Data Analysis Tools and Practice(Using R)

2019.04.12

ggplot2画图II

北京大学 软件与微电子学院
School of Software and Microelectronics, Peking University

Huiping Sun(孙惠平)
sunhp@ss.pku.edu.cn

课堂测试时间

课堂测试06

先用电脑完成
*40*分钟　然后誊抄纸上

- 1、查看数据集mtcars，根据要求作图：

  (1)分别使用qplot、ggplot函数画出mpg和wt关系的散点图；

  (2)使用三种方式画出mpg列的直方图,同时在使用qplot和ggplot时指定每个小圆柱体的宽度是4；

  (3)使用三种绘图函数画出mpg变量的密度曲线。

- 2、使用datasets包中的数据集pressure，查看其数据并按要求画图：

  (1)请画出pressure和temperature关系的曲线图；

  (2)分别使用qplot和ggplot画出pressure和temperature关系的散点图和折线图。

- 3、使用datasets中的数据集ToothGrowth，完成如下的绘图要求：

  (1)以supp变量作为分类,分别使用三种绘图函数画出len变量的箱型图。

- 4、使用ggplot2包中数据集mpg，完成练习：

  (1)使用mpg数据集定义一个 ggplot对象，表示hwy与cty的关系；

  (2)画一个散点图，指定颜色有year列来指定，并在上边绘图的基础上画出平滑的拟合曲线；

  (3)继续使用(1)中定义的ggplot对象画散点图，使用class来指定颜色，displ指定大小，透明度；
  指定为0.5,position指定为抖动，在散点图的基础上添加拟合曲线；

  (4)使用qplot画出hwy与cty的关系的散点图，并根据year变量分面，同时添加拟合曲线。

上次课程内容回顾

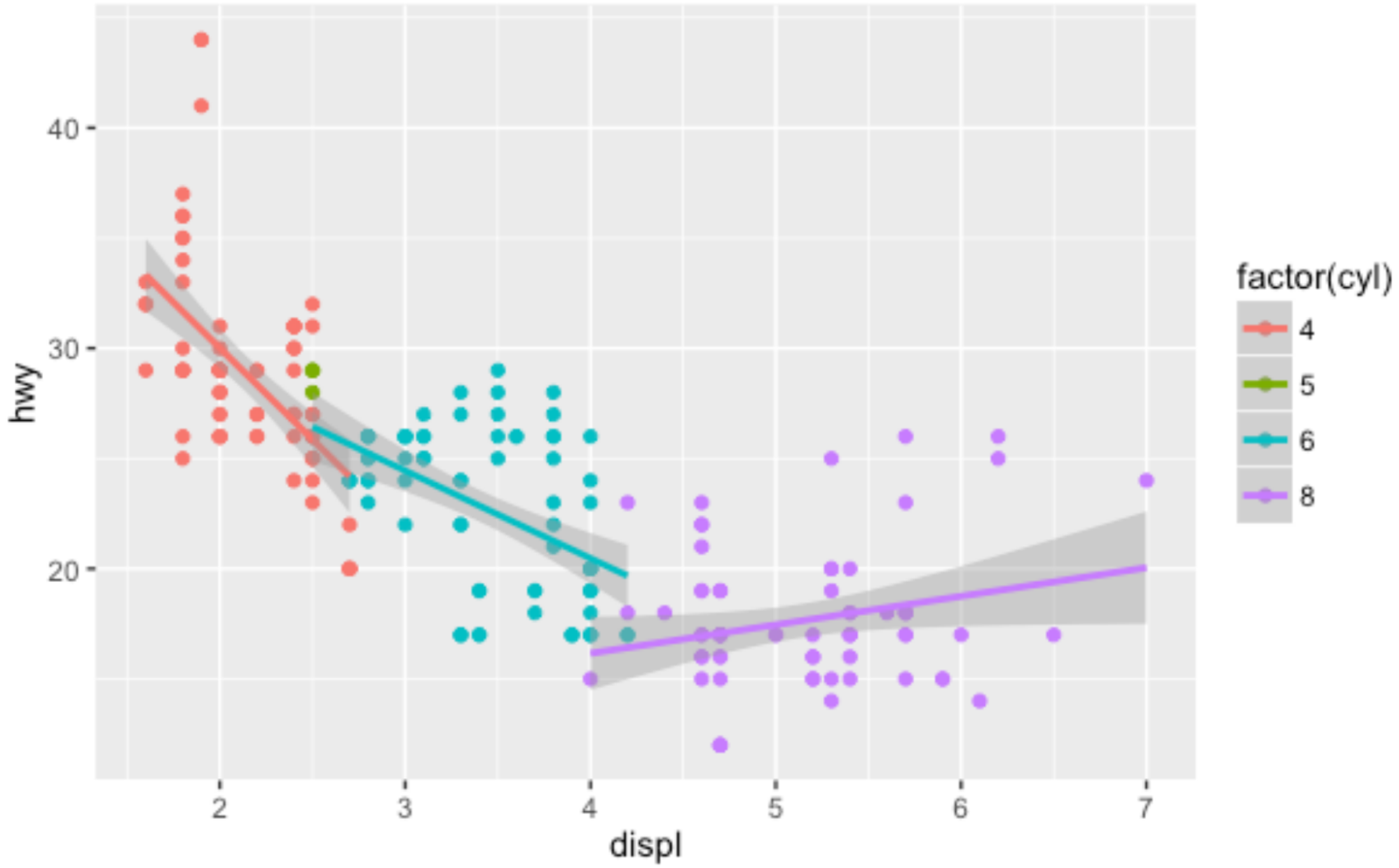
- ggplot2
- qplot()：
  - ✴ data; log; colour; shape; alpha;
- geom：
  - ✴ point; smooth; jitter; boxplot;path; line; histogram; freqpoly; density; bar;
  - ✴ binwidth; fill; weight; scale_y_continous(); smooth;
- facets：
- ggplot():
  - ✴ +; %+%; layer(); geom_xxx(); stat_xxx(); aes(); group;

工具箱

CH5

图层用途

- 展示数据本身
- 展示数据的统计摘要
- 添加额外的元数据、上下文信息和注解



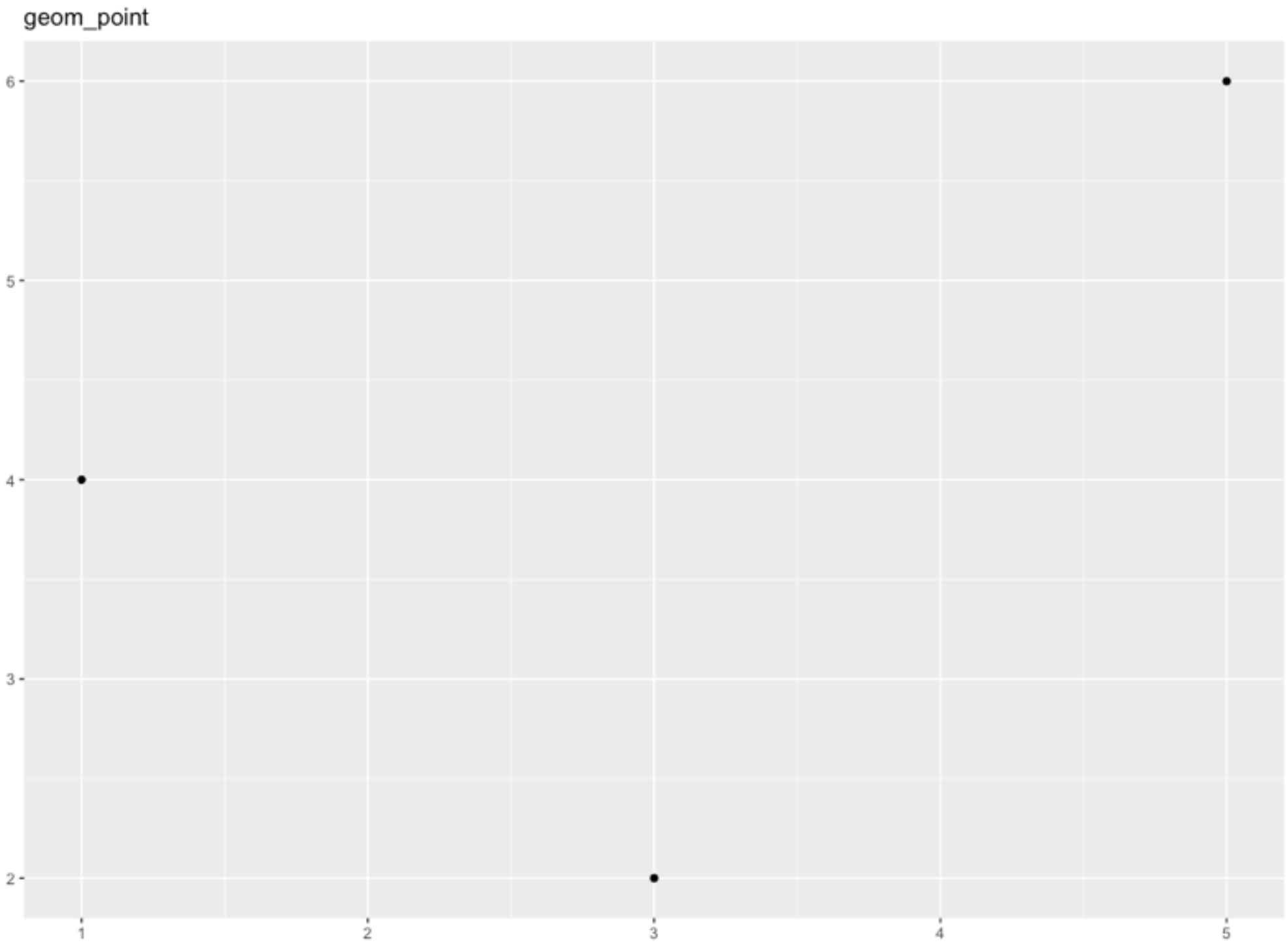
基本图形类型

- geom_area()：面积图
- geom_bar(stat="identity")：条形图
- geom_line()：线条图
- geom_point()：散点图
- geom_text()：添加标签
- geom_tile()：色深图、水平图

```
> 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)
```

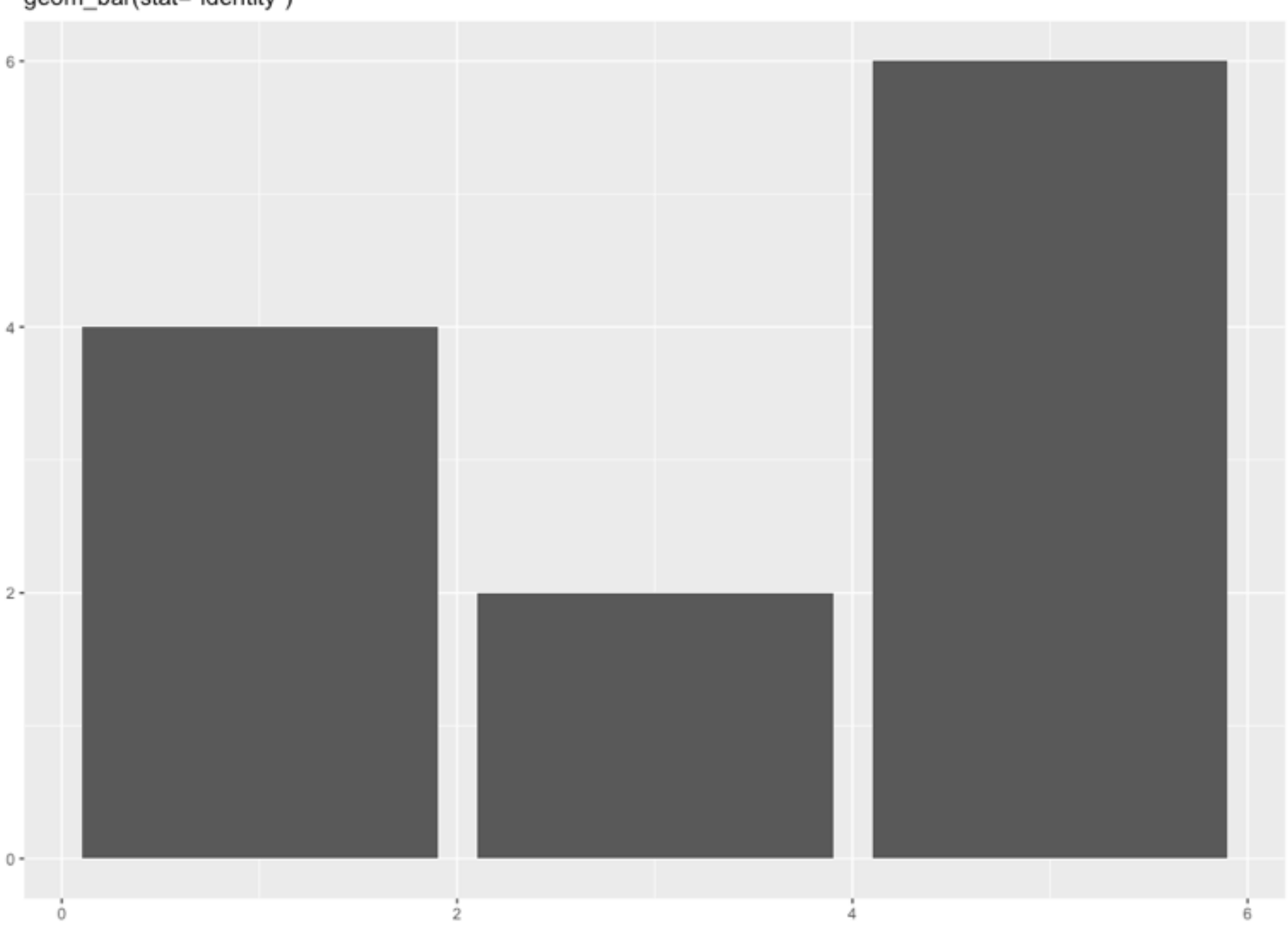
散点图

```
> p + geom_point() + labs(title = "geom_point")
```



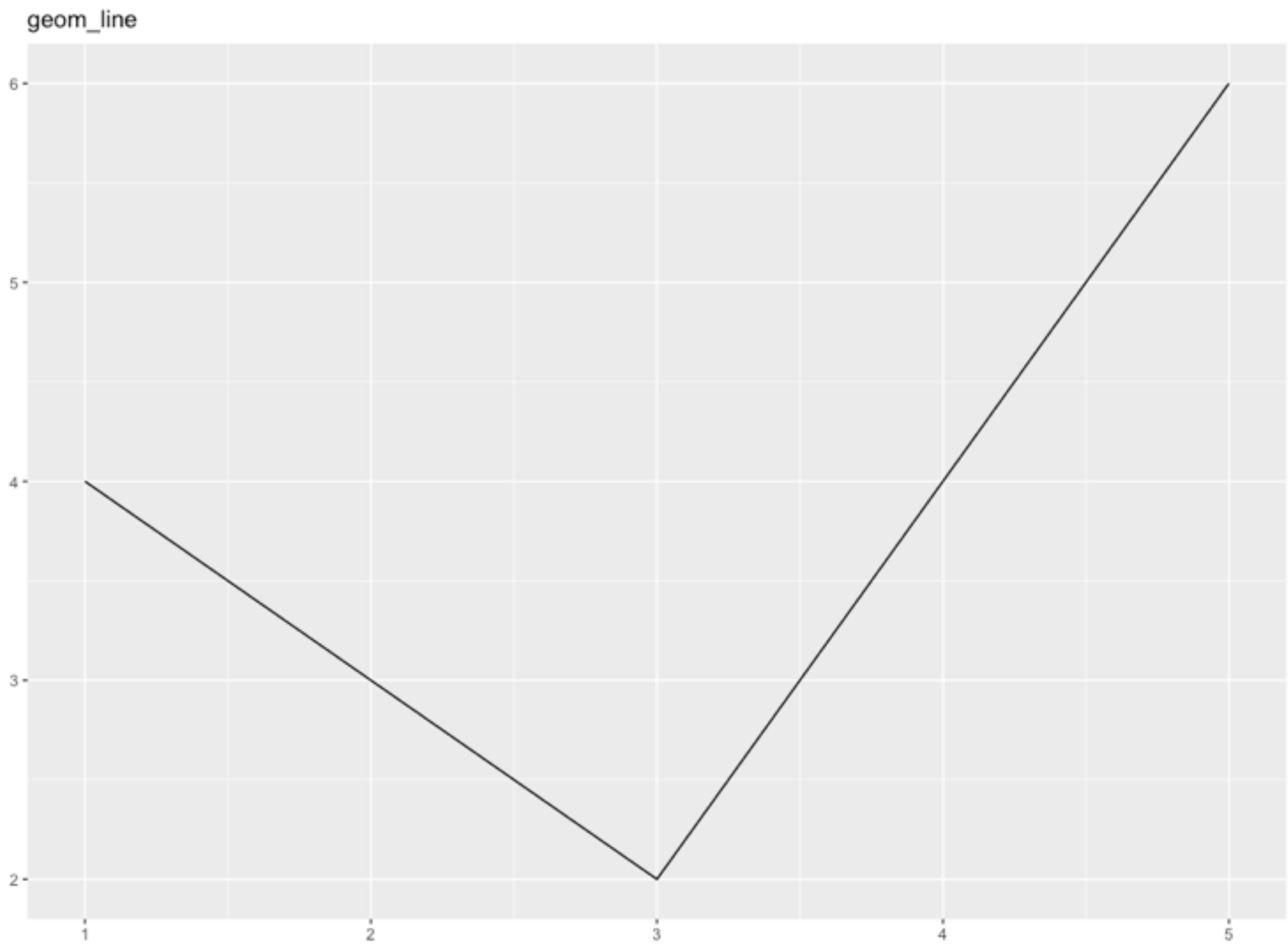
条形图

```
> p + geom_bar(stat="identity") +
+   labs(title = "geom_bar(stat=\"identity\")")
```



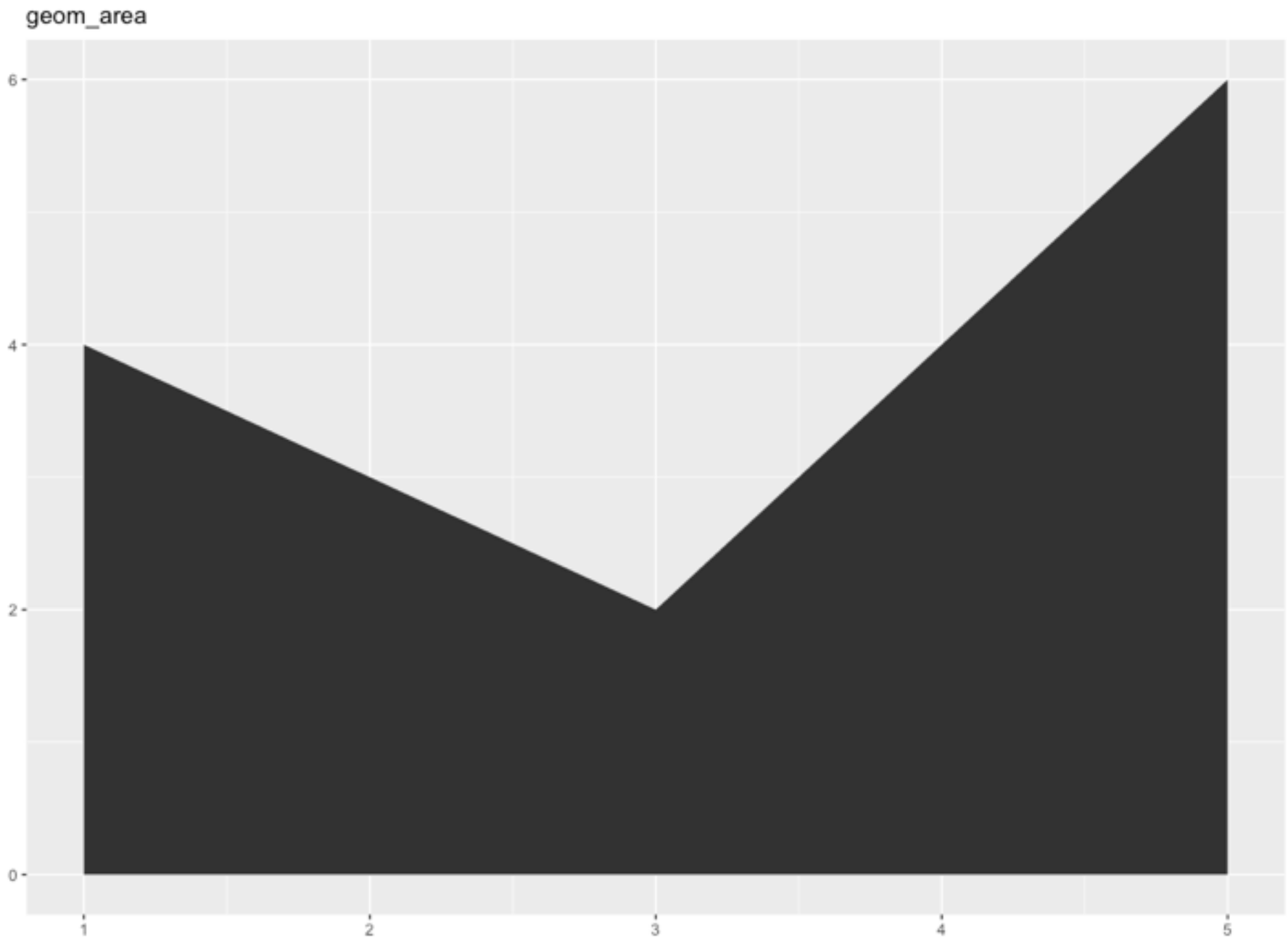
线条图

```
> p + geom_line() + labs(title = "geom_line")
```



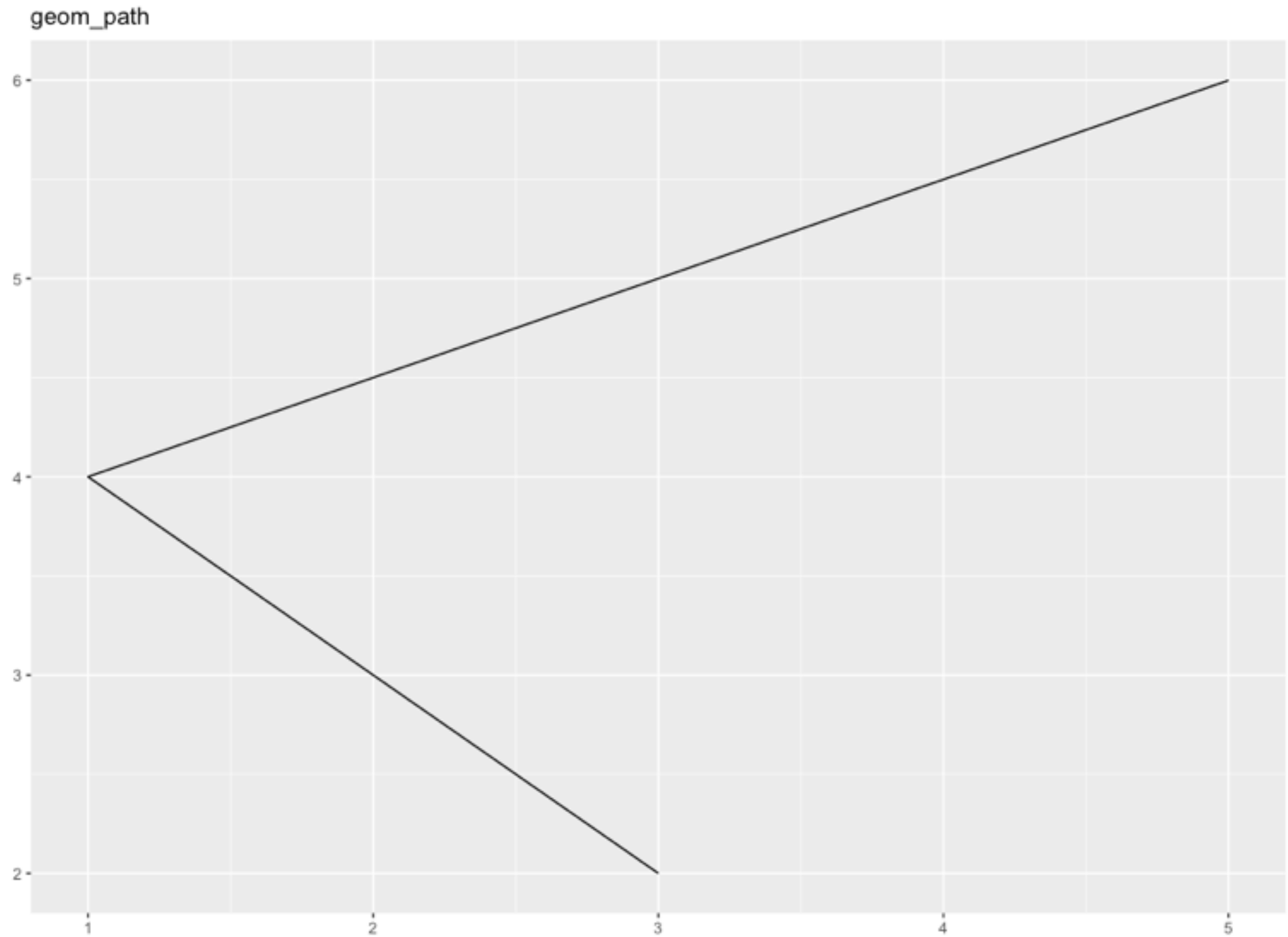
面积图

```
> p + geom_area() + labs(title = "geom_area")
```



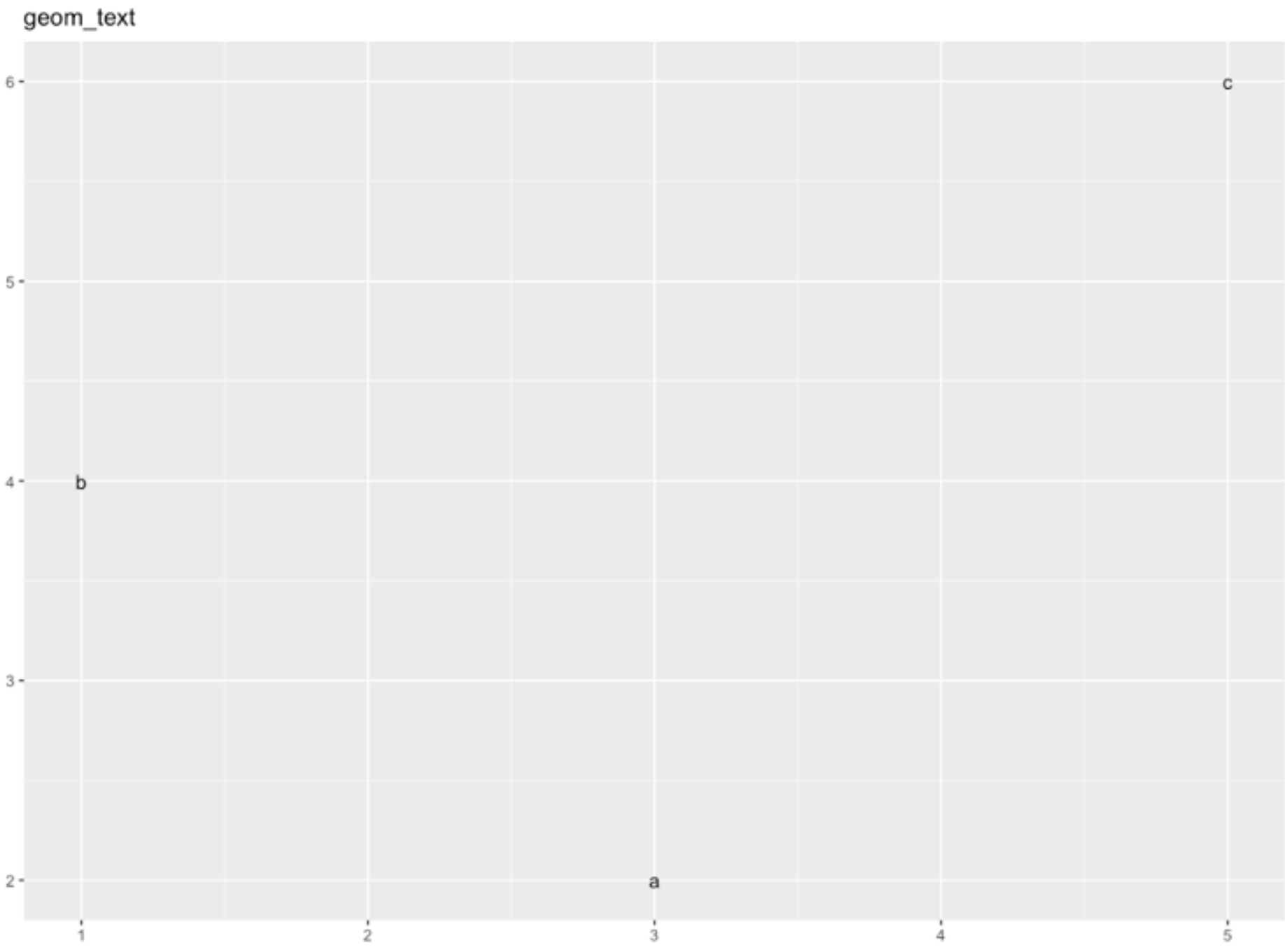
路径图

```
> p + geom_path() + labs(title = "geom_path")
```



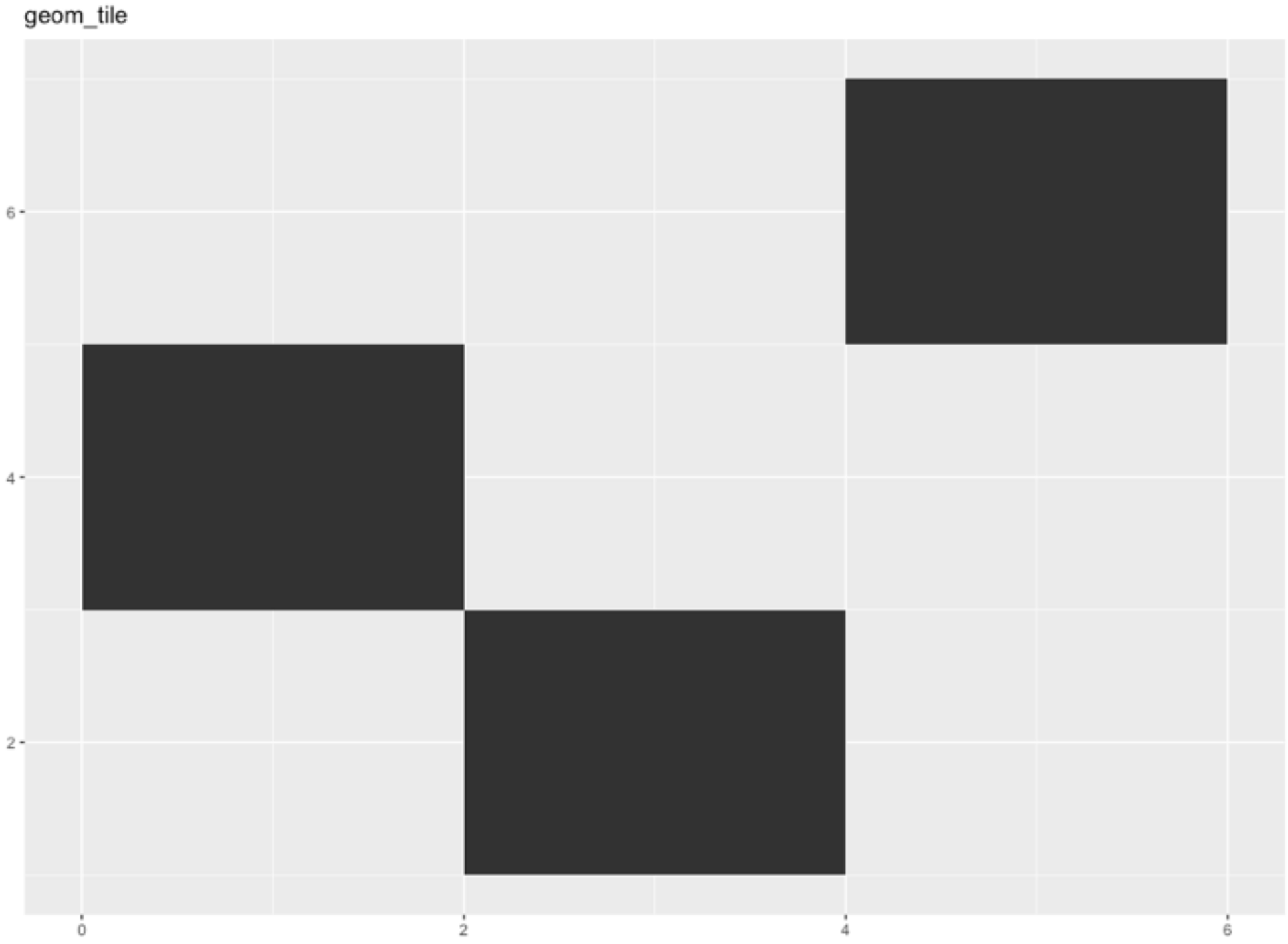
添加标签

```
> p + geom_text() + labs(title = "geom_text")
```



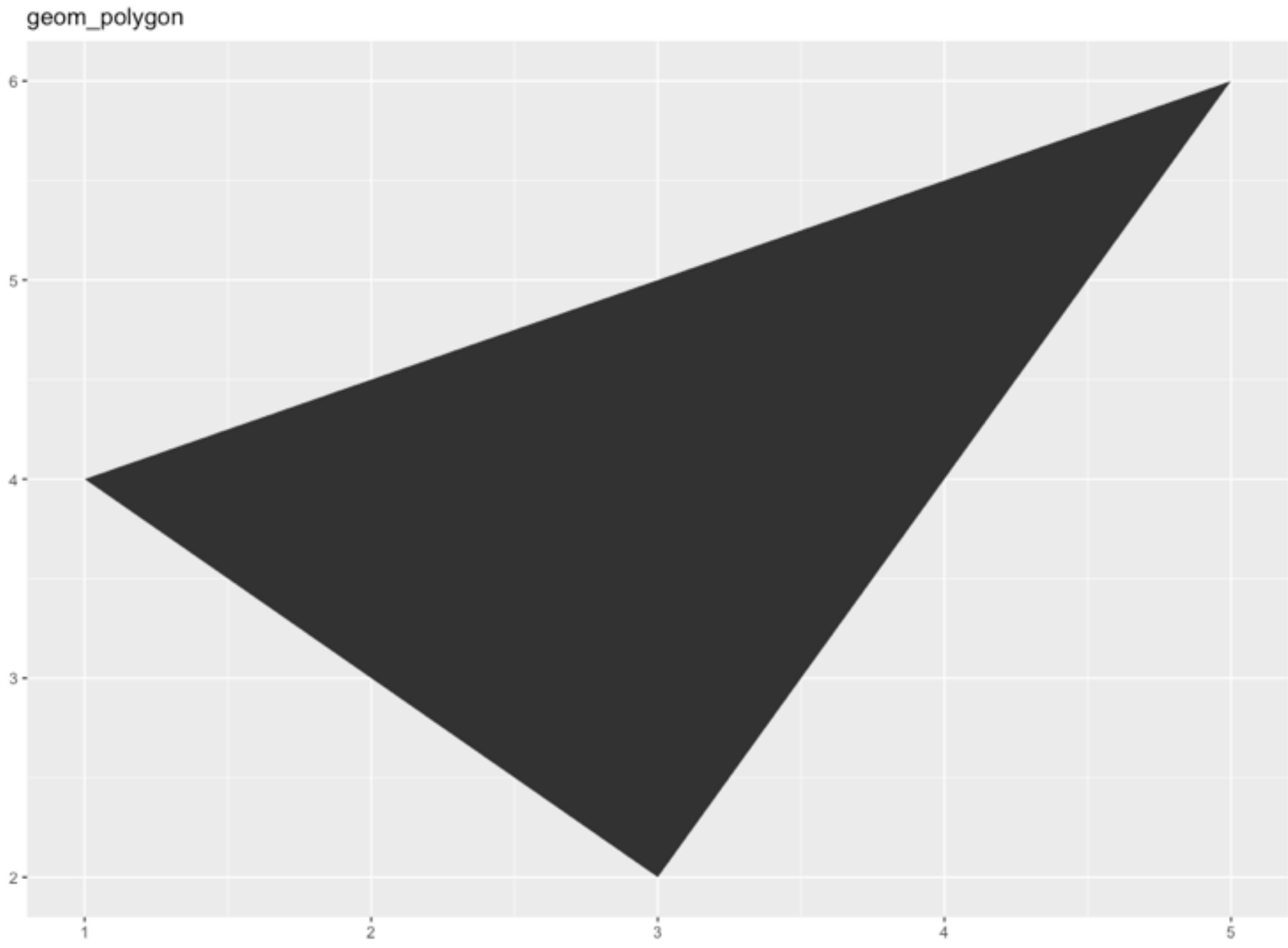
色深图 / 水平图

```
> p + geom_tile() + labs(title = "geom_tile")
```



多边形图

```
> p + geom_polygon() + labs(title = "geom_polygon")
```



钻石数据集

carat	cut	color	clarity	depth	table	price	x	y	z
0.2	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.2	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.2	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.2	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.2	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.2	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48

carat：克拉重量

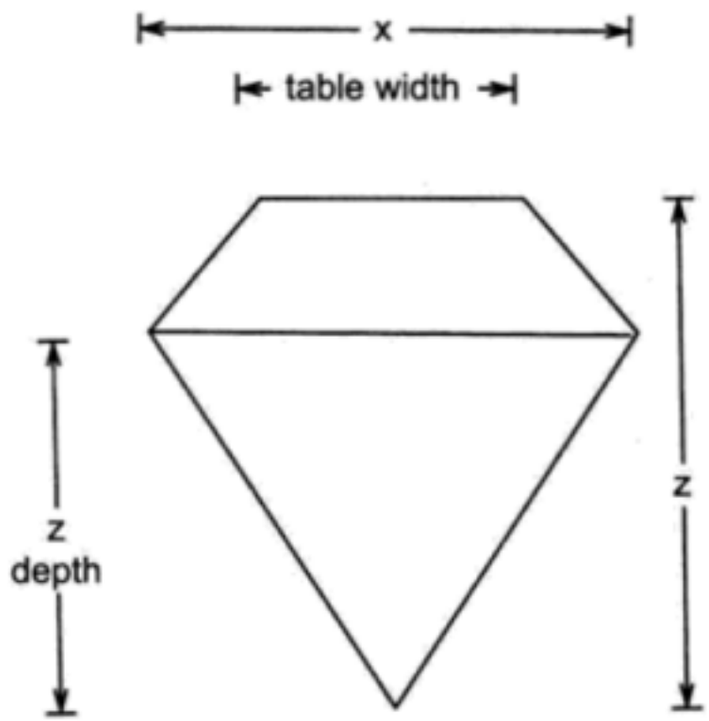
cut：切工

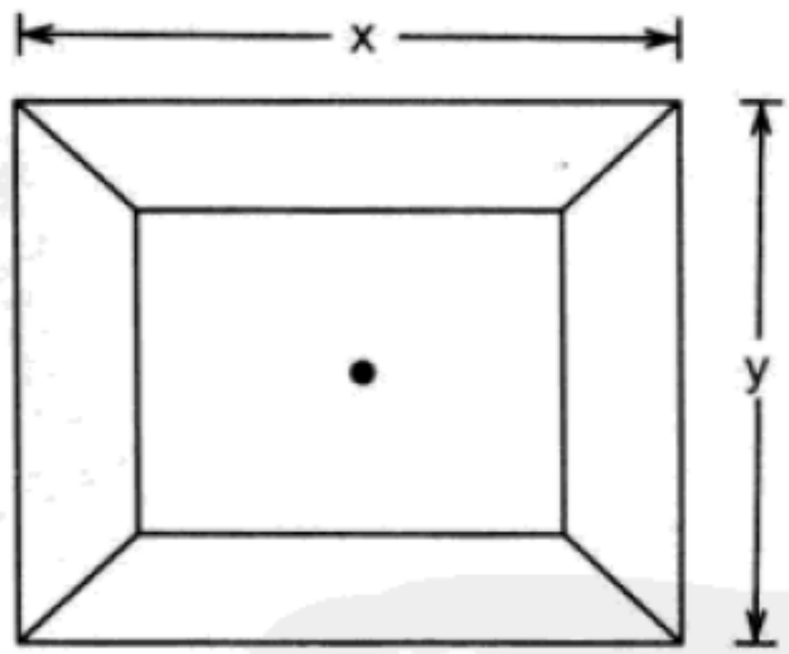
color：颜色

clarity：净度

depty：深度

table：钻面宽度

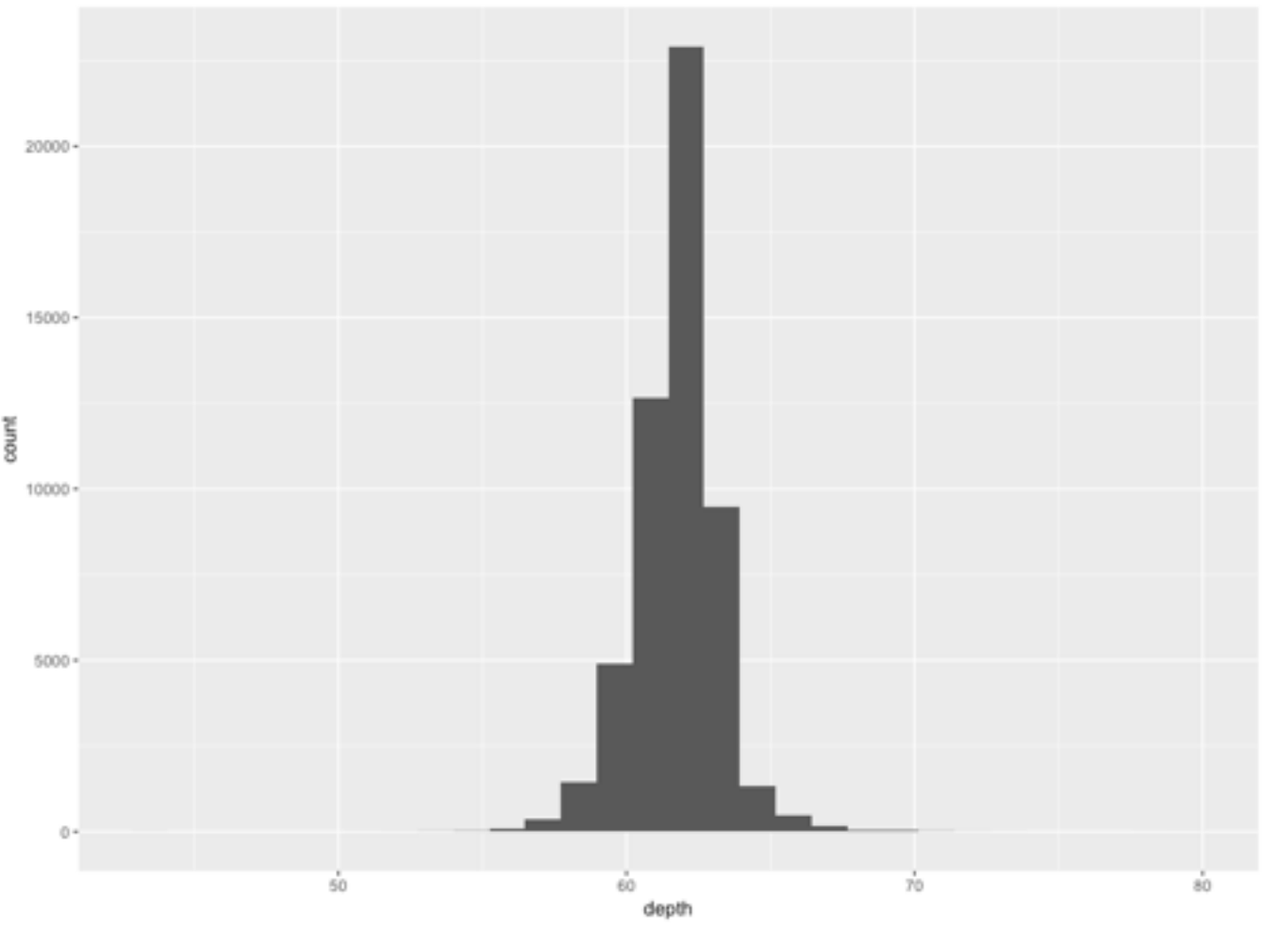


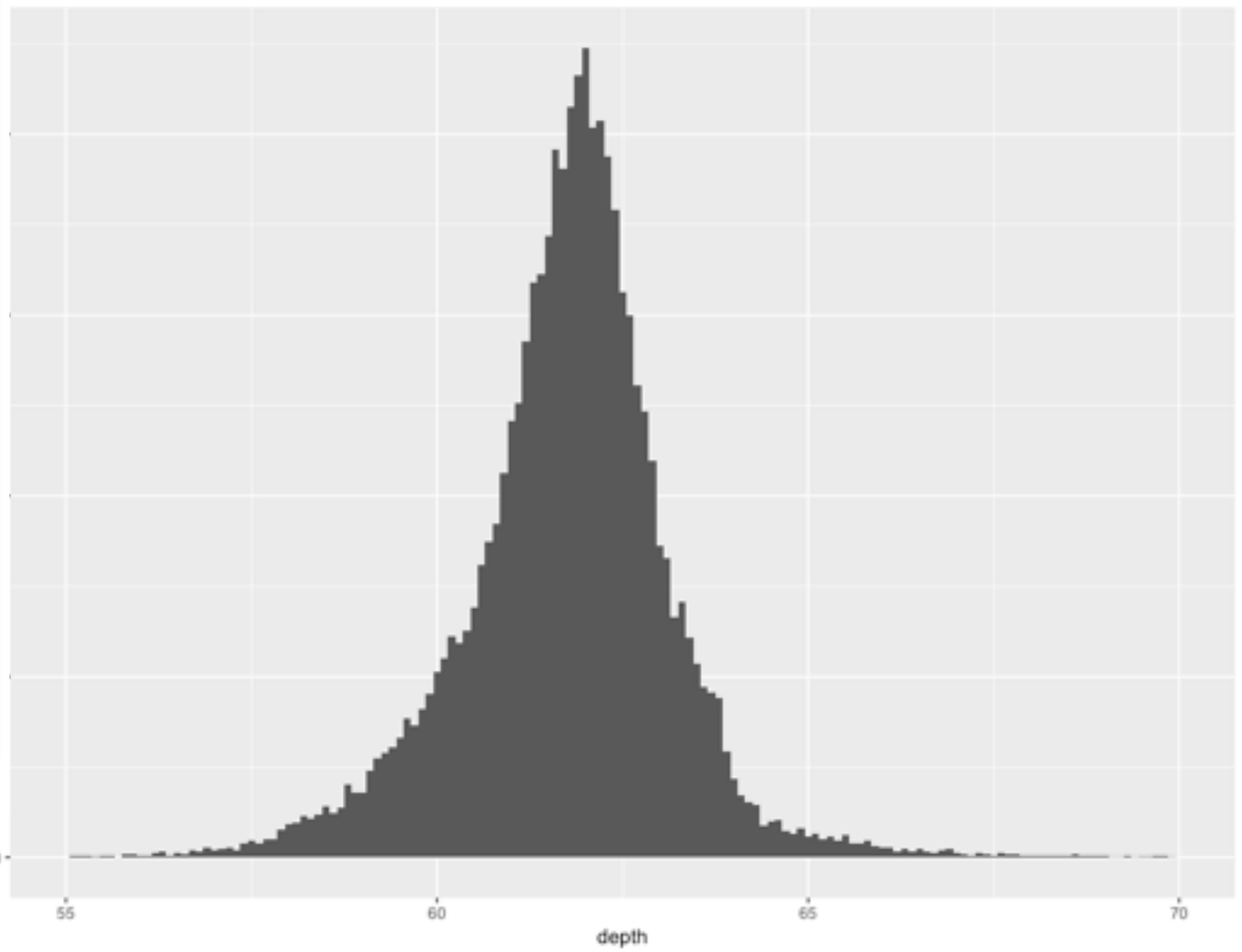


depth = z depth / z * 100
table = table width / x * 100

展示数据分布

```
> qplot(depth, data=diamonds, geom="histogram")
```

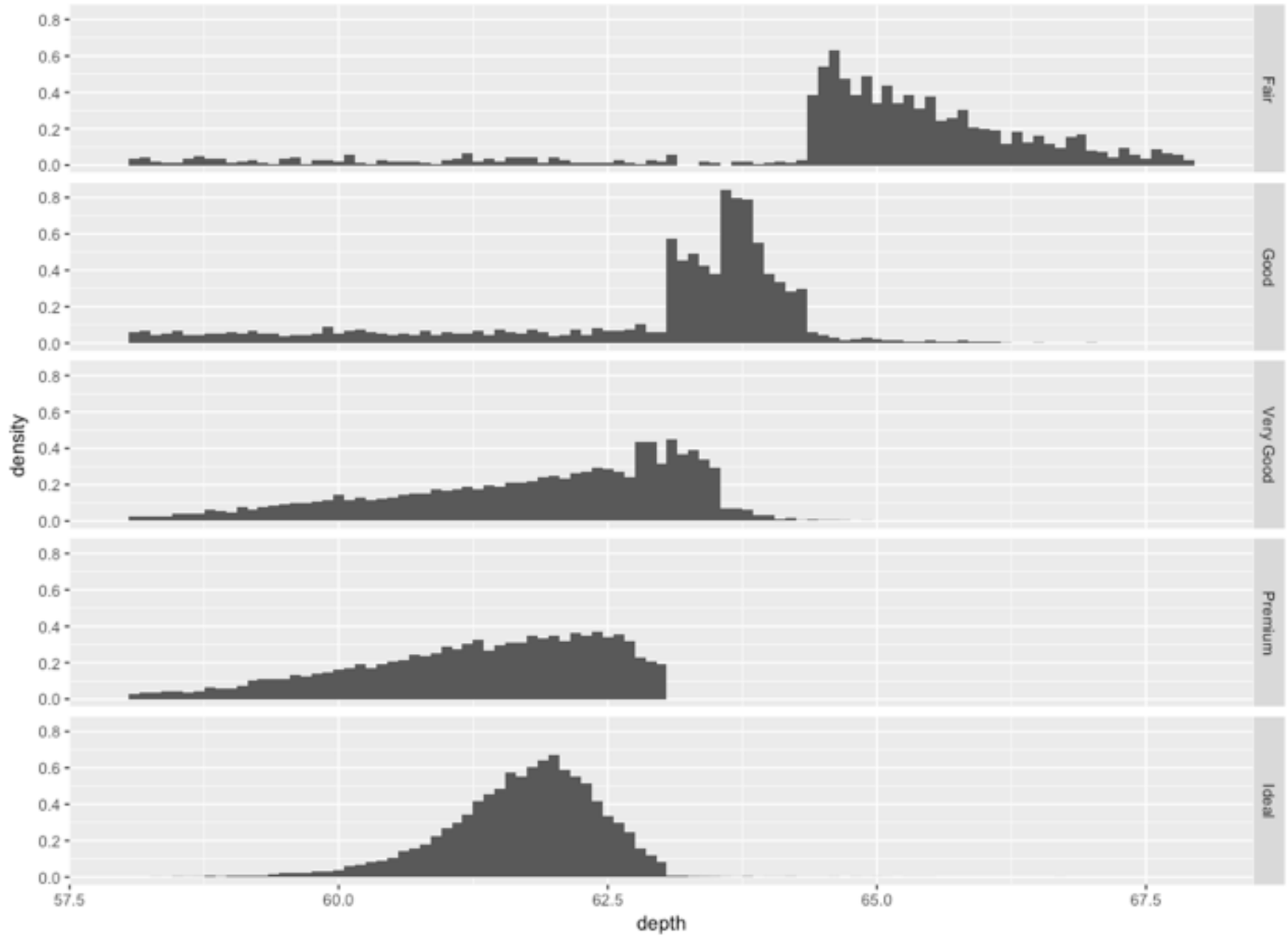





```
> qplot(depth, data=diamonds, geom="histogram", xlim=c(55, 70), binwidth=0.1)
```

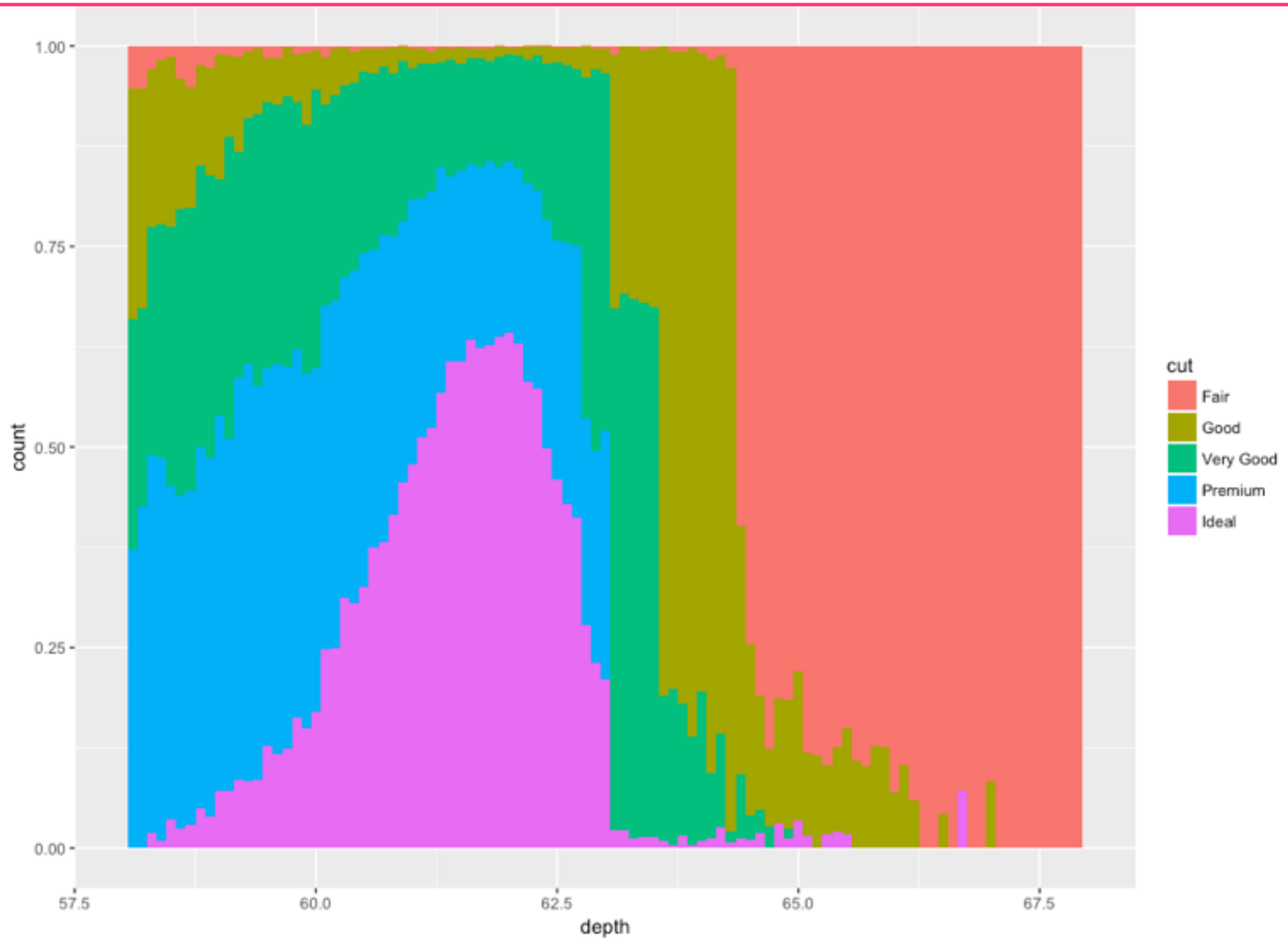
分面直方图

```
> depth_dist <- ggplot(diamonds, aes(depth)) + xlim(58, 68)
> depth_dist +
+   geom_histogram(aes(y = ..density..), binwidth = 0.1) +
+   facet_grid(cut ~ .)
```



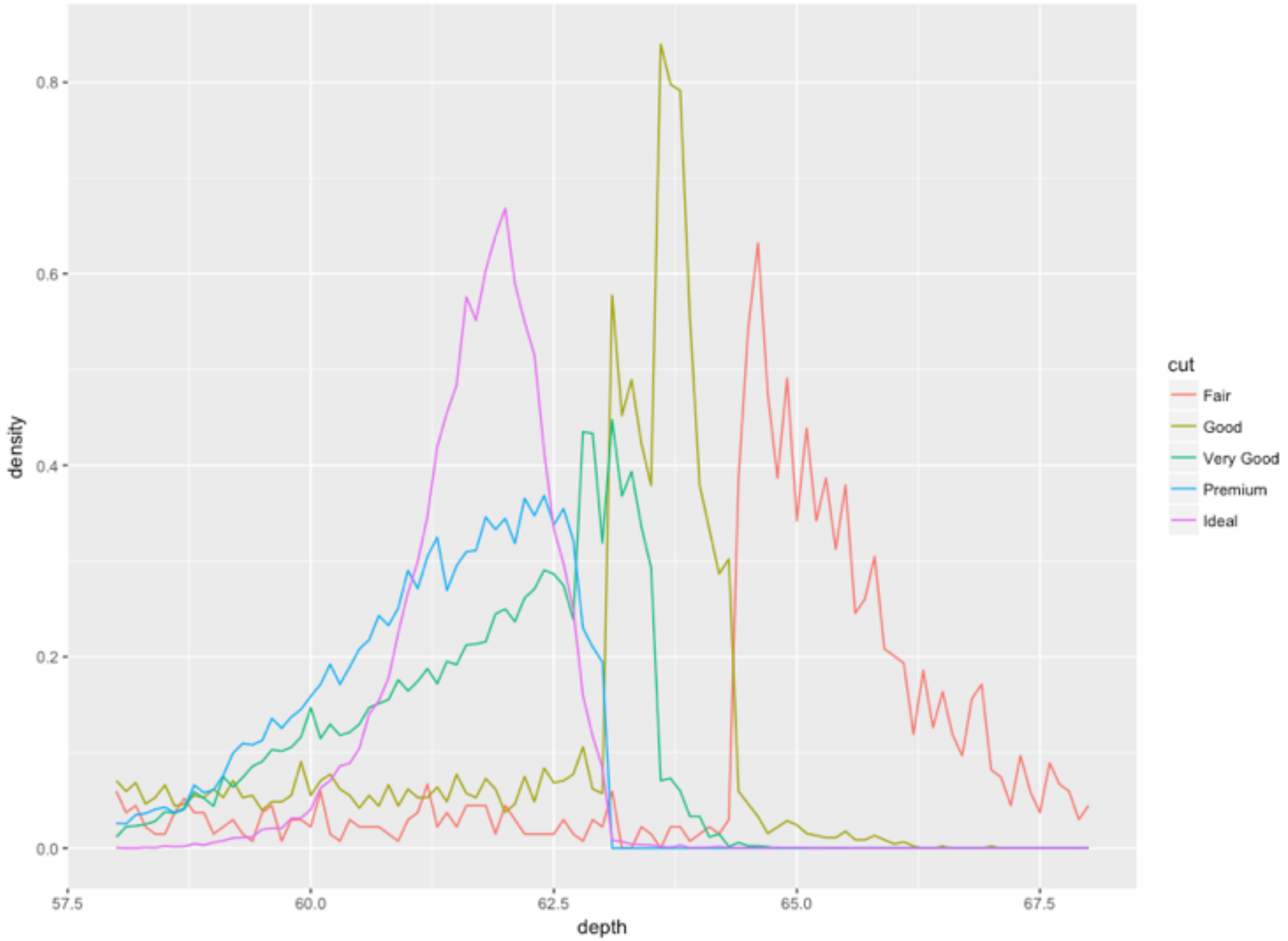
频率多边形图

```
> depth_dist + geom_histogram(aes(fill = cut), binwidth = 0.1,
+   position = "fill")
```



条件密度图

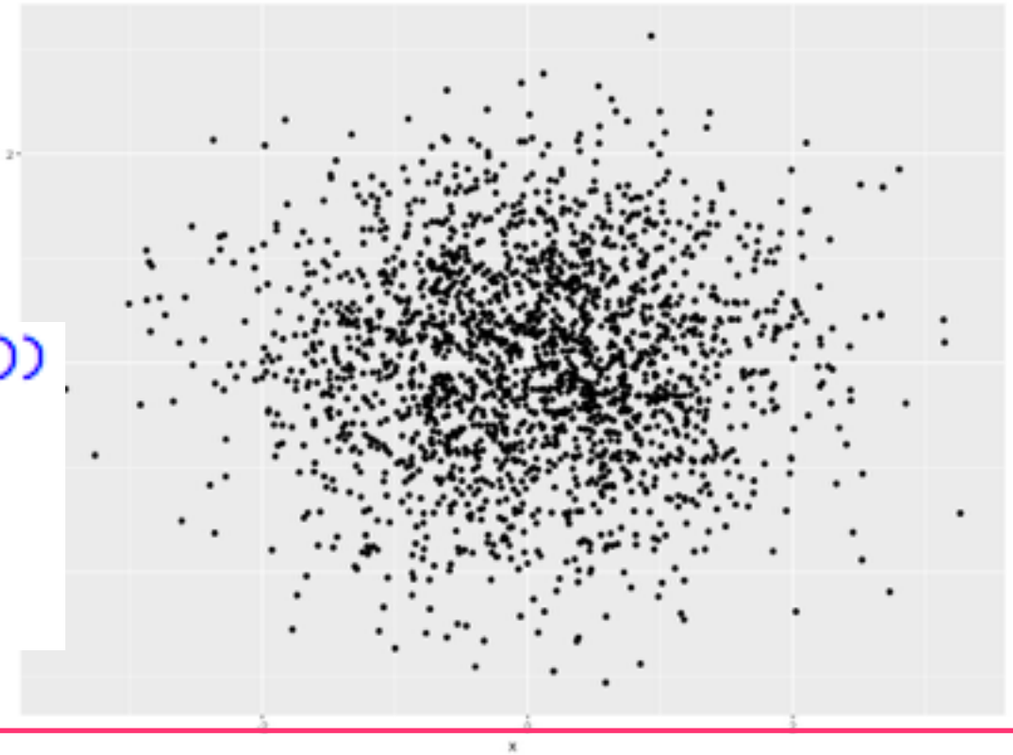
```
> depth_dist + geom_freqpoly(aes(y = ..density.., colour = cut),
+    binwidth = 0.1)
```

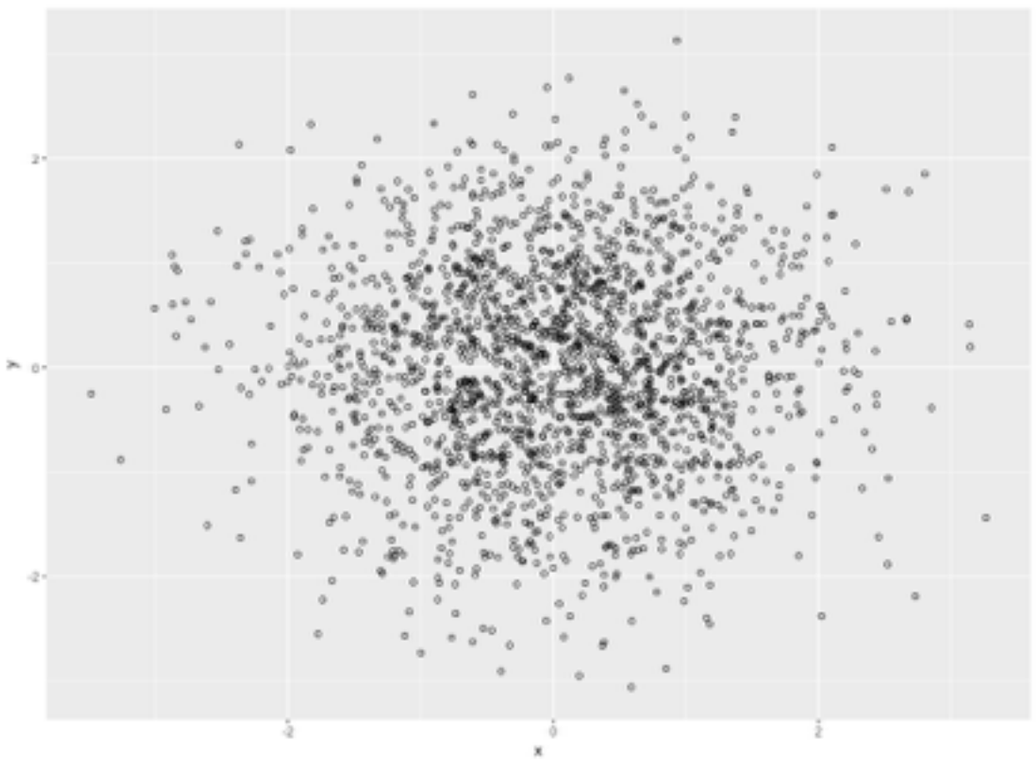


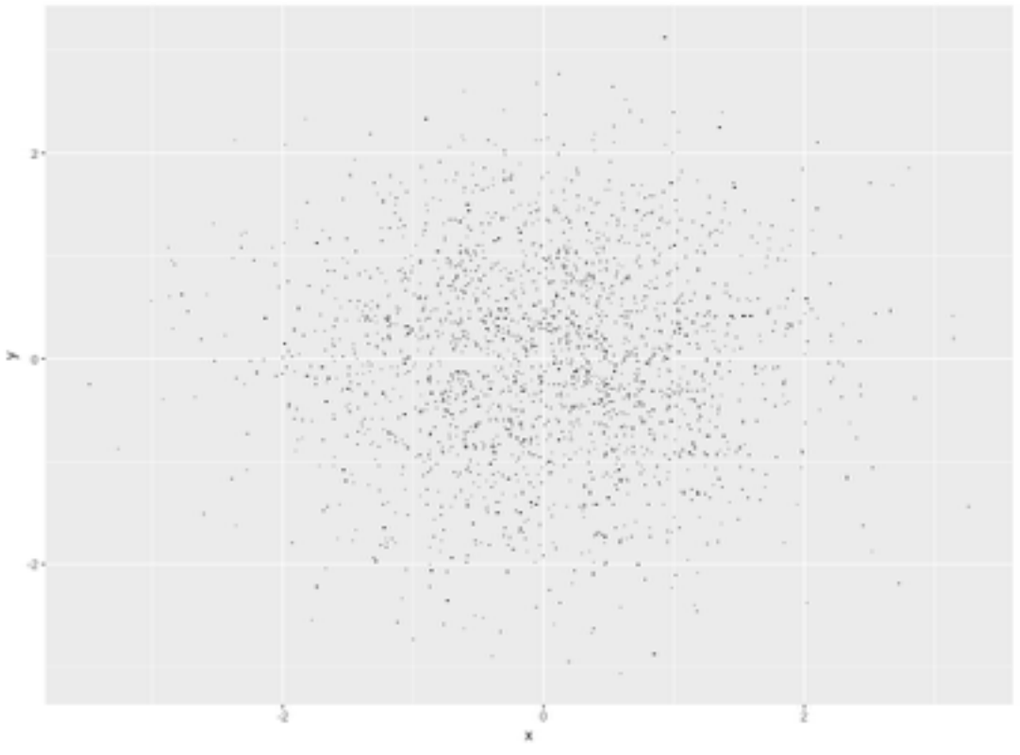
遮盖绘制

使用点的大小

```
> 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
```



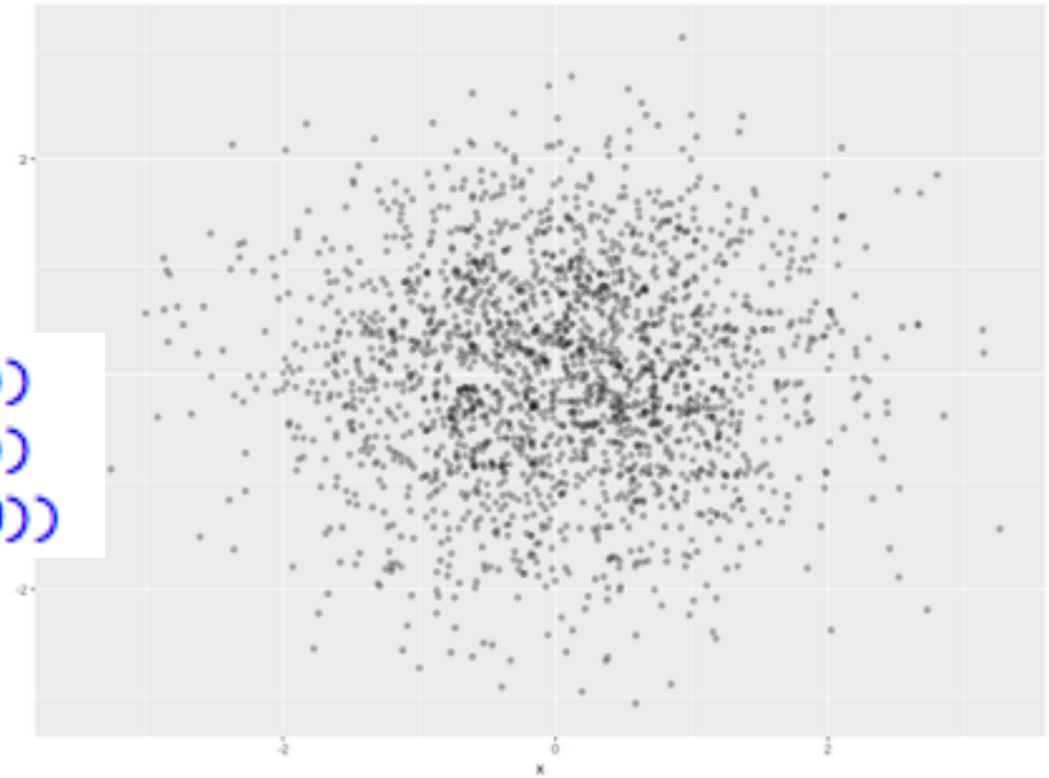


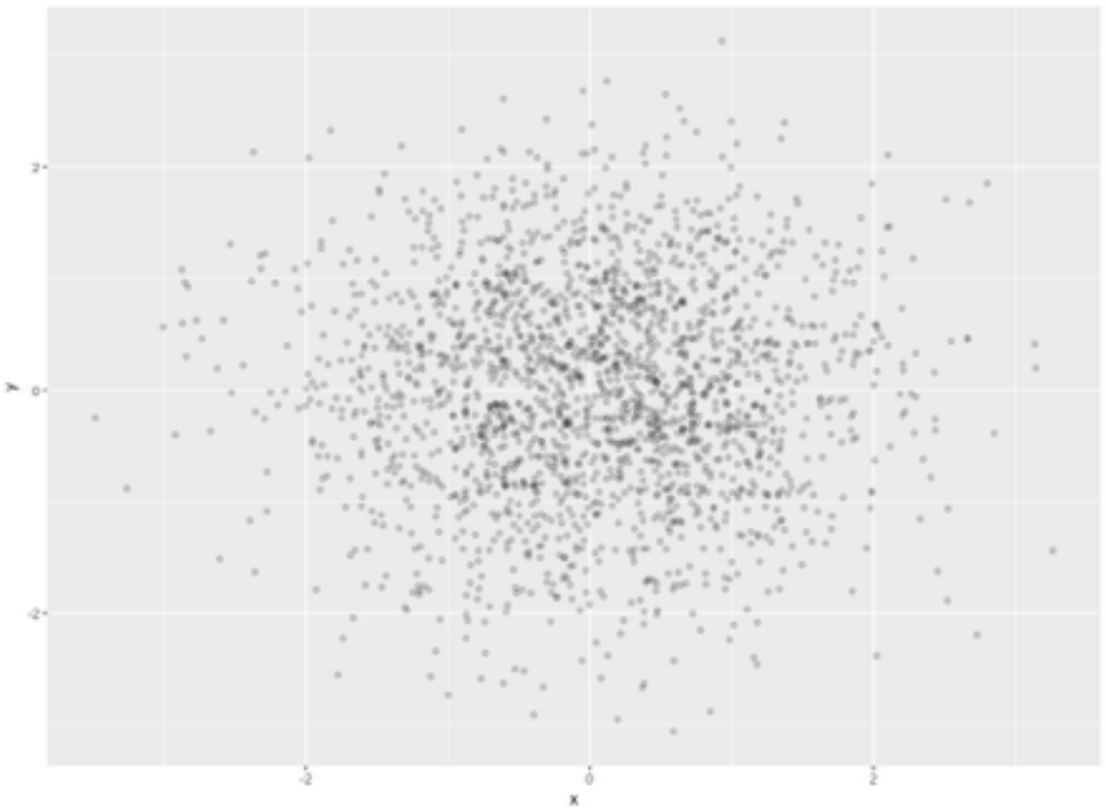


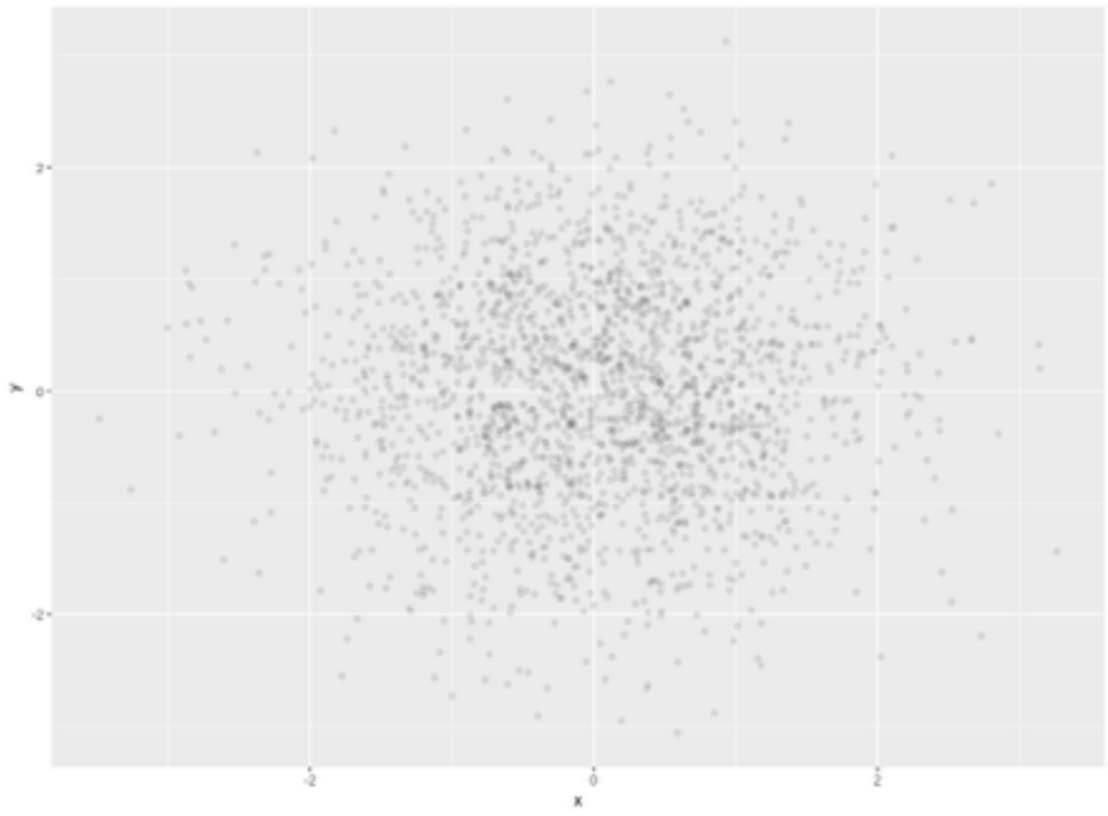
遮盖绘制

使用点的透明度

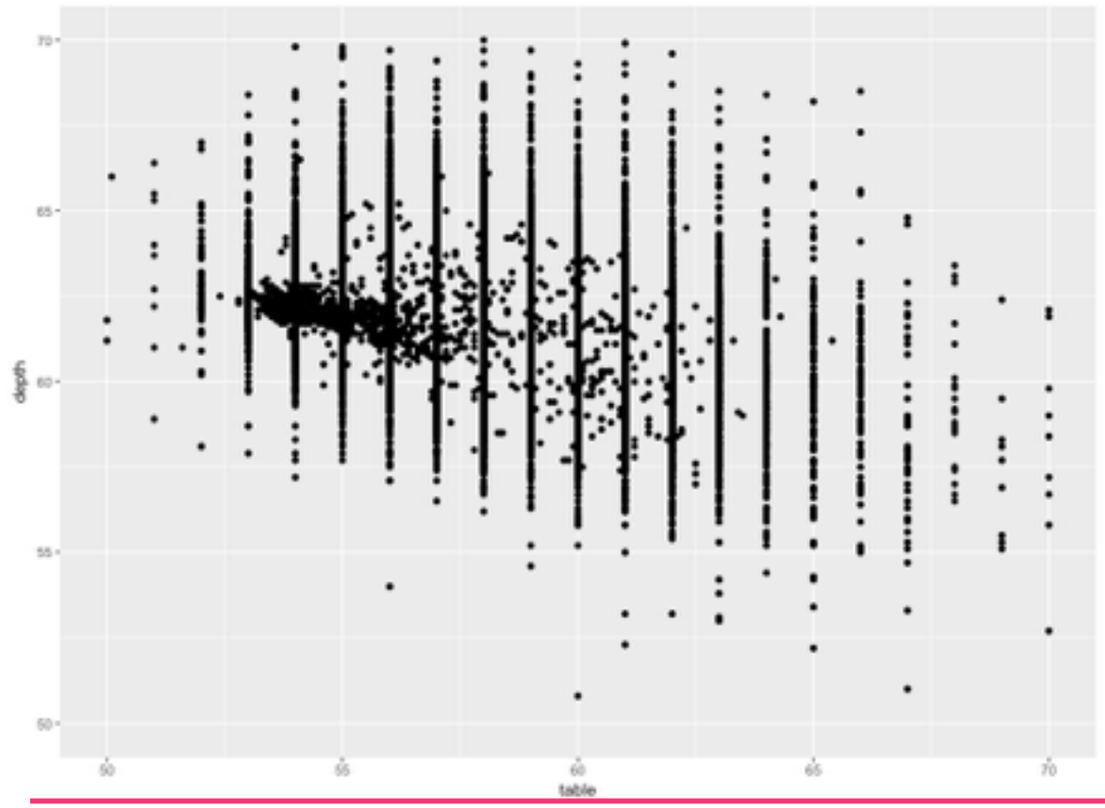
```
> norm + geom_point(colour = alpha("black", 1/3))
> norm + geom_point(colour = alpha("black", 1/5))
> norm + geom_point(colour = alpha("black", 1/10))
```







遮盖绘制

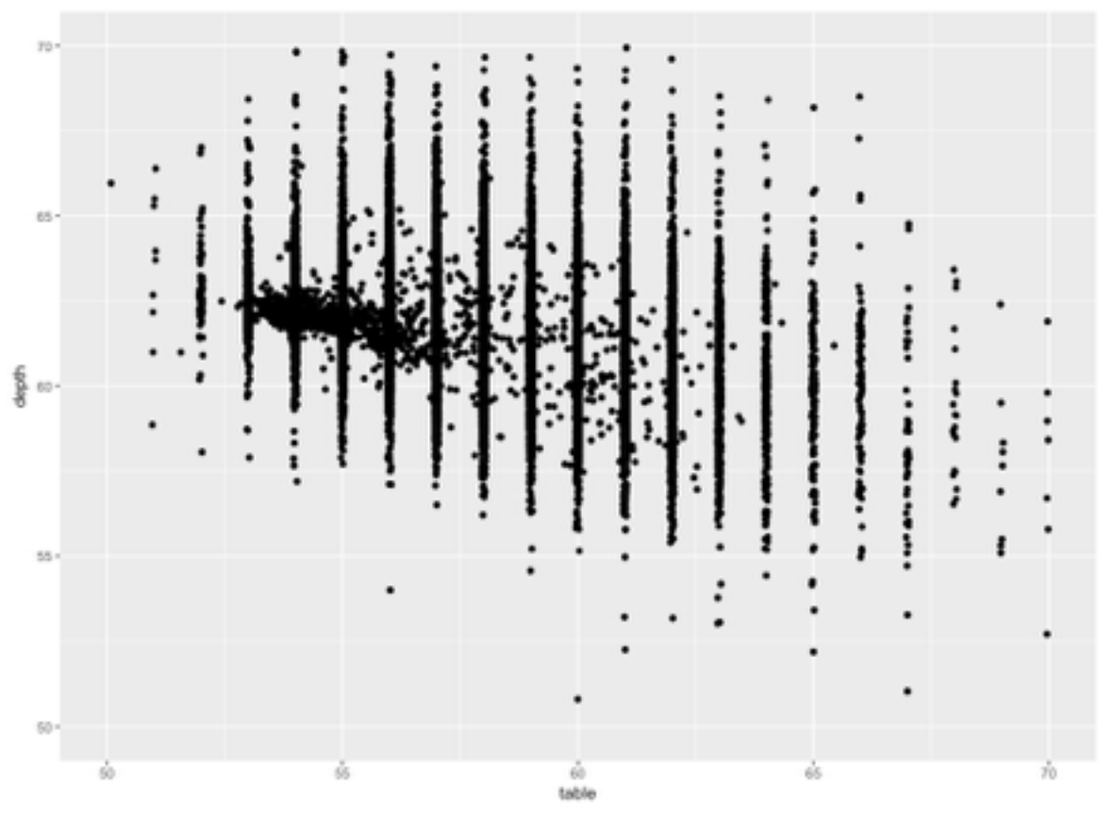


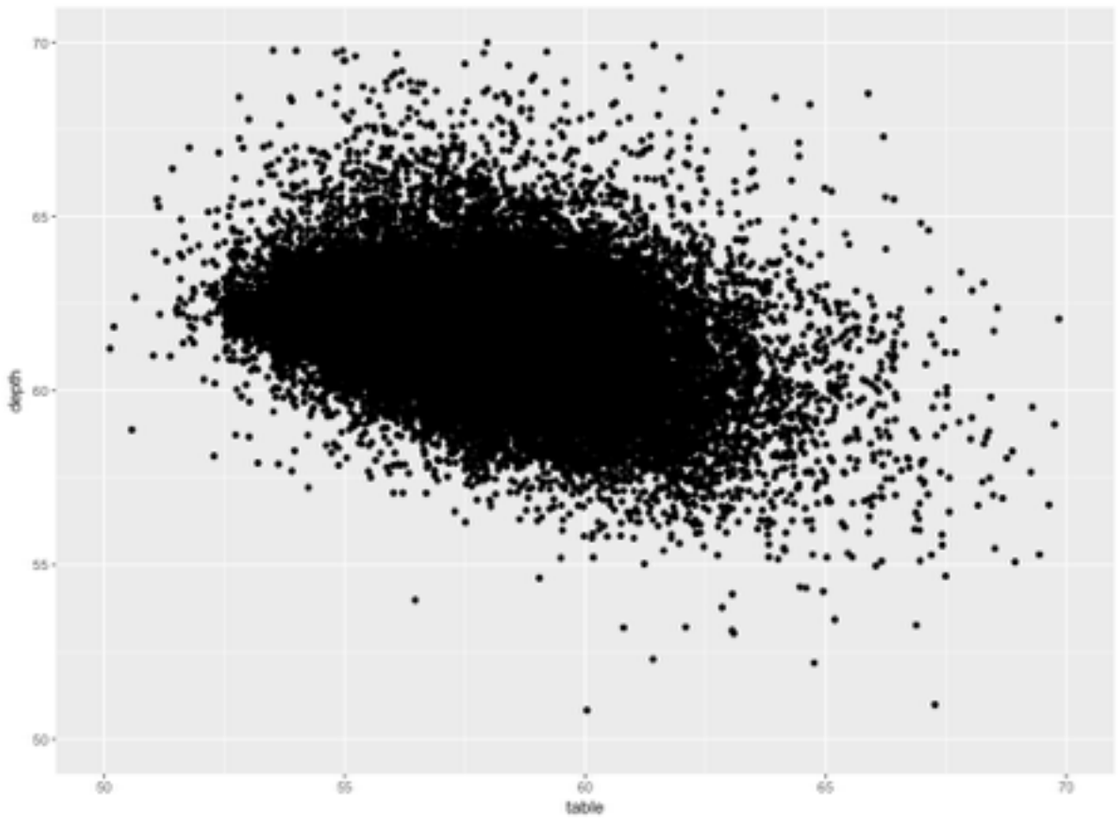
使用随机扰动

```
> td <- ggplot(diamonds, aes(table, depth)) +
+   xlim(50, 70) + ylim(50, 70)
> td + geom_point()
```

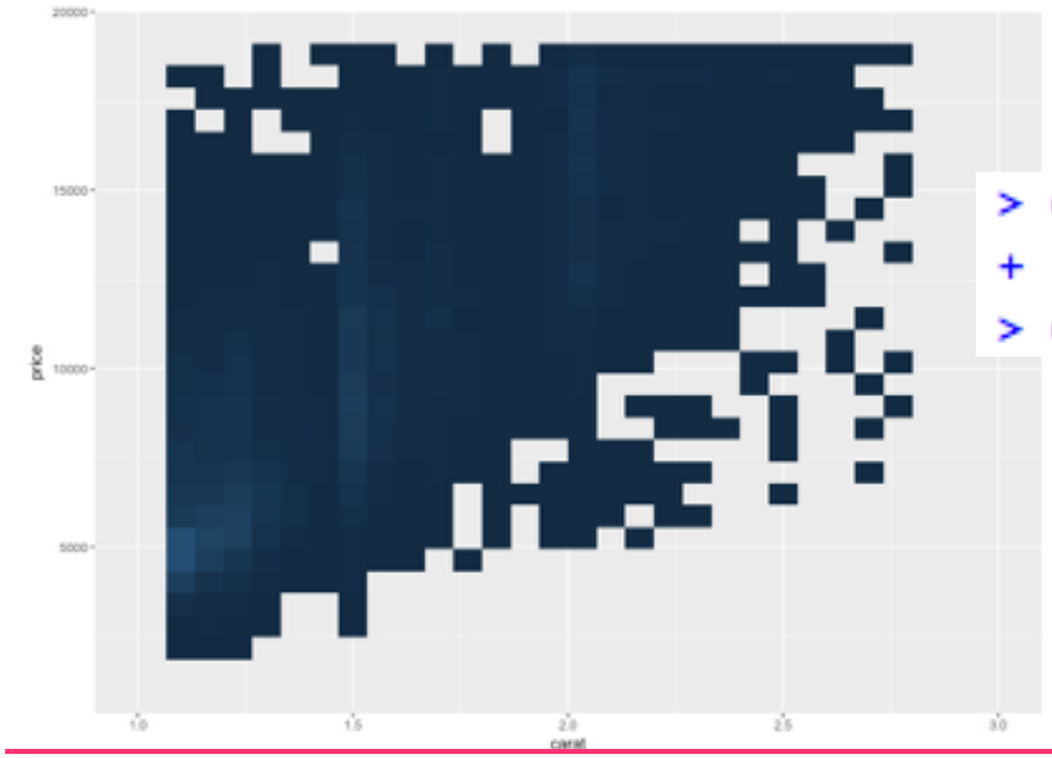
```
> td + geom_jitter()
```

```
> jit <- position_jitter(width = 0.5)
> td + geom_jitter(position = jit)
```





遮盖绘制

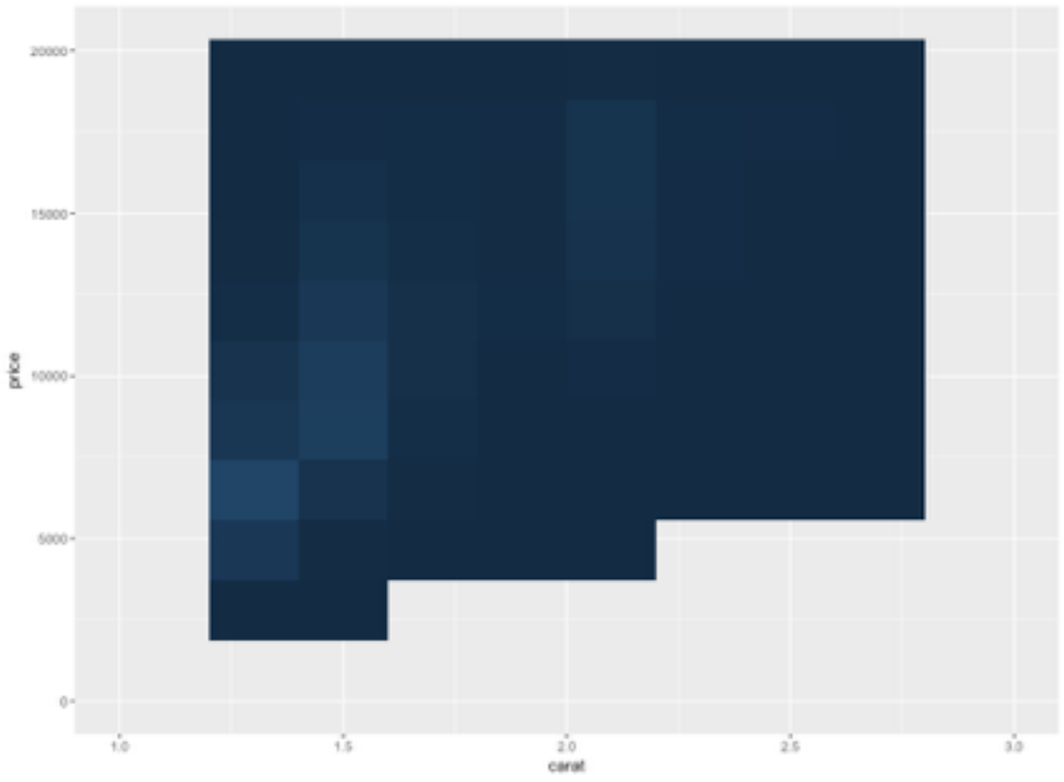


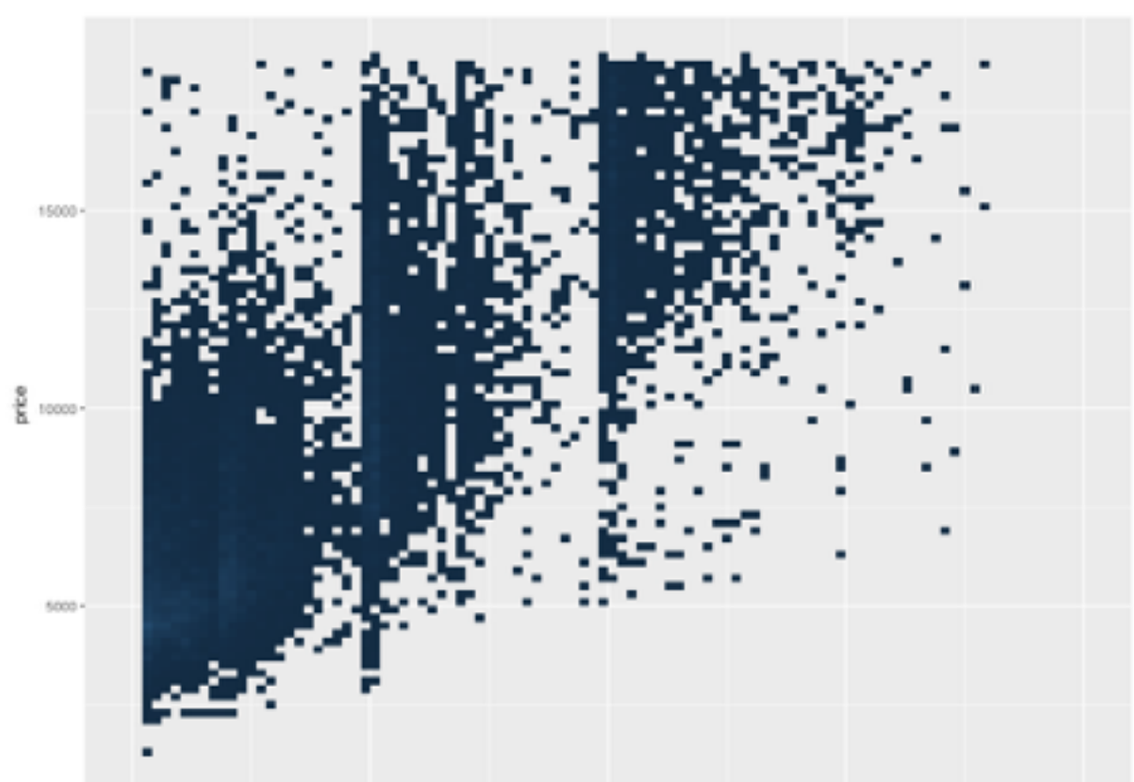
使用分箱计数

```
> d <- ggplot(diamonds, aes(carat, price)) + xlim(1,3) +
+   theme(legend.position = "none")
> d + stat_bin2d()
```

```
> d + stat_bin2d(bins = 10)
```

```
> d + stat_bin2d(binwidth=c(0.02, 200))
```



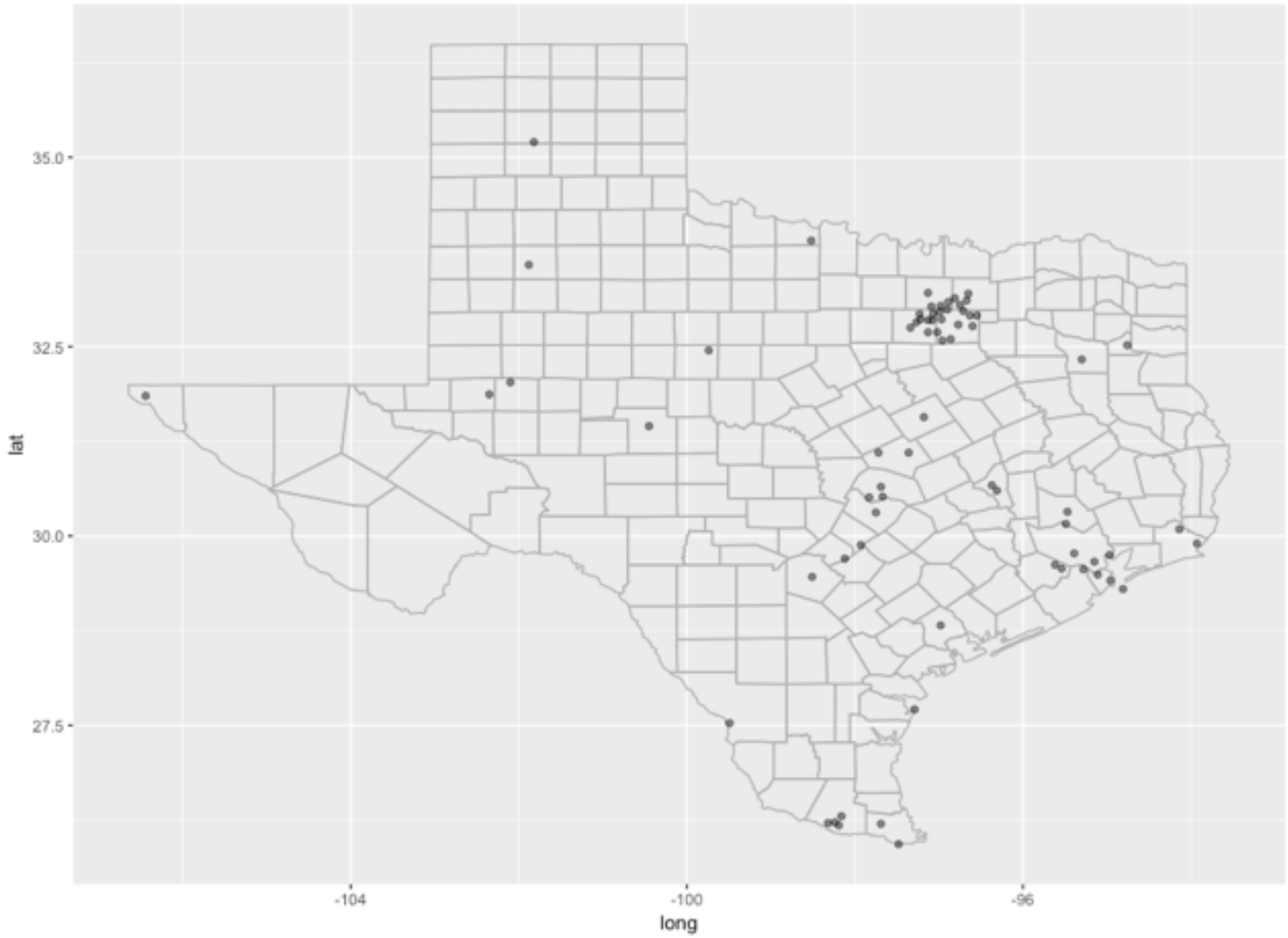


地图

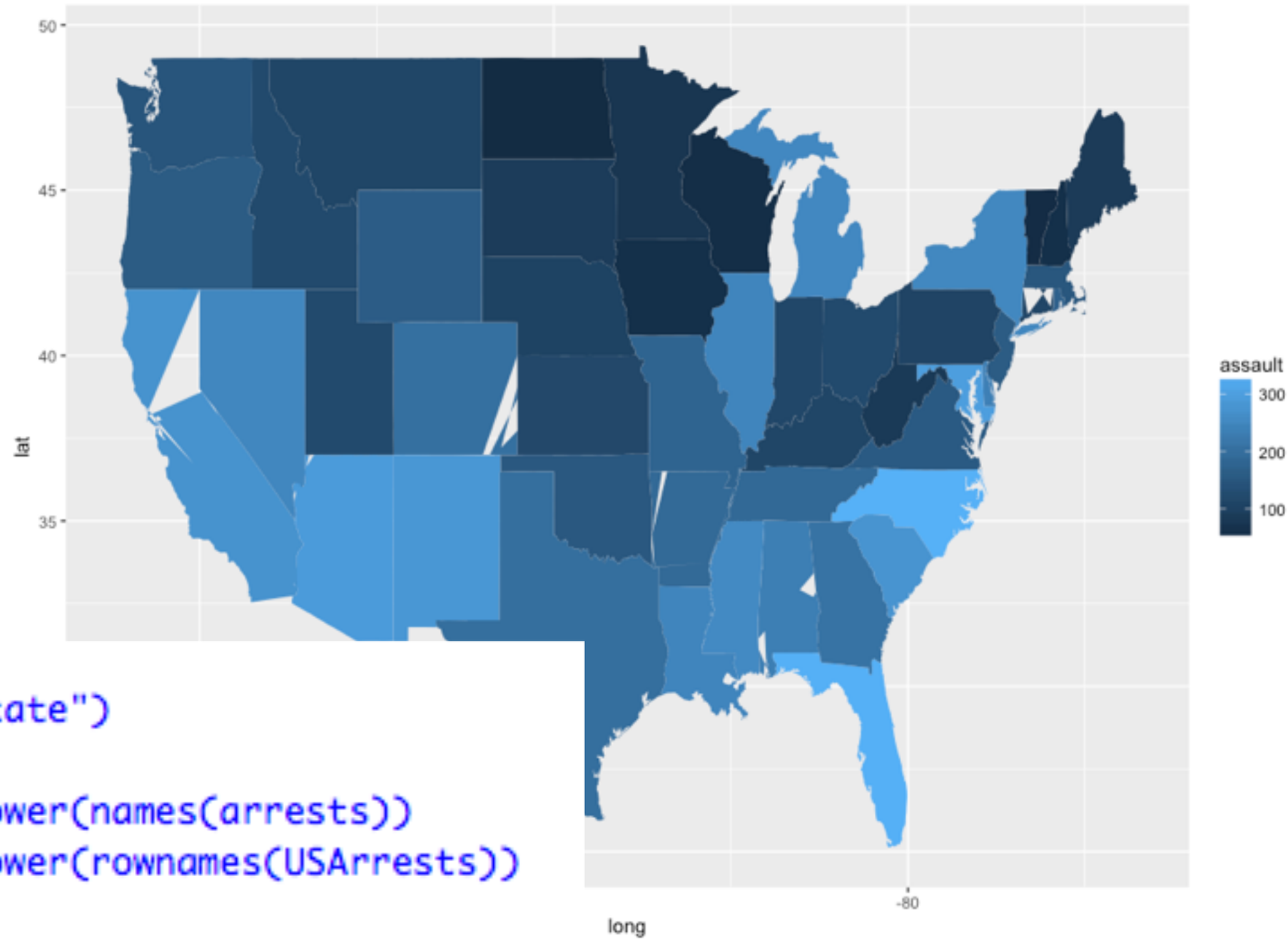
```
> library(maps)
> data(us.cities)
> big_cities <- subset(us.cities, pop > 500000)
> qplot(long, lat, data = big_cities) + borders("state", size = 0.5)
```

地图

```
> tx_cities <- subset(us.cities, country.etc == "TX")
> ggplot(tx_cities, aes(long, lat)) +
+   borders("county", "texas", colour = "grey70") +
+   geom_point(colour = alpha("black", 0.5))
```



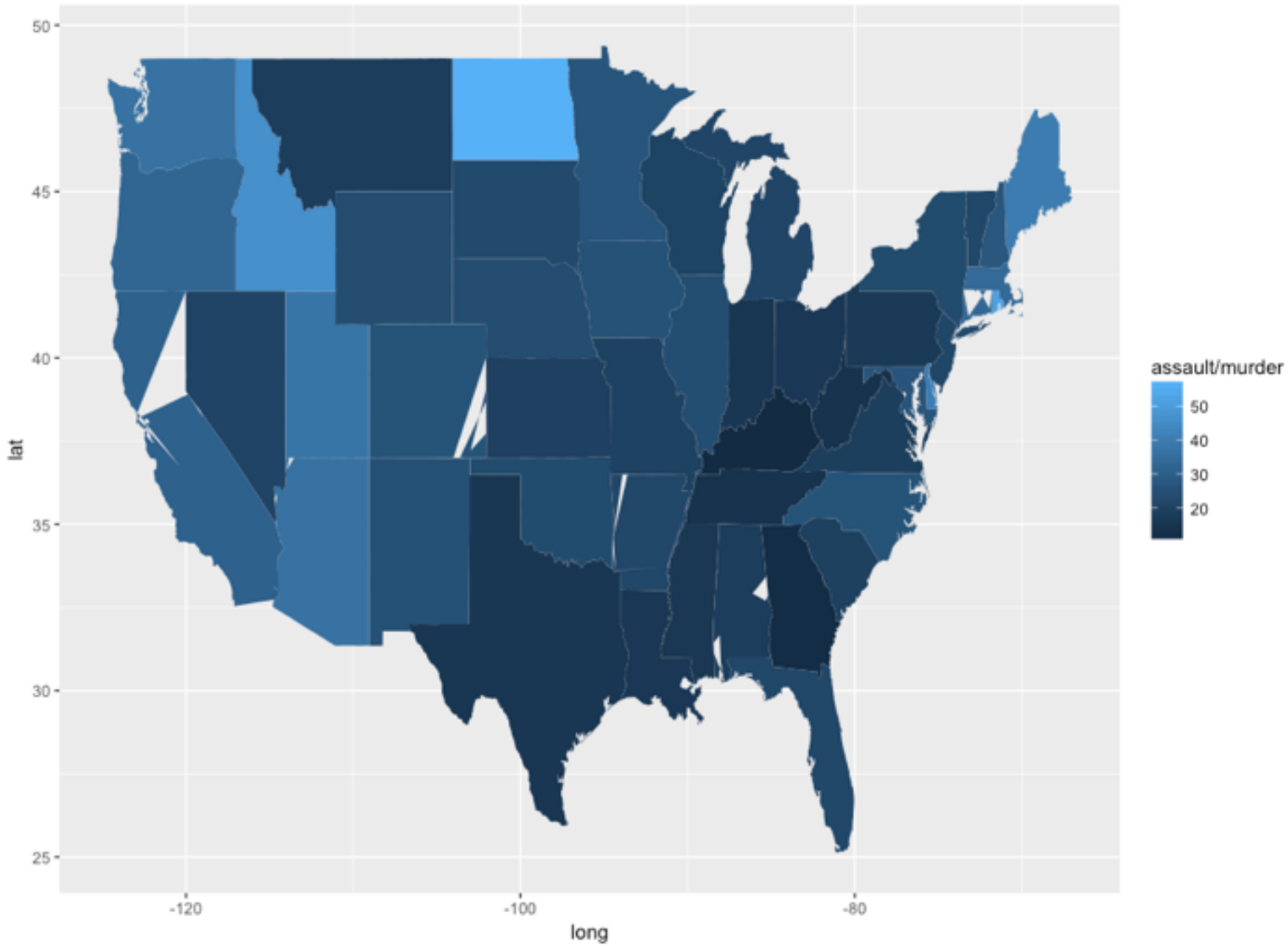
地图




```
> library(maps)
> states <- map_data("state")
> arrests <- USArrests
> names(arrests) <- tolower(names(arrests))
> arrests$region <- tolower(rownames(USArrests))
>
> choro <- merge(states, arrests, by = "region")
> choro <- choro[order(choro$order), ]
> qplot(long, lat, data = choro, group = group,
+   fill = assault, geom = "polygon")
```

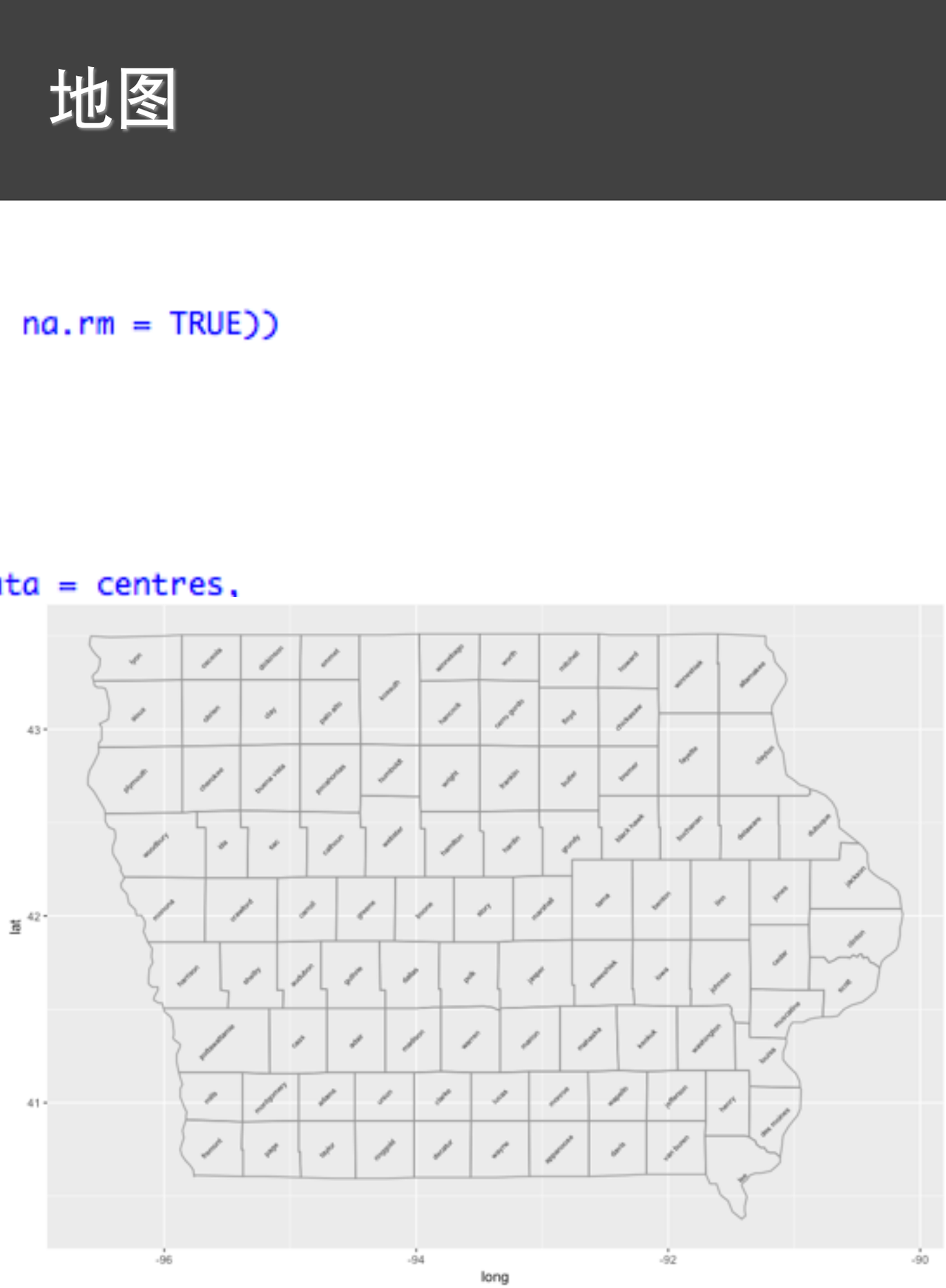
地图

```
> qplot(long, lat, data = choro, group = group,
+   fill = assault / murder, geom = "polygon")
```



地图

```
> library(plyr)
> ia <- map_data("county", "iowa")
> mid_range <- function(x) mean(range(x, na.rm = TRUE))
> centres <- ddply(ia, .(subregion),
+   colwise(mid_range, .(lat, long)))
> ggplot(ia, aes(long, lat)) +
+   geom_polygon(aes(group = group),
+     fill = NA, colour = "grey60") +
+   geom_text(aes(label = subregion), data = centres,
+     size = 2, angle = 45)
```



标度、坐标系和图例

CH6

定位

CH7

提问时间！

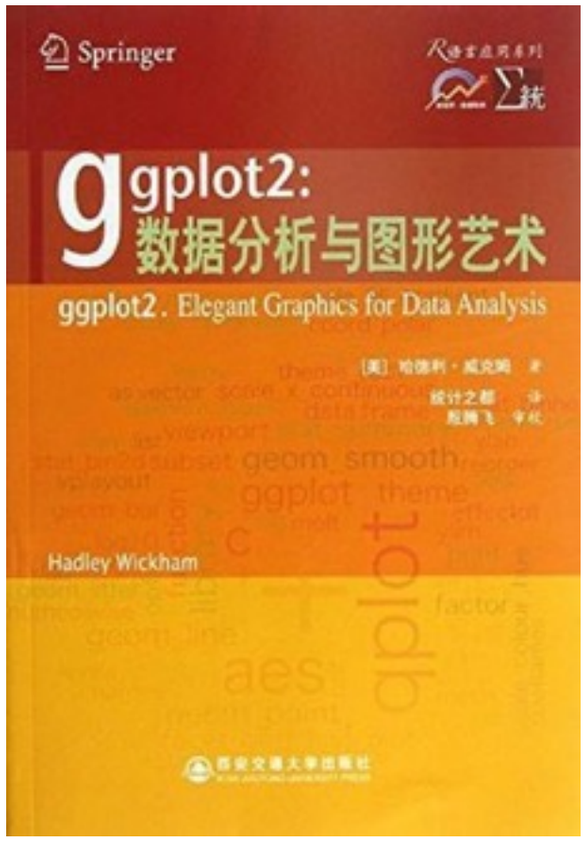
孙惠平

sunhp@ss.pku.edu.cn

练习

练习 - 0030

- ggplot2的4-7章，熟悉所有例子。
- R数据可视化手册的6-13章，熟悉所有例子。
- 教材RIA（第二版）的第19章，熟悉所有例子。
- 看R Graphs Cookbook所有章节



大作业 – 0003

- 完成大作业0003
- 执行结果pdf文件，代码Rmarkdown文件

谢谢！

孙惠平

sunhp@ss.pku.edu.cn