R语言之数据可视化

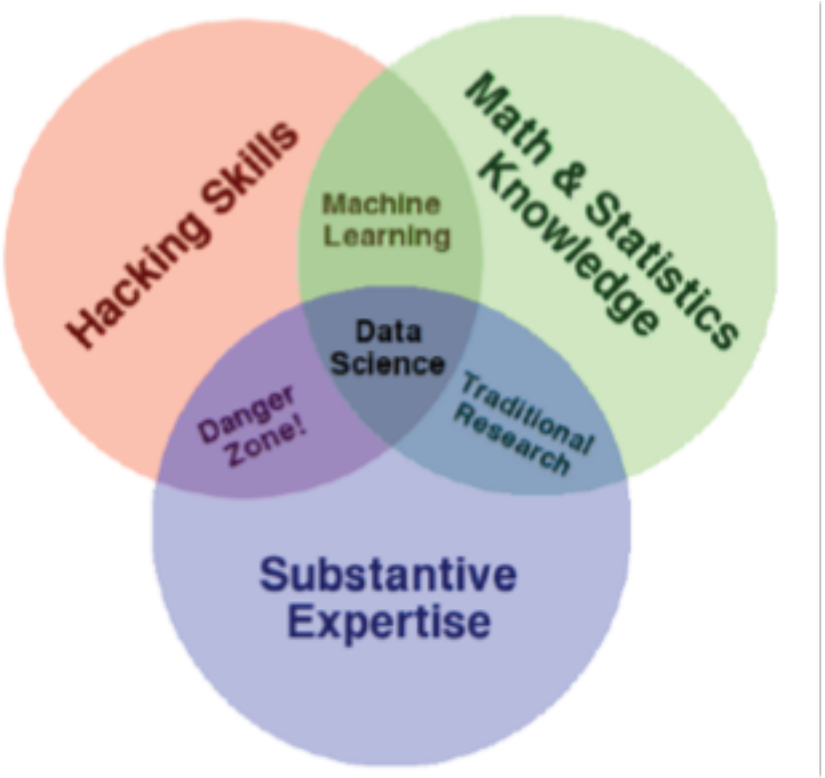
数据科学家需要具备的知识和技能

了解数据的特征

数据可视化: R的绘图系统

制作并发布报告

数据科学家需要具备哪些知识与技能？



Drew Conway: http://drewconway.com/the-lab/

数据科学家的分类

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

O'Reilly Strata Survey (http://radar.oreilly.com/)

http://radar.oreilly.com/2013/06/theres-more-than-one-kind-of-data-scientist.html

R语言体系主要包括什么？

- 掌握数据分析的整个流程、核心知识和技能
  - 入门: 动机 + R的应用举例 + R安装
  - 基础: 数据结构 + 数据操纵
  - 数据可视化: 绘图 + 探索性分析 + R Markdown
  - 统计分析
  - 机器学习: 回归 / 分类 / 聚类 / ...
  - 数据产品开发

完整的数据分析流程

定义研究问题
定义理想的数据集
确定能够获取什么数据
获取数据
清理数据

探索性分析 (数据可视化!!!)
统计分析/建模(机器学习)等

解释结果 (数据可视化!!!)
挑战结果(有没有其他可能?)
书写报告(Reproducible原则)

假设驱动 (Hypothesis Driven)
vs.
数据驱动 (Data Driven)

数据基础

- 观测(observation)、变量(variable)、数据矩阵(data matrix)

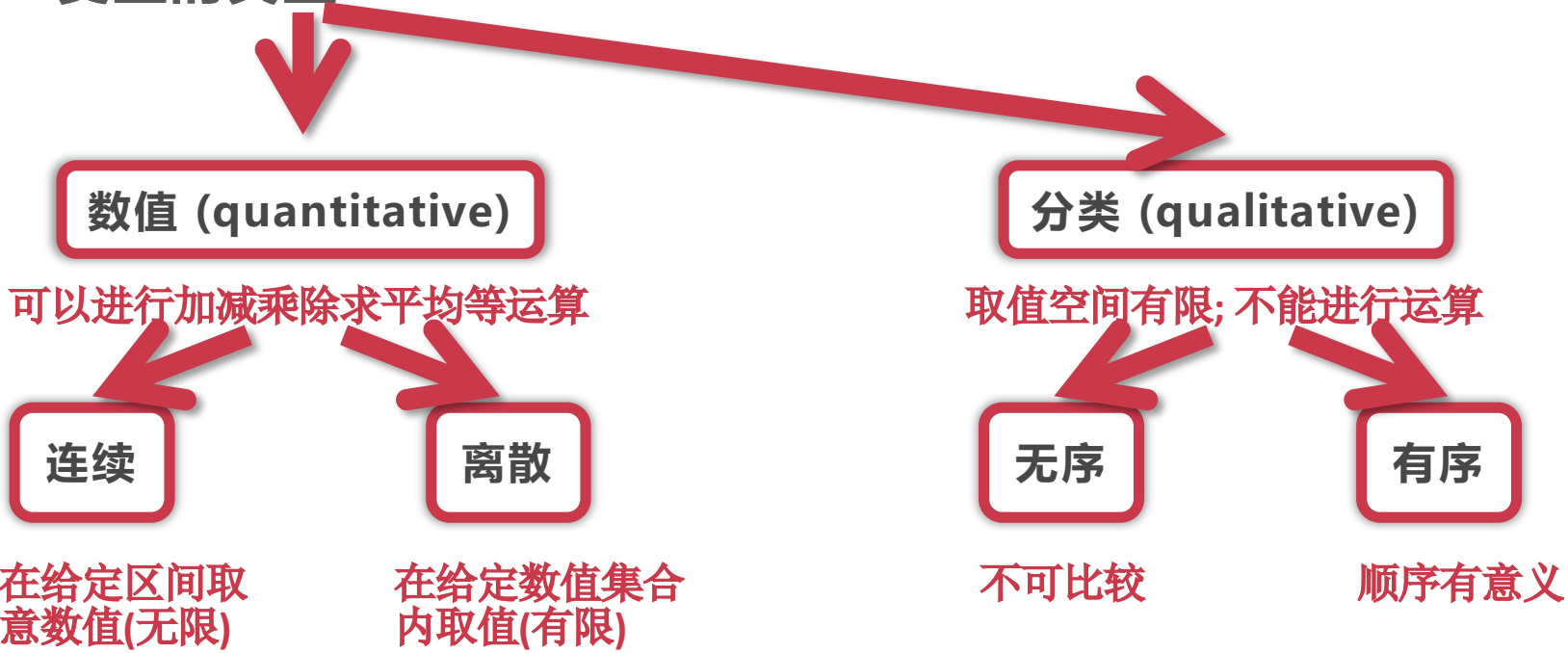
Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6

一次观测

一个变量

数据基础

- 变量的类型


数值 (quantitative)
分类 (qualitative)
可以进行加减乘除求平均等运算
取值空间有限; 不能进行运算
连续
离散
无序
有序
可在给定区间取任意数值(无限)
在给定数值集合内取值(有限)
不可比较
顺序有意义

数据基础

- 变量间的关系 (对应不同的可视化方法和统计分析方法)
  - 两个数值变量
  - 两个分类变量
  - 一个数值变量、一个分类变量

数值变量的特征和可视化

- 数据集中趋势的测量 (measures of center)
  - 均值(mean)、中位数(median)、众数(mode)

1 9 2 8 3 9 4 5 7 6

均值 = (1+9+2+8+3+9+4+5+7+6) / 10 = 5.4

中位数 = 排序后位于正中间的一个数 或 位于正中间的两个数的均值 = 5.5

众数= 出现次数最多的数 = 9

数值变量的特征和可视化

- 数据**分散趋势**的测量 (measures of spread)
  - 值域(range: max-min)、方差(variance)、标准差(standard variance)、四分位距(interquartile range)

1 9 2 8 3 9 4 5 7 6

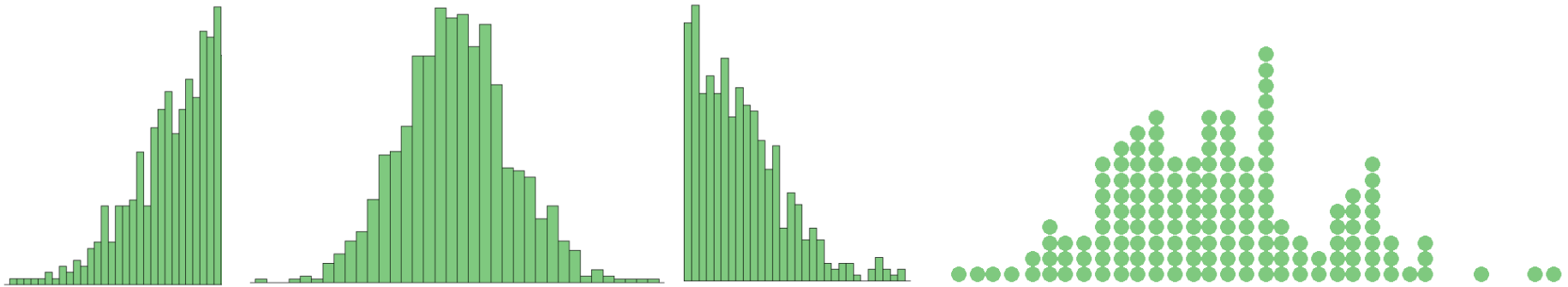
方差 = $\frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}$ = 8.27　　**标准差** = 方差开方 = 2.88

数值变量的特征和可视化

- 稳健统计量 (robust statistics)
  - 是: 中位数、四分位差 (受极端值影响小)
  - 否: 均值、标准差、值域 (受极端值影响大)

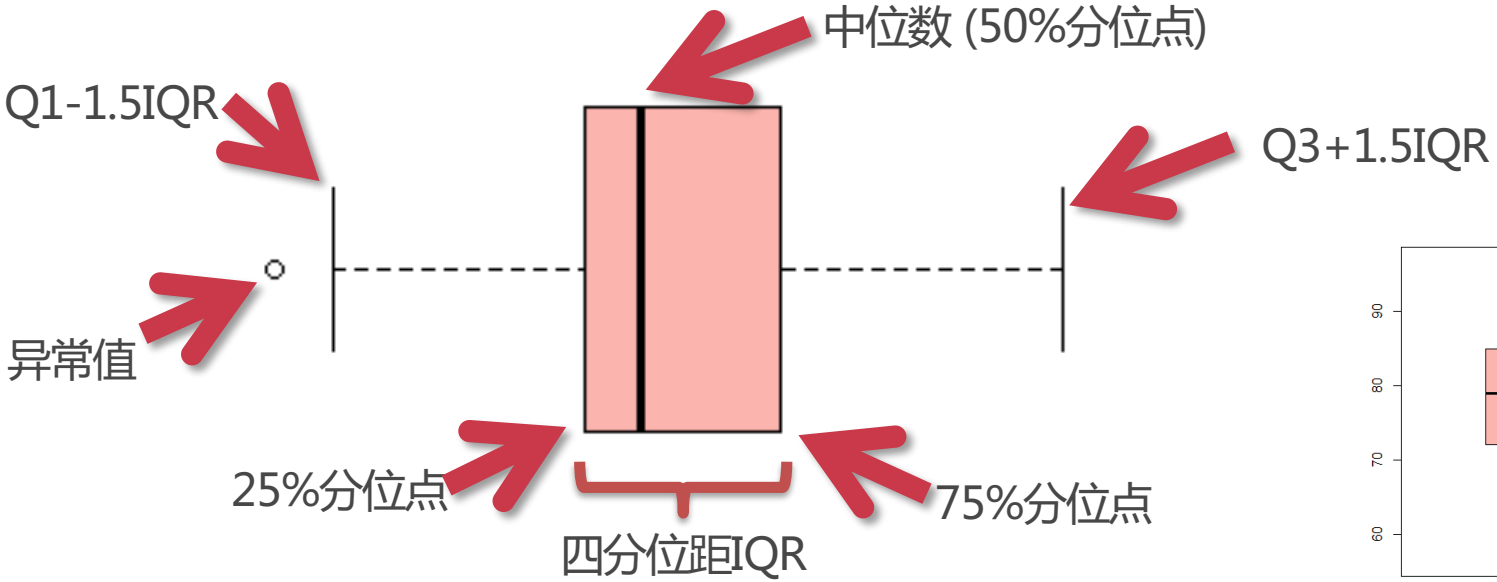
数值变量的特征和可视化

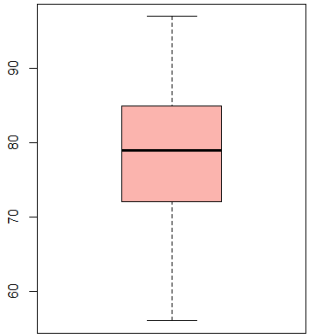
- **一个变量的可视化**
  - 柱状图 (histogram)、点图 (dot plot) (分布)



数值变量的特征和可视化

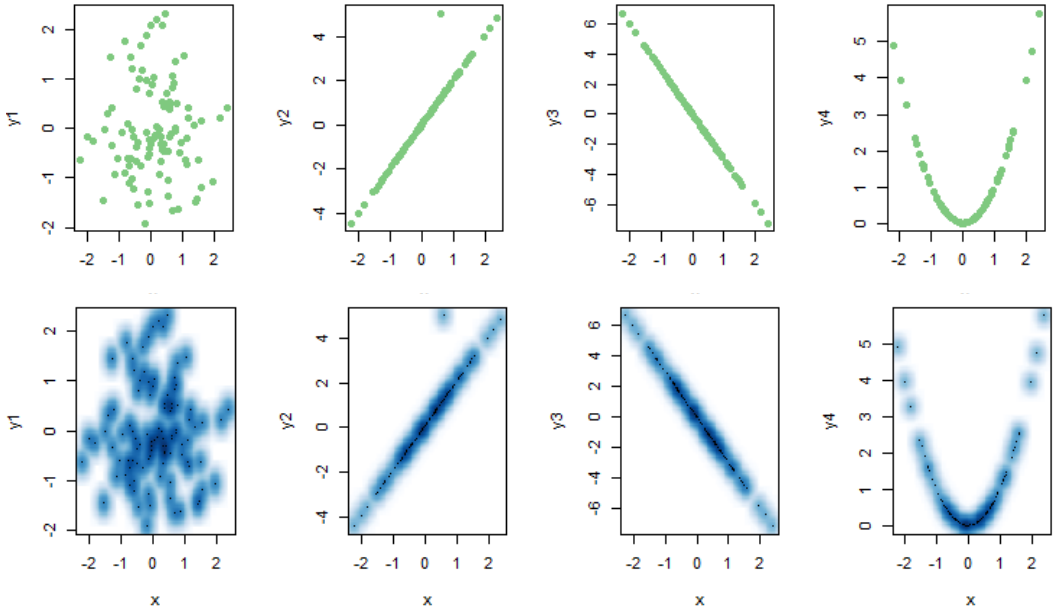
- **一个变量的可视化**
  - **箱图 (box plot) (中位数、分位点、极端值)**





数值变量的特征和可视化

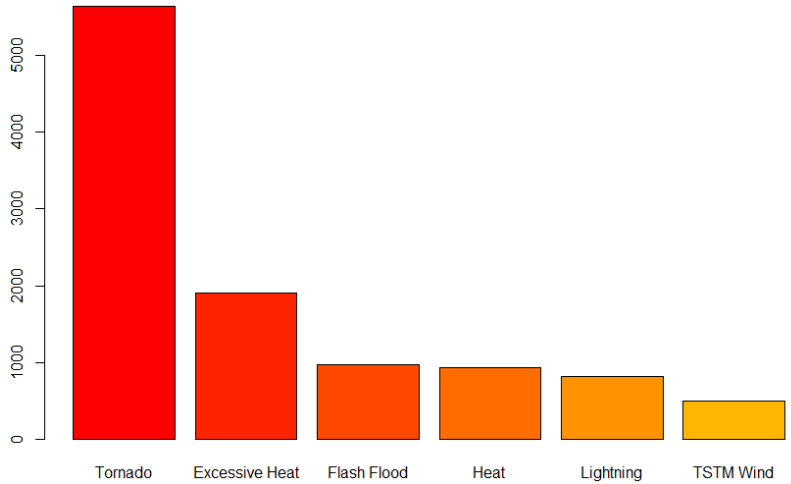
- **两个变量的关系**
  - 散点图 (scatter plot): 方向、形状、强度、极端值



分类变量的特征和可视化

- **一个分类变量的可视化**
  - 频率表(frequency table)、条形图 (bar plot)

Fatalities (thousand persons)	Counts	Frequencies
Tornado	5633	52.3%
Excessive Heat	1903	17.7%
Flash Flood	978	9.1%
Heat	937	8.7%
Lightning	816	7.5%
TSTM Wind	504	4.7%
Total	10771	100%



分类变量的特征和可视化

- **两个分类变量的关系**
  - 关联表(contingency table)、相对频率表(relative frequencies)

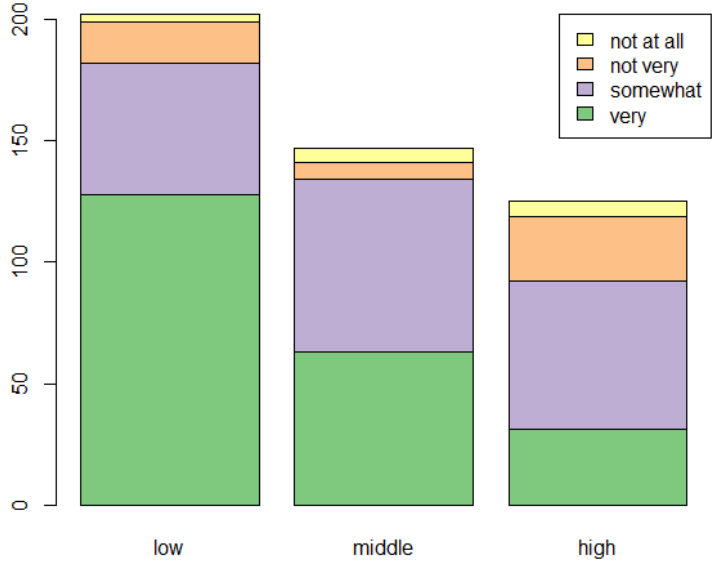
		Age		
		Child	Adult	Total
Survive	No	0	122	122
	Yes	6	197	203
	Total	6	319	325

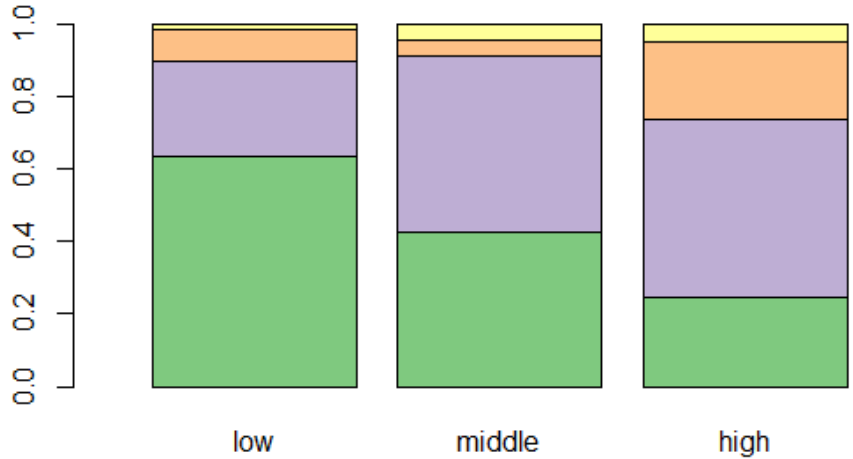
0% 38%
100% 62%

两个变量是互依的
(dependent != causal)

分类变量的特征和可视化

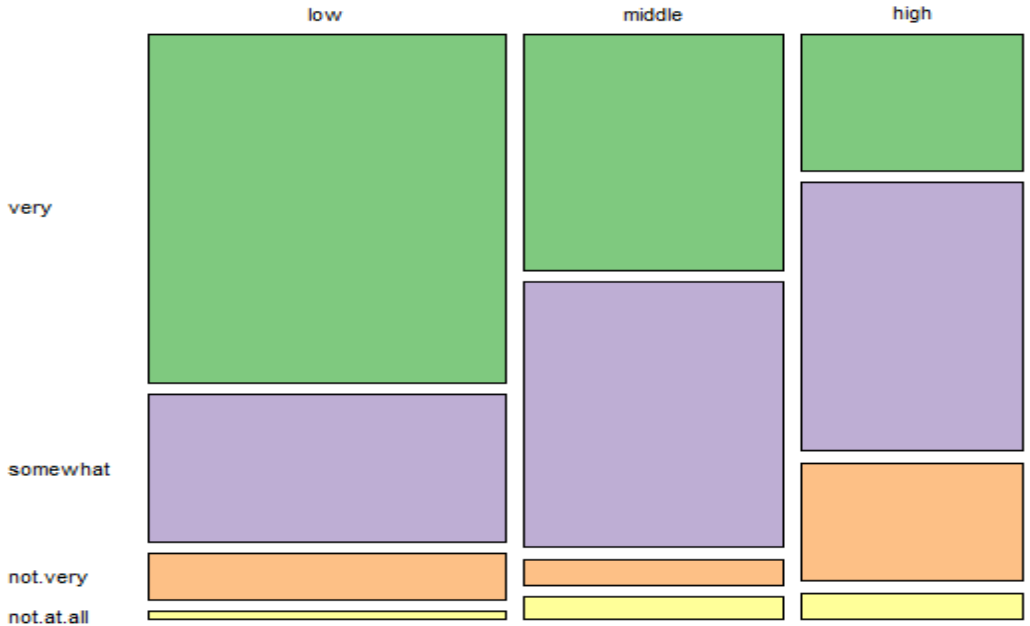
- **两个分类变量的关系**
  - 分段条形图、相对频率分段条形图





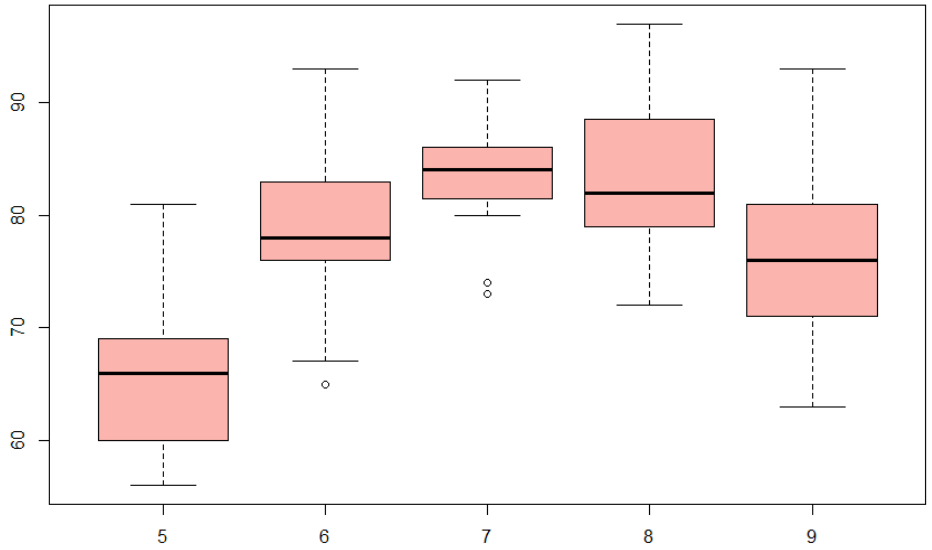
分类变量的特征和可视化

- **两个分类变量**的关系
  - 马赛克图(mosaicplot)



分类变量的特征和可视化

- **一个分类变量、一个数值变量的关系**
  - 并排箱图 (side-by-side box plot)



R的三大绘图系统

- **基本绘图系统 (Base Plotting System)**
  - 艺术家的调色板: 绘图始于空白帆布
    - 需要事先计划; 直观地实时反映绘图和分析数据的逻辑
  - 两步 = 图 + 修饰/添加 = 执行一系列函数
  - 适于绘制2D图

R的三大绘图系统

- **Lattice绘图系统 (Lattice Plotting System)**
  - 绘图 = 使用一次函数调用(一次成图)
  - 特别适用于观测变量间的**交互**: 在变量z的不同水平，变量y如何随变量x变化

R的三大绘图系统

- ggplot2绘图系统 (ggplot2 Plotting System)
  - The Grammar of Graphics
    - 图: 动词、名词、形容词等
    - 数据映射到几何客体(points/lines/bars)的美学属性(颜色/形状/大小)
  - 基本绘图系统 + Lattice绘图系统
    - 自动处理标题/文字说明/空间等, 但也允许通过添加注释进行修改

基本绘图系统

- 绘图函数(graphics包)
  - plot / hist / boxplot / points / lines / text / title / axis
  - 调用函数会启用一个图形设备（如果没有正在运行的图形设备）并设备上绘图
    - 基本绘图系统 + 屏幕设备

基本绘图系统

- **plot()**
  - plot(x, y, ….)
  - 重要参数: xlab / ylab / lwd / lty / pch / col
  - ?par
- **par()**
  - 用于设置全局参数 (作用于R中的所有plot绘图)
    - bg / mar / las / mfrow / mfcol
  - 这些参数可以在每次plot之前进行修改

Lattice绘图系统

- 绘图函数
  - lattice包
    - xyplot / bwplot / histogram / stripplot / dotplot / splom / levelplot / contourplot
    - 格式: xyplot(y ~ x | f * g, data)
    - panel函数，用于控制每个面板内的绘图
  - grid包
    - 实现了独立于base的绘图系统
    - lattice包是基于grid创建的; 很少直接从grid包调用函数

Lattice绘图系统

- **Lattice与Base的重要区别**
  - Base绘图函数直接在图形设备上绘图
  - Lattice绘图函数返回trellis类对象
    - 打印函数真正执行了在设备上绘图
    - 命令执行时,trellis类对象会被自动打印,所以看起来就像是lattice函数直接完成了绘图

ggplot2绘图系统

- 层 (Layer)

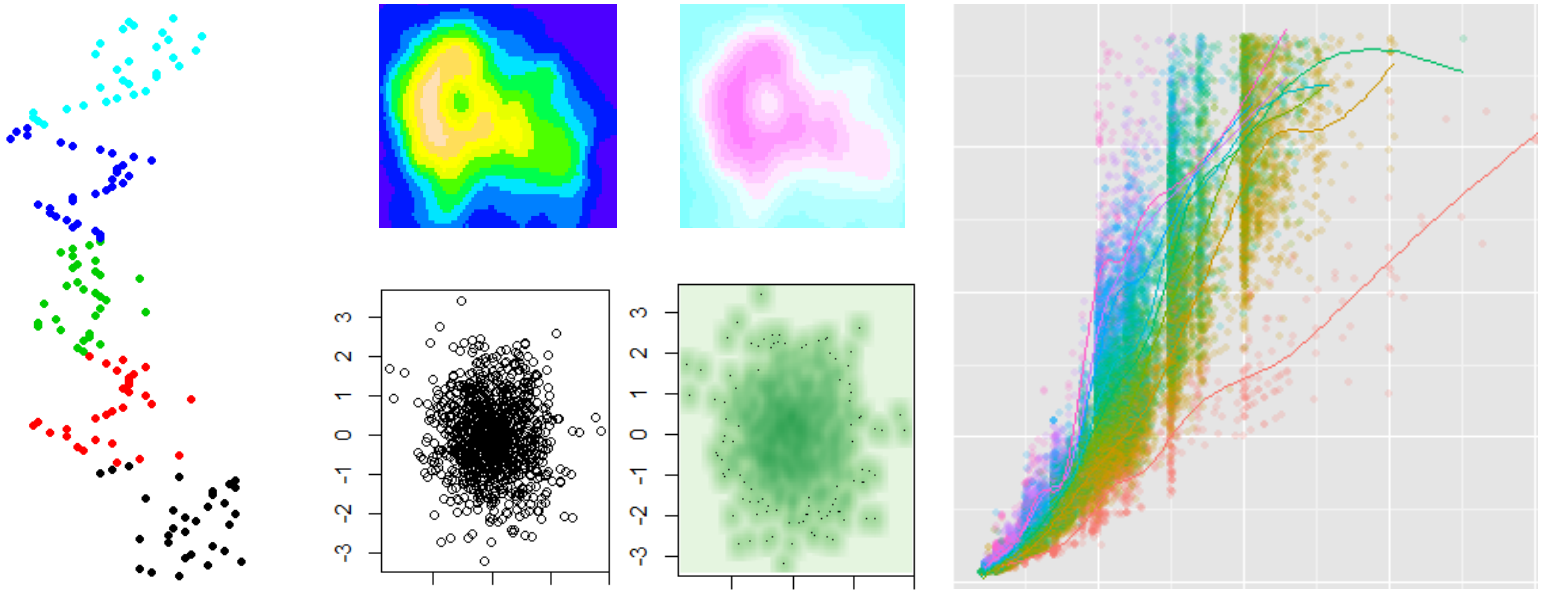
Data	感兴趣的变量 (data frame)
Aesthetics	x-axis / y-axis / color / fill / size / labels / alpha / shape / linear width / linear type
Geometries	point / line / histogram / bar / boxplot
Facets	columns / rows
Statistics	binning / smoothing / descriptive / inferential
Coordinates	cartesian / fixed / polar / limits
Themes	non-data ink

ggplot2绘图系统

- 绘图函数
  - qplot()
    - 类似于Base系统的plot(), 参数包含aesthetics/geom/facet…
    - 隐藏了绘图实现的细节
  - ggplot()
    - 是核心,可以实现qplot()无法实现的功能
    - 调用ggplot()本身并不能实现绘图, 要在其基础上添加层(如geom_point())才可以

R语言绘图之颜色

- 颜色的重要性

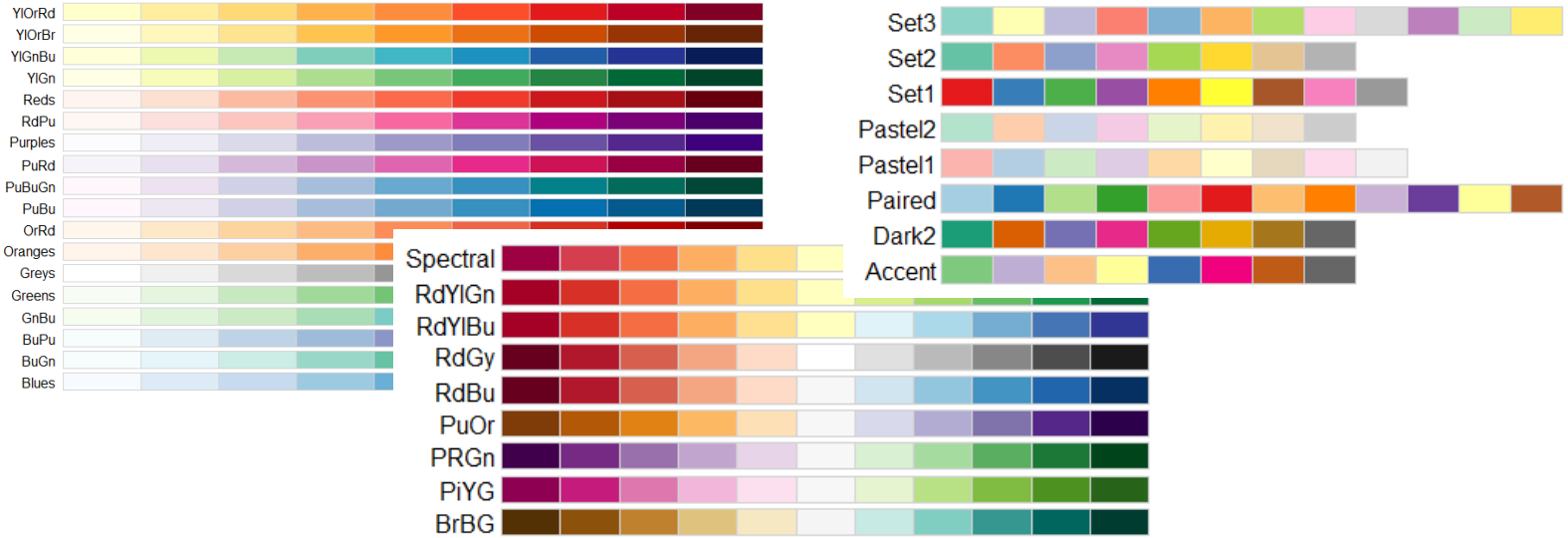


R语言绘图之颜色

- grDevice包
  - colorRamp() & colorRampPalette()
  - 颜色名字可使用colors()获取

R语言绘图之颜色

- RColorBrewer包
  - 三类调色板: sequential / diverging / qualitative
  - 调色板信息可与colorRamp/colorRampPalette结合使用



R支持的图形设备

- 什么是图形设备
  - 屏幕设备(探索性分析常用): 电脑屏幕
    - windows() on Windows / quartz() on Mac / xll() on Unix or Linux
  - 文件设备(打印/文章用图常用)
    - 向量格式 (vector format): PDF
    - 位图 (bitmap format): PNG/JPEG/TIFF/BMP
  - grDevices包
    - 包含了实现各种图形设备的代码
    - ?Devices 如PDF / PNG / BMP

R支持的图形设备

- 生成图形的两种途径
  - 调用绘图函数(默认使用屏幕) → 屏幕设备显示图形 → 进一步修饰图形
  - 明确指定图形设备 → 调用绘图函数 (如果指定的是文件设备则无法在屏幕上看到图形) → 进一步修饰图形 → 关闭图形设备 dev.off()
  - 可同时打开多个设备，但一次只能在一个设备上绘图
    - dev.cur() / dev.set()

R支持的图形设备

- **拷贝图形**
  - 多个设备之间互相拷贝: dev.copy
    - 拷贝到PDF文件: dev.copy2pdf
  - 注意: 拷贝的结果可能与原图有出入

探索性数据分析

- 目的
  - 了解数据的特征
  - 寻找数据中存在的模式

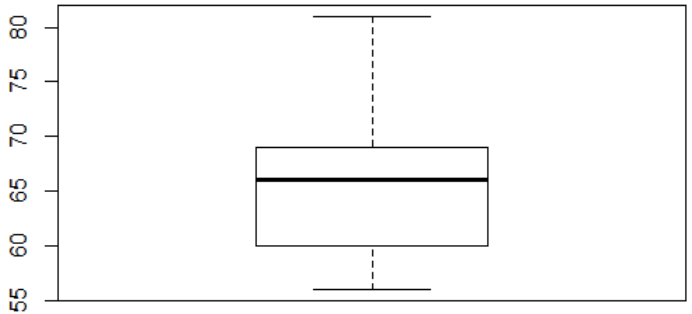
- 为什么离不开图？
  - 了解数据特征、找到数据中的模式、形成分析策略
  - 与数字互相验证、帮助发现错误、用于交流结果

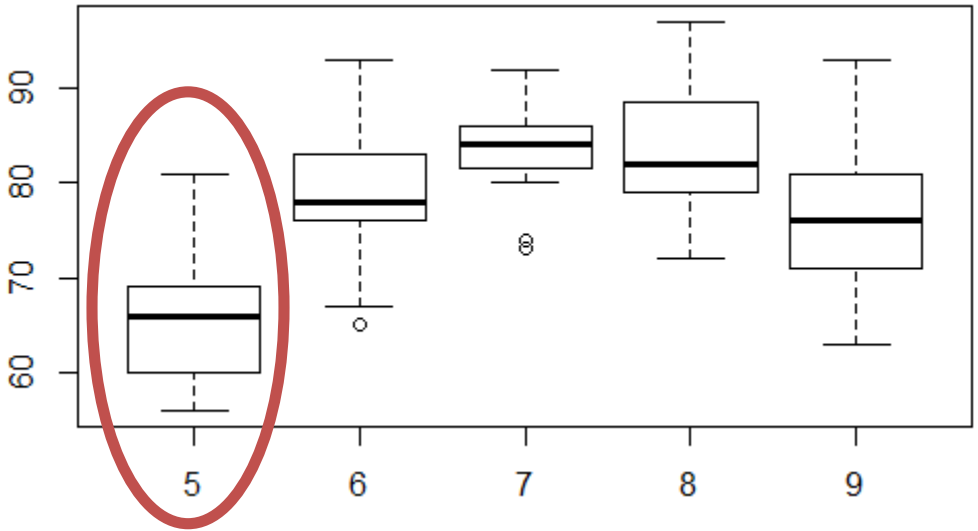
绘图前请思考

- 在哪儿绘图 (屏幕？文件？)
- 如何使用图 (屏幕呈现？网页呈现？文章用图？)
- 用于绘图的数据量的大小？
- 是否需要动态调整图的大小？
- 用哪个绘图系统 (Base / Lattice / ggplot2)？一般三者不混用

分析性作图的六大原则

- 凸显比较 (谁和谁比？)





分析性作图的六大原则

- 凸显机制 (你认为可能的原因？)

分析性作图的六大原则

- 凸显多元性 (>2个变量、逃离扁平)

分析性作图的六大原则

- 整合证据
  - 整合文字、数字、图、表等
  - 用多种方式显示数据的特征

分析性作图的六大原则

- 使用适当的图标、尺度等
  - 完备性、一图胜千言
- 内容是王道

R Markdown

- 可重复性研究(Reproducible Research) 的工具
  - 重复(Replication)的缺点: 没钱/没时间/研究的独特性
  - 让数据和分析过程透明

获取帮助

- 如何问问题
  - 操作系统、版本、哪一步产生的错误、预期是什么、得到的结果是什么、其他有用的信息
  - 例如：Win7 R 3.2.0 lm() “seg fault on large data frame”

- Google