FCU ePaper

逢甲大學學生報告 ePaper

房屋價格分析：以高雄市左營區為例

Analysis of Housing Prices: A Case Study of Zuoying District, Kaohsiung City

作者：許芸華
系級：運物四乙
學號：D1160661
開課老師：周進華

課程名稱：Python 入門與行銷資料科學
開課系所：行銷學系
開課學年：114 學年度 第 1 學期

商學院

中文摘要

近年，隨著高雄市發展重心北移、台積電設廠及三鐵共構之交通優勢，左營區已成為北高雄商業與交通核心，房地產市場結構隨之發生改變。本研究旨在探討民國 109 年 11 月至 114 年 11 月間，高雄市左營區住宅大樓、華廈及公寓之房價波動與影響因素，研究透過 Python 網路爬蟲技術，自「內政部不動產交易實價查詢服務網」擷取共 10,993 筆有效交易數據，進行資料清洗、特徵工程與離群值處理，以確保數據品質。

本研究結合敘述性統計、地理空間分析與機器學習演算法，比較線性迴歸、決策樹、隨機森林及梯度提升樹四種模型對房價的預測能力。研究結果顯示，隨機森林模型表現最佳，其決定係數(R^2)達 0.8409，均方根誤差(RMSE)為 35,778 元/坪，顯著優於傳統線性迴歸模型之決定係數結果 0.53，說明房價影響因素間存在高度非線性關係。

特徵重要性分析指出，「屋齡」為影響左營區房價之關鍵因素，重要性佔 35.85%，其次為「交易年份」之 25.11%與「地理位置」，經緯度合計為 16.59%。空間分析則說明以高鐵左營站為核心之同心圓價格梯度現象：核心區，如高鐵特區、蓮池潭首排之單價最高，平均落在 35 萬元/坪以上；巨蛋商圈及核心市區次之；外圍與老舊社區則價格相對較低。本研究透過數據，量化各項房屋屬性與外部環境對價格的影響，驗證新屋溢價與地段效應，並建立房價預測模型。

關鍵字：房價預測、機器學習、隨機森林、實價登錄、左營區

Abstract

In recent years, driven by the northward shift of Kaohsiung City's development focus, the establishment of TSMC's manufacturing facilities, and the transportation advantages of the three-railway terminal, Zuoying District has emerged as the commercial and transportation hub of Northern Kaohsiung. Consequently, the structure of its real estate market has undergone significant transformation. This study aims to investigate housing price fluctuations and determinants for residential high-rises, mid-rise elevator buildings, and walk-up apartments in Zuoying District from November 2020 to November 2025. Utilizing Python web crawling techniques, 10,993 valid transaction records were extracted from the Ministry of the Interior's Real Estate Actual Transaction Price Query Service. Data cleaning, feature engineering, and outlier processing were performed to ensure data quality.

Integrating descriptive statistics, geospatial analysis, and machine learning algorithms, this study compares the predictive capabilities of four models: Linear Regression, Decision Tree, Random Forest, and Gradient Boosting Trees. The results indicate that the Random Forest model achieved the best performance, yielding a coefficient of determination (R^2) of 0.8409 and a Root Mean Square Error (RMSE) of 35,778 TWD/ping. This significantly outperforms the traditional Linear Regression model (R^2 = 0.53), demonstrating the existence of highly non-linear relationships among housing price determinants.

Feature importance analysis identifies "Building Age" as the most critical factor influencing housing prices in Zuoying District, accounting for 35.85% of importance, followed by "Transaction Year" (25.11%) and "Geographic Location" (Latitude and Longitude combined, totaling 16.59%). Spatial analysis reveals a concentric price gradient centered on the Zuoying High Speed Rail Station. The core zone, including the HSR special zone and the Lotus Pond waterfront, commands the highest unit prices, averaging over 350,000 TWD/ping. This is followed by the Kaohsiung Arena commercial district and the city center, while peripheral areas and older communities exhibit relatively lower prices. Through data analysis, this study quantifies the impact of various property attributes and external environmental factors on prices, verifying the new house premium and location effects, and establishes a robust housing price prediction model.

Keyword：Housing Price Prediction, Machine Learning, Random Forest, Real Estate Actual Transaction Price, Zuoying District

目　次

圖目錄

表目錄

第一章 緒論

1.1 研究背景與動機

近年來，高雄市都市發展重心呈現北移趨勢，帶動房地產市場出現結構性的價格推升，其中北高雄的發展尤為引人注目。然而，促成左營近幾年房價變化之因，主要源於產業轉型、交通運輸及總體經濟三者交織所影響。高雄作為南台灣半導體產業廊帶，隨者民國 110 年台積電宣布於鄰近的楠梓區設廠，此一重大產業利多不僅僅重塑楠梓的產業地貌，更產生外溢效應，使鄰近且生活機能更為完善的左營區成為科技業人才居住選擇之一。隨著高科技人才進駐，市場對左營區的住宅需求從在地剛性需求擴大至科技移民的置產需求，隨之而來的便是水漲船高的房價。隨著台灣一日生活圈的成熟，左營區匯聚高鐵左營站、高雄捷運紅線、台鐵左營站三鐵共構成為一大優勢，大幅縮短南北商務與通勤的距離，使得左營成為外地置產客、北客南下投資以及高階商務人士的首選熱區。交通便利的地理優勢加上發展成熟的漢神巨蛋商圈，已然確立其作為北高雄商業與交通核心的樞紐地位。

其中，過往研究多聚焦於宏觀區域均價，較少探討個別房屋條件對價格的影響。本研究取得之資料集涵蓋完整的建物屬性與精確的地理坐標，不僅能分析地段帶來之增值，且能探討在房價高漲的時代，消費者對於屋齡新舊、有無管理等居住品質之願付價格是否也隨之改變。

綜上所述，左營區在產業紅利與交通運輸的驅使下，房地產市場結構已發生明顯的變化。因此，本研究旨在探討民國 109 年至 114 年共 5 年期間，左營區房價波動，其中分析包含不同地區(如高鐵特區、巨蛋商圈及左營舊城)之價格差異、五至六年間房價成長幅度與多重變數衝擊之關聯，冀能透過數據分析，作為未來市場判斷的參考依據。

1.2 研究目的

本研究之目的可歸納如下：

一、針對近五年之交易資料，藉由計算各年度及各季度之平均單價、總價與交易量，繪製量價走勢圖，以釐清在疫情前後、台積電設廠期間，左營區大樓、華廈與公寓產品是否呈現顯著價量結構改變。

二、利用經緯度與路名資訊，將左營區細分為不同分區，比較不同路段與生活圈的房價表現並分析交通節點對周邊房價的影響範圍與強度。

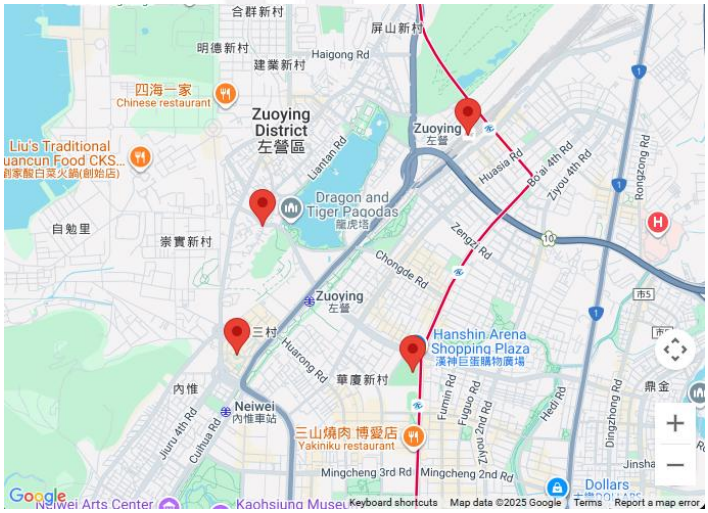
三、根據現有爬蟲獲得之資料，透過模型評估與參數的調整，預測民國 114 年後房價之走勢及發展潛力，期望研究成果能為自住購屋者、置產投資人提供決策參考。

1.3 研究範圍與限制

一、時間範圍:本研究之資料時間軸設定為民國 109 年 11 月至民國 114 年 11 月，共計約 5 年。

二、區域範圍：鎖定於高雄市左營區，根據內政部不動產交易實價查詢服務網透過爬蟲抓取到之路名與經緯度資料，主要分為以下三個區域進行分析，如下圖 1 紅色地標所標示處。

1. 高鐵特區：以高鐵左營站為核心，其核心地標為高鐵左營站、新光三越左營店，路線涵蓋高鐵路、華夏路等周邊路段，為三鐵共構的交通樞紐，周邊商圈發展較新。
2. 巨蛋商圈：以漢神巨蛋為核心，其核心地標為漢神巨蛋購物廣場、高雄巨蛋體育館、捷運巨蛋站，路線涵蓋博愛二路、富民路等路段，為北高雄最繁華的商業中心。
3. 左營舊城與果貿生活圈：包含左營大路、中華一路等早期發展區域，其核心地標為蓮池潭、鳳山縣舊城、果貿圓環社區等，具有濃厚的歷史文化氣息。



資料來源：Google Map 地圖擷取

圖 1 主要區域範圍

三、研究對象：考量資料的一致性與統計分析之代表性，本研究僅鎖定住宅大樓(11 層以上具電梯)、華廈(10 層以下具電梯)及公寓(5 層以下，無電梯）三種建物型態，排除透天厝、店面、商辦、工廠等數據，以確保分析結果能精確反映多數民眾之居住市場行情。

四、研究限制：本研究在資料蒐集與分析過程中，受限於數據本身特性與外部環境變數，存在以下限制：

1. 本研究使用之資料集僅包含大樓、華廈與公寓等分層住宅，未納入透天厝之交易紀錄，因此分析結果較適用於反映現代化集合住宅之市場趨勢，無法完全代表左營區整體房地產全貌。
2. 房地產價格受多重屬性影響，雖然本研究已納入屋齡、樓層、車位等變數，但部分影響價格之不可觀察變數，如是否包含裝潢或為毛胚屋此類可能導致部分交易價格偏離模型預測，但無法透過實價登錄資料完全量化，為本研究統計上之限制。
3. 本研究主要依賴量化數據進行分析，然而房地產市場往往受投資客心理預期影響甚鉅。例如台積電設廠消息傳出當下之市場氛圍，雖可透過價格跳漲觀察，但具體預期心理強度難以直接數值化，僅能透過質性描述加以輔助說明。

1.4 研究內容

本研究透過 Python 爬取內政部不動產交易實價查詢服務網高雄市左營區之資料，預先將資料進行清理，包含清理缺失值、離群值，以分析及預測更為精準的房屋價格，最後透過迴歸系列模型進行訓練，並調整參數優化模型提高預測數年後房價之準確性，本研究之研究內容可分為以下幾點：

一、**樣本資料結構與敘述性統計：**說明數據收集與清理步驟，包括如何處理異常值、填補缺失數據，確保模型訓練及分析所需數據一致和完整，並計算全區單價與總價的平均值、中位數、標準差，排除極端值干擾。

二、**時序趨勢分析：**將資料依時間序列展開，探討價格隨時間的變化，觀察各季度的變化，探討是否有特定季節或政策發布點造成交易量瞬間萎縮。

三、**建物屬性特徵之影響分析：**利用屋齡、車位、樓層等欄位，透過散佈圖或折線圖之圖表形式，深入分析影響房價的微觀因素。

1.5 研究流程

研究流程圖如圖 2 所示，研究詳細步驟如下：

一、**確立研究主題：**確認研究所需使用之高雄房價資料背景、研究動機、後續資料分析欲使用方法及界定研究範圍等，本次以高雄市左營區之房價作為研究主題。

二、**蒐集左營區房價數據**：透過 Python 程式碼自內政部不動產交易實價查詢服務網中搜尋欄位查詢「高雄市左營區」且將年份設定為 109 至 114 年，透過爬蟲抓取數據，並匯入至新 excel 檔中以利後續分析，內容包含屋齡、交易日期、總樓層、經緯度等作為分析主要資料。

三、**資料前處理**：將所有欄位資料按照所設定之順序進行排列，並將有缺漏及離群數值用平均值取代，避免進行資料分析及預測時因資料缺漏或離群值過大影響準確性。

四、**針對個別變數進行編碼**：所爬取之資料大多為文字資料，為方便後續資料分析及機器學習演算發法分析之方便及需要，將文字欄位成數值型態，如："有"轉換為 1，"無"則轉換為 0。

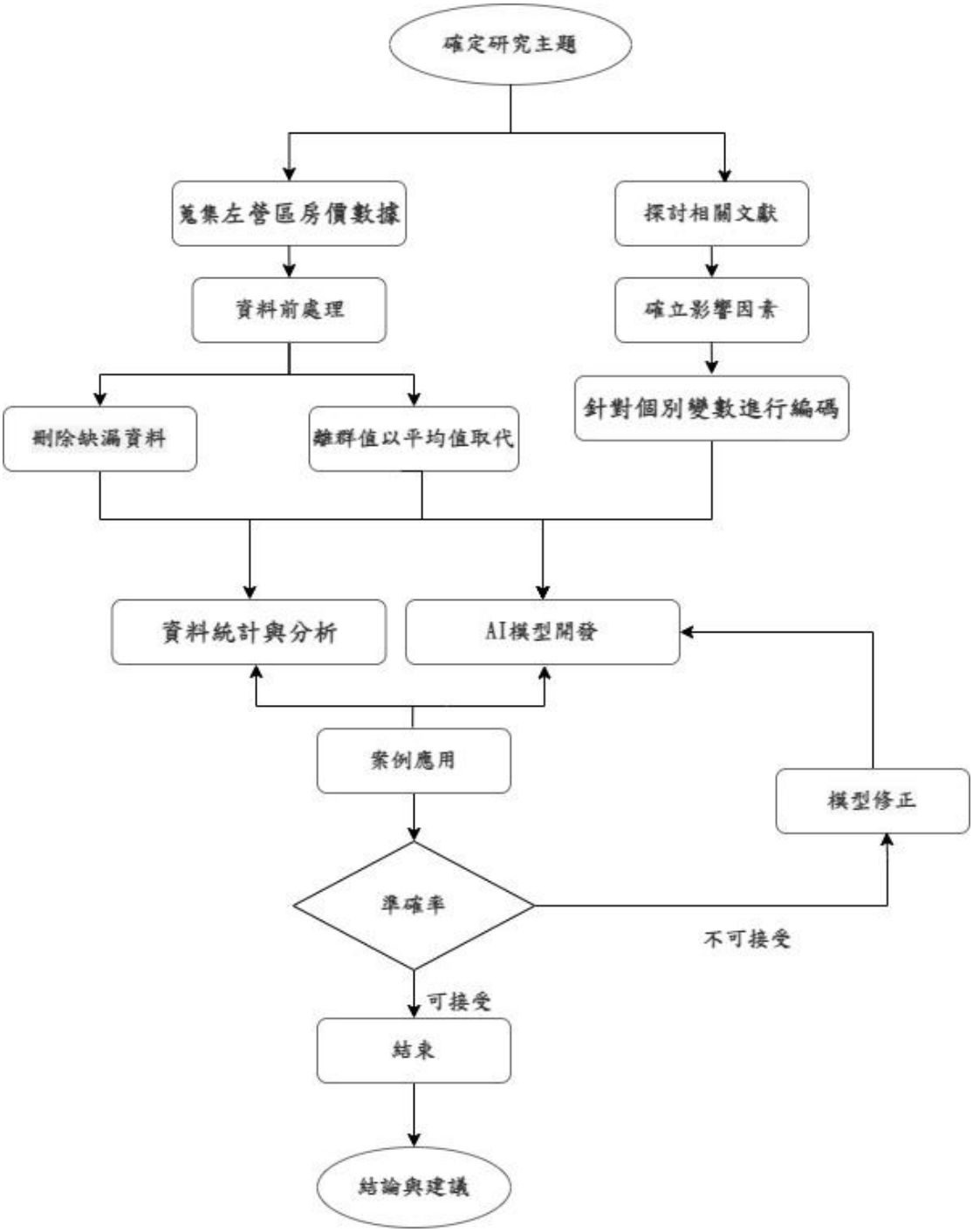
五、**刪除無需使用資欄位**：若該欄位無分析價值、缺漏值過多且無法以平均值進行差補或資料量過少之情形，則考慮將該欄位刪除，以針對重點欄位進行分析。

六、**資料統計與分析**：將透過敘述性統計、時序趨勢分析、空間差異分析、屬性特性分析等方法探討平均行情、繪製房價走勢圖、比較三大生活圈之價格梯度與漲幅差異、探討屋齡折舊對價格之影響力等面向。

七、**建構機器學習模型**：使用 Random Forest、邏輯迴歸等不同機器學習模型進行比較，加入測試集與訓練集資料進行多次迭代以優化模型參數，並針對不同影響因素建構模型。

八、**案例應用**：將分析完的結果及訓練好的模型應用在實際案例中進行預測並評估準確率，並判斷模型準確率是否可接受，若不可接受則需重新訓練模型。模型訓練完畢後，透過評估指標對建構出模型進行評估以確保誤差值越小，讓預測數值更為準確。

九、**結論與建議**：歸納左營區房市之發展特徵，並針對自住客、投資者及未來市場展望提出具體建議。



資料來源：研究者自製

圖 2 研究流程

第二章 文獻回顧

2.1 影響房地產價格之因素

邱司杰（2014）[3]利用內政部實價登錄以及房仲網站等多源資料，建構房地產交易資料庫與標準化收集流程，確保房價資料的長期與穩定取得。透過網頁與視覺化技術，發展靜態與互動式圖表及房價地圖，協助使用者掌握各區房價水準、結構與時間趨勢。以實價登錄交易資料為基礎，進行資料清理與異常值處理後，運用 SPSS 線性迴歸與半對數模型，建立區位別房價預測模型，並以坪數、屋齡、樓層、建物型態與格局等特徵作為解釋變數，同時以 SSE 等指標檢驗模型誤差與配適度。研究結果顯示，面積、區位與屋齡等變數對房價具顯著影響，建構之模型除可作為買賣雙方訂價與政府課稅參考，亦能輔助偵測價格異常案件並觀察房市走勢，提供後續自動估價與時間序列分析之基礎。

邱恩華（2017）[4]以內政部實價登錄資料，蒐集 2012 至 2016 年屏東市住宅交易資料，運用多元線性迴歸分析房屋屬性與區位因素對不同建物型態（透天厝、大樓華廈、公寓）單價之影響。結果顯示，三種建物中屋齡皆對單價具有顯著負向效果，顯示老屋折舊與維護成本會壓低價格；透天厝之建物面積對單價為正向，但在大樓華廈與公寓則呈負向，反映透天以大家庭需求為主，而大樓、公寓多為小家庭空間偏好。

Owusu-Manu 等人（2019）[2]，利用來自加納三個地區（Accra、Kumasi、Takoradi 及 Tema）房地產代理商的 270 筆住宅交易資料，應用回歸模型（Hedonic Regression）估計住宅屬性對房價之相對影響。結果顯示，位置（住宅等級）為最強決定因素：一級住宅區相較五級區，房價高 156%；四級區高 50%，凸顯優質社區規劃、基礎設施（如道路、排水）及執法對房價之關鍵作用。其他顯著屬性包括臥室數（每增加一間貢獻 16%）、樓層數（每增加一層 13%）、總樓地板面積、土地大小、房屋年齡（>5 年降 11%）及豪華裝潢（貢獻 14%），而衛浴數及豪華配件影響較小。模型解釋力達 70%，無多共線性問題，證實加納住宅市場異質性強，受都市化與供給不足影響。

Chen.和 Hsu.（2020）[1]以高雄市為案例，利用 2011 至 2015 年 174,913 筆住宅交易資料，建構價格模型（Hedonic Price Model），探討不同類型多模式鐵路站（單模式、雙模式、多模式）對住宅價格彈性之影響。結果顯示，高雄捷運站周邊 1km 內，每增加 100m 距離，平均住宅價格下降新台幣 25.8 萬元；雙對數模型下，距離每增加 1%，價格彈性為-0.067%。

2.2 機器學習應用於房價預測

邱國祥（2020）[5]基於台中市實際登錄房價數據，採用多種機器學習模型進行房價預測，方法包括多元線性迴歸、正則化迴歸（Lasso 與 Ridge 迴歸）、隨機森林及 XGBoost 等演算法。研究利用 2018 至 2019 年的房產資料，共 16,328 筆交易紀錄，特徵涵蓋總價、坪價、房間數、屋齡等多個維度，透過逐步特徵選擇、Lasso 迴歸特徵篩選等方法，將 2,642 個原始特徵降維至更精簡的特徵集合。研究結果說明，XGBoost 在 1,000 個估計器配置下表現最優，均方根誤差（RMSE）為 215.50，平均絕對百分比誤差（MAPE）為 24.40%，優於線性迴歸與隨機森林迴歸。

許雅晶（2022）[6]以新北市林口區 2020 至 2021 年實價登錄與經濟指標資料為基礎，建構房價預測模型，探討機器學習在區域房價估算之應用與準確度。研究首先整理實價登錄與國內各業生產指數資料，萃取房數、廳數、衛數、屋齡、坪數、有無車位、地址特徵及批發零售、運輸、金融、不動產服務等指標作為解釋變數。採用隨機森林回歸、支持向量回歸、K 近鄰回歸、決策樹回歸及集成學習之 Stacking，比較不同模型之測試集評分與平均百分比誤差。結果顯示隨機森林與 Stacking 表現最佳（R^2約 0.93，誤差約 12%），但整體誤差仍不如以複回歸並先行切割一般住宅與豪宅之既有模型，顯示資料分群與極端值處理對預測精度具關鍵影響。

陳玟寧（2022）[7]透過內政部實價登錄公開資料，建立台北市房價機器學習預測模型，比較 Lasso、Ridge、Elastic Net、SVR、KRR 等迴歸演算法之預測表現。結果顯示，迴歸式機器學習適合用於連續型房價預測，其中以支持向量迴歸（SVR）表現最佳，核嶺迴歸（KRR）緊追其後，正規化迴歸雖精度略低但運算速度快。透過最佳子集迴歸，研究識別出建物移轉總面積、主建物面積、房間數、車位總價、主要建材、鄉鎮市區與交易標的等七項為影響房價最重要之變數，並以 2016 至 2020 年資料建模、2021 年新資料驗證，證實模型具穩定預測能力。

林于軒（2025）[8]以堆疊式集成學習架構構建房價預測模型，整合實際登錄價格（特徵 A）、環境外部性（特徵 B）、市場情緒（特徵 C）與總體經濟因素（特徵 D）等多維度特徵。採用線性迴歸、隨機森林與 LightGBM 作為基礎模型，LightGBM 作為中介模型進行集成。研究透過時間序列交叉驗證（TSCV）評估模型在不同特徵組合下的預測績效。實驗結果顯示，堆疊式集成模型整體優於單一基礎模型，當特徵 A、B、C 組合時，平均絕對百分比誤差（MAPE）達 9.40%；加入特徵 D 後，MAPE 進一步降低至 9.14%，決定係數（R^2）達 0.94。

第三章 資料預處理

本研究為取得高雄市左營區房地產交易之歷史數據，採用 Python 程式語言透過爬蟲套件工具，爬取政府公開數據庫對資料進行抓取，以下將對資料抓取、資料清洗與預處理、特徵工程與非結構化文字轉換、類別變數處理與編碼、地址資料清洗與空間特徵提取及遺失值處理等資料處理步驟詳細說明。

3.1 資料抓取

一、資料來源：本研究使用數據來源為內政部地政司建置之「內政部不動產交易實價查詢服務網」，該平台提供具公信力之不動產交易資訊，涵蓋交易標的、價格、面積、屋齡及地理位置等關鍵變數，選用之查詢條件為「高雄市左營區」，時間範圍為民國 109 年 11 月至民國 114 年 11 月，共計五年。

二、資料蒐集方法：為有效獲取大量結構化數據，研究透過 Python 程式語言配合以下關鍵套件進行自動化蒐集。

1. **請求模組：**利用 requests 套件發送 HTTP GET 請求至內政部伺服器，為避免伺服器將程式誤判為惡意攻擊或機器人，程式碼中特別設置「請求標頭」，將 User-Agent 參數設定為一般瀏覽器，模擬真實使用者之瀏覽行為，確保連線請求能順利通過伺服器驗證。
2. **錯誤處理機制：**考量網路傳輸可能不穩定，程式設計中包含例外處理 Try-Except 機制。針對 HTTP 錯誤、連線超時或 JSON 解析失敗等狀況進行偵測，確保爬蟲在遇到網路波動時能輸出錯誤提示而非直接崩潰，提升資料蒐集的穩定性。

三、爬取資料內容：透過分析網站回傳之 JSON 格式，研究所爬取之資料包含房屋特徵與交易資訊等多維度欄位，原始資料經解析後轉換為 Pandas DataFrame 結構。

四、資料轉換與欄位編碼：由於內政部原始資料庫之欄位名稱皆為英文代碼，如 AA11,tp,lat 等缺乏可讀性且較不直觀，為利後續統計分析與機器學習模型之建構，本研究進行以下資料轉換工作。

1. **欄位名稱映射：**將原始英文設定代碼轉換為具語義之中文欄位名稱，例如將「tp」轉換為「總價」、「p」轉換為「單價」、「g」轉換為「屋齡」等，共計處理約 40 個欄位，確保分析變數欄位名稱較為易讀。
2. **格式統一與儲存：**完成欄位更名後，使用 Pandas 套件將整理完畢之 DataFrame 輸出為 csv 格式檔案，在儲存過程中，特別指定編碼格式為 utf-8-sig，以解決 Excel 開啟 csv 檔時常見的繁體中文亂碼問題，主

要爬取原始欄位、轉換後欄位名稱及資料說明如下對照表 1 所示。

表 1 實價登錄資料欄位定義與說明表

原始欄位	轉換後欄位名稱	資料說明
AA11	主要用途	住家用、商業用、工業用
AA12	主要建材	鋼筋混凝土 RC、鋼骨 SC
a	地址	交易標的之完整門牌地址
b	建物型態	住宅大樓、華廈、公寓、透天厝
bn	社區_建案名稱	交易標的所屬社區大樓或建案名稱
bs	主建物占比	主建物面積占總移轉面積之百分比
city	縣市代碼	E 代表高雄市
commid	社區代碼	系統賦予該社區的唯一識別碼
cp	車位價格	車位之單獨交易價格
e	交易日期	不動產買賣契約簽訂之日期
el	電梯	建物是否設有電梯設備
es	公設比	共有部分面積占總面積之比例
f	樓層	交易標的所處之樓層
fi	總樓層	該棟建物之地上總樓層數
g	屋齡	建物建築完成日至交易日之年數
id	系統編號	實價登錄系統內部之案件流水號
j	土地筆數	該次交易包含之土地筆數
k	建物筆數	該次交易包含之建物棟數
l	車位筆數	該次交易包含之車位數量
lat	緯度	交易標的之地理緯度座標
lon	經度	交易標的之地理經度座標
m	管理組織	社區是否有管理委員會或管理組織
mark	標記	特殊交易註記（如：急買急賣）
msg	單價備註	針對單價計算方式之補充說明
note	備註	特殊情況之說明(如：含裝潢)
p	單價	建坪單價，通常已扣除車位價格
parkmain	車位類別	坡道平面、升降機械
pimg	圖片	交易標的之格局圖或外觀圖連結
pu	使用分區	住宅區、商業區

punit	單價單位	單價之計價單位
r	路寬	建物臨路之道路寬度
reid	關聯 ID	關聯案件之識別碼
s	總面積	包含主建物、附屬建物及公設
sq	交易 ID	該筆交易之唯一識別碼
t	交易標的	房地(土地+建物)、房地+車位
town	行政區代碼	A02 代表左營區
tp	總價	交易之總金額
tunit	總價單位	總價之計價單位
type	資料類型	買賣、預售屋
unit	單位	面積之計量單位（平方公尺、坪）
v	格局	3 房 2 廳 2 衛

五、刪除無用欄位：原始爬取之實價登錄資料集共包含 41 個欄位，雖然資訊詳盡，但其中包含大量與房價預測模型無關之行政資訊，如：系統編號、關聯 ID、重複性描述，如：交易標的、縣市代碼以及非結構化的文字備註。過多的無效變數不僅會增加運算負擔，更可能引入雜訊干擾後續分析結果，因此將以下三類無效欄位刪除。

1. **行政管理代碼：**如 id（系統編號）、reid（關聯 ID）、town（行政區代碼，因本研究已鎖定左營區，故此欄位無變異性）。
2. **冗餘或重複資訊：**如 city（縣市代碼）、punit（單價單位）、tunit（總價單位）。
3. **非結構化或低相關資訊：**如 note（備註，多為雜亂文字）、pimg（圖片網址）、mark（標記）。

經篩選後，最終保留對房價具備解釋力之 17 個關鍵變數，作為本研究之分析資料集，資料如下表 2 所示。

表 2 特徵篩選與保留欄位說明表

欄位名稱	欄位名稱說明
主要用途	住
地址	左營區文守路 137 巷 0041 號#左營區文守路 137 巷 0041 號
建物型態	住宅大樓(11 層含以上有電梯)
主建物占比	56.14%
交易日期	114/10/11
電梯	有
電梯	十六層/二十層
屋齡	37
車位筆數	2
緯度	22.66248
經度	120.3044
管理組織	有
單價	408,975
使用分區	住家用
總面積	37.04
總價	9,600,000
格局	3 房 2 廳 2 衛

3.2 資料清洗與預處理

原始爬取之實價登錄資料雖然豐富，但存在格式不一致、包含非數值字元及極端值等問題，為確保後續統計分析與機器學習模型之準確性，本研究透過 Pandas 套件對資料進行清洗。

一、欄位重整與排序：為提升資料之易讀性與邏輯性，首先剔除與房價分析無直接相關之冗餘欄位，並將保留之 17 個關鍵變數依照「地理位置」、「物理屬性」至「價格資訊」之邏輯順序重新排列，重整後之欄位順序如下：緯度、經度、地址、交易日期、屋齡、總面積、樓層、格局、車位筆數、電梯、管理組織、建物型態、主要用途、使用分區、主建物占比、單價、總價。

二、**日期格式標準化:**原始資料中之「交易日期」採中華民國紀年,如:109/05/20,且部分格式為字串，為利於後續之時間序列分析及季度劃分，本研究定義轉換函式，將年份加上 1911 換算為西元紀年，並將欄位統一轉換為 datetime 時間物件格式，如：2020-05-20，若遇格式錯誤之資料，則強制轉換為 NaT 並予以剔除，以確保時間軸之正確性。

三、**數值型別轉換與清洗：**在原始數據中，「總面積」、「單價」與「總價」等關鍵連續變數包含千分位逗號，如：15,000 且被儲存為物件型別，無法直接進行數學運算。本研究透過字串處理函式移除逗號，並將其強制轉換為浮點數或整數型態，此步驟確保後續計算平均單價、總價漲幅以及相關係數矩陣之可行性。

四、**離群值檢測與處理：**考量房地產資料常存在極端值，例如特殊豪宅、包含大量土地之交易或登錄錯誤之案件，此類數據將提高整體平均值並降低模型預測力，因此研究針對「總面積」變數採用四分位距法進行離群值篩選，步驟如下：

1. 計算四分位數：計算總面積之第一四分位數(Q_1)與第三四分位數(Q_3)。
2. 定義四分位距：$IQR = Q_3 - Q_1$。
3. 設定門檻值：定義離群值上限為$Q_3 + 1.5 \times IQR$。
4. 排除極端值：將總面積超過上限之樣本視為異常極端值予以剔除。

經上述 IQR 規則過濾後，不僅保留絕大多數正常交易樣本，亦有效排除坪數過大，可能為商場或特殊地產之極端案件，提升樣本之代表性與同質性，資料處理前後對照如下表 3 所示。

表 3 資料預處理前後對照表

處理項目	處理前範例	處理後範例	目的
日期轉換	1090520	2020-05-20	統一時間軸，利於時序作圖
數值清理	1,250	1250	去除逗號，轉為數值以利運算
離群值處理	總面積 200 坪	該筆資料已被刪除	避免極端豪宅影響整體平均

3.3 特徵工程與非結構化文字轉換

原始實價登錄資料中，「樓層」與「格局」欄位皆以複合式字串與中文數字記載，無法直接應用於統計模型，為提取其中隱含的定量資訊，本研究透過Python程式進行以下兩項關鍵的特徵工程處理：

一、樓層資訊之結構化與數值轉換：「樓層」變數原始格式為「移轉樓層/總樓層」，如：八層/十四層，且採中文數字書寫，考量「所在樓層」影響視野與採光，而「總樓層」反映建物規模與法規配置，兩者對價格之影響機制不同，故需拆分為獨立變數，處理步驟如下：

1. 拆分變數：以「/」符號作為分隔符，將原始欄位拆解為「物件樓層」與「總樓層數」兩個衍生欄位。
2. 中文數字轉碼：建構轉換函式，將中文數字（如一、二...十）轉換為阿拉伯數字（1, 2...10）。
3. 語意邏輯處理：針對透天、特殊物件或意指整棟交易出現之「全」字，將其「物件樓層」數值直接替換為「總樓層數」，以符合實際交易之物理現狀。

二、格局資訊之正則表達式提取：變數原始格式為3房2廳2衛，為量化房屋之內部配置價值，本研究運用正規化進行特徵提取：

1. **特徵拆解：**設定(\d+)房、(\d+)廳、(\d+)衛等匹配模式，分別提取字串中「房」、「廳」、「衛」前方的數字，生成三個獨立數值型欄位。
2. **缺失值填補：**若原始資料中缺乏特定配置，如套房物件僅有1房1衛、缺廳，程式將自動捕捉缺失值並填補為0，確保資料矩陣之完整性。

經上述處理後，原本無法運算之文字描述已轉化為具備次序性或數值性之量化指標，將大幅提升後續房價預測模型之解析能力，轉換後結果如下表4所示。

表4 欄位拆分

樓層	物件樓層	總樓層數	格局	房	廳	衛
十層/十四層	10	14.0	3房2廳2衛	3	2	2
十四層/十五層	14	15.0	2房1廳1衛	2	1	1
五層/十三層	5	13.0	3房2廳2衛	3	2	2

3.4 類別變數處理與編碼

實價登錄資料中包含多項類別型變數，為使其能應用於機器學習演算法，研究針對二元變數與多元變數分別採取不同的處理策略，並進行雜訊清洗以減少模型維度。

一、二元變數之標籤編碼：針對僅有兩種狀態之屬性欄位，包括「電梯」與「管理組織」，研究採用標籤編碼將文字轉換為數值，轉換規則為將「有」編碼為 1，將「無」編碼為 0。

二、建物型態與用途之清洗與簡化：原始資料中，「建物型態」常包含括號補充說明，如：住宅大樓(11 層含以上有電梯)，為統一分類標準，運用字串處理移除括號及其內容，將其簡化為「住宅大樓」、「華廈」、「公寓」等類別。此外，針對「主要用途」欄位，剔除「其他」、「工」等非典型住宅用途之樣本，以確保分析對象之同質性。

三、多元變數之獨熱編碼：針對「建物型態」與「主要用途」等無次序性之多元類別變數，若直接編碼為 1, 2, 3 可能導致模型誤判其大小關係。因此，本研究採用獨熱編碼。

1. 虛擬變數生成：將每個類別轉換為獨立的二元欄位，如：建物型態_公寓、建物型態_華廈。
2. 預防多重共線性：為避免虛擬變數陷阱，在編碼過程中設定 drop_first=True，即刪除第一個參考類別，確保變數間彼此獨立。

3.5 地址資料清洗與空間特徵提取

地址欄位隱含極具價值之地段資訊，為從非結構化的地址字串中提取路段特徵，對資料進行清洗，步驟如下：

一、地址字串標準化：原始地址欄位常包含重複資訊或系統註記，如「#」符號，本研究透過字串切割，僅保留「#」符號後之有效地址，並移除「左營區」等冗餘行政區名稱，將地址簡化為如「重仁路 249 號十樓」之精簡格式。

二、路名特徵提取：路段往往決定房價基本行情，如：博愛二路 vs. 巷弄，運用正規表達式針對清洗後之地址進行樣式匹配。

1. 提取規則：優先抓取以「路」、「街」、「大道」結尾之字串；若無，則次級抓取以「巷」結尾之字串。
2. 特徵生成：成功提取之資訊儲存於衍生欄位「路名」，作為後續分析路段價差之關鍵變數。

3.6 缺失值處理

在完成上述清洗與特徵工程後，針對資料集中仍存有之缺失值，本研究依據變數特性採取兩種處理機制：

一、**中位數填補**：針對「屋齡」變數，考量其為連續變數且數據分佈可能呈現偏態，直接刪除將導致樣本數量過少。因此，計算全樣本屋齡之中位數進行填補，選擇在於中位數較不易受極端老屋或新建案之影響，能更穩健地代表中心趨勢。

二、**刪除法：**針對難以準確估計或屬性極為關鍵之變數，包括「主建物占比」、「單價」、「格局」、「樓層資訊(物件/總樓層)」及「主要用途」，若發生缺失狀況，顯示該筆交易資訊本身不完整或登錄有誤，為避免錯誤填補造成模型偏誤，將直接刪除含有上述缺失值之樣本。

地址資料及樓層資料拆分後，資料集包含 23 個欄位，涵蓋地理資訊(緯度、經度、地址)、交易資訊(交易日期、單價、總價)、建物特徵(屋齡、總面積、樓層、格局、建物型態)、以及配套設施(電梯、管理組織、車位筆數)等多個維度，經過上述資料清理及缺失值處理之過程，原資料清理前所抓取之數據資料為 12174 筆，資料清理後刪除無效資料筆數為 1181 筆，留下具有分析價值之數據為 10993 筆資料，後續將針對所抓取及清理後之資料進行分析。

第四章 資料分析

4.1 地理空間分析

為解高雄市左營區近 5 年房屋總價之分布情況，研究首先採用 Python 程式語言作為資料分析工具，透過直方圖針對目標變數房屋總價進行分析，以視覺化方式呈現。首先利用 Pandas 套件讀取原始資料集，並透過 Matplotlib 設定中文字體 Microsoft JhengHei 以確保圖表標籤顯示正確，並透過 Seaborn 套件中的 histplot 函式進行繪製，為清楚呈現數據分布，將組距 bins 設定為 50，同時透過核密度估計參數，以平滑曲線疊加於直方圖上，藉此觀察房價集中趨勢與偏態情形。

由圖 3 總價分布圖可得知，X 軸為房屋總價，單位是 TWD(新台幣)，其刻度顯示為科學記號10^7，即千萬，如 0.5 代表 500 萬，1.0 代表 1000 萬、Y 軸為頻率，即在特定價格區間內的房屋數量；藍色曲線為核密度估計曲線，用於觀察分佈趨勢。由圖中可發現其為一右偏分佈，大部分房價主要集中於左側價格較低區域，右側尾部較高價格之房價則趨於減少，圖中之價格高峰主要落在 700 萬至 900 萬台幣之間，說明此價格區間之物件較多亦是市場主要交易價格，且由於呈現右偏分布，通常會導致平均數大於中位數大於眾數，使得平均價格被右側高價格之物件提高。

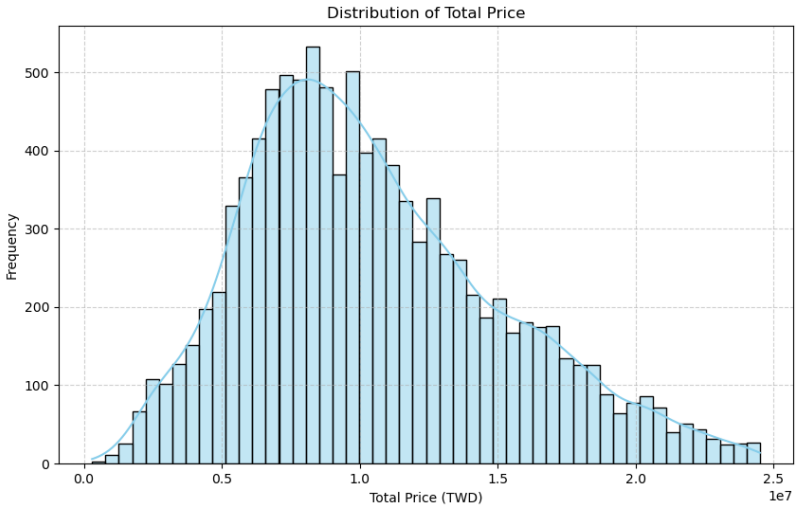


圖 3 房屋總價分布情形

下圖 4 以直方圖呈現單價之分布情形，橫軸為單價（元/坪），縱軸為交易次數。其中，紅色虛線代表平均單價 271,017 元/坪，綠色虛線代表中位數單價 264,374 元/坪，兩者的差距約 6,643 元/坪，顯示分布略呈右偏，存在少數高價物件拉高整體平均值。此右偏分布為房地產市場典型特徵，反映市場中主流產品集中在中等價位，而高端豪宅作為小眾市場雖然數量有限但對平均值產生顯著影響。

從整體分布形態來看，單價分布呈現明顯單峰集中特徵，最高峰落在 25 萬至 30 萬元/坪區間，此區間交易頻次接近 700 筆，為市場交易最活躍之價格帶。此峰值區間占據全部交易約 20%的比例，代表高雄房地產市場的主流產品定位。以統計學角度，此峰值略低於平均值與中位數所在位置，顯示市場需求最旺盛價格帶集中在中等偏下的區間，反映高雄作為南部都會區，整體房價水平較台北、新竹等北部城市為低，市場以服務中等收入家庭為主。

主流價格帶從 20 萬至 35 萬元/坪，涵蓋整體交易約 60-70%的比例，為高雄房地產市場的核心區段，此區間內分布相當集中，形成高聳的峰值，顯示市場供需在此達到最佳平衡點。而 20 萬至 25 萬元/坪對應左營區外圍、三民區、鳳山區等次級地段，或是屋齡 15 至 25 年中古大樓、25 萬至 30 萬元/坪則對應左營核心區之中古屋或新興重劃區之新成屋、30 萬至 35 萬元/坪多為左營精華地段。

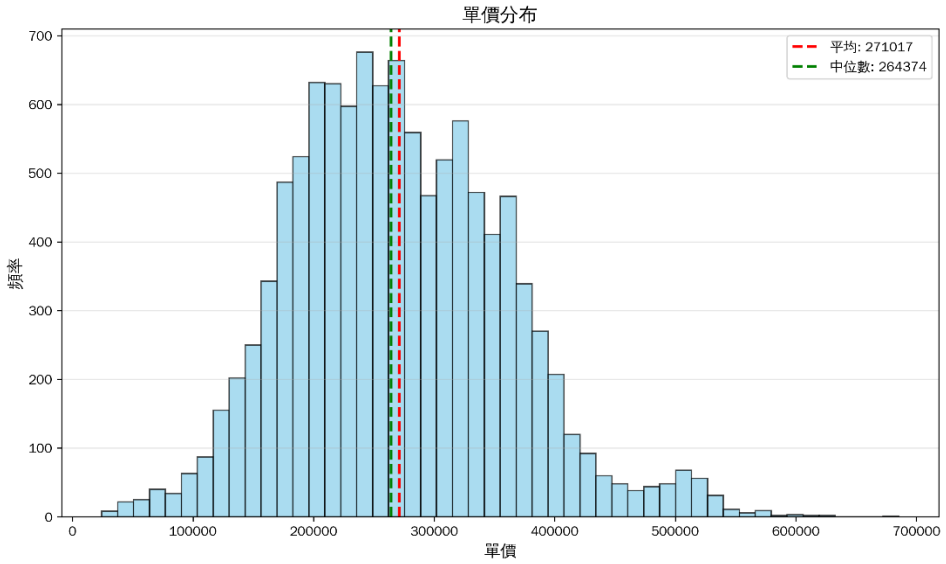


圖 4 房屋單價分布情形

圖 4 右側有一次波峰，研究將針對其價格約略落於 50 萬左右之區間進行分析，由下圖 5 平均單價最高路段 Top10 可得知，單價前三名分別為環潭路路段、重信路路段及保靖街路段。發現環潭路單價約為 56 萬可能為「京城建設-京城天湖」建案，該區域少數擁有蓮池潭完整首排景觀之新大樓，稀有性極高，因此單價高於周邊行情；重信路可能為「鑫龍騰建設-鑫悅或鑫高鐵系列」，重信路位於高鐵特區核心為典型高鐵捷運宅；保靖街可能為「城揚建設-珈柏麗」建案，其為近期交屋的新成屋，位於漢神巨蛋商圈北側，屬於該區域高價指標案。

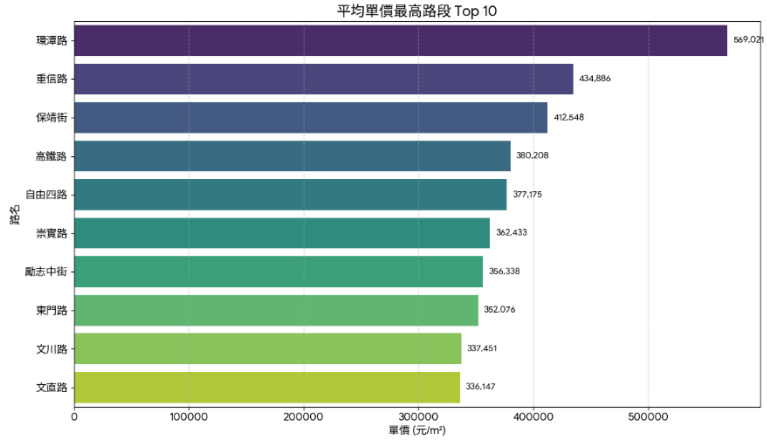


圖 5 高單價房屋路段分佈

下圖 6 為屋齡與單價之散佈圖，用於分析房屋老舊程度如何影響成交價格，X 軸為屋齡，範圍從 0 年之新成屋到約 55 年之老屋、Y 軸為單價，單位為TWD/m^2，紅色虛線為趨勢線，顯示整體數據的平均走向。

由圖中可得知紅色趨勢線明顯向右下方傾斜呈現負相關，證實屋齡越高，單價通常越低，隨著屋齡增加，建築物本身的價值折舊，導致整體交易單價下滑。X 軸 0 至 10 年區間數據點非常密集且分佈範圍極廣，分佈範圍自 20 萬至 60 萬以上，顯示新成屋是市場主要交易，且價格受建案品牌、地段、影響大，因此價格變異數大；其中，約略於 21 至 22 年出現垂直高單價柱狀分布，推測此類物件可能位於較佳地段因而不受屋齡折舊所影響；當屋齡超過 30 年，大部分交易單價皆落於紅色趨勢線下方，顯示 30 年以上老屋，建築價值已所剩無幾，房價主要由土地價值支撐，因此價格波動較小。

圖 6 屋齡與單價散佈圖

最後，由下圖 7 之相關係數矩陣結果進行說明，當顏色呈現深紅即數值接近 1.0，代表兩者呈現正相關，A 越高，則 B 隨之越高；當顏色呈現深藍即數值接近 -1.0，代表兩者呈現負相關，A 越高，B 則反之越低；當顏色呈現淺白即數值接近 0，代表兩者無關連，A 變高變低，B 則不太受影響。

圖中呈現資訊可發現，總價與總面積之相關係數達 0.68，呈現正相關，說明房屋總價多寡與坪數大小有相關聯；單價及總面積之相關係數卻是-0.04，推測房屋總面積大並不代表單價較高，但總價一定會因為房屋面積大而導致價格疊加。相關係數矩陣圖中屋齡與單價之相關係數為-0.42 為圖中深藍色最明顯區塊之一，同樣證實圖 6 之分析結果，說明房屋屋齡增加，使得整體單價下滑。其中，車位比數與總價間之相關係數達 0.5，比樓層高度之影響大，可推測有車位說明該物件總價不低，因都市地區寸土寸金，因此有車位之建案不論是大坪數亦或是公寓社區，通常總價皆不斐。

房間數與單價之相關係數為-0.28 呈現負相關，說明房間越少，單價反而越高，房間越多單價反而越低，推測可能為現今市場小宅化趨勢，少子化使得民眾無須買太大的房，建商為房屋得以出售，將坪數縮小拉高單價，以至於買 1 房或套房之民眾雖總價低但換算單價卻相對提高。最後，總樓層數與單價之相關係數為 0.29 呈現正相關，顯示當樓層約高單價通常越高。

圖 7 相關係數矩陣分析

4.2 機器學習預測

以下將透過機器學習對左營區房價進行分析，利用迴歸分析測試四種機器學習模型以預測房價，包含線性迴歸、決策樹、隨機森林及 Gradient Boosting Trees，並透過均方根誤差（RMSE）、平均絕對誤差（MAE）與決定係數（R^2）等指標評估模型效能。亦透過時間序列分析，觀察房價與交易量的月度變化趨勢，識別市場週期與季節性特徵，最後透過綜合統計分析，自多維度呈現市場的整體樣貌，包括價格分布、特徵關聯性與設施影響等。

一、機器學習方法介紹

1. 線性與非線性迴歸

線性迴歸和非線性迴歸兩者皆為常用的統計建模方法，用於描述目標變量與自變量之間的關係。線性迴歸模型結構簡單、便於解釋，且計算速度快、效率高，適合數據中線性關係明顯的情況。而非線性迴歸能捕捉更複雜的數據模式，處理自由度更高的非線性結構，因此適合機器學習中的複雜迴歸問題，應用範圍涵蓋非線性動態系統建模等場景，其特徵統整如表 5 所示。

表 5 線性與非線性回歸統整

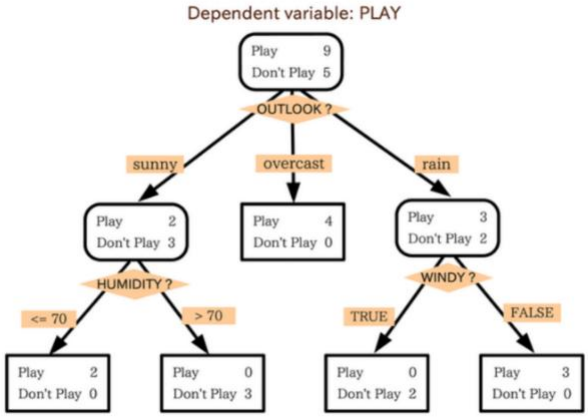
特徵	線性迴歸	非線性迴歸
解釋性	較容易且直觀	較為複雜
訓練方法	基於解析解（如最小二乘法）或梯度下降方法即可	需要非線性優化算法（如梯度下降、牛頓插值法等），但有局部最小值問題
擬合程度	若過於簡單將導致欠擬合	容易過擬合，需要使用正則化來抑制
泛化能力	對於簡單數據集的泛化能力較強，但對於非線性數據或高噪聲數據表現較差	泛化能力較強，但可能存在過擬合風險，需要正則化或調參數

2. 決策樹（Decision Tree）

常被用於處理分類及迴歸的問題，而 Classification and Regression Tree（CART）是最常見的演算法之一，其他不同種類的決策樹模型如 ID3、C4.5 及 CHAID，則分別應用於不同的類別分類問題，主要體現在分枝好壞的評估方式上。當數據的輸出為連續型數值時，該模型被稱為迴歸樹，透過逐步展開樹的結構，並以葉節點的均值作為預測值。從根節點開始，根據樣本的

某一特徵進行測試，將數據分配至對應的子節點，每個子節點表示該特徵的一個取值。這一過程持續進行，直到到達葉節點，完成樹的構建，其運作方式如圖 8 所示。

決策樹通過對所有特徵及其對應值進行切分，找到最適合的分枝並向下拓展。樹的深度越深，決策規則越複雜，模型對數據的擬合程度越高。然而，若數據中存在雜訊，過深的樹結構將導致過擬合。因此單一迴歸樹往往難以充分應用於實際場景，因此可以利用如 Boosting 架構的集成學習方法，通過結合多棵迴歸樹來提升模型性能，進一步衍生出隨機森林和 XGBoost 等更穩定的模型，有效處理複雜數據及過擬合的風險。

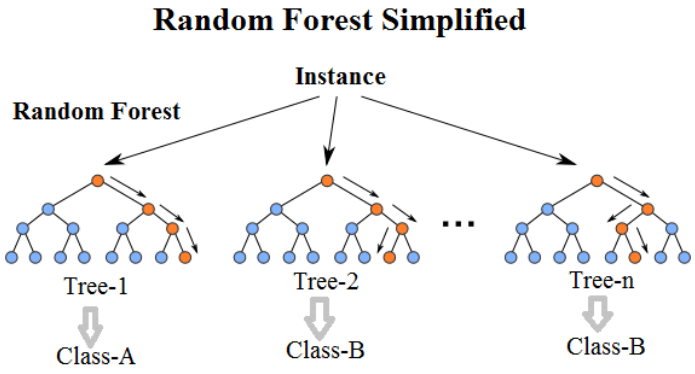


資料來源：Sunil Ray(2024)

圖 8 決策樹的運作模式

3. 隨機森林(Random Forest)

本研究採用迴歸隨機森林以提升模型的準確性與穩定性，並有效避免單一決策樹可能造成的過擬合問題。迴歸隨機森林特別適合處理非線性關係與高維度數據，捕捉多種影響因素中旅行時間的複雜模式，亦可對噪聲等狀況具有良好的容錯性，顯著提升模型的泛化能力。在單一決策樹中，若樹的最大深度設置過大，模型容易過擬合，隨機森林的結構特性使其易於擴展到大規模數據的處理，能在兼顧準確性與穩定性的基礎上應用於更多情境，而架構如圖 9 所示。

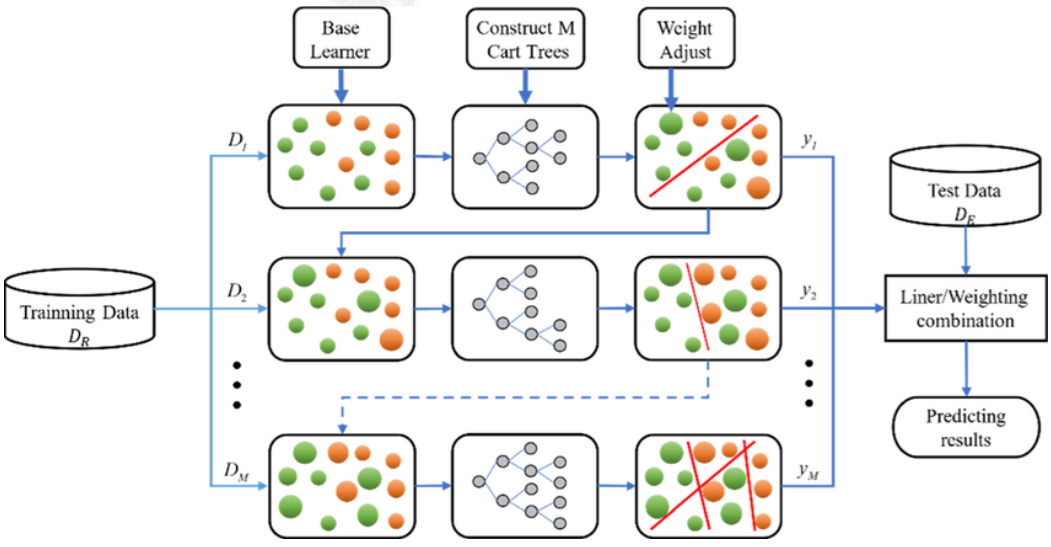


資料來源：資工心理人的理財筆記

圖 9 隨機森林的運作模式

4. 梯度提升樹（Gradient Boosting Trees）

為一種序列化的集成學習方法，在每一步迭代中，針對前一步模型的預測殘差訓練新的弱學習器，並將其輸出以加權方式疊加至整體模型，透過梯度下降最小化損失函數，最終形成高精確度的強學習器。在迴歸問題中，最終預測值為所有弱學習器預測結果的加權和；在分類問題中，則可透過投票機制決定最終類別。由於每棵新樹僅需擬合目前殘差，梯度提升樹能夠有效修正前序模型的不足，並在控制學習率和樹的深度等超參數下，兼顧擬合能力與過擬合風險，此方法在結構化資料的表現穩定，且對於異常值和非線性關係具備良好適應性，其架構如圖 10 所示。



資料來源：Z. C. He(2023)

圖 10 梯度提升樹的運作模式

二、房價多變數迴歸分析

3.1 透過既有的資料進行相關係數矩陣、直方圖等分析，本章節將針對過去 5 年共計 10993 筆歷史數據進行預測，研究將透過機器學習模型之隨機森林與線性迴歸演算法，將屋齡、總面積、經緯度坐標、樓層數及車位狀況等關鍵特徵納入訓練。分析面向包括估算房屋總價與單價、結合時間序列分析，根據歷史成交趨勢推估未來一至兩年的房價走勢等，以下將針對分析結果進一步說明。

下圖 8 為根據現有數據資料對未來高雄市左營區之房價進行預測，X 軸為該筆交易發生之時間，時間範圍自 2021 年至預測 2026 年房價，Y 軸左側為該筆交易單價範圍為 0 至 70 萬、右側長條圖為屋齡，其中圖中點之顏色亦說明該房屋屋齡，當點接近淺藍色或白色，對應右側長條圖屋齡大約接近 30 至 50 年，代表該房屋屋齡較老且此類點大致呈現於圖中下方；反之，當圖中點呈現深藍色，對應右側長條圖屋齡大約為新成屋至 10 年，代表該房屋較新，且這些點大致呈現於圖中上方。

圖中紅色實線、紫色實線與綠色虛線為模型計算出之預測路徑，因資料預測考慮屋齡及樓層，而將不同情境呈現出來。其中紅色實線為假設該房屋為全新屋即屋齡 0 年且位於高樓層，預測出結果大約落於紅線上，該坪單價大約 40 萬至 45 萬；綠色虛線為市場平均預測線，假設情境為平均屋齡及平均樓層，此為市場大眾行情，若買一屋況普通樓層適中之房屋，價格大約落於綠色虛線，單坪大約 30 萬左右；最後，子色實線則是假設房屋為 31 年舊房且位於 2 樓之低樓層，若預算有限僅能買老舊公寓，模型預測房價大約落於紫色實線上，單坪房價大約 20 萬左右。

根據預測結果R^2為 0.43，說明模型解釋力為 43%，因此預測僅加入時間、屋齡及樓層三個變數，對於整體若要預測房價尚有需多為考量到之處，如路段差異、幾房幾室幾廳等欄位，此外，生活機能指標如距離捷運站、高鐵站或商圈遠近與社區管理品質亦是影響房價的關鍵因素。目前模型尚未納入上述變數，導致約模型解釋力的偏低、價格變異無法被解釋，仍需依賴更多細部特徵數據才能精準預測。

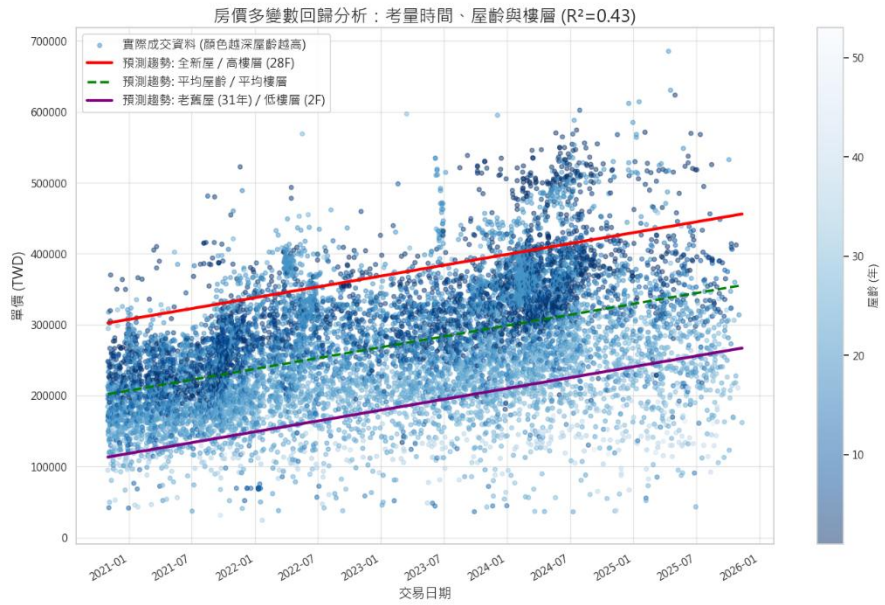


圖 11 房價多變數迴歸分析

三、 房價預測模型比較

下圖 12 為線性迴歸、決策樹、隨機森林及梯度提升樹四種模型之預測結果，圖中左上角為線性迴歸模型，R^2為 0.5327，說明模型只能解釋約一半的房價變異，另外 47%的變異無法被模型捕捉，而線性迴歸表現不佳之原因為線性假設與實際房價形成機制不符。線性模型假設每個特徵對房價的影響是獨立且恆定的，例如屋齡每增加一年，房價下降固定金額；面積每增加一坪，房價上升固定金額，然而實際情況遠比這複雜。屋齡對房價的影響並非線性，前 10 年因新屋溢價使得折舊較慢，10 至 20 年折舊較快，20 年以上因房價降價空間有限而趨緩；此外，地理位置與屋齡的交互作用也很重要，市中心的老屋仍可能比郊區新屋貴，這種交互效應線性模型無法捕捉，且線性模型對於離群值較為敏感，極端高價或低價的物件會影響係數估計。

右上角為決策樹之預測結果，相較於線性迴歸，決策樹之R^2提升至 0.7685，決策樹之預測結果之所以較線性迴歸佳，在於其分段常數的建模方式，決策樹將特徵空間分割成許多小區域，每個區域內給予一個固定預測值，此方式能夠逼近任意複雜之非線性函數，例如，若屋齡小於 5 年且緯度大於 22.65 之市中心且總面積大於 40 坪，則預測單價 40 萬/坪、若屋齡小於 5 年且緯度小於 22.60 之郊區且總面積大於 40 坪，則預測單價 28 萬/坪。這兩條規則顯示，同樣是新的大坪數物件，在不同地點的價格差異巨大，決策樹透過分支結構自動捕捉此類地點與屋齡的交互效應。

然而，決策樹 77%之解釋力雖然尚佳，但仍有提升空間。單一決策樹之問題在於過度擬合風險，為在訓練集上達到高準確度，樹可能建立過深、分割過細，學習到訓練資料中雜訊而非真正的規律，導致在測試集上表現可能欠佳。此外，

決策樹對訓練資料之微小變化較敏感，若訓練資料稍有不同，可能產生完全不同之樹結構，此不穩定性限制其實用性。

左下角為隨機森林之預測結果，其表現為四種模型中最佳，R^2提升至 0.8409，相較於單一之決策樹容易受到訓練資料微小變化影響，隨機森林的預測更加穩定可靠。隨機森林之所以能達到最佳效能，在於其集成學習的優勢，透過建立 100 棵決策樹並平均其預測，隨機森林有效降低單一樹的變異性。每棵樹可能在某些樣本上預測過高或過低，但平均後誤差相互抵消。同時，由於每棵樹皆基於相同的資料分布，所有樹皆學到資料中的規律、系統性之正確模式被保留下來。此外，隨機特徵選擇使得每棵樹關注不同的特徵組合，增加樹的多樣性，進一步提升集成效果。

最後右下角為梯度提升樹之預測結果，其模型預測表現與隨機森林接近，R^2為 0.8369，在解釋力上，梯度提升樹僅略低於隨機森林 0.4 個百分點，在實務應用中這個差距幾乎可以忽略，而梯度提升樹略遜於隨機森林的可能原因為：一，本研究設定的樹深度最大深度 5 層，可能因此限制其表現，若增加深度可能提升效能，但也增加過度擬合風險；二，梯度提升樹對參數較為敏感，學習率、樹數量、深度等參數都需要仔細調整，研究使用的參數可能未達最優。

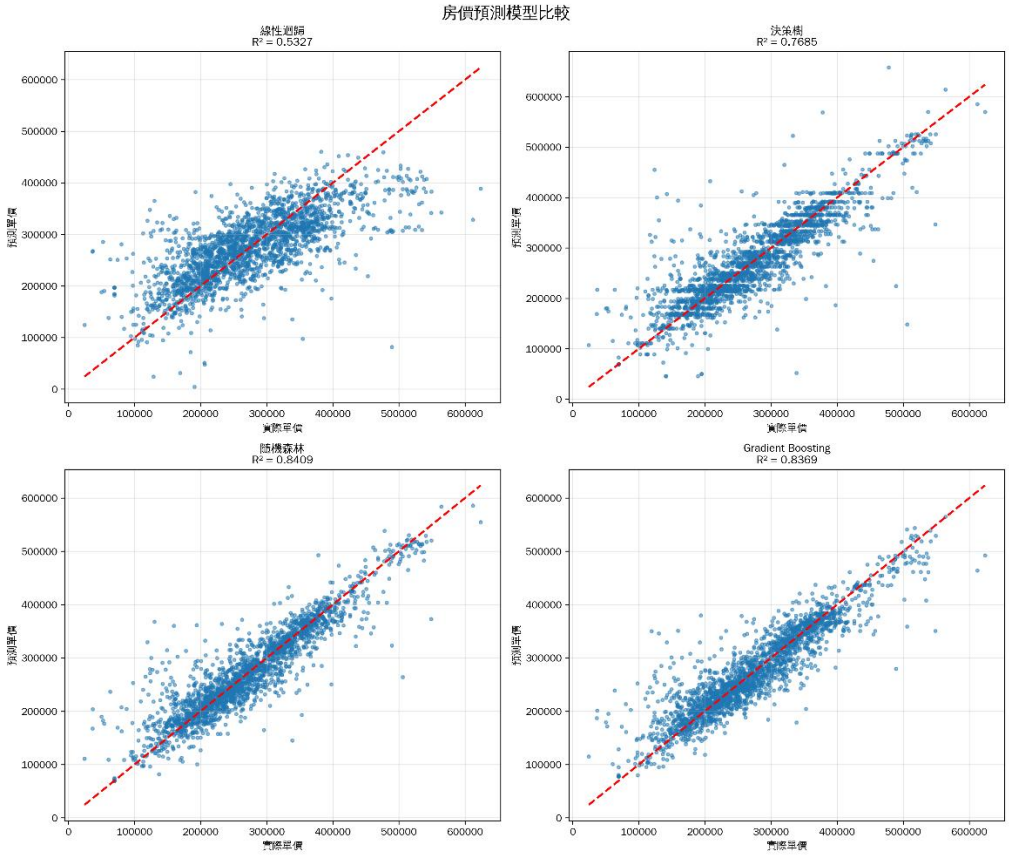


圖 12 房價預測模型比較

四、模型性能比較

為更直觀地呈現各模型之評估結果，以下將分別展示四個模型在 RMSE、MAE 與 R^2三個指標上的表現。三種指標從不同角度評估模型的預測能力，RMSE 為均方根誤差，因為對大誤差給予較重的懲罰，適合用於評估極端預測錯誤的風險；MAE 為平均絕對誤差，其為所有預測誤差的簡單平均，更容易理解與溝通；R^2為決定係數，反映模型的整體解釋能力，為評估模型優劣最常用之綜合指標。

下圖 13 之左圖為 RMSE 之評估結果，其用於衡量預測值與實際值之間的平均偏差程度，以下為計算公式，單位為「元/坪」：

$$RMSE = \sqrt{[\sum \frac{(實際值 - 預測值)^2}{樣本數}]}$$

RMSE 越小代表模型預測越準確，如當 RMSE =35,778 元/坪，說明著模型預測值平均偏離實際值約 3.6 萬元/坪，以一間 30 坪的房屋計算，相當於總價誤差約 107 萬元。其中，線性迴歸之結果為 61,314.19 元/坪、決策樹為 43,158.46 元/坪、隨機森林為 35,778.01 元/坪，梯度提升樹為 36,229.66 元/坪，線性迴歸之結果明顯高於其他三個模型，幾乎是隨機森林的 1.7 倍，顯示其預測誤差遠大於非線性模型。

下圖 13 之中圖為 MAE 之評估結果，其同樣衡量預測誤差，但採用絕對值而非平方，以下為計算公式，單位為「元/坪」：

$$\text{MAE} = \sum \frac{|實際值 - 預測值|}{樣本數}$$

MAE 更直觀易懂，其直接表示平均每筆預測偏離實際值多少。MAE =23,733 元/坪，表示平均誤差約 2.4 萬元/坪，對於一間 30 坪的房屋，相當於總價誤差約 71 萬元。其中，線性迴歸之結果為 45,692.80 元/坪、決策樹為 28,073.39 元/坪、隨機森林為 23,733.20 元/坪，梯度提升樹為 24,891.96 元/坪，MAE 之排名與 RMSE 完全一致，但相對差距略小。線性迴歸之 MAE 約是隨機森林的 1.9 倍，差距仍然顯著但不如 RMSE 的 1.7 倍大，顯示線性迴歸不僅平均誤差大，且存在一些極端之預測失誤。

下圖 13 之右圖為R^2之評估結果，其為一無單位之比率，取值範圍通常在 0 到 1 之間，以下為計算公式：

$$R^2 = 1 - (\frac{殘差平方和}{總平方和})$$

R^2表示模型能夠解釋多少百分比的資料變異，如R^2 =0.8409 表示模型能解釋 84.09%的房價變異，剩餘 15.91%是模型無法解釋的隨機因素或未納入的變數，當R^2越接近 1 表示模型解釋性越強，一般而言，當R^2大於 0.7 則被認為為不錯之模型結果。其中，線性迴歸之結果為 0.5327、決策樹為 0.7685、隨機森林為 0.8409，梯度提升樹為 0.8369。

由上述三種評估指標之結果，線性迴歸於 RMSE 與 MAE 之誤差皆最大，R^2之解釋力最弱，主要歸咎於實際之房價市場並非簡單線性關係且各因素之間會互相影響；隨機森林在三種指標表現皆最佳，其原因在於隨機森林透過 Bootstrap 抽樣，即每棵樹使用不同的訓練子集與隨機特徵選擇，即每次分割只考慮部分特徵，此差異性確保樹不會全部犯相同的錯誤，使得平均效果更好。

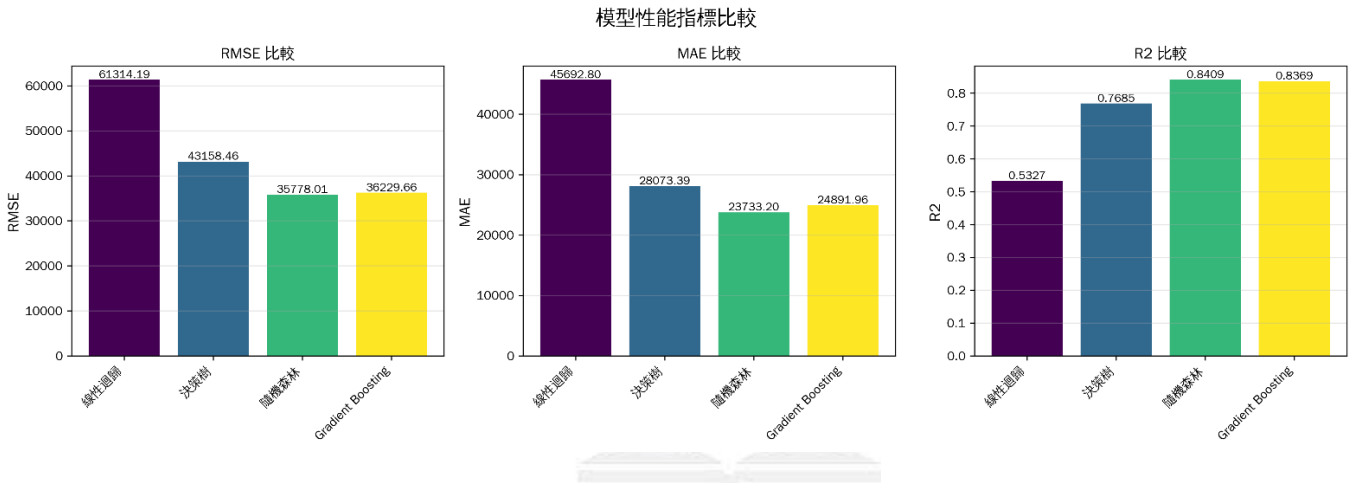


圖 13 模型性能指標比較

五、 影響房價關鍵因素

為解哪些因素是影響房地產價格之核心問題，下圖 14 利用隨機森林模型的特徵重要性分析功能，計算資料拆分前 17 個特徵對房價預測之影響程度，特徵重要性的計算原理是評估每個特徵在所有決策樹中被用於分割時帶來的預測改善程度，數值越高表示該特徵對預測結果的影響越大。

分析結果顯示，屋齡為影響房價最關鍵之因素，重要性高達 35.85%，遠超過其他所有特徵，說明新屋與舊屋價格差異極為顯著，每增加一年屋齡，房價平均下降約 1-1.5%，若以投資角度而言，意味房屋折舊效應明顯；對於購屋者，若預算有限，選擇 15-20 年的中古屋可以較低成本取得良好的地點與生活機能，但需評估後續的維護成本；對於建商，此結果強調新建案的時間價值，應盡快銷售以獲取最大利潤。

交易年份排名第二，重要性為 25.11%，反映房地產市場受總體經濟環境、貨幣政策與供需變化的顯著影響。房價隨時間波動，2020 至 2025 年間整體呈現上漲趨勢，但也存在短期的回檔與調整。地理位置由緯度與經度兩個特徵表達，兩者的重要性分別為 10.36%與 6.23%，合計 16.59%，排名第三與第五，此結果驗證房地產市場「地點、地點、地點」定律，在高雄地區，市中心與郊區的價格落差可達一倍以上，好的地點不僅帶來生活便利性，也是房價長期保值的關鍵。

總面積排名第四，重要性為 7.85%，較大的面積通常伴隨較高的單價，其原因可能為大坪數物件多位於較優質的社區，或是市場對於空間的溢價效應。其他特徵如房數、廳數、衛浴數、主建物占比等，重要性皆在 2-5%之間，屬於次要影響因素。

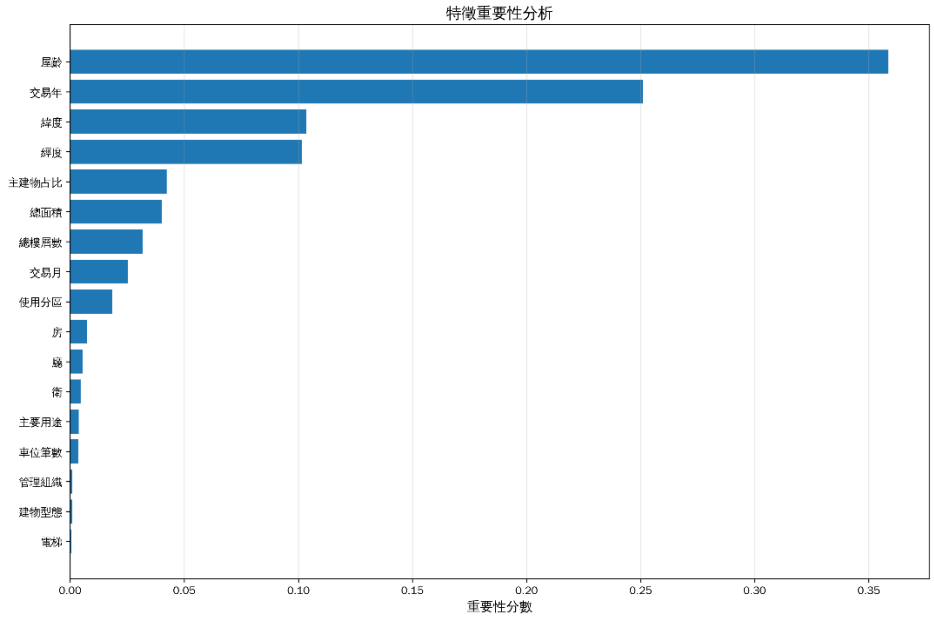


圖 14 影響房價關鍵因素

六、 左營區價格區分圖

下圖 15 為房地產交易的地理空間分布，透過顏色編碼呈現不同價格區間物件分布特徵，左營區涵蓋經度 120.28 至 120.32 度、緯度 22.66 至 22.71 度的範圍，單價自 24,291 元/坪至 685,478 元/坪，呈現高度的價格異質性。圖中將交易資料依單價分為六個區間並以不同顏色標示，從低價之深紅色至高價之深綠色，直觀呈現左營區房價的空間分層現象。

從整體分布可發現，左營區房價呈現明顯同心圓模式，以新左營高鐵站為核心向外擴散形成價格梯度，圖中深綠色點，其單價超過 40 萬元/坪，主要集中在核心區域，包括蓮池潭東側的保靖街、環潭路一帶，以及高鐵站前的重信路、重立路等精華地段。此類高價區共有 723 筆交易，占總數 7.1%，平均屋齡僅 11.1 年，代表近十年新建的優質大樓，享有絕佳的景觀視野或交通便利性，其中，最高價物件達 68.5 萬元/坪，位於蓮池潭第一排的景觀豪宅，充分展現稀缺景觀資源的溢價效應。

淺綠色與黃色點為單價 30 至 40 萬元/坪，形成第二圈，涵蓋崇實路、文川路、大中二路精華路段等左營核心市區，此價格帶共有 3,112 筆交易，占總數 28.3%，為左營區重要的中高價市場。其中，崇實路以 265 筆高價交易領先，作為貫穿左營核心區的南北向主要幹道，沿線生活機能完善、交通便利；文川路則以學區優勢見長，周邊有明德國中、左營高中等優質學校，吸引重視子女教育的家庭進駐，此價格帶的物件平均屋齡約 14 年，面積約 40 坪，為中高收入家庭的主流選擇。

橘色點為單價 20 至 30 萬元/坪，其分布更為廣泛，涵蓋北左營的桃子園路、大中二路外圍路段、文自路邊緣區域等，此價格帶共有 4,117 筆交易，占總數 37.5%，是左營區最大的市場區段。其中，桃子園路以 545 筆交易量居所有路段之冠，反映北左營與農 16 重劃區的市場活躍度。此區域距離市中心稍遠，但因重劃區持續開發、基礎建設逐步完善，吸引許多首購族與年輕家庭進駐。

深紅色點單價低於 20 萬元/坪，散布在左營區最外圍，與楠梓區、仁武區交界的邊緣地帶，以及老舊社區，此價格帶有 2,171 筆交易，占總數 19.7%，平均屋齡高達 24.7 年，多為 25 年以上的老舊公寓或華廈，此類物件雖然單價低廉，但因交通不便、屋況老舊、生活機能簡陋，主要吸引預算極度有限的買家或等待都更的投資客。

整體而言，左營區房價的空間分布清楚反映地理位置、交通便利性、生活機能與建物品質等因素的綜合影響，以高鐵站為中心，房價隨距離增加而遞減，形成由核心到邊陲的價格梯度。

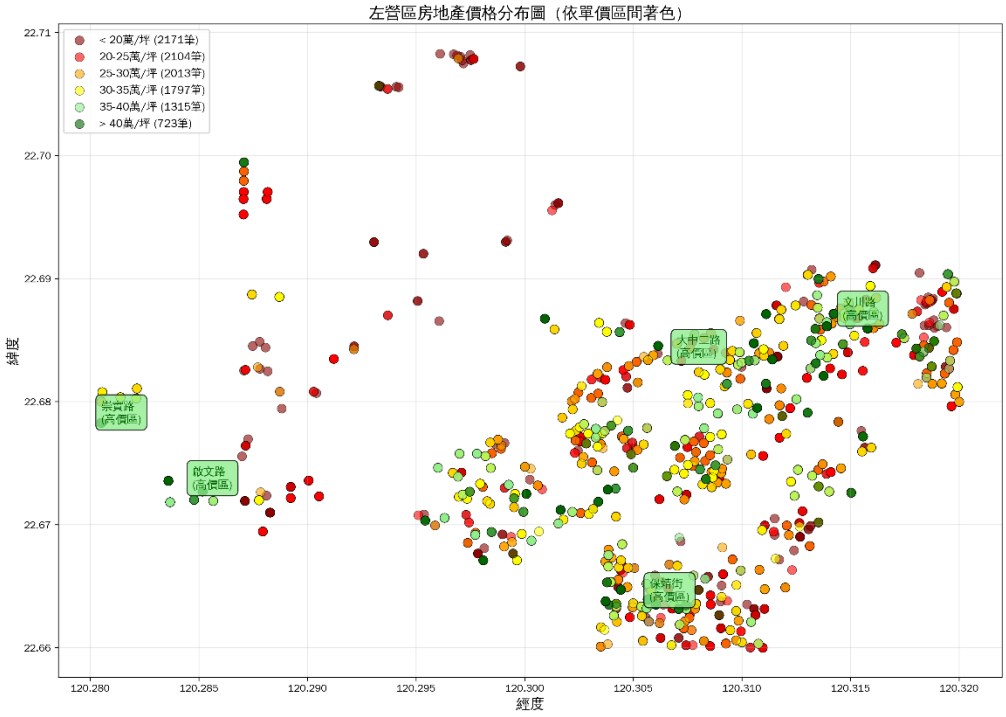


圖 15 左營區房地產價格分布圖

第五章 結論與建議

5.1 結論

本研究透過 Python 爬蟲技術取得高雄市左營區民國 109 年 11 月至 114 年 11 月共五年期間 10,993 筆房地產交易資料，運用敘述性統計、地理空間分析與機器學習模型等方法，系統性探討左營區房地產市場的價格結構、影響因素與發展趨勢，並透過線性迴歸、決策樹、隨機森林與梯度提升樹四種模型的比較分析，深入理解房價形成機制並識別關鍵影響因素。

從市場結構來看，左營區房地產市場呈現典型右偏分布，平均單價 271,017 元/坪，中位數 264,374 元/坪，主流價格帶集中在 20 萬至 35 萬元/坪，占整體交易約六成至七成。在模型預測效能方面，隨機森林模型表現最佳，R^2達 0.8409，RMSE 為 35,778 元/坪，MAE 為 23,733 元/坪，能夠解釋 84%的房價變異。相較之下，線性迴歸僅達 53%解釋力，決策樹為 77%，梯度提升樹為 84%，非線性模型相較於線性模型的效能有效提升，說明房價為非線性特徵，透過隨機森林、梯度提升樹此類集成學習方法，相較於單一決策樹的 R^2提升約 9 至 10%，顯示透過結合多個模型能有效降低預測誤差。

關鍵影響因素分析顯示，屋齡為影響房價最關鍵因素，重要性高達 35.85%，遠超過其他所有特徵，量化驗證「新屋溢價、舊屋折舊」的市場現象。交易年份排名第二，重要性為 25.11%，反映房地產市場受總體經濟環境、貨幣政策與供需變化的顯著影響。地理位置由緯度與經度表達，合計重要性 16.59%，驗證「地點、地點、地點」的房地產定律。

空間分布分析揭示左營區房價呈現明顯的同心圓模式，以新左營高鐵站為核心向外擴散形成價格梯度。核心圈範圍為 1.5 公里內單價多在 35 萬元/坪以上，包括蓮池潭景觀帶、高鐵站前核心區、博愛商圈精華地段，最高價達 68.5 萬元/坪。第二圈範圍約 1.5 至 3 公里，單價約 25 至 35 萬元/坪，涵蓋崇實路、文川路、大中二路精華路段等左營核心市區。外圍圈 3 公里以上，單價多在 25 萬元/坪以下，包括北左營、邊緣地帶與老舊社區。

整體而言，本研究透過大數據分析與機器學習技術，全面解析左營區房地產市場的運作邏輯，識別影響房價的關鍵因素，建立準確的預測模型。

5.2 建議

1. 整合多元外部資料源以擴充特徵維度，本研究主要依賴實價登錄資料，雖然涵蓋建物基本屬性，但仍缺乏許多重要的影響因素，後續研究可串接政府開放資料平台取得環境品質指標，包括各區域的空氣品質指數、噪音分貝數、犯罪率統計、公園綠地面積等，此類環境因素雖然在傳統分析中較少被量化，但對居住品質與房價有顯著影響。
2. 納入社群媒體與網路評價資料進行情感分析，可爬取 PTT、Dcard 等社群平台上關於左營區各社區的討論文章，透過自然語言處理與情感分析技術，量化社區的網路評價分數。
3. 本研究聚焦於微觀的建物特徵，較少討論總體經濟環境，後續可建立迴歸模型探討供需關係如何影響價格，更全面理解房價波動的深層機制。

參考文獻

[1]Chen, Y.-J., & Hsu, C.-K. (2020). **Comparison of Housing Price Elasticities Resulting from Different Types of Multimodal Rail Stations in Kaohsiung, Taiwan**. *International Real Estate Review*, 23(3), 417–432.

[2]Owusu-Manu, D.G., Edwards, D.J., Donkor-Hyiaman, K.A., Asiedu, R.O., Hosseini, M.R., & Obiri-Yeboah, E. (2019).**Housing attributes and relative house prices in Ghana**.*International Journal of Building Pathology and Adaptation*, 37(5), 733–746.

[3]邱司杰（2014），基於實價登錄的房價模型研究，國立交通大學網路工程研究所碩士論文。

[4]邱恩華（2017），探討影響屏東市房價之因素，國立屏東大學國際貿易學系碩士論文。

[5]邱國祥（2020），以多元線性迴歸與機器學習模型預估不動產價格-以台中市實價登錄為例，國立中興大學應用數學系所碩士論文。

[6]許雅晶（2022），房價預測模型-以新北市林口區為例，實踐大學資訊科技與管理系碩士論文。

[7]陳玟寧（2022），迴歸機器學習應用於房價預測—以台北市實價登錄為例，明志科技大學工業工程與管理系碩士論文。

[8]林于軒（2025），應用多維特徵與機器學習的房價預測模型，中原大學資訊管理學系碩士論文。