HW2: Multiple Linear Regression

一. 作業目標：

本次作業延伸自 HW1，目標是讓同學能夠以實際資料集進行「多元線性迴歸 (Multiple Linear Regression)」的完整分析，並遵循 CRISP-DM 流程完成從資料理解、建模到評估的全過程。

二. 作業內容：

1. 資料來源：
   - 1.1. 至 Kaggle 選擇一個具有 10 至 20 個特徵 (features) 的公開資料集。
   - 1.2. 類型不限（資安主題）。
   - 1.3. 請明確標示資料集來源與連結。
2. 分析任務：
   - 2.1. 使用線性迴歸 (Linear Regression) 模型進行預測。
   - 2.2. 可嘗試單純線性迴歸、多元線性迴歸或 Auto Regression。
   - 2.3. 必須執行 **特徵選擇 (Feature Selection)** 與 **模型評估 (Model Evaluation)**。
   - 2.4. 結果部分需包含請提供預測圖(加上信賴區間或預測區間)
3. CRISP-DM 流程說明：
   - Business Understanding
   - Data Understanding
   - Data Preparation
   - Modeling
   - Evaluation
   - Deployment
4. AI 協助要求：
   - 4.1. 所有與 ChatGPT 的對話請以 pdfCrowd 或其他方式**須匯出為 PDF**。
   - 4.2. 請**使用 NotebookLM** 對網路上同主題的解法進行研究，並撰寫一份 100 字以上的摘要，放入報告中。
   - 4.3. 請在報告中明確標示「GPT 輔助內容」與「NotebookLM 摘要」。
5. 繳交內容：
   - 5.1. 主程式：5114050013_hw2.py /.ipynb
   - 5.2. 報告檔： PDF，需包含以下內容：
     - ① 按照 CRISP-DM 說明的分析流程
     - ② GPT 對話過程（pdfCrowd 匯出）
     - ③ NotebookLM 研究摘要
     - ④ 網路上主流或更優解法之比較與說明

   註 1：以上檔案與資料夾，**請壓縮為學號命名的一個zip**（例如：5114050013_hw2.zip）上傳。

   註 2：(optional) 若上傳至 GitHub，或是以colab撰寫，需提供連結，並在 README.md 中整理流程與成果。

三. 評分標準：

1. 文件說明（50%）
   1.1. CRISP-DM 流程完整且邏輯清楚（25%）
   1.2. 包含 GPT 對話與 NotebookLM 摘要（15%）
   1.3. 有明確說明資料集來源與研究脈絡（10%）
2. 結果呈現（50%）
   2.1. 模型正確可執行，具特徵選擇與評估（25%）
   2.2. 結果合理、美觀且具有說服力（15%）
   2.3. 呈現出Kaggle名次(若有) / 預測結果評估(預測圖、評估指標)（10%）

[Datasets Source URL]
https://www.kaggle.com/datasets/atharvasoundankar/global-cybersecurity-threats-2015-2024

[my GitHub website URL]
https://github.com/candice-wu/Cybersecurity_HW_02_Multiple_Linear_Regression.git

[my Streamlit Demo-site URL]
(To be confirmed)

💻 Cybersecurity Threat Financial Loss Prediction

Goal：預測各項資安事件造成的財務損失（Financial Loss in Million $）
Framework：CRISP-DM + Streamlit Deployment

一. 資料集來源與連結

◉ Dataset Source：Kaggle

◉ Subject：
Global Cybersecurity Threats (2015 ~ 2024)

◉ Website link：
https://www.kaggle.com/datasets/atharvasoundankar/global-cybersecurity-threats-2015-2024

◉ Datasets Introduction：

① Scope：
The Global Cybersecurity Threats Dataset (2015 ~ 2024) provides extensive data on cyberattacks, malware types, targeted industries, and affected countries.
It is designed for threat intelligence analysis, cybersecurity trend forecasting, and machine learning model development to enhance global digital security.

② Data Overview - Features｜Target (Please kindly refer to below table)

✳ Target：設定「Financial Loss (in Million $)」作為目標變數 (y)，其他欄位來預測經濟損失，目的：探討「哪些因素會影響資安事件的損失金額」。

✳ Features：除了 Target 的其它 9 個欄位。

Column Name	Data Type	Description	Sample Value
Country	object	發生國家	China, UK, Germany
Year	int	發生年份 (2015–2024)	2019
Attack Type	object	攻擊類型	Ransomware, DDoS, Phishing
Target Industry	object	攻擊目標產業	IT, Finance, Retail
Financial Loss (in Million $)	float	經濟損失 (百萬美金)	80.53
Number of Affected Users	int	受影響用戶數	773169
Attack Source	object	攻擊來源	Hacker Group, Nation-state
Security Vulnerability Type	object	弱點類型	Zero-day, Weak Passwords
Defense Mechanism Used	object	防禦機制	VPN, Firewall, Antivirus
Incident Resolution Time (in Hours)	int	事件處理時間 (小時)	63

二. CRISP-DM 實作架構

流程階段	目標	主要任務	對應方法 / Python 模組
1. 業務理解 (Business Understanding)	明確定義分析目標與應用場景	1. 定義研究問題：「哪些因素最影響資安事件的經濟損失？」 2. 說明資料應用價值與決策意義	N/A
2. 數據理解 (Data Understanding)	了解資料結構與特徵分布	1. 使用 df.info()、df.describe()、df.isnull().sum() 2. 使用 seaborn 與 matplotlib 畫出關鍵變數分布、箱型圖、熱力圖	pandas seaborn matplotlib
3. 數據準備 (Data Preparation)	清理、轉換與特徵工程	1. 缺失值處理、Outlier 檢查 2. 類別特徵轉為 One-Hot Encoding 3. 數值標準化 (StandardScaler) 4. 訓練/測試資料切分 (train_test_split)	pandas sklearn.preprocessing
4. 建模 (Modeling)	建立預測模型	1. 使用 LinearRegression 建立多元線性模型 2. 可嘗試 Auto Regression (若以年份為時間序列) 3. 使用 RFE 進行特徵選擇	sklearn.linear_model sklearn.feature_selection
5. 評估 (Evaluation)	檢驗模型表現與合理性	1. 計算 R^2、RMSE、MAE 2. 繪製「實際 vs 預測」散點圖、殘差圖 3. 繪製含信賴區間的預測圖	sklearn.metrics matplotlib
6. 部署 (Deployment)	展示可執行、具互動性成果	1. 將訓練好的模型儲存 (joblib.dump) 2. 使用 Streamlit 實作互動式預測應用： ① 使用者輸入攻擊屬性 → 即時預測財務損失 ② 顯示信賴區間與圖形化	joblib streamlit

1. 業務理解（Business Understanding）
   - ◉ 目標說明
     - ✳ 探討哪些因素會影響資安事件造成的財務損失。
     - ✳ 建立可預測 Financial Loss 的迴歸模型。
   - ◉ 任務重點
     - ✳ 明確定義 Target：Financial Loss (in Million $)。
     - ✳ 其餘欄位為 features（9 欄），需進行類別與數值型特徵處理。

2. 數據理解（Data Understanding）
   - ◉ 任務
     - ✳ 載入資料集（Global_Cybersecurity_Threats_2015-2024.csv）
     - ✳ 瞭解欄位資訊、缺失值與統計摘要
     - ✳ 視覺化資料分布與關聯性

   - ◉ 程式碼

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# 載入資料集
df = pd.read_csv("Global_Cybersecurity_Threats_2015-2024.csv")
df.head()

# 檢查欄位資訊與缺失值
df.info()
df.isnull().sum()

# 基本統計摘要
df.describe()

# 可視化欄位分布與關聯
plt.figure(figsize=(8,5))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap="coolwarm")
plt.title("Feature Correlation Heatmap")
plt.show()
```

3. 數據準備（Data Preparation）
   - ◉ 任務
     - ✳ 缺失值處理
     - ✳ One-Hot Encoding 類別特徵
     - ✳ 標準化數值欄位
     - ✳ 資料切分（Train/Test）

   - ◉ 程式碼

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# 設定 target 與 features
target = "Financial Loss (in Million $)"
X = df.drop(columns=[target])
y = df[target]
```

```
# 類別欄位轉換
X = pd.get_dummies(X, drop_first=True)

# 數值標準化
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 訓練/測試集切分
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
random_state=42)
```

4. 建模（Modeling）
   - ◉ 任務
     - ✳ 建立模型
       - ① 單純線性迴歸模型
       - ② 多元線性迴歸模型
       - ③ Auto Regression（若以年份為時間序列）
     - ✳ 使用 Recursive Feature Elimination (RFE) 進行特徵選擇
   - ◉ 程式碼

```
from sklearn.linear_model import LinearRegression
from sklearn.feature_selection import RFE

model = LinearRegression()
rfe = RFE(model, n_features_to_select=10)
rfe.fit(X_train, y_train)

selected_features = X.columns[rfe.support_]
selected_features

# 使用選擇後特徵重新訓練模型
model.fit(X_train[:, rfe.support_], y_train)
```

5. 評估（Evaluation）- 模型評估與結果視覺化
   - ◉ 任務
     - ✳ 指標 - 評估模型表現
       - ① R^2
       - ② RMSE
       - ③ MAE
     - ✳ 圖表建議
       - ① 預測 vs 實際損失值散點圖
       - ② 殘差圖

③ 各特徵重要性條狀圖
④ 加上 95% 信賴區間預測圖

◉ 程式碼

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
import numpy as np

y_pred = model.predict(X_test[:, rfe.support_])

r2 = r2_score(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
mae = mean_absolute_error(y_test, y_pred)

print(f"R²: {r2:.3f}, RMSE: {rmse:.3f}, MAE: {mae:.3f}")

# 可視化
plt.figure(figsize=(6,6))
sns.scatterplot(x=y_test, y=y_pred)
plt.xlabel("Actual")
plt.ylabel("Predicted")
plt.title("Actual vs Predicted Financial Loss")
plt.show()
```

6. 部署（Deployment）

◉ 任務

✳ 將訓練好的模型儲存（joblib.dump）
✳ 使用「Streamlit」建立互動式預測應用。
✳ 提供輸入欄位（如：攻擊類型、產業別、受害人數等），即時顯示預測損失金額與信賴區間。

◉ 程式碼

```
import joblib

# 儲存模型
joblib.dump(model, "cyber_risk_model.pkl")
joblib.dump(scaler, "scaler.pkl")

print("✅ 模型與 scaler 已儲存，可於 Streamlit 應用中載入使用。")
```

三. Streamlit 互動頁面設計範例（邏輯結構）

3-1. 以下是建議的頁面結構與互動流程，可直接照此在 .py 或 .ipynb 中實作。

塊	功能說明	Streamlit 元件建議
Header 區	顯示標題與專案說明	st.title() st.markdown()
Sidebar 區	使用者輸入控制	st.sidebar.selectbox() st.sidebar.slider() st.sidebar.number_input()
Main 區（預測結果）	顯示模型預測與評估指標	st.metric() st.write() st.pyplot()
圖表區	視覺化預測結果與信賴區間	matplotlib / plotly
模型資訊區	顯示 R^2、RMSE、MAE 等指標	st.table() 或 st.write()

3-2. Wireframe

```
------------------------------------------------------
💻 Cybersecurity Threat Loss Prediction App
------------------------------------------------------
[ Sidebar ]
Year: [2023]
Attack Type: [Ransomware]
Target Industry: [Finance]
...
------------------------------------------------------
預測損失金額： $85.37 Million
R² = 0.82, RMSE = 6.13
[ 散點圖 / 信賴區間圖 / 殘差圖 ]
------------------------------------------------------
```

Appendix

A. NotebookLM 研究摘要（至少 100 字）

- 搜尋主題：「Multiple Linear Regression for cybersecurity dataset」
- 重點：網路上主流或更優解法之比較與說明
- 摘要：

B. GPT 對話記錄摘要（PDFCrowd 其它方式匯出 PDF）