國立交通大學

資訊學院 資訊學程

碩士論文

巨量資料：公開資料與房仲網的房價分析

Big Data:Open Data and Realty Website Analysis

研究生：陳珍華

指導教授：袁賢銘 教授

中華民國一〇三年六月

巨量資料:公開資料與房仲網的房價分析

Big Data:Open Data and Realty Website Analysis

研究生: 陳珍華　　Student : Zhen-Hua Chen

指導教授: 袁賢銘　　Advisor : Shyan-Ming Yuan

國立交通大學

資訊學院 資訊學程

碩士論文

A Thesis

Submitted to College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Computer Science

June 2014

Hsinchu, Taiwan, Republic of China

中華民國一〇三年六月

巨量資料:公開資料與房仲網的房價分析

研究生:陳珍華　　指導教授:袁賢銘

國立交通大學 資訊學院 資訊學程碩士班

摘　要

多數人購屋資訊來自親友介紹，房仲網，和實價登錄網。不過這些資料分散在不同地方，缺少直接比較的資訊。本實驗用新竹縣實價登錄資料來建立分析房價模型。理解資料後，將不相關的房屋過濾掉，比如，篩選掉商業用途的辦公大樓。使用 K-means 分群，得到房仲網的平均價格比開放資料高的結論。使用「坪數」與「屋齡」，來看房仲網和開放資料的比值，並找實際物件來驗證。 在 Ubuntu 上，安裝 Apache 網頁伺服器和 MySQL 資料庫來架設網站。使用 HTML 和 PHP 語言來編寫網頁。資料庫字集設定為 UTF8-Unicode，來處理內容為中文的房價資料。定期自動到政府資料開放平台和 Yahoo 奇摩房地產網，取得不動產買賣實價登錄資料和房仲網資料，由 Python 處理後，自動匯入資料庫，申請網路空間，提供房價分析服務 。本服務提供： 顯示同縣市同坪數同屋齡房仲網與開放資料的比較資訊， 總價元比值「平均值」和「標準差」。加入 Google Analytics 來收集使用者的瀏覽行為 。最後使用問卷取得使用者的意見回饋。本實驗對於購屋的消費者，提供房屋議價空間的資訊。

關鍵詞：房價分析，房仲網，實價登錄網。

Big Data:Open Data and Realty Website Analysis

Student: Zhen-Hua Chen Advisor:Shyan-Ming Yuan

Degree Program of Computer Science

National Chiao Tung University

ABSTRACT

Information for buying a house is from friends, real-agent-web and register-real-price data for most people. However ,those data are in different places.There are no direct comparison. I build a model by using register-real-price data of Hsinchu County.First ,I observe data and delete irrelevant data.For example ,I delete office buildings for commercial purposes. Second,I use K-means clustering to get the conclusion.The average price of real-agent-web is higher than the average price of register-real-price .Third,I calculate ratios of real-agent-web 's price to register-real-price's price by the conditions of 「square feet」 and「age of building」. Fourth,I find some real instances to support the experiment.Fifth,I install Apache and MySQL in Ubuntu and write HTML and PHP. I use the UTF-8 character set to process Chinese words in the house-price data. I write a shell script.It can get data from data.gov.tw and tw.house.yahoo.com termly and automatically. I write Python code to process data.The program imports them to the database automatically.I apply for a web space in order to provide the service that analyzes house prices.The system compares the house-price information of real-agent-web and register-real-price in same counties,「square feet」 and「age of building」. It also shows mean and standard deviation of price's ratios.I use 「Google Analytics」 to observe user's browsing behavior .I get users' feedback by questionnaires. In conclusion,the analysis of house prices is useful for consumers.

Keywords：House,Price,Analyze.

誌　謝

能夠順利完成碩士論文，首先我要感謝我的指導教授，在研究上給予我足夠的自由發揮空間，循序漸進地給予適當的建議，指引我完成實驗，從中學習到遇到問題，如何去學習解決問題的能力。感謝口試委員們給予我許多改善論文的建議。也非常感謝 DCS 實驗室的學長姐和同學們，有你們的支持和鼓勵，使我在研究生生涯中感到溫暖和開心。接著我要感謝我爸爸，媽媽，弟弟，有你們的陪伴，使我可以一路安心求學。接著感謝交大給予我這個友善的環境，可以在這裡盡情去學習，感謝我在交大所選修過課的老師和助教們，你們經驗豐富的授課和指導，給予我一生中寶貴的知識寶藏。也感謝我的朋友們，在這一條努力的路上，感謝大家的幫忙，謝謝你們

陳珍華 謹誌
2014 年 6 月

Table of Contents

Index of Tables

Illustration Index

一 、緒論

1.1 引言

3C 產品普及於大眾，這些數位科技產品隨時隨地都在產生資料，Twitter 的推文，Facebook 上按讚次數，LinkedIn 等社群網站[8]每天都在產生大量數據。

分析巨量資料的好處，例如，亞馬遜解散書評團隊[2]就是電腦自動化取代專家的例子，推薦系統比專家更能讓讀者購買更多的書籍。挖掘何時在電子商務網站上購買商品是最省錢。唱片公司用預測演算法，對可成為暢銷歌曲砸下重金宣傳[3]。

1.2 研究動機與目的

購買房子對大多數人而言，可能是一生中，支出最大筆的交易。如何選擇價格合理且符合需求的房子，是一個重要課題。以往購買房屋的資訊來源，價格資訊單方面掌握在房仲業者手上，或者靠親友口耳親傳來預估房價。近年來網路對於資訊傳播有開放的貢獻，房價實價登錄網把以往資訊獨佔於房仲業者的房屋成交價格，開放在網路上，但這些資料是分散在不同網頁上，消費者缺少綜觀比較房價的資訊。比如說，網路上的房仲業者除了提供賣屋的價格，也提供實價登錄價格，但只有單方面查詢的功能，無法知道賣方資訊和政府開放資料兩者間的差異。倘若能整合開放資料和房仲網，將可幫助消費者在購買房屋時，有數據依據的議價參考。 本實驗目的是從房價實價登錄資料集和房仲網資料中，找出購屋的議價空間。網路上實價登錄網和房仲網的資料是分開查詢的，並沒有一起比較的功能。本論文將兩者放在一起比較，並將結果以網站來提供服務。

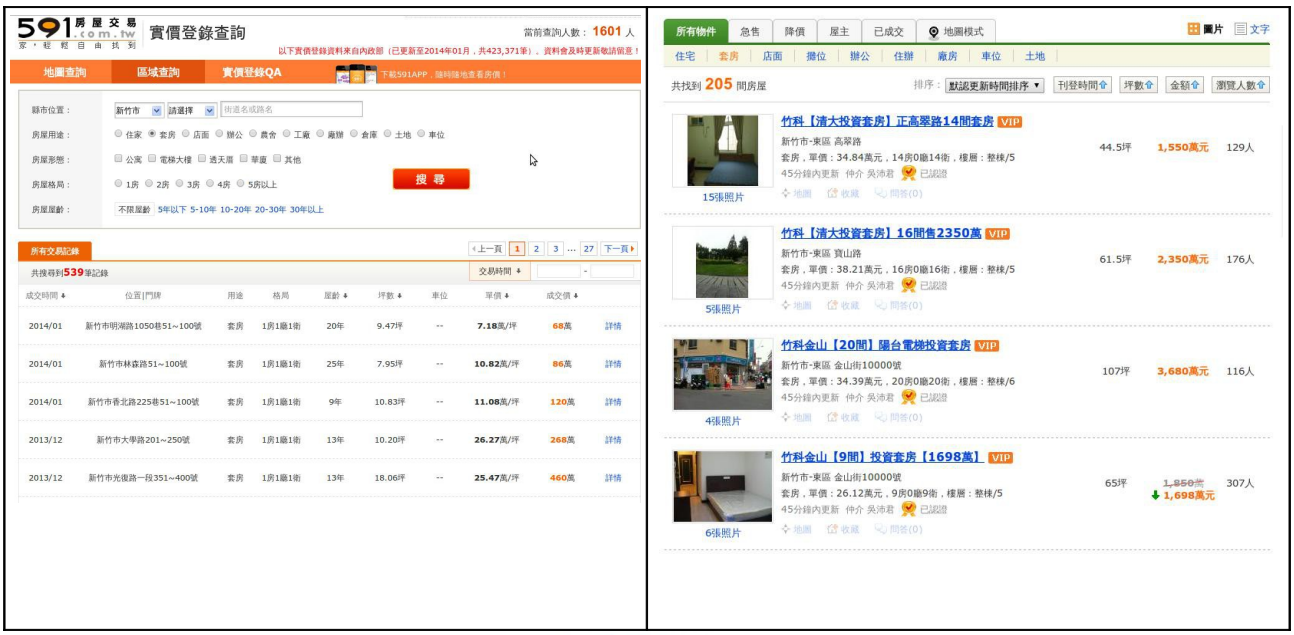


Illustration 1: 房仲網的「賣屋物件查詢」網頁

Illustration 2: 房仲網的「實價登錄查詢」網頁

資料來源：591 房屋交易網

1.3 論文架構

實驗分兩部分，第一部分是新竹市房屋數據為來建置分析模型。第二部分建置網站最新的房價分析服務。最後以問卷取得回饋。

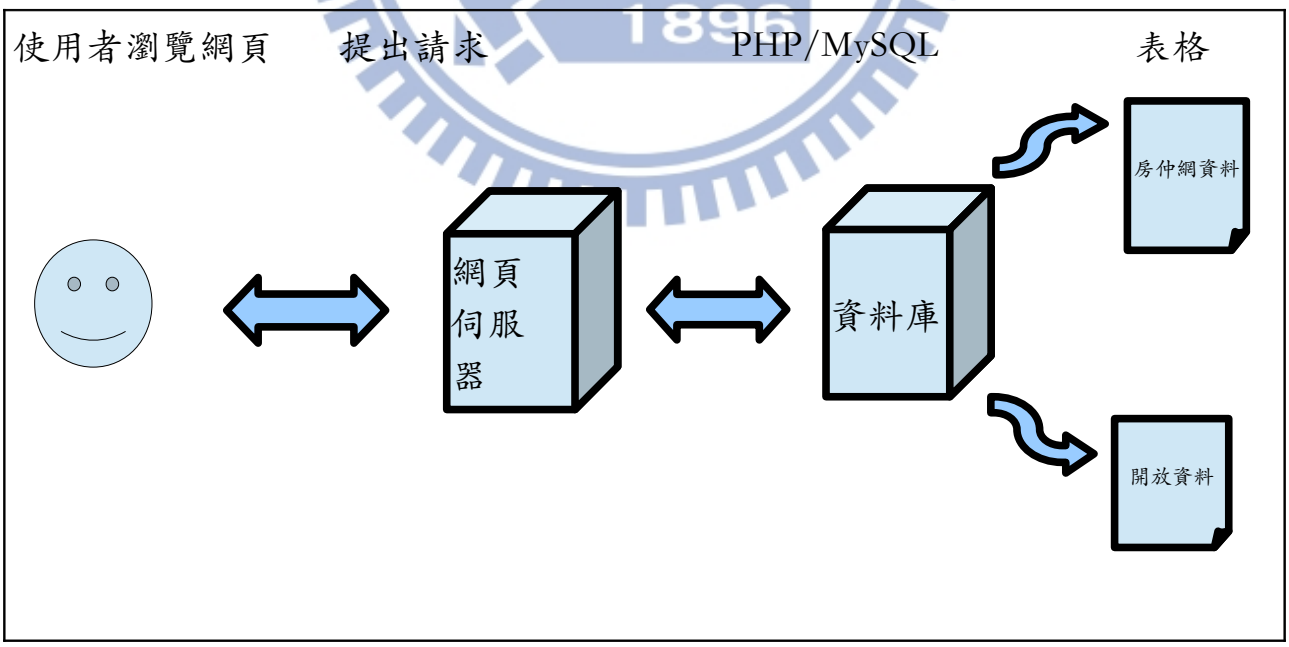


Illustration 3: 房價分析服務的系統架構

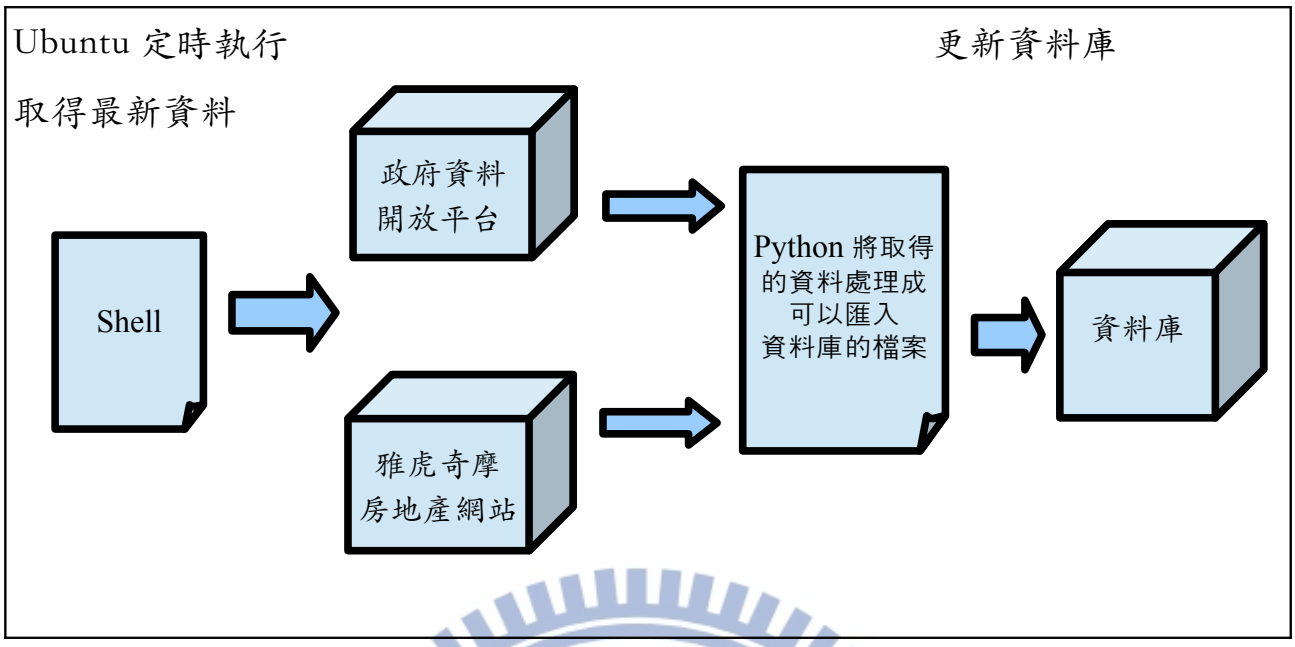


Illustration 4: 定期自動更新資料庫流程

二、研究背景與文獻探討

2.1 巨量資料分析帶來好處

藉著分析巨量資料，而挖掘出有價資訊的例子越來越多。舉例來說，「美國沃瑪百貨分析過去商品販賣紀錄，在颱風來臨前，在賣場大量擺放小甜餅，提高業績。」[4]，「公司分析歷史記錄，將廣告傳單寄給購買機率高的消費者，而避免將宣傳費用放在購買機率極低的消費者上。」[5]

2.2 政府資料開放平台

許多在過去就存在的數據，但目前尚未被找出新的用途和價值。舉例來說，政府資料開放平台有全國郵局 ATM 分佈資料集，紀錄全國郵局 ATM 的經緯度，是否有自動補摺機，是否有自動提款功能，ATM 設定在郵局裡面還是外面。這些資料自從郵局成立以來就一直存在，這些開放數據用文字或數字來描述事物的特性，以 csv 或 xml 的形式放在平台上，使用這些開放資料，並找出新價值，新商機，促進商業和科技發展，是值得發展的議題。

2.3 資料探勘

資料探勘可以用來分析處理海量資料。選擇有潛在價值的資料集，讓合適的模型來處理，得到有價值的結果。在本實驗中，先將新竹縣開放房價進行資料前處理，並建立模型。使用 K-means 來為「開放房價集」和「房仲網」進行坪數和價格上的分群，觀察各群的平均價格。進一步找出房仲網是開放資料價格的幾倍，而得知購屋時的議價空間。

2.4 程式語言

在本實驗中，使用 R 語言[9]來分析新竹縣房價資料，建置分析房價模型。使用 LAMP(Linux，Apache ，MySQL，PHP)架設網站[6]，將實驗結果以網頁的方式呈現，申請網路空間，將寫好的網頁上傳的開放網路上 ，並撰寫 Shell Scripts 定期自動執行撰寫的 Python 程式，取得政府資料開放平台和「雅虎奇摩房地產網」的資料，處理成可匯入 MySQL 資料庫的檔案，上傳到網頁資料夾下，PHP 檔案[10]更新資料庫。並以網頁來呈現各縣市的房仲網和開放資料的價格比。來提供房價分析服務 。

三、研究方法

以新竹縣實價登錄資料來建立分析模型，計算數值變數與房價的相關係數，找出最相關的欄位，得到「坪數」是影響房價的條件 。透過資料探勘K-means 分群得到房仲網的平均房價高於與開放資料價格的結論 。設計實驗驗證「屋齡」是影響房價的條件。電腦根據相同的「坪數」和「屋齡」，找出實際例子。 最後定期自動將開放資料與房仲網的比較匯入資料庫裡，提供的房價分析服務 。

3.1 取得房價實價登錄資料和房仲網的資料集

近年來政府推動房價實價登錄，讓不動產交易資訊透明化。也讓這些不動產交易價格開放在網路上，讓民眾可以瀏覽房屋的成交資訊。本實驗從政府資料開放平台上取得全台實價登錄房屋資料集。為了知道與房仲網的價差，也撰寫程式爬取房仲網的賣房資訊。

Table 1

房屋資料來源來自「雅虎奇摩房地產網」上的「東森房屋」,「台灣房屋」,「台灣搜屋網」,「好房 HouseFun」銷售房屋的資訊。	解壓縮後的政府開放房價資料以全台各縣市為區隔，為 csv 檔案，在本論文簡稱為開放資料。

Illustration 5: 實價登錄房屋資料集

Illustration 6: 房仲網

資料來源：雅虎奇摩房地產網

3.2 資料前處理與觀察

3.2.1 資料前處理

「真實資料多不完整。透過資料前處理來改善。」[1]。資料前處理是資料分析中，需要花費最多時間和心力的步驟。巨量資料裡的資料結構是多樣的。也就是說，巨量資料不像過去以往的關聯式資料庫裡的資料，保證欄位裡的資料

格式是一致的，或者單純文字描述，巨量資料可能是圖像，多媒體影音，甚至可能同一欄位裡，有多種格式來描述資料。舉例來說，填寫表單的家用電話號碼欄位，可能有地區號碼，也可能沒有附上地區號碼。或者因填表人沒有室內電話，所以無法填寫，導致此欄位變成空白。因此在執行演算法之前，需要觀察資料集的特性，確定想要解決的問題為何，在將不必要的資料清除，將資料整理格式一致，將缺失的資料給予適當的表示值，以減少資料處理時，遇到不明的錯誤，而影響實驗結果的準確度。

在實驗中，下載開放資料集為 big5 編碼，而透過 Ubuntu 解壓縮後瀏覽的編碼是 utf8，會導致 R 語言無法讀取某幾個中文字，在讀檔時候發生錯誤。需要人工對此中文字修正後，才可以正常讀檔。例如，修正「六張黴段 1~50 地號」為「 六張段 1~50 地號」。「鋼腳嬰段 151~200 地號」修正為「腳段 151~200 地號」。而網站系統以 Python 來處理自動更新的資料，在讀檔時同樣地遇到某幾個無法辨識中文字而中斷執行，解決的方式是，以忽略這些 big5 無法辨識的字，繼續讀取其他資料。

```
file = open(fileName,'r',encoding='Big5hkscs',errors='ignore')
```

Illustration 7: Python 的程式碼

3.2.2 觀察

用 R 語言觀察[7]開放資料中新竹縣不動產資料集裡的價格是怎樣分佈?

```
file <-file("../Documents/R.data/RealEstate/J_lvr_land_A.CSV" , encoding
="big5")

realEstate <- read.csv(file)

x <-realEstate[,'總價元']

x2<-c("最高價"=max(x),"平均價格"=mean(x),"中間的價格"=median(x),"最低價
格"=min(x))
```

Illustration 8: R 程式碼

Table 2: 結果

最高價	平均價格	中間的價格	最低價格
一億六千萬	一千萬	768 萬	一萬元

最低價格一萬元? 這是怎樣的物件呢？最高價一億六千萬元。查看不動產交易標的是由哪些物件組成的？

```
obj <- levels(realEstate[,'交易標的'])
table(realEstate[,'交易標的'])
```

Illustration 9: R 觀察物件組成

Table 3: 結果

交易標的	數量
土地	181
房地(土地+建物)	168
房地(土地+建物)+車位	372
車位	2

此資料集包含土地或車位等物件，不過車位或土地不是本實驗想要分析的目標。本實驗的目的，是找出房屋的議價空間，給想購買居住用房屋的消費者價格上的參考。單只有土地或車位，並不符合實驗目的。所以將新竹縣不動產資料集過濾，資料集內容只包括房地(土地+建物)和房地(土地+建物)+車位 ，只有兩種交易類型物件。

```
landHouseIndex<-which(realEstate[,'交易標的'] == '房地(土地+建物)')
landHouseParkingSpaceIndex<-which(realEstate[,'交易標的'] == '房地(土地+建物)+車位')
index<- c(landHouseIndex,landHouseParkingSpaceIndex)
house<- realEstate[index,]
```

Illustration 10: R 過濾不必要的資訊

Table 4: 結果

交易標的	數量
房地(土地+建物)	168
房地(土地+建物)+車位	372

Table 5: 過濾後得到新竹縣房屋資料集，價格範圍

最高價	平均價格	中間的價格	最低價格
一億一千萬	一千萬	八百萬	二十萬

去除土地和車位的物件後，最低價由一萬元提高到二十萬元。最高價由一億六千萬元降到一億一千萬元，中間價在八百萬元。但平均價仍然在一千萬元。 二十萬到一億一千萬這個價格範圍有點過大，再觀察新竹縣房屋資料集的其他欄位，看是否能找出影響房價的關鍵特性。

3.3 相關係數

新竹縣的房屋資料集有二十六個欄位來描述房屋。

```
colnames(house)
```

Illustration 11: R 查看資料集欄位

Table 6: 房屋的欄位名稱

"鄉鎮市區"	"交易標的"	"車位總價元"	"非都市土地使用分區"	"車位移轉總面積平方公尺"
"都市土地使用分區"	"有無管理組織"	"車位類別"	"交易年月"	"土地區段位置或建物區門牌"
"交易筆棟數"	"移轉層次"	"總樓層數"	"建物型態"	"非都市土地使用編定"
"主要用途"	"主要建材"	"建築完成年月"	"單價每平方公尺"	"土地移轉總面積平方公尺"

"建物現況格局.房"	"建物現況格局.廳"	"建物現況格局.衛"	"建物現況格局.隔間"	"建物移轉總面積平方公尺"
"總價元"				

多達二十六種欄位。每個欄位的特性不同，想要找出對房價有怎樣的影響，依據變數類型，來選擇處理的方式。舉例來說，數值變數本身是數字，變數間彼此可以互相比較大小。比如說，面積平方公尺。數值變數適合與房價計算相關係數，來看此變數與房價是否有相關。為了找出哪一個欄位最能影響房價，將數值資料的欄位與房價，計算兩變數的相關係數。

```
cor(house['單價每平方公尺'],house['總價元'], use = "na.or.complete")
cor(house['建物移轉總面積平方公尺'],house['總價元'], use = "na.or.complete")
cor(house['土地移轉總面積平方公尺'],house['總價元'], use = "na.or.complete")
cor(house["交易年月"],house['總價元'], use = "na.or.complete")
cor(house["建物現況格局.房"],house['總價元'], use = "na.or.complete")
cor(house["建物現況格局.廳"],house['總價元'], use = "na.or.complete")
cor(house["建物現況格局.衛"],house['總價元'], use = "na.or.complete")
cor(house["建築完成年月"],house['總價元'], use = "na.or.complete")
cor(house["車位總價元"],house['總價元'], use = "na.or.complete")
cor(house["車位移轉總面積平方公尺"],house['總價元'], use = "na.or.complete")
```

Illustration 12: R 計算相關係數

Table 7: 與價格的相關係數

單價每平方公尺	建物移轉總面積	土地移轉總面積	交易年月	建物現況格局-房	建物現況格局-廳	建物現況格局-衛
0.52	0.84	0.39	0.03	0.10	-0.22	0.02
建築完成年月	車位移轉總面積平方公尺	車位總價元				

0.30	-0.02	0.30				

價格與數值欄位的相關係數，0 代表沒有相關，1 代表完全相關。「建物移轉總面積」的相關係數是最高 0.84，，表示最相關。所以在數值資料欄位還是以「建物移轉總面積」為影響房價的關鍵。 在實驗中處理資料時，會把

「建物移轉總面積」當作一個關鍵條件。住宅通常用「建物移轉總面積」來表示坪數，比如說，在總樓數是五層的公寓裡，二樓住屋的「建物移轉總面積」有五十坪，那麼它的「土地移轉總面積」為五十除以五為十坪。

3.4 過濾不需要的名目資料

有些變數本身是名目資料，代表只是一種記號區別，比如說房屋集的「都市土地使用分區」，分為「住」，或「商」。名目資料本身就是一種分類，並沒有因為變數的代表數值高，就表示優於數值低的變數。舉例來說，將「住」表示為 0，「商」用表示為 1，雖然 0 小於 1，並不表示「住」小於「商」，名目變數間無法藉由數值大小來作比較。在開放房屋資料集中，有些欄位是名目資料，因此無法藉由與房價計算相關係數，來決定此欄位是否會影響房價。所以在過濾欄位為名目資料時，根據研究目的，來刪減不是研究目標的物件。

Table 8:「都市土地使用分區」的名目資料

"住"	"其他"	"商"	"工"	"農"

在篩選資料時，需謹慎考慮欲分析物件為何，才能根據目的選擇合適的篩選條件。本實驗目的，為購買住宅的消費者，找出合理的房價。所以在新竹縣的房屋資料集裡，挑選「都市土地使用分區」欄位為「住」和「其他」的房屋。而商業，工業，和農業用途的房屋，則不在本實驗分析範圍裡。

在房屋資料集內過濾掉商業用途的房屋。挑選「主要用途」為「住家用」，「住商用」，「見其他登記事項」的房屋。

Table 9:「主要用途」的名目資料

"住商用"	"住家用"	"商業用"	"工業用"	"市場攤位"	"見其他登記事項"	"農舍"

挑選建物型態為「住宅大樓」，「公寓」，「其他」，「套房」，「華廈」，「透天厝」的房屋。

Table 10: 「建物型態」的名目資料

"住宅大樓(11 層含以上有電梯)"	"其他"	"套房(1 房 1 廳 1 衛)"	"店面(店鋪)"	"透天厝"
"公寓(5 樓含以下無電梯)"	"農舍"	"華廈(10 層含以下有電梯)"	"辦公商業大樓"	

在開放資料集中，「建物移轉總面積」是以「平方公尺」為單位，而房仲網的資料集則是以「坪數」表示，為了方便與房仲網的資料進行面積的比較。先將開放資料的面積，由平方公尺乘上 0.3025 轉換成坪數，加上「坪數」欄位到開放資料集裡。

本實驗分析房屋目標，是給剛入門購買房屋的消費者。先購買小坪數的住宅，等以後年收入增加，有足夠的資金和儲蓄，有能力購買更寬敞的房屋。或是家中成員增加，有購買較大坪數房屋需求，再進行小坪數房屋換購大坪數的房屋。所以本實驗分析物件設定為二十五坪以下的房屋，因此將新竹縣的房屋資料集，篩選留下坪數二十五坪以下的物件。

Table 11: 新竹縣開放資料集的價格範圍

	最高價	平均價格	中間的價格	最低價格
開放資料	388 萬	191 萬	170 萬	66 萬

最低價由二十萬提昇至六十六萬，平均價格由一千萬將到一百九十一萬，中間價格由八百萬降至一百七十萬，最高價由一億一千萬降至三百八十八萬。因為物件經過篩選後，新竹縣房屋集內的物件為二十五坪以下的住宅，價格差距變得比較集中。在進行分析時。會比較容易找出影響住屋價格的關鍵。同樣地，以 Python 語言在房仲網，取得新竹縣房屋集內為二十五坪以下的住宅資料。觀察房仲網資料集，可以看出房仲網不管是最高價，最低價，平均和中間價格都高於開放資料集。

Table 12: 房仲網新竹縣房屋資料集的價格範圍

	最高價	平均價格	中間的價格	最低價格
房仲網	3360 萬	412 萬	366 萬	68 萬

3.5 用 K-means 演算法分群房屋資料集

由相關係數得知，「建物轉移面積」是影響房價關鍵，經計算後得到「坪數」。過濾後的開放資料集為新竹縣二十五平以下的住宅。

將開放房價資料和房仲網房價資料，依照坪數和價格用 K-means 分成 3 群:「低房價群」,「中房價群」,「高房價群」。來進行分析比較。

開放房價集分成 3 群：	房仲網房價資料集分成 3 群：
「開放房價低房價群」	「房仲網低房價群」
「開放房價中房價群」	「房仲網中房價群」
「開放房價高房價群」	「房仲網高房價群」

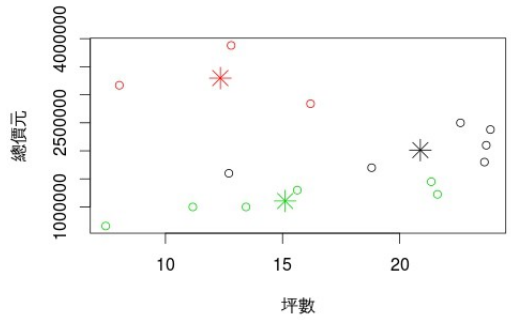
Illustration 13: 分群示意圖

為探討開放房價資料和房仲網的房價的關係，讓房屋坪數和價格條件相似的物件，分成各群來觀察。將開放房價資料和房仲網的房價資料分別作 K-means 分群 。

```
areaPrice<-cbind(livehouse['坪數'],livehouse['總價元'])
km <- kmeans(areaPrice,3)
```

Illustration 14: R K-means

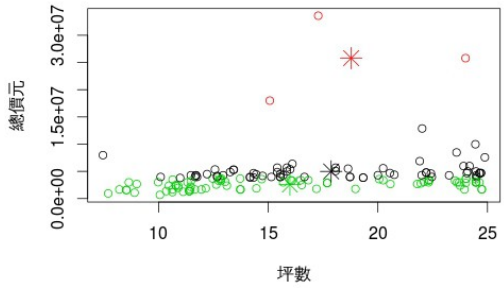
開放房價資料分群



開放房價資料各群平均價格

低房價群	中房價群	高房價群
110 萬	201 萬	330 萬

房仲網的房屋分群



房仲網各群平均價格

低房價群	中房價群	高房價群
257 萬	499 萬	2580 萬

Illustration 15: K-means 分群後的結果

可以看出房仲網的房價，不管低中高價群的平均，都大於開放房價資料。得到房仲網房價高於開放資料的結論。

3.6 「坪數」與「屋齡」分群

由圖可以看出，在相同坪數下，房仲網的物件比開放資料的價格高。

黑圈：政府開放資料集(15 個物件)， 紅圈：房仲網資料集(149 個物件)

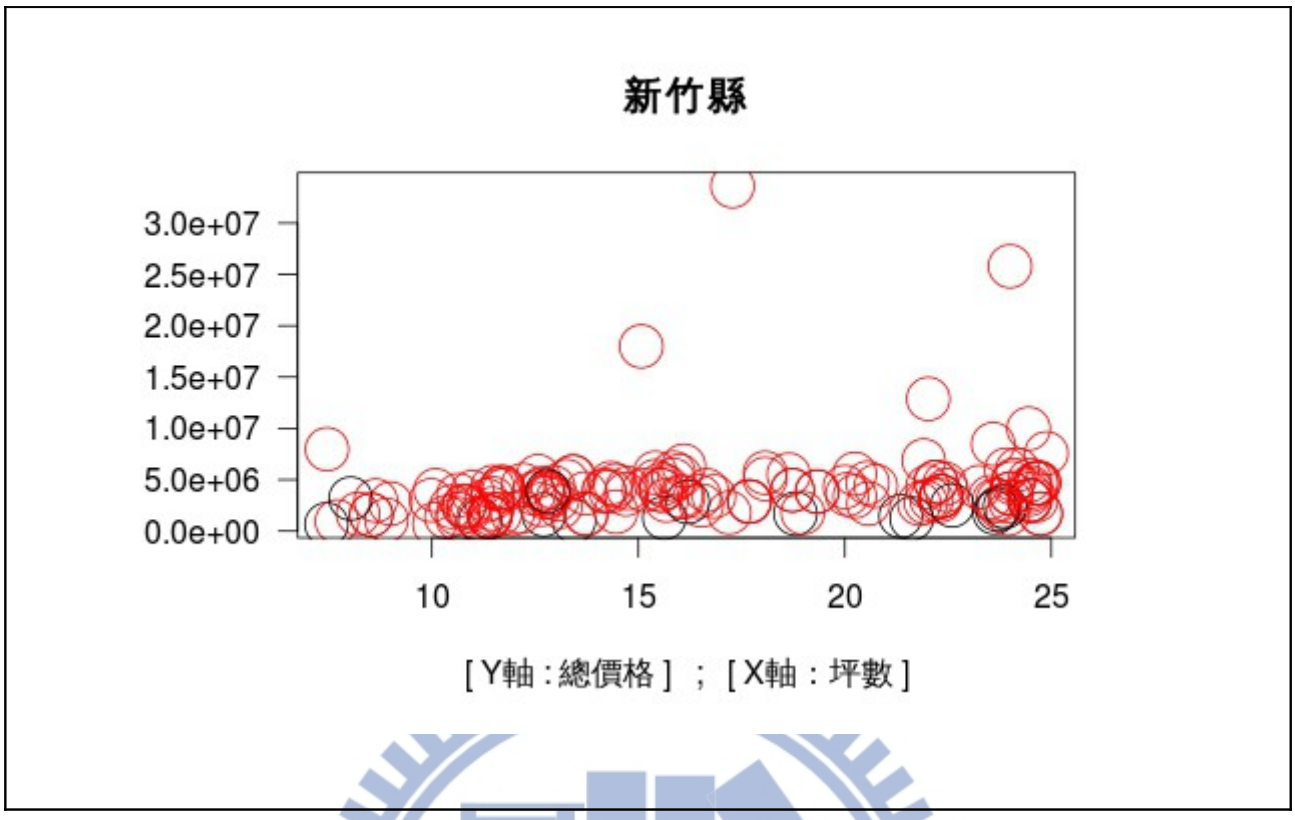


Illustration 16: 新竹縣二十五坪以下住宅的比較

新竹縣每五坪為一個區間分群，探討每個區間，房仲網定價是開放資料成交價格的幾倍。

Table 13: 坪數分群表

"城市名"	"坪數最小值"	"坪數最大值"	"房仲網平均價格"	"公開資料平均價格"	"房仲網/公開資料"
"新竹縣"	"0"	"4"	"NaN"	"NaN"	"NaN"
"新竹縣"	"5"	"9"	"2612222.22222222"	"1915000"	"1.3640847113432"
"新竹縣"	"10"	"14"	"3095636.36363636"	"1870000"	"1.65542051531356"
"新竹縣"	"15"	"19"	"5387500"	"1946666.66666667"	"2.76755136986301"
"新竹縣"	"20"	"24"	"4680000"	"1909166.66666667"	"2.45133129637713"

新竹縣沒有四坪以下的住宅，所以房仲網和開放資料的平均價格均為 NaN。

五~九坪住宅，房仲網是開放資料的 1.4 倍。

十~十四坪住宅，房仲網是開放資料的 1.7 倍。

十五～十九坪住宅，房仲網是開放資料的 2.8 倍。

二十~二十四坪住宅，房仲網是開放資料的 2.5 倍。

由分析實驗，得知「坪數」是影響價格的因素。除了「坪數」的因素外，假設「屋齡」也是影響價格的因素，爲了驗證這項假設，對「屋齡」和價格做實驗。用 R 語言計算開放資料的屋齡。開放資料集有「建築完成年月」，為了計算屋齡，將今年民國一〇三年扣掉「建築完成年」得到「屋齡」。把屋齡資料新增欄位存進開放資料集內。爬取房仲網的資料已有屋齡資訊，房仲網的屋齡不需要再經過計算。將屋齡以十年為單位，來探討新屋和中古屋的定價與成交價的價差。

紅圈：政府開放資料集(202 個物件) 黑圈：房仲網資料集(1258 個物件)

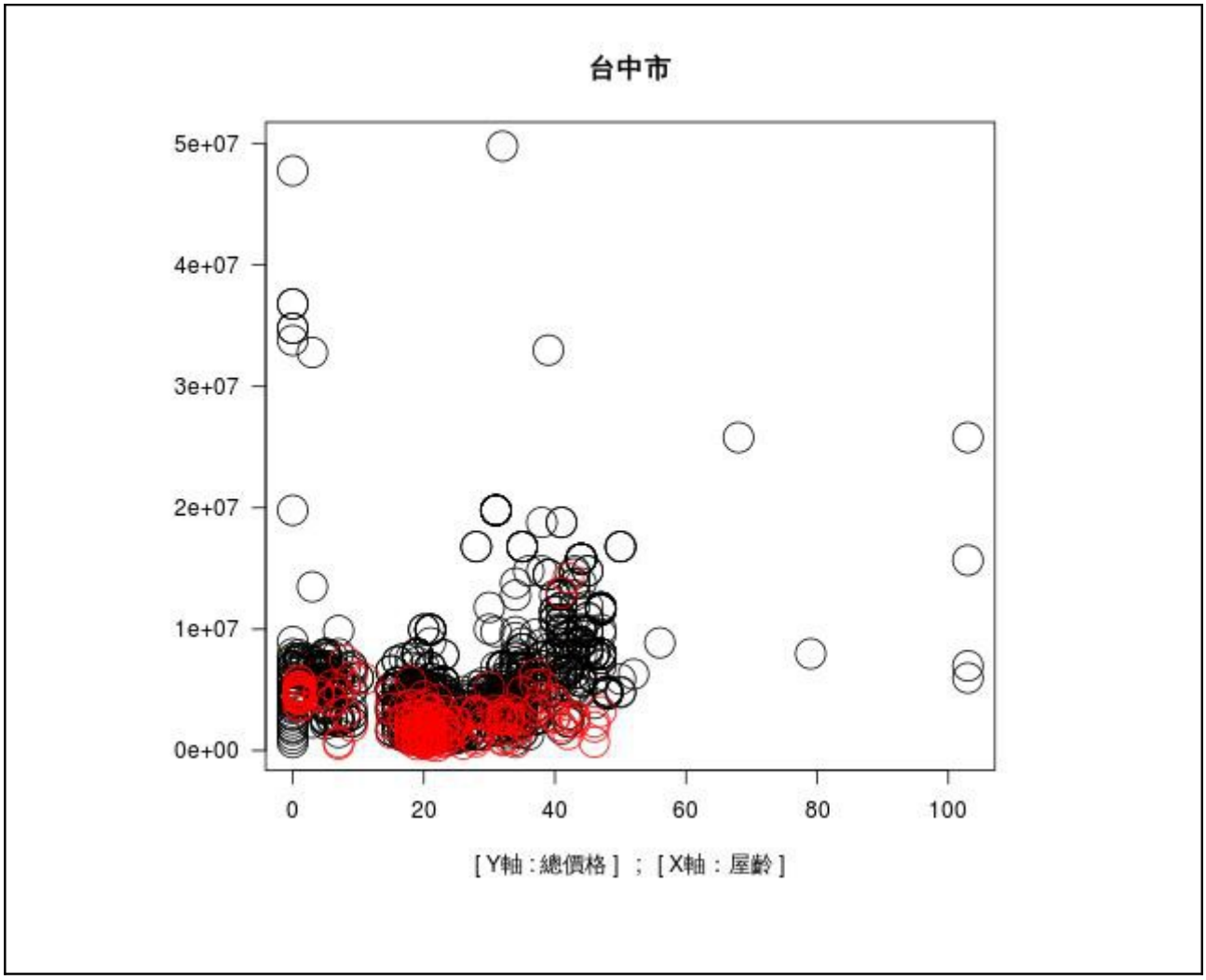


Illustration 17: 台中市二十五坪以下住宅的比較。

Table 14: 屋齡分群，房仲網與開放資料的平均價格比。

"城市名"	"屋齡最小值"	"屋齡最大值"	"房仲網平均價格"	"公開資料平均價格"	"房仲網/公開資料"
"台中市"	"0"	"9"	"6646569.76744186"	"4627368.42105263"	"1.4363606185327"
"台中市"	"10"	"19"	"3278859.64912281"	"2165857.14285714"	"1.51388546559328"
"台中市"	"20"	"29"	"3265626.53562654"	"1920970.14925373"	"1.69998817362945"
"台中市"	"30"	"39"	"5784128.113879"	"2990740.74074074"	"1.93401187708648"
"台中市"	"40"	"49"	"9077051.28205128"	"3927989.625"	"2.31086437303186"

屋齡九年以下，房仲網定價是開放資料成交價格的 1.4 倍。

屋齡十～十九年，房仲網是開放資料的平均價格 1.5 倍。

屋齡二十～二十九年，房仲網定價是開放資料成交價格的 1.7 倍。

屋齡三十～三十九年，房仲網定價是開放資料成交價格的 1.9 倍。

屋齡四十～四十九年，房仲網定價是開放資料成交價格的 2.3 倍。

屋齡小的議價空間不像中古屋來的大，中古屋比新屋有更多的議價空間。 所以「屋齡」的確是影響房價的關鍵條件之一。

3.7 實際案例

撰寫程式比對，開放資料集與房仲網資料集，找出同縣市，同屋齡，同坪數的房屋，來驗證實驗分析。以台中市為例。

Table 15: 台中市的實際案例。

"土地區段位置或建物區門牌"	"建物型態"	"建物現況格局"	"坪數"	"屋齡"	"總價元"	"資料來源"	"房仲網/公開資料"
"臺中市北區健行路401~450號"	"套房(1房1廳1衛)"	"0 房 0 廳 0 衛"	10.705475	19	1180000	"政府資料開放平台"	1
"台中市北區健行路"	"大樓"	"1 房 0 廳 1 衛"	10.5	19	1680000	"房仲網"	1.4237288136
"臺中市南屯區東興西街1~50號"	"套房(1房1廳1衛)"	"0 房 0 廳 0 衛"	17.935225	23	1750000	"政府資料開放平台"	1
"台中市南屯區東興西街"	"公寓"	"1 房 0 廳 1 衛"	18.23	23	2980000	"房仲網"	1.7028571429
"臺中市西屯區至善路101~150號"	"住宅大樓(11層含以上有電梯)"	"2 房 0 廳 2 衛"	14.45345	21	1710000	"政府資料開放平台"	1
"台中市西屯區至善路"	"大樓"	"2 房 0 廳 2 衛"	14.43	21	2980000	"房仲網"	1.7426900585
"台中市西屯區至善路"	"大樓"	"2 房 0 廳 2 衛"	14.45	21	2980000	"房仲網"	1.7426900585
"台中市西屯區至善路"	"大樓"	"2 房 0 廳 2 衛"	14.45	21	2980000	"房仲網"	1.7426900585
"臺中市西區民生路151~200號"	"住宅大樓(11層含以上有電梯)"	"1 房 1 廳 2 衛"	23.567775	25	2400000	"政府資料開放平台"	1
"台中市西區民生路"	"公寓"	"1 房 1 廳 1 衛"	23.45	25	3880000	"房仲網"	1.6166666667
"台中市西區民生路"	"透天"	"1 房 1 廳 1 衛"	23.45	25	3880000	"房仲網"	1.6166666667
"臺中市南區學府路146巷1~50號"	"住宅大樓(11層含以上有電梯)"	"1 房 1 廳 1 衛"	10.50885	18	1200000	"政府資料開放平台"	1
"台中市南區學府路"	"透天"	"1 房 0 廳 1 衛"	10.13	18	1780000	"房仲網"	1.4833333333
"臺中市西區均安街51~100號"	"公寓(5樓含以下無電梯)"	"1 房 0 廳 1 衛"	7.8771	28	740000	"政府資料開放平台"	1
"台中市西區均安街"	"公寓"	"1 房 1 廳 1 衛"	7.88	28	1250000	"房仲網"	1.6891891892
"台中市西區均安街"	"透天"	"1 房 1 廳 1 衛"	7.87	28	1280000	"房仲網"	1.7297297297

到目前為止，完成分析建模的工作，接著為將實驗結果呈現，而進行架

設網站的工作。

3.8 建置定時自動更新房價分析系統

實驗找出實際案例，並確認房仲網的價格的確比開放資料來得高。對購屋者而言，兩者價格比值是有參考價值的，因此以網頁的方式呈現兩者比值。

架設網站，撰寫程式定時自動從政府資料開放平台和房仲網取得最新資料，經過 Python 處理成可匯進 MySQL 的資料庫裡的檔案格式，shell 會上傳這些檔案到網頁資料夾下，並定期自動執行 update.php 來更新資料庫裡的資料。前端以網頁呈現同縣市下，同坪數同屋齡的房屋價格比值和資訊。

在首頁選擇縣市，按下查詢。 由 ShowBoth.php 比對兩資料集，找出同坪數同屋齡的房屋的價格比。同時也顯示比值的平均價格和標準差。

設定 Ubuntu 定時執行 shell 檔案，更新資料庫。

1. 處理政府開放平台的實價登錄資料的流程：
   (1) 自動到平台下載檔案。
   (2) 將檔案解壓縮。
   (3) Python 讀檔處理成資料庫可讀的 csv 檔。
   (4) 上傳 csv 檔案到本機端 www 資料夾下和開放網路空間資料夾下。
   (5) 執行 PHP 讀 csv 檔案，更新資料庫裡的開放資料。
2. 處理房仲網資料的流程：
   (1) Python 爬取房仲網的網頁，將資料處理成資料庫可讀的 csv 檔。
   (2)上傳 csv 檔案到本機端 www 資料夾下和開放網路空間資料夾下。
   (3) 執行 PHP 讀 csv 檔案，更新資料庫裡的房仲網資料。

Table 16: 放程式碼的 Ubuntu 路徑表

處理爬文房仲網 /usr/lib/python3.2/ParserWebsite/
本機端網頁資料夾 /var/www/HouseAnalyse/
開放網頁資料夾 ftp://bigdata.net16.net/public_html/

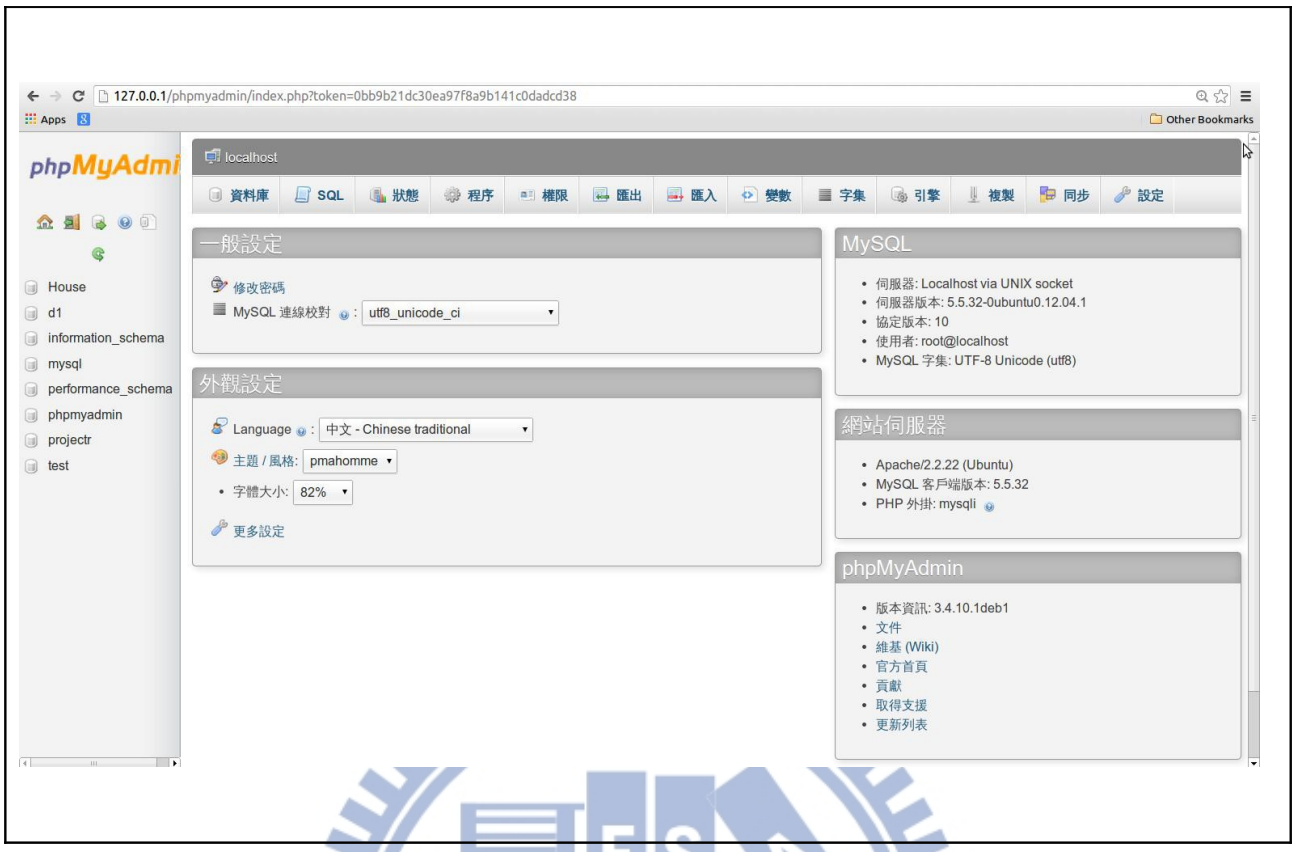


Illustration 18: 用 phpMyAdmin 操作 MySQL 資料庫。

在本機端安裝開發環境後，在 MySQL 資料庫裡，建立 House 資料庫，兩個資料表分別為，房仲網資料表，開放資料表。每個資料表有七個欄位。分別是「土地區段位置或建物區門牌」，「建物型態」，「建物現況格局」，「坪數」，「屋齡」，「總價元」，「資料來源」欄位。

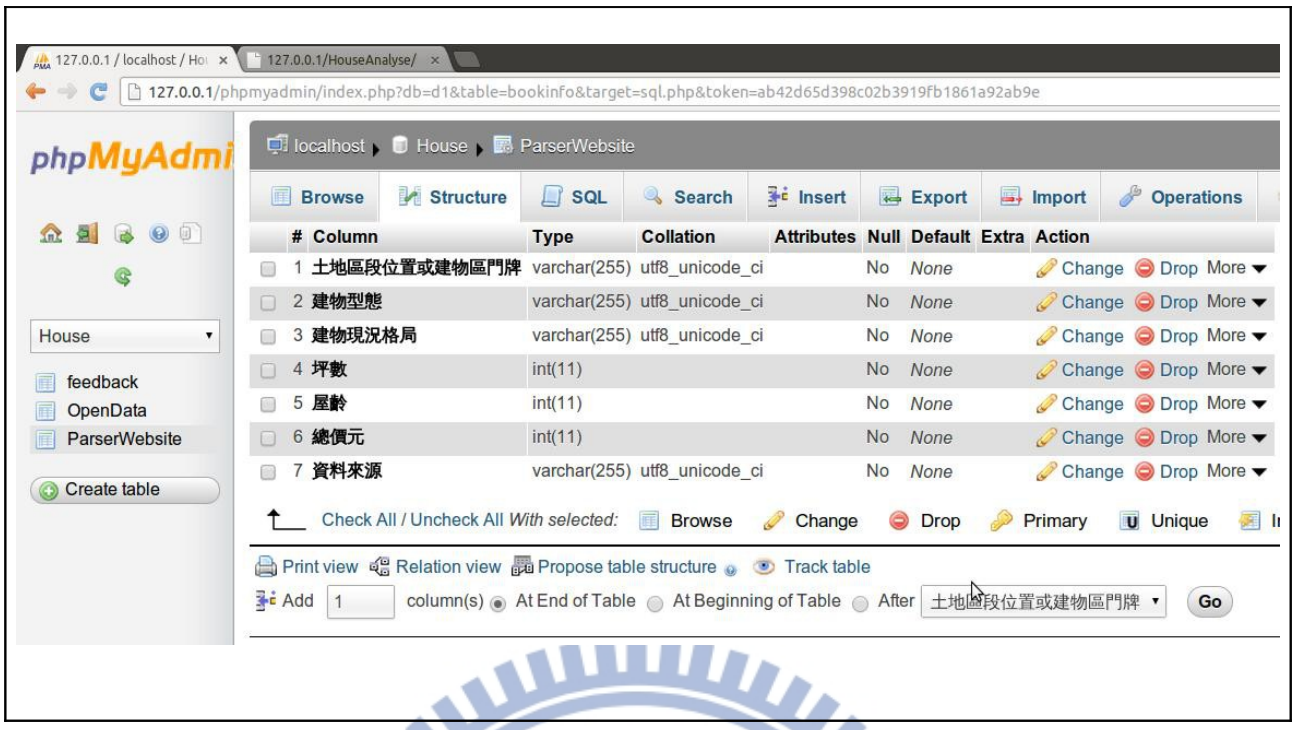


Illustration 19: 資料表的七個欄位

3.8.1 Python 處理「實價登錄資料」的流程

Python 從「政府開放平台」下載的「實價登錄資料」，處理成可匯入資料庫的開放資料表，有七個欄位資料的 csv 檔案。程式碼在以下路徑 /usr/lib/python3.2/ParserWebsite/processOpenData.py。選擇台灣北中南東代表縣市來做房價分析，讀台北市，新竹市，台中市，高雄市，宜蘭縣的實價登錄資料。開放資料集原本就有「土地區段位置或建物區門牌」，「建物型態」和「總價元」欄位，直接使用即可。需要經過計算後的欄位有：

(1)「建物現況格局」：將「建物現況格局.房」，「建物現況格局.廳」，「建物現況格局.衛」欄位的值，合併為表示幾房幾廳幾衛的字串。

```
temp = line.split(",")
#建物現況格局 QRS 17~19
# 1 房 1 廳 1 衛
status =temp[16] +"房" + temp[17] + "廳"+ temp[18] +"衛"
```

Illustration 20: Python 程式碼

(2)「坪數」：1 平方公尺= 0.3025 坪。將「建物轉移面積」做轉換。

(3)「屋齡」：取得目前的西元年-1911，得到民國年。

```
thisYear = datetime.date.today().timetuple().tm_year -1911
```

Illustration 21: Python 程式碼

將目前民國年-建築完成年，得到屋齡。

```
age = thisYear -math.ceil( int(age)/10000 )
```

Illustration 22: Python 程式碼

(4)「資料來源」：加上「政府資料開放平台」字串。

把七個欄位「土地區段位置或建物區門牌」,「建物型態」,「建物現況格局」,「坪數」,「屋齡」,「總價元」,「資料來源」用逗號為分隔存成存成一行字串，處理過後的字串放進總內容裡。

```
newRow = address + "," + buildingType + "," + status + "," + str(area) + "," +
str(age) + "," + str(price) + "," + "政府資料開放平台,\n"
 newContent = newContent + newRow
```

Illustration 23: Python 程式碼

最後存檔成'OpenData.csv'。

```
file = open('OpenData.csv','w',encoding = 'UTF-8')
file.write(newContent)
file.close()
```

Illustration 24: Python 程式碼

3.8.2 Python 爬取雅虎奇摩房地產網頁的流程

程式碼放在路徑 /usr/lib/python3.2/ParserWebsite/parserWebsite.py 底下。

(1) 定義找房屋總比數的 class ，以網頁所定義的 tag 和 attrs 來找尋。

```
class houseNumParser(HTMLParser):
        if tag == 'div' and attrs == [('class','yui3-u-1-3 summary')]:
```

Illustration 25: Python 程式碼

(2) 爬文取得房仲網的房屋總筆數。

```
webOjbNumString = webOjbNumParser.metadata[0]
startIndex = webOjbNumString.index('共')
endIndex = webOjbNumString.index('筆')
webOjbTotalNum =
float( webOjbNumString[startIndex+1:endIndex] )
```

Illustration 26: Python 程式碼

(3) 計算總頁數，房仲網每頁顯示 10 筆，總頁數為總筆數除以十取整數上限。

```
webOjbNum = math.ceil(webOjbTotalNum/10)
```

Illustration 27: Python 程式碼

(4) 得到爬取目標網頁的網址， 將頁數依序黏貼到缺少頁數的網址後。

```
#string='https://...&page='
#url='https://...&page=1'
    for p in range(webOjbNum):
        p+=1
        url=string+str(p)
        webUrl.append(url)
```

Illustration 28: Python 程式碼

(5) 定義找房屋資訊的 class ，以網頁的 tag 和 attrs 。

```
class myparser(HTMLParser):
        if tag == 'div' and attrs == [('class','yui3-g info-detail')]：
```

Illustration 29: Python 程式碼

(6) 由爬取目標的網址，取得網頁內容

```
for webUrlStr in webUrl:
        data = urllib.request.urlopen(webUrlStr)
        contentList.append(data.read().decode('utf_8'))
```

Illustration 30: Python 程式碼

(7) 從網頁內容取出房屋資訊

```
for content in contentList:
    Parser.feed(content)
```

Illustration 31: Python 程式碼

(8) 從房屋資訊取想要的資訊

```
for string in Parser.metadata:
    address = string[3:]+','
```

Illustration 32: Python 程式碼

(9) 將房屋資訊以「土地區段位置或建物區門牌」,「建物型態」,「建物現況格局」,「坪數」,「屋齡」,「總價元」,「資料來源」存 csv 檔。

```
file = open(fileName,'a',encoding='UTF-8')
oneRow = address + buildingType +status + area + age + price +
source
file.write(oneRow)
```

Illustration 33: Python 程式碼

3.8.3 將 csv 檔匯進 MySQL 的資料庫

(1) Python 處理房仲網資料集以及開放資料後，分別存.CSV 檔，各上傳到網路空間和本機端資料夾下。

(2) 更新資料庫的 updateDB.php，

本機端的程式碼在 /var/www/HouseAnalyse/updateDB.php，

開放網路的程式碼在

ftp://bigdata.net16.net/public_html/updateDB.php 。

(3) 將更新目標設定為資料庫裡的房仲網和開放資料表

```
$targetArray = array('ParserWebsite','OpenData');
$target = $targetArray[$i];
```

Illustration 34: PHP 程式碼

(4) 清空舊的 TABLE 指令

```
$sql = "delete from " . $target;
```

Illustration 35: PHP 程式碼

(5) 讀取 CSV 檔

```
$fhandle = fopen($csvFile, "r")
$rowData = fgetcsv($fhandle,0)
```

Illustration 36: PHP 程式碼

(6) 連結資料庫，設定 UTF-8 字集，來處理中文字。

```
$link =mysql_connect(資料庫網址,帳號,密碼);
mysql_select_db("House");
mysql_query('SET CHARACTER SET UTF8;');
mysql_query("SET NAMES 'utf8'");
```

Illustration 37: PHP 程式碼

(7) MySQL 指令更新資料庫的表格

```
$sql = "INSERT INTO ". $target ." (土地區段位置或建物區門牌, 建物型態, 建物現況格局, 坪數, 屋齡,總價元,資料來源) VALUES " . $temp_sql;
$result=mysql_query($sql,$link);
```

Illustration 38: PHP 程式碼

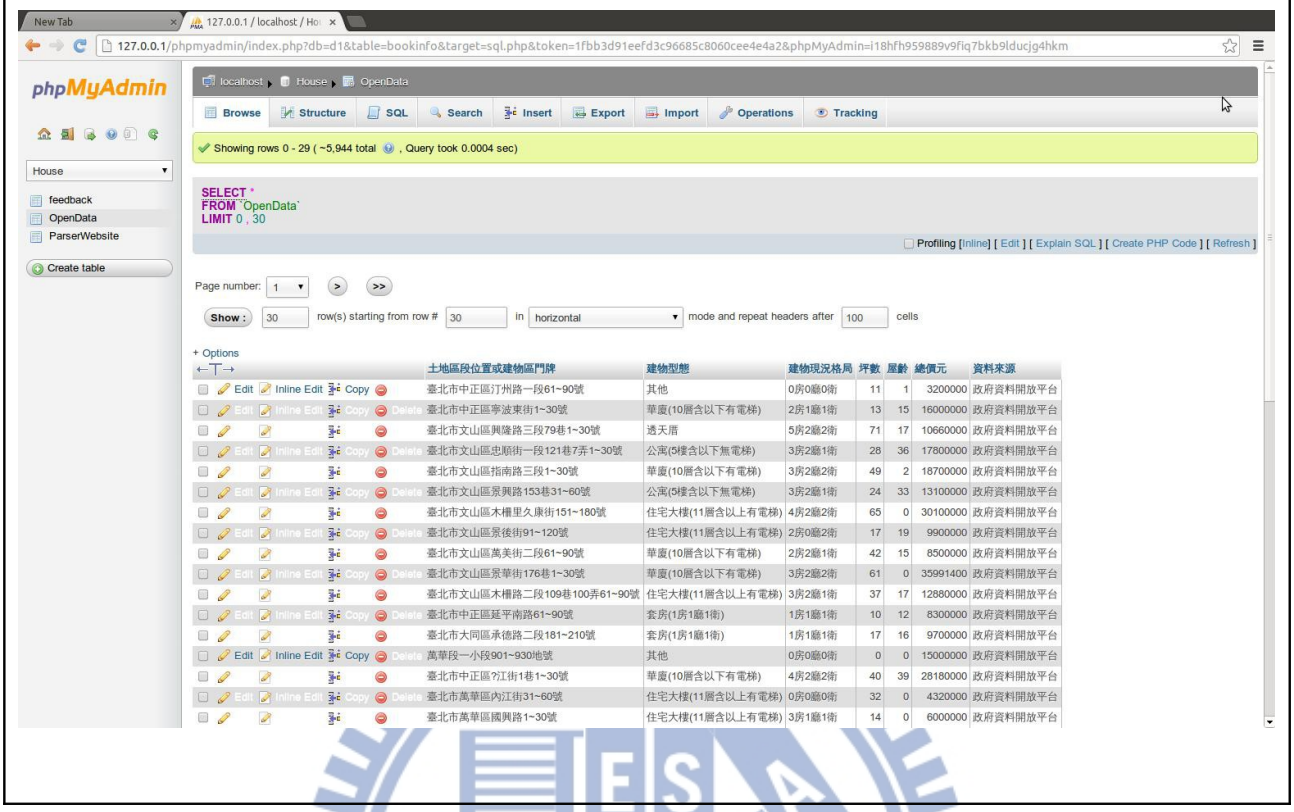


Illustration 39: 資料庫更新資料後的截圖

3.8.4 資料的總筆數

一筆資料代表一個房屋物件，有地址，價格，坪數，屋齡等資訊。開放資料有 4806 筆，房仲網有 5609 筆，兩者加起來共一萬多筆的資料。

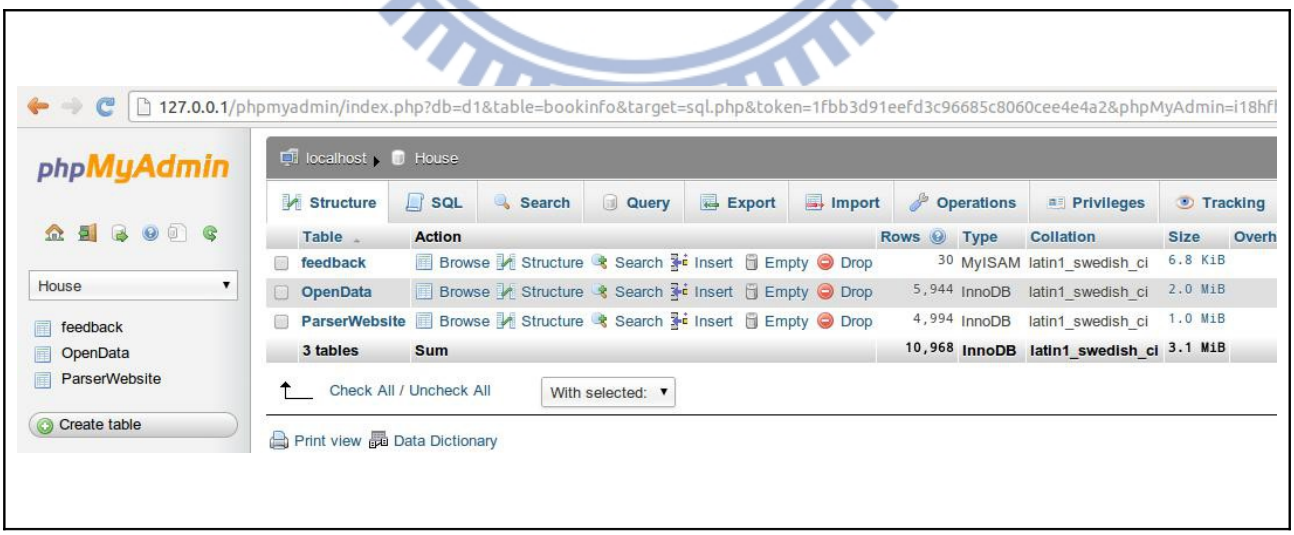


Illustration 40

3.8.5 網站首頁

顯示各縣市的選項，讓使用者選擇。

```
<html>
<link rel="stylesheet" href="style.css"/>
<h1>房價分析</h1>
<body>
<form action="ShowBoth.php" method="post">
<select name="cityName" size ="1">
 <option value="北市信義區">台北市信義區</option>
 <option value="北市大安區">台北市大安區</option>
 <option value="北市萬華區">台北市萬華區</option>
 <option value="新竹市">新竹市</option>
 <option value="中市">台中市</option>
 <option value="高雄市">高雄市</option>
 <option value="宜蘭縣">宜蘭縣</option>
</select>
<tr><input type="submit" value="查詢"></tr>
</body>
</html>
```

Illustration 41: HTML 程式碼

下拉式選單，選擇縣市送出，SQL 指令從資料庫取出房仲網和開放資料。

Illustration 42: 首頁截圖。

3.8.6 顯示查詢的結果

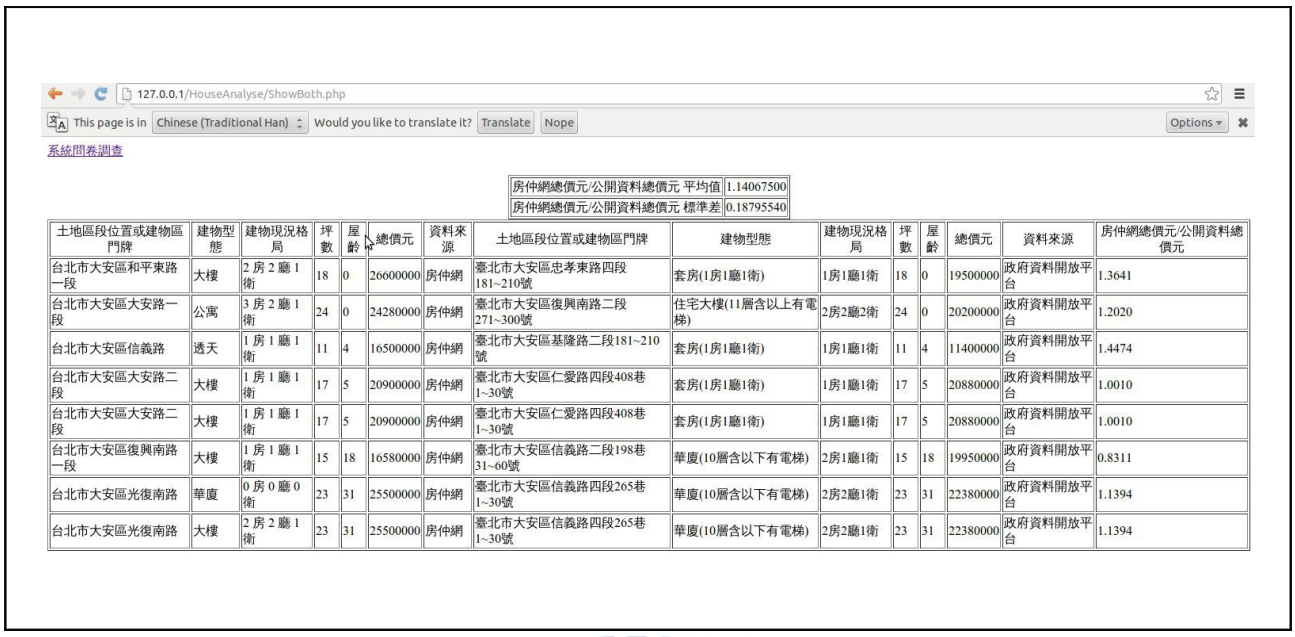
檔名為 ShowBoth.php，比較兩資料集，取同縣市同坪數同屋齡的物件，並將兩者的總價相除，得到比值。

```
SELECT P.*, O.*,P.總價元 / O.總價元
        FROM OpenData AS O, ParserWebsite AS P
        WHERE O.屋齡 = P.屋齡 AND O.坪數 = P.坪數 AND (O.土地區段位置或建物區門牌 LIKE '%".$cityName."%' AND P.土地區段位置或建物區門牌 LIKE '%".$cityName."%')
```

Illustration 43: MySQL 指令

```
#---平均值--標準差---SQL 指令
SELECT AVG(P.總價元 / O.總價元) ,STD(P.總價元 / O.總價元)
        FROM OpenData AS O, ParserWebsite AS P
        WHERE O.屋齡 = P.屋齡 AND O.坪數 = P.坪數 AND (O.土地區段位置或建物區門牌 LIKE '%".$cityName."%' AND P.土地區段位置或建物區門牌 LIKE '%".$cityName."%')
```

Illustration 44: 比值的平均值和標準差



土地區段位置或建物區門牌	建物型態	建物現況格局	坪數	屋齡	總價元	資料來源	土地區段位置或建物區門牌	建物型態	建物現況格局	坪數	屋齡	總價元	資料來源	房仲網總價元/公開資料總價元
台北市大安區和平東路一段	大樓	2 房 2 廳 1 衛	18	0	26600000	房仲網	臺北市大安區忠孝東路四段181~210號	套房(1房1廳1衛)	1房1廳1衛	18	0	19500000	政府資料開放平台	1.3641
台北市大安區大安路一段	公寓	3 房 2 廳 1 衛	24	0	24280000	房仲網	臺北市大安區復興南路二段271~300號	住宅大樓(11層含以上有電梯)	2房2廳2衛	24	0	20200000	政府資料開放平台	1.2020
台北市大安區信義路	透天	1 房 1 廳 1 衛	11	4	16500000	房仲網	臺北市大安區基隆路二段181~210號	套房(1房1廳1衛)	1房1廳1衛	11	4	11400000	政府資料開放平台	1.4474
台北市大安區大安路二段	大樓	1 房 1 廳 1 衛	17	5	20900000	房仲網	臺北市大安區仁愛路四段408巷1~30號	套房(1房1廳1衛)	1房1廳1衛	17	5	20880000	政府資料開放平台	1.0010
台北市大安區大安路二段	大樓	1 房 1 廳 1 衛	17	5	20900000	房仲網	臺北市大安區仁愛路四段408巷1~30號	套房(1房1廳1衛)	1房1廳1衛	17	5	20880000	政府資料開放平台	1.0010
台北市大安區復興南路一段	大樓	1 房 1 廳 1 衛	15	18	16580000	房仲網	臺北市大安區信義路二段198巷31~60號	華廈(10層含以下有電梯)	2房1廳1衛	15	18	19950000	政府資料開放平台	0.8311
台北市大安區光復南路	華廈	0 房 0 廳 0 衛	23	31	25500000	房仲網	臺北市大安區信義路四段265巷1~30號	華廈(10層含以下有電梯)	2房2廳1衛	23	31	22380000	政府資料開放平台	1.1394
台北市大安區光復南路	大樓	2 房 2 廳 1 衛	23	31	25500000	房仲網	臺北市大安區信義路四段265巷1~30號	華廈(10層含以下有電梯)	2房2廳1衛	23	31	22380000	政府資料開放平台	1.1394

Illustration 45: 呈現的網頁畫面

3.8.7 申請網站空間

在 000webhost.com 免費網站空間，申請帳號。透過電子郵件信箱啟動此帳號。申請網址名稱 www.bigdata.net16.net。將網頁使用 FTP 上傳。設定 MySQL 的帳號密碼後，登入 phpMyAdmin 頁面，匯入資料庫的資料表，完成後端資料庫的建置。之後透過 script shell 連上 FTP 上傳 csv 檔案，定期自動更新資料庫。

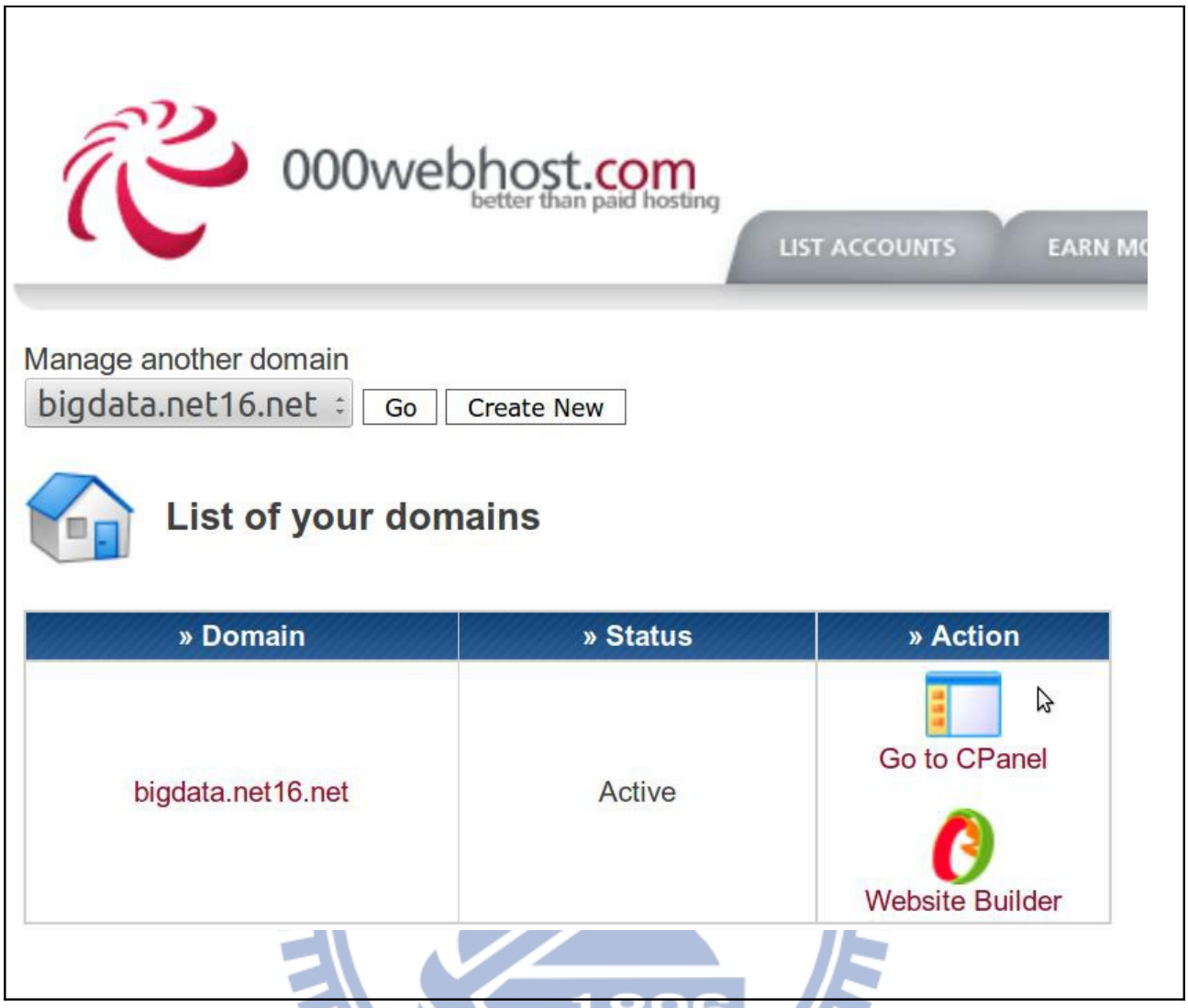


Illustration 46: 申請的網站空間操作介面

使用者用瀏覽器連上網址，即可在開放網路上，使用房價分析的服務。

www.bigdata.net16.net/ShowBoth.php

系統問卷調查

房仲網總價元/公開資料總價元 平均值	2.74600000
房仲網總價元/公開資料總價元 標準差	2.19734917

土地區段位置或建物區門牌	建物型態	建物現況格局	坪數	屋齡	總價元	資料來源	土地區段位置或建物區門牌	建物型態	建物現況格局	坪數	屋齡	總價元	資料來源	房仲網總價元/公開資料總價元
台北市信義區光復南路	公寓	0 房 1 廳 1 衛	13	6	25000000	房仲網	臺北市信義區忠孝東路五段781~810號	住宅大樓(11層含以上有電梯)	1房1廳1衛	13	6	10000000	政府資料開放平台	2.5000
台北市信義區信義路五段	大樓	1 房 1 廳 1 衛	12	8	16880000	房仲網	臺北市信義區忠孝東路五段451~480號	其他	0房0廳0衛	12	8	2750000	政府資料開放平台	6.1382
台北市信義區信義路五段	大樓	1 房 1 廳 1 衛	12	8	16880000	房仲網	臺北市信義區忠孝東路五段451~480號	其他	0房0廳0衛	12	8	2750000	政府資料開放平台	6.1382
台北市信義區忠孝東路五段	大樓	1 房 1 廳 1 衛	13	9	23880000	房仲網	臺北市信義區信義路四段395巷6弄1~30號	住宅大樓(11層含以上有電梯)	0房0廳2衛	13	9	20000000	政府資料開放平台	1.1940
台北市信義區忠孝東路五段	大樓	1 房 1 廳 1 衛	13	9	20500000	房仲網	臺北市信義區信義路四段395巷6弄1~30號	住宅大樓(11層含以上有電梯)	0房0廳2衛	13	9	20000000	政府資料開放平台	1.0250
台北市信義區忠孝東路五段	大樓	1 房 1 廳 1 衛	19	20	18000000	房仲網	臺北市信義區忠孝東路五段481~510號	辦公商業大樓	0房0廳0衛	19	20	14780000	政府資料開放平台	1.2179
台北市信義區信義路五段	大樓	1 房 0 廳 1 衛	12	20	9280000	房仲網	臺北市信義區松山路421~450號	住宅大樓(11層含以上有電梯)	1房1廳1衛	12	20	9200000	政府資料開放平台	1.0087

Illustration 47: 申請的網址查詢畫面

3.8.8 自動更新的 Shell

程式碼在路徑/usr/lib/python3.2/ParserWebsite/autoParserWeb_cloud.sh 更新開放資料的步驟：

(1) 刪除舊的開放資料檔案，新建資料夾，進入此資料夾下。

```
#!/bin/bash
sudo rm -r OpenData
sudo mkdir OpenData
cd OpenData/
```

Illustration 48: Shell Script

(2) 到政府資料開放平台下載壓縮檔案 lvr_landAcsv.zip 。

```
sudo wget -c -O
/usr/lib/python3.2/ParserWebsite/OpenData/lvr_landAcsv.zip
"http://data.gov.tw/iisi/logaccess?dataUrl=http%3A%2F
%2Flvr.land.moi.gov.tw%2Fopendata
%2Flvr_landAcsv.zip&type=CSV&nid=6213"
```

Illustration 49: Shell Script

(3) 將檔案解壓縮 ，離開資料夾回上一層。

```
sudo unzip lvr_landAcsv.zip
cd ..
```

Illustration 50: Shell Script

(4) 刪除舊檔案，產生新檔案。

```
sudo rm  OpenData.csv
sudo python3  processOpenData.py
```

Illustration 51: Shell Script

(5) 刪除本機端舊的檔案，將新檔案上傳到本機端 www 的資料夾裡。

```
sudo rm  /var/www/HouseAnalyse/OpenData.csv
sudo cp  OpenData.csv /var/www/HouseAnalyse
```

Illustration 52: Shell Script

(6) 將新檔案用 ftp 上傳開放網路的資料夾裡。

```
sudo curl -u 帳號:密碼  --upload-file  OpenData.csv
ftp://bigdata.net16.net/public_html/
```

Illustration 53: Shell Script

更新房仲網檔案的步驟：

(7) 刪除舊的房仲網檔案。

```
sudo rm  ParserWebsite.csv
```

Illustration 54: Shell Script

(8)爬取房仲網資料，產生可匯入資料庫的房仲網檔案。

```
sudo python3  parserWebsite.py
```

Illustration 55: Shell Script

(9) 刪除本機端舊的檔案，將新檔案上傳到本機端 www 的資料夾裡。

```
sudo rm  /var/www/HouseAnalyse/ParserWebsite.csv
sudo cp  ParserWebsite.csv /var/www/HouseAnalyse
```

Illustration 56: Shell Script

(10) 將新檔案用 ftp 上傳開放網路的資料夾裡。

```
sudo curl -u 帳號:密碼 --upload-file  ParserWebsite.csv
ftp://bigdata.net16.net/public_html/
```

Illustration 57: Shell Script

(11) 執行本機端更新 php 檔案，更新本機端資料庫。

```
sudo wget "http://127.0.0.1/HouseAnalyse/updateDB.php"
```

Illustration 58: Shell Script

(12) 執行開放網路更新 php 檔案，更新開放網路資料庫。

```
sudo wget "http://www.bigdata.net16.net/updateDB.php"
```

Illustration 59: Shell Script

3.8.9 使用 crontab 指令，設定更新的日期

因為政府開放平台是每月 1 號和 16 號更新實價登錄資料 ，所以設定 crontab 指令，在每月 2 號和 17 號凌晨三點自動取得最新檔案並更新資料庫。

sudo vim /etc/crontab

```
0 3   2 * *  root  cd /usr/lib/python3.2/ParserWebsite/ && sh
/usr/lib/python3.2/ParserWebsite/autoParserWeb_cloud.sh


0 3   17 * *  root  cd /usr/lib/python3.2/ParserWebsite/ && sh
/usr/lib/python3.2/ParserWebsite/autoParserWeb_cloud.sh
```

Illustration 60: Shell Script

設定完畢，將 cron 服務重開，更新剛編輯的 crontab 檔案。

```
cd /etc/init.d
sudo service cron restart
```

Illustration 61: Shell Script

完成設定資料庫定期自動更新的流程。

3.8.10 申請 Google Analytics

在 Google Analytics 申請網頁服務 ，網址為 http://www.google.com/intl/zh-TW/analytics/ ，來觀察瀏覽者的瀏覽行為。

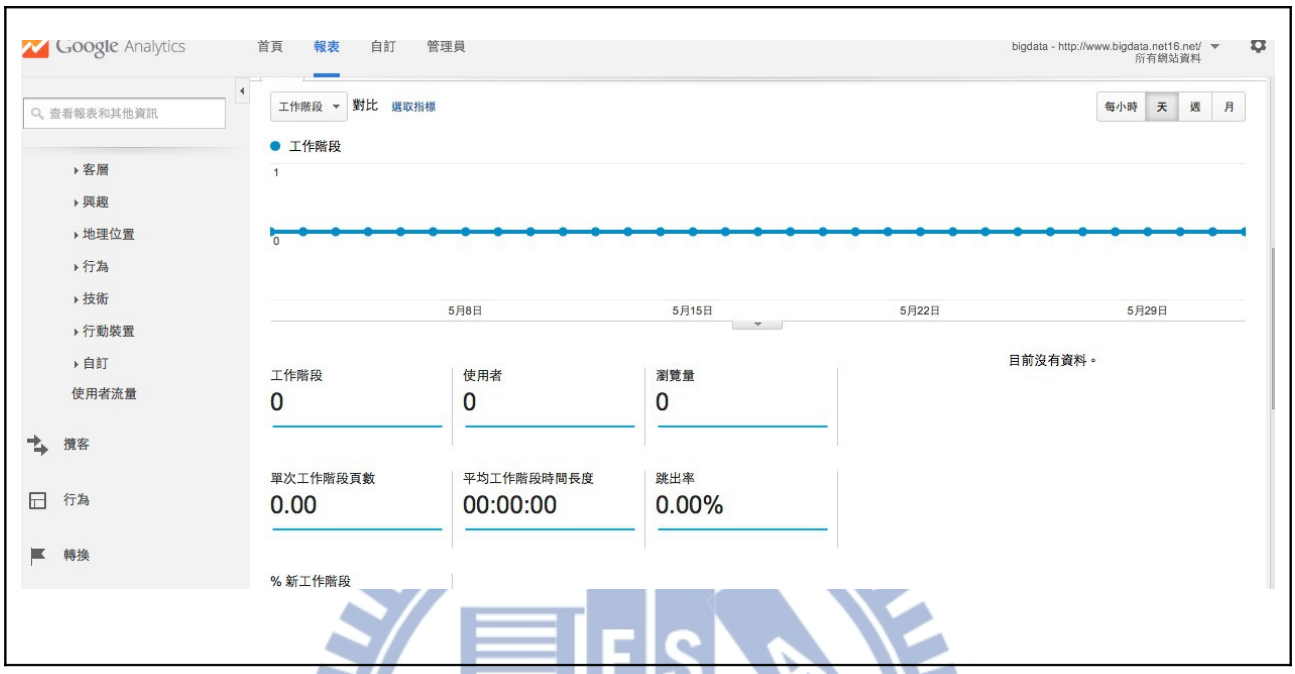


Illustration 62: 剛申請 Google Analytics 完的畫面

四、實驗結果

在政府資料開放平台取得的實價登錄資料，以新竹縣的實價登錄資料來建立分析模型。 將不需要的資訊去除，比如說，純粹只是車位的物件，或者是用途是辦公大樓的房屋。因為本實驗主要是研究購買住宅的物件，所以將非住宅用的不動產刪除。計算變數間的相關係數，得「坪數」最影響房價的條件。接著使用資料探勘的 K-means 方法，根據坪數和價格來分群這些住屋，得到低房價群，中價位群，高價位群的平均價格。發現房仲網的各群平均價格都比開放資料來的高。爲了了解房仲網的價格與開放資料的價差，進一步以每五坪「坪數」因素來分群房價 。假設「屋齡」會影響房價，並以每十年「屋齡」來分群，最後實驗發現中古屋比新屋有更大的議價空間。 用電腦比對找出兩資料集內，條件類似的實際物件來驗證開放資料集和房仲網資料集在價格上的差異。定期自動化取得房仲網和公開資料，以網頁的方式呈現各縣市房價實際例子，讓消費者可以透過本服務，找出最新的房仲網和開放資料的價差。在網站上加入 Google Analytics 來收集使用者的瀏覽行為 。最後，對使用者做系統使用後的問卷調查，觀察實際上，消費者的購屋意願和本服務的實用性 。

4.1 房價分析服務網站首頁

Illustration 63: 網站首頁

4.2 查詢顯示頁面

房仲網總價元/公開資料總價元 平均值	1.75101039
房仲網總價元/公開資料總價元 標準差	0.98992207

土地區段位置或建物區門牌	建物型態	建物現況格局	坪數	屋齡	總價元	資料來源	土地區段位置或建物區門牌	建物型態	建物現況格局	坪數	屋齡	總價元	資料來源	房仲網總價元/公開資料總價元
台中市西區均安街	公寓	1 房 0 廳 1 衛	9	0	1680000	房仲網	臺中市南區工學一街151~180號	其他	0房0廳0衛	9	0	740000	政府資料開放平台	2.2703
台中市西區均安街	公寓	1 房 0 廳 1 衛	9	0	1680000	房仲網	臺中市南區工學一街151~180號	其他	0房0廳0衛	9	0	670000	政府資料開放平台	2.5075
台中市豐原區豐勢路二段	透天	3 房 2 廳 1 衛	22	3	6980000	房仲網	臺中市西屯區西屯路二段241~270號	住宅大樓(11層含以上有電梯)	1房1廳1衛	22	3	5780000	政府資料開放平台	1.2076
台中市西屯區文心路三段	大樓	1 房 1 廳 1 衛	18	5	6580000	房仲網	臺中市西屯區太原路一段181~210號	華廈(10層含以下有電梯)	1房1廳1衛	18	5	4050000	政府資料開放平台	1.6247
台中市北區五常街	大樓	1 房 1 廳 1 衛	9	5	2200000	房仲網	臺中市北區五常街361~390號	套房(1房1廳1衛)	1房1廳1衛	9	5	2100000	政府資料開放平台	1.0476
台中市潭子區中山路一段	大樓	2 房 1 廳 1 衛	22	6	4680000	房仲網	臺中市南屯區永春東路991~1020號	住宅大樓(11層含以上有電梯)	2房2廳2衛	22	6	4600000	政府資料開放平台	1.0174
台中市西屯區文華路	大樓	1 房 1 廳 1 衛	22	6	6280000	房仲網	臺中市南屯區永春東路991~1020號	住宅大樓(11層含以上有電梯)	2房2廳2衛	22	6	4600000	政府資料開放平台	1.3652
台中市西屯區文華路	大樓	1 房 1 廳 1 衛	22	6	5980000	房仲網	臺中市南屯區永春東路991~1020號	住宅大樓(11層含以上有電梯)	2房2廳2衛	22	6	4600000	政府資料開放平台	1.3000
台中市西屯區文華路	華廈	1 房 1 廳 1 衛	22	6	5980000	房仲網	臺中市南屯區永春東路991~1020號	住宅大樓(11層含以上有電梯)	2房2廳2衛	22	6	4600000	政府資料開放平台	1.3000
台中市西屯區國安一路	大樓	0 房 0 廳 0 衛	22	6	5580000	房仲網	臺中市南屯區永春東路991~1020號	住宅大樓(11層含以上有電梯)	2房2廳2衛	22	6	4600000	政府資料開放平台	1.2130
台中市北屯區瀋陽路二段	大樓	1 房 1 廳 1 衛	22	6	6980000	房仲網	臺中市南屯區永春東路991~1020號	住宅大樓(11層含以上有電梯)	2房2廳2衛	22	6	4600000	政府資料開放平台	1.5174
台中市西屯區文華路	華廈	1 房 1 廳 1 衛	22	6	5980000	房仲網	臺中市南屯區永春東路991~1020號	住宅大樓(11層含以上有電梯)	2房2廳2衛	22	6	4600000	政府資料開放平台	1.3000

Illustration 64: 查詢頁面

4.3 Google Analytics 來收集使用者的瀏覽行為。

Illustration 65: Google Analytics 分析畫面

4.4 請瀏覽者填寫對本系統的意見回饋。

Illustration 66: 網頁的問卷調查畫面

問卷結果：

(1) 樣本數目: 30 人。

(2) 性別: 男 :77% , 女 : 23% 。

(3) 年齡: 30 歲以下：70% ， 30~65 歲 : 30% 。

(4) 目前或未來有購屋的計畫:

有: 77 % ，無: 17% ，未表示：7% 。

Illustration 67

(5) 買屋的消息來源：

親友介紹＋房屋仲介＋網路：27% ，網路：20% ， 房屋仲介+網路：20% ，親友介紹＋房屋仲介：10% ，親友介紹：7% ，房屋仲介：7% ， 親友介紹＋網路：7% ，房屋仲介＋網路＋其他：3%。

Illustration 68

(6) 在哪種網路平台取得資訊:

房屋仲介網＋房屋實價登錄網：77% ,房屋仲介網：13% , 房屋實價登錄網：7% , 房屋仲介網＋房屋實價登錄網＋部落客：3%。

Illustration 69

(7) 將同屋齡同坪數「房仲網」和「開放資料（政府資料開放平台）」在同一列顯示，對購屋時的參考:

0%認為沒有幫助，77%認為有幫助，20%認為很有幫助，3%未表示。

Illustration 70

(8) 顯示房屋資訊（如坪數,屋齡,價格），對購屋時的參考：

0%認為沒有幫助，53%認為有幫助，47%認為很有幫助，0%未表示。

Illustration 71

(9) 顯示同屋齡同坪數的「房仲網總價元/開放資料總價元」，對購屋時的參考：

3%認為沒有幫助，80%認為有幫助，17%認為很有幫助，0%未表示。

Illustration 72

(10) 顯示「房仲網總價元/開放資料總價元」平均值和標準差，對購屋時的參考：

13%認為沒有幫助，63%認為有幫助，17%認為很有幫助，7%未表示。

Illustration 73

(11) 顯示台灣北中南東縣市(如台北市信義區,台中市,高雄市，宜蘭縣) 的房仲網與開放資料比較，對購屋時的參考：

10%認為沒有幫助，80%認為有幫助，10%認為很有幫助，0%未表示。

Illustration 74

五 、結論

5.1 結論

購屋消費者，以往獲得資訊管道是透過房仲業者，或者是消費者累積經驗。不過因網路的發展而將資訊有效地傳播，現在人們可透過房仲網和政府資料開放平台查詢想要的房屋資訊。不過這兩種資料集的資料量大，且缺少整合性比較兩者價差的功能。本實驗結合開放資料和房仲網，比對兩者的房價價差。避免因為房仲業者因掌握獨占資訊，開出價格較高的房屋定價。買方也可藉由資訊開放透明化來得到議價空間。

首先以新竹縣房價來建立分析模式，由相關係數知道「坪數」是影響價格的關鍵，藉由 K-means 分群房仲網和開放資料，得到各群平均價格。實驗結果發現房仲網提供的價格高於開放資料，爲了進一步了解房仲網價格會比開放資料的價格高多少，對房價資料進行更細步的分群實驗 。每五坪為一單位來分群。發現在不同坪數下，房仲網對開放資料的比值會各有不同。看來影響房價的因素不會單只有「坪數」，爲了找出其它影響房價的條件，假設「屋齡」是影響房價的關鍵 ，進行屋齡分群實驗，發現中古屋的確有比新屋有更多的議價空間。爲了找出實際的案例，用電腦比對的方式，在房仲網資料集與開放資料集內，找同縣市內，同屋齡，同坪數的房屋物件，顯示其房屋物件資訊，來驗證實際結果。最後為了提供房價分析的服務，使用 LAMP (Linux，Apache ，MySQL，PHP)架設網站。系統會定期自動到政府資料開放平台和雅虎奇摩房地產網取得實價登錄資料和房地產資料。用 Python 語言將取得的資料處理成可匯入資料庫的檔案，自動上傳到本機端和開放網路端的網頁資料夾下，在以 PHP 檔案自動寫入 MySQL 資料庫。網頁顯示同縣市下，房仲網和開放資料的同坪數同屋齡的物件 。計算房仲網與開放資料總價比值，以及比值的「平均值 」和「標準差 」。 消費者在購買房子時，可以依據本系統所提供的價差比值，平均值和標準差，做有數據基礎的議價。申請 Google Analytics 的服務，觀察到本網站瀏覽者的瀏覽行為。最後請瀏覽過網站的使用者，填寫使用後的建議和回饋。問卷調查發現，超過八成的使用者認為本系統對購買房屋是有幫助，顯示房價分析對於人們的生活是有助益的。

本系統直接在同一網頁呈現房價比較資訊，節省消費者需要到不同網站

查詢的心力。藉由網路服務，讓資訊傳播不再受限人事時地物的限制，讓消費者在購屋時，取得議價資訊更加便利。

5.2 未來工作

- 每次更新資料，將原本資料保留，來與新資料比對加上時間因素來看房價漲跌的情況。
- 房屋區分建物類型來比較價格，比如說，將電梯大樓，和公寓做區分。
- 考慮將房屋與房屋加車位分成兩群在進行各群內部物件的比較。
- 系統顯示房屋的資訊欄位可在多增加一些資訊，比如說，交易的日期。
- 選擇同地段的物件來進行價格比較。坪數和屋齡等資訊可設成搜尋條件。
- 依照房仲網跟政府開放資料的價格比值，顯示這個地區房仲網是否開價過高。房屋價格可加上逗號，分辨閱讀。房屋地址可考慮做成超連結連GoogleMap。加強網頁排版與美化，達到好的預覽效果。
- 提供手機版的房價查詢。
- 目前只提供台灣北中南東代表縣市查詢，未來可擴大提供全台其他縣市的房價查詢。
- 將免費的網路空間移轉到付費平台，以減少網頁廣告。

參考文獻

[1]Jiawei Han，Micheline Kamber ，資料探勘：概念與方法，52 頁，王派洲譯

[2]麥爾荀伯格(Viktor Mayer-Schönberger)，庫基耶(Kenneth Cukier)著 ，大數據，75 頁，林俊宏譯，一版，天下文化，台北市，民國一O 二年三月。

[3]麥爾荀伯格(Viktor Mayer-Schönberger)，庫基耶(Kenneth Cukier)著 ，大數據，85 頁，林俊宏譯，一版，天下文化，台北市，民國一O 二年三月。

[4]麥爾荀伯格(Viktor Mayer-Schönberger)，庫基耶(Kenneth Cukier)著 ，大數據，79 頁，林俊宏譯，一版，天下文化，台北市，民國一O 二年三月。

[5]David Olson，Yong Shi，資料探勘 Introduction to Business Data Mining，302 頁，郭志隆，張芳菱譯，初版，美商麥格羅．希爾，民國九十七年 。

[6]翁卓立著，Linux 進化特區：Ubuntu 13.04 從入門到精通，初版，電腦人文化，台北市，民國一O 二年八月。

[7]Michael Milton 著，深入淺出資料分析，楊仁和譯，初版，歐萊禮，台北市，民國九十九年。

[8]Matthew A. Russell 著，社群網站的資料探勘，師蓉，胡為君譯，初版，歐萊禮，台北市，民國一O 二年五月。

[9]陳景祥著，R 軟體應用統計方法，初版，東華，台北市，民國九十九年九月。

[10]陳會安，PHP+ MySQL 與 jQuery Mobile 跨行動裝置網站開發，初版，碁峰資訊，台北市，民國一O 二年六月。