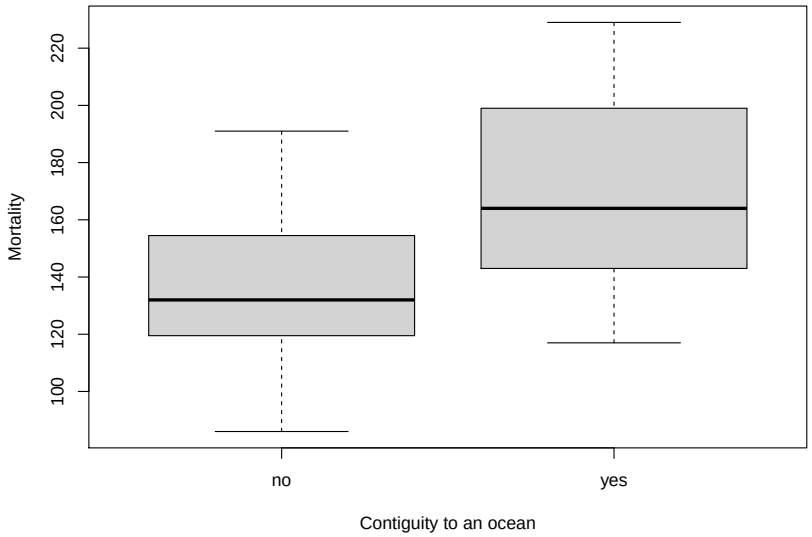

Figure 2.2 Parallel boxplots of malignant melanoma mortality rates by contiguity to an ocean.

is to formally estimate the density function of a variable and then plot the resulting estimate; details of density estimation are given in Chapter 8 but for the ocean and non-ocean states the two density estimates can be produced and plotted as shown in Figure 2.3 which supports the impression from Figure 2.2. For more details on such density estimates we refer to Chapter 8.

Now we might move on to look at how mortality rates are related to the geographic location of a state as represented by the latitude and longitude of the center of the state. Here the main graphic will be the scatterplot. The simple xy scatterplot has been in use since at least the eighteenth century and has many virtues – indeed according to Tufte (1983):

> The relational graphic – in its barest form the scatterplot and its variants – is the greatest of all graphical designs. It links at least two variables, encouraging and even imploring the viewer to assess the possible causal relationship between the plotted variables. It confronts causal theories that x causes y with empirical evidence as to the actual relationship between x and y.

Let's begin with simple scatterplots of mortality rate against longitude and mortality rate against latitude which can be produced by the code preceding Figure 2.4. Again, the `layout` function is used for partitioning the plotting

```
R> dyes <- with(USmelanoma, density(mortality[ocean == "yes"]))
R> dno <- with(USmelanoma, density(mortality[ocean == "no"]))
R> plot(dyes, lty = 1, xlim = xr, main = "", ylim = c(0, 0.018),
+       xlab = "Mortality")
R> lines(dno, lty = 2)
R> legend("topleft", lty = 1:2, legend = c("Coastal State",
+         "Land State"), bty = "n")
```

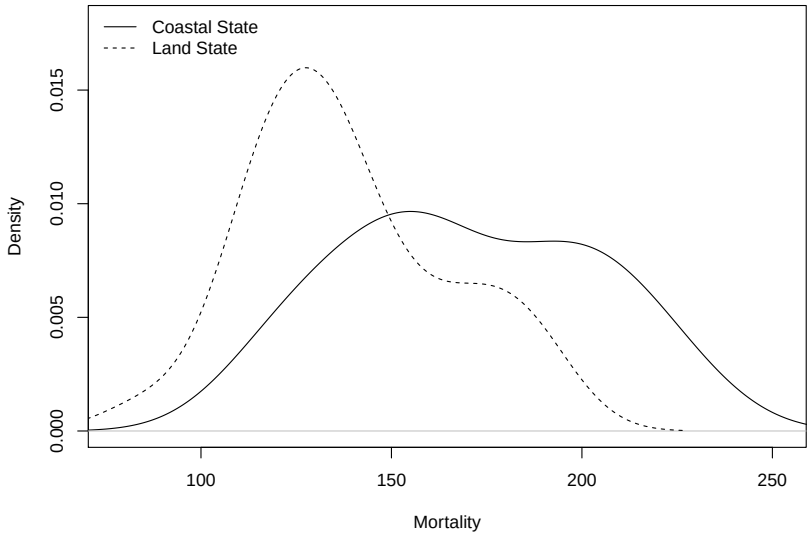

Figure 2.3 Estimated densities of malignant melanoma mortality rates by contiguity to an ocean.

device, now resulting in two side-by-side plots. The argument to `layout` is now a matrix with only one row but two columns containing the numbers one and two. In each cell, the `plot` function is called for producing a scatterplot of the variables given in the *formula*.

Since mortality rate is clearly related only to latitude we can now produce scatterplots of mortality rate against latitude separately for ocean and non-ocean states. Instead of producing two displays, one can choose different plotting symbols for either states. This can be achieved by specifying a vector of integers or characters to the `pch`, where the ith element of this vector defines the plot symbol of the ith observation in the data to be plotted. For the sake of simplicity, we convert the `ocean` factor to an *integer* vector containing the numbers one for land states and two for ocean states. As a consequence, land states can be identified by the dot symbol and ocean states by triangles.

```
R> layout(matrix(1:2, ncol = 2))
R> plot(mortality ~ longitude, data = USmelanoma,
+       ylab = "Mortality", xlab = "Longitude")
R> plot(mortality ~ latitude, data = USmelanoma,
+       ylab = "Mortality", xlab = "Latitude")
```

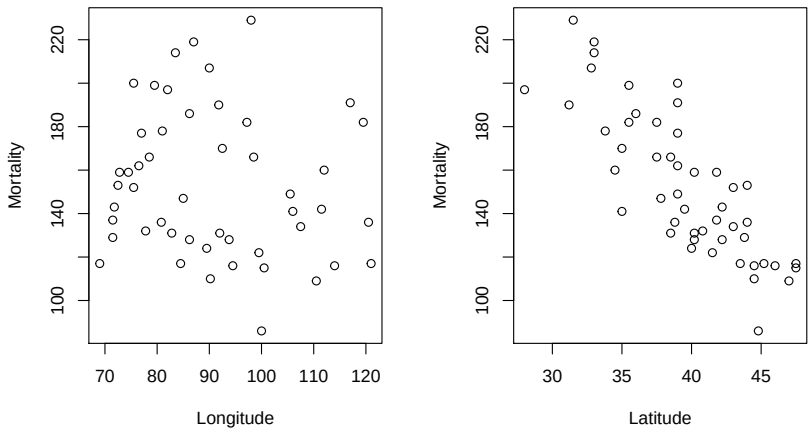

Figure 2.4 Scatterplot of malignant melanoma mortality rates by geographical location.

It is useful to add a legend to such a plot, most conveniently by using the `legend` function. This function takes three arguments: a string indicating the position of the legend in the plot, a character vector of labels to be printed and the corresponding plotting symbols (referred to by integers). In addition, the display of a bounding box is anticipated (`bty = "n"`). The scatterplot in Figure 2.5 highlights that the mortality is lowest in the northern land states. Coastal states show a higher mortality than land states at roughly the same latitude. The highest mortalities can be observed for the south coastal states with latitude less than 32°, say, that is

```
R> subset(USmelanoma, latitude < 32)
```

	mortality	latitude	longitude	ocean
Florida	*197*	*28.0*	*82.0*	*yes*
Louisiana	*190*	*31.2*	*91.8*	*yes*
Texas	*229*	*31.5*	*98.0*	*yes*

Alternatively, we also may simply want to look at a color-coded map of the United States, where each state is plotted in a color that corresponds to its mortality rate. It is fairly simple to set-up such a plot using the **sp** family of packages (Pebesma and Bivand, 2013). We start with loading a map of the mainland states, basically a number of polygons:

```
R> plot(mortality ~ latitude, data = USmelanoma,
+       pch = (1:2)[ocean], ylab = "Mortality",
+       xlab = "Latitude")
R> legend("topright", legend = c("Land state", "Coast state"),
+         pch = 1:2, bty = "n")
```

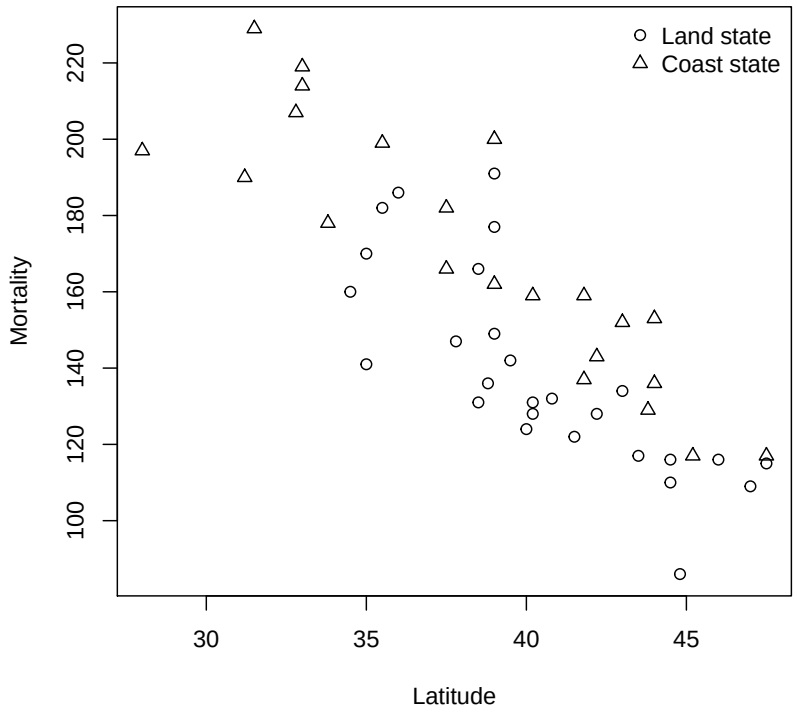

Figure 2.5 Scatterplot of malignant melanoma mortality rates against latitude.

```
R> library("sp")
R> library("maps")
R> library("maptools")
R> states <- map("state", plot = FALSE, fill = TRUE)
```

It is of course important to match the mortality rates to the corresponding state. We therefore create unique names of the states in lower-case letters for both the polygons and the mortality data

```
R> IDs <- sapply(strsplit(states$names, ":"), function(x) x[1])
R> rownames(USmelanoma) <- tolower(rownames(USmelanoma))
```

```
R> spplot(us2, "mortality", col.regions = rev(grey.colors(100)))
```

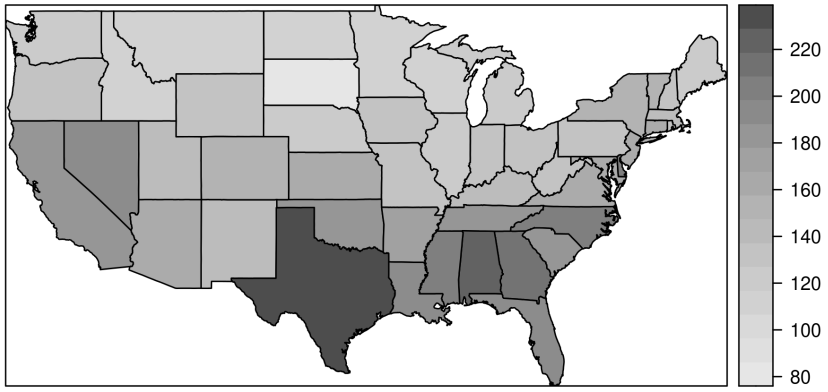

Figure 2.6 Map of the United States of America showing malignant melanoma mortality rates.

Now we are ready to merge these two objects into a so-called *SpatialPolygonsDataFrame* object. We first create a *SpatialPolygons* object from the map in the correct reference system (WGS84, in our case) and then merge the polygons with the data

```
R> us1 <- map2SpatialPolygons(states, IDs=IDs,
+      proj4string = CRS("+proj=longlat +datum=WGS84"))
R> us2 <- SpatialPolygonsDataFrame(us1, USmelanoma)
```

The resulting object `us2` can now be plotted using the `spplot` function, see Figure 2.6. The colors correspond to the mortality rate, as shown in the color legend to the right of the map. We see that darker grey values corresponding to higher mortality rates appear in the southern costal states, both on the east and the west coast in good agreement with our earlier results.

Up to now we have primarily focused on the visualization of continuous variables. We now extend our focus to the visualization of categorical variables.

2.3.2 Chinese Health and Family Life

One part of the questionnaire the Chinese Health and Family Life Survey focuses on is the self-reported health status. Two questions are interesting for us. The first one is 'Generally speaking, do you consider the condition of your health to be excellent, good, fair, not good, or poor?'. The second question is

```
R> barplot(xtabs(~ R_happy, data = CHFLS))
```

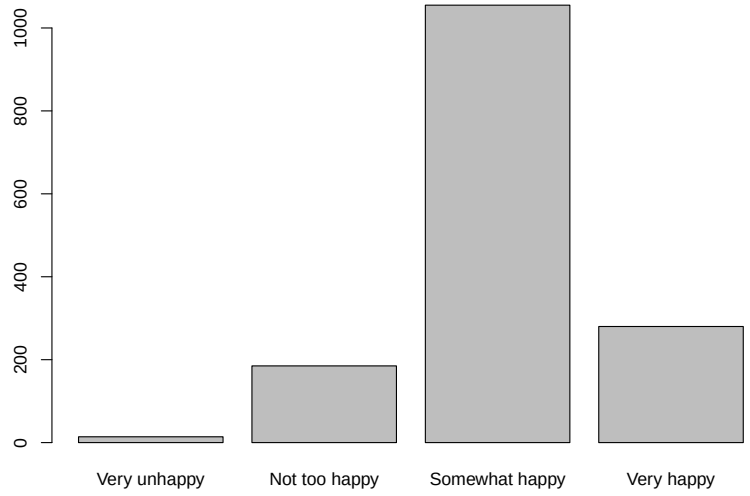

Figure 2.7 Bar chart of happiness.

'Generally speaking, in the past twelve months, how happy were you?'. The distribution of such variables is commonly visualized using barcharts where for each category the total or relative number of observations is displayed. Such a barchart can conveniently be produced by applying the `barplot` function to a tabulation of the data. The empirical density of the variable `R_happy` is computed by the `xtabs` function for producing (contingency) tables; the resulting barchart is given in Figure 2.7.

The visualization of two categorical variables could be done by conditional barcharts, i.e., barcharts of the first variable within the categories of the second variable. An attractive alternative for displaying such two-way tables are *spineplots* (Friendly, 1994, Hofmann and Theus, 2005, Chen et al., 2008); the meaning of the name will become clear when looking at such a plot in Figure 2.8.

Before constructing such a plot, we produce a two-way table of the health status and self-reported happiness using the `xtabs` function:

```
R> xtabs(~ R_happy + R_health, data = CHFLS)
```

```
                R_health
R_happy          Poor Not good Fair Good Excellent
  Very unhappy      2        7    4    1         0
  Not too happy     4       46   67   42        26
  Somewhat happy    3       77  350  459       166
  Very happy        1        9   40   80       150
```

```
R> plot(R_happy ~ R_health, data = CHFLS, ylab = "Happiness",
+       xlab = "Health")
```

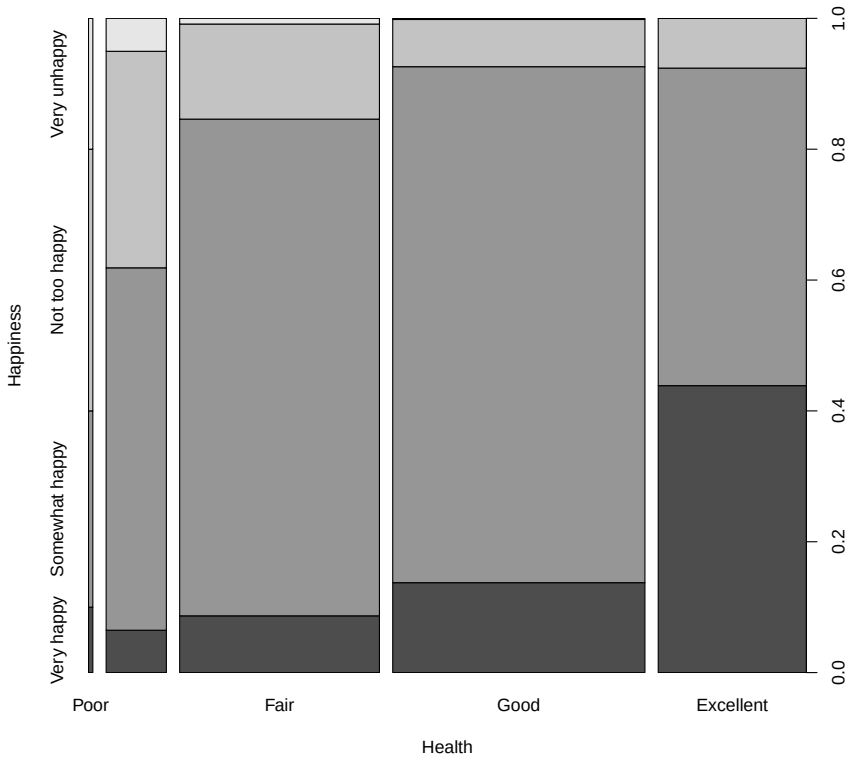

Figure 2.8 Spineplot of health status and happiness.

A *spineplot* is a group of rectangles, each representing one cell in the two-way contingency table. The area of the rectangle is proportional with the number of observations in the cell. Here, we produce a mosaic plot of health status and happiness in Figure 2.8.

Consider the right upper cell in Figure 2.8, i.e., the 150 very happy women with excellent health status. The width of the right-most bar corresponds to the frequency of women with excellent health status. The length of the top-right rectangle corresponds to the conditional frequency of very happy women given their health status is excellent. Multiplying these two quantities gives the area of this cell which corresponds to the frequency of women who are both very happy and enjoy an excellent health status. The conditional frequency of very happy women increases with increasing health status, whereas the conditional frequency of very unhappy or not too happy women decreases.

```
R> layout(matrix(1:2, ncol = 2))
R> plot(R_happy ~ log(R_income + 1), data = CHFLS,
+       ylab = "Happiness", xlab = "log(Income + 1)")
R> cdplot(R_happy ~ log(R_income + 1), data = CHFLS,
+         ylab = "Happiness", xlab = "log(Income + 1)")
R>
```

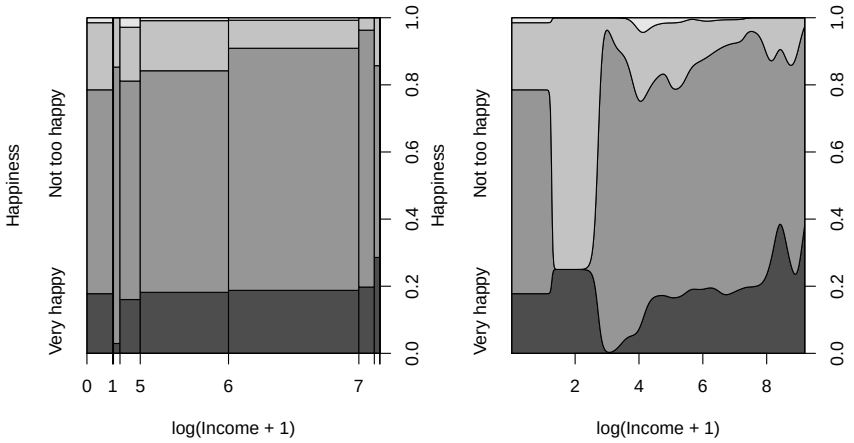

Figure 2.9 Spinogram (left) and conditional density plot (right) of happiness depending on log-income.

When the association of a categorical and a continuous variable is of interest, say the monthly income and self-reported happiness, one might use parallel boxplots to visualize the distribution of the income depending on happiness. If we were studying self-reported happiness as response and income as independent variable, however, this would give a representation of the conditional distribution of income given happiness, but we are interested in the conditional distribution of happiness given income. One possibility to produce a more appropriate plot is called *spinogram*. Here, the continuous x-variable is categorized first. Within each of these categories, the conditional frequencies of the response variable are given by stacked barcharts, in a way similar to spineplots. For happiness depending on log-income (since income is naturally skewed we use a log-transformation of the income) it seems that the proportion of unhappy and not too happy women decreases with increasing income whereas the proportion of very happy women stays rather constant. In contrast to spinograms, where bins, as in a histogram, are given on the x-axis, a *conditional density plot* uses the original x-axis for a display of the conditional density of the categorical response given the independent variable.

For our last example we return to scatterplots for inspecting the associa-

tion between a woman's monthly income and the income of her partner. Both income variables have been computed and partially imputed from other self-reported variables and are only rough assessments of the real income. Moreover, the data itself is numeric but heavily tied, making it difficult to produce 'correct' scatterplots because points will overlap. A relatively easy trick is to jitter the observation by adding a small random noise to each point in order to avoid overlapping plotting symbols. In addition, we want to study the relationship between both monthly incomes conditional on the woman's education. Such conditioning plots are called *trellis* plots and are implemented in the package **lattice** (Sarkar, 2014, 2008). We utilize the `xyplot` function from package **lattice** to produce a scatterplot. The formula reads as already explained with the exception that a third *conditioning* variable, `R_edu` in our case, is present. For each level of education, a separate scatterplot will be produced. The plots are directly comparable since the axes remain the same for all plots.

The plot shown in Figure 2.10 reveals several interesting issues. Some observations are positioned on a straight line with slope one, most probably an artifact of missing value imputation by linear models (as described in the data dictionary, see the documentation `?CHFLS`). Four constellations can be identified: both partners have zero income, the partner has no income, the woman has no income or both partners have a positive income.

For couples where the woman has a university degree, the income of both partners is relatively high (except for two couples where only the woman has income). A small number of former junior college students live in relationships where only the man has income, the income of both partners seems only slightly positively correlated for the remaining couples. For lower levels of education, all four constellations are present. The frequency of couples where only the man has some income seems larger than the other way around. Ignoring the observations on the straight line, there is almost no association between the income of both partners.

2.4 Summary of Findings

Using relatively straightforward graphical techniques only on the two sets of data considered in this chapter we have been able to uncover a number of important features of each data set;

Melanoma mortality Mortality is related only to the latitude of a state not to its longitude, mortality is higher for costal states than for land states, and the highest mortality is observed in the south costal states with latitude less than 32 degrees.

Health and family life We saw that happiness depends on health status. Women reported to be very happy more often when they also reported a good or excellent health status. The dependency of happiness on the income of the women seems to be less clear, but we conclude that, conditional on education, the income of wives and their husbands is highly correlated.

```
R> library("lattice")
R> xyplot(jitter(log(R_income + 0.5)) ~
+         jitter(log(A_income + 0.5)) | R_edu, data = CHFLS,
+         pch = 19, col = rgb(.1, .1, .1, .1),
+         ylab = "log(Wife's income + .5)",
+         xlab = "log(Husband's income + .5)")
```

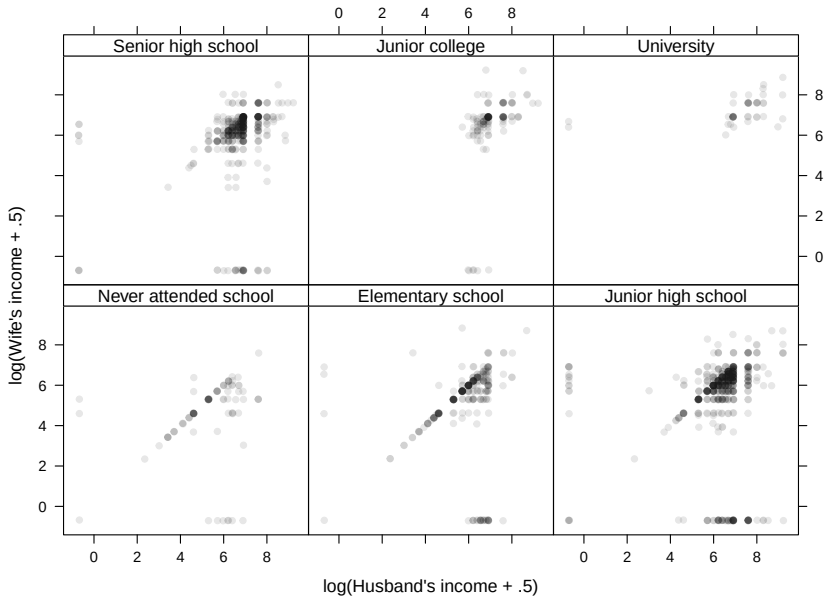

Figure 2.10 Scatterplot of jittered log-income of wife and husband, conditional on the wife's education.

2.5 Final Comments

Producing publication-quality graphics is one of the major strengths of the R system and almost anything is possible since graphics are programmable in R. Naturally, this chapter can be only a very brief introduction to some commonly used displays and the reader is referred to specialized books, most important Murrell (2005), Sarkar (2008), and Chen et al. (2008). Interactive 3D-graphics are available from package **rgl** (Adler and Murdoch, 2014).

Exercises

Ex. 2.1 The data in Table 2.1 are part of a data set collected from a survey of household expenditure and give the expenditure of 20 single men and 20 single women on four commodity groups. The units of expenditure are Hong Kong dollars, and the four commodity groups are

`housing` housing, including fuel and light,

`food` foodstuffs, including alcohol and tobacco,

`goods` other goods, including clothing, footwear, and durable goods,

`service` services, including transport and vehicles.

The aim of the survey was to investigate how the division of household expenditure between the four commodity groups depends on total expenditure and to find out whether this relationship differs for men and women. Use appropriate graphical methods to answer these questions and state your conclusions.

Table 2.1: `household` data. Household expenditure for single men and women.

housing	food	goods	service	gender
820	114	183	154	female
184	74	6	20	female
921	66	1686	455	female
488	80	103	115	female
721	83	176	104	female
614	55	441	193	female
801	56	357	214	female
396	59	61	80	female
864	65	1618	352	female
845	64	1935	414	female
404	97	33	47	female
781	47	1906	452	female
457	103	136	108	female
1029	71	244	189	female
1047	90	653	298	female
552	91	185	158	female
718	104	583	304	female
495	114	65	74	female
382	77	230	147	female
1090	59	313	177	female
497	591	153	291	male
839	942	302	365	male
798	1308	668	584	male
892	842	287	395	male
1585	781	2476	1740	male
755	764	428	438	male
388	655	153	233	male
617	879	757	719	male
248	438	22	65	male
1641	440	6471	2063	male

Table 2.1: `household` data (continued).

housing	food	goods	service	gender
1180	1243	768	813	male
619	684	99	204	male
253	422	15	48	male
661	739	71	188	male
1981	869	1489	1032	male
1746	746	2662	1594	male
1865	915	5184	1767	male
238	522	29	75	male
1199	1095	261	344	male
1524	964	1739	1410	male

Ex. 2.2 The data set shown in Table 2.2 contains values of seven variables for ten states in the US. The seven variables are

`Population` population size divided by 1000,

`Income` average per capita income,

`Illiteracy` illiteracy rate (% population),

`Life.Expectancy` life expectancy (years),

`Homicide` homicide rate (per 1000),

`Graduates` percentage of high school graduates,

`Freezing` average number of days per below freezing.

With these data

1. Construct a scatterplot matrix of the data labeling the points by state name (using function `text`).
2. Construct a plot of life expectancy and homicide rate conditional on average per capita income.

Ex. 2.3 Mortality rates per $100,000$ from male suicides for a number of age groups and a number of countries are given in Table 2.3. Construct side-by-side box plots for the data from different age groups, and comment on what the graphic tells us about the data.

Table 2.3: `suicides2` data. Mortality rates per $100,000$ from male suicides.

	A25.34	A35.44	A45.54	A55.64	A65.74
Canada	22	27	31	34	24
Israel	9	19	10	14	27
Japan	22	19	21	31	49

Table 2.3: `suicides2` data (continued).

	A25.34	A35.44	A45.54	A55.64	A65.74
Austria	29	40	52	53	69
France	16	25	36	47	56
Germany	28	35	41	49	52
Hungary	48	65	84	81	107
Italy	7	8	11	18	27
Netherlands	8	11	18	20	28
Poland	26	29	36	32	28
Spain	4	7	10	16	22
Sweden	28	41	46	51	35
Switzerland	22	34	41	50	51
UK	10	13	15	17	22
USA	20	22	28	33	37

Ex. 2.4 Flury and Riedwyl (1988) report data that give various length measurements on 200 Swiss bank notes. The data are available from package **mclust** (Fraley et al., 2014); a sample of ten bank notes is given in Table 2.4.

Table 2.4: `banknote` data (package **mclust**). Swiss bank note data.

Length	Left	Right	Bottom	Top	Diagonal
214.8	131.0	131.1	9.0	9.7	141.0
214.6	129.7	129.7	8.1	9.5	141.7
214.8	129.7	129.7	8.7	9.6	142.2
214.8	129.7	129.6	7.5	10.4	142.0
215.0	129.6	129.7	10.4	7.7	141.8
214.4	130.1	130.3	9.7	11.7	139.8
214.9	130.5	130.2	11.0	11.5	139.5
214.9	130.3	130.1	8.7	11.7	140.2
215.0	130.4	130.6	9.9	10.9	140.3
214.7	130.2	130.3	11.8	10.9	139.7
⋮	⋮	⋮	⋮	⋮	⋮

Use whatever graphical techniques you think are appropriate to investigate whether there is any 'pattern' or structure in the data. Do you observe something suspicious?

Ex. 2.5 The data in Table 2.5 were originally derived from a study reported

Table 2.2: `USstates` data. Socio-demographic variables for ten US states.

Population	Income	Illiteracy	Life.Expectancy	Homicide	Graduates	Freezing
3615	3624	2.1	69.05	15.1	41.3	20
21198	5114	1.1	71.71	10.3	62.6	20
2861	4628	0.5	72.56	2.3	59.0	140
2341	3098	2.4	68.09	12.5	41.0	50
812	4281	0.7	71.23	3.3	57.6	174
10735	4561	0.8	70.82	7.4	53.2	124
2284	4660	0.6	72.13	4.2	60.0	44
11860	4449	1.0	70.43	6.1	50.2	126
681	4167	0.5	72.08	1.7	52.3	172
472	3907	0.6	71.64	5.5	57.1	168

in Vuilleumier (1970) which investigated numbers of bird species in isolated 'islands' of paramo vegetation in the northern Andes. The aim of the study was to investigate how the number of species (`N`) is related to four other variables, `AR` (area of 'island' in thousands of square km), `EL` (elevation in thousands of m), `Dec` (distance from Ecuador in km) and `DNI` (distance to the nearest 'island' in km). Begin by constructing a scatterplot matrix of the data differentiating the islands on each panel by a different plotting symbol and on each diagonal panel showing the histogram of the associated variable. What can you conclude from this plot about how N is related to the other four variables?

Table 2.5: `birds` data. Birds in paramo vegetation.

	N	AR	EL	Dec	DNI
Chiles	36	0.33	1.26	36	14
LasPapas	30	0.50	1.17	234	13
Sumapaz	37	2.03	1.06	543	83
Tolima	35	0.99	1.90	551	23
Paramillo	11	0.03	0.46	773	45
Cocuy	21	2.17	2.00	801	14
Pamplona	11	0.22	0.70	950	14
Cachira	13	0.14	0.74	958	5
Tama	17	0.05	0.61	995	29
Batallon	13	0.07	0.66	1065	55
Merida	29	1.80	1.50	1167	35
Perija	4	0.17	0.75	1182	75
SantaMarta	18	0.61	2.28	1238	75
Cende	15	0.07	0.55	1380	35

Bibliography

Adler, D. and Murdoch, D. (2014), ***rgl**: 3D Visualization Device System (OpenGL)*, URL `http://rgl.neoscientists.org`, R package version 0.93.996.

Chen, C., Härdle, W., and Unwin, A., eds. (2008), *Handbook of Data Visualization*, Berlin, Heidelberg: Springer-Verlag.

Flury, B. and Riedwyl, H. (1988), *Multivariate Statistics: A Practical Approach*, London, UK: Chapman & Hall.

Fraley, C., Raftery, A. E., and Wehrens, R. (2014), ***mclust**: Model-based Cluster Analysis*, URL `http://www.stat.washington.edu/mclust`, R package version 4.3.

Friendly, M. (1994), "Mosaic displays for multi-way contingency tables," *Journal of the American Statistical Association*, 89, 190–200.

Hofmann, H. and Theus, M. (2005), "Interactive graphics for visualizing conditional distributions," Unpublished Manuscript.

Murrell, P. (2005), *R Graphics*, Boca Raton, Florida, USA: Chapman & Hall/CRC.

Pebesma, E. and Bivand, R. (2013), ***sp**: Classes and Methods for Spatial Data*, URL `http://CRAN.R-project.org/package=sp`, R package version 1.0-14.

Sarkar, D. (2008), *Lattice: Multivariate Data Visualization with R*, New York, USA: Springer-Verlag.

Sarkar, D. (2014), ***lattice**: Lattice Graphics*, URL `http://CRAN.R-project.org/package=lattice`, R package version 0.20-27.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, Connecticut: Graphics Press.

Vuilleumier, F. (1970), "Insular biogeography in continental regions. I. The northern Andes of South America," *The American Naturalist*, 104, 373–388.

Wilkinson, L. (1992), "Graphical displays," *Statistical Methods in Medical Research*, 1, 3–25.