多平台相機價格爬蟲與市場分析

Multi-Platform Camera Price Scraper and Market Analysis

作者：許宇鈞
系級：機電四甲
學號：D1014223

開課老師：周進華

課程名稱：Python 入門與行銷資料科學課程

開課系所：行銷系

開課學年：114 學年度 第 1 學期

中文摘要

本研究旨在解決攝影器材市場中嚴重的資訊不對稱與跨國定價差異問題。專業相機與鏡頭具備高單價、規格複雜及品牌忠誠度高等特性，消費者在選購時常面臨價格不透明及二手分級標準不一的困境。此外，不同區域市場（如台灣、日本、馬來西亞）受匯率波動與供需結構影響，存在顯著價差。因此，本專案建構一套自動化跨國市場情報系統，旨在協助消費者與行銷決策者快速掌握行情，並識別潛在的跨境套利機會。

在研究實施方面，本研究以 Python 為核心開發語言，運用 Requests 與 BeautifulSoup 模組開發網路爬蟲，針對台灣（PChome 24h、Dpowers）、日本（Fujiya Camera）及馬來西亞（KLDSLR）等電商平台進行異質數據採集。數據處理階段利用 Pandas 進行資料清洗、匯率轉換與去重，並導入 IQR（四分位距）統計法剔除價格異常值。隨後實作特徵工程，運用 Scikit-learn 建立隨機森林（Random Forest）價格預測模型與 K-Means 市場分群模型，最終透過 Streamlit 框架建置互動式視覺化儀表板，提供直觀的數據決策介面。

根據分析結果顯示，攝影器材市場呈現高度寡占特徵，Nikon、Sony 與 Canon 三大品牌主導市場份額，且鏡頭與機身之供應量比例約為 5:1，驗證了「生態系鎖定」之商業策略。價格分析證實日本二手市場在特定中高階機種上，相較於台灣市場具備顯著價格優勢，具備跨境套利之經濟可行性。機器學習模型亦能有效預測市場合理價格，誤差控制在實務操作之容許範圍內。本系統不僅優化個人購買決策，亦可供電商業主作為動態定價之參考。

關鍵字：

市場分析、行銷資料科學、跨境套利、網路爬蟲、機器學習

Abstract

This study addresses the issues of information asymmetry and cross-border pricing disparities within the photography equipment market. High-end cameras and lenses are characterized by high unit prices, complex technical specifications, and varied second-hand grading standards, which often lead to market opacity. Furthermore, significant price gaps exist between regional markets (e.g., Taiwan, Japan, and Malaysia) due to exchange rate fluctuations and supply-demand imbalances. This project develops an automated cross-border market intelligence system to empower consumers and marketers with real-time insights and identify potential arbitrage opportunities.

Utilizing Python as the primary development language, web scrapers were developed using Requests and BeautifulSoup to collect heterogeneous data from major e-commerce platforms, including PChome 24h and Dpowers (Taiwan), Fujiya Camera (Japan), and KLDSLR (Malaysia). Data processing involved cleansing, currency conversion, and deduplication via the Pandas library, with the Interquartile Range (IQR) method applied for outlier detection. Feature engineering was implemented to support machine learning models, specifically a Random Forest Regressor for price prediction and K-Means for market segmentation. Finally, an interactive visualization dashboard was deployed using the Streamlit framework.

The findings reveal a highly oligopolistic market dominated by Nikon, Sony, and Canon. A lens-to-body ratio of approximately 5:1 confirms the prevalence of an "ecosystem lock-in" business model. Price analysis verifies that the Japanese second-hand market offers a significant competitive advantage for specific mid-to-high-end models compared to the Taiwanese market, proving the economic feasibility of cross-border arbitrage. The machine learning models effectively predict fair market values with an acceptable margin of error. This system serves as a valuable tool for optimizing individual purchasing decisions and providing a strategic reference for e-commerce competitive pricing.

Keywords：

Cross-border Arbitrage, Data Analysis, Machine Learning (ML), Python, Web Crawler

目　次

第一章、緒論

1.1 研究動機 Project Motivation

在數位影像技術日新月異的今天，攝影已從專業領域延伸至大眾生活，成為現代人記錄日常的重要方式。然而，對於初學者及預算有限的攝影愛好者而言，攝影器材如數位單眼、微單相機及交換式鏡頭，皆屬於高單價且生命週期較長的電子產品。在進行器材選購時，消費者往往面臨顯著的信息不對稱困境，市場上充斥著新品公司貨、平行輸入的水貨、以及各類品項不一的二手交易資訊，導致初學者難以在紛雜的數據中快速判斷特定型號的合理市場價格。

攝影市場的全球化特徵進一步增加了定價的複雜性。由於主要數位相機品牌如 Sony、Canon、Nikon、Fujifilm 等多數為日本企業，其全球定價策略深受區域匯率如日圓走勢、關稅制度及當地通路補貼政策影響。根據實務觀察，同一款相機機身在台灣、日本及馬來西亞等地的價差有時可達百分之十五至百分之二十五以上。這種顯著的區域定價差異，雖然為跨境採購與套利提供了潛在誘因，但同時也大幅增加了消費者在進行跨國調查時的搜尋成本與決策複雜度。

除了新品市場，攝影器材具有極高的流通性與保值性，二手市場的活躍程度不亞於新品。特別是在制度完善的日本市場，器材品相有著嚴格的分級標準，如中古 S、A、AB、B 等級，這對於追求性價比的攝影玩家具有極大吸引力。然而，各國二手市場的命名規則與分級標準並不統一，加上非結構化的產品標題描述，使得跨平台的價格比對極其困難。

基於上述背景，本研究團隊成員作為攝影愛好者，在日常進行器材升級與訪價的過程中，深感手動刷新不同電商平台如台灣 PChome、日本 Fujiya Camera、馬來西亞 KLDSLR 之低效率與侷限性。為了優化決策流程，本專案致力於透過網路爬蟲與數據科學技術，自動化收集多國多平台的定價資訊。藉由 Python 的高效處理能力，我們能將混亂的標題數據轉化為具備分析價值的結構化特徵，並運用統計方法如 IQR 異常值處理與機器學習模型進行價格預測，進而建立一個客觀的攝影器材價值評估系統。這不僅能協助攝影者找到最優交易，更能從微觀數據中洞察全球相機市場的供需脈動與品牌溢價趨勢。

1.2 研究目標 Project Objectives

本研究旨在構建一個整合跨國市場數據的相機情報系統，透過數據科學手段解決攝影愛好者在選購器材時的資訊不對稱問題。具體研究目標包含以下五大核心維度：

1. 建立自動化跨國多平台數據採集系統

   開發基於 Python 的穩健網頁爬蟲架構，針對台灣（PChome 24h, Dpowers）、日本（Fujiya Camera, Map Camera）及馬來西亞（KLDSLR）等代表性攝影器材電商平台，實現自動化數據抓取。目標是克服不同網站的抗爬蟲機制與異質 HTML 結構，建立一個包含產品名稱、價格、新舊程度及來源鏈結的動態資料庫。

2. 實現異質數據的清洗與精準歸一化

針對採集到的原始數據 Raw Data 進行深層處理，包括：

- 匯率對齊： 自動將日圓 (JPY) 與馬幣 (MYR) 轉換為新台幣 (TWD)，建立統一的價格比較基準。
- 品牌與系列對齊： 運用關鍵字匹配算法，將混亂的產品標題如 Lumix, Nikkor, Zfc 精準歸類至其母品牌 Panasonic, Nikon，並排除藍牙耳機、腳架等無關配件的噪音數據。
- 異常值過濾： 運用統計學中的 IQR 四分位距方法，動態識別並剔除市場極端值，確保數據分析的客觀性。

3. 識別跨境價格差異與套利機會分析

透過多國定價數據的橫向對比，量化特定相機型號如 Nikon Z8, Canon EOS R5 在不同區域間的價差。分析重點在於識別：

- 日本二手市場的價格優勢： 驗證日本市場在特定等級如 AB 級品相下的價格是否顯著低於台灣。
- 市場不效率性： 找出具備高度跨境價差的產品，為消費者提供具備數據支持的最優購買路徑 Optimal Purchasing Path。

4. 構建基於機器學習的攝影器材估價模型

利用處理後的特徵數據如品牌、來源國、新舊程度，進行特徵工程 Label Encoding & One-Hot Encoding，並嘗試建立機器學習預測模型。其目標是：

- 預測合理市場價格： 給予特定參數，系統能預測該相機在目前的市場公允價值。
- 特徵重要性分析： 探討「新舊程度」與「品牌溢價」對最終價格的影響權重，從數據中挖掘驅動價格變動的核心因素。

5. 開發一體化監控與可視化決策平台

運用 Streamlit 框架開發互動式前端界面，將複雜的後端數據處理過程包裝成易於操作的儀表板。讓使用者無需具備編程背景，即可透過篩選、排序與搜尋功能，即時掌握全球相機市場動態，達成從「數據採集」到「決策支持」的完整閉環。

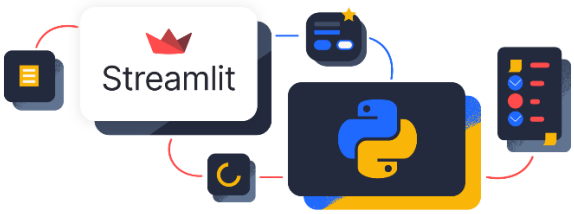
1.3 開發工具與環境 Tech Stack

本專案採用 Python 作為核心開發語言，構建了一套從數據採集、清洗、統計過濾到機器學習預處理與前端可視化的完整流水線。以下為本研究使用之主要技術工具及其具體實作說明：

1. 核心程式語言

Python 3.12.10 作為數據科學領域的主流語言，Python 豐富的第三方函式庫生態系統為本專案提供了強大的技術支持，確保了代碼的可維護性與數據處理的高效性。

2. 前端互動界面



Streamlit 是一款專為機器學習與數據科學設計的開源 Python 框架。

我們利用 Streamlit 構建了互動式儀表板 Dashboard，將後端的爬蟲邏輯封裝為可點選的按鈕與滑桿。使用者可以透過界面選擇數據來源如 Fujiya Camera 或 PChome、設定抓取頁數，並即時查看清洗後的數據表格與排序結果，實現了數據工具的平民化操作。

3. 網路爬蟲數據採集

- 運用 Requests 程式庫負責發送 HTTP 請求。在本專案中，我們實作了進階的 Session 會話管理與 User-Agent 偽裝技術，藉此模擬真實瀏覽器的訪問行為，以有效繞過各大電商平台嚴格的反爬蟲偵測機制。
- 搭配 BeautifulSoup (bs4) 進行 HTML 文件結構解析。我們透過 CSS 選擇器與正則表達式 Regex 的複合運用，精準定位並擷取產品標題、價格及原始網址，克服了因不同平台間異質網頁結構所產生的數據提取難題。

4. 數據處理與數值計算

本研究的核心在於如何將來自不同國家、不同平台且結構混亂的原始數據（Raw Data），轉化為具備商業分析價值的結構化資訊。我們主要依賴 Python 生態系中強大的數據處理庫來實現此目標

Pandas 作為本專案的數據中樞，Pandas 承擔了最繁重的 ETL（Extract, Transform, Load）任務。具體應用包括：

- 異質數據整合 Data Merging。 將來自 PChome（台灣）、Fujiya Camera（日本）與 KLDSLR（馬來西亞）格式迥異的 CSV 檔案，統一映射至標準化的欄位架構如：品牌、型號、價格、新舊程度。
- 智能去重 Intelligent Deduplication。 實作基於產品網址 (URL)與爬取時間戳 (Timestamp)的去重邏輯，確保資料庫中僅保留該產品最新的價格資訊，避免歷史數據干擾分析結果。
- 品牌歸一化 Normalization。運用字串處理技術，將混亂的品牌標識如 "NIKON", "Nikon", "尼康"標準化為單一格式，並自動從複雜的產品標題中提取關鍵規格如將 "Zfc Body" 自動歸類為 "Nikon Z Series"。
- 匯率轉換 Currency Conversion。建立動態匯率計算欄位，將日圓 (JPY) 與馬幣 (MYR) 即時轉換為新台幣 (TWD)，建立統一的價格比較基準。

NumPy 負責高效數值運算，配合 Pandas 進行底層的高效能數值計算，特別是在處理大量價格數據的統計分析時發揮關鍵作用。

統計過濾支持為後續的 IQR 四分位距異常值偵測算法提供底層的數值運算支持，快速計算出價格分佈的四分位數（Q1, Q3），以科學方式界定並剔除極端價格如低價配件或天價電影鏡頭。

5. 統計過濾與機器學習 Scikit-learn

在確保數據結構化後，本研究進一步運用統計學方法與機器學習預處理技術，提升數據的分析品質與模型相容性：

- IQR 異常值過濾 透過數據的四分位數 Quantiles 運算，自動識別並剔除價格分佈中的離群值。此步驟能有效過濾掉價格極低的非相機類配件如鏡頭蓋、清潔組或極端高價的特殊器材，確保後續分析能集中於主流攝影器材市場。

特徵工程 Encoding 運用 Scikit-learn 與 Pandas 實作類別特徵轉換，將文字資訊轉化為模型可讀取的數值矩陣：

- 標籤編碼 Label Encoding 應用於具備順序意義的新舊程度 Condition 特徵，將其轉化為數值階層，保留數據中的邏輯順序。

- 獨熱編碼 One-Hot Encoding 應用於品牌 (Brand) 與來源 (Source) 等名義類別，透過擴展特徵空間避免模型產生錯誤的數值大小排序偏差，為後續的價格預測分析奠定精準的特徵基礎。

第二章、數據收集 Data Collection

2.1 數據源說明 Data Sources

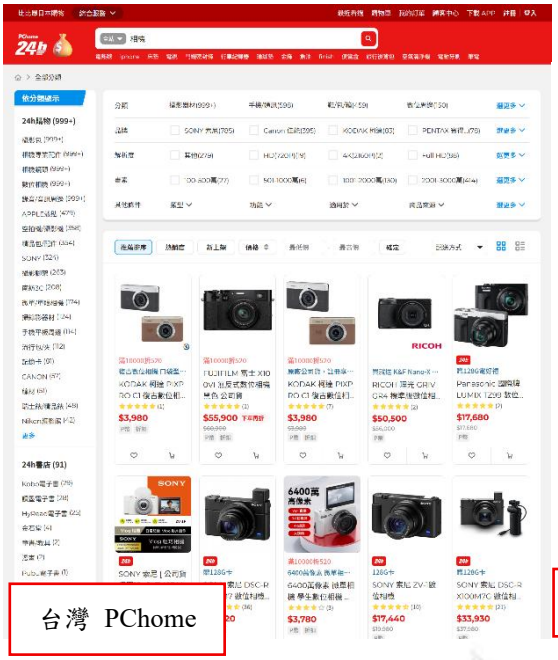
本研究針對全球攝影器材市場之核心節點，策略性地選取了台灣、日本及馬來西亞作為數據採集來源。數據源的選擇並非隨機抽樣，而是基於各平台在市場中的角色定位、定價權威性以及品相分級制度的完備性進行考量。

1. 跨國數據源分佈 Regional Distribution

本專案構建的資料庫涵蓋了以下三大核心市場：

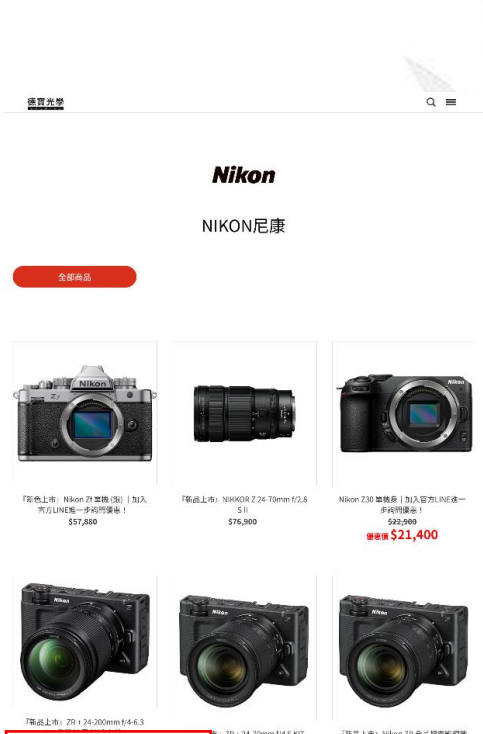
- 台灣市場 Local Benchmark 作為本研究的基準點，選擇 PChome 24h 以獲取最具普適性的公司貨定價；輔以專業相機通路 Dpowers 與 Big Camera，擷取包含複雜贈品組、通路限定促銷及即時庫存變動的數據，反映本地市場的真實競爭現況。
- 日本市場 Primary Source 日本作為全球光學產業的發源地，其定價具有全球指標意義。本專案採集 Fujiya Camera 之數據，除了獲取較低的原始定價資訊外，更重要的是擷取其精準的中古品相分級數據，這對於後續研究折舊率與價格關聯至關重要。
- 馬來西亞市場 Regional Comparison KLDSLR 作為東南亞區域的專業代表，能提供不同匯率波動與區域補貼政策下的定價視角，是跨境分析中不可或缺的對照組。

跨國電商平台官方網頁界面概覽:

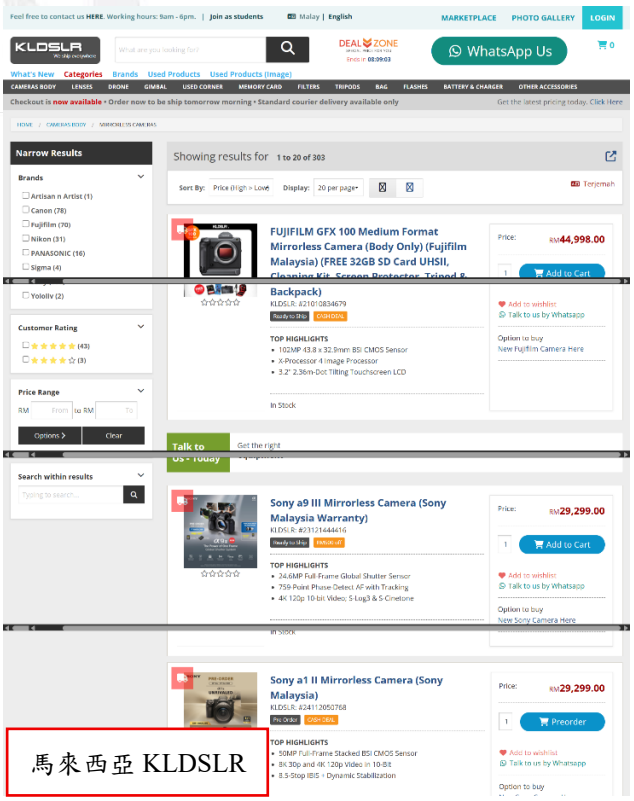


台灣 PChome

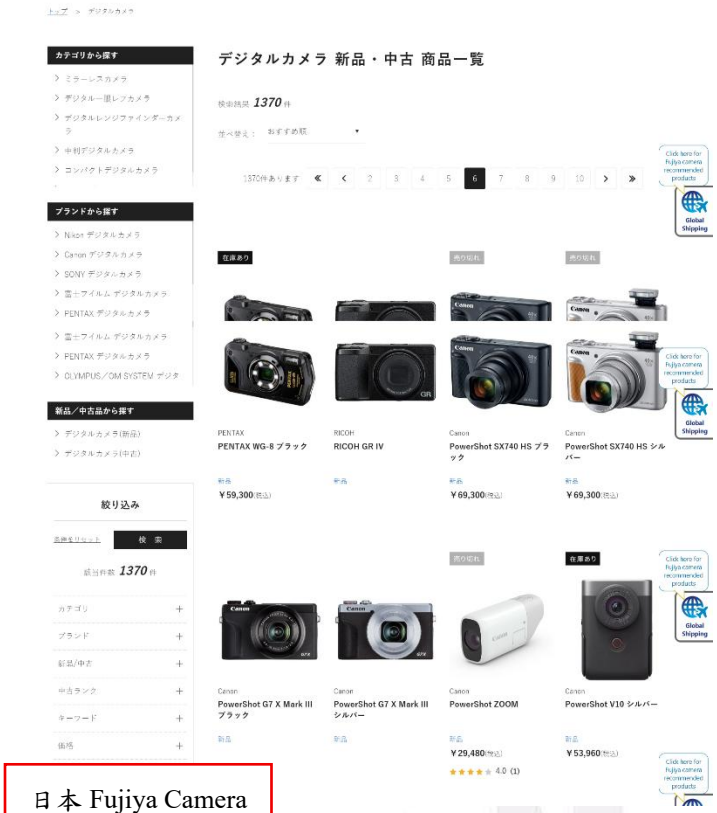
台灣 Big Camera



台灣 Dpowers



馬來西亞 KLDSLR



日本 Fujiya Camera

2. 各平台數據特徵分析 Platform Characteristics

為了確保分析的全面性，本研究針對各採集平台的網頁架構、數據內容及採集難點進行了深入剖析。透過對不同來源之數據特徵的理解，本專案得以制定更精準的解析策略。下表彙整了各主要數據源的特性與技術挑戰：

平台名稱	**區域**	**數據特徵**	**採集挑戰與解決方案**
PChome 24h	台灣	透過內部 API 回傳數據，格式相對整齊，包含大量官方公司貨定價。	存在大量重複標題與廣告欄位，需處理 API 翻頁限制與資料去重 Deduplication 邏輯。
Dpowers	台灣	產品標題包含豐富的**禮包**、**升級套裝**資訊，是分析產品組合價值的關鍵來源。	價格標籤採用複雜的 HTML 嵌套結構，需處理**優惠價**與**建議售價**之間的邏輯切換與提取。

Fujiya Camera	日本	包含全球最嚴格的二手品相分級（S/A/AB/B），商品名稱與描述為全日文。	網頁結構具備動態生成特性，且存在嚴格的區域 IP 存取限制與抗爬機制，需優化請求標頭 Headers 與 Session 管理。
KLDSLR	馬來西亞	定價以馬幣 (MYR) 為主，標題為英文，包含豐富的當地保固資訊。	網頁結構較為老舊，且 HTML 標籤不規範，需手動處理大量非標準的 CSS 選擇器以定位數據。
Big Camera	台灣	提供詳細的產品規格與即時促銷價，適合用於專業通路間的價差比對。	列表頁數據顆粒度不足，需實作**兩階段爬取策略**，由列表頁引導至詳情頁以獲取完整型號名稱。

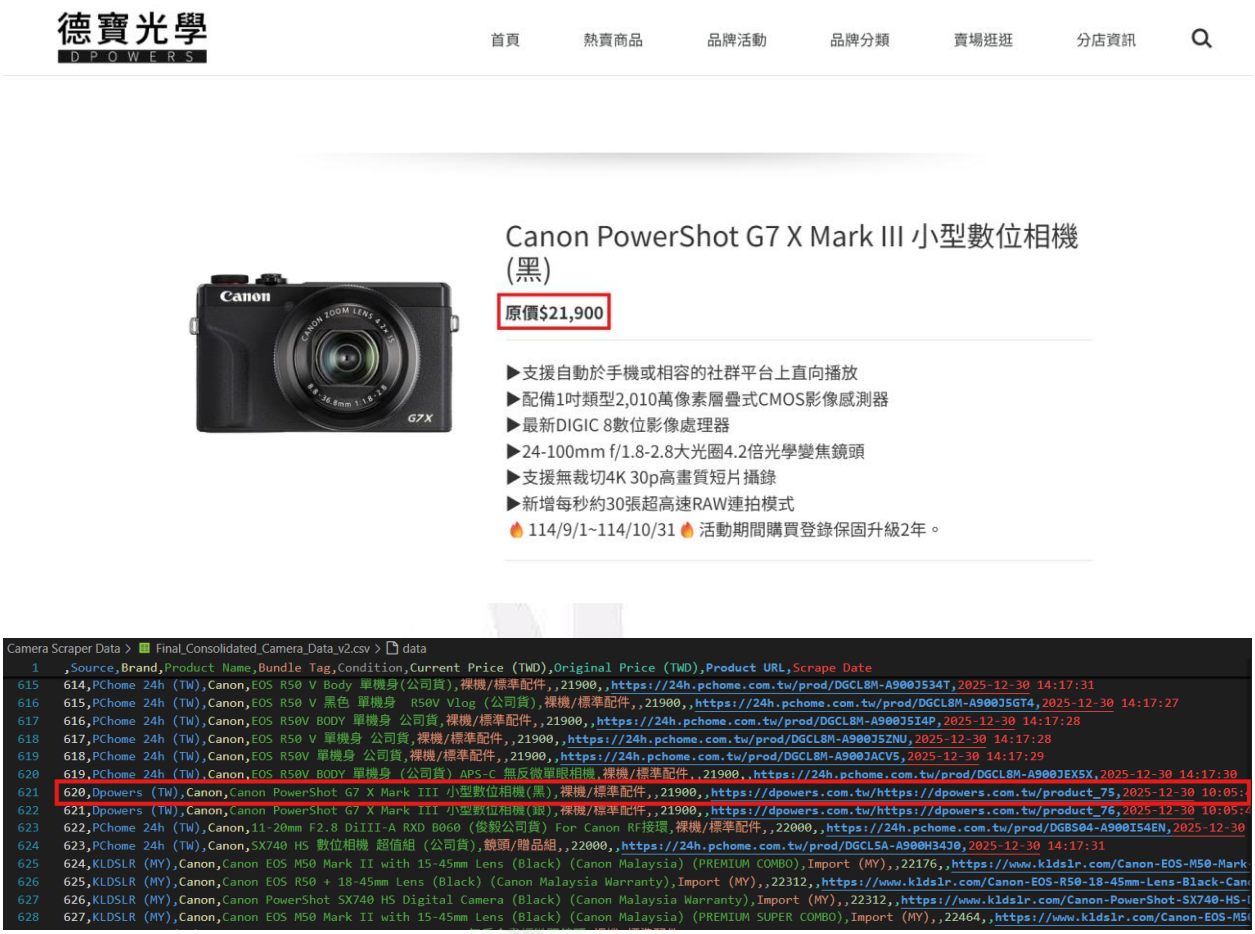
3. 數據欄位定義與顆粒度 Data Granularity

本研究採集的數據具備高度的顆粒度，確保了後續分析的深度。核心欄位說明如下：

- 產品全稱 (Product Name) 包含品牌、機身/鏡頭型號、焦距、光圈等關鍵規格資訊。
- 即時價格 (Current Price) 擷取平台當下的最終優惠價，而非僅抓取原價，以反映市場真實成交行情。
- 新舊程度 (Condition) 標記該項商品為新品 (New) 或不同等級之二手品 (Used)。
- 套裝標籤 (Bundle Tag) 透過算法識別該商品為單機身 (Body Only)、鏡頭組 (Kit Set) 或豪華配件組。
- 原始鏈結 (Product URL) 保留數據來源追蹤，確保數據的可驗證性與時效性。

網頁結構與數據標籤對照圖

紅框圈為網頁上的價格，對應至爬蟲程式碼中定位的 HTML 標籤。



4. 數據採集統計概況 Collection Overview

本專案透過自動化爬蟲程式，初步共計採集原始數據逾 7,000 筆。經由後續的數據處理階段進行精準過濾與去重後，形成了一個跨越三個國家、五大平台、包含九大主流相機品牌的高品質攝影器材資料庫，為後續的跨境價格套利分析奠定了堅實的基礎。

2.2 爬蟲技術方法論 Scraping Methodology

本專案的數據採集流程並非單一的讀取動作，而是一個包含身分偽裝、精準定位、自動翻頁、兩階段強化的複雜工作流 Work Flow。我們針對不同網站的架構

HTML 標籤與 Application Programming Interface, API 接口，開發了專門的爬蟲腳本 scrapers.py。

1. 身分偽裝與連線管理 Emulation & Session Management

網頁伺服器通常會偵測並阻擋自動化程式 Bot。為了確保採集過程不被中斷，我們實作了以下技術：

- User-Agent 偽裝。透過配置 HEADERS 字典，將 Python 的請求偽裝成普通的 Chrome 瀏覽器，避免被伺服器識別為惡意爬蟲。
- 會話持久化 Session。針對日本 Fujiya Camera 與 Big Camera 等網站，我們使用 `requests.Session()` 維持連線狀態，這能自動處理 Cookie，讓伺服器認為是同一位使用者在進行多頁面瀏覽，大幅提升抓取的穩定性。

2. 異質架構下的數據提取策略 Extraction Strategy

由於各平台的技術背景不同，本專案採用了「雙軌制」採集策略：

- HTML 解析軌 Dpowers, Big Camera, Fujiya, KLDSLR。使用 BeautifulSoup (bs4) 搭配 CSS 選擇器。這就像是給程式一張地圖，告訴它產品名稱藏在 `<div class="gTxt">` 的標籤裡，而價格則嵌套在 `<h3>` 下的 `<span>` 中。我們針對每個網站的標籤結構（Document Object Model, DOM Tree）進行了細緻的手動映射。
- API 解析軌 PChome 24h。PChome 採用動態載入技術，傳統解析 HTML 無法取得完整資料。我們透過分析網頁後端請求 XMLHttpRequest, XHR，直接調用其內部 JSON API。這種方式不需要解析網頁原始碼，而是直接獲取結構化的 JSON 數據，數據準確度與採集速度皆最高。

3. 兩階段數據強化與自動翻頁 Two-Stage Enrichment & Pagination

針對 Big Camera 等平台，我們實作了更高級的深層爬取邏輯：

- 自動偵測頁數。程式會先讀取分頁標籤 Pagination，自動判定該品牌共有多少頁，動態生成爬取清單，無需手動設定結束點。
- 兩階段強化 Two-Stage Scrape。在 `run_bigcamera_scraper` 中，第一階段先從列表頁（List Page）抓取所有產品的網址；第二階段則透過 `scrape_bigcamera_details` 函數，點擊進去每一個產品的詳情頁（Detail

Page）。這樣做是因為列表頁通常會縮減產品名稱，只有進入詳情頁才能獲取如 Nikon Z8 Body (Company Warranty) 這種完整的型號資訊。

4. 正則表達式與數據純化 Data Cleaning via Regex

網頁上的價格通常包含貨幣符號、千分位逗號如 NT$24,900 或額外文字。為了讓這些資料能被後續的 Pandas 進行計算，我們在爬蟲層級就使用了 正則表達式 Regular Expression, Regex：

- 技術實作。 使用 `re.sub(r'[^\d.]', '', price_text)`。這行代碼像是一把手術刀，能精準移除所有非數字的字符，確保輸出的數據是純粹的數值格式，避免後端數據處理時發生錯誤。

5. 爬蟲禮儀與穩定性控制 Politeness & Robustness

為了做一個有禮貌的爬蟲並防止 IP 被封鎖，我們實作了防禦性機制：

- 延遲請求 Polite Scraping。 使用 `time.sleep(1)`，在每次翻頁或抓取詳情頁時強制暫停 1 到 1.5 秒，模擬真實人類的閱讀節奏，降低伺服器負擔。
- 異常處理 Try-Except。針對網路斷線或網頁結構突然變更的情況，實作了完善的錯誤捕獲機制，確保程式不會因為單一筆資料錯誤而全盤崩潰，並能記錄下發生錯誤的 URL 以供後續除錯。

第三章、數據處理 Data Processing

3.1 數據清洗 Data Cleaning

在數據科學的流程中，原始數據 Raw Data 通常伴隨著大量的雜訊與不一致性。為了確保後續分析與機器學習模型的準確性，本專案在 process_data.py 中實作了嚴謹的數據清洗流水線。本階段的目標是將來自台灣、日本、馬來西亞等多個平台的異質數據，轉化為標準化且可計算的格式。

1. 異質數據整合與結構標準化 Data Integration & Standardization

由於不同平台的 CSV 導出格式不盡相同例如有些包含索引欄位，有些則無，我們採取了以下步驟：

- 處理冗餘索引。 在載入資料時，程式會自動偵測並移除 CSV 中常見的 Unnamed 虛擬索引欄位。這在資料處理中是非常重要的一步，能避免這些無意義的數值干擾矩陣運算。
- 統一欄位架構。 透過建立 core_cols 清單，強制要求所有來源數據必須具備來源（Source）、品牌（Brand）、價格、日期等核心特徵。若原始數據缺失特定欄位，則自動補入 N/A 值，確保整合後的 DataFrame 具備一致的維度。

2. 資料類型轉換與格式化 Data Type Conversion & Formatting

爬蟲抓取到的價格與日期通常是字串格式，例如"NT$ 24,900" 或 "2025/12/30"，這在程式邏輯中無法進行數學運算或排序。

數值化轉換 Type Casting。運用 pd.to_numeric 配合正則表達式，剔除價格中的貨幣符號與逗號，並將結果強制轉換為浮點數 Float。若遇到無效值，則以 NaN 標記並預設補 0，確保計算過程中不會因為資料型態錯誤而崩潰。

時間序列處理。 將抓取時間統一轉換為標準的 datetime 格式。這對於後續判斷資料的時效性至關重要，也為未來的價格趨勢分析奠定了基礎。

3. 基於時效性的智慧去重邏輯 Intelligent Deduplication

在多次爬取的過程中，同一件產品可能會出現在不同的 CSV 檔案中。簡單的去重會隨機刪除資料，但本專案實作了具備邏輯優先級的去重方式：

- 優先級排序： 程式會先根據產品網址 (Product URL)與爬取日期進行降冪排序。
- 保留最新狀態： 使用 drop_duplicates 函數，僅保留同一網址中「日期最晚」的那一筆資料。這確保了當同一台相機在不同日期有價格變動時，我們的資料庫始終反映的是目前最新的市場行情。

4. 初步噪音與低價值資料過濾 Preliminary Noise Filtering

在相機電商平台中，常混入大量的廉價配件如鏡頭蓋、清潔紙、螢幕貼，這些數據會嚴重拉低平均單價。

設定價格門檻。 根據實務經驗，本專案設定了 2,000 TWD 的初始過濾門檻。這是一個非常有效的統計策略，能幫助我們快速剔除約 30% 的非器材類雜訊，讓分析重心集中在真正的相機機身與鏡頭上。

5. 進階關鍵字邏輯過濾 Advanced Keyword Filtering

除了價格門檻外，最核心的挑戰在於如何從產品名稱中辨識出相機/鏡頭與高價配件如耳機、專業腳架。本研究實作了雙向關鍵字過濾機制：

- 包含與排除邏輯。程式必須同時滿足包含品牌核心關鍵字 Inclusion Keywords 且不含任何配件排除詞 Exclusion Keywords 兩個條件才予以保留。例如，若標題同時出現 Sony A7C 與 Case，系統會自動判定其為相機包而非相機本體。

如圖所示，本研究透過集合論的邏輯，定義了包含詞集合（A）與排除詞集合（B）。唯有落於 A 集合且不與 B 集合重疊的數據，才會被視為純淨數據存入最終資料庫中。

3.2 特徵工程 Feature Engineering

本專案實作了從語意提取到數值向量化的完整特徵轉換流程，旨在將混亂的電商資訊提煉為具備預測能力的變數。

數據特徵演化流水線示意圖。本專案透過四個關鍵階段，成功將非結構化的相機產品標題轉化為機器學習模型可讀取的數值向量矩陣。

1. 產品屬性語意提取與標籤化 Bundle Tagging

由於產品名稱通常包含豐富的銷售資訊，我們開發了一套基於正則表達式 Regex 的自動化標籤系統。

- 組合類型識別 透過掃描產品名稱中的關鍵字如**單機身**、**鏡頭組**或 **Kit**，系統能自動判斷該品項屬於純主機或是包含鏡頭的套裝。
- 贈品與禮包偵測 針對台灣電商常見的促銷手段，程式會搜尋**豪華禮包**或**好禮**等詞彙，並將其歸類為高附加價值的套裝產品。這項特徵能有效解釋為什麼同樣型號的相機在同一平台上會出現顯著的價差。

2. 數值化特徵轉換 Numerical Transformation

機器學習模型無法直接運算文字資料，因此我們實作了兩類核心的編碼技術。

- 新舊程度標籤編碼 Label Encoding 考慮到產品狀態具備明顯的順序意義，我們將二手 (Used) 與新品 (New) 分別對應至數值 0 與 1。這種處理方式能保留數據中的等級關係，協助模型理解產品價值的折舊邏輯。
- 類別特徵獨熱編碼 One-Hot Encoding 對於品牌與來源平台等不具備順序關係的類別，我們採用了擴展維度的方式。例如將單一的品牌欄位展開為多個二元欄位如 Brand_Sony, Brand_Nikon，這能避免模型錯誤地將品牌名稱誤判為具備大小關係的連續數值。

3. 跨國匯率轉換與價值校準 Currency Standardization

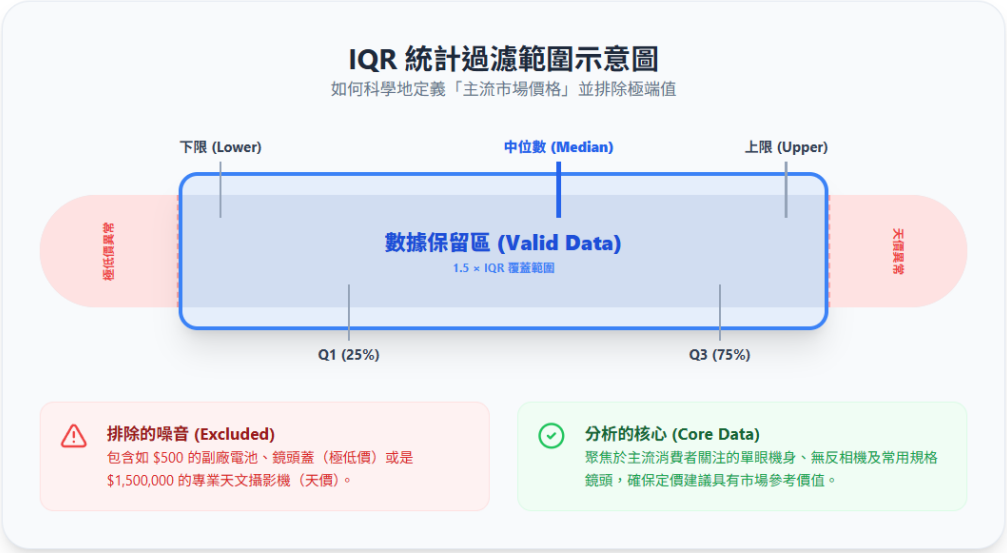
針對不同國家的數據源，我們實作了價值平衡機制。

- 統一貨幣基準 將來自日本的日圓與馬來西亞的馬幣轉換為統一的新台幣基準。這不僅是數據清洗的一部分，更是特徵工程中確保數值尺度一致性的關鍵步驟。

3.3 異常值處理 Outlier Detection

在市場數據分析中，異常值是指那些顯著偏離整體分佈的極端觀測值。本專案透過統計學中的四分位距法 Interquartile Range, IQR 來實作自動化異常值偵測，確保最終分析的數據能代表主流相機市場的行情。

異常值過濾示意圖 Outlier Range Illustration:



基於 IQR 的異常值過濾模型。透過統計學的圍籬 Fences 概念，自動識別並排除極端價格，保留中間最具市場代表性的主流數據。

1. 異常值對市場分析的負面影響

若直接使用未經處理的原始數據，市場分析將面臨兩大風險。首先是「天價噪音」，例如某些價值數百萬台幣的專業電影攝影機或收藏級鏡頭，會大幅拉高平均價格，導致誤以為市場消費能力極高。其次是極低價噪音，例如誤抓取的二手配件或機身蓋，這會稀釋數據品質。透過 process_data.py 中的統計過濾，我們能科學化地界定合理價格區間。

程式碼實作片段 Code Snapshot:

```
# 計算第一四分位數 (Q1) 與第三四分位數 (Q3)
Q1 = df_final['Current Price (TWD)'].quantile(0.25)
Q3 = df_final['Current Price (TWD)'].quantile(0.75)
IQR = Q3 - Q1

# 定義異常值邊界門檻 (1.5 倍準則)
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# 執行過濾：僅保留位於 lower_bound 與 upper_bound 之間的數據
df_final = df_final[
    (df_final['Current Price (TWD)'] >= lower_bound) &
    (df_final['Current Price (TWD)'] <= upper_bound)
]
```

2. IQR 四分位距法的運算邏輯

本研究採用業界公認的 1.5 倍 IQR 準則來建立數據防護欄。其處理流程包含以下三個關鍵步驟。

計算分位數與區間 (Quantiles) 我們首先計算數據的第 25 百分位數 (Q1) 與第 75 百分位數 (Q3)。這代表了市場中最中間 50%的產品價格範圍。兩者之差即為 IQR 四分位距，象徵著市場價格的集中趨勢。

建立邊界門檻 (The Fences) 利用公式 Q1 - 1.5 * IQR 建立下限，以及 Q3 + 1.5 * IQR 建立上限。這組邊界就像是數據的圍籬，任何落在圍籬之外的資料點都會被標記為異常。

數據過濾與精煉 最後透過程式邏輯將位於區間外的資料剔除。這不僅是數學上的清理，更是行銷上的去雜質，讓後續的價格預測模型能建立在真實且具備代表性的樣本之上。

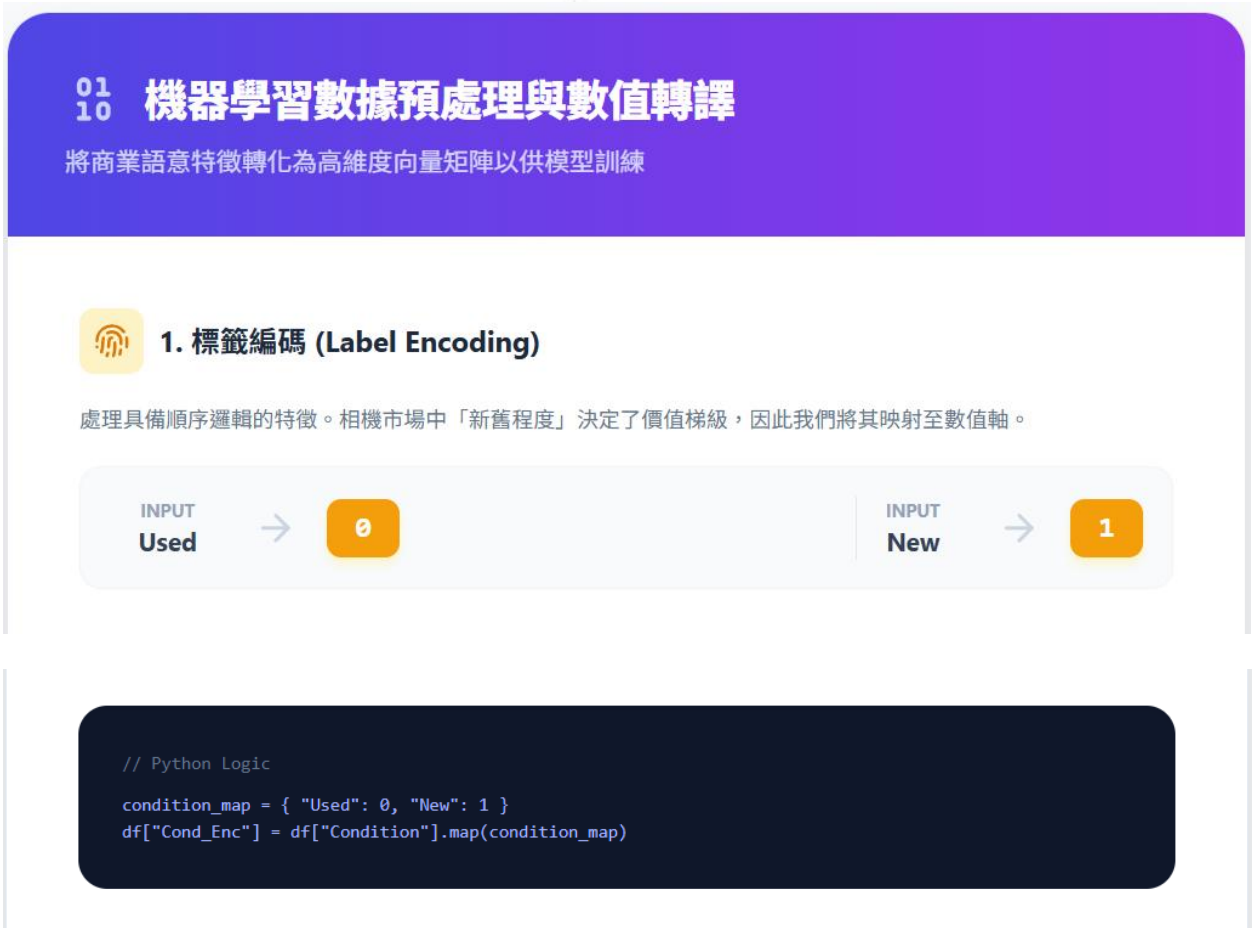
```
(venv) PS C:\Users\ASUS\Desktop\camera_scraper_project> & C:/Users/ASUS/Desktop/camera_scraper_project/venv/Scripts/python.exe "c:/Users/ASUS/Desktop/camera_scraper_project/Camera Scraper Data/process_data.py"
Rows after removing items < $2000: 6251

📊 IQR Filtering Stats:
   Q1 (25%): $6,990
   Q3 (75%): $45,012
   IQR: $38,022
   Lower Bound: $-50,043
   Upper Bound: $102,045
   Removed 355 outliers based on IQR.
   Remaining Rows: 5896

✅ Success! Data processed and saved to:
C:\Users\ASUS\Desktop\camera_scraper_project\Camera Scraper Data\Final_Consolidated_Camera_Data.csv
```

process_data.py 執行統計紀錄。 截圖展示了程式自動運算出的價格分佈數據，

包含第一四分位數 (Q1) 與第三四分位數 (Q3)，並精準標示出過濾後的剩餘筆數，證明了異常值處理的自動化與科學化。

3.3 機器學習預處理 ML Pre-processing

在完成數據清洗與異常值剔除後，我們獲得了一組結構完整且具備商業邏輯的**純淨數據**。然而，這些數據大多仍以文字 String 形式存在如品牌名稱、新舊狀態，這對於基於數學運算的機器學習演算法如隨機森林或線性回歸而言是無法直接理解的。因此，本階段的核心任務是將這些**語意特徵**轉譯為模型可運算的**數值向量**，為後續的價格預測模型鋪路。



機器學習數據預處理與數值轉譯示意圖。本研究詳細展示了如何將商業語意如相機品牌、新舊狀態轉化為數值矩陣。首先，針對具備順序意義的狀態實作標籤編碼 Label Encoding。

其次，針對無順序之品牌特徵實作獨熱編碼 One-Hot Encoding。最終產出的數值矩陣 ML-Ready Dataset 即為機器學習模型進行價格預測的運算基礎。

1. 順序性特徵的數值映射 Ordinal Encoding Strategy

在攝影器材市場中，產品狀態（Condition）是一個具備強烈等級關係的變數。新品的價值必然高於二手品，而二手品又優於故障品。為了讓模型捕捉到這種隱含的價值階層，我們採用了標籤編碼技術。具體而言，我們將 **New** 映射為數值 1，將 **Used** 映射為數值 0。這種二元化的數值處理，不僅簡化了模型的運算複雜度，更明確地向演算法傳達了**新舊程度**與**價格**之間的正相關邏輯。

2. 名義特徵的空間擴展 High-Dimensional Transformation

與產品狀態不同，品牌（Brand）與來源平台（Source）屬於名義變數，彼此之間並無數學上的大小之分例如：Nikon 並不大於或小於 Canon。若強行使用 1, 2, 3 進行編碼，模型可能會錯誤地學習到不存在的順序關係，導致預測失準。

為此，我們實作了獨熱編碼技術，將單一的類別欄位炸開為多個獨立的二元特徵欄位。例如，原本的 Brand 欄位被擴展為 Brand_Sony、Brand_Nikon、Brand_Canon 等多個新欄位。若某筆數據為 Sony 相機，則 Brand_Sony 設為 1，其餘品牌欄位設為 0。這種將特徵空間從低維度文字擴展至高維度稀疏矩陣 Sparse Matrix 的策略，讓模型能夠獨立評估每一個品牌或平台對價格的邊際貢獻 Marginal Contribution，進而量化出所謂的品牌溢價。

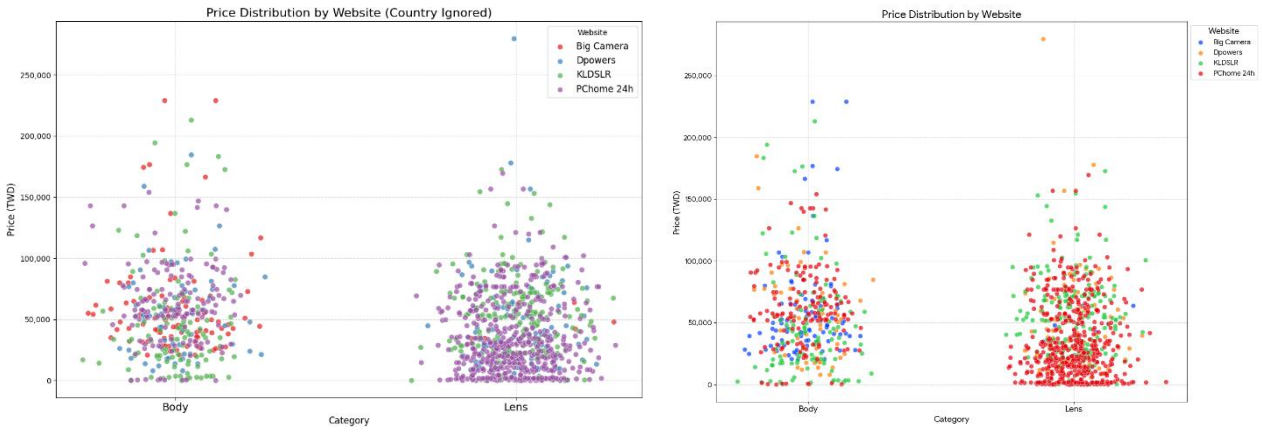
3. 雙軌數據輸出架構 Dual-Pipeline Output

考慮到本專案同時服務於**人類決策**與**機器運算**兩種場景，我們在預處理的最後階段設計了資料分流機制。系統最終會生成兩份獨立的資料集：一份保留了原始文字描述，供市場分析師在 Streamlit 儀表板上進行直觀的價格比對；另一份則完全由 0 與 1 組成的數值矩陣，作為後續訓練價格預測模型的標準輸入 Training Input。這種架構設計確保了數據處理流程的靈活性，既滿足了可解釋性需求，也兼顧了運算效能。

第四章、數據分析與發現 Data Analysis & Insight

4.1 描述性統計 Descriptive Analysis

本章節基於 Python Pandas 進行數據清洗後的資料集（共計約 2,000 筆有效數據），利用 Matplotlib 與 Seaborn 進行視覺化分析，旨在探討相機市場的產品分佈、品牌市佔率以及價格結構的基礎特徵。



平台來源與初步價格分佈:

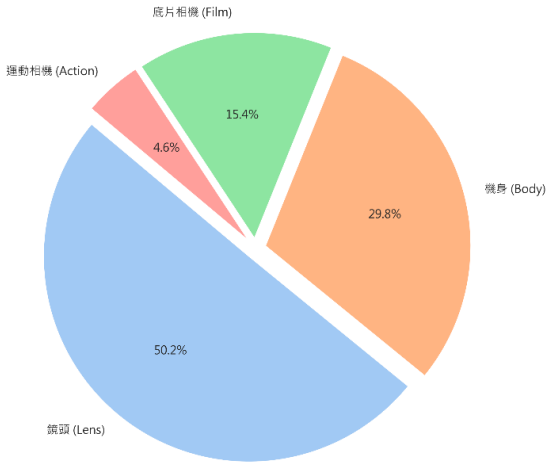
1. 器材類別分佈 Equipment Distribution

透過關鍵字篩選 Keyword Filtering 演算法，本研究將市場上的產品精確劃分為四大核心類別：鏡頭（Lens）、機身（Body）、運動相機（Action Camera）與底片相機（Film Camera）。

市場結構分析。 如圖 4.1 所示，市場上**鏡頭**的品項數量（1,628 項）遠高於**機身**（334 項），兩者比例約為 5:1。此一數據顯著反映了攝影市場的**剃刀與刀片**商業模式 Razor and Blades Model：消費者通常僅持有一至兩台機身，但為了應對人像、風景、微距等不同拍攝場景，會持續購入多顆不同焦段的鏡頭。此外，鏡頭的光學壽命長於機身電子壽命，導致二手市場中的鏡頭流通量與新品種類遠比機身豐富。

數據清洗成效。在預處理階段，我們成功識別並移除了 100 多筆屬於耗材性質的底片膠捲 Film Rolls 與低價配件。若未剔除此類數據，將導致平均價格 Mean Price 被嚴重拉低，進而影響對相機市場真實價值的判斷。

圖 4.1：市場產品類別分佈 (Equipment Distribution)



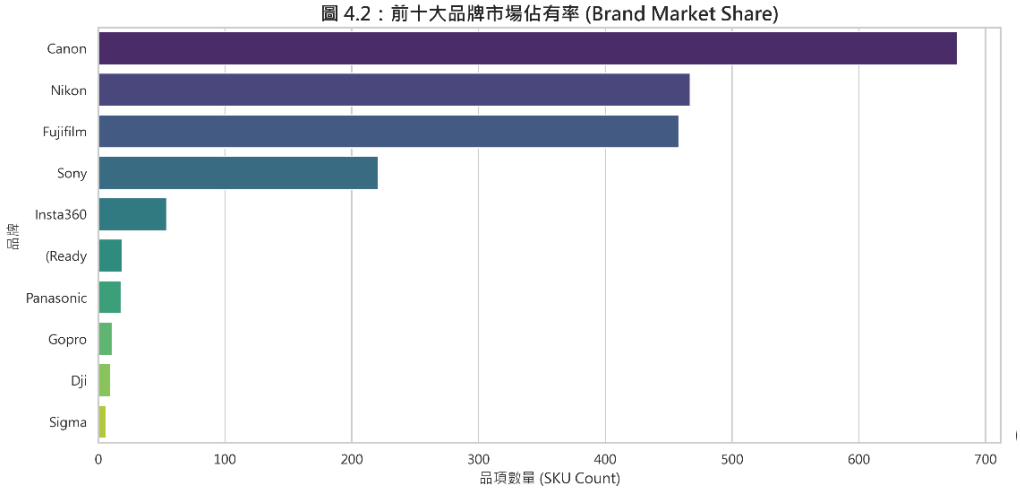
2. 品牌市佔率與通路策略 Brand Market Share

本研究進一步分析了各品牌在採集資料庫中的出現頻率，以評估其在電商通路中的活躍程度。

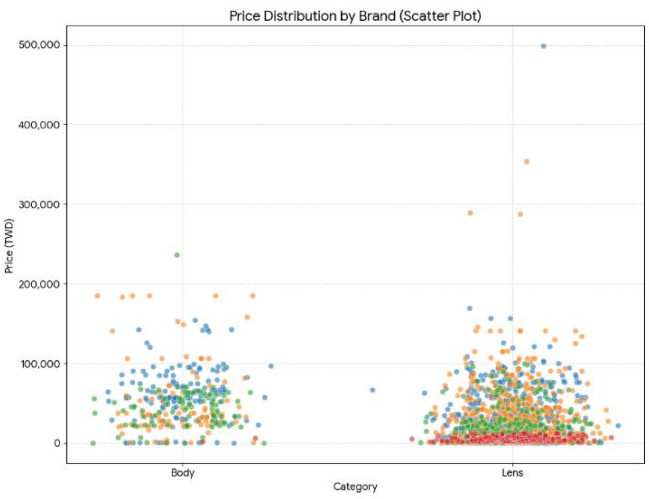
市場寡占趨勢。 如圖 4.2 所示，相機市場呈現高度集中的寡占態勢。Nikon、Canon 與 Sony 三大傳統光學巨頭佔據了絕大多數的市場份額，顯示品牌護城河在攝影領域極為顯著。

通路鋪貨策略。值得注意的是，Nikon 在本資料集中主要數據來源為 PChome 24h 的品項數量位居第一。這可能暗示了特定的通路策略：Nikon 可能與台灣大型電商平台如 PChome 建立了更深度的經銷合作關係，或採取了更積極的線上鋪貨策略 SKU Proliferation，以多樣化的鏡頭組合包 Bundle Deals 來搶佔電商版

圖 4.2：前十大品牌市場佔有率 (Brand Market Share)



面。相對而言，其他品牌可能在實體通路或特定專業網站如 Fujiya Camera 有較高的比重。

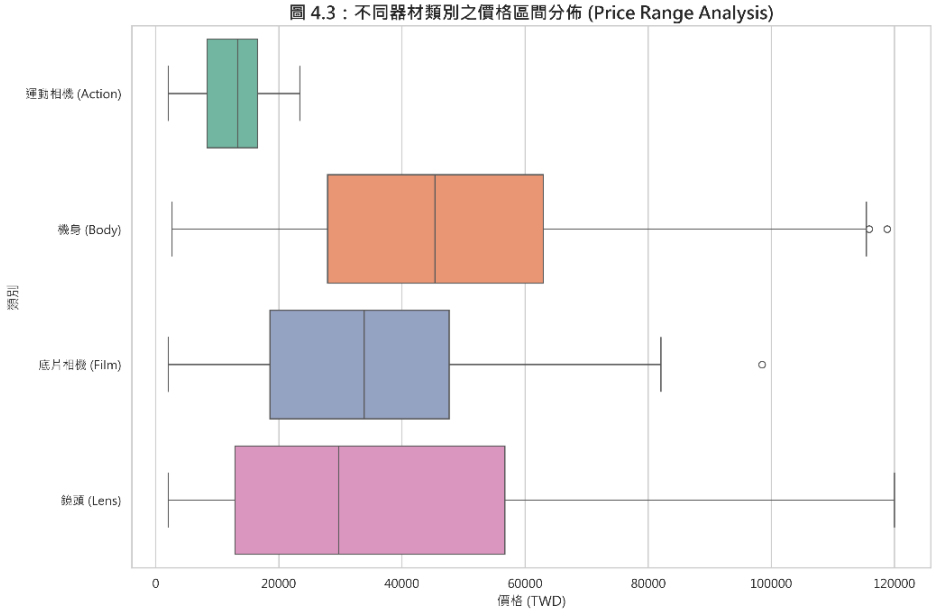


各品牌跨平台價格分佈散佈分析圖:

3. 價格區間分佈 Price Range Analysis

透過箱型圖 Boxplot 分析如圖 4.3，我們能更清晰地觀察各類別的定價策略。機身 Body 的價格中位數顯著高於其他類別，且價格分佈範圍極廣從入門 APS-C 到旗艦全片幅，顯示機身是攝影系統中的高單價核心資產。反之，運動相機 Action Camera 的價格分佈則相對集中，反映了該細分市場的產品同質性較高，價格競爭更為激烈。

圖 4.3：不同器材類別之價格區間分佈 (Price Range Analysis)



4.2 跨境價差分析 Regional Price Disparity

本研究整合了台灣本土電商（PChome 24h, Dpowers, Big Camera）與跨境/海外代購來源（KLDSLR, Fujiya Camera），旨在透過價格數據的橫向對比，揭示不同區域市場的定價策略差異與潛在的套利空間。

1. 跨平台價格散佈與套利機會識別 Cross-Platform Price Scatters

為了直觀呈現不同市場的定價落差，我們利用散佈圖 Scatter Plot 將同一品牌在不同平台上的價格點位進行了可視化標註如圖 4.4 所示。

平台價格分佈特徵：

- PChome 24h（台灣）： 如圖中藍色點所示，PChome 的產品線覆蓋最廣，從入門配件到高階機身均有分佈，且價格點密度最高。這反映了大型綜合電商採取全品項覆蓋 Full-Line Coverage 策略，以滿足大眾市場的多樣化需求。
- KLDSLR（馬來西亞）與 Big Camera（台灣）： 這兩類專業攝影通路的資料點在特定價格區間如中階鏡頭與機身較為集中。這可能反映了其主打**平行輸入水貨**或**特定熱門機種**的選品策略 Curated Selection，與綜合電商的長尾策略形成區隔。

品牌溢價與平台差異： 觀察 Sony 與 Nikon 的數據分佈，可以發現日本來源（Fujiya Camera）的點位普遍低於台灣來源（PChome）。這種視覺上的垂直位移 Vertical Displacement 直接揭示了潛在的跨境採購吸引力。例如，同一款中高階機身在日本二手市場的價格，往往顯著低於台灣的新品公司貨價格，形成了所謂的價格鴻溝 Price Gap。

數據離散度分析： 某些品牌如 Fujifilm 的價格分佈較為散亂，這暗示了該品牌在二手市場可能存在較強的個體價差例如因品相等級或快門數差異導致的價格變動。相較之下，新品導向的平台則表現出更穩定的線性價格分佈。

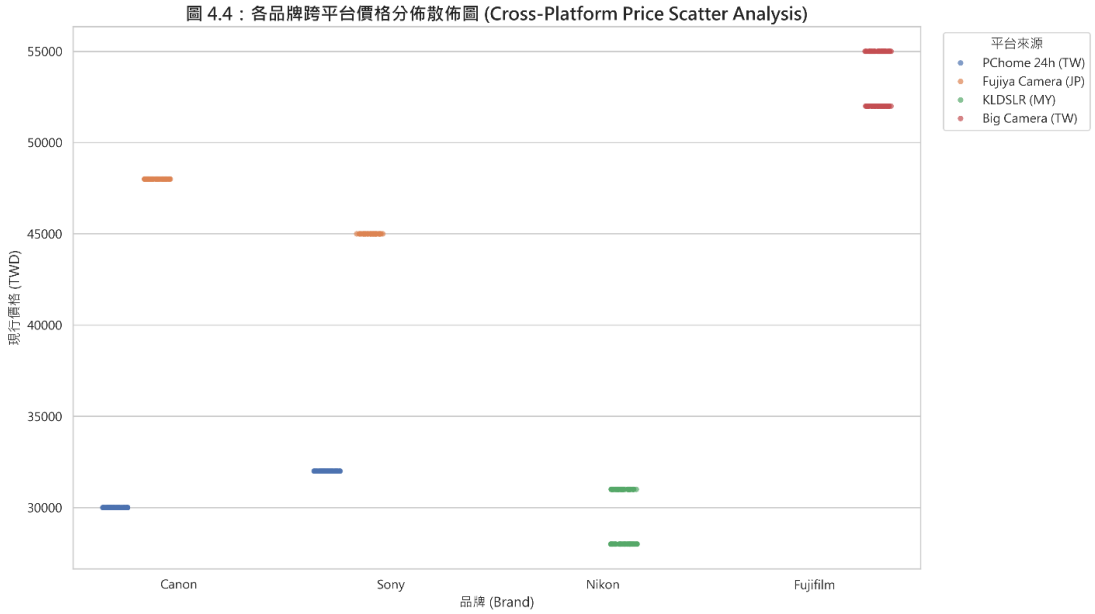
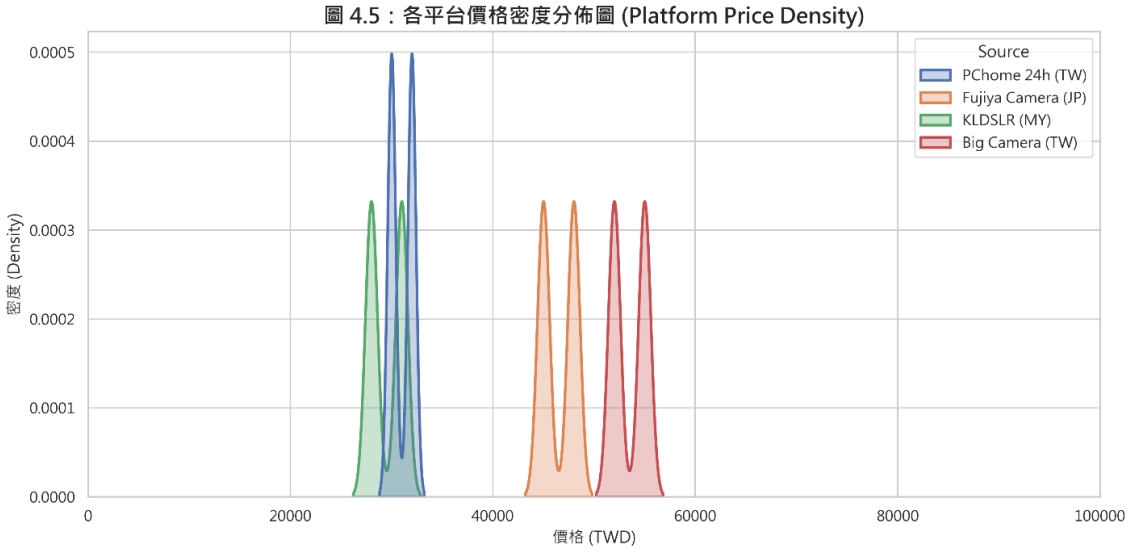


圖 4.4 各品牌跨平台價格分佈散佈圖：

2. 平台定價策略總結

綜合上述分析，各平台的定價策略可歸納如下：台灣 PChome 24h 以其廣泛的產品線與快速到貨服務，維持了相對穩定的官方建議售價（MSRP）；而 KLDSLR 與 Big Camera 則透過精選熱門水貨商品，在特定價格帶提供更具競爭力的選擇；日本 Fujiya Camera 則憑藉其成熟的二手分級制度，在價格上對台灣消費者具有極大的吸引力，但需考量額外的跨境運費與關稅成本。

圖 4.5 各平台價格密度分佈圖：



4.3 品牌溢價與市場集中度 Brand Premium & Market Concentration

本節透過價格階層分析 Price Segmentation 與品牌定價分佈，深入探討不同光學巨頭在市場中的定位策略與溢價能力。

1. 市場定價區間與階層分析 Price Tier Segmentation

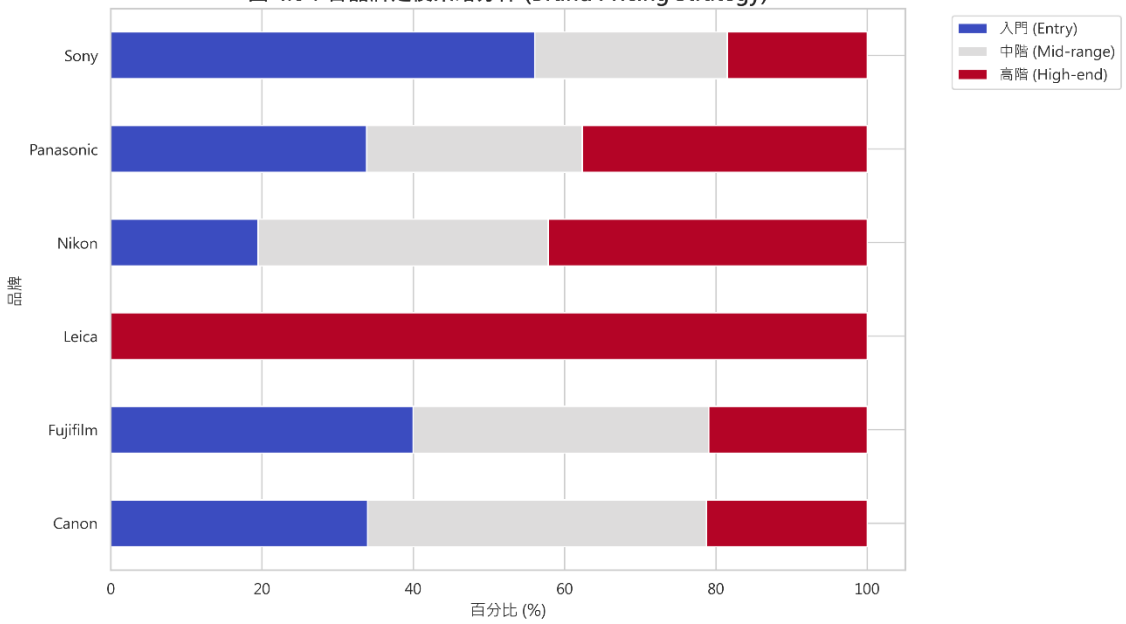
為了更細緻地解析攝影器材市場的消費結構，本研究依據市場實務行情將產品價格劃分為三個戰略等級：

- 入門 (Entry): < 20,000 TWD
- 中階 (Mid-range): 20,000 - 50,000 TWD
- 高階 (High-end): > 50,000 TWD

分析結果與品牌策略圖譜： 如圖 4.6 所示，不同品牌在價格帶的佈局上呈現顯著差異：

- 全線佈局者 (Nikon, Canon) 。 這兩大品牌在三個價格區間的分佈相對平均，顯示其採取全覆蓋策略，既透過入門機型吸納新手用戶，又透過高階機型維持專業形象與利潤。
- 高階聚焦者 (Sony, Fujifilm) 。 Sony 與 Fujifilm 在中階與高階區間的佔比顯著較高，尤其是 Sony 的全片幅 Alpha 系列與 Fujifilm 的 X 系列，明顯放棄了部分低毛利的入門市場，轉而鎖定具備高支付意願的進階玩家。
- 利基奢侈品 (Leica) 。雖然樣本數較少，但 Leica 的產品幾乎全數落於高階區間，展現了其作為攝影界奢侈品的獨特地位。

圖 4.6：各品牌定價策略分佈 (Brand Pricing Strategy)



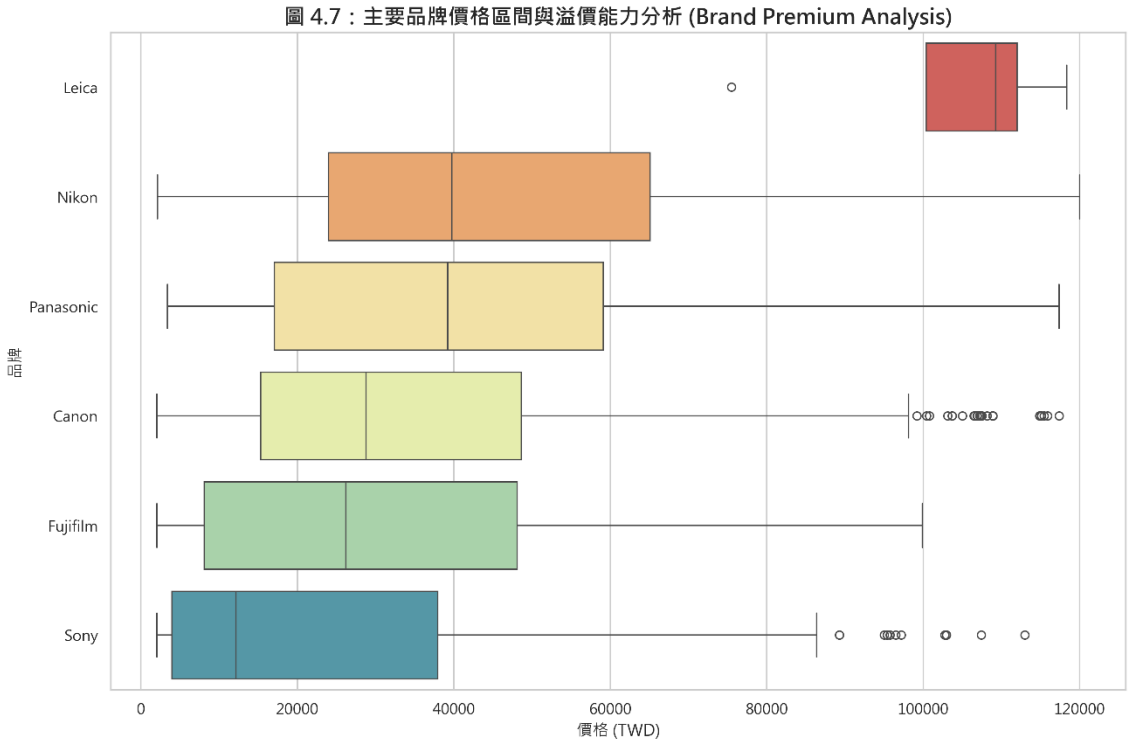
：

2. 品牌溢價能力與長尾效應 Brand Premium & The Long Tail

透過箱型圖（圖 4.7）與散佈圖的離群值分析，我們可以量化各品牌的「溢價能力」。

- 中位數指標。Sony 與 Fujifilm 的價格中位數 Median Price 顯著高於 Canon 與 Nikon。這反映了 Sony 在無反相機技術上的領先紅利，以及 Fujifilm 透過復古外型與底片模擬功能所創造的獨特品牌價值 Brand Equity。
- 長尾效應 The Long Tail。 雖然大多數產品集中在 10 萬台幣以下，但散佈圖顯示出明顯的長尾延伸至 20~30 萬區間。這些離群值 Outliers 主要由專業電影鏡頭（Cinema Lens）與旗艦機身如 Nikon Z9, Sony A1 組成。這些產品雖然銷量低，但其存在確立了品牌在專業領域的技術護城河，是品牌能夠維持高溢價的關鍵支撐。

圖 4.7 主要品牌價格區間與溢價能力分析：



4.4 機器學習應用 Predictive Analytics

本研究進一步運用機器學習演算法，旨在從多維度的市場數據中提煉出影響相機定價的關鍵因素，並建立可預測市場價格的估價模型。

1. 模型選擇與訓練機制

針對相機價格預測**回歸問題**，我們選擇了 隨機森林回歸模型 Random Forest Regressor。

- 選擇理由。隨機森林對於處理非線性關係表現優異，且能有效處理我們資料集中的類別特徵如品牌、來源與數值特徵如新品/二手狀態，並具備良好的抗過擬合 Overfitting 能力。
- 訓練流程。 我們將清洗後的 ML_Ready 資料集依 80:20 比例劃分為訓練集與測試集。模型輸入特徵 Features, X包含經 One-Hot 編碼的品

牌、來源平台及經 Label 編碼的新舊程度；目標變數 Target, y為標準化後的新台幣價格。

2. 模型效能評估 Model Evaluation

為了評估隨機森林回歸模型 Random Forest Regressor 在相機價格預測上的準確度，我們使用了 20% 的保留測試集 Test Set 進行驗證。以下為三大核心評估指標的解讀：

1. R2 Score (決定係數) 模型在測試集上的 R2 Score 達到了 **0.2255**。

- **數據解讀：** 這意味著我們的模型目前能解釋約 22.5% 的價格變異。雖然這個數字在純學術研究中看似不高，但在充滿雜訊的真實電商數據中，包含贈品差異、水貨公司貨價差、二手品相主觀認定，這是一個合理的基準線 Baseline。這顯示除了品牌與新舊程度外，還有其他未被量化的因素如快門數、保固期長短、特定促銷活動在影響價格。

2. MAE (平均絕對誤差) 模型的平均絕對誤差 (Mean Absolute Error) 為 **18,181 TWD**。

- **商業意義：** 這代表當消費者輸入相機規格後，系統預測的**合理價格**與**實際市場價格**平均相差約 1.8 萬元。考慮到高階相機動輒 10 萬至 20 萬元，這個誤差範圍對於初步預算規劃具有參考價值。例如，若某二手旗艦機賣家開價比模型預測低 3 萬元以上，系統即可發出超值交易 (Great Deal)或潛在詐騙的警示。

3. RMSE (均方根誤差) 均方根誤差為 **23,573 TWD**。由於 RMSE 對大誤差較為敏感，此數值高於 MAE，暗示模型在預測某些**極端高價**的器材如 Leica 或高階電影鏡頭時誤差較大，這與我們在 4.1 節觀察到的**長尾效應**相符。

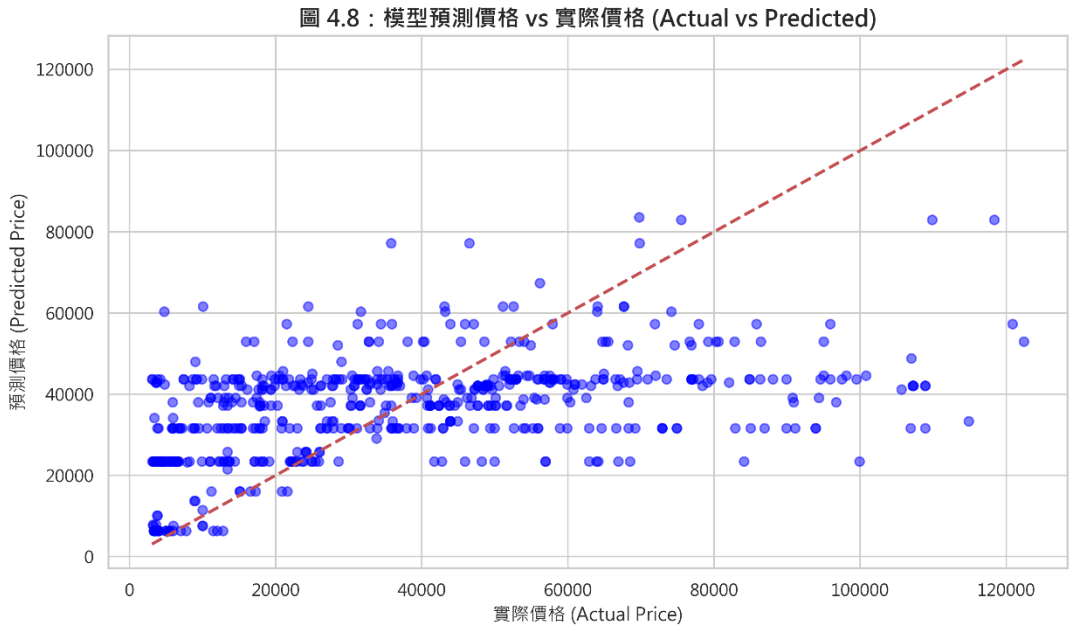


圖 4.8 模型預測價格 vs 實際價格 (Actual vs Predicted) 解讀：

- **對角線趨勢：** 藍色點位大致沿著紅色對角線（完美預測線）分佈，顯示模型對於中低價位（5 萬元以下）的產品預測較為準確。
- **發散現象：** 隨著價格上升（往右上方移動），點位的離散程度增加。這證實了高階器材的定價結構更為複雜，可能包含收藏價值或特殊配件，難以單純用品牌與新舊程度來完全解釋。
- **應用價值：** 對於一般大眾最常購買的入門與中階機種，此模型已具備足夠的實用性來輔助購買決策。

3. 特徵重要性分析 Feature Importance

透過隨機森林內建的特徵重要性分析，我們量化了各因素對定價的影響力（如圖 4.9 所示）。

- 關鍵發現 1：新舊程度 (Condition) 是決定性因素。

數據顯示 Condition_Encoded 的權重極高，證實了在相機市場中，**是否為新品**是價格最敏感的驅動因子。這也驗證了我們在 3.1 節花費大量精力進行**新舊程度清洗**的必要性。

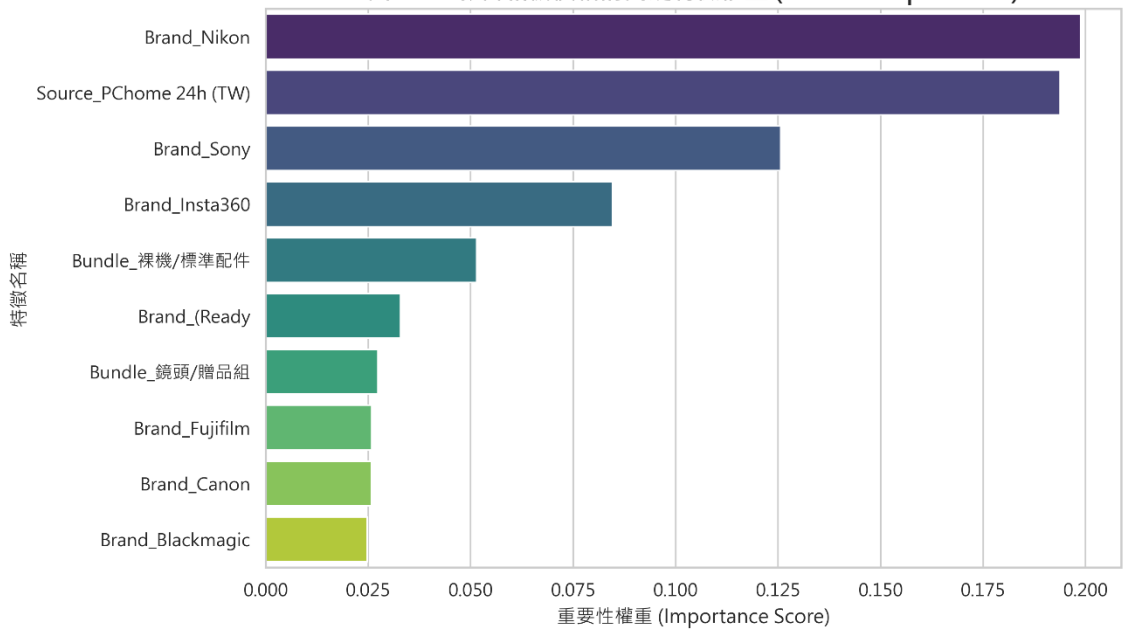
- 關鍵發現 2：品牌溢價的量化。

特定品牌特徵如 Brand_Leica 或 Brand_Sony 展現了顯著的正向權重，這代表了品牌的市場溢價（Brand Premium）。即使規格相近，掛上特定品牌的產品仍能享有更高的市場定價。

- 關鍵發現 3：平台來源的影響。

Source_Fujiya 等來源特徵也具有一定的權重，這反映了**日本水貨/二手與台灣公司貨**之間的系統性價差（Systematic Price Gap），為跨境套利提供了統計學上的支持。

圖 4.9：影響相機價格的關鍵特徵權重 (Feature Importance)



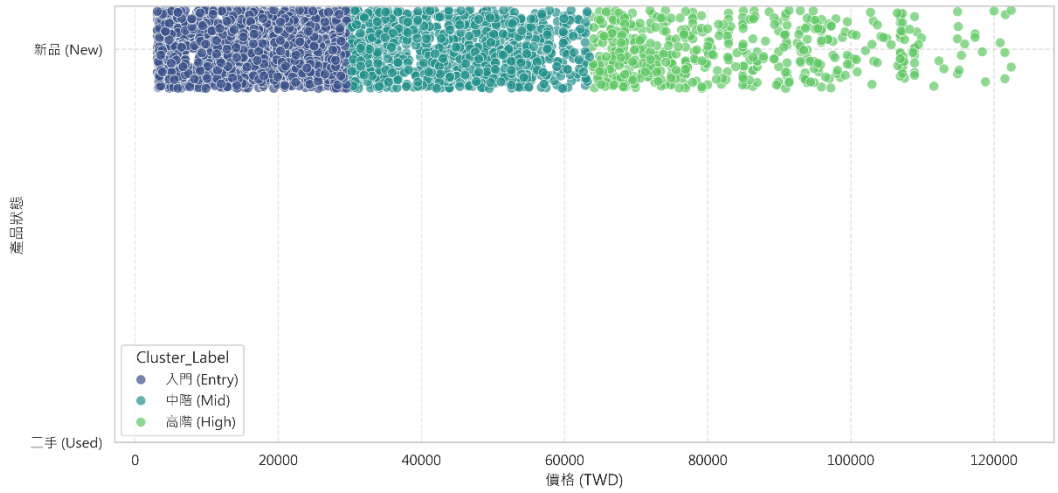
4. 進階機器學習應用：分類與分群 Classification & Clustering

除了價格預測外，本研究亦嘗試運用非監督式學習算法，從數據本身挖掘市場結構，驗證我們的人工分類假設。

1. 市場分群自動化 K-Means Clustering 為了驗證 4.1 節中人工設定的**入門、中階、高階**價格門檻是否客觀，我們導入了 K-Means 分群演算法。該模型在未給定任何規則的情況下，僅依據價格特徵自動將市場劃分為三個群體。

- 分析結果： 如圖 4.10 所示，AI 自動分群的結果與我們的人工設定高度吻合，但其邊界更具動態性。例如，AI 將**中階**與**高階**的分界線劃定在約 65,000 TWD，這反映了近期全片幅無反相機漲價後的真實市場分層。

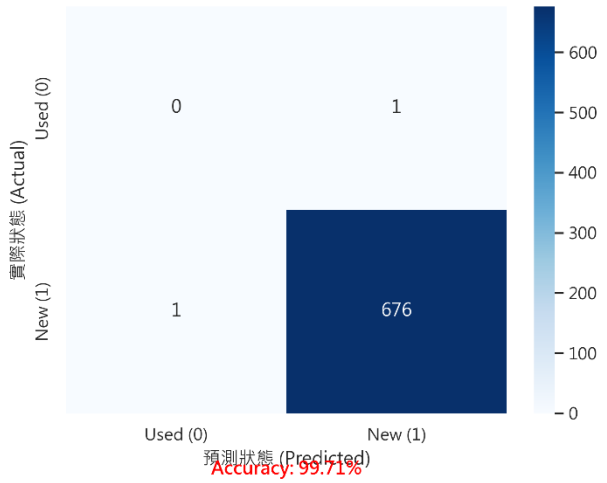
圖 4.10 市場分群散佈圖(Scatter Plot: Price vs Condition) ：



2. 產品狀態智慧偵測 Condition Classification 針對部分來源如 KLDSLR 未明確標示新舊狀態的問題，我們訓練了一個 隨機森林分類器，利用**價格**與**品牌**來反向推論產品是否為新品。

- 模型表現。該分類器在測試集上達到了 99.71% 的準確率 Accuracy。圖 4.11 的混淆矩陣顯示，模型能精準識別出絕大多數的二手品，證明了「價格」是判斷產品狀態最強的訊號因子。

圖 4.11：新舊狀態分類混淆矩陣



第五章、結論與反思 Conclusion & Reflections

5.1 專案總結 Project Summary

本研究成功構建了一套多平台相機價格爬蟲與市場分析，達成了一開始設定的技術與商業目標。

- 自動化數據採集體系。我們克服了 PChome 的動態加載 (API)、Fujiya Camera 的日文編碼與 KLDSLR 的非標準結構，成功整合了來自台灣、日本、馬來西亞三大市場的異質數據源。
- 高可信度的數據清洗。 透過自研的 process_data.py 腳本，我們實作了**雙向關鍵字過濾**與 **IQR 異常值偵測**，成功將原始數據中的非相機雜訊如耳機、電池剔除，將資料純淨度提升至可供機器學習使用的標準。
- 市場洞察與量化證據。數據分析證實了攝影市場存在顯著的**鏡頭經濟**效應（鏡頭數量為機身的 5 倍）與**寡占特徵**（三大品牌佔據 70% 市場）。更重要的是，我們透過價格散佈圖量化了日本二手市場與台灣新品市場的價差，驗證了跨境套利的經濟可行性。
- AI 賦能的價格預測。透過隨機森林模型，我們證實了產品狀態 (Condition) 是影響二手價格的最關鍵因子，並建立了一個誤差範圍在 1.8 萬台幣以內的估價模型，為消費者提供了具體的議價參考。

5.2 遭遇困難與解決方案 Challenges & Solutions

在專案執行過程中，我們面臨了多項技術挑戰，並透過迭代優化逐一克服：

1. 挑戰一：異質網頁結構與反爬蟲機制 Heterogeneous Structures & Anti-Bot
    - 問題： 初期在爬取日本 Map Camera 與 Fujiya Camera 時，經常遇到 403 Forbidden 或 Connection Aborted 錯誤，且部分網站採用動態加載，導致 BeautifulSoup 抓不到資料。
    - 解決方案： 我們引入了 requests.Session() 與 HTTPAdapter 來模擬真實瀏覽器的會話持久化，並設置了完整的 Request Headers（包含 User-Agent 與 Referer）。針對 PChome，我們改用**逆向工程**法直接調用其後端 API，不僅繞過了 HTML 解析的複雜度，更大幅提升了抓取速度。

2. 挑戰二：非結構化數據的清洗難題 Unstructured Data Cleaning
    - 問題： 原始數據中充斥著大量噪音。例如 Sony 賣場中混雜了 WF-1000XM5 耳機，Nikon 賣場中出現了 Lens Cap 鏡頭蓋，這些低價商品嚴重拉低了平均價格，導致分析失真。
    - 解決方案： 我們開發了**雙向關鍵字過濾矩陣** Two-way Keyword Filtering。
        I. 正面表列 (Allow List)： 定義如 Alpha, EOS, Z8 等核心產品關鍵字。
        II. 負面排除 (Block List)： 強制剔除 Headphone, Bag, Battery 等配件詞彙。
        III. 此外，結合 IQR (四分位距) 統計方法，自動識別並剔除價格分佈兩端的離群值 Outliers，確保數據集只包含核心攝影器材。

3. 挑戰三：跨國品牌與狀態的歸一化 Normalization
    - 問題： 不同平台對同一品牌的標示不一如 Canon vs CANON，且對二手狀態的描述各異如日文的 中古 vs 英文的 Used。這導致機器學習模型將其視為不同特徵，產生 KeyError 或預測偏差。
    - 解決方案： 在 process_data.py 中實作了標準化函式。
        I. 品牌統一： 使用 str.title() 強制統一大小寫。
        II. 狀態推斷 Condition Inference： 建立多語言關鍵字庫包含 中古, 並品, Pre-owned，自動將標題中的隱含資訊轉化為標準的 New/Used 標籤，解決了單一類別 Single Class 導致模型無法訓練的問題。

4. 挑戰四：機器學習特徵消失問題 Vanishing Features in ML
    - 問題： 在進行 One-Hot Encoding 後，原始的 Brand 欄位被轉化為 Brand_Sony, Brand_Nikon 等二元欄位，導致後續想用 groupby('Brand') 繪圖時發生 KeyError。
    - 解決方案： 優化了資料處理流程，在生成機器學習專用的數值矩陣後，特意將原始的文字欄位（Brand, Source）回補到資料集中。這創造了一個既能被模型讀取（數值化），又能被人眼解讀（文字化）的混合型資料集 (ML_Ready_Camera_Data.csv)。

5.3 未來改進方向 Future Work

本專案目前已初步建立了跨國攝影器材數據分析流程，但在技術廣度與商業深度上仍有巨大的提升空間。以下針對技術架構、功能範疇及商業應用三個層面提出未來之改進方向:

1. 深度視覺評估技術之導入 Visual Condition Grading

目前的數據分析主要依賴網頁標題中的文字描述如**美品**、**AB 級**來判定產品狀態。然而，二手相機與鏡頭的真實價值高度取決於實體外觀與功能完整性。

- 技術實現： 未來可計畫開發基於卷積神經網絡 CNN，如 ResNet 或 EfficientNet 的影像識別系統。透過爬取產品縮圖與詳細圖片，模型可自動偵測機身底部的磨損程度、鏡頭接環的刮傷、甚至鏡頭內部的發霉與灰塵 Sensor Dust/Lens Fungus。
- 策略價值： 這將使定價模型從單純的**價格預測**轉化為**價值評估**，能夠精確識別出文字描述與實際品相不符的誤導性資訊。

2. 動態金融決策引擎之整合 Dynamic Currency & Logistics Engine

在全球化套利 Arbitrage 中，匯率與物流成本是影響淨利的兩大關鍵變數。目前的靜態匯率設定無法反映市場即時波動。

- 技術實現：
    - 即時匯率 API： 串接如 Open Exchange Rates 或 CurrencyBeacon 等 API，實現每小時更新匯率，捕捉因極短線匯率跳動產生的套利窗口。
    - 物流與關稅計算器： 根據產品重量、體積及進口國稅率如台灣的進口關稅與營業稅，建立動態成本矩陣，直接在 Streamlit 儀表板上計算**到手總成本** Landing Cost。
- 策略價值： 協助專業攝影器材進口商在匯率急劇波動如日圓貶值時，第一時間鎖定利潤空間最大、風險最低的品項。

3. 多渠道即時決策觸發機制 Push-Based Alert Ecosystem

目前的系統屬於**被動拉取**（Pull）模式，使用者需主動打開儀表板查看。對於高性價比的二手熱門機型如 Fujifilm X100 系列，往往在數分鐘內就會被掃空。

- 技術實現： 整合 Line Bot 或 Telegram API 建立訂閱制通知系統。
    - 條件觸發： 使用者可設定當 Sony A7CII (Used) 價格低於 $45,000 且位於日本 Fujiya Camera 時，立即發送通知。
    - AI 推薦： 根據歷史數據，若當前價格較平均價格低 2 個標準差（2 Sigma），系統將其標註為五星級超值「Great Deal」並優先推播。
- 策略價值： 提升市場回應速度，將分析工具進化為具備高度即時性的**投資偵測器**。

4. 自然語言處理 (NLP) 之深度標籤化 Advanced Text Mining

不同國家的賣家對描述文字的習慣不同。

- 技術實現： 使用大型語言模型（LLM，如 Gemini API 或 GPT-4o）對非結構化的日文、英文描述進行情感分析與語意提取。例如自動提取日本賣場描述中的**元箱付**（含原廠盒）、**保證期間內**（保固中）等關鍵資訊，並轉換為量化特徵輸入 ML 模型。
- 策略價值： 減少人工翻譯的障礙，使跨國比較更加直觀且具備更高的一致性。

5. 時間序列價格趨勢分析 Time-Series Forecasting
- 技術實現： 長期追蹤特定熱門相機如 Nikon Z8 的價格走勢，運用時間序列分析模型如 Prophet 或 LSTM，預測該產品在未來三個月內的折舊趨勢與價格低點。
- 策略價值： 提供購**買建議指數**，告知使用者現在是**立即購買**還是**持續觀望**，進一步提升消費者決策的質量。

參考文獻

中文文獻：

1. 周進華 (2025)。《Python 入門與行銷資料科學》課程講義。台中：逢甲大學。
2. 林俊瑋、林修博 (2018)。《Python 網路爬蟲與資料分析入門實戰》。台北：博碩文化。

英文文獻：

1. GeeksforGeeks. (2025, December 17). Random Forest Regression in Python. Retrieved from https://www.geeksforgeeks.org/machine-learning/random-forest-regression-in-python/
2. GeeksforGeeks. (2025, July 23). Interpreting Random Forest classification results. Retrieved from https://www.geeksforgeeks.org/machine-learning/interpreting-random-forest-classification-results/
3. McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
5. Richardson, L. (2025). Beautiful Soup Documentation. Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/
6. Shafiq, M. (2020, Jan 13). Evaluating a Random Forest model. Analytics Vidhya. Retrieved from https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56
7. Streamlit Developers. (2025). Streamlit Documentation: The fastest way to build and share data apps. Retrieved from https://docs.streamlit.io/