

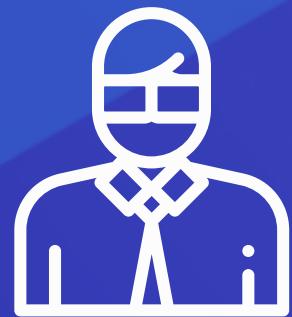


Day 9

# 資料清理數據前處理

EDA :

## 離群值(Outlier)及其處理



出題教練

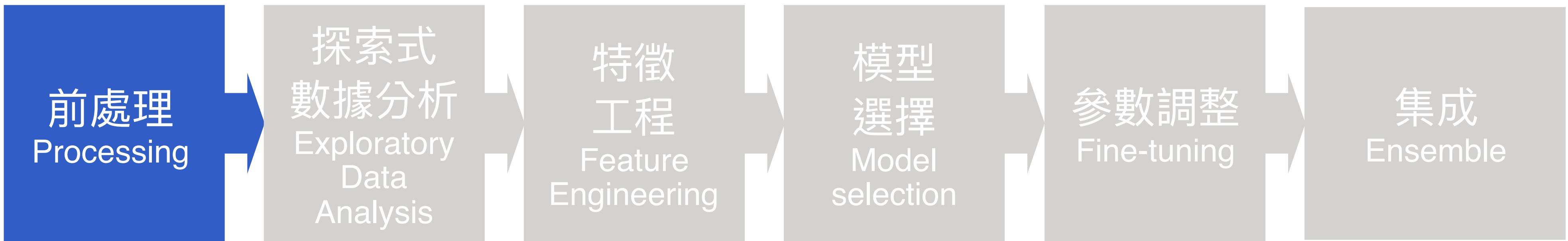
游為翔 / 杜靖愷



# 知識地圖 EDA：離群值(Outlier)及其處理

## 機器學習前處理

### 監督式學習 Supervised Learning



### 非監督式學習 Unsupervised Learning

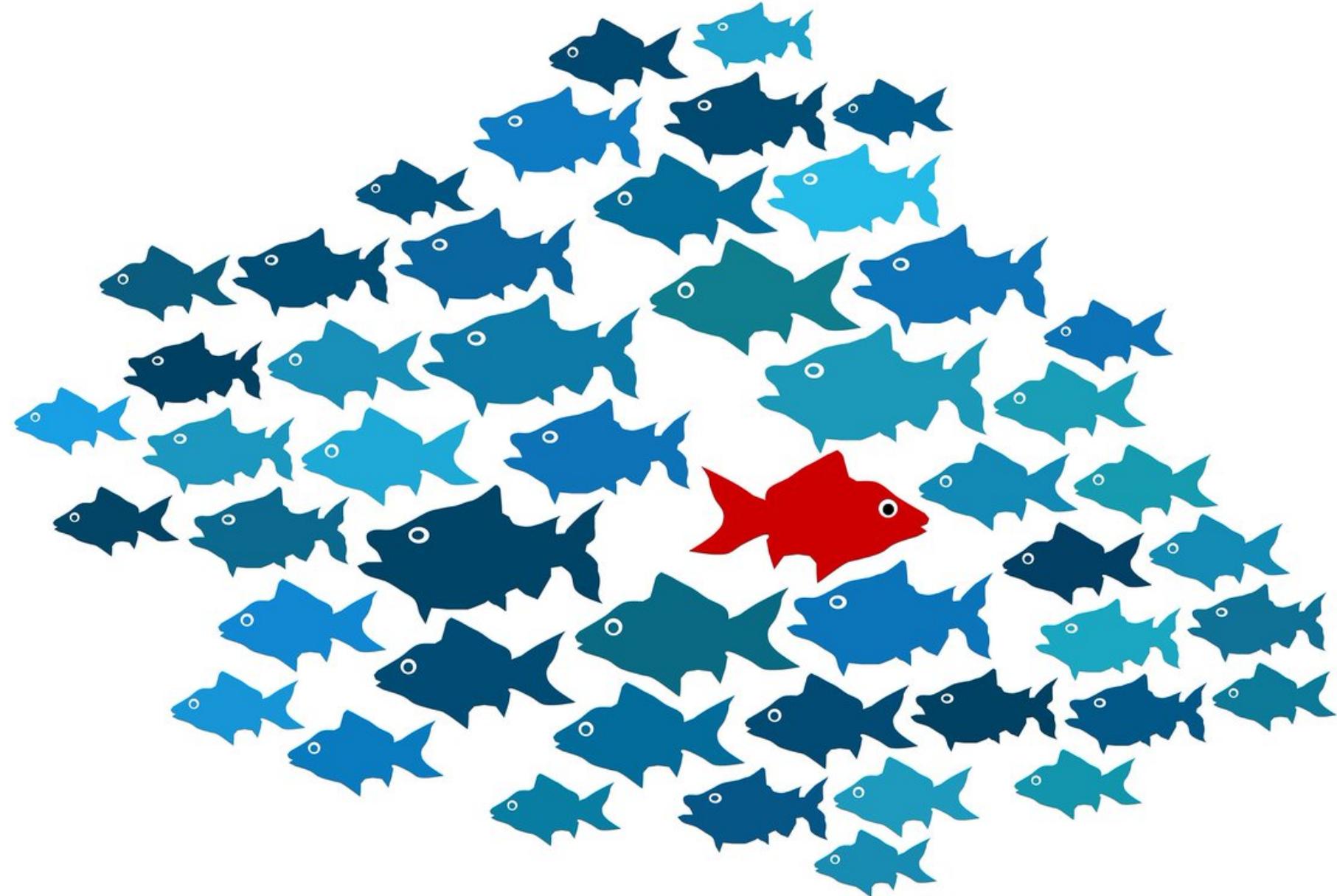


### 前處理 Processing



# 本日知識點目標

- 了解什麼是例外值 (outlier)
- 學會如何透過資料探勘方法找到例外值



圖片來源: [Sergio Santoyo](#)

## Dell電腦標價錯誤



Dell UltraSharp™ 2007FP 20" 液晶顯示器 高階平面顯示器含數位 DVI-D/類比/S-video/ Composite 輸入

原價 ..... NTD 13,200  
線上折扣 ..... NTD 7,000

線上折後價 ..... NTD 6,200

包括增值税和運費

優惠



Dell E2009W 20吋寬螢幕平面顯示器

原價 ..... NTD 7,999  
線上折扣 ..... NTD 7,000

線上折後價 ..... NTD 999

包括增值税和運費

優惠

我要自選配備

我要自選配備

1

## 異常值 (Outliers) 出現的可能原因

- 所以未知值，隨意填補 (約定俗成的代入)  
如年齡 = -1 或 999, 電話是 0900-123-456

2

- 可能的錯誤紀錄/手誤/系統性錯誤

如某本書在某筆訂單的銷售量 = 1000 本

3

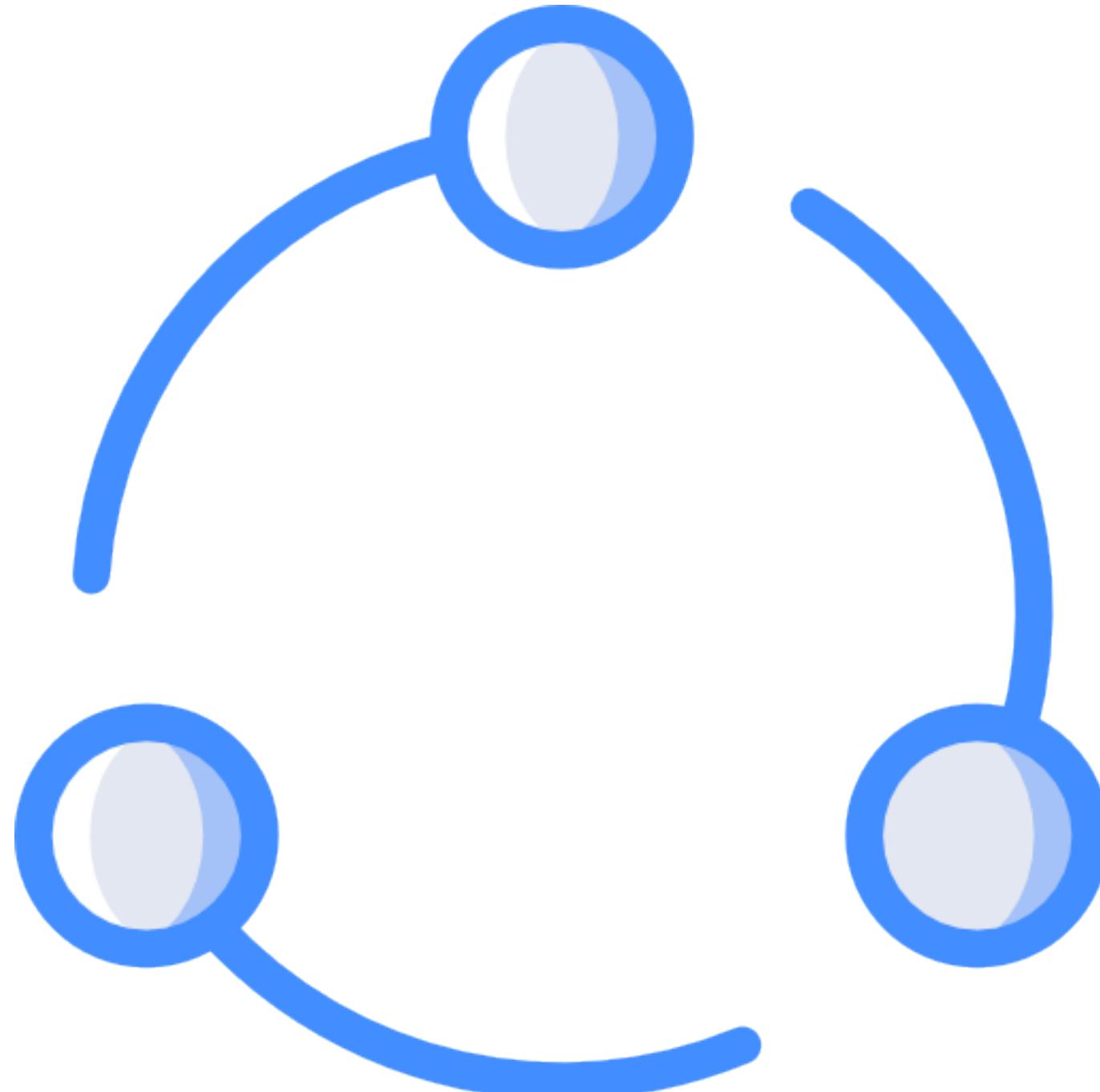
## 檢查 Outliers 的流程與方法

- 盡可能確認每一個欄位的意義 (但有些競賽資料不會提供欄位意義)
- 透過檢查數值範圍 (五值、平均數及標準差) 或繪製散點圖 (scatter)、分布圖 (histogram) 或其他圖檢查是否有異常。

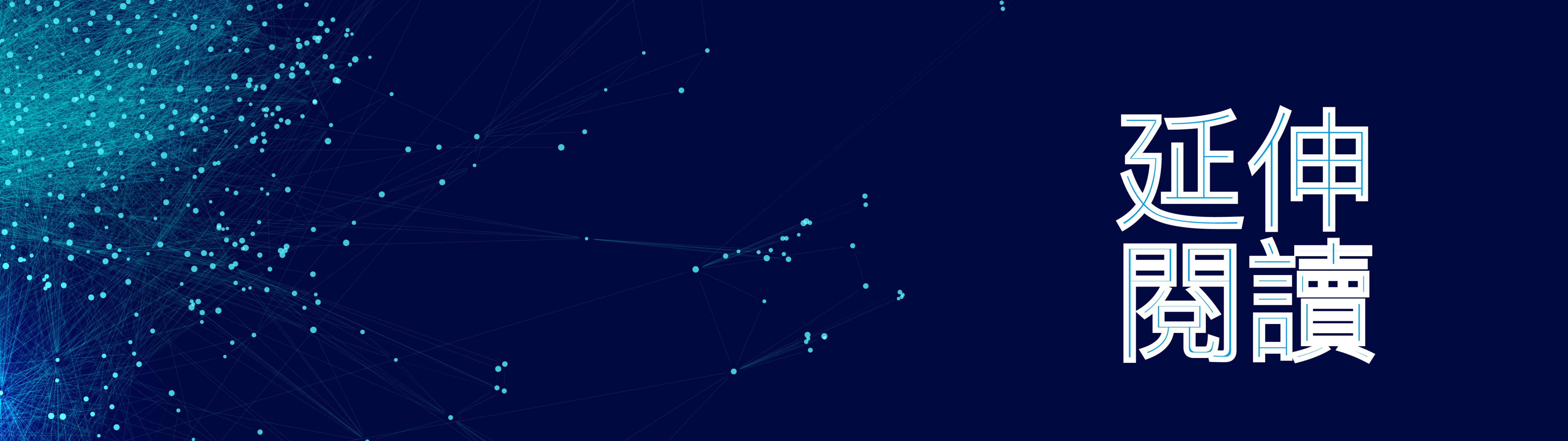
## 對 Outliers 的處理方法

- 新增欄位用以紀錄異常與否
- 填補 (取代)
- 視情況以中位數, Min, Max 或平均數填補(有時會用 NA)

# 重要知識點複習



- 檢查異常值的方法
  - 統計值：如平均數、標準差、中位數、分位數
  - 畫圖：如直方圖、盒圖、次數累積分布等
- 處理異常值
  - 取代補值：中位數、平均數等
  - 另建欄位
  - 整欄不用



# 延伸 閱讀

除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有  
多餘時間，可再補充延伸閱讀文章內容。

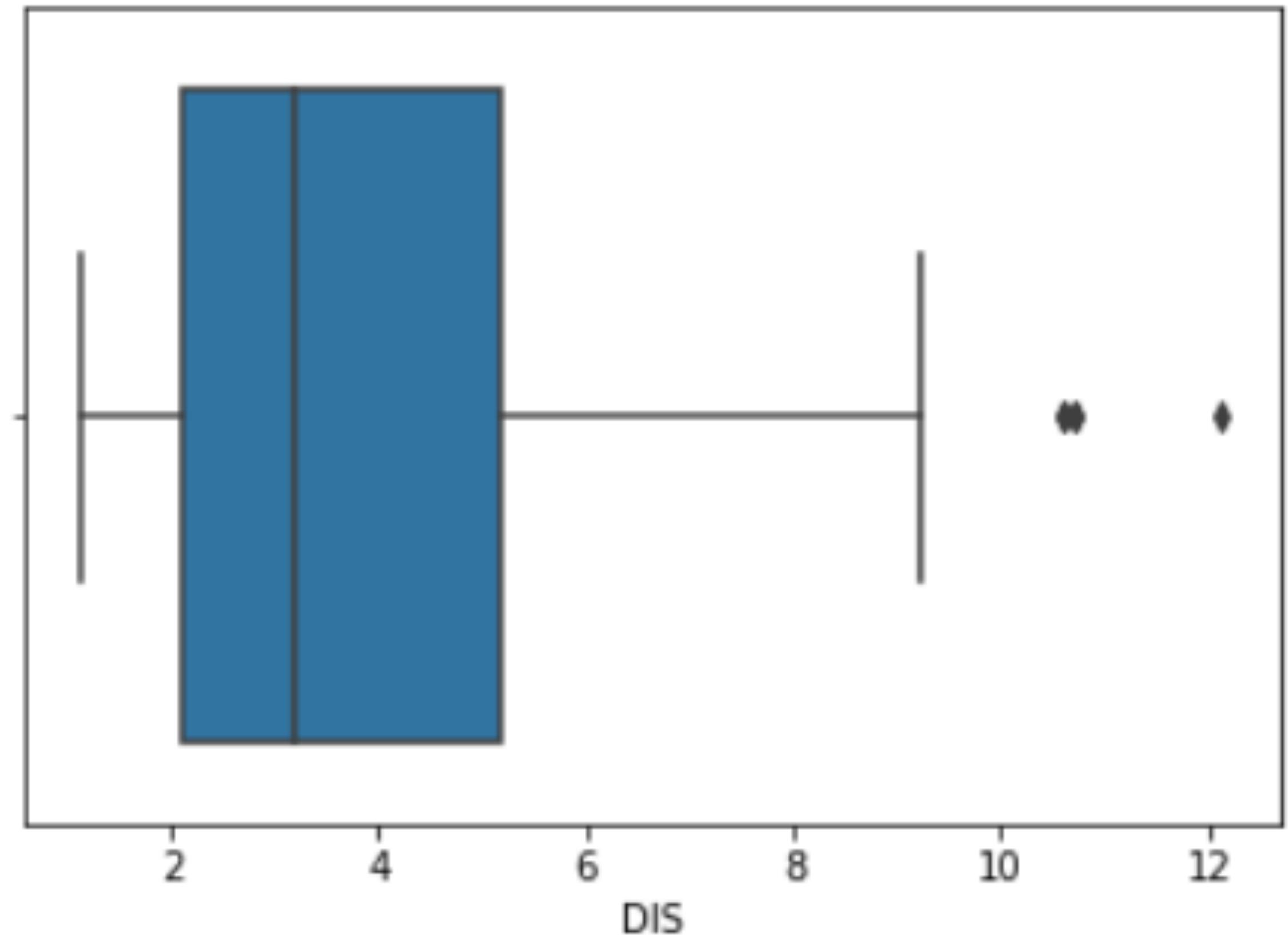
# 推薦延伸閱讀

Ways to Detect and Remove the Outliers

網頁連結

閱讀重點：

- 視覺方法 - boxplot, scatter plot
- 統計方法 - zscore, IQR



# 推薦延伸閱讀

## How to Use Statistics to Identify Outliers in Data

### 網頁連結

- 閱讀重點：

- 標準差與容忍範圍
- 個標準差：涵蓋 68% 數據
- 個標準差：涵蓋 95% 數據
- 個標準差：涵蓋 99.7% 數據
- 舉例來說，假設一個數字超過平均值 + 3 個標準差，那代表這個樣本點非常罕見！（所以要不是很特別，就是它的發生來自某種問題）

```
2 from numpy.random import seed
3 from numpy.random import randn
4 from numpy import mean
5 from numpy import std
6 # seed the random number generator
7 seed(1)
8 # generate univariate observations
9 data = 5 * randn(10000) + 50
10 # calculate summary statistics
11 data_mean, data_std = mean(data), std(data)
12 # identify outliers
13 cut_off = data_std * 3
14 lower, upper = data_mean - cut_off, data_mean + cut_off
15 # identify outliers
16 outliers = [x for x in data if x < lower or x > upper]
17 print('Identified outliers: %d' % len(outliers))
18 # remove outliers
19 outliers_removed = [x for x in data if x >= lower and x <= upper]
20 print('Non-outlier observations: %d' % len(outliers_removed))
```

1 Identified outliers: 29  
2 Non-outlier observations: 9971



解題時間

It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

