Task #04 作業 01【簡答題】資料前處理與資料清理

1. Python/Pandas 中有兩個空的型態「NaN」和「None」，請問這兩個是一樣的嗎？還是有什麼差別？

NaN表示一個非數值(Not a Number)，如0/0, ∞/∞…，在 Pandas 中被視為一個浮點數。通常出現在數值運算中。

None表示缺少一個數值，在 Pandas 中被視為一個 Python 物件。通常出現在物件運算中。

2. 在資料清理的過程中，我們可以使用填補空值的方法避免遺失值帶來的錯誤。實務上會建議「連續數值的話補平均，離散數值的話補眾數」，請問為什麼？

補平均及補眾數皆屬於能夠保持資料的整體分布特性，不會引起太大的偏差。

3. 承上題，也有人會建議可以利用統計內差/回歸或機器學習的方法進行空值填補，你覺得這是一種好方法嗎？為什麼？

利用統計內插回歸或ML的方法，也許對於簡單的補平均值或補眾數更能保持資料分布的特性，但由於計算成本可能較高(如計算時間)，對於使用者的專業知識技能要求也較高，整體所付出的成本較高，因此不會說是不好的方法，但應該視情況而定經過評估時間等成本後，再考慮是否使用此類型補值方法。

4. 可以請你解釋一下「Label-Encoding」和「One-Hot-Encoding」分別用於什麼情境？

Label-Encoding適用於當此欄位的值具有階層性或具有順序關係時使用。

如：年紀、名次…等

One-Hot-Encoding適用於欄位的值不相關或不具有順序關係時使用。

如：各種水果、不同種品牌…等。

5. 請問如果以下幾種字串資料，你應該要用哪一種編碼方法處理？或是其他建議？

手機號碼：0988123123、0937666666 ...

郵遞區號：300、807 ...

銀行帳號末五碼：33136、25466 ...

因為手機號碼、郵遞區號及銀行帳號這種欄位的值並無順序相關或有其他相關性，因此會建議使用One-Hot-Encoding方式進行編碼。