大数据核心课程

商务大数据分析

西安交通大學管理学院
THE SCHOOL OF MANAGEMENT
XI'AN JIAOTONG UNIVERSITY

2. 探索性数据分析

刘跃文 博士

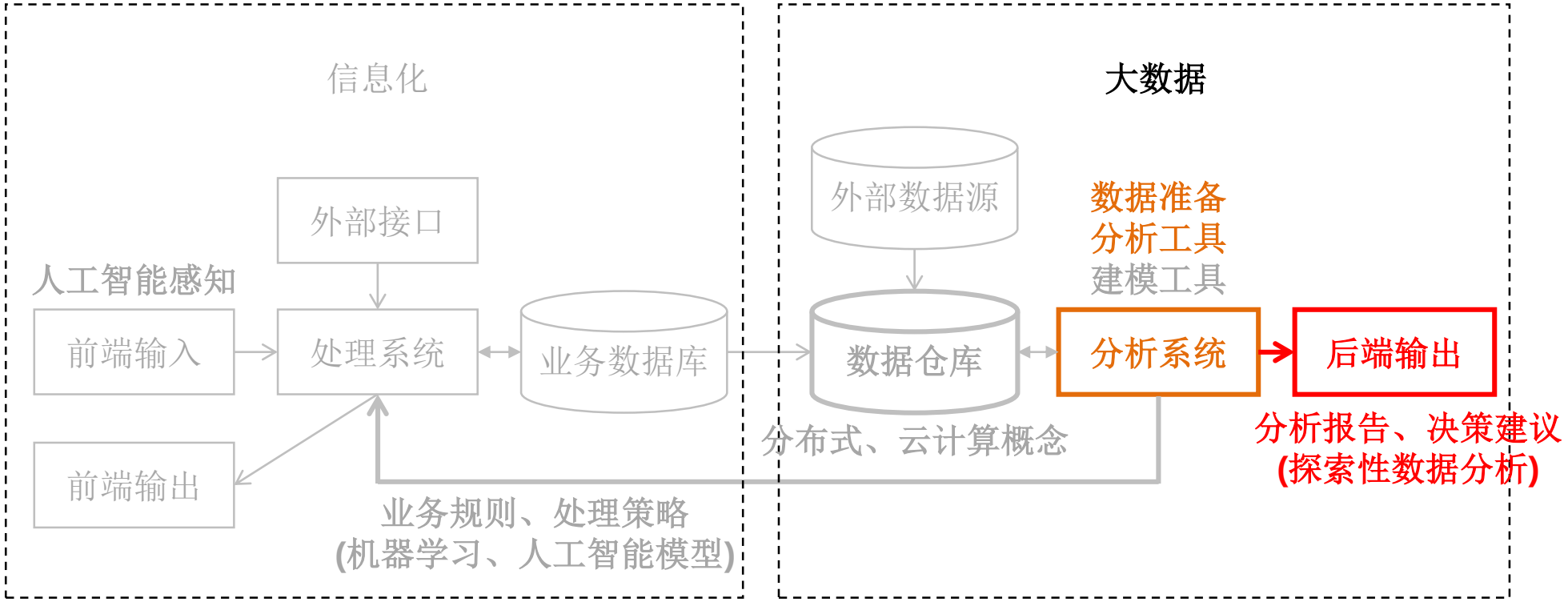
教授、博士生导师

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

V2.3, 2021-9-25

知识地图



提纲

2.1 数据的来源与种类
2.2 静态数据探索
2.3 动态数据探索
2.4 数据分析的目标
2.5 时刻注意数据质量
2.6 撰写数据分析报告

大数据核心课程

商务大数据分析 >>2. 探索性数据分析

2.1 数据的来源与种类

刘跃文 博士

教授、博士生导师

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

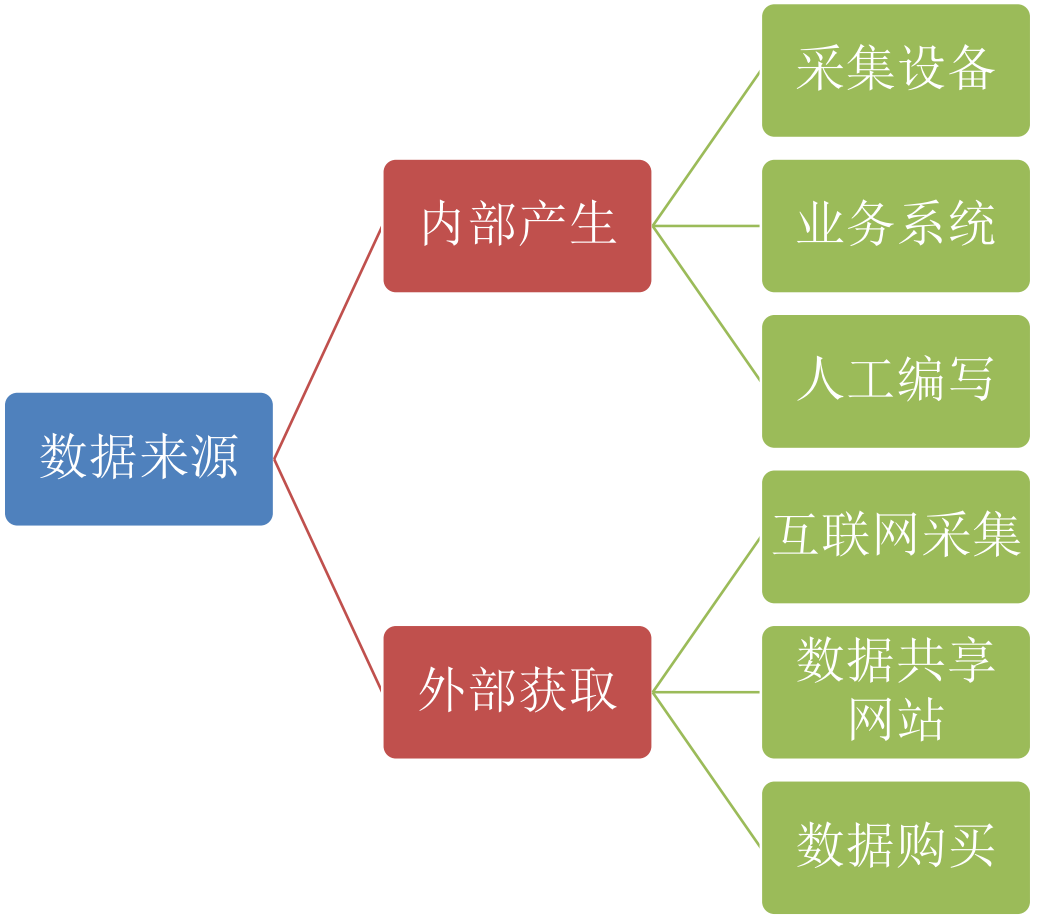
V2.2, 2021-9-8

提纲

1. 数据的来源
2. 数据的种类

1. 数据的来源

1. 数据的来源



例如：摄像头、NFC标签、雷达、各类传感器

例如：生产系统、进销存系统、人资系统、扫码系统、OA等

例如：各类登记表、各类文档、知识分享等

例如：抓取的微博、淘宝、京东等平台上的数据（仅包括能公开访问的数据）

例如：各类数据竞赛平台、科研平台、政府数据开放平台上的数据

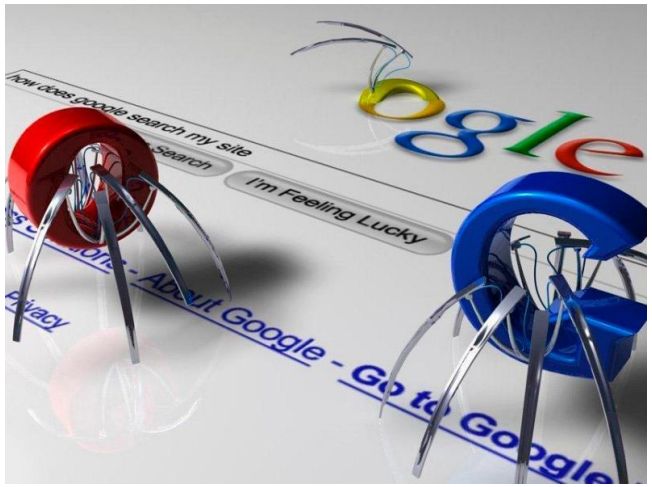
例如：从网上购买号码归属地表、数据交易平台等

开放数据平台

- 政府数据开放平台
  - 美国政府数据开放平台
  - 中国政府数据开放平台
- 数据竞赛平台
  - Kaggle
  - 阿里天池
  - https://www.zhihu.com/question/36374964
- 科研数据共享平台
- 互联网公司开放API接口
  - 高德/百度地图数据下载

2. 网络数据采集简介 [参考]

- 网页数据采集器
  - 网络爬虫
  - Crawler
  - Spider



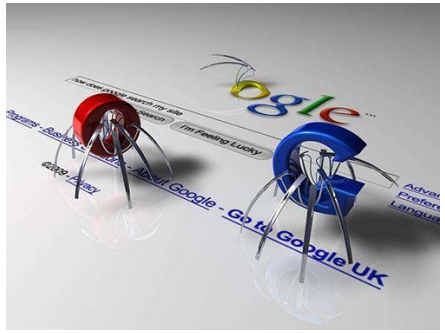
- 主要步骤：

下载页面 → 内容解析 → 数据存储

互联网上，人类只是少数派 [参考]

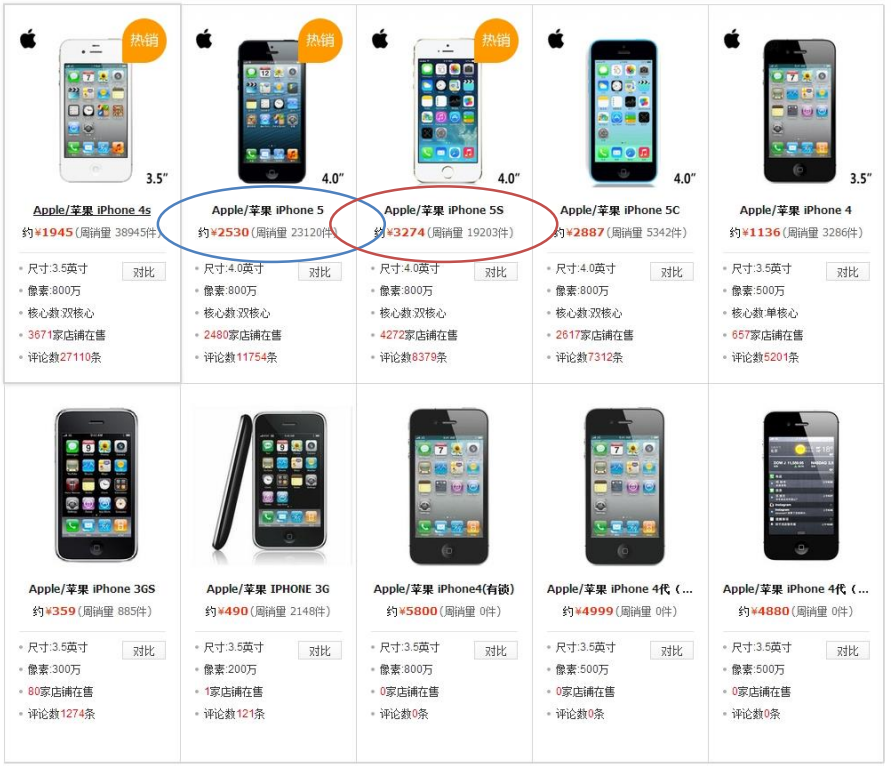
网络数据采集原理 [参考]

- 模拟客户端：假装你是浏览器
- HttpClient 是 Apache Jakarta Common 下的子项目，可以用来提供高效的、最新的、功能丰富的支持 HTTP 协议的客户端编程工具包，并且它支持 HTTP 协议最新的版本和建议。
  1. 创建 HttpClient 的实例
  2. 创建某种连接方法的实例，在这里是GetMethod。在 GetMethod 的构造函数中传入待连接的地址
  3. 调用第一步中创建好的实例的 execute 方法来执行第二步中创建好的 method 实例
  4. 读 response
  5. 释放连接。无论执行方法是否成功，都必须释放连接
  6. 对得到后的内容进行处理
- 可以声称是爬虫（例如百度爬虫），但是可能会被拒绝
  - http://www.taobao.com/robots.txt



内容解析：寻找规律 [参考]

```
2450
2451            </div>
2452          </div>
2453        <!--列表扩展的部分 over-->
2454      </li>
2455                                  <li>
2456            <div class="pic">
2457                  <a href="detail.htm?cat=1512&amp;spuid=207729829#content
2458" target="_blank"><img data-ks-lazyload="http://img.taobaocdn.com/bao/uploaded/i7/T1CcYUFkthXXXtkpnc_095556.jpg_180x180.jpg" alt="" /></a>
2459            </div>
2460                                        <i class="hot"></i>
2461                                    <div class="title">
2462                                <p class="general"><a href="detail.htm?cat=1512&amp;spuid=207729829#content
2463" title="Apple/苹果 iPhone 5"><strong><b>Apple/苹果 iPhone 5</b></strong></a></p>
2464                                  <p><font>约</font><span>¥</span><em>2530</em></span>(周销量 23120件)</p>
2465                  </div>
2466                  <div class="message">
2467                        <div class="compare J_listCompare" data-spuid="207729829" data-spu='{"id":207729829,"name":"Apple/苹果 iPhone
2468
2469
2470
2471
2472
2473                                  <p><a target="_blank" href="detail.htm?cat=1512&amp;spuid=207729829#item-container
2474"><span>2480</span>家店铺在售</a></p>
2475                                  <p><a target="_blank" href="detail.htm?cat=1512&amp;spuid=207729829#spu_comments
2476">评论数<span>11754</span>条</a></p>
2477                  </div>
2478              <!--列表扩展的部分 start-->
2479          <div class="itemTag">
2480            <div class="item-T">
2481              <p class="item-Tcon">
2482
2483            </div>
2484          </div>
2485        <!--列表扩展的部分 over-->
2486      </li>
2487                                  <li>
2488            <div class="pic">
2489                  <a href="detail.htm?cat=1512&amp;spuid=229361412#content
2490" target="_blank"><img data-ks-lazyload="http://img.taobaocdn.com/bao/uploaded/i1/T1S33bFjxgXXbWfpPc_095625.jpg_180x180.jpg" alt="" /></a>
2491            </div>
2492                                        <i class="hot"></i>
2493                                    <div class="title">
2494                                <p class="general"><a href="detail.htm?cat=1512&amp;spuid=229361412#content
2495" title="Apple/苹果 iPhone 5S"><strong><b>Apple/苹果 iPhone 5S</b></strong></a></p>
2496                                  <p><font>约</font><span>¥</span><em>3274</em></span>(周销量 19203件)</p>
2497                  </div>
2498                  <div class="message">
2499                        <div class="compare J_listCompare" data-spuid="229361412" data-spu='{"id":229361412,"name":"Apple/苹果 iPhone
2500
2501
2502
2503
2504
2505                                  <p><a target="_blank" href="detail.htm?cat=1512&amp;spuid=229361412#item-container
2506"><span>4272</span>家店铺在售</a></p>
2507                                  <p><a target="_blank" href="detail.htm?cat=1512&amp;spuid=229361412#spu_comments
2508">评论数<span>8379</span>条</a></p>
2509                  </div>
2510              <!--列表扩展的部分 start-->
2511          <div class="itemTag">
2512            <div class="item-T">
2513              <p class="item-Tcon">
2514
2515            </div>
2516          </div>
2517
```

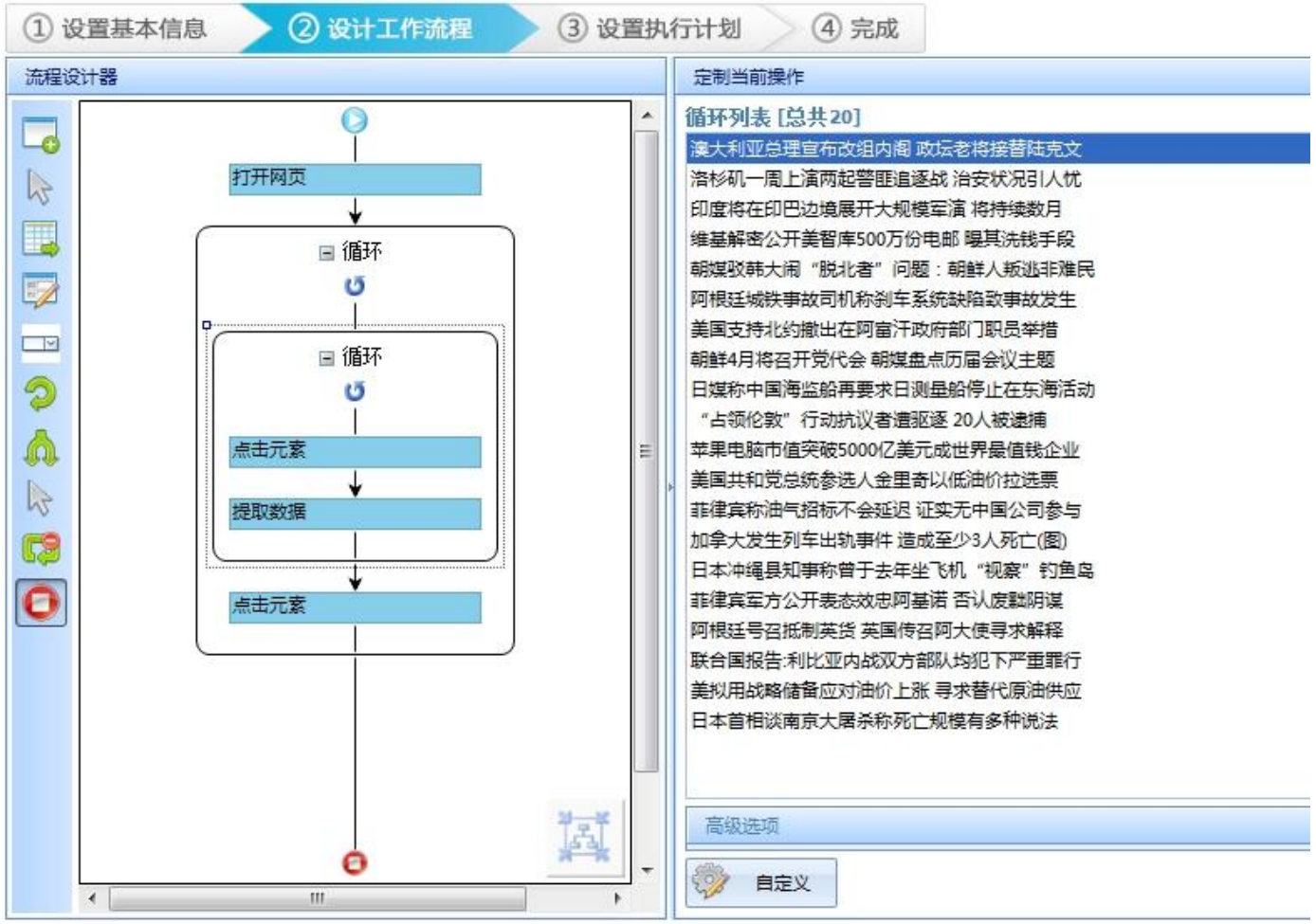


数据保存 [参考]

- 文本文件
  - 优点：简单直接
  - 缺点：文件较大时效率较低、查找不方便

- 数据库
  - 优点：去重、查找方便、上一步输出可作为下一步输入、数据采集过程的良好控制
  - 缺点：写程序稍微麻烦一点

数据采集器——八爪鱼采集器 [参考]



2. 数据的种类

1. 数据的种类

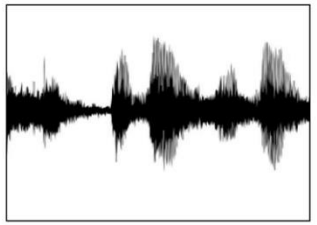
- 按照数据是否结构化
  - 结构化数据：行数据，存储在数据库里，可以用二维表结构来逻辑表达实现的数据（规整好的表格）。
  - 非结构化数据：视频、图像、音频、文本、文档、网页等。
  - 半结构化数据：具有一定的结构性，但结构变化很大，如XML。

结构化数据
Structured Data

Size	#bedrooms	···	Price (1000$s)
2104	3		400
1600	3		330
2400	3		369
⋮	⋮		⋮
3000	4		540

User Age	Ad Id	···	Click
41	93242		1
80	93287		0
18	87312		1
⋮	⋮		⋮
27	71244		1

非结构化数据
Unstructured Data



Audio

Image

Four scores and seven years ago...

Text

结构化数据可用于分析；半结构化数据可用于查询；非结构化数据要先结构化（人工智能）。

计量经济学的分类法 [参考]

- 从时空的角度来看：
  - 横截面数据(Cross-sectional data)：在某一时点收集的不同对象的数据。
  - 时间序列数据(Time-series data)：对同一对象在不同时间连续观察所取得的数据。
  - 纵向数据(Longitudinal data)或面板数据(Panel data)：是截面数据与时间序列综合起来的一种数据资源。

Cross Sectional Data

Time Series Data

Panel Data

Query1

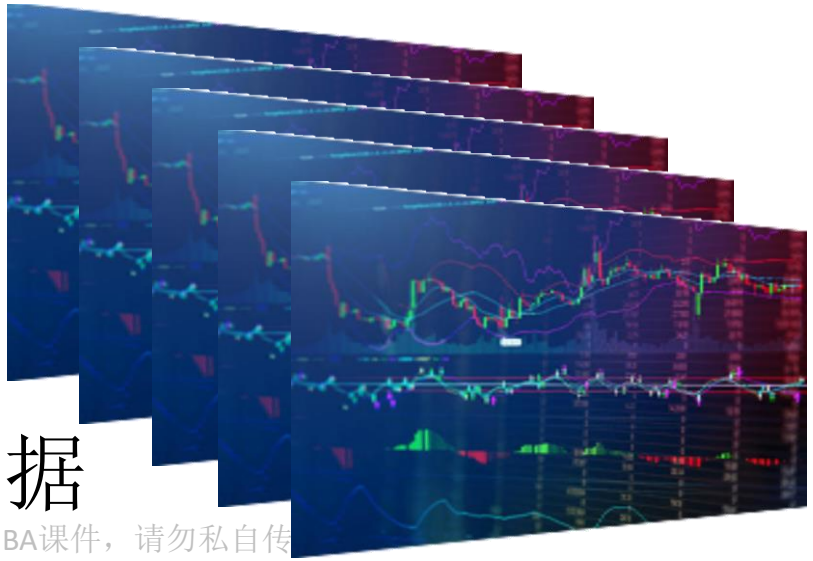
date	四环路以内	四至五环路	五至六环路	六环路以外
2009-02-01	14226	14447	7206	6703
2009-03-01	15124	14797	8258	6836
2009-04-01	15593	15204	8941	7226
2009-05-01	16510	15744	9308	7550
2009-06-01	16967	15748	9296	7603
2009-07-01	17478	16663	9430	7886
2009-08-01	18113	16868	9749	7847
2009-09-01	19109	17205	9980	7878
2009-10-01	19750	17391	10314	801

以股票数据为例 [参考]

横截面数据

时序数据

=



面板数据

从商业数据分析的角度[重点]

- 从商业数据分析（用途）的角度：
  - 静态数据（横截面数据）：登记数据、注册数据、基本信息
  - 动态数据（不是面板数据）：生产、交易类数据 {ID、时间、地点（ID）、事件、属性}
- 静态与动态数据举例：
  - 车辆登记信息，车辆出行记录（如高速收费数据）
  - 会员信息，会员交易记录
  - 银行账户信息，银行交易流水
  - 电话号码注册信息，通话记录

银行卡开户业务申请表

年 月 日

客户信息	申请人资料	中文姓名		英文名		性别	出生日期	
		证件类型		证件号码				
		通讯地址						
	监护人	中文姓名		英文名		性别		
		与申请人关系		电话：				
	储蓄帐户	开户						
	存款信息	币种		金额				
银行记录	交易时间：	经办人：						
客户确认	本人同意贵行根据所申请的项目服务。							
	申请人或监护人签名：							

华夏银行

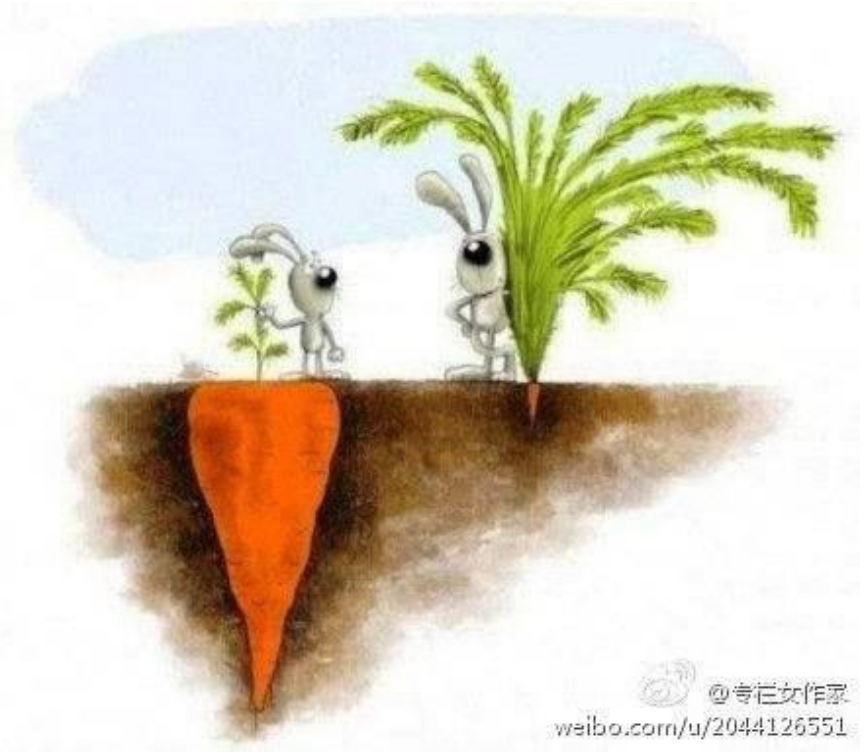
华夏卡帐户交易明细对帐单

:6226301510313668 持卡人姓名:刘晶兰 持卡人客户号:020000000000000991081 2011年03月10日

交易卡号	交易地点	交易日期	流水号	摘要	借贷	发生额
6032202001010210167	华夏银行广州分行珠江	20100916	229	代发工资	存入	3015.00
6032202001010210167	华夏银行广州分行珠江	20100916	389	代发工资	存入	200.00
6032202001010210167	华夏银行广州分行珠江	20100916	385143	金卡取款	支出	3000.00
6032202001010210167	华夏银行广州分行珠江	20100919	534	代发工资	存入	1500.00
6032202001010210167	华夏银行广州分行珠江	20100920	385386	金卡取款	支出	1500.00
6032202001010210167	华夏银行广州分行	20100921	102269	结息	存入	0.25
6032202001010210167	华夏银行广州分行	20100921	102269	扣税	支出	0.00
6032202001010210167	华夏银行广州分行珠江	20100924	385530	金卡取款	支出	200.00
6032202001010210167	华夏银行广州分行珠江	20101019	364	代发工资	存入	3015.00
6032202001010210167	华夏银行广州分行珠江	20101019	498	代发工资	存入	200.00
6032202001010210167	华夏银行广州分行珠江	20101020	387112	金卡消费	支出	162.00
6032202001010210167	华夏银行广州分行珠江	20101020	387116	金卡取款	支出	3000.00
6032202001010210167	华夏银行广州分行珠江	20101103	746765	卡存款	存入	3987.00
6032202001010210167	华夏银行广州分行珠江	20101108	388263	金卡取款	支出	2500.00
6032202001010210167	华夏银行广州分行珠江	20101109	259	代发工资	存入	3015.00
6032202001010210167	华夏银行广州分行珠江	20101109	374	代发工资	存入	200.00
6032202001010210167	华夏银行广州分行珠江	20101110	388356	金卡取款	支出	3000.00
6032202001010210167	华夏银行广州分行珠江	20101113	388511	金卡取款	支出	1800.00
6032202001010210167	华夏银行广州分行珠江	20101201	750068	卡存款	存入	3379.00

变量存储类型

- 从计算机存储角度看变量类型
- **数值类**
  - 整数型（Integer）：存储整型数
  - 实数型（Real）：存储小数
- **字符串类**
  - 字符串型（String）：存储字符串型数据
- **时间类**
  - 时间型（Time）：存储持续时间数据
  - 日期型（Date）：存储日期数据
  - 时间戳型（Time Stamp）：存储时间点数据



从数据分析视角看变量的类型 [重点]

视角1

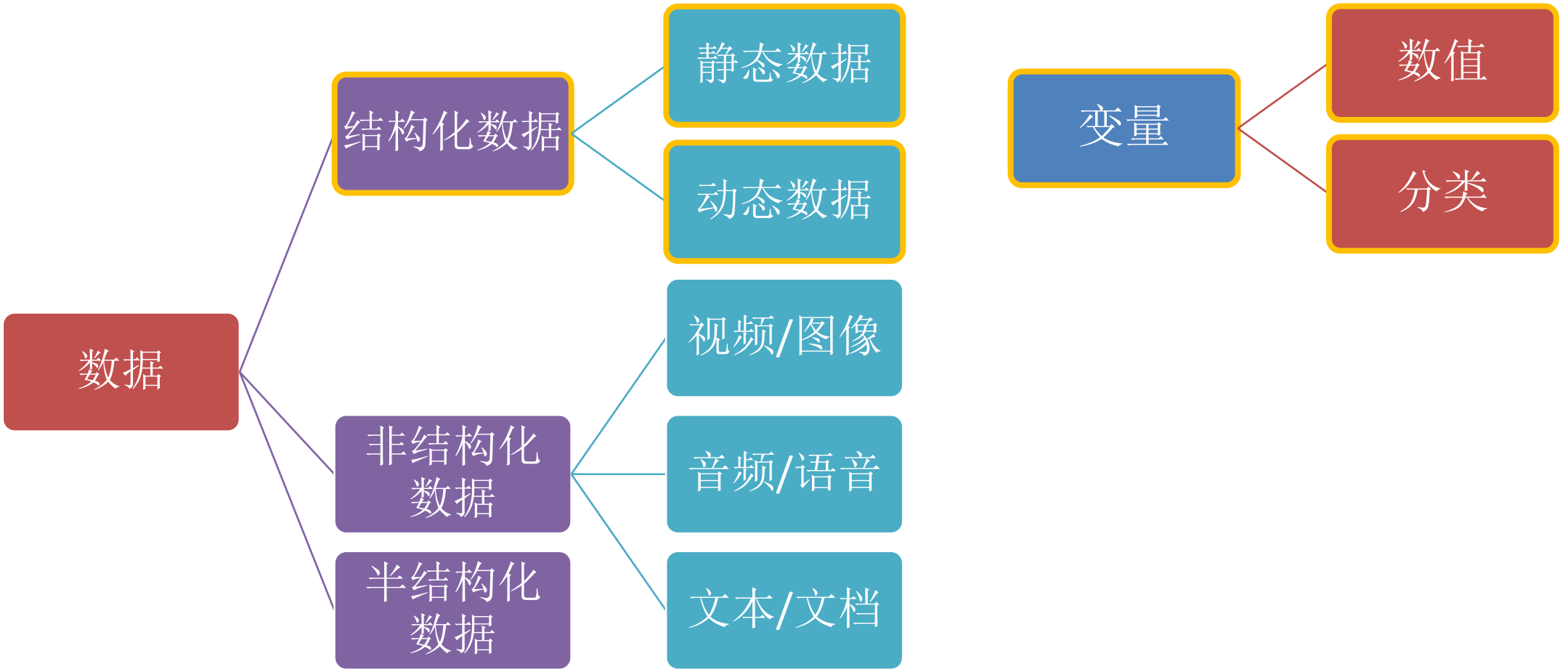
视角2

数据涵义 / 存储类别	数值（度量）		分类（维度）				
	不受限数值	受限数值	有序分类	无序二分类	无序有限分类	无序无限分类（ID 无类型）	日期时间
数值	银行余额	成绩 年龄 体重	客户等级	性别编码 是否会员	省份编码 民族编码	手机号码	日期时间
字符串	银行余额	成绩 年龄 体重	客户等级	性别 是否会员	省份 民族 城市	手机号码 身份证号码 姓名	日期时间
日期时间							日期时间

注：红字为建议的存储类型。

问题：哪个视角更适用于数据分析？

数据类型小结



大数据核心课程

商务大数据分析 >>2. 探索性数据分析

2.2 静态数据探索

刘跃文 博士

教授、博士生导师

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

V2.2, 2021-9-8

提纲：静态数据探索方法

- “单变量”数据探索：数据分布、统计指标、现状描述
- 两变量关系数据探索：关联关系、对比分析
- 三变量关系数据探索：分组关联关系对比、多层级对比分析、可视化交互

- **学习目标：分析数据的思路。**
- Tableau只是实现目标的工具和手段。

了解数据基本情况

- 待分析的数据为无分行分列的简单表格。
- 有分行或分列的数据需要进一步处理。

- 需要说明如下情况：
  - 从哪里获取的数据？
  - 数据反映了什么业务情况？
  - 需求方要实现什么目标？
  - 有多少条数据？（行数）
  - 数据有多少个字段？（列数）

品名	上月末库存		本月购进		本月原材料加权平	
	数量	金额	数量	金额	数量	单价
材料01	100.00	120.00	41.00	748.00	141.00	6.16
材料02	101.00	121.00	28.00	437.00	129.00	4.33
材料03	102.00	122.00	32.00	426.00	134.00	4.09
材料04	103.00	123.00	35.00	492.00	138.00	4.46
材料05	104.00	124.00	-	-	104.00	1.19
材料06	105.00	125.00	-	-	105.00	1.19
材料07	106.00	126.00	-	-	106.00	1.19
材料08	107.00	127.00	-	-	107.00	1.19
材料09	108.00	128.00	17.00	306.00	125.00	3.47
材料10	109.00	129.00	-	-	109.00	1.18
材料11	110.00	130.00	-	-	110.00	1.18
材料12	111.00	131.00	-	-	111.00	1.18
材料13	112.00	132.00	-	-	112.00	1.18
材料14	113.00	133.00	-	-	113.00	1.18
材料15	114.00	134.00	-	-	114.00	1.18
材料16	115.00	135.00	-	-	115.00	1.17
合计	1,720.00	2,040.00	153.00	2,409.00	1,873.00	2.38

	序号	事 项	责任人	初始计划完成时间		是否完成	未完成原因	解决方法（若无解决方法请将问题复制到例会议题里）	更改计划完成时间		备注
				开始时间	完成时间				开始时间	完成时间	
本周工作总结	1										
	2										
	3										
	4										
	5										
	1										
	2										
	3										
	4										
	5										

1. 单变量数据探索

数据分布、现状描述

“单变量”数据探索

- 不存在真正的“单变量”数据分析
- 隐藏的变量：**记录数**
- “单变量”数据探索事实上是：“**变量-记录数**”探索

- 描述性统计
- 分类型变量的探索
- 数值型变量的探索

ID	日期	货物	金额	记录数
2	5-17	大衣	300	1
2	5-18	裙子	200	1
4	3-3	皮衣	500	1
5	4-15	裤子	100	1

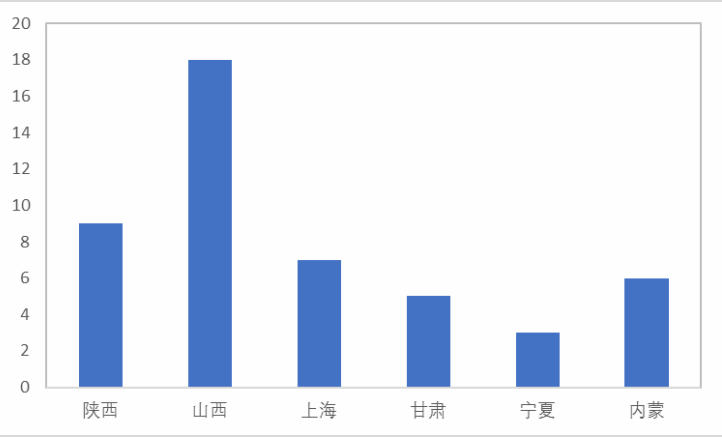
- 关于Tableau的教程，参考
- https://onlinehelp.tableau.com/current/pro/desktop/zh-cn/default.htm

1. 分类型变量的探索

- **分析目标：研究数据在分类上的分布情况。**
- **二分类型、少数分类型变量：研究数据在分类上的分布情况**
- 图形的选择：
  - 柱图（分类数量较少/分类标签文本较短）
  - 条形图（分类数量较多/分类标签文本较长）
- 图形的排序：
  - 按照数量大小（一般情况）
  - 按照分类顺序（有序分类等）

慎用折线图

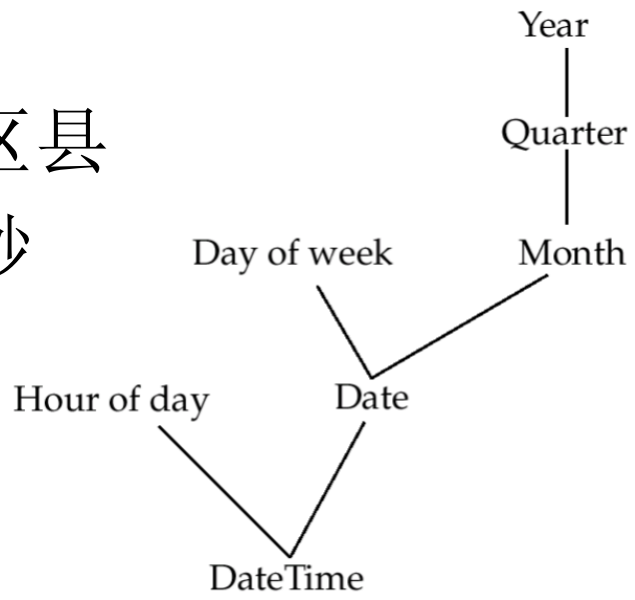
- 数据一般都是离散的，当把离散的数据连起来时，要特别小心。
- 分类变量不可用折线图，折线图反映的是趋势，避免造成误会。



层次型分类变量

- **例如：**
- 学校→学院→系→班→组
- 大洲→国家→地区→省→市→区县
- 年→季度→月→日→时→分→秒

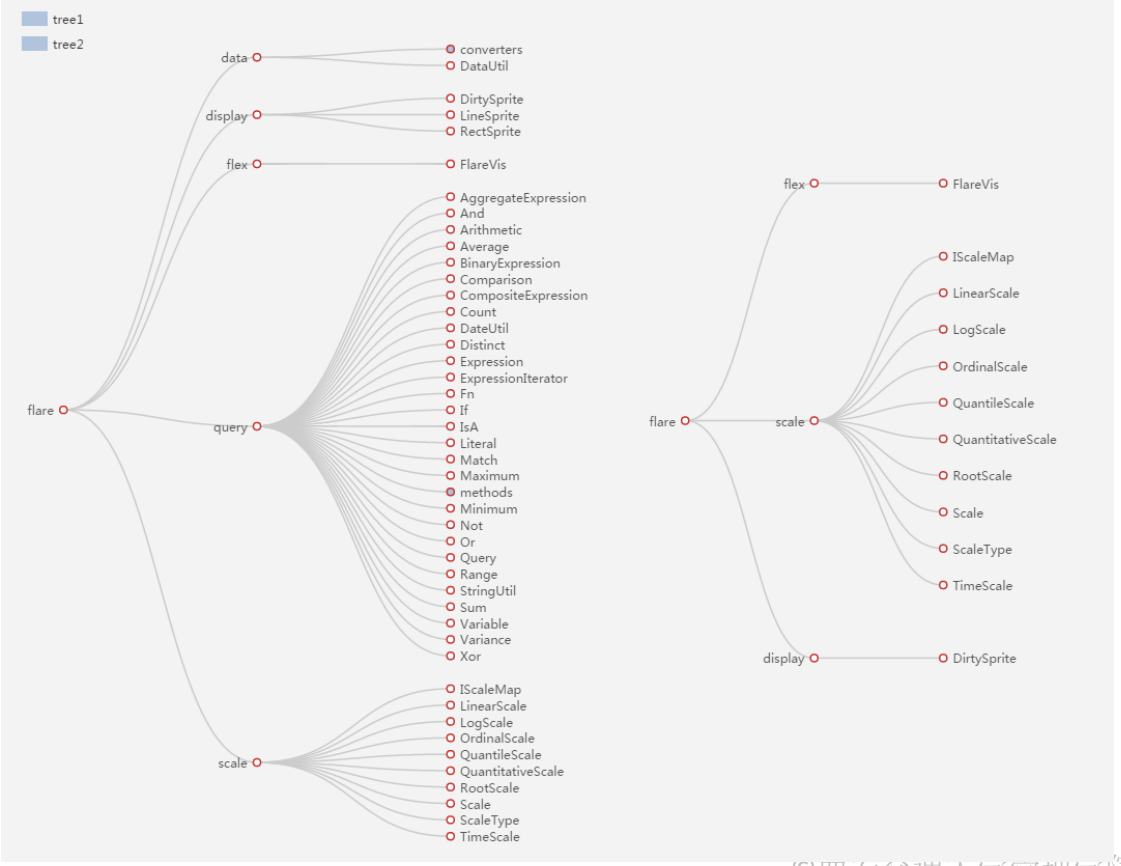
- 图形选择：
  - 分面板展示图形
  - 树形图/树状图
  - 可交互图表

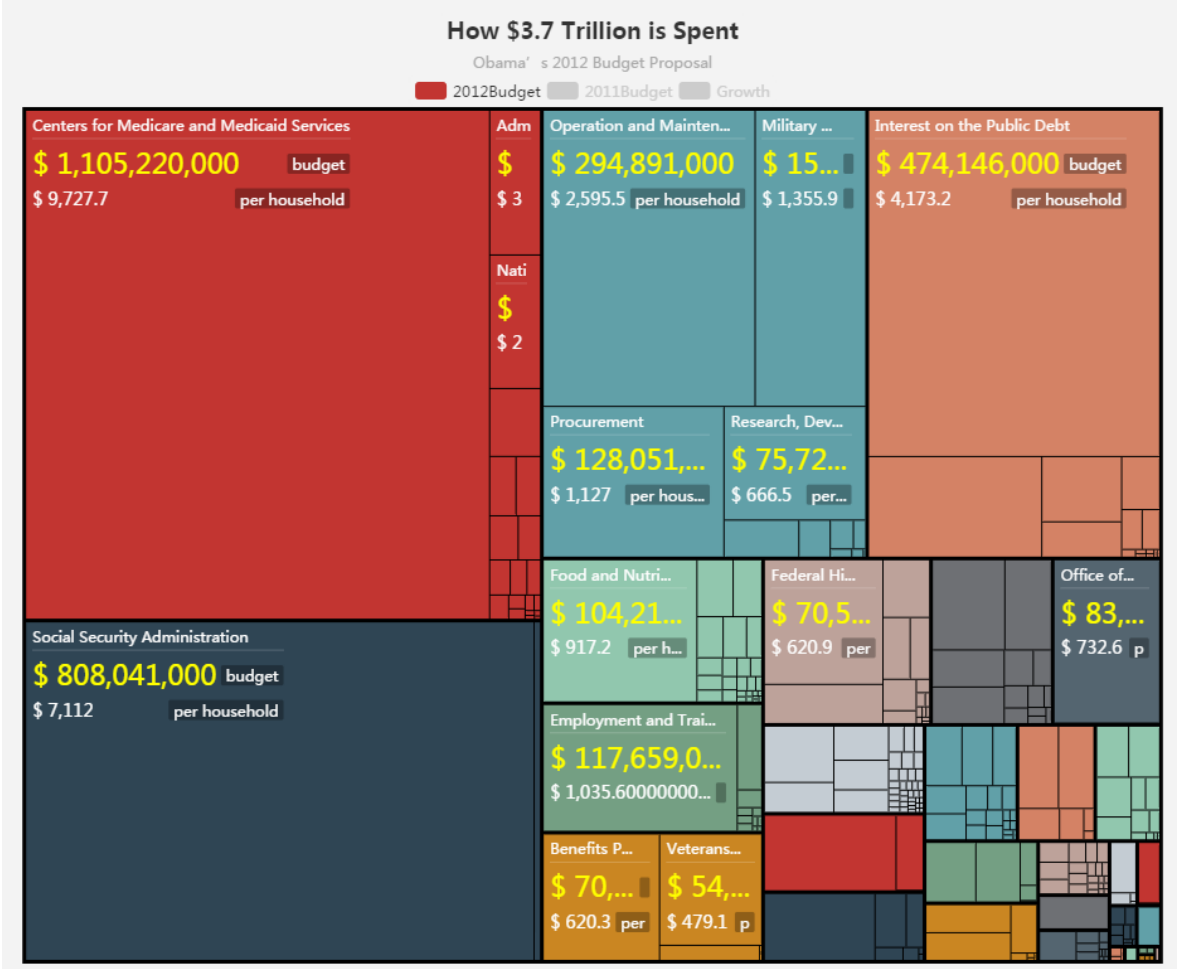


a) Time Hierarchy

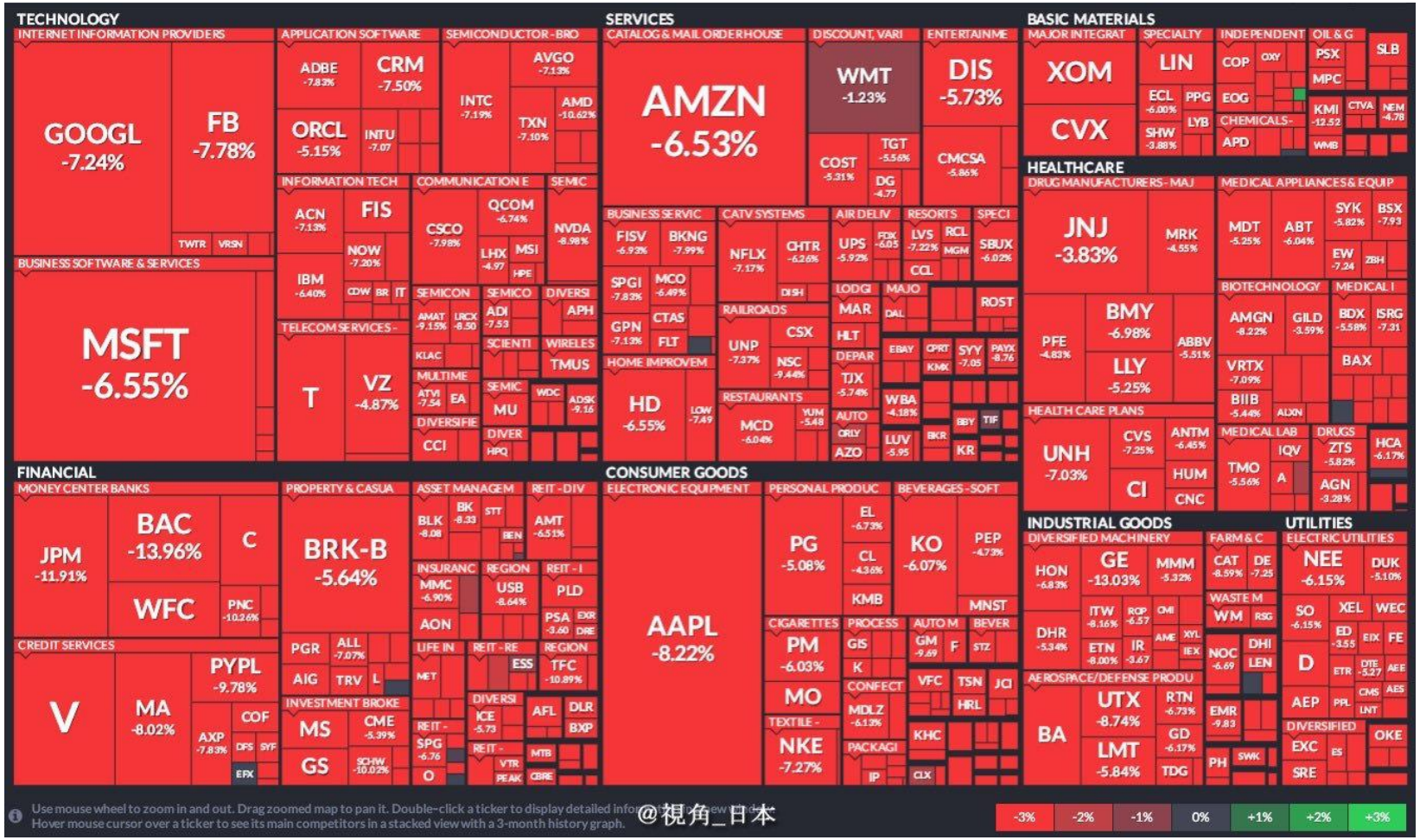
b) Location Hierarchy

- 树形图





例如：
大盘指数的
构成及表现



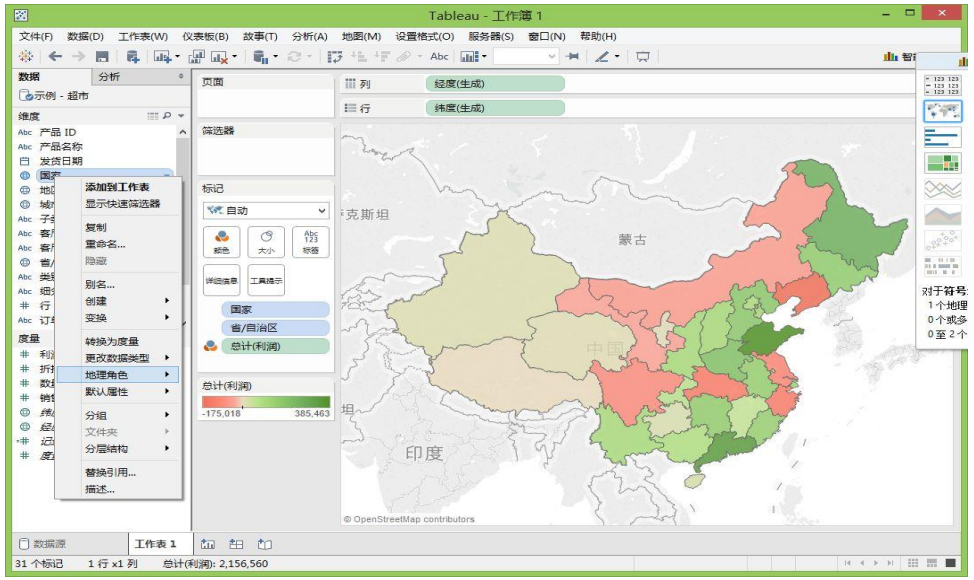
- 交互方式（BI）：
  - 上卷 Roll up：从细粒度到粗粒度
  - 钻取 Drill down：从粗粒度到细粒度

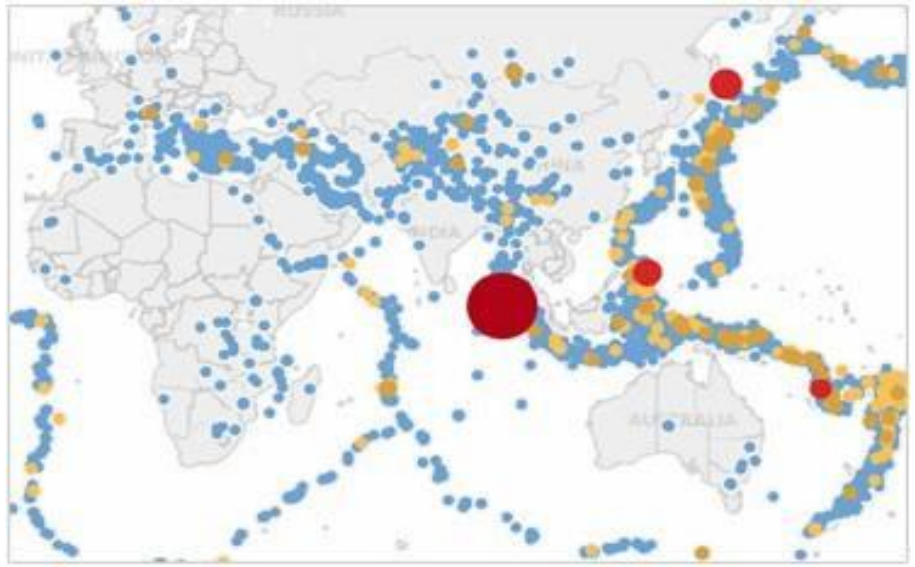
clothes_size: **all**

category	*item_name*	*color*				
		dark	pastel	white	total	
womenswear	skirt	8	8	10	53	
	dress	20	20	5	35	
	subtotal	28	28	15		88
menswear	pants	14	14	28	49	
	shirt	20	20	5	27	
	subtotal	34	34	33		76
total		62	62	48		164

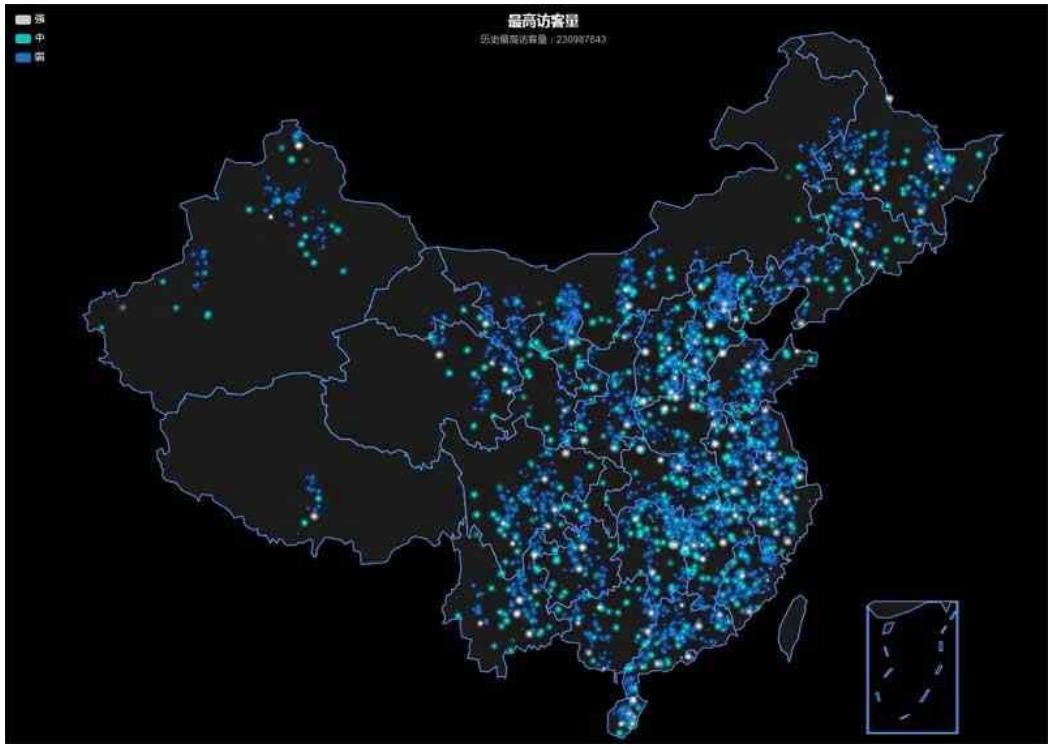
地理位置类型的分类变量

- **研究数据的空间分布**
- 填色地图（区域分布、地理位置数据能拼成完整地图）
- 描点地图（轨迹分布、地理位置数据不能拼成完整地图）



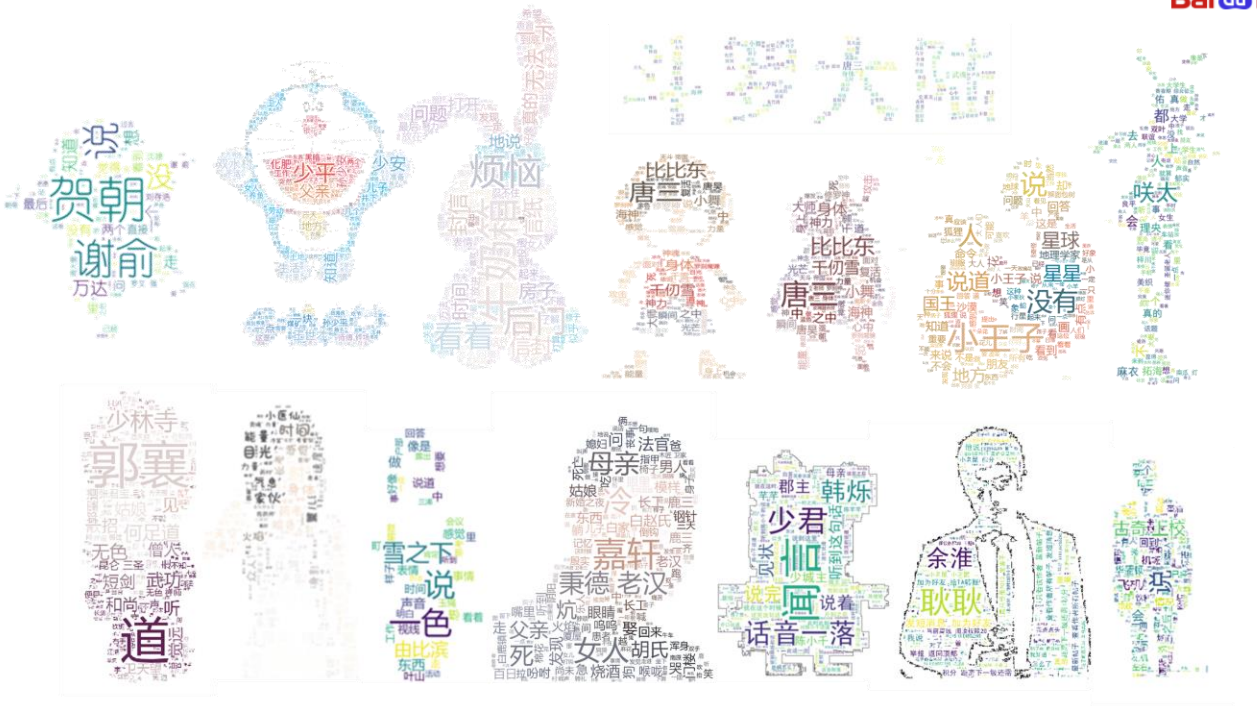


- 一些更多的地图可视化分析



文字型的分类变量

- 词云：词语的面积反映了其频次/统计量。
- 词云工具：Python/R/互联网工具



分类数量特别多怎么办？

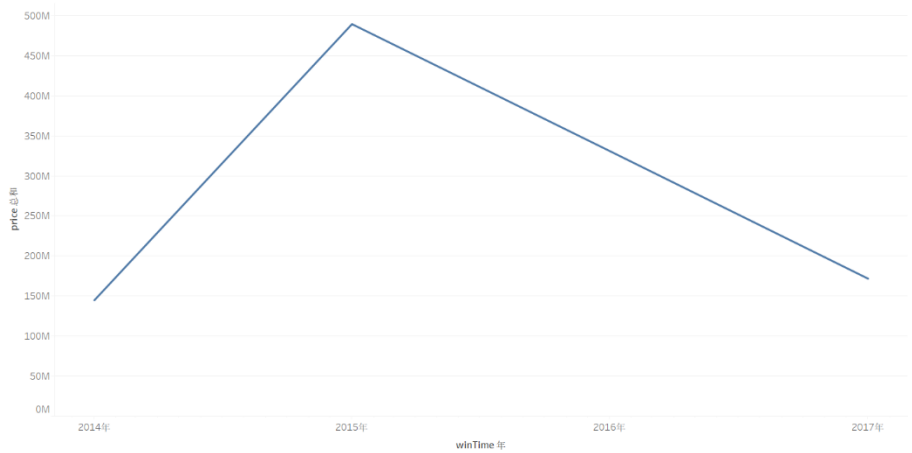
- **多数分类型、无限分类型变量**
- 方法1：
  - 只取前若干名，转为有限分类
  - 排除/筛选器
- 方法2：
  - **在有限的面积上展示较多的分类数量**
  - 使用饼图
  - 使用树状图
  - 使用“词云”等新图形

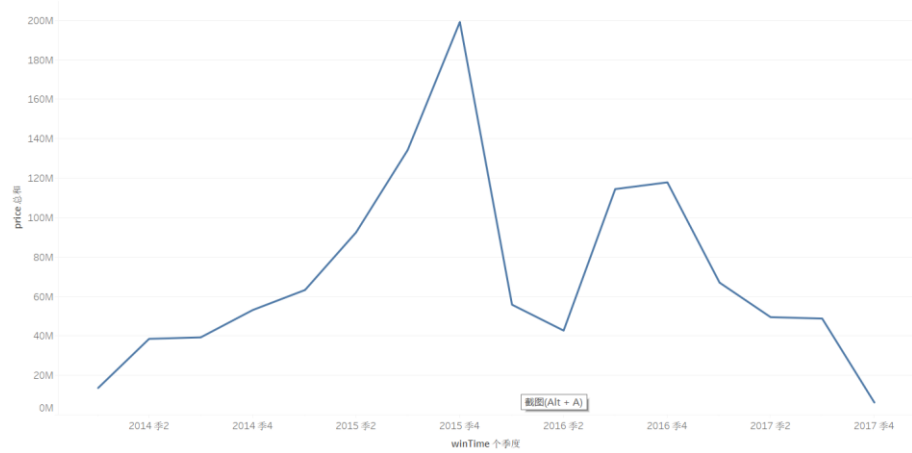
日期时间分类变量

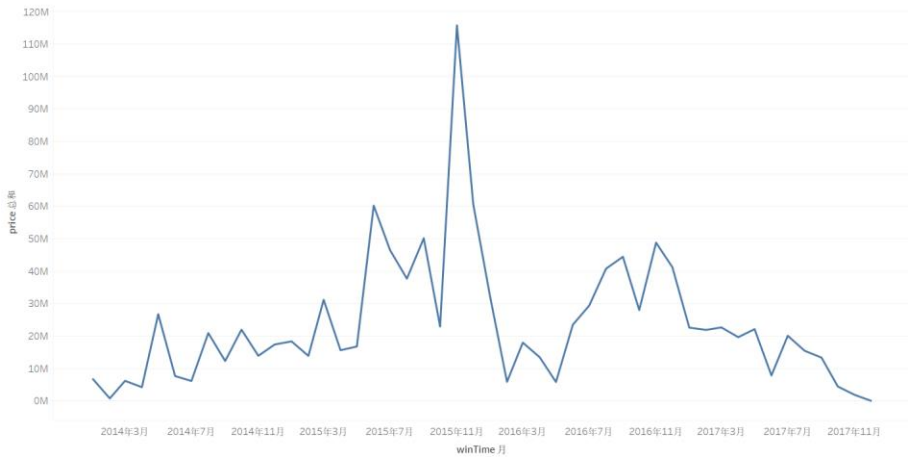
- 日期、时间是有序分类变量

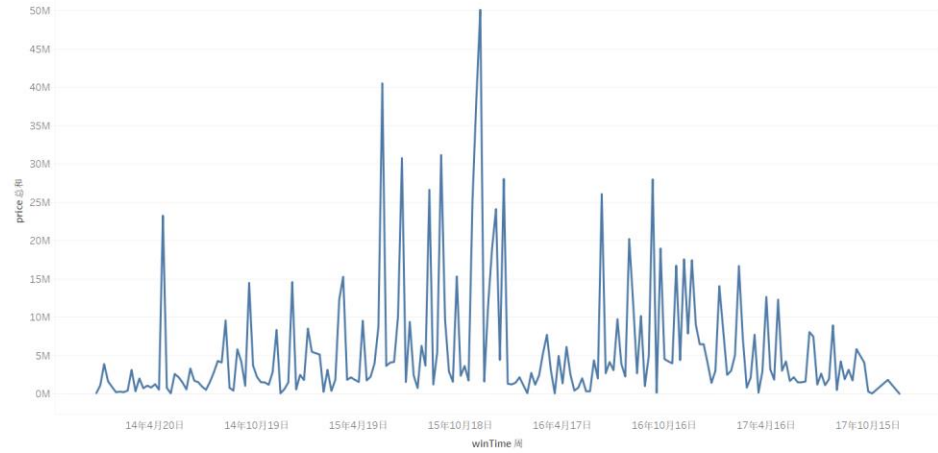
- 拆分方法一：完整日期/时间（具体时间点），用于研究发展趋势（折线图/柱状图）
  - 年，年季度，年月，年月日，年月日小时，年月日小时分

- 拆分方法二：某一日期/时间元素，用于研究周期性（折线图/柱状图）
  - 每年的周期性：季度，月，月日，周数，周数日
  - 每月的周期性：日（月初/月末效应）
  - 每周的周期性：星期几
  - 每天的周期性：小时，小时分钟，小时分钟秒



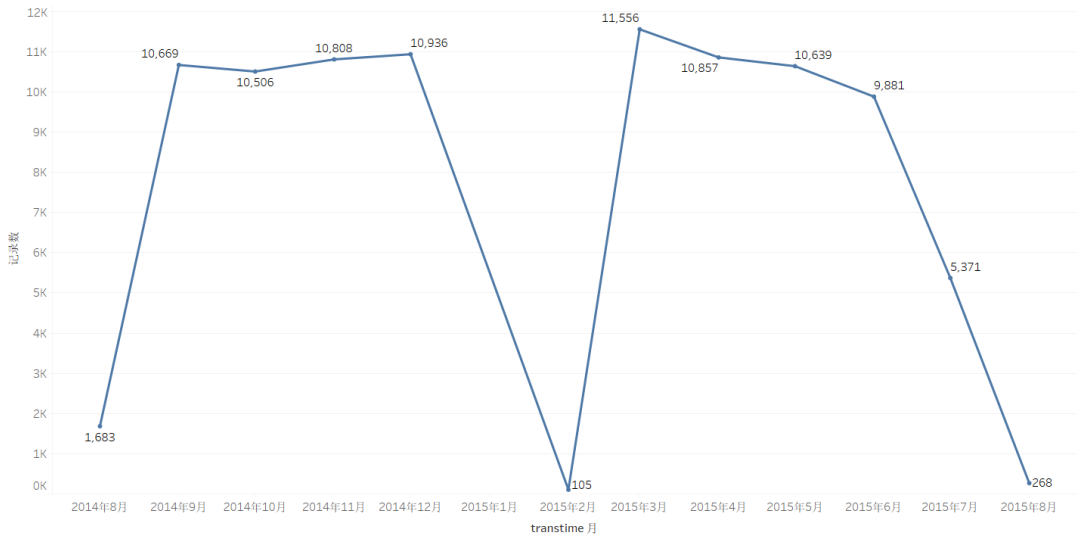


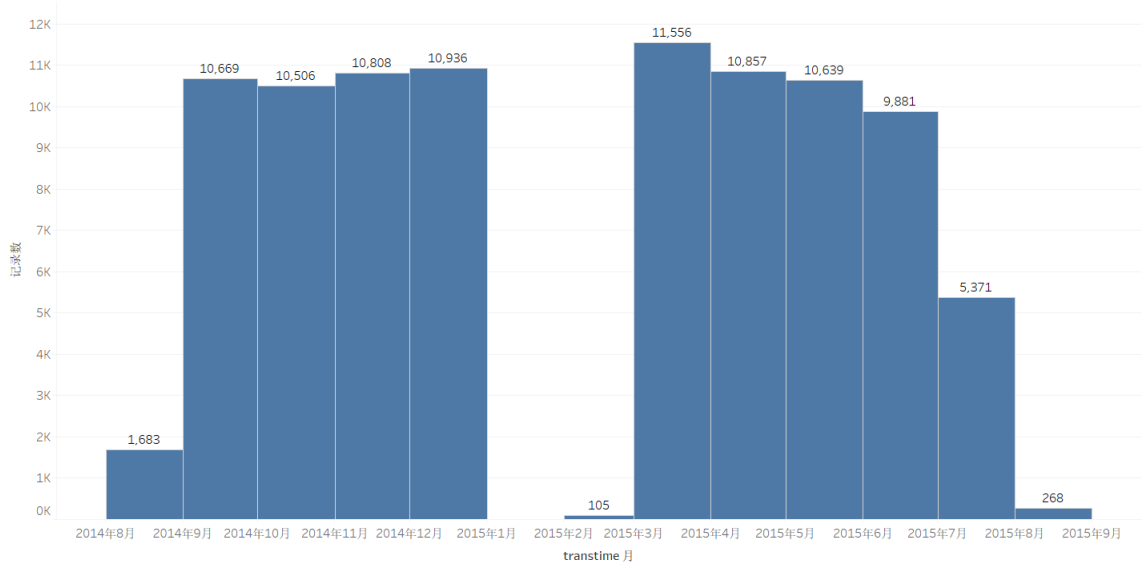




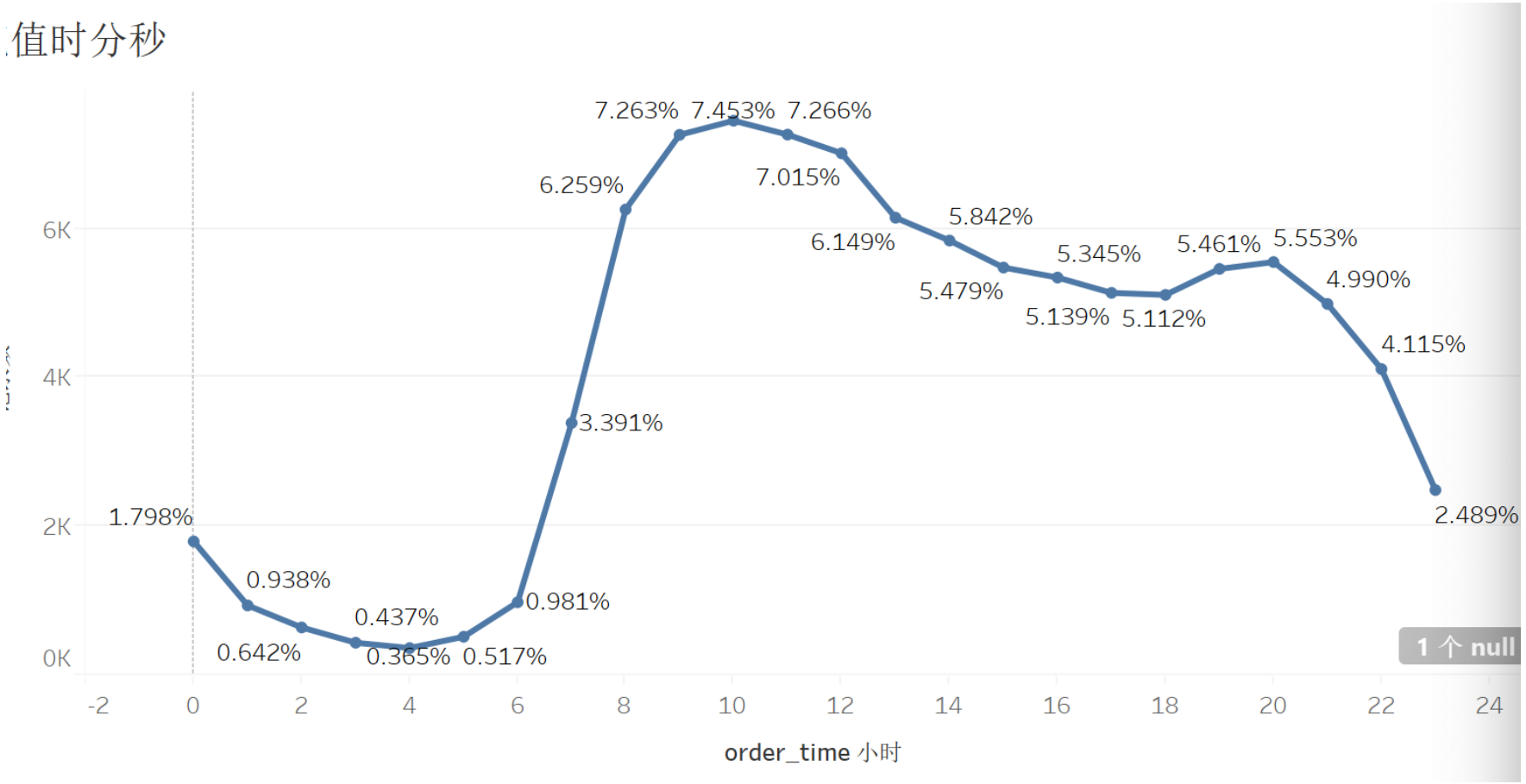
慎用折线图

- 时间变量如中间断开，避免直接用折线图，以免造成错觉。
- 使用折线图时，确保日期是连续的（或为没有数据的日期补0）。

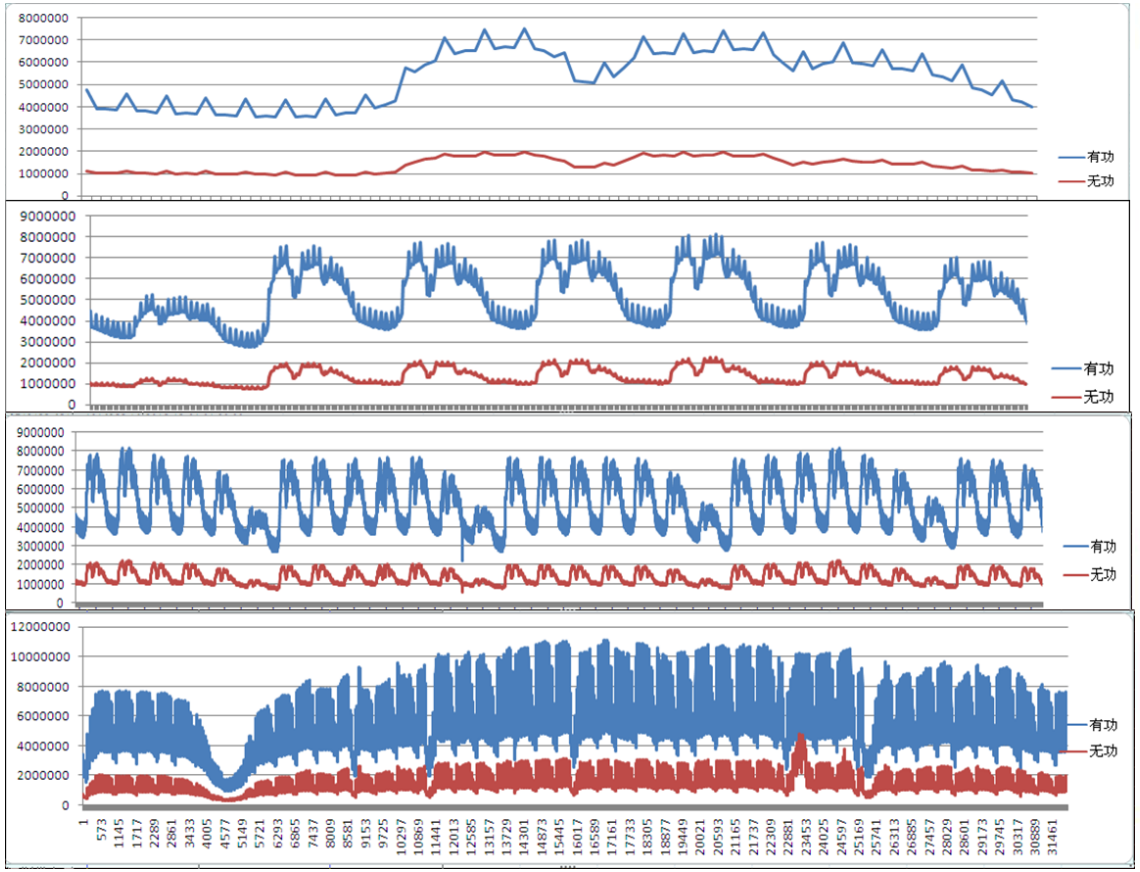


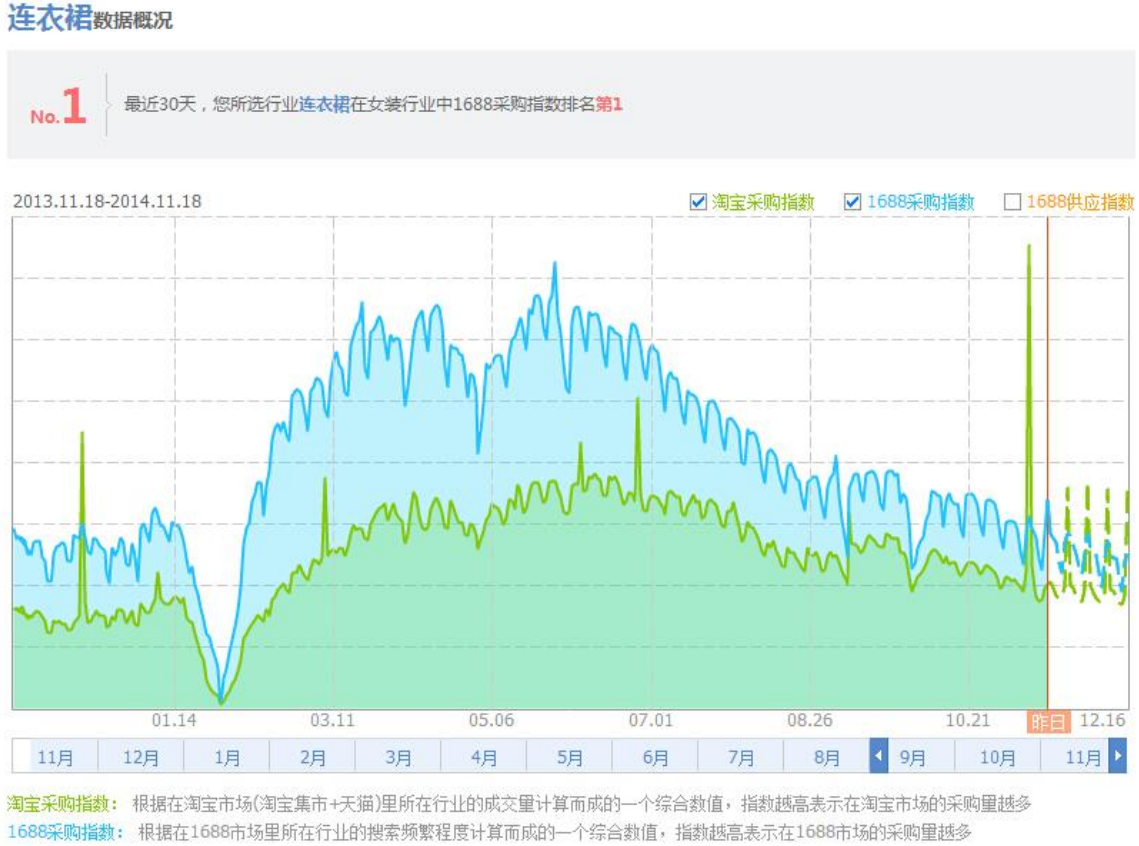


每日充值规律图



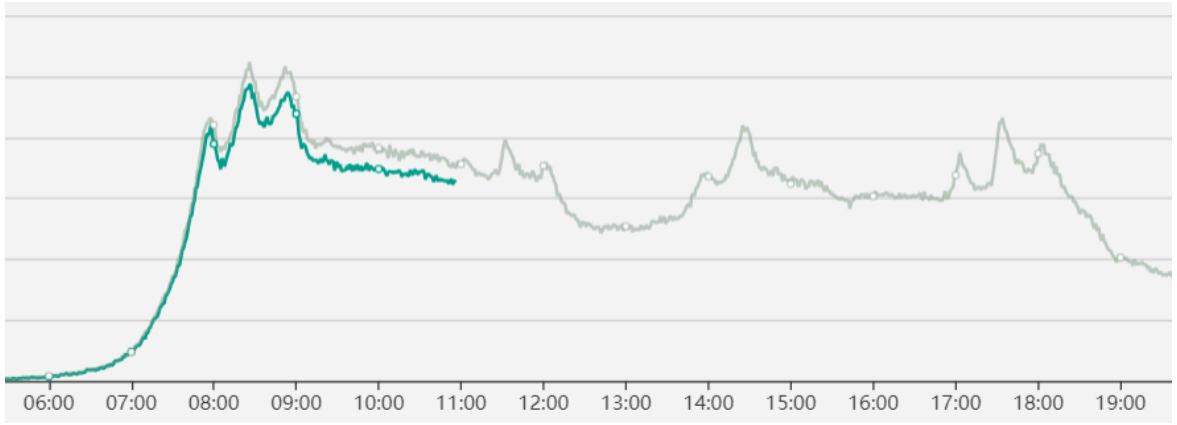
多种业务的时间规律图

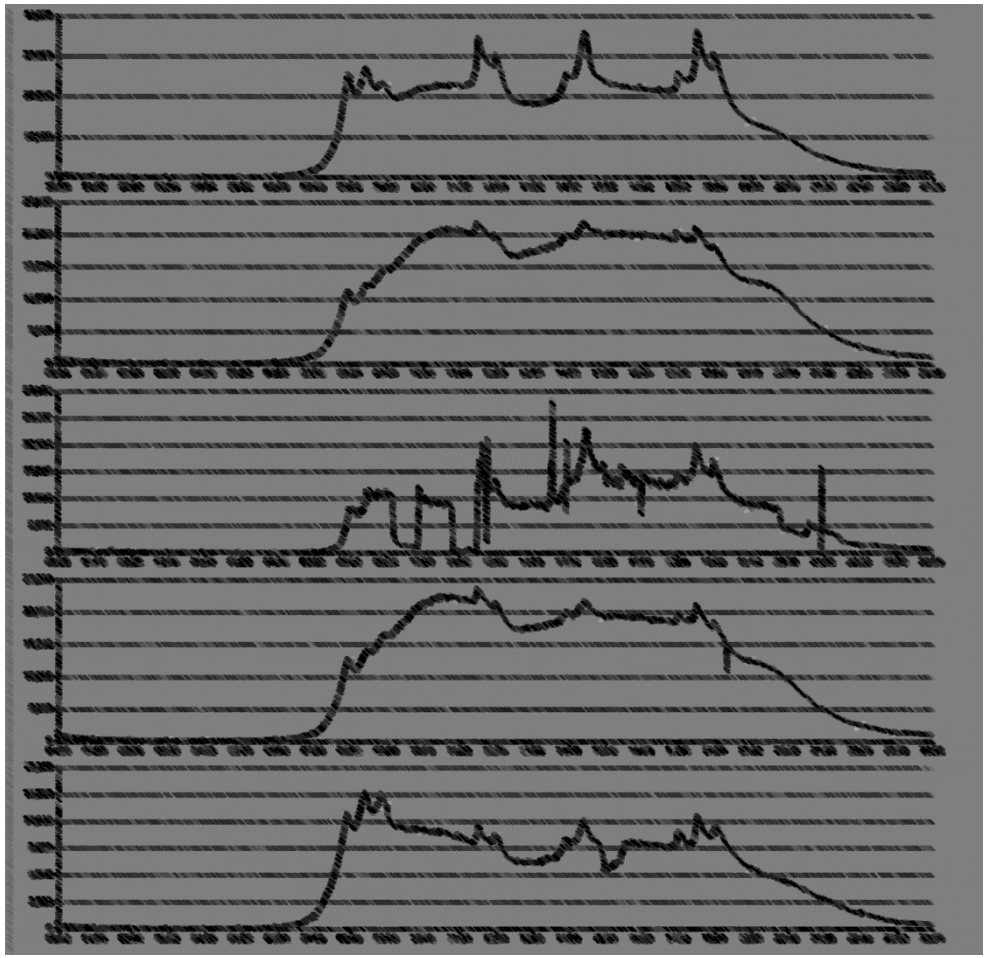




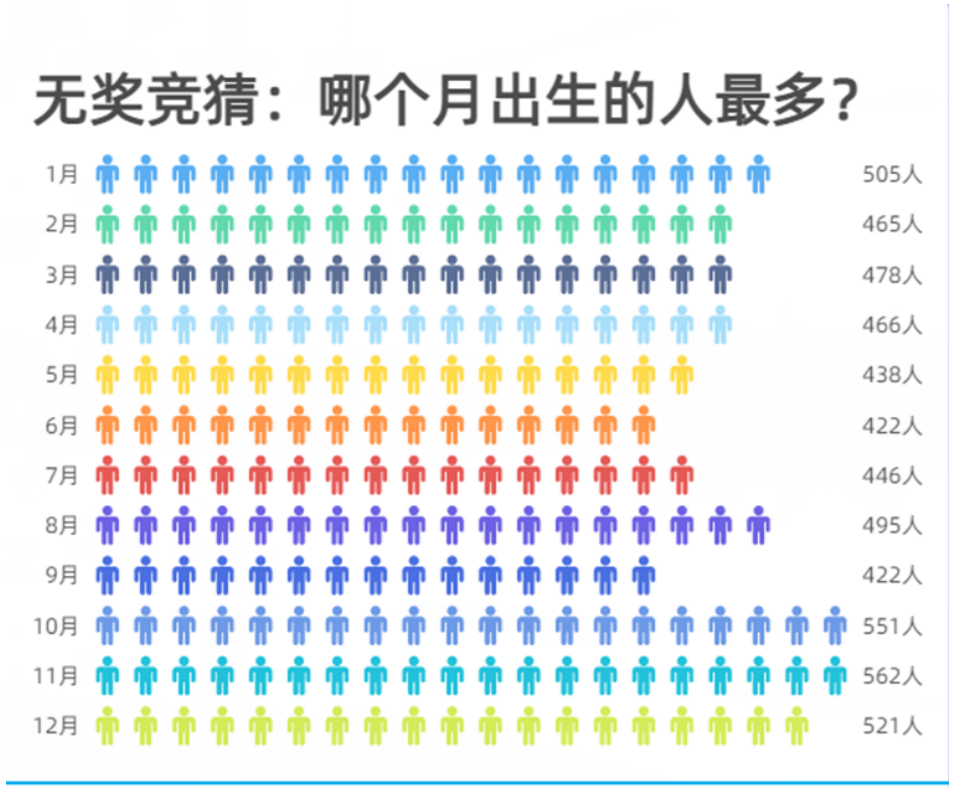
利用周期性规律判断数据异常

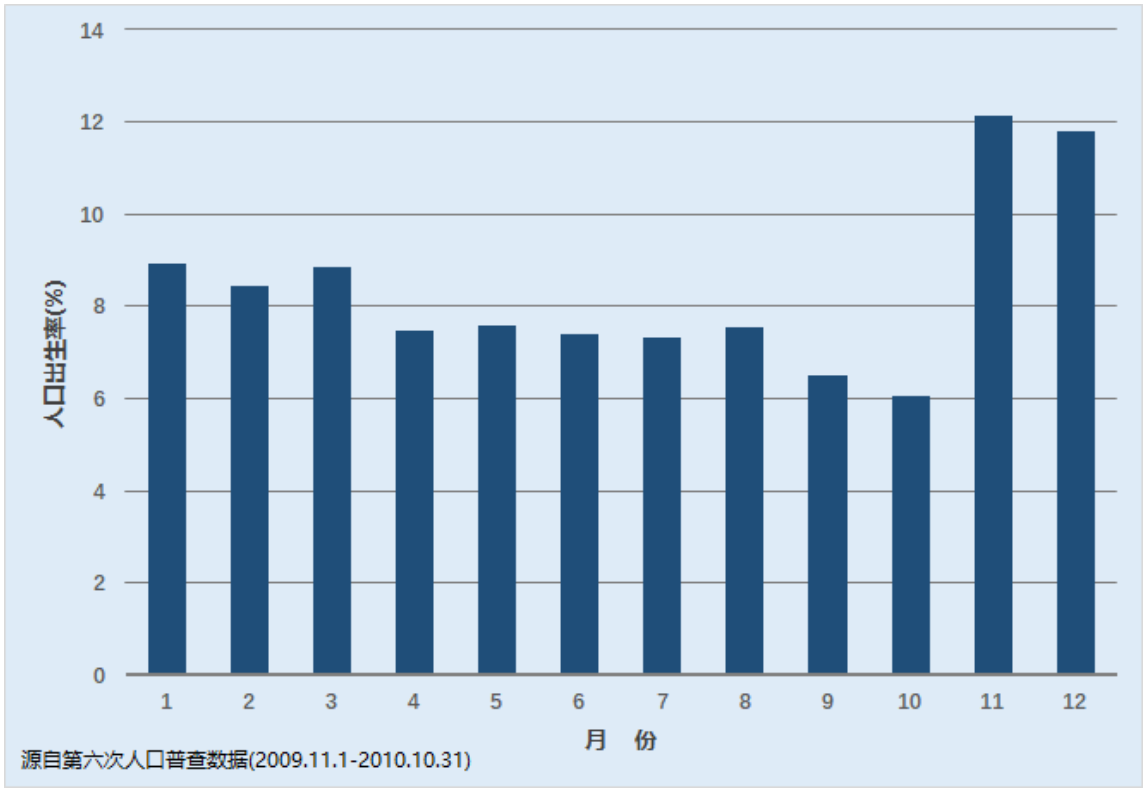
- 利用时间分布曲线识别异常。
- 正常而言，每天的规律都可以参考历史分布曲线。





从周期中识别规律





2. 数值型变量的探索-描述性统计

- 分析目标：数值型变量的均值、范围等描述性统计与分布研究。

数据涵义 / 存储类别	数值（度量）		分类（维度）				日期时间（维度）
	不受限数值	受限数值	有序分类	无序二分类	无序有限分类	无序无限分类（ID 无类型）	
数值	银行余额	成绩 年龄 体重	客户等级	性别编码 是否会员	省份编码 民族编码	手机号码	日期时间
字符串	银行余额	成绩 年龄 体重	客户等级	性别 是否会员	省份 民族 城市	手机号码 身份证号码 姓名	日期时间
日期时间							日期时间
处理方式							

数据涵义 / 存储类别	数值（度量）		分类（维度）				日期时间（维度）
	不受限数值	受限数值	有序分类	无序二分类	无序有限分类	无序无限分类（ID 无类型）	
数值	银行余额	成绩 年龄 体重	客户等级	性别编码 是否会员	省份编码 民族编码	手机号码	日期时间
处理方式	计数 均值 求和 最大 最小 方差	计数 均值 求和 最大 最小 方差	计数 分类计数 最大 最小	计数 分类计数	计数 主要分类计数	计数	计数

描述性统计 Descriptive Statistics

- 只能对度量型变量（注意不是数值型）进行描述性统计。

Table 1
Descriptive characteristics of variables.

Variable	*Description*	*Mean*	*SD.*	*Min.*	*Max.*
Age	*Mobile phone users' age, classified into five groups: (17-25)*, (26-35), (36-50), (51-60) or (61-)*	42.80	13.80	18.00	129.00
Gender	*Mobile phone users' gender: male or female**	–	–	–	–
AQI	*The daily Air Quality Index, classified into four groups: good*, light, moderate, or heavy*	144.54	63.22	62.00	355.00
Highest temperature	*The daily highest temperature (℃)*	11.00	4.10	3.00	22.00
Weather	*The weather of a day: sunny*, cloudy or rainy*	–	–	–	–
Day of the week	*The day of the week: Monday*, Tuesday, Wednesday, Thursday, Friday, Saturday, or Sunday*	–	–	–	–
RG	*The radius of gyration, reflecting the travel distance (kilometer)*	1.94	2.86	0.00	80.66
NP	*The number of visited places, reflecting the travel area*	27.49	37.55	1.00	1264.00

* This category is the benchmark group in the regression models.

Yuquan Xu, **Yuewen Liu***, Xiangyu Chang, Wei Huang, (2021) How does air pollution affect travel behavior? A big data field study. Transportation Research Part D: Transport and Environment, 99, 103007.

2015-06-09 第3185期

22025 人跟帖

平均工资反映不了普遍收入

导语 日前，北京公布2014年全市职工平均工资，为77560元，月平均工资为6463元。不出意料，又有大批网友吐槽“被平均”、“拖后腿”。正如很多人所说，拿平均工资来反映社会普遍收入状况是断然不够的，年复一年地只让人们看平均工资是一种忽悠。不仅如此，由于我国社保基数、住房公积金缴交基数都由平均工资决定，这实质上造成了让低收入者更贫困、以及劫贫济富的效果，这种状况理所应当需要改变。

平均工资反映不了普遍收入状况，起码应该公布中位数

“张家有财一千万，九个邻居穷光蛋，平均起来算一算，个个都是张百万”

每当统计局公布平均工资的时候，这首打油诗都注定会被网友们反复提起，而今年的“统计局数据”又比往年来得令人乍舌：2014年全国平均工资4.99万，月均四千元出头，这也就罢了；首都北京的平均工资达到了77560元，月平均工资为6463元，这就很让人艳羡了；尤其是，全市城镇非私营单位就业人员年平均工资为102268元，月均达8522元——对于大部分网友而言，这岂止是“拖后腿”，简直是拍马也追不上，“这得怎么个平均法才能把我的收入平均到这么高啊？”

Jeff Bezos made 1.2 million times the median Amazon employee in 2017

学院
EMENT
ERSITY

均值其实不一定靠谱！
谨慎使用平均值。

[Photo: Public Domain Pictures/Pexels]

BY AINSLEY HARRIS 2 MINUTE READ

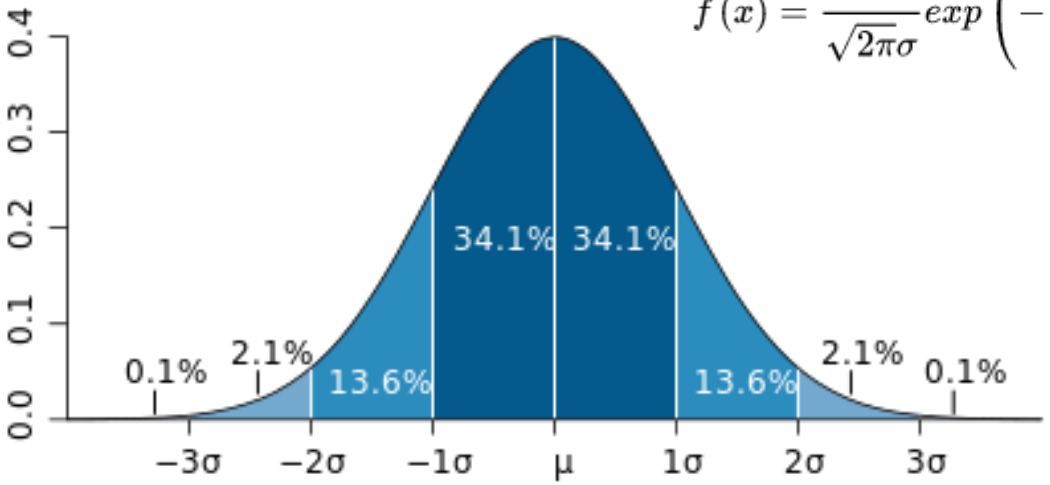
Last year, Amazon CEO Jeff Bezos's net worth ballooned by $35.1 billion. Also last year, the median Amazon employee made $28,446. Do the math, and Bezos made 1.2 million times his typical employee in 2017.

That's not the ratio Amazon reports in its annual proxy filing, released yesterday. There, the company calculates median pay vs. CEO compensation using Bezos's salary—a comparatively modest $1,681,840. Based on that figure, Bezos made just 59 times his median employee.

（1）非正态分布不能用平均值

- 正态分布 Normal Distribution
- $\frac{\text{最大值}}{\text{最小值}}$ 这个比值是有限的
- 例如：身高、体重、跑步速度等的分布

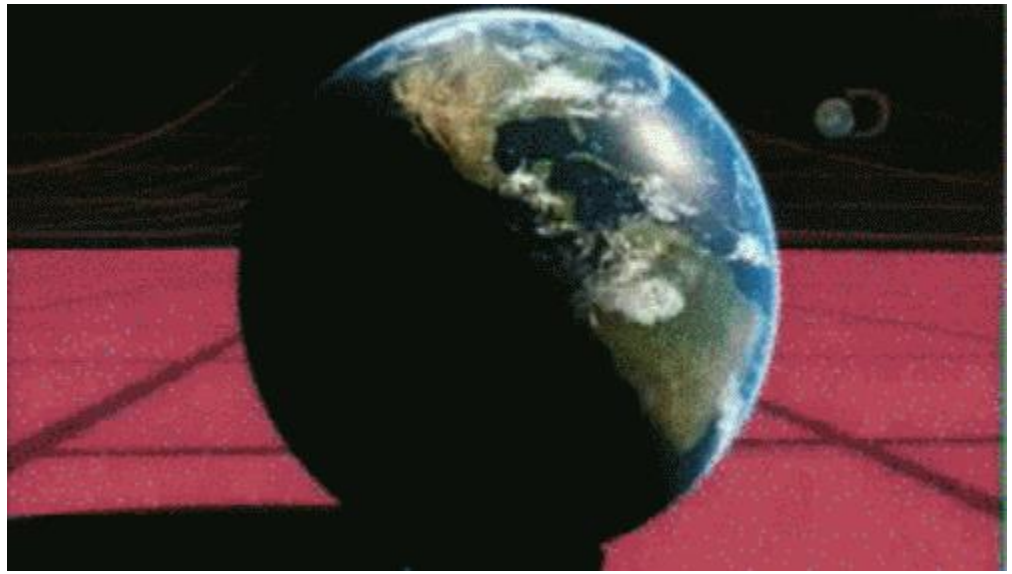
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$



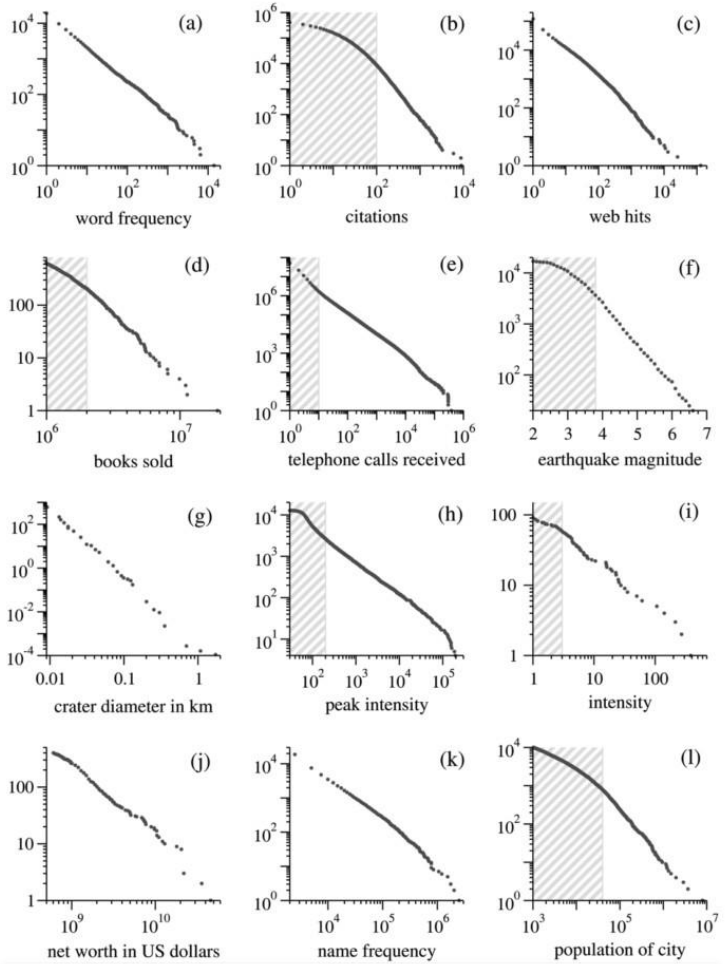
幂律分布

- 幂律分布 类似于
  - 马太效应
  - 长尾理论
  - 二八法则
  - 齐夫法则 (Zipf)
  - 帕累托分布
  - 基尼系数

$$f(x) = cx^{-\alpha-1}, x \to \infty$$



地球 太阳 海山二 参宿四 大犬座



https://www.bilibili.com/video/BV1VW411Z7ps 财富是如何分布在美国的 | 数据可视化

https://www.bilibili.com/video/BV1n4411h7my 世界财富分配*

全球财富分布

@轻松学图表 汉化作品

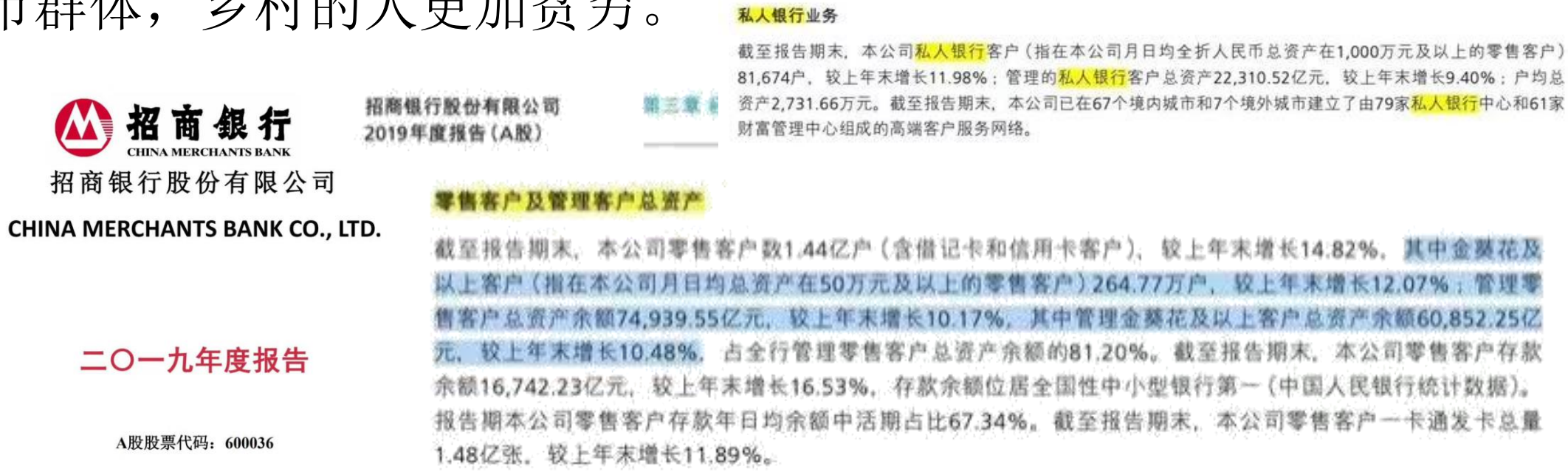
翻译：@油杀臭干

字幕：@油杀臭干

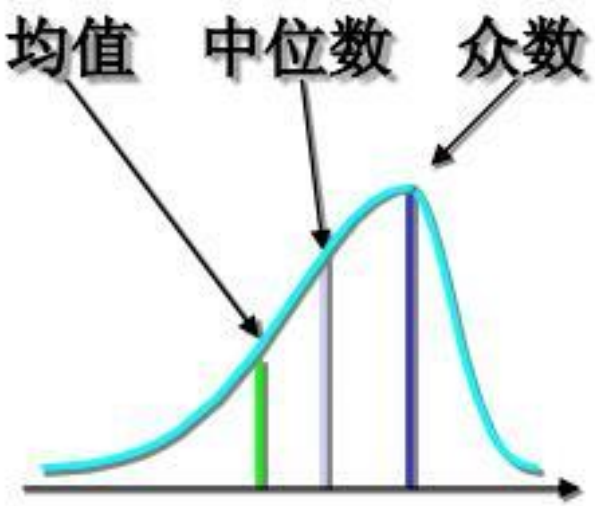
原视频：http://www.therules.org

招行2019年报

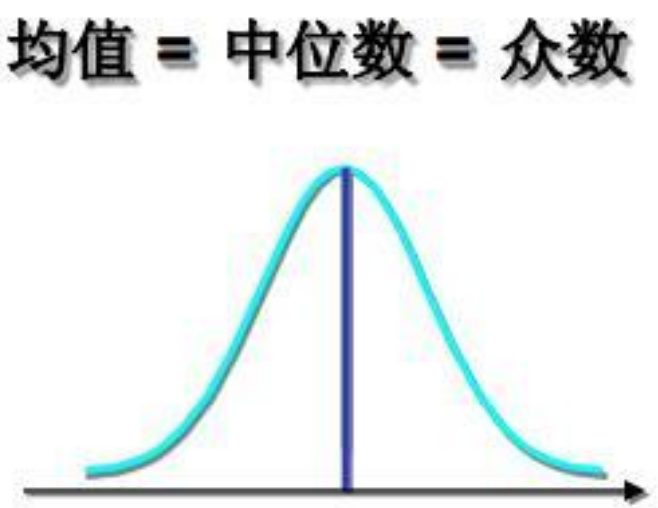
- 根据招行2019年报，1.84%的储户拥有81.2%的存款；而0.056%的储户拥有29.8%的存款。真实的财富分布可能更加极端，因为这只是存款，富有的人更有可能拥有其它资产；招行反映的只是城市群体，乡村的人更加贫穷。



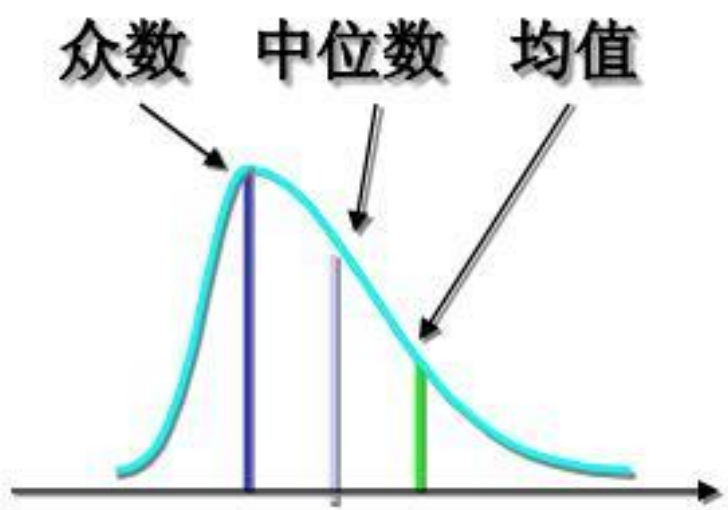
解决方案：中位数、众数



左偏分布

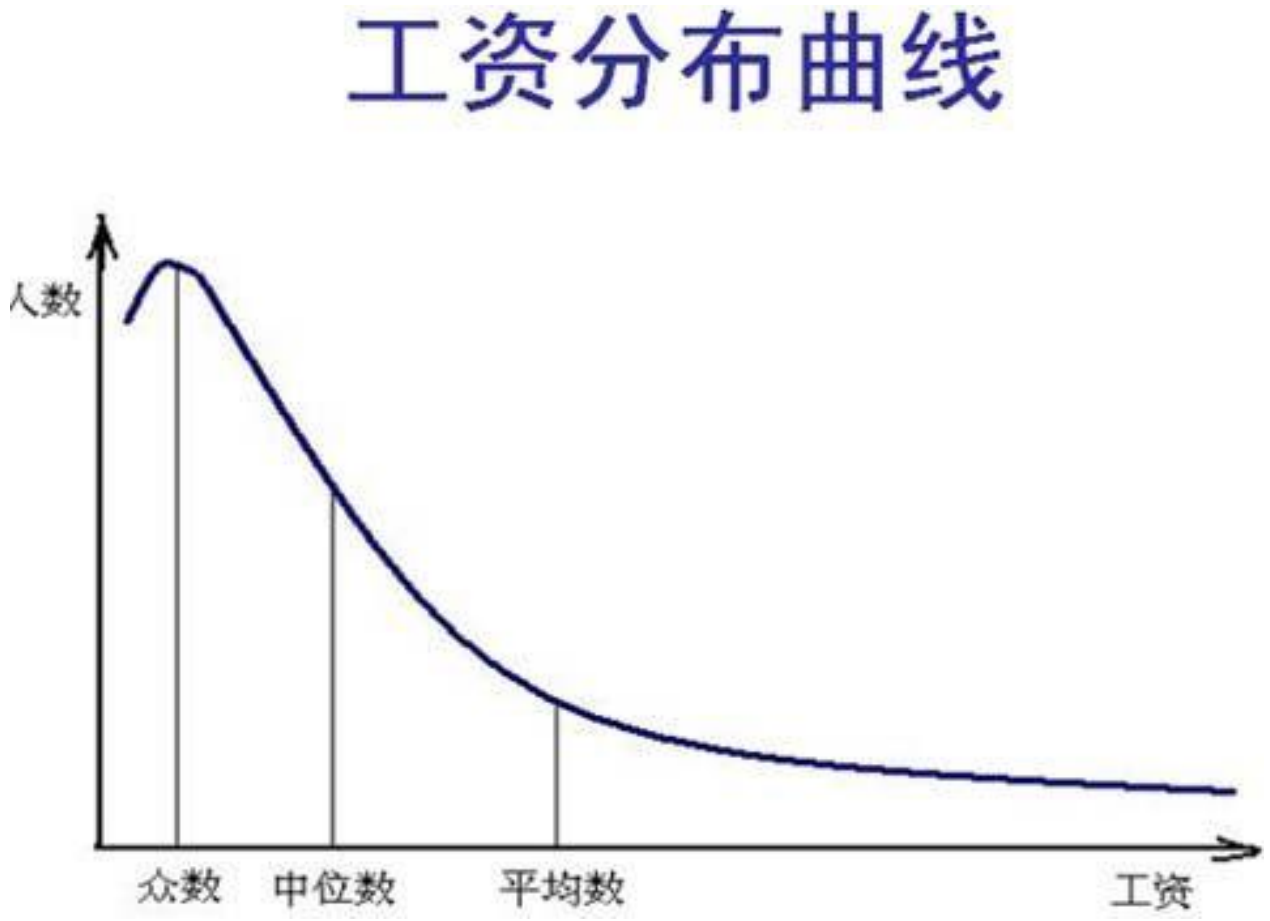


对称分布



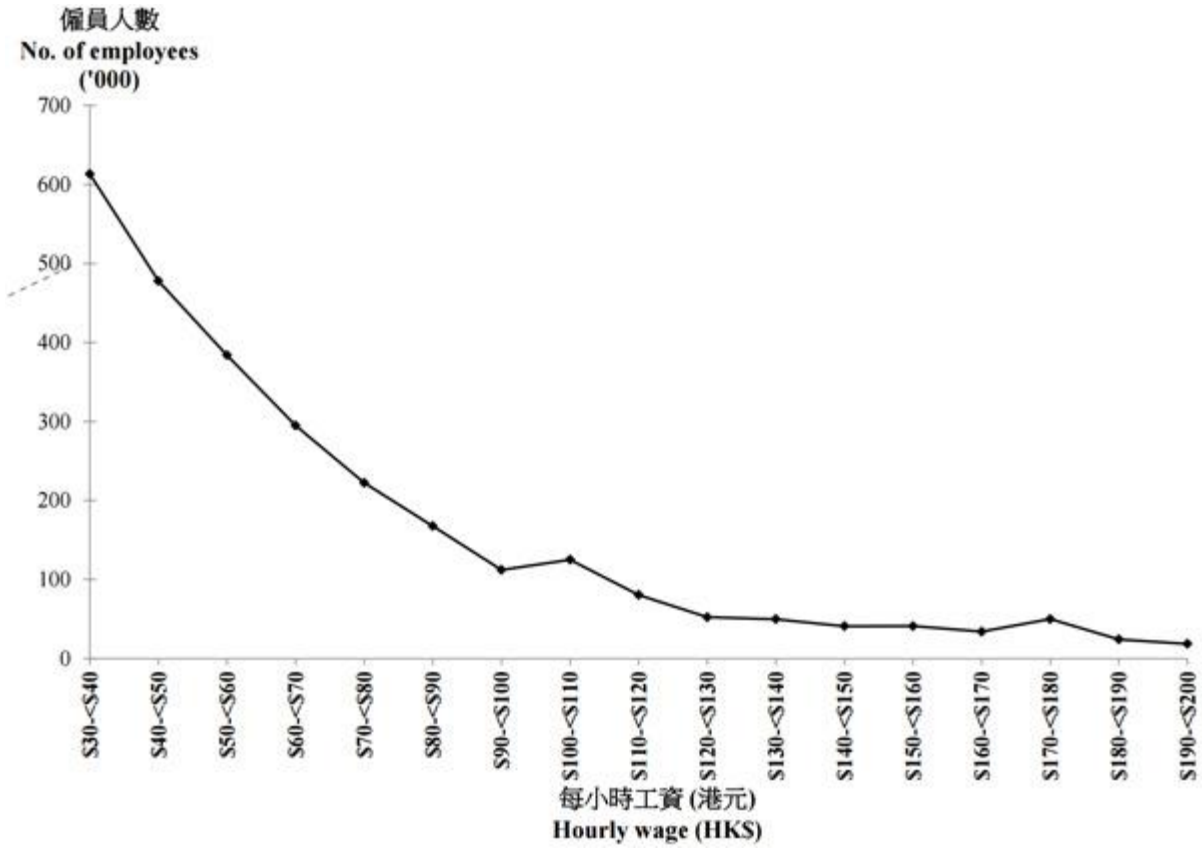
右偏分布

例如：收入/财富分布的非正态性



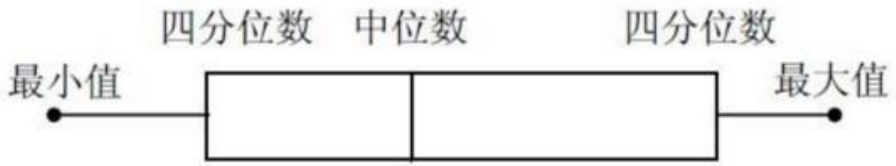
香港“收入及工时按年统计调查”中的收入分布

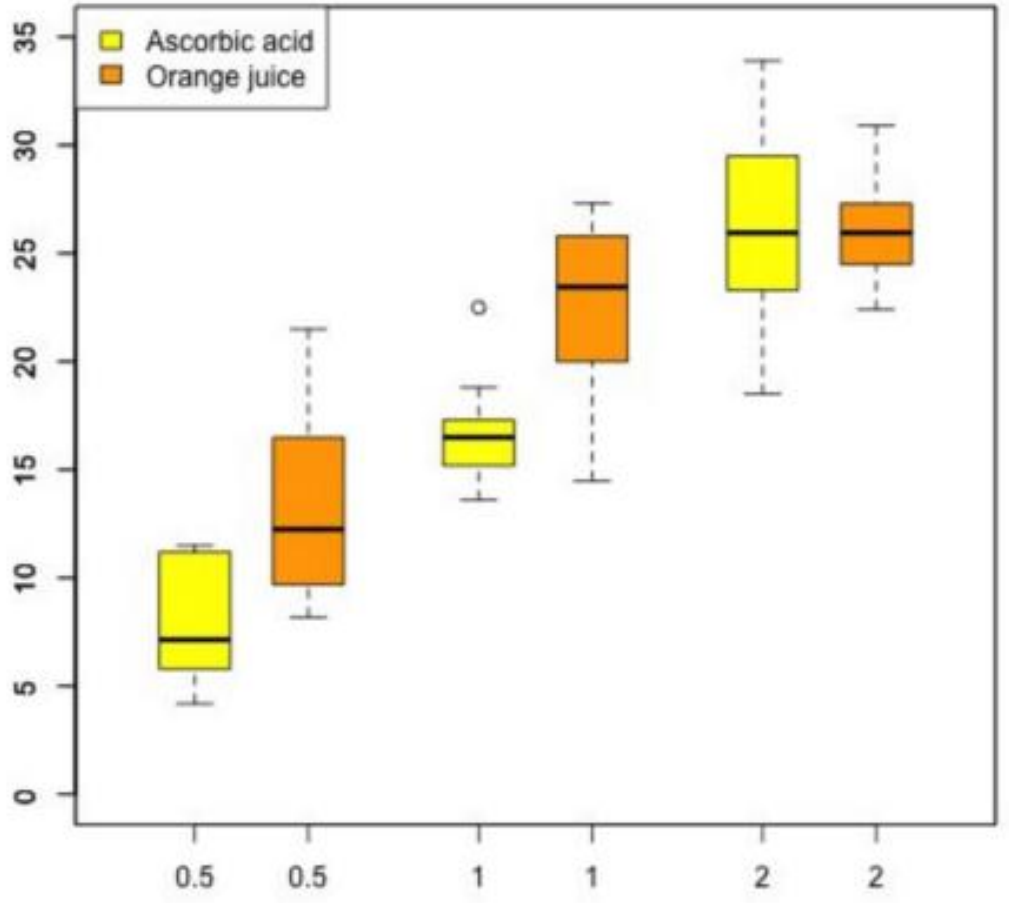
	2014 年 5 月至 6 月 May – Jun 2014	2013 年 5 月至 6 月 May – Jun 2013
每月工資分布（港元）：所有僱員 Monthly wage distribution (HK$) : All employees		
第十個百分位數 10th percentile	8,000	7,700
第二十五個百分位數 25th percentile	10,500	10,000
第五十個百分位數 50th percentile	14,800	14,100
第七十五個百分位數 75th percentile	23,000	22,000
第九十個百分位數 90th percentile	37,600	36,200



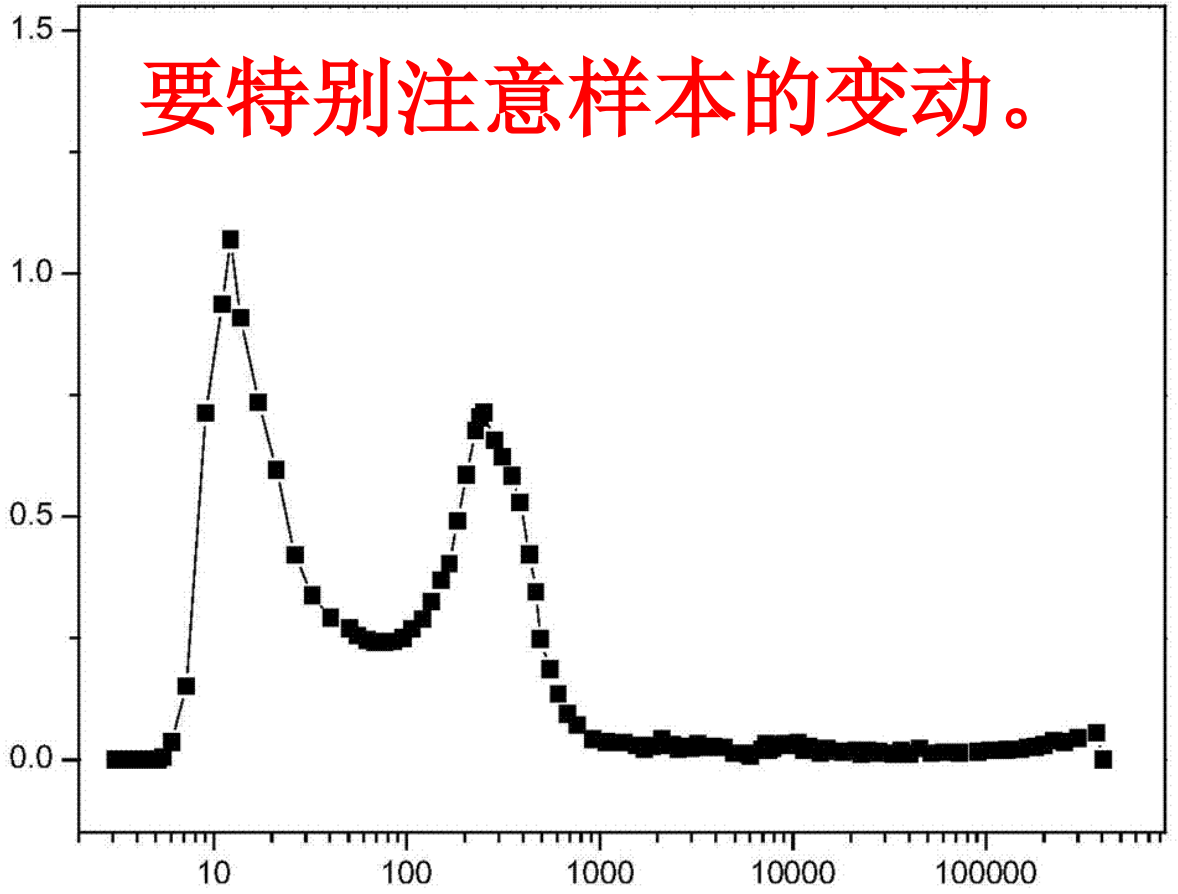
分位数与盒须图

- 分位数
  - 中位数、四分位数
- 分布分析
  - 分区间（分箱）
  - 频次统计
- 箱线图



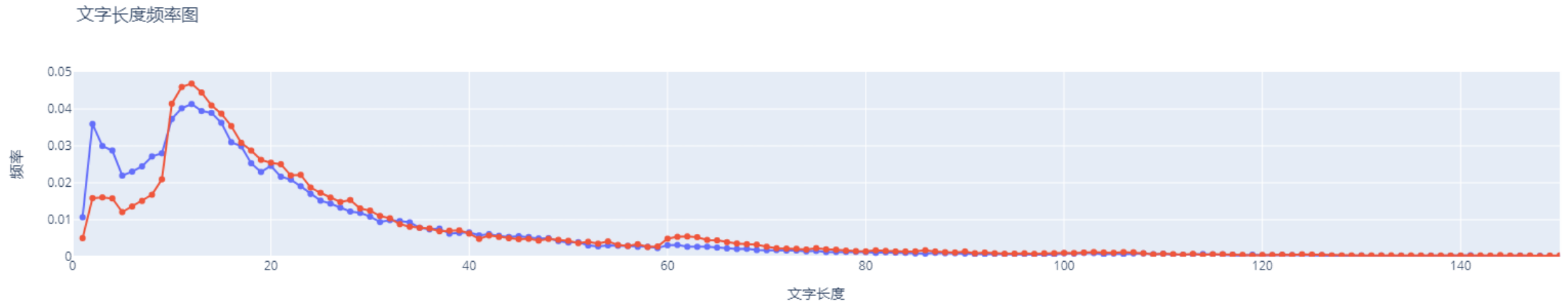


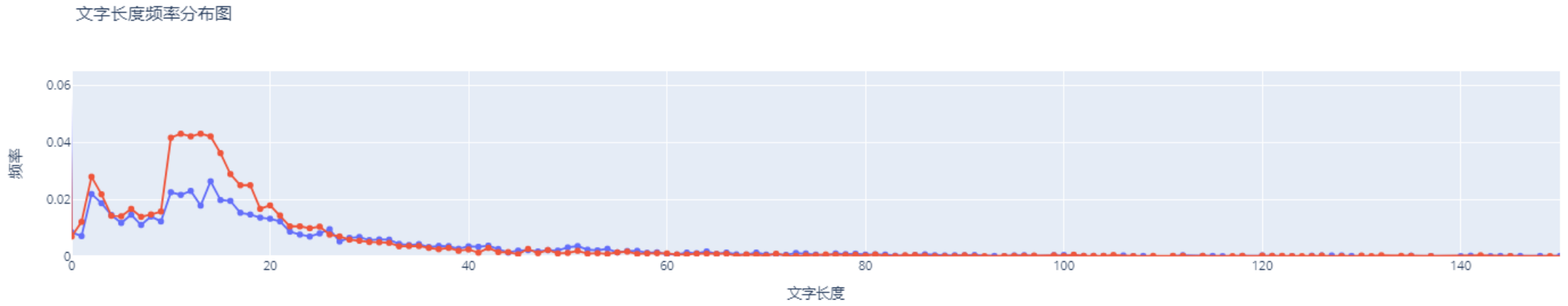
（2）多个样本混合体不能用平均值



要特别注意样本的变动。

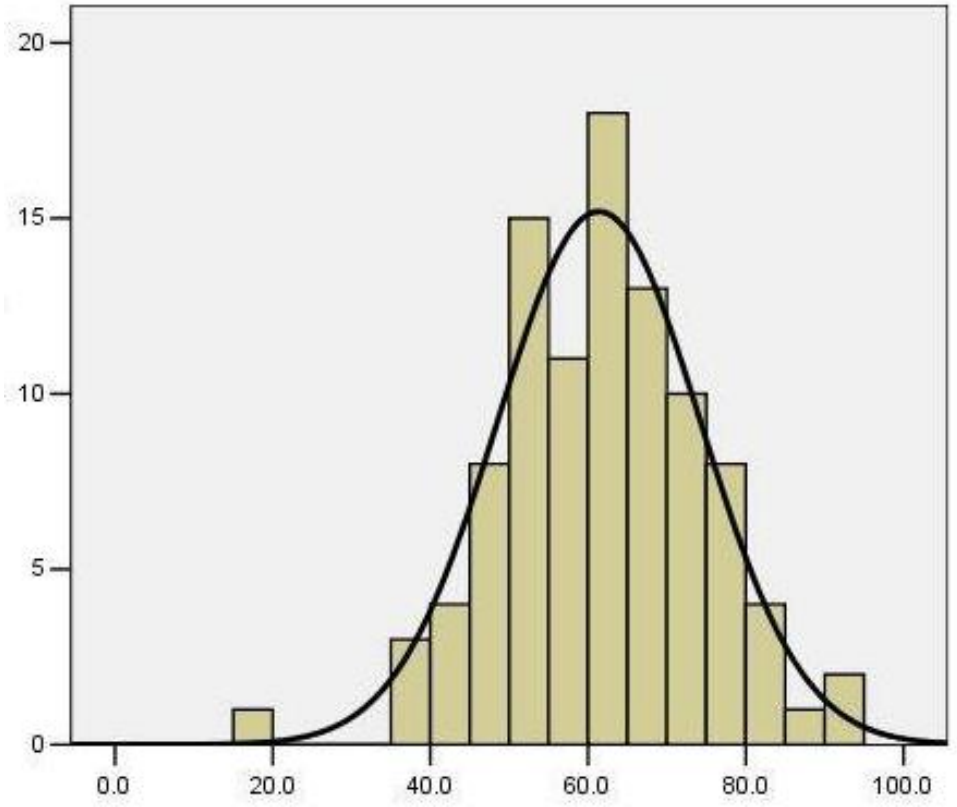
部分互联网类业务的数据分布

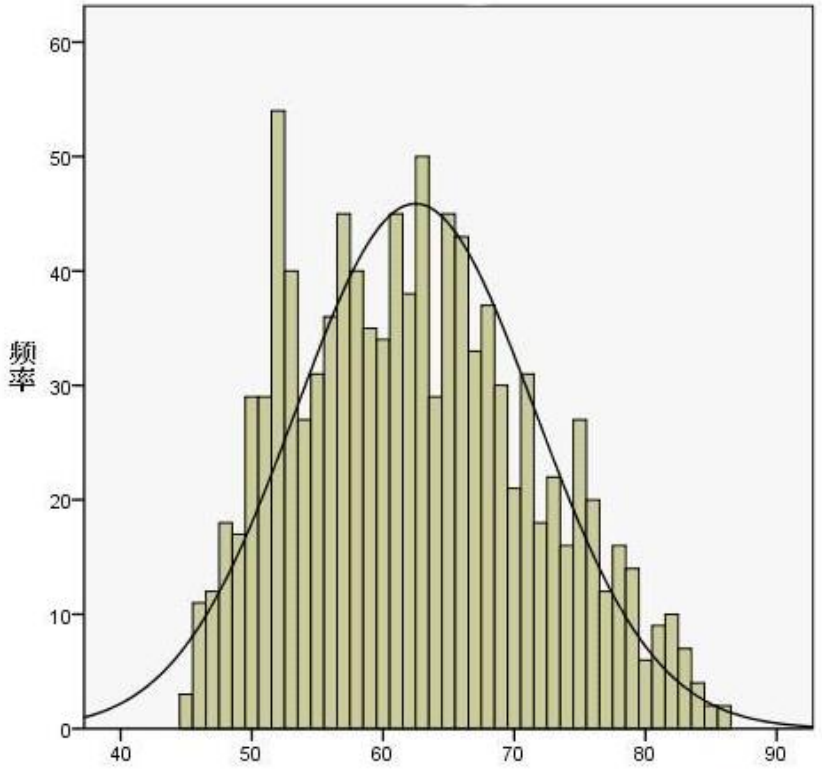




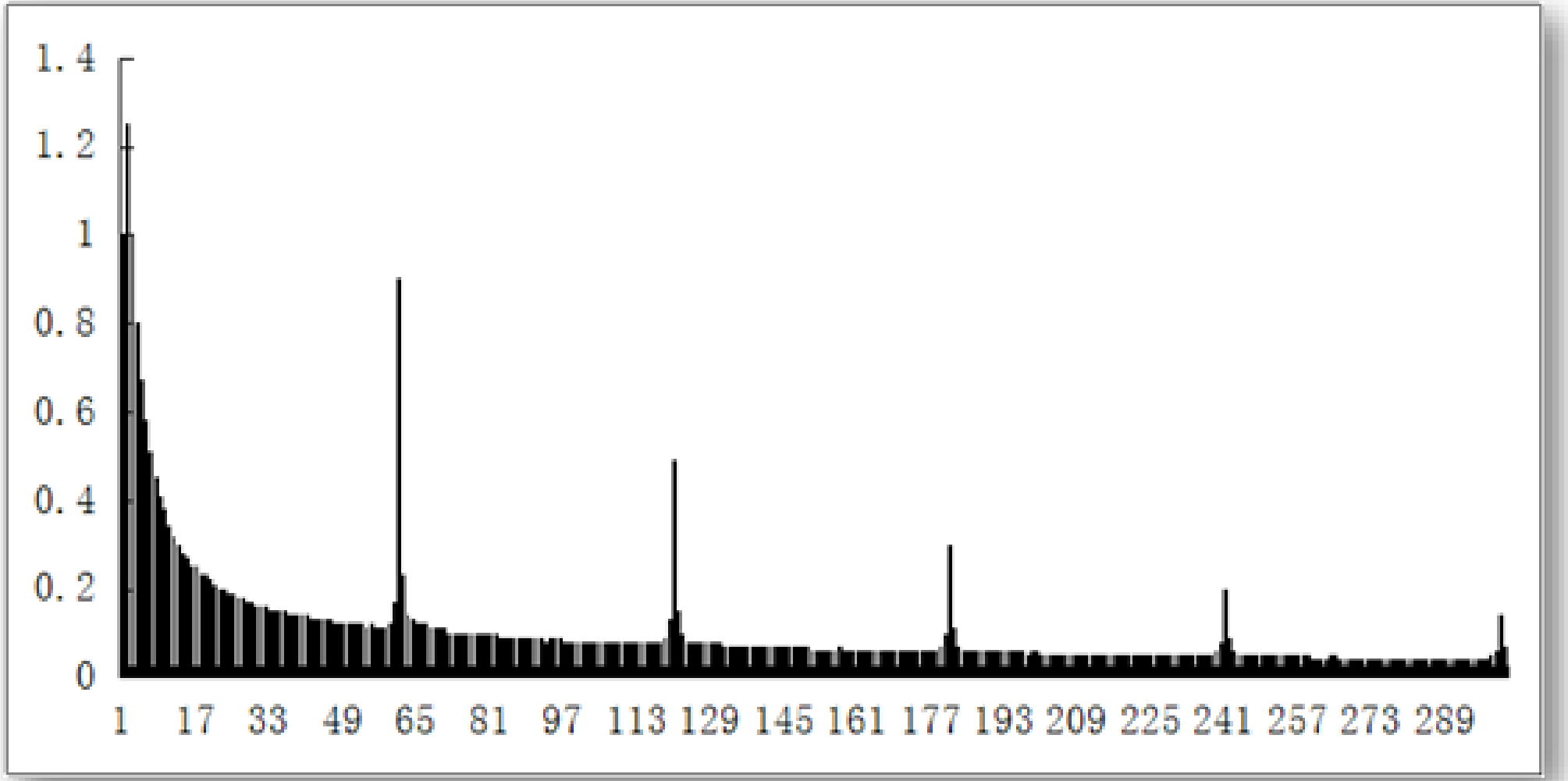
解决方案：直方图展示数据分布细节

- 直方图
- 分箱大小

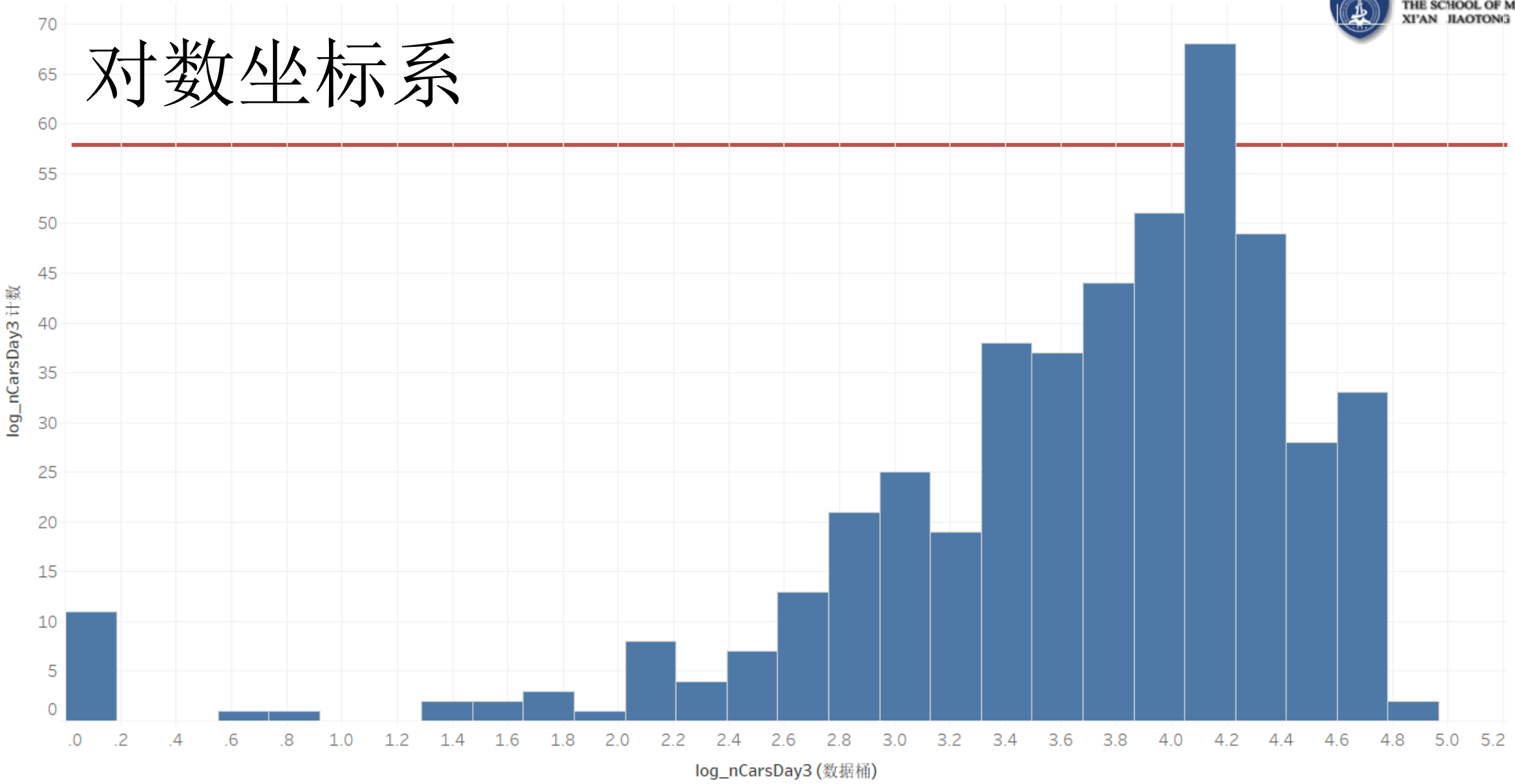




不平滑的分布线与“异常”



对数坐标系



Log_nCarsDay3 (数据桶) 的 log_nCarsDay3 计数 的趋势。

“单变量”数据探索小结

- 分类型变量
  - 饼、柱、条、词云、树、地图
  - 排序、其它、层次
  - 日期时间型变量转为有序分类变量处理
  - 慎用饼图和折线图

- 数值型变量
  - 描述性统计
  - 箱线图
  - 直方图

查看数据的分布

2. 双变量数据探索：关联关系

1. 双变量探索分类

- 双变量探索：**两种变量的关系探索**
- 问题：时间类型的变量不用考虑吗？

	数值	分类
数值	相关系数 散点图	
分类	分组比较 分类--统计指标 分类--分布	分组比较 二维表

2. 相关≠因果

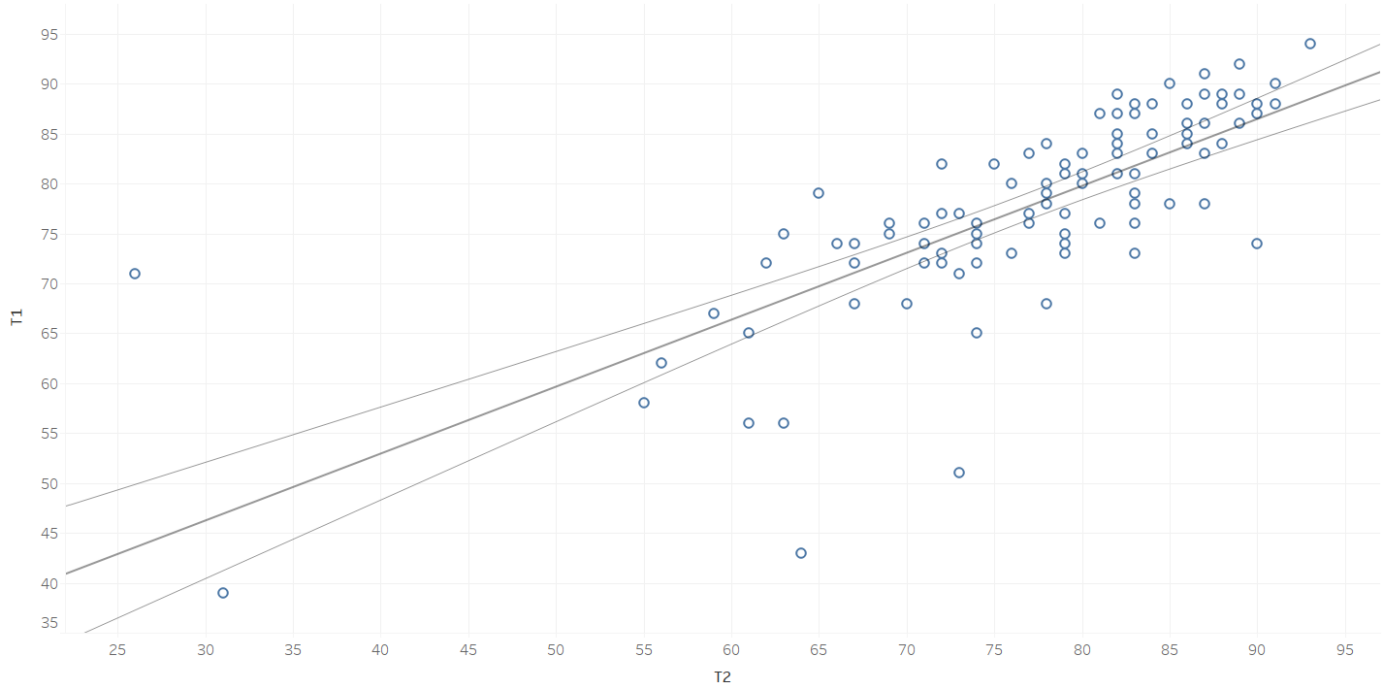
- 《爱上统计学》一书给了一个例子：在美国中西部的一个小镇，地方警察局局长发现冰淇淋消费量越多，犯罪率就越高。这个例子中，冰淇淋消费量和犯罪率是正相关的，但并不意味着冰淇凌消费的增多导致了犯罪率的上升，更不可能通过减少冰淇凌的销售来降低犯罪率。
- 事实上，存在某个变量同时和冰淇淋消费量、犯罪率相关，这个变量就是室外温度。当室外气温变暖，如在夏天，就会有更多犯罪（白天更长，人们多开窗口等）。而因为天气变暖，人们更享受吃冰淇凌的乐趣。相对地，在又长又黑暗的寒冬，冰淇凌的消费就减少，同时犯罪也越少。

“发现没有，中国武汉、日本东京、韩国大邱、美国西雅图，这几个新冠病毒肆虐的城市都以樱花闻名。结论，COVID-19的中间宿主是樱花。”

群里的一位data scientist：“这个例子，基本代表了作为数据分析师的我每天的工作内容。”

3. 数值 vs 数值

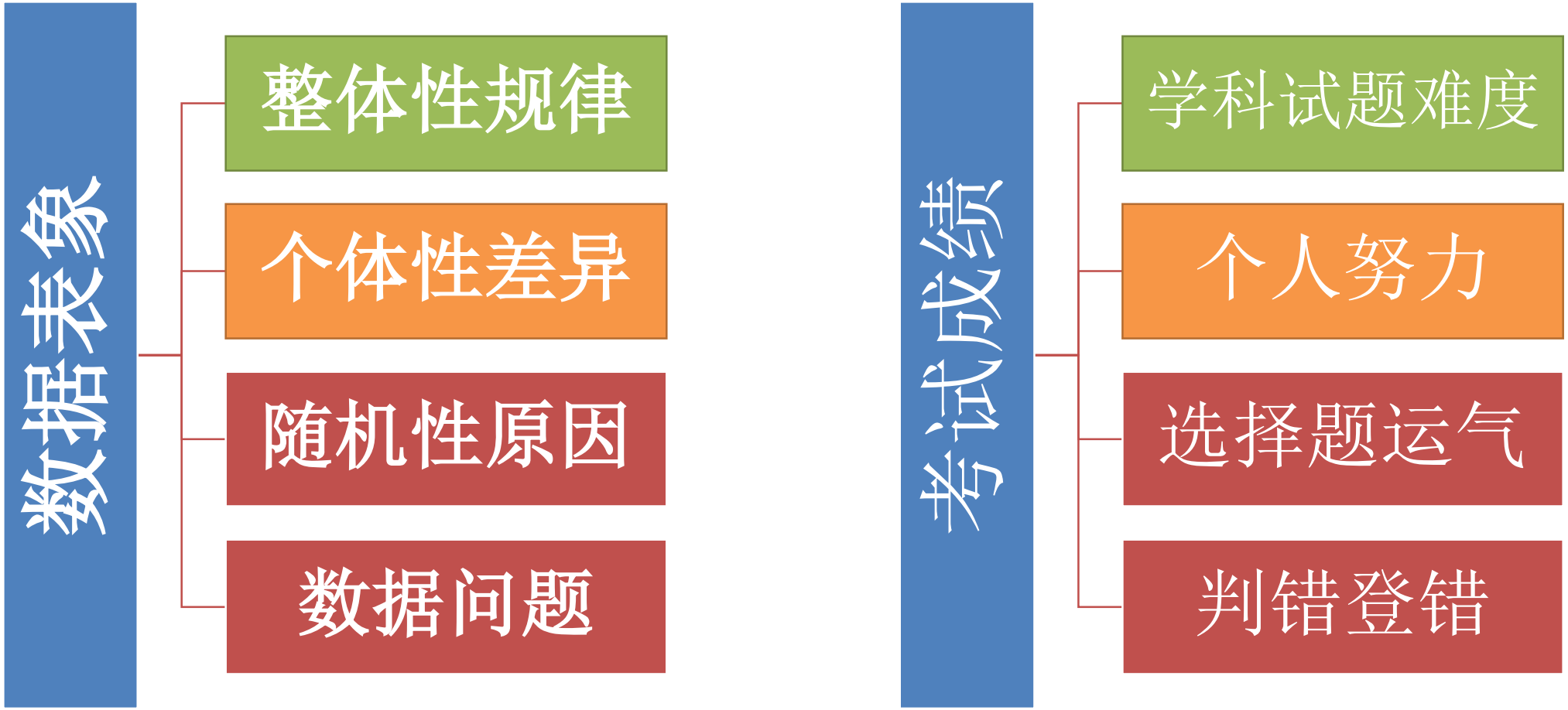
- 分析目标：**两个数值之间是否存在关系/存在何种关系？**
- 散点图，趋势线，趋势线公式，置信区间，曲线拟合



回归分析

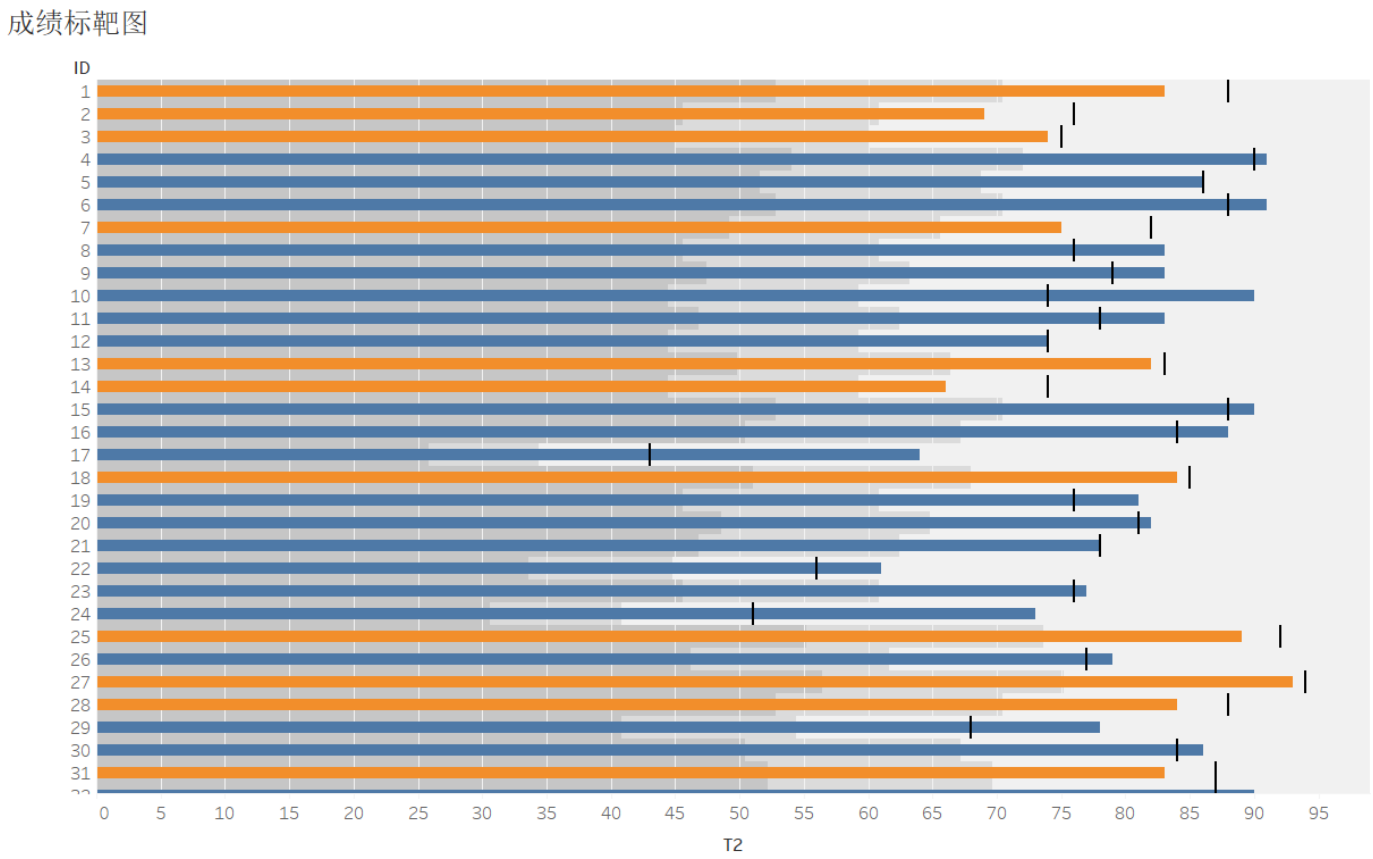
一元线性回归分析

数据表象背后的原因



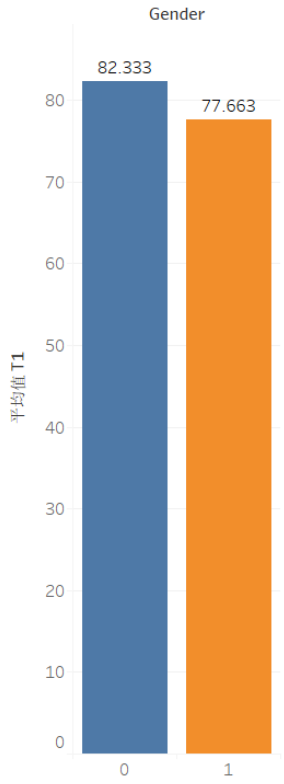
特例：标靶图/靶心图

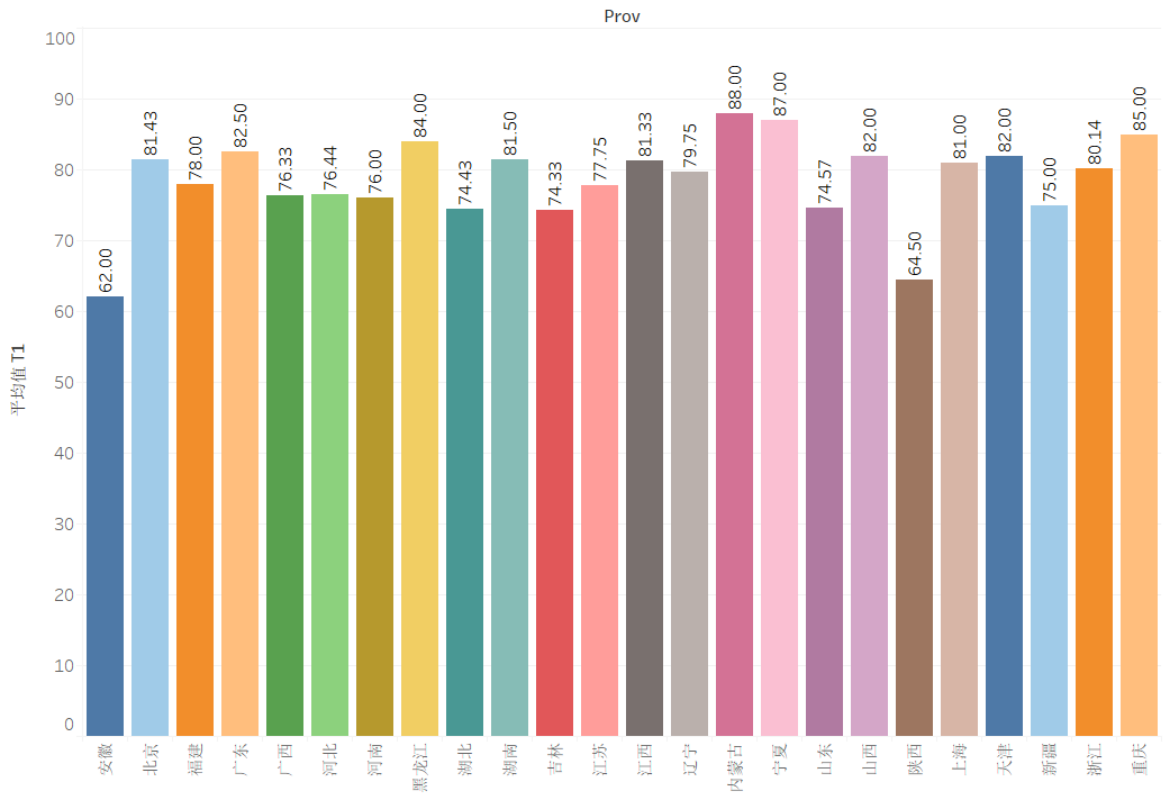
- 分析目标：其中有一个是参考值（如历史值、目标值）时，与参考值作比较。



4. 数值 vs 分类

- 分析目标：各个分类的数值比较。
- 最简单的办法：简单统计指标（如均值）分类比较。

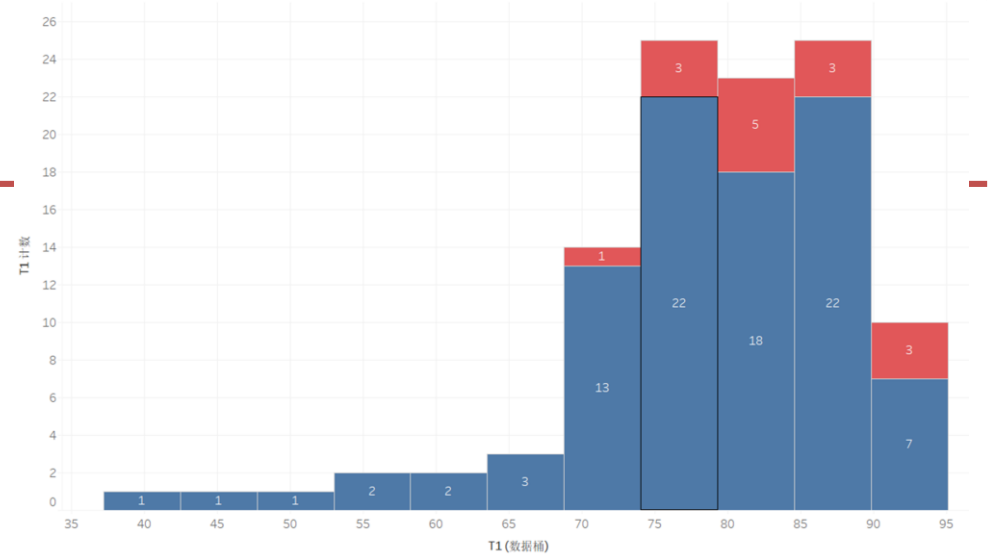




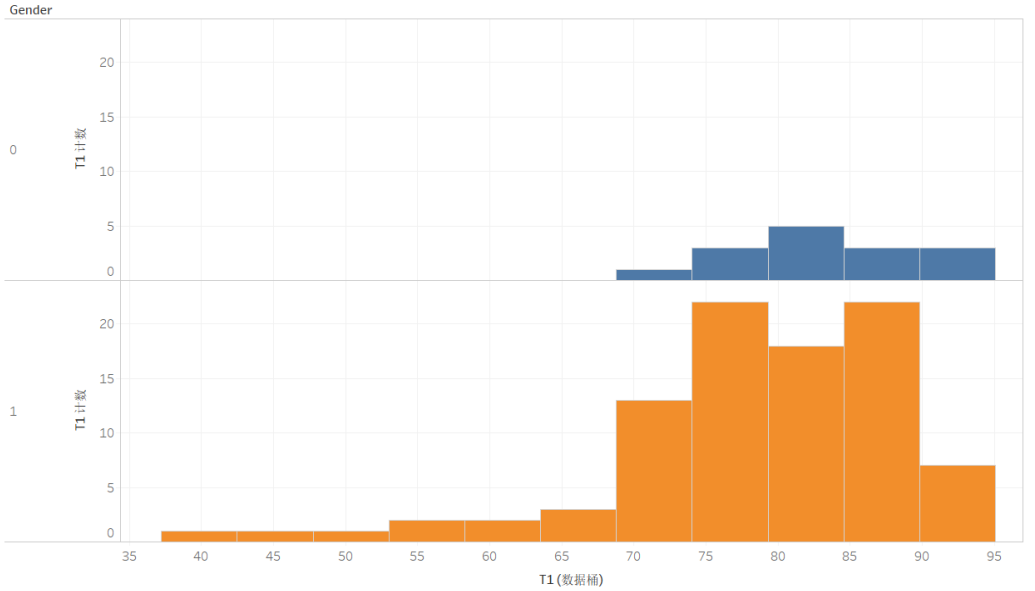
T检验
ANOVA

直方图分类比较

- 直方图分类比较：
  - 对细节观察更为深入
  - 适用于分类较少的情况
  - 分类数量过多时没有现实意义

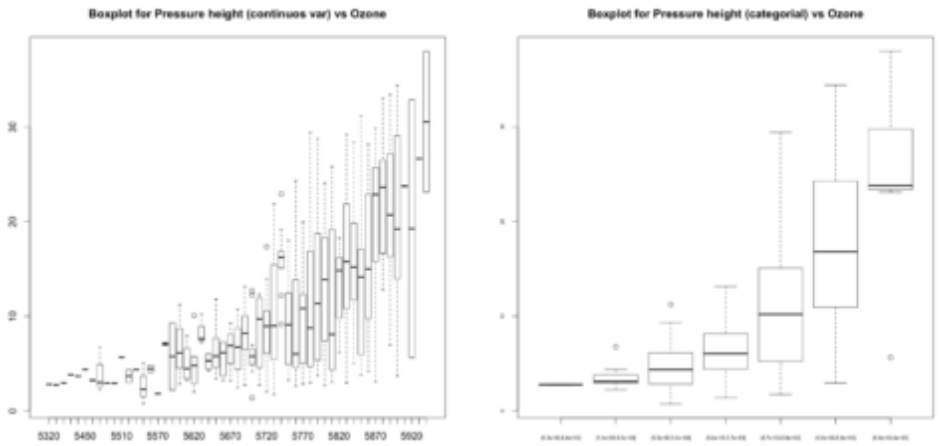


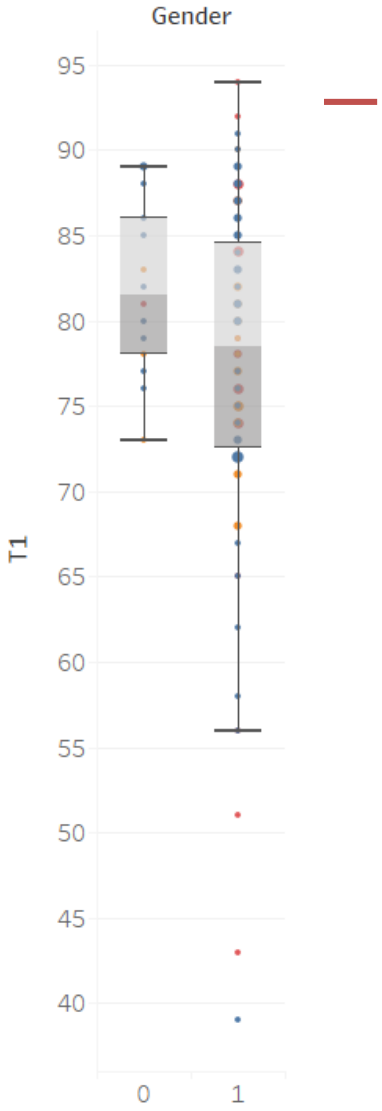
- 可用颜色、分区、面板等来表示新增的分类变量



箱线图分类比较

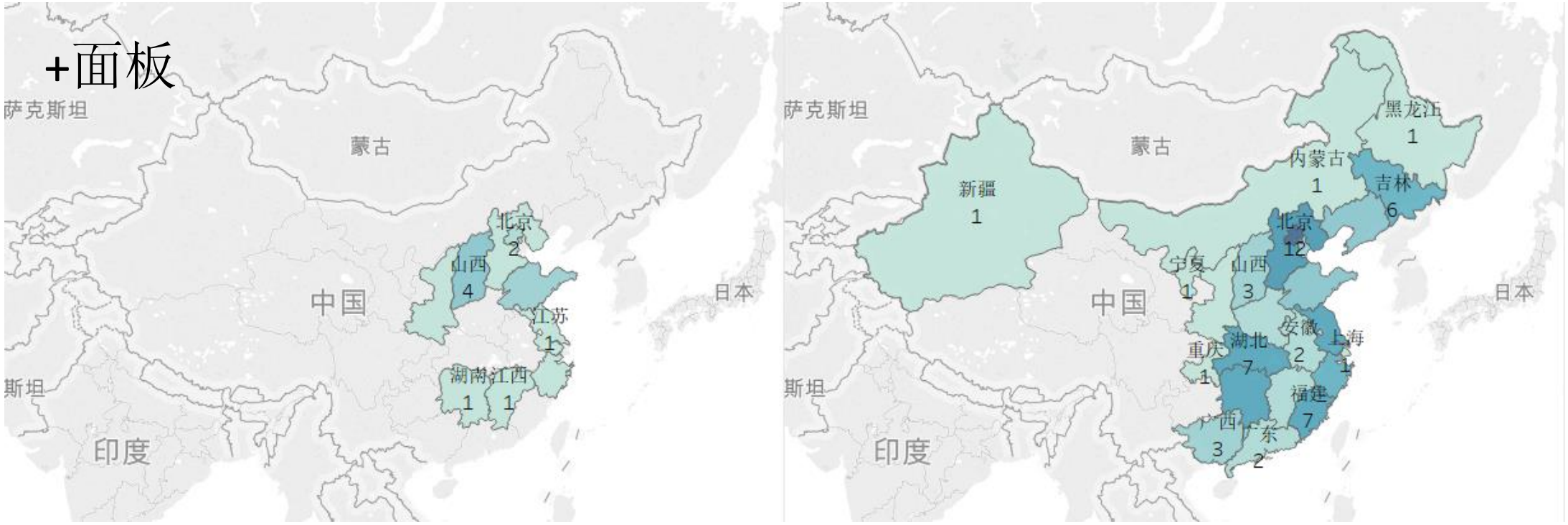
- 适用于分类较多的情况、且能观察数据分布。
- 给定一个连续变量后，**离群值**可以认为是哪些超出1.5倍四分位距的观测点。四分位距是0.25分位数和0.75分位数的差。
- 通过箱线图来检测离群点，在须轴以外的点就是离群点。

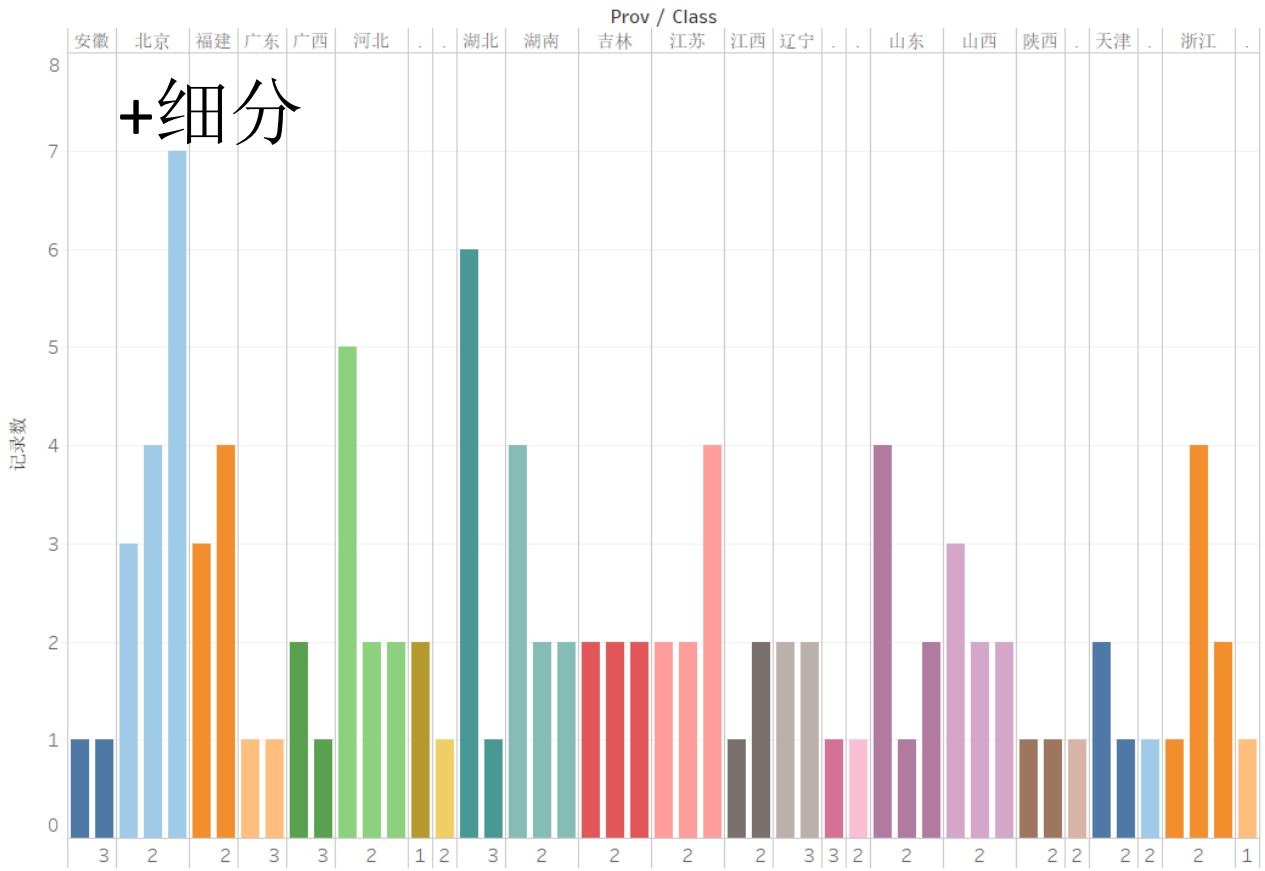


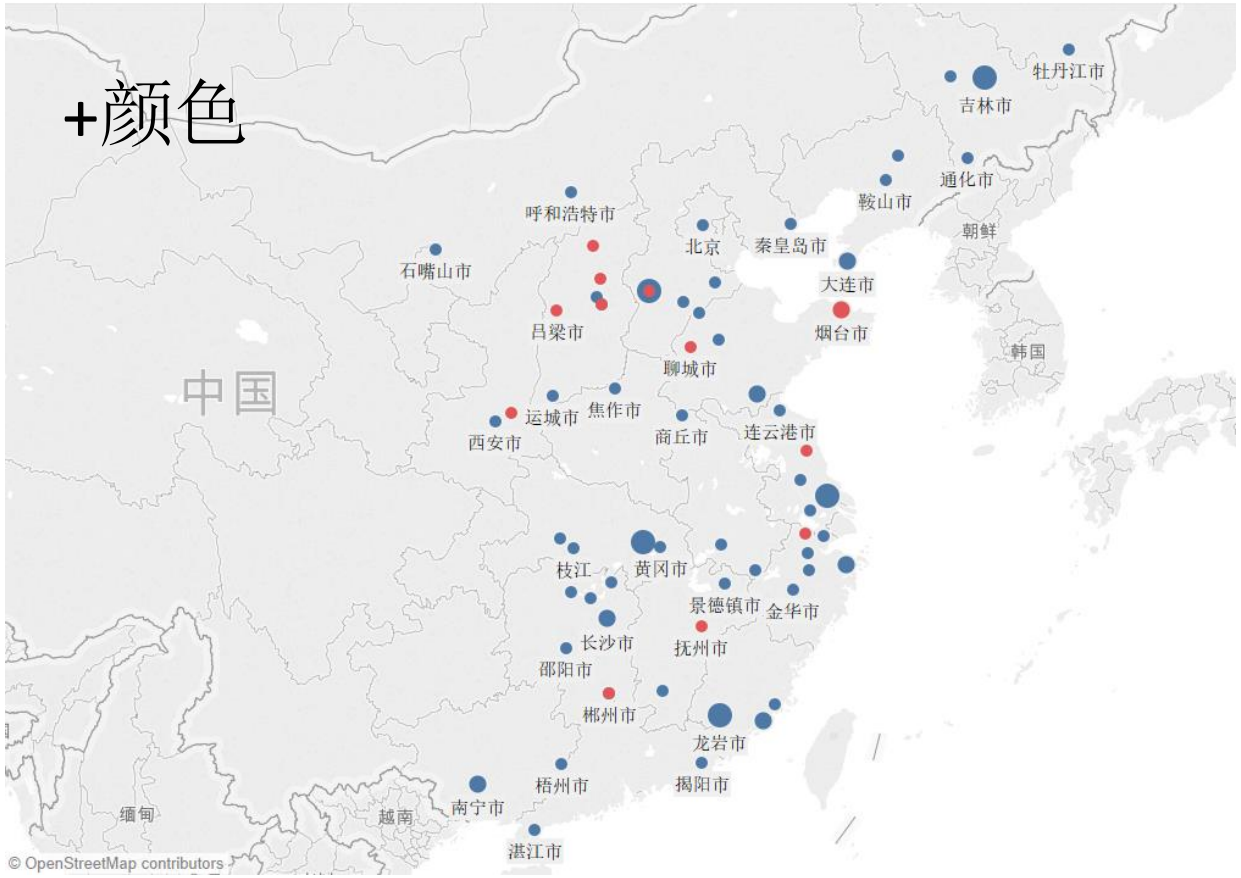


5. 分类 vs 分类

- 分析目标：数据在两个维度上的联合分布。
- 可视化：可以在分类单变量分析的基础上叠加第二个分类变量：可用颜色、分区、面板、页面等来表示新增的分类变量。

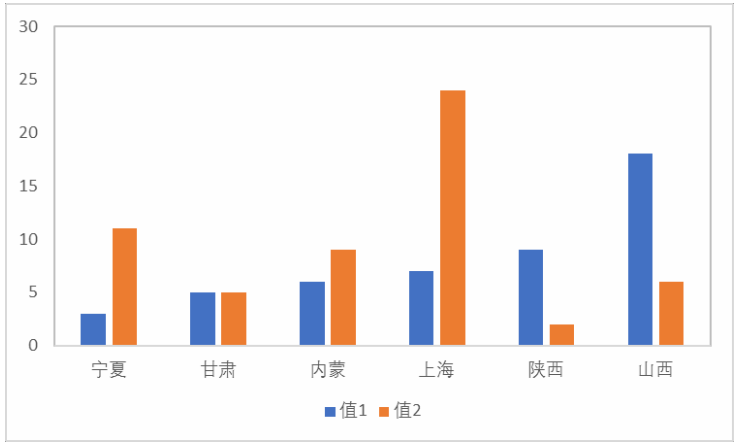


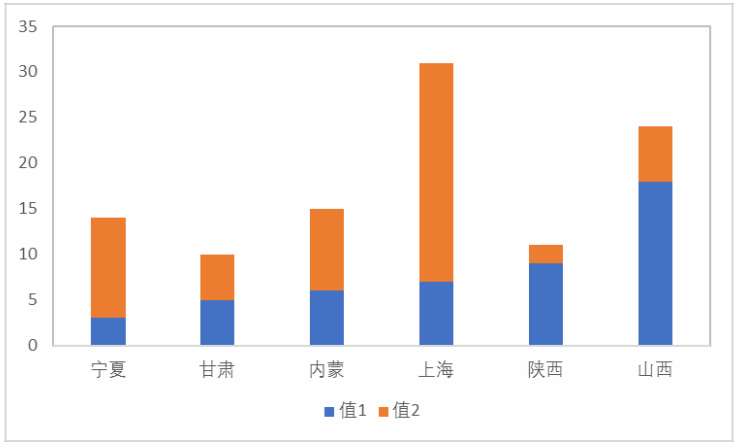


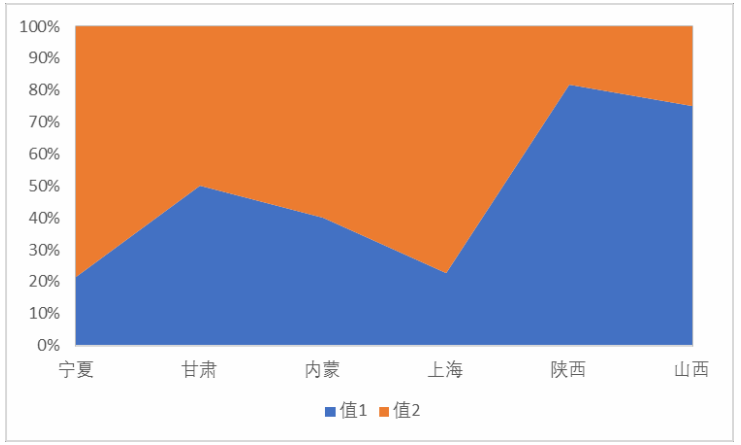


慎用面积图/堆积图

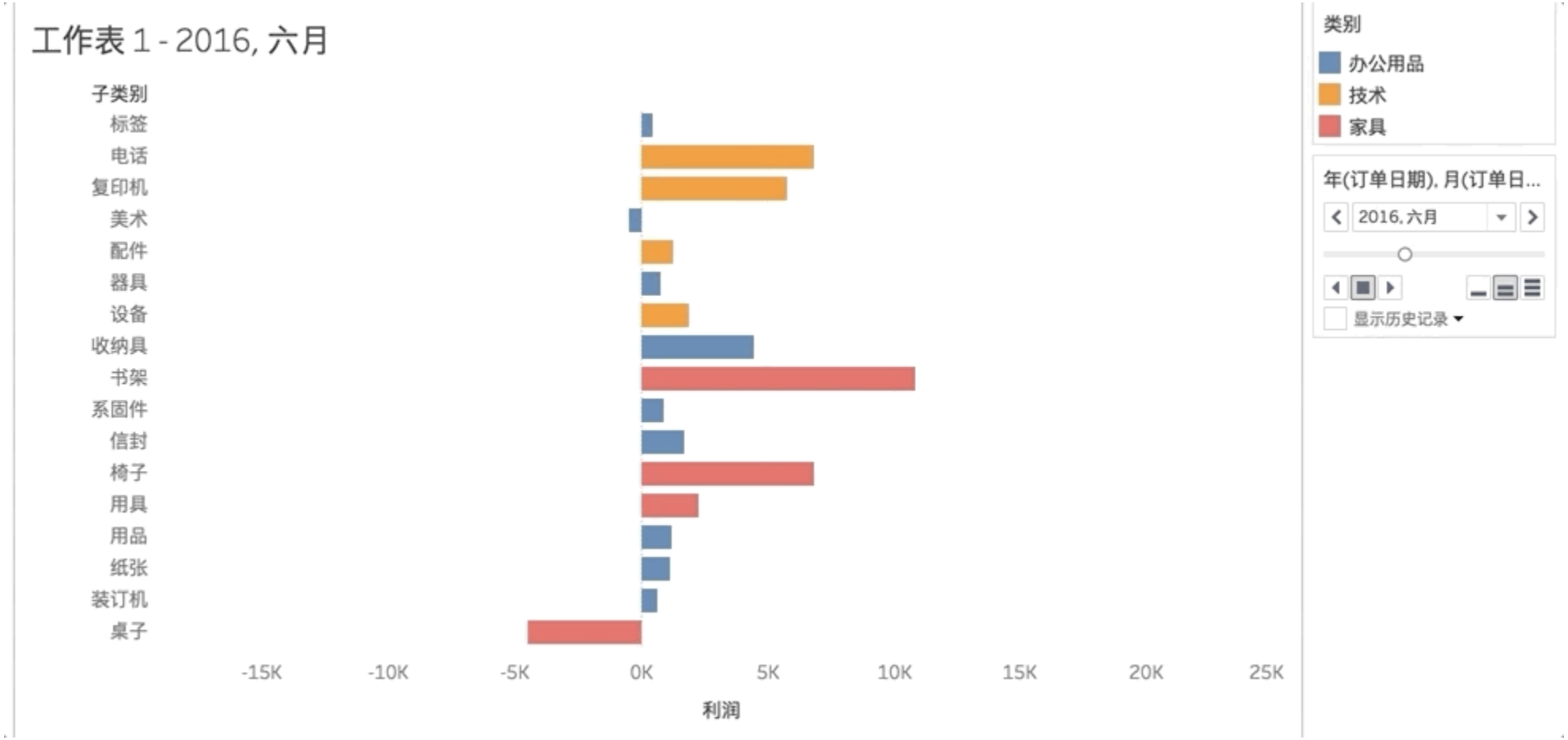
- 当横坐标是分类变量时，使用面积图，分类变量中间的部分是毫无意义的，且容易误导人。





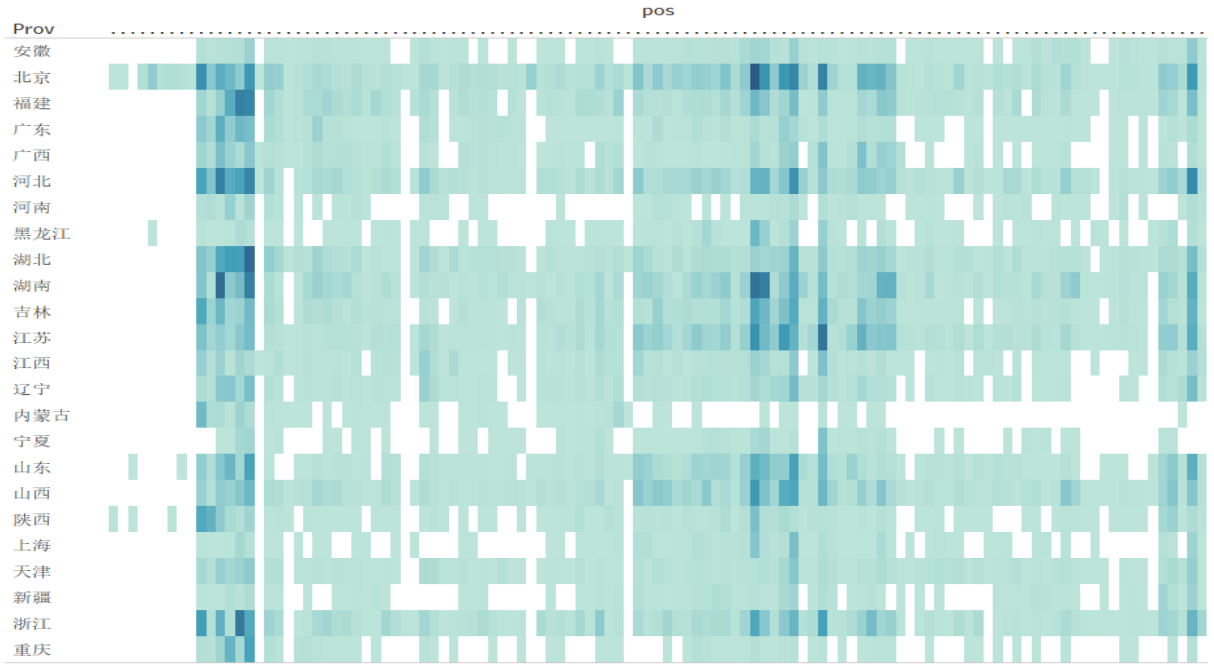


+页面：让图动起来

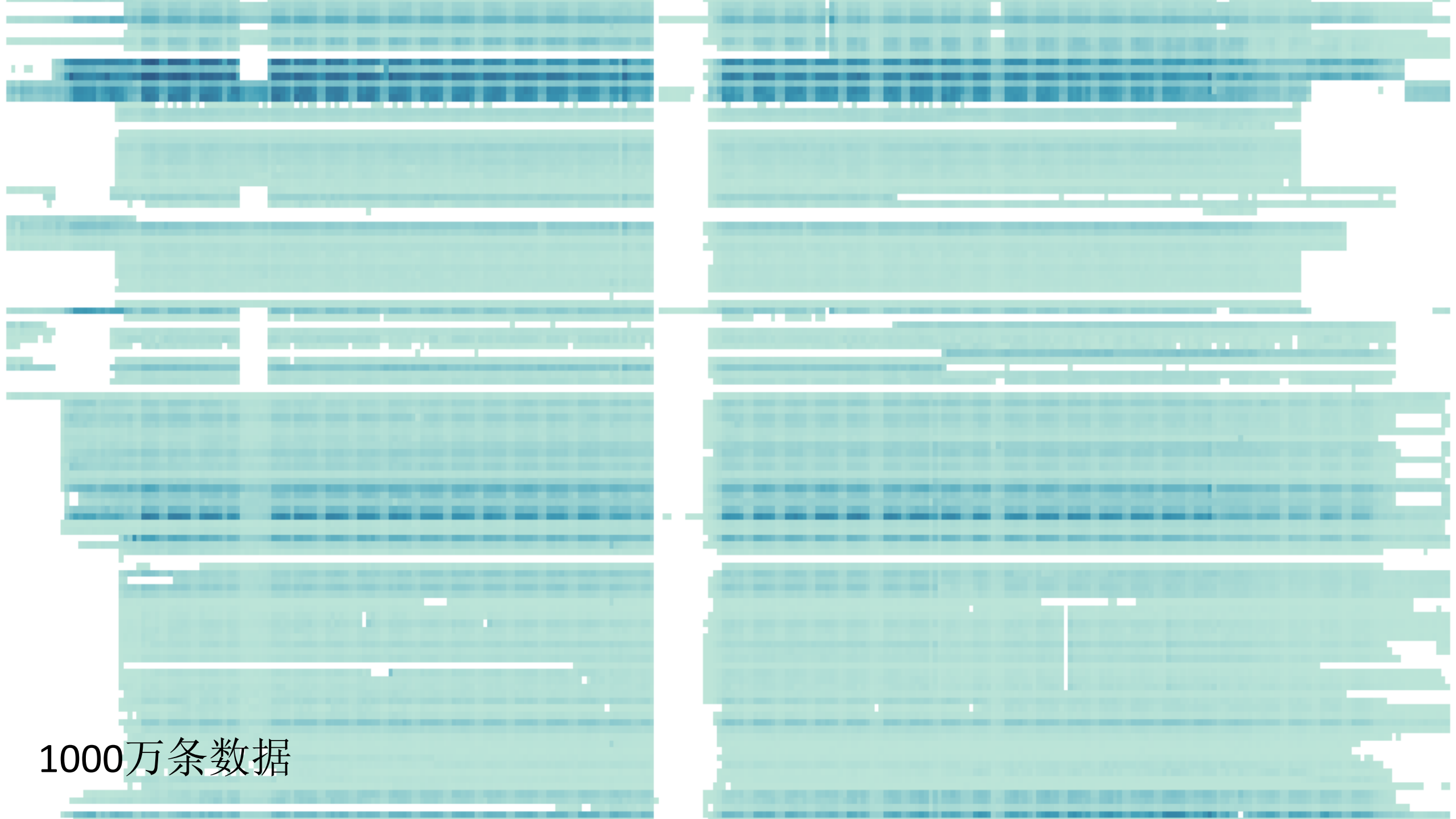


二维表

- 分析目标：数据质量评估、周期/规律识别
- 特例：时间-对象二维表



	Class		
Prov	1	2	3
安徽	1		1
北京	3	4	7
福建	3	4	
广东	1		1
广西	2		1
河北	5	2	2
河南	2		
黑龙江		1	
湖北	6		1
湖南	4	2	2
吉林	2	2	2
江苏	2	2	4
江西	1	2	
辽宁	2		2
内蒙古			1
宁夏		1	
山东	4	1	2
山西	3	2	2
陕西	1	1	
上海		1	
天津	2	1	
新疆		1	
浙江	1	4	2
重庆	1		


1000万条数据

3. 三变量可视化分析及可视化交互

数据分布、现状描述

1. 三变量及以上的探索——叠加

- 分析目标：关系比较、分布比较、规律探索

- 可视化分析方法：底图+叠加
  - 选好底图：以单变量及双变量图为底图
  - 叠加数值型变量：颜色（热力）、尺寸、大小
  - 叠加分类型变量：颜色、面板（分行/列）、页面

- 可视化分析策略：
  - 三个变量的组合可能性非常多，没必要一一尝试，大部分也没有意义。
  - 基于业务理解，探索有价值的分析视角。

2. 三变量组合的可能性

- 分类、分类、分类
- 分类、分类、数值
- 分类、数值、数值
- 数值、数值、数值

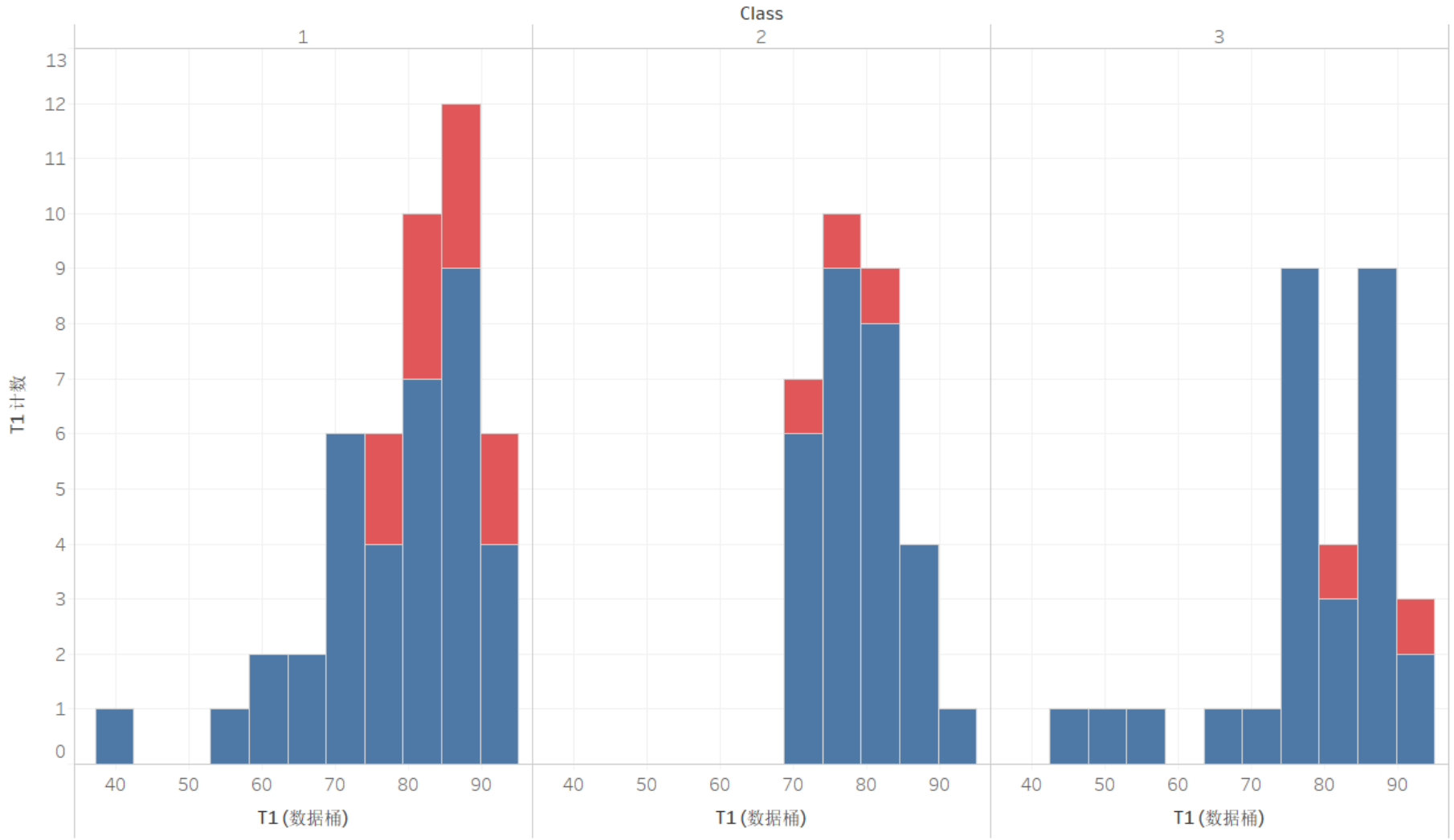
- 尽管只有4个大类，但是拆分方法很多，取决于要实现的目标和变量的具体情况，例如：
- 分类、分类、分类：
  - 分类（分布图）+ 分类、分类：叠加颜色、面板、页面等
  - 分类、分类（二维表）+ 分类：叠加面板、页面等

分类、分类、分类（二维表+柱状图）

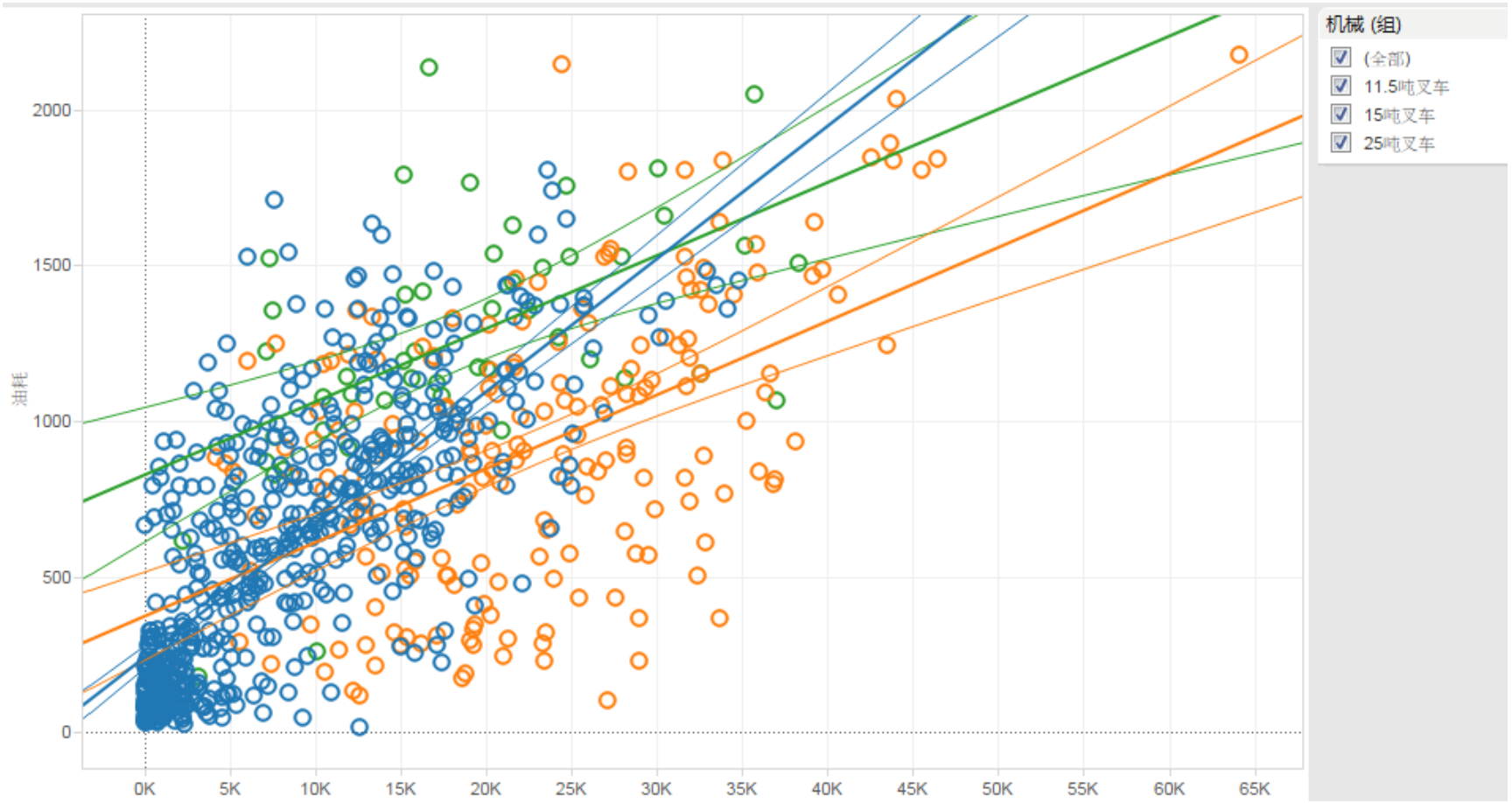
Prov

Class	Gender	安徽	北京	福建	广东	广西	河北	河南	黑龙..	湖北	湖南	吉林	江苏	江西	辽宁	内蒙..	宁夏	山东	山西	陕西	上海	天津	新疆	浙江	重庆
1	1																								
	0																								
2	1																								
	0																								
3	1																								
	0																								

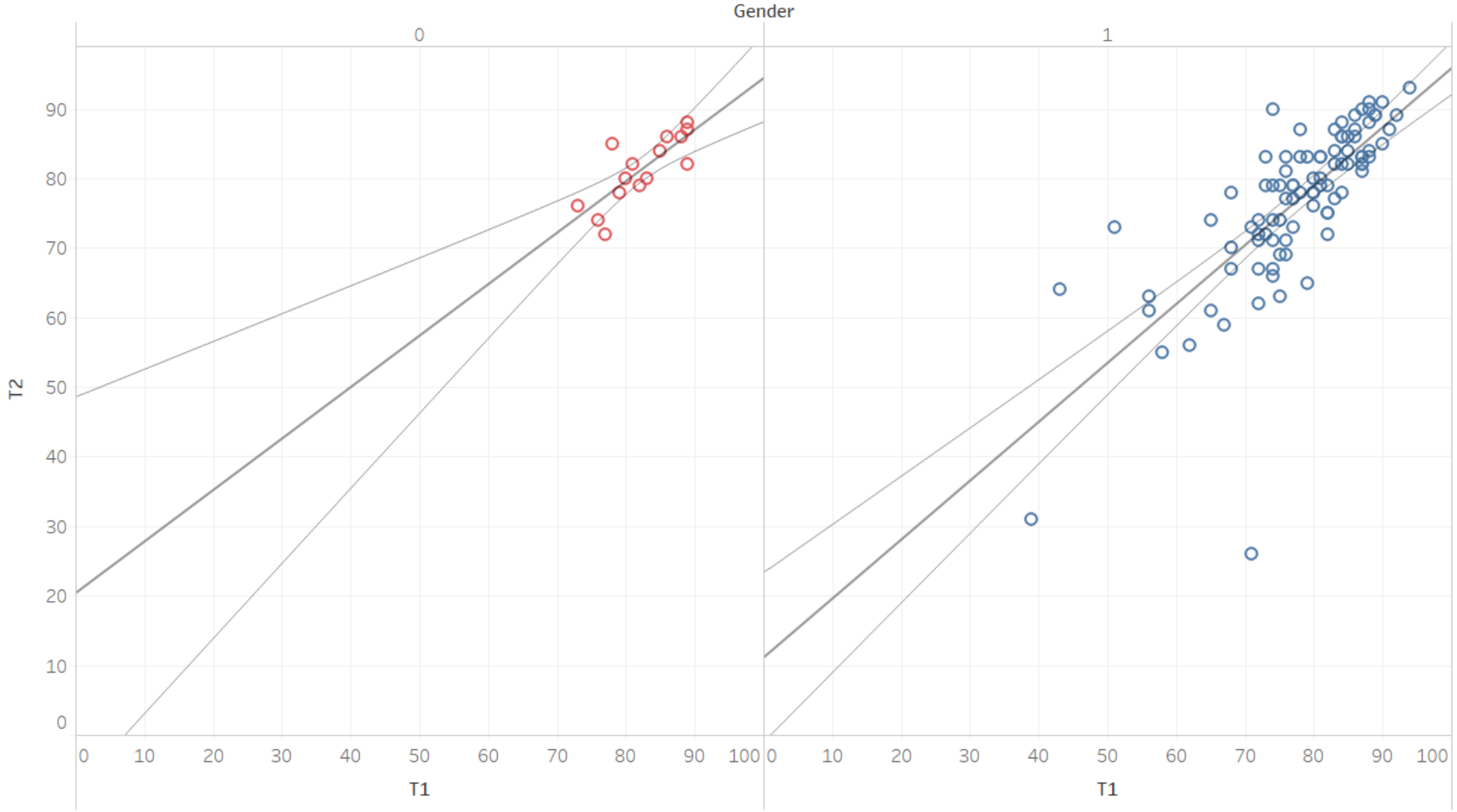
分类、分类、数值（直方图+颜色+面板）



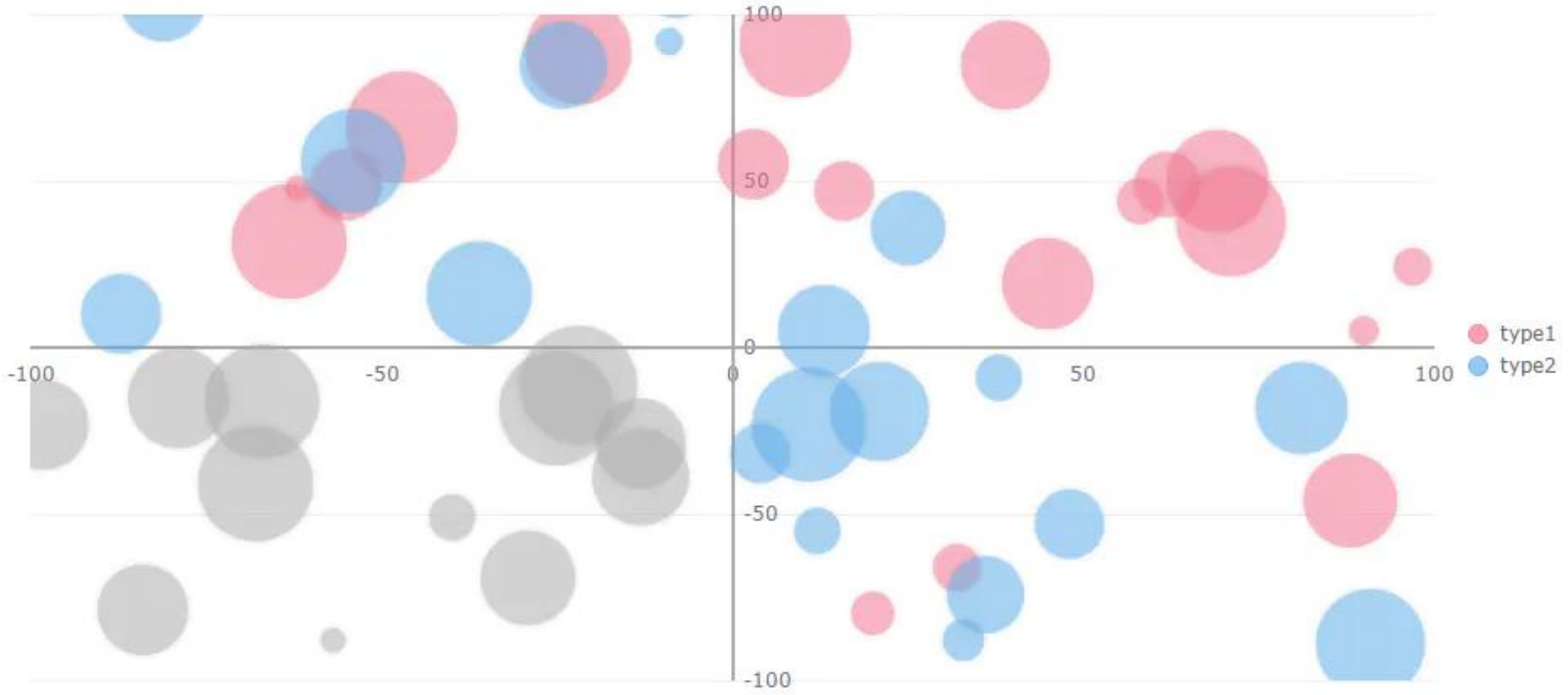
示例：数值、数值、分类（颜色）

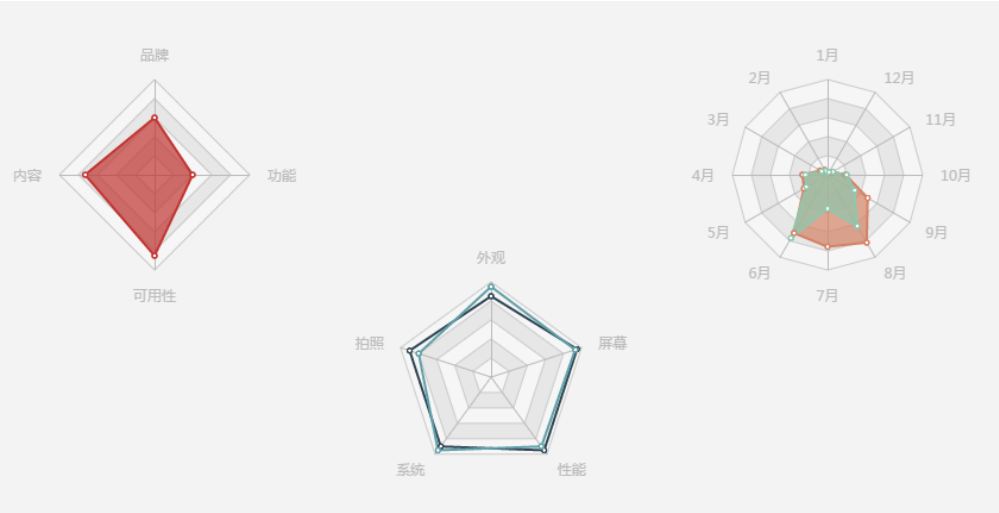


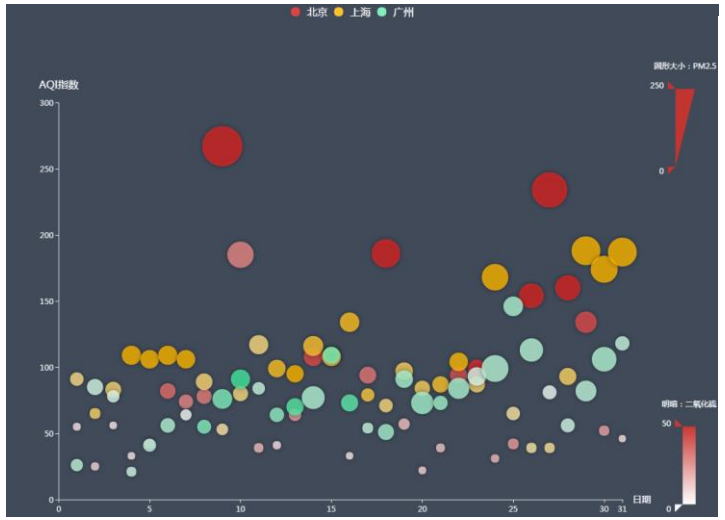
数值、数值、分类（面板）



数值、数值、数值（大小）：气泡图







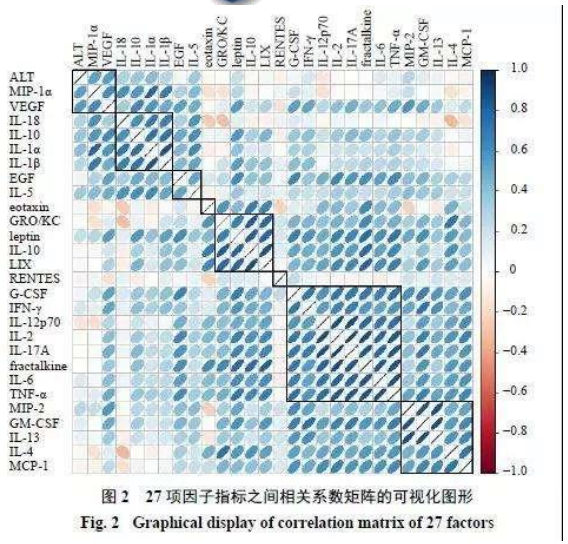


图 2　27 项因子指标之间相关系数矩阵的可视化图形

Fig. 2　Graphical display of correlation matrix of 27 factors

画更多的变量，可以吗？

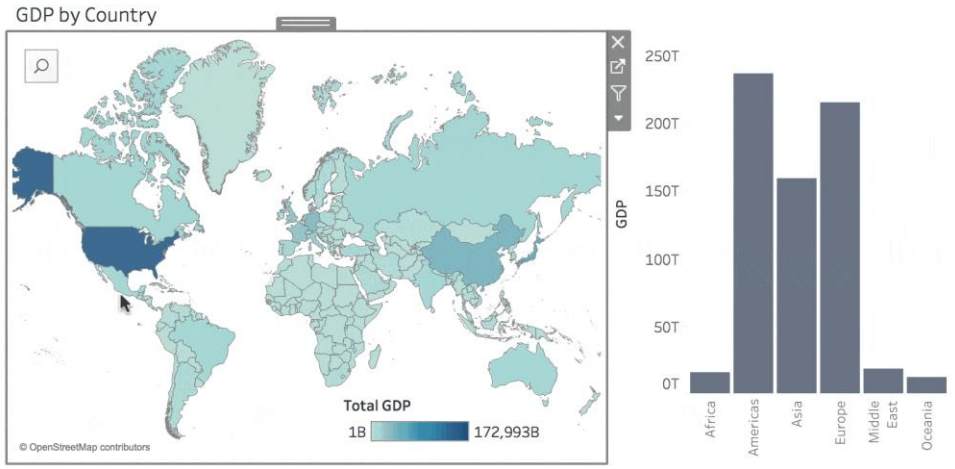
叠加、再叠加 or 复杂的图形

画更多的变量，非常困难！

多元统计分析/机器学习模型

3. Tableau操作及可视化

- 打开数据、数据类型、生成图表
- 地图、词云、树状图
- 多层次变量
- 动态图表：页面
- 查看数据及导出数据
- 生成变量（函数的使用）、使用参数
- 趋势线及时间序列外推
- 筛选器及过滤器
- 从Worksheet到Dashboard到Story
- Dashboard的交互性：作为过滤器



西安交通大学管理学院
THE SCHOOL OF MANAGEMENT
XI'AN JIAOTONG UNIVERSITY

2.3 动态数据探索

刘跃文 博士

教授、博士生导师

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

V2.2, 2021-9-8

1. 动态数据的基本构成

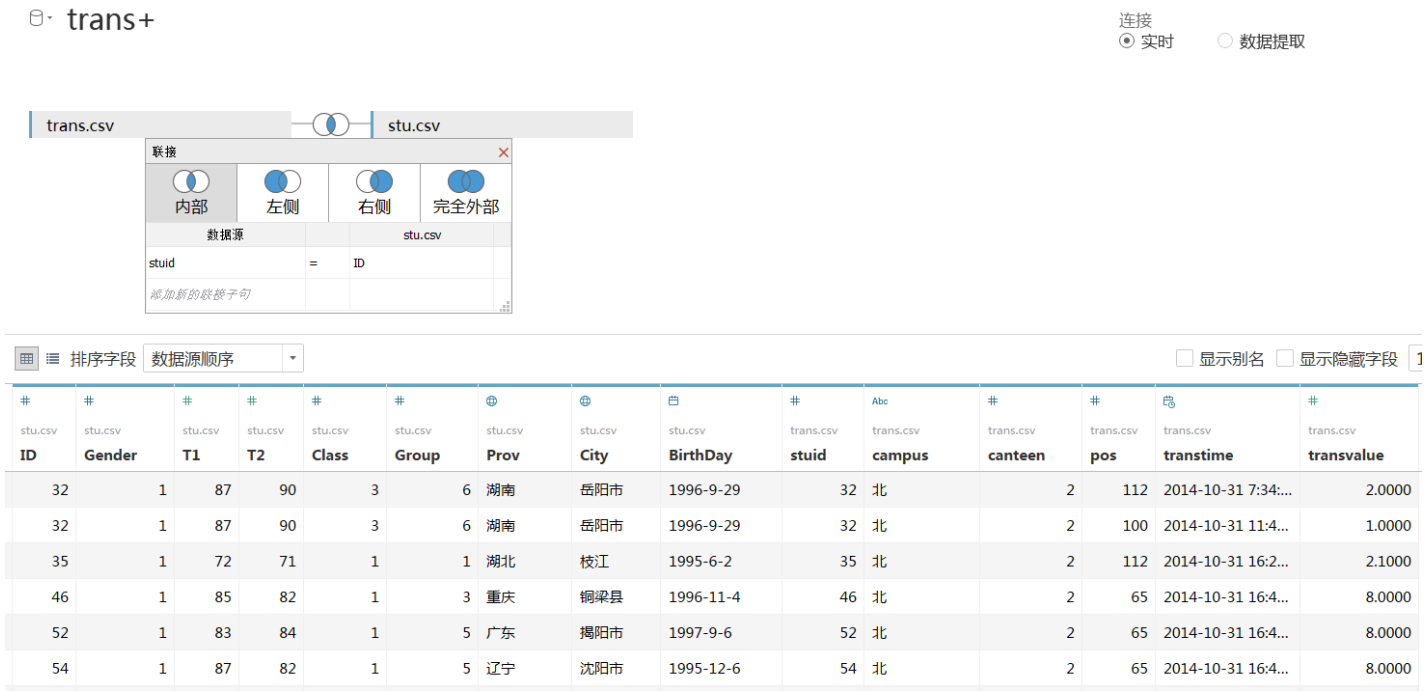
- 动态数据的基本构成：
  - 记叙文6要素：人物、时间、地点、事件起因、经过、结果
  - 动态数据5要素：{ID、时间、地点、事件、属性}
  - 例：学生卡号，时间，POS机号，消费类型，金额

- 5个要素中，4+个维度，1+个度量
  - 普遍存在的度量：数据条数

- 所有详细的账目都是动态数据

动态数据的简单扩展

- 动态数据的扩展方法：关联静态数据
- 关联静态数据后可以获得扩展的维度及度量

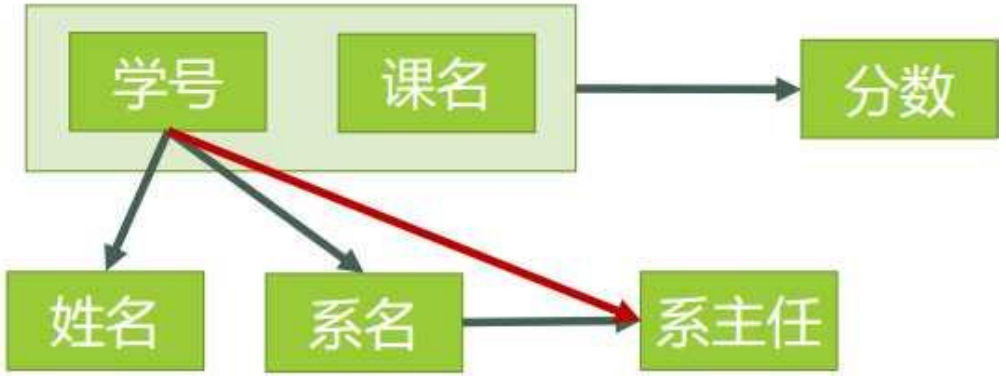


stu.csv ID	stu.csv Gender	stu.csv T1	stu.csv T2	stu.csv Class	stu.csv Group	stu.csv Prov	stu.csv City	stu.csv BirthDay	trans.csv stuid	trans.csv campus	trans.csv canteen	trans.csv pos	trans.csv transtime	trans.csv transvalue
32	1	87	90	3	6	湖南	岳阳市	1996-9-29	32	北	2	112	2014-10-31 7:34:...	2.0000
32	1	87	90	3	6	湖南	岳阳市	1996-9-29	32	北	2	100	2014-10-31 11:4...	1.0000
35	1	72	71	1	1	湖北	枝江	1995-6-2	35	北	2	112	2014-10-31 16:2...	2.1000
46	1	85	82	1	3	重庆	铜梁县	1996-11-4	46	北	2	65	2014-10-31 16:4...	8.0000
52	1	83	84	1	5	广东	揭阳市	1997-9-6	52	北	2	65	2014-10-31 16:4...	8.0000
54	1	87	82	1	5	辽宁	沈阳市	1995-12-6	54	北	2	65	2014-10-31 16:4...	8.0000

信息化与大数据的数据结构区别

- 信息化--业务数据库：范式要求、数据拆小。
- 大数据--分析数据库：分析需要、字段更多。

学号	姓名	系名	系主任	课名	分数
1022211101	李小明	经济系	王强	高等数学	95
1022211101	李小明	经济系	王强	大学英语	87
1022211101	李小明	经济系	王强	普通化学	76
1022211102	张莉莉	经济系	王强	高等数学	72
1022211102	张莉莉	经济系	王强	大学英语	98
1022211102	张莉莉	经济系	王强	计算机基础	88
1022511101	高芳芳	法律系	刘玲	高等数学	82
1022511101	高芳芳	法律系	刘玲	法学基础	82



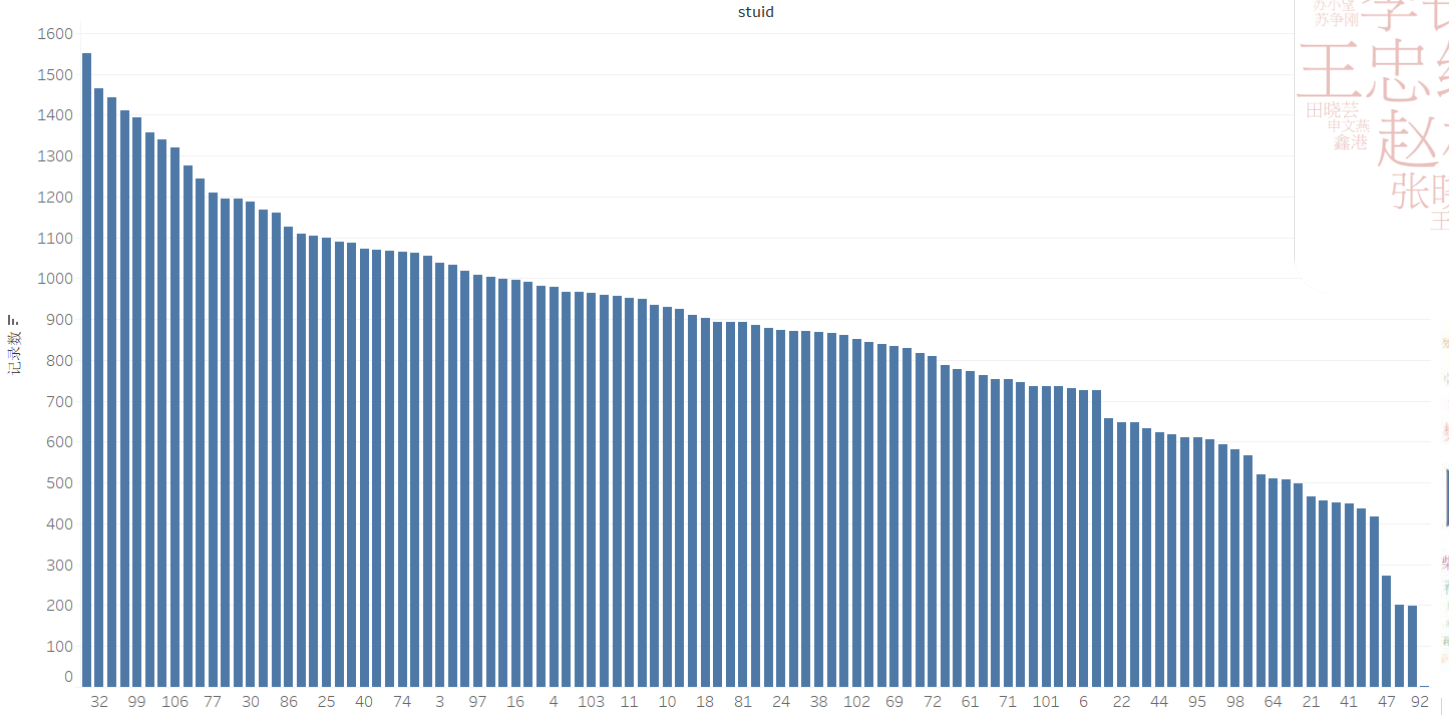
2. 单变量分析

- **动态数据分析，不要放过任何一个变量。**
- **实际业务远远超出我们的想象（业务理解、数据理解的关键）。**

- 分类型变量
  - 人、地、事、物、时间、组织……

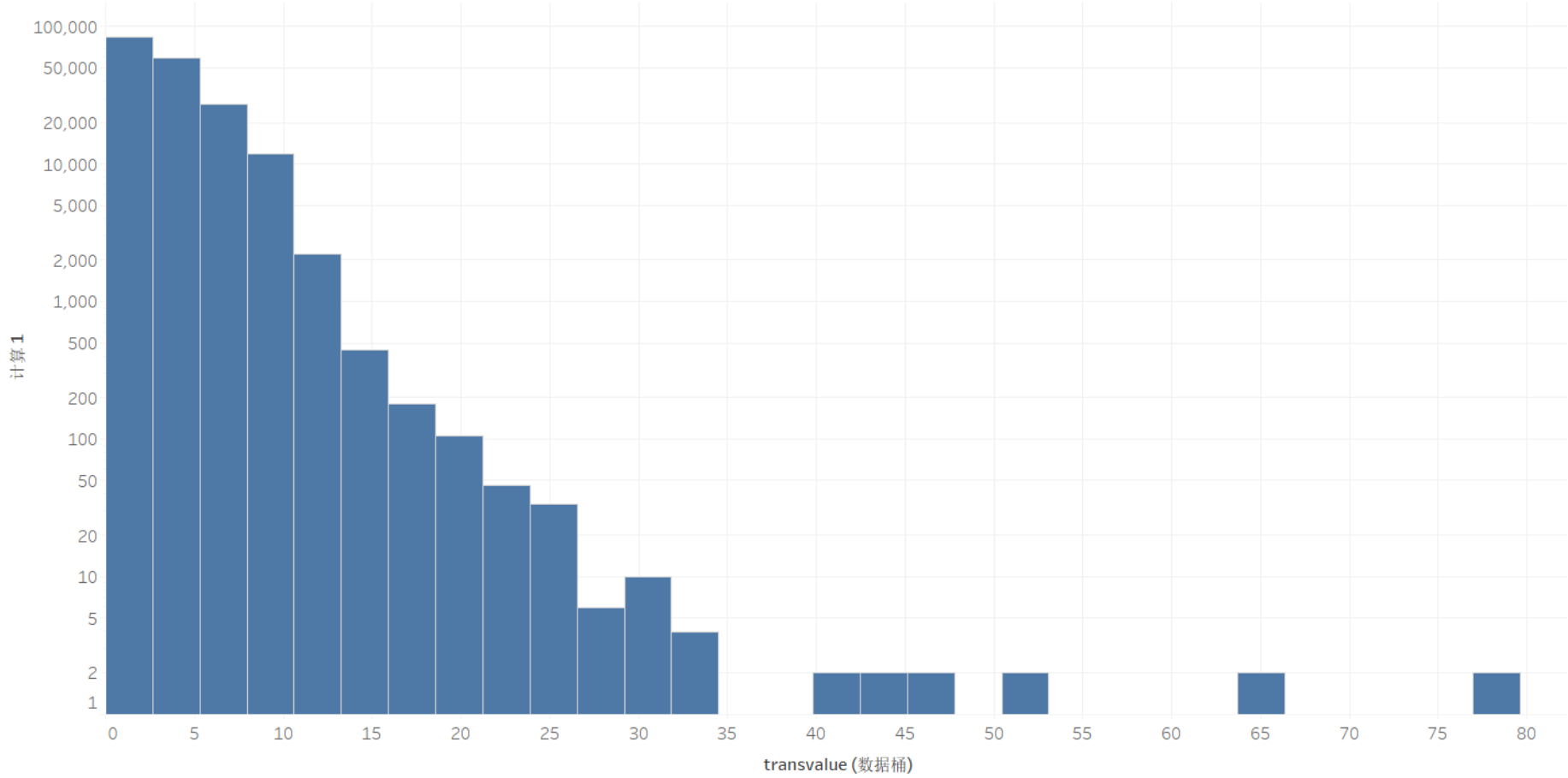
- 数值型变量
  - 金额、次数

一般而言，人员分布极度不均匀

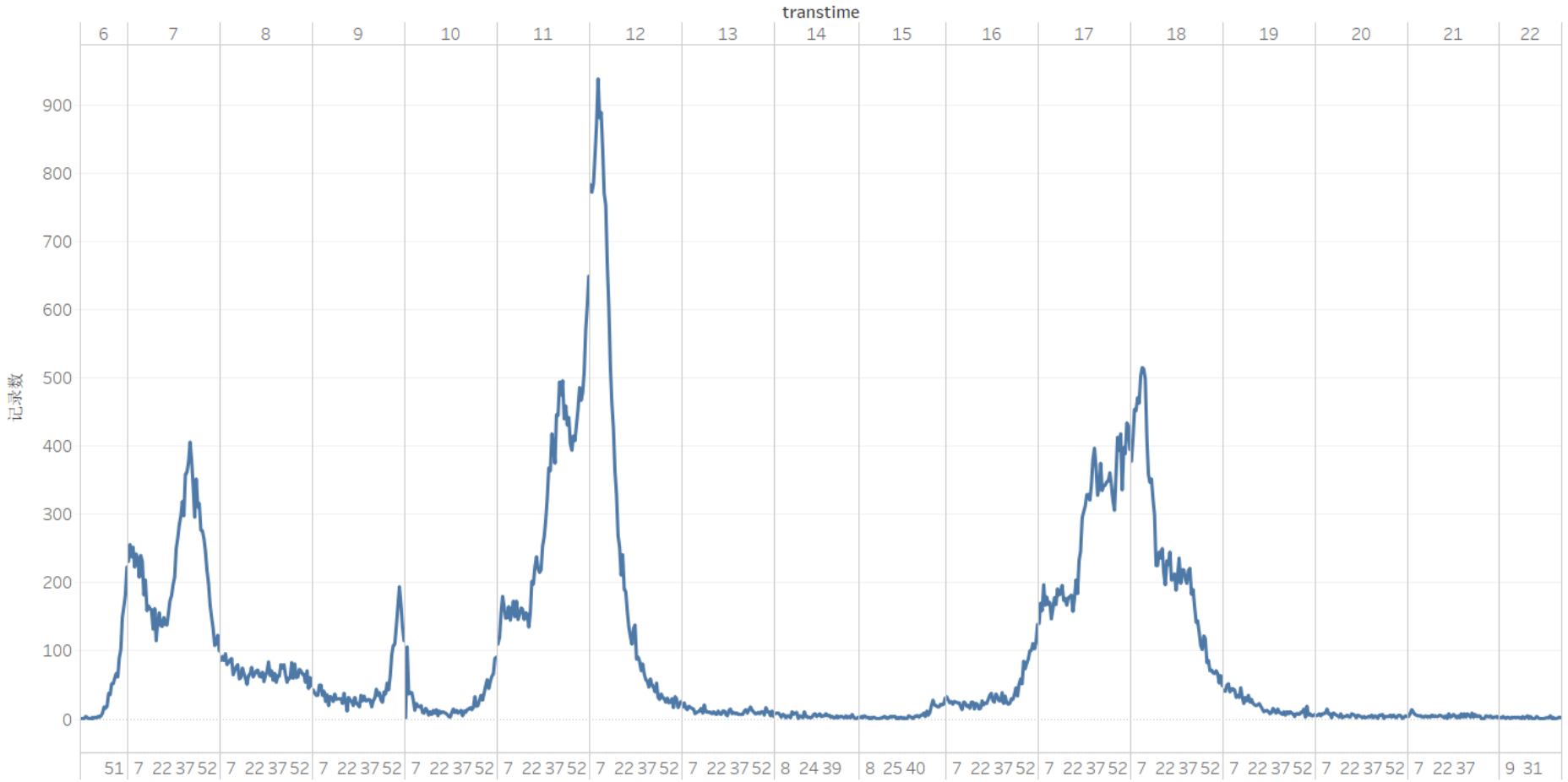


区域分布极度不平衡

交易金额分布极度不均等



周期性规律



自定义地图上的分布展示

- 任意背景作为地图
  - 选择任意图片作为地图
  - 创建两个字段
  - 标记每个点的横纵坐标
  - 绘制散点图

3. 双变量分析

对业务关联的深度探索：**大胆假设、小心求证**

	时间	人	地	事	属性
时间					
人					
地					
事					
属性					

- 以销售数据为例：
- 人-事：什么人擅长卖什么货物
- 事-地：什么地点消费什么货物的数量较大
- 事-时间：货物销量的时间周期（淡旺季分析）
- 人-属性：销售人员的销售额分析
- ……

- 关联规则：
- 时间-时间：在什么时候购买的，还会在什么时候购买
- 地-地：什么地方购买了，什么地方也会购买
- 事-事：购买什么的，还会购买什么
- ……

例如：偏好

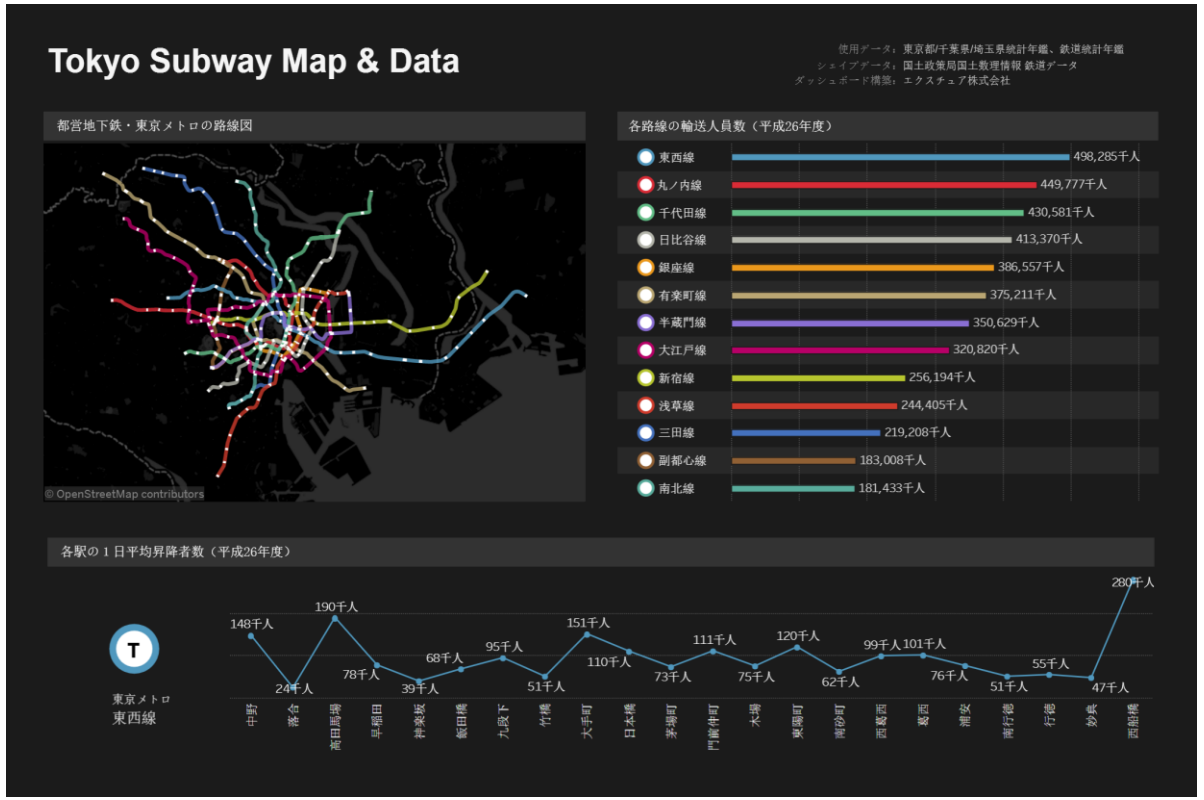
	pos								
stuid	10	11	12	13	14	15	67	68	74
1				276					
4	335								
6	257								
10					374				
13			282						
31									302
40	276								
46				296		391			
49					324				
51	353	304							
70								253	
77							271		
93				272		306			
99			482			249			
100			246		238				

4. 多变量分析

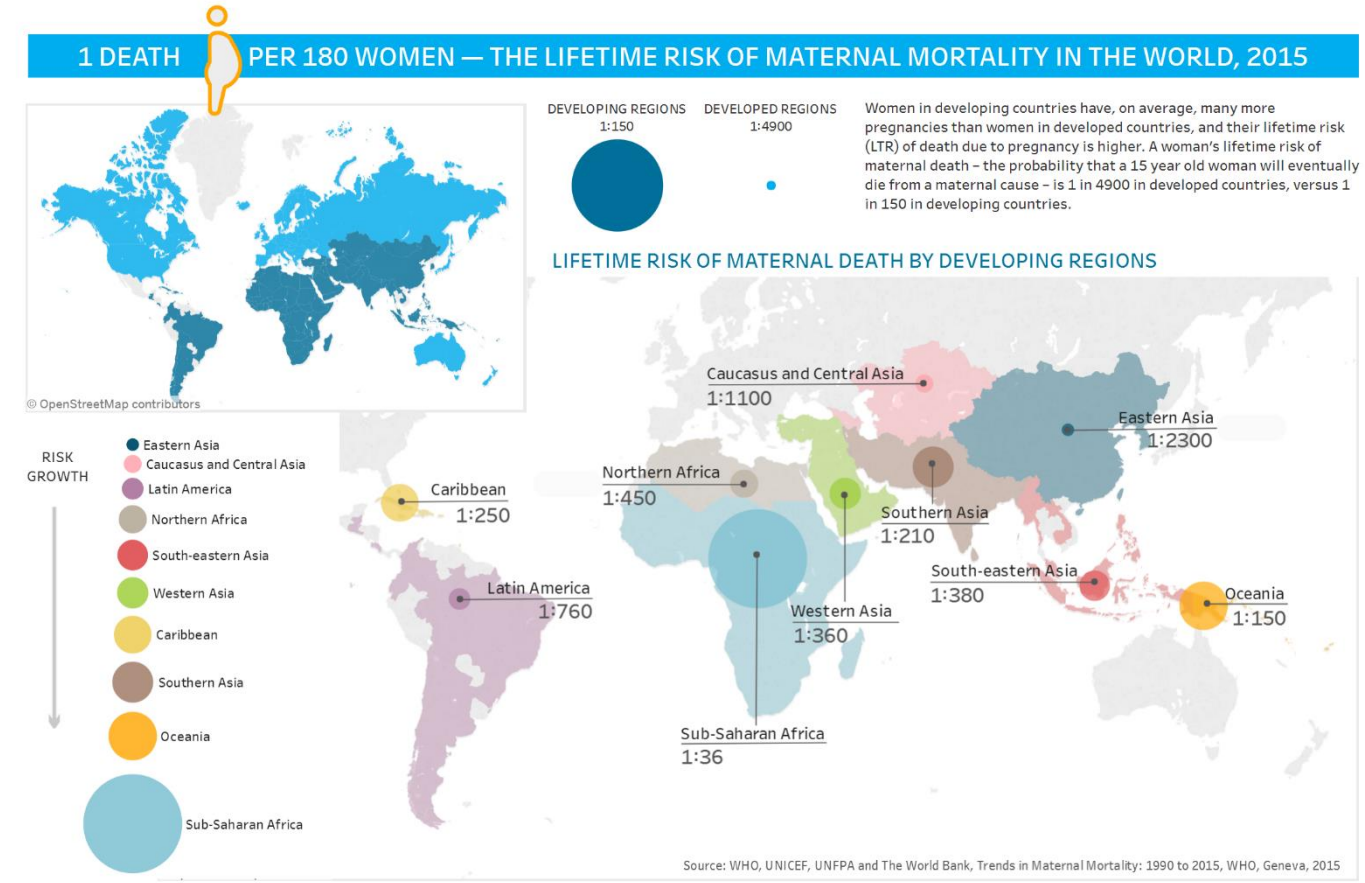
- 在单变量/双变量分析的基础上，增加一个维度
- +时间：类别
- +地：类别
- +人：类别
- +事：类别
- +属性：颜色/大小

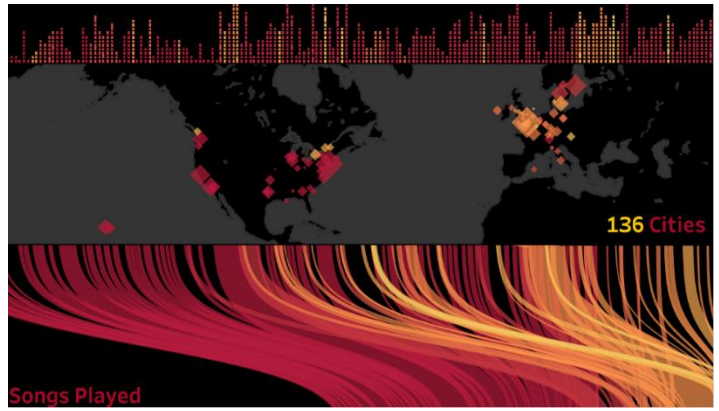
Tableau的可视化 [参考]

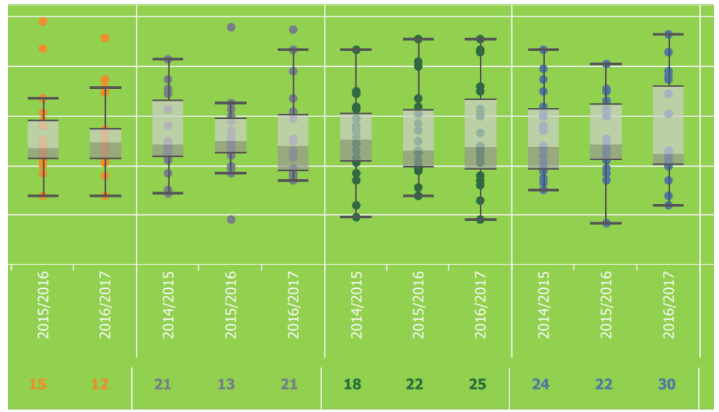
- 更多高级图表 https://public.tableau.com/s/gallery



[参考]

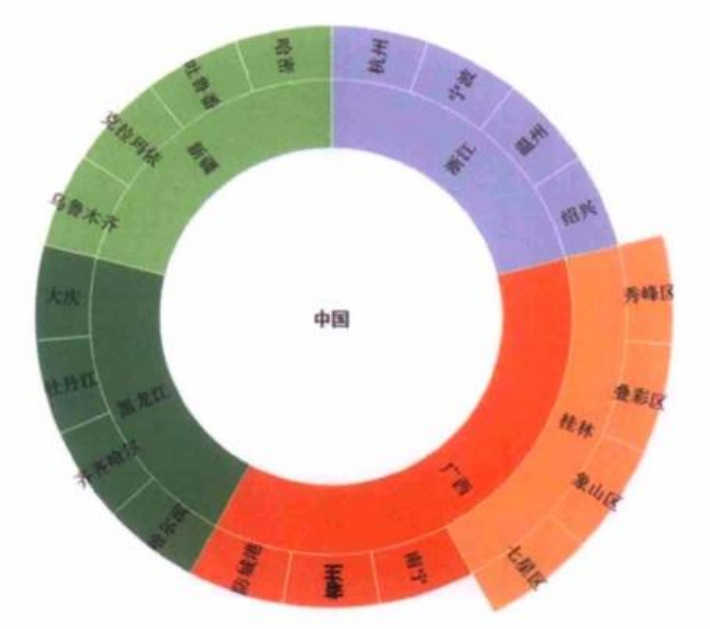


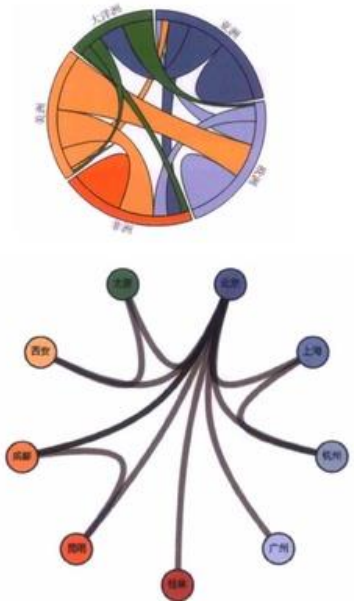


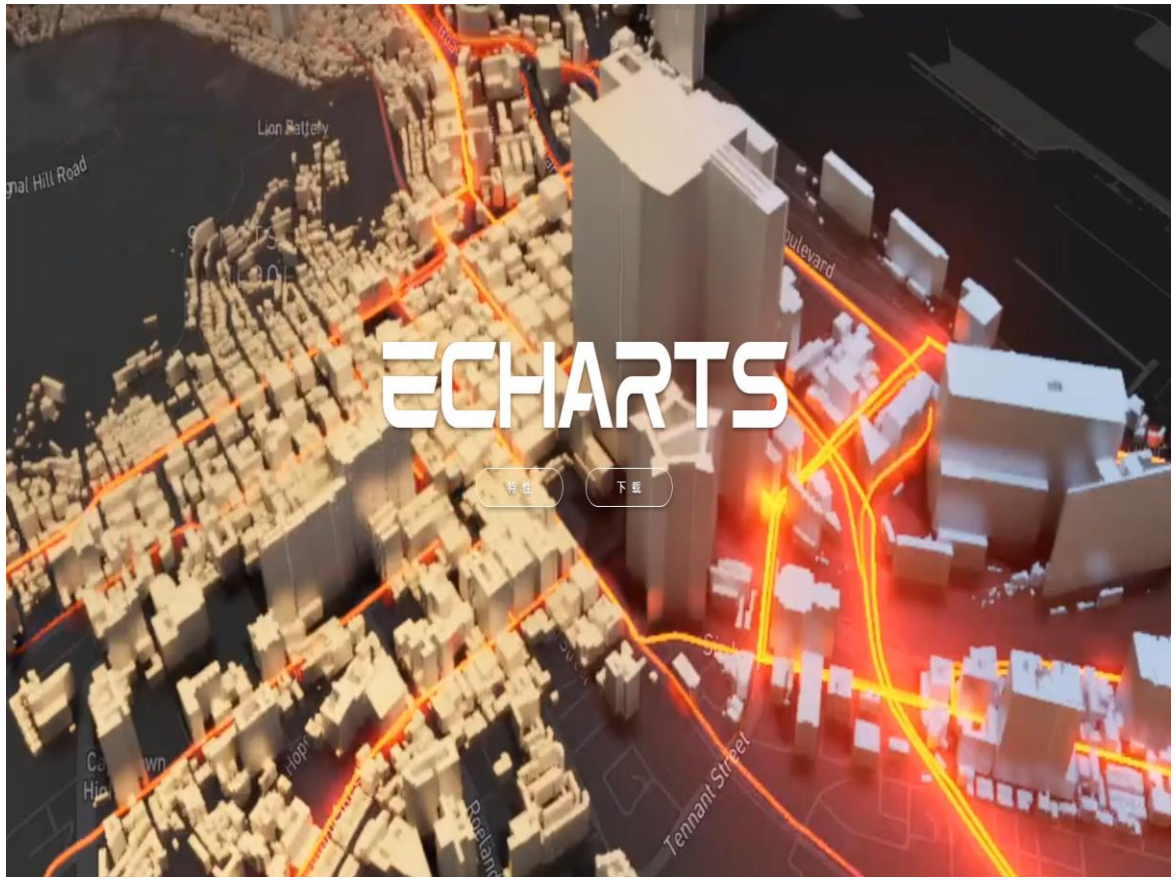


探索性数据分析/可视化工具很多

- BI工具：PowerBI，帆软，其它
- 开源可视化脚本：eCharts，D3
- 编程语言：R，Python







画图的目的是什么？

数据分析！

大数据核心课程

商务大数据分析 >>2. 探索性数据分析

2.4 数据分析的目标

刘跃文 博士

教授、博士生导师

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

V2.2, 2021-9-8

1. 大数据分析的目标

描述现状 → 识别异常 → 发现规律 → 预测未来

价值挖掘：知识发现

知识是大家以前不曾掌握的内容

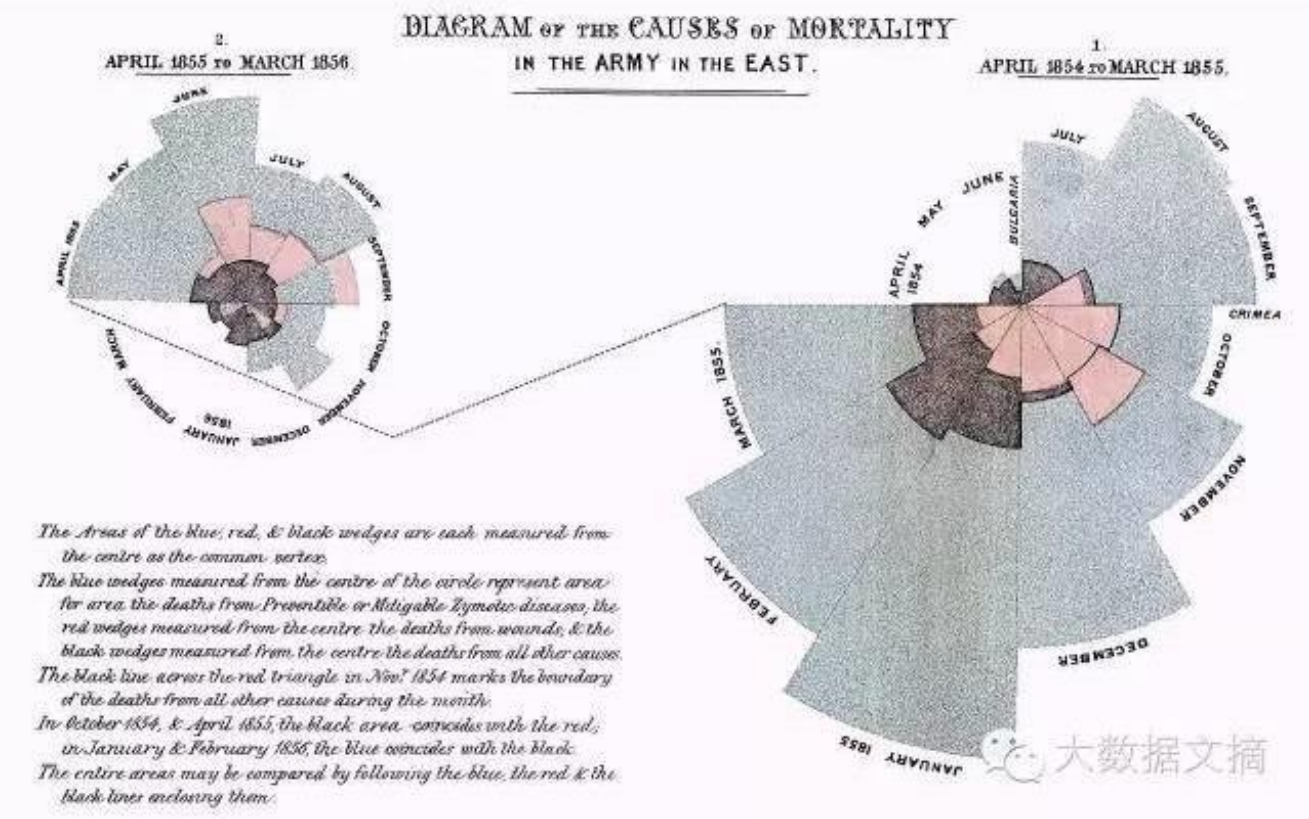
辅助线和文字

Tableau + Power point

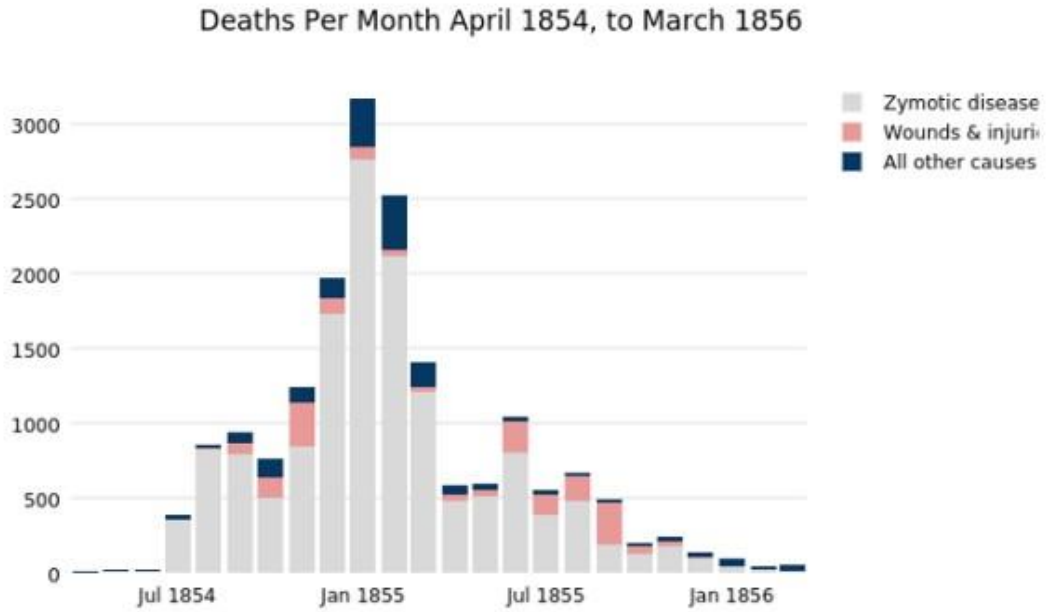
2. 描述现状

- 要描述大家不知道的现状（机会很少）
- 量化描述，而不是定性描述
- 要对现状进行提炼、总结
- 用更直观的方式让大家认知

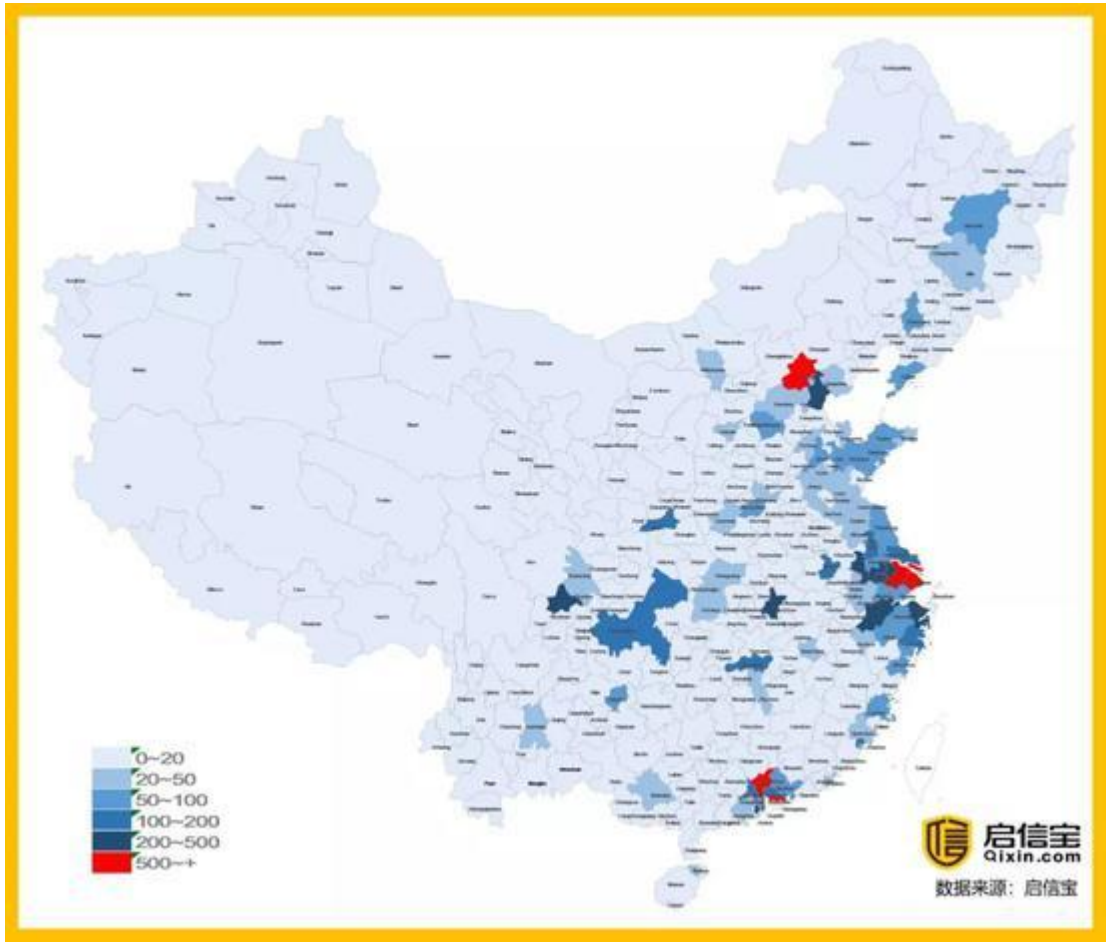
南丁格尔玫瑰图



深入一线：
发现别人未曾注意的现状

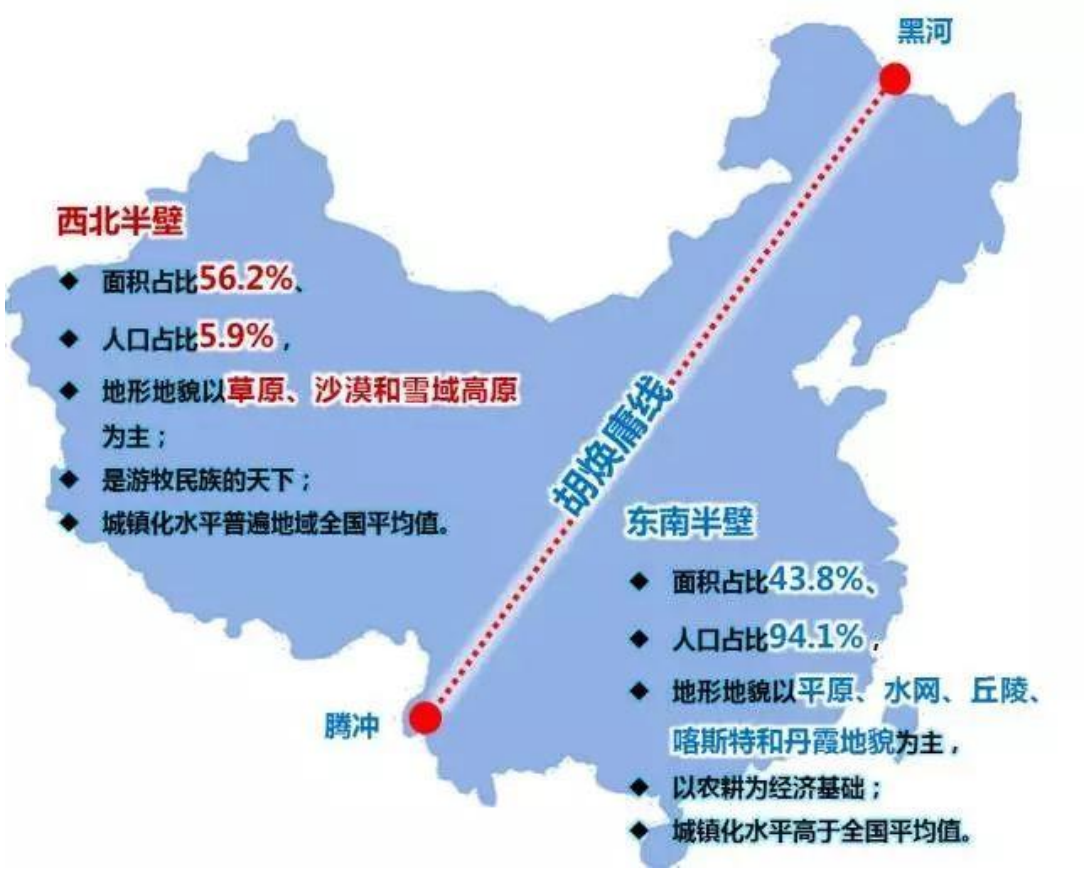


地理分布现状

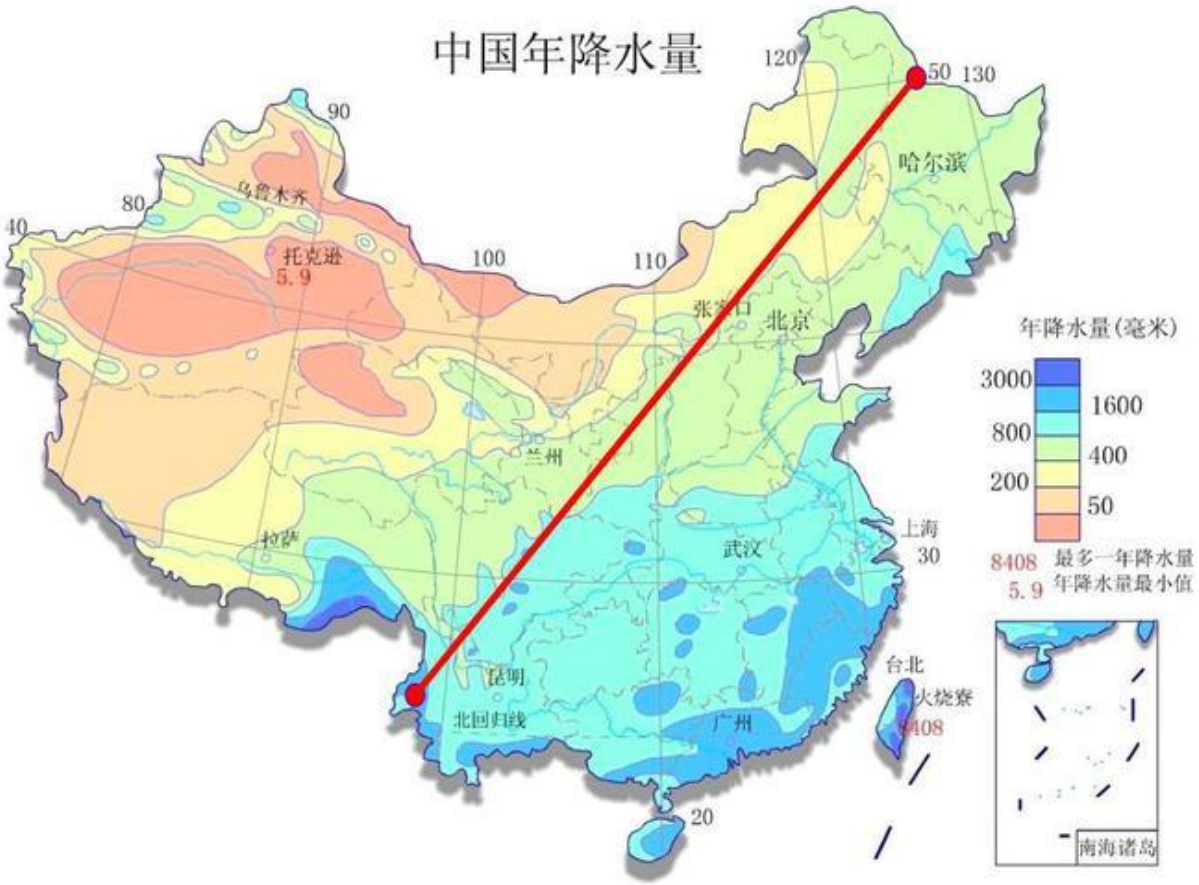


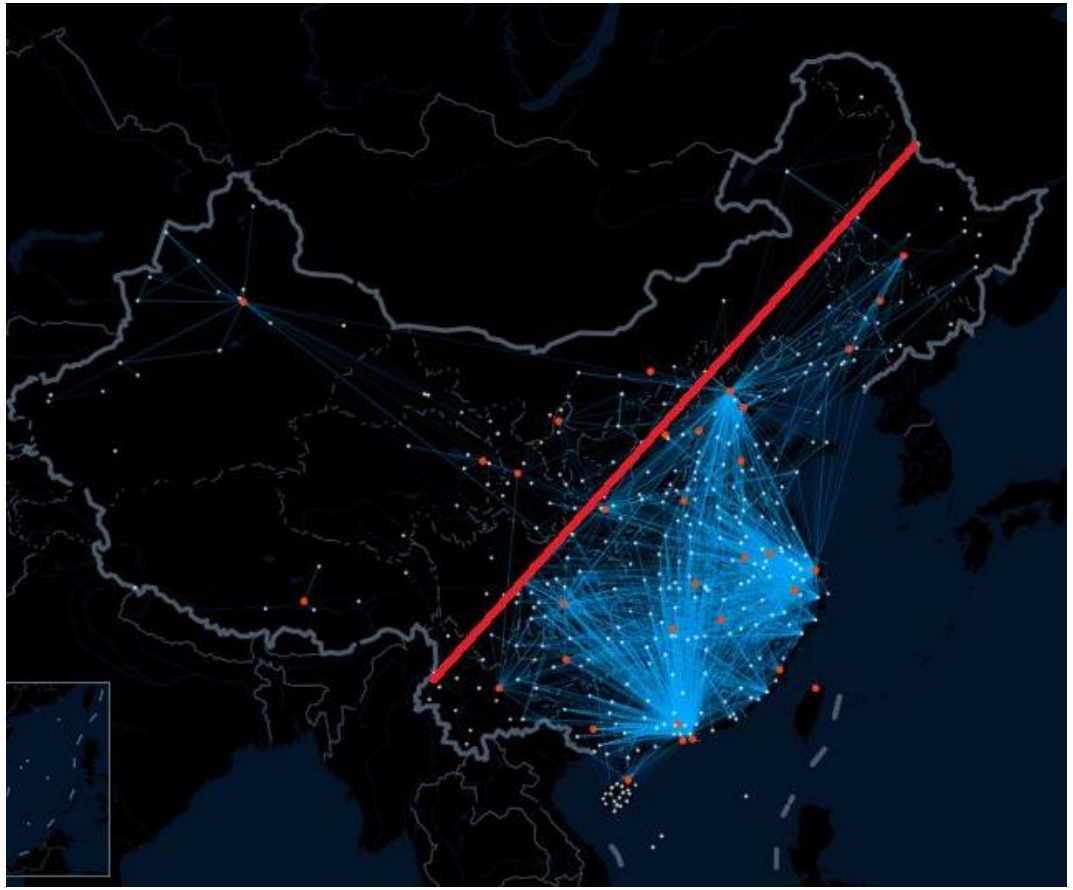
胡焕庸线：量化的价值

- 胡焕庸线主要描述了中国人口密度在不同地区的分布，并由此得出中国第一张人口密度图。这张人口密度图被附在其于1935年发表在《地理学报》上的论文《中国之人口分布》之后。
- 已故经济地理学家、人文地理学家、中科院院士吴传钧曾这样回忆他的老师：“当时中国总人口估计有4.75亿，他（胡焕庸）以1点表示2万人，根据掌握实际情况将2万多个点子落实到地图上，再以等值线画出人口密度图。”
- 多年后，美国学者将之称为“胡焕庸线”。
- 自古以来，中国东南地狭人稠、西北地广人稀似乎早成事实，但没有人对这种模糊的认识加以有力的佐证。瑷珲ài huī—腾冲线的出现则廓清了这一分界，影响深远，成为研究和决策的重要参考依据。



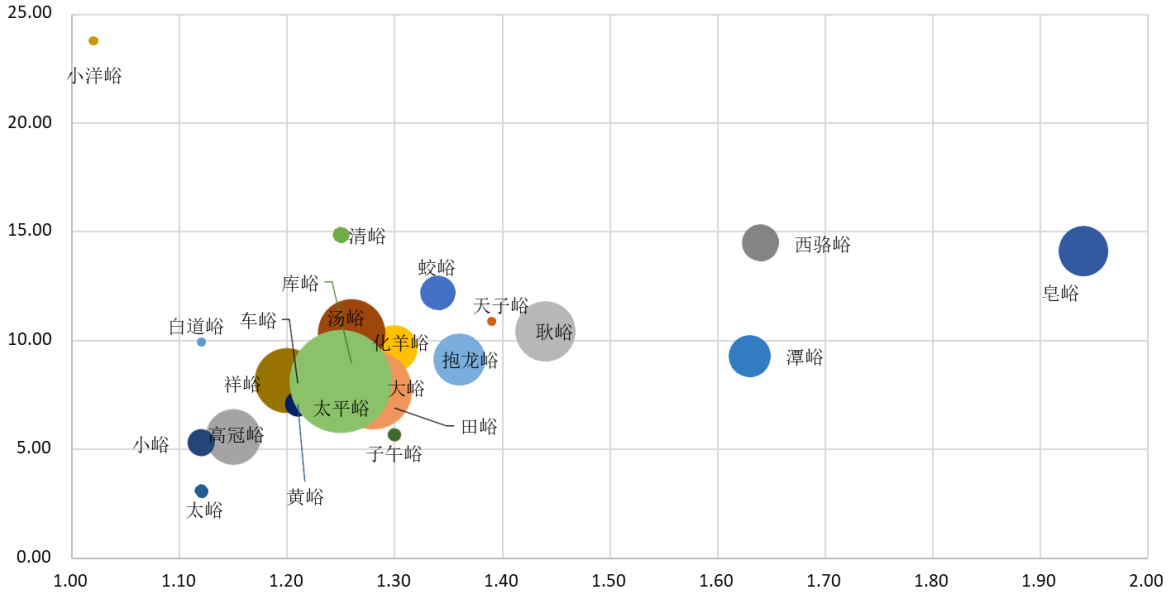
情况是不是变了？

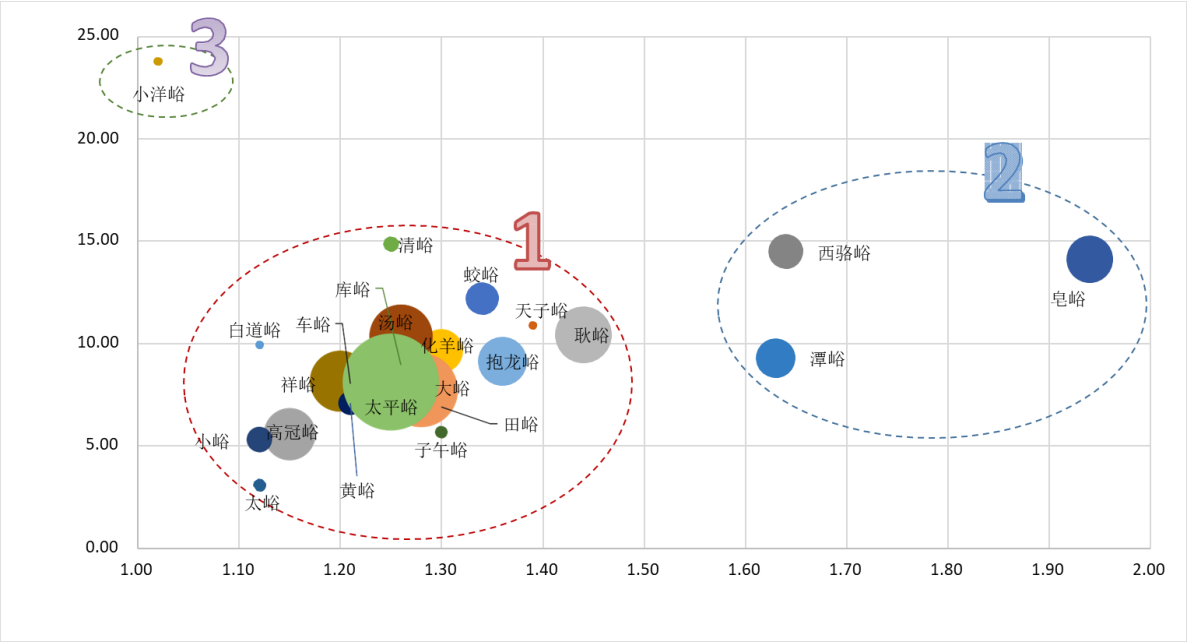




对结果分类总结

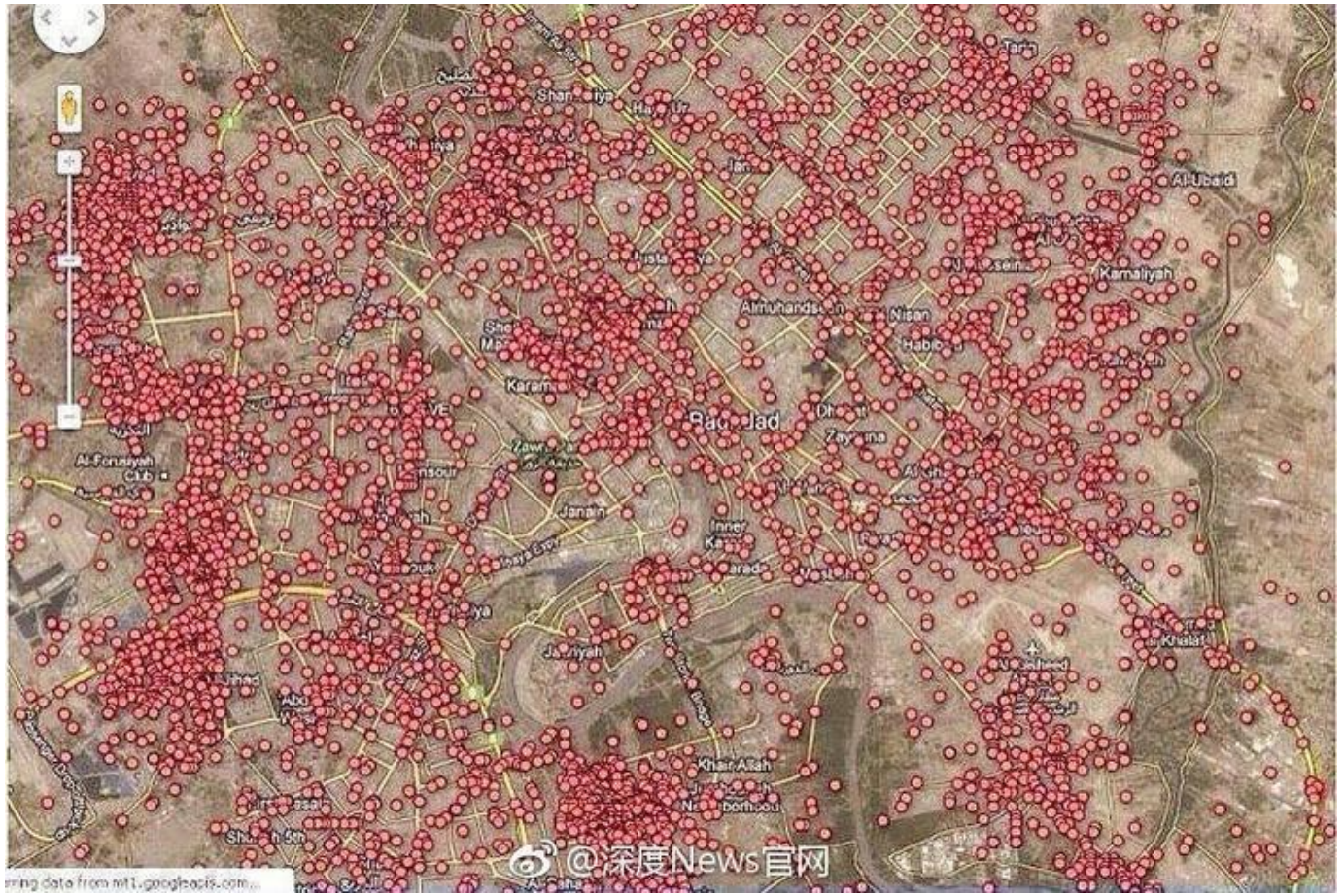
- 对结果分类是一种很好的总结和提炼方式。





更有冲击力的展示：伤亡地图

- 2010年10月23日《卫报》利用维基解密的数据做了一篇“数据新闻”。
- 将伊拉克战争中所有的人员伤亡情况均标注于地图之上。地图上一个红点便代表一次死伤事件，鼠标点击红点后弹出的窗口则有详细的说明：伤亡人数、时间，造成伤亡的具体原因。
- 密布的红点多达39万，显得格外触目惊心。一经刊出立即引起朝野震动，推动英国最终做出撤出驻伊拉克军队的决定。



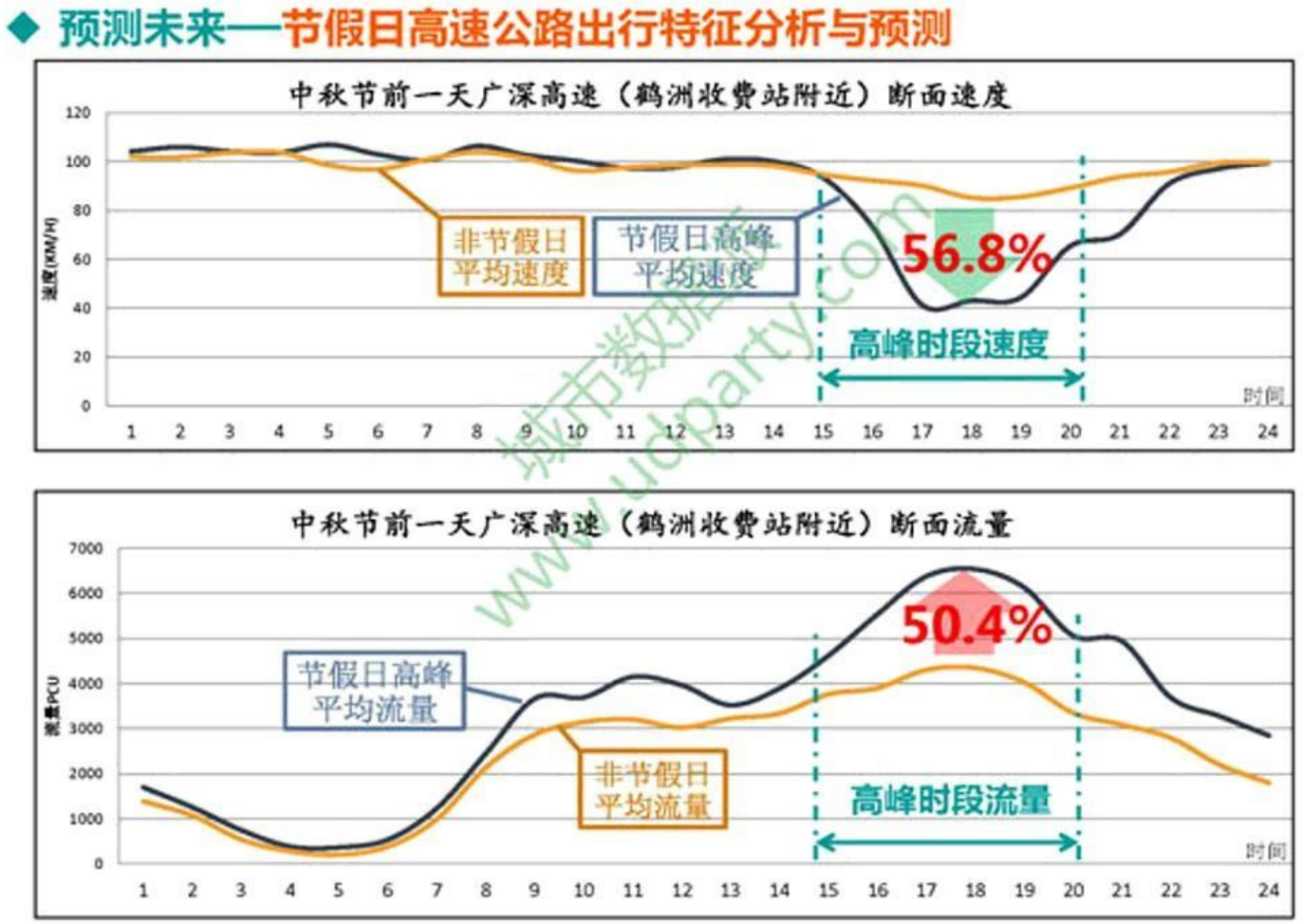
3. 识别异常

- 异常是相对于正常的

- 寻找参考系：
  - 绝对值
  - 历史值
  - 横向比较值
  - 其它数据体系（对账）
  - **理论**

量化异常

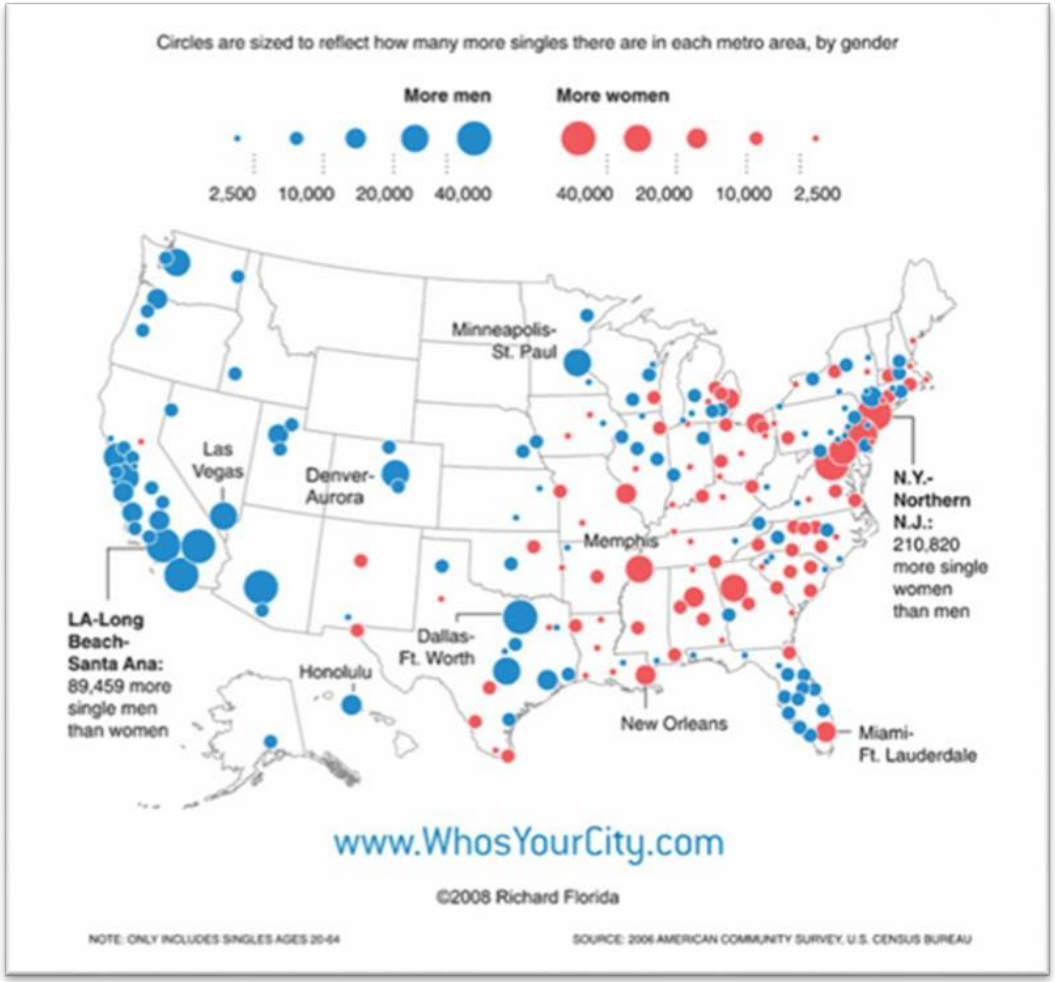
- 从量化中获取更加深刻和准确的认知。
- 量化本身就是创造知识。
- ~~差不多先生/女士~~：差不多、大概、应该、比较好、比较差……



4. 规律识别

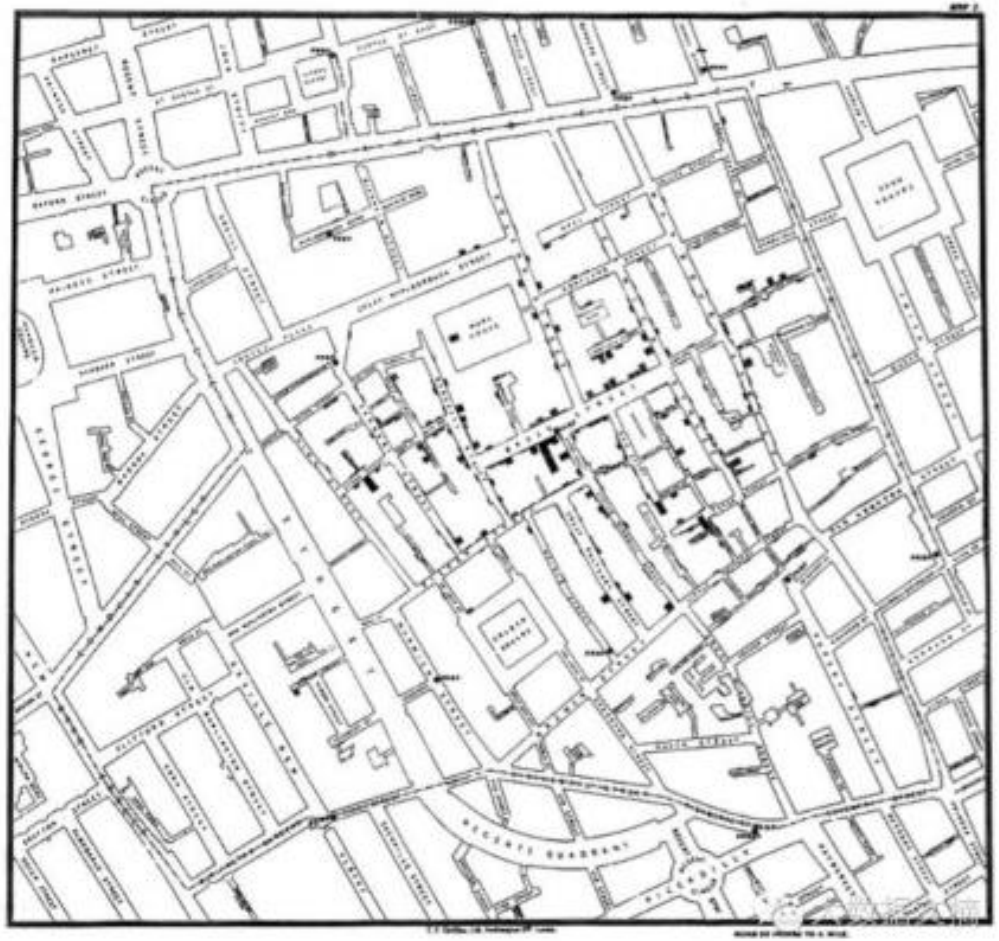
- 什么是规律？
  - Hypothesis / 假设 / 理论。
  - 具有一定的预测性。
  - 如果XXX，那么XXX。

- 规律三问：
  - 是什么？
  - 为什么？
  - 如何利用规律应对？



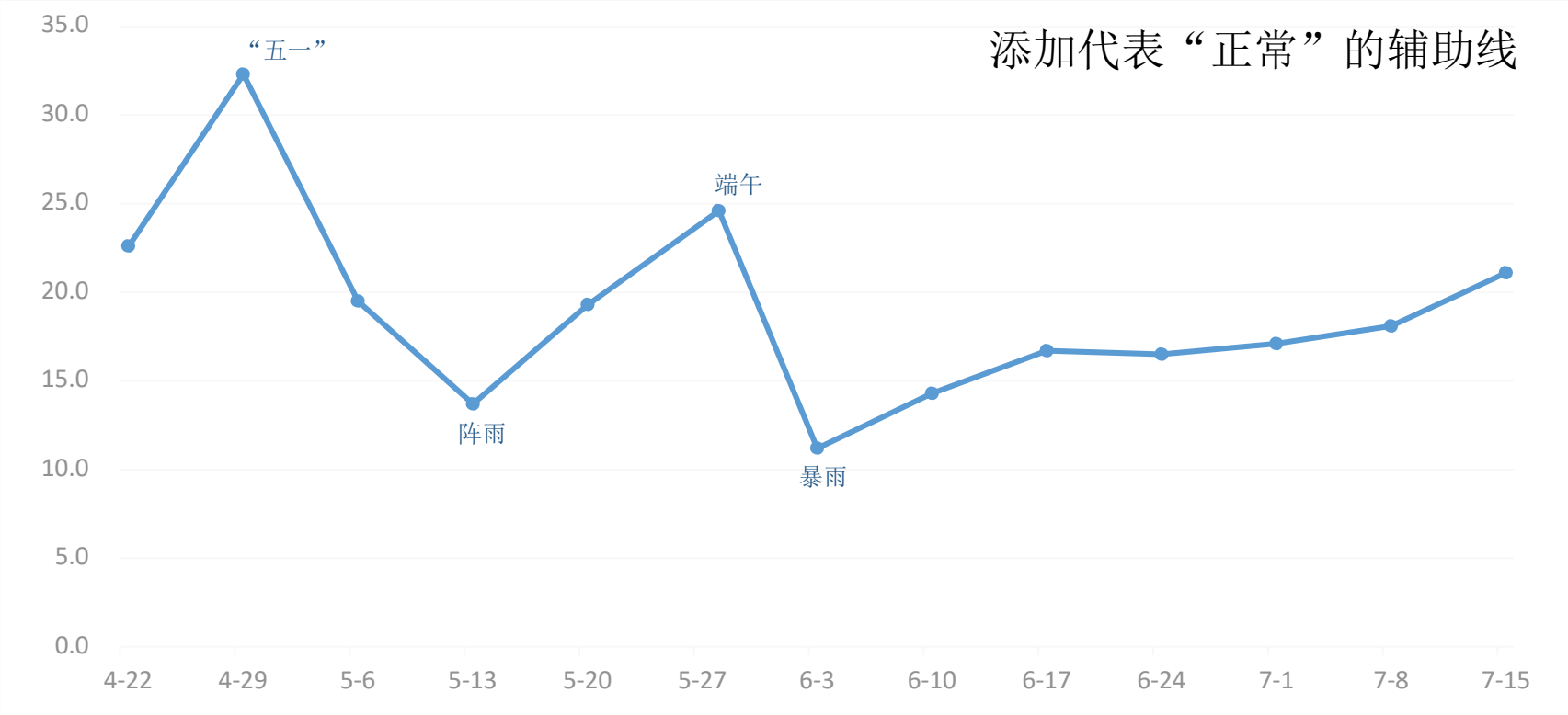
空间分布的规律

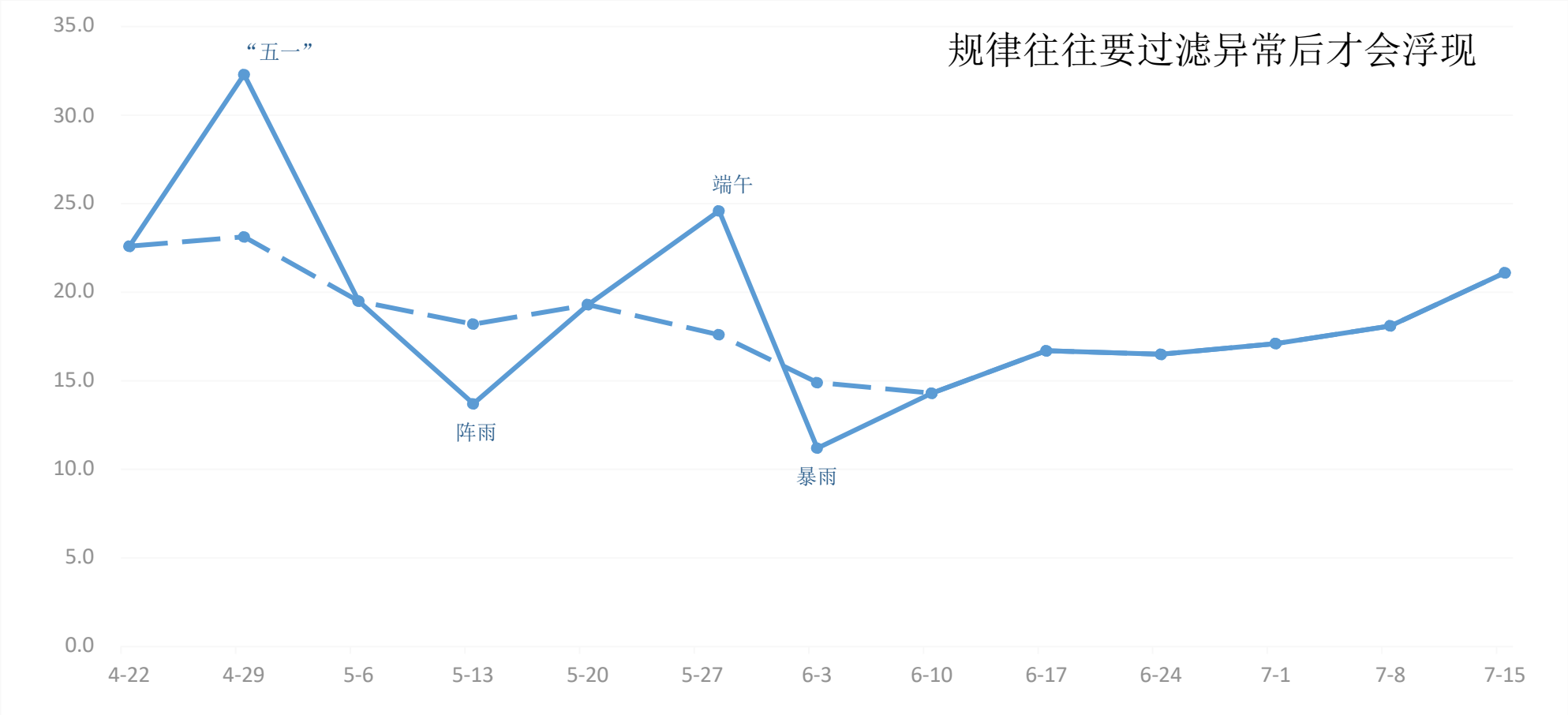
- 约翰·雪诺的地图，展示了1854年伦敦霍乱爆发时的发病源头。线条代表街道。黑色的长条代表了所在街区死亡的人数。圆点代表抽水泵。特别注意在宽街(Broad Street)上的抽水泵周围的死亡人数相对集中。雪诺用他的这幅地图佐证了他极富争议的理论：霍乱是由被污染的饮用水传播开来的。当政府关闭了宽街上的水泵，霍乱的蔓延也平息了。引发霍乱的病菌最终由德国物理学家罗伯特·科赫(Robert Koch)在1883年分离出来。



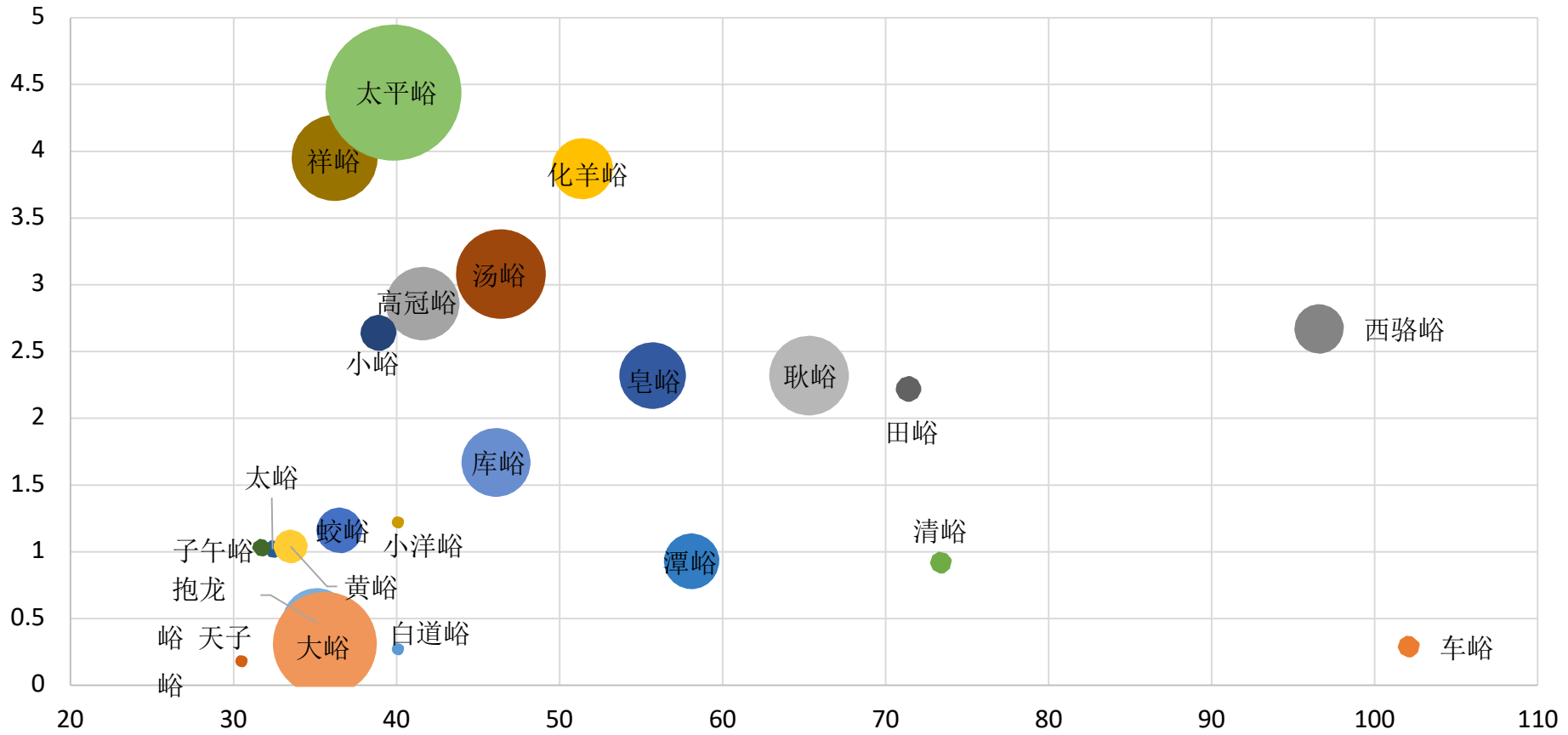
1854年伦敦霍乱

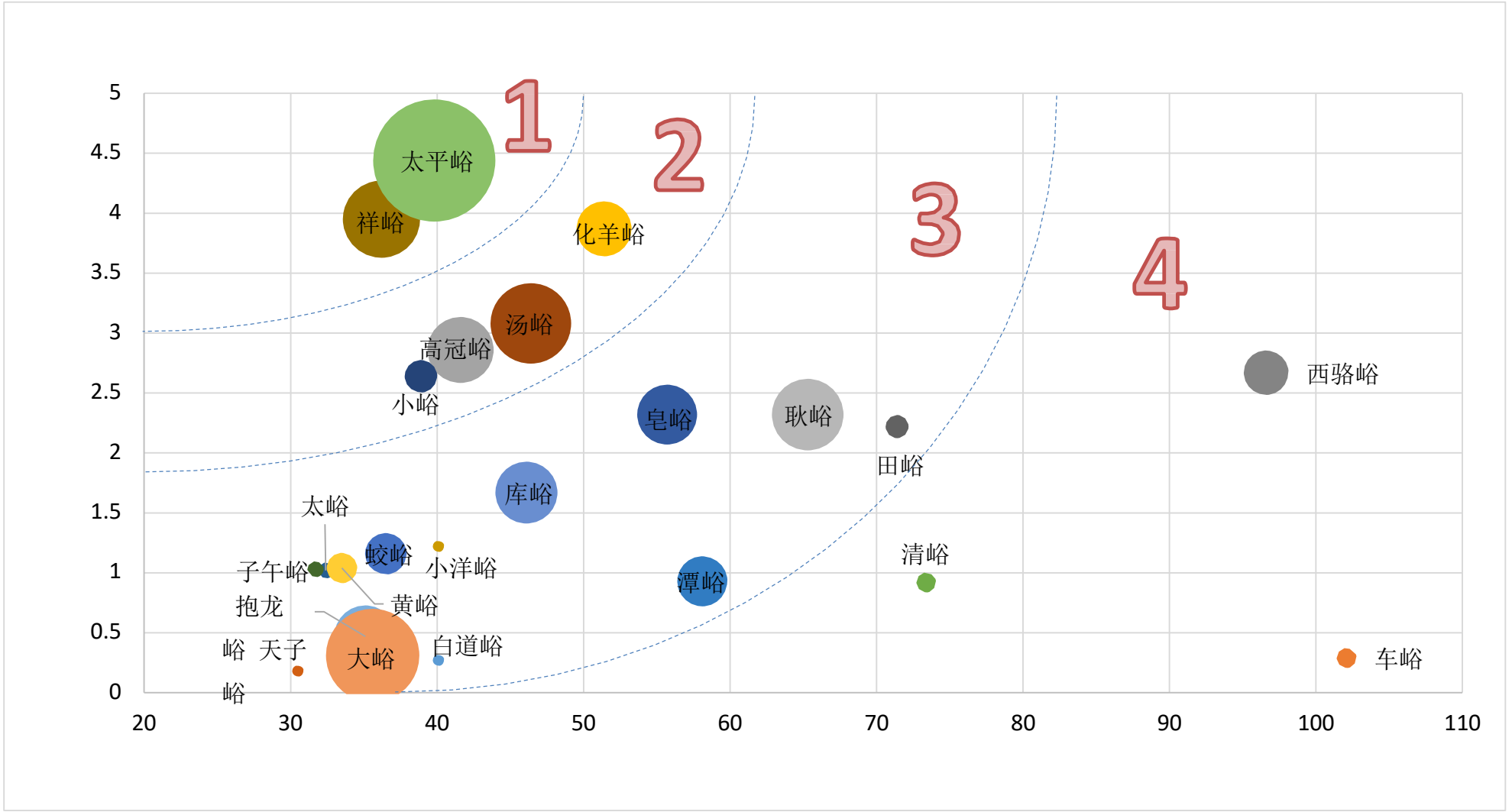
时间发展的规律





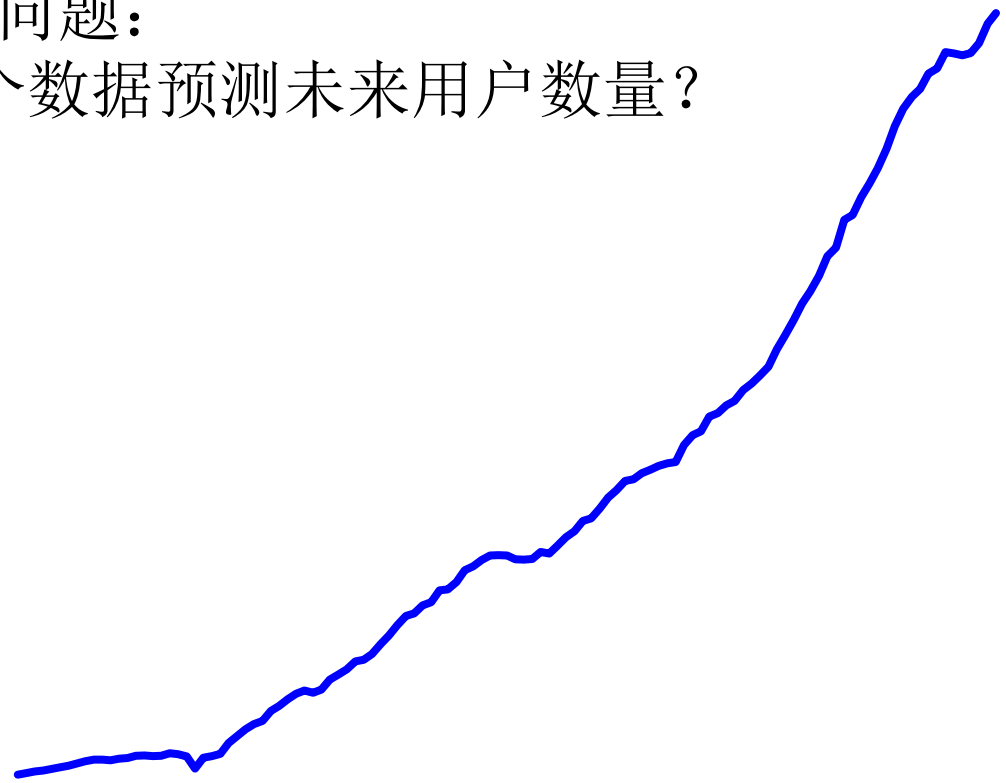
逻辑关系的规律



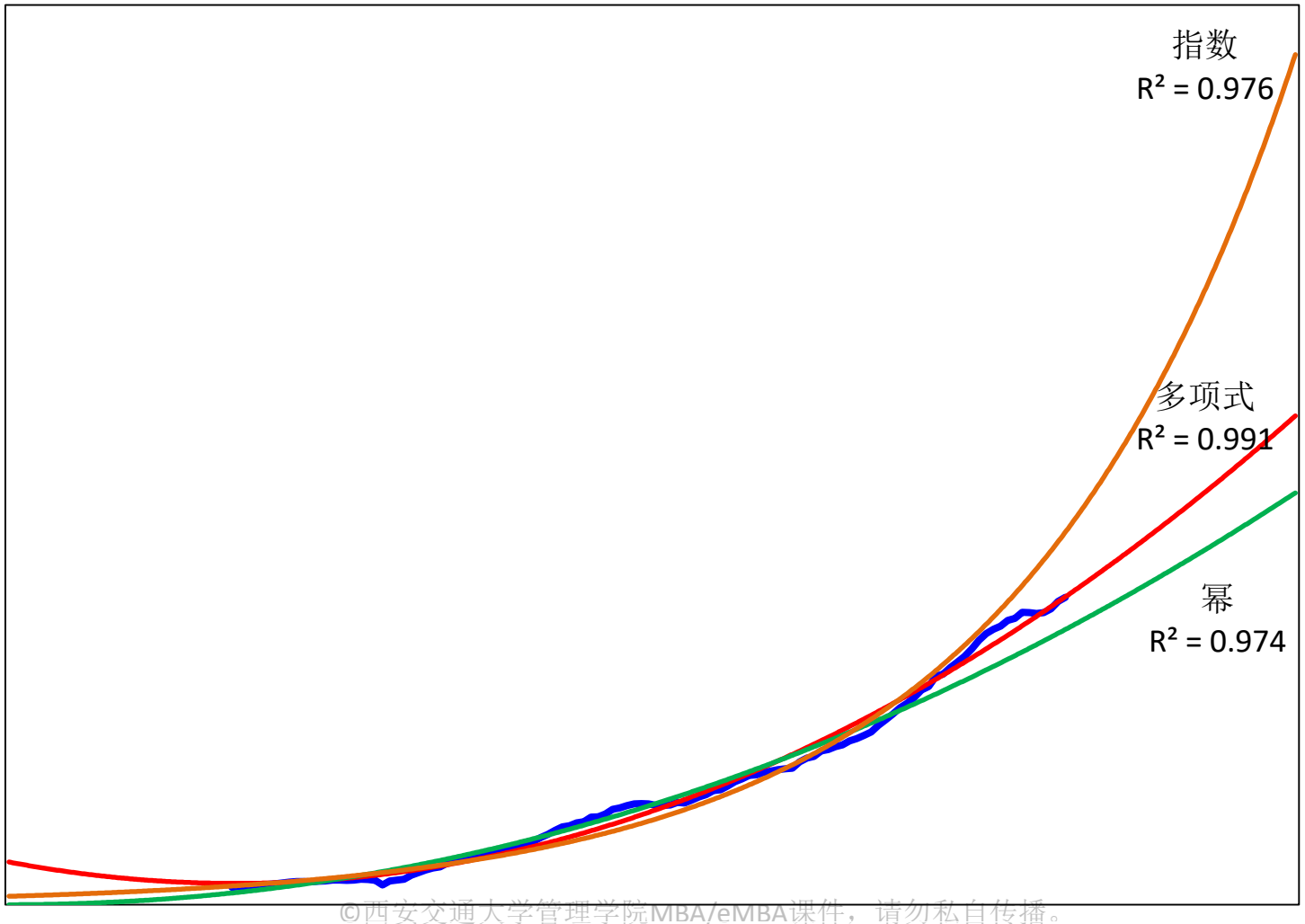


5. 案例：MAU预测

提出问题：
140个数据预测未来用户数量？

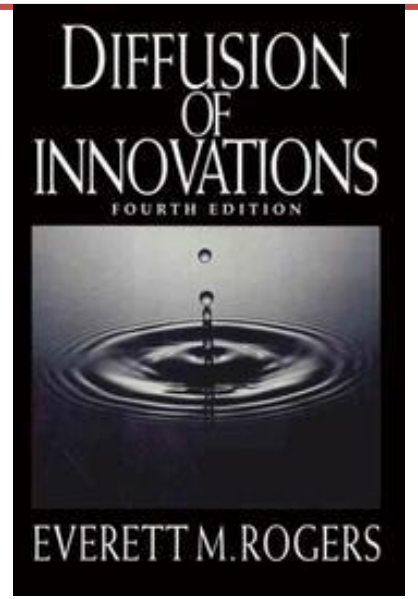


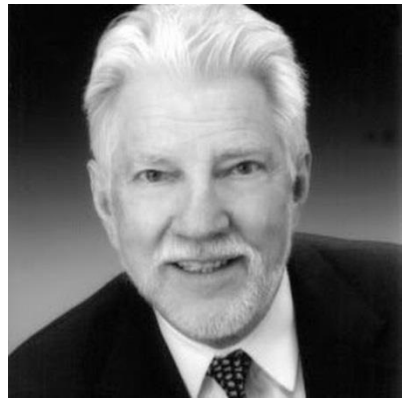
能否通过曲线拟合得到“预测”值？



（1）创新扩散理论

- Everett M. Rogers
  - March 6, 1931 – October 21, 2004
- Rogers, E.M. (1995). Diffusion of innovations (4th edition). The Free Press. New York.
  - Published 5th edition in 2003
  - Cited 48280 times, counted by Google Scholar (2013/5/5)
  - “As per the Social Science Citation Index, Diffusion of Innovations is the second most cited book in the social sciences.” (Stacks & Salwen 2009)





Innovation Diffusion 创新扩散

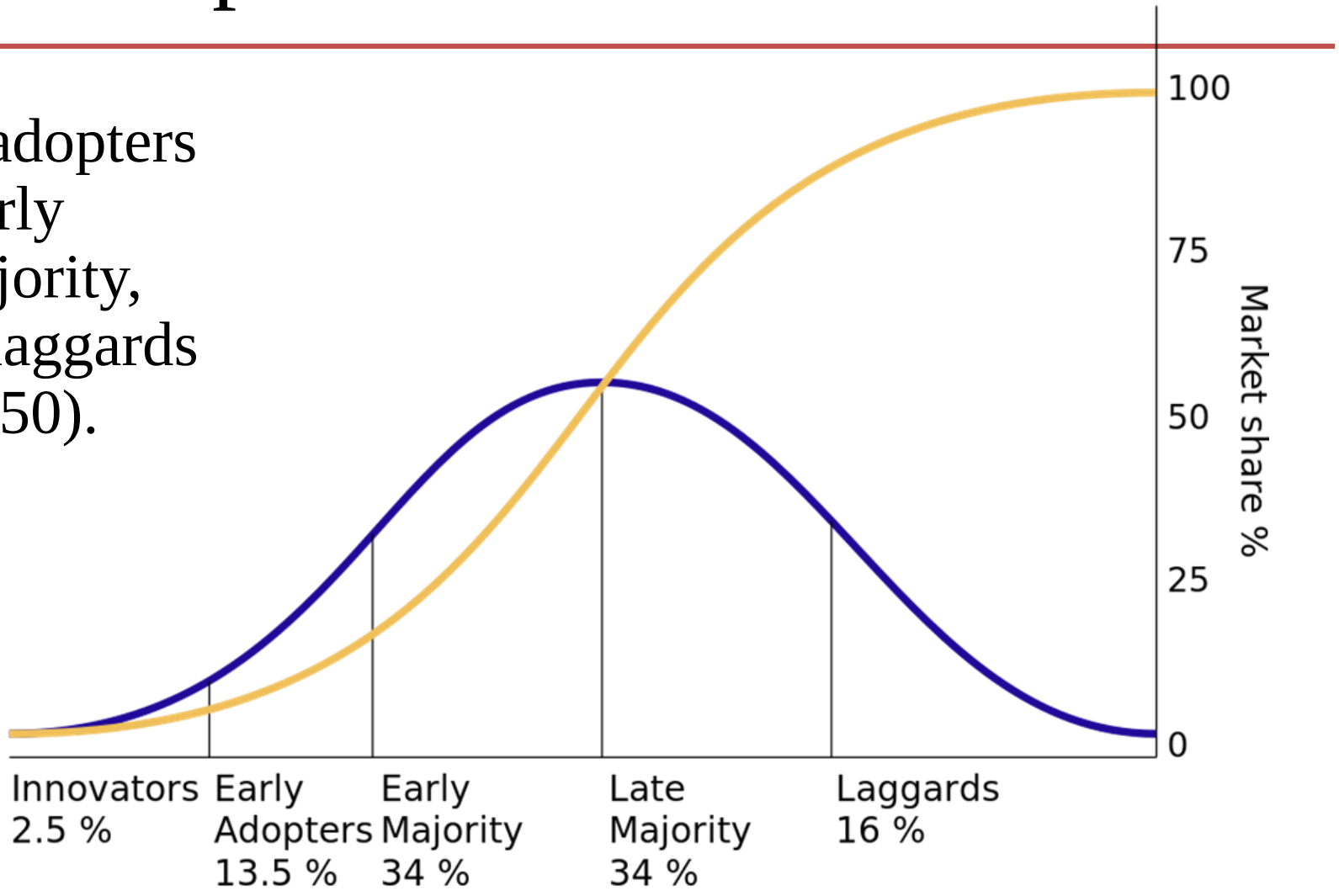
- Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system.

- Rogers (1962) espoused the theory that there are four main elements that influence the spread of a new idea: the innovation, communication channels, time, and a social system.
- This process relies heavily on human capital.
- The innovation must be widely adopted in order to self-sustain.

- Diffusion of Innovations manifests itself in different ways in various cultures and fields and is highly subjective to the type of adopters and innovation-decision process.

Categories of Adopters

- The categories of adopters are: innovators, early adopters, early majority, late majority, and laggards (Rogers 1962, p. 150).



Adopter category	Definition
Innovators	Innovators are the first individuals to adopt an innovation. Innovators are willing to take risks, youngest in age, have the highest social class, have great financial liquidity, are very social and have closest contact to scientific sources and interaction with other innovators. Risk tolerance has them adopting technologies which may ultimately fail. Financial resources help absorb these failures. (Rogers 1962 5th ed, p. 282)
Early adopters	This is the second fastest category of individuals who adopt an innovation. These individuals have the highest degree of opinion leadership among the other adopter categories. Early adopters are typically younger in age, have a higher social status, have more financial lucidity, advanced education, and are more socially forward than late adopters. More discrete in adoption choices than innovators. Realize judicious choice of adoption will help them maintain central communication position (Rogers 1962 5th ed, p. 283).
Early Majority	Individuals in this category adopt an innovation after a varying degree of time. This time of adoption is significantly longer than the innovators and early adopters. Early Majority tend to be slower in the adoption process, have above average social status, contact with early adopters, and seldom hold positions of opinion leadership in a system (Rogers 1962 5th ed, p. 283)
Late Majority	Individuals in this category will adopt an innovation after the average member of the society. These individuals approach an innovation with a high degree of skepticism and after the majority of society has adopted the innovation. Late Majority are typically skeptical about an innovation, have below average social status, very little financial lucidity, in contact with others in late majority and early majority, very little opinion leadership.
Laggards	Individuals in this category are the last to adopt an innovation. Unlike some of the previous categories, individuals in this category show little to no opinion leadership. These individuals typically have an aversion to change-agents and tend to be advanced in age. Laggards typically tend to be focused on "traditions", likely to have lowest social status, lowest financial fluidity, be oldest of all other adopters, in contact with only family and close friends.

创新扩散理论的价值？

- 识别潜在客户。

（2）创新扩散模型—Bass模型

- Frank M. Bass
  - Bass 1969 "A New Product Growth for Model Consumer Durables." Management Science
  - INFORMS members have voted the “Bass Model” paper as one of the Top 10 Most Influential Papers published in the 50-year history of Management Science in connection with the 50th anniversary of the journal. (Bass 2004)
  - Bass, F. M. (2004). "A New Product Growth for Model Consumer Durables." Management Science 50: 1825-1832.
  - Bass, F. M. (2004). "Comments on "A New Product Growth for Model Consumer Durables."." Management Science 50: 1833-1840.

1926-2006

- 创新系数 模仿系数
  - **P(T)=p+(q/m)Y(T)**

- **P(T)=p+(q/m)Y(T)** , where p and q/m are constants and Y(T) is the number of previous buyers.
- "Fortunately, there exists a closed-form solution to the differential equation in the time domain. If the coefficient of imitation is greater than the coefficient of innovation the solution rises to a peak and then declines."

$$f(T) = ((p+q)^2/p)[e^{-(p+q)T}/(q/pe^{-(p+q)T}+1)^2],$$

and

$$S(T) = (m(p+q)^2/p)[e^{-(p+q)T}/(q/pe^{-(p+q)T}+1)^2].$$

Figure 4 Actual Sales and Sales Predicted by Regression Equation

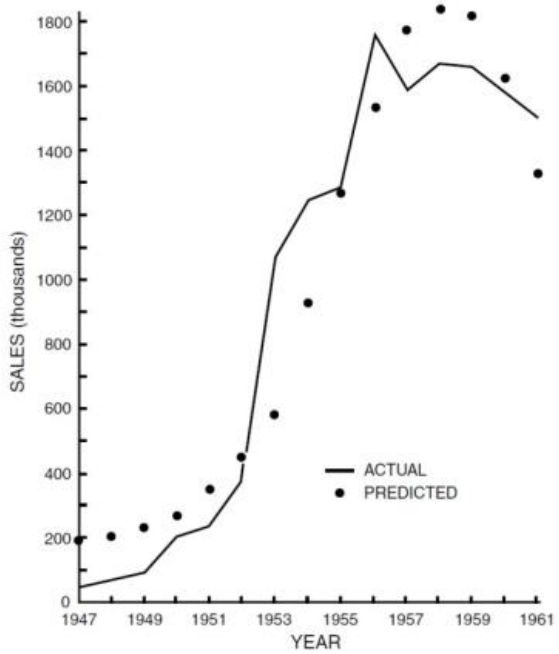
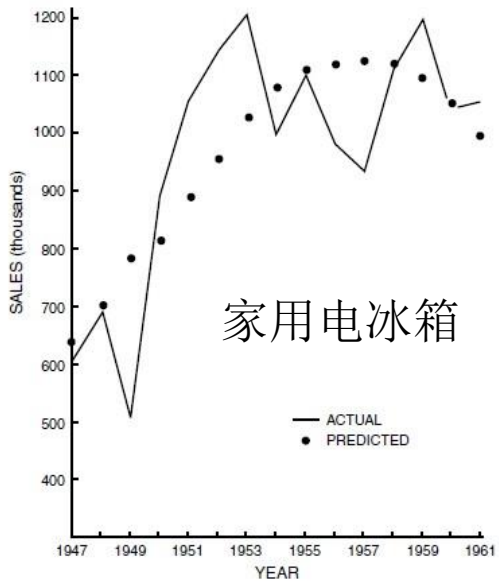
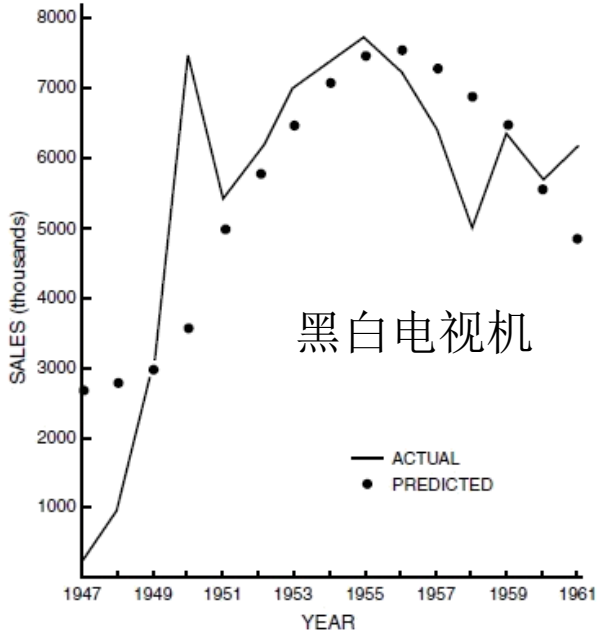


Figure 5 Actual Sales and Sales Predicted by Regression Equation (Home Freezers)



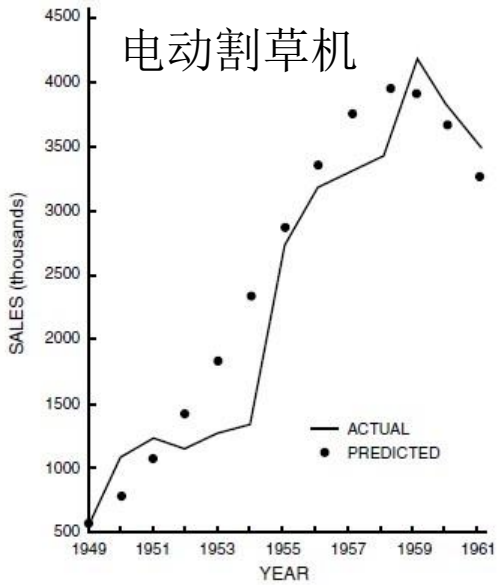
家用电冰箱

Figure 6 Actual Sales and Sales Predicted by Regression Equation (Black & White Television)



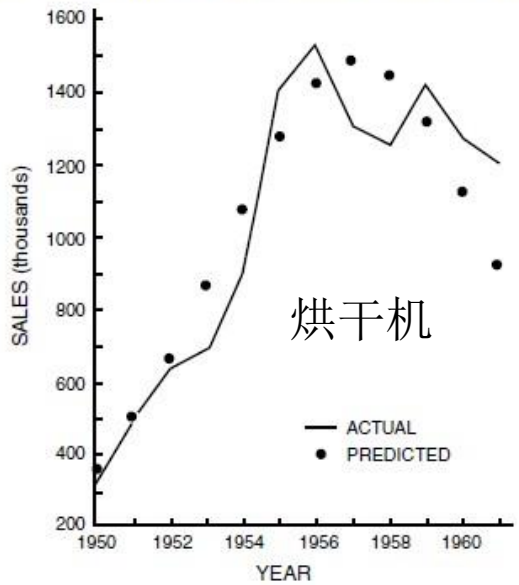
黑白电视机

Figure 7 Actual Sales and Sales Predicted by Model (Power Lawnmowers)



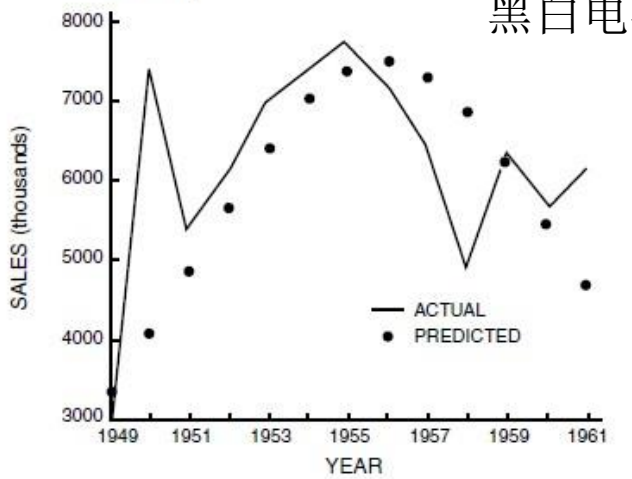
电动割草机

Figure 8 Actual Sales and Sales Predicted by Model (Clothes Dryers)



烘干机

Figure 9 Actual Sales and Sales Predicted by Model (Black & White Television)

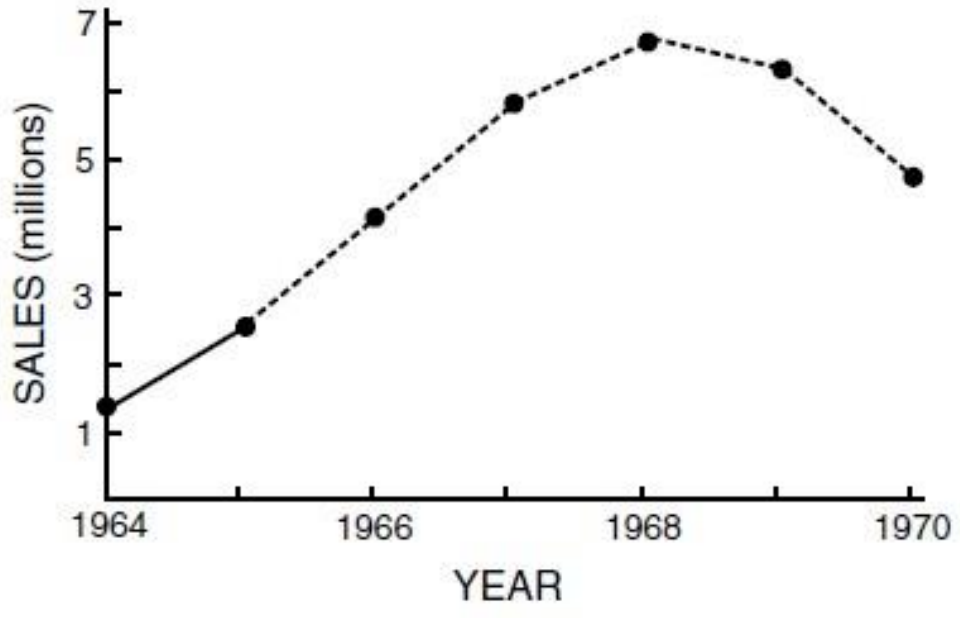


黑白电视机（预测）

- “彩色电视机销量到1968年达到顶峰，约670万台。”(Bass 2004 Comment)

Sales (Millions of Units)	Year
0.7	1963
1.35	1964
2.50	1965

Figure 10 Projected Sales—Color Television



沉迷偷菜游戏 小心“偷”出病来 危害身心健康

南方日报 来源：南方日报 作者：欧旭江 2010年10月19日10:55

能够保持对网络等新事物的兴趣本来是件好事，但如果沉迷其中的话，必然会危害身心健康

哈尔滨市的田女士今年63岁，自从去年冬天开始迷上了上网“偷菜”，整天坐在电脑前盯着自己的“地”，窥视别人的“田”。后来，田女士突然出现头疼、头晕等症状，家人带她到医院检查，她竟患上脑动脉硬化。医生表示，这与她缺乏运动、上网过度、超负荷用脑有很大关系。

“偷菜游戏”自从在网络上出现后，立刻吸引了众多网友加入偷菜行列中来，而“今天你偷菜了吗?”甚至成为人们的问候语，其火爆程度可见一斑。而随着年轻网民的热情退去，目前这种简单的游戏又在老年网民中流行。他们有的为从别人的菜地里偷劳动果实，长时间地坐在电脑前玩游戏，不少人更是半夜起来实施偷菜计划。这种过度劳累让身体吃不消，有的病倒入院。

除此之外，“偷菜”引发的负面报道不断出现，像有人沉迷“偷菜”引起离婚，有人网上“偷菜”不过瘾直接到菜地偷，还有人“偷菜”偷成强迫症……有鉴于此，近日传出文化部工作人员称偷菜游戏可能被取缔或改良。

- QQ等级制度更新（2011年11月）：
  - 八种基础加速：
    - 在线时长满5小时 最高可加速1.3天
    - 与5个好友或群互动 可加速0.1天
    - 发送50条消息 可加速0.1天
    - 连续登录达6天 可加速0.1天
    - 使用最新版QQ 可加速0.1天
    - QQ和Q+同时在线5小时 可加速0.2天
    - 非隐身时长满2小时 可加速0.2天
    - 使用5个Q+应用 可加速0.1天
  - 五种服务额外加速：
    - QQ电脑管家 额外1天加速
    - 超级QQ 最高1.9倍加速
    - 使用QQ输入法 额外0.1天加速
    - 腾讯微博 最高0.2天加速
    - QQ会员 最快1.8倍加速

来源 <http://news.mydrivers.com/1/209/209339.htm>

QQ等级加速方式悄然升级 最多3.5天＋3.7倍

2011-11-15 17:23:22 144872 人阅读 作者：上方文Q 编辑：上方文Q [复制链接] [我要爆料]

没等到新的QQ2011 SP版本，腾讯就已经悄然升级了QQ等级加速方式，现在打开“我的资料面板”中“我的等级”标签页即可看到新的加速说明。

大数据核心课程

商务大数据分析 >>2. 探索性数据分析

2.5 时刻留心数据质量

刘跃文 博士

教授、博士生导师

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

V2.2, 2021-9-8

提纲

- 什么是数据质量
- 数据质量的维度
- 拥抱不完美

1. 什么是数据质量？

- 是对数据的客观评估。
- 是对数据现状的认识和认同。

- **不是对数据无休止的诉求。**
- **不是抱怨、牢骚和不做工作的借口。**

2、数据质量的维度

- 数据质量（Wang & Strong 1996, JMIS；ISO8000）

- 核心指标：
  - 正确性/准确性
  - 完整性
  - 一致性
  - 及时性

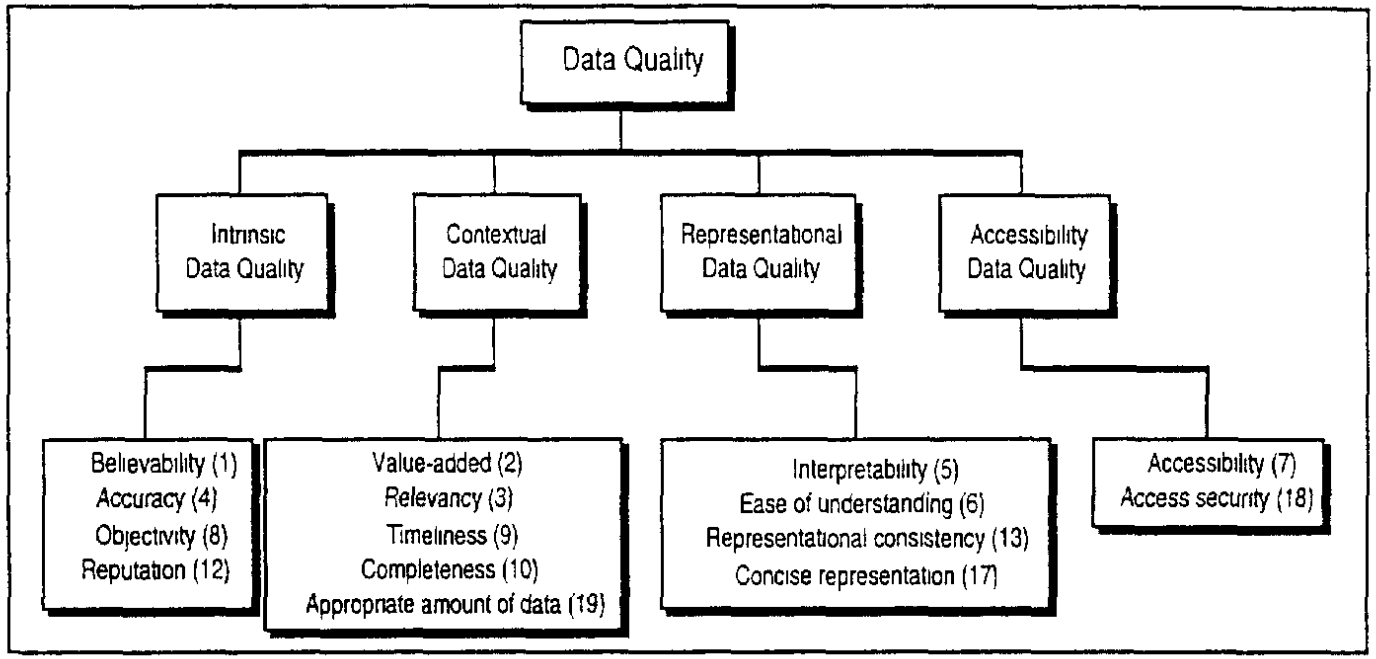


Figure 2. A Conceptual Framework of Data Quality

正确性/准确性（Correctness, Accuracy）

- 与客观世界不符
  - 识别错误：车牌、人像识别
  - **录入**错误：
    - 有意错误、无意错误
    - 页面填写、基站位置、核查电话号码
    - 录入错误与素质无关：某软件的操作编码
    - 录入错误受考核影响：批量输入假数据
  - 系统错误：
    - 系统**时间**、传输错误：某食堂数据的时间

- 因为不知道真实情况是什么，正确性有时很难评判！

- 现实中的数据准确性，一般都有问题。
  - 某互联网公司的好友操作记录；某酒店的会员记录：性别；

ID	FriendID	ActionTime	ActionTime	Flag
2010006	2020699	20 年 7 月 14 日	13:32:55	1
2010006	2020699	20 年 8 月 8 日	23:41:19	1
2010006	2020699	20 年 8 月 16 日	16:12:10	1
2010006	2020699	20 年 8 月 16 日	16:42:23	1
2010006	2020699	20 年 8 月 16 日	16:42:54	1
2010006	2020699	20 年 8 月 16 日	16:43:30	1
2010006	2020699	20 年 8 月 16 日	17:03:16	1
2010006	2020699	20 年 8 月 17 日	12:33:59	1
2010006	2020699	20 年 8 月 19 日	13:09:15	1
2010006	2020699	20 年 8 月 19 日	14:51:21	1
2010006	2020699	20 年 8 月 19 日	15:21:46	1
2010006	2020699	20 年 8 月 20 日	4:15:47	1
2010006	2020699	20 年 8 月 20 日	13:19:16	1
2010006	2020699	20 年 8 月 20 日	15:05:35	1
2010006	2020699	20 年 8 月 21 日	2:52:34	2

	gender	freq
1	M	12773970
2	F	6479097
3		791564
4	N	4362
5	0	1020
6	1	119
7	#M	9
8		1
9	19790522	1
10	#0449	1

完整性（Completeness）

- 字段（列）完整性：
  - 应该有值的字段没有空缺（较好评判）。
- 记录（行）完整性：
  - 没有缺失的记录（通过分布分析来判断）。

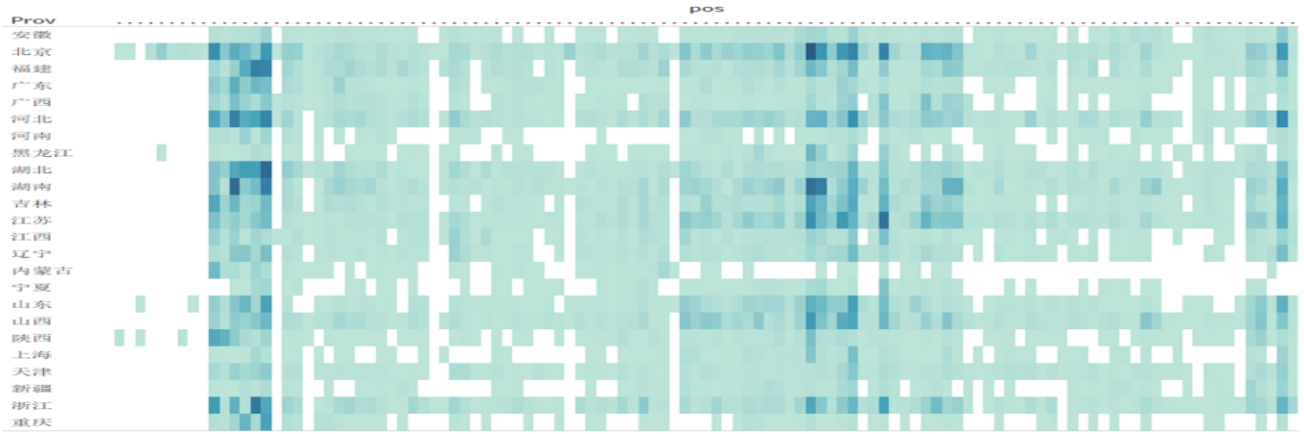
列/字段完整性

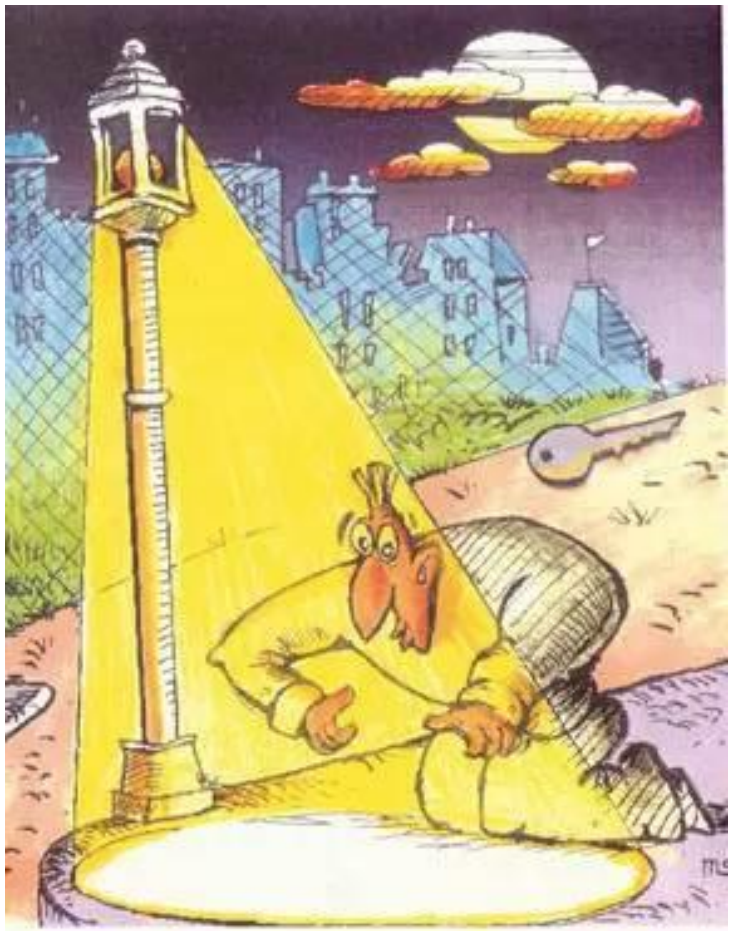
- 如何从来源消灭空字段？
  - 程序员的手段：下拉框、强制与默认值
  - 一定奏效吗？
    - 如果是比较多的选项，如地域、行业？你会怎么做？
- 设计者与操作者的博弈
- ID缺失怎么办？

记录完整性

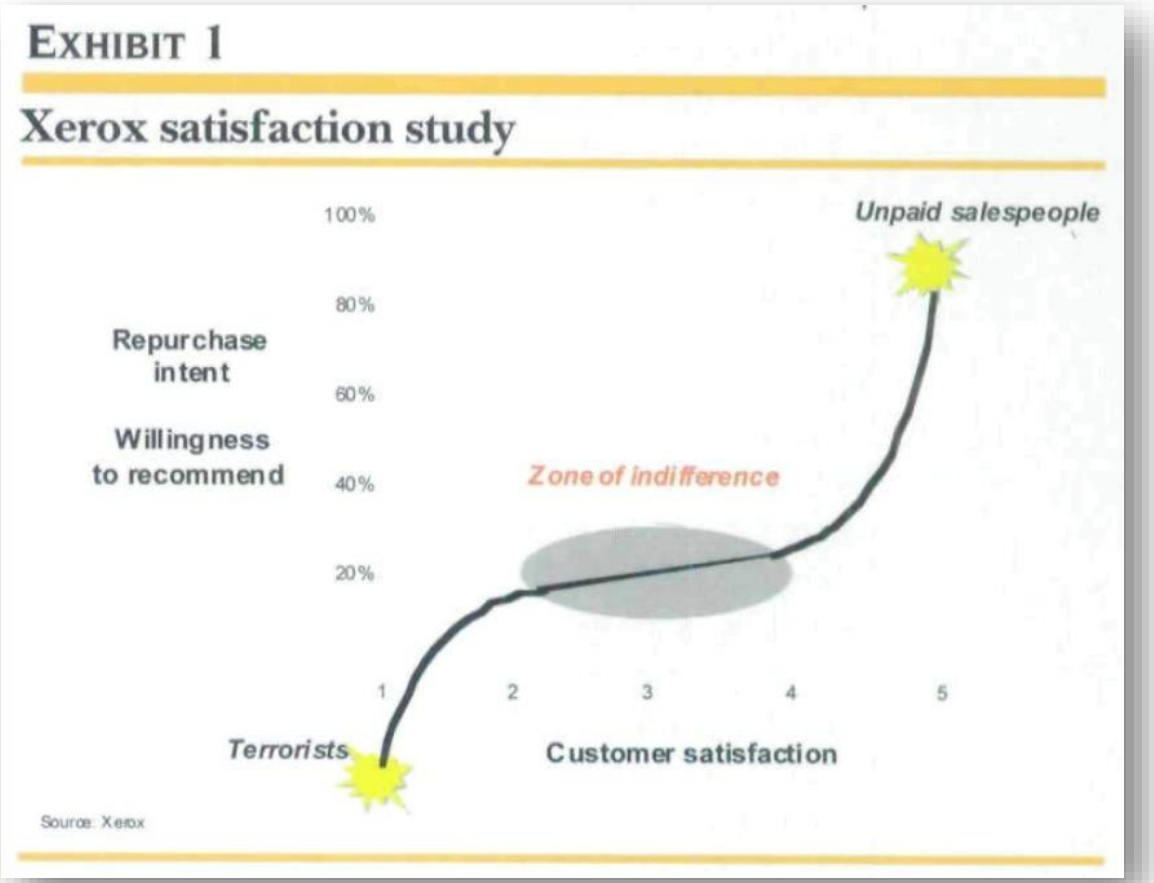
- Response Bias 有偏的样本
  - 只有负责任/感兴趣的人才会填写数据
  - 只有正常运转的设备才会上报数据

- 记录完整性的一种分析方法：二维表

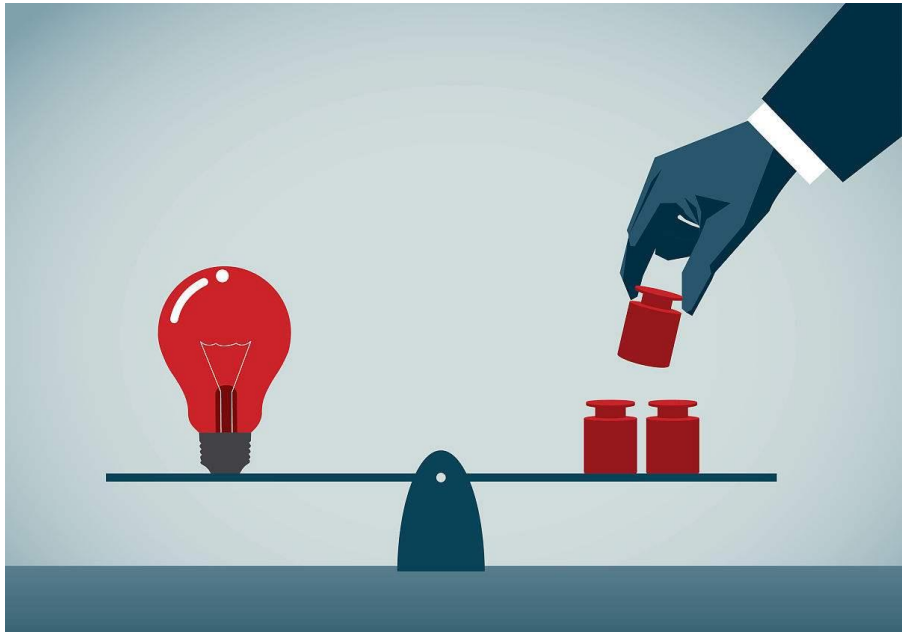




案例：用户评价的可靠度



- 如何解决正确性、完整性的问题？



- **发现**数据正确性、完整性问题
  - 多源数据比较（对账）
  - 分布分析、二维表
- **解决**数据正确性、完整性问题
  - 机制设计与博弈
  - 责任到人，及时反馈（信息、收益、惩罚）
  - 实事求是（数据的正确性、完整性不可能达到100%）

一致性

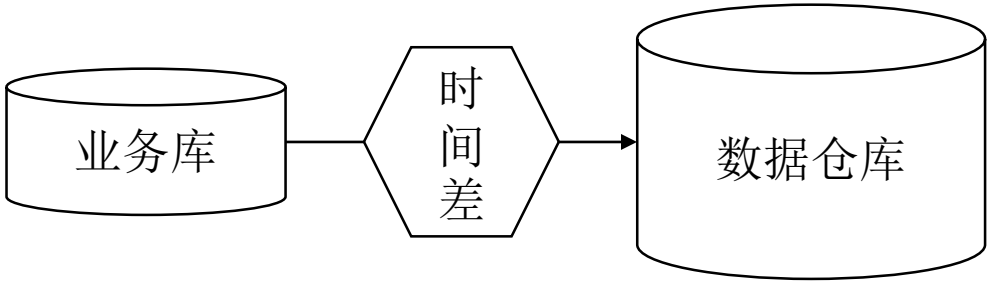
- 一致性：相同的内容，数据应当相同
- 大学的“名称”
  - 科大、中科大、中国科大、中国科技大学、USTC、中国科学技术大学
  - 交大、西交大、西安交大、西安交通大学、仙交大、香蕉大、XJTU

- 解决一致性问题：
- 程序员的手段：
  - 下拉框、复选框（尽可能地标准化）
- 数据分析师的手段：
  - 实体识别、文本分析、语义网络
  - 例：ERP系统中的物料名称

及时性（Timeliness）

- 及时性：数据及时入库
- 业务数据库：实时性、高并发、其它单位
- 数据仓库（汇聚库/分发库）：大容量、低并发

- 从业务库到数据仓库：
  - 批量转存
  - 存在时间差



- 大数据不能过分苛求原始数据的及时性。

3、拥抱不完美

- 数据质量问题：
  - 正确/准确性，完整性，一致性，及时性

- 数据质量**不可能**完美！
- 接受不完美的数据，构建**可用**的模型！

- 大数据的特点：
  - 从低质量数据中加工高质量信息/知识！

西安交通大学管理学院
THE SCHOOL OF MANAGEMENT
XI'AN JIAOTONG UNIVERSITY

2.6 编写数据分析报告

刘跃文 博士

教授、博士生导师

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

V2.2, 2021-9-8

1. 数据分析报告的撰写

- 数据分析报告的目的是：给别人带来“知识”

- 不能只是罗列数据分析结果

- 要努力：
- 描述现状
- 识别并量化异常状况
- 探索数据中的规律

明确数据分析报告的用户

- 数据分析报告的用户是谁？
  - A 基层员工
  - B 基层主管
  - C 中层领导
  - D 高层领导

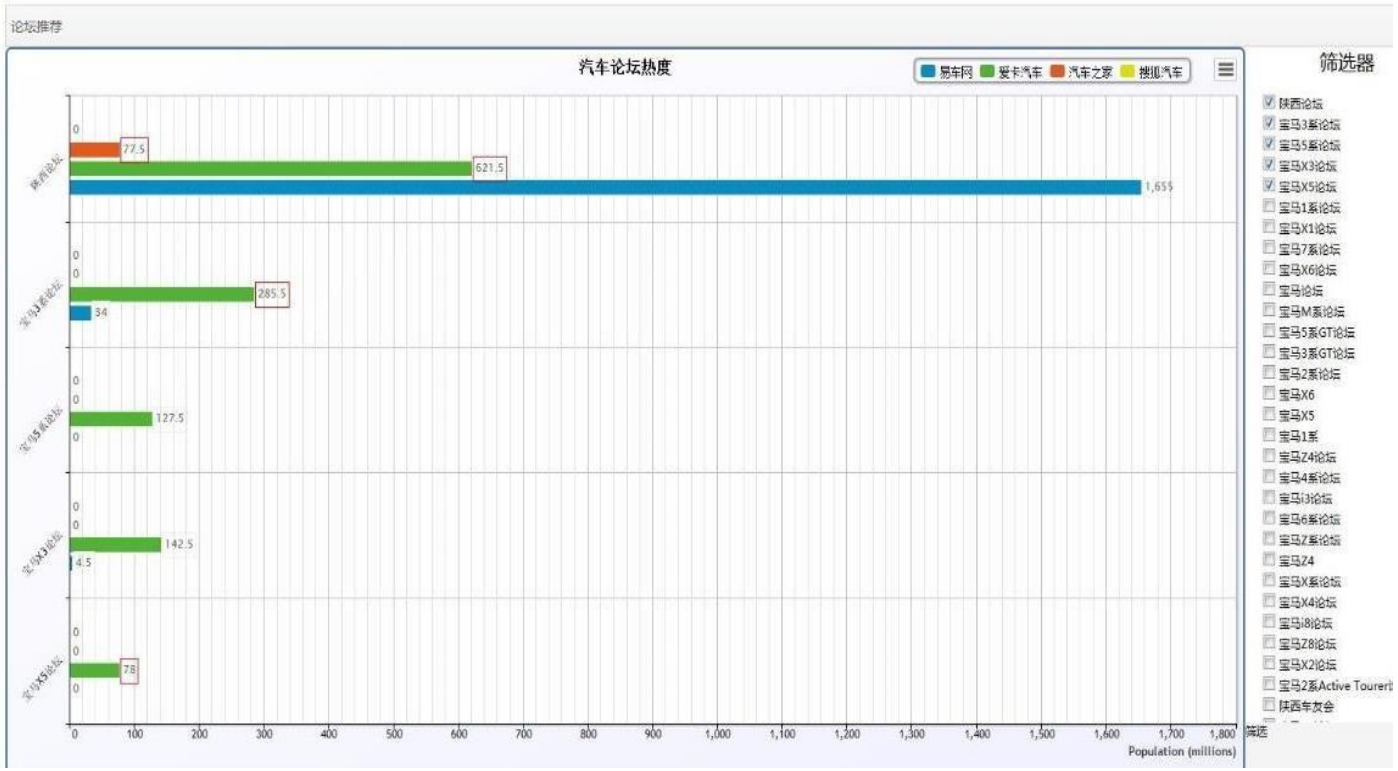
- 不同用户的需求并不相同；一般而言，越基层越需要详细数据，越高层越需要汇总的结论。
- 不同任务的需求也并不相同。

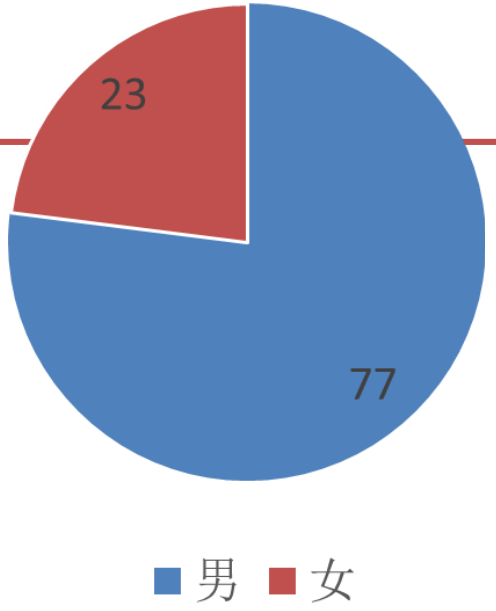
数据分析报告的写作技巧

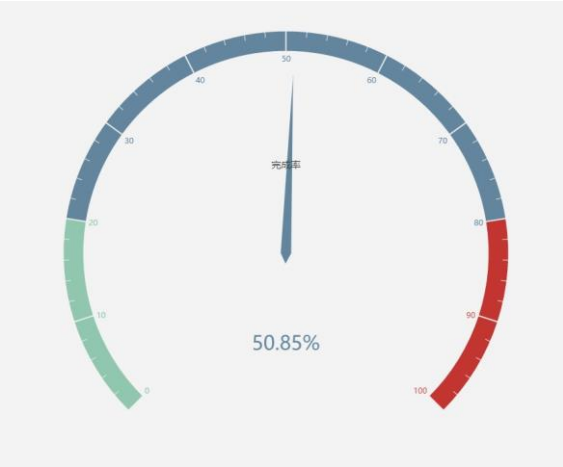
1. 明确目标用户和任务；
2. 理清层次与逻辑；
3. 言之有物，有理有据；不要罗列数据与图表；
   （切记：不要为了画图而画图；有时数字和表格更直白）
4. 识别亮点：为用户新增了哪些信息和知识；
5. 增加必要的文字说明，解释清楚图表内容
6. 提炼结论：提炼摘要及Bullets

合理的图表布局

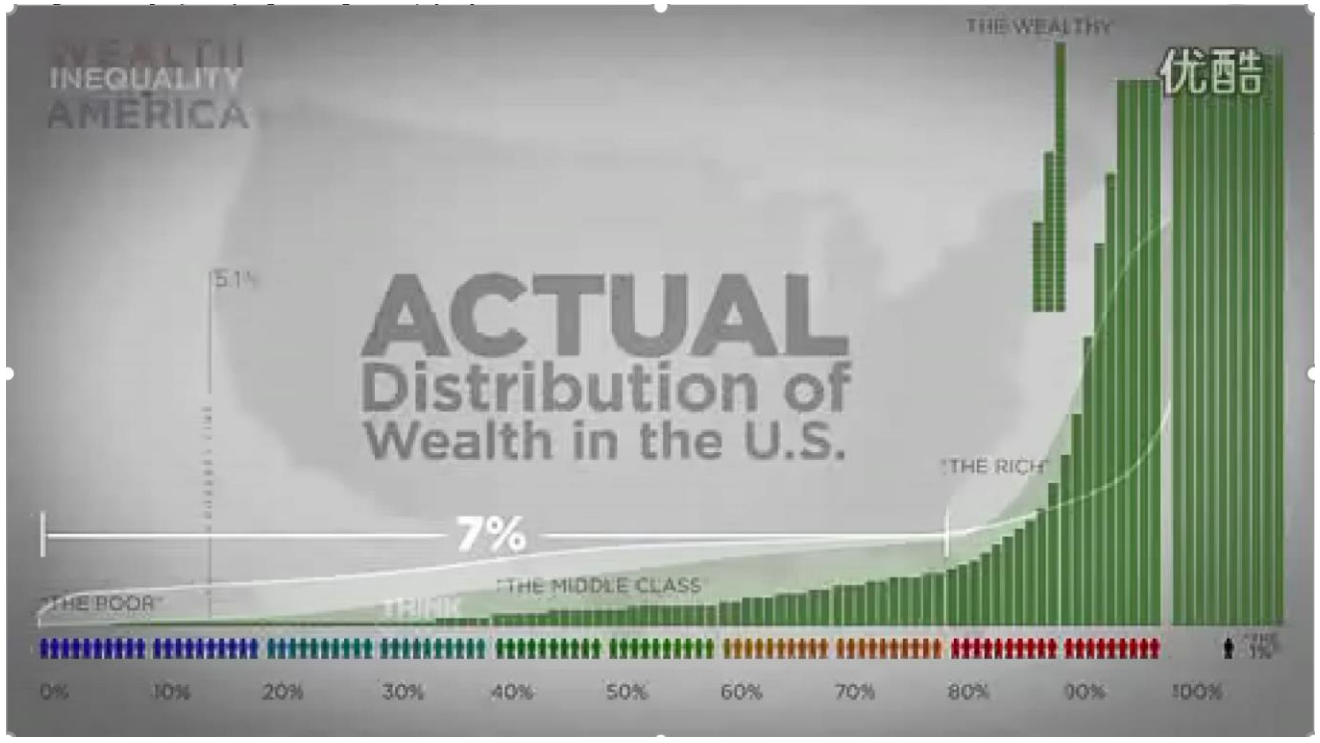
- 避免过低的信息密度（过于浪费空间）







如何处理特别长的数字？



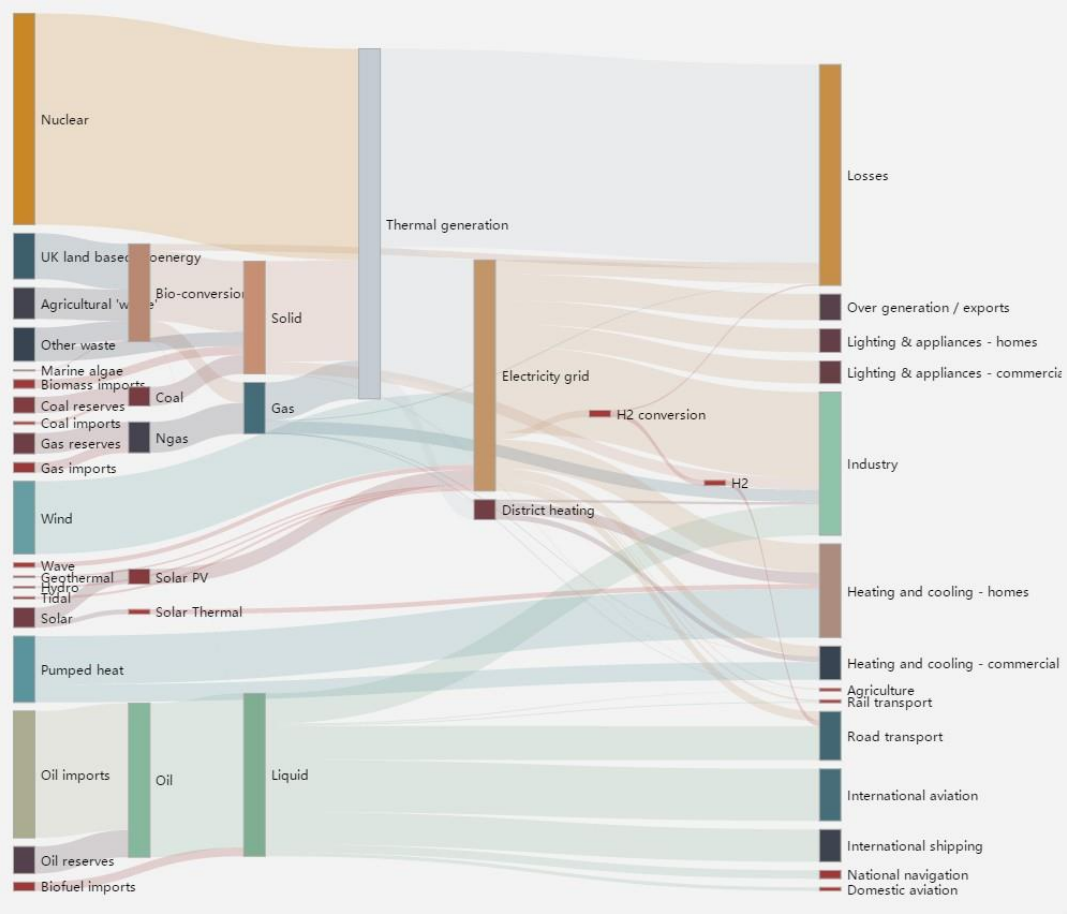
图表格式不能错乱

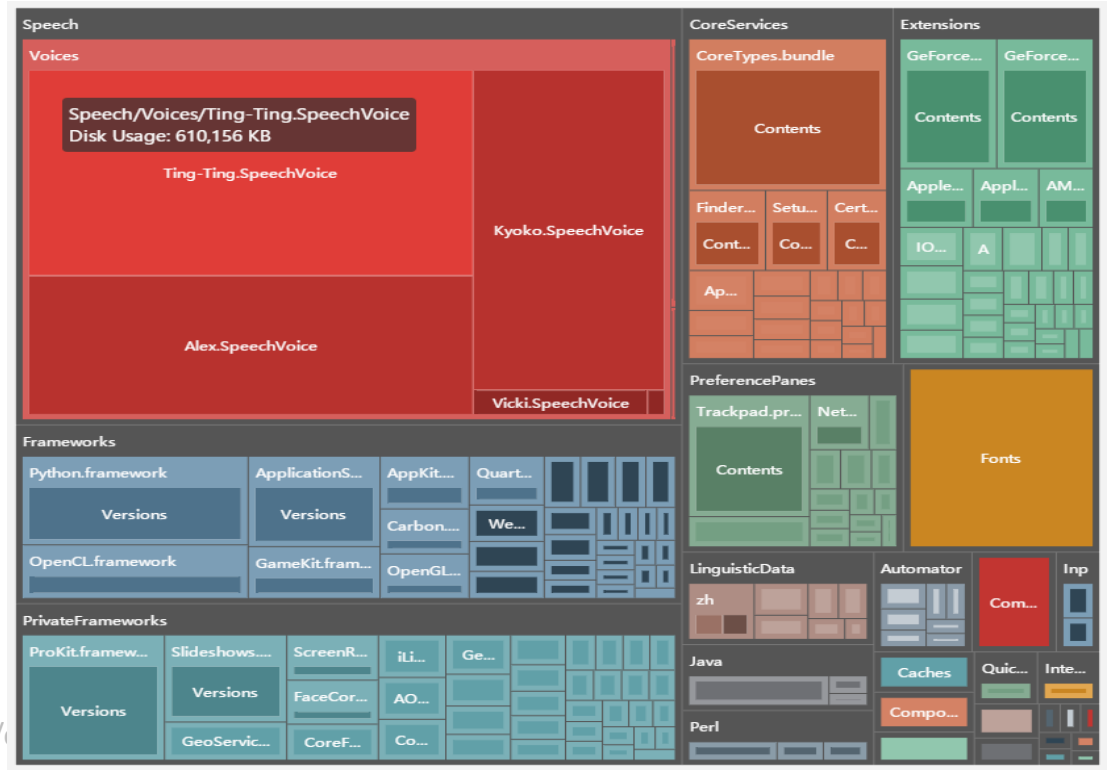
经销商	促销	活动
奥迪		
陕西奥诚	151	73
陕西庞大乐业	114	68
陕西新丰泰	110	9
陕西新丰泰博奥	77	135
宝马		
陕西金花	210	148
西安荣宝	99	128
西安顺宝行	32	74
奔驰		
西安利之星	131	113
西安庞大兴驰	93	29
西安新丰泰之星	63	56
西安之星	34	35
雷克萨斯		
西安钧盛雷克萨斯	222	228
西安元丰雷克萨斯	86	44
路虎		
陕西惠通陆华	78	107
陕西天华	66	101

经销商	促销	活动
奥迪		
陕西奥诚	151	73
陕西庞大乐业	114	68
陕西新丰泰	110	9
陕西新丰泰博奥	77	135
宝马		
陕西金花	210	148
西安荣宝	99	128
西安顺宝行	32	74
奔驰		
西安利之星	131	113
西安庞大兴驰	93	29
西安新丰泰之星	63	56
西安之星	34	35
雷克萨斯		
西安钧盛雷克萨斯	222	228
西安元丰雷克萨斯	86	44
路虎		
陕西惠通陆华	78	107
陕西天华	66	101

经销商	促销	活动
奥迪		
陕西奥诚	151	73
陕西庞大乐业	114	68
陕西新丰泰	110	9
陕西新丰泰博奥	77	135
宝马		
陕西金花	210	148
西安荣宝	99	128
西安顺宝行	32	74
奔驰		
西安利之星	131	113
西安庞大兴驰	93	29
西安新丰泰之星	63	56
西安之星	34	35
雷克萨斯		
西安钧盛雷克萨斯	222	228
西安元丰雷克萨斯	86	44
路虎		
陕西惠通陆华	78	107
陕西天华	66	101

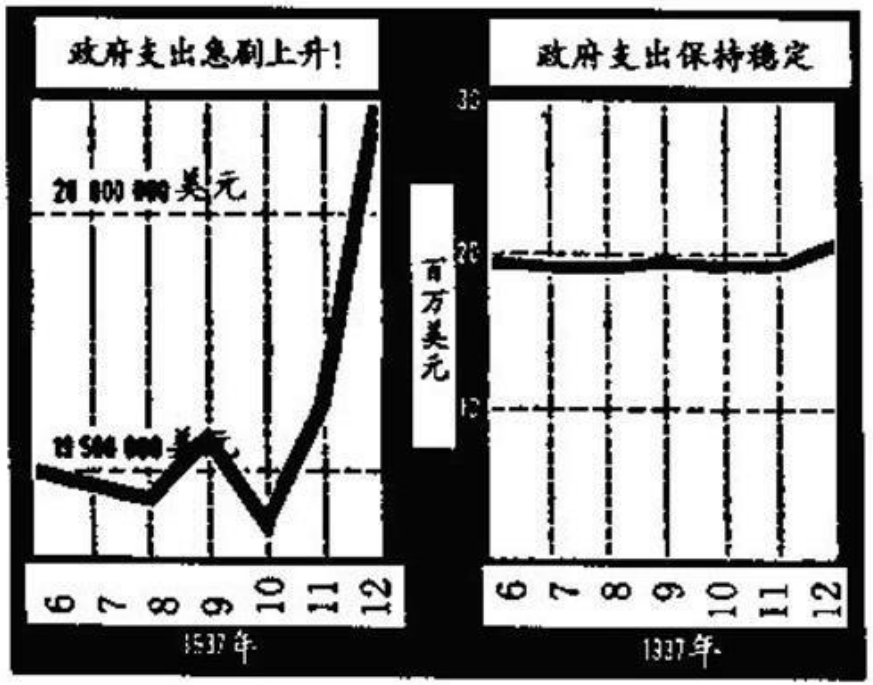
新颖、直白的图表类型

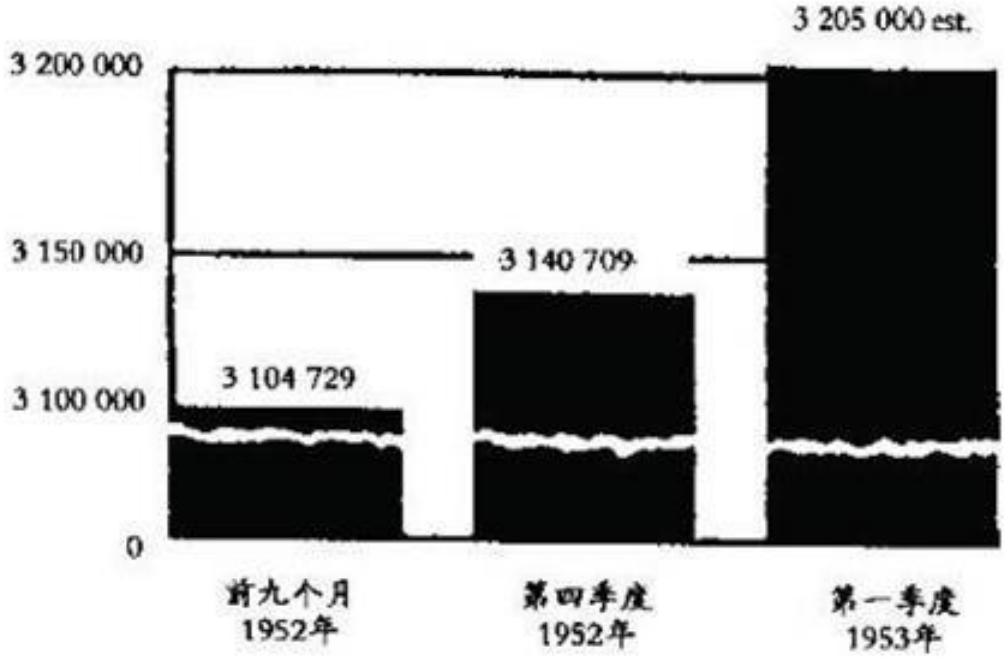




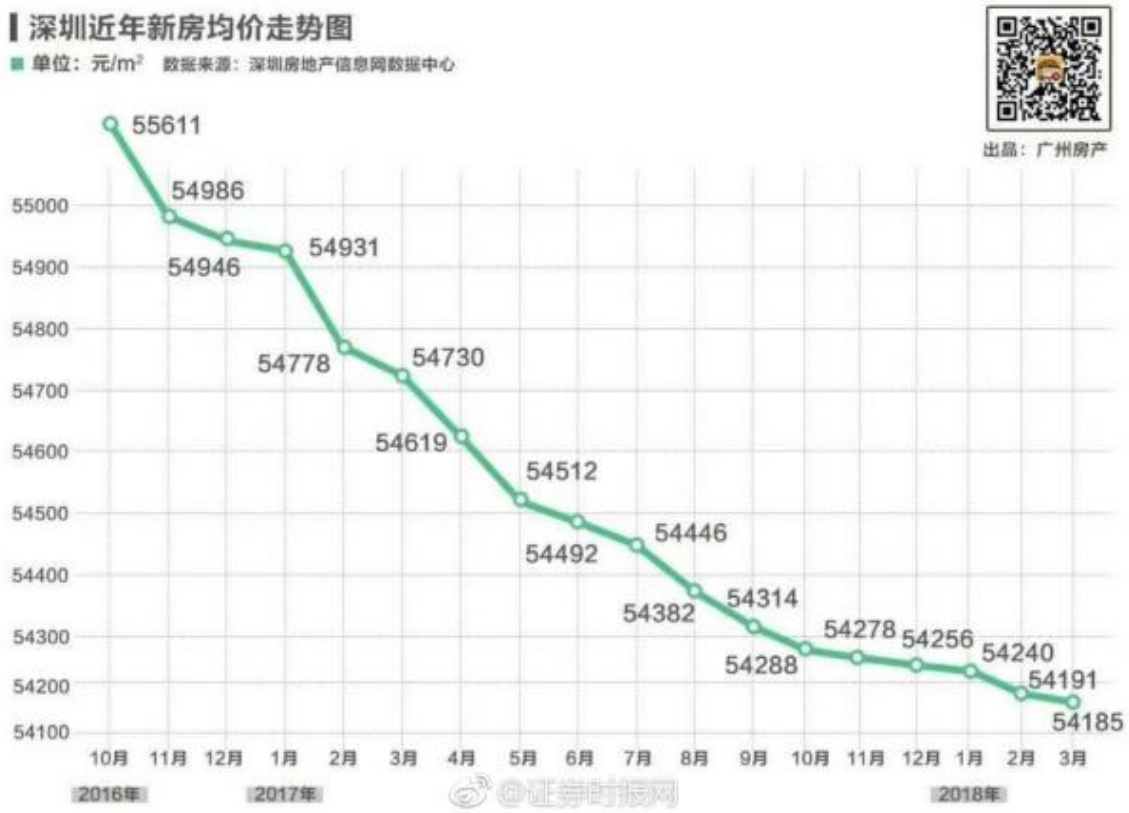
2. 不能刻意去误导别人

- 图表正确地传达客观事实。
- 《统计数字会撒谎》





案例：深圳房价大幅度下跌



- 原话是这样说的：“显然是调控在这里起着非常重要的作用，就是不让这个价格往上走，这个势头还是比较猛的”。
- 并不是针对三月均价下跌6元这个具体数字说势头猛的，而是针对十八个月来的调控形势说的。
- http://sz.news.fang.com/open/28153331.html

案例：中国队真牛

在长达84年20届世界杯决赛周历史上，只有三支国家队战胜过中国国家队，分别是巴西、土耳其和哥斯达黎加，即使是巴西这样的世界强队也仅战胜过中国一次，而中国国家队也从未在世界杯十二码大战中失利过，从来没有一支球队能够在世界杯上击败过中国队两次，而且中国队在世界杯上失球数远少于巴西和防守见称的意大利队，在过去84年里中国失球数只有9球，除此之外世界上除了巴西，中国是另外一支在胸前有五粒星的球队。

案例：人均住房面积

导读 preface

日前， 由北大中国社会科学调查中心完成的《中国民生发展报告2012》 显示，目前中国家庭平均住房面积为116.4平方米，人均住房面积为36.0平方米 。很多人在知道这个调查结果后都惊呼自己“又被平均了”。对于这个数字，很多人表示难以置信，有人无奈自嘲“对不住全国人民，我又拖后腿了”。

全国家庭平均住房面积116.4㎡，你被平均了吗？

543 拖后腿了　51 超过了

报告称全国家庭平均住房面积116.4㎡

概述： 日前，北京大学召开中国家庭动态跟踪调查研讨会，发布由北大中国社会科学调查中心完成的《中国民生发展报告2012》。调查显示，2011年全国家庭现住房完全自有率为84.7%。全国家庭的平均住房面积为116.4平方米，人均住房面积为36.0平方米 ……[详细]

国家	人均住房面积	国家	人均住房面积
美国	67㎡	英国	35.4㎡
意大利	43㎡	法国	35.2㎡
荷兰	40.82㎡	西班牙	25.8㎡
德国	39.4㎡	韩国	19.8㎡
中国	36㎡	日本	19.6㎡

部分国家人均住房面积参考

案例：自媒体

首先是报纸刊登了美少女中考状元吴思齐的照片，照片中的主体是她一个人。该报道的主题意在突出少女学霸多才多艺。

几天之后，有自媒体公布了这个照片的完整版：

哈哈！我们在报纸上看到的美少女状元，原来也早恋(´･ɪ･`)

画面中两个人牵着手，看上去非常幸福呢~ɐ(˘ ᴗ ˘)ɔ

又过了几天，又有人放出了**完整版的完整版**：

什么早恋！他们当时正在拍毕业照做游戏好嘛！我们的美少女仍然是少女好嘛！

事实上，如果你仔细对比三张照片，人物的细节动作都是不一样的，换句话说这不是截图报道，而是在同一次活动上的三张照片。

纵观这件不起眼的小事，我们会发现：

纸媒拍照后截取了人物主题，意在突出美少女个人，传播积极向上的正能量；

某自媒体为了带流量带关注，发布了让人出乎意料的牵手版本，**但绝对故意隐瞒了学生们做游戏的环境；**

直到有人站出来澄清，发布了最终的极限反杀照片，告诉大家“你们都被骗了”。

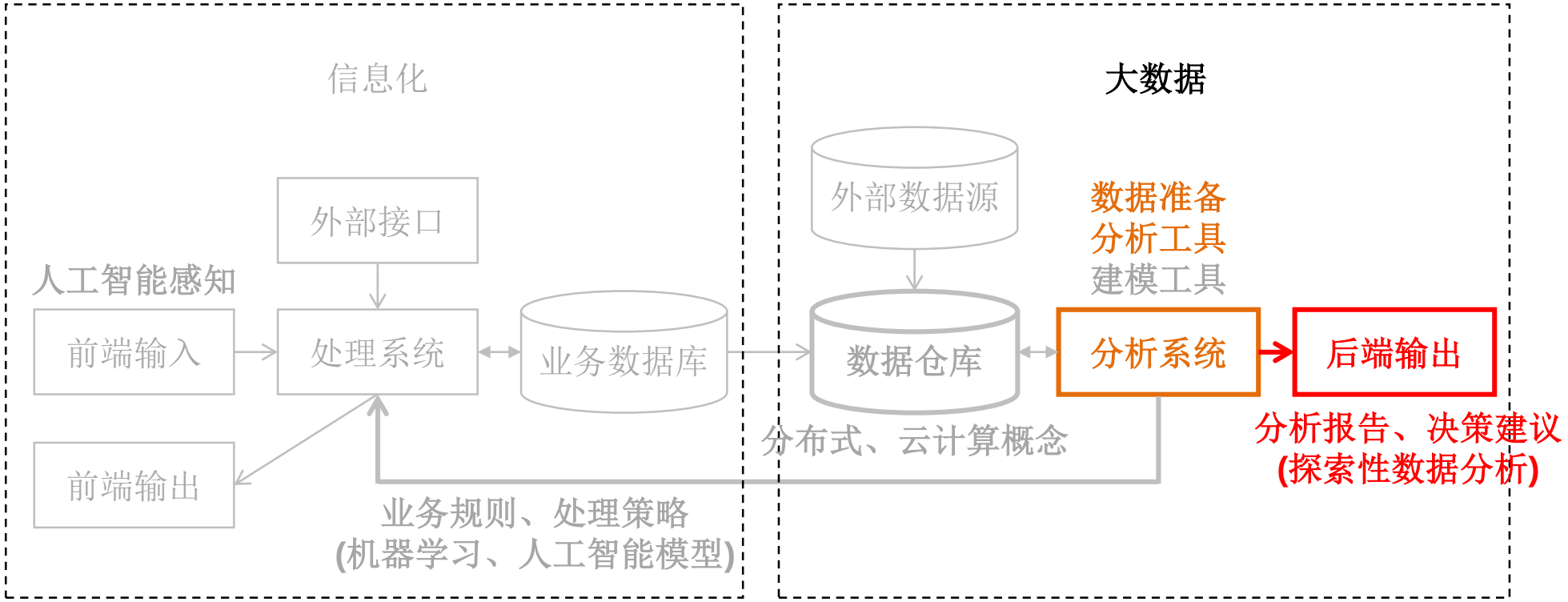
“你看到的，都是媒体想让你看到的。”一个活生生的例子摆在眼前，大家这下知道媒体的恐怖之处了吧。很多时候我们不要以为看到的就是事实，自己想的就是最正确的，其实我们早就被潜移默化地影响了思考。

为什么？

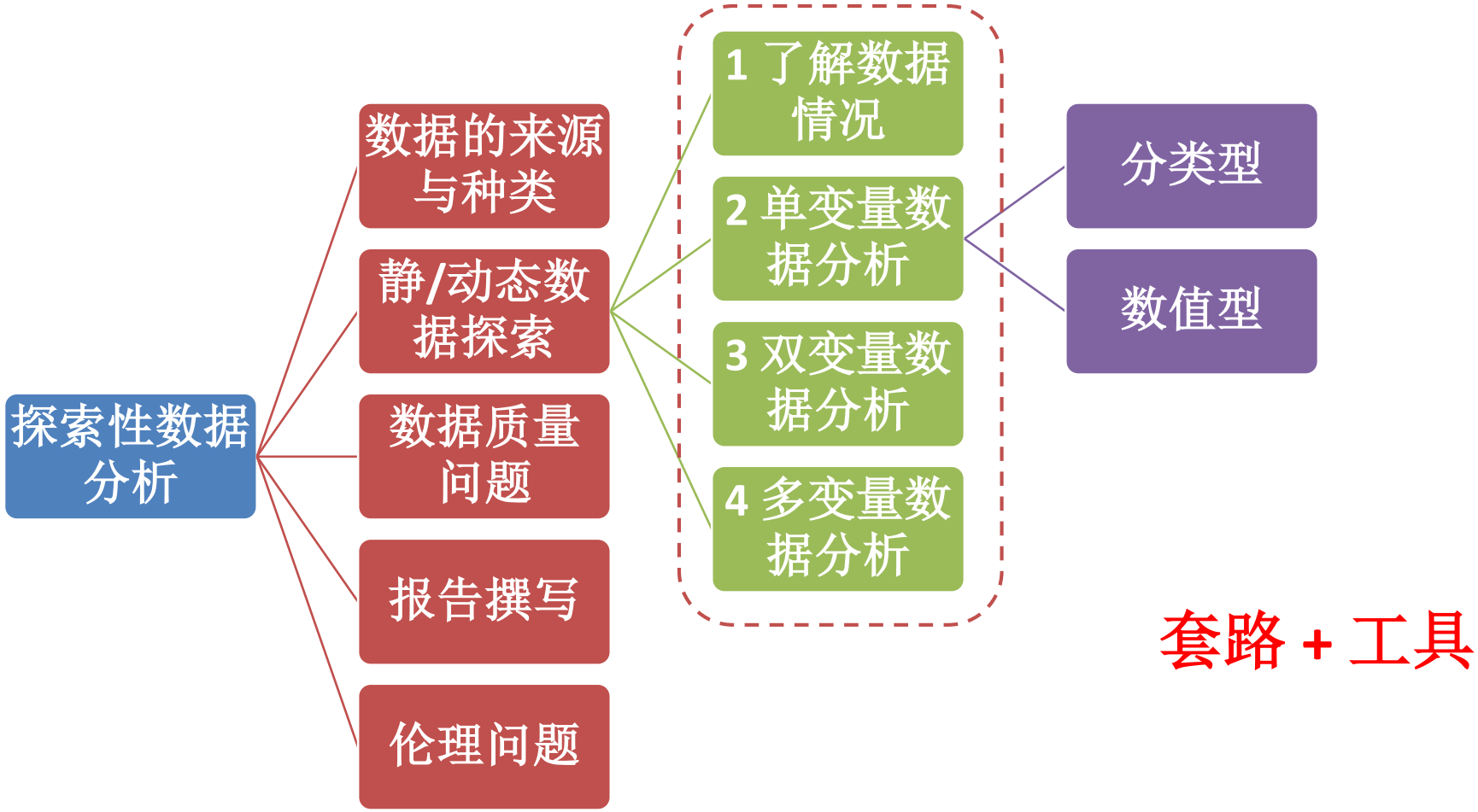
- 因为这些报告虚报数据？歪曲事实？
- **其实并没有。非常客观地陈述了事实。**

- 基于数据的误导，有两类：
- （1）只报告一部分数据，形成直观印象，其余部分由读者自行脑补。
- （2）使用一些不恰当的统计指标。比较常见的是，在非正态分布时，使用均值、p值。

知识地图



知识体系



套路 + 工具

谢谢！