ggplot2工具箱

本节内容

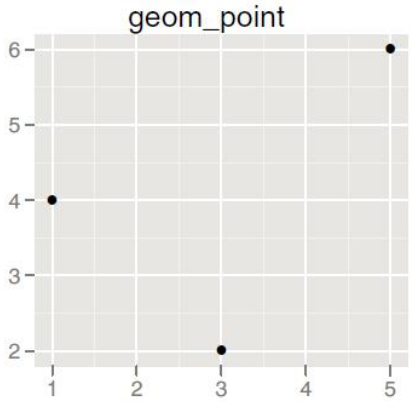
- 之前已经介绍了图层含义以及添加图层的方法，本节概述可用的几何对象和统计变换。包括：
  - 基本数据类型
  - 展示分布
  - 散点图中的遮盖绘制问题
  - 统计摘要
  - 添加图形注解
  - 绘制含权数据

图层叠加策略

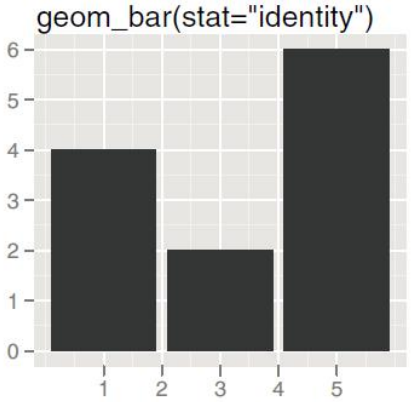
- 图层有三种：
  1. 用于展示数据本身（**data**）。
     - 辨识数据的整体结构、局部结构、离群点。
     - 在数据探索的初始阶段，本层通常是唯一的图层。
  2. 用于展示数据的统计摘要（**summary**）
     - 用于展示模型的预测效果。
     - 绘制在数据层之上。
  3. 用以添加额外的元数据（**metadata**）、上下文信息和注解
     - 元数据层展示背景上下文，帮助我们理解原始数据。元数据既可以作为前景也可以作为背景。地图经常作为空间数据的背景层。背景元数据不应影响主数据展示，因此它往往被放置在主数据下层，配色不能突出，“想看就看看到，不想看就看不到”。
     - 元数据也可以用来强调数据中的重要特征，比如为离群点加上解释性的标签。这时元数据图层是最后绘制的图层。

基本图形类型（1/2）

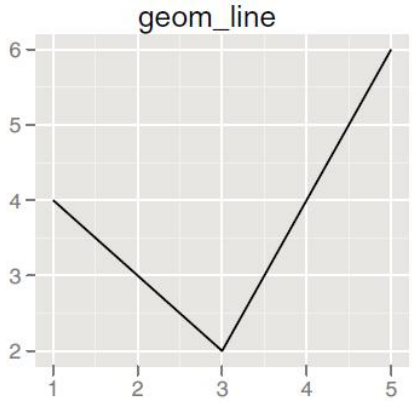
散点图



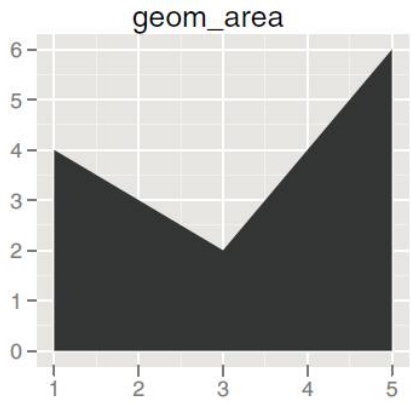
条形图



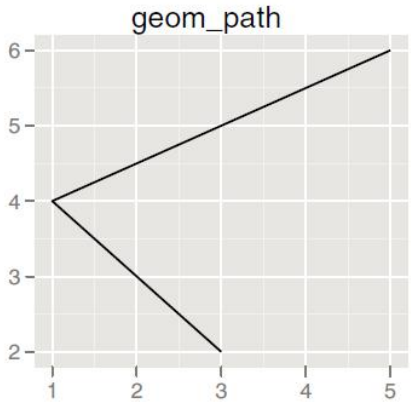
线条图



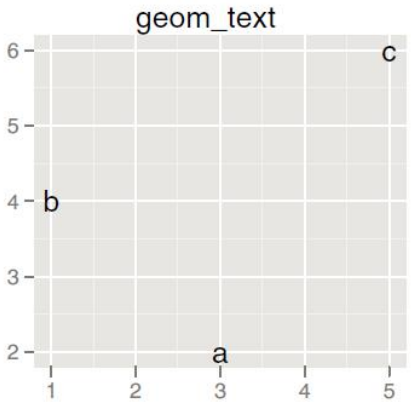
面积图



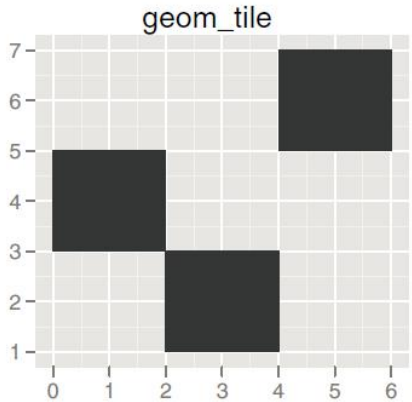
路径图



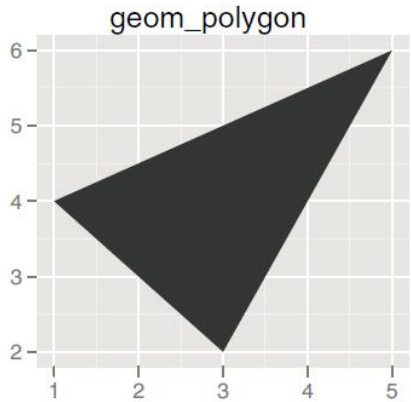
带标签的散点图



色深图/水平图



多边形图



基本图形类型（2/2）

- 上页图的代码

```
df <- data.frame(
  x = c(3, 1, 5),
  y = c(2, 4, 6),
  label = c("a","b","c")
)
p <- ggplot(df, aes(x, y, label = label)) +
  xlab(NULL) + ylab(NULL)
p + geom_point() + labs(title = "geom_point")
p + geom_bar(stat="identity") + labs(title = "geom_bar(stat=\"identity\")")
p + geom_line() + labs(title = "geom_line")
p + geom_area() + labs(title = "geom_area")
p + geom_path() + labs(title = "geom_path")
p + geom_text() + labs(title = "geom_text")
p + geom_tile() + labs(title = "geom_tile")
p + geom_polygon() + labs(title = "geom_polygon")
```

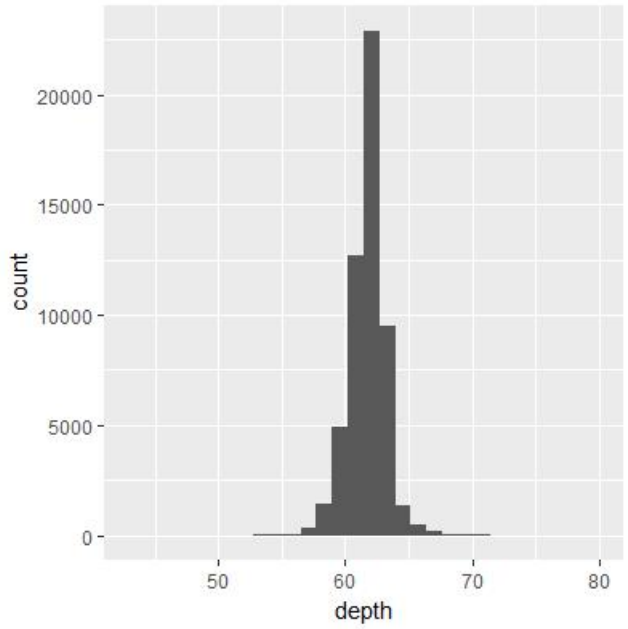
展示数据分布

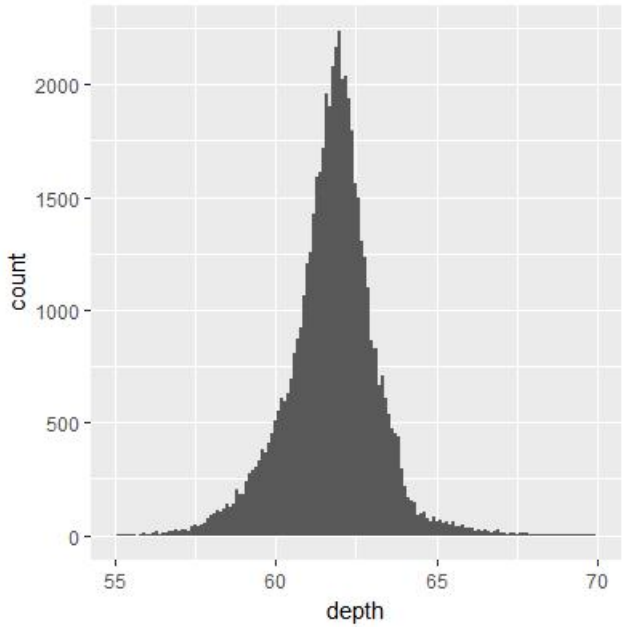
展示数据分布

- 有一些几何对象可以用于展示数据的分布，具体使用哪种取决于分布的维度、分布是连续型还是离散型、条件分布还是联合分布。

直方图

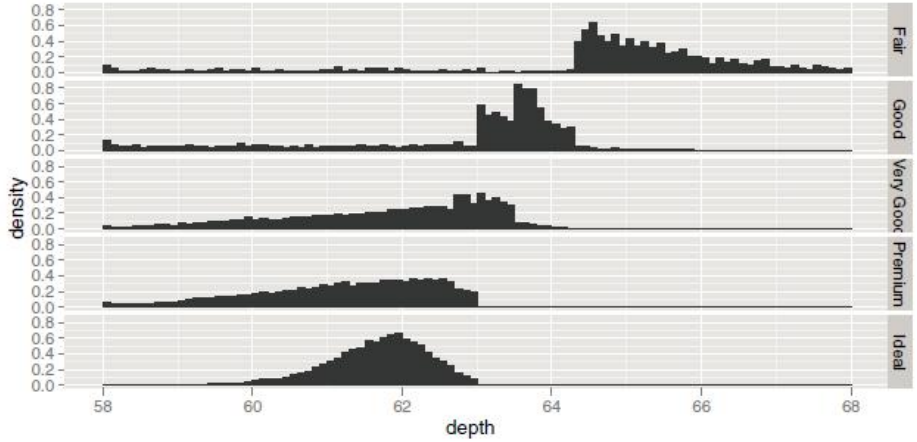
- 对于一维连续型分布，最重要的几何对象是直方图。
  - qplot(depth, data=diamonds, geom="histogram")#左图
  - qplot(depth, data=diamonds, geom="histogram", **xlim=c(55, 70)**, **binwidth=0.1**) #右图：对x轴进行放大，并选取了更小的组距

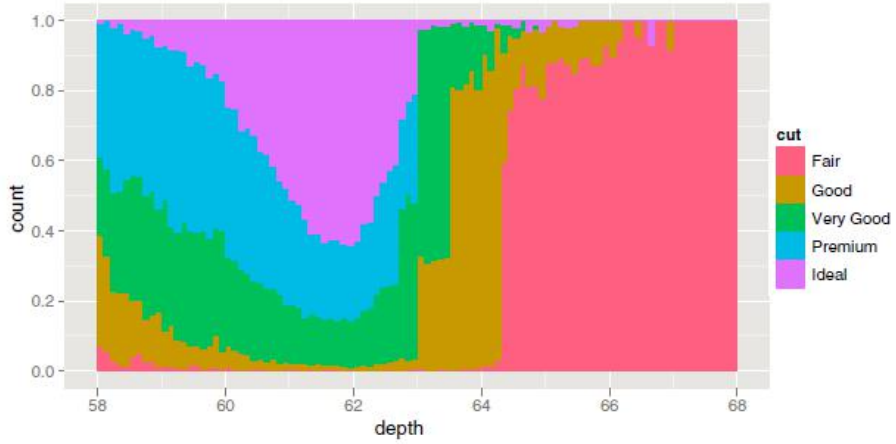


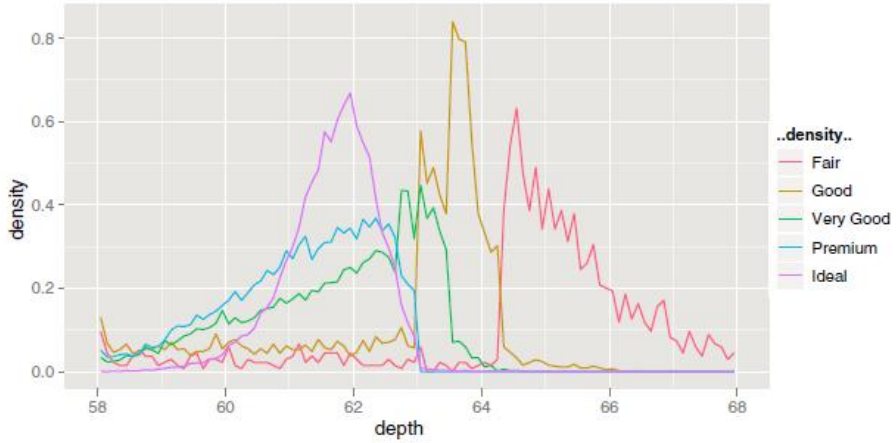


分布的跨组比较

- 有多种方式可以实现分布的跨组比较：同时绘制多个小的直方图、条件密度图、频率多边形。详见下页。







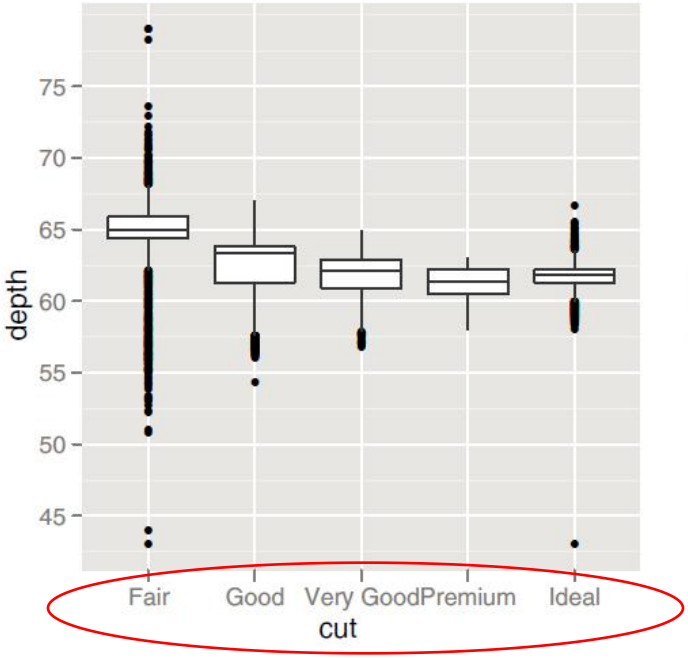
- 钻石数据切割和深度分布的三种视图：
  - 分面直方图、
  - 条件密度图、
  - 频率多边形图。
- 它们均显示出：随着钻石质量提高，分布逐渐向左并愈发对称。
- 代码：
  - depth_dist <- ggplot(diamonds, aes(depth)) + xlim(58, 68)
  - depth_dist + geom_histogram(aes(y = ..density..), binwidth = 0.1) + facet_grid(cut ~ .)
  - depth_dist + geom_histogram(aes(fill = cut), binwidth = 0.1, position = "fill")
  - depth_dist + geom_freqpoly(aes(y = ..density.., colour = cut), binwidth = 0.1)

几何对象（geom）和统计变换（stat）

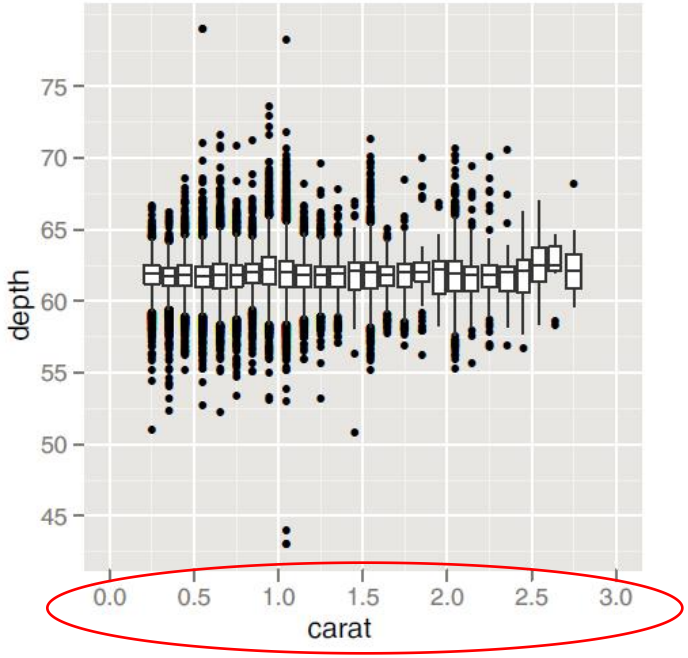
- 和分布相关的许多几何对象都是以几何对象（geom）/统计变换（stat）的形式成对出现的。
- 这些几何对象中大多数本质都是别名：一个基本几何对象结合一个统计变换，即可绘制出想要的图形。
- 实例
  - 箱线图
  - geom_jitter：离散型分布上添加随机噪声避免遮盖绘制
  - geom_density：基于核平滑方法得到的频率多边形。

箱线图

- 例：
  - qplot(cut, depth, data=diamonds, geom="boxplot")#左图：离散型
  - library(plyr)
  - qplot(carat, depth, data=diamonds, geom="boxplot", **group =** round_any(carat, 0.1, floor), xlim=c(0, 3))#右图：连续型



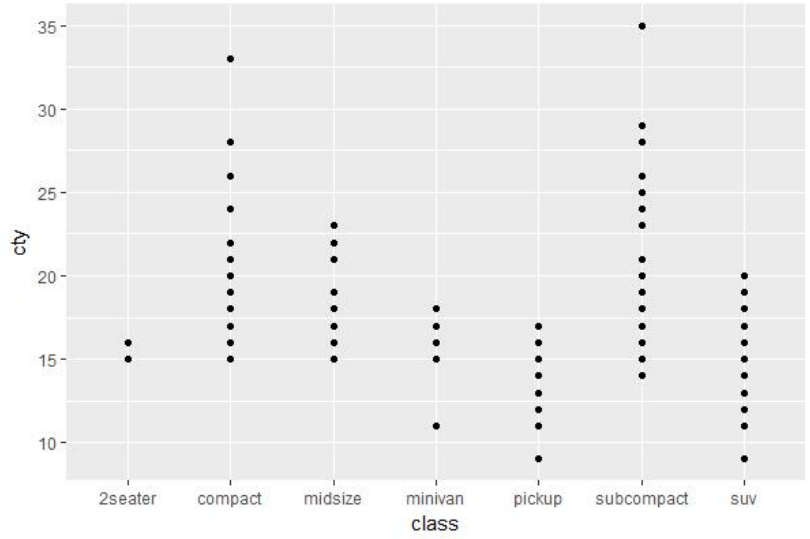
离散型

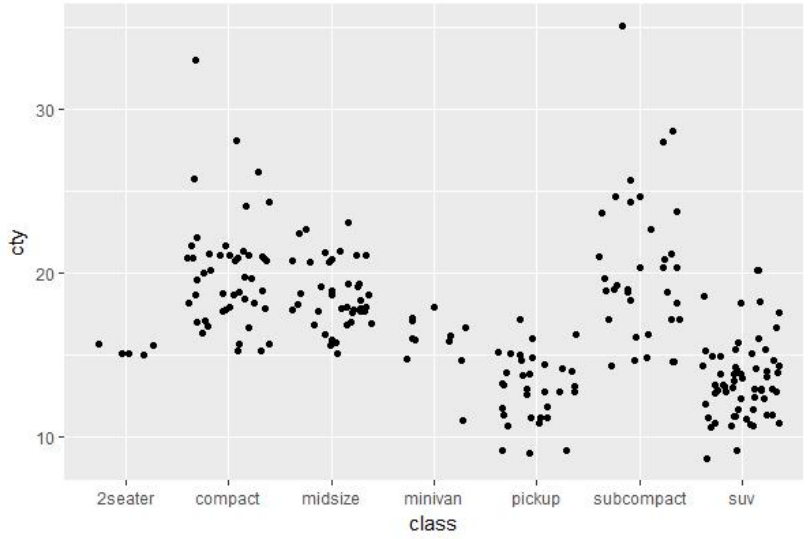


对于**连续型变量**，必须设置**group**图形属性以便得到多个箱线图。

geom_jitter

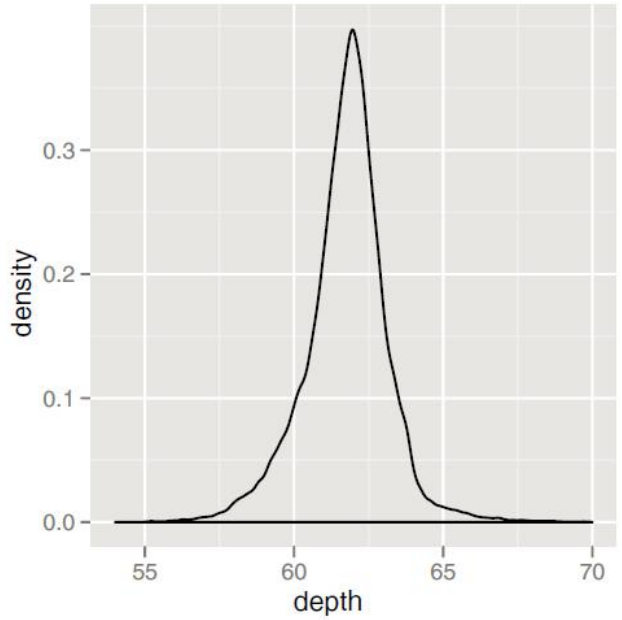
- geom_jitter：离散型分布上添加随机噪声避免遮盖绘制。例：
  - qplot(class, cty, data=mpg) #左：常规散点图
  - qplot(class, cty, data=mpg, **geom="jitter"**)#右

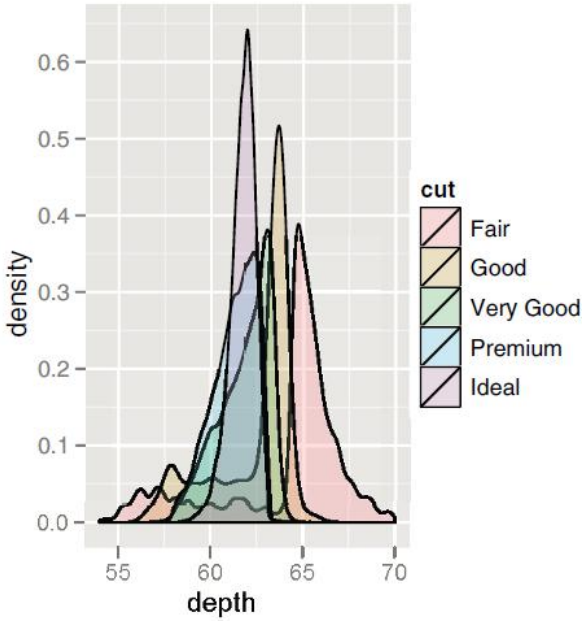




geom_density

- geom_density：基于核平滑方法得到的频率多边形。仅在已知潜在的密度分布为平滑、连续且无界的时候使用这种密度图。例：
  - qplot(depth, data=diamonds, geom="density", xlim = c(54, 70))#左
  - qplot(depth, data=diamonds, geom="density", xlim = c(54, 70), fill = cut, alpha = I(0.2)) #右图





处理遮盖绘制问题

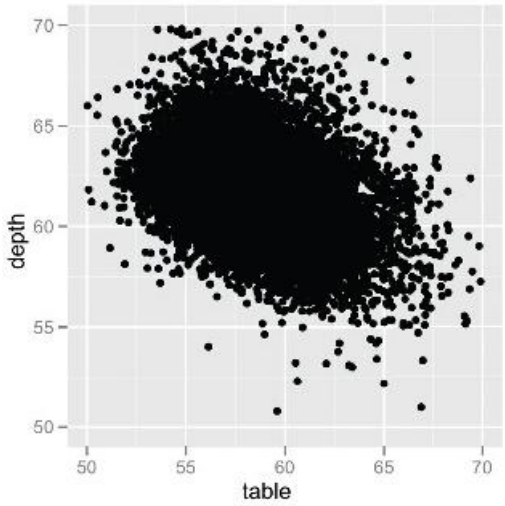
小规模的遮盖绘制

- 当数据量很大时，散点图中的点会出现重叠现象。（在“数据可视化基础篇”我们研究过这个问题）
- **小规模**的遮盖绘制问题可以通过**绘制更小的点**加以缓解。例：
  - df <- data.frame(x = rnorm(2000), y = rnorm(2000))
  - norm <- ggplot(df, aes(x, y))
  - norm + geom_point()
  - norm + geom_point(shape = 1)
  - norm + geom_point(shape = ".") # Pixel sized

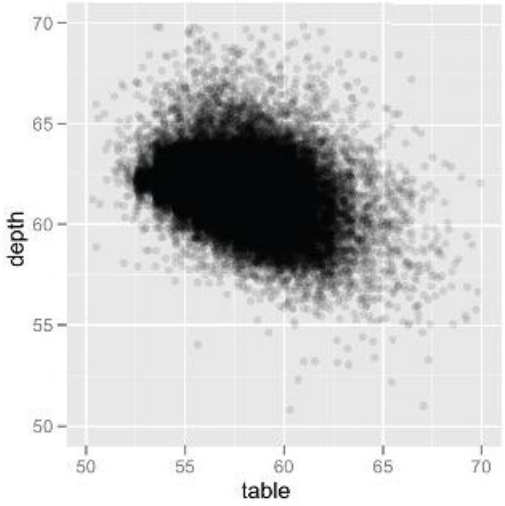
大数据集的遮盖绘制：α混合

- 大数据集可以使用α混合（调整透明度）让点呈现透明效果。
- α是一个比值，代表一个位置需要多少个点才能变成完成不透明。R中最小的透明度是1/256，所以对于严重的遮盖绘制，这种方法效果并不会太好。例：
  - td <- ggplot(diamonds, aes(table, depth)) + xlim(50, 70) + ylim(50, 70)
  - jit <- position_jitter(width = 0.5) #对于**离散性的数据，首先可以在点上增加随机扰动来减轻重叠**，之后再用α混合。
  - td + geom_jitter(position = jit)
  - td + geom_jitter(position = jit, colour = **alpha**("black", **1/10**))
  - td + geom_jitter(position = jit, colour = **alpha**("black", **1/50**))
  - td + geom_jitter(position = jit, colour = **alpha**("black", **1/200**))
  - #图在下页

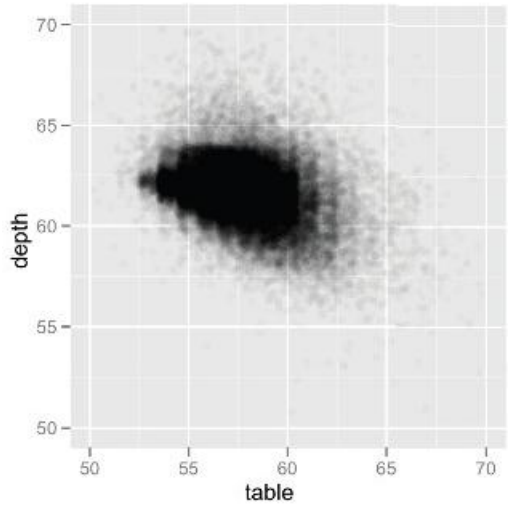
geom jitter with horizontal jitter of 0.5



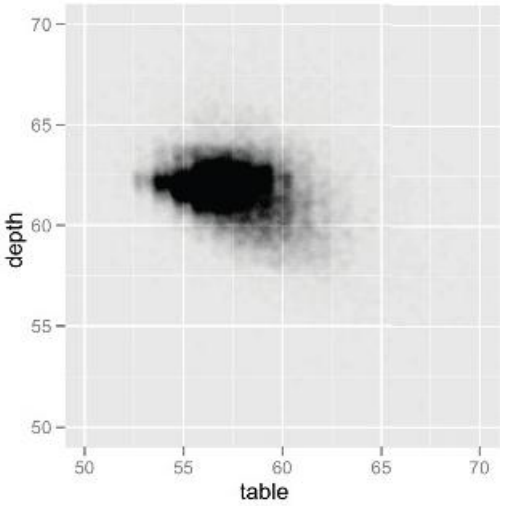
alpha of 1/10



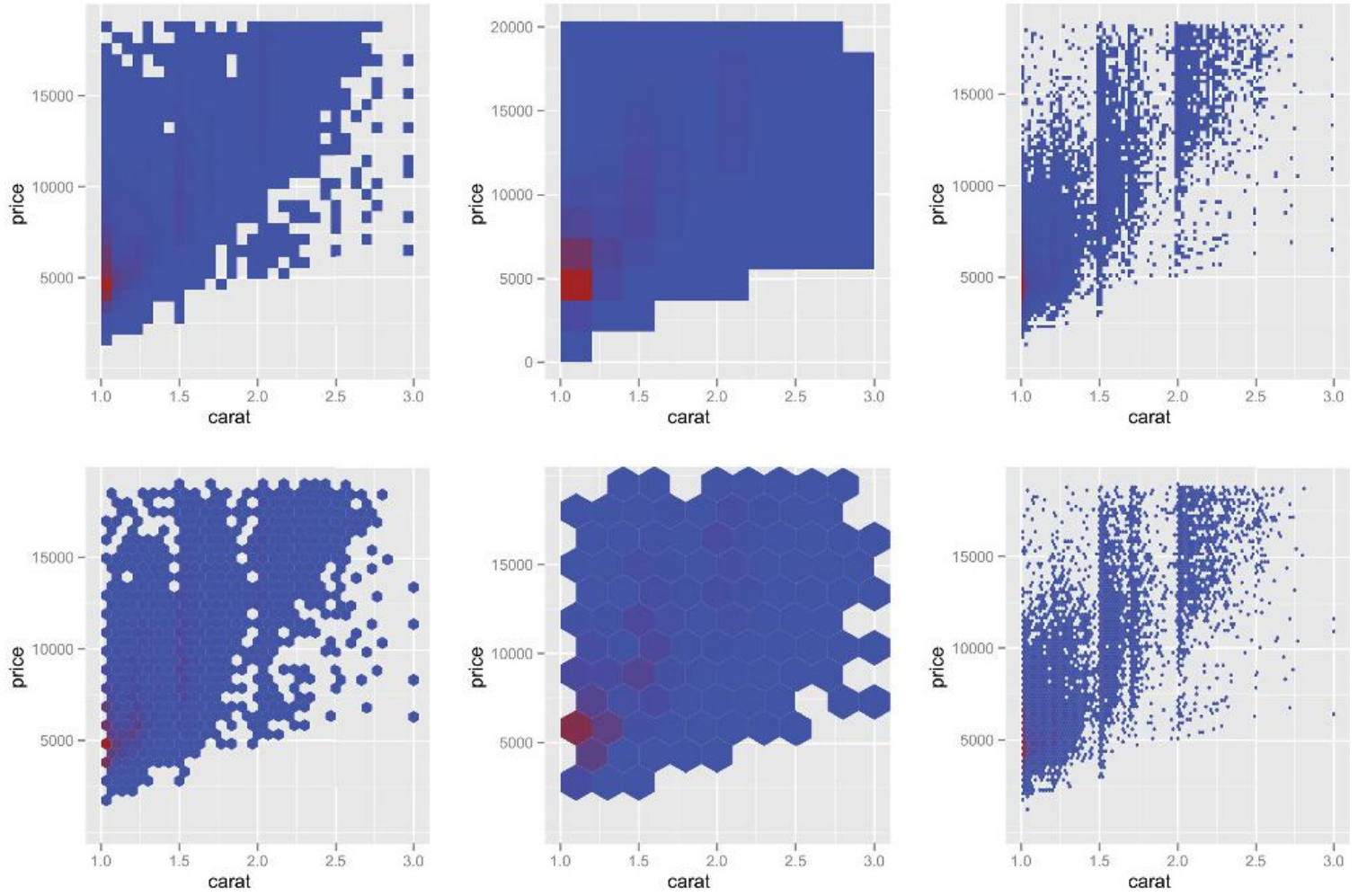
alpha of 1/50



alpha of 1/200



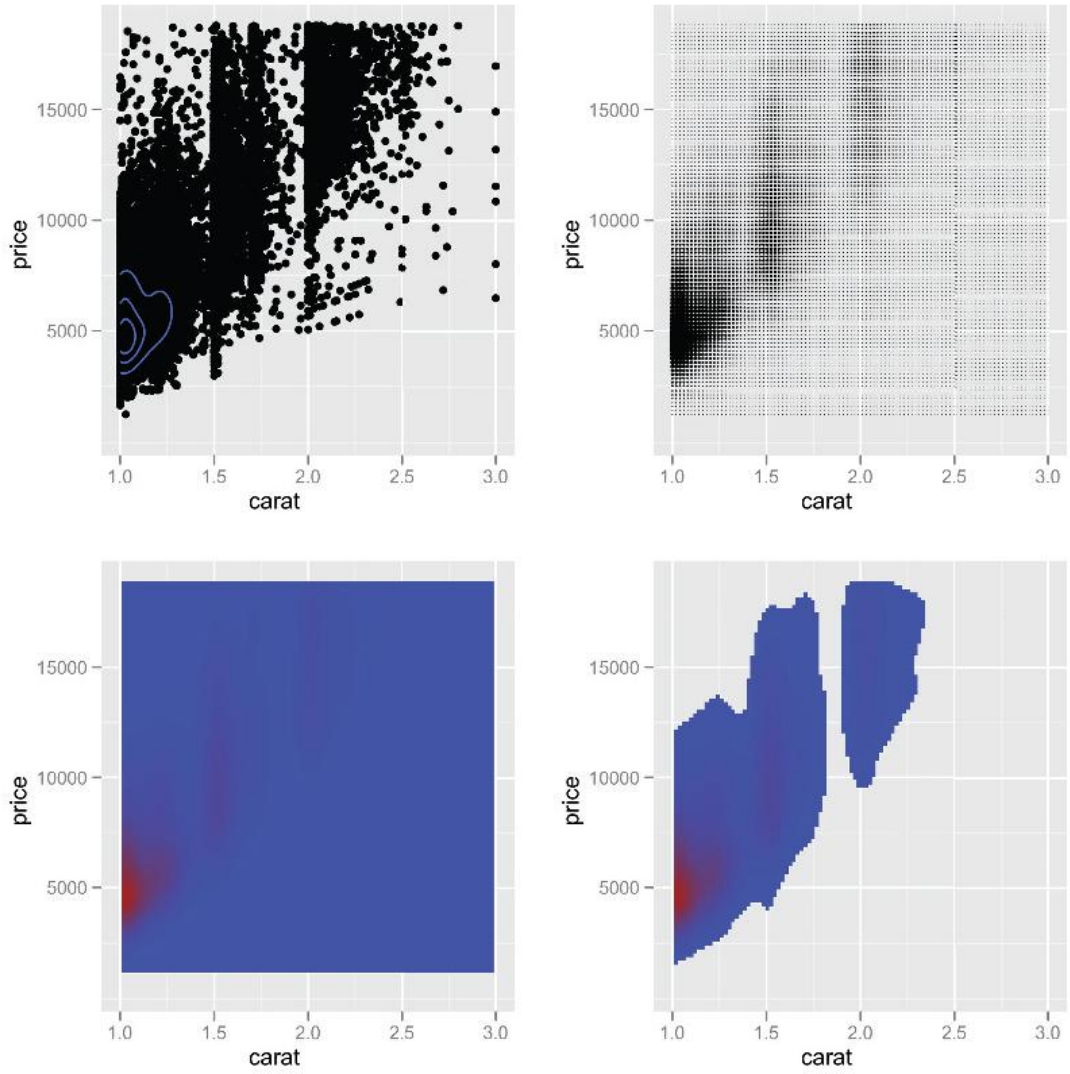
大数据集的遮盖绘制：直方图的二维推广（1/2）



大数据集的遮盖绘制：直方图的二维推广（2/2）

- 上页图形对应代码：
  - d <- ggplot(diamonds, aes(carat, price)) + xlim(1,3) +labs(legend.position = "none")
  - d + stat_bin2d()
  - d + stat_bin2d(bins = 10)
  - d + stat_bin2d(binwidth=c(0.02, 200))
  - d + stat_binhex()
  - d + stat_binhex(bins = 10)
  - d + stat_binhex(binwidth=c(0.02, 200))

大数据集的遮盖绘制：密度估计



代码：

- d <- ggplot(diamonds, aes(carat, price)) + xlim(1,3) + labs(legend.position = "none")
- d + geom_point() + geom_density2d()
- d + stat_density2d(geom = "point", aes(size = ..density..), contour = F) + scale_size_area()
- d + stat_density2d(geom = "tile", aes(fill = ..density..), contour = F)
- last_plot() + scale_fill_gradient(limits = c(1e-5,8e-4))

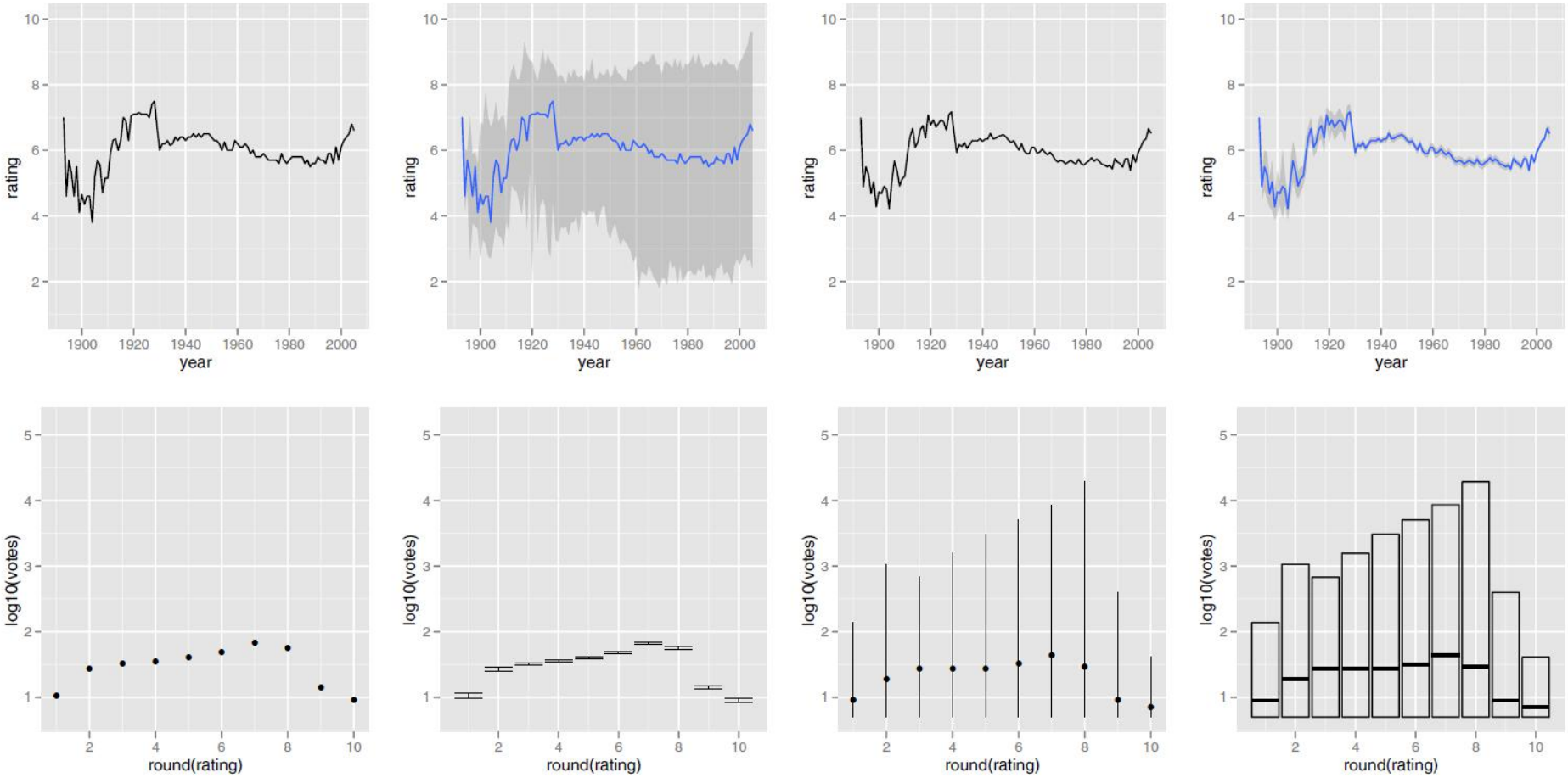
统计摘要

统计摘要

- 对于每个x取值，计算对应y值的统计摘要通常很有用。
- 在ggplot2，可以利用图形属性ymin，y和ymax，通过**stat_summary()**汇总统计y的条件分布。
- 使用stat_summary()时，可以：
  - 为每个参数制定**单独的摘要计算函数**，或
  - 使用**统一的摘要计算函数**。

stat_summary示例（1/2）

上面四幅图展示了**连续型**变量x的：
中位数曲线、median_hilow()曲线和平滑带、均值曲线、mean_cl_boot()和平滑带



下面四幅图展示了**离散型**变量x的：mean()均值点、mean_cl_normal()均值点和误差棒、median_hilow()中位数点和值域、median_hilow()中位数点和值域条

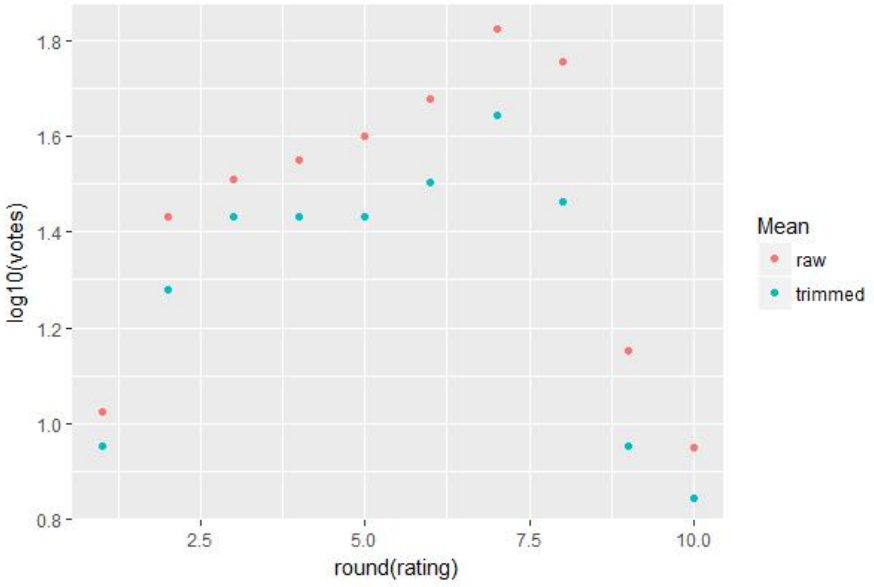
stat_summary示例（2/2）

- 上页代码（具体语法下面介绍）：
  - library(ggplot2movies)
  - **m** <- ggplot(movies, aes(year, rating)) #后面会用到
  - m + stat_summary(fun.y = "median", geom = "line")
  - m + stat_summary(fun.data = "median_hilow", geom = "smooth")
  - m + stat_summary(fun.y = "mean", geom = "line")
  - m + stat_summary(fun.data = "mean_cl_boot", geom = "smooth")
  - **m2** <- ggplot(movies, aes(round(rating), log10(votes))) #后面会用到
  - m2 + stat_summary(fun.y = "mean", geom = "point")
  - m2 + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar")
  - m2 + stat_summary(fun.data = "median_hilow", geom = "pointrange")
  - m2 + stat_summary(fun.data = "median_hilow", geom = "crossbar")

单独的摘要计算函数

- 参数fun.y，fun.ymin和fun.ymax能够接受简单的**数值型摘要计算函数**。该函数能传入一个数值向量并返回一个数值型结果，如mean(), median(), min(), max()。例：
  - midm <- function(x) mean(x, trim = 0.5)
  - m2 + **stat_summary(**aes(colour = "trimmed"), **fun.y = midm**, geom = "point"**)** +**stat_summary(**aes(colour = "raw"), fun.y = mean, geom = "point"**)** + scale_colour_hue("Mean")

geom指定为散点图



统一的摘要计算函数

- fun.data可以支持更复杂的摘要计算函数。
- 下例利用了自己编写的摘要计算函数，函数返回了一个vector。

```
iqr <- function(x, ...) { #这里我们不关心函数意义
  qs <- quantile(as.numeric(x), c(0.25, 0.75), na.rm = T)
  names(qs) <- c("ymin", "ymax")
  qs
}
m + stat_summary(fun.data = "iqr", geom="ribbon")
```

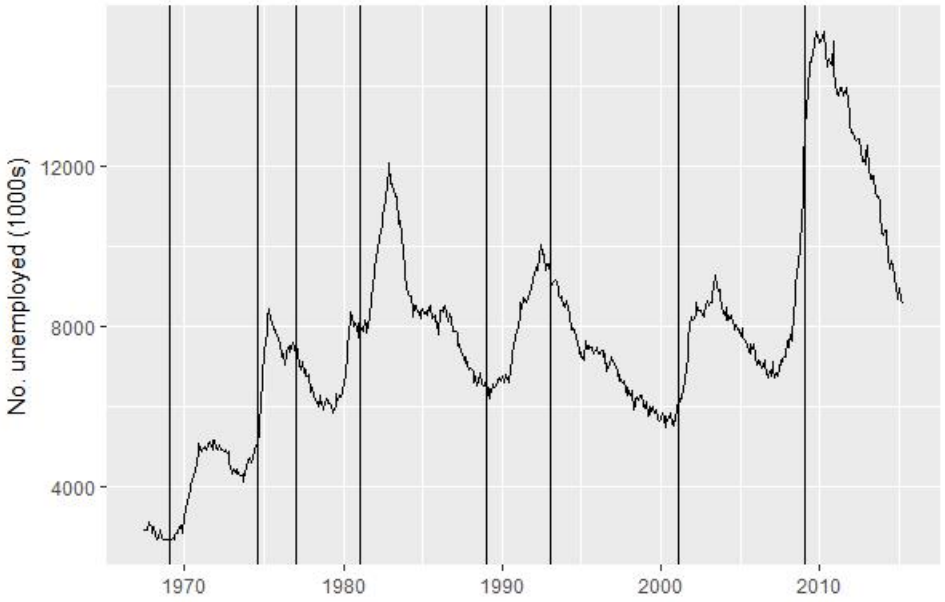
添加图形注解

图形注解

- 在使用额外的标签注解图形时要记住：这些注解仅仅是额外的数据而已。
- 添加图形注解有两种基本方式：
  - 逐个添加
    - 适合少量的、图形属性多样化的注解。只要为相应的图形属性设置好对应的值即可。
  - 批量添加
    - 添加多个具有类似属性的注解，将它们放在数据框中并一次添加完成。
- 例子：为美国经济数据中加入总统信息。

例一：添加垂直线
（按不同总统划分）

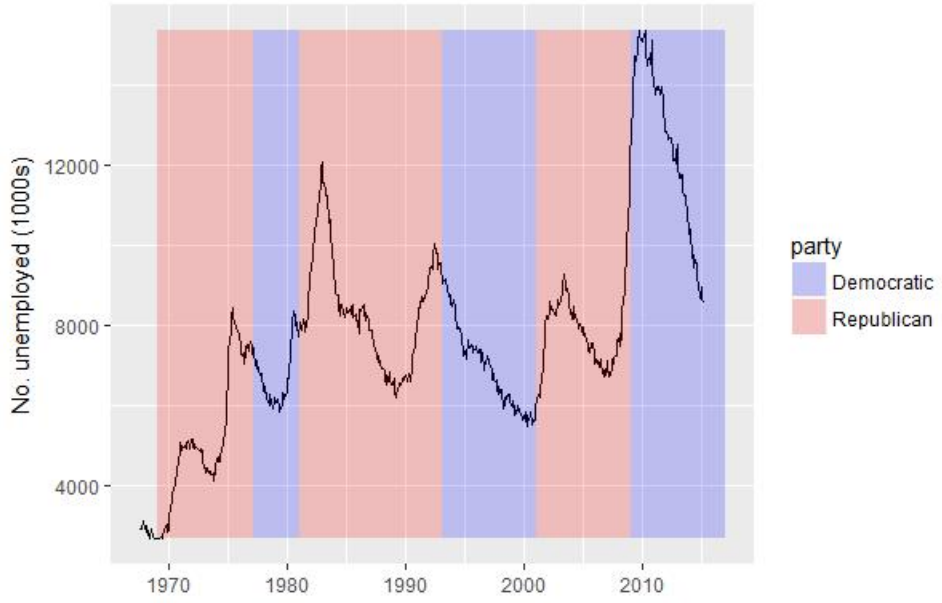
- (unemp <- qplot(date, unemploy, data=economics, geom="line", xlab = "", ylab = "No. unemployed (1000s)"))
- presidential <- presidential[-(1:3),]
- yrng <- range(economics$unemploy)
- xrng <- range(economics$date)
- unemp + **geom_vline**(aes(xintercept = as.numeric(start)), data = presidential)



例二：强调图形中感兴趣的区域

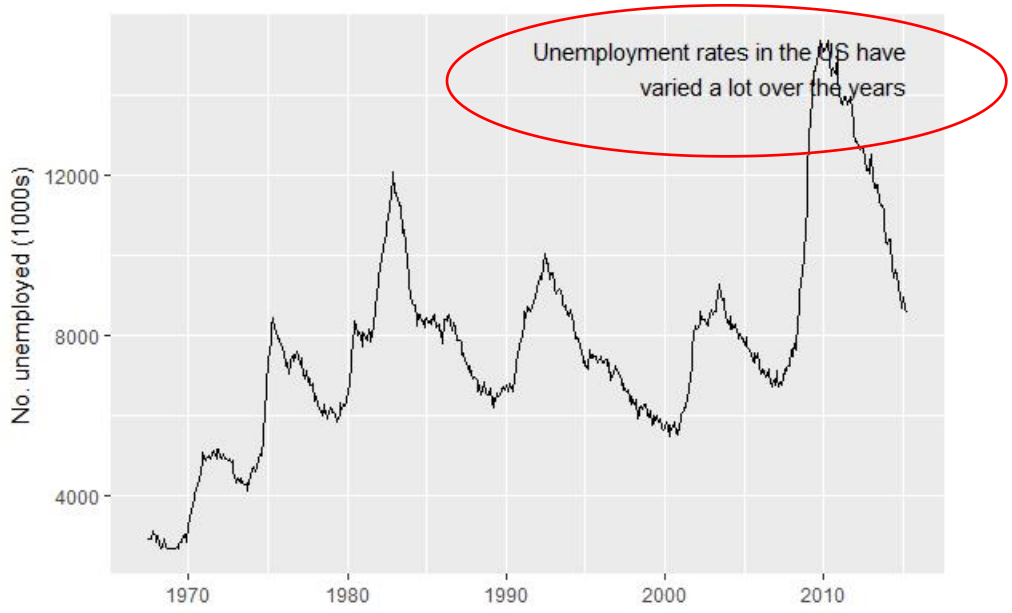
（按总统的党派标记）

- unemp + **geom_rect**(aes(NULL, NULL, xmin = start, xmax = end, fill = party), ymin = yrng[1], ymax = yrng[2], data = presidential) + scale_fill_manual(values = alpha(c("blue", "red"), 0.2))



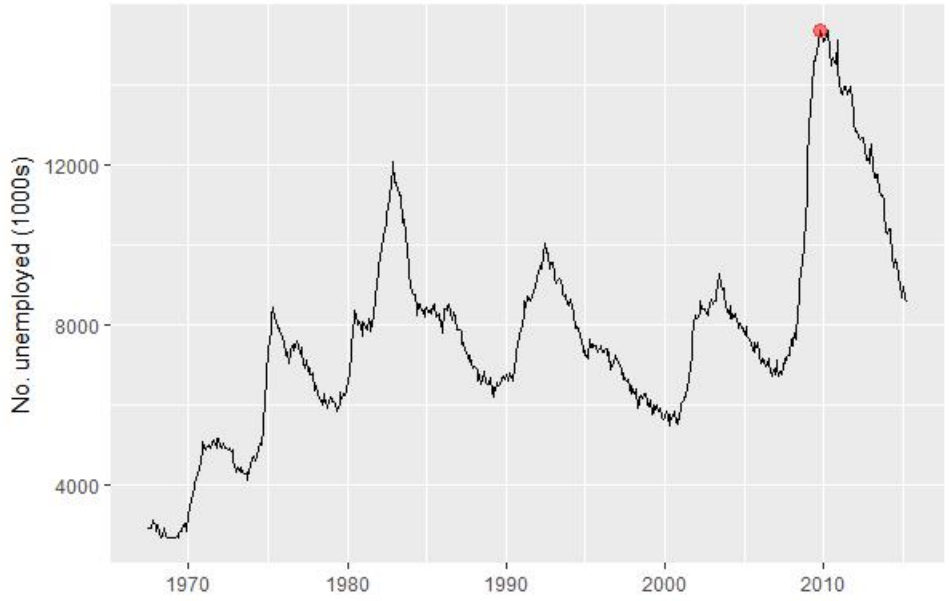
例三：文本注解

- caption <- paste(strwrap("Unemployment rates in the US have varied a lot over the years", 40), collapse="\n")
- unemp + **geom_text**(aes(x, y, label = caption), data = data.frame(x = xrng[2], y = yrng[2]), hjust = 1, vjust = 1, size = 4)



例四：强调某个点
（最大失业率的点）

- highest <- subset(economics, unemploy == max(unemploy))
- unemp + **geom_point**(data = highest, size = 3, colour = alpha("red", 0.5))



含权数据

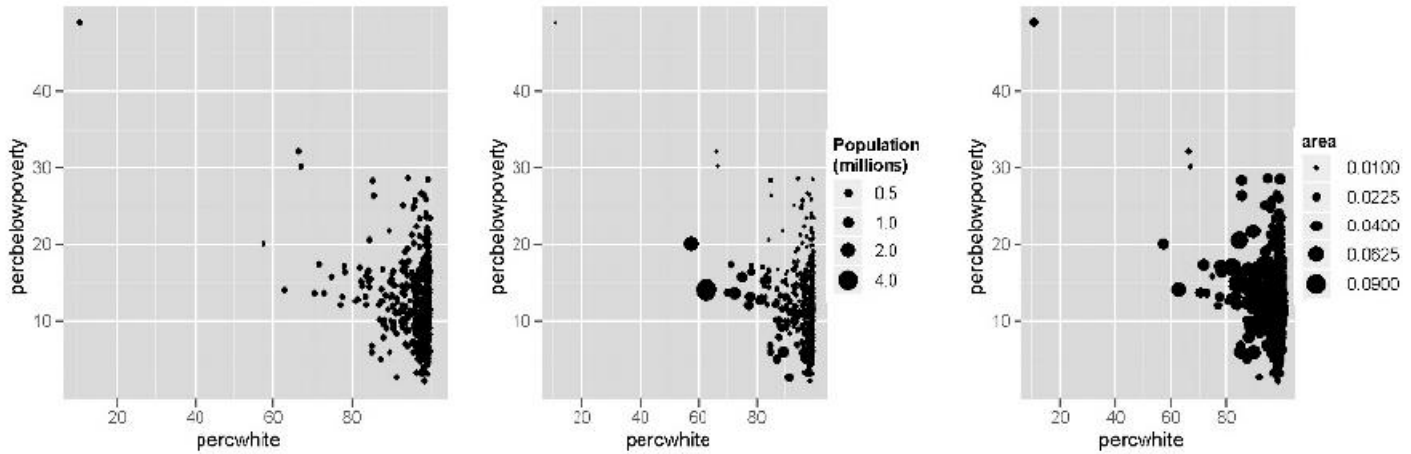
数据集

- 2000年美国人口普查中，中西部各州的统计数据。
  - 比例型数据
    - 白种人比例
    - 贫困线以下人口比例
    - 大学学历人口比例
  - 每个郡的基本信息
    - 面积
    - 人口总数
    - 人口密度
- 有些数据可以**作为权重使用**
  - **总人数**：与原始的绝对数配合使用
  - **总面积**：研究地缘效应

使用点的大小展示权重

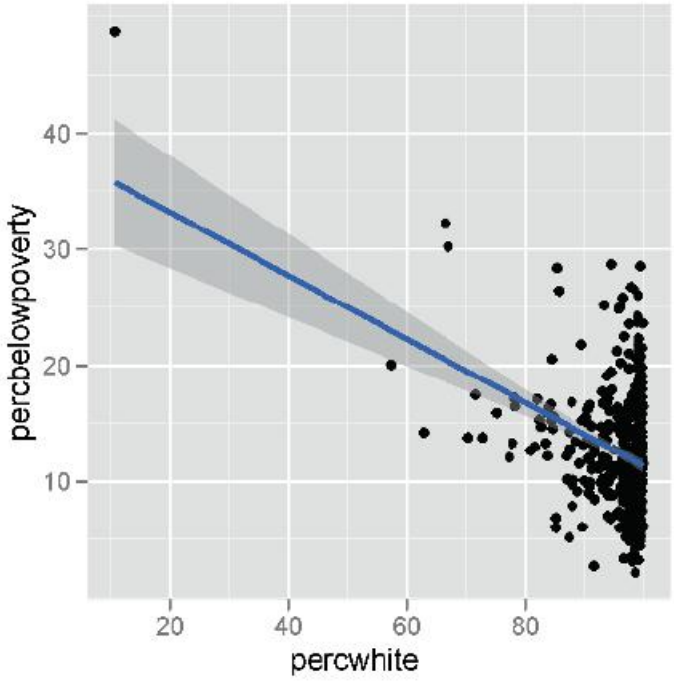
白种人比例　贫困线以下人口比例

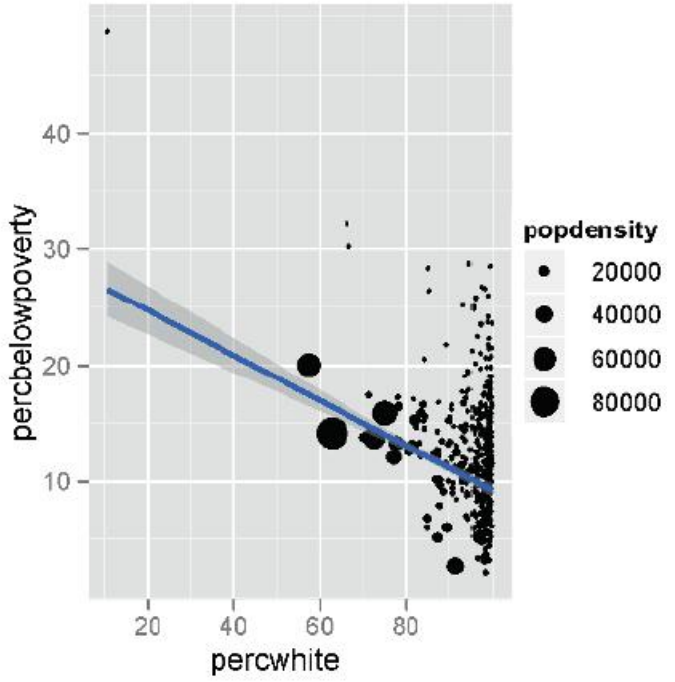
- qplot(percwhite, percbelowpoverty, data = midwest) **#左图：无权重**
- qplot(percwhite, percbelowpoverty, data = midwest, **size = poptotal / 1e6)** + scale_size_area("Population\n(millions)", breaks = c(0.5, 1, 2, 4)) **#中图：以人口数量为权重。注：灰色部分用作对图形优化，可不写。**
- qplot(percwhite, percbelowpoverty, data = midwest, **size = area**) + scale_size_area() **#右图：以面积为权重**



未考虑权重的最优拟合曲线和以人口数量为权重的最优拟合曲线

- lm_smooth <- geom_smooth(method = lm, size = 1)
- qplot(percwhite, percbelowpoverty, data = midwest) + lm_smooth
- qplot(percwhite, percbelowpoverty, data = midwest, weight = popdensity, size = popdensity) + lm_smooth #weight图形属性表示权重，它会被直接传递给汇总计算函数。权重变量会影响统计汇总结果。





含权重信息的直方图

- 例：贫困线以下人口比例直方图
  - qplot(percbelowpoverty, data = midwest, binwidth = 1) #左图，无权重信息
  - qplot(percbelowpoverty, data = midwest, **weight =** poptotal, binwidth = 1) + ylab("population") #右图，人口为权重。