台北市房價影響因素分析

Business Analytics (113-1)
Final Report

Group 19

組員:
工管系 B10701203 呂欣融
會計所 R12722031 杜奕寧
財金所 R12723006 林姝延
商研所 R13741058 黃茂殷
經濟系 B09607059 蔡宜芳

專案摘要

我們的報告使用台北市房地產市場交易資料，探討影響房屋交易價格的主要因素，透過 EDA、回歸建模和優化，分析每坪單價與總價的決定因素。我們進行了數據收集、資料清理、變數分析，並建立高解釋力的回歸模型，分析並解釋各變數對房價的影響程度。

在資料清理過程中，我們排除空值、異常值（如親友交易），對部分變數做格式轉換以及前處理，包含建立「豪宅」變數、公設比計算、把建物面積和樓層等變數級距化。同時針對右偏分布的連續變數進行對數轉換，處理數據分布不均的問題。

透過EDA，我們觀察到以下結果：

- 區域特性：大安區、中正區及松山區的房價顯著高於其他區域，這些地區的房價可能受明星學區及豪宅因素影響較大。
- 建築特徵：
    - 有電梯、管理組織的建物價格比較高。
    - 住宅大樓與華廈價格高於透天厝與公寓。
    - 商業用途建物價格高於工業與農業用途。
- 變數觀察：建物面積和房價呈正相關，公設比和屋齡之間存在負向相關性。

根據我們的最終模型結果，得到以下觀察：

- 房價主要影響因素：建物移轉面積、屋齡、是否為豪宅、公設比、電梯及樓層數為顯著影響房價的變數。
- 模型解釋力：總價模型的 R-squared 達 0.8164，解釋力高於單價模型，我們推測變數對總價的影響更直接。
- 共線性問題：電梯與建物型態之間的共線性經過優化後顯著降低。

而根據我們的模型檢測結果，殘差圖呈現均勻分布，但尾端偏離，我們推測可能是因為沒有納入到某些外部變數。另外 QQ plot 也顯示部分殘差沒有完全服從常態分布，我們也推測可能是沒有考慮到政策、交通便利性等外部因素的影響。

我們的研究結論為：
房價受到多重因素影響，包含建物面積、屋齡與是不是豪宅等，而未來的研究建議可以納入外部政策或環境因子來提升模型的預測準確度。

一、研究動機

我們觀察到台北市房價高且波動幅度大，也了解到台北市房地產市場長期以來受到許多因素影響，包含地理位置、建物特徵、設備設施、交易時間及市場需求等，因此希望透過研究分析來了解並掌握房價波動因素。我們的研究動機包括：

1. 針對不同區域、建物特徵與交易資料進行分析，找出影響房價的主要因素。
2. 透過分析結果幫助政府制定更合理的房屋政策，解決市場不均衡問題。
3. 提供買賣雙方和投資者更有參考價值的房屋市場分析，幫助做出理性決策。

二、研究目標

1. 透過資料清理 EDA，檢視變數之間的相關性和特徵。
2. 建立回歸模型，分別針對「每坪單價」和「總價」進行解釋與預測。
3. 解決共線性問題，優化模型，提高模型的解釋能力和準確性。
4. 提出研究結果和限制，並探討未來可納入的外部變數。

透過研究，我們希望能提供一個全面且清楚的房價分析框架，為房地產市場相關利害關係人提供參考依據。

三、資料說明

我們的研究使用內政部不動產成交案件實際資訊資料供應系統所提供的公開資料，資料來源為 網站連結，選取 113年第3季臺北市 的房屋交易數據進行分析。資料涵蓋台北市各行政區的房屋交易數據，包含多個變數：

1. 應變數：
    - 房屋每坪單價（單價元平方公尺）：房屋每單位面積的交易價格。
    - 房屋總價（總價元）：房屋的總交易金額。
2. 自變數：
   包含房價影響的重要變數，包括：
    - 地理與交易資訊：鄉鎮市區、交易標的、交易年月日
    - 土地與建物資訊：土地移轉總面積平方公尺、建物移轉總面積平方公尺、建築完成年月
    - 建物特徵：建物型態、主要用途、移轉層次、總樓層數
    - 內部格局：建物現況格局-房、建物現況格局-廳、建物現況格局-衛
    - 設備設施：車位類別、有無管理組織、電梯
    - 價格資訊：單價元平方公尺、總價元

四、資料處理

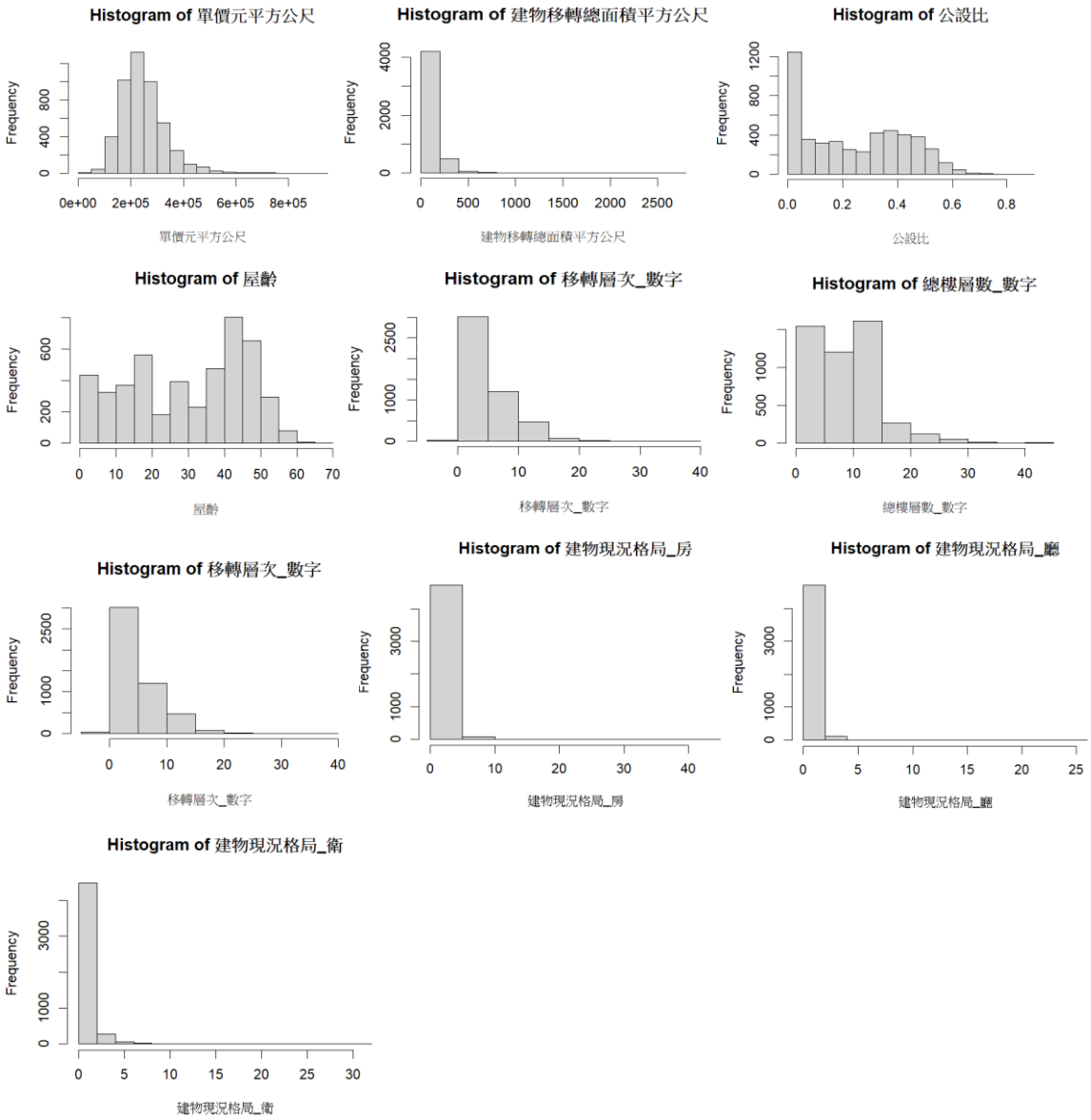
1. 新增欄位：
   - 豪宅：
     - 台北市政府（豪宅房屋稅）：
       - 高級住宅之認定為，房屋為鋼筋混凝土以上構造等級，用途為住宅，經按戶認定房地總價在新臺幣8,000萬元以上，且在90年7月1日以後建築完成者。
       - 以原資料中的主要建材、主要用途、總價元、建築完成年欄位篩選
     - 央行(豪宅限貸令)跟財政部(豪宅交易稅)還有不同定義，但感覺市政府篩選條件更精細，因此採用此標準。
       - 參考資料：台北市稅捐稽徵處
   - 公設比：
     - 共有 / 總建物面積 *100% = (總建物 − 主建物 − 附屬建物 − 陽台) / 總建物面積 *100%
       - 以原資料中的總建物面積、主建物面積、附屬建物面積、陽台面積欄位進行計算
     - 商辦/住商大樓公設比本來就較高(公廁或茶水間)。
     - 住家高公設比：可能有泳池或健身房等設施，或樓層較高(8層樓以上要有兩座逃生梯)。同總建物面積下，私人可使用空間會較少。
       - 參考資料：信義房屋
   - 屋齡：
     - 先將交易年月日轉換成日期型態，再以(交易年月日-建築完成年月日)計算。
     - 剔除負數(預售屋)。
2. 資料型別轉換：
   - 建物移轉總面積以級距呈現：
     - 小：0~20 坪 (66.1157平方公尺)
     - 中：21~50 坪 (165.2893平方公尺)
     - 大：50 坪以上
   - 移轉層次、總樓層數：
     - 先剔除雜亂的文字(平台、夾層、騎樓)，再提取樓層的中文數字，並從string轉為 int (若一次買賣多層，則取最低樓層)。包含地下層及一樓者填入 0，整棟交易者取總樓層平均數填入 2.9615。
     - 將移轉層次再轉為級距呈現：
       - <0：地下層
       - 0：一層與地下層
       - 1~5：一到五層
       - 6~10：六到十層
       - 11~15：十一到十五層
       - 16~20：十六到二十層
       - >20：二十層以上

        - *2.9615：全
- 有無管理組織、電梯：
    - 將「有」、「無」改為1, 0表示。

3. 排除可能的離群值：
    - 從原資料備註欄為中篩選掉“親友”的買賣，以更好的反映市場交易價。
4. 刪除空值

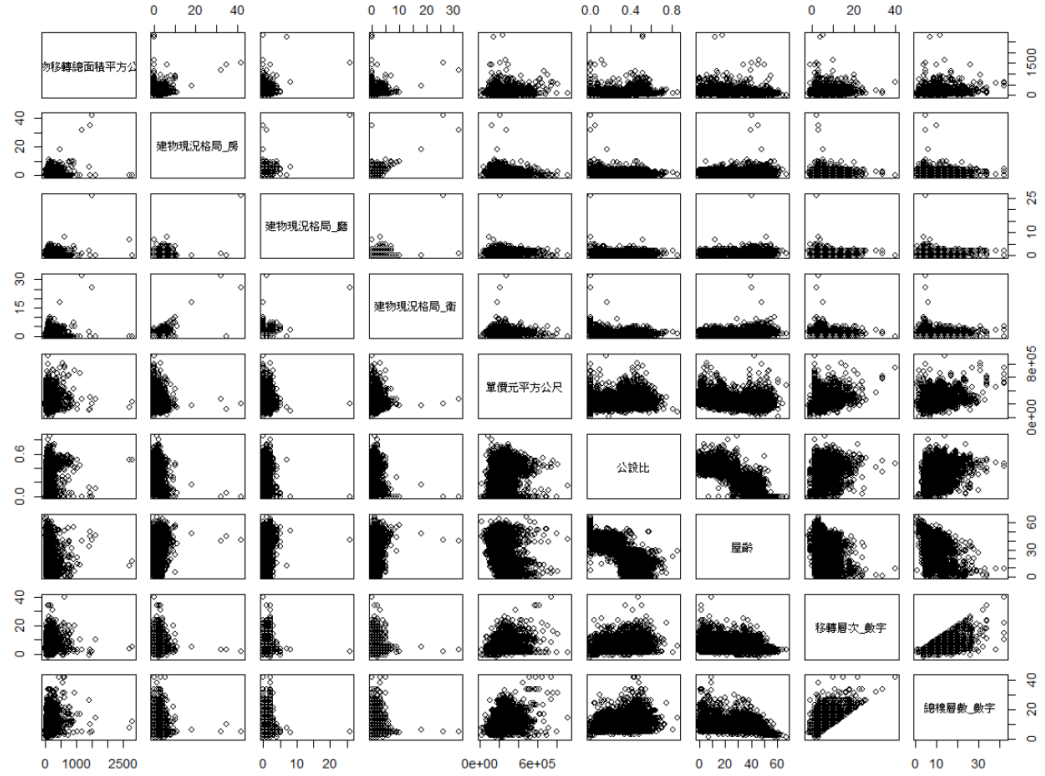
五、EDA

首先，針對連續變數繪製直方圖，看是否有需要做 transform



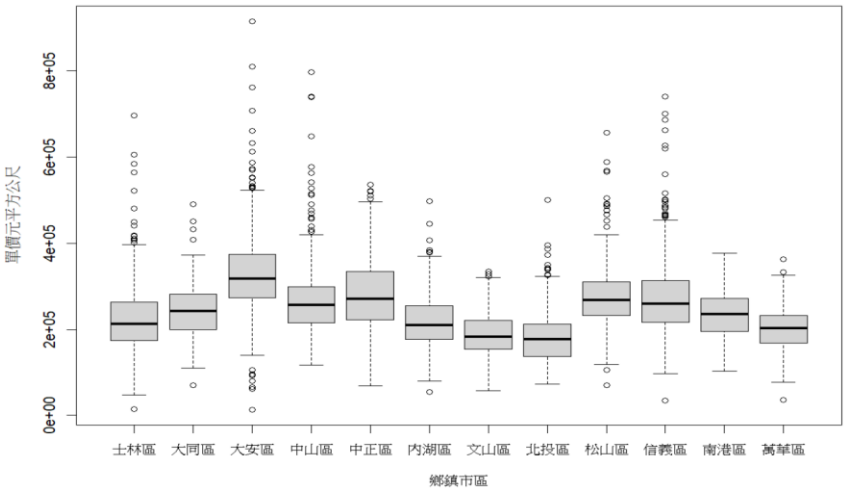
結果顯示，除了屋齡之外，大部分連續變數都呈現右偏，因此建模時可考慮對這些右偏變數做 log 轉換。

接著，使用散佈圖呈現各變數之間的相互關係。



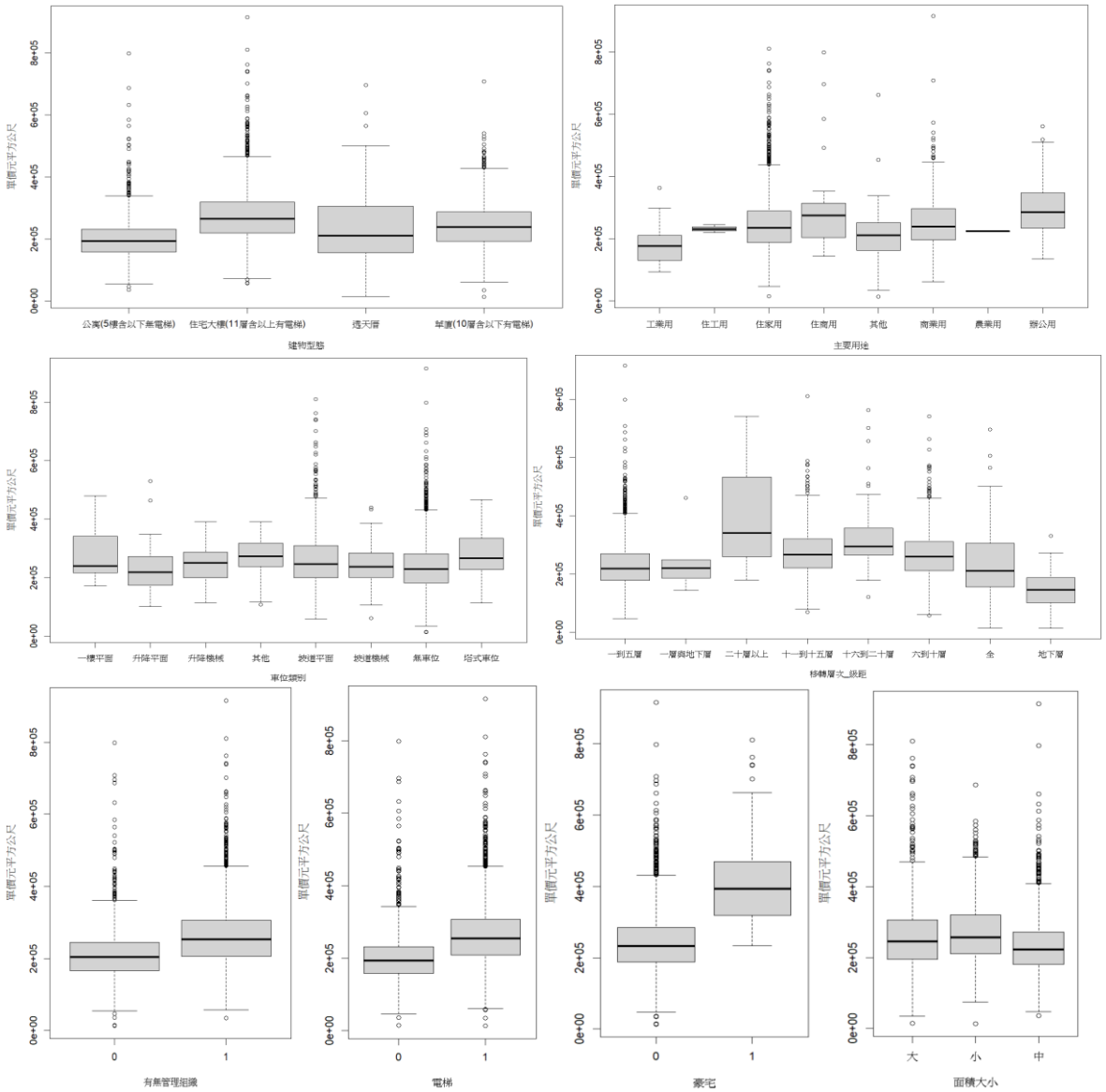
觀察結果顯示，單價元平方公尺沒有與任何一個連續變數呈現明顯的趨勢，但可以看到像是公設比與屋齡之間存在明顯負向相關性，建模時可以考慮兩者的共線性關係。

再來針對類別變數繪製箱型圖。首先觀察各**鄉鎮市區**在**單價元平方公尺**上的分佈特徵。



從圖中可看到不同鄉鎮市區的中位數有明顯的差異。大安區、中正區、松山區的中位數明顯高於其他區域，顯示這些區域為高房價區域，與本組在期末初報告預測這三個區域可能因為位於明星國中所在區域而房價較高的結果相同。此外，士林區、大安區、信義區的異常值（圈點）顯著多且分布在高價區域，可能是因為這些區域內存在豪宅；萬華區、文山區等低價區域的異常值則較少，並且異常值的價格也不算極端。

下圖為其他類別變數在**單價元平方公尺**上的分佈特徵：



值得注意的是，有管理組織、有電梯的建築房價較高；觀察**建物型態**可以看見住宅大樓與華廈的房價較透天厝與公寓高；觀察**主要用途**可以看見有商業用途的建築價位也會較高。

六、建立模型

模型採用兩個不同的應變數進行建模，分別交易的每坪單價與總價，並都採用相同的自變數開始建立模型，後續會測試增加與減少變數是否有效提高模型的解釋能力

第一部分：單價元平方公尺
第一個模型lm1包含了資料集中所有資料內涵不重複的變數，剩下級距類型的資料會在後續評估是否比起原始數值更能提高模型的解釋能力

```
#模型1，考量資料內容不重複的變數
lm1=lm(單價元平方公尺~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +主要用途+車位類別+電梯+公設比+屋齡+移轉層次_數字+總樓層數_數字
       +豪宅+鄉鎮市區+有無管理組織,data=data)
summary(lm1) #Multiple R-squared:  0.5077,  Adjusted R-squared:  0.5037
```

```
Residual standard error: 60640 on 4775 degrees of freedom
Multiple R-squared:  0.5077,    Adjusted R-squared:  0.5037
F-statistic: 126.3 on 39 and 4775 DF,  p-value: < 2.2e-16
```

此模型的 R-squared 為 0.5077，幾乎所有變數都有顯著影響，同時為了後續解讀容易，將鄉鎮市區變數的基礎改為係數最小的北投區。車位類別雖然係數上改為"其他"較好，但解讀上還是以"無車位"比較適合。而主要用途係數上來看雖然應改為"農業用"，但根據 boxplot 的呈現覺得並不適合，因此還是維持原本的"工業用"。

第二個模型 lm2 考量到房價或許最初會隨房屋年齡而降低，但後續會逐漸回升，如同上課的例題，因此加入屋齡的平方作為新變數。

```
#加上屋齡的平方
lm2=lm(單價元平方公尺~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +主要用途+車位類別+電梯+公設比+屋齡+屋齡2+移轉層次_數字+總樓層數_數字
       +豪宅+鄉鎮市區+有無管理組織,data=data)
summary(lm2) #Multiple R-squared:  0.5259,  Adjusted R-squared:  0.5219
anova(lm2,lm1) #顯著
```

```
車位類別塔式車位   -2780.372   6002.081  -0.463 0.643217
電梯              -80706.622  29181.396  -2.766 0.005702 **
公設比            -49099.999  11417.931  -4.300 1.74e-05 ***
屋齡               -5195.054    251.054 -20.693  < 2e-16 ***
屋齡2                 59.767      4.415  13.536  < 2e-16 ***
移轉層次_數字        341.408    293.041   1.165 0.244056
總樓層數_數字       2796.438    357.219   7.828 6.05e-15 ***
```

```
Residual standard error: 59510 on 4774 degrees of freedom
Multiple R-squared:  0.5259,    Adjusted R-squared:  0.5219
F-statistic: 132.4 on 40 and 4774 DF,  p-value: < 2.2e-16
```

結果顯示屋齡的平方的確也對每平方公尺的單價有顯著影響，能夠提高解釋能力，且執行 anova 測試後確認新模型與舊模型間有顯著差異，因此後續分析會保留屋齡的平方。

接下來考量可能的交乘項

1. 鄉鎮市區與主要用途：我們預期即使是同樣用途的建物，在不同的地區也可能會對價格有不同的影響程度，因此考量了兩者間的交乘項。

```
##地區與用途:同樣的用途在不同的地區也可能會有不同的影響程度
lm3=lm(單價元平方公尺~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +車位類別+電梯+公設比+屋齡+屋齡2+移轉層次_數字+總樓層數_數字
       +豪宅+主要用途*鄉鎮市區+有無管理組織,data=data)
summary(lm3) #Multiple R-squared:  0.5402,  Adjusted R-squared:  0.5316
             #加入交乘項後共線性太高或類別變數太多，無法執行vif。不考慮。
anova(lm3,lm2) #顯著
```

加入了兩者的交乘項後，雖然模型的解釋能力有提高，且也和前一個模型有顯著差異，不過於確認共線性時卻發生了共線性太高或類別變數過多的問題，無法執行 VIF，因此後續不考量兩者間的交乘項。

2. 樓層與電梯：我們認為，樓層越高的住戶通常越需要電梯，因此也加入了電梯變數與移轉樓層、總樓層數的交互項，不過發現交互項的共線性太高了，並不是個良好的模型。後續分析中也不會考量本交乘項。

```
##樓層與電梯:樓層越高的住戶越需要電梯
lm3b=lm(單價元平方公尺~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
        +車位類別+公設比+屋齡+屋齡2+電梯*(移轉層次_數字+總樓層數_數字)
        +豪宅+主要用途+鄉鎮市區+有無管理組織,data=data)
summary(lm3b) #Multiple R-squared:  0.5486, Adjusted R-squared:  0.5446
              #加入交乘項後共線性太高或類別變數太多，無法執行vif。不考慮。
```

再來我們考慮了變數的轉換，將適合取 log 且分佈呈現右偏的變數作轉換，確認模型的解釋能力是否有變高。做了轉換的變數包含建物移轉總面積平方公尺、移轉層次_數字、總樓層數_數字。同時因為移轉的層次包含地下室，因此只選資料集中"移轉層次_數字" > 0 的樣本進行回歸，會少 34 個樣本，對於資料的影響不大。

```
#取log，考量變數轉換，將移轉層數限定>0會損失34個樣本
lm4=lm(單價元平方公尺~log(建物移轉總面積平方公尺)+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +車位類別+電梯+公設比+屋齡+屋齡2+log(移轉層次_數字)+log(總樓層數_數字)
       +豪宅+主要用途+鄉鎮市區+有無管理組織,data=data[data$移轉層次_數字>0,])
summary(lm4) #Multiple R-squared:  0.538, Adjusted R-squared:  0.5341
             #面積與移轉層次的顯著程度提高
```

```
Residual standard error: 58540 on 4740 degrees of freedom
Multiple R-squared:  0.538,     Adjusted R-squared:  0.5341
F-statistic:   138 on 40 and 4740 DF,  p-value: < 2.2e-16
```

經過變數轉換的模型解釋能力有稍微提高，且經過轉換的"移轉總面積平方公尺"與"移轉層次"的顯著程度提高。

接著我們考慮將面積與樓層變數轉為級距，觀察是否在某一個級距之上或之下，對於單價有明顯的影響，以及是否能提高解釋能力。最初的級距一樣包含建物移轉總面積平方公尺、移轉層次_數字、總樓層數_數字，不過檢查共線性時發現共線性太高無法執行，並在逐項測試後發現原因來自於"移轉層次_數字"。所以排除本變數的級距，以原始格式進行回歸。

```
#使用級距代表移轉層次與面積
lm5=lm(單價元平方公尺~面積大小+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +車位類別+電梯+公設比+屋齡+屋齡2+移轉層次_級距+log(總樓層數_數字)
       +豪宅+主要用途+鄉鎮市區+有無管理組織,data=data)
summary(lm5) #Multiple R-squared:  0.5355,  Adjusted R-squared:  0.531
             #有效果，但沒有比取LOG好，不過lm5共線性太高，而無法計算VIF。
             #分別替換執行後發現是"移轉層次_級距"的原因。
```

```
Residual standard error: 58940 on 4768 degrees of freedom
Multiple R-squared:  0.5355,    Adjusted R-squared:  0.531
F-statistic: 119.5 on 46 and 4768 DF,  p-value: < 2.2e-16
```

顯著程度相比於原始格式的確也有稍微提高，但解釋能力還是以對數轉換的模型更好。所以我們將進一步以 lm4 為基礎，逐一刪除顯著程度不高的變數。

```
#以lm4為基礎刪減變數
lm6 <- update(lm4, . ~ . - 有無管理組織)
summary(lm6) #Multiple R-squared:  0.5379,  Adjusted R-squared:  0.5341
anova(lm6,lm4) #不顯著，可剃除

lm7 <- update(lm6, . ~ . - 建物現況格局.廳)
summary(lm7) #Multiple R-squared:  0.5379,  Adjusted R-squared:  0.5342
anova(lm7,lm6) #不顯著，可剃除

lm8 <- update(lm7, . ~ . - 車位類別)
summary(lm8) #Multiple R-squared:  0.5379,  Adjusted R-squared:  0.5342
anova(lm8,lm7) #接近1%的顯著，不可剃除
```

我們針對 lm4 中三個顯著程度不高的變數，逐一依照顯著程度由低至高刪除，並比較刪除前後的模型是否有顯著差異。 根據分析的結果，我們發現有無管理組織與房內餐廳客廳的數量對於每平方公尺的價格並無顯著影響。另外雖然相比於相同條件下無車位的交易標的，每一種車位類型對於每平方公尺的價格的顯著程度幾乎都不高，不過整體來說還是有接近1%的顯著程度，因此不從模型中刪除。

最後模型為: 單價元平方公尺 ~ log(建物移轉總面積平方公尺) + 建物型態 + 建物現況格局.房 + 建物現況格局.衛 + 車位類別 + 電梯 + 公設比 + 屋齡 + 屋齡^2 + log(移轉層次_數字) + log(總樓層數_數字) + 豪宅 + 主要用途 + 鄉鎮市區。Adjusted R-square 為 0.5342。

第二部分：交易總價

最初的模型包含了資料集中所有資料內涵不重複的變數。

```
##以總價為y----
lm1.1=lm(總價元~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +主要用途+車位類別+電梯+公設比+屋齡+移轉層次_數字+總樓層數_數字
       +豪宅+鄉鎮市區+有無管理組織,data=data)
summary(lm1.1) #Multiple R-squared:  0.817, Adjusted R-squared:  0.8155
```

```
Residual standard error: 15030000 on 4775 degrees of freedom
Multiple R-squared:  0.817,     Adjusted R-squared:  0.8155
F-statistic: 546.6 on 39 and 4775 DF,  p-value: < 2.2e-16
```

相較於每平方公尺的價格，相同的變數更能夠解釋總價，再來一樣加入屋齡的平方，我們預期和先前的回歸式會有相同的結果。

```
#加上屋齡的平方
lm2.1=lm(總價元~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +主要用途+車位類別+電梯+公設比+屋齡+屋齡2+移轉層次_數字+總樓層數_數字
       +豪宅+鄉鎮市區+有無管理組織,data=data)
summary(lm2.1) #Multiple R-squared:  0.8182,  Adjusted R-squared:  0.8167
anova(lm2.1,lm1.1) #顯著
```

電梯	-14445922	7348455	-1.966	0.04937 *
公設比	-13142013	2875262	-4.571	4.98e-06 ***
屋齡	-428059	63220	-6.771	1.43e-11 ***
屋齡2	6210	1112	5.585	2.46e-08 ***
移轉層次_數字	163646	73794	2.218	0.02663 *
總樓層數_數字	1152255	89955	12.809	< 2e-16 ***
豪宅	49764733	1668831	29.820	< 2e-16 ***

```
Residual standard error: 14990000 on 4774 degrees of freedom
Multiple R-squared:  0.8182,    Adjusted R-squared:  0.8167
F-statistic: 537.1 on 40 and 4774 DF,  p-value: < 2.2e-16
```

加入屋齡平方後的模型解釋能有提高，且新舊模型間的差異為顯著。屋齡的平方也如同預期的為正數，因此我們後續會接續以本模型為基礎增減變數。

我們對於總價的模型也考慮了一樣的交乘項，不過不管哪一個的結果都和先前一樣。雖然交乘項本身有顯著性，也提高了模型的解釋能力。但也有共線性太高或是類別變數過多的問題，因此加入了交乘項的兩個模型也不在我們的考慮範圍內。

```
#考量交乘項
##地區與用途:同樣的用途在不同的地區也可能會有不同的影響程度
lm3.1=lm(總價元~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +車位類別+電梯+公設比+屋齡+屋齡2+移轉層次_數字+總樓層數_數字
       +豪宅+主要用途*鄉鎮市區+有無管理組織,data=data)
summary(lm3.1) #Multiple R-squared:  0.8228,  Adjusted R-squared:  0.8195
               #加入交乘項後共線性太高或類別變數太多，無法執行vif。不考慮。
anova(lm3.1,lm2.1) #顯著

##樓層與電梯:樓層越高的住戶越需要電梯，
lm3.1b=lm(總價元~建物移轉總面積平方公尺+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
        +車位類別+公設比+屋齡+屋齡2+電梯*(移轉層次_數字+總樓層數_數字)
        +豪宅+主要用途+鄉鎮市區+有無管理組織,data=data)
summary(lm3.1b) #Multiple R-squared:  0.8198, Adjusted R-squared:  0.8182
                #交乘項共線性太高。不考慮。
```

對於經過對數轉換的模型，解釋能力意外的大幅將低。不僅 R-squared 減少至 0.6296，經過轉換的變數顯著程度也下降了，所以我們不採用這個模型，而是以第二個模型繼續分析。

```
#取log，考量變數轉換，將移轉層數限定>0會損失34個樣本
lm4.1=lm(總價元~log(建物移轉總面積平方公尺)+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
       +車位類別+電梯+公設比+屋齡+屋齡2+log(移轉層次_數字)+log(總樓層數_數字)
       +豪宅+主要用途+鄉鎮市區+有無管理組織,data=data[data$移轉層次_數字>0,])
summary(lm4.1) #Multiple R-squared:  0.6296,  Adjusted R-squared:  0.6265
               #面積顯著程度提高，移轉層次的顯著程度降低
```

```
電梯                6803019    10502804    0.648 0.517189
公設比             -2119847     4191475   -0.506 0.613054
屋齡                -558171       90350   -6.178 7.04e-10 ***
屋齡2                 9367        1614    5.805 6.87e-09 ***
log(移轉層次_數字)   20160      548237    0.037 0.970669
log(總樓層數_數字)  9717574     1669146    5.822 6.20e-09 ***
豪宅               72251133     2329126   31.021  < 2e-16 ***
```

```
Residual standard error: 21400000 on 4740 degrees of freedom
Multiple R-squared:  0.6296,    Adjusted R-squared:  0.6265
F-statistic: 201.4 on 40 and 4740 DF,  p-value: < 2.2e-16
```

同樣的以級距類型的資料替代原始變數進行回歸後，發現模型的解釋能力也大幅下降了。所以我們也不考慮轉換過的變數，而是以第二個模型為基礎，逐一刪除不需要的變數。

```
#使用級距代表移轉層次與面積
lm5.1=lm(總價元~面積大小+建物型態+建物現況格局.房+建物現況格局.衛+建物現況格局.廳
         +車位類別+電梯+公設比+屋齡+屋齡2+移轉層次_級距+log(總樓層數_數字)
         +豪宅+主要用途+鄉鎮市區+有無管理組織,data=data)
summary(lm5.1) #Multiple R-squared:  0.5266,  Adjusted R-squared:  0.5225
               #顯著程度還在，但解釋能力也大幅下降。
```

```
Residual standard error: 24130000 on 4768 degrees of freedom
Multiple R-squared:  0.5291,    Adjusted R-squared:  0.5245
F-statistic: 116.5 on 46 and 4768 DF,  p-value: < 2.2e-16
```

我們同樣的挑選出有無管理組織與車位類別這三個顯著程度較不明顯的變數來刪除，並比對刪除前後的模型是否有顯著差異。前面曾被檢驗的"建物現況格局.廳"變數如今有一定程度的顯著程度，因此不考慮刪除。

```
#以lm2.1為基礎刪減變數
lm6.1 <- update(lm2.1, . ~ . - 有無管理組織)
summary(lm6.1) #Multiple R-squared:  0.8181,  Adjusted R-squared:  0.8166
anova(lm6.1,lm2.1) #有10%的顯著，不可剃除

lm7.1 <- update(lm2.1, . ~ . - 車位類別)
summary(lm7.1) #Multiple R-squared:  0.8177,  Adjusted R-squared:  0.8164
anova(lm7.1,lm6.1) #不顯著，可剃除
```

檢驗過後，管理組織本身的 p-value，即 anova 測試的 p-value 都為 0.07027，雖然不高，但我們認為還沒到需要刪除的程度。而車位類別則不顯著，可以從模型中排除。最後的模型如下：

總價元 ~ 建物移轉總面積平方公尺 + 建物型態 + 建物現況格局.房 + 建物現況格局.衛 + 建物現況格局.廳 + 主要用途 + 電梯 + 公設比 + 屋齡 + 屋齡2 + 移轉層次_數字 + 總樓層數_數字 + 豪宅 + 鄉鎮市區 + 有無管理組織

後續將以 lm7 與 lm7.1 作進一步的分析、解讀與模型評估。

lm7 VIF 測試:

首先對 lm7 進行 VIF 檢定，檢定結果如下：

```
> vif(lm7)
                                    GVIF Df GVIF^(1/(2*Df))
log(建物移轉總面積平方公尺)    2.370834  1        1.539751
建物型態                     820.475713  3        3.059691
建物現況格局.房                2.770125  1        1.664369
建物現況格局.衛                2.396287  1        1.547995
車位類別                       3.778187  7        1.099599
電梯                         233.481293  1       15.280095
公設比                         6.602709  1        2.569574
屋齡                          22.761960  1        4.770950
屋齡2                         23.408382  1        4.838221
log(移轉層次_數字)             1.685568  1        1.298294
log(總樓層數_數字)             8.541677  1        2.922615
豪宅                          1.256671  1        1.121014
主要用途                      1.722827  7        1.039619
鄉鎮市區                      1.825889 11        1.027745
```

由結果顯示在電梯、屋齡、屋齡2這三個項目中存在著較高的共線性。

當中又以電梯項為最高(15.280095)，因此優先考慮此變數，藉由資料中的變數評估可推測電梯可能與建物型態有高度相關(已被包含)，所以要考慮對這兩個數據做轉換或是取捨。

結果如下：

1. 直接刪除電梯這個變數：

```
> lm7_1<-lm(單價元平方公尺 ~ log(建物移轉總面積平方公尺) +
+             建物型態 + 建物現況格局.房 + 建物現況格局.衛 +
+             車位類別  + 公設比 + 屋齡 + 屋齡2 + log(移轉層次_數字) +
+             log(總樓層數_數字) + 豪宅 + 主要用途 + 鄉鎮市區,
+           data = data[data$移轉層次_數字 > 0, ])
> vif(lm7_1)
                                   GVIF Df GVIF^(1/(2*Df))
log(建物移轉總面積平方公尺)  2.369262  1        1.539241
建物型態                    11.771634  3        1.508248
建物現況格局.房              2.769001  1        1.664031
建物現況格局.衛              2.341968  1        1.530349
車位類別                     3.725498  7        1.098497
公設比                       6.599708  1        2.568990
屋齡                        22.761944  1        4.770948
屋齡2                       23.394836  1        4.836821
log(移轉層次_數字)           1.685352  1        1.298211
log(總樓層數_數字)           8.522540  1        2.919339
豪宅                         1.256055  1        1.120739
主要用途                     1.710673  7        1.039094
鄉鎮市區                     1.819822 11        1.027589
```

2. 直接將電梯與建物型態交乘後共線性太高或類別變數太多，無法執行 VIF

```
> lm7_2<-lm(單價元平方公尺 ~ log(建物移轉總面積平方公尺) +
+             建物型態*電梯 + 建物現況格局.房 + 建物現況格局.衛 +
+             車位類別  + 公設比 + 屋齡 + 屋齡2 + log(移轉層次_數字) +
+             log(總樓層數_數字) + 豪宅 + 主要用途 + 鄉鎮市區,
+           data = data[data$移轉層次_數字 > 0, ])
> vif(lm7_2)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
錯誤發生在 vif.default(lm7_2): there are aliased coefficients in the model
```

3. 利用建物型態_電梯 <- interaction (建物型態,電梯) 做轉換

```
> data$建物型態_電梯 <- interaction(data$建物型態, data$電梯)
> lm7_3<-lm(單價元平方公尺 ~ log(建物移轉總面積平方公尺) + 建物型態_電梯
+             + 建物現況格局.房 + 建物現況格局.衛 +
+               車位類別  + 公設比 + 屋齡 + 屋齡2 + log(移轉層次_數字) +
+               log(總樓層數_數字) + 豪宅 + 主要用途 + 鄉鎮市區,
+             data = data[data$移轉層次_數字 > 0, ])
> vif(lm7_3)
                                GVIF Df GVIF^(1/(2*Df))
log(建物移轉總面積平方公尺)  2.370834  1        1.539751
建物型態_電梯               12.641916  4        1.373177
建物現況格局.房              2.770125  1        1.664369
建物現況格局.衛              2.396287  1        1.547995
車位類別                     3.778187  7        1.099599
公設比                       6.602709  1        2.569574
屋齡                        22.761960  1        4.770950
屋齡2                       23.408382  1        4.838221
log(移轉層次_數字)           1.685568  1        1.298294
log(總樓層數_數字)           8.541677  1        2.922615
豪宅                         1.256671  1        1.121014
主要用途                     1.722827  7        1.039619
鄉鎮市區                     1.825889 11        1.027745
```

比較 lm7_1 與 lm7_3，兩者都讓模型中 VIF 的最高值顯著下降，又 lm7_3 的 R-squared 較高

```
summary(lm7)  ##Multiple R-squared:  0.5379,  Adjusted R-squared:  0.5342
summary(lm7_1)##Multiple R-squared:  0.5367,  Adjusted R-squared:  0.5331
summary(lm7_3)##Multiple R-squared:  0.5379,  Adjusted R-squared:  0.5342
```

4. 考量其他共線性較高之項目:剩下 VIF 值較高的部分為屋年與屋齡2，考量屋齡2為屋齡^2本身即存在一定的共線性，且進行 anova 後發現屋齡2顯著不為 0，因此暫時不做更動，在目前皆不超過 5 的情況下暫不進行變換，因此接續依 lm7_3 繼續做分析。

5. 尋找 outliers or influential point並去除

(一) 利用 cooks distance，studentized residuals 及 leverage 進行檢測

```
> cooklm7_3 <-cooks.distance(lm7_3)
> plot(cooklm7_3, xlab="ID number")
> cooklm7_3[cooklm7_3 > 1.0]
<NA>
  NA
> cooklm7_3[cooklm7_3> 0.5]
<NA>
  NA
```

```
> ##leverage
> levlm7_3 <- hatvalues(lm7_3)
> k <- 3; n <- nrow(lm7_3)
> ( thr3 <- 3*(k+1)/n )
numeric(0)
> ( thr2 <- 2*(k+1)/n )
numeric(0)
> plot(levlm7_3, xlab="ID Number", ylab="Leverage")
> abline(h=thr3, lty=2); abline(h=thr2,lty=3)
> levlm7_3[levlm7_3 > thr3]
named numeric(0)
```

```
> reslm7_3 <- rstandard(lm7_3)
> hist(reslm7_3, breaks=20)
> reslm7_3[reslm7_3<(-3) | reslm7_3>3]
      138       266       290       301       322       429       703       824      1039      1040
 5.995235  3.045262  3.397008  3.314679  6.986890  3.728351  4.550567  3.132645  6.654371  3.598766
     1051      1055      1078      1119      1120      1123      1252      1267      1268      1288
-3.612437  6.661314  4.040747  5.716895  5.757069  3.516803  5.370212  3.707358  8.683976  3.976075
     1349      1405      1496      1497      1514      1598      1812      <NA>      2001      2291
 3.613677  3.058202  4.628787  4.628787  3.126406  5.265812 -4.687216        NA  3.227972  4.846644
     2420      2452      2712      2774      3180      3345      3531      3593      3638      3659
 3.898526 -3.175111  3.604203  4.382223  5.893606  4.052965 -3.047381  5.504618  4.080435 -3.052973
     3669      3742      3840      3892      3897      3920      4069      4079      4428      4435
 9.599319  6.260849  4.512686  4.684756 -3.981615  3.359122  3.091681  4.041595  5.591863  3.491100
     4452      4458      4463      4536      4615      4636
 3.392867  5.028475  4.839499 -4.120175  3.448848  3.081313
```

```
> ##remove outlier
> outlier_indices <- which(reslm7_3 < -3 | reslm7_3 > 3)
> data_clean <- data[-outlier_indices, ]
> any(is.na(data_clean))
[1] FALSE
```

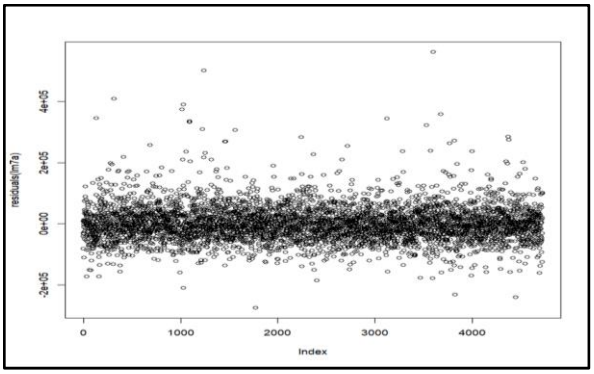
透過以上的步驟進行離群值的刪除，並以整理過後的資料繼續進行分析

```
lm7a  <- lm(單價元平方公尺 ~ log(建物移轉總面積平方公尺) + 建物型態_電梯
            + 建物現況格局.房 + 建物現況格局.衛 +
              車位類別  + 公設比 + 屋齡 + 屋齡2 + log(移轉層次_數字) +
              log(總樓層數_數字) + 豪宅 + 主要用途 + 鄉鎮市區,
            data = data_clean[data_clean$移轉層次_數字 > 0, ])
```

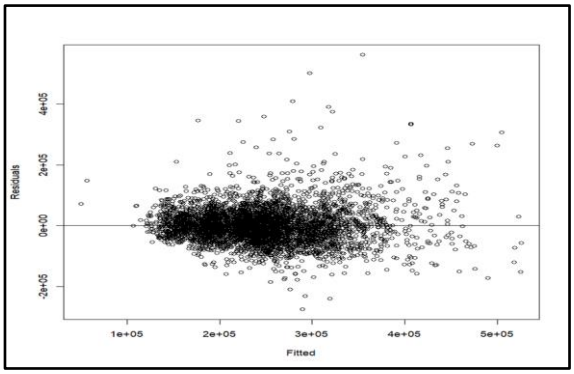
6. 比較原本模型與去除 outliers 之後的模型：
原模型 (lm7)：
Multiple R-squared: 0.5379, Adjusted R-squared: 0.5342
調整後模型lm(7a)：
Multiple R-squared: 0.5403, Adjusted R-squared: 0.5365
可以發現在調整後，Multiple R-squared 與 Adjusted R-squared：0.5365，皆略微提升。

7. 評估：
接下來將對建立的模型建立一些基本的測試與評估，評估是否符合統計的 3+1 個假設，並評估我們建立的模型。

(1) 殘值的 Scatter plot

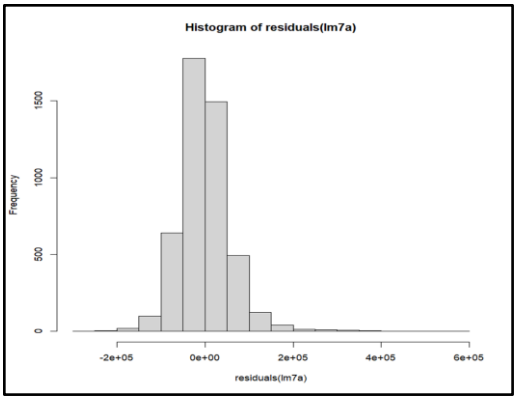


(2) Residual plot

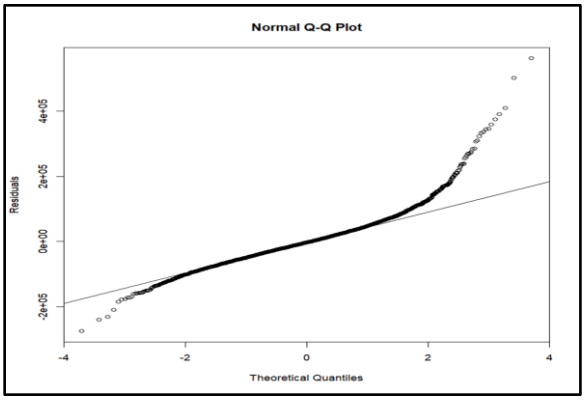


(1) 為本模組的殘差散佈圖，可以觀察到殘差大多分布於 0 的上下接近平均值，僅有極少數殘差偏離較大。整體而言，殘差未呈現明顯的模式或異常值分布，亦無特別高或特別低的情形，因此符合殘差期望值為 0 的假設。
(2) 從圖中可以看出，殘差在水平線附近大致呈現均勻分布，但在圖形右側，有部分殘差未完全服從均勻分布，且少數值明顯偏離水平線。針對這點，可以進一步嘗試進行變數轉換或刪除某些自變數項目，以優化模型。但整體來說，模型仍大致符合假設。

(3) 殘差直方圖



(4) QQ plot



(3) 整體而言，數據分布大致呈現接近對稱的鐘型分布，因此可以認為符合常態分布的假設。
(4) 從 QQ plot 的結果來看，殘差在線的兩端偏離較多，在首尾部分可以明顯觀察到殘差的偏離，這可能表示模型還尚未納入某些重要變數或交互項或可以嘗試其他的變數組合，來改善此殘差分布出現異常的狀況。

8. 嘗試優化模型:

經過變數轉換後發現，將模型中的 log(移轉層次_數字) 替換為 移轉層次 + (移轉層次_數字)^2，並移除總樓層數變數(由原本檔案中的資料判讀，兩個樓層數的變數可能有較高的相關性)後，模型的 Multiple R-squared 和 Adjusted R-squared 均有所提升，顯示模型的解釋能力得到了增強。

```
lm7a.12<-lm(單價元平方公尺 ~ log(建物移轉總面積平方公尺)+ 建物型態_電梯
            + 建物現況格局.房 + 建物現況格局.衛 +
              車位類別 + 公設比 + 屋齡 + 屋齡2+ 移轉層次_數字+sq移轉層次_數字+ 袁宅 + 主要用途 + 鄉鎮市區,
            data = data_clean[data_clean$移轉層次_數字 > 0, ])
```

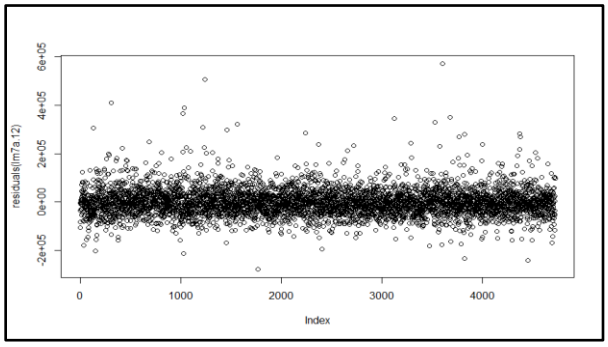
Multiple R-squared: 0.5481, Adjusted R-squared: 0.5445

原模型為：

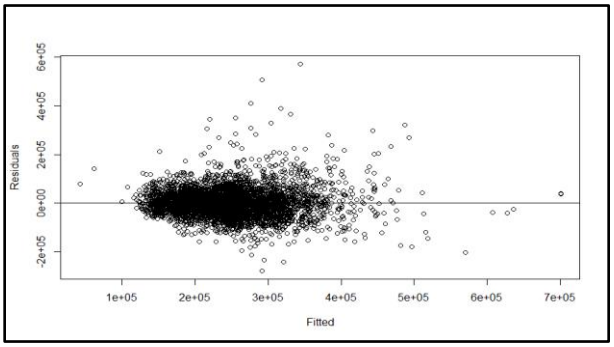
Multiple R-squared: 0.5403, Adjusted R-squared: 0.5365

模型評估：

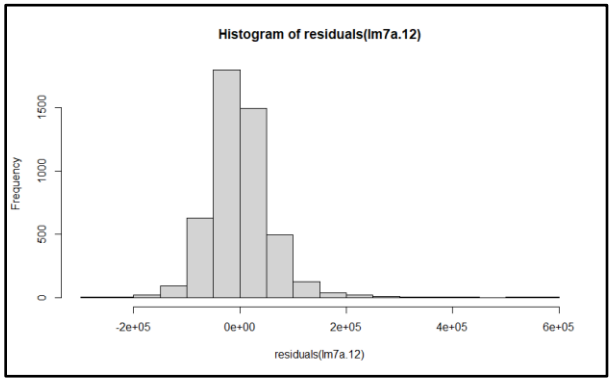
(1) 殘值的 Scatter plot



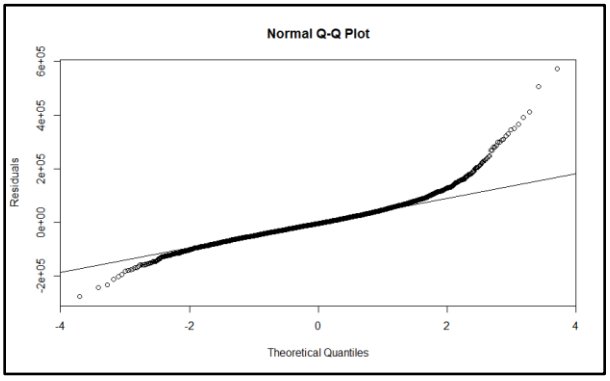
(2) Residual plot



(3) 殘差直方圖



(4) QQ plot



在模型評估的 (3+1) 檢定中前面3個與原模型相似，但在 qqplot 部份仍舊沒有明顯改善:

結論：

經過嘗試對各變數進行多種轉換（如倒數和平方）與刪減部分變數後，問題仍未改善。除了模型 lm7a.12 外，其餘模型在 R-squared 的提升幅度並不顯著，圖形表現也未見明顯改善。因此，推測可能有其他影響每平方公尺單價元的關鍵變數，例如:如政策因素、交通環境或都市翻展計畫等未被納入資料集。最終，選擇採用模型 lm7a.12 作為分析基礎。

新模型 lm7.1 VIF 測試：

首先對 lm7.1 進行 VIF 檢定，檢定結果如下：

```
> vif(lm7.1)
                          GVIF Df GVIF^(1/(2*Df))
建物移轉總面積平方公尺   1.576496  1        1.255586
建物型態               608.820138  3        2.911263
建物現況格局.房          2.938458  1        1.714193
建物現況格局.衛          2.397079  1        1.548250
建物現況格局.廳          1.570358  1        1.253139
主要用途                 1.784988  7        1.042255
電梯                   234.272041  1       15.305948
公設比                   5.112751  1        2.261139
屋齡                    22.401503  1        4.733023
屋齡2                   22.536744  1        4.747288
移轉層次_數字            1.861599  1        1.364404
總樓層數_數字            4.835971  1        2.199084
豪宅                     1.314206  1        1.146388
鄉鎮市區                 1.711634 11        1.024730
有無管理組織             3.438170  1        1.854230
```

同樣地結果顯示在電梯、屋齡、屋齡2這三個項目中存在著較高的共線性

當中又以電梯項為最高(15.305948)，因此優先考慮此變數，藉由資料中的變數評估可推測電梯可能與建物型態有高度相關(已被包含)，所以要考慮對這兩個數據做轉換或是取捨。

結果如下：

1. 直接刪除電梯這個變數：

```
> lm7.11<-lm(總價元 ~ 建物移轉總面積平方公尺 +
+              建物型態 + 建物現況格局.房 + 建物現況格局.衛 +
+              建物現況格局.廳 + 主要用途 + 公設比 +
+              屋齡 + 屋齡2 + 移轉層次_數字 + 總樓層數_數字 +
+              豪宅 + 鄉鎮市區 + 有無管理組織, data = data)
> vif(lm7.11)
                          GVIF Df GVIF^(1/(2*Df))
建物移轉總面積平方公尺  1.575172  1        1.255059
建物型態                10.254379  3        1.473957
建物現況格局.房          2.938430  1        1.714185
建物現況格局.衛          2.338649  1        1.529264
建物現況格局.廳          1.549916  1        1.244956
主要用途                 1.772709  7        1.041741
公設比                   5.105342  1        2.259500
屋齡                    22.400097  1        4.732874
屋齡2                   22.503337  1        4.743768
移轉層次_數字            1.861596  1        1.364403
總樓層數_數字            4.835931  1        2.199075
豪宅                     1.312765  1        1.145760
鄉鎮市區                 1.706639 11        1.024594
有無管理組織             3.432940  1        1.852820
```

2. 直接將電梯與建物型態交乘後共線性太高或類別變數太多，無法執行 VIF

3. 利用建物型態_電梯 <- interaction (建物型態,電梯) 做轉換

```
> data$建物型態_電梯 <- interaction(data$建物型態, data$電梯)
> lm7.13<-lm(總價元 ~ 建物移轉總面積平方公尺 +
+              建物型態_電梯 + 建物現況格局.房 + 建物現況格局.衛 +
+              建物現況格局.廳 + 主要用途 + 公設比 +
+              屋齡 + 屋齡2 + 移轉層次_數字 + 總樓層數_數字 +
+              豪宅 + 鄉鎮市區 + 有無管理組織, data = data)
> vif(lm7.13)
                          GVIF Df GVIF^(1/(2*Df))
建物移轉總面積平方公尺  1.576496  1        1.255586
建物型態_電梯           10.985025  4        1.349274
建物現況格局.房          2.938458  1        1.714193
建物現況格局.衛          2.397079  1        1.548250
建物現況格局.廳          1.570358  1        1.253139
主要用途                 1.784988  7        1.042255
公設比                   5.112751  1        2.261139
屋齡                    22.401503  1        4.733023
屋齡2                   22.536744  1        4.747288
移轉層次_數字            1.861599  1        1.364404
總樓層數_數字            4.835971  1        2.199084
豪宅                     1.314206  1        1.146388
鄉鎮市區                 1.711634 11        1.024730
有無管理組織             3.438170  1        1.854230
```

比較 lm7.11 與 lm7.13，兩者都讓模型中 VIF 的最高值顯著下降，又 lm7.13 的 R-squared 較高

```
summary(lm7.1)##Multiple R-squared:  0.8177,  Adjusted R-squared:  0.8164
summary(lm7.11)##Multiple R-squared:  0.8175, Adjusted R-squared:  0.8163
summary(lm7.13)##Multiple R-squared:  0.8177, Adjusted R-squared:  0.8164
```

4. 與前面一樣考量其他共線性較高之項目:剩下vif值較高的部分為屋年與屋齡2，考量屋齡2為屋齡^2本身即存在一定的共線性，且進行anova後發現屋齡2顯著不為 0，因此暫時不做更動，在目前皆不超過 5 的情況下暫不進行變換，因此接續依 lm7.13 繼續做分析。

5. 尋找 outliers or influential point並去除

(一) 利用 cooks distance，studentized residuals 及 leverage 進行檢測

```
> cooklm7.13 <-cooks.distance(lm7.13)
> plot(cooklm7.13, xlab="ID number")
> cooklm7.13[cooklm7.13 > 1.0]
    <NA>      3989
      NA 1.008853
> cooklm7.13[cooklm7.13> 0.5]
    <NA>      3989
      NA 1.008853
```

```
> levlm7.13 <- hatvalues(lm7.13)
> k <- 3; n <- nrow(lm7.13)
> ( thr3 <- 3*(k+1)/n )
numeric(0)
> ( thr2 <- 2*(k+1)/n )
numeric(0)
> plot(levlm7.13, xlab="ID Number", ylab="Leverage")
> abline(h=thr3, lty=2); abline(h=thr2,lty=3)
> levlm7.13[levlm7.13 > thr3]
named numeric(0)
```

```
> reslm7.13 <- rstandard(lm7.13)
> hist(reslm7.13, breaks=20)
> reslm7.13[reslm7.13<(-3) | reslm7.13>3]
       37        165        192        263        433        703        715        768        799
-3.351637  -4.491100  -3.007492  -3.567546   4.351340   3.325292  -3.696873   3.684866   6.589164
     1039       1055       1119       1120       1238       1288       1355       1419       1496
 6.872570   6.773499   9.074330   9.124482  -3.596124   6.481236   3.691440   3.399999   8.900876
     1497       1598       1812       <NA>       2092       2095       2127       2155       2173
 8.900876  13.197426  -4.400771         NA  -3.964514   3.220416  -3.271475 -12.982656  -3.491348
     2299       2322       2420       2423       2774       3121       3189       3471       3473
-4.836568  -3.194096  12.350897  -3.039426  21.267459  -3.850657  -3.371610  -3.514813  -3.864406
     3512       3742       3769       3812       3832       3840       3844       3891       3892
-3.004254  10.296944   3.496079  -3.059282   4.858800  11.782965   8.021854  -6.196699  23.873535
     3989       4302       4318       4368       4372       4383       4397       4424       4452
-13.789210  -5.568633   4.339195   4.954158  -3.008082  -6.777922   3.747294  -3.128711  11.181460
     4535       4554       4584       4726       4812
-3.574853  11.574988  -3.861017   5.842142  -3.195228
```

```
outlier_indices <- which(reslm7.13 < -3 | reslm7.13 > 3)
data_clean2 <- data[-outlier_indices, ]
lm7b <- lm(總價元 ~ 建物移轉總面積平方公尺 +
             建物型態_電梯 + 建物現況格局.房 + 建物現況格局.衛 +
             建物現況格局.廳 + 主要用途 + 公設比 +
             屋齡 + 屋齡2 + 移轉層次_數字 + 總樓層數_數字 +
             豪宅 + 鄉鎮市區 + 有無管理組織, data = data_clean2)
```

透過以上的步驟進行離群值的刪除，並以整理過後的資料繼續進行分析。

6. 比較原本模型與去除 outliers 之後的模型:

原模型(lm7.13)：

Multiple R-squared: 0.8177, Adjusted R-squared: 0.8164

調整後模型lm(7b) ：

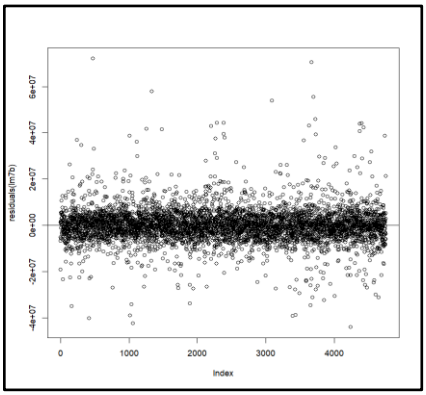
Multiple R-squared: 0.8933, Adjusted R-squared: 0.8926

可以發現在調整後，Multiple R-squared 與 Adjusted R-squared，皆有不小的提升。

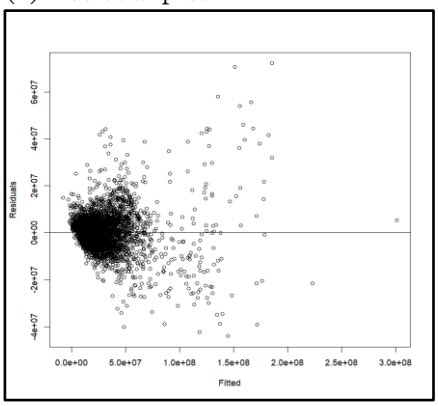
7. 評估：

接下來將對建立的模型建立一些基本的測試與評估，評估是否符合統計的 3+1 個假設，並評估我們建立的模型。

(1) 殘值的 Scatter plot：

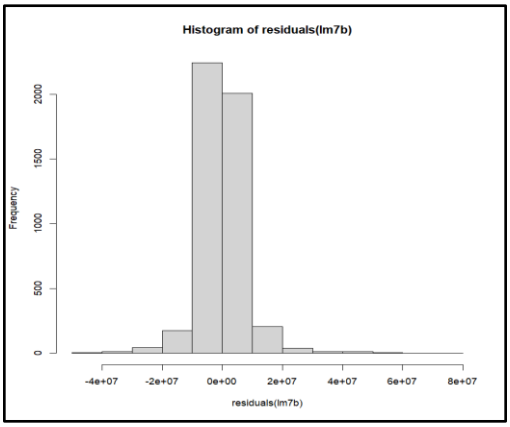


(2) Residual plot：

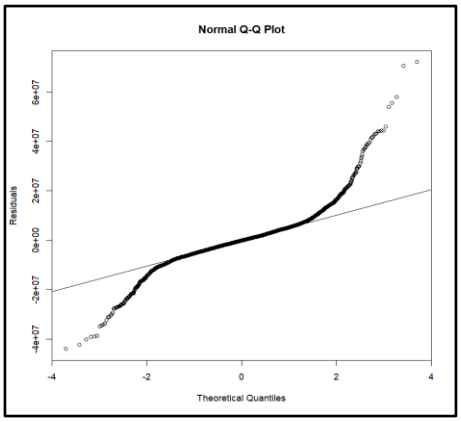


(1) 為本模型的殘差散佈圖，可以觀察到大部分殘差均勻分布於 0 的上下，未呈現明顯的特徵或異常的高低值情形，因此符合殘差期望值為 0 的假設。
(2) 從圖中可以看出，殘差大部分集中於圖形的左側，但在右側出現部分值未完全服從均勻分布，此外少部分數值明顯偏離水平線部分。針對這種情況，可以嘗試進行變數轉換或刪除部分自變數項目，以進一步調整和優化模型。

(3) 殘插直方圖



(4) QQ plot



(3) 整體而言，數據大致呈現接近對稱分布，但主要集中於中間部分，兩端的分布相對較少。
(4) 首尾部分的殘差呈現明顯偏離，這可能表示模型尚未納入某些重要的變數或交互項，導致殘差分布出現異常。需進一步檢視模型結構。

8. 小結:
經過嘗試對各個變數進行轉換（如倒數和平方）以及刪除某些自變數後，問題仍未改善。R-squared 的提升幅度並不顯著，且 3+1 檢定中的圖形表現也未見明顯改善。因此，推測可能存在其他對總價元有影響的關鍵變數，例如:政策因素、交通環境或都市計畫等，這些變數未被納入現有的資料集中。

七、研究結論

我們透過對台北市房屋交易資料進行分析，利用資料清理、EDA 和回歸建模，探討影響房價的因素，提出具有解釋力和分析價值的模型。以下是研究主要發現和未來研究方向。

1.主要研究發現：

- 影響房價的主要因素：
    - 建物面積：建物移轉總面積和房價呈顯著正相關，面積越大，房價越高，表示市場對大面積住宅的需求可能較高。
    - 屋齡：屋齡和房價之間存在非線性關係，新房屋因設計與建材較新而有較高價格，而較高屋齡的房屋可能因為稀缺性或改建的潛力，價格逐漸回升。

    - 豪宅：符合豪宅標準的建物價格顯著提升，表示台北高端住宅市場大、有強勁的消費客群。
    - 建物型態與設備：住宅大樓和華廈價格遠高於公寓與透天厝，有電梯和管理組織的建物價格也顯著提升，表示現代化設備和便利性是重要的買房考量因素。
    - 區域特徵：大安區、中正區和松山區的房價明顯高於其他區域，反映出明星學區、交通便利性和商業發展程度對房價的影響力。另外，豪宅和異常值也主要集中在士林、大安、信義等高價的區域。

- 回歸模型的解釋力：
    - 我們分別建構了「每坪單價」和「總價」的回歸模型，兩者均能有效解釋房價變異。其中，總價模型的解釋能力較高，R-squared 達到 0.8164，表示變數對總價的影響更為直接與穩定。
    - 屋齡平方項的加入解釋了屋齡和房價之間的非線性關係，提升了模型的準確度。
- 共線性問題與優化：
    - 模型檢測過程當中，我們發現電梯與建物型態之間存在共線性問題。透過變數交互作用的處理，降低了共線性影響並確保模型穩定性。
    - 儘管我們進一步移除了部分離群值，模型的 Adjusted R-squared 提升至 0.8926，但仍然沒有完全改善殘差的尾端偏離問題。

2.研究限制和未來研究方向

- 殘差分布與變數局限：
    - 雖然模型已達到較高的解釋力，但殘差圖與 QQ plot 顯示尾端偏離，可能是因為沒有納入政策、交通環境及都市發展計畫等外部變數，導致模型預測仍存在一定誤差。
    - 未來研究可以整合更全面的資料，例如政府土地政策、都市更新計畫、交通便利性及學區分布等，來進一步提升模型的解釋力和預測力。
- 資料面向擴展：
    - 我們的研究只針對台北市的房價資料進行分析，沒有考慮外部縣市或時間序列因素。後續研究可以擴展到其他城市或加入時間維度，探討房價的長期趨勢和整體市場動態。
- 非量化因素考量：
    - 房價可能也同時受到市場心理、人口結構變化和投資行為等非量化因素的影響，這些因素比較難透過回歸模型直接納入，未來可以結合其他建模方法（如機器學習）進一步做分析。