

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 1 頁，共 12 頁

## 單選題 50 題 (佔 100%)

D	<p>1. 當我們在清理資料時，經常需要處理字串 (String) 相關資料，請問在 Python 語言中，已知 string = 'Hello World'，關於處理資料的結果，下列何者「不」正確？</p> <p>(A) string.split(' ')會取得['Hello', 'World']</p> <p>(B) string.replace('Hello', 'World')會取得'World World'</p> <p>(C) string[:8]會取得'Hello Wo'</p> <p>(D) string[8]會取得'o'</p>																																
D	<p>2. 參考附圖，關於 R 語言遺缺值 (Missing Value)，下列敘述何者「不」正確？</p> <pre>&gt; summary(Cars93[24:27])</pre> <table><thead><tr><th>Luggage.room</th><th>Weight</th><th>Origin</th><th>Make</th></tr></thead><tbody><tr><td>Min. : 6.00</td><td>Min. :1695</td><td>USA :48</td><td>Acura Integra: 1</td></tr><tr><td>1st Qu.:12.00</td><td>1st Qu.:2620</td><td>non-USA:45</td><td>Acura Legend : 1</td></tr><tr><td>Median :14.00</td><td>Median :3040</td><td></td><td>Audi 100 : 1</td></tr><tr><td>Mean :13.89</td><td>Mean :3073</td><td></td><td>Audi 90 : 1</td></tr><tr><td>3rd Qu.:15.00</td><td>3rd Qu.:3525</td><td></td><td>BMW 535i : 1</td></tr><tr><td>Max. :22.00</td><td>Max. :4105</td><td></td><td>Buick Century: 1</td></tr><tr><td>NA's :11</td><td></td><td></td><td>(Other) :87</td></tr></tbody></table> <p>(A) 在資料蒐集的過程中，由於外在環境的影響，資料有可能出現缺失的遺漏值</p> <p>(B) 面對大數據的資料分析，遺漏值的資料儘可能不要直接刪除</p> <p>(C) Cars93 資料物件的 Luggage.room 欄位有遺漏值</p> <p>(D) sum(is.na(Cars93\$Luggage.room))執行結果為 14</p>	Luggage.room	Weight	Origin	Make	Min. : 6.00	Min. :1695	USA :48	Acura Integra: 1	1st Qu.:12.00	1st Qu.:2620	non-USA:45	Acura Legend : 1	Median :14.00	Median :3040		Audi 100 : 1	Mean :13.89	Mean :3073		Audi 90 : 1	3rd Qu.:15.00	3rd Qu.:3525		BMW 535i : 1	Max. :22.00	Max. :4105		Buick Century: 1	NA's :11			(Other) :87
Luggage.room	Weight	Origin	Make																														
Min. : 6.00	Min. :1695	USA :48	Acura Integra: 1																														
1st Qu.:12.00	1st Qu.:2620	non-USA:45	Acura Legend : 1																														
Median :14.00	Median :3040		Audi 100 : 1																														
Mean :13.89	Mean :3073		Audi 90 : 1																														
3rd Qu.:15.00	3rd Qu.:3525		BMW 535i : 1																														
Max. :22.00	Max. :4105		Buick Century: 1																														
NA's :11			(Other) :87																														
B	<p>3. 參考附圖，關於 Python 語言 re 套件，下列敘述何者正確？</p> <pre>import re</pre> <p>x = "台股今(20)日封關，去年第4季在半導體領漲帶動下，上市公司市值水漲船高，豬年封關上市公司總市值達37兆元，一年市值增加6兆元、年增23%。"</p> <p>(A) re.findall("\d{1,3}", x)，執行結果為['20', '37', '23']</p> <p>(B) re.findall("\d{1,3} 兆元", x)，執行結果為['37 兆元', '6 兆元']</p> <p>(C) re.findall("\d{1,2} 日", x)，執行結果為['(20)日']</p> <p>(D) re.findall("\d%", x)，執行結果為['23%']</p>																																
C	<p>4. 下列何者「不」屬於資料前處理 (Data Preprocessing) 的步驟之一？</p> <p>(A) 資料清理 (Cleaning)</p> <p>(B) 資料整合 (Integration)</p> <p>(C) 資料建模 (Modeling)</p> <p>(D) 資料轉換 (Transformation)</p>																																
C	<p>5. 正規表示式可運用於多處，在匯入檔案時，使用下列何者可搜尋出任何空白字元？</p> <p>(A) \n</p>																																

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 2 頁，共 12 頁

	(B) \r (C) \s (D) \t																																			
B	<p>6. 參考附圖，在 R 語言中，若希望 iris 資料集依照第一順序 Petal.Width 欄位由大至小排序，第二順序 Petal.Length 欄位由小至大排序，應使用下列何者程式指令？</p> <table><thead><tr><th>Sepal.Length</th><th>Sepal.Width</th><th>Petal.Length</th><th>Petal.Width</th><th>Species</th></tr></thead><tbody><tr><td>6.7</td><td>3.3</td><td>5.7</td><td>2.5</td><td>virginica</td></tr><tr><td>6.3</td><td>3.3</td><td>6.0</td><td>2.5</td><td>virginica</td></tr><tr><td>7.2</td><td>3.6</td><td>6.1</td><td>2.5</td><td>virginica</td></tr><tr><td>5.8</td><td>2.8</td><td>5.1</td><td>2.4</td><td>virginica</td></tr><tr><td>6.3</td><td>3.4</td><td>5.6</td><td>2.4</td><td>virginica</td></tr><tr><td>6.7</td><td>3.1</td><td>5.6</td><td>2.4</td><td>virginica</td></tr></tbody></table> <p>(A) iris[order(iris\$Petal.Width, iris\$Petal.Length), ] (B) iris[order(-iris\$Petal.Width, iris\$Petal.Length), ] (C) iris[order(iris\$Petal.Length, iris\$Petal.Width), ] (D) iris[order(-iris\$Petal.Length, iris\$Petal.Width), ]</p>	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	6.7	3.3	5.7	2.5	virginica	6.3	3.3	6.0	2.5	virginica	7.2	3.6	6.1	2.5	virginica	5.8	2.8	5.1	2.4	virginica	6.3	3.4	5.6	2.4	virginica	6.7	3.1	5.6	2.4	virginica
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species																																
6.7	3.3	5.7	2.5	virginica																																
6.3	3.3	6.0	2.5	virginica																																
7.2	3.6	6.1	2.5	virginica																																
5.8	2.8	5.1	2.4	virginica																																
6.3	3.4	5.6	2.4	virginica																																
6.7	3.1	5.6	2.4	virginica																																
D	<p>7. 參考附圖，關於 Python 語言 pandas 套件中，使用 aggregate 函數之執行結果為何？</p> <pre>In [1]: import pandas as pd ...: import numpy as np ...: df = pd.DataFrame([[2, 1, 1], ...:                    [True, 2, 6], ...:                    [6, 3, 9], ...:                    [True, np.nan, 2]], ...:                    columns=['A', 'B', 'C'])  In [2]: df Out[2]:    A    B  C 0  2  1.0  1 1 True  2.0  6 2  6  3.0  9 3 True NaN  2  In [3]: type(df) Out[3]: pandas.core.frame.DataFrame  In [4]: df.agg(['sum', 'mean'])</pre> <table><tr><td>(A)</td><td>A</td><td>B</td><td>C</td></tr><tr><td>sum</td><td>8.0</td><td>6.0</td><td>18.0</td></tr><tr><td>mean</td><td>2.0</td><td>2.0</td><td>4.5</td></tr></table>	(A)	A	B	C	sum	8.0	6.0	18.0	mean	2.0	2.0	4.5																							
(A)	A	B	C																																	
sum	8.0	6.0	18.0																																	
mean	2.0	2.0	4.5																																	

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 3 頁，共 12 頁

	<div><div>(B)</div><table><tr><td></td><td>A</td><td>B</td><td>C</td></tr><tr><td>sum</td><td>8.0</td><td>6.0</td><td>18.0</td></tr><tr><td>mean</td><td>2.0</td><td>1.5</td><td>4.5</td></tr></table><div>(C)</div><table><tr><td></td><td>A</td><td>B</td><td>C</td></tr><tr><td>sum</td><td>10.0</td><td>6.0</td><td>18.0</td></tr><tr><td>mean</td><td>2.5</td><td>1.5</td><td>4.5</td></tr></table><div>(D)</div><table><tr><td></td><td>A</td><td>B</td><td>C</td></tr><tr><td>sum</td><td>10.0</td><td>6.0</td><td>18.0</td></tr><tr><td>mean</td><td>2.5</td><td>2.0</td><td>4.5</td></tr></table></div>		A	B	C	sum	8.0	6.0	18.0	mean	2.0	1.5	4.5		A	B	C	sum	10.0	6.0	18.0	mean	2.5	1.5	4.5		A	B	C	sum	10.0	6.0	18.0	mean	2.5	2.0	4.5
	A	B	C																																		
sum	8.0	6.0	18.0																																		
mean	2.0	1.5	4.5																																		
	A	B	C																																		
sum	10.0	6.0	18.0																																		
mean	2.5	1.5	4.5																																		
	A	B	C																																		
sum	10.0	6.0	18.0																																		
mean	2.5	2.0	4.5																																		
B	<div>8. 關於 R 語言常見的敘述性統計函式，下列何者代表標準差？</div> <div><div>(A) var()</div><div>(B) sd()</div><div>(C) diff()</div><div>(D) prod()</div></div>																																				
D	<div>9. 關於 Python 語言，敘述結尾應使用下列何種符號？</div> <div><div>(A) .</div><div>(B) ,</div><div>(C) :</div><div>(D) 無</div></div>																																				
D	<div>10. 關於遺缺值（Missing Value）處理的方式，下列敘述何者「不」正確？</div> <div><div>(A) 直接刪除含有遺缺值的資料或欄位</div><div>(B) 利用機器學習方法進行補值</div><div>(C) 常數（0/-1）或通用值（unknown）填補遺缺值</div><div>(D) 刪除整欄變數</div></div>																																				
D	<div>11. 關於降維的好處，下列敘述何者「不」正確？</div> <div><div>(A) 減少運算時間與儲存空間</div><div>(B) 移除共線性資料能有效提高線性模型的效能</div><div>(C) 當資料維度降至 2~3 維時，能很容易的直接視覺化展示資料分佈</div><div>(D) 降維後的資料集訊息量增加，不會減少</div></div>																																				
C	<div>12. 在進行資料分析時，經常會遇到類別型（Categorical）與數值型（Numerical）資料，關於這兩類型的資料處理方式，下列敘述何者「不」正確？</div> <div><div>(A) 數值型資料可透過平均值、變異數或畫出分配圖等方式觀察其特性</div><div>(B) 類別型資料可根據其類別轉換成對應數字，如 0, 1, 2, ...，以便</div></div>																																				

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 4 頁，共 12 頁

	<p>後續建構模型時直接使用</p> <p>(C) 數值型、類別型資料皆可使用直方圖 (Histogram) 觀察其分布狀況</p> <p>(D) 數值型資料標準化 (Standardization) 後，此資料的分佈圖圖形保持不變，只是標準差和平均數變為 1 與 0</p>
C	<p>13. 主成份分析 (Principal Component Analysis, PCA) 是一種常見的將資料降維擷取特徵的方法，關於主成份分析，下列敘述何者「不」正確？</p> <p>(A) 是一種線性降維度的方法</p> <p>(B) 是通過投影將高維度的資料映射到低維的空間中</p> <p>(C) 主成份分析的計算方式考慮誤差 (Error)，相關矩陣的對角線數值皆為 1</p> <p>(D) 可以基於資料共變異數 (Covariance) 矩陣進行計算</p>
B	<p>14. Sklearn (Scikit-learn) 是基於 Python 語言的機器學習工具，請問下列何者為 sklearn 預處理 (Preprocessing) 模組中，可以將各變數中心化與標準差化的類別函數？</p> <p>(A) RobustScaler()</p> <p>(B) StandardScaler()</p> <p>(C) MaxAbsScaler()</p> <p>(D) QuantileTransformer()</p>
B	<p>15. 關於類別型 (Categorical) 資料的處理方式，下列敘述何者正確？</p> <p>(A) 標籤編碼 (Label Encoding)：將類別由雜湊函數定應到一組數字，需調整雜湊函數對應值的數量</p> <p>(B) 單熱編碼 (One-hot Encoding)：主要是為了改良數字大小沒有意義的問題，將不同的類別分別獨立為一欄</p> <p>(C) 均值編碼 (Mean Encoding)：類似於流水號，依序將新出現的類別依序編上新代碼</p> <p>(D) 特徵雜湊 (Feature Hash)：取用兩個相關欄位，使用目標值的平均值，取代原本的類別型特徵</p>
A	<p>16. 下列對於分散式儲存 (Distributed Storage) 的敘述何者最「不」適當？</p> <p>(A) 方便維護資料</p> <p>(B) 易於擴充系統</p> <p>(C) 資料處理效率不一定較高</p> <p>(D) 避免單一機器故障影響系統運作</p>
D	<p>17. 下列何種方法無法透過單一次 MapReduce 實現？</p> <p>(A) 計算資料筆數</p>

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 5 頁，共 12 頁

	(B) 計算資料裡有幾個偶數 (C) 計算資料總和 (D) 計算資料平均數
A	18. 關於 Hadoop 與 Spark，下列敘述何者「不」正確？ (A) Spark 數據處理速度較 MapReduce 來的慢 (B) Spark 可佈署在 Hadoop Yarn（資源排程）上 (C) Spark 是一種分散式計算平台，使用 scala 語言編寫 (D) Hadoop 是一種具分散式管理、儲存、計算的生態系統
A	19. 關於 MySQL 資料庫進行資料操作時，下列何種作法較「不」恰當？ (A) 直接下 SQL 語法，對特定資料表一次刪除 6000 萬筆資料 (B) 若需要定期拉取大量歷史資料，可以將該工作排程設定在伺服器非尖峰用量時刻 (C) 建立新資料表（table）之前，妥善規劃各欄位（column）的資料格式 (D) 做好資料庫帳號權限管理，以維護資料庫安全
C	20. ETL 為數據分析工作中「獲取資料（Extract, E）→資料轉換（Transform, T）→資料存儲（Load, L）」的處理流程。關於使用 Python 環境進行 ETL 流程，下列敘述何者「不」正確？ (A) Extract：透過 pymongo 模組，連結後端之產品會員 MongoDB 資料庫，獲取近 30 天內註冊會員的基本資料 (B) Transform：透過 pandas 模組進行產品季銷售資訊處理，以彙整出各產品近 1 年的銷售成長率 (C) Load：透過 numpy 模組進行 csv 資料存讀取 (D) ETL：以 logging 模組建立工作日誌、並透過串接通訊軟體（如 Slack, Telegram）機器人 API，建立 ETL 工作進度監測機器人
B	21. 資料 A：6, 10, 9, 6, 2, 7，請問資料 A 的中位數為何？ (A) 7 (B) 6.5 (C) 6 (D) 8
B	22. 下列何種統計圖形可用於類別型資料分析（Categorical Data Analysis）？ (A) 直方圖（Histogram） (B) 長條圖（Bar Chart） (C) 散佈圖（Scatter Plot） (D) 莖葉圖（Stem-and-leaf diagram）

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 6 頁，共 12 頁

D	23. 關於盒鬚圖 (Box Plot)，下列何種統計量「不」會被顯示出來？ (A) 第一四分位數 (B) 中位數 (C) 第三四分位數 (D) 標準差
B	24. 關於算術平均數的特性，下列敘述何者「不」正確？ (A) 容易受到極端值影響 (B) 位於中央位置的資料值 (C) 與所有資料差異的總和為 0 (D) 與資料差異的平方和不一定等於 0
B	25. 「樣本或母體中出現次數最多，且大於 1 次的數值」，請問上述為下列何者數值之定義？ (A) 平均數 (B) 眾數 (C) 變異數 (D) 百分位數
C	26. 關於統計假設，下列敘述何者「不」正確？ (A) 虛無假設 (Null Hypothesis) 是研究者希望推翻之假設 (B) 虛無假設一般使用 $H_0$ 表示 (C) 事實上虛無假設為真，結果確拒絕虛無假設，此錯誤稱為第二型錯誤 (type II error) (D) 允許第一型錯誤的最大機率稱為顯著水準 (Significance Level)
D	27. 某研究人員想檢定 3 大品牌飲料的平均銷售量是否相同，於是隨機抽取 5 個地區進行調查，關於單因子變異數分析完全隨機設計，下列敘述何者正確？ (A) 處理方法 (3 大品牌飲料) 的自由度為 3 (B) 隨機誤差的自由度為 8 (C) 總和的自由度為 15 (D) f 檢定值的自由度為 $F(2, 12)$
C	28. 附圖為 8 位男性的身高 (cm) 與體重 (kg) 資料。請問若要比較身高資料與體重資料之分散程度，下列何種統計值較為恰當？ male_heights = [164, 188, 196, 180, 183, 162, 184, 192] male_weights = [65, 95, 87, 109, 95, 77, 74, 82] (A) 標準差 (Standard Deviation) (B) 四分位距 (Interquartile Range) (C) 變異係數 (Coefficient of Variation) (D) 算術平均數 (Arithmetic Mean)

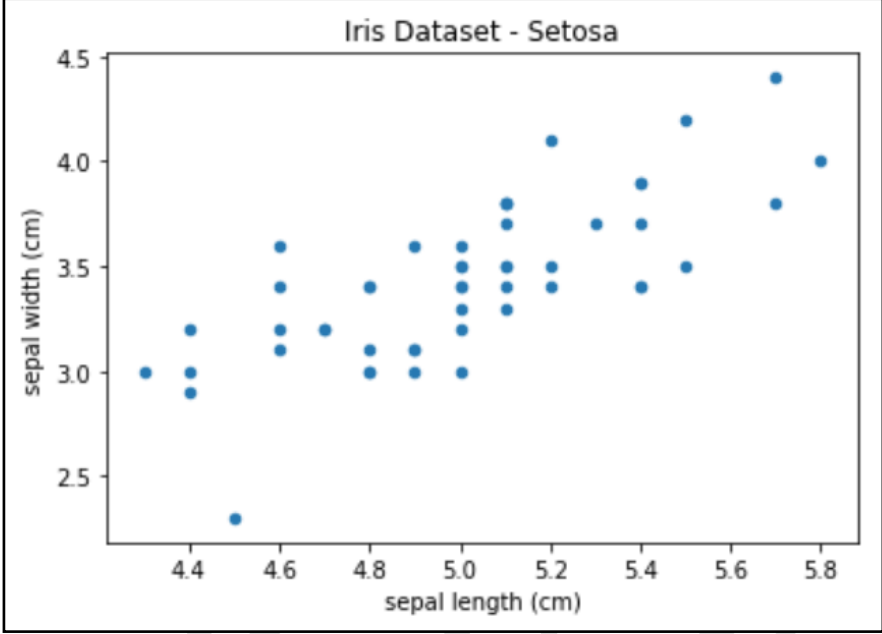


# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 7 頁，共 12 頁

D	<p>29. 附圖之鳶尾花資料集中，表示 setosa 花之各樣本萼片長度 (sepal length) 與萼片寬度 (sepal width) 點分佈圖。若計算其皮爾遜相關係數 (Pearson Correlation Coefficient) 所得到之值為 <math>r</math>，則其 <math>r</math> 值應較符合下列何者選項？</p>  <p>(A) 因資料點過少，無法計算其 <math>r</math> 值          (B) <math>r</math> 值為 0          (C) <math>r</math> 值小於 0          (D) <math>r</math> 值大於 0</p>
C	<p>30. 關於單一變量的 (Univariate) 統計量數，下列敘述何者正確？</p> <p>(A) 變異係數 (Coefficient of Variation) 不適用於量化變數          (B) 四分位距 (Inter-quartile Range) 可由類別變數的次數分佈進行計算          (C) 熵係數 (Entropy Coefficient) 可用於檢視類別變數次數分佈的異質性          (D) 異質性 (Heterogeneity) 最低時集中度 (Concentration) 最低；而異質性最高時集中度達到最高</p>
C	<p>31. 有一個數列 [1,2,3,4,5,7,20]，若找出此數列中的離群值 (Outlier)，下列計算何者「不」必要？</p> <p>(A) 計算此數列的平均數          (B) 計算此數列的標準差          (C) 計算此數列的峰度係數 (Kurtosis)          (D) 將此數列標準化</p>
D	<p>32. 有一調查想將受訪者分類，變項名稱與資料如附圖。請問若要使用 K-means 分類，下列計算何者「不」必要？</p>

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 8 頁，共 12 頁

	<p>(id, 身高, 體重, 年齡, 性別)</p> <p>[</p> <p>(1, 160, 60, 16, 'Male'),</p> <p>(2, 170, 67, 20, 'Male'),</p> <p>(3, 183, 55, 28, 'Female'),</p> <p>(4, 155, 46, 17, 'Male'),</p> <p>(5, 140, 70, 12, 'Female'), ...</p> <p>]</p> <p>(A) 將各變項標準化</p> <p>(B) 隨機挑出一筆資料做為初始值</p> <p>(C) 給定預計分類群組數</p> <p>(D) 計算自變項之間的相關性</p>
D	<p>33. 下列何者最容易受到離群值 (Outlier) 的影響？</p> <p>(A) 總合 (Sum)</p> <p>(B) 眾數 (Mode)</p> <p>(C) 中位數 (Median)</p> <p>(D) 平均數 (Mean)</p>
D	<p>34. 下列何者為非監督式學習 (Unsupervised Learning) 方法？</p> <p>(A) K 近鄰法 (K-nearest Neighbor)</p> <p>(B) 支援向量機 (Support Vector Machine)</p> <p>(C) 邏輯迴歸 (Logistic Regression)</p> <p>(D) K 平均法 (K-means)</p>
A	<p>35. 下列何者最適合使用集群分析 (Clustering Analysis) ？</p> <p>(A) 找出使用者行為特性</p> <p>(B) 預測股票走向</p> <p>(C) 進行天氣預測</p> <p>(D) 地震發生預測</p>
C	<p>36. 參考附圖，在 R 語言執行繪圖結果中，輸入哪一個函數，其繪圖會以 一列二行方式呈現？</p>

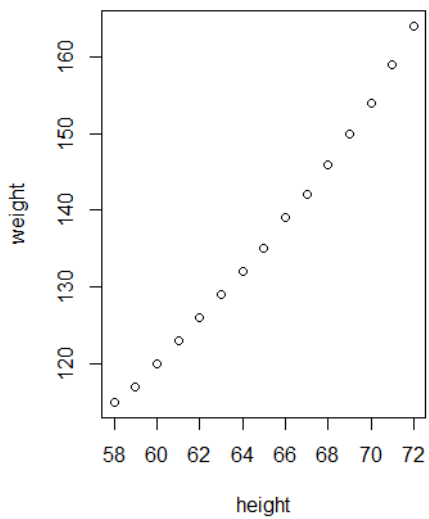
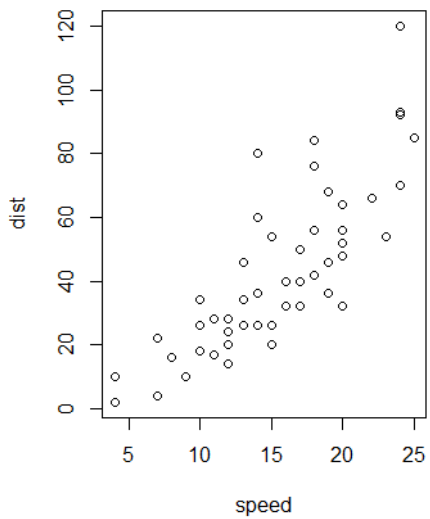


# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 9 頁，共 12 頁

	<div style="display: flex; justify-content: space-around;">   </div> <p>(A) <code>par(mar=c(1, 2))</code>            (B) <code>par(mai=c(2, 1))</code>            (C) <code>par(mfrow=c(1, 2))</code>            (D) <code>par(mfg=c(1, 2))</code></p>
C	<p>37. 關於空間密度集群算法 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN)，下列敘述何者「不」正確？</p> <p>(A) 屬於相同分群的資料會聚集在一個密度較高的區域內            (B) 不同的分群則是由一個密度較低的區域分隔            (C) 從密度點 (Density Point) 可到達的鄰近地區的點就被分成不同集群            (D) 如果某個點鄰近地區的點數小於可達區域的最小點個數 (Reachability Minimum Number of Points)，則此點為邊界點 (Border Point)</p>
A	<p>38. 關於機器學習 (Machine Learning)，下列敘述何者「不」正確？</p> <p>(A) 非監督式學習 (Unsupervised Learning) 運用的資料需被定義，因此資料需要事先標籤            (B) 階層式集群法 (Hierarchical Clustering) 屬於非監督式學習            (C) 迴歸分析 (Regression Analysis) 是屬於監督式學習            (D) 空間密度集群算法 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN) 不用事先提供集群個數</p>
C	<p>39. 若想探討學校前、中、後段班學生，其數學成績是否有顯著差異，應採用何種統計方法？</p> <p>(A) 相關分析 (Correlation Analysis)            (B) 迴歸分析 (Regression Analysis)</p>

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 10 頁，共 12 頁

	<p>(C) 變異數分析 (ANOVA)</p> <p>(D) 卡方檢定 (Chi-square Test)</p>
B	<p>40. 下列何者「不」屬於機器學習 (Machine Learning) ?</p> <p>(A) 強化學習 (Reinforcement Learning)</p> <p>(B) 非強化學習 (Un-Reinforcement Learning)</p> <p>(C) 監督式學習 (Supervised Learning)</p> <p>(D) 非監督式學習 (Un-Supervised Learning)</p>
D	<p>41. 下列何者「不」是決策樹 (Decision Tree) 產生的基本演算法?</p> <p>(A) ID3 (Iterative Dichotomiser)</p> <p>(B) C5.0</p> <p>(C) CART (Classification and Regression Trees)</p> <p>(D) 貝氏分類 (Bayesian Classification)</p>
D	<p>42. 關於集群分析 (Clustering Analysis)，下列敘述何者正確?</p> <p>(A) 組內平方和 (Within Sum of Squares, WSS) 越大越好</p> <p>(B) 組間平方和 (Between Sum of Squares, BSS) 越小越好</p> <p>(C) 總平方和 (Total Sum of Squares, TSS) 滿足 <math>TSS &gt; WSS + BSS</math></p> <p>(D) 陡坡圖 (Screen Plot) 的 Y 軸表示組內平方和，通常隨著集群個數增加而減少</p>
D	<p>43. 關於關聯規則 (Association Rule)，下列敘述何者正確?</p> <p>(A) 關聯規則的目標之一是找出小於最小支援度 (Support) 的規則</p> <p>(B) 關聯規則的目標之一是找出小於最小信賴度 (Confidence) 的規則</p> <p>(C) 關聯規則的 Apriori 演算法是假設項目集是頻繁項目集 (Frequent Itemset)，則包括它的子集合也不是頻繁項目集</p> <p>(D) 關聯規則的目標之一是找出提升度 (Lift) 大於 1 的規則</p>
B	<p>44. R 語言中，常使用 GLM (Generalized Linear Model) 進行廣義線性模型分析，如果需要修改預設模型，可設定 family 參數。依變數表示計算某事件發生的次數，則關於 family 參數的設定，下列敘述何者正確?</p> <p>(A) 參數 family = binomial()</p> <p>(B) 參數 family = poisson()</p> <p>(C) 參數 family = gaussian()</p> <p>(D) 參數 family = quasibinomial()</p>
B	<p>45. 關於強化學習 (Reinforcement Learning)，下列敘述何者「不」正確?</p> <p>(A) 透過不斷反覆正確與錯誤的學習，結果就會越來越精確</p>

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 11 頁，共 12 頁

	<p>(B) 我們可以標註一些資料，讓機器知道哪些是正確、那些是錯誤的</p> <p>(C) 強化學習必須因應環境的變動、隨之改變原有的作法</p> <p>(D) 機器透過每一次與環境互動來學習，以取得最大化的預期利益</p>									
C	<p>46. 關於訓練機器學習（Machine Learning）模型，下列敘述何者「不」正確？</p> <p>(A) 資料清理、特徵萃取、特徵選擇、選取方法都是重要的過程</p> <p>(B) 機器學習是實現人工智慧的其中一種方式</p> <p>(C) 假設檢定為機器學習的一種方法</p> <p>(D) 特徵萃取（Feature Extraction）是從資料中挖出可以用的特徵</p>									
D	<p>47. 下列何者「不」是常見的機器學習任務？</p> <p>(A) 迴歸（Regression）</p> <p>(B) 降維（Dimensionality Reduction）</p> <p>(C) 分類（Classification）</p> <p>(D) 因果（Causality）</p>									
D	<p>48. 如附圖所示，關於混淆矩陣（Confusion Matrix）與其延伸指標，下列敘述何者正確？</p> <table border="1"><tr><td></td><td>實際 YES</td><td>實際 NO</td></tr><tr><td>預測 YES</td><td>TP</td><td>FP</td></tr><tr><td>預測 NO</td><td>FN</td><td>TN</td></tr></table> <p>(A) 混淆矩陣常使用於線性迴歸模型，如預測下個月某地點房價數字</p> <p>(B) 準確率（precision）= <math>(TP + TN) / (TP + FP + FN + TN)</math></p> <p>(C) 預測正確率（Accuracy）適合用於所有模型，並不會出現指標失準的情況</p> <p>(D) 召回率（Recall）= <math>(TP) / (TP + FN)</math></p>		實際 YES	實際 NO	預測 YES	TP	FP	預測 NO	FN	TN
	實際 YES	實際 NO								
預測 YES	TP	FP								
預測 NO	FN	TN								
A	<p>49. 關於線性迴歸（Linear Regression）模型，下列敘述何者「不」正確？</p> <p>(A) 線性迴歸最多只能以 2 個自變數進行模型建立</p> <p>(B) 在使用線性迴歸模型之前，首先要確認資料分布上是否具有線性相關</p> <p>(C) 對於資料內的離群值（Outlier），可先進行排除、避免其造成模型失準</p> <p>(D) 線性迴歸模型屬於監督式學習的一種</p>									
C	<p>50. 關於監督式學習（Supervised Learning）模型，下列敘述何者「不」正確？</p> <p>(A) 可藉由觀察特徵（Features）間的相關係數，挑選出相關性較高的特徵</p>									

# 111 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：111 年 05 月 28 日

第 12 頁，共 12 頁

	<p>(B) 觀察某數值特徵的分佈直方圖 (Histogram)，確認是否有離群值</p> <p>(C) 監督式學習僅能用於建立連續數值資料 (Continuous Data) 之預測模型，無法用於類別資料 (Nominal Data) 之分類預測</p> <p>(D) 若資料中某類別型特徵的分類數量只有 5 個，可考慮透過單熱編碼 (One-hot Encoding) 將該類別型資料轉換成數值型資料</p>
--	--