

## chapter 35

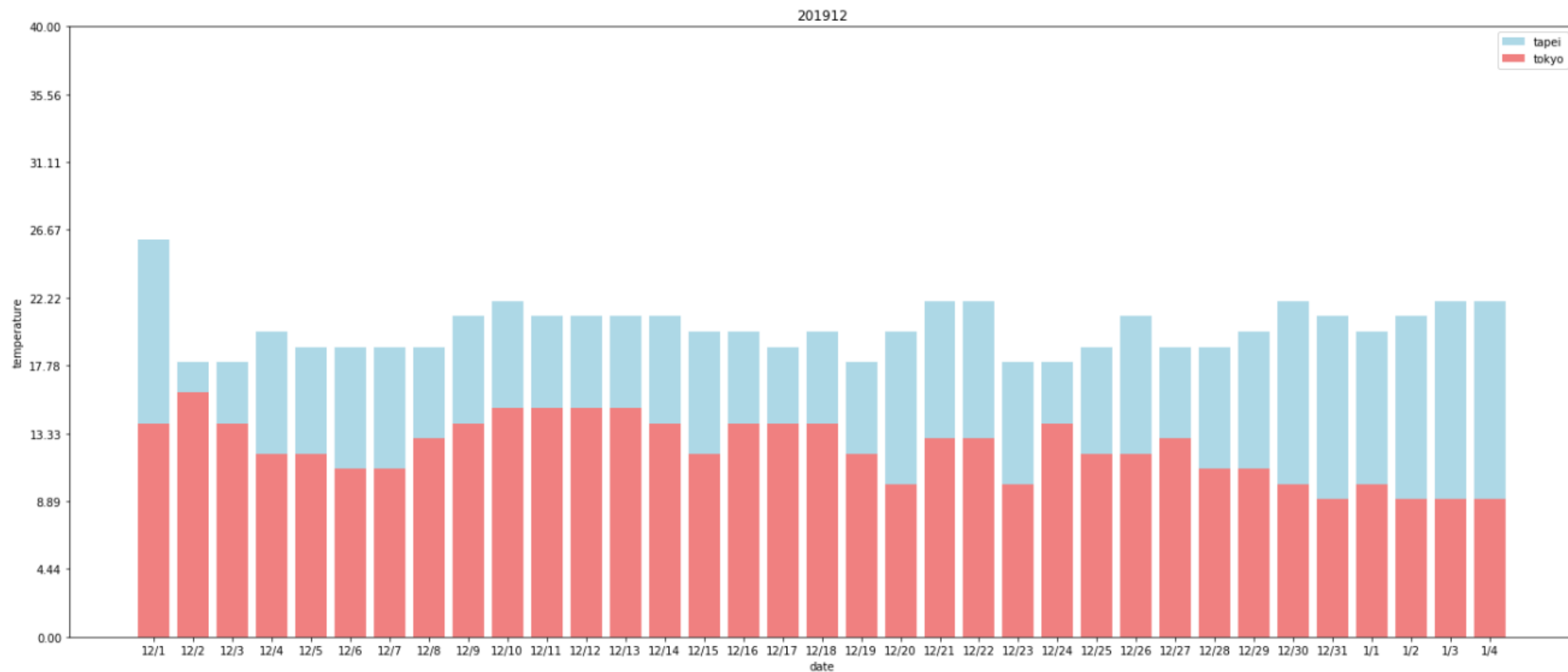
# 資料視覺化(柱狀圖與直方圖)

---

- 柱狀圖/直方圖
- 大樂透數據分析
- NBA數據分析

# 柱狀圖

- 適合進行數據間的比較




## ■ 基本使用方式

```
plt.bar(left, height)
```



left為x軸資料  
height為y軸資料(實際資料)

```
plt.bar(left, height, format)
```

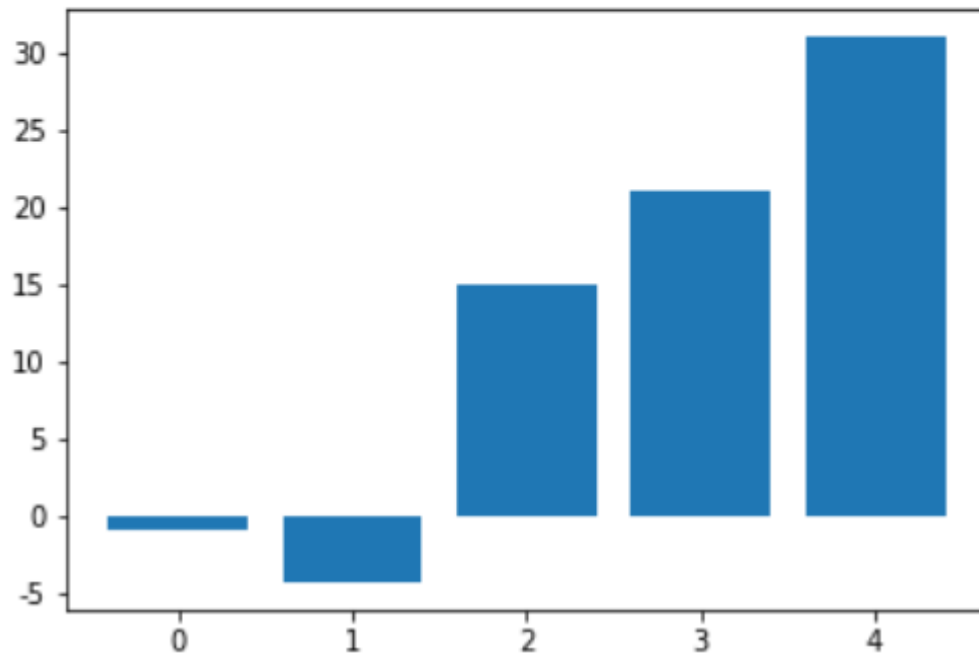


format.. 可以傳入其他格式

## ■ 範例

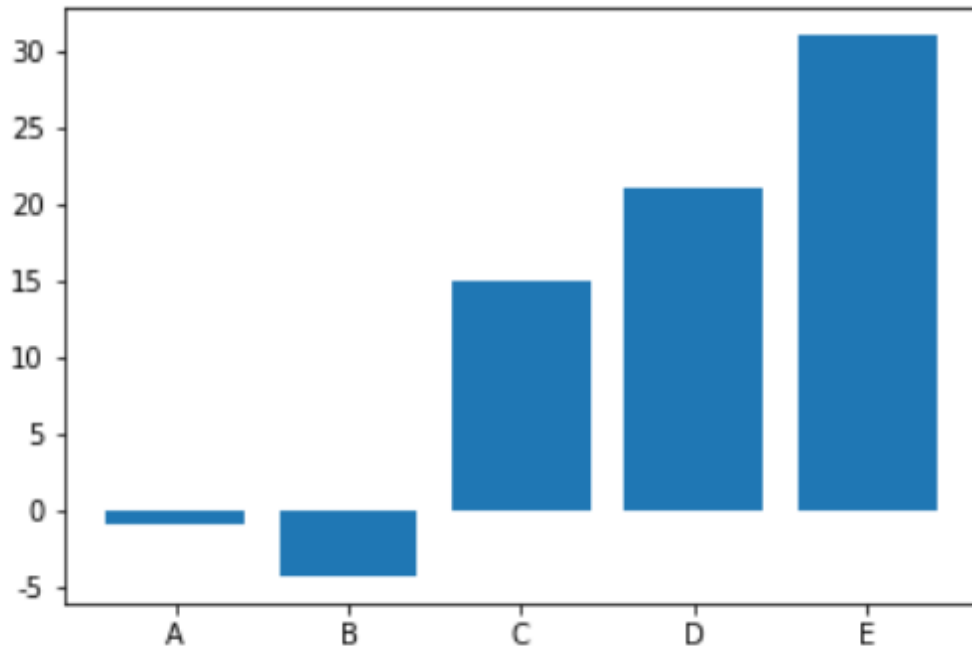
```
1 data=[-1,-4.3,15,21,31]
2
3
4 plt.bar(range(len(data)),data)
5
6 plt.show()
```

← 一定需要left,height的資料，且長度要一致



## ■ 更改X軸資料

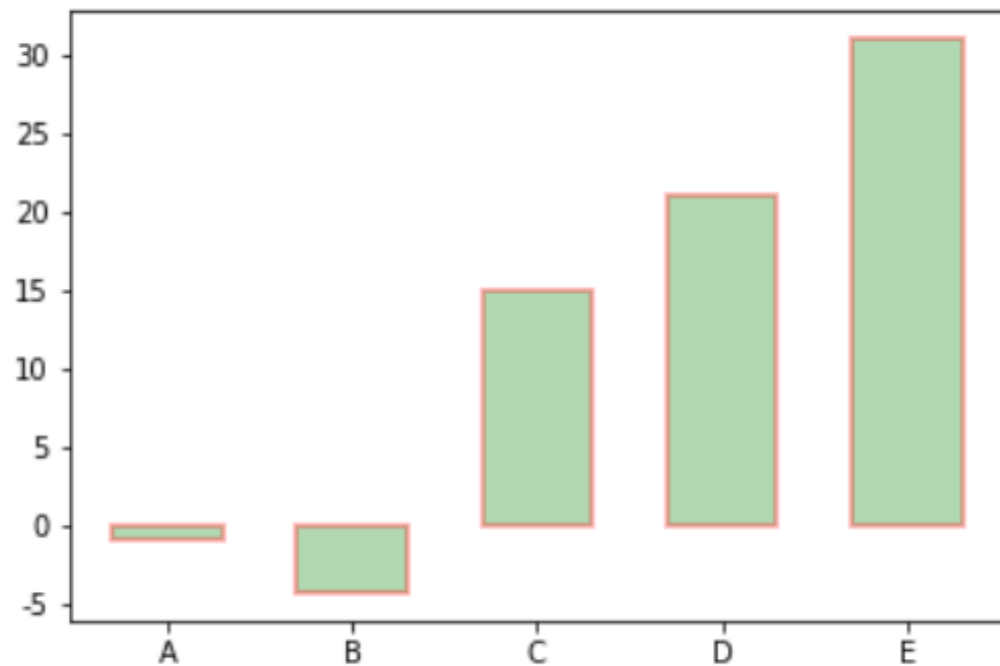
```
1 data=[-1,-4.3,15,21,31]
2
3
4 plt.bar(['A','B','C','D','E'],data)
5
6
7 plt.show()
```



## ■ 其他參數

- 1. `left`：x軸的位置序列，一般採用`range`函數產生一個序列
- 2. `height`：y軸的數值序列，柱形圖的高度，需要展示的數據
- 3. `alpha`：透明度
- 4. `width`：為柱形圖的寬度，預設為0.8
- 5. `color` or `facecolor`：柱形圖填充的顏色
- 6. `edgecolor`：圖形邊緣顏色
- 7. `label`：圖像代表的含義
- 8. `linewidth` or `linewidths` or `lw`：邊緣或線的寬度

```
1 data=[-1,-4.3,15,21,31]
2
3
4 plt.bar(['A','B','C','D','E'],data,alpha=0.3,width=0.6,edgecolor='red',color='green',
5         linewidth=2.0)
6
7 plt.show()
```



## ■ 程式練習

- 讀取weather.csv
- 展現天氣柱狀圖
- 使用index當作x座標資料
- 使用tokyo行資料當作y座標資料
- color 設定為"lightblue"
- label 設定為"tokyo"

```
1 import pandas as pd
2
3 df1=pd.read_csv('weather.csv',index_col=0)
4
5 df1
```

	tokyo	taipei
12/1	14	26
12/2	16	
12/3	14	18
12/4	12	20
12/5	12	19
12/6	11	19
12/7	11	19
12/8	13	19
12/9	14	21
12/10	15	22
12/11	15	21
12/12	15	21
12/13	15	21
12/14	14	21
12/15	12	20

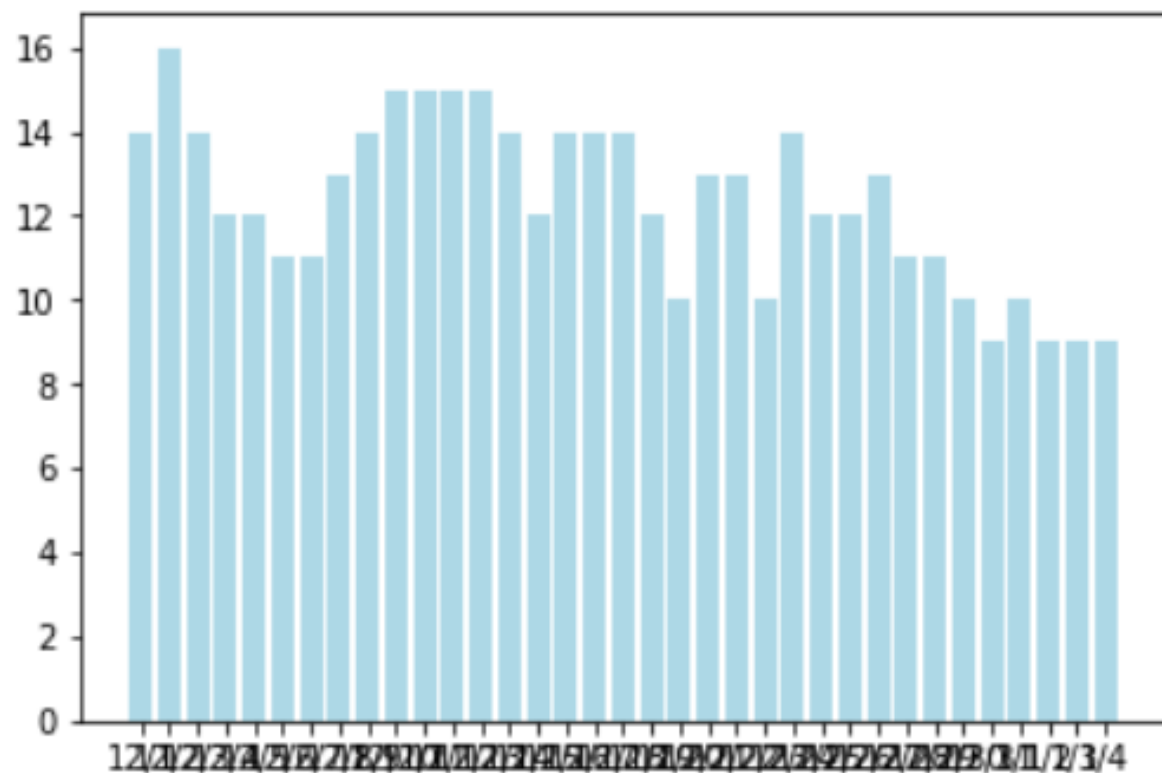
tokyo為y軸  
座標資料

使用index當作  
x座標資料



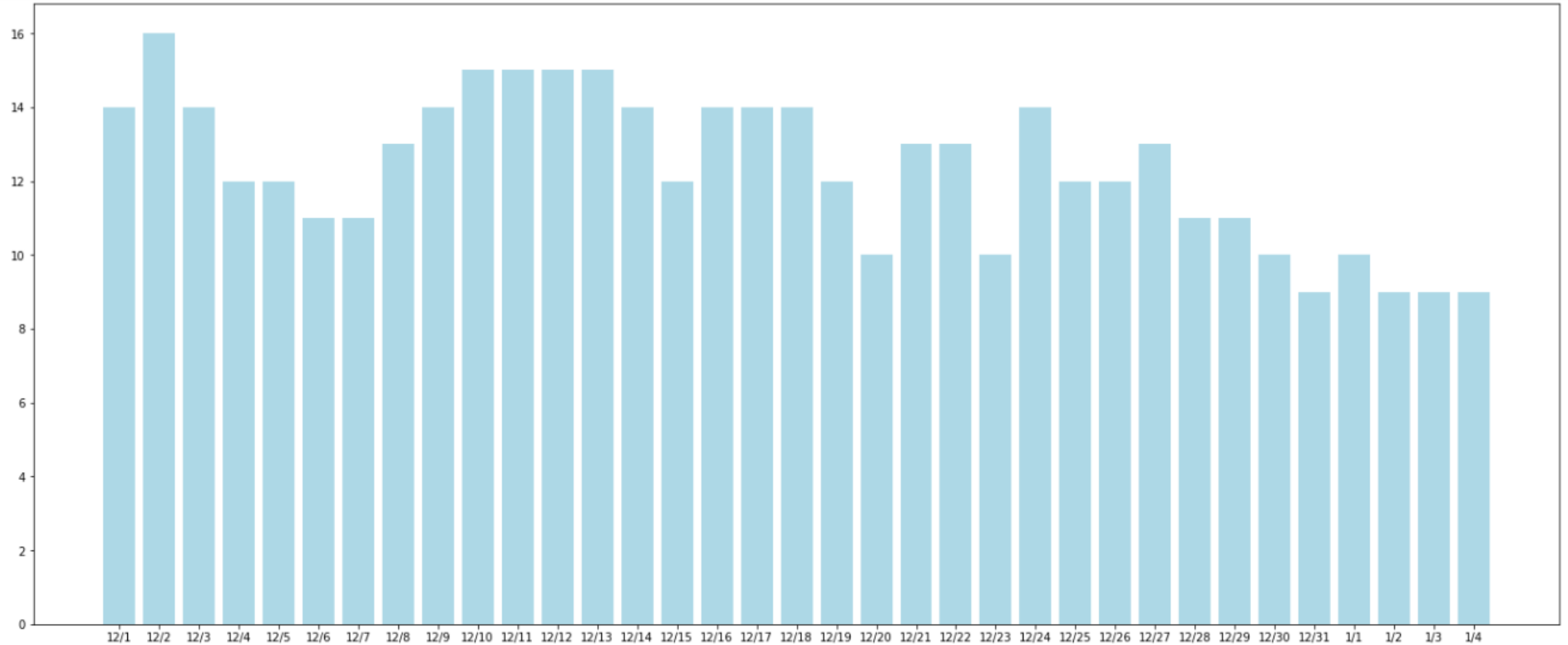
## ■ 預設Size問題

```
1 plt.bar(df1.index,df1['tokyo'],color='lightblue',label='tokyo')  
2  
3 plt.show()
```



## ■ 更改尺寸

■ `plt.figure(figsize=(24,10))`

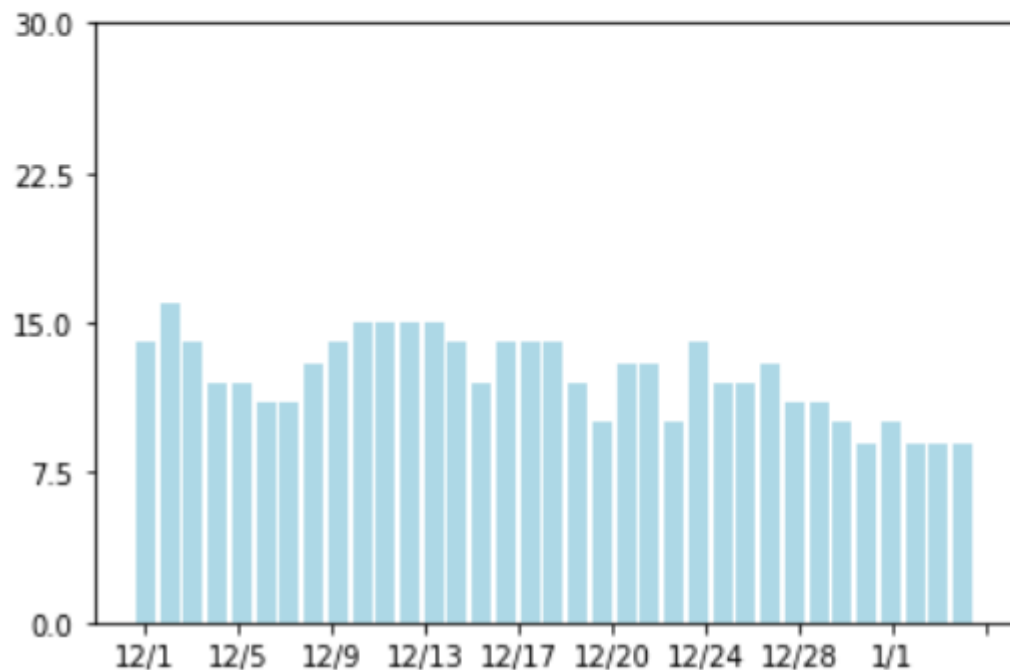


## ■ 更改刻度

### ■ 使用xticks, yticks

```
1 import numpy as np
2 #plt.figure(figsize=(24,10))
3 plt.bar(df1.index,df1['tokyo'],color='lightblue',label='tokyo')
4 #設定x刻度
5 plt.xticks(np.linspace(0,len(df1.index),10))
6 plt.yticks(np.linspace(0,30,5))
7 plt.show()
8
```

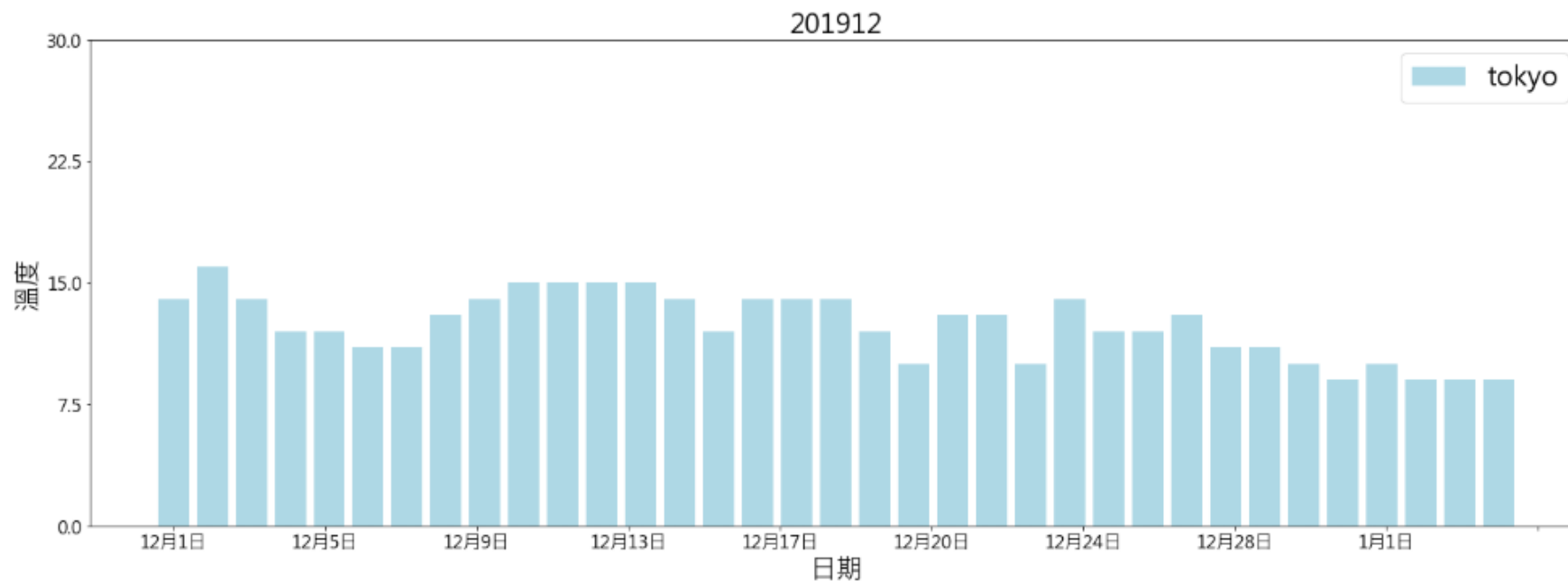
搭配np.linspace重新設定刻度



## ■ 設定FontSize



```
1 import numpy as np
2
3 plt.figure(figsize=(24,8))
4
5 plt.bar(df1.index,df1['tokyo'],color='lightblue',label='tokyo')
6 #設定x刻度
7 plt.xticks(np.linspace(0,len(df1.index),10),fontsize=16)
8 plt.yticks(np.linspace(0,30,5),fontsize=16)
9 plt.xlabel('日期',fontsize=24)
10 plt.ylabel('温度',fontsize=24)
11 plt.title('201912',fontsize=24)
12 plt.legend(fontsize=24)
13 plt.show()
```



## ■ 使用barh橫向顯示

```
1 import numpy as np
2
3 plt.figure(figsize=(24,10))
4
5 plt.barh(df1.index,df1['tokyo'],color='lightblue',label='tokyo',height=0.5)
6 #設定x 刻度
7 plt.xticks(np.linspace(0,len(df1.index),10))
8 plt.yticks(np.linspace(0,30,5))
9 plt.ylabel('date')
10 plt.xlabel('temperature')
11 plt.title('201912')
12 plt.legend()
13 plt.show()
```

搭配np.linspace重新設定刻度



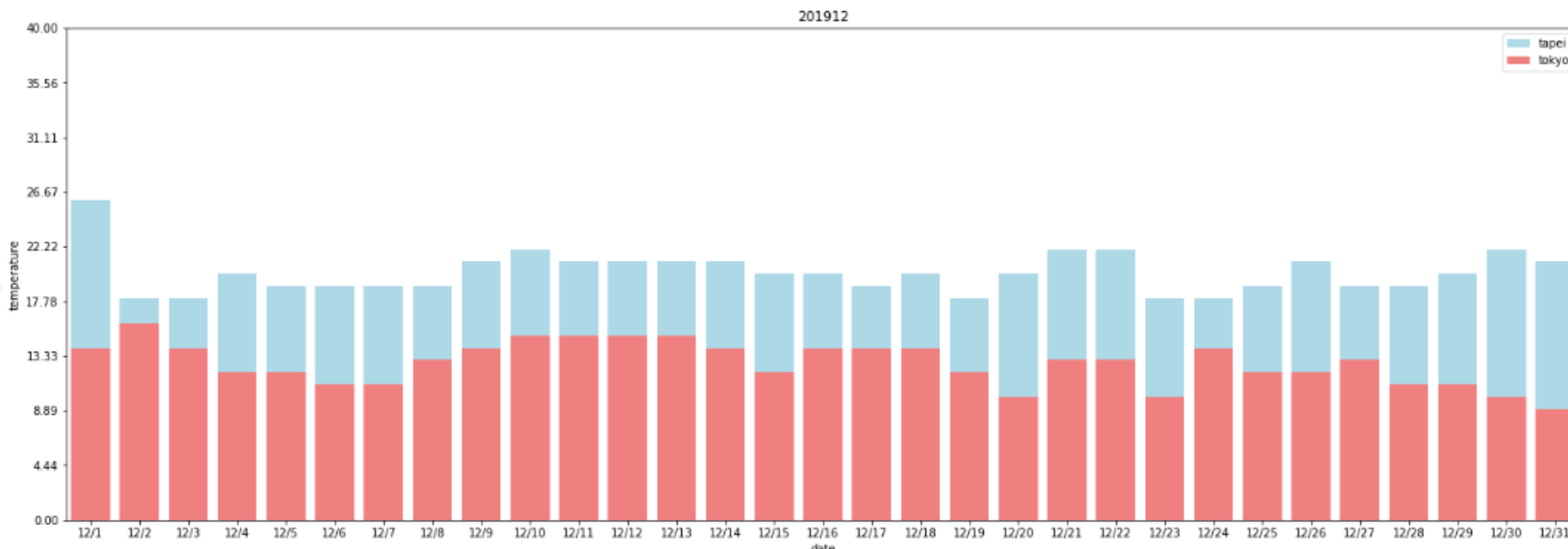
# ■ 多直條圖疊加顯示

先後順序顯示

設定yticks  
xlim

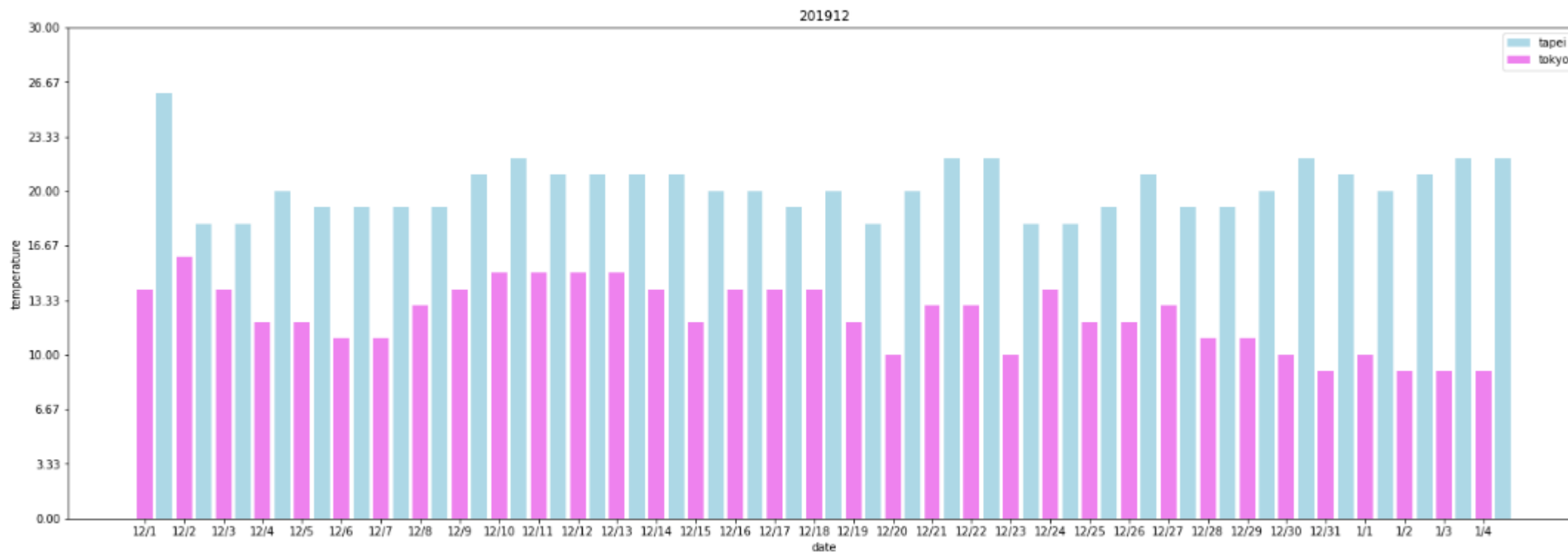
```
1 import numpy as np
2
3 plt.figure(figsize=(24,8))
4 plt.bar(df1.index,df1['taipei'],color='lightblue',label='taipei')
5 plt.bar(df1.index,df1['tokyo'],color='lightcoral',label='tokyo')
6 #設定y刻度(溫度區分成10格)
7 plt.yticks(np.linspace(0,40,10))
8 plt.xlim(-0.5,30.5)
9 #plt.xticks(np.linspace(0,len(df1.index),20))
10 plt.xlabel('date')
11 plt.ylabel('temperature')
12 plt.title('201912')
13 plt.legend()
14 plt.savefig('weather_bar.png')
15 plt.show()
```

可改變alpha值  
顯示



## ■ 雙柱狀圖(間隔顯示)

- 將X軸座標加倍
- 穿插顯示



## ■ 將X軸顯示

### ■ 設定在index為偶數上

```
1 index=range(len(df1.index)*2)
2
3 index
```

取兩倍長度

```
range(0, 70)
```

```
1 plt.figure(figsize=(24,8))
2
3 plt.bar(index[1::2],df1['taipei'],color='lightblue',label='taipei')
4 plt.bar(index[0::2],df1['tokyo'],color='violet',label='tokyo')
5 #設定y刻度(溫度區分成10格)
6 plt.yticks(np.linspace(0,30,10))
7 #xticks(用偶數來顯示座標上)
8 plt.xticks(index[0::2],df1.index)
9 plt.xlabel('date')
10 plt.ylabel('temperature')
11 plt.title('201912')
12 plt.legend()
13 plt.savefig('wheather_bar_1.png')
14 plt.show()
```

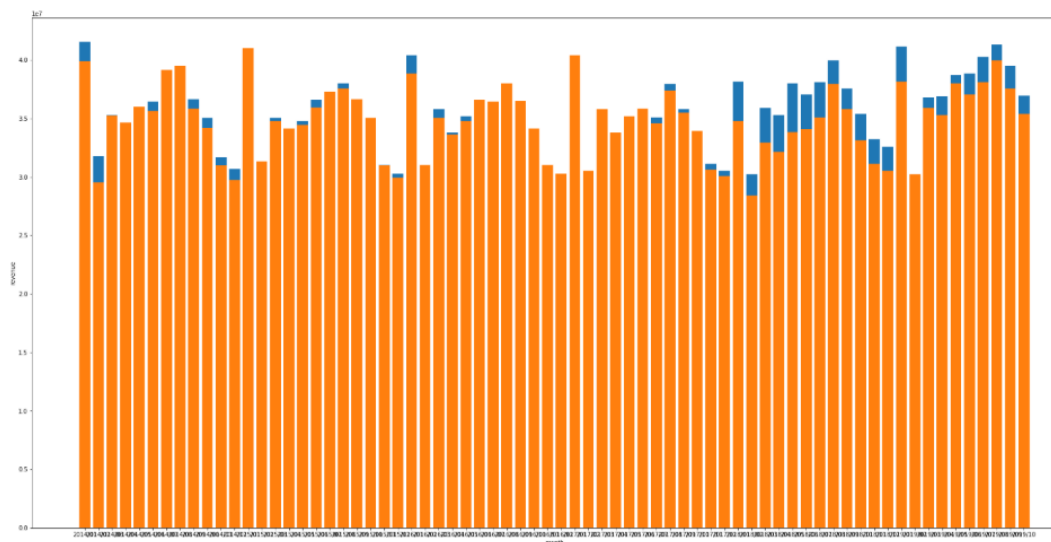
穿插顯示

設定X座標  
並顯示以偶數為主

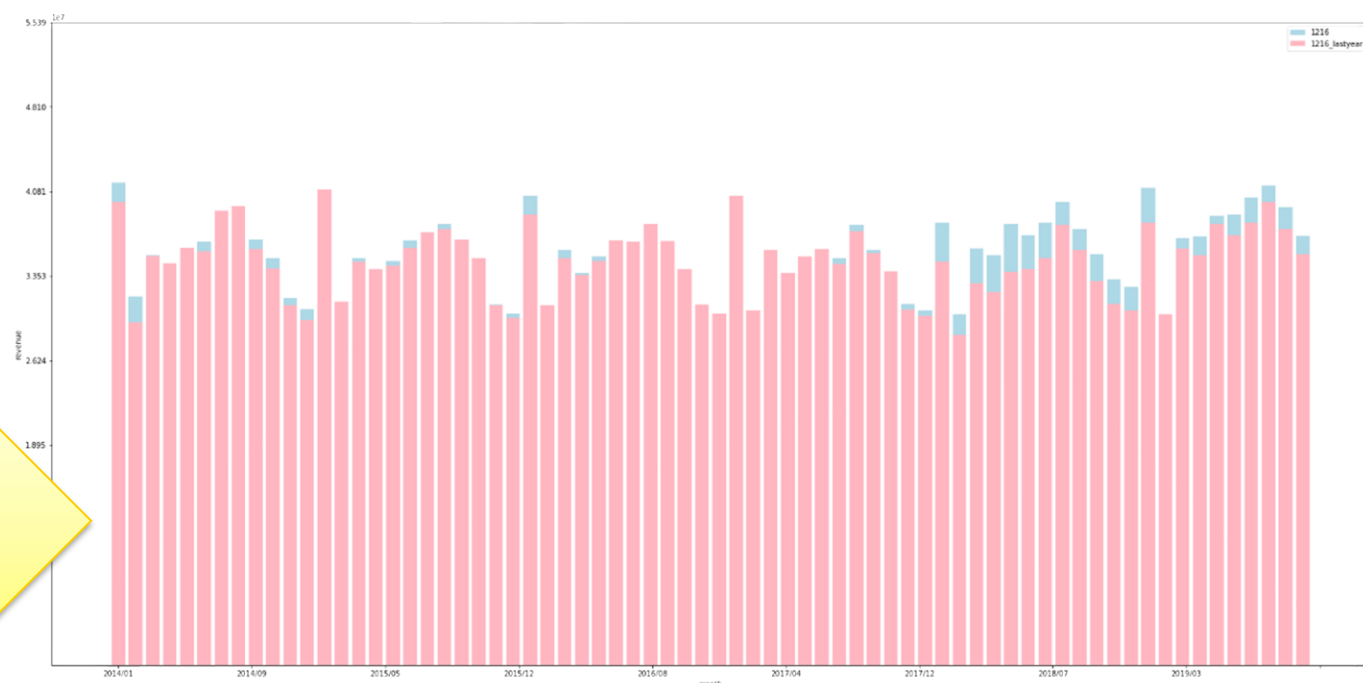


## ■ 程式練習

- 讀取統一營收表\_201910.csv
- 輸出疊加柱狀圖(今年跟去年營收比較)
- 修改顯示效果



修改後輸出



## ■ 修改練習

- `plt.figure(figsize=(32, 12))`
- 今年營收顯示
  - `color='lightblue', label='1216'`
- 去年營收顯示
  - `color=pink', label='1216 last year'`
- y軸使用`yticks`方法重新設定刻度
  - 使用最高營收x1.3及最低營收x0.3 設定六個區間
  - 使用`linspace`
- x軸使用`xticks`設定重新設定刻度
  - 設定10個刻度日期
  - 使用`linspace`

## ■ 讀取用程式碼

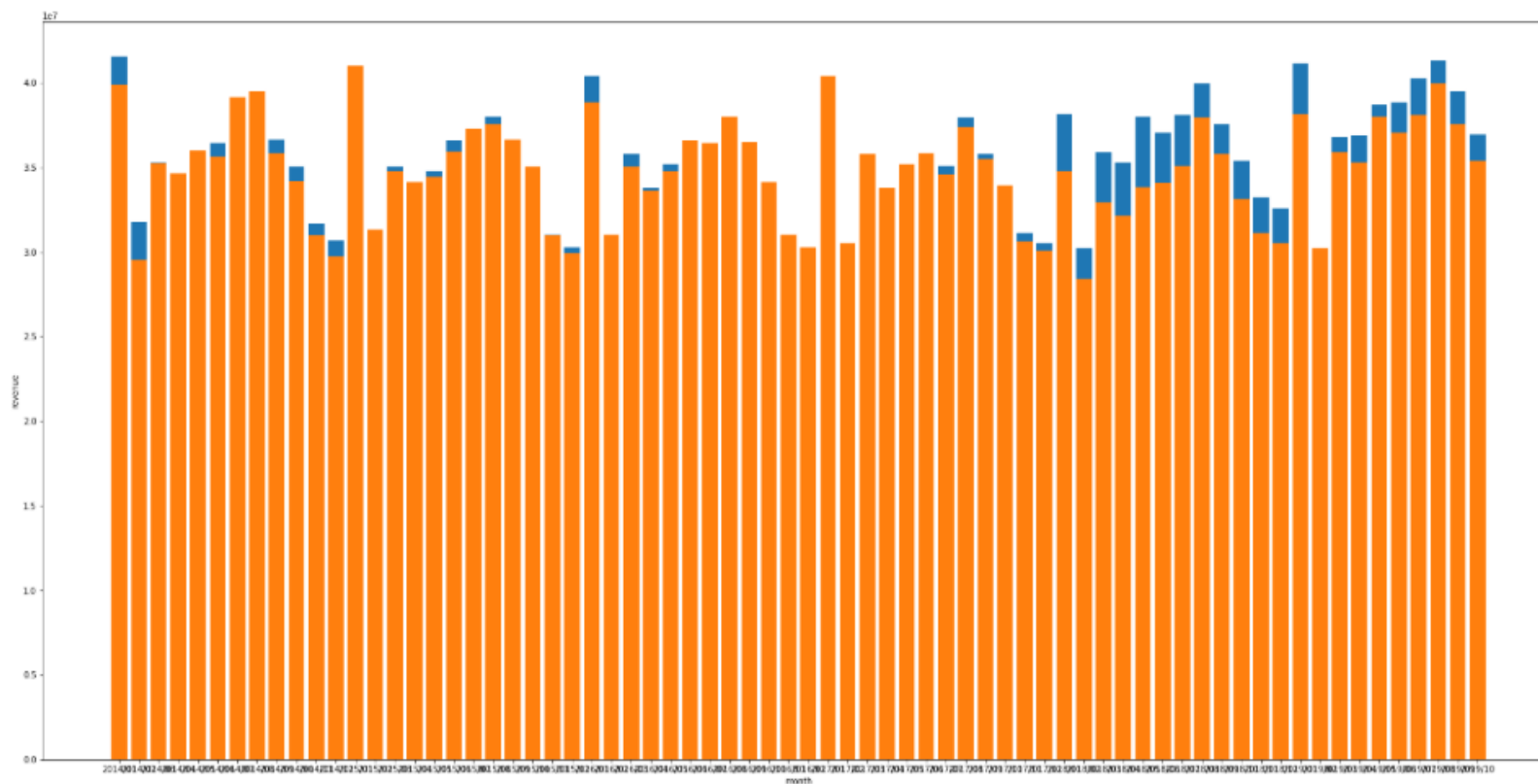


```
1 import pandas as pd
2
3 df1=pd.read_csv('統一營收表_201910.csv',engine='python',index_col=0,encoding='utf-8-sig',thousands=",")
4
5 df1=df1[::-1]
6
7 df1
```

2018/02	30250748	-20.7	28430864	6.4	68412720	63202920	8.2
2018/03	35875544	18.6	32926232	9.0	104288264	96129152	8.5
2018/04	35313668	-1.6	32149960	9.8	139601936	128279112	8.8
2018/05	37995652	7.6	33849168	12.3	177597584	162128288	9.5
2018/06	37048408	-2.5	34105064	8.6	214645984	196233344	9.4
2018/07	38128164	2.9	35072216	8.7	252774144	231305552	9.3
2018/08	39945628	4.8	37940820	5.3	292719776	269246368	8.7
2018/09	37539416	-6.0	35792012	4.9	330259200	305038400	8.3
2018/10	35398600	-5.7	33118532	6.9	365657792	338156928	8.1
2018/11	33229600	-6.1	31140540	6.7	398887392	369297472	8.0
2018/12	32590936	-1.9	30552612	6.7	431478336	399850080	7.9
2019/01	41144408	26.2	38161968	7.8	41144408	38161968	7.8
2019/02	29826476	-27.5	30250748	-1.4	70970888	68412720	3.7

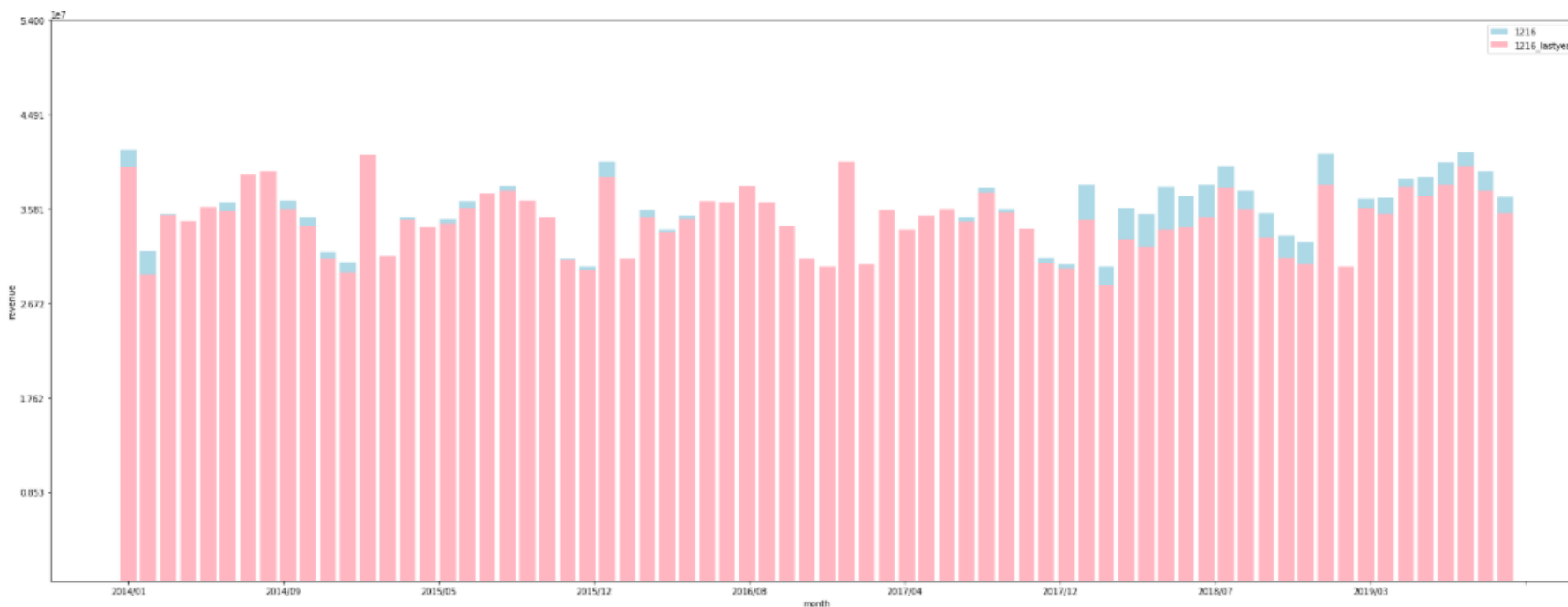
## ■ 原始修改用程式碼

```
1 import numpy as np
2
3 plt.figure(figsize=(32,16))
4
5 plt.bar(df1.index,df1['當月營收'])
6
7 plt.bar(df1.index,df1['去年同期營收'])
8
9 plt.xlabel('month')
10 plt.ylabel('revenue')
11
12 plt.show()
```



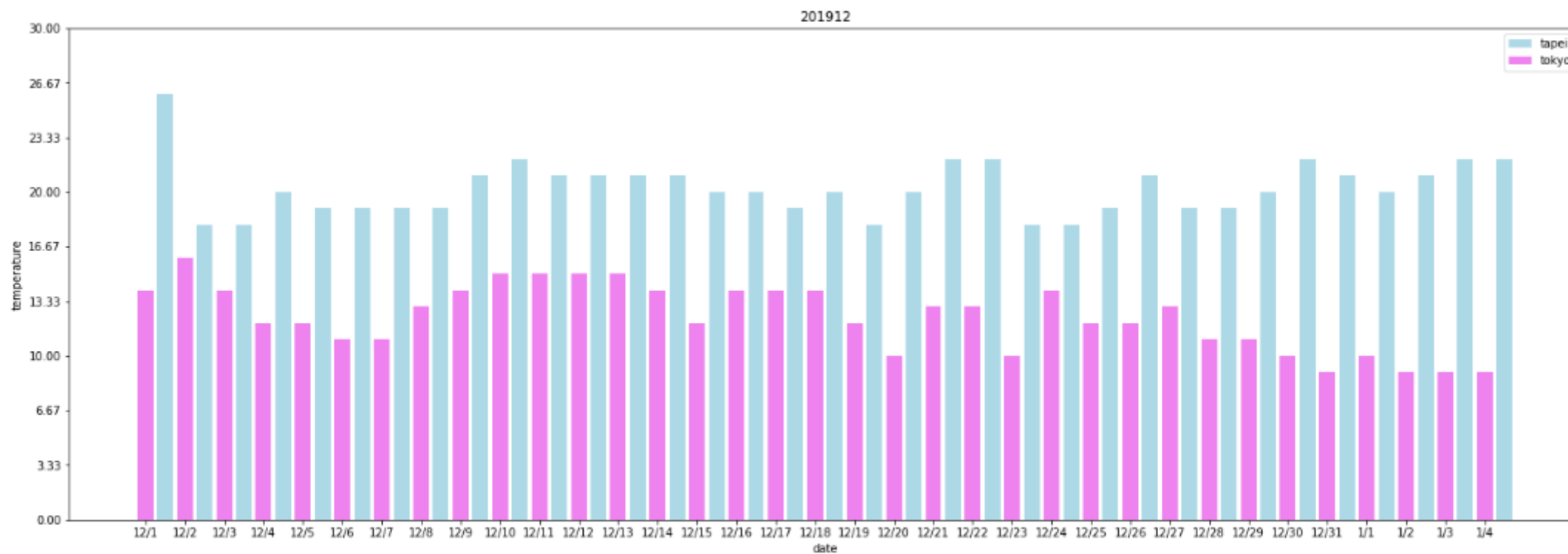
## ■ 修改後程式碼

```
1 import numpy as np
2
3 plt.figure(figsize=(32,12))
4 plt.bar(df1.index,df1['當月營收'],color='lightblue',label='1216')
5 plt.bar(df1.index,df1['去年同月營收'],color='lightpink',label='1216_lastyear')
6 #設定y刻度(營收區分成10格)
7 plt.yticks(np.linspace(df1['當月營收'].min()*0.3,df1['當月營收'].max()*1.3,6))
8 #資料總長度，切分格數
9 plt.xticks(np.linspace(0,len(df1.index),10))
10 plt.xlabel('month')
11 plt.ylabel('revenue')
12 plt.legend()
13
14 plt.show()
```



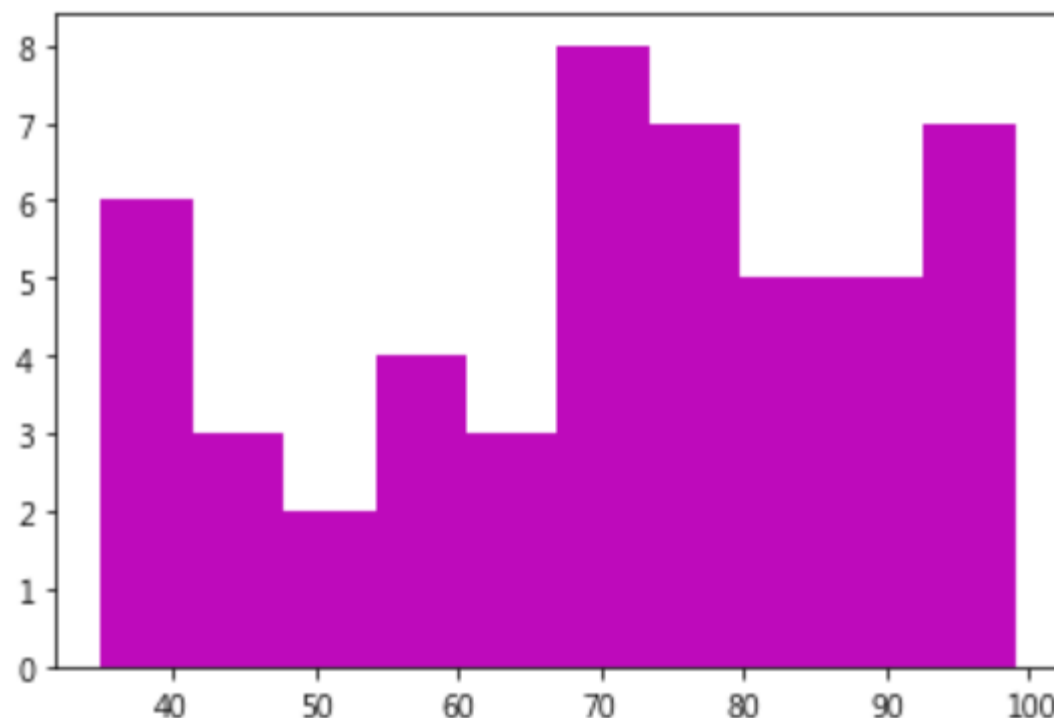
## ■ 程式練習

### ■ 修改顯示為雙柱狀圖



# 直方圖

- 主要是用於顯示數值分布的情形，是一種次數分配表，方便看出數據出現的頻率。



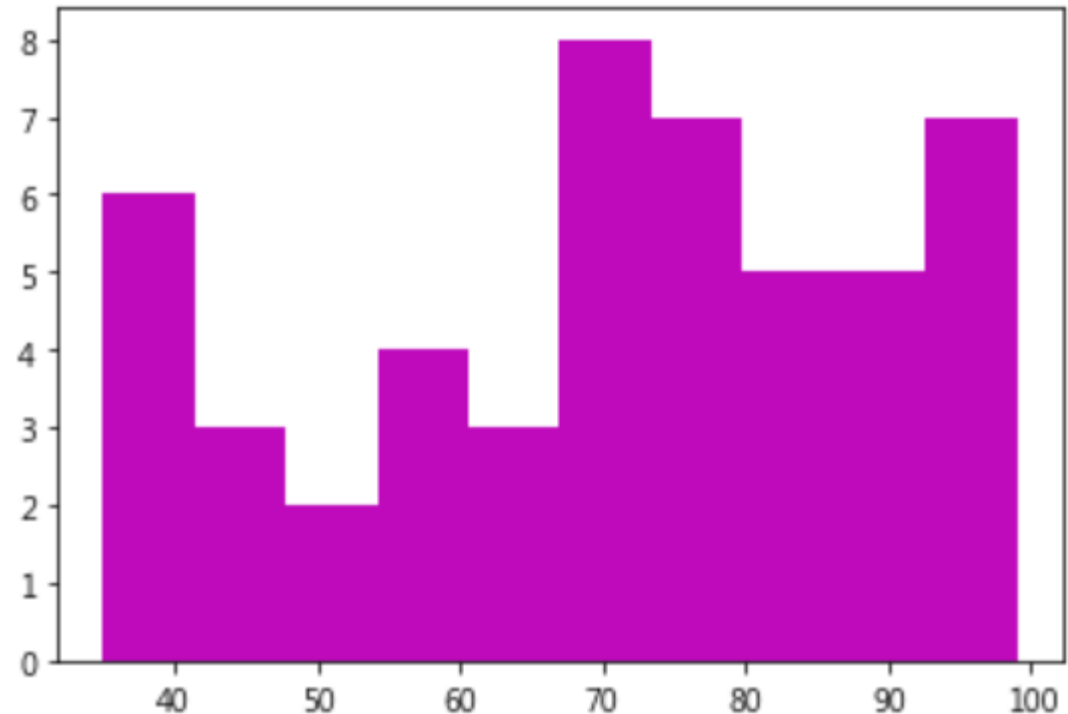
## ■ 基本使用方式

```
plt.hist(datas, nums, format)
```

datas → 資料  
nums → 分割的區域  
format → 其他格式

nums → x軸分割的區間

datas → y軸出現的頻率





## ■ 產生50組數據

```
import numpy as np
```

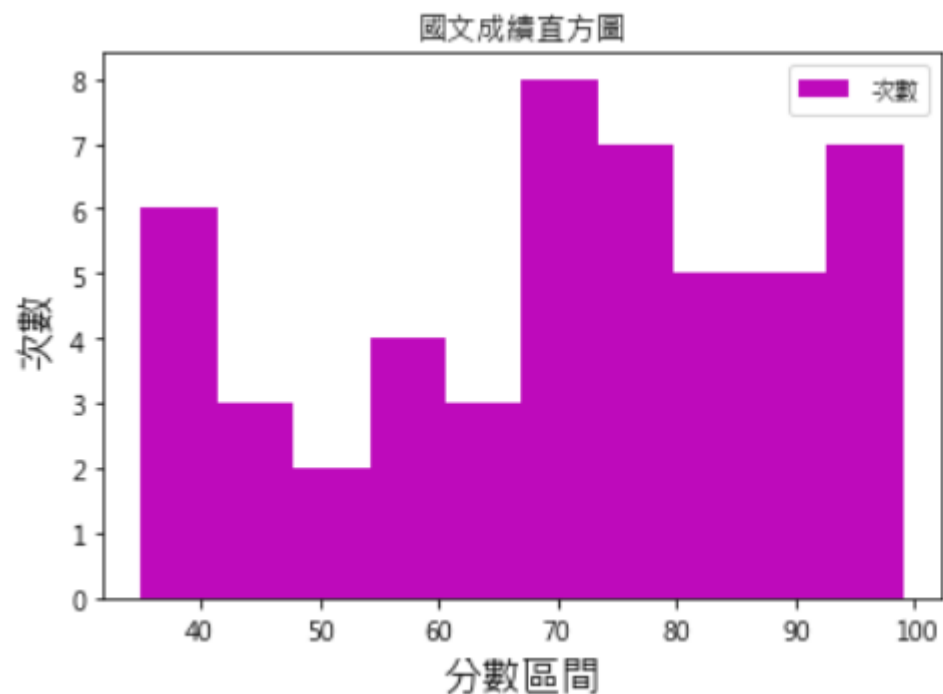
```
scores=np.random.randint(35,101,50)
```

```
scores|
```

```
y([69, 40, 86, 67, 56, 90, 98, 78, 86, 38, 93, 69, 91, 64, 55, 54, 72,  
    85, 43, 57, 74, 62, 79, 96, 87, 99, 35, 39, 42, 68, 68, 90, 92, 61,  
    78, 47, 55, 38, 99, 83, 76, 81, 75, 41, 77, 94, 70, 51, 98, 72])
```

## ■ 設定X軸為10個分數區間

```
1 nums=10
2
3 plt.hist(scores,nums,color='m',label='次數')
4
5 plt.xlabel('分數區間',fontsize=16)
6 plt.ylabel('次數',fontsize=16)
7
8 plt.title('國文成績直方圖')
9 plt.legend()
10 plt.show()
11
```



x軸分割為10個分數區間

## ■ 取得分數分割區間及出現次數

```
n=plt.hist(scores,nums,color='m',label='次數')  
print(n)
```

```
print('數據分割區域:{}'.format(list(n[1])))  
print('數據出現次數:{}'.format(list(n[0])))
```

數據分割區域:[35.0, 41.4, 47.8, 54.2, 60.6, 67.0, 73.4, 79.80000000000001, 86.2, 92.6, 99.0]

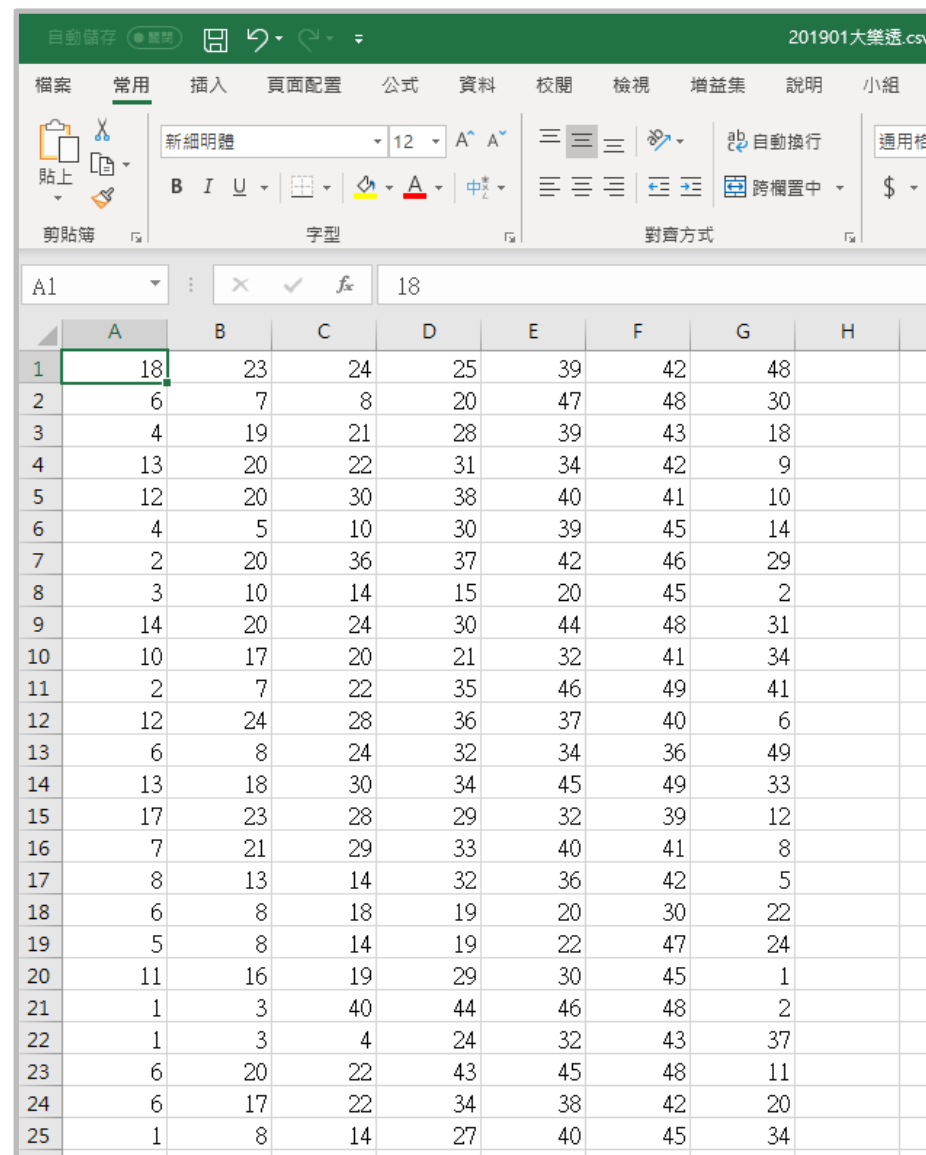
數據出現次數:[6.0, 3.0, 2.0, 4.0, 3.0, 8.0, 7.0, 5.0, 5.0, 7.0]

```
1 print(scores[scores>=92.6])
```

[98 93 96 99 99 94 98]

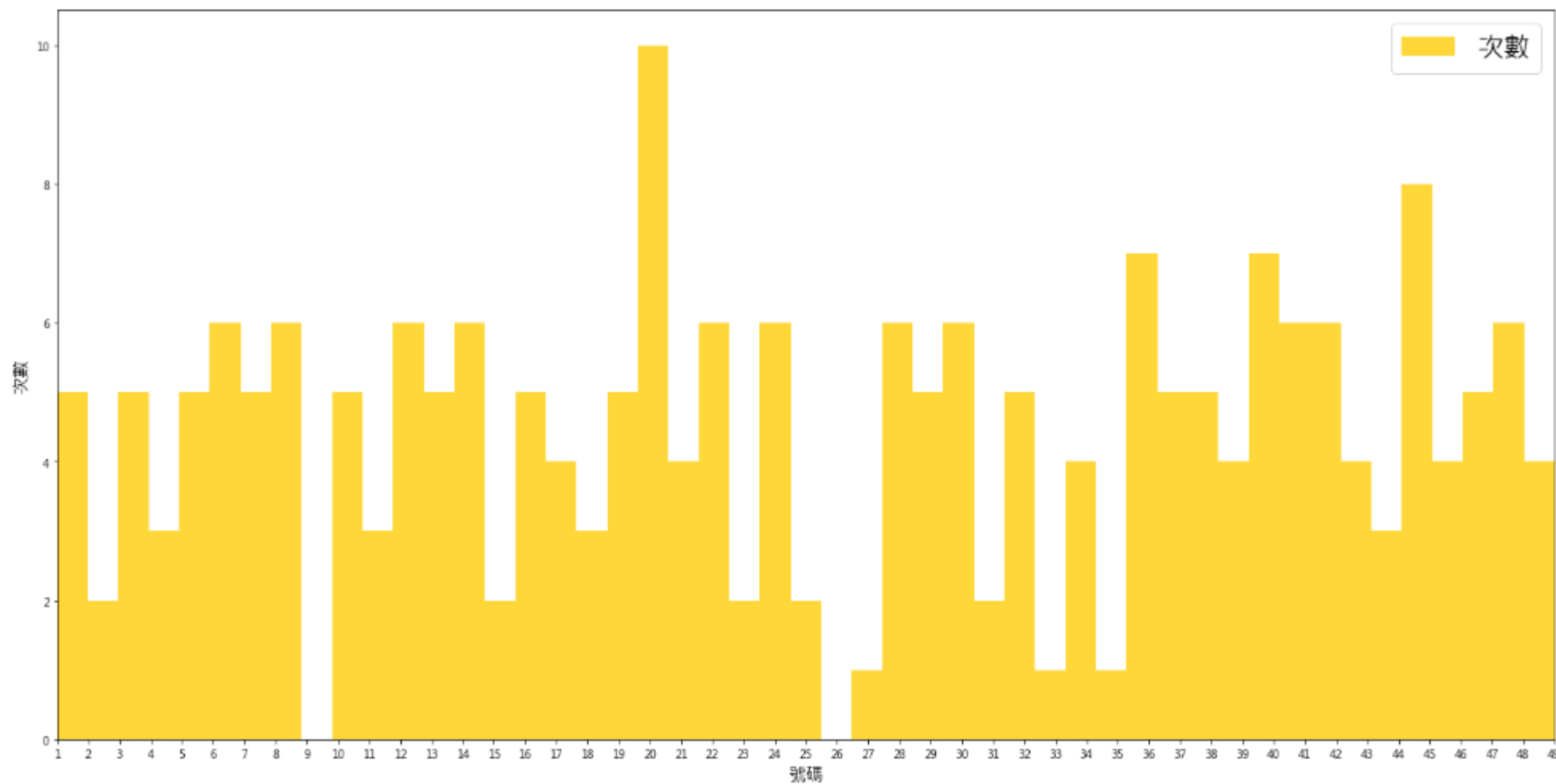
## ■ 程式練習

- 讀取201901大樂透.csv
- 進行直方圖的繪製
- 顯示號碼出現頻率



	A	B	C	D	E	F	G	H
1	18	23	24	25	39	42	48	
2	6	7	8	20	47	48	30	
3	4	19	21	28	39	43	18	
4	13	20	22	31	34	42	9	
5	12	20	30	38	40	41	10	
6	4	5	10	30	39	45	14	
7	2	20	36	37	42	46	29	
8	3	10	14	15	20	45	2	
9	14	20	24	30	44	48	31	
10	10	17	20	21	32	41	34	
11	2	7	22	35	46	49	41	
12	12	24	28	36	37	40	6	
13	6	8	24	32	34	36	49	
14	13	18	30	34	45	49	33	
15	17	23	28	29	32	39	12	
16	7	21	29	33	40	41	8	
17	8	13	14	32	36	42	5	
18	6	8	18	19	20	30	22	
19	5	8	14	19	22	47	24	
20	11	16	19	29	30	45	1	
21	1	3	40	44	46	48	2	
22	1	3	4	24	32	43	37	
23	6	20	22	43	45	48	11	
24	6	17	22	34	38	42	20	
25	1	8	14	27	40	45	34	

## ■ 繪製直方圖



## ■ 出現頻率數據

號碼

```
[ 5.  2.  5.  3.  5.  6.  5.  6.  0.  5.  3.  6.  5.  6.  2.  5.  4.  3.  
  5. 10.  4.  6.  2.  6.  2.  0.  1.  6.  5.  6.  2.  5.  1.  4.  1.  7.  
  5.  5.  4.  7.  6.  6.  4.  3.  8.  4.  5.  6.  4.]
```

號碼1 出現5.0次  
號碼2 出現2.0次  
號碼3 出現5.0次  
號碼4 出現3.0次  
號碼5 出現5.0次  
號碼6 出現6.0次  
號碼7 出現5.0次  
號碼8 出現6.0次  
號碼9 出現0.0次  
號碼10 出現5.0次  
號碼11 出現3.0次  
號碼12 出現6.0次  
號碼13 出現5.0次  
號碼14 出現6.0次

## ■ 資料整理

### ■ 讀取資料

```
1 import numpy as np
2
3 numbers=np.genfromtxt('201901大樂透.csv', delimiter=',',encoding='utf-8-sig')
4
5 numbers
```

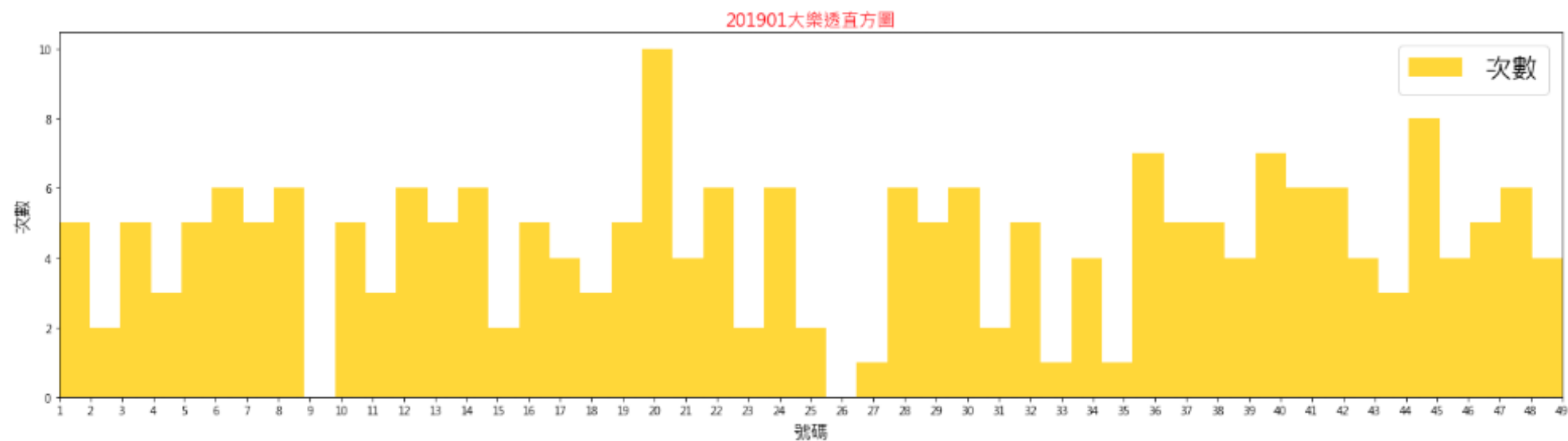
使用numpy 套件  
讀取csv檔案  
並得到ndarray陣列

1. 分離出一般號碼
2. 將二維資料重新朔型為一維
3. 使用astype轉型成int

```
1 x_number=numbers[:, :-1]
2 x_number=x_number.reshape(x_number.shape[0]*x_number.shape[1]).astype(int)
3
4 # x_number=[int(i) for i in x_number[0]]
5
6 print(x_number)
```

```
[18 23 24 25 39 42  6  7  8 20 47 48  4 19 21 28 39 43 13 20 22 31 34 42
12 20 30 38 40 41  4  5 10 30 39 45  2 20 36 37 42 46  3 10 14 15 20 45
14 20 24 30 44 48 10 17 20 21 32 41  2  7 22 35 46 49 12 24 28 36 37 40
 6  8 24 32 34 36 13 18 30 34 45 49 17 23 28 29 32 39  7 21 29 33 40 41
 8 13 14 32 36 42  6  8 18 19 20 30  5  8 14 19 22 47 11 16 19 29 30 45
 1  3 40 44 46 48  1  3  4 24 32 43  6 20 22 43 45 48  6 17 22 34 38 42
 1  8 14 27 40 45 12 13 16 21 42 43  3 11 15 25 31 48  5  7 22 37 40 45
 5 10 12 20 28 36  7 16 17 36 47 48  1 12 16 37 41 47  6 13 14 19 24 38
 1  3 12 16 29 49  5 28 38 40 41 47 11 28 29 38 41 49 10 36 37 44 45 46]
```

```
1  nums=49
2
3  plt.figure(figsize=(24,6))
4
5  n=plt.hist(x_number,nums,color='gold',label='次數')
6  plt.xlabel('號碼',fontsize=16)
7  plt.ylabel('次數',fontsize=16)
8  plt.xlim(1,49)
9  plt.xticks(range(1,nums+1))
10 plt.legend(fontsize=24)
11 plt.title('201901大樂透直方圖',color='red',fontsize=16)
12 plt.show()
13
```





```
1 x=n[0].astype(int)
2
3 for i in range(1,nums+1):
4     print('號碼{} 出現{}次'.format(i,x[i-1]))
5
```

號碼1 出現5次  
號碼2 出現2次  
號碼3 出現5次  
號碼4 出現3次  
號碼5 出現5次  
號碼6 出現6次  
號碼7 出現5次  
號碼8 出現6次  
號碼9 出現0次  
號碼10 出現5次  
號碼11 出現3次  
號碼12 出現6次  
號碼13 出現5次  
號碼14 出現6次  
號碼15 出現2次  
號碼16 出現5次  
號碼17 出現4次  
號碼18 出現3次  
號碼19 出現5次  
號碼20 出現10次  
號碼21 出現4次  
號碼22 出現6次  
號碼23 出現2次  
號碼24 出現6次  
號碼25 出現2次  
號碼26 出現0次  
號碼27 出現1次  
號碼28 出現6次  
號碼29 出現5次  
號碼30 出現6次

## ■ NBA 資料分析



```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3
4 df1=pd.read_csv('GSW_players_stats_2017_18.csv')
5 df1
```

	No.	Player	Pos	Ht	Wt	Birth Date	Nationality	Exp	College	PTS/G	AST	TRB	3P
0	2	Jordan Bell	C	6-9	224	January 7, 1995	us	R	University of Oregon	4.6	1.8	3.6	0.0
1	25	Chris Boucher	PF	6-10	182	January 11, 1993	lc	R	University of Oregon	0.0	0.0	1.0	0.0
2	18	Omri Casspi	SF	6-9	225	June 22, 1988	il	8	NaN	5.7	1.0	3.8	0.2
3	4	Quinn Cook	PG	6-2	184	March 23, 1993	us	1	Duke University	9.5	2.7	2.5	1.4
4	30	Stephen Curry	PG	6-3	190	March 14, 1988	us	8	Davidson College	26.4	6.1	5.1	4.2
5	35	Kevin Durant	PF	6-9	240	September 29, 1988	us	10	University of Texas at Austin	26.4	5.4	6.8	2.5
6	23	Draymond Green	PF	6-7	230	March 4, 1990	us	5	Michigan State University	11.0	7.3	7.6	1.1
7	9	Andre Iguodala	SF	6-6	215	January 28, 1984	us	13	University of Arizona	6.0	3.3	3.8	0.5
8	15	Damian Jones	C	7-0	245	June 30, 1995	us	1	Vanderbilt University	1.7	0.1	0.9	0.0
9	34	Shaun Livingston	PG	6-7	192	September 11, 1985	us	12	NaN	5.5	2.0	1.8	0.0
10	5	Kevon Looney	C	6-9	220	February 6, 1996	us	2	University of California, Los Angeles	4.0	0.6	3.3	0.0
11	0	Patrick McCaw	SG	6-7	185	October 25, 1995	us	1	University of Nevada, Las Vegas	4.0	1.4	1.4	0.3
12	1	JaVale McGee	C	7-0	270	January 19, 1988	us	9	University of Nevada, Reno	4.8	0.5	2.6	0.0
13	27	Zaza Pachulia	C	6-11	270	February 10, 1984	ge	14	NaN	5.4	1.6	4.7	0.0
14	11	Klay Thompson	SG	6-7	215	February 8, 1990	us	6	Washington State University	20.0	2.5	3.8	3.1
15	3	David West	C	6-9	250	August 29, 1980	us	14	Xavier University	6.8	1.9	3.3	0.0
16	6	Nick Young	SG	6-7	210	June 1, 1985	us	10	University of Southern California	7.3	0.5	1.6	1.5

## ■ 群聚分析

- 使用Pos欄位資料(找出該隊伍位置的相關人數)

```
1 df_pos=df1.groupby('Pos')
2 df_pos.mean()
```

	No.	Wt	PTS/G	AST	TRB	3P
Pos						
C	8.833333	246.500000	4.550000	1.083333	3.066667	0.000000
PF	27.666667	217.333333	12.466667	4.233333	5.133333	1.200000
PG	22.666667	188.666667	13.800000	3.600000	3.133333	1.866667
SF	13.500000	220.000000	5.850000	2.150000	3.800000	0.350000
SG	5.666667	203.333333	10.433333	1.466667	2.266667	1.633333

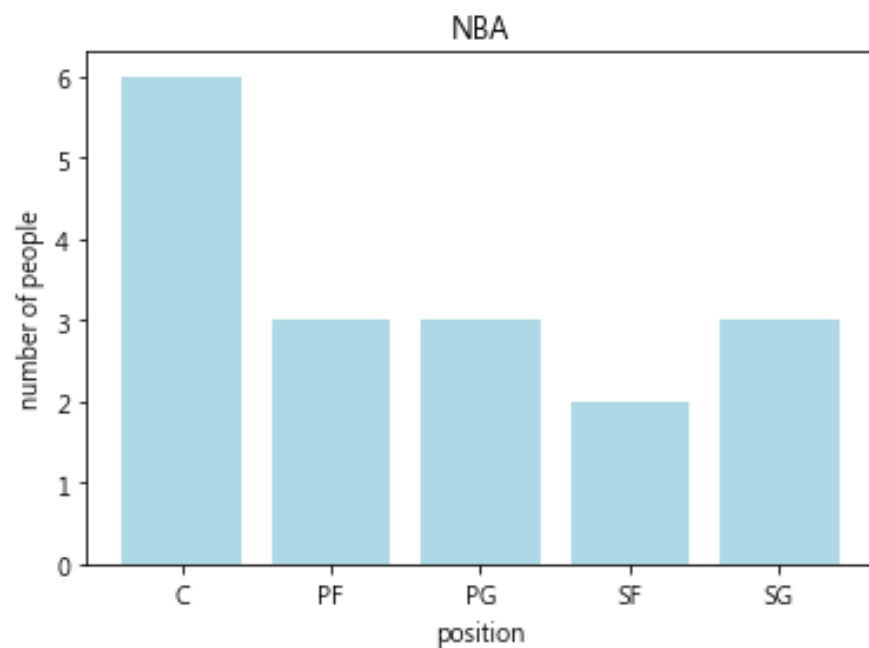
再使用Player欄位  
進行計數

```
1 df_pos_count=df_pos['Player'].count()
2
3 df_pos_count
```

Pos	
C	6
PF	3
PG	3
SF	2
SG	3
Name: Player, dtype: int64	

- 進行繪製
  - 使用柱狀圖

```
1 plt.bar(df_pos_count.index,df_pos_count,color='lightblue')
2 plt.ylabel('number of people')
3 plt.xlabel('position')
4 plt.title('NBA')
5
6 plt.show()
7
```

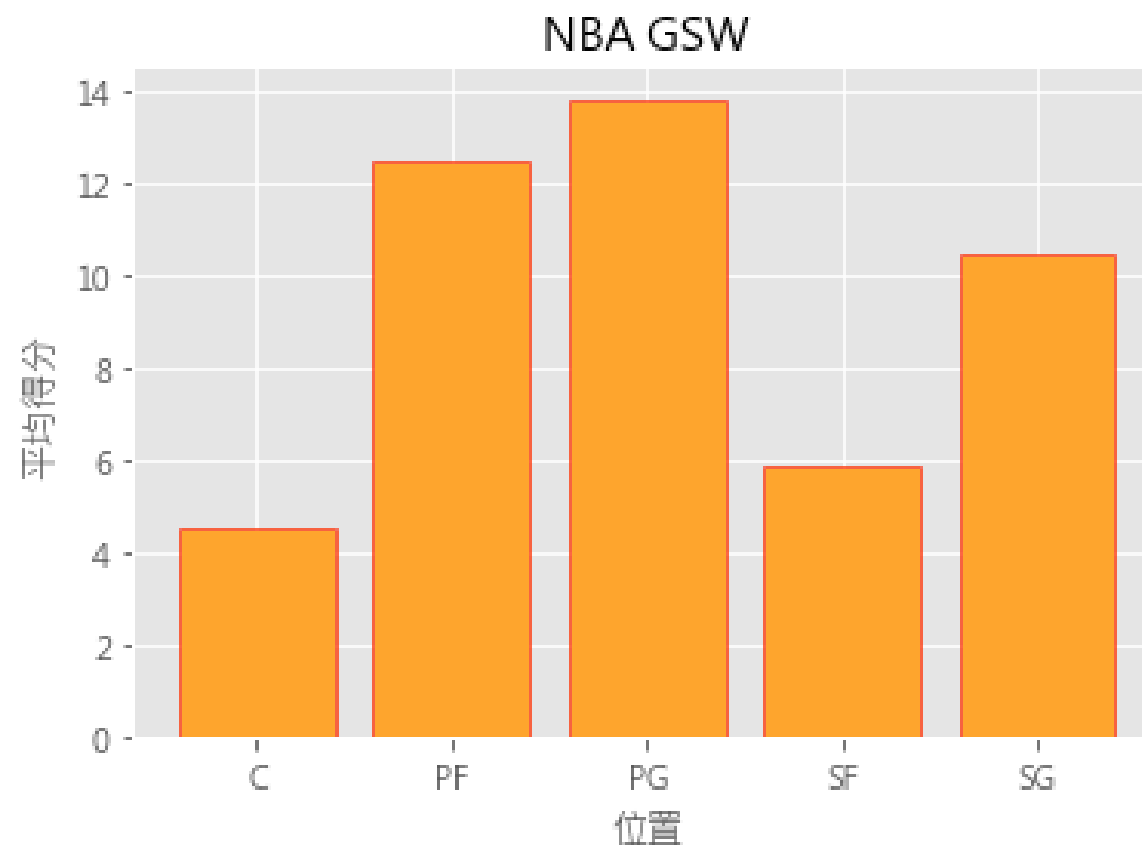


## ■ 程式練習

### ■ 輸出每個位置的平均得分

#### ■ `df1.groupby('Pos')['PTS/G']`

College	PTS/G	AST	TRB	3P
University of Oregon	4.6	1.8	3.6	0.0
University of Oregon	0.0	0.0	1.0	0.0
NaN	5.7	1.0	3.8	0.2
Duke University	9.5	2.7	2.5	1.4
Davidson College	26.4	6.1	5.1	4.2
University of Texas at Austin	26.4	5.4	6.8	2.5
Michigan State University	11.0	7.3	7.6	1.1
University of Arizona	6.0	3.3	3.8	0.5
Vanderbilt University	1.7	0.1	0.9	0.0
NaN	5.5	2.0	1.8	0.0
University of California, Los Angeles	4.0	0.6	3.3	0.0
University of Nevada, Las Vegas	4.0	1.4	1.4	0.3
University of Nevada, Reno	4.8	0.5	2.6	0.0
NaN	5.4	1.6	4.7	0.0
Washington State University	20.0	2.5	3.8	3.1
Xavier University	6.8	1.9	3.3	0.0
University of Southern California	7.3	0.5	1.6	1.5



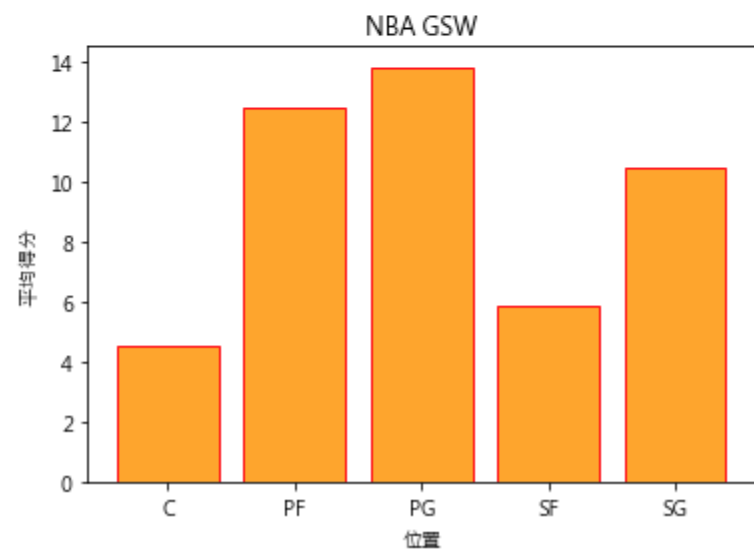
## ■ 程式碼



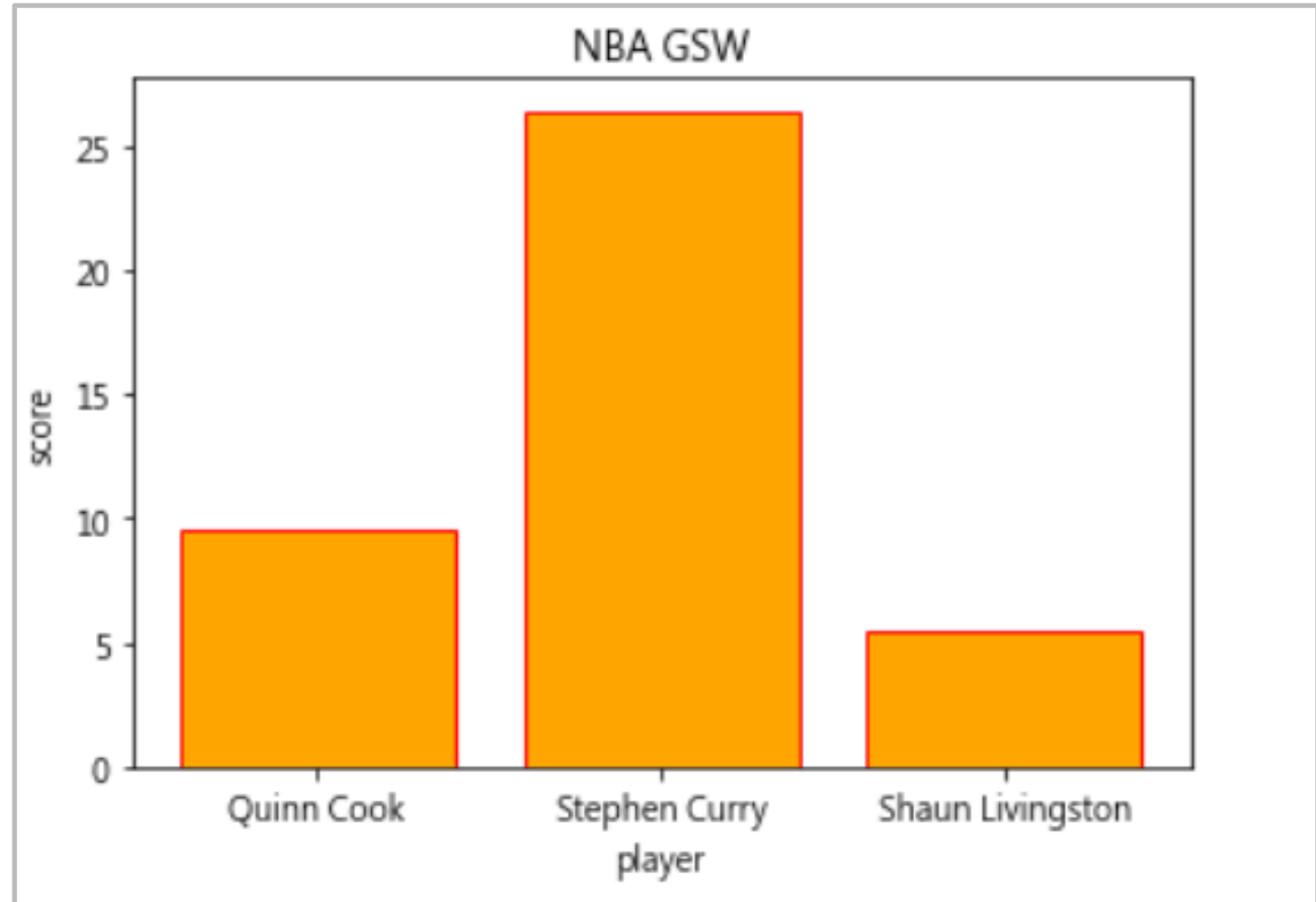
```
1 df_points=df1.groupby('Pos')['PTS/G'].mean()  
2  
3 df_points
```

```
Pos  
C      4.550000  
PF     12.466667  
PG     13.800000  
SF      5.850000  
SG     10.433333  
Name: PTS/G, dtype: float64
```

```
1 plt.bar(df_points.index,df_points,color='orange',edgecolor='red')  
2 plt.ylabel('平均得分')  
3 plt.xlabel('位置')  
4 plt.title('NBA GSW')  
5 plt.show()
```



- 程式練習
  - 輸出PG位置人員得分分布



## ■ 程式練習

- 取得各位置得分最高數據

