数据的描述

数据的描述—外在美

描述分析简介

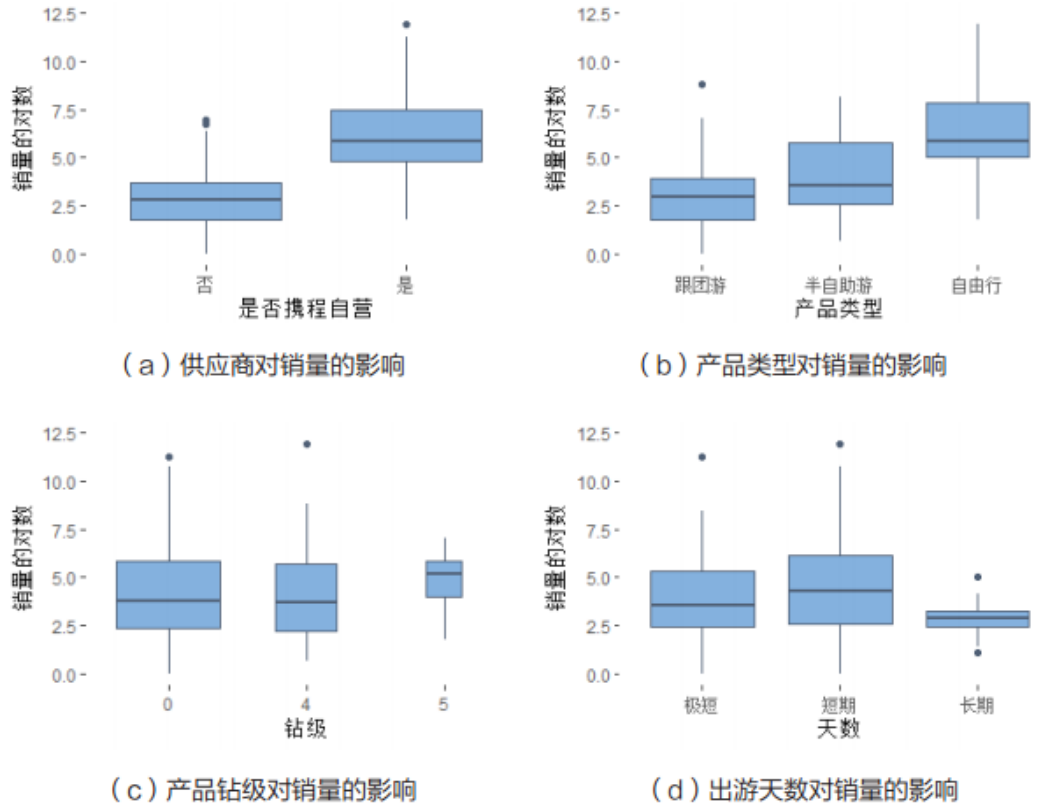
描述分析是数据分析报告中非常重要的环节，主要内容包括：

1. 用**统计图**初步展示数据。
2. 用**统计表**及各种**统计指标**对数据进行描述。
3. 适当**解读**描述的结果。单独展示统计图的意义不大，要学会根据统计图表“讲故事”。

描述分析的整体规范

描述分析的整体规范需要注意**篇幅**、**排版**和**逻辑**

1. 篇幅：注意控制篇幅，有所取舍
2. 排版：图文穿插，简洁美观
3. 逻辑：**归纳分组**，逻辑清楚



（a）供应商对销量的影响　（b）产品类型对销量的影响

（c）产品钻级对销量的影响　（d）出游天数对销量的影响

图 5-11　部分基本属性对产品销量的影响

归纳分组展示统计图

描述分析的整体规范 — 示例

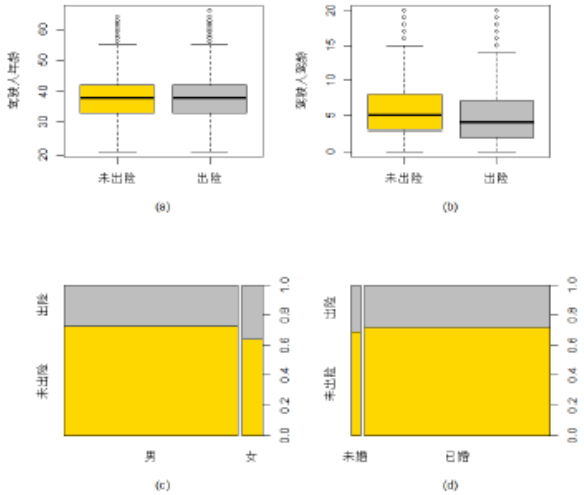


图 3-1：驾驶人因素描述统计图汇总

注：(a) 驾驶人年龄分组箱线图；(b) 驾驶人驾龄分组箱线图；(c) 驾驶人性别棘状图；(d) 驾驶人婚姻状况棘状图。

（二）自变量：汽车因素

案例数据中汽车因素包括六个变量：汽车车龄、发动机引擎大小、是否进口车、所有者性质、是否有固定车位和是否有防盗装置。

首先将车龄变量和引擎大小变量进行离散化处理，即将车龄为 1 年的看作是新车，车龄大于 1 年的看作是旧车；将引擎小于等于 1.6 升的车看作是普通级，引擎大于 1.6 升的看作是中高级。从图 3-2 可以看出，新车出险率更高，普通级车辆出险率更高。因此可以初步判定汽车车龄和车辆级别会影响出险行为。

从图 3-3 可以看出，有防盗装置、有固定车位、进口车以及私人车的出险率略高。值得注意的是，样本量在有无防盗装置、有无固定车位、是否进口车和所有者性质的不同水平之间，分配并不均匀。因此，这种差异是否显著，需要借助后续建模结果进行判断。

规范排版的示例

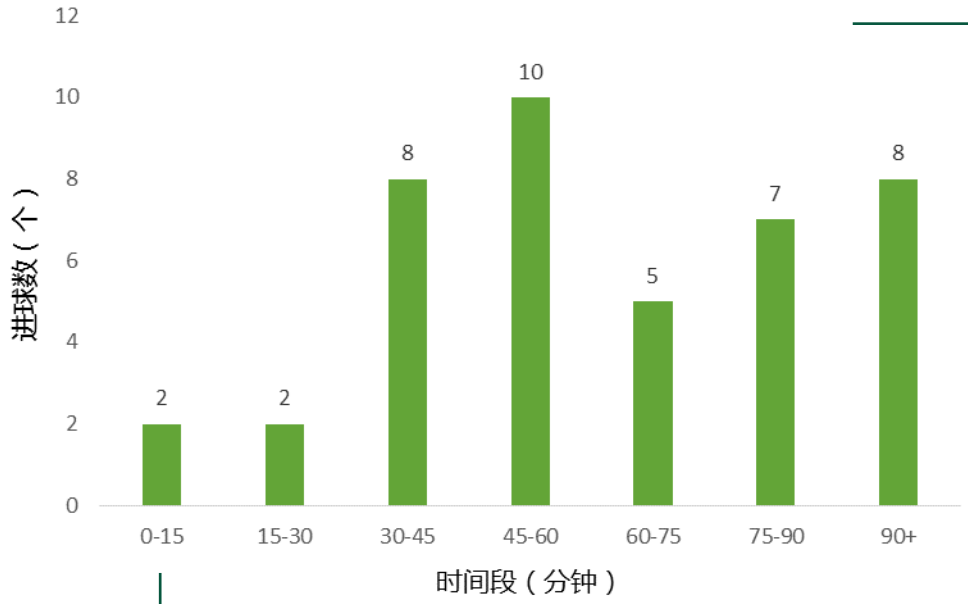
归纳分组，4个统计图放在一组进行展示；节省空间，逻辑清楚

一页报告纸，图文穿插，排版合理

描述分析的整体规范 — 示例

- **横、纵轴要标注清楚，**如果有**单位**的话，需要注明

- 图的标题，横轴纵轴等，出现的**文字要统一和准确**



进球时间分布柱状图

- **图的比例要协调**，别太胖别太瘦，别太高也别太矮

- **图的内容要正确、简明**，避免出现不必要的标签、背景等

- 注意**图的配色**。不要精挑细选一组非常难看的配色！

- 画完图要有**适当的评述**，尤其是在报告里

- 要有图标题，一般在图的下方，标题要简洁明了；报告中的统计图要有标号

描述分析的整体规范 — 示例

作表基本原则（以变量说明表格为例）

表1：二手房数据变量说明

- 要有**表格标题**，一般在表格上方；报告中的表格需有**标号**

变量类型		变量名	详细说明	取值范围	备注
因变量		单位面积房价	单位：万元/平方米	1.83~14.98	
自变量	内部因素	房屋面积	单位：平方米	30.06~299.00	
		卧室数	单位：个	1~5	
		厅数	单位：个	0~3	
		所属楼层	定性变量 共3个水平	低楼层、中楼层、高楼层	相对楼层
	区位因素	所属城区	定性变量 共6个水平	朝阳区、东城区、丰台区、 海淀区、石景山区、西城区	
		是否邻近地铁	定性变量 共2个水平	1代表邻近地铁 0代表不邻近地铁	82.89% 邻近地铁
		是否学区房	定性变量 共2个水平	1代表学区房 0代表非学区房	30.22% 是学区房

- 标明**因变量、自变量**
- 自变量**合理分组**
- 表格中元素为**中文，表意明确（如变量名称）**，切忌直接粘贴code中的英文
- 需要标注**表头**
- 在报告中需对表格内容有简单的文字说明
- 表格中**统一字体、样式**，注意排版不要一行只放单独的一个文字
- 如有小数数字，保留2位小数为宜
- 内容上，以变量说明表格为例，需要对变量进行简单说明，如单位、定量或定性等；定量变量给出范围，定性变量给出取值水平

数据的描述——内在美

好看的统计图
都是相似的
难看的统计图
各有各的丑

作图基本步骤

1 明确数据的含义（数据类型、采集方式、单位等）

- 定性数据：性别、种族等
- 定量数据：工资、房价等

2 找到合适的工具进行描述分析（可视化分析）

3 对描述分析的结果做适当的评述

“准确”是最起码的要求

准确：能够使用正确的统计图去描述不同类型的数据

- 单个变量
  - 定性变量：柱状图、条形图、饼图、环形图
  - 定量变量：直方图、箱线图
  - 时间序列变量：折线图
- 两个变量
  - 两个定性变量：堆积柱状图
  - 两个定量数据：散点图
  - 一个定性一个定量：分组箱线图
- 多个定量数据：相关系数矩阵图

Barplot

柱状图

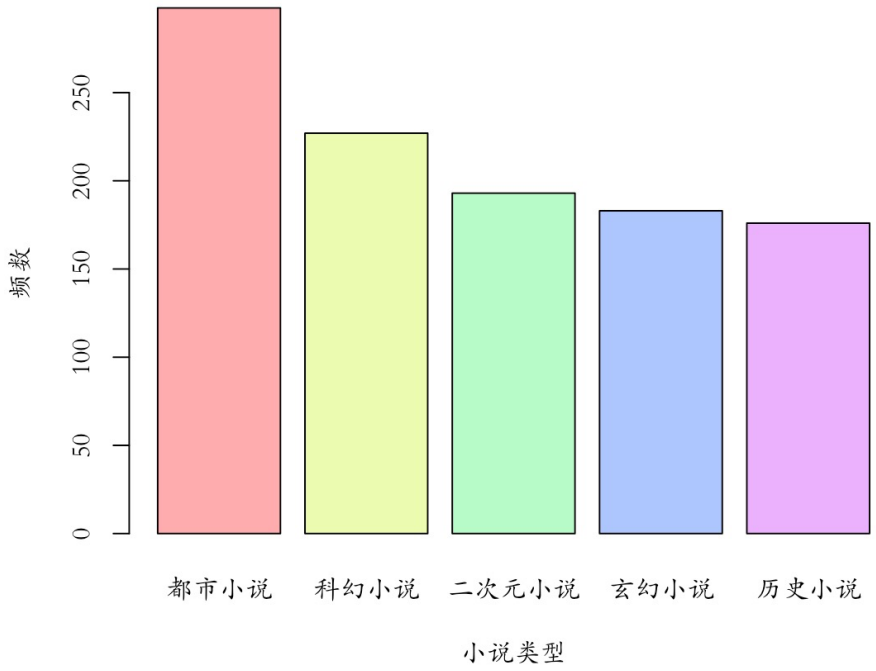
柱状图

柱状图 – 定义

针对 **定性数据**（如：小说类型）

柱子代表 **类别**（都市小说、科幻小说、二次元小说……）

柱子的高度是这个类别的 **频数** 或者 **百分比**



示例与修改

图 4-12 借款用户信用等级分布图

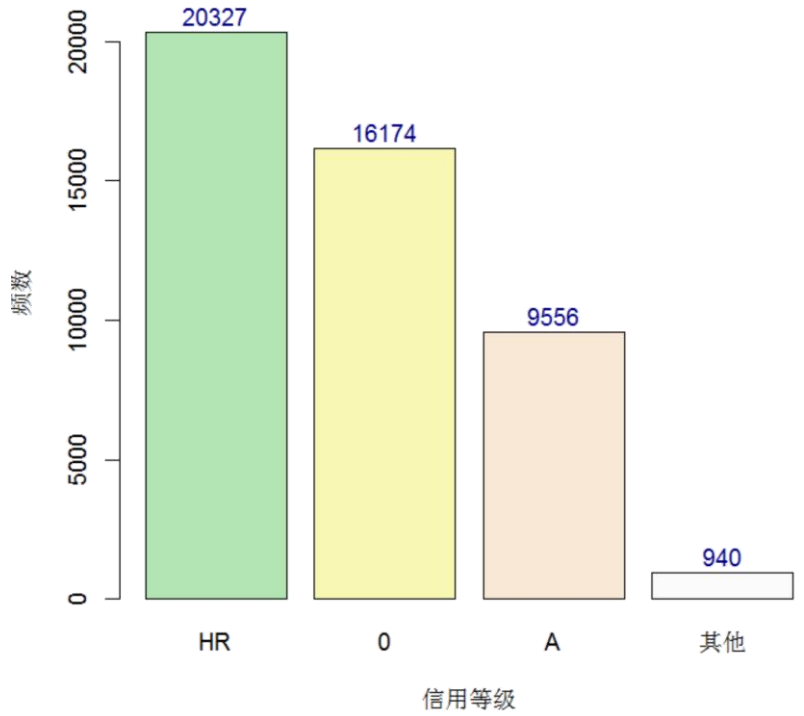


图 xx 借款用户信用等级分布柱状图

存在的问题：

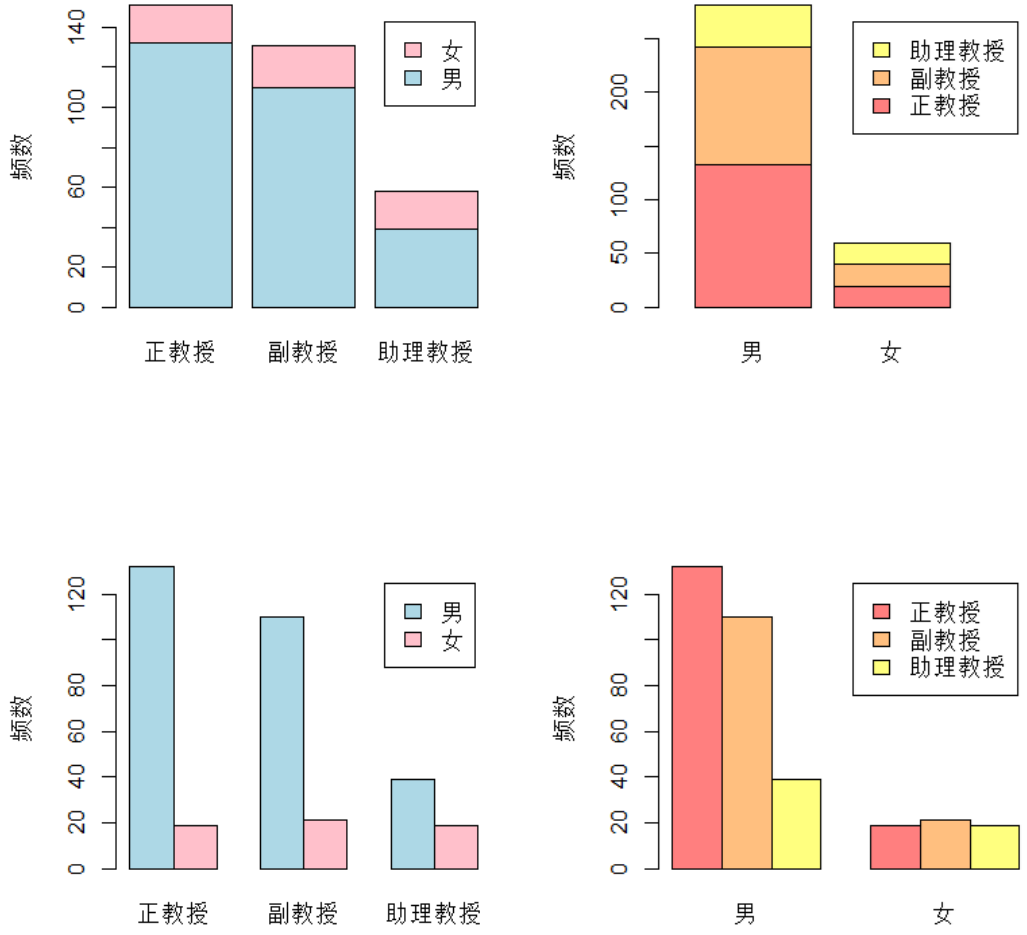
1. 最高的柱子和最矮的柱子相差太多，影响美观。改进办法：按照某种顺序排序；将频数较小的类别合并。
2. 柱子之间没有空隙。
3. 图表题不准确：“分布图”是一种很笼统的称呼，应该准确地叫“柱状图”。

堆积柱状图

堆积柱状图 – 定义

展示两个 **定性数据**（性别&职称）的 **频数分布**

- **注1**：也可用于展示
  - 一个定性数据和一个定量数据
  - 两个定量数据
  - 需要将定量数据离散化
- **注2**：不适合在柱子上标注交叉频数，会显得混乱



示例与修改

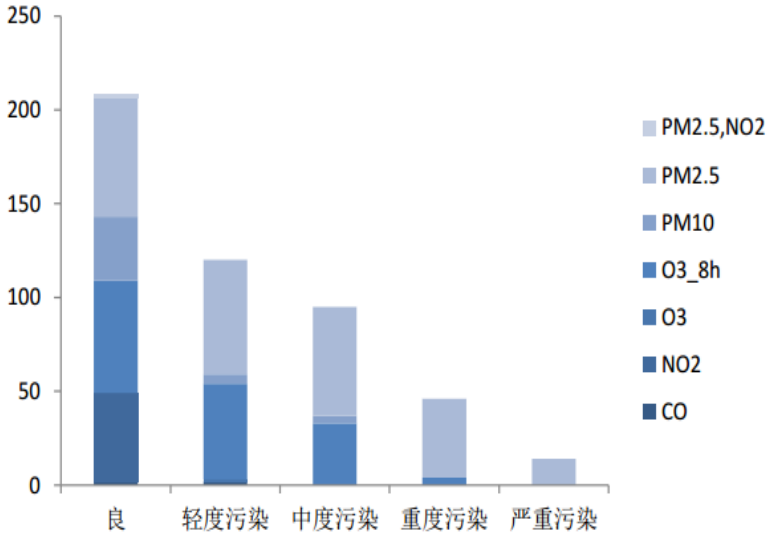


图 5-1　北京市不同空气质量指数类别下首要污染物分布图

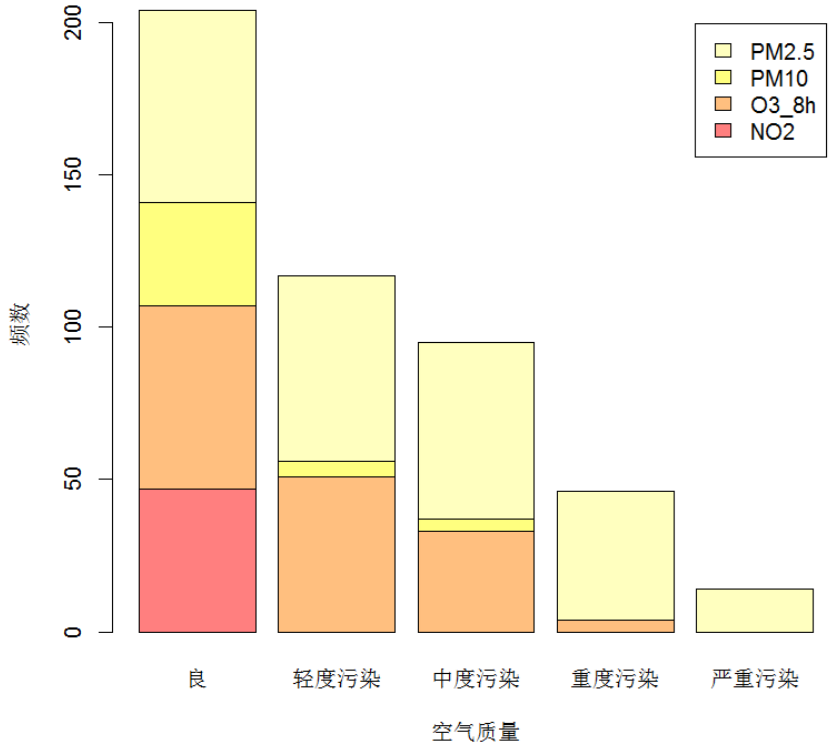


图 xx 北京市空气污染物分布柱状图

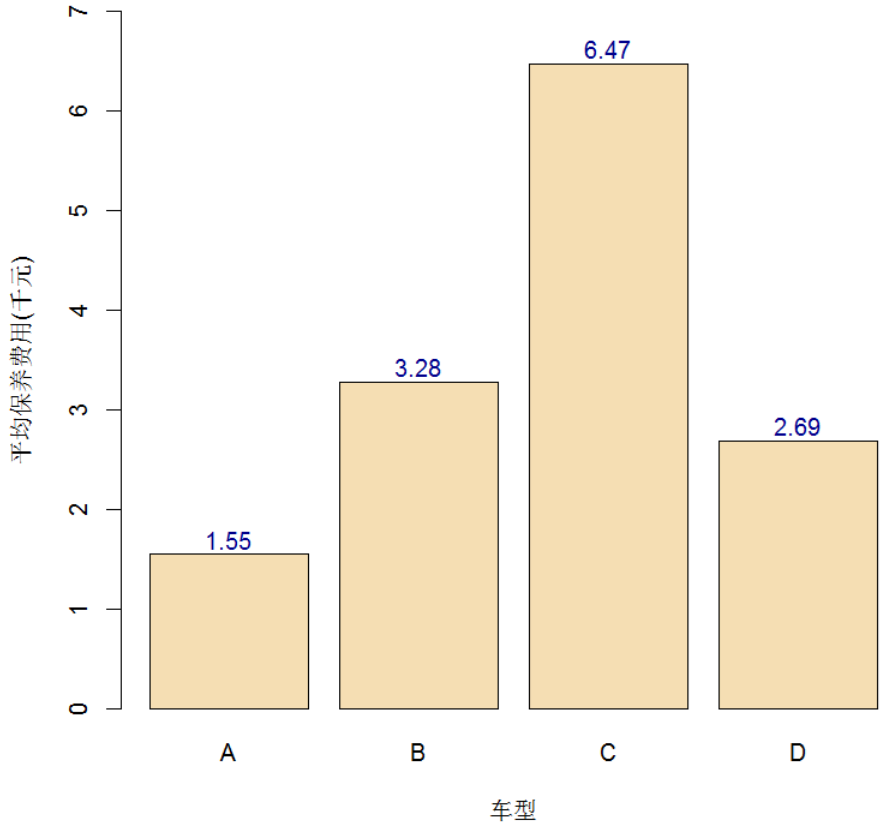
存在的问题：

1. 柱状图涉及的颜色过多，且区分度较差，让读者无法对应。
2. 柱子只显示了4中颜色，但是图例却有7种污染物。
3. 查看原始的频数分布表，有些类别频数太低导致没有显示。

柱状图的其他用处

柱状图的其他用处

- 展示常用的 **统计量**，如均值
- 假设样本数据包含1000辆车，四种车型（A、B、C和D）
- 现在想比较，不同车型在2015年全年的 **平均保养花销**
- 柱高代表平均保养花销



Pie Chart

饼图

饼图有多可怕

你想象自己画的饼图

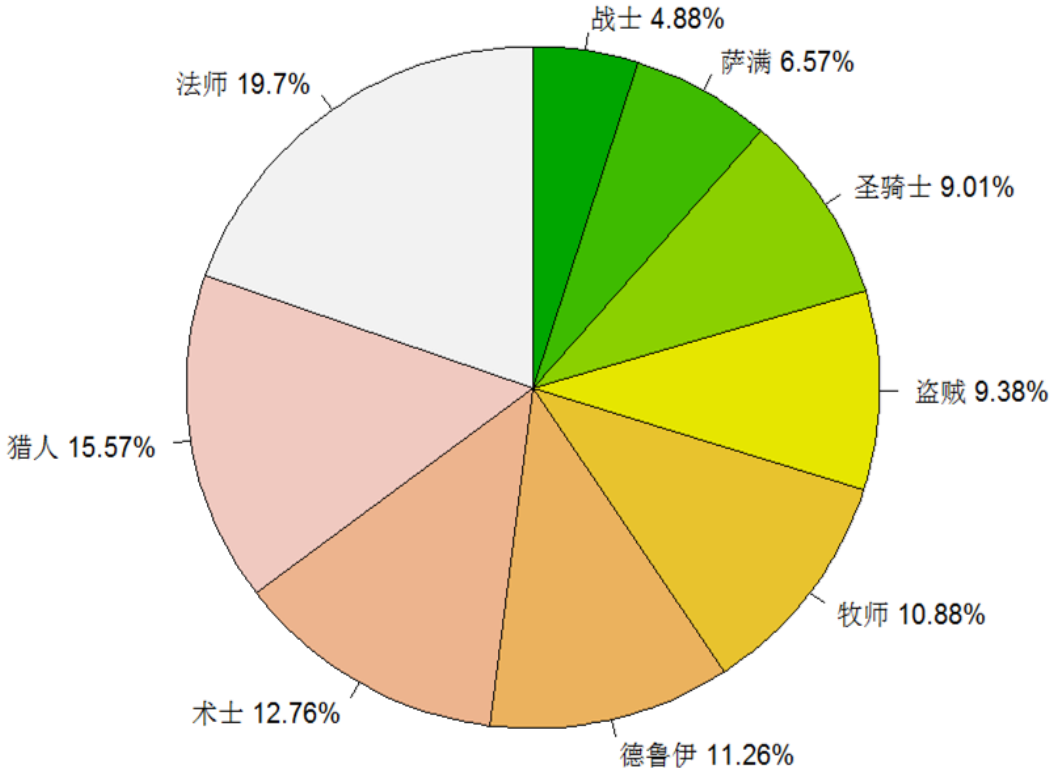
实际上你画的饼图

老师眼中你画的饼图

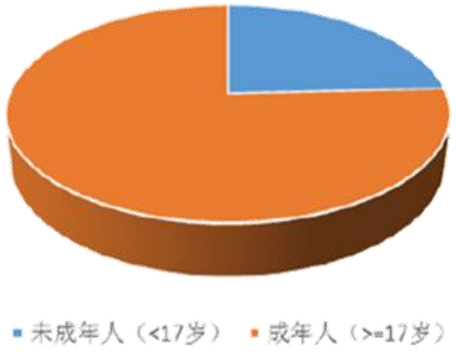
饼图

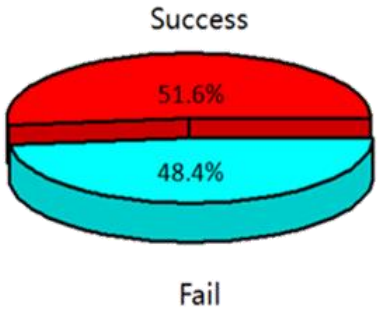
饼图 – 定义

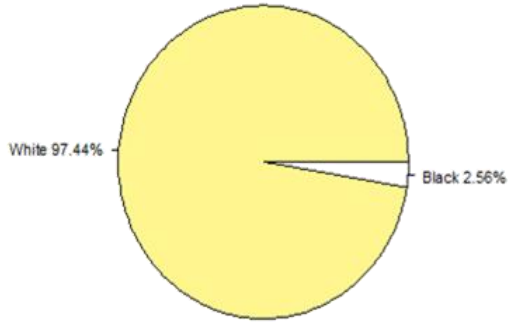
- 针对 **定性数据**
- 柱状图多用于展示频数
- 饼图多用于展示 **百分比**

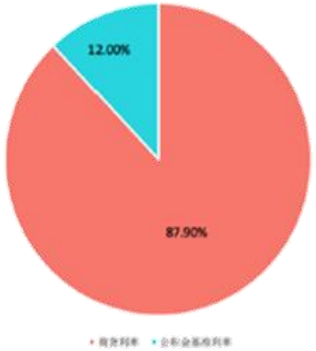


示例与修改



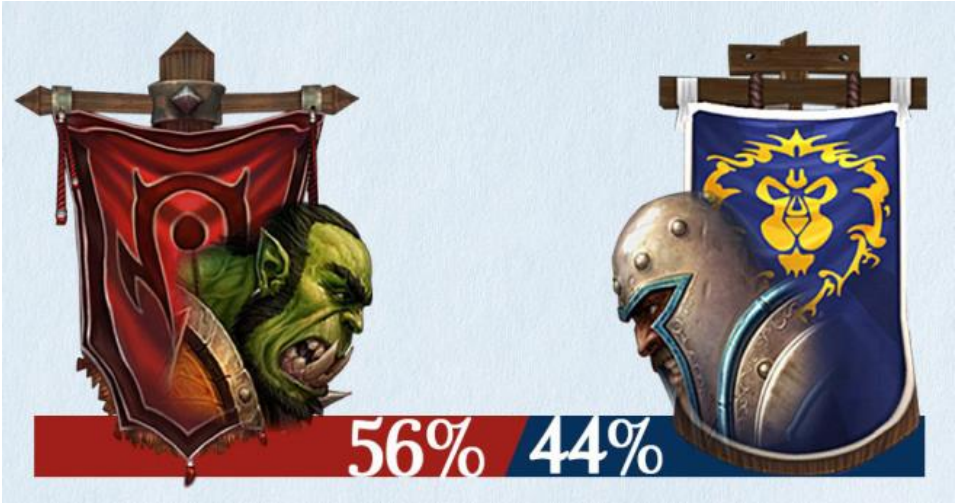




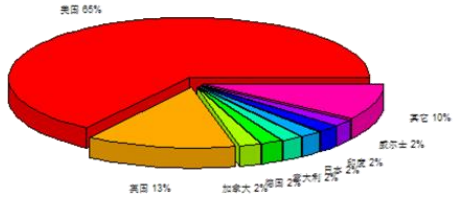


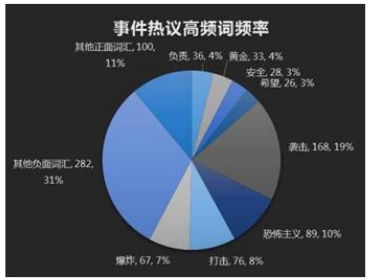
问题与解决办法：

1. 只有两个类别，信息量太少，导致饼图不美观。
2. 在PPT中，没有必要占篇幅汇报这样的饼图。
3. 建议直接在报告中进行文字汇报（例如右上）：“样本数据中，成功的比例为51.6%”。



示例与修改





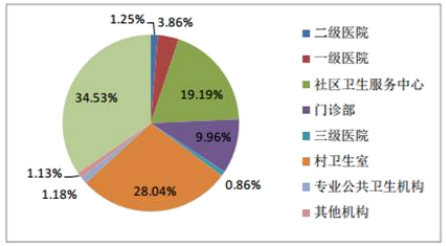
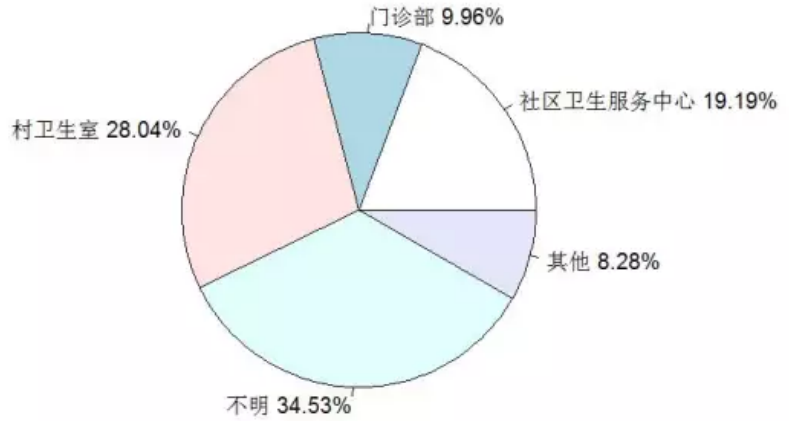


图 3-4　2014 年北京市医疗卫生机构饼状图

问题与解决办法：

1. 类别过多，眼花缭乱。
2. 将占比较少的归类，叫做“其他”，或者绘制条形图。
3. “标签”标注在饼的旁边，方便对应，而非做成图例。无需同时标注频数和百分比。



图示 2014年北京市医疗卫生机构分布饼图

注：“其他”包括二级医院（1.25%）、一级医院（3.86%）、三级医院（0.86%）、专业公共卫生机构（1.18%）和其他机构（1.13%）

左侧“右下角”饼图的改进效果

没事儿别画饼图！

引自R软件饼图的HELP：

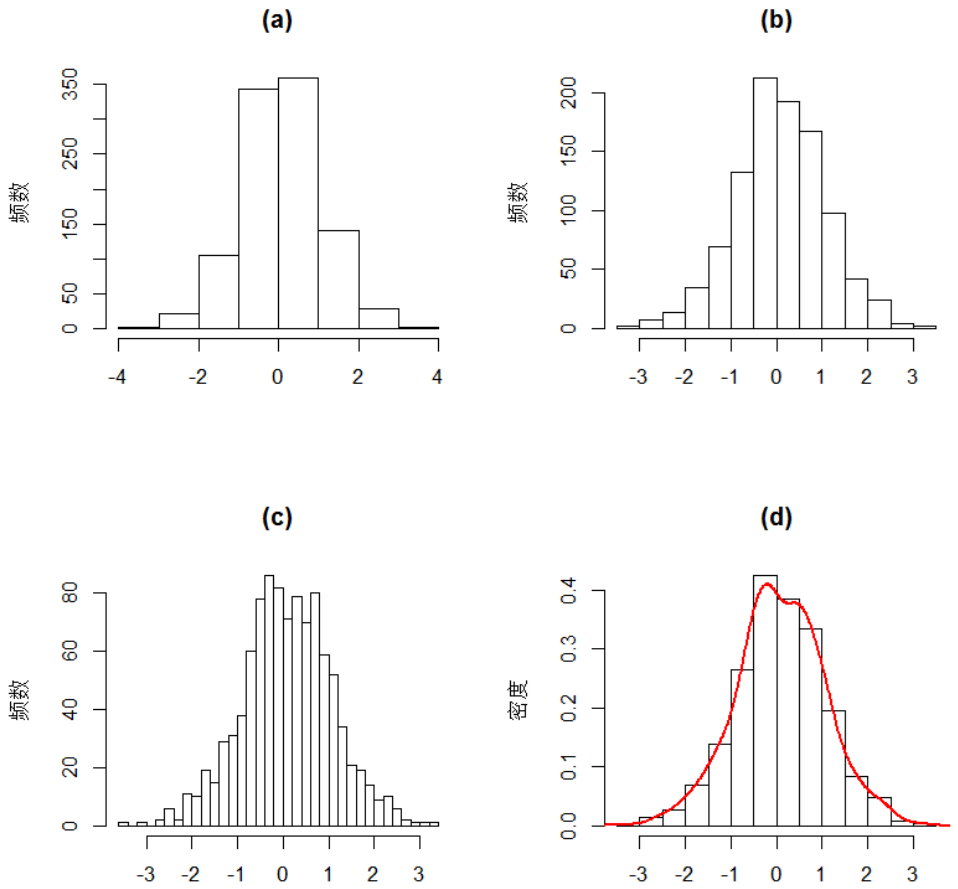
Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

没事儿别画饼图！

Histogram

直方图

直方图

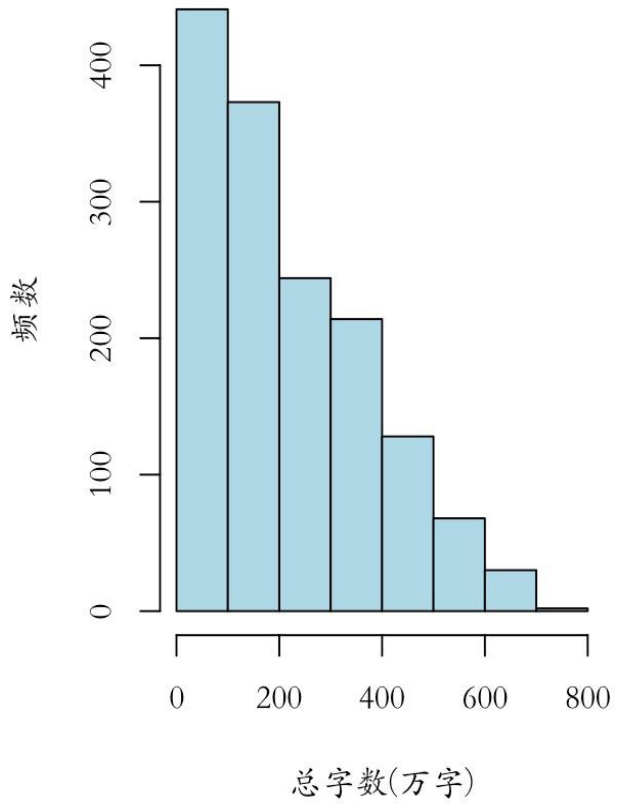


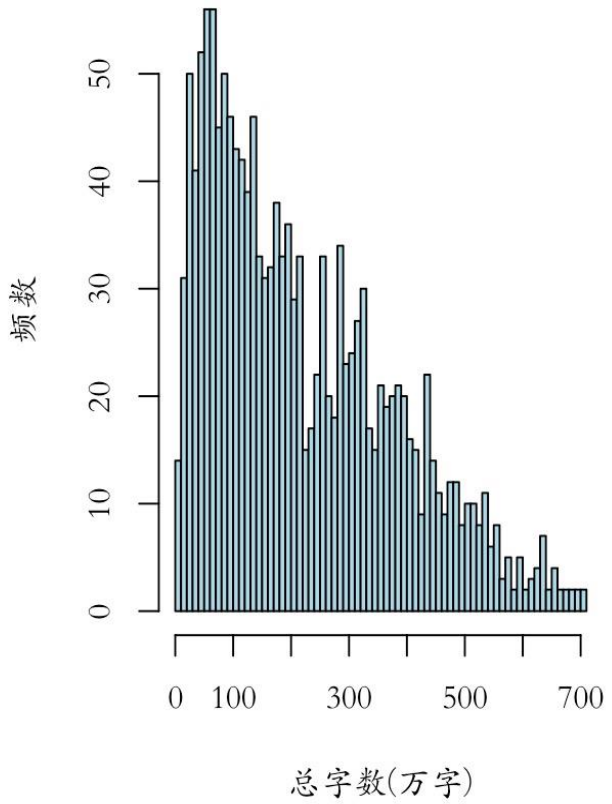
直方图 – 定义

- **针对定量数据**
- **横轴**是实数轴，被分成了许多连续的区间
- **纵轴**，有两种处理方式，一是代表频数，如图(a)—(c)；二是代表密度，如图(d)
- **频数**：数一数落在相应区间内的样本数。从(a)到(c)，区间越来越“窄”，数据的分布形态也被展示的越来越“细”

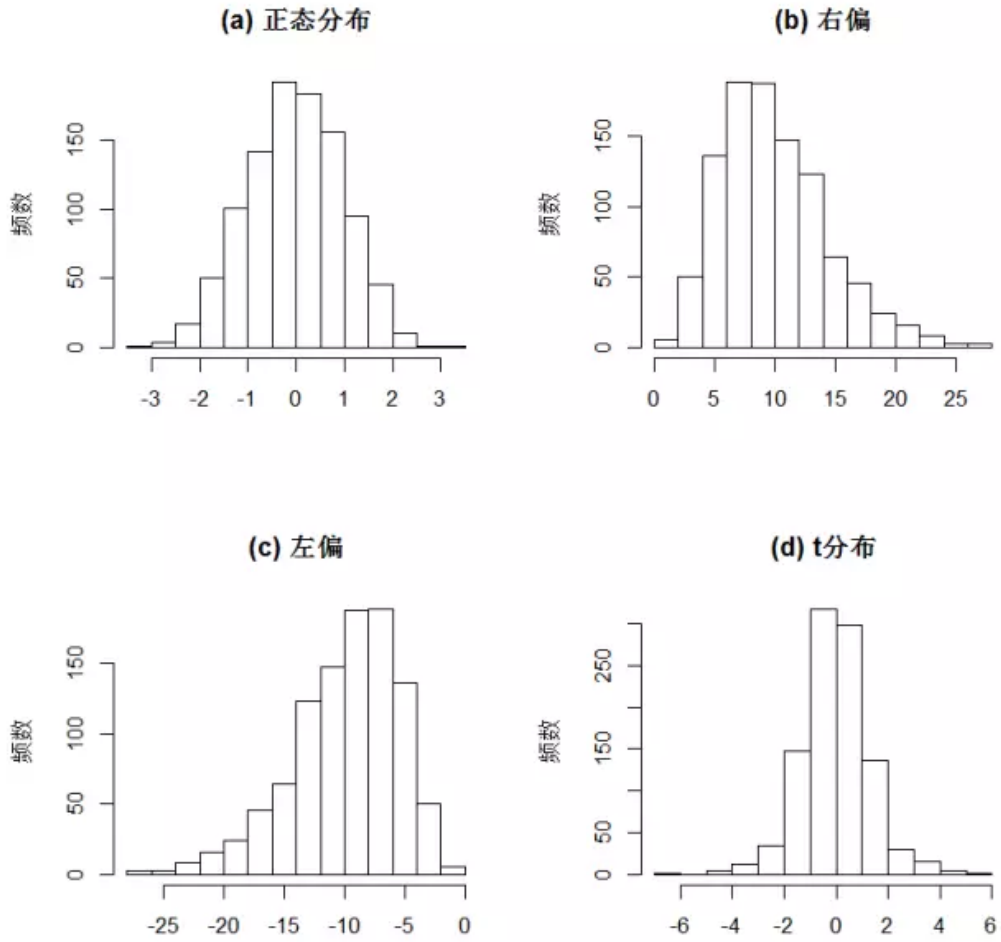
直方图

小说数据


频数
0
100
200
300
400
0
200
400
600
800
总字数(万字)


频数
0
10
20
30
40
50
0
100
300
500
700
总字数(万字)

直方图的应用



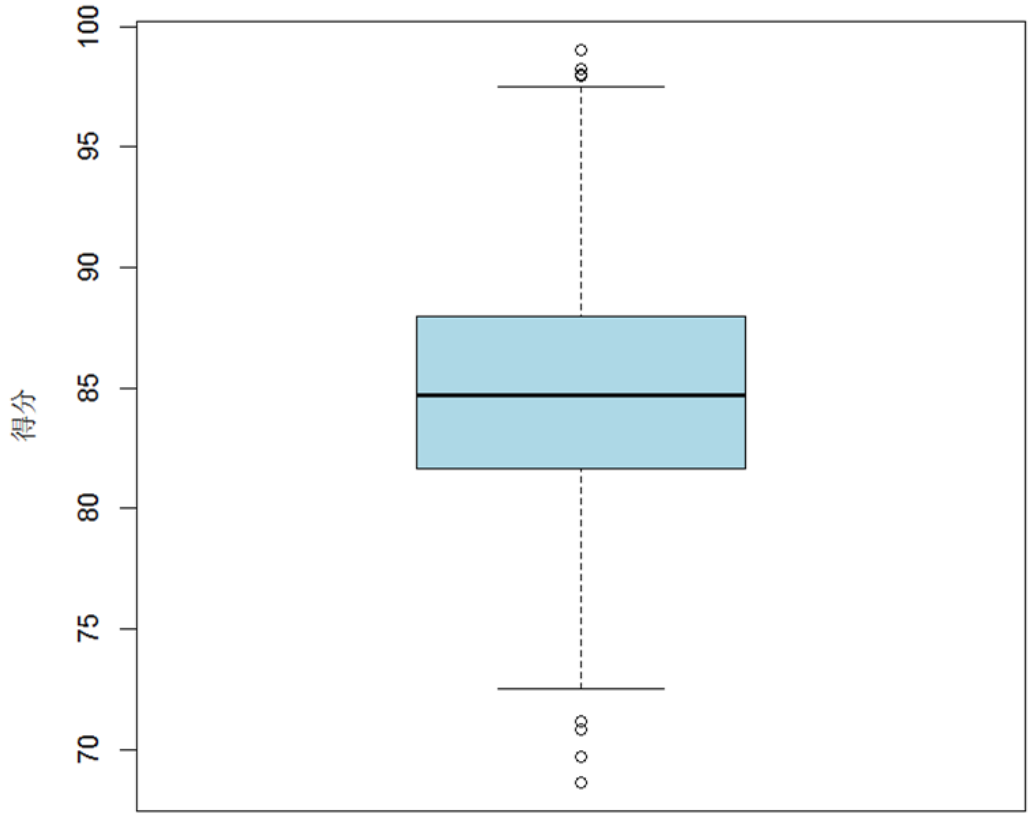
直方图 – 应用

- 观察数据分布的**形态**
- 直方图的**“尾巴”**在哪里，就是往哪里偏，仿佛新娘婚纱的拖尾。
- **t分布**的直方图，呈现非常典型的**“厚尾”**现象，两边各拖着一个小“尾巴”。相比正态分布，有更大的可能产生极端值。

Boxplot

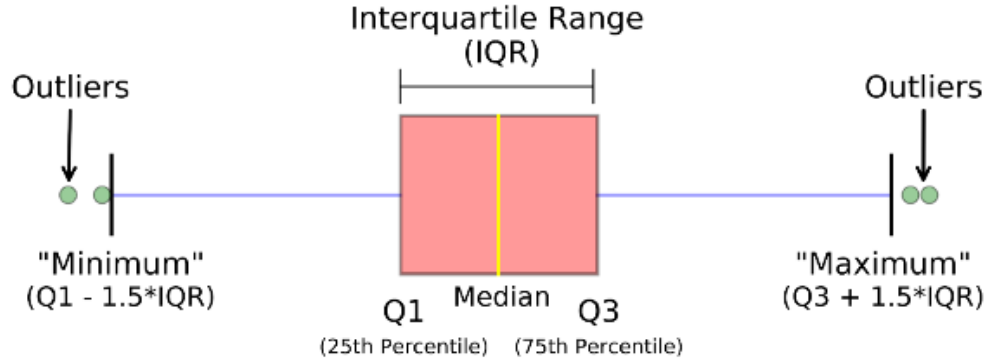
箱线图

箱线图



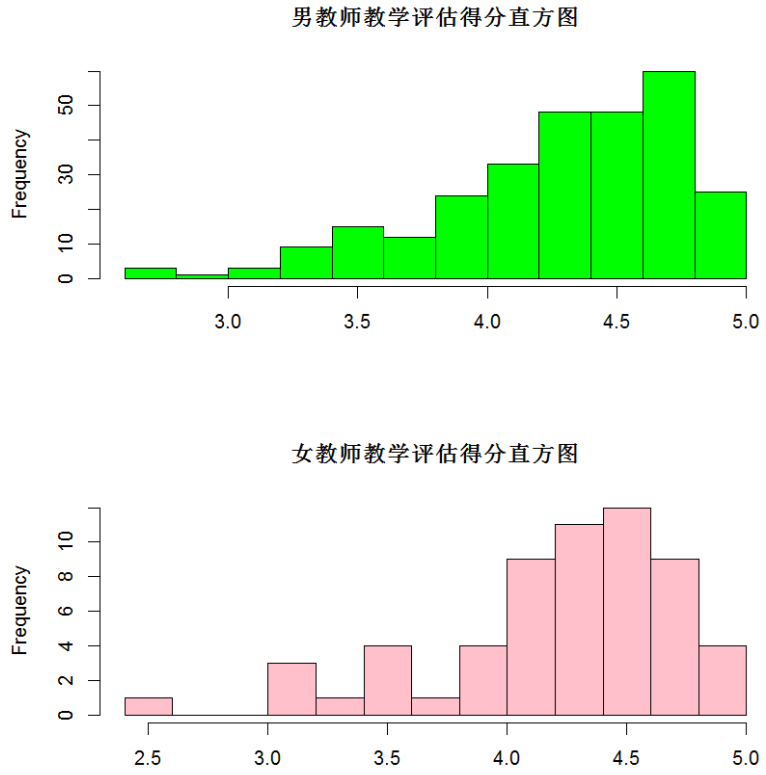
箱线图 – 定义

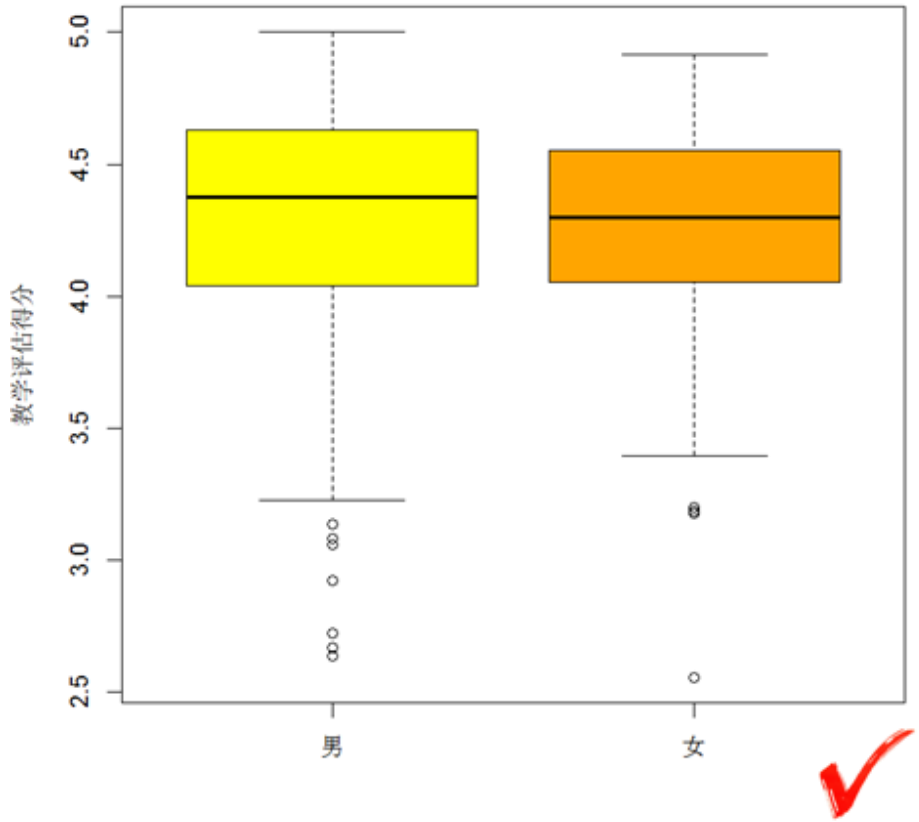
- 中线是**中位数**，代表了样本数据的**平均水平**
- 箱子的上下线，分别是数据的**上、下四分位数**
- 箱子包含了50%的数据，箱子的宽度在一定程度上反映了数据的**波动程度**



箱线图的应用

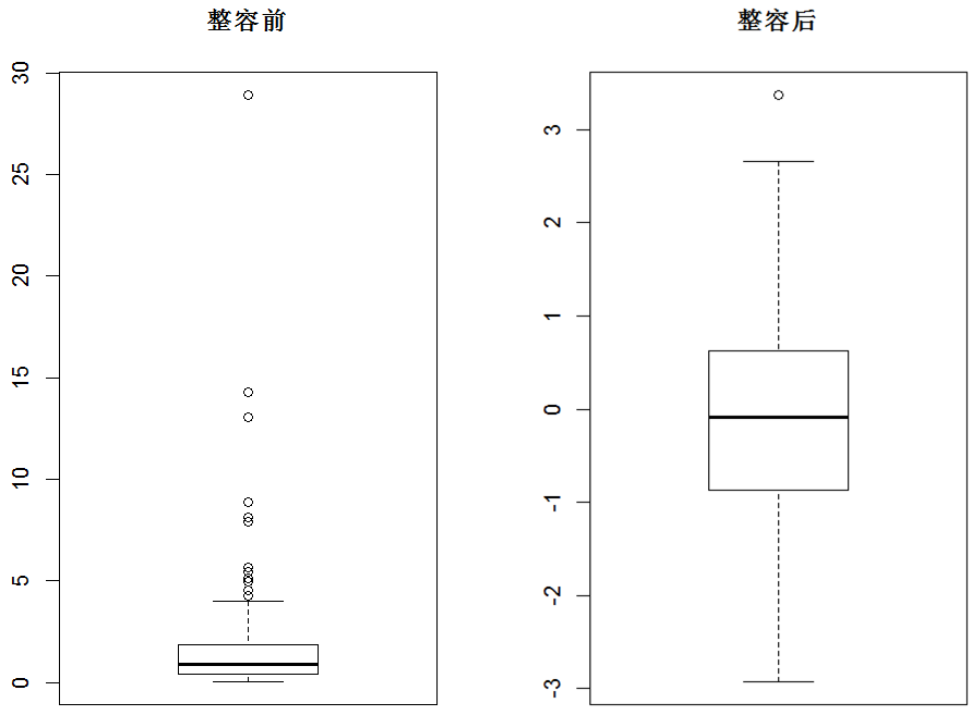
箱线图最好的应用是分组做比较





推荐一款整容神器：对数变换

箱子被压得很“扁”怎么办?



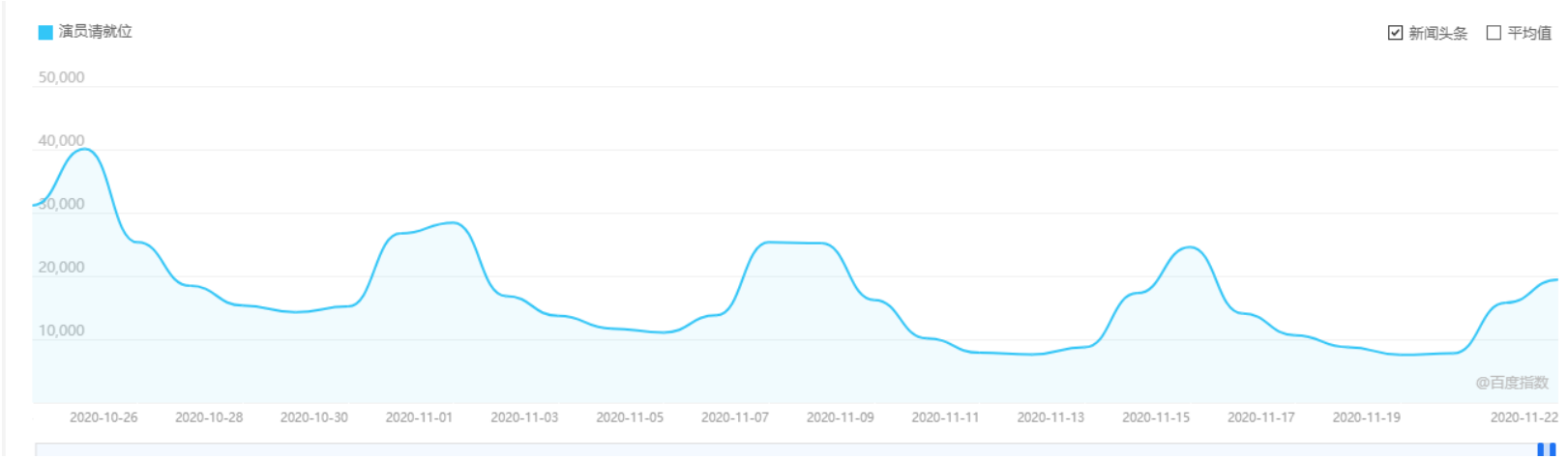
- 数据中的异常或者非常偏态的分布都会导致箱子被压扁。
- 如果数据取值为正数（如工资、房价等），那么可以尝试做**对数变换**。
- **对数变换**：画图界的整容神器，专治各种不对称分布、非正态分布和异方差现象等。

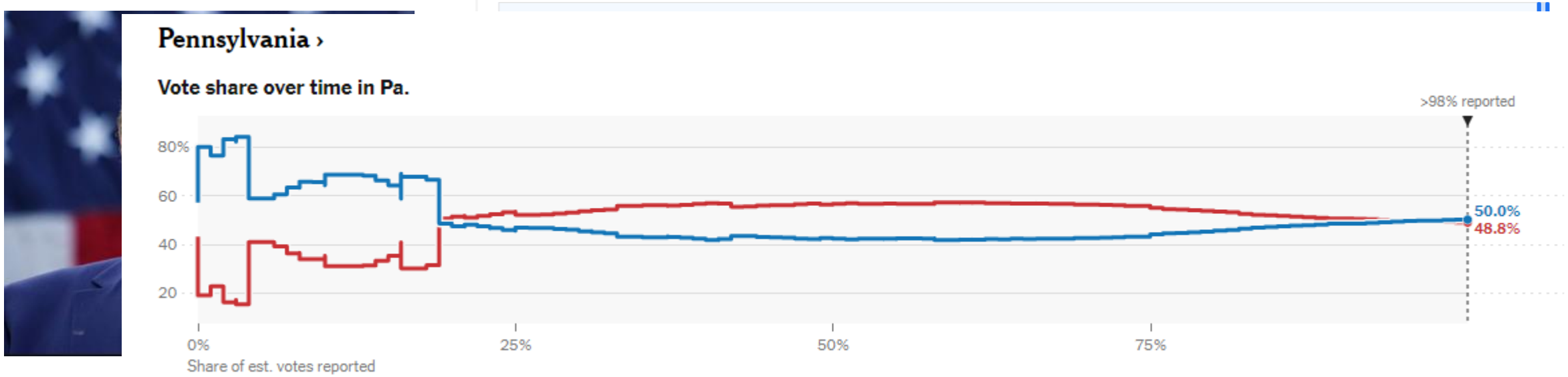
Line Chart

折线图

折线图

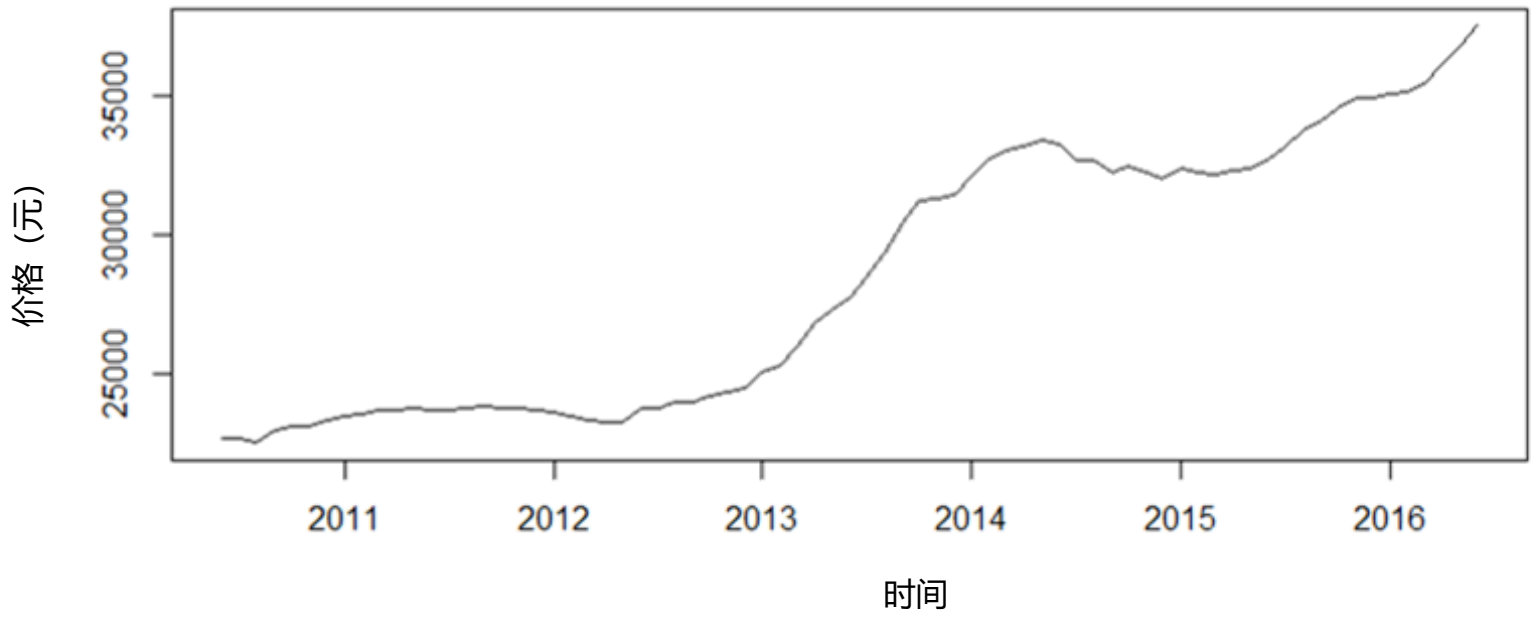
- 横轴：时间
- 纵轴：指标的取值





折线图的应用

折线图 – 应用



1 看趋势

指标随着时间的变化，呈现递增、递减、还是持平的趋势。

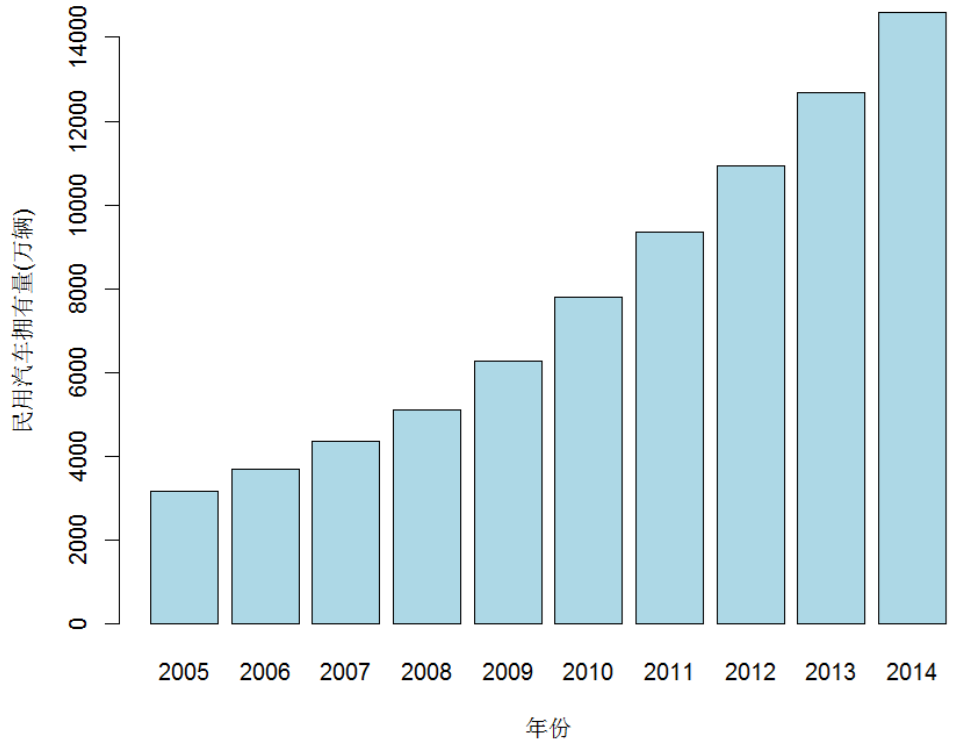
2 看周期

指标的取值，是否呈现一定的周期规律。

3 看突发事件

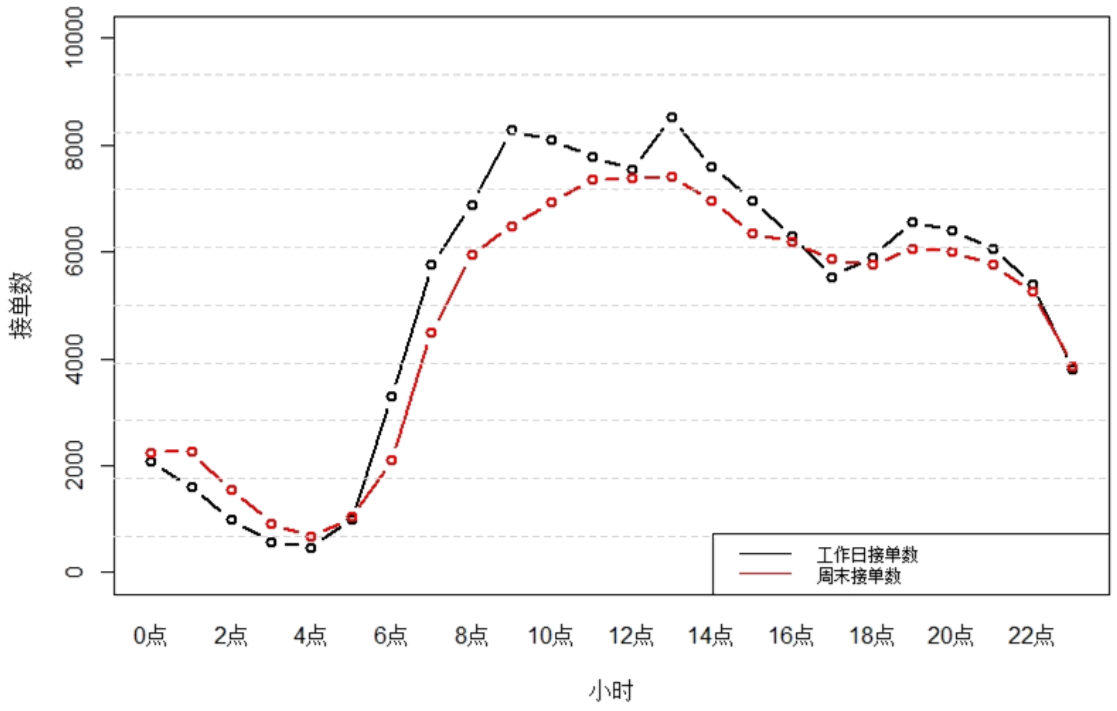
指标的取值，是否因为某个时间的发生，出现波峰或者波谷。

两点说明 2/1



- 有时候，经济指标的变化趋势，惯用**柱状图**，而非折线图。
- 左图是民用汽车拥有量随时间变化的柱状图，柱高代表民用汽车拥有量，本质上跟折线图一个道理。

两点说明 2/2

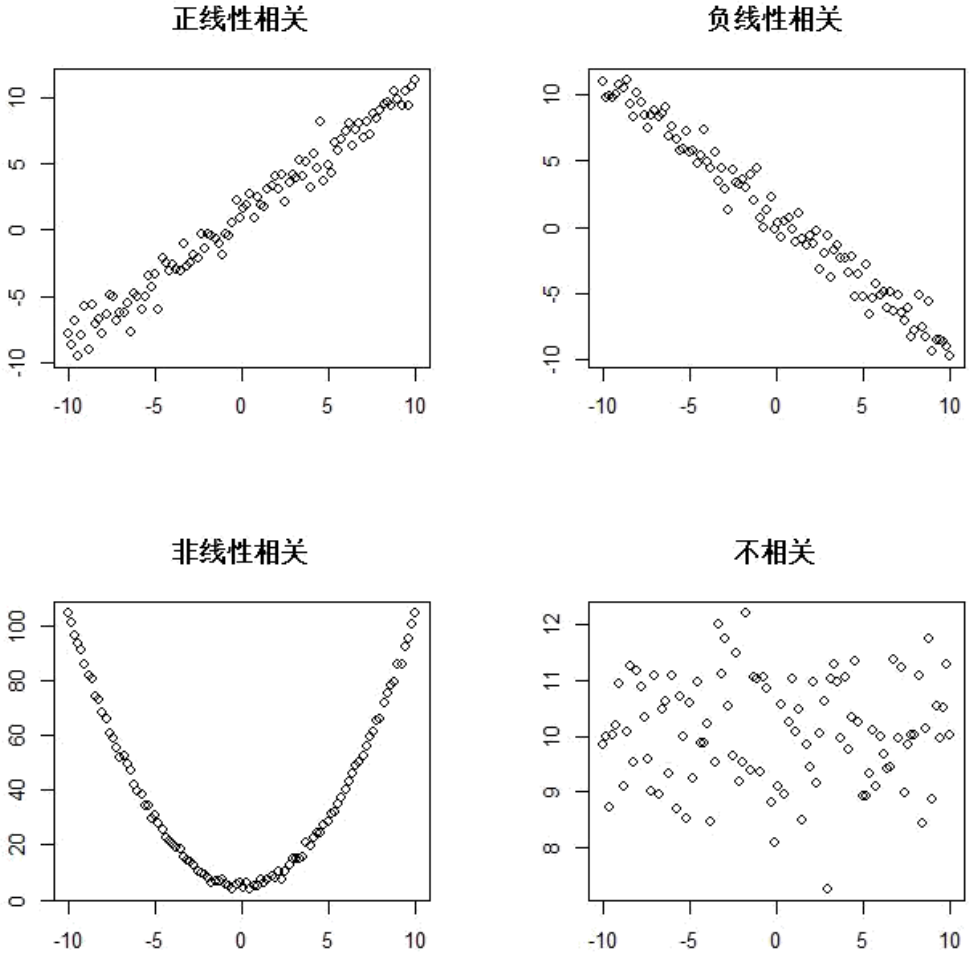


- 平时 vs. 周末的出行规律**对比**
- 工作日和周末出租车每小时接单数变化趋势基本相同
- 在上午8点到下午2点的时段，出租车工作日接单数大于周末接单数；
- 在凌晨时段，周末的接单数多于工作日

Scatter Plot

散点图

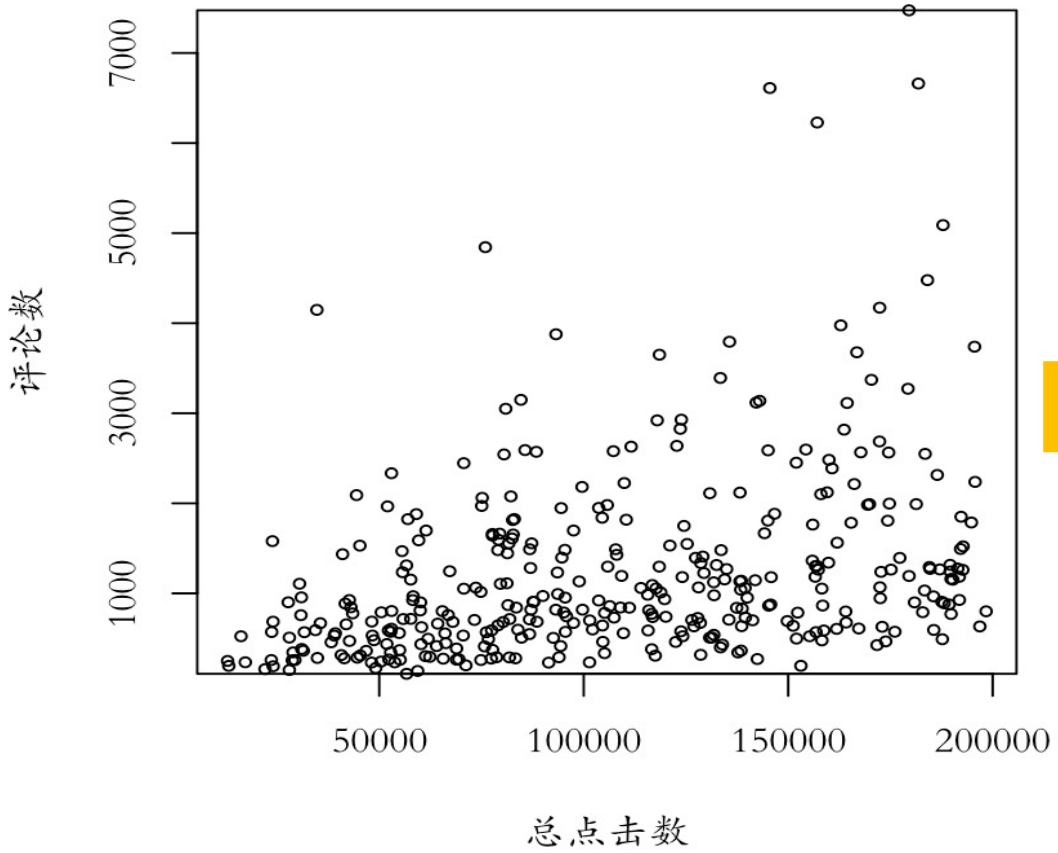
散点图

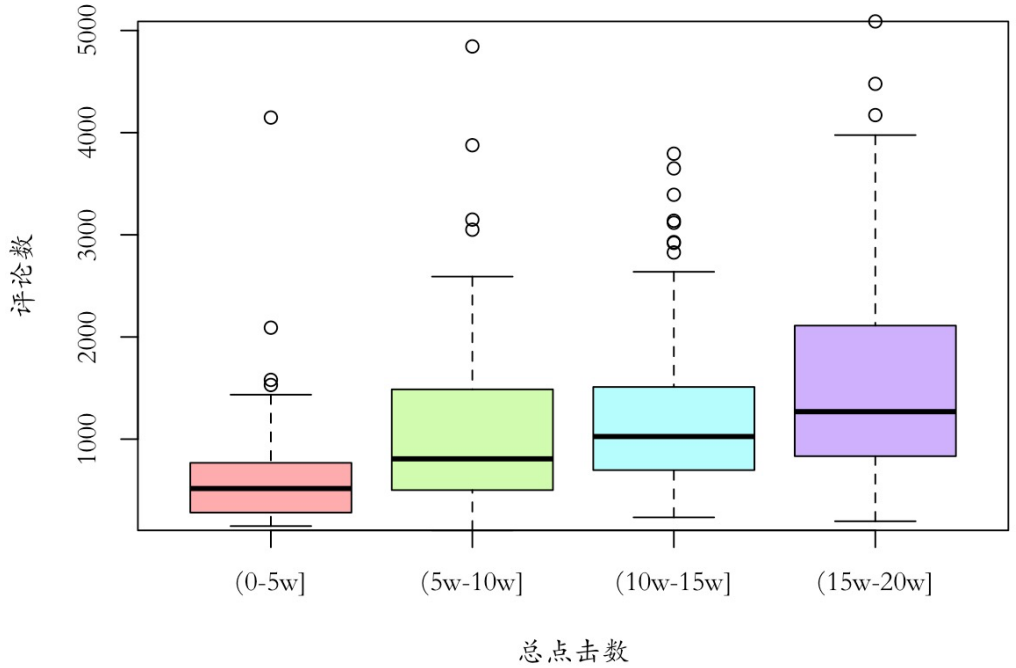


散点图 – 定义

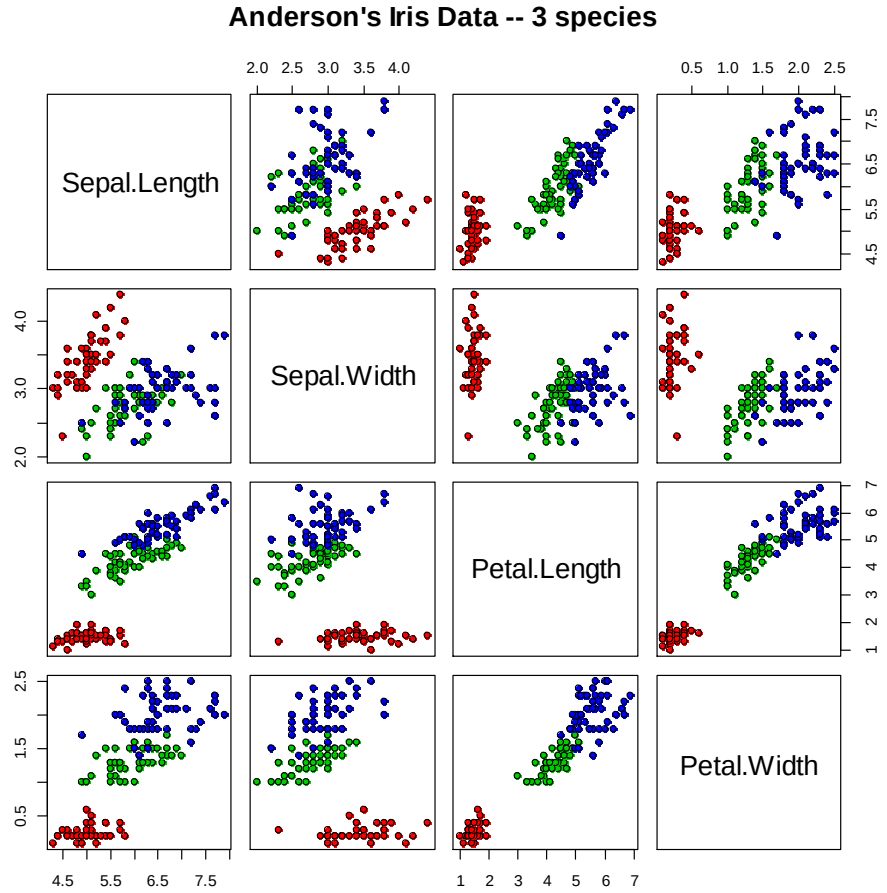
- 散点图是针对两个**定量**变量所做的统计图
- 从散点图上，可以解读两个变量的相关关系
  - 线性相关关系（正向、负向）
  - 非线性相关关系
  - 不相关

现实中的散点图


评论数
1000
3000
5000
7000
50000
100000
150000
200000
总点击数

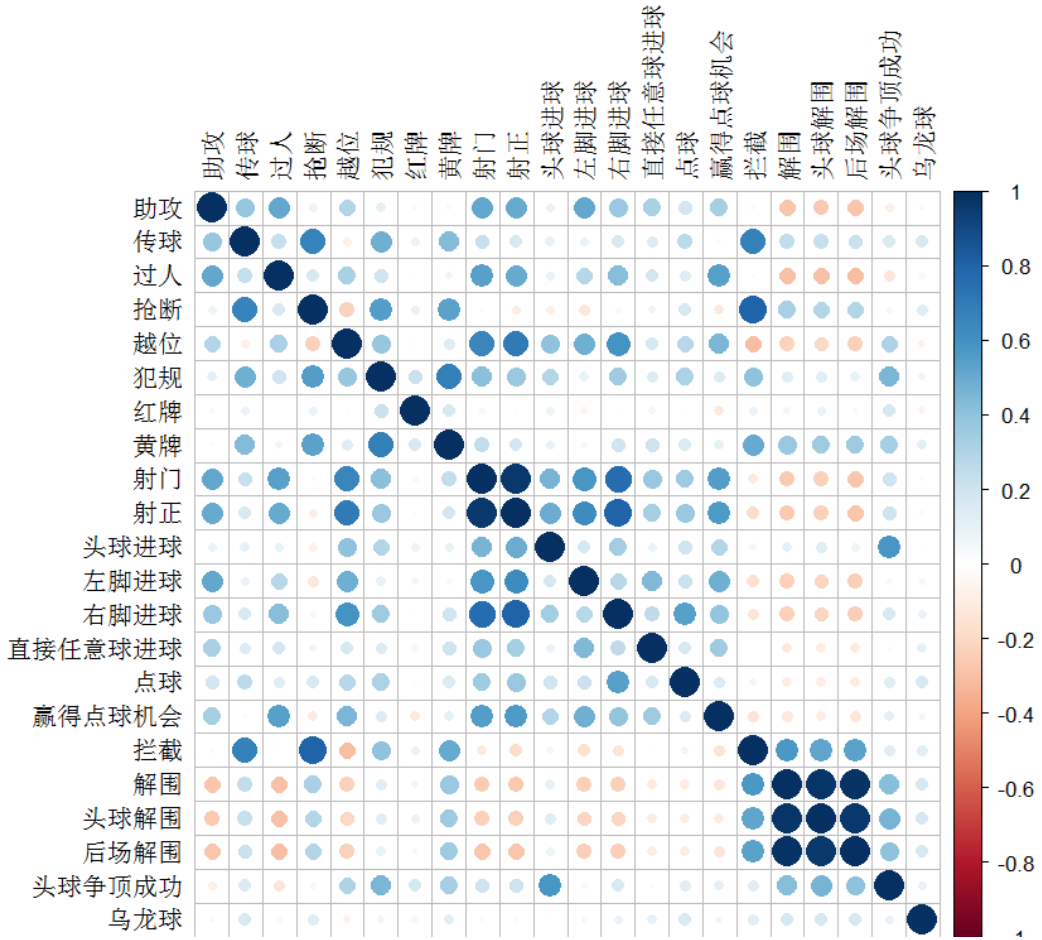

评论数
1000
2000
3000
4000
5000
(0-5w]
(5w-10w]
(10w-15w]
(15w-20w]
总点击数

散点图 – 扩展 2/1



- 针对多个定量数据
- 成对画散点图
- 输出**散点图“矩阵”**
- **注意：**变量不易过多，否则散点图矩阵会非常混乱

散点图 – 扩展 2/2

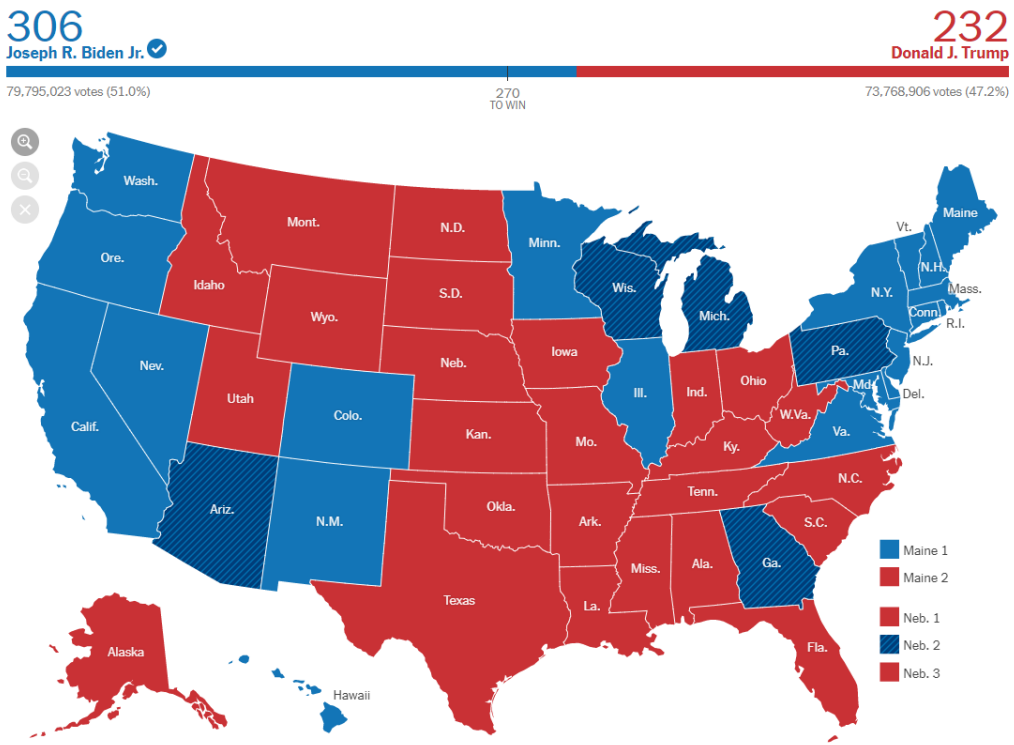


- **相关系数矩阵图**
- 计算两两变量之间的线性相关系数（取值范围-1到1）
- 线性相关系数的取值越接近1，说明正相关性越强；越接近-1，说明负相关性越强；接近0，说明不存在线性相关关系

其他统计图形

美国总统大选

		Hillary Clinton DEMOCRAT	Donald J. Trump REPUBLICAN	Change from 2012
Sex	Male	41%	53%	5 ◀
	Female	54%	42%	1 ▶
Race	White	37%	58%	1 ◀
	Black	88%	8%	7 ◀
	Hispanic/Latino	65%	29%	8 ◀
	Asian	65%	29%	11 ◀
	Other	56%	37%	1 ◀
Age	18-29	55%	37%	5 ◀
	30-44	50%	42%	1 ▶
	45-64	44%	53%	5 ◀
	65 and over	45%	53%	4 ▶
Education	High school or less	45%	51%	
	Some college/associate degree	43%	52%	10 ◀
	College graduate	49%	45%	8 ▶
	Postgraduate study	58%	37%	8 ▶
Education by race	White college graduates	45%	49%	10 ▶
	White without a college degree	28%	67%	14 ◀
	Nonwhite college graduates	71%	23%	
	Nonwhite without a college degree	75%	20%	



体育中的数据分析

收藏 | 33182 | 1230

克里斯蒂亚诺·罗纳尔多 (葡萄牙足球运动员)

编辑

同义词 C罗一般指克里斯蒂亚诺·罗纳尔多（葡萄牙足球运动员）

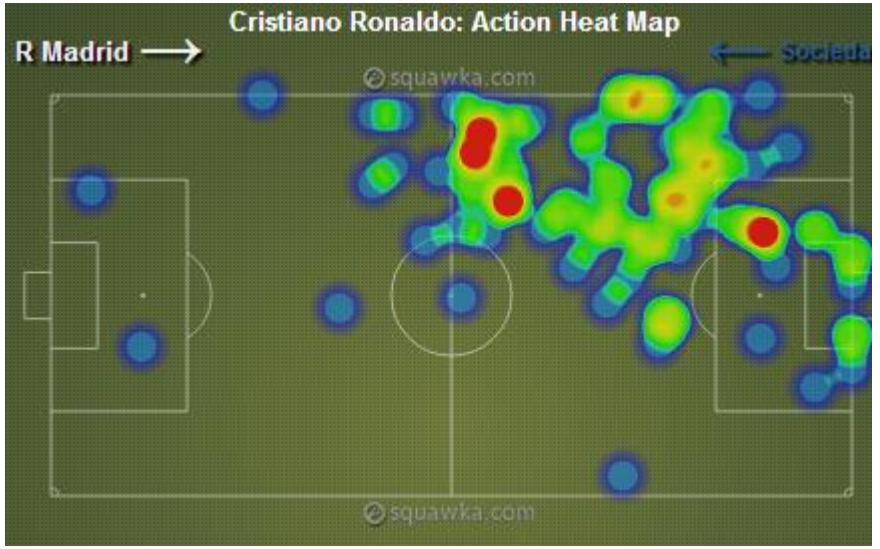
声明 本词条已参考体育类词条编辑指南进行整理。点击了解本词条维护团队。

克里斯蒂亚诺·罗纳尔多·多斯·桑托斯·阿维罗（Cristiano Ronaldo dos Santos Aveiro），简称"C罗"，1985年2月5日出生于葡萄牙马德拉岛丰沙尔，葡萄牙职业足球运动员，司职边锋、中锋，效力于意大利尤文图斯足球俱乐部，并身兼葡萄牙国家男子足球队队长。[1-2]

C罗出道于里斯本竞技。2003年加盟英超曼联，期间获得了英格兰足球超级联赛冠军、欧洲冠军联赛冠军、世俱杯冠军等十个赛事冠军。2009年6月以身价9600万欧元转会至西甲皇马，期间获得了4次欧洲冠军联赛冠军、2次西甲联赛冠军、3次世俱杯冠军等十六个赛事冠军。C罗效力皇马9年时间，438场比赛贡献450球、131次助攻，以场均1.03球的进球率成为皇马历史上进球率最高的球员。[3]

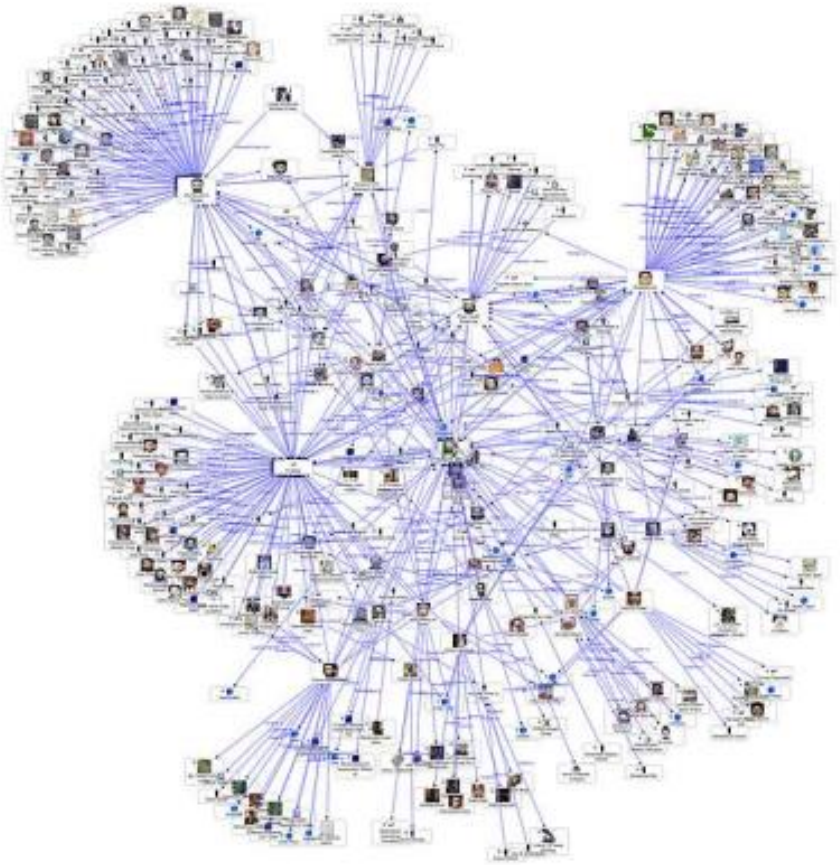
C罗职业生涯保持着多项个人记录，包括欧洲五大联赛个人总进球记录、皇马俱乐部个人总进球记录、欧冠联赛个人总进球记录、欧洲国家队个人总进球记录等。C罗已5次获得金球奖、3次获得世界足球先生、4次获得欧洲金靴奖、7次获得欧冠最佳射手等个人荣誉。

克里斯蒂亚诺·罗纳尔多图册



社交网络数据

Terrorist

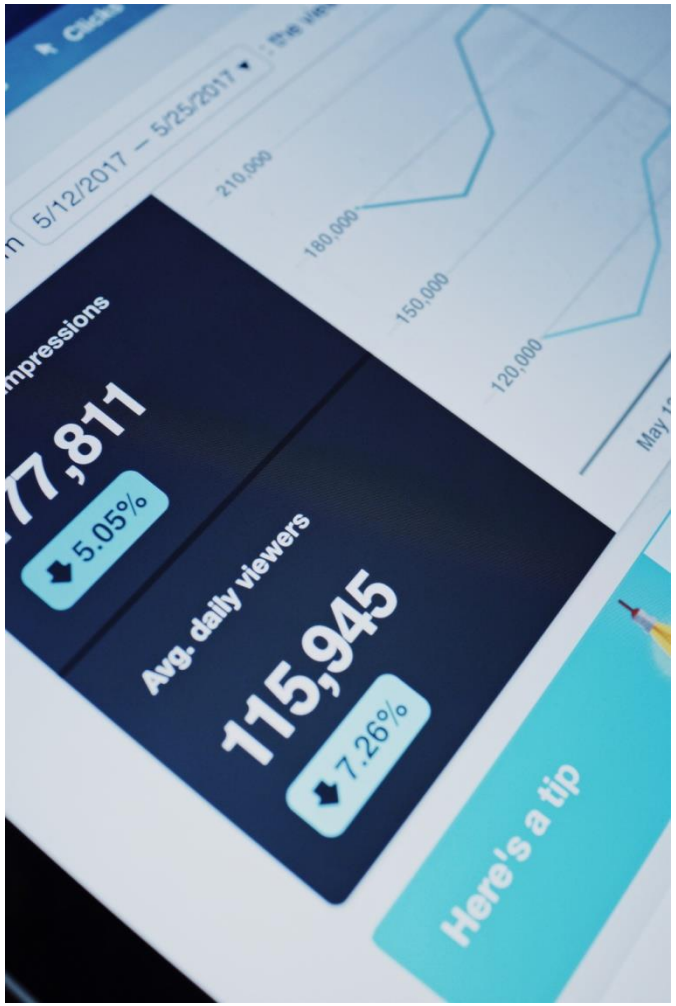


写好描述性文字

写好描述性文字

描述性文字的两个层次

1. **客观陈述：**描述统计图所展示的现象。
   例如：直方图的分布形态、柱状图各个类别的频数等。
2. **合理推断：**解读统计图背后的原因，推测数据为何呈现某种规律。



写好描述性文字

示范一：对二手房单位面积房价进行描述

本案例所关心的因变量是单位面积房价（单位：万元 / 平方米）。从直方图中可以看出，单位面积房价是呈现右偏分布的，如图 2-10 所示。具体来说，单位面积房价的均值为 6.12 万元 / 平方米、中位数为 5.74 万元 / 平方米。这一现象符合人们对于房价的基本认知，即存在少数天价房，从而拉高了房价的平均水平。在本案例中，单位面积房价的最小值为 1.83 万元 / 平方米，所对应的房屋是某地的一套两居室，总面积 100.83 平方米；最大值为 14.99 万元 / 平方米，所对应的房屋是某地的一套三室一厅，总面积 77.40 平方米。

“整体”陈述，用语准确

“细节”补充，引起读者兴趣

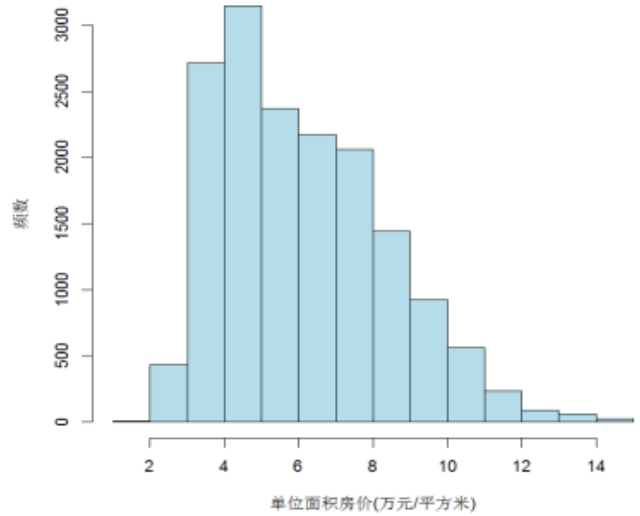


图 2-10　二手房单位面积房价直方图

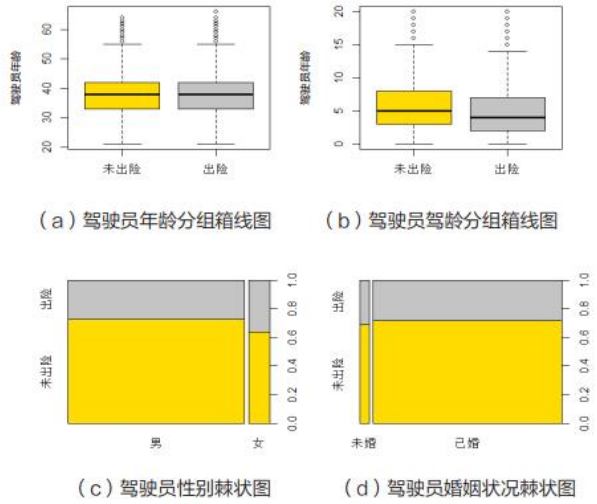
写好描述性文字

示范二：对驾驶员因素进行描述

本案例有一个明确的因变量：驾驶员是否出险。

所有的描述分析都聚焦在对比出险和未出险驾驶员的特征。

驾驶员因素共包含 4 个变量：驾驶员年龄、驾驶员驾龄、驾驶员性别和驾驶员婚姻状况。



（a）驾驶员年龄分组箱线图　（b）驾驶员驾龄分组箱线图

（c）驾驶员性别棘状图　（d）驾驶员婚姻状况棘状图

图 2-11　驾驶员因素描述统计图汇总

通过图 2-11，能够得到以下结论。

（1）驾驶员年龄：从图 2-11（a）所示的箱线图可以看出，出险和未出险驾驶员年龄的平均水平（中位数）和波动水平的差异并不明显。

（2）驾驶员驾龄：从图 2-11（b）所示的箱线图可以看出，出险驾驶员驾龄的平均水平（中位数）要明显低于未出险驾驶员，说明新手驾驶员更有可能出险。

（3）驾驶员性别和婚姻状况：从图 2-11（c）和图 2-11（d）所示的棘状图可以看出，女性驾驶员的出险率更高，但样本量远小于男性驾驶员；未婚驾驶员出险率略高，但样本量远小于已婚驾驶员。

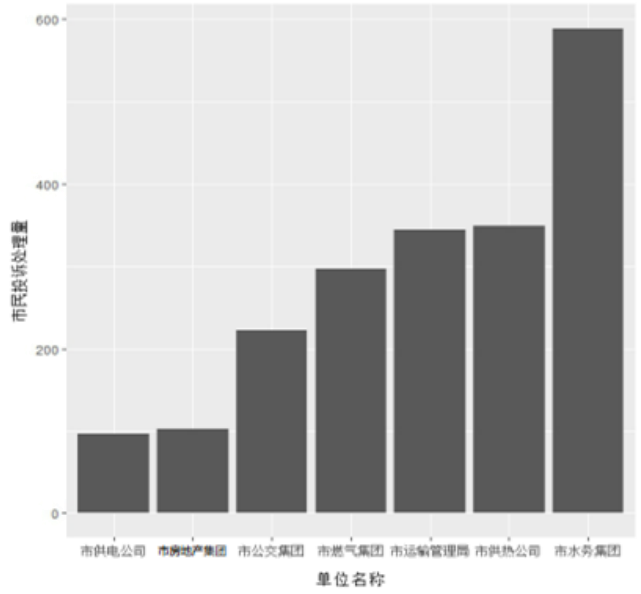
初步的结论是驾驶员的性别和婚姻状况可能对出险行为有影响，这种影响也可能是由于数据本身的样本量差异形成的。

注意用语。描述分析部分，不要使用“显著”！

写好描述性文字

示范三：各政府部门的市民投诉量

老王作为便民服务电话后台中心的负责人，想方设法提高建议和投诉信息的分类效率。老王想，最近流行数据分析，那么，数据分析的方法能不能解决自己的问题呢？首先他从刚刚过去的 12 月份的处理记录中提取了 2000 条被正确分类的建议、投诉信息，包含市民建议或投诉的文本记录，以及最终受理的政府部门。他想要看看投诉主要集中在哪些部门，于是对记录中各政府部门的受理数量进行了统计。当年 12 月，市水务集团被投诉的最多，而市供电公司与市房地产集团收到的投诉最少，如图 2-12 所示。老王猜测，这可能是因为该城市为北方的某省会城市。这时候城市的气温极低，水管容易破裂，造成街道、楼梯与住房等地方结冰，影响人们的正常生活，故投诉较多。



情景代入式写法，强调业务问题

合理推断

客观陈述

THANK YOU