实验报告

赵若涵 2023200231

文件目录

code/hw4_crawler_scraping_data.ipynb ： 爬虫收集数据

code/hw4_EDA.ipynb：数据清洗与探索性数据分析

code/hw4_regression_model.ipynb：建模与绘图比较

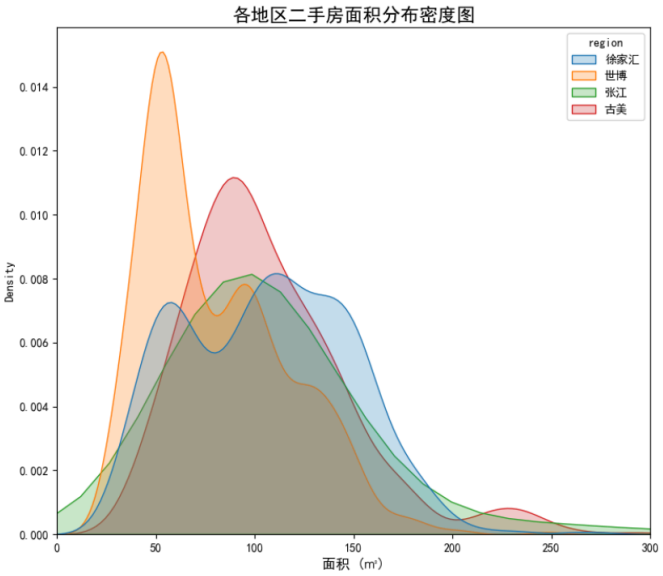
data/：由于不上传data，略，实际在本地应有四个地区的esf、zu数据，以及在数据清洗时保存的all、clean、merged version data

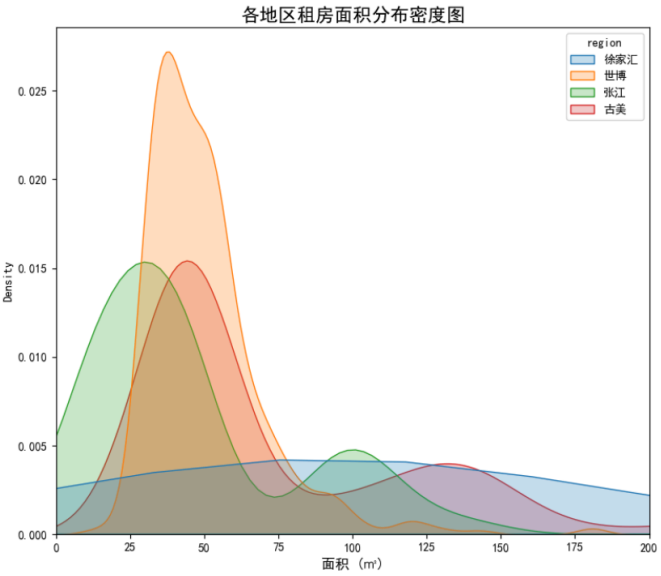
探索性数据分析

1. 数据描述性统计

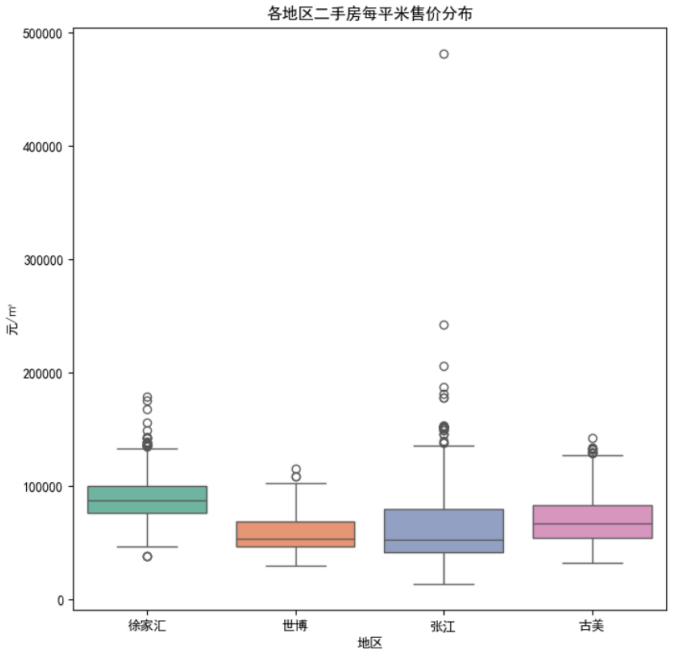
```
各地区二手房每平米售价 (price_per_sqm) 描述统计:
         count          mean           std      min       25%      50%  \
region
世博       1200.0  58987.040000  15549.880278  29811.0  46821.50  53734.0
古美        504.0  71896.569444  20911.861046  32357.0  54457.75  67531.5
张江       1200.0  64443.458333  33004.038670  14057.0  42033.25  52769.0
徐家汇      1120.0  89349.368750  18058.055795  38758.0  76746.25  87391.0

             75%       max
region
世博       69194.75  115606.0
古美       83636.00  143000.0
张江       79514.00  481602.0
徐家汇      99868.00  178663.0

各地区租房每平米月租金 (rent_per_sqm) 描述统计:
         count        mean        std        min         25%         50%  \
region
世博       1200.0  103.410549  23.524285  22.000000   90.000000  101.562500
古美        582.0   83.191572  24.801580  55.172414   67.391304   74.418605
张江       1200.0   80.536479  32.318853  27.500000   63.190341   73.043478
徐家汇      1197.0  125.089235  30.419353   1.000000  104.545455  121.212121

                75%         max
region
世博       116.127225  195.744681
古美        90.452261  166.666667
张江        90.000000  200.000000
徐家汇      140.350877  270.000000
```

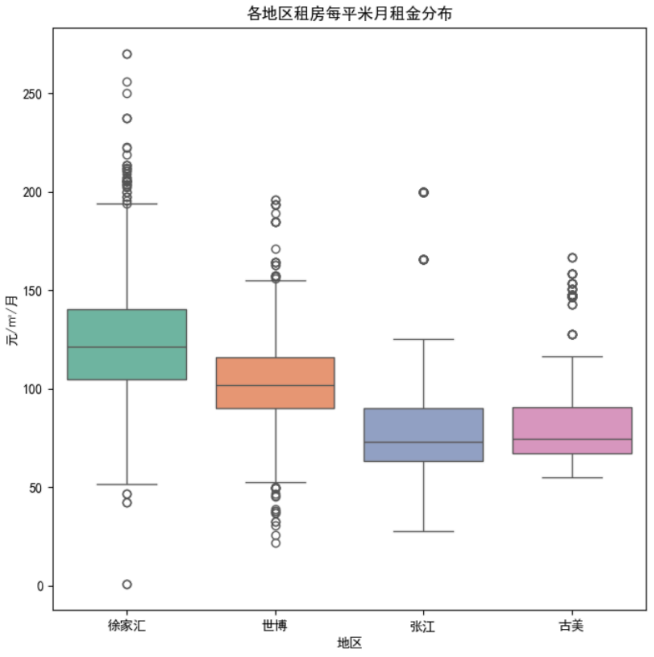
2. 各地区房屋免积分布密度



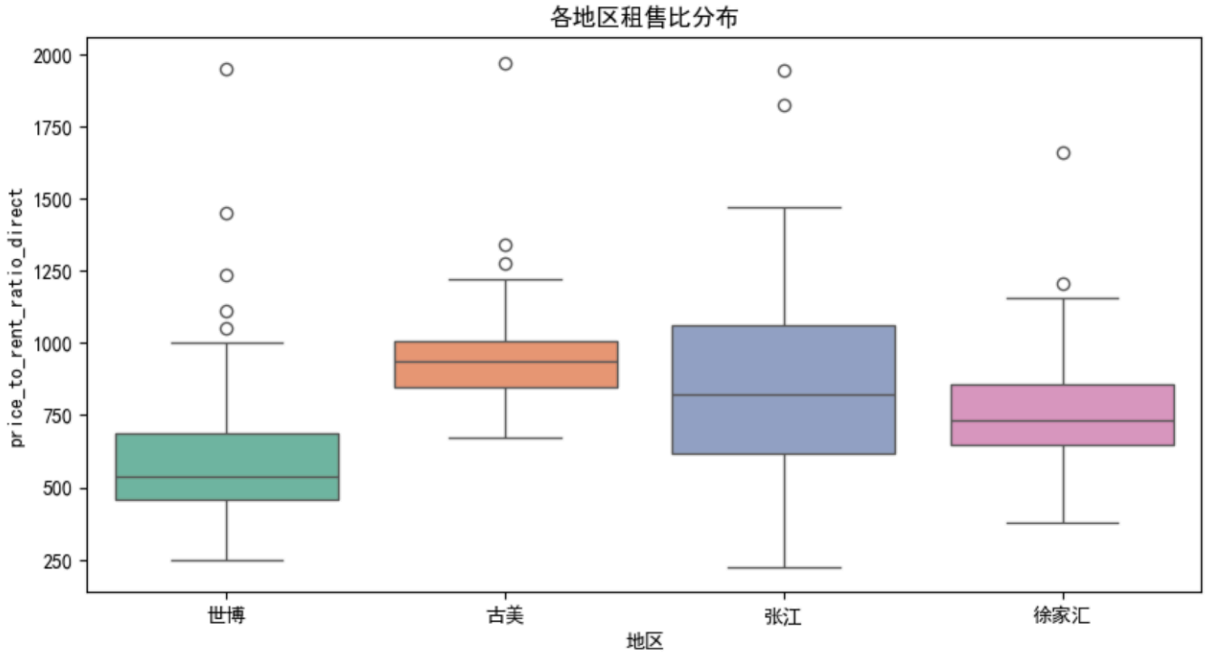


3. 各地区售价与租金箱线图





4. 各地区租售比箱线图



观测到一些离群点，经核实应为信息发布者个人偏好或误差，非数据收集与处理问题。

三种租售比方法的比较与选择

我们利用四个地区（徐家汇、人民广场、虹桥、静安寺）的数据，通过三种方法计算了租售比。

1. **直接计算法**:
    - **优点**: 真实反映市场抽样，尤其是在地区层面进行宏观对比时（如 Figure A Part 1），能清晰地展示出不同板块的投资回报概况。
    - **缺点**: 在小区层面，依然受样本量影响较大。
2. **基础模型法 (`单价 ~ 面积 + 小区`)**:
    - **优点**: 模型在更大的数据集上进行学习，使得对每个小区的固定效应估计更准确，平滑噪声的效果更好。
    - **缺点**: 未显式利用“地区”信息。它只是把“静安寺的A小区”和“虹桥的B小区”看作两个独立的类别变量，没有认识到它们在地理和经济上的关联与差异。
3. **高级模型法 (`单价 ~ ... + 面积:地区`)**:
    - **优点**: 这是本次分析中最具洞察力的模型。通过引入**面积和地区的交互项**，模型不仅学到了每个小区的基准价格，还学到了**面积对价格的影响在不同地区是不同的**。例如，模型可能会发现，在静安寺，面积增加10平米带来的单价下降幅度，要小于在虹桥的下降幅度。这更符合经济直觉，因为核心区的空间价值更高。这使得模型的预测更为精准（Adj. R^2更高）。
    - **缺点**: 模型复杂性更高。

应该信任哪一个？——基于样本量和分析目标考量

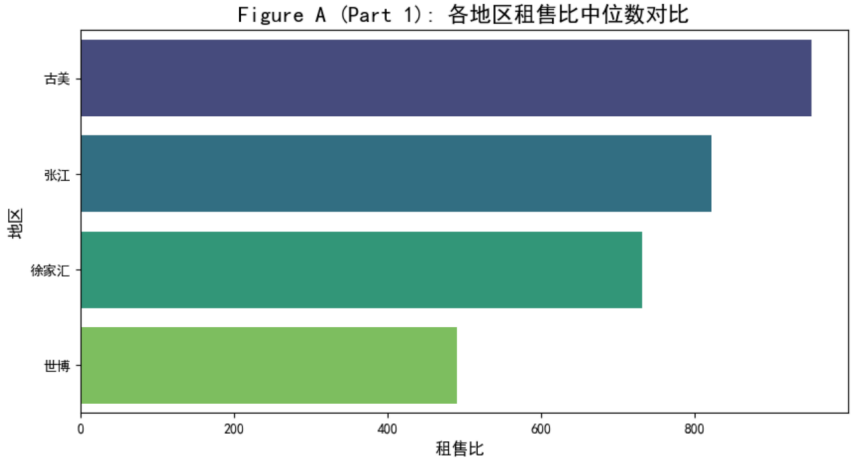
- **进行宏观地区对比时**:
    - **直接计算法（聚合到地区层面）非常可靠**。它直接回答了“从现有数据看，哪个地区的平均租售比最高？”
- **评估具体小区的投资价值时**:
    - **对于样本量充足的小区**: **直接计算法** 依然是黄金标准，因为它最真实。

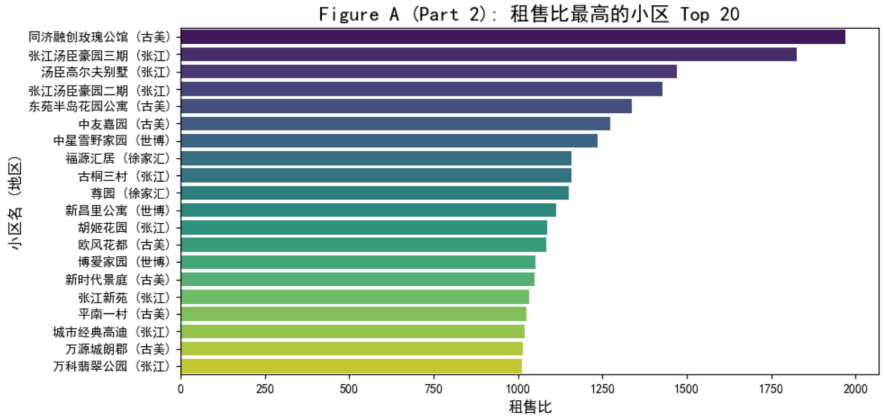
- **对于样本量稀疏的小区: 高级模型法 (方法三) 是最佳选择**。它的优势在于，即使一个小区的样本很少，模型也能利用从**同地区其他小区**以及**其他地区**学习到的“面积-价格”规律，给出一个最合理的推断。它不仅“借鉴”了小区本身的数据，还“借鉴”了其所在地区的整体模式。

总结: 拥有多地区数据后，我们的分析从单一层级（小区）提升到了多层级（地区-小区）。高级模型通过引入交互项，出色地利用了这一数据结构优势，提供了最强的预测能力和最深刻的洞察。因此，在需要对稀疏数据进行推断时，**高级模型的结果最值得信赖**。

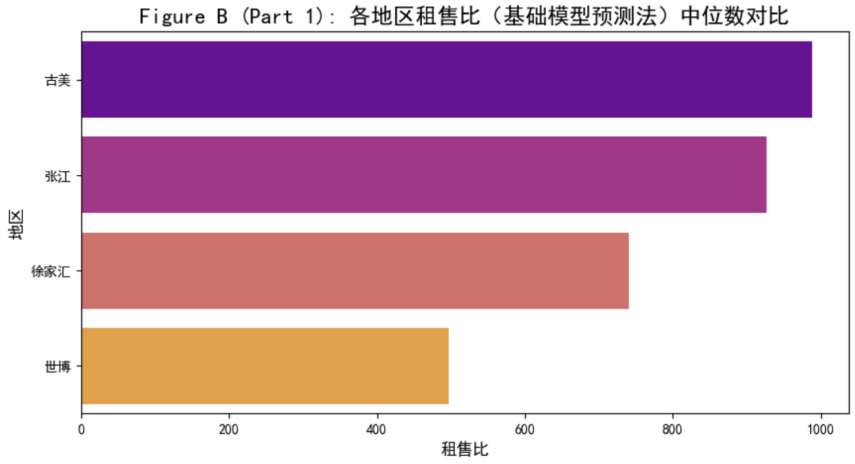
结果展示

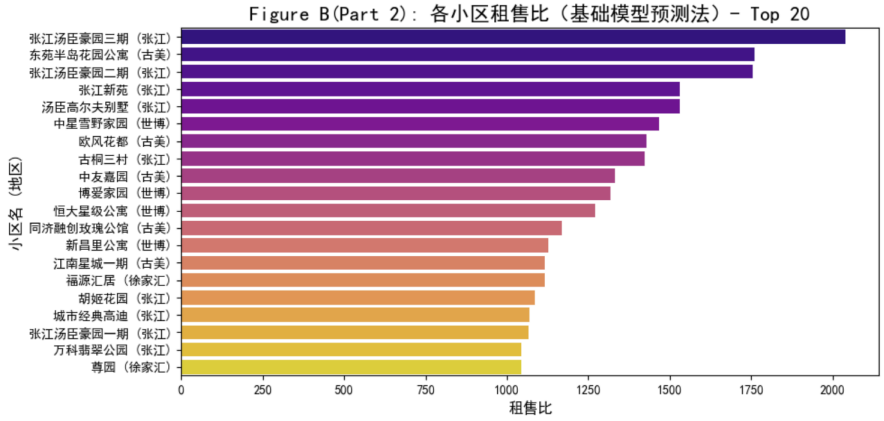
1. direct median



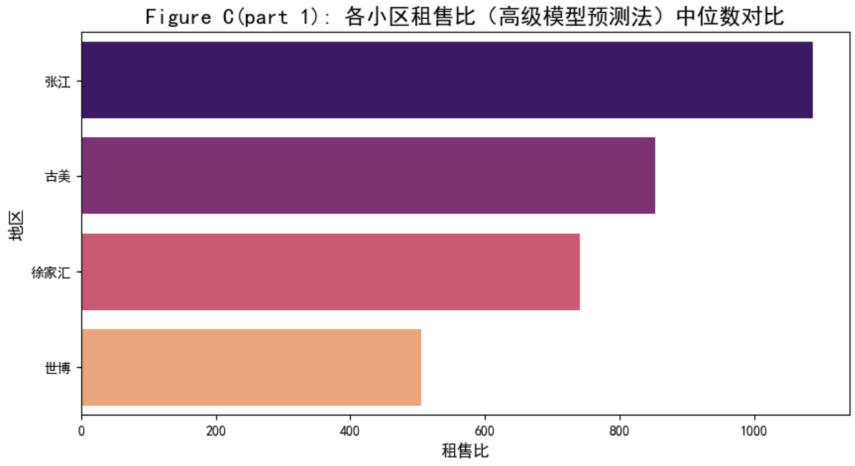


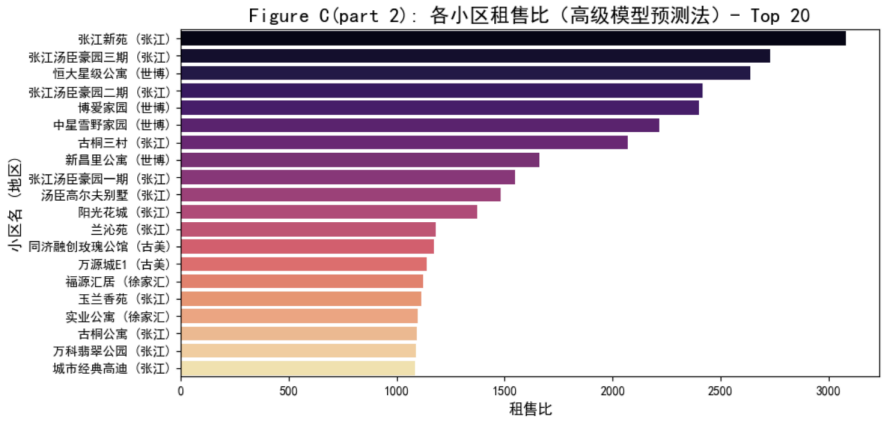
2. 简单线性回归





3. 加入交互项和二次项的回归





4. 模型R2对比

```
              Model  R-squared  Adj. R-squared
0   Model 1 (Price)   0.876976        0.867767
1  Model 1+ (Price)   0.884283        0.875405
2    Model 2 (Rent)   0.627515        0.605476
3   Model 2+ (Rent)   0.674752        0.655037
```

5. 预测结果展示

	region	community_name	esf_count	zu_count	price_to_rent_ratio_direct	price_to_rent_ratio_model	price_to_rent_ratio_model_plus
0	世博	万科金色城品	3	4	620.051538	675.919208	757.601910
1	世博	上南一村	24	34	435.958308	452.572156	472.452156
2	世博	上南七村	10	20	383.443333	385.484391	397.563047
3	世博	上南三村	20	14	417.436385	406.791311	421.715983
4	世博	上南九村	71	46	538.598898	534.149228	555.895053
5	世博	上南二村	13	48	408.724364	406.778211	423.942914
6	世博	上南五村	13	39	438.813187	427.439760	436.849788
7	世博	上南八村	29	8	478.246339	412.271255	415.634921
8	世博	上南六村	37	42	395.513500	405.863564	399.241007