任务完成情况以及数据分析报告

任务一：数据爬取完整（Homework3-1）

代码部分的 1.1-1.8 为房产数据的爬取；2.1-2.7 为租房数据的爬取

任务二：小组数据整合（Homework3-2）

1.整合数据

小组在群内进行了分享，个人各自进行了整合数据。

2.数据描述（Homework3-3 第一问）

代码部分的 3.1-3.2 是对数据进行清洗与描述。一是把重复值和残缺值删掉，二是分组进行房产数据和租房数据的描述性统计。结果如下：

房价数据： 原始2898条， 清洗后854条
租金数据： 原始2997条， 清洗后626条

各区块房价数据描述性统计：
苏州桥 (区块 1)：

	m2	total_price
count	169.000000	1.690000e+02
mean	88.281538	6.646331e+06
std	60.300798	3.897075e+06
min	26.800000	2.550000e+06
25%	56.700000	4.300000e+06
50%	66.910000	5.280000e+06
75%	91.280000	7.500000e+06
max	337.400000	2.690000e+07

万柳 (区块 2)：

	m2	total_price
count	281.000000	2.810000e+02
mean	186.679431	3.162324e+07
std	103.081810	2.612303e+07
min	44.170000	3.790000e+06
25%	124.000000	1.525000e+07
50%	156.000000	2.290000e+07
75%	237.950000	3.980000e+07
max	745.000000	1.940000e+08

北太平庄 (区块 3)：

	m2	total_price
count	117.000000	1.170000e+02
mean	132.097009	1.142299e+07
std	88.001812	7.494976e+06
min	26.500000	1.750000e+06
25%	78.400000	5.900000e+06
50%	95.150000	8.500000e+06
75%	150.550000	1.558000e+07
max	491.050000	2.950000e+07

世纪城 (区块 4)：

	m2	total_price
count	287.000000	2.870000e+02
mean	164.537491	1.892066e+07
std	52.627178	8.905256e+06
min	57.000000	1.000000e+06
25%	131.400000	1.370000e+07
50%	164.000000	1.700000e+07
75%	184.730000	2.225000e+07
max	494.340000	8.000000e+07

各区块租金数据描述性统计：
苏州桥 (区块 1)：

	m2	rent_price
count	211.000000	211.000000
mean	65.919431	8516.146919
std	28.124183	3241.072734
min	8.000000	2100.000000
25%	54.000000	6760.000000
50%	63.000000	7630.000000
75%	78.000000	9700.000000
max	254.000000	20000.000000

万柳 (区块 2)：

	m2	rent_price
count	127.000000	127.000000
mean	238.708661	46724.330709
std	203.788371	36884.105636
min	42.000000	5500.000000
25%	113.500000	15700.000000
50%	200.000000	32000.000000
75%	321.500000	68000.000000
max	1501.000000	160000.000000

北太平庄 (区块 3)：

	m2	rent_price
count	159.000000	159.000000
mean	69.493082	9045.993711
std	44.102922	5340.695178
min	10.000000	1700.000000
25%	50.000000	6600.000000
50%	57.000000	7400.000000
75%	81.000000	9950.000000
max	367.000000	45000.000000

世纪城 (区块 4)：

	m2	rent_price
count	129.000000	129.000000
mean	154.604651	18693.643411
std	45.482555	10126.197147
min	57.000000	7500.000000
25%	127.000000	13000.000000
50%	161.000000	15980.000000
75%	173.000000	20000.000000
max	323.000000	70000.000000

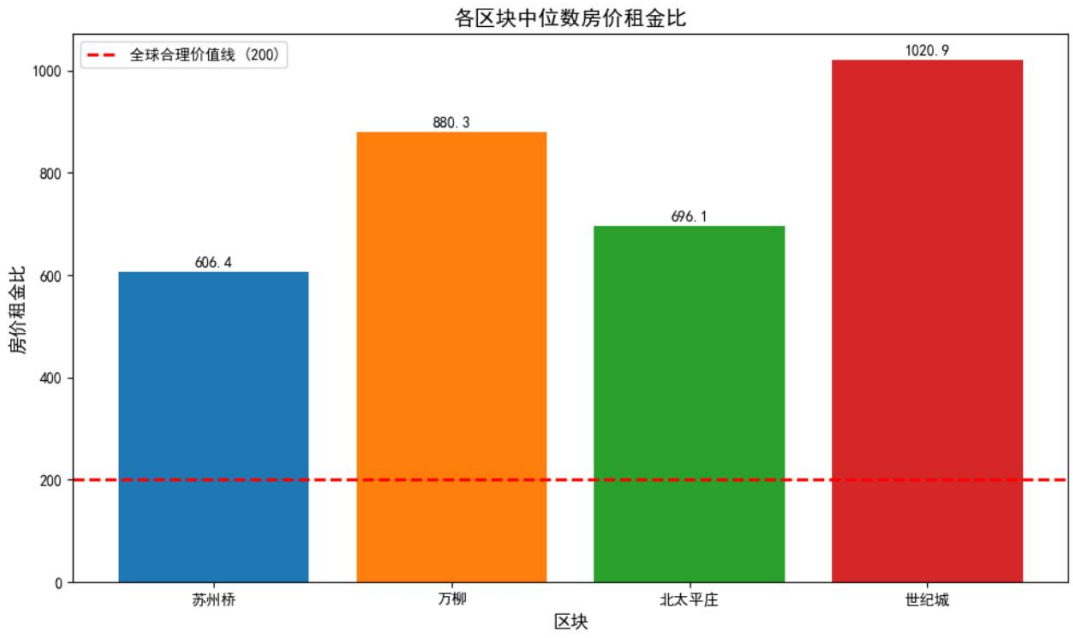
从生成内容可以看出，各区块的房价和租金数据均在合理范围内，未发现明显的异常值。数据分布符合北京市房产市场的实际情况，最大值和最小值均在可接受的经济学范围内。

任务三：回归分析

1.计算原始数据房价租金比并绘制图 A（Homework3-3 第二问）

代码部分的 4.1-4.2 分别进行了计算和柱状图展示。结果如下：

```
苏州桥：
  中位数房价每平方米：78092.00
  中位数租金每平方米：128.79
  房价租金比：606.36
万柳：
  中位数房价每平方米：144043.00
  中位数租金每平方米：163.64
  房价租金比：880.26
北太平庄：
  中位数房价每平方米：90498.00
  中位数租金每平方米：130.00
  房价租金比：696.14
世纪城：
  中位数房价每平方米：109727.00
  中位数租金每平方米：107.48
  房价租金比：1020.94
```



2.基础线性回归模型分析并绘制图 B（Homework3-4）

代码部分的 5.1-5.3 进行了基础回归模型的建立、预测、指标可视化。

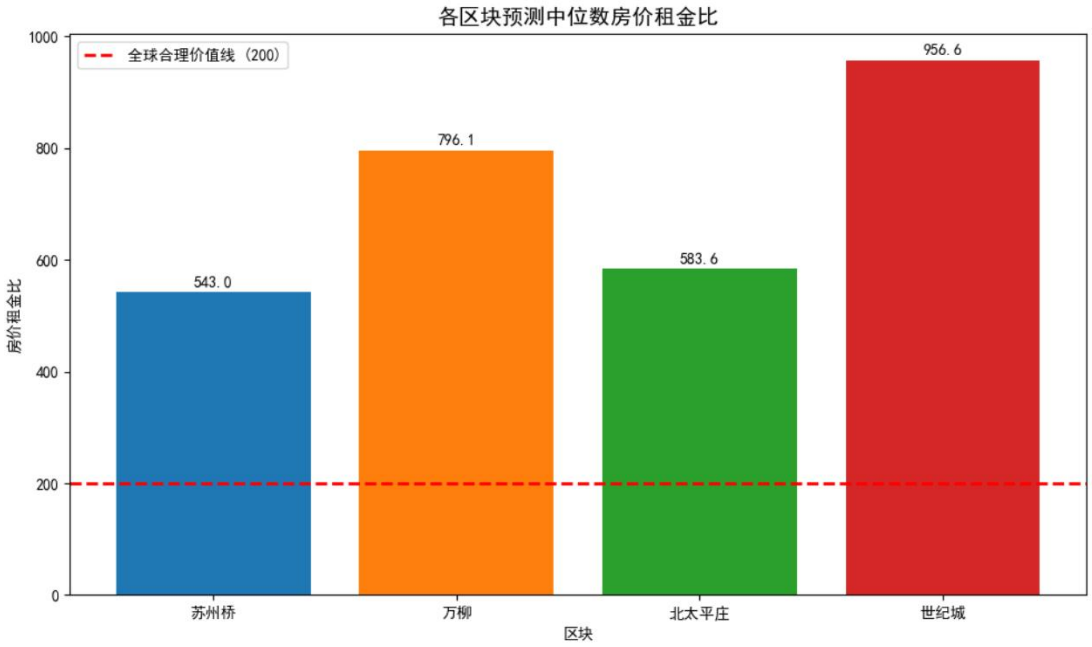
建立的两个回归模型分别为：

$$\text{price}/_{m^2} = \beta_0 + \beta_1 \times m^2 + \beta_2 \times location + \varepsilon$$

$$\text{rent}/_{m^2} = \beta_0' + \beta_1' \times m^2 + \beta_2' \times location + \varepsilon'$$

结果如下：

```
苏州桥：
  预测房价中位数：74554.19 元/㎡
  预测租金中位数：137.31 元/㎡
  预测房价租金比：542.95
万柳：
  预测房价中位数：148434.45 元/㎡
  预测租金中位数：186.44 元/㎡
  预测房价租金比：796.14
北太平庄：
  预测房价中位数：80589.95 元/㎡
  预测租金中位数：138.10 元/㎡
  预测房价租金比：583.57
世纪城：
  预测房价中位数：112843.31 元/㎡
  预测租金中位数：117.96 元/㎡
  预测房价租金比：956.61
预测完成
```



3.增强线性回归模型分析并绘制图 C（Homework3-5）

代码部分的 6.1-6.4 进行了增强线性回归模型的建立、预测、指标可视化。

（1）建立的两个增强回归模型分别为：

$$\text{price}/_{m^2} = \beta_0 + \beta_1 \times m^2 + \beta_2 \times location + \beta_3 \times \left(m^2\right)^2 +$$
$$\beta_4 \times m^2 \cdot location + \beta_5 \times \text{location} \cdot \text{distance} + \beta_6 \times m^2 \cdot \text{distance} + \varepsilon$$
$$\text{rent}/_{m^2} = \beta_0' + \beta_1' \times m^2 + \beta_2' \times location + \beta_3' \times \left(m^2\right)^2 +$$
$$\beta_4' \times m^2 \cdot location + \beta_5' \times \text{location} \cdot \text{distance} + \beta_6' \times m^2 \cdot \text{distance} + \varepsilon'$$

增强模型的特征包括了：一次项（m^2，$distance$，$location$），非线性项（$(m^2)^2$）交互项（$m^2 \times group$，$group \times distance$，$m^2 \times distance$）。

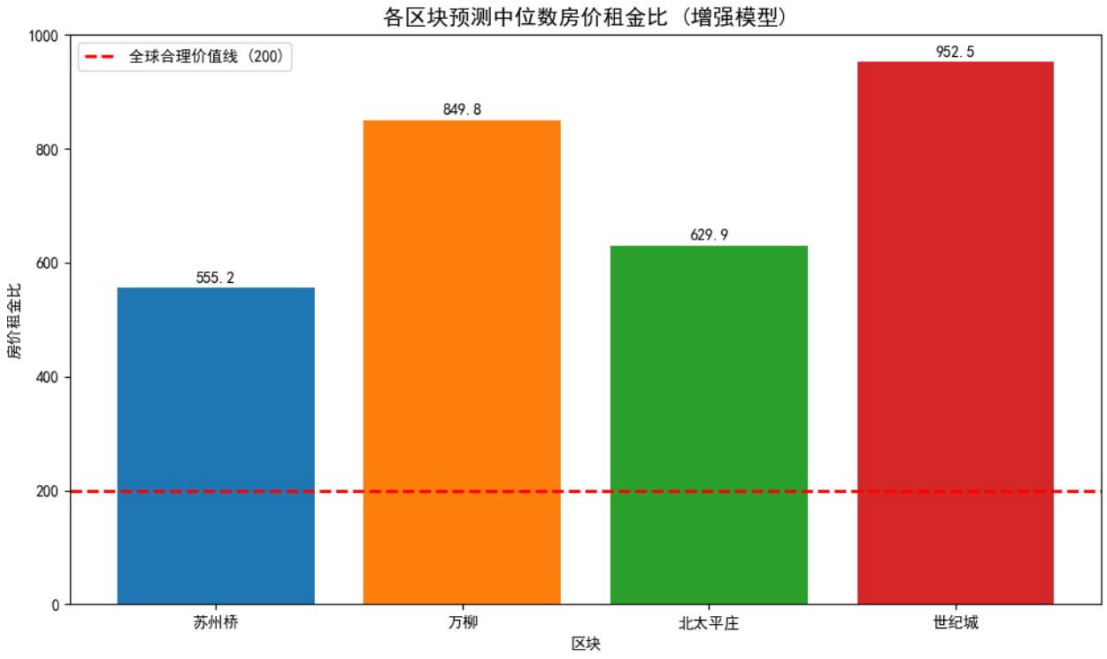
（2）两个模型 R^2 的比较如下所示：

```
模型R²比较:
房价模型 - 原始: 0.5764, 增强: 0.6796
租金模型 - 原始: 0.1953, 增强: 0.4193
```

很明显增强的模型 R^2 更大，这是因为模型更灵活，更能捕捉复杂模式，例如非线性项使得可以拟合一条抛物线，比直线灵活，能够湾区贴近非线性的数据点。引入交互项考虑了不同自变量之间的联合效应，从而做出更精准的预测，进一步减少残差。

（3）指标计算和可视化如下所示：

```
苏州桥:
  增强模型预测房价中位数: 78592.74 元/㎡
  增强模型预测租金中位数: 141.55 元/㎡
  增强模型预测房价租金比: 555.21
万柳:
  增强模型预测房价中位数: 142453.17 元/㎡
  增强模型预测租金中位数: 167.64 元/㎡
  增强模型预测房价租金比: 849.77
北太平庄:
  增强模型预测房价中位数: 88291.78 元/㎡
  增强模型预测租金中位数: 140.17 元/㎡
  增强模型预测房价租金比: 629.91
世纪城:
  增强模型预测房价中位数: 111275.45 元/㎡
  增强模型预测租金中位数: 116.83 元/㎡
  增强模型预测房价租金比: 952.49
```

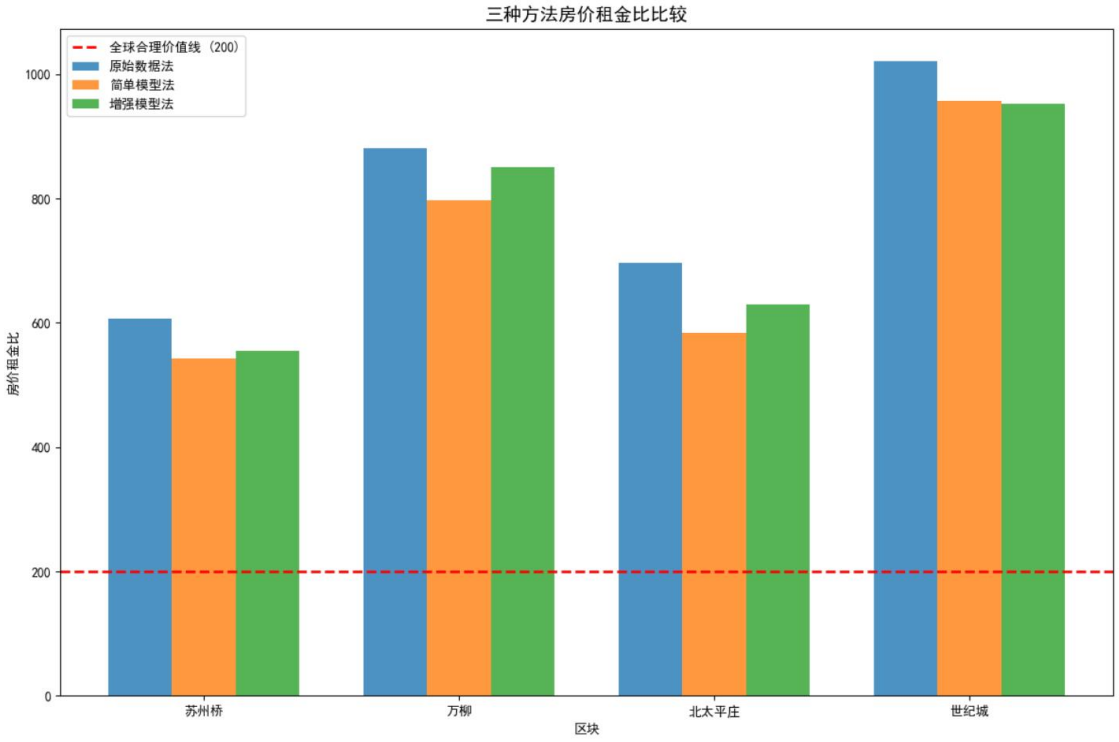


任务四：一些自行探索

1. 在代码部分的 6.5-6.6，我还对比了原始数据和两种模型的预测情况，绘制了原始数据、基础模型、增强模型所计算的指标的堆叠条形图。根据三种方法房价租金比比较图分析，我们可以得到：

（1）所有方法计算出的房价租金比均显著高于全球合理价值线（200），表明该区域房产市场可能存在高估现象。

（2）各区块比值排序一致，万柳地区房价租金比最高，苏州桥次之，世纪城和北太平庄相对较低。这种普遍高于合理水平的情况说明这些房产更适合自住而非租赁投资。



总结

总体来说，本研究报告是基于 Python 数据分析课程的作业要求，通过完整的爬虫数据采集、数据清洗、建模分析流程，对北京市四个主要区块的房产市场进行深入研究，完整展示了 Python 在数据科学项目中的全流程应用。