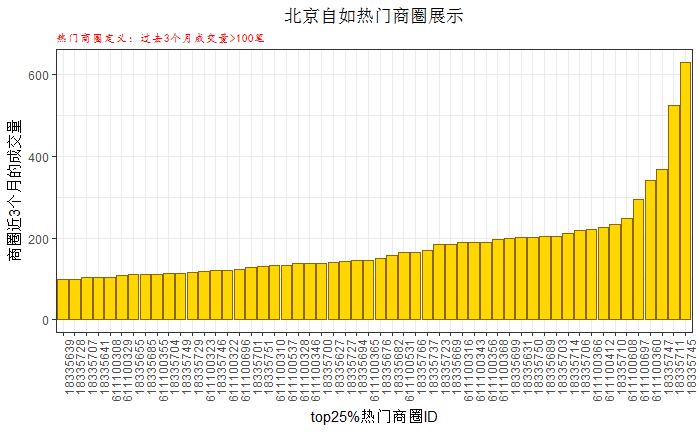
# 北京自如调研

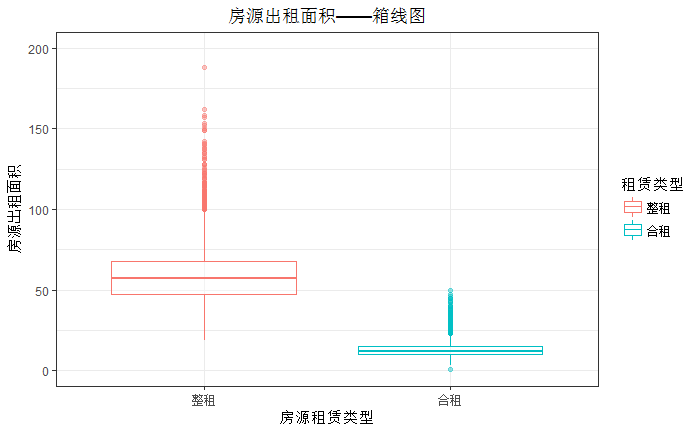
数据来源：商家=自如；城市=北京；时间范围：sign\_time:2018年；pt=20180315

## 热门商圈

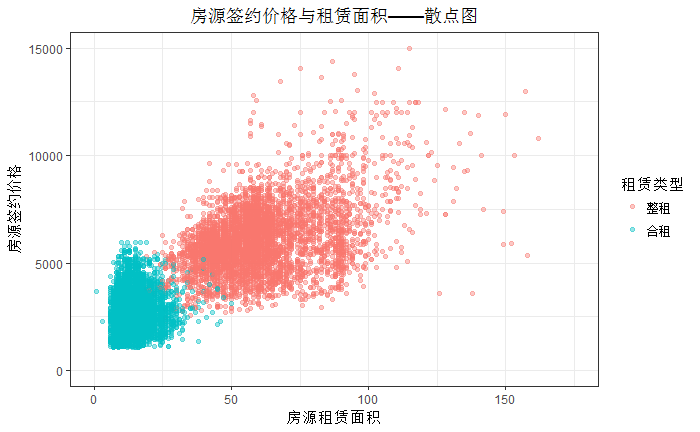


## 所有商圈统计数据

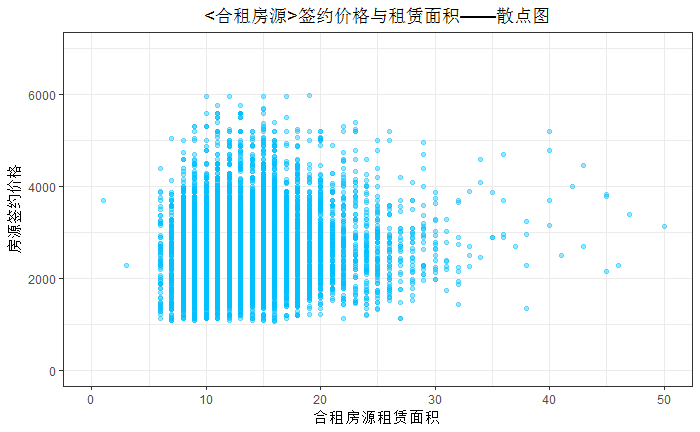
### 出租面积 箱线图



### 签约价格~出租面积 散点图



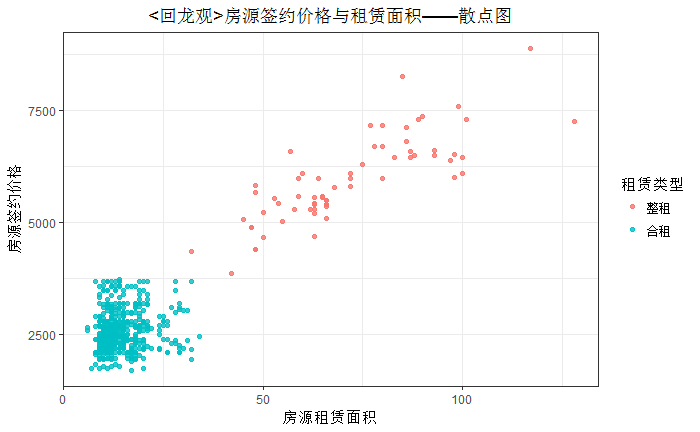
### 【合租放大】签约价格~出租面积 散点图



从北京整体来看：合租房源的成交价格和出租面积的相关系数：**0.1** 所以合租房源的成交价格不能用单位价格来衡量； 整租房源的成交价格和出租面积的相关系数：**0.44**。

## 回龙观商圈统计数据

选择了自如在2018年3个月交易量最大的商圈：**回龙观**（共629个交易记录）。



在**回龙观**这一个商圈内：合租房源的**成交价格**与**出租面积**的相关系数：**0.09。**

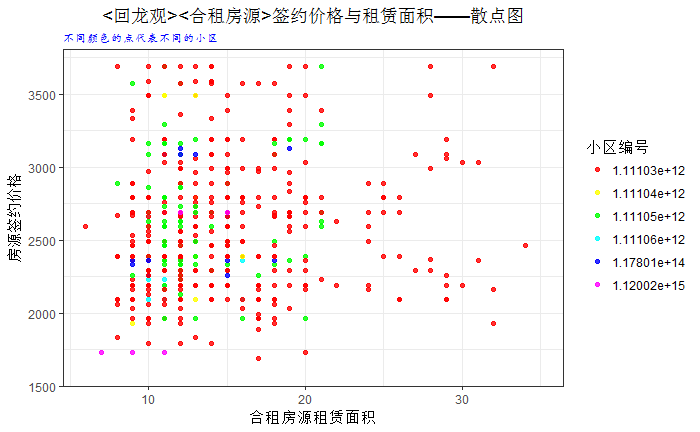
整租房源的**成交价格**和**出租面积**的相关系数：**0.81。**

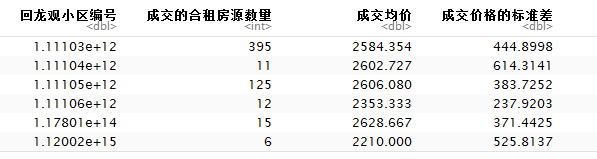
结论：同一个商圈内，自如**整租**的价格与**面积**有高度**线性相关**性。**合租价格与面积线性不相关。**

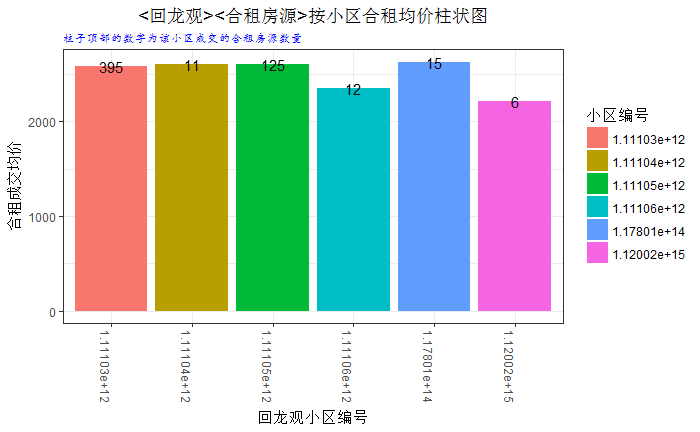
**既然合租的价格与面积不相关，那么影响合租价格的因素是什么呢？**

**考虑：1. 小区 2. 卧室数量**

### 回龙观+合租+小区







从直观上来看：回龙观商圈内部，各小区的合租均价没有显著差别。

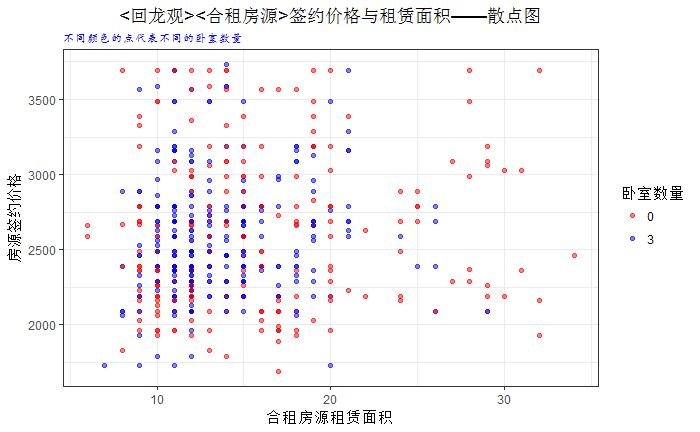
从假设检验的角度来看：

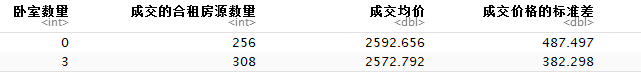
单因素方差分析的结果：P值=0.13，所以在0.05的显著性水平下，我们可以认为各个小区的合租均价无明显差别。

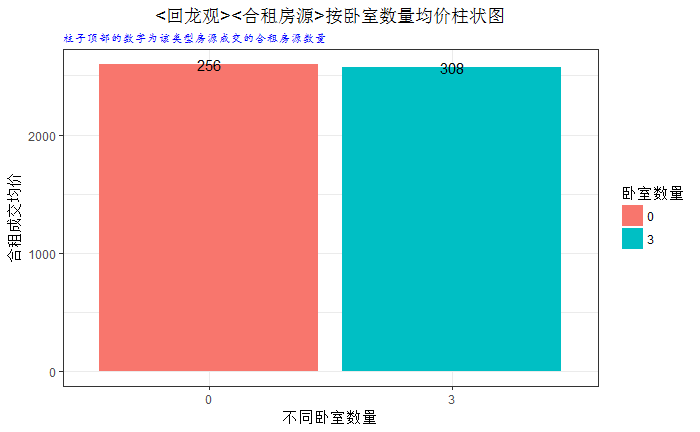


### 回龙观+合租+卧室数量

卧室数量只有两个水平：0间卧室 或者 3间卧室







从直观上来看：回龙观商圈内部，不同卧室数量房源的合租均价没有显著差别。

从假设检验的角度来看：

单因素方差分析的结果：P值=0.588，P值很大，我们可以肯定的认为回龙观的合租价格不受房源卧室数量的影响。



# 第三方品牌调研

## Adhoc提数

表： ods.ods\_rent\_plat\_rent\_house\_info\_da （基础信息表）

数据来源：商家=去除自如和链家普租；时间范围：sign\_time:2018年；pt=20180329000000项目所在目录：C:\Users\lianjia\Desktop\junjia\_model\_lianjia\other-diaoyan </br>

Sql的where语句;

WHERE house\_type not in('107500000001','107500000006','107500000007','107500000008','107500000016','107500000017')

and app\_source\_brand!='200301001000' and app\_source\_brand!='200302002000'

and app\_source!='200100000001'and pt=20180329000000

过滤条件：去掉爬虫数据源；其他房屋类型；自如和链家这两个品牌。

## 数据描述

Adhoc提取出来的数据的维数是：48437 33

保存在文件642709.csv中。

### 城市

关于房源所在城市的情况如下：

北京市 成都市 杭州市 南京市 上海市 深圳市 天津市 武汉市 重庆市 NA's

10603 26717 3982 177 1395 82 132 5325 1 23

可研究的城市：北京和成都。

### 商圈

关于房源所在商圈的情况如下：

length(table(data0$hdic\_bizcircle\_id)) #统计这10个城市一共有599个商圈。

其中成都一共有135个商圈。北京有216个商圈。

要研究商圈，一定要在具体的城市下面再研究。

### 品牌

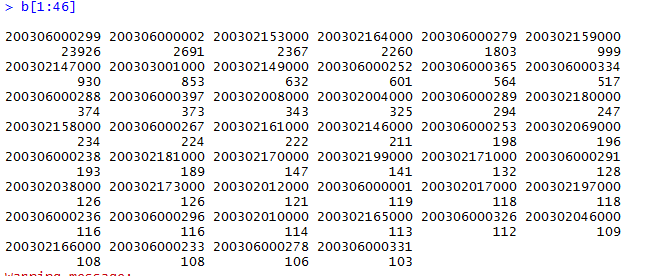
关于房源所在品牌（app\_source\_brand）情况如下：

在全部城市统计的品牌数目是：223个。

其中 挂牌房源数目大于100的品牌共有46家。

这46个品牌的频数统计图见下图：

可以发现：除了一家品牌编号为：200306000299的品牌的频数是2万多以外，其余的都是在2000到100之间。



### 挂牌价和成交价

关于房源的挂牌价和成交价：

rent\_price\_listing：本数据集的挂牌价格是全的

rent\_price\_trans ：本数据集的成交价是有缺失的：缺失的记录显示为0.

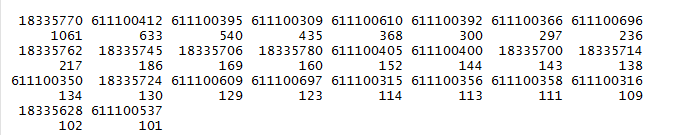
## 北京(房源挂牌数目：10603 )

### 热门商圈

全部品牌覆盖的北京商圈数目：216个商圈

定义北京的热门商圈标准： 在本数据集中挂牌量>100单的商圈。

依此标准得到的热门商圈数量：26个（占总数216的12%）。商圈id及对应的挂牌房源数见下图：

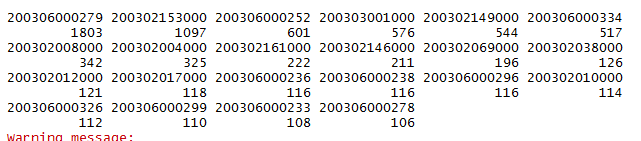


### 热门品牌

布局北京的品牌数量是：147家（本数据集全部品牌数：223家）

定义北京的热门品牌标准： 在本数据集中挂牌量>100单的品牌。

依此标准得到的北京热门品牌数量：22个（占总数147的15%）。品牌id及对应的挂牌房源数见下图：



### 【热门品牌\*热门商圈】列联表

样本范围：我们的样本选取的是：热门商圈 and 热门品牌，共有**5108**条这样的记录。（北京地区总记录数：**10603**），说明我们划分出热门商圈和热门品牌的概念很有效。（12%的商圈\*15%的品牌：占了总体50%的数据）。再在其基础上保留**合租类型**的房源数据： 共**3270**条。

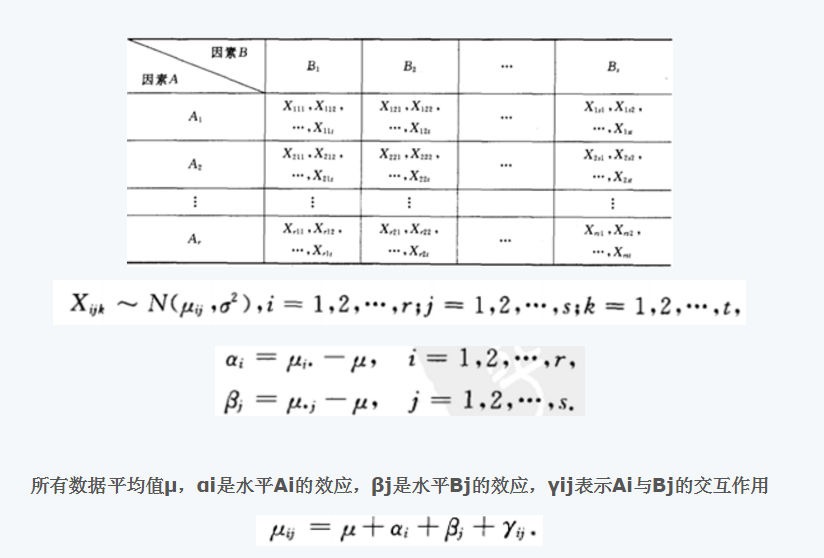
下面就要进行品牌这个因素和商圈这个因素的 双因素方差分析。研究的因变量是：合租房的挂牌价。

我们希望通过双因素方差分析来回答：在北京，品牌因素和商圈因素能否显著影响合租房源的挂牌价格？这个问题。

品牌因素：有19个水平 （为什么不是22品牌水平了？因为删掉了一些非合租的数据，导致有些品牌被删掉了）

商圈因素：有26个水平

双因素方差分析的思路是什么？如下图所示：我们有A（品牌）和B（商圈） 两个影响因变量（合租价格）因素。



双因素方差分析有个前提假设：

假设每个组（指的是每个【品牌\*商圈】组，所以我们case中：一共是26行\*19列 = 494个组），组内的样本来自同一个正态总体；并且这494个组的正态分布同方差。

所以做方差检验的第一步：得到这个26\*19的列联表。见表1，表1的19列代表19个热门品牌；表1的26行代表26个热门商圈；表1中单元格内的数字是组内频数。由于表1存在大量的NA，为了更清晰的展示表格内容，非NA数字都添加了红色色阶。表1中的所有频数的加和=3270.

表1：北京商圈（26个）\*品牌（19个）样本数量列联表

注：每个单元格中的数字代表该品牌在该商圈挂牌的合租房源数量。



从表中得到的结论：品牌并不是在所有的商圈都布局。品牌在北京各商圈的覆盖情况呈现出稀疏的商圈布局。并且根据热门商圈和热门品牌的概念可知： 表1在原始的最大的列联表（216\*147维）中，已经是红色覆盖最密集的区域了。

并且：从上表中找不出来全部填充为红色的子矩形。这说明我们的数据（商圈\*品牌矩阵）过于稀疏了。

如果是这样的情况，那么我们就无法在其基础上做有效的方差分析。

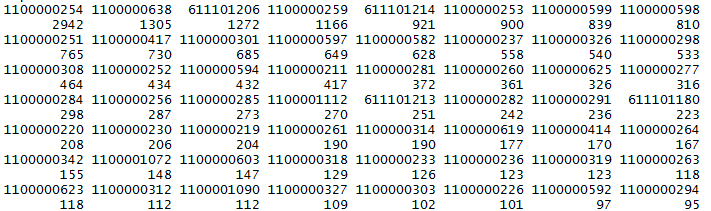
## 成都(房源挂牌数目：26717 )

### 热门商圈

全部品牌覆盖的成都商圈数目：135个商圈

定义成都的热门商圈标准： 在本数据集中挂牌量>100单的商圈。

依此标准得到的热门商圈数量：54个（占总数135的40%）。商圈id及对应的挂牌房源数见下图：

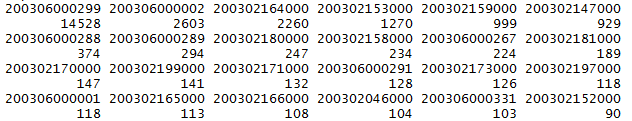


### 热门品牌

布局成都的品牌数量是：67家（本数据集全部品牌数：223家）

定义成都的热门品牌标准： 在本数据集中挂牌量>100单的品牌。

依此标准得到的成都热门品牌数量：23个（占总数67的34%）。品牌id及对应的挂牌房源数见下图：



### 【热门品牌\*热门商圈】列联表

样本范围：我们的样本选取的是：成都热门商圈 and 热门品牌，共有**22787**条这样的记录。（成都地区总记录数：**26717**），说明我们划分出热门商圈和热门品牌的概念很有效。（40%的商圈\*34%的品牌：占了总体85%的数据）。再在其基础上保留**合租类型**的房源数据： 共**20826条**。

表2的22列代表22个热门品牌；表2的54行代表54个热门商圈；表2中单元格内的数字是组内频数。由于表2存在大量的NA，为了更清晰的展示表格内容，非NA数字都添加了红色色阶。表2中的所有频数的加和=**20826**。

表2：成都热门商圈（54个）\*品牌（22个）样本数量列联表

注：每个单元格中的数字代表该品牌在该商圈挂牌的合租房源数量。



从上表中可以看出：成都的品牌中，有两个覆盖商圈非常广的品牌：

对北京22个热门品牌和成都的23个热门品牌取交集：只有"200302153000" "200306000299"这两个共同的热门品牌。说明不同的品牌战略侧重点在不同的城市。

## 调研：同一个品牌在同一个商圈内的合租价格是否同质。

思考：在商圈和品牌都相同的情况下，影响合租价格的因素有哪些？

以品牌号：'200306000299' 商圈号：'611101214' 为研究范围，来看一下影响它价格的因素是什么？

这个研究范围的数据集包含的字段有：（黄色字段为因变量）

[1] "hdic\_resblock\_id" （小区id） "floor\_num\_real" （物理楼层） "floor\_total" （总楼层）"house\_area" （房屋建筑面积）

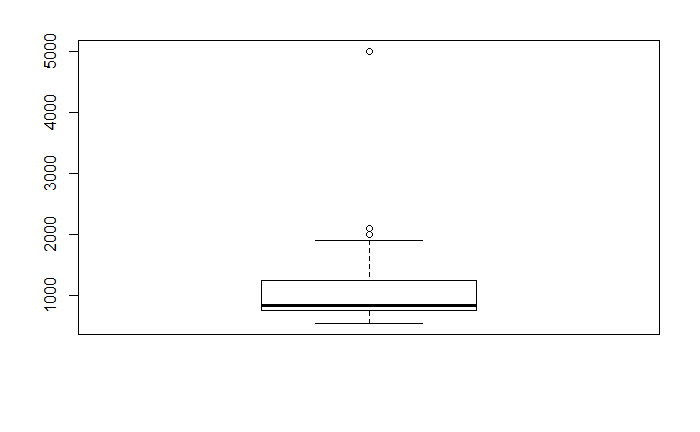
[5] "rent\_area"（出租面积） "frame\_hall\_num"（厅数量） "frame\_bedroom\_num" （卧室数量） "frame\_bathroom\_num"（卫生间数量）

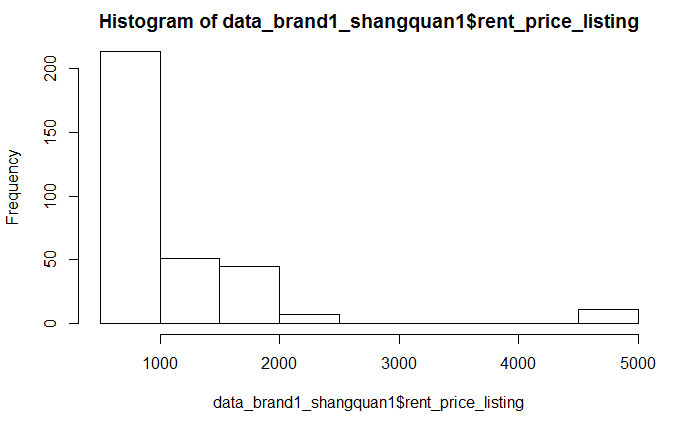
[9] "rent\_price\_listing" "rent\_price\_trans" "decoration\_level" "house\_status"

因变量rent\_price\_listing的分位数是;



分布形态展示：箱线图和频率分布直方图如下：



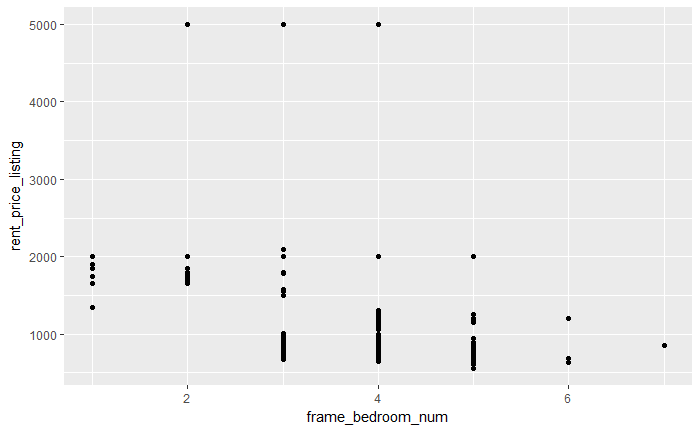


离群点：

认为合租租赁价格高于2000元的，都是价格上的离群点。

试过用其他自变量对挂牌价格进行回归建模, R2= 0.2 ;所以还是先做可视化吧。

挂牌价格 – 卧室数量 散点图



判断：卧室数量是不是影响挂牌价格的显著因素。