中国石油大学(北京)克拉玛依校区
CHINA UNIVERSITY OF PETROLEUM - BEIJING AT KARAMAY

多元统计分析

Applied Multivariate Statistical Analysis

肖 磊，2026 年 3 月 10 日

Course Logistics

- 课程描述
  - Part I Descriptive Techniques (描述性方法)
    - ✓ Boxplots (箱线图)
    - ✓ Histograms (直方图)
    - ✓ Kernel Densities (核密度)
    - ✓ Scatterplots (散点图)
    - ✓ Chernoff-Flury Faces (脸谱图)
    - ✓ Andrews' Curves (安德鲁斯曲线)
    - ✓ Parallel Coordinate Plots (平行坐标图)
    - ✓ Hexagon Plots (六边形图)
    - ✓ Boston Housing (波士顿房价数据)

Course Logistics

- 课程描述
  - Part II Multivariate Random Variables (多元随机变量)
    - ✓ A Short Excursion into Matrix Algebra (矩阵代数初探)
    - ✓ Moving to Higher Dimensions (高维中的相关概念)
    - ✓ Multivariate Distributions (多元分布)
    - ✓ Theory of the Multinormal (多元正态分布理论)
    - ✓ Theory of Estimation (估计理论)
    - ✓ Hypothesis Testing (假设检验)

Course Logistics

- 课程描述
  - Part III Multivariate Techniques (多元方法)
    - ✓ Regression Models (回归模型)
    - ✓ Variable Selection (变量选择)
    - ✓ Decomposition of Data Matrices by Factors (数据矩阵的因子分解)
    - ✓ Principal Components Analysis (主成分分析)
    - ✓ Factor Analysis (因子分析)
    - ✓ Cluster Analysis (聚类分析)
    - ✓ Discriminant Analysis (判别分析)
    - ✓ Correspondence Analysis (对应分析)
    - ✓ Canonical Correlation Analysis (典型相关分析)
    - ✓ Multidimensional Scaling (多维标度分析)
    - ✓ Conjoint Measurement Analysis (联合度量分析)

Course Logistics

- 预期目标：课程结束时应具备以下能力
  - 使用 $\mathfrak{R}$ 对多维数据进行可视化.
  - 掌握描述一个数据集相关关系的基本概念，尤其是用于定义汇总统计量的矩阵运算.
  - 掌握多元统计分析中常用的基本概率方法.
  - 掌握多元正态分布的基本理论，了解两个伴随分布：Wishart 分布, Hoteling 分布.
  - 能够利用极大似然理论对多元正态总体进行统计推断 (推导估计量、置信区间和假设检验).

Course Logistics

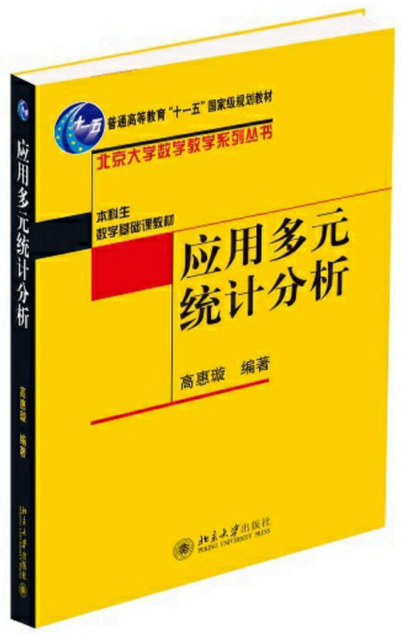
- 预期目标：课程结束时，同学们应具备以下能力
  - 会应用多元统计的方法分析多维数据集，如
    - Regression models (回归模型)
    - Variable selection (变量选择)
    - Decomposition of Data Matrices by Factors (数据矩阵的因子分解)
    - Principal Components Analysis (主成分分析)
    - Factor Analysis (因子分析)
    - Cluster Analysis (聚类分析)
    - Discriminant Analysis (判别分析)
    - Correspondence Analysis (对应分析)
    - Canonical Correlation Analysis (典型相关分析)

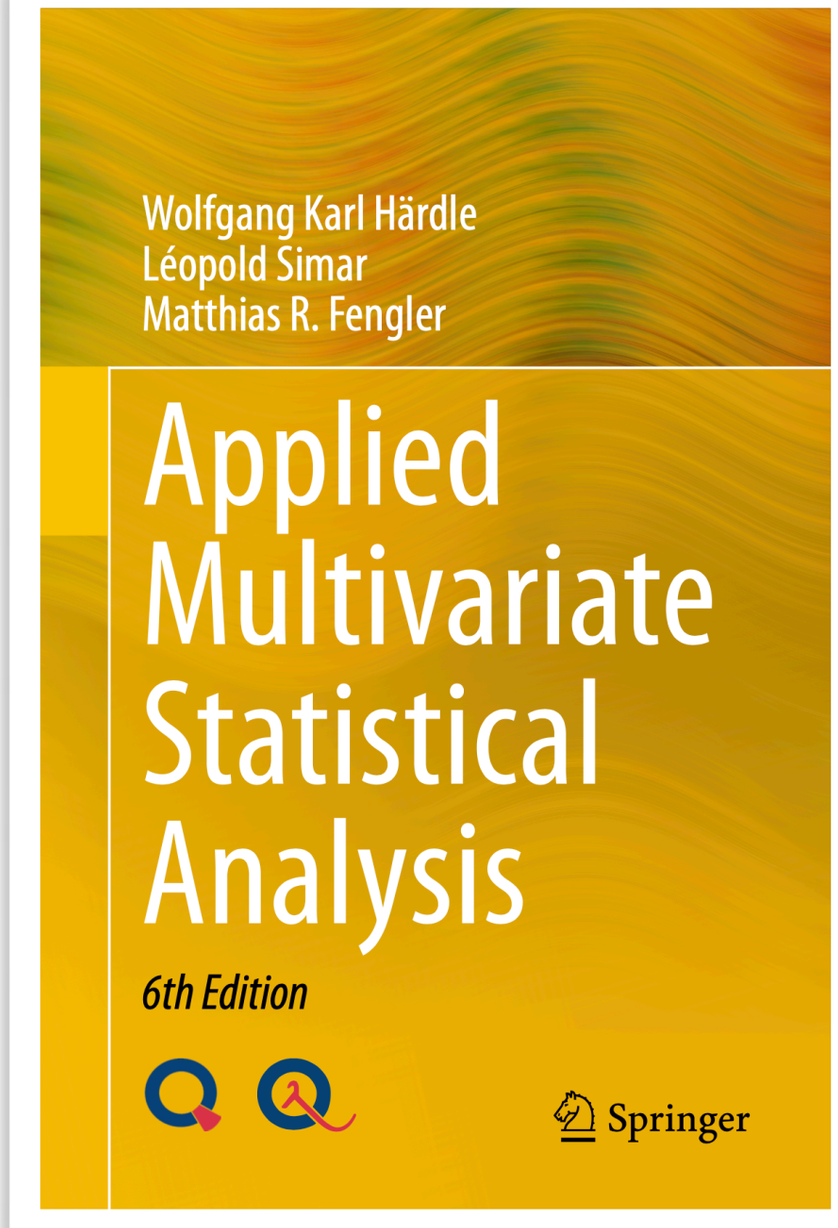
Course Logistics

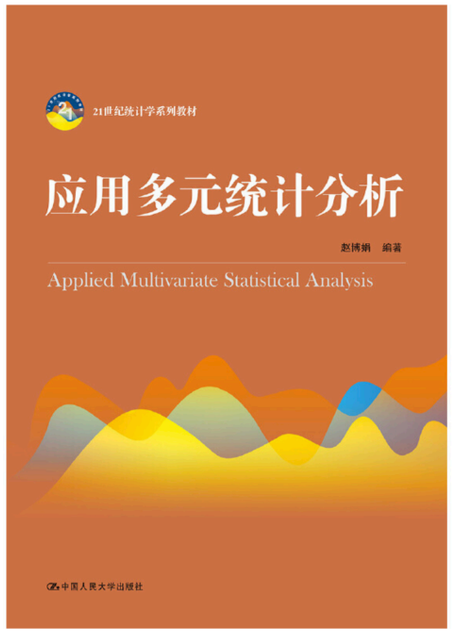
- 基本信息:
  - 作业：每周五布置本周的课后作业，一周时间完成.
  - 测验：根据授课内容进展情况安排 5 次.
  - 报告：教学内容完成后进行.
  - 答疑：通过 QQ、微信答疑，或预约时间答疑.
  - 邮箱: xiaolei@cup.edu.cn

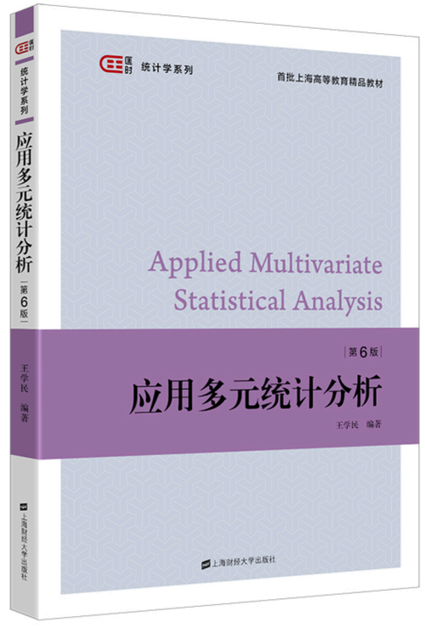
Course Logistics

- 教材与参考书：









Course Logistics

- 成绩结构：课程最终成绩由以下各部分按相应权重确定
  - 过程性考核成绩 (50%).
  - 期末结课报告 (50%)

作业使用 LaTeX 编辑完成，生成 PDF 格式后提交至指定邮箱. 作业题目当中每一问赋 2 分：按时完成提交且答案基本正确得 2 分；按时完成提交且答案部分正确得 1 分；未按时提交或答案完全错误得 0 分.
每位同学可以有一次迟交作业的机会，但不得超过指定时间的三日.

- 过程性考核：作业 (60%)、考勤 (15%)、阶段测验 (25%).
  - 作业成绩：课程全部教学 12 周的周作业成绩的平均分，其中

$$\text{每周作业成绩} = \frac{\text{周作业得分}}{\text{该周作业总分}} \times 100$$

  - 考勤：每缺勤 1 次扣除 2.5 分，缺勤次数超过 6 次者取消考试资格.

    每请假 1 次扣 1 分，请假次数超过 8 次者取消考试资格.

  - 阶段测验：安排 5 次阶段测验，每次测验满分 5 分，5 次测验的成绩之和为阶段测验的成绩.

Part I Descriptive Techniques

描述性方法

Chapter 1 Comparison of Batches (分类比较)

- 概述
  - Boxplots (箱线图)
  - Histograms (直方图)
  - Kernel Densities (核密度)
  - Scatterplots (散点图)
  - Chernoff-Flury Faces (脸谱图)
  - Andrews' Curves (安德鲁斯曲线)
  - Parallel Coordinate Plots (平行坐标图)
  - Hexagon Plots (六边形图)

Data (数据)

- 多元统计分析
  - 高维当中数据的分析与推断 (建模).
- 向量 (变量): $\boldsymbol{X} = \left(X_1, X_2, \ \ldots, X_p\right) \in \mathbb{R}^p$
- 数据: $\boldsymbol{X}$ 的 n 个观测结果, $\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ \ldots, \boldsymbol{x}_n\right)$, 其中

$$\begin{matrix} X_1 & X_2 & \cdots & X_p \\ \downarrow & \downarrow & \cdots & \downarrow \end{matrix}$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{matrix} \longleftarrow \ \boldsymbol{x}_1^{\mathrm{T}} \\ \longleftarrow \ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \longleftarrow \ \boldsymbol{x}_n^{\mathrm{T}} \end{matrix}$$

$$\boldsymbol{x}_i^{\mathrm{T}} = \left(x_{i1}, x_{i2}, \ \ldots, x_{ip}\right) \ , \quad \boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} , \quad i = 1, 2, \ \ldots, n$$

- 第一步：如何观察数据
  - 描述性方法

Data (数据)

- 通过描述性方法我们可以回答以下问题
  - X 中的某个变量比其它变量更为分散?
  - X 中是否有些变量用来对数据进行分类?
  - X 的各变量中是否存在异常值 (outliers)?
  - 数据的分布是否接近正态分布?
  - X 中变量的低维线性组合是否有非正态的表现?

Boxplots (箱线图)

- 表示变量分布的一种可视化方法.
  - 位置 (location)， 偏度 (skewness)， 分散程度 (spread)， 尾部长度 (tail length)， 离群点 (outlying points)
- 五数总括 (Five-Number Summary) 的一种图示.
  - 最小值 $\left(x_{\min}\right)$, 下四分位数 $\left(F_L\right)$, 中位数 (M), 上四分位数 $\left(F_U\right)$, 最大值 $\left(x_{\max}\right)$.

数据: $x_1, x_2, \ldots, x_n \Longrightarrow$ 顺序统计量: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

$$x_{\min} = x_{(1)}$$

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & n \text{ even} \end{cases}$$

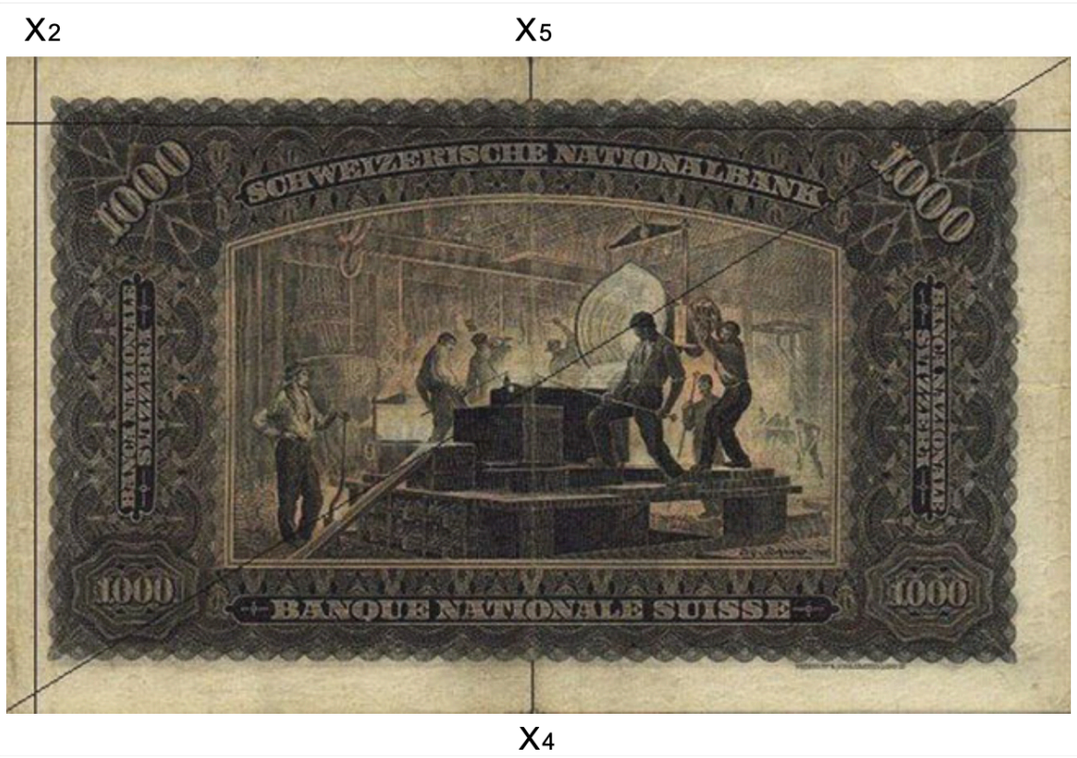
$$x_{\max} = x_{(n)}$$

Boxplots (箱线图)

- 瑞银钞票数据

```
library(mclust)
data(banknote)
str(banknote)
head(banknote)
```

```
> str(banknote)
'data.frame':   200 obs. of  7 variables:
 $ Status  : Factor w/ 2 levels "counterfeit",..: 2 2 2 2 2
 $ Length  : num  215 215 215 215 215 ...
 $ Left    : num  131 130 130 130 130 ...
 $ Right   : num  131 130 130 130 130 ...
 $ Bottom  : num  9 8.1 8.7 7.5 10.4 9 7.9 7.2 8.2 9.2 ...
 $ Top     : num  9.7 9.5 9.6 10.4 7.7 10.1 9.6 10.7 11 10 ...
 $ Diagonal: num  141 142 142 142 142 ...
> head(banknote)
   Status Length  Left Right Bottom  Top Diagonal
1 genuine  214.8 131.0 131.1    9.0  9.7    141.0
2 genuine  214.6 129.7 129.7    8.1  9.5    141.7
3 genuine  214.8 129.7 129.7    8.7  9.6    142.2
4 genuine  214.8 129.7 129.6    7.5 10.4    142.0
5 genuine  215.0 129.6 129.7   10.4  7.7    141.8
6 genuine  215.7 130.8 130.5    9.0 10.1    141.4
```



- X_1 = 钞票的长度
- X_2 = 钞票的宽度 (左侧)
- X_3 = 钞票的宽度 (右侧)
- X_4 = 钞票内框到下边界的距离
- X_5 = 钞票内框到上边界的距离
- X_6 = 中心图片对角线的长度

Boxplots (箱线图)

- 瑞银钞票数据

```
x = banknote$Length
summary(x) # 五数总括 + 平均值
```

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.617   2.617   2.617   2.617   2.617   2.617
```

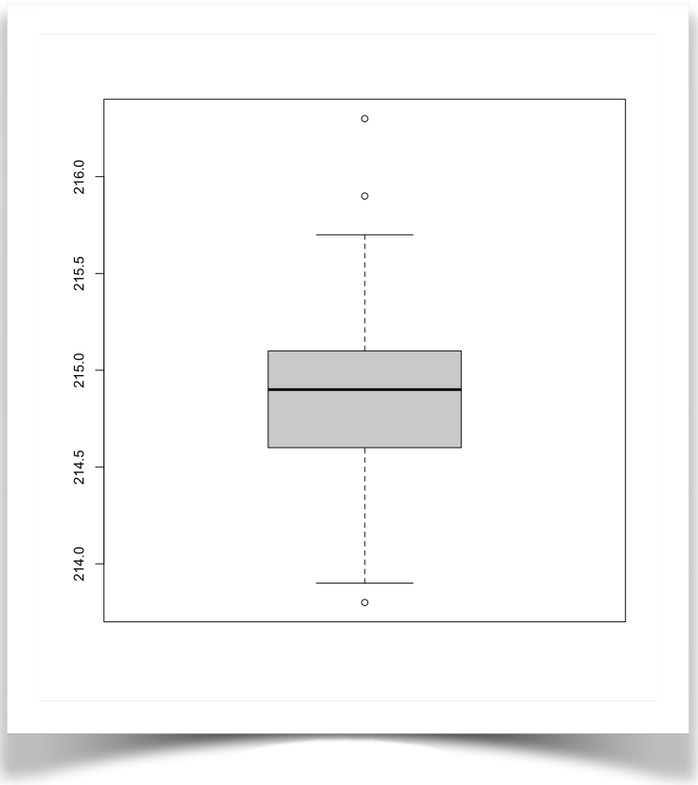
```
quantile(x, probs = c(0, 0.25, 0.5, 0.75, 1)) # 五数总括
```

```
> quantile(x, probs = c(0, 0.25, 0.5, 0.75, 1))
      0%      25%      50%      75%     100%
2.616544 2.616544 2.616544 2.616544 2.616544
```

Boxplots (箱线图)

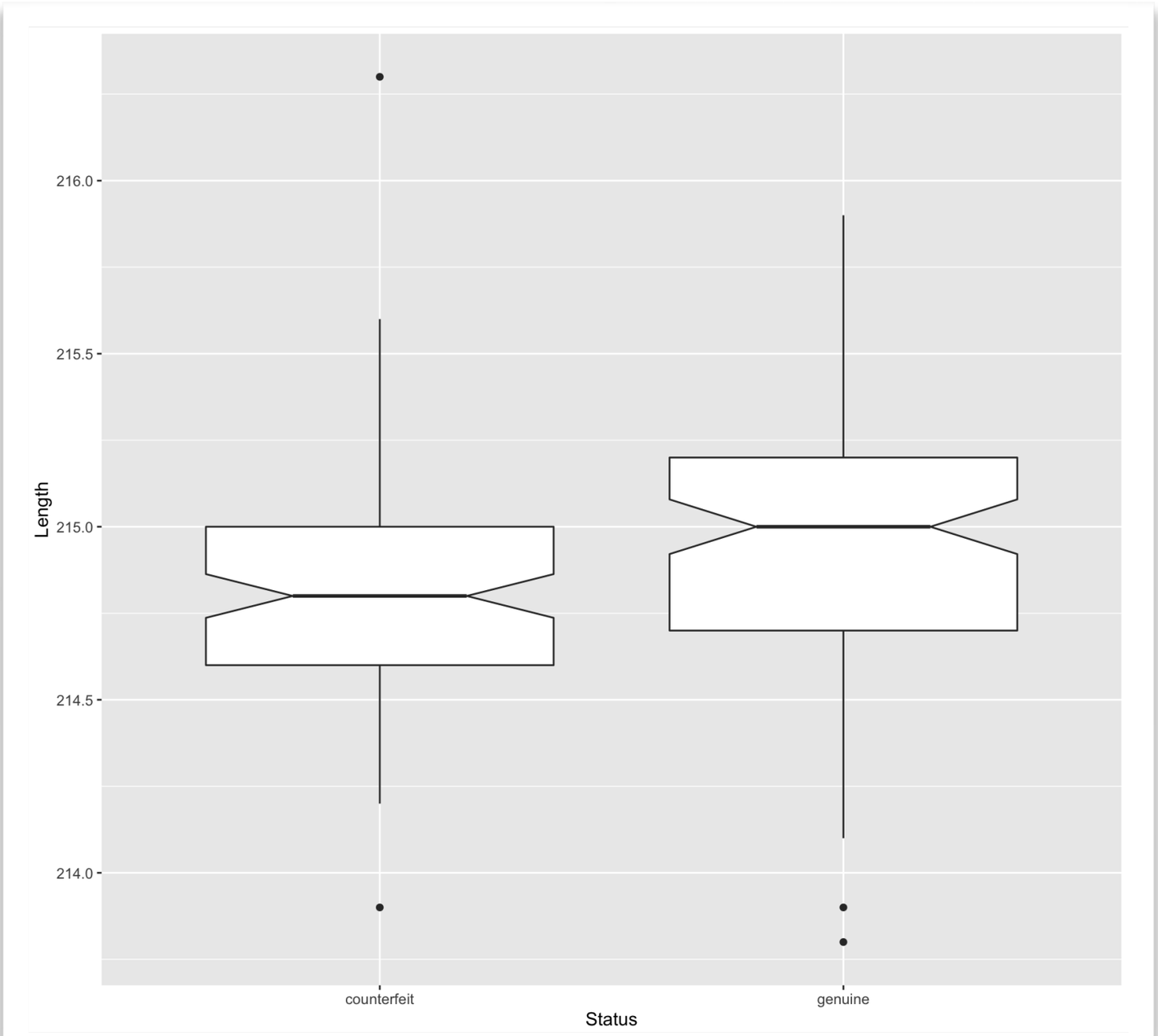
- 箱线图的构造
  - 以 F_L 和 F_U 为边界画一个矩形框 (即，有 50% 数据位于框内)
  - 在框内中位数的位置画一条实线.
  - 从矩形框的上、下端至最远的非异常值点 (默认为四分位差的1.5倍远处) 画一条虚线.
  - 画 “∘” 表示位于 $F_{U\,L} \pm 1.5\left(F_U - F_L\right)$ 之外的异常值点.

boxplot(x)



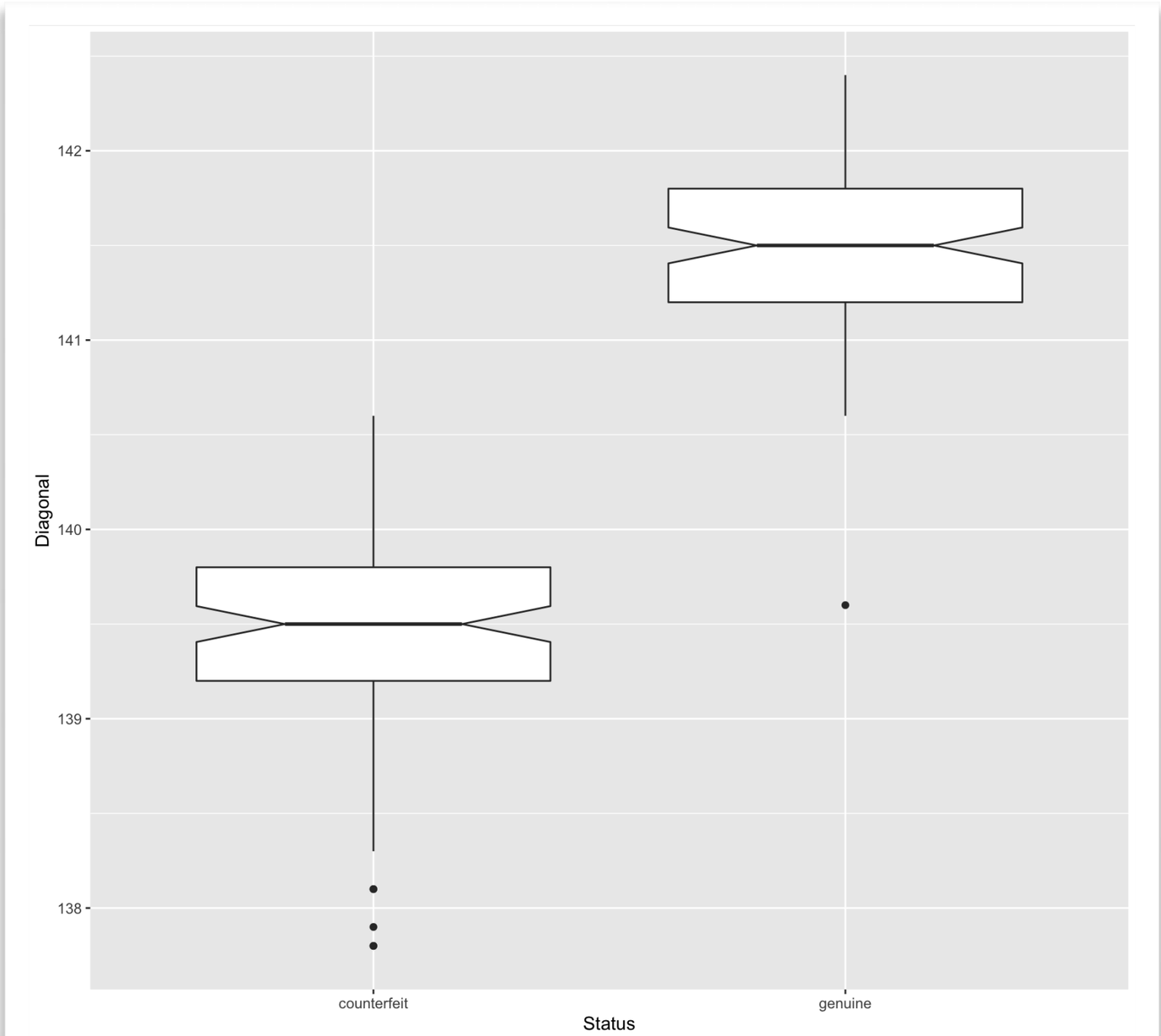
Boxplots

```
library(ggplot2)
ggplot(data = banknote, aes(x = Status, y = Length)) +
  geom_boxplot(notch = TRUE)
```



Boxplots

```
ggplot(data = banknote, aes(x = Status, y = Diagonal)) +
  geom_boxplot(notch = TRUE)
```



Histograms (直方图)

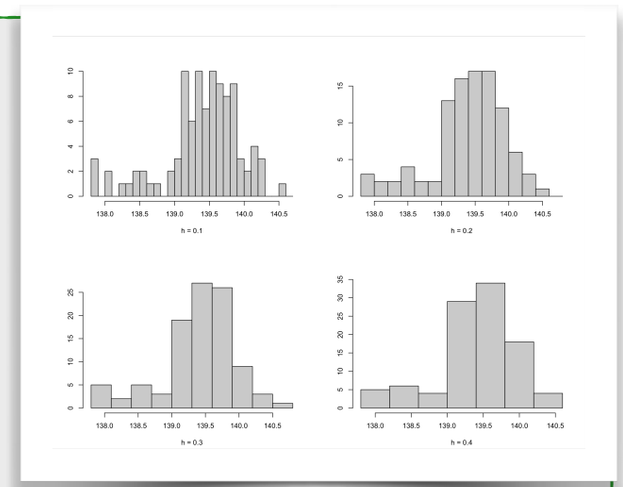
- 直方图是对密度的估计，它对数据的分布提供了一种很好的图示.
- 其思想是通过计算以 x_0 为初始点的一系列连续区间 (bin) 内观测数据的个数来局部表示数据密度.
- 长为 h 的区间：$B_j\left(x_0,\, h\right) = \left[x_0 + (j-1)h,\, x_0 + jh\right)$ ， $j \in \mathbb{Z}$.
- 数据 $x_1,\, x_2,\, \ldots,\, x_n$ 看作是来自密度为 f 的独立同分布 (i.i.d.) 的一个样本.

- 直方图定义如下:

$$\widehat{f_h}\,(x) = \frac{1}{nh}\sum_{j\in\mathbb{Z}}\sum_{i=1}^{n} \mathrm{I}\left\{x_i \in B_j\left(x_0,\, h\right)\right\} \cdot \mathrm{I}\left\{x \in B_j\left(x_0,\, h\right)\right\}$$

Histograms (直方图)

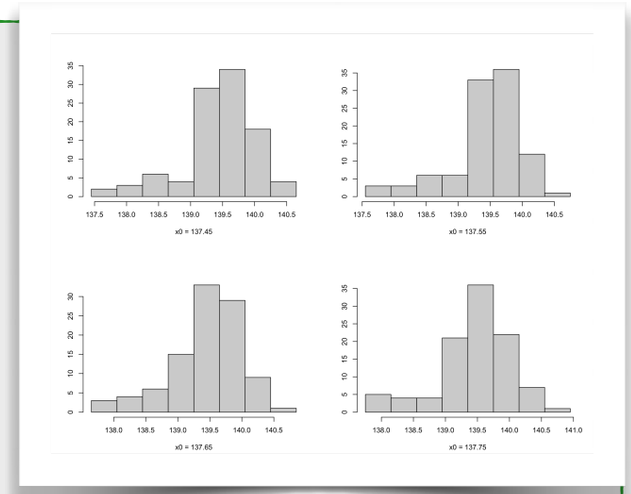
```
par(mfrow=c(2, 2))
x_0 = 137.8
h1 = 0.1
breaks.vec1 = seq(from = x_0, by = h1, length.out = 30)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec1,
     xlab = 'h = 0.1', ylab = '', main = '')
h2 = 0.2
breaks.vec2 = seq(from = x_0, by = h2, length.out = 16)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec2,
     xlab = 'h = 0.2', ylab = '', main = '')
h3 = 0.3
breaks.vec3 = seq(from = x_0, by = h3, length.out = 11)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec3,
     xlab = 'h = 0.3', ylab = '', main = '')
h4 = 0.4
breaks.vec4 = seq(from = x_0, by = h4, length.out = 8)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec4,
     xlab = 'h = 0.4', ylab = '', main = '')
```



$$h_{\text{opt}} = \left(\frac{24\sqrt{\pi}}{n} \right)^{1/3}$$

Histograms (直方图)

```
par(mfrow=c(2, 2))
x_0 = 137.45
breaks.vec5 = seq(from = x_0, by = h4, length.out = 9)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec5,
     xlab = 'x0 = 137.45', ylab = '', main = '')
x_0 = 137.55
breaks.vec6 = seq(from = x_0, by = h4, length.out = 9)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec6,
     xlab = 'x0 = 137.55', ylab = '', main = '')
x_0 = 137.65
breaks.vec7 = seq(from = x_0, by = h4, length.out = 9)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec7,
     xlab = 'x0 = 137.65', ylab = '', main = '')
x_0 = 137.75
breaks.vec8 = seq(from = x_0, by = h4, length.out = 9)
hist(banknote$Diagonal[banknote$Status == 'counterfeit'], breaks = breaks.vec8,
     xlab = 'x0 = 137.75', ylab = '', main = '')
```



Histograms (直方图)

- 直方图估计的主要困难
  - 区间长度 (bin-width) h 的确定，它决定了直方图的形状. $h_{\text{opt}} = \left(\frac{24\sqrt{\pi}}{n}\right)^{1/3}$
  - 区间起始点 x_0 的选择，它在某种程度上也影响直方图的形状.
  - 因观测数据被其所在区间的中点代替而导致的信息损失.
  - 通常情况下密度函数是光滑的，但直方图不光滑.

  - 事实上，直方图还可以表示成

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{I}\left(\left|x - x_i\right| \le \frac{h}{2}\right).$$

Kernel Densities (核密度)

- 核密度估计量的一般形式

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

- 常用核函数 $K(u)$:

$$\text{Uniform}: \quad K(u) = \frac{1}{2}\mathrm{I}\left(|u| \le 1\right)$$

$$\text{Triangle}: \quad K(u) = \left(1 - |u|\right)\mathrm{I}\left(|u| \le 1\right)$$

$$\text{Espanechnikov}: \quad K(u) = \frac{3}{4}\left(1 - u^2\right)\mathrm{I}\left(|u| \le 1\right)$$

$$\text{Quartic(Biweight)}: \quad K(u) = \frac{15}{16}\left(1 - u^2\right)^2\mathrm{I}\left(|u| \le 1\right)$$

$$\text{Gaussian}: \quad K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right) = \varphi(u)$$

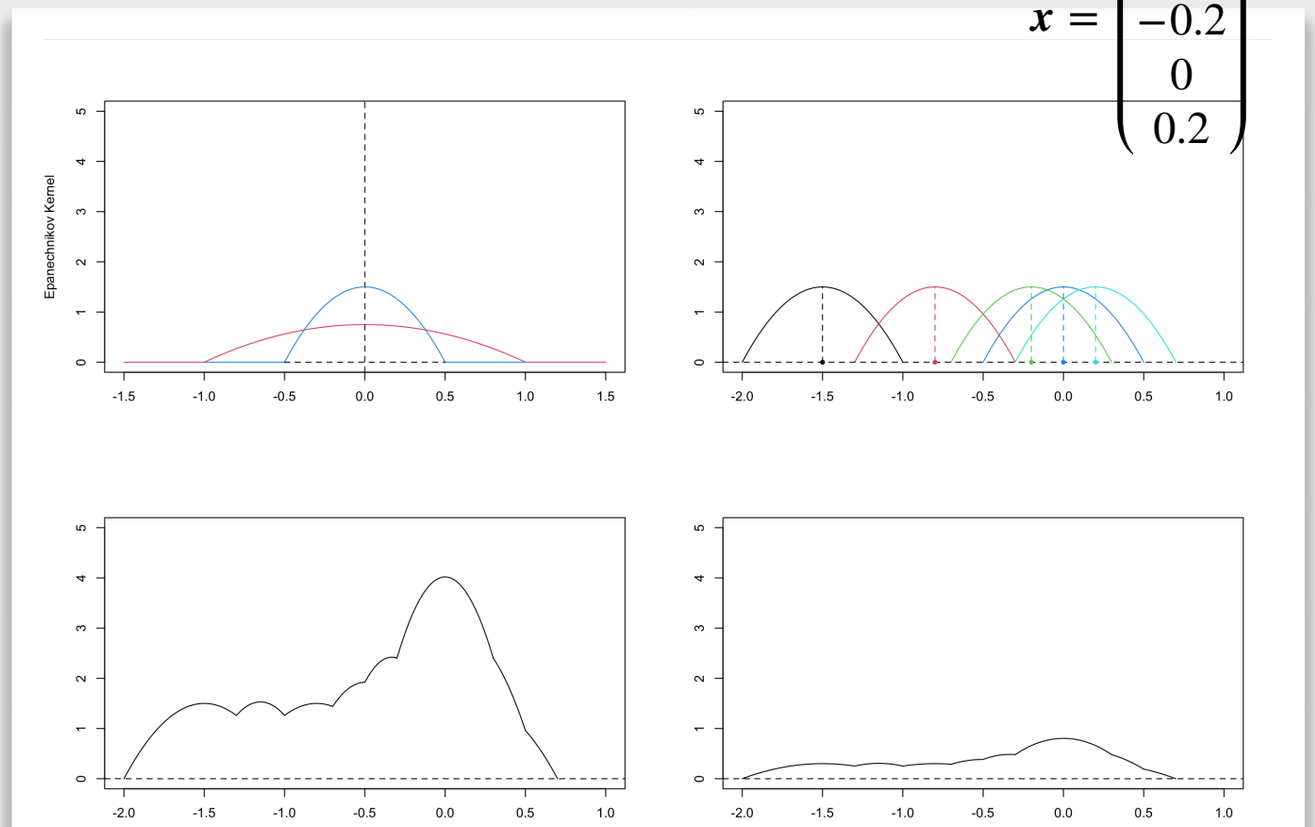
Kernel Densities (核密度)

```
par(mfrow=c(2, 2))
s = function(x) (3/4) * (1-x^2)
h = 1/2
g = function(x) (1/h) * s(x/h)
curve(s, -1, 1, xlim=c(-1.5, 1.5), ylim=c(0, 5), xlab="", ylab="Epanechnikov Kernel", col=2)
lines(c(-1.5, -1), c(0, 0), col=2)
lines(c(1, 1.5), c(0, 0), col=2)
abline(v=0, lty=2)
curve(g, -h, h, col=4, add=TRUE)
lines(c(-1, -0.5), c(0, 0), col=4)
lines(c(0.5, 1), c(0, 0), col=4)
lines(c(-0.5, 0.5), c(0, 0), lty=2)

x = c(-1.5, -0.8, -0.2, 0, 0.2)
f = function(t) (3/4) * (1-t^2)
h = 1/2
g1 = function(t) (1/h) * f((t-x[1])/h)
g2 = function(t) (1/h) * f((t-x[2])/h)
g3 = function(t) (1/h) * f((t-x[3])/h)
g4 = function(t) (1/h) * f((t-x[4])/h)
g5 = function(t) (1/h) * f((t-x[5])/h)
curve(g1, x[1]-h, x[1]+h, xlim=c(-2, 1), ylim=c(0, 5), xlab="", ylab="", lty=1)
curve(g2, x[2]-h, x[2]+h, add=TRUE, col=2, lty=1)
curve(g3, x[3]-h, x[3]+h, add=TRUE, col=3, lty=1)
curve(g4, x[4]-h, x[4]+h, add=TRUE, col=4, lty=1)
curve(g5, x[5]-h, x[5]+h, add=TRUE, col=5, lty=1)
abline(h=0, lty=2)
for (i in 1:5) lines(c(x[i], x[i]), c(0, 1.5), lty=2, col=i)
for (i in 1:5) points(x[i], 0, cex=0.8, pch=16, col=i)

h1 = function(t) g1(t)
curve(h1, -2, -1.3, xlim=c(-2, 1), ylim=c(0, 5), xlab="", ylab="")
h2 = function(t) (g1(t) + g2(t))
curve(h2, -1.3, -1, add=TRUE)
h3 = function(t) g2(t)
curve(h3, -1, -0.7, add=TRUE)
h4 = function(t) (g2(t) + g3(t))
curve(h4, -0.7, -0.5, add=TRUE)
h5 = function(t) (g2(t) + g3(t) + g4(t))
curve(h5, -0.5, -0.3, add=TRUE)
h6 = function(t) (g3(t) + g4(t) + g5(t))
curve(h6, -0.3, 0.3, add=TRUE)
h7 = function(t) (g4(t) + g5(t))
curve(h7, 0.3, 0.5, add=TRUE)
h8 = function(t) g5(t)
curve(h8, 0.5, 0.7, add=TRUE)
abline(h=0, lty=2)

f1 = function(t) (1/5) * g1(t)
curve(f1, -2, -1.3, xlim=c(-2, 1), ylim=c(0, 5), xlab="", ylab="")
f2 = function(t) (1/5) * (g1(t) + g2(t))
curve(f2, -1.3, -1, add=TRUE)
f3 = function(t) (1/5) * g2(t)
curve(f3, -1, -0.7, add=TRUE)
f4 = function(t) (1/5) * (g2(t) + g3(t))
curve(f4, -0.7, -0.5, add=TRUE)
f5 = function(t) (1/5) * (g2(t) + g3(t) + g4(t))
curve(f5, -0.5, -0.3, add=TRUE)
f6 = function(t) (1/5) * (g3(t) + g4(t) + g5(t))
curve(f6, -0.3, 0.3, add=TRUE)
f7 = function(t) (1/5) * (g4(t) + g5(t))
curve(f7, 0.3, 0.5, add=TRUE)
f8 = function(t) (1/5) * g5(t)
curve(f8, 0.5, 0.7, add=TRUE)
abline(h=0, lty=2)
```

$$K\left(\frac{x - x_i}{h}\right)$$

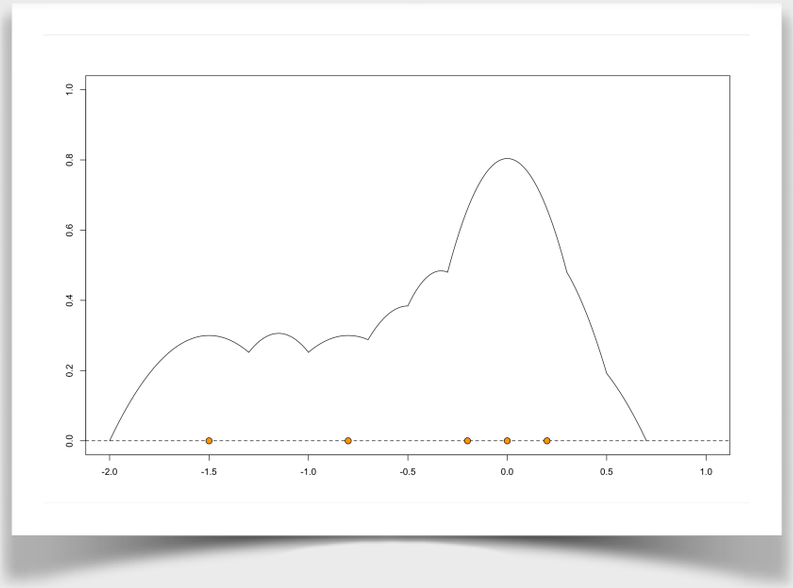
$$K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}(|u| \leq 1)$$

$$\boldsymbol{x} = \begin{pmatrix} -1.5 \\ -0.8 \\ -0.2 \\ 0 \\ 0.2 \end{pmatrix}$$



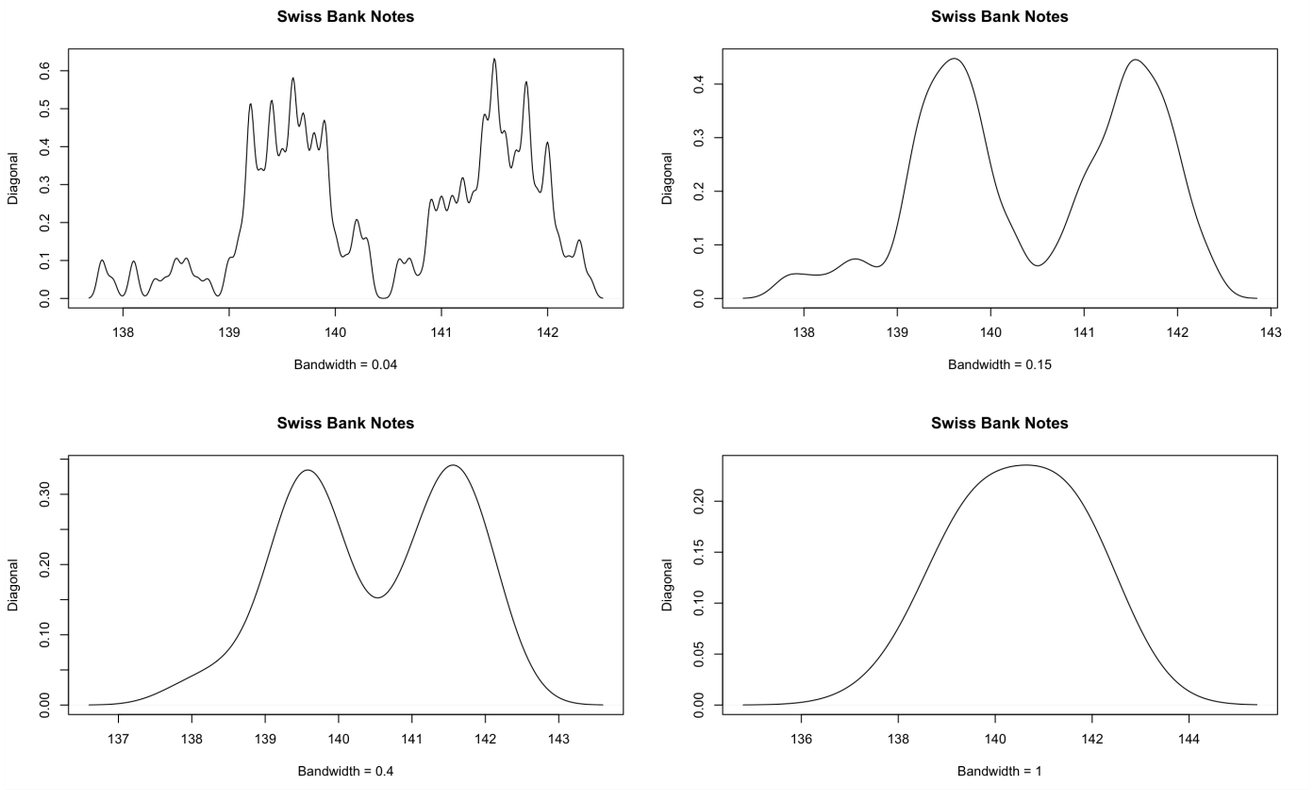
Kernel Densities

```
par(mfrow=c(1, 1))
f1 = function(t) (1/5) * g1(t)
curve(f1, -2, -1.3, xlim=c(-2, 1), ylim=c(0, 1), xlab="", ylab="")
f2 = function(t) (1/5) * (g1(t) + g2(t))
curve(f2, -1.3, -1, add=TRUE)
f3 = function(t) (1/5) * g2(t)
curve(f3, -1, -0.7, add=TRUE)
f4 = function(t) (1/5) * (g2(t) + g3(t))
curve(f4, -0.7, -0.5, add=TRUE)
f5 = function(t) (1/5) * (g2(t) + g3(t) + g4(t))
curve(f5, -0.5, -0.3, add=TRUE)
f6 = function(t) (1/5) * (g3(t) + g4(t) + g5(t))
curve(f6, -0.3, 0.3, add=TRUE)
f7 = function(t) (1/5) * (g4(t) + g5(t))
curve(f7, 0.3, 0.5, add=TRUE)
f8 = function(t) (1/5) * g5(t)
curve(f8, 0.5, 0.7, add=TRUE)
abline(h=0, lty=2)
for (i in 1:5) points(x[i], 0, cex=1.5, pch=21, bg='orange')
```



Example: Swiss bank notes

```
par(mfrow=c(2, 2))
plot(density(banknote$Diagonal, bw = 0.04), xlab='Bandwidth = 0.04', main='Swiss Bank Notes',
     ylab='Diagonal')
plot(density(banknote$Diagonal, bw = 0.15), xlab='Bandwidth = 0.15', main='Swiss Bank Notes',
     ylab='Diagonal')
plot(density(banknote$Diagonal, bw = 0.4), xlab='Bandwidth = 0.4', main='Swiss Bank Notes',
     ylab='Diagonal')
plot(density(banknote$Diagonal, bw = 1), xlab='Bandwidth = 1', main='Swiss Bank Notes',
     ylab='Diagonal')
```



Choice of bandwidth h (带宽的选择)

- 经验法则:
  - 对正态核函数

$$h_G = \frac{1.06\ \hat{\sigma}}{\sqrt[5]{n}}$$

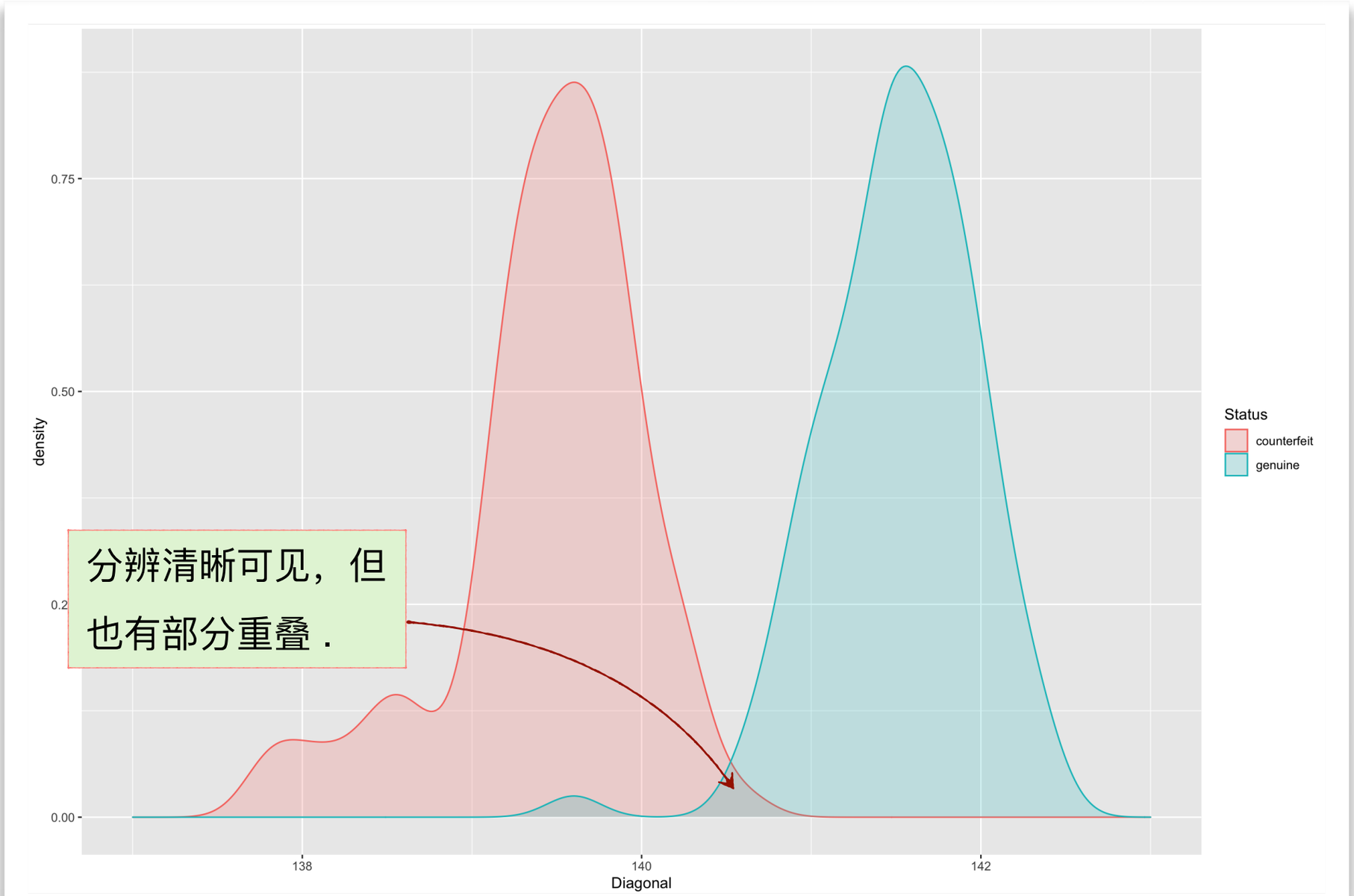
$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

  - 对二次核函数

$$h_Q = 2.62 \cdot h_G = \frac{2.7772 \cdot \hat{\sigma}}{\sqrt[5]{n}}$$

Example: Swiss bank notes (瑞银钞票)

```
library(ggplot2)
ggplot(banknote, aes(Diagonal, fill = Status, color = Status)) +
  geom_density(alpha = 0.2) +
  xlim(137, 143)
```

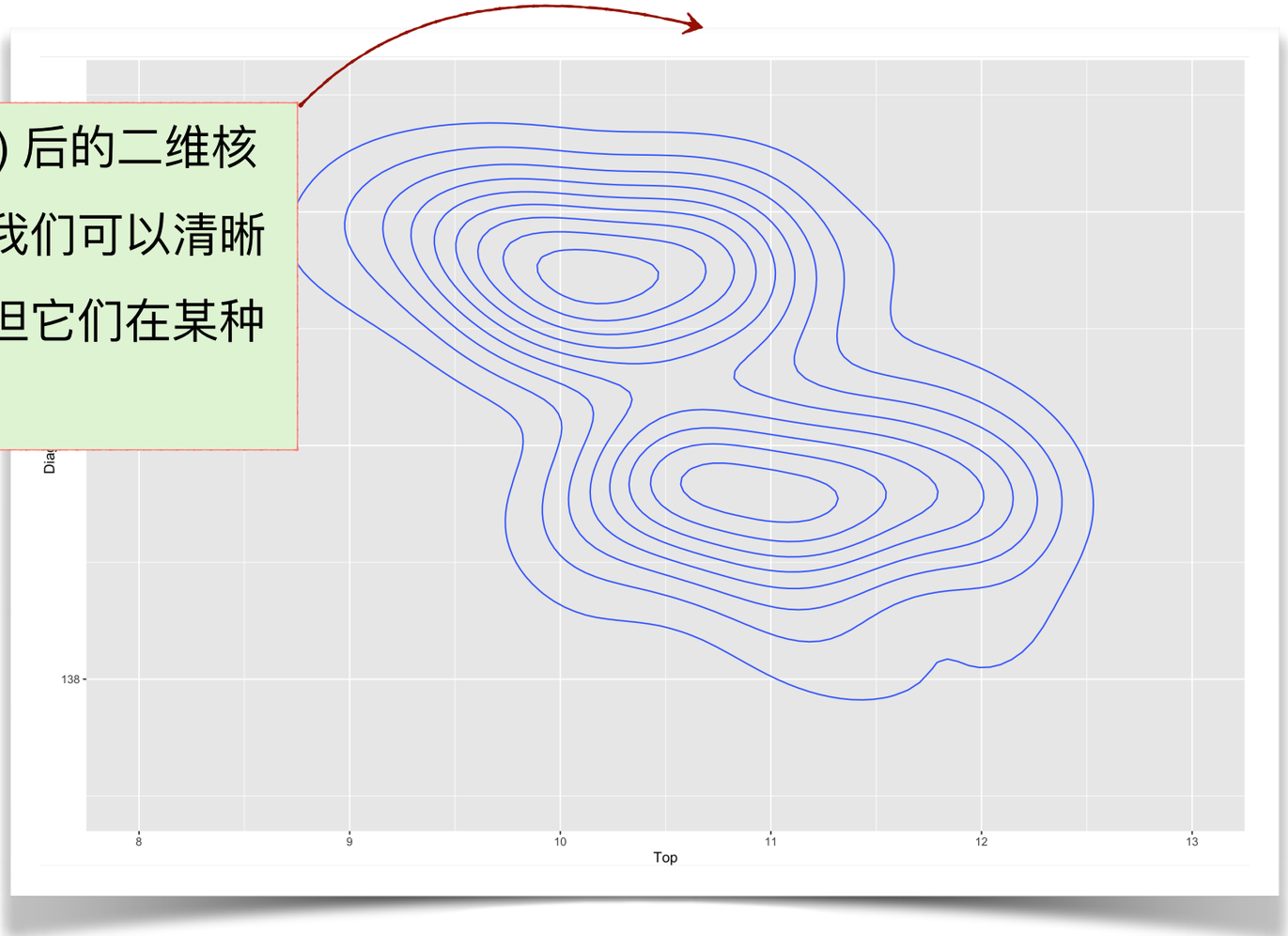


Example: Swiss bank notes (瑞银钞票)

- 问题：除使用对角线数据外，能否再利用其它一或两个变量以更好分辨真钞、伪钞?

```
ggplot(banknote, aes(x = Top, y = Diagonal)) +
  xlim(8, 13)  +
  ylim(137, 143) +
  geom_density_2d(na.rm = TRUE)
```

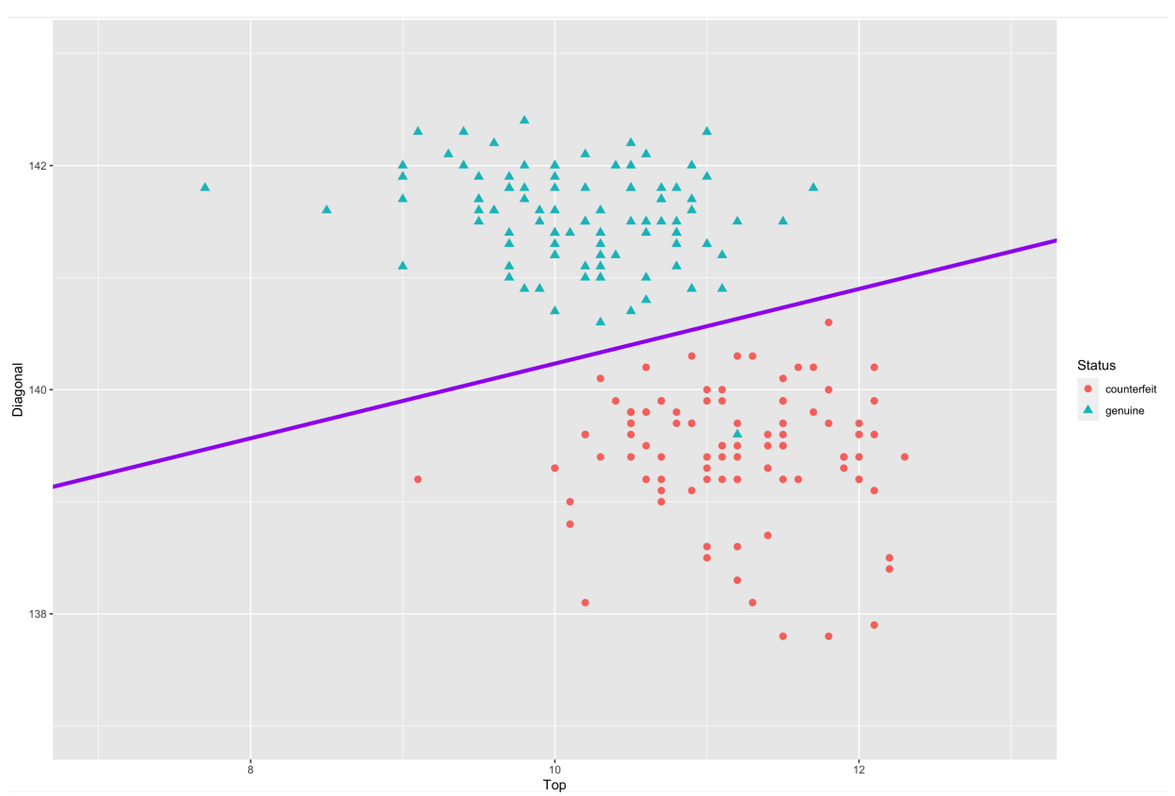
这是添加一个变量 (Top) 后的二维核密度估计的等高线图，我们可以清晰看到两个不同的分布，但它们在某种程度上仍然有重叠.



Scatterplots (散点图)

- 散点图是变量之间的二维或三维图形. 它有助于我们理解一个数据集当中各个变量之间的相互关系.

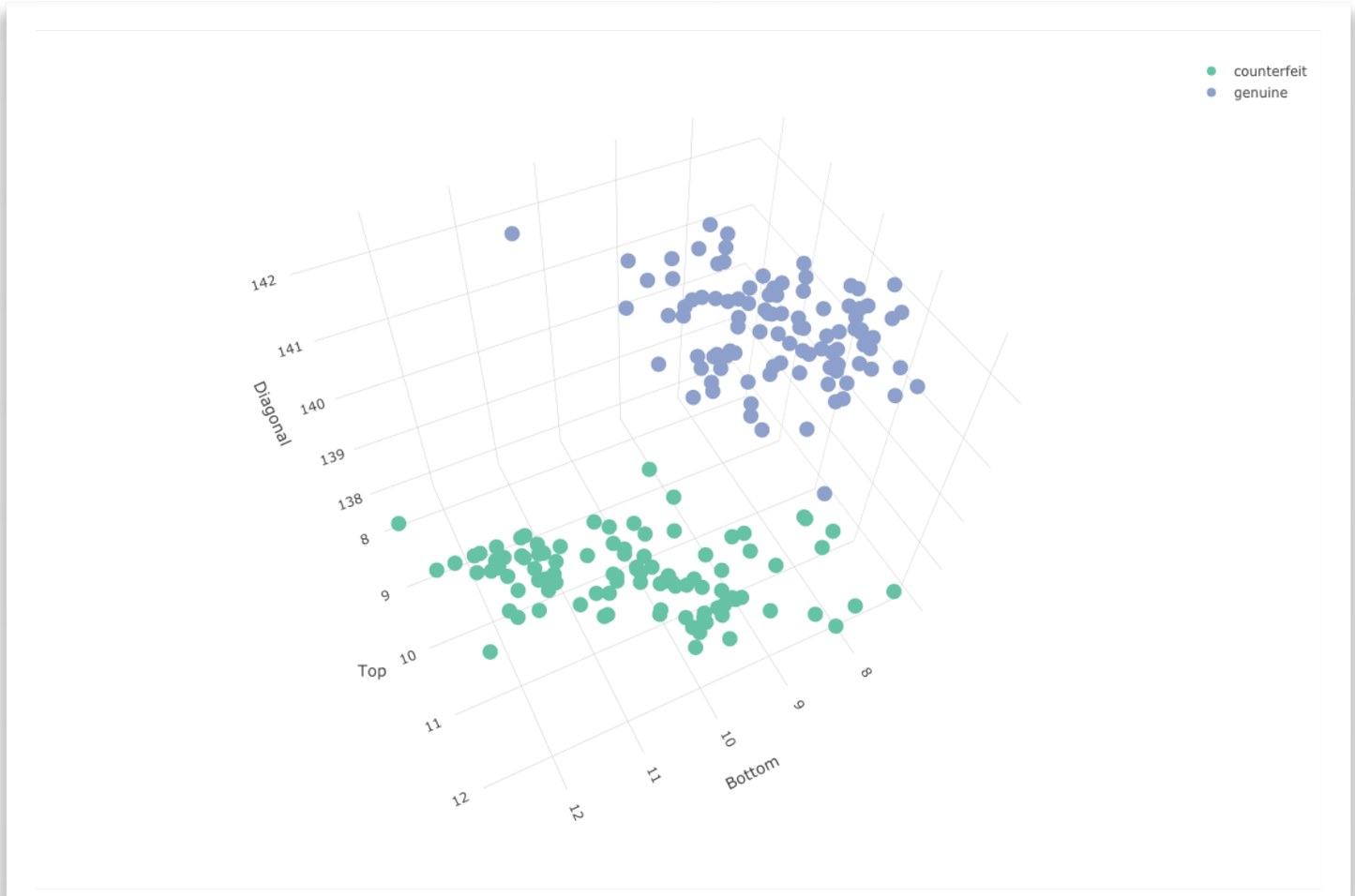
```
ggplot(banknote, aes(x = Top,  y = Diagonal, colour = Status, shape = Status)) +
  xlim(7, 13)  +
  ylim(137, 143) +
  geom_point(size = 2.5) +
  geom_abline(intercept = 136.9, slope = 1/3, colour = 'purple',size = 1.5)
```



Scatterplots (散点图)

- 散点图是变量之间的二维或三维图形. 它有助于我们理解一个数据集当中各个变量之间的相互关系.

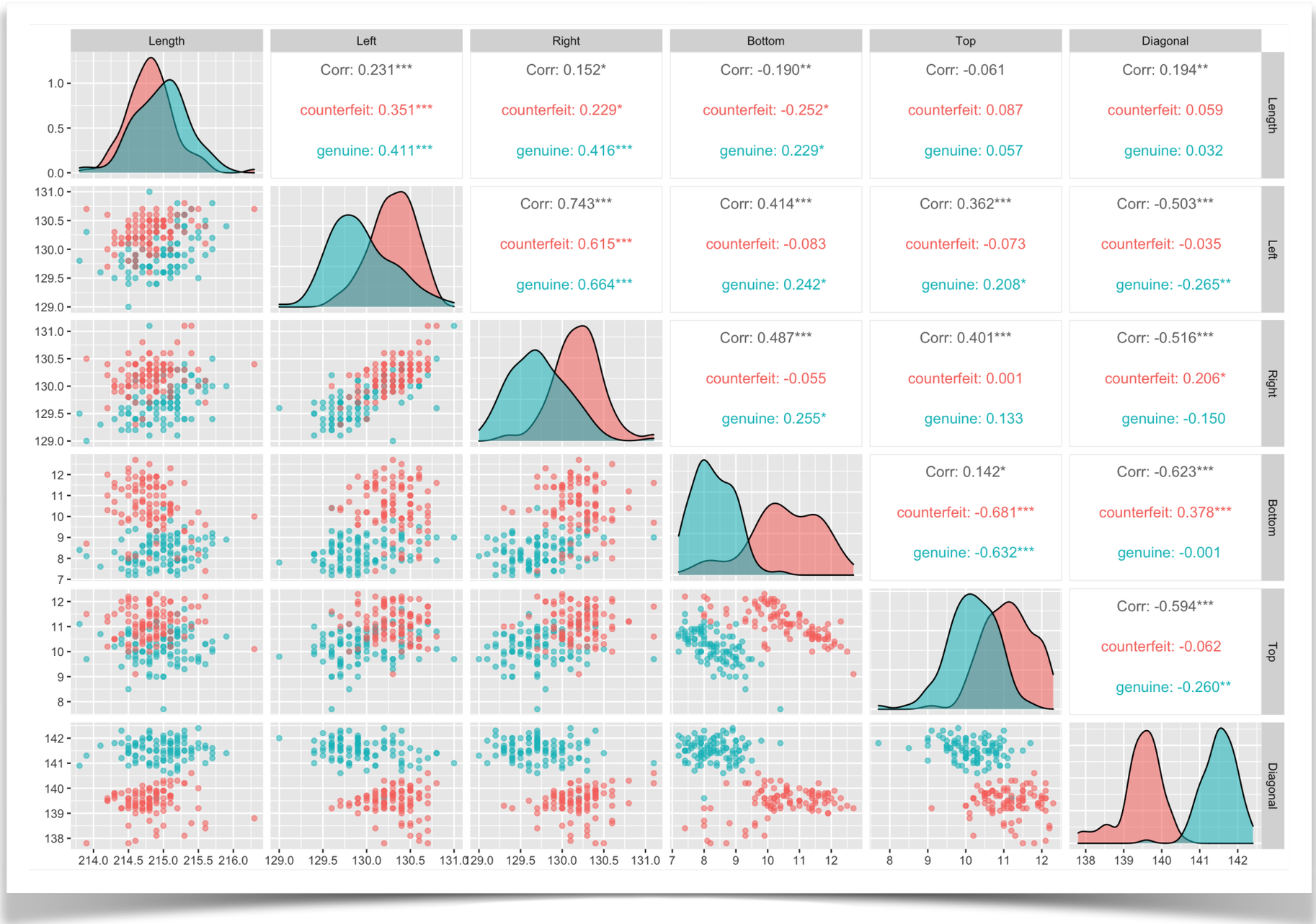
```
library(plotly)
plot_ly(banknote, x =~ Bottom, y =~ Top, z =~ Diagonal) %>%
  add_markers(color =~ Status, symbol =~ Status)
```



Scatterplots (散点图)

- 散点图矩阵 (scatterplot matrix)：所有变量两两之间二维散点图的矩阵.

```
library(GGally)
ggpairs(banknote, columns = 2:7, ggplot2::aes(colour=Status, alpha = 0.2))
```



Chernoff-Flury Faces (脸谱图)

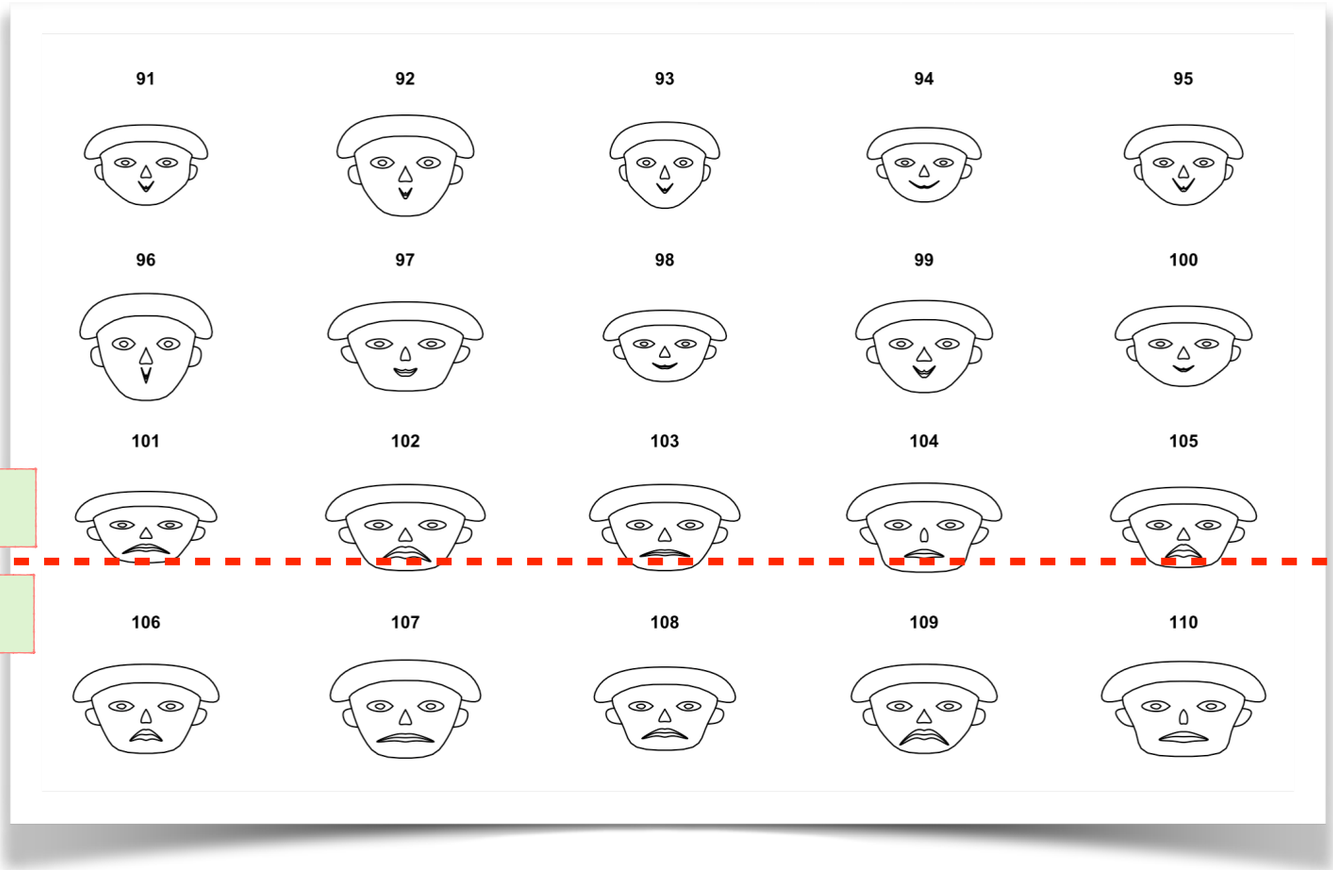
- 遵循 Flury 和 Riedwyl (1988) 描述的设计思路，用到的特征如下：

1	右眼尺寸	2	右眼瞳孔大小
3	右眼瞳孔的位置	4	右眼的倾斜程度
5	右眼的水平位置	6	右眼的垂直位置
7	右眉弯曲程度	8	右眉的浓密程度
9	右眉的水平位置	10	右眉的垂直位置
11	右上发线	12	右下发线
13	脸右侧线	14	右侧头发的黑暗程度
15	右侧头发的倾斜程度	16	鼻子右侧线
17	嘴右侧大小	18	嘴右侧的倾斜程度
19 ~ 36	与 1 ~ 18 相同，关于左侧		

  - 首先，对应于某个面部特征的变量，作变换使其取值位于 $(0, 1)$ 区间内，最小值对应 0、最大值对应 1.
  - 因此，面部特征变量的极端值对应的是 “高兴” 或者 “开心” 的表情.
  - 深色头发可编码为 1，而金发则可编码为 0，依此类推.

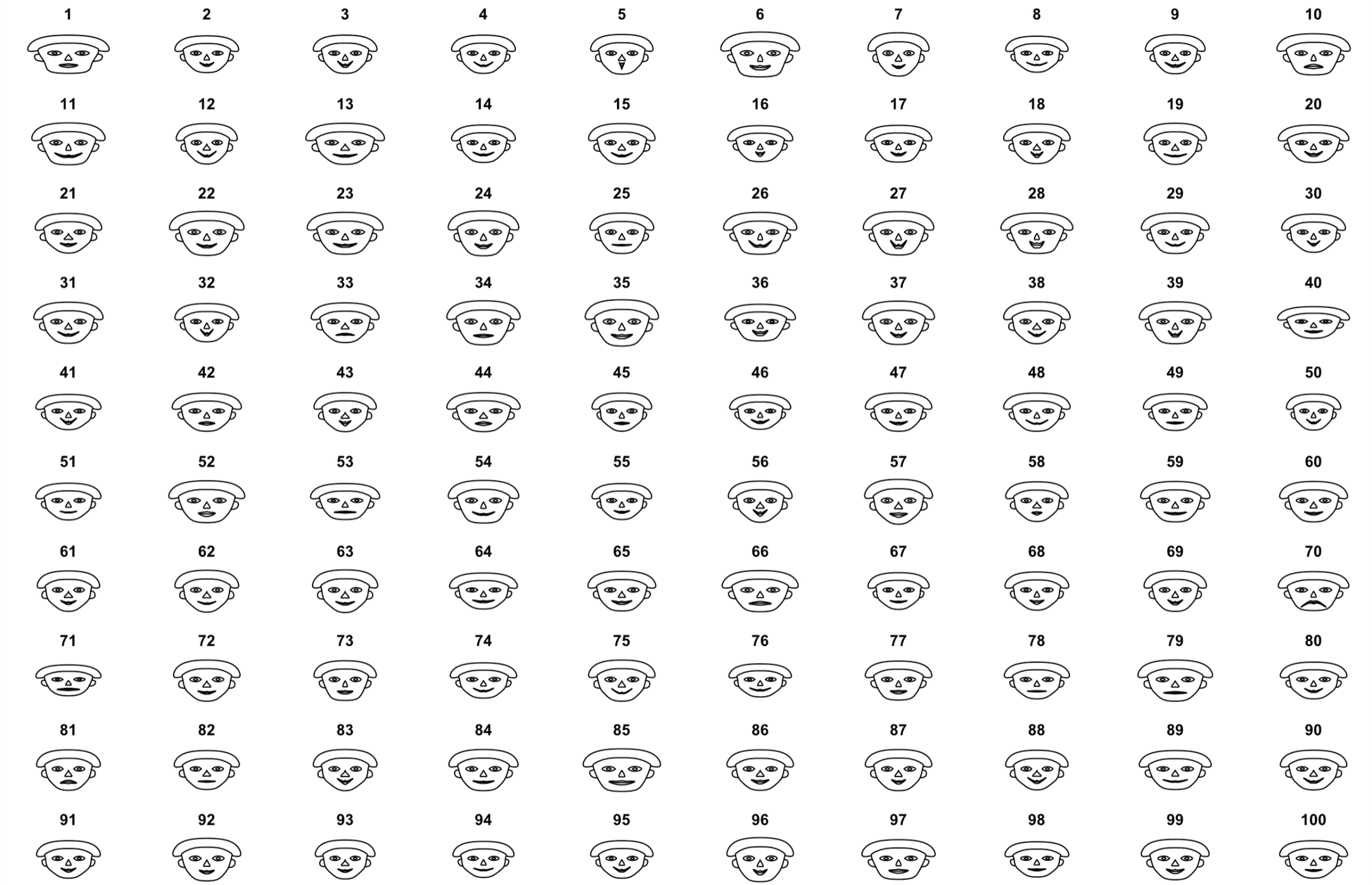
Chernoff-Flury Faces (脸谱图)

```
library(TeachingDemos)
faces(banknote[91:110, 2:7], fill = TRUE)
```



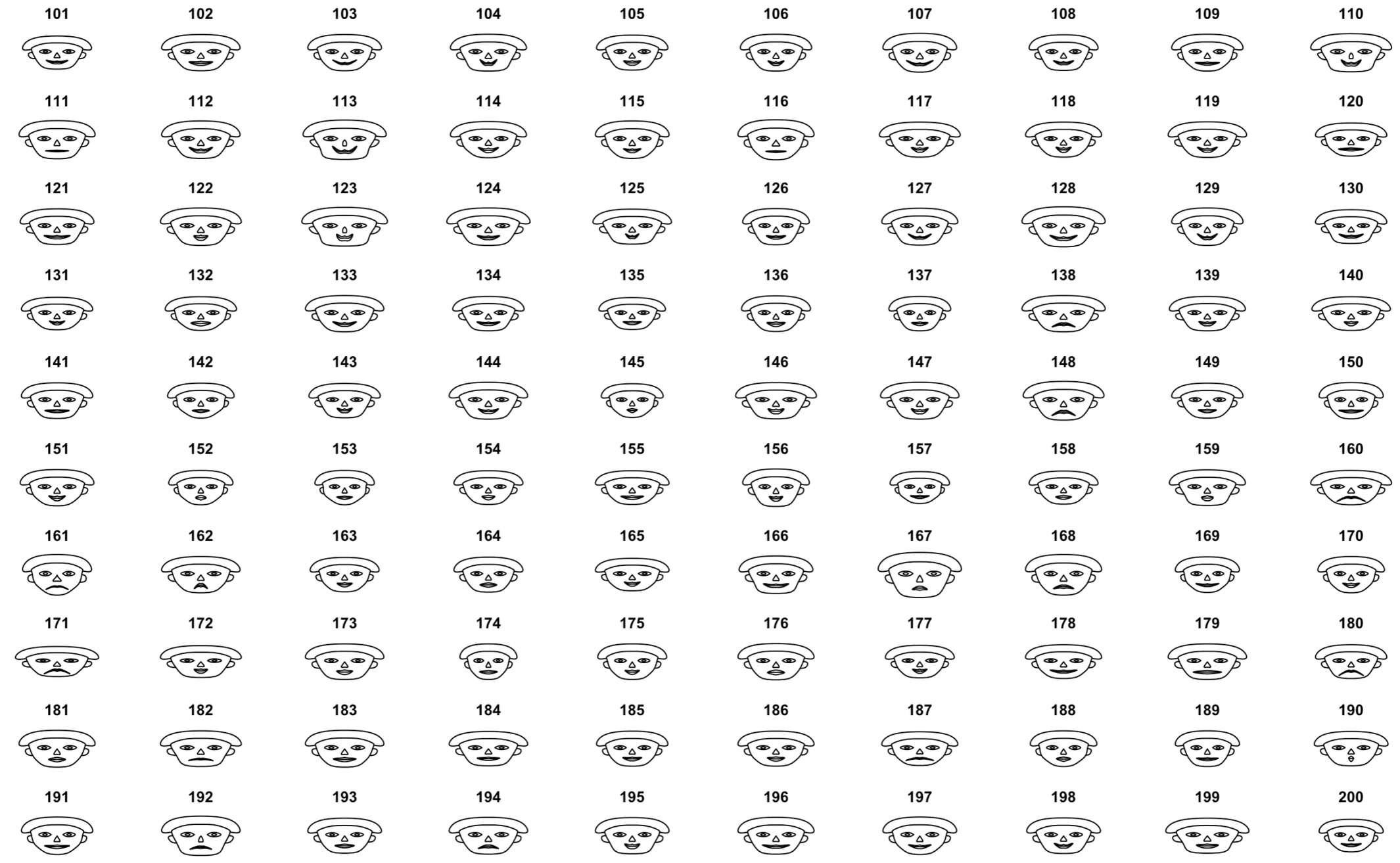
Chernoff-Flury Faces (脸谱图)

```
faces(banknote[1:100, 2:7], ncol=10, fill = TRUE) # 真钞
```



Chernoff-Flury Faces (脸谱图)

```
faces(banknote[101:200, 2:7], ncol=10, fill = TRUE) # 假钞
```



Chernoff-Flury Faces (脸谱图)

- The function `faces()` in package `aplpack`.

Failed to install into my computer.

Andrews' Curves (安德鲁斯曲线)

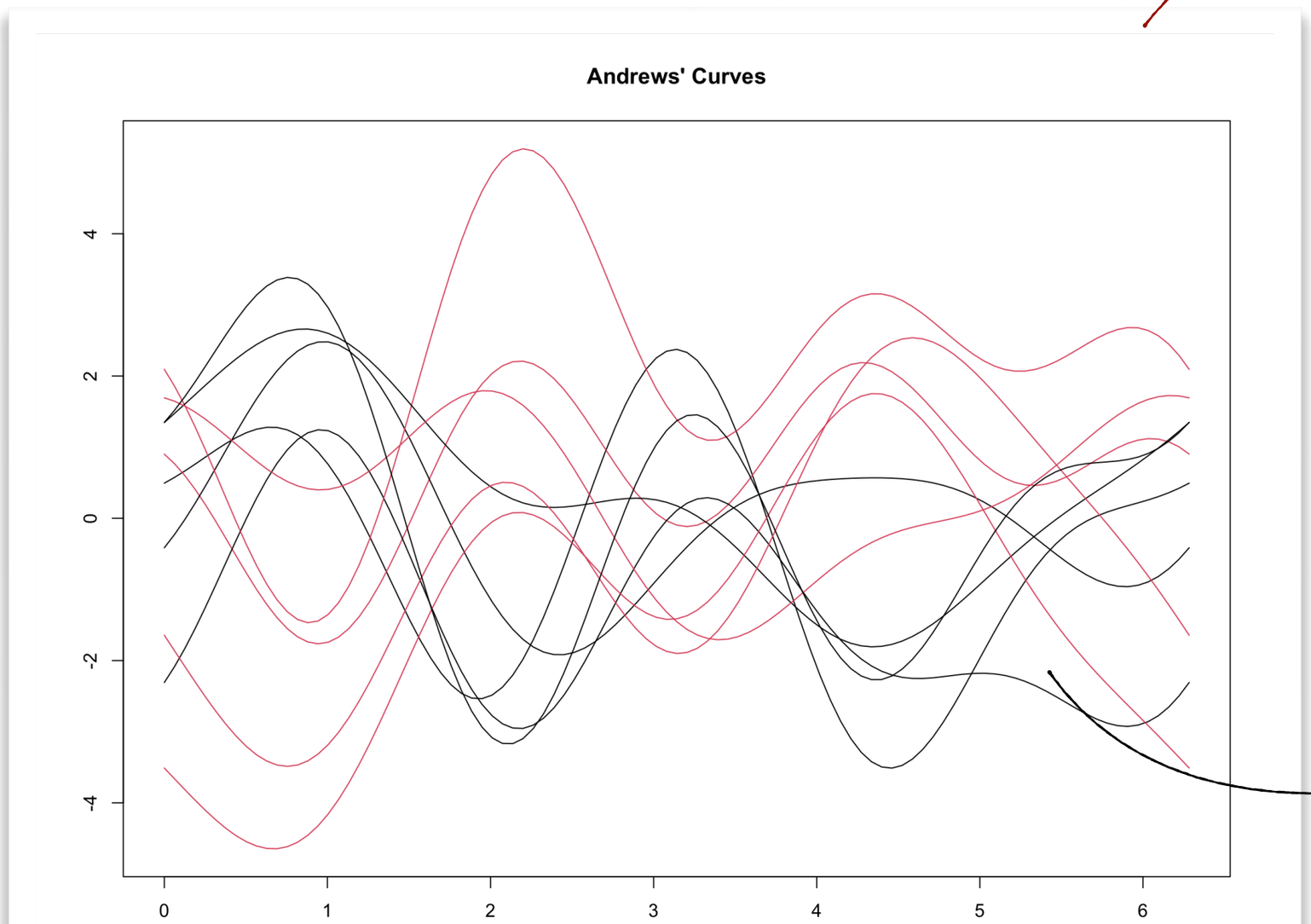
- 将多元变量的每一个观测值 $\boldsymbol{x}_i = \left(x_{i1},\ x_{i2},\ \ldots,\ x_{ip}\right)$ 转换为如下所示曲线

$$f_i(t) = \begin{cases} \dfrac{1}{\sqrt{2}}x_{i1} + x_{i2}\sin(t) + x_{i3}\cos(t) + \cdots + x_{i,\,p-1}\sin\left(\dfrac{p-1}{2}t\right) + x_{ip}\cos\left(\dfrac{p-1}{2}t\right), & p \text{ 为奇数} \\ \dfrac{1}{\sqrt{2}}x_{i1} + x_{i2}\sin(t) + x_{i3}\cos(t) + \cdots + x_{ip}\sin\left(\dfrac{p}{2}t\right), & p \text{ 为偶数} \end{cases}$$

观测值表示的是所谓傅立叶级数的系数 ($t \in [-\pi,\ \pi]$).

Andrews' Curves (安德鲁斯曲线)

```
library(pracma)
A1 = as.matrix(banknote[96:105, 2:7])
andrewsplot(scale(A1), f = banknote$Status[96:105], style = 'cart')
```

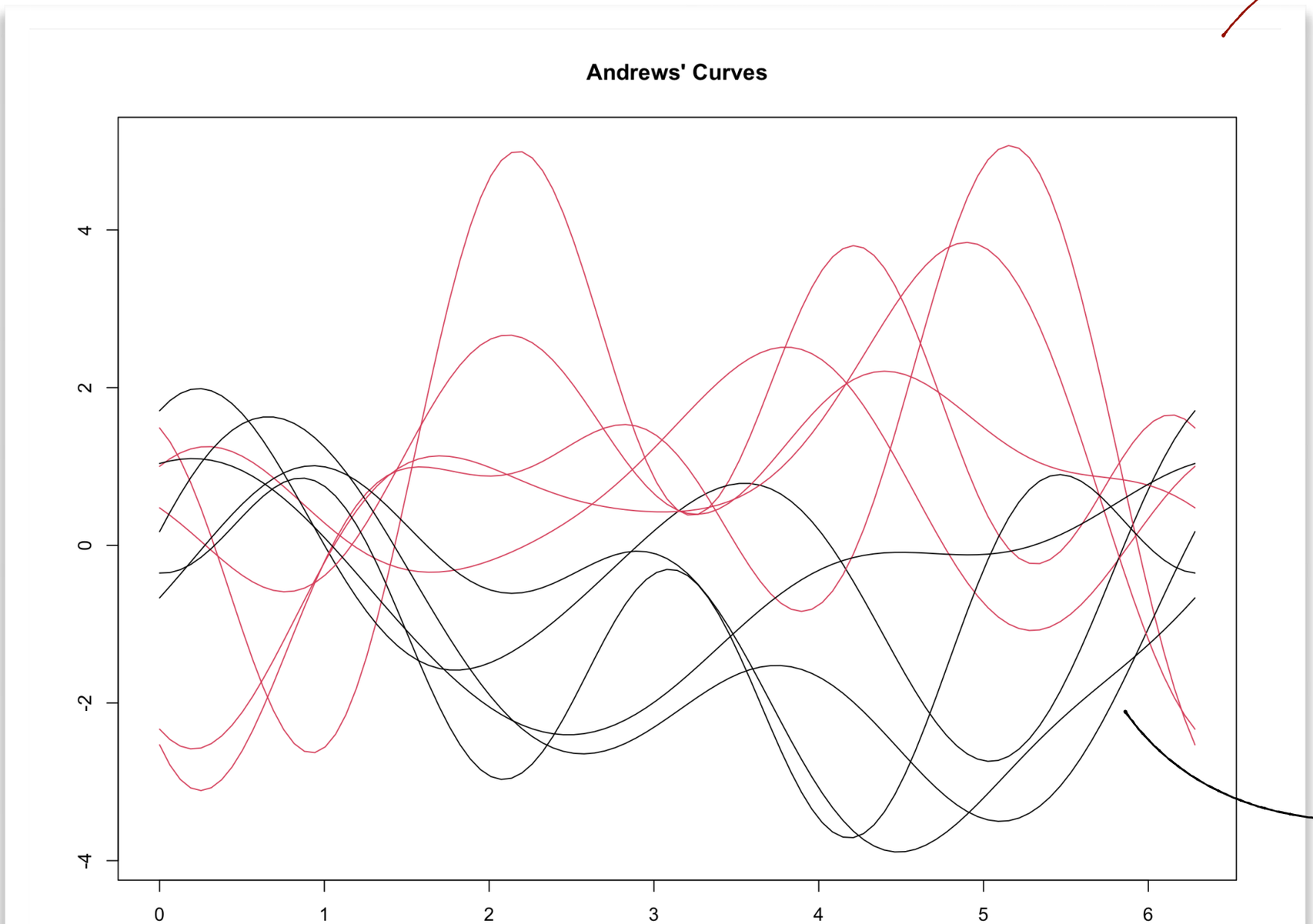


真钞

假钞

Andrews' Curves (安德鲁斯曲线)

```
A2 = as.matrix(banknote[96:105, 7:2]) # 变量次序不同
andrewsplot(scale(A2), f = banknote$Status[96:105], style = 'cart')
```

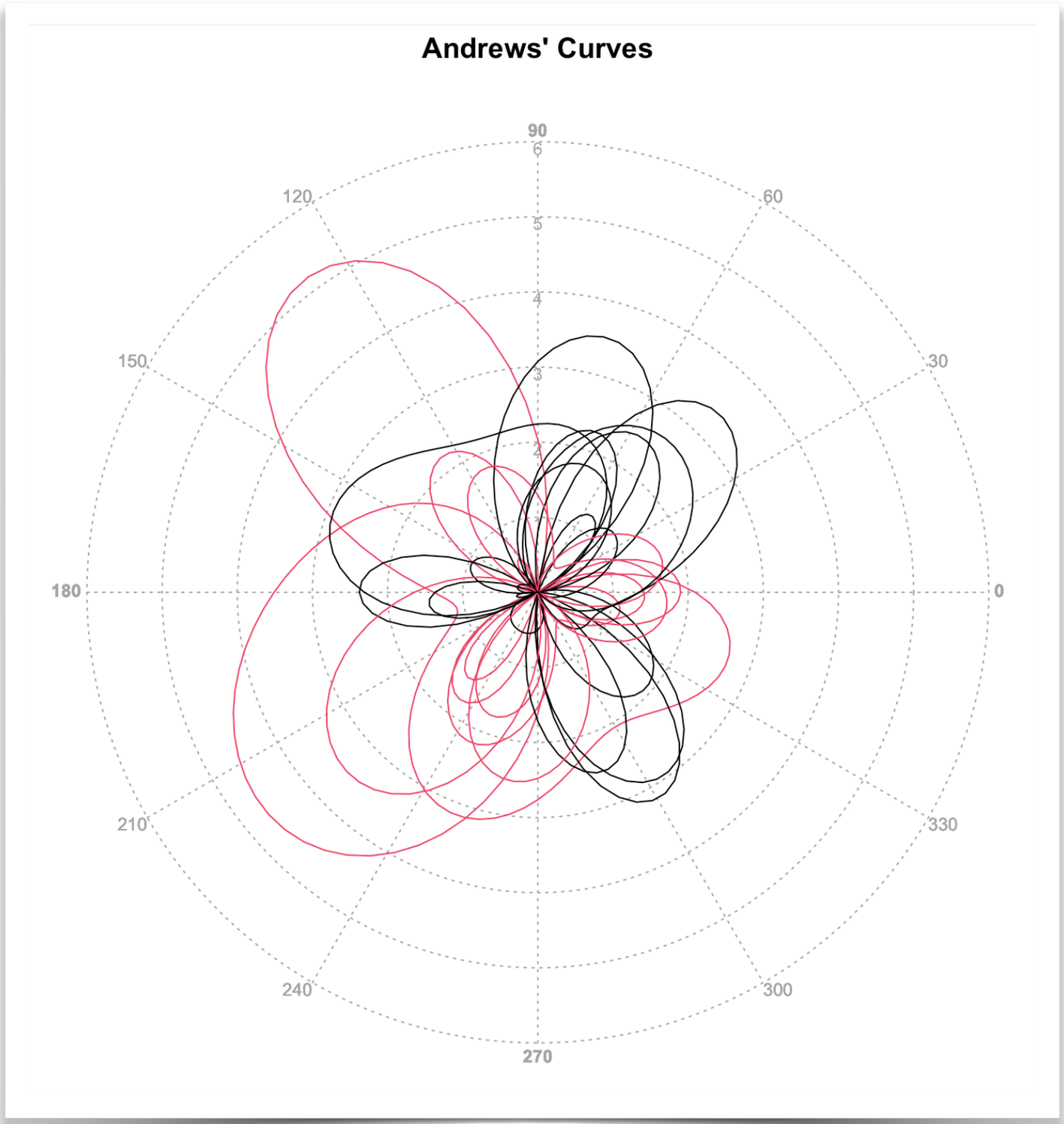


真钞

假钞

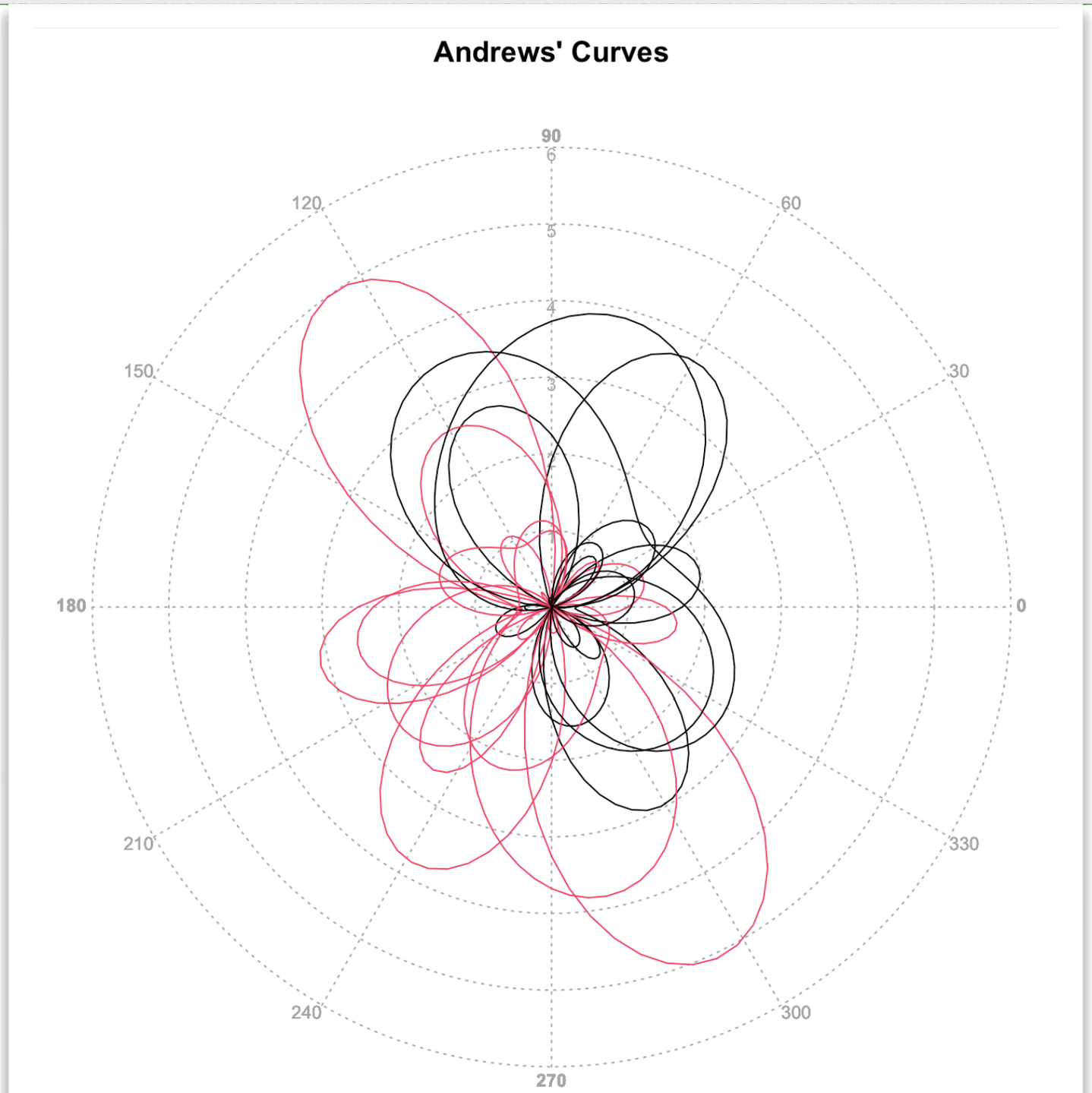
Andrews' Curves (安德鲁斯曲线)

```
andrewsplot(scale(A1), f = banknote$Status[96:105], style = 'pol')
```



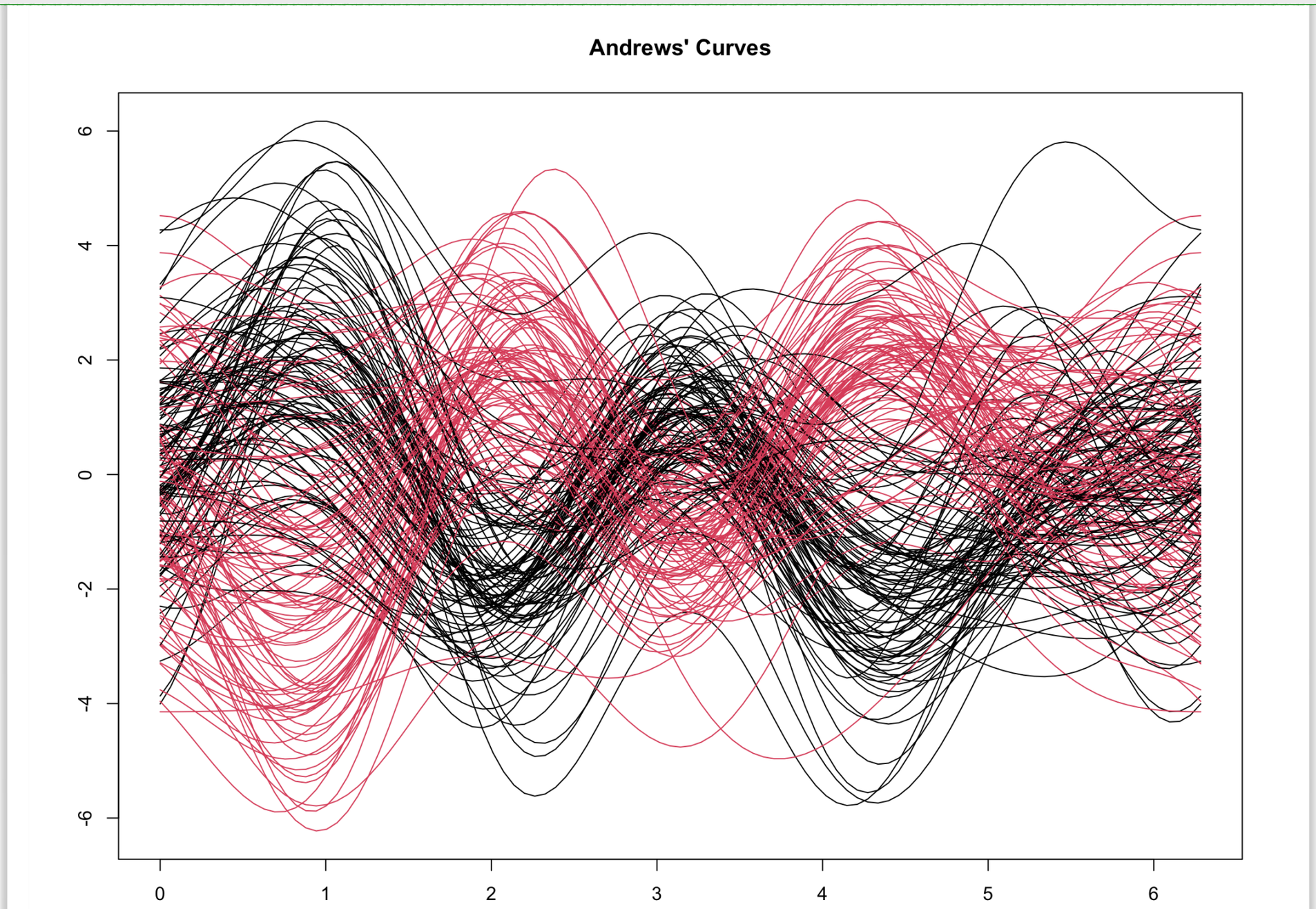
Andrews' Curves (安德鲁斯曲线)

```
andrewsplot(scale(A2), f = banknote$Status[96:105], style = 'pol')
```



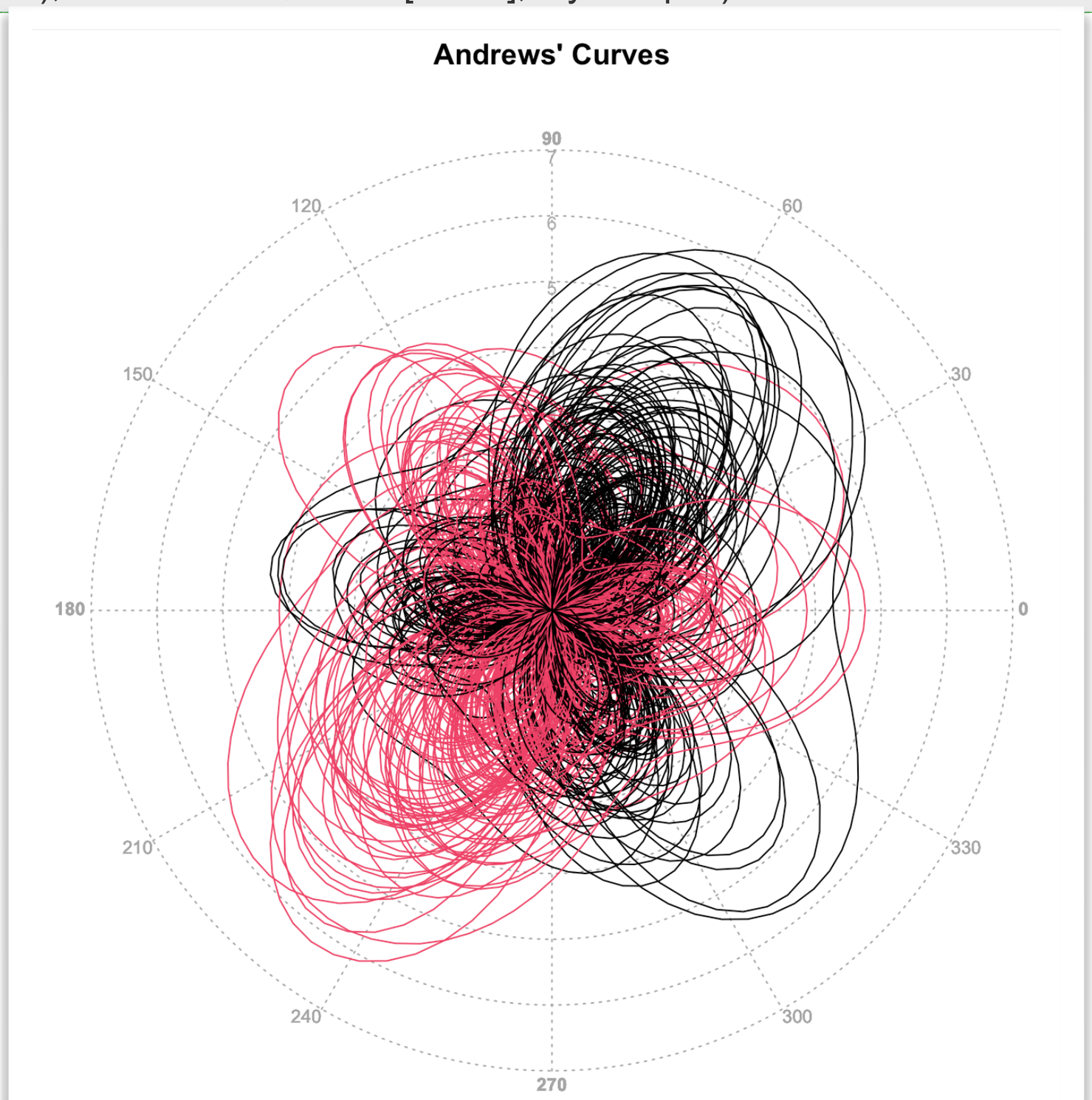
Andrews' Curves (安德鲁斯曲线)

```
A3 = as.matrix(banknote[1:200, 2:7])
andrewsplot(scale(A3), f = banknote$Status[1:200], style = 'cart')
```



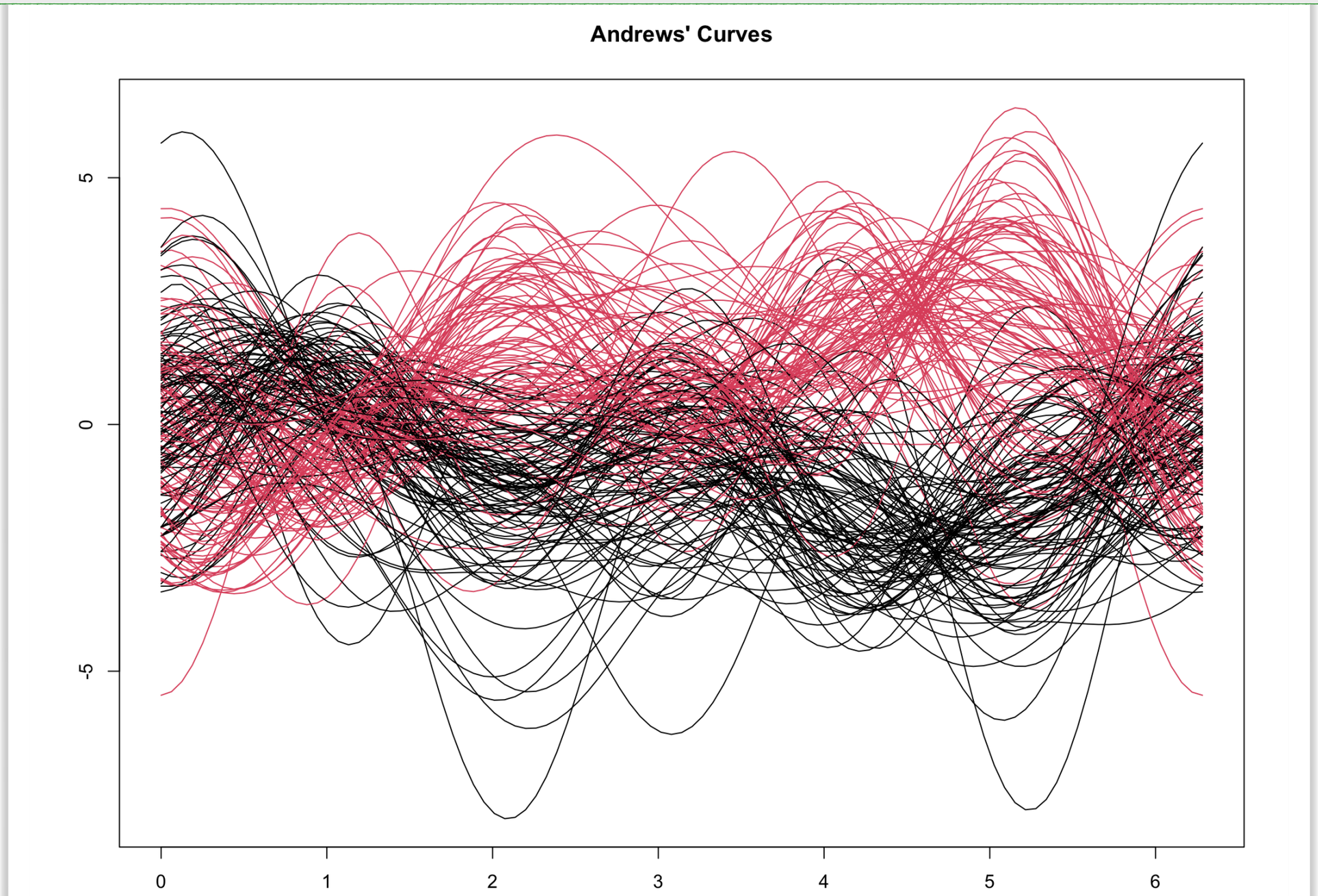
Andrews' Curves (安德鲁斯曲线)

```
andrewsplot(scale(A3), f = banknote$Status[1:200], style = 'pol')
```



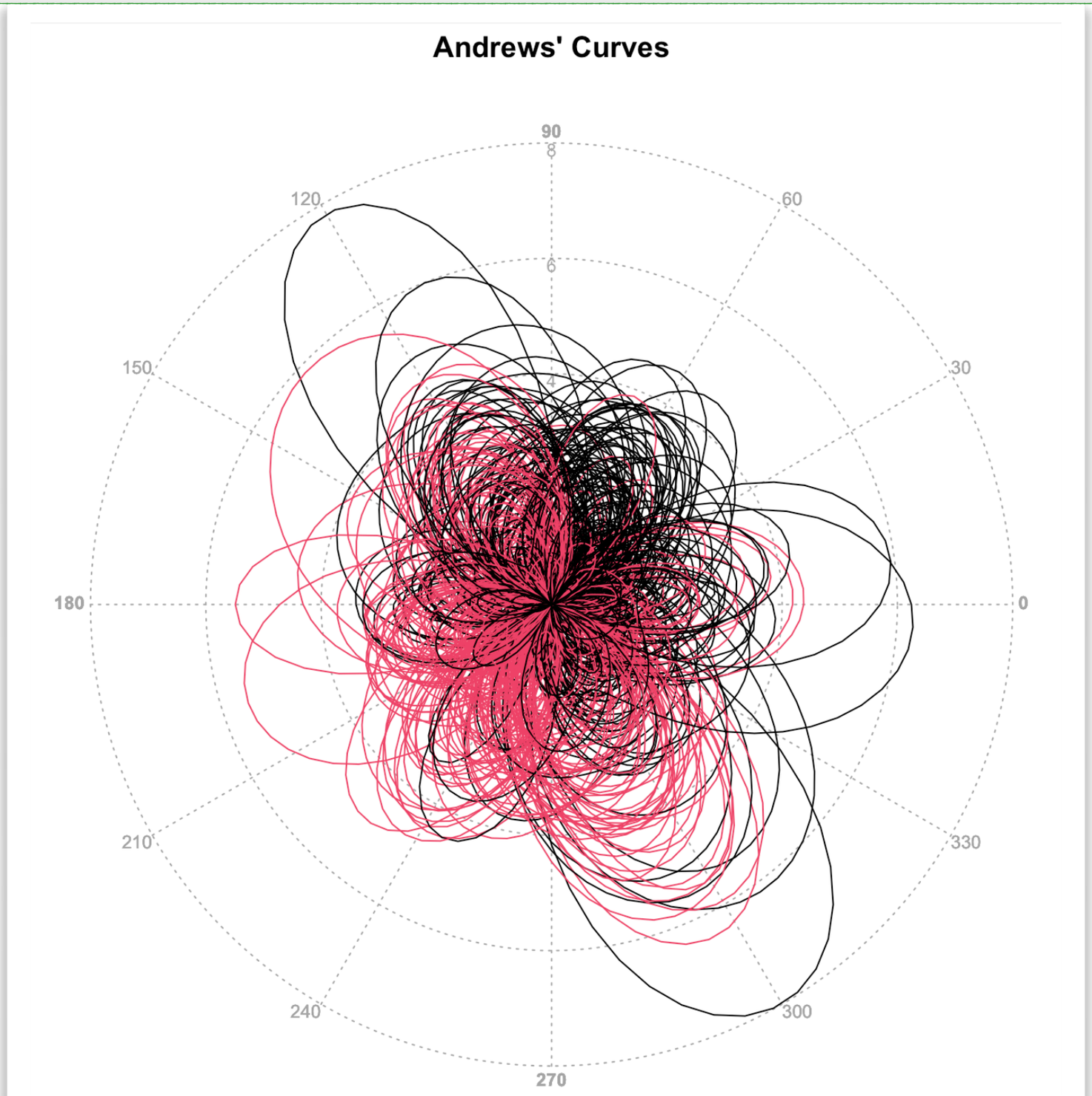
Andrews' Curves (安德鲁斯曲线)

```
A4 = as.matrix(banknote[1:200, 7:2])
andrewsplot(scale(A4), f = banknote$Status[1:200], style = 'cart')
```



Andrews' Curves (安德鲁斯曲线)

```
andrewsplot(scale(A4), f = banknote$Status[1:200], style = 'pol')
```

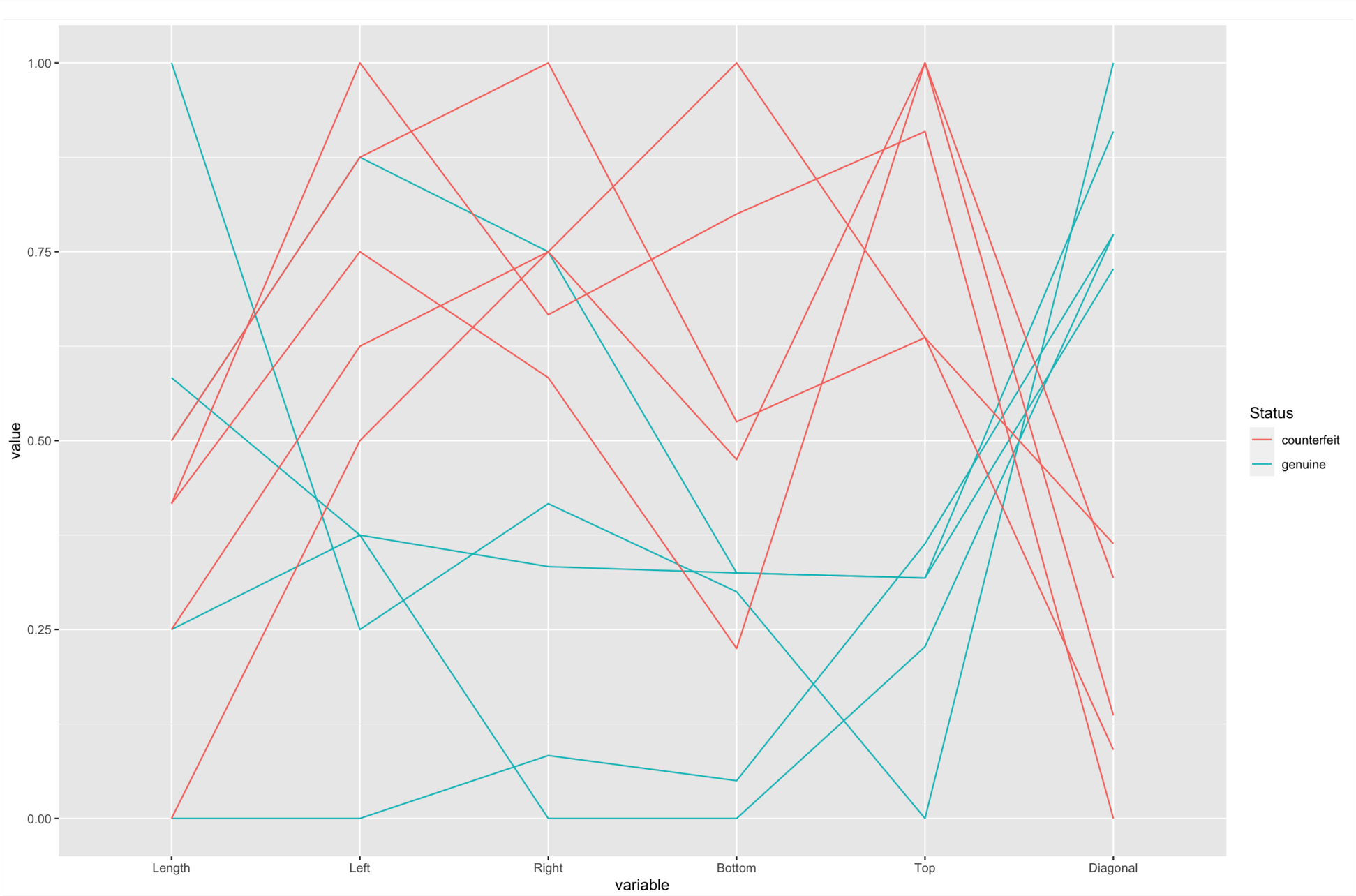


Parallel Coordinate Plots (平行坐标图)

- 平行坐标图是用来表示高维数据的一种方法.
  - 它在相互平行的坐标轴上绘制数据点，并用直线将它们连接起来.
  - 先将所有变量的取值变换为 $\max = 1$、$\min = 0$.
  - 画一条水平线，其第 j 个位置表示第 j 个变量.
  - 变换后的变量值 x_{ij} 对应于第 j 个纵轴上的一点.
  - 但是，它对变量的顺序也很敏感.
  - 位于 `GGally` 包中的函数 `ggparcoord()` 可作平行坐标图.

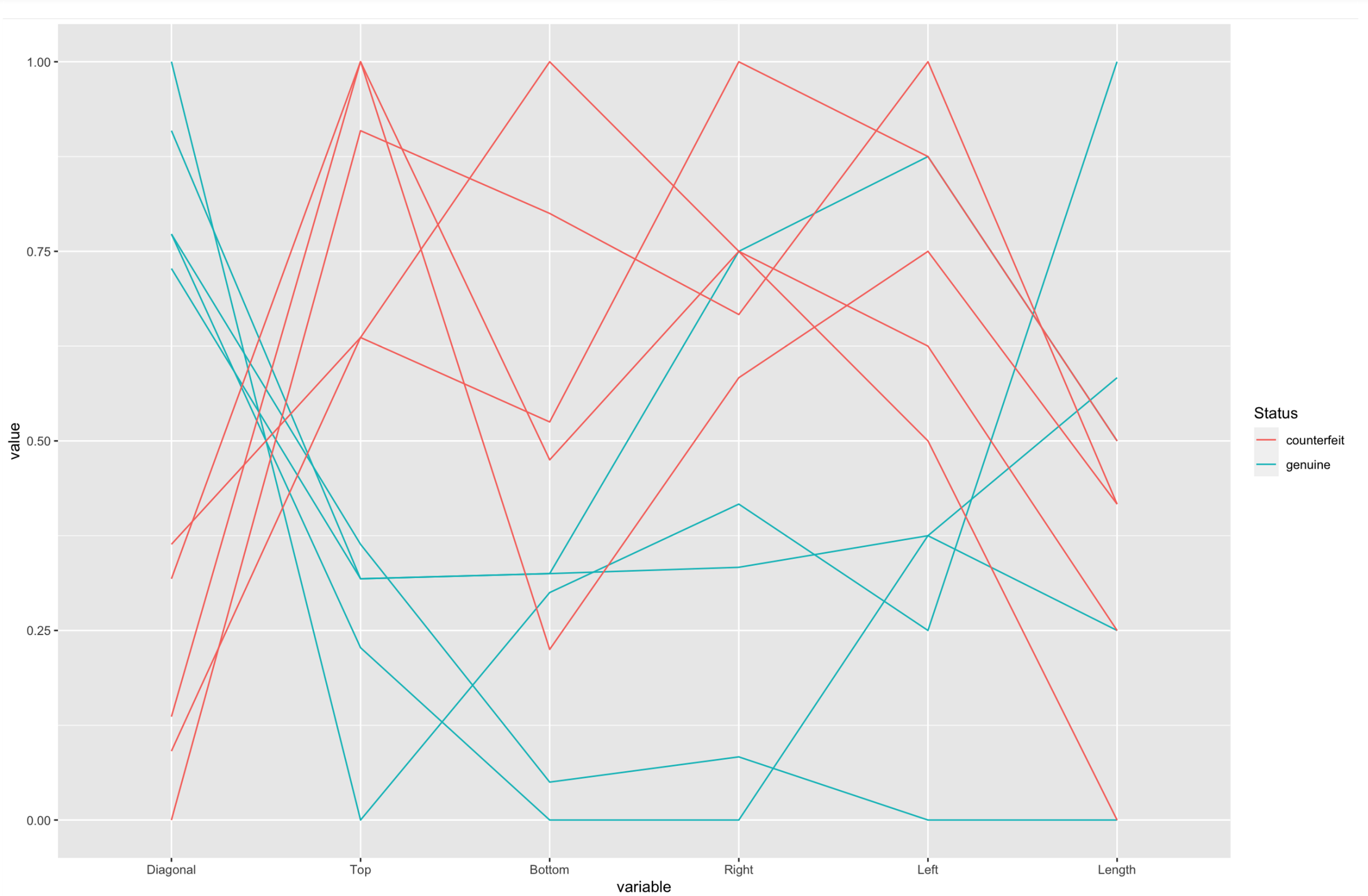
Parallel Coordinate Plots (平行坐标图)

```
library(GGally)
banknote.sub = banknote[96:105,]
ggparcoord(data = banknote.sub, columns = 2:7, groupColumn = 1, scale = 'uniminmax')
```



Parallel Coordinate Plots (平行坐标图)

```
ggparcoord(data = banknote.sub, columns = 7:2, groupColumn = 1, scale = 'uniminmax')
```



Parallel Coordinate Plots (平行坐标图)

```
ggparcoord(data = banknote, columns = 2:7, groupColumn = 1, scale = 'uniminmax')
```

Parallel Coordinate Plots (平行坐标图)

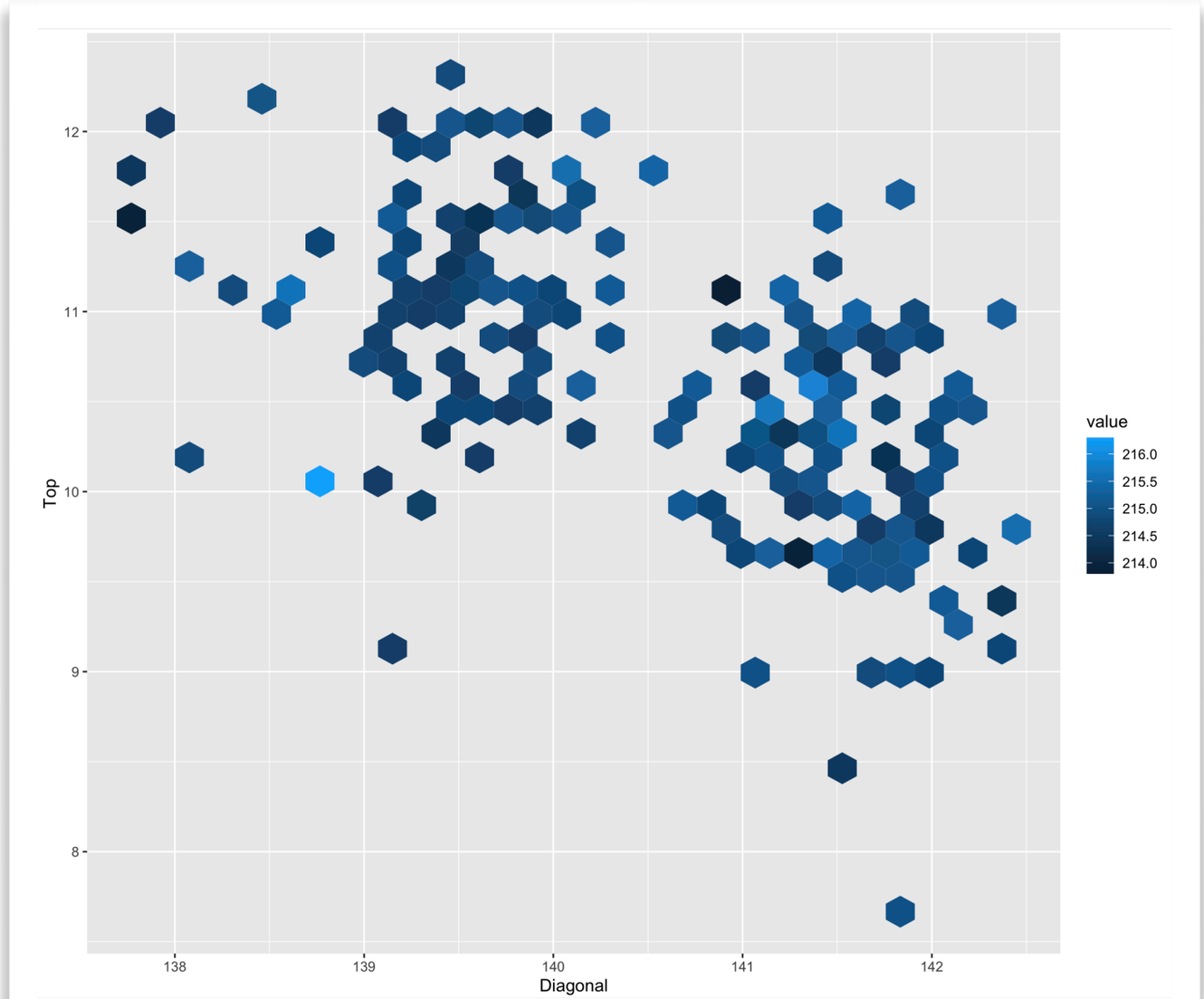
```
ggparcoord(data = banknote, columns = 7:2, groupColumn = 1, scale = 'uniminmax')
```

Hexagon Plots (六边形图)

- 在几何学中，六边形是有六条边和六个顶点的一个多边形.
- 六边形图是一种以六边形为边界的二元直方图.
- 当观测值数量 n 很大时，这是对数据集的结构很有用的一种可视化方法.
  - 将 xy 平面上 $\left(x \text{ 的取值区间}\ ,\ y \text{ 的取值区间}\right)$ 对应的区域用由六边形的规则网格镶嵌.
  - 计算落入每一个六边形当中观测点的个数.
  - 对于计数结果 > 0 的六边形，使用颜色渐变或按计数比例改变半径来画六边形.
- 即使 $n \geq 10^6$，该方法也能非常快速有效地显示数据集的结构.

Hexagon Plots (六边形图)

```
ggplot(data = banknote, aes(x = Diagonal, y = Top, z = Length)) +
  stat_summary_hex()
```



Hexagon Plots (六边形图)

?diamonds

diamonds {ggplot2} R Documentation

Prices of over 50,000 round cut diamonds

Description

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

Usage

Format

A data frame with 53940 rows and 10 variables:

price
: price in US dollars ($326–$18,823)

carat
: weight of the diamond (0.2–5.01)

cut
: quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color
: diamond colour, from D (best) to J (worst)

clarity
: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

x
: length in mm (0–10.74)

y
: width in mm (0–58.9)

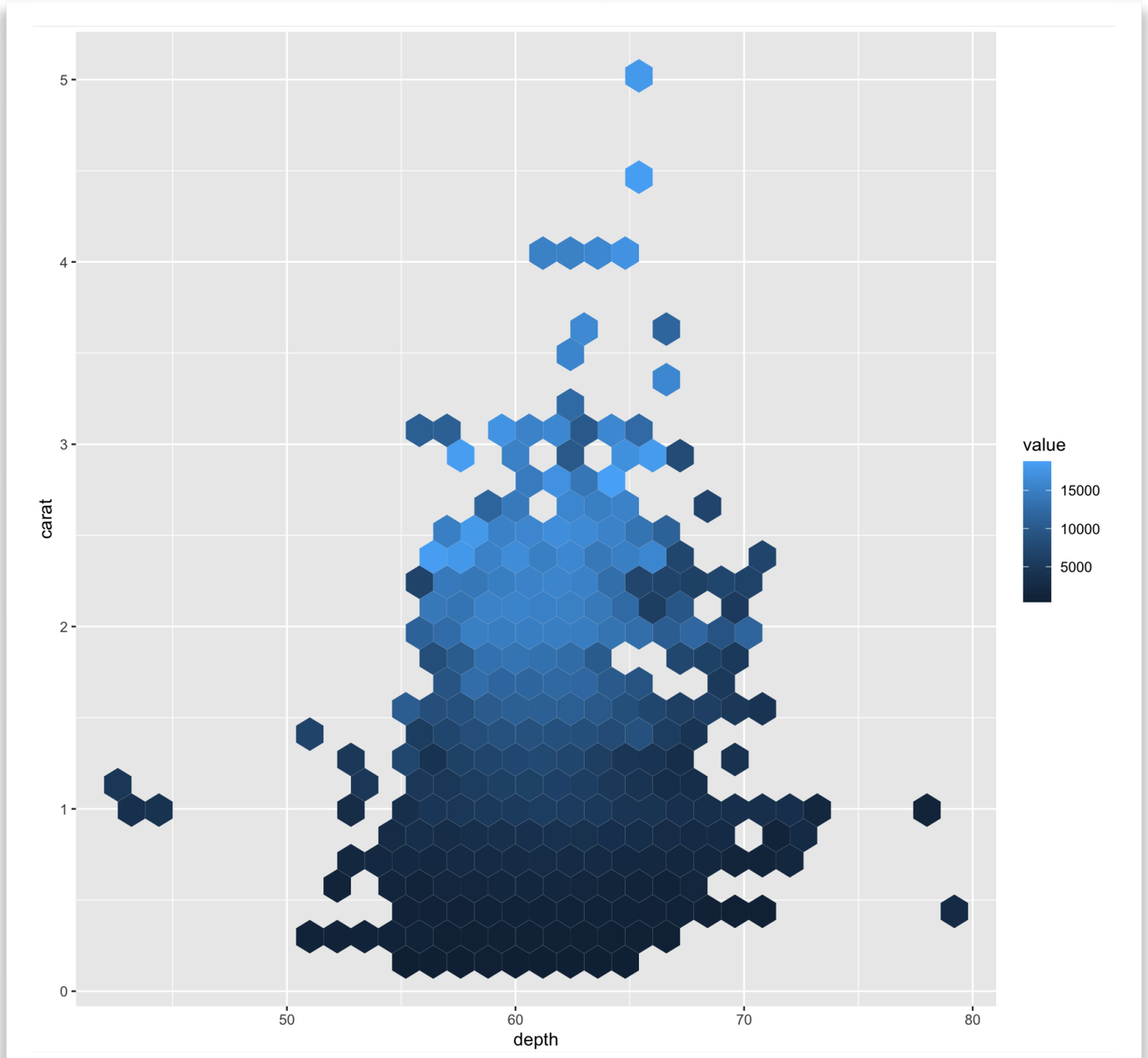
z
: depth in mm (0–31.8)

depth
: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)

table
: width of top of diamond relative to widest point (43–95)

Hexagon Plots (六边形图)

```
ggplot(data = diamonds, aes(x = depth, y = carat, z = price)) +
  stat_summary_hex()
```



Boston Housing (波士顿房屋)

- 波士顿房屋数据集由 Harrison 和 Rubinfeld 于 1978 收集.
- 目的：研究空气质量是否对房价有影响.
- 我们将用该数据集验证多元统计分析的许多方法.

```
library(MASS)
str(Boston)
```

14 个变量

506 个观测值

```
> str(Boston)
'data.frame':   506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Boston Housing (波士顿房屋)

- 波士顿房屋数据集由 Harrison 和 Rubinfeld 于 1978 收集.
- 目的：研究空气质量是否对房价有影响.
- 我们将用该数据集验证多元统计分析的许多方法.

X_{14}：自住房屋价格的中位数(单位: 1000美元)

X_4：Charles 河 (1 指河边，0 为其它)

X_6：每个住宅的平均房间数

X_8：到波士顿五个就业中心的加权距离

X_{10}：每一万美元的全额财产税税率

head(Boston)

```
> head(Boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

X_1：人均犯罪率

X_2：划作大型住宅用地的比例

X_3：非零售业务占地的比例

X_5：一氧化氮浓度

X_7：1940年之前建造的自住房的比例

X_9：辐射状高速公路的可达性指数

X_{11}：生师比

X_{12}：$1000\left(B-0.63\right)^2 \mathrm{I}(B<0.63)$

其中 B是非裔美国人的比例

X_{13}：社会低层人口的百分比

Boston Housing (波士顿房屋)

- 由平行坐标图可以看出什么?

x_{12} 与 x_{14} 强相关，因为 x_{12} 的下侧几乎无青色的线条.

```
library(GGally)
c = Boston$medv > median(Boston$medv) # 按 X14 是否大于其中位数分为两类
boston = cbind(Boston, c)
ggparcoord(data = boston, columns = 1:14, groupColumn = 15, scale = 'uniminmax')
```

一些变量呈现强相关.

x_{11} 与 x_{14} 强相关，因为 x_{11} 的下侧几乎都是青色的线条:

说明x_{11}的低值(低生师比)会导致 x_{14} 取值较大(高房价).

x_{13} 与 x_{14} 呈负相关.

$X_{14} > \text{median}\,(X_{14})$

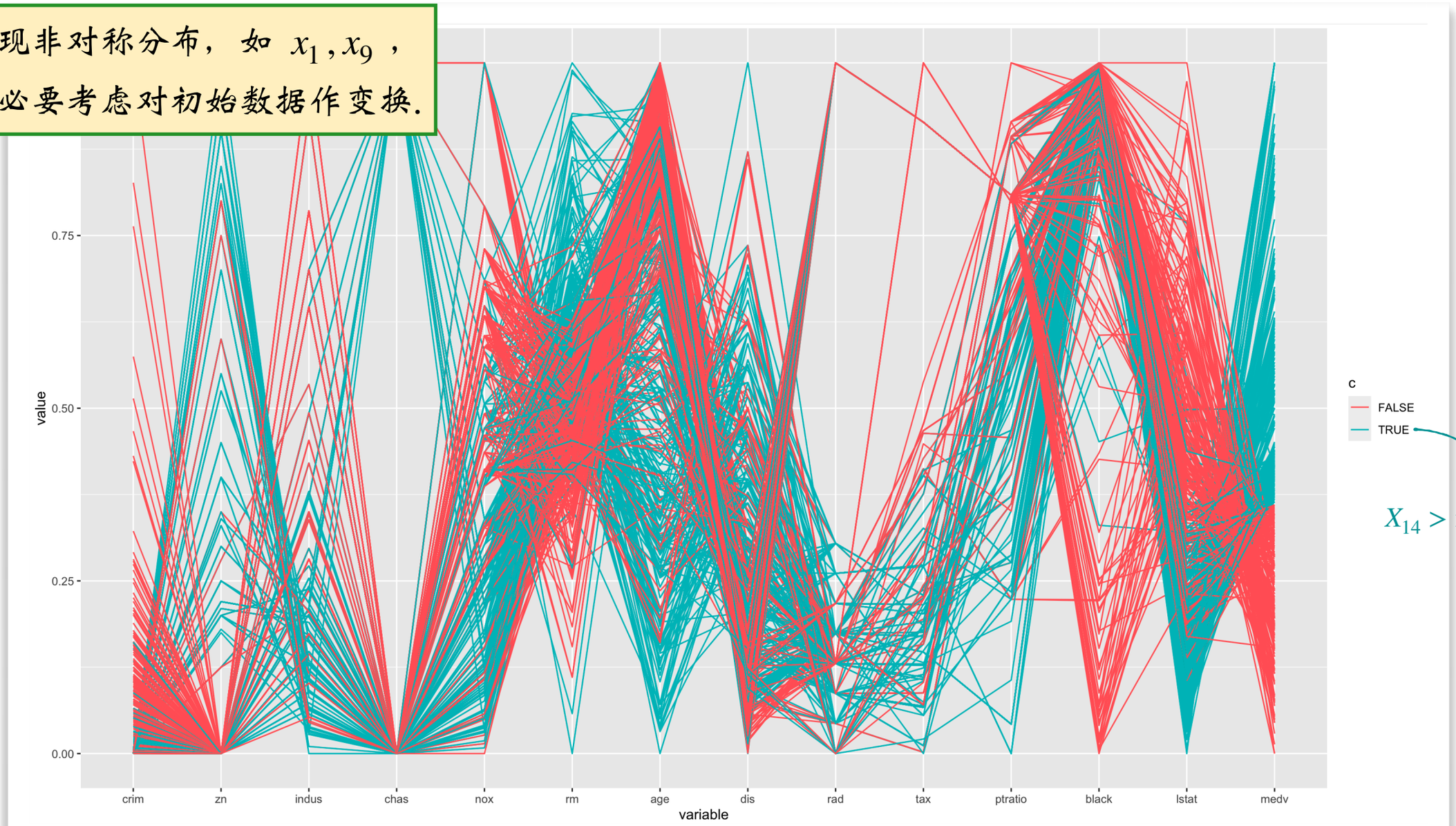
Boston Housing (波士顿房屋)

- 由平行坐标图可以看出什么?

```
library(GGally)
c = Boston$medv > median(Boston$medv) # 按 X14 是否大于其中位数分为两类
boston = cbind(Boston, c)
ggparcoord(data = boston, columns = 1:14, groupColumn = 15, scale = 'uniminmax')
```

所有变量的取值均变换到$[0,1]$区间.

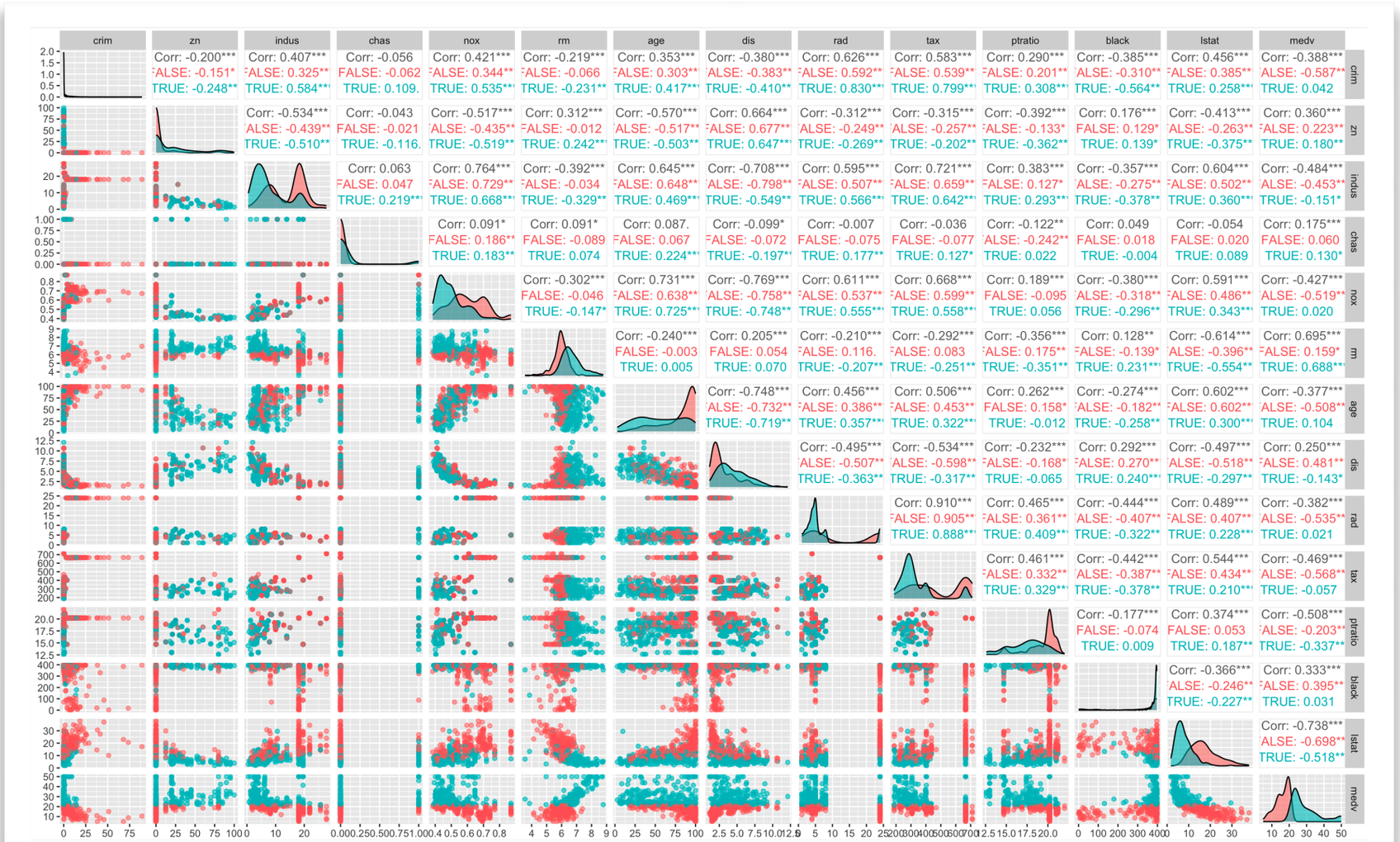
变量呈现非对称分布，如 x_1, x_9 ，说明有必要考虑对初始数据作变换.



Boston Housing (波士顿房屋)

- 散点图矩阵
  - 平行坐标图的局限：多条线段互相堆叠在一起.
  - 可以将全部14个变量画在一张散点图矩阵中，但很难看出其中的规律性.

```
ggpairs(boston, columns =1:14, ggplot2::aes(colour = c, alpha = 0.2))
```



Boston Housing (波士顿房屋)

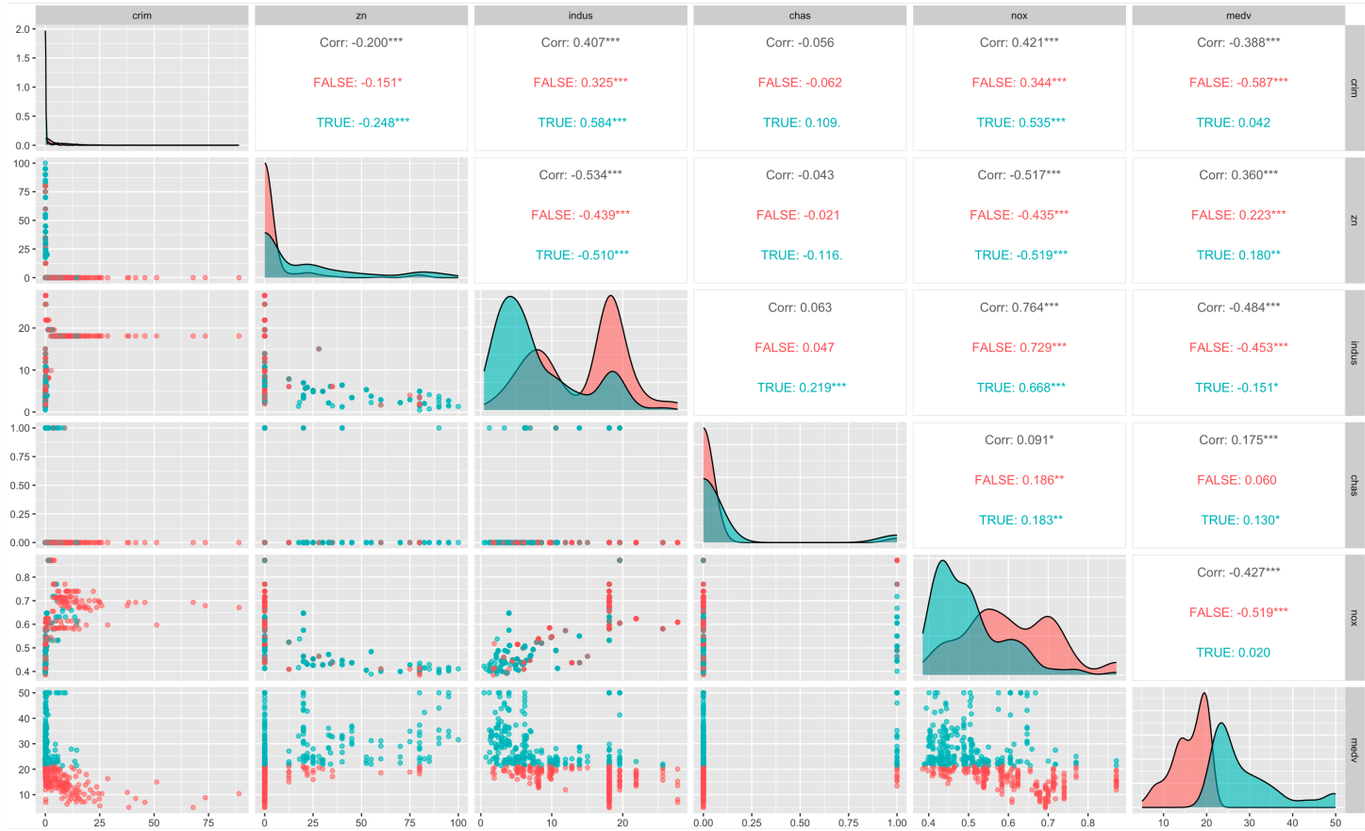
- 散点图矩阵
  - 平行坐标图的局限：多条线段互相堆叠在一起.
  - 可以将全部14个变量画在一张散点图矩阵中，但很难看出其中的规律性.
  - 变量 $X_4, X_6, X_{11}, X_{12}, X_{13}, X_{14}$ 的散点图矩阵.

```
ggpairs(boston, columns = c(4, 6, 11, 12, 13, 14), ggplot2::aes(colour = c, alpha = 0.2))
```

Boston Housing (波士顿房屋)

- 散点图矩阵
  - 平行坐标图的局限：多条线段互相堆叠在一起.
  - 可以将全部14个变量画在一张散点图矩阵中，但很难看出其中的规律性.
  - 变量 $X_1, X_2, X_3, X_4, X_5, X_{14}$ 的散点图矩阵.

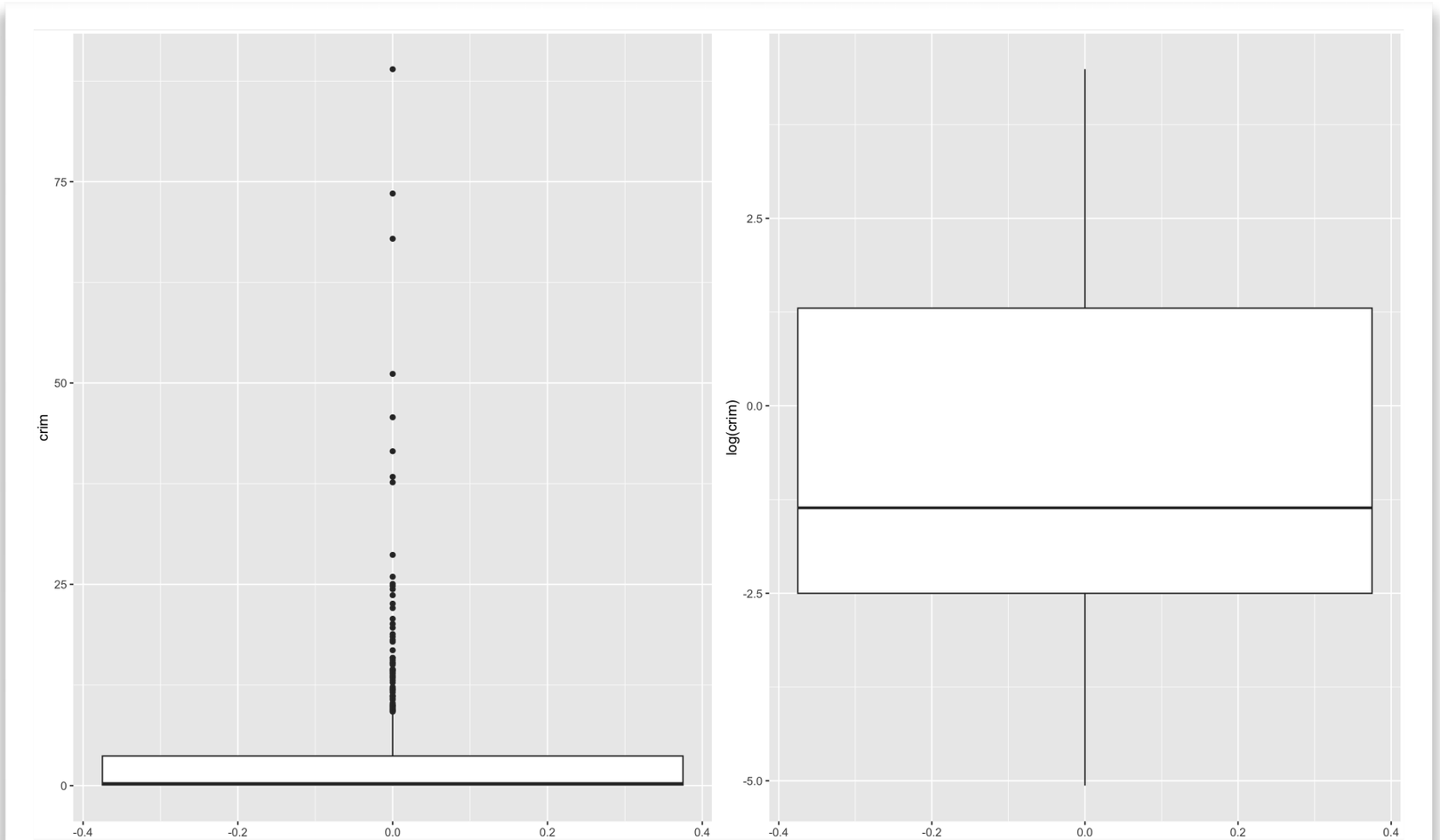
```
ggpairs(boston, columns = c(1:5, 14), ggplot2::aes(colour = c, alpha = 0.2))
```



Boston Housing (波士顿房屋)

- 人均犯罪率 X_1
  - 对数变换之后，该变量的分布更为对称.

```
library(magrittr)
library(gridExtra)
Box_X1 = ggplot(data = boston, aes(y = crim)) + geom_boxplot(notch = FALSE) # 变量 X_1 的箱线图
Box_Log_X1 = ggplot(data = boston, aes(y = log(crim))) + geom_boxplot(notch = FALSE) # 变量 X_1 取对数之后的箱线图
grid.arrange(Box_X1, Box_Log_X1, ncol = 2)
```

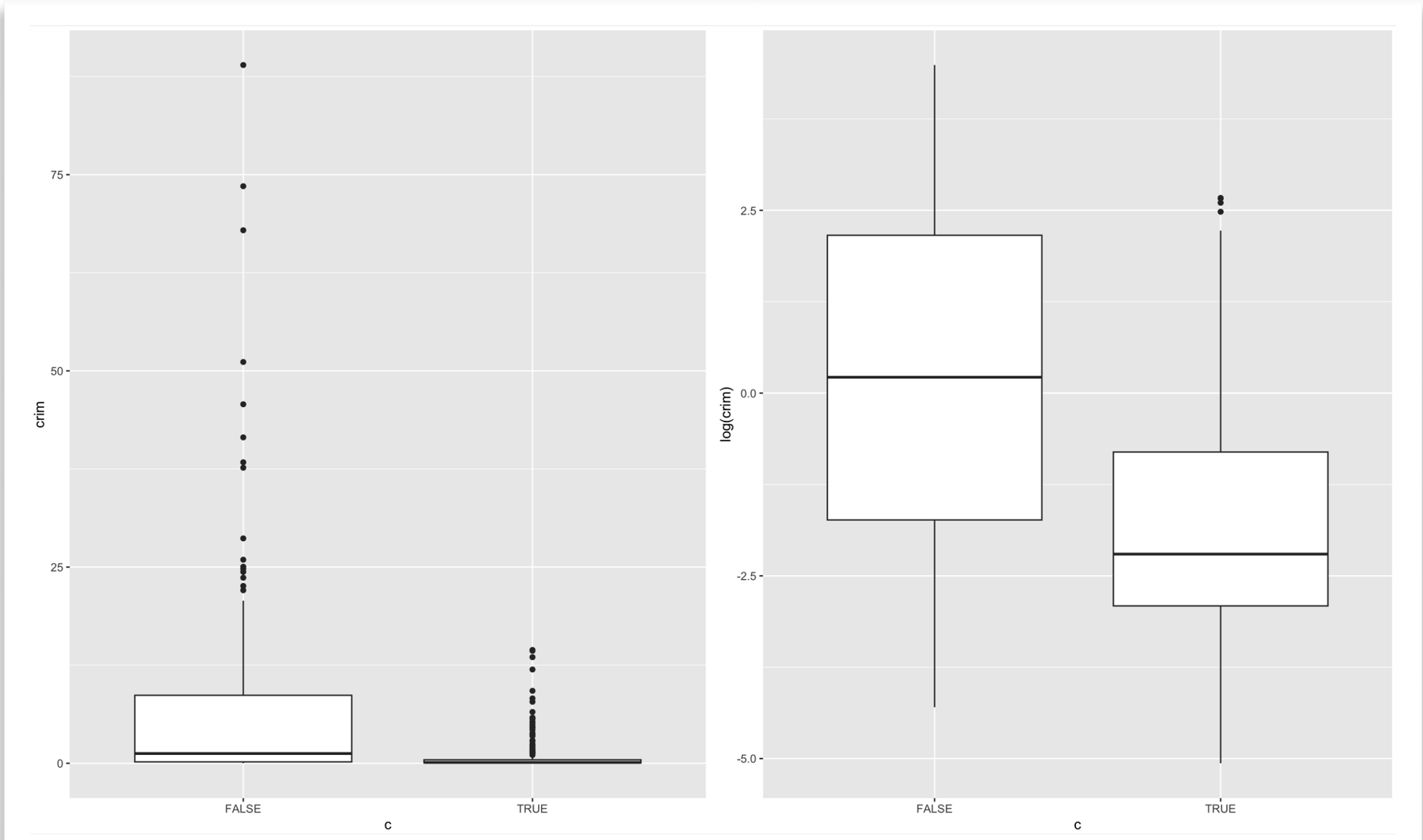


Boston Housing (波士顿房屋)

可能存在两个均值不同的子分布?

- 人均犯罪率 X_1
  - 对数变换之后，该变量的分布更为对称.

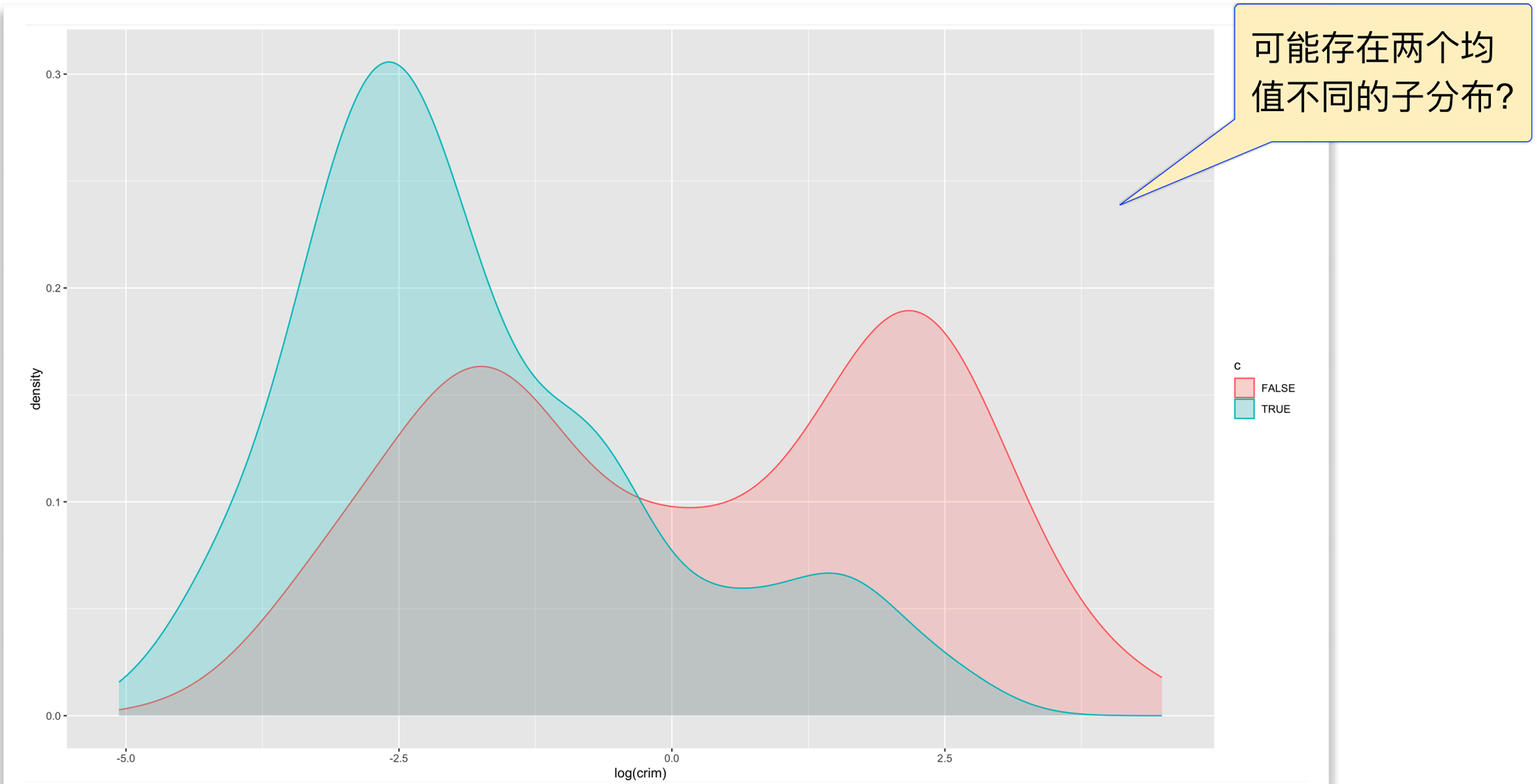
```
Boxc_X1 = ggplot(data = boston, aes(x = c, y = crim)) + geom_boxplot(notch = FALSE) # 变量 X_1 依房价的分类箱线图
Boxc_Log_X1 = ggplot(data = boston, aes(x = c, y = log(crim))) + geom_boxplot(notch = FALSE) # 变量 X_1 取对数之后依房价的分类箱线图
grid.arrange(Boxc_X1, Boxc_Log_X1, ncol = 2)
```



Boston Housing (波士顿房屋)

- 人均犯罪率 X_1
  - 对数变换之后依房价的分类核密度图.

```
ggplot(boston, aes(log(crim), fill = c, color = c)) +
  geom_density(alpha = 0.2) # 对数变换之后依房价的分类核密度图
```

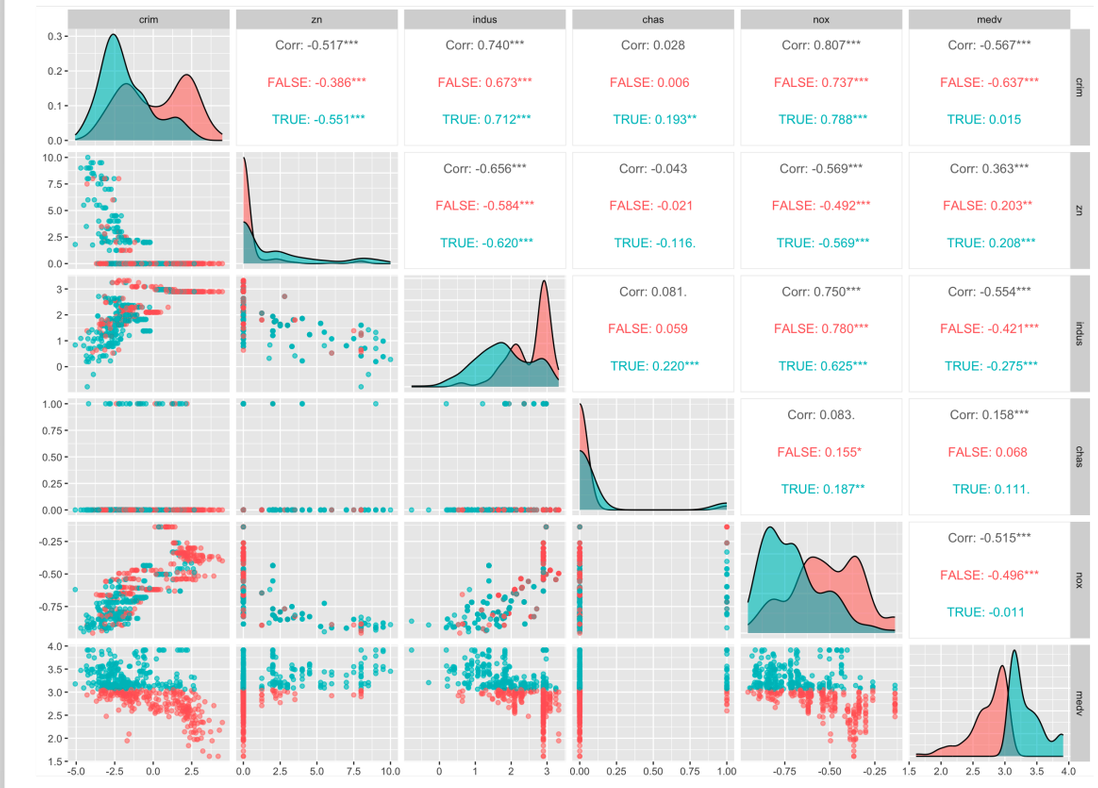


Boston Housing (波士顿房屋)

第一列散点图并未清晰呈现出 $\log(X_1)$ 两个不同总体的现象.

- 人均犯罪率 X_1
  - 变量 $X_1, X_2, X_3, X_4, X_5, X_{14}$ 变换之后的散点图矩阵.

```
# 变量 1:5, 14 变换之后的散点图矩阵.
Log_X1 = log(Boston$crim)
X2 = Boston$zn / 10
Log_X3 = log(Boston$indus)
Log_X5 = log(Boston$nox)
Log_X14 = log(Boston$medv)
boston_sub = data.frame(crim = Log_X1, zn = X2, indus = Log_X3, chas = Boston$chas, nox = Log_X5, medv = Log_X14, c = boston$c)
ggpairs(boston_sub, columns = 1:6, ggplot2::aes(colour = c, alpha = 0.2))
```

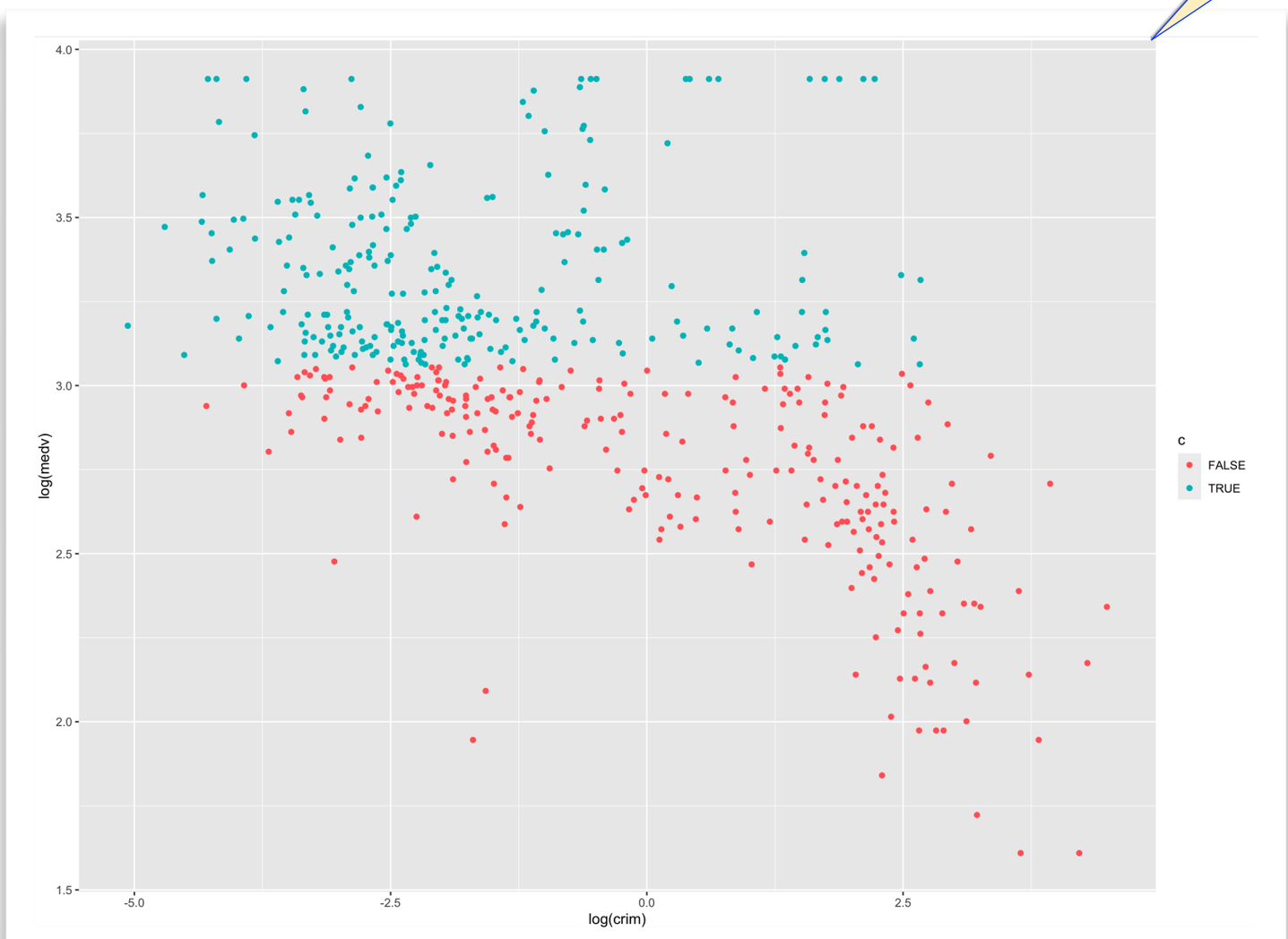


Boston Housing (波士顿房屋)

- 人均犯罪率 X_1
  - 变量 $\log X_1$ 与 $\log X_{14}$ 的散点图.

```
ggplot(boston_sub, aes(x = crim, y = medv, color = c)) +
  geom_point() +
  xlab("log(crim)") +
  ylab("log(medv)")
```

$\log X_1$ 与 $\log X_{14}$ 呈现较强的负相关.



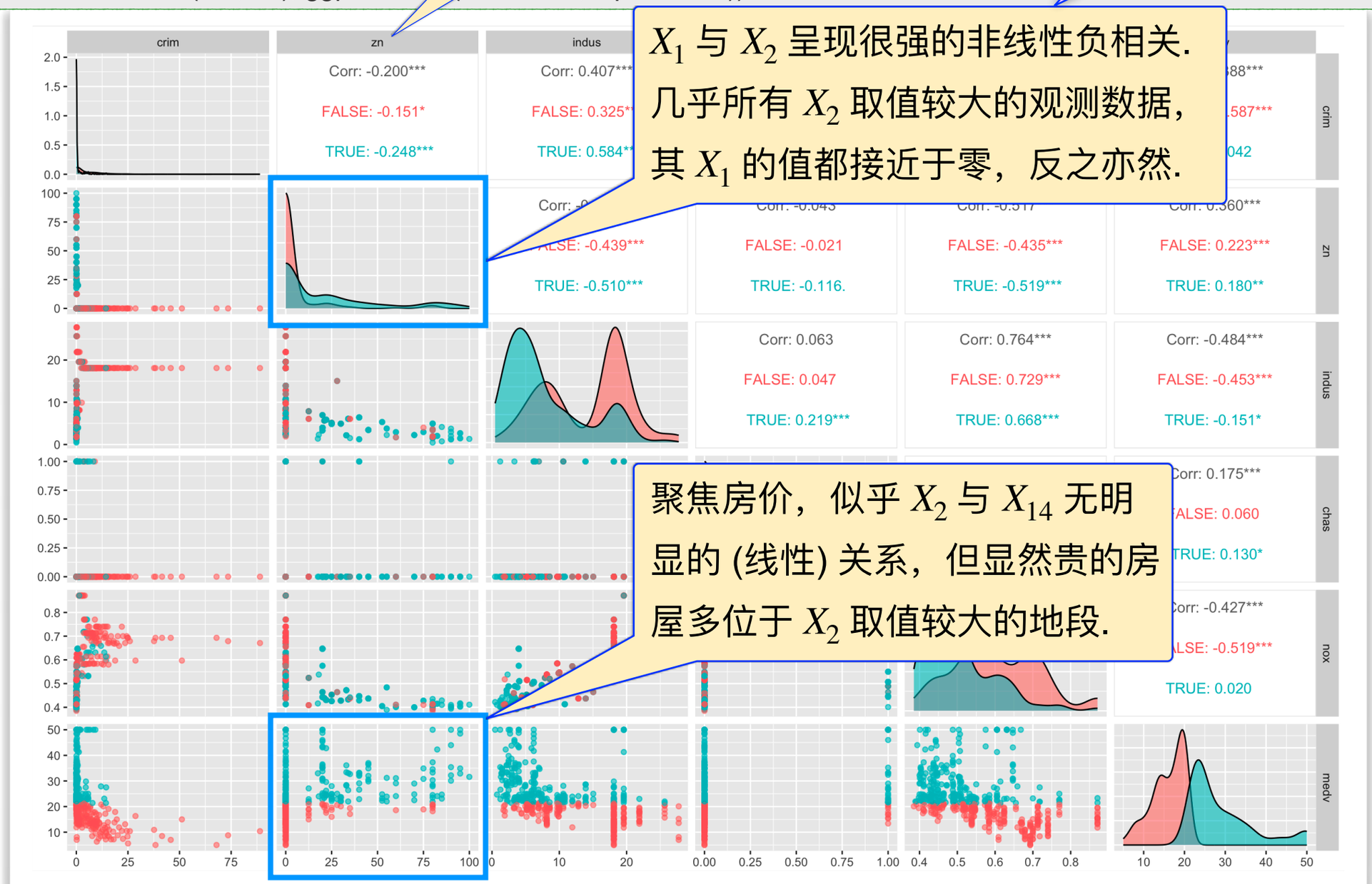
Boston Housing (波士顿房屋)

- 划作大型住宅用地的比例 X_2
  - 变量 $X_1, X_2, X_3, X_4, X_5, X_{14}$ [illegible].

许多观测数据中的 $X_2 = 0$.

可能缘于这些地段的位置，如城区的犯罪率更高，同时不太可能有大片地划作住宅用.

```
ggpairs(boston, columns = c(1:5, 14), ggplot2::ae[illegible]olour = c, alpha = 0.2))
```



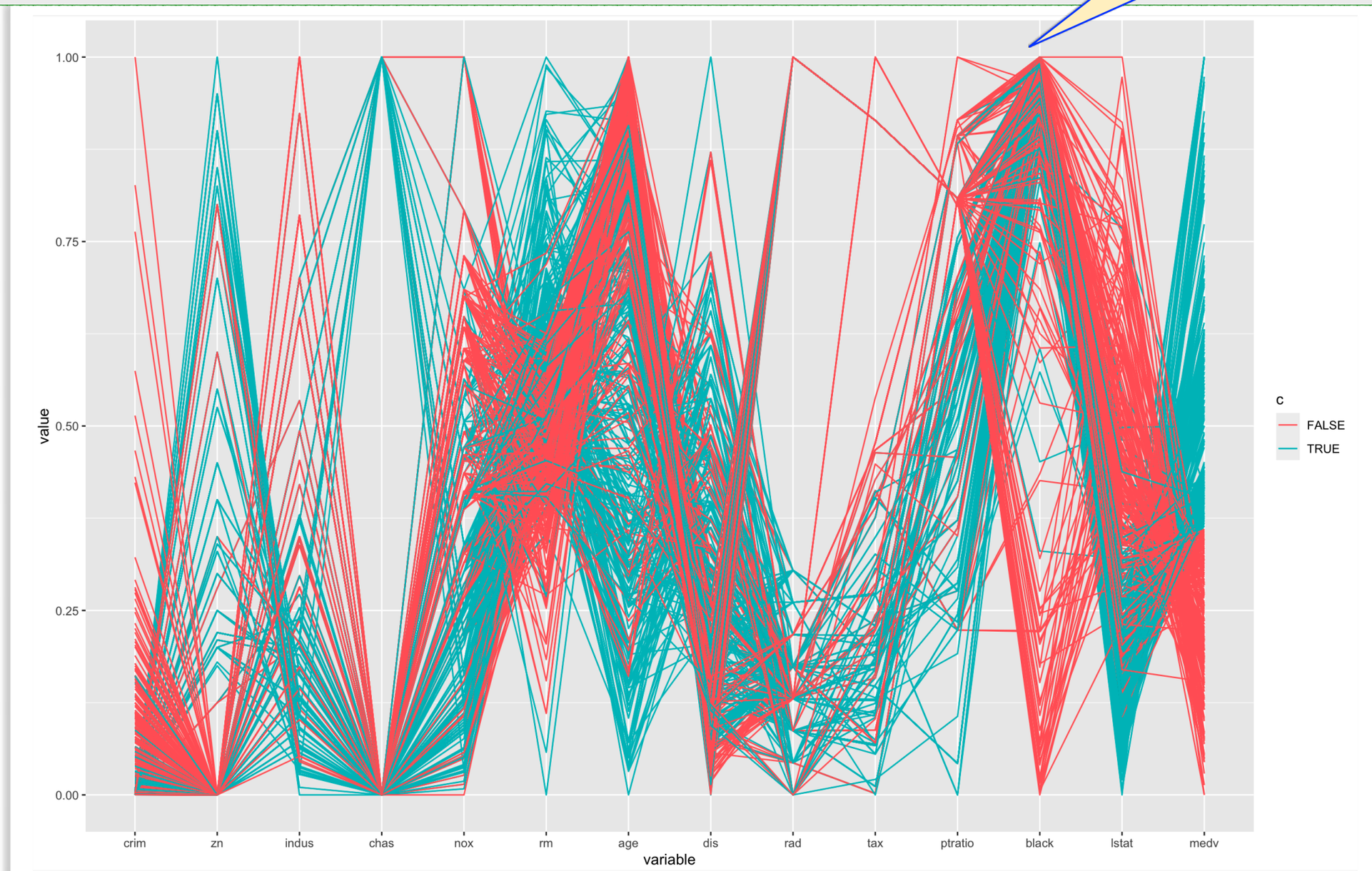
X_1 与 X_2 呈现很强的非线性负相关. 几乎所有 X_2 取值较大的观测数据，其 X_1 的值都接近于零，反之亦然.

聚焦房价，似乎 X_2 与 X_{14} 无明显的 (线性) 关系，但显然贵的房屋多位于 X_2 取值较大的地段.

Boston Housing (波士顿房屋)

- 非零售业务占地的比例 X_3
  - Boston 房屋数据集的平行坐标图.

```
ggparcoord(data = boston, columns = 1:14, groupColumn = 15, scale = 'uniminmax')
```

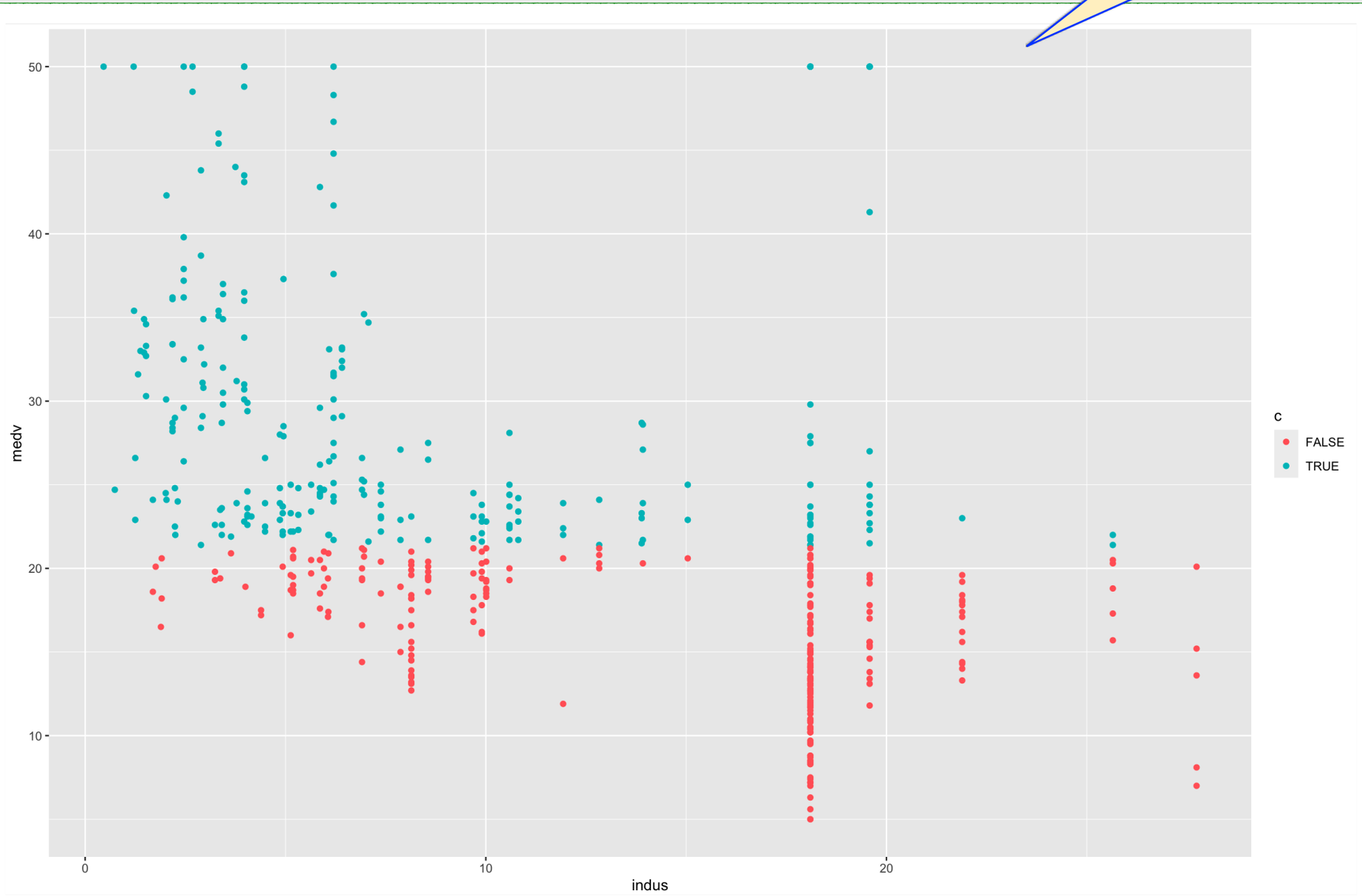


Boston Housing (波士顿房屋)

- 非零售业务占地的比例 X_3
  - 变量 X_3 与 X_{14} 的散点图.

X_3 与 X_{14} 呈现明显负相关.

```
ggplot(boston, aes(x = indus, y = medv, color = c)) + geom_point()
```

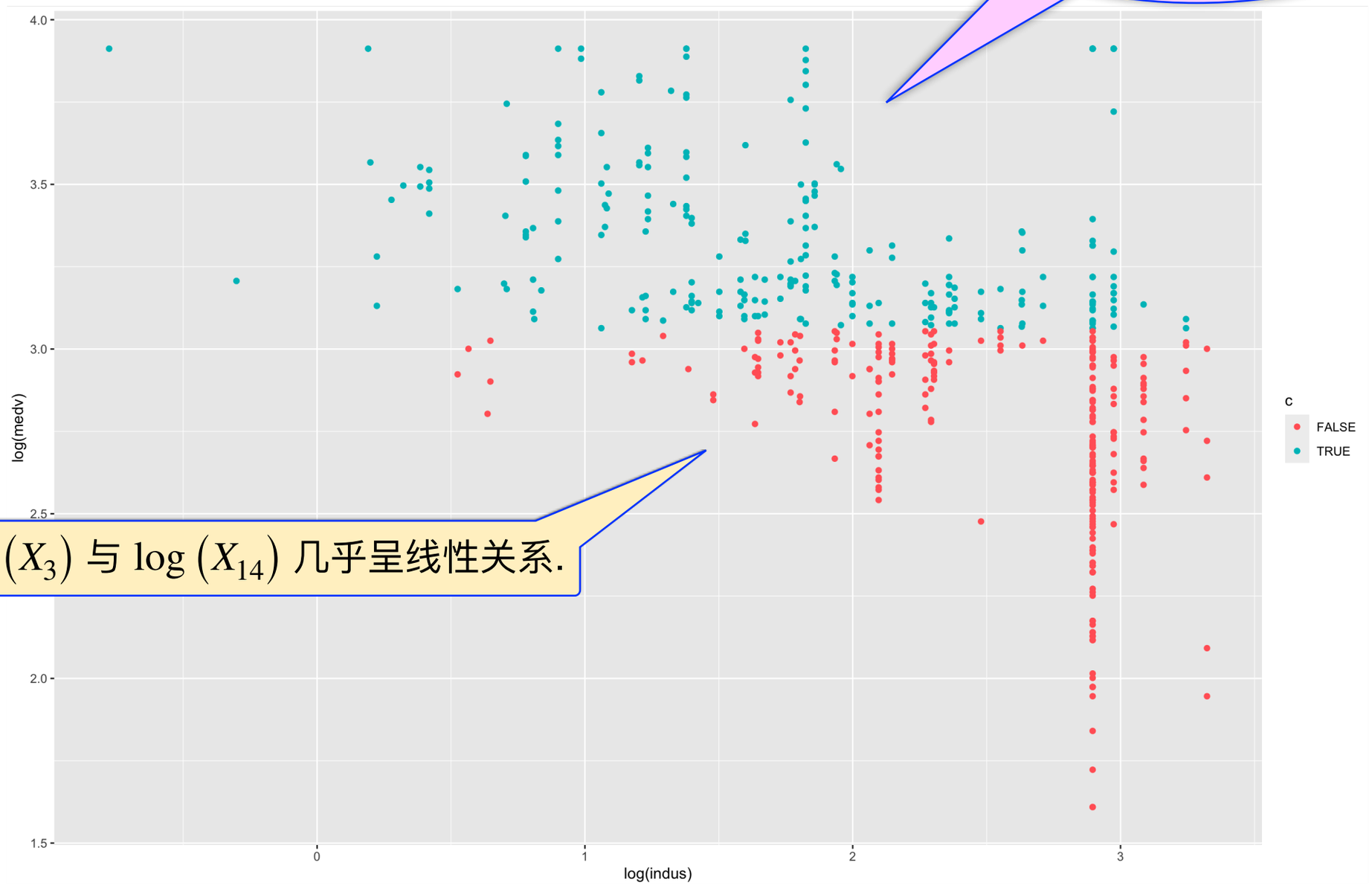


Boston Housing (波士顿房屋)

- 非零售业务占地的比例 X_3
  - 变量 X_3 与 X_{14} 经对数变换之后的散点图.

缘于非零售业务会引发噪音等污染. 因此，在作线性回归分析预测 X_{14} 时，需将 X_3 作为一个解释变量.

```
ggplot(boston, aes(x = log(indus), y = log(medv), color = c)) + geom_point()
```

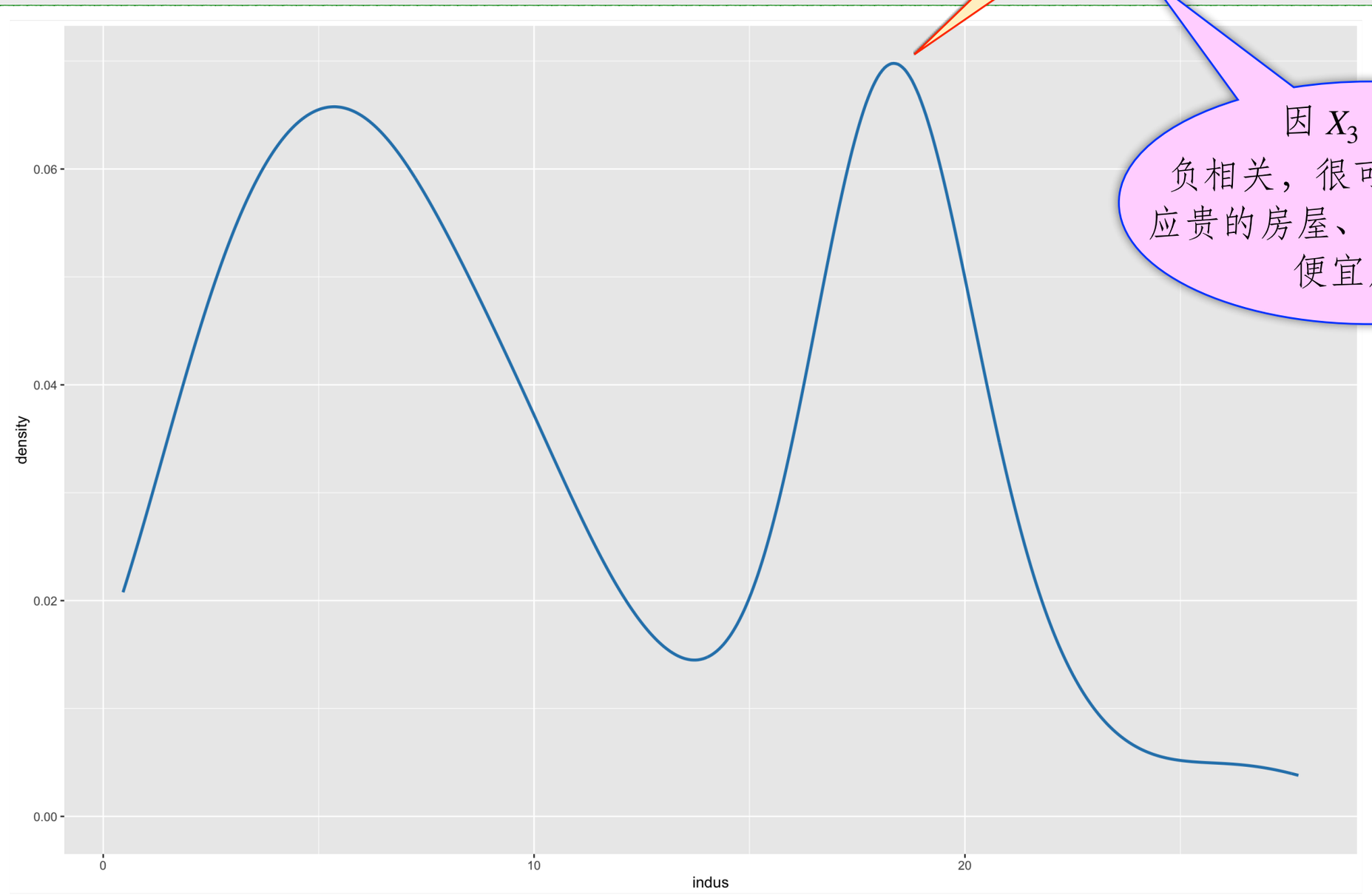


$\log(X_3)$ 与 $\log(X_{14})$ 几乎呈线性关系.

Boston Housing (波士顿房屋)

- 非零售业务占地的比例 X_3
  - 变量 X_3 的核密度图.

```
ggplot(boston, aes(indus)) + geom_density(linewidth = 1, colour = 'steelblue')
```



双峰表明观测数据可分为两类.

因 X_3 与 X_{14} 负相关，很可能某一类对应贵的房屋、另一类则对应便宜房屋.

Boston Housing (波士顿房屋)

- 虚拟变量 Charles 河 X_4
  - Boston 房屋数据集的平行坐标图.

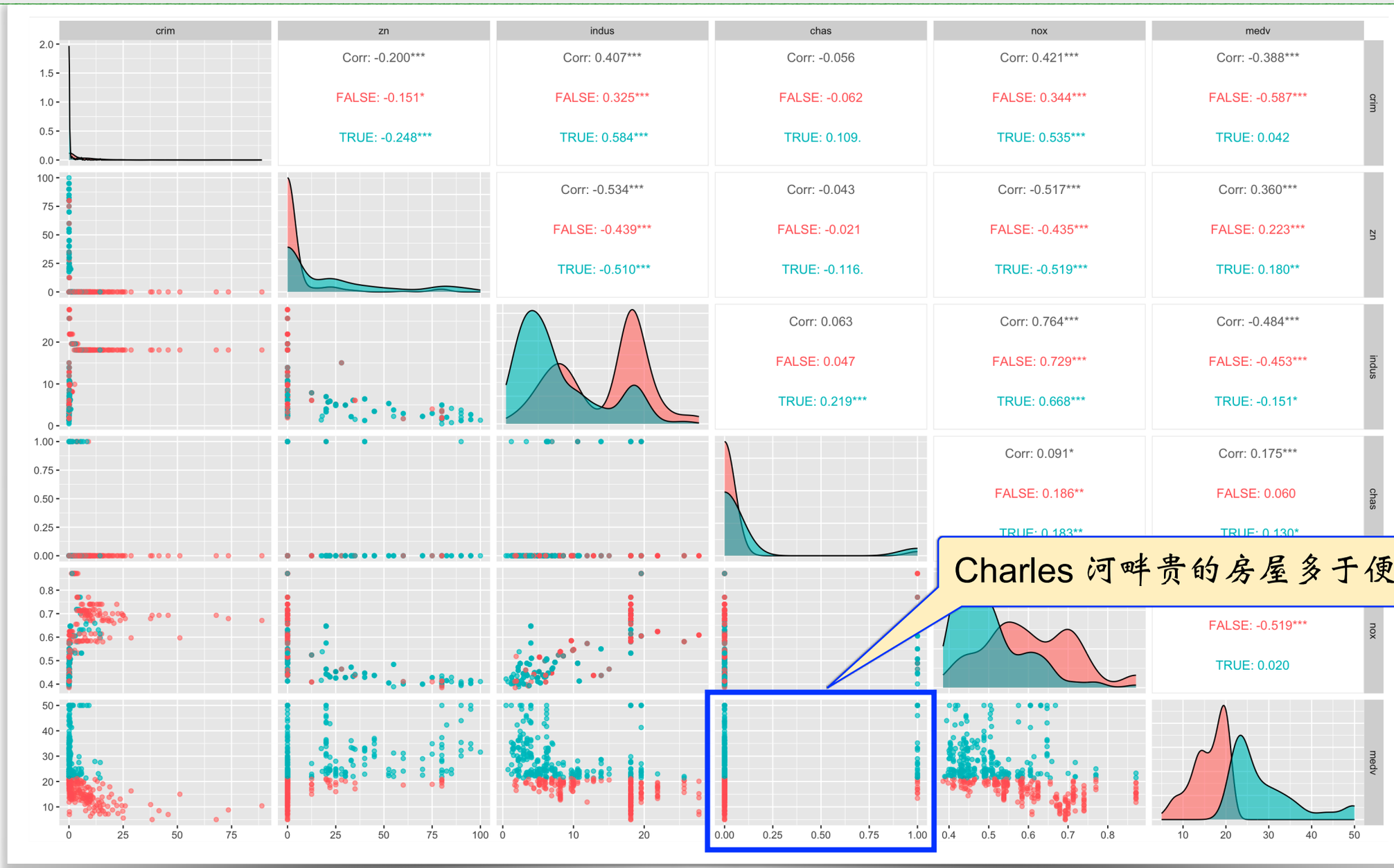
Charles 河畔贵的房屋多于便宜的房屋.

```
ggparcoord(data = boston, columns = 1:14, groupColum = 15, scale = 'uniminmax')
```

Boston Housing (波士顿房屋)

- 虚拟变量 Charles 河 X_4
  - 变量 $X_1, X_2, X_3, X_4, X_5, X_{14}$ 的散点图矩阵.

```
ggpairs(boston, columns = c(1:5, 14), ggplot2::aes(colour = c, alpha = 0.2))
```



Boston Housing (波士顿房屋)

- 虚拟变量 Charles 河 X_4
  - 尽管如此，我们可能还是会怀疑靠近河流真的会影响房价吗?
  - 校正后的原始数据集：增加了五个变量.

```
library(mlbench)
data(BostonHousing2)
head(BostonHousing2)
```

自住房屋价格校正后的中位数(单位: 1000美元)

```
> head(BostonHousing2)
      town tract      lon      lat medv cmedv    crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
1   Nahant  2011 -70.9550 42.2550 24.0  24.0 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
2 Swampscott  2021 -70.9500 42.2875 21.6  21.6 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
3 Swampscott  2022 -70.9360 42.2830 34.7  34.7 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
4 Marblehead  2031 -70.9280 42.2930 33.4  33.4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
5 Marblehead  2032 -70.9220 42.2980 36.2  36.2 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
6 Marblehead  2033 -70.9165 42.3040 28.7  28.7 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
```

普查区域的纬度

普查区域的经度

人口普查区域

城镇名称

Boston Housing (波士顿房屋)

- 虚拟变量 Charles 河 X_4
  - 尽管如此，我们可能还是会怀疑靠近河流真的会影响房价吗?
  - 观察原始数据集，会清晰看到观测值当中 $X_4 = 1$ 的住宅区都很接近.

```
subset(BostonHousing2, chas == 1)
```

Charles 河畔房价高可能纯属偶然.

高房价可能是由其它因素造成的，如生师比或非零售企业面积的比例.

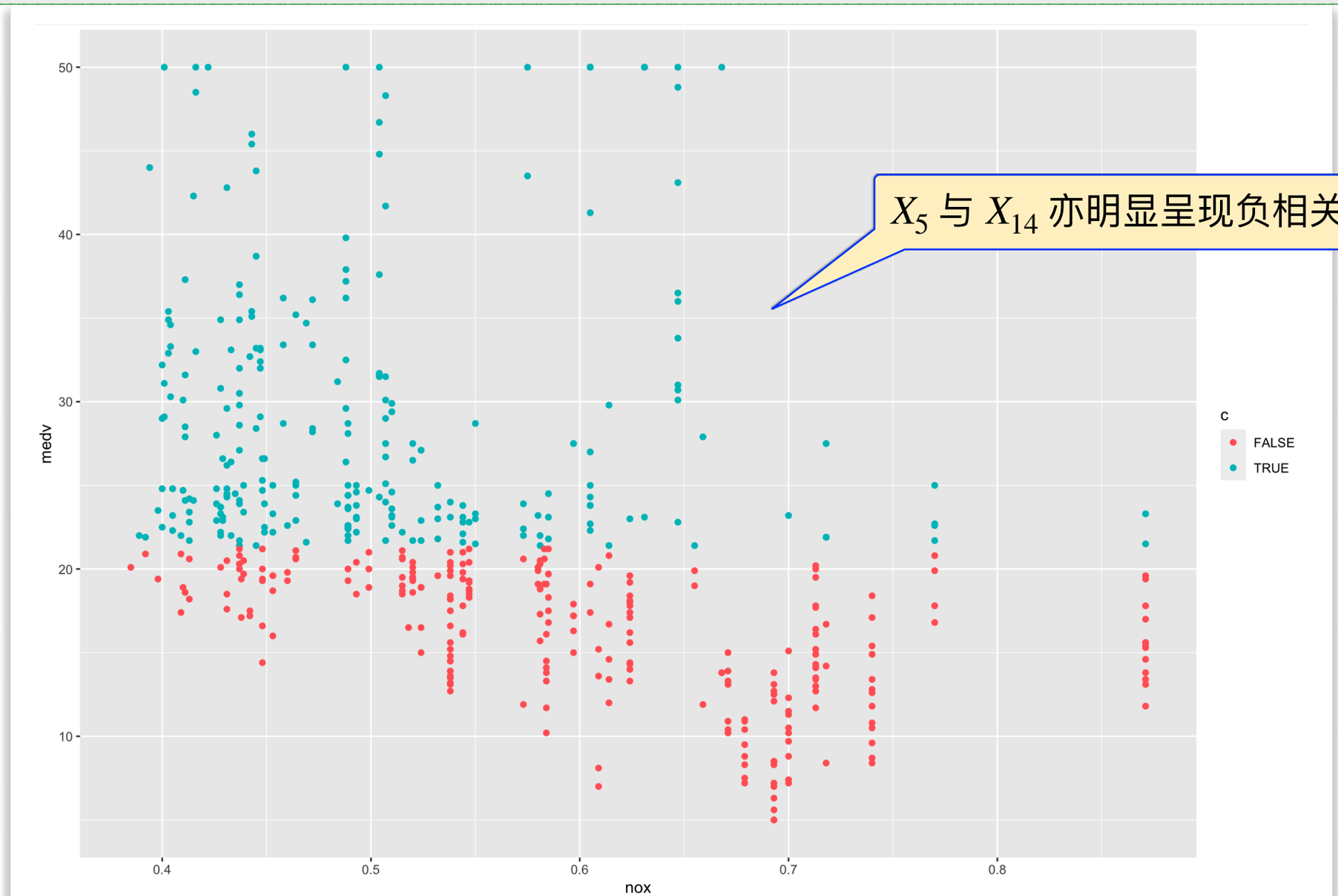
```
> subset(BostonHousing2, chas == 1)
```

	town	tract	lon	lat	medv	cmedv						dis	rad	tax	ptratio	b	lstat
143	Cambridge	3521	-71.0480	42.2222	13.4	13.4	3.3					1.3216	5	403	14.7	396.90	26.82
153	Cambridge	3531	-71.0590	42.2170	15.3	15.3			0	19.58	1	0.8710	5.012	88.0	1.6102	5 403	14.7	343.28	12.12
155	Cambridge	3533	-71.0662	42.2162	17.0	17	1.41385	0	19.58	1	0.8710	6.129	96.0	1.74	5 403	14.7	321.02	15.12	
156	Cambridge	3534	-71.0680	42.2150	15.6	15.6	3.53501	0	19.58	1	0.8710	6.152	82.6	1.7455		14.7	88.01		
161	Cambridge	3539	-71.0700	42.2214	27.0	27.0	1.27346	0	19.58	1	0.6050	6.250	92.6	1.7984	5 4				
163	Cambridge	3541	-71.0770	42.2250	50.0	50.0	1.83377	0	19.58	1	0.6050	7.802	98.2	2.0407	5 403				
164	Cambridge	3542	-71.0815	42.2250	50.0	50.0	1.51902	0	19.58	1	0.6050	8.375	93.9	2.1620	5 403				
209	Waltham	3684	-71.1503	42.2220	24.4	24.4	0.13587	0	10.59	1	0.4890	6.064	59.1	4.2392	4 27				
210	Waltham	3685	-71.1430	42.2217	20.0	20.0	0.43571	0	10.59	1	0.4890	5.344	100.0	3.8750	4 2				
211	Waltham	3686	-71.1435	42.2177	21.7	21.7	0.17446	0	10.59	1	0.4890	5.960	92.1	3.8771	4 2				
212	Waltham	3687	-71.1380	42.2216	19.3	19.3	0.37578	0	10.59	1	0.4890	5.404	88.6	3.6650	4 27				
213	Waltham	3688	-71.1335	42.2250	22.4	22.4	0.21719	0	10.59	1	0.4890	5.807	53.8	3.6526	4 277				
217	Watertown	3701	-71.1166	42.2230	23.3	23.3	0.04560	0	13.89	1	0.5500	5.888	56.0	3.1121	5 276				
219	Watertown	3703	-71.0933	42.2220	21.5	21.5	0.11069	0	13.89	1	0.5500	5.951	93.8	2.8893	5 276	16.4			
220	Watertown	3704	-71.1060	42.2185	23.0	23.0	0.11425	0	13.89	1	0.5500	6.373	92.4	3.3633	5 276	16.4	393.74	10.50	
221	Newton	3731	-71.1100	42.2137	26.7	26.7	0.35809	0	6.20	1	0.5070	6.951	88.5	2.8617	8 307	17.4	391.70	9.71	
222	Newton	3732	-71.1210	42.2166	21.7	21.7	0.40771	0	6.20	1	0.5070	6.164	91.3	3.0480	8 307	17.4	395.24	21.46	
223	Newton	3733	-71.1250	42.2134	27.5	27.5	0.62356	0	6.20	1	0.5070	6.879	77.7	3.2721	8 307	17.4	390.39	9.93	
235	Newton	3745	-71.1320	42.2142	29.0	29.0	0.44791	0	6.20	1	0.5070	6.726	66.5	3.6519	8 307	17.4	360.20	8.05	
237	Newton	3747	-71.1485	42.2110	25.1	25.1	0.52058	0	6.20	1	0.5070	6.631	76.5	4.1480	8 307	17.4	388.45	9.54	
270	Dedham	4021	-71.0980	42.1540	20.7	20.7	0.09065	20	6.96	1	0.4640	5.920	61.5	3.9175	3 223	18.6	391.34	13.65	
274	Dedham	4025	-71.1170	42.1510	35.2	35.2	0.22188	20	6.96	1	0.4640	7.691	51.8	4.3665	3 223	18.6	390.77	6.58	
275	Needham	4031	-71.1305	42.1675	32.4	32.4	0.05644	40	6.41	1	0.4470	6.758	32.9	4.0776	4 254	17.6	396.90	3.53	
277	Needham	4033	-71.1405	42.1632	33.2	33.2	0.10469	40	6.41	1	0.4470	7.267	49.0	4.7872	4 254	17.6	389.25	6.05	
278	Needham	4034	-71.1495	42.1730	33.1	33.1	0.06127	40	6.41	1	0.4470	6.826	27.6	4.8628	4 254	17.6	393.45	4.16	
283	Wellesley	4044	-71.1775	42.1735	46.0	46.0	0.06129	20	3.33	1	0.4429	7.645	49.7	5.2119	5 216	14.9	377.07	3.01	
284	Dover	4051	-71.1730	42.1475	50.0	50.0	0.01501	90	1.21	1	0.4010	7.923	24.8	5.8850	1 198	13.6	395.52	3.16	
357	Boston Allston-Brighton	1	-71.0830	42.2172	17.8	17.8	8.98296	0	18.10	1	0.7700	6.212	97.4	2.1222	24 666	20.2	377.73	17.60	
358	Boston Allston-Brighton	2	-71.0950	42.2120	21.7	21.7	3.84970	0	18.10	1	0.7700	6.395	91.0	2.5052	24 666	20.2	391.34	13.27	
359	Boston Allston-Brighton	3	-71.1007	42.2100	22.7	22.7	5.20177	0	18.10	1	0.7700	6.127	83.4	2.7227	24 666	20.2	395.43	11.48	
364	Boston Allston-Brighton	8	-71.0865	42.2150	16.8	16.8	4.22239	0	18.10	1	0.7700	5.803	89.0	1.9047	24 666	20.2	353.04	14.64	
365	Boston Back Bay	101	-71.0590	42.2098	21.9	21.9	3.47428	0	18.10	1	0.7180	8.780	82.9	1.9047	24 666	20.2	354.55	5.29	
370	Boston Back Bay	108	-71.0497	42.2125	50.0	50.0	5.66998	0	18.10	1	0.6310	6.683	96.8	1.3567	24 666	20.2	375.33	3.73	
371	Boston Beacon Hill	201	-71.0422	42.2144	50.0	50.0	6.53876	0	18.10	1	0.6310	7.016	97.5	1.2024	24 666	20.2	392.05	2.96	
373	Boston Beacon Hill	203	-71.0397	42.2182	50.0	50.0	8.26725	0	18.10	1	0.6680	5.875	89.6	1.1296	24 666	20.2	347.88	8.88	

Boston Housing (波士顿房屋)

- 一氧化氮浓度 X_5
  - 变量 X_5 与 X_{14} 的散点图

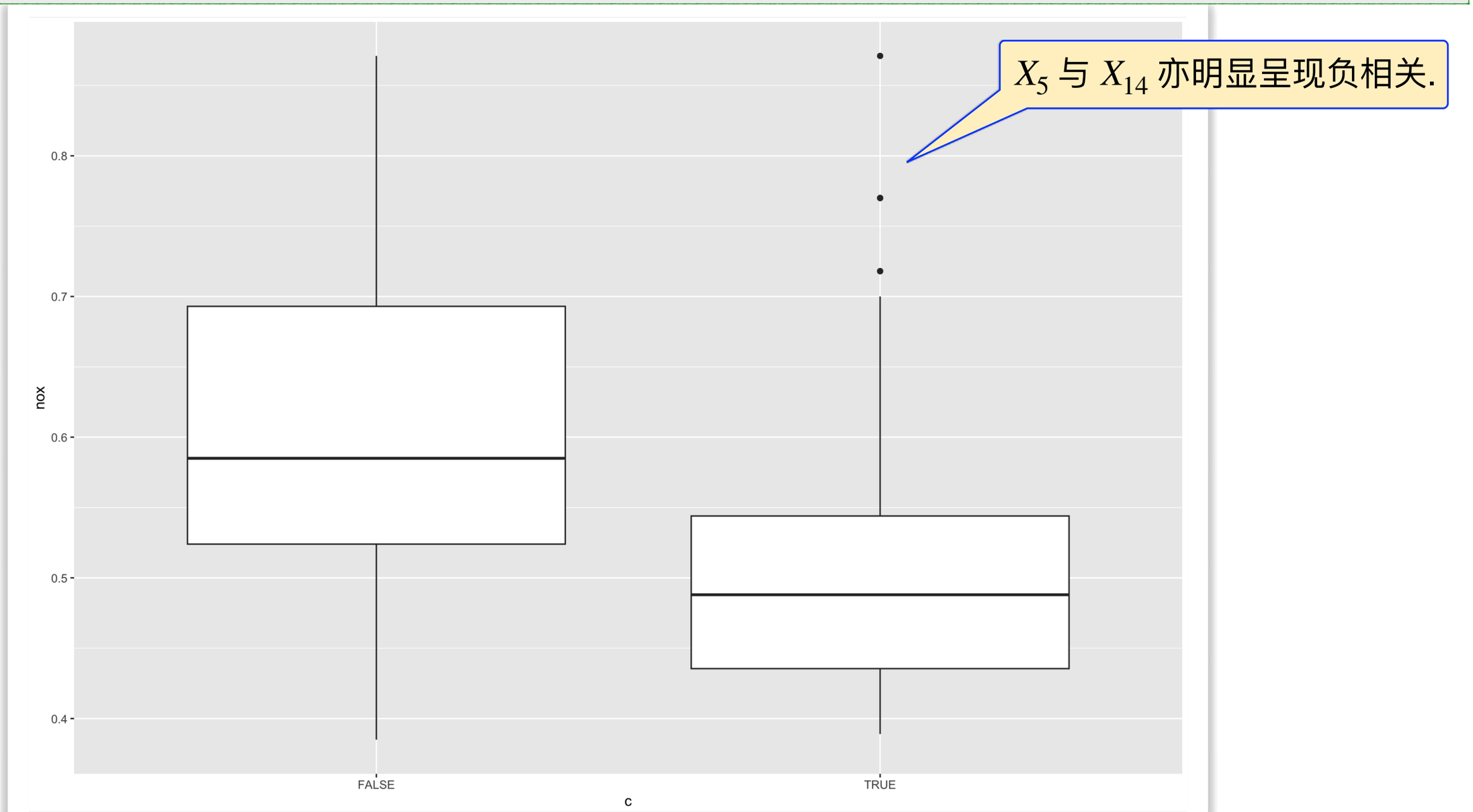
```
ggplot(boston, aes(x = nox, y = medv, color = c)) + geom_point()
```



Boston Housing (波士顿房屋)

- 一氧化氮浓度 X_5
  - 变量 X_5 依房价的分类箱线图

```
ggplot(data = boston, aes(x = c, y = nox)) + geom_boxplot(notch = FALSE)
```



Boston Housing (波士顿房屋)

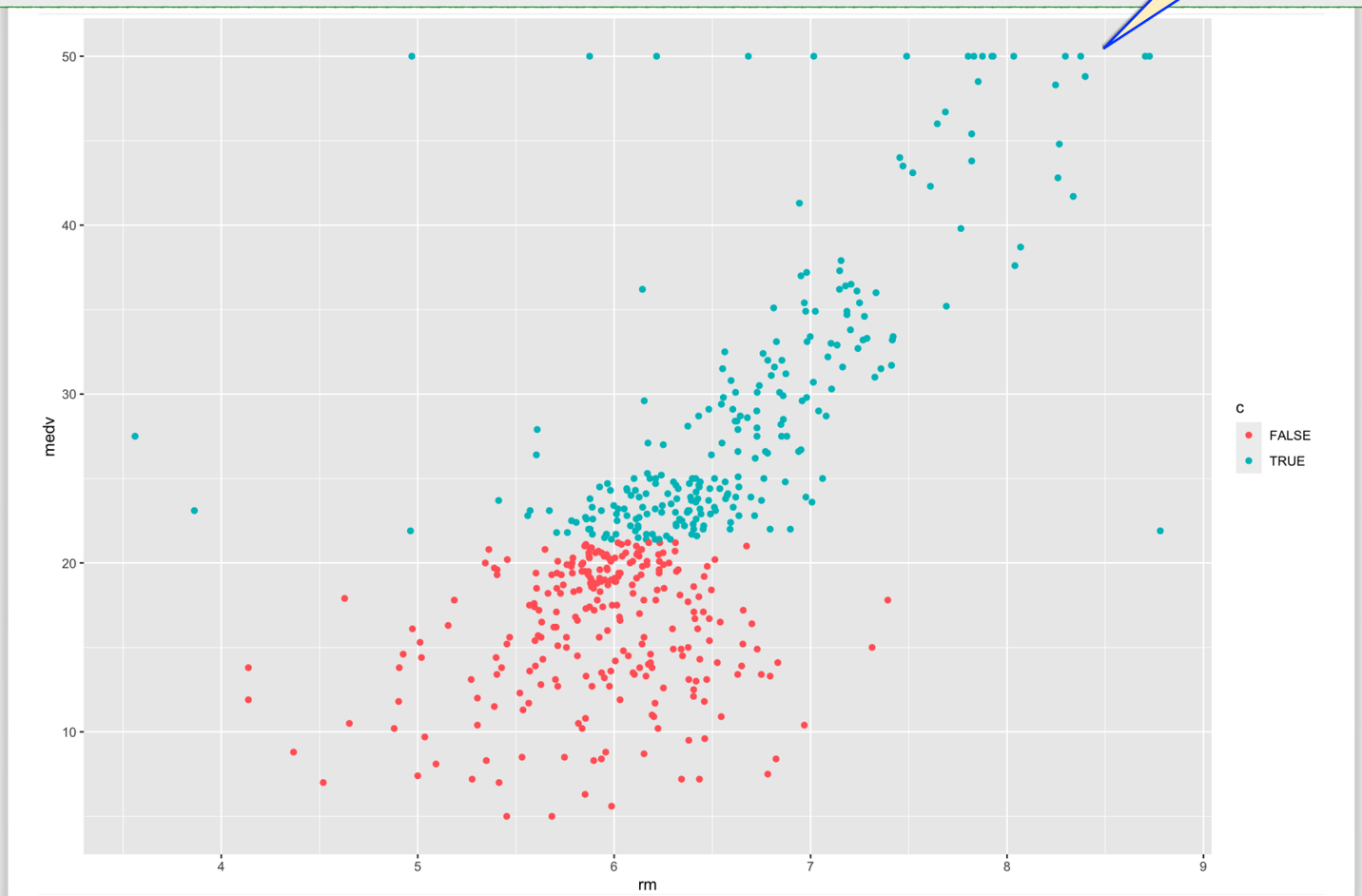
- 一氧化氮浓度 X_5
  - 因为研究的主要目的是分析污染对房价的影响，所以在考虑变量 X_5 能否作为房价 X_{14} 的一个解释变量时应特别小心.
  - 变量 X_5 不能作为解释变量的一种可能是人们不喜欢生活在氮氧化物排放量高的区域. 一氧化氮主要由汽车、工厂和住宅供暖的排放所致.
  - 然而，我们可以想到的是，除一氧化氮之外，还有许多其它合理的原因不选择住在城区或工业区. 例如，噪声污染可能是影响房价的一个更好的解释变量.
  - 由于氮氧化物的排放通常伴随着噪声污染，使用 X_5 作为 X_{14} 的一个解释变量就可能导致我们作出错误的结论：即人们选择远离氮氧化物，而实际上他们想远离的是噪声污染.

Boston Housing (波士顿房屋)

- 住宅的平均房间数 X_6
  - 住宅的房间数量是衡量房屋大小的一个可能的指标，所以，我们会预期 X_6 与 X_{14} 强相关.
  - 变量 X_6 与 X_{14} 的散点图.

X_6 与 X_{14} 明显呈正相关.

```
ggplot(boston, aes(x = rm, y = medv, color = c)) + geom_point()
```

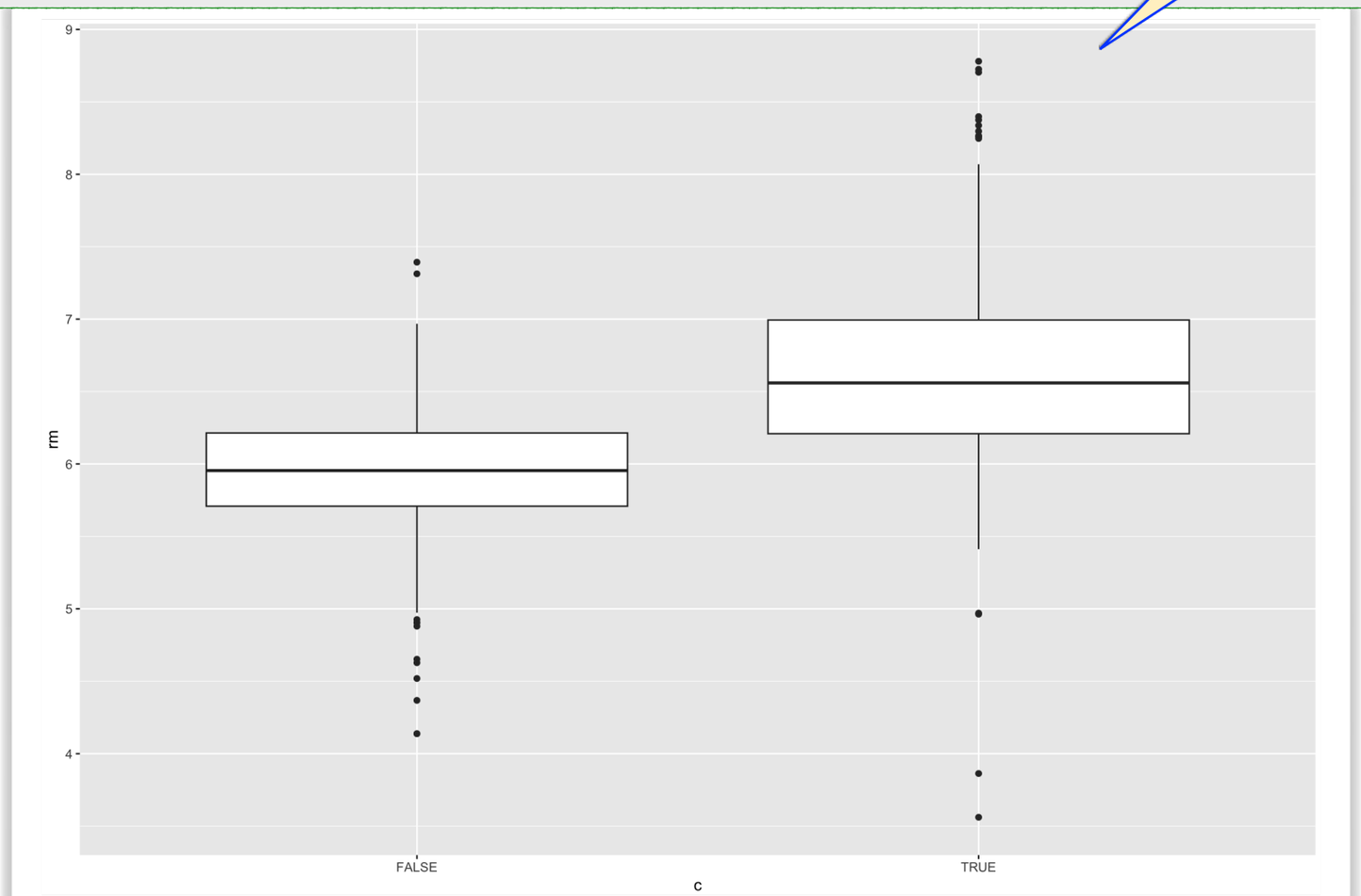


Boston Housing (波士顿房屋)

- 住宅的平均房间数 X_6
  - 住宅的房间数量是衡量房屋大小的一个可能的指标，所以，我们会预期 X_6 与 X_{14} 强相关.
  - 变量 X_6 依房价的分类箱线图.

X_6 与 X_{14} 明显呈正相关.

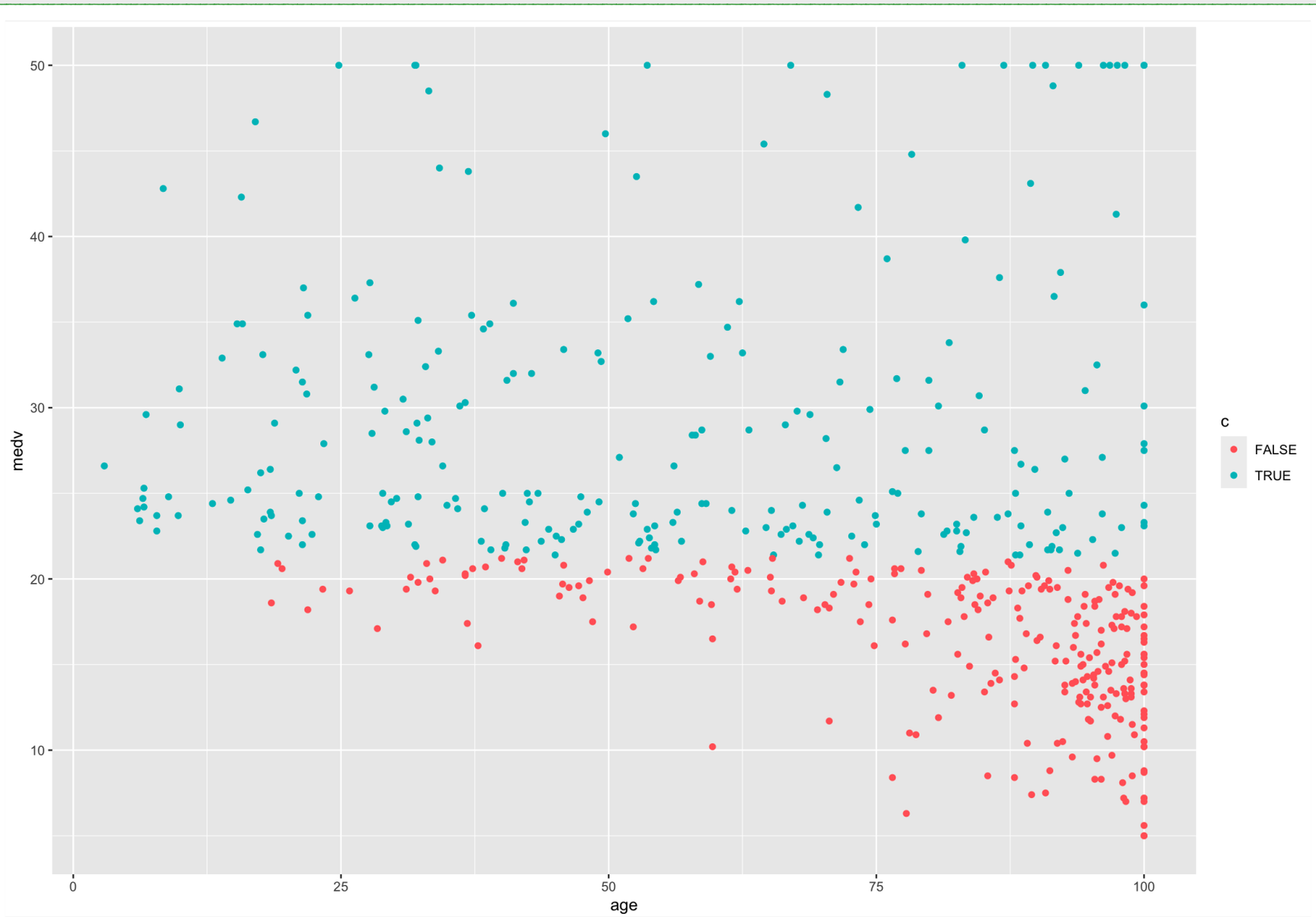
```
ggplot(data = boston, aes(x = c, y = rm)) + geom_boxplot(notch = FALSE)
```



Boston Housing (波士顿房屋)

- 1940年之前所建住房的比例 X_7
  - 变量 X_7 与 X_{14} 的散点图如下，无很清晰的关系可见.

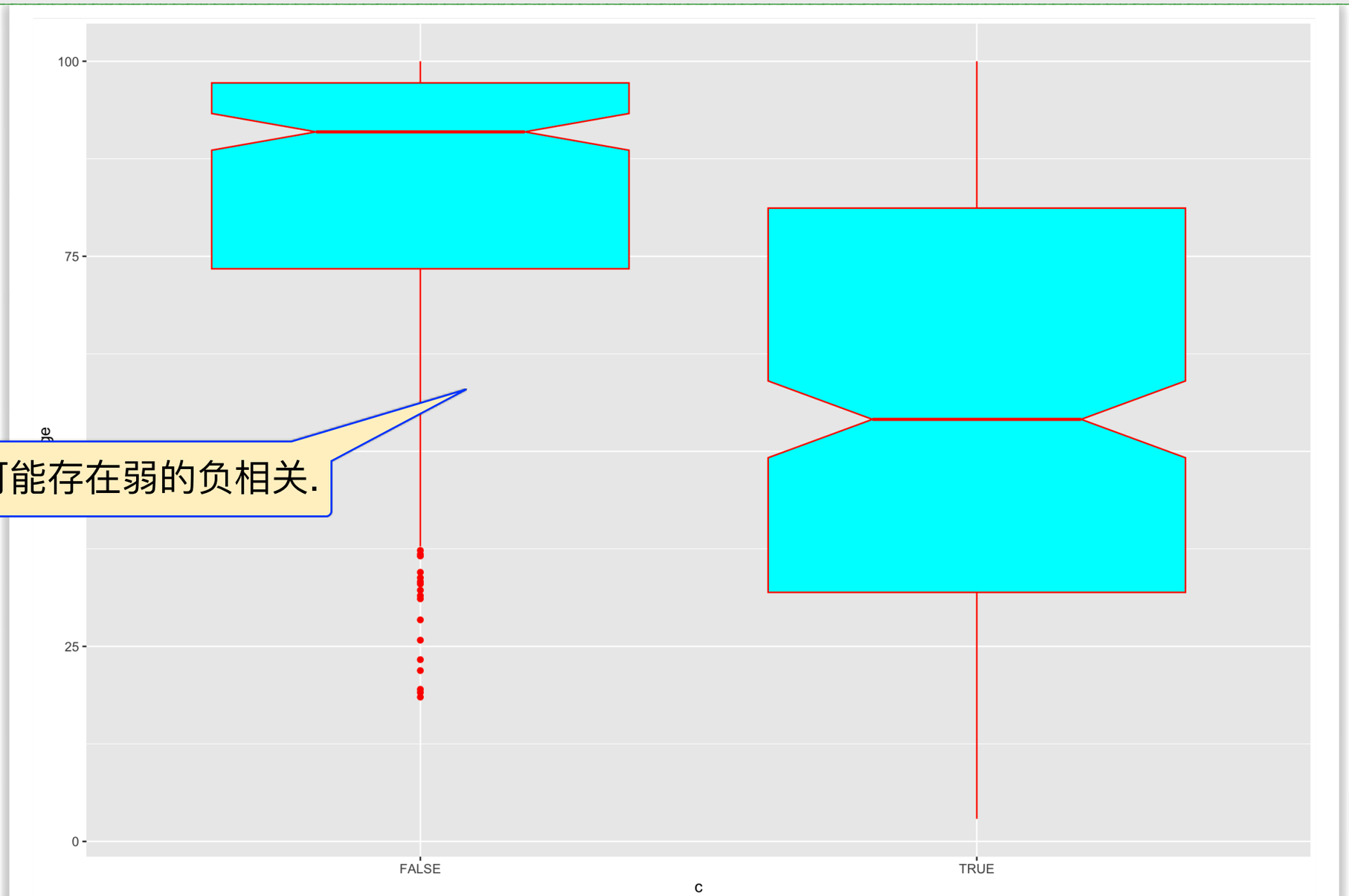
```
ggplot(boston, aes(x = age, y = medv, color = c)) + geom_point()
```



Boston Housing (波士顿房屋)

- 1940年之前所建住房的比例 X_7
  - 变量 X_7 依 X_{14} 的分类箱线图如下.

```
ggplot(data = boston, aes(x = c, y = age)) + geom_boxplot(notch = TRUE, fill = "pink1", colour = "blue")
```



X_7 与 X_{14} 可能存在弱的负相关.

Boston Housing (波士顿房屋)

- 1940年之前所建住房的比例 X_7
  - 相关性不那么清楚的现象可能是因以下两种相反效果所致.

    ① 一方面，如果老房状况不好，则房价应该低.

    ② 另一方面，人们因喜欢其空间大、设计传统，房价反而会高.
  - 无论如何，房龄影响房屋价格 X_{14} 似乎是合理的.

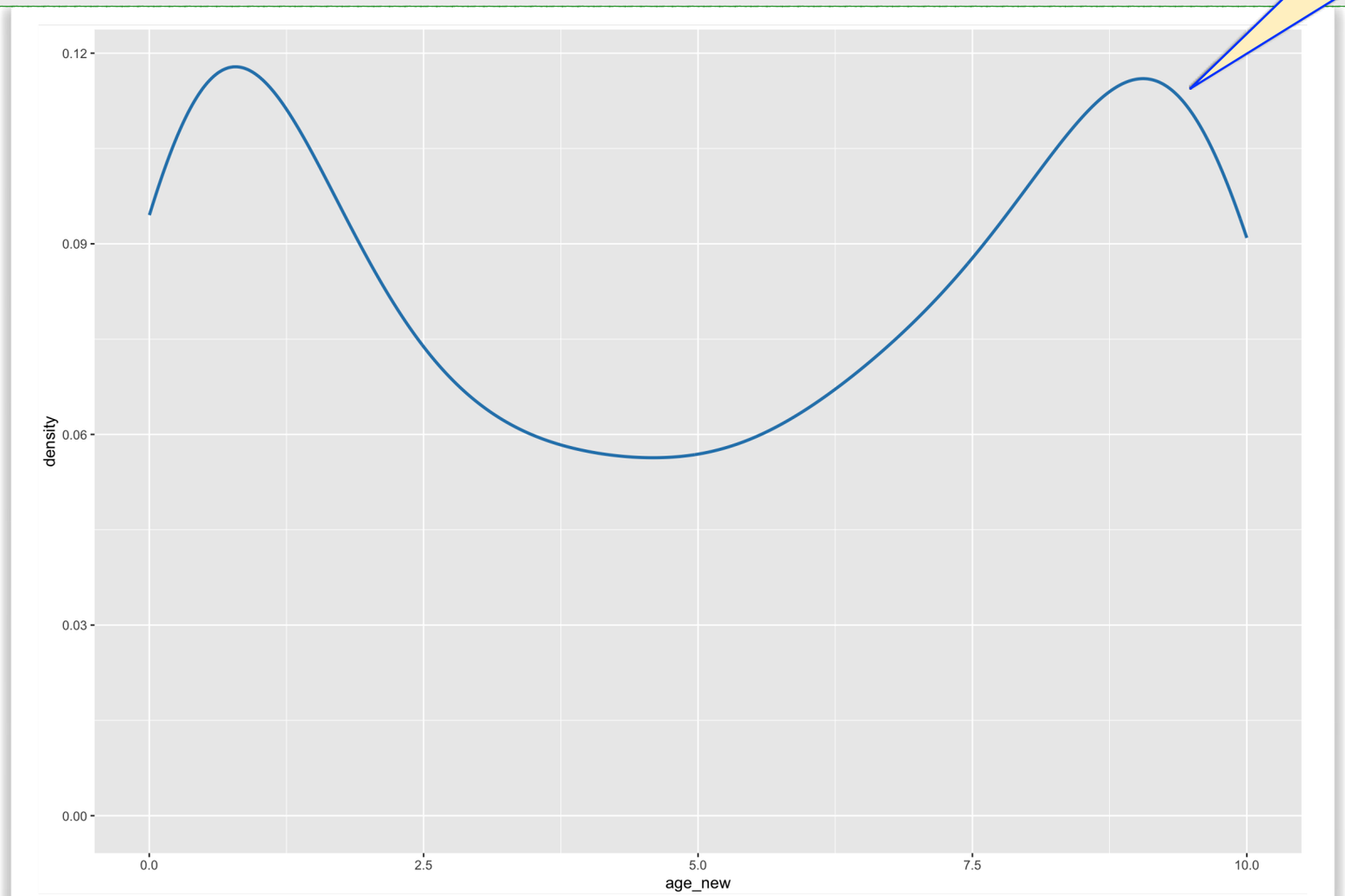
Boston Housing (波士顿房屋)

- 1940年之前所建住房的比例 X_7
  - 变量 X_7 的变换：$\widetilde{X}_7 = \dfrac{x_7^{2.5}}{10000}$. 变换后的核密度图如下.

此时双峰分别对
应房价高、低两组的现象
并不明显.

$\widetilde{X}_7$ 明显呈双峰.

```
Y7 = Boston$age^(2.5) / 10000
boston_sub = cbind(boston_sub, age_new = Y7)
ggplot(boston_sub, aes(age_new)) + geom_density(linewidth = 1, colour = 'steelblue')
```

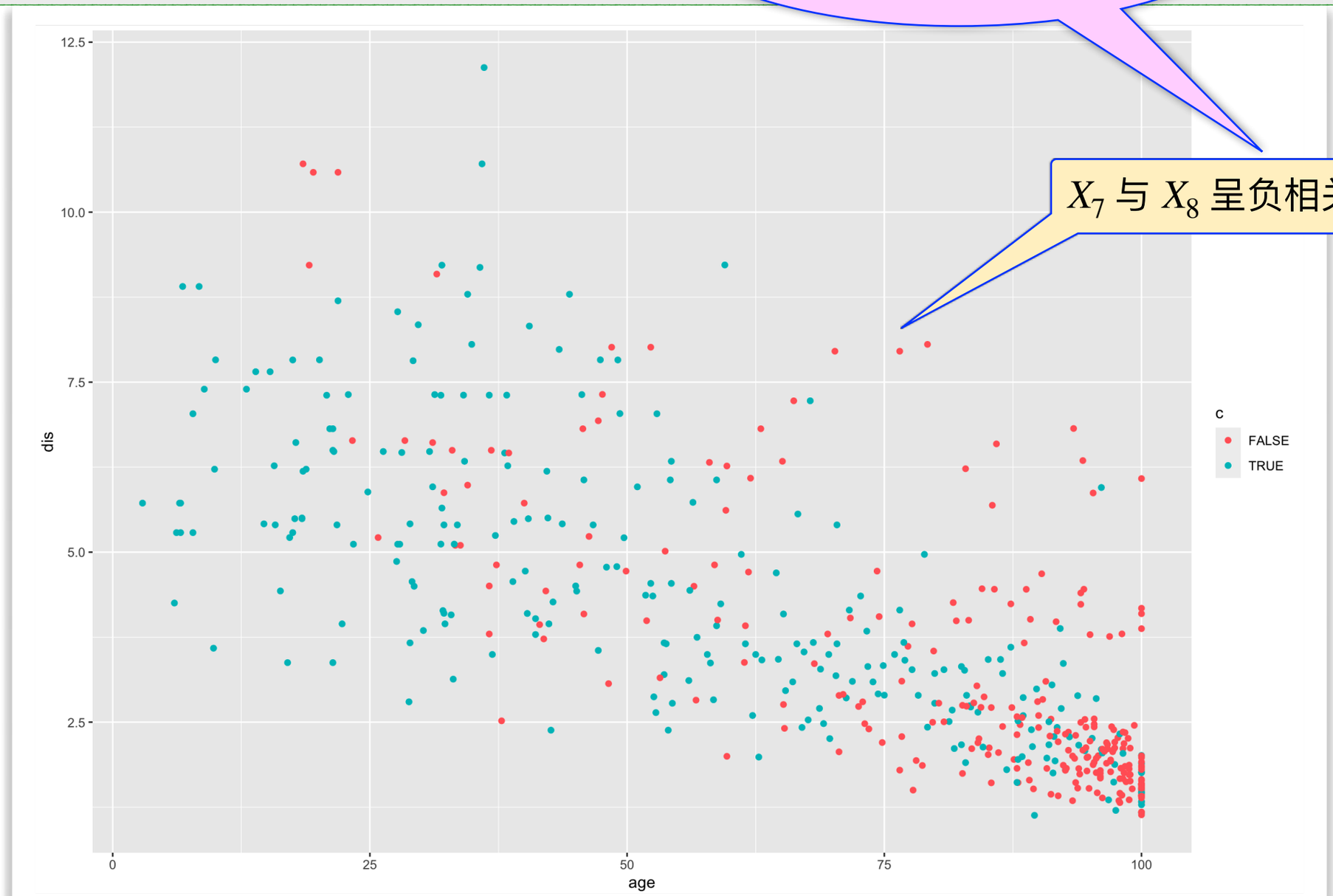


Boston Housing (波士顿房屋)

- 1940年之前所建住房的比例 X_7
  - 进一步考察变量 X_7 与 X_8 的散点图.

可能反映了波士顿市区随时间推移的发展方式；房龄低的区域离就业中心和工业设施更远.

```
ggplot(boston, aes(x = age, y = dis, color = c)) + geom_point()
```

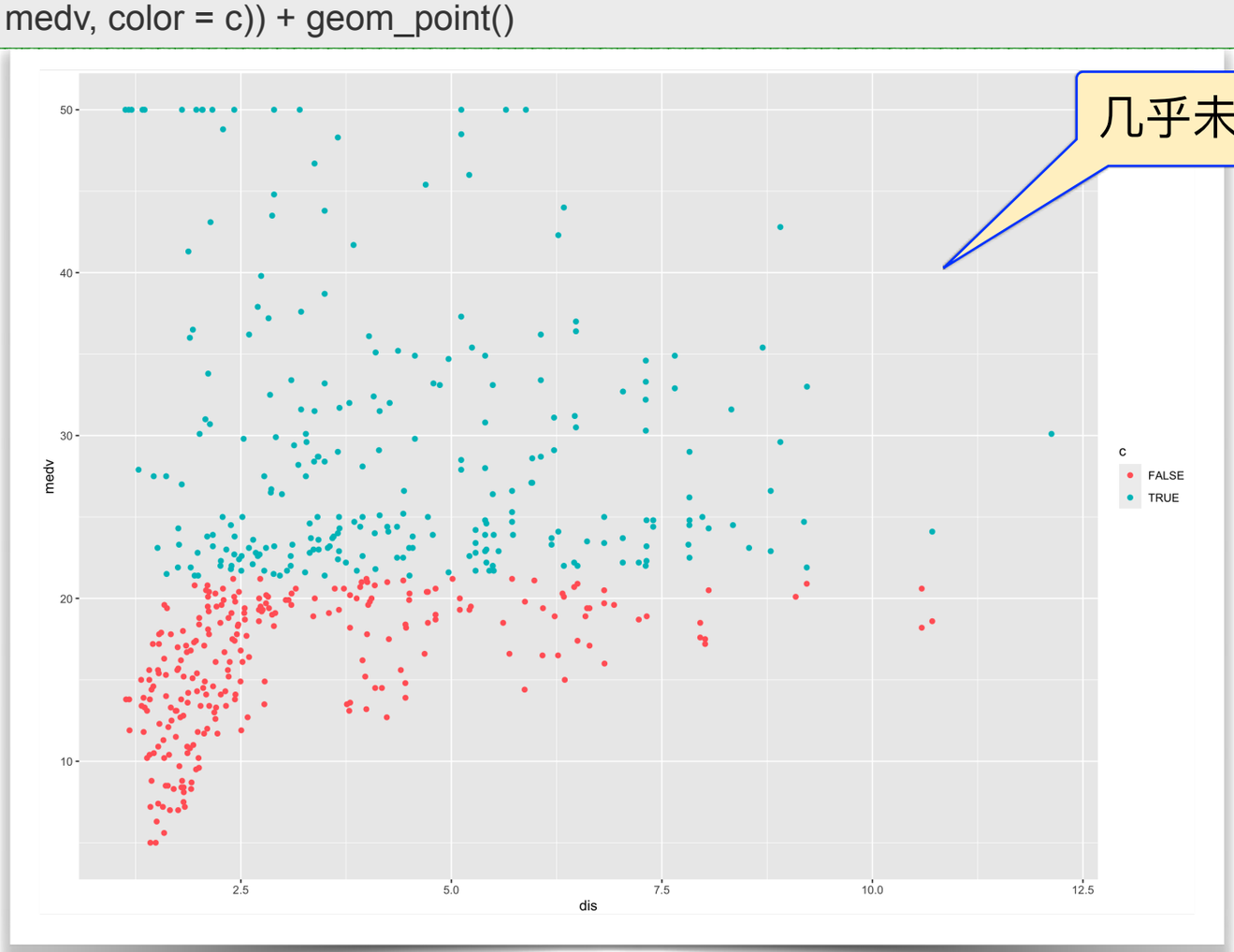


X_7 与 X_8 呈负相关.

Boston Housing (波士顿房屋)

- 至波士顿五个就业中心的加权距离 X_8
  - 由于大多数人喜欢住在离工作地点近的地方，我们预期房屋至就业中心的距离与房价之间存在负相关.
  - 变量 X_8 与 X_{14} 的散点图.

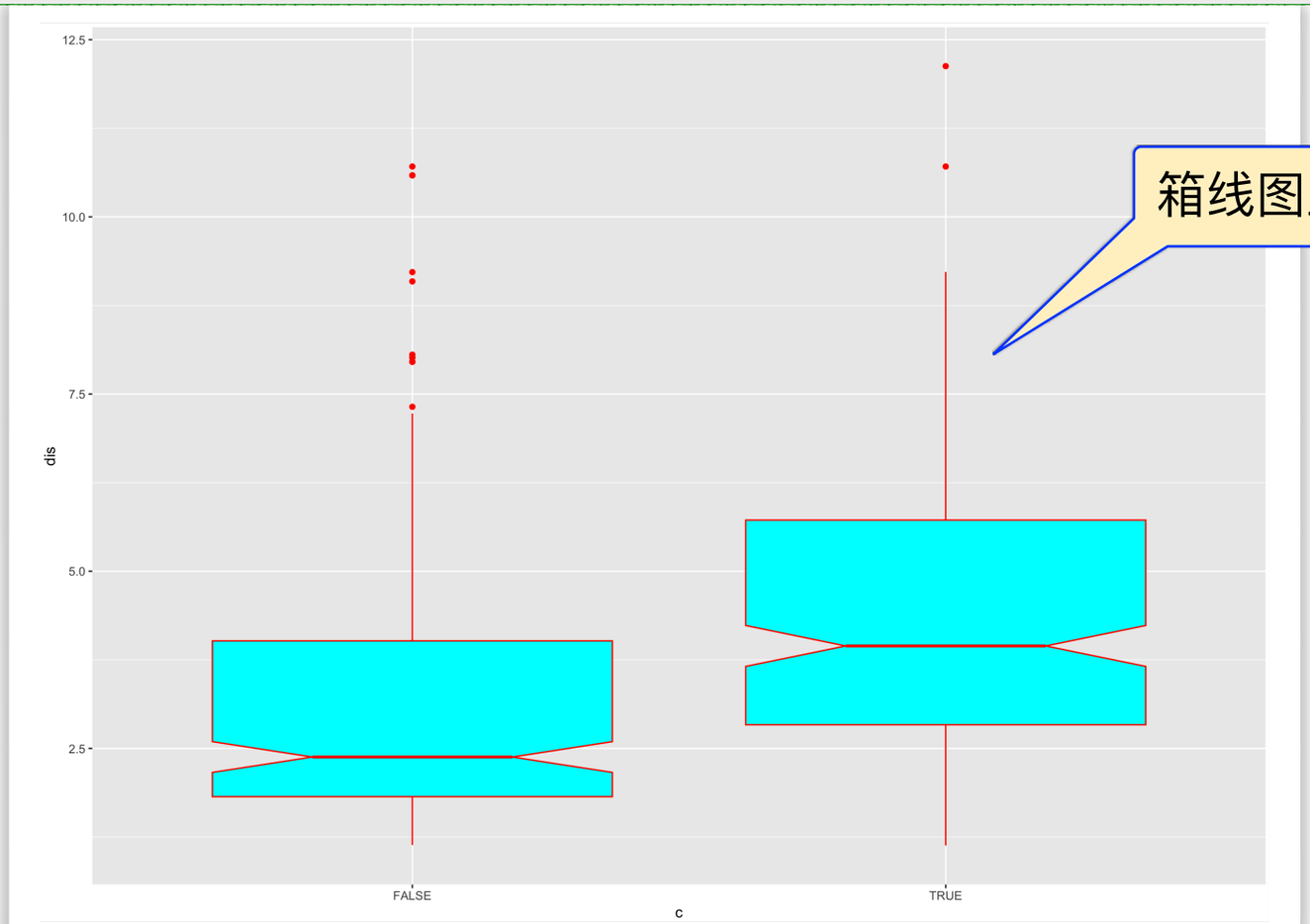
```
ggplot(boston, aes(x = dis, y = medv, color = c)) + geom_point()
```



Boston Housing (波士顿房屋)

- 至波士顿五个就业中心的加权距离 X_8
  - 由于大多数人喜欢住在离工作地点近的地方，我们预期房屋至就业中心的距离与房价之间存在负相关.
  - 变量 X_8 依 X_{14} 的分类箱线图.

```
ggplot(data = boston, aes(x = c, y = dis)) + geom_boxplot(notch = TRUE, fill = "cyan", colour = "orangered2")
```



Boston Housing (波士顿房屋)

- 至波士顿五个就业中心的加权距离 X_8
  - 由于大多数人喜欢住在离工作地点近的地方，我们预期房屋至就业中心的距离与房价之间存在负相关.
  - 同样，这里可能也有两个相反的效应在起作用.

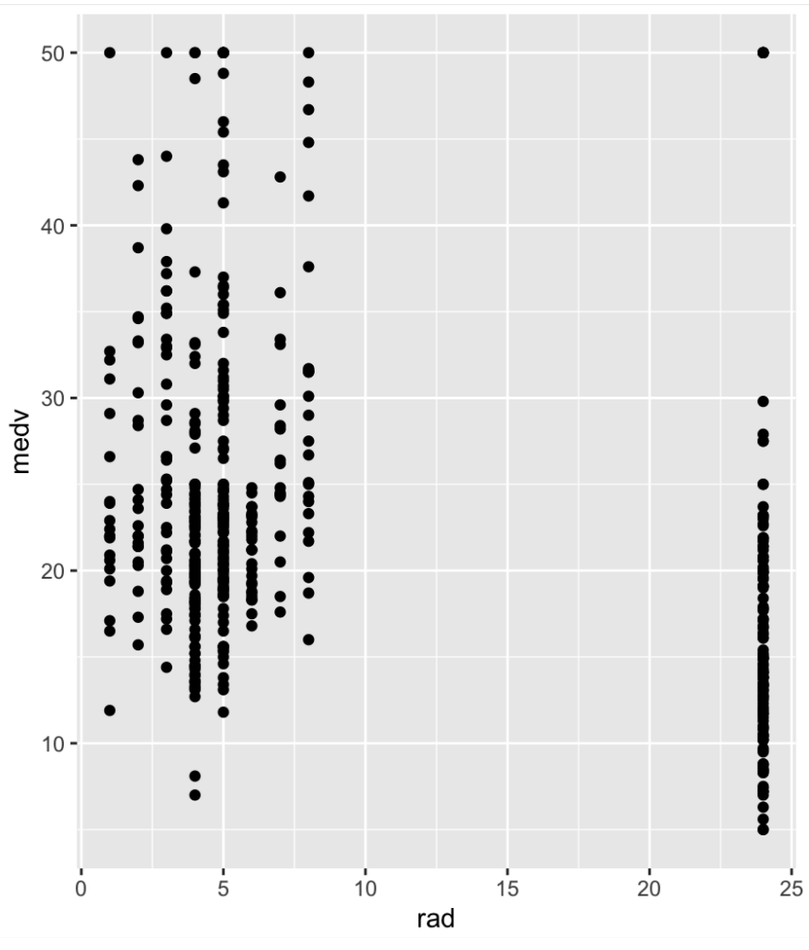
    ① 一方面，靠近就业中心居住可能污染会严重一些.

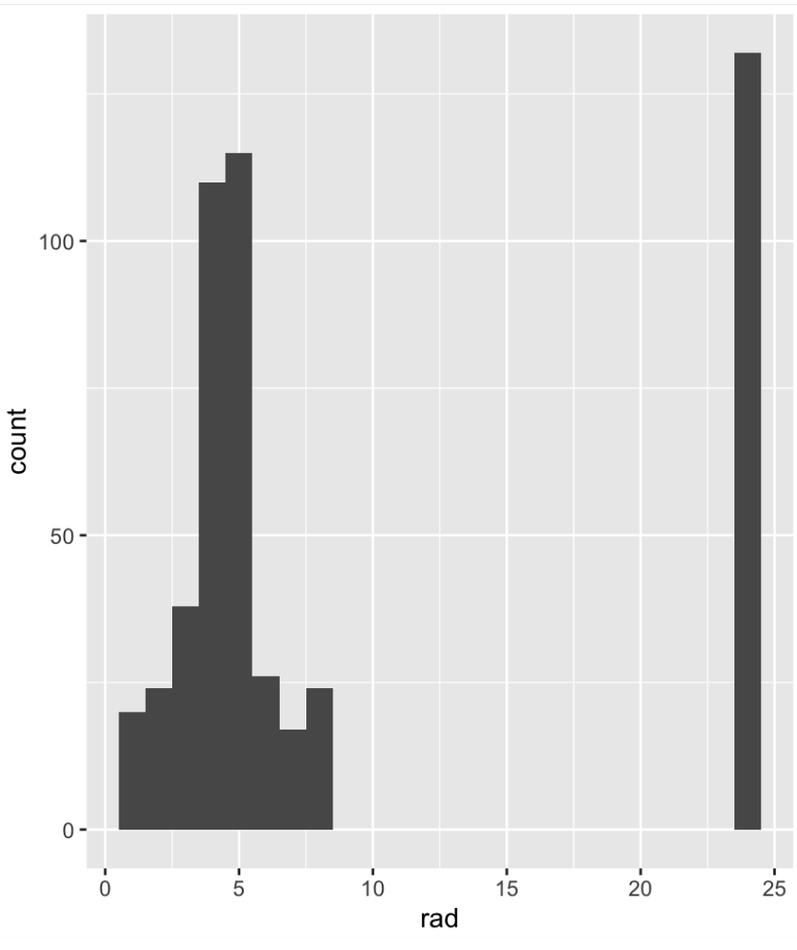
    ② 另一方面，上班距离近.

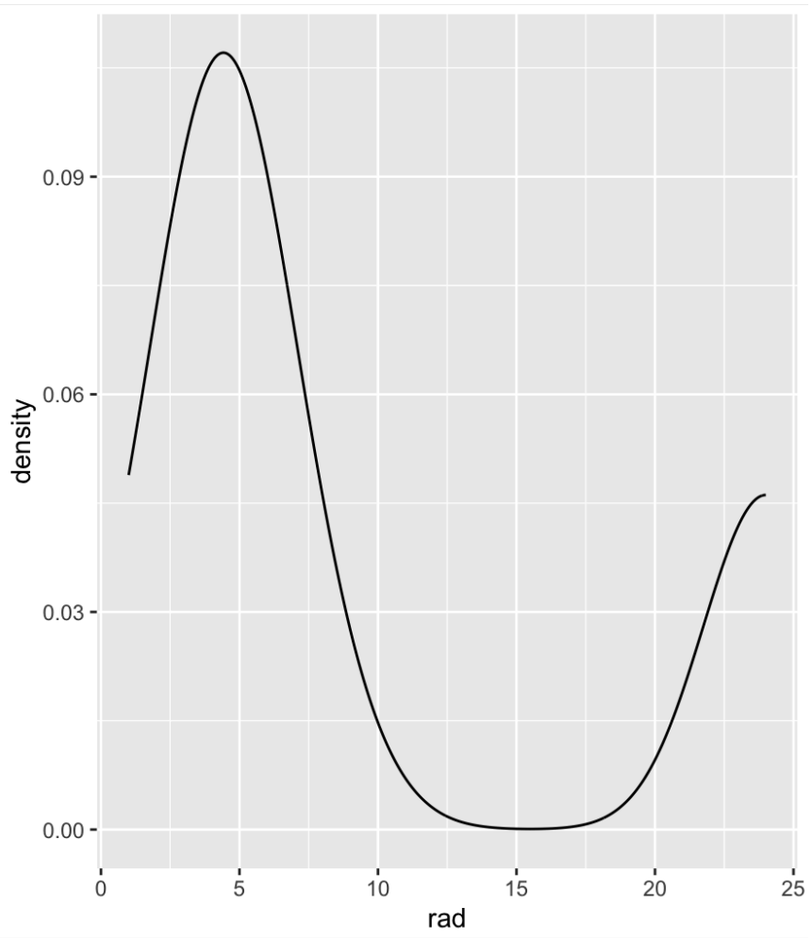
Boston Housing (波士顿房屋)

- 辐射状高速公路的可达性指数 X_9
  - 变量 X_9 的散点图、直方图、核密度图均呈双峰.

```
p9_1 = ggplot(boston, aes(x = rad, y = medv, color = c)) + geom_point()
p9_2 = ggplot(boston, aes(x = rad)) + geom_histogram(binwidth = 1)
p9_3 = ggplot(boston, aes(x = rad)) + geom_density()
grid.arrange(p9_1, p9_2, p9_3, ncol = 3)
```



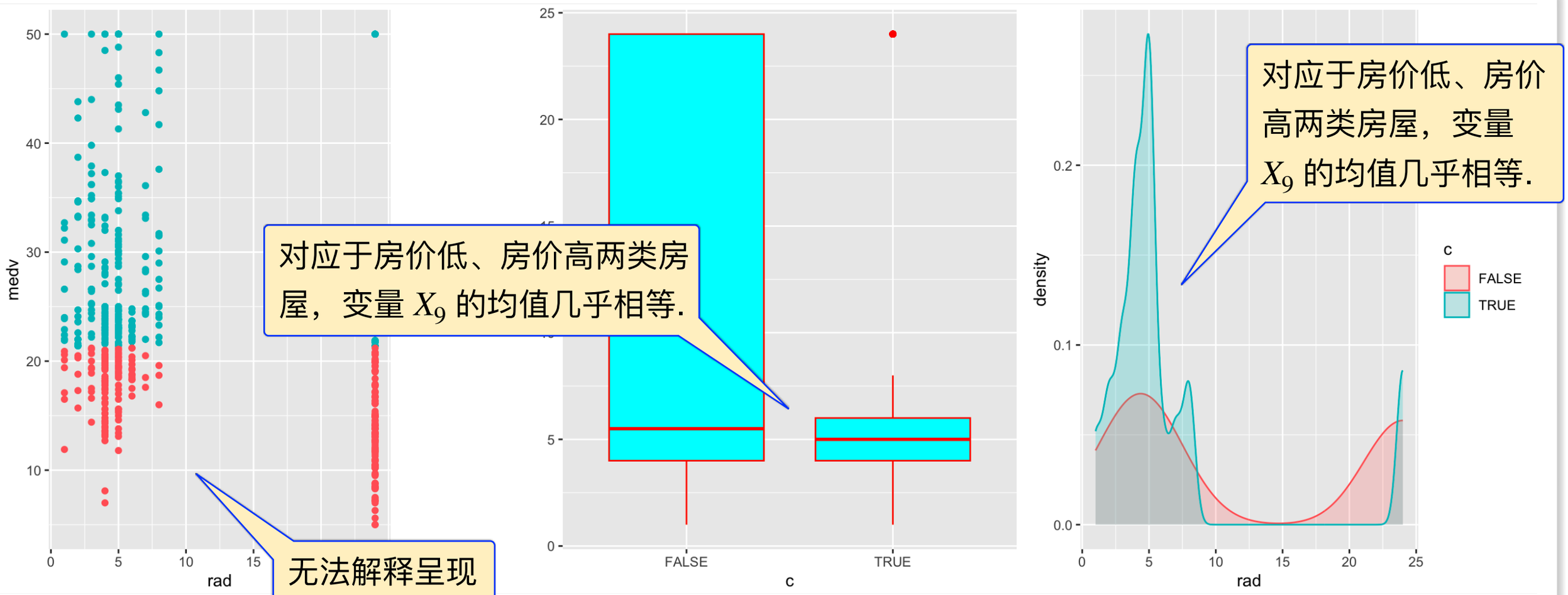




Boston Housing (波士顿房屋)

- 辐射状高速公路的可达性指数 X_9
  - 变量 X_9 依 X_{14} 的散点图、箱线图、核密度图.

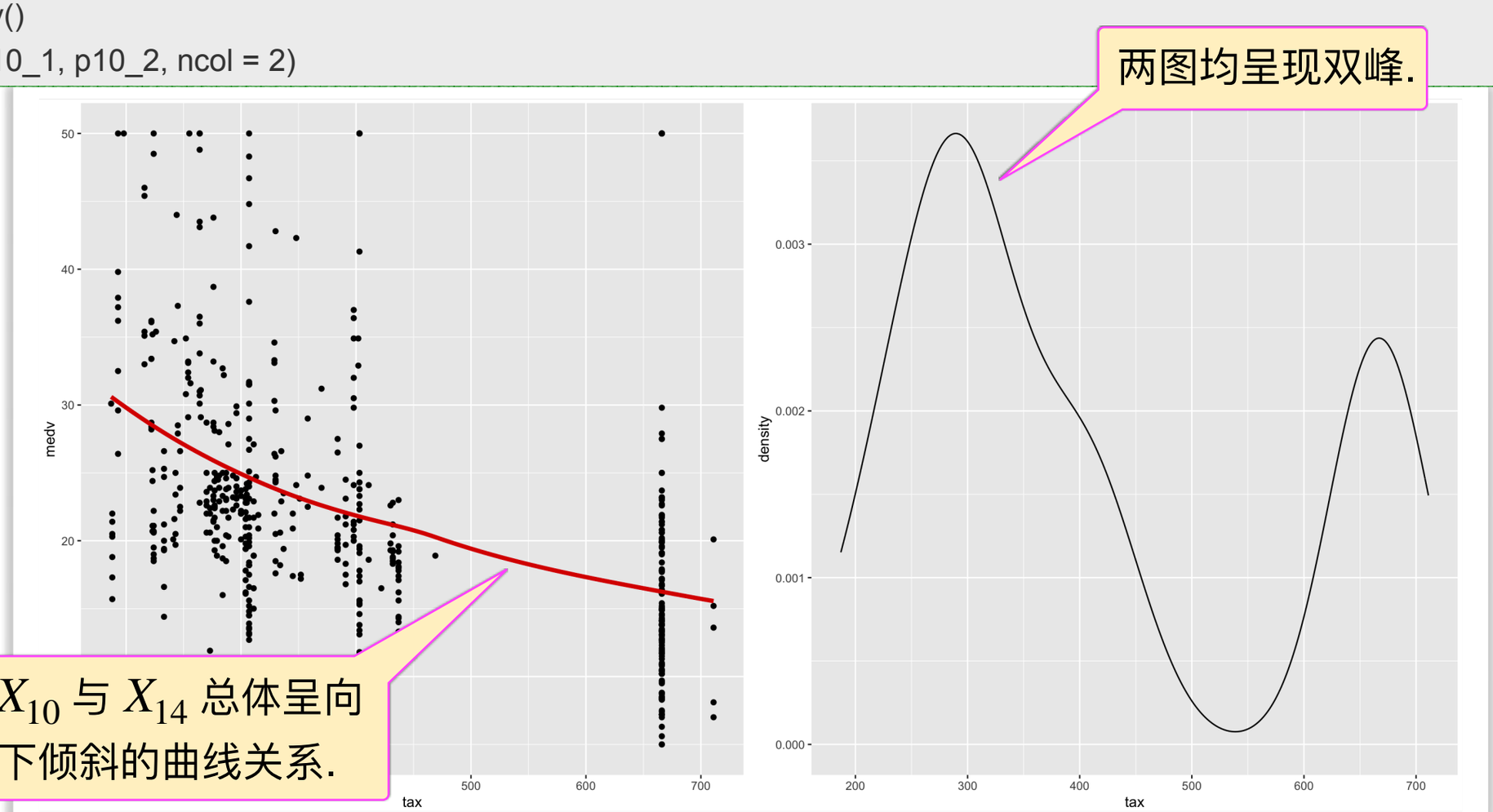
```
p9_4 = ggplot(boston, aes(x = rad, fill = c)) + geom_histogram(binwidth = 1) + facet_wrap(~ c, ncol = 1)
p9_5 = ggplot(boston, aes(x = c, y = rad)) + geom_boxplot(notch = FALSE, fill = "cyan", colour = "orangered2")
grid.arrange(p9_1, p9_4, p9_5, ncol = 3)
```



Boston Housing (波士顿房屋)

- 每一万美元的全额财产税税率 X_{10}
  - 变量 X_{10} 与 X_{14} 的散点图，X_{10} 的核密度图.

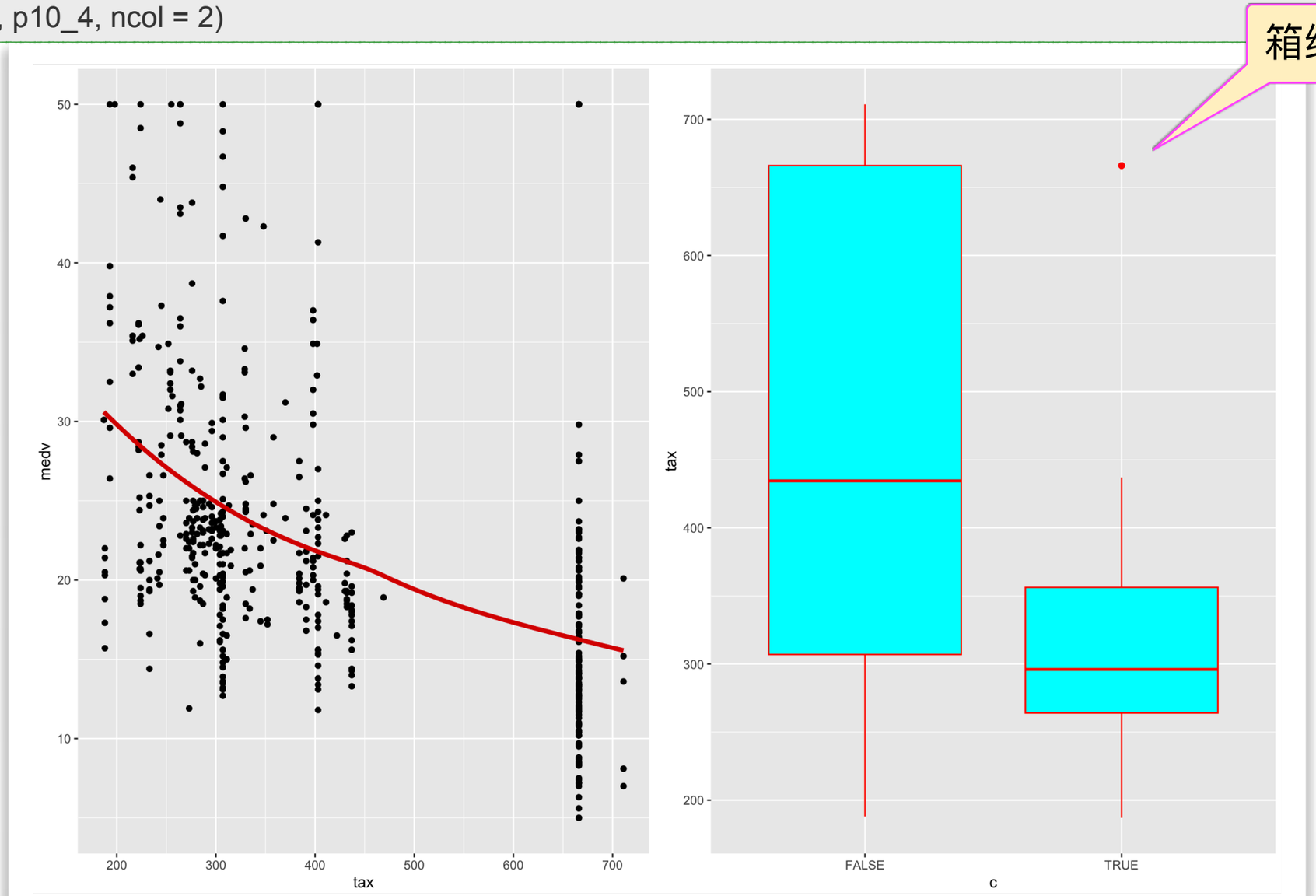
```
p10_1 = ggplot(boston, aes(x = tax, y = medv)) +
  geom_point() +
  geom_smooth(method = 'loess', se = FALSE, span = 0.85, linewidth = 1.5, colour = 'red3')
p10_2 = ggplot(boston, aes(x = tax)) +
  geom_density()
grid.arrange(p10_1, p10_2, ncol = 2)
```



Boston Housing (波士顿房屋)

- 每一万美元的全额财产税税率 X_{10}
  - 变量 X_{10} 关于 X_{14} 的散点图，X_{10} 的箱线图.

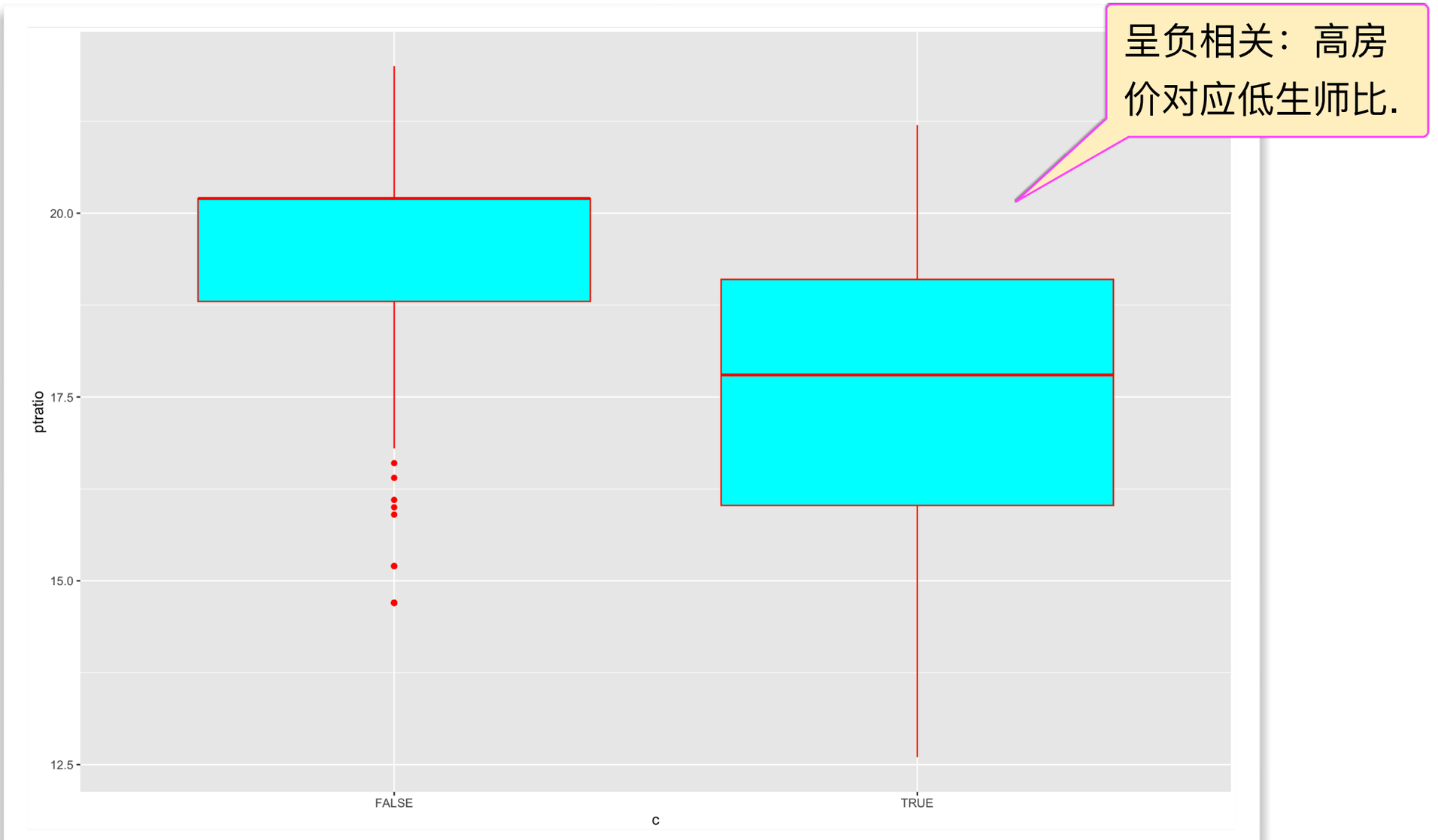
```
p10_3 = ggplot(boston, aes(x = c, y = tax)) +
  geom_boxplot(notch = FALSE, fill = "cyan", colour = "orangered2")
grid.arrange(p10_3, p10_4, ncol = 2)
```



Boston Housing (波士顿房屋)

- 生师比 X_{11}
  - 变量 X_{11} 依 X_{14} 的分类箱线图.

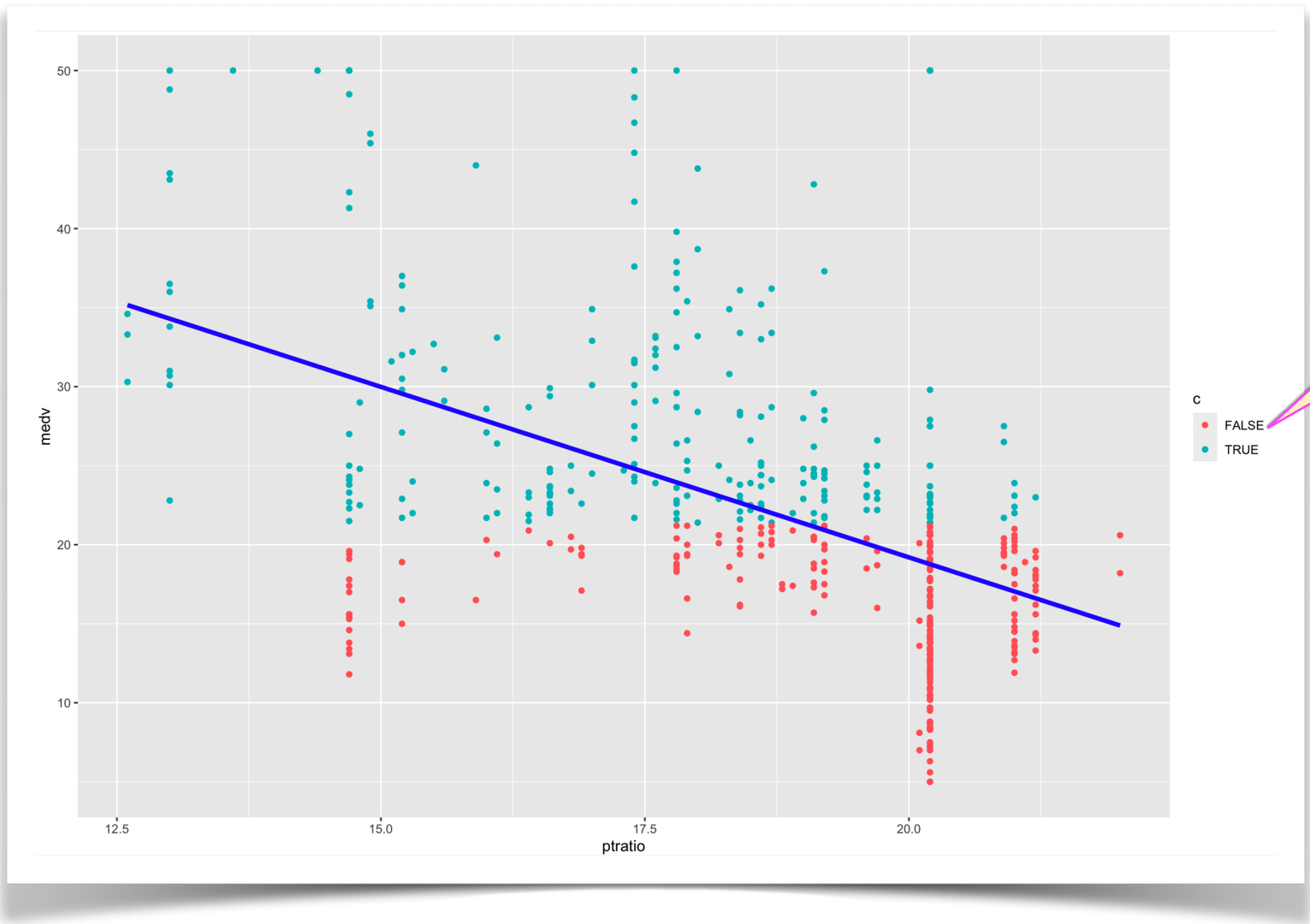
```
p11_1 = ggplot(boston, aes(x = c, y = ptratio)) + geom_boxplot(notch = FALSE, fill = "cyan", colour = “orangered2")
p11_1
```



Boston Housing (波士顿房屋)

- 生师比 X_{11}
  - 变量 X_{11} 与 X_{14} 的分类散点图.

```
p11_2 = ggplot(boston, aes(x = ptratio, y = medv, colour = c)) +
  geom_point() +
  geom_smooth(method = 'lm', linewidth = 1.5, colour = 'blue', se = FALSE)
p11_2
```



呈负相关：高房价对应低生师比.

Boston Housing (波士顿房屋)

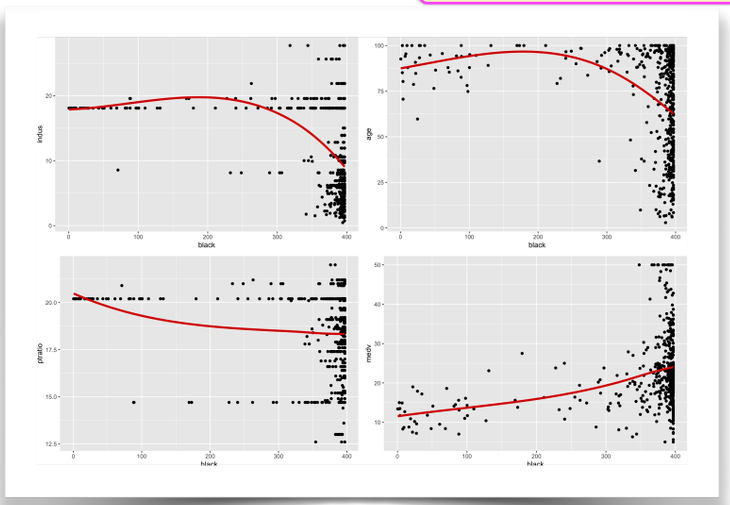
- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 变量 X_{12} 与 X_3, X_7, X_{11}, X_{14} 的散点图.

$$= \begin{cases} 1000\left(B - 0.63\right)^2, & B < 0.63 \\ 0, & B \geq 0.63 \end{cases}$$

```
p12_3 = ggplot(boston, aes(x = black, y = indus)) +
  geom_point() +
  geom_smooth(method = 'loess', se = FALSE, span = 0.95, linewidth = 1.5, colour = 'red3')
p12_7 = ggplot(boston, aes(x = black, y = age)) +
  geom_point() +
  geom_smooth(method = 'loess', se = FALSE, span = 0.95, linewidth = 1.5
p12_11 = ggplot(boston, aes(x = black, y = ptratio)) +
  geom_point() +
  geom_smooth(method = 'loess', se = FALSE, span = 0.95, linewidth = 1.5, colour = 'red3')
p12_14 = ggplot(boston, aes(x = black, y = medv)) +
  geom_point() +
  geom_smooth(method = 'loess', se = FALSE, span = 0.95, linewidth =
grid.arrange(p12_3, p12_7, p12_11, p12_14, ncol = 2)
```

X_{12} 与 X_3, X_7, X_{11} 均呈非线性负相关.

X_{12} 与 X_{14} 呈正相关.



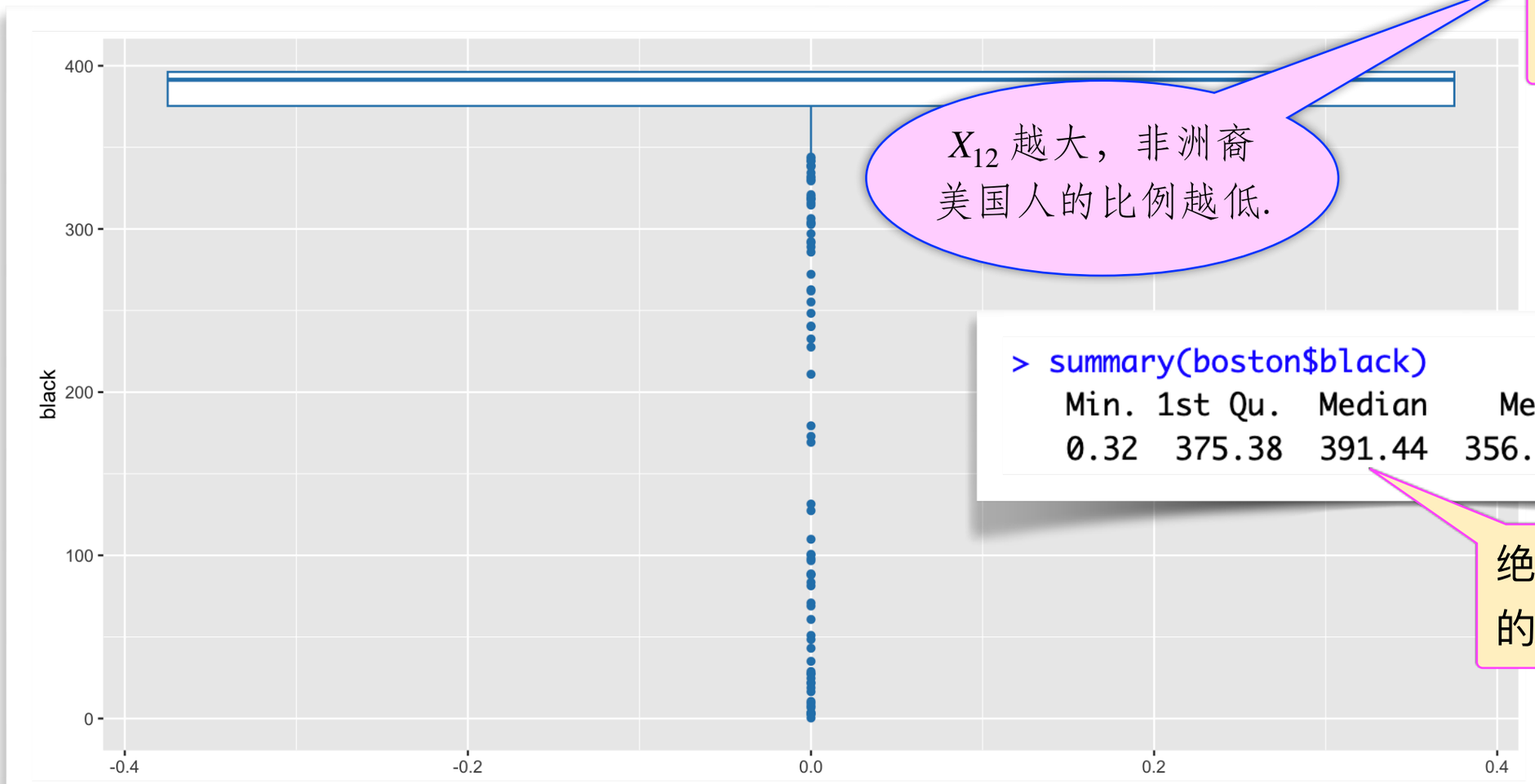
Boston Housing (波士顿房屋)

- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 变量 X_{12} 的箱线图、描述性统计量的值.

$$= \begin{cases} 1000\left(B - 0.63\right)^2, & B < 0.63 \\ 0, & B \geq 0.63 \end{cases}$$

因 $B < 0.63$，这意味着绝大部分住宅区域的 B 接近于 0.

```
ggplot(boston, aes(y = black)) +
  geom_boxplot(notch = FALSE, fill = "cyan", colour = "orangered2")
summary(boston$black)
```



X_{12} 越大，非洲裔美国人的比例越低.

```
> summary(boston$black)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.32  375.38  391.44  356.67  396.23  396.90
```

绝大部分住宅区域 X_{12} 的值位于390附近.

Boston Housing (波士顿房屋)

- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 变量 X_{12} 的箱线图、描述性统计量的值.
  - 变量 X_{12} 取值位于上、下四分位数之外的个数.

$$= \begin{cases} 1000\left(B - 0.63\right)^2, & B < 0.63 \\ 0, & B \geq 0.63 \end{cases}$$

因 $B < 0.63$，这意味着绝大部分住宅区域的 B 接近于 0.

```
dim(Boston)
sum(boston$black < 375)
sum(boston$black > 397)
```

```
> dim(boston)
[1] 506  15
> sum(boston$black < 375)
[1] 125
> sum(boston$black > 397)
[1] 0
```

X_{12} 越大，非洲裔美国人的比例越低.

- 观测数据 405 至 470 的 X_{12}

说明这些住宅区非裔美国人的比例大于0.

```
> boston$black[405:470]
 [1] 329.46 384.97 370.22 332.09 314.64 179.36   7.68  35.05  28.79 210.97  88.27  27.25  21.57
[14] 127.36  16.45  48.45 318.75 319.98 291.55   2.52   3.65   7.68  24.65  18.82  96.73  60.72
[27]  83.45  81.33  97.95 100.19 100.63 109.85  27.49   9.32  68.95 396.90 391.45 385.96 395.69
[40] 386.73 240.52  43.06 318.01 388.52 396.90 304.21   0.32 355.29 385.09 375.87   6.68  50.92
[53]  10.48   3.50 272.21 396.90 255.23 391.43 396.90 393.82 396.90 334.40  22.01 331.29 368.74
[66] 396.90
```

有相当多的 X_{12} 值远低于390.

```
> summary(boston$black)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.32  375.38  391.44  356.67  396.23  396.90
```

绝大部分住宅区域 X_{12} 的值位于390附近.

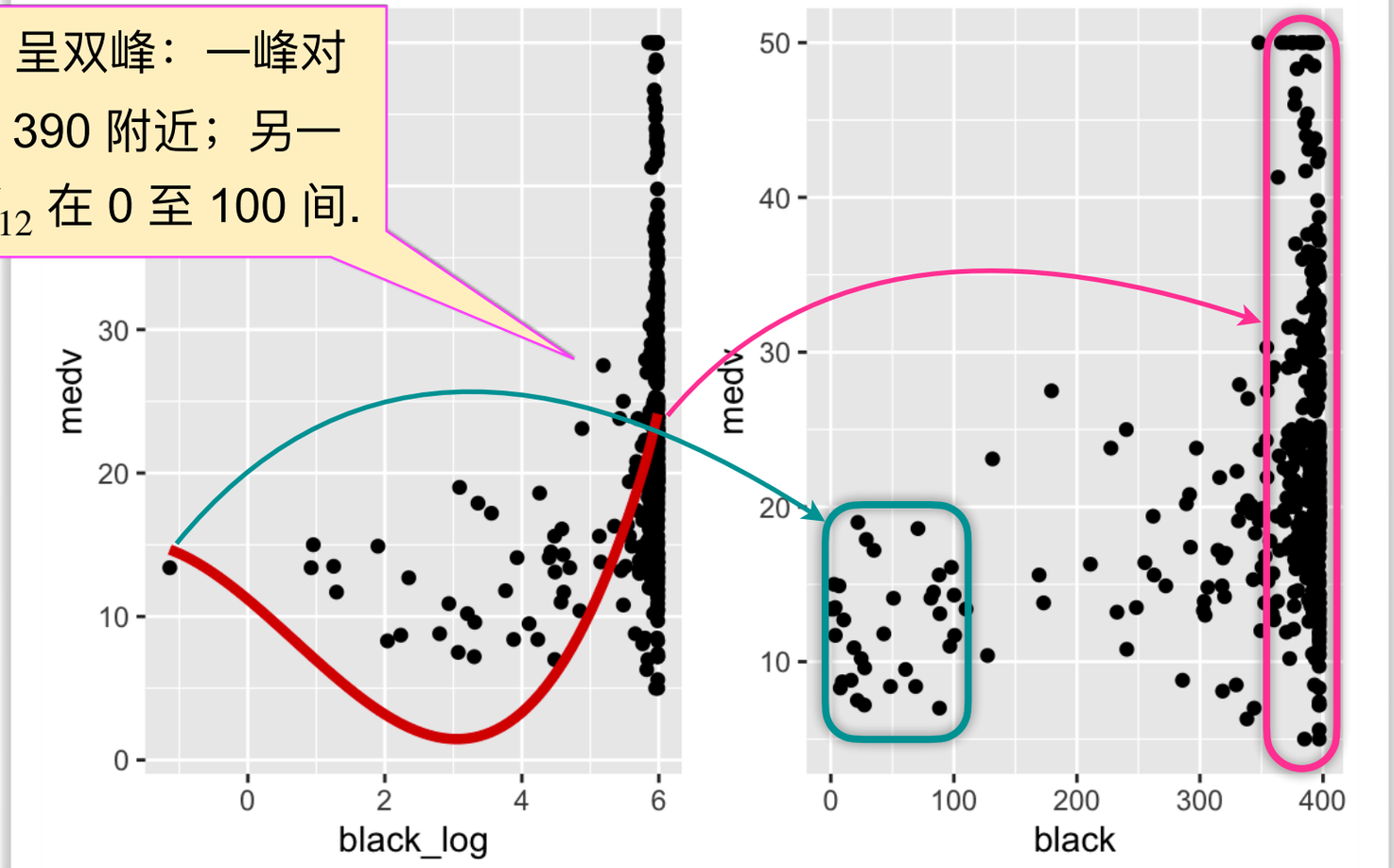
Boston Housing (波士顿房屋)

- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 变量 $\log\left(X_{12}\right)$ 的散点图.

$$= \begin{cases} 1000\left(B - 0.63\right)^2, & B < 0.63 \\ 0, & B \geq 0.63 \end{cases}$$

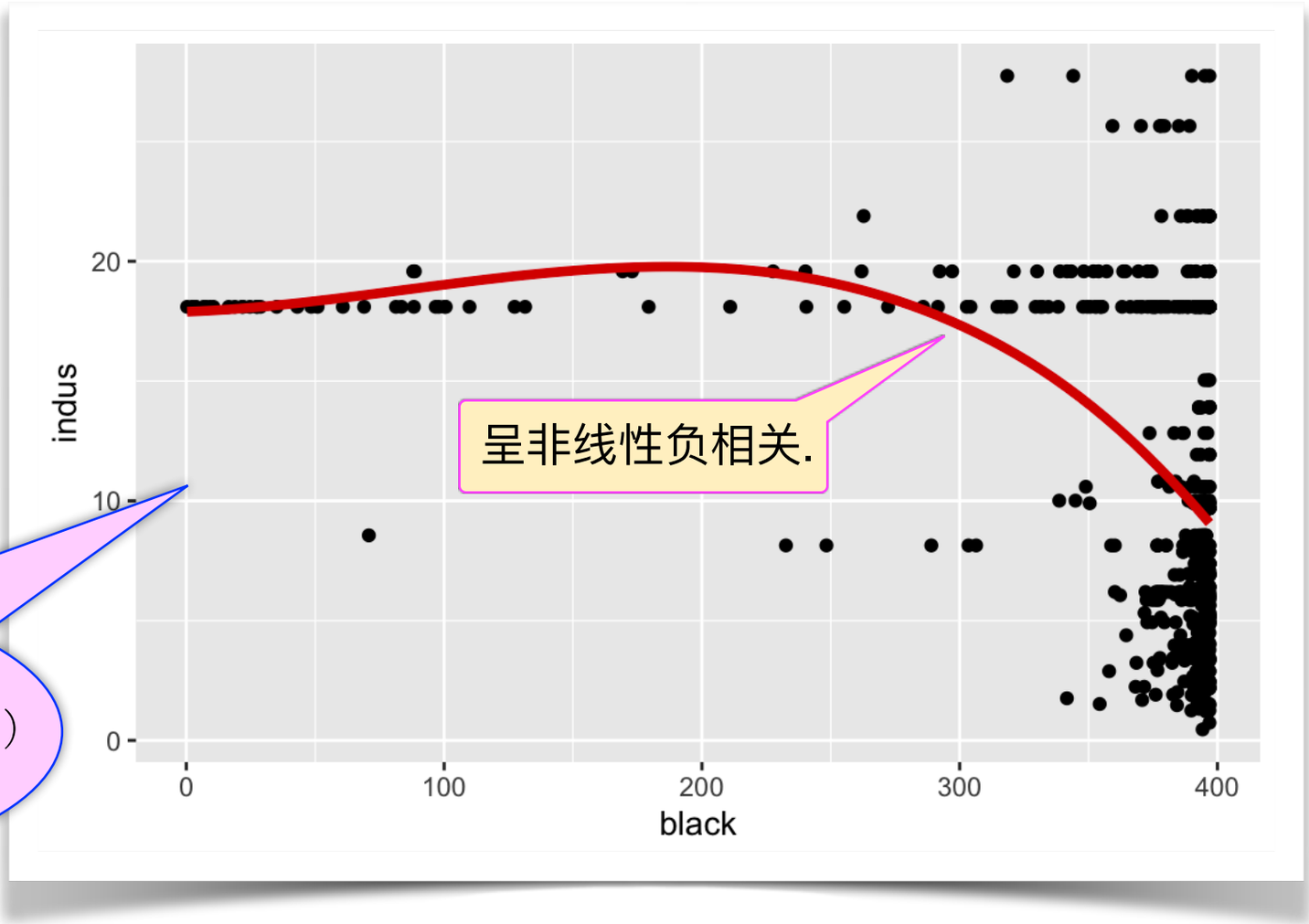
```
black_log = log(boston$black)
boston = cbind(boston, black_log = black_log)
p12_log = ggplot(boston, aes(x = black_log, y = medv)) + geom_point() +
  geom_smooth(method = 'loess', se = FALSE, span = 0.95, linewidth = 1.5, colour = ‘red3')
p12_medv = ggplot(boston, aes(x = black, y = medv)) + geom_point()
grid.arrange(p12_log, p12_medv, ncol = 2)
```

$\log\left(X_{12}\right)$ 呈双峰：一峰对应 X_{12} 在 390 附近；另一峰对应 X_{12} 在 0 至 100 间.



Boston Housing (波士顿房屋)

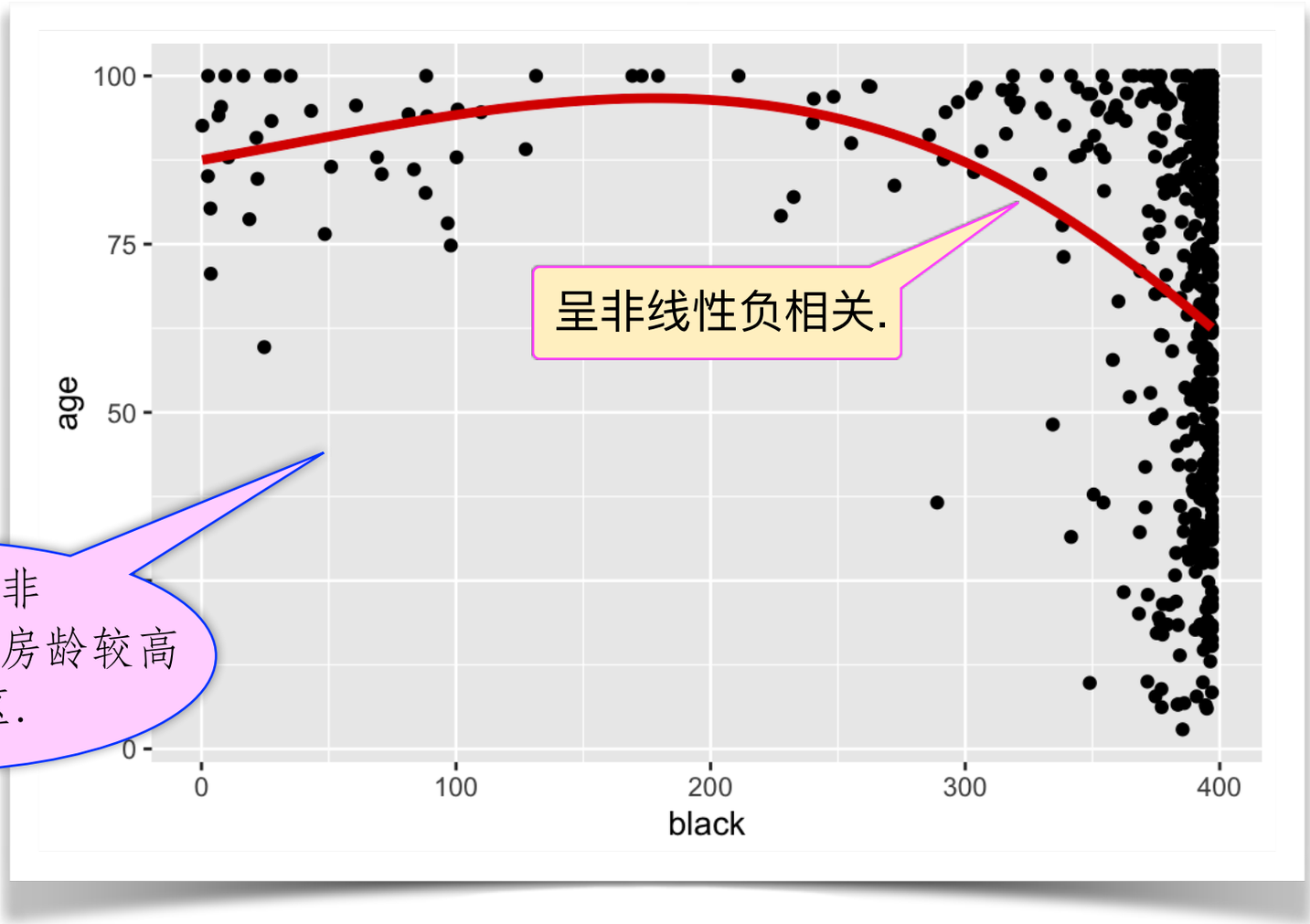
- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 当 X_{12} 与某个变量呈正相关时，则非裔美国人的实际比例与该变量呈负相关，反之亦然.
  - 变量 X_{12} 与 X_3 的散点图.



这意味着非裔美国人生活在非零售商业用地(工业区)比例很高的地区.

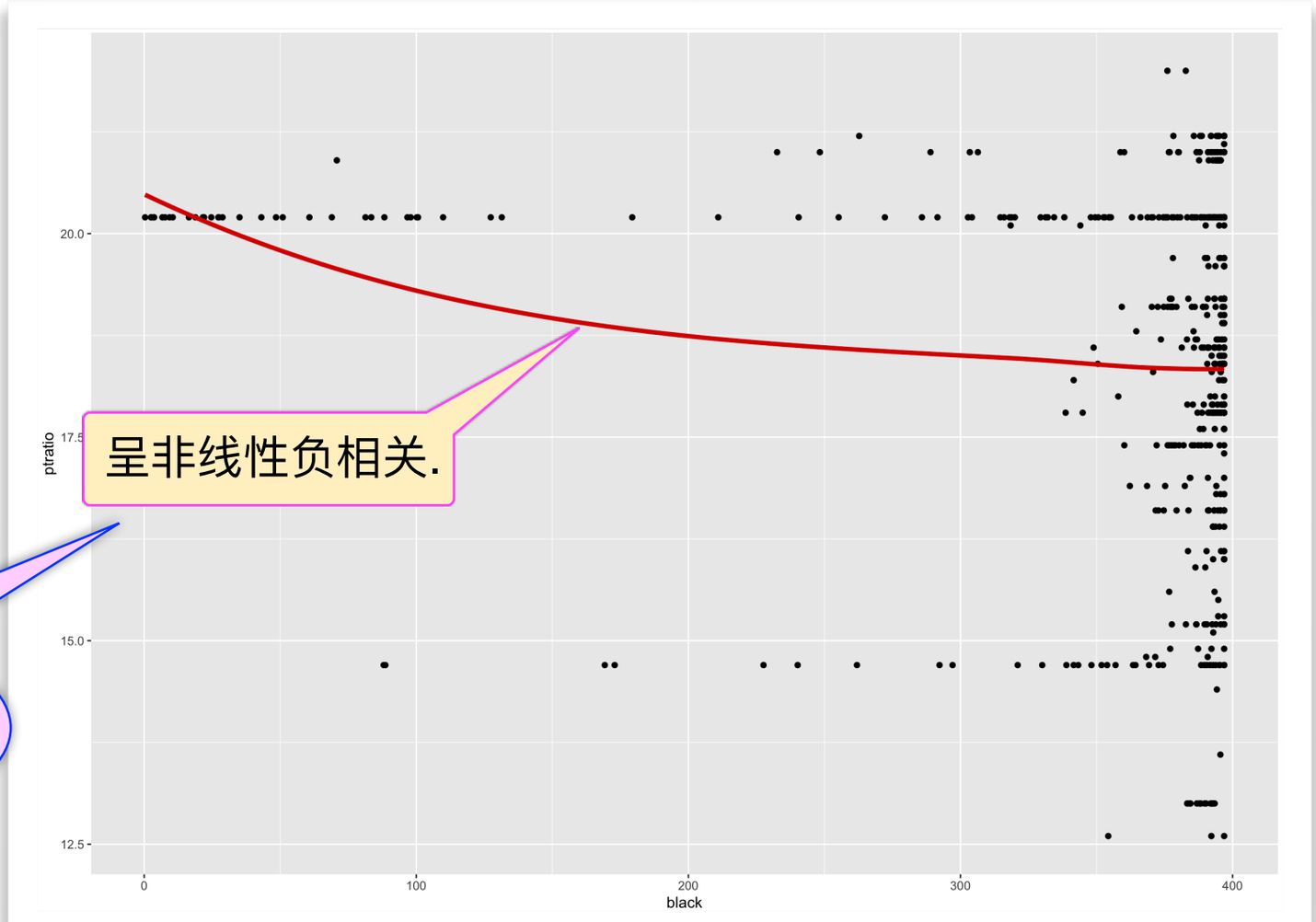
Boston Housing (波士顿房屋)

- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 当 X_{12} 与某个变量呈正相关时，则非裔美国人的实际比例与该变量呈负相关，反之亦然.
  - 变量 X_{12} 与 X_7 的散点图.



Boston Housing (波士顿房屋)

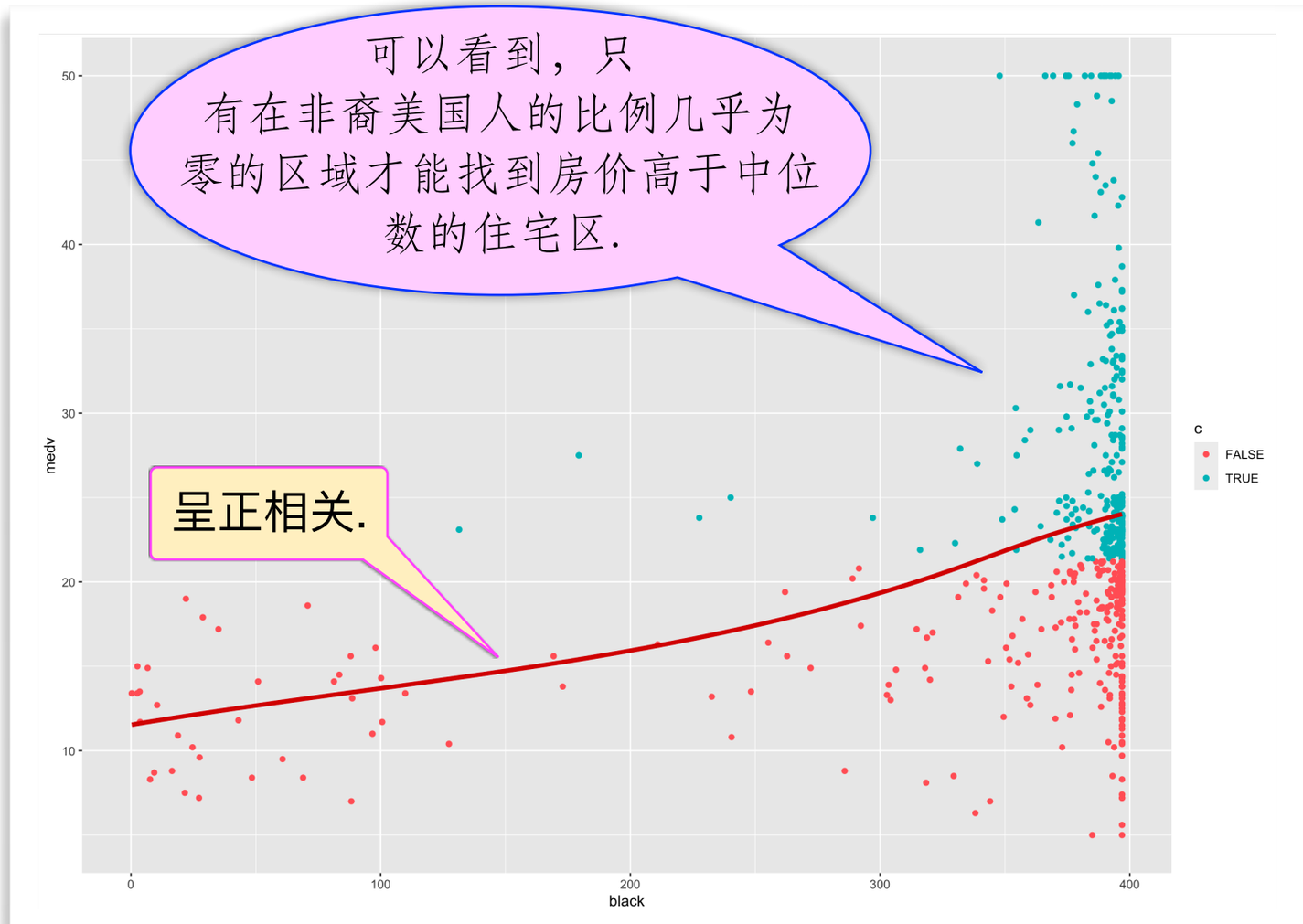
- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 当 X_{12} 与某个变量呈正相关时，则非裔美国人的实际比例与该变量呈负相关，反之亦然.
  - 变量 X_{12} 与 X_{11} 的散点图.



这意味着非裔美国人生活在生师比高的住宅区.

Boston Housing (波士顿房屋)

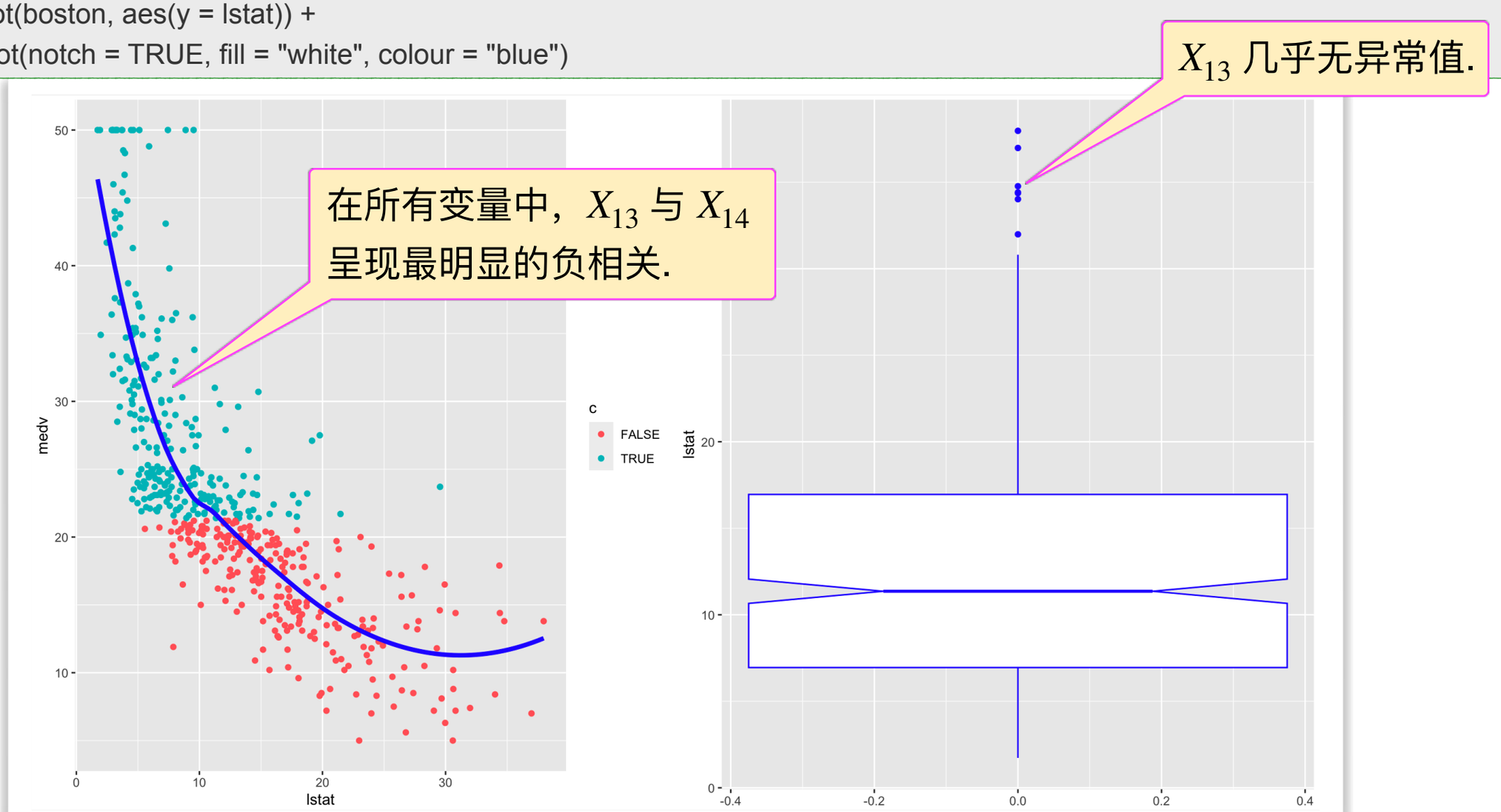
- 变量 $X_{12} = 1000\left(B - 0.63\right)^2 \times \mathrm{I}(B < 0.63)$，其中 B 是非洲裔美国人的比例.
  - 当 X_{12} 与某个变量呈正相关时，则非裔美国人的实际比例与该变量呈负相关，反之亦然.
  - 变量 X_{12} 与 X_{14} 的散点图.



Boston Housing (波士顿房屋)

- 社会低层人口的百分比 X_{13}
  - 变量 X_{13} 与 X_{14} 的散点图、X_{13} 的箱线图.

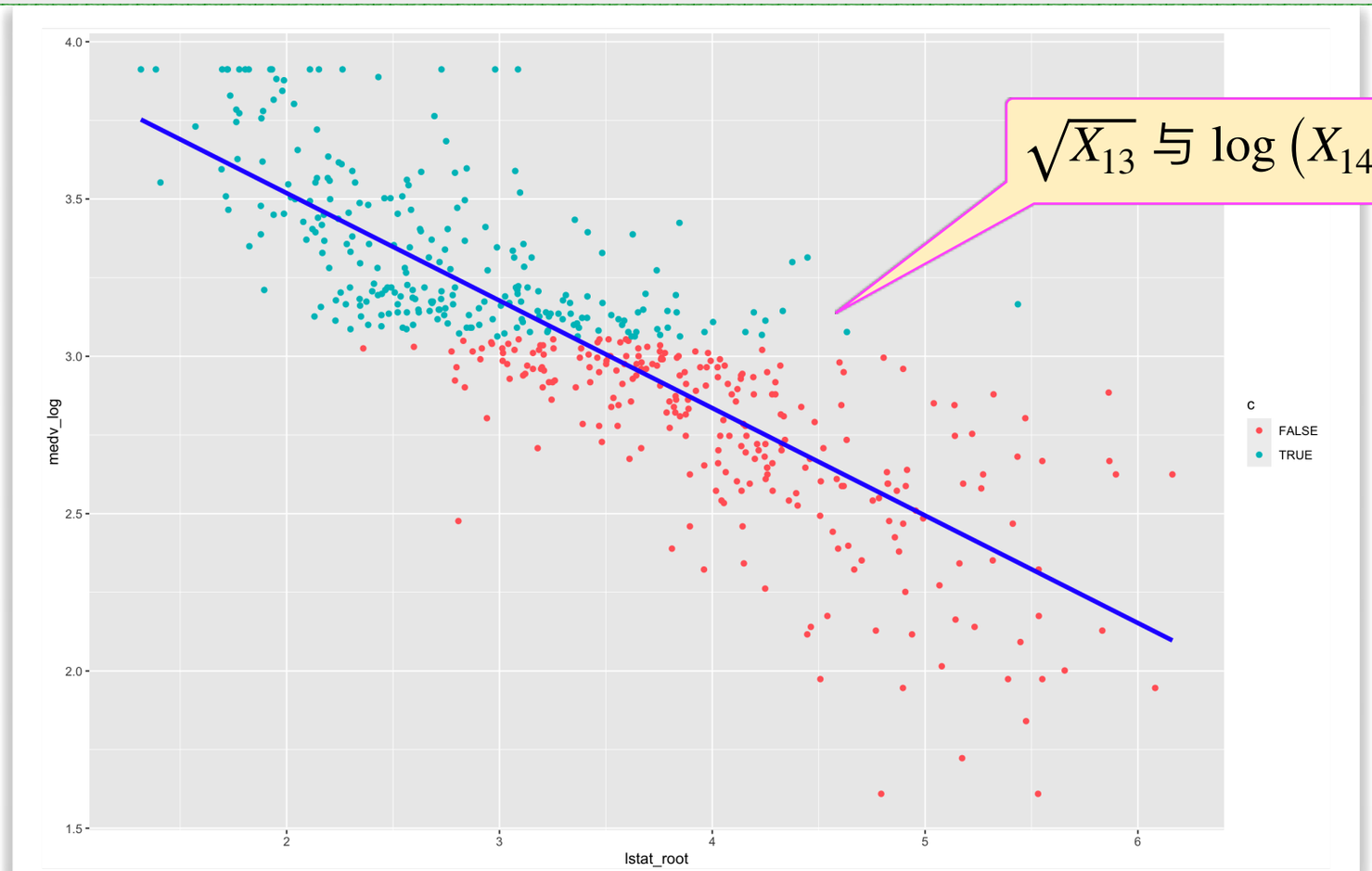
```
ggplot(boston, aes(x = lstat, y = medv, colour = c)) +
  geom_point() +
  geom_smooth(method = 'loess', se = FALSE, span = 0.8, linewidth = 1.5, colour = ‘blue')
p13_2 = ggplot(boston, aes(y = lstat)) +
  geom_boxplot(notch = TRUE, fill = "white", colour = "blue")
```



Boston Housing (波士顿房屋)

- 社会低层人口的百分比 X_{13}
  - 变量 $\sqrt{X_{13}}$ 与 $\log\left(X_{14}\right)$ 的散点图.

```
lstat_root = sqrt(boston$lstat)
medv_log = log(boston$medv)
boston = cbind(boston, lstat_root = lstat_root, medv_log = medv_log)
ggplot(boston, aes(x = lstat_root, y = medv_log, colour = c)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE, linewidth = 1.5, colour = 'blue')
```



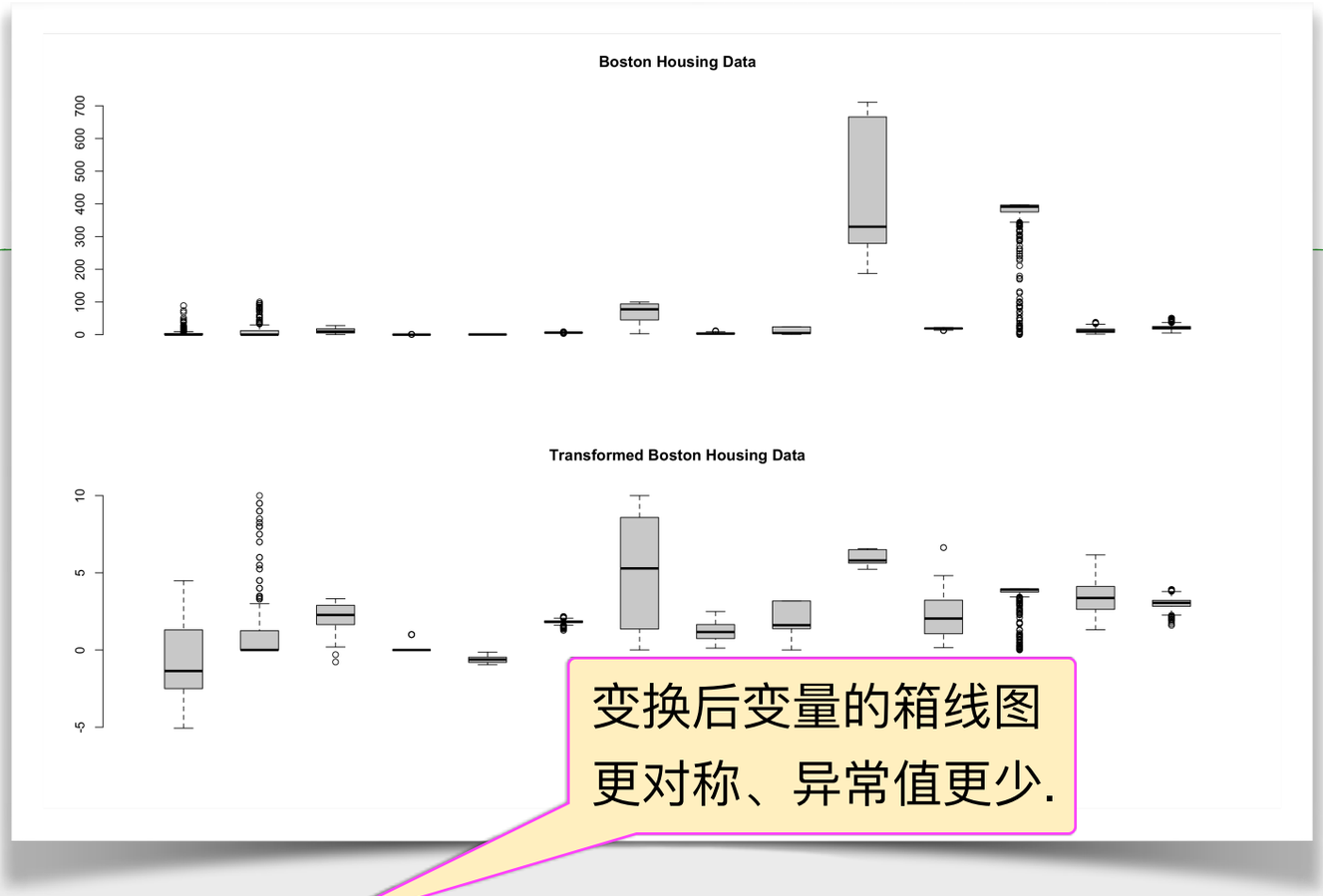
Boston Housing (波士顿房屋)

- 变量变换
  - 由于大多数变量都呈现出不对称，且左侧的密度更高，所以建议进行以下变换：

$$\widetilde{X}_1 = \log\left(X_1\right)$$
$$\widetilde{X}_2 = \frac{X_2}{10}$$
$$\widetilde{X}_3 = \log\left(X_3\right)$$
$$\widetilde{X}_4 = X_4，\text{不作变换，因其为二值变量}$$
$$\widetilde{X}_5 = \log\left(X_5\right)$$
$$\widetilde{X}_6 = \log\left(X_6\right)$$
$$\widetilde{X}_7 = \frac{X_7^{2.5}}{10000}$$
$$\widetilde{X}_8 = \log\left(X_8\right)$$
$$\widetilde{X}_9 = \log\left(X_9\right)$$
$$\widetilde{X}_{10} = \log\left(X_{10}\right)$$
$$\widetilde{X}_{11} = \frac{\exp(0.4 \times X_{11})}{1000}$$
$$\widetilde{X}_{12} = \frac{X_{12}}{100}$$
$$\widetilde{X}_{13} = \sqrt{X_{13}}$$
$$\widetilde{X}_{14} = \log\left(X_{14}\right)$$

Boston Housing (波士顿房屋)

- 变量变换
  - 变换前后各个变量的箱线图.




```
trans_X1 = log(boston$crim)
trans_X2 = boston$zn / 10
trans_X3 = log(boston$indus)
trans_X5 = log(boston$nox)
trans_X6 = log(boston$rm)
trans_X7 = boston$age^(2.5) / 10000
trans_X8 = log(boston$dis)
trans_X9 = log(boston$rad)
trans_X10 = log(boston$tax)
trans_X11 = (exp(0.4 * boston$ptratio)) / 1000
trans_X12 = boston$black / 100
trans_X13 = sqrt(boston$lstat)
trans_X14 = log(boston$medv)
trans_boston = data.frame(trans_X1, trans_X2, trans_X3, boston$chas, trans_X5, trans_X6,
                          trans_X7, trans_X8, trans_X9, trans_X10, trans_X11, trans_X12,
                          trans_X13, trans_X14)
names(trans_boston) = names(Boston)
par(mfrow = c(2, 1))
boxplot(Boston, main = 'Boston Housing Data', axes = FALSE, boxwex = 0.5)
axis(2)
boxplot(trans_boston, main = 'Transformed Boston Housing Data', axes = FALSE, boxwex = 0.5)
axis(2)
```