


第4章: 描述性统计分析与可视化



Python数据科学：全栈技术详解

讲师：Ben

自我介绍

- 天善商业智能和大数据社区 讲师 – Ben
- 天善社区 ID - Ben_Chang
- <https://www.hellobi.com> – 学习过程中有任何相关的问题都可以提到技术社区数据挖掘版块。

主要内容

- 描述性统计与探索型数据分析
- 描述性方法大全
- Python绘图
- 统计制图原理
- 数据可视化



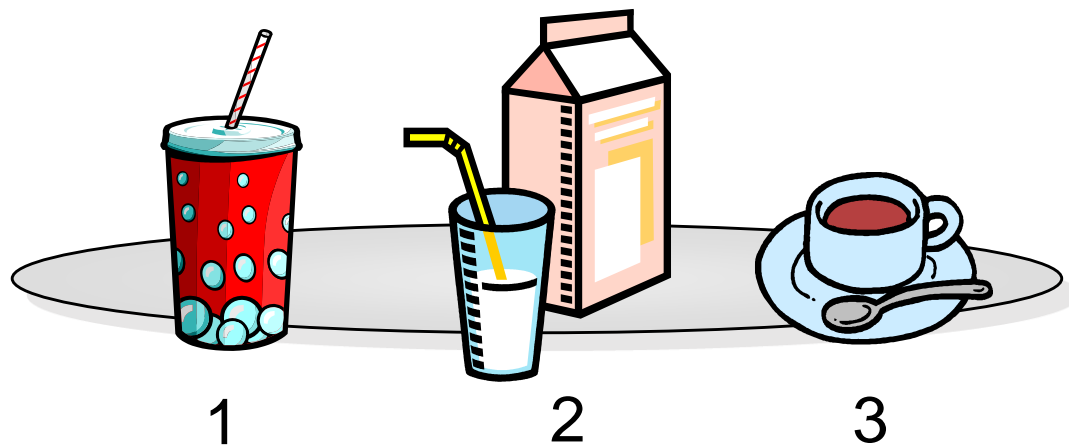
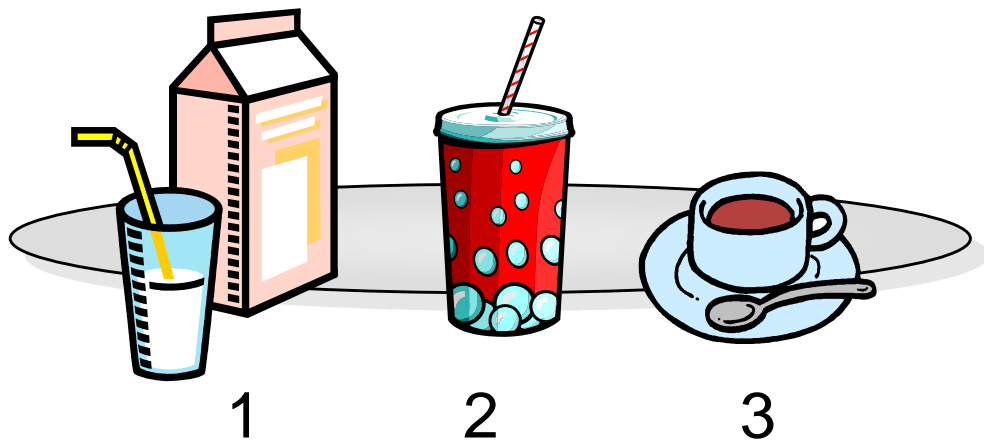
4.1 描述性统计与探索型数据分析

认识变量度量类型



- 在数据分析之前，先明确变量的度量类型(名义、等级、连续)

- 变量: 比如饮料类型



比如饮料包装大小



小



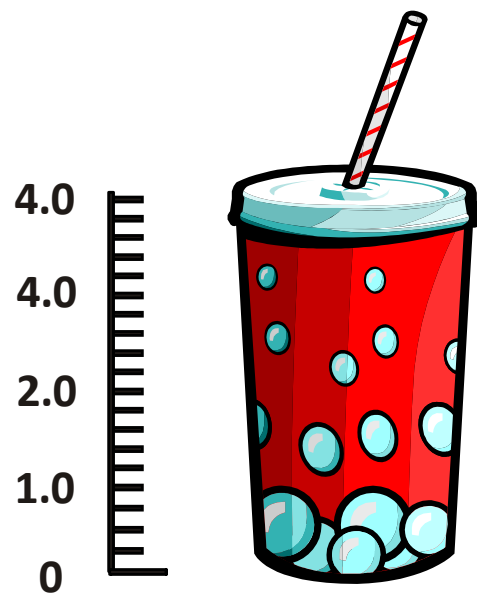
中



大

连续变量

饮料的体积



比例数据

饮料的温度



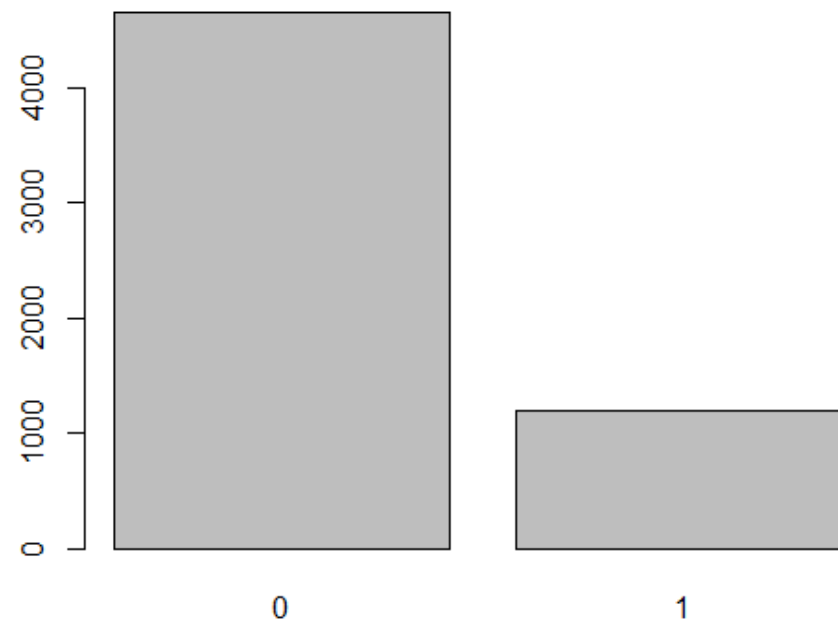
间隔数据

描述名义变量的分布

频数表

	频次	百分比
正常	4648	79.50
违约	1197	20.40

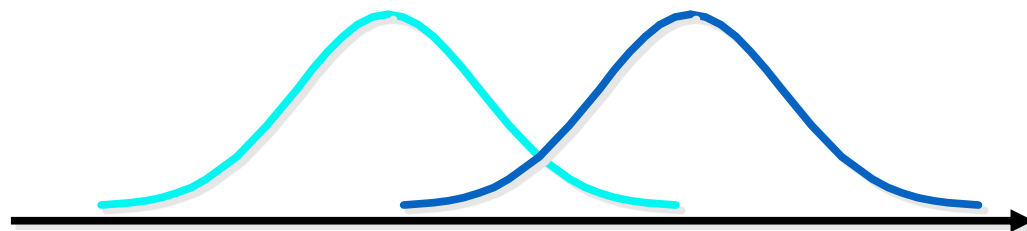
柱形图



描述连续变量的分布

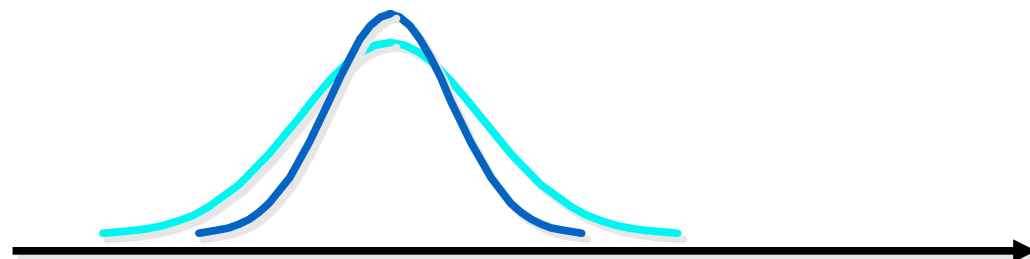
- 需要对**变量进行分布探索**,并了解以下情况:

集中趋势
(位置)



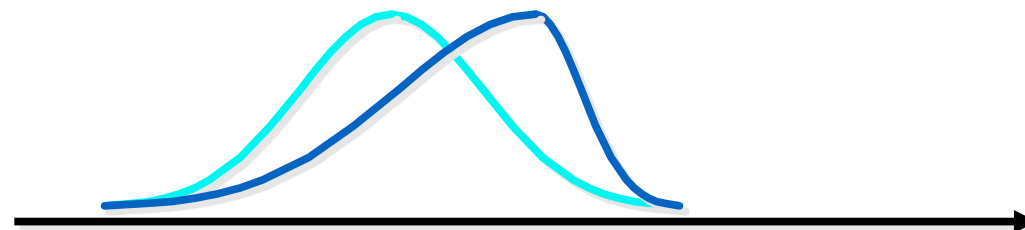
什么统计量可以概括这个变量?

离中趋势
(分散程度)



这个统计量的概括能力有多强?

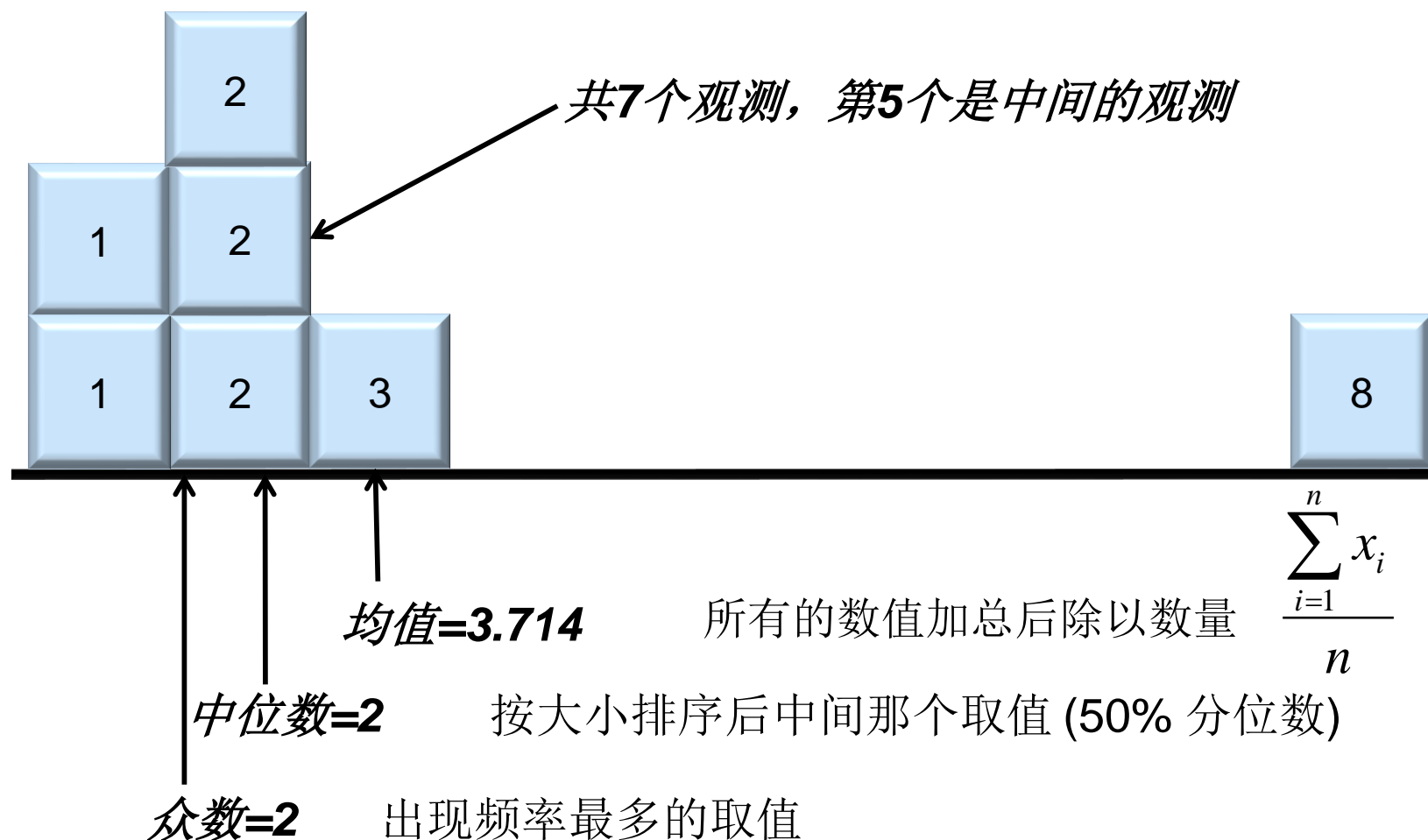
偏态和峰态
(形状)



选择哪个统计量更“科学”?

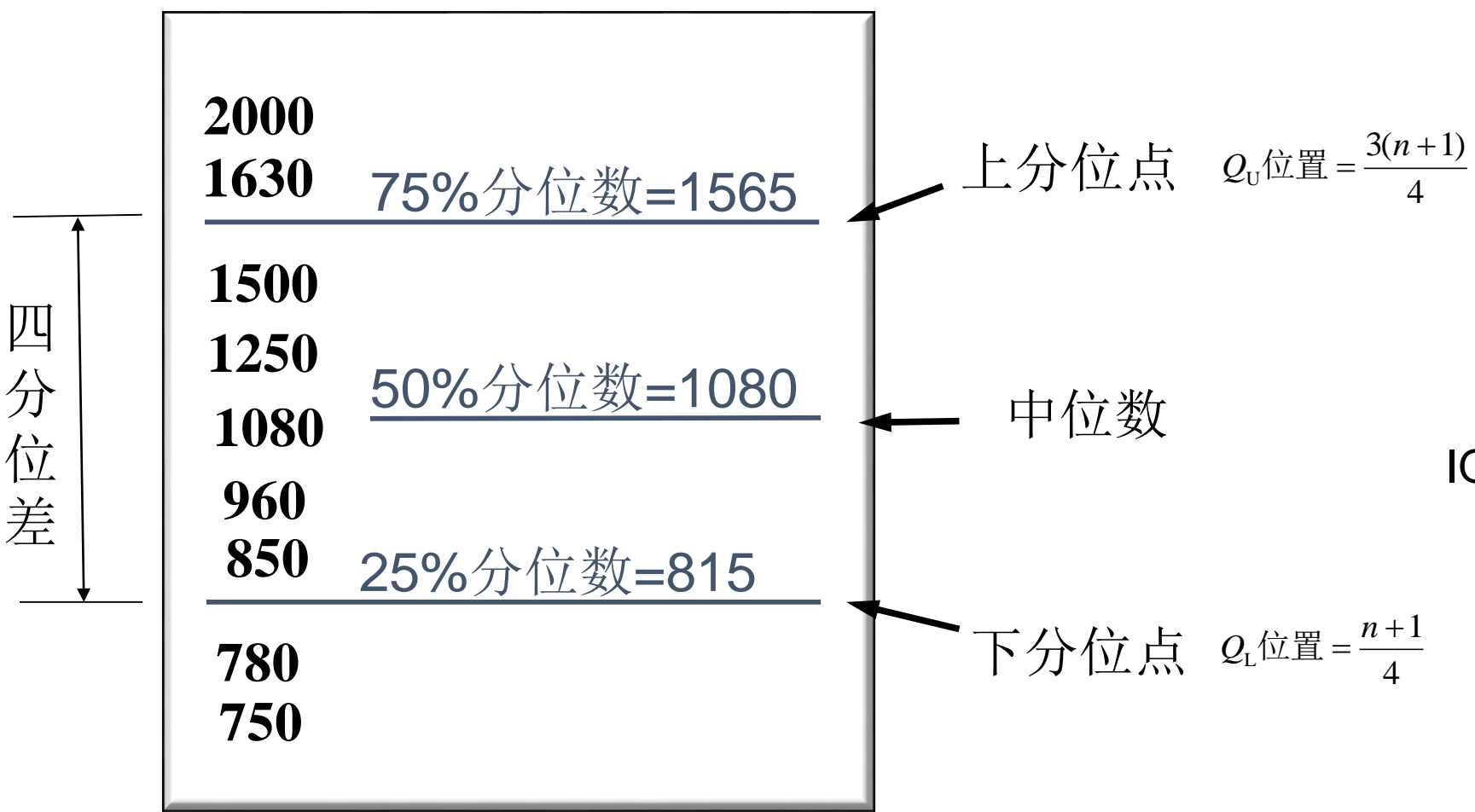
连续数据的位置

中心的度量- 均值、中位数、众数

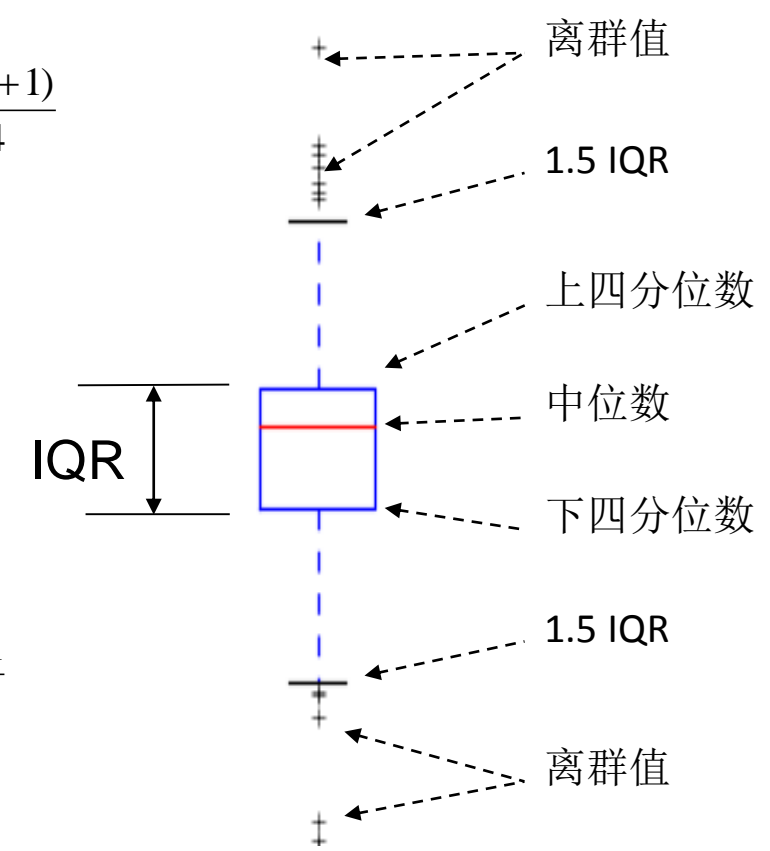


连续数据的位置

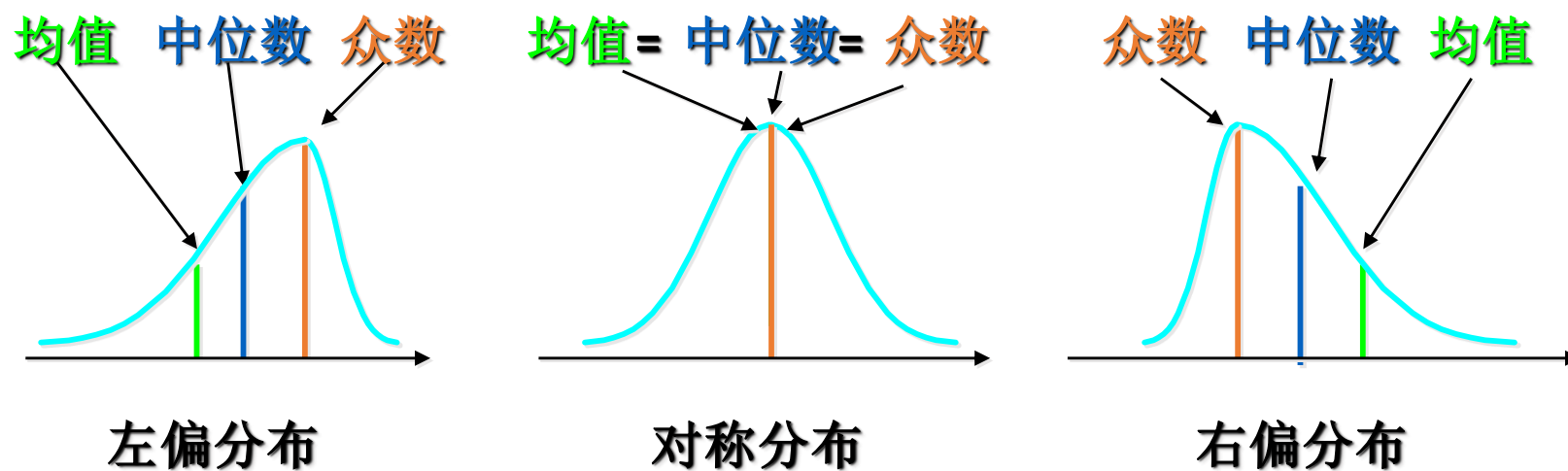
百分位数



盒须图



众数、中位数和平均数的关系



数据的离散程度

极差(Range)

极差=最大值-最小值

四分位差(IQR)

四分位差=上分位数-下分位数

平均绝对偏差(Mean Absolute Deviation)

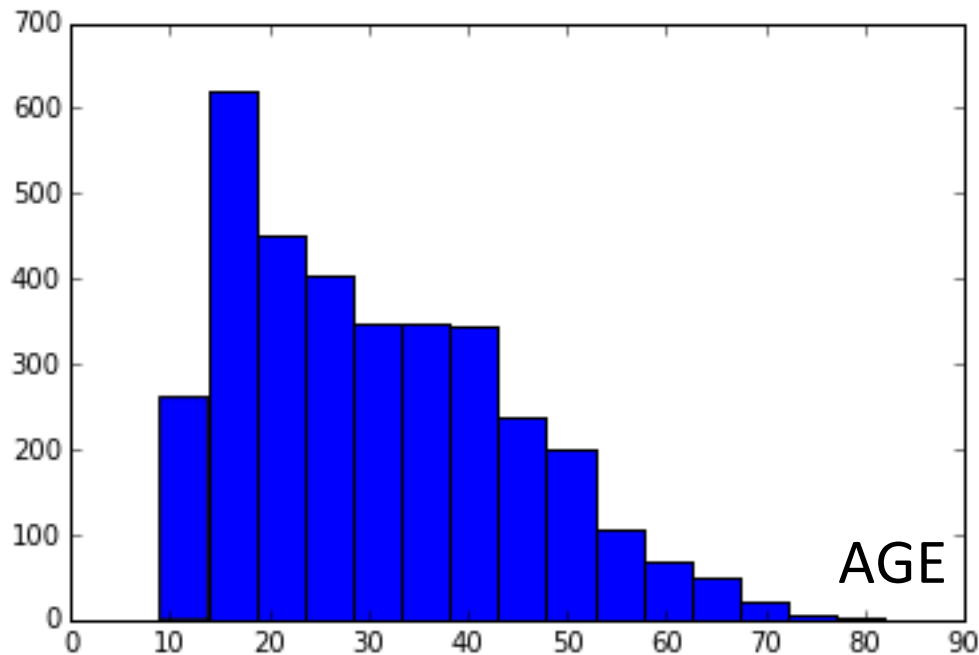
$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

方差(Variance)和标准差(Standard Deviation)

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

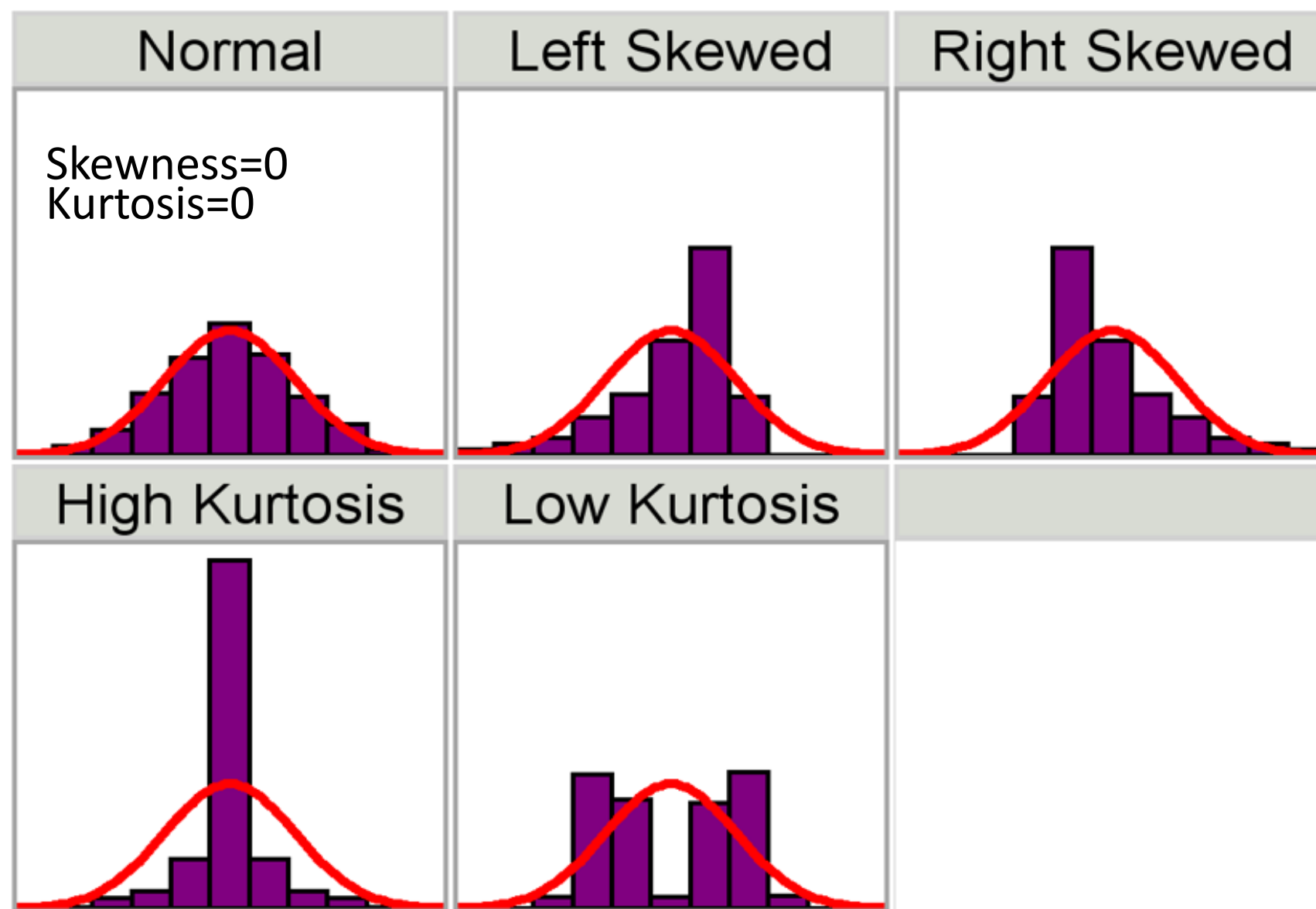
$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

描述连续变量的分布展现-直方图

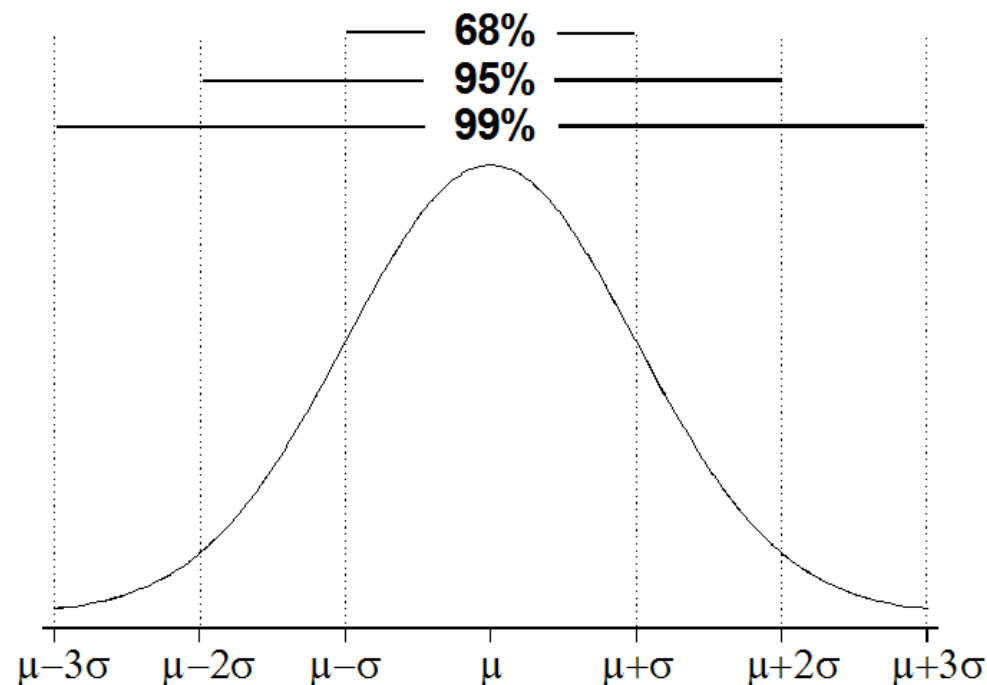


- 直方图常用于了解数据的分布形状
- 一般情况下，横轴为连续变量的分段进行等宽离散后的值，纵轴为频次；
- 每个柱的宽度可不相同，纵轴也可以不是频次，通过bins和normed参数可进行相应设置

描述连续变量的分布形态-偏态与峰度



正态分布



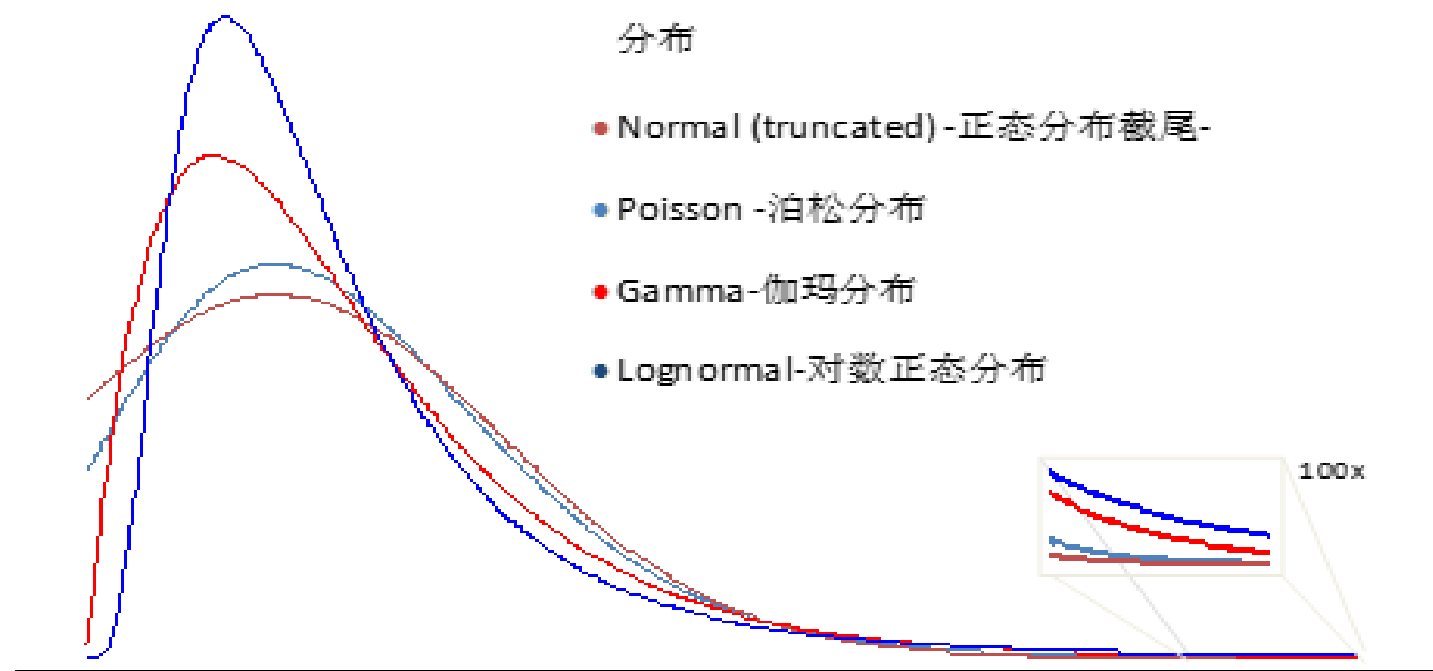
对称 (symmetric). 关于均值左右对称分布.

均值和标准差的代表性 (**fully characterized**) 只要知道其均值和标准差, 这个变量的分布情况就完全知道了.

倒钟形.

均值 = 中位数 = 众数.

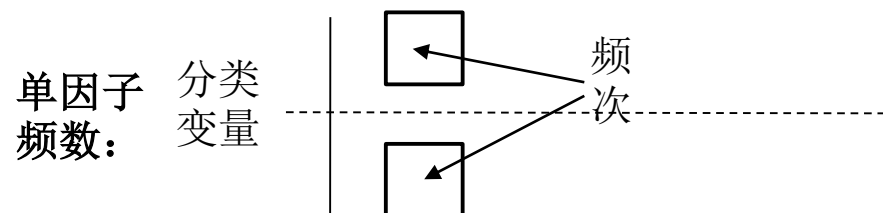
其它常见连续分布形式



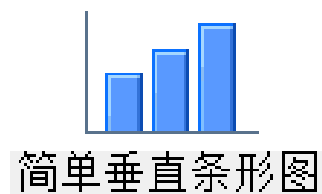
其中对数正态分布在统计分析中运用最为广泛，顾名思义，这种类型的分布在取对数之后服从正态分布。因为其具有这样的良好属性，在精确度要求并不严格的统计分析中，经常对偏态分布首先进行对数转换。

4.2 描述统计方法大全

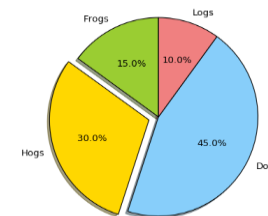
描述统计的总结



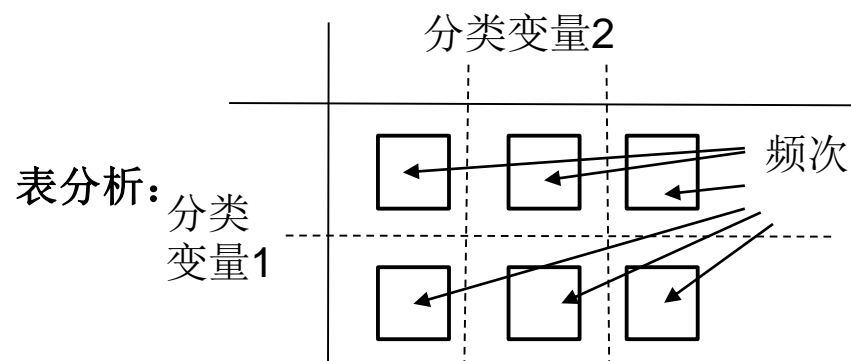
```
snd.district.value_counts()
```



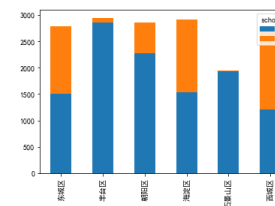
```
snd.district.value_count  
s().plot(kind = 'bar')
```



```
snd.district.value_count  
s().plot(kind = 'pie')
```



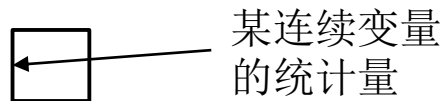
```
pd.crosstab(snd.district,snd.school)
```



```
pd.crosstab(snd.district,  
snd.school).plot(kind =  
'bar')
```

描述统计的总结

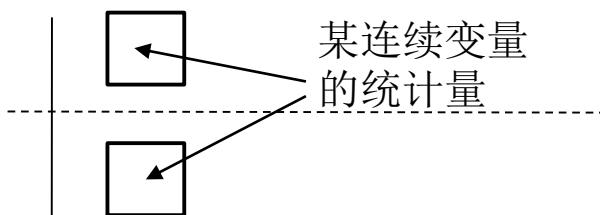
单连续
变量描
述:



```
snd.price.agg(['mean','median','sum','std',  
skew'])
```

分类汇
总:

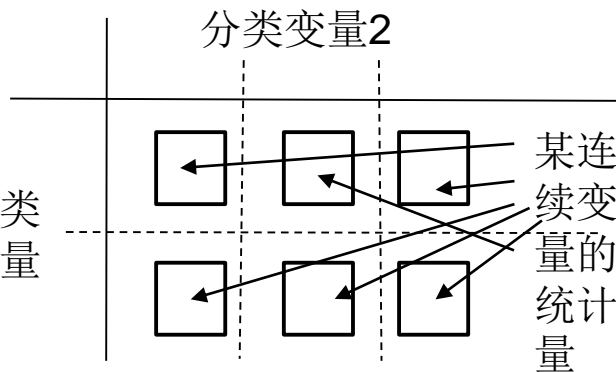
分类
变量



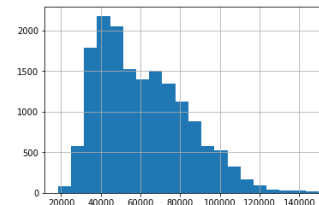
```
snd.price.groupby(snd.district).sum()
```

汇总表:

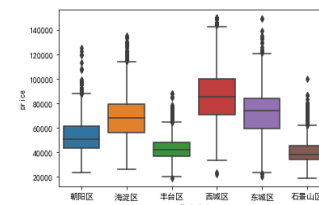
分类
变量
1



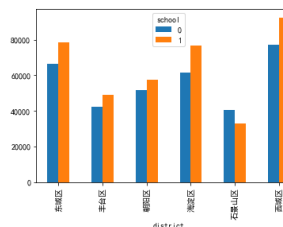
```
snd.pivot_table(values='price', index='district',  
columns='school', aggfunc=np.mean)
```



```
snd.price.hist(bins=20)
```

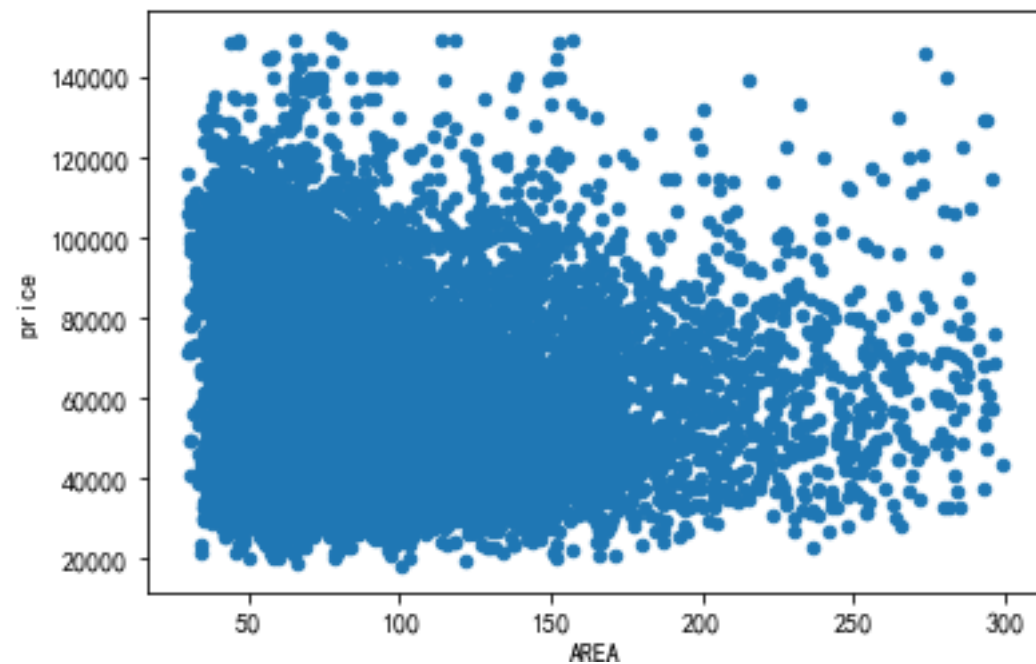


```
sns.boxplot(x = 'district', y  
= 'price', data = snd)
```



描述统计的总结

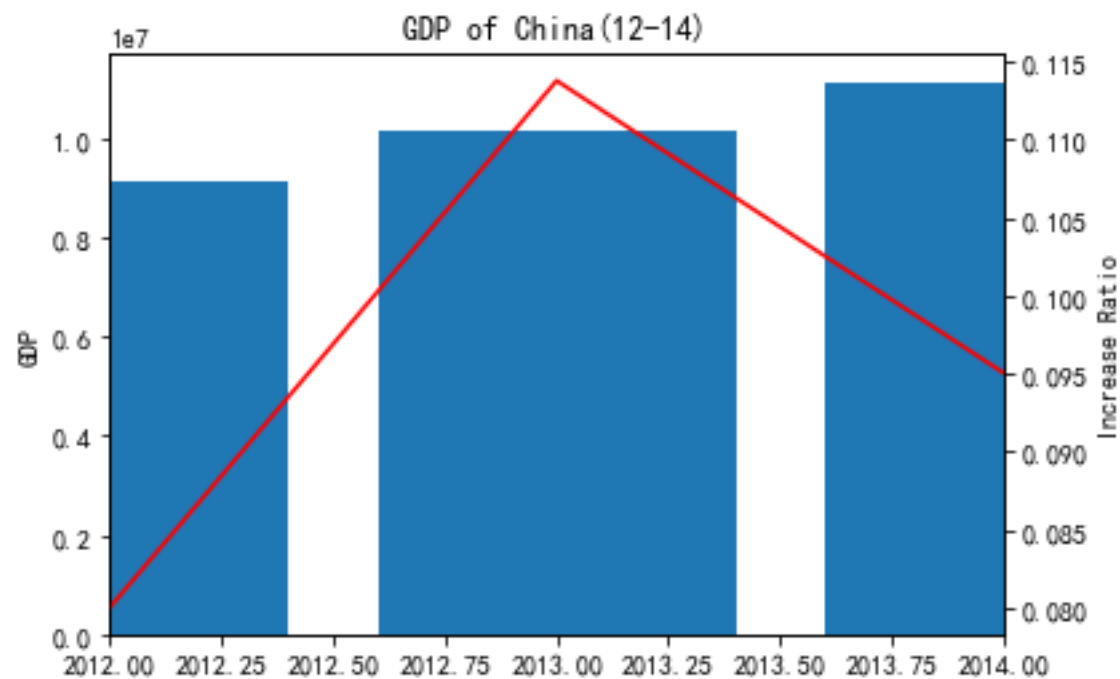
两个连续变量:



```
snd.plot.scatter(x = 'AREA', y = 'price')
```

描述统计的总结

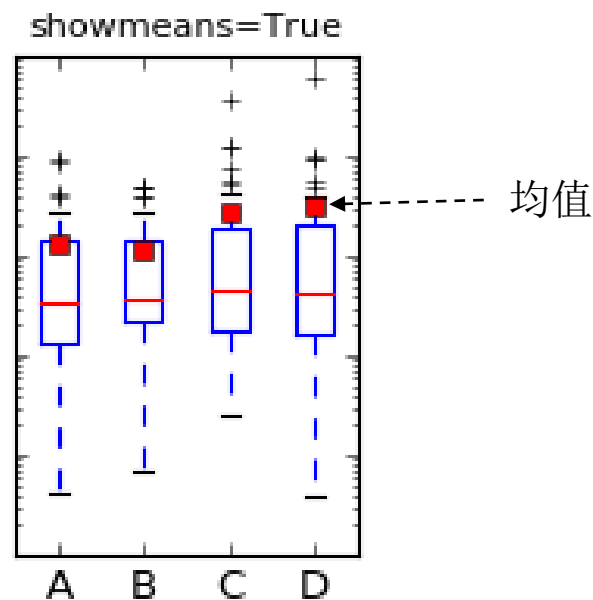
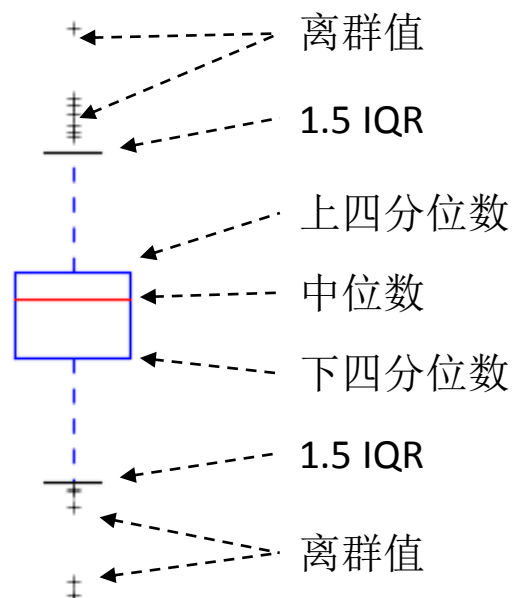
时间与
两个连
续变量:



描述统计的总结

分类与
单个连续
变量:

- 盒须图 (Box-plot) , 用于显示一组数据分散情况的统计图, 可以显示中位数、均值、四分位数、离群值等信息
- 常用于多组之间数据分布的比较



描述统计的总结

分类与
单个连续变量:

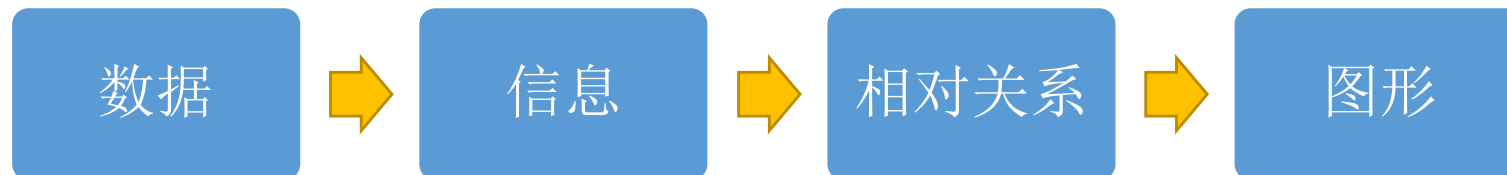
盒须图能够提供某变量分布以及异常值的信息，其通过分位数来概括某变量的分布信息从而比较不同变量的分布。盒须图的基本元素包括：

- IQR：变量上下四分位数之间的数据，这个范围代表了数据中间50%的数据。
- 中位数位置：中位数位置即代表变量中位数在总体分布中的位置。
- 1.5IQR：上下1.5IQR表示上下1.5倍IQR范围的数据，其能够提供中位数左右95%的置信区间的数据。可以直观的从盒须图中看出超出95%置信区间范围的数据，即异常值。
- 不同变量的盒须图比较时，可以通过中位数位置来比较两变量数据的中位数差异状况。

4.3 制图原理

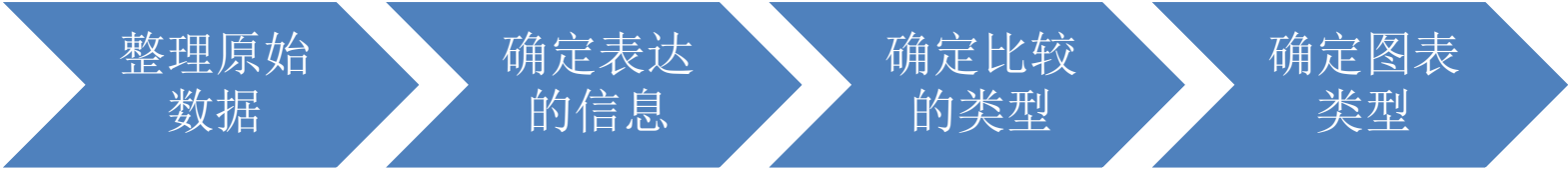
图形展示三步

- 第一步 明确你要表达的信息
- 第二步 确定相对关系
- 第三步 选择图表形式



图形展示三步：第一步

以数据为基础制作图表的步骤

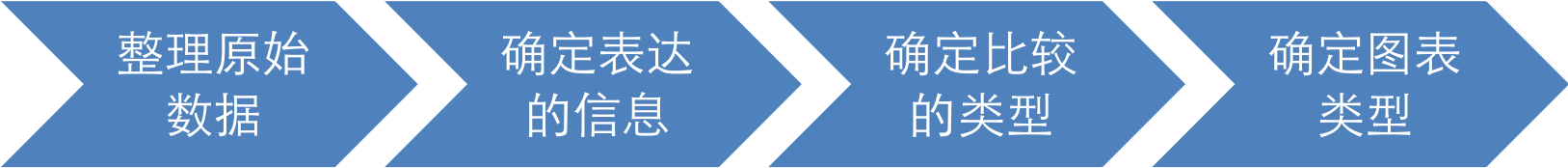


年份	销售员	市场	销售额	利润
2010	赵	东	267310	32117
2010	钱	南	295000	38171
2010	李	南	291520	35639
2010	周	南	316470	41241
2010	郑	南	296340	29595
2011	钱	西	275680	27857
2011	孙	西	298030	36228
2011	周	北	314990	38385
2011	吴	东	337040	44445
2011	郑	东	303160	24127

整理好的规整数据
是做后续分析的基础。

图形展示三步：第二步

以数据为基础制作图表的步骤



时间 序列	年份	销售额(元)
	2010	1466640
	2011	1528900
	2012	2420480
	2013	2140000

分析以往年份销售额的变化趋势,预测下一年的销售额。

区域 比较	区域	销售额(元)
	东	35000
	西	15000
	南	52000
	北	23000

比较不同区域的销售情况,为绩效考核提供依据。

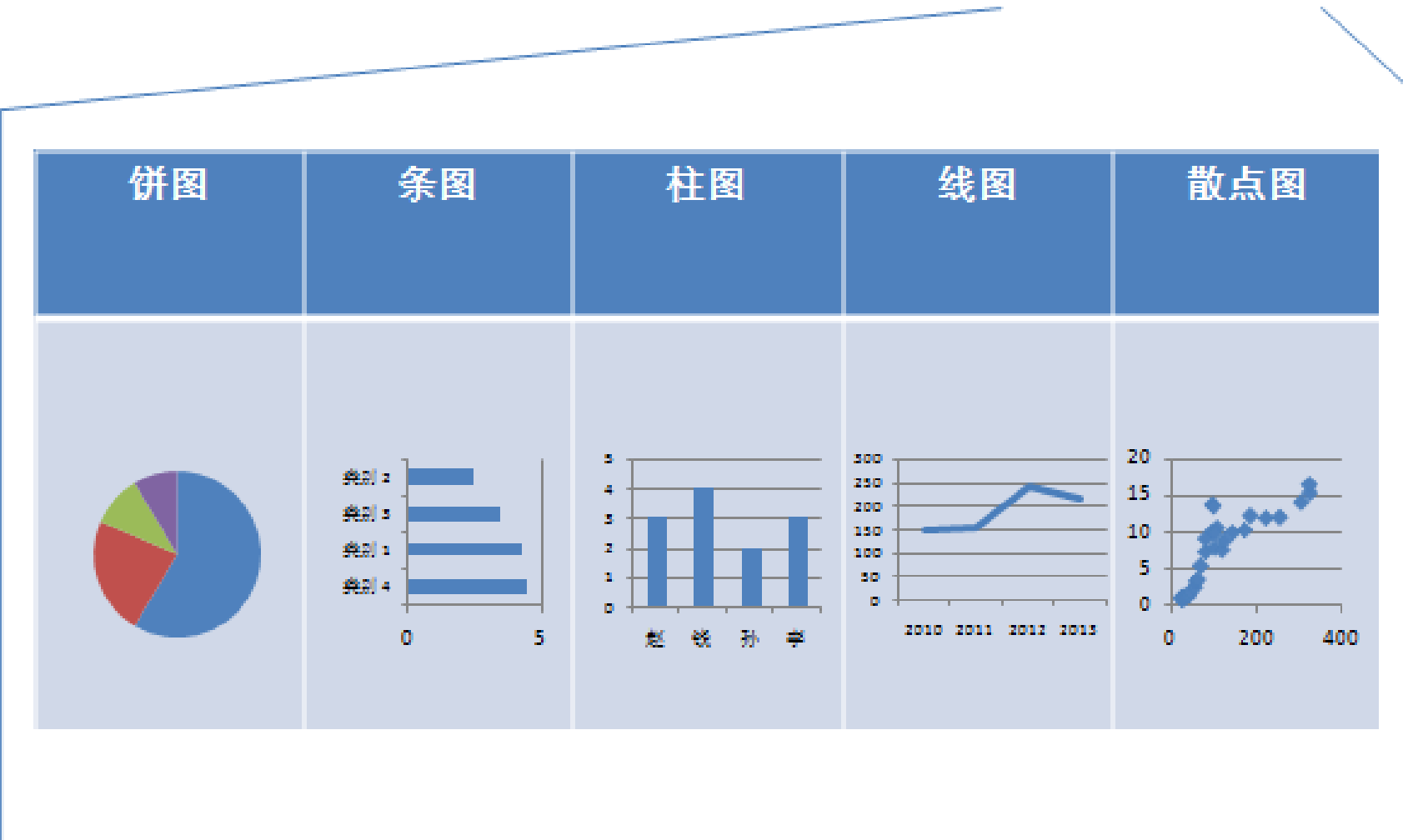
图形展示三步：第三步



比较类型	说明	举例
成分	各个部分占整体百分比的大小	当年不同区域销售额占整个公司销售额的百分比
排序	不同元素的排序	当年哪个区域的销售额最高,哪个最低
频率分布	单变量在不同数值上的频数或百分比	当年哪个区域的销售额上亿元了
时间序列	变量在时间纬度上的变化	整个公司近几年的销售额变化情况
关联性	两种可变因素之间的关系	公司的销售额变化和国家宏观经济情况（GDP）变化是否相关


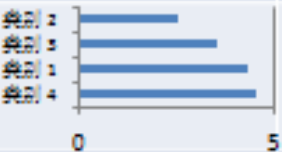



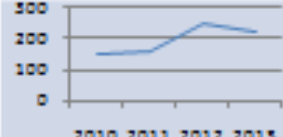

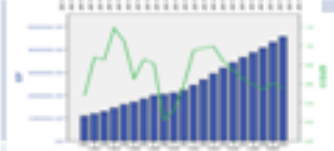
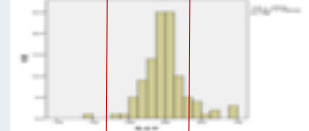
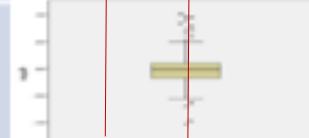
图形展示三步：第三步

以数据为基础制作图表的步骤



图形展示三步：第三步

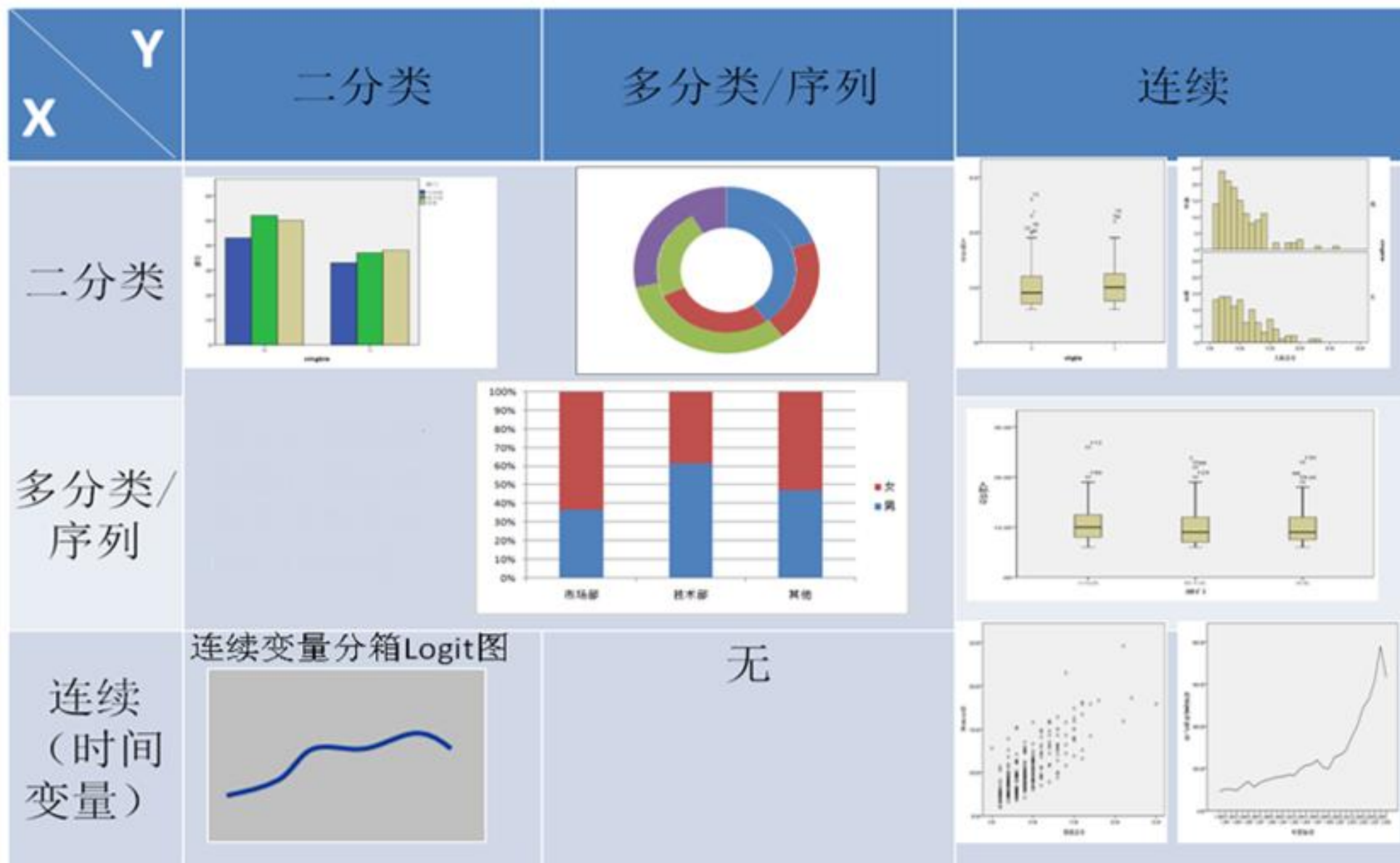
通过不同的图表表达特定对比

	成分	排序	频率(分类)	分布(连续)	时间趋势
饼形图					
条形图 柱形图					
线形图					
高低图					
双轴图					
直方图					
箱形图					

- “条图”与“柱图”具有细微的差异，两个图形经常回混用。
- “双轴图”由于存在两个Y轴，容易被混淆，所以广被病垢。R的专业绘图包ggplot中甚至明确的提出不支持双轴图。其实“双轴图”在使用中有其特殊的用途和约定俗成的使用情景。其中的X轴为时间，左Y轴为指标水平（也称作绝对量,比如GDP）的刻度，右Y轴为指标增长率（也称作变化率,比如GDP增长率）。而且用柱形代表水平，用线形代表增长率。

图形展示三步：第三步

表达关联性的不同图表



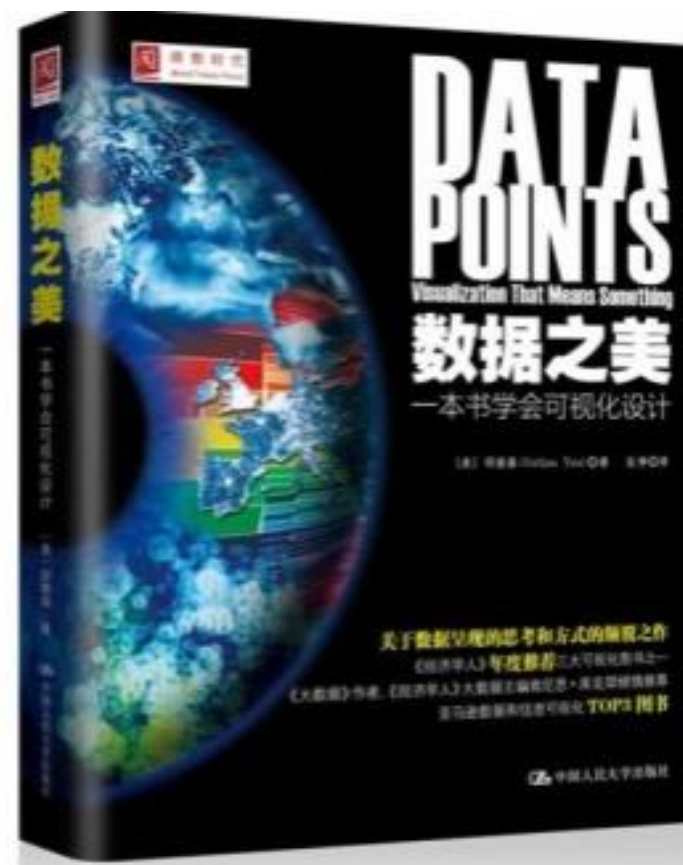
图形部分说明

统计图属于描述性统计，是对统计汇总表的形象性展示。**EXCEL**虽然提供了比**R**更多的图表功能。但是严格来说，**EXCEL**并不能直接做出统计图，它需要在个体记录原始数据的基础之上进行统计汇总，然后根据汇总数据进行作图。而**R**的作图功能是直接基于个体记录的原始数据进行绘图。

每类统计图是为了满足特定叙述目的而出现的，其类似于语言，有其明确的定义与叙述方式。复杂叙述目的的实现是通过综合运用每类统计图，而不是创造出复杂的图。好的统计图可以使阅读者在仅阅读标题关键字和图形，不用注意任何坐标轴标题、刻度和附注的情况下顺利地理解需要表达的意义。

统计图分为表述性统计图和检验统计图，前者是对某些变量分布、趋势的描述，大量出现在工作报告中和统计报告中，比如饼图、条图。后者是对特定统计检验和统计量的形象展示，仅出现在特定统计报告中，一般不在工作报告中出现，比如直方图和箱形图，**P-P**图，**ROC**曲线。不过这个界限有些模糊，比如箱形图已开始是统计图，但后来人们觉得其表现连续函数和分类变量的关系时很直观，所以也被广泛的用于工作报告中。

推荐一本书



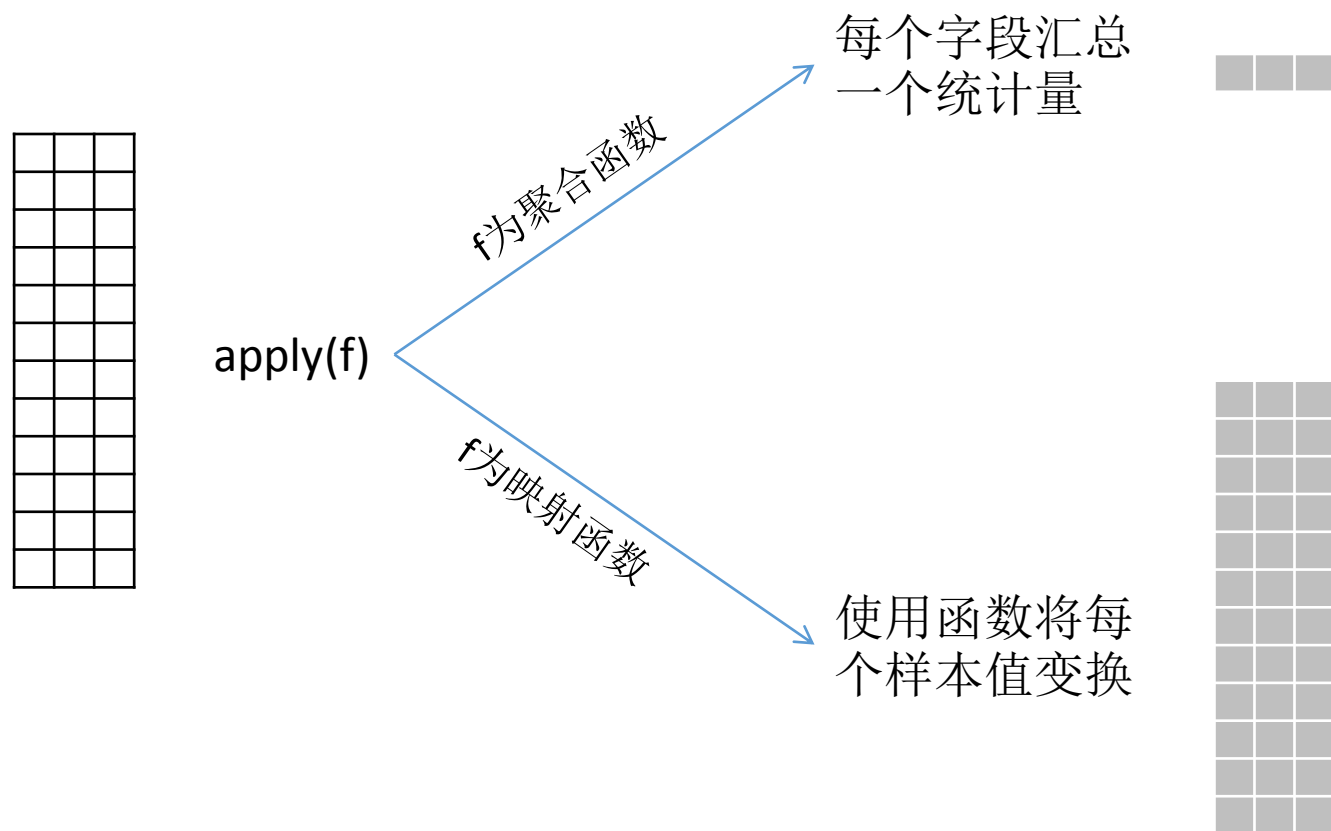
数据之美：一本书学会可视化设计



4.4 附录: apply\map\groupby 及其它相关功能

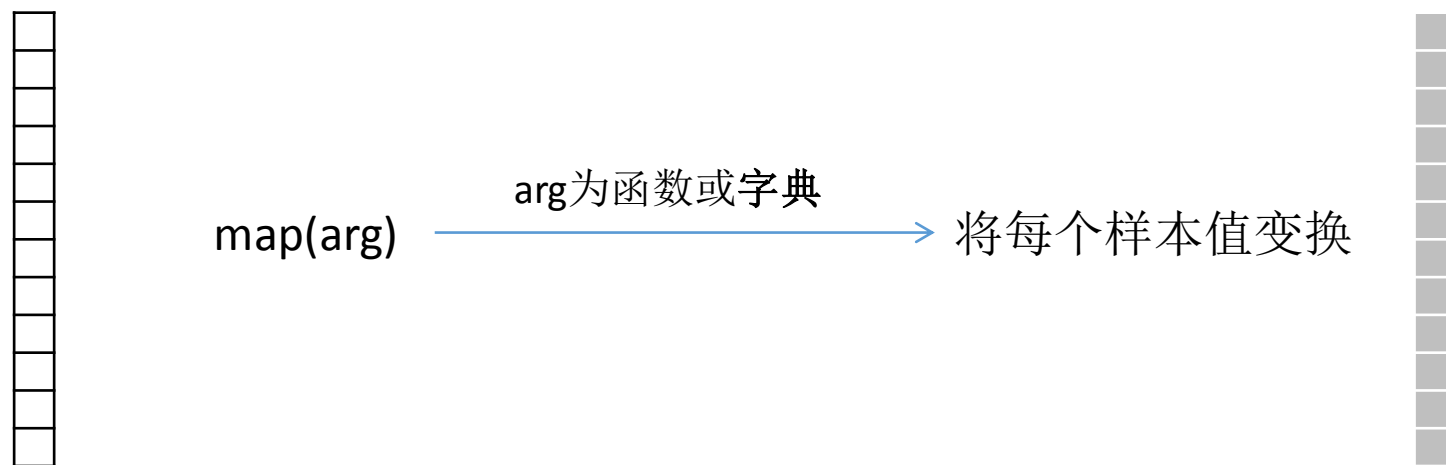
- 日常的数据分析当中经常需要生成报表，其应用广泛，结果简单易懂：
 - 聚合(汇总)：pandas提供了比reduce更强大的汇总方法-apply
 - 映射：使用“广播”或使用与列表、数组相类似的方法-map
 - 分组汇总：使用groupby按字段分组，再使用aggregate进行汇总
 - 交叉表：多个字段交叉，汇总频次、均值等
 - 其它：transform、agg等等

- 可将函数应用到每个字段，根据函数的不同可以用于“聚合”数据或“映射”数据



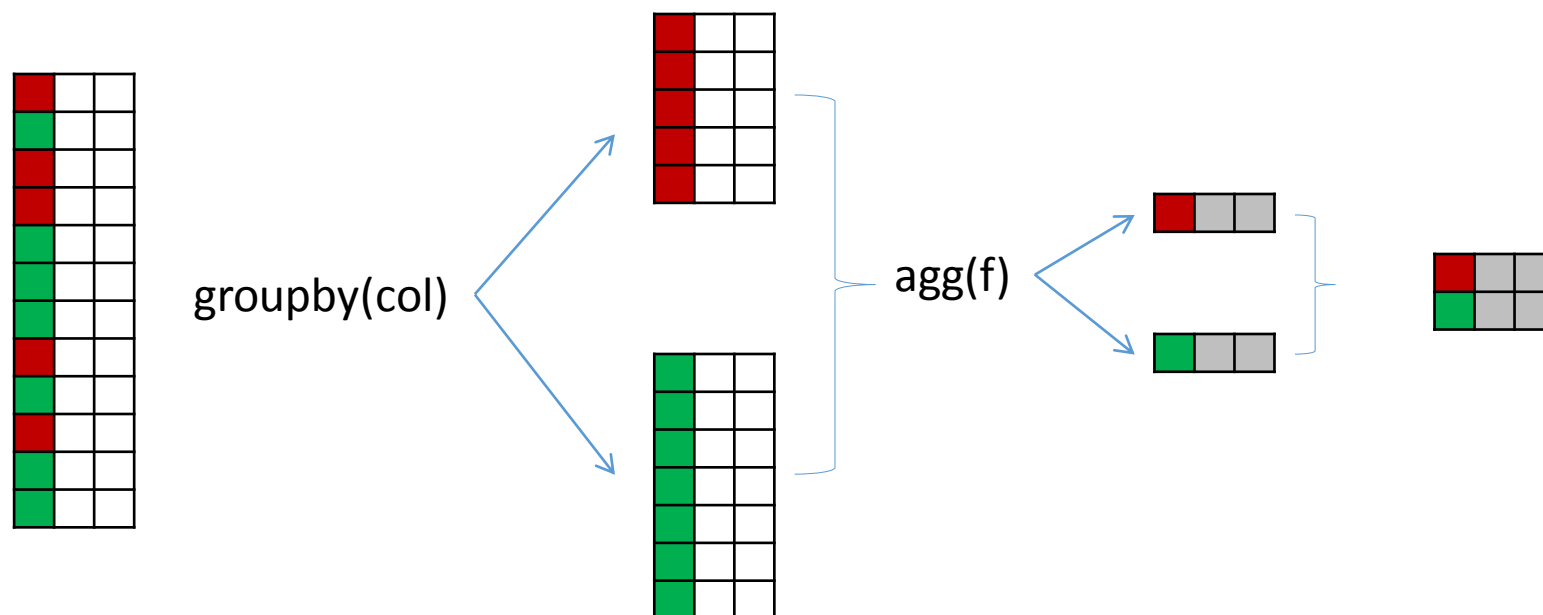
map

- map方法可以将某个字段的每个值使用函数进行变换，类似于内建的map方法，仅能对单独字段（series）使用
- 可以使用字典进行map，例如将“物品编号-物品名”字典传入，可将编号字段映射成名称字段

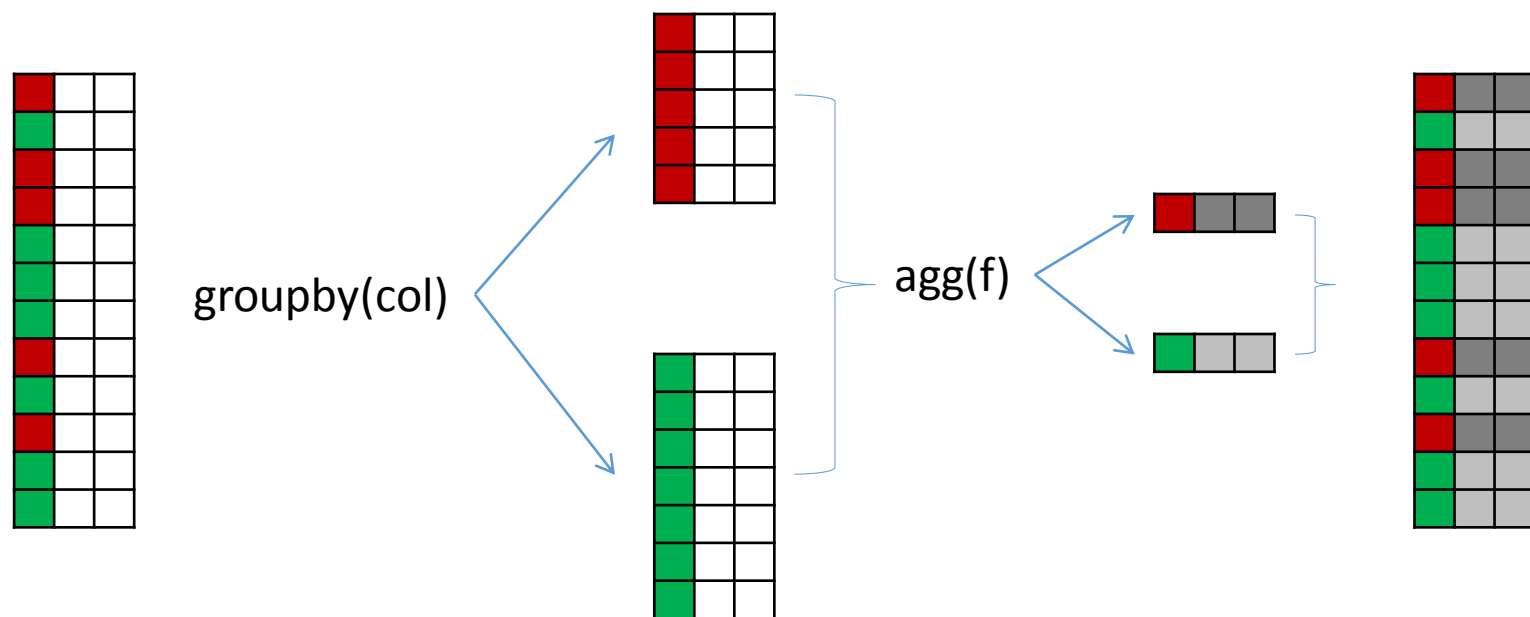


groupby和aggregate

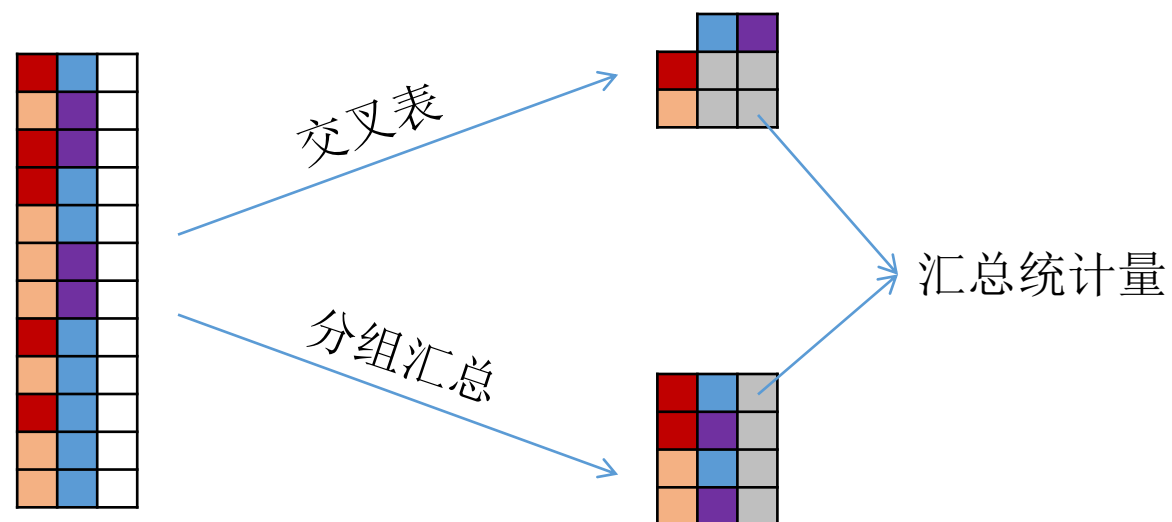
- 使用groupby将数据分组，使用aggregate将每个分组进行聚合，类似还可以使用agg和apply
- 用于分组的字段需要与待聚合的字段有同样长度，可以**按多个字段进行分组**，也可以一次聚合多个汇总字段



- 使用groupby将数据分组，使用transform将每个分组进行聚合，聚合的结果返回到原数据集中



- 按照多个字段汇总统计量可以使用交叉表，其结果的可读性高于分组汇总表
- 可以交叉两个或多个字段



—— 秦路主讲 ——
七周成为数据分析师
七周为期，Get一条数据分析师职业黄金通道！



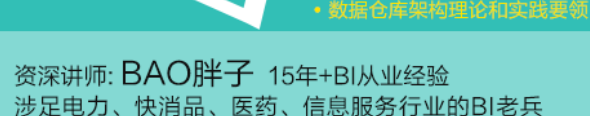
—— Python ——
数据分析与挖掘
集Python爬虫、数据采集、数据处理、数据分析与数据挖掘于一体，打造Python全栈工程师
主讲老师: 韦玮
VIP会员群+在线答疑+录播复习+1年反复观看



案例为师, 实战为王
开启Python机器学习之路
科学规划全套课程体系，从入门到进阶，从理论到技巧，嵌入丰富课程案例讲解，逐步推进
讲师: 唐宇迪 深度学习领域多年一线实践研究专家



**独一无二的
数据仓库**建模指南系列教程升级版
• 从企业视角进行数据规划以及数据仓库模型的搭建
• 高质量的数据库模型和技巧，以及丰富的例子
• 数据仓库架构理论和实践要领
资深讲师: BAO胖子 15年+BI从业经验
涉足电力、快消品、医药、信息服务行业的BI老兵



业务知识一站通
技术+业务，挣钱有门路！
—— 讲师: 陈文 ——



自己动手 丰衣足食
Python3网络爬虫实战案例
— 循序渐进，案例为王，诠释全面，思路制胜 —
讲师: 崔庆才 北航硕士，百万级热度爬文博主



讲师 丘祐玮
人人都爱数据科学家
Python数据科学精华实战课程



数据分析报告制作
秘籍升级版
讲师: 陈丹奕 知乎大神，前百度资深数据分析师



先机致胜 破冰AI
—— 深度学习模型/框架与实战 ——
讲师: 唐宇迪 同济大学硕士
深度学习领域多年一线实践研究专家



BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI