Data Preprocessing

- 关于数据
- 为什么要预处理数据？
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据规约
- 小结

Motivation

- 动机
  - ✓ 更好的理解数据
  - ✓ 对数据有一个全局了解

- 描述性数据汇总
  - ✓ **Central tendency (集中趋势)**
  - ✓ **Dispersion (散布性)**

度量中心趋势(1)

- **Mean (均值，algebraic measure):**
  - ✓ 算术平均数:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - ✓ 加权算术平均数:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

e.g.薪水和分数

度量中心趋势(2)

- **Median (中位数, holistic measure)**
  - ✓ 奇数个数值的中间那个值，或者是偶数个数值的中间两个值的平均值
    - ***Data*** **57 55 85 24 33 49 94 2 8 51 71 30 91 6 47 50 65 43 41 7**
    - ***Ordered Data***
      - **2 6 7 8 24 30 33 41 43 *47 49* 50 51 55 57 65 71 85 91 94**
    - **Median 48**

- **众数**
  - ✓ 数据中出现频率最高的数值

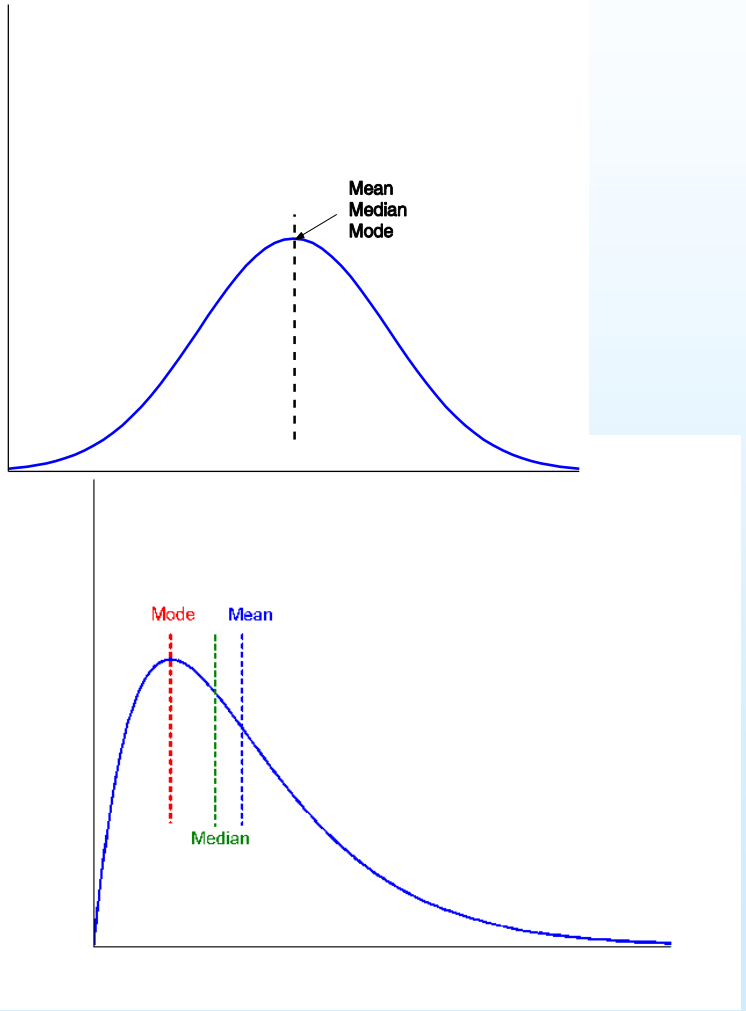
1亿
7000
6000
5000
4000
3000
2000

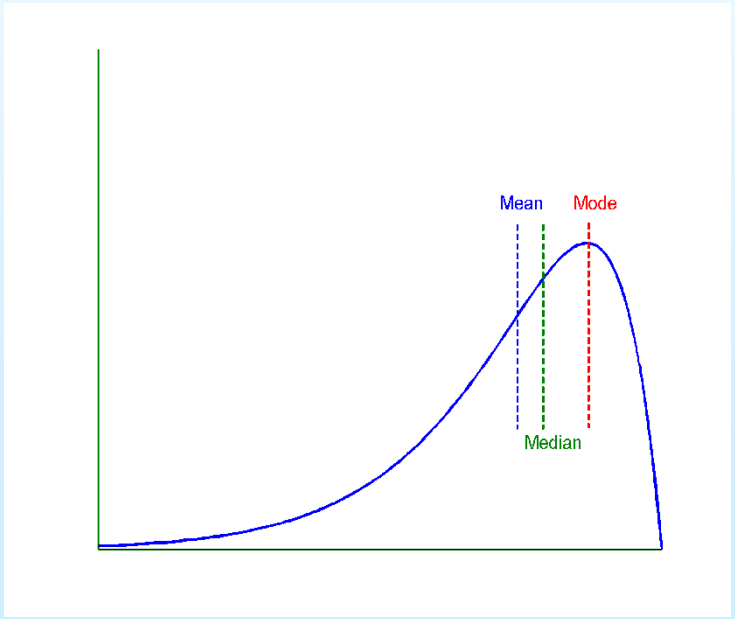
平均数 ←

中位数 ←

对称的数据v.s有偏数据

- 对称数据、左偏数据和右偏数据的中位数、均值和众数

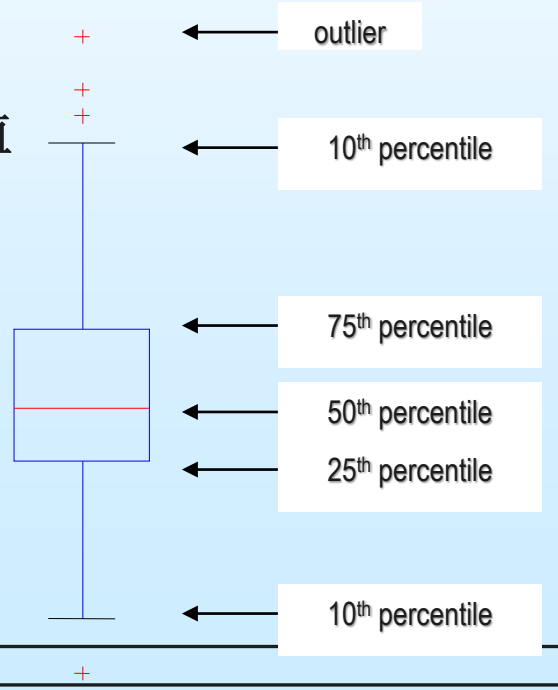




度量数据的离散程度(1)

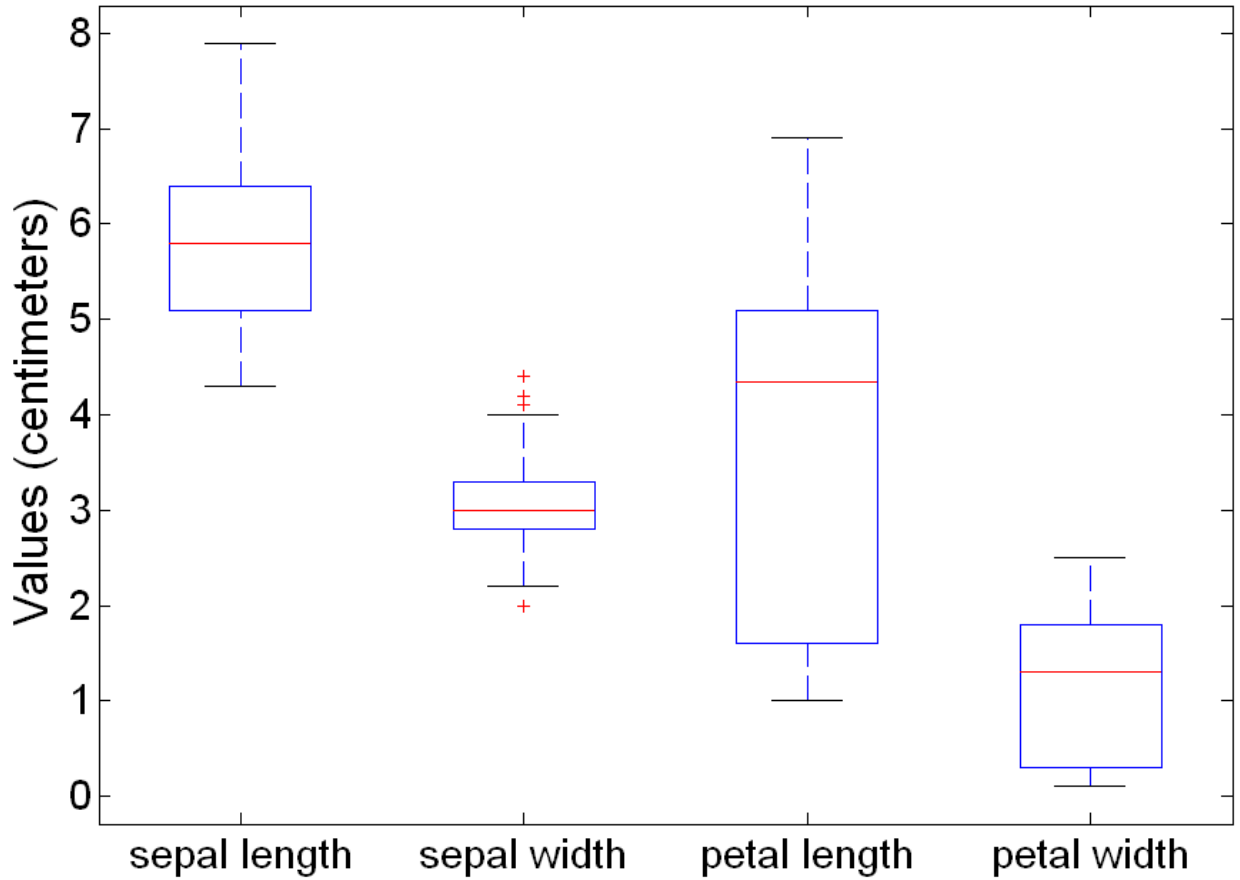
四分位数，离散点和盒图

- ✓ **Quartiles(4分位数)**: Q_1 (25th percentile), Q_3 (75th percentile)
- ✓ **Inter-quartile range（中间四分位数）**: $IQR = Q_3 - Q_1$
- ✓ **Five number summary（五数概括）**: min, Q_1, M, Q_3, max
- ✓ **Boxplot（盒图）**:两端是四分位数，中位数被标记出来，外边界，并且分别绘制出离散点
- ✓ **Outlier**:通常，比**1.5**倍的**IQR**的值高/低的值



度量数据的离散程度(1)

- 盒图可用于比较不同的属性



度量数据的离散程度(2)

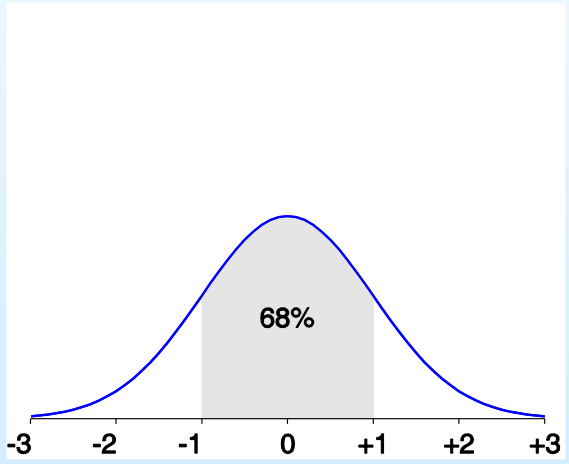
- **Variance（方差） and standard deviation（标准差）**
  - ✓ **Variance s^2: (代数的，可伸缩的计量)**

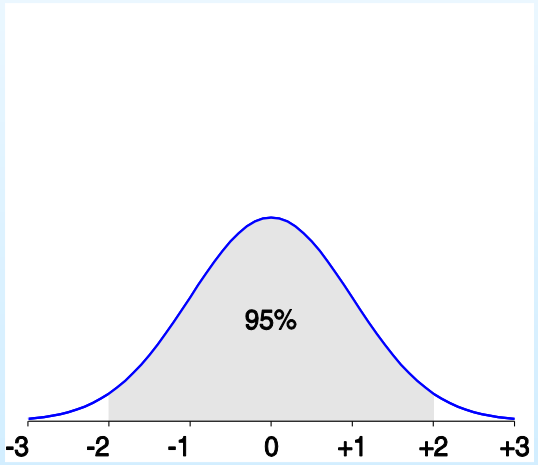
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

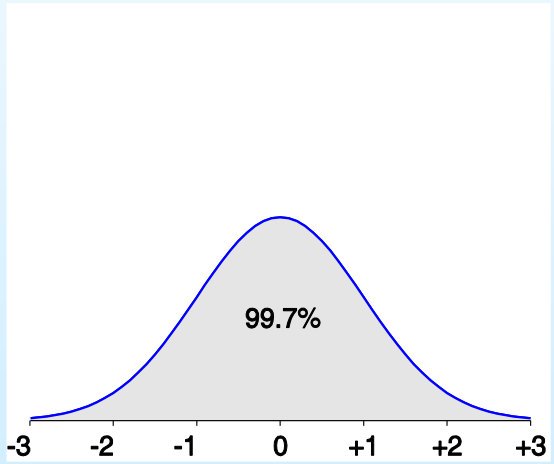
  - ✓ **Standard deviation s 方差s^2的平方根**

正态分布曲线的特性

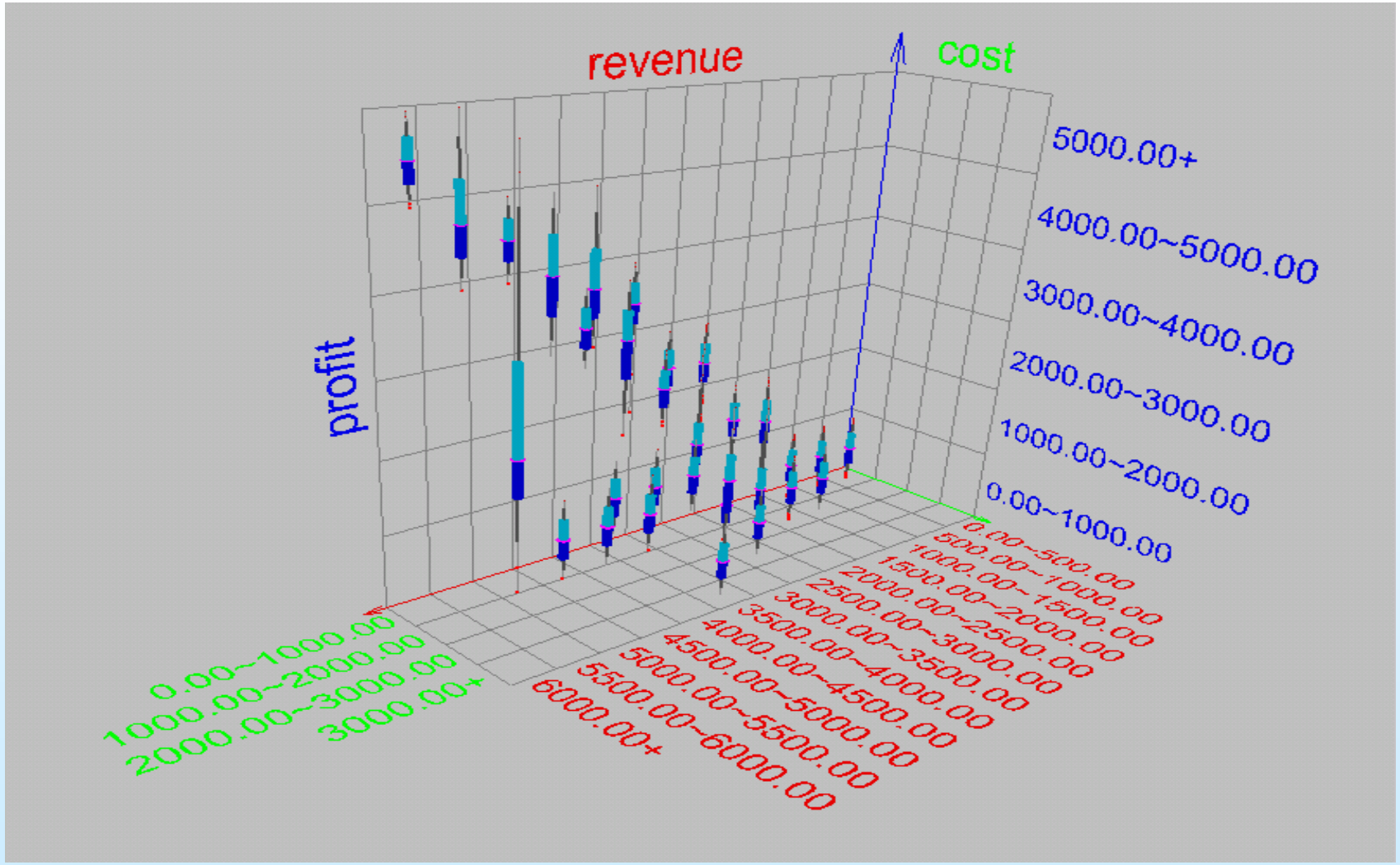
- 正态分布曲线
  - ✓ 从**μ–σ** 到 **μ+σ:**包含了**68%**的测量值（**μ**：均值，**σ**：标准差）
  - ✓ 从**μ–2σ** 到 **μ+2σ:**包含了**95%**的测量值
  - ✓ 从**μ–3σ** 到**μ+3σ:**包含了**99.7%**的测量值







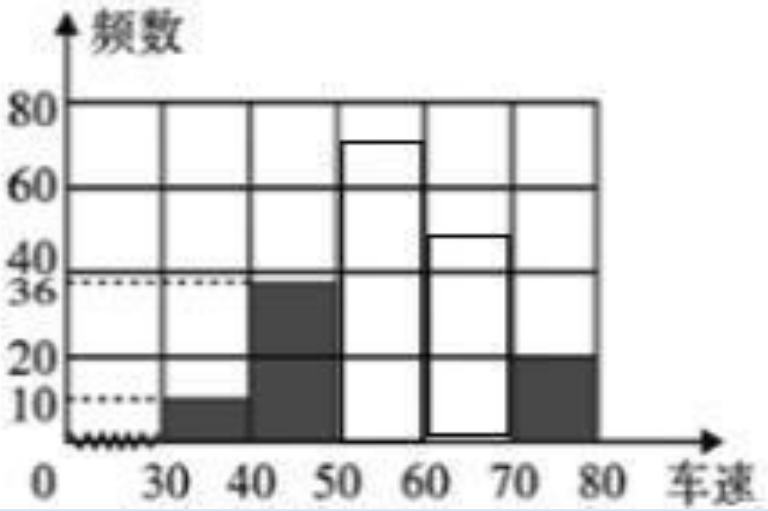
数据散布性的可视化：盒图分析



直方图分析

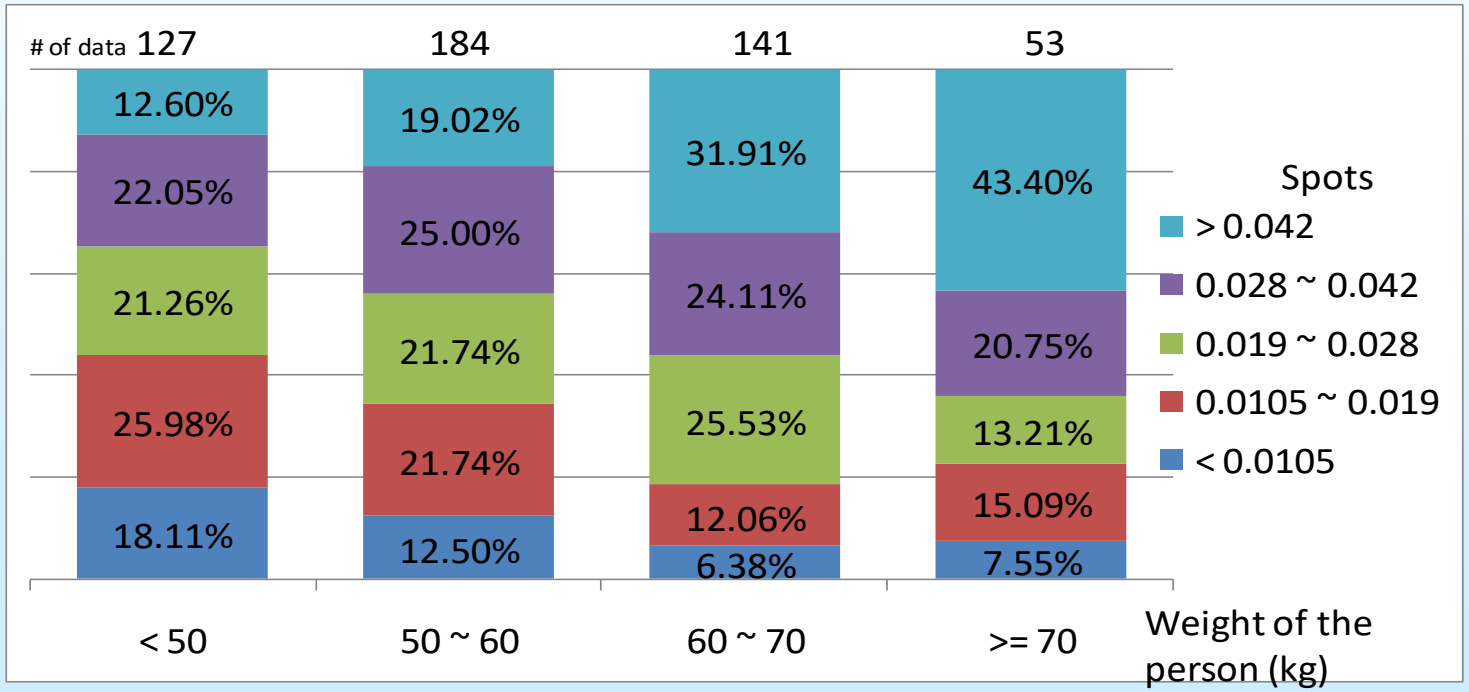
- 图表展示了基本统计的类描述
  - ✓ **Frequency histograms**（频率直方图）
    - 一个单变量图解法
    - 由一组矩形组成，这些矩形反映了给定数据中所呈现的类别的计数或者是频率

车速（千米时）	频　数	百分比
30≤x<40	10	5%
40≤x<50	36	18%
50≤x<60	78	39%
60≤x<70	56	28%
70≤x<80	20	10%
总　计		100%



直方图分析

- 图表展示了基本统计的类描述
  - ✓ **Frequency histograms**（频率直方图）
    - 一个单变量图解法
    - 由一组矩形组成，这些矩形反映了给定数据中所呈现的类别的计数或者是频率



Quantile Plot（分位数图）

- 展示所有的数据（允许用户同时评估整体行为和不寻常事件）
- 绘制分位数信息
  - ✓ 对于一个数据x_i，数据被升序排列，f_i代表小于或等于x_i的数据在全部数据中所占的百分比

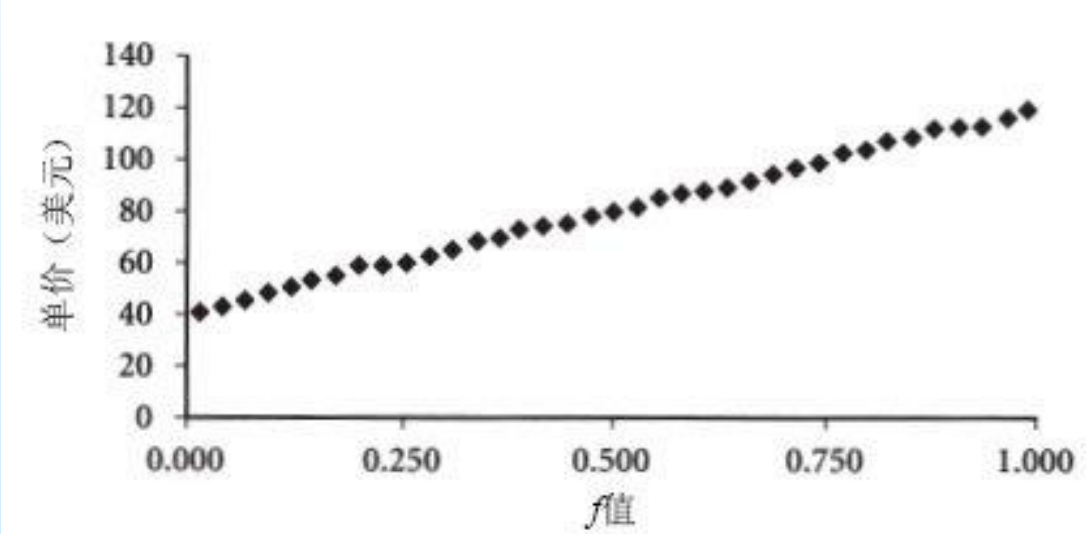


图2-5　表2-1单价数据的分位数图cnrepair.com

Scatter plot(散布图)

- 提供了一个先看看二元数据的群集和离群点等的途径
- 每对值都被当作一对坐标并在平面上用点绘出来

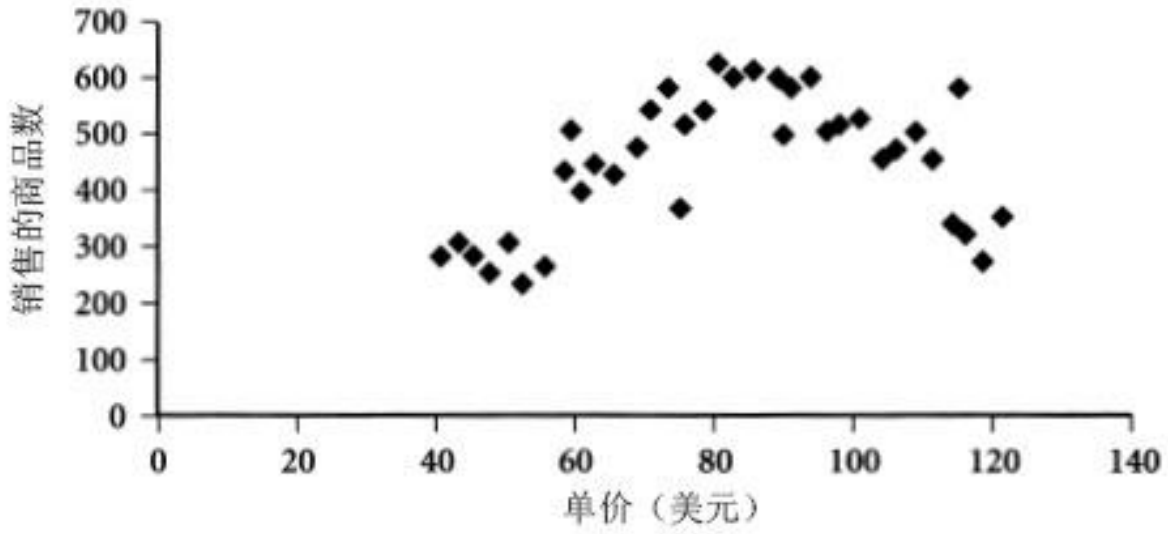
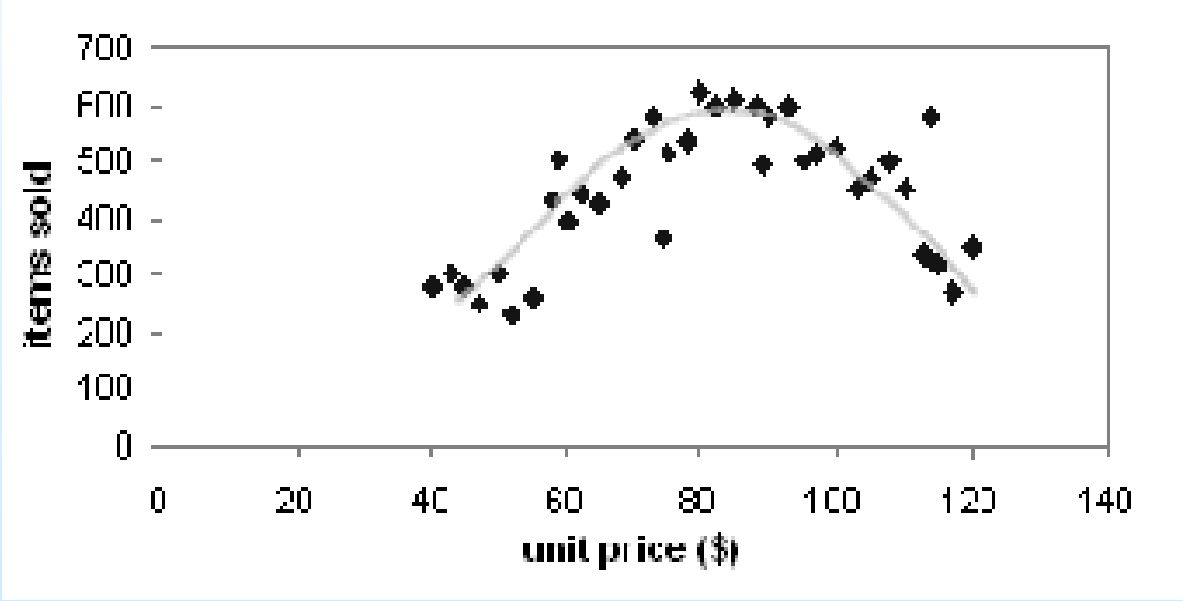


图2-7　表2-1中数据的散布图

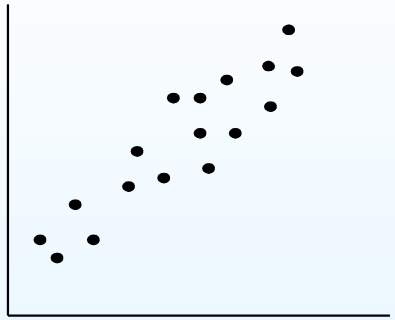
Loess Curve（局部回归曲线）

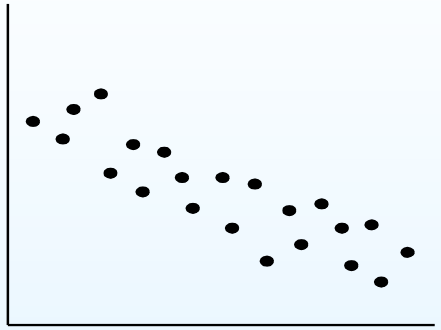
- 给散点图添加一条平滑的曲线，为模式的依赖性提供更好的展示
- 局部回归曲线通过设置两个参数拟合：一个平滑参数，和回归拟合的多项式程度

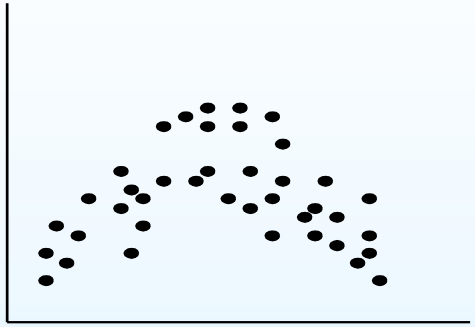


相关数据

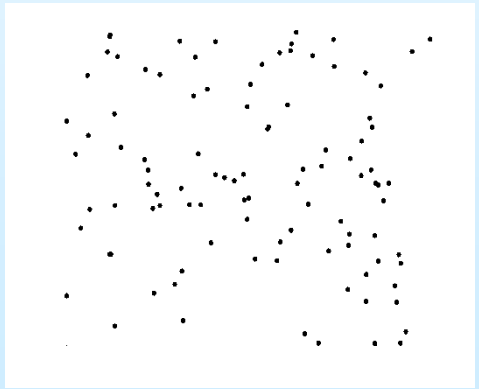
- 正相关和负相关数据

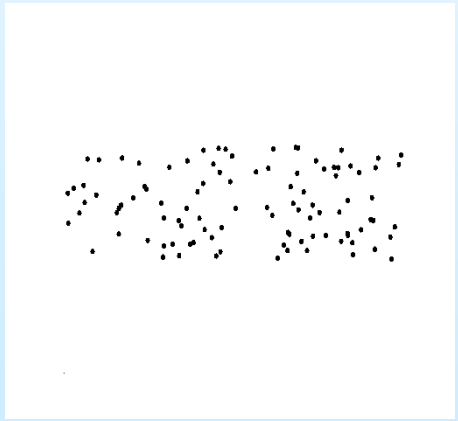


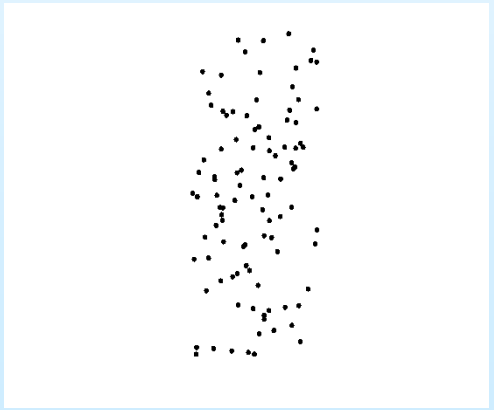




- 不相关数据







散布图矩阵

鸢尾花属性的散布图矩阵

散布图矩阵

鸢尾花属性的散布图矩阵