数据可视化

第五讲：可视化方法-基本图形

Our structure

Part 1
什么是可视化
可视化的重要性

Part 2
可视化的理论基础

- 视觉感知
  - 格式塔理论八大原则
  - 可视化编码：视觉通道
- 数据基础
  - 数据与可视化密不可分
  - 可视化贯穿每一个环节
- 图形完整性
- 六原则
- 图形重新设计：Lie factor，Data-ink ratio，Data density

Part 3
可视化应用及基于Python的数据可视化实操

- Python可视化工具介绍
- 基本图形及复杂图形的可视化
- 时空数据可视化
- 文本数据可视化
- 网络和层次数据可视化

可视化任务

- 迭代的数据探索过程

来源：大数据可视化，周苏、王文编著，清华大学出版社，2016年

目录

- 数据变换
- 统计图表

数据变换

- 标准化
- 平滑化
- 采样
- 分箱
- 降维
- 聚类

数据变换

- 目的
  - 更好地解决特定问题
  - 提供更多的可视化设计选择

- 举例 – 数值型温度
  - 发现温度变化的异常值 – 数值型温度
  - 分析全球温度是否升高 – 数值型温度差
  - 判断水温是否适合洗澡 – 序数型(hot, warm, cold)

数据归一化

目的

根据分布映射数据

颜色/尺寸/坐标位置编码

归一化区间：

[-1，1]

[0，1]

数据变换

线性变换

$$y = \frac{x - MinValue}{MaxValue - MinVlue}$$

[0, 1]

反正切变换

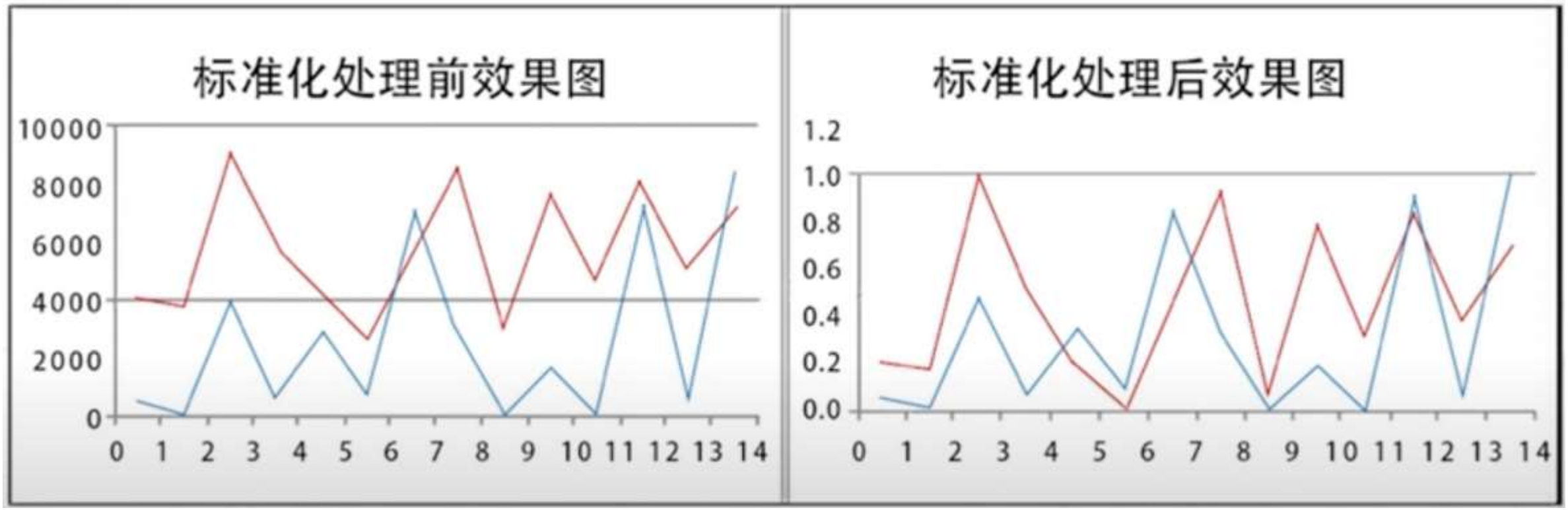
$$y = \frac{\arctan(x) * 2}{\pi}$$

[−1, 1]

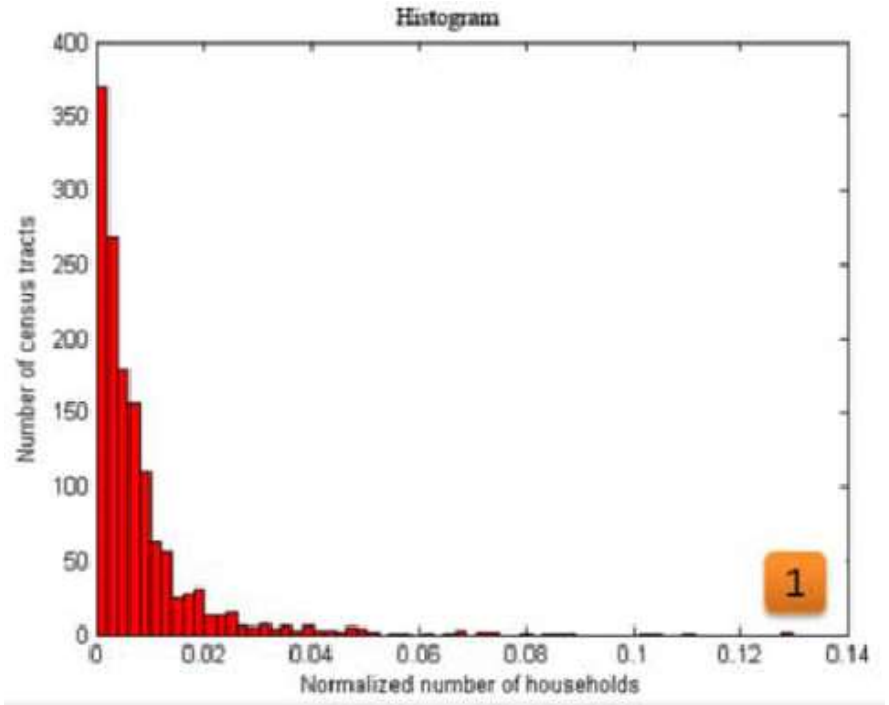
对数变换

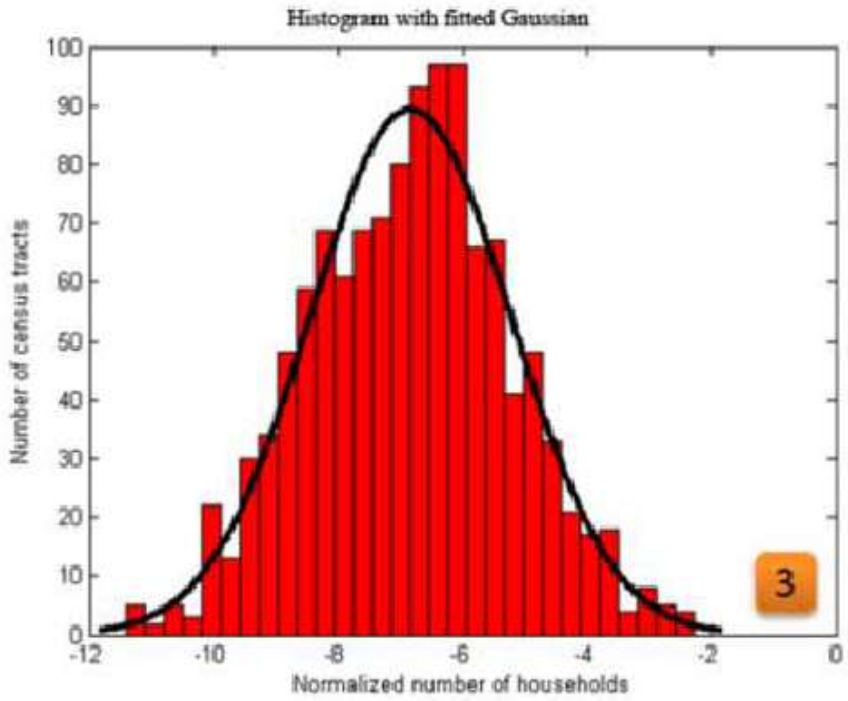
$$y = \log_{10}(x)$$

可以自定义变换函数


标准化处理前效果图
10000
8000
6000
4000
2000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
标准化处理后效果图
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

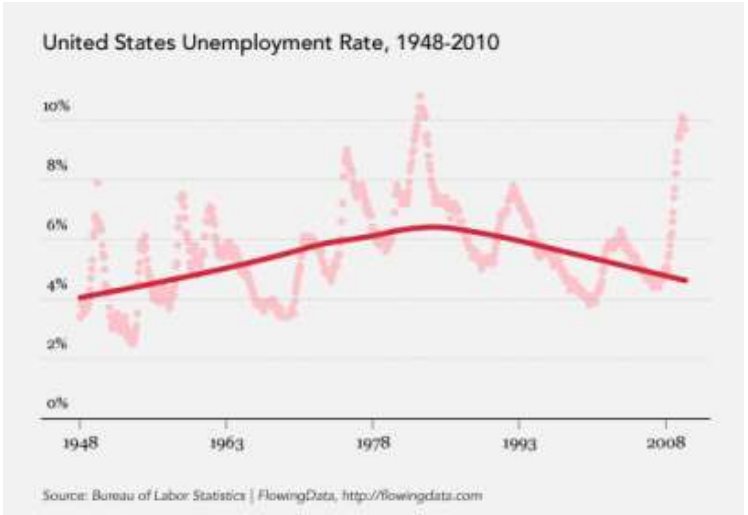
指数变换

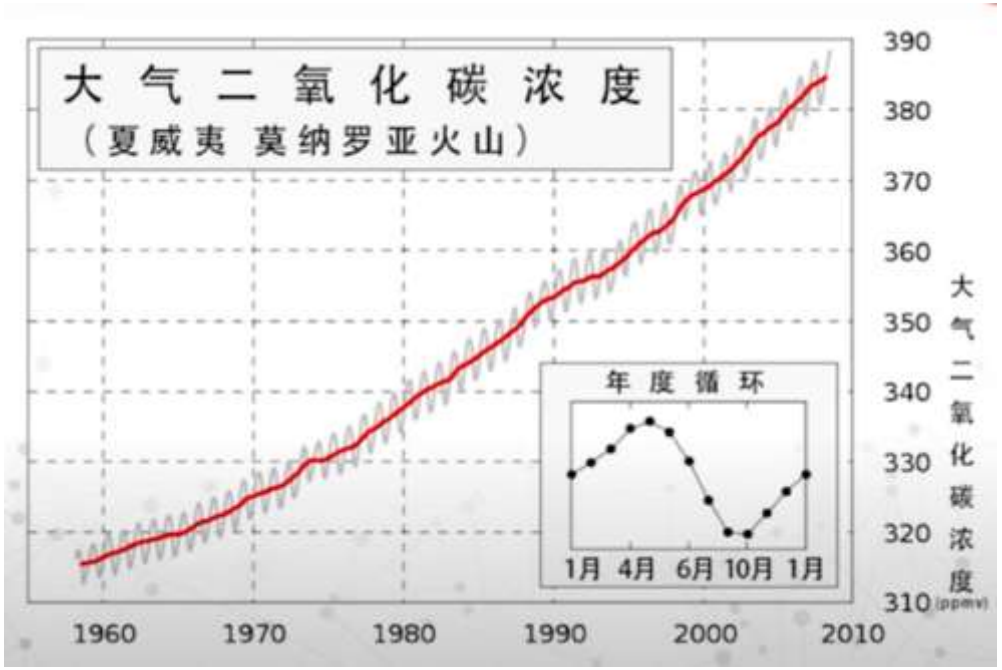

Histogram
Number of census tracts
Normalized number of households
0
50
100
150
200
250
300
350
400
0.02
0.04
0.06
0.08
0.1
0.12
0.14
1


Histogram with fitted Gaussian
Number of census tracts
Normalized number of households
0
10
20
30
40
50
60
70
80
90
100
-12
-10
-8
-6
-4
-2
0
3

曲线拟合/数据光滑化

- 将数据转化成平滑连续的曲线
  - 将数据从“微小的细节”中转移到“更高层面的趋势观察和判断
- 目的：展示数据趋势





曲线拟合

- 不同的拟合方式：表达并观测趋势（*劫富济贫*）
  - 线性回归

$$\min_{\vec{x}} \sum_{i=1}^{n} (y_m - y_i)^2.$$

    - PLSR（partial least squares regression，偏最小二乘拟合）
    - LOESS（Locally weighted scatterplot smoothing）
  - 指数函数
  - 多项式曲线
  - 自定义方程

统计采样

什么是统计采样？

从统计分布中选出的样本

用于近似原分布中的特征

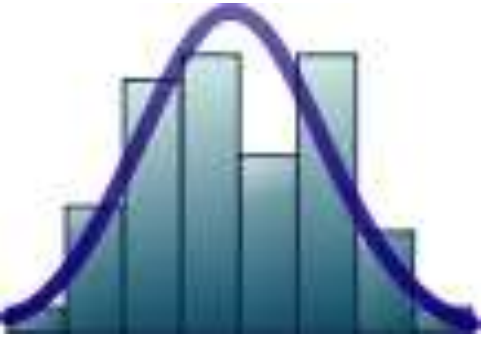
影响采样的因素

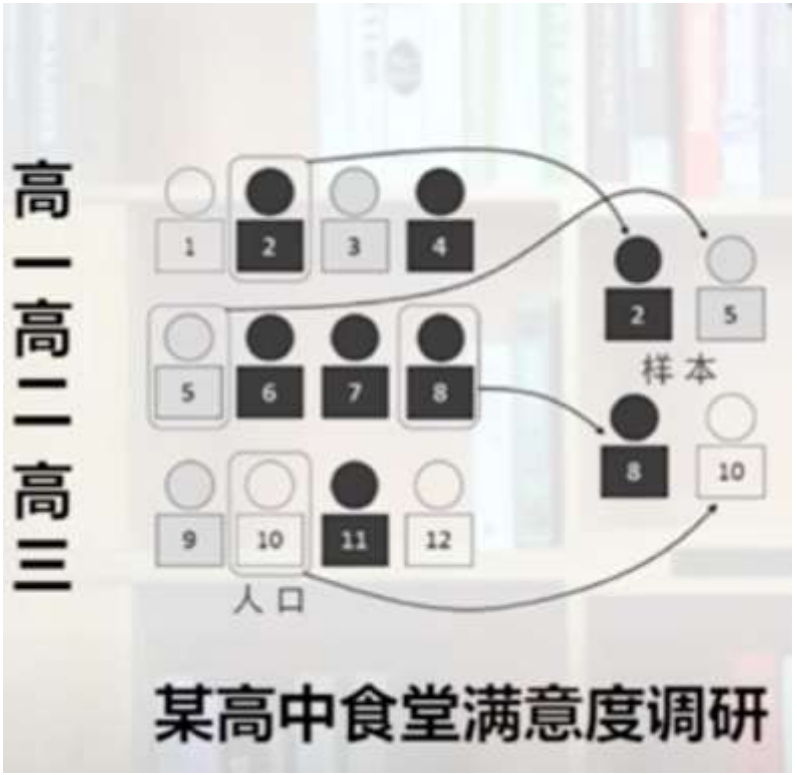
分布本身的特性

数据的测量精度

是否需要分析样本细节（样本精细度）

采样成本



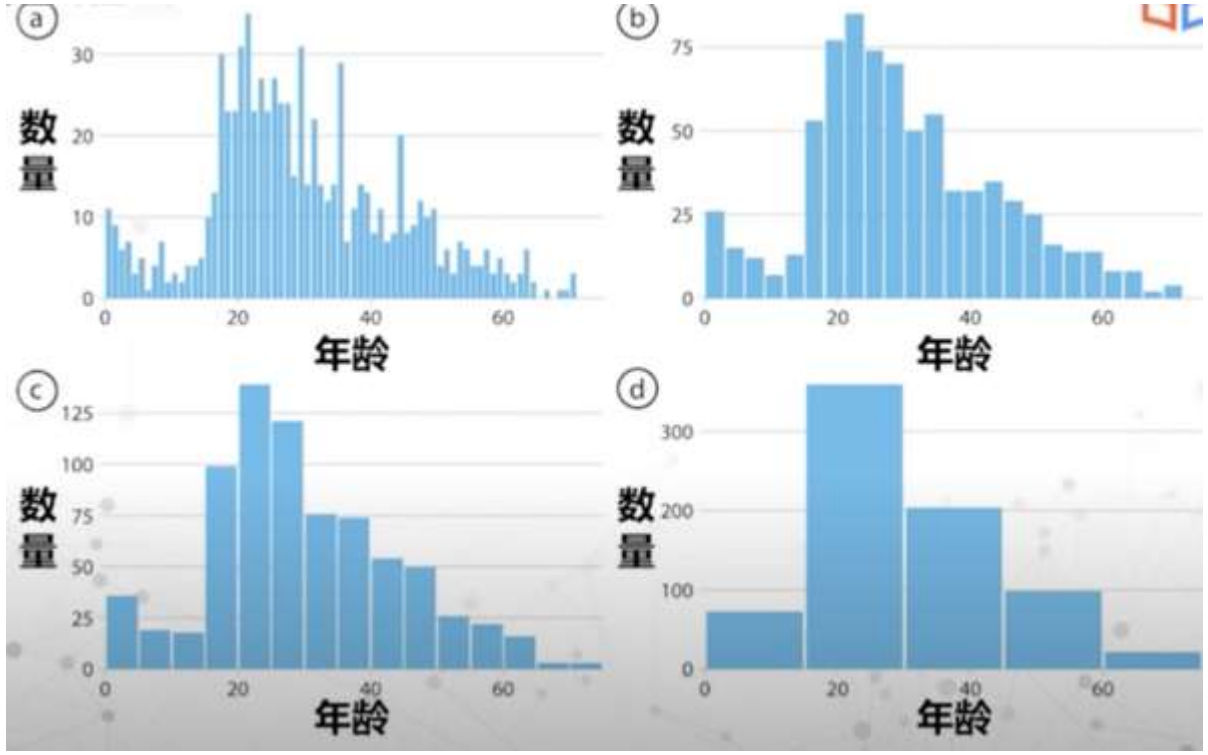


分箱

- 连续数据离散化
  - 将一些连续值分组装进一些“小箱子”的方法

- 单一维度 ->分箱
- 多个维度

数据降维

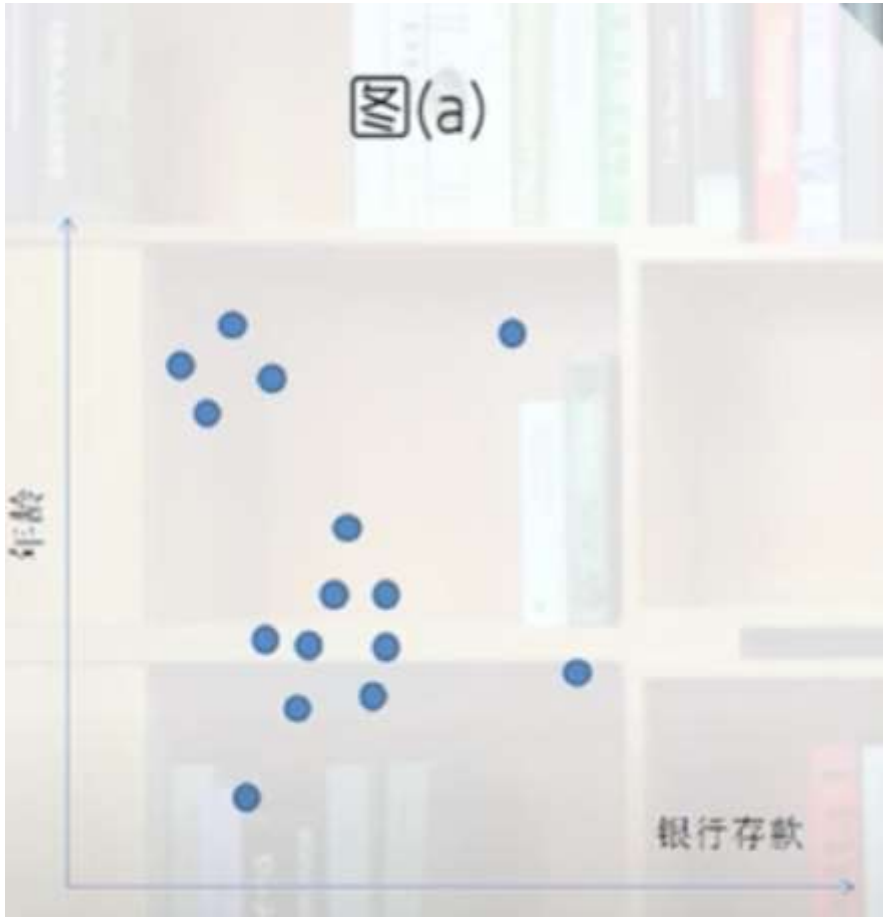


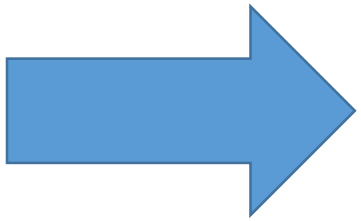
数据降维*

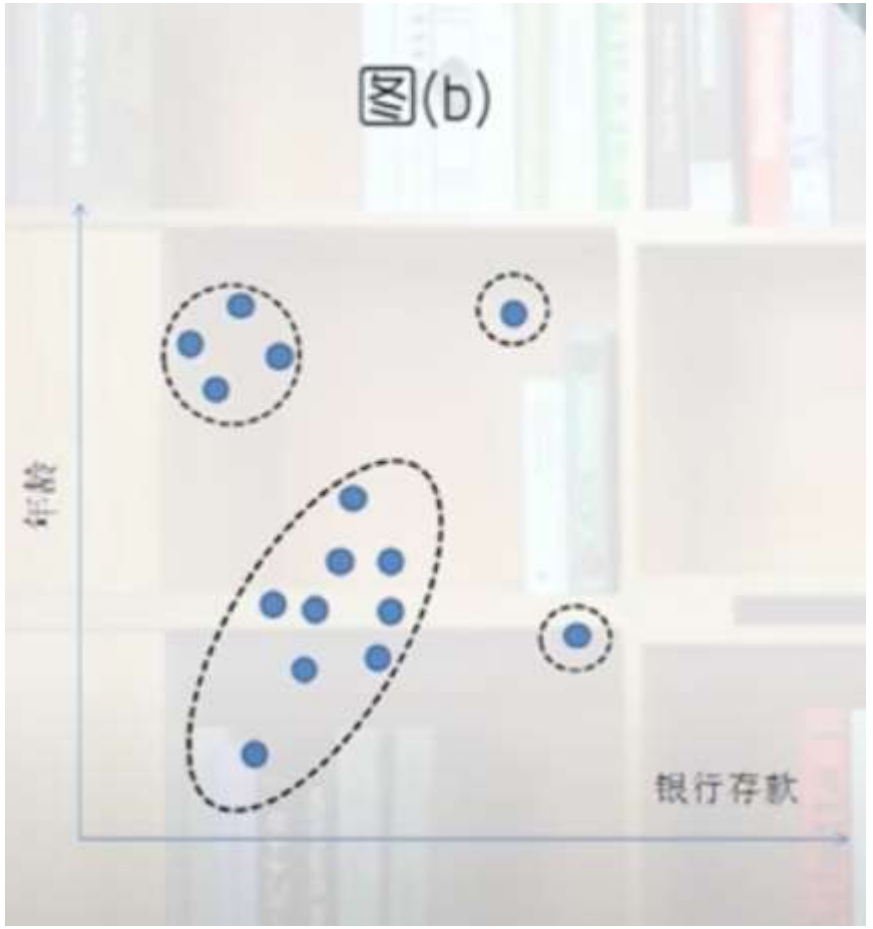
把数据从多维的空间投影到二维或者三维的空间

- 线性方法，如：
  - 主元分析（Principal Components Analysis，PCA）
  - 多维尺度标记（Multidimensional Scaling，MDS）
- 非线性方法，如：
  - T分布随即近邻潜入（t-SNE）
  - 自组织网络（Self-Organizing Map，SOM）
  - 等距特征映射（ISOMAP）

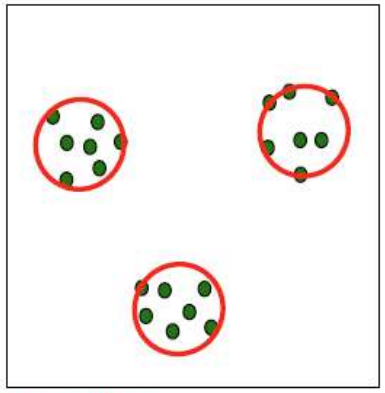
数据聚类



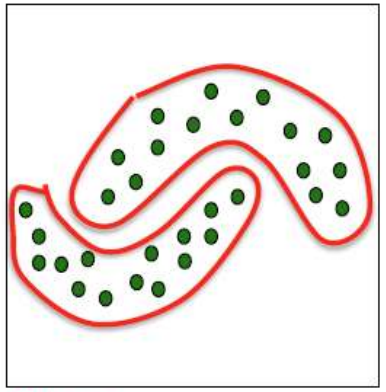




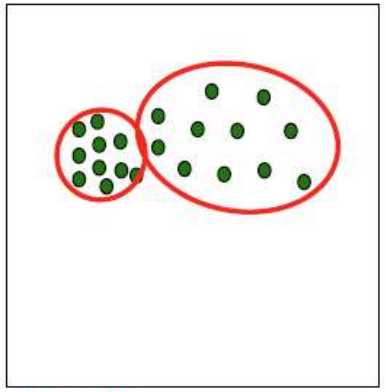
聚类方法



Location



Shape



Density

聚类方法：

K-means聚类

Expectation-Maximization Clustering (EM) *

Gaussian Mixture Model (GMM)*

Spectral Clustering*

Hierarchical Clustering*

K-Means聚类

- K-means

  随机产生K个中心位置

  将每个数据点归为距离最近的中心位置所属的类

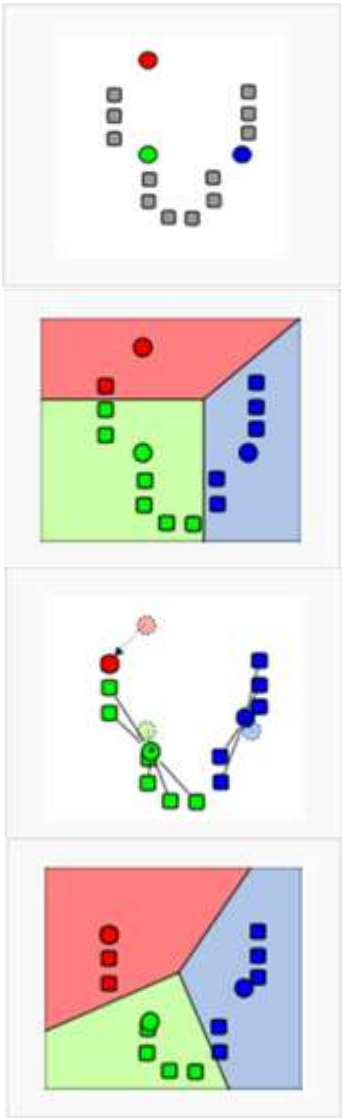
  根据新的类别划分重新计算中心位置

  回到第二步，直到满足一定约束

- K-medoids – 改进

  中心位置必须在数据点所在位置上

  中心位置满足“到类内所有数据点的距离之和最小”

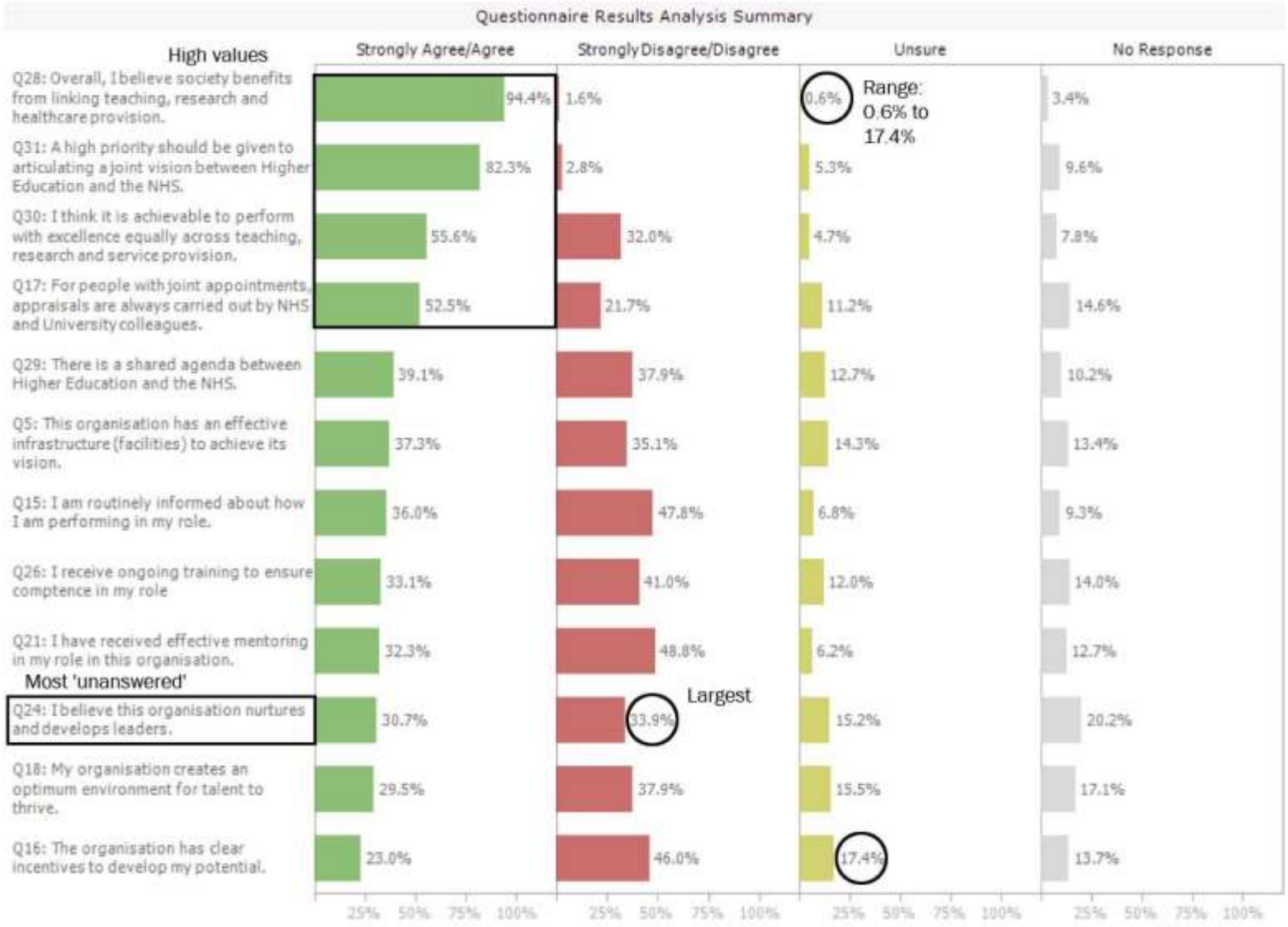


目录

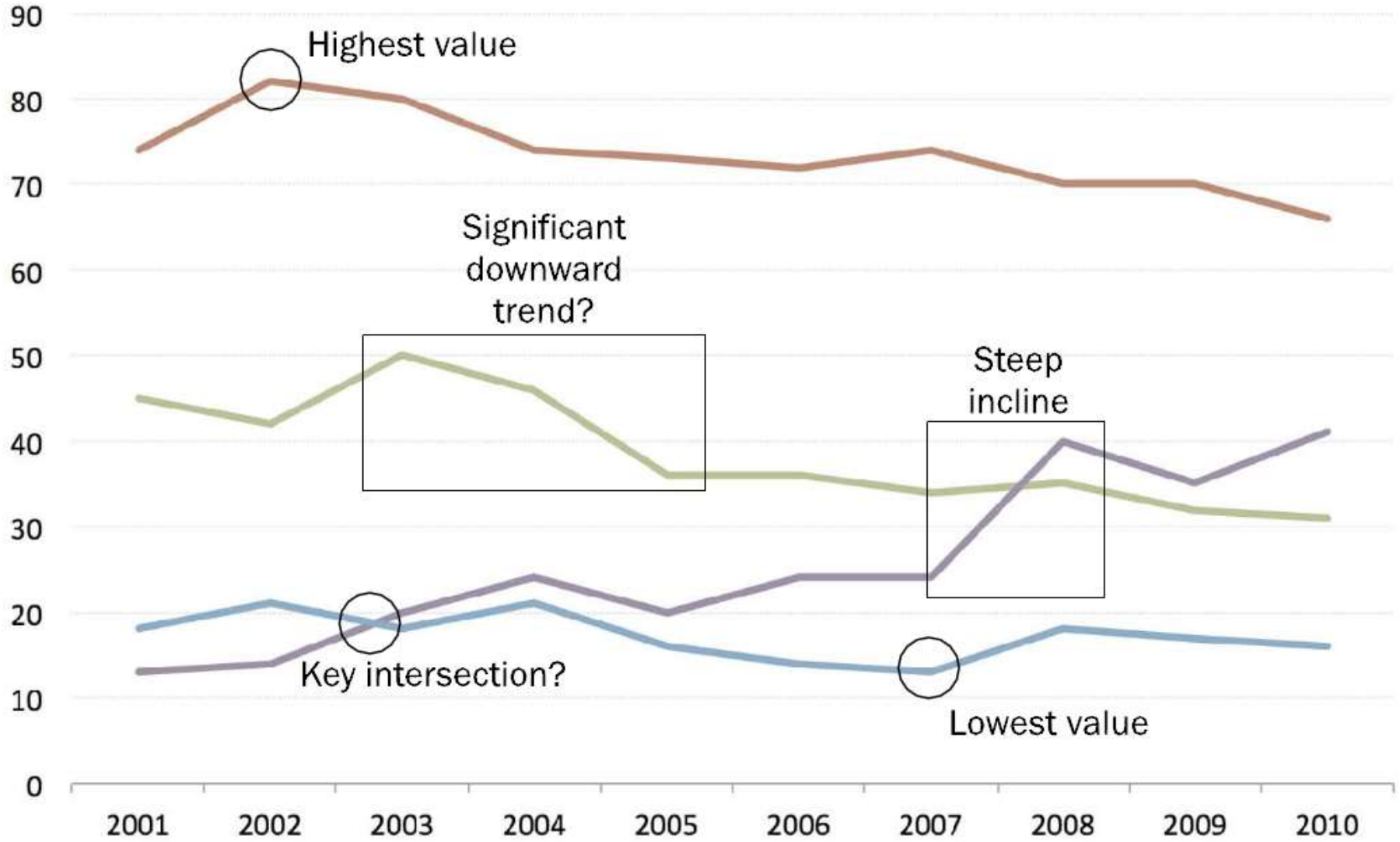
统计图表

- 作用
  - 比较与比例
  - 趋势和模式
  - 关系

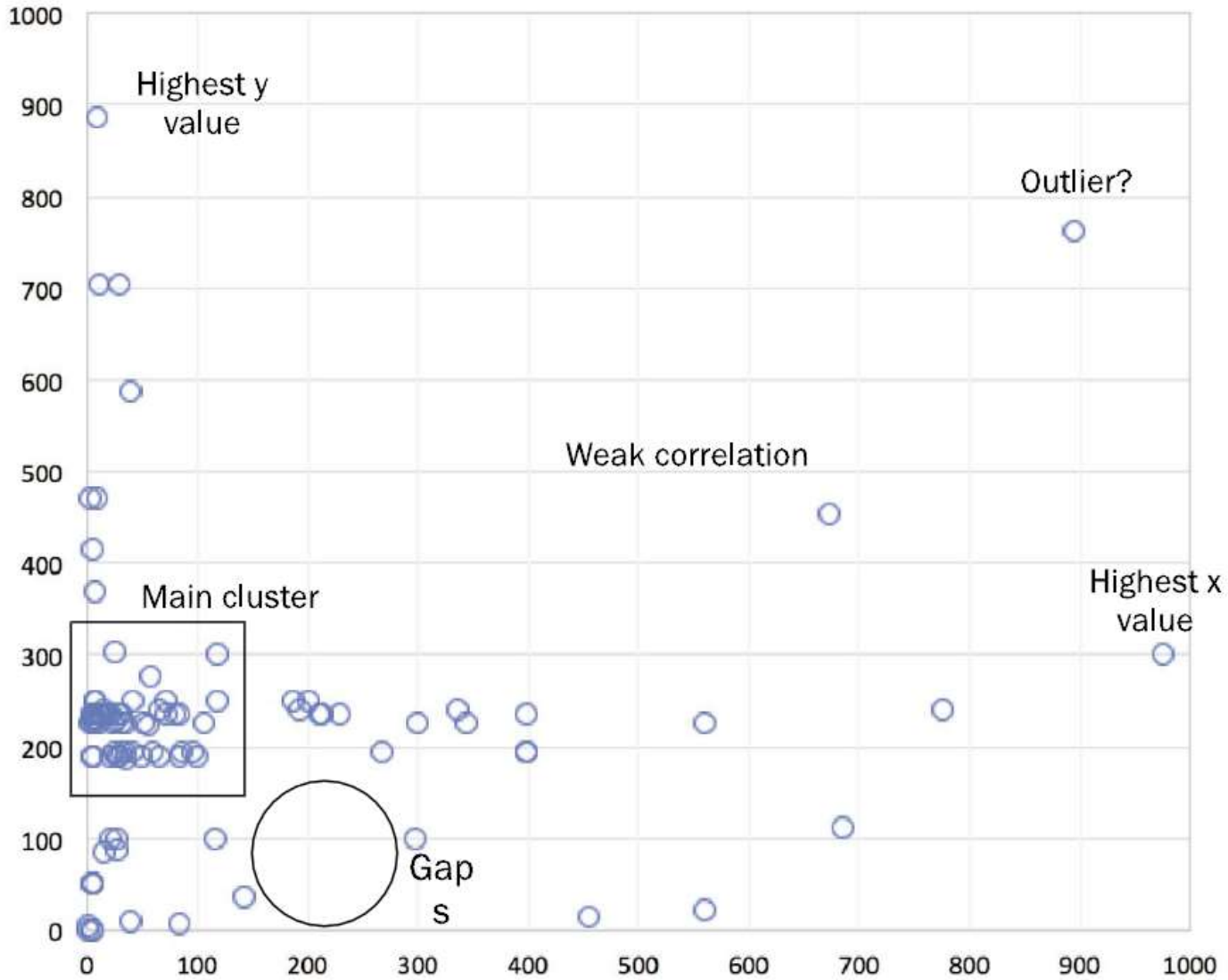
统计图表 – 比较与比例


Questionnaire Results Analysis Summary
High values
Strongly Agree/Agree
Strongly Disagree/Disagree
Unsure
No Response
Q28: Overall, I believe society benefits from linking teaching, research and healthcare provision.
94.4%
1.6%
0.6%
Range: 0.6% to 17.4%
3.4%
Q31: A high priority should be given to articulating a joint vision between Higher Education and the NHS.
82.3%
2.8%
5.3%
9.6%
Q30: I think it is achievable to perform with excellence equally across teaching, research and service provision.
55.6%
32.0%
4.7%
7.8%
Q17: For people with joint appointments, appraisals are always carried out by NHS and University colleagues.
52.5%
21.7%
11.2%
14.6%
Q29: There is a shared agenda between Higher Education and the NHS.
39.1%
37.9%
12.7%
10.2%
Q5: This organisation has an effective infrastructure (facilities) to achieve its vision.
37.3%
35.1%
14.3%
13.4%
Q15: I am routinely informed about how I am performing in my role.
36.0%
47.8%
6.8%
9.3%
Q26: I receive ongoing training to ensure comptence in my role
33.1%
41.0%
12.0%
14.0%
Q21: I have received effective mentoring in my role in this organisation.
32.3%
48.8%
6.2%
12.7%
Most 'unanswered'
Largest
Q24: I believe this organisation nurtures and develops leaders.
30.7%
33.9%
15.2%
20.2%
Q18: My organisation creates an optimum environment for talent to thrive.
29.5%
37.9%
15.5%
17.1%
Q16: The organisation has clear incentives to develop my potential.
23.0%
46.0%
17.4%
13.7%
25% 50% 75% 100%

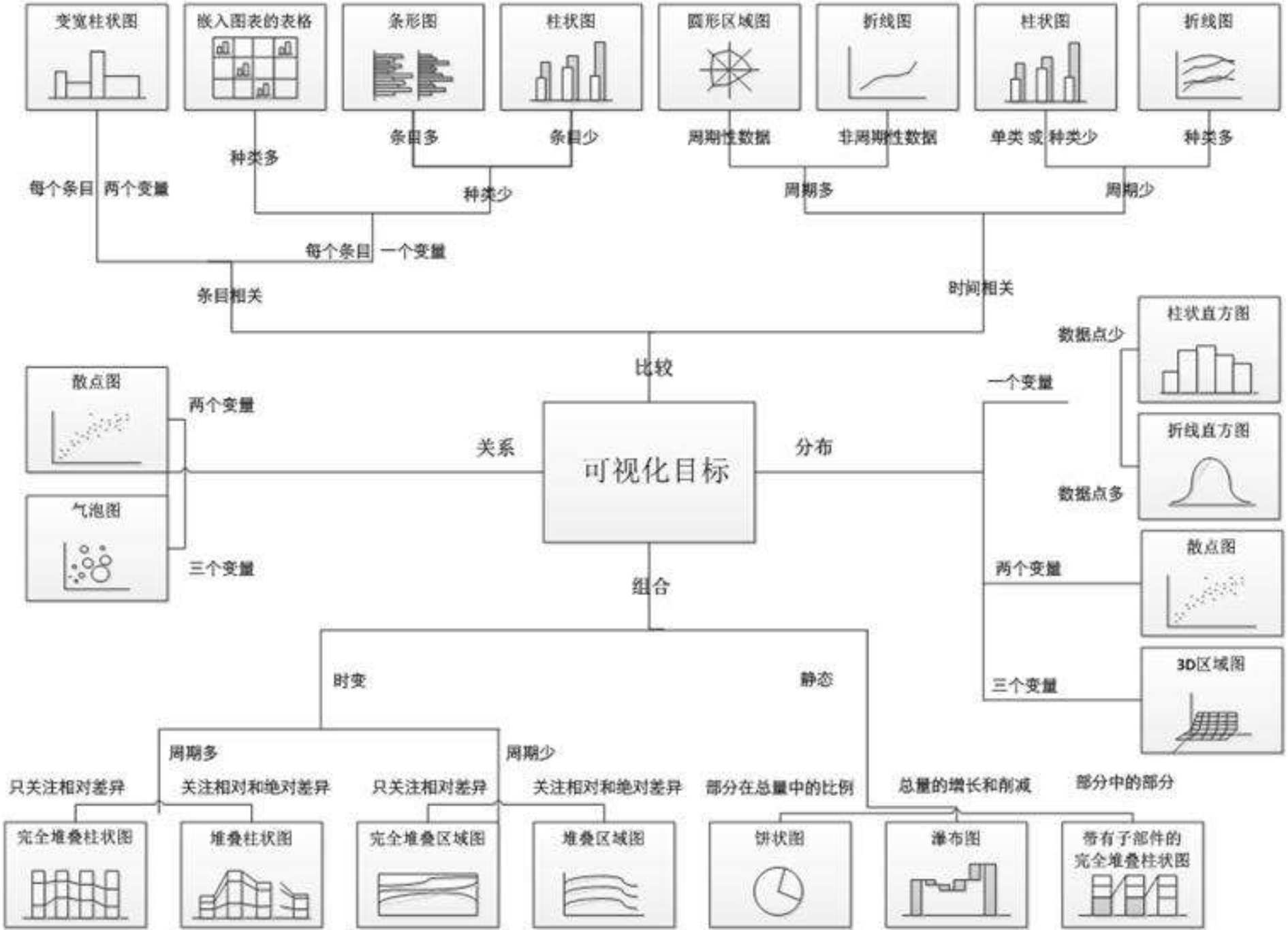
统计图表 - 趋势与模式


90
80
70
60
50
40
30
20
10
0
Highest value
Significant downward trend?
Steep incline
Key intersection?
Lowest value
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

统计图表 - 关系

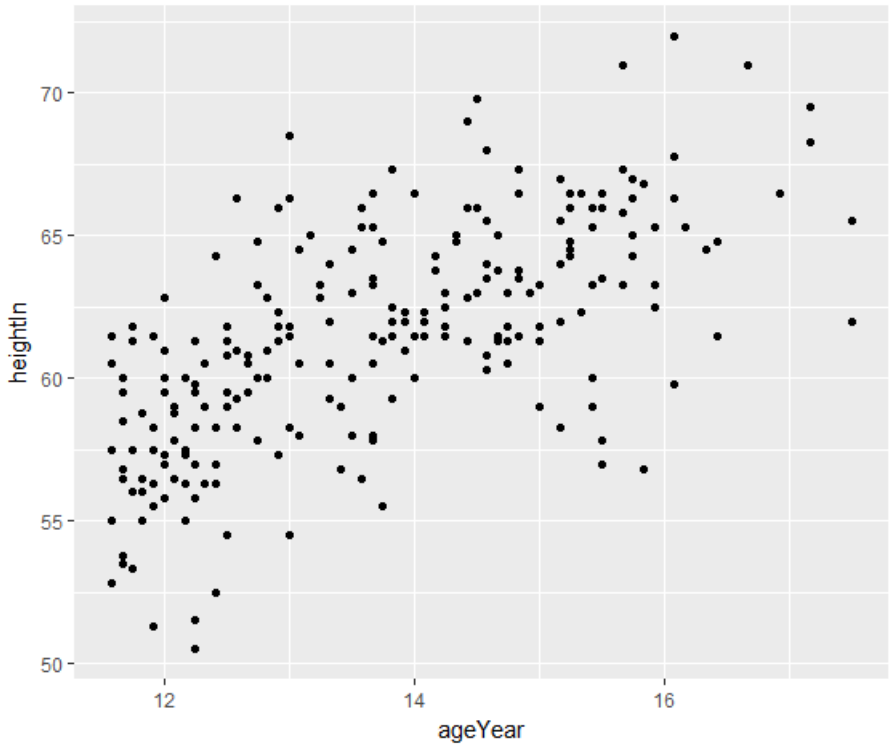

1000
900
800
700
600
500
400
300
200
100
0
0
100
200
300
400
500
600
700
800
900
1000
Highest y value
Outlier?
Weak correlation
Main cluster
Highest x value
Gap
s

统计图表


可视化目标
比较
关系
分布
组合
条目相关
时间相关
每个条目 两个变量
每个条目 一个变量
种类多
种类少
条目多
条目少
周期性数据
非周期性数据
周期多
周期少
单类 或 种类少
种类多
散点图
两个变量
气泡图
三个变量
一个变量
数据点少
柱状直方图
数据点多
折线直方图
两个变量
散点图
三个变量
3D区域图
时变
静态
周期多
周期少
只关注相对差异
关注相对和绝对差异
只关注相对差异
关注相对和绝对差异
部分在总量中的比例
总量的增长和削减
部分中的部分
完全堆叠柱状图
堆叠柱状图
完全堆叠区域图
堆叠区域图
饼状图
瀑布图
带有子部件的
完全堆叠柱状图

关系：散点图

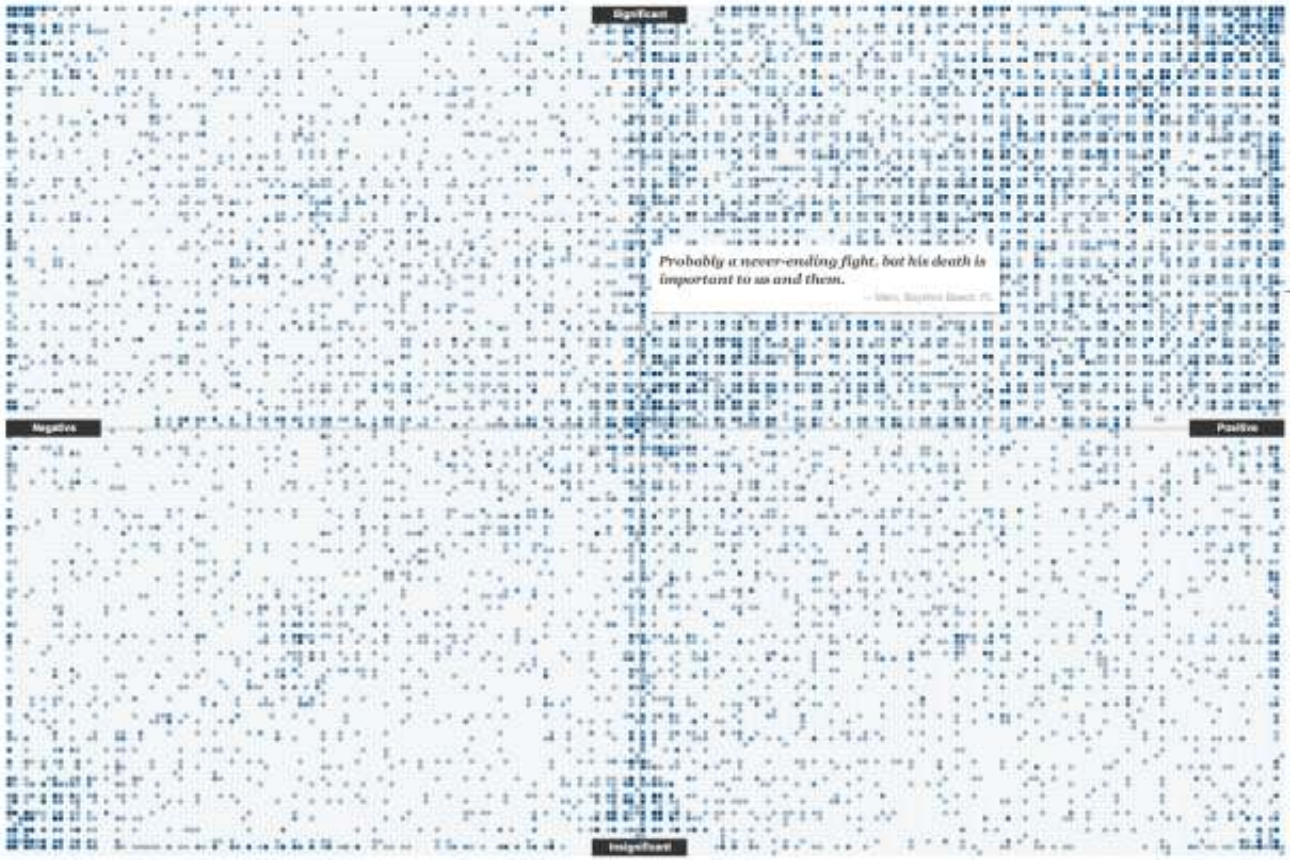
- 散点图是表示二维数据的标准方法，探讨变量之间关系



本拉登之死

意义重大

坏事



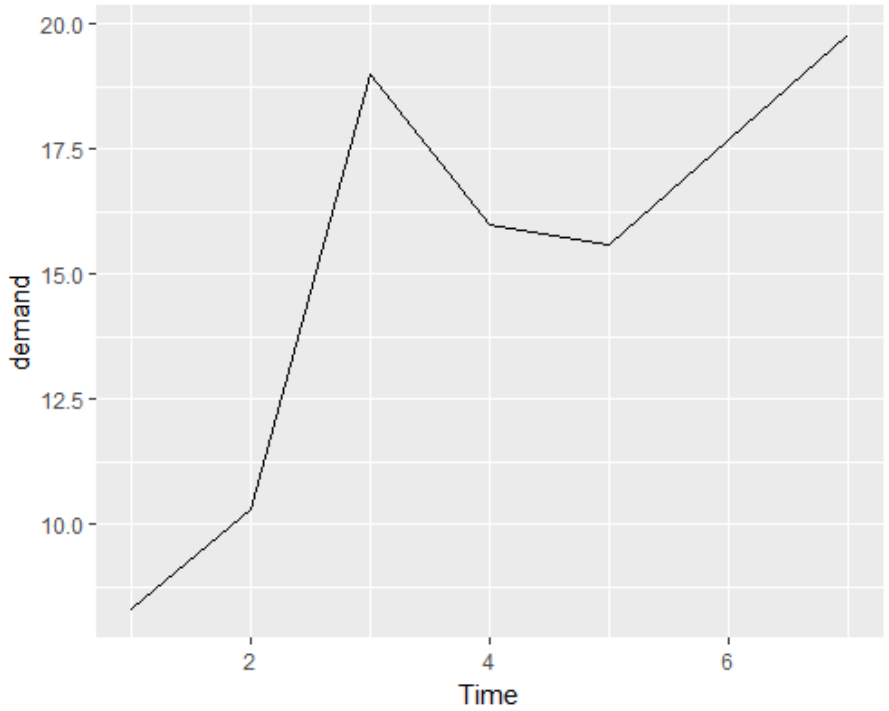
好事

毫无意义

http://www.nytimes.com/interactive/2011/05/03/us/20110503-osama-response.html

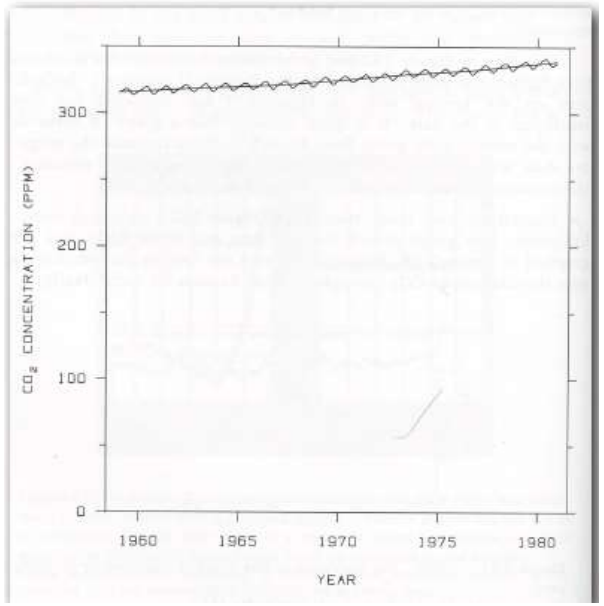
比较：折线图

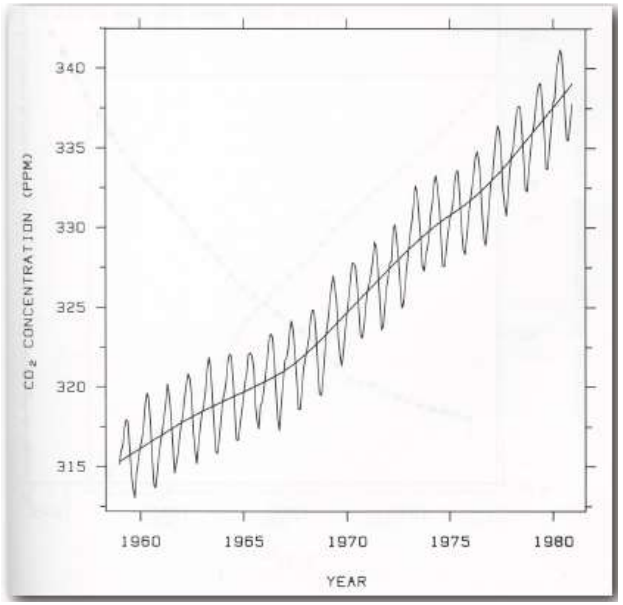
- 使用直线段来连接一系列点



比较：折线图

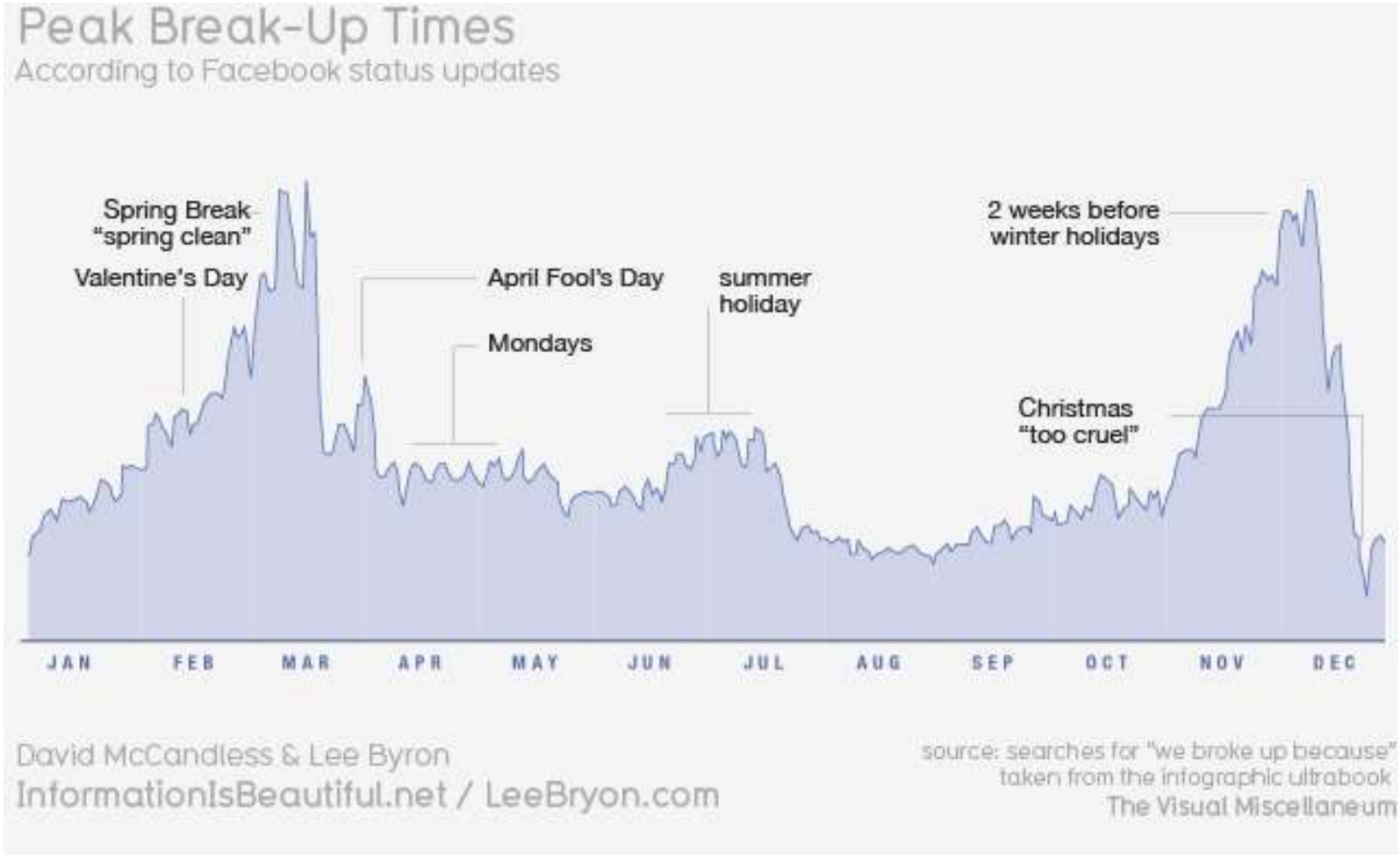
- 宽高比



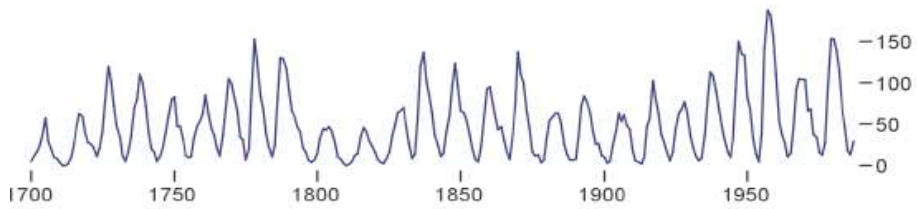


Multi-Scale Banking to 45º, 宽高比：左7.87，右1.17

折线图


Peak Break-Up Times
According to Facebook status updates
Spring Break
"spring clean"
Valentine's Day
April Fool's Day
Mondays
summer
holiday
2 weeks before
winter holidays
Christmas
"too cruel"
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
David McCandless & Lee Byron
InformationIsBeautiful.net / LeeBryon.com
source: searches for "we broke up because"
taken from the infographic ultrabook
The Visual Miscellaneum

统计图对比



折线图

重量级
同时表达数据走势和分布

Sparkline

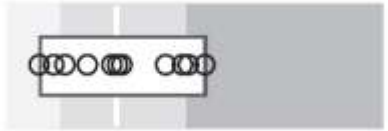
轻量级
只表达数据走势

By Edward Tufte

Bandline

数据走势 + 模糊分布

数据分布



中量级？
表达数据走势和分布

By Stephen Few

比较：柱状图/条形图

- 采用长方形的形状和颜色编码数据的属性

每个国家消费了多少啤酒？

每人每周消耗的瓶数

柱状图

- 注意事项
  - 尺度
  - 使用零点为基准点
  - 不必要的三维设计

堆叠柱状图

- 每根直柱内部也可用像素图方式编码
  - 分解整体，比较局部信息

每个柱子代表一个整体

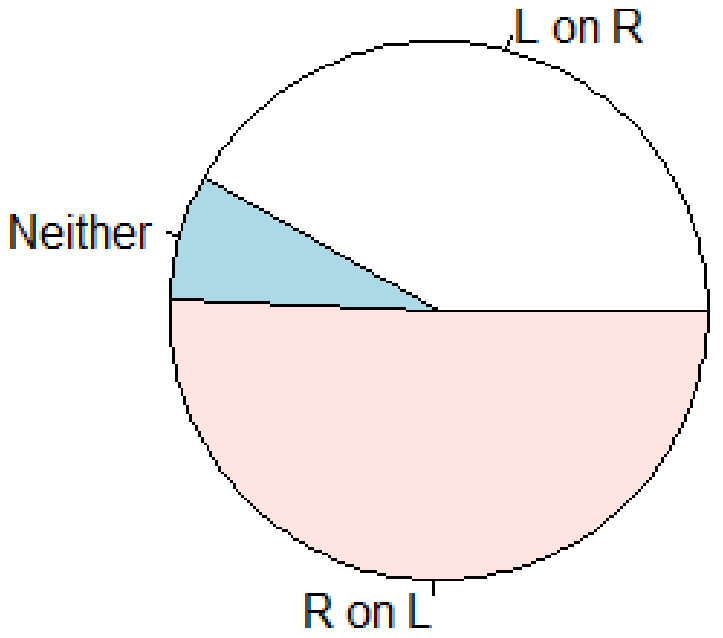
柱子中不同颜色代表不同类别

堆叠柱状图

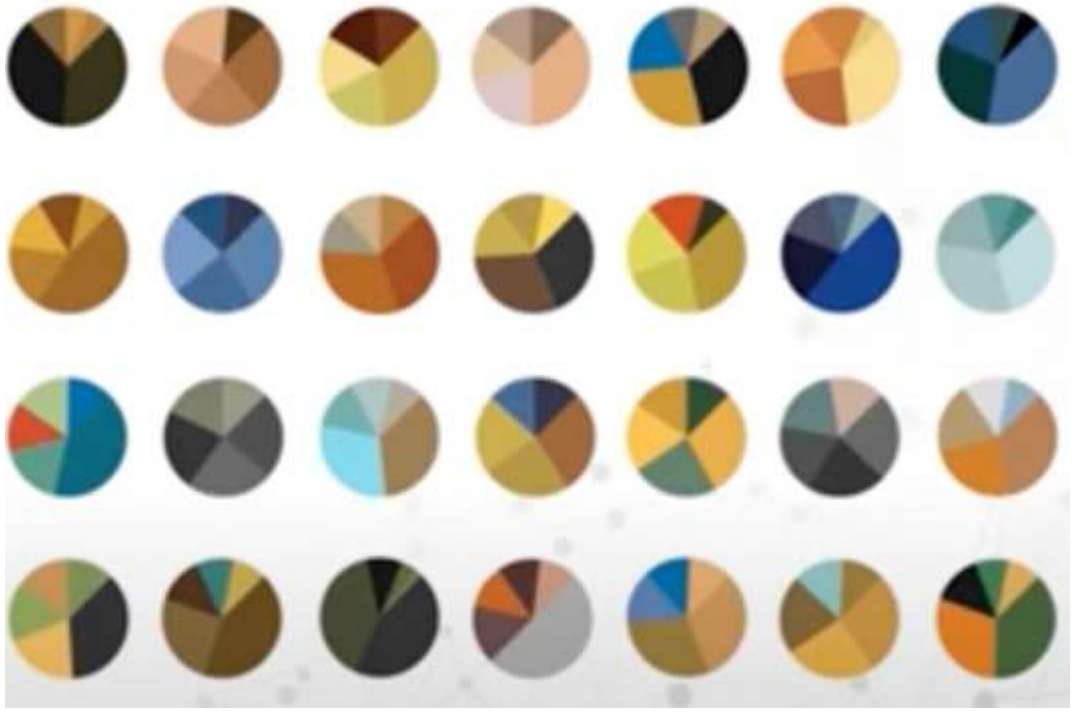

Espresso
Espresso
[ess-press-oh]
Milk Foam
Espresso
Espresso Macchiato
[ess-press-oh mock-e-ah-toe]
Whipped Cream
Espresso
Espresso con Panna
[ess-press-oh kon pawn-nah]
Milk Foam
Steamed Milk
Espresso
Caffé Latte
[caf-ay lah-tey]
Steamed Milk
Espresso
Flat White
Milk Foam
Steamed Half-And-Half
Espresso
Cafe Breve
[caf-ay brev-ay]
Milk Foam
Steamed Milk
Espresso
Cappuccino
[kapp-oo-chee-noh]
Whipped Cream
Steamed Milk
Chocolate Syrup
Espresso
Caffé Mocha
[caf-ay moh-kuh]
Water
Espresso
Americano
[uh-mer-i-kan-oh]

饼图

- 用环状方式呈现各分量在整体中的比例，是环状树图的基础
  - 采用了饼干的隐喻
  - 避免3D

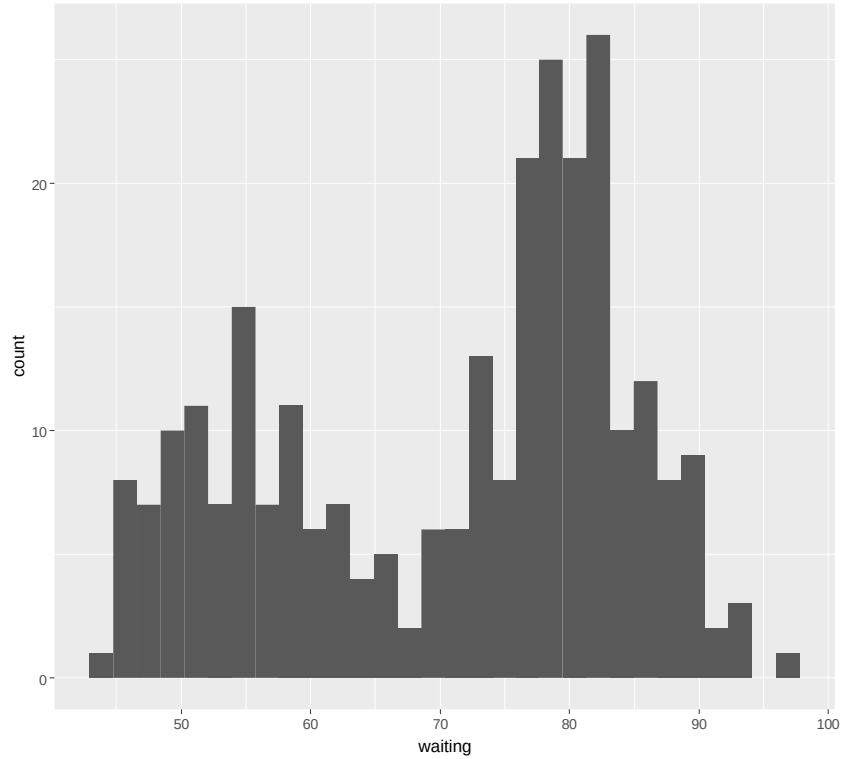


饼图：梵高的作品



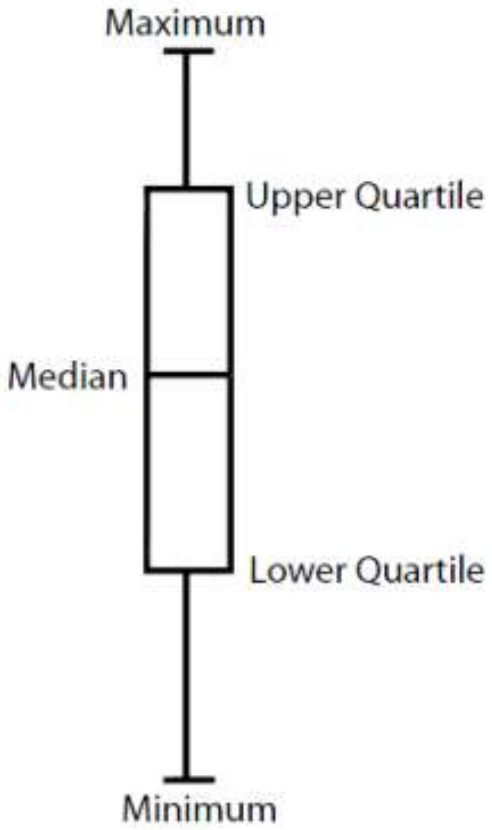
分布：直方图

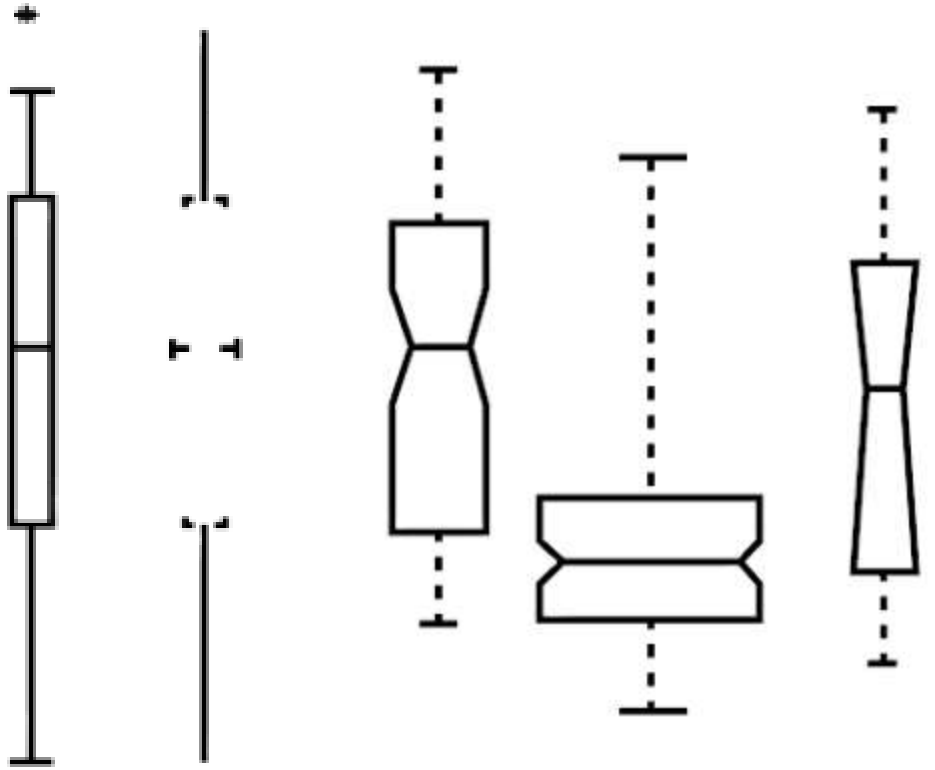
- 对数据集的某个数据属性的频率统计
- 呈现数据分布、离群值和分布的模态



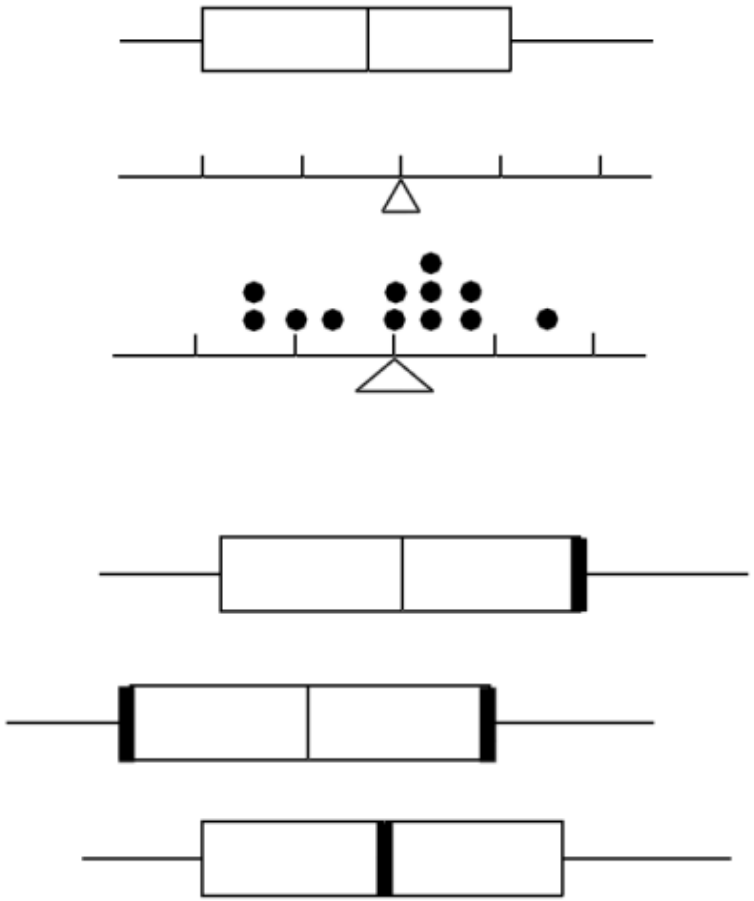
分布：箱线图/盒须图

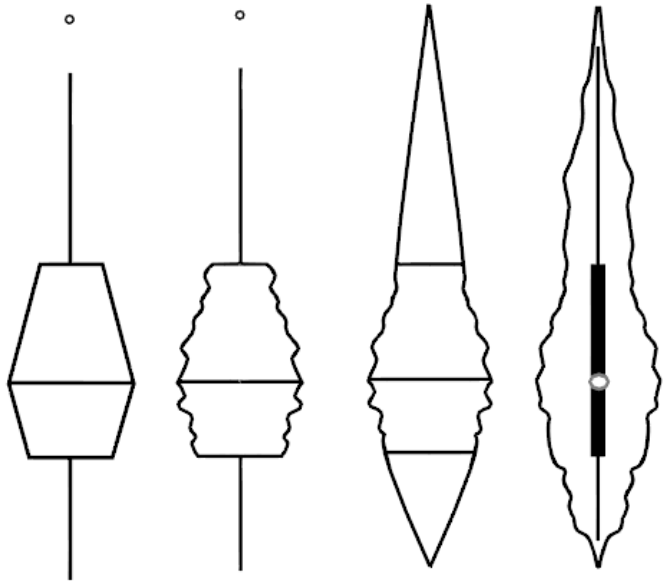
- 也叫盒须图。长方形盒子表示数据的大概范围



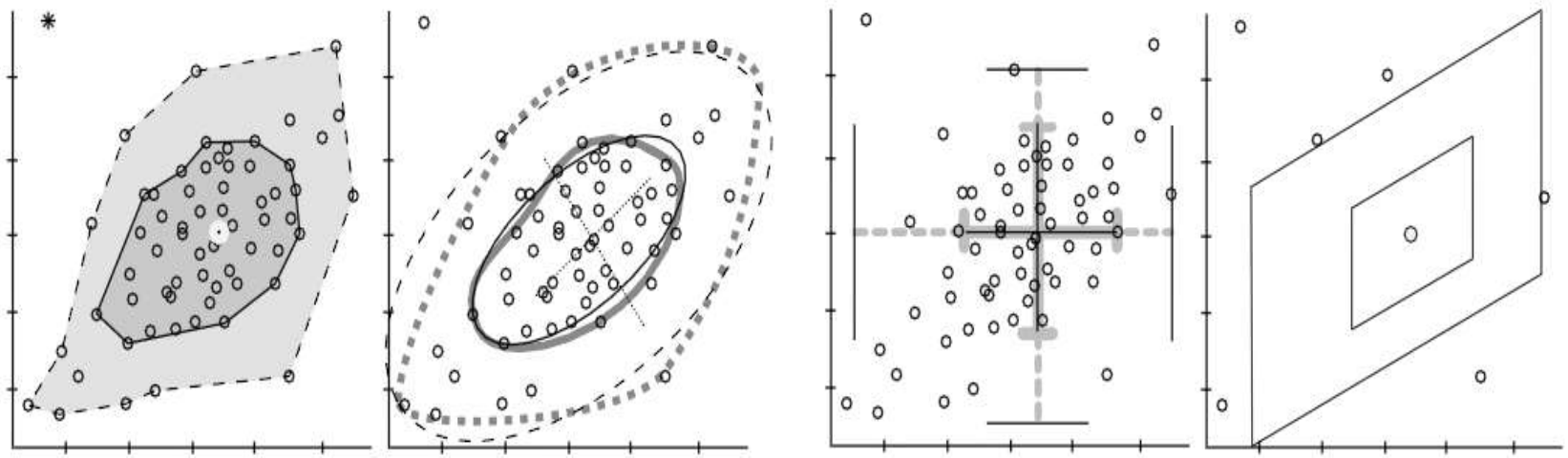


盒须图变种





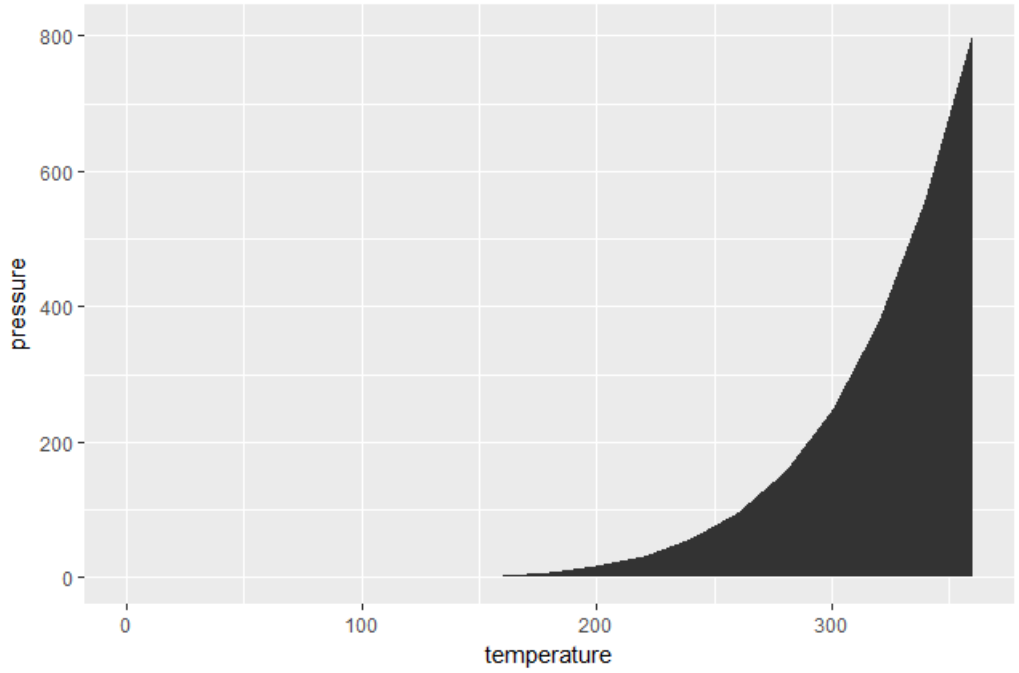
盒须图变种



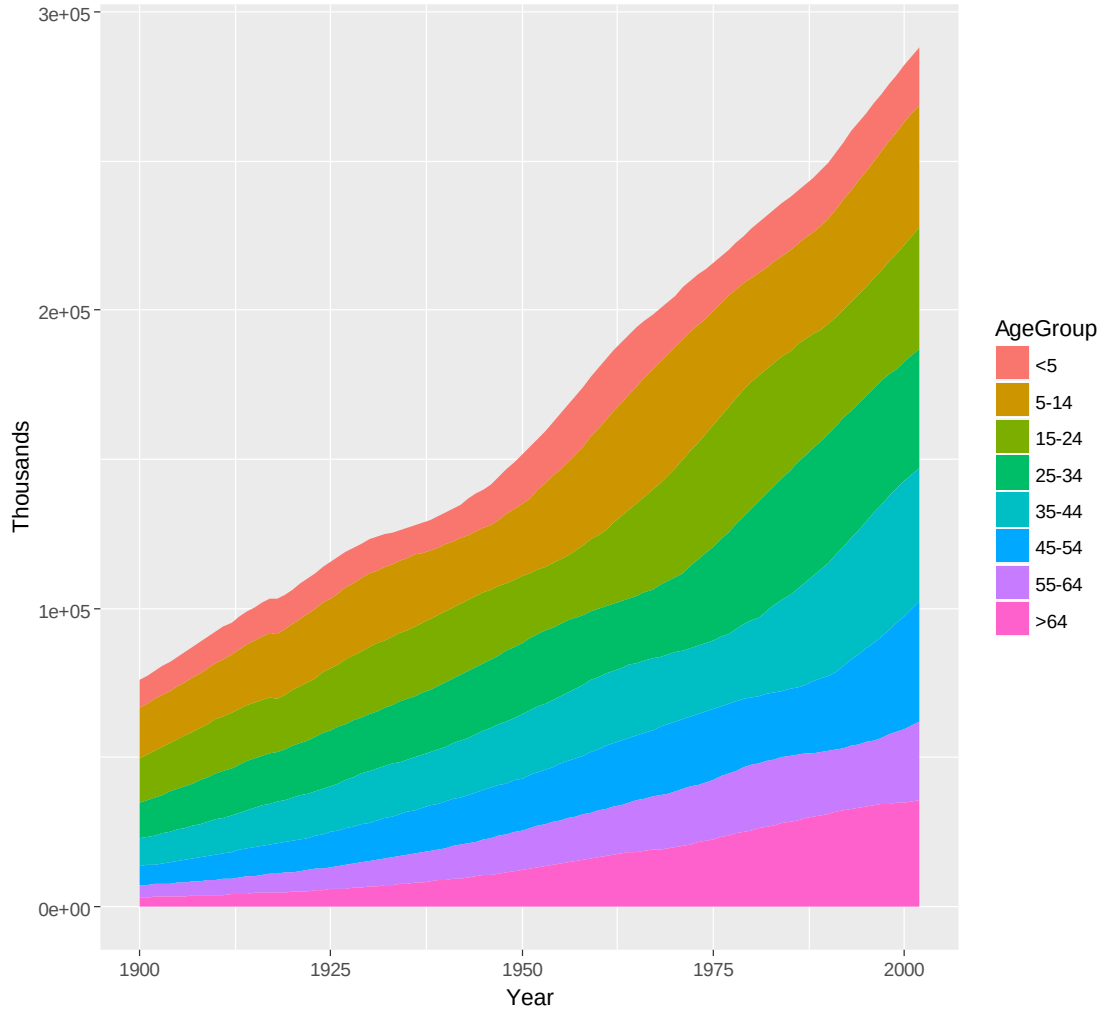
2D Box Plot　　Relplot　　Rangefinder Box Plot　　Bag Plot

面积图

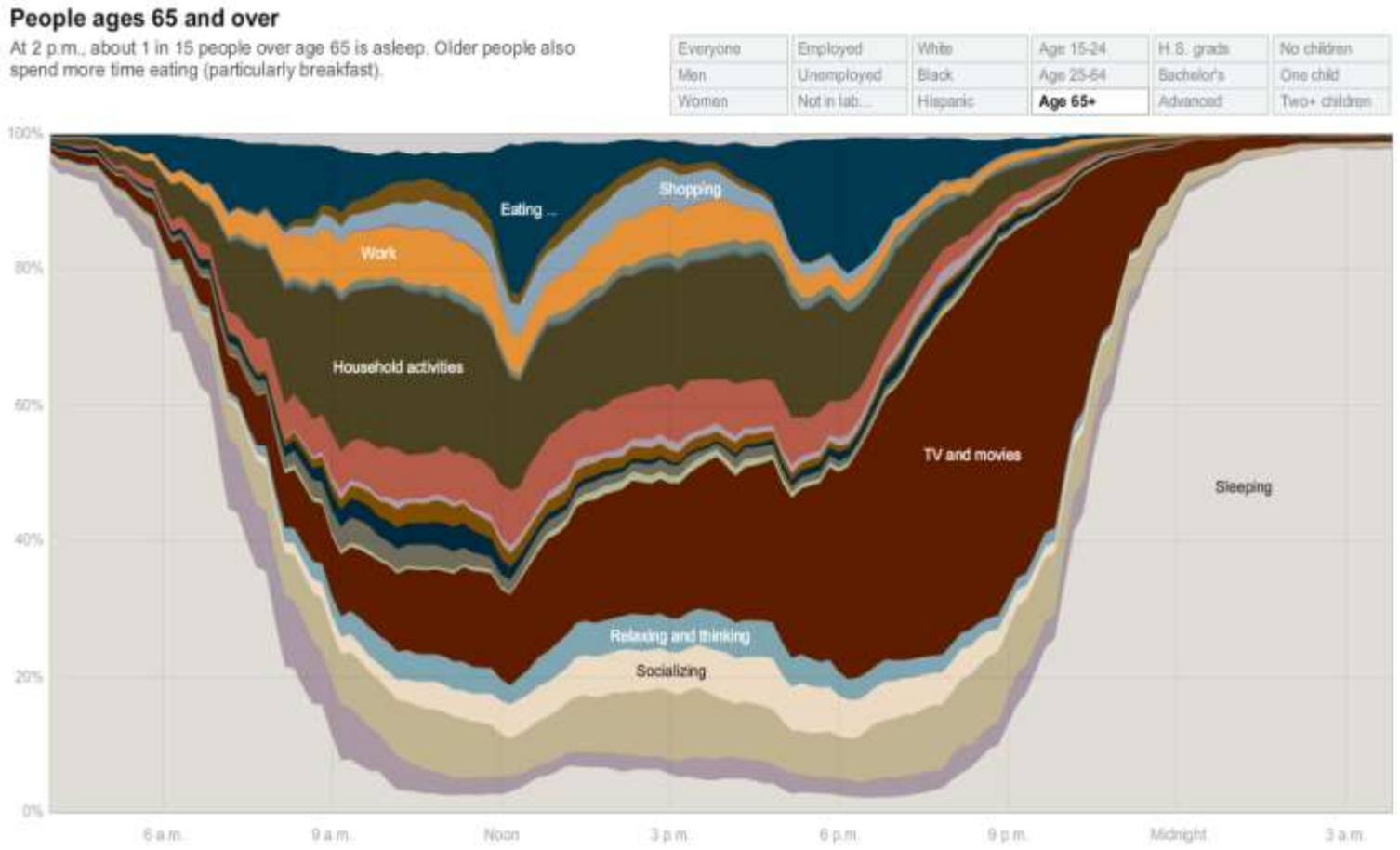
- 强调数据随时间而变化的程度，引起对总值趋势的注意



堆积面积图


3e+05
2e+05
1e+05
0e+00
Thousands
1900
1925
1950
1975
2000
Year
AgeGroup
<5
5-14
15-24
25-34
35-44
45-54
55-64
>64

堆积面积图



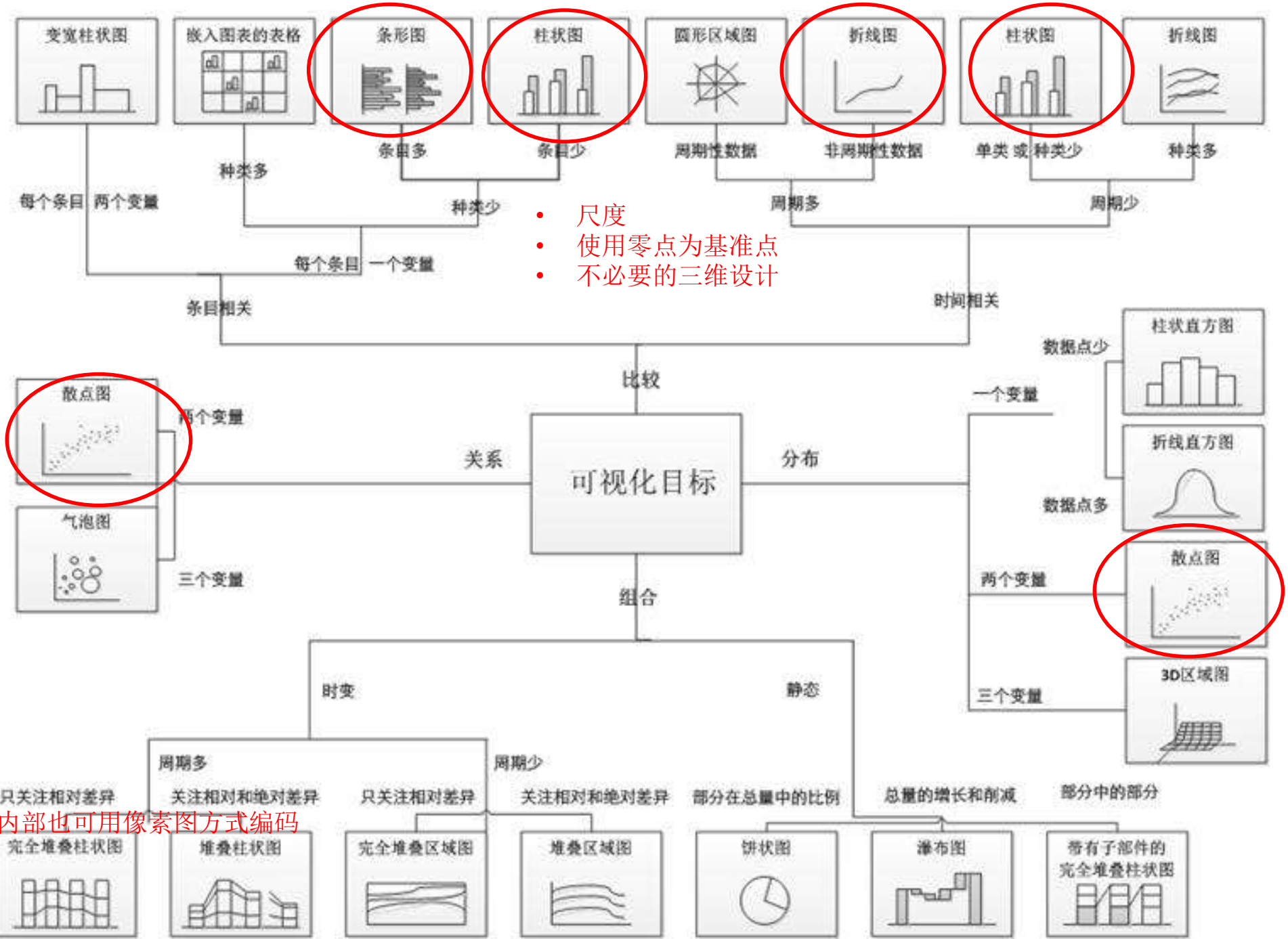
http://www.nytimes.com//interactive/2009/07/31/business/20080801-metrics-graphic.html?hp

交互的数据可视化

- 交互的可视化可以进一步增强用户对于数据的探索和认知，鼓励用户参与
- D3 (*Interactive Data Visualization for the Web)*, Processing.js
- Plotly
- Bokeh…
- https://scnetworkviz.github.io/SCNetworkViz/

统计图表

注意宽高比



- 尺度
- 使用零点为基准点
- 不必要的三维设计

每根直柱内部也可用像素图方式编码

数据变换的目的是简化数据、降低显示数据的尺寸和复杂度

- 归一化
- 曲线拟合
- 统计采样
- 降维，etc.