天氣與人工智慧
機器學習實作

資料清理數據前處理 - 2

2020/10/28

Subject

- EDA
  - 資料類型介紹
  - 資料分布
  - 檢查outlier

知識地圖 機器學習前處理 欄位的資料類型介紹及處理

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

前處理 Processing → 探索式數據分析 Exploratory Data Analysis → 特徵工程 Feature Engineering → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

- 分群 Clustering
- 降維 Dimension Reduction

前處理 Processing

資料讀取 → 格式調整 → 填補缺值 → 去離群值 → 特徵縮放

本節重點

- 瞭解 pandas datafreame 欄位的基本資料類型

資料類型

資料的欄位變數一般可分

01

- 離散變數: 只能用整數單位計算的變數
  - 房子的房間數量、性別、國家
- 連續變數: 在一定區間內可以任意取值的變數
  - Ex: 測量的身高、飛機起飛到降落所花費的時間、車速

Pandas DataFrame中 最常見的欄位資料類型有三種

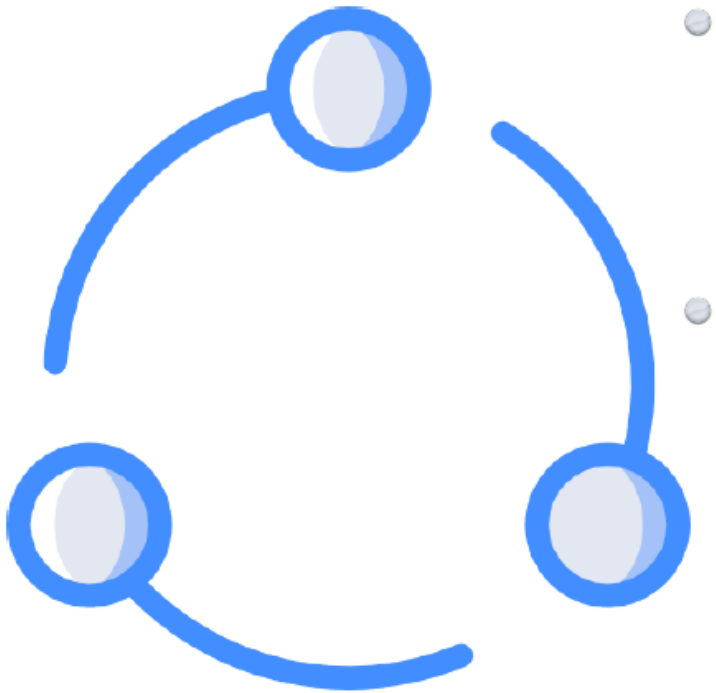
02

- float64 ： 浮點數，可表示離散或連續變數
- int64 ： 整數，可表示離散或連續變數
- object ： 包含字串，用於表示類別型變數

當然還有日期、boolean 等等不同的格式，實務中遇到再 google 就好

03

重要知識點複習



- 拿到資料的第一步，通常就是看我們有什麼，觀察有什麼欄位，這些欄位代表什麼意義、以什麼樣的資料類型來儲存
- 資料原來是字串/類別的話，如果要做進一步的分析時（如訓練模型），一般需要轉為數值的資料類型，轉換的方式通常有兩種
  - Label encoding：使用時機通常是該資料的不同類別是有序的，例如該資料是年齡分組，類別有小孩、年輕人、老人，表示為 0, 1, 2 是合理的，因為年齡上老人 > 年輕人、年輕人 > 小孩
  - One Hot encoding：使用時機通常是該資料的不同類別是無序的，例如國家

網頁連結

這邊完整的列舉了 Pandas 所有的類別型態，同學大概知道有哪些即可，若有需要深入了解的，我們在後面的課程會再提及。

Pandas所支持的數據類型

1. float
2. int
3. bool
4. datetime64 [ns]
5. datetime64 [ns,tz]
6. timedelta [ns]
7. category
8. object

默認的數據類型是int64,float64.

將資料轉換成編碼

- Label encoding: 把每個類別 mapping 到某個整數，不會增加新欄位
- One Hot encoding: 為每個類別新增一個欄位，用 0/1 表示是否

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

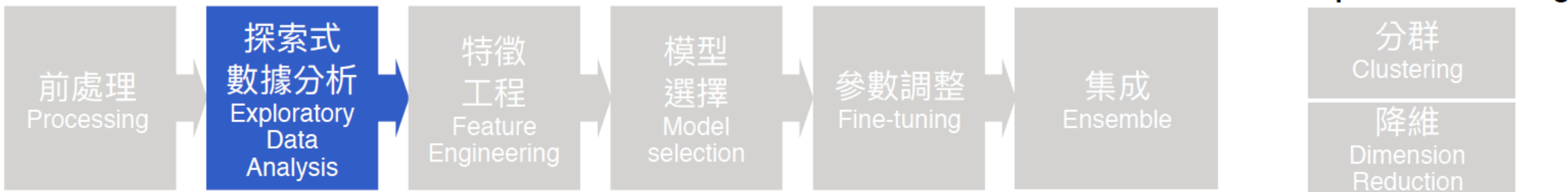

解題時間
It’s Your Turn

知識地圖 探索式數據分析 EDA 資料分布

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

非監督式學習
Unsupervised Learning



探索式數據分析 Exploratory Data Analysis (EDA)

統計值的視覺化

- 相關係數
- 核密度函數
- 離散化

繪圖排版

- 常用圖形
- 模型體驗

本節重點

- 瞭解如何利用基本的統計數值以及畫圖來瞭解資料

EDA - 統計量化的方式？

以單變量分析來說，量化的分析方式可包含

計算集中趨勢

- 平均值 Mean
- 中位數 Median
- 眾數 Mode

計算資料分散程度

- 最小值 Min
- 最大值 Max
- 範圍 Range
- 四分位差 Quartiles
- 變異數 Variance
- 標準差 Standard deviation

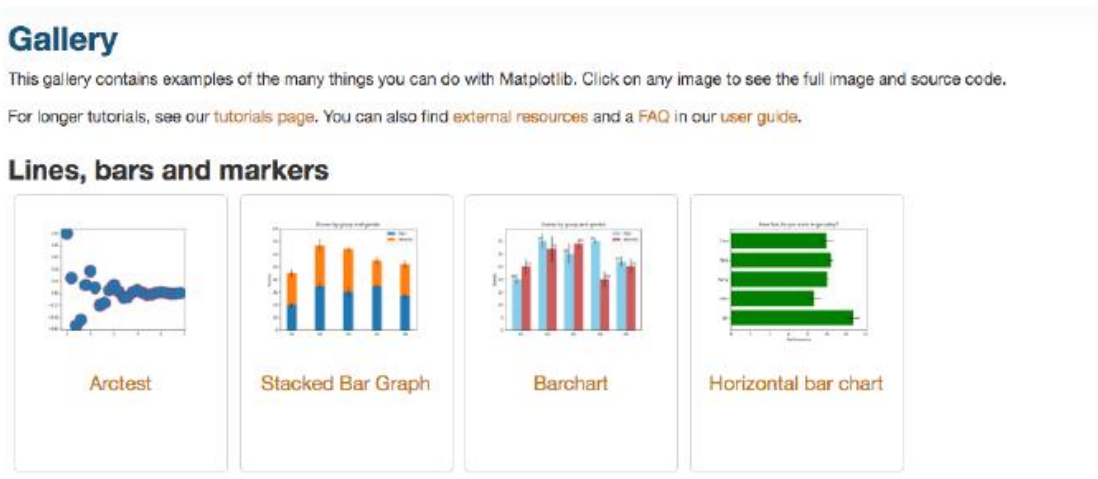
基本上使用上述統計特徵就可以讓我們初步了解資料的樣子，並且觀察是否有異樣

EDA視覺化的方式？

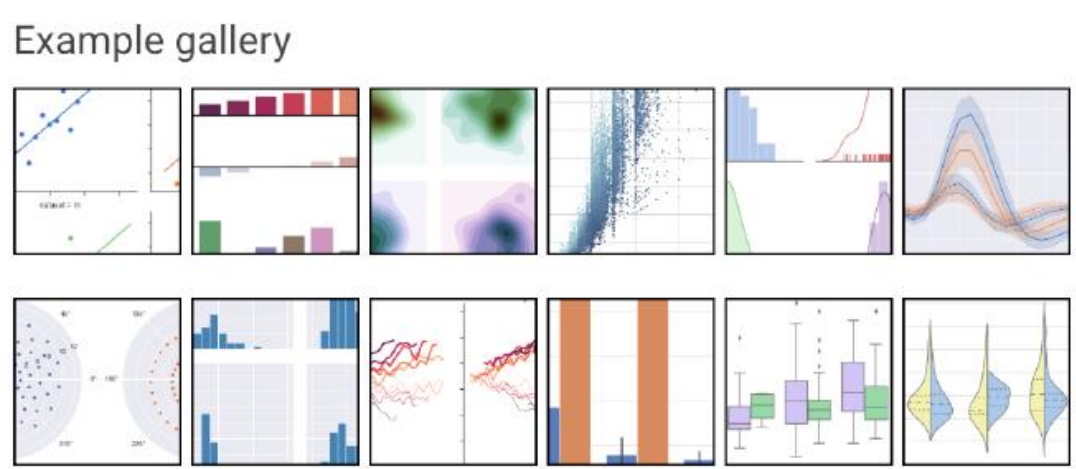
有句話「一畫勝千言」，除了數字，視覺化的方式也是一種很好觀察資料分佈的方式，可參考 python 中常用的視覺化套件

畫圖沒靈感的時候可以到這兩個套件的範例網頁逛逛！

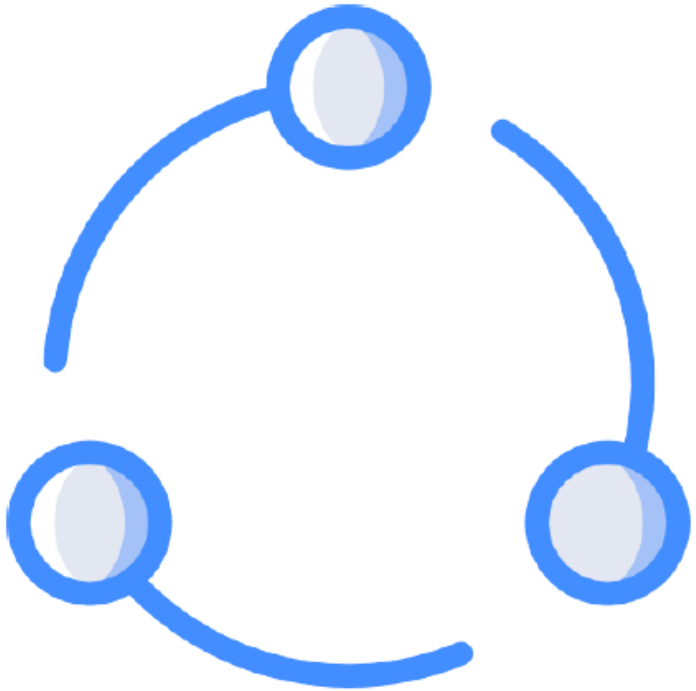
matplotlib



seaborn



重要知識點複習



- 資料大部分時候都是非常多的，我們沒辦法用眼睛一筆一筆都看完，平均值、標準差、最大最小值等統計數值能幫助我們迅速對資料有初步的了解。
- 了解統計數值後，把資料的圖畫出來除了能夠更全面地了解資料，也能幫我們快速觀察到異常的地方
- pandas 有許多已經寫好用來做以上這些觀察的函數，熟悉這些函數的使用能加速觀察資料的過程


解題時間
It’s Your Turn

知識地圖 機器學習前處理 Outlier 及處理

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

前處理 Processing → 探索式數據分析 Exploratory Data Analysis → 特徵工程 Feature Engineering → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

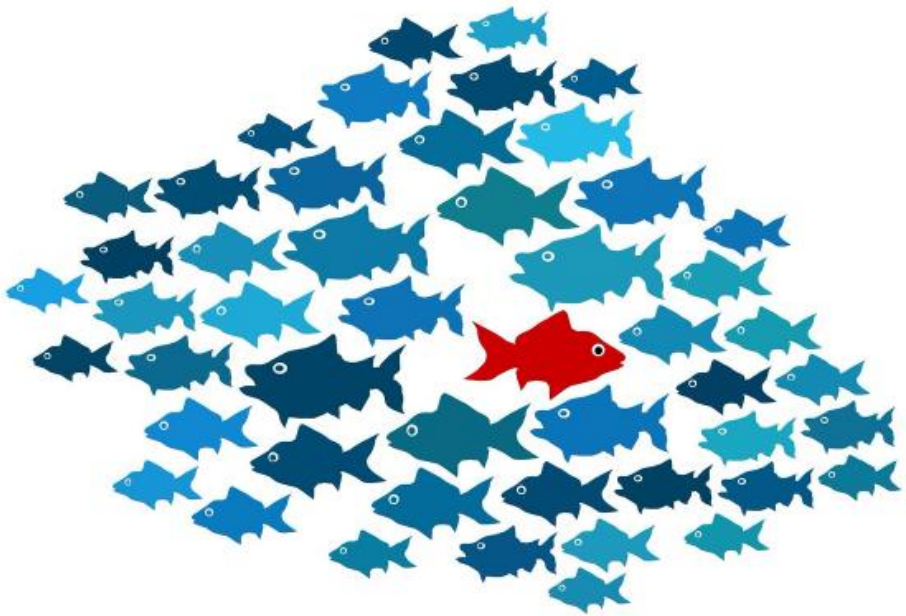
- 分群 Clustering
- 降維 Dimension Reduction

前處理 Processing

資料讀取 → 格式調整 → 填補缺值 → **去離群值** → 特徵縮放

本節重點

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



圖片來源: Sergio Santoyo

1 異常值 (Outliers) 出現的可能原因

1.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

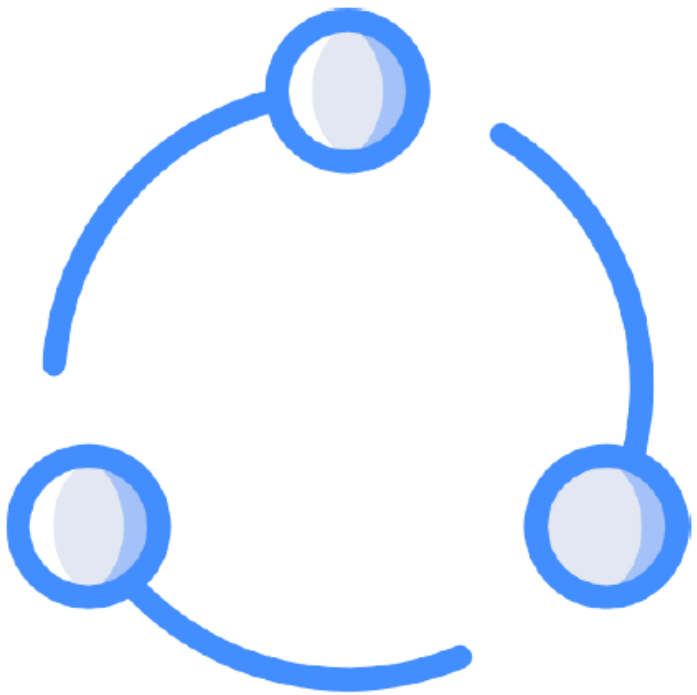
2 檢查 Outliers 的流程與方法

- 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。

3 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

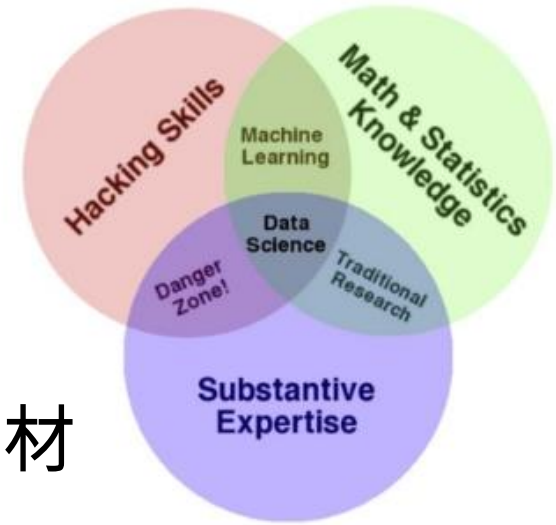


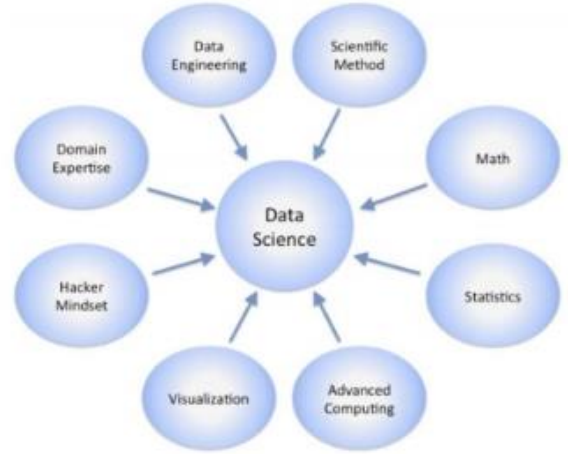
- 檢查異常值的方法
  - 統計值：如平均數、標準差、中位數、分位數
  - 畫圖：如直方圖、盒圖、次數累積分布等
- 處理異常值
  - 取代補值：中位數、平均數等
  - 另建欄位
  - 整欄不用


解題時間
It’s Your Turn

参考資料

敘述統計與機率分布

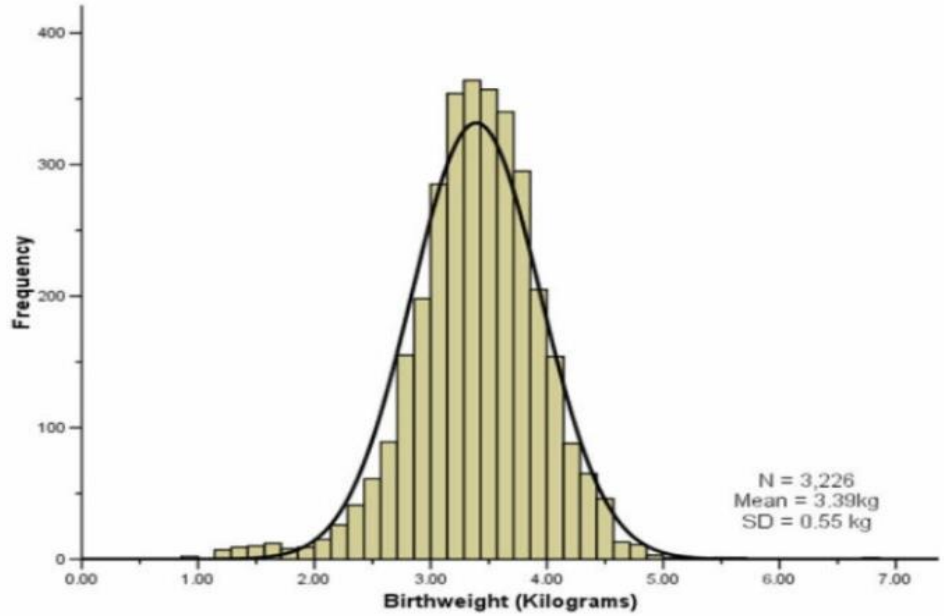




- 吳漢銘老師 AI學校經理人班教材
- [網頁連結]
- 要做出足夠深入的 EDA，對於統計的理解是必須的，這份教材可以提供同學了解統計觀念的機會，但是這份教材的範圍太廣，牽涉到太多預備知識，並不適合同學完整閱讀，只建議在不熟悉名詞時，回頭當作工具書參考即可。

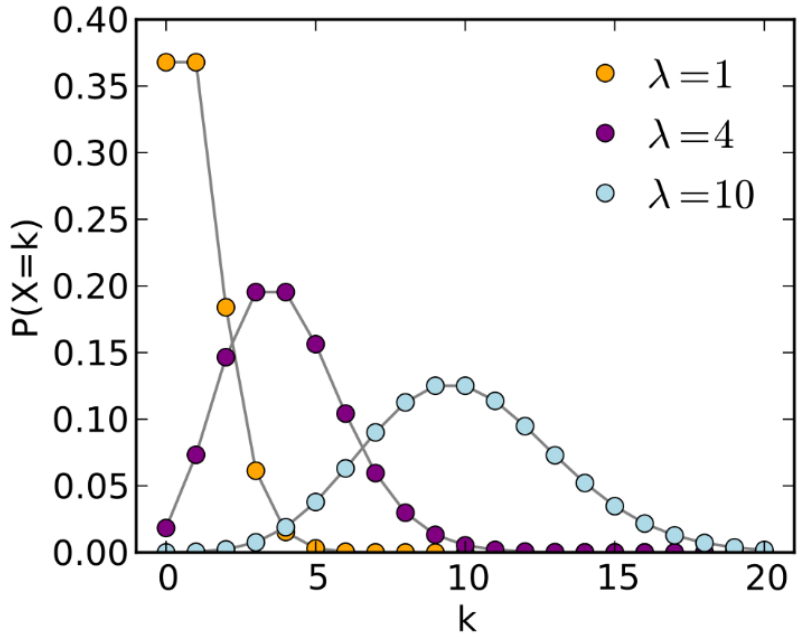
常見的統計分佈 (英文)

- 網頁連結
- 這個網頁描述了幾個常見的分布：常態分布 / 二項式分布 / 卜瓦松分布，其中常態分布是我們最常使用到的，這個網頁建議同學大致上知道常態分布的形狀即可，至於機率密度函數等其他相關知識，可以等到有需要時再查詢。



統計分佈清單 (英文)

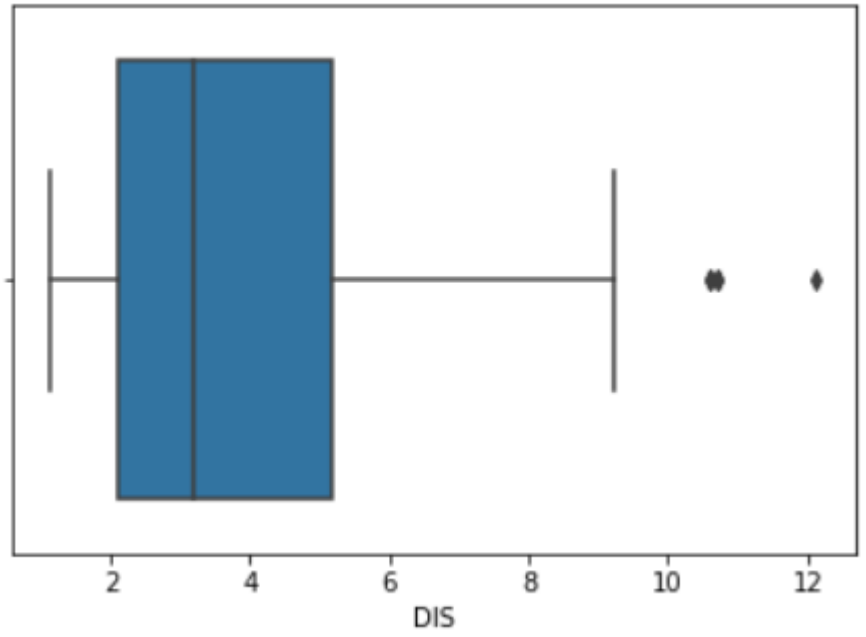
- 網頁連結
- 維基百科上有更完整的統計模型清單, 包含離散與連續分布，不過當清單到這麼完整的時候，就更不可能全部讀完，建議同學也是當作查詢工具即可。(附圖為卜瓦松分佈 Poisson Distribution)



Ways to Detect and Remove the Outliers

閱讀重點：

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



How to Use Statistics to Identify Outliers in Data

閱讀重點：

- 標準差與容忍範圍
  - 1 個標準差: 涵蓋 68% 數據
  - 2 個標準差: 涵蓋 95% 數據
  - 3 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)

```
from numpy.random import seed
from numpy.random import randn
from numpy import mean
from numpy import std
# seed the random number generator
seed(1)
# generate univariate observations
data = 5 * randn(10000) + 50
# calculate summary statistics
data_mean, data_std = mean(data), std(data)
# identify outliers
cut_off = data_std * 3
lower, upper = data_mean - cut_off, data_mean + cut_off
# identify outliers
outliers = [x for x in data if x < lower or x > upper]
print('Identified outliers: %d' % len(outliers))
# remove outliers
outliers_removed = [x for x in data if x >= lower and x <= upper]
print('Non-outlier observations: %d' % len(outliers_removed))
```

```
Identified outliers: 29
Non-outlier observations: 9971
```