数据格式化工具：

1. Google Refine
2. Mr. Data Converter: 可以把Excel中的数据转化为XML、JSON等其他格式。

开箱可用的可视化工具：

1. Excel
2. Google Spreadsheets
3. Many Eyes，单词云（Wordle），缺点：数据是存储在公共空间里的，不能分析敏感数据。
4. Tableau Software 软件较新，图形界面漂亮。桌面版需要收费，免费版本同Many Eyes一样，数据是公开的。
5. Your.flowingdata 主要针对个人数据方面。

编程工具：

1. Python
2. PHP
3. Processing：需要使用到Java小应用程序
4. Flash和ActionScript
5. HTML、JavaScript和CSS：Protovis是一款免费开源的可视化函数库。
6. R

绘图软件：

1. Adobe Illustrator：商业软件，设计业界的标准。
2. Inkscape：开源软件，Illustrator的替代品。

地图绘制工具：

1. Google Map
2. ArcGIS 桌面软件，不需要写代码，能在用户界面中完成所有事情。
3. Modest Maps：Flash和ActionScript的区块拼接地图函数库。
4. Polymaps 类似于Javascript版本的Modest Maps
5. R

## 有关时间趋势的可视化：

### 时间中的离散点：

1. 柱形

R语言首先生成图，然后使用Illustrator进行图像美化。

Tips：①处理的数据都是正数时，要让柱形图的数值轴从0开始，否则会让人们难以从视觉上比较各柱形的高度。

②在选择柱形间隔时需要小心。如果间隔的宽度和柱形宽度相近，视觉上就会产生振动效应，给人感觉柱形和间隔的角色发生了互换。

1. 堆叠柱形图

与常规柱形图相似，区别在于矩形彼此堆叠。如果数据存在子分类，且各子分类之和有意义，我们就可以用到堆叠柱形图。

1. 圆点

即散点图，更加能体现出“流”的感觉。

#### 延续性数据

延续性数据表现的是不断变化的现象。

1. 将之前的散点图用线相连
2. 阶梯图

适用于那些长期停留在某个值上，然后突然增长或者衰退的数据。例如：银行利率、邮费等。

1. 平滑和估算

在散点图上加上拟合曲线，这样就可以对未来趋势进行预测了。

## 有关比例的可视化

时间序列数据是以时间为分组依据的。在比例数据中同样也存在分组依据，只不过它是按类别、子类别和群体进行划分的。

### 整体中的各个部分

1. 饼图

良好组织数据，不要将一个饼图分成太多块。

颜色也是一个重要的因素，我们应该让较大的楔形使用更深的颜色，而较小的楔形使用较浅的颜色。

关于颜色的选择，可以参考这两个网站：http://colorbrewer2.org和http://0to255.com.

1. 面包圈图

面包圈图的中心一般用于放置标签或者其他内容。

1. 可交互的堆叠柱状图

堆叠柱状图不仅仅可以表现时间数据，也可用来表现类别数据。

4. 层级和矩形

板块层级图（treemap）是一种基于面积的可视化方法，通过每一个板块（通常是矩形）的尺寸大小进行度量。它最适合展示的是层级结构，即树状结构的数据。

该图对数据格式的要求：每一行需要一个独立的id、矩形尺寸的衡量依据、以及数据所属的父类别。还有第四个可选的属性，用来给矩形着色。

### 带时间属性的比例

实例：同样的民意调查在一年中的每一个月都会举行一次；某地区的人口年龄比例随时间的变化等。

这里当然也可以用之前介绍过的堆叠柱状图，下面介绍一些新的表现形式。

1. 堆叠面积图

将多个时间序列表从下往上堆叠，填满空白的区域，就得到一个堆叠面积图。

堆叠面积图存在缺点：非底层的变化趋势可能会难以识别，因为每一个数据点的位置都受到了它下方点的影响。所以有时候用直接的时间序列图来表现分布反而会更直观一些。

两种图各有千秋，前者重整体，后者重个体。选择何种图表应该反映出你希望表达的重点是什么。如果页面上足够的空间，你甚至也可以把两种图都放上去。

### 小结

比例分布数据和其他数据类型的主要区别在于它们体现的是整体中的各个部分。每一个单独的数值都有其意义，而各部分相加得到的子集或者总和也是如此。

如果数据不多，那么最好用饼图，当然也可以用面包圈图。如果数据有多个类别的多个数值，可以考虑堆叠柱形图，而非多个饼图。如果你想探索其中随时间而变化的模式，则可以使用堆叠面积图，或者使用传统的时间序列图。

## 有关关系的可视化

不同变量之间的关系包括因果关系和关联关系，前者很难通过数量进行证实，因此也难以用图表描述，事实上，研究人员可能要花费数年时间来弄清楚这些事情。但对于后者，我们能够很轻易地发现事物间的关联。这能引导我们进行更加深入的探究和分析。例子：每升牛奶的价格和每升汽油的价格。

### 关联性

关联关系可以帮助我们发现事物之间可能的因果关系。但大多数情况下这种预测并不一定准确。例如：汽油价格和世界人口都在逐年增长，但这并不意味着降低汽油价格，人口增长就会减缓。

1. 散点图

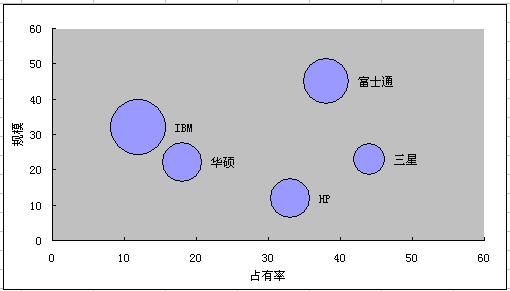
设计建议：散点图最后一般要加一个LOESS拟合曲线，这样才能清晰地显示出二者之间的关联。一般要强调曲线（加粗），弱化散点（颜色调淡），从而将读者的注意力吸引到曲线上面。

2. 散点图矩阵

这种方法在数据探索阶段特别有用。对于多个变量，画出所有的配对散点图。这样有利于我们从中发现其中有关联的变量组。

3. 气泡图

一般气泡图指的是带有“气泡”维度的散点图。这种图表类型的优势在于它便于我们一次比较3个变量，一个变量是x轴，一个变量是y轴，而第三个则通过气泡的面积大小来体现。



## 分布

1. 茎叶图

这是一种老式的分布图表。

2. 直方图

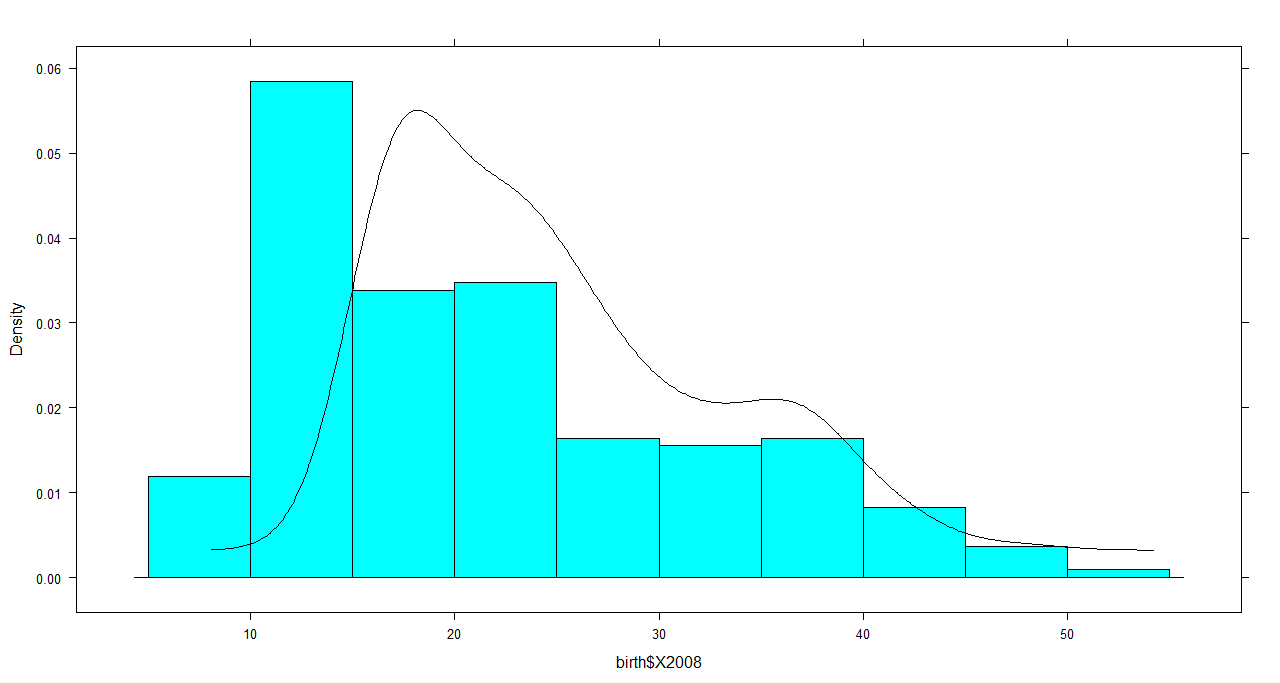
直方图是更加图像化的茎叶图。

建议：在直方图上说明分布的关键点，比如平均值、最大值和最小值等。

3. 密度图

虽然直方图的数值轴是延续性的，但整个分布依然被分成了数个矩形。每一个柱形代表的都是一些条目的集合。我们不能看出在每个柱形内部的变化。这是我们就要采取密度图来对分布的细节变化进行可视化。

另外，密度图可以和直方图进行统一，即画到一张图里面。这样既能照顾到整体，又能照顾到部分，如下图所示。



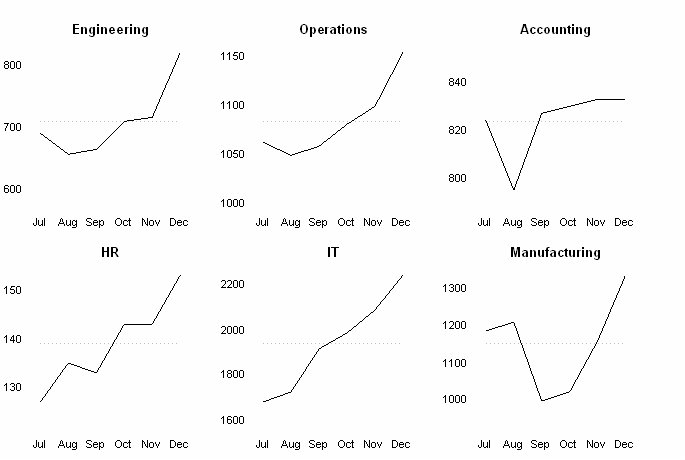
## 对照和比较

通常，将多个分布同时进行比较会带来更大帮助。如比较每一年的整个世界人口的自然增长率分布。

1. 直方图矩阵

2. 系列组图

将大量小图表归于一起的技巧通常被称作“系列组图”。这种图表方便我们多个群组和分类之间及其内部比较。此外，将有规律的多个图表整合到一起，还会显得更有组织性。



系列组图其实也可以在一张图上画出来，但可能存在如下缺点：1. 各个变量的值的范围不同，导致高度相差很大。2. 最终生成的这张图较复杂。这时不妨尝试一下系列组图，说不定有意想不到的效果。

## 发现差异