Download and clean data

- 功能敘述：到http://plvr.land.moi.gov.tw/DownloadOpenData下載台北市的不動產交易資料，整理資料後再輸出檔案。
- 使用 Python(Pandas)
- 整個程式共三個檔 main.py, downloadData.py 和 cleanData.py。
- 資料轉換規則由 CATEGORY_MAP 給定，不在範圍內的資料視為 invalid data，不會被輸出，另外要標記資料中的 outlier(要輸出)，以下為 CATEGORY_MAP 的部分截圖：

```
CATEGORY_MAP = {'主要建材': {'': 0,
'加強磚造': 1,
 '土木造': 2,
'土磚石混合造': 3,
'土造': 4,'壁式預鑄鋼筋混凝土造': 5,
             '木造': 6,
'石造': 7,
 '磚造': 8,
'見使用執照': 9,
 '見其他登記事項': 10,
 '鋼筋混凝土加強磚造': 11,
 '鋼筋混凝土造': 12, '鋼骨混凝土造': 13,
'鋼骨鋼筋混凝土造': 14,'鐵造': 15,'預力混凝土造': 16},

'主要用途': {'': 0,'住商用': 1,
'住家用': 2,
'住工用': 3,'停車空間': 4,'商業用': 5,
'工商用': 6,
'工業用': 7,'見使用執照': 8,
             '見其他登記事項': 9,
'農舍': 10},

'建物型態': {'': 0,
'住宅大樓(11層含以上有電梯)': 1,'倉庫': 2,
'公寓(5樓含以下無電梯)': 3,'其他': 4,'套房(1房1廳1衛)': 5,
'工廠': 6,
    '店面(店鋪)': 7,'廠辦': 8,'華廈(10層含以下有電梯)': 9,
'辦公商業大樓': 10,
 '農舍': 11,
'透天厝': 12},
```

以下說明此程式的運作:

1. 此程式將會在此內政部不動產成交案件的網站下載台北市的資料，共有**不動產買賣**、**預售屋買賣**和**不動產租賃**三種類型的檔案，而一年又依四季分為四段，所以一年份的資料共有 12 個檔案。

在主程式 main.py 中執行 processingData(102)，會下載台北市 102 年四季的不動產買賣、預售屋買賣以及不動產租賃的資料。(若執行 processingData(102, 107)則會下載 102 到 107 共 6 年的資料)

2. 下載完檔案後會將要留下的資料依 CATEGORY_MAP 的規定轉換格式，並刪去和預設格式不符合的資料，最後輸出整理好的 csv 檔，而被刪去的資料會依刪除的原因整理成 invalidDataLog.txt 和 outlierLog.txt 兩個檔案一併輸出。

下圖為台北市 102 年第 1 季不動產買賣資料的部分截圖(未整理資料)：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	鄉鎮市區	交易標的	土地區段	土地移轉	都市土地	非都市土	非都市土	交易年月	交易筆棟	移轉層次	總樓層數	建物型態	主要用途	主要建材
2	The villag	transactior	land sector	land shiftir	the use zor	the non-m	non-metro	transactior	month an	transactior	shifting le	total floor	building st	main use
3	文山區	房地(土	台北市文	16.4	住			1011012	土地1建	七層	十四層	住宅大樓	住家用	鋼筋混凝
4	文山區	房地(土	台北市文	52.66	住			1011009	土地1建	六層，電	七層	華廈(10層	住家用	鋼筋混凝
5	文山區	土地	興隆段四	1.04	其他			1011026	土地3建	物0車位0		其他		
6	文山區	房地(土	台北市文	23.67	住			1011011	土地1建	一層，平	四層	公寓(5樓	住家用	鋼筋混凝
7	文山區	土地	興泰段二	53	其他			1011019	土地1建	物0車位0		其他		
8	文山區	房地(土	台北市文	14.26	住			1011020	土地2建	八層	九層	華廈(10層	住家用	鋼筋混凝
9	文山區	房地(土	台北市文	10.68	住			1011001	土地0建	十一層	十六層	套房(1房	住家用	鋼筋混凝
10	文山區	房地(土	台北市文	30.9	住			1011008	土地1建	五層	七層	華廈(10層	住家用	鋼筋混凝
11	文山區	房地(土	台北市文	20.46	商			1010919	土地1建	騎樓，一	五層	公寓(5樓	商業用	鋼筋混凝
12	大同區	房地(土	台北市大	17.81	商			1011005	土地3建	七層	十三層	住宅大樓	住家用	鋼筋混凝
13	萬華區	房地(土	台北市萬	26.49	住			1011015	土地1建	三層	十二層	住宅大樓	國民住宅	鋼筋混凝
14	中正區	房地(土	台北市中	11.28	住			1011017	土地1建	一層	七層	華廈(10層	見其他登	鋼筋混凝
15	萬華區	土地	雙園段二	40.72	其他			1011024	土地4建	物0車位0		其他		
16	萬華區	土地	華中段一	205	其他			1011024	土地1建	物0車位0		其他		
17	中正區	房地(土	台北市中	137.99	商			1010904	土地1建	十二層	十四層	辦公商業	商業用	鋼筋混凝
18	萬華區	房地(土	台北市萬	16.4	住			1011002	土地1建	二層	五層	公寓(5樓	住家用	鋼筋混凝

下圖為整理後的情況：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	鄉鎮市區	交易標的	土地區段	土地移轉	都市土地	非都市土	非都市土	交易年月	交易筆棟	移轉層次	總樓層數	建物型態	主要用途	主要建材
2	10	房地(土	台北市文	16.4	1			1011012	土地1建	七層	十四層	1	2	12
3	10	房地(土	台北市文	52.66	1			1011009	土地1建	六層，電	七層	9	2	12
4	10	土地	興隆段四	1.04	2			1011026	土地3建	物0車位0		4	0	0
5	10	房地(土	台北市文	23.67	1			1011011	土地1建	一層，平	四層	3	2	12
6	10	土地	興泰段二	53	2			1011019	土地1建	物0車位0		4	0	0
7	10	房地(土	台北市文	14.26	1			1011020	土地2建	八層	九層	9	2	12
8	10	房地(土	台北市文	10.68	1			1011001	土地0建	十一層	十六層	5	2	12
9	10	房地(土	台北市文	30.9	1			1011008	土地1建	五層	七層	9	2	12
10	10	房地(土	台北市文	20.46	3			1010919	土地1建	騎樓，一	五層	3	5	12
11	8	房地(土	台北市大	17.81	3			1011005	土地3建	七層	十三層	1	2	12
12	2	房地(土	台北市中	11.28	1			1011017	土地1建	一層	七層	9	9	12
13	12	土地	雙園段二	40.72	2			1011024	土地4建	物0車位0		4	0	0
14	12	土地	華中段一	205	2			1011024	土地1建	物0車位0		4	0	0
15	2	房地(土	台北市中	137.99	3			1010904	土地1建	十二層	十四層	10	5	12
16	12	房地(土	台北市萬	16.4	1			1011002	土地1建	二層	五層	3	2	12
17	8	房地(土	台北市大	12.85	3			1010916	土地1建	二層	十二層	10	5	12
18	8	房地(土	台北市大	19.19	3			1010916	土地1建	一層，騎	十二層	10	5	12
19	8	房地(土	台北市大	59.16	3			1010916	土地1建	地下層	十二層	10	5	12

3. 以下為整理檔案的規則：
   - 在整理資料時，invalid data 是在將字串轉換為數字代號時，不在 category_map 中的資料，invalid data 不會被輸出，執行完會有一個 invalidDataLog.txt，其中記錄了整理資料時的 invalid data，以及輸出檔的資料筆數。
   - 輸出的檔案多加了一個"outlier" 欄，一般資料的值為 0，outlier 的值為 1，far outlier 的值為 2。

     Outlier 的定義如下：在**鄉鎮市區**和**建物型態**的分群下，**單價(元/平方公尺)**數值太極端的資料，極端數值的範圍是由 Interquartile Range (IQR)所定義，數值低於 Q1-1.5IQR 或高於 Q3+1.5IQR 者視為 outlier，其中數值低於 Q1-3IQR 或高於 Q3+3IQR 者視為 far outlier。

下圖為 102 至 107 年不動產買賣平均價錢圖表：