1 文件目录和说明

在代码尚未运行时，文件目录如下。

```
project
├── chromedriver.exe          // chrome驱动
├── scrape_house_price.ipynb // 爬取房屋价格信息
├── scrape_rent_price.ipynb  // 爬取租赁价格信息
├── data_research.ipynb       // 异常值筛查、FigA绘制、中位数计算
└── modeling.ipynb            // 模型回归（简单&PROMAX模型）、FigB、FigC
```

代码运行后（稍后会介绍运行顺序），会自动创建文件夹，将数据在文件夹中保存。全部运行后的项目目录如下。

```
project
├── chromedriver.exe          // chrome驱动
├── scrape_house_price.ipynb // 爬取房屋价格信息
├── scrape_rent_price.ipynb  // 爬取租赁价格信息
├── data_research.ipynb       // 异常值筛查、FigA绘制、中位数计算
├── modeling.ipynb            // 模型回归（简单&PROMAX模型）、FigB、
├── figures     // 作业要求中的图像
|   ├── figure_A.png
|   ├── figure_B.png
|   └── figure_C.png
├── house_price     // 爬取的房屋价格
|   ├── house_price_huilongguan.csv
|   ├── house_price_huoying.csv
|   ├── house_price_shahecsv
|   └── house_price_tiantongyuan.csv
├── house_rent      // 爬取的租赁价格
|   ├── house_rent_huilongguan.csv
|   ├── house_rent_huoying.csv
|   ├── house_rent_shahecsv
|   └── house_rent_tiantongyuan.csv
├── cleaned_data    // 数据清理后的价格
|   ├── cleaned_price_回龙观.csv
|   ├── cleaned_price_霍营.csv
|   ├── cleaned_price_沙河.csv
|   ├── cleaned_price_天通苑.csv
|   ├── cleaned_rent_回龙观.csv
|   ├── cleaned_rent_霍营.csv
|   ├── cleaned_rent_沙河.csv
|   └── cleaned_rent_天通苑.csv
├──predicted_data   // 利用模型预测的价格
|   ├── price_data_with_predictions.csv
|   └── price_data_with_predictions.csv
```

2 运行复现说明

代码没有默认只爬取20页数据，而是一直爬取所有页的信息直到末页，具体实现方式详见代码。

若要得到最终所有结果，应当运行所有ipynb且按照以下顺序。

```
1. scrape_house_price.ipynb // 爬取房屋价格信息
2. scrape_rent_price.ipynb  // 爬取租赁价格信息
3. data_research.ipynb      // 异常值筛查、FigA、数据清理
4. modeling.ipynb           // 模型回归（简单&PROMAX模型）、FigB、FigC
```

运行失败时可能出现的问题：

1. 兼容性问题：由于电脑的库版本不匹配等，可以将电脑所有库更新后重试。
2. chrome_driver不匹配：由于新版selenium会自动下载对应版本的驱动，但是由于速度过慢我主动下载了驱动并利用本地驱动启动，这样可能会造成chrome的版本不匹配等问题，所以无法运行，只需要安装对应版本的chrome即可。

3 代码描述

1. 代码中含有详尽的注释，每一步都有对应的注释，包含每一步的目的和操作。
2. 由于ipynb的特性，我的所有运行结果都呈现在了代码中，所以代码中也包含了运行痕迹和最终结果。

4 模型描述和回归结果

4.1 简单回归模型

我们有 4 个位置：`loc1`, `loc2`, `loc3`, `loc4`。
在做虚拟变量编码时，以 `loc1` 为基准，因此模型中只保留 `loc2`, `loc3`, `loc4`。

模型公式为：

$$\text{price\_m2/rent\_m2} = \beta_0 + \beta_1 \cdot \text{area} + \beta_2 \cdot loc2 + \beta_3 \cdot loc3 + \beta_4 \cdot loc4 + \epsilon$$

解释：

- β_0：截距，表示在 `loc1` 且面积为 0 时的基准房价。
- β_1：面积的系数，表示面积每增加 1 单位，房价的平均变化量（对所有位置相同）。
- $\beta_2, \beta_3, \beta_4$：位置虚拟变量的系数，表示不同位置相对于基准位置 `loc1` 的固定差异。
- ϵ：误差项。

4.2 PROMAX回归模型

在考虑交互项后，模型允许 **不同位置的面积效应不同。**
依然以 `loc1` 为基准，模型变为：

$$\text{price\_m2/rent\_m2} = \beta_0 + \beta_1 \cdot \text{area} + \beta_2 \cdot loc2 + \beta_3 \cdot loc3 + \beta_4 \cdot loc4 + \beta_5 \cdot (\text{area} \times loc2) + \beta_6 \cdot (\text{area} \times loc3) + \beta_7 \cdot (\text{area} \times loc4) +$$

解释：

- β_0：在 `loc1` 且面积为 0 时的基准房价。
- β_1：在基准位置 `loc1` 下，面积对房价的影响（斜率）。
- $\beta_2, \beta_3, \beta_4$：不同位置相对于 `loc1` 的固定差异。
- $\beta_5, \beta_6, \beta_7$：交互项系数，表示不同位置下面积对房价的额外影响。
  - 例如：`area × loc2` 的系数 β_5 表示在 `loc2` 时，面积对房价的边际效应相对于 `loc1` 的差异。
- ϵ：误差项。

4.3 回归模型效果

4.3.1 简单回归模型

- 售卖数据 **R²**: 0.166
- 租赁数据 **R²**: 0.431

解释

- R² 值较低，说明模型只能解释部分的房价变动，大部分信息没有被捕捉到。
- 原因在于：该模型假设 **所有位置的面积效应是相同的**，忽略了不同地段的房价随面积变化的差异。

4.3.2 PROMAX回归模型

- 售卖数据 **R²**: 0.204
- 租赁数据 **R²**: 0.653

解释

- R² 提升幅度很大，说明模型能解释大部分房价变动，拟合效果显著提高。
- 改进的关键是：**引入了面积与位置的交互项**，允许不同位置对面积的敏感度不同。
  - 原因可能是在靠近市中心位置，面积对价格的边际影响可能更大；而在郊区，面积的影响可能相对较小。
  - 交互项很好地捕捉到了这种差异，因此拟合效果更好。

4.3.3 总结

- **简单模型**：只能提供一个“平均”的面积效应，忽视了不同地段的差异 → 导致低 R^2。
- **交互模型**：允许每个位置都有自己的面积斜率，更贴合实际房价规律 → 大幅提升了预测性能。

5 租售比数据分析

5.1 初始真实数据（未删减）

地区	租售比中位数
回龙观	531.17
霍营	560.22
沙河	683.30
天通苑	513.89

特点：

- 数据未删减，包含所有真实样本。
- 租售比数值合理，反映市场真实水平。
- 样本量大，代表性好。

5.2 简单回归模型预测数据

地区	租售比中位数
回龙观	555.14
天通苑	468.48
沙河	580.24
霍营	753.61

特点：

- 与真实数据接近，数值合理。
- 模型假设面积对房价的影响在所有地区相同。
- 删除了异常值，因此预测较为平稳，但可能略低或略高于真实值。

5.3 加入交互项的回归模型预测数据

地区	租售比中位数
回龙观	624.12
天通苑	625.74
沙河	622.57
霍营	1222.94

特点：

- 预测趋势上能反映面积与位置的交互效应。
- 某些地区（如霍营）预测值明显高于真实水平，可能受样本删除或模型放大效应影响。
- 相对排序在部分地区合理，但绝对值偏高。

5.4 数据可信度分析

1. **基于样本量和绝对数值**：
   - 初始真实数据样本量大，包含异常值，反映市场全貌 → 可信度高。
   - 简单回归模型预测数据基于删减样本，数值接近真实水平，整体可信。
   - 交互模型预测数据绝对值偏高，尤其霍营，可能夸大了实际租售比。
2. **基于趋势和相对差异**：
   - 简单模型无法捕捉面积在不同地区的差异，对趋势解释能力有限。
   - 交互模型能体现不同地区面积效应差异，趋势合理，但数值需校正。
3. **综合判断**：
   - **绝对值参考**：以 **初始真实数据** 为主，保证数值合理性。
   - **趋势和相对排序参考**：可参考 **交互模型预测数据**，了解面积与位置交互效应对租售比的影响。

5.5 总结

1. 可以对交互模型的预测值进行 **比例或缩放校正**，使绝对值接近真实水平。
2. 将初始真实数据与交互模型趋势结合使用，既保留数值可信度，又参考交互效应。
3. 简单回归模型可作为快速参考，但不足以解释区域间面积效应差异。