数据科学导论第 5 讲——数据可视化

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 04 月 07 日

中国传媒大学
COMMUNICATION UNIVERSITY OF CHINA

目录

DEMO

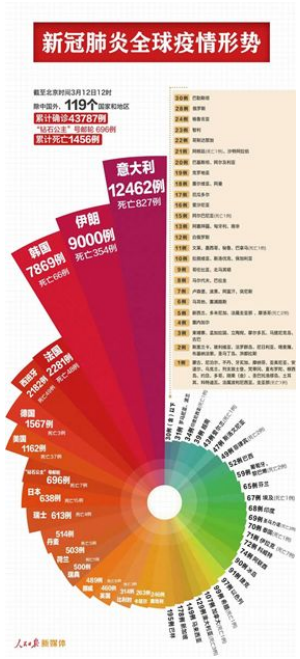


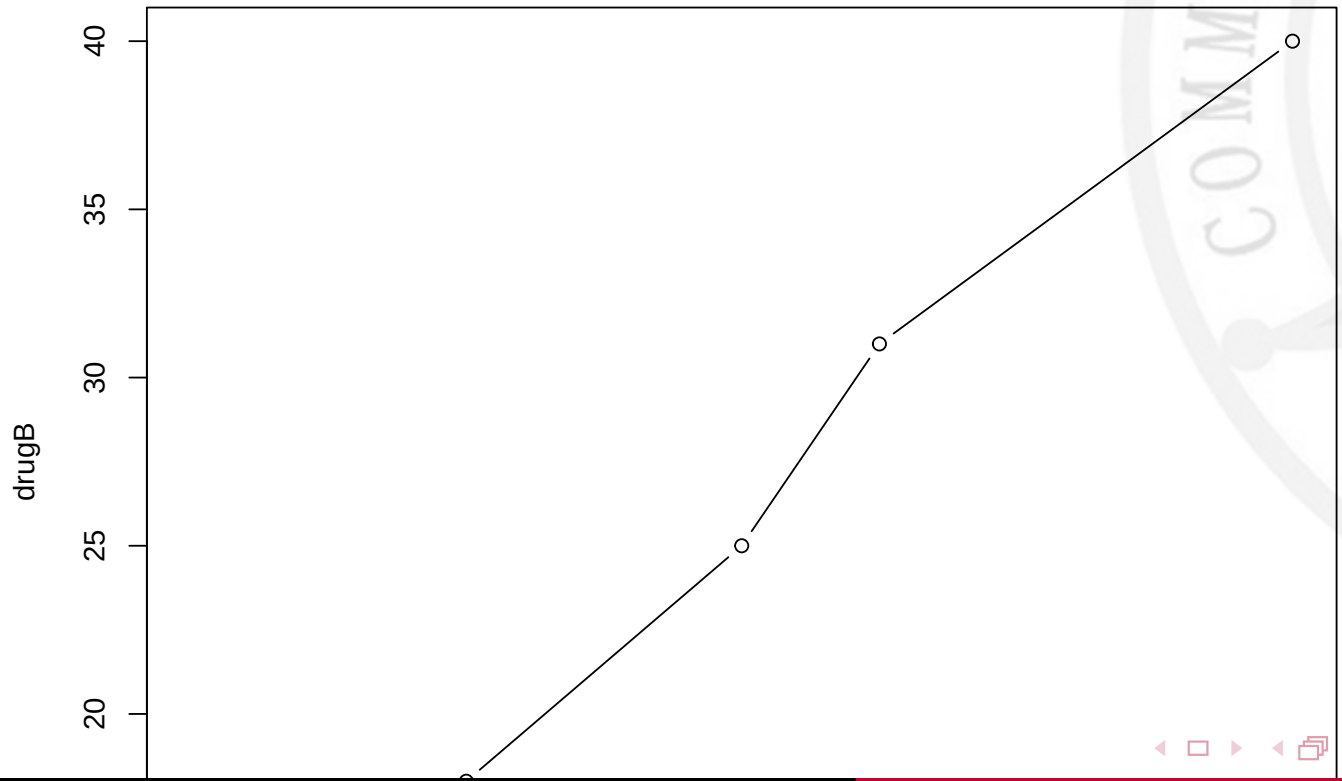
图 1: 南丁格尔玫瑰图

图形初阶

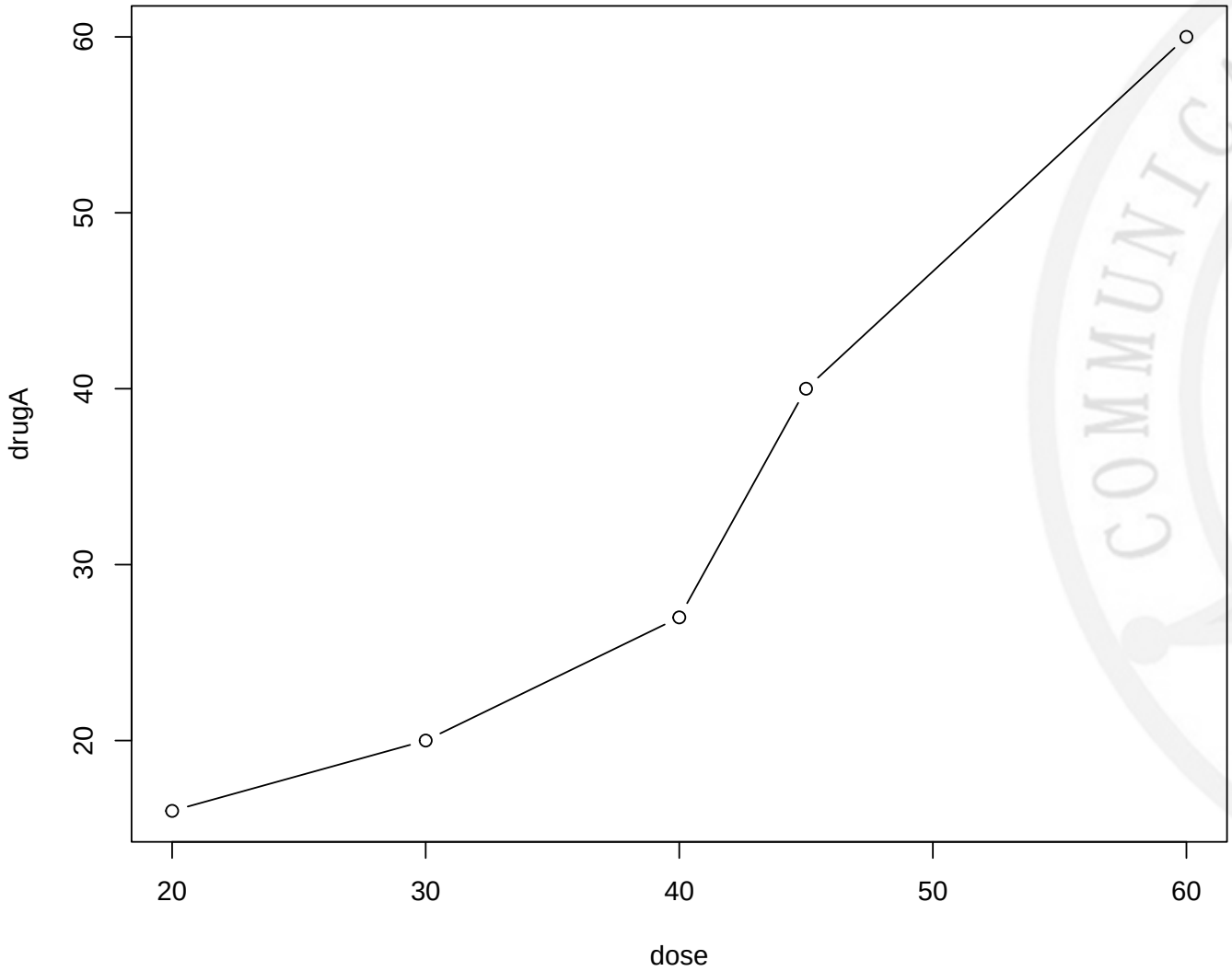
一个例子

- 病人 A 和 B 对两种药物五个剂量水平上的响应情况

```
dose <- c(20, 30, 40, 45, 60);
drugA <- c(16, 20, 27, 40, 60)
drugB <- c(15, 18, 25, 31, 40);plot(dose, drugB, type="
```

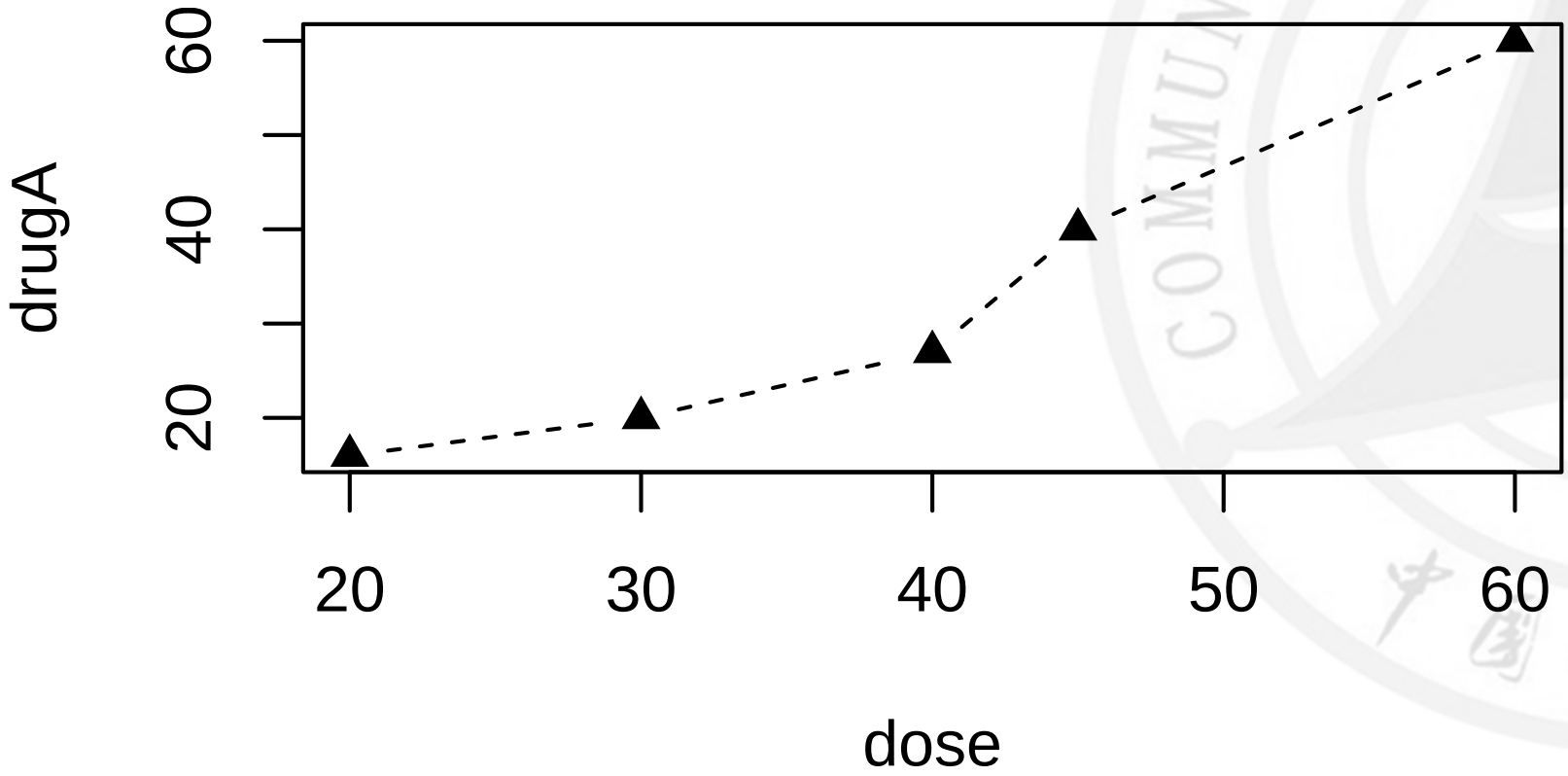


```r
plot(dose, drugA, type="b")
```



图形参数

```
opar <- par(no.readonly=TRUE)
par(lty=2, pch=17)# 线条类型和点符号
plot(dose, drugA, type="b")
```




```
par(opar)
```

图形参数示例

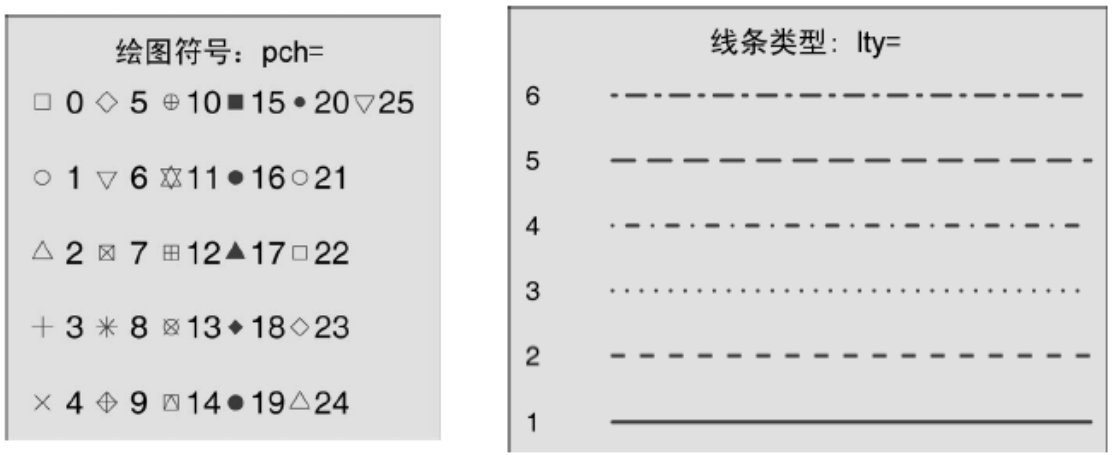


图 2: 图形参数

plot(dose, drugA, type="b", lty=3, lwd=3, pch=15, cex=2)

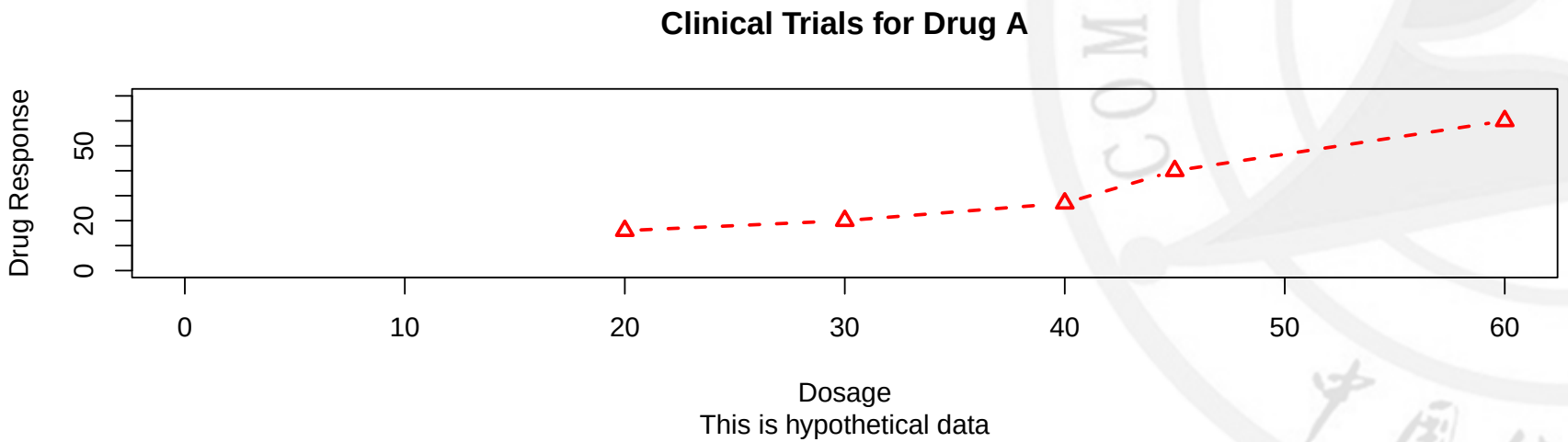
图形属性

- 颜色参数：col，col.axis，col.lab,col.main,col.sub,fg,bg
- 文本属性：cex,cex.axis,cex.main
- 字体属性：font,font.axis,font.lab,font.main
- par(font.lab=3, cex.lab=1.5, font.main=4, cex.main=2)
- 图形尺寸

1. pin 以英寸表示的图形尺寸（宽和高）
2. mai 以数值向量表示的边界大小，顺序为“下、左、上、右”，单位为英寸
3. mar 以数值向量表示的边界大小，顺序为“下、左、上、右”，单位为英分 。默认值为 c(5, 4, 4, 2) + 0.1

添加文本、自定义坐标轴和图例

```
plot(dose, drugA, type="b", col="red", lty=2, pch=2, lw
main="Clinical Trials for Drug A",sub="This is hypothet
xlab="Dosage", ylab="Drug Response",xlim=c(0, 60), ylim
```

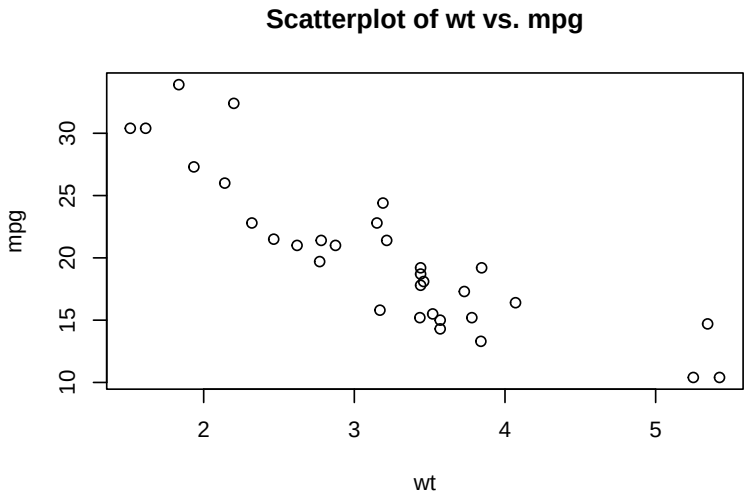


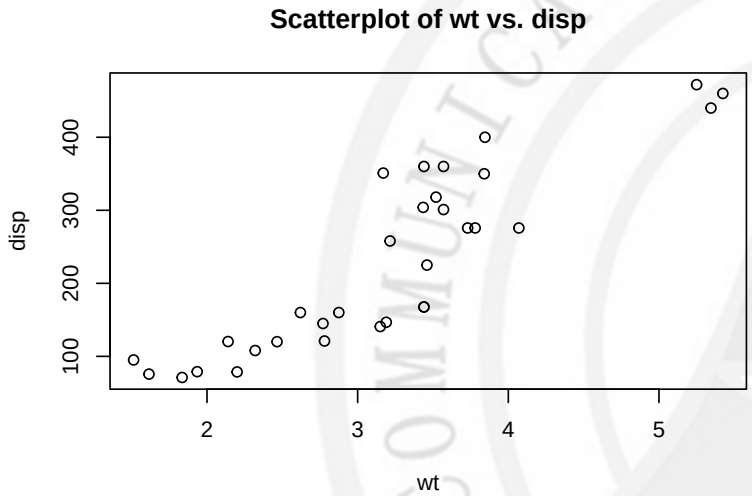
添加文本、自定义坐标轴和图例

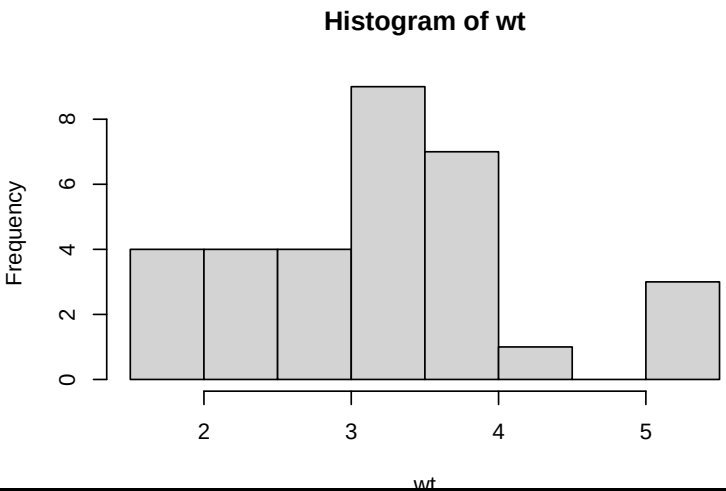
- 标题：title()
- 坐标轴：axis()
- 参考线：abline()
- 图例：legend()
- 文本标注：text()
- 数学标注：plotmath() or demo(plotmath)

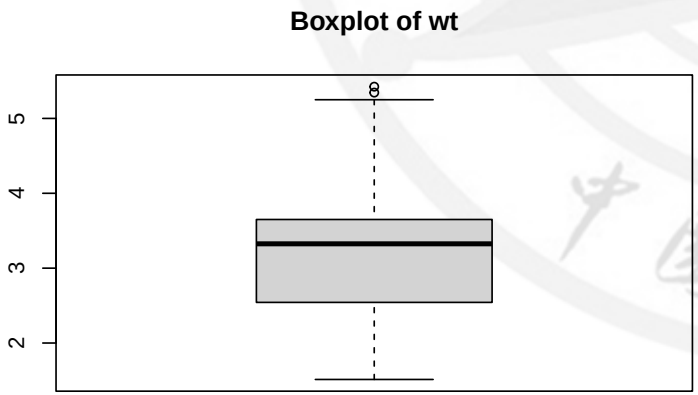
图形组合

- 函数： par() ,par(mfrow=c(2,2))





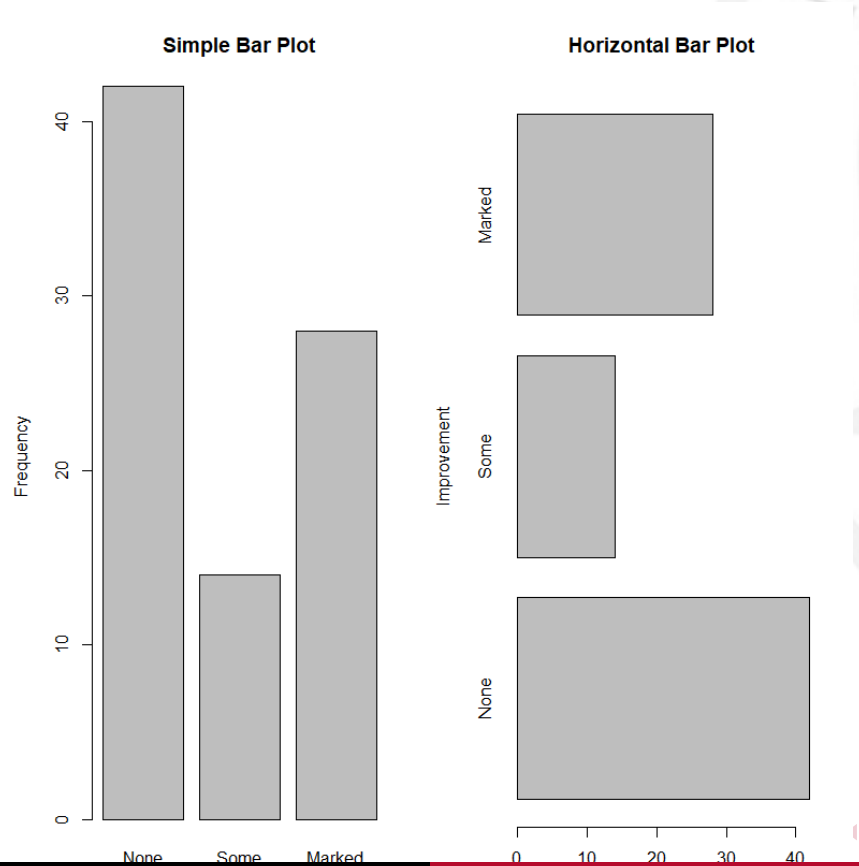




基本图形

条形图

- 条形图通过垂直的或水平的条形展示了类别型变量的分布（频数）
- 函数 barplot() 的最简单用法是：barplot(height)

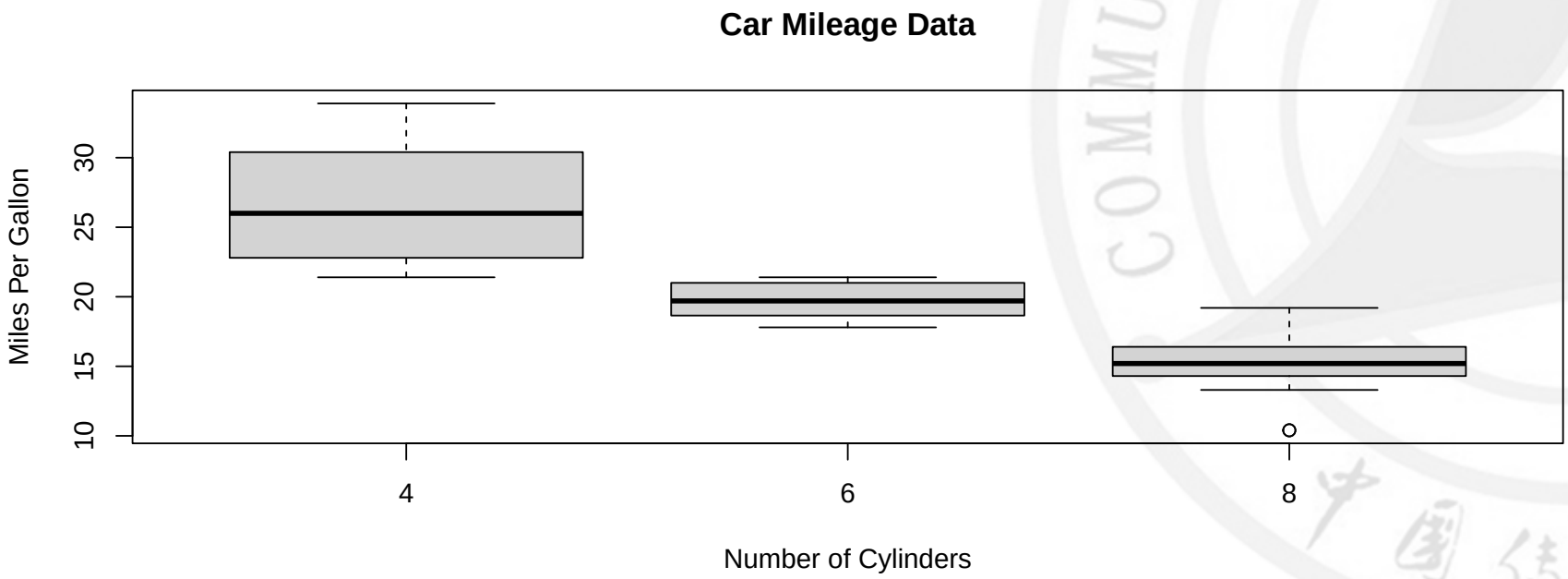


箱线图

- 箱线图（又称盒须图）通过绘制连续型变量的五数总括，即最小值、下四分位数（第 25 百分位数）、中位数（第 50 百分位数）、上四分位数（第 75 百分位数）以及最大值，描述了连续型变量的分布。
- 箱线图能够显示出可能为离群点（范围 ±1.5*IQR 以外的值，IQR 表示四分位距，即上四分位数与下四分位数的差值）的观测。例如：boxplot(mtcars$mpg, main="Box plot", ylab="Miles per Gallon")

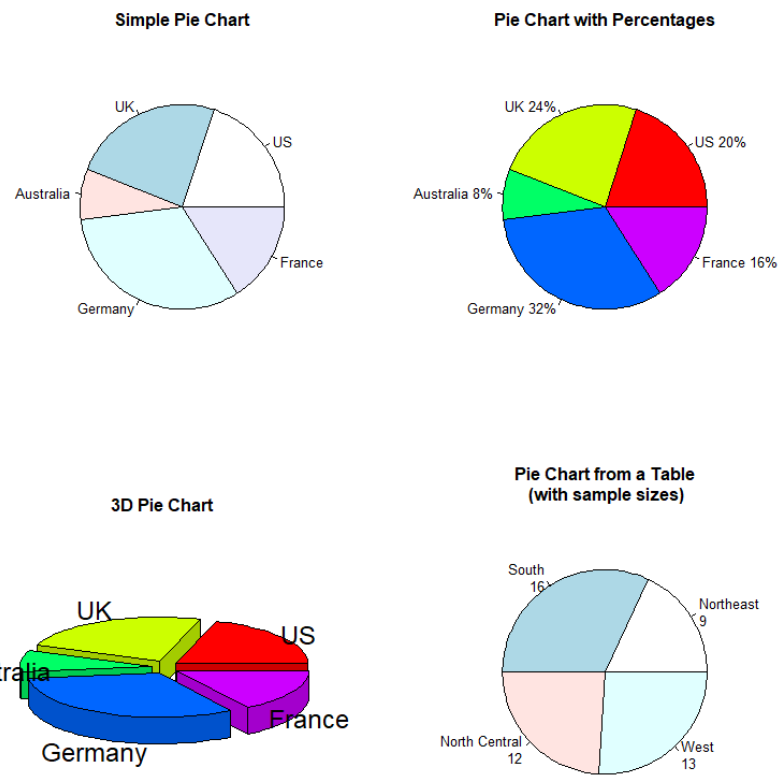
箱线图

```
boxplot(mpg ~ cyl, data=mtcars,main="Car Mileage Data"
xlab="Number of Cylinders",ylab="Miles Per Gallon")
```



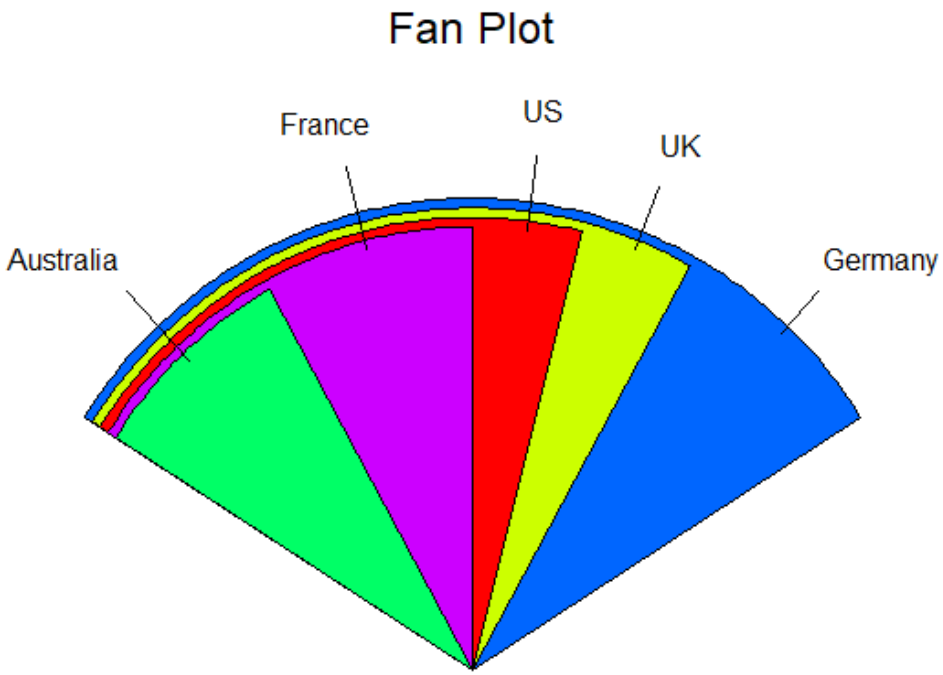
饼图

- 饼图表示同一变量不同水平所占的比例：pie(x, labels),x 非负数值向量，每个扇形的面积，而 labels 则是表示各扇形标签的字符型向量。



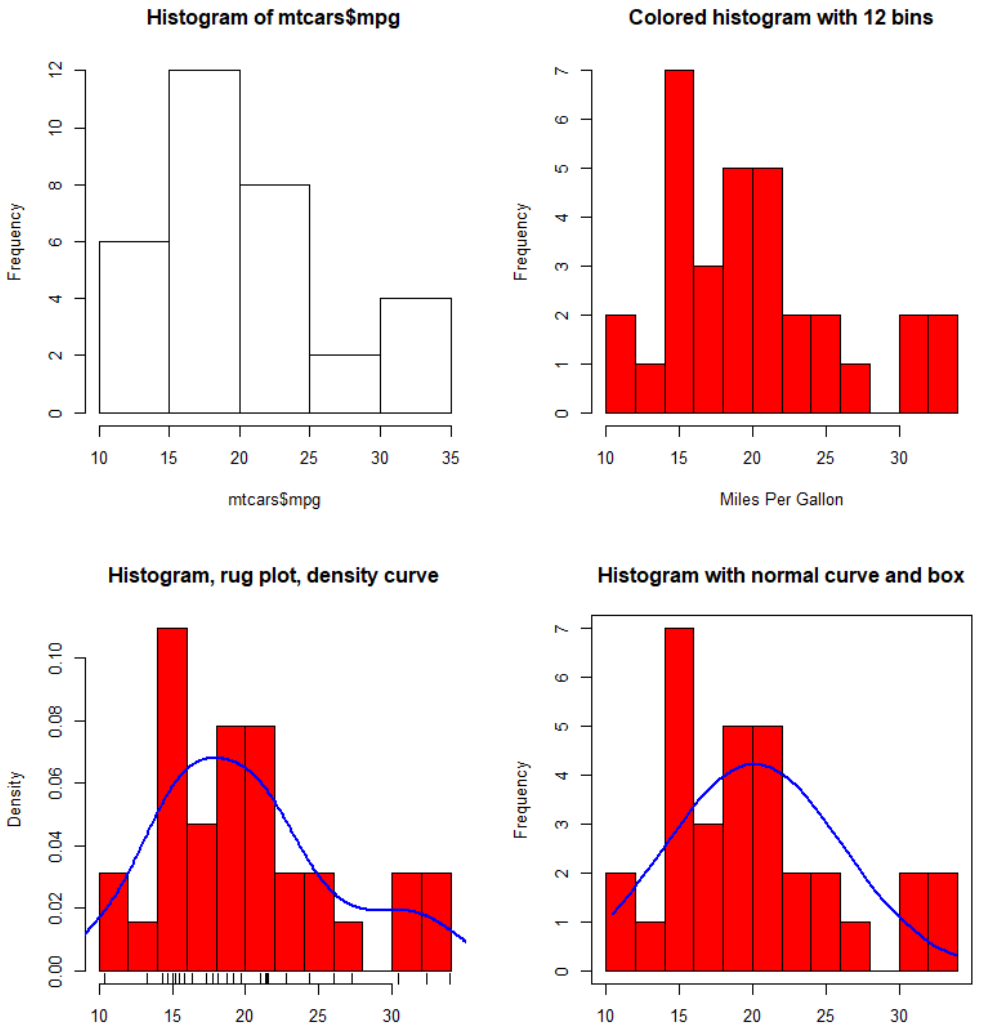
扇形图

- 扇形图是通过 plotrix 包中的 fan.plot() 函数实现的。



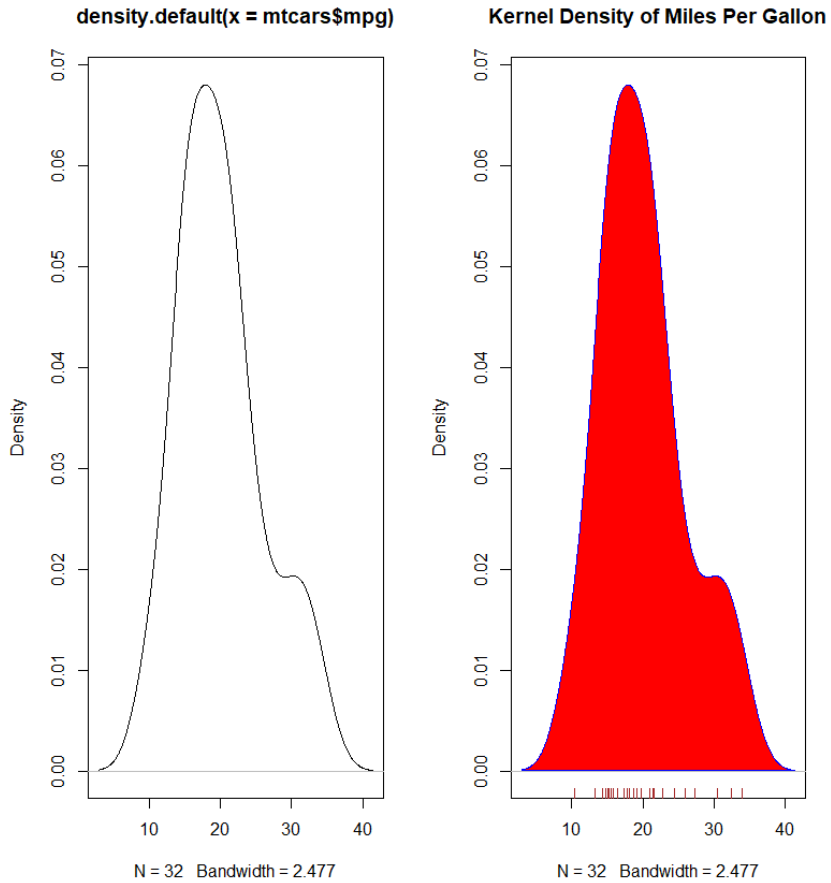
直方图

- 连续型变量的分布：hist(x), x 是一个由数据值组成的



核密度图

- 一种用来观察连续型变量分布的有效方法。绘制密度图的方法:plot(density(x))



点图

- 点图提供了一种在简单水平刻度上绘制大量有标签值的方法: dotchart(x, labels=)
- x 是一个数值向量，而 labels 则是由每个点的标签组成的向量

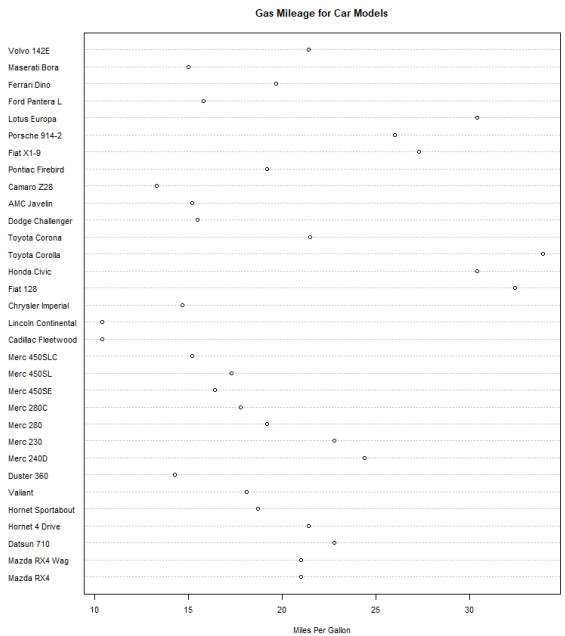


图 6: 点图

散点图

- 来描述两个连续型变量间的关系

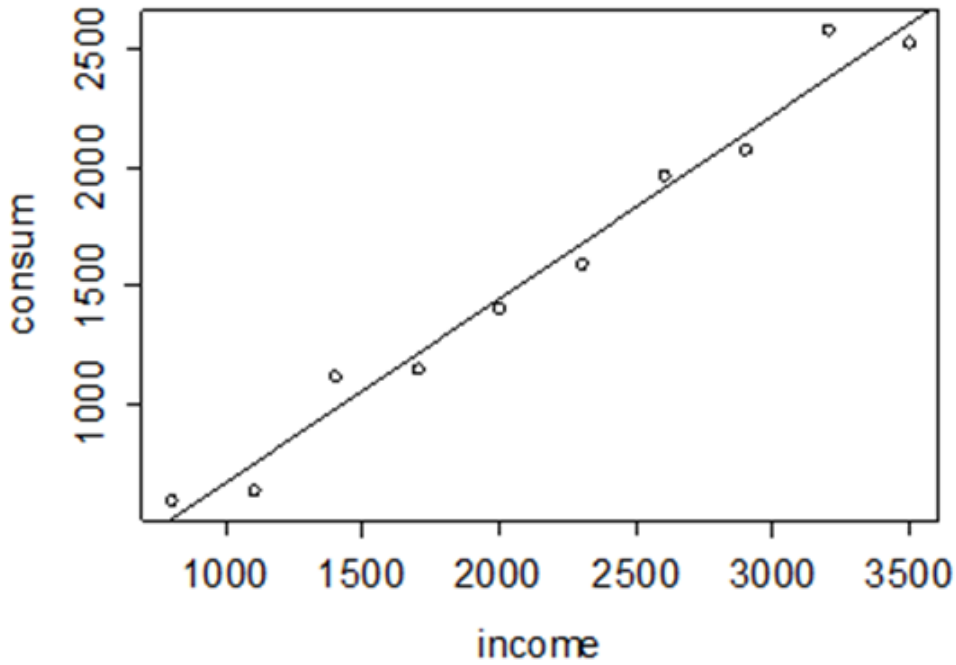
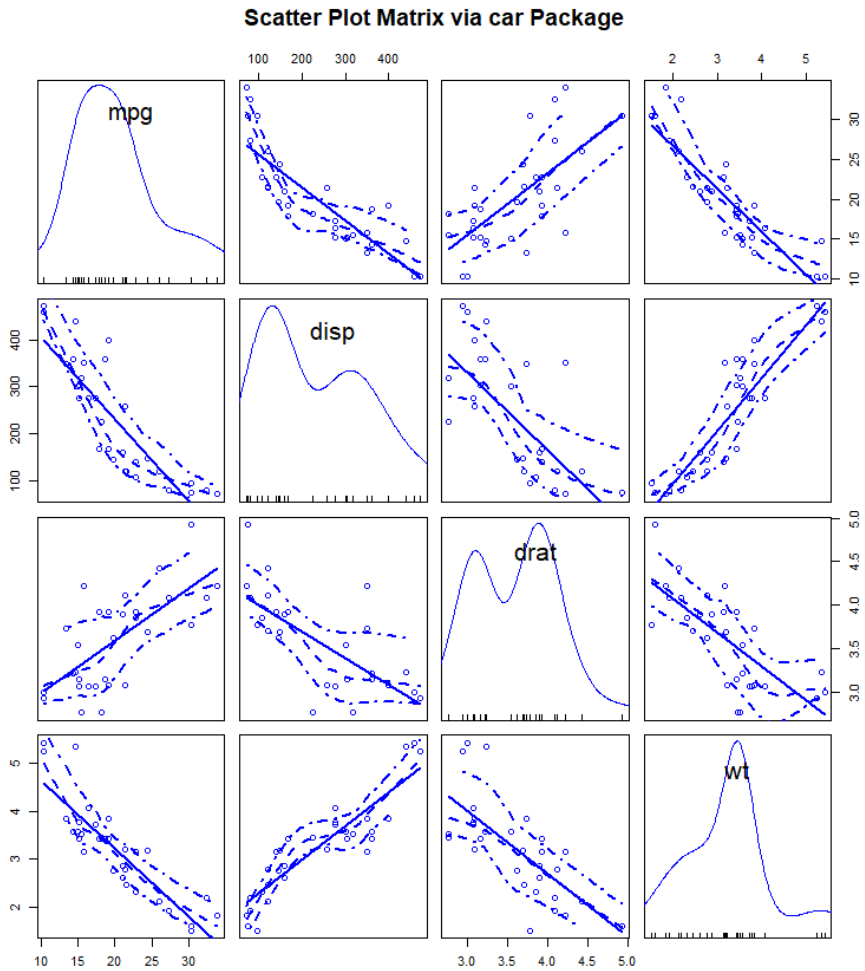


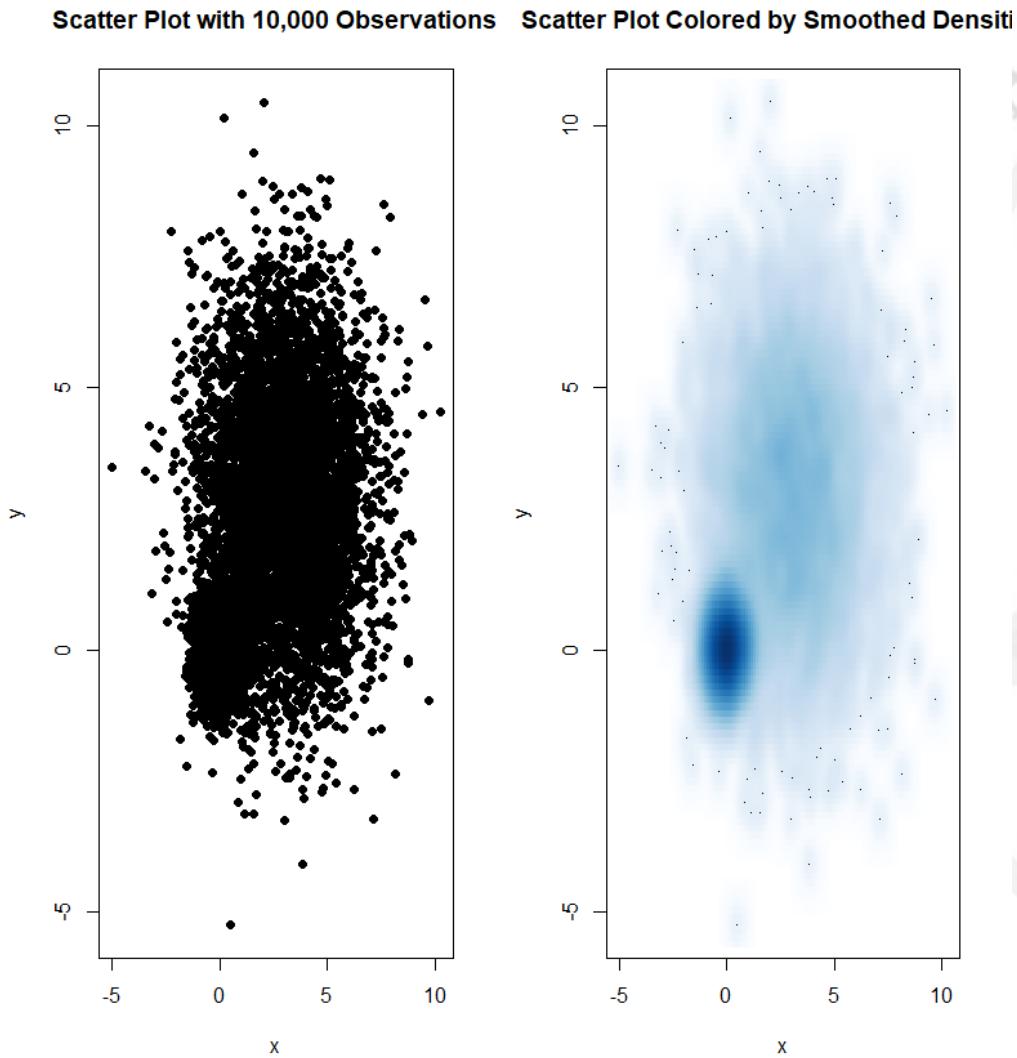
图 7: 散点图

散点图矩阵

- 来描述多个变量中任意两个连续型变量间的关系



高密度散点图


Scatter Plot with 10,000 Observations
Scatter Plot Colored by Smoothed Densiti
y
x
-5
0
5
10

ggplot2

ggplot2 包含以下几个概念

- 数据（Data）和映射（Mapping）
- 标度（Scale）
- 几何对象（Geometric）
- 统计变换（Statistics）
- 坐标系统（Coordinate）
- 图层（Layer）
- 分面（Facet）
- 数据（Data）和映射（Mapping）。

ggplot2

- 标度（Scale）标度负责控制映射后图形属性的显示方式。具体形式上来看是图例和坐标刻度。Scale 和 Mapping 是紧密相关的概念。
- 几何对象（Geometric）几何对象代表我们在图中实际看到的图形元素，如点、线、正方块等多边形。
- 统计变换（statistics）对原始数据进行某种统计计算，例如对二元散点图加上一条回归线或者置信区间登记。
- 分面（Facet）条件绘图，将数据按某种方式分组，然后分别绘图。分面就是控制分组绘图的方法和排列形式。

ggplot2 图书推荐

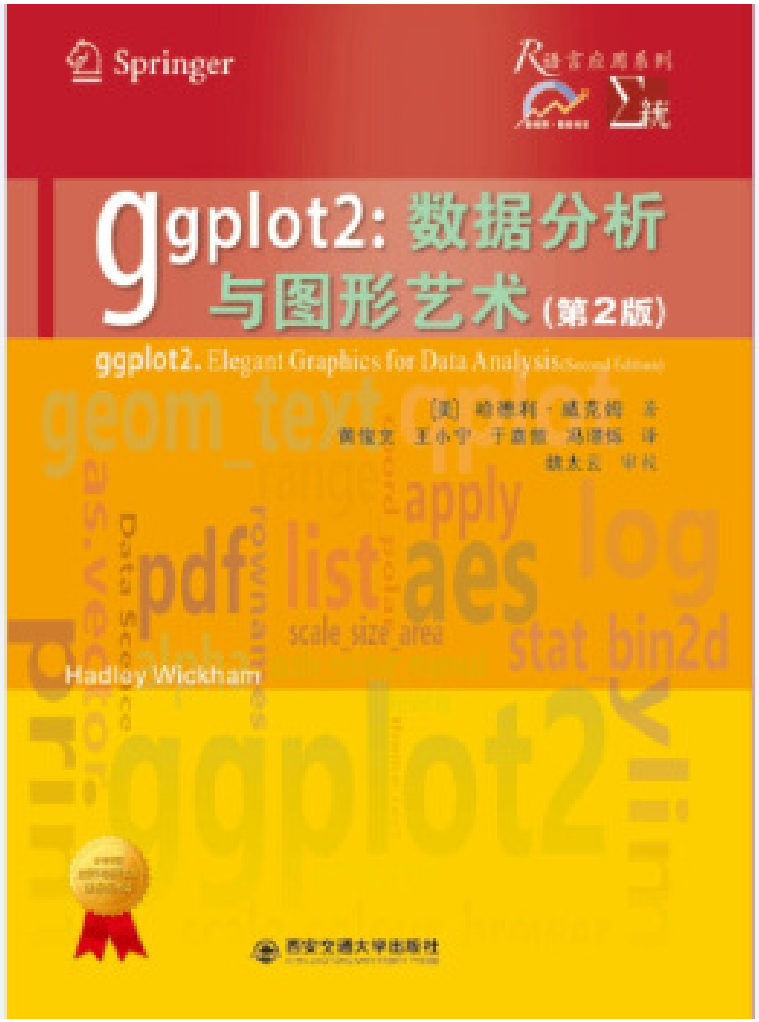


图 10: ggplot2：数据分析与图形艺术第 2 版

本周推荐

1. 一本书：《一个数学家的辩白》，哈代，人民邮电出版社，2020
2. 一部电影：《The Joy of Stats(统计学的乐趣)》，2016
3. 练习：《R 语言实战（第 2 版）》，第 7 章代码实现

谢　谢!