PlotBasic

2018 年 6 月 22 日

1 視覺化工具

1.1 使用函式庫

1. Pandas: 我們的特徵資料表達的方式，Pandas 也跟 Matplotlib 有一個很好的連結，可以快速地把 DataFrame 轉成圖
2. Matplotlib: 所有視覺化工具的鼻祖，很多進階的視覺化函式庫都是基於 matplotlib 建立的
3. Seaborn: 進階的視覺化函式庫之一，有一些好用的函式讓你快速地建立一個複雜圖

1.2 介紹

這個章節我們不講一些比較數學的作圖，我們希望讓你了解在做機器學習或者深度學習的時候，我們常用的一些前置的作圖，讓我們了解我們資料特徵之間的關係或者是特徵和標籤之間的關係。

1.3 安裝方法

請用命令列或者 pycharm 安裝好 matplotlib 和 seaborn 函式庫

1.4 官方文件和例子

1. seaborn: https://seaborn.pydata.org/
2. matplotlib: https://matplotlib.org/

1.5 目標

我們使用 Kaggle 的鐵達尼號資料集來教你繪圖的基本操作

1.6 資料集位置

https://www.kaggle.com/c/titanic/data

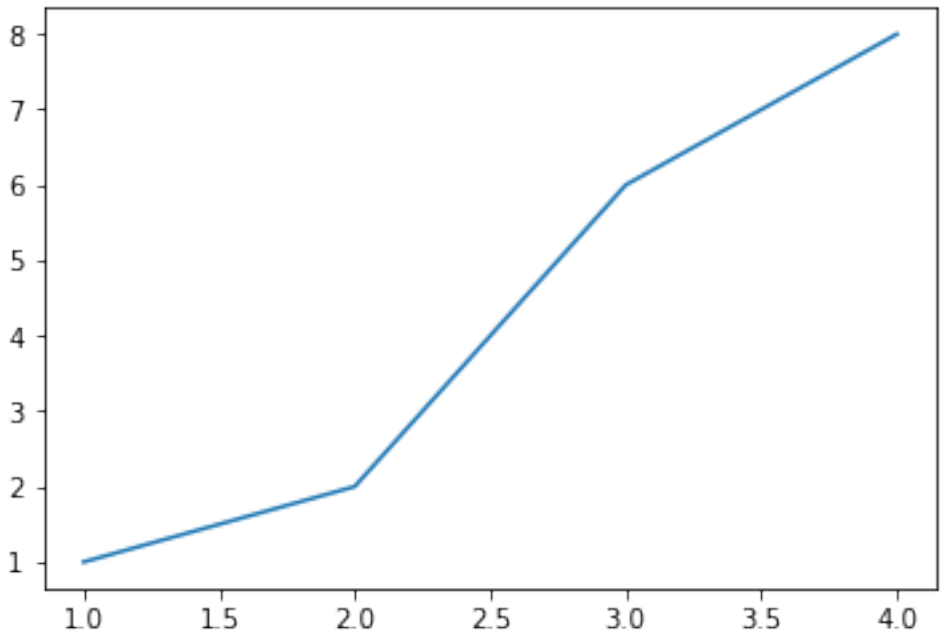
1. 需要登入才能下載
2. 只取裡面的 train.csv 來做繪圖

1.7 Matplotlib 最基本概念

1. 最簡單的繪圖就是給好 x 軸的 list 和給好 y 軸的 list, 然後就可以繪圖了
2. 流程 plot() -> show()
3. plot 的時候你必須把所以想在一張圖上的東西畫好
4. 所有的東西畫好在呼叫 show()
5. (純粹習慣) 大家習慣在 import 以後改叫 plt

```
In [1]: import matplotlib.pyplot as plt
```

```
In [2]: x_list = [1, 2, 3, 4]
        y_list = [1, 2, 6, 8]
        plt.plot(x_list, y_list)
        plt.show()
```



1.8 選擇線條類型和顏色

繪圖的第三個參數可以以一組類型 + 顏色來客製化

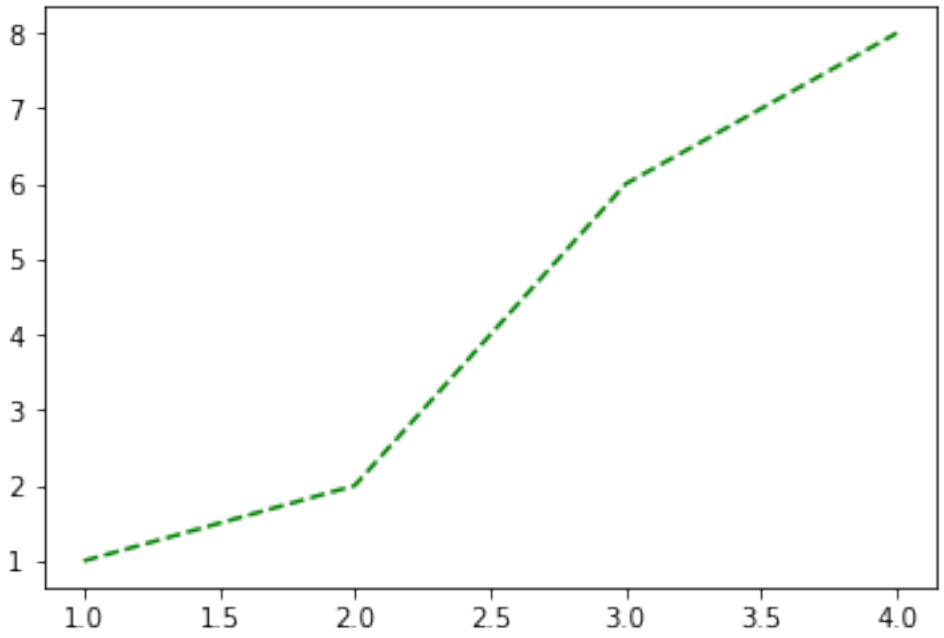
1. 線條

character	description
'-'	solid line style
'--'	dashed line style
'-.'	dash-dot line style
':'	dotted line style
'.'	point marker
','	pixel marker
'o'	circle marker
'v'	triangle_down marker
'^'	triangle_up marker
'<'	triangle_left marker
'>'	triangle_right marker
'1'	tri_down marker
'2'	tri_up marker
'3'	tri_left marker
'4'	tri_right marker
's'	square marker
'p'	pentagon marker
'*'	star marker
'h'	hexagon1 marker
'H'	hexagon2 marker
'+'	plus marker
'x'	x marker
'D'	diamond marker
'd'	thin_diamond marker
'|'	vline marker
'_'	hline marker

2. 顏色

character	color
‘b’	blue
‘g’	green
‘r’	red
‘c’	cyan
‘m’	magenta
‘y’	yellow
‘k’	black
‘w’	white

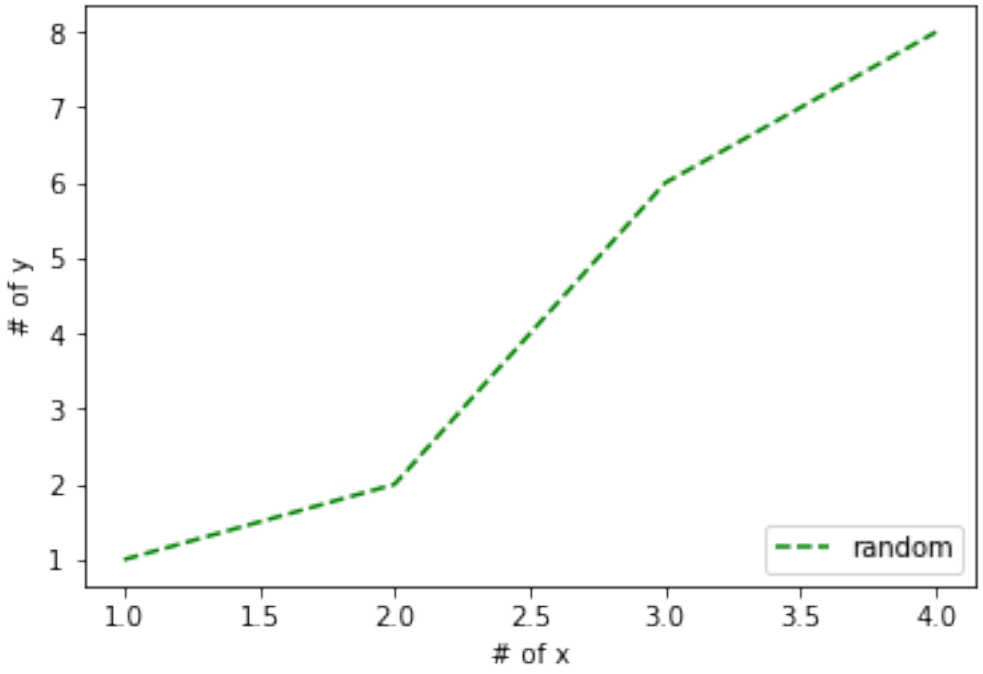
```
In [3]: x_list = [1, 2, 3, 4]
        y_list = [1, 2, 6, 8]
        # dash-line + green color
        plt.plot(x_list, y_list, "--g")
        plt.show()
```



1.9 加上標籤

1. 在 plot 的時候加上參數 label
2. 記得要用 legend 放在對的位置上，不然一樣看不到
3. 可以用 xlabel() 和 ylabel() 加上 x 軸標籤和 y 軸標籤

```
In [4]: x_list = [1, 2, 3, 4]
        y_list = [1, 2, 6, 8]
        # dash-line + green color
        plt.plot(x_list, y_list, "--g", label = "random")
        plt.legend(loc = "lower right")
        plt.xlabel("# of x")
        plt.ylabel("# of y")
        plt.show()
```



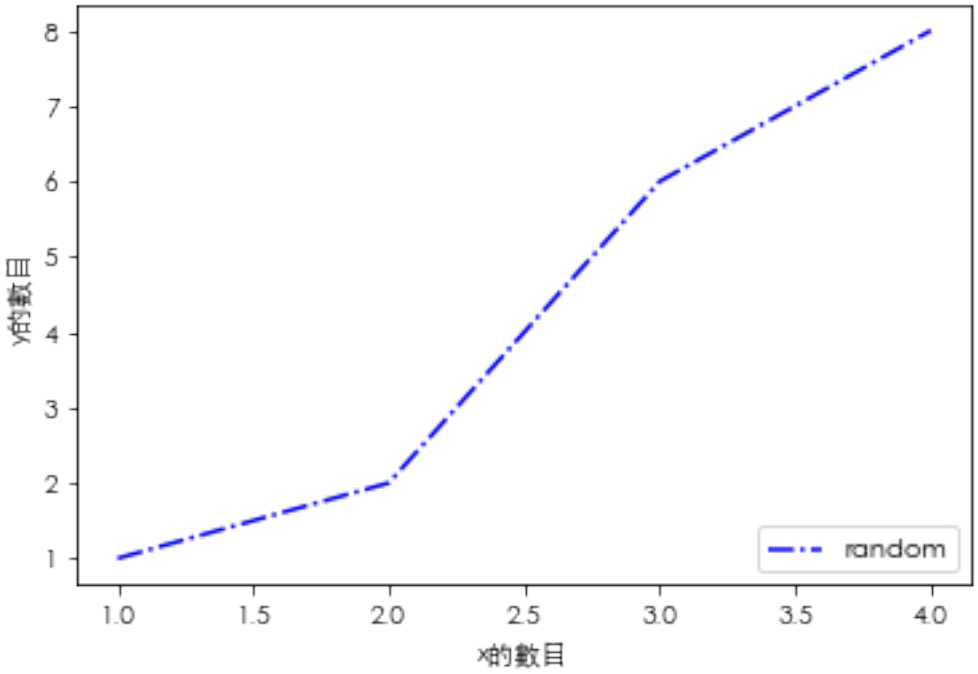
1.10 (盡量不要) 中文顯示

1. Matplotlib 預設只有外文字體，對於外文字體的顯示也比較漂亮一點
2. 如果你真的萬不得已，想要使用中文字體來顯示，要特別將預設的字體修改成中文字體

1.11 使用字體

STHeiti(蘋果黑體): Mac 內建字型，由於這裡 OS 是使用 Mac，所以我選擇了 Mac 內建的字型

```
In [5]: import matplotlib
        matplotlib.rcParams['font.sans-serif'] = 'STHeiti'
        x_list = [1, 2, 3, 4]
        y_list = [1, 2, 6, 8]
        # dash-dot + blue color
        plt.plot(x_list, y_list, "-.b", label = "random")
        plt.legend(loc = "lower right")
        plt.xlabel("x 的數目")
        plt.ylabel("y 的數目")
        plt.show()
```



1.12 鐵達尼號 + Seaborn

1. 使用 Pandas 先將鐵達尼號的資料讀取出來
2. Survied 欄位: 0 代表無法倖存的乘客 1 代表倖存的乘客
3. 以下是每個欄位代表的意思

```
In [6]: import pandas as pd
        # 為了顯示的漂亮, 我刻意的把印出來的 row 只顯示 20 個和 column 只顯示十個
        # 大家練習的時候可以去掉下面兩行
        pd.set_option('display.max_rows', 20)
        pd.set_option('display.max_columns', 10)

        df = pd.read_csv("train.csv", encoding = "utf-8")
        df

Out[6]:      PassengerId  Survived  Pclass  \
        0              1         0       3
        1              2         1       1
        2              3         1       3
        3              4         1       1
        4              5         0       3
        5              6         0       3
        6              7         0       1
        7              8         0       3
        8              9         1       3
        9             10         1       2
        ..           ...       ...     ...
        881          882         0       3
        882          883         0       3
        883          884         0       2
        884          885         0       3
        885          886         0       3
        886          887         0       2
        887          888         1       1
        888          889         0       3
        889          890         1       1
        890          891         0       3

                                                      Name     Sex     ...     \
        0                                 Braund, Mr. Owen Harris    male     ...
        1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female     ...
        2                                 Heikkinen, Miss. Laina  female     ...
        3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female     ...
```

4				Allen, Mr. William Henry	male	...	
5				Moran, Mr. James	male	...	
6				McCarthy, Mr. Timothy J	male	...	
7				Palsson, Master. Gosta Leonard	male	...	
8				Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	...	
9				Nasser, Mrs. Nicholas (Adele Achem)	female	...	
..				...	...	...	
881				Markun, Mr. Johann	male	...	
882				Dahlberg, Miss. Gerda Ulrika	female	...	
883				Banfield, Mr. Frederick James	male	...	
884				Sutehall, Mr. Henry Jr	male	...	
885				Rice, Mrs. William (Margaret Norton)	female	...	
886				Montvila, Rev. Juozas	male	...	
887				Graham, Miss. Margaret Edith	female	...	
888				Johnston, Miss. Catherine Helen "Carrie"	female	...	
889				Behr, Mr. Karl Howell	male	...	
890				Dooley, Mr. Patrick	male	...	

	Parch	Ticket	Fare	Cabin	Embarked
0	0	A/5 21171	7.2500	NaN	S
1	0	PC 17599	71.2833	C85	C
2	0	STON/O2. 3101282	7.9250	NaN	S
3	0	113803	53.1000	C123	S
4	0	373450	8.0500	NaN	S
5	0	330877	8.4583	NaN	Q
6	0	17463	51.8625	E46	S
7	1	349909	21.0750	NaN	S
8	2	347742	11.1333	NaN	S
9	0	237736	30.0708	NaN	C
..	...	...	...	...	...
881	0	349257	7.8958	NaN	S
882	0	7552	10.5167	NaN	S
883	0	C.A./SOTON 34068	10.5000	NaN	S
884	0	SOTON/OQ 392076	7.0500	NaN	S
885	5	382652	29.1250	NaN	Q
886	0	211536	13.0000	NaN	S
887	0	112053	30.0000	B42	S

```
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

1.13 Import 函式庫以及利用 Notebook

1. 由於 Seaborn 是基於 Matplotlib 的函式庫，所以正常寫法下，你一樣得在最後一行加上 plt.show() 印出圖形
2. 在 Jupyter Notebook 裡使用的時候其實可以透過他提供給我們的簡便工具來簡化我們使用
3. (重要，且只有 Notebook 可以用) 加上%matplotlib inline 這行的話，你就可以在每次做圖的時候少打 plt.show()

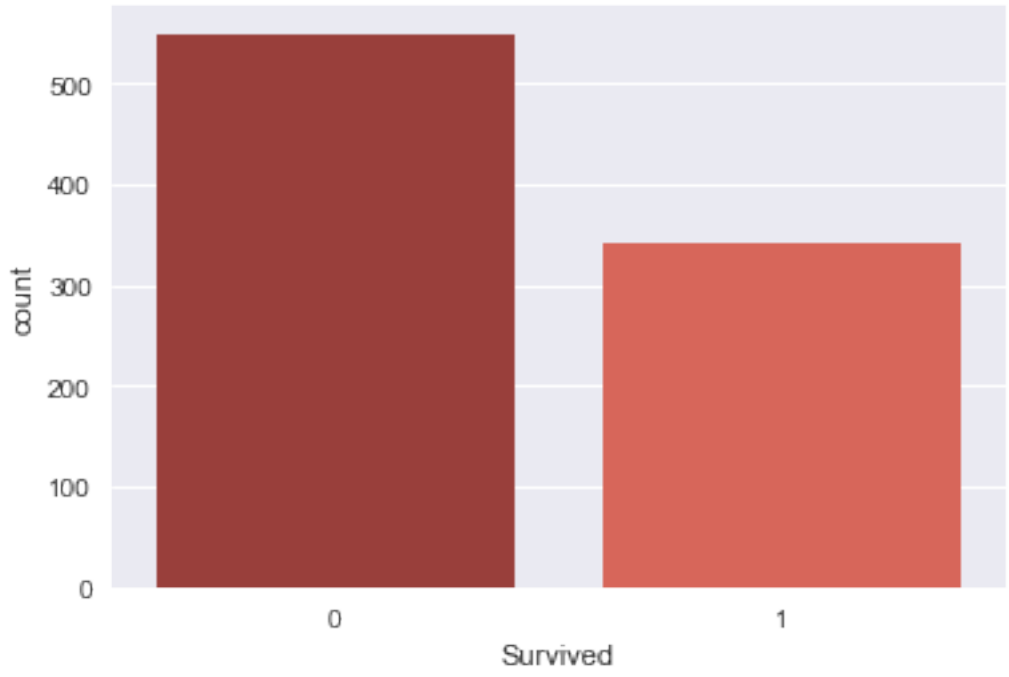
```
In [7]: # 大家這裡習慣給他改名成 sns
        import seaborn as sns
        # 這行只有在 Jupyter Notebook 可以使用
        %matplotlib inline
        # sns 每次會跳出 remove_na 的 warning
        # 我為了頁面的美化 不讓 warning 印出，但讀者不一定要過濾掉
        import warnings
        warnings.filterwarnings('ignore')
```

1.14 數量圖

1. 數量圖是在類別做圖的時候一個很好用的圖示工具，統計各個類別分別有多少數量
2. 使用 seaborn.countplot，我們通常選擇只設置其中一軸，另一軸就是數量
3. 你可以藉由 palette 這個參數來選擇一下你喜歡的整體色系
4. palette: https://seaborn.pydata.org/generated/seaborn.color_palette.html#seaborn.color_palette

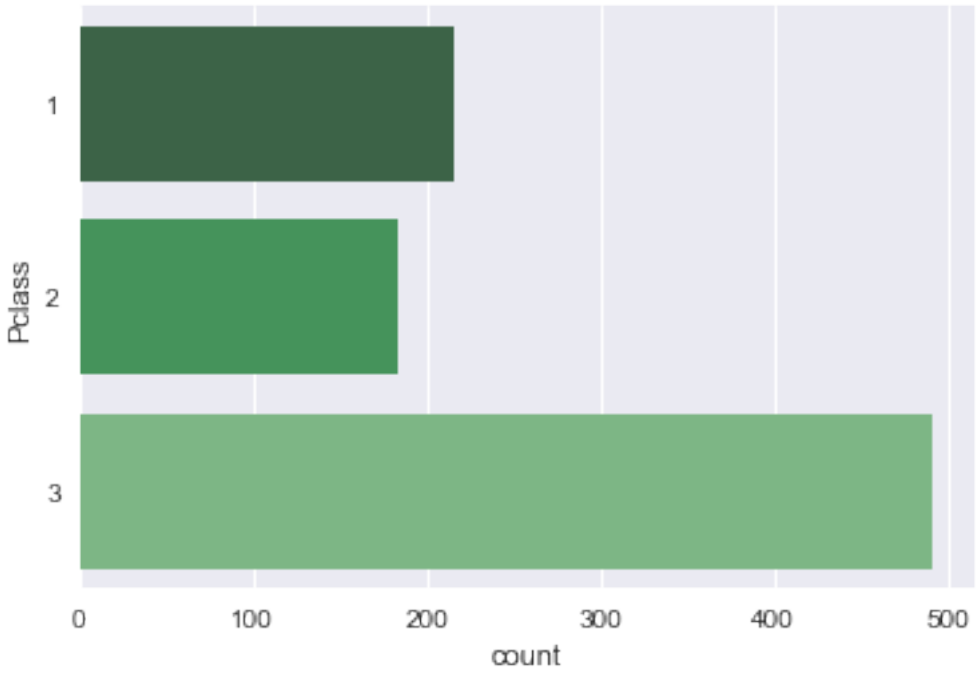
```
In [8]: # 針對一個類別做數量圖，由於我們 inline 了 matplotlib，所以不需要 plt.show()
        # 設置 x 軸的往上長的長條圖
        sns.countplot(x = df["Survived"], palette = "Reds_d")
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x114c73128>
```



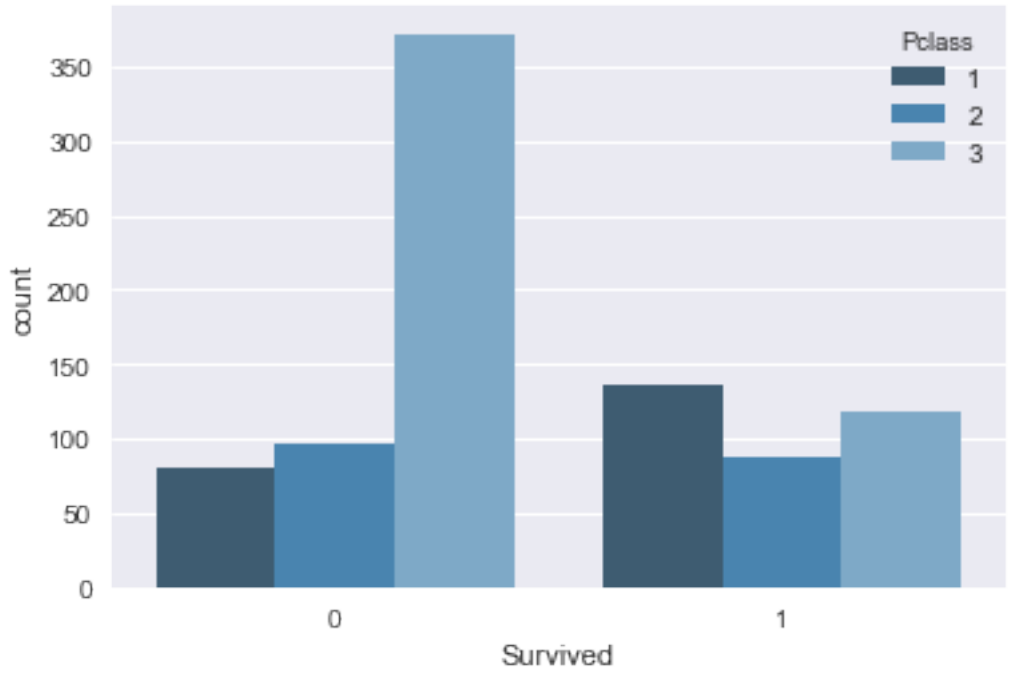

```
In [9]: # 針對一個類別做數量圖, 由於我們 inline 了 matplotlib, 所以不需要 plt.show()
        # 設置 y 軸的往右長的長條圖
        # 畫出各艙等的人數
        sns.countplot(y = df["Pclass"], palette = "Greens_d")

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x114c9d208>
```




```
In [10]: # 結合上面兩個，把一個區域的長條加入第二個特徵，統計第二個特徵＋第一個特徵的數目
         # hue 裡面放的就是第二個特徵
         # 你找到了其中一個相關性：第三艙等的存活率稍微低了一點
         sns.countplot(x = df["Survived"], hue = df["Pclass"], palette = "Blues_d")

Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x114d068d0>
```

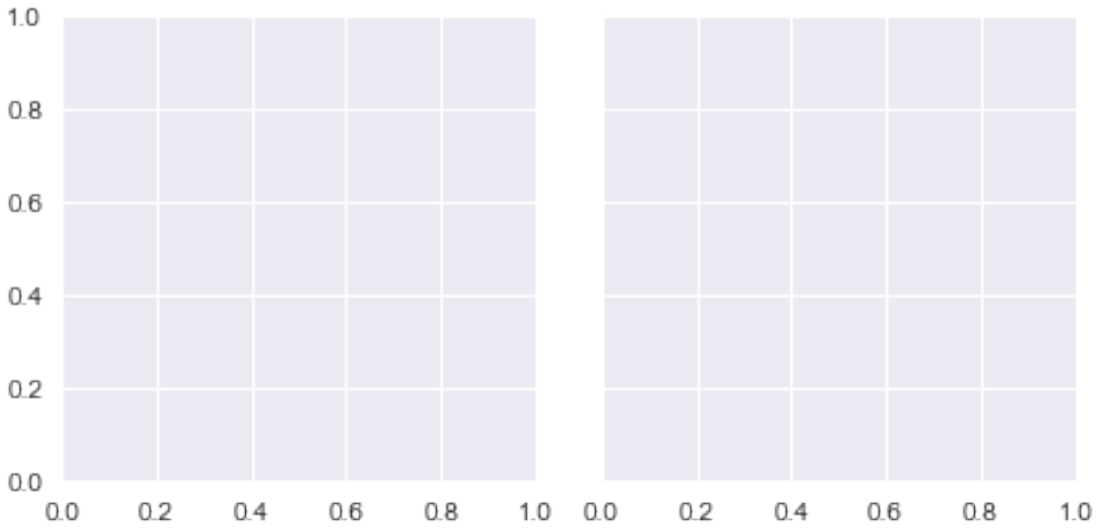


1.15 FacetGrid 網格圖 + 分布圖

1.15.1 FacetGrid 網格圖

1. FacetGrid 網格圖是你可以固定變數的一種圖形，你設定的欄位會把所有可能的值拿出來，變出這麼多圖出來
2. 可以設定 row 和 col，你的圖的數量就等於 row 可能數 * col 可能數
3. 使用 FacetGrid 來創造，記得最後要給你的每個圖一個畫圖方式 (ex. 分布圖)
4. FacetGrid 是一個跟 pandas 非常友善的設計，你可以將 DataFrame 直接丟給第一個參數
5. 使用 FacetGrid 的主因之一是我們要做出分布圖 (x 軸是連續的)，但連續的東西我們沒辦法像上面的 countplot 分成三條來看，會看不出連續的趨勢

```
In [11]: # 對每一個網格裡的圖做下面的初始化
         # 由於剛剛已經設定完 data, 這裡直接給他欄位的名稱當 x 軸即可
         fg = sns.FacetGrid(df, col = "Survived")
```

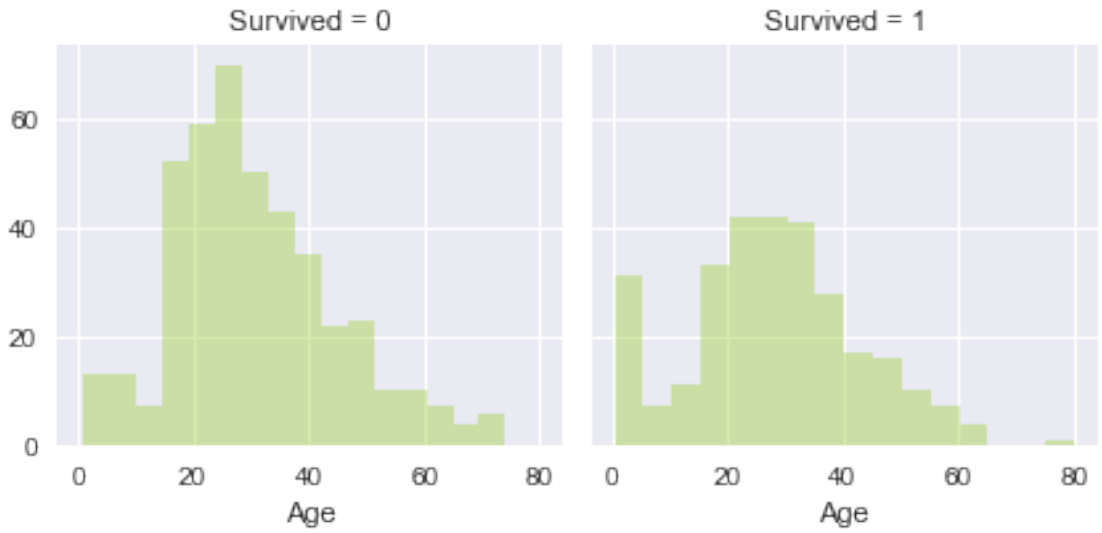


1.15.2 DistPlot 分布圖

1. 通常我們使用了網格圖以後，我們需要對每個圖初始化，我們常用的就是分佈圖
2. 分布圖要傳入一個一個維度的群集，ex. 一個 Series
3. 他會把你傳入的一維群集轉換成 x 軸的區間
4. 跟 Countplot 一個很大的不同是我們的 x 軸是連續的，不是分類型的
5. 你可以給你的分布圖一個顏色: https://matplotlib.org/users/colors.html

```
In [14]: fg = sns.FacetGrid(df, col = "Survived")
         fg.map(sns.distplot, 'Age', color = "yellowgreen", kde = False)
```

```
Out[14]: <seaborn.axisgrid.FacetGrid at 0x1152dea90>
```


Survived = 0
Survived = 1
60
40
20
0
0
20
40
60
80
Age
Age