HW2 — Bike Sharing Demand Prediction

1. CRISP-DM 流程說明

Business Understanding（商業理解）

- 目標：預測每日租借量 `cnt`，以作為決策參考（例如營運安排）。
- Notebook 中以 RMSE 與 R^2 作為評估指標，並在 Evaluation 階段印出其數值。

Data Understanding（資料理解）

- Notebook 會讀取 `day.csv` (`pd.read_csv('day.csv')`) 並印出 `df.shape` 與 `df.head()`，供人工檢視欄位與樣本數。

Data Preparation（資料準備，實作細節）

- 移除欄位：`instant`, `dteday`, `casual`, `registered`（在 code cell 中以 `cols_to_drop` 處理）。
- 特徵處理：
  - 類別欄位（例如 `season`, `yr`, `mnth`, `holiday`, `weekday`, `workingday`, `weathersit`）使用 `OneHotEncoder`。
  - 數值欄位使用 `StandardScaler()`。
  - 使用 `ColumnTransformer` 建立 preprocessor，並透過 `preprocessor.fit_transform(X)` 取得轉換後矩陣與 `feature_names`。

Modeling（建模，實作細節與參數）

- 特徵選擇：
  - 使用 LassoCV：`LassoCV(alphas=np.logspace(-4,2,100), cv=5, random_state=0, max_iter=10000)` 進行擬合。
  - 以 Lasso 得到的係數做絕對值閾值過濾：`selected_mask = np.abs(lasso.coef_) >= 200.0`。
- 最終訓練：建立 `Pipeline(preprocessor, ColumnSelector(selected_mask), LinearRegression())`，並執行 `train_test_split(X, y, test_size=0.2, random_state=42)` 後訓練與預測。

Evaluation（評估，實作的檢查與輸出）

- 指標：計算 `mse`, `rmse`, `r2` 並在 notebook 中印出（數值依執行而定）。
- 殘差檢視：列出測試集上 abs(residual) 前 5 名項目。

- 預測區間（實作方式）：
    - 使用 OLS 理論近似：重建設計矩陣（含截距），計算 RSS、估計 sigma_hat，利用 XtX 的擬逆計算每個測試點的預測標準誤，再以 t 分布取得 95% 區間（程式中以 `np.linalg.pinv` 與 `scipy.stats.t` 實作）。
    - 程式亦嘗試使用 statsmodels 的 `sm.OLS` 與 `get_prediction(...).summary_frame()`；若回傳 `obs_ci_lower`/`obs_ci_upper`，則覆寫為該區間以供繪圖。

Plots

- 散佈圖：Actual vs Predicted（實際 vs. 預測）。
- 排序圖：將測試樣本依實際值排序後繪製 Actual、Predicted 與 95% Prediction Intervals（若 lower/upper 已計算）。

Deployment

- 確認完 `requirements.txt` 中的套件皆已安裝後，執行 `7114056007.ipynb`。

2. 資料集來源與研究脈絡

- 資料來源：於連結中下載 `day.csv` https://www.kaggle.com/code/gauravduttakiit/bike-sharing-multiple-linear-regression/input (https://www.kaggle.com/code/gauravduttakiit/bike-sharing-multiple-linear-regression/input)
- 研究目的：建立可解釋的線性模型以了解影響每日租借量的重要特徵，並提供預測與其區間估計，協助營運決策。

3. 檔案說明

於 https://github.com/Aweeeeei/AIOT-HW2/tree/main (https://github.com/Aweeeeei/AIOT-HW2/tree/main) 可下載所有檔案。

- `day.csv`：原始 dataset（CSV），Notebook 會直接讀取此檔案。請確保它放在同一資料夾以便 notebook 正常執行。
- `7114056007_hw2.ipynb`：包含資料處理、Lasso-based 特徵選擇、LinearRegression 訓練、評估與繪圖步驟（如上所述）。
- `log.txt`：將 Chatgpt 產生的執行檔，透過 Copilot 逐步修改至符合要求的紀錄。
- `requirements.txt`：Notebook 所需之套件。