

# Data Science

探討多面向因素的房價預測- 以台北市為例

---

Group 4

R26121073 陳沛群 (組長)

RE6121029 張立勳

RE6121037 李易庭

RE6124035 黃亮臻

Instructor: 鄭順林 教授



# Target Problem, Importance & Novelty

- 以不動產資訊、環境機能、總體經濟三大面向來探討房價與變數間的關係，並試圖以此建構成一房價預測模型。
- Who Matters: 有購屋意願並有能力購屋的人
- Why: 利用房價預測模型預測房價未來走勢，作為不動產交易的重要依據

# Outline

- Literature Review
- Data Description
- Data Visualization
- Data Preprocessing
- Model Fitting and Prediction
- Conclusions and Future Work



# Literature Review

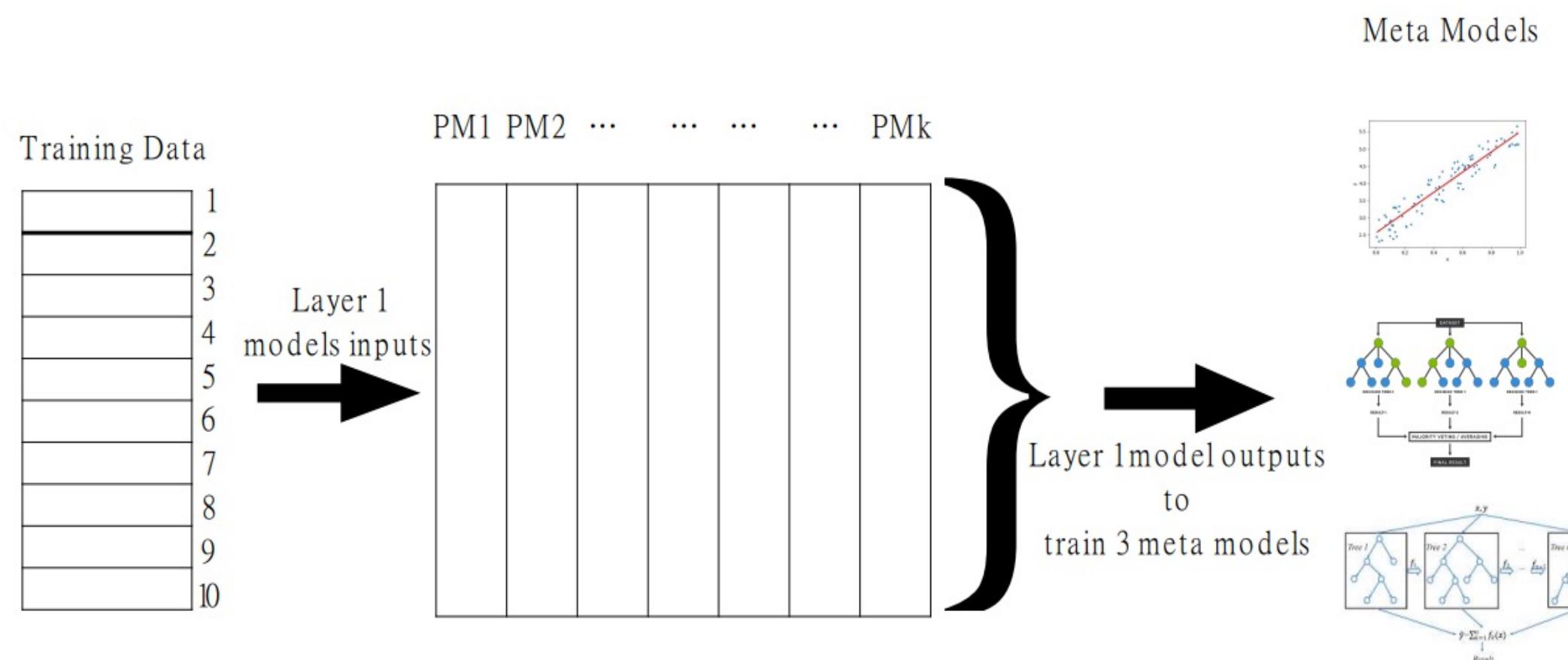


# 應用時價登陸已聚類方法之堆疊泛化房價預測模型

## 黃允亭, 政治大學經濟學系研究所

- 目的：使用交叉驗證之遞迴特徵消除法，挑選特徵後，再進行預測。
- 方法：Stacking

第一層用Lasso、KNN、Decision Tree，第二層用XGBoost、Random Forest、Linear Regression，並使用到k-means分群建模，減少房價的異質性。



基模型	評估標準	元模型		
		Linear Regression	RandomForest	XGBoost
KNN	MAPE	0.1245	0.1321	0.1257
	MSE	74046200	78007407	73939563
	R2_score	0.7551	0.7420	0.7555
Lasso	MAPE	0.2033	0.2052	0.2032
	MSE	164871765	173555606	168558798
	R2_score	0.4548	0.4261	0.4426
Decision Trees	MAPE	0.1036	0.1062	0.1075
	MSE	51670047	52461190	53480275
	R2_score	0.8291	0.8265	0.8231

# Duan et al. (2021) Addressing the macroeconomic and hedonic determinants of housing prices in Beijing Metropolitan Area ,China

- 目的：考量總體經濟因素與享樂性(Hedonic)因素，探討影響北京房價的原因
- 方法：向量自我回歸模型(VAR) & 地理加權回歸模型(GWR)
- 結果：

VAR模型：貨幣供應量對長期房屋價格有正面影響，貸款利率有反面影響，而消費者物價指數對房價動態沒有顯著影響。

GWR地理加權迴歸模型：享樂性(Hedonic)因素的影響在不同地理位置有顯著的差異，  
 $R-squared=0.654$  (只考慮享樂性因素，沒有綜合考慮總經變數)

- All of the variables are stationary. VAR model Variance Decomposition:

貨幣供給: 20.2% 貸款利率: 15.6% 消費者物價指數: 4.0% 平均房價: 60.1 %)

# 以總經變數預測不動產價格之模型

## 吳岱蓉, 清華大學計量財務金融學系研究所

$$ATE_{it} = \gamma_{0it} + \gamma_{1it}ATE_{it-1} + \gamma_{2it}ATE_{it-2} + \gamma_{3it}M1B_{it-1} + \gamma_{4it}CPI_{it-1} + \gamma_{5it}CCI_R_{it-1} + \gamma_{6it}D(UMI_{it-1}) + \gamma_{7it}D(CL_{it-1}) + \varepsilon_{it}$$

ATE<sub>it</sub>為第i個都市在第t期的平均處理效應

$\gamma_{0it}$ 為截距項

ATE<sub>it-1</sub>與ATE<sub>it-2</sub>分別為第i個都市在第t期的落後一期及二期的平均處理效應

M1B<sub>it-1</sub>為第i個都市在第t期落後一期的對數貨幣供給額

CPI<sub>it-1</sub>為第i個都市在第t期落後一期的對數消費者物價指數

CCI\_R<sub>it-1</sub>為第i個都市在第t期落後一期的對數營建股價指數報酬率

D(UMI<sub>it-1</sub>)為第i個都市在第t期取差分後之落後一期的對數失業率

D(CL<sub>it-1</sub>)為第i個都市在第t期取差分後之落後一期的對數建築貸款餘額

表 3 台北市房價指數模型實證結果

變數	係數	標準差	T 統計量	P 值
C	0.405774	0.93868	0.432281	0.6676
TAIPEI(-1)	0.703159	0.145517	4.832154	0***
TAIPEI(-2)	0.173524	0.148494	1.168561	0.2489
CCI_R(-1)	-0.00033	0.000417	-0.79523	0.4307
CPI(-1)	-0.27942	0.620461	-0.45034	0.6547
M1B(-1)	0.022285	0.104007	0.214268	0.8313
D(UMI(-1))	-0.23662	0.278014	-0.8511	0.3993
D(CL(-1))	1.434686	0.918654	1.561726	0.1255
R-squared	0.913351	Adjusted R-squared	0.899566	

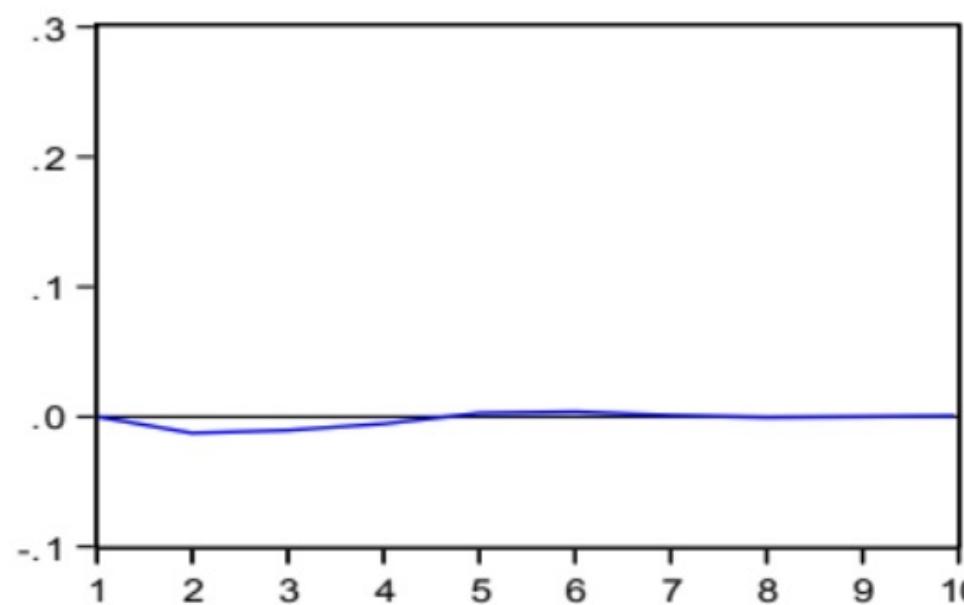
# 房貸利率、新成屋房價與建商股價關聯性之研究

## 蕭曉文, 中正大學財務金融研究所

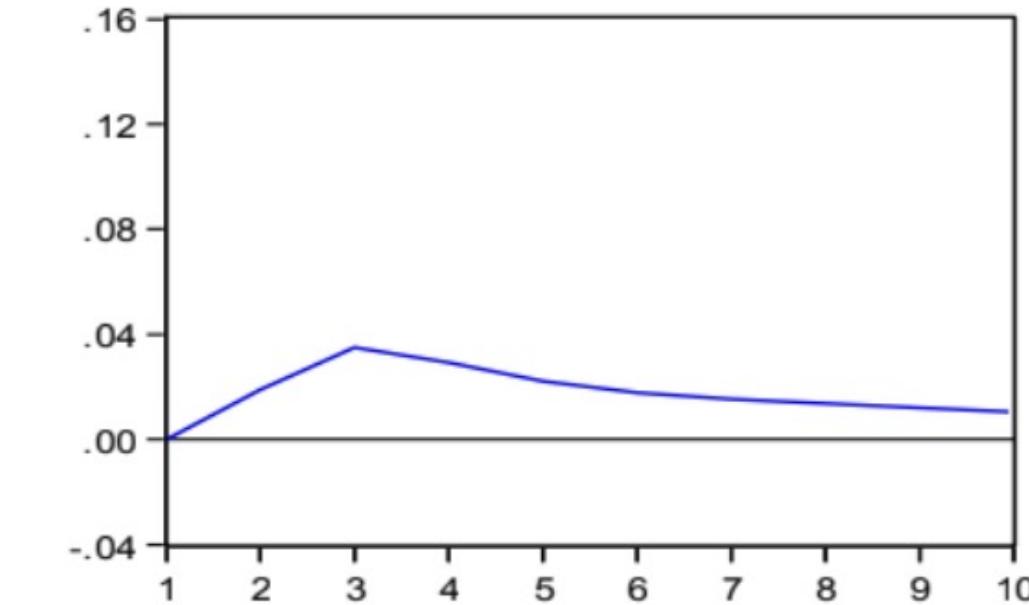
- 目的：探討房貸利率、新成屋房價與建商股價的關聯性
- 方法：針對全國、高雄、台北三個地方以代表性建商股價、房價、以及五大銀行平均房貸利率建立自回歸(VAR)模型
- 結果：做Granger因果關係檢定、衝擊反應分析如下

4.3.2.c 應變數：台北市房價變動率

	Chi-sq	p-value
華固股價變動率	0.800699	0.6701
房貸利率	10.87452	0.0044



華固股價變動對於台北市房價之衝擊反應



利率變動對台北市房價之衝擊反應

# Data Description & Visualization



**總體經濟**

**不動產資訊**

**環境因素**

**房價**

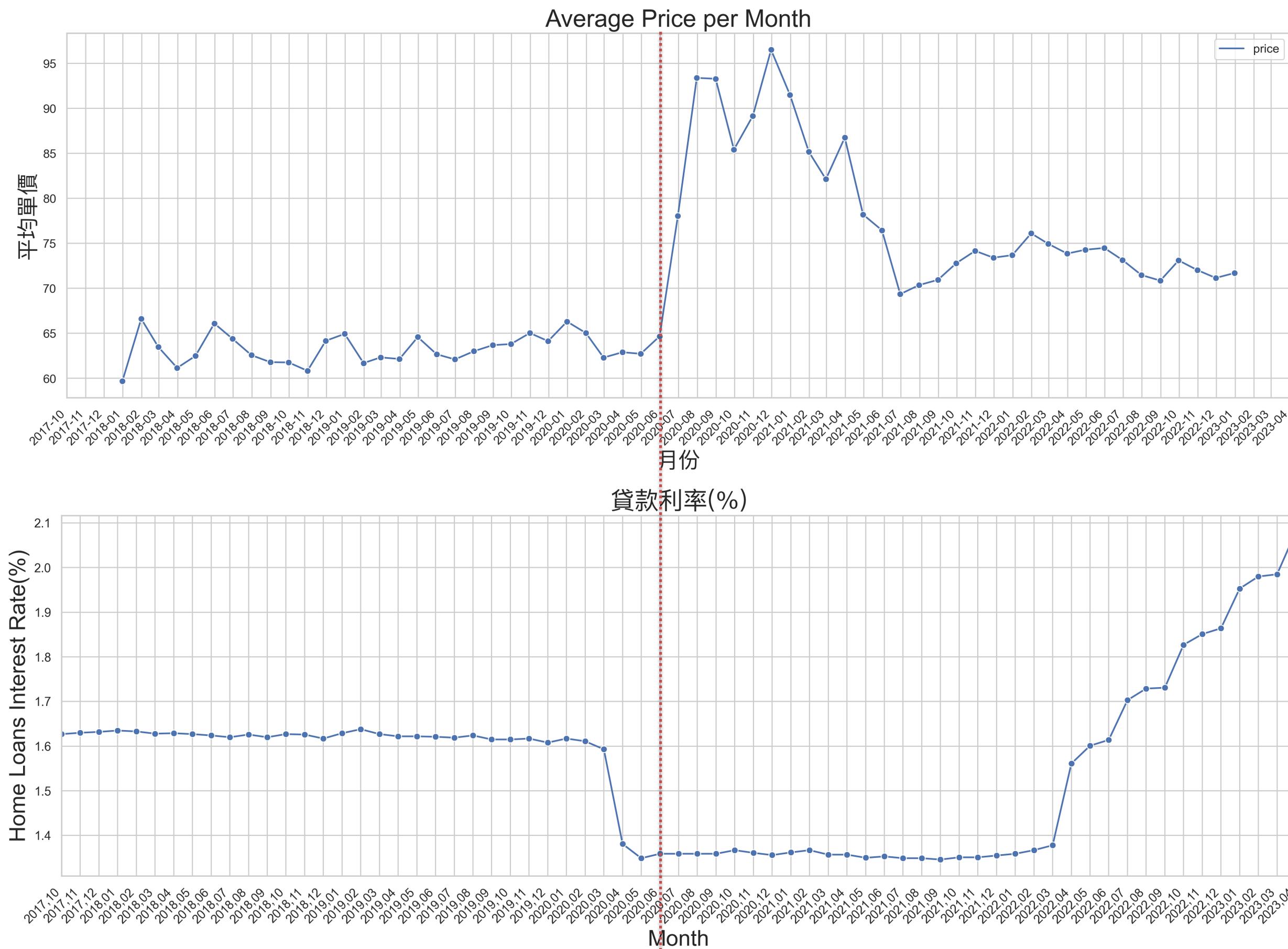
**(萬/坪)**

# 總體經濟資料

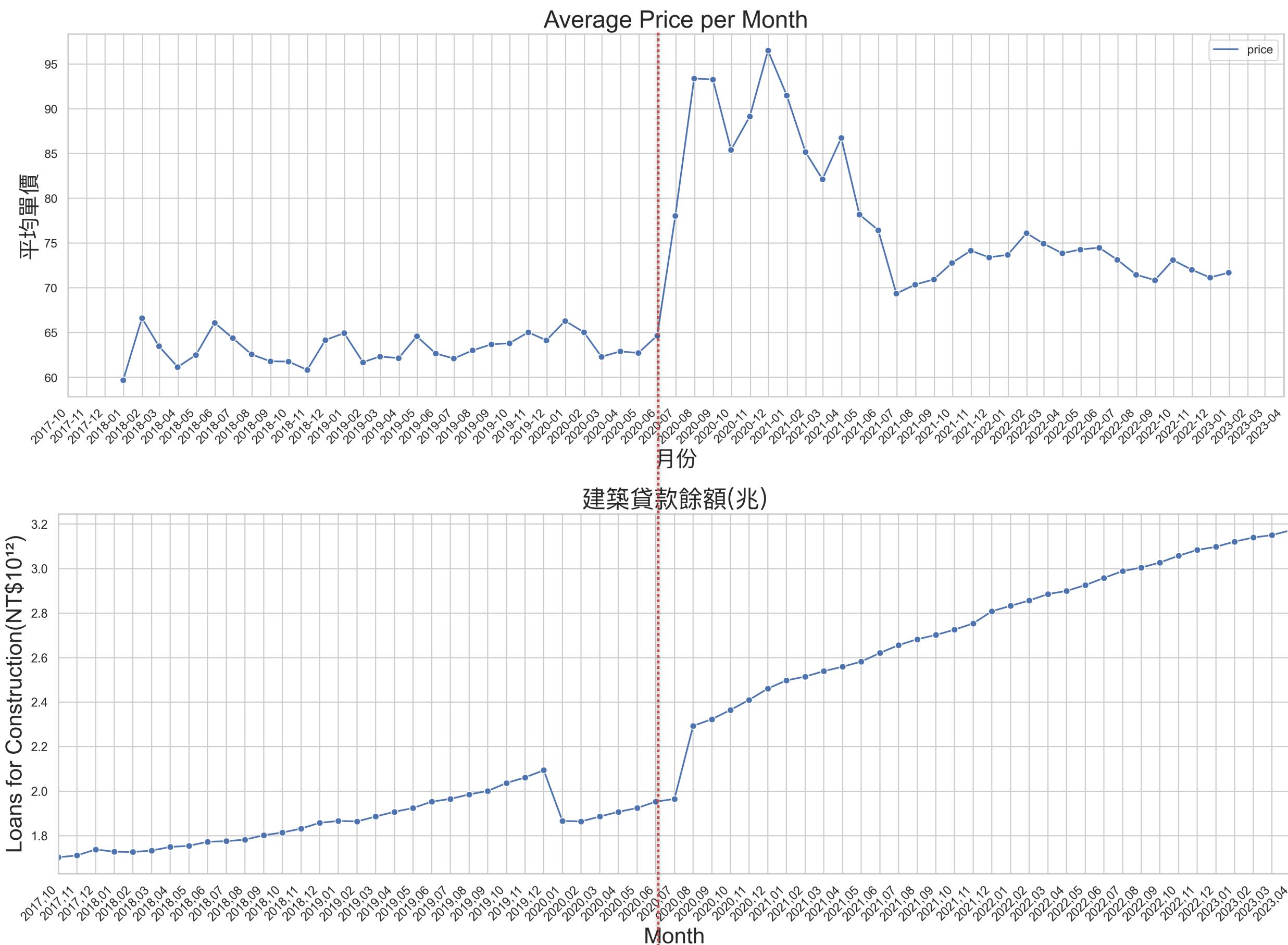
- 貨幣供給額：採用中央銀行定義之M1B當作指標
- 建築貸款餘額：建築業、從事建築投資之個人，因承做房屋興建投資所辦理之購地、興建房屋及周轉金貸款
- 房貸利率：五大銀行(台灣銀行、合作金庫銀行、土地銀行、華南銀行及第一銀行)新承做放款金額與利率
- 消費者物價指數：衡量一般家庭購買消費性商品及服務之價格水準變動情形
- 失業率：失業人口佔勞動人口之比例

資料來源：中華民國中央銀行、行政院主計處

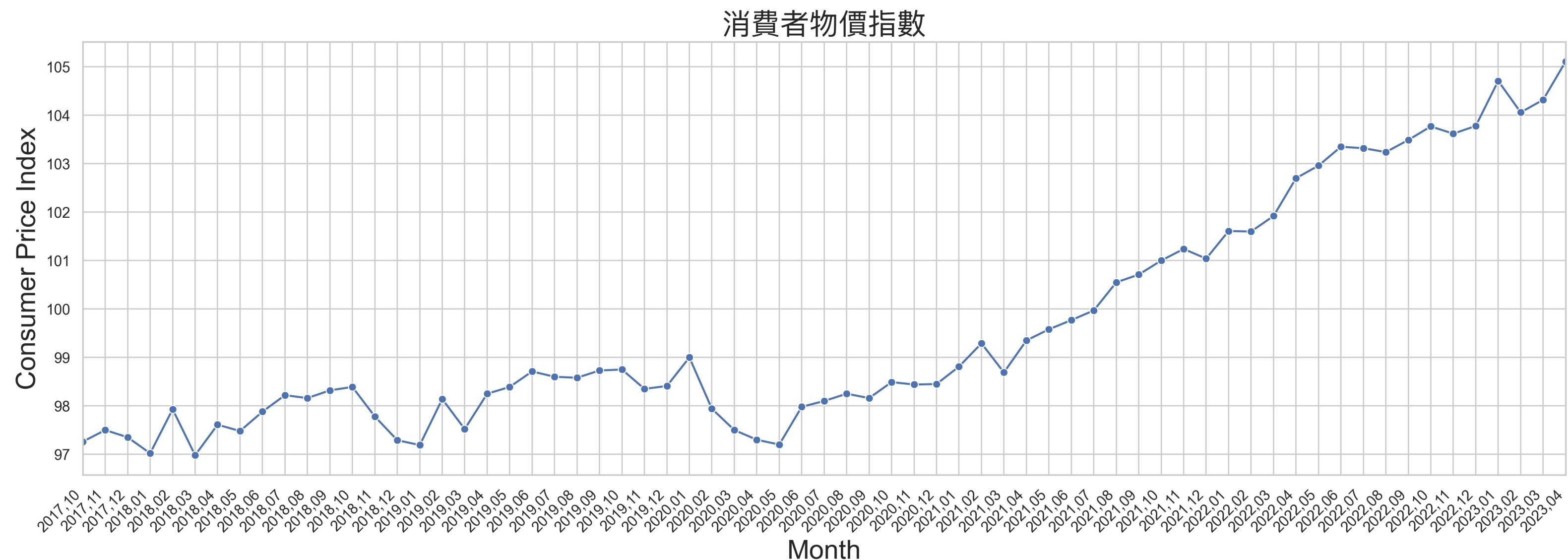
## • 每月平均單價(萬) vs 貸款利率



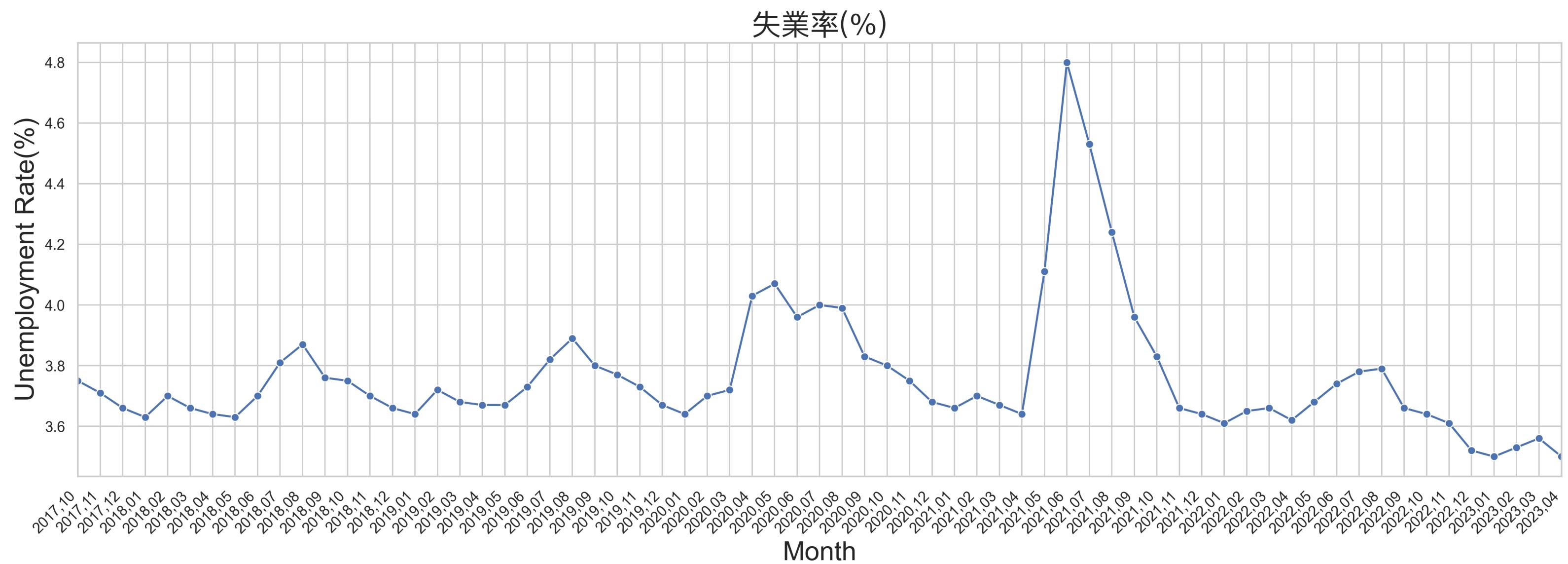
## • 每月平均單價(萬) vs 建築貸款餘額(兆)



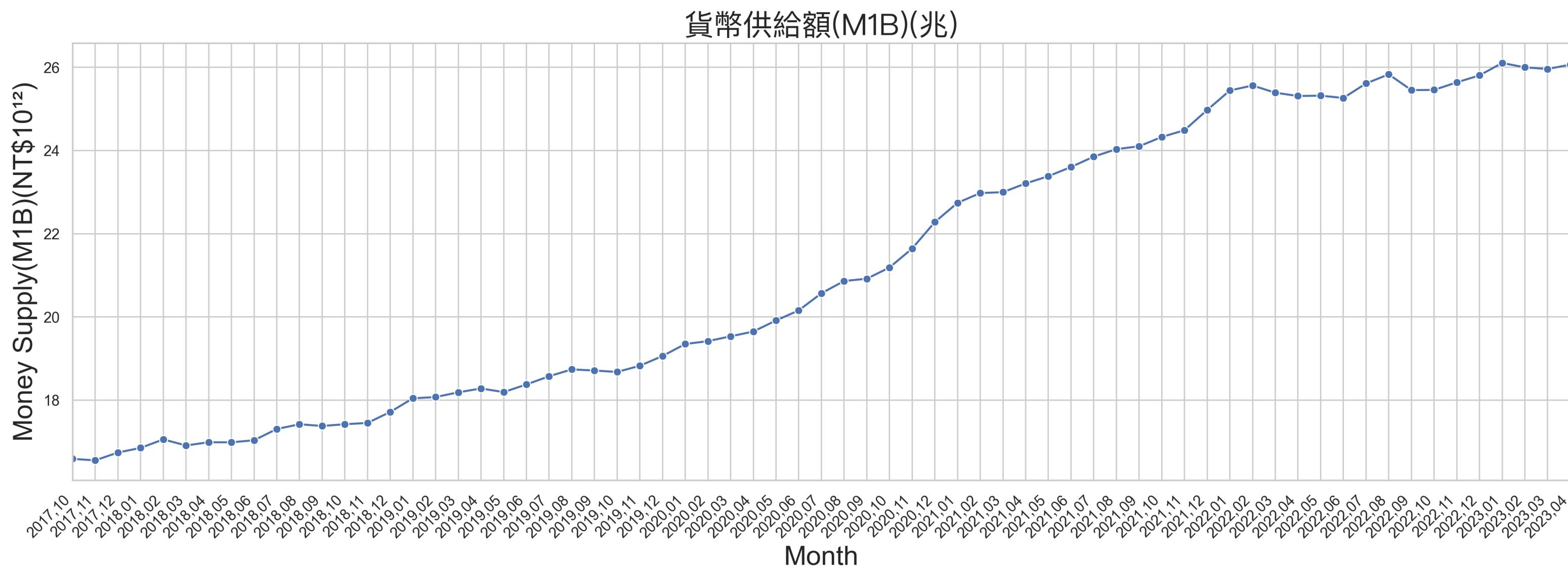
- 消費者物價指數趨勢圖



- 失業率趨勢圖



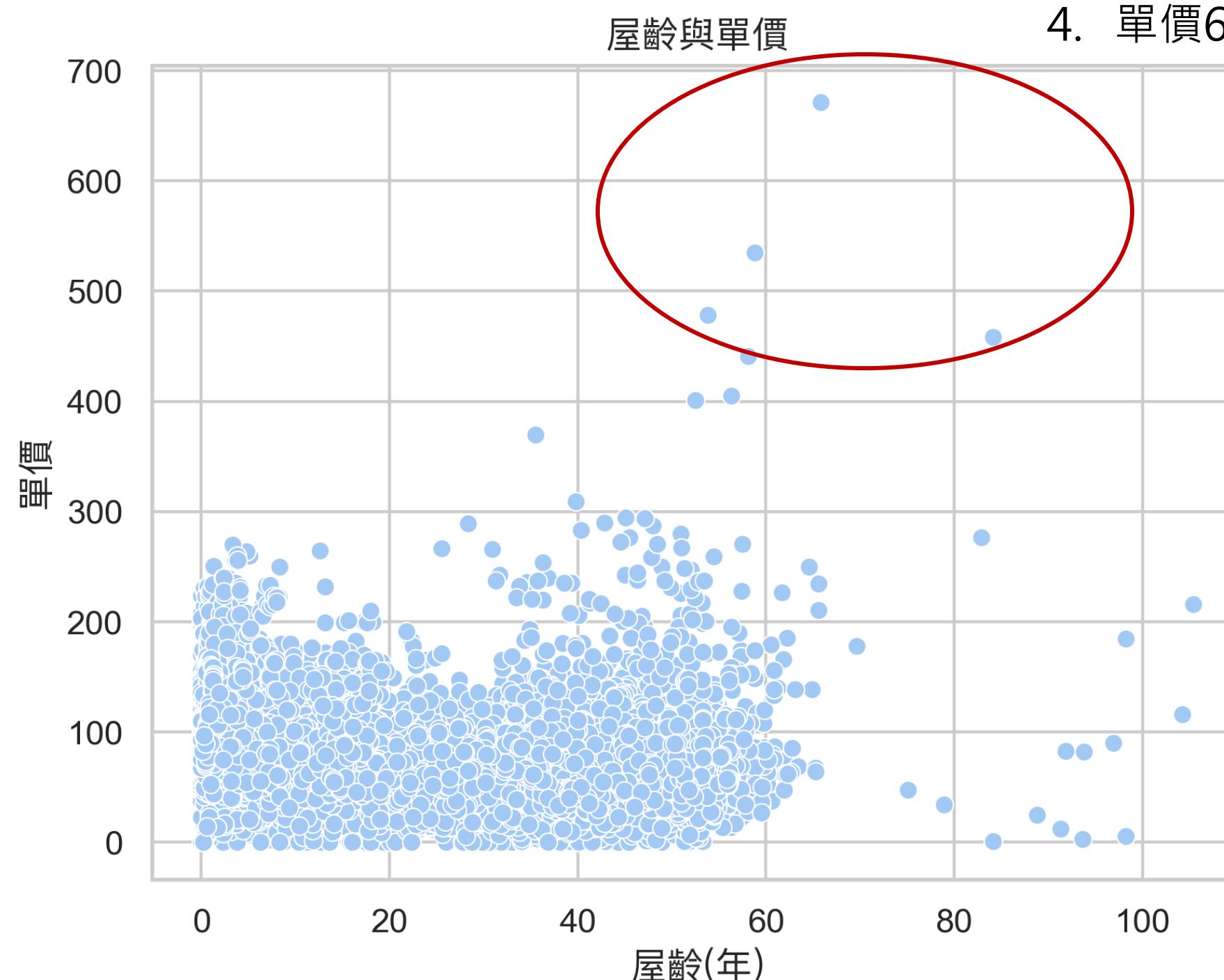
- 貨幣供給額(M1B)(兆)趨勢圖



# 實價登錄資料

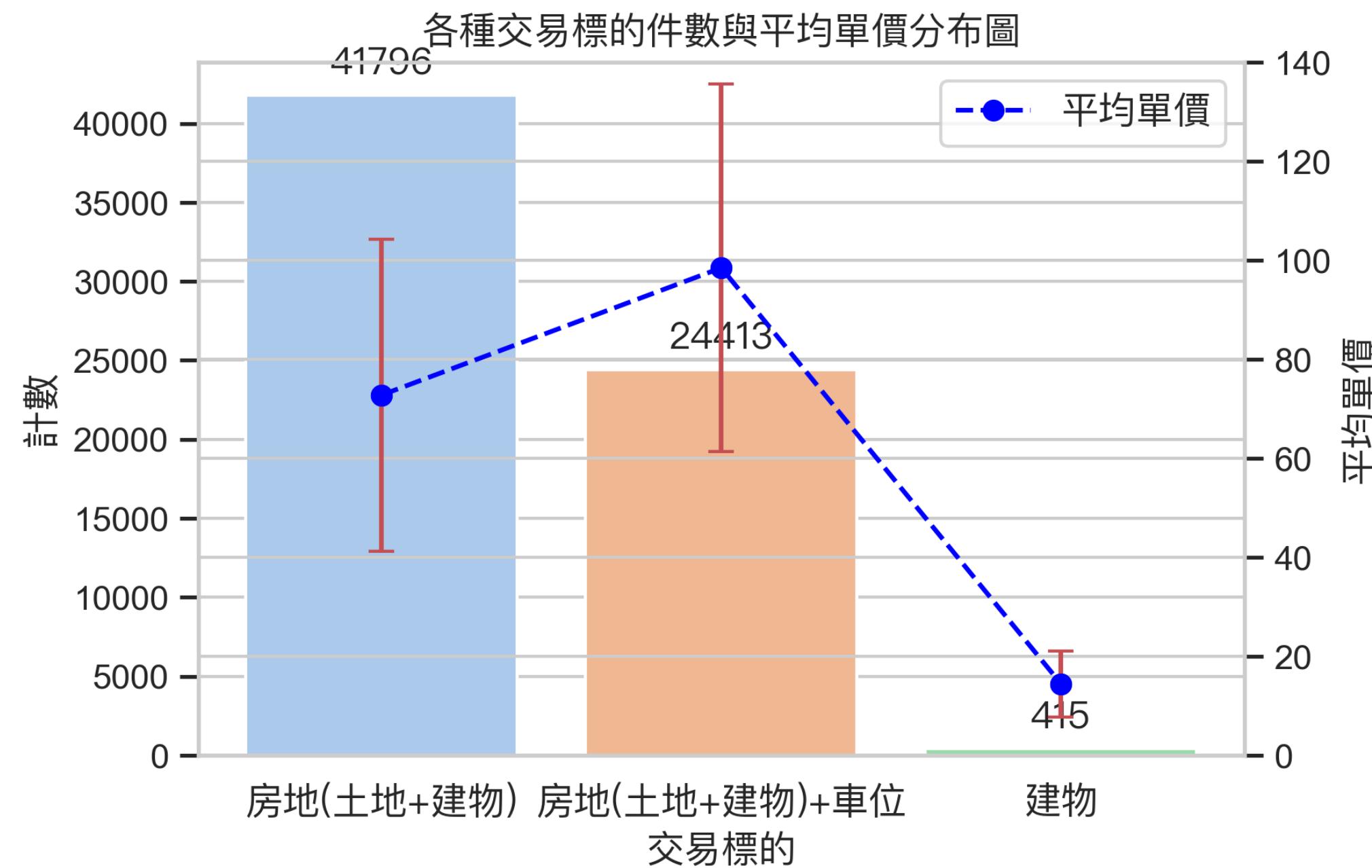
- 基本資訊：交易日、交易類別、屋齡、經/緯度
- 計數：房地(土地、建物、車位)、格局、樓層
- 面積：土地移轉坪數、建物移轉坪數、車位總面積(坪)
- 總價：車位總價(萬)
- 類別變數：主要建材、交易類別、建物型態、都市土地使用分區、主要用途、鄉鎮市區
- 有無：管委會、臨路、頂樓
- 備註：毛胚屋、特殊交易關係

## • 屋齡與單價

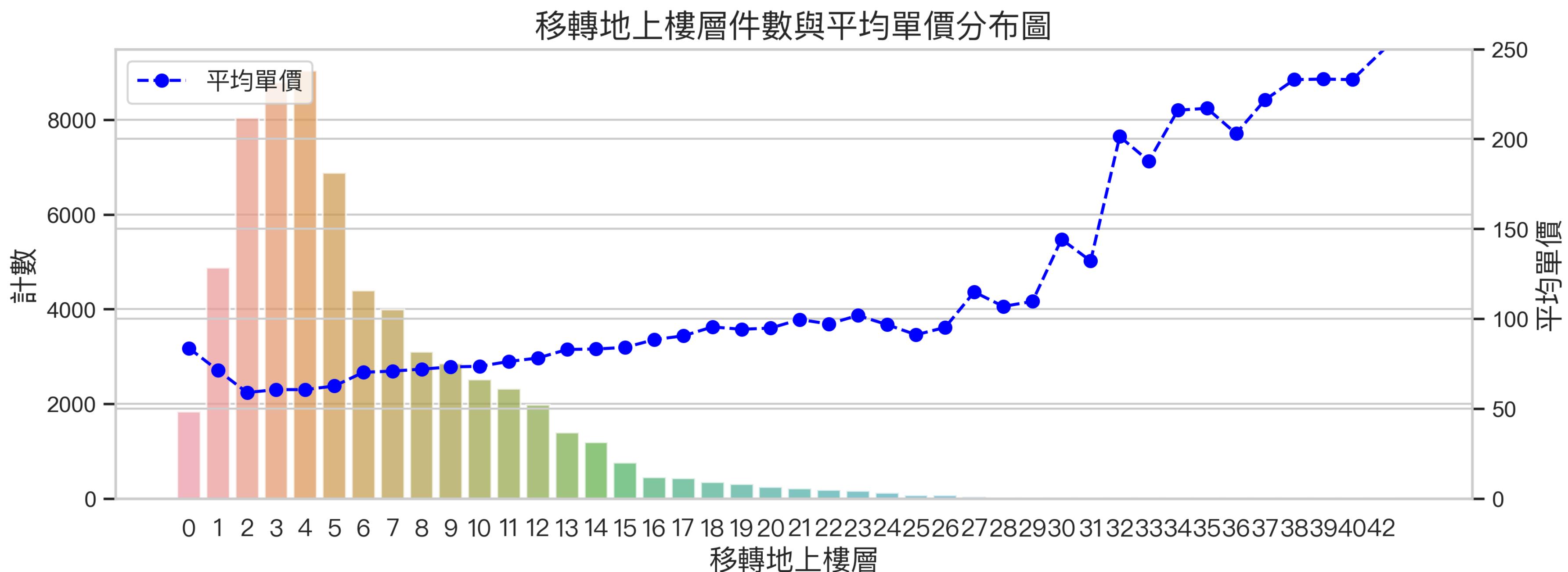


1. 單價458.6: 士林區, 透天厝, 磚造
2. 單價478.6: 士林區, 透天厝, 加強磚造
3. 單價534.0: 中山區, 透天厝, 加強磚造
4. 單價671.2: 中山區, 透天厝, 加強磚造

- 房地與平均單價分布圖

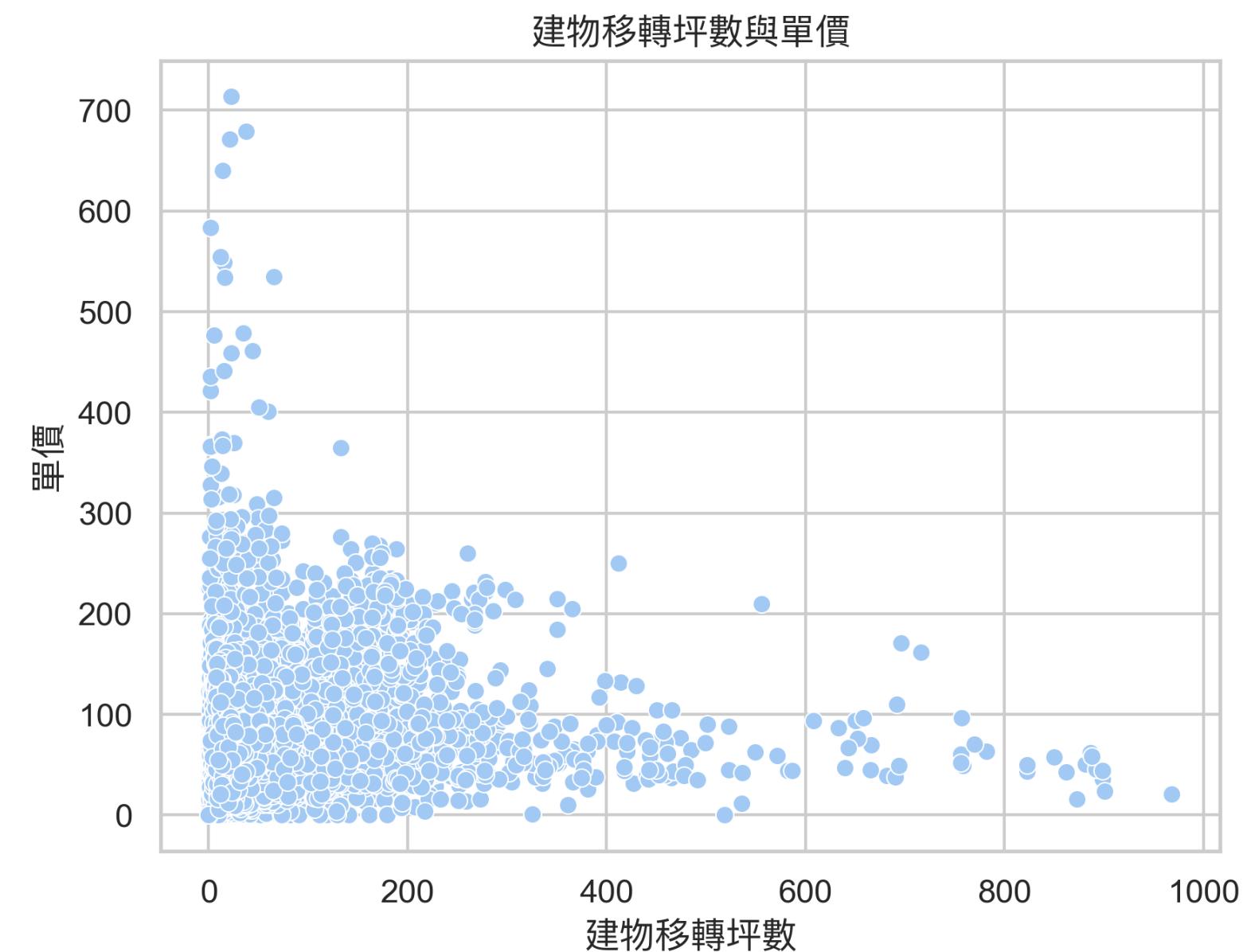
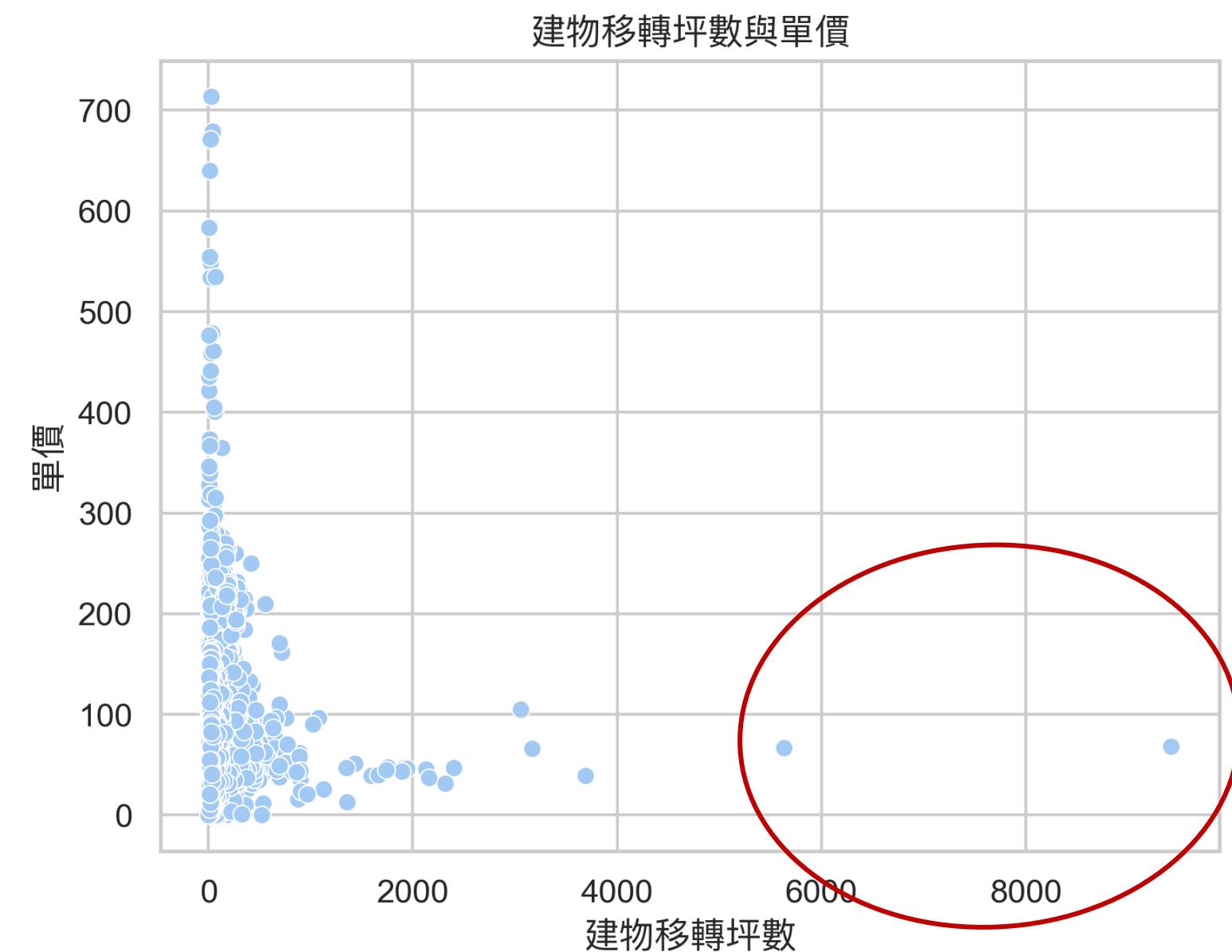


- 移轉地上樓層件數與平均單價分布圖



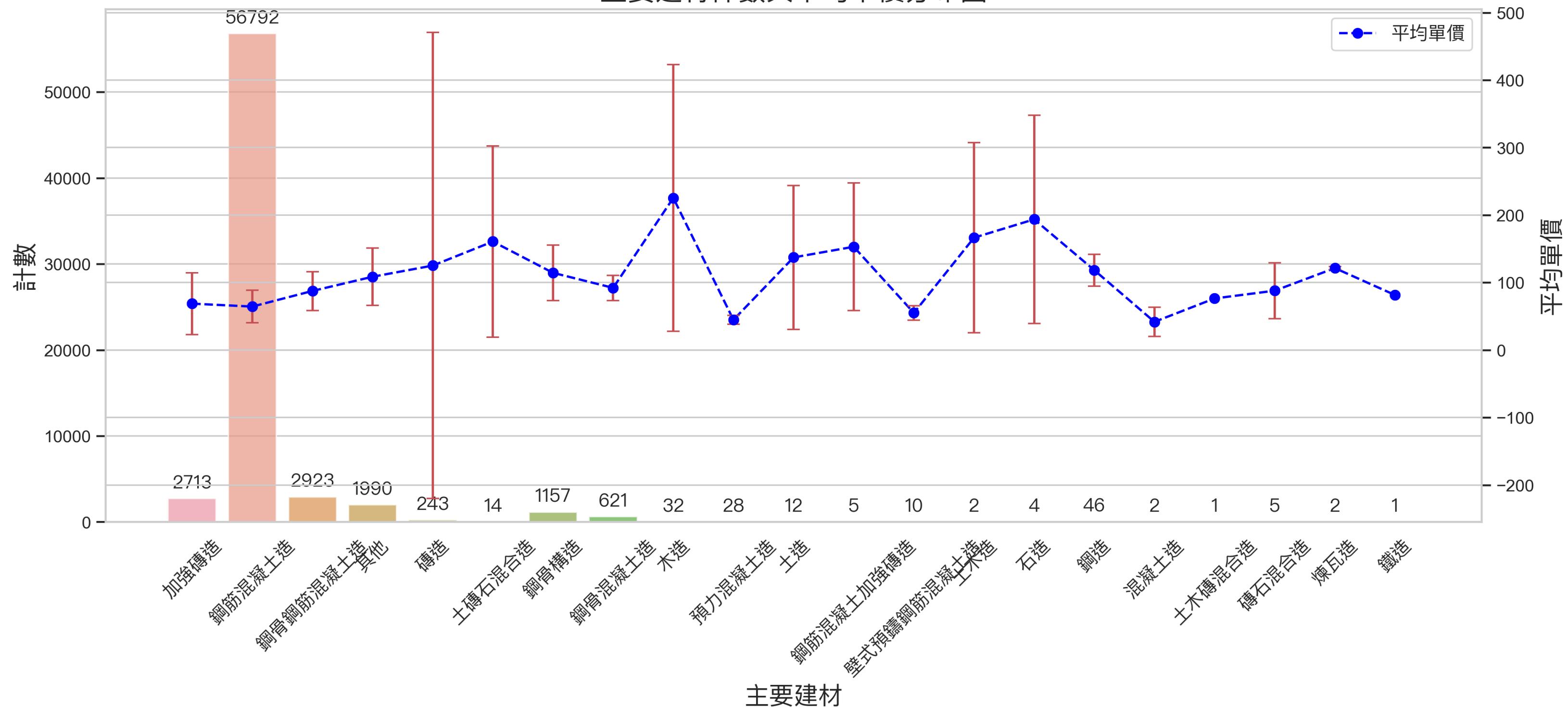
## • 建物移轉坪數與單價

1. 建物移轉坪數9423: 辦公商業大樓
2. 建物移轉坪數5630: 辦公商業大樓, 停車空間



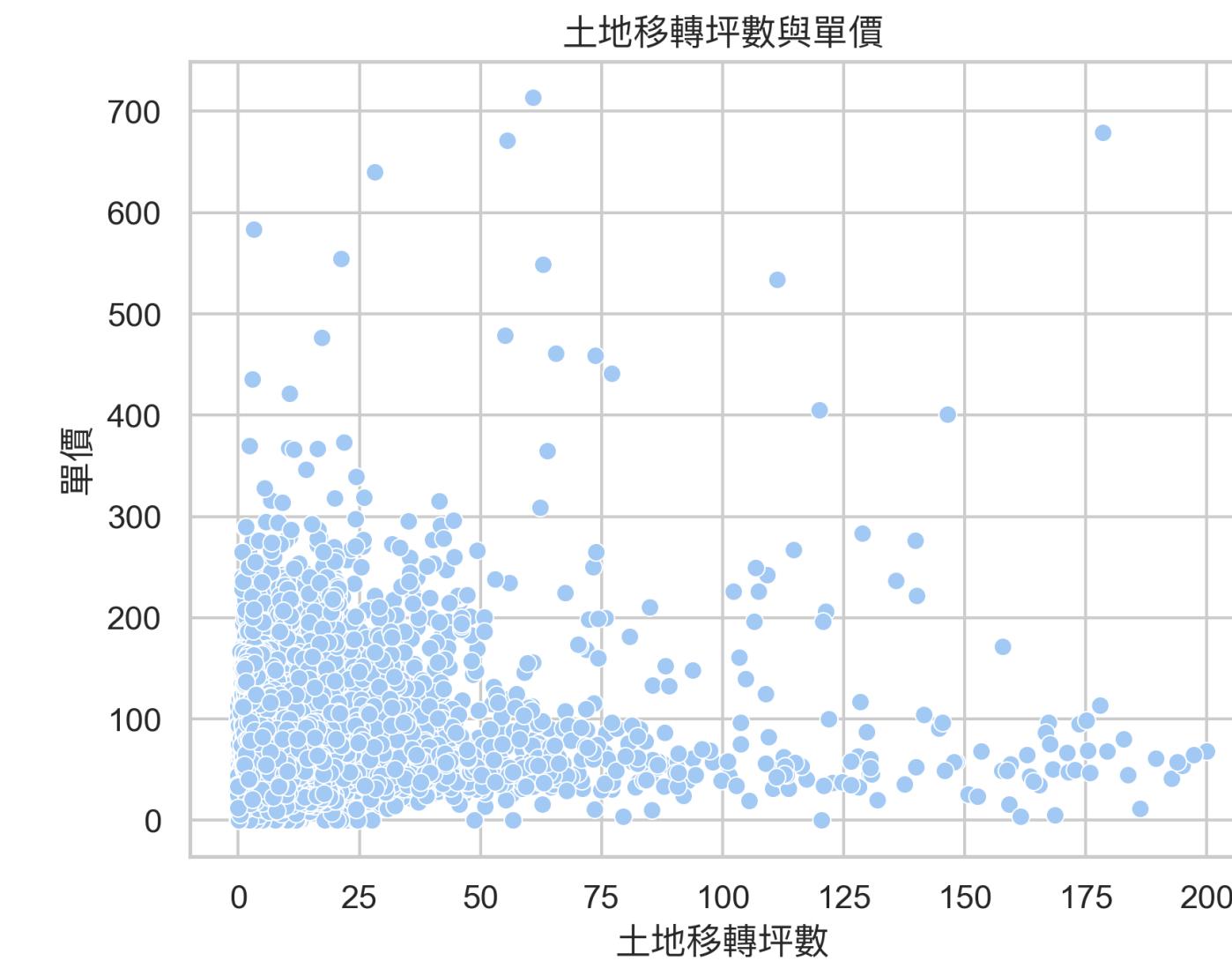
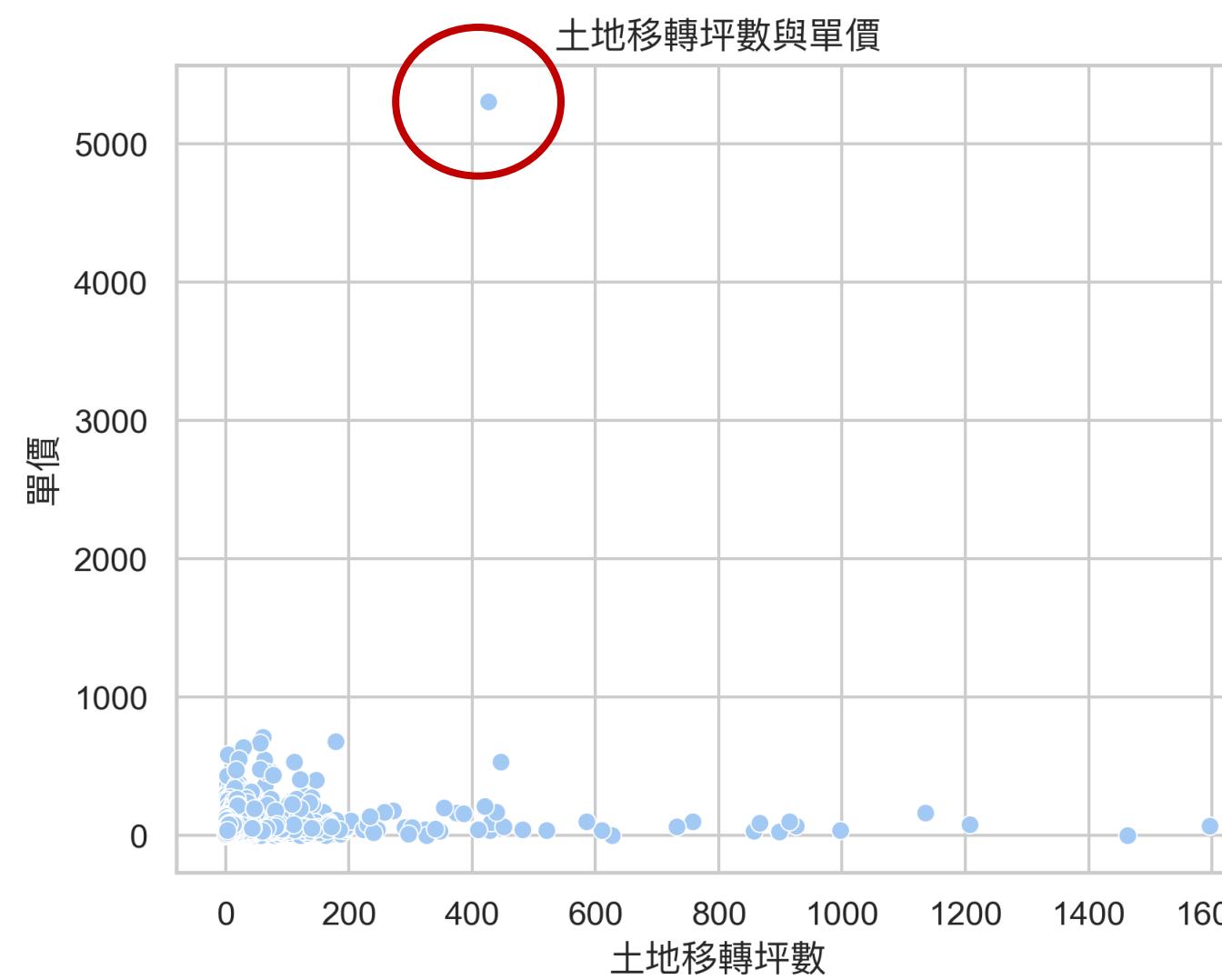
## • 主要建材與平均單價分布圖

主要建材件數與平均單價分布圖



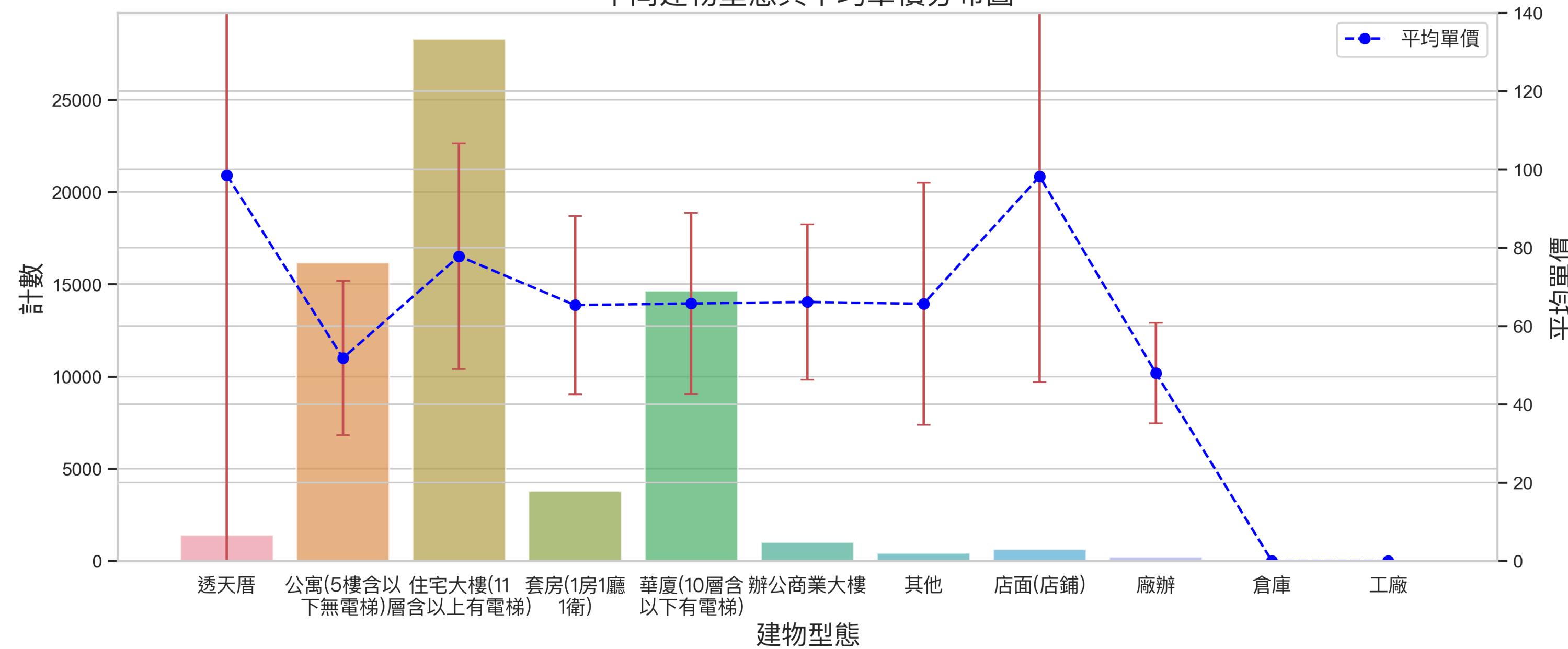
## • 土地移轉坪數與單價

屋齡84年 11坪透天厝 5筆土地交易 磚造 特殊關係交易

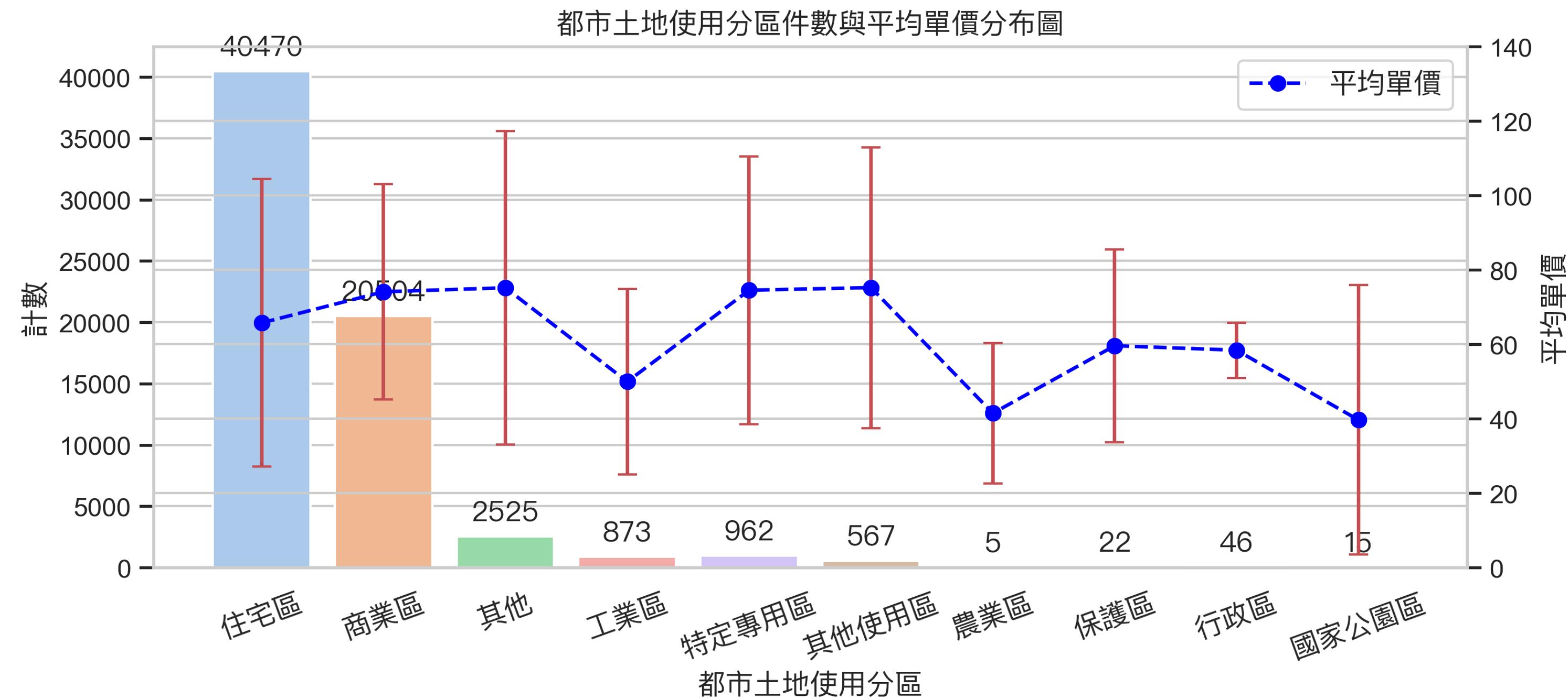


## • 建物型態與平均單價分布圖

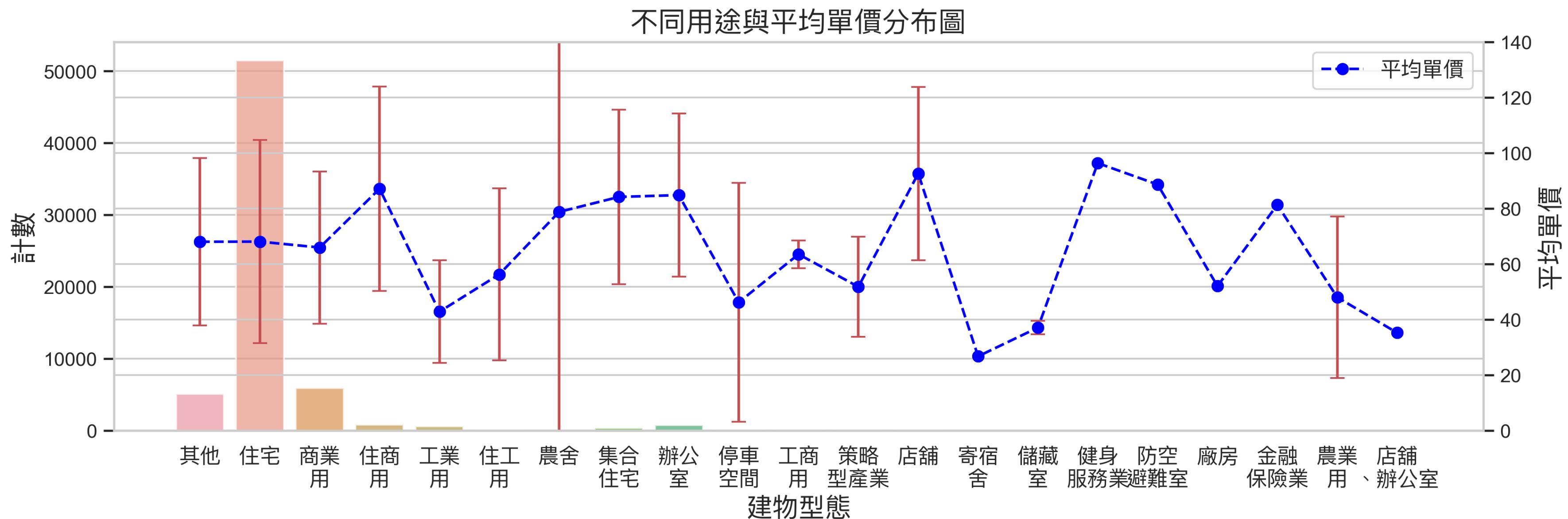
不同建物型態與平均單價分布圖



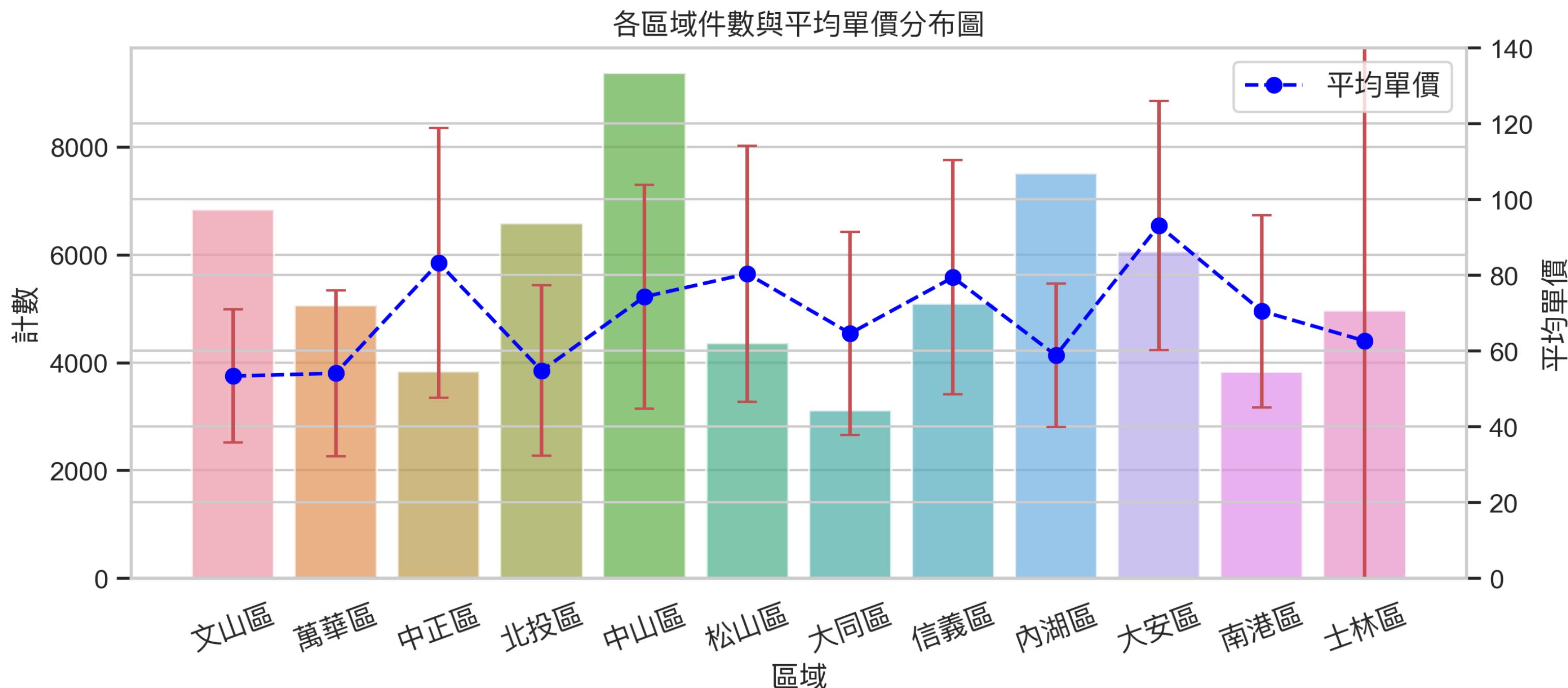
- 都市土地使用分區件數與平均單價分布圖



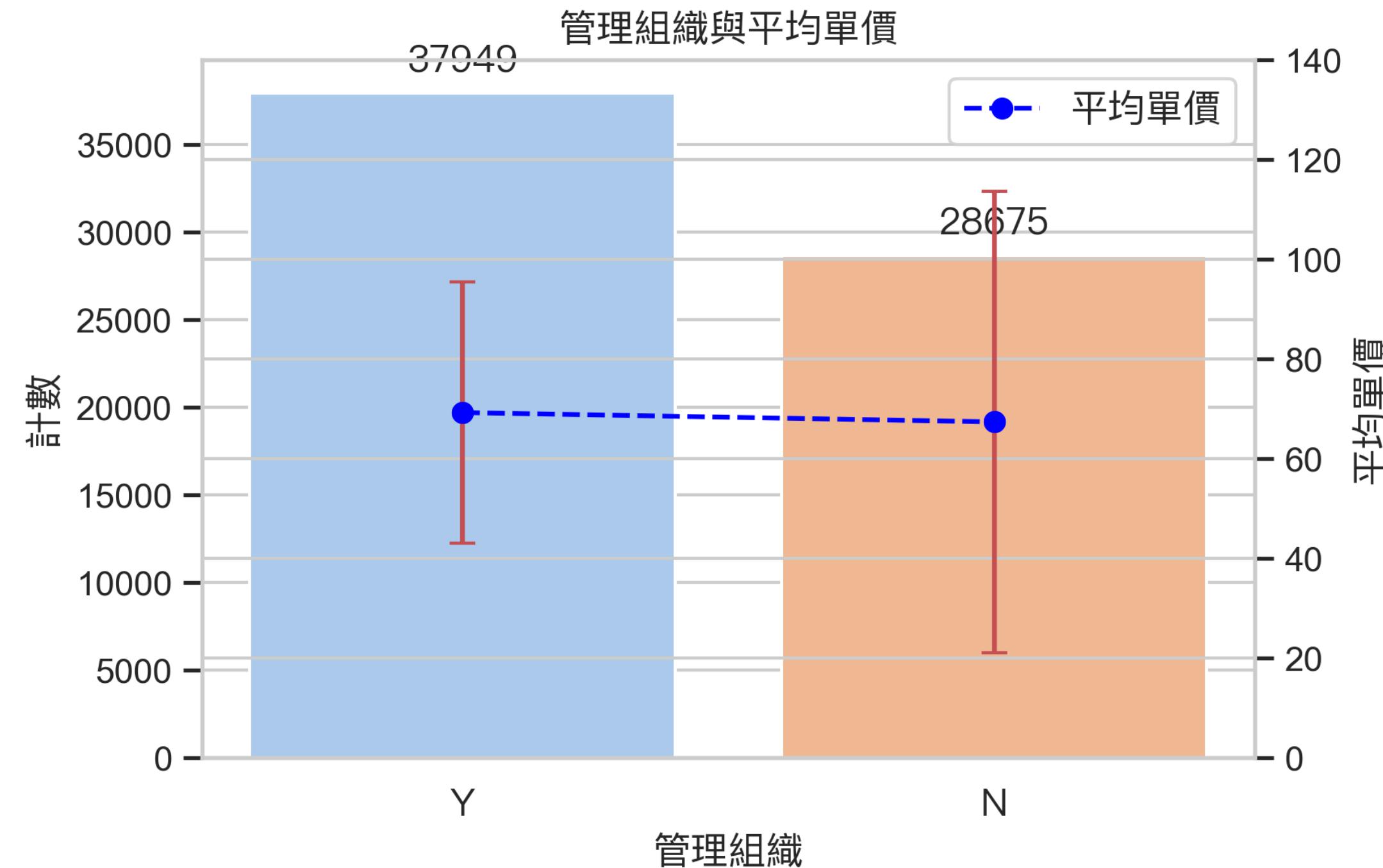
## • 主要用途與平均單價分布圖



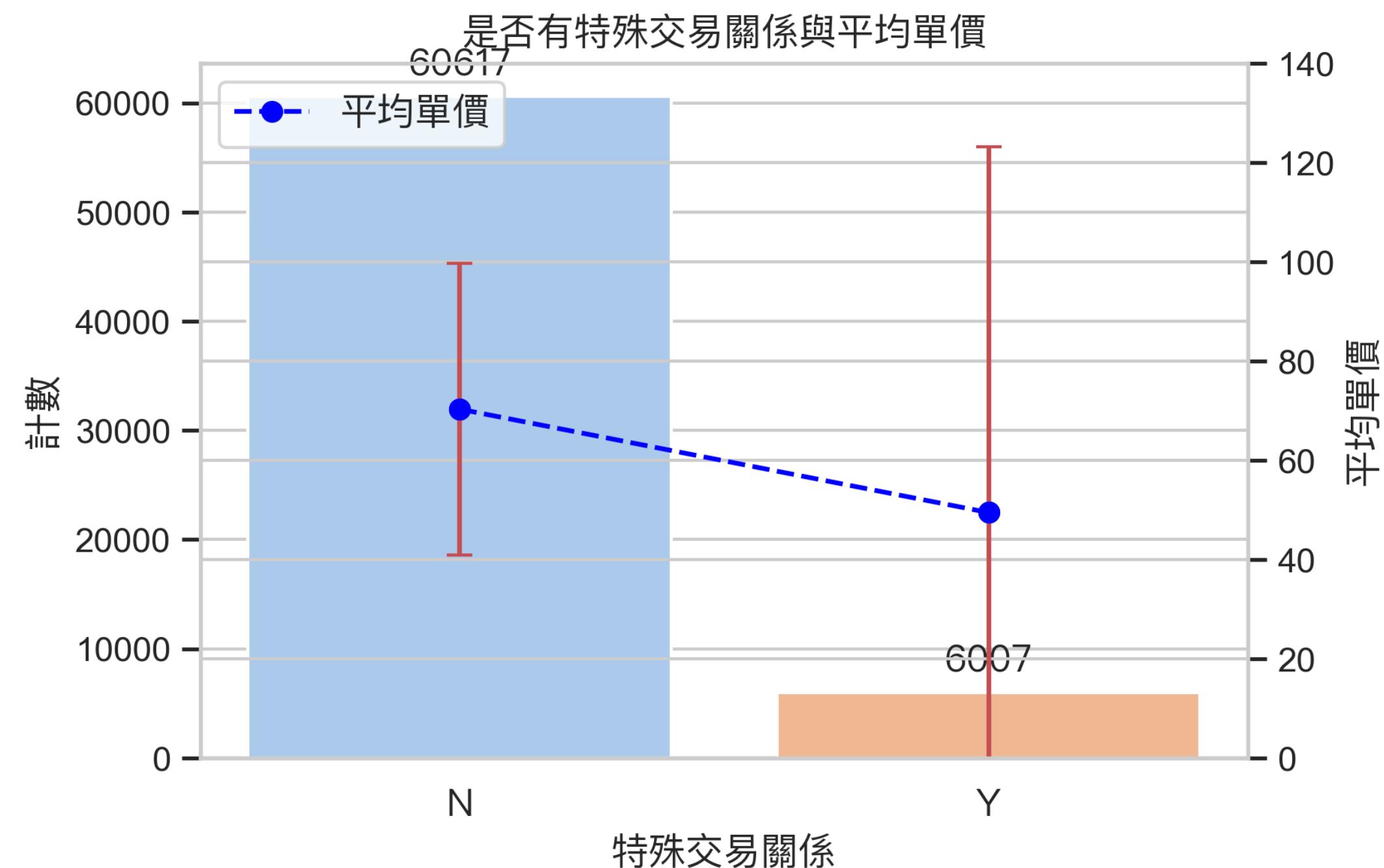
- 各區域件數與平均單價分布圖



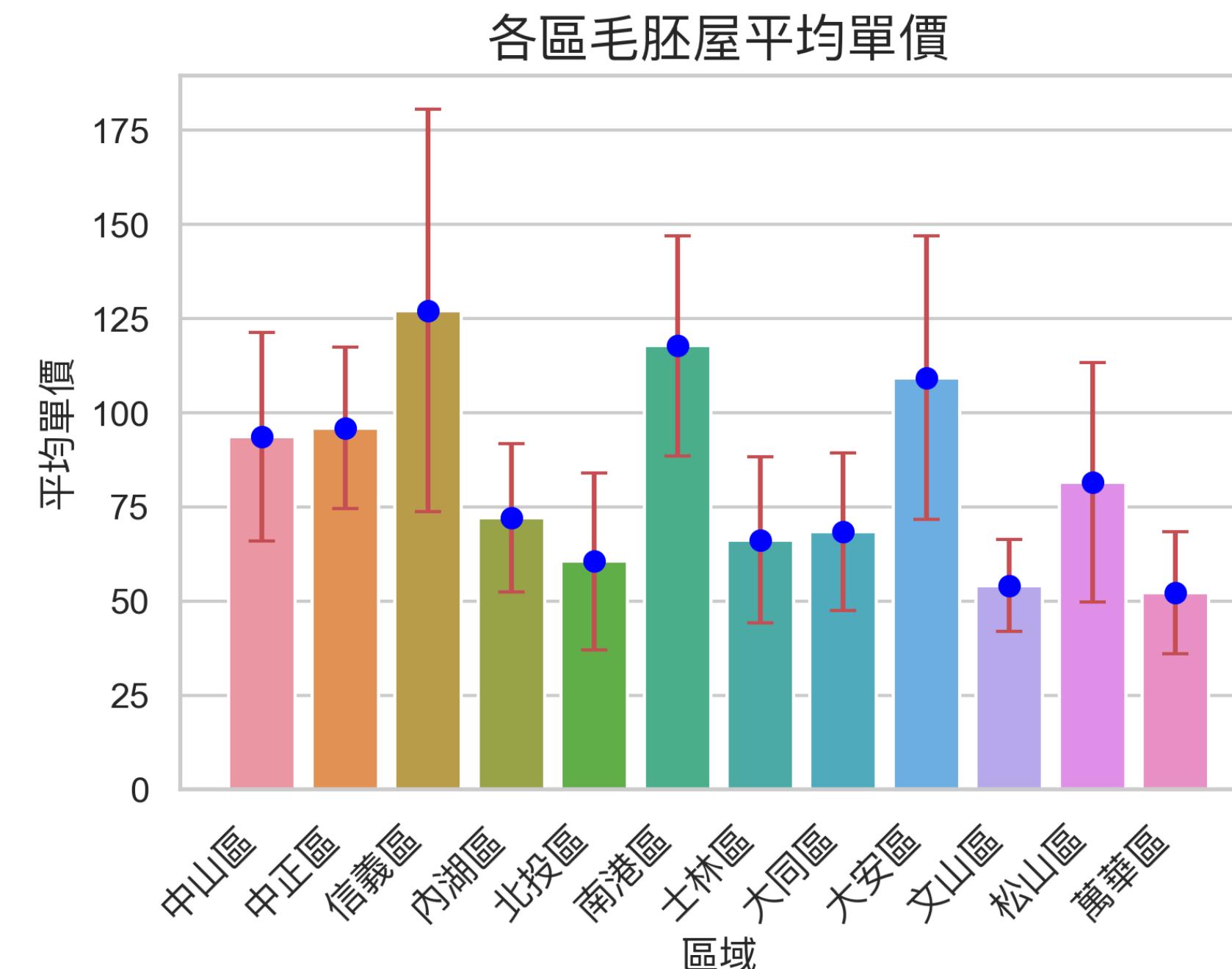
- 是否有管理組織件數與平均單價分布圖



- 是否有特殊交易關係與平均單價



- 各區毛胚屋平均單價



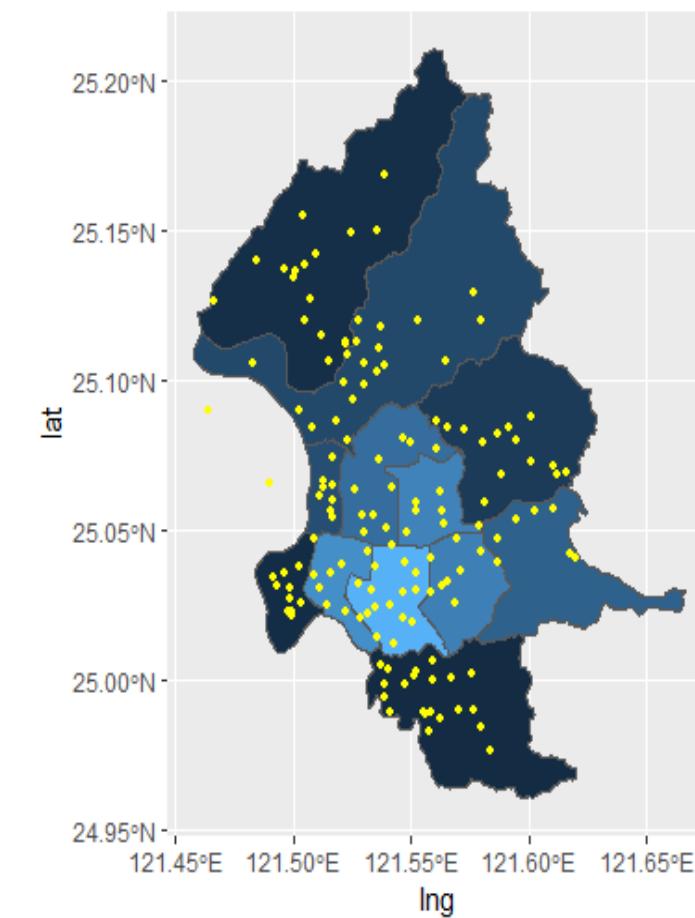
# 鄰近環境資料

- 交通：捷運站、公車站
- 生活機能：便利商店、郵局、銀行
- 教育環境：國小、國中、高中、大學
- 醫療服務：醫院、中醫/西醫診所、牙醫

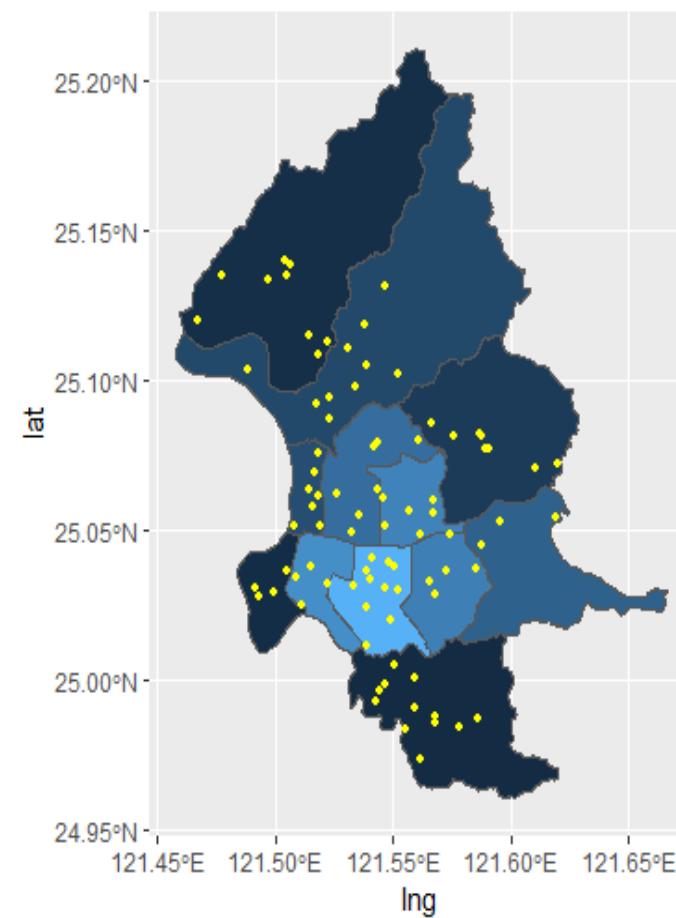
資料來源：政府資料開放平臺

- 教育環境 vs 每坪單價

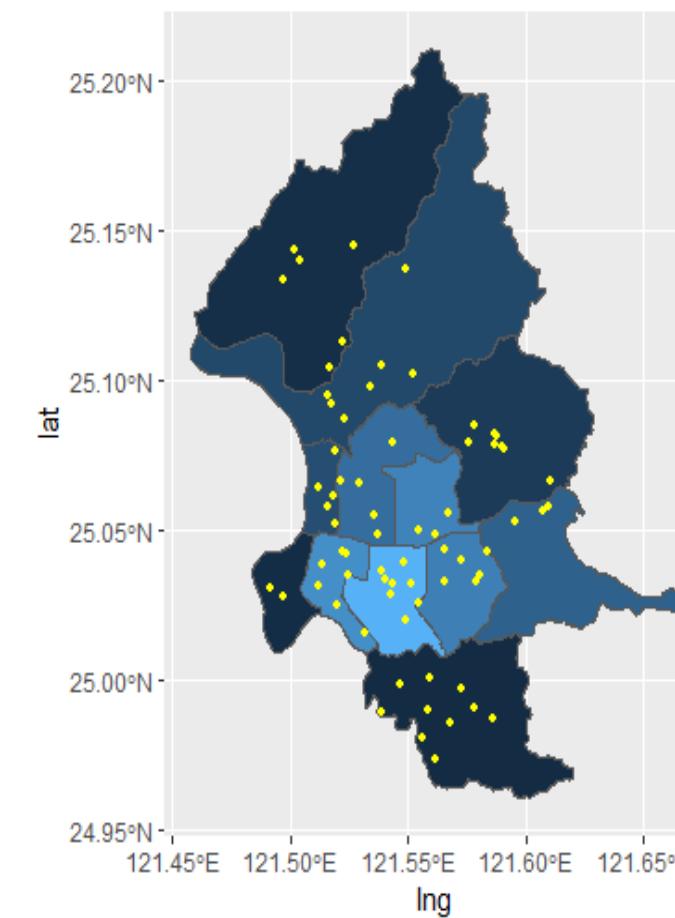
台北市小學圖



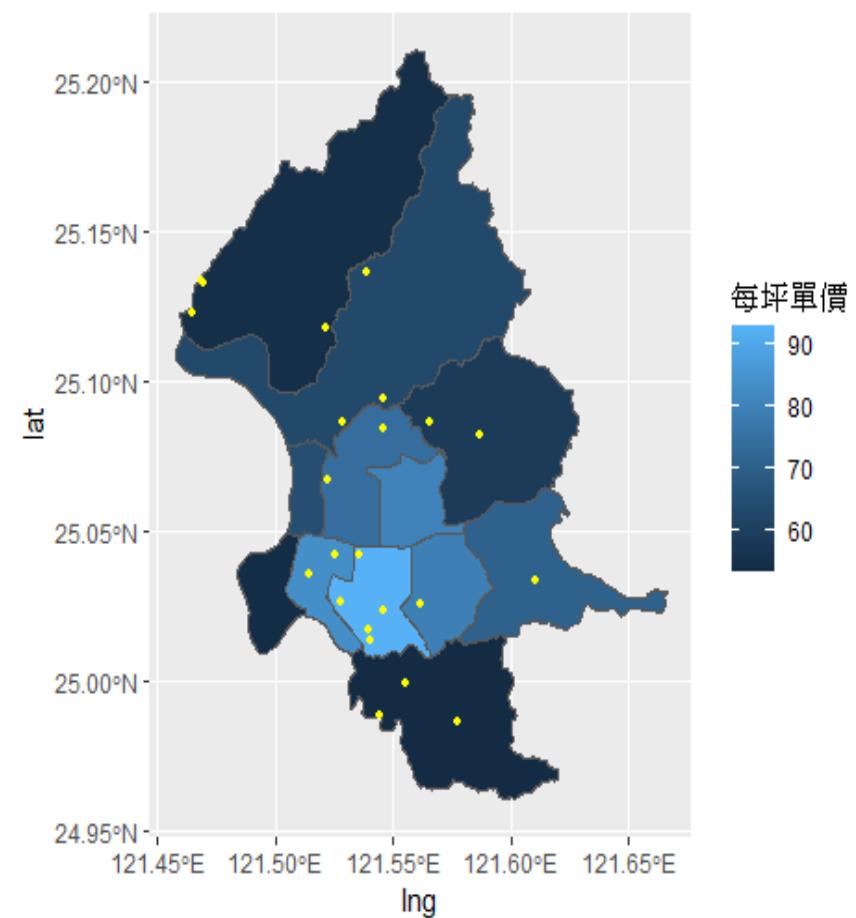
台北市國中圖



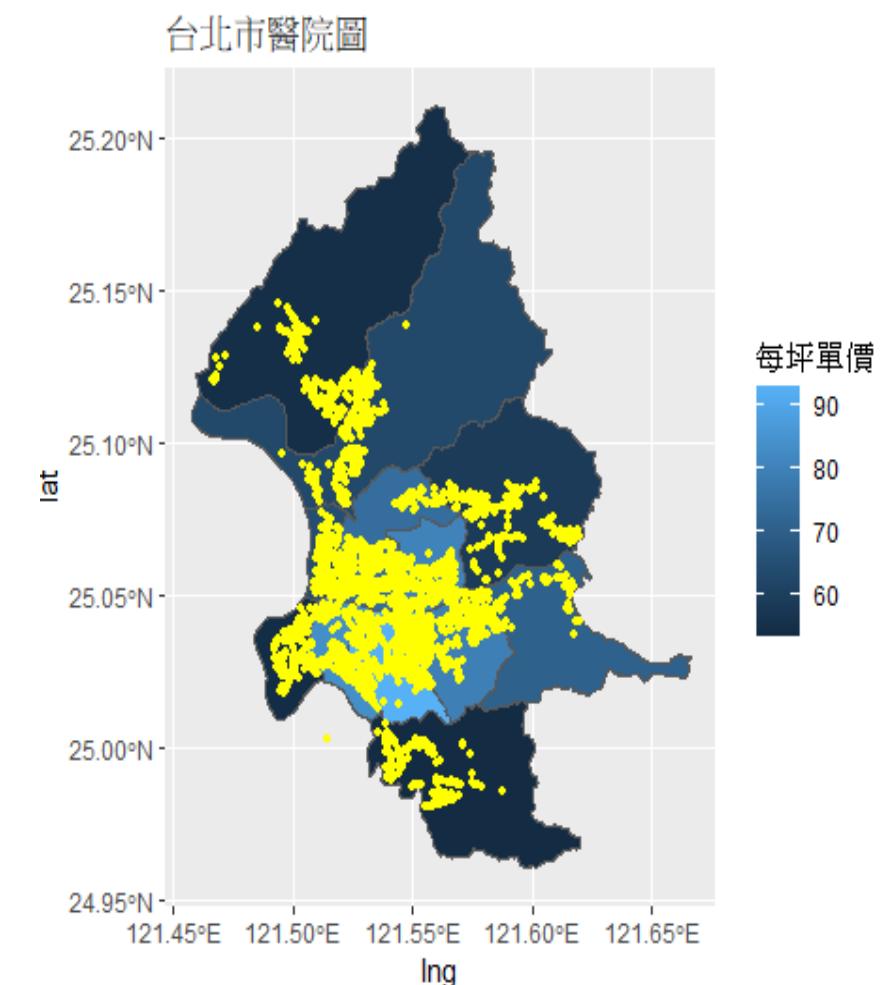
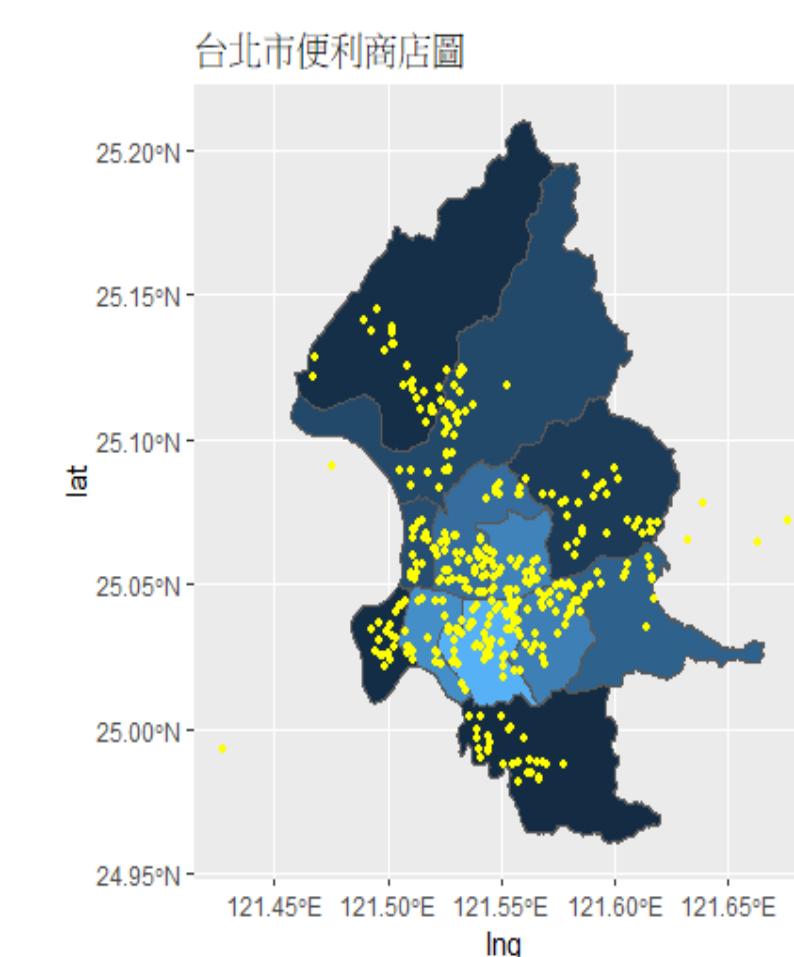
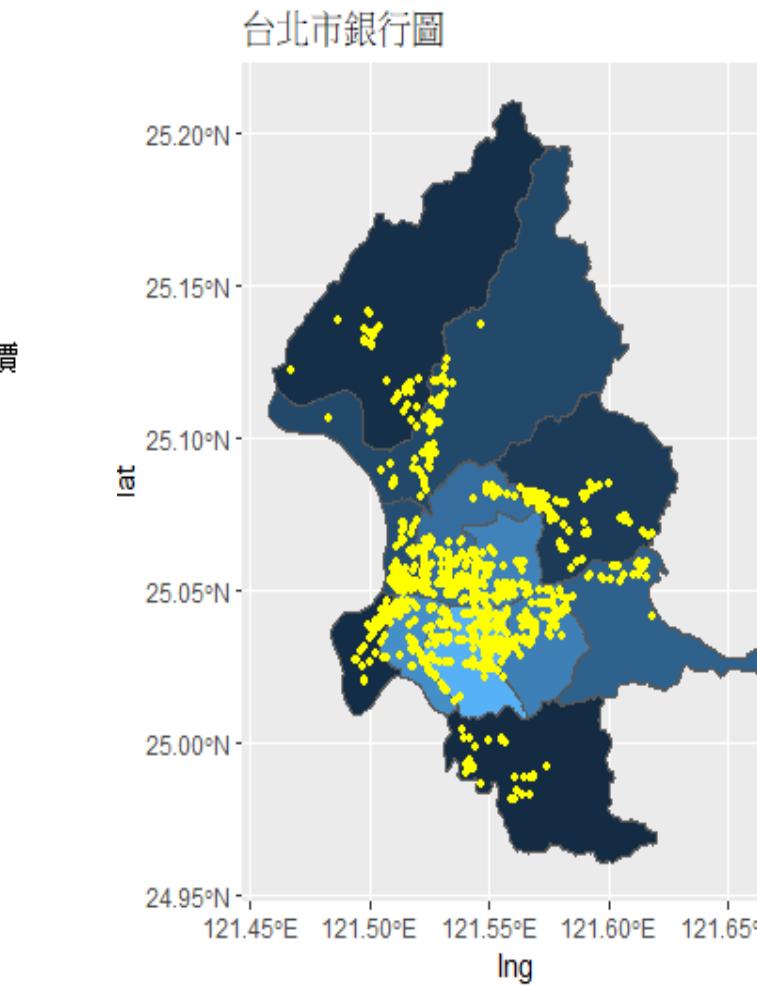
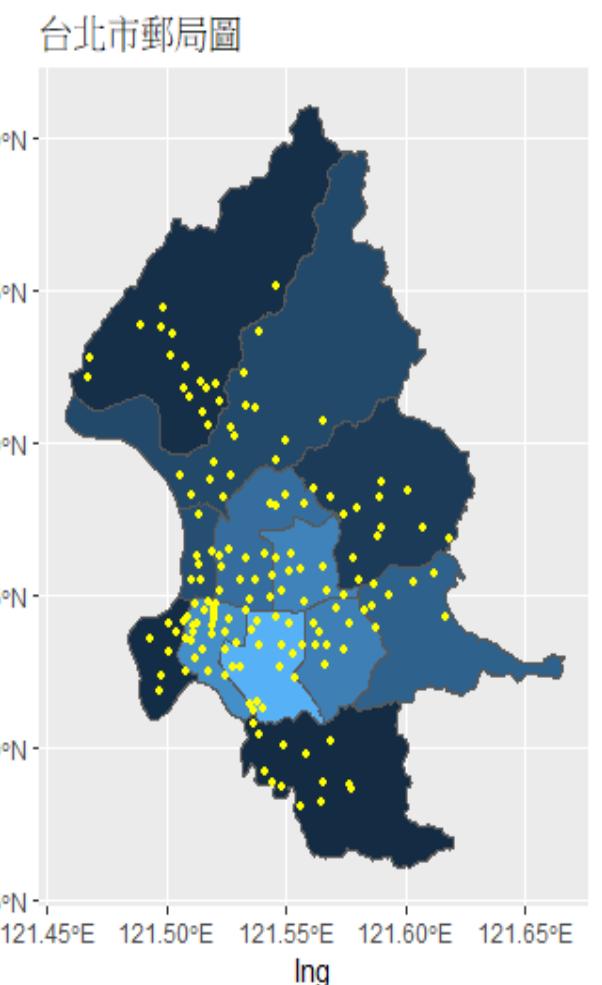
台北市高中圖



台北市大學圖



- 生活機能、醫療服務 vs 每坪單價



# Data Preprocessing



# 變數來源

實價登錄

鄰近環境

總體經濟

歷史房價

# 實價登錄資料

原始資料：  
台北市 2017.12.01 ~ 2022.12.31

**共 70823 筆, 54 個變數**

**交易標的**  
僅保留：土地+建物+車位、土地+建物  
(刪除3889筆)

**屋齡遺失值**  
將該筆交易日 - 建築完成日期作為屋齡  
(補齊3885筆)

**交易別**  
根據備註欄說明，調整為預售屋或毛胚屋

- 刪除特殊交易關係之資料  
(6092筆)
- 刪除有車位、無申報面積價錢之資料  
(3895筆)

**類別變數做 One Hot Encoding**

**KNN 補遺失值**  
屋齡(10489筆)、車位總面積(1012筆)、最低樓層,最高樓層(158筆)

**共 57017 筆, 88 個變數**

# 鄰近環境資料

最短距離(m)

醫院	國小	郵局
牙醫	國中	銀行
中醫	高中	捷運
西醫	大學	便利商店
		公車站

考慮數量

1000公尺內牙醫數量

1000公尺內中醫數量

1000公尺內西醫數量

500公尺內公車站數量

500公尺內便利商店數量

共新增 18 個變數

# 經濟變數

- 加入過去三年，每三個月一期的總體經濟變數

## 建築貸款餘額

建築貸款餘額 $t-3$ , 建築貸款餘額 $t-6$ , ..... 建築貸款餘額 $t-33$ , 建築貸款餘額 $t-36$

## 消費者物價指數

消費者物價指數, 消費者物價指數 $t-6$ , ..... 消費者物價指數 $t-33$ , 消費者物價指數 $t-36$

## 貨幣供給額

貨幣供給額 $t-3$ , 貨幣供給額 $t-6$ , ..... 貨幣供給額 $t-33$ , 貨幣供給額 $t-36$

## 失業率

失業率 $t-3$ , 失業率 $t-6$ , ..... 失業率 $t-33$ , 失業率 $t-36$

## 貸款利率

貸款利率 $t-3$ , 貸款利率 $t-6$ , ..... 貸款利率 $t-33$ , 貸款利率 $t-36$

共新增 60 個變數

# 歷史房價

- 加入過去兩年，該區每月平均房價(單價)

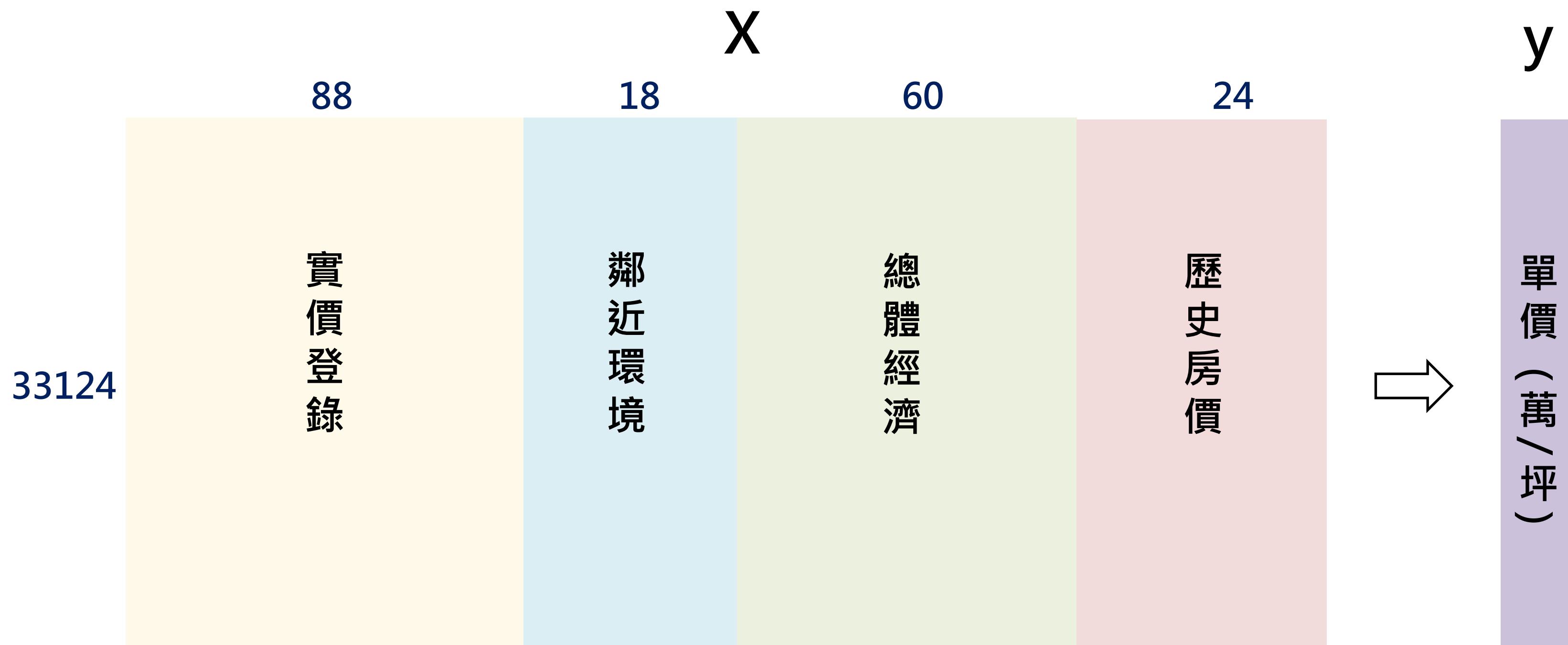
歷史房價

該區平均房價 $t_{-1}$ , 該區平均房價 $t_{-2}$ , ..... 該區平均房價 $t_{-23}$ , 該區平均房價 $t_{-24}$

- 將實價登錄資料由 2017.12 ~ 2022.12 限縮至 2019.12 ~ 2022.12  
**(原 57017 筆調整為 33142 筆)**
- 若當月該區無房屋交易，以全區的平均房價為基準插補

共新增 24 個變數

# Input & Output



共33142筆, 190個變數

# Before Model

- 將變數做 Min-Max Normalization
- 按時間劃分資料集
  - Train (80%): 2019.12.01 ~ 2022.06.15
  - Test (20%): 2022.06.16 ~ 2022.12.31

備註: 測試集中會有部分資料的歷史房價參照測試集

# Description

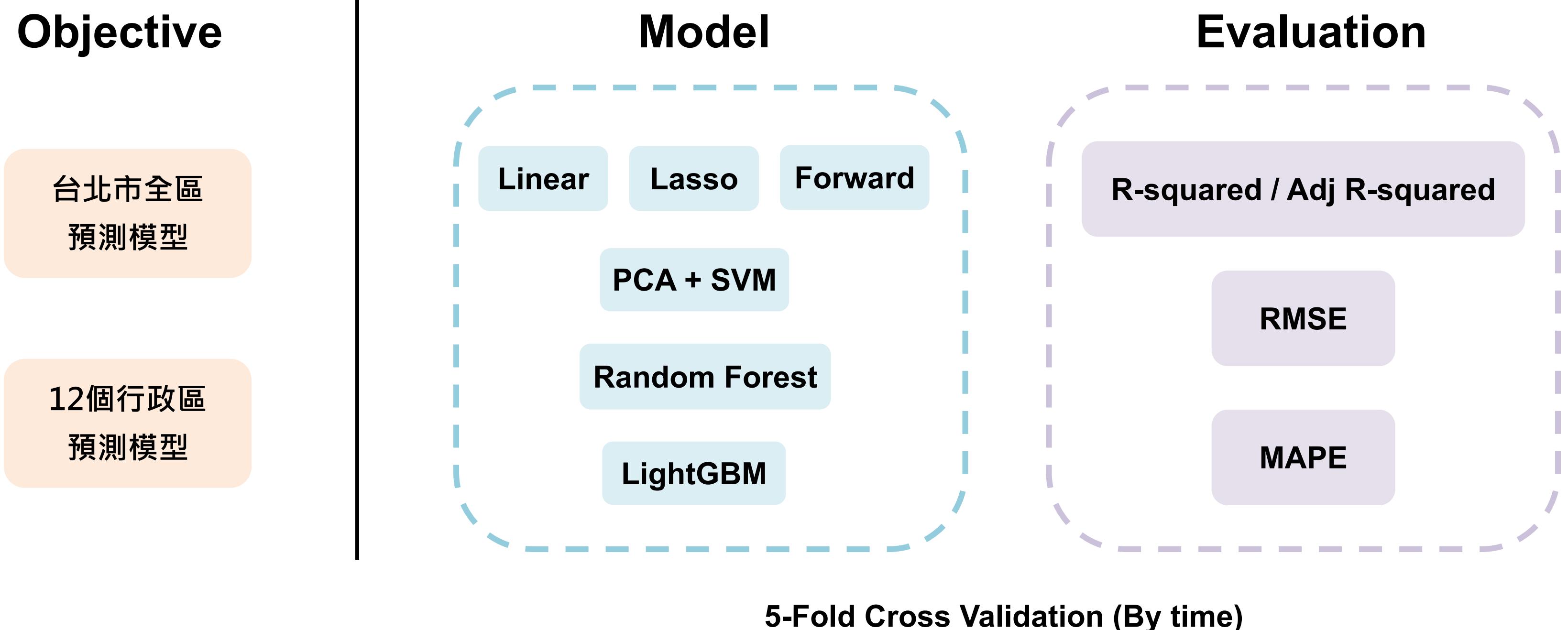
全區					
Training			Testing		
平均單價	標準差	數量	平均單價	標準差	數量
75.28	30.33	26539	75.35	28.10	6603

分區	Training			Testing		
	平均單價	標準差	數量	平均單價	標準差	數量
中山區	79.33	29.61	3881	78.17	25.32	1007
中正區	90.56	35.28	1480	91.71	30.42	344
信義區	86.24	29.26	2150	88.63	28.93	511
內湖區	62.88	17.80	2616	65.45	18.04	647
北投區	58.16	24.23	2361	64.12	27.34	688
南港區	79.33	21.32	2088	74.96	24.27	314
士林區	66.18	27.19	1842	69.85	29.68	542
大同區	70.61	24.36	1031	74.08	21.12	324
大安區	103.23	31.63	2511	103.04	27.51	653
文山區	59.60	17.22	2958	57.00	14.40	607
松山區	93.23	37.53	1701	86.67	25.42	411
萬華區	59.48	19.35	1920	58.72	15.66	555

# Modeling



# Model & Evaluation



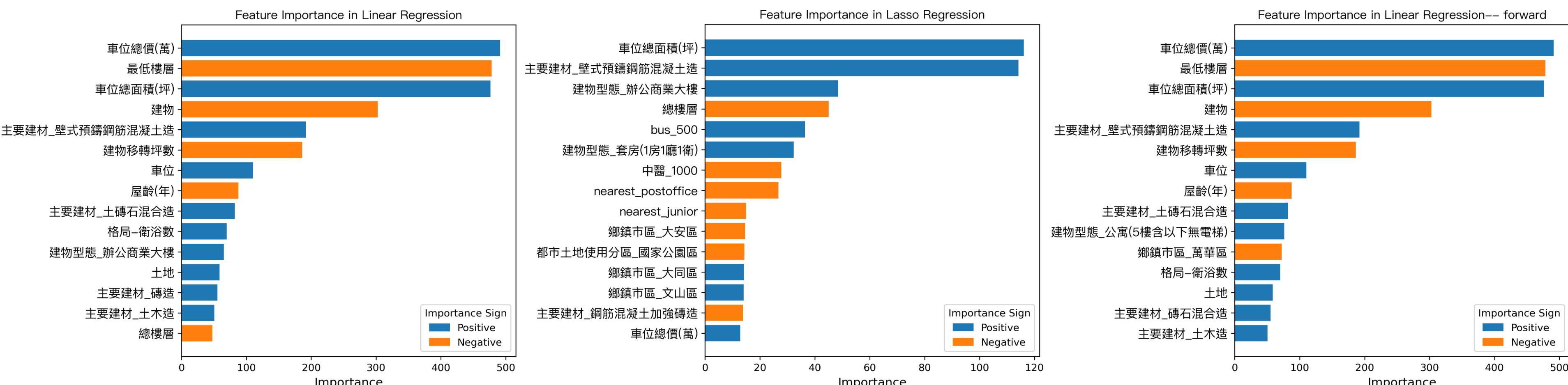
# Linear Model: 台北市全區

全區	Training			Testing			#p
	Adj R2	RMSE	MAPE	Adj R2	RMSE	MAPE	
Linear	0.65	17.85	0.16	0.56	18.27	0.18	190
Lasso	0.61	18.79	0.17	0.58	18.16	0.18	85
Forward	0.65	17.85	0.18	0.56	18.32	0.18	149

## Hyperparameters

- Lasso:  
Lambda=0.0283

Linear, Forward: 車位、建材、建物型態、屋齡  
Lasso: 車位、建材、建物型態、區域、生活機能



# Linear Model: 分區

分區	Linear					
	Training			Testing		
	Adj R2	RMSE	MAPE	Adj R2	RMSE	MAPE
中山區	0.63	17.68	0.14	-3.41	47.95	0.57
中正區	0.72	17.59	0.14	-8.06	61.53	0.38
信義區	0.68	15.92	0.13	-2.63	43.81	0.40
內湖區	0.66	10.04	0.11	-3.33	31.62	0.27
北投區	0.77	11.10	0.14	-4.74	55.80	0.36
南港區	0.81	8.97	0.09	-5.81	40.02	0.28
士林區	0.58	16.63	0.15	-1.78	39.98	0.37
大同區	0.67	12.63	0.12	-5.52	34.87	0.24
大安區	0.68	17.30	0.14	-4.31	53.49	0.34
文山區	0.64	9.98	0.11	-0.40	14.16	0.20
松山區	0.83	14.57	0.13	-1.31	28.44	0.21
萬華區	0.64	11.08	0.12	-4.33	29.39	0.26

分區	Lasso					
	Training			Testing		
	Adj R2	RMSE	MAPE	Adj R2	RMSE	MAPE
中山區	0.48	21.24	0.17	0.46	18.25	0.17
中正區	0.58	22.42	0.18	0.04	28.05	0.22
信義區	0.56	19.09	0.15	0.31	22.30	0.16
內湖區	0.56	11.67	0.13	-0.46	20.67	0.18
北投區	0.61	15.06	0.19	0.64	15.92	0.19
南港區	0.76	10.27	0.10	0.55	14.48	0.15
士林區	0.38	21.15	0.20	0.36	22.82	0.21
大同區	0.63	14.30	0.13	0.16	17.54	0.17
大安區	0.57	20.47	0.16	0.33	21.57	0.16
文山區	0.59	10.96	0.12	0.51	9.56	0.15
松山區	0.77	17.74	0.15	0.40	18.28	0.16
萬華區	0.40	14.90	0.14	0.37	12.12	0.16

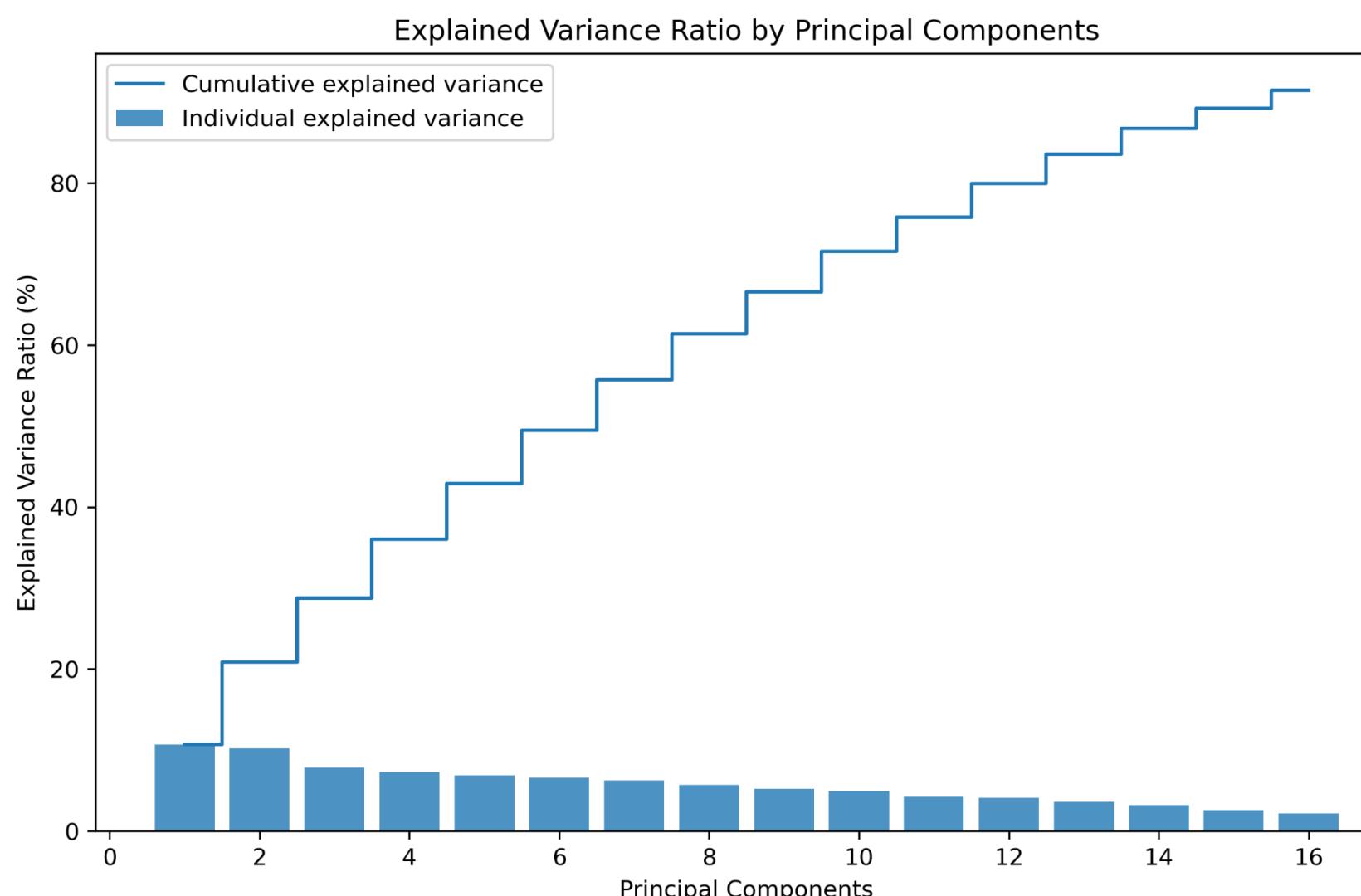
分區	Forward					
	Training			Testing		
	Adj R2	RMSE	MAPE	Adj R2	RMSE	MAPE
中山區	0.63	17.68	0.14	-32.81	138.24	1.79
中正區	0.73	17.59	0.14	-5.44	62.37	0.36
信義區	0.69	15.92	0.13	-1.14	36.94	0.32
內湖區	0.67	10.04	0.11	-2.75	31.41	0.25
北投區	0.78	11.10	0.14	-16.20	102.47	1.79
南港區	0.81	8.97	0.09	-10.09	63.63	0.78
士林區	0.60	16.63	0.15	-27.76	139.90	2.09
大同區	0.70	12.63	0.12	-36.30	101.75	1.19
大安區	0.68	17.30	0.14	-3.27	50.61	0.27
文山區	0.65	9.98	0.11	-1.48	20.37	0.36
松山區	0.84	14.57	0.13	-4.49	49.55	0.48
萬華區	0.65	11.08	0.12	-90.78	132.32	2.07

# PCA+SVM: 台北市全區

全區	Training			Testing			#PC
	R2	RMSE	MAPE	Adj R2	RMSE	MAPE	
SVM	0.01	30.20	0.29	-0.04	28.58	0.30	16

## Hyperparameters

- rbf kernel
- C=1
- Gamma=10
- Epsilon=0.1



# PCA+SVM: 分區

分區	SVM						
	Training			Testing			# PC
	R2	RMSE	MAPE	R2	RMSE	MAPE	
中山區	0.02	29.26	0.23	-0.05	25.68	0.23	16
中正區	0.09	33.48	0.23	-0.08	30.91	0.27	15
信義區	-0.01	29.36	0.25	-0.09	29.73	0.23	16
內湖區	-0.02	17.96	0.22	-0.10	18.70	0.21	14
北投區	0.10	22.88	0.29	-0.12	28.59	0.34	14
南港區	0.03	20.90	0.24	-0.08	24.58	0.31	17
士林區	0.03	26.69	0.23	-0.13	31.04	0.30	15
大同區	0.05	23.51	0.21	-0.21	22.71	0.24	15
大安區	0.03	31.07	0.22	-0.08	28.20	0.21	15
文山區	0.13	16.01	0.22	-0.05	14.57	0.25	17
松山區	0.02	36.93	0.23	-0.07	25.79	0.18	15
萬華區	0.03	19.02	0.22	-0.04	15.76	0.24	15

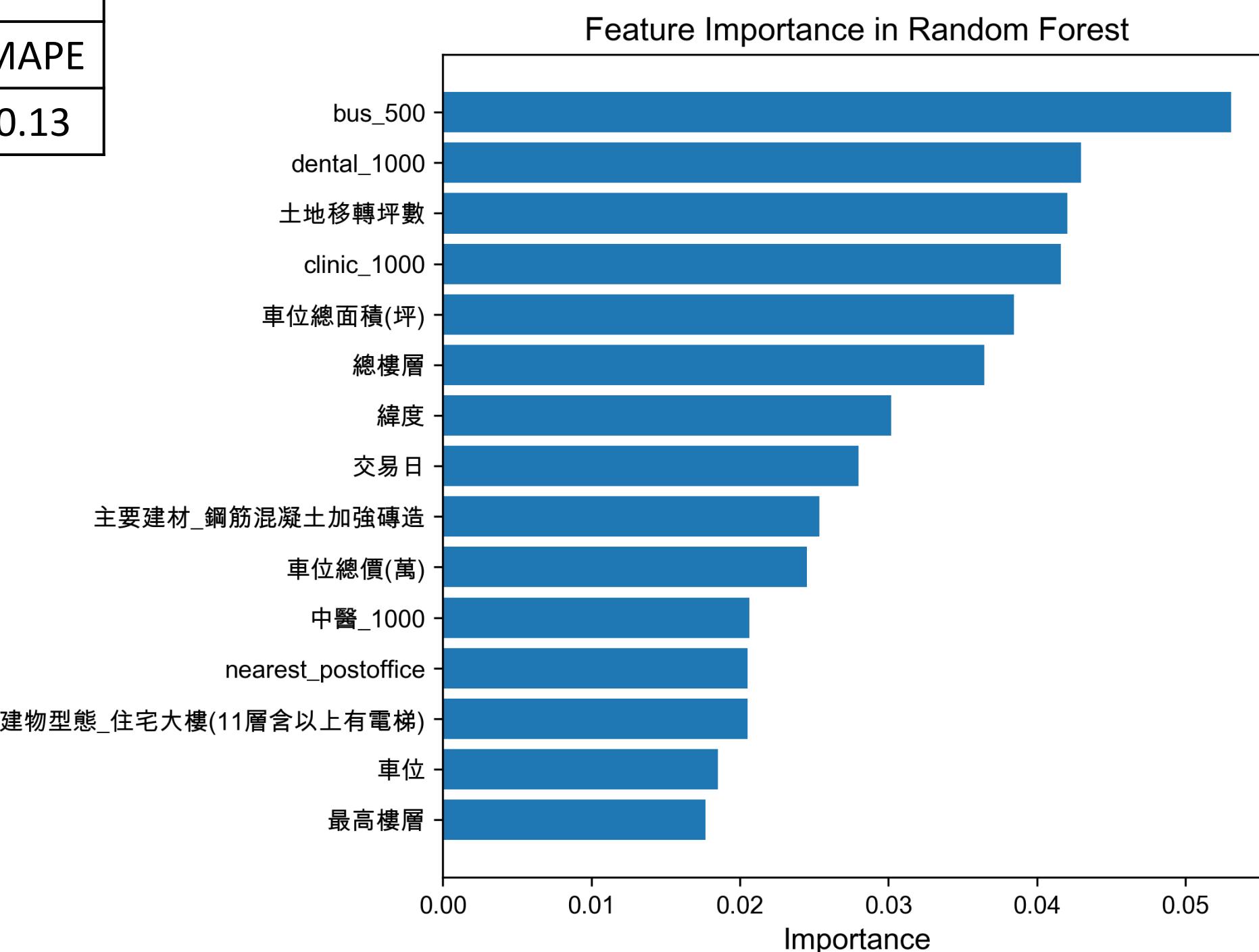
# Random Forest: 台北市全區

不分區	Training			Testing		
	R2	RMSE	MAPE	R2	RMSE	MAPE
RF	0.97	5.12	0.04	0.74	14.21	0.13

Hyperparameters

- n\_estimators=200
- max\_depth=None
- min\_samples\_split=2
- min\_samples\_leaf=1

生活機能、坪數、車位、建材、建物型態、交易日



# Random Forest: 分區

分區	Random Forest					
	Training			Testing		
	R2	RMSE	MAPE	R2	RMSE	MAPE
中山區	0.93	7.93	0.05	0.72	13.49	0.12
中正區	0.97	6.20	0.04	0.69	16.91	0.15
信義區	0.96	5.81	0.04	0.63	17.58	0.13
內湖區	0.95	3.97	0.04	0.53	12.38	0.13
北投區	0.89	8.21	0.08	0.67	15.72	0.15
南港區	0.96	4.00	0.04	0.75	12.22	0.13
士林區	0.81	12.00	0.10	0.51	20.76	0.17
大同區	0.93	6.57	0.05	0.40	16.32	0.16
大安區	0.94	7.61	0.05	0.64	16.54	0.12
文山區	0.95	3.78	0.03	0.44	10.81	0.17
松山區	0.97	6.68	0.06	0.66	14.74	0.13
萬華區	0.86	7.28	0.05	0.66	9.13	0.12

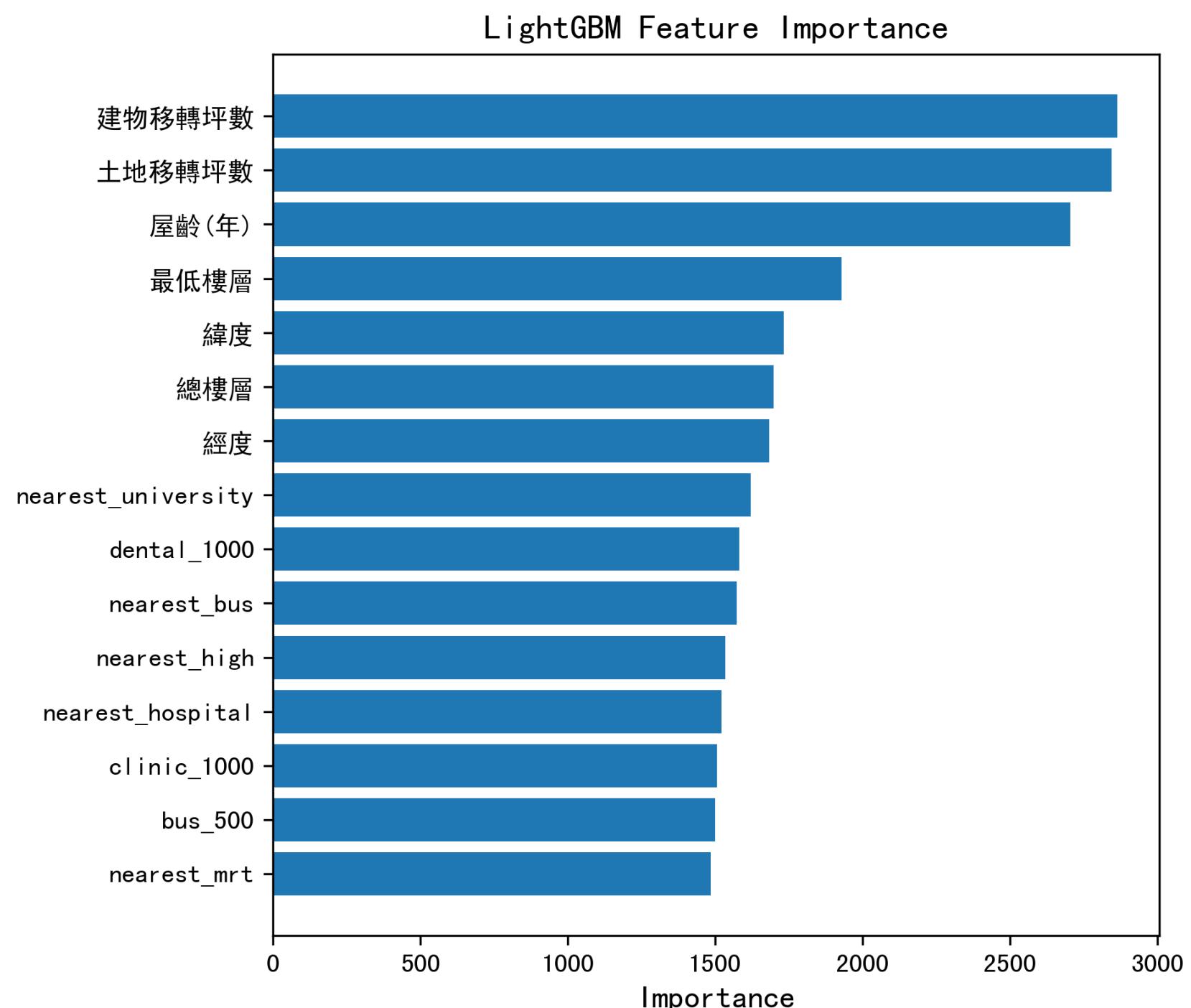
# LightGBM: 台北市全區

建物、土地移轉坪數、屋齡、樓層、經緯度、生活機能

不分區	Training			Testing		
	R2	RMSE	MAPE	R2	RMSE	MAPE
LightGBM	0.97	5.08	0.05	0.82	11.91	0.11

## Hyperparameters

- boosting\_type: gbdt
- colsample\_bytree: 0.8
- lambda\_l1: 0.1
- lambda\_l2: 0.1
- learning\_rate: 0.05
- max\_depth: 5
- n\_estimators: 2000
- num\_leaves: 40
- objective: regression
- subsample: 0.8



# LightGBM: 分區

分區	LightGBM					
	Training			Testing		
	R2	RMSE	MAPE	R2	RMSE	MAPE
中山區	0.99	2.98	0.03	0.79	11.62	0.10
中正區	1.00	0.79	0.01	0.71	16.40	0.14
信義區	1.00	1.66	0.02	0.73	15.15	0.12
內湖區	1.00	1.25	0.02	0.56	12.03	0.12
北投區	1.00	1.56	0.02	0.72	14.58	0.14
南港區	1.00	1.09	0.01	0.80	10.78	0.11
士林區	0.99	2.23	0.03	0.58	19.33	0.14
大同區	1.00	0.65	0.01	0.52	14.57	0.14
大安區	0.99	2.27	0.02	0.74	13.98	0.11
文山區	0.96	3.23	0.04	0.58	9.36	0.13
松山區	1.00	0.70	0.01	0.72	13.36	0.11
萬華區	0.99	1.84	0.02	0.66	9.10	0.11

# Summary: 台北市全區

全區	Testing		
	R2	RMSE	MAPE
Linear	0.56	18.27	0.18
Lasso	0.58	18.16	0.18
Forward	0.56	18.32	0.18
SVM	-0.04	28.58	0.30
RF	0.74	14.21	0.13
LightGBM	0.82	11.91	0.11

# Summary: 分區

分區	Testing																	
	中山區			中正區			信義區			內湖區			北投區			南港區		
	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE
Linear	-3.41*	47.95	0.57	-8.06*	61.53	0.38	-2.63*	43.81	0.40	-3.33*	31.62	0.27	-4.74*	55.80	0.36	-5.81*	40.02	0.28
Lasso	0.46*	18.25	0.17	0.04*	28.05	0.22	0.31*	22.30	0.16	-0.46*	20.67	0.18	0.64*	15.92	0.19	0.55*	14.48	0.15
Forward	-32.81*	13.24	1.79	-5.44*	62.37	0.36	-1.14*	36.94	0.32	-2.75*	31.41	0.25	-16.20*	102.47	1.79	-10.09*	63.63	0.78
SVM	-0.05	25.68	0.23	-0.08	30.91	0.27	-0.09	29.73	0.23	-0.10	18.70	0.21	-0.12	28.59	0.34	-0.13	31.04	0.30
RF	0.72	13.49	0.12	0.69	16.91	0.15	0.63	17.58	0.13	0.53	12.38	0.13	0.67	15.72	0.15	0.75	12.22	0.13
LightGBM	0.79	11.62	0.10	0.71	16.40	0.14	0.73	15.15	0.12	0.56	12.03	0.12	0.72	14.58	0.14	0.80	10.78	0.11

分區	Testing																	
	士林區			大同區			大安區			文山區			松山區			萬華區		
	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE
Linear	-1.78*	39.98	0.37	-5.52*	34.87	0.24	-4.31*	53.49	0.34	-0.40*	14.16	0.20	-1.31*	28.44	0.21	-4.33*	29.39	0.26
Lasso	0.36*	22.82	0.21	0.16*	17.54	0.17	0.33*	21.57	0.16	0.51*	9.56	0.15	0.40*	18.28	0.16	0.37*	12.12	0.16
Forward	-27.76*	139.90	2.09	-36.30*	101.75	1.19	-3.27*	50.61	0.27	-1.48*	20.37	0.36	-4.49*	49.55	0.48	-90.78*	132.32	2.07
SVM	-0.13	31.04	0.30	-0.21	22.71	0.24	-0.08	28.20	0.21	-0.05	14.57	0.25	-0.07	25.79	0.18	-0.04	15.76	0.24
RF	0.51	20.76	0.16	0.40	16.32	0.16	0.64	16.54	0.12	0.44	10.81	0.17	0.66	14.74	0.13	0.66	9.13	0.12
LightGBM	0.58	19.33	0.14	0.52	14.57	0.14	0.74	13.98	0.11	0.58	9.36	0.13	0.72	13.36	0.11	0.66	9.10	0.11

\*: Adjusted R-squared

# **Conclusions & Future Work**



# Conclusions

- 在分區與不分區模型中，LightGBM皆表現最佳
- 在分區模型中，各地區模型表現差異不大，唯大同區、士林區、內湖區、文山區模型表現較差
- 就整體而言，不分區模型表現較分區模型好

# Future Work

- 先分群再配適模型
- 新增更多資料
- 擴增至直轄市六都模型，並比較模型差異，找出潛在影響因子

# Reference

- 應用時價登陸已聚類方法之堆疊泛化房價預測模型, 黃允亭, 政治大學經濟學系研究所
- 以總經變數預測不動產價格之模型, 吳岱蓉, 清華大學計量財務金融學系研究所
- Duan et al. (2021) “Addressing the macroeconomic and hedonic determinants of housing prices in Beijing Metropolitan Area, China”
- 房貸利率、新成屋房價與建商股價關聯性之研究, 蕭曉文, 中正大學財務金融研究所
- TEJ API <https://api.tej.com.tw/columns.html?idCode=TWN/AAPRTRAN>
- 內政地理資訊圖資雲整合服務平台 <https://www.tgos.tw/tgos/Addr/Compare>
- 政府資料開放平台 <https://data.gov.tw/>
- 中華民國中央銀行全球資訊網 <https://www.cbc.gov.tw/tw/mp-1.html>
- 行政院主計處中華民國統計資訊網 <https://www.stat.gov.tw/cl.aspx?n=3563>

# Thank you!



# Appendix



# 實價登錄資料-完整資料預處理(1)

變數	敘述	資料維度
.	實價登錄提供之原始台北市資料	(70823, 54)
交易標的	保留: 房地(土地+建物)+車位、房地(土地+建物) 刪除: 建物、土地、車位	(67045, 54)
移轉樓層 移轉地上樓層 移轉單一地上樓層(棟)	三個變數改為最低樓層、最高樓層兩個變數	(67045, 53)
屋齡	若屋齡為遺失值，則將該筆的交易日-建築完成日期當作屋齡。 (原為15693個遺失值，補後仍有11808個)	(67045, 53)
交易別	根據備註欄說明，調整為預售屋或毛胚屋	(67045, 53)
.	根據備註欄說明，刪除特殊交易關係的資料	(60953, 53)
建物型態	刪除為倉庫工廠的資料	(60912, 53)
主要用途	留下住商、工業、辦公室類別，其餘併去「其他類」	(60912, 53)
.	刪除有車位，沒有申報面積價錢的資料 (影響單價)	(57017, 53)
經度、緯度	遺失7191筆。使用地址欄位資訊，於內政地理資訊圖資雲整合服務平台補齊	(57017, 53)
交易日	轉為自Unix紀元(1970年1月1日)以來的秒數	(57017, 53)
.	刪除遺失值過多，或重複資訊的變數 (見下頁)	(57017, 29)
.	針對類別變數做One-Hot Encoding	(57017, 88)
.	KNN補遺失值: 屋齡(10489)、車位總面積(1012)、最低樓層(158)、最高樓層(158)	(57017, 88)

# 實價登錄資料-完整資料預處理(2)

刪除遺失值過多變數：

- 非都市土地使用分區 (NA: 70690)
- 非都市土地使用編定(NA: 70823)
- 交易日期異常註記 (NA: 70823)
- 無效門牌/地號註記 (NA: 70204)
- 無交易價格註記 (NA: 70787)

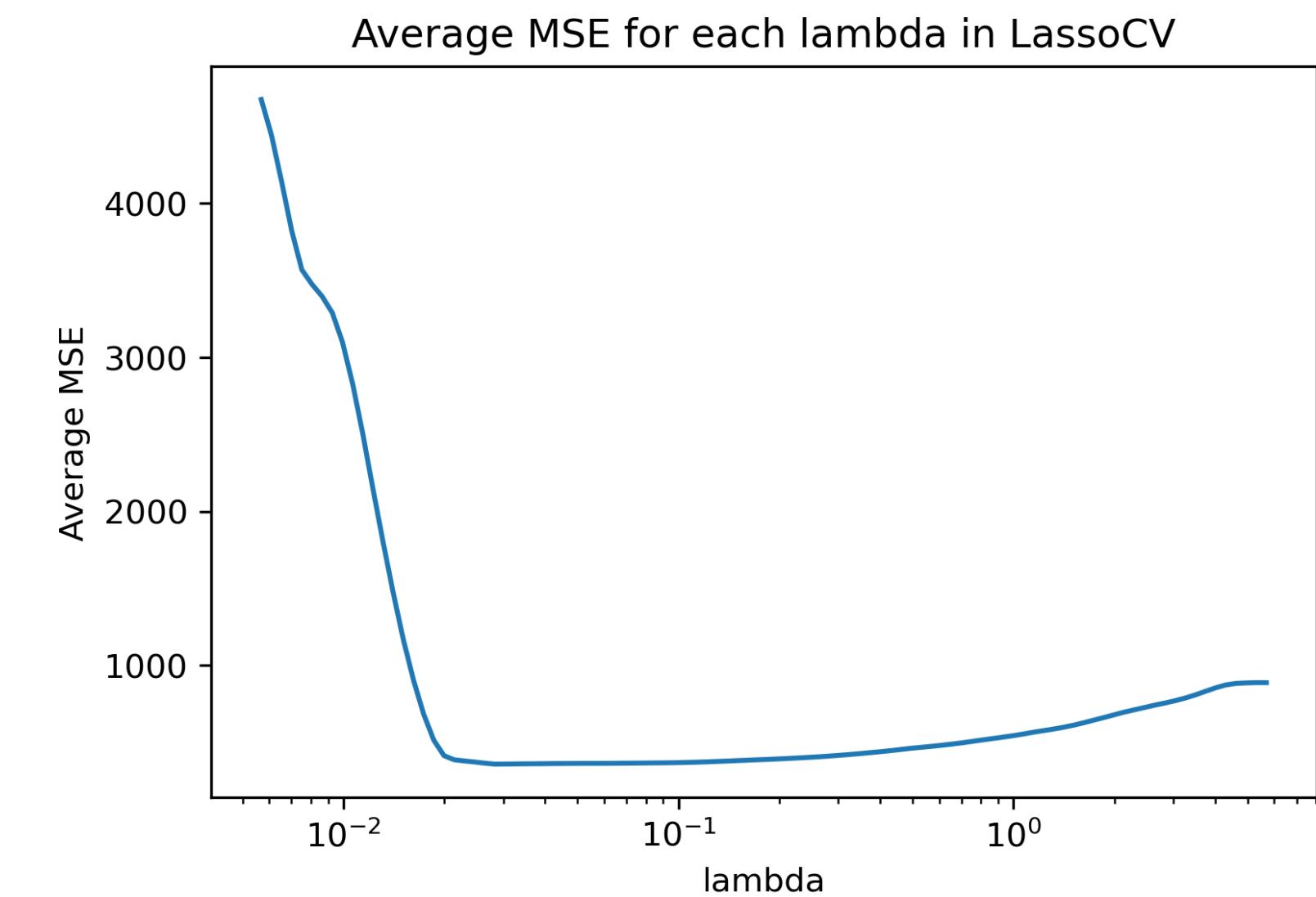
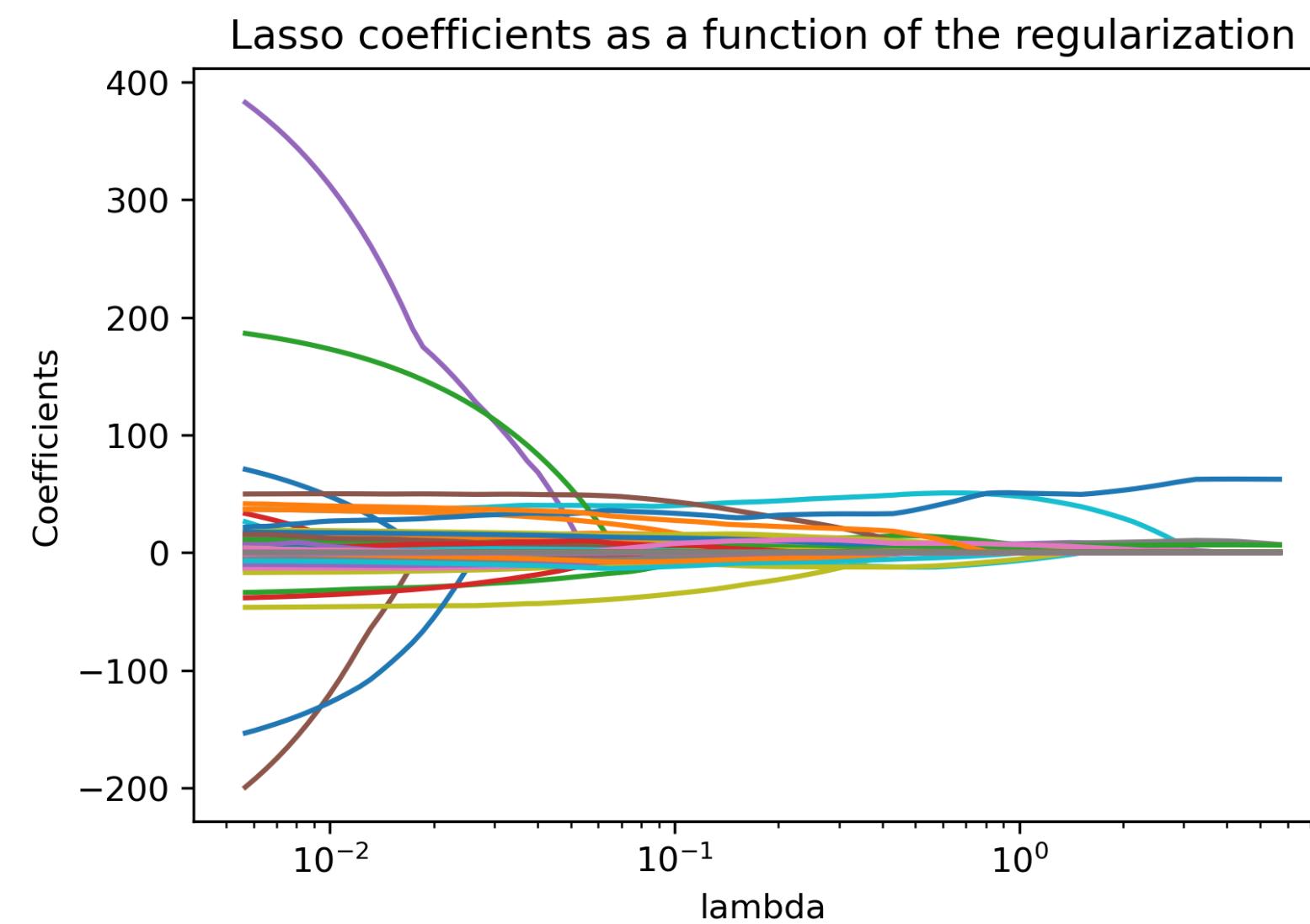
刪除為分類註記的變數：

- 非都市土地使用分區註記
- 交易別註記
- 交易標的註記
- 都市土地使用分區註記
- 主要用途註記
- 車位類別註記
- 主要建材註記
- 建物型態註記
- 非都市土地使用編定註記

無使用之變數

- 建造日期
- 隔間
- 公告日
- 縣市
- 案碼
- 鄉鎮市區碼
- 備註
- 總價
- 車位類別
- 地址欄位

# Lasso(全區)- 選擇lambda



# 分區SVM, RF使用參數

SVM	HyperParameters
中山區	rbf, C: 10, gamma: 10, epsilon: 0.1
中正區	rbf, C: 10, gamma: 10, epsilon: 0.1
信義區	rbf, C: 1, gamma: 1, epsilon: 0.3
內湖區	rbf, C: 0.01, gamma: 0.01, epsilon: 0.4
北投區	rbf, C: 10, gamma: 10, epsilon: 0.5
南港區	rbf, C: 10, gamma: 0.1, epsilon: 0.6
士林區	rbf, C: 10, gamma: 10, epsilon: 0.7
大同區	rbf, C: 10, gamma: 1, epsilon: 0.8
大安區	rbf, C: 10, gamma: 10, epsilon: 0.9
文山區	rbf, C: 10, gamma: 10, epsilon: 0.10
松山區	rbf, C: 10, gamma: 10, epsilon: 0.11
萬華區	rbf, C: 10, gamma: 10, epsilon: 0.12

RF	HyperParameters
中山區	n_estimators=200, max_depth=None, min_samples_split=5, min_samples_leaf=1
中正區	n_estimators=100, max_depth=20, min_samples_split=2, min_samples_leaf=1
信義區	n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1
內湖區	n_estimators=100, max_depth=20, min_samples_split=2, min_samples_leaf=1
北投區	n_estimators=50, max_depth=None, min_samples_split=10, min_samples_leaf=2
南港區	n_estimators=50, max_depth=None, min_samples_split=5, min_samples_leaf=1
士林區	n_estimators=50, max_depth=10, min_samples_split=2, min_samples_leaf=2
大同區	n_estimators=50, max_depth=20, min_samples_split=5, min_samples_leaf=1
大安區	n_estimators=100, max_depth=None, min_samples_split=5, min_samples_leaf=1
文山區	n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1
松山區	n_estimators=100, max_depth=20, min_samples_split=5, min_samples_leaf=1
萬華區	n_estimators=50, max_depth=20, min_samples_split=2, min_samples_leaf=2

# 分區LightGBM使用參數

LightGBM	HyperParameters
中山區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0, 'lambda_l2': 0.5, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 2000, 'num_leaves': 40}
中正區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0.5, 'lambda_l2': 0.5, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 2000, 'num_leaves': 20}
信義區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0, 'lambda_l2': 0.5, 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, 'num_leaves': 20}
內湖區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0.5, 'lambda_l2': 0.5, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 2000, 'num_leaves': 20}
北投區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0, 'lambda_l2': 0.1, 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, 'num_leaves': 20}
南港區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0.5, 'lambda_l2': 0.5, 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, 'num_leaves': 20}
士林區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0.1, 'lambda_l2': 0.5, 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, 'num_leaves': 20}
大同區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0.5, 'lambda_l2': , 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, 'num_leaves': 20}
大安區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, lambda_l1': 0.1, lambda_l2': 0.5, 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, num_leaves': 20}
文山區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0, 'lambda_l2': 0, 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, 'num_leaves': 20}
松山區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0, 'lambda_l2': 0, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 2000, 'num_leaves': 40}
萬華區	{'bagging_fraction': 0.8, 'feature_fraction': 0.8, 'lambda_l1': 0.1, 'lambda_l2': 0, 'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 2000, 'num_leaves': 20}

# Project Work Records



# Project Work Records(1)

- Date: 10/2(一) 13:00-14:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 決定主題
  - 收集資料(實價登錄)
  - 討論可用的外部資料
  - 討論可用的資料分析方法

# Project Work Records(2)

- Date: 10/5(四) 13:00-15:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 查找相關文獻
  - 收集資料(環境、總經變數)
  - 討論報告內容
  - 製作報告

# Project Work Records(3)

- Date: 10/19(四) 13:00-15:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 修改主題、查找相關文獻
  - 整理資料
    - 保留台北市資料
    - 保留交易標的為房地(土地+建物)、房地(土地+建物)+車位的資料
    - 保留主要用途為住商、工業、辦公室的資料，其餘類別合併去其他
    - 刪除有特殊交易關係、建物、倉庫工廠、有車位沒申報的資料
    - 刪除遺失值過多、有註記的變數
    - 更新交易別變數(NA但備註欄出現預售屋、毛胚屋)
    - 合併變數(移轉樓層、移轉地上樓層、移轉單一地上樓層合併成最低、最高樓層)

# Project Work Records(4)

- Date: 10/26(四) 13:00-15:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 討論文獻內容
  - 整理資料
    - 實價登錄
      - 插補經緯度遺失值(利用地址欄位變數轉換經緯度)
      - 刪除地址欄位、車位類別變數
      - 轉換日期變數為自Unix紀元（1970年1月1日）以來的秒數
      - 插補所有變數遺失值(以knnimputer插補)
      - 轉換所有類別變數(以One Hot Encoding的方式)

# Project Work Records(4)

- 環境變數
  - 新增最近距離變數: 西醫、中醫、牙醫、大型醫院、捷運、學校、郵局、便利商店
  - 新增周遭數量(m)變數: 西醫(1000)、中醫(1000)、牙醫(1000)、便利商店(500)、公車站(500)
- 總經變數
  - 新增總經變數(lag為3、6、...、36個月): 建築貸款餘額、貨幣供給額、消費者物價指數、失業率、貸款利率

# Project Work Records(5)

- Date: 11/2(四) 13:00-17:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 整理文獻大綱
  - 資料視覺化(實價登錄、環境、總經變數)
  - 討論報告內容
  - 製作報告

# Project Work Records(6)

- Date: 11/9(四) 13:00-15:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 修改資料視覺化圖形
  - 整理資料
    - 保留2019.12-2022.12的資料
    - 新增歷史變數(lag為1、2、...、24個月): 房價(單價)
    - 插補歷史變數遺失值(以同期平均插補)
    - 刪除歷史變數為0的資料
    - 正規化處理所有自變數(以Min-Max Scale的方式)

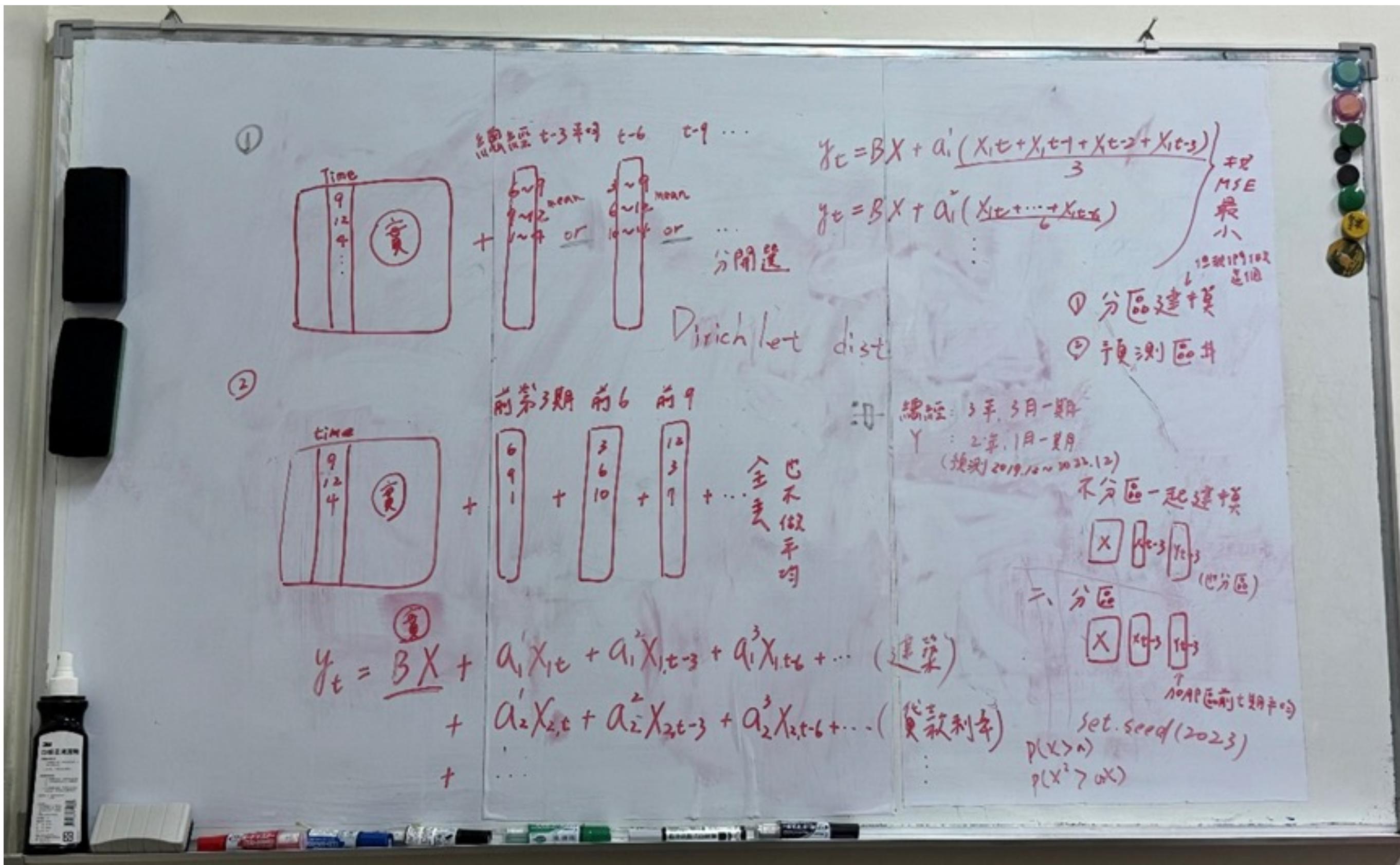
# Project Work Records(7)

- Date: 11/23(四) 13:00-18:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 分割訓練(80%)、測試(20%)資料集
  - 訓練分區與不分區模型: 參數選擇採Grid Search的方式、CV採5 fold
    - Linear Model(分別以Lasso和Forward的方式挑變數): Lambda
    - Random Forest: n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf
    - LightGBM: n\_estimators, num\_leaves, min\_child\_sample
    - SVM: C, Gamma
  - 討論模型結果(以Adjusted R-squared和MSE綜合衡量)
  - 討論變數重要度
  - 討論報告內容、製作報告

# Project Work Records(8)

- Date: 12/14(四) 13:00-15:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 重新分割訓練(80%)、測試(20%)資料集(依時間分割)
    - 訓練集: 2019.12.01-2022.06.15
    - 測試集: 2022.06.16-2022.12.31
  - 重新訓練分區與不分區模型: 參數選擇採Grid Search的方式、CV採5 fold(依時間分割)
    - Linear Model(分別以Lasso, Lasso, Forward的方式挑變數): Lambda
    - Random Forest: n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf
    - LightGBM: max\_depth, num\_leaves, lambda\_l1, lambda\_l2
    - SVM: C, Gamma, epsilon

# Project Work Records(8)



# Project Work Records(9)

- Date: 12/21(四) 13:00-15:00
- Place: 62514研究室
- Participants: 陳沛群、張立勳、黃亮臻、李易庭
- Note Taker: 張立勳
- Content:
  - 修改及新增資料視覺化的圖片
  - 討論模型結果(以Adjusted R-squared, R-squared, RMSE, MAPE綜合衡量)
  - 討論變數重要度
  - 討論報告內容
  - 彙整過往報告內容
  - 製作報告

# Workload

黃亮臻(數據所)

- Data Visualization: 實價登錄資料
- Data Preprocessing: 實價登錄資料
- Modeling : Linear Model
- Final PPT整理、報告

陳沛群(統計所)

- Data Preprocessing : Economic Features, FillNa, Delete Outlier
- Modeling : SVM, RandomForest
- Proposal1, Final PPT整理、報告

李易庭(數據所)

- Data Visualization: Environmental Features
- Data Preprocessing : Environmental Features
- Modeling : LightGBM
- Proposal3 PPT整理、報告

張立勳(數據所)

- Data Preprocessing : Economic Features, Historic Features, Normalization
- Group Meetings Recording
- Proposal2 PPT整理、報告