111 年度中級巨量資料分析師能力鑑定試題

科目 1：資料分析與資料科學

考試日期：111 年 8 月 20 日

單選題 50 題（佔 100%）

D	1. 參考附圖，關於 R 語言使用 jsonlite 套件匯入 JSON 資料，下列敘述何者正確？  (A) 匯入後 df 資料物件為資料框（data.frame） (B) df 資料物件的元素長度為 9 (C) df$cwbopendata$location 資料物件為矩陣（matrix） (D) nrow(df$cwbopendata$location)結果為 438
C	2. 關於 ETL（Extract-Transform-Load），下列敘述何者「不」正確？ (A) 建置或更新資料倉儲（Data Warehouse）中的內容時所需的過程 (B) Extract：從資料來源處擷取所需之數據資料 (C) Transform：針對結構資料轉換，非結構資料則無法處理 (D) Load：最後將已作適當轉換過的數據資料載入到目的地
C	3. 可延伸標記式語言（Extensible Markup Language, XML）是一種標記式語言，被廣泛用來作為跨平台之數據互動的形式。設計 XML 是用來傳送和攜帶數據資訊，而非用於表現和展示資料，下列何項敘述違背所謂結構良好（well-formed）的 XML 文件？ (A) 每個 XML 元素必須有一個名稱 (B) XML 屬性名稱大小寫有別 (C) XML 屬性僅能出現在結尾標籤中，不得出現在起始標籤 (D) XML 元素的巢狀結構必須正確
D	4. 政府資料開放平臺（data.gov.tw）的檔案格式中，CSV（Comma-Separated Values）為常見格式之一。請問下列何者「並非」CSV 的特性？ (A) 儲存兩個維度的陣列資料 (B) 欄與欄之間以逗號分隔

(C) 列與列之間以換行分隔
(D) 不支援中文資料

A　5. 關於 ETL（Extract-Transform-Load）載入（Load），下列敘述何者「不」正確？
(A) 考慮到系統效能，通常採一筆一筆資料載入，確保資料的完整性
(B) 考量資料的完整性，先將資料載入到暫存區（Temp）或階段區（Stage），之後等資料都到位了之後，再由其他的 ETL 作業把資料一併載入到資料倉儲或資料市集
(C) 資料最後載入的目的地通常是資料倉儲（Data Warehouse）或是資料市集（Data Mart）
(D) 程式或工具對於載入介面的可擴充性及多樣性，是需要考慮的重點之一

C　6. 參考附圖，請問程式碼_____之處，應填入選項中哪一個 pandas 函數，才能得到如附圖下表之結果？
有一 pandas DataFrame 格式的變數 df，其資料內容如下：

	顧客編號	滿意度
0	A0001	非常滿意
1	A0002	非常滿意
2	A0003	不滿意
3	A0004	普通
4	A0005	非常不滿意

```
scoreDict = {
    "非常滿意": 3,
    "滿意": 1,
    "普通": 0,
    "不滿意": -1,
    "非常不滿意": -3
}
df['滿意度分數'] = df.滿意度._____(scoreDict)
```

	顧客編號	滿意度	滿意度分數
0	A0001	非常滿意	3
1	A0002	非常滿意	3
2	A0003	不滿意	-1
3	A0004	普通	0
4	A0005	非常不滿意	-3

(A) translate
(B) apply

	(C) map (D) groupby

D	7. 參考附圖，關於 R 語言使用 dplyr 套件進行資料分析，下列敘述何者正確？
	`> library(dplyr)`
	`> summarise(group_by(iris, Species), mean(Petal.Width))`
	(A) group_by 函數會依照 Species 進行遞增排序
	(B) group_by 函數會依照 Species 進行遞減排序
	(C) summarise 函數與 summary 函數功能相同，皆會顯示資料摘要
	(D) summarise 函數會依照 Speciess 群組資料，計算各群組的 Petal.Width 平均值

A	8. 參考附圖，關於 Python 語言使用 re 模組進行資料分析時，下列敘述何者正確？
	`import re`
	`mystr = '02-1234-5678'`
	`phoneRegex = re.compile(r'(\d{2})-(\d{4}-\d{4})')`
	`myresult = re.search(phoneRegex, mystr)`
	(A) myresult.group(0)結果為'02-1234-5678'
	(B) myresult.group(1)結果為'1234'
	(C) myresult.group(2)結果為'5678'
	(D) myresult.group(-1)結果為'1234-5678'

B	9. 參考附圖，R 語言中，已知 mystr 為串列（list），希望取出圖中串列結果，下列選項之 lapply 函數敘述何者正確？
	`> mystr`
	`[[1]]`
	`[1] "111" "6"  "1"`
	`[[2]]`
	`[1] "111" "6"  "2"`
	`[[3]]`
	`[1] "111" "6"  "3"`
	`> lapply( _____ )`
	`[[1]]`
	`[1] "1"`
	`[[2]]`
	`[1] "2"`
	`[[3]]`
	`[1] "3"`

(A) mystr, "(", 3

(B) mystr, "[", 3

(C) mystr, ")", 3

(D) mystr, "]", 3

C

10. 參考附圖，Python 語言中，當使用 pandas 與 numpy 模組進行資料分析時，下列敘述何者正確？

```
import pandas as pd
import numpy as np
df = pd.DataFrame({
    'A' : pd.Categorical(["train","train","test","test"]),
    'B' : [1,2,1,2],
    'C' : [1,2,3,np.nan]})
df_measure = df[['B','C']]
df_group = df.groupby('A')
```

(A) pd.isnull(df_measure).sum()結果是 1

(B) df_measure.mean()結果是計算每列(row)的平均值

(C) df_group.sum()結果是依群組計算各行(column)的數值總計

(D) df_group.agg(max)結果是依群組計算最小值

B

11. 請問執行附圖程式碼後，下列哪一個選項內的使用者 ID「不」在 df_final 的使用者 ID 欄位當中？

有一 pandas DataFrame 格式的變數 df1，其資料內容如下：

	使用者ID	試驗A_分數
0	A	50
1	B	70
2	D	83
3	E	54
4	G	96

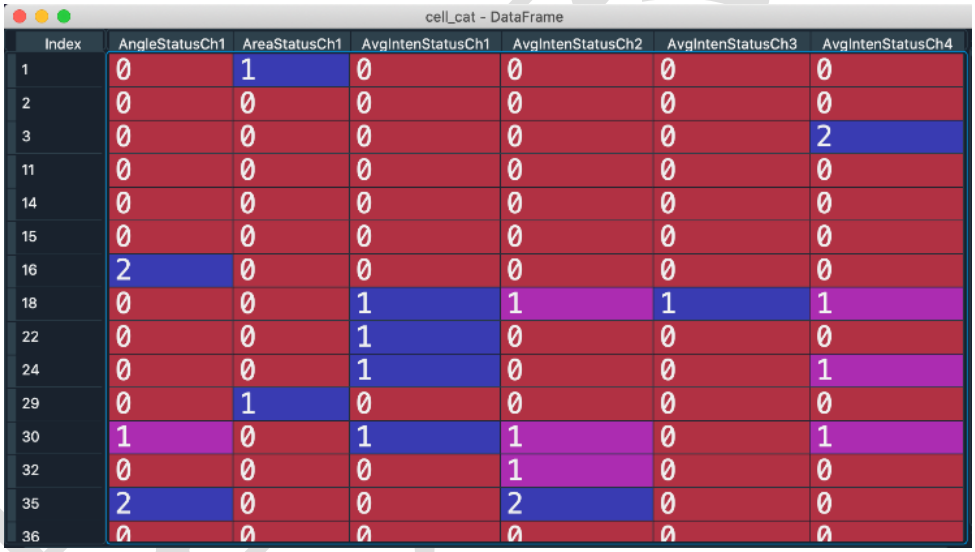
另一 pandas DataFrame 格式的變數 df2，其資料內容如下：

	使用者ID	試驗B_分數
0	A	40
1	B	55
2	C	35
3	D	74
4	E	66

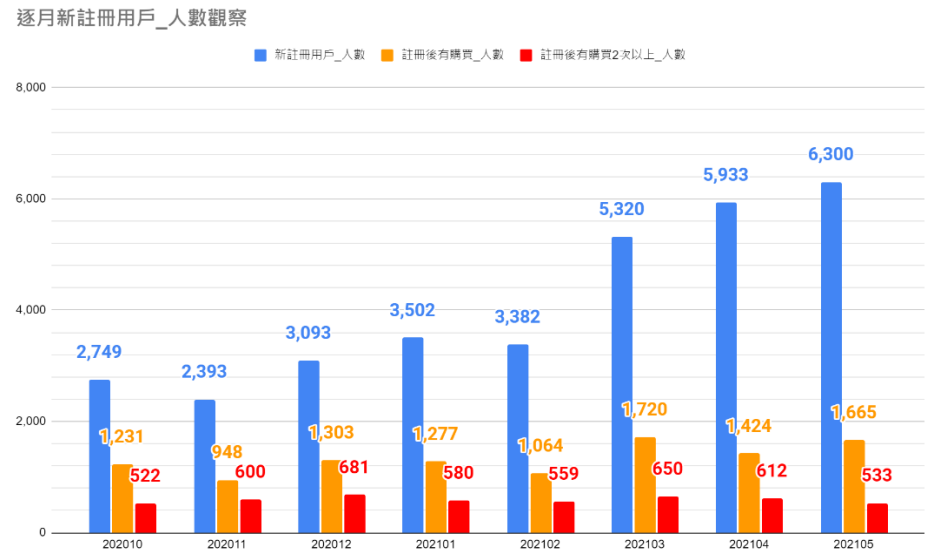
df_final = df1.merge(df2, on="使用者 ID", how="left")

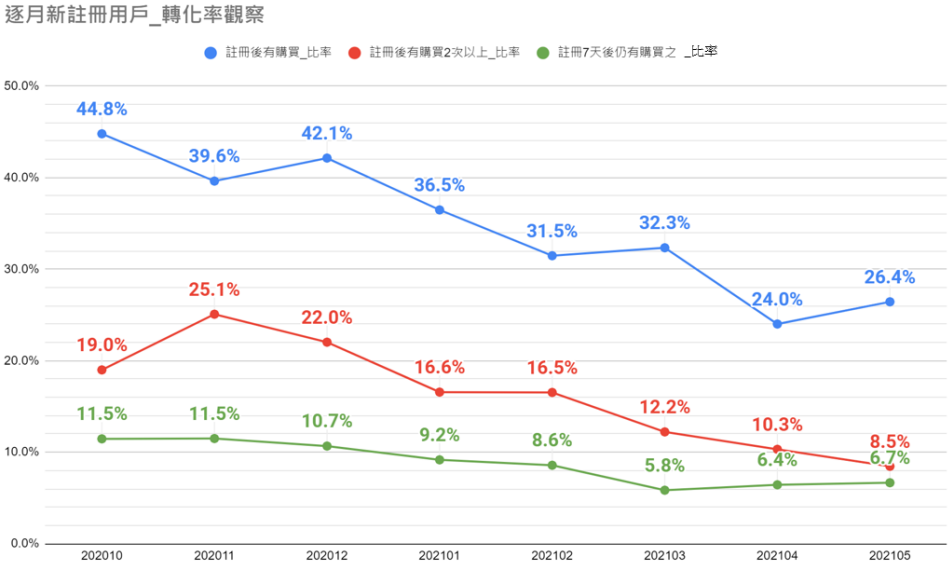
df_final

(A) A

	(B) C (C) G (D) E
D	12. 資料清理是指發現並糾正資料中的錯誤，關於資料清理的方法，下列敘述何者「不」正確？ (A) 驗證資料的正確性 (B) 遺缺值（missing value）的處理 (C) 異常值的處理 (D) 迴歸係數的處理
B	13. 附圖為 pandas 資料表（DataFrame）cell_ca t 的部份內容，請問下列選項何者為產製全部變量之次數分佈表的正確指令？  cell_cat - DataFrame (A) cell_cat.value_counts() (B) cell_cat.apply(lambda x: x.value_counts(), axis=0) (C) cell_cat.to_freq() (D) cell_cat.select_dtypes(include="object")
D	14. 關於主成分分析（Principal Components Analysis, PCA）於特徵提取（feature extraction）之主要用途，下列敘述何者正確？ (A) 提取重要特徵後不能以圖像視覺化呈現多變量資料 (B) 將低度相關的預測變數矩陣 X，轉換成相關且量多的潛在變項集合 (C) 將最相關的訊息與無關的雜訊結合 (D) 將問題領域中的數個變數，組合成單一或數個具訊息力的特徵變數
D	15. 關於主成分分析（Principal Component Analysis, PCA）與奇異值分解（Singular Value Decomposition, SVD）的比較，下列敘述何者正確？

<table>
<tr><td></td><td>(A) 當變量個數大於觀測值個數時，可以使用 PCA 進行維度縮減
(B) SVD 只有分解出資料矩陣橫列或縱行其一的基底向量（Basis）
(C) PCA 提供資料矩陣之縱行與橫列的基底向量
(D) SVD 較 PCA 更一般化</td></tr>
<tr><td>D</td><td>16. 關於長條圖（Bar Char）與直方圖（Histogram），下列敘述何者「不」正確？
(A) 直方圖的橫軸變數為數值型連續變數（Continuous Variable）
(B) 長條圖的橫軸變數為類別型離散變數（Discrete Variable）
(C) 直方圖的組距是有順序的，所以不可相互置換，而長條圖則無順序，可以置換
(D) 從長條圖可以看出中位數、眾數的大約位置</td></tr>
<tr><td>D</td><td>17. 關於盒鬚圖（Box Plot），下列敘述何者「不」正確？
(A) 顯示一組數據分布情況資料的統計圖
(B) 它能顯示出一組數據的最大值、最小值及四分位數
(C) 盒鬚圖可用來了解資料的偏態（Skewness）
(D) 盒鬚圖盒子的位置居左表示資料為左偏</td></tr>
<tr><td>B</td><td>18. 附圖為某電商平台於 2020 年 10 月至 2021 年 5 月份之新註冊用戶統計數據。請問關於選項中對此二圖表的敘述與解讀（請參考欄位定義與兩圖表），下列敘述何者「不」正確？
<table>
<tr><th>欄位</th><th>定義</th></tr>
<tr><td>新註冊用戶_人數</td><td>註冊於該月份的用戶人數。</td></tr>
<tr><td>註冊後有購買_人數</td><td>註冊於該月份、且有過購買紀錄的用戶人數。</td></tr>
<tr><td>註冊後有購買2次以上_人數</td><td>註冊於該月份、且有過2筆以上購買紀錄的用戶人數。</td></tr>
<tr><td>註冊7天後仍有購買之 _人數</td><td>註冊於該月份、且於註冊滿7日後仍有購買紀錄的用戶人數。</td></tr>
<tr><td>註冊後有購買_比率</td><td>=(註冊後有購買_人數 / 新註冊用戶_人數)</td></tr>
<tr><td>註冊後有購買2次以上_比率</td><td>=(註冊後有購買2次以上_人數 / 新註冊用戶_人數)</td></tr>
<tr><td>註冊7天後仍有購買之 _比率</td><td>=(註冊7天後仍有購買之 _人數 / 新註冊用戶_人數)</td></tr>
</table></td></tr>
</table>





(A) 此平台之逐月新註冊用戶有正成長趨勢，並於 2021 年 3 月份出現月增 57%的成長

(B) 由圖表來看，此平台的逐月獲利與新註冊人數，皆呈現正成長趨勢

(C) 由圖表來看，此平台每 100 個新註冊用戶、只有不到 12 個用戶會於註冊後 7 天仍有購買行為

(D) 儘管 2021 年 5 月的首購用戶數量、相較 2020 年 10 月成長了約 400 人；但此平台用戶的首購率卻下降了接近 20%。顯示平台對於設定目標對象、促使用戶購買的規劃上可能出現了問題

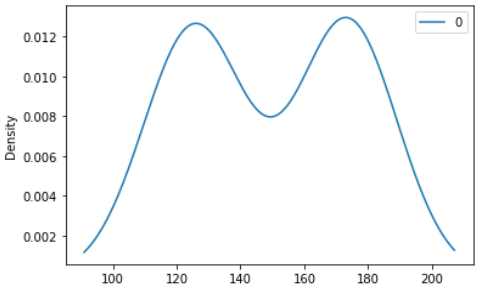
C

19. 請問下列選項中的圖表，何者較「符合」附圖程式碼進行核密度估計繪圖（Kernel Density Estimation, KDE）的結果？

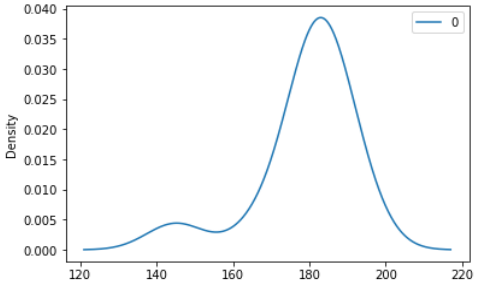
```
import pandas as pd
```

```
df = pd.DataFrame(
    [140, 143, 145, 145, 148, 153, 190, 192, 193, 193, 195, 198]
)
df.plot.kde()
```

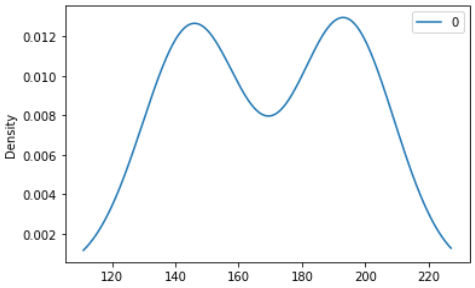
(A)



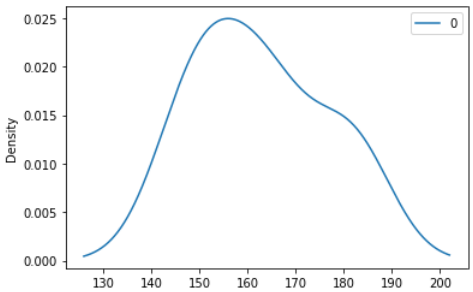
(B)



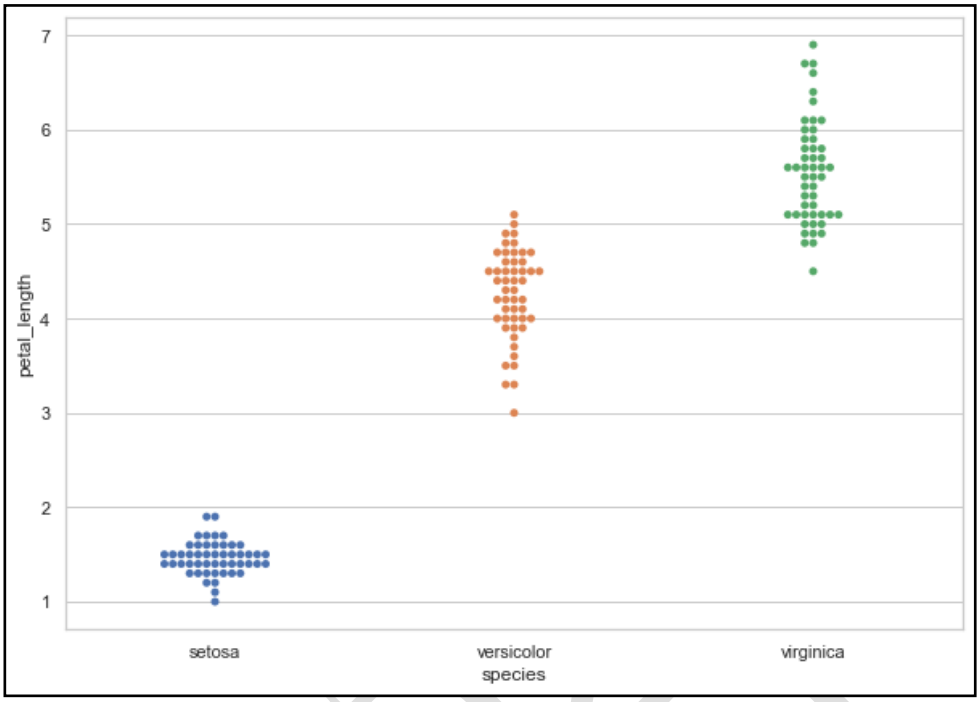
(C)



(D)



D	20. 附圖為鳶尾花資料集（iris dataset）所繪製而成的分布圖。關於該數據與圖表，下列敘述何者「不」正確？（x 軸為三種鳶尾花品種；y 軸為花瓣長度、單位：公分）



(A) setosa 物種的花瓣長度，大部分在 1-2 公分之間
(B) versicolor 物種的花瓣長度，平均較 setosa 物種來的更長
(C) virginica 物種最大的花瓣長度，可以達到接近 7 公分
(D) versicolor 物種最大的花瓣長度，不可能長到 virginica 物種最小的花瓣長度

A

21. 關於 R 語言模型參數調校使用 caret 套件，下列敘述何者「不」正確？
(A) 若想要自訂參數調校過程，可以 train()函數改變 trainControl()函數的控制參數
(B) 資料建模的相關參數有兩種：一種可直接利用資料估計其值的模型參數，另一種是不易從資料中估計的超參數
(C) train()函數結合重抽樣方法評估不同模型參數對績效的影響，並從中選出最佳模型
(D) 套件{caret}中 train()函數可針對不同模型給予參數調整的範圍

D

22. 有 4 種交叉驗證方法，分別為(1) 留一驗證法（leave-one-out cross-validation, LOOCV）、(2) 5 折（5-fold）交叉驗證、(3) Bootstrap、(4) 10 折（10-fold）交叉驗證。請問在一個約 1000 筆資料集的訓練過程，下列交叉驗證方法執行時間排序，何者正確？
(A) (4) > (2) > (1) > (3)
(B) (1) > (3) > (4) > (2)
(C) (3) > (4) > (2) > (1)
(D) (1) > (4) > (2) > (3)

D	23. 在影像分類中，假設圖片有 0,1,2 三個類別，請問我們該如何對附圖進行單熱編碼（one-hot encoding）？  (A) A=0, B=0, C=0 (B) A=1, B=1, C=1 (C) A=0, B=1, C=1 (D) A=0, B=1, C=0
D	24. k 折交叉驗證（k-fold cross-validation）是機器學習中常用來驗證訓練出來的模型好壞的一種方法，請問以下敘述何者正確？ (A) 當 k=10 時是指將數據集分成 10 份，其中 7 份做為訓練，剩下 3 份做驗證 (B) 通常會重複 k 次以上，再取其中 k-1 次的結果進行平均來評估模型準確率 (C) 資料依照類別排序後，依序將資料分成 10 份 (D) 留一驗證法（leave-one-out cross-validation, LOOCV）也是一種 k-fold cross-validation
A	25. Generative model 與 Discriminative model 是兩種不同類型的模型，Generative model 可以透過統計的方法，根據所觀測的資料來建立近似原始資料分布的統計模型，因此可以用在模擬上，下列何者「不」是 Generative model？ (A) Logistic regression (B) HMM（Hidden Markov Model） (C) GMM（Gaussian Mixture Model） (D) Naïve Bayes
C	26. 關於生成對抗網路（Generative Adversarial Network, GAN）進行超解析度成像，下列敘述何者「不」正確？ (A) 讓兩個神經網路相互博弈的方式進行學習 (B) 透過自己相互對抗的生成與鑑別網路，大幅減少資料量的需求 (C) 屬監督式學習訓練方法 (D) 模型嘗試訓練出比原圖更高解析度的圖像
B	27. 關於分類問題，不同類樣本數相差太大時，下列何種做法最「不」適

	合？ (A) 依比例重複對數量少的樣本抽樣，生成數量相當的樣本 (B) 直接進行分類，可以最大限度利用資料 (C) 從數量較多的樣本抽樣，使樣本數與樣本數較少的一方相符 (D) 給予各樣本與數量呈反比的權重
A	28. 訓練神經網路模型時，有時會遇到 Loss function 出現 NaN，下列何種做法最「不」恰當？ (A) 提高 Learning Rate，使其較快收斂 (B) 將輸入值作 Normalization (C) 設置 Gradient Clipping，限制梯度範圍 (D) 檢查輸入值，確保資料中不含 NaN
B	29. 關於梯度消失（Gradient Vanishing），下列敘述何者「不」正確？ (A) 當神經網路作反向傳播（Back-Propagation, BP），梯度由後往前傳，梯度不斷減小，最後變為零 (B) 可用 sigmoid function 作為 activation function 解決 gradient vanishing (C) 可用 ReLU function 作為 activation function 防止 gradient vanishing (D) gradient vanishing 會造成靠近輸出層的隱藏層（Hidden Layer）權重得不到更新
A	30. 實務上常見各類樣本分佈差距大的不平衡學習（Imbalanced Learning）情境，關於不平衡學習的處理方式，下列敘述何者「不」正確？ (A) 運用過度抽樣（Oversampling）解決之，此種方法可避免模型過度配適（Overfitting） (B) 運用正負樣本的懲罰權重來解決，若分析建模的算法支援樣本權重設定，此方法是簡單有效的解決途徑 (C) 以薈萃式學習（Ensemble Learning）集成模型解決，形成模型預測能力良好的森林（Forest） (D) 進行特徵選取（Feature Selection）來解決類別不平衡問題，透過變數的選取來提高模型績效
B	31. 就非監督式學習（unsupervised learning）而言，評估集群（cluster）優劣的一種方式是計算群內樣本的相似性（similarity）。當我們持續形成更多群時，群內相似性向上攀升，將樣本切分為更細的集群，請問此操作可能會發生什麼問題？ (A) 配適不足（underfitting） (B) 過度配適（overfitting）

	(C) 配適良好（well-fitted） (D) 配適狀況不明（unknown）
B	32. 在迴歸分析當中，最常用的迴歸係數估計方法是普通最小平方法（Ordinary Least Squares, OLS），不過 OLS 常常被錯誤理解或誤用，關於 OLS 失靈的狀況，「不」包括下列何項？ (A) 預測變量個數大於樣本數 (B) 預測變量間相關性（correlation）不足 (C) 預測變量矩陣存在共線性（collinearity） (D) 某一變量是其他變量的線性組合
A	33. 關於線性相依（linearly dependent）、線性獨立（linearly independent）、正交（orthogonality）與相關（correlation），下列敘述何者正確？ (A) 如果 X 與 Y 線性獨立，則兩者無關/正交 (B) 如果 X 與 Y 無關/正交，則兩者線性獨立 (C) 如果 X 與 Y 線性相依，則兩者相關 (D) 如果 X 與 Y 線性相依，則兩者非正交
D	34. 關於因素分析（factor analysis）的概念，下列敘述何者「不」正確？ (A) 因素分析是利用少數幾個因素來解釋一群彼此有關係存在的變數 (B) 因素分析的每個變數除了受共同因素（common factor）影響外，也包含獨特因素（specific factor）存在 (C) 因素分析的應用在於從一群變數中找出少數幾個具代表性的變數，以便作為進一步的統計分析用 (D) 資料經因素分析後不能以簡化後的因素對個體作分析
A	35. 收集學生的國文、英文、統計、經濟、會計成績進行主成分分析（Principal Components Analysis, PCA），計算出 5 個特徵值，分別為：λ_1=3.148，λ_2=1.352，λ_3=0.351，λ_4=0.122，λ_5=0.037，第 1 主成分，解釋全體變數的變異數比例為何？ (A) 62.9% (B) 90.0% (C) 87.3% (D) 67.7%
B	36. 蒐集氣象觀測站的溫度、水量、風速…等 10 項氣象指標以主成分法做因素分析。10 個特徵值與解釋變量如附圖，下列敘述何者「不」正確？

	Eigenvalue	Cumulative Eigenvalue	Cumulative %

λ_1	4.540	4.540	45.4%
λ_2	2.674	7.214	72.1%
λ_3	0.959	8.173	81.7%
λ_4	0.576	8.749	87.5%
λ_5	0.503	9.252	92.5%
λ_6	0.422	9.674	96.7%
λ_7	0.179	9.853	98.5%
λ_8	0.081	9.934	99.3%
λ_9	0.058	9.992	99.9%
λ_{10}	0.008	10.000	100.0%

(A) 第 1 個特徵值可解釋全體變數的變異數 45.4%

(B) 若以解釋能力 90%為標準，需選入 4 個特徵值（λ_1、λ_2、λ_3、λ_4）

(C) 雖然第 3 個特徵值小於 1，但前 3 個特徵值的可解釋全體變數的變異數 80%以上，仍可考慮選入

(D) 選擇 2 個因素（λ_1、λ_2）可獲得 72.1%的解釋能力

B

37. 下列何者是較穩健（Robust）的相關性衡量方法？

(A) 肯德爾（Kendall）相關係數法

(B) 最小共變異數判別式法（Minimum Covariance Determinant, MCD）

(C) 皮爾森（Pearson）相關係數法

(D) 史皮爾曼（Spearman）相關係數法

B

38. 關於 k 近鄰（k-nearest neighbors）分類法，下列敘述何者「不」正確？

(A) 須留意預測變數的尺度

(B) 能容忍遺缺值

(C) 不須對資料有任何分布上的假設

(D) 需選擇適合的 k

C

39. 以年收入（X_1）和房子坪數（X_2）做區別變數，辨別家庭有無投資股票。分別蒐集 30 個有投資股票與 30 個無投資股票的家庭資料。相關數據資料如附圖，請問下列何者「不」正確？

無投資股票的平均數向量（以$\overline{x_1}$為例 60.00 為平均年收入；42.87 為房子平均坪數）與共變量數矩陣

$$\overline{x_1} = \begin{pmatrix} 60.00 \\ 42.87 \end{pmatrix} \quad s_1 = \begin{bmatrix} 68.83 & 45.76 \\ 45.76 & 61.98 \end{bmatrix}$$

有投資股票的平均數向量與共變量數矩陣

	$\overline{x_2} = \begin{pmatrix}75.13\\55.03\end{pmatrix}$ $s_2 = \begin{bmatrix}50.12 & 33.58\\33.58 & 57.34\end{bmatrix}$ (A) 綜合共變易矩陣： $s = \frac{1}{2}(s_1 + s_2) = \begin{bmatrix}59.47 & 39.67\\39.67 & 59.66\end{bmatrix}$ (B) 兩條線性區別函數為 $d_1(x) = 0.9517x_1 + 0.0857x_1 - 31.0808$ $d_2(x) = 1.1645x_1 + 0.1481x_1 - 48.5156$ (C) 假設有一個家庭的年收入和房子坪數分別為 81 和 59，這個家庭會被歸類在沒有投資股票組 (D) 當區別規則 d2(x) > d1(x)，則將之歸類在第 2 群體
D	40. 關於分群（clustering）演算法，下列敘述何者正確？ (A) k-means 演算法進行分群前，可先不決定 k 值 (B) 資料量夠大就不需人為定義分群數 (C) 分群的效果與資料數量、群集數量都無關 (D) 訓練模型所使用的資料不需要包含類別標籤
A	41. 關於深度學習的說明，下列敘述何者「不」正確？ (A) 利用多層神經網路來分析數據，重點是事先給定特徵值 (B) 不斷地調整某個網路中的每個參數，直到找到一個參數組合使之有效運作 (C) 具忍受有雜訊的數據，可分析影像、影片等多維度且複雜的數據 (D) 神經元個數相同的卷積神經網路表現會比一般深度神經網路來得出色，大幅降低了需要訓練的參數量
C	42. 使用 k-means 分群法（k-means clustering algorithm）與歐氏距離（Euclidean distance），將附圖資料分成三群，何者會自成一群？

ID	X	Y
p1	0	0
p2	5	9
p3	3	2
p4	0	3
p5	13	5
p6	4	10
p7	2	2
p8	7	10
p9	3	11
p10	6	9

(A) p7

	(B) p9 (C) p5 (D) p2
A	43. 關於優化神經網路準確度（Accuracy）的方法，下列敘述何者「不」正確？ (A) 將訓練總 Epoch 數減少 (B) 將每層神經網路內之神經元數量加多 (C) 將訓練總 Epoch 數提高 (D) 使用資料增強之技巧
D	44. 考慮購物網站的銷售資料集時，若使用集群法（clustering）進行銷售分析，下列敘述何者正確？ (A) k-means 集群法（k-means clustering）的結果是同一集群內的樣本點具有高度的差異性 (B) k-medoid 集群法（k-medoid clustering）與 k-means 集群法（k-means clustering）比較時，前者較容易受到異常值或極端值的影響 (C) 凝聚階層法（agglomerative hierarchical）是一種切割式集群法（partitional clustering） (D) 在 k-medoid 集群法（k-medoid clustering）中，側影係數（Silhouette Coefficient）為正數且數值較大時，表示該資料分派到較合適的集群
D	45. 關於使用支援向量機（Support Vector Machines, SVM）的核函數（kernel function）於處理分類問題時，下列敘述何者正確？ (A) 核函數的目的是將原始資料轉換至低維度空間中，方便進行分類 (B) 使用較複雜的核函數轉換，其分類正確率一定較高 (C) 如果是不平衡資料（Imbalanced Dataset）時，可以考慮使用較小反正規化參數（Regularization Parameter）C 值 (D) 考慮 x 與 y 二個向量，則徑向基函數核（Radial Basis Function Kernel）為 $K(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
B	46. 考慮使用 iris 資料集，輸入變數為 Sepal.Length（花萼長度）、Sepal.Width（花萼寬度）、Petal.Length（花瓣長度）與 Petal.Width（花瓣寬度），輸出變數為 Species（物種）。使用 keras 等模組進行多層感知器（Multilayer Perceptron）分析，參考附圖 Python 語言結果，下列敘述何者「不」正確？

```
import numpy as np
import pandas as pd

from keras.models import Sequential
from keras.layers import Dense
from keras.utils import to_categorical
```

```
model = Sequential()
model.add(Dense(6, input_shape=(4,), activation="relu"))
model.add(Dense(6, activation="relu"))
model.add(Dense(3, activation="softmax"))
```

```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 6)                 30

 dense_1 (Dense)             (None, 6)                 42

 dense_2 (Dense)             (None, 3)                 21

=================================================================
Total params: 93
Trainable params: 93
Non-trainable params: 0
_________________________________________________________________
```

(A) 輸入層有 4 個特徵
(B) 第 1 個隱藏層的權重參數有 42 個
(C) 第 2 個隱藏層的神經元個數為 6 個
(D) 輸出層的啟動函數（activation function）為 softmax 函數

C　47. 考慮企業分析不同廣告費用（youtube，facebook，newspaper）對銷售額（sales）的影響，參考附圖 R 語言使用 lm 函數分析結果，下列敘述何者「不」正確？

```
> summary(marketing)
    youtube          facebook       newspaper          sales
 Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92
 1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45
 Median :179.70   Median :27.00   Median : 30.90   Median :15.48
 Mean   :176.45   Mean   :27.82   Mean   : 36.66   Mean   :16.83
 3rd Qu.:262.59   3rd Qu.:43.62   3rd Qu.: 54.12   3rd Qu.:20.88
 Max.   :355.68   Max.   :59.52   Max.   :136.80   Max.   :32.40
> summary(lm(sales ~ ., data=marketing))

Call:
lm(formula = sales ~ ., data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6539  -1.0505   0.2733   1.4182   3.3793

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5561270  0.3728689   9.537   <2e-16 ***
youtube      0.0455313  0.0013923  32.702   <2e-16 ***
facebook     0.1891022  0.0086111  21.960   <2e-16 ***
newspaper   -0.0006339  0.0058512  -0.108    0.914
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.018 on 196 degrees of freedom
Multiple R-squared:  0.8977,    Adjusted R-squared:  0.8961
F-statistic:   573 on 3 and 196 DF,  p-value: < 2.2e-16
```

(A) 資料沒有遺漏值

(B) 調整後判定係數為 0.8961

(C) newspaper 變數對整體線性模型最具有影響

(D) 全部資料其有 200 筆

B

48. 關於集群分析（cluster analysis），下列敘述何者「不」正確？

(A) 集群分析是依據個體間的相似性，將資料分群，使群內差異小，群間差異大

(B) 集群分析主要有兩種形式，分別為 k-means 分群（k-means clustering）和分層分群（hierarchical clustering），這兩種方式皆需在一開始就決定好分群數

(C) 集群分析與其他分類分析，如判別分析（discriminant analysis）不同之處在於分組結果完全由資料所導出，各群的特性事前未知

(D) 集群分析的變數只能使用連續（continuous）變數，不能使用類別（categorical）變數

D

49. 關於迴歸分析（regression analysis），下列敘述何者「不」正確？

(A) 在預測的研究中需將變數區分為反應變數和解釋變數

(B) 迴歸分析是利用兩個或多個數量變數間的關係，使反應變數的值可以用一個或多個解釋變數的值加以預測的方法

(C) 參數皆為一次的模型稱為線性模型，一般以 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{(p-1)} X_{(p-1)} + \varepsilon$ 的形式表達

	(D) 殘差分析（residual analysis）無法用來評估迴歸模型的預測品質
B	50. 請問下列敘述何者「不」正確？ (A) 當應變數和自變數皆屬連續型資料，可用迴歸分析來探討其關係；若當應變數是分類時，可用判別分析（discriminant analysis）來探討其因果關係 (B) 判別分析（discriminant analysis）的分析資料不需要任何的假設條件 (C) 判別分析與集群分析都是針對每個觀察值的個體做分類，但兩者的差異在於判別分析已知每個觀察值所屬群體，而集群分析則否 (D) 判別分析（discriminant analysis）可以用來找出對於應變數有解釋能力的自變數