乐居二手房数据一览

一. 数据预览

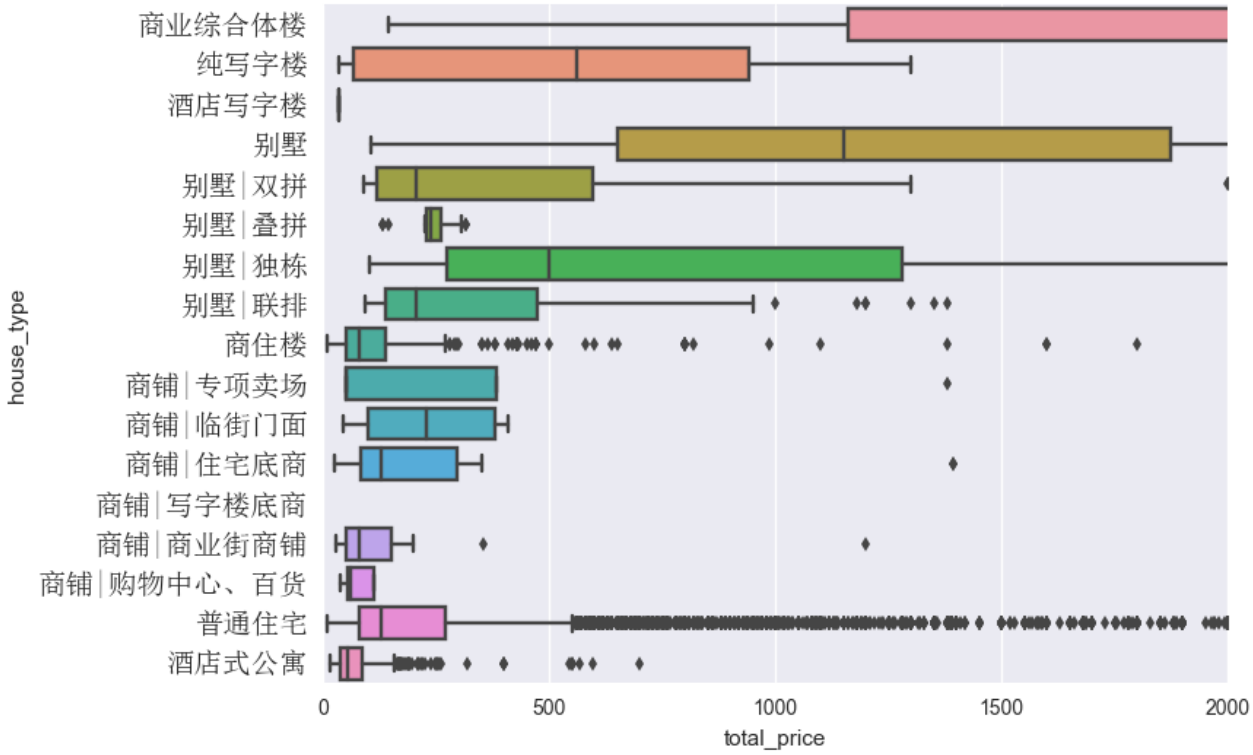
清洗过后的乐居网站数据

标题	总价	单价	房屋类型	产权	面积	装修	朝向	物业费	地址	小区
西樵山 广东4大名山之	300	12000	16	4	250	5	6	2.2	佛山市南海区	观山上岛
户型方正 超 高实用率	118	12688	16	4	93	6	7	3.1	广州黄埔科学	万科里享家
星河湾海怡半岛 高层南	1300	47445	16		274	2	7	4	番禺区洛西岛	星河湾海怡
四季花园 证够五年 地	230	30263	16		76	2	3	1.43		四季花园
哇塞，这么好的房子，	270	34177	16	4	79	4	6	2.8	白云广州大道	君华香柏广
少有省级学位 东川楼栋	550	53398	16	1	103	6	6	1.2	东川路	白云一街
增城中海联智汇城 lof	40	9302	16	4	43	5	7	3	增城区荔新大	中海联智汇

开发商	物业	绿化率	容积率	卧室	客厅	浴室	楼层	总楼层
佛山市南海西樵豪景有限	广东中奥物业管理有限	25	1	5	2	3	4	4
广州市金仑房地产开发有	广州万科物业有限公司	35	3	3	2	1	17	34
广州番禺海怡房地产开发	广州市星河湾物业管理	35	3.1	4	2	3	17	23
广州市四季花园房地产发	大都会物业管理公司	33.1	3.04	2	1	1	7	18
广州君华地产置业有限公	广州君华高力物业管理	30	3.4	2	2	1	15	30
		30	2	3	2	1	4	9
广州市锦鑫房地产开发有	戴德梁行	20	2.2	1	2	2	24	29
佛山市三水恒福兴达房地产开发有限公司		35	2.5	3	2	2	18	30

共有 17764 条不重复的有效数据

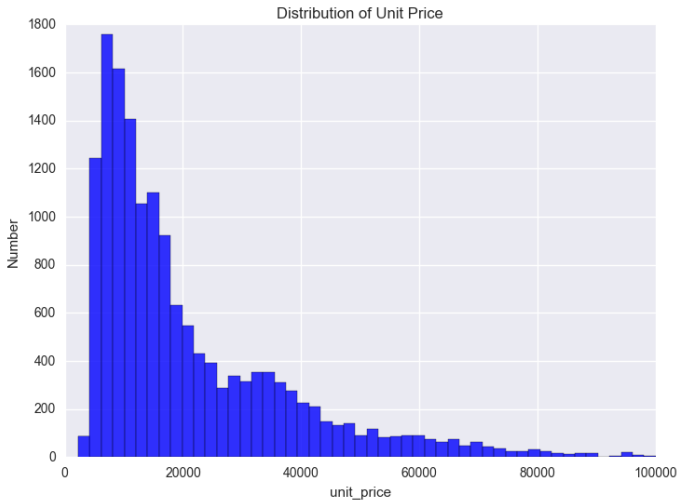
二手房类型分布

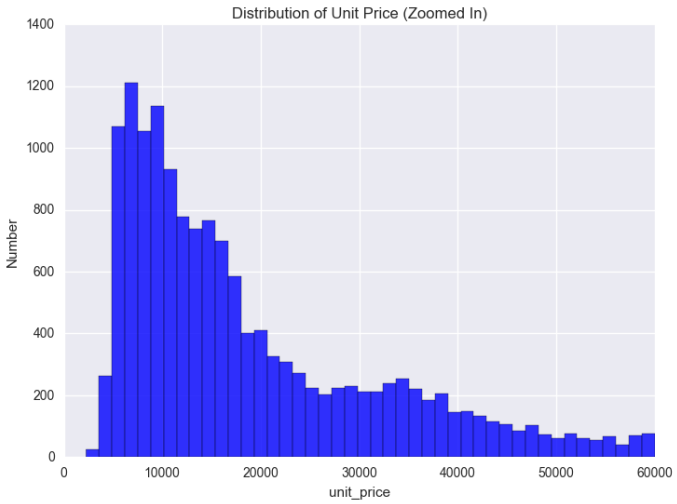


房屋类型很多，但其中主要是普通住宅有 15553 条数据，占所有数据的 87.5%，其次是酒店式公寓，占 6%。因此单挑普通住宅出来，略去其他干扰数据之后再进行进一步分析。

二．数据的主要分布特征

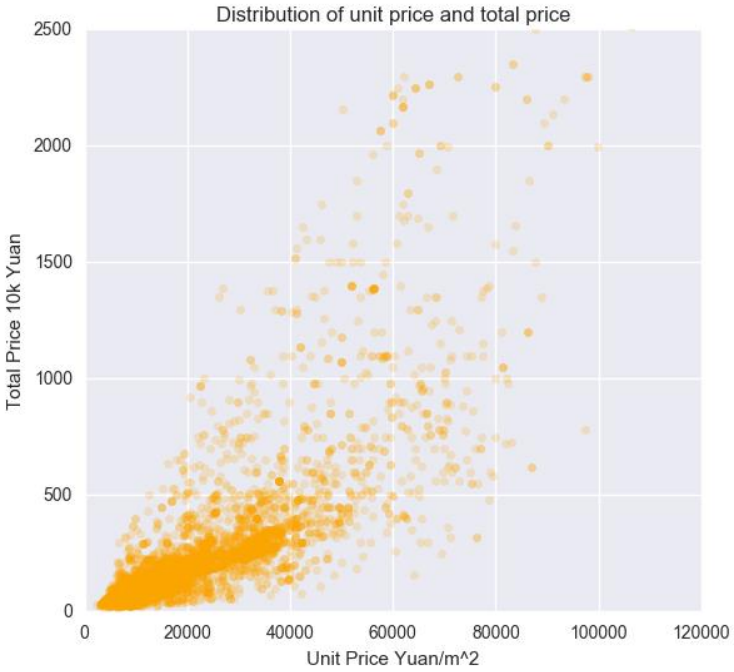
1. 房屋每平米单价的分布（左）及单价 6 万以内部分的放大图（右）

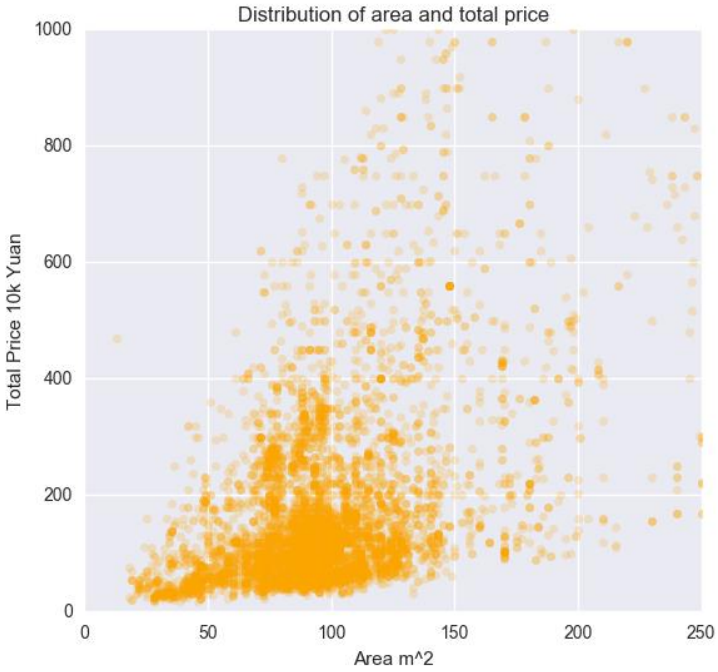




出售的普通住宅的每平米单价主要集中在四万元以内，其中在两万元以内的分布非常密集

2. 每平米单价与总价分布的散点图（左）及房屋面积与总价分布的散点图（右）

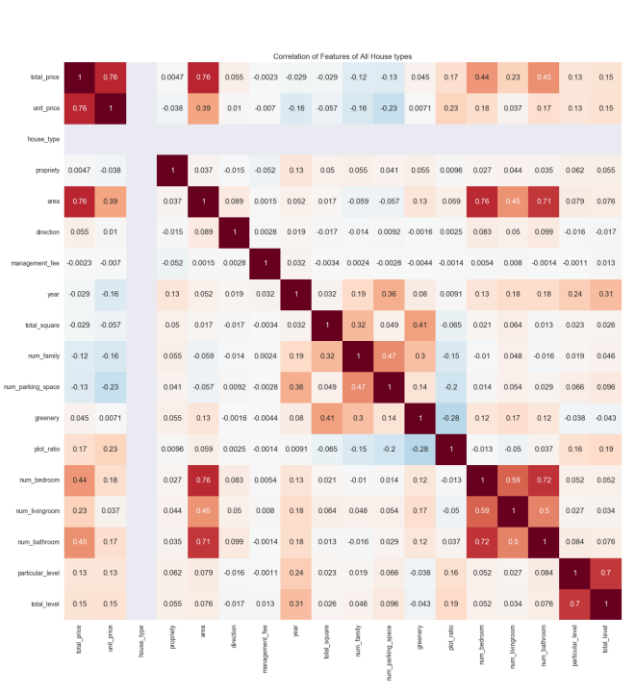


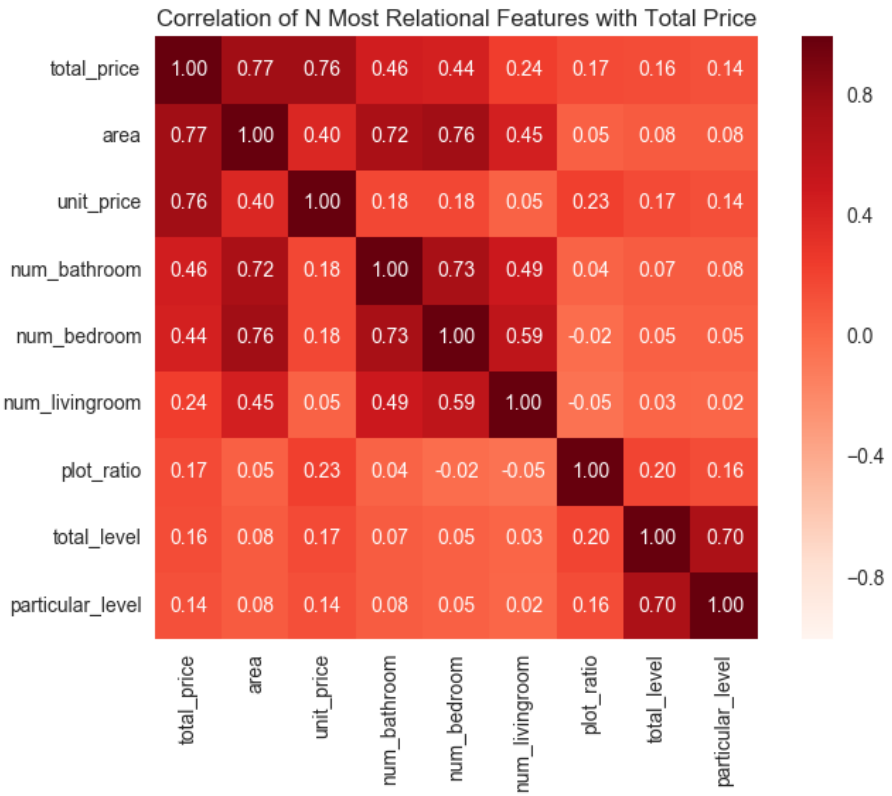


显然左图呈现比较符合常理的每平米单价越高，房屋总价可能越高，两者呈现很好的线性相关性；其中在数据密集区上方与线性分布数据偏移较远的数据说明可能有其他的因素影响房屋的总价。由于乐居的二手房数据包含众多的房屋类型（商铺、别墅、商品房等）因此很有可能是不同的房屋类型造成了小部分数据不满足房屋总价与单价近似线性相关的关系。

对右图取出数据密集分布区（房屋面积小于 250 平米，价格小于 1000 万，可以过滤留下绝大部分的商住房），发现房屋面积与总价同样也有较强的相关性（相关性可达 0.76，见下图）

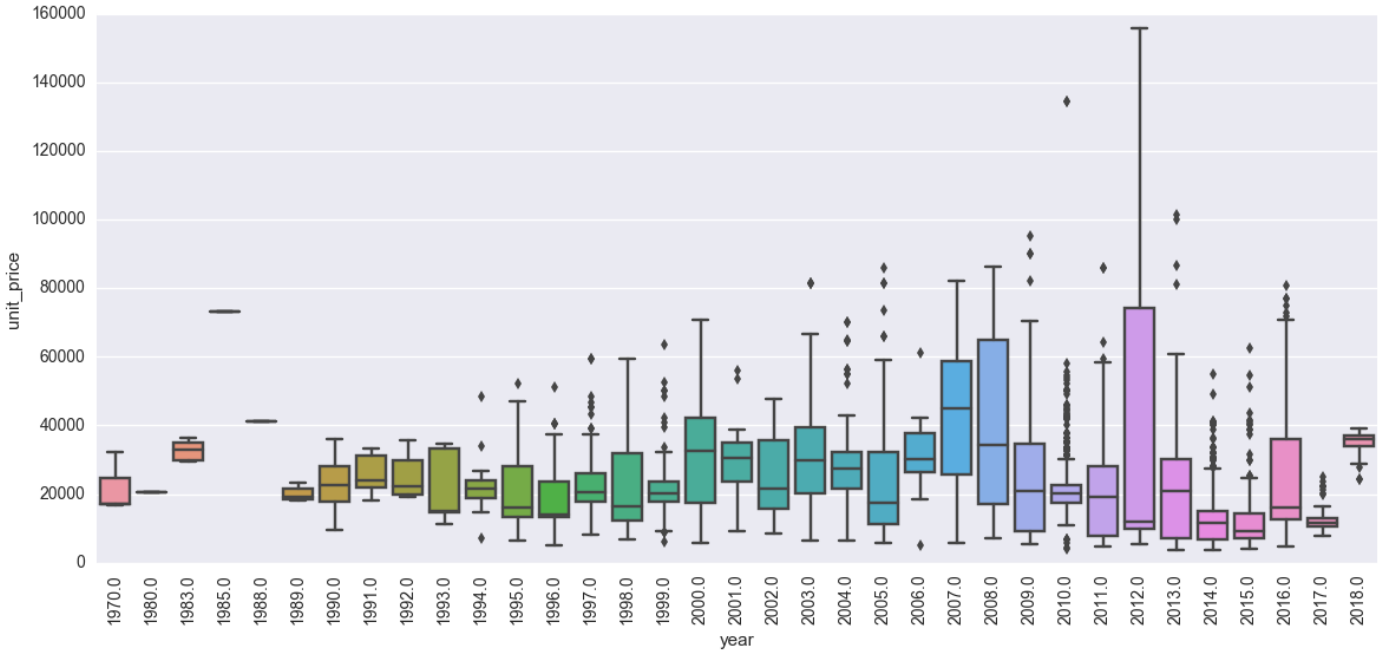
3. 总价与各变量之间的相关性的热点图（左）以及相关性最大的九个非分类变量组成的热点图（右）





房屋总价与总面积以及每平米单价的相关性最高，分别为 0.77 与 0.76；与房屋的卧室数量和卫浴的数量的相关性相近，可能由于两者之间本身有很强的相关性，可视为同一指标，有趣的是，客厅的数量和总价的关系与前两者相差较远，客厅不是影响房屋定价的主要因素。

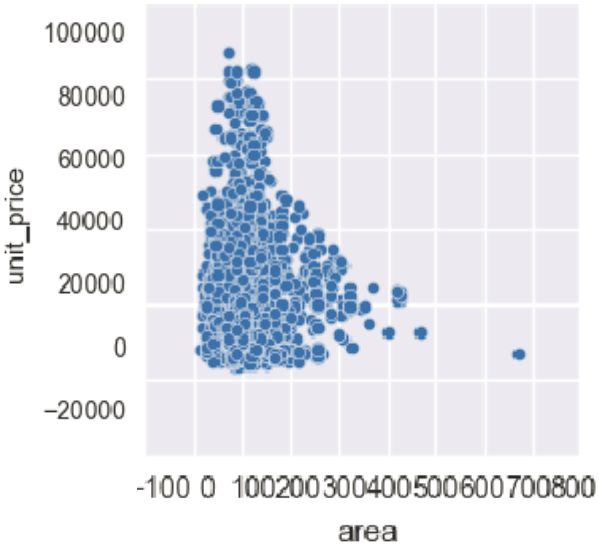
4. 住宅的单价与年份的关系



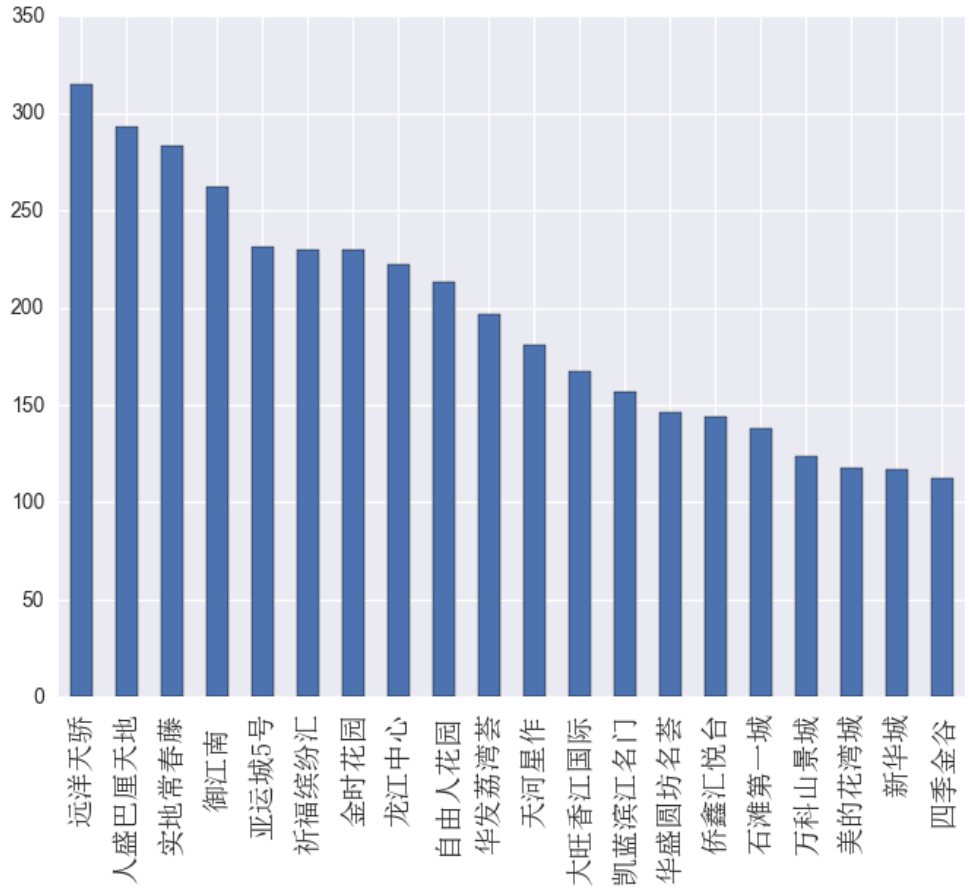
比较意外的，乐居上住宅的每平米单价并不随着房龄的增大而减小，而是体现出一个较弱的负相关（置信度不高），怀疑可能有其他因素影响房龄与单价之间的关系。

5. 过滤掉九个非分类变量中标定性相近的指标，画出它们的配对散点图

有意思的是单价与面积的散点图
房屋的面积越大，面积较小的房屋主要集中在最左侧
它们的分布呈直线话，价格波动区间很大，这些
房屋可能位于市区的中心区，价格受地段等影响
较大。而面积较大的房屋，每平米单价则更趋向于
分布在较低的水平，这些房屋可能不在市区最繁华
的地带



6. 在售的房子所在小区分布（数量的前二十）



7. 房源的地区分布