**Task #05 作業 01【簡答題】特徵工程與資料視覺化**

1. 請問「資料清理」、「資料視覺化」與「特徵工程」在資料前處理中扮演的角色是什麼？

資料清理：確保數據質量，處理缺失值、異常值、重複值等提高數據完整性和一致性。

資料視覺化：利用圖表呈現資料的分布情形及關聯性，幫助探索數據的趨勢及結構等等。

特徵工程：針對有意義的資料進行特徵萃取、轉換、正規化標準化等，使得更好的描述內在結構及模式。

2. 你有聽過什麼叫做「不平衡資料集 （Unbalanced Data）」嗎？試著舉例生活中可能存在不平衡特性的資料集。

(在這裡回答)

Unbalanced Data代表在一群不同類別中的資料有分布不均的狀況，或是都集中在某一類別上，

例如在針對垃圾郵件檢測的時候，通常垃圾郵件數量會比正常郵件數量少，或是在一些醫療罕見疾病上的患者數量很少，健康人群很多，這樣的樣本數量差距就會很大。

3. [長尾理論](https://zh.wikipedia.org/wiki/%E9%95%BF%E5%B0%BE)是商業世界中一種銷售的趨勢，意旨熱門的商品銷售數量會遠大於非熱門的產品，其銷售成果如下。請問如果你的資料中遇到的這種偏移的現象，該如何處理？

一張含有 圖表, 行, 文字, 繪圖 的圖片

自動產生的描述

1. 對於數據進行特徵縮放，利用Z-score將數據縮放為標準常態分布形式，或是進行對數或平方根的轉換
2. 對於熱門的商品進行Undersampling，減少數據量，對於冷門商品進行Oversampling以增加數據量。
3. 或是針對不同商品分配不同權重，使得冷門商品也可以在數據中獲得關注。

4.特徵工程有一種方法稱為「裝箱法（ Binning、 Bucketization ）」，主要用於將連續資料轉成離散的型態。例如，我們可以將連續的年紀欄位（例如：3, 18, 22, 16, 28）變成「小孩」、「少年」、「中年」和「老年」四種可能的離散資料。請問什麼時候適合採用裝箱法進行資料的處理呢？

對於連續資料數據如身高年紀或是音量分貝、薪水階級等等的數據都可以透過裝箱法將連續型態數據轉換成離散型態，減少數據的複雜性，避免模型overfitting等問題產生。

5. 請問以下這個資料是否存在什麼問題？可以怎麼做調整？

一張含有 文字, 螢幕擷取畫面, 數字 的圖片

自動產生的描述

1. 使用者編號對於數據集並沒有直接的幫助，可以將這資料drop掉。
2. 對於性別可以進行one-hot encoding轉換
3. 對於身份證字號因字首開頭僅有E字母，可能造成樣本數不足。
4. 身分證字號欄位可以利用特徵工程取出有意義的數據。
5. 對於年紀可以用特徵轉換將他們對應到年齡的區間分群，減少數據複雜性。