**ML 房價資料分析**

**(1) 目標變數（Target）與特徵（Features）**

* Target: 要預測的東西：SalePrice（成交價）。
* Features(hypothesis): 能幫忙預測的資訊：坪數、屋齡、社區、廚房品質…每一欄就是一個特徵。

**(2) 為什麼做 EDA（探索式資料分析）**

就像先「認識病人」再開藥：先看價錢分布、有哪些缺值、哪些特徵跟價錢關係強，再決定怎麼處理。

做完 EDA，模型會更穩、分數更好，也比較能解釋。

**(3) 需要產生的圖**

* 目標分布（直方圖）：看價錢是長尾還是平均。房價常「右偏」（貴的房子很少但非常貴）。
* log1p：對 SalePrice 做 log(1+price)，把「右偏」拉回來，線性回歸會更穩。
* 缺值 Top-20：哪些欄位常缺資料？要訂規則怎麼補。
* 類別欄位出現次數 Top-30：哪些分類很雜？One-Hot 之後欄位會暴增，要有心理準備。
* 連續特徵 vs 價格（散點圖）：看「坪數↑ 價格是否↑」。
* 分組圖（盒鬚圖/中位數條圖）：像 OverallQual（整體品質）和 Neighborhood（社區）對價錢影響有多大。
* 皮爾森相關（熱力圖 / Top-20）：量化「線性」關係強不強（|r| 越大越強）。
* 衍生特徵關係圖：把多欄位合成更有意義的指標（例如總面積 TotalSF、總衛浴 TotalBath），常比原始欄位更好用。
* 離群值 (Outliers)：特別大/特別小的異常點，會影響線性回歸。通常會「截尾（封頂）」或只標記、或在訓練時排除少量極端點**。**