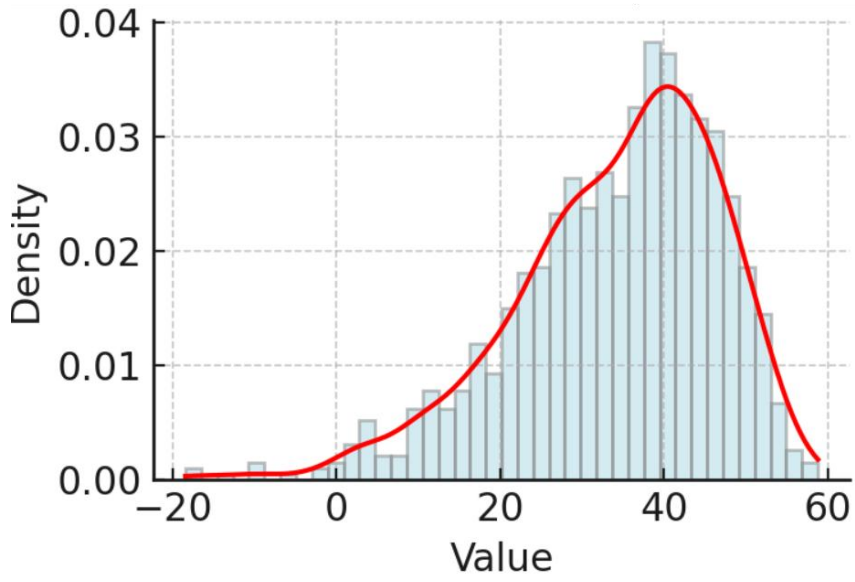
114 年第二次 AI 應用規劃師-中級能力鑑定【公告試題】

第二科：大數據處理分析與應用

考試日期：114 年 11 月 08 日 試題公告日期：114 年 11 月 20 日

答案	題目
D	1. 若某數據點的 Z 分數（Z-Score）= 2，請問代表下列哪一種意涵？ (A)代表該數據點之原始數值為 2； (B)該數據點比平均值低 2 個標準差； (C)代表數據為異常值； (D)該數據點比平均值高 2 個標準差
B	2. 使用 Python 的 pandas 套件處理各商品銷售數據（變數為 df）時，若需計算「總銷售額」欄位的敘述性統計量（如平均值、標準差等），應使用下列哪一種語法？ (A)df['總銷售額'].sum()； (B)df['總銷售額'].describe()； (C)df['總銷售額'].sort_values()； (D)df['總銷售額'].stats()
A	3. 附圖為某資料之分佈圖，此圖資料之偏態（Skewness）值較有可能為下列哪個選項？  (A)Skewness < 0； (B)Skewness > 0； (C)Skewness = 0； (D)無法計算 Skewness
B	4. 累積分佈函數（Cumulative Distribution Function, CDF）可用於描述隨機變數的機率分佈特性，其數學定義為下列何者？ (A)機率密度函數（Probability Density Function, PDF）的平均值； (B)機率密度函數（Probability Density Function, PDF）的積分；

答案	題目
	(C)機率密度函數（Probability Density Function, PDF）的離散總和； (D)機率密度函數（Probability Density Function, PDF）的標準差
B	5. 在進行資料前處理時，若使用 Label Encoding 將類別變數轉換為數字型態，下列何者為最常見的潛在風險？ (A)無法處理缺值； (B)會引入類別之間的虛假順序關係； (C)無法擴展至新資料； (D)記憶體佔用過高
C	6. 在進行資料分析時，會遇到類別型（Categorical）與數值型（Numerical） 資料格式。關於這兩種資料格式的處理，下列敘述何者不正確？ (A)One-Hot 編碼（One-Hot Encoding）會將類別變數轉換為多維二元向量，適用於無序（Nominal）類別資料，但在高基數（High Cardinality）特徵下可能造成維度爆炸問題； (B)標籤編碼（Label Encoding）會以整數表示不同類別，若應用於無序（Nominal）資料，可能導致模型誤將編碼值解讀為具數值大小關係的特徵； (C)標準化（Standardization）透過將資料平移與縮放，使其平均值為 0、標準差為 1，可在多數距離型演算法中改善收斂速度，並同時將數值範圍壓縮至 0 至 1 之間； (D)對連續變數進行分箱（Binning）可提升模型可解釋性，但若分段方式未依據資料分佈特性設計，可能導致資訊損失或邊界偏誤
C	7. 在資料庫的 ACID 特性中，下列何者為「原子性（Atomicity）」的正確定義？ (A)所有資料欄位必須為相同型別； (B)每次交易需以批次方式執行； (C)交易不可分割，需完全成功或完全失敗； (D)系統會自動同步交易資料至所有節點
B	8. 資料科學家為分析顧客行為，利用現有欄位「銷售金額」與「瀏覽次數」，計算出新變數「銷售金額/瀏覽次數」。此動作屬於下列哪一類特徵工程方法？ (A)特徵選擇（Feature Selection）； (B)特徵衍生（Feature Derivation）；

答案	題目
	(C)特徵轉換（Feature Transformation）； (D)分箱處理（Binning）
C	9. 在進行數值特徵的標準化（Normalization）時，若資料中存在極端值（Outliers），下列哪一種方法最適合使用？ (A)Min-Max 正規化（Min-Max Scaling）； (B)Z-score 標準化（Z-score Normalization）； (C)穩健縮放（Robust Scaling）； (D)標準分箱（Standard Binning）
C	10 下列哪一種情境最適合應用異常偵測（Anomaly Detection）技術？ (A)根據歷史銷售資料預測特定商品在旺季期間是否會出現供貨短缺，以提前調整庫存策略； (B)透過信用風險模型預測顧客是否可能發生違約，以輔助核貸決策； (C)即時分析金融交易資料流，偵測與平常交易行為明顯不同的可疑交易紀錄； (D)監控線上服務平台的使用者登入次數，預測次日的登入量變化趨勢
C	11 若一家公司需即時監控大量物聯網裝置的異常行為，下列哪一種組合最適合此應用？ (A)傳統關聯式資料庫+圖形視覺化； (B)批次資料處理+雲端備份； (C)大數據平台+即時資料分析技術； (D)Word 文件+手動標註
A	12 在處理分類問題時，若某一類樣本數明顯少於其他類別，研究人員可能採用隨機過採樣（Random Oversampling）以平衡資料比例，此方法最常造成下列哪一種問題？ (A)增加過擬合風險； (B)降低模型的收斂速度； (C)減少資料總筆數數量； (D)導致訓練資料欄位缺失
D	13 下列何者為同態加密（Homomorphic Encryption）技術的核心特性？ (A)將資料轉換為匿名識別碼以隱藏身分； (B)對資料進行標準化處理以提升模型精度； (C)自動偵測與排除異常值； (D)可直接在加密狀態下進行數據運算

答案	題目
D	14 某組資料共 10 項標籤如下： A, A, A, A, A, B, B, B, B, B 若該標籤僅有 A、B 兩種，請問這組資料的「正規化吉尼不純度（Normalized Gini impurity)」為何？ (A)0； (B)0.42； (C)0.84； (D)1
C	15 某家客服中心統計資料發現，平均每小時會接到約 20 通顧客來電，但每分鐘的來電數量不固定，可能為 0、1、2 通不等。這些來電事件彼此獨立，且在短時間內，發生的機率與時間長短成正比。若要以機率模型描述「每分鐘接到幾通來電」的機率分佈，下列哪一種最適合使用？ (A)均勻分佈（Uniform distribution)； (B)指數分佈（Exponential distribution)； (C)卜瓦松分佈（Poisson distribution)； (D)常態分佈（Normal distribution)
A	16 某金融科技公司以 Z 分數（Z-Score）監控交易金額異常狀況。若交易金額平均為新台幣 2,000 元，標準差為 400 元，某筆交易金額為 3,200 元，且公司以$\|Z\| \geq 3$ 判定為異常值（Outlier)，下列判斷何者最為正確？ (A)該筆交易的 Z 分數為 3，應標記為異常值； (B)該筆交易的 Z 分數為 2.5，屬於合理變異範圍； (C)該筆交易的 Z 分數為 2，顯示模型標準差估計過高； (D)該筆交易的 Z 分數為 1.5，無須納入異常檢測
B	17 某電商公司欲利用顧客行為資料建立消費預測模型，其中「會員等級」欄位包含「一般、白金、黑卡」三種類別。若模型採用梯度提升樹（Gradient Boosting Tree）演算法，資料科學家在進行特徵編碼時應特別注意下列何種情況？ (A)應優先採用獨熱編碼（One-Hot Encoding)，以減少類別之間的相依性與記憶體使用量； (B)直接使用標籤編碼（Label Encoding）可能使模型誤判類別間存在順序關係，導致特徵重要性偏誤； (C)使用目標編碼（Target Encoding）會自動消除過擬合

答案		題目
		（Overfitting）風險； (D)若類別數量較少，建議先使用主成分分析（Principal Component Analysis, PCA）進行降維
A	18	某人工智慧團隊使用分散式資料庫（Distributed Database）儲存模型訓練資料，並在更新訓練樣本時啟用多節點交易。若其中一個節點在交易過程中發生錯誤，但系統仍確保整體資料不會出現部分更新、最終狀態維持一致，下列何者最能說明此現象？ (A)系統透過原子性（Atomicity）確保交易必須全部成功或全部回復（Rollback）； (B)系統透過一致性（Consistency）確保交易完成後資料符合完整性規則； (C)系統透過隔離性（Isolation）避免多筆交易同時存取或修改相同資料； (D)系統透過持久性（Durability）確保交易一旦提交，其結果將永久保留於資料庫中
C	19	某製造企業導入上萬台物聯網（IoT）感測器以進行設備健康監測。系統需在毫秒級回應異常事件，並同時將完整資料保留於雲端供後續 AI 模型訓練與分析。若企業希望兼顧即時性、資料完整性與可擴展性，下列哪一種資料流程設計最符合此目標？ (A)感測器 → 雲端 API Gateway → 分散式資料庫→ 批次特徵工程（→ 模型推論； (B)感測器 → MQTT Broker → 雲端資料倉儲→ 即時儀表板→ 模型再訓練； (C)感測器 → 邊緣運算節點→ 流式資料處理框架（Stream Processing Framework）→ 雲端資料湖→ 模型推論； (D)感測器 → 本地快取層→ RESTful API → 雲端報表系統）→ 模型批次更新
D	20	某銀行計畫將信用風險評估模型部署至雲端平台，以便即時分析客戶交易行為。由於涉及大量敏感金融資料，銀行要求雲端服務商在不解密原始資料的情況下仍能執行模型運算。為達成此目標，最適合採用下列哪一項技術？ (A)在上傳資料前進行匿名化（Anonymization），僅保留可識別代碼供比對使用；

答案	題目
	(B)利用雜湊（Hash）函數轉換資料，以確保模型可追蹤但無法還原個資； (C)採用資料本地化（Data Localization）策略，將所有模型訓練限制於內部伺服器中； (D)透過同態加密（Homomorphic Encryption），讓雲端系統能直接在加密資料上執行運算，解密後結果與原始資料一致
B	21 某資料分析師設計在業務績效報告時，希望單一頁面中同時呈現多區域、不同產品線的銷售趨勢變化，並確保主管能在短時間內掌握整體資料走向。若依據 Edward Rolf Tufte 的數據密度（Data Density） 原則，下列哪一種設計方式最能符合該概念？ (A)將每個區域的銷售資料分成多張獨立折線圖，以避免資訊重疊； (B)使用顏色區分產品線，於同一圖表中整合多區域趨勢線，保持比例一致且標註清晰； (C)移除所有輔助線與標籤，僅保留主要折線以凸顯趨勢； (D)將資料轉換為表格形式，確保數值精確呈現並取代圖表視覺化
D	22 某投資研究員希望分析四檔科技類股（A、B、C、D）每日報酬率的變化趨勢，以判斷這些股票之間是否存在高度相關性與共變動性，並評估投資組合分散風險的程度。若研究員希望以單一圖表快速呈現各股票間的關聯強度與方向，下列哪一種視覺化呈現方式最適合？ (A)為每檔股票各自繪製直方圖（Histogram）以比較報酬率分佈； (B)針對任兩檔股票繪製散佈圖並加上趨勢線（Regression Line）； (C)使用雙軸折線圖（Dual-axis Line Chart）同時顯示四檔股價變化； (D)熱力圖（Heatmap）配合相關係數矩陣（Correlation Matrix）
C	23 某研究團隊以單樣本 t 檢定（one-sample t-test）檢驗「新行銷策略後的平均月銷售額是否與原本的 100 萬元不同」，顯著水準設定為 α =0.05。檢定結果顯示：p 值=0.08，且 95%信賴區間為 [95 萬元, 108 萬元]。根據上述結果，下列敘述何者正確？ (A)因 p 值< 0.05，可拒絕虛無假設； (B)若顯著水準改為 0.10，仍不顯著； (C)因 100 萬元落在信賴區間內，無法拒絕虛無假設； (D)信賴區間寬度僅與顯著水準有關
A	24 某企業建置生成式 AI 系統，利用大量客服紀錄與產品評論資料訓練語言模型，以自動生成客服回覆與知識摘要。由於資料來源多樣，且包含非結

答案	題目
	構化文字、影像與表格資訊，團隊希望在不降低模型效能的前提下，提升資料處理效率與一致性，下列哪一種資料處理策略最適合？ (A)建立資料湖（Data Lake）結構，並以 Apache Spark 或 Ray 進行分散式資料預處理與特徵抽取，再串接至模型訓練管線（Pipeline）； (B)採用單節點高效能伺服器搭配批次處理模式，集中執行資料清理與格式轉換； (C)將所有文字資料轉換為向量，並以資料庫索引方式直接餵入語言模型訓練； (D)使用生成式模型先行自動清理資料內容，再將結果輸入至下游訓練流程
A	25 某電商資料團隊繪製顧客單筆消費金額的箱型圖後發現：四分位距（IQR）範圍極小，但上鬚線拉得很長，且在高金額區域有多筆離群值。若希望協助行銷部門依據消費層級設計分群策略，下列哪一種視覺化方式最有助於凸顯不同消費層級間的差異？ (A)以對數刻度繪製箱型圖或長條圖，放大高金額消費族群的變化差異； (B)移除所有離群值，確保資料呈現集中分布； (C)採用等距分箱（Equal-Width Binning）方式分群； (D)改以折線圖（Line Chart）觀察時間變化趨勢
C	26 某串流影音平台運用關聯規則學習（Association Rule Learning）分析用戶的觀影行為，發現若使用者觀看了科幻影集，則有較高機率接著觀看超級英雄電影。分析顯示，同時觀看這兩種類型的使用者約佔全部觀影紀錄的 12%，而觀看科幻影集的使用者中，有 50%也觀看了超級英雄電影，該規則的提升度（Lift）為 1.8。根據上述資訊，下列哪一項推論最為正確？ (A)支持度（Support）過低，代表此規則不具任何商業價值； (B)提升度（Lift）大於 1 表示兩種類型內容無關，僅屬於隨機重疊； (C)信賴度（Confidence）為 50%，代表觀看科幻影集者有明顯傾向觀看超級英雄電影； (D)同時觀看比例僅 12%，代表兩種類型互相排斥
D	27 某金融科技公司分析每日上億筆交易資料，以監控客戶轉帳金額分佈與異常波動。由於資料量極大，為兼顧效率與準確度，團隊決定採用「近似分位數（Approximate Quantile）」方法進行資料摘要統計。下列何者最能正確反映該技術的核心目的？

答案	題目
	(A)確保每個分位值的結果完全精確，即使計算時間較長； (B)利用機器學習模型預測分位數位置，以減少統計計算量； (C)僅能對結構化資料進行批次處理，無法應用於即時資料流； (D)在可容忍誤差範圍內，快速估算分位值以支援即時分析
A	28 若在高維度（>500 維）的資料上應用 DBSCAN（Density-Based Spatial Clustering of Applications with Noise）演算法，卻發現所有資料點皆被判定為雜訊（Noise），下列何者為最有可能的原因？ (A)高維下距離變化趨同，導致 ε（Epsilon）閾值選擇失效； (B)使用錯誤的距離函數（Distance Function）； (C)MinPts 參數設得太小； (D)資料過度標準化導致特徵消失
D	29 某團隊在開發風險評估模型時，使用主成分分析（Principal Component Analysis, PCA）進行降維。輸入資料包含三個數值欄位：「交易金額（單位：新台幣）」、「交易次數（次／月）」與「年齡（歲）」，其數值量級分別約為 10^5、10^1與 10^2。分析人員直接將原始數據帶入 PCA，結果第一主成分（PC1）幾乎完全由「交易金額」主導。下列哪一項作法或判斷最合理？ (A)這是正常現象，金額本身變異較大，應主導主要成分； (B)若改用特徵選擇法，可自動解決變數量級問題； (C)可刪除「交易金額」欄位以平衡各主成分的影響； (D)在進行 PCA 前應先進行標準化（Standardization），以避免因數值尺度差異造成特徵偏誤
C	30 某行銷團隊想了解「廣告預算」與「銷售金額」之間的關聯程度。經繪製散佈圖後發現兩者呈現明顯線性趨勢，且資料中無明顯離群值（Outliers）。若希望衡量兩者之間線性關係的強度與方向，下列哪一種方法最適合？ (A)均方根誤差（Root Mean Squared Error, RMSE）； (B)共變異數（Covariance）； (C)皮爾森相關係數（Pearson Correlation Coefficient）； (D)平均絕對誤差（Mean Absolute Error, MAE）
B	31 某電商團隊觀察到，每位顧客對廣告推播的點擊行為可視為一次伯努利試驗（Bernoulli Trial），單次點擊成功機率為 p=0.4。當推播對象擴增至 5,000 位顧客時，團隊想快速預估「成功點擊總數」的分佈情形，以進行

答案	題目
	模型效能模擬與預測。若希望以常態分佈（Normal Distribution）近似原始分佈，下列哪一項判斷最為合理？ (A)因樣本數極大，可直接以常態分佈近似二項分佈（Binomial Distribution）； (B)只有當 np 與 n(1-p) 皆大於 5 時，才能以常態分佈作近似； (C)常態近似只適用於 p=0.5 的情況； (D)無論樣本數多大，二項分佈都不能以常態分佈近似
A	32 某電信公司導入生成式 AI 客服系統，利用過去對話紀錄與用戶行為資料訓練語言模型，在資料治理與合規審查過程中，團隊發現模型可能會在回答中生成包含真實姓名、電話或交易資訊的內容。為確保系統符合個資法及生成式 AI 的安全與隱私要求，下列哪一項作法最符合實務可行及法規原則？ (A)在訓練資料前進行資料匿名化（Anonymization）或偽匿名化（Pseudonymization）處理，並建立輸出內容稽核機制； (B)改以強化學習（Reinforcement Learning）微調模型，使模型學習避免產出真實資訊； (C)採用同態加密（Homomorphic Encryption）以加密所有文字輸入，確保模型無法辨識任何個資； (D)僅設定模型回覆時不顯示用戶姓名，即可視為隱私防護完成
D	33 某金融機構的量化分析師在建立資產風險評估模型時，發現報酬率資料分佈明顯非對稱，且出現多次極端損失事件，使得傳統假設常態分佈的模型無法準確反映真實風險。若希望在不依賴常態分佈假設的前提下，採取更能捕捉資料極端情況的建模策略，下列哪一種方法最為合適？ (A)採用線性迴歸模型（Linear Regression Model），以常態分佈殘差（Residuals）為基礎進行推估； (B)使用平均數（Mean）與標準差（Standard Deviation）估計波動範圍； (C)將資料裁剪至 $\pm 3\sigma$ 範圍內以排除異常值影響； (D)採用分位數回歸模型（Quantile Regression Model），聚焦於尾部分位（Tail Quantiles）以評估極端風險
B	34 在圖形資料庫（Graph Database）中建模社群平台資料時，若每筆「按讚」行為都包含時間戳記（Timestamp）與裝置類型（Device Type）等資訊。若希望同時保留使用者與貼文之間的互動關係，並能有效查詢「按

答案	題目
	讚」的行為屬性，下列哪一種設計方式最為合適？ (A)將「按讚」視為節點（Node），與使用者建立邊（Edge）； (B)將「按讚」資訊作為邊的屬性（Property）儲存，連結使用者與被按讚的貼文節點； (C)把「按讚」資訊直接寫入使用者節點中作為屬性； (D)建立「按讚紀錄表」並將資料存入關聯式資料庫
B	35 某企業欲建構知識圖譜（Knowledge Graph），以整合內部的研究報告、專利資料與專家知識，並支援語意查詢與關聯推理。若希望模型能具備良好的語意擴展性與高效推理能力，下列哪一種圖模型設計最為合適？ (A)僅以節點（Node）與邊（Edge）表示，所有資訊存放於節點屬性中； (B)將資料結構建為 RDF（Resource Description Framework）三元組（Subject－Predicate－Object）； (C)使用文件型資料庫儲存內容，並以標籤（Tag）連接節點； (D)採用關聯式資料庫儲存對應關係，並搭配預建索引加速查詢
B	36 某研究人員欲使用線性迴歸模型（Linear Regression Model）分析變數 Y 與 X 之間的關係，但發現 Y 的分佈明顯右偏，且其變異數隨 X 的增大而增加。為滿足模型假設並提升配適效果，下列哪一種前處理方法最為合適？ (A)對 X 進行標準化（Standardization）； (B)對 Y 進行 Box－Cox 轉換（Box－Cox Transformation）； (C)對資料進行一次差分（First Differencing）； (D)將 Y 中變異較大的樣本移除
C	37 若開發一個用於罕見疾病自動診斷的分類模型，目前資料集中確診樣本僅佔不到 1%，且因為標註成本高，短期內無法取得更多資料。在此情況下，若希望提升模型對少數類的偵測能力，同時避免過擬合，下列哪一種策略最為合理？ (A)對少數類進行隨機過採樣（Random Oversampling）； (B)對多數類進行欠採樣（Random Undersampling）； (C)使用 SMOTE（Synthetic Minority Over-sampling Technique）生成合成少數類樣本後再訓練分類模型； (D)僅使用現有資料調整模型決策閾值（Decision Threshold）以提升召回率
B	38 一家製造廠評估新生產線推出後，產品良率是否較原生產線提升。工程師

答案	題目
	分別從兩條生產線各抽樣 100 件產品，原生產線良率為 95%，新生產線為 97%。若欲檢定兩條生產線良率的差異是否具有統計意義，下列哪一種方法最為合適？ (A)雙樣本平均數 t 檢定（Two-sample t-test）； (B)雙比例 Z 檢定（Two-proportion Z-test）； (C)卡方檢定（Chi-square test）； (D)變異數分析（ANOVA）
D	39 若評估一個新開發的腫瘤分類模型，其資料集中有 80%的樣本來自良性病例。若直接使用 5-fold 交叉驗證（Cross-Validation） 進行模型評估，可能導致模型效能評估出現偏差，為避免此問題，下列哪一種作法最合適？ (A)降低 K 值以減少交叉驗證次數； (B)改為使用拔靴法（Bootstrap）； (C)調整測試集使良性樣本比例更高，以模擬真實分佈； (D)使用分層交叉驗證（Stratified K-Fold Cross-Validation），以確保每折類別比例一致
B	40 請參考附圖，下列虛擬程式碼（pseudocode）最可能是在描述何種驗證法？ Input： - data_set：包含 N 筆資料的資料集 - model_training_function：用來訓練模型的函式 - model_evaluation_function：用來評估模型的函式（如計算誤差或準確率） Output： - 平均評估指標（如平均準確率或平均誤差） Algorithm： 1. 初始化評估指標列表 metrics = [] 2. 對 i = 1 到 N： a. 將第 i 筆資料作為測試集 test_data b. 將其餘 N-1 筆資料作為訓練集 train_data c. 使用 model_training_function 在 train_data 上訓練模型 d. 使用訓練好的模型對 test_data 做預測，計算評估指標 metric_i e. 將 metric_i 加入 metrics 3. 計算 metrics 的平均值 mean_metric 4. 回傳 mean_metric (A) Hold-out 驗證（Hold-out Validation）；

答案	題目
	(B)留一交叉驗證 LOOCV（ Leave-One-Out Cross Validation）； (C)K-fold 交叉驗證（K-fold Cross Validation）； (D)拔靴法（Bootstrap）驗證
A	41 請參考附圖，下列虛擬程式碼（pseudocode）最可能是在描述何種演算法？ Input： - data_points：N 筆資料，每筆資料有 D 個特徵 - X：要分成的群數 Output： - clusters：每筆資料所屬的群編號 - centroids：每個群的中心點 Algorithm： 1. 隨機選擇 X 個資料點作為初始中心 2. 重複以下步驟直到收斂： a. 分群： 對每個資料點，計算它到每個中心的距離 將資料點指派給距離最近的中心 b. 更新中心： 對每個群： 計算該群中所有資料點的平均值 將群中心更新為這個平均值 3. 當群中心不再變動時，停止 回傳每筆資料的群編號 clusters，以及最後的群中心 centroids (A)K-means 分群（K-means Clustering）； (B)高斯混合模型分群（Gaussian Mixture Model Clustering）； (C)階層式分群（Hierarchical Clustering）； (D)DBSCAN 分群（Density-based Spatial Clustering of Applications with Noise Clustering）
C	42 考慮某生產線每小時出現瑕疵品的個數符合卜瓦松分佈（Poisson Distribution），已知平均每小時產生 5 個瑕疵品，附圖程式碼展示資料處理，請問下列敘述何者正確？

答案	題目
	`` import numpy as np from scipy.stats import poisson lambda_poisson = 5 print(poisson.pmf(5, lambda_poisson)) ``` (A) lambda_poisson = 5 表示每小時最多 5 個瑕疵品； (B) poisson.pmf(5, lambda_poisson) 表示小於 5 個瑕疵品的機率； (C)卜瓦松分佈的適用條件為事件彼此獨立，且平均發生率固定； (D) poisson.cdf(10, 5) 表示大於或等於 10 個瑕疵品的機率
	一間遊戲市場研究公司正在分析全球電子遊戲銷售情況，並準備建立一份「熱銷遊戲銷售報告」。分析師取得了一份名為 vgsales.csv 的資料集，內容包含了全球銷量超過 10 萬份的電子遊戲清單。研究團隊希望透過這份資料，了解不同年份、平台與地區的銷售趨勢。資料集的欄位說明如下，請根據下述資料情境回答以 43~47 題。 Name：遊戲名稱 Platform：遊戲平台（如 PS4、X360、Wii 等） Year：發售年份 Genre：遊戲類型（如 Action、Sports、Role-Playing 等） Publisher：發行商名稱 NA_Sales / EU_Sales / JP_Sales / Other_Sales：各地區銷售量（單位：百萬份） Global_Sales：全球總銷售量（單位：百萬份） 資料的欄位概觀如下： ``` import pandas as pd import matplotlib.pyplot as plt import seaborn as sns # 載入資料 data = pd.read_csv("vgsales.csv") data.head() ``` Rank Name Platform Year Genre Publisher NA_Sales EU_Sales JP_Sales Other_Sales Global_Sales 0 1 Wii Sports Wii 2006.0 Sports Nintendo 41.49 29.02 3.77 8.46 82.74 1 2 Super Mario Bros. NES 1985.0 Platform Nintendo 29.08 3.58 6.81 0.77 40.24 2 3 Mario Kart Wii Wii 2008.0 Racing Nintendo 15.85 12.88 3.79 3.31 35.82 3 4 Wii Sports Resort Wii 2009.0 Sports Nintendo 15.75 11.01 3.28 2.96 33.00 4 5 Pokemon Red/Pokemon Blue GB 1996.0 Role-Playing Nintendo 11.27 8.89 10.22 1.00 31.37

答案	題目
B	43 `data['Year']` 0 2006.0 1 1985.0 2 2008.0 3 2009.0 4 1996.0 ... 16593 2002.0 16594 2003.0 16595 2008.0 16596 2010.0 16597 2003.0 Name: Year, Length: 16598, dtype: float64 分析師在載入資料後，檢視 Year 欄位的資料型態，發現它是 float64，而非一般年份常用的整數。他想了解這樣的情形為什麼會發生。請問下列哪些原因可能導致這種狀況？ 原因 A：CSV 檔中 Year 欄位有缺失值(NaN)，導致 Pandas 自動將整欄轉為浮點數。 原因 B：CSV 檔中的年份資料原本是字串(如 "2006")，Pandas 轉換時出錯而變成浮點數。 原因 C：Pandas 預設會將所有數值型態讀取為 float64，不論資料是否為整數。 原因 D：CSV 檔中的年份資料可能包含小數點(例如 2006.0)，因此被視為浮點數。 (A)原因 B、原因 C； (B)原因 A、原因 D； (C)原因 A、原因 B、原因 D； (D)原因 C、原因 D
D	44 研究團隊接下來想要將 Year 欄位轉換為整數型態，以便後續進行年份趨勢分析。考慮到資料中可能包含缺失值（NaN），請選出最合適的轉換方式。 (A)data['Year'] = data['Year'].astype(int)； (B)data['Year'] = data['Year'].fillna(0).astype(int)； (C)data['Year'] = data['Year'].fillna(1).astype(int)； (D)data['Year'] = data['Year'].astype('Int64')；
A	45 為了觀察各遊戲平台的市場表現，分析師想要統計每個平台的全球銷售總額，並以長條圖呈現。請選出最能正確實現此分析的程式碼。 (A)data.groupby("Platform")["Global_Sales"].sum().plot(kind="bar")； (B)data.groupby("Platform")["Global_Sales"].count().plot(kind="bar")；

答案	題目
	(C)data["Platform"].value_counts().plot(kind="bar")； (D)data.groupby("Platform")["Global_Sales"].mean().plot(kind="bar")
C	46 團隊希望比較北美、歐洲、日本及其他地區的整體銷售比例，並使用 seaborn 套件以長條圖的形式進行可視化分析。請選出能正確顯示這些地區銷售總額比例的程式碼。 (A)sns.countplot(x=["NA_Sales","EU_Sales","JP_Sales","Other_Sales"], data=data)； (B)sns.lineplot(x="Platform", y=["NA_Sales","EU_Sales","JP_Sales","Other_Sales"], data=data)； (C)sns.barplot(x="variable", y="value", data=pd.melt(data, value_vars=["NA_Sales","EU_Sales","JP_Sales","Other_Sales"]), estimator=sum)； (D)sns.histplot(data[["NA_Sales","EU_Sales","JP_Sales","Other_Sales"]])
B	47 研究團隊想要知道在北美地區（NA）銷售成績最好的遊戲前五名，並希望以 seaborn 的條狀圖呈現結果。請選出能正確完成這項分析的程式碼。 (A)sns.barplot(x="NA_Sales", y="Name", data=data.head(5))； (B)sns.barplot(x="Name", y="NA_Sales", data=data.nlargest(5, "NA_Sales"))； (C)sns.lineplot(x="Name", y="NA_Sales", data=data.nlargest(5, "NA_Sales"))； (D)sns.countplot(x="Name", y="NA_Sales", data=data)
	使用銷售資料集(marketing.csv)進行迴歸分析，附圖程式碼展示資料載入與處理，請回答後續 48~50 題。 `import pandas as pd` `df = pd.read_csv("marketing.csv")` 下圖顯示資料集的前 5 筆資料與相關資訊。

答案	題目
	以下為 df.head() 與 df.describe() 的輸出結果： `>>> df.head()` `youtube facebook newspaper sales` `0 276.12 45.36 83.04 26.52` `1 53.40 NaN 54.12 12.48` `2 20.64 55.08 83.16 11.16` `3 181.80 49.56 70.20 22.20` `4 216.96 12.96 70.08 15.48` `>>> df.describe()` `youtube facebook newspaper sales` `count 200.000000 199.000000 200.000000 200.000000` `mean 176.451000 27.820101 36.664800 16.827000` `std 103.025084 17.808410 26.134345 6.260948` `min 0.840000 0.000000 0.360000 1.920000` `25% 89.250000 11.940000 15.300000 12.450000` `50% 179.700000 27.000000 30.900000 15.480000` `75% 262.590000 43.680000 54.120000 20.880000` `max 355.680000 59.520000 136.800000 32.400000`
D	48 根據上述結果，下列何者正確？ (A)資料集個數為 199 筆，變數個數為 4 個； (B)sales 變數的中位數是 16.827； (C)facebook 變數的第三四分位數(Q3)是 11.94； (D)youtube 變數的第一四分位數(Q1)是 89.25
C	49 參考下圖計算各變數的遺漏值(NaN)個數結果，下列何者正確？ `youtube 0` `facebook 1` `newspaper 0` `sales 0` `dtype: int64` 選項 A: df.isnull().sum() 選項 B: df.isNaN().sum() 選項 C: df.isna().sum() 選項 D: df.isnan().sum() (A)選項 D； (B)選項 B、選項 C、選項 D； (C)選項 A、選項 C；

答案	題目
	(D)選項 A、選項 B、選項 C
B	50 考慮資料集已經填補遺漏值，參考下圖執行結果，下列何者正確？

```
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
X = df[['youtube', 'facebook', 'newspaper']]
y = df['sales']
reg =    空格1
print(reg.coef_)
X2 = sm.add_constant(X)
model_sm =    空格2
print(model_sm.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.898
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     573.0
Date:                Sat, 20 Sep 2025   Prob (F-statistic):           1.03e-96
Time:                        19:37:30   Log-Likelihood:                -422.21
No. Observations:                 200   AIC:                             852.4
Df Residuals:                     196   BIC:                             865.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.5561      0.373      9.537      0.000       2.821       4.291
youtube        0.0455      0.001     32.702      0.000       0.043       0.048
facebook       0.1891      0.009     21.960      0.000       0.172       0.206
newspaper     -0.0006      0.006     -0.108      0.914      -0.012       0.011
```

A：空格 1 完整語法 reg = LinearRegression().fit(y, X)

B：空格 1 完整語法 reg = LinearRegression().fit(X, y)

C：print(reg.coef_) 結果為包括截距項等 4 個係數值

D：空格 2 完整語法 sm.OLS(X2, y).fit()

E：model_sm 迴歸模型的所有迴歸係數在 α=0.05 之下具有顯著的解釋力

F：截距項係數值為 3.5561

(A)B、C、F

(B)B、F

(C)A、C、D、F

(D)B、E

《以下空白》