資料探索(ED) : 了解、清理和準備數據，多次迭代步驟 4 到 7，才能提出我們的改進模型
步驟 : 1.變量識別 2.單變量分析 3.雙變量分析 4.缺失值處理 5.異常值處理 6.變量變換 7.變量創建

(一) 數據探索和準備步驟 :

1. 變量識別 : Step1 , 變數種類 : 預測變量（輸入）& 目標（輸出）變量
   Step2 , 數據類型 : 字串、數值 ; 變數類別 : 分類、連續
2. 單變量分析 : 分類變量 : 1. 頻率表
   2. 每個類別下的百分比，用 Count 和 Count% 指標來衡量
   3. 條形圖
   連續變量 : 1. 集中趨勢：平均、中位數、眾數、最小值、最大值
   2. 離散度量：範圍、四分位數、IQR、變異數、標準差、偏度、峰度
   3. 直方圖、箱型圖

NOTE : 單變量分析也用於突出缺失值和異常值。

3. 雙變量分析 : 在最開始定義的顯著水平上，尋找關聯和分離
   連續&連續 :1. 散點圖 : 線性或非線性的關係，不能代表強度
   2. 相關性 : -1：完全負線性相關、+1：完全正線性相關、0：無相關性
   Correlation = Covariance(X,Y) / Sqrt (Var(X) * Var(Y)) , $-1 \leq Cor \leq 1$
   Pearson Correlation : Excel : CORREL()、SAS : PROC CORR
   分類&分類 : 1. 雙向表 : row : 一個變量的類別，column : 另一個變量的類別
   每個組合中觀察值的計數(count)或計數百分比(count%)
   2. 堆積柱形圖：視覺化雙向表
   3. 卡方檢驗：檢驗變量之間關係的統計顯著性
   檢驗樣本中的證據是否足夠強大
   卡方是基於雙向表中一個或多個類別的預期數和觀察數之間的差異
   機率 0 : 兩個變量相依；機率 1 : 兩個變量獨立
   機率小於 0.05：變量之間的關係在 95%的信心下顯著

檢驗兩個分類變量獨立性的卡方檢驗統計量：

$\chi^2 = \sum \frac{(O-E)^2}{E}$, O 為實際觀察數 , E 為虛無假設下的預期數 , $E = \frac{row\ total * column\ total}{sample\ size}$

關係檢定力的統計測度：Cramer’ s V : for Nominal Categorical Variable (名義分類變量)
Mantel-Haenszed Chi-Square : for ordinal categorical variable (有序分類變量)
分類&連續 : 1. 為分類變量的每個級別繪製箱型圖
2. 如果級別數量較少，將不會顯示統計顯著性

統計顯著性(平均值) : Z - test : $z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, 如果 Z 的機率小，則兩個平均值差異顯著

T - test : $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$, $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, 兩類別觀察數均小於 30 時用

ANOVA : 比較兩組以上的平均值