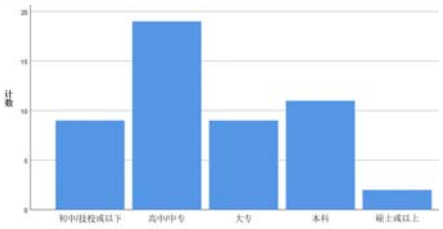
统计图的分类框架

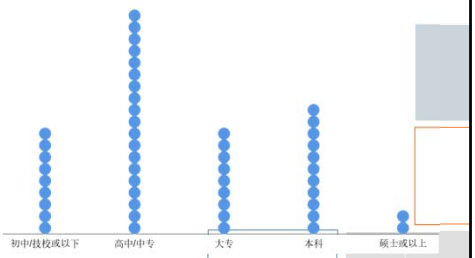
- 统计图的分类方法有许多种，但作为数据分析和呈现的工具，最好使用和统计学体系最为贴近的方法对其加以分类
- 首先按照其呈现变量的数量，将统计图大致分为单变量图、双变量图、多变量图等
- 随后再根据相应变量的测量尺度进行更细的区分
- Tableau、python等新兴的数据可视化工具出于各种考虑，提供了一些新式的图形，他们并不完全符合标准的统计绘图要求，对这些图形的使用应当谨慎，注意不要因此冲淡分析主题

单个-分类变量-呈现原始频数

- 简单条图
  - 按照分类区分直条，直条高度代表频数大小

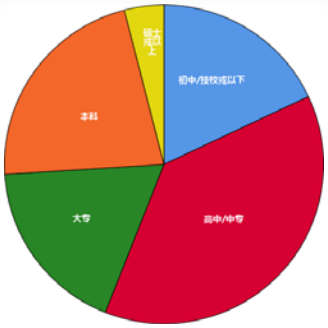


- 点图
  - 类似于简单条图，但是用散点代表每一个案例



单个-分类变量-呈现数据构成比

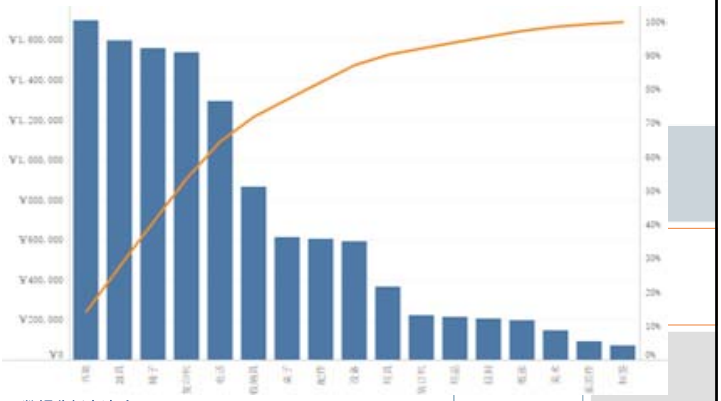
- 饼图
  - 饼块大小代表频数/构成比大小



- 分段条图/百分条图
  - 按照分类区分颜色，条段大小代表频数/构成比大小

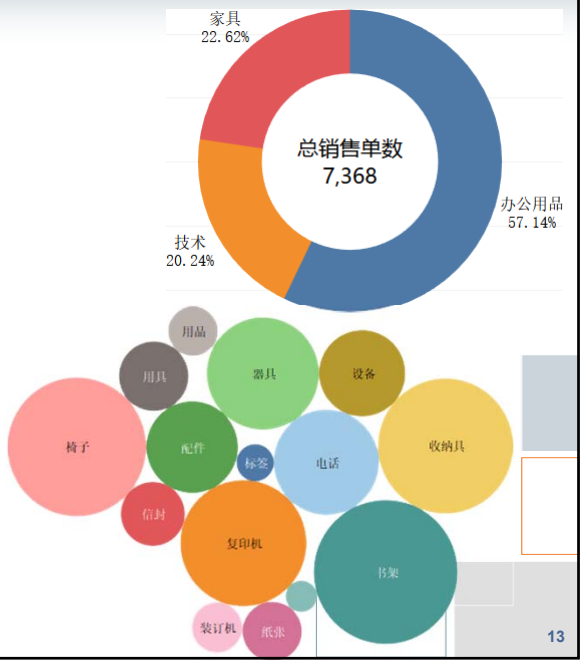
单个-分类变量-呈现原始频数/构成比

- Pareto图：条图和线图的组合
  - 直条代表绝对数值，按照降序排列
  - 折线代表累积百分比，因此呈现为上升速度减慢的曲线
  - 两方面的数据相结合，就可以迅速确定业务类别中最关键的部分



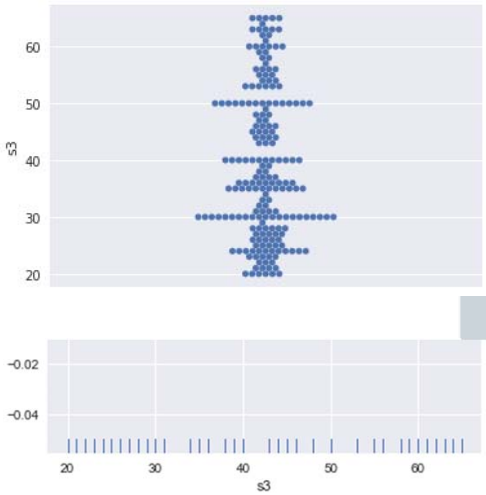
单个-分类变量-各种特殊图形

- 圆环图
  - 可提供汇总信息等附加信息
  - 非标准统计图形
- 气泡图
  - 用气泡大小代表频数/构成比大小
  - 违背了统计图形应当便于对比数据的基本原则，很好看，但需要控制使用



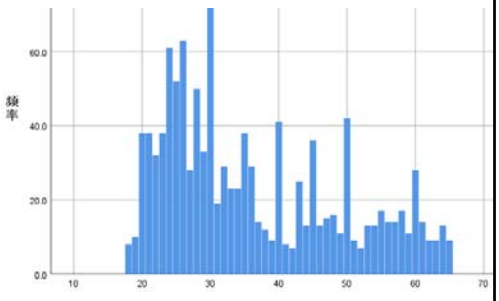
单个-数值变量-显示原始数据分布

- 条带图（Strip Plot）
  - 以散点的形式展示原始数据分布
  - 绘图时会加入随机扰动以改善显示效果
- 地毯图（Rug Plot）
  - 只显示原始数值位置，不显示频数
  - 单独使用无价值，一般作为辅助图形出现



单个-数值变量-显示汇总后数据分布

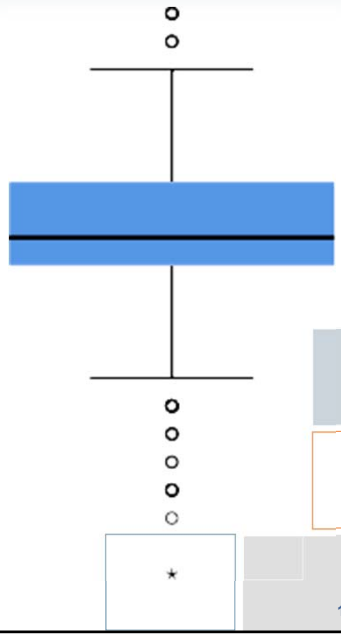
- 直方图
  - 对数值进行分组频数汇总，呈现整个取值区间上的数据分布基本特征



- 茎叶图
  - 对数值进行分组频数汇总，呈现整个取值区间上的数据分布基本特征
  - 比直方图更详细，可提供离群值等信息
  - 一般是纯文本图形

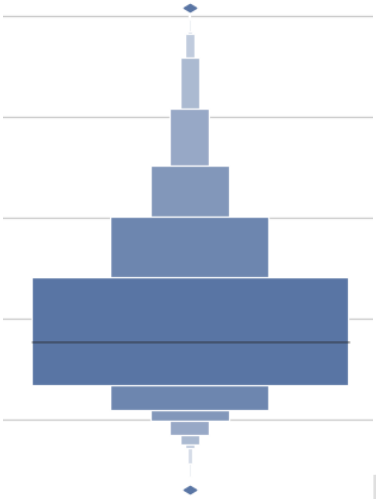
单个-数值变量-显示汇总后数据分布

- 箱图
  - 使用百分位数体系刻画整个取值区间
    - 箱体最中间的粗线为P50（中位数）
    - 方框上下界为P25和P75（四分位数）
  - 与四分位数（即方框上下界）的距离超过1.5倍四分位间距（即方框长度）的都会被定义为离群值，该距离界限以内的第一个数值位置在图中以线段表示
    - 界限外的离群数据用散点加以表示
    - 超过3倍则定义为极端值
  - 所有数值均未超界时，该线段就是最大/最小值



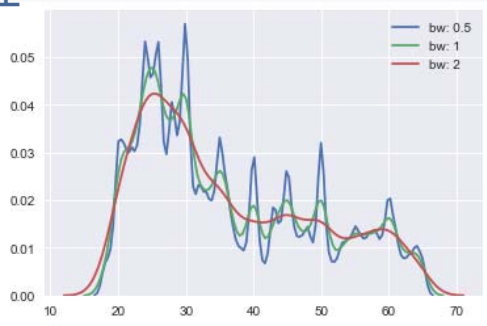
单个-数值变量-显示汇总后数据分布

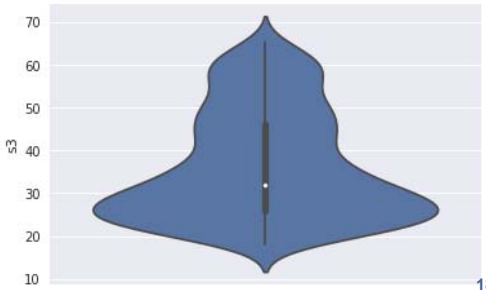
- 增强箱图
  - 对于大样本数据，箱图只显示IRQ和离群值，显然提供的信息不够丰富
  - 增强箱图则考虑提供更丰富的百分位数信息



单个-数值变量-显示数据分布特征

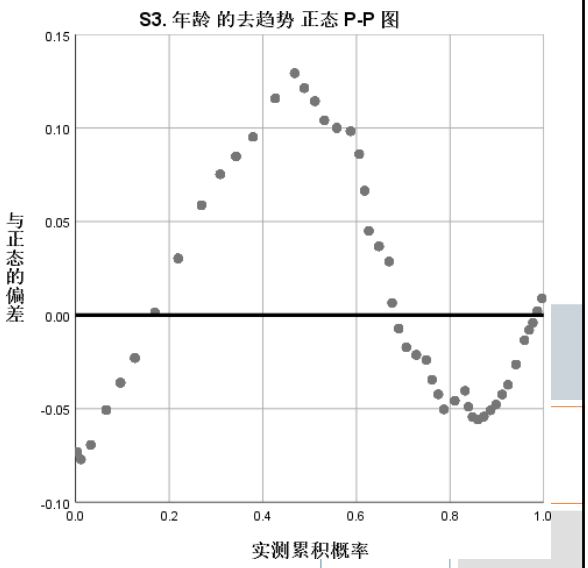
- KDE图（核密度估计）
  - 基于样本数据本身拟合最适合的分布曲线
  - 曲线形状会受到参数取值的影响
  - 不宜单独使用，以免产生误解
    - 例如和直方图联合使用
- 提琴图（Violin Plot）
  - KDE图和箱图的结合
  - KDE图以对称形式绘制
  - 也可以进行其他同类图形的组合





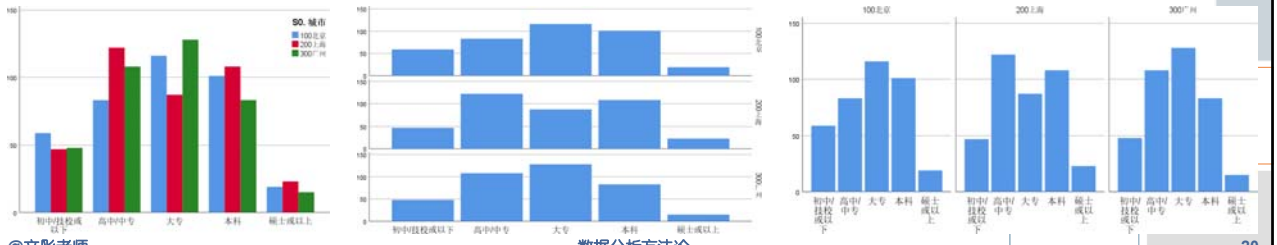
单个-数值变量-和假设的分布特征相比较

- P-P图
  - 将变量的实际累积分布概率与所指定的理论累积分布概率分别作为横、纵坐标而绘制的散点图，用于直观地检测样本数据是否符合某一概率分布
  - 如果被检验的数据符合所指定的分布，则代表样本数据的点应当基本在代表理论分布的对角线上
- Q-Q图
  - 图形用途与P-P图相同，但使用分位数进行散点的绘制



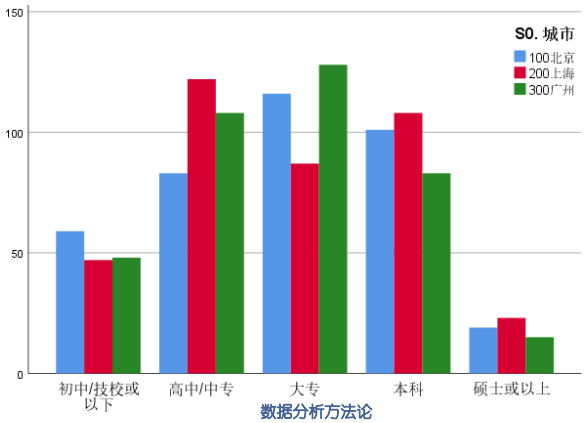
分类变量 vs. 分类变量

- 核心问题在于如何对组合单元格的频数/百分比进行呈现
- 在研究问题可以区分变量主次的情况下，可以将其中一个作为分组变量（分类轴），然后对另一个分类变量进行呈现
  - 分组图：在同一图形内分组分别绘图
  - 行面板图：按类别分为不同行单独绘图
  - 列面板图：按类别分为不同列单独绘图



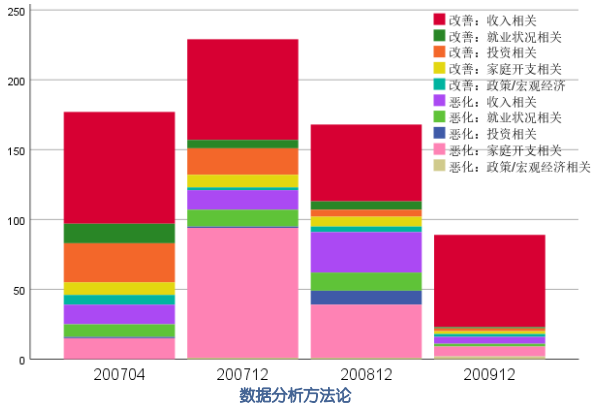
分类变量 vs. 分类变量

- 分组条图
  - 呈现两个分类变量各种类别组合下的频数状况
  - 便于对所有细分后各个单元格做直接对比，但难以呈现各变量边际分布的状况



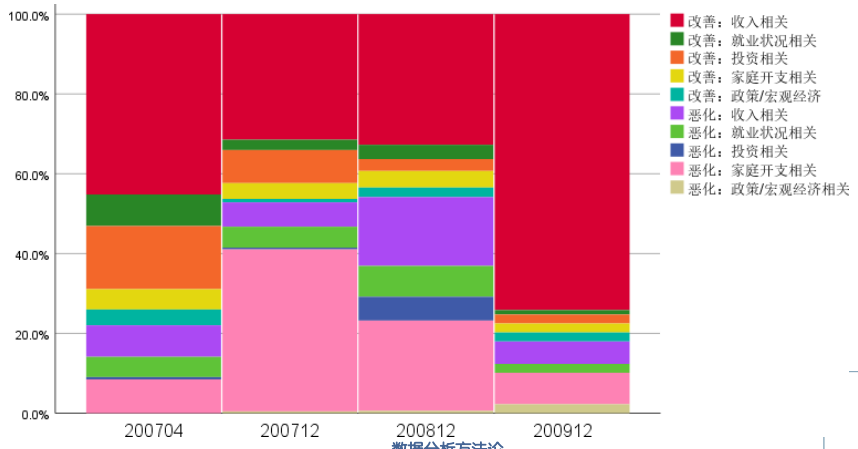
分类变量 vs. 分类变量

- 分段条图（堆积条图）
  - 突出一个分类变量各类别的频数，在此基础上表现两个类别的组合频数
  - 相对而言较难呈现出细分后各个单元格的直接对比信息



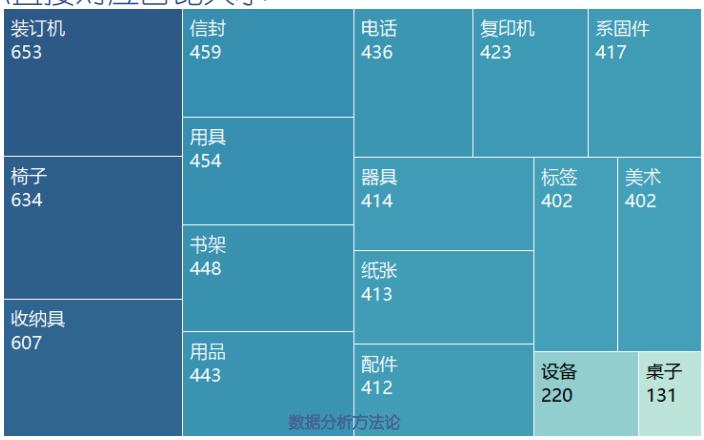
分类变量 vs. 分类变量

- 百分条图（马赛克图）
  - 呈现在一个变量不同类别下，另一个变量各类别的百分比变化情况



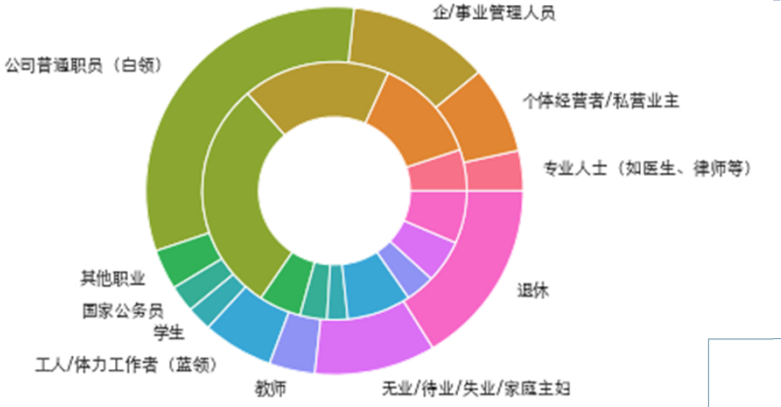
分类变量 vs. 分类变量

- 树状图
  - 将两个分类变量置于同等地位，显示各个组合单元格所占的频数/百分比
  - 单元格面积直接对应占比大小



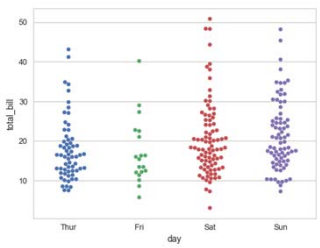
分类变量 vs. 分类变量

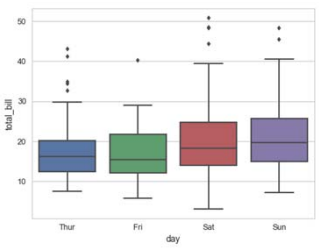
- 复合饼图/圆环图
  - 采用圆环套叠的方式，每层圆环代表一个大分类
  - 圆环内部的饼块/片段大小代表该大分类中某个小分类的频数/构成比大小

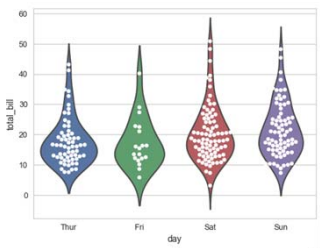


数值变量 vs. 分类变量-显示原始数值/汇总后特征

- 直接按照分类变量分组，各组内分别呈现数值变量信息
  - 分组条带图
  - 分组箱图
  - 分组提琴图
  - 同类图形组合

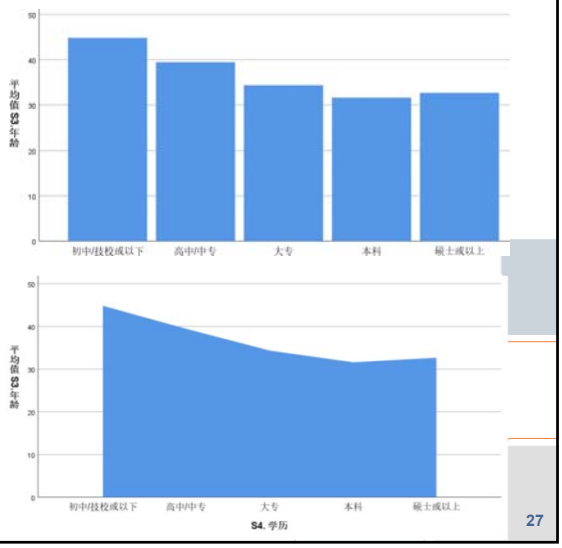






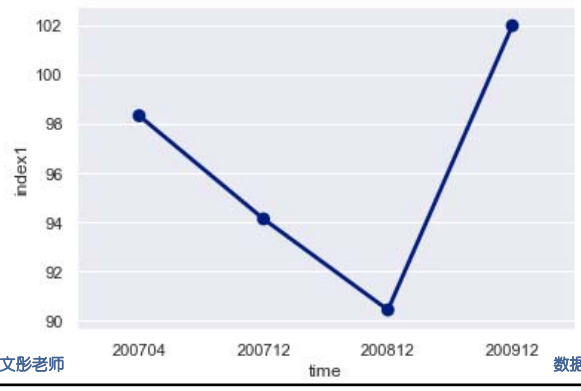
数值变量 vs. 分类变量-显示数值汇总指标

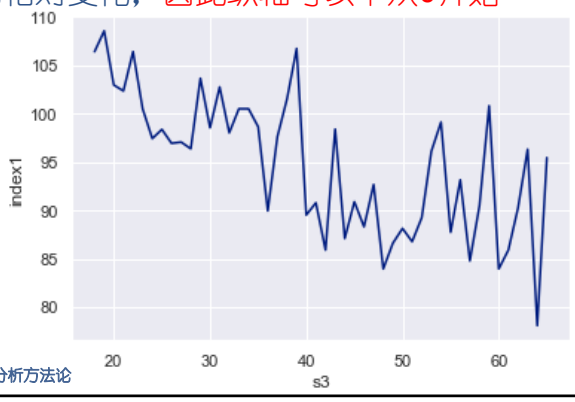
- 数值变量只需要呈现对应汇总数值，分组图形被高度简化
- 简单条图
  - 用直条高度反映数值汇总指标在各类间的绝对数值差异
  - 可在汇总指标基础上加绘可信区间等
  - 误差图/区间图：重在显示区间
- 面积图：基于条图直接衍生而来
  - 使用场景基本和条图相同



数值变量 vs. 分类变量-显示数值汇总指标

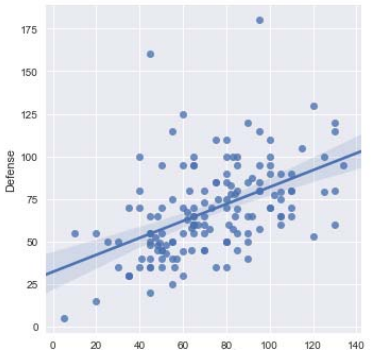
- 线图
  - 呈现有序分类自变量（或者时间变量）的影响
    - 也可扩展至连续变量使用，但注意纵轴显示的是汇总信息！
  - 重点在于呈现各类别间汇总指标的相对变化，因此纵轴可以不从0开始

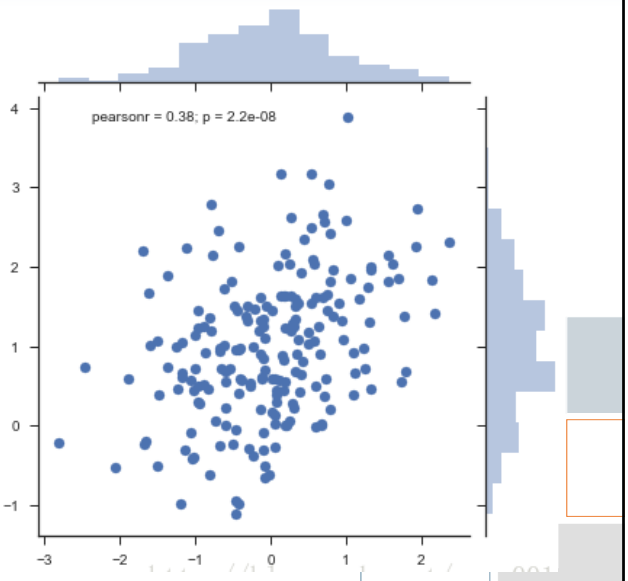




数值变量 vs. 数值变量-呈现原始数值分布

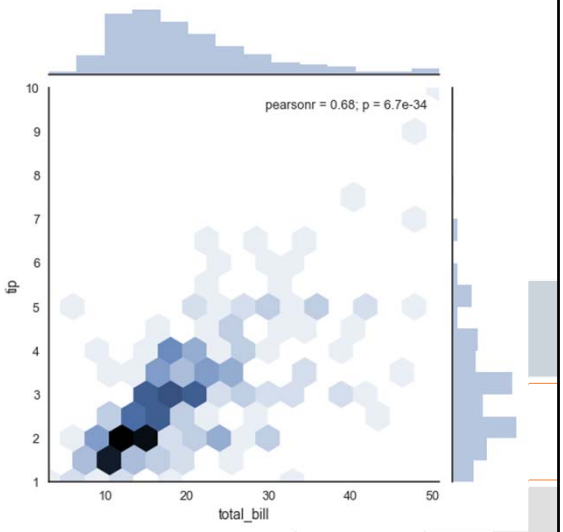
- 散点图
  - 呈现两变量在数量上的关联特征
  - 可加绘回归曲线及可信区间
  - 可加绘两变量各自的分布曲线





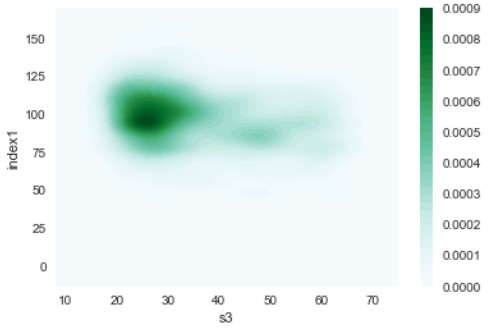
数值变量 vs. 数值变量-对原始数值做适当汇总

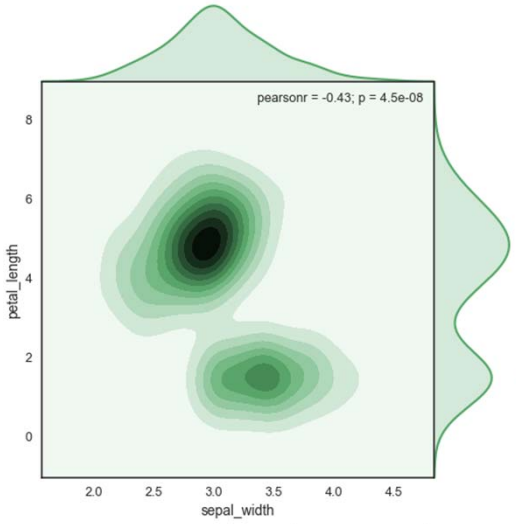
- Hexplot/Sunflower图
  - 本质上是分组汇总在双变量图中的应用
  - 同样可加绘回归曲线及可信区间等



数值变量 vs. 数值变量-对原始数值做适当汇总

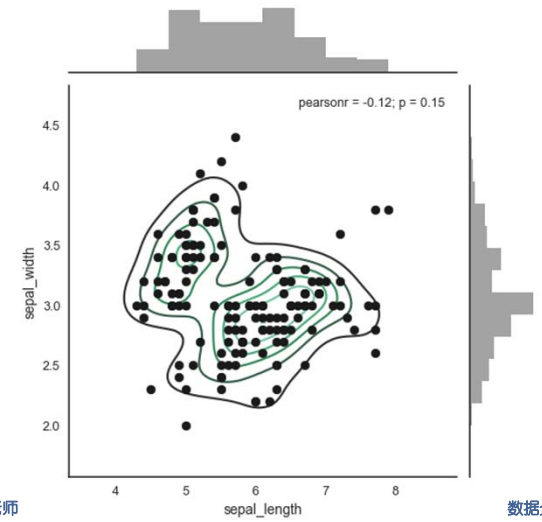
- KDE图/等高线图
  - 本质上是核密度估计在双变量中的应用
  - 同样可加绘回归曲线及可信区间等
  - 可绘制为连续颜色变化

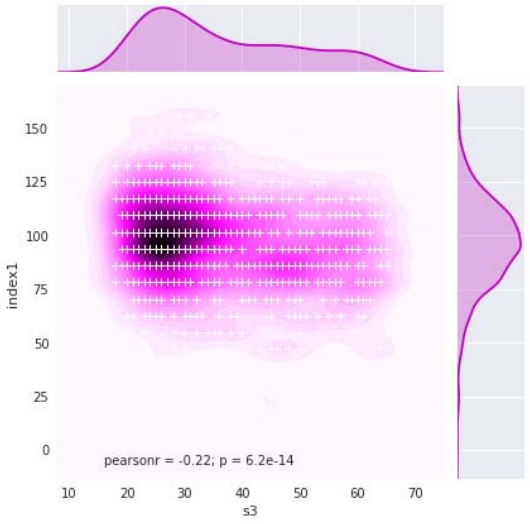




数值变量 vs. 数值变量-对原始数值做适当汇总

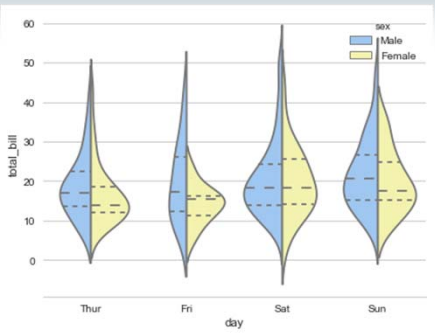
- 前述各种图形的组合

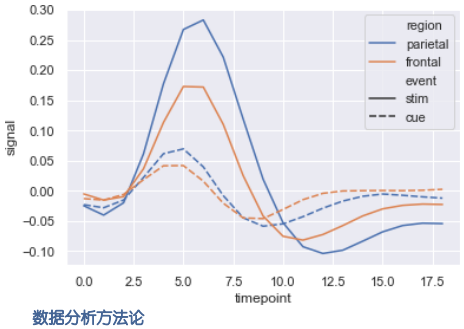




双变量图 + 更多的分类变量

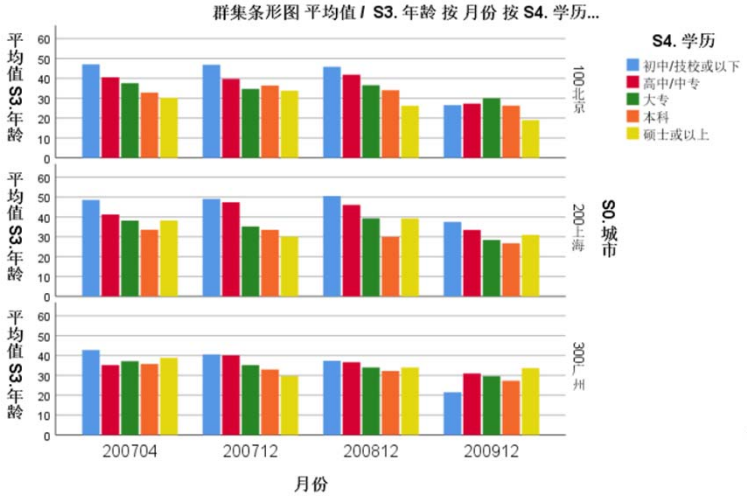
- 可以考虑提供Z轴作为新的分类轴
  - 但这显然不是首选方案
- 采用图例对二维图进行扩充
  - 一般首选颜色图例
  - 分组线图、条图、直方图等
- 采用更多的图形元素对信息继续进行扩充
  - 直条/线段宽度
  - 线形/散点形状
  - 直条填充图案





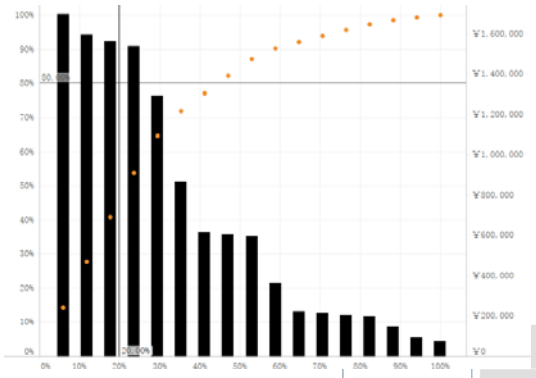
双变量图 + 更多的分类变量

- 行面板/列面板图组/更复杂的不等距图组组合



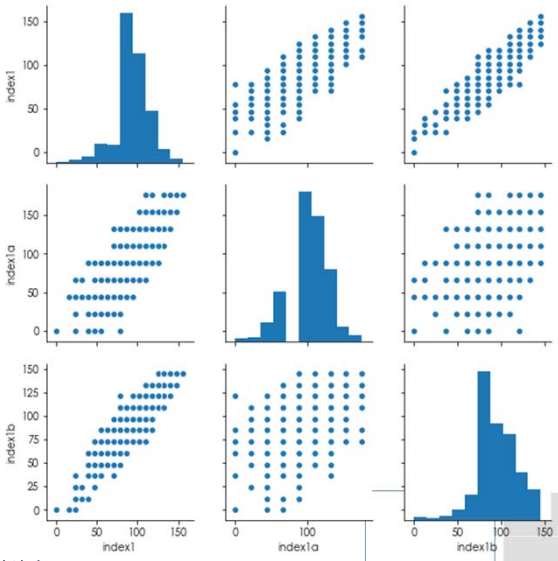
多个连续变量相互比较/同时呈现

- 将多个连续变量看成一个变量，虚拟一个新的分组变量用于区分
  - 绘图需求直接转化为“双变量图 + 更多的分类变量”类型
- 组合统计图：根据实际需要自行设计
  - 条图/线图/面积图可自由进行组合
  - 最常见的是线图/条图组合
- 双轴图：
  - 提供两个纵轴尺度，便于对比数值相差较大的两个指标



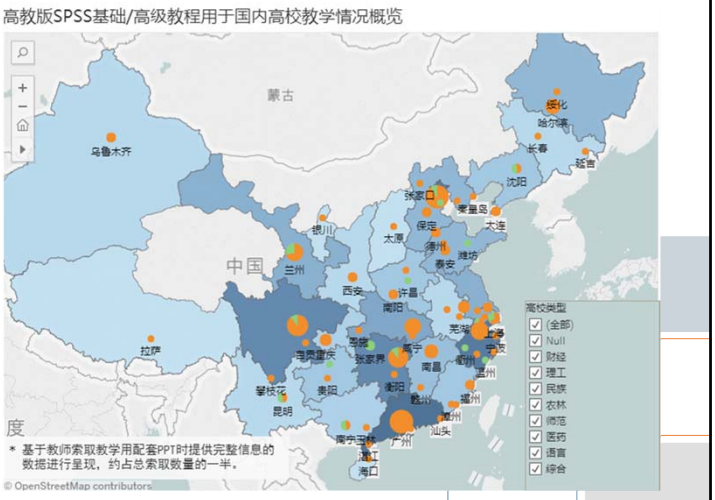
多个连续变量间的关系呈现

- 三维图形，但直接观察比较困难
  - 不作为首选方案
- 图形矩阵
  - 主对角线呈现单变量分布
  - 非主对角线呈现联合分布
  - 也可绘制为非对称矩阵



与地图数据相结合：统计地图

- 相对而言Tableau的功能更为专业
- 可自定义地图数据



其余更复杂的图形：根据需求自行组合设计

- 各种异化的条图
  - 甘特图：反映项目进展是否按照时间计划进行
  - 标靶图：在条图的基础上增加了目标值，可反映任务完成情况
  - 人口金字塔、漏斗图、k线图、瀑布图。。。
- 异化的频数图
  - 词云：用于直观反映各词汇在语料库中的出现频次
  - 热图：用颜色代表每个单元格的频数多少
- 雷达图、凹凸图。。。