知識點回顧

延伸閱讀

D22 結合 Pandas 與 Matploglib 進行進階資料視覺化練習





- 資料集輸入與處理

import pandas as pd

先行導入相關的套件

import numpy as np import seaborn as sns

import matplotlib as mpl

瞭解有關資料集屬性

import matplotlib.pyplot as plt from mpl_toolkits.mplot3d import Axes3D

量、列名稱、它們的數據類型和空值數。 資料集的處理

• 有時候無法從資料集明確的看出資料的屬性與因子的相互關係,要針對資料做處理 瞭解數據集

• 我們可以使用 info()或是 descript() 方法瞭解有關資料集屬性的更多資訊。特別是行和列的數

要瞭解數據集的統計摘要,即記錄數、平均值、標準差、最小值和最大值,我們使用 describe()

可以使用 info() 方法瞭解有關資料集屬性的更多資訊。

特別是行和列的數量、列名稱、它們的數據類型和空值數。 df.info()

處理缺失值 • df = pd.get_dummies

<class 'pandas.core.frame.DataFrame'>

Int64Index: 6497 entries, 0 to 4897 Data columns (total 13 columns): # Column

• df. describe()

0 fixed_acidity 6497 non-null float64 1 volatile_acidity 6497 non-null float64 2 citric_acid 6497 non-null float64 3 residual_sugar 6497 non-null float64 4 chlorides 6497 non-null float64 5 free_sulfur_dioxide 6497 non-null float64 6 total_sulfur_dioxide 6497 non-null float64 ity 6497 non-null float64 6497 non-null float64 density рΗ 9 sulphates 6497 non-null float64 10 alcohol 6497 non-null float64 11 quality 6497 non-null int64 12 color 6497 non-null object dtypes: float64(11), int64(1), object(1) memory usage: 710.6+ KB

• 可視化是數據分析的一個固有部分,因為它用於以簡單而有效的方式傳達我們的發現。

• 繪製強大趨勢、圖表和各種其他統計圖表的技術,可幫助人們輕鬆瞭解有關數據的資訊

Non-Null Count Dtype

熱力圖 # 聯合圖

Hist 直方圖

什麼是可視化?

直方圖

citric acid

1.0

200

- 1.0

- 0.8

- 0.6

- 0.4

- 0.2

- 0.0

- -0.2

- -0.4

free sulfur dioxide

2000

2000

15

0.0

2000

fixed acidity

直方圖使用 PANDAS 來可視化所有數值數據。 在垂直軸上計數,在水平軸上使用值範圍。

hist 函數通過將所有屬性繪製在一起使操作變得簡單。

• 這種對數據的瞭解反過來又有助於預測和模型構建。

• 針對紅白酒的資料集,我們會依序使用

df_all.hist(bins=10, color='lightblue',edgecolor='blue',linewidth=1.0,

xlabelsize=8, ylabelsize=8, grid=False)

density

1.00

f, ax = plt.subplots(figsize=(10, 6))

citric_acid - 0.32

residual_sugar - -0.11

total_sulfur_dioxide - -0.33 -0.41

density -

sulphates -

free_sulfur_dioxide - -0.28 -0.35 0.13

chlorides

f.subplots_adjust(top=0.93)

bottom, top = ax.get_ylim()

1.02

1.04

2000

plt.tight_layout(rect=(0, 0, 1.2, 1.2))

chlorides alcohol 1000 2000

quality residual sugar 2000 1000 2500 sulphates total sulfur dioxide volatile acidity 2000 -2000 1000 1.5 1.0 400 1.0 熱力圖 熱力圖是數據的二 - D可視化,其中兩個要素之間的關係量級由色調表示。 熱圖中的梯度根據屬性之間的相關性強度而變化。 在下面的示例中,高度相關的屬性的陰影比其餘屬性暗。 # Plotting heatmap

ax.set_ylim(bottom + 0.5, top - 0.5) title= f.suptitle('Correlation Heatmap for wine attributes', fontsize=12)

Correlation Heatmap for wine attributes fixed_acidity - 1 -0.41 volatile_acidity - 0.22

0.039 -0.13

-0.2 0.14

0.2

0.23 0.056 -0.19

夏 0.75 ·

0.50

1.75 -

1.50

1.25

0.27 0.096

0.14 0.039

-0.28

0.26 -0.33 -0.27 0.045 -0.15 -0.24 0.012

0.55 0.36 0.026 0.032

-0.19 -0.28

0.13

-0.2

-0.28

0.2 | 0.096 | -0.33 | 0.056 | -0.01 | 0.086

0.72 0.026 0.15 0.19 0.18 0.055

0.26 0.19

0.032 -0.24 -0.28 -0.27 -0.041

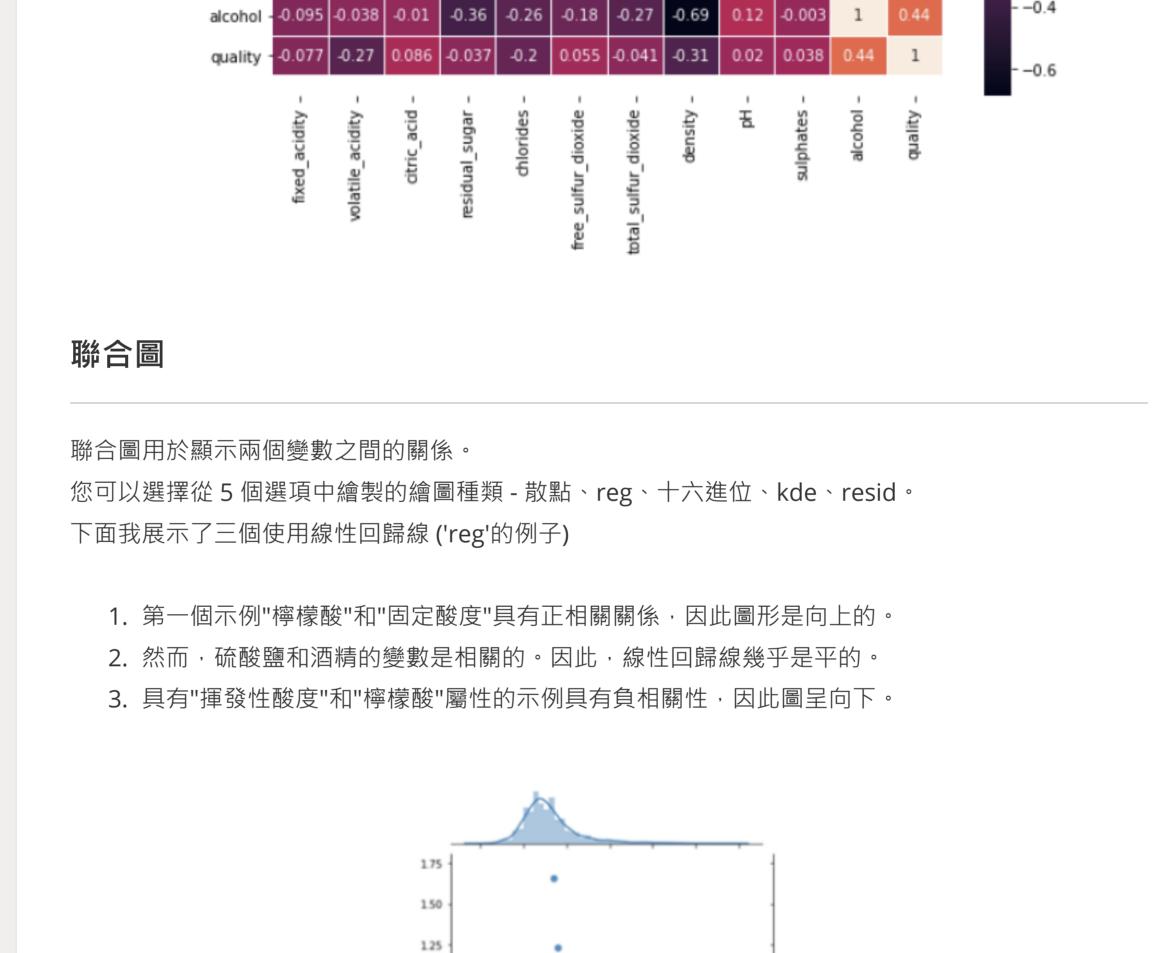
0.012 0.26 -0.69 -0.31

-0.003 0.038

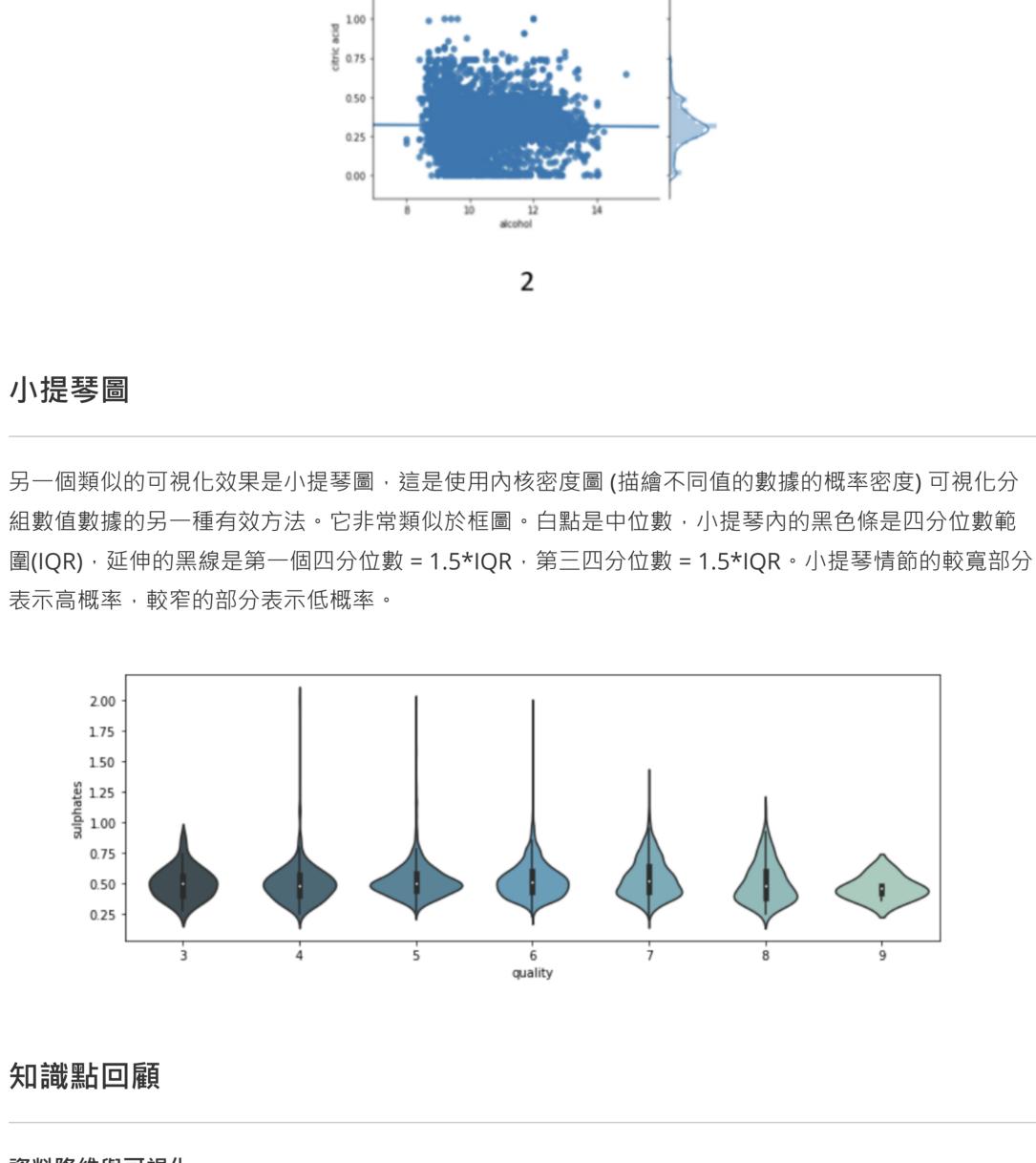
0.19

-0.27 | -0.19 | -0.36 | -0.037

b = sns.heatmap(df_all.corr(), annot=True, linewidths=.05, ax=ax)



10 fixed acidity



為了讓人們容易理解高維資料的分佈情況及降低後續特徵提取演算量,最常用的方式就是將資料「降

維(Dimensionality Reduction)」到二維或三維空間再進行觀察,亦可看做是將資料從高維度重新投 影(Projection)至低維度空間 作用:

目的:

小提琴圖

2.00

1.75

1.50

第 1.25

읔 1.00

0.75

0.50

0.25

知識點回顧

資料降維與可視化

網站:<u>歐尼克斯實境互動工作室</u>

000000000000 11111////////

MNIST Dataset

555555555555 88888888888888888 99999999999

易於觀察資料集的內容,尤其在經過降維之後有沒有更好,以手寫辨識資料庫為例

【AI HUB專欄】如何應用高維資料可視化一眼看穿你的資料集

PYTHON 數據可視化 Pandas - Python 讀取 csv 檔、excel 檔及文字檔 txt 的工具套件。 Seaborn 的分布圖是整合了 matplotlib 的直方圖與密度圖(kdeplot)的功能,增加了地毯效果 (rug) 用來觀測分布,同時也可以使用 fit 參數去擬合分配圖形。 我們可以使用 Matplotlib, Seaborn, Pandas 處理龐大的數據集 • 條形圖 (Bar plot):條形圖也可稱為柱狀圖,通常用在數值的顯示或者比較。 • 直方圖(Hist plot):用於頻率分佈,y 軸表示頻率分佈(數值或者比率),hist 函數柱體個數預 設 bins=10,且預設圖中會有網格線。 • 散點圖能夠顯示 2 個維度上每組數據的值。可以顯示觀察數據分布情形,描述數據的相關性 • 核密度圖顯示數值變量的分佈,它非常類似於直方圖。 • 熱力圖是一個以顏色變化來顯示數據的矩陣。簡單來說,就是用依據數字的不同,使用不同的

t-SNE

Python如何快速創建強大的探索性數據分析可視化 網站:<u>kknews.cc</u> 如何創建默認配對圖以快速檢查我們的數據,以及如何自定義可視化以獲取更深入的洞察力

列:

延伸閱讀

顏色來呈現數據。

• 內容包含 散點圖,內核密度,箱型圖 • 利用配色與關鍵字,進一步分析 data Seaborn

4 Afghanistan

• 針對sns.pairplot() 有深入的分析

• 針對使用PairGrid進行自定義

country continent year life_exp pop gdp_per_cap O Afghanistan Asia 1952 28.801 8425333 779.445314 1 Afghanistan Asia 1957 30.332 9240934 820.853030 2 Afghanistan Asia 1962 31.997 10267083 853.100710 3 Afghanistan Asia 1967 34.020 11537966 836.197138

首先,我們需要知道我們有什麼數據。我們可以將社會經濟數據加載並查看

每行數據代表一個國家在一年內的觀察結果,列中包含變量(這種格式的數據 稱為整理數據)。有2個分類專欄(國家和大洲區域)和4個數字專欄。這些專 欄:life_exp是幾年出生時的預期壽命,pop是人口,gdp_per_cap是以國際美 元為單位的人均國內生產總值。

Asia 1972 36.088 13079460 739.981106 头亲号 / 不難測的描

雖然後面我們將使用分類變量進行著色,但seaborn中的默認對圖僅繪製了數 字列。創建默認配對圖非常簡單:我們加載seaborn庫並調用pairplot函數: # Seaborn visualization library

import seaborn as sns

Create the default pairplot

sns.pairplot(df)

下一步:閱讀範例與完成作業