

0509 Lecture

- Review
 - Midterm and HK2 are graded
 - HK3 solution on web for reference
- Today
 - 從政府資料平台抓 CSV 檔 → DataFrame
 - NumPy : ndarray
 - Pandas : Series/DataFrame
- Next week 5/16
 - Project team member

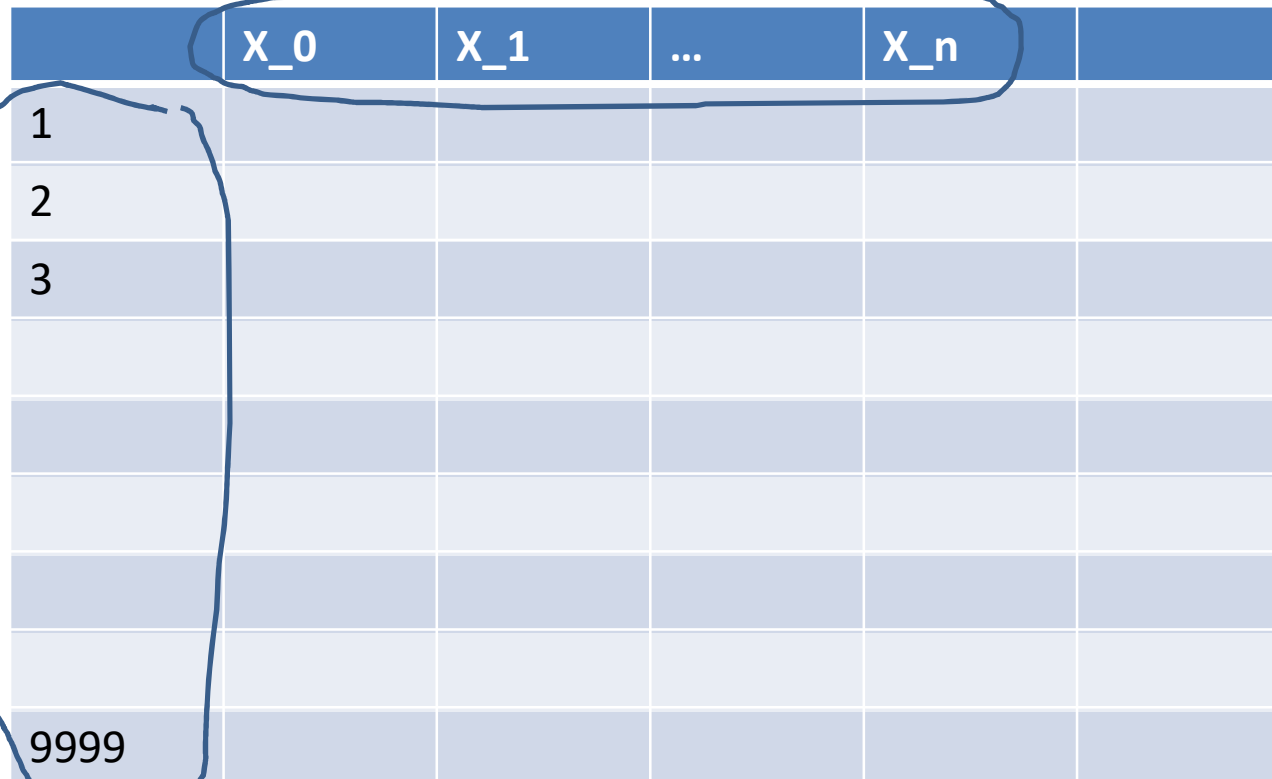
Project

- 3 種主要資料格式 : html, json, csv
 - 政府資料平台(json, csv): <https://data.gov.tw/>
 - 腸病毒資料集: FGU-Class/Project/Health_data_0.ipynb
 - 大家可以任選資料集....
 - Preprocessing: 清理, 轉換
 - Visualization: 折線圖, 長條圖, ...
-
- 我們可以從資料中學習到什麼?
 - Learning : regression or classification
 - Classification: e.g. spam filter
 - Regression : e.g. 房價 股價 predictor
 - google : Hello World - Machine Learning Recipes
 - google : The 7 Steps of Machine Learning



Goal : From dataset to generate the following DataFrame

There are n data features:
columns



	X_0	X_1	...	X_n	
1					
2					
3					
9999					

index

Maybe there are tons of data !! Big data !!

Ex1 :健保門診及住院就診人次統計-腸病毒

首頁 » 資料集 » 健保門診及住院就診人次統計-腸病毒

健保門診及住院就診人次統計-腸病毒

資料集評分:

★★★★☆

平均 3 (2 人次投票)

資料集描述:

各縣市、年齡別、年週之腸病毒門診及住院就診人次統計

主要欄位說明:

年、週、就診類別、年齡別、縣市、腸病毒健保就診人次

資料資源:

CSV

JSON



檢視資料

腸病毒



檢視資料

腸病毒

提供機關:

衛生福利部疾病管制署

提供機關聯絡人:

王先生 (02-23959825#4032)

更新頻率:

每月

授權方式:

政府資料開放授權條款-第1版

計費方式:

免費

上架日期:

2015/05/01

CSV 格式

<https://data.gov.tw/dataset/14590>

Ex1 :健保門診及住院就診人次統計-腸病毒

年,週,就診類別,年齡別,縣市,腸病毒健保就診人次,健保就診總人次

2008,14,住院,0-2,台中市,0,105

2008,14,住院,0-2,台北市,2,151

2008,14,住院,0-2,台東縣,0,14

2008,14,住院,0-2,台南市,0,20

2008,14,住院,0-2,宜蘭縣,0,44

2008,14,住院,0-2,花蓮縣,0,17

2008,14,住院,0-2,金門縣,0,1

2008,14,住院,0-2,屏東縣,0,19

2008,14,住院,0-2,苗栗縣,0,1

2008,14,住院,0-2,桃園市,0,141

2008,14,住院,0-2,高雄市,2,87

2008,14,住院,0-2,基隆市,0,21

▷ CSV 格式

逗號分隔值（Comma-Separated Values，CSV，有時也稱為字元分隔值，因為分隔字元也可以不是逗號），其檔案以純文字形式儲存表格資料（數字和文字）。（維基百科，自由的百科全書）

Ex1 :健保門診及住院就診人次統計-腸病毒

DataFrame

Read CSV file into DataFrame

```
In [20]: df1=pandas.read_csv("NHI_EnteroviralInfection.csv")
```

```
In [21]: df1[:10]
```

columns

Out[21]:

Slicing 取 row :
前面 10 rows

index

	年	週	就診類別	年齡別	縣市	腸病毒健保就診人次	健保就診總人次
0	2008	14	住院	0-2	台中市	0	105
1	2008	14	住院	0-2	台北市	2	151
2	2008	14	住院	0-2	台東縣	0	14
3	2008	14	住院	0-2	台南市	0	20
4	2008	14	住院	0-2	宜蘭縣	0	44
5	2008	14	住院	0-2	花蓮縣	0	17
6	2008	14	住院	0-2	金門縣	0	1
7	2008	14	住院	0-2	屏東縣	0	19
8	2008	14	住院	0-2	苗栗縣	0	1
9	2008	14	住院	0-2	桃園市	0	141

Ex1 :健保門診及住院就診人次統計-腸病毒

- 取欄:
 - ex: `df1['縣市']`

Ex1 :健保門診及住院就診人次統計-腸病毒

```
In [22]: df1.info()
```

總共有 111319 筆資料

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 111319 entries, 0 to 111318  
Data columns (total 7 columns):  
年                111319 non-null int64  
週                111319 non-null int64  
就診類別          111319 non-null object  
年齡別            111319 non-null object  
縣市              111319 non-null object  
腸病毒健保就診人次  111319 non-null int64  
健保就診總人次    111319 non-null int64  
dtypes: int64(4), object(3)  
memory usage: 5.9+ MB
```