

Project

- 任選一個資料集
- 上台報告日期: 6/20, 6/27
- 報告前請將全部已執行出結果的 .ipynb檔, 存入usb , copy 到教室的老師電腦
- 同學互評與老師一起評分
- 評分標準
 - 資料擷取匯入, 產生dataframe: 20%
 - 每個欄位說明: 10%
 - 資料集統計分析, 相關性分析(correlation): 20%
 - 聚合函數(groupby)、樞紐分析(pivot table): 20%
 - 使用Matplotlib 作圖: 20%
 - 資料分析與心得洞察結果: 10%

Project example

[新竹市不動產實價登錄資訊-買賣案件| 政府資料開放平臺](https://data.gov.tw/dataset/67502)

<https://data.gov.tw/dataset/67502>

鄉鎮市區、交易標的、土地區段位置/建物區段門牌、土地移轉總面積[平方公尺]、使用分區或編定、非都市土地使用分區、非都市土地使用地、交易年月、交易筆棟數、移轉層次、總樓層數、建物型態、主要用途、主要建材、建築完成年月、建物移轉總面積[平方公尺]、現況格局-房、現況格局-廳、現況格局-衛、現況格局-隔間、有無管理組織、總價[元]、單價[元/平方公尺]、車位...¶

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43456 entries, 0 to 43455
Data columns (total 27 columns):
P1LA_CF46      42881 non-null object
P1MA_CASEFLAG  43456 non-null object
P1MA_DD09      43456 non-null object
P1LA_CArea     43456 non-null float64
P1LA_C11_1     37214 non-null object
P1LA_C12_1     5986 non-null object
P1LA_C12_2     5794 non-null object
(沒有資料行名稱) 43456 non-null int64
(沒有資料行名稱)1 43456 non-null object
P1JD_14_1     36516 non-null object
P1LA_F13      37664 non-null object
P1MA_BUILD5    43456 non-null object
P1LA_F11      37566 non-null object
P1LA_F12      37659 non-null object
(沒有資料行名稱)2 35754 non-null float64
P1LA_FArea     43456 non-null float64
P1MA_BUILD1    35252 non-null object
P1MA_BUILD2    34042 non-null object
P1MA_BUILD3    35494 non-null object
P1MA_BUILD4    43456 non-null object
P1MA_MANAGE    43456 non-null object
P1MA_TOTPRICE  43456 non-null int64
MeanPrice      43456 non-null int64
P1PA_PARK_1    21905 non-null object
P1PA_PARKAREA  43456 non-null float64
P1PA_PARKPRICE 43456 non-null int64
P1MA_NOTE      43456 non-null object

```

鄉鎮市區、交易標的、土地區段位置/建物區段門牌、土地移轉總面積[平方公尺]、使用分區或編定、非都市土地使用分區、非都市土地使用地、交易年月、交易筆棟數、移轉層次、總樓層數、建物型態、主要用途、主要建材、建築完成年月、建物移轉總面積[平方公尺]、現況格局-房、現況格局-廳、現況格局-衛、現況格局-隔間、有無管理組織、總價[元]、單價[元/平方公尺]、車位...¶

少於 43456 的 欄位表示有 Missing value, Nan, 遺失值

取有興趣的欄位

```
dfmo=df[ ["P1LA_CF46", "(沒有資料行名稱)2", "P1LA_FArea", "P1MA_BUILD1", "P1MA_BUILD2", "P1MA_BUILD3", "P1MA_TOTPRICE", "P1MA_BUILD5" ] ]
```

↑
建物型態

用 `value_counts()` 去看有哪些,及有多少個

建物型態

```
dfmo["P1MA_BUILD5"].value_counts()
```

Out[16]:

住宅大樓	17158
透天厝	6448
華廈	6065
土地	5776
套房	3340
公寓	2551
店面	747
車位	624
辦公商業大樓	519
其他	70
廠辦	63
工廠	57
農舍	37
倉庫	1

Name: P1MA_BUILD5, dtype: int64

鄉鎮市區

```
dfmo["P1LA_CF46"].value_counts()
```

Out[14]:

東區	23329
北區	12013
香山區	7539

Name: P1LA_CF46, dtype: int64

Project 其他資料集

- 美國開放資料平台
(<https://www.data.gov/>)
- 加州大學爾灣分校機器學習資料庫
(<http://archive.ics.uci.edu/ml/>)
- Stanford Large Network Dataset Collection
(<https://snap.stanford.edu/data/>)
- Kaggle (<https://www.kaggle.com/>)

被 google 買走

你聽過 Kaggle 嗎？Google 買下知名機器學習社群，加速推廣雲端AI

如何取 dataframe row/column 位置

- df.loc

`dfmo.loc[2, "P1MA_BUILD1"]`

`dfmo[:5]`

	P1LA_CF46	(沒有資料行名稱)2	P1LA_FArea	P1MA_BUILD1	P1MA_BUILD2	P1MA_BUILD3	P1MA_TOTPRICE
0	香山區	7209.0	114.860	3房	1廳	1衛	8500000
1	東區	NaN	75.900	3房	2廳	1衛	5408000
2	東區	NaN	0.000	NaN	NaN	NaN	13770000
3	東區	10102.0	103.851	1房	1廳	1衛	6900000
4	香山區	8705.0	28.392	1房	NaN	1衛	520000

df.loc : explicit index

df.iloc : implicit index, as if it is a simple Numpy array

如何修改欄位名稱

```
dfmo[:5]
```

	P1LA_CF46	(沒有資料行名稱)2	P1LA_FArea	P1MA_BUILD1	P1MA_BUILD2	P1MA_BUILD3	P1MA_TOTPRICE
0	香山區	7209.0	114.860	3房	1廳	1衛	8500000
1	東區	NaN	75.900	3房	2廳	1衛	5408000
2	東區	NaN	0.000	NaN	NaN	NaN	13770000
3	東區	10102.0	103.851	1房	1廳	1衛	6900000
4	香山區	8705.0	28.392	1房	NaN	1衛	520000

```
dfmo=dfmo.rename(columns={'P1LA_CF46':'鄉鎮市區',  
                          '(沒有資料行名稱)2':'建築完成年月',  
                          'P1LA_FArea':'建物移轉總面積[平方公尺]'})
```

```
dfmo[:6]
```

	鄉鎮市區	建築完成年月	建物移轉總面積[平方公尺]	P1MA_BUILD1	P1MA_BUILD2	P1MA_BUILD3	P1MA_TOTPRICE
0	香山區	7209.0	114.860	3房	1廳	1衛	8500000
1	東區	NaN	75.900	3房	2廳	1衛	5408000
2	東區	NaN	0.000	NaN	NaN	NaN	13770000
3	東區	10102.0	103.851	1房	1廳	1衛	6900000
4	香山區	8705.0	28.392	1房	NaN	1衛	520000
5	東區	7106.0	110.400	4房	2廳	1衛	5428000

一些常用dataframe 指令

- 刪除欄位
 - `df_new=df.drop(“欲刪除的欄位 “ , axis=1)`
- 刪除row
 - `df=df.drop([rowi rowj ...])`
- 刪除所有有遺失值 的 rows.
 - `df_new=df.dropna()`
- 用 `df.describe()` 去看各欄的
- `count,mean,std,min,max` 統計資料
- 用 `df.info()`: a concise summary of a DataFrame

取值

- 單欄:
- 多欄:

```
col=pd1[ 'Hk1' ]  
print(col)
```

```
John    A  
Mary    B  
susan   A  
Peter   C  
Lin     A  
Name: Hk1, dtype: object
```

```
cols=pd1[['Hk1','HK2']]  
print(cols)
```

```
      Hk1  HK2  
John    A    A  
Mary    B    A  
susan   A    A  
Peter   C    B  
Lin     A    A
```

取值

- Examine the raw data array using the values attributes

```
dfmo.values
```


```
Out[14]:
```

```
array( [['香山區', 7209.0, 114.86, ..., '1衛', 8500000, '透天厝'],  
        ['東區', nan, 75.9, ..., '1衛', 5408000, '公寓'],  
        ['東區', nan, 0.0, ..., nan, 13770000, '土地'],  
        ...,  
        ['北區', nan, 0.0, ..., nan, 5793523, '土地'],  
        ['香山區', nan, 0.0, ..., nan, 2880001, '土地'],  
        ['北區', nan, 103.5, ..., '2衛', 8500000, '透天厝'] ], dtype=object)
```

如何新增欄位

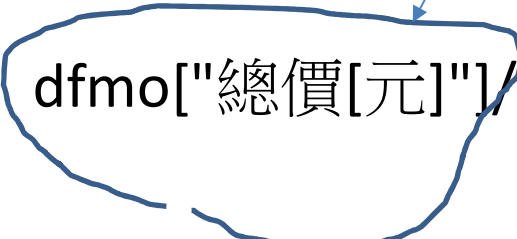
新增 list

`dfmo['現況格局-房'] = t_room`

A blue oval highlights the variable 't_room' in the code snippet. A blue arrow points from the text '新增 list' (Add new list) to this oval.

用 既有的欄位

`dfmo["總價[百萬元]"] = dfmo["總價[元]"]/1000000`

A blue oval highlights the expression 'dfmo["總價[元]"]/1000000' in the code snippet. A blue arrow points from the text '用 既有的欄位' (Use existing column) to this oval.