时序模型文档

2253551 李沅衡

第1步：数据分析与探索

数据导入结果如下：

```
[1] 0
    datesold     postcode        price propertyType     bedrooms
           0            0            0            0            0
'data.frame':   29580 obs. of  5 variables:
 $ datesold    : chr  "2007-02-07 00:00:00" "2007-02-27 00:00:00" "2007-03-07 00:00:00" "2007-03-09 00:00:00" ...
 $ postcode    : int  2607 2906 2905 2905 2906 2905 2607 2606 2902 2906 ...
 $ price       : int  525000 290000 328000 380000 310000 465000 399000 1530000 359000 320000 ...
 $ propertyType: chr  "house" "house" "house" "house" ...
 $ bedrooms    : int  4 3 3 4 3 4 3 4 3 3 ...
   datesold           postcode        price         propertyType         bedrooms
 Length:29580       Min.   :2600   Min.   :  56500   Length:29580       Min.   :0.00
 Class :character   1st Qu.:2607   1st Qu.: 440000   Class :character   1st Qu.:3.00
 Mode  :character   Median :2615   Median : 550000   Mode  :character   Median :3.00
                    Mean   :2730   Mean   : 609736                      Mean   :3.25
                    3rd Qu.:2905   3rd Qu.: 705000                      3rd Qu.:4.00
                    Max.   :2914   Max.   :8000000                      Max.   :5.00
```

```
数据概览：
1. 时间范围： 2007-02-07 至 2019-07-27
2. 房产类型分布：

house  unit
24552  5028

3. 卧室数量分布：

    0     1     2     3     4     5
   30  1627  3598 11933 10442  1950
```

1. 缺失值检查

运行结果：

```
总缺失值数量: 0
    datesold     postcode        price propertyType     bedrooms
           0            0            0            0            0
```

- `总缺失值数量: 0`：
  - **含义**：整个数据集中没有任何缺失值。
  - **重要性**：这意味着所有的数据行都完整，适合进行后续的分析和建模。无需进行缺失值填补或删除操作，节省了数据预处理的时间和精力。
- `colSums(is.na(data))`：
  - **含义**：逐列计算缺失值的数量。输出显示每一列（`datesold`、`postcode`、`price`、`propertyType`、`bedrooms`）的缺失值数量均为0。
  - **重要性**：确认了数据框中每个变量（列）都没有缺失值，进一步确保数据的完整性和质量。

2. 数据框结构（`str(data)` 的输出）

运行结果：

```
'data.frame':   29580 obs. of  5 variables:
 $ datesold    : chr  "2007-02-07 00:00:00" "2007-02-27 00:00:00" "2007-03-07 00:00:00" "2007-03-09 00:00:00" ...
 $ postcode    : int  2607 2906 2905 2905 2906 2905 2607 2606 2902 2906 ...
 $ price       : int  525000 290000 328000 380000 310000 465000 399000 1530000 359000 320000 ...
 $ propertyType: chr  "house" "house" "house" "house" ...
 $ bedrooms    : int  4 3 3 4 3 4 3 4 3 3 ...
```

解读：

在数据框结构中，数据集包含29,580条记录和5个变量：`datesold`为字符型，记录销售日期和时间，需转换为日期格式以便时间序列分析；`postcode`为整数型，表示房产所在的邮政编码，范围从2600到2914，可用于地理分组分析；`price`为整数型，房价范围广泛，从56,500到8,000,000，存在显著的高价异常值，可能需要进一步检查和处理；`propertyType`为字符型，表示房产类型，建议转换为因子类型以便分类处理；`bedrooms`为整数型，卧室数量从0到5不等。

3. 数据摘要（`summary(data)` 的输出）

运行结果：

```
 datesold            postcode       price          propertyType
 bedrooms
 Length:29580       Min.   :2600   Min.   :  56500   Length:29580       Min.
:0.00
 Class :character   1st Qu.:2607   1st Qu.: 440000   Class :character   1st
Qu.:3.00
 Mode  :character   Median :2615   Median : 550000   Mode :character   Median
:3.00
                    Mean   :2730   Mean   : 609736                     Mean
:3.25
                    3rd Qu.:2905   3rd Qu.: 705000                     3rd
Qu.:4.00
                    Max.   :2914   Max.   :8000000                     Max.
:5.00
```

解读：

对于`datesold`（销售日期），数据类型为字符型，未显示进一步统计。

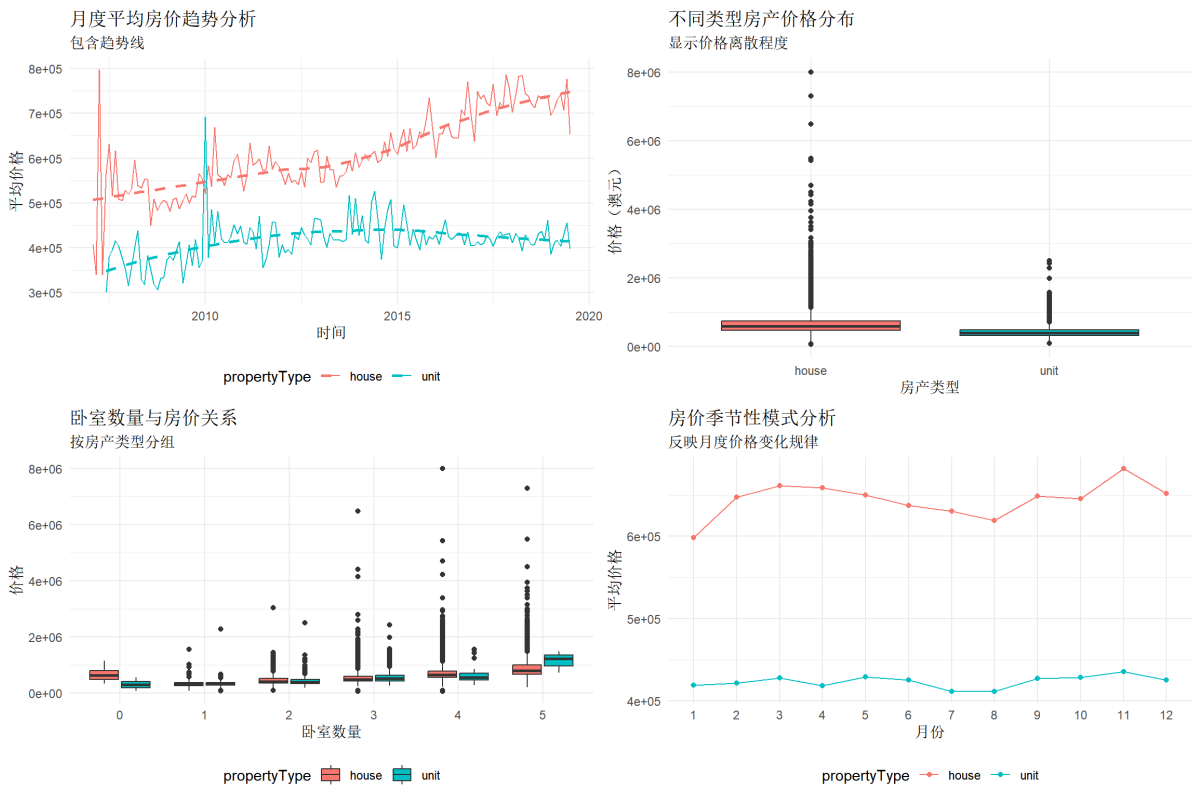
对于`postcode`（邮政编码），最小值为2600，第一四分位数为2607，中位数为2615，均值为2730，第三四分位数为2905，最大值为2914，显示邮政编码的分布较为均匀，涵盖了多个区域。

对于`price`（价格），最小值为56,500，第一四分位数为440,000，中位数为550,000，均值为609,736，第三四分位数为705,000，最大值为8,000,000，表明价格的分布非常分散，存在一些高价异常值。

对于`propertyType`（房产类型），数据类型为字符型，未显示进一步统计。

对于`bedrooms`（卧室数量），最小值为0，第一四分位数为3，中位数为3，均值为3.25，第三四分位数为4，最大值为5，表明大部分房产的卧室数量集中在3到4之间，且有部分房产的卧室数为0，可能需要进一步调查。

时间序列趋势可视化分析：



1. **月度平均房价趋势分析**：
   该图显示了从2007年到2019年不同房产类型（house和unit）的月度平均房价变化趋势。整体来看，house的房价呈现出逐年上升的趋势，同时波动较大，尤其在2008年和2016年有明显的起伏；而unit的房价则较为平稳，价格增长幅度相对较小。此图反映出house在市场中受到更大的波动影响，而unit表现更为稳定。
2. **不同类型房产价格分布**：
   该图展示了house和unit两种房产类型的价格分布情况。可以看到，house的价格中位数显著高于unit，但house的价格分布更广，具有较多的高价离群点，而unit的价格整体集中于较低的范围，离散性较小。箱线图说明了两种房产类型的市场特性：house的价格多样性更大，unit的市场价格相对统一。
3. **卧室数量与房价关系**：
   该图分析了卧室数量（bedrooms）与房价之间的关系，并按房产类型（house和unit）分组。可以看出，随着卧室数量的增加，房价总体呈现上升趋势，尤其是house类型的价格增幅更为显著。但在卧室数量增加到4或5时，价格的波动范围显著扩大，表明高卧室数量的房产价格具有更高的离散性和不确定性。
4. **房价季节性模式分析**：
   该图展示了不同月份房价的季节性变化特征。对于house类型，房价在3月和11月达到高峰，显示出一定的季节性波动规律，而unit类型的房价变化较为平稳，未表现出明显的季节性特征。这表明house可能受到市场供需变化的影响更大，而unit的需求相对均衡。

第3步：平稳性检验与纯随机性检验

AR/MA 模型要求时间序列是平稳的。可以使用 **ADF（Augmented Dickey-Fuller）检验** 来检查平稳性。检验结果如下：

```
=== 房产类型: house ===

```

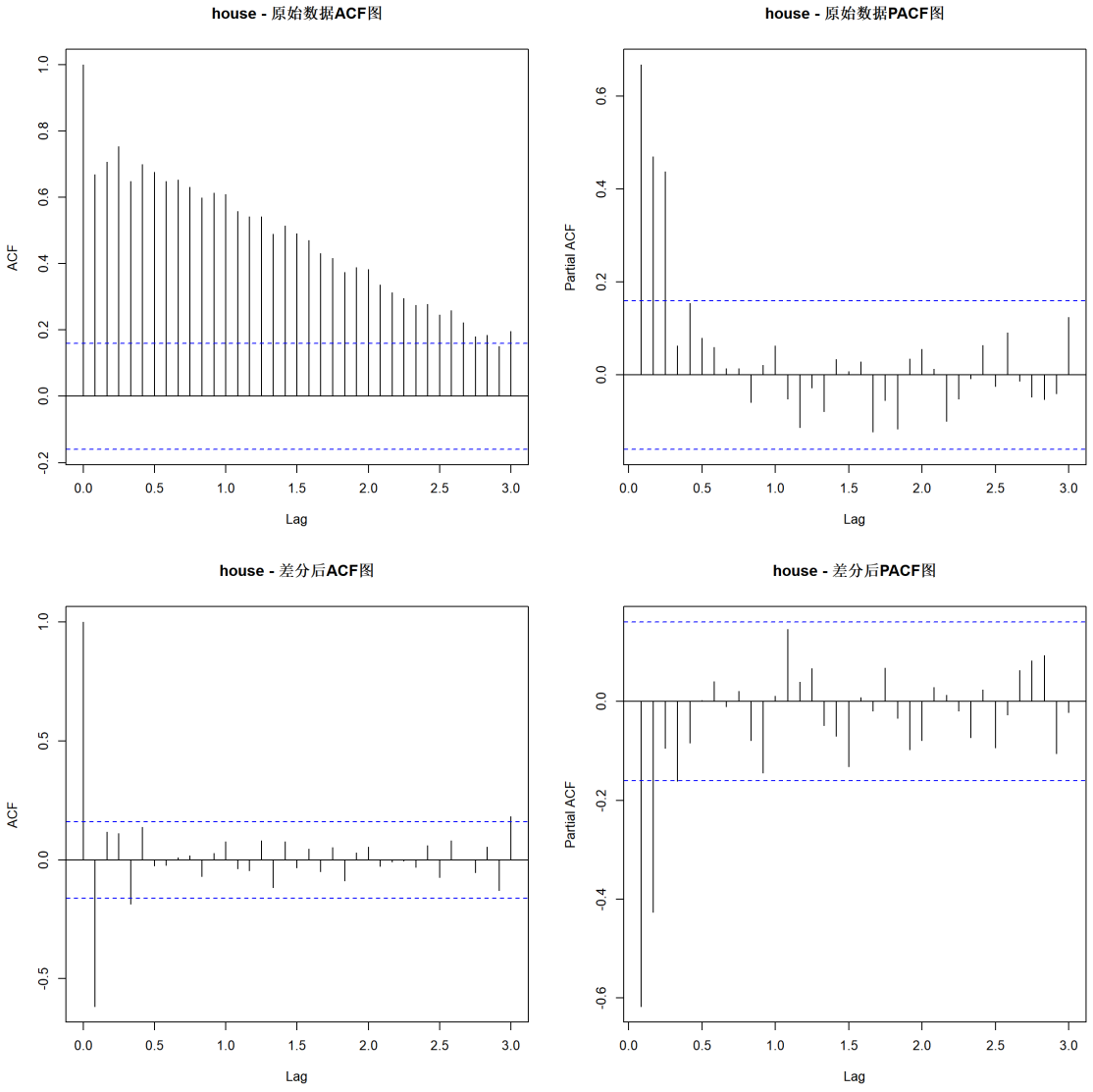
```
ADF检验结果:

    Augmented Dickey-Fuller Test

data:  ts_price
Dickey-Fuller = -2.6569, Lag order = 5, p-value = 0.3027
alternative hypothesis: stationary


时间序列不平稳，需要差分处理

一阶差分后的ADF检验结果:

    Augmented Dickey-Fuller Test

data:  ts_diff
Dickey-Fuller = -7.6141, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary


=== 房产类型: unit ===

ADF检验结果:

    Augmented Dickey-Fuller Test

data:  ts_price
Dickey-Fuller = -2.6001, Lag order = 5, p-value = 0.3266
alternative hypothesis: stationary


时间序列不平稳，需要差分处理

一阶差分后的ADF检验结果:

    Augmented Dickey-Fuller Test

data:  ts_diff
Dickey-Fuller = -6.7776, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

房产类型: `house`

1. **原序列 ADF 检验结果**:
   - **Dickey-Fuller = -2.6569**，**p-value = 0.3027**。
   - **解读**：p-value 大于 0.05，表明原始时间序列不是平稳的。这意味着 `house` 房产的月均价格随时间可能存在趋势或季节性变化，需要进一步处理。
2. **差分后的 ADF 检验结果**:
   - **Dickey-Fuller = -7.6141**，**p-value = 0.01**。

- **解读**：p-value 小于 0.05，表明差分后的时间序列是平稳的。通过一阶差分（`diff(ts_pt)`），成功消除了原序列中的非平稳性（例如趋势），使数据适合进一步的时间序列建模。

3. `house` **价格的ACF和PACF图**



在**原始数据部分**，滞后 1 的竖条显著高且超出蓝线，说明昨天的数据与今天的数据显著相关。ACF 图显示出自相关系数逐步缓慢衰减，表明时间序列存在趋势性或**非平稳性**。

PACF 图中滞后 1 处有显著的偏自相关系数，其后逐步减弱。这说明原始数据中滞后 1 是主要影响因素，可能适合低阶的 AR 模型（如 AR(1)），但由于趋势性，原始数据不够平稳，需要进一步差分处理。

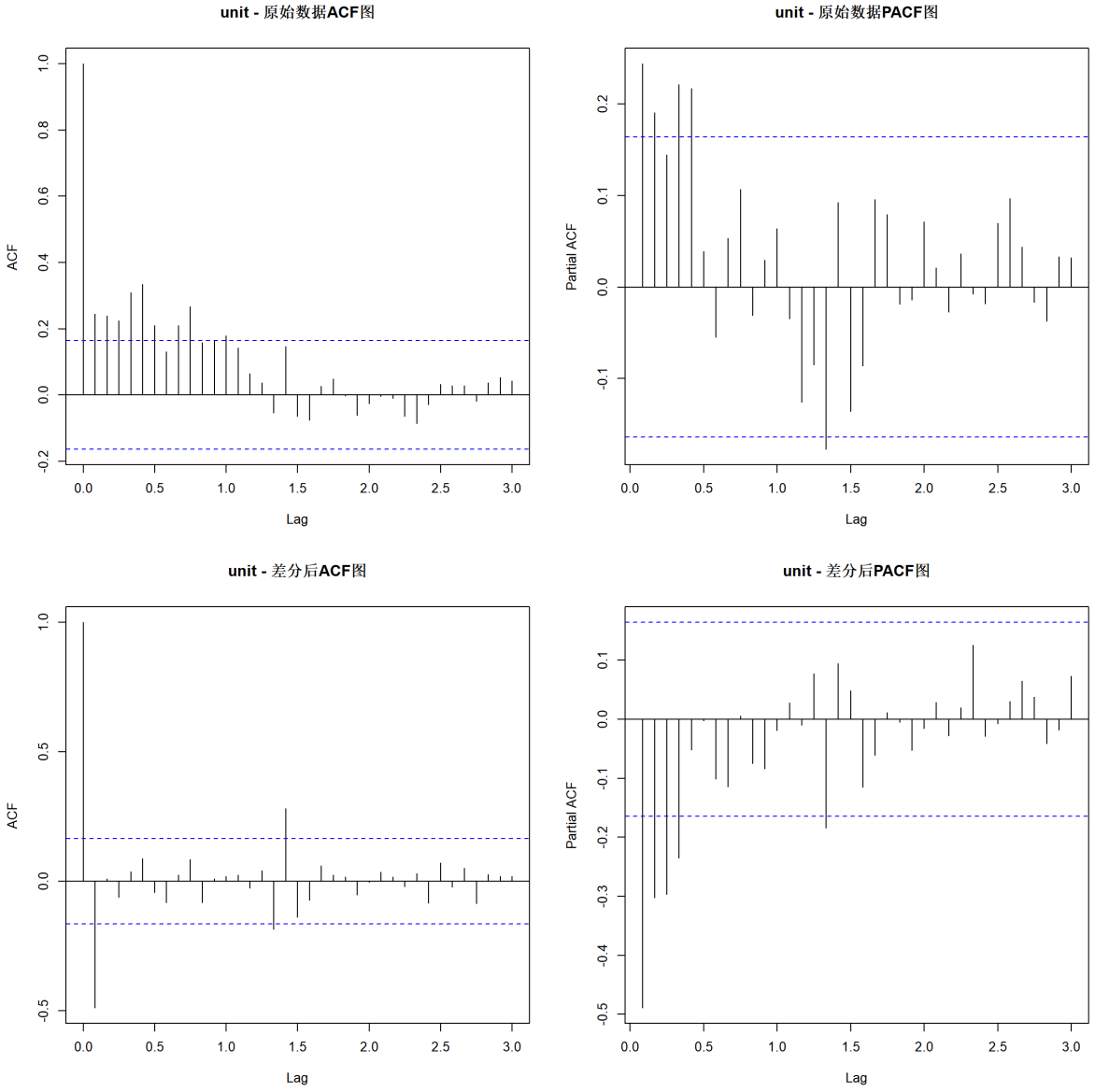
在**差分后的部分**，ACF 图快速衰减至接近零，表明差分后序列趋于平稳，移除了大部分趋势性或季节性影响。同时，PACF 图显示只有少数滞后点显著，滞后 1 处最为明显，进一步支持了差分后序列适合使用 AR(1) 模型进行建模。

房产类型: `unit`

1. **原序列 ADF 检验结果**：
   - **Dickey-Fuller = -2.6001，p-value = 0.3266。**
   - **解读**：与 `house` 类似，p-value 大于 0.05，表明 `unit` 房产的原始时间序列也不是平稳的。这可能是因为 `unit` 的价格数据同样受趋势或季节性变化影响。
2. **差分后的 ADF 检验结果**：

- **Dickey-Fuller = -6.7776，p-value = 0.01。**
- **解读**：差分后的时间序列的 p-value 小于 0.05，说明序列已经达到平稳状态。通过一阶差分去除了非平稳性，为后续建模奠定了基础。

3. `unit`**价格的ACF和PACF图**



在**原始数据部分**，滞后 1（Lag = 1）处的竖条较高，且超出蓝线，表明数据在滞后 1 上有显著相关性。ACF 图显示出自相关系数逐步缓慢衰减，表明时间序列存在趋势性或非平稳性。而 PACF 图中滞后 1 处有显著的偏自相关系数，其后逐步减弱，这说明原始数据的主要相关性集中在滞后 1，可能适合低阶的 AR 模型（如 AR(1)），但由于趋势性，原始数据不够平稳，需要进一步差分处理。

在**差分后的部分**，ACF 图中的自相关系数在滞后 1 之后迅速衰减至接近零，表明差分操作成功去除了序列中的趋势性，使时间序列趋于平稳。同时，PACF 图在滞后 1 处的偏自相关系数仍然显著，而其他滞后点接近零，表明差分后的时间序列适合使用 AR(1) 模型进行建模。

纯随机性检验

```
=== 房产类型: house 的纯随机性检验 ===

Box-Ljung检验结果:

    Box-Ljung test

```

```
data:  ts_price
X-squared = 829.11, df = 12, p-value < 2.2e-16


Box-Pierce检验结果）:

    Box-Pierce test

data:  ts_price
X-squared = 784.25, df = 12, p-value < 2.2e-16


=== 房产类型: unit 的纯随机性检验 ===

Box-Ljung检验结果:

    Box-Ljung test

data:  ts_price
X-squared = 95.675, df = 12, p-value = 3.886e-15


Box-Pierce检验结果）:

    Box-Pierce test

data:  ts_price
X-squared = 90.593, df = 12, p-value = 3.786e-14
```

对于房产类型 `house` 的时间序列，Box-Ljung 检验和 Box-Pierce 检验均显示其显著拒绝纯随机性假设（p-value < 2.2e-16）。Box-Ljung 检验的统计量为 X2=829.11X^2 = 829.11，而 Box-Pierce 检验为 X2=784.25X^2 = 784.25，这表明 `house` 的时间序列中存在显著的非随机性，可能包含趋势或季节性模式。由于 p-value 远小于常见显著性水平（如 0.05），可以确定该序列不是由纯随机过程生成的。

对于房产类型 `unit` 的时间序列，Box-Ljung 检验统计量为 X2=95.675X^2 = 95.675，Box-Pierce 检验统计量为 X2=90.593X^2 = 90.593，对应的 p-value 分别为 3.886e-15 和 3.786e-14，均显著小于 0.05。这同样表明 `unit` 的时间序列拒绝纯随机性假设，但其非随机性程度不如 `house` 显著。从检验结果看，`unit` 的序列可能存在较弱的趋势或其他结构性特征，但依然需要进一步建模来捕捉这些规律。

第4步：数据集划分

将数据按照时间范围拆分为训练集和测试集，以便后续模型的训练和评估。其中：

- **训练集**：数据时间范围从序列开始至 **2018年7月**。
- **测试集**：数据时间范围为 **2018年8月** 至 **2019年7月**。

通过为每种房产类型（`propertyType`）标记数据集所属类别（训练或测试），可以在不混淆数据的情况下对模型进行评估。

第5步：建立ARIMA模型

house 类型的ARIMA模型

```
=== 建立 house 的ARIMA模型 ===

模型摘要:
Series: ts_train
ARIMA(2,1,0)(1,0,2)[12]

Coefficients:
          ar1      ar2     sar1     sma1     sma2
      -1.0570  -0.6796  0.8720  -0.8102  0.1873
s.e.   0.0792   0.0924  0.0873   0.1538  0.1495

sigma^2 = 2.052e+09:  log likelihood = -1664.3
AIC=3340.61   AICc=3341.26   BIC=3358.13

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE
ACF1
Training set 3455.585 44307.66 31066.72 0.2011234 5.210925 0.6941345
0.07406755

    Ljung-Box test

data:  Residuals from ARIMA(2,1,0)(1,0,2)[12]
Q* = 24.127, df = 19, p-value = 0.1913

Model df: 5.   Total lags used: 24


预测误差指标:
MAE: 39867.65
RMSE: 48469.68
MAPE: 5.644582 %
```

该模型是为 `house` 类型房产时间序列建立的 ARIMA 模型，最终选择的结构是 ARIMA(2,1,0)(1,0,2)[12]，即包含 2 阶自回归项、1 阶差分以及季节性成分。模型的系数估计（如 ar1、ar2、sar1 等）均具有较小的标准误差，表明模型参数估计较为可靠。模型的 AIC 为 3340.61，BIC 为 3358.13，表明模型拟合的优度和复杂度在较低范围内平衡。训练集的误差指标（如 RMSE=44307.66，MAE=31066.72）显示模型在训练数据上的拟合较好，且 Ljung-Box 检验的 p-value 为 0.1913，大于显著性水平 0.05，表明模型的残差为白噪声，没有显著的自相关性。

在预测误差方面，模型的 MAE 为 39867.65，RMSE 为 48469.68，MAPE 为 5.64%，表明模型在测试集上的预测性能较为准确。尤其是 MAPE 小于 6%，显示预测值与实际值的相对误差较小，总体而言，该模型可以较好地捕捉 `house` 类型房产的时间序列特征并进行有效预测。

unit 类型的ARIMA模型

```
=== 建立 unit 的ARIMA模型 ===

模型摘要:
Series: ts_train
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.8481
s.e.   0.0473

sigma^2 = 1.922e+09:  log likelihood = -1574.09
AIC=3152.19   AICc=3152.28   BIC=3157.92

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE
ACF1
Training set 3432.629 43510.21 28095.86 0.0398565 6.600268 0.6718125
-0.0266724

    Ljung-Box test

data:  Residuals from ARIMA(0,1,1)
Q* = 21.7, df = 23, p-value = 0.5384

Model df: 1.   Total lags used: 24


预测误差指标:
MAE: 17474.35
RMSE: 21952.41
MAPE: 4.09615 %
```

该模型是为 `unit` 类型房产时间序列建立的 ARIMA(0,1,1) 模型，表示该序列经过一阶差分后，使用一个移动平均项（MA1）进行建模。模型的 MA1 系数为 -0.8481，标准误差较小（0.0473），表明参数估计具有较高的准确性。模型的 AIC 值为 3152.19，BIC 为 3157.92，显示模型相对简单且拟合较好。训练集的误差指标（如 RMSE = 43510.21，MAE = 28095.86）显示模型在训练数据上的表现较稳定，且 Ljung-Box 检验的 p-value 为 0.5384，大于显著性水平 0.05，说明残差为白噪声，没有显著的自相关性。

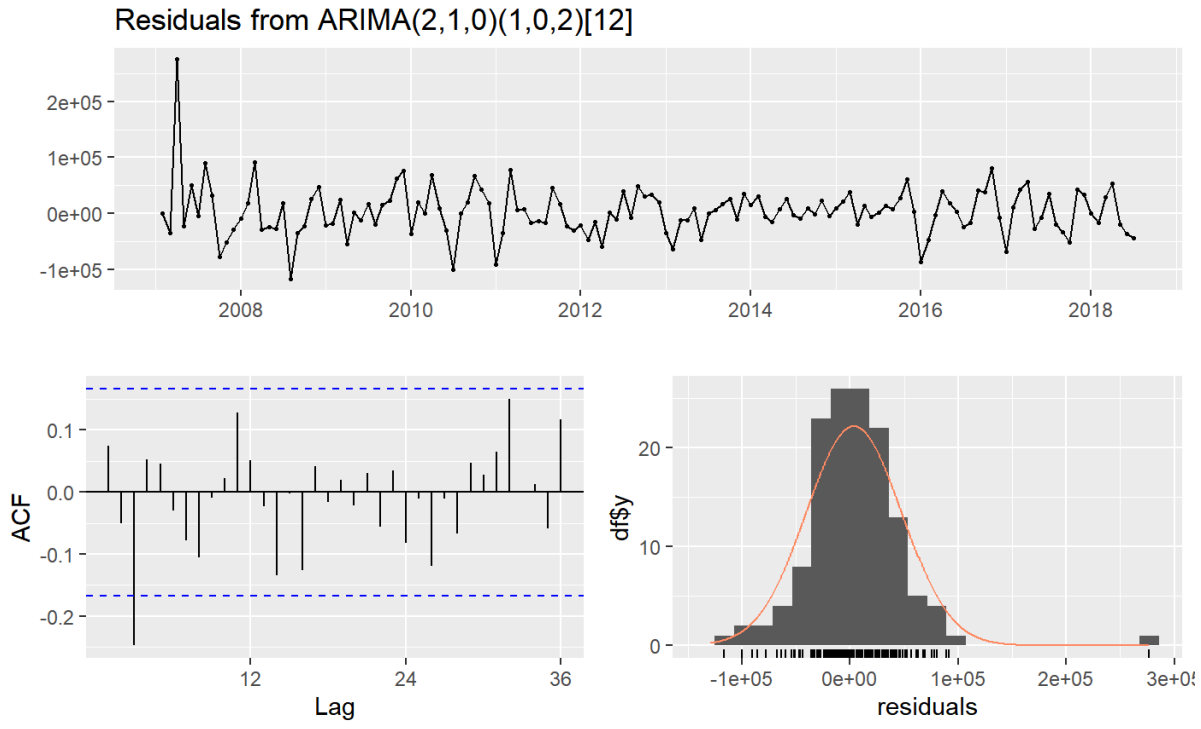
在预测方面，模型的误差指标表现较好，MAE 为 17474.35，RMSE 为 21952.41，MAPE 为 4.10%，表明模型在测试集上的预测性能较为精准。特别是 MAPE 小于 5%，说明预测值与实际值的相对误差较小。总体来看，该模型能够有效捕捉 `unit` 类型房产时间序列的特性，具有良好的预测能力，同时保持模型的简单性和稳定性。

第6步：预测结果与模型评估

house 类型的预测结果

	A	B	C	D	E
1	Month	Actual	Predicted	Lower_95	Upper_95
2	2018/8/1	713360.85	737179.1	648385.93	825972.26
3	2018/9/1	739834.32	757097.74	668160.48	846035.01
4	2018/10/1	732847.11	748444.89	653300.72	843589.06
5	2018/11/1	733213.02	794711.06	684053.57	905368.55
6	2018/12/1	739387.21	758706.46	647881.04	869531.87
7	2019/1/1	695679.17	721139.06	601917.35	840360.78
8	2019/2/1	711578	768111.96	642921.52	893302.39
9	2019/3/1	729635.59	767448.29	640890.2	894006.38
10	2019/4/1	743605.69	767037.8	633301.8	900773.81
11	2019/5/1	707236.01	768911.16	631742.33	906080
12	2019/6/1	777167.53	754181.95	614345.03	894018.87
13	2019/7/1	652604.88	765620.5	620356.77	910884.23
14					

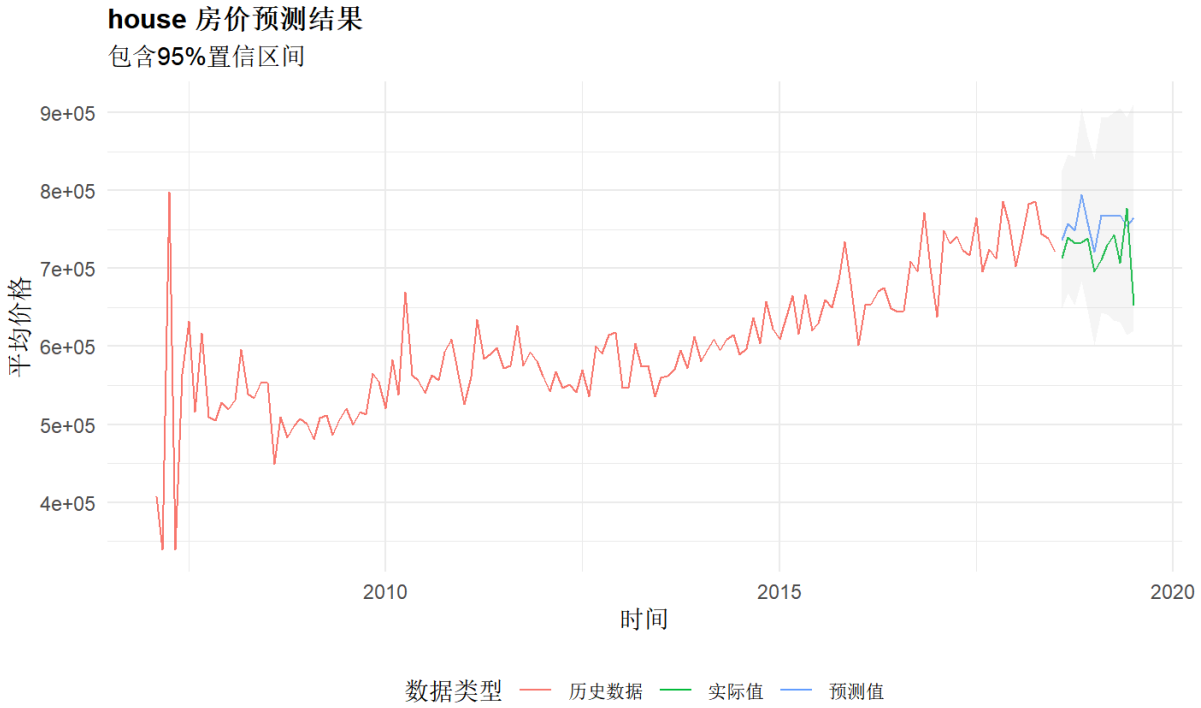
Residuals from ARIMA(2,1,0)(1,0,2)[12]



1. **残差时间序列图**：图中显示了模型残差随时间的变化情况。残差值在零附近随机波动，没有显著的趋势或周期性，表明模型捕捉到了时间序列中的主要特征。虽然在部分时间点（例如 2010 年）出现了较大的波动，但整体残差表现相对稳定。
2. **自相关函数（ACF）**：残差的自相关系数在所有滞后下基本落在 95% 置信区间（蓝色虚线）范围内，表明残差不存在显著的自相关性。结合 Ljung-Box 检验结果，说明模型的残差接近白噪声，验证了模型的合理性。

3. **残差分布直方图**：残差直方图显示残差的分布接近对称的正态分布，红色曲线为拟合的正态分布曲线。虽然尾部有少量偏离，但总体来看残差的分布与正态分布较为一致，这支持模型的假设，即误差项为正态分布。

house房价预测与实际值对比

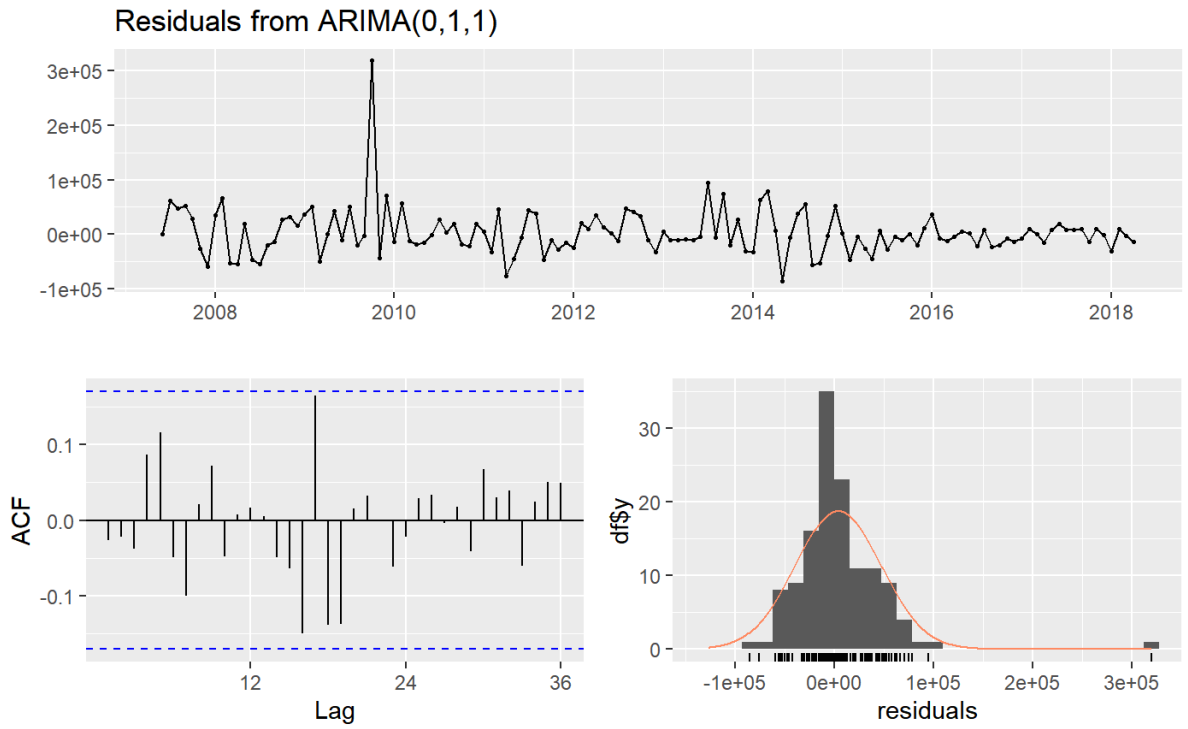


图中显示了 house 类型房产的历史平均价格（红色曲线）、预测值（蓝色曲线）和实际测试集值（绿色曲线），并包含了预测的 95% 置信区间（灰色阴影部分）。可以看到，历史数据呈现出逐步上升的趋势，并伴有一定的波动。预测值与实际测试值的变化趋势基本一致，且大多数测试点均落在预测值的置信区间内，表明模型对于 house 类型房产的短期价格预测具有较好的准确性

unit类型的预测结果

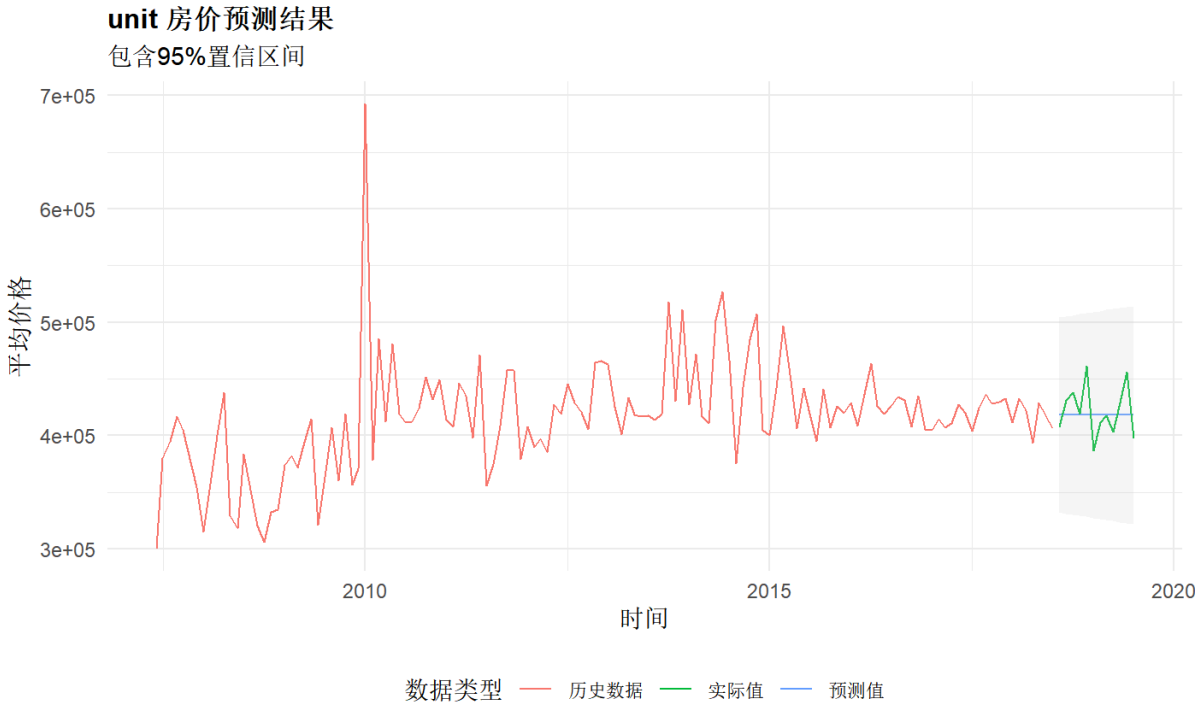
	A	B	C	D	E	F
1	Month	Actual	Predicted	Lower_95	Upper_95	
2	2018/8/1	407890.23	418056.02	332119.05	503992.99	
3	2018/9/1	431110.36	418056.02	331132.67	504979.38	
4	2018/10/1	437801.99	418056.02	330157.35	505954.69	
5	2018/11/1	418685	418056.02	329192.73	506919.31	
6	2018/12/1	461145.1	418056.02	328238.48	507873.56	
7	2019/1/1	386341.46	418056.02	327294.26	508817.79	
8	2019/2/1	411525.42	418056.02	326359.76	509752.29	
9	2019/3/1	417760.17	418056.02	325434.68	510677.36	
10	2019/4/1	403325.91	418056.02	324518.76	511593.28	
11	2019/5/1	429084.62	418056.02	323611.72	512500.32	
12	2019/6/1	456118.97	418056.02	322713.31	513398.73	
13	2019/7/1	397410.68	418056.02	321823.28	514288.76	
14						

Residuals from ARIMA(0,1,1)



1. **残差时间序列图**：该图显示了残差值在时间上的变化。大部分残差波动集中在零附近，且没有明显的趋势或周期性，说明模型对数据的拟合较好。然而，在某些时间点（例如 209 年）存在显著的异常波动，可能是由于极端事件或噪声的影响。
2. **自相关函数（ACF）**：ACF 图用于检测残差序列中的自相关性。大多数滞后点的自相关系数都落在 95% 置信区间（蓝色虚线）范围内，表明残差基本没有显著的自相关性，进一步支持了模型的白噪声假设。
3. **残差分布直方图**：该图显示了残差的分布情况，并叠加了一条拟合的正态分布曲线。直方图的形状接近对称的钟形，表明残差符合正态分布假设。尽管尾部略有偏离，但整体分布特征良好。

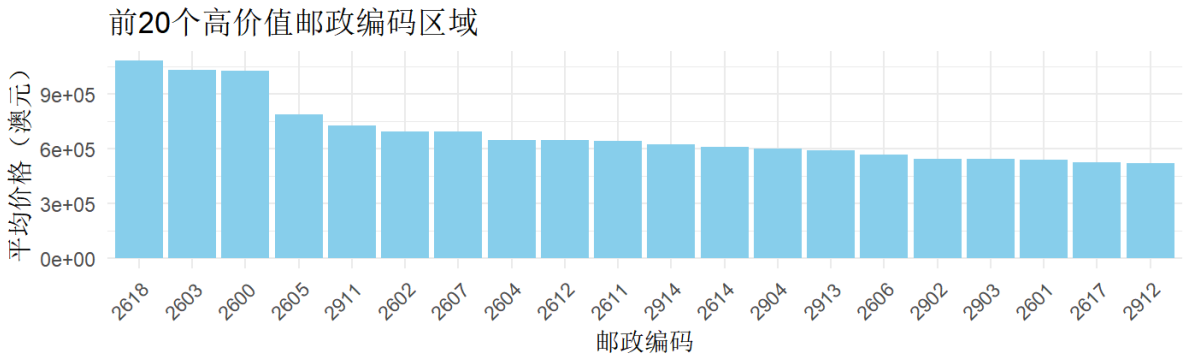
unit房价预测与实际值对比

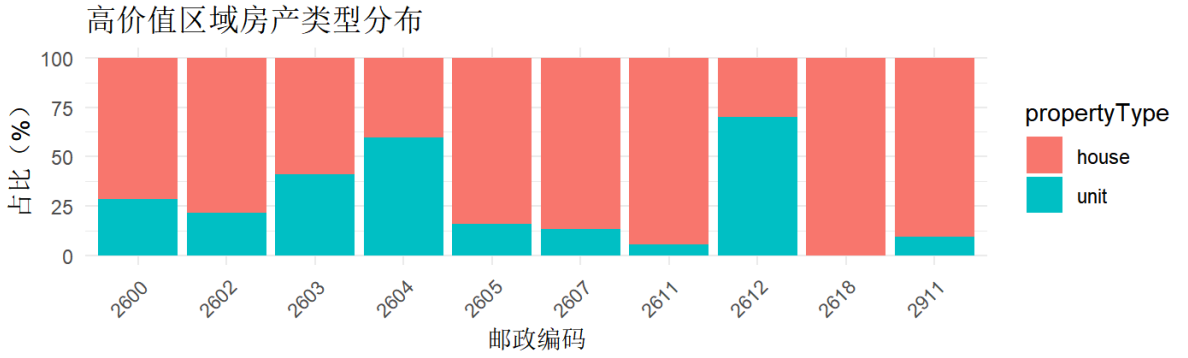


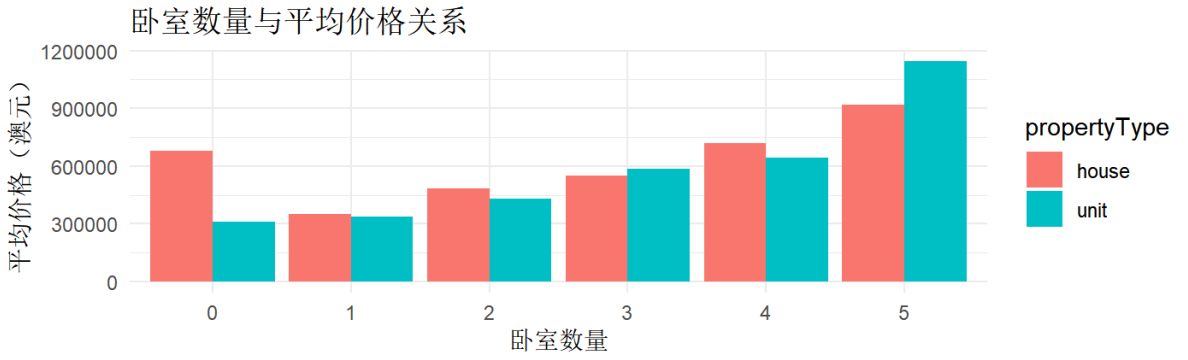
图中展示了 unit 类型房产的历史平均价格（红色曲线）、预测值（蓝色曲线）和实际测试集值（绿色曲线），同样包含 95% 的置信区间（灰色阴影部分）。历史数据相比 house 类型波动更大，且在整体水平上较低。预测值与实际测试值的变化趋势接近，且多数测试值落在预测的置信区间范围内，说明模型在 unit 类型房产价格预测中表现稳定，能够合理反映市场趋势和价格变化。

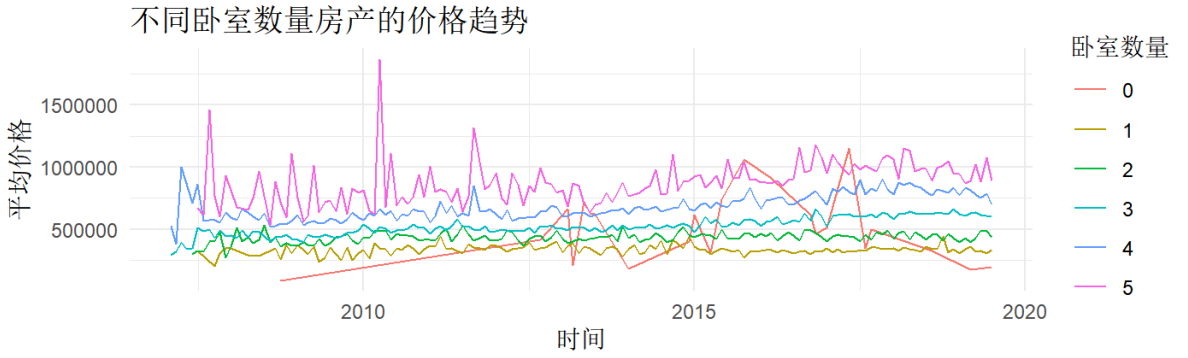
第7步：提出建议

7.1 邮政编码与卧室数量影响分析









7.2 提出建议

1. **House 聚焦高端市场与长期自住需求**：House 类型房产价格普遍较高，适合定位于高端市场或满足长期自住需求的客户。房产中介可以将 House 的核心优势（如更大的居住面积、更好的私密性、独立庭院等）作为宣传重点，吸引注重生活品质的家庭购房者。同时，对于高价值区域的 House，可以推荐作为长期投资的优质选择，因其增值潜力较大。
2. **Unit 主攻低价投资市场**：Unit 类型房产价格相对较低，适合年轻人、小型家庭或投资者购买。房产中介可以将 Unit 的便利性（如靠近交通枢纽、商业中心）、较低的维护成本，以及稳定的租赁需求作为卖点吸引客户。此外，购房者如果关注租金回报率，可以重点考虑热门区域的 Unit 作为投资组合的一部分。
3. **根据区域差异推荐房型**：在郊区或地广人稀的区域，House 往往更符合市场需求，因其提供更多居住空间和更好的生活品质；而在市区或商业密集区，Unit 更受欢迎，因其价格低且交通便利。因此，房产中介在推广时应根据区域特点进行精准推荐，例如在市区优先推广 Unit，而在郊区重点宣传 House。
4. **针对高价值区域的投资机会**：从邮政编码的高价值区域分析中可以看出，房价较高的区域集中在特定邮政编码，例如`2618`、`2603`等。这些区域可能由于良好的地理位置或高需求而保持较高的房价水平。因此，房产中介可以优先在这些区域集中资源，例如推广高端房产项目或吸引高净值客户群体。购房者则需要根据自身预算，考虑是否在这些区域进行投资，特别是在这些区域房价稳定增长的情况下。
5. **房型与卧室数量的选择**：对于卧室数量的分析表明，卧室较多（如4-5间）的房产通常价格较高，但在某些区域或房型中，其价格涨幅可能趋于平缓。如果购房者注重性价比，选择卧室数量适中的房型（如3间）可能会是更好的选择。而房产中介可以通过分析区域内需求，推荐符合购房者预算和需求的房型，优化房产匹配。

第8步：模型可能存在的不足与给出优化建议

模型可能存在的不足

1. **外部影响因素未纳入建模**：目前模型仅基于时间序列数据构建，未考虑影响房价的关键外部因素（如政策变动、利率、经济增长、人口迁移等），这可能导致模型在面对实际市场波动时失去鲁棒性。
2. **过于依赖线性假设**：ARIMA模型假设时间序列具有线性特性，而实际房价可能存在复杂的非线性规律，例如周期性变化、突发事件等，这些复杂模式未能被捕捉。
3. **季节性特征捕捉不充分**：尽管使用了SARIMA模型处理季节性，但从结果来看，季节性波动未被完全捕捉，特别是在某些特定月份的价格波动上，预测误差较大。
4. **模型对区域和房型特性的细化不足**：不同邮政编码和房型对房价的影响存在差异，但目前的模型未对这些特性进行充分的分组建模，仅进行了全局性分析，可能导致预测结果缺乏针对性。
5. **训练数据和测试数据分布存在偏差**：训练数据和测试数据可能在时间上具有不同的特性（如政策或市场环境的变化），可能影响模型在测试数据上的预测性能。

优化建议

1. **引入外部影响因素**：在现有模型基础上，结合外部变量（如利率、通胀率、人口流入流出、基础设施建设等），构建更复杂的ARIMAX模型或基于机器学习的方法进行建模，提高预测的解释力和准确性。
2. **采用非线性建模方法**：引入非线性时间序列模型（如Prophet、LSTM）或混合模型（如将ARIMA与深度学习结合），以捕捉房价的非线性变化规律和复杂的动态特性。

3. **强化季节性分析**：通过调整SARIMA模型的季节性参数，进一步优化季节性部分的拟合。此外，采用傅里叶变换或小波分析等方法对时间序列中的季节性特征进行更细致的处理。
4. **细分建模**：对不同邮政编码、房型和卧室数量的组合进行分组建模。基于区域特性（如高房价区域和低房价区域）分别构建模型，这样可以提高模型的适用性和精确度。
5. **集成多模型方法**：通过结合不同模型（如ARIMA与ETS、随机森林或支持向量机）的预测结果，采用集成学习的方式提升模型的稳定性和整体预测性能。
6. **改进数据分割策略**：通过时间序列交叉验证（Time Series Cross-Validation）技术，进一步优化训练集和测试集的划分方式，从而减少因数据分布差异导致的性能偏差。