

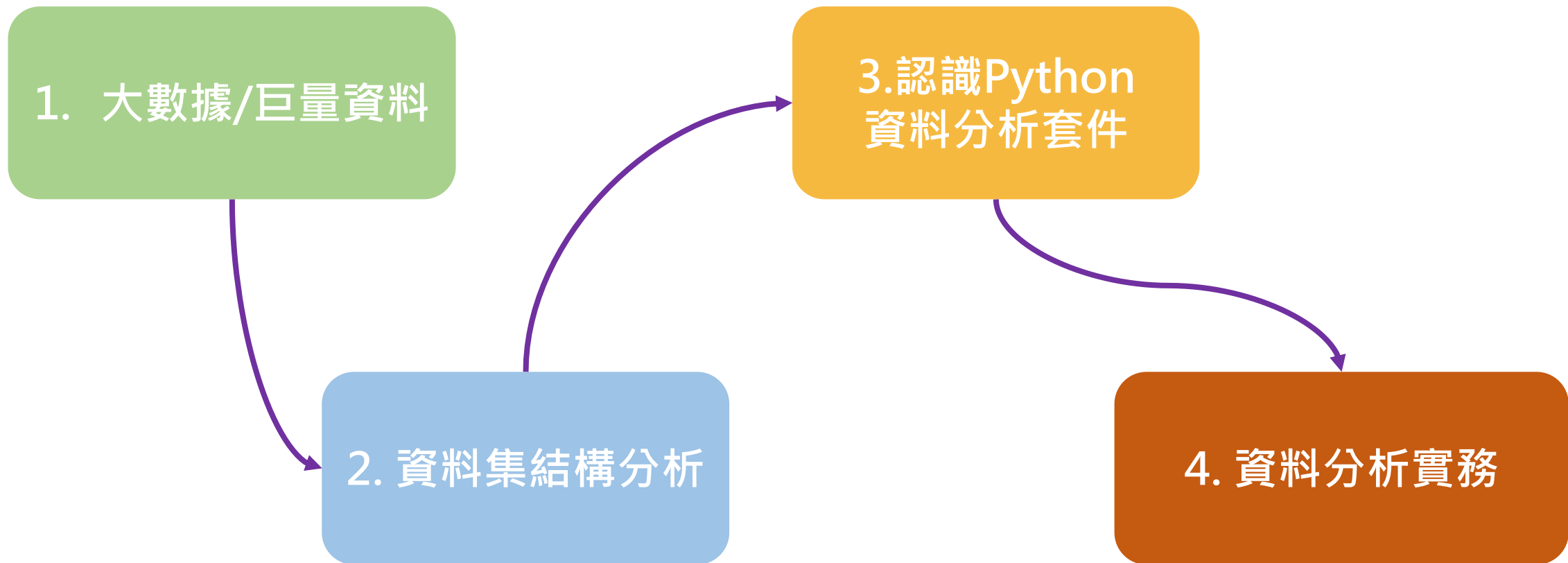
資料分析

SDPMLab, NCKU

軟體開發與流程管理實驗室

國立高雄師範大學

課程大綱



大數據/巨量資料

SDPMLab, NCKU

軟體開發與流程管理實驗室

國立高雄師範大學

特徵	解釋
種類(Variety)	意指大數據多樣化的資料類型，數據來源多樣，包括結構化數據、非結構化數據和半結構化數據。
數量 (Volume)	意指資料量，巨型資料動輒數十TB，甚至數百PB
速度 (Velocity)	是指接收資料的速率 (有時也含括處理資料的速率)。

大數據/巨量資料 - 簡介

大數據或巨量資料是指更龐大且複雜的資料集。種類更多樣化 (**variety**)、數量不斷增加 (**volume**) 且產生速度越來越快 (**velocity**) 的數據。以上三個特徵又稱為「三個 V」。

大數據/巨量資料 – 優點與挑戰

優點：

- 大數據能讓您擁有更充足的資訊，進而獲得更完整的答案。
- 更完整的答案代表資料更為有用，因為可以採用各種不同的方法來解決問題。

挑戰：

- 巨量資料的資料量無比龐大，如何提升資料儲存速度及效率。
- 從巨量資料內獲得「乾淨的資料」需耗費大量成本，資料科學家平均須投入 50% 到 80% 的時間整理並準備資料，這樣資料才能真正派上用場。
- 總結以上，巨型資料的挑戰面臨**整合、管理、分析**，三大挑戰。

大數據/巨量資料 – 應用領域

大數據/巨量資料現今使用領域與案例非常廣，例如：

- 機器學習，訓練模型。
- 客戶體驗，更加清楚地洞察客戶體驗，更容易的保留客戶族群。
- 產品開發，預測客戶需求來規劃、生產及啟動新產品。
- 預測性維護，在發生問題前及早分析故障跡象。



資料集 結構分析

SDPMLab, NKNU
軟體開發與流程管理實驗室
國立高雄師範大學

資料集結構分析 – 資料說明



- 請同學至該連結 <https://shorturl.at/jpC01> 下載 edu_bigdata_imp1.csv 與教學用開放資料說明.pdf 檔案。
- 該檔案為**結構化資料**檔案，欄位明確，資料內容完整度尚可。
- 該檔案為教育型受眾分析資料，內部包含了 397 位使用者 (共 **287,137** 筆資料) 在某平台上學習時各類操作的紀錄 (log)，未包含任何隱私資料。

資料集結構分析 - 了解資料欄位

- 請同學開啟edu_bigdata_mp1.csv 與教學用開放資料說明.pdf 檔案。
- 開啟CSV檔後可能會發現檔案中部分欄位值有
"#####" 符號存在，只需將欄位拉大即可解決該問題。
- CSV檔為學生於學習平台上所觀看的學習影片與問題作答之紀錄，檔案欄位可搭配教學用開放資料說明.pdf 查看。
- CSV檔案資料來自於兩個不同的學習平台，即 dp001 與 dp002，我們可依據欄位的名稱判斷資料來自哪個平臺。



資料集結構分析 - 了解資料欄位

資料說明範例：

- PseudoID：為每個學生的獨立編號，可作為條件，判斷該生相關操作行為。
- dp001_review_sn：為教學影片唯一編號，可搭配PseudoID找出學生以學習之影片。

1. 使用者基本資料

- PseudoID: 使用者流水號

2. dp001:影片基本資料

- dp001_indicator: 能力指標 (格式: X-X-XX-YYY 其中 X 為列指標)
- dp001_video_item_sn: 影片 ID

3. dp001: 使用者瀏覽影片時的基本資料

- dp001_review_sn: 影片瀏覽序號
- dp001_review_start_timestamp: 影片瀏覽開始的秒數
- dp001_review_end_timestamp: 影片瀏覽結束的秒數
- dp001_review_start_time: 影片瀏覽開始的時間戳記
- dp001_review_end_time: 影片瀏覽結束的時間戳記
- dp001_review_finish_rate: 影片瀏覽的進度
- dp001_review_seme_year: 影片瀏覽學年度

認識Python資料 分析套件

SDPMLab, NCKU

軟體開發與流程管理實驗室

國立高雄師範大學

認識Python資料分析套件

- Pandas為基於**NumPy**所產生的Python軟體庫，主要用於**資料操縱**和**分析**，並提供資料結構與運算操作方法。
- 在Pandas中主要有兩種資料結構，分別為**Series** 和 **DataFrame**。
- 支持多種資料格式的導入和導出，如 CSV, Excel, SQL, JSON 等。
- 可針對缺失、重複、異常資料做處理。
- 提供多種API可供使用者操作。

Pandas



認識Python資料分析套件 - 資料結構

Series:

- 一維數據結構：Series 是一個一維的數據結構，用來存儲一維的數據（類似於一維陣列）。
- 帶標籤：每個元素都有一個唯一的標籤（或稱為索引），可以是數字或其他自定義的標籤。
- 數據類型單一：Series 中的所有數據都應屬於同一數據類型（如整數、浮點數、字符串等）。

```
import pandas as pd  
s = pd.Series([1, 2, 3, 4])  
print(s)
```

```
0    1  
1    2  
2    3  
3    4  
dtype: int64
```

Pandas



認識Python資料分析套件 - 資料結構

DataFrame :

- 二維數據結構：DataFrame 是一個二維的表格型數據結構，**具有行和列的標籤**。
- 可存儲多種數據類型：DataFrame 的不同列**可以存儲不同類型的數據**（如整數、字符串、日期等）。
- 靈活操作：DataFrame 支持各種操作，如選擇、過濾、排序、合併、分組、聚合等
- 類似於 Excel 表格：DataFrame 的形式和 Excel 的表格或 SQL 的表非常相似，容易理解和使用。

```
import pandas as pd
taiwan = {'city': ['台北市', '新北市', '桃園市', '台中市', '台南市', '高雄市'],
          'pop': [2631083, 4024539, 2255753, 2816741, 1878845, 2773401],
          'area': [271.7997, 2052.5667, 1220.9540, 2214.8968, 2191.6531, 2951.8524],}
s = pd.DataFrame(taiwan)
print(s)
```

	city	pop	area
0	台北市	2631083	271.7997
1	新北市	4024539	2052.5667
2	桃園市	2255753	1220.9540
3	台中市	2816741	2214.8968
4	台南市	1878845	2191.6531
5	高雄市	2773401	2951.8524

Pandas



認識 Python資料分析套 件 - 常用 方法

讀取檔案：read_csv(檔案位置 , 其他設定)：用於讀取檔案，讀取後的檔案格式為DataFrame，此外還提供read_excel等相關方法。

取回資料：類似於dist的取值方式，使用DataFrame[欄位名稱]，即可獲得該欄位的所有資料，也可搭配布林判斷實作資料篩選。

過濾不存在之資料：欄位中多少會有缺失或缺少數值，可透過dropna()方法過濾這些資料。

尋找不重複的資料：drop_duplicates()或unique()方法皆可快速過濾出欄位中不重複的數值，其中drop_duplicates可對DataFrame和Series結構操作，unique只能對Series操作。

計算資料筆數：value_counts()可針對Series結構進行筆數計算，並返回Series結構。

認識Python資料 分析套件 - 常用 方法

取得最大值的索引：`idmax()`，該方法可取得DataFrame和Serise的最大索引。

取得最大值：`max()`，該方法可取得DataFrame和Serise中的最大值。

日期轉換：`to_datetime()`，欄位值有時會包含日期結構的純文字，此時可利用`to_datetime`方法將字串轉為日期型別。

取得日期：`dt.date`，該方法主要使用於Serise結構中的日期型別，並取得日期部分(即年月日)。

更多方法可參閱

<https://pandas.pydata.org/docs/reference/index.html>

認識Python資料分析套件 – 常用方法

讀取檔案

```
import pandas as pd
df = pd.read_csv('/content/edu_bigdata_impl.csv', encoding='big5')
```

過濾不存在之資料

```
import pandas as pd
import numpy as np
taiwan = {'city': ['台北市', '新北市', '桃園市', '台中市', '台南市', np.nan],
          'pop': [2631083, 4024539, 2255753, 2816741, 1878845, np.nan],
          'area': [271.7997, 2052.5667, 1220.9540, 2214.8968, 2191.6531, np.nan],}
s = pd.DataFrame(taiwan)
s = s.dropna()
print(s)
```

	city	pop	area
0	台北市	2631083.0	271.7997
1	新北市	4024539.0	2052.5667
2	桃園市	2255753.0	1220.9540
3	台中市	2816741.0	2214.8968
4	台南市	1878845.0	2191.6531

取回資料

```
import pandas as pd
taiwan = {'city': ['台北市', '新北市', '桃園市', '台中市', '台南市', '高雄市'],
          'pop': [2631083, 4024539, 2255753, 2816741, 1878845, 2773401],
          'area': [271.7997, 2052.5667, 1220.9540, 2214.8968, 2191.6531, 2951.8524],}
s = pd.DataFrame(taiwan)
print(s['city'])
```

```
0  台北市
1  新北市
2  桃園市
3  台中市
4  台南市
5  高雄市
Name: city, dtype: object
```

尋找不重複的資料

```
import pandas as pd
import numpy as np
taiwan = {'city': ['台北市', '台北市', '桃園市', '台中市', '台南市'],
          'pop': [2631083, 4024539, 2255753, 2816741, 1878845],
          'area': [271.7997, 2052.5667, 1220.9540, 2214.8968, 2191.6531],}
s = pd.DataFrame(taiwan)
s = s['city'].unique()
print(s)
```

```
['台北市' '桃園市' '台中市' '台南市']
```

認識Python資料分析套件 – 常用方法

計算資料筆數

```
import pandas as pd

s = pd.Series(['cat', 'dog', 'cat', 'dog', 'bird', 'cat'])

value_counts = s.value_counts()

print(value_counts)
```

```
cat    3
dog    2
bird   1
dtype: int64
```

```
import pandas as pd

df = pd.DataFrame({
    'A': [1, 5, 3, 8],
    'B': [3, 2, 7, 6],
    'C': [2, 9, 4, 5]
})

# 使用 max() 找到每列的最大值
max_values_in_columns = df.max()

print(f"The maximum values in the columns are:\n{max_values_in_columns}")
```

```
The maximum values in the columns are:
A    8
B    7
C    9
dtype: int64
```

取得最大值的索引

```
import pandas as pd

df = pd.DataFrame({
    'A': [1, 5, 3, 8],
    'B': [3, 2, 7, 6],
    'C': [2, 9, 4, 5]
})

# 使用 idxmax() 找到每列中最大值的索引
index_of_max_in_columns = df.idxmax()

print(f"The indices of the maximum values in the columns are:\n{index_of_max_in_columns}")
```

```
The indices of the maximum values in the columns are:
A    3
B    2
C    1
dtype: int64
```

取得最大值

認識Python資料分析套件 - 常用方法

日期轉換

```
import pandas as pd

date = pd.to_datetime('2023-10-19')
print(date)
```

2023-10-19 00:00:00

取得日期

```
import pandas as pd

s = pd.Series(pd.to_datetime(['2023-10-19 08:00:00', '2023-10-19 09:00:00', '2023-10-19 10:00:00']))

# 使用 .dt.date 取得日期
dates = s.dt.date

# 結果
print(dates)
```

```
0    2023-10-19
1    2023-10-19
2    2023-10-19
dtype: object
```



資料分析實務

SDPMLab, NKNUN

軟體開發與流程管理實驗室

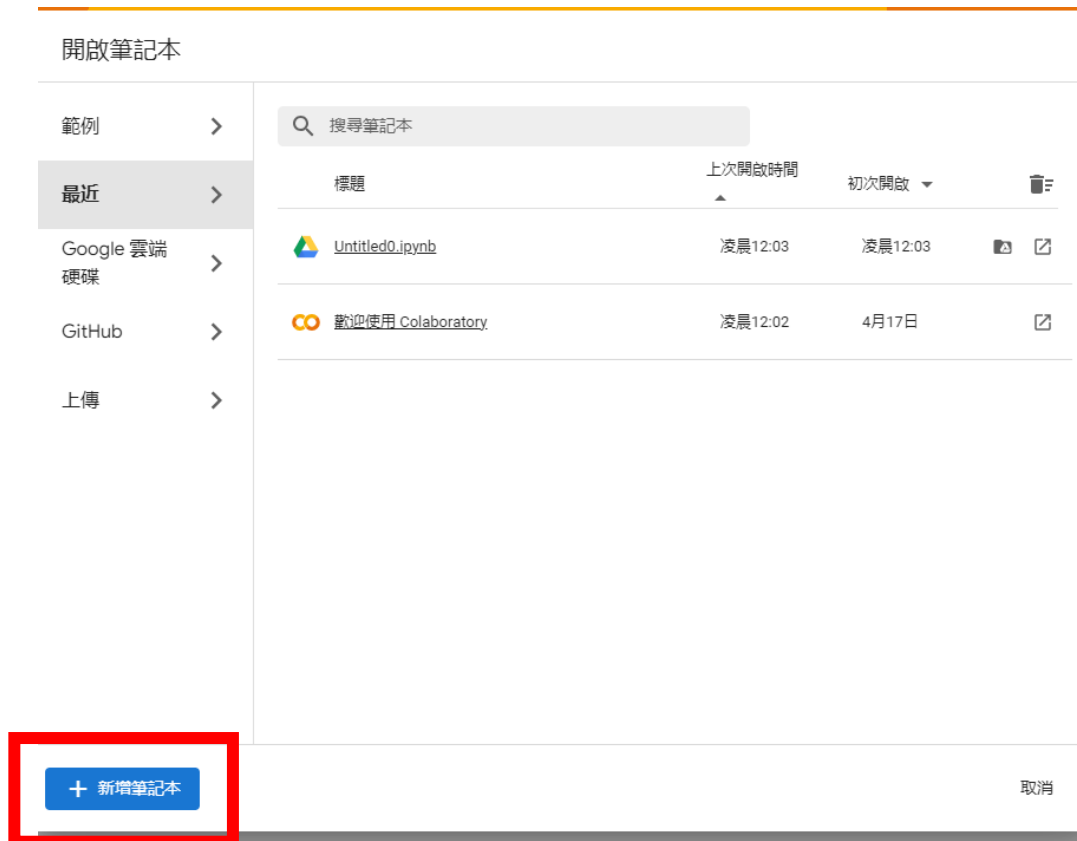
國立高雄師範大學

資料分析實務

- 請同學下載 edu_bigdata_imp1資料檔練習題.pdf 檔案。
- 檔案內總共有10題練習題，我們將使用Python於Colab平台上進程式撰寫。
- 請同學使用瀏覽器開啟 <https://colab.research.google.com/?hl=zh-tw> 或直接搜尋“google colab”。

資料分析實務 – 使用Colab

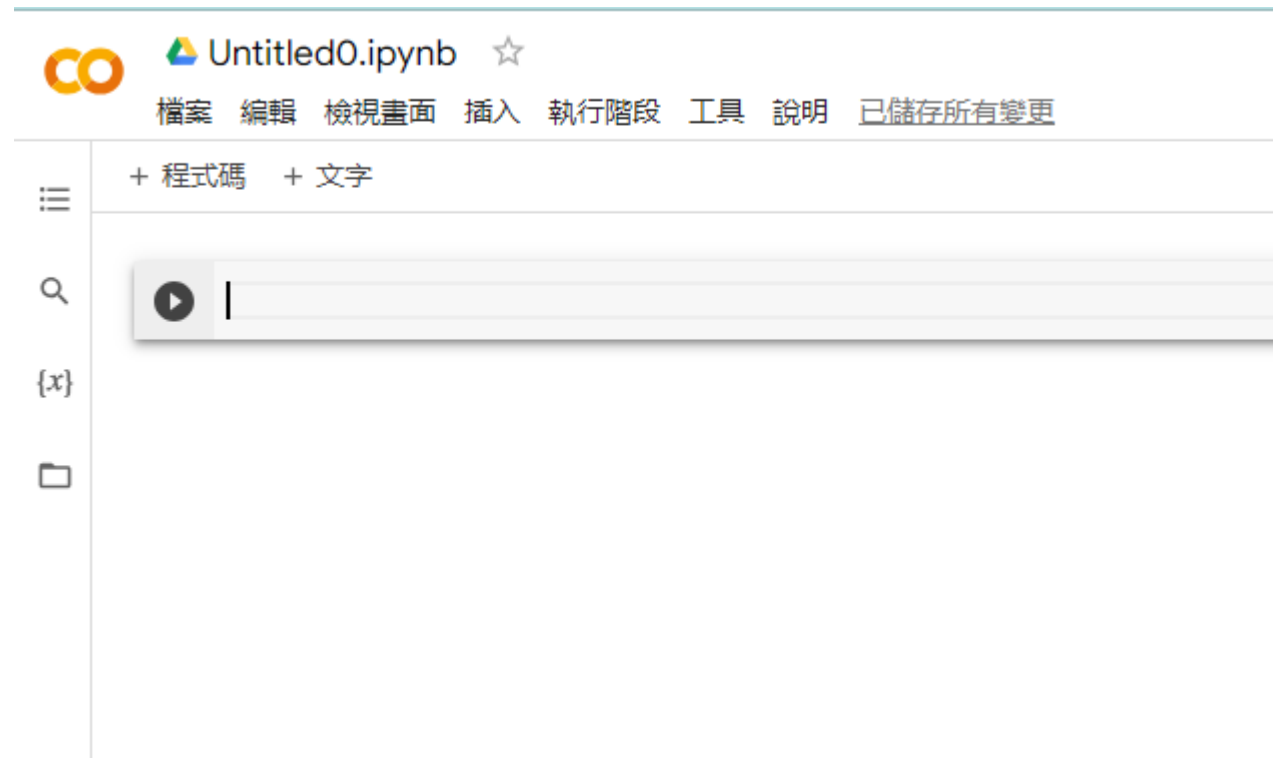
STEP 1 : 新增筆記本



The screenshot shows the 'Open Notebook' (開啟筆記本) interface in Google Colab. On the left, there is a sidebar with navigation options: '範例' (Examples), '最近' (Recent), 'Google 雲端硬碟' (Google Drive), 'GitHub', and '上傳' (Upload). The '最近' section is active, displaying a table of recent notebooks. The table has columns for '標題' (Title), '上次開啟時間' (Last opened time), and '初次開啟' (First opened). Two notebooks are listed: 'Untitled0.ipynb' and '歡迎使用 Colaboratory'. At the bottom left, a blue button with a plus icon and the text '+ 新增筆記本' (Add new notebook) is highlighted with a red rectangle. At the bottom right, there is a '取消' (Cancel) link.

標題	上次開啟時間	初次開啟
Untitled0.ipynb	凌晨12:03	凌晨12:03
歡迎使用 Colaboratory	凌晨12:02	4月17日

STEP 2 : 獲得開發環境



The screenshot shows the Google Colab development environment for a notebook titled 'Untitled0.ipynb'. The top bar includes the Colab logo, the notebook title, and a star icon. Below this is a menu bar with options: '檔案' (File), '編輯' (Edit), '檢視畫面' (View), '插入' (Insert), '執行階段' (Runtime), '工具' (Tools), '說明' (Help), and '已儲存所有變更' (Save all changes). The main workspace is divided into two sections: '+ 程式碼' (Code) and '+ 文字' (Text). The 'Code' section is active, showing a code editor with a play button icon and a cursor. The left sidebar contains icons for navigation: a menu icon, a search icon, a variable icon '{x}', and a file icon.

資料分析實務 – 使用Colab

STEP 3 : 上傳edu_bigdata_mp1.csv 至Colab，後續將使用到該檔案



題號	題目	答案
請依據 39 號學生的行為紀錄回答下列問題		
1.1	於 dp001 平臺總共進行幾次 <u>不重複</u> 的影片瀏覽的學習紀錄?	
1.2	於 dp001 平臺瀏覽影片時，總共進行幾次 <u>不重複</u> 的檢核點作答?	

資料分析實務 － 程式撰寫

- 我們將以 edu_bigdata_imp1資料檔練習題.pdf 的題目進行解題。
- 簡報中將以第一大題為範例，帶各位進行初步分析。

資料分析實務 – 程式撰寫

問題 1-1

將問題進行簡化，理解成白話文就是，**39號學生看過幾部不同的學習影片**，並用影片編號和學生編號進行實作。

解題思路

1. 讀取csv檔案。
2. 先過濾出39號學生的資料。
3. 從39號學生的資料中，以影片編號進行不重複的搜尋。
4. 即得所解。

```
# 1-1

# 引入Pandas
import pandas as pd

# 讀取csv檔案
df = pd.read_csv('/content/edu_bigdata_impl.csv', encoding='big5', low_memory=False)

# 根據 'PseudoID' 為 39 過濾 DataFrame
df_filtered = df[df['PseudoID'] == 39]

# 從過濾後的 DataFrame 中找到 'dp001_review_sn' 欄位的不重複值
unique_values = df_filtered['dp001_review_sn'].unique()

# 印出不重複值得長度，即可得知39號學生看過幾部影片
print(len(unique_values))
```

資料分析實務 – 程式撰寫

- 首先引入pandas套件
- 使用pandas的read_csv方法讀取檔案，並由一個參數承接讀取後的DataFrame。
- 過濾出PseudoID為39的學生。
- 將過濾後的DataFrame中的dp001_review_sn欄位進行不重複值的過濾。
- 印出結果長度。

資料分析實務 – 程式撰寫

問題 1-2

將問題進行簡化，理解成白話文就是，**39號學生做過幾個不同的檢核做答**，並用檢核點試題編號和學生編號進行實作。

解題思路

1. 讀取csv檔案。
2. 先過濾出39號學生的資料。
3. 排除欄位中包含NA或NAN的無效資料
4. 從39號學生的資料中，以檢核點試題編號進行不重複的搜尋。
5. 即得所解。

```
# 1-2
import pandas as pd

df = pd.read_csv('/content/edu_bigdata_impl.csv', encoding='big5', low_memory=False)

df_filtered = df[df['PseudoID'] == 39]

# 過濾掉 'dp001_question_sn' 欄位中包含 NaN 值的行
df_filtered = df_filtered.dropna(subset=['dp001_question_sn'])

# 從過濾後的 DataFrame 中找到 'dp001_question_sn' 欄位的不重複值
unique_values = df_filtered['dp001_question_sn'].unique()

# 打印出不重複值
print(len(unique_values))
```

資料分析實務 – 程式撰寫

- 首先引入pandas套件
- 使用pandas的read_csv方法讀取檔案，並由一個參數承接讀取後的DataFrame。
- 過濾出PseudoID為39的學生。
- 將首次過濾後的資料，使用dropna對dp001_question_sn欄位進行無效值過濾。
- 再將DataFrame中的 dp001_question_sn 欄位進行不重複值的過濾。
- 印出結果長度。

SDPMLab, NKTU

軟體開發與流程管理實驗室

國立高雄師範大學

實作時間



資料分析實務 –實作時間

1. 請實作第一大題並將解答紀錄下來。
2. 可參考第一大題的實作過程，接續實作第二大題並將解答紀錄下來。



請依據 281 號學生的行為紀錄回答下列問題

2.1 於 dp001 平臺總共瀏覽過哪些不重複的影片且對應的能力指標為何？

2.2 於 dp001 平臺共有幾次的練習題作答紀錄正確率是 100？

資料分析實務一 第二大題解析

2-1 解析

- 過濾出PseudoID為281的學生資料。
- 題目要求取出不重複的影片，透過影片編號進行不重複過濾。
- 印出影片編號與其對應的能力指標。

請依據 281 號學生的行為紀錄回答下列問題

2.1 於 dp001 平臺總共瀏覽過哪些不重複的影片且對應的能力指標為何?

2.2 於 dp001 平臺共有幾次的練習題作答紀錄正確率是 100?

資料分析實務 – 第二大題解析

2-2 解析

- 過濾出PseudoID為281的學生資料與練習正確率欄位為100的資料。
- 印出結果長度即為答案。

SDPMLab, NCKU

軟體開發與流程管理實驗室

國立高雄師範大學

實作時間





資料分析實務－實作時間

- 請搭配簡報中**常用方法**章節介紹的幾個方法實作第三大題。

請依據 71 號學生的行為紀錄回答下列問題

3.1 於 dp001 平臺的瀏覽影片時，操作行為名稱為「暫停」總共有幾次？

3.2 分別於哪幾天進入 dp001 平臺？

資料分析實務 – 第三大題解析

3-1 解析

- 以PseudoID為71的學生資料與影片操作的行為名稱為“暫停 (資料內為 **paused**)” 進行過濾。
- 印出過濾結果的長度。

請依據 71 號學生的行為紀錄回答下列問題

3.1 於 dp001 平臺的瀏覽影片時，操作行為名稱為「暫停」總共有幾次？

3.2 分別於哪幾天進入 dp001 平臺？

資料分析實務 – 第三大題解析

3-2 解析

- 過濾出PseudoID為71的學生資料。
- 將影片瀏覽開始的時間與影片瀏覽結束的時間進行類型轉換。
- 將上述欄位轉換為年月日格式。
- 找出不重複的值
- 印出結果。

實作時間

SDPMLab, NKTU

軟體開發與流程管理實驗室

國立高雄師範大學



資料分析實務－實作時間

- 請搭配簡報中**常用方法**章節介紹的幾個方法實作第四大題。



請依據全體學生的行為紀錄回答下列問題

- 4.1 請找出在 dp001 平臺中，最多影片瀏覽行為的影片序號
-
- 4.2 請找出在 dp002 平臺中，操作資源的知識架構分類中為「十二年國民基本教育類」總共有幾筆？
-
- 4.3 請找出在 dp002 平臺中，前 3 個最常發生的操作行為名稱
-
- 4.4 請找出在 dp002 平臺中，操作資源的知識架構分類中為「校園職業安全」總共有幾筆？
-

資料分析實務一 第四大題解析

4-1 解析

- 取得影片瀏覽序號欄位，相同值的資料筆數。
- 取得出現次數最多的值及出現次數。
- 印出結果

請依據全體學生的行為紀錄回答下列問題

- 4.1 請找出在 dp001 平臺中，最多影片瀏覽行為的影片序號
- 4.2 請找出在 dp002 平臺中，操作資源的知識架構分類中為「十二年國民基本教育類」總共有幾筆？
- 4.3 請找出在 dp002 平臺中，前 3 個最常發生的操作行為名稱
- 4.4 請找出在 dp002 平臺中，操作資源的知識架構分類中為「校園職業安全」總共有幾筆？

資料分析實務一 第四大題解析

4-2 解析

- 過濾出操作資源的知識架構分類為十二年國民基本教育類的資料。
- 印出結果長度即為答案。
- 備註：請注意資料中操作資源的知識架構分類的值，欄位中的值有標點符號存在，過濾時需注意。

請依據全體學生的行為紀錄回答下列問題

4.1 請找出在 dp001 平臺中，最多影片瀏覽行為的影片序號

4.2 請找出在 dp002 平臺中，操作資源的知識架構分類中為「十二年國民基本教育類」總共有幾筆？

4.3 請找出在 dp002 平臺中，前 3 個最常發生的操作行為名稱

4.4 請找出在 dp002 平臺中，操作資源的知識架構分類中為「校園職業安全」總共有幾筆？

資料分析實務一 第四大題解析

4-3 解析

- 從**操作行為名稱**欄位過濾掉NA或NAN的資料。
- 為欄位中的值進行計數。
- 印出出現次數前三多的值及其出現次數

請依據全體學生的行為紀錄回答下列問題

4.1 請找出在 dp001 平臺中，最多影片瀏覽行為的影片序號

4.2 請找出在 dp002 平臺中，操作資源的知識架構分類中為「十二年國民基本教育類」總共有幾筆？

4.3 請找出在 dp002 平臺中，前 3 個最常發生的操作行為名稱

4.4 請找出在 dp002 平臺中，操作資源的知識架構分類中為「校園職業安全」總共有幾筆？

資料分析實務 – 第四大題解析

4-4 解析

- 過濾出操作資源的知識架構分類為校園職業安全的資料。
- 印出結果長度即為答案。
- 備註：請注意資料中操作資源的知識架構分類的值，欄位中的值有標點符號存在，過濾時需注意。

資料分析

SDPMLab, NKNUN

軟體開發與流程管理實驗室

國立高雄師範大學