

# 不動產價格因子分析



# 目錄

**01**

目的和資料來源

資料視覺化

**02**

**03**

模型建置

優化模型

**04**

**05**

分析結果

學習心得

**06**

# 目錄

**01**

目的和資  
料來源

資料視覺化

**02**

**03**

模型建置

優化模型

**04**

**05**

分析結果

學習心得

**06**

## 研究目的

- 探討**不動產因子**對於**價格**的影響力
- 透過手邊資料探索可能**原因**
- 嘗試找出能夠用**不動產因子**預測價格的**模型**

- **二手房地產交易資訊：**

- 內政部不動產實價登入查詢: <https://lvr.land.moi.gov.tw/>

- **座標位置資訊：**

- 台灣電子地圖服務網: <https://www.map.com.tw/>

- **其他欄位蒐集:**

- Shopping mall 資訊: TripAdvisor (<https://www.tripadvisor.com.tw/>)
- Tapei School: [臺北市府教育局-相關連結-所屬學校 \(gov.taipei\)](#)
- MRT: [臺北大眾捷運股份有限公司 \(metro.taipei\)](#)
- 醫院: Google 搜尋
- Train: [交通部臺灣鐵路管理局 \(railway.gov.tw\)](#)

## • 股市：

- 台灣證交所：

<https://www.twse.com.tw/zh/>

## • 外匯：

- 台灣期交所：

<https://www.taifex.com.tw/cht/index>

- 每日外幣參考匯率查詢

# 目錄

**01**

目的和資料來源

資料視覺化

**02**

**03**

模型建置

**05**

分析結果

優化模型

**04**

學習心得

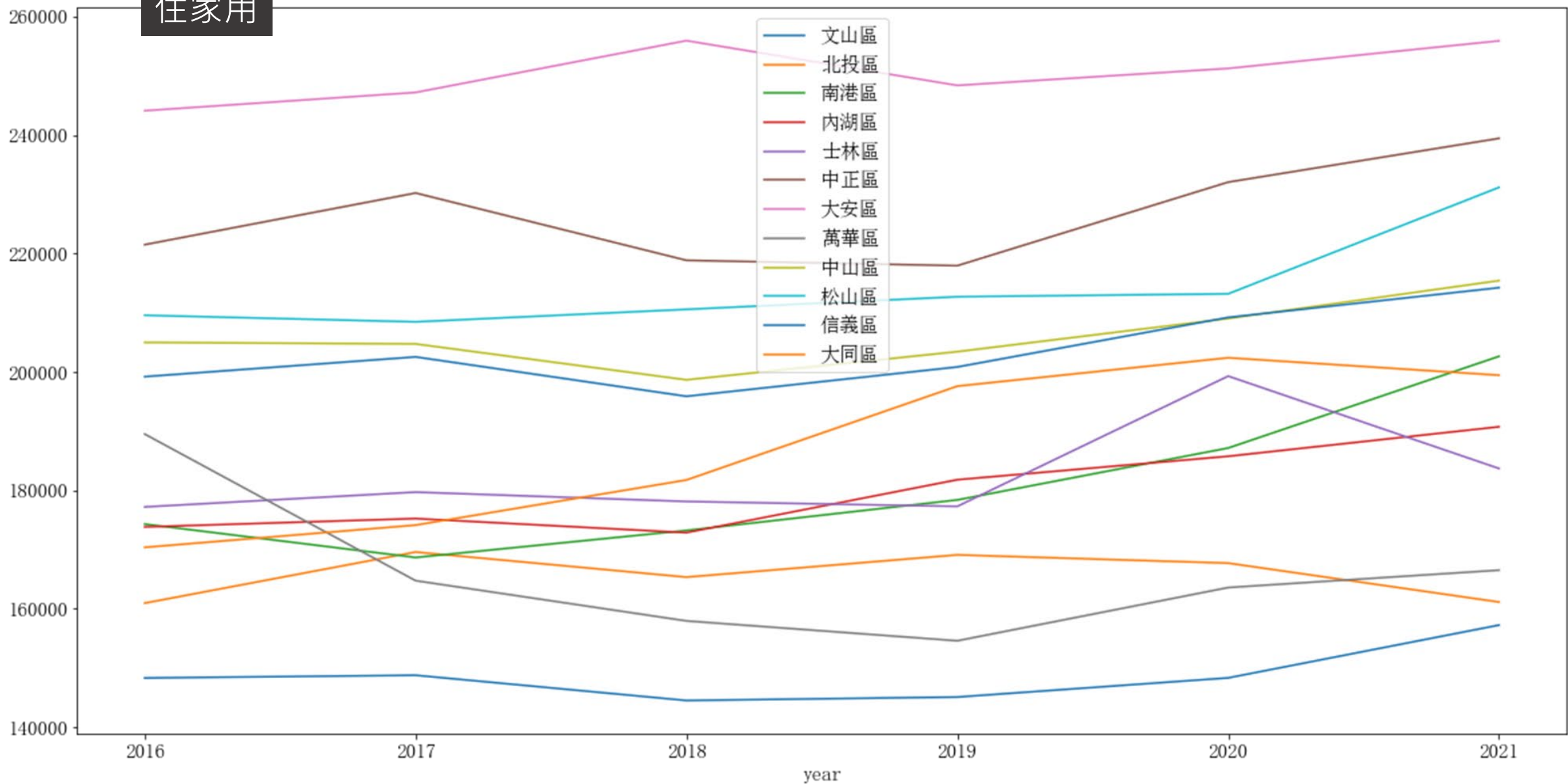
**06**



不同用途不同地段的價格差異為何？

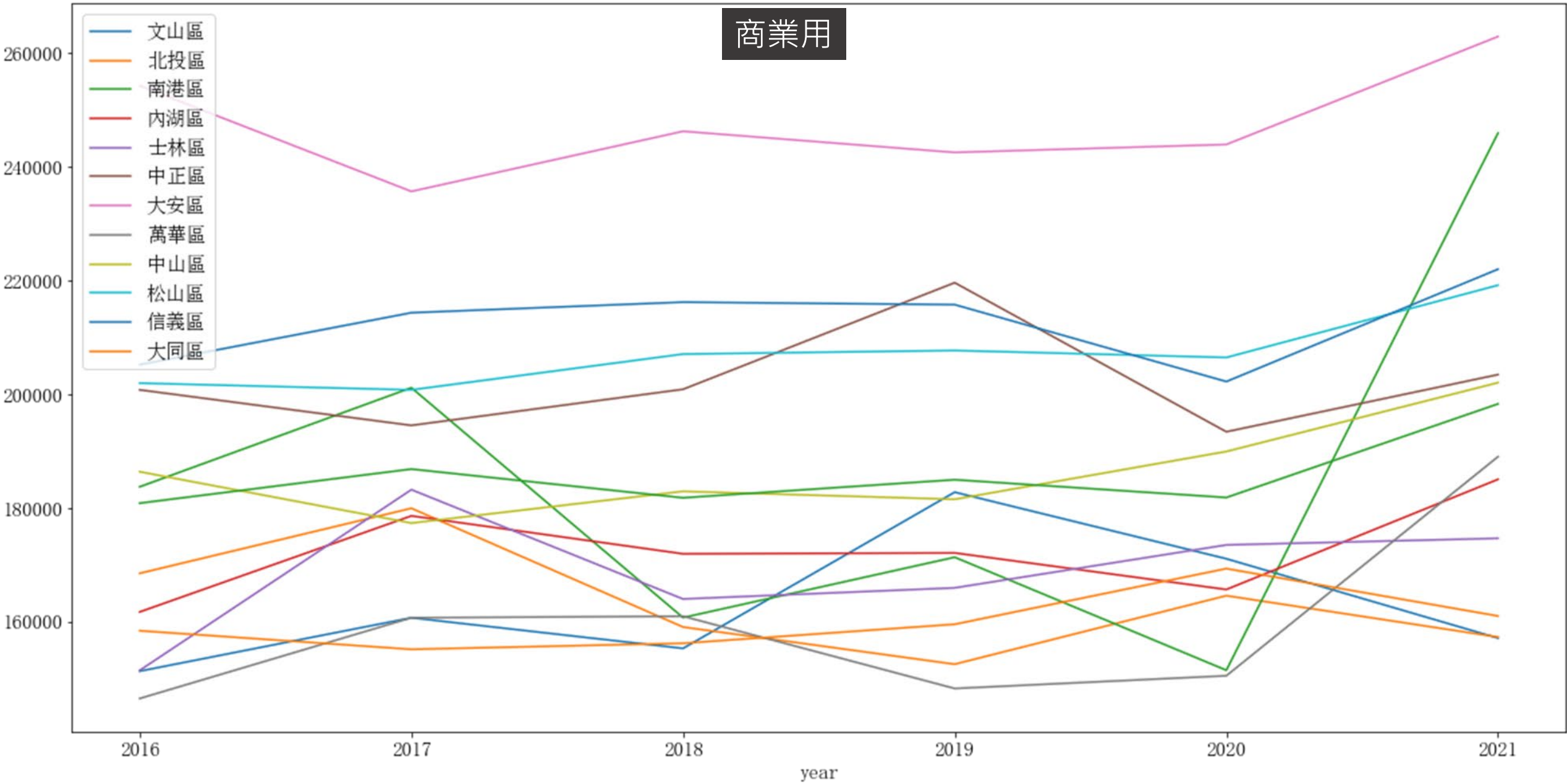


住家用



| Year                                                         | 大安區    | 中山區    | 南港區    | 松山區    | 萬華區    | 大同區    | 北投區    | 士林區    | 中正區    | 內湖區    | 文山區    | 信義區    |
|--------------------------------------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 2016                                                         | 244126 | 204981 | 174292 | 209565 | 189492 | 170359 | 160939 | 177194 | 221496 | 173814 | 148296 | 199199 |
| 2017                                                         | 247211 | 204727 | 168648 | 208457 | 164732 | 174118 | 169572 | 179690 | 230222 | 175224 | 148755 | 202534 |
| 2018                                                         | 255960 | 198657 | 173218 | 210563 | 157933 | 181743 | 165329 | 178113 | 218847 | 172847 | 144493 | 195876 |
| 2019                                                         | 248395 | 203420 | 178375 | 212698 | 154577 | 197592 | 169090 | 177286 | 217947 | 181785 | 145074 | 200835 |
| 2020                                                         | 251265 | 209017 | 187141 | 213189 | 163571 | 202388 | 167697 | 199291 | 232057 | 185745 | 148308 | 209205 |
| 2021                                                         | 255925 | 215423 | 202632 | 231171 | 166503 | 199441 | 161124 | 183681 | 239451 | 190726 | 157228 | 214227 |
| <div>1 大安區 2 中正區 3 南港區</div> <div>成長最快 衰退 衰退 倒數第二 最後一名</div> |        |        |        |        |        |        |        |        |        |        |        |        |

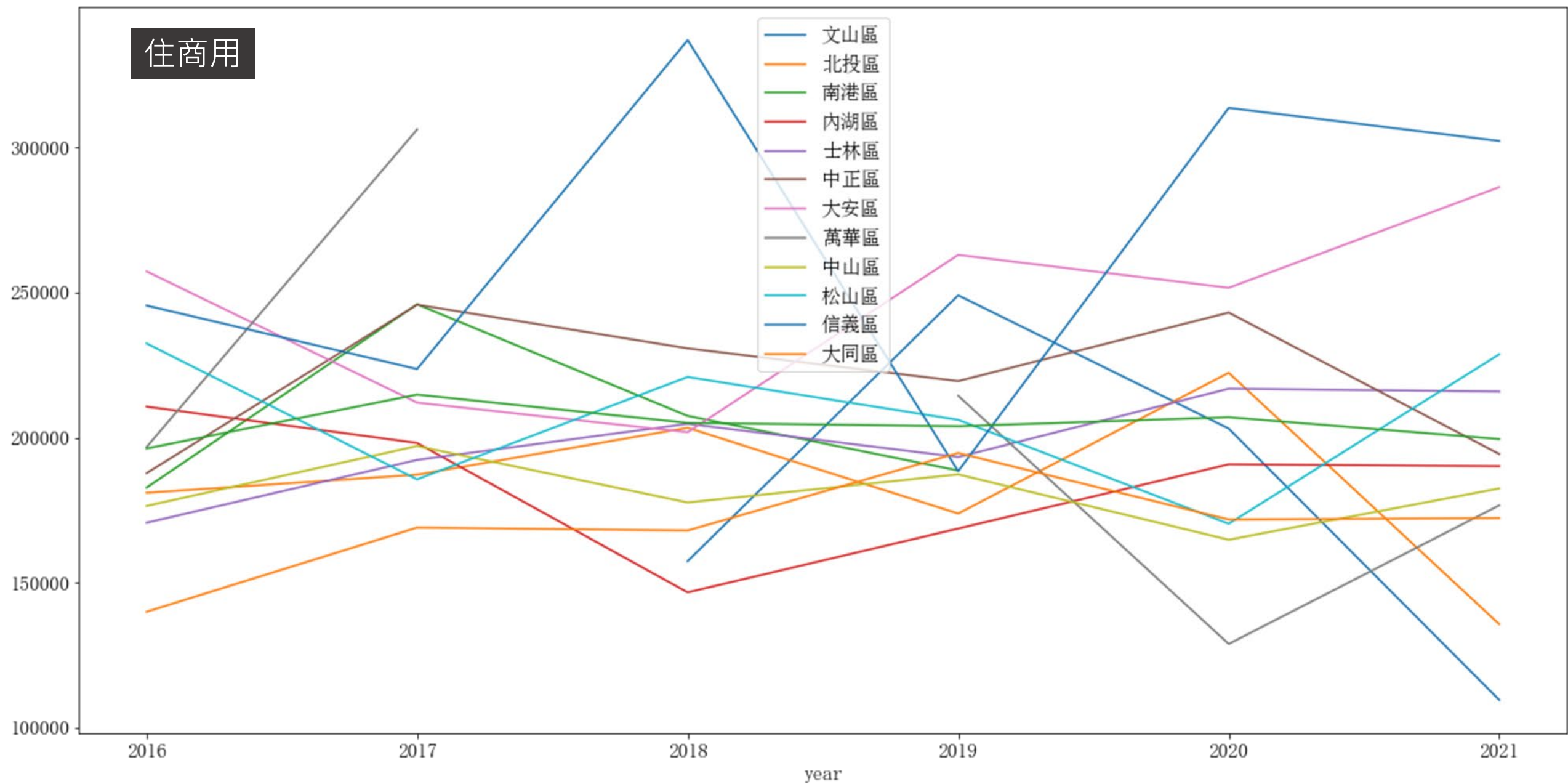
商業用





住商用

- 文山區
- 北投區
- 南港區
- 內湖區
- 士林區
- 中正區
- 大安區
- 萬華區
- 中山區
- 松山區
- 信義區
- 大同區



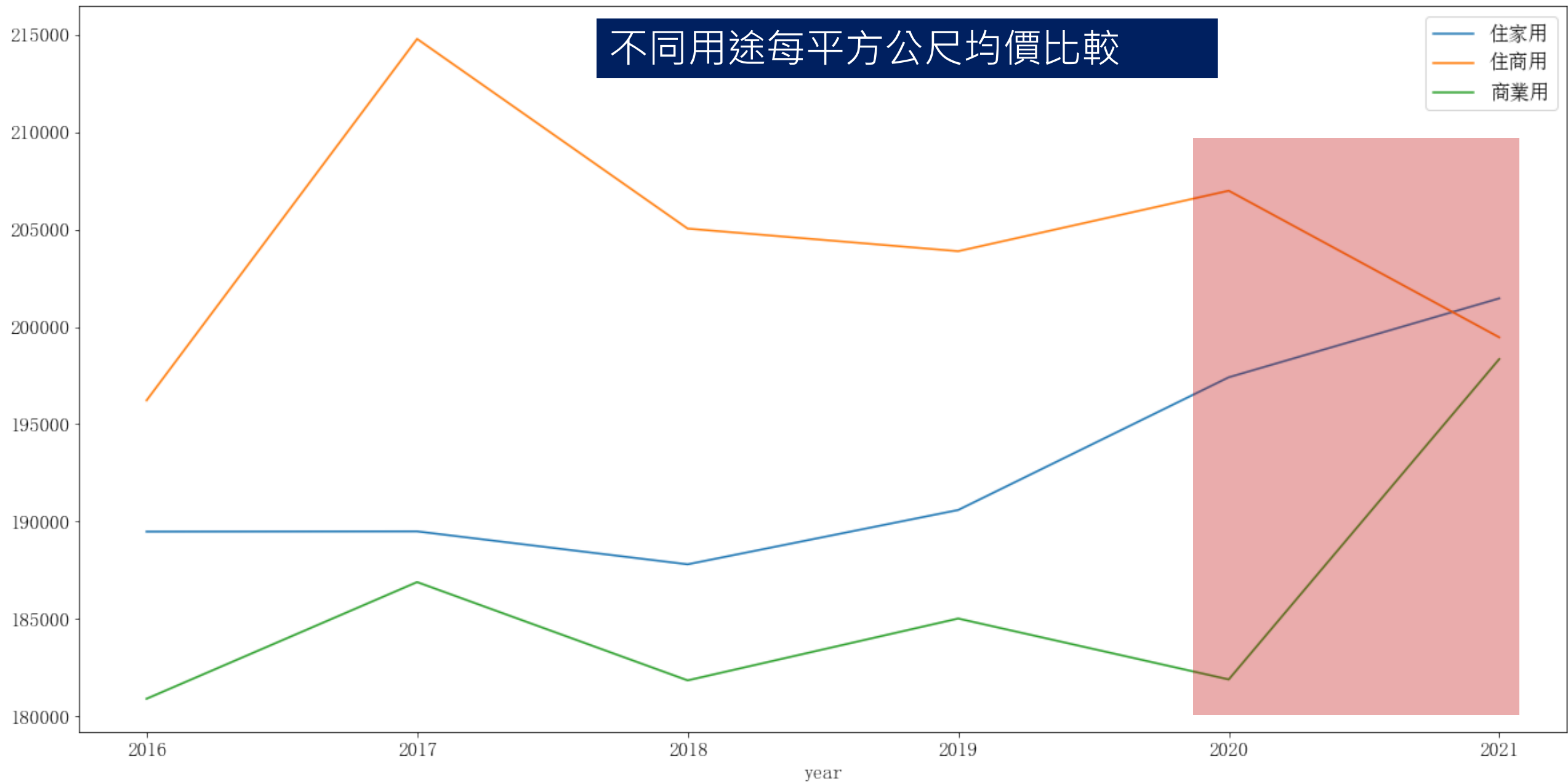


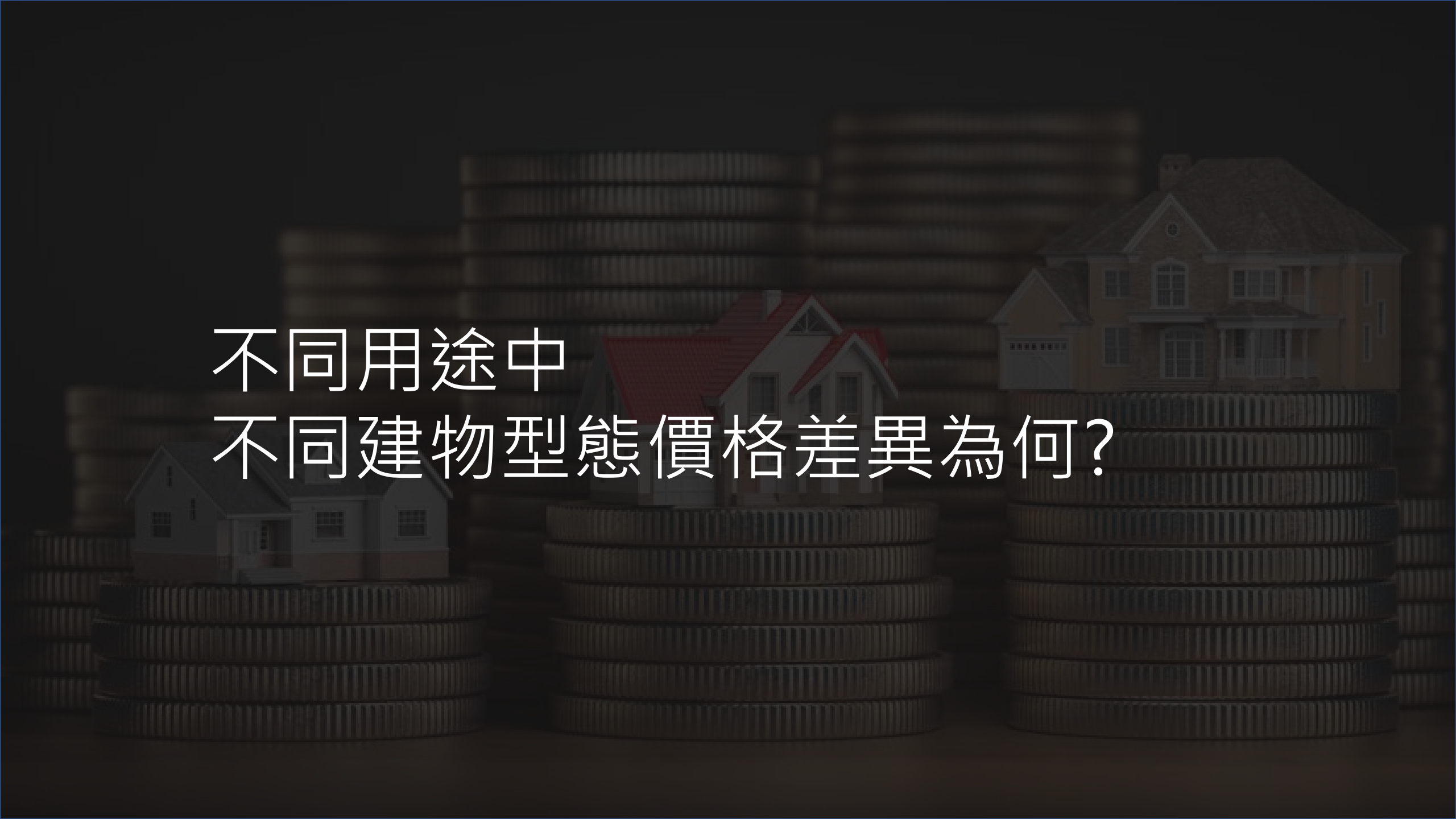
| Year | 大安區    | 中山區    | 南港區    | 松山區    | 萬華區    | 大同區    | 北投區    | 士林區    | 中正區    | 內湖區    | 文山區    | 信義區    |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 2016 | 257272 | 176426 | 182745 | 232442 | 196714 | 139953 | 180952 | 170631 | 187714 | 210648 | 173666 | 245479 |
| 2017 | 212031 | 197066 | 245902 | 185590 | 306168 | 168954 | 187197 | 192288 | 245768 | 198113 |        | 223614 |
| 2018 | 201818 | 177610 | 207451 | 220872 |        | 167963 | 203346 | 204704 | 230740 | 146635 | 157406 | 336960 |
| 2019 | 262960 | 187303 | 188625 | 206095 | 214399 | 194669 | 173795 | 193283 | 219455 | 168621 | 249011 | 188395 |
| 2020 | 251599 | 164770 |        | 170271 | 128854 | 171756 | 222318 | 216823 | 243055 | 190765 | 203082 | 313605 |
| 2021 | 286304 | 182456 |        | 228705 | 176633 | 172231 | 135647 | 215873 | 194308 | 190141 | 109541 | 302202 |
| 倒數第二 |        |        |        |        |        |        | 成長最快   |        | 衰退     |        | 最後一名   |        |



# 不同用途每平方公尺均價比較

- 住家用
- 住商用
- 商業用

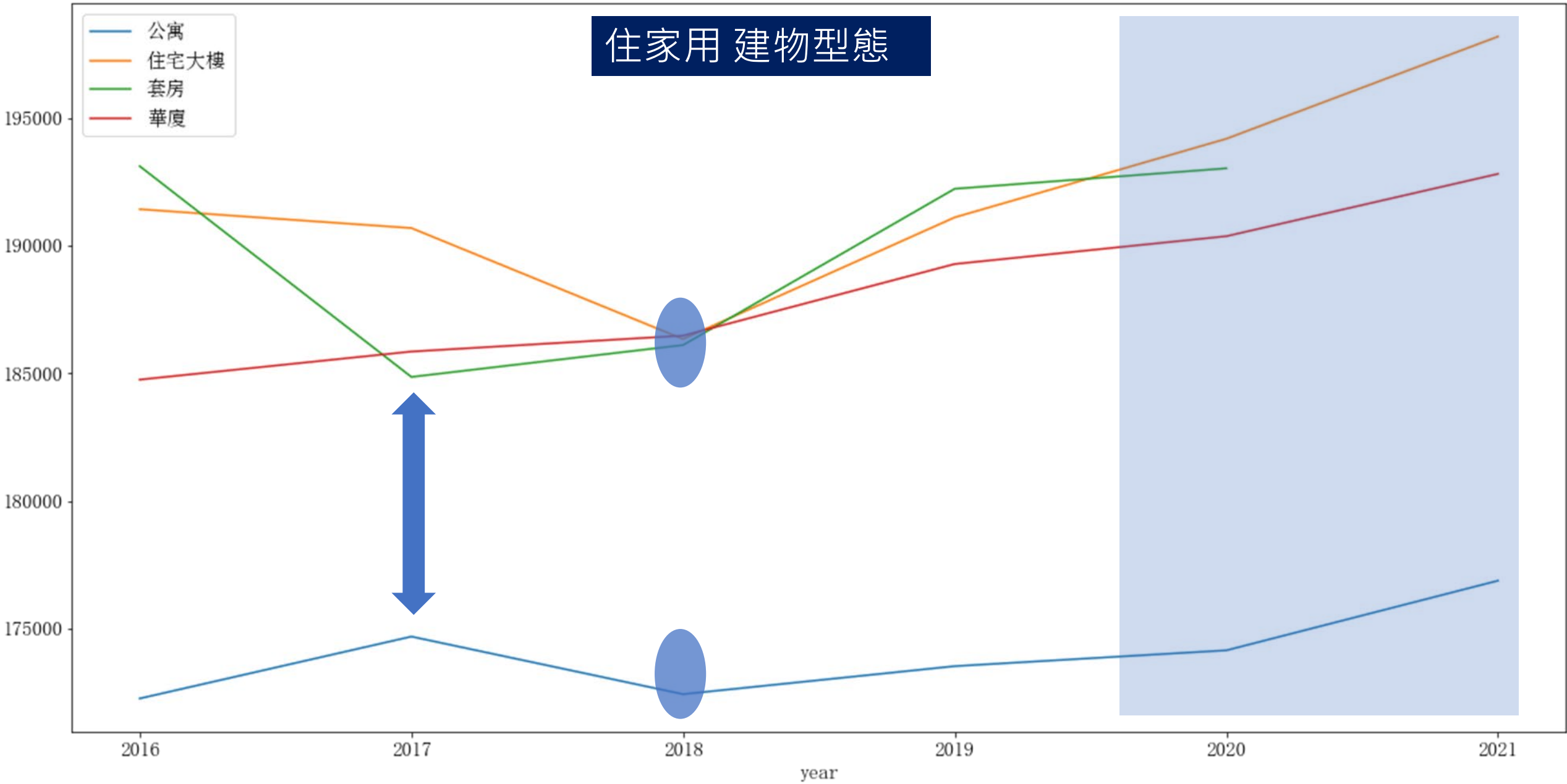


The background of the image is dark and features several stacks of gold coins of varying heights. Interspersed among these stacks are small, detailed models of houses. One house has a red roof, while others are in shades of grey and brown. The overall composition suggests a connection between real estate and finance.

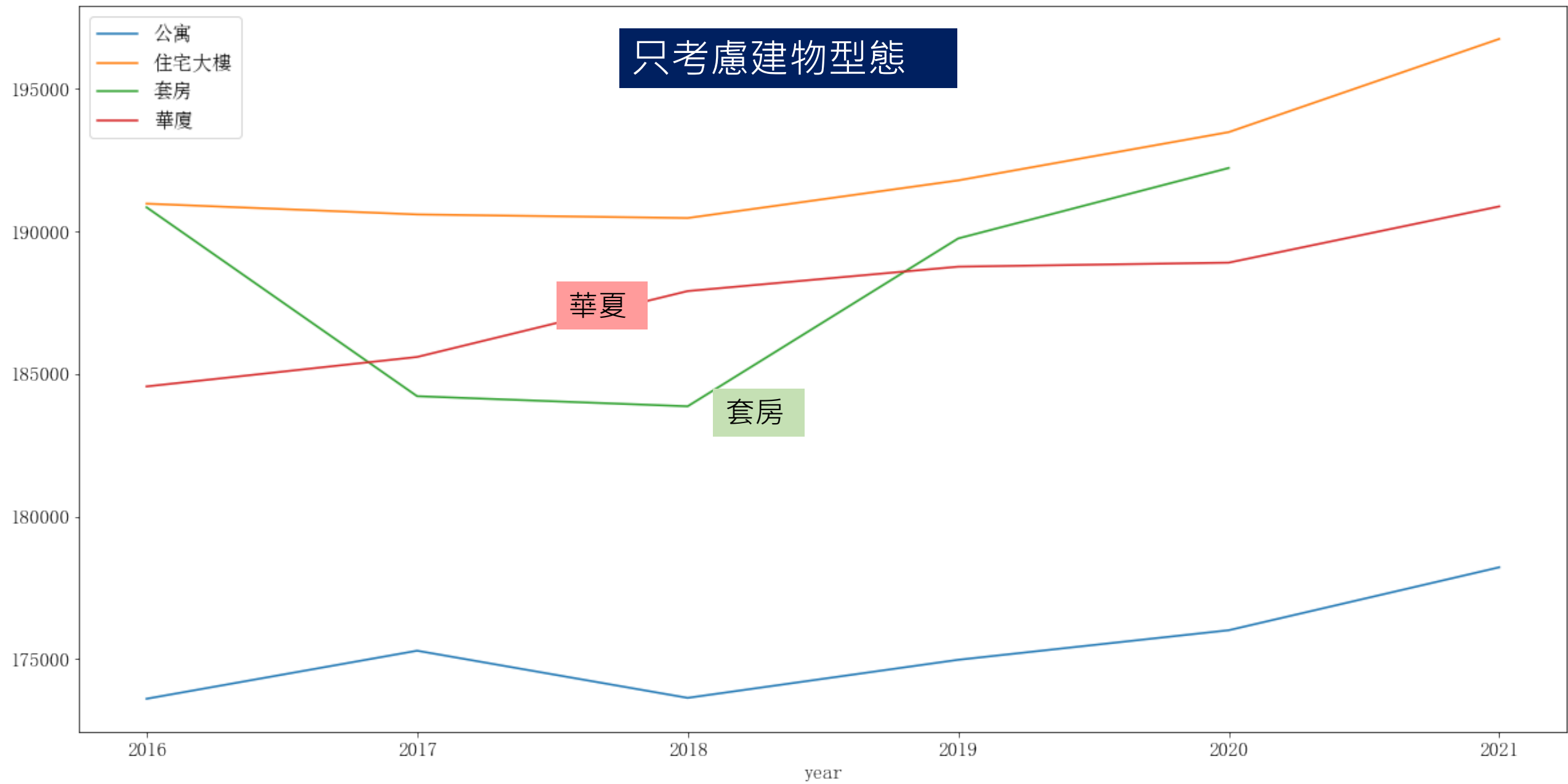
不同用途中  
不同建物型態價格差異為何？



# 住家用 建物型態



| Year | 公寓     | 住宅大樓   | 套房     | 華廈     |
|------|--------|--------|--------|--------|
| 2016 | 172256 | 191424 | 193109 | 184748 |
| 2017 | 174682 | 190686 | 184853 | 185848 |
| 2018 | 172425 | 186338 | 186102 | 186471 |
| 2019 | 173522 | 191104 | 192224 | 189277 |
| 2020 | 174142 | 194184 | 193026 | 190366 |
| 2021 | 176872 | 198188 |        | 192808 |



| Year | 公寓     | 住宅大樓   | 套房     | 華廈     |
|------|--------|--------|--------|--------|
| 2016 | 173609 | 190979 | 190852 | 184567 |
| 2017 | 175295 | 190601 | 184225 | 185601 |
| 2018 | 173640 | 190475 | 183870 | 187912 |
| 2019 | 174976 | 191796 | 189759 | 188768 |
| 2020 | 176013 | 193490 | 192231 | 188910 |
| 2021 | 178222 | 196760 |        | 190883 |

# 資料欄位

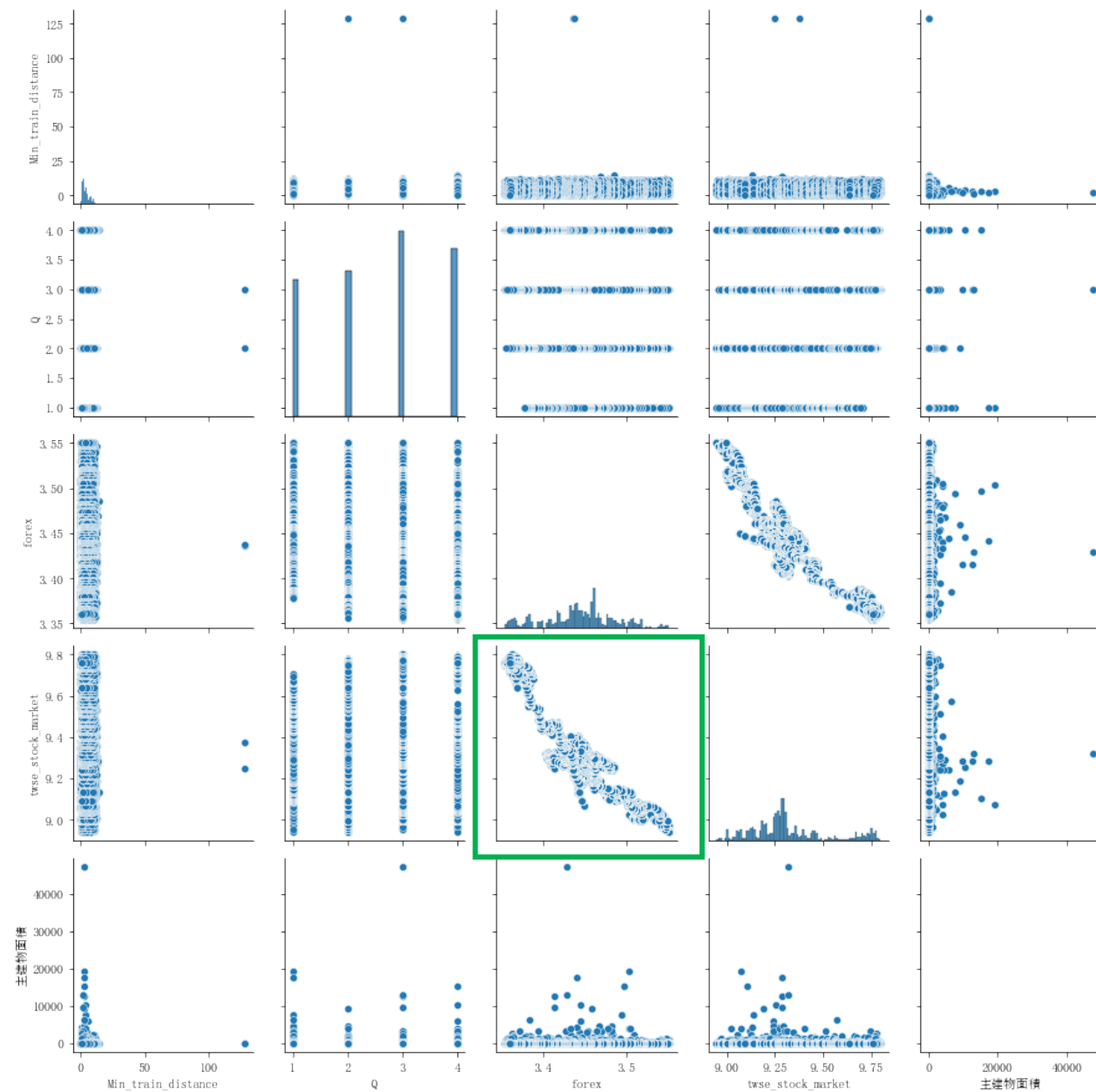
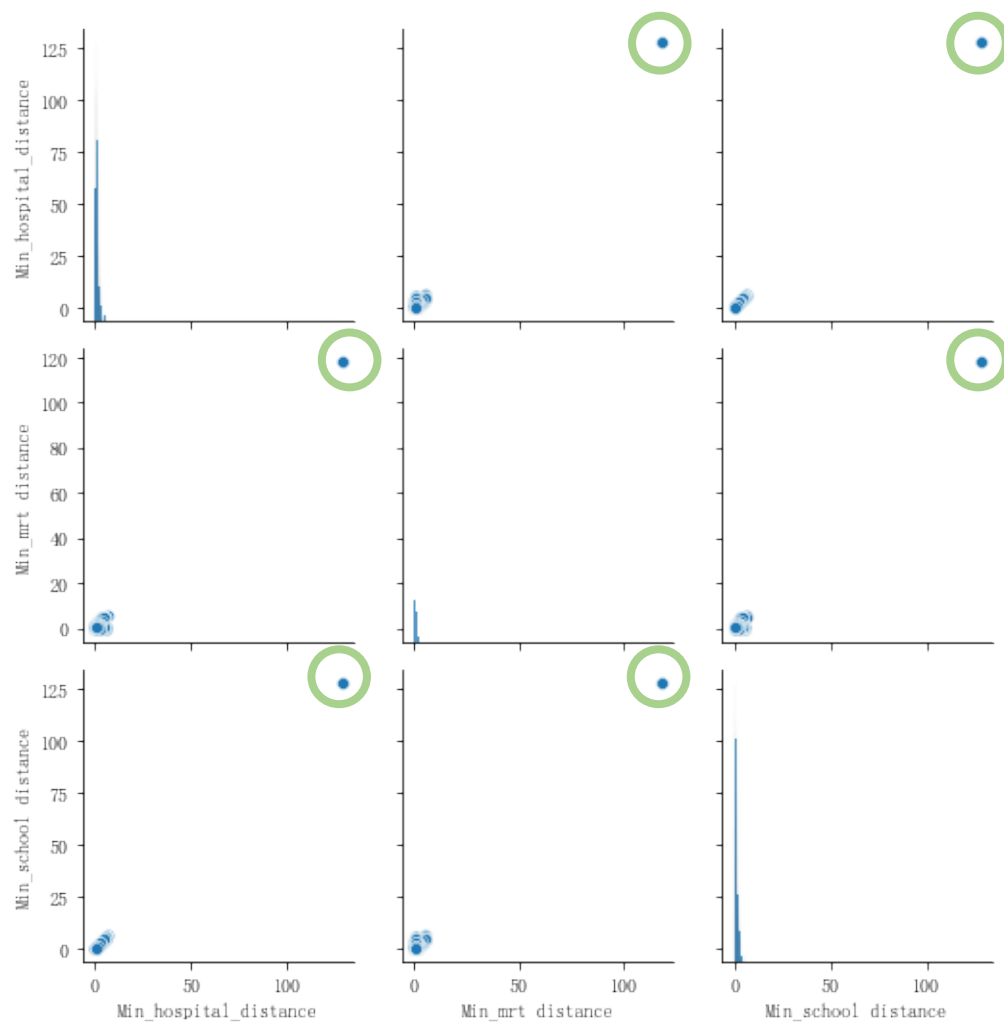
平均數比中位數大：右偏  
平均數比中位數小：左偏

- 欄位合併後 一共 235個欄位
- 欄位合併依據：[台北市土地使用分區管理規則](#)

距離單位：公里  
大部分資料右偏

|                            | count | mean   | std    | min      | 25%    | 50%    | 75%    | max    |
|----------------------------|-------|--------|--------|----------|--------|--------|--------|--------|
| Min_hospital_distance      | 51987 | 1.2399 | 1.1387 | 0.01902  | 0.6448 | 1.0354 | 1.5894 | 128.15 |
| Min_mrt distance           | 51987 | 0.618  | 0.8635 | 3.95E-13 | 0.3301 | 0.4955 | 0.716  | 118.17 |
| Min_school distance        | 51987 | 1.2399 | 1.1387 | 0.01902  | 0.6448 | 1.0354 | 1.5894 | 128.15 |
| Min_shopping_mall distance | 51987 | 0.9531 | 1.0761 | 0        | 0.3687 | 0.7274 | 1.3231 | 127.66 |
| Min_train_distance         | 51987 | 3.7371 | 2.6928 | 0.09457  | 1.7105 | 3.1487 | 5.0273 | 128.93 |
| Q                          | 51987 | 2.6053 | 1.0945 | 1        | 2      | 3      | 4      | 4      |
| forex                      | 51987 | 3.4434 | 0.0424 | 3.35449  | 3.4185 | 3.4448 | 3.4673 | 3.5507 |
| twse_stock_market          | 51987 | 9.3155 | 0.1968 | 8.94442  | 9.1929 | 9.2828 | 9.3851 | 9.8001 |
| 主建物面積                      | 51987 | 93.085 | 296.88 | 0        | 42.27  | 75.48  | 104.88 | 47356  |

# 資料欄位



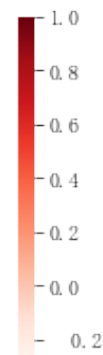
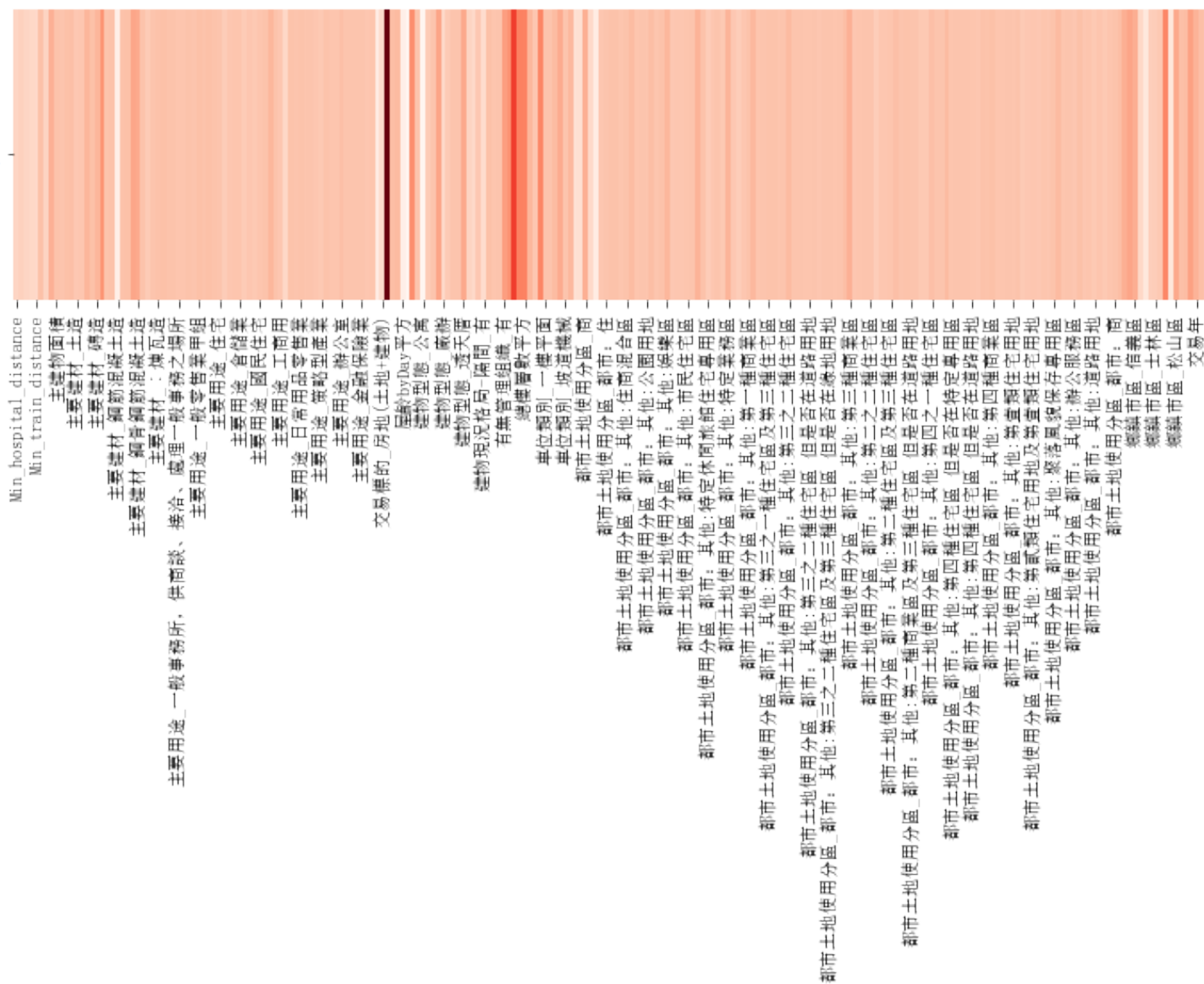
★ 價位資訊全部取Log 處理：

外匯、大盤指數、車位總價元、單價元平方公尺、總價元

★ Correlation Matrix 以單價元平方公尺作為 Y



### Correlation Matrix



|              |          |
|--------------|----------|
| 總價元          | 0.512204 |
| 總樓層數         | 0.266778 |
| 總樓層數平方       | 0.260582 |
| 車位總價元        | 0.247987 |
| 鄉鎮市區_大安區     | 0.242526 |
| 建物型態_住宅大樓    | 0.215392 |
| 移轉層次         | 0.204168 |
| 主要建材_見其他登記事項 | 0.192655 |
| 有無管理組織_有     | 0.152641 |
| 電梯_有         | 0.141430 |
| 建物型態_店面      | 0.126154 |
| 鄉鎮市區_中正區     | 0.125924 |
| 主要建材_鋼骨造     | 0.120082 |
| 鄉鎮市區_松山區     | 0.113124 |



| 構造類別        |             | 103年2月1日實施<br>單位:元/平方公尺 | 調整後<br>單位:元/平方公尺(依臺北市<br>營造工程物價指數調高<br>4.35%，無條件捨去個位數) |
|-------------|-------------|-------------------------|--------------------------------------------------------|
| 加強磚造及輕型鋼架構造 |             | 7,080                   | 7,380                                                  |
| 鋼筋混凝土造      | 一至五層建築物     | 8,180                   | 8,530                                                  |
|             | 六至八層建築物     | 10,620                  | 11,080                                                 |
|             | 九至十二層建築物    | 12,220                  | 12,750                                                 |
|             | 十三至十五層建築物   | 14,660                  | 15,290                                                 |
|             | 十六至二十層建築物   | 15,390                  | 16,050                                                 |
|             | 二十一至二十五層建築物 | 16,170                  | 16,870                                                 |
|             | 二十六至三十層建築物  | 16,980                  | 17,710                                                 |
|             | 三十一層以上建築物   | 17,810                  | 18,580                                                 |
| 鋼骨鋼筋混凝土造    | 十層以下建築物     | 15,150                  | 15,800                                                 |
|             | 十一至十五層建築物   | 15,910                  | 16,600                                                 |
|             | 十六至二十層建築物   | 16,700                  | 17,420                                                 |
|             | 二十一至二十五層建築物 | 17,540                  | 18,300                                                 |
|             | 二十六至三十層建築物  | 18,420                  | 19,220                                                 |
|             | 三十一至三十五層建築物 | 19,350                  | 20,190                                                 |
|             | 三十六層以上建築物   | 20,310                  | 21,190                                                 |

資料來源：[台北市建築公會](#)

|                 |             |           |           |           |
|-----------------|-------------|-----------|-----------|-----------|
| 鋼骨構造            | 十層以下建築物     |           | 18,570    | 19,370    |
|                 | 十一至十五層建築物   |           | 19,500    | 20,340    |
|                 | 十六至二十層建築物   |           | 20,480    | 21,370    |
|                 | 二十一至二十五建築物  |           | 21,490    | 22,420    |
|                 | 二十六至三十建築物   |           | 22,570    | 23,550    |
|                 | 三十一至三十五建築物  |           | 23,710    | 24,740    |
|                 | 三十六層以上建築物   |           | 24,890    | 25,970    |
| 土地改良物及<br>雜項工作物 | 挖方          | 立方公尺      | 150       | 150       |
|                 | 填方          | 立方公尺      | 230       | 240       |
|                 | 圍牆          | 公尺        | 2,200     | 2,290     |
| 擋土牆(公尺)         | 砌卵石         | 公尺        | 1,950     | 2,030     |
|                 | 鋼筋混<br>凝土   | 三公尺以下     | 4,020     | 4,190     |
|                 |             | 超過三公尺至五公尺 | 4,750     | 4,950     |
|                 |             | 超過五公尺至八公尺 | 7,690     | 8,020     |
|                 |             | 超過八公尺以上   | 19,910    | 20,770    |
| 排水溝(公尺)         | 五十公分以下      |           | 720       | 750       |
|                 | 超過五十公分至一百公分 |           | 1,950     | 2,030     |
|                 | 超過一百公分以上    |           | 3,310     | 3,450     |
|                 |             |           | 其他以實際造價計算 | 其他以實際造價計算 |

備註：

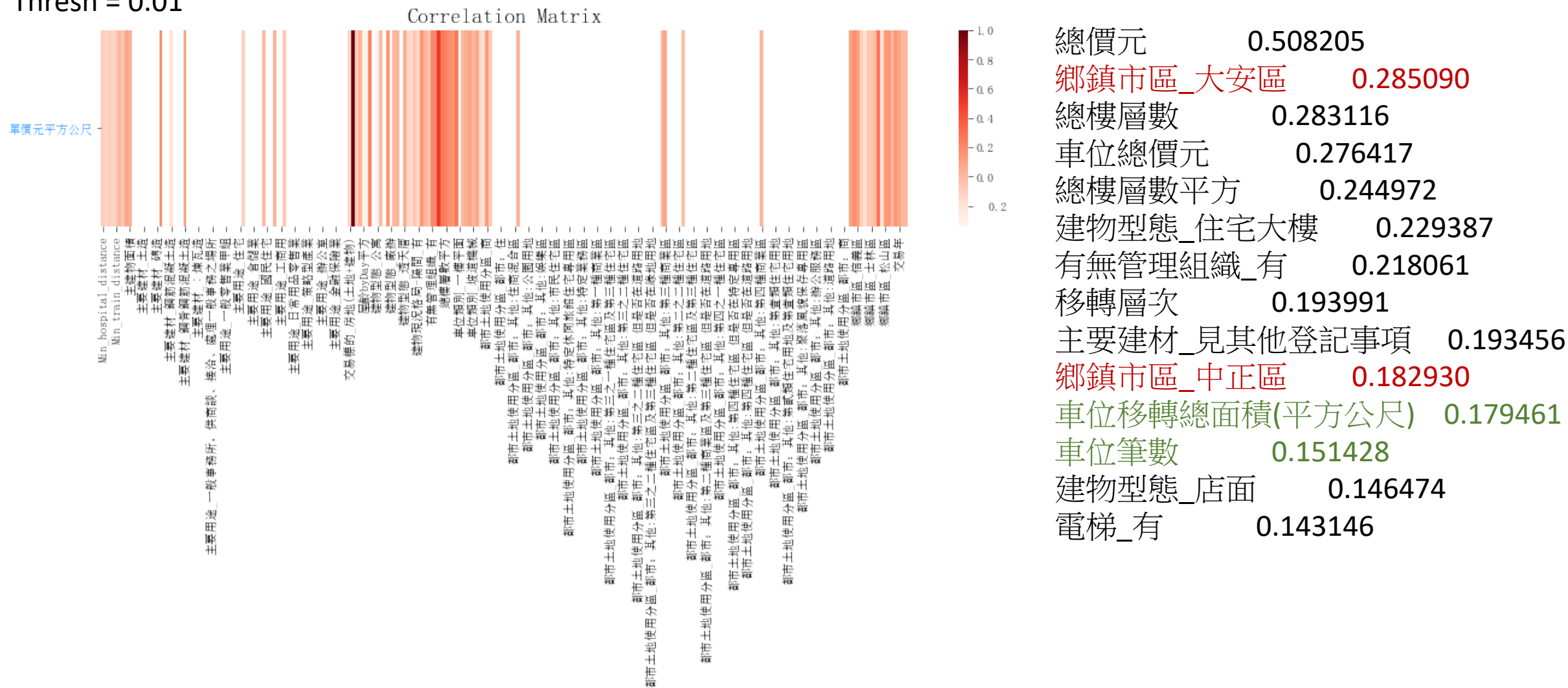
1. 本表未列之工程項目，以實際施工所需之工程費用為準。

原樣本總數 : 51987

Thresh後樣本總數 : 37598

Thresh後樣本總數：**37598**

Thresh = 0.01



# 目錄

**01**

目的和資料來源

資料視覺化

**02**

**03**

模型建置

**05**

分析結果

優化模型

**04**

學習心得

**06**



# PLS model

|                                                                                   |                                                                                                                 |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| <p>方法</p> <p>把X Y 投影到新的平面去做分析</p> <p>X、Y之間需是線性關係</p>                              | <p>適用時機</p> <ol style="list-style-type: none"><li>1. 當預測變數量大於樣本量且OLS產生係數標準誤高或完全失效</li><li>2. 預測變量高度共線</li></ol> |
| <p>超參數</p> <p>唯一參數為<br/>n_components，</p> <p>代表要保留的<br/>components，<br/>預設值為2</p> | <p>模型介紹</p> <p><b>PLS1</b> 是指只有一個因變量的偏最小二乘模型，<br/>而 <b>PLS2</b> 是指具有多個因變量的模型</p>                                |

# PLS model

$$\mathbf{X} = \mathbf{Z}\mathbf{V}^T + \mathbf{E} \quad (16.2)$$

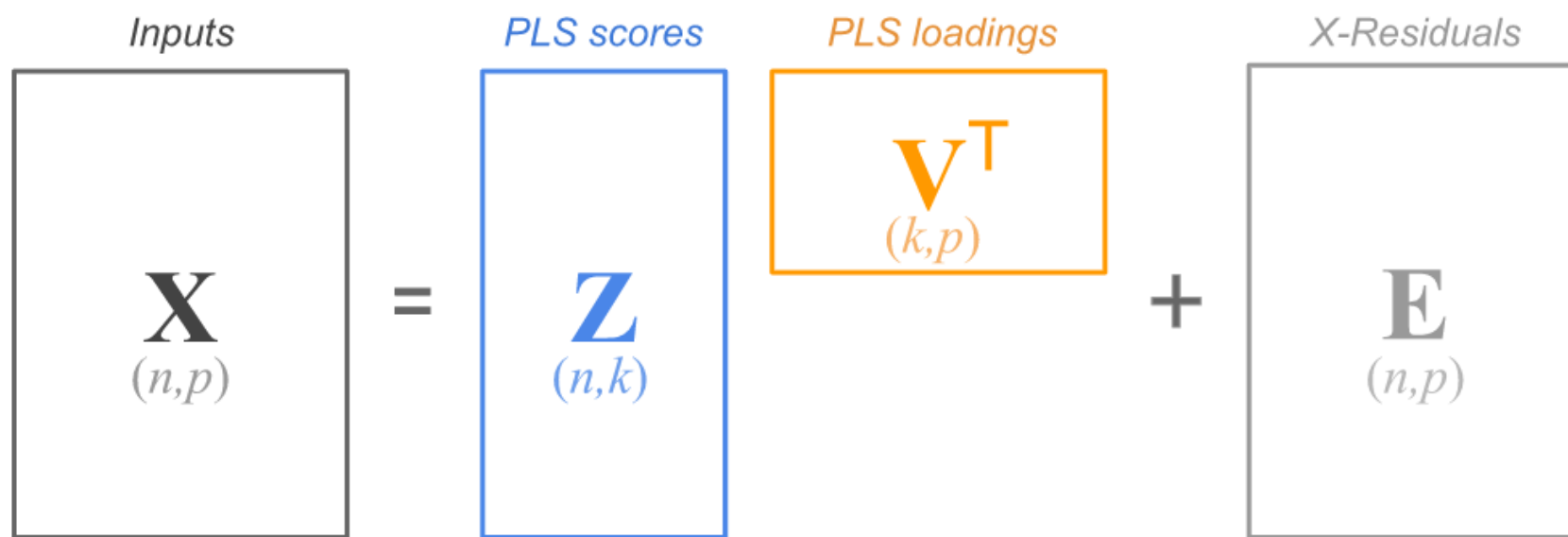


Figure 16.1: Matrix diagram for inputs

# PLS model

$$\mathbf{y} = \mathbf{Z}\mathbf{b} + \mathbf{e} \quad (16.3)$$

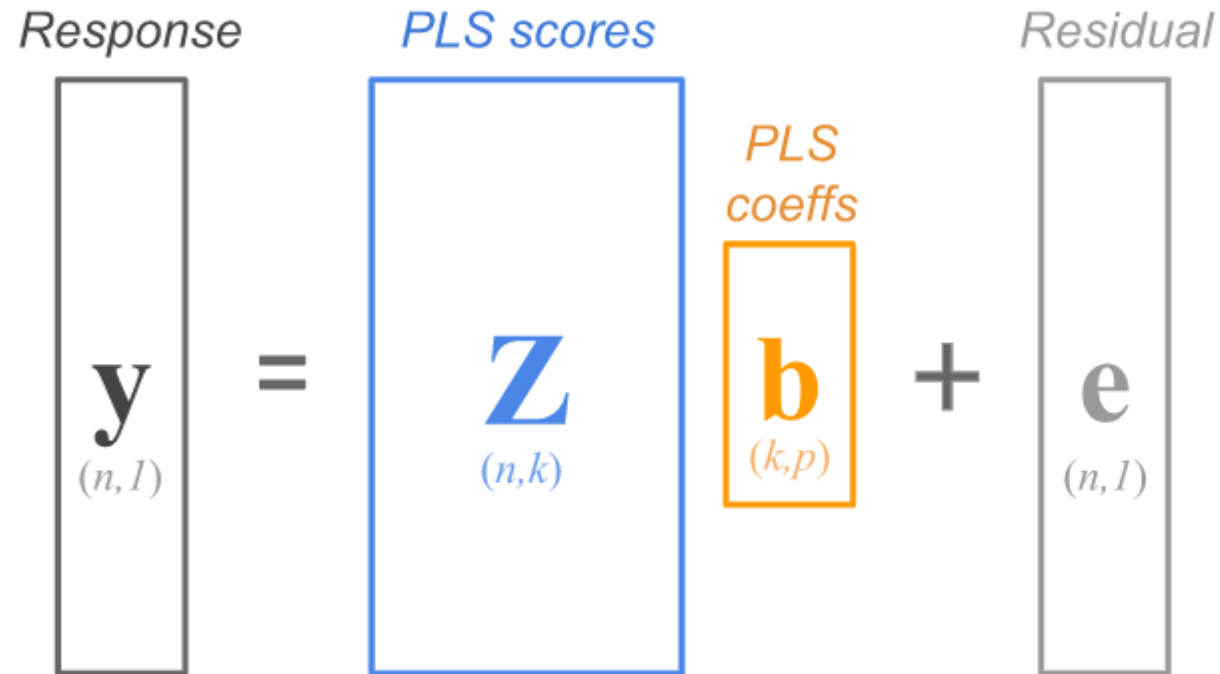


Figure 16.2: Matrix diagram for response

# PLS model

$$\tilde{\mathbf{w}}_1 = (\text{cov}(\mathbf{x}_1, \mathbf{y}), \dots, \text{cov}(\mathbf{x}_p, \mathbf{y}))$$



$$\tilde{\mathbf{w}}_1 = \mathbf{X}^T \mathbf{y} / \mathbf{y}^T \mathbf{y}$$



$$\mathbf{w}_1 = \frac{\tilde{\mathbf{w}}_1}{\|\tilde{\mathbf{w}}_1\|}$$

$$\mathbf{x}_1 \longrightarrow \tilde{w}_{11} = \text{cov}(\mathbf{x}_1, \mathbf{y})$$

$$\mathbf{x}_2 \longrightarrow \tilde{w}_{21} = \text{cov}(\mathbf{x}_2, \mathbf{y})$$

⋮

$$\mathbf{x}_j \longrightarrow \tilde{w}_{j1} = \text{cov}(\mathbf{x}_j, \mathbf{y})$$

⋮

$$\mathbf{x}_p \longrightarrow \tilde{w}_{p1} = \text{cov}(\mathbf{x}_p, \mathbf{y})$$

# PLS model

$$\mathbf{z}_1 = \mathbf{X}\mathbf{w}_1 = \mathbf{X}\mathbf{w}_1 / \mathbf{w}_1^T \mathbf{w}_1$$



$$\mathbf{v}_1 = \mathbf{X}^T \mathbf{z}_1 / \mathbf{z}_1^T \mathbf{z}_1$$

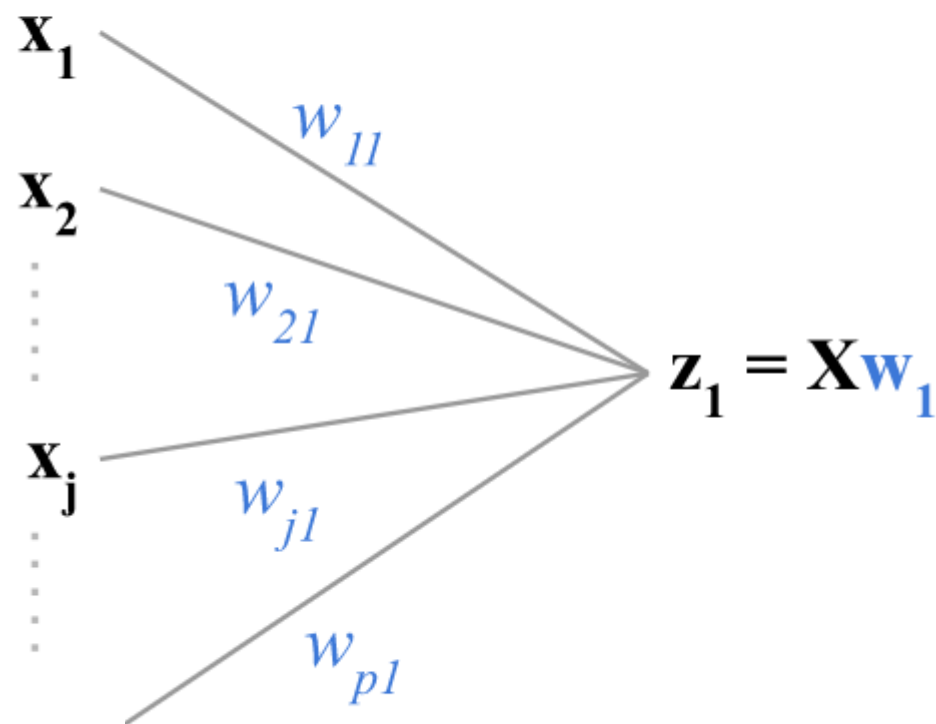
$$b_1 = \mathbf{y}^T \mathbf{z}_1 / \mathbf{z}_1^T \mathbf{z}_1$$



$$\mathbf{X}_1 = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{z}_1 \mathbf{v}_1^T$$

$$\mathbf{y}_1 = \mathbf{y} - b_1 \mathbf{z}_1$$

$$\mathbf{z}_1 = w_{11}\mathbf{x}_1 + \cdots + w_{p1}\mathbf{x}_p$$



(deflation)





# PLS model

0. We start by setting  $\mathbf{X}_0 = \mathbf{X}$ , and  $\mathbf{y}_0 = \mathbf{y}$ .

Repeat for  $h = 1, \dots, r = \text{rank}(\mathbf{X})$ :

1. Start with weights  $\tilde{\mathbf{w}}_h = \mathbf{X}_{h-1}^T \mathbf{y}_{h-1} / \mathbf{y}_{h-1}^T \mathbf{y}_{h-1}$

2. Normalize weights:  $\mathbf{w}_h = \tilde{\mathbf{w}}_h / \|\tilde{\mathbf{w}}_h\|$

3. Compute PLS component:  $\mathbf{z}_h = \mathbf{X}_{h-1} \mathbf{w}_h / \mathbf{w}_h^T \mathbf{w}_h$

4. Regress  $\mathbf{y}_h$  onto  $\mathbf{z}_h$ :  $b_h = \mathbf{y}_h^T \mathbf{z}_h / \mathbf{z}_h^T \mathbf{z}_h$

5. Regress all  $\mathbf{x}_j$  onto  $\mathbf{z}_h$ :  $\mathbf{v}_h = \mathbf{X}_{h-1}^T \mathbf{z}_h / \mathbf{z}_h^T \mathbf{z}_h$

6. Deflate (residual) predictors:  $\mathbf{X}_h = \mathbf{X}_{h-1} - \mathbf{z}_h \mathbf{v}_h^T$

7. Deflate (residual) response:  $\mathbf{y}_h = \mathbf{y}_{h-1} - b_h \mathbf{z}_h$

實際情況

$$\hat{\mathbf{y}} = \mathbf{Z}\mathbf{b} \longrightarrow \hat{\mathbf{y}} = \mathbf{X}\mathbf{b}^*$$

# PLS model

```
1  function PLS1( $X, y, l$ )
2   $X^{(0)} \leftarrow X$ 
3   $w^{(0)} \leftarrow X^T y / \|X^T y\|$ , an initial estimate of  $w$ .
4   $t^{(0)} \leftarrow X w^{(0)}$ 
5  for  $k = 0$  to  $l$ 
6       $t_k \leftarrow t^{(k)T} t^{(k)}$  (note this is a scalar)
7       $t^{(k)} \leftarrow t^{(k)} / t_k$ 
8       $p^{(k)} \leftarrow X^{(k)T} t^{(k)}$ 
9       $q_k \leftarrow y^T t^{(k)}$  (note this is a scalar)
10     if  $q_k = 0$ 
11          $l \leftarrow k$ , break the for loop
12     if  $k < l$ 
13          $X^{(k+1)} \leftarrow X^{(k)} - t_k t^{(k)} p^{(k)T}$ 
14          $w^{(k+1)} \leftarrow X^{(k+1)T} y$ 
15          $t^{(k+1)} \leftarrow X^{(k+1)} w^{(k+1)}$ 
16     end for
17     define  $W$  to be the matrix with columns  $w^{(0)}, w^{(1)}, \dots, w^{(l-1)}$ .
        Do the same to form the  $P$  matrix and  $q$  vector.
18      $B \leftarrow W(P^T W)^{-1} q$ 
19      $B_0 \leftarrow q_0 - P^{(0)T} B$ 
20     return  $B, B_0$ 
```

# PLS model

Tuning  
hyperparameter

- For  $k = 1, 2, \dots, r = \text{rank}(\mathbf{X})$ 
  - For  $q = 1, \dots, Q$ 
    - fit PLSR model  $h_{k,q}$  with  $k$  PLS-scores on  $\mathcal{D}_{\text{train}-q}$
    - compute and store  $E_{\text{eval}-q}(h_{k,q})$  using  $\mathcal{D}_{\text{eval}-q}$
  - end for  $q$
  - compute and store  $E_{cv_k} = \frac{1}{Q} \sum_q E_{\text{eval}-q}(h_{k,q})$
- end for  $k$
- Compare all cross-validation errors  $E_{cv_1}, E_{cv_2}, \dots, E_{cv_r}$  and choose the smallest of them, say  $E_{cv_{k^*}}$
- Use  $k^*$  PLS scores to fit the (finalist) PLSR model:
$$\hat{\mathbf{y}} = b_1 \mathbf{z}_1 + b_2 \mathbf{z}_2 + \dots + b_{k^*} \mathbf{z}_{k^*} = \mathbf{Z}_{1:k^*} \mathbf{b}_{1:k^*}$$
- Remember that we can reexpress the PLSR model in terms of the original predictors: 
$$\hat{\mathbf{y}} = (\mathbf{X} \mathbf{W}_{1:k^*}^*) \mathbf{b}_{1:k^*} = \mathbf{X} \mathbf{b}_k^*$$

# PLS model

## 參數設定和步驟

1. 使用預設參數  $n = 2$
2. 資料切成 70% 為訓練，30% 測試
3. 擬合訓練資料
4. 分別計算訓練跟測試的 RMSE 和得分 (coefficient)

RMSE train = 0.30101165899325705

RMSE test = 0.3673149931726019

Coefficient:

PLS train: 0.4426842691130447

PLS test: 0.3444534141293333

overfitting

|              |           |
|--------------|-----------|
| 鄉鎮市區_大安區     | 0.071129  |
| 建物型態_透天厝     | 0.043130  |
| 建物型態_店面      | 0.041096  |
| 鄉鎮市區_中正區     | 0.037273  |
| 鄉鎮市區_松山區     | 0.036969  |
| 車位總價元        | 0.034377  |
| 鄉鎮市區_信義區     | 0.032175  |
| 主要建材_見其他登記事項 | 0.025678  |
| 主要建材_磚造      | 0.021011  |
| 鄉鎮市區_中山區     | 0.019430  |
| 交易標的_建物      | -0.053162 |
| 都市土地使用分區_未公告 | -0.050808 |
| 鄉鎮市區_文山區     | -0.047980 |
| 鄉鎮市區_北投區     | -0.042632 |
| 主要建材_鋼筋混凝土造  | -0.037421 |
| 建物型態_公寓      | -0.034903 |
| 屋齡byDay      | -0.034130 |
| 建物型態_廠辦      | -0.027276 |
| 鄉鎮市區_內湖區     | -0.025690 |
| 都市土地使用分區_工   | -0.024052 |

# 目錄

**01**

目的和資料來源

資料視覺化

**02**

**03**

模型建置

**05**

分析結果

優化模型

**04**

學習心得

**06**

增加資料集：  
手邊已有資料已全部  
用到，無法再增加

1

優化模型超參數

2

訓練資料增加閾值  
避免訓練雜訊

3

優化方向

5

換一個模型試試看

4

簡化模型變數

增加資料集：  
手邊已有資料已全部  
用到，無法再增加

1

2

優化模型超參數

優化方向

3

訓練資料增加閥值  
避免訓練雜訊

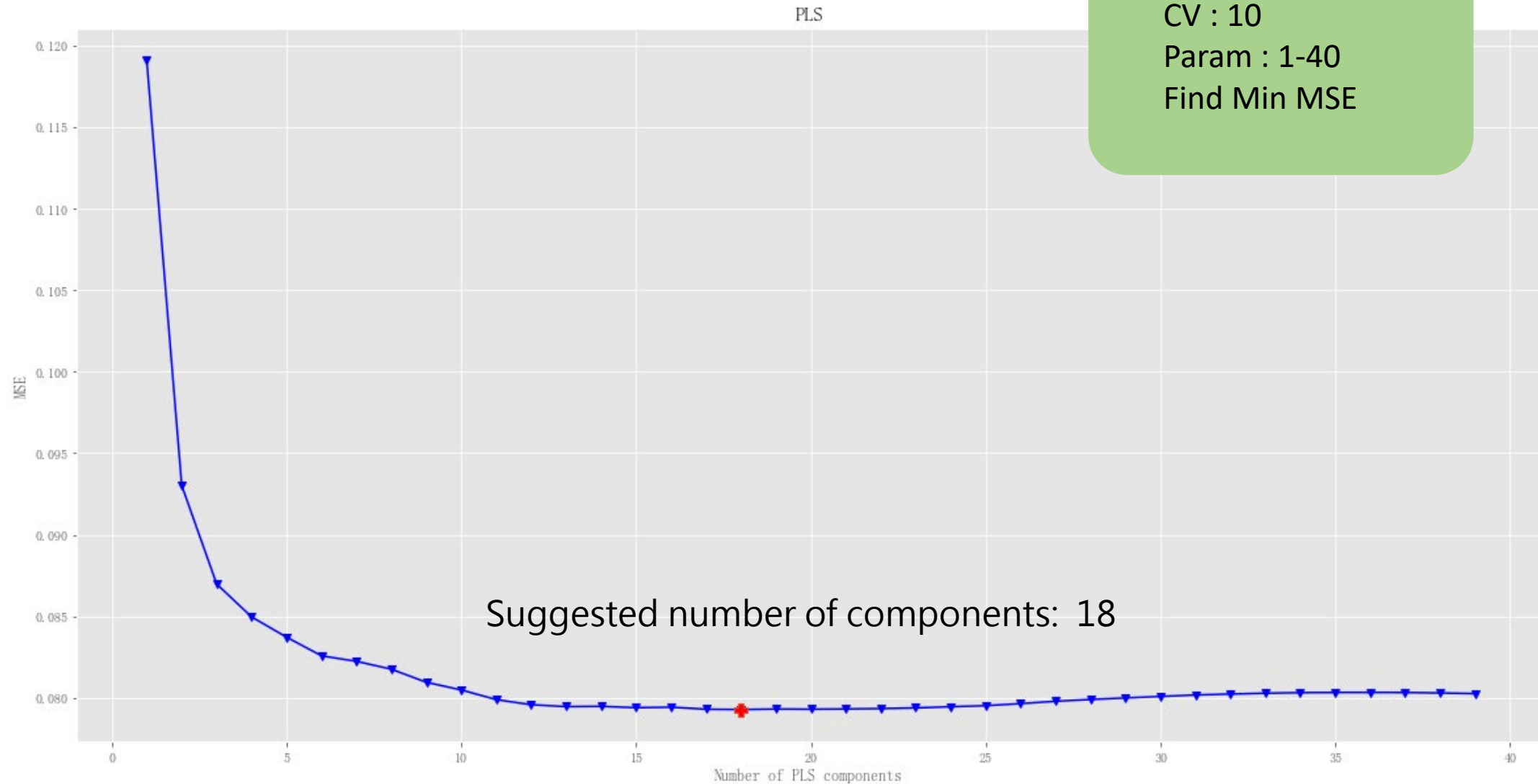
5

換一個模型試試看

4

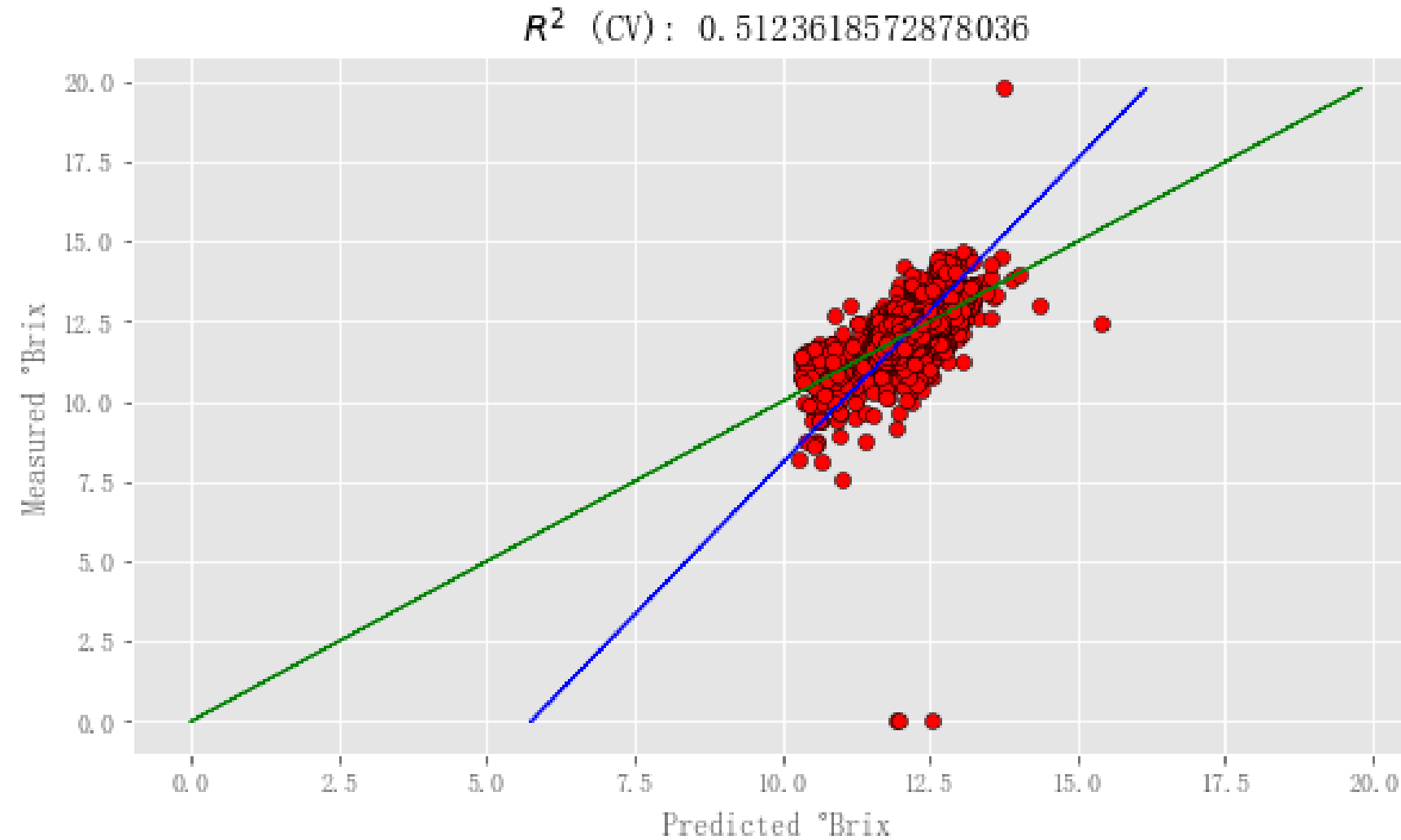
簡化模型變數

# PLS tuning hyperparameter





# PLS tuning hyperparameter



R2 calib: 0.526  
R2 CV: 0.512  
MSE calib: 0.077  
MSE CV: 0.079

# PLS 優化後結果

RMSE train = 0.30101165899325705  
RMSE test = 0.3673149931726019

Coefficient:

PLS train: 0.4426842691130447

PLS test: 0.3444534141293333

測試資料 RMSE 下降 4.8 %

測試資料 Coefficient 上升 17.7%

RMSE train = 0.27760985810473326

RMSE test = 0.34982948805965125

Scoring ( coefficient )

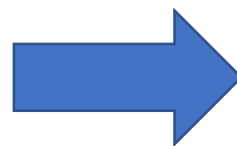
PLS Train: 0.5259715198648702

PLS Test: 0.40538059608892896

# Grid Search

參數設定

n\_components 從 1 到 100，每次增加2  
CV = 10，使用 neg\_mean\_squared\_error



推薦Param

PLSRegression(n\_components=17)

經測試後，n\_components = 17  
的RMSE 和 coefficient 相較 18  
在Test上表現並沒有比較好

選擇 n = 18 作為最佳參數

# Grid Search

| param_n_components | split0_test_score | split1_test_score | split2_test_score | split3_test_score | split4_test_score | split5_test_score | split6_test_score | split7_test_score |
|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 17                 | -0.067            | -0.061            | -0.074            | -0.100            | -0.061            | -0.085            | -0.064            | -0.107            |
| 19                 | -0.067            | -0.061            | -0.074            | -0.100            | -0.061            | -0.085            | -0.064            | -0.107            |
| 21                 | -0.068            | -0.061            | -0.074            | -0.100            | -0.061            | -0.084            | -0.065            | -0.107            |

| split8_test_score | split9_test_score | mean_test_score | std_test_score | rank_test_score |
|-------------------|-------------------|-----------------|----------------|-----------------|
| -0.107            | -0.065            | -0.079          | 0.018          | 1               |
| -0.107            | -0.065            | -0.079          | 0.018          | 2               |
| -0.107            | -0.065            | -0.079          | 0.018          | 3               |



# PLS 最佳解結果

鄉鎮市區\_大安區 0.152485

屋齡byDay平方 0.127302

鄉鎮市區\_松山區 0.110909

鄉鎮市區\_中山區 0.110833

鄉鎮市區\_信義區 0.102128

鄉鎮市區\_中正區 0.100352

車位總價元 0.061349

建物型態\_店面 0.060416

建物型態\_透天厝 0.054341

土地移轉總面積平方公尺 0.048870

屋齡byDay -0.235223

交易標的\_建物 -0.078655

建物移轉總面積平方公尺 -0.048189

主要建材\_鋼筋混凝土造 -0.033468

建物型態\_公寓 -0.021057

總樓層數 -0.018593

建物型態\_廠辦 -0.017314

建物現況格局-房 -0.016961

主要用途\_農舍 -0.016096

都市土地使用分區\_工 -0.013712

## 優化方向

增加資料集：  
手邊已有資料已全部  
用到，無法再增加

1

優化模型超參數

2

3

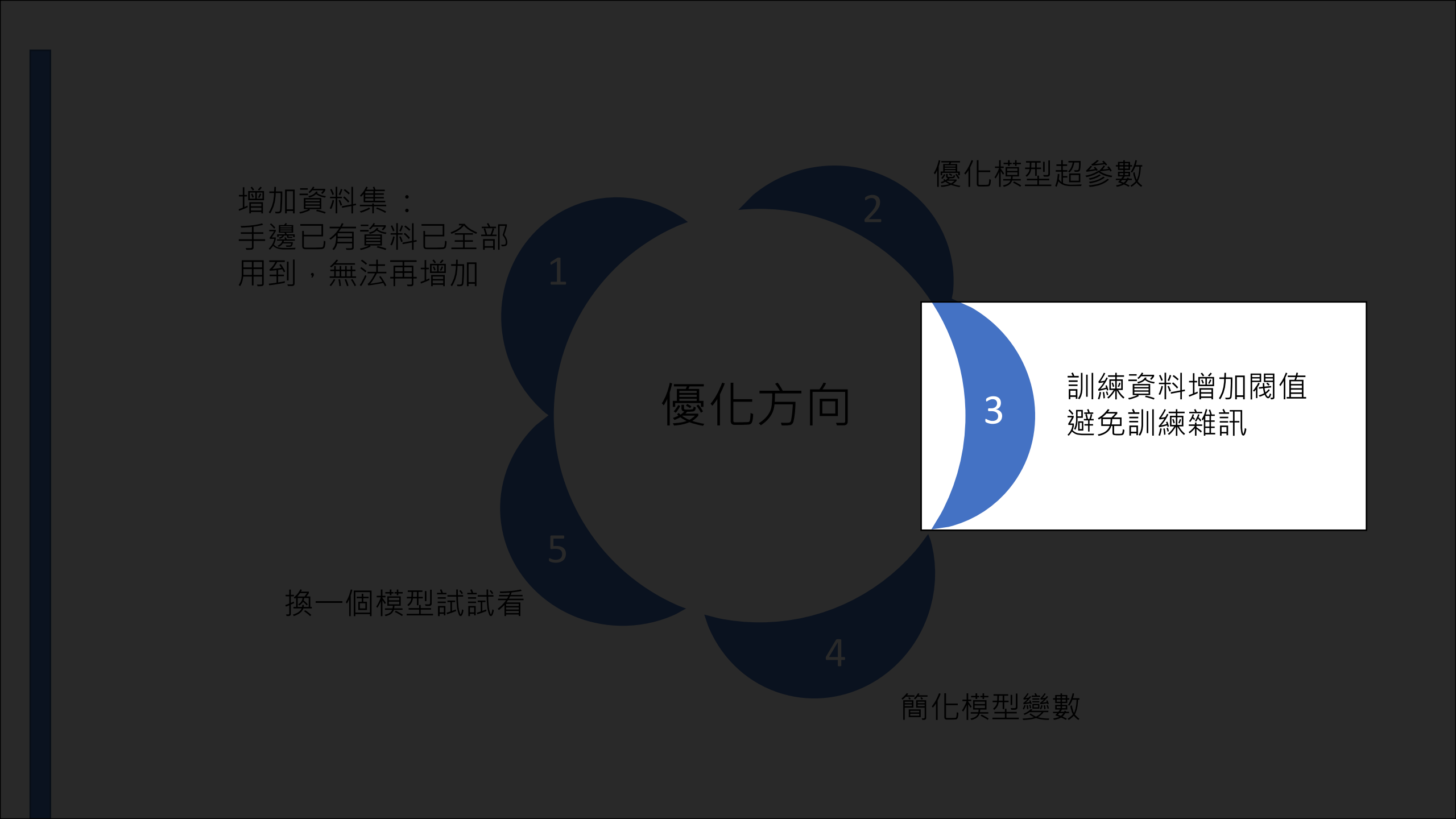
訓練資料增加閾值  
避免訓練雜訊

5

換一個模型試試看

4

簡化模型變數



為什麼要增加閾值，把低於或高於(1-閾值)資料剔除？

- 資料非正向分布，且極端偏向某一邊，透過閾值讓資料在犧牲部分資料代價上使模型更趨於正向分布
- 避免訓練到雜訊

## 為什麼只對訓練資料做閥值去除

- 測試資料為模擬真實情境，若去除極端值變成只針對正常情況做預測，但正常情況發生機率為未知，需要知道極端事件對於整體模型影響到底多大



# Train data

36390 → 26496

Default 參數 = 2

RMSE Train = 0.24658478450616242

RMSE Test = 0.5387436769424702

Coefficient:

PLS train: 0.42857576606955183

PLS test: **-0.41023217877268725**



個人優化後 參數 = 39

RMSE Train = 0.2128253536329557

RMSE Test = 0.8944057287644921

Coefficient :

PLS train: 0.5743302070707275

PLS test: **-2.886828626672633**

# Train data

36390 → 26496

Grid Search 參數 = 47

RMSE Train = 0.21282415302816318

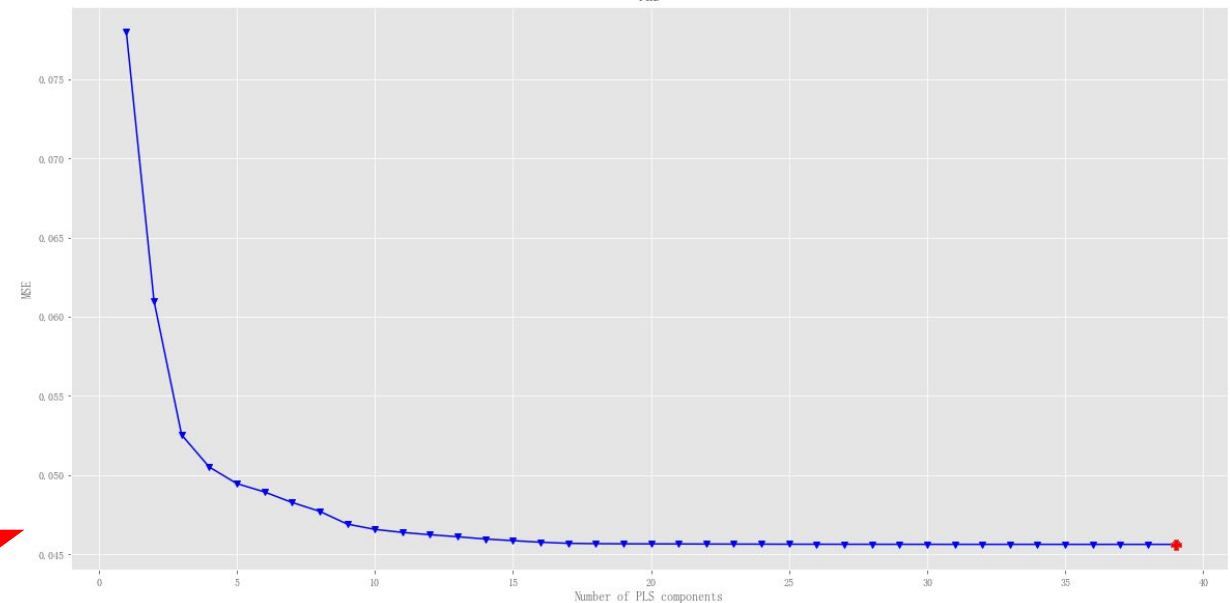
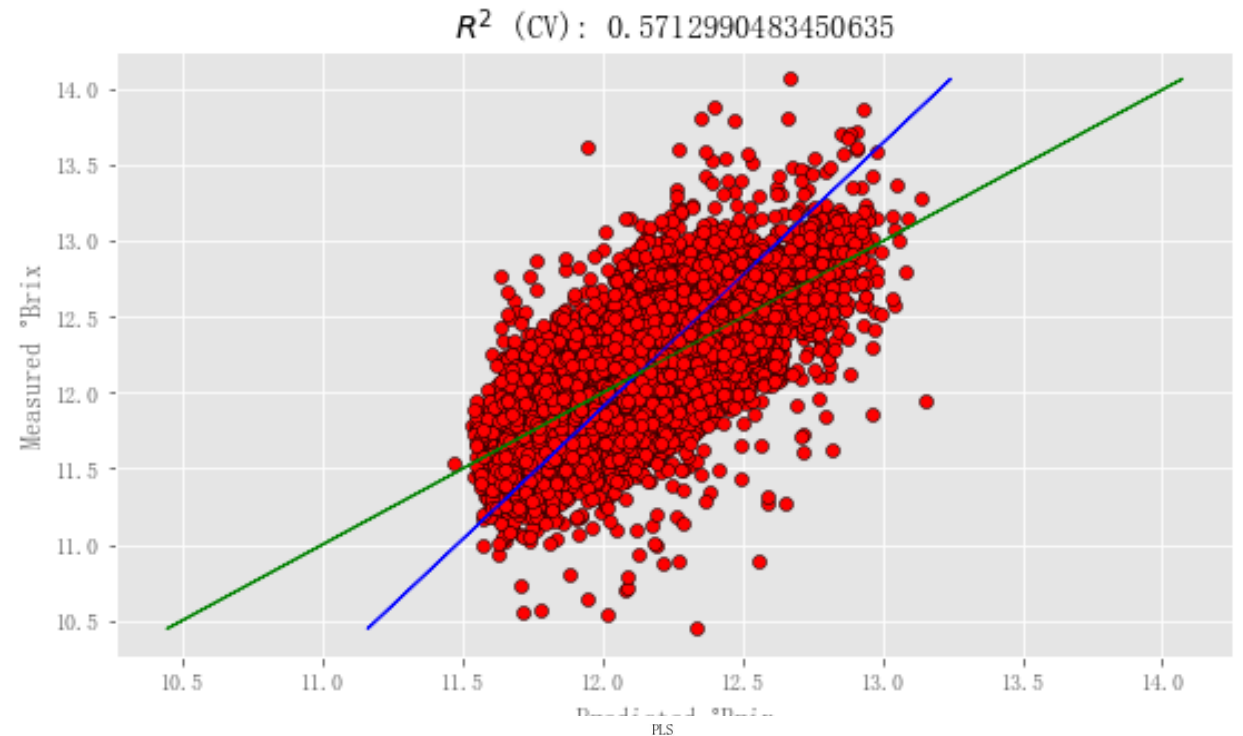
RMSE Test = 0.8928577493849114

Coefficient:

PLS train: 0.5743350096916859

PLS test: **-2.8733861280629736**

此案例不適合用Threshold來優化模型



增加資料集：  
手邊已有資料已全部  
用到，無法再增加

1

優化模型超參數

2

優化方向

3

訓練資料增加閾值  
避免訓練雜訊

5

換一個模型試試看

4

簡化模型變數



# 模型存在嚴重過擬合

根據前述研究，只取在沒有Thresh下的最優解模型中相關性絕對值前二十名作為輸入變數

最優解模型：n\_components = 18

鄉鎮市區\_大安區、鄉鎮市區\_中山區、鄉鎮市區\_松山區  
鄉鎮市區\_信義區、鄉鎮市區\_中正區、鄉鎮市區\_大同區  
鄉鎮市區\_內湖區、鄉鎮市區\_南港區、鄉鎮市區\_士林區

建物型態\_透天厝、建物型態\_店面  
總樓層數平方、屋齡byDay、屋齡byDay平方  
主要建材\_鋼筋混凝土造、twse\_stock\_market、車位總價元  
交易標的\_建物、建物移轉總面積平方公尺、土地移轉總面積平方公尺

# 先使用前次最優解來跑模型

n\_components = 18  
RMSE\_Train = 0.28989664577177654  
RMSE\_Test = 0.35915527921687046

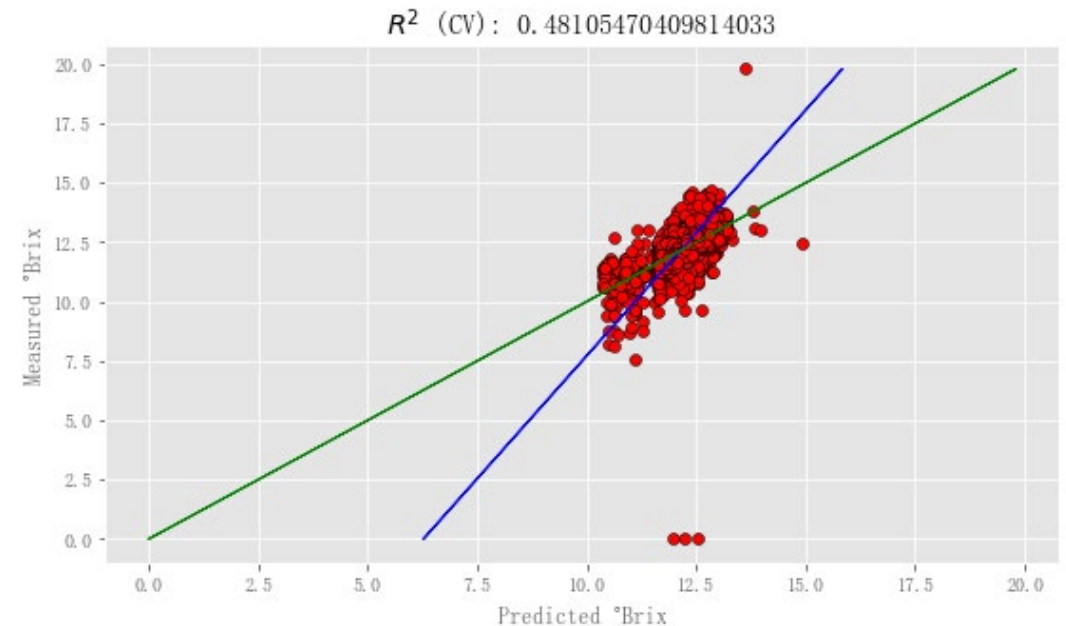
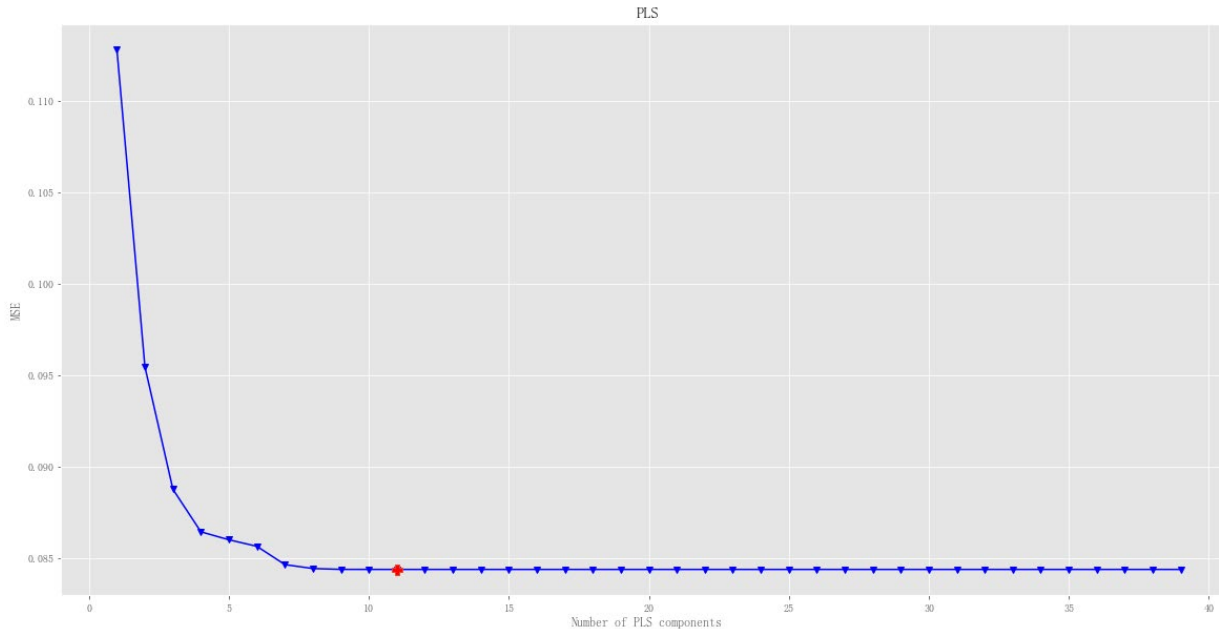
Coefficient:  
PLS train: 0.4830827230104663  
PLS test: 0.37325517432360544



MSE 優化

n\_component = 11  
RMSE\_Train = 0.2898967727969102  
RMSE\_Test = 0.35915659953996276

Coefficient  
PLS train: 0.48308227001108506  
PLS test: 0.3732505662488521



# Grid Search

| param_n_components | split0_test_score | split1_test_score | split2_test_score | split3_test_score | split4_test_score | split5_test_score |
|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 14                 | -0.07             | -0.07             | -0.08             | -0.11             | -0.07             | -0.09             |
| 16                 | -0.07             | -0.07             | -0.08             | -0.11             | -0.07             | -0.09             |
| 18                 | -0.07             | -0.07             | -0.08             | -0.11             | -0.07             | -0.09             |
| 20                 | -0.07             | -0.07             | -0.08             | -0.11             | -0.07             | -0.09             |
| 22                 | -0.07             | -0.07             | -0.08             | -0.11             | -0.07             | -0.09             |

| split6_test_score | split7_test_score | split8_test_score | split9_test_score | mean_test_score | std_test_score | rank_test_score |
|-------------------|-------------------|-------------------|-------------------|-----------------|----------------|-----------------|
| -0.07             | -0.11             | -0.11             | -0.07             | -0.08           | 0.02           | 1               |
| -0.07             | -0.11             | -0.11             | -0.07             | -0.08           | 0.02           | 2               |
| -0.07             | -0.11             | -0.11             | -0.07             | -0.08           | 0.02           | 3               |
| -0.07             | -0.11             | -0.11             | -0.07             | -0.08           | 0.02           | 3               |
| -0.07             | -0.11             | -0.11             | -0.07             | -0.08           | 0.02           | 3               |



更新不動

# Grid Search

RMSE Train = 0.2898966457718158

RMSE Test = 0.35915528644669664

Coefficient

PLS Train: 0.4830827230103262

PLS Test: 0.3732551490907525

n components = 14 的結果



鄉鎮市區\_大安區 0.167762

屋齡byDay平方 0.134055

鄉鎮市區\_中山區 0.129583

鄉鎮市區\_松山區 0.122227

鄉鎮市區\_信義區 0.110485

鄉鎮市區\_中正區 0.107895

建物型態\_透天厝 0.066140

建物型態\_店面 0.061416

總樓層數平方 0.054553

鄉鎮市區\_士林區 0.050471

屋齡byDay -0.240747

交易標的\_建物 -0.088918

建物移轉總面積平方公尺 -0.034559

主要建材\_鋼筋混凝土造 -0.027232

鄉鎮市區\_南港區 0.028422

twse\_stock\_market 0.036292

土地移轉總面積平方公尺 0.038121

鄉鎮市區\_大同區 0.040174

車位總價元 0.044155

鄉鎮市區\_內湖區 0.046163

增加資料集：  
手邊已有資料已全部  
用到，無法再增加

優化模型超參數

2

1

優化方向

3

訓練資料增加閾值  
避免訓練雜訊

5

換一個模型試試看

4

簡化模型變數



# LSTM

資料有時間順序，LSTM  
專門針對序列模型，是否  
會有更好的結果？

建立4層LSTM  
三個DropOut皆設0.2

資料除了原先已取LOG的  
資料外，全部取LOG

有負數欄位採平移取LOG

Model: "sequential"

| Layer (type)        | Output Shape   | Param # |
|---------------------|----------------|---------|
| =====               |                |         |
| lstm (LSTM)         | (None, 1, 128) | 185344  |
| dropout (Dropout)   | (None, 1, 128) | 0       |
| lstm_1 (LSTM)       | (None, 1, 64)  | 49408   |
| dropout_1 (Dropout) | (None, 1, 64)  | 0       |
| lstm_2 (LSTM)       | (None, 1, 64)  | 33024   |
| dropout_2 (Dropout) | (None, 1, 64)  | 0       |
| lstm_3 (LSTM)       | (None, 1, 64)  | 33024   |
| dense (Dense)       | (None, 1, 1)   | 65      |
| =====               |                |         |

# LSTM

對所有資料按照年月日排序

Total params: 300,865

Trainable params: 300,865

Non-trainable params: 0

找最小MSE

Epoch = 30

Batch\_size = 2000

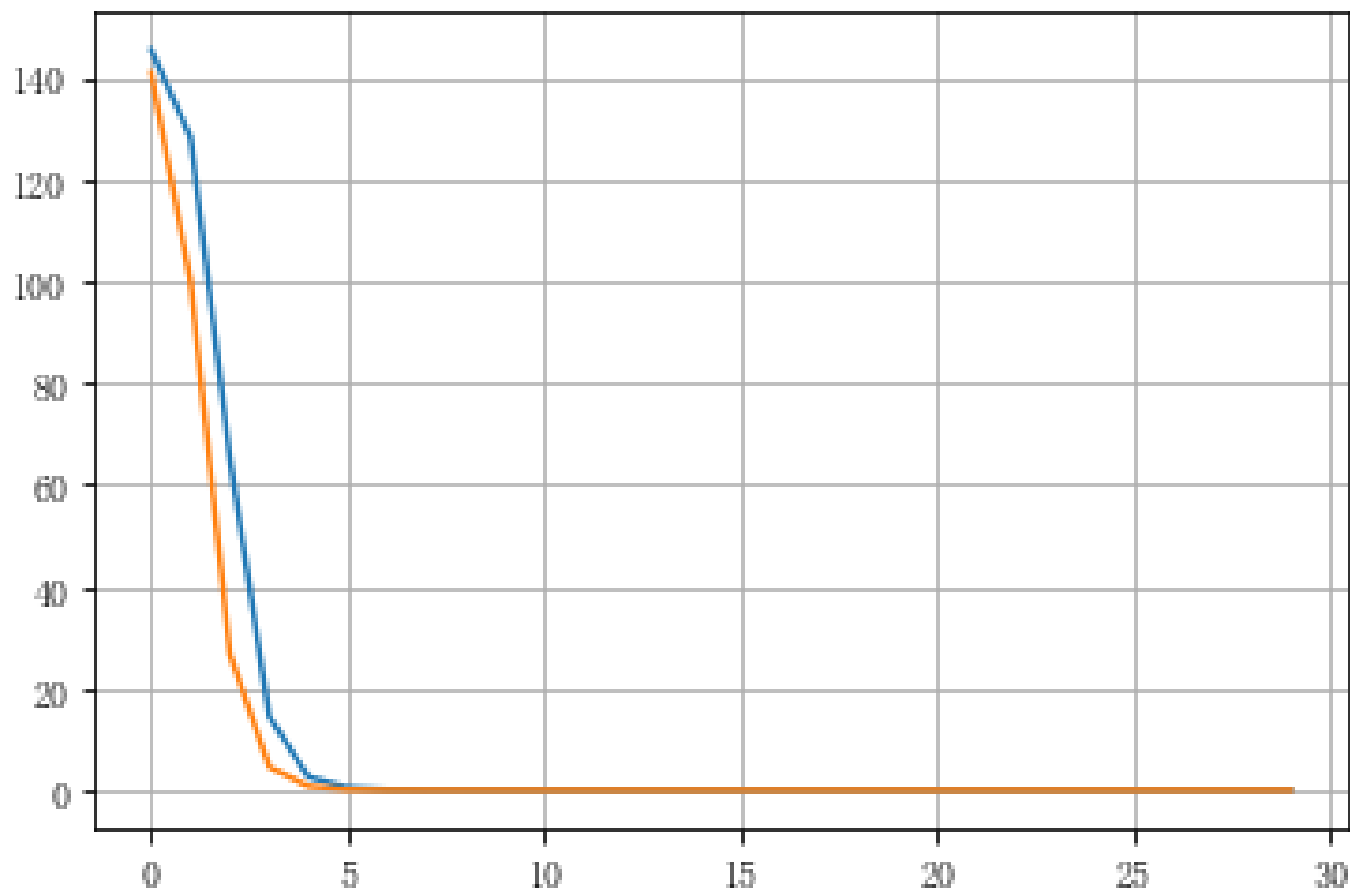
未避免訓練未來資料

直接用Train data 20%作為  
驗證資料

Lr = 0.01, decay = 1e-6

Train Score: 0.418265 RMSE

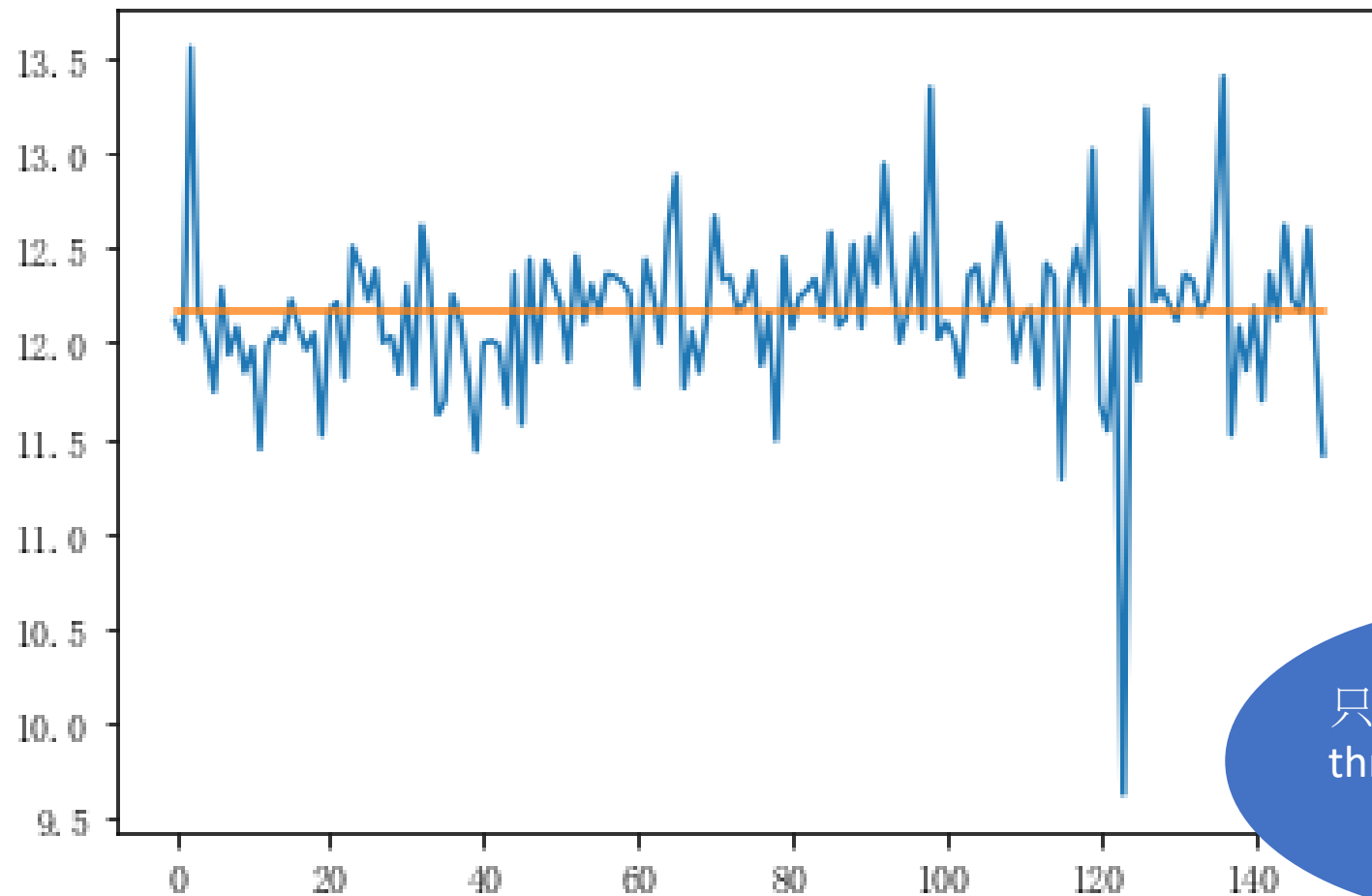
Test Score: **0.423383** RMSE





# LSTM

## First 150 Prediction



只考慮PLS without  
thresh 和縮減模型  
結果

# 目錄

**01**

目的和資料來源

資料視覺化

**02**

**03**

模型建置

優化模型

**04**

**05**

分析結果

學習心得

**06**

本次分析主要考量沒有Thresh後的相關性跟在此下的PLS模型，縮減模型作為輔助

台北房子單價主要是跟地段有關，其他關係影響並沒有想像中的深

在相關性矩陣中：離 Shopping mall 最短距離、離醫院最短距離、離學校最短距離、離MRT最短距離、離火車最短距離皆呈現極弱(大於-0.1)負相關性，但在PLS模型中MRT跟火車皆呈現極弱正相關性，其他維持弱負相關性，根據實務，個人比較偏好在PLS模型中呈現的結果，極弱相關很可能與台北通車便利有關

屋齡無論在相關性矩陣還是PLS模型中，皆呈現弱負相關性，代表房屋老舊對於台北的房子單價還是有影響力

比較令人感到奇怪的是建物移轉總面積、總樓層數這兩個在相關性矩陣為正向關係的欄位，在優化後的PLS模型皆呈現負向相關，甚至在縮減模型中，建物移轉總面積仍舊是負相關，如果以PLS模型為準，這可能隱含移轉總面積跟居住所蓋的樓層高低對於房價並沒有絕對的影響，甚至有可能是減項

股市對於房價也有正向拉抬的作用，這點可以從無論是相關性矩陣還是PLS模型皆為正向得到證實



交易標的為建物會對於房子每平方公尺單價造成負向相關，在相關性矩陣、縮減模型、未縮減模型中得到證實，這點我覺得蠻有趣的，可能是沒有涵蓋到土地？

# 最後，大安區真香

# 目錄

**01**

目的和資料來源

資料視覺化

**02**

**03**

模型建置

優化模型

**04**

**05**

分析結果

學習心得

**06**

由於這次是一個比較大型的專案，在過程中需要耗費大量的記憶體，因此架構的好壞非常重要，前期有大半時間都在切資料跟優化架構，很多時候程式在小資料程序中是可以的，但當跑大型資料程序時，就會報很多錯，今年又剛好碰到大停電事件，基於程式碼並沒有匯出機制，致使中間過程爬取的資料全數遺失，當然，也很感謝這次的事件，讓我重新優化程式架構。

一直以來都很想知道台北房價是不是就如傳統認知一般，跟火車站、**MRT**、醫院這些距離是有相關性，是否有決定房價的關鍵因子，由於單就總房價不夠客觀，所以，在這次分析中只以單價元平方公尺來做衡量，此外，針對類別型變數，由於大多都是無序的，因此全部都採用**one-hot**轉換，當然也有做一些其他處理，像是對於價格取**log**。

就這次的結果來看，火車站、**MRT**雖然是正相關，但仍非決定性因子，對於價格最直接的因素還是地段，而在醫院、**Shopping mall**甚至是學校皆呈現負向相關，這些結果可能說明了台北在社會福利還有消費上的便利性。

以前比較常聽到**OLS**，但在這次的專案中碰到**PLS**，對於我來說是新的模型，也很開心能夠藉此碰觸到**PLS**理論跟用法

原本就有計畫在今年做一個比較大的專案，恰巧碰到這次課程需要，順便複習一下程式語法，對我來說收穫頗豐，謝謝教授