GROUP 6
Presentation

Data Processing:

1. 对自然语言特征，利用大模型分析情绪，量化为数值类特征（transformers库pipeline函数）
2. 将"房屋朝向"转化为"南北通透"
3. 用两次交易时间作差来填充"房屋年限"缺失值 44510→ 25464
4. 生成"内外面积比"

Feature Engineering:

1. 删除Outlier：变种IQR法，设置了最低阈值
2. 将类别型特征分成低基数（<=10）与高基数（>10）两类，进行独热编码时忽略低频类别
3. 交互项只在低基数类别与数值型之间产生
4. 将填补缺失值的操作封装在pipeline中，训练时仅对训练集进行操作，防止数据泄露

Metrics		In sample	Out of sample	Cross-validation	Kaggle score
OLS	Sales	412715.4213	440446.8167	441117.8479	**67.23**
	Rent	111315.0966	113437.9438	113364.9896	
Lasso	Sales	413727.1808	440771.2288	441704.7651	**67.21**
	Rent	111394.0193	113444.4333	113412.9368	
Ridge	Sales	412715.4335	440446.8295	441117.8579	**67.23**
	Rent	111315.0967	113437.9443	113364.9899	

李书桓
2023200055

1. 基于地理聚类的缺失值填充

使用K近邻算法依据经纬度坐标进行空间聚类，对缺失值进行智能化填充。

2. 目标变量的对数变换

对偏态分布的目标变量进行对数变换，使其更符合正态分布。

3. 多源文本信息的结构化提取

从“房屋优势”、“周边配套”、“交通出行”等多个非结构化文本特征中，提取出结构化信息。

4. 客户反馈的情感分析

对文本评论进行情感打分，将主观的定性反馈转化为可量化的情感指数，捕捉潜在的用户情绪因素。

Metrics		In sample	Out of sample	Cross-validation	Kaggle score
OLS	Sales	607527.82	572441.1	609769.42	**66.12**
	Rent	127811.26	127885.04	128074.34	
Lasso	Sales	665657.77	634359.47	666906.25	**59.32**
	Rent	138983.25	139308.66	139175.67	
Ridge	Sales	607540.74	572476.17	609778.49	**66.12**
	Rent	127799.57	127951.17	128061.48	
ElasticNet	Sales	635912.38	603200.61	637290.66	**62.43**
	Rent	132606.44	132423.27	132821.5	

朱奕宁

2023200057

陈纳川
2023200094

Data Processing:

1. 计算“得房率”（套内面积/建筑面积），根据中位数计算缺失套内面积的预测值
2. 使用正则表达式提取“厅室卫”、“梯户”、“楼层”等特征
3. 将含有多个特征的文字（商品房/使用权）,根据 keywords_list 拆分属性
4. 对于水电暖等缺失值，使用“板块”-> “区域” -> “城市”训练集众数（中位数）分层填充（经济位置特点）
5. 使用 K-means 聚类方法对每个城市内样本进行分类（地理位置特点）
6. 提取交易年份和交易月份（季节性），生成分类变量

Feature Engineering:

1. 删除Outlier：缩尾法，将 Q1 和 Q99 外的 Outlier 替换
2. 计算梯户比、厅户比等比率类特征
3. 对于|skew|>1 的数值类数据，进行对数转化，对于双峰分布等另类数据进行分享操作（K-means）
4. 计算每个城市内样本到中心点的距离和距离的平方（使用样本经纬度中位数作为中心点）
5. 使用循环构建特征，使用 Ridge 和 LASSO 预测（受算力约束未实现）

Metrics		In sample	Out of sample	Cross-validation	Kaggle score
OLS	Sales	349859.7959	347307.4272	353300.2521	**74.14**
	Rent	89231.9096	90369.2374	90431.1436	
Lasso	Sales	505194.8406	499899.4255	504794.9683	**68.23**
	Rent	125104.5866	123339.3151	125154.9308	
Ridge	Sales	349859.7974	347221.3134	353359.2267	**73.96**
	Rent	89231.91	90397.7409	90445.4053	

1.K-Fold目标编码

针对“城市”、“板块”等高基数特征，仅用训练集计算“板块组合”的均价；同时在训练集内部使用6折交叉验证编码，严格防泄露。

2.K-Means地理聚类

函数仅在训练集经纬度上 fit K-Means模型（50簇），再用这个拟合好的模型去 predict 所有数据（含测试集）的地理簇标签

3.LassoCV自动化特征筛选

在725个工程特征上，用 LassoCV 在完整训练集上拟合，通过6折CV自动找到最佳 alpha，并筛选出92个核心特征用于建模。

4.K-Means分箱

仅在训练集上让模型自动学习“房龄”、“楼层”的5个自然断点，再用此分箱器 transform 全量数据,捕捉其对价格的非线性阶梯效应

5.防数据泄露插补

所有的处理函数都传入 n_train 参数，仅在训练集 (iloc[:n_train]) 上计算中位数，再用这个固定值 fillna 全量数据。

6.系统性非线性变换

批量对高偏度特征应用对数变换，并自动生成“室”、“厅”、“得房率”等多项式交互项以捕捉组合效应。

删除异常值后，训练集：price 103264行 rent 98240行

Metrics		In sample	Out of sample	Cross-validation	Kaggle score
OLS	Sales	471087.256	469931.274	471480.134	65.25
	Rent	114202.194	113185.516	114128.144	
Lasso	Sales	471265.017	469349.701	471381.119	67.36
	Rent	113995.175	112894.116	113860.242	
Ridge	Sales	471339.828	469500.234	471480.134	65.37
	Rent	114267.411	113161.249	114126.872	

黄鑫
2023200098