把数据化“宽”为“长”

概述

- ggplot2进行数据分组时必须根据行，而不能根据列。例如可以把钻石根据颜色分组，却不能把钻石的克拉数和价格变量分为两组。所以有时需要把“宽”的数据变成“长”的数据，即变量不是放在各个列上，而是排成一列，每个变量都占其中的几行。
- 可以使用reshape2包的melt()。

多重时间序列

- 数据集：economics：包含失业人数（unemploy）和失业周数的中位数（uempmed）。预判这两个变量可能相关。
- 将原数据（左表）melt成“长”的形式（右表）。ggplot2用“长”数据做图更方便。

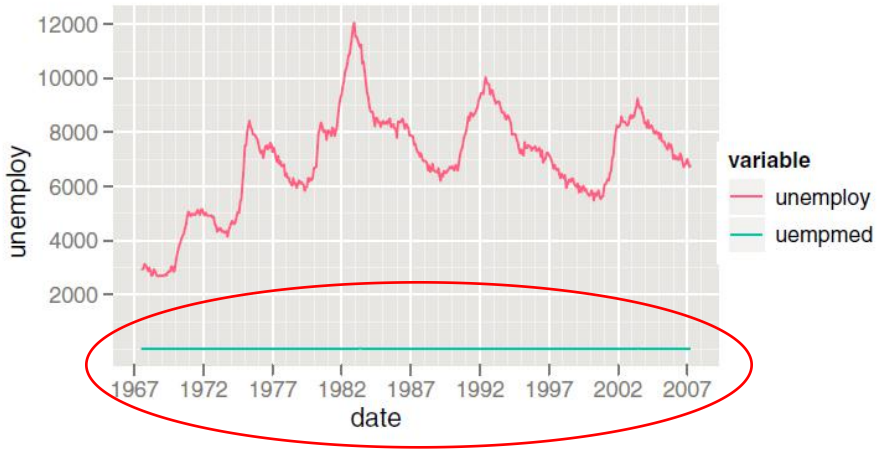
date	pce	pop
1967-06-30	508	198,712
1967-07-31	511	198,911
1967-08-31	517	199,113
1967-09-30	513	199,311
1967-10-31	518	199,498
1967-11-30	526	199,657

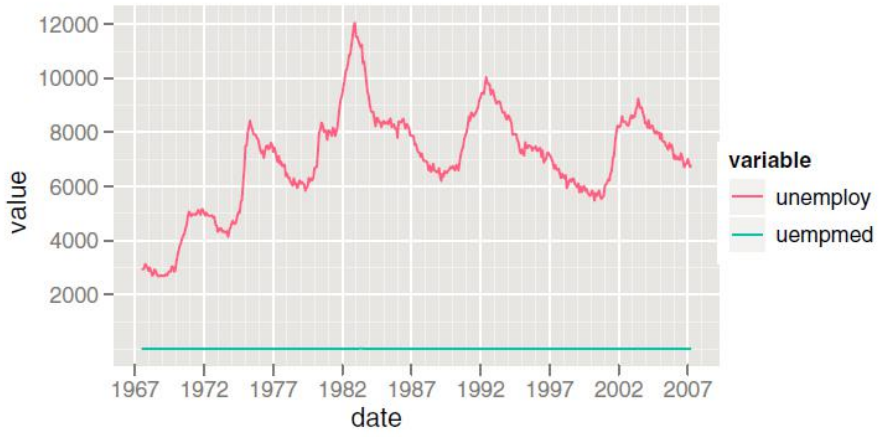
date	variable	value
1967-06-30	pce	508
1967-07-31	pce	511
1967-08-31	pce	517
1967-09-30	pce	513
1967-10-31	pce	518
1967-11-30	pce	526
1967-06-30	pop	198,712
1967-07-31	pop	198,911
1967-08-31	pop	199,113
1967-09-30	pop	199,311
1967-10-31	pop	199,498
1967-11-30	pop	199,657

```
ggplot(economics, aes(date)) +
  geom_line(aes(y = unemploy, colour = "unemploy")) +
  geom_line(aes(y = uempmed, colour = "uempmed")) +
  scale_colour_hue("variable") #左图

emp <- melt(economics, id = "date",
  measure = c("unemploy", "uempmed"))
qplot(date, value, data = emp, geom = "line", colour = variable) #右图
```

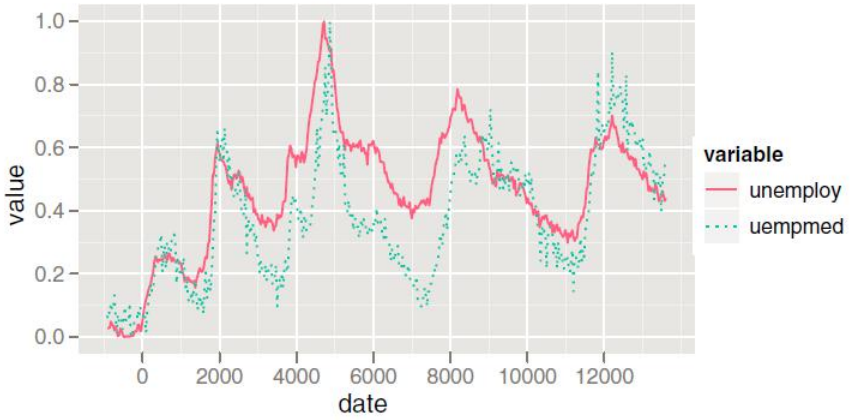
- 两种方法画出的图是一样的。但它们有一个共同问题：两个变量取值差异太大，所以uempmed成为了图形底部的一条平坦的线。ggplot2不允许画带有两个不同坐标轴的图，因为这样的图有误导性。
- 有两个解决方法（见下页）：
  1. 把数据标度调整到相同范围
  2. 使用自由标度的分面

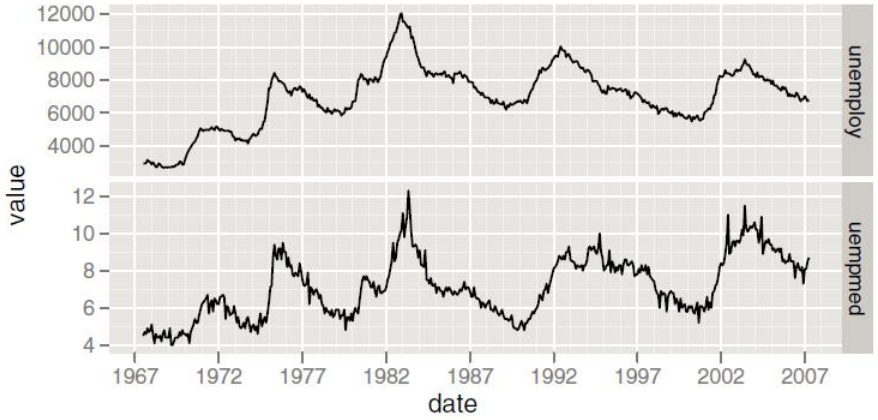




```
library(plyr)
range01 <- function(x) {
  rng <- range(x, na.rm = TRUE)
  (x - rng[1]) / diff(rng)
}
emp2 <- ddply(emp, .(variable), transform, value = range01(value))
qplot(date, value, data = emp2, geom = "line",
  colour = variable, linetype = variable) #左图：把数据标度调整到相同范围。

qplot(date, value, data = emp, geom = "line") +
  facet_grid(variable ~ ., scales = "free_y") #右图：使用自由标度的分面，把数据画在不同的分面上。
```





平行坐标图

- 对于“长”数据，可以画平行坐标图。
- 平行坐标图以varaible变量为x轴表示变量名，以value为y轴表示变量取值。
- 例：下页平行坐标图使用840部电影的用户评分数据。数据集共10个变量，每个变量对应电影在1到10分的比例。这个数据集各个变量标度已经是统一的，所以不必再标准化。

```
library(reshape2)
popular <- subset(movies, votes > 1e4)
ratings <- popular[, 7:16]
ratings$.row <- rownames(ratings)
molten <- melt(ratings, id = ".row")
pcp <- ggplot(molten, aes(variable, value, group = .row))
pcp + geom_line() #左上：观测数目太多，看不出比例
pcp + geom_line(colour = alpha("black", 1 / 20)) #右上：使用透明度，但并不明显
jit <- position_jitter(width = 0.25, height = 2.5)
pcp + geom_line(position = jit) #左下：使用扰动，效果不明显
pcp + geom_line(colour = alpha("black", 1 / 20), position = jit) #右下：使用扰动和透明度
```