

# 學習分析工具實務應用

## \_20241112

指導教授 周克行 博士



國立臺灣大學  
National Taiwan University



Educational : **Finance PHD**

Current Job : **Postdoctoral Researcher**

Experience:

2023: AI education with Google Bard

2023: Carbon Blockchain for agriculture plan, presentation for president

2021:The 14th崇越 Paper Awards, Invited to meet the President at the Presidential Palace on 2021/12/23

SSCI paper:

**Ke-Hsin Chou**, Min-Yuh Day\*, and Chien-Liang Chiu, "Do Bitcoin news information flow and return volatility fit the sequential information arrival hypothesis and the mixture of distribution hypothesis?" *International Review of Economics and Finance*, Volume 88, November 2023, pp. 365-385. [SSCI] [Impact Factor@2024: **7.2, Q1**, (26/111 = 23.4%, Business, Finance); (69/380 = 18.1%, Economics)][Taiwan Ministry of Science and Technology Finance: Level A-Journal]

Kao, Yu-Sheng, Min-Yuh Day, and **Ke-Hsin Chou\***. "A comparison of bitcoin futures return and return volatility based on news sentiment contemporaneously or lead-lag." *The North American Journal of Economics and Finance* 72 (2024): 102159. [SSCI] [Impact Factor@2022: **3.6, Q2**, (40/111 = 36%, Business, Finance); (100/380 = 26.3%, Economics)]

Chiu, Chien-Liang, and **Ke-Hsin Chou\***. "The soft commodities multiple bubbles tests: evidence from the New York Futures Markets." *Applied Economics Letters* (2020): 1-6.

崇越  
論文大賞  
管理學界奧斯卡

TSC 崇越論文大賞  
TSC Thesis Award



2021 第十四屆崇越論文大賞總統接見

# TA

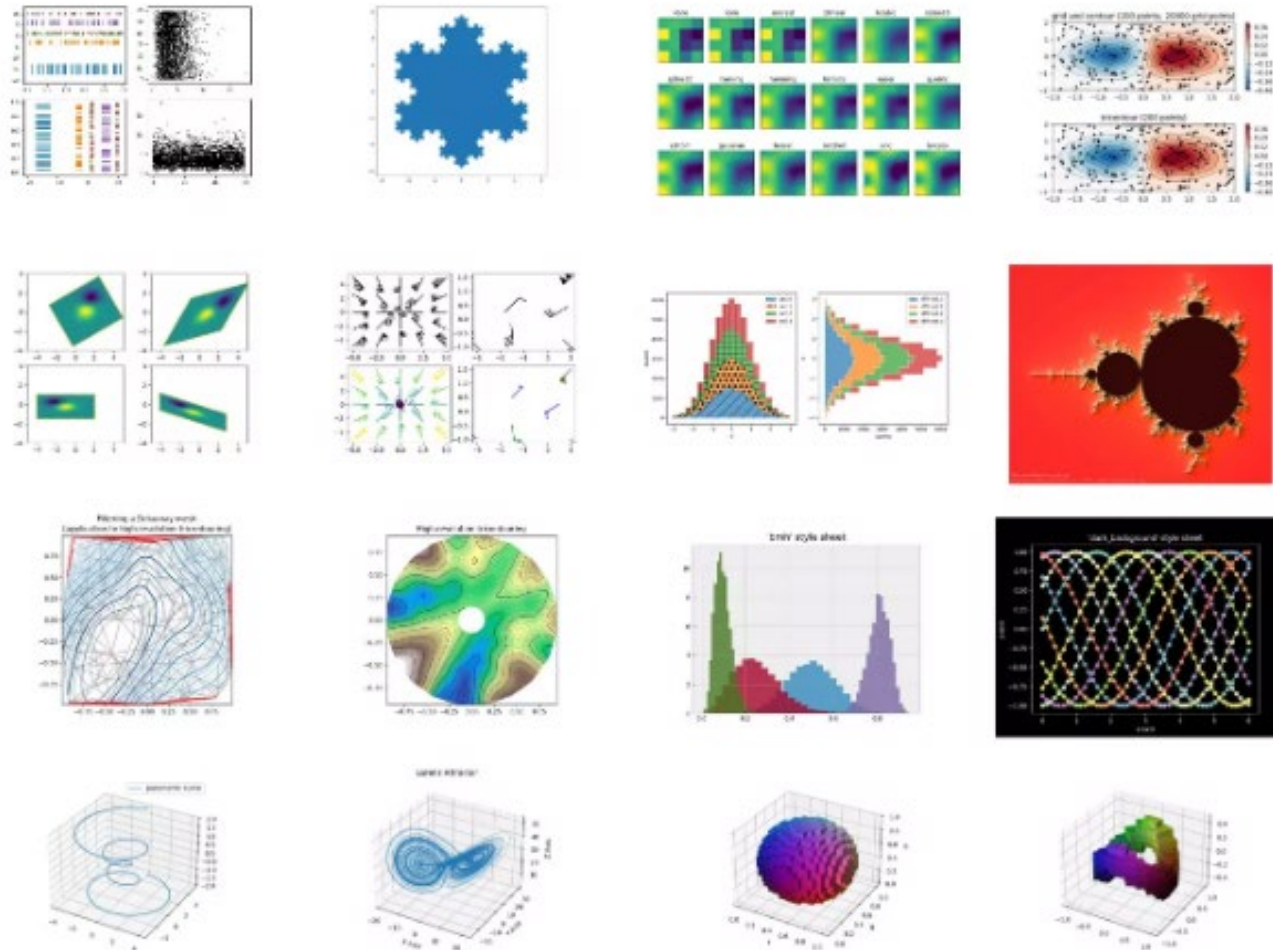
姓名	系級	修過的教育大數據學 程課	擔任過TA的課	email
李俊融	企管 大四	學習分析工具實務應 用	112-2學習分析工具實務 應用	<a href="mailto:21211256asd0810@gmail.com">21211256asd0810@gmail.c om</a>
黃羿寧	教育 大四	大數據程式設計 教育大數據專題製作	無	<a href="mailto:40900101eshirley@gmail.com">40900101eshirley@gmail.co m</a>
簡珮軒	科技系大四		助教（111-2）	peggygirl0202@gmail.com

matplotlib 是 Python 的一個**第三方函式庫**，是相當重要且受歡迎的資料視覺化函式庫，matplotlib 可以根據數據資料，繪製直方圖、元餅圖、折線圖...等各種圖表，也能和其他 Python 的資料處理函式庫 ( NumPy、Pandas...等 ) 互相搭配，進行更複雜的視覺圖表繪製。

# matplotlib



# matplotlib 支援的圖表類型



# 安裝 matplotlib 函式庫

```
!pip install matplotlib
```

## **import matplotlib**

要使用 matplotlib 必須先 import matplotlib 模組。

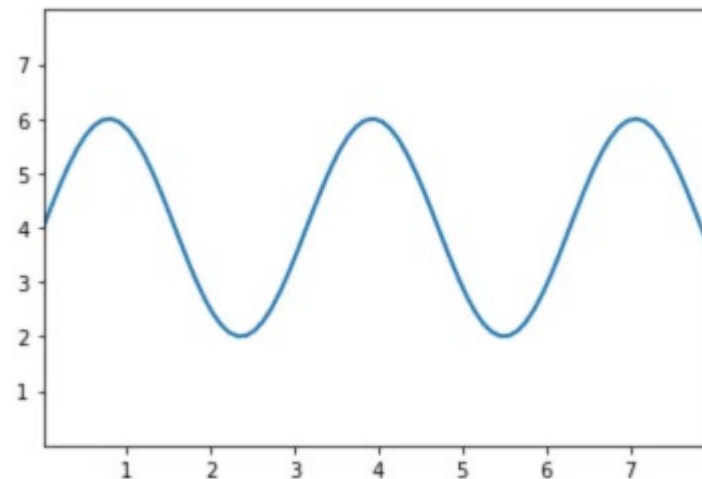
透過 matplotlib 繪製圖表時，大部分的情況會使用 pyplot 模組，通常會將其獨立命名為 plt。

```
import matplotlib.pyplot as plt
```



# 簡單感受一下 matplotlib

- `import matplotlib.pyplot as plt`
- `import numpy as np`
- `x = np.linspace(0, 10, 100)` # 產生 0~10 總共 100 連續數字
- `y = 4 + 2 * np.sin(2 * x)` # 使用 NumPy 的廣播方式，產生 sin 數值的陣列 y
- `fig, ax = plt.subplots()`
- `ax.plot(x, y, linewidth=2.0)` # 繪製折線圖
- `ax.set(xlim=(0, 8), xticks=np.arange(1, 8),` # 設定座標軸
- `ylim=(0, 8), yticks=np.arange(1, 8))`
- `plt.show()` # 顯示圖表



# 改變圖片顯示的尺寸

```
import matplotlib.pyplot as plt
import matplotlib.image as img
import os
image = img.imread('aa.jpg')
plt.figure(figsize=(10,10)) # 改變圖表尺寸
plt.imshow(image)
plt.show()
```





# 3D 圖表

```
import matplotlib.pyplot as plt
```

```
fig = plt.figure(figsize=(6,6))
```

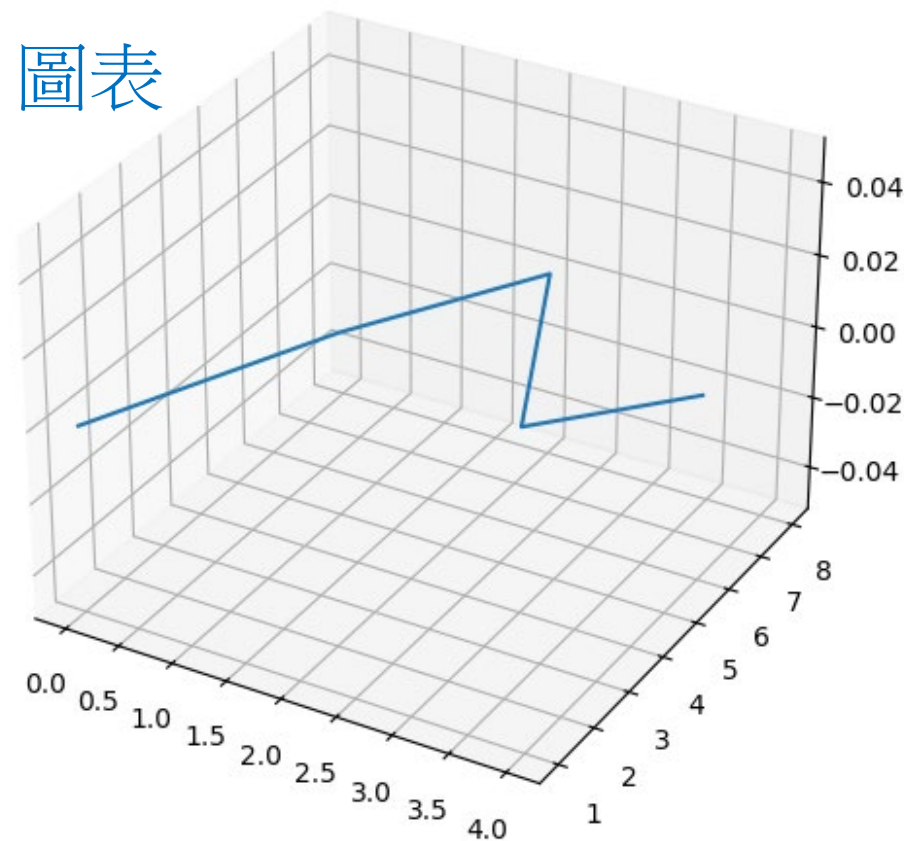
```
ax = plt.subplot(projection='3d') # 設定為 3D 圖表
```

```
x = range(5)
```

```
y = [1,5,8,4,6]
```

```
ax.plot(x,y)
```

```
plt.show()
```



# 折線圖

```
import matplotlib.pyplot as plt
```

```
x = [1,2,3,4,5]
```

```
# 畫出顏色紅色、圓形錨點、虛線、粗細2、資料點大小6的線條
```

```
#plt.plot(x,color='r', marker='o', linewidth=2, markersize=6)
```

```
plt.plot(x)
```

```
plt.show()
```

# 參數說明

參數	說明
x	必填，第一組數據 ( x 軸 )。
y	第二組數據 ( y 軸 )。
color	線條或資料點的顏色 ( 除了十六進位色碼，也可填入顏色代碼，例如 r、g、b、m、c、y...等，參考： <a href="#">color 列表</a> )。
marker	資料點樣式，預設無資料點 ( 資料點樣式代碼為 .、,、o、v...等，參考： <a href="#">markers 列表</a> )。
linewidth	線條粗細，預設 2。
markersize	資料點大小，預設 6。

# 用DataFrame

```
import numpy as np
import pandas as pd
dates = pd.date_range("20130101", periods=100)
df = pd.DataFrame(np.random.randn(100, 4), index=dates,
columns=list("ABCD"))
df
```

# 作業練習1

```
import matplotlib.pyplot as plt
```

```
x = #請問這裡要怎麼改?
```

```
# 畫出顏色紅色、圓形錨點、虛線、粗細2、資料點大小6的線條
```

```
plt.plot(x,color='r', marker='o', linewidth=2, markersize=6)
```

```
#plt.plot(x)
```

```
plt.show()
```

# 長條圖

```
import matplotlib.pyplot as plt
```

```
x = [1,2,3,4,5]
```

```
h = [10,20,30,40,50]
```

```
color = ['r','b','g','y','m'] # 顏色數據
```

```
label = ['a','b','c','d','e'] # 標籤數據
```

```
plt.bar(x,h,color=color,tick_label=label,width=0.5) # 加入顏色、標籤  
和寬度參數
```

```
plt.show()
```

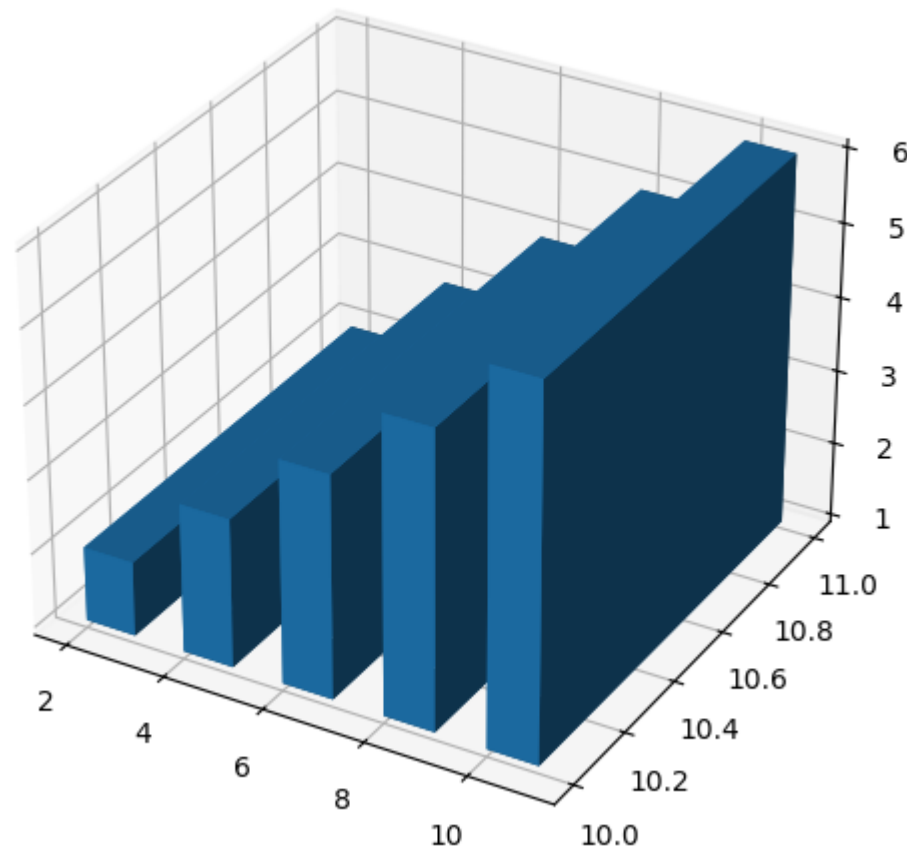
# 圓餅圖

```
import matplotlib.pyplot as plt  
x = [1,2,3,4,5]  
plt.pie(x, radius=1.5, labels=x)  
plt.show()
```



# 3D 柱狀長條圖

```
import matplotlib.pyplot as plt  
fig = plt.figure(figsize=(6,6))  
ax = plt.subplot(projection='3d')  
x = [2,4,6,8,10]  
y = 10  
z = 1  
ax.bar3d(x,y,z,dx=1,dy=1,dz=[1,2,3,4,5])  
plt.show()
```



# 作業練習2

請把資料改成以下**pandas** 產生的資料, 畫出長條圖, 圓餅圖, **3D** 柱狀長條圖

```
import numpy as np
import pandas as pd
dates = pd.date_range("20130101", periods=100)
df = pd.DataFrame(np.random.randn(100, 4), index=dates,
columns=list("ABCD"))
```

可以問ChatGPT 或 Google Gemini

# 上星期AI prompt 分享

萬用 prompt:

1. 你是人類，我是ChatGPT 你即將要求我幫你寫一個（任務），請問你會如何下命令給我

2. GPT生出的 prompt 在貼回去GPT

# 皮爾森相關係數(Pearson's correlation coefficient)

$$\rho = \frac{x \text{ 和 } y \text{ 的共變異數}}{x \text{ 的標準差} \times y \text{ 的標準差}}$$

$$\text{共變異數(covariance): } \text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

$$\text{變異數(variance): } \text{var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)^2$$

$$\text{標準差(standard deviation): } \text{std}(x) = \sqrt{\text{var}(x)}$$

它是兩個變量的共變異數與其標準差的乘積之比；因此，它本質上是共變異數的歸一化度量，因此結果始終具有介於-1和1之間的值。

假設有兩個變數 $(x_i, y_i)$ ,  $i=1, \dots, n$ ，一般網路看到的相關係數的公式定義如下：

$$\rho = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2 \sum_{i=1}^n (y_i - \mu_y)^2}}$$

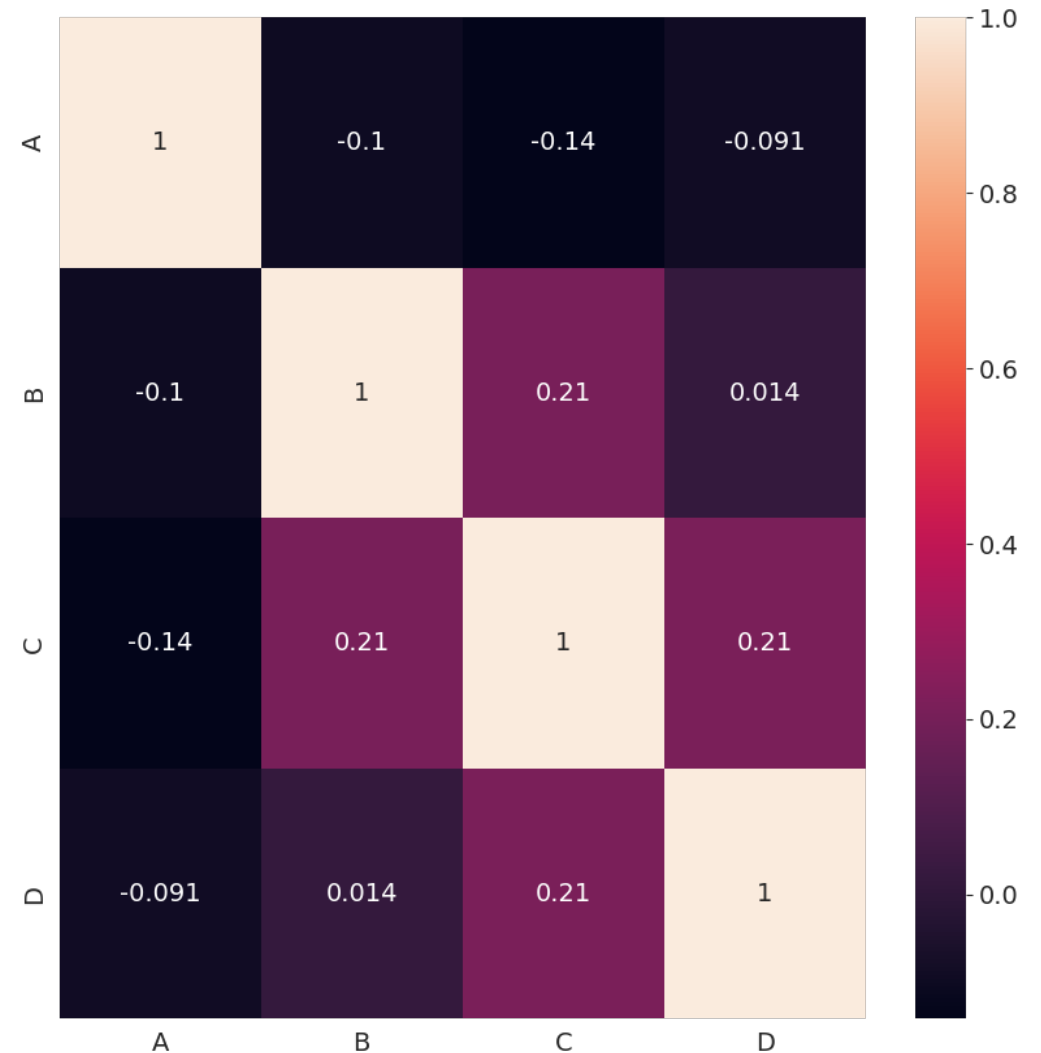
# 相關係數圖

```
import pandas as pd
import seaborn as sns
```

```
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

```
# Generate a correlation matrix for the selected columns
corrMatrix = df.corr()
```

```
# Plot the correlation matrix using a heatmap
plt.figure(figsize=(10,10))
sns.heatmap(corrMatrix, annot=True)
plt.show()
```



# ANOVA (Analysis of Variance)變異數分析

**ANOVA (Analysis of Variance)** 檢定是一種統計方法，用於比較兩組或多組之間的均值，來檢查它們之間是否有顯著的差異。這種方法經常被用於檢測不同組別之間的影響是否顯著。通常，**ANOVA**被用來解答以下問題：

- 各組的均值是否相同？
- 各組之間的差異是否顯著大於組內變異？



```
import numpy as np
from scipy import stats

# 三個班學生的成績
group1 = [85, 86, 88, 75, 78, 94]
group2 = [92, 94, 89, 96, 93, 95, 76, 50]
group3 = [78, 81, 79, 84, 77, 80]

# 使用 scipy.stats.f_oneway 進行單因子 ANOVA 檢定
f_statistic, p_value = stats.f_oneway(group1, group2, group3)

# 輸出結果
print("F 統計量:", f_statistic)
print("p-value:", p_value)

# 解釋結果
alpha = 0.05 # 顯著性水平
if p_value < alpha:
    print("拒絕原假設：不同組別之間的均值存在顯著差異。")
else:
    print("未能拒絕原假設：不同組別之間的均值沒有顯著差異。")
```

# 變異數分析會有什麼幫助？

- 為什麼這項分析很有用呢？
- 這是因為當您了解了每個自變量的均值與其他自變量有什麼不同時，您就可以開始研究並知道其中哪個自變量與您的因變量（如登陸頁面的點擊量）有關係，並且了解是什麼因素在驅動這個行為

# 用處

- ANOVA 檢定包含因變量和自變量。在 ANOVA 中，因變量（**Dependent Variable**）是連續數據，通常是我們希望比較的測量結果，而自變量（**Independent Variable**）是分類變量，用於將數據分組（如不同治療方法或教育方式）。ANOVA 的目的是檢查這些自變量是否對因變量的均值有顯著影響。
- 例如，假設我們有一個情境：我們想研究不同的教學方法對學生考試成績的影響。這裡“學生考試成績”是因變量，而“教學方法”是自變量。

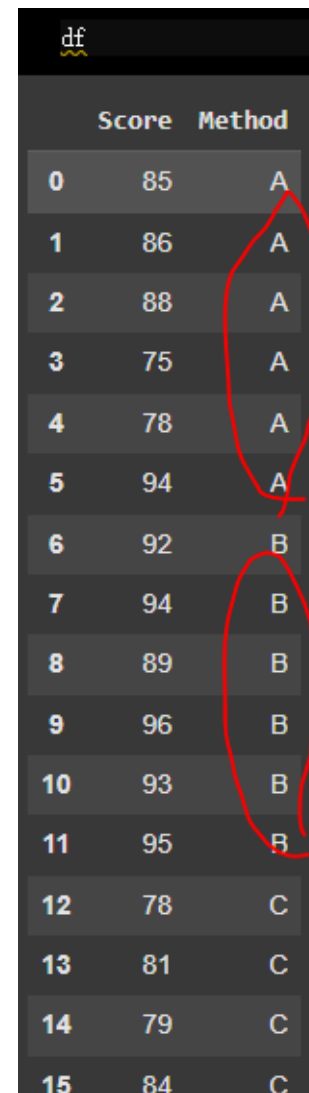
```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

# 創建數據集
data = {
    'Score': [85, 86, 88, 75, 78, 94, 92, 94, 89, 96, 93, 95, 78, 81, 79, 84, 77, 80],
    'Method': ['A', 'A', 'A', 'A', 'A', 'A', # 教學方法 A
               'B', 'B', 'B', 'B', 'B', 'B', # 教學方法 B
               'C', 'C', 'C', 'C', 'C', 'C'] # 教學方法 C
}

# 將數據轉換為 DataFrame
df = pd.DataFrame(data)

# 使用 statsmodels 中的 ols 方法進行單因子 ANOVA 分析
model = ols('Score ~ C(Method)', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

# 輸出 ANOVA 表
print(anova_table)
```



	Score	Method
0	85	A
1	86	A
2	88	A
3	75	A
4	78	A
5	94	A
6	92	B
7	94	B
8	89	B
9	96	B
10	93	B
11	95	B
12	78	C
13	81	C
14	79	C
15	84	C

	sum_sq	df	F	PR(>F)
C(Method)	552.111111	2.0	13.848941	0.000391
Residual	299.000000	15.0	NaN	NaN

	離均差平方和(SS)	自由度(DF)	F (檢定)	P (顯著)
組間	SSB (組間變異)	DFB=K-1 (組別-1)	MSB/MSW	查表
組內	SSW (組內變異)	DFW=(N-1)-(K-1)=N-K	MSW	
全體	SST (總變異)	DFT=N-1 (樣本數-1)		

**自由度**代表在計算某個統計量時可以自由變化的數據點的數量, ex:  $X_1 + X_2 + X_3 + 1 = 10$

一個正式的報告

# 1. 先把資料的圖畫出來





## 2. 在使用統計模型說明資料之間的關係

### ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	$\eta^2$
drug	3.45	2	1.73	18.61	0.00009	0.71
Residuals	1.39	15	0.09			