统计描述性分析

引言

- 常言道，一图胜千言。统计图更是如此。
- 我们将讲述如何画基本的统计图形，即对数据的描述分析。
- 描述分析占据着整个数据分析过程中的重要地位——建模前，它是观察数据、发现问题、识别异常与规律的有力武器；建模后，它是总结规律、表现结论、为更多人传递信息的生动方式。

- 因此了解学习基本的作图方法，无疑会对我们的数据分析大有裨益。
- 最常用的图表类型，也就是柱状图、箱线图、散点图、折线图、直方图和饼图。它们是针对不同变量类型，不同变量个数展现时可能用到的工具。

- 我们采用某二手房数据作为示范

变量类型		变量名	详细说明
因变量		单位面积房价	单位：万元/平方米
自变量	内部因素	房屋面积	单位：平方米
		卧室数	单位：个
		厅数	单位：个
		所属楼层	定性变量 共3个水平
		所属城区	定性变量 共6个水平
	区位因素	是否邻近地铁	定性变量 共2个水平
		是否学区房	定性变量 共2个水平

一、为一个变量画图

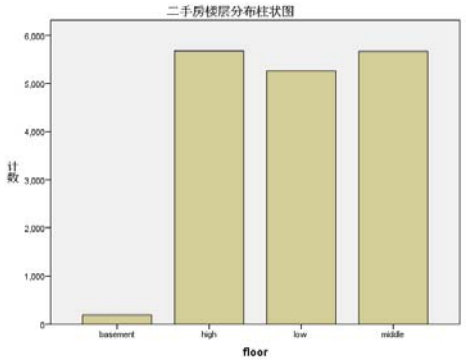
1.一个定性变量

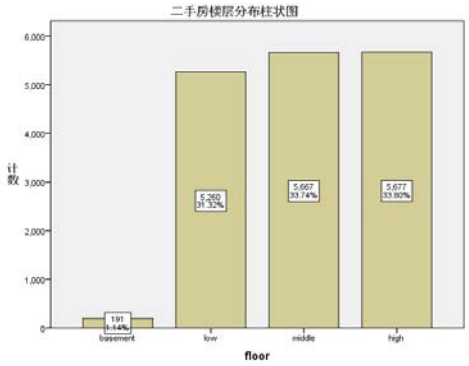
所谓定性变量，就是性别、国籍这类描述一个事物的特性的变量，他们的取值只能是离散的，比如男女，比如中国、英国等，描述这类型变量的图形有两种：柱状图和饼图。

- **柱状图（条形图）**

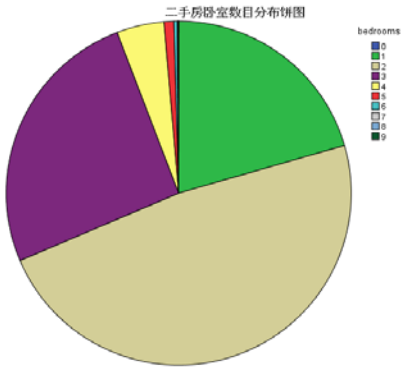
---柱状图适合展示一个定性变量的频数分布，也可用来观察不同类别样本的分布。

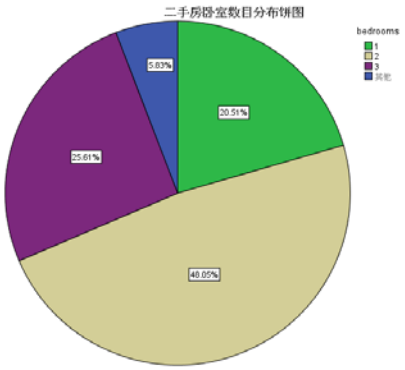
---比如在二手房数据中，楼层就是一个有不同取值的定性变量，如果我们想看看不同楼层的频数分布，便可以画一个柱状图。





- 饼图

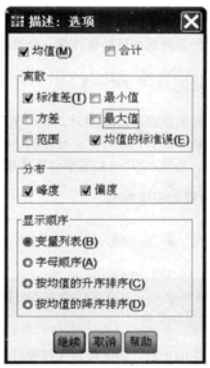




2.一个定量变量

所谓定量变量，就是可以取连续数值的变量，比如年龄、收入等。

- 描述性分析



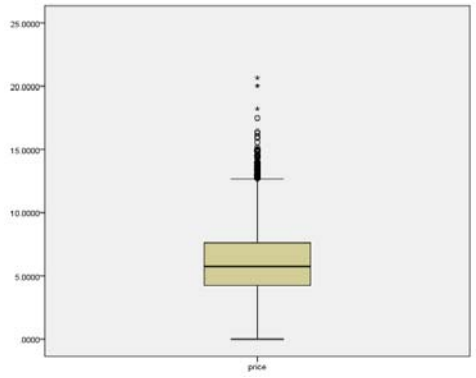
描述统计

	个案数	最小值	最大值	平均值	标准差	偏度		峰度	
	统计	统计	统计	统计	统计	统计	标准误差	统计	标准误差
price	16795	.0000	20.6667	6.083526	2.3441549	.578	.019	.465	.038
有效个案数（成列）	16795								

描述统计

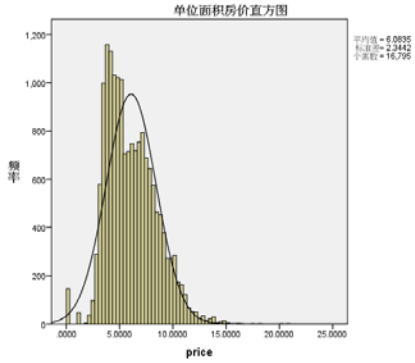
	个案数	最小值	最大值	平均值	标准差
AREA	16795	7.80	18779.50	95.8066	157.56002
有效个案数（成列）	16795				

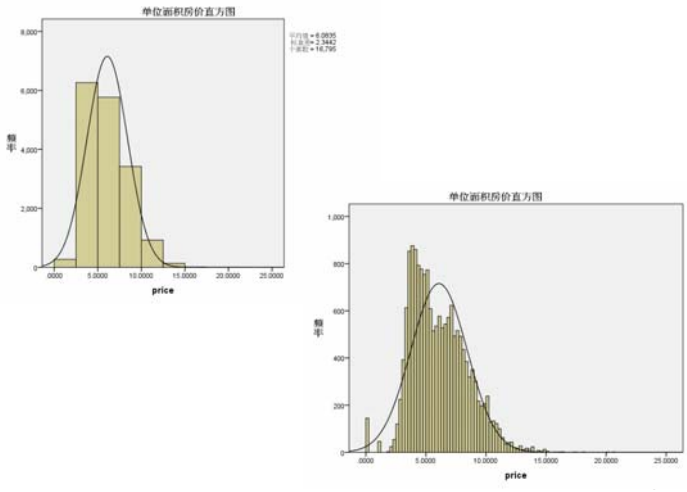
·箱线图



·直方图

---对于横截面数据来说，最重要的可能就是它的分布。观察分布最常用的工具就是：直方图





---不同的分组可以观察到不同的细节特征:

左图较粗糙，我们大致可以看到单位面积房价在5万左右的房子最多，因此在左图中产生了一个小高峰；

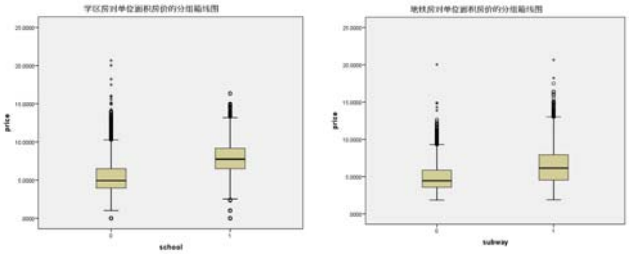
右图则展示了更多的细节，除了我们在5万以下的地方（大概3.8万左右）出现了小高峰，还可以看出在5之后出现了另外一个房价的小高峰区，大约在7万每平米左右.

---组数越多，我们可以观察到的细节就越丰富。

二、如果你要为两个变量画图

1. 定性变量及定量变量

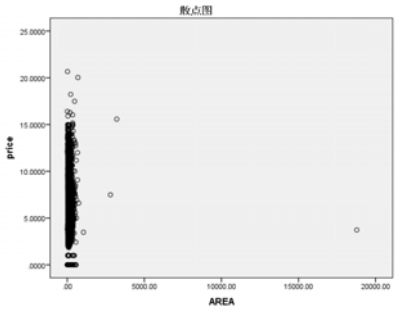
- 探索定性与定量变量之间的关系是数据分析中很常见的需求，比如我们想比较不同教育水平的收入差异，想比较不同地段的房价差异，想比较在电视剧里是好人活的集数多还是坏人活的集数多，这些就是在某个分类变量的标准下，比较另一个定量变量的表现。
- 分组箱线图，一种好用直观的工具，能够让我们一目了然，看清楚两组的对比及各自的数据分布。



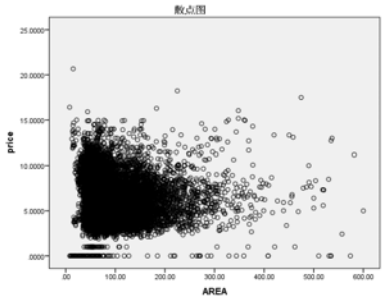
可以很清晰的看到二手房中的地铁房、学区房，单位面积房价也比其他房子高一截儿，尤其在学区房与非学区房之间这种差别就更加明显！

2.两个定量变量

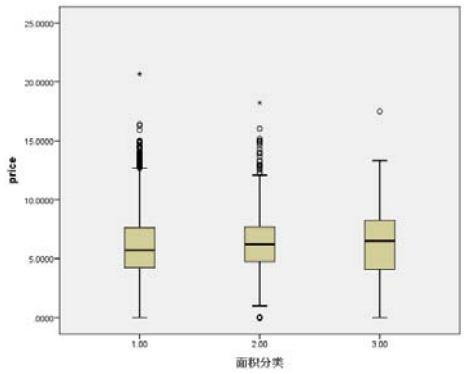
- 想探究两个定量变量之间的关系，最常用的就是散点图，它为我们观察这两变量的相关方向及相关程度提供了直观的阐释。



去除较大的异常值(AREA>600)后画图

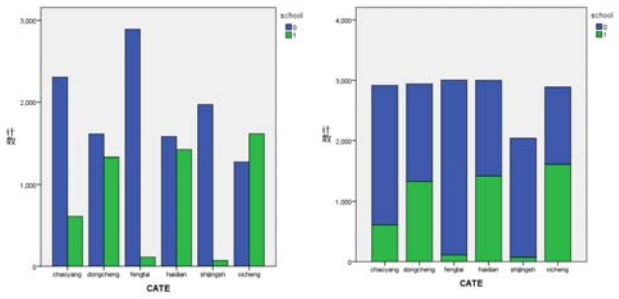


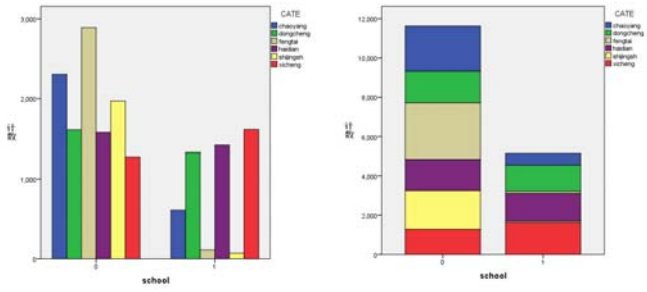
- 从这幅图中模模糊糊可以看出一点正相关的迹象，但相关程度并不高，遇到这种情况呢，我们可以考虑把某个连续变量离散化，也就是把它分组，变成定性变量，比如这里我们将房屋面积离散化，再跟房价做箱线图，可能这时候相关关系就慢慢显现出来了。



3. 两个定性变量

- 定性变量可以用柱状图来表示各个水平的取值大小，而两个定性变量可以采用柱状图的变形——堆积柱状图和并列柱状图。





小结

可能影响单位面积房价的因素：

- 区位因素：城区、地铁、学区
- 内部因素：卧室数、是否有客厅、面积、楼层