对第五组模型代码的审阅报告

一、代码优点

本组运行了代码，确认代码实现了数据加载、预处理、特征选择、模型训练（OLS 回归）及预测输出等功能。代码在以下方面有较为突出的表现：

1.数据处理较全面

(a) 文本信息结构化：通过正则表达式从原始字段（如楼层、朝向、户型）中精准提取数值信息（如“低楼层”拆分为 relative_height 和 total_floors），并利用 cn2an 库处理中文数字转换（如“两梯四户”转为数值比例），极大提升了数据可用性。

(b) 缺失值智能填充：针对不同场景采用多样化策略，如：分类变量（如“别墅类型”）用“非别墅”填充 NaN；连续变量（如“套内面积”）通过计算训练集的“建筑面积-套内面积”平均比例补全；文本描述字段（如“核心卖点”）统一标记为“有描述/无描述”，避免信息丢失。

2.地理层级关系有所创新

(a) 树状结构构建：通过 get_location_relation_df 函数建立城市、区域、板块、小区的层级关系表，动态处理测试集中的新地点。

(b) 邻居分配策略：采用“上级法”为测试集新地点分配训练集中的邻居（如新板块归属同一区域的其他板块），并在虚拟变量上加权（1/m），增强模型适用性，避免因新类别导致的预测失效。

3.特征设计丰富

(a) 非线性特征扩展：对连续变量（如面积、房间数）生成多项式、对数、倒数等非线性项（如 log(area_gross+1)^3），捕捉潜在复杂关系。

(b) 交互项设计：生成环线与城市虚拟变量的交互项（如 location1_北京*ring_三环），体现地域特异性，提升模型解释力。

4.模型优化细节

(a) 目标变量变换：对右偏的房价取对数拟合，符合线性模型假设，提升预测精度。

(b) 特征筛选：通过分类型变量（如朝向、装修）的虚拟变量化，充分挖掘类别信息。

二、数据处理

对于数据处理部分，我们认为该代码存在如下几方面问题：

1. 异常值处理

(a) 极端值处理

在房产数据集中存在两种极端值，一部分是高价异常值，可能来源于核心地段的豪华住宅或特殊产权房产，另一部分是具有特殊用途的极小户型（如学区房指标房）和超大户型，因此面积-价格关系呈现非线性分段特征。代码并未考虑这种特性，而是直接使用原始数据训练，这对 OLS 回归线的斜率有显著影响。

(b) **缺失值处理**

代码简单使用均值填充缺失的套内面积，忽略了不同楼盘、不同建筑结构的面积比等差异，建议根据不同的数据类型选用多种数据填充方式，并进行合理性检验。

2. **特征工程**

(a) **地理位置处理**

代码采用城市→区域→板块→小区四级编码，在测试集中有大量小区未出现在训练集，于是代码采用了借用邻居信息的方法，这种特征处理虽然较为创新，但也会导致信息过于冗余。四级编码中下级地理信息与上级存在较强的相关性，前 3 个主成分已能解释大部分的地理信息方差。

(b) **特征选择机制缺失**

代码并未考虑地理空间的连续性特征，单一地使用小区邻居信息借用的方法，忽略了一些较为重要的空间关系，比如与地铁站的距离（直线距离与路网距离差异）、周边配套设施的空间密度和区域边界效应（如行政区交界处的价格突变）等，市场季节性等关键特征也被忽略，这会导致经济含义的缺失。

(c) **特征维度爆炸**

样本数与特征数比例失调，当前数据存在严重的维度灾难问题,特征矩阵达到 4059 维，而训练样本为 84133 条，样本特征比仅约为 20:1。这会导致模型自由度过高，容易记忆噪声，存在着一定的过拟合风险，训练集 R^2 虚高，验证集表现波动大，也会带来计算效率的问题，

三、样本划分与数据泄漏问题

在对模型代码进行审阅与验证的过程中,我们发现其在样本划分与评估流程中存在若干关键性问题，可能对模型结果的稳定性与泛化能力构成潜在威胁，具体如下：

1、该项目未按照要求将 train 数据集，划分为 80%训练集和 20%的测试集。这一规范性操作的缺失使得模型评估可能基于全量数据完成，从而引发数据泄漏。数据泄漏会导致模型在训练阶段意外获取测试数据的信息，从而显著夸大其在测试阶段的表现，进而误导模型的实际预测能力判断。

2、项目中仅采用了简单的线性回归模型进行建模，未使用 Ridge、LASSO 和与 ElasticNet 模型。这些方法不仅可以提升模型对高维特征数据的适应能力，还能有效缓解多重共线性问题，提升模型在未见数据上的泛化性能。缺乏对不同正则化策略的比较与调优，削弱了模型选择的科学性与鲁棒性。

3、在验证过程中发现，项目未使用交叉验证对模型进行稳健性评估。交叉验证有助于减少样本划分带来的偶然性偏差，提高模型评估的可靠性。在交叉验证缺位的情况下，模型表现极有可能受限于特定划分结构，难以推广至其他样本。

4、该项目在模型性能评估方面亦存在不完整之处。其结果分析部分缺少如平均绝对误差（MAE）和均方根误差（RMSE）等关键指标。这些指标是衡量回归模型预测误差的重要工具，能够从不同角度量化模型拟合程度。缺失这些指标会妨碍对模型预测精度的系统性评价。

四、问题总结及建议

以下是对于代码中可能存在的一些问题的总结和建议：

1、代码层面：

(a) **可读性与规范性欠缺**：代码整体的可读性较差，主要体现在缺少必要的注释。这使得对于其他代码使用者和后续维护来说，理解代码逻辑和功能实现存在较大困难。此外，部分代码的写作不够规范，例如采用'+'号来拼接字符串、多次重复调用 DataFrame 等操作，这不仅降低了代码的可读性，还可能引入潜在的错误和风险。因此，建议添加必要的注释，并更规范，更简洁的按照规范的编程风格进行修改和完善。

(b) **代码可操作性不足**：部分代码可操作性不强。比如在选择特征变量时，无法简单的选用希望使用的特征，或者舍弃不合适的特征，难以满足模型调试过程中对特征变量进行精细化调整的需求。建议新增相应的代码模块，以便能够更方便和更精确的调试模型

2、模型层面：

(a) **模型选择单一**：仅使用了 Linear 线性模型进行训练和预测，这种单一的模型选择限制了模型对复杂数据关系的拟合能力。相较于 Ridge 和 Lasso，Linear 模型容易出现过拟合，导致模型表现不佳，但模型并没有这方面的考虑。建议添加 Ridge 和 Lasso 模型。

(b) **没有进行交叉验证**：未对模型进行交叉验证，这使得模型的评估结果可能存在较大的偏差和不确定性。缺乏交叉验证会导致无法正确评估模型的预测效果，无法即时识别过拟合导致模型性能下降。

(c) **模型评估不全面**：缺少对模型预测效果的评估，代码中没有输出模型评价指标。应该输出相关的预测指标，如 MAE、MSE、RMSE、R^2等，这些指标有助于评估模型的拟合情况等，以有助于对比不同的模型的拟合效果。建议加上相关输出。

3、特征工程层面：

(a) **特征文件未充分利用**： 没有利用 rent 和 detail 两个文件。但其中有许多重要信息，比如租金、绿化率等特征，对房价预测有重要影响。

(b) **特征工程不够细致，还有许多信息未识别**：特征工程的处理相对简单，还有进一步的提升空间。例如可以新增文本识别类任务，从文本数据中提取有价值的信息并转化为特征（如附近是否有地铁站、医院等信息），这些信息对于房间预测有重要影响。

4、其它建议：

(a) **高次项的选择不合理**：模型大量使用了高次项，但部分高次项过于复杂（比如取对数后再取三次方项），这不仅增加了模型的复杂度和计算负担，还可能导致过拟合使模型性能下降。同时，从经济学角度来看，这些复杂的高次项缺乏明确的经济学含义，使得难以对模型参数进行直观上的解释。建议更加谨慎地选择高次项。

(b) **代码运行效率较低**：相较于同等预测水平的代码，本代码运行所需内存大，运行时间偏长。我们认为主要问题在于，使用了过多的无效特征，模型的特征多达 4000 个，但实际上，如果采用更精确的特征，仅需 400 个左右就可以达到相同的预测水平。建议对特征进行更精细的筛选，去除冗余和不相关的特征，以提高代码的执行效率。

(c) **部分代码可能复杂度与实际效果不匹配：**通过寻找上一级地址实现训练集中缺失地址数据的补全比较有创意。 但代码实现较为复杂，且运行时间长内存占用大，但对实际预测结果的影响可能较小。对于缺失小区信息的处理完全可以用更加简便的方式处理（如采用 groupby，mean 函数取同类平均），这种方法更效率更高，并且也可以达到近似的效果。建议在编写代码时同时考虑资源占用和实现效率，采用更加高效的代码。

小组分工：

1、代码运行与问题讨论：小组共同完成。

2、报告写作：

第一部分：朱堃琳

第二部分：王成林

第三部分：张沛渊

第四部分：唐汇宸

3、文本整合：唐汇宸