

Python 系列 – MIMIC-IV Demo 描述性統計

鄭哲宇 副統計分析師

進行任何資料分析工作時，我們都需要先載入並且預先處理好資料，取得正確並且整齊的資料，是資料分析不可或缺的第一個步驟。上期 eNews 使用 MIMIC-IV Demo 資料作為範例，介紹如何如何使用 Python 讀取資料及簡單處理資料的方法。本期 eNews 將接著示範使用 Python 中常用的函數，計算資料的描述性統計。藉由觀察資料的描述性統計，我們可以檢驗資料的正確性，例如資料的分布情形是否符合我們對資料的預期，同時透過描述性統計了解欲分析資料的樣貌。

第 1 部分 下載並載入 MIMIC-IV Demo data


1.1 下載 MIMIC-IV Demo data

eNews 第 48 期中，對 MIMIC 資料庫有詳細的介紹，並提到目前 MIMIC 官方有釋出 100 筆病患資料做為 demo 檔案。讀者可以前往下列網址：<https://physionet.org/content/mimic-iv-demo-omop/0.9/>，捲動至網頁下半部後，點選”Download the ZIP file”來下載資料檔。

Files

Total uncompressed size: 73.0 MB.

Access the files

 [Download the ZIP file \(10.3 MB\)](#)

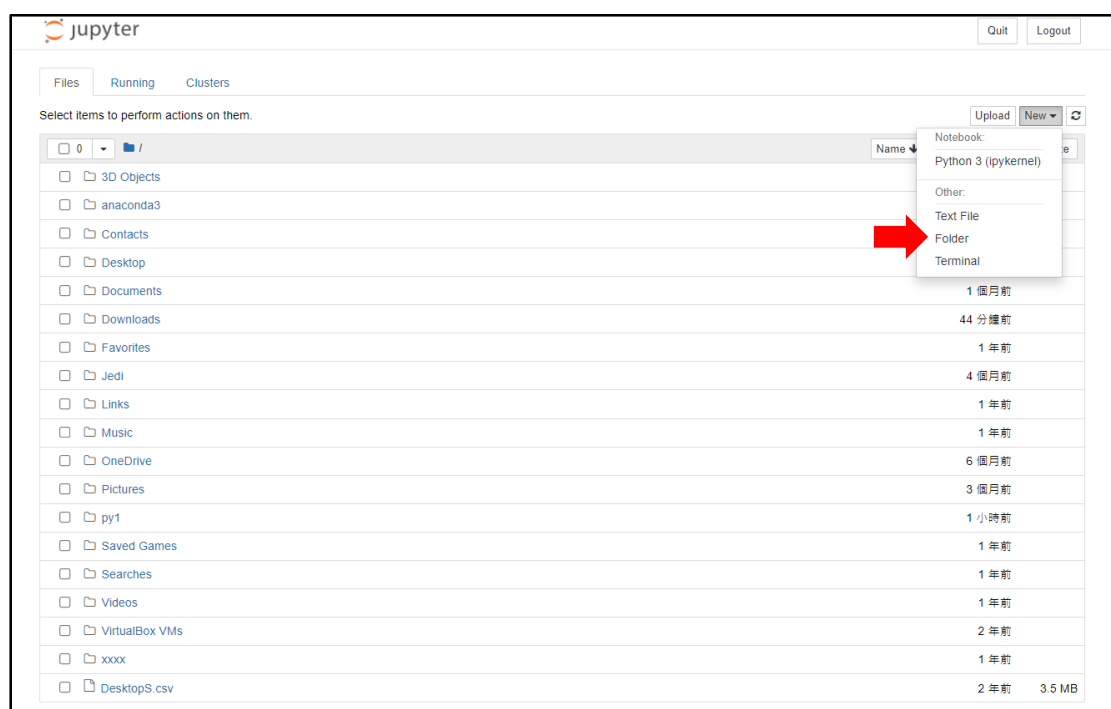
- Access the files using the Google Cloud Storage Browser [here](#). Login with a Google account is required.
- Access the data using the Google Cloud command line tools (please refer to the [gsutil](#) documentation for guidance):

```
gsutil -m -u YOUR_PROJECT_ID cp -r gs://mimic-iv-demo-omop-0.9.physionet.org DESTINATION
```
- [Request access](#) using Google BigQuery.
- Download the files using your terminal: `wget -r -N -c -np https://physionet.org/files/mimic-iv-demo-omop/0.9/`

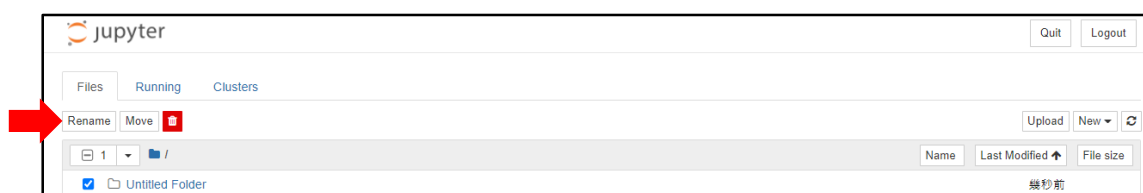
1.2 開啟 Jupyter Notebook

本文將使用 eNews 第 47 期所介紹安裝的 Jupyter Notebook 作為示範 Python 程式碼的界面，並使用 eNews 第 49 期介紹的 Python 程式庫 Pandas 來讀取資料及進行資料分析。建議尚未安裝 Jupyter Notebook 的讀者可先參考 eNews 第 47 期介紹方式安裝，若想瞭解 Pandas 讀取及檢視資料的方法，可參考 eNews 第 49 期。

1. 開啟 Jupyter Notebook 後，可依照下列步驟建立存程式碼的新資料夾。



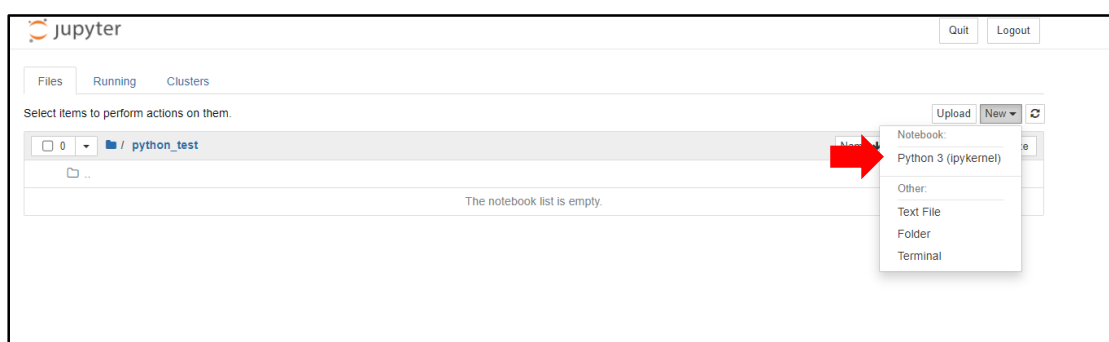
2. 建立新資料夾



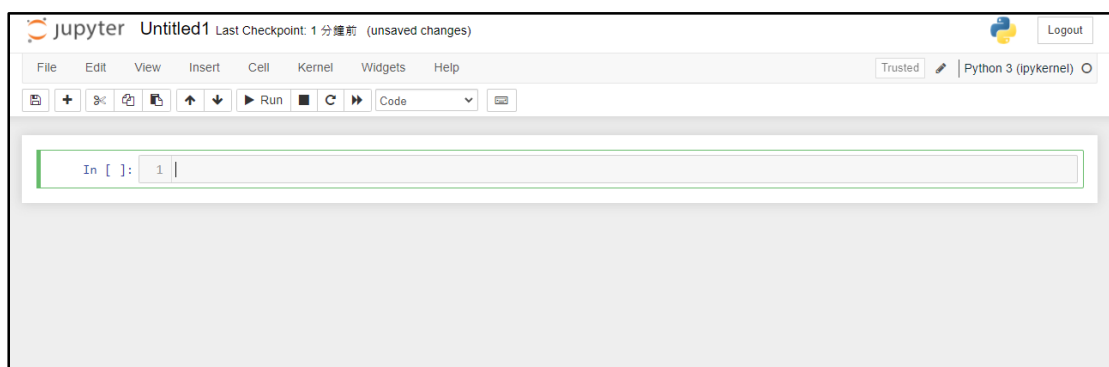
3. 重新命名新資料夾



4. 建立新資料夾後，就可以新增 Python 程式碼檔案



5. 建立新的程式碼之後，可以開始撰寫 Python 程式碼



6. 匯入 pandas 程式庫

```
1 #匯入pandas程式庫
2 import pandas as pd
```

1.3 載入 MIMIC-IV Demo data

Pandas 程式庫當中，提供許多工具方便讀取各種資料格式，本文以 MIMIC-IV Demo data 作為範例，示範如何讀取 CSV 格式的資料檔。如下圖，在函數 `pd.read.csv()` 中，放入欲讀取檔案之路徑，就可以在 Python 中順利讀取資料夾中的 `drug_exposure.csv` 檔案，並將資料儲存為一個名叫 `drug_exposure` 的變數。

單純執行變數 `drug_exposure`，可以看到程式中列出了我們剛剛讀取的檔案內容，如果沒有正確顯示，則要再回頭檢查 `pd.read.csv()` 中，填寫的檔案之路徑是否有錯誤。`drug_exposure` 資料檔案中，儲存了資料中每一筆用藥紀錄、劑量等等相關欄位。

```

1 #載入路徑中的資料
2 drug_exposure = pd.read_csv("C:/Users/biostat/Desktop/1_omop_data_csv/drug_exposure.csv")

```

```

1 #檢視drug_exposure變數
2 drug_exposure

```

	drug_exposure_id	person_id	drug_concept_id	drug_exposure_start_date	drug_exposure_start_datetime	drug_exposure_end_date	drug_e
0	294884377115777655	1741351032930224901	40166274	2177-07-16	2177-07-16 22:00:00	2177-07-17	
1	-3609243742606366340	1741351032930224901	40166274	2177-07-17	2177-07-17 19:00:00	2177-07-18	
2	-6865345241721388581	1741351032930224901	40166274	2177-07-15	2177-07-15 19:00:00	2177-07-16	
3	-826223020394544622	1741351032930224901	40166274	2177-07-21	2177-07-21 22:00:00	2177-07-22	
4	2417954811860157314	1741351032930224901	40166274	2177-07-18	2177-07-18 23:00:00	2177-07-19	
...
18224	-5144476626986792845	3192038106523208432	36249735	2136-08-05	2136-08-05 01:00:00	2136-08-05	
18225	2583812317283757784	3192038106523208432	36249735	2136-08-08	2136-08-08 20:00:00	2136-08-11	
18226	5344866293695870763	3192038106523208432	19127213	2136-08-05	2136-08-05 01:00:00	2136-08-11	
18227	-8090179442343534710	3192038106523208432	19127213	2136-08-09	2136-08-09 09:00:00	2136-08-11	
18228	923805161291343268	3192038106523208432	19127213	2136-08-09	2136-08-09 21:00:00	2136-08-11	

18229 rows x 23 columns

第 2 部分 MIMIC-IV Demo data 描述性統計

2.1 篩選資料欄位

當我們讀取原始資料時，原始資料可能包含了許多我們不需要的資訊，又或者當我們想要從眾多的資料中，找出符合特定條件的某些觀察值時，便可以使用 Pandas 程式庫中資料框的篩選功能。

首先，我們使用下列程式碼篩選出 `drug_exposure` 資料框中，`drug_concept_id` 和 `quantity` 兩個欄位，並儲存為名叫 `data1` 的資料框。接著，在 `data1` 資料框中，篩選出 `drug_concept_id` 為 711620 資料，儲存為名叫 `data2` 的資料框。關於資料框中，`drug_concept_id` 為 711620 資料代表的詳細意義，可以到 OMOP-CDM 通用資料模型的官方網站 <https://www.ohdsi.org/data-standardization/the-common-data-model/>，內有更詳盡的說明。

```
1 data1 = drug_exposure[['drug_concept_id', 'quantity']]
2
```

```
1 data2 = data1[data1['drug_concept_id']==711620]
2
3 data2
```

	drug_concept_id	quantity
167	711620	7.5
168	711620	5.0
169	711620	10.0
170	711620	5.0
171	711620	7.5
172	711620	10.0
173	711620	10.0
708	711620	5.0
709	711620	5.0

2.2 描述性統計

2.2.1 概括描述統計

Pandas 程式庫中提供了一個簡便的方法，讓我們可以快速瞭解資料的描述性統計，我們可以使用 `.describe()` 來檢視 `data2` 資料框中，`quantity` 的描述性統計如下圖，執行的結果依序為：`count` 資料筆數、`mean` 平均數、`std` 標準差、`min` 最小值、`25%`二十五百分位數、`50%`五十百分位數、`75%`七十五百分位數、`max` 最大值。

```
1 data2['quantity'].describe()
count    35.000000
mean     7.500000
std      2.572479
min      5.000000
25%      5.000000
50%      7.500000
75%     10.000000
max     15.000000
Name: quantity, dtype: float64
```

2.2.2 平均數、中位數、眾數

我們可以使用下列程式碼來分別檢視 `data2` 資料框中，`quantity` 的平均數、中位數、眾數的數值如下圖。

```
1 #平均數
2 data2['quantity'].mean()
7.5
```

```
1 #中位數
2 data2['quantity'].median()
7.5
```

```
1 #眾數
2 data2['quantity'].mode()
0    5.0
dtype: float64
```

2.2.3 變異數、標準差

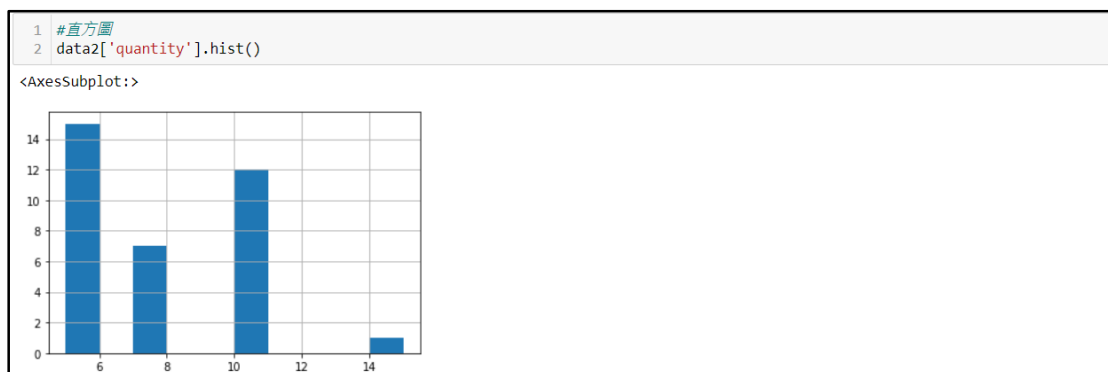
我們可以使用下列程式碼來分別檢視 `data2` 資料框中，`quantity` 的變異數、標準差的數值如下圖。

```
1 #變異數
2 data2['quantity'].var()
6.617647058823529
```

```
1 #標準差
2 data2['quantity'].std()
2.572478771376323
```

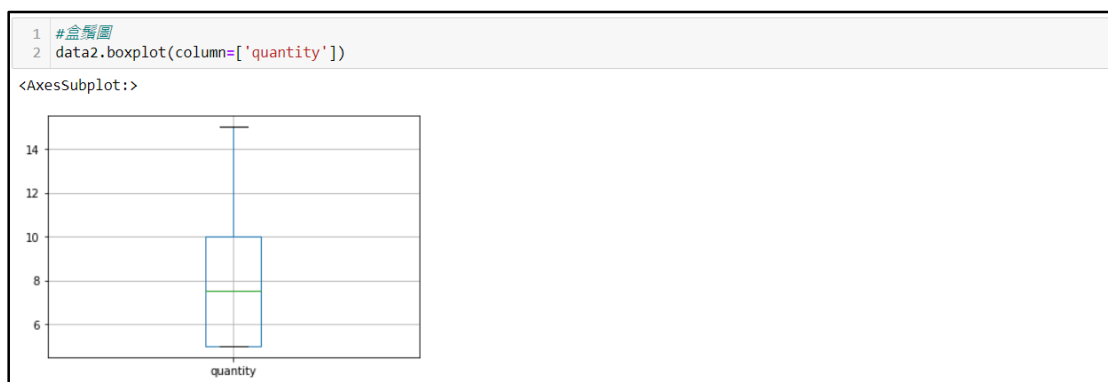
2.2.4 直方圖

除了描述統計的數字外，我們可以應用資料視覺化的方式，將數據轉化為圖表，讓我們更直覺得瞭解資料的分布情形。例如我們可以將 data2 資料框中 quantity 變數，應用下列程式碼繪製直方圖。



2.2.5 盒鬚圖

盒鬚圖使用四分位數來描繪出資料的分布情形，盒型本體的上限表示資料的第三個四分位數，而盒型本體的下限則是資料的第一個四分位數，盒型的中線標示出資料的中位數。我們可以使用下列程式碼繪製出盒鬚圖。



第 3 部分 總結

本文以 MIMIC-IV Demo 資料為範例，示範從資料下載，到使用 python 程式碼讀取、檢視資料的步驟，並且介紹如何使用 Pandas 程式庫計算資料的描述性統計。通過檢視描述性統計輔以長條圖、盒鬚圖等視覺化圖形，我們可以首先掌握欲分析資料的分布情形，並且檢查資料中是否有不合理或者超乎預期得數值。因此檢視描述性統計是所有分析最重要的第一步，也可以確保我們使用正確的資料，再進行更進一步的資料分析。