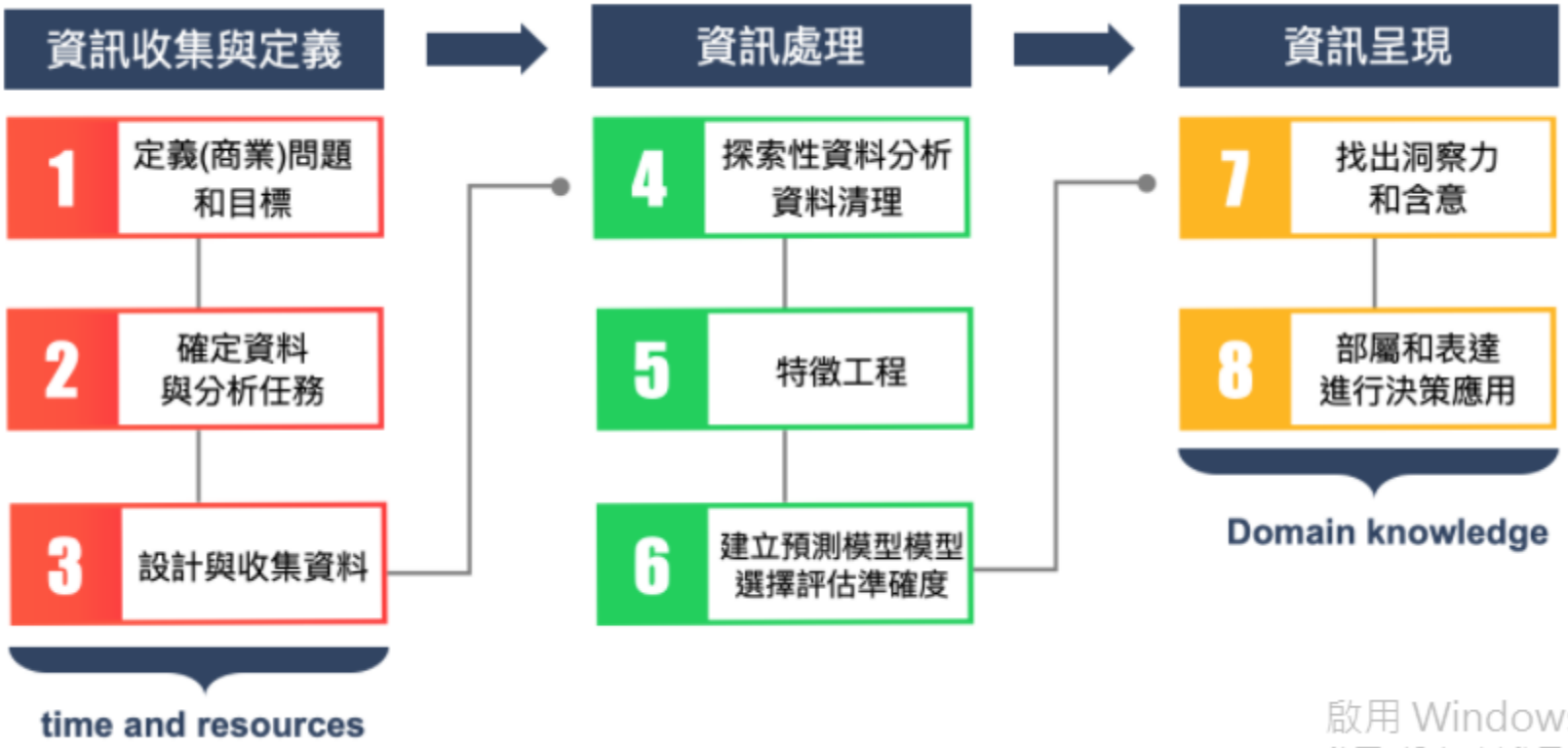
機器學習實作教育訓練

Agenda

- 資料科學的實作流程
- 探索性資料分析
  - 資料態樣
  - 遺失值
  - 異常值
  - 相關性
- 特徵工程
  - 過濾法
  - 包裝法
  - 嵌入法
- 預測模型
  - Logistic reg
  - Decision Tree
  - SVM
  - …

1 資料科學實作流程

“數據和特徵決定了機器學習的上限，
而模型和算法只是逼近這個上限而已”



啟用 Windows
移至 [設定] 以啟用 Wi

1 資料科學實作流程

“數據和特徵決定了機器學習的上限，而模型和算法只是逼近這個上限而已”

1.確認資料態樣
　　型態(object、int、float)、欄位的意義與分布
2.檢測 異常值 和 缺失值
3.發掘 特徵變數之間的關係
　(1) 特徵變數與目標變數之間的關係
　(2)除目標變數外，特徵變數彼此之間的關係
4.提取或找出 重要的特徵變數 [特徵工程]
5.選擇合適的模型 (機器學習)-> xgboost

鐵達尼資料為例

目標變數

變數(Variable)	意義(Meaning)	資料型態
PassengerId	乘客編號	數字
Survived	是否存活 (1:活/0:死)	離散
Pclass	票務艙 (1:Upper,2:Middle,3:Lower)	離散
Name	姓名	字串
Sex	性別	離散
Age	年齡	連續
SibSp	在船上兄弟姊妹配偶的人數 (定義家庭關係，非家庭關係不納入)	連續
Parch	在船上父母和子女的人數 (定義家庭關係，非家庭關係不納入)	連續
Ticket	船票號碼	字串
Fare	乘客票價	字串
Cabin	船艙號碼	離散
Embarked	登船港口	離散

確認資料態樣

```
from pandas_profiling import ProfileReport
report = ProfileReport(df)
```

(jupter titanic)

Overview

Overview | Warnings 16 | Reproduction

Dataset statistics

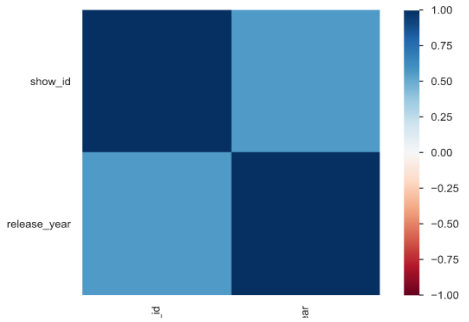
Number of variables	12
Number of observations	6234
Missing cells	3036
Missing cells (%)	4.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	584.6 KiB
Average record size in memory	96.0 B

Variable types

Numeric	2
Categorical	10

Correlations

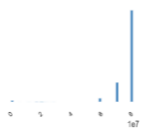
Pearson's r | Spearman's ρ | Kendall's τ | Phik (φk) | Cramér's V (φc) | Toggle correlation descriptions



Variables

show_id
Real number (ℝ≥0)
UNIQUE

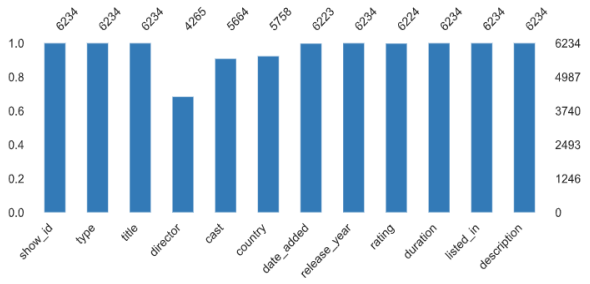
Distinct	6234	Mean	76703679.32
Distinct (%)	100.0%	Minimum	247747
Missing	0	Maximum	81235729
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	48.8 KiB



Toggle details

Missing values

Count | Matrix | Heatmap | Dendrogram



A simple visualization of nullity by column

2 探索性資料分析

遺失值

WHAT

遺失類型	定義	例子
完全隨機缺失 (Missing Completely at Random)	缺失資料與該變數的真實值無關，與其他變數的數值也無關。	老師走在路上，不小一掉了幾張考卷，所以成績的遺失和成績無關，和學生其他變數也無關。
條件隨機缺失 (Missing at Random)	缺失資料與其他變數有關。	一群學生的體重，發現有些人的體重缺失，後來發現女生不喜歡寫體重資訊，因此體重的資訊遺失與性別相關。
非隨機缺失 (Missing not an Random)	缺失資料依賴於該變數本身。	有分收入的問卷調查，通常薪資高的人，不喜歡填有關收入資訊，所以收入資料缺失和收入高低有關。

WHY

有些程式與演算法，不容許有遺失值出現
不能影響整體的估計，填補數與已有的數的分佈、特徵應符合

HOW

可以分為刪除和補值兩類。

(1)刪除：刪除有過多缺失資料的變數 (通常超過 60% 就會刪除)

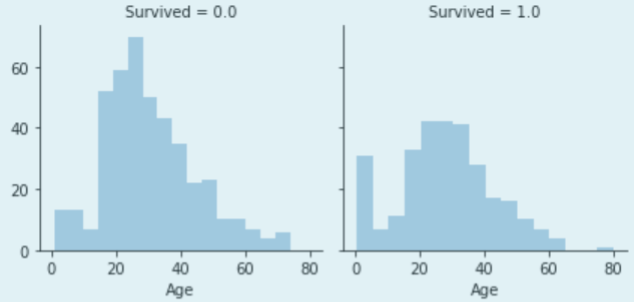
(2)補值：

1.給定一個固定值去填補遺失值(ex : 0)
2.由後往前補值 或 由前往後補值(時間性相關適用)
3.用現有的資料取平均值、中位數、眾數等進行補值
4.用預測方法補值，迴歸或機器學習

2 探索性資料分析

年齡變數的補值(Day35)

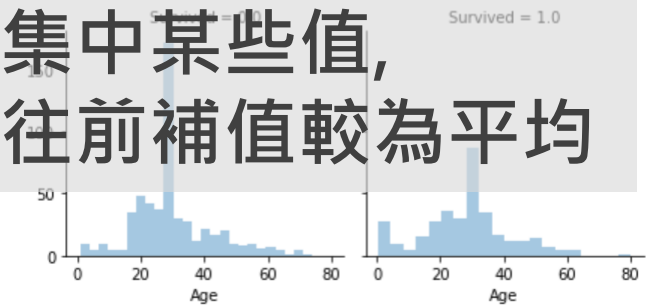
原資料分布



由平均值補值會集中某些值,
由前往後或由後往前補值較為平均

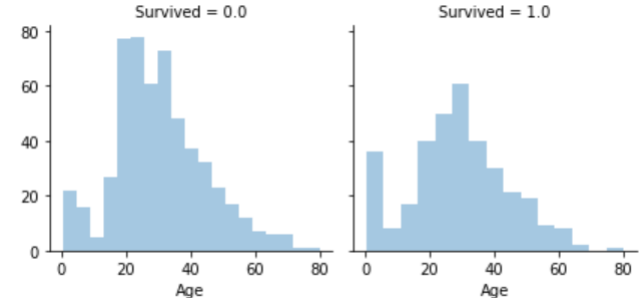
1.用平均值補

```
df['Age'] = df['Age'].fillna(df['Age'].mean())
```



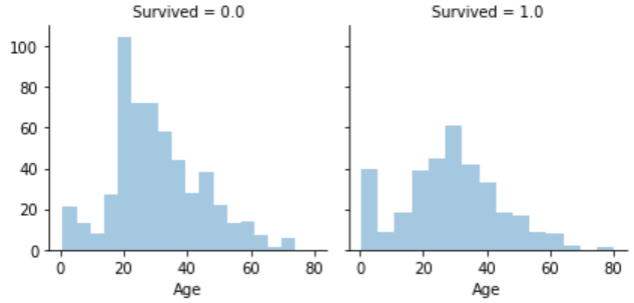
2.由前往後補

```
df['Age'] = df['Age'].fillna(method='bfill')
```



3.由後往前補

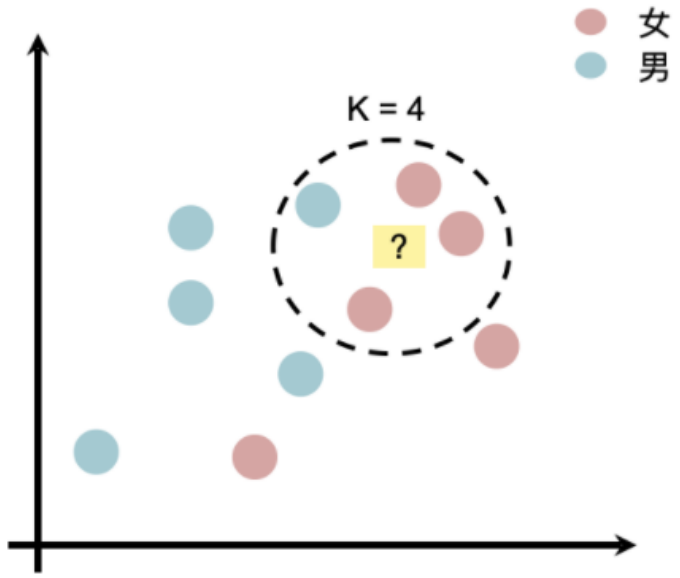
```
df['Age'] = df['Age'].fillna(method=  'pad')
```



4.用KNN補值(Day37)

什麼是 KNN？

- K-Nearest Neighbor(KNN) 是一種無須機率分配的假設下的演算法，跟距離預測值最近的 k 個數值，來估計預測值。
- 以下圖為例， 要預測這個人是男生或女生，以這個例子來看，透過 KNN (k=4) 的預測值為女生



- 由於類別型資料，透過投票來決定預測值，因此 k 建議以奇數為主，避免掉平手的問題。

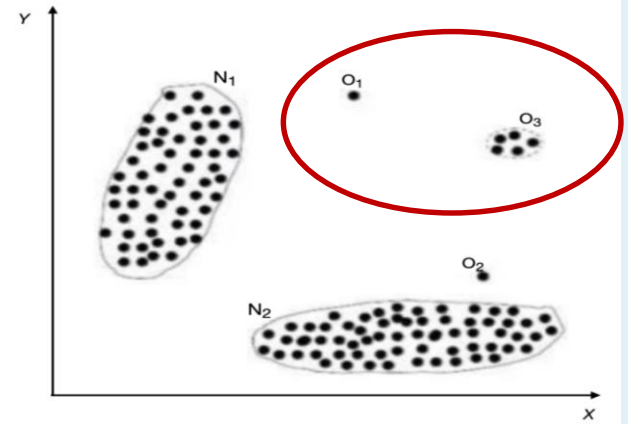
KNN 的三個步驟：
1.計算距離
2.尋找最近 k 組數值
3.類別型態資料以多數決
　數值型態用 k 組資料的統計值 (如平均)

2 探索性資料分析

異常值

WHAT

1.偏離樣本整體數據的值，所以是採用一種相對的概念。
2.也被稱為離群值、新奇、噪聲、偏差和例外。



WHY

因為出現異常值的原因不同，如輸入錯誤、故意離群、自然異常等
了解背後的原因才能決定處理方式

HOW

1.刪除異常值：由於數據輸入錯誤、數據處理錯誤或異常值數目很少

2.數據轉換：例如對數據取對數可以減少極端值的變化

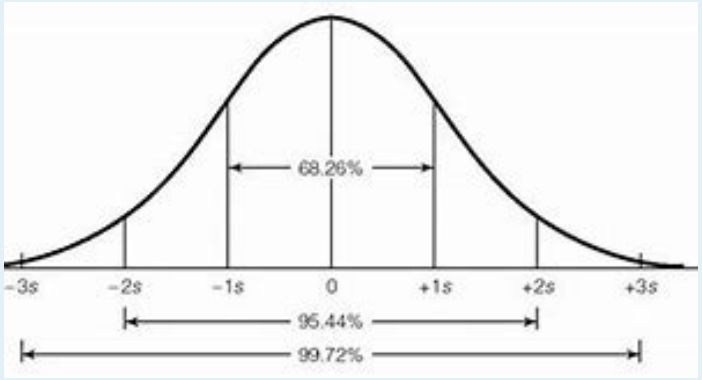
3.分離對待：如果異常值的數目比較多，在統計模型中我們應該對它們分別處理

4.替換：替換缺失值，我們也可以替換異常值。我們可以使用均值、中位數、眾數替換方法。

5.聚類：我們也可以用決策樹直接處理帶有異常值的數據（決策樹基本不會受到異常值和缺失值的影響），或是對不同的觀測值分配權重。

票價的異常值檢測

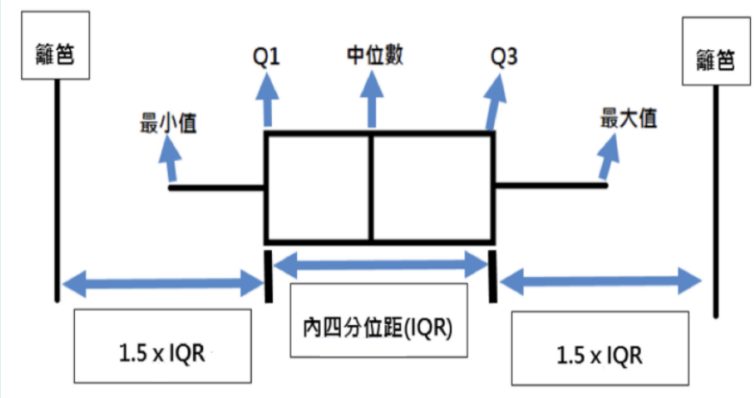
1. 3σ 原則

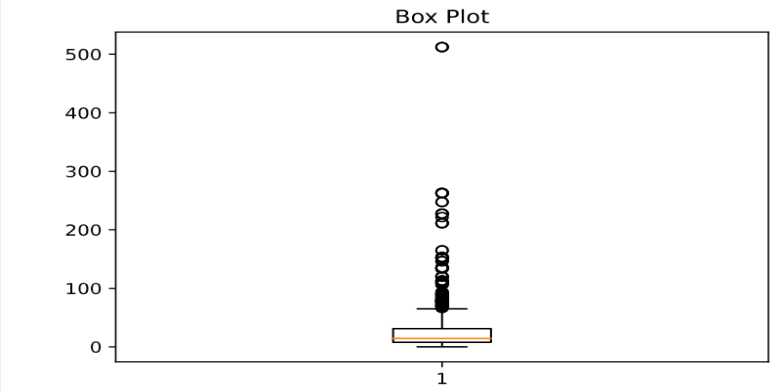



```
In [4]: #1.最大值來看為異常值，超過1.5的IQR,也超過三倍標準差範圍外
        df_train['Fare'].describe()

Out[4]: count    891.000000
        mean      32.204208
        std       49.693429
        min        0.000000
        25%        7.910400
        50%       14.454200
        75%       31.000000
        max      512.329200
        Name: Fare, dtype: float64
```

2.盒鬚圖 (箱型圖)

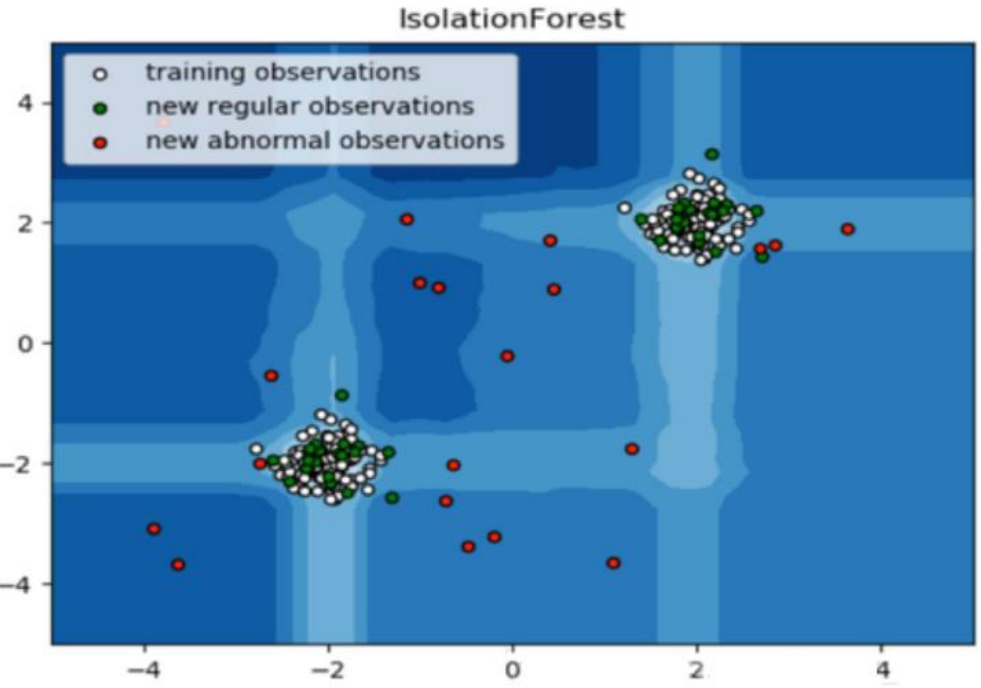


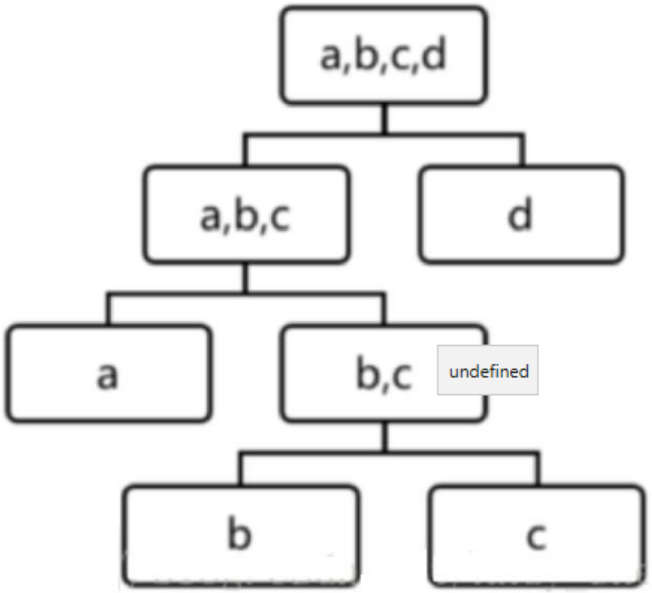


異常值檢測

3.異常檢測模型
異常檢測模型是針對整體樣本中的異常資料進行分析和挖掘，以便找到其中的異常個案和規律

孤立森林IsolationForest為例：





b 和c 的高度为3，a的高度是2，d的高度是1。可以看到d 最有可能是异常，因为其最早就被孤立（isolated）了。

「企業戶每日活存餘額異常監測」為例

IsolationForest

```
from sklearn.ensemble import IsolationForest

data = df3[['amt']]

# train isolation forest
model = IsolationForest(contamination=outliers_fraction)
model.fit(data)
df3['anomaly2'] = pd.Series(model.predict(data))
a = df3.loc[df3['anomaly2'] == -1, ['date', 'amt']] #anomaly
print(f'異常個數: {len(a)}')
print(f'異常值: {a}')
```

```
異常個數: 9
異常值:            date        amt
27   2020-01-30   0.595584
118  2020-04-30   1.983515
122  2020-05-04  -1.234657
150  2020-06-01  -0.587087
169  2020-06-20  10.240702
171  2020-06-22  -5.318800
172  2020-06-23  -4.907934
173  2020-06-24   2.665800
178  2020-06-29  -2.196487
```

2 探索性資料分析

相關性

WHY

挖掘變數之間的關係，有助提升模型效度

HOW

1.年齡與票價
(連續 vs 連續)

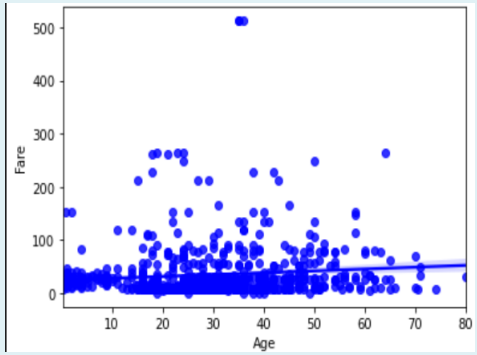
Pearson 相關係數=0.096

相關係數

低度線性相關

相關係數範圍	變數之間關聯程度
1	完全線性相關
0.7-0.99	高度線性相關
0.4-0.69	中度線性相關
0.1-0.39	低度線性相關
< 0.1	無線性相關

圖示觀察

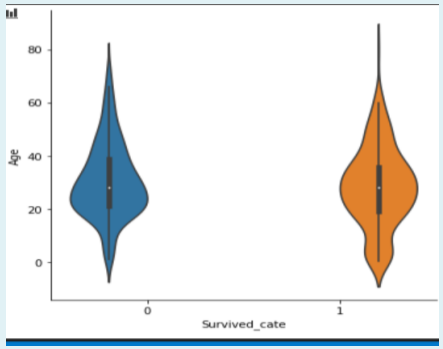


2.年齡與存活
(連續 vs 離散)

eta-squared=0.005

低度相關

η^2	Interpretation
0.00 < 0.01	Negligible
0.01 < 0.06	Small
0.06 < 0.14	Medium
0.14 <= 1.00	Large

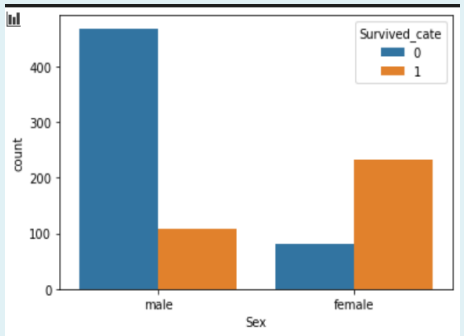


3.性別與存活
(離散 vs 離散)

Cramer’ s V 係數=0.543

高度相關

df*	negligible	small	medium	large
1	0 ~ .10	.10 ~ .30	.30 ~ .50	.50 or more
2	0 ~ .07	.07 ~ .21	.21 ~ .35	.35 or more
3	0 ~ .06	.06 ~ .17	.17 ~ .29	.29 or more
4	0 ~ .05	.05 ~ .15	.15 ~ .25	.25 or more
5	0 ~ .05	.05 ~ .13	.13 ~ .22	.22 or more



3 特徵工程

WHAT

WHY

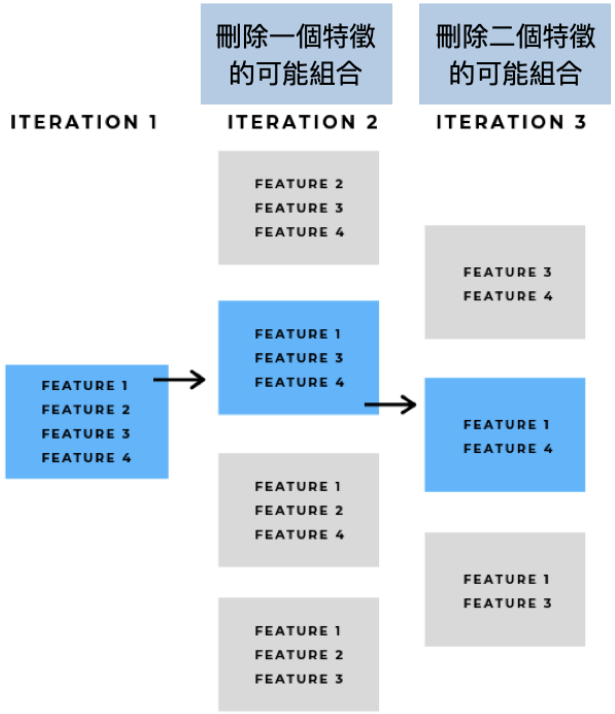
透過這些特徵能把目標做清楚的分類與預測，藉此改善模型性能。
特徵選擇是特徵工程裡的一個重要問題，能剔除不相關或冗餘的特徵，從而達到減少特徵個數，提高模型精確度，減少運行時間的目的。

HOW

一、特徵工程包含衍生和添加。
衍生：以現有收集的資料為主，透過探索性分析，了解資料與目標之間的關係後，產生出特徵。X^2
添加：現有收集資料以外的資訊

二、具體特徵選擇方法
1.Filter：過濾法，按照發散性或者相關性對各個特徵進行評分，設定閾值或者待選擇閾值的個數，選擇特徵。
2.Wrapper：包裝法，根據目標函數（通常是預測效果評分），每次選擇若干特徵，或者排除若干特徵。
3.Embedded：嵌入法，先使用某些機器學習的算法和模型進行訓練，得到各個特徵的權值係數，根據係數從大到小排序選擇特徵。類似於Filter方法，但是是通過訓練來確定特徵的優劣。

Wrapper：包裝法，根據目標函數（通常是預測效果評分），每次選擇若干特徵，或者排除若干特徵。(Day40)



運用 python 執行包裝法

Step 1：sex 離散型要先轉成數值型態

Step 2：根據目標變量是連續或離散，來決定判斷的準則。
離散型，SVC(kernel="linear")

Step 3：設定 RFE 裡面的參數

- n_features_to_select：最後要選擇留下多少特徵。
- Step：刪除法，每一部刪除多少特徵。

Step 4：.fit(x.y)：每一步都依不同的特徵組合建立模型，判斷最終要選擇那些特徵

Step 5：透過 support_ 呈現包裝法搭配 SVC 下，選擇最好的特徵，用 True 來表示

Step 6：透過 ranking_ 呈現每個特徵對於模型的重要性，1 代表被選重的特徵，2 代表次之重要的特徵，依此類推