**國立空中大學**

**「Python 程式設計與實務應用」**

**期末小組專題**

**「房價預測器」**

**應用程式開發文件**

導師：賴素純

班級：ZZZ002

組別：第11組

組名：8089

組長：112222911詹秉蒼

組員（按tronclass平台組員順序）：

112224868蔡惠婷

112224871蔡婷羽

113122776張心齡

113122877陳銘泓

113170973余誼姍

**目 錄**

[一、 專題主題 1](#_Toc185025559)

[二、 專題簡介 1](#_Toc185025560)

[三、 小組分工 2](#_Toc185025561)

[四、 開發環境 3](#_Toc185025562)

[五、 系統開發流程 4](#_Toc185025563)

[（一）需求分析 4](#_Toc185025564)

[（二）定義功能的輸出、輸入與邏輯規則 5](#_Toc185025565)

[（三）開發項目分工與制定完成日期 7](#_Toc185025566)

[（四）時程圖（甘特圖） 8](#_Toc185025567)

[（五）軟體開發程序 9](#_Toc185025568)

[（六）版本控制協作流程 10](#_Toc185025569)

[（七）開發挑戰與問題解決 11](#_Toc185025570)

[（八）整合測試與問題修正 11](#_Toc185025571)

[六、 程式設計主要技巧 12](#_Toc185025572)

[（一）導入模組 12](#_Toc185025573)

[（二）類別方法封裝 12](#_Toc185025574)

[（三）HTTP訪問 13](#_Toc185025575)

[（四）檔案解壓縮與解析 14](#_Toc185025576)

[（五）資料庫批次寫入 15](#_Toc185025577)

[（六）圖形化使用者介面（Graphical User Interface, GUI） 16](#_Toc185025578)

[（七）數據驗證 17](#_Toc185025579)

[七、 原始碼重點解說 18](#_Toc185025580)

[（一）資料下載與解壓縮（lib.DownloadFile.py） 18](#_Toc185025581)

[（二）資料庫連接與CRUD操作（lib.MySQL.py） 20](#_Toc185025582)

[（三）寫入資料庫（lib.CreateLvrData.py） 21](#_Toc185025583)

[（四）GUI介面（資料搜集） 23](#_Toc185025584)

[（五）數據驗證 24](#_Toc185025585)

[（六）資料庫語句生成（Select.py） 26](#_Toc185025586)

[（七）預測模型（predictive\_model.py） 28](#_Toc185025587)

[（八）結果輸出 30](#_Toc185025588)

1. **專題主題**

房價預測器

1. **專題簡介**

本專案開發之軟體，透過爬取內政部實價登錄資料，使用最真實的成交資料作為依據，以統計學方法科學地預測房價行情。

在現今臺灣房價飛漲，房價所得比不斷上升的年代，購屋已不易，又若購屋前沒有做足功課及進行相關統計分析，不熟悉欲購置區域的行情，就容易受騙，以高於行情的價格成交。

購屋前的資料蒐集及統計分析費時又費力，而透過本軟體，購屋者可藉由篩選特定條件的房屋交易資料，並輸入欲購置的時間及物件坪數等相關資料，就可以快速得到欲購買標的之預測房價。有了房屋預測價格，購屋時便不易遭屋主或房仲之話術欺騙，導致買貴。

本軟體可減輕購屋者購屋前資料收集分析的負擔，增加決策正確性及速度，並提供議價依據（議價籌碼）。

1. **小組分工**

|  |  |  |
| --- | --- | --- |
| 工作項目 | | 人員 |
| 提案發想 | | 蔡婷羽（112224871） |
| 需求規格化 | | 蔡婷羽（112224871） |
| 專案統籌規劃 | | 詹秉蒼（112222911） |
| 需求分析 | | 全體組員 |
| 程式撰寫 | GUI介面、數據驗證 | 陳銘泓（113122877） |
| 資料表設計與優化 | 詹秉蒼（112222911） |
| 交易資料下載 | 詹秉蒼（112222911） |
| 交易資料檔彙整與分類 | 余誼姍（113170973）  張心齡（113122776） |
| 資料庫查詢語句與條件產出 | 蔡惠婷（112224868） |
| 預測模型與演算法 | 蔡婷羽（112224871） |
| 整合程式碼 | 詹秉蒼（112222911） |
| 應用程式開發文件編纂 | | 詹秉蒼（112222911）  蔡婷羽（112224871）  陳銘泓（113122877）  蔡惠婷（112224868） |
| 使用手冊製作 | | 余誼姍（113170973）  陳銘泓（113122877）  蔡婷羽（112224871） |
| 執行檔編譯（Windows） | | 陳銘泓（113122877） |
| 執行檔編譯（MacOS） | | 蔡婷羽（112224871） |
| 整合測試 | | 全體組員 |
| 報告投影片製作 | | 張心齡（113122776） |
| 專題報告 | | 張心齡（113122776） |
| 宣傳影片/海報製作 | | 余誼姍（113170973） |

|  |  |
| --- | --- |
| 模組開發者 | 模組名稱 |
| 陳銘泓（113122877） | root.GUI.py |
| 蔡婷羽（112224871） | root.predictive\_model.py |
| 蔡惠婷（112224868） | root.Select.py |
| 余誼姍（113170973） | root.lib.DataFormatting.py（主程式未使用）  root.output.\* （主程式未使用） |
| 詹秉蒼（112222911） | root.lib.Api.py  root.lib.DownloadFile.py  root.lib.CrateLvrData.py  root.lib.MySQL.py  root.lib.Tools.py  root.lib.params.py |
| 張心齡（113122776） | root.lib.opendata\_rearrange.py（主程式未使用） |

1. **開發環境**

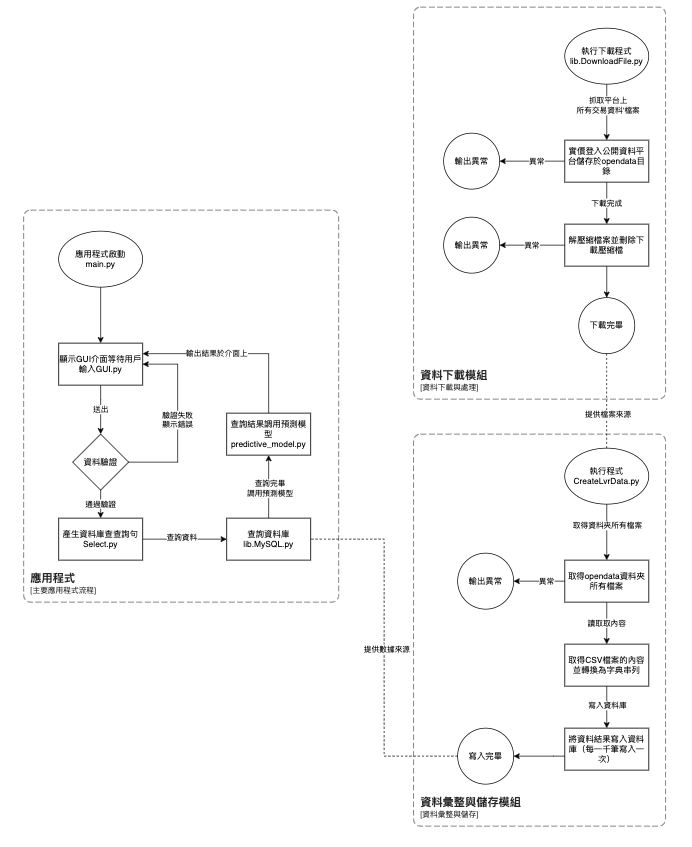
* 語言：Python 3.12.7
* 依賴套件：
  + 檔案下載：requests
  + 檔案解壓縮：zipfile
  + CSV檔案處理：os、sys、csv、datetime、collections、json
  + 資料庫操作：mysql.connector
  + GUI介面與資料驗證：tkinter、customtkinter、functools. Partial
  + 預測模型：pandas、numpy
* 開發系統：Windows、MacOS
* 流程圖：Draw.io
* 資料庫：MySQL 8.1.30
* 版本控制：Git、GitHub
* IDE：Microsoft Visual Studio Code

1. **系統開發流程**

## **（一）需求分析**

由全體組員進行討論，對於提案內容進行流程圖繪製、功能分解、使用套件、分配開發人員、制定函數、類別接口與參數相關工作分派。

系統流程圖如下：

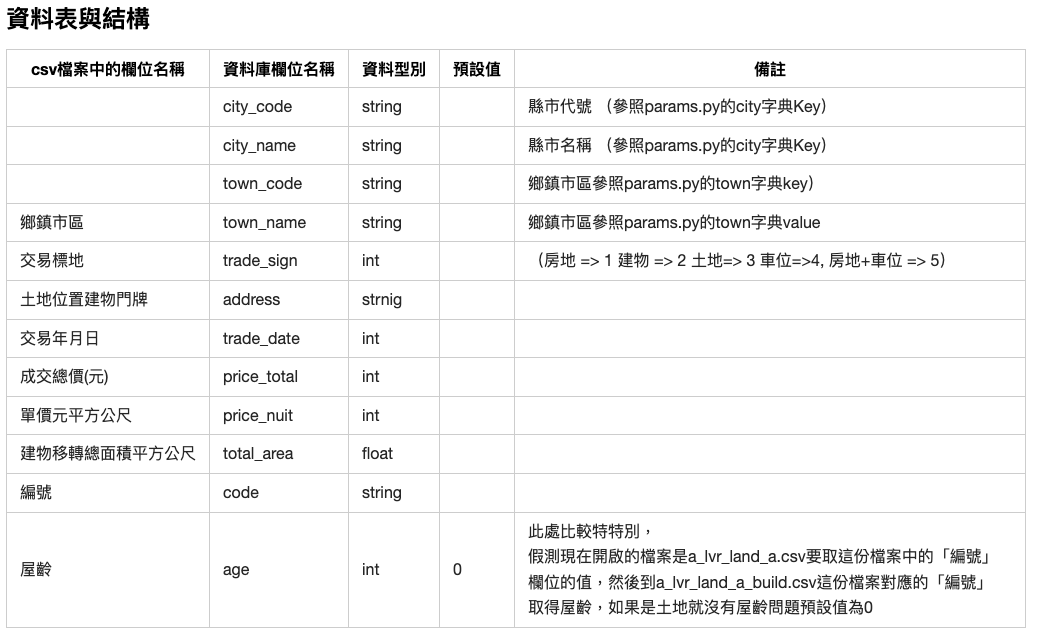


## **（二）定義功能的輸出、輸入與邏輯規則**

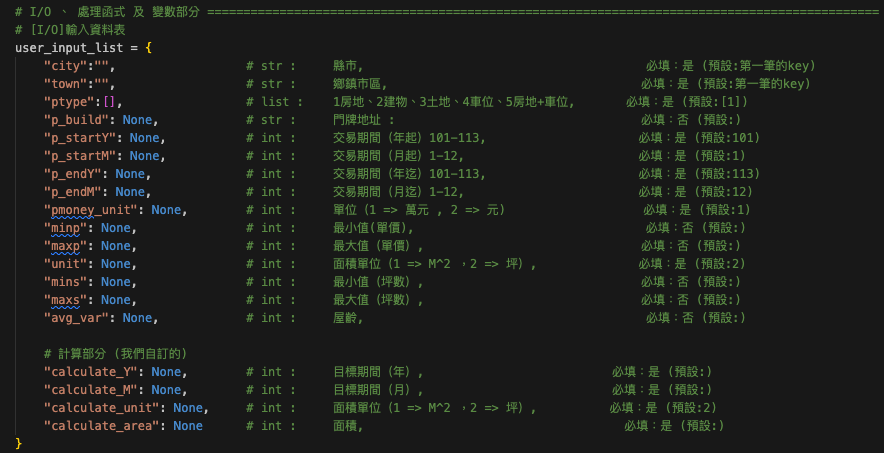
資料下載邏輯：



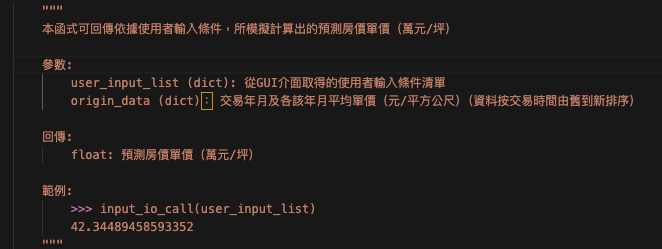
CSV資料轉換為資料庫格式與注意事項：



GUI介面提供的輸出內容：



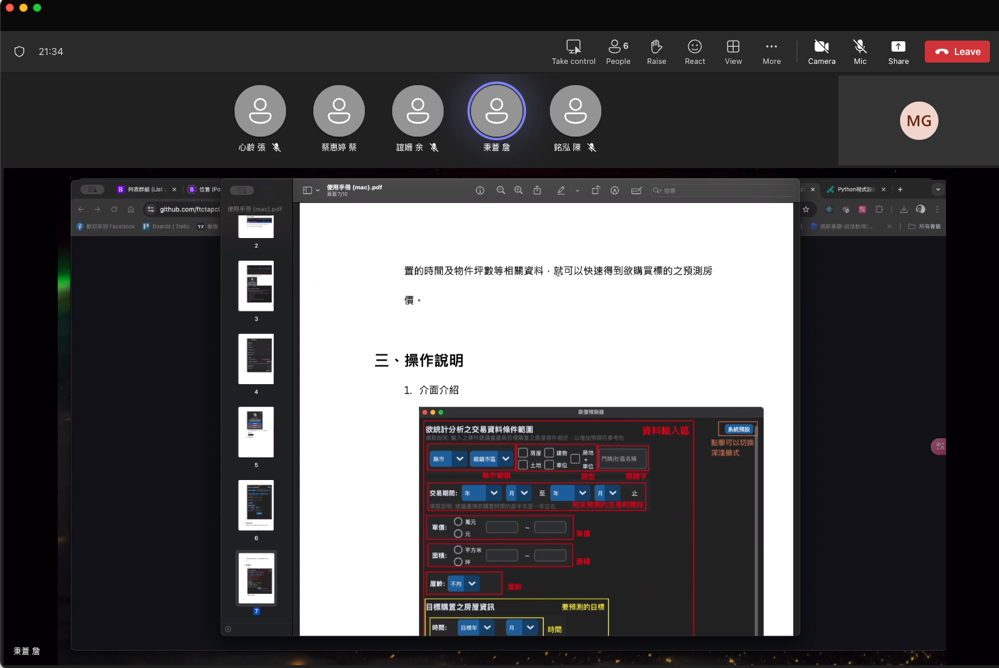
預測模型的輸入與輸出：



## **（三）開發項目分工與制定完成日期**

於每次會議討論工作項目、分配工作並追蹤進度，會議紀錄如下表：

|  |  |
| --- | --- |
| 會議時間 | 討論內容 |
| 113年11月5日  19：00～20：00 | 1. 流程圖解說 2. 討論分工 3. 討論資料型態 4. 相關知識教學（網際網路、requests、API等概念） 5. 初步擬定作業時程 |
| 113年11月19日  21：00～22：20 | 1. 運行一遍程式的運作結果（由組長看完後合併到develop分支，並由組長統一展示） 2. 各程式間的串接輸出與輸入的分析（確認所有屬性是否都存在、型別是否正常） 3. 指派程式串接事務，以及完成時間（1人，串接有問題，原開發者協助除錯與調整） 4. 著手進行文件、報告相關事宜（共同討論如何分工） 5. 作業時程規劃 |
| 113年12月3日  19：00～19：35 | 1. 未盡事宜工作分配（開發文件統整及使用手冊撰寫事宜） 2. 報告投影片展示及說明 |
| 113年12月13日  21：00～21：45 | 1. 共同檢視文件（開發文件、使用手冊）有無缺漏項目，並就不足部分分配撰寫、修改工作 2. 討論更新後程式碼於macOS之程式編譯及發佈事宜（Windows版本已更新並發布） 3. 修改後報告投影片展示及說明 4. 討論小組專題介紹影片或製作海報事宜 |



▲會議畫面截圖

## **（四）時程圖（甘特圖）**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 時間  任務名稱 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 題目發想、提案報告撰寫、需求分析 |  |  |  |  |  |  |  |  |  |  |  |  |
| 需求規格化、程式碼撰寫、開發 |  |  |  |  |  |  |  |  |  |  |  |  |
| 測試（單元測試、整合測試、跨平台測試） |  |  |  |  |  |  |  |  |  |  |  |  |
| 報告投影片製作 |  |  |  |  |  |  |  |  |  |  |  |  |
| 面授報告 |  |  |  |  |  |  |  |  |  |  |  |  |
| 開發文件、使用手冊、宣傳影片或海報製作 |  |  |  |  |  |  |  |  |  |  |  |  |

註：1表示專案進行第一週（10/13～10/19）；2表示專案進行第二週（10/20～10/26）；以下以此類推。

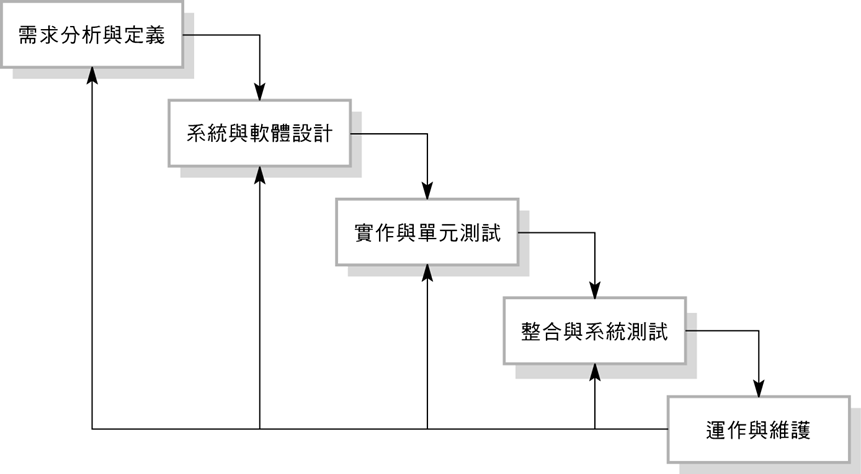
## **（五）軟體開發程序**

軟體開發的程序有四種：瀑布式、漸進式、需求規格化、組合式的軟體開發程序。

本專案採傳統階梯式的軟體開發程序，亦即瀑布式的軟體開發程序，下圖為階梯式的軟體開發模型，也就是Waterfall model。

此軟體開發程序，講求軟體生命週期（Software life cycle），將軟體發展劃分為明確階段。此開發模式為標準的程序，為大多數人所接受，有利於軟體專案的管理。

由前揭甘特圖可見得，本專案開發程序中，有明確的階段。在經過系統需求分析後，進行需求規格化，接著進行系統開發實作，實作的過程中，亦同時進行單元測試，實作完畢後，進行整合測試，故屬瀑布式的軟體開發程序。

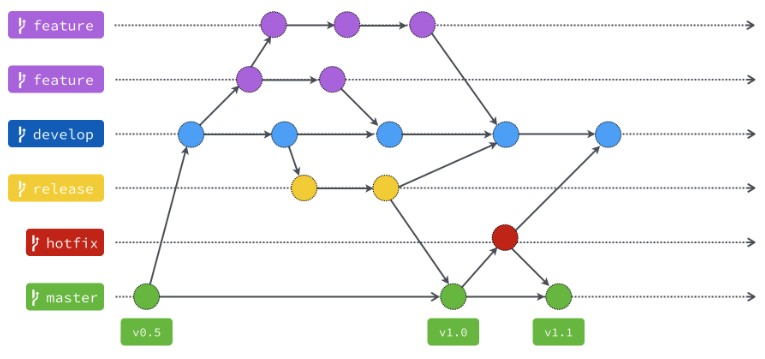


(圖片來源：<https://wayne265265.pixnet.net/blog/post/113080214>)

## **（六）版本控制協作流程**

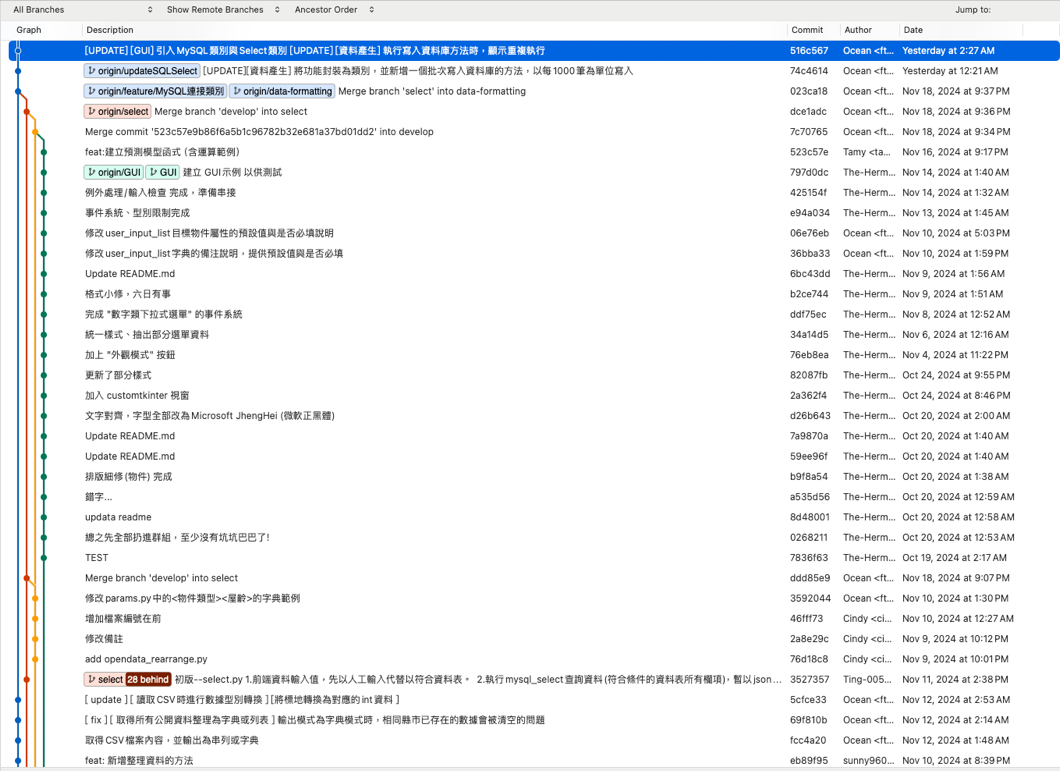
採用現今最流行的GitFlow流程，將Branch（分支）狀態系統化。

組員開發各自所負責之功能時，於各自之feature分支進行開發，功能開發完畢後，合併至develop分支，待功能測試完畢，整體程式可正常執行，即將develop分支合併至main分支（master分支），即成為穩定上線版本。



（圖片來源：<https://gitbook.tw/chapters/gitflow/why-need-git-flow>）

下圖為本專案版本控制協作流程歷程節錄：



## **（七）開發挑戰與問題解決**

**問題一**：實價登入網站的版本更新，無法獲取資料來源。

**解決方式：**使用爬蟲抓取「[不動產成交案件實際資訊資料供應系統](https://plvr.land.moi.gov.tw/Index)」中的公開資料，來解決無法即時獲取交易資料問題，作為解決方案。

**問題二**：下載之數據量高達460萬餘筆，使用資料庫儲存時，遭遇效能問題。

**解決方式：**透過資料庫的分析語句工具對相關的查詢語句，進行索引的優化與調教。

## **（八）整合測試與問題修正**

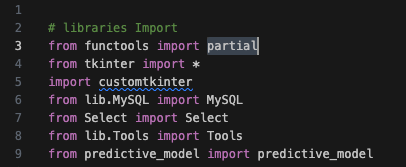
由組員各自在本機上進行完整的測試，組員使用之作業系統包含Windows及macOS，故測試包含跨平台測試。

測試階段遇到問題反映給原作者，由原作者完成修正後，再進行測試。

1. **程式設計主要技巧**

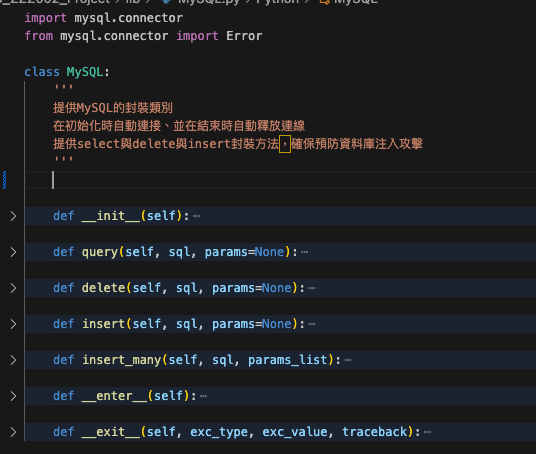
## **（一）導入模組**

導入Python原生、第三方模組加速建構，以及組員開發之各實作程式，使開發工作可分工各自進行，不互相干擾。



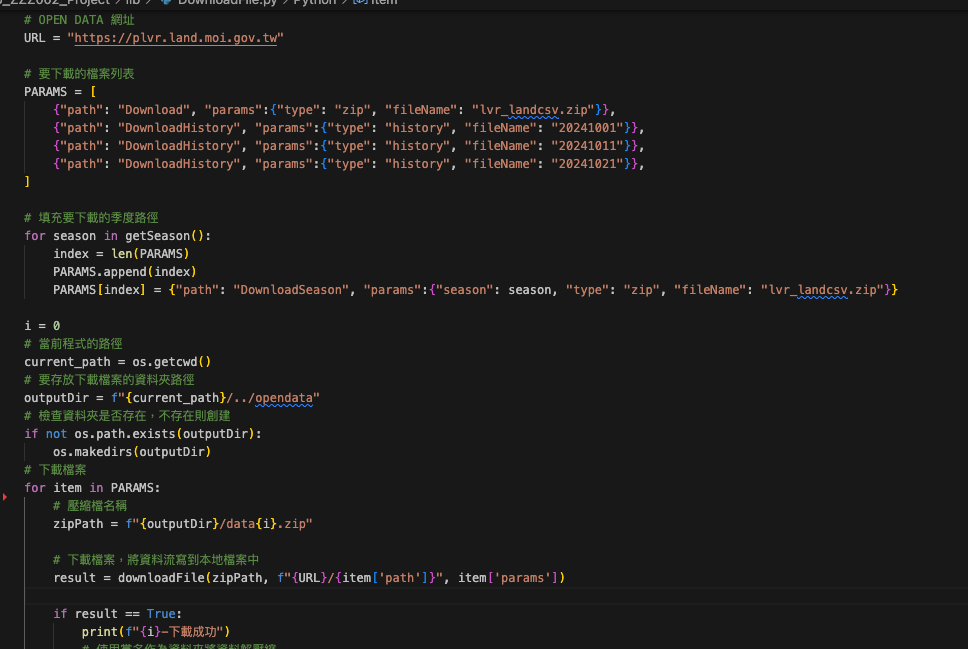
## **（二）類別方法封裝**

下圖將資料庫的新增（C）、查詢（R）、刪除（D）、更新（U）各方法封裝為類別：



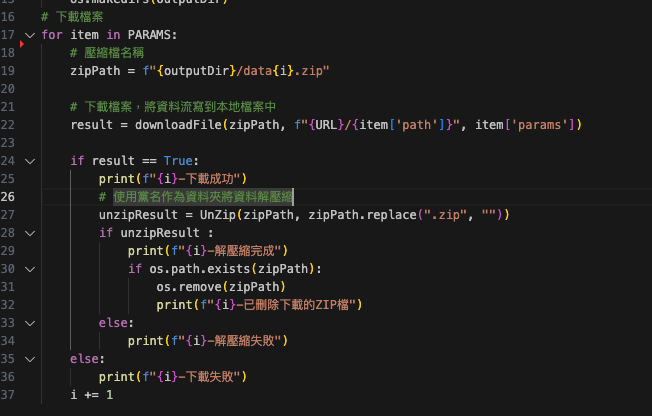
## **（三）HTTP訪問**

使用requests套件，下載「[不動產成交案件實際資訊資料供應系統](https://plvr.land.moi.gov.tw/Index)」中的交易資料。下圖為先產生所有欲下載的連結，透過http GET方法將所有公開檔案下載到本地端。

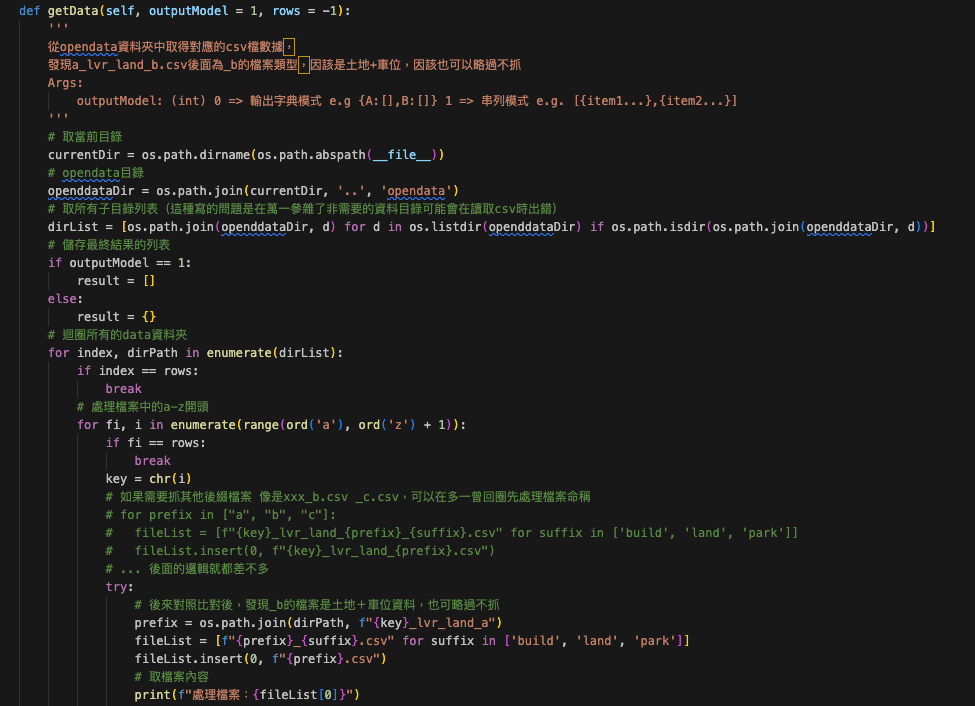


## **（四）檔案解壓縮與解析**

將下載的檔案（壓縮檔進行解壓縮）完成後刪除，避免佔用過多空間。



使用CSV套件以取得檔案內容，並彙整成資料庫所需格式。



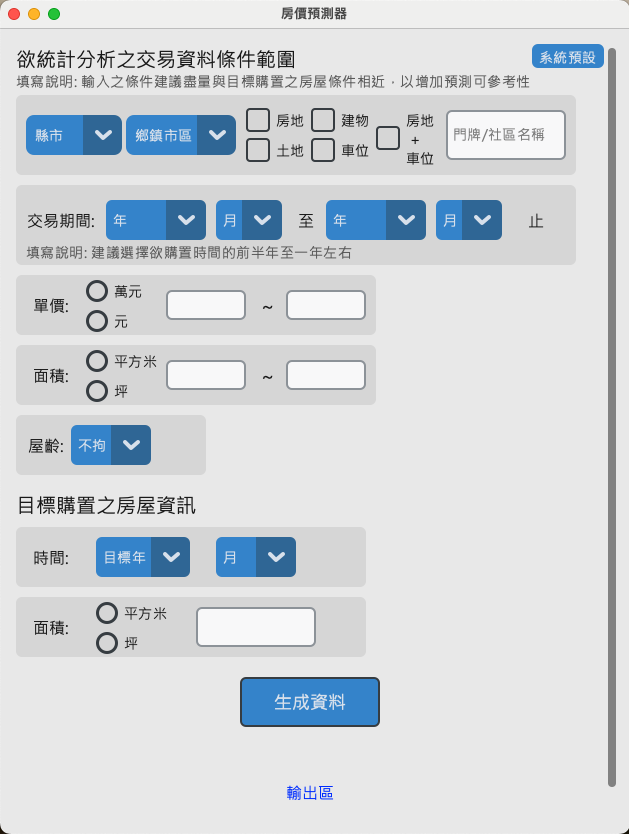
## **（五）資料庫批次寫入**

因搜集到的數據量超過百萬筆（約460萬餘筆），若一次性寫入資料庫，將造成大量數據庫開銷，故採用每一千筆批次寫入，使資料庫的IO達到較良好的效能。



## **（六）圖形化使用者介面（Graphical User Interface, GUI）**

透過圖形化使用者介面搜集之用戶對資料的篩選條件，以此為基礎搜尋資料庫中之對應資料，並生成預測結果。



## **（七）數據驗證**

提醒用戶必要輸入的數據，以及進行數據型態的控制與驗證。



1. **原始碼重點解說**

## **（一）資料下載與解壓縮（lib.DownloadFile.py）**

主要目的用於下載自101年起至當下之實價登錄資料，並將其解壓縮後儲存於本地端。主要邏輯片段：

|  |
| --- |
| 函數主要產生101年到當年的年+季以[明國年] S [第幾季]為返回格式。  透過迴圈產生出對應公開資料平台的下載點URL。  以字典串列形式返回，字典中包含檔名、路徑、類型。    將資料解壓縮後，再刪除壓縮檔以節省儲存空間。 |

## **（二）資料庫連接與CRUD操作（lib.MySQL.py）**

主要用於封裝mysql.connector套件中的資料操作方法，方便重複調用，以及於每次操作資料庫完畢後，自動釋放連線資源。

|  |
| --- |
| 封裝CRUD Create 、Read、Update、Delete資料操作方法，並再提供一個針對批次寫入的方法。    透過\_\_enter\_\_、\_\_exit\_\_在結束時自動釋放資料庫連線。 |
|  |

## **（三）寫入資料庫（lib.CreateLvrData.py）**

主要用於提取所有下載的CSV檔案中的內容，並整理為可寫入資料庫的數據型態，將資料儲存於MySQL資料庫中。

|  |
| --- |
| CreateLvrData.getData()方法：  取得lib.DownloadFile.py下載的CSV檔案並從中提取數據，進行數據的資料轉換。    CreateLvrData.getCsv()方法：  對CSV中的數據進行預處理與轉換。    CreateLvrData.insertSQL ()方法：  將取得的數據以每批一千筆的的速率寫入資料庫。 |

## **（四）GUI介面（資料搜集）**

|  |
| --- |
| 使用tkinter及customtkinter套件製作GUI介面，允許用戶通過GUI介面圖像化輸入資料，介面中物件程式碼範本如下：    使用元件包括: customtkinter的CTkScrollableFrame、CTkFrame、CTkButton、CTkLabel、CTkOptionMenu、CTkCheckBox、CTkRadioButton及CTkEntry。  使用者填寫資料時，會通過回調函數將資料寫入user\_input\_list： |

## **（五）數據驗證**

|  |
| --- |
| 當使用者按下[生成資料]按鈕時，會開始檢查輸入資料是否正確、完整：    檢查函數分別驗證每項資料是否合規，若不符合規則，則判斷是否有預設值。若有，則填入預設值；若無，則對使用者進行文字警告：    如果資料填寫完整且合規，則呼叫資料庫及預測模型： |

## **（六）資料庫語句生成（Select.py）**

主要用於接收GUI介面所搜集到的用戶輸入資料，並產生對應的資料庫查詢語句

|  |
| --- |
| Select.createQuery()方法：  處理傳入的數據，產生對應的資料庫查詢語句，若數據沒有值或者為None時，將不會產生對應的欄位查詢語句。  定義函數def adjust\_trade\_date()：  為使條件式欄位「trade\_date」值能符合資料表而進行轉換成完整日期格式。    定義函數def save\_to\_json()：  將最後查詢資料存入json檔，用於預測模型。    定義函數 def createQuery()：  利用if-elif-else查詢資料表，先判斷以下欄位名稱後，再因應欄位屬性進行其它判斷。 |

## **（七）預測模型（predictive\_model.py）**

用於接收透過資料庫查詢語句挑出的資料，並使用統計方法，運算出預測不動產單價後，將其回傳。

|  |
| --- |
| predictive\_model.predictive\_model()方法：  本方法有兩個參數，第一個參數為由GUI介面取得之使用者輸入條件清單；第二個參數為透過資料庫查詢語句挑出的資料（交易年月及各該年月平均單價（元/平方公尺）（資料按交易時間由舊到新排序））。  引入用於數據分析與處理之第三方函式庫Pandas及NumPy進行不動產成交價分析與預測。Pandas為用於高階資料結構（如DataFrame）和資料處理之工具；NumPy用於數值計算，提供陣列操作和進階的數學函式。    本方法將交易時間作為解釋變數（自變數），將不動產交易單價作為被解釋變數（依變數），透過統計學之「最小平方法」，求出簡單線性迴歸式，亦即不動產單價預測公式，再透過使用者輸入的欲購置時間，以產生對應該時點的預測不動產單價。 |

## **（八）結果輸出**

|  |
| --- |
| output\_show()函數取得回傳值後，判斷回傳值是計算結果(正數)或錯誤代碼(負數)，並以此輸出結果至GUI:    錯誤範例: |