

資料科學 DATA SCIENCE

資料分析 | 數據分析 (Data Analysis)
資料視覺化 (data Visualization)

陳怡芬 yfchen@gapps.fg.tp.edu.tw 2017





數據科學家

- 所謂數據科學家就是：運用**數據和科學**，**創造新東西**的人
- 數據科學家這個職位的頭銜則是 2009 年由 Natahn Yau 首次提及的，他認為數據科學家就是能夠從大型數據集中析取出數據，並提供某些可供非數據專家使用的東西的人。



數據科學家、創業家 Mike Driscoll 則認為數據有三個性感之處：建模、轉換、可視化。

而一種比較有詩意的表述方式是：數據科學家好比是哥倫布遇上科倫坡，目光如炬的探險家與懷疑一切的大偵探的合體。

而在《數據科學家：二十一世紀最性感的職業》一文中，設計 LinkedIn 的「你可能認識的人」功能的數據科學家 Jonathan Goldman 的工作，也許是對數據科學家工作方式的最好詮釋：首先構建理論、印證預感，然後尋找出模式，對應該推出某人的哪一個網絡做出預測。

數據科學家

- 數據科學家就是採用科學方法、運用數據挖掘工具尋找新的數據洞察的工程師。
- 科學辦法就是構思假設、測試想法、精心設計實驗、經由他人驗證，這些是他們從統計身上掌握的知識，經科學訓練出來的經驗；而工具的運用則是來自其工程經驗，或者更確切地說，來自於其計算機科學與編程背景。
- 最好的數據科學家是產品與流程的創新者，有時候還是新的數據挖掘工具的開發者。



如何成為 資料科學家

The background features a dark, almost black, field with dynamic, flowing shapes. On the left side, there are vibrant green, leaf-like or wave-like forms that curve upwards and outwards. On the right side, there are bright orange and yellow, wave-like forms that curve downwards and outwards, creating a sense of movement and contrast against the dark background.

DATA ANALYSIS

資料分析

DATA ANALYSIS





業務導覽

● 教育統計查詢窗口

- 教育統計查詢網
- 大專校院學科標準分類查詢系統
- 大專校院系所特色及新生註冊率查詢系統
- 大專校院校務資訊公開平台
- 高中職學生比查詢系統
- 高級中等學校地理資訊查詢系統
- 偏遠地區國中小地理資訊查詢系統

● 教育消費支出調查

105年應用統計分析

首頁 > 統計分析與出版品 > 分析與研究 > 應用統計分析 > 105年應用統計分析

發佈時間	標題	公告單位
105-05-18	104學年度原住民教育概況分析	統計處
105-05-17	未來5年(105~109學年)公立國中小教師人數推估分析報告	統計處
105-05-06	104學年大專校院新生註冊率變動分析	統計處
105-04-27	104學年新住民子女就讀國中小人數分布概況統計分析	統計處
105-04-25	104學年度各級教育統計概況分析	統計處
105-04-20	105~120 學年度大專校院大學1年級學生人數預測分析報告	統計處
105-04-20	105~120 學年度高級中等教育階段學生人數預測分析報告	統計處
105-04-20	105~120 學年度國民教育階段學生人數預測分析報告	統計處
105-02-01	99-101學年度大專校院畢業生就業薪資巨量分析 (完整報告)	統計處

教育部統計處



經貿議題

我國經貿現況

國際貿易情勢分析

每月國際貿易情勢分析

歷年國際貿易情勢分析

貿易統計資料查詢系統

貿易統計重要參考指標(摺頁卡)

我國貿易統計

各國貿易統計

相關網站資訊連結

目前位置：首頁 > 經貿議題 > 國際貿易情勢分析 > 每月國際貿易情勢分析

回上一頁

每月國際貿易情勢分析

日期範圍：

標題查詢：

查詢

項目	標題	文章公布日期	最新檢視日期
1	2016國際貿易情勢分析12月號 	2017-01-09	2017-01-09
2	2016年11月號國際貿易情勢分析 	2016-12-22	2016-12-22
3	2016年10月號國際貿易情勢分析 	2016-11-30	2016-11-30



文化統計



進
階
搜
尋

[關於文化統計](#)

[訊息與新聞](#)

[調查與研究](#)

[文化統計指標](#)

[統計研究分析](#)

[歷年文化統計資料查詢](#)

最新文化統計數據

- + 中央政府文化支出預算：
31,785(百萬元) - 2015年
- + 中央政府文化支出預算占總
預算比率：1.64(%) -
2015年
- + 文化部及所屬機關(構)預算
數：16,741(百萬元) -
2015年
- + 文化部及所屬機關(構)文化
預算占文化支出預算之比
率：52.67(%) - 2015年
- + 文化部及所屬機關(構)公款
補助國內團體經費：
3,720(百萬元) - 2015年
- + 文化部及所屬機關(構)公款
補助縣市政府經費：
1,569(百萬元) - 2015年

[首頁](#) > [統計研究分析](#)

統計研究分析

[意見回饋與訂閱電子報](#)

日期	標題
2017/3/20	國內外文化產業訊息及趨勢分析雙月報(106年第1期)
2017/1/17	國內外文化產業訊息及趨勢分析雙月報(105年第6期)
2016/11/16	國內外文化產業訊息及趨勢分析雙月報(105年第5期)
2016/9/13	國內外文化產業訊息及趨勢分析雙月報(105年第4期)
2016/7/13	國內外文化產業訊息及趨勢分析雙月報(105年第3期)
2016/5/23	國內外文化產業訊息及趨勢分析雙月報(105年第2期)
2016/5/20	國內外文化產業訊息及趨勢分析雙月報(105年第1期)
2016/1/26	國內外文化產業訊息及趨勢分析雙月報(104年第6期)



兩人一組學習任務

定義問題- 收集資料 – 統計分析資料 – 資料視覺化 – 解釋資料 – 提供建議與行動方案

開放資料 (Open data) 指的是一種經過挑選與許可的資料，這些資料不受著作權、專利權，以及其他管理機制所限制，可以開放給社會公眾，任何人都可以自由出版使用，不論是要拿來出版或是做其他的運用都不加以限制。

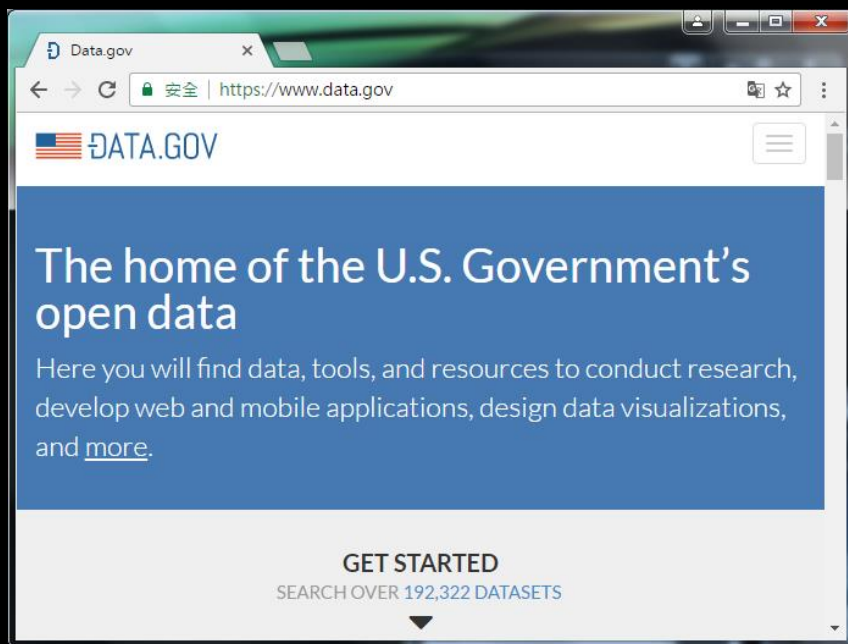
Open data 運動希望達成的目標與開放原始碼、內容開放、開放獲取。Open data 背後的核心思想由來已久，但 Open data 這名詞直到近代才出現，拜網際網路崛起而為人所知。

OPEN DATA

開放資料

OPEN DATA

- 美國國家政府資料開放平臺 <https://www.data.gov/>
- 政府資料開放平臺 <http://data.gov.tw/>
- 臺北市政府資料開放平台 <http://data.taipei/>



DATA ANALYSIS TOOLS

Excel, Matlab, Matplotlib in python, R.....



DATA ANALYSIS IN EXCEL

- Range 範圍
- Formulas and Functions 公式與函式
- Ribbon 標籤頁
- Workbook 活頁簿
- Worksheets 活頁紙
- Format Cells 儲存格
- Find & Select 尋找
- Data Validation 資料驗證

- Sort 排序
- Filter 篩選
- Conditional Formatting 條件格式化
- Charts 圖表
- Pivot Tables 樞紐分析表
- Tables 表格
- What-If Analysis 假設分析



DATA VISUALIZATION

資料視覺化

TED : David McCandless – 資料視覺化的美麗

David McCandless:

David McCandless : 資料 視覺化的美麗

TEDGlobal 2010 · 17:56 · Filmed Jul 2010

31 subtitle languages

View interactive transcript



Add to list



Like



Download



Rate

David McCandless將全世界的軍隊開銷、媒體活動、Facebook的狀態更新等複雜的資料，轉化成簡單又美麗的圖表。他提倡人們應該使用設計工具，來整理當今過多的資訊，找出獨特的模式與關連性，也許就能改變我們對這個世界的看法。

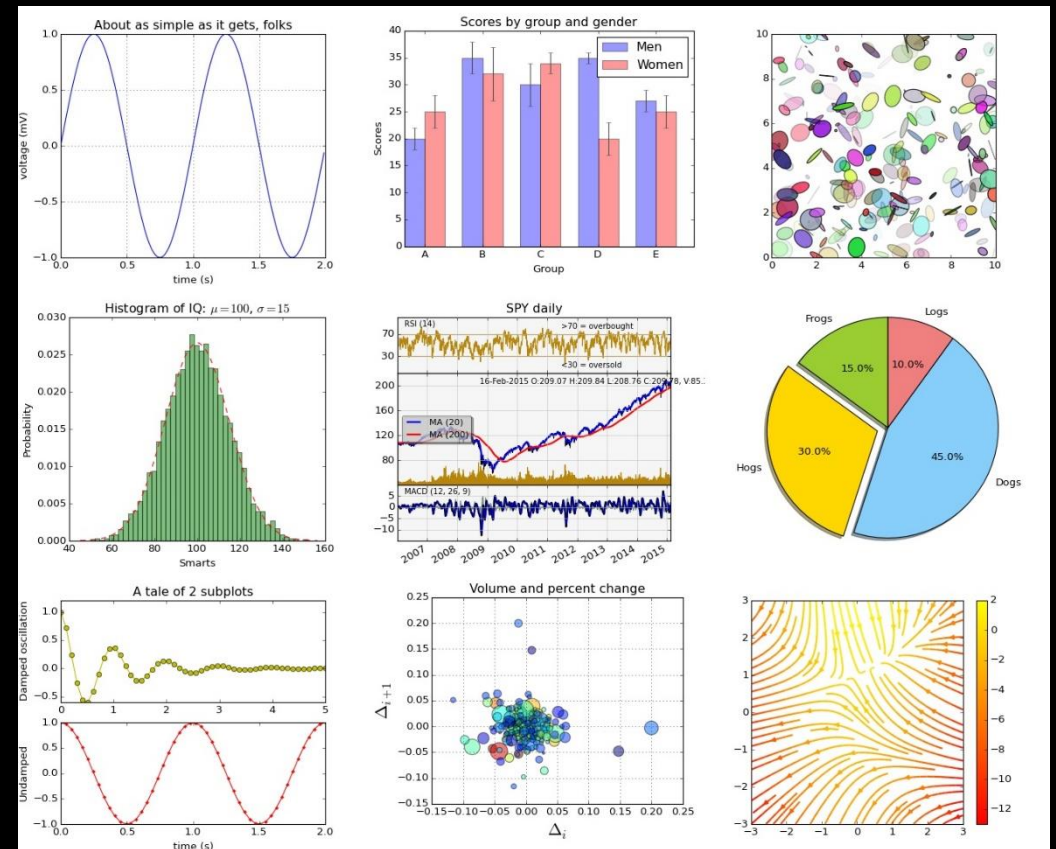
DATA ANALYSIS WITH PYTHON

<https://pythonprogramming.net/data-analysis-python-pandas-tutorial-introduction/>



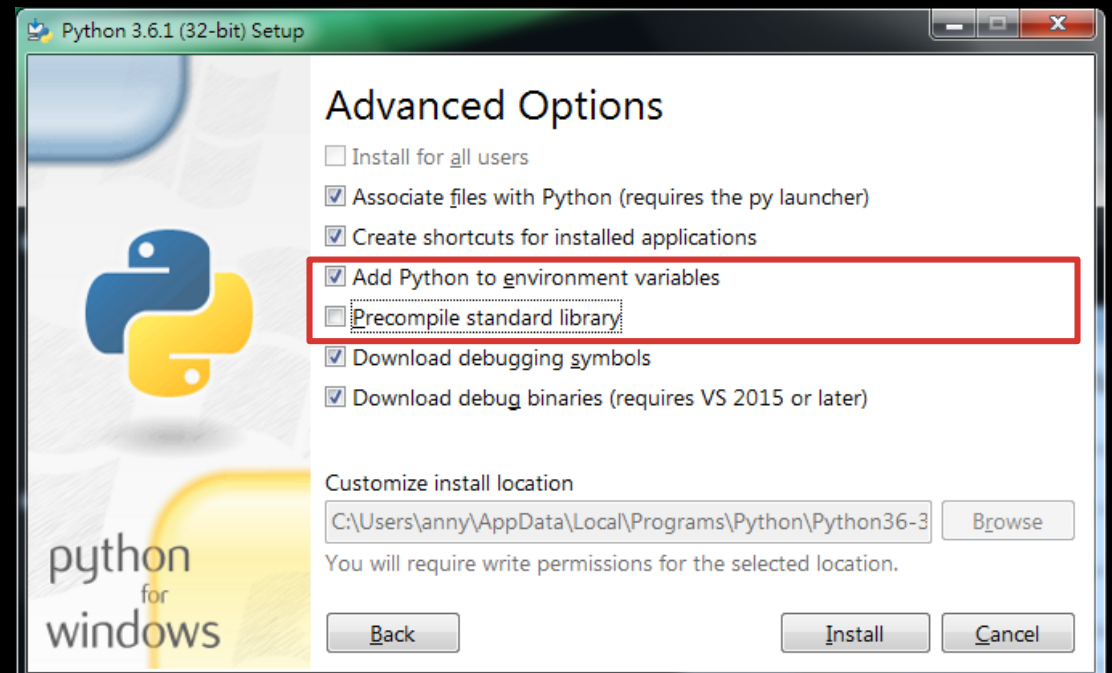
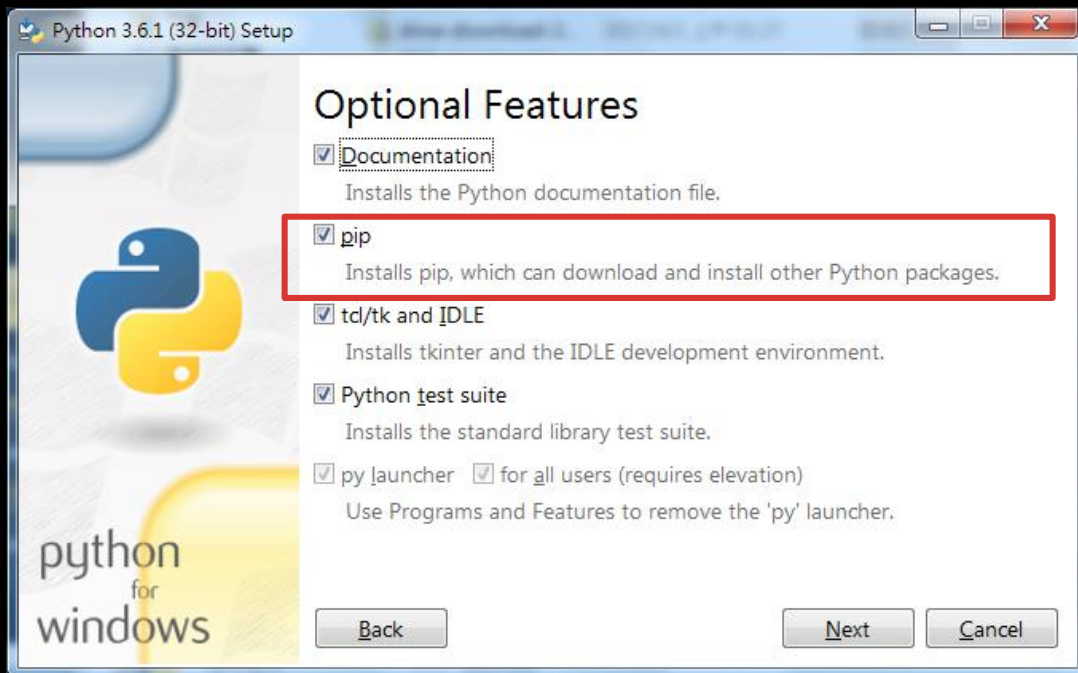
DATA VISUALIZATION VIA MATPLOTLIB PLOTTING

- line graphs 折線圖
- scatter plots xy散佈圖
- bar charts 長條圖
- pie charts 圓餅圖
- stack plots 堆疊圖
- 3D graphs
- geographic map graphs



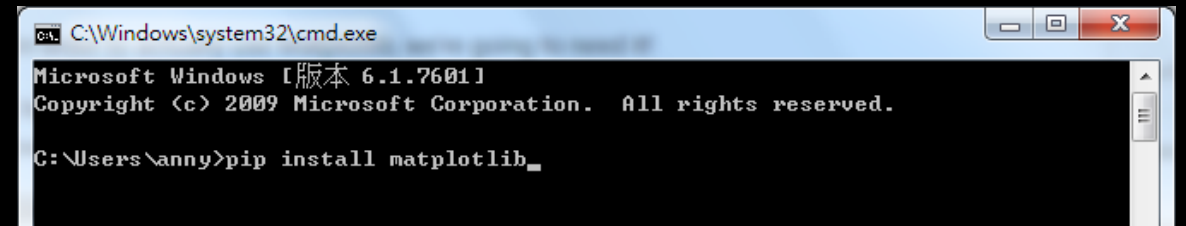
GETTING START-1

- Install **python 3.6.1** <https://www.python.org/>
 - Optional Features: 選項設定 (全選)
 - Advanced Options: 進階選項(全選)



GETTING START-2

- **cmd** (開啟命令提示字元視窗)
- **pip install matplotlib**



```
C:\Windows\system32\cmd.exe
Microsoft Windows [版本 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\anny>pip install matplotlib_
```

INTRODUCTION TO MATPLOTLIB AND BASIC **LINE**

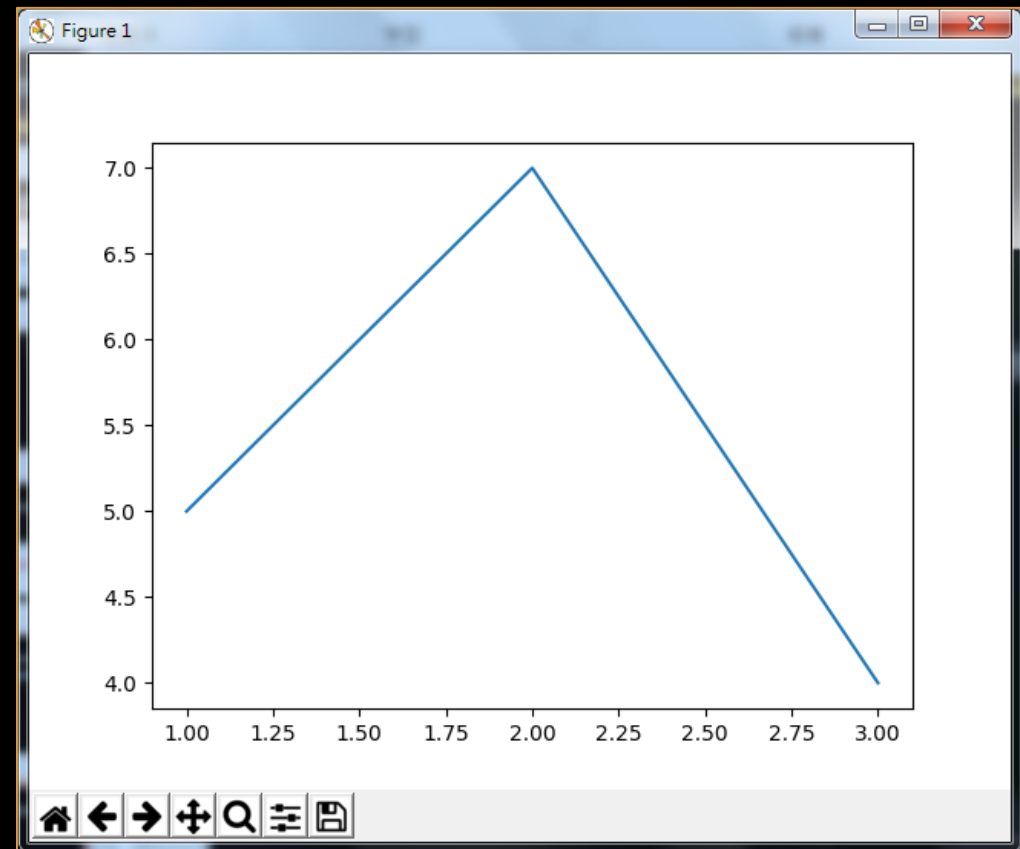
```
#plot1.py
```

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

```
plt.plot([1,2,3],[5,7,4])
```

```
plt.show()
```

Try1:修改 x,y 的 value list , 觀察變化



LEGENDS, TITLES, AND LABELS WITH MATPLOTLIB

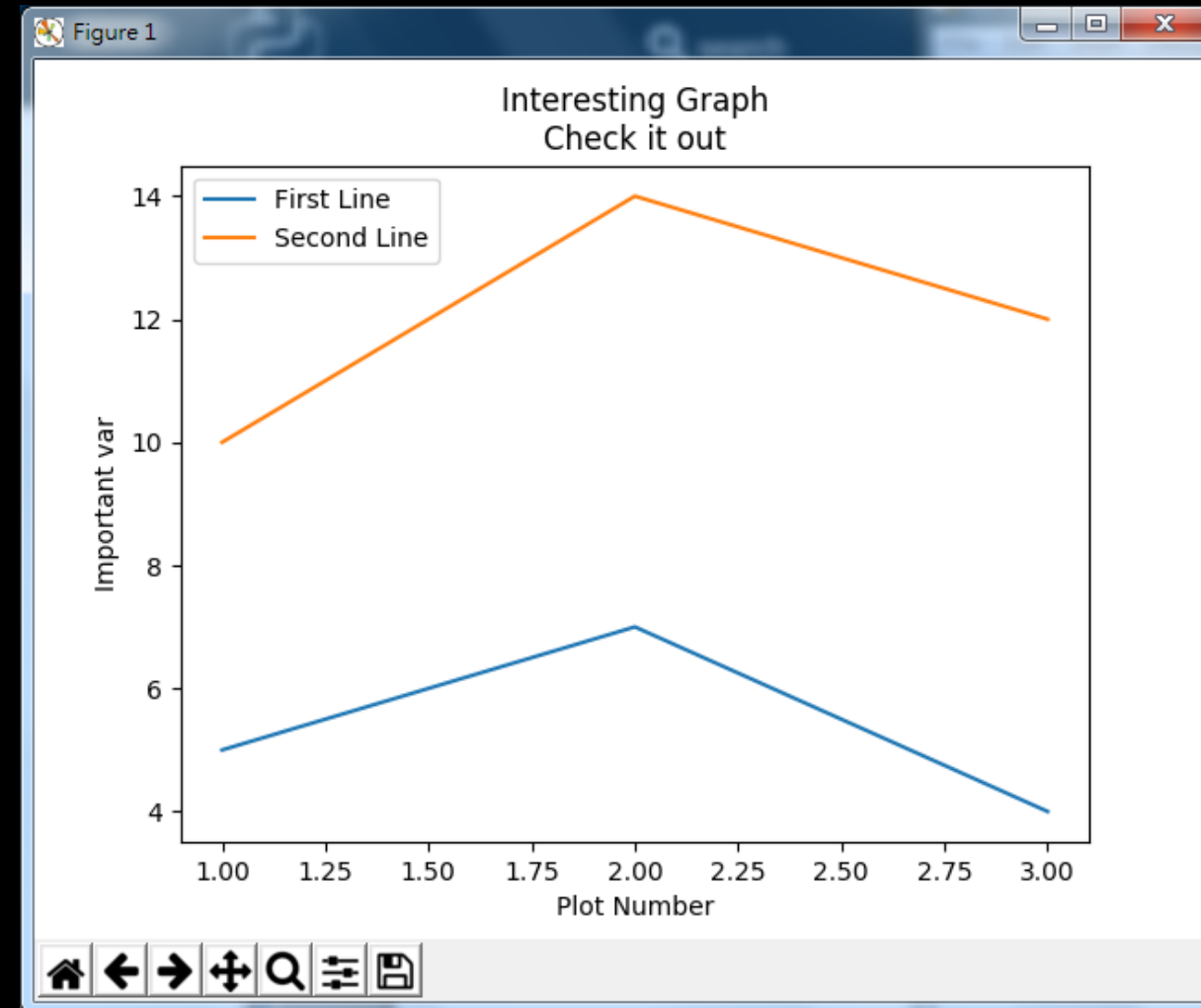
```
#plot2.py
import matplotlib.pyplot as plt

x = [1,2,3]
y = [5,7,4]

x2 = [1,2,3]
y2 = [10,14,12]

plt.plot(x, y, label='First Line')
plt.plot(x2, y2, label='Second Line')

plt.xlabel('Plot Number')
plt.ylabel('Important var')
plt.title('Interesting Graph\nCheck it out')
plt.legend()
plt.show()
```



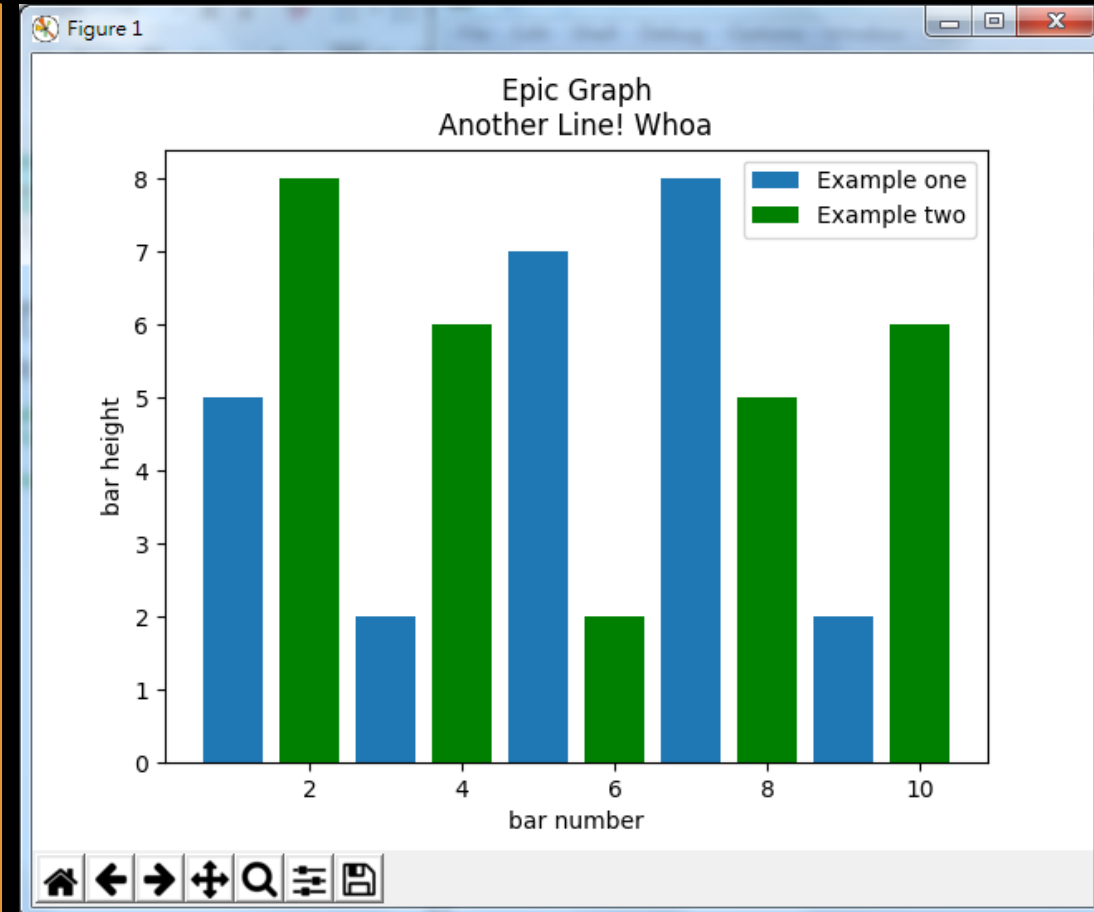
BAR CHARTS WITH MATPLOTLIB

```
#plot3.py
import matplotlib.pyplot as plt

plt.bar([1,3,5,7,9],[5,2,7,8,2], label="Example one")
plt.bar([2,4,6,8,10],[8,6,2,5,6], label="Example two", color='g')

plt.legend()
plt.xlabel('bar number')
plt.ylabel('bar height')
plt.title('Epic Graph\nAnother Line! Whoa')

plt.show()
```



HISTOGRAMS WITH MATPLOTLIB

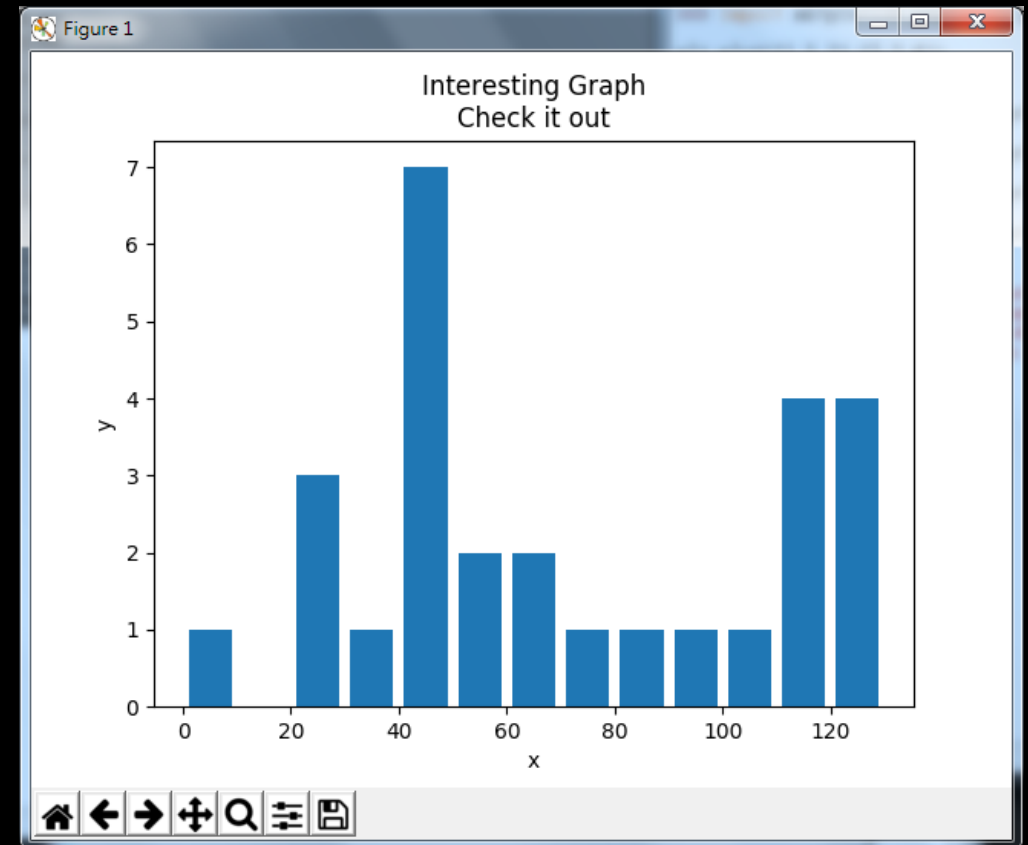
```
#plot4.py
#Histogram
import matplotlib.pyplot as plt

population_ages =
[22,55,62,45,21,22,34,42,42,4,99,102,110,120,121,122,130,11
1,115,112,80,75,65,54,44,43,42,48]

bins = [0,10,20,30,40,50,60,70,80,90,100,110,120,130]

plt.hist(population_ages, bins, histtype='bar', rwidth=0.8)
plt.xlabel('x')
plt.ylabel('y')
plt.title('Interesting Graph\nCheck it out')
plt.legend()

plt.show()
```



<https://pythonprogramming.net/bar-chart-histogram-matplotlib-tutorial/?completed=/legends-files-labels-matplotlib-tutorial/>

SCATTER PLOTS WITH MATPLOTLIB

```
#plot5.py  
#Scatter plots
```

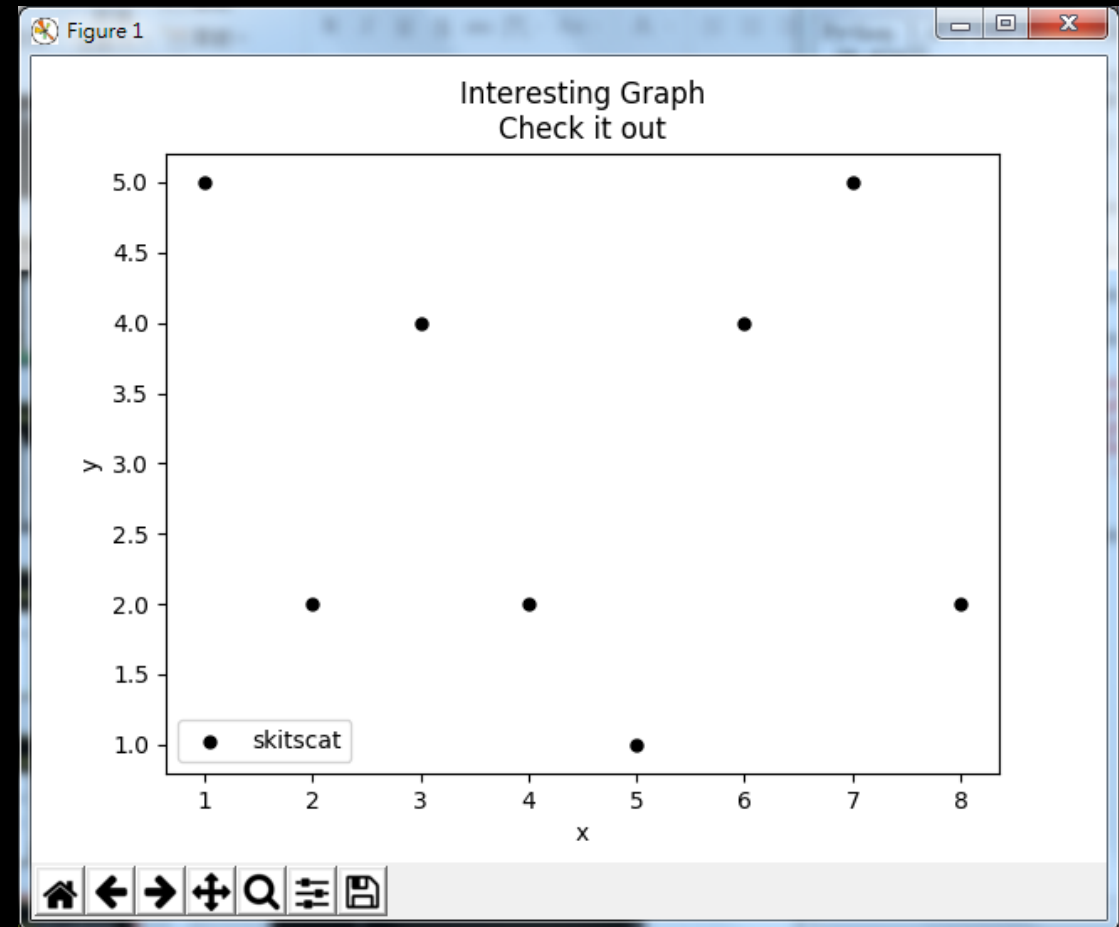
```
import matplotlib.pyplot as plt
```

```
x = [1,2,3,4,5,6,7,8]  
y = [5,2,4,2,1,4,5,2]
```

```
plt.scatter(x,y, label='skitscat', color='k', s=25, marker="o")
```

```
plt.xlabel('x')  
plt.ylabel('y')  
plt.title('Interesting Graph\nCheck it out')  
plt.legend()
```

```
plt.show()
```

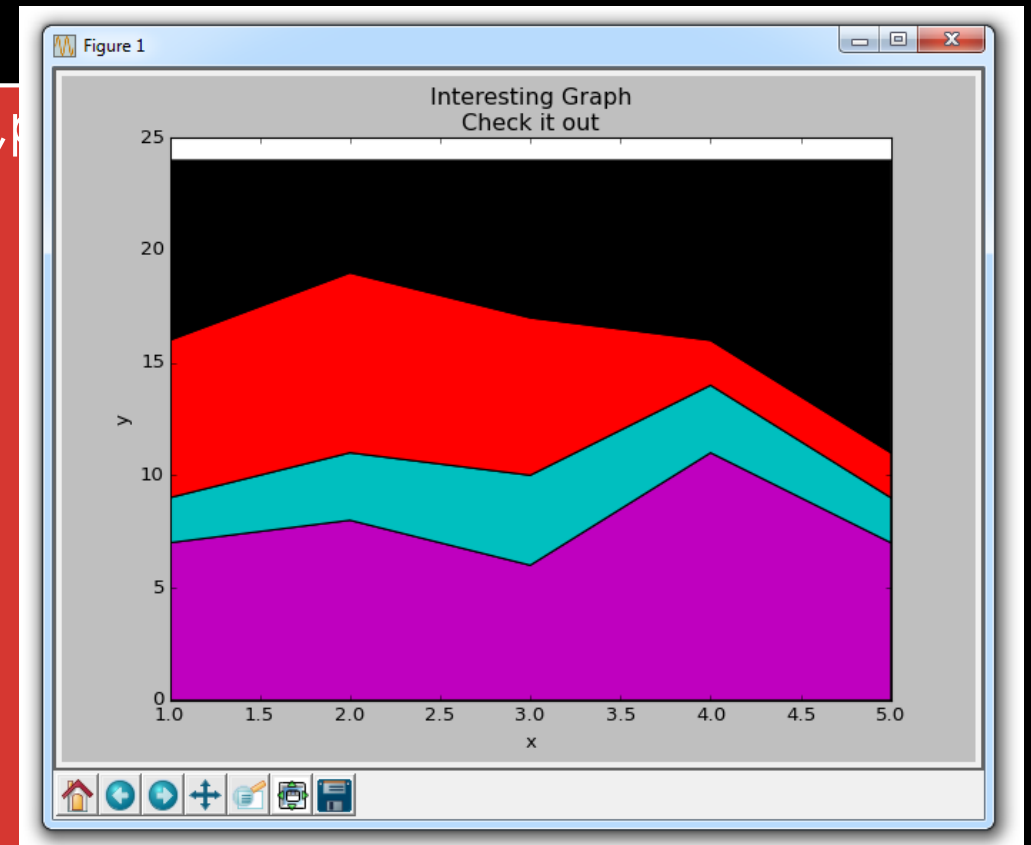


<https://pythonprogramming.net/scatter-plot-matplotlib-tutorial/?completed=/bar-chart-histogram-matplotlib-tutorial/>

STACK PLOTS WITH MATPLOTLIB

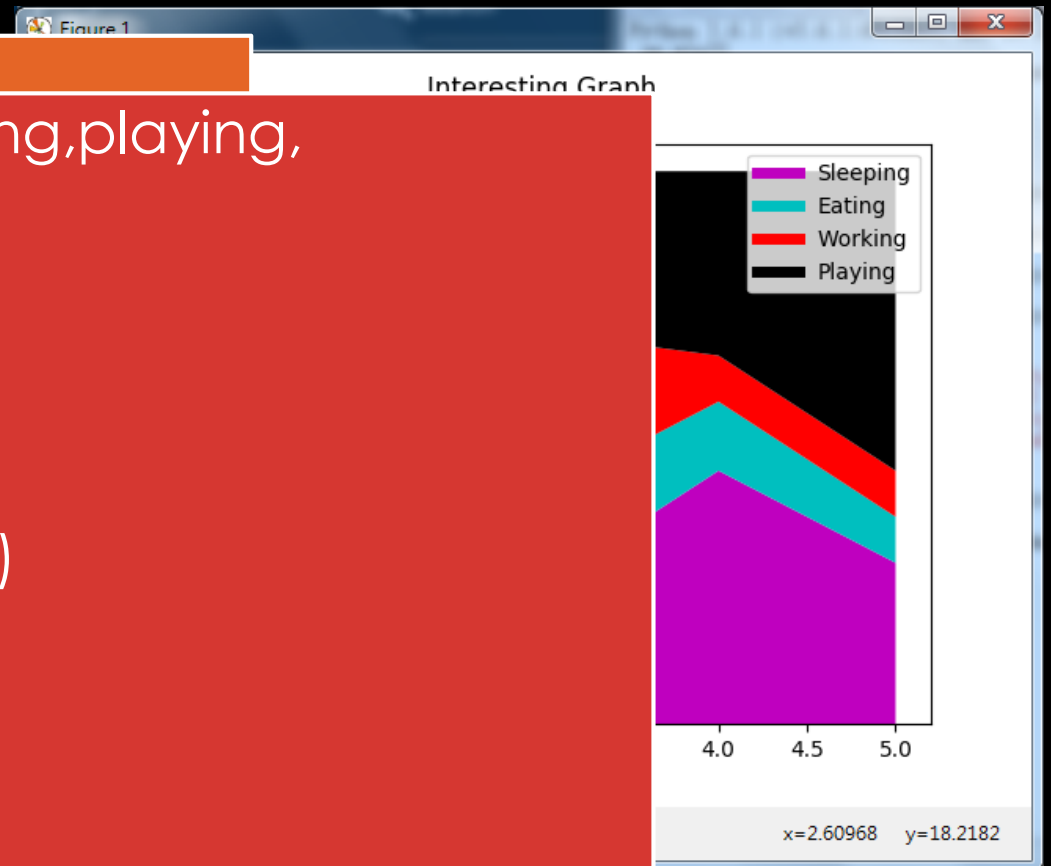
```
#plot6.py
# plt.stackplot(days, sleeping,eating,working,
colors=['m','c','r','k'])

plt.xlabel('x')
plt.ylabel('y')
plt.title('Interesting Graph\nCheck it out')
plt.legend()
plt.show()
```



STACK PLOTS WITH MATPLOTLIB

```
#plot6.py
#
plt.stackplot(days, sleeping,eating,working,playing,
              colors=['m','c','r','k'])
plt.xlabel('x')
plt.ylabel('y')
plt.title('Interesting Graph\nCheck it out')
plt.legend()
plt.show()
```



PIE CHARTS WITH MATPLOTLIB

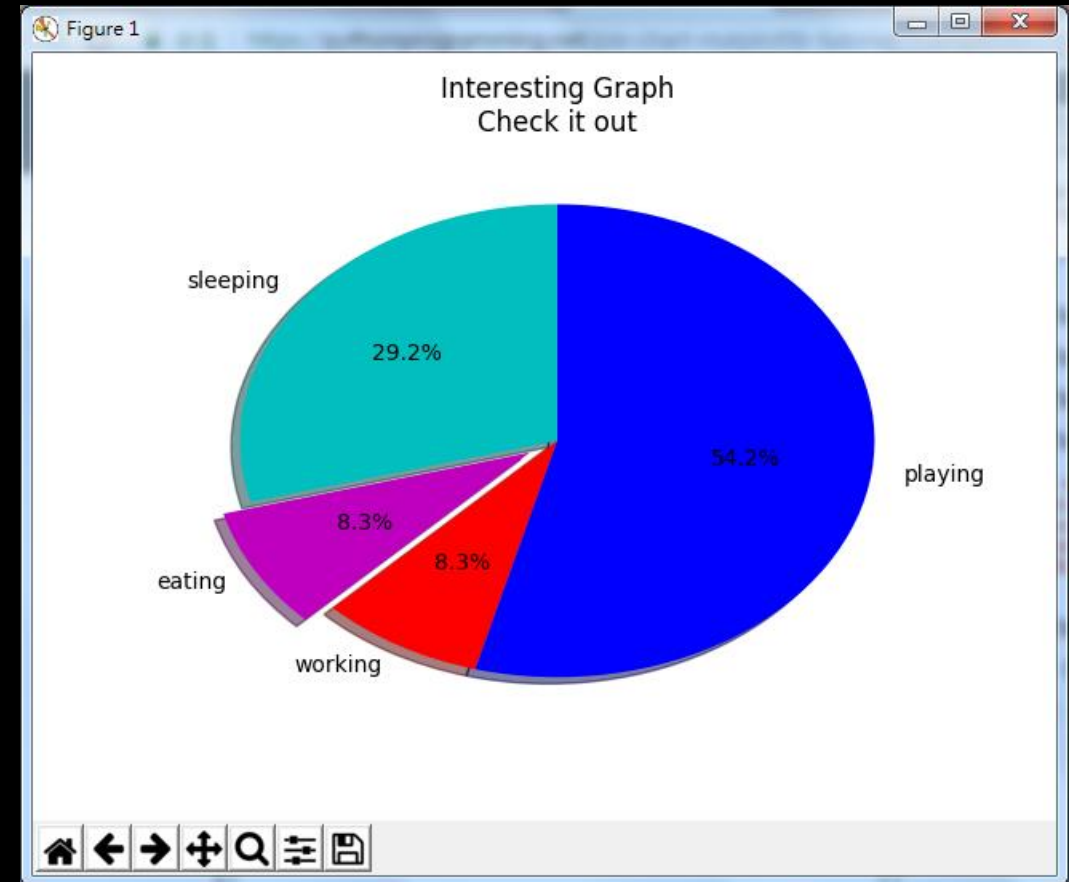
```
#plot7.py
#pie chart

import matplotlib.pyplot as plt

slices = [7,2,2,13]
activities = ['sleeping','eating','working','playing']
cols = ['c','m','r','b']

plt.pie(slices,
        labels=activities,
        colors=cols,
        startangle=90,
        shadow=True,
        explode=(0,0.1,0,0),
        autopct='%1.1f%%')

plt.title('Interesting Graph\nCheck it out')
plt.show()
```

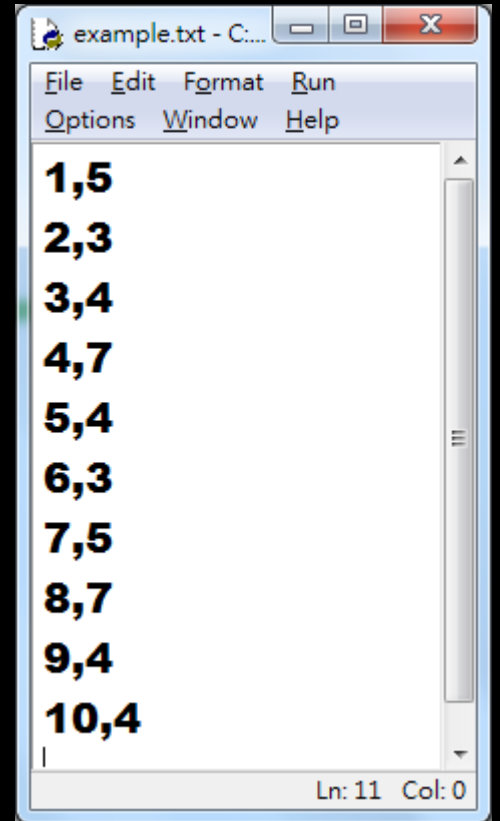


LOADING DATA FROM FILES FOR MATPLOTLIB

```
#plot&
#load plt.plot(x,y, label='Loaded from file!')
plt.xlabel('x')
import plt.ylabel('y')
import plt.title('Interesting Graph\nCheck it out')
plt.legend()

x = []
y = [] plt.show()

with o
plot
for m
x
y
```



```
example.txt - C...
File Edit Format Run
Options Window Help
1,5
2,3
3,4
4,7
5,4
6,3
7,5
8,7
9,4
10,4
Ln: 11 Col: 0
```

Example.txt

LOADING DATA FROM FILES FOR MATPLOTLIB

```
plot8_loadFile.py - C:/Users/anny/Desktop/Anny/DataAnalysis/plot8_loadFile.py (3.6.1)
File Edit Format Run Options Window Help

#plot8.py
#load data from file

import matplotlib.pyplot as plt
import csv

x = []
y = []

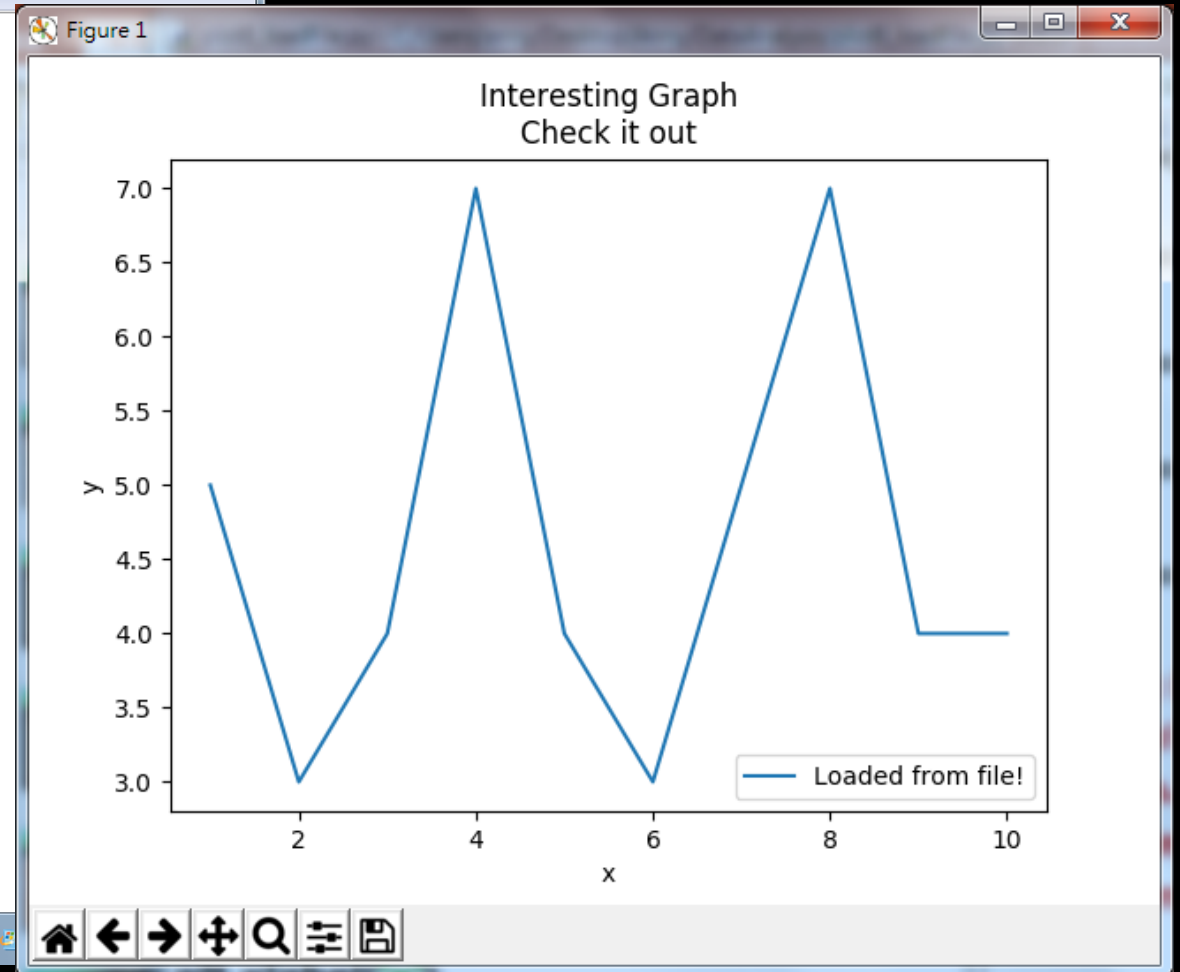
with open('example.txt','r') as csvfile:
    plots = csv.reader(csvfile, delimiter=',')
    for row in plots:
        x.append(int(row[0]))
        y.append(int(row[1]))

plt.plot(x,y, label='Loaded from file!')
plt.xlabel('x')
plt.ylabel('y')
plt.title("Interesting Graph\nCheck it out")
plt.legend()
plt.show()
```

```
example.txt - C:...
File Edit Format Run
Options Window Help

1,5
2,3
3,4
4,7
5,4
6,3
7,5
8,7
9,4
10,4

Ln: 11 Col: 0
```



USING THE **NUMPY** MODULE TO LOAD OUR FILES

- **cmd** (開啟命令提示字元視窗)
- **pip install numpy**

NumPy(Numeric Python) 是Python語言的一個擴充程式庫。支援高階大量的維度陣列與矩陣運算，此外也針對陣列運算提供大量的[數學函式函式庫](#)。

USING THE **NUMPY** MODULE TO LOAD OUR FILES

```
plot9_loadFile_numpy.py - C:/Users/anny/Desktop/Anny/DataAnalysis/plot9_loadFile_numpy.py (3.6.1)
File Edit Format Run Options Window Help
#plot9
#load file using numpy

import matplotlib.pyplot as plt
import numpy as np

x, y = np.loadtxt('example.txt', delimiter=',', unpack=True)
plt.plot(x,y, label='Loaded from file!')

plt.xlabel('x')
plt.ylabel('y')
plt.title('Interesting Graph\nCheck it out')
plt.legend()
plt.show()

Ln: 15 Col: 0
```

