



Introduction to Data Analysis

# 数据分析基础

中国地质大学（北京）信息工程学院

叶山

[yes@cugb.edu.cn](mailto:yes@cugb.edu.cn)

# 数据的组织和管理

(继续)

# 数据组织和管理

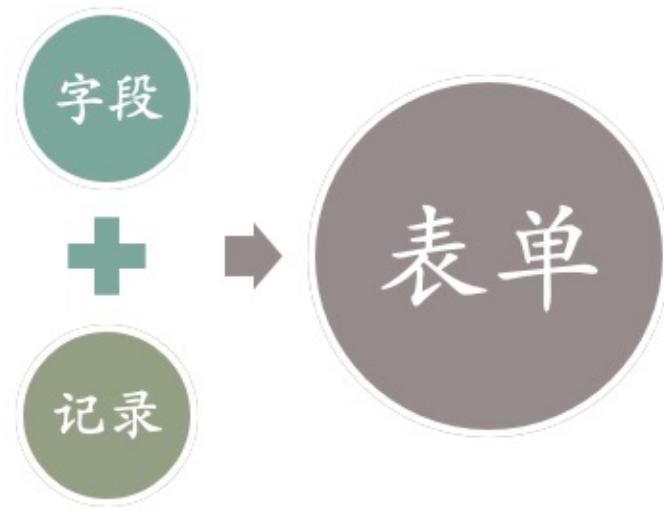
数据可视化相关的数据组织



# 数据表单

字段也叫属性，是事物或现象的某种特征，在统计学中称为变量。

序号	日期	销量	销售额	销售商品数	动销率
1	2018/9/9	1811	81944.8	55	54.46%
2	2018/9/8	713	30967.2	34	33.66%
3	2018/9/7	186	8892.1	31	30.69%
4	2018/9/6	240	10185.1	22	21.78%
5	2018/9/5	230	9347.2	21	20.79%
6	2018/9/4	467	19602.8	32	31.68%
7	2018/9/3	412	18180.2	34	33.66%



记录也叫数据点、数据值、数据元或变量值，是事物或现象某种特征的具体表现。

# 数据表单

## 设计原则

1. 数据表由标题行和数据构成。
2. 第一行为表的列标题（字段），列标题不能重复。
3. 从第二行起是数据部分，每一行数据称为一个记录，原则上不允许出现空自行和空白列。
4. 在数据表中不能出现跨越多行或多列的单元格。
5. 数据表单优先以一维的形式存储，但在实际工作中，所接触到的数据往往是以二维表格的形式存在的，这时应尽量将二维表转换为一维表的形式进行存储。

Date	Start Time	Lunch Start	Lunch End	End Time	Overtime	Hourly Rate (USD)	Total Hours Worked (h)	Amount Earned
4/1/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/2/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/3/2019	8:00am	1:00pm	1:30pm	5:00pm	2	50	10	500
4/4/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/5/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/6/2019	9:00am	2:00pm	2:30pm	6:00pm	0	50	8	400
4/7/2019	9:00am	2:00pm	2:30pm	6:00pm	0	50	8	400
4/8/2019	8:00am	1:00pm	1:30pm	5:00pm	1	50	9	450
4/9/2019	8:00am	1:00pm	1:30pm	5:00pm	1	50	9	450
4/10/2019	8:00am	1:00pm	1:30pm	5:00pm	1	50	9	450

AL	Montgomery	Huntsville	Dec 14, 1819	5,024,279	52,420	135,767
AK	Juneau	Anchorage	Jan 3, 1959	733,391	665,384	1,723,337
AZ	Phoenix		Feb 14, 1912	7,151,502	113,990	295,234
AR	Little Rock		Jun 15, 1836	3,011,524	53,179	137,732
CA	Sacramento	Los Angeles	Sep 9, 1850	39,538,223	163,695	423,967
CO	Denver		Aug 1, 1876	5,773,714	104,094	269,601

# 数据表单

## 优点

- 灵活度好，处理数据时的自由度高
- 有齐全的配套软件
- 对用户的技术要求低
- 软件对电子表单文件的变种（CSV、TSV等格式）支持度好

## 缺点

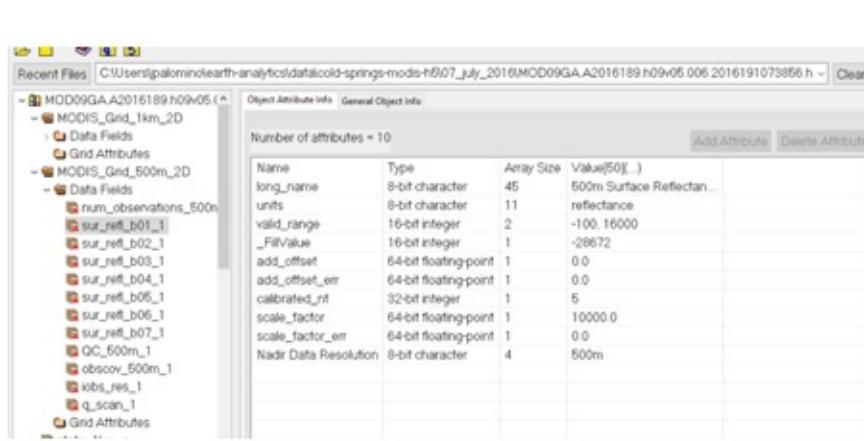
- 数据冗余率高
- 不易访问调用、查错改错
- 难以添加约束
- 安全性低
- 缺少数据类型和元数据，需要语义解释

# 半结构化文件

- 标记语言格式
  - 为方便通用型数据存储和交换，数据导向型的应用程序会采用标记语言格式来组织数据，比如XML（Extensible Markup Language，可扩展标记语言）。
- 科研领域
  - 一些科学领域使用特定的半结构化文件以满足特殊领域知识的表达和处理需求。
    - HDF（层次型数据结构）：美国国家超级计算应用中心创建的多对象文件格式。
    - JSON：轻量级的文本数据交换格式，使用JavaScript语法来描述数据对象，但是JSON可跨语言、跨平台使用。多用于互联网前端。

```
<studentsList>
```

```
  <student id="1">
    <firstName>Greg</firstName>
    <lastName>Dean</lastName>
    <certificate>True</certificate>
    <scores>
      <module1>70</module1>
      <module12>80</module12>
      <module3>90</module3>
    </scores>
  </student>
  <student id="2">
    <firstName>Wirt</firstName>
    <lastName>Wood</lastName>
    <certificate>True</certificate>
    <scores>
      <module1>80</module1>
      <module12>80.2</module12>
      <module3>80</module3>
    </scores>
  </student>
</studentsList>
```

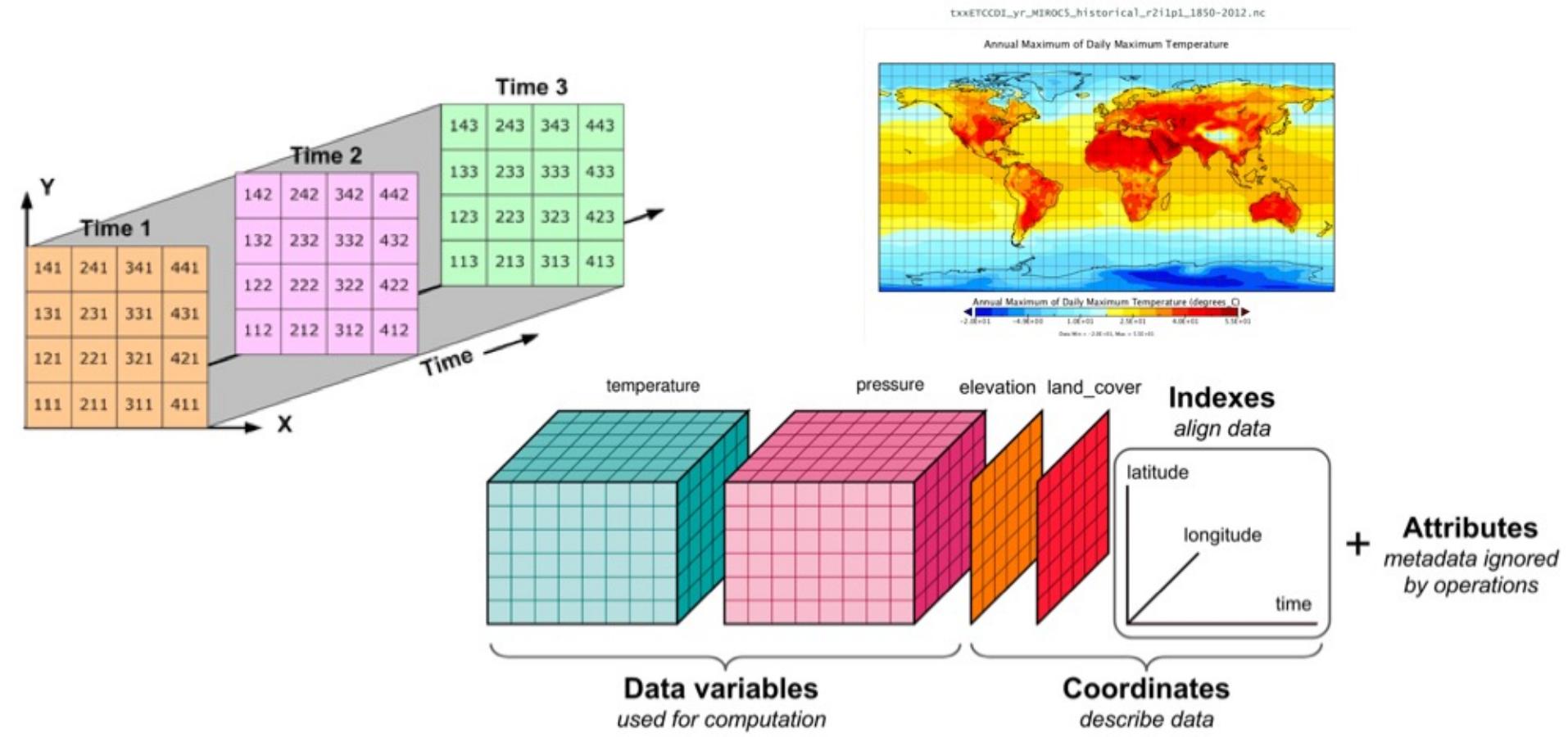


```
[
  {
    "1_param_1": "value_1",
    "1_param_2": "value_2",
    "1_nested": {
      "1_sub_param_1": "sub_value_1",
      "1_sub_param_2": "sub_value_3"
    }
  },
  {
    "2_param_1": "value_1",
    "2_param_2": "value_2",
    "2_nested": {
      "2_sub_param_1": "sub_value_1",
      "2_sub_param_2": "sub_value_3"
    }
  },
  {
    "3_param_1": "value_1",
    "3_param_2": "value_2",
    "3_nested": {
      "3_sub_param_1": "sub_value_1",
      "3_sub_param_2": "sub_value_3"
    }
  }
]
```

# 半结构化文件

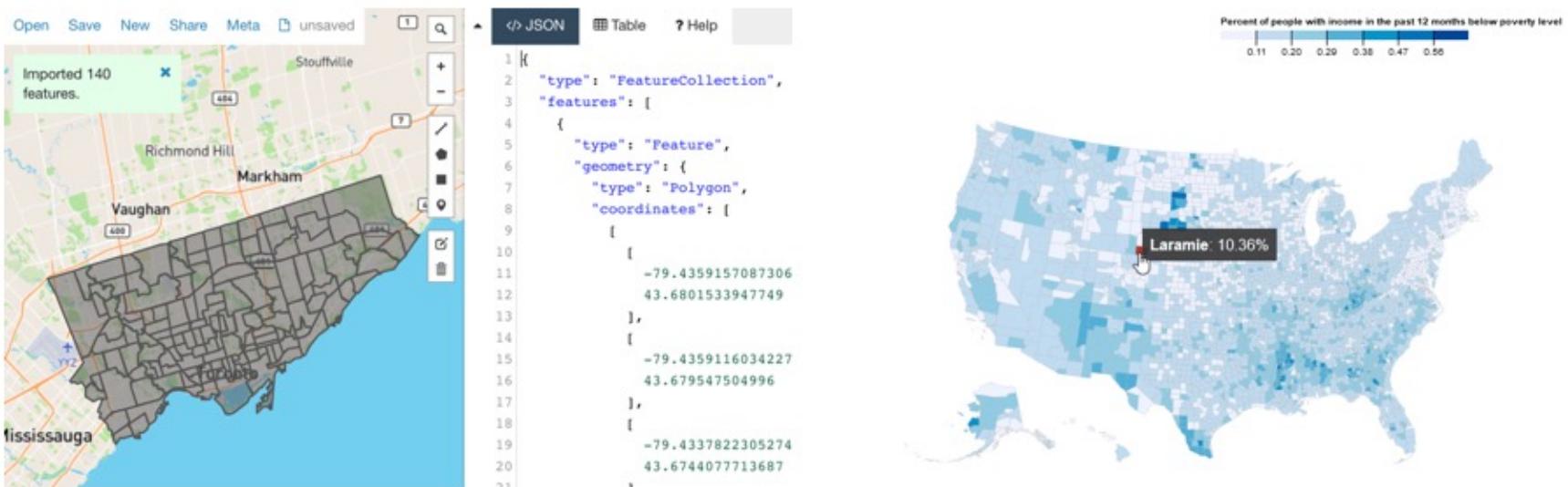
NetCDF（网络通用数据格式）：NOAA气候类数据，是多维时间序列。有三个基本维度（经度、纬度、时间）、可根据需求扩展维度（气温、降水、湿度等）。

广泛应用于大气科学、水文、海洋学、环境模拟、地球物理等领域。



## 半结构化文件

**GeoJSON**和**TopoJSON**: JSON格式的扩展形式，常用于时空分析、网页端地理可视化（Leaflet、D3.js）



Leaflet效果展示：<https://yeshan-geo.github.io/examples/anomaly/map.html>

D3.js效果展示：<https://yeshan-geo.github.io/examples/geojob/geojob.html>

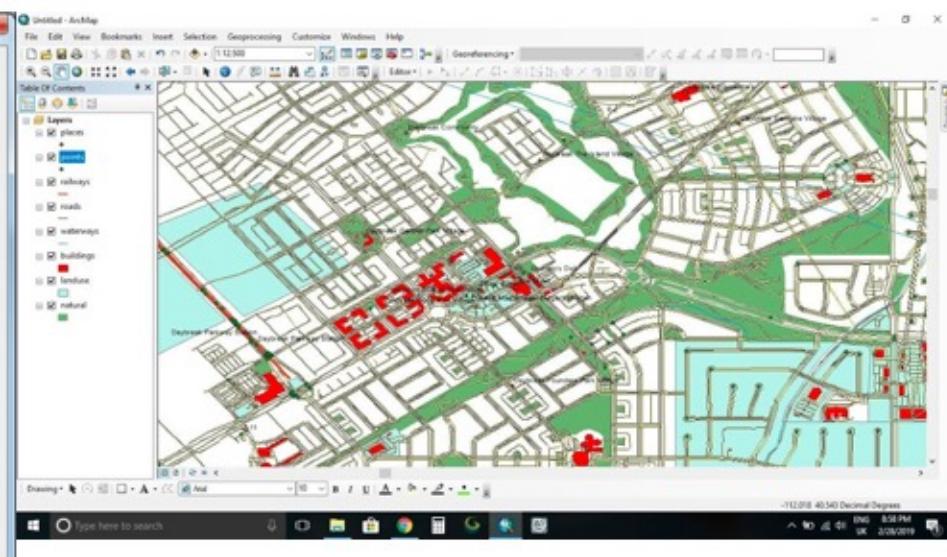
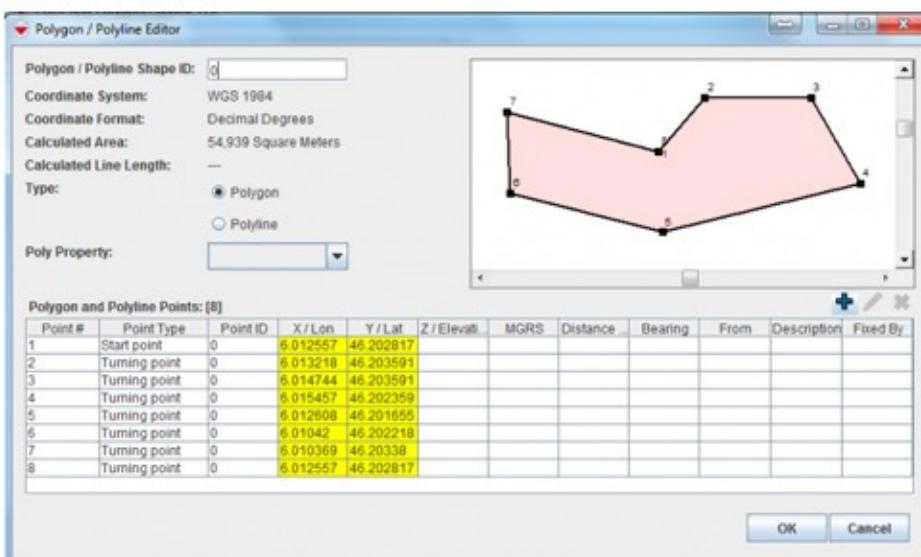
# 半结构化文件

Shapefile: Esri公司开发的地理信息矢量数据格式

将地理要素（道路、河流、城市等）存储为原始几何形状（点、线、多边形），然后把形状与属性连接起来，创建地理数据。

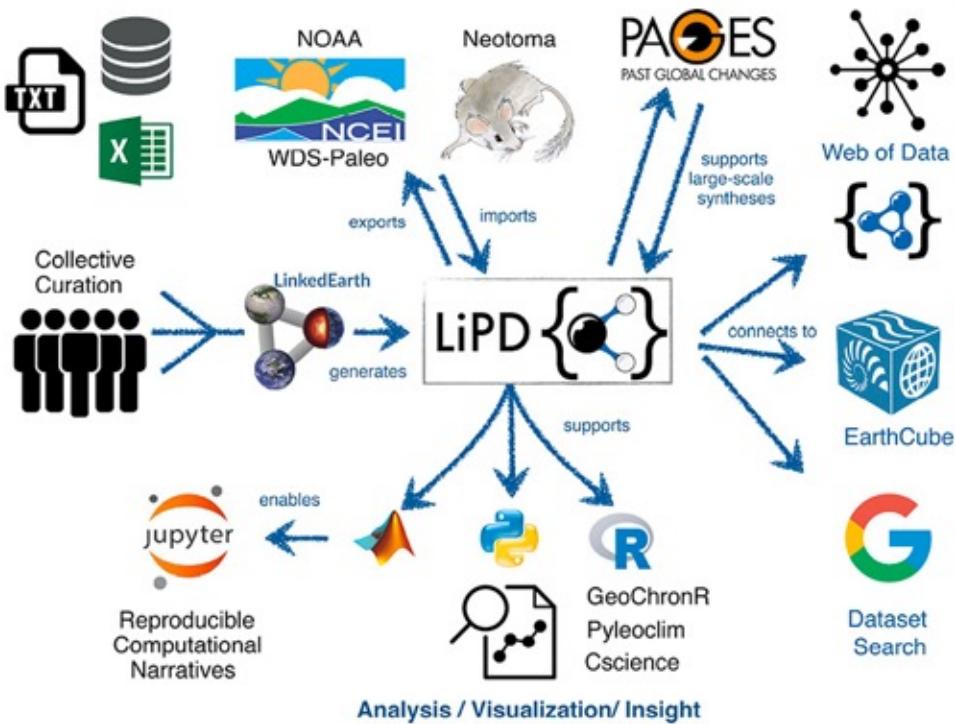
由多个文件配合完成任务：

- .shp: 地理要素的几何形状（端点坐标）
- .shx: 位置索引，方便快速搜索
- .dbf: 要素的属性（列联表）
- .prj: 地理参考系和投影



# 半结构化文件

LipD ( Linked Paleo Data ) : 古气候学领域使用的一种新型数据格式，它旨在简化古气候数据的共享和分析过程，提高数据的重复使用率。



Notes <notes that describe the table as a whole>							
Variables	variableName	variableType	Units	ProxyObservationType	InferredVariableType	TakenAtDepth	InferredFrom
<variable 1>							
<variable 2>							
<variable 3>							
<variable 4>							
<variable 5>							
<variable 6>							
<variable 7>							
<variable 8>							
<variable 9>							
<variable 10>							
<variable 11>							
<variable 12>							
<variable 13>							
<variable 14>							

**OnProxyObservationProperty:** What type are the measuredVariables? Leave blank for inferredVariables and use OnInferredVariableProperty instead). For instance, if the columns contain values for Mg/Ca measurements, the type is Mg/Ca. Use the list or add your own.

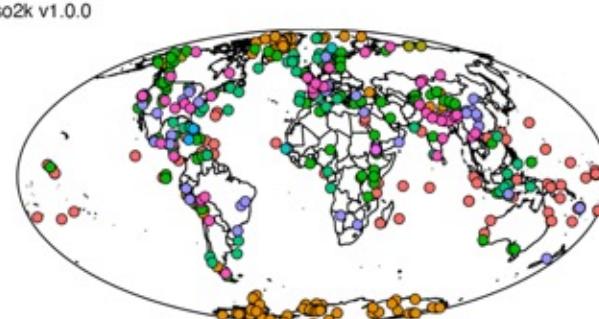
Notes	variableName	variableType	Units	ProxyObservationType	InferredVariableType	TakenAtDepth	InferredFrom
<variable 1>							
<variable 2>							
<variable 3>							
<variable 4>							
<variable 5>							
<variable 6>							
<variable 7>							
<variable 8>							
<variable 9>							
<variable 10>							
<variable 11>							
<variable 12>							
<variable 13>							
<variable 14>							

**Data** Paste Data Table below starting in Column A. Do not paste the headers <variable 1> they will be automatically generated from the metadata

**Missing Value:** <missing value> The value or character string used as a placeholder for missing values

<variable 1> <variable 2> <variable 3> <variable 4> <variable 5> <variable 6> <variable 7>

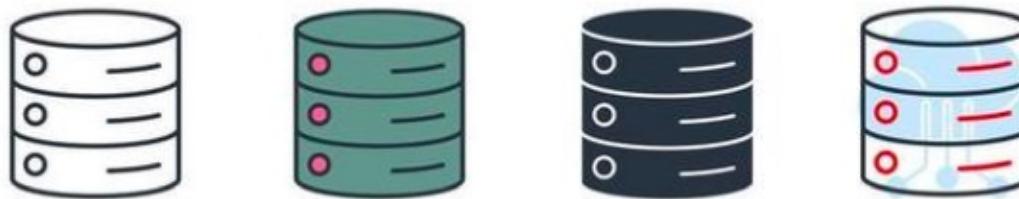
<variable 1 data> <variable 2 data> <variable 3 data> <variable 4 data> <variable 5 data> <variable 6 data> <variable 7 data>



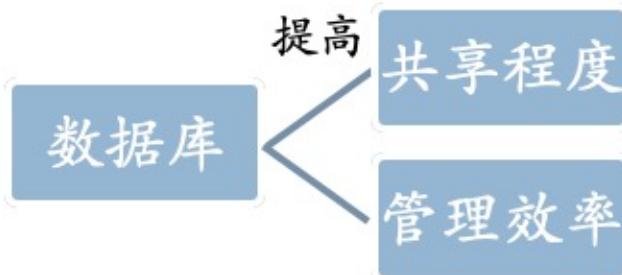
archiveType
Coral
Glaciocene
Groundice
LakeSediment
MarineSediment
MolluskShells
Sclerosponge
Speleothem
TerrestrialSediment
Wood

# 数据库

- 数据库是数据组织的高级形式。
  - 本质：存储在计算设备内的有组织的、共享的、统一管理的数据集合。
- 数据库中保存数据结构
  - 描述数据间的内在联系，便于增删查改
  - 保证数据的独立性、可靠性、安全性与完整性

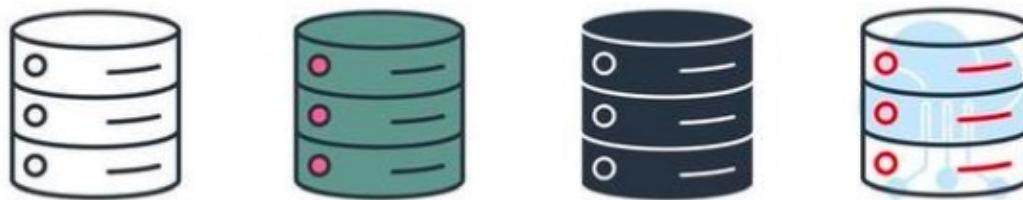


# 数据库



## 数据库分类

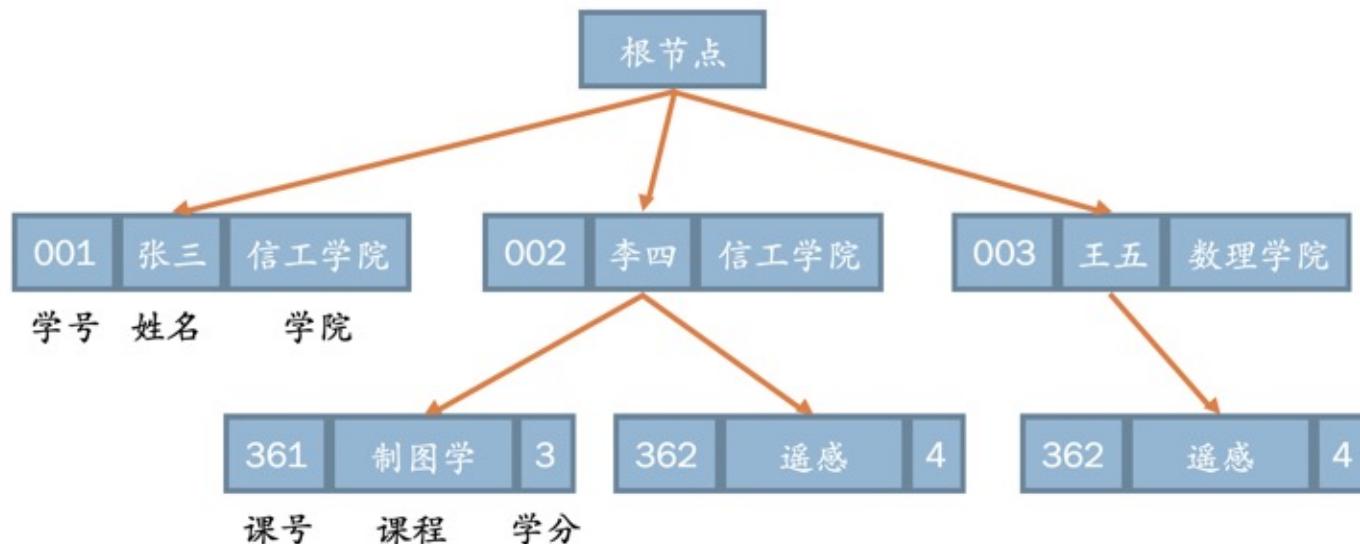
- 分层型数据库 (hierarchical database)
- 关系型数据库 (relational database)
- 面向对象型数据库 (object-oriented database)



# 数据库

## 分层型数据库

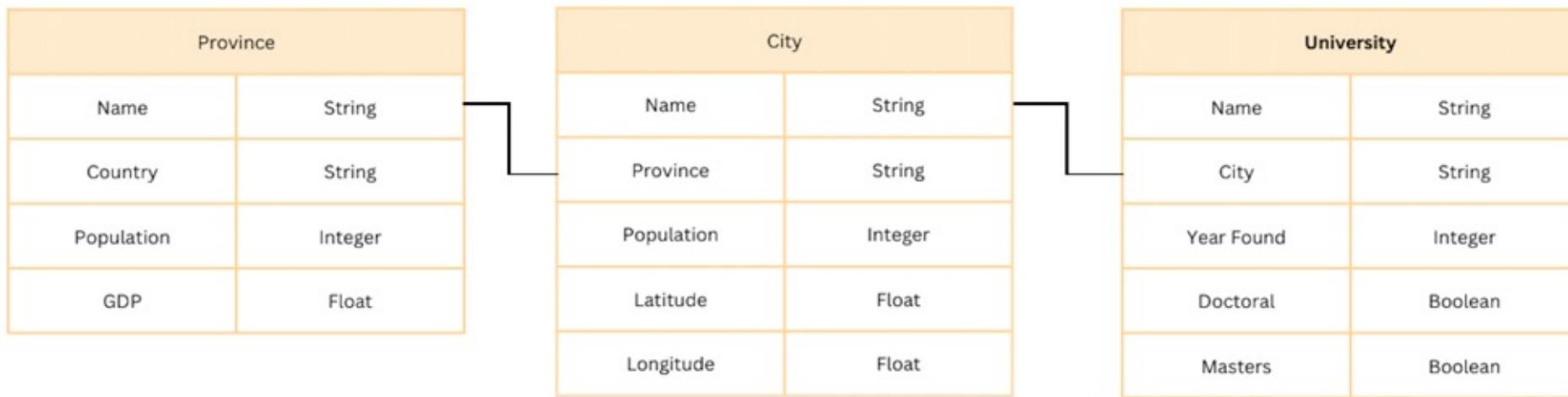
- 数据以记录的形式存储，组织成树状结构，其中一个父节点可以有多个通过链接连接的子节点。
- 适用范围：内部结构清楚、简单、静态的数据。
- 优点：适用于磁带等线性数据存储介质、易于添加和删除信息、访问速度快、支持通过一对多关系工作的系统。
- 缺点：不灵活；无法应对复杂逻辑关系。



# 数据库

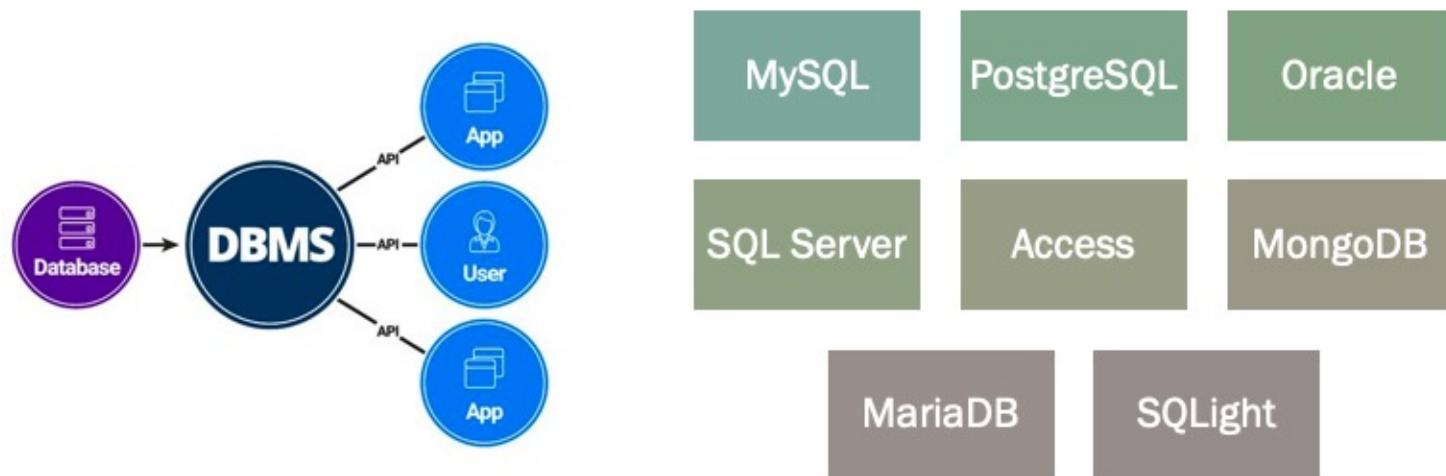
## 关系型数据库

- 当前最常用的数据库类型。
  - 基于数据关系模型，由多个二维表组成，每个表为一个实体类型
  - 表中每一行有唯一标识符，称之为键（主键 key）
  - 不同表之间的逻辑连接称作关系（或联系）
- 优点：灵活提取数据；复杂逻辑关系；ER图（Entity-Relationship，实体-关系图）展示数据的关系和交互。
- 缺点：结构复杂、维护难度较高



# 数据库

- 数据库管理系统（DBMS）：管理数据库的软件，用户可管理数据库中的数据，或管理数据库的结构。
- 标准化查询语言（Structured Query Language/SQL）：关系型数据库的结构化查询语言

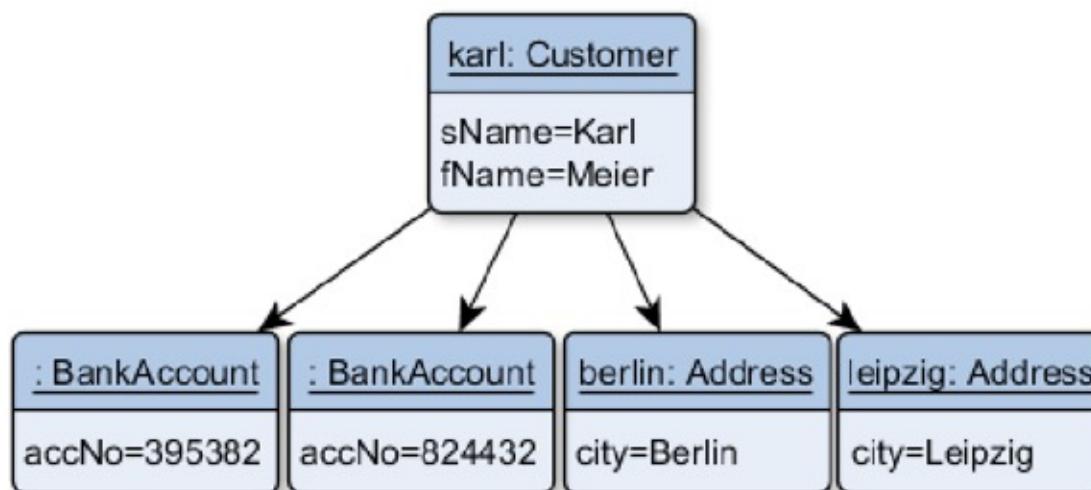


SQL读法：可以读作sequel，也可以三个字母分开读

# 数据库

## 面向对象数据库

- 借用面向对象编程（OOP）的理念设计。多用于Web开发领域，和面向对象的语言（C#、Java、JavaScript、Python等）有较好的融合度。



The diagram shows three relational database tables:

Customers		
ID	SName	FName
1	Karl	Meier

Addresses		
ID	City	CID
1	Berlin	1
2	Leipzig	1

BankAccounts		
ID	AccNo	CID
1	395382	1
2	824432	1

# Pop Quiz

回答正确得2分  
回答错误得1分  
缺席不得分

1. 以下哪一项为非概率抽样?  
A. 分层抽样 | B. 整群抽样 | C. 定额抽样 | D. 简单随机抽样
2. 以下哪种方法有利于保护隐私和数据脱敏?  
A. 分箱法 | B. 合并法 | C. 蒙特卡洛模拟 | D. 自举抽样
3. 以下哪一项**不属于**数据清洗的目标之一?  
A. 代表性 | B. 时效性 | C. 一致性 | D. 完整性
4. 最简单的数据组织管理形式是?  
A. 数据库 | B. 半结构化数据文件 | C. 表单 | D. 字符
5. 以下哪种抽样方法有利于找到具有某些特征的特殊受访人群?  
A. 整群抽样 | B. 滚雪球抽样 | C. 偶遇抽样 | D. 系统抽样



# 数据分析



# 数据分析

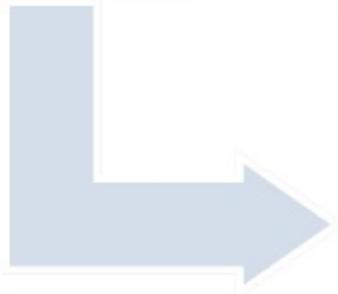
- 用适当的统计分析方法，对数据进行分析研究和概括总结，提取有用信息并形成结论的过程。
- 目的：在大量杂乱的数据中提炼出有价值的信息，总结出所研究对象的规律。



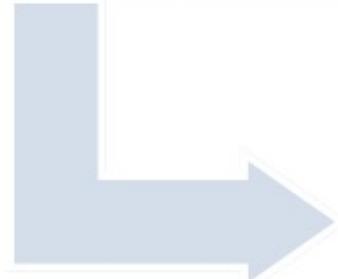
“见人所未见”

# 数据分析

描述性数  
据分析



验证性数  
据分析



探索性数  
据分析

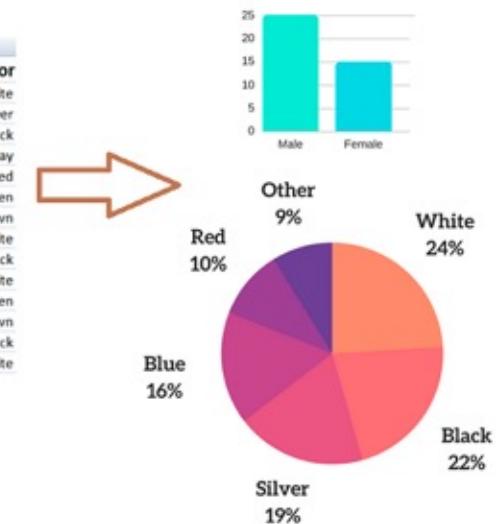
# 描述性数据分析

## 描述性数据分析 (Descriptive Data Analysis, DDA)

- 分析数据的表面特征，包括描述样本的个体特征及其所代表的总体的特征，是一种初级数据分析。
- 常用方法
  - 平均分析法
  - 对比分析法
  - 交叉分析法
  - 分组分析法
  - 结构分析法

	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2		1	22	M
3		2	37	F
4		3	45	F
5		4	62	F
6		5	28	M
7		6	45	M
8		7	88	F
9		8	61	M
10		9	95	M
11		10	27	M
12		11	39	F
13		12	43	M
14		13	55	F
15		14	59	F

RAW DATA

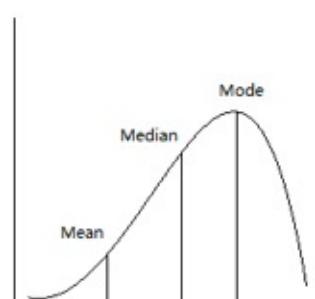


Descriptive Statistics

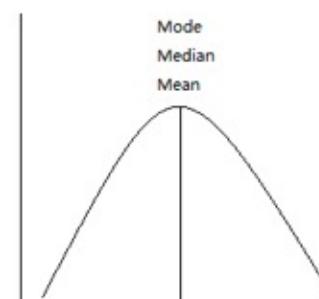
# 描述性数据分析

## 平均分析法

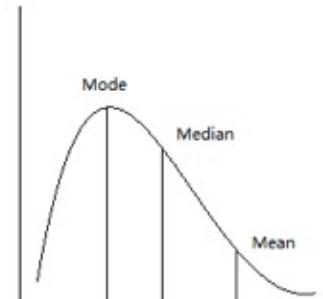
- 计算平均值，观察对象所处的位置和发展水平。
- 数值平均值
  - 所有数据都参与计算得到的平均数
    - 算术平均数（加减关系）
    - 几何平均数（乘除关系，比如产品合格率、银行利率）
  - 操作时可以加权。
- 位置平均数
  - 按照数据的大小顺序或出现的频率选出代表值
    - 中位数
    - 众数



Left skew



Normal Distribution

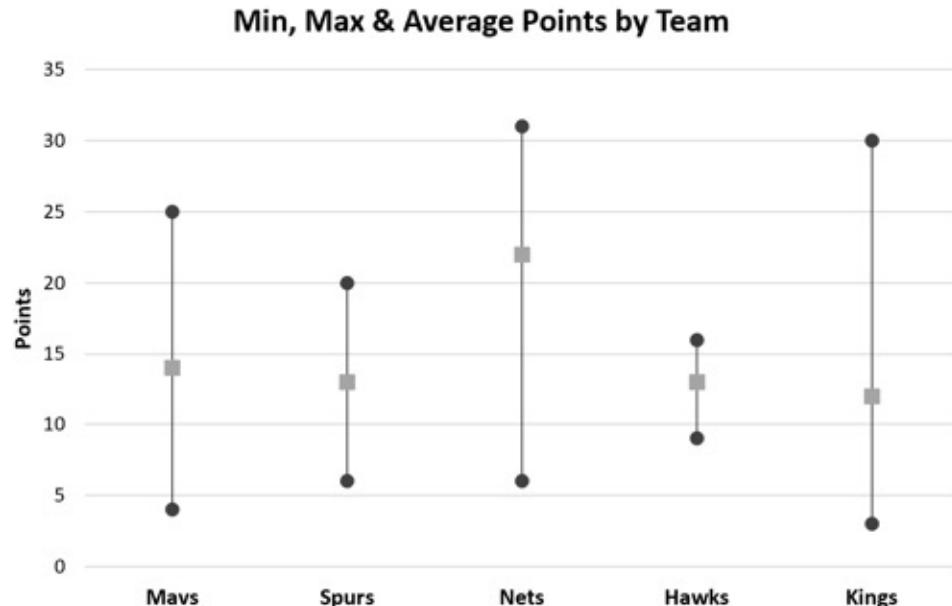


Right skew

# 描述性数据分析

## 对比分析法

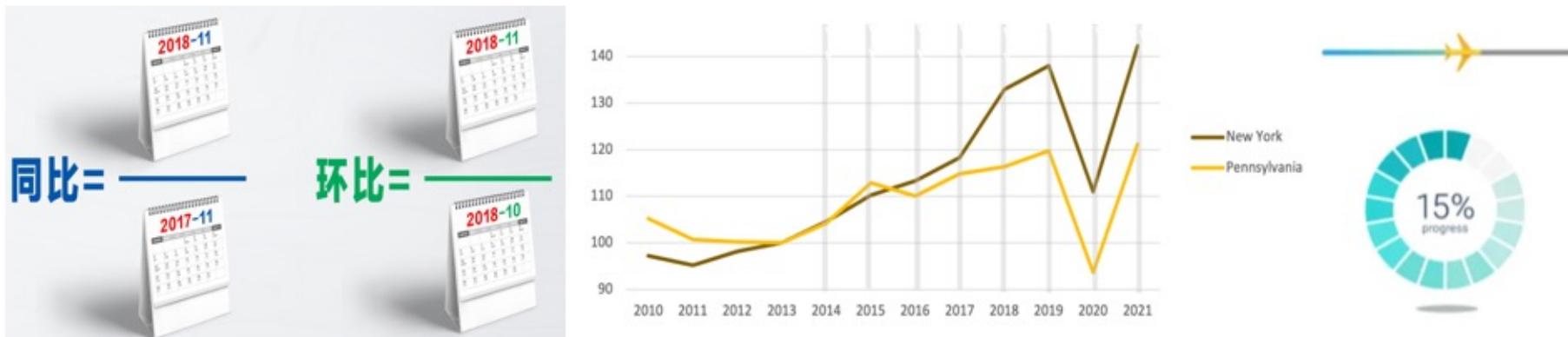
- 将两个或两个以上相互联系的指标数据进行比较，分析其变化情况，从而了解事物的本质特征和发展规律。
- 主要体现在比较对象的规模大小，水平高低，速度快慢等关系的异同情况。
- 对比形式
  - 绝对数
  - 相对数
  - 平均数
  - 百分比
  - 频率



# 描述性数据分析

## 对比分析法

- **时间周期**: 选择不同时间的指标作为对比标准。
  - 同比: 与上一年的同期进行对比, 体现季节周期。
  - 环比: 与本年度的前一时期对比, 分析相邻时间的变化
- **空间指标**: 在同一个时间周期内, 选择不同空间指标数据进行比较。空间不一定是地理概念, 也可以是抽象概念。
- **计划标准**: 分析发展进度和完成率。



# 描述性数据分析

## 交叉分析法

- 两个变量之间的相互关系。
- 常用于市场调研，能让复杂的数据关系简单化。

		What is Your Favorite Baseball Team?			
		Toronto	Boston	New York	
Cross tabulation		Blue Jays	Red Socks	Yankees	Row Totals
In What City Do You Reside?	Boston, MA	11	33	7	51
	Row Percent	21.57%	64.71%	13.73%	34.93%
	Montreal, Canada	23	14	9	46
	Row Percent	50.00%	30.43%	19.57%	31.51%
	Montpellier, VT	22	13	14	49
	Row Percent	44.90%	26.53%	28.57%	33.56%
	Column totals	56	60	30	146
	Column Percent	38.36%	41.10%	20.55%	100.00%

# 描述性数据分析

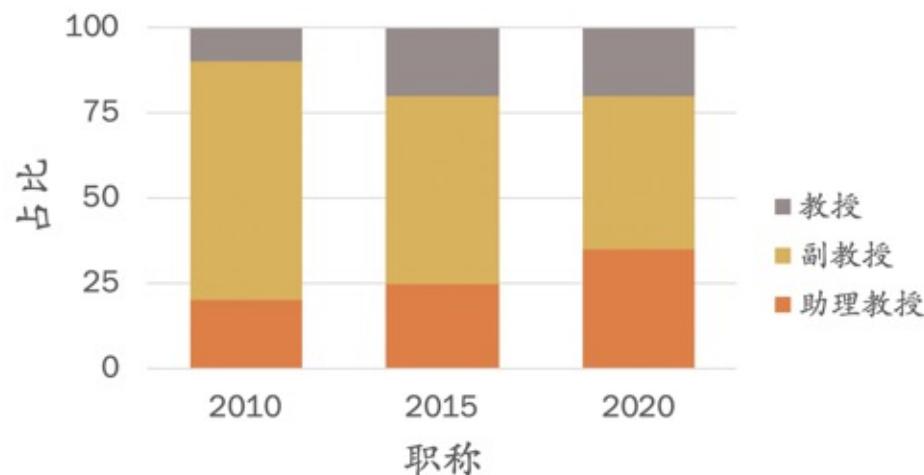
## 分组分析法

- 将总体数据按照某一特征划分成若干个组再加以分析。
- 属性指标
  - 定性描述，不能进行运算，分组逻辑简单
  - 对于复杂问题需用统计分类法多层次划分
- 数量指标
  - 计量数据，可进行运算
  - 根据数量指标将总体划分为若干个性质不同的部分，以分析数据的分布特征和内部联系。
    - 单项式分组：适用于离散型数据，且数据值不多、变动范围较小的情况
    - 组距式分组：在数据变化幅度较大时，将数据总体划分为若干个区间，区间内数据性质相同

# 描述性数据分析

## 结构分析法

- 又称比重分析法
- 在分组分析法的基础上，计算各组成部分占总体的比重
- 特点：强调占比、操作简单
- 用于了解总体中某一部分的重要程度



# 验证性数据分析

## 验证性数据分析 (Confirmatory Data Analysis, CDA)

- 通过数据分析对已有假设模型进行验证，旨在验证假设（证实或证伪）
- 需要提前给出明确的研究问题和假设，并选择适当的统计方法来验证假设。
- 常见的验证性数据分析方法：

参数检验

非参数检验

相关分析

回归分析

因子分析

# 验证性数据分析

## 参数检验

- 基于样本数据来验证关于总体分布参数的假设。
- 当总体分布已知时，根据样本数据对总体分布的统计参数进行推断。
- 常见的参数检验方法
  - T检验
  - 方差分析ANOVA

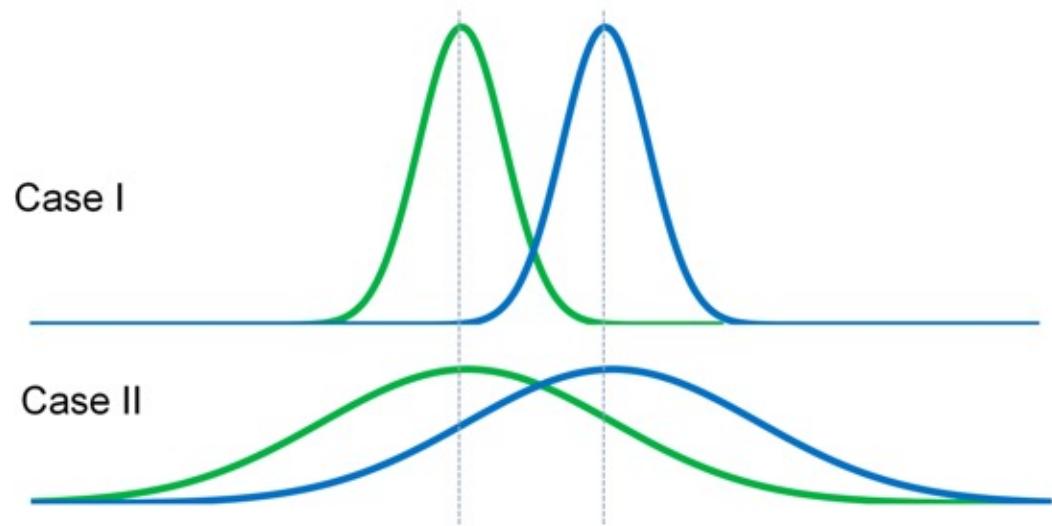
# 验证性数据分析

## t检验 (Student's T-test)

- 用于比较两组的平均值，并确定它们之间的差异是否真实存在。它是一种统计假设检验。



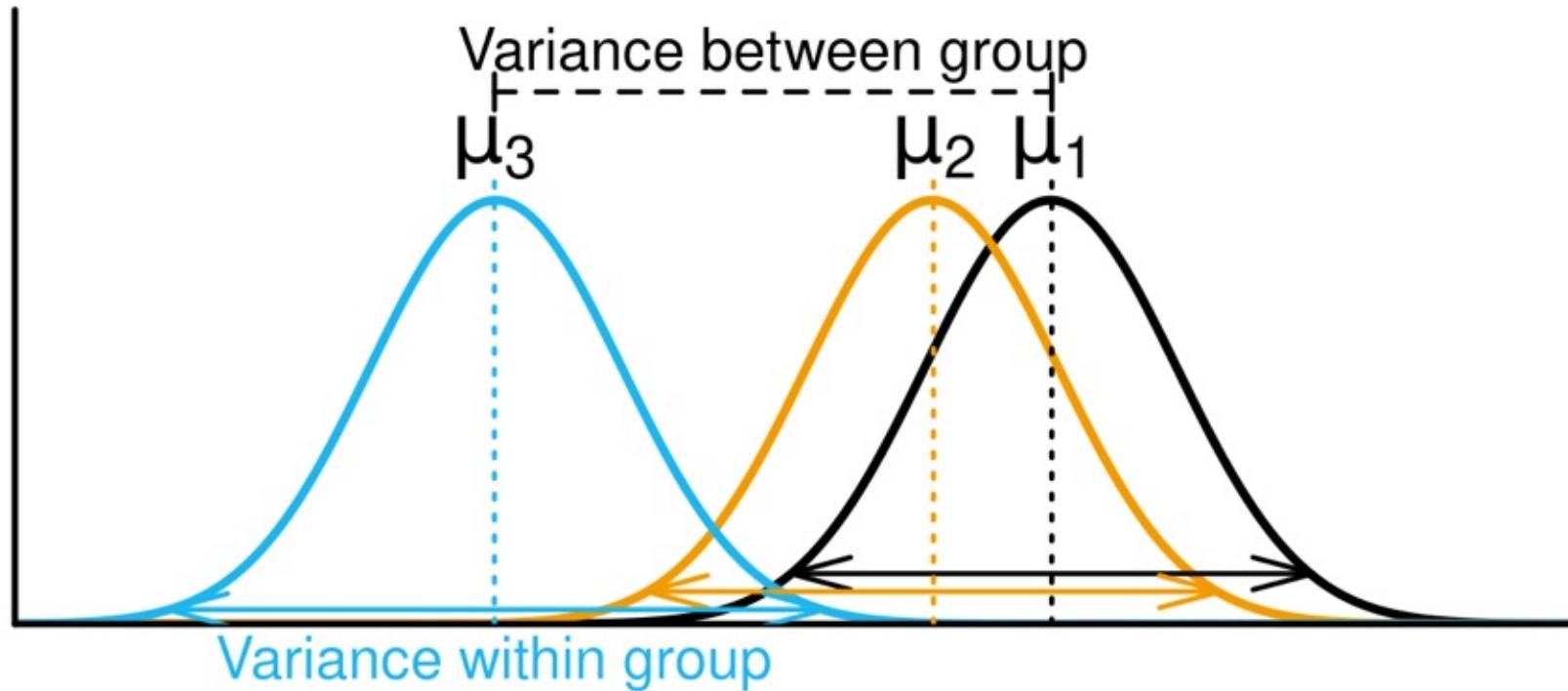
威廉·戈塞 (William T. Gosset)  
都柏林 Guinness 酿酒厂  
笔名 Student



# 验证性数据分析

## 方差分析 (Analysis of variance, ANOVA)

- 用于比较三个或更多组之间的均值差异
- 应用于实验数据分析
- 分为单因素ANOVA和多因素ANOVA



# 验证性数据分析

## 非参数检验

- 在总体方差未知或知道甚少的情况下，利用样本数据对总体分布形态进行推断的方法。在推断过程中不涉及有关总体分布的参数。
- 卡方检验( $\chi^2$ )：用于检验观察频率与期望频率的差异。

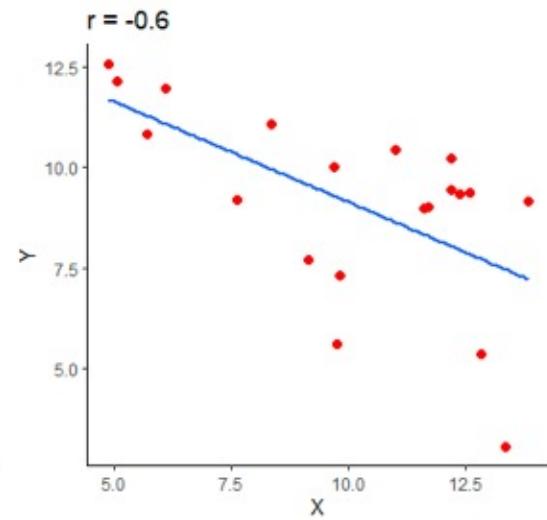
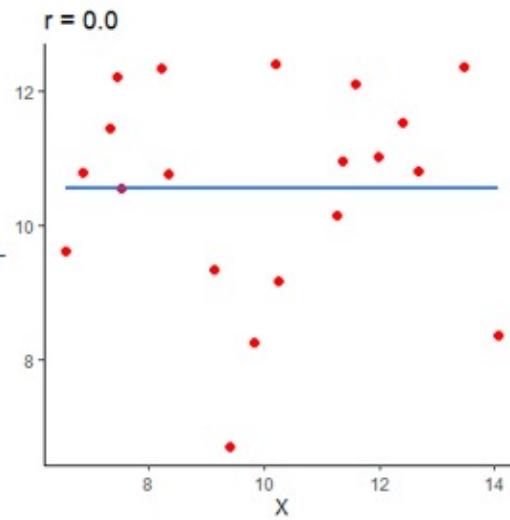
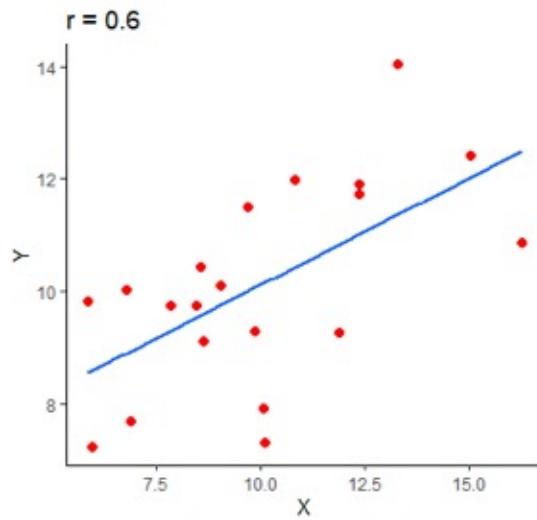


A/B TESTING

# 验证性数据分析

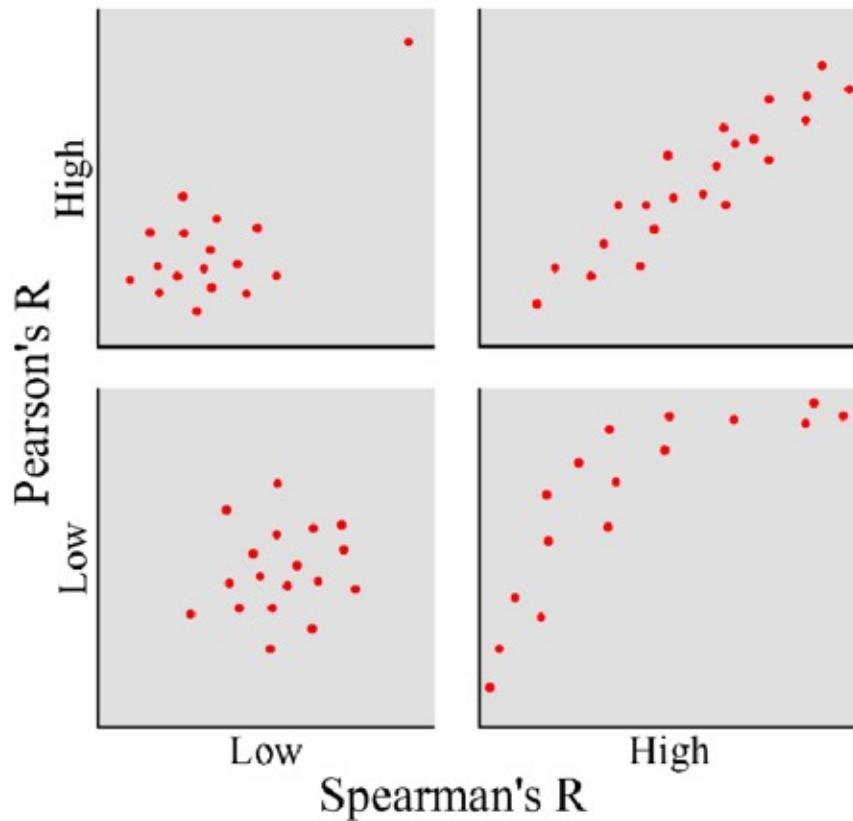
## 相关分析

- 用于评估两个变量之间的关联程度。
- 皮尔森相关系数 (Pearson's  $r$ )
  - 用于度量两组数据的变量X和Y之间的线性相关的程度。
  - 结果取值范围[-1,1]。



# 验证性数据分析

- 斯皮尔曼等级相关系数 (Spearman's rho)
  - 衡量两个变量的相关性的非参数指标。
  - 非参数：数据之间的排位。



# 验证性数据分析

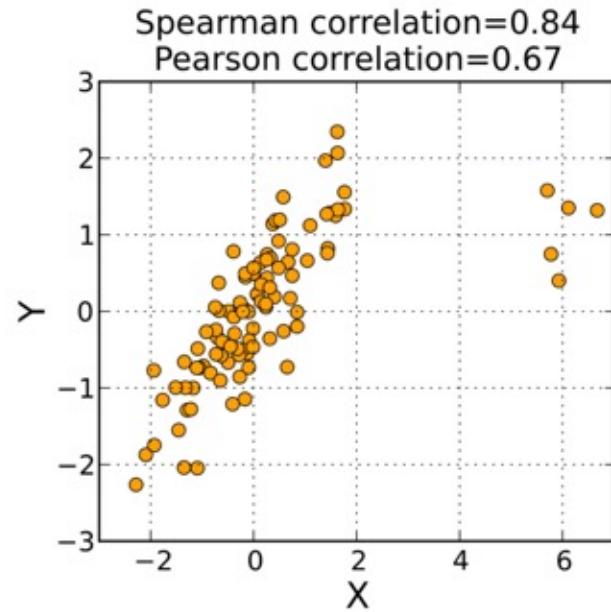
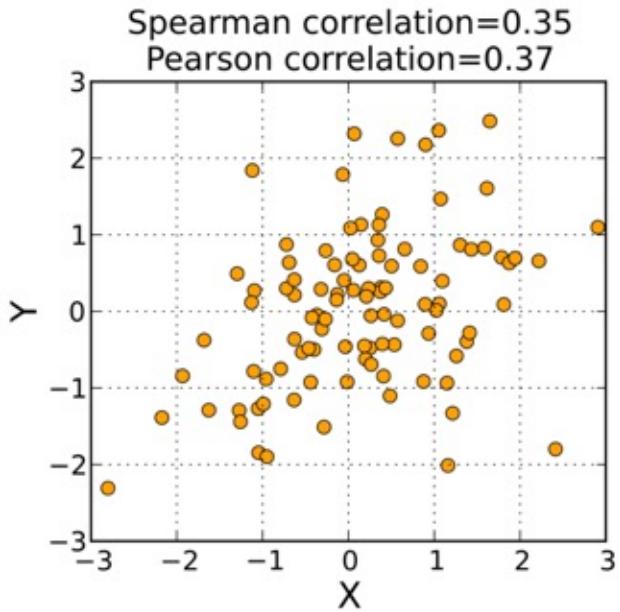
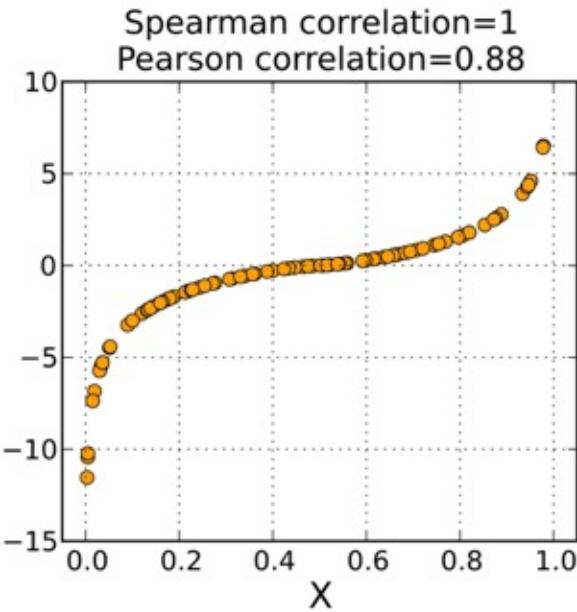
## Pearson's r

- 数据是不是排布在一条上升或下降的直线上（是否成比例地增长或减少）？
- 统计效力高于斯皮尔曼（要求“排成直线”）
- 要求连续型变量、符合正态分布
- 数据有离群值、长尾分布时，效果较差

## Spearman's rho

- 数据是不是在递增或递减（不一定成比例）？
- 不一定要连续型变量、不一定要正态分布
- 数据有离群值或呈长尾分布时，结果相对稳健

# 验证性数据分析



Spearman's rho为1，表明两个变量是单调相关的，但它们之间的相关关系可能并非线性的。相较而言，其Pearson's r并不是1。

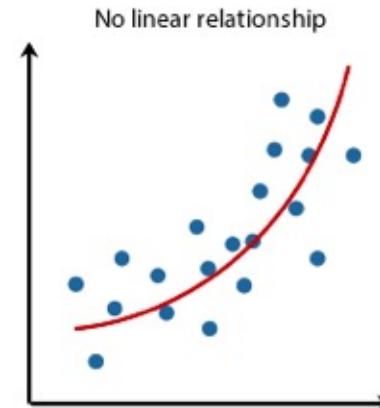
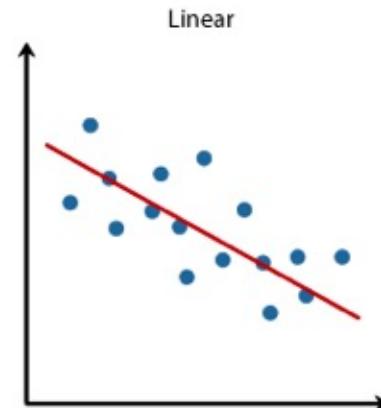
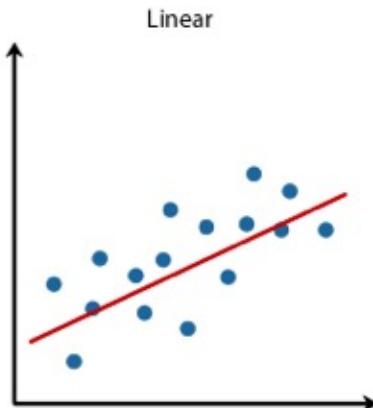
当数据大致呈椭圆分布且没有明显的离群点时，Pearson's r和Spearman's rho接近。

对于样本中的离群点，Pearson's r更加敏感。

# 验证性数据分析

## 回归分析

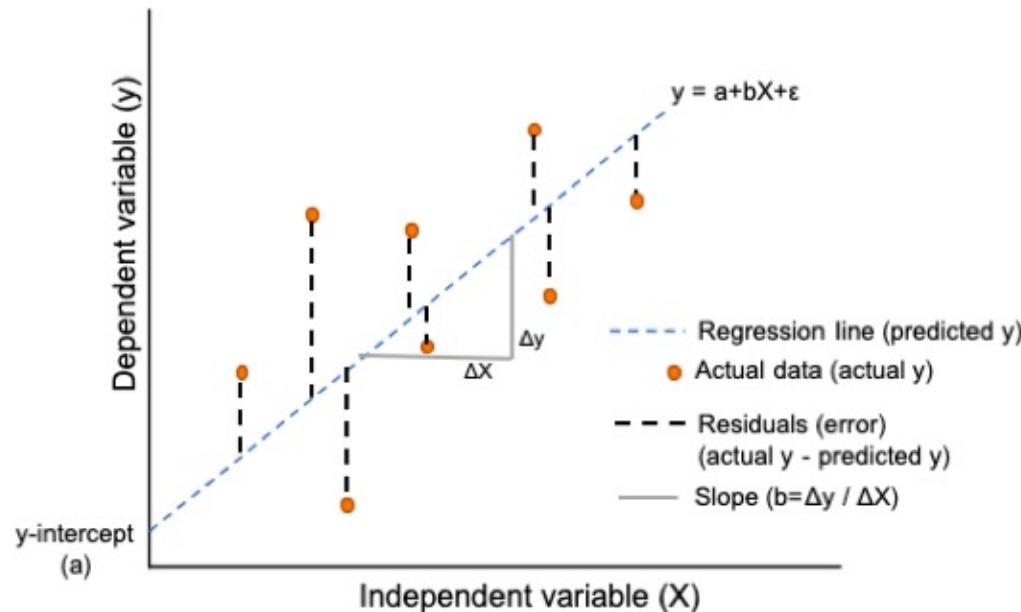
- 用于研究自变量与因变量之间的关系，并预测因变量的值。
- 回归的几何意义：用一条具有代表性的直线或曲线（高维空间的超平面）来拟合输入和输出的数据点。
- 在统计学中，回归分析指确定两种或两种以上变量间相互依赖的定量关系的。
- 按照自变量和因变量之间的关系类型，可分为线性回归和非线性回归。



# 验证性数据分析

## 线性回归

- $y = a + bx + \epsilon$ , 其中  $y = a + bx$  为回归线,  $\epsilon$  为误差 (服从均值为 0 的正态分布)
- 假设目标值与特征之间存在线性相关。通过考虑误差  $\epsilon$  的大小, 构建损失函数。
- 当损失函数输出值 (误差) 最小时, 计算参数  $a$  和  $b$ , 获得线性方程。



# 验证性数据分析

## 因子分析

- 用于探索数据中的潜在因子结构，并将观测到的变量归纳到少量的潜在因子中。

	Factor 1	Factor 2	Factor 3
	0.98	-0.04	0.02
	0.95	-0.01	0.18
	0.92	0.04	-0.05
	0.62	0.17	-0.33
	-0.03	0.97	-0.01
	-0.02	0.96	0.09
	-0.07	0.95	0.09
	-0.09	0.19	0.96
	0.03	0.09	0.95
	-0.19	-0.42	-0.09

# 探索性数据分析

## 探索性数据分析 (Exploratory Data Analysis, EDA)

- 运用计量手段从大量数据中发现未知且有价值的信息的过程
- 侧重于在数据之中发现新的特征和规律。
- 常规的工作内容
  - 检查数据（对数据进行清洗）
  - 对数据进行描述（可视化、查看数据的分布）
  - 考察变量之间的关系（比较数据变量，培养对数据的直觉，对数据进行总结）
- 常见的分析方法
  - 回归分析
  - 相关分析
  - 多维尺度分析

# 探索性数据分析

## 探索性数据分析步骤

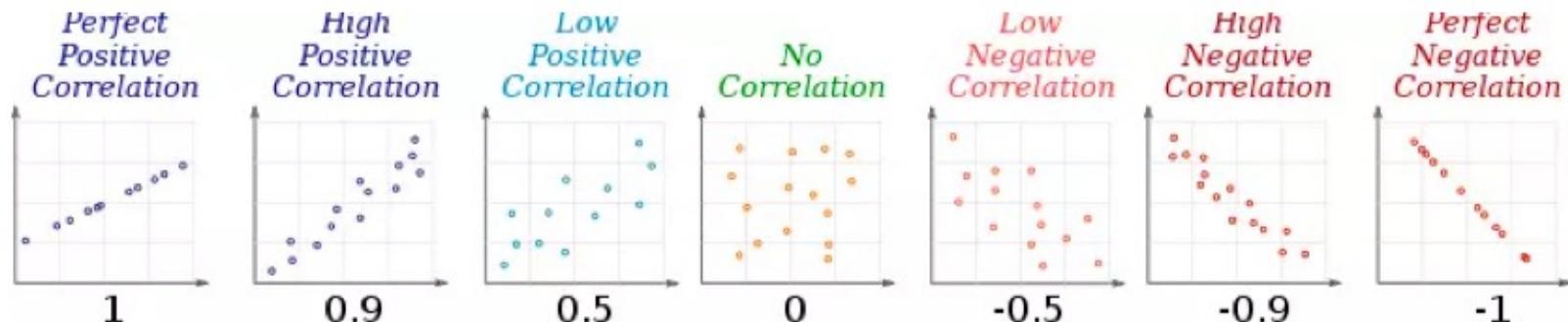
- 1. 检查数据
  - - 是否有缺失值? 是否有异常值? 是否有重复值?
  - - 样本是否均衡? 是否需要抽样?
  - - 变量是否需要转换?
  - - 是否需要增加新的特征?
- 2. 可视化
  - 1) 连续变量
    - 描述性统计量: 平均值, 中位数, 众数, 极值, 四分位数, 标准差
    - 频数分布表, 直方图, 箱线图、小提琴图
  - 2) 无序型离散变量
    - 各个变量出现的频数和占比
    - 频数分布表, 柱形图, 条形图, 茎叶图, 饼图
  - 3) 有序型离散变量
    - 各个变量出现的频数和占比
    - 频数分布表, 堆积柱形图, 堆积条形图

# 探索性数据分析

## 探索性数据分析步骤

- 3. 考察变量之间的关系：

- 1) 连续变量和连续变量 (Continuous & Continuous)：
  - 两个连续变量：散点图
  - 多个连续变量：散点图矩阵、相关系数矩阵、热图、气泡图
  - 量化指标：Pearson's r, Spearman's rho



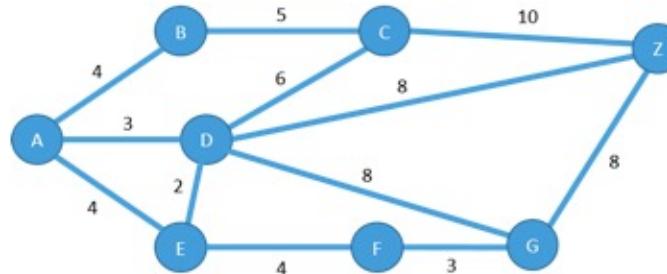
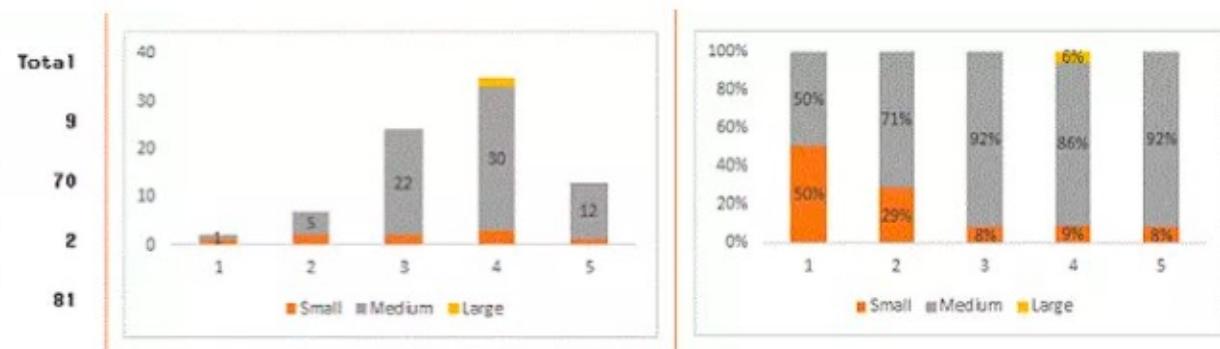
# 探索性数据分析

## 探索性数据分析步骤

- 2) 离散变量和离散变量 (Discrete & Discrete)：
  - 两个离散变量：交叉分组表 (crosstab)、复合柱形图、堆积柱形图、饼图
  - 多个离散变量：图（节点之间的边及权重显示关系的强弱）
  - 量化指标：卡方检验

		Product Category				
		1	2	3	4	5
Frequency	Row Pct	1	2	2	3	1
Small	11.11	22.22	22.22	33.33	11.11	
Medium	1.43	7.14	31.43	42.86	17.14	
Large	0.00	0.00	0.00	100.00	0.00	
Total	2	7	24	35	13	

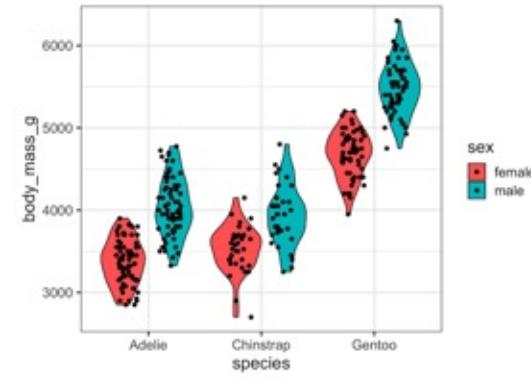
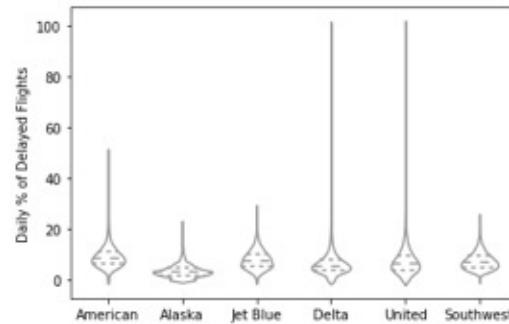
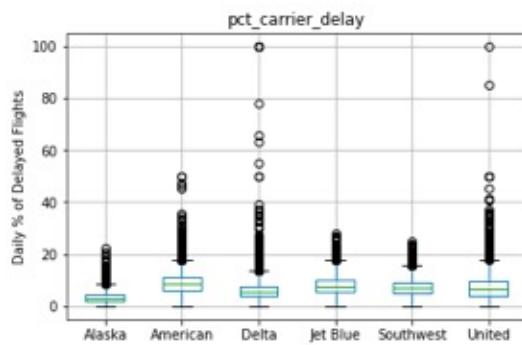
Frequency Missing = 77



# 探索性数据分析

## 探索性数据分析步骤

- 3) 离散变量和连续变量 (Discrete & Continuous) :
  - 直方图, 箱线图, 小提琴图
  - 量化指标: 独立样本t检验中的t统计量, 和相应的p值 (两个变量), ANOVA中的 $\eta^2$  (三个变量及以上)



# 探索性数据分析

## Chart Suggestions—A Thought-Starter

