学一点数据分析-常见工具

第四节

三、数据分析工具简介

现有的数据分析工具种类繁多，常见的工具有如下几种：

1□Excel

Excel不仅可以用于对数据进行过滤清理，也可以用于对数据进行分析，使用Excel能够满足很大一部分数据新闻的分析需求。Excel拥有强大的附加扩展工具库，因而在简便性与功能性两个方面实现了平衡。

2□SPSS

3□R语言

R语言是一款开源的编程语言，基于统计学界中广泛使用的S语言开发而成，主要用于统计分析、绘图、数据挖掘。它的语言形式和语法都非常简单，输入的命令能够立即被执行。

4□python

5. 其他工具

SAS（Statistical Analysis System）

Stata：特点是采用命令操作，程序容量较小，计算速度快，统计分析方法齐全。Minitab较为简单易懂和畅销

PASW Modeler：包含基本的数据挖掘算法功能，可以做高级数据分析和数据挖掘应用工具。

数据分析工具的种类很多，也无所谓好坏之分，建议读者按照自己的专业水平找到其中一至两种适合自己的工具熟练掌握即可。

在《国家报》，
Excel去清洗、组织和分析数据；
谷歌电子表格去发布、连接像谷歌Fusion Tables、Junar开放数据平台这样的服务；
*Junar用于分享我们的数据，并嵌入我们的文章和博客里；
*Tableau用于发布我们的交互式数据的可视化；
*Qlikview，一个非常快速的商业智能工具，我们用它来分析、筛选大型数据集；
*NitroPDF用来把PDF文件转换成文档和Excel文件；
*谷歌Fusion Tables用于地图可视化。

国家报（阿根廷）— 安赫利卡·佩拉塔·拉莫斯(Angélica Peralta Ramos)

我用的工具是Excel，它可以处理大部分CAR（计算机辅助报道）问题，并具有简单易学、大多数记者可快速掌握的优点。当需要合并表时，我通常使用Access，但会把合并后的表导出到Excel，做进一步的工作。我使用ESRI的ArcMap做地理分析，它很强大并且被收集地理编码数据的机构所使用。 TextWrangler在快速分析文本数据的布局及分隔方面很强大，并能用规则的表达式进行复杂的搜索和替换。当需要如线性回归这样的统计技术时，我用SPSS，它有一个友好的操作菜单。对于确实繁重的工作，比如处理数百万计的记录、需要认真筛选和程序化变量转换的数据集，我用SAS软件。

沃尔特·克朗凯特新闻学院 — 史蒂夫·多伊格(Steve Doig)

我们选择的工具包括Python和Django，用于破解、抓取和操控数据；PostGIS，QGIS和MapBox工具箱，用于建设复杂的网络地图。 我们正在考虑选择R语言还是NumPy+ MatPlotLib做探索性数据分析的工具，虽然目前我们最喜欢的数据工具是自主研发的CSVKit。我们所做的一切或多或少都是在云端部署的。
芝加哥论坛报 — 布莱恩·博耶(Brian Boyer)

四、警惕“数据陷阱”

1□忽略基数的百分比

忽略基数的百分比是没有意义的。正如《统计数据的真相》一书的作者瓦尔特·克莱默所说：“一个百分数不但提供了一些信息，而且同时也掩盖了一些信息。如果我们在左边给出两个数字，也就是分子与分母，那么在等式的右边就可以得到一个唯一的结果。这就是所谓的信息。例如比例数字1/5、7/35、117/585，所有这些相对数表示的都是一样的百分率20%，尽管分子（分母也一样）的差异性非常大。如果某人想要隐瞒某个目的，那么他会更愿意以百分数的形式来表示。”

2013年某报报道《洛阳离婚率三年升三倍》，其中提及“洛阳市民政局社会事务科一负责人说，离婚率一般是按照（离婚人数:全市人口数）比例测算出来的，2010年测算的是1.1%，2011年测算的就是2.5%，去年就是3.3%”。记者据此得出了“三年升三倍”的结论。这看起来似乎无可厚非，但实际上这是基于增长率的增长率，因为三年的人口基数各不相同，而且人们容易因为这样的数据而忽视其背后的基础数据，即2013年洛阳的离婚人数到底有多少？

一位市民从工作人员手中接过离婚证

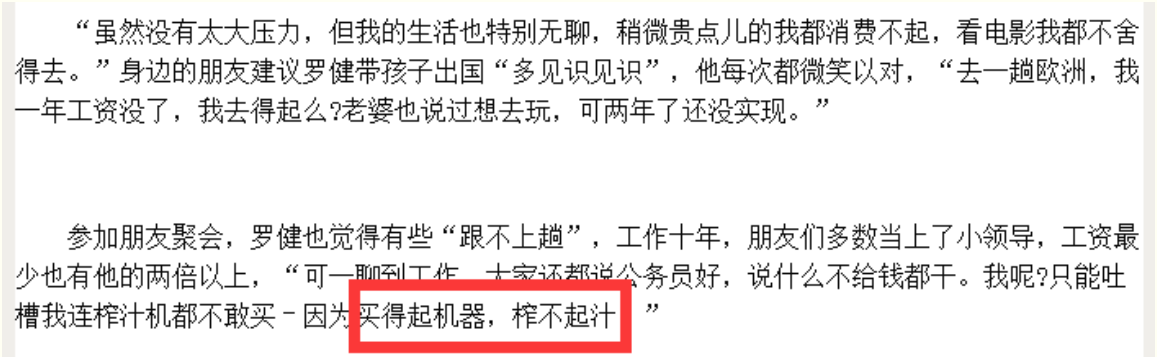
四、警惕“数据陷阱”

2□缺乏代表性的均值

前面曾提及均值容易受到极端值的影响，例如，为了调查“双十一”购物节对大学生的影响，我的学生收集了100份问卷，对所有男生样本当日的购买总额计算了均值，结果得到的数据是1000多元，看起来超过了一般大学男生的购买能力，甚至比女生的样本数据还要高，这让很多人感到费解。实际这个问题恰恰出在不能简单地做均值上，因为根据学生的描述，他们在样本中发现有两位男生购买了3 000多元的电子产品，这两个人的极端值影响了整个均值的计算结果。这样计算出来的均值就不能反映平均水平了。

但是现实中还是有很多统计数据是以均值的方式呈现的，比如有媒体报道2013年北京市职工年平均工资为69 521元，月平均工资为5 793元。这样的均值很容易掩盖不同行业、不同地区、不同机构职工工资之间存在的庞大差距，让低收入人群感觉自己“被增收”，而高收入者则感觉“被低估”。

所以在使用均值时，通常需要结合标识离散程度的数据，以展现相对全面的分布描述。



“虽然没有太大压力，但我的生活也特别无聊，稍微贵点儿的我都消费不起，看电影我都不舍得去。”身边的朋友建议罗健带孩子出国“多见识见识”，他每次都微笑以对，“去一趟欧洲，我一年工资没了，我去得起么?老婆也说过想去玩，可两年了还没实现。”

参加朋友聚会，罗健也觉得有些“跟不上趟”，工作十年，朋友们多数当上了小领导，工资最少也有他的两倍以上，“可一聊到工作，大家还都说公务员好，说什么不给钱都干。我呢?只能吐槽我连榨汁机都不敢买－因为买得起机器，榨不起汁。”

四、警惕“数据陷阱”

3□仅供参考的趋势

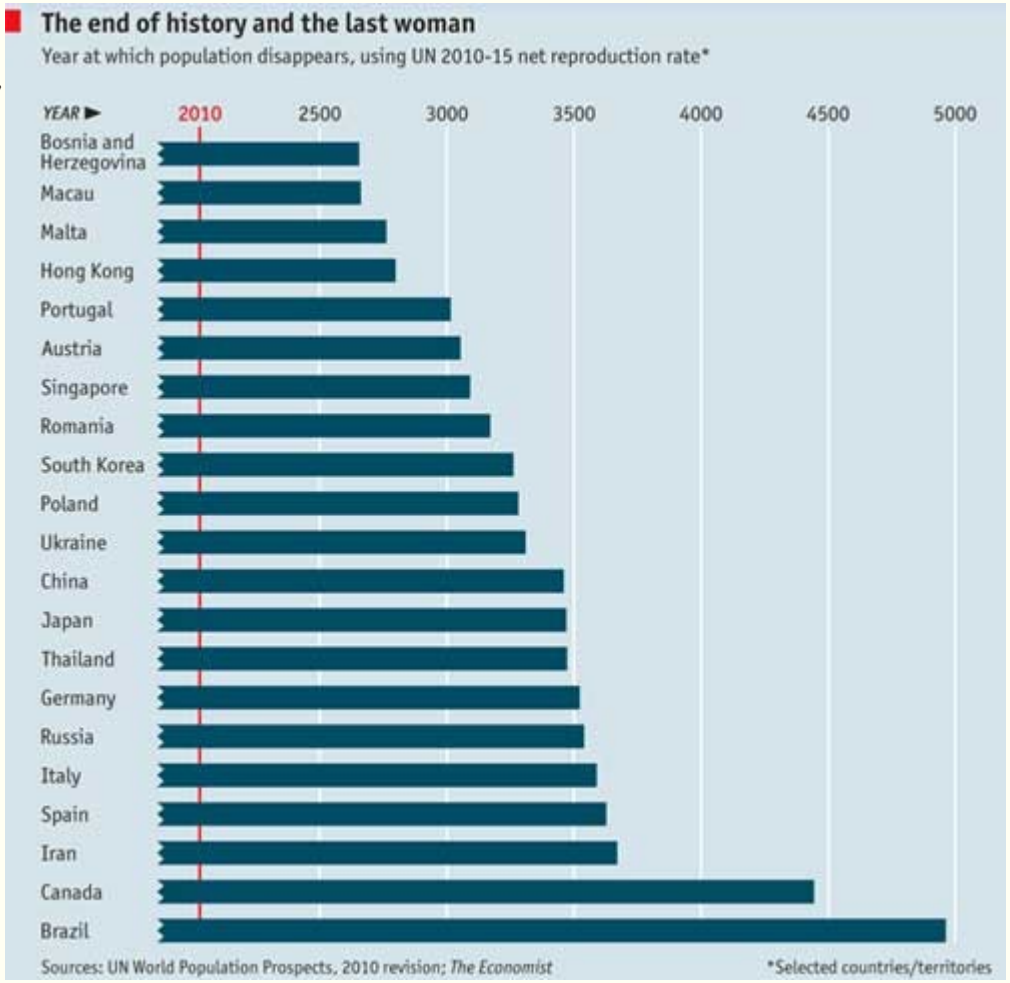
“趋势”是媒体非常关注的数据内容之一，但是“趋势”往往是统计学家按照之前一定时段内的发展规律而做出的分析与预测。如果当下影响“趋势”的因素在未来突然消失了，或是有其他因素对“趋势”做出影响，那么“趋势”则并不能如统计学家预期的产生那么大的作用。

案例：《经济学人》图表报道《历史的终结与最后一位女性》

2011年8月，《经济学人》网站曾以《历史的终结与最后一位女性》为名，做了一篇图表报道。报道聚焦女性选择单身的普遍潮流，从理论上推测女性生育率降低，将导致全球很多国家和地区的女孩越来越少，最后导致女性只剩一个。图表展现了距离最后一个女性，各个国家（或地区）还有多长时间。

这篇报道的价值不在于预测趋势的准确性，因为大家都知道没有一个国家或地区会让问题真正变得那么糟，在可以预见的未来中，会有无数的促进生育率提高的政策发布影响这种趋势，如中国自2014年开始全面推出了单独二胎政策就会使其线条大大拉长。

当然也没有必要认为报道毫无价值，因为这篇报道以一种独特的角度让人们正视女性选择单身潮流和低生育率可能带来的社会影响，具有发人深省的意义。



四、警惕“数据陷阱”

4□缺乏可比性的对比

在对数据进行比较时，首先应该考虑的是，它们之间真的具有可比性。如由于通货膨胀率的影响，两个时间段的收入数字不能简单进行对比。由于人口数量的不同，两个地区的犯罪率也不能直接加以对比。记者应当将无法直接对比的数据转化为有可比性的数据，比如购买力指数、每万人中案发率等。其次，还需要考虑到不能通过简单的对比就仓促下结论。2013年10月，央视推出了报道批评星巴克在华谋取暴利，央视记者在报道中称，对比北京、伦敦、纽约、孟买的星巴克一款中杯拿铁咖啡的价格，北京的27元最贵，而一杯中杯拿铁咖啡的物料成本不足4元。报道推出后遭到很大的质疑，专家和网民都指出这种对比不科学，因为竞争行业的定价问题不能简单以成本来计算，毛利高不一定是售价高决定的，可能还受到人工成本等因素的影响。

5□将相关关系等同因果关系

前面曾解释过相关关系，但是要注意相关关系可能只是一种偶然现象，并不能将之直接等同于因果关系来报道。比如曾有外媒研究中国城乡的房价和人口中的男性比例之间的关系，发现两者存在一定程度的正相关关系，即男性比例越高，城乡的房价也随之增高。报道可以展现这种相关关系，但是这样的问题显然不能简单地以该地男性居多作为原因来解释，还需要尽量分析更多的社会因素。

本章小结

数据会“说话”吗？当然，它看似沉默无声，但只要你能与之沟通，就可以发现新闻事件和话题背后那些有趣的数据，它们构建起一个信息庞大而又丰富多彩的世界。我们所要做的是学会从哪些维度切入去观察它们、思考它们、理解它们、分析它们，甚而呈现它们，让更多读者也和你一同感知我们所处的奇妙世界。

很多数据新闻实际在对数据找到新闻分析的角度之后，就直接以可视化方式呈现了，这无可厚非，部分经典的数据新闻产品就是这样做的。因为这些作品收集了非常完整的数据，且对数据进行了不同角度的分类，最后选择了最适合的可视化方式呈现。

还有一些数据新闻并不满足于收集、分类和呈现数据，而是需要进一步进行统计分析和数据挖掘，操作这种数据新闻就需要对数据进行预处理，保障数据的质量，其中包括统一数据格式，将之导入数据库，做数据清理。

以Excel为例，介绍了进行数据导入和数据清理的步骤和要求，Excel只是文件管理系统，相比数据库而言功能要弱很多，但其界面友好便于入门者学习操作。对于有一定计算机背景的读者而言，可以尝试功能更强大的关系数据库来做数据分析的预处理工作。

本章还介绍了一些基本的数据分析方法和工具，并提醒读者数据也会骗人，在做报道时，更需谨慎对待数据，核实数据。

Data Journalism and Tableau:

Using data for more power in storytelling

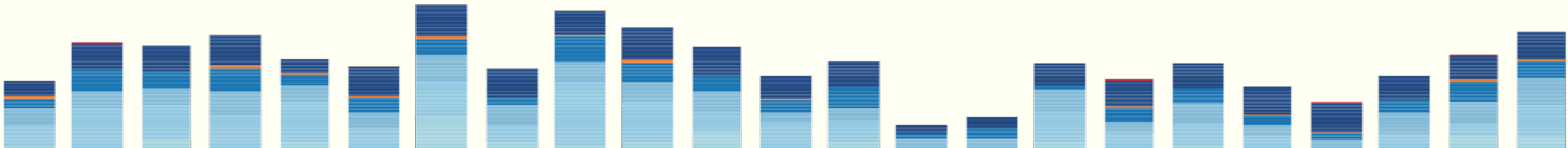
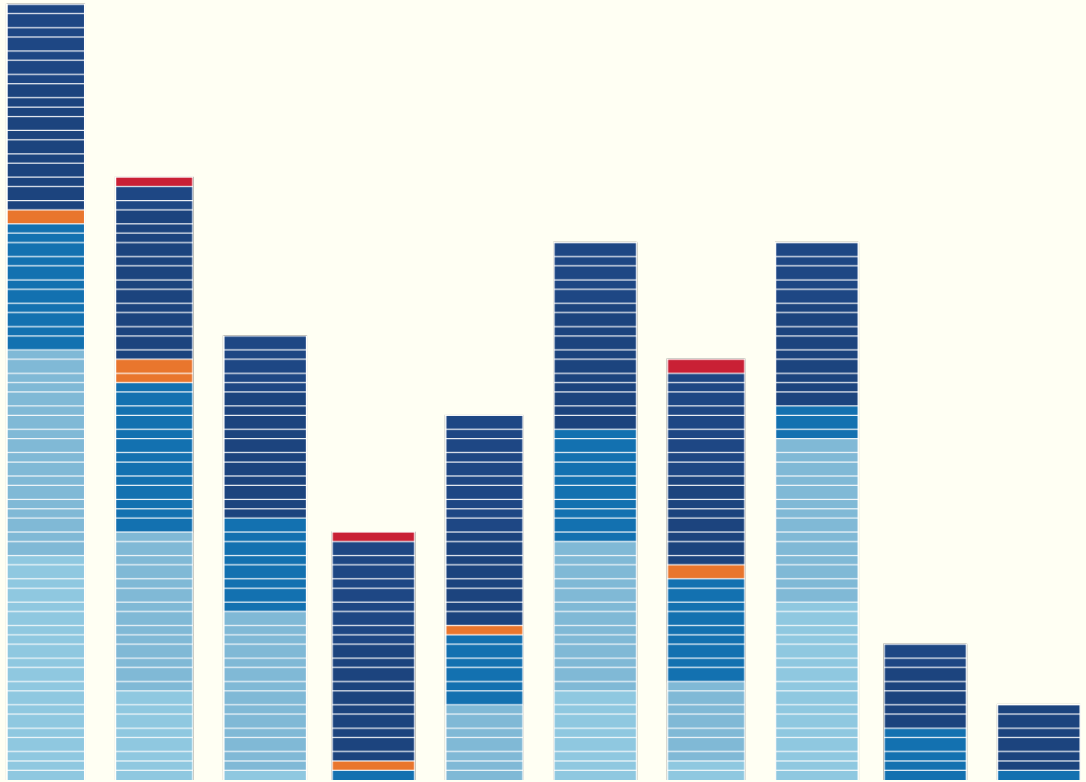


Tableau journalism basics



- Import data
- Designing a Dashboard
- Layout and design
- Filter Actions
- Formatting
- Publishing
- Assignment: design your own dashboard


+ableau®