清华大学数据库技术与应用

数据可视化 II

授课教师：计算机系王健楠

授课学期：2026年（春季）

课程议程

课程议程

清華大學 Tsinghua University

01 多变量关系

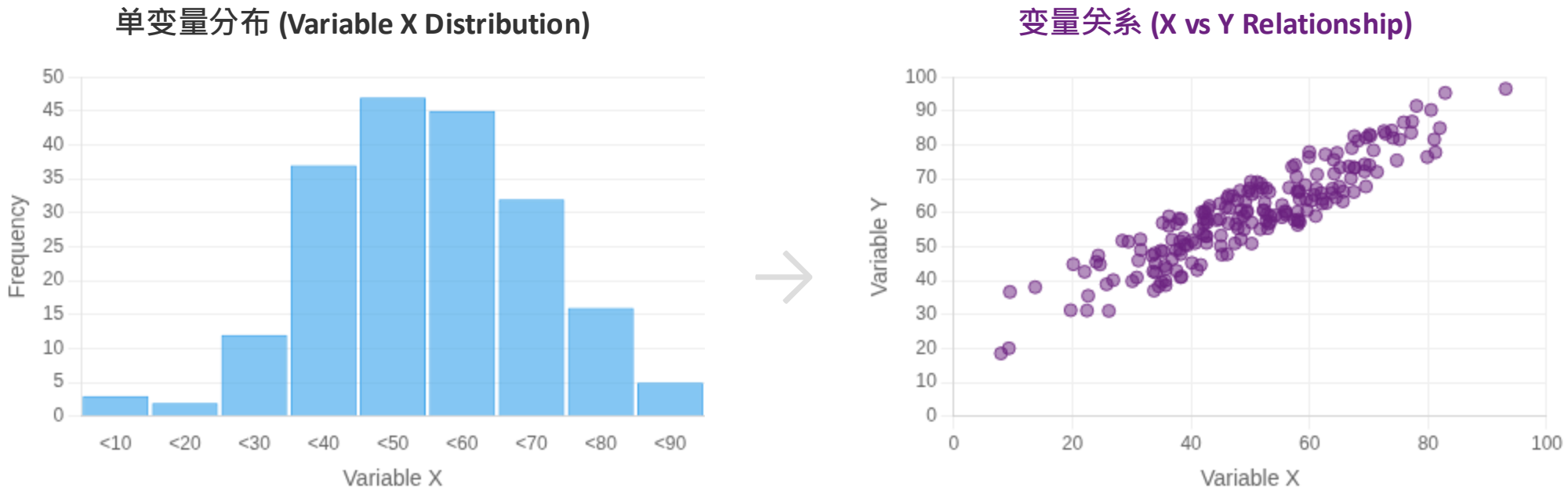
02 数据转换

03 可视化原则

从“分布”到“关系”

回顾：单变量分布 —— **之前主要关注**单个变量的特征可视化（直方图、密度图、箱线图）

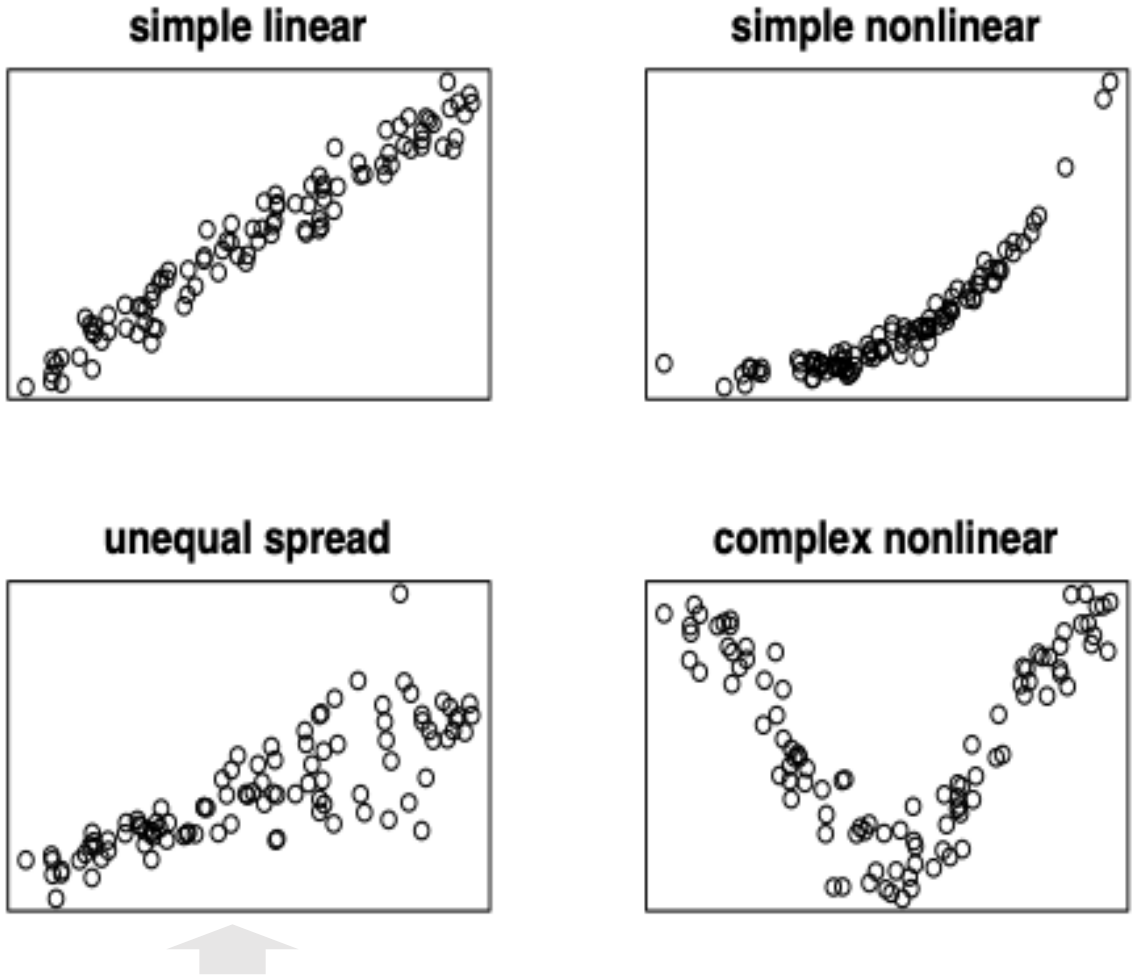
进阶：变量间关系 —— 现在的重点是可视化**两个或更多**变量**之**间的关系（联合变化）



散点图

散点图用于揭示**两个数值变量**之间的关系。

- 图中的左侧关系更接近线性，右侧则明显更偏非线性。

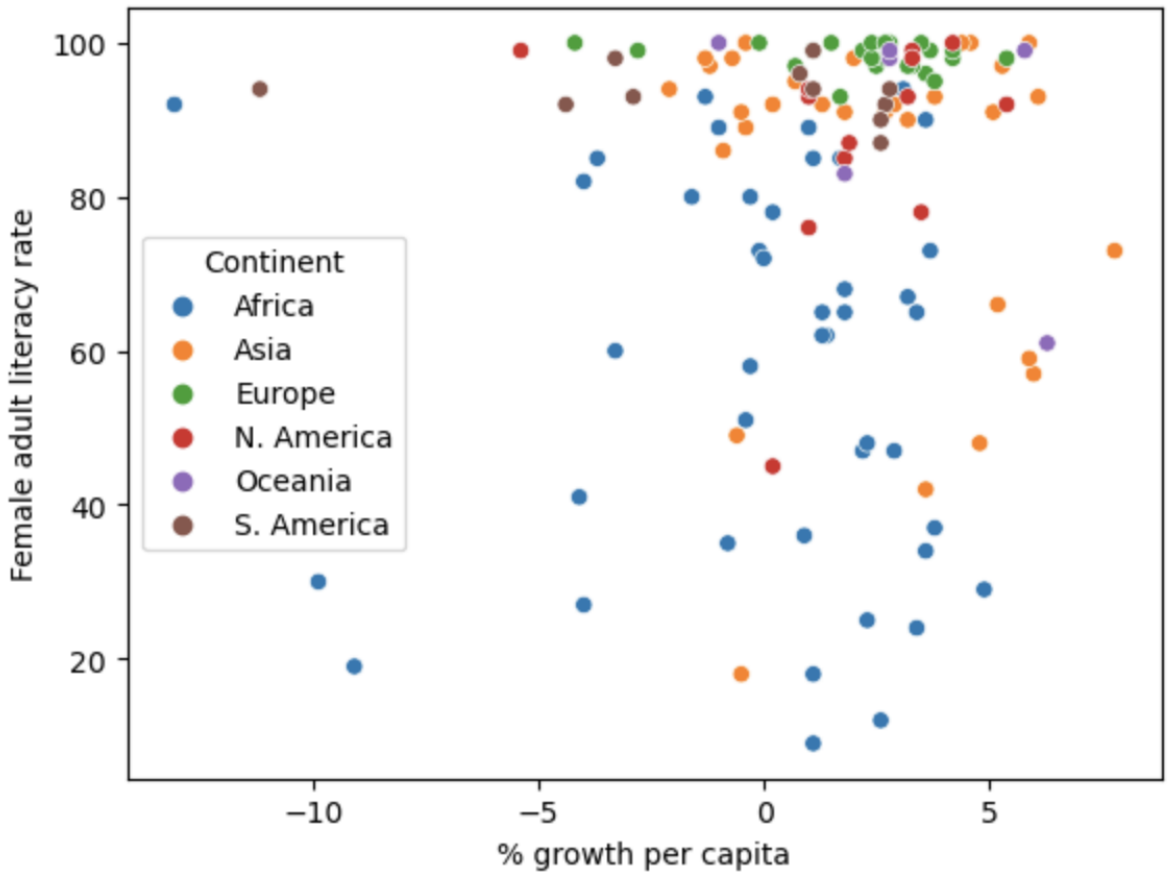


这种关系看起来是线性的，但随着 x 变大，数据的分散程度也在增大。

散点图的基本用法

一张最基本的散点图，需要明确三件事：

- 横轴画哪个变量
- 纵轴画哪个变量
- **一个点究竟代表什么**观测
  - **如果**还要编码类别信息，可以**再引入**颜色（hue）




```
plt.scatter(x_values, y_values)
```

```
sns.scatterplot(data=df, x="x_column", \
                y="y_column", hue="hue_column")
```

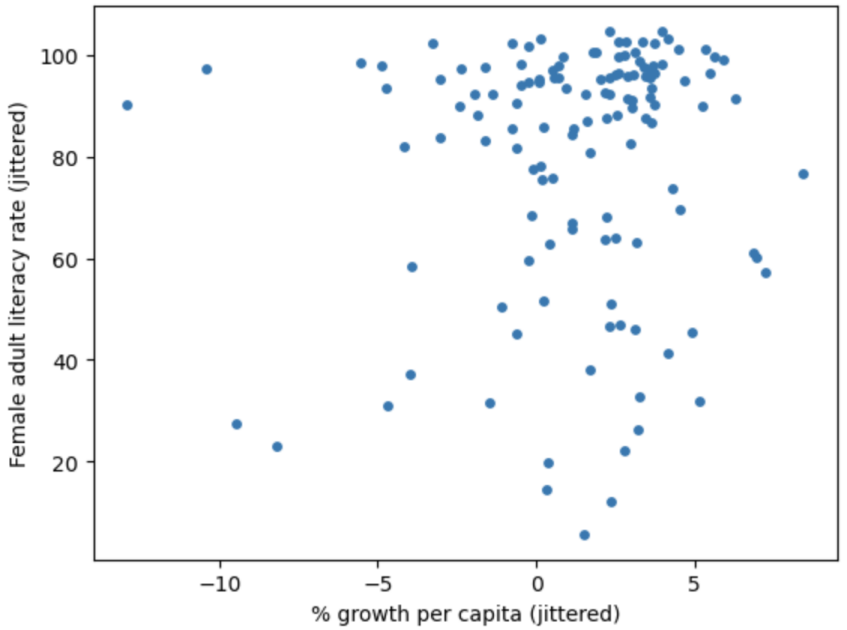
过度重叠（Overplotting）

问题：点太多，堆在一起看不清?

- **当很多点堆叠在一起**时，散点图会变得难以阅读，这**就是**过度重叠。
- **常**见处理方法是加入轻微抖动（jittering），让原本重**合的点稍微分开**。
- 这样不会改变总体趋势，却能显著提升单个样本的可辨识度。

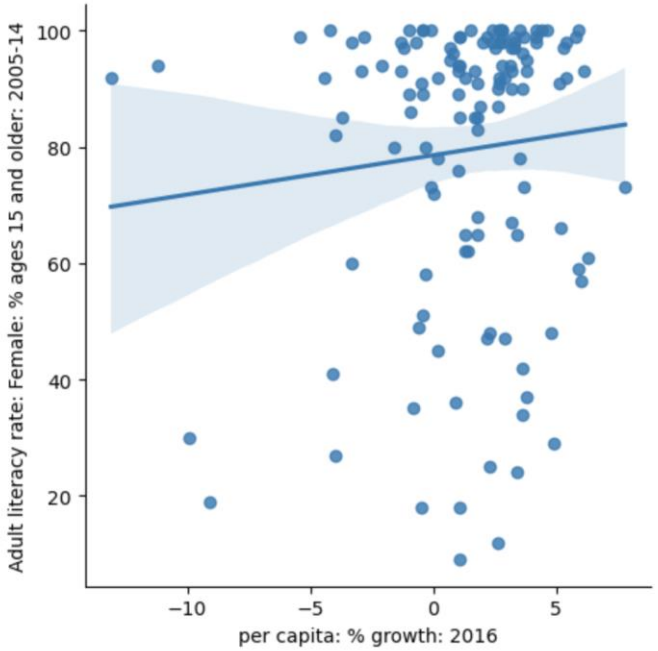
```
x_noise = np.random.uniform(-1, 1, len(wb))
y_noise = np.random.uniform(-5, 5, len(wb))

plt.scatter(wb['% growth'] + x_noise, \
            wb['Literacy rate: Female'] + y_noise, \
            s=15);
```

补充：减小点大小也有帮助；其中`s`**控制 marker 的尺寸。**



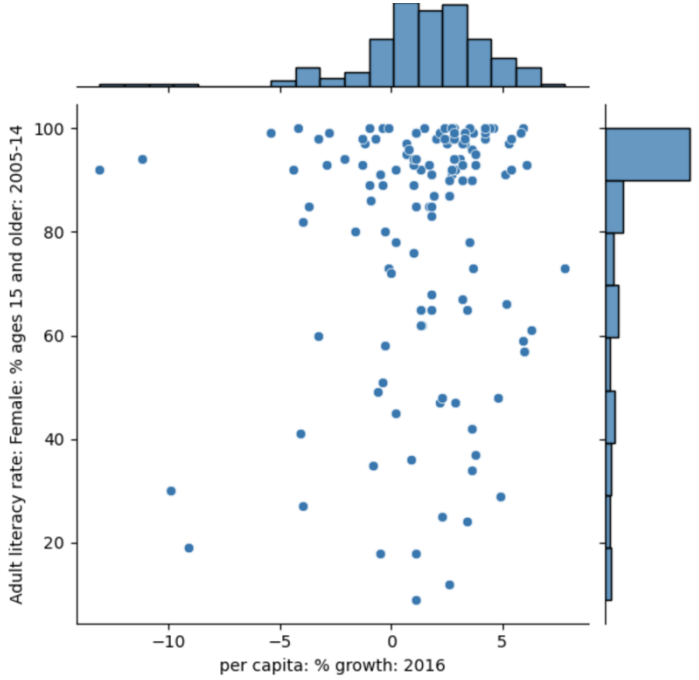
散点图的替代方案

如果只画散点还不够，Seaborn 还提供了更高层次的关系图接口。




```
sns.lmplot(data=df, \
x="x_column", y="y_column")
```

lmplot：**直接叠加回**归线与置信区间。



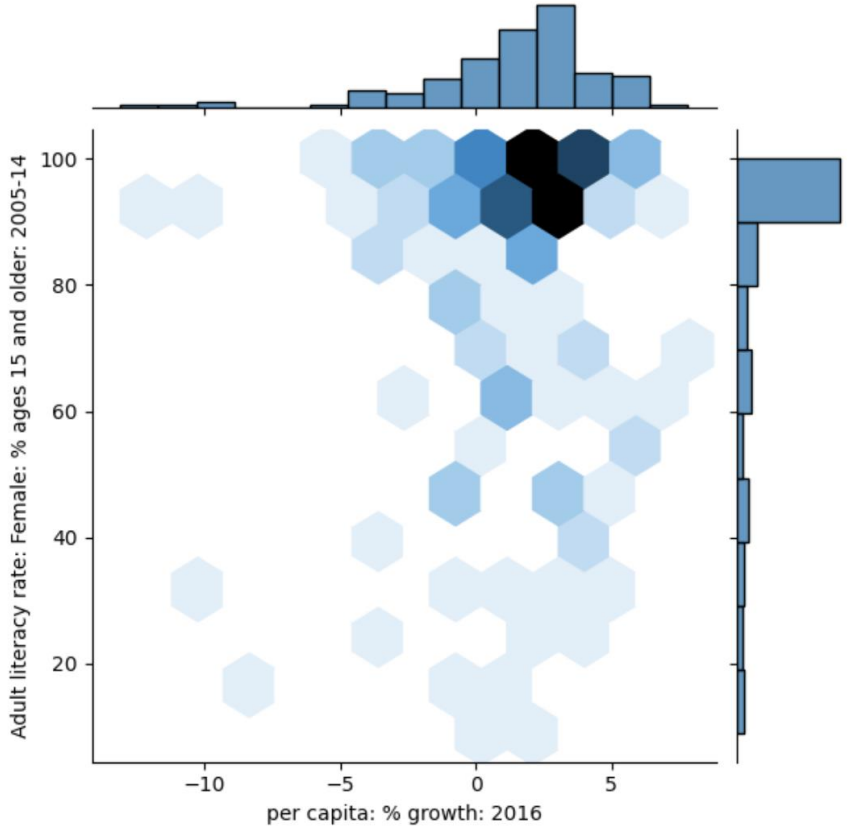

```
sns.jointplot(data=df, \
x="x_column", y="y_column")
```

Hex 图（六边形分箱图）

当点太多时，不一定非要逐点绘制。

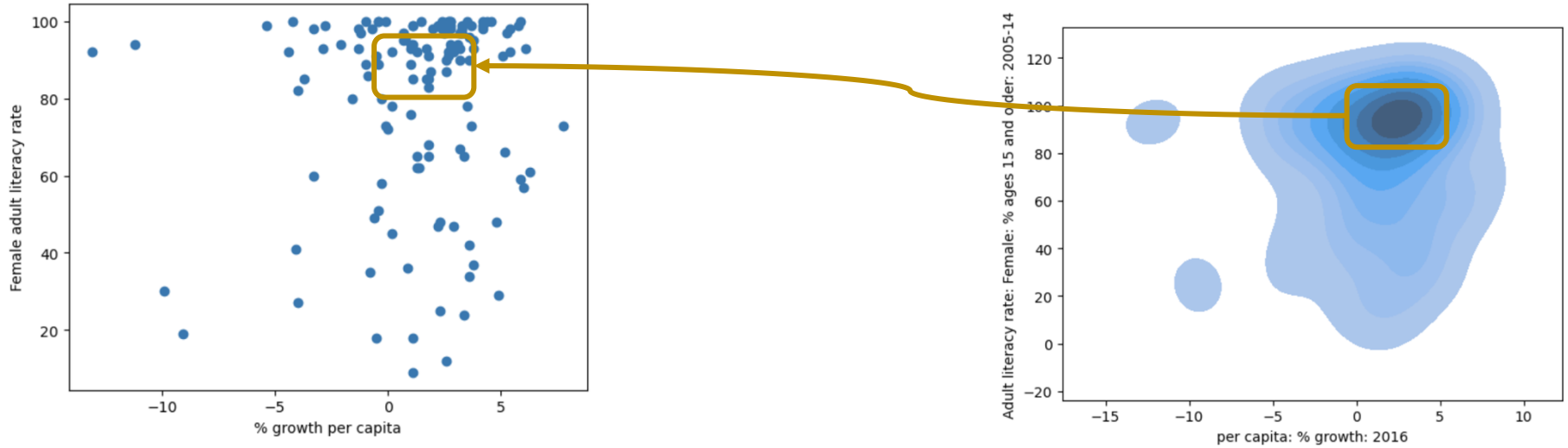
- Hex 图可以看作二维直方图。
- **它把 x-y 平面划分成**许多六边形格子，再统计每个格**子里的点数**。
- 颜色越深，通常表示该区域的数据越密集。

```
sns.jointplot(data=df, x="x_column", \
y="y_column", kind="hex")
```



密度等高线图

- **密度等高**线图可以理解为二维 KDE 图。
- **它**类似地形图：同一条等高线上的位置代表相近的数据密度，颜色越深，表示这一带的样本越多。



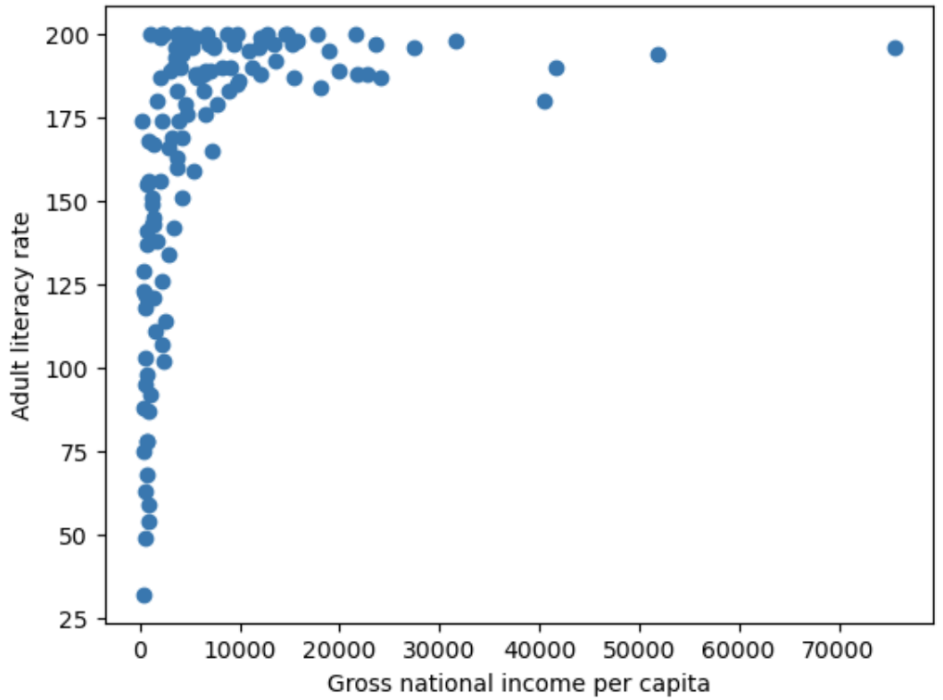
原始散点图　　转换为二维密度图

```
sns.kdeplot(data=df, x="x_column", y="y_column", fill=True)
```

课程议程

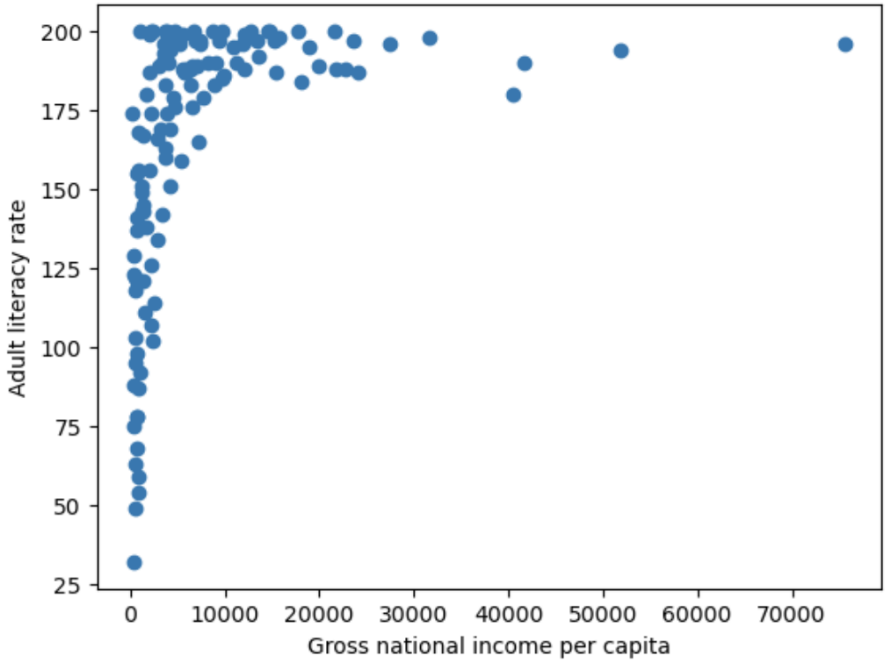
可视化目标

- 回顾一下，可视化至少有两个目标：帮助我们自己理解数据，以及帮助别人理解结论。
- 图形的选择和绘图前的数据准备方式，会直接影响这两个目标能否实现。
- 有些散点图即使做了抖动，信息仍然挤在一起，难以读出清晰关系。
- 因此，我们常常会先对数据做变换，再去可视化。

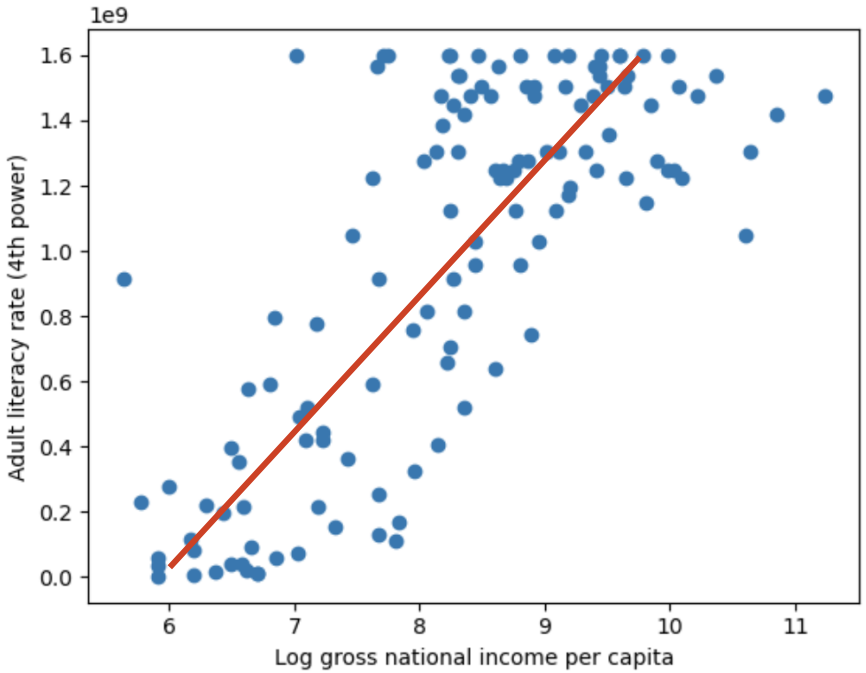


线性化（Linearization）

- 做变量变换时，一个常见目标是把关系“线性化”，**也就是**让 x **和** y **更接近直**线关系。
- 为什么要这样做？因为线性关系最容易解释；斜率和截距都有清晰含义，而且后续线性模型也更容易发挥作用。



原始关系

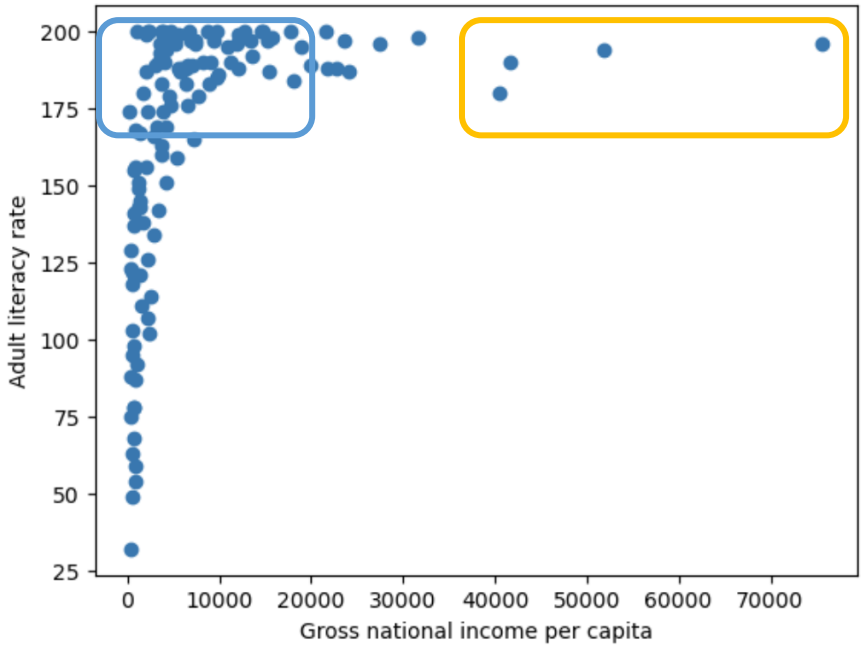


更接近线性的关系更容易解释

数据变换的应用

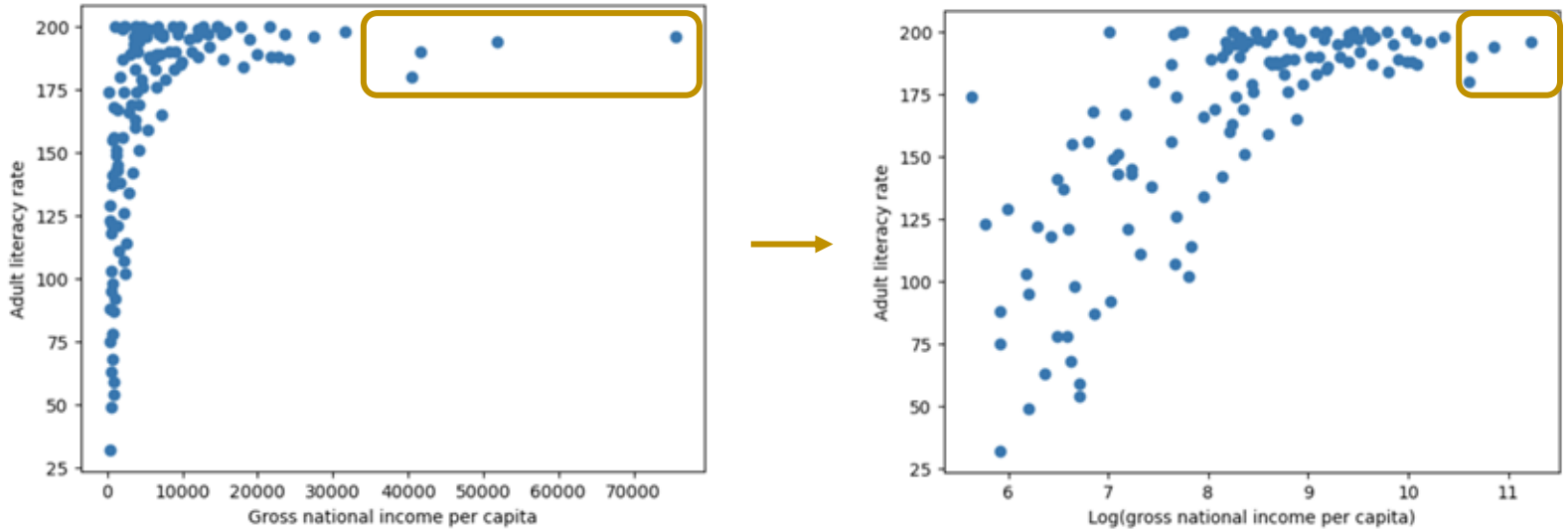
这张图为什么看起来不像线性关系？

- 原因 1：**少数特别大的 x 值把横**轴拉得很长，压缩了大多数样本所在的区域。
- 原因 2：许多**较大的 y 值**挤在一起，使纵轴上半**部分的信息很**难分开。
- **也就是**说，不一定是“**关系本身非**线性”，**也可能是坐**标尺度把结构挤变形了。



数据变换的应用

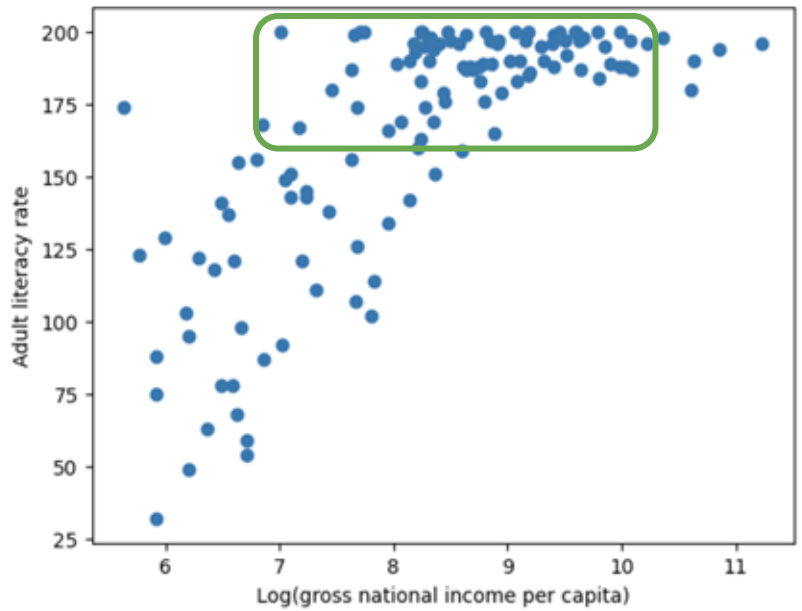
- **先看** x 轴的问题：少数极大的 x 值把横轴尺度拉开了。
- **一个自然的解决**办法是对 x **做** log 变换
  - **大数取**对数后会缩小得更明显
  - **小数**变化相对没那么剧烈



数据变换的应用

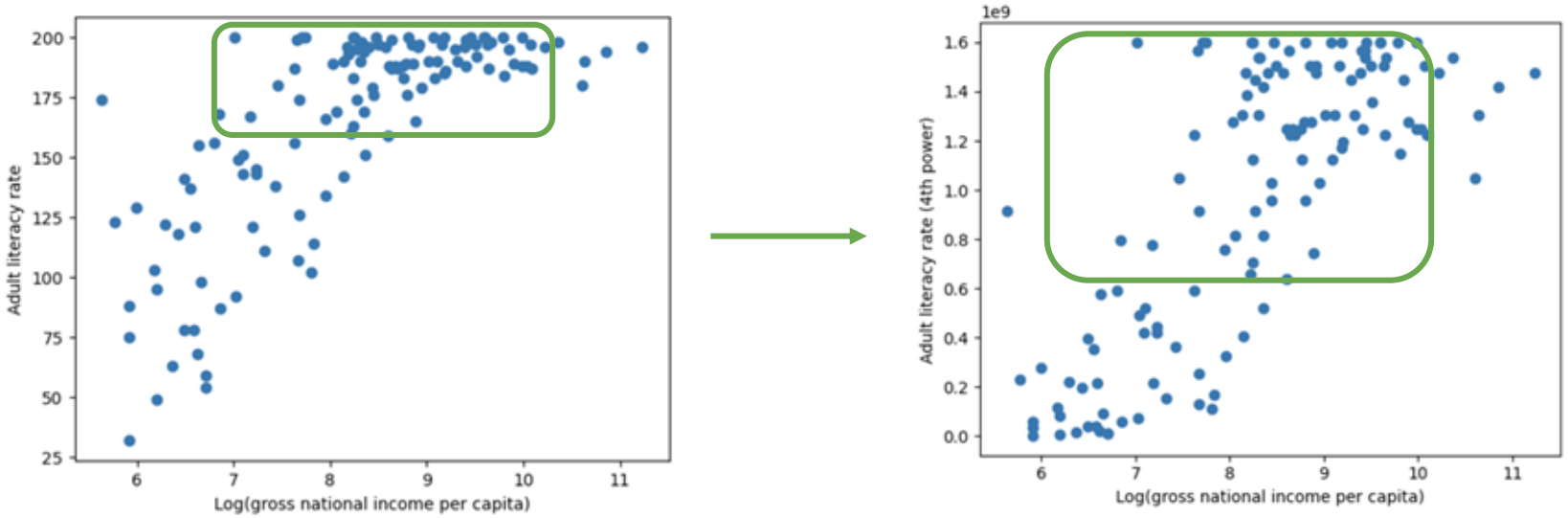
解决了 x 轴之后，y 轴的问题仍然存在。

- 可以看到，**高端的 y 值仍然明显扎堆**，导致纵向差异被压缩。
- 这说明除了 x 轴外，y 轴本身也需要进一步的尺度调整。
- 下一步，我们就考虑如何把较大的 y 值拉开。



数据变换的应用

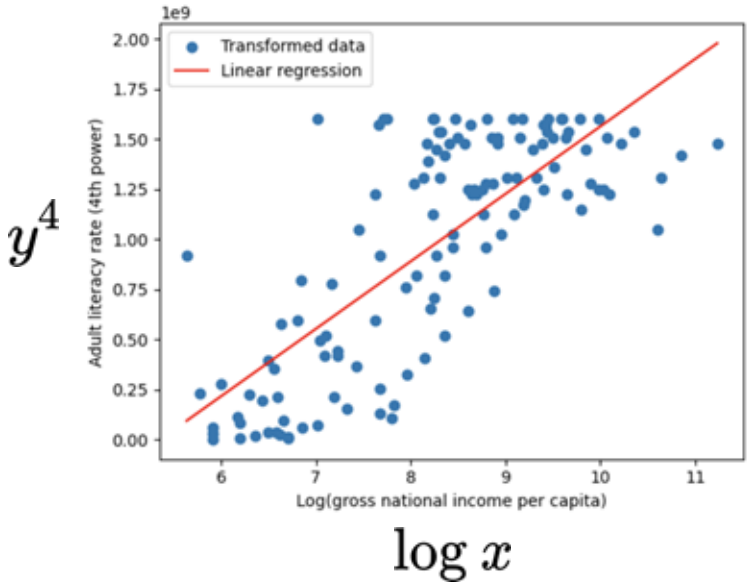
- 对 y 的一个常见处理方式是做幂变换，例如平方、立方或更高次幂。
- 大数经过幂变换后会增长得更快，因此更容易被拉开；小数的相对变化则没有那么明显。

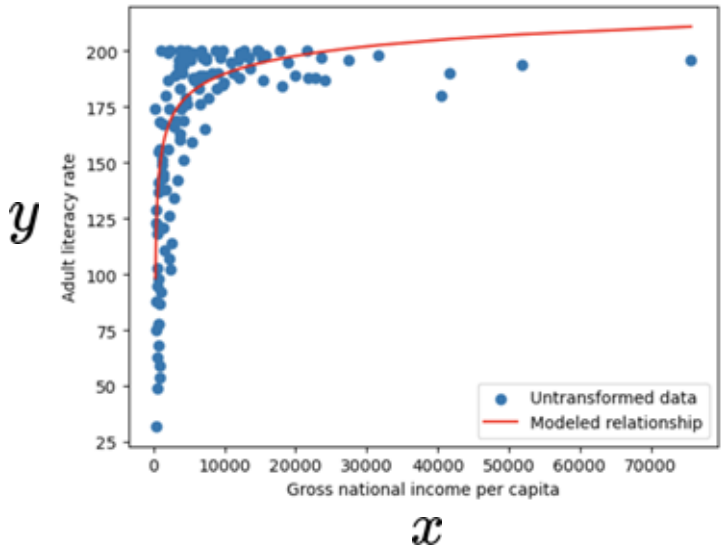


如何理解变换后的数据

- 现在，变换后的变量之间已经呈现出更接近线性的关系。
- 这说明：原始 x 和 y 之间其实存在一个可解释的底层关系，只是原始坐标把它隐藏了。

$$y^4 = m(\log x) + b \longrightarrow y = [m(\log x) + b]^{1/4}$$

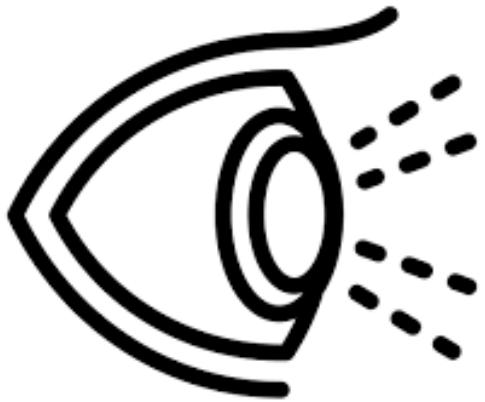


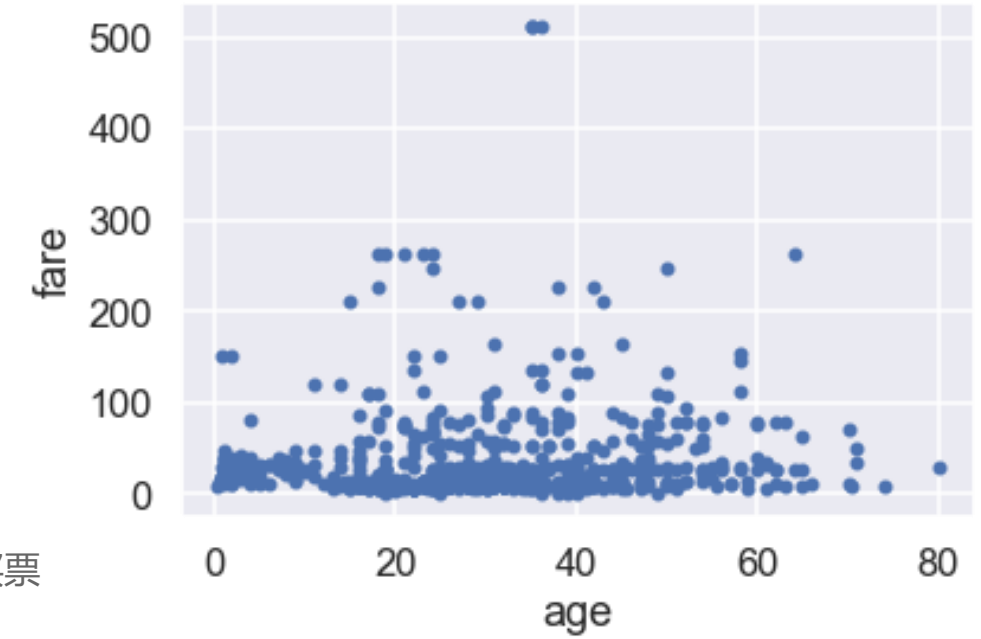


课程议程

可视化原则

可视化是给人看的

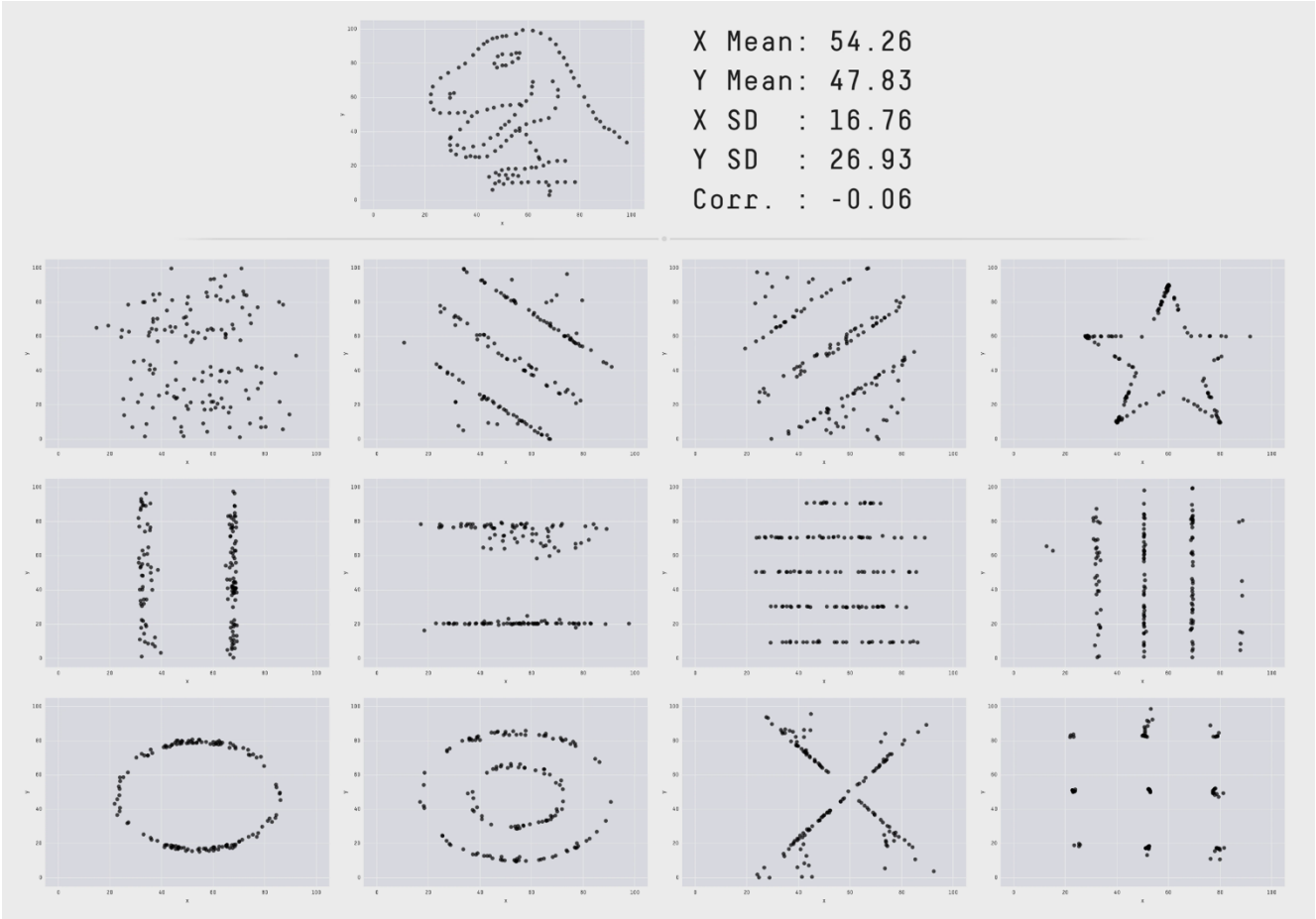




看起来年纪更大的乘客并没有在泰坦尼克号上花更多钱买票

可视化比摘要统计量更有表现力

- **每个数据集的均**值、标准差、相关系数都一样，**但**图形形状完全不同。
- 统计量和可视化互相补充，而不是二选一。



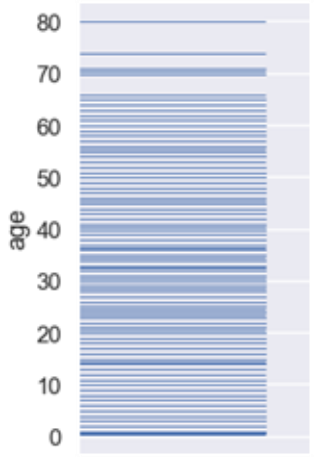
https://www.autodesk.com/research/publications/same-stats-different-graphs

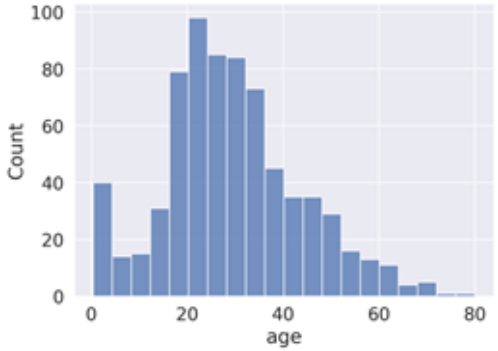
信息通道

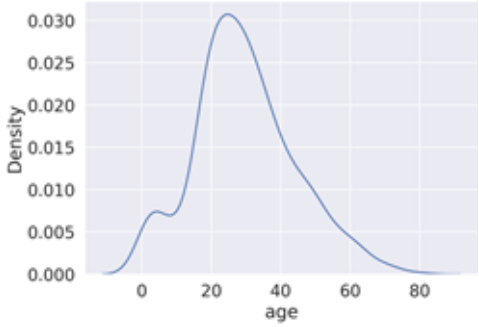
充分利用人的视觉感知系统

- 同一份数据可以有很多种画法。
- 理解图形中的信息通道，有助于我们拆解“一个图到底是如何表达数据的”。

	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0

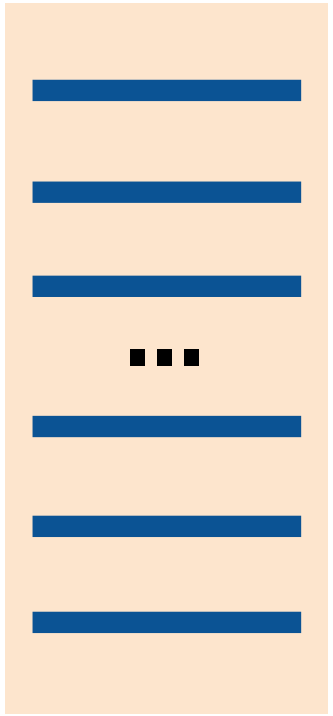


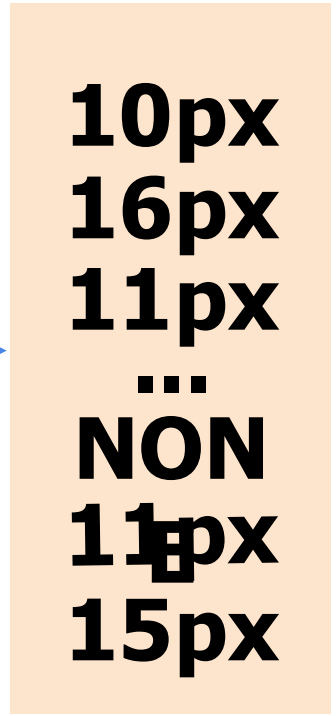


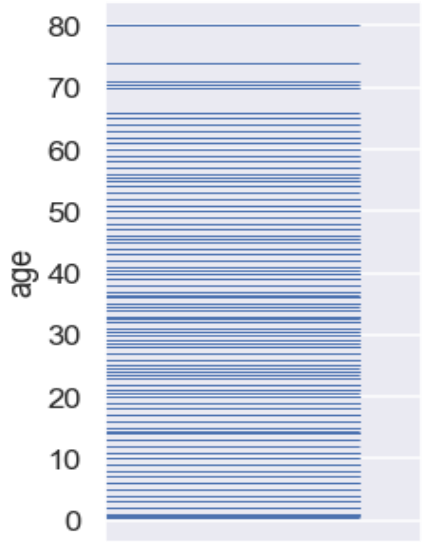


地毯图：编码 1 个变量

	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0





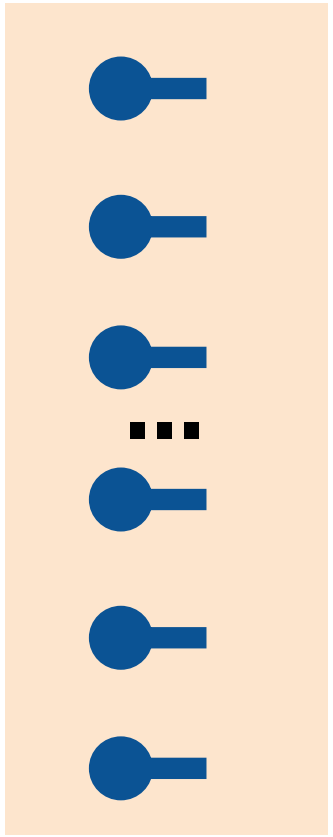


标记
(表示一个数据点)

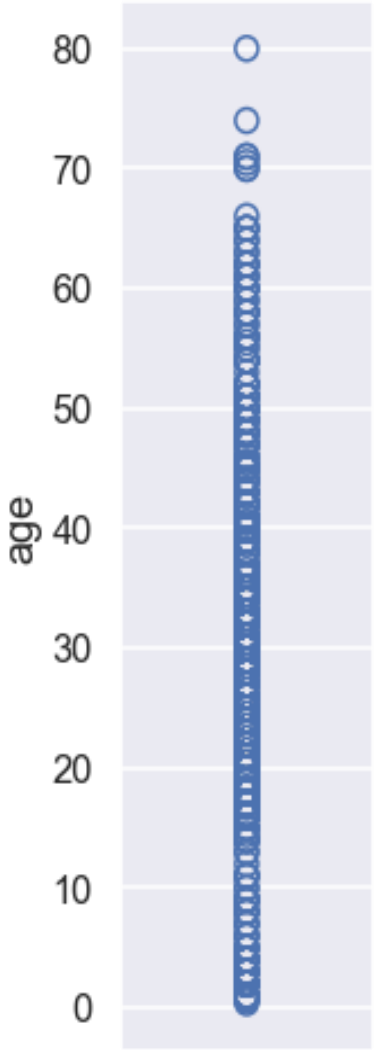
编码
(把数据值映射到视觉位置)

地毯图：不同的标记

	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0



10px
16px
11px
...
NON
11px
15px

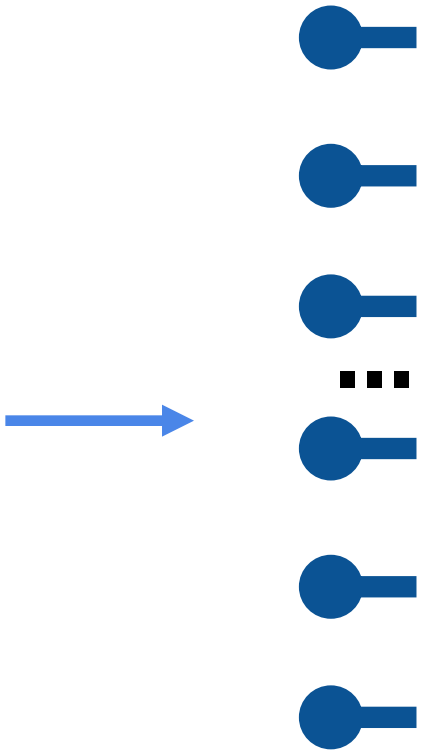


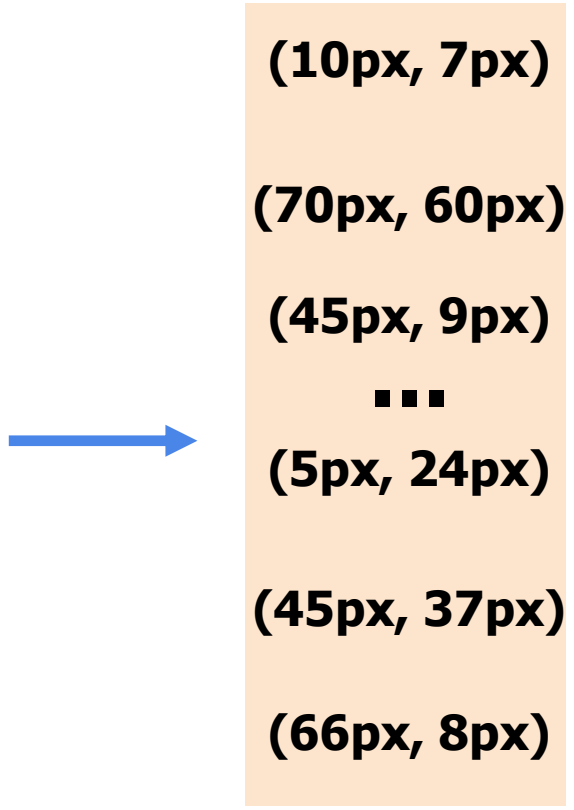
标记
(表示一个数据点)

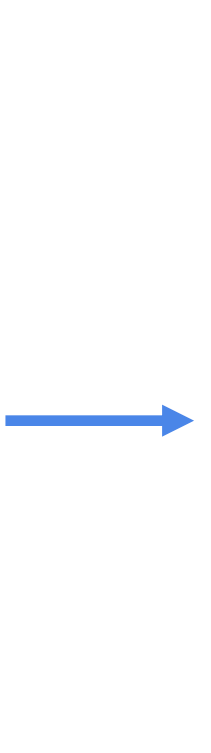
编码
(把数据值映射到视觉位置)

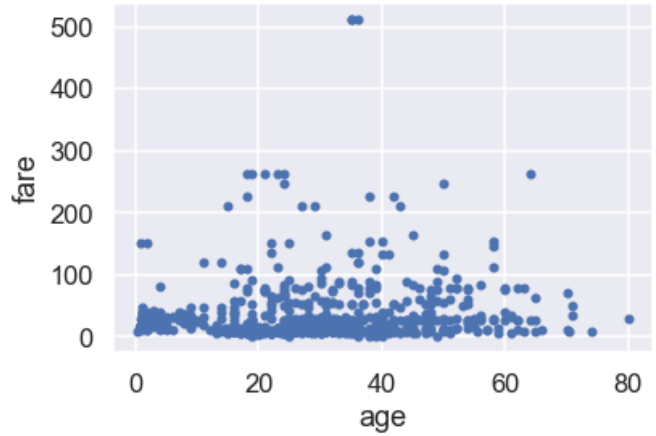
散点图：编码 2 个变量

	age	fare
0	22.0	7.25
1	38.0	71.28
2	26.0	7.92
...	...	...
888	NaN	23.45
889	26.0	30.00
890	32.0	7.75









标记
(表示一个数据点)

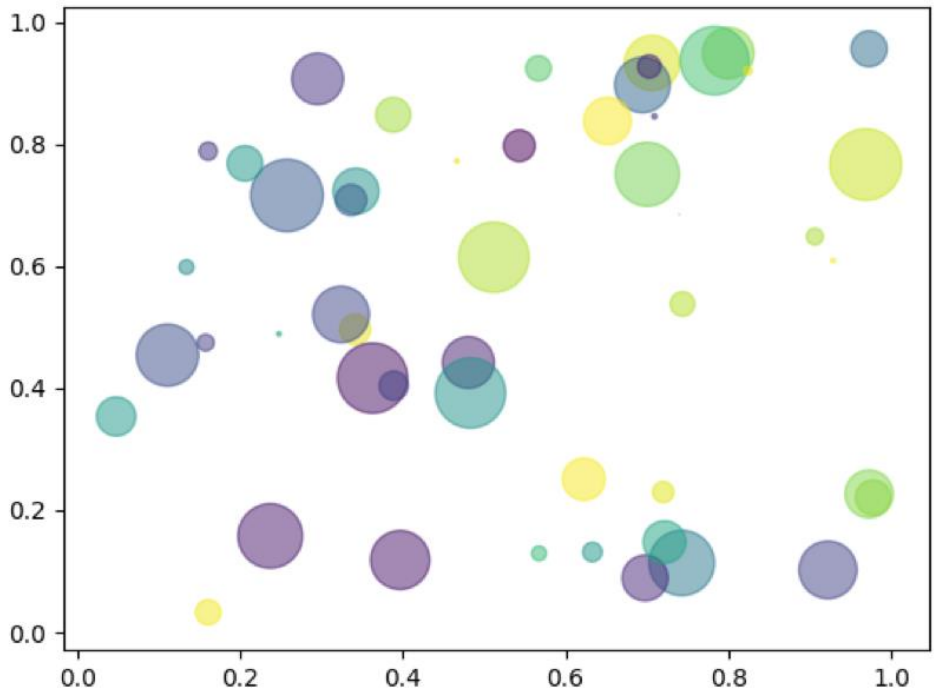
编码
(把数据值映射到视觉位置)

再进一步：编码 3 个以上变量

清華大學 Tsinghua University

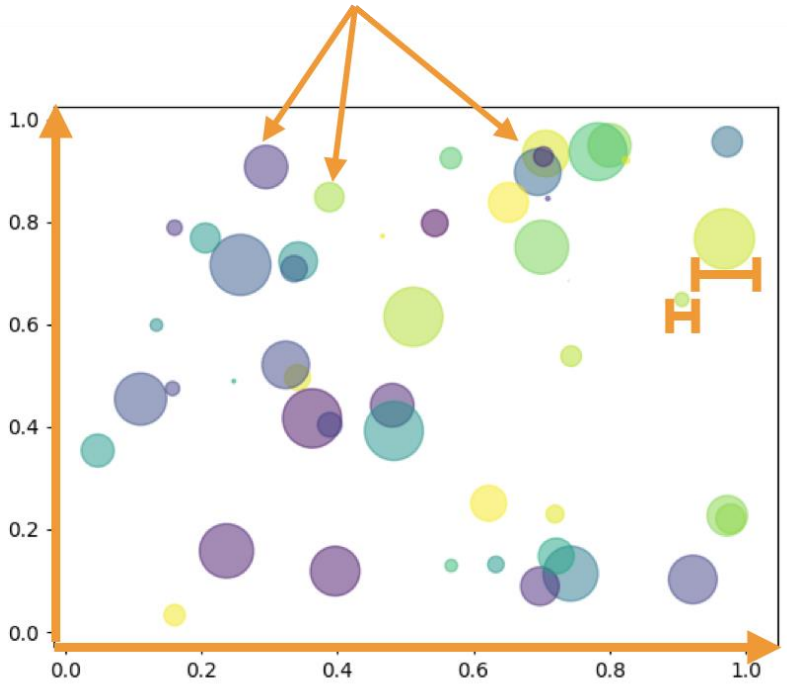
这个图除了 x 和 y 之外，还额外编码了哪些信息通道？

也就是：我们到底同时表达了几个变量？



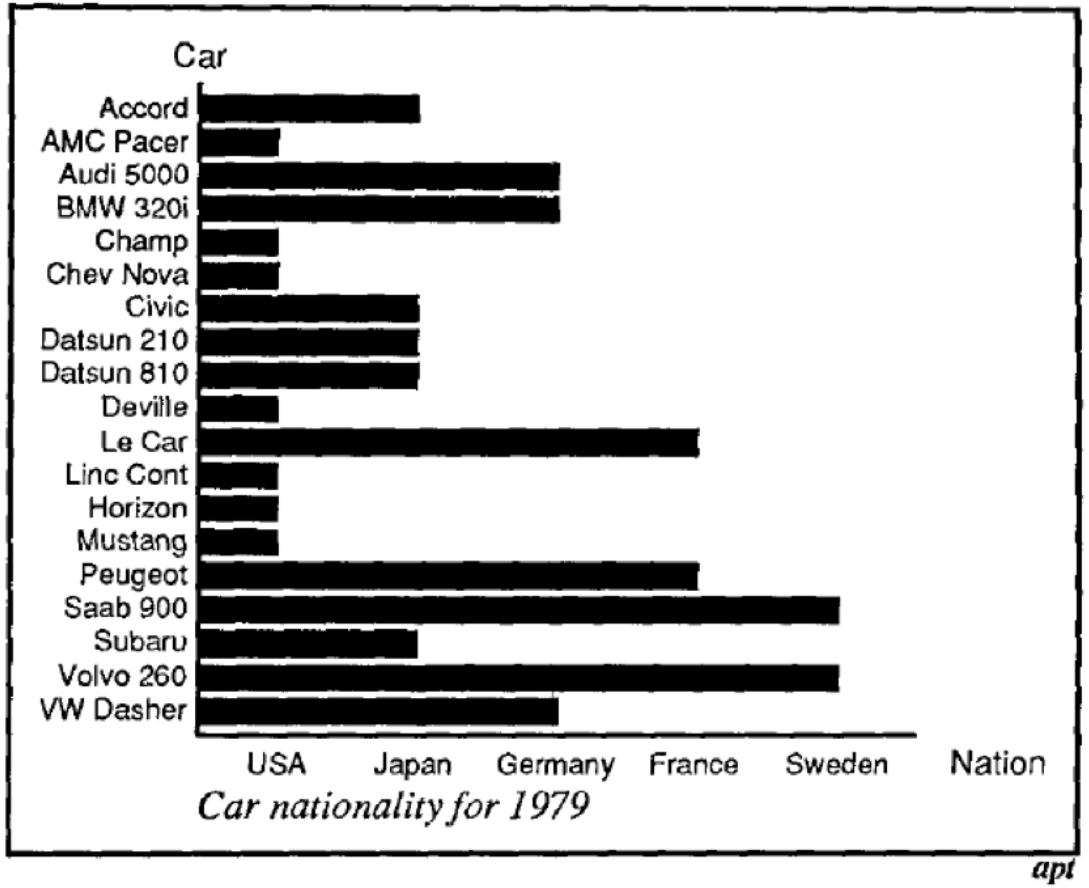
答案：这里编码了 4 个变量

- 还可以继续增加**这张图编码了 4 个变量：x 位置、y 位置、面积、颜色。**
- **通道，例如形状、**边框颜色、透明度、纹理等。
- **但通道不是越多越好。通道增加后，**读者的认知负担也会同步增加。



编码误用：长度

- 可视化里有很多地方会“**画得出来，但表达得不好**”。
- 下面这个例子就滥用了**长度通道**，导致读者容易产生错误直觉。
- 接下来我们会依次讨论：如何更合理地使用：
  - x/y
  - 颜色、
  - 标记
  - 条件化
  - 上下文

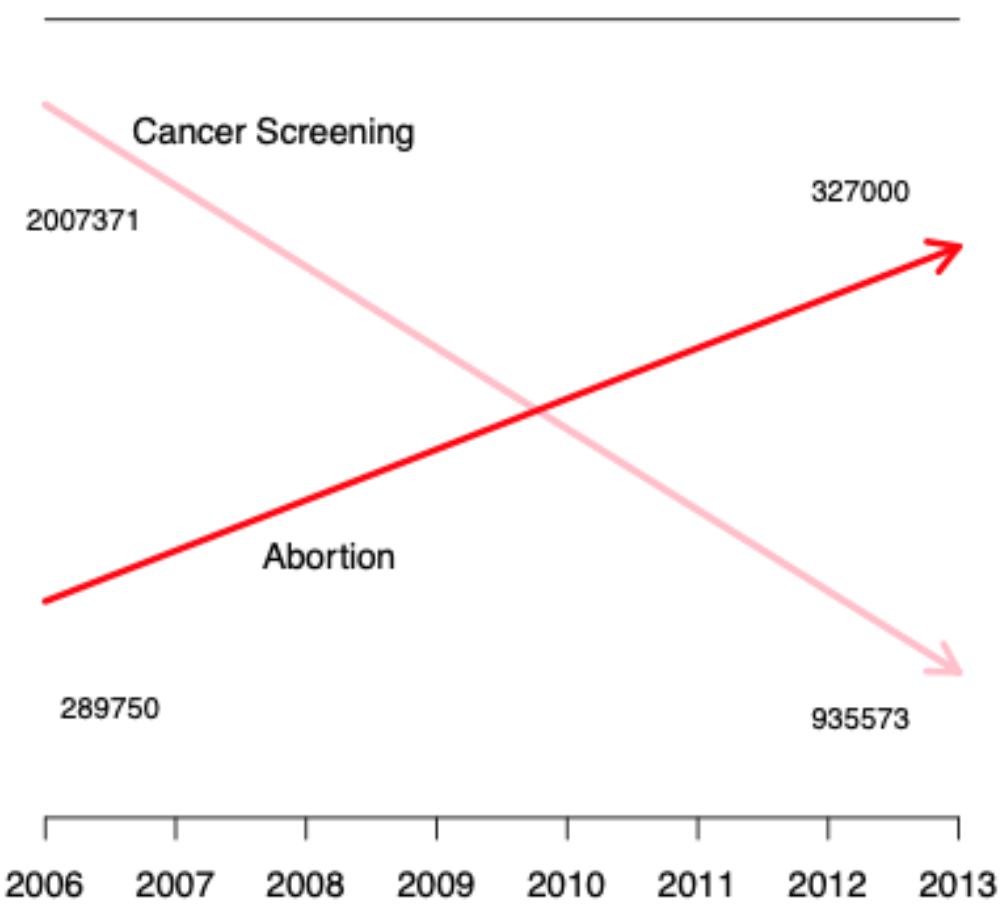


https://info.luddy.indiana.edu/~katy/S637-S11/Mackinlay86.pdf

善用 X/Y 通道

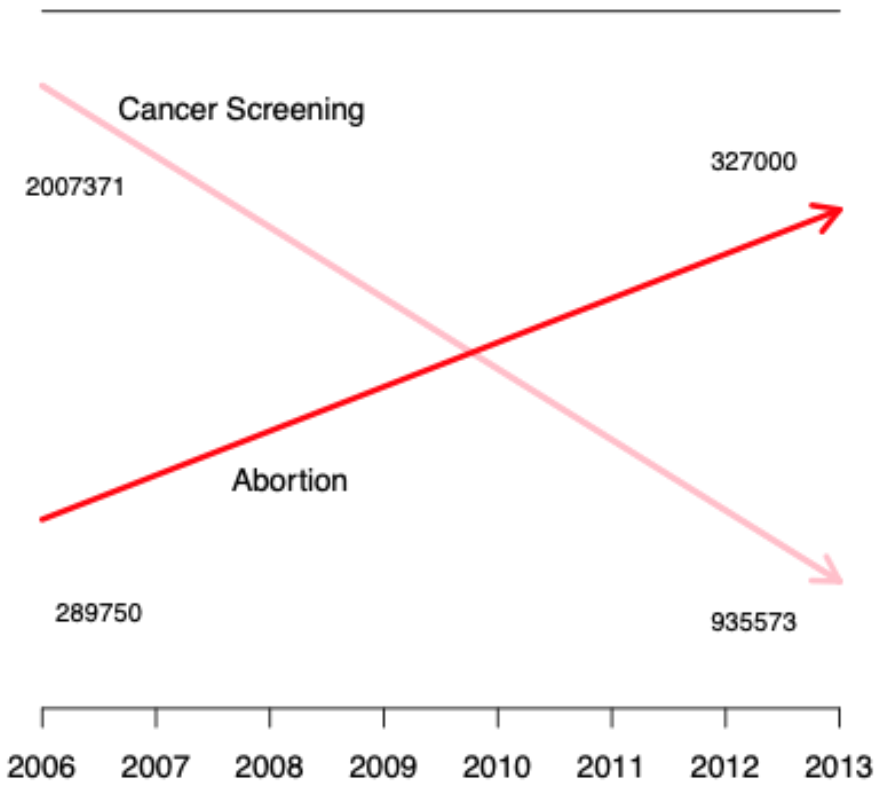
案例：Planned Parenthood 听证会

- 2025年，Planned Parenthood 被指控从流产胎儿组织中牟利
- 国会议员展示了这张图
  - 这图画的是什么?
  - 它想传达什么信息?
  - 有没有可疑之处?



保持坐标轴刻度一致

- 这张图的问题在于，两条线对应了完全不同的纵轴刻度。
- 结果是：明明数值更小的量，视觉上却显得更大

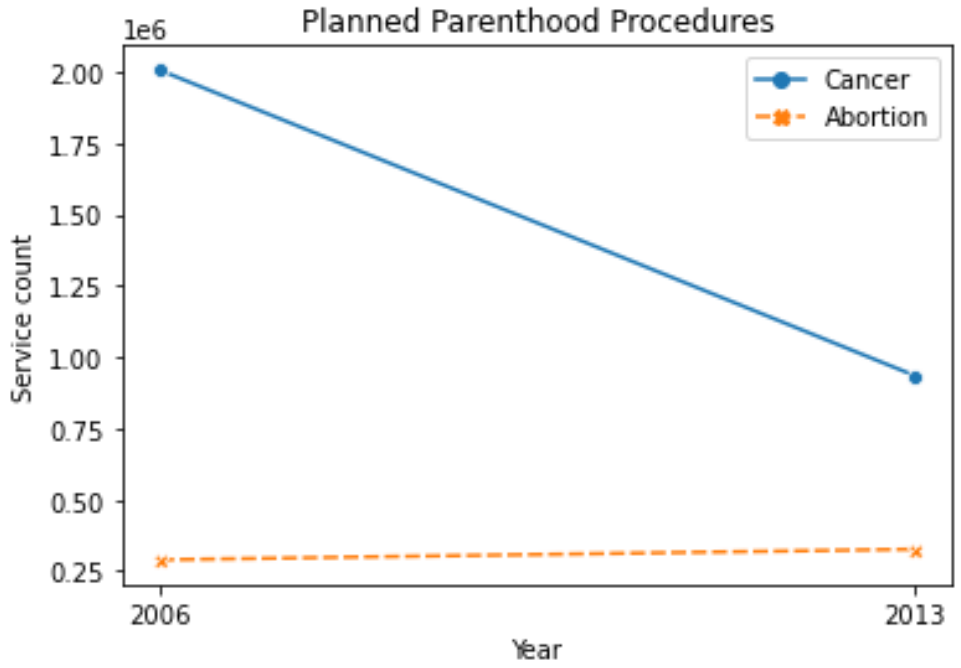


原则：同轴同义的数据，应尽量使用相同尺度。

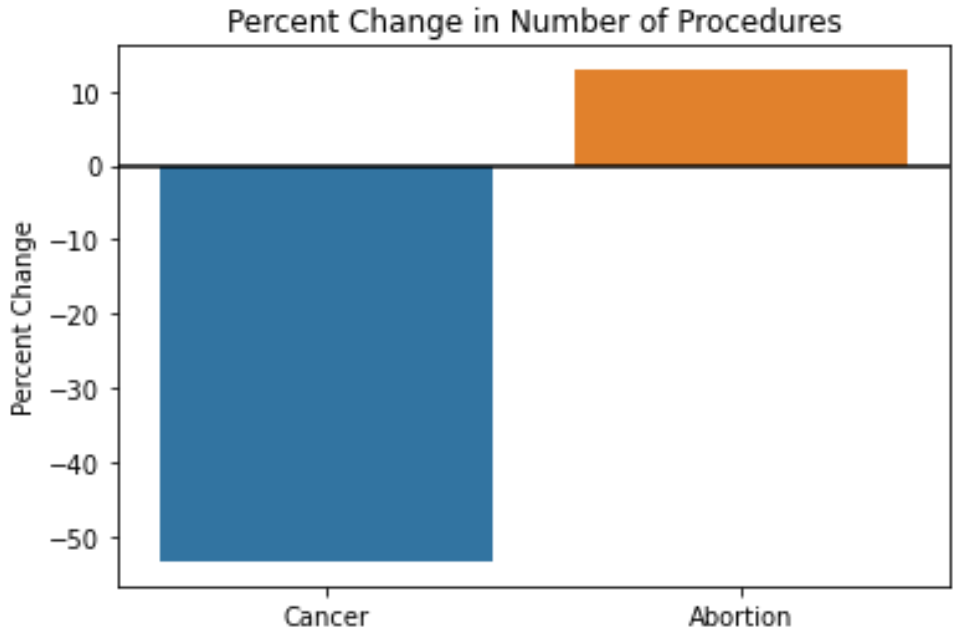
比较“同类”数据时一定要看尺度

- 上图把所有数据画在同一尺度下，可以清楚看到癌症筛查数量远高于堕胎数量。
- 下图改画百分比变化，也能表达“筛查减少、堕胎增加”，**但不会**误导读者。

```
sns.lineplot(data=ppdf, markers=True)
```



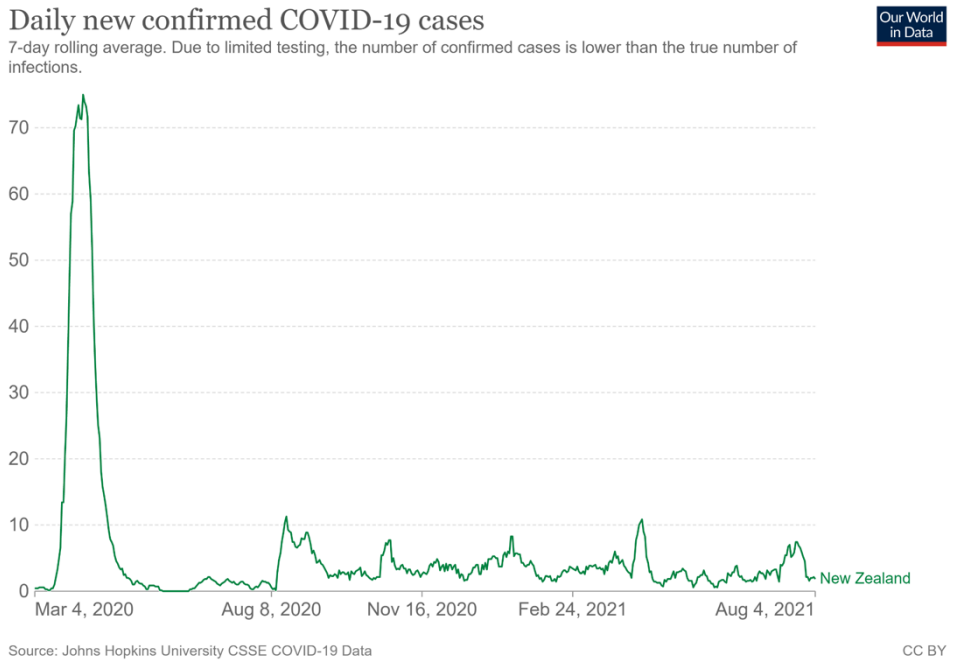
同一尺度下比较原始数量

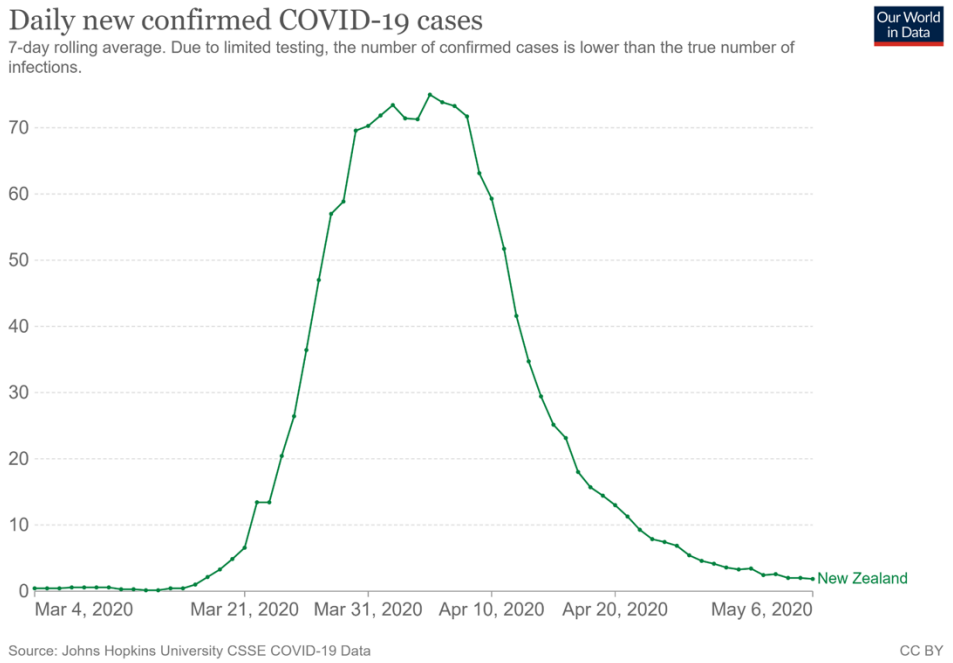


改看百分比变化

让数据真正显现出来

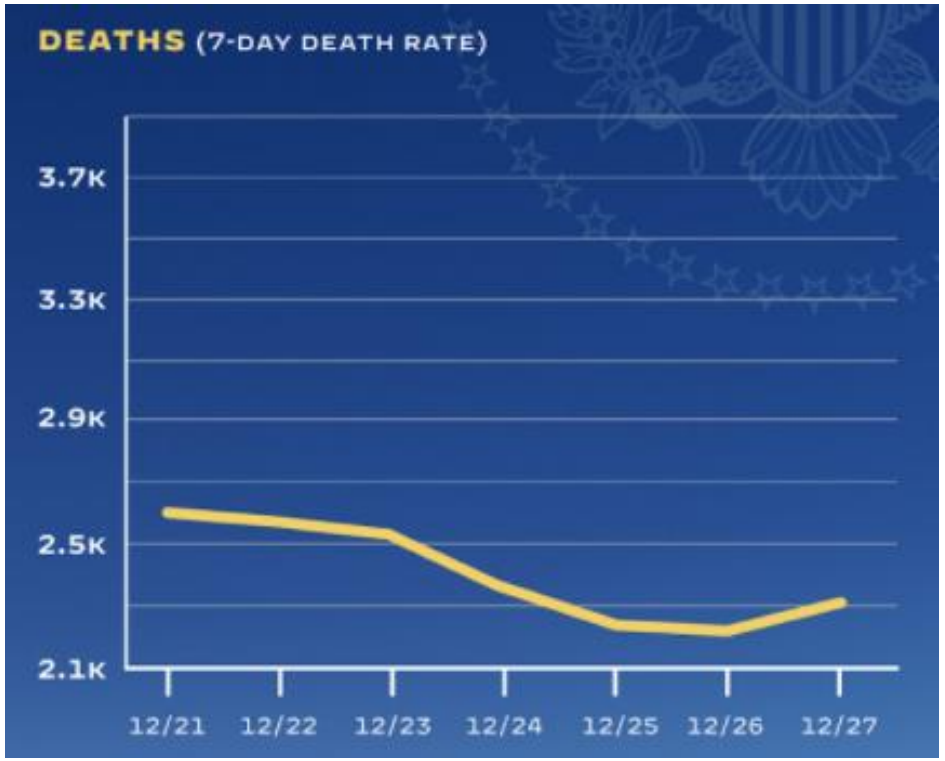
- 建议之一：坐标轴范围要尽量让关键信息充满图面，而不是把重要部分挤成一团。
- 不必一次展示全部数据；如果你真正关心的是局部区域，适当放大是完全合理的。





让数据真正显现出来

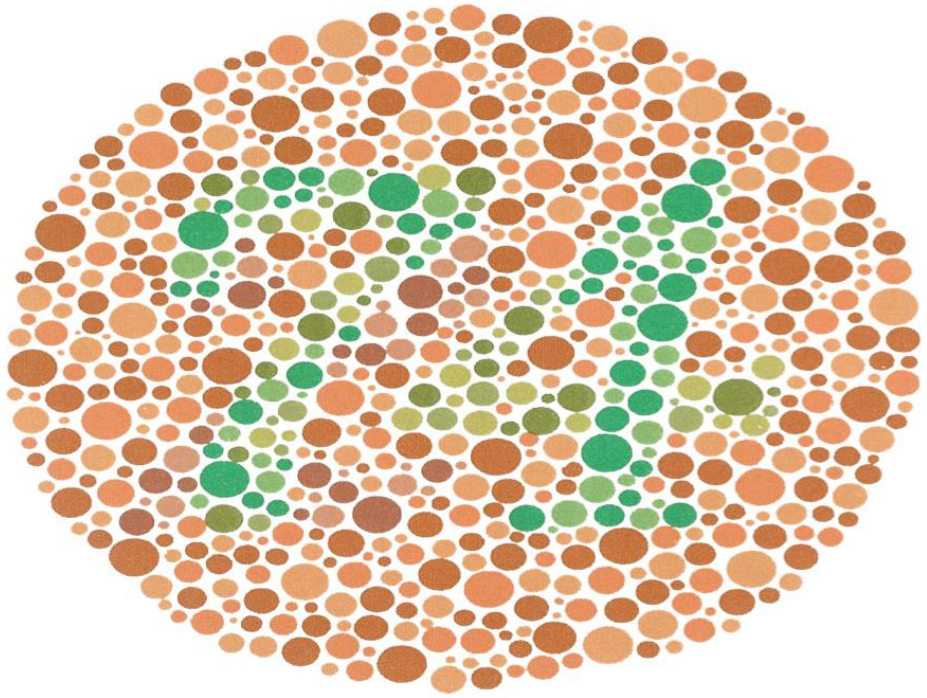
- 建议之一：坐标轴范围要尽量让关键信息充满图面，而不是把重要部分挤成一团。
- 不必一次展示全部数据；如果你真正关心的是局部区域，适当放大是完全合理的。
- 不好的例子（美国白宫COVID1-9）：
  - 就是 y 轴上出现了一个让人摸不着头脑的最大值。



善用颜色通道

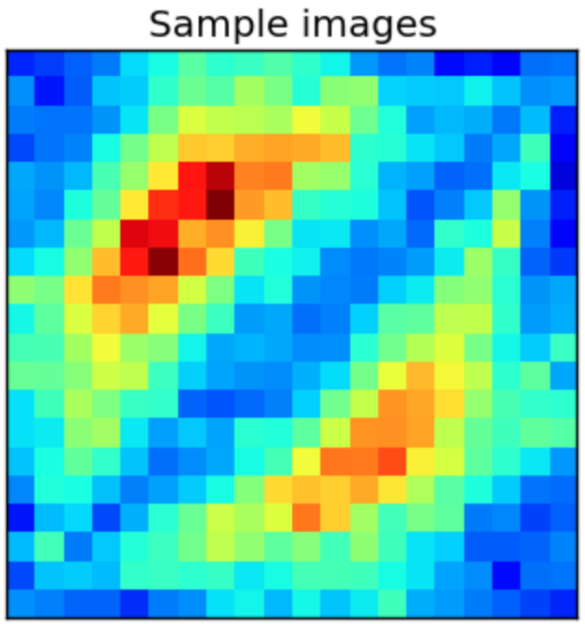
颜色感知

- 挑一组既协调又好读的颜色，其实并不容易。
- 颜色设计不仅关乎美观，也关乎可读性，因为不同人的色觉感知并不完全相同。
- 一个很实用的做法是用 Color Oracle 之类的工具去模拟不同色觉条件下的效果。

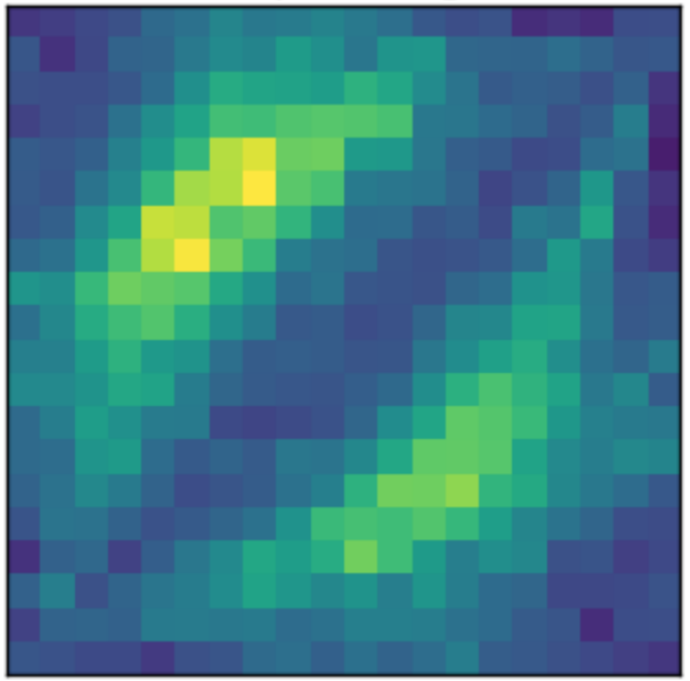


色彩图（Colormaps）

- 同样是把连续数值映射成颜色，不同色图带来的感知效果差异很大。



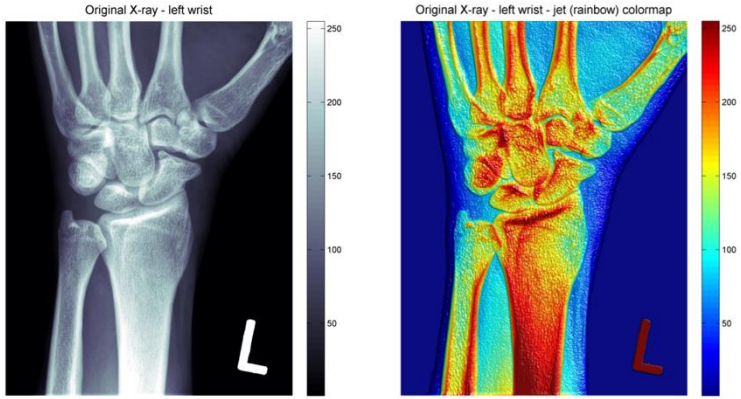
Jet

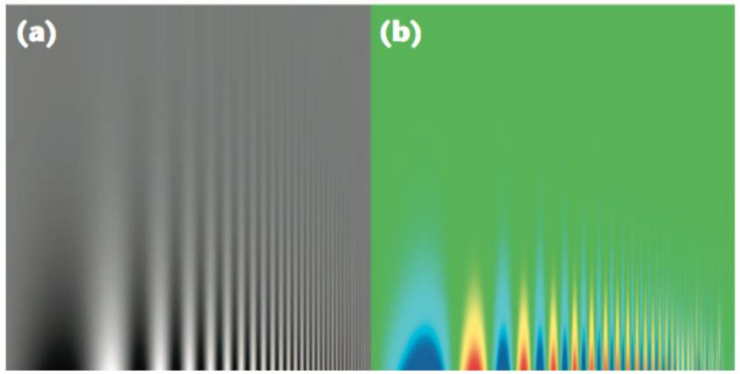


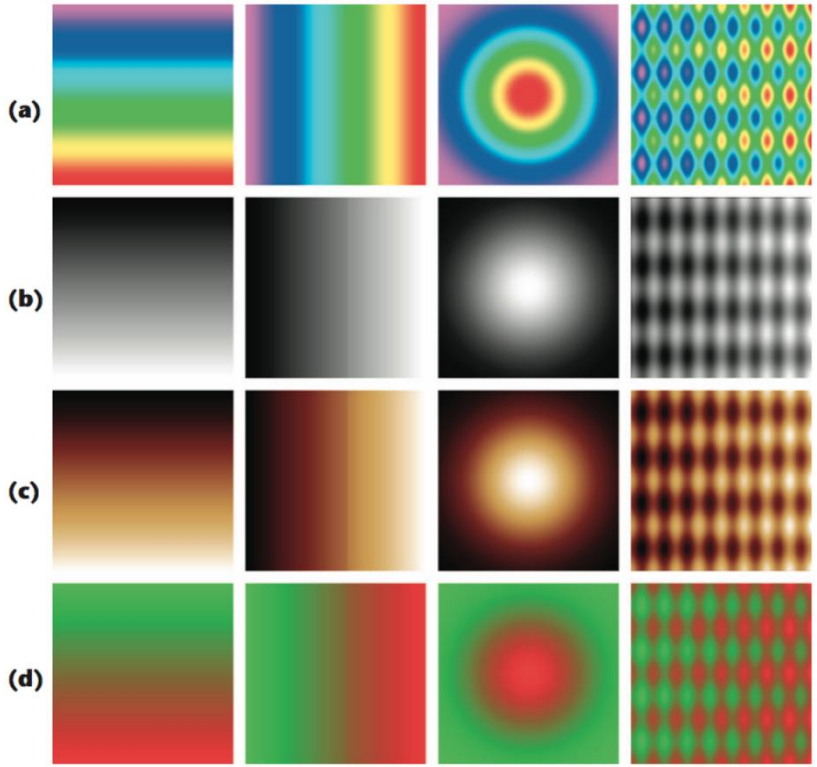
Viridis

Jet / Rainbow 色彩图会误导读者

- **彩虹色**图看起来很鲜艳，但会人为制造并不存在的边界和突变。
- 这也是为什么很多可视化文献都明确批评 rainbow / jet 色图。

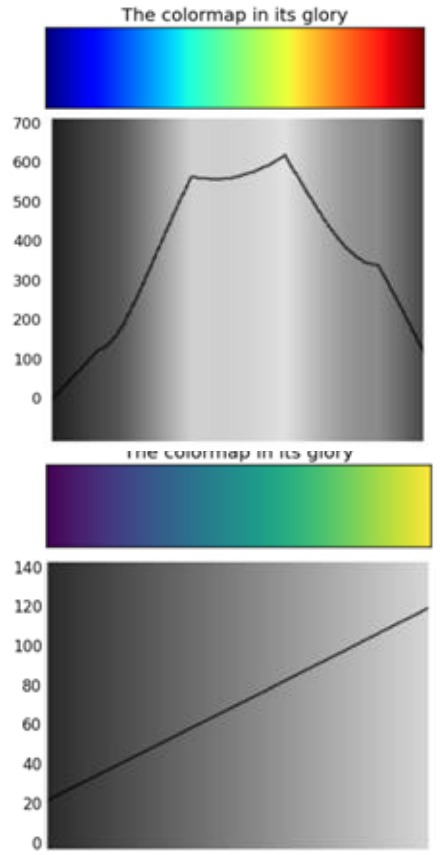






使用感知均匀的色图

- 所谓“感知均匀”，是指数据从 0.1 到 0.2 的变化，和从 0.8 到 0.9 的变化，在视觉上应有近似同样的强度。
- 旧版 matplotlib 默认的 Jet 远远做不到这一点；现在默认的 Viridis 就好得多。
- 另外，也要尽量避免红绿组合，因为红绿色盲非常常见。



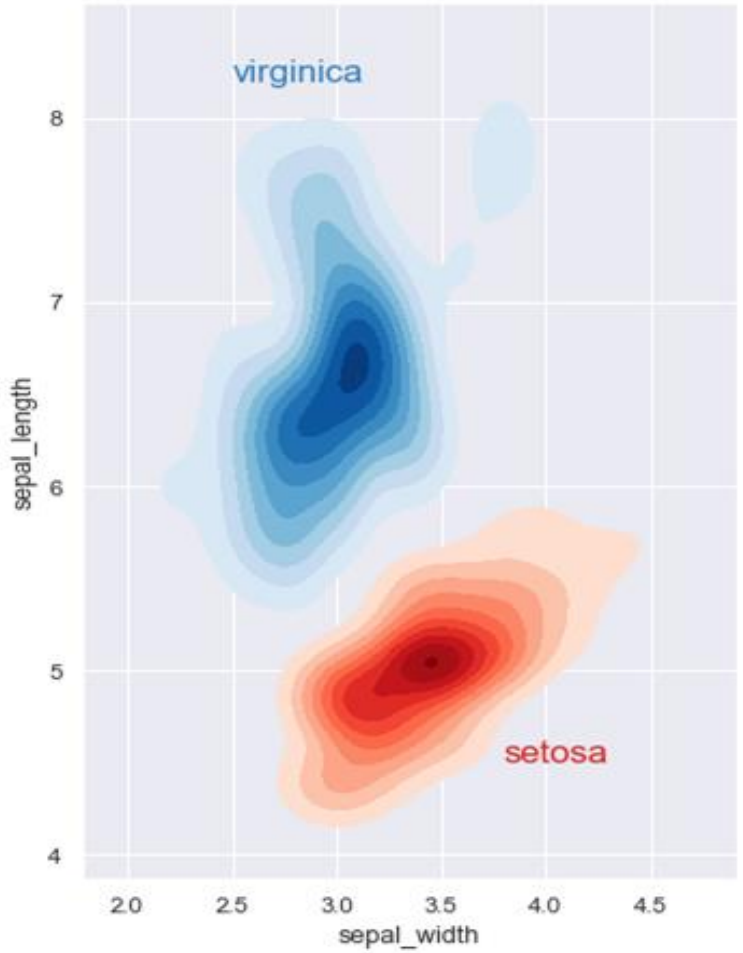
Bounces all over

Slope is constant

x-axis is color, y-axis is “lightness”

用颜色突出数据类型

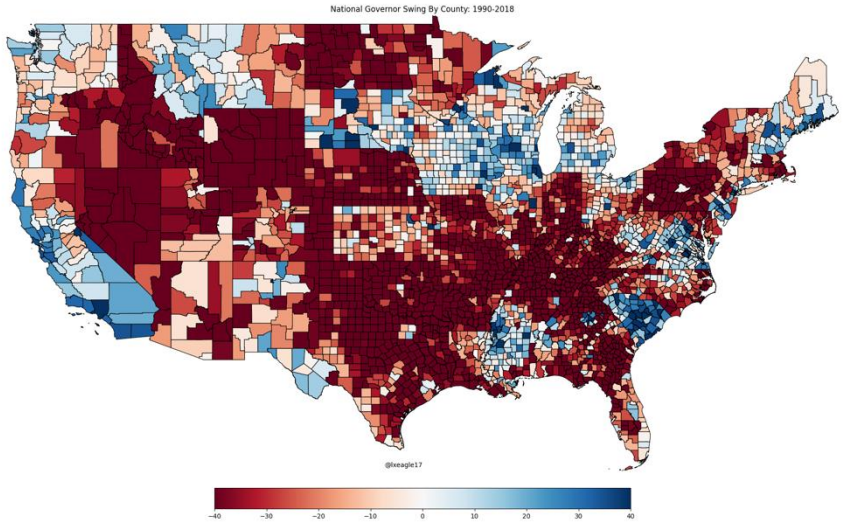
- **定性变量：**应选择容易区分类别的离散配色，因为类别之间没有“**高低次序**”。
- **定量变量：**则更适合用能表达强弱变化的连续配色。
- **有些**图里这两类需求会同时出现，因此更要小心颜色是否传达了错误含义。



定量数据：顺序色图 vs 发散色图

- **如果数据从低到高**单向变化，适合用顺序色图（sequential）。
- **如果中**间值最中性、两端都值得强调，则更适合用发散色图（diverging）。

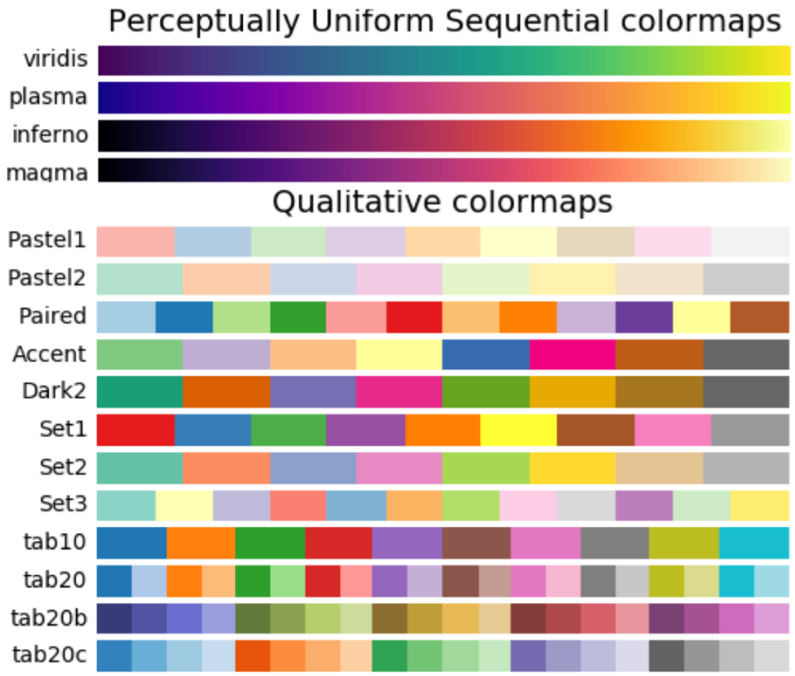
顺序色图

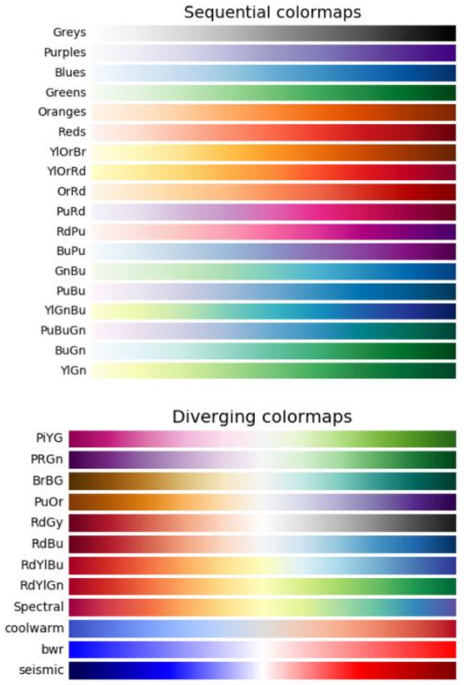


发散色图

Matplotlib 默认色图

- Matplotlib 自带很多色图，但“**能用**”**不等于**“**适合当前任**务”。
- 选色图时，首先要问的是：它是否和你的数据类型、比较目标相匹配？

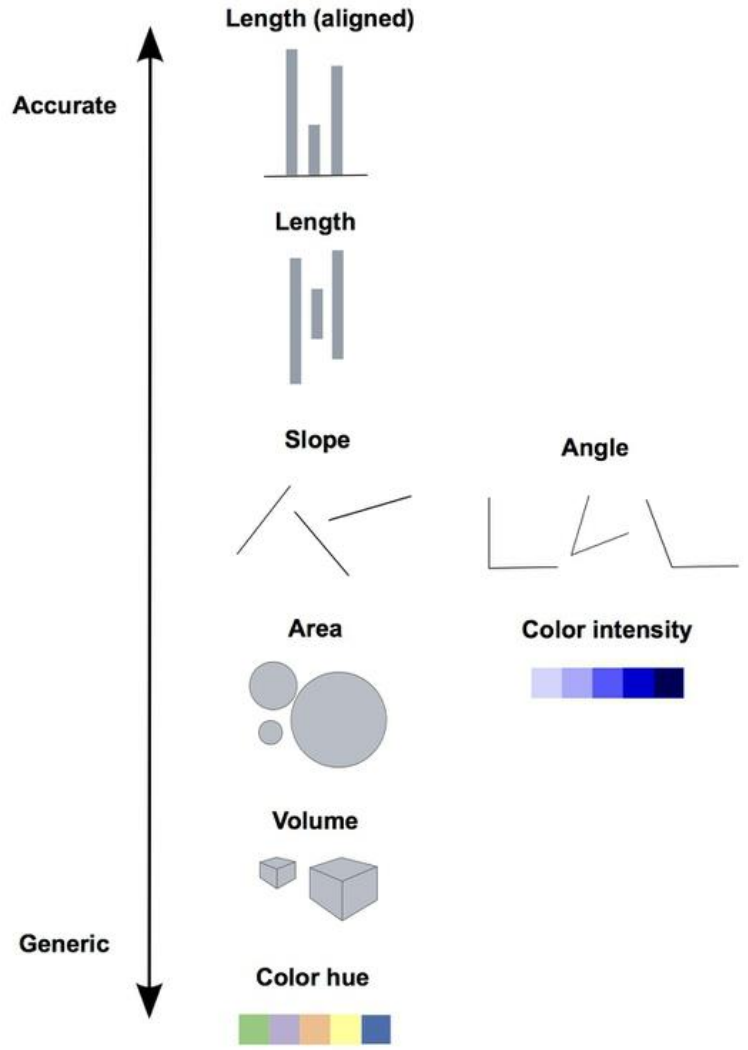




善用图形标记

标记的感知精度

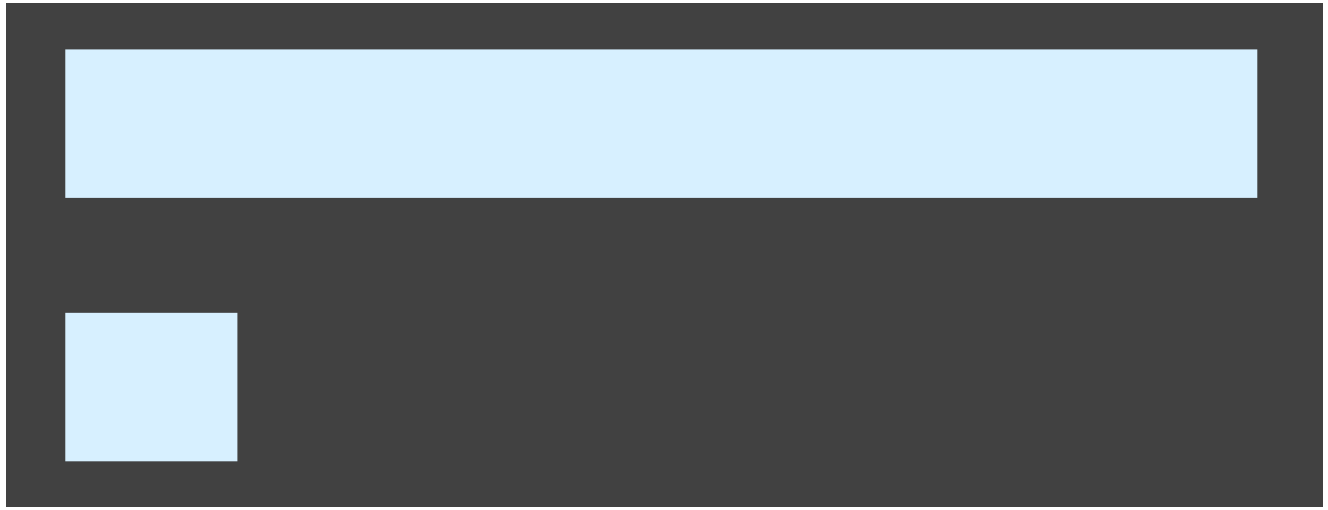
- 我们对数值大小的判断精度，会随着图形标记类型不同而明显变化。
- 有的通道很容易比较，有的则很难。设计图时，应优先选择人眼更擅长的通道。



你觉得长条是短条的几倍?

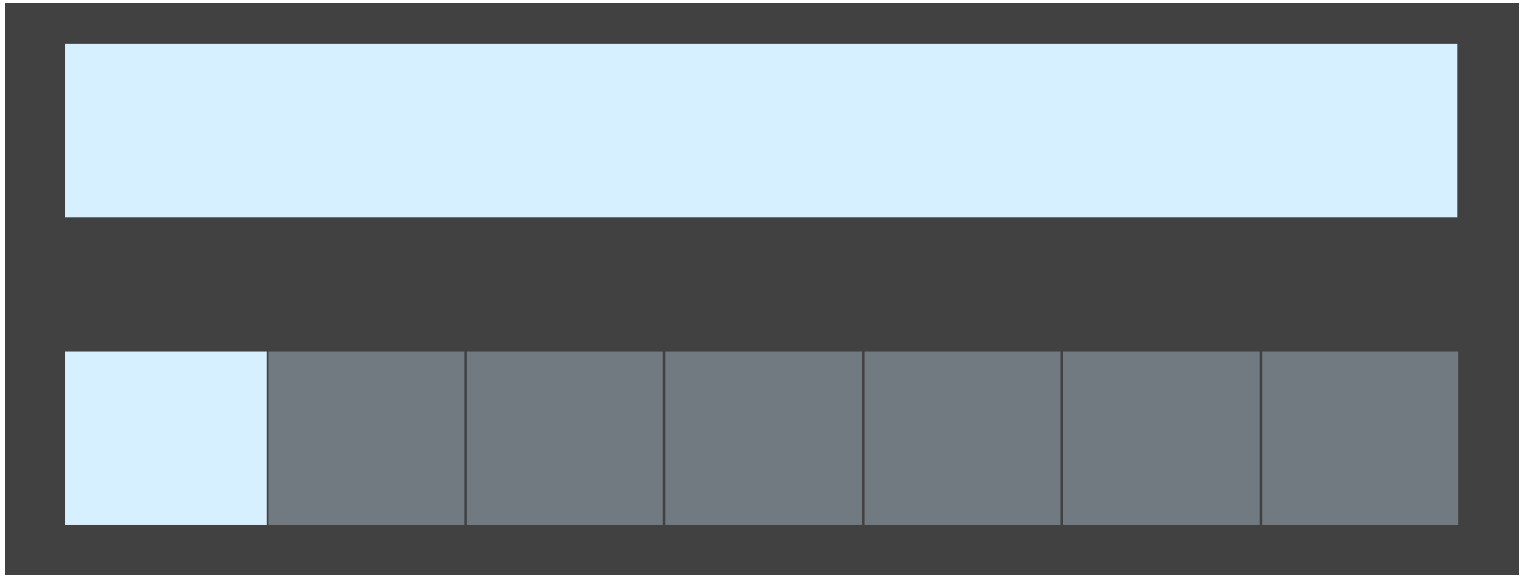
清華大學 Tsinghua University

请直觉判断：长条大约是短条的多少倍？



实际上：长条长度是短条的 7 倍

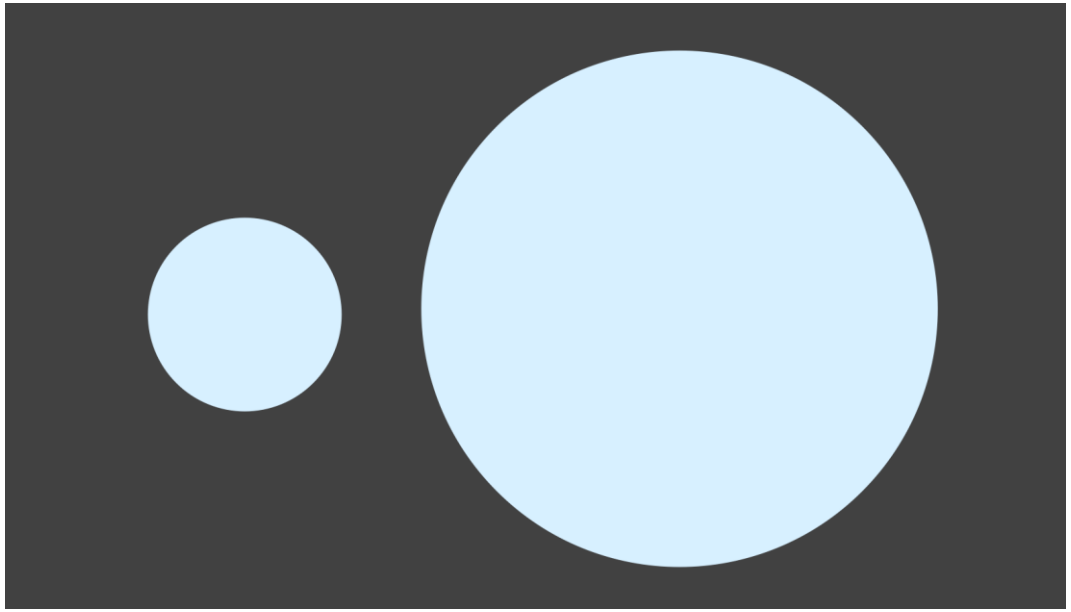
- 大多数人对长度的相对判断会比较接近真实值。
- 这也是为什么条形图在比较数量大小时通常非常有效。



你觉得大圆是小圆的几倍?

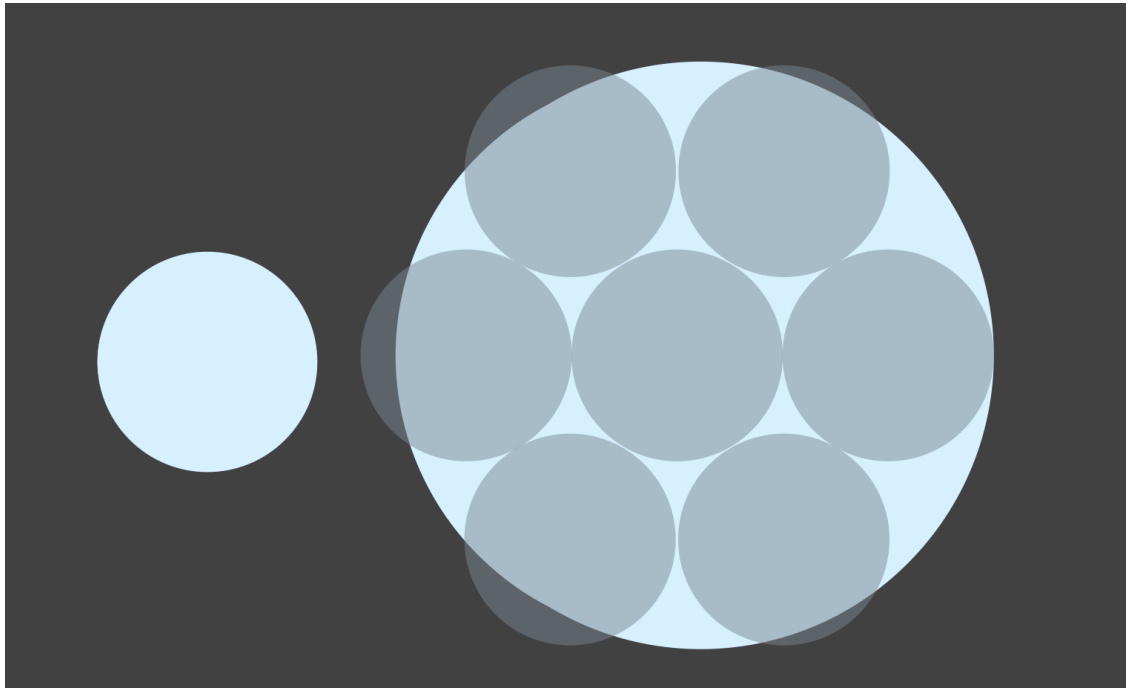
清華大學 Tsinghua University

再直觉判断一下：大圆看起来是小圆的多少倍？



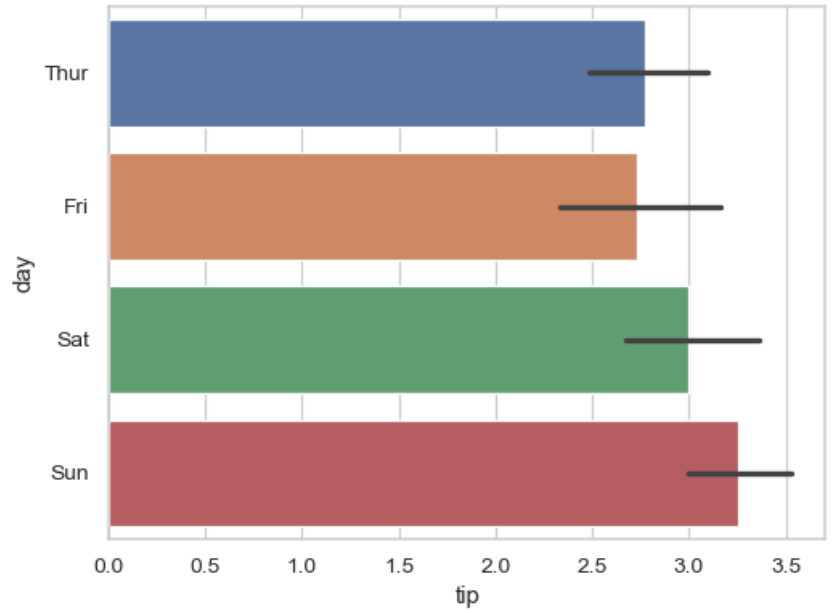
实际上：大圆面积是小圆的 7 倍

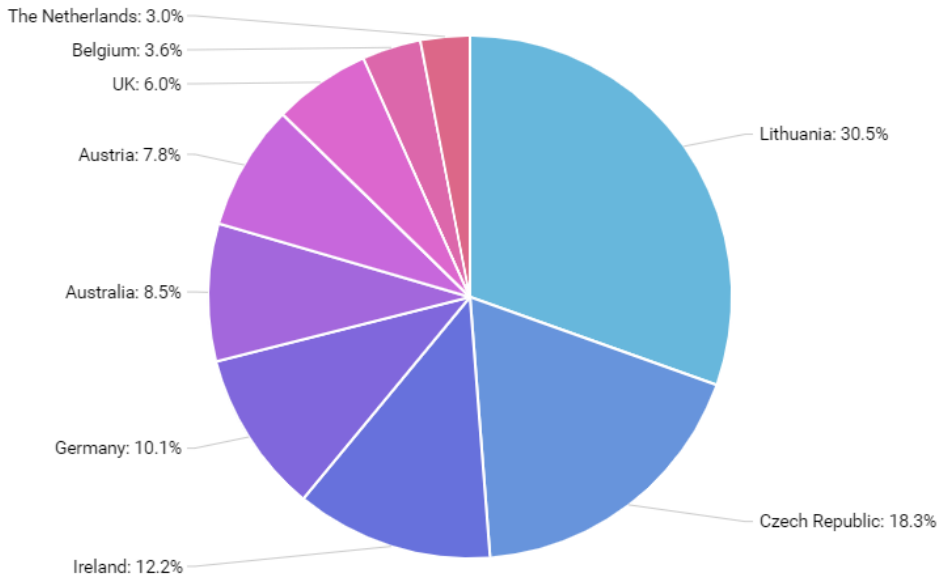
- **即使真**实倍数和上一页相同，人对面积的判断误差通常会更大。
- **所以，如果任**务是精确比较，面积往往不如长度可靠。



长度易分辨，角度难分辨

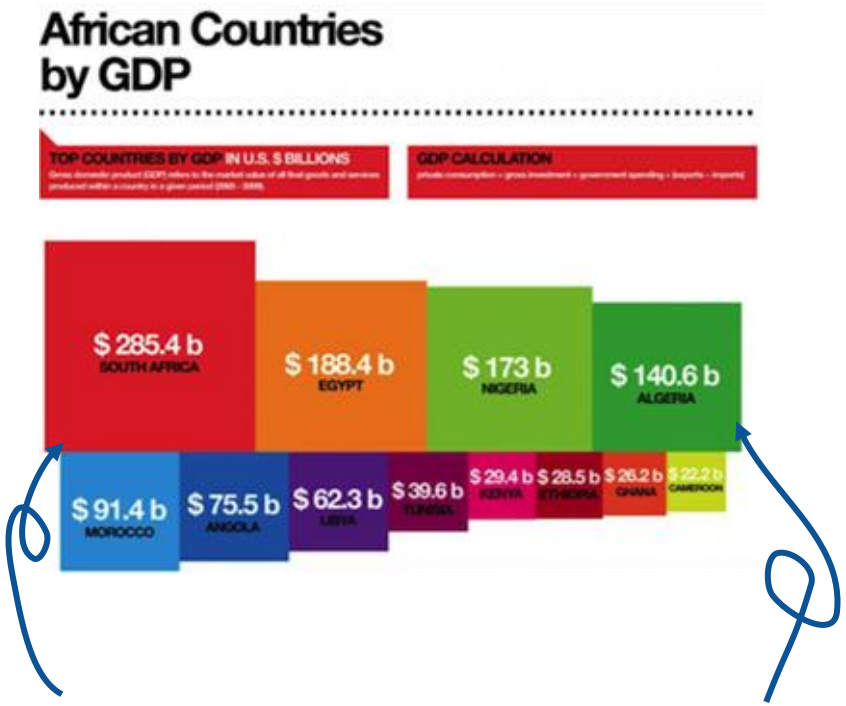
- 长度判断通常比角度判断更准确。
- 这也是为什么在很多场景下，条形图比饼图更适合比较大小。

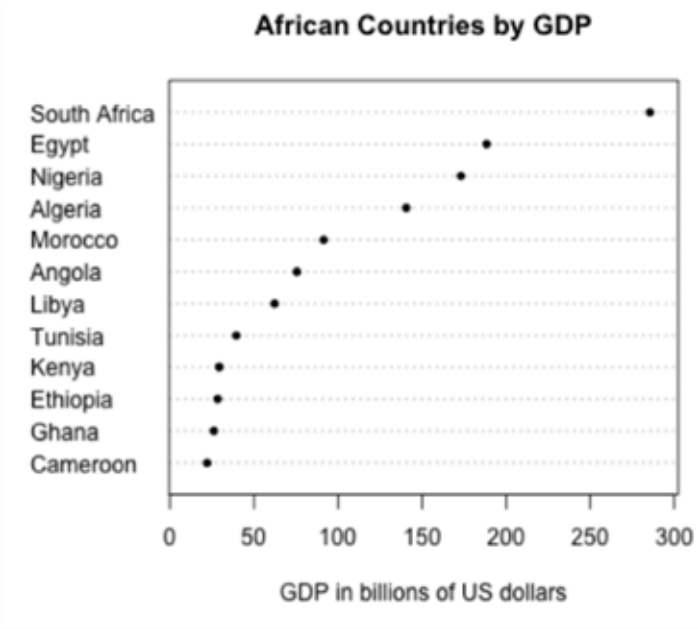




不推荐

面积也不容易准确比较

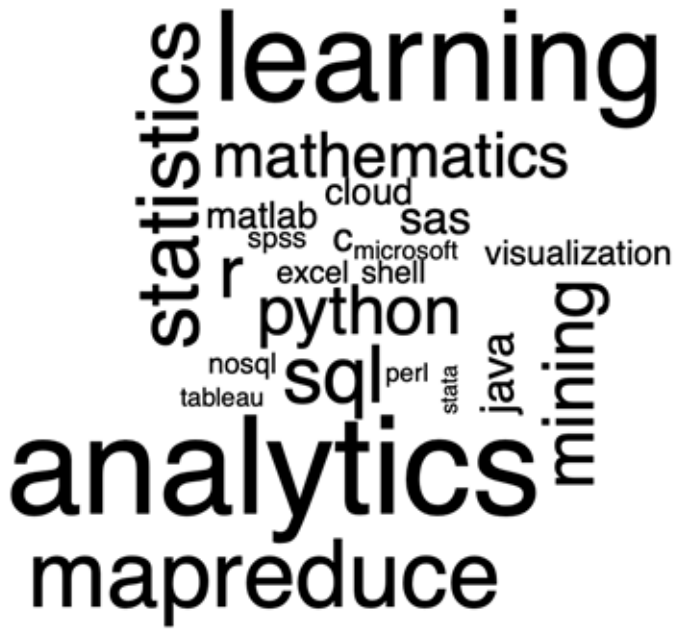


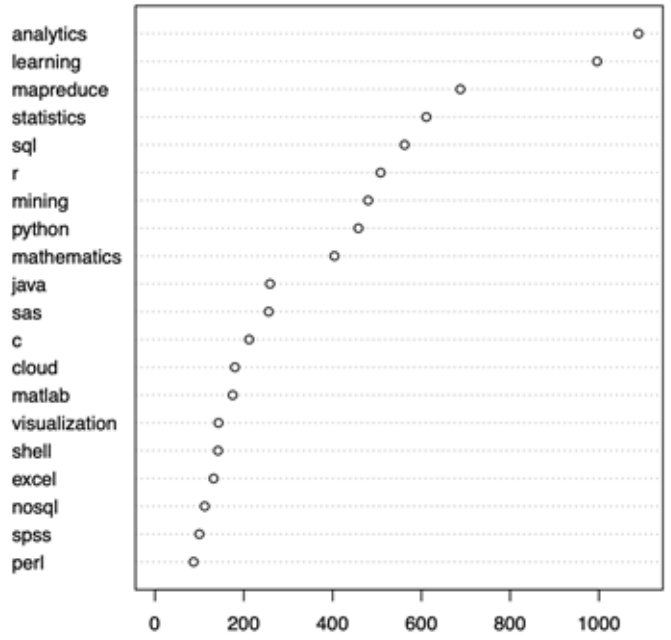


- 当图用面积来表示数值时，读者很难准确判断倍数关系。
- 例如：南非 GDP 是阿尔及利亚的两倍，但从面积上并不容易一眼看出来。

面积难比较，词云也一样

- 词云的问题和面积图类似：我们很难准确比较不同词占据的面积。
- 如果目标只是制造“主题印象”，词云还能用；如果要做定量比较，就不合适。

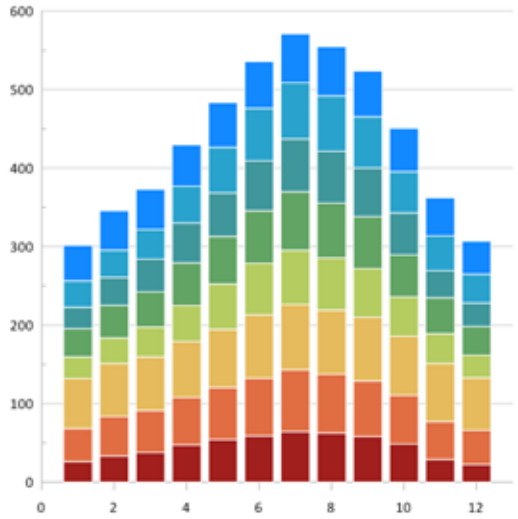




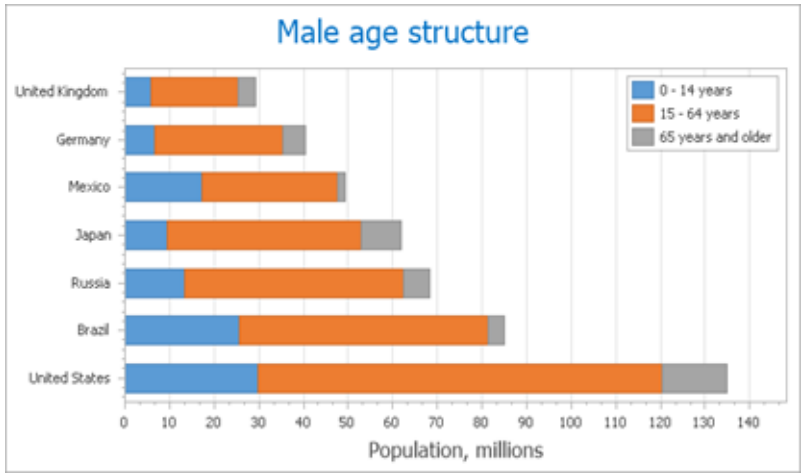
避免“易懂”的基线

- **堆叠条形**图、面积图等图形常常让比较变难，因为基线在不断变化。
- **当基**线“**移**动**起来**”**之后**，**同**样长度的比较就没那么直接了。

在第一张图中：
- 最上面的蓝色柱子长度其实都大致相同。
- 但这一点并不容易立刻看出来！

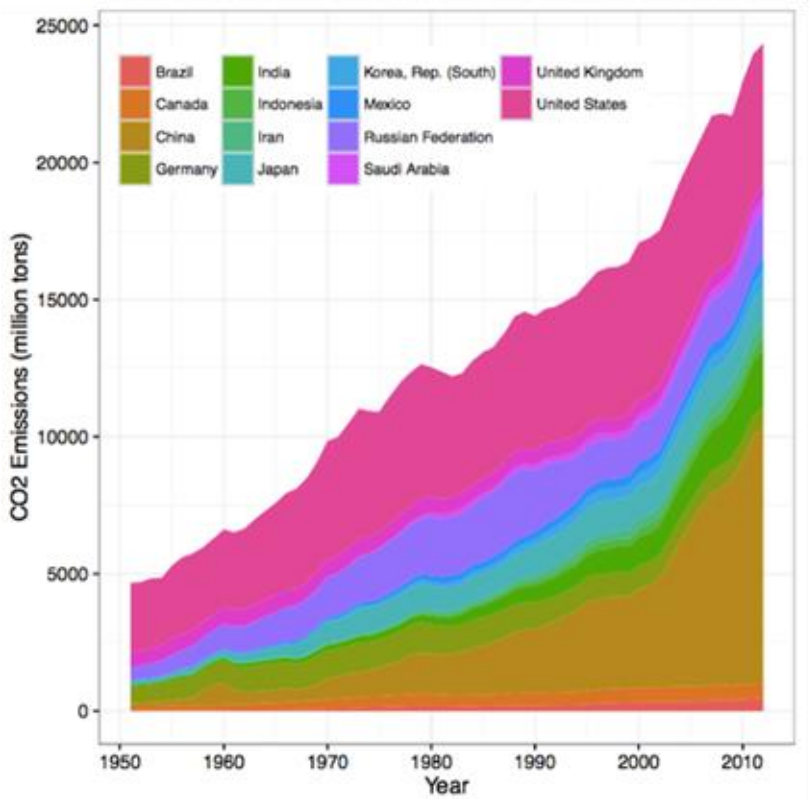


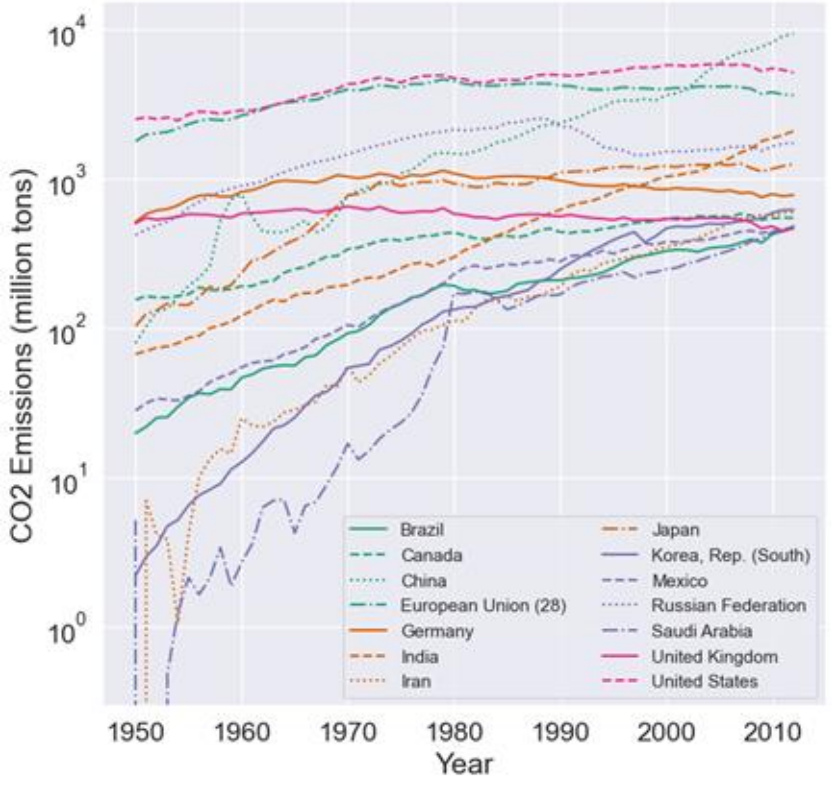
在第二张图中：
- 比较德国和墨西哥 15–64 岁男性 的数量是比较困难的。



改成折线图后更容易比较

- 有时只要换一种标记方式，比较任务就会立刻变简单。
- 这里改成折线图后，趋势和相对差异都更容易被读出来。



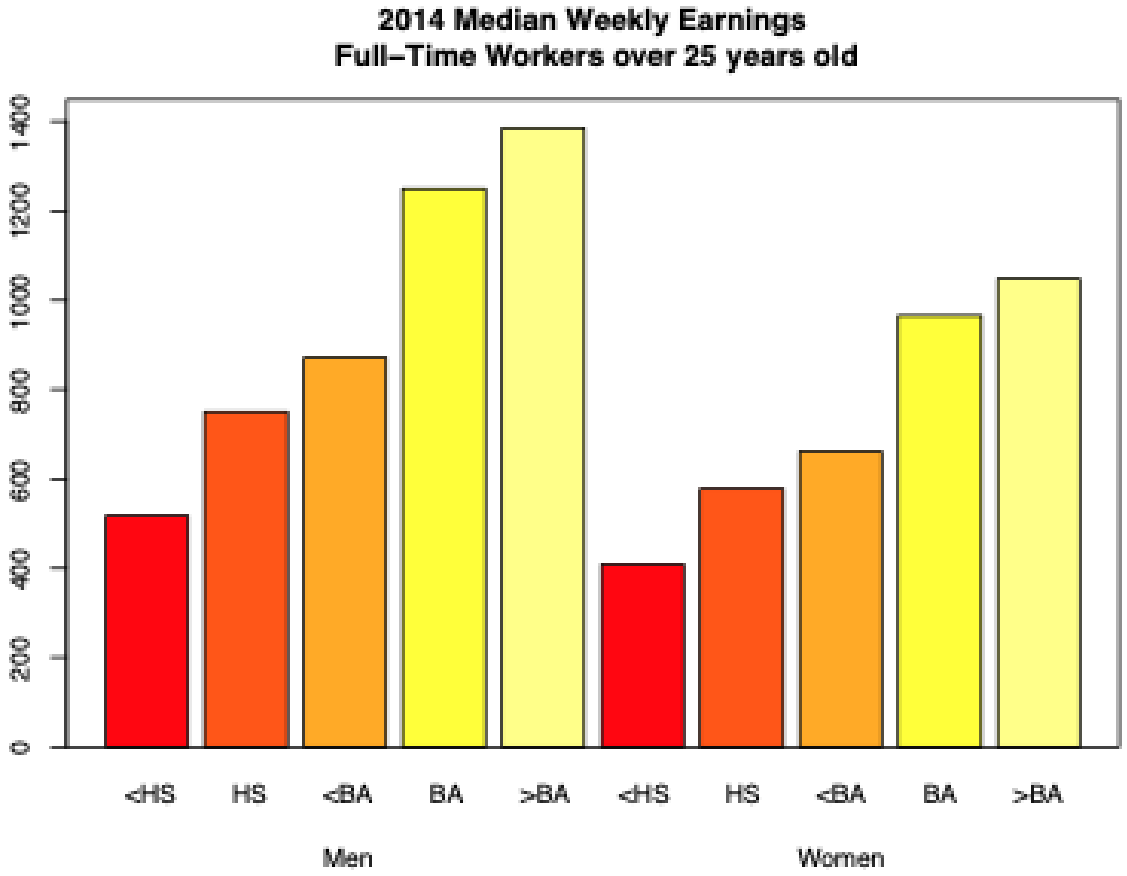


条件化与并列比较

用条件化帮助比较

这组数据来自美国劳工统计局，展示了不同教育水平下男女的周收入中位数。

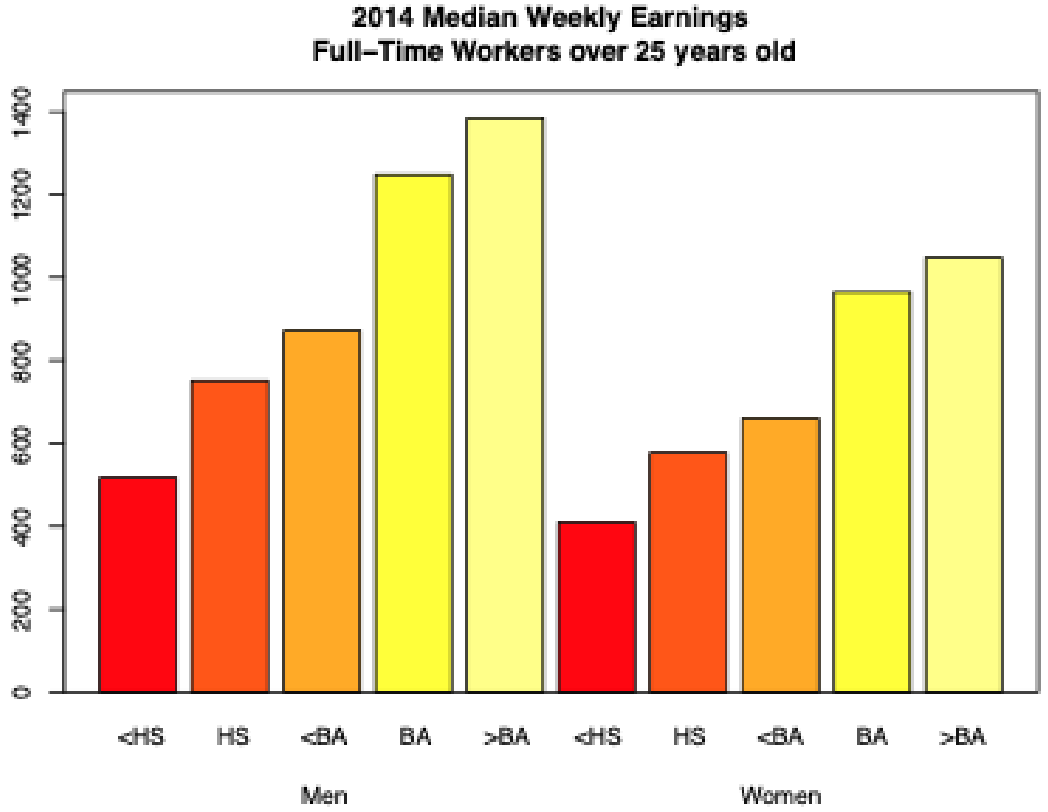
- 用这张图，哪些比较是容易做出来的?
- 哪些比较才是最有意思、最重要的?



用条件化帮助比较

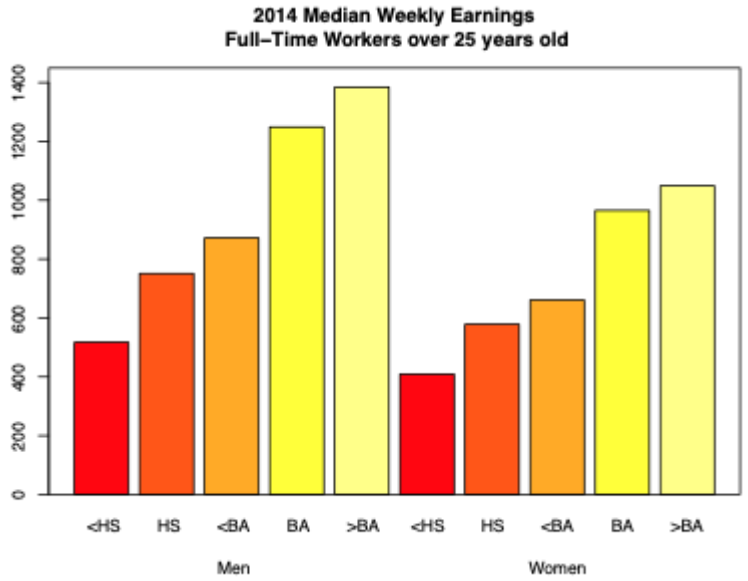
- **当前**这张图很容易看出：教育水平越高，收入通常也越高。
- **但它并不方便我**们比较男女之间的收入差距，这反而可能是更重要的问题。

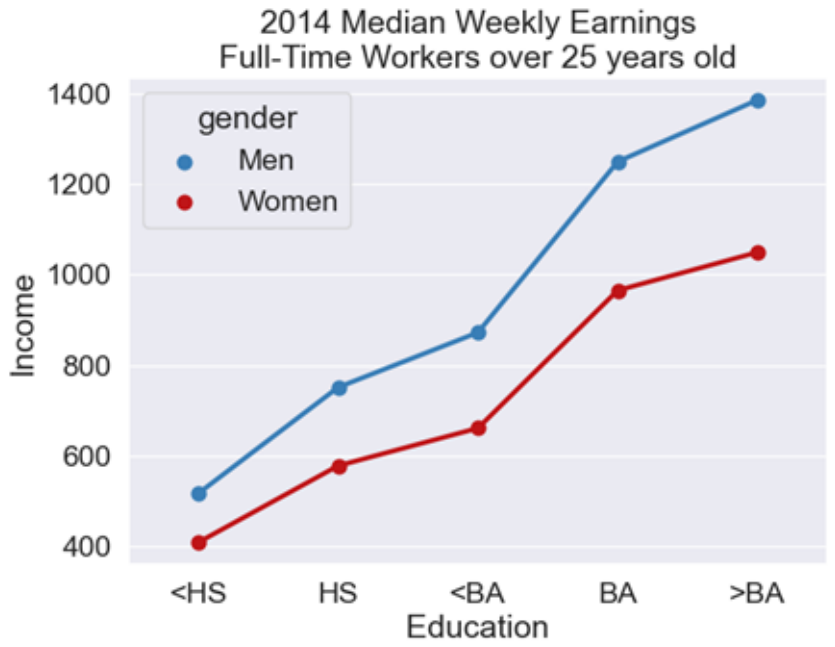
- **怎样改图，才能让这个“难比较”的任务变容易？**



拆成两条线后，性别差距更清楚

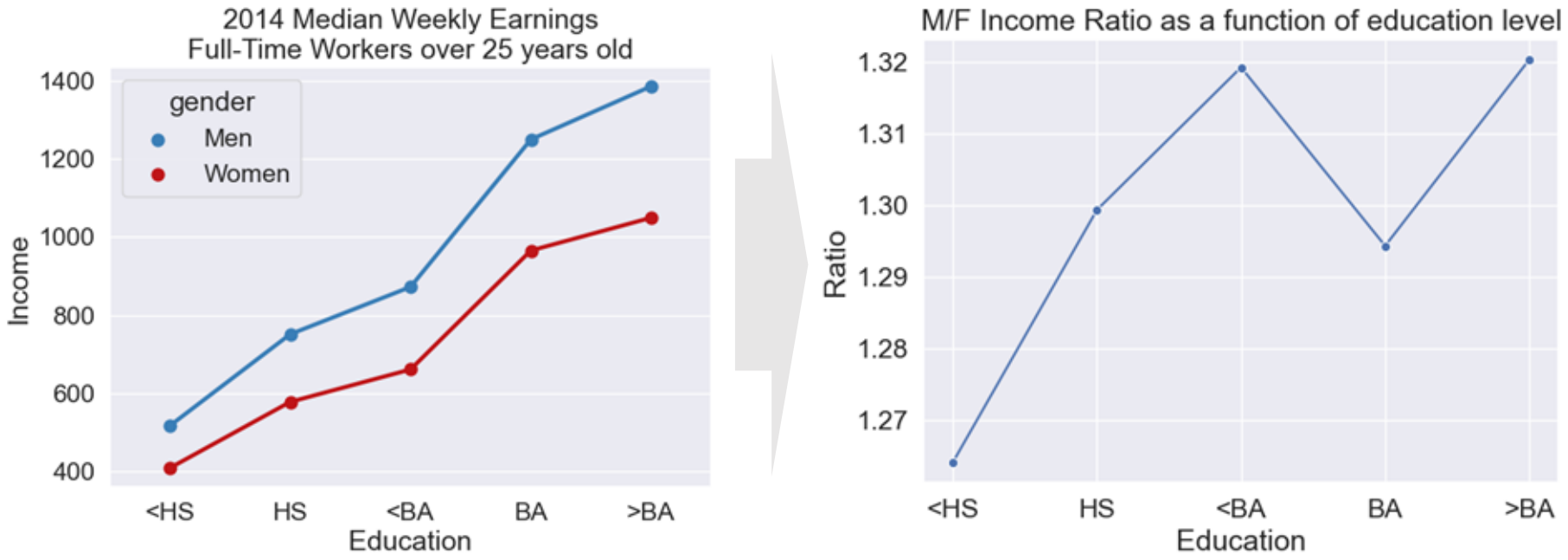
- **把男性和女性拆成两条**线后，性别差距就变成了可以直接观察的对象。
- 这就是条件化的作用：把**“原本不好比较”的关系拆开**，让比较任务变简单。





教育程度越高，收入差距如何变化？

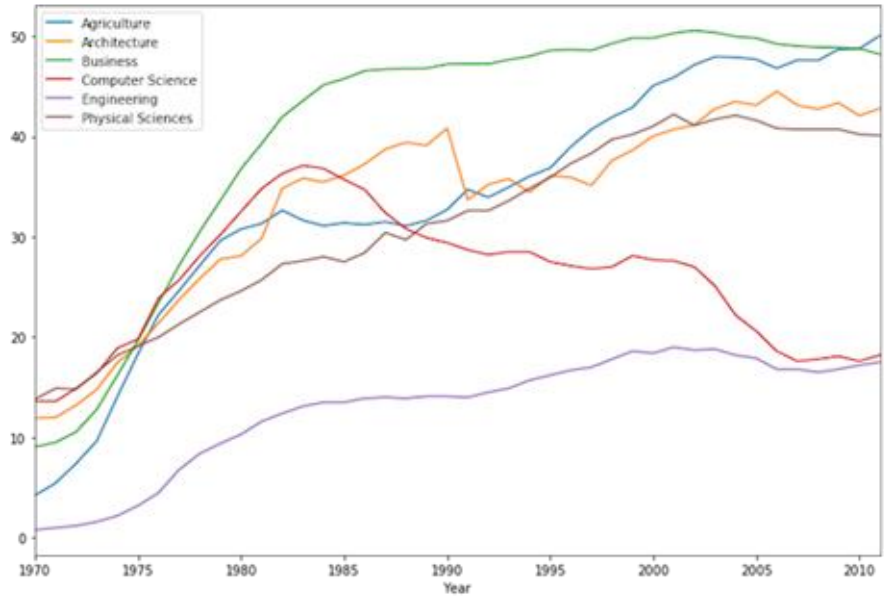
- 当图表已经把性别分开后，我们就能更容易追问：收入差距会不会随教育程度增加而扩大？
- 这类问题正是条件化与 groupby 结合后最容易回答的。

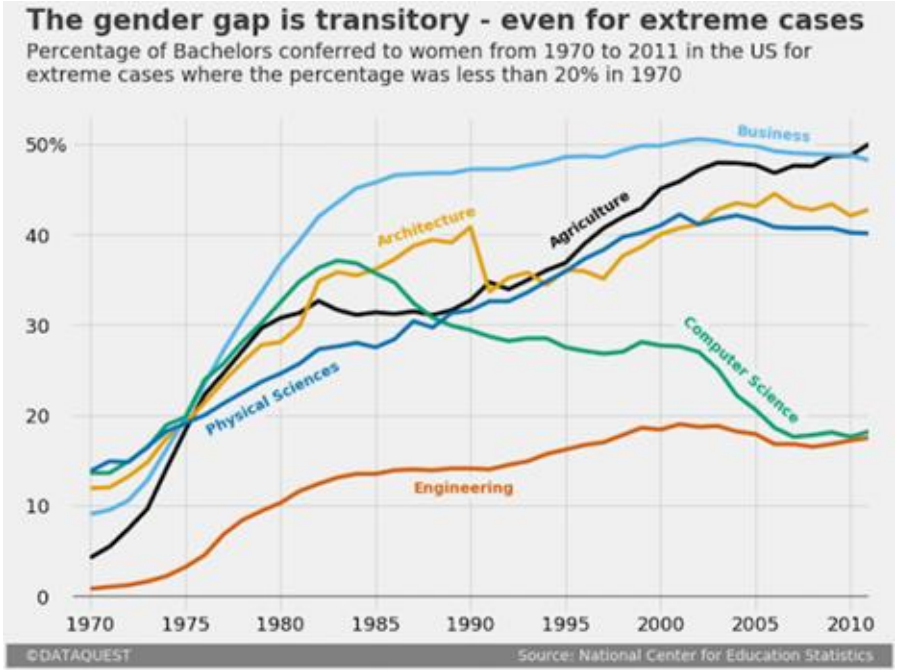


语境与发表表达

为“可发表”做准备

- **教学代**码能画出图，不代表图已经适合直接发表或放进报告。
- **真正面向**读者的图，需要把上下文、结论和解释一起准备好。





发表级图表：把上下文直接放进图里

一张可用于正式发表的图需要具备：

- **信息明确的标题**（要传达结论，而不是只描述图的形式）。
- 例如，用 **“年长乘客在机票上的花费更高”**，而不是 **“价格与年龄的散点图”**。
- **坐标轴标签。**
- 针对重要数值添加**参考线、标记和标签。**
- 适当时加入**图例。**
- 用于说明数据内容的**图注。**

发表级图表：图注

一图胜千言，但你真正想表达的一千个字，不一定都已经写在图里 。很多时候，我们需要借助图注来帮助讲清楚故事：

- 完整且自包含
- 说明图中画了什么
- 引导读者注意重要特征
- 说明从图中得出的结论

总结

清華大學 Tsinghua University

可视化需要深思熟虑！

单变量分布：**地毯**图、直方图、密度图、箱线图、小提琴图

多变量联合分布：**散点**图、六边形图、等高线图

可视化工具： seaborn **和** matplotlib，**以及** Pandas **内置**绘图

数据转换：Log 变换，幂变化

可视化原则：**信息通道**，**X/Y通道**，颜色，图形标记，条件化，语境