台北的獨立套房租金研究

一、動機

對於大多數的人來說，通常要預測或分析跟房屋有關的資料都會是售價，但幾乎沒有人去探討租金的問題，也鮮少看到相關的研究內容，而這就是我想要做這個系統的原因。

我想任何與價格有關的事物都會有被哄抬的可能性，如果有了這個系統，對於研究相關內容的學者來說，取得相關資料以及數據將會更加的方便；而對於一般民眾像是房東、租客等等，則可以透過這個系統去判斷該租金是否已超過市場的正常範圍，進而去達到調整租金或者是另選別處居住等等的目的。

二、目的

希望這個系統能夠讓大眾查看房屋的相關資料、租金的合理走向以及各項房屋資料相互的影響關係，讓大眾更有自主判斷租金合理性以及目前租房市場的趨勢。

三、相關工作

目前市面上都沒有出現能夠供大眾使用並查看的相關系統，因此對於這個系統所需的相關工作，就只有找到合適的資料來源網站，並確認該網站能夠只用動態爬蟲的方式取得網站。

四、系統描述

1. 開發的包（或模塊）和工具
   (1) Selenium：自動化網頁操作，用於模擬用戶行為（如點擊、滾動）並從動態網頁中提取數據。
   (2) Pandas：創建、操作和合併數據框。
   (3) Re（正則表達式）：從文本中提取特定模式的數據。
   (4) Seaborn：高級數據可視化庫，用於創建統計圖表，如熱圖。
   (5) Matplotlib：基礎數據可視化庫，用於創建各種圖表。
   (6) Scikit-learn：機器學習庫，用於數據預處理、模型訓練和評估。
2. 系統功能和主要技術
   (1) 自動化數據收集：
      a. 使用 Selenium 自動化瀏覽器操作，從 591 租屋網頁中提取房屋租賃信息，包括房屋名稱、租金、面積和距離捷運站的距離。
      b. 動態滾動頁面和點擊按鈕以獲取多頁數據。
   (2) 數據處理：
      a. 使用 Pandas 進行數據處理，轉換數據格式，處理缺失值。
      b. 使用正則表達式從文本中提取數字信息。
   (3) 數據分析和建模：

a. 使用 Scikit-learn 進行數據拆分和標準化。

b. 使用線性回歸模型對租金進行預測，並評估模型的性能（包括均方根誤差和決定係數）。

(4) 數據可視化：

a. 使用 Matplotlib 和 Seaborn 創建圖表，包括實際值與預測值的散點圖和特徵相關性的熱圖。。

3. 描述開發過程(遇到的問題)

(1) 進入網頁後無法抓取想要的資料

https://www.youtube.com/watch?v=ZOxIpMTgaXY&t=982s

後來想到之前的作業就有因為視窗未顯示的區域抓不到內容，所以想說要把網頁移到最下方，讓資料全都顯示出來，所以參考了上方網址的影片，成功控制網頁到最下方，也解決了進入網頁無法抓資料的問題

(2) 無法透過 class 名稱得到想要的資料

https://medium.com/marketingdatascience/%E5%8B%95%E6%85%8B%E7%B6%B2%E9%A0%81%E7%88%AC%E8%9F%B2%E7%AC%AC%E4%BA%8C%E9%81%93%E9%8E%96-selenium%E6%95%99%E5%AD%B8-%E5%A6%82%E4%BD%95%E4%BD%BF%E7%94%A8find-element-s-%E5%8F%96%E5%BE%97%E7%B6%B2%E9%A0%81%E5%85%83%E7%B4%A0-%E9%99%84python-%E7%A8%8B%E5%BC%8F%E7%A2%BC-b66920fc8cab

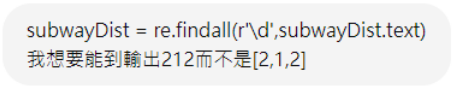
因為 class 名稱為 item-tip subway 中間有空白字元，所以上網搜尋是不是因為有空白字元的關係，導致無法透過一般 By.CLASS_NAME 作使用，在上方連結的網頁內容中提到如果遇到遇到空白字元，須將空白字元改成”.”，並更改為 By.CSS_SELECTOR 即可。根據網站的描述後進行修正確實成功抓取與捷運站的距離

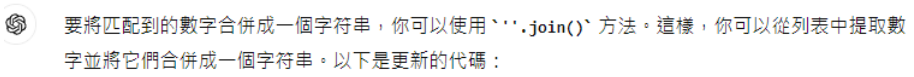
(3) 只要文字中的數字，其餘文字需省略不輸出

https://hackmd.io/@aaronlife/regular-expression

利用關鍵字在網路上搜尋後，使用了正規表達式來處理這個問題，而使用的方法是參考上方網站中的內容，因為只需要保留純數字的部分，所以用 r'\d'來完成需求。

因為要抓取的數字是離捷運站的距離，但透過正規表達式得到的結果會將每個數字分開，因此透過 ChatGPT 得知 join()這個方法，以下為對話內容以及改正的效果





由於我只更改有加上 join()的部分，其餘皆沒有更改，所以 ChatGPT 給的完整程式碼就不截圖，只附上自己程式碼的前後差別

```
subwayDist = re.findall(r'\d',subwayDist.text) #[2,1,2]
subwayDist = ''.join(re.findall(r'\d', subwayDist.text)) #212
```

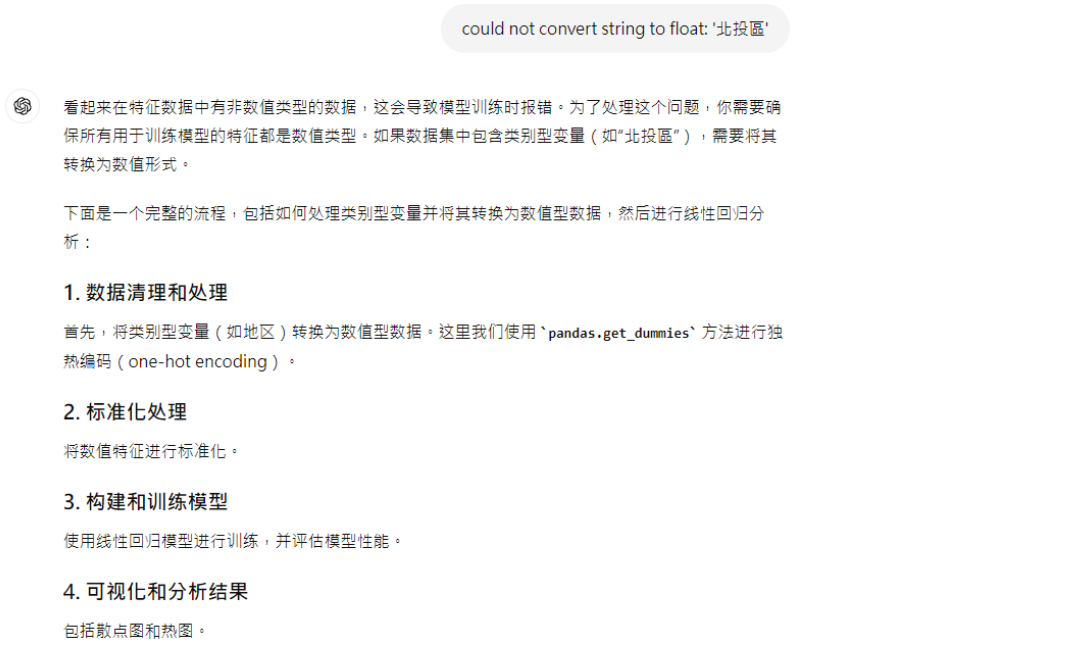
上方會以陣列形式輸出距離，而下方則會以我所要求的樣子輸出

(4) 抓取下一頁資料時，不斷把前一頁資料覆蓋

根據 ChatGPT 所提供的方式實際做更改，確實解決了問題，成功留下前一頁的資料並繼續記錄不覆蓋

以上為根據 ChatGPT 提供的解決辦法，實際有使用的 ChatGPT 範例

(5) 資料內的中文讓模擬無法進行



由於資料內有行政區，造成了分析資料時的錯誤，根據 ChatGPT 所給的建議做調整

```
# 將類別變數轉換為獨熱編碼
all_house_data = pd.get_dummies(all_house_data, columns=['Name'], drop_first=True)
```

將行政區都更改為獨熱編碼，即可正常運行

五、結論和反思

在整個系統的運行過程中，我一直都還沒有辦法克服的就是 RMSE 值的部分，從一開始用 KNN 分析，再換成使用四分線法、決策樹等等，但不管是 RMSE 值還是 MSE 值都出現極大的數字，這使得分析的結果可以說是沒什麼參考價值，後來改成了線性回歸的模型，才成功將 RMSE 值降到 7000 以下，中間有改了好幾次數據的總量，但依舊沒有太大的效果。

如果說準確度是這個系統很明顯的缺點，但我想系統所產生的熱圖就會是這個系統的優點，雖然沒辦法精準，但卻可以透過熱圖看出每個特徵間的關係，以及相互影響的程度，對於參考的部分，我想是有發揮它的功用的。

未來我希望能夠再多增加其他的特徵，像是是否有租房補助、室內設施、學區等等的其他條件，以及增加各區個別分析的部分，這樣就可以根據各區的所得收入以及其他條件去做到更詳盡的分析。

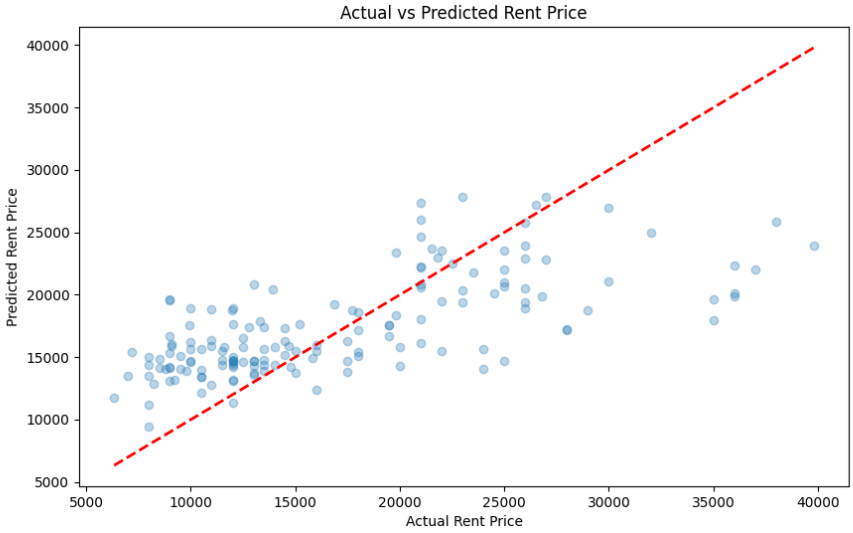
以下為結果及圖表展示：

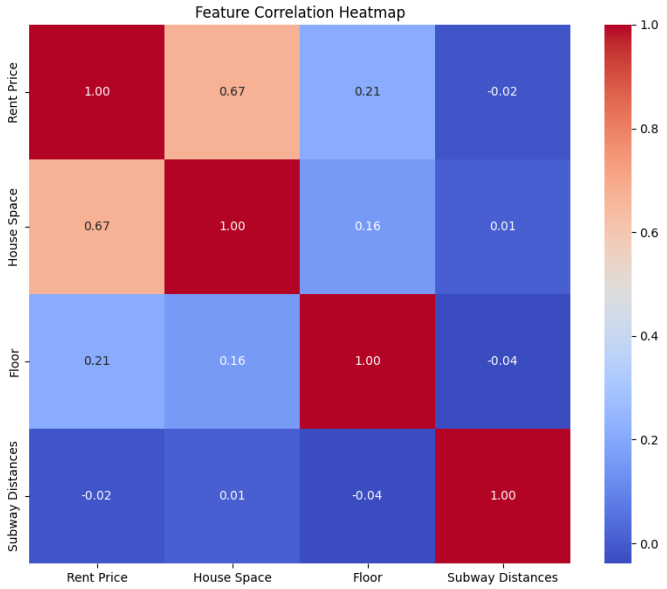
爬蟲的資料輸出(區域、租金、坪數、樓層、跟捷運站的距離)
區域的部分為大安區、內湖區、士林區、文山區、北投區的所有資料
但因其他設定關係，目前只能統一輸出最後一個北投區

	Name	Rent	Price	House Space	Floor	Subway Distances
0	北投區	16000		9.5	10	544
1	北投區	12499		7	5	289
2	北投區	10999		8	1	1484
3	北投區	20000		16	1	238
4	北投區	13000		7	4	461
..	...	...		...	...	...
766	北投區	7200		8	1	393
767	北投區	6333		4	1	757
768	北投區	41850		11	4	197
769	北投區	11500		5	5	1078
770	北投區	7500		5	1	822

```
缺失值檢查：
Name                0
Rent Price          0
House Space         0
Floor               0
Subway Distances    0
dtype: int64
Rent Price           int64
House Space         object
Floor               object
Subway Distances    object
dtype: object
```

```
Train RMSE: 6450.695519471907
Test RMSE: 5727.751244627665
Train R2: 0.4667207356191595
Test R2: 0.4415300950792064
```





六、資料來源

租金資料 https://rent.591.com.tw/