【项目10】 房价影响因素挖掘

作业要求

1、数据清洗、整合

要求：

① 将“house\_rent”、“house\_sell”分别读取

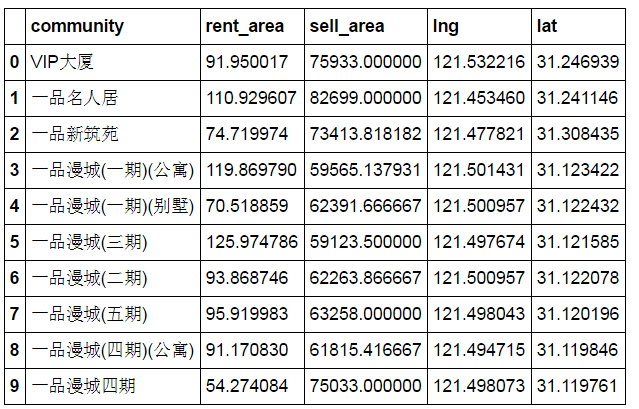
1. 分别计算平方米建筑面积的月租金、每平方米建筑面积的房价
2. 将数据按照小区名合并

提示：

① 删除缺失值

② 按照小区做均值分析

整理后数据大概11532条



2、计算“房屋售租比”，并做初步判断

要求：

① 计算指标

1. 绘制直方图、箱型图看“售租比”的一个数据分布情况

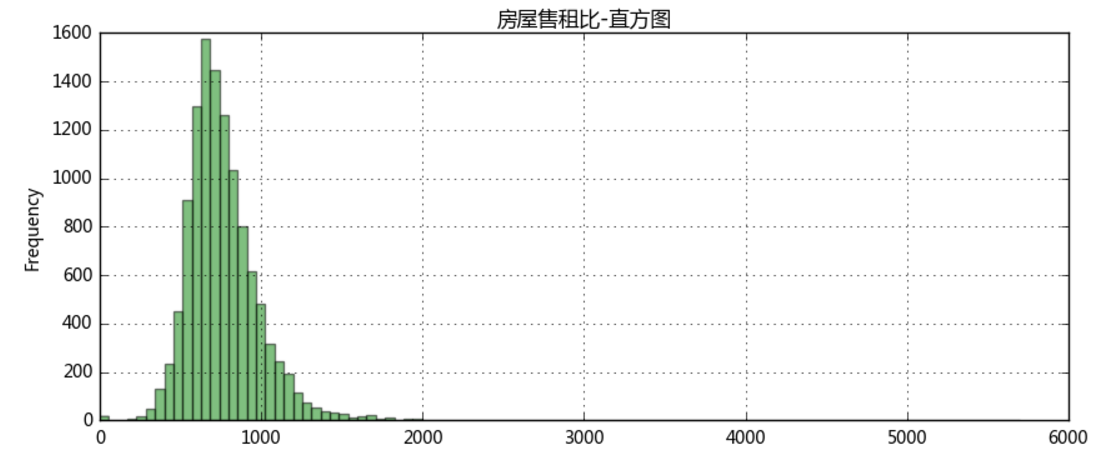
提示：

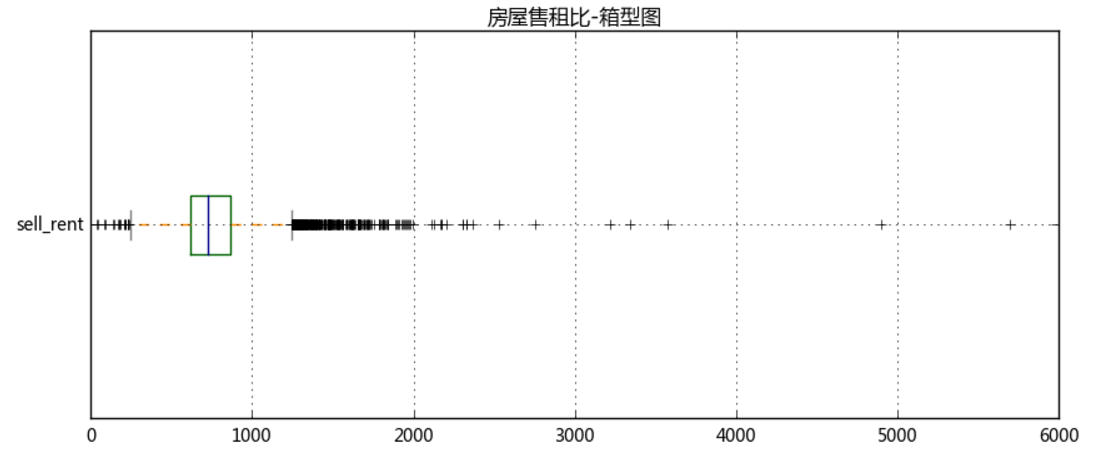
① “房屋售租比”=“每平方米建筑面积的房价”/“每平方米建筑面积的月租金”

② 直方图bins数量大于80来作图

\*\* 一个房子，100平，2万单价 → 2000000万，首付600000,贷款1400000

\*\* 租金 8800，20000/（8800/100）





一个结论 →

如果按照中位数725个月来看，假设从25岁研究生一毕业，就立刻全款买了一套房，然后放出去收租，那么约85岁时，这60年来陆陆续续所收的租金总数就可以达到了25岁时买房所付的钱啦！（这里还没考虑净现值折算问题）

① 如果仅靠租金收入的话，上海全市平均回收投资需要725个月。而这种格局的维持，必须有赖于购房者对上海的房价上升的持续预期。→ 也就是说，在上海，投资房产绝不是利率收益，而是预期收益（其实城市越小租金回报反而越高）

② 上海不是这么看住房投资的（绝大多数房子房租收入跑不赢商贷利率，否则按照这个思路看那肯定是买一个亏一个）

3、上海市人口密度、路网密度、餐饮价格和“房屋每平米均价”是否有关系呢？

要求：

① 首先，导出整理好的数据，并qgis中绘制空间格网图，查看房屋每平米均价、房屋每平米租金及售租比数据的空间分布

② 第二，空间统计，分别按照格网对人口密度、路网密度、餐饮价格进行指标统计并标准化

③ 第三，加载上海中心点point空间数据，计算每个网格到市中心距离

④ 第四，将空间格网的“房屋每平米均价”按照距市中心的距离排序，并制作散点图，看看能否挖掘出什么信息

\*\*\* 这里市中心点坐标为：lng-353508.848122，lat-3456140.926976 (投影坐标系)

提示：

① 导出csv数据，用dataframe.to\_csv()

② qgis加载数据后，以“net\_population”为格网数据做空间统计

③ 注意qgis数据都为投影坐标系

④ 人口密度指标 → 已有“net\_population”数据

路网密度指标 → 以格网为空间单元，计算道路长度

餐饮价格指标 → 以格网为空间单元，计算餐饮设施的人均均价数据

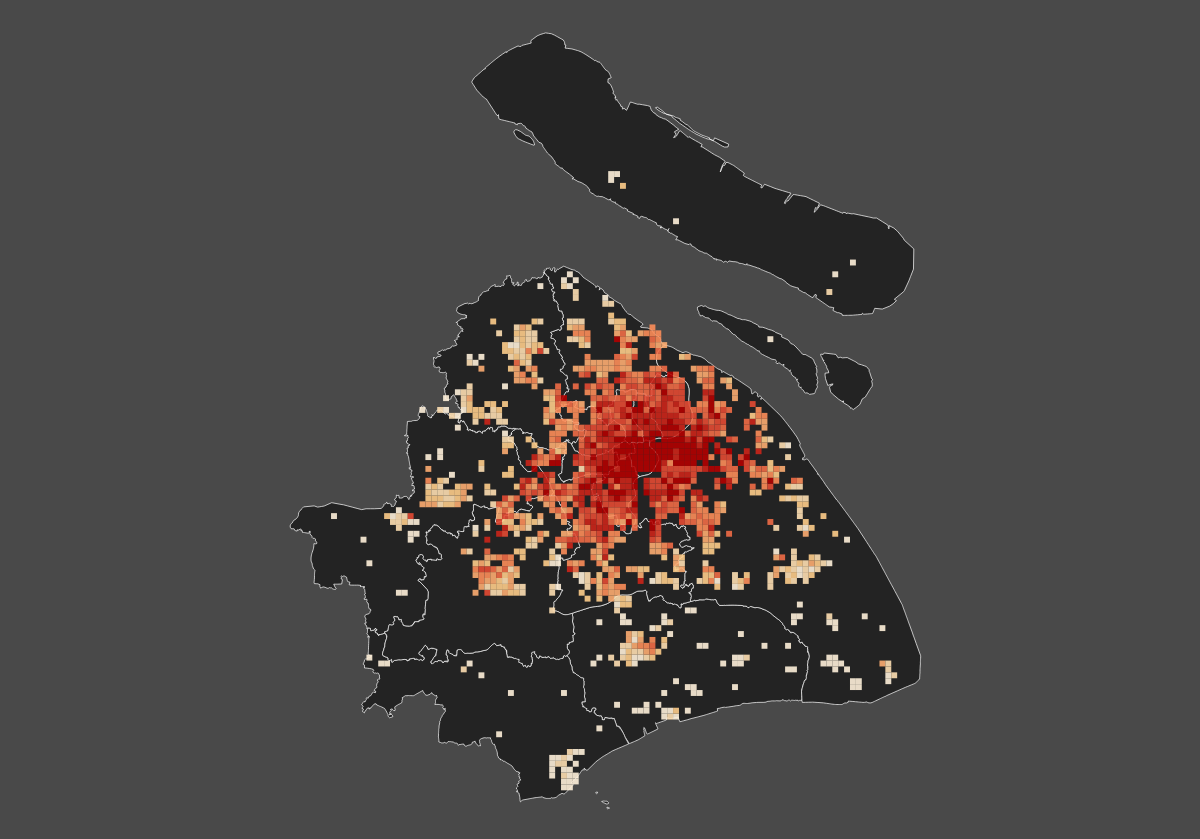
\*\*\* 最后数据导入python中，标准化得分至0-1区间

\*\*\* 导入数据后要填充空值为0

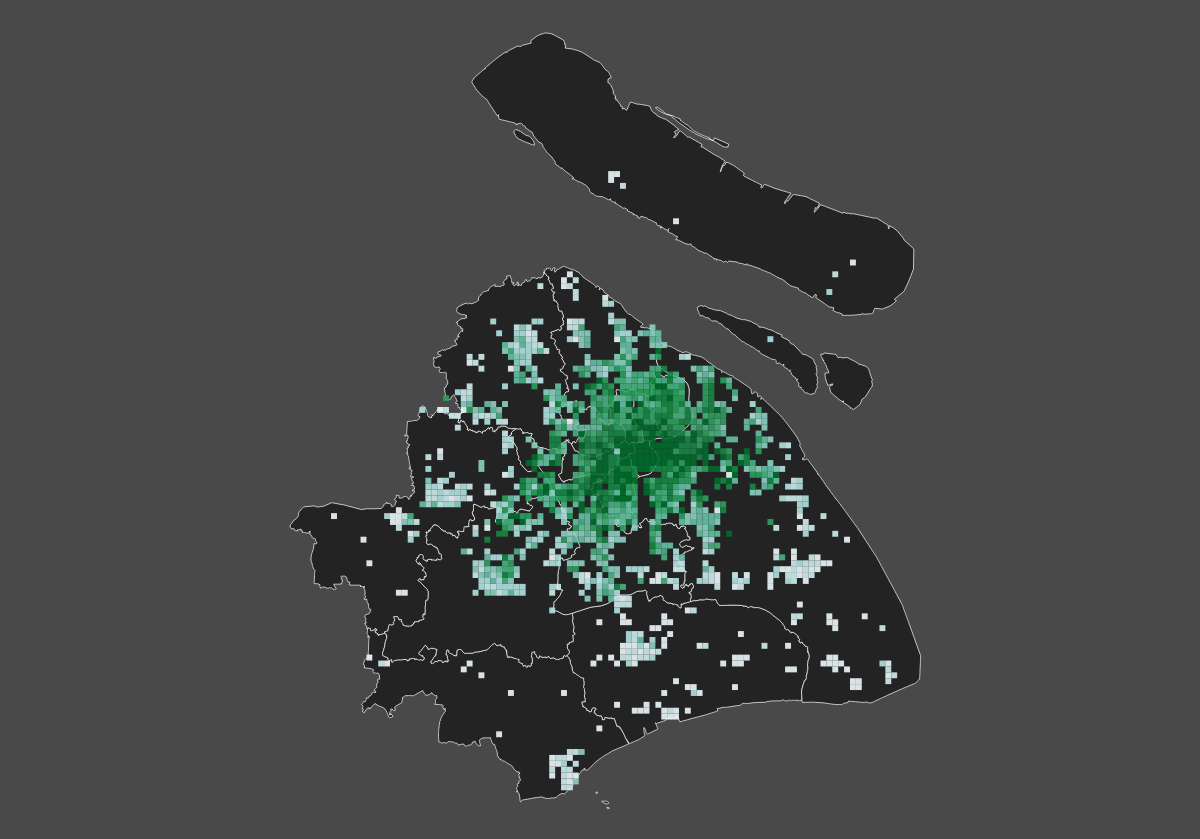
\*\*\* qgis中可以用结果net数据作为下一个分析数据，以此将统计结果汇总在一张属性表内

\*\*\* 格网数据在导出前，先转为点数据，并计算经纬度，这里用投影经纬度，好依据中心点坐标计算离市中心距离

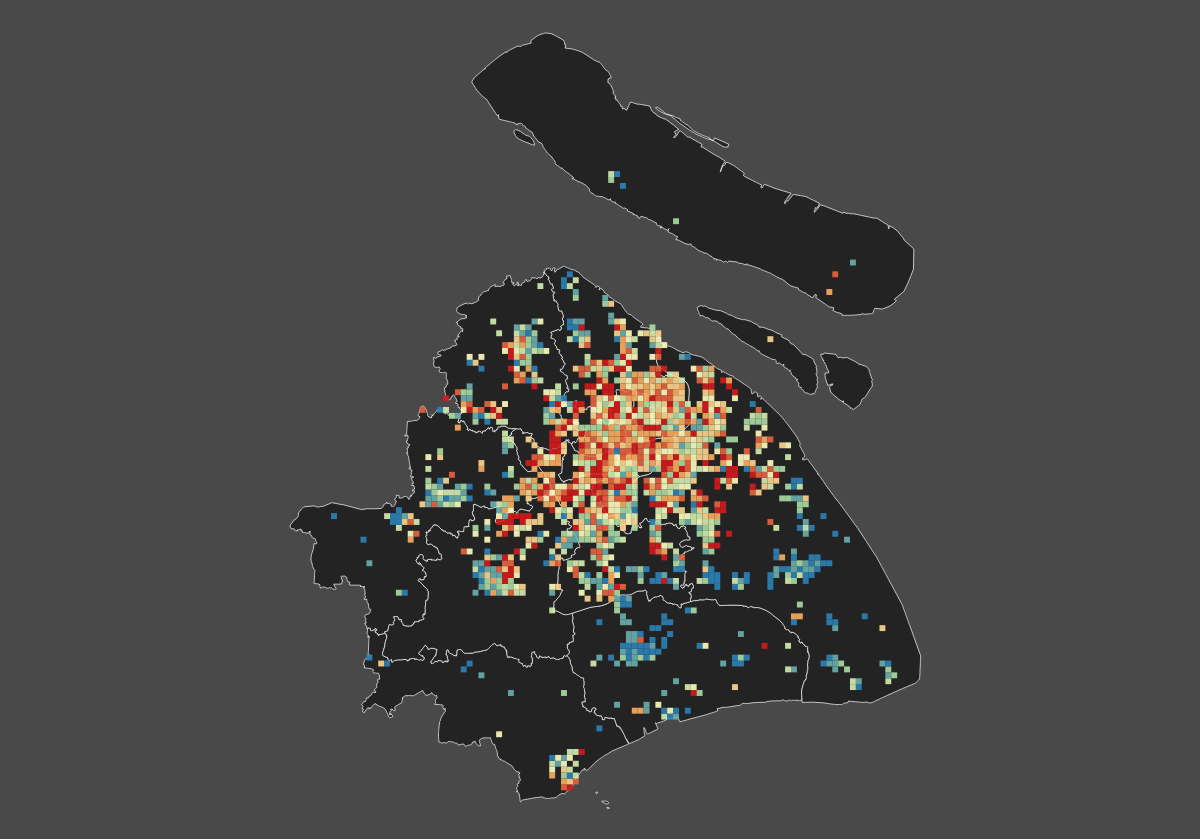
⑤ 清洗数据，去除“售租比”为0的数据



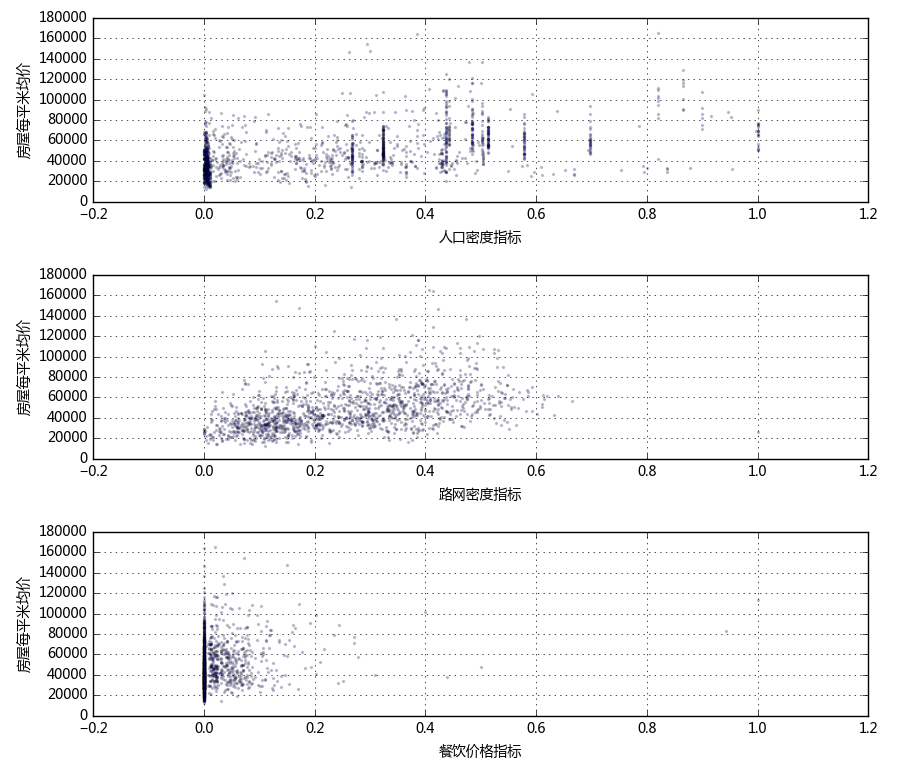
房价空间分布

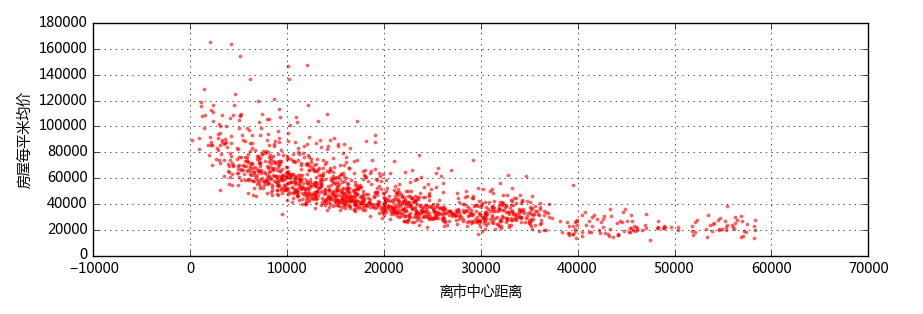


租金空间分布



售租比空间分布





结论

① 在上海全市层面，“离市中心距离”与“房屋每平米均价”相关性最强

② “人口密度”及“路网密度”和“房屋每平米均价”为中等相关

③ “餐饮价格”与“房屋每平米均价”为弱相关

④ “房屋每平米均价”数据的离散程度却和空间距离有关 → “房屋每平米均价”越靠近市中心越离散，越远离市中心则越收敛

4、按照离市中心距离每10km，分别再次判断人口密度、路网密度、餐饮价格和“房屋每平米均价”的相关程度

要求：

① 按照空间距离分别迭代计算三指标和“房屋每平米均价”的关系

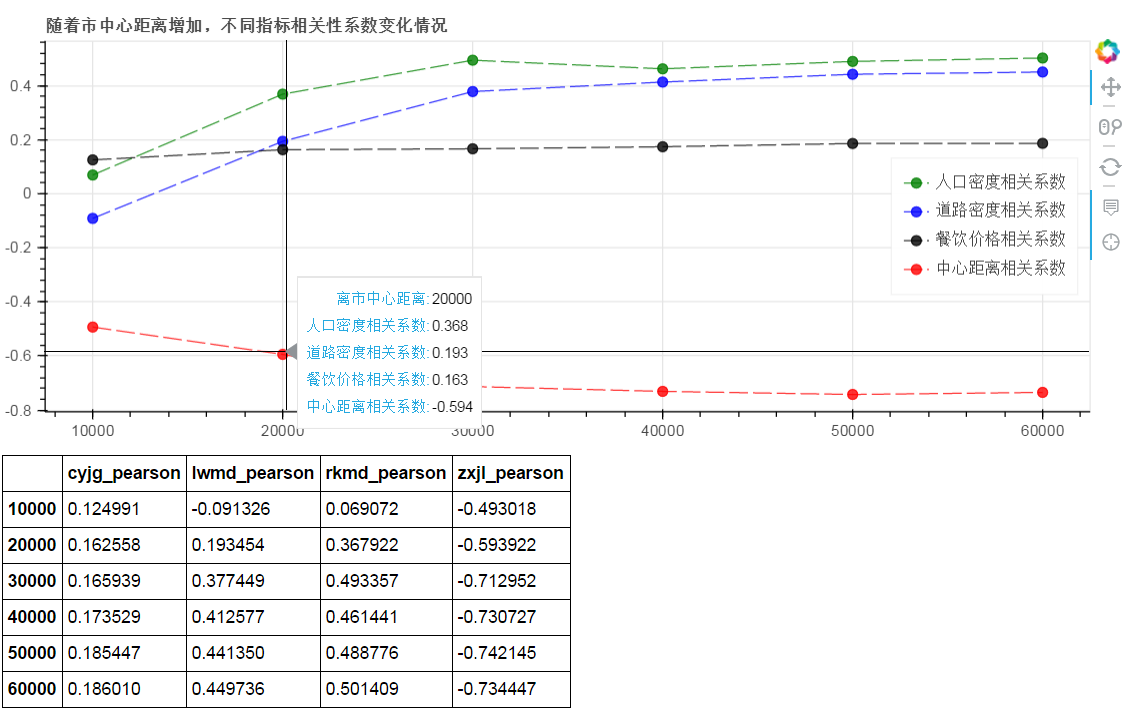
② 绘制折线图查看：随着市中心距离增加，不同指标相关系系数变化情况

\*\*\* 建议用bokeh制图

提示：

① 用for循环迭代空间距离，然后筛选数据并计算相关性

② bokeh可以通过多次调用figure.line()来绘制多条折线图



结论

① “人口密度”、“道路密度”、“离市中心距离”和“房屋均价”有着明显的相关性，而“餐饮价格”和“房屋均价”相关性较弱

② 随着离市中心的距离越远，指标的相关性在数据上体现更明显，而这个分界线大概在20-30km处，这正是上海中心城区和郊区的分界 → 上海房价市场的“中心城区-郊区”分化特征

③ 中心城区的房产市场对指标因素的影响更加敏锐，而郊区则更迟钝 → 越靠近市中心，影响因素越复杂