

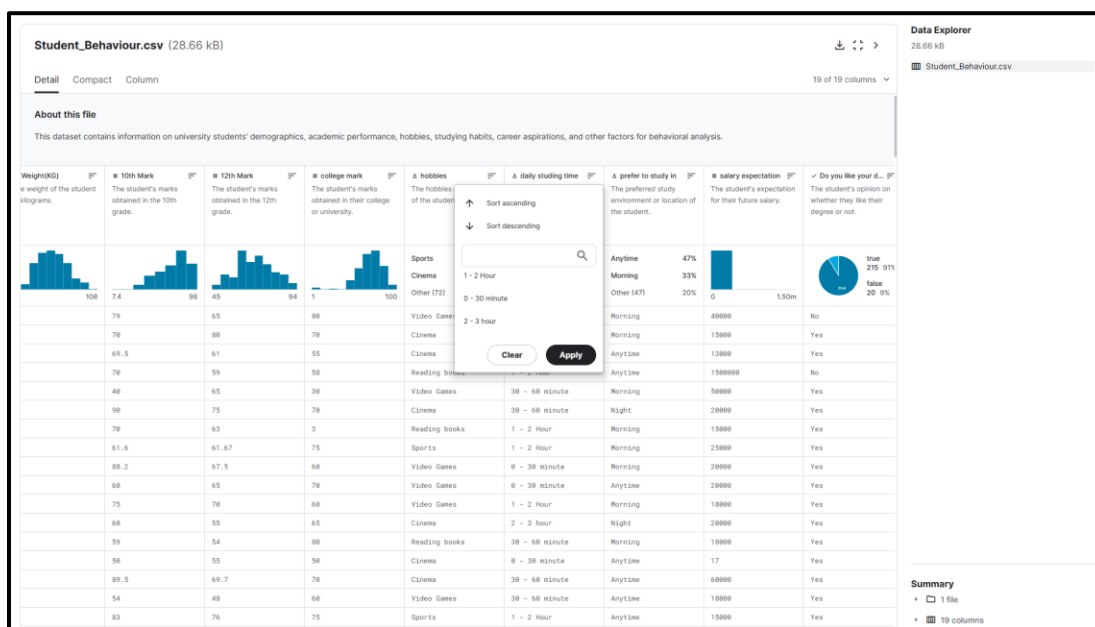
資料科學在教育上的應用—期中作業

1. 請問你挑選的哪一份資料集？請根據「欄位介紹」簡單描述這份資料集提供哪些資料並說明每一個欄位的意義。

資料集 1：Student Behavior。

欄位介紹：

- 1) Certification Course: 學生是否完成認證課程（回答 Yes / No）
- 2) Gender: 學生性別（Male / Female）。
- 3) Department: 學生就讀系所或是學習領域（此資料集大多為 BCA / Commerce）。
- 4) Height (CM): 學生的身高，公分為單位。
- 5) Weight (KG): 學生的體重，公斤為單位。
- 6) 10th Mark: 學生 10 年級（高一）的分數。
- 7) 12th Mark: 學生 12 年級（高三）的分數。
- 8) College Mark: 學生大學的分數。
- 9) Hobbies: 學生的興趣嗜好（此資料集大多為 Sports / Cinema）。
- 10) Daily Studying Time: 學生每日學習量。
- 11) Prefer to Study in: 學生喜歡的學習地點或著環境（此資料集只有 Anytime / Morning / Night）。
- 12) Salary Expectation: 學生對於未來薪資的期望。
- 13) Do you like your degree?: 學生是否喜歡的科系。
- 14) Willingness to pursue a career based on their degree: 學生對於科系未來職業的意願程度。
- 15) Social Media & Video: 學生花在社群影片平台的時間。
- 16) Traveling Time: 學生上學時間。
- 17) Stress Level: 學生自認的壓力程度。
- 18) Financial Status: 學生財務狀況。
- 19) Part-time Job: 學生是否有兼職工作。



圖一：欄位介紹資料

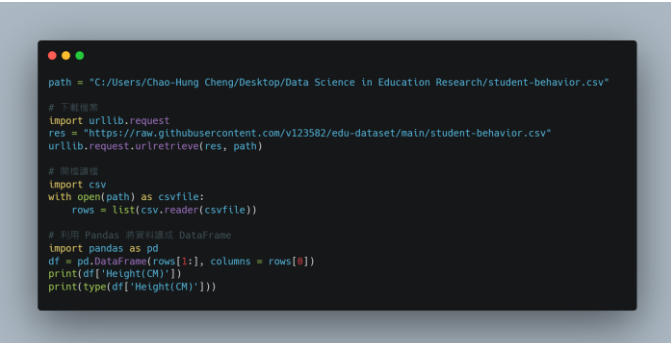
2. 請根據你挑選的資料集，從定義出五個資料可以回答的「假設」。

- 1) 通勤時間是否影響成績？
- 2) 每日學習量是否影響成績？
- 3) 學生喜歡的學習環境是否影響成績？
- 4) 學生是否喜歡自己的科系是否影響成績？
- 5) 學生花在社群平台的時間是否影響成績？

3. 請利用 Python 中的 Requests 下載「資料網址」並利用 Pandas 分析，請附上使用到的 Python 程式碼與執行結果：

(1) 利用 Requests 將資料下載，並且利用 Pandas 將資料讀成 DataFrame

請見 GitHub：<https://github.com/twdanielcheng/Data-Science-in-Education-Research/blob/master/Q3-1.py>。



```

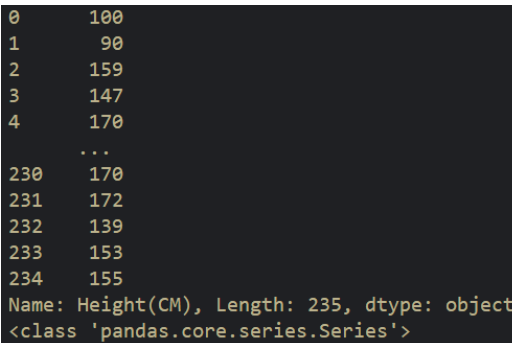
path = "C:/Users/Chao-Hung Cheng/Desktop/Data Science in Education Research/student-behavior.csv"

# 下載檔案
import urllib.request
res = "https://raw.githubusercontent.com/v123582/edu-dataset/main/student-behavior.csv"
urllib.request.urlretrieve(res, path)

# 解碼檔案
import csv
with open(path) as csvfile:
    rows = list(csv.reader(csvfile))

# 利用 Pandas 將資料讀成 DataFrame
import pandas as pd
df = pd.DataFrame(rows[1:], columns = rows[0])
print(df['Height(CM)'])
print(type(df['Height(CM)']))

```



```

0      100
1       90
2      159
3      147
4      170
...
230     170
231     172
232     139
233     153
234     155
Name: Height(CM), Length: 235, dtype: object
<class 'pandas.core.series.Series'>

```

(2) 這份資料集有多少筆資料和多少個欄位？

資料筆數 235、欄位數量 19

程式碼請見 GitHub：<https://github.com/twdanielcheng/Data-Science-in-Education-Research/blob/master/Q3-2.py>



```

path = "C:/Users/Chao-Hung Cheng/Desktop/Data Science in Education Research/student-behavior.csv"

# 下載檔案
import urllib.request
res = "https://raw.githubusercontent.com/v123582/edu-dataset/main/student-behavior.csv"
urllib.request.urlretrieve(res, path)

# 解碼檔案
import csv
with open(path) as csvfile:
    rows = list(csv.reader(csvfile))

# 利用 Pandas 將資料讀成 DataFrame
import pandas as pd
df = pd.DataFrame(rows[1:], columns = rows[0])
# print(df['Height(CM)'])
# print(type(df['Height(CM)']))

# 列出資料數量、欄位數量
print("資料筆數", df.shape[0])
print("欄位數量", df.shape[1])

```



資料筆數 235
欄位數量 19

(3) 哪些是數值欄位？哪些是類別欄位？

程式碼請見 GitHub：<https://github.com/twdanielcheng/Data-Science-in-Education-Research/blob/master/Q3-3.py>

數值欄位為黃底，類別欄位為藍底

- 1) Certification Course: 學生是否完成認證課程（回答 Yes / No）。
- 2) Gender: 學生性別（Male / Female）。
- 3) Department: 學生就讀系所或是學習領域（此資料集大多為 BCA / Commerce）。
- 4) Height (CM): 學生的身高，公分為單位。
- 5) Weight (KG): 學生的體重，公斤為單位。
- 6) 10th Mark: 學生 10 年級（高一）的分數。
- 7) 12th Mark: 學生 12 年級（高三）的分數。
- 8) College Mark: 學生大學的分數。

- 9) Hobbies: 學生的興趣嗜好（此資料集大多為 Sports / Cinema）。
- 10) Daily Studying Time: 學生每日學習量。
- 11) Prefer to Study in: 學生喜歡的學習地點或著環境（此資料集只有 Anytime / Morning / Night）。
- 12) Salary Expectation: 學生對於未來薪資的期望。
- 13) Do you like your degree?: 學生是否喜歡的科系。
- 14) Willingness to pursue a career based on their degree: 學生對於科系未來職業的意願程度。
- 15) Social Media & Video: 學生花在社群影片平台的時間。
- 16) Traveling Time: 學生上學時間。
- 17) Stress Level: 學生自認的壓力程度。
- 18) Financial Status: 學生財務狀況。
- 19) Part-time Job: 學生是否有兼職工作。

```
path = "C:/Users/Chao-Hung Cheng/Desktop/Data Science in Education Research/student-behavior.csv"

# 下載檔案
import urllib.request
res = "https://raw.githubusercontent.com/v123582/edu-dataset/main/student-behavior.csv"
urllib.request.urlretrieve(res, path)
import io
import requests

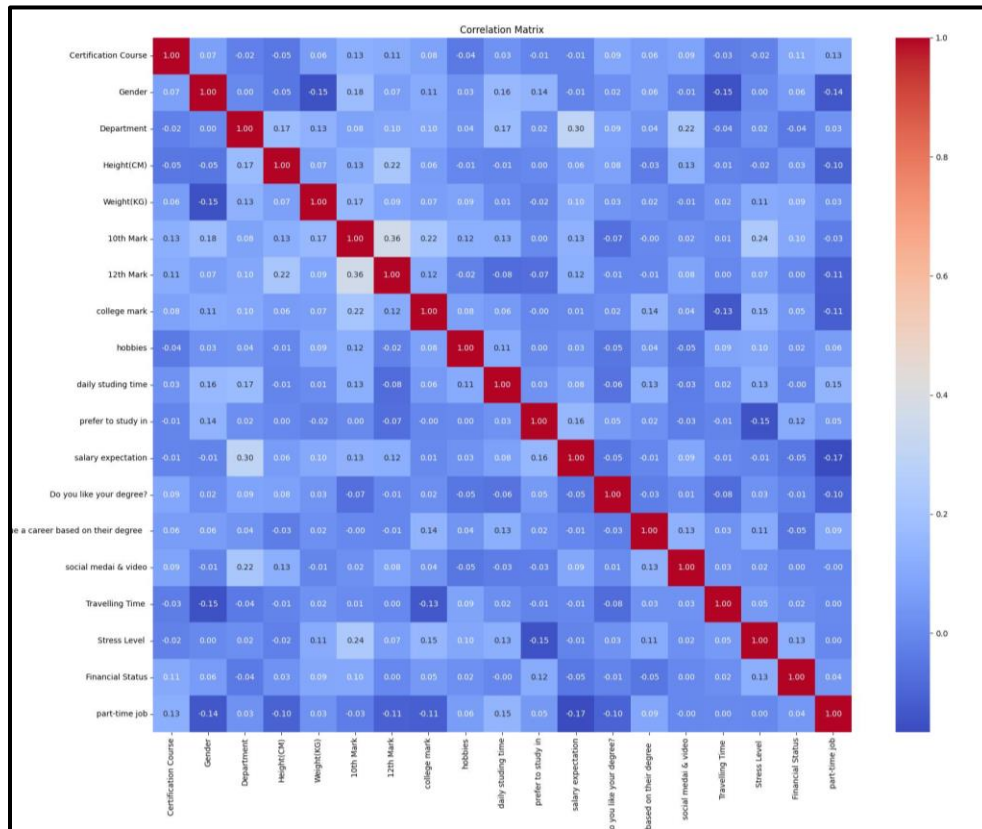
import pandas as pd
df = pd.read_csv(io.StringIO(requests.get(res).content.decode('utf-8')))
num_rows, num_columns = df.shape

# object 為類別欄位、非 object 為數值欄位
print(df.dtypes)
```

Certification Course	object
Gender	object
Department	object
Height(CM)	float64
Weight(KG)	float64
10th Mark	float64
12th Mark	float64
college mark	float64
hobbies	object
daily studing time	object
prefer to study in	object
salary expectation	int64
Do you like your degree?	object
willingness to pursue a career based on their degree	object
social medai & video	object
Travelling Time	object
Stress Level	object
Financial Status	object
part-time job	object

(4) 請畫出相關係數矩陣並觀察結果？

<https://github.com/twdanielcheng/Data-Science-in-Education-Research/blob/master/Q3-4.py>



```
import urllib.request
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import requests
import io

path = "C:/Users/Chao-Hung Cheng/Desktop/Data Science in Education Research/student-behavior.csv"

# 下載檔案
res = "https://raw.githubusercontent.com/v123582/edu-dataset/main/student-behavior.csv"
urllib.request.urlretrieve(res, path)

# 讀取資料
df = pd.read_csv(io.StringIO(requests.get(res).content.decode('utf-8')))
df_encoded = df.apply(lambda x: x.factorize()[0])

correlation_matrix = df_encoded.corr()

plt.figure(figsize=(20, 16))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f")

plt.title("Correlation Matrix")
plt.savefig("image.jpg")
plt.show()
```