大数据挖掘软件系列课之：R语言 数据可视化

肖韬

数据分析部分之：数据可视化

本部分内容对应参考书《A Handbook of Statistical Analyses Using R》(HSAUR书)第2章

通过两个数据——美国的恶性黑素瘤数据（US Malignant Melanoma）和中国人的健康与家庭生活调查数据（Chinese Health and Family Life）来介绍一些基本的但是重要的从数据中获取信息的图形展示技巧。

USmelanoma: 美国的恶性黑素瘤数据

Fisher and Belle (1993)报告了美国的每个州在1950–1969时间段内白人男性因为恶性黑素瘤致死的死亡率。这个数据包括各个州的因恶性黑素瘤致死的死亡数，每个州的中心点的经度和纬度，以及一个二元变量指示这个州是否临海。感兴趣的问题是：

- 临海和不临海的州的死亡率的比较
- 死亡率是如何受到纬度和经度的影响的？

CHFLS: 中国健康与家庭生活调查数据

- “中国健康与家庭生活调查”选择对中国的60个农村和城市进行抽样调查，以求在考虑地理分布的情况下尽可能的全面的反映当代中国的社会经济情况。
- 在每个采样地点通过官方登记处随机采集了83个成年人的信息，这些成年人的年龄范围在20至64岁之间，总共采集了5000个成年人的信息。这里，我们将我们的注意力放在有男性伴侣的妇女身上，她们在我们的数据表中有以下一些变量：

CHFLS: Chinese Health and Family Life Survey

R_edu：接受采访的妇女的受教育程度

R_income：接受采访的妇女的月收入（元）

R_health：接受采访的妇女的在过去一年里的健康状况

R_happy：接受采访的妇女的在过去一年里的快乐程度

A_edu：接受采访的妇女的男性伴侣的受教育程度

A_income：接受采访的妇女的男性伴侣的月收入（元）

这里，我们只侧重于用图形的方式来展示中国妇女以及她们的男性伴侣的一些健康与社会经济变量之间的关系。

初始的数据分析

据Chambers et al. (1983), “没有任何一种统计工具能比一个恰当选择的图形展示方式更好更强大”:

- 与其他展示方式比较起来，没有什么比一个经过精心设计的图表可以更加吸引观众的注意力并吊起观众的兴趣。
- 通过图表可视化展示的关系更容易被理解和记忆。
- 使用图表可以节省时间，因为统计方法的测量结果可以通过使用图表一目了然的展示出来。
- 相对与纯数据或者文字描述的信息，图表可以提供一个对所反映的问题的更全面、更综合、更均衡的理解。
- 图表可以揭示隐藏的事实和关系，可以激发和辅助思索和探索。

A Word of Warning

Carl Sagan（在他写的书《Contact》里）给出以下警示：

> 人类，非常善于察觉直观的细微模式，但却又同时非常不善于在头脑中想象出完全抽象的细微模式。

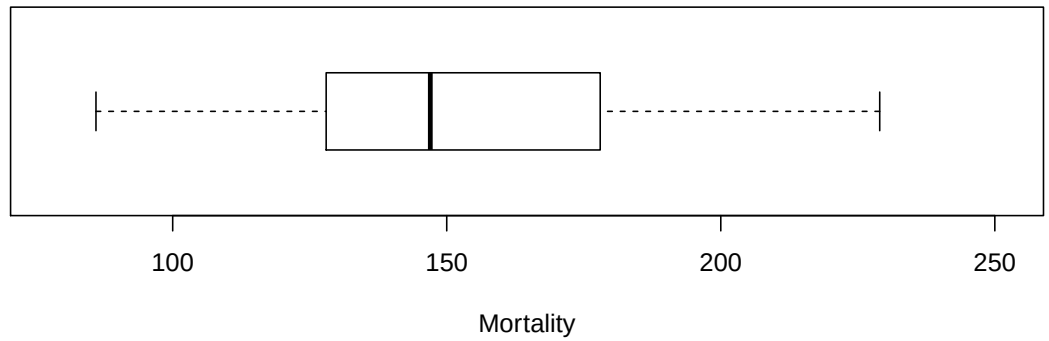
在恶性黑素瘤数据上使用箱线图（boxplot）和直方图（histogram）

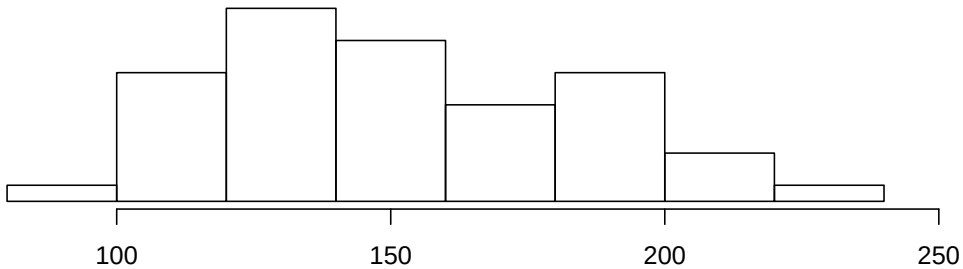
我们可以开始检查恶性黑色素瘤数据——通过构建直方图和箱线图来展示所有的死亡率数据。当使用这个相对简单的技术来展示两个图时，我们必须确保x轴在两个图中相同。如何做到这一点呢？ 我们可以通过计算死亡率数据的范围，然后稍后通过xlim参数在两种图形的画图函数中指定范围：

```
R> #import Melanoma Datasets
R> #USmelanoma<-read.csv("R/Datasets/USmelanoma.csv")
R> xr <- range(USmelanoma$mortality) * c(0.9, 1.1)
R> xr
```

```
[1]  77.4 251.9
```

```
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)
```

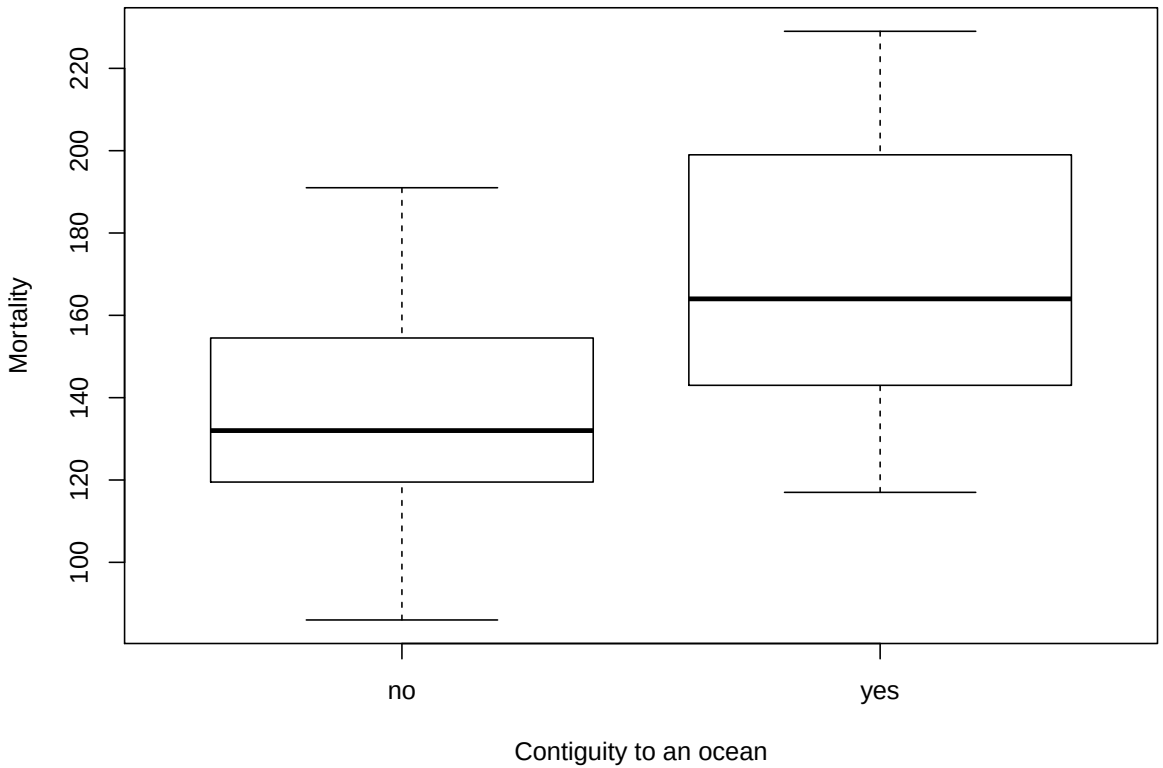




Malignant Melanoma: 州与州之间的比较

查看死亡率的分布特征是能带给我们一些信息，但是更有用的信息是对临海与不临海的州的死亡率的比较。我们可以对两种不同类型的州分别画出它们的直方图或者箱线图，并且并列起来进行比较。我们可以使用boxplot函数来比较一个连续性变量（这里是"死亡率"）的分布函数在一个离散型（属性型）变量的不同级别上的差异。

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

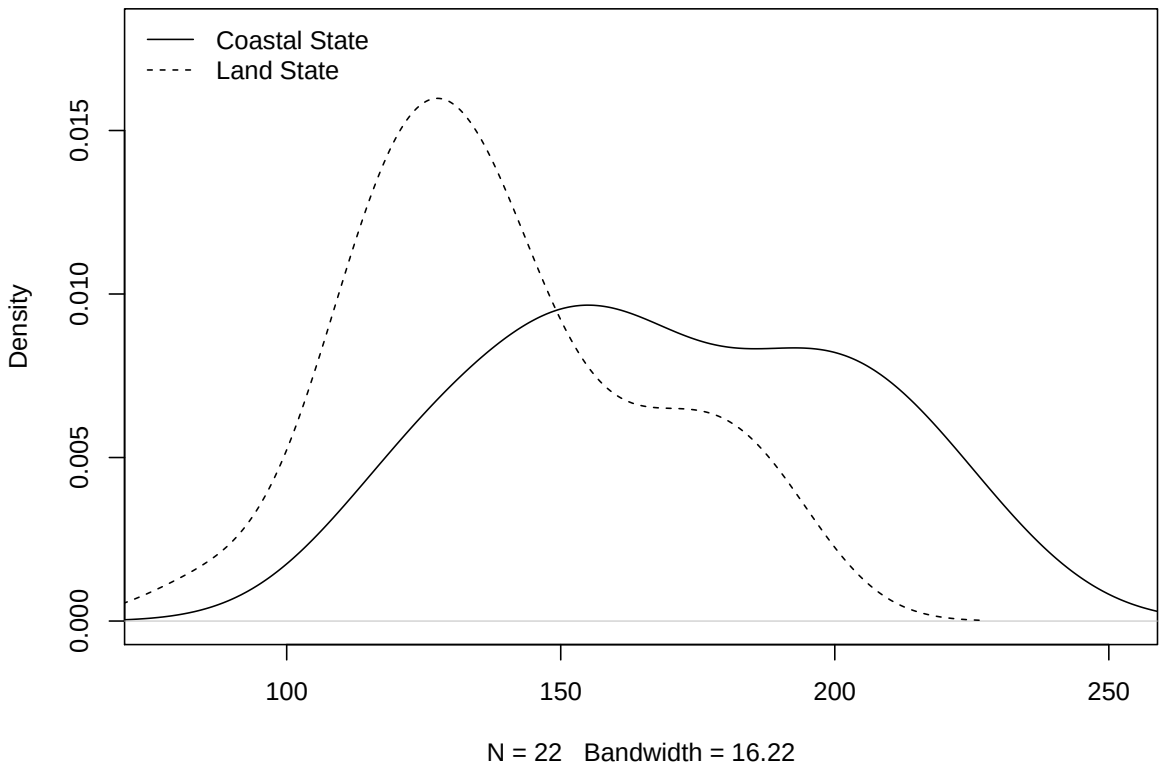


恶性黑素瘤: 分布密度图

直方图通常用于两个目的：计数和显示 变量的分布；Wilkinson (1992)说，“它们对两个目的的实现都不是很有效，因为直方图通常会误导的显示分布——它们依赖于所画区间的数量。

另一种方法是较严谨的直接估计变量的概率密度函数，然后把估计的概率密度函数绘制出来。

```
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))
R> lines(dno, lty = 2)
R> legend("topleft", lty = 1:2, legend = c("Coastal State",
+         "Land State"), bty = "n")
```

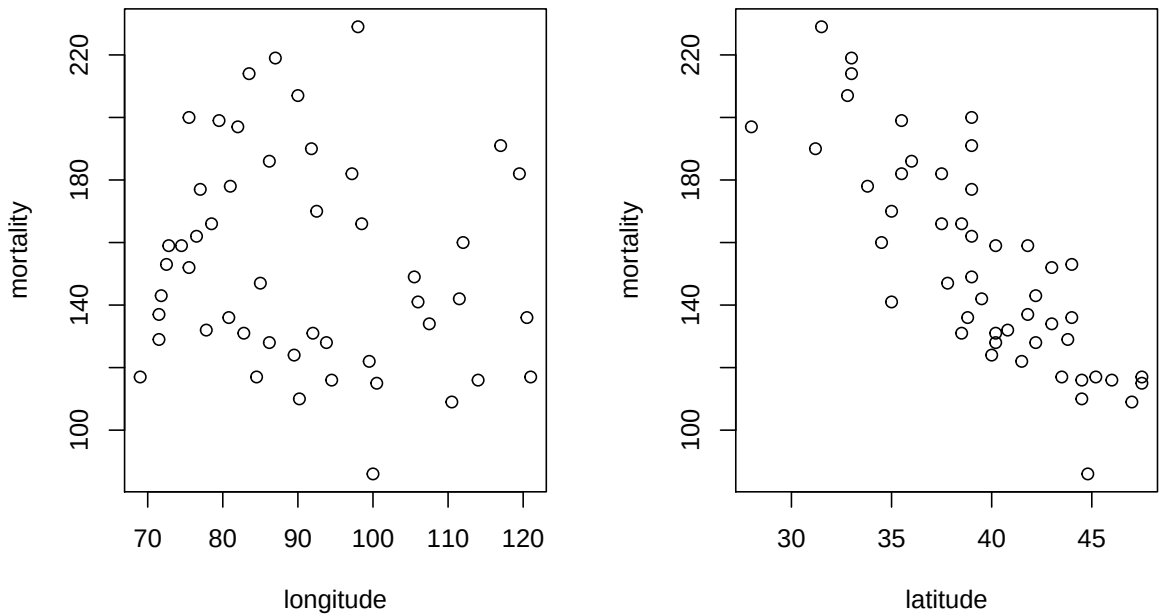


恶性黑素瘤: 与地理位置的关系

现在，我们继续来查看死亡率是如何与地理位置（通过州中心点的纬度和经度来表征）相关联的。这里我们使用的主要图表是散点图（scatter plot）。简单的$x-y$型关系图这种散点图至少在18世纪就开始被使用。据Tufte (1983)，这种图形展示方式具有以下优点：

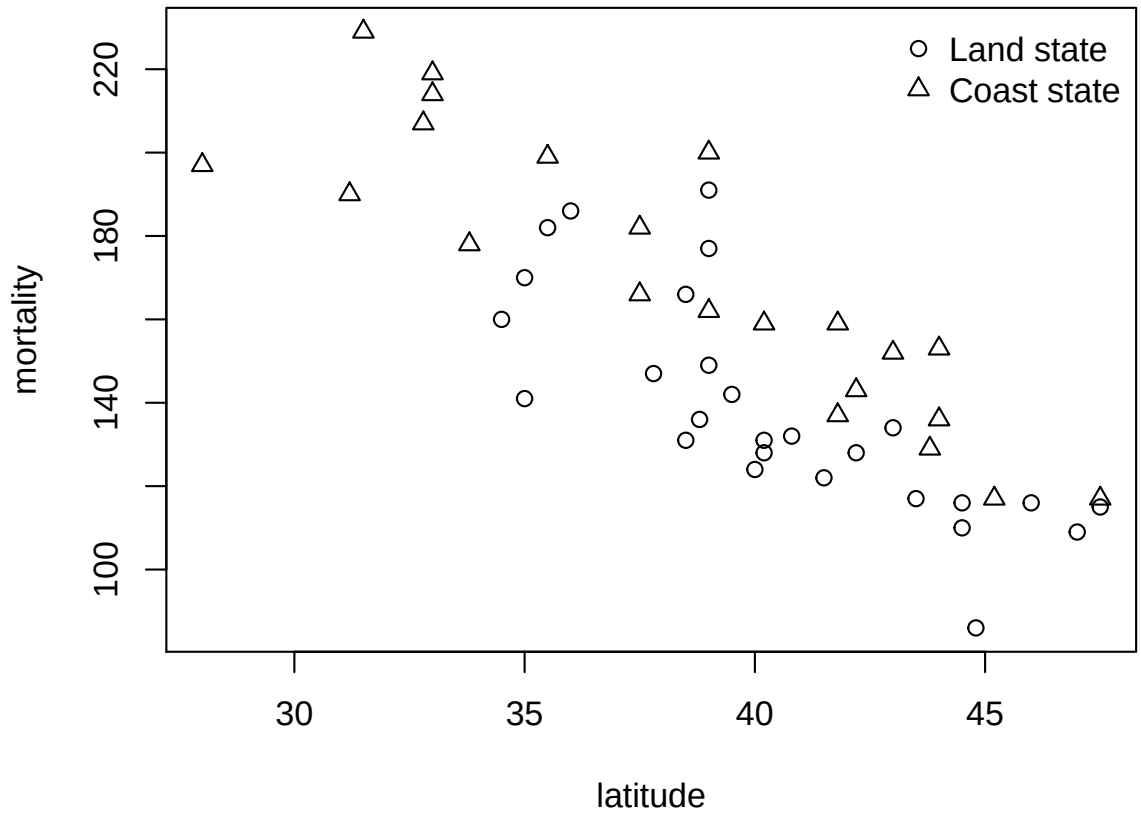
> $x-y$型关系图，是一种最简单的形式的散点图，这种图是所有图形展示方式设计中最有价值的一种；关系图能将至少两个变量进行比较，它能激励甚至请求观看者对这些变量之间的因果关系进行评价–它将观察到的x和y之间的关联关系进行直接展示并拷问观察者：是否会有x导致了y的因果关系？

```
R> layout(matrix(1:2, ncol = 2))
R> plot(mortality ~ longitude, data = USmelanoma)
R> plot(mortality ~ latitude, data = USmelanoma)
```



由于死亡率显然仅与纬度相关，我们现在可以对临海和非临海的州分别绘制展示死亡率和纬度的关系的散点图。我们不必像上图那样用两个图片来分别显示临海州和非临海州的图，而是可以利用不同的绘图符号将对应两种类型的州的散点图绘制在同一幅图片中。

```
R> plot(mortality ~ latitude, data = USmelanoma,
+      pch = as.integer(USmelanoma$ocean))
R> legend("topright", legend = c("Land state",
+      "Coast state"), pch = 1:2, bty = "n")
```



从这张散点图可看出北部的州，死亡率最低。如果纬度相同，沿海的州会比陆地州的死亡率更高。拥有最高死亡率的州是那些南部的维度低于可以观察到南沿海州的死亡率低于32°的临海州：

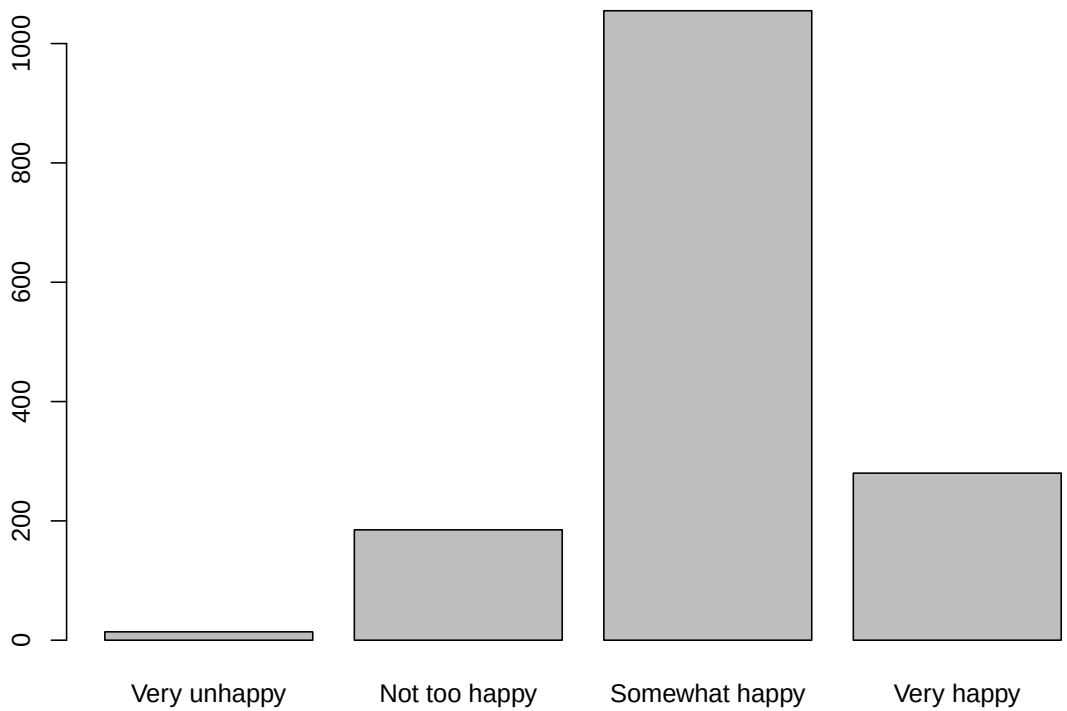
```
R> subset(USmelanoma, latitude < 32)
```

	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*

中国人的健康与家庭生活调查

- “中国人的健康与家庭生活”的调查问卷的一部分侧重于自我报告的健康状态。
- 第一个问题是：“总的来说，你认为你的健康状况是属于以下哪种：优秀、良好、一般、不好、糟糕？”
- 第二个问题是：“总的来说，在过去的12个月中，你有多快乐：非常快乐、一般快乐、不是很快乐、非常不快乐？”.
- 这样的类别变量（或者叫属性型变量，categorical variable）的数据分布情况通常使用柱状图（bar charts）来图形化显示变量每一个类别（category）的总读数或者相对其他类别的相对读数。

```
R> #import Chinese Health and Family Life Survey Dataset
R> #CHFLS<-read.csv("R/Datasets/CHFLS.csv")
R> barplot(xtabs(~ R_happy, data = CHFLS))
```



中国人的健康与家庭生活调查: 两个变量的关系

两个类别变量的可视化可以用条件柱状图完成，即 以第二个变量的每一个类别作为条件分别绘制第一个变量的柱状图。显示此类二维图表的一种有吸引力的替代方法 是*spineplots* (Friendly, 1994, Hofmann and Theus, 2005, Chen et al., 2008)。

在构建这种图之前，我们使用xtabs函数生成一个自我报告的幸福感和健康状况这两个变量的二维频数列联表：

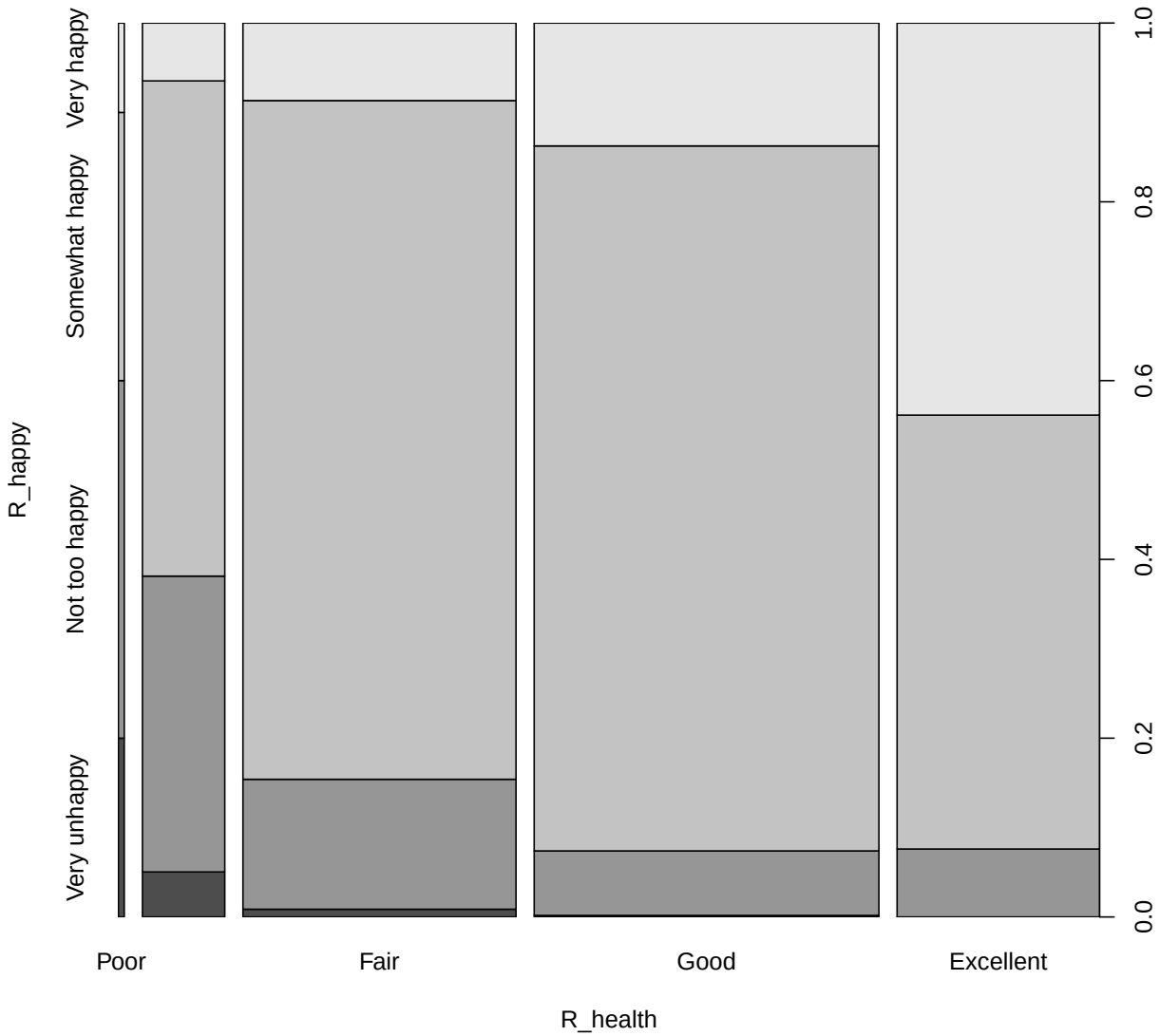
```
R> xtabs(~ R_happy + R_health, data = CHFLS)
```

	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

Chinese Health and Family Life: spineplots

脊柱图（*spineplot*）是一组矩形，每个矩形代表双向列联表中的一个单元格。每个矩形的面积与每个单元格内观察值频数成正比。以下我们绘制了关于健康状况和幸福感两个变量的脊柱图：

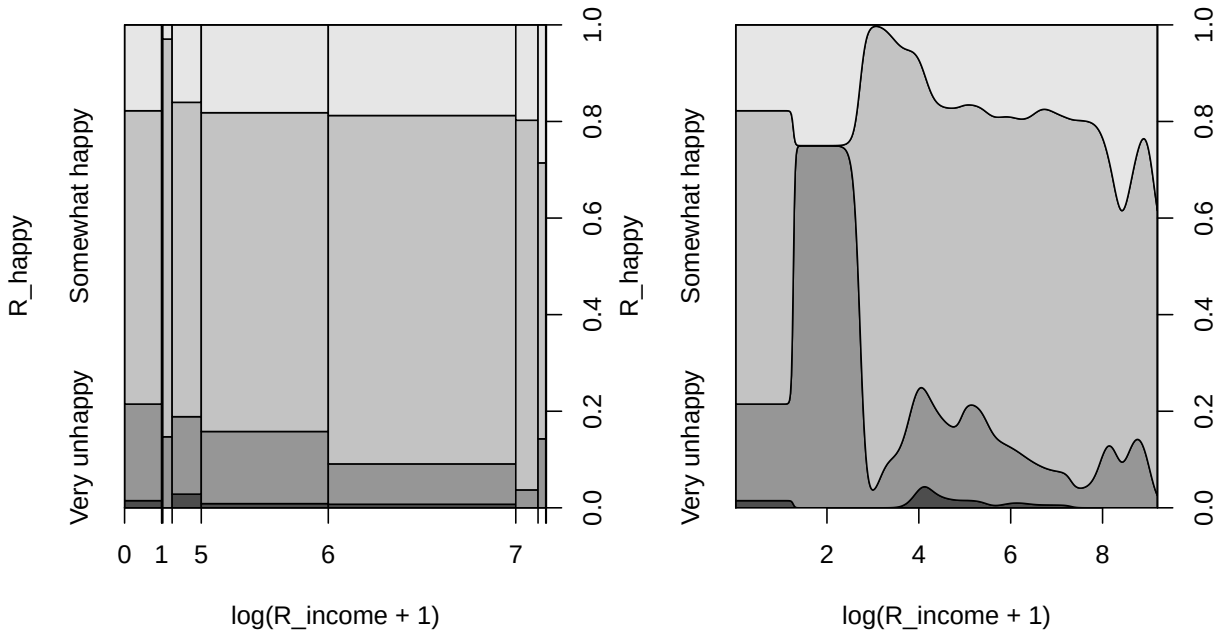
```
R> plot(R_happy ~ R_health, data = CHFLS)
```



Chinese Health and Family Life: spinogram

当我们想要可视化一个分类变量和一个连续变量的关系时，应该如何做呢？比如说，我们对妇女的月收入与其自我报告的幸福感之间是什么样的关系感兴趣：比如，在不同的收入值的条件下，幸福感的分布是如何变化的。因为收入值是连续分布的，如果我们将连续分布的收入值划分成不同的区间（这样连续型分布的收入值被转变为一种离散的分类型变量），那么我们就可以使用类似于可视化两个分类型变量的脊柱图（spineplot）的方式来可视化这种关系，我们把这种图称为脊髓图（*spinogram*）。这时，连续的自变量x会首先被分类。在以每个类别作为满足条件时，因变量（响应变量）y的频率由堆叠的柱状图给出，与脊柱图（spineplot）类似。以下除了脊髓图外，我们也通过*cdplot*函数给出使用连续型而不是分类型的月收入作为x轴的条件密度图作为对比。

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



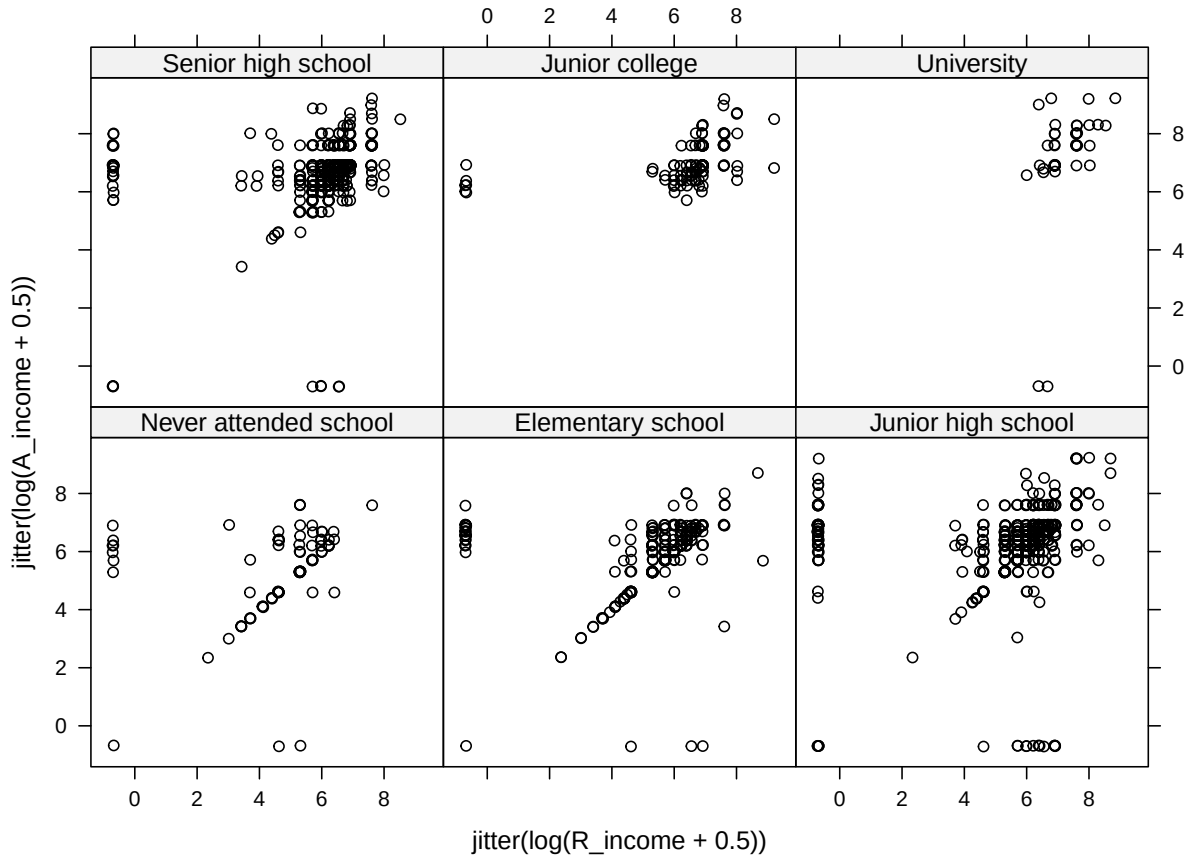
Chinese Health and Family Life: conditional plots

另外两个流行的画图扩展包:

- library(lattice)
- library(ggplot2)

作为最后一个示例，我们使用*lattice* (Sarkar, 2014, 2008)包中的格状图（*trellis plots*）函数*xyplot*来可视化在妇女的不同教育水平（这个变量是个离散型分类型的变量）的条件下，妇女的月收入与她的男性伴侣的月收入（两个变量都是连续型的变量）之间的关系。

```
R> xyplot(jitter(log(A_income + 0.5)) ~
+     jitter(log(R_income + 0.5)) | R_edu, data = CHFLS)
```



从图中可以识别出四个“星座”：伴侣双方都是没有收入，男性伴侣没有收入，女性没有收入，或者伴侣双方都有正收入。

Summary

能制作出版物级别的图形是R的主要优势之一 。可以说，在R中几乎可以制作任何图形，因为 在R中进行图形制作是编程的。本讲义的内容只提供了一些对常见关系进行可视化的常用图形，大家如果有兴趣，可以阅读一些专业的介绍在R中如何制图的书籍，比如： cite HSAUR：Murrell2005， cite HSAUR：Sarkar2008和 Chen et al. (2008)。另外值得一提的是， *rgl* (Adler and Murdoch, 2014)可以制作交互式的3D图形。

References

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

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, London: Chapman & Hall/CRC.

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

Fisher, L. D. and Belle, G. V. (1993), *Biostatistics. A Methodology for the Health Sciences*, New York, USA: John Wiley & Sons.

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.

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.

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